Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise Visual de Dados Educacionais: Um Estudo de Caso das Disciplinas Introdutórias de Programação da UnB

Luiza Aguiar Hansen

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientadora
Prof.a Dr.a Maristela Terto de Holanda

Coorientador
Prof. Dr. Vinícius Ruela Pereira Borges

Brasília
2021

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise Visual de Dados Educacionais: Um Estudo de Caso das Disciplinas Introdutórias de Programação da UnB

Luiza Aguiar Hansen

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof.a Dr.a Maristela Terto de Holanda (Orientadora)
CIC/UnB

Prof. Dr. Thiago Paulo Faleiros
CIC/UnB

Prof.a Dr.a Dilma Da Silva
Texas A&M University

do Programa de Pós-graduação em Informática

Brasília, 15 de setembro de 2021

Dedicatória

Dedico este trabalho a minha família e amigos que me apoiaram nas horas mais difíceis e sorriram comigo nas horas mais alegres. Sem eles, este trabalho e muitos dos meus sonhos não se realizariam.

Agradecimentos

Gostaria de agradecer aos orientadores, Professora Doutora Maristela Holanda e Professor Doutor Vinícius Borges, pelo auxílio e disponibilidade para auxiliar no desenvolvimento da pesquisa e da escrita deste trabalho. Ao SIGRA pelo material disponibilizado, essencial para a realização do estudo e a todos os professores que se disponibilizaram a responder o questionário de avaliação dos algoritmos de visualização. Agradeço também a todos que me acompanharam nessa trajetória, principalmente aqueles que contribuíram na revisão do texto: Tereza Cristina de Melo Aguiar, Inácio Cauduro Hansen, Wânia Mara de Melo Aguiar e Rafael Aires de Alencar Lucas da Silva.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

As técnicas de visualização auxiliam na análise e na compreensão de conjuntos de dados, de forma a evidenciar características e padrões de informações. Na literatura, a análise de dados educacionais é empregada em tarefas como predição de performance e identificação de perfis dos estudantes e monitoramento de sistemas educacionais com o objetivo de contribuir para a melhoria da qualidade de ensino. Nas disciplinas introdutórias de computação, evidencia-se o alto número de reprovação e abandono de alunos, o que caracteriza um cenário apropriado para estudo e investigação utilizando técnicas de análise e visualização de dados.

Este trabalho estuda e avalia os algoritmos de visualização para que professores e gestores educacionais possam tomar decisões. Os algoritmos foram aplicados em um estudo de caso de três disciplinas introdutórias de computação da UnB, levando em consideração fatores sociais e acadêmicos dos alunos. As visualizações foram avaliadas a partir de um questionário aplicado a professores e gestores educacionais. A partir do resultado, observou-se que os questionados se sentiram mais seguros ao utilizar algoritmos já conhecidos, como o gráfico de pizza e o gráfico de barras. Dentre os selecionados, o diagrama de sankey, o *treemap* e o gráfico de violino eram os menos conhecidos pelos questionados. Ademais, o gráfico de barras foi o algoritmo em que as informações foram identificadas de forma mais rápida e correta. Por fim, de forma a extrair conhecimento das visualizações selecionadas, foram utilizados os dados dos alunos referentes ao gênero, à forma de entrada no curso e à disciplina cursada.

Palavras-chave: visualização de dados, questionário, disciplinas introdutórias, UnB

Abstract

Data visualization techniques supports analysis and understanding of datasets, which emphasize the observation of characteristics and patterns. In the literature, the analysis of educational data is used in tasks such as performance prediction and identification of student profiles, as also monitoring educational systems in order to improve the quality of education. In introductory computing subjects, there is a high number of students who fail and dropout the courses, which makes it an appropriate scenario for using data analysis and visualization techniques.

This paper studies visualization algorithms in order to help teachers and educational managers to make decisions that support the learning. The algorithms were applied in a case study of three introductory computing subjects at UnB and were evaluated through a questionnaire applied to teachers and educational managers. The results show that the respondents felt more secure when using familiar algorithms, such as pie chart and bar chart. Among the selected visualization, sankey chart, treemap, and violin chart were the least known by the respondents. Furthermore, the bar chart was the algorithm where the information was identified quickly and correctly most of the time. Finally, in order to extract knowledge from the selected visualizations, student data regarding gender, entrance form to university and subject studied were analysed.

Keywords: data visualization, survey, introductory courses, UnB

Sumário

1 Introdução **1**
1.1 Objetivos . . . 2
1.2 Estrutura do Documento . . . 2

2 Fundamentação Teórica **4**
2.1 Estudo de Dados . . . 4
2.1.1 Desafio dos Dados Educacionais . . . 5
2.1.2 Fundamentos de Estatística . . . 5
2.2 Técnicas de Visualização . . . 7
2.2.1 Visualização de Dados . . . 8
2.2.2 Visualização de Informação . . . 11
2.2.3 Visualizações não Classificadas por Lengler & Eppler . . . 12

3 Revisão da Literatura **16**
3.1 Método de Pesquisa: Processo de Mapeamento Sistemático . . . 16
3.2 Trabalhos Relacionados . . . 18
3.3 Considerações Finais . . . 20

4 Metodologia **22**
4.1 Definição do Conjunto de Dados . . . 22
4.1.1 Dados do SIGRA . . . 23
4.1.2 Dados do Questionário . . . 29
4.2 Estudo das Visualizações . . . 31
4.3 Aplicação do Estudo de Caso . . . 33
4.3.1 Domínio de Informações Gerais . . . 33
4.3.2 Domínio de Informações Acadêmicas . . . 39
4.3.3 Domínio de Informações de Percepção . . . 46
4.4 Realização de Experimentos . . . 55
4.5 Considerações finais . . . 55

5 Questionário **56**
5.1 Questões Iniciais 56
5.2 Questões Elaboradas 57
5.2.1 Domínio de Informações Gerais 57
5.2.2 Domínio de Informações Acadêmicas 60
5.2.3 Domínio de Informações de Percepção 64
5.3 Resultados 68
5.3.1 Informações Gerais 68
5.3.2 Informações Acadêmicas 69
5.3.3 Informações de Percepção 70
5.4 Considerações Finais 70

6 Aplicação do Resultado **71**
6.1 Domínio de Informações Gerais 71
6.2 Domínio de Informações Acadêmicas 76
6.3 Domínio de Informações de Percepção 79

7 Conclusão **84**

Referências **87**

Apêndice **91**

A Informações Detalhadas dos Dados do Formulário **92**

Lista de Figuras

2.1 Algoritmos da categoria visualização de dados. 10
2.2 Algoritmos da categoria visualização de informação. 13
2.3 Visualizações não classificadas por Lengler & Eppler. 15

3.1 Metodologia do processo de mapeamento sistemático. 17
3.2 Trabalhos de acordo com o nível educacional informado. 19

4.1 Metodologia do projeto. 22
4.2 Modelo do banco de dados relacional utilizando informações do SIGRA. . . 24
4.3 Tabela para armazenar as respostas do questionário. 30
4.4 Visualizações da Pergunta 1 do domínio de informações gerais. 34
4.5 Visualizações da Pergunta 2 do domínio de informações gerais. 35
4.6 Visualizações da Pergunta 3 do domínio de informações gerais. 36
4.7 Visualizações da Pergunta 4 do domínio de informações gerais. 39
4.8 Visualizações da Pergunta 1 do domínio de informações gerais. 41
4.9 Visualizações da Pergunta 2 do domínio de informações acadêmicas. 42
4.10 Visualizações da Pergunta 3 do domínio de informações acadêmicas. 43
4.11 Visualizações da Pergunta 4 do domínio de informações acadêmicas. 44
4.12 Visualizações da Pergunta 5 do domínio de informações acadêmicas. 46
4.13 Visualizações da Pergunta 1 do domínio de informações de percepção. . . . 48
4.14 Visualizações da Pergunta 2 do domínio de informações de percepção. . . . 49
4.15 Visualizações da Pergunta 3 do domínio de informações de percepção. . . . 50
4.16 Visualizações da Pergunta 4 do domínio de informações de percepção. . . . 51
4.17 Visualizações da Pergunta 5 do domínio de informações de percepção. . . . 52
4.18 Visualizações da Pergunta 6 do domínio de informações de percepção. . . . 54

6.1 Comparação entre a quantidade de alunos por curso e as respectivas notas 72
6.2 Quantidade de aprovações e reprovações. 73
6.3 Semestre em que os alunos cursam a disciplina. 74
6.4 Quantidade de vezes que os alunos cursam a disciplina. 75

6.5 Evolução das notas dos alunos. 77
6.6 Métricas das notas dos alunos que não ingressaram na universidade por meio do sistema de cotas. 77
6.7 Métricas das notas dos alunos que ingressaram na universidade por meio do sistema de cotas. 77
6.8 Evolução da quantidade de aprovação dos alunos que não ingressaram na universidade por meio do sistema de cotas. 78
6.9 Evolução da quantidade de aprovação dos alunos que ingressaram na universidade por meio do sistema de cotas. 78
6.10 Evolução da quantidade de créditos dos alunos que não ingressaram na universidade por meio do sistema de cotas. 78
6.11 Evolução da quantidade de créditos dos alunos que ingressaram na universidade por meio do sistema de cotas. 79
6.12 Comparação entre os alunos que não ingressaram na universidade por meio do sistema de cotas e que fizeram a disciplina pela primeira vez com os que cursaram novamente. 79
6.13 Comparação entre os alunos que ingressaram na universidade por meio do sistema de cotas e que fizeram a disciplina pela primeira vez com os que cursaram novamente. 80
6.14 Questões sobre pertencimento do questionário aplicado aos alunos de APC. 80
6.15 Relação entre a área de trabalho dos pais com o nível de ensino dos mesmos. 81
6.16 Informação sobre se os alunos tiveram contato prévio com programação antes da disciplina. 82
6.17 Quais as linguagens de programação mais conhecidas pelos alunos. 82
6.18 Relação entre quantidade de horas de estudo fora da disciplina com a percepção de que precisam de mais horas de estudos que os colegas. 83

Lista de Tabelas

4.1 Resumos dos atributos da tabela “Disciplina”. 25
4.2 Resumos dos atributos da tabela “Aluno”. 26
4.3 Resumos dos atributos da tabela “Aluno_Curso”. 28
4.4 Resumos dos atributos da tabela “Alunos_Curso_Disciplina”. 29
4.5 Número de publicações que utiliza cada algoritmo de visualização. 31
4.6 Análise de cenário dos algoritmos de visualização estudados. 32
4.7 Informações do domínio de informações gerais. 34
4.8 Resultado do domínio de informações acadêmicas. 40
4.9 Resultado do domínio de informações de percepção. 47

5.1 Nível de conhecimento dos questionados sobre as visualizações. 57
5.2 Resultado dos algoritmos da Pergunta 1 do domínio de informações gerais. 58
5.3 Resultado dos algoritmos da Pergunta 2 do domínio de informações gerais. 58
5.4 Resultado dos algoritmos da Pergunta 3 do domínio de informações gerais. 59
5.5 Resultado dos algoritmos da Pergunta 4 do domínio de informações gerais. 60
5.6 Resultado dos algoritmos da Pergunta 1 do domínio de informações acadêmicas. 61
5.7 Resultado dos algoritmos da Pergunta 2 do domínio de informações acadêmicas. 62
5.8 Resultado das questões elaboradas da Pergunta 3 do domínio de informações acadêmicas. 63
5.9 Resultado dos algoritmos da Pergunta 4 do domínio de informações acadêmicas. 63
5.10 Resultado dos algoritmos da Pergunta 5 do domínio de informações acadêmicas. 64
5.11 Resultado dos algoritmos da Pergunta 1 do domínio de informações de percepção. 65
5.12 Resultado dos algoritmos da Pergunta 2 do domínio de informações de percepção. 66

5.13 Resultado dos algoritmos da Pergunta 3 do domínio de informações de percepção. 66
5.14 Resultado dos algoritmos da Pergunta 4 do domínio de informações de percepção. 67
5.15 Resultado dos algoritmos da Pergunta 5 do domínio de informações de percepção. 67
5.16 Resultado dos algoritmos da Pergunta 6 do domínio de informações de percepção. 68

Capítulo 1

Introdução

Com o avanço da tecnologia na era da informação, observa-se o crescimento da demanda por cursos de graduação na área de computação [1], apesar da programação de computadores não ser matéria ensinada com frequência em escolas do ensino médio. No entanto, programar exige o entendimento da sequência de ações dos algoritmos e suas respectivas consequências. Assim, uma vez que a maioria dos alunos entra em contato com programação apenas na universidade, muitos apresentam dificuldade e reprovam nas disciplinas introdutórias. Somado a isso, problemas relacionados à cultura, como forma de entrada, fatores financeiros, gênero e até mesmo o curso, podem influenciar na disposição dos alunos, impactando diretamente o sucesso ou o insucesso nas primeiras disciplinas de programação [2]. Isto posto, a média dos alunos que abandonam tais disciplinas, durante o ensino superior, no mundo é de 30% [3] [4].

Dessa forma, as pesquisas no âmbito acadêmico têm crescido nos últimos anos, concentrando-se em descobrir e propor novos métodos, técnicas e procedimentos conducentes ao aprendizado. Essas pesquisas enfocam no planejamento e gerenciamento dos cursos, além do fornecimento de soluções alternativas para superar desafios e dificuldades na estrutura educacional, de forma a melhorar o sistema de ensino [5].

Os dados para as pesquisas educacionais podem ser obtidos por meio de registros escolares ou de sistemas desenvolvidos para auxiliar a educação dos alunos, como: sistemas de gerenciamento de aprendizagem (LMS - *Learning Management System*), de tutoria inteligente ou de ensino à distância [6]. Entretanto, a quantidade de dados aumenta continuamente, o que dificulta processos de análise manual pelos especialistas. Dessa forma as técnicas de mineração de dados se fazem necessárias para o estudo e exploração das bases, em busca de padrões e conhecimentos implícitos [7] [8].

No entanto, os padrões extraídos por meio das técnicas de mineração podem não se apresentar de maneira intuitiva, assim como os resultados podem não revelar padrões de interesse. Alternativamente, a visualização de dados tem como objetivo represen-

tar graficamente informações de um determinado domínio de aplicação, de forma que a representação visual gerada explore a capacidade de percepção humana, facilitando a interpretação e a compreensão das informações apresentadas, além de gerar novos conhecimentos [9] [10]. De modo geral, essa técnica auxilia a análise e o entendimento de um conjunto de dados, por meio dos mecanismos de interação e de geração de representações gráficas, que evidenciam a observação de características e padrões nos dados.

O emprego de técnicas de visualização em análise de dados proporciona diversas vantagens ao usuário na descoberta de conhecimento, tais como [11] [12] [13]: identificar grandes quantidades de dados de modo compreensível e coeso; agilizar o entendimento de informações pelo uso da visão, que é o sentido humano que possui maior capacidade de captação de informações por unidade de tempo; facilitar o reconhecimento de padrões e de relações, possibilitando a exploração de valores atípicos; e contribuir para o envolvimento do usuário ao transmitir uma mensagem, que pode gerar impacto, funcionar como extensão da memória humana e ainda auxiliar no processo cognitivo.

1.1 Objetivos

Esse trabalho tem como objetivo o estudo de técnicas de visualização no contexto educacional com foco nas primeiras disciplinas de programação oferecidas pelo Departamento de Ciência da Computação da UnB, de forma a facilitar a descoberta de conhecimento por parte dos professores e gestores educacionais. Este estudo leva em consideração fatores sociais e acadêmicos dos alunos.

Com a finalidade de alcançar o objetivo geral deste trabalho, os seguintes objetivos específicos foram definidos:

1. Delimitação dos dados educacionais a serem utilizados, incluindo a forma de armazenamento e suas características;
2. Exploração das técnicas de visualização mais apropriadas para os dados educacionais disponíveis;
3. Geração das visualizações a partir dos dados do estudo de caso;
4. Validação das visualizações, de forma a determinar se as visualizações selecionadas comunicam a informação de forma clara e coesa.

1.2 Estrutura do Documento

Este documento é composto pelos seguintes capítulos:

- Capítulo 2: apresenta os conceitos básicos utilizados neste trabalho, enfatizando conceitos relacionados aos dados educacionais, técnicas estatísticas e os algoritmos de visualização utilizados no projeto;
- Capítulo 3: discorre sobre o estado da arte, incluindo o método de pesquisa utilizado para obter os artigos existentes na literatura relacionados ao tema deste projeto, além de expor os trabalhos que justificam a escolha do tema;
- Capítulo 4: apresenta a metodologia seguida para a execução deste projeto, incluindo um resumo dos dados selecionados, o estudo das visualizações, a validação dos algoritmos aplicados ao estudo de caso e por fim a realização de experimentos;
- Capítulo 5: expõe os resultados alcançados a partir da aplicação do questionário, com o objetivo de validar as visualizações selecionadas;
- Capítulo 6: aborda a descoberta de conhecimento utilizando os algoritmos selecionados em três assuntos distintos - questão de gênero na computação, a entrada na universidade através das cotas e a percepção dos alunos em cada disciplina.
- Capítulo 7: conclui o trabalho ao sumarizar as atividades realizadas, descrever as principais limitações e contribuições do projeto, bem como trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo revisa os fundamentos teóricos utilizados para o estudo das técnicas de visualização aplicadas no contexto educacional. A Seção 2.1 aborda os principais conceitos relacionados aos dados, bem como as particularidades e os desafios no manuseio das informações educacionais. Além disso, nessa seção são abordados conceitos de estatística para análise de dados. Por fim, a Seção 2.2 conceitua visualização de informação e seus processos, como também apresenta os algoritmos de visualização.

2.1 Estudo de Dados

No campo da tecnologia da informação, os dados são símbolos que representam as propriedades de objetos e eventos e a informação consiste em dados processados, cujo processamento visa aumentar a sua utilidade [14]. O dado é mensurável, coletado, reportado e analisado, sendo possível visualizá-lo em gráficos, imagens ou qualquer ferramenta de análise.

No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é responsável pelo levantamento e divulgação dos dados no contexto educacional, com o intuito de contribuir para o desenvolvimento econômico e social do país [15] [16]. Dados educacionais podem ser coletados por meio do censo escolar realizado pelo INEP ou por meio de ferramentas da própria escola, sendo armazenados em Banco de Dados (BD).

Bancos de dados são definidos como coleções de dados inter-relacionados que representam informações de um domínio específico [17]. Dentre as várias formas de representação, o modelo relacional é muito utilizado, consistindo em uma matriz bidimensional simples composta por elementos. Sua vantagem é a fácil inserção e recuperação prática de conjuntos de dados. Um conjunto de dados $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$ é definido por N instâncias multidimensionais, em que cada instância $\mathbf{x_i} = \{x_{i,1}, ...x_{i,m}\}$ é descrita por m atributos,

sendo interpretada como um vetor m-dimensional, de forma que o conjunto de dados $\mathbf{X}$ define um espaço m-dimensional [18].

No modelo relacional, baseado em tabelas, as instâncias estão associadas com as linhas e os atributos com as colunas. Além disso, cada instância é associada a uma chave primária (PK - *primary key*), essa chave é o identificador único de um registro. Já a chave estrangeira (FK - *foreign key*) é utilizada no relacionamento entre tabelas, de forma a referenciar registros externos.

2.1.1 Desafio dos Dados Educacionais

Dados educacionais são utilizados no monitoramento dos sistemas educacionais, considerando o acesso, a permanência e a aprendizagem dos alunos. Dessa forma, as pesquisas com dados educacionais contribuem para a criação de políticas públicas voltadas para a melhoria da qualidade da educação e dos serviços oferecidos à sociedade pela escola [15]. No entanto, é necessário entender os desafios impostos pelo uso desses dados, a fim de tornar as pesquisas mais eficientes.

As dificuldades se apresentam desde a obtenção dos dados, que exibem, geralmente, pouca similaridade entre os campos, resultante da escassa padronização entre as instituições de ensino. Ademais, quando obtidos, frequentemente sofrem perda de informações decorrente de mapeamentos imprecisos e de sua baixa qualidade, levando à informações incorretas, incompletas ou inesperadas [19]. Além disso, cabe ressaltar a preocupação com a exposição de dados e divulgação de informações sigilosas [20], principalmente no contexto educacional, em que grande parte das informações são de jovens.

Por fim, os dados educacionais apresentam como característica o alto número de atributos e instâncias com diferentes granularidades [21], o que exige uma seleção das variáveis para melhorar a performance e simplificar os algoritmos, além de reduzir o custo computacional e fornecer um melhor entendimento sobre os resultados encontrados.

2.1.2 Fundamentos de Estatística

A análise de dados é a atividade de transformar um conjunto de dados com o objetivo de obter conhecimento. Nesse sentido, são aplicadas técnicas estatísticas para mensurar comportamentos, tendências, variações e resultados. Esta seção faz uma breve abordagem teórica dos fundamentos da estatística.

A expectativa matemática de uma variável aleatória X fornece um valor único que atua como representante ou média dos valores de X, sendo chamado, por essa razão, de medida de tendência central [22]. A média mescla, de maneira mais uniforme, os valores

mais baixos e os mais altos de uma lista, sendo usada em outros cálculos de probabilidade, tais como desvio padrão, covariância e correlação.

A expectativa de um conjunto de dados X é determinada por μ_X, ou apenas μ. Para uma variável aleatória discreta com N valores $x_1, ..., x_n$, o valor esperado, μ, é definido pela Equação (2.1). Já o grau de dispersão entre os dados é calculado por meio do desvio padrão (σ), configurado através da Equação (2.2).

$$\mu = \mathrm{E}[X] = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2.2}$$

O desvio padrão varia de 0 a 1, em que um valor próximo a 1 indica que os dados estão mais espalhados e um valor mais baixo assinala dados mais homogêneos que tendem a estar próximos da média ou do valor esperado.

A covariância, *Cov(X,Y)*, é determinada como a medida da variabilidade em conjunto com o grau de associação [23]. A covariância pode ser definida a partir da Equação (2.3), em que X e Y são dois conjuntos de dados distintos com valores $x_1, x_2, ..., x_N$ e $y_1, y_2, ..., y_N$, respectivamente. Além disso, $p(x_i, y_i)$ é a probabilidade de ocorrer o par (x_i, y_i) e μ_i^{var} é a média para os valores da variável $var \in \{x, y\}$.

$$Cov(X, Y) = \sum_{i=1}^{N}[(x_i - \mu_i^x)(y_i - \mu_i^y)p(x_i, y_i)] \tag{2.3}$$

Quando o valor de X for maior do que a média e o valor de Y também tender a ser maior, a covariância será positiva, uma vez que as variáveis aleatórias são associadas positivamente. Em contrapartida, quando X é maior do que μ_x, mas Y tende a ser menor do que a média, e vice-versa, a covariância será negativa, sendo as variáveis aleatórias associadas negativamente.

Por último, a correlação é uma técnica estatística que busca indicar a força e a direção na relação entre duas ou mais variáveis [24]. Essas variáveis estão relacionadas quando a ocorrência de mudanças no valor de uma provoca alterações no valor da outra. O coeficiente de correlação utilizado para otimização da análise de dados é o "coeficiente de correlação de Pearson", denotado por r, também chamado de produto-momento, calculado a partir da Equação (2.4), em que $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_n$ são os valores medidos das variáveis X e Y, assim como μ_x e μ_y são as médias respectivas.

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)} * \sqrt{\sum_{i=1}^{N}(y_i - \mu_y)}} \tag{2.4}$$

O coeficiente é avaliado em escala sem unidade e pode conter valores desde -1, passando por 0 e até +1. Valores de r maiores do que 0 indicam correlação positiva, à medida que valores de r menores do que 0 apresentam correlação negativa. Caso o valor seja de r seja igual a 0, então a correlação é neutra. Assim, se o coeficiente for positivo, à medida em que um valor x cresce ou decresce, o valor y varia no mesmo sentido. Se o coeficiente for negativo, à medida em que um valor x cresce, o valor y decresce e vice-versa. Por fim, se o coeficiente for neutro, não existe nenhuma associação linear [25]. Ao definir o coeficiente de correlação, é possível utilizá-lo em diversas técnicas de visualização tradicionais, gerando uma análise voltada para a relação de variação entre os fatores.

2.2 Técnicas de Visualização

A visualização de dados consiste na comunicação de informações utilizando representações gráficas [26]. Khan & Khan em [27] delimitaram as técnicas de visualização em três principais finalidades: visualização científica, visualização de software e visualização da informação. A visualização científica tem a finalidade de facilitar o entendimento dos fenômenos físicos dos dados, modelos matemáticos, entre outros. Já a visualização de software auxilia na aprendizagem e visualização do funcionamento de programas, o que facilita o entendimento de sistemas complexos. Por último, a visualização da informação representa graficamente informações, de forma a facilitar comparações, reconhecimentos de padrões e detecção de alterações, ou seja, a exibição de informações por meio de imagens, figuras, estruturas gráficas e qualquer outro tipo de diagrama. Esta classificação chama a atenção do usuário para as características dos dados e assim potencializa a apropriação da informação [28] [29] [30].

A utilização de técnicas de visualização de informação permite reunir um maior número de informações em apenas uma imagem, auxiliando na interpretação e descoberta de conhecimento. Desta forma, com o crescimento da geração de dados, têm-se demandado o uso frequente de algoritmos de visualização que possibilitam a exploração visual progressiva e iterativa, contribuindo para o entendimento e a análise dos dados [31]. Desta forma, novas técnicas são criadas e aprimoradas de forma a solucionar problemas específicos de visualização [32].

As técnicas de visualização de informação foram organizadas por Lengler & Eppler [33] de forma a agrupar os principais métodos em seis categorias: visualização de dados, visu-

alização de informação, visualização de conceito, visualização de estratégia, visualização de metáforas e visualização composta. As técnicas de visualização de dados incluem as representações visuais ou quantitativas de dados na forma esquemática. A visualização de informações implica em algoritmos interativos com objetivo de aumentar a cognição, ou seja, dados são transformados em imagens dentro de um espaço na tela. A visualização conceitual abrange os métodos para elaborar conceitos qualitativos, ideias, planos e análises. A visualização de estratégia trata das visualizações complexas de análise, desenvolvimento, formulação, comunicação e implementação de estratégias organizacionais. A visualização de metáforas promove a organização e a estruturação das informações, gerando uma percepção sobre a representação de informações através de características chaves da metáfora empregada. Finalmente, a visualização composta apresenta diferentes métodos gráficos em um esquema.

A Seção 2.2.1 apresenta e explicita os algoritmos de visualização de dados, enquanto que a Seção 2.2.2 expõe os algoritmos de visualização de informações. Além das duas categorias, foram estudados os algoritmos de visualização que não são incluídos na divisão de Lengler & Eppler, sendo estes detalhados na Seção 2.2.3. Em cada seção são apresentados exemplos dos algoritmos de visualização, nos quais foram utilizados os dados educacionais obtidos do SIGRA (Sistemas Acadêmicos da Universidade de Brasília) para geração dos gráficos.

2.2.1 Visualização de Dados

Segundo a classificação de algoritmos de Lengler & Eppler [33] a visualização de dados contém representações visuais de dados quantitativos em forma esquemática. Os algoritmos deste grupo são usados para todas as finalidades, mas principalmente para obter uma visão geral dos dados.

- **Gráfico de Pizza (*Pie Chart*)** - O objetivo deste algoritmo é descrever a proporção numérica, ou seja, dada uma frequência relativa f_i, será apresentado uma fatia de circunferência de $360 * f_i$ associado a observação i. A vantagem do gráfico de pizza é que este permite comparações de frequências em relação a mesma variável, além de resumir os resultados em percentuais e enfatizar descobertas gerais. Entretanto, este gráfico não permite visualizar comparações ou evoluções temporais. A Figura 2.1a apresenta um exemplo do gráfico de pizza.
- **Gráfico de Barras (*Bar Chart*)** - Este algoritmo representa a frequência absoluta ou relativa das observações, que podem ser categóricas ou numéricas. Todavia, o gráfico de barras não costuma ser aplicado em dados contínuos. A preeminência deste algoritmo situa-se nas comparações diretas, possibilitando comparações

rápidas, além de economizar espaço na apresentação. Esta visualização não deve ser utilizada quando houver muitos grupos a serem comparados ou quando a variável representada possuir muitas categorias. O gráfico de barras é representado na Figura 2.1g.

- **Gráfico de Linha (*Line Chart*)** - Este algoritmo exibe informações como uma série de pontos ligados por segmentos de linhas e é frequentemente utilizado para dados temporais e para medir a evolução dos dados. O gráfico de linha é mais eficaz do que o gráfico de barras ao apresentar cinco ou mais pontos, porém, é menos eficaz ao enfatizar as diferenças em poucos períodos. A Figura 2.1c apresenta o gráfico de linha.

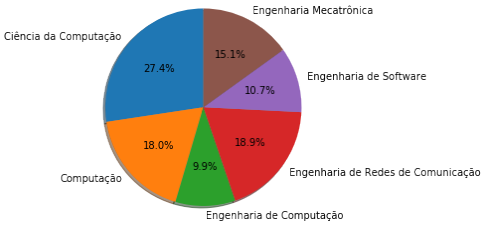
- **Gráfico de Área (*Area Chart*)** - Esta visualização é semelhante ao gráfico de linha na forma com que os dados são plotados e conectados. Contudo a área entre a linha e o eixo X é preenchida com cores para enfatizar a magnitude. O gráfico de área possibilita a visualização de tendências e de mudanças de volume ao longo do tempo, fornecendo uma imagem clara e concisa do direcionamento do desempenho de cada grupo. Entretanto, ler valores exatos de um gráfico de área não é uma tarefa simples, desta forma, ao renderizar diversas categorias, o entendimento dos dados é dificultado. A Figura 2.1d exemplifica o gráfico de pizza.

- **Histograma** - Também é conhecido como Diagrama de Dispersão de Frequências e consiste em uma representação gráfica de dados divididos em classes, com o objetivo de conferir como um processo se comporta em relação a suas especificidades. Assim, o diagrama permite visualizar de que forma os dados se distribuem pelos diversos valores observados. O gráfico de histograma é apresentado na Figura 2.1e.

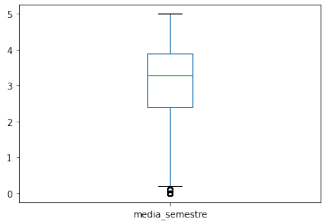
- **Gráfico de dispersão (Scatterplot)** - Coordenadas cartesianas são utilizadas para exibir valores de um conjunto de dados em duas dimensões. Os dados são exibidos como uma coleção de pontos, em que cada ponto é marcado por duas variáveis (x,y). O valor de uma variável determina a posição no eixo X e o valor de outra determina a posição no eixo Y. Desta forma, o gráfico de dispersão é utilizado em conjuntos de dados emparelhados, de forma a visualizar as correlações entre dois conjuntos. O gráfico de dispersão é ilustrado na Figura 2.1f.

- **Gráfico de Caixa (*Box-plot*)** - É utilizado para avaliar a distribuição empírica dos dados, apresentando informações sobre os quartis, o limite superior e inferior, a localização, a dispersão, assimetria, comprimento da cauda e *outliers*. Porém, não é uma visualização popular e a falta de conhecimento pode levar a conclusões erradas. Além disso, este gráfico não é recomendado para amostras menores que

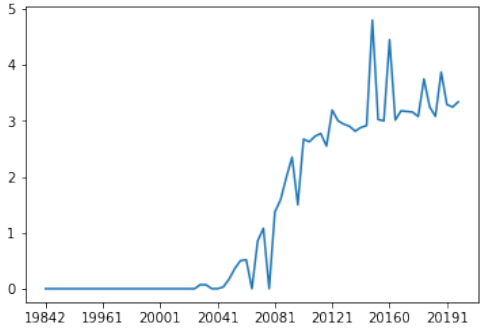
10, pois a ferramenta utiliza 5 medidas de dados, o que deixaria o gráfico de caixas pouco informativo. A Figura 2.1b apresenta um exemplo desta visualização.



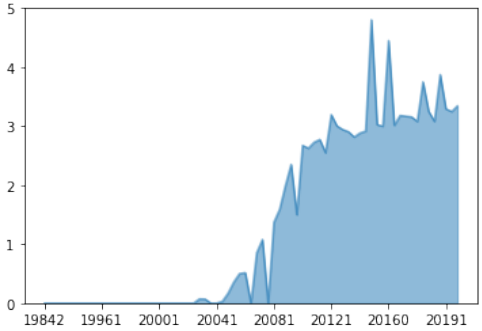
(a) Exemplo utilizando gráfico de pizza.



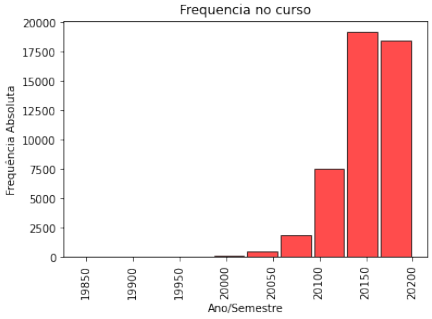
(b) Exemplo utilizando gráfico de caixa.



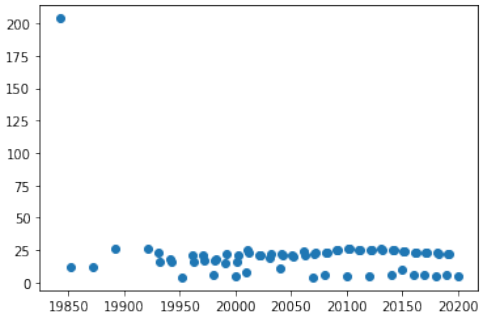
(c) Exemplo utilizando gráfico de linhas.



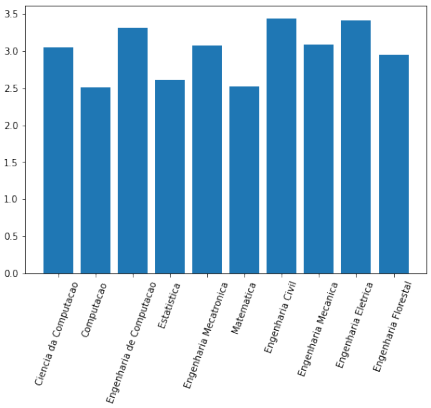
(d) Exemplo utilizando gráfico de área.



(e) Exemplo utilizando histograma.



(f) Exemplo utilizando gráfico de dispersão.



(g) Exemplo utilizando gráfico de barra.

Figura 2.1: Algoritmos da categoria visualização de dados.

2.2.2 Visualização de Informação

A visualização de informações contém as representações visuais interativas, ou seja, os dados são mapeados para o espaço de tela e a imagem pode ser alterada, aumentando a cognição dos usuários.

- **Gráfico de Radar (*Radar Chart*)** - Este método é utilizado para visualização de dados multivariados na forma de um gráfico bidimensional de três ou mais variáveis quantitativas representadas em eixos a partir do mesmo ponto. O gráfico de radar é adequado para selecionar *outliers* e semelhanças. Todavia, este gráfico utiliza ângulos, o que torna a posição pouco informativa, pelo fato de o ser humano não estar acostumado a comparar visualmente o comprimento de diferentes raios. A Figura 2.2a apresenta um exemplo do gráfico de radar.
- **Coordenadas Paralelas (*Parallel Coordinates*)** - O algoritmo é utilizado para comparar os valores de dados que são de tipos ou magnitudes completamente diferentes em uma única visualização. No gráfico de coordenadas paralelas, cada eixo vertical está associado a uma variável e as instâncias de dados são representadas por poli-linhas que interceptam esses eixos em uma posição determinada pelo valor do atributo associado ao eixo. O gráfico de coordenadas paralelas é indicado para identificar valores discrepantes ou padrões com base em fatores de métricas relacionadas e localizar pontos de cruzamento. Porém, cada eixo pode ter, no máximo, dois vizinhos, limitando o número de variáveis a serem comparadas. Este gráfico é apresentado na Figura 2.2g.
- **Linha do Tempo (*Timeline*)** - Este gráfico permite a exibição de uma lista de eventos em ordem cronológica, funcionando tanto para análises quanto para apresentar visualmente uma visão no tempo. Desta forma, é possível compreender os eventos e estabelecer suas relações em um determinado tópico. A Figura 2.2c exemplifica o gráfico de linha do tempo.
- **Diagrama de Venn (*Venn Diagram*)** - Os diagramas são usados na matemática para simbolizar graficamente propriedades, axiomas e problemas relativos aos conjuntos e sua teoria. Desta forma, utilizar o Diagrama de Venn facilita o entendimento das operações básicas de conjuntos - inclusão e pertinência, união e intersecção, diferença e conjunto complementar. Este diagrama tem forte apelo ao usuário, possibilita a organização visual, facilita a comparação entre dois ou mais grupos, além de auxiliar na solução de problemas matemáticos e no raciocínio lógico. Contudo, construir um diagrama de Venn para mais de 3 conjuntos leva a impreci-

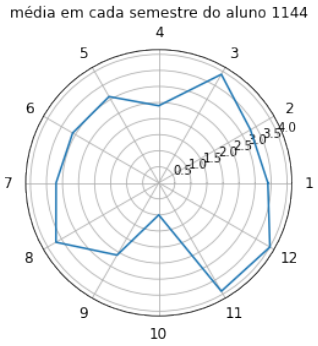
são, devido à necessidade de diminuir/alongar alguns conjuntos para a visualização. O Diagrama de Venn é ilustrado na Figura 2.2e.

- **Diagrama de Sankey (*Sankey Diagram*)** - Este diagrama é utilizado no mapeamento dos fluxos de energia e massa entre processos. Por isso, é uma representação visual de um fluxo envolvendo transferência de uma propriedade física de uma etapa para a outra, tendo as quantidades conservadas dentro dos limites definidos. A Figura 2.2b apresenta um exemplo do diagrama de Sankey.
- **Gráfico de Mapa (*Data Map*)** - O gráfico é utilizado para mostrar valores em um fundo que é frequentemente geográfico, ou seja, o mapa de um país, continente ou região. O gráfico de mapa utiliza escalas de cores e tamanhos para representar informações, destacando visualmente áreas de alto valor. O gráfico de mapa é exibido na Figura 2.2f.
- **Gŕafico Treemap (Treemap Chart)** - Esta é uma técnica de exibição de dados hierárquicos que utiliza retângulos aninhados, ou seja, cada filho de um nodo é representado por um retângulo contido no retângulo do nodo pai. Desta forma, a área total é desmembrada de forma proporcional entre todos os retângulos conforme a importância do dado, fazendo uso eficiente do espaço, além de possibilitar a visualização de padrões ao utilizar cores e tamanhos. Este gráfico é ilustrado na Figura 2.2d.

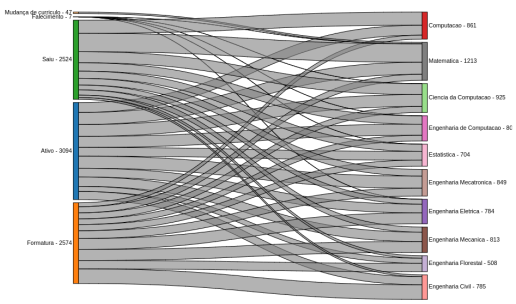
2.2.3 Visualizações não Classificadas por Lengler & Eppler

Novos algoritmos são desenvolvidos a cada dia com o intuito de facilitar o entendimento e a análise de dados. Desta forma, os algoritmos apresentados nessa seção não foram mapeados no estudo de Lengler & Eppler, apesar de serem utilizados de forma a potencializar a apropriação da informação:

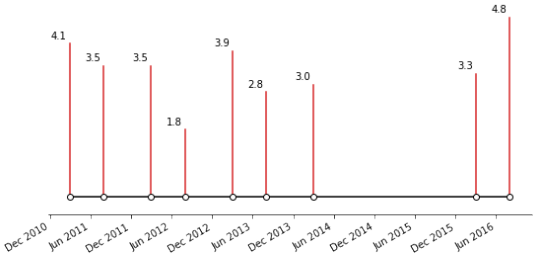
- **Mapa de Calor (*Heatmap*)** - Este gráfico é aplicado em dados geográficos, demográficos ou comportamentais, em que cada cor da célula corresponde ao valor de correlação de duas determinadas variáveis. O mapa de calor utiliza da cor e da disposição em matriz para demonstrar como o fenômeno está agrupado/varia no espaço. Representar os dados em formato de matriz facilita a extração do valor exato, mesmo em grandes conjuntos de dados. Da mesma forma, as cores auxiliam a rápida leitura de padrões, em escala quantitativa. O mapa de calor é ilustrado na Figura 2.3e.



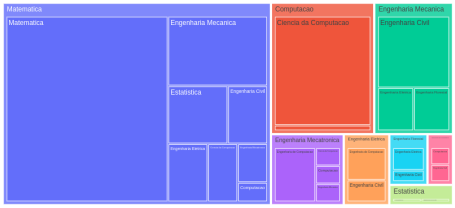
(a) Exemplo utilizando gráfico de radar.



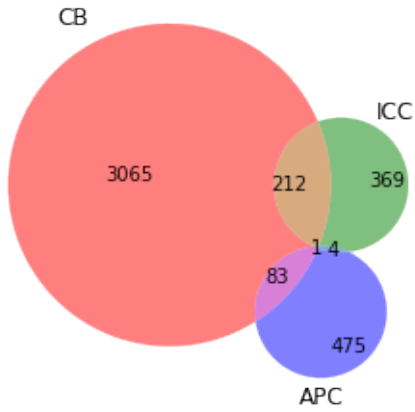
(b) Exemplo utilizando diagrama de Sankey.



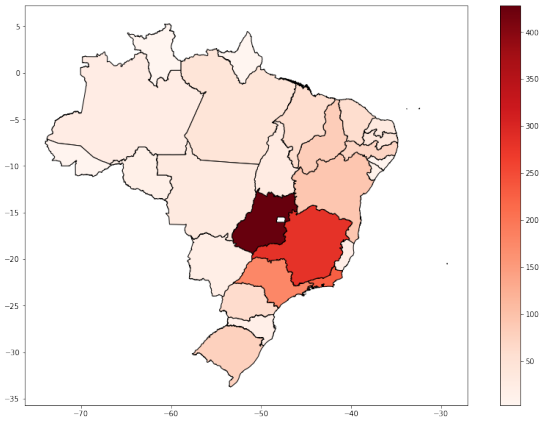
(c) Exemplo utilizando linha do tempo.



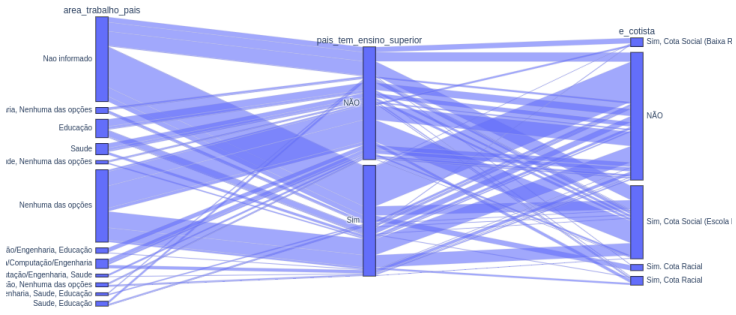
(d) Exemplo utilizando treemap.



(e) Exemplo utilizando diagrama de Venn.



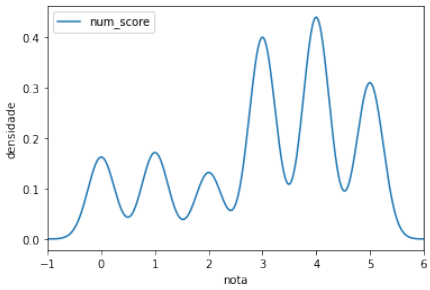
(f) Exemplo utilizando gráfico de mapa.



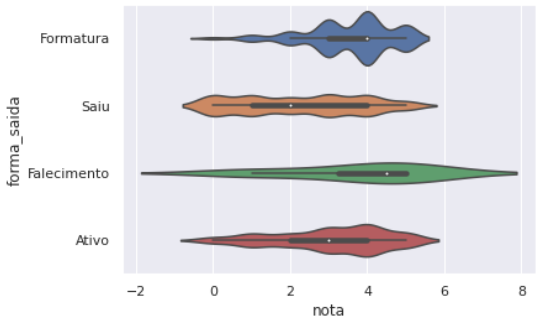
(g) Exemplo utilizando coordenadas paralelas.

Figura 2.2: Algoritmos da categoria visualização de informação.

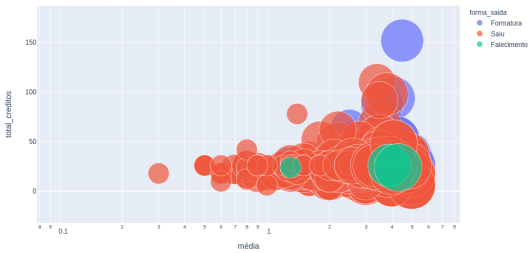
- **Gráfico de Violino (*Violin Plot*)** - A visualização da distribuição de dados quantitativos em variáveis categóricas é possibilitada por meio deste gráfico, de forma que a comparação é facilitada. O gráfico de violino é semelhante ao gráfico de caixa (*box-plot*), mas com a possibilidade de visualizar a densidade de probabilidade dos dados, suavizados por um estimador de densidade de Kernel. A Figura 2.3b apresenta um exemplo desse gráfico.
- **Gráfico de Bolhas (*Bubble Chart*)** - É um gráfico multi-variável semelhante ao gráfico de dispersão e ao gráfico de área proporcional. São exibidas três dimensões do dado ao utilizar um sistema de coordenadas cartesianas para plotar pontos utilizando duas variáveis X e Y, cada uma representando um eixo. Além disso, a área do círculo permite inferir uma terceira variável, enquanto que as cores podem ser usadas para distinguir entre categorias ou para representar uma variável adicional. Porém, assim como os gráficos de área proporcional, o tamanho dos círculos tem como base a área, não o raio ou diâmetro, o que pode levar a interpretações errôneas pelo sistema visual humano. O gráfico de bolhas é utilizado para comparar e exibir as relações entre círculos categorizados, enquanto que a imagem do gráfico pode ser usada para analisar padrões e correlações. A Figura 2.3c apresenta o gráfico de bolhas.
- **Gráfico de Rede (*Network*)** - Uma rede de nós, vértices e linhas de ligação é utilizado por este gráfico para representar as conexões e auxiliar na visualização do relacionamento entre um grupo de entidades. Nesse sentido, este gráfico mostra como a informação flui, onde os componentes existem na rede e como estes interagem. O gráfico de rede é ilustrado na Figura 2.3d.
- **Gráfico de Densidade (*Density Plot*)** - Este gráfico descreve o padrão de uma distribuição. A área sob a curva indica a frequência das observações que se enquadram naquele intervalo. O gráfico de densidade difere do histograma ao reduzir o tamanho das barras e traçar uma linha na parte superior das mesmas, além de utilizar a medida de frequência em vez de contagens. Este gráfico é exemplificado na Figura 2.3a.



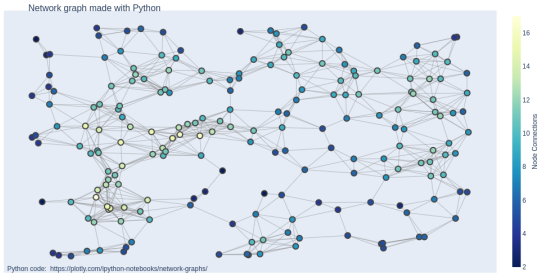
(a) Exemplo utilizando gráfico de densidade.



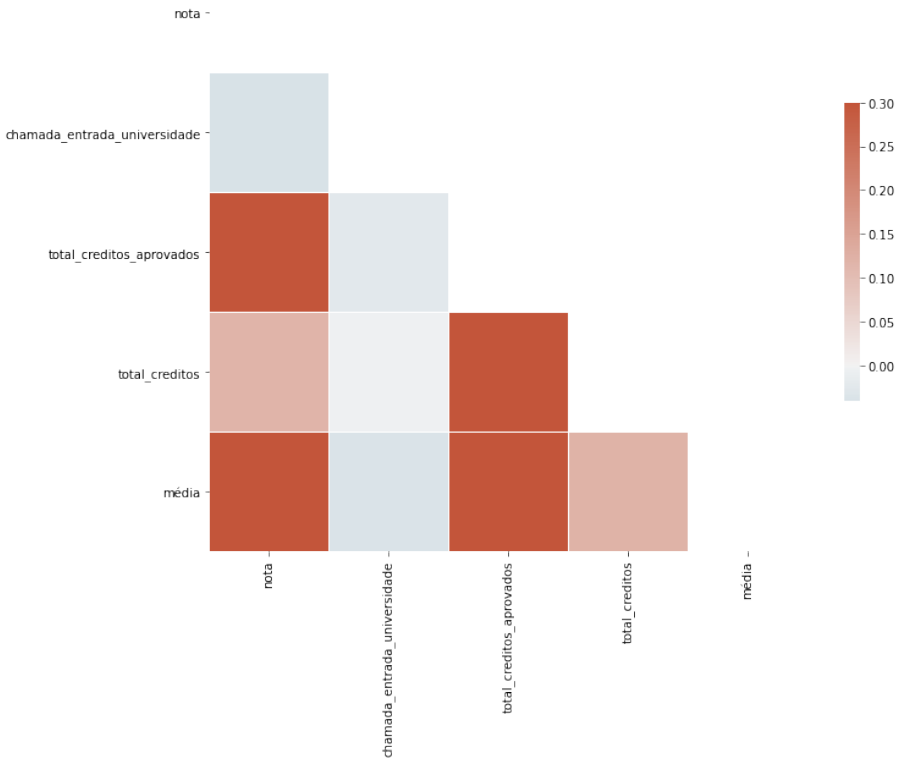
(b) Exemplo utilizando gráfico de violino.



(c) Exemplo utilizando gráfico de bolhas.



(d) Exemplo utilizando gráfico de rede.



(e) Exemplo utilizando o mapa de calor.

Figura 2.3: Visualizações não classificadas por Lengler & Eppler.

Capítulo 3

Revisão da Literatura

Este capítulo tem como objetivo revisar os trabalhos literários que utilizam temáticas similares ao presente estudo. Inicialmente, na Seção 3.1, é estabelecida uma metodologia para identificação de trabalhos acadêmicos aderentes ao estudo. Em seguida, na Seção 3.2, são apresentados os trabalhos mais relevantes sobre sistema de análise visual em dados educacionais, bem como as diferentes formas de avaliação de visualização de dados educacionais encontradas na literatura. Por último, a Seção 3.3 apresenta as considerações finais.

3.1 Método de Pesquisa: Processo de Mapeamento Sistemático

Com o objetivo de obter os trabalhos existentes na literatura acadêmica sobre o tema "análise visual no contexto educacional", foi desenvolvida uma abordagem sistemática dividida em três etapas: Busca por Trabalhos, Primeiro Ciclo e Segundo Ciclo. Esta metodologia, além de elencar os trabalhos mais aderentes ao tema, apresenta resultados não enviesados [34], permitindo avaliar e sintetizar outras contribuições [35]. A Figura 3.1 apresenta o fluxo das três etapas da metodologia.

A primeira etapa da metodologia consiste na obtenção dos trabalhos resultantes da aplicação de uma *string* de busca em bases de dados. Para tanto, foi necessário definir as palavras que fazem parte da busca e quais bases de dados são utilizadas. Determinou-se que seriam empregados o Scopus e o Web of Science (WoS) como fontes de pesquisa e que a construção da *string* de busca utilizaria o sistema VOSviewer, além da opinião de especialistas. O VOSviewer é uma ferramenta de software para construção e visualização de redes bibliométricas [36].

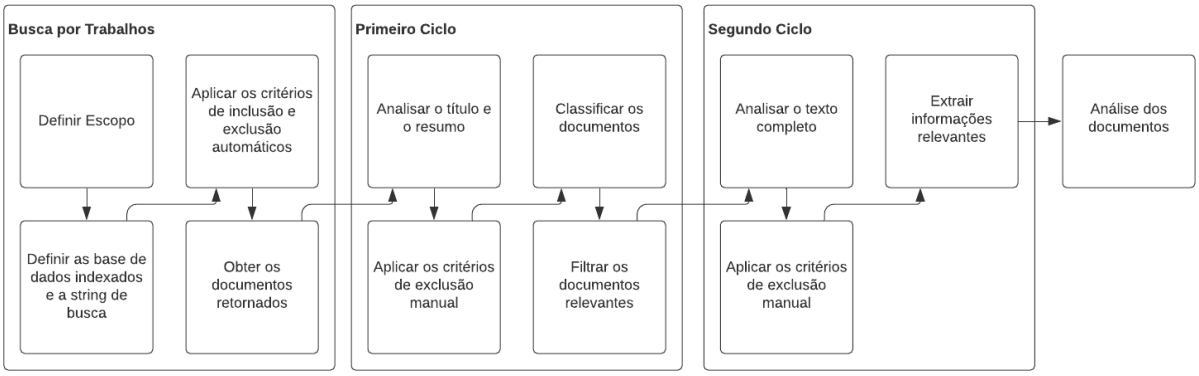


Figura 3.1: Metodologia do processo de mapeamento sistemático.

A definição da *string* de busca teve início com a aplicação de palavras-chave, relacionadas ao tema, nas bases de dados definidas. Os termos iniciais foram "visualization" e "educational data". O VOSviewer foi, então, utilizado para descobrir novas palavras relacionadas, aumentando a *string* de busca. As palavras previamente selecionadas foram submetidas aos especialistas que identificaram as mais relevantes e propuseram outros termos não previstos anteriormente. Ao final do processo, a *string* definida foi a seguinte:

(("visualization techniques" OR "visualisation techniques" OR "visualization approach" OR "visualisation approach" OR "algorithm visualization" OR "algorithm visualisation" OR "data visualization" OR "data visualisation" OR "information visualization" OR "information visualisation" OR "visualization tool" OR "visualisation tool" OR "visual analytics")
AND
("educative" OR "educational dataset" OR "educational data" OR education)
AND
(student* OR undergraduate*))

Com o intuito de utilizar apenas trabalhos relevantes para a pesquisa, foi necessário estabelecer critérios de exclusão automáticos para diminuir a incidência dos falso-positivos resultantes da aplicação da *string*. Para tanto foram aplicados os seguintes filtros: 1) Somente trabalhos da última década (2010 até 2019) seriam analisados; e 2) Somente trabalhos da área de computação e ciências sociais seriam considerados. Ao final da Busca por Trabalhos, que integra a aplicação da *string* de busca nas bases de dados e o emprego dos critérios de exclusão, 823 trabalhos resultaram do Scopus e 205 do WoS, totalizando 1028 documentos.

A etapa do Primeiro Ciclo teve como objetivo detectar e filtrar manualmente os trabalhos relacionados à visualização de análises em dados educacionais. Assim, foram lidos o título e o resumo dos 1028 trabalhos obtidos ao final da fase Busca por Trabalhos e

selecionados os relevantes ao tema. O Segundo Ciclo consistiu em analisar os artigos selecionados na etapa anterior e extrair as informações relevantes para a realização do estudo da literatura. De forma a realizar o proposto na etapa, os trabalhos foram lidos na íntegra.

Durante a leitura dos textos, em ambas as fases, foram empregados os critérios de exclusão manual, onde foram removidos: 1) artigos em duplicidade; 2) trabalhos classificados como livros inteiros; 3) trabalhos em idiomas divergentes de português/inglês; 4) trabalhos não encontrados na íntegra; e 5) trabalhos não classificados como artigos completos. Os critérios de exclusão manual identificaram 175 trabalhos duplicados, 21 classificados como revisão de literatura, 17 identificados como livros e 6 trabalhos que possuíam o título e o resumo em inglês, mas o restante do texto estava em espanhol. Além disso, 16 trabalhos não eram artigos completos e 44 não estavam disponíveis. Estes documentos foram excluídos da análise.

Ao final da metodologia, 128 trabalhos foram selecionados no tocante a área de análise visual de dados educacionais, dos quais as seguintes informações foram extraídas: *Type* - sobre o que se trata o trabalho analisado, LMS - se analisa alguma LMS (*Learning Management System*), API - se utiliza alguma API (*Application Programming Interface*) para análise dos dados, *Open* - se possui algoritmo aberto, *Algorithm* - quais algoritmos de visualização são utilizados, *Data Mining* - se é usado e quais são os algoritmos de mineração de dados, *Source* - qual a fonte dos dados analisados, *Major* - se teve e qual foi o curso superior analisado, *Educational Level* - qual nível educacional analisado (ensino fundamental, graduação ou pós-graduação) e *Class* - se teve e qual foi a disciplina escolhida para análise. Estas informações facilitam na classificação e análise dos textos, sem a necessidade de releitura do trabalho.

A revisão de literatura completa em conjunto com a análise dos resultados pode ser encontrada no artigo [37] "A Literature Study of Visual Analysis in an Educational Context" publicado no evento "Frontiers in Education" (FIE), em outubro de 2020. Na próxima seção serão apresentados alguns dos trabalhos obtidos utilizando a metodologia descrita.

3.2 Trabalhos Relacionados

Sistemas dedicados à análise visual têm sido utilizados majoritariamente para apoiar os professores e promover qualidade ao ensino. A partir da metodologia apresentada na seção anterior, foi possível identificar uma série de trabalhos com essa temática.

Essa & Ayad [38] descreveram o sistema *Student Success System* (S3), que dispõe de análises visuais holísticas do progresso acadêmico dos alunos. São utilizados modelos preditivos e aprendizagem de máquina para prever alunos em risco de reprovação, além

de algoritmos de visualização para possibilitar percepções diagnósticas e gerenciar intervenções. O sistema S3 apresenta uma lista dos alunos, associando cada um deles a um indicador de risco. Para cada aluno, são apresentados o progresso acadêmico e os fatores de risco, por meio de visualizações tais como quadrante de risco, gráfico de pontos interativo, gráfico de ganhos e perdas e um sociograma que ilustra a comunicação e interação dos alunos.

Sistemas de análise visual são empregados em diversos níveis educacionais, como no sistema CareerVis [39] que auxilia na escolha da qualificação superior para alunos do ensino médio utilizando formas visuais simples. O sistema apresenta: o fluxo principal de cursos em uma faculdade; os gráficos de dispersão destacando áreas e profissões; o contexto de outras faculdades e profissões; e as características detalhadas das profissões no contexto de todas as outras. Já em [40] foi apresentado um sistema desenvolvido na Universidade de Ruse, Bulgária, que objetiva a condução da formação de alunos de doutorado em um alto nível acadêmico. O sistema permite que a situação de cada aluno de doutorado seja determinada com precisão e pontualidade, além de manter os portfólios detalhados de cada participante e facilitar a geração de várias ferramentas para visualização do progresso. O sistema já está em produção, sendo utilizado por 312 alunos de doutorado, 189 candidatos e 216 supervisores.

Em uma avaliação dos trabalhos selecionados pela metodologia, que informam o nível educacional, constatou-se que a grande maioria - um total de 61 trabalhos - realiza análise em dados de graduação, conforme ilustra a Figura 3.2. Entretanto, apenas 30 trabalhos utilizaram dados específicos de disciplinas para análise, das quais apenas 7 eram introdutórias [41] [42] [43] [44] [45] [46] [47].

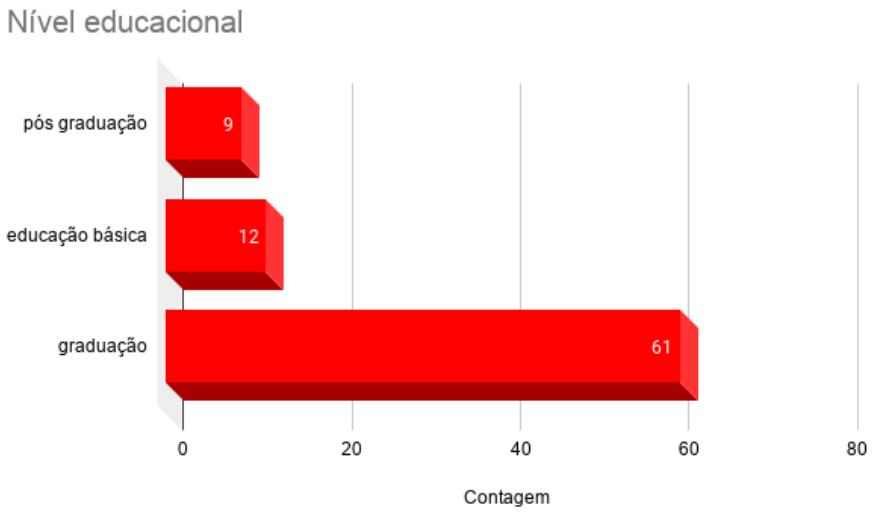


Figura 3.2: Trabalhos de acordo com o nível educacional informado.

Em relação à avaliação do sistema de análise visual, Culligan et al. [48] utilizou um questionário aplicado aos usuários para avaliar o sistema *VEAP, the Visualisation Engine and Analyser for PreSS#*. Este sistema tem como foco os professores e possui três funcionalidades principais: 1) gerar visualizações dos dados produzidos e obtidos pelo *PreSS#*; 2) apresentar os dados individuais dos alunos; e 3) realizar análises sobre os resultados dos alunos. Por outro lado, a avaliação do sistema DAVID [49], foi realizada a partir de um estudo de caso utilizando dados de um curso oferecido pelo SENAC (*Serviço Nacional de Aprendizagem Comercial*), no qual foram identificados pontos que necessitavam de melhorias no ensino, com vista a elaborar uma nova edição do curso.

Silva et al. [50] aplicaram um questionário em 7 professores para avaliar os gráficos da ferramenta de visualização de dados desenvolvida. Esta ferramenta utiliza resultados de técnicas de mineração de dados para análise. A avaliação utilizou a escala Likert para as respostas: concordo plenamente, concordo, neutro, discordo e discordo plenamente. As perguntas estavam divididas nas categorias: utilidade percebida, facilidade de uso percebida e intenção de uso. Já em [51] foi realizado um estudo de três técnicas de avaliação de visualizações hierárquica - "Nielsen", "Bastien e Scapin" e uma técnica proposta pelos autores. 14 alunos do Instituto de Informática da UFRGS, da disciplina de Comunicação Homem-Computador, do sexto semestre do Curso de Ciência da Computação e 4 avaliadores mais experientes foram divididos em três grupos, cada um responsável por utilizar uma das avaliações selecionadas. Segundo os resultados, mais problemas de usabilidade foram identificados utilizando a técnica proposta e apenas 1 aluno não conseguiu identificar todos os problemas.

3.3 Considerações Finais

A literatura é bem extensa no âmbito de sistemas de análise visual de dados no contexto educacional, com variações nos algoritmos utilizados. Entretanto, nenhum dos trabalhos encontrados estudou, selecionou e validou diversas visualizações de forma a utilizar a mais apropriada para descoberta de conhecimento. Além disso, quando se trata de dados referentes às disciplinas introdutórias em programação, são poucos os trabalhos encontrados, ainda que evidências na literatura indiquem que essas disciplinas são importantes para a continuidade dos alunos no curso [3] [4] [2].

Diferente do conteúdo apresentado na literatura, a pesquisa desenvolvida tem como contribuição o estudo prévio de diversas visualizações, bem como a validação das escolhas com professores e gestores educacionais. Os algoritmos selecionados foram também aplicados, com o intuito de descoberta de conhecimento, em dados das disciplinas introdutórias de computação da Universidade de Brasília - Introdução à Ciência da Computação (ICC),

disciplina para alunos que não são do Departamento de Ciência da Computação; Algoritmos e Programação de Computadores (APC), ofertada para alunos do departamento; e Computação Básica (CB), antiga denominação da disciplina APC.

Capítulo 4

Metodologia

Este capítulo apresenta os processos necessários para a execução deste trabalho, tendo sido estruturado em 4 fases: definição do conjunto de dados, estudo dos fundamentos de visualização, elaboração e definição das visualizações e, por fim, realização de experimentos, análise dos resultados e ajustes finais. A Figura 4.1 ilustra as etapas que são detalhadas nas Seções 4.1, 4.2, 4.3 e 4.4, respectivamente.

Figura 4.1: Metodologia do projeto.

4.1 Definição do Conjunto de Dados

O objetivo dessa fase é realizar um estudo e uma triagem dos dados, selecionando aqueles que são relevantes para o trabalho. Estes dados não devem sobrecarregar o usuário com informações desnecessárias, tampouco impossibilitar a tomada de decisão assertiva por apresentarem poucas informações. Desta forma, são utilizadas duas fontes distintas de dados anonimizados, por questão de confidencialidade: o SIGRA (Sistemas Acadêmicos da Universidade de Brasília) e o questionário aplicado no início e no final do semestre. Os fundamentos desses dados são apresentados nas Seções 4.1.1 e 4.1.2 respectivamente. O acesso a esses dados foi aprovado pelo comitê de ética com o parecer número 4.283.719, de 17/09/2020.

4.1.1 Dados do SIGRA

Os dados extraídos do SIGRA possuem informações referentes aos alunos da Universidade de Brasília por um período que abrange desde o segundo semestre de 1984 até o semestre de verão de 2020. Os registros totalizam 181.491 informações pessoais e acadêmicas de 8.052 alunos distintos nos seguintes cursos da UnB: Ciência da Computação, Computação (Licenciatura), Engenharia Civil, Engenharia de Computação, Engenharia Elétrica, Engenharia Florestal, Engenharia Mecatrônica, Engenharia Mecânica, Estatística e Matemática.

Tratamento de dados é qualquer operação realizada utilizando um conjunto de dados, podendo incluir coleta, recebimento, reprodução, extração, armazenamento, entre outras. Desta forma, após a extração das informações provenientes do SIGRA foi necessário realizar análises para corrigir as inconsistências nos dados.

Das ações tomadas, evidencia-se: a capitalização do nome da disciplina; a retirada de espaços em branco nos campos de identificação do aluno, nome do curso, período de entrada na universidade e nome da disciplina; e os tratamentos para valores indefinidos ("NaN"), em que foi estipulado, para os alunos sem endereço, o valor 'Nao informado', para os alunos sem cota, o valor 'Nao' e para os alunos que não possuíam forma de saída, o valor 'Ativo'; o período de saída dos alunos ativos foi marcado como '0'; por fim, os alunos que não saíram por formatura ou falecimento, foram agregados na forma de saída 'Saiu'.

O último tratamento dos dados foi devido a uma mudança no currículo do curso de Ciência da Computação, Computação, Engenharia Elétrica, Estatística e Matemática, em que os dados informavam que os alunos estavam ativos tanto no currículo antigo, quanto no currículo novo. Portanto, foi criado uma nova forma de saída chamada "Mudança de currículo" e foi declarada a saída desses alunos do antigo currículo no ano da mudança, variando de curso para curso.

Ao realizar a avaliação dos dados, foram encontradas divergências nas disciplinas, em que a mesma disciplina tinha quantidade de créditos distintos. Entretanto essa variação ocorria apenas em alguns alunos cujas as notas estavam registradas como TR (trancamento parcial de matrícula: concessão automática), TJ (trancamento parcial de matrícula: excepcional e justificado), CC (crédito concedido), AP (aprovado) e DP (dispensado). Desta forma, foram excluídos os registros com essas menções, ficando apenas SS (nota de 10 a 9), MS (nota de 8,9 a 7), MM (nota de 6,9 a 5), MI (nota de 4,9 a 3), II (nota de 2,9 a 1) e SR (nota de 0,9 a 0). Os dados originais possuíam 18.1491 registros, sendo que as instâncias inconsistentes, 12,3%, foram removidas no pré-processamento, resultando em 15.9381 registros. Desta forma, foi possível enumerar as notas, criando um novo campo com os valores numéricos das mesmas, variando de 0 a 5.

Para o armazenamento das informações, foi implementado o modelo de banco de dados relacional apresentado na Figura 4.2. O modelo criado contém quatro tabelas: “Disciplina”, “Aluno”, “Aluno_Curso” e “Aluno_Curso_Disciplina”. As tabelas e os dados de cada uma são explicados abaixo.

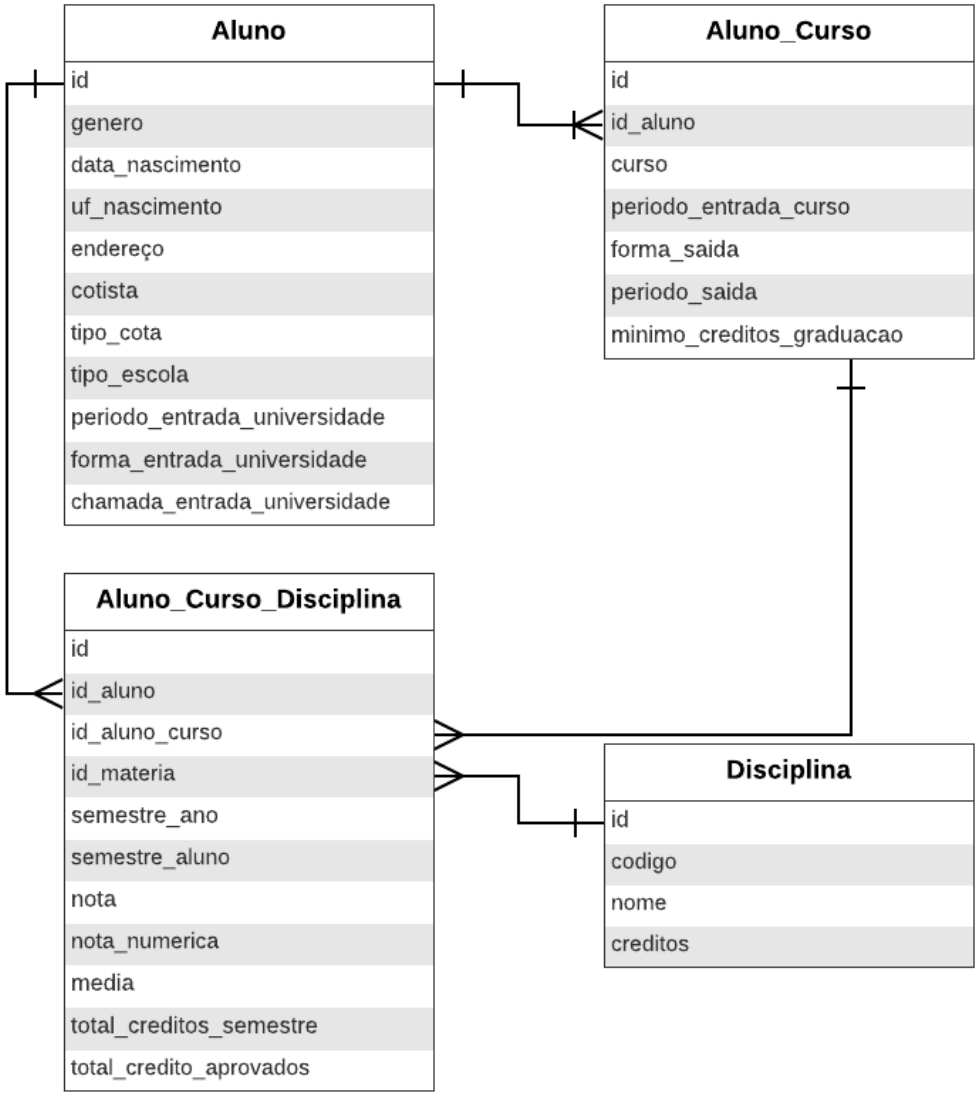


Figura 4.2: Modelo do banco de dados relacional utilizando informações do SIGRA.

Disciplina

A tabela “Disciplina” contém informações gerais sobre as disciplinas da universidade, os detalhes sobre os atributos são apresentados abaixo e resumidos na Tabela 4.1.

- Id - chave primária gerada automaticamente;

- Código - informa o código da disciplina;
- Nome - informa o nome da disciplina;
- Créditos - informa a quantidade de créditos que a disciplina vale;

Tabela 4.1: Resumos dos atributos da tabela “Disciplina”.

Campo	**Tipo do campo**	**Total de categorias**	**Descrição do campo**
código	numérico/ categórico	132	código da disciplina.
nome	categórico	132	nome da disciplina.
créditos	numérico	-	informa a quantidade de créditos da disciplina.

Aluno

A tabela “Aluno” contém informações gerais sobre os estudantes e seus atributos são apresentados abaixo e resumidos na Tabela 4.2.

- Id - chave primária que identifica o aluno;
- Gênero - informa o gênero do aluno, podendo ser ‘M’ para masculino ou ‘F’ para feminino;
- Data nascimento - informa a data de nascimento do aluno no seguinte formato: DD/MM/AAAA, em que DD são os dois dígitos da data, MM são os dois dígitos do mês e AAAA são os quatro dígitos do ano;
- UF de nascimento - representa a sigla da Unidade da Federação onde o aluno nasceu, as opções são os 27 estados do Brasil e o Distrito Federal;
- Endereço - informa o endereço do aluno no formato *string*;
- É cotista - informa se o aluno entrou pelo sistema de cotas, em que as opções são: “Sim”, caso o aluno tenha entrado, ou “Nao”, caso o aluno não seja cotista;
- Tipo de cota - informa se o aluno entrou por cota e qual tipo de cota. Este campo possui 9 categorias distintas: “Escola Púb Alta Renda-PPI-PCD”, “Escola Púb Baixa Renda-Não PPI”, “Escola Púb Baixa Renda-PPI-PCD”, “Escola Púb. Alta Renda-Não PPI”, “Escola Pública Alta Renda-PPI”, “Escola Pública Baixa Renda-PPI”, “Indígena”, “Nao” e “Negro”;

- Tipo de escola - informa o tipo de escola na qual o aluno estudou antes de entrar na universidade, podendo ser “Particular” ou “Publica”;
- Período de entrada na universidade - é um campo numérico, em que os quatro primeiros dígitos são o ano de entrada na universidade e o último representa o semestre;
- Forma de entrada na universidade - representa a forma como o aluno ingressou na faculdade. Este campo possui 13 categorias distintas: “Acordo Cultural-PEC-G”, “Convenio - Andifes”, “Convenio-Int”, “Enem UnB”, “Matricula Cortesia”, “Portador Diploma Curso Superior”, “Programa de Avaliacao Seriada (PAS)”, “Refugiado”, “SISU-Sistema de Selecao Unificada”, “Transferencia Facultativa”, “Transferencia Obrigatoria”, “Transferencia-Convenio” e “Vestibular”;
- Chamada de entrada na universidade - informa por qual chamada o aluno entrou na universidade.

Tabela 4.2: Resumos dos atributos da tabela “Aluno”.

Campo	**Tipo do campo**	**Total de categorias**	**Descrição do campo**
Gênero	categórico	2	informa o gênero do aluno
Data de nascimento	data (DD/MM/AAAA)	-	informa a data de nascimento
UF de nascimento	categórico	28	indica a sigla da Unidade da Federação onde o aluno nasceu
Endereço	string	-	informa o endereço do aluno
É cotista	categórico	2	informa se o aluno é cotista ou não
Tipo de cota	categórico	9	informa qual o tipo de cota
Tipo de escola	categórico	2	expressa o tipo de escola na qual o aluno estudou antes de entrar na universidade
Período de entrada na universidade	categórico/ numérico	-	informa o período em que o aluno entrou na universidade
Forma de entrada na universidade	categórico	13	indica a forma como o aluno ingressou na faculdade
Chamada de entrada na universidade	numérico	-	indica a chamada em que o aluno ingressou na faculdade

Aluno_Curso

Na UnB os alunos podem entrar em um curso e, futuramente, mudar de opção, de forma que os cursos dos quais um aluno participou são armazenados na tabela "Aluno_Curso". Os atributos relativos ao curso do aluno são descritos a seguir e resumidos na Tabela 4.3.

- Id - chave primária gerada de forma automática;
- Id_aluno - chave estrangeira que referencia a tabela "Aluno", identificando o código do aluno;
- Curso - representa os cursos que os alunos cursaram/estão cursando. É um dado categórico com 10 opções - Ciência da Computação, Computação, Engenharia Civil, Engenharia de Computação, Engenharia Elétrica, Engenharia Florestal, Engenharia Mecânica, Engenharia Mecatrônica, Estatística e Matemática;
- Período de entrada no curso - informa o ano e o semestre que o aluno entrou no curso. É um campo numérico, em que os quatro primeiros dígitos identificam o ano de entrada e o último identifica o semestre;
- Forma de saída - identifica a forma com que o aluno saiu do curso, podendo ser: "Saiu" - se o aluno saiu do curso por alguma forma de desligamento; "Ativo" - se o aluno ainda está no curso; "Formatura" - se o aluno terminou o curso; "Falecimento" - se o aluno faleceu durante o curso; e "Mudança de currículo" - se o curso mudou de currículo enquanto o aluno estava ativo;
- Período de saída - representa o período que o aluno saiu do curso. É um campo numérico, em que os quatro primeiros dígitos são o ano de saída da opção e o último representa o semestre;
- Mínimo de créditos para graduação - representa o número de créditos para a formatura de um curso.

Aluno_Curso_Disciplina

Por último, a tabela "Aluno_Curso_Disciplina" se relaciona com as demais, ou seja, informa a disciplina que foi cursada pelo aluno durante o curso. Os atributos relativos a essa tabela são descritos abaixo e resumidos na Tabela 4.4.

- Id - chave primária gerada automaticamente;
- Id_aluno - chave estrangeira que referencia a tabela "Aluno", identificando o código do aluno;

Tabela 4.3: Resumos dos atributos da tabela “Aluno_Curso”.

Campo	**Tipo do campo**	**Total de categorias**	**Descrição do campo**
curso	categórico	10	informa os cursos que os alunos cursaram/estão cursando.
período de entrada no curso	categórico/ numérico	-	expressa o ano e o semestre em que o aluno entrou no curso..
forma de saída	categórico	5	refere-se a forma com que o aluno saiu do curso.
período de saída	categórico/ numérico	-	informa o período que o aluno saiu do curso.
minimo de créditos para a formatura	numérico	-	indica o número de créditos para a formatura de um curso.

- Id_aluno_curso - chave estrangeira que se relaciona à chave primária da tabela “Alunos_Curso”;
- Id_materia - chave estrangeira que referencia a tabela “Disciplina”, representando o identificador da disciplina;
- Semestre/Ano - informa o período em que o aluno cursou determinada disciplina. É um campo numérico, no qual os quatro primeiros dígitos identificam o ano e o último o semestre;
- Semestre do aluno - é um campo numérico, representando o semestre do aluno em números ordinais;
- Nota - identifica a nota do aluno em determinada disciplina. As notas são representadas pelas categorias: SR, II, MI, MM, MS, SS, no qual para ser considerado aprovado na disciplina o aluno precisa tirar SS, MS ou MM. As notas SR, II e MI significam que o aluno foi reprovado;
- Nota numérica - representa as notas transformadas em números, em que ‘SS’: 5, ‘MS’: 4, ‘MM’: 3, ‘MI’: 2, ‘II’: 1 e ‘SR’: 0;
- Média no semestre - informa a média geral do semestre, considerando todas as disciplinas que o aluno fez no mesmo semestre. É um dado numérico que varia de 0.0 até 5.0;
- Total de créditos no semestre - identifica a quantidade de créditos cursados pelos alunos em um mesmo semestre;

- Total de créditos aprovados no semestre - informa a quantidade de créditos das disciplinas onde o aluno foi considerado aprovado em um mesmo semestre.

Tabela 4.4: Resumos dos atributos da tabela "Alunos_Curso_Disciplina".

Campo	**Tipo do campo**	**Total de categorias**	**Descrição do campo**
nota	categórico	6	identifica a nota que o aluno tirou na respectiva disciplina
nota numérica	numérico	-	identifica a nota que o aluno tirou na respectiva disciplina em número
média no semestre	numérico	-	representa a média do aluno durante o semestre
semestre do aluno	numérico	-	indica qual é o semestre que o aluno cursou esta disciplina
semestre/ano	categórico/numérico	-	identifica o ano e o semestre que o aluno cursou a disciplina
total de créditos no semestre	numérico	-	indica a quantidade de créditos cursados no semestre
total de créditos aprovados no semestre	numérico	-	indica a quantidade de créditos das disciplinas aprovadas no semestre

4.1.2 Dados do Questionário

Para complementar os dados acadêmicos com dados de perfil demográfico dos alunos, a fim de utilizar algoritmos de visualização para diferentes tipos de dados, foi aplicado um questionário no início e no final de cada período letivo, nas turmas de APC e ICC, onde os alunos assinaram virtualmente o TCLE (Termo de Consentimento Livre e Esclarecido) de forma voluntária. O objetivo do questionário é colher informações atualizadas dos alunos presentes em cada turma. As perguntas incluem dados pessoais, percepção do curso, experiência em programação, autoavaliação e sugestões para melhorar a disciplina. A visualização desses dados e discussão dos padrões encontrados considerou as respostas do segundo semestre de 2019 e do primeiro semestre de 2020, totalizando 343 registros distintos.

Nos dados do formulário foram realizados os seguintes tratamentos: os valores vazios ("*null*") foram substituídos por "Nao informado"; a separação das respostas de múltipla

escolha foram padronizadas pelo caracter vírgula; palavras “No’ foram substituídas por “Não”; assim como “No gosto” foi corrigido para “Não gosto”.

Para o armazenamento dos dados foi proposta a tabela apresentada na Figura 4.3 que contém as respostas às perguntas do questionário. Os atributos relativos a essa tabela são descritos no Apêndice A.

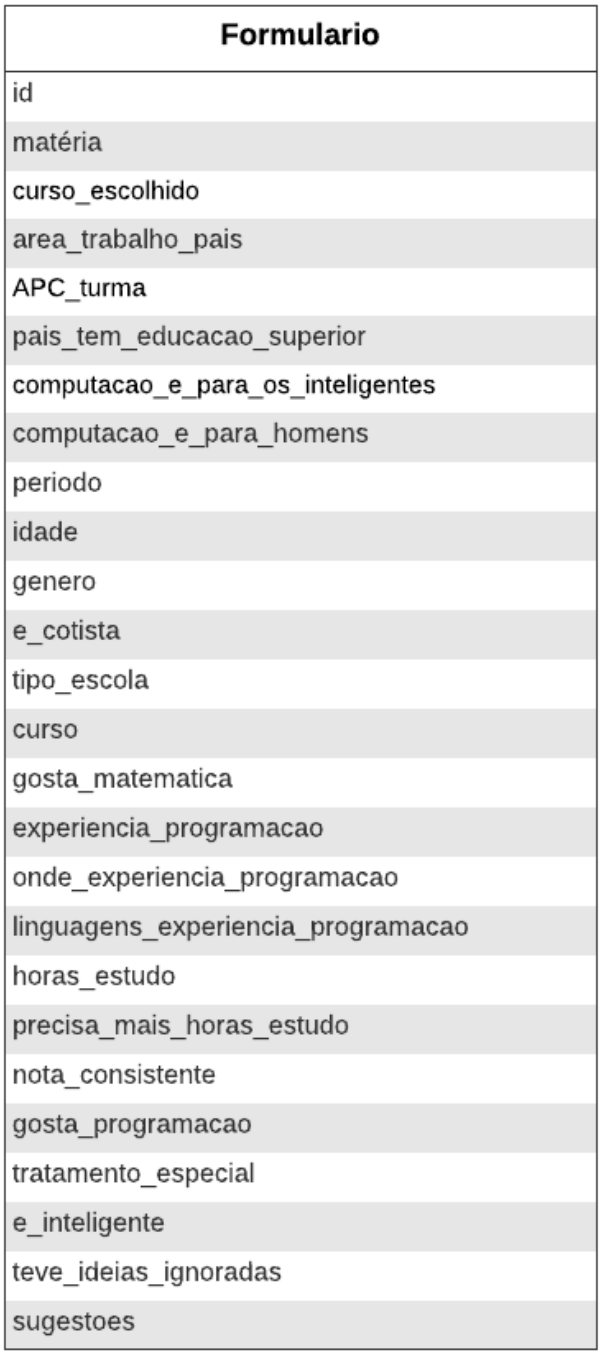


Figura 4.3: Tabela para armazenar as respostas do questionário.

4.2 Estudo das Visualizações

A segunda etapa da metodologia compreende a exploração e delimitação dos algoritmos de visualização e das ferramentas utilizadas para organização, exploração e reorganização dos dados, bem como o *layout* de apresentação das visualizações no espaço disponível da tela. O estudo de visualização de dados contempla uma grande variedade de técnicas, de forma que é preciso selecionar aquelas que possibilitam a interpretação dos padrões existentes no conjunto de dados e que possam auxiliar na descoberta de conhecimento.

Segundo o levantamento de trabalhos na revisão da literatura apresentada na Seção 3.1 e no artigo [37], os algoritmos mais citados nos trabalhos são aqueles expostos na Tabela 4.5. Para a escolha das técnicas de visualização estudadas nesta pesquisa, foram analisadas as mais empregadas nos processos de análise visual no contexto educacional.

Tabela 4.5: Número de publicações que utiliza cada algoritmo de visualização.

Algoritmos de visualização	**Número de publicações**
Gráfico de Barras	46
Gráfico de Linhas	24
Gráfico de Calor	14
Gráfico de Pizza	14
Gráfico de Rede	11
Gráfico de Dispersão	11
Linha do Tempo	10
Gráfico de Caixa	7

Ao comparar o resultado encontrado na literatura com a divisão de Lengler & Eppler [33], observa-se que, dos 8 algoritmos mais utilizados nos trabalhos, 5 pertencem ao grupo de visualização de dados - gráfico de barras, gráfico de linhas, gráfico de pizza, gráfico de dispersão e gráfico de caixas. Ademais, foram identificadas algumas aplicações do grupo de visualização de informação, tais como o gráfico de radar, coordenadas paralelas, linha do tempo e o *treemap*. Observa-se também que, segundo o estudo da literatura, a visualização baseada em linha do tempo está entre as mais utilizadas. Dessa forma, neste trabalho são abordadas duas categorias estabelecidas por Lengler & Eppler: visualização de dados e visualização de informação. Os algoritmos de cada grupo foram especificados e detalhados na Seção 2.2.1 e 2.2.2, respectivamente.

Entretanto, outros algoritmos identificados no mapeamento da literatura não são apresentados na divisão de Lengler & Eppler. Em vista disso, propõe-se utilizar os algoritmos dos dois grupos como base, porém, estendendo o estudo também aos algoritmos encontrados na literatura, objetivando a descoberta de conhecimento em dados educacionais. Esses algoritmos foram expostos na Seção 2.2.3.

No estudo dos 19 algoritmos selecionados foram identificados: os tipos de dados representados nas visualizações e suas características, as vantagens e desvantagens de cada gráfico, além de quais são os gráficos semelhantes, ou seja, aqueles que podem transmitir a mesma mensagem. A Tabela 4.6 apresenta um resumo das técnicas estudadas e uma análise do cenário de aplicação. A partir desse estudo, foi possível eleger os algoritmos mais apropriados para cada tipo de informação.

Tabela 4.6: Análise de cenário dos algoritmos de visualização estudados.

Algoritmo	**Análise de cenário**
Gráfico de pizza	- visualizar relação entre os valores - visualizar relação de um valor único com o total
Gráfico de linhas	- visualizar tendências ou mudanças ao longo do tempo
Gráfico de barras	- comparar diversos valores
Gráfico de área	- comparar as tendências de volume em série de tempo - enfatizar a magnitude da alteração ao longo do tempo - chamar a atenção para o valor total entre uma tendência.
Histograma	- demonstrar uma distribuição de frequências
Gráfico de dispersão	- verificar se existe uma relação entre causa e efeito entre duas variáveis
Gráfico de caixa	- comparar o intervalo e distribuição de grupos de dados numéricos
Gráfico de radar	- comparar membros de uma dimensão em uma função de várias métricas
Coordenadas paralelas	- mostrar a comparação de elementos de dados em um agrupamento
Linha do tempo	- apresentar uma sequência de eventos em ordem cronológica e linear
Diagrama de venn	- mostrar visualmente os relacionamentos entre os conjuntos de dados - organizar informações
Diagrama de sankey	- mostrar quantidades específicas - localizar as contribuições mais significativas para um fluxo geral
Gráfico de mapa	- comparar valores e mostrar categorias entre regiões geográficas
Treemap	- trabalhar com grandes quantidades de dados estruturados hierarquicamente
Mapa de calor	- identificar padrões
Gráfico de violino	- comparar a distribuição de uma determinada variável
Gráfico de bolhas	- mostra a relação entre valores numéricos
Gráfico de redes	- mostra os componentes de uma rede e como eles interagem
Gráfico de densidade	- descrever uma variável numérica

4.3 Aplicação do Estudo de Caso

A terceira fase corresponde à elaboração, definição e especificação das visualizações. Nesta fase foi executado o caso de uso que compreende as disciplinas introdutórias de computação com o objetivo de validar e realizar os ajustes necessários. Desta forma, três domínios foram identificados - informações acadêmicas, informações de percepção e informações gerais. Em todos os segmentos, para a escolha dos algoritmos de visualização, foram selecionados candidatos utilizando o estudo realizado na segunda fase da metodologia.

4.3.1 Domínio de Informações Gerais

O domínio de informações gerais é importante para visualizar, de forma completa, as informações históricas da turma. Este domínio apresenta as respostas para as seguintes perguntas:

1. Qual a quantidade de aprovações/reprovações nas disciplinas?
2. Em qual semestre os alunos cursam a disciplina?
3. Quantas vezes os alunos cursam a disciplina?
4. Quais os cursos dos alunos que fizeram pela primeira vez a disciplina em comparação com os que fazem novamente?

Para responder cada pergunta, algoritmos foram selecionados para apresentar os dados obtidos do SIGRA. Os dados e os algoritmos selecionados estão resumidos na Tabela 4.7.

Pergunta 1: Qual a quantidade de aprovações/reprovações nas disciplinas?

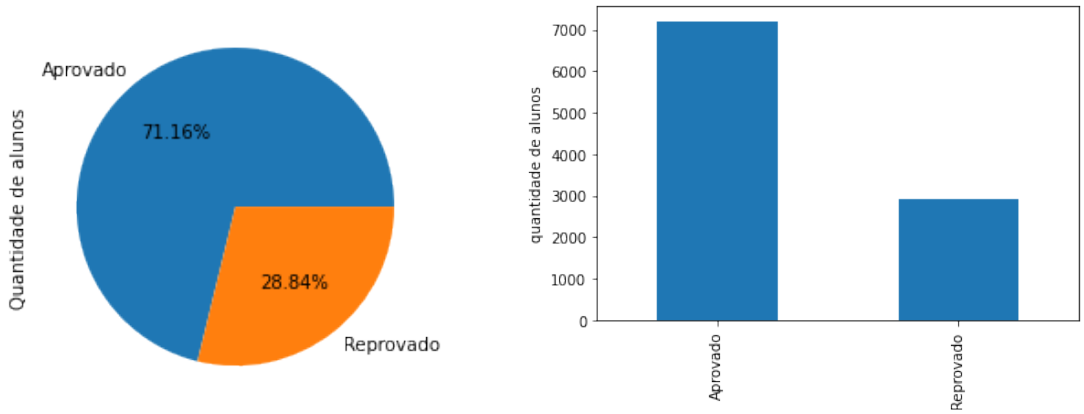
De forma a verificar se os alunos costumam ser aprovados na disciplina, as notas dos alunos foram utilizadas para gerar as visualizações. Notas iguais ou maiores que 3 entraram na categoria “Passou”, enquanto que as notas menores foram classificadas no grupo “Reprovou”. Nesse sentido, o gráfico de pizza e o gráfico de barras foram escolhidos por apresentarem de forma clara dados com poucas categorias, como mostram as visualizações nas Figuras 4.4a e 4.4b. É possível observar que 71% dos alunos foram aprovados nas disciplinas do estudo de caso, ou seja, mais de 7000 alunos.

Pergunta 2: Em qual semestre os alunos cursam a disciplina?

Para entender o perfil dos alunos que cursam a disciplina - se a maioria está no início da faculdade ou no final - foi utilizado o semestre em que este cursou a disciplina. As visualizações candidatas foram: gráfico de barras, gráfico de densidade, histograma e o

Tabela 4.7: Informações do domínio de informações gerais.

Perguntas de pesquisa	**Atributos**	**Algoritmos**
Pergunta 1	nota	Gráfico de Pizza
		Gráfico de Barras
Pergunta 2	semestre do aluno	Gráfico de Barras
		Treemap
		Gráfico de Densidade
		Histograma
Pergunta 3	quantidade de vezes que o aluno estava relacionado com a disciplina	Gráfico de Barras
		Treemap
		Gráfico de Densidade
		Histograma
Pergunta 4	curso	Coordenadas Paralelas
	nota	Gráfico de Barras
	quantidade de vezes que o aluno estava relacionado com a disciplina	Gráfico de Caixa
		Gráfico de Violino
		Gráfico de Barras e Gráfico de Linhas



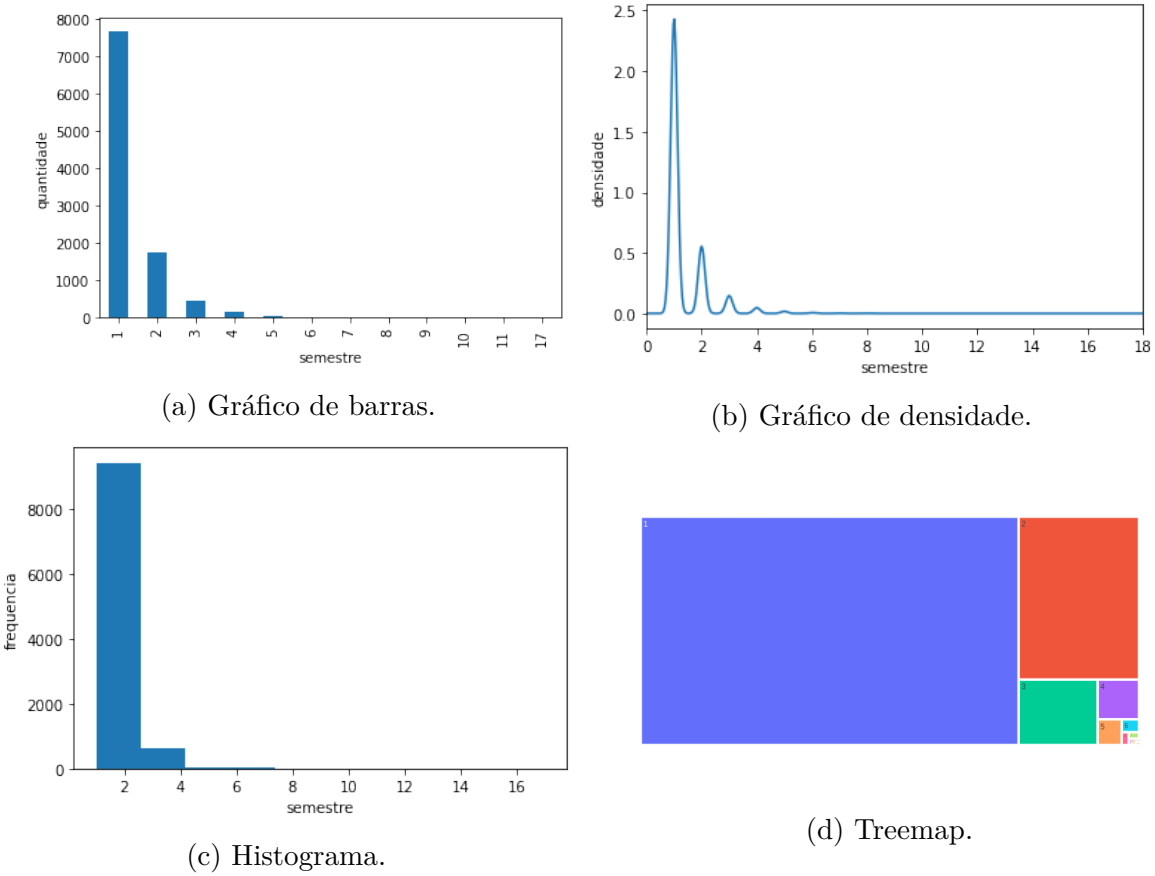
(a) Gráfico de pizza.

(b) Gráfico de barras.

Figura 4.4: Visualizações da Pergunta 1 do domínio de informações gerais.

treemap. Essas visualizações foram escolhidas por não ser possível prever o limite para o semestre do aluno, que pode assumir valores dispersos.

A partir das visualizações das Figuras 4.5a, 4.5b, 4.5c e 4.5d é possível observar que a maioria dos alunos cursam a disciplina no primeiro semestre, ao passo que o valor cai mais do que a metade para alunos cursando durante o segundo semestre do curso. Além disso, o semestre máximo que algum aluno já cursou as disciplinas foi no décimo sétimo. Como foram analisadas as disciplinas introdutórias de programação, era esperado que os alunos estivessem nos semestres iniciais da universidade.



(a) Gráfico de barras.

(b) Gráfico de densidade.

(c) Histograma.

(d) Treemap.

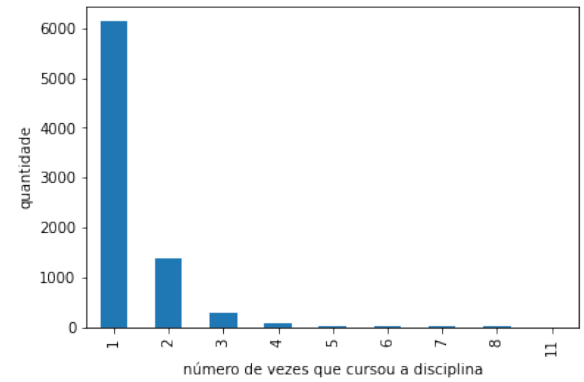
Figura 4.5: Visualizações da Pergunta 2 do domínio de informações gerais.

Pergunta 3: Quantas vezes os alunos cursam a disciplina?

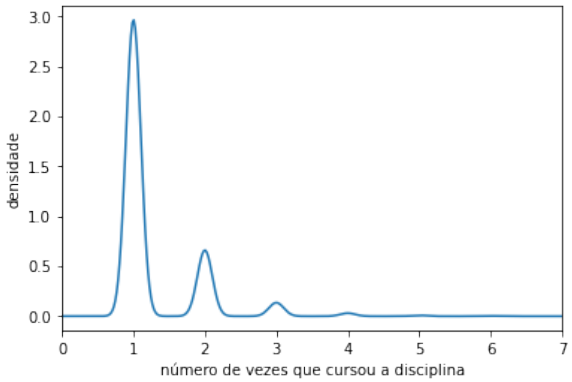
A terceira pergunta verifica quantas vezes o aluno acaba cursando a disciplina até passar ou desistir. Da mesma forma que a pergunta anterior, não é possível especificar a variação dos valores e, por essa razão, as visualizações candidatas são as mesmas - gráfico de barras, de densidade, histograma e treemap. Ao aplicar os dados nos algoritmos selecionados, foram geradas as visualizações apresentadas nas Figuras 4.6a, 4.6b, 4.6c e 4.6d. Isto posto, observa-se que a maioria dos alunos cursa a disciplina apenas uma vez e poucos cursam mais do que cinco vezes. Além disso, a quantidade máxima que um aluno frequentou a disciplina foi onze vezes.

Pergunta 4: Quais os cursos dos alunos que fizeram pela primeira vez a disciplina em comparação com os que fazem novamente?

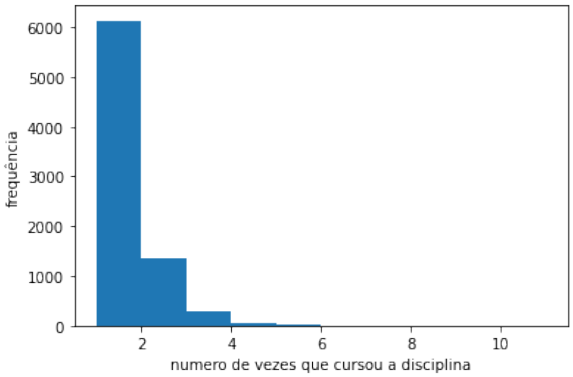
Para visualizar a relação entre os alunos que fizeram a disciplina pela primeira vez com os que fizeram mais de uma vez, foram utilizados os seguintes atributos: o curso dos alunos, as notas obtidas e a vez que o aluno cursou a disciplina. Os dados relacionados



(a) Gráfico de barras.



(b) Gráfico de densidade.



(c) Histograma.

(d) Treemap.

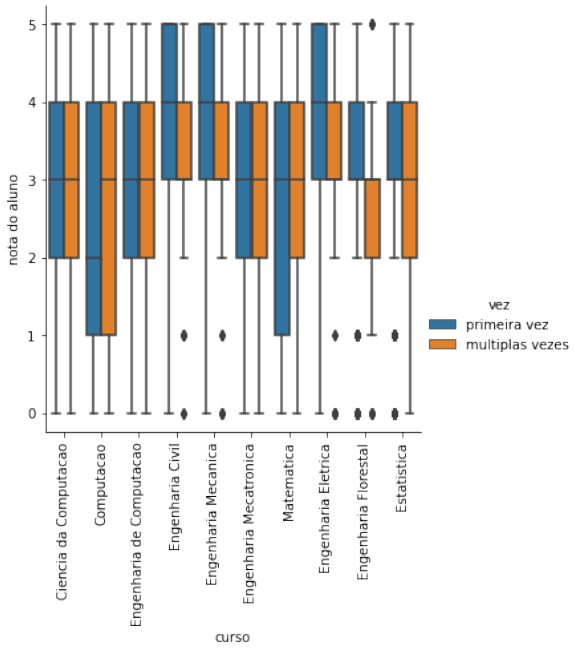
Figura 4.6: Visualizações da Pergunta 3 do domínio de informações gerais.

à primeira vez foram agrupados em uma categoria, enquanto que os outros dados foram para a categoria “outras vezes”.

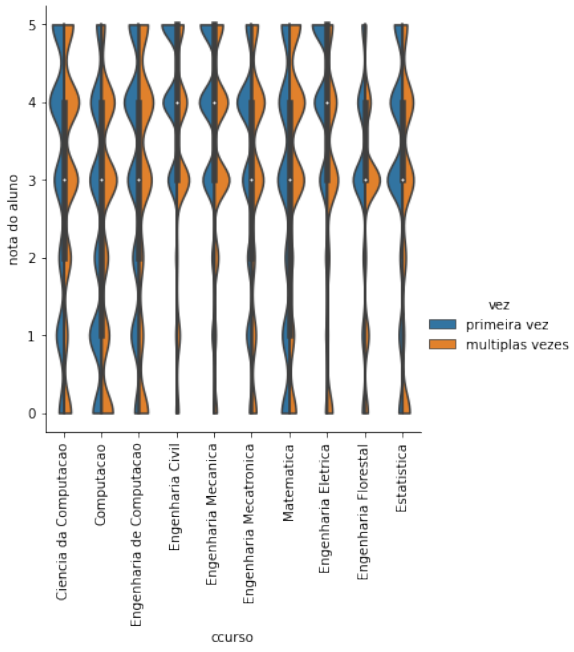
O gráfico de violino e o gráfico de caixas organizam grandes quantidades de dados numéricos, desta forma, foram utilizados como candidatos para comparar as notas das duas categorias, especificando os cursos. Já os gráficos de barras agruparam, em duas formas distintas, a quantidade de alunos de cada curso, segmentado pela categoria de vez. O algoritmo de coordenadas paralelas foi selecionado como candidato por ter como ponto positivo a comparação entre valores de dados que são de tipos e magnitudes diferentes. Por fim, foi utilizado um gráfico de barras em conjunto com o gráfico de linhas, em que as linhas representam as médias das notas dos alunos de cada curso, enquanto que as barras apresentam a quantidade de alunos, ambos segmentados pela categoria “vez cursando a disciplina”.

A partir da análise das visualizações das Figuras 4.7a, 4.7b, 4.7c, 4.7d, 4.7e e 4.7f, observa-se que o curso de Matemática possui a maior quantidade de alunos cursando as disciplinas analisadas, tanto pela primeira vez, quanto mais de uma vez. Já o curso que possui as maiores notas é Engenharia Civil, seguido por Engenharia Elétrica e Engenharia

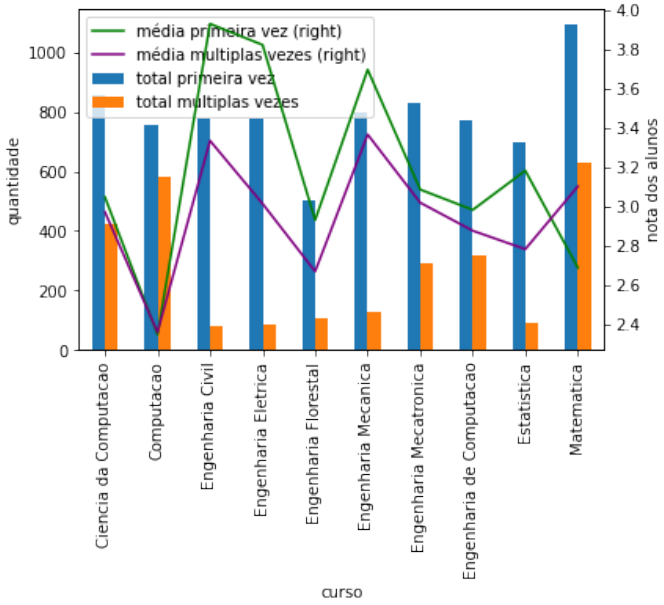
Mecânica. Os alunos que fazem pela primeira vez as disciplinas costumam tirar notas maiores do que os alunos que fazem novamente, exceto pelos alunos de Matemática. Além disso, o curso de Computação é o que possui a maior porcentagem de alunos fazendo novamente a disciplina, assim como é o curso com a menor média.



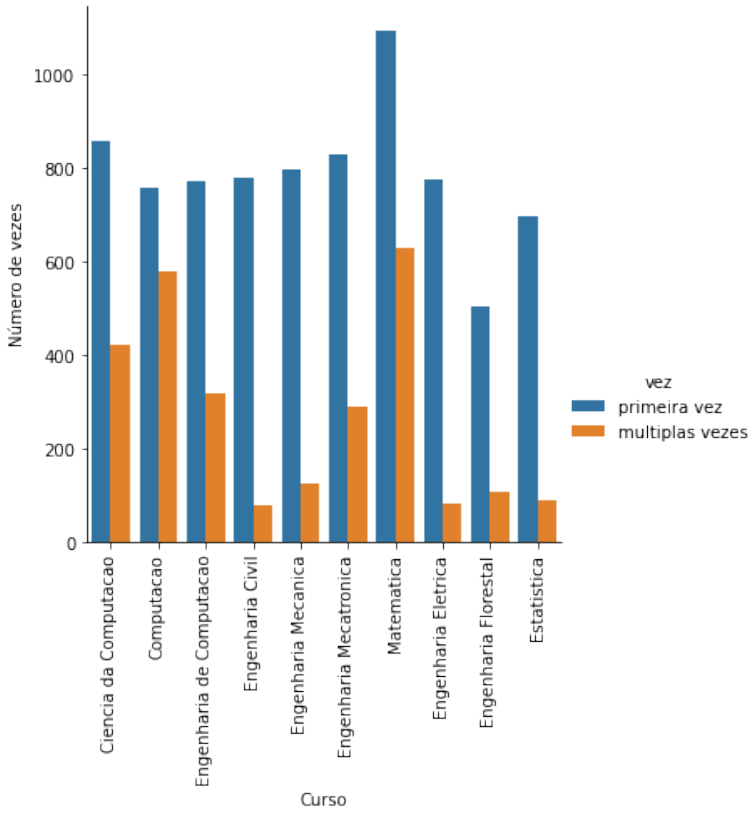
(a) Gráfico de caixas.



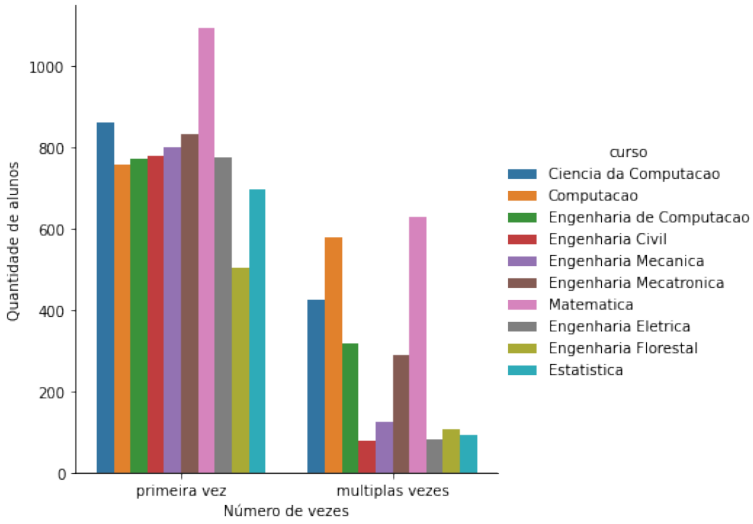
(b) Gráfico de violino.



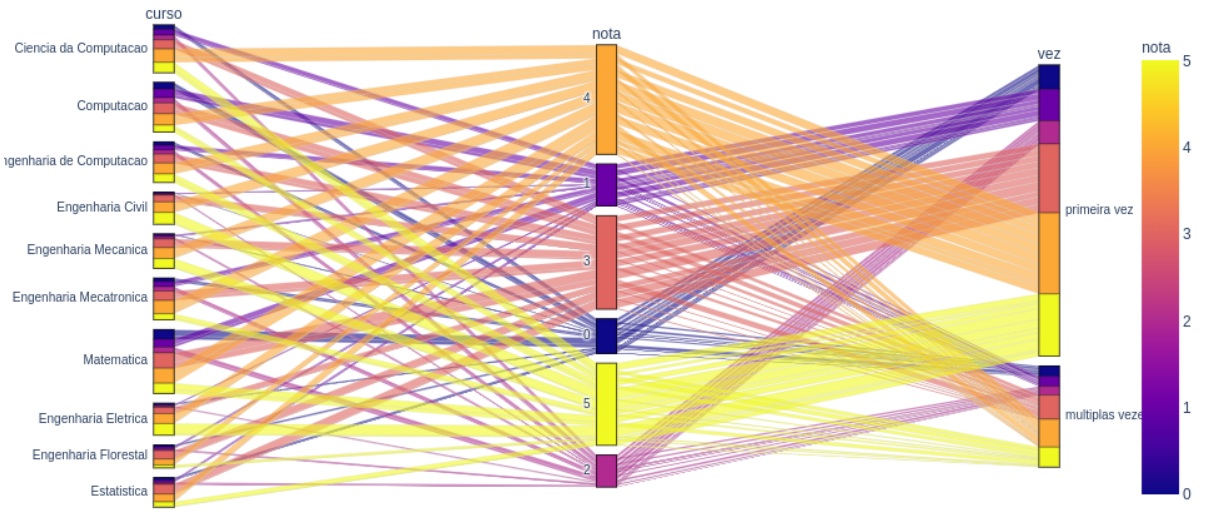
(c) Gráfico de barras com gráfico de linhas.



(d) Gráfico de barras agrupado pelo curso do aluno.



(e) Gráfico de barras agrupado pela vez em que o aluno cursou a disciplina.



(f) Coordenadas paralelas.

Figura 4.7: Visualizações da Pergunta 4 do domínio de informações gerais.

4.3.2 Domínio de Informações Acadêmicas

O domínio de informações acadêmicas descreve os resultados dos alunos na disciplina, expondo aos professores e gestores, o nível de dificuldade e se é necessário mais atenção aos estudantes. Esse domínio busca responder às seguintes perguntas:

1. Qual a evolução das notas dos alunos?
2. Qual a relação entre a quantidade de créditos cursados no semestre pelos alunos com o número de aprovações?
3. Quais são as métricas das notas dos alunos?
4. Qual a evolução da quantidade de alunos aprovados comparando com o total?
5. Qual a relação entre as notas e os créditos dos alunos que fizeram pela primeira vez a disciplina com as dos alunos que fazem novamente?

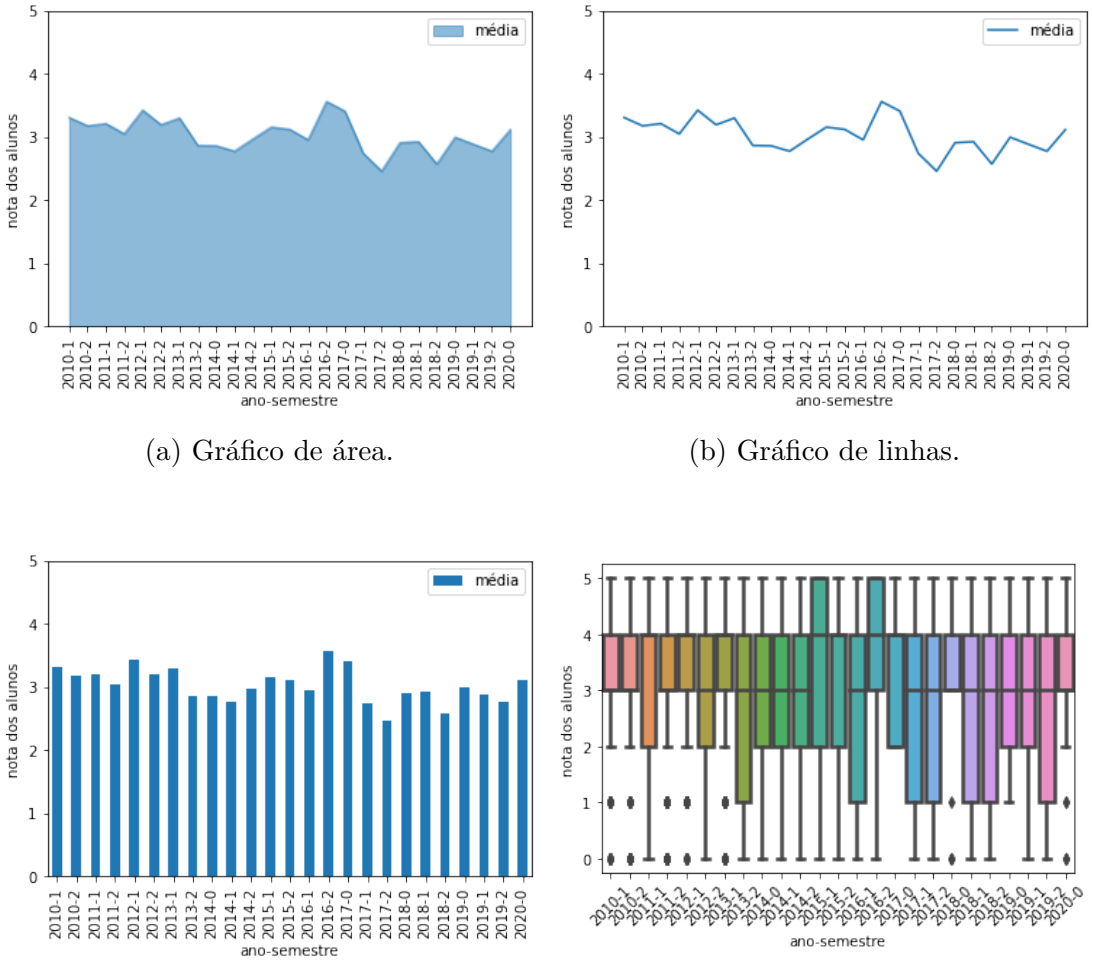
Assim como o domínio de informações gerais, as informações acadêmicas utilizaram os dados provenientes do SIGRA para responder às perguntas de pesquisa. Essa seleção é apresentada na Tabela 4.8 e detalhadas nas seções a seguir.

Tabela 4.8: Resultado do domínio de informações acadêmicas.

Perguntas	**Atributos**	**Algoritmos**
Pergunta 1	notas	Gráfico de Linhas
		Gráfico de Área
	semestre/ano	Gráfico de Barras
		Gráfico de Caixas
Pergunta 2	créditos cursados	Gráfico de Linhas e Gráfico de Área
	notas	
	créditos aprovados	Gráfico de Linhas e Gráfico de Barras
	semestre/ano	
Pergunta 3	notas	Gráfico de Densidade
		Gráfico de Caixas
		Gráfico de Violino
		Gráfico de Barras
		Gráfico de Pizza
		Gráfico de Radar
		Histograma
Pergunta 4	total de alunos	Gráfico de Área
		Gráfico de Linhas
	semestre/ano	Gráfico de Barras
Pergunta 5	créditos cursados	Coordenadas Paralelas
	semestre/ano	Gráfico de Bolhas
	nota	Gráfico de Linhas
	créditos aprovados	Gráfico de Linhas e Gráfico de Barras
	quantidade de vezes que cursou a disciplina	Gráfico de Linhas e Gráfico de Área

Pergunta 1: Qual a evolução das notas dos alunos?

De forma a visualizar a evolução das notas dos alunos, foram utilizados, do sistema SIGRA, as informações de semestre e as respectivas notas, de forma a obter o progresso das médias. Os gráficos de linhas, de área e de barras são utilizados para visualizações no tempo e, por esse motivo, foram selecionados como candidatos. Em relação ao gráfico de caixas, ao plotar vários em um plano cartesiano, obtém-se a ideia de progresso, além de evidenciar mais informações do que apenas a média, tais como: localização, dispersão, assimetria, comprimento da cauda e medidas discrepantes. A partir das visualizações das Figuras 4.8a, 4.8b, 4.8c e 4.8d, observa-se que a média variou, mas a oscilação não foi expressiva. A maior média foi no segundo semestre de 2016 (2016-2).



(a) Gráfico de área.

(b) Gráfico de linhas.

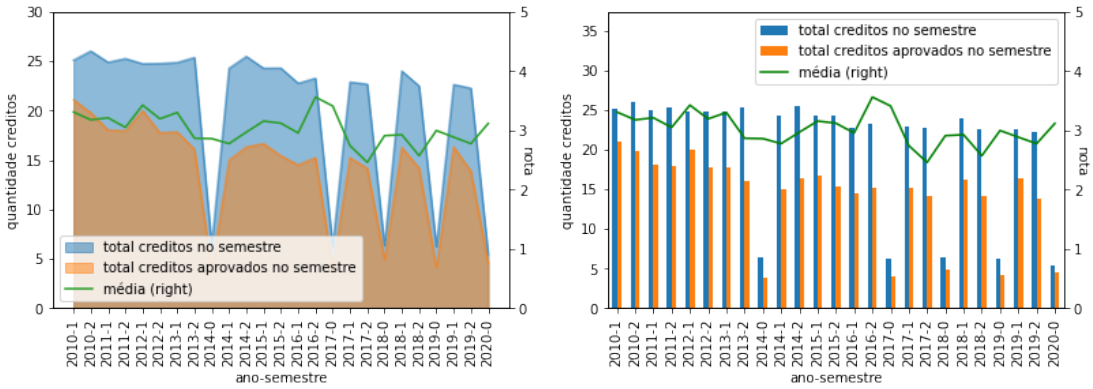
(c) Gráfico de barras.

(d) Gráfico de caixas.

Figura 4.8: Visualizações da Pergunta 1 do domínio de informações gerais.

Pergunta 2: Qual a relação entre a quantidade de créditos cursados no semestre pelos alunos com o número de aprovações?

De forma a visualizar a relação entre a quantidade de créditos cursados com a quantidade de créditos aprovados, foram utilizados os atributos referentes às notas e aos créditos dos alunos. Nesse sentido, o gráfico de linhas foi combinado com o gráfico de área e de barras, apresentados nas Figuras 4.9a e 4.9b. É possível verificar que nos semestres de verão (2014-0, 2017-0, 2018-0, 2019-0 e 2020-0) os alunos costumam cursar menos disciplinas, entretanto costuma haver um maior percentual de aprovações, ou seja, do total de créditos cursados, os alunos possuem mais créditos aprovados.

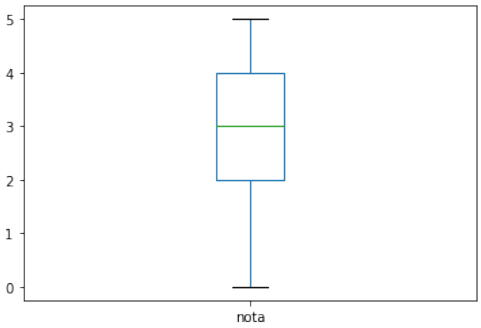


(a) Gráfico de área e gráfico de linhas. (b) Gráfico de barras e gráfico de linhas.

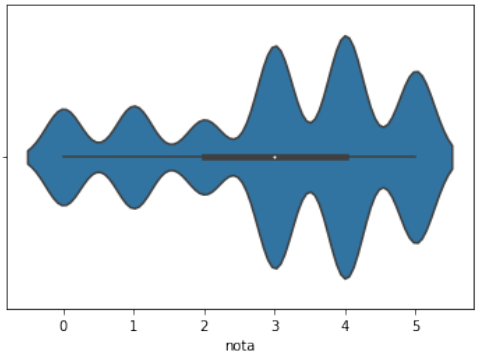
Figura 4.9: Visualizações da Pergunta 2 do domínio de informações acadêmicas.

Pergunta 3: Quais são as métricas das notas dos alunos?

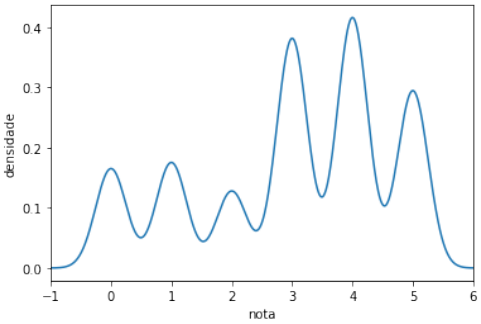
Métricas são medidas quantificáveis usadas para analisar o resultado de um processo. Dessa maneira, o objetivo desta pergunta de pesquisa é avaliar as notas dos alunos de forma a obter informações para tomada de decisões. Tanto o gráfico de barras, de pizza e de radar apresentam informações de forma clara e simples para o usuário, entretanto a informação transmitida é apenas sobre a quantidade. O gráfico de caixas apresenta medidas de estatísticas descritivas como o mínimo, máximo, primeiro quartil, segundo quartil ou mediana e o terceiro quartil, além dos valores discrepantes, quando existentes. O gráfico de violino apresenta as mesmas informações que o gráfico de caixas, somado à informação de densidade de probabilidade. O gráfico de densidade descreve o padrão de distribuição das notas, assim como o histograma. A partir das visualizações das Figuras 4.10a, 4.10b 4.10c, 4.10d, 4.10e, 4.10f e 4.10g, é possível identificar os quartils, o primeiro é próximo de 2, o segundo (a média) no valor 3, o terceiro é na nota 4. Além disso, a partir da distribuição das notas é possível perceber que as maiores notas são 3 (MM), 4 (MS) e 5 (SS).



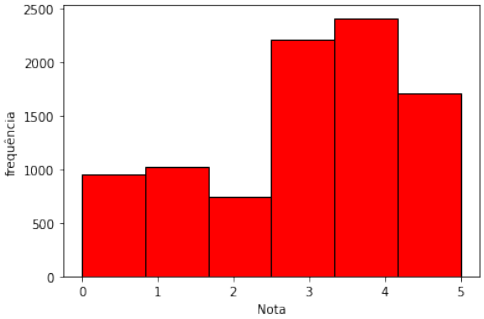
(a) Gráfico de caixas.



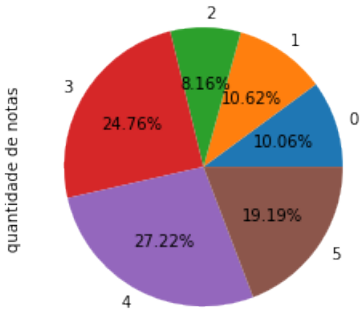
(b) Gráfico de violino.



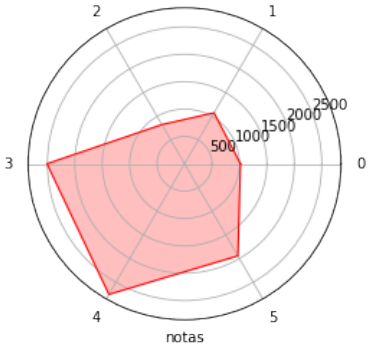
(c) Gráfico de densidade.



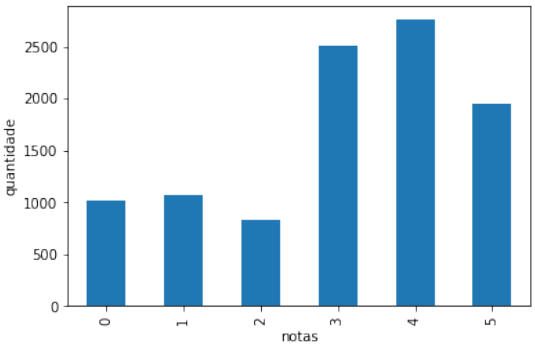
(d) Histograma.



(e) Gráfico de pizza.



(f) Gráfico de radar.

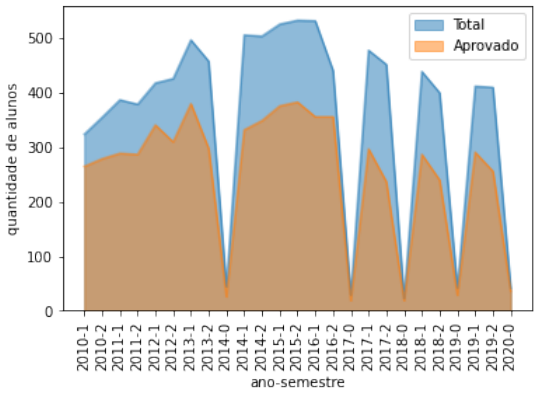


(g) Gráfico de barras.

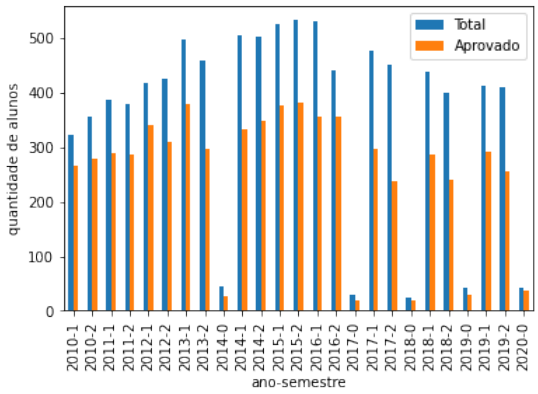
Figura 4.10: Visualizações da Pergunta 3 do domínio de informações acadêmicas.

Pergunta 4: Qual a evolução da quantidade de alunos aprovados comparando com o total?

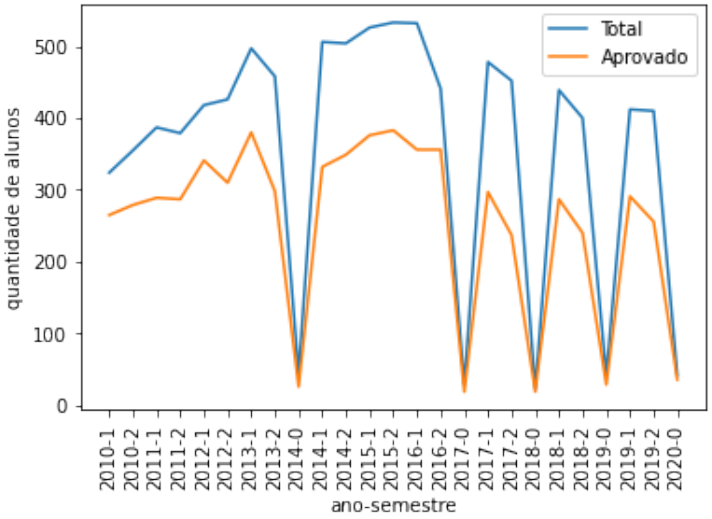
Esta pergunta tem dois objetivos - analisar a evolução no tempo e comparar a quantidade de alunos aprovados em relação ao total de alunos que cursaram a disciplina. Por utilizar duas categorias (estudantes aprovados e o total), apenas os gráficos de linhas, de área e de barras foram selecionados por apresentarem a evolução no tempo. A partir das visualizações das Figuras 4.11a, 4.11b, e 4.11c, observa-se que uma menor quantidade de alunos frequentou a disciplina nos semestres de verão, mas nesses semestres a proporção de alunos aprovados é maior. Já o semestre com maior número de alunos foi 2015-2.



(a) Gráfico de área.



(b) Gráfico de barras.

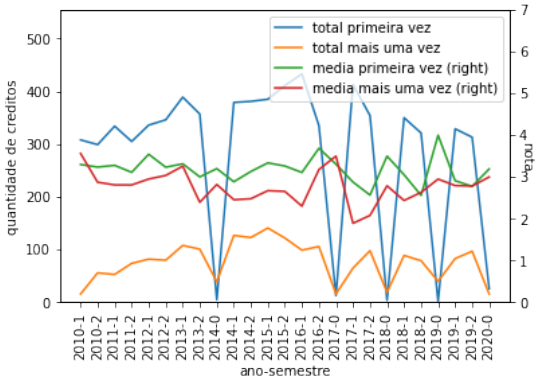


(c) Gráfico de linhas.

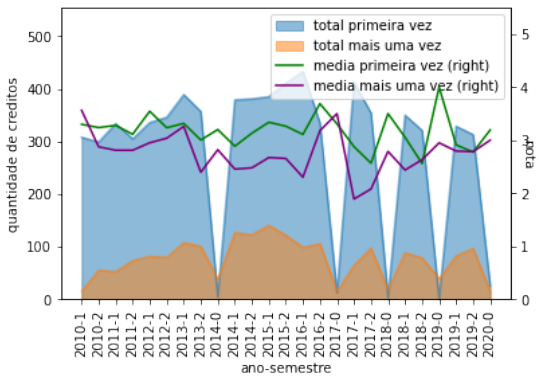
Figura 4.11: Visualizações da Pergunta 4 do domínio de informações acadêmicas.

Pergunta 5: Qual a relação entre as notas e os créditos dos alunos que fizeram pela primeira vez a disciplina com as dos alunos que fazem novamente?

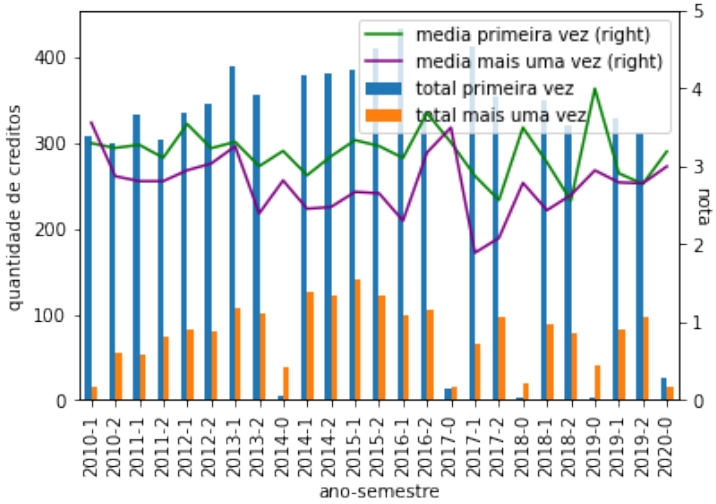
Foram utilizados os seguintes atributos para responder a quarta pergunta: créditos cursados e aprovados, nota dos alunos, quantidade de vezes que o aluno fez a disciplina e o semestre. O gráfico de linhas com o gráfico de barras e com o gráfico de área são utilizados de forma a identificar todos os atributos selecionados, bem como o gráfico de bolhas e o gráfico de coordenadas paralelas. As Figuras 4.12a, 4.12b, 4.12c, 4.12d e 4.12e apresentam as visualizações geradas, em que é possível verificar que os alunos que fazem as disciplinas nos semestre de verão, são em maioria, alunos fazendo novamente a disciplina. Entretanto, esses alunos costumam pegar menos créditos totais e possuem média de notas menores.



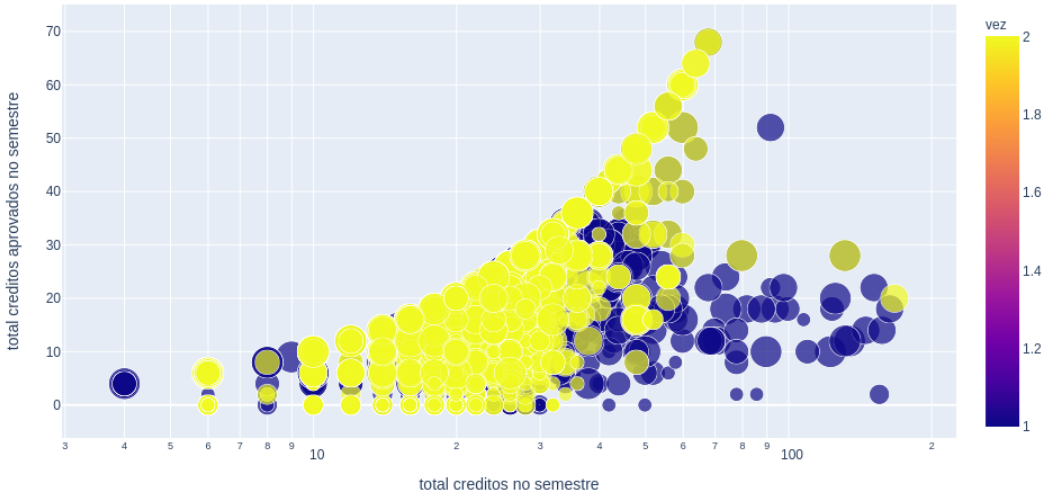
(a) Gráfico de linhas.



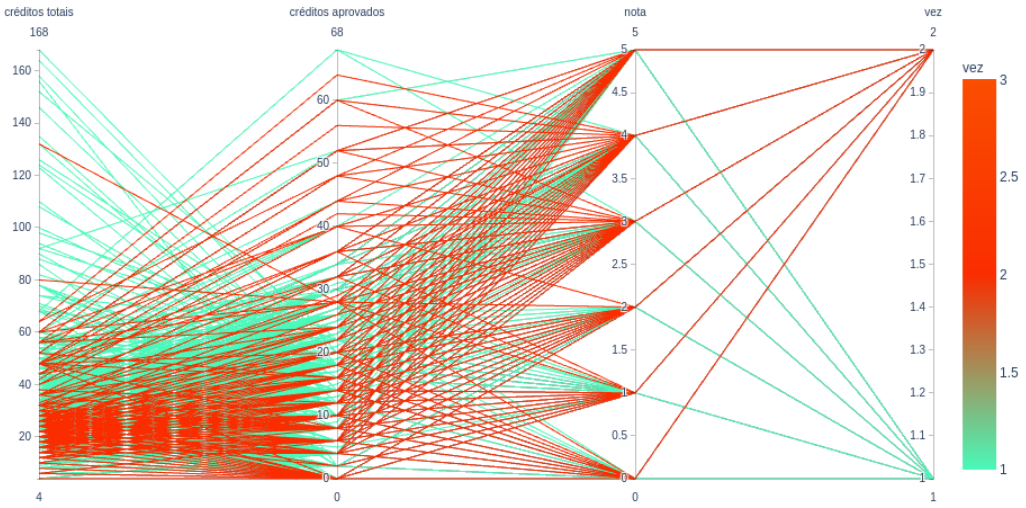
(b) Gráfico de área e gráfico de linhas.



(c) Gráfico de barras e gráfico de linhas.



(d) Gráfico de bolhas.



(e) Coordenadas paralelas.

Figura 4.12: Visualizações da Pergunta 5 do domínio de informações acadêmicas.

4.3.3 Domínio de Informações de Percepção

As informações de percepção são importantes para a descoberta de conhecimento por parte do usuário acerca da compreensão do aluno a respeito do curso e de sua situação, a fim de direcionar os estudos de forma mais eficaz. Esse domínio busca responder às seguintes perguntas:

1. Os alunos já entram na disciplina com contato prévio com programação?

2. Com qual linguagem os alunos já tiveram contato?
3. Os alunos costumam escutar que Engenharia/Computação é para pessoas super inteligentes?
4. Os alunos costumam escutar que Engenharia/Computação é para homens?
5. Qual a relação entre área de trabalho dos pais com o nível educacional dos pais?
6. Qual a relação entre a quantidade de horas estudadas com a percepção de que os alunos precisam ou não de mais horas de estudos do que os colegas?

Para responder às perguntas do domínio de informações de percepção, utilizou-se informações provenientes do questionário. Os dados e os algoritmos selecionados para responder cada uma das questões são apresentados nas seções abaixo e sintetizados na Tabela 4.9.

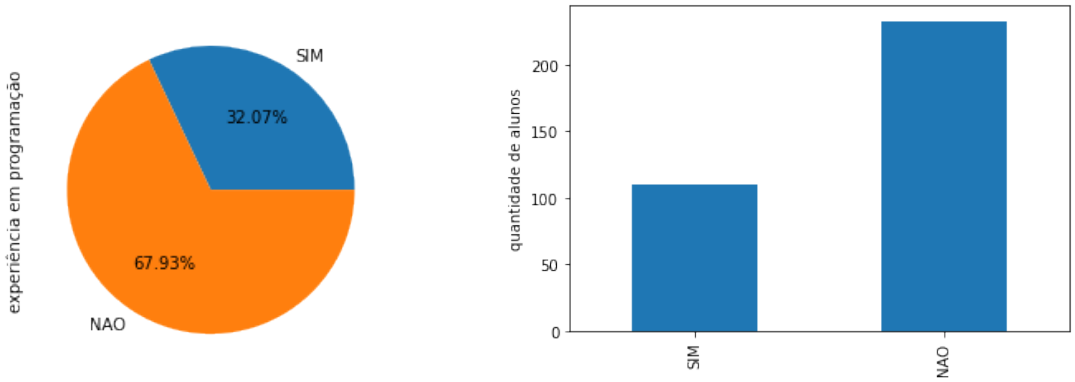
Tabela 4.9: Resultado do domínio de informações de percepção.

Perguntas de pesquisa	**Atributos**	**Algoritmos**
Pergunta 1	experiência em programação	Gráfico de Pizza
		Gráfico de Barras
Pergunta 2	linguagem de experiência	Gráfico de Pizza
		Gráfico de Barras
Pergunta 3	escutam que computação é para pessoas inteligentes	Gráfico de Pizza
		Gráfico de Barras
		Gráfico de Barras Empilhadas
Pergunta 4	escutam que matemática é para homens	Gráfico de Pizza
		Gráfico de Barras
		Gráfico de Barras Empilhadas
Pergunta 5	área de trabalho dos pais	Coordenadas Paralelas
	país tem educação superior	Diagrama de Sankey
		Gráfico de Barras
Pergunta 6	horas estudadas	Coordenadas Paralelas
	precisa de mais horas de estudo	Diagrama de Sankey
		Gráfico de Barras

Pergunta 1: Os alunos já entram na disciplina com contato prévio com programação?

De forma a verificar o conhecimento prévio em programação com que os alunos entram na disciplina, foram analisadas as respostas da questão do questionário: “Você já tinha

experiência em programação". A resposta pode ser "Sim" ou "Não", desta forma, para representar os dados categóricos foram selecionados o gráfico de pizza e o de barras como candidatos. As Figuras 4.13a e 4.13b apresentam as visualizações geradas. Em ambas as visualizações é fácil perceber que a maior parte dos alunos não teve contato com programação antes da disciplina. Entretanto, no gráfico de barras é possível visualizar a diferença absoluta, enquanto que no gráfico de pizza observa-se a diferença em porcentagem.



(a) Gráfico de pizza.

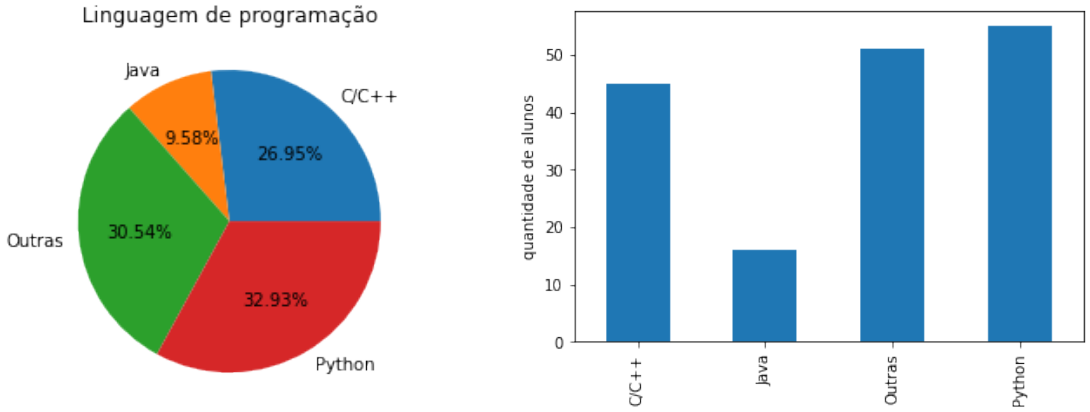
(b) Gráfico de barras.

Figura 4.13: Visualizações da Pergunta 1 do domínio de informações de percepção.

Pergunta 2: Com qual linguagem os alunos já tiveram contato?

A segunda pergunta complementa a anterior, em que foi verificado se os alunos tinham experiência em programação antes da disciplina. A presente pergunta tem como objetivo identificar quais as linguagens de programação que os alunos tiveram contato. Nesse sentido, foram utilizadas as respostas da seguinte questão do questionário: "Se você tinha experiência em programação antes dessa disciplina, informe em quais linguagens". A resposta do aluno poderia ser uma ou mais dentre as opções: "C/C++", "Java", "Python" e "Outras". Para apresentar os dados categóricos, os gráficos de pizza e de barras foram selecionados como candidatos e são apresentados nas Figuras 4.14a e 4.14b.

Entre as visualizações geradas, o gráfico de barras permite que o usuário visualize mais facilmente a ordem das linguagens de programação que são mais conhecidas. Com uma quantidade maior de categorias, o gráfico de pizza acaba tornando a comparação mais demorada. Nos gráficos obtidos, é possível perceber que a linguagem Python é a mais conhecida.



(a) Gráfico de pizza.

(b) Gráfico de barras.

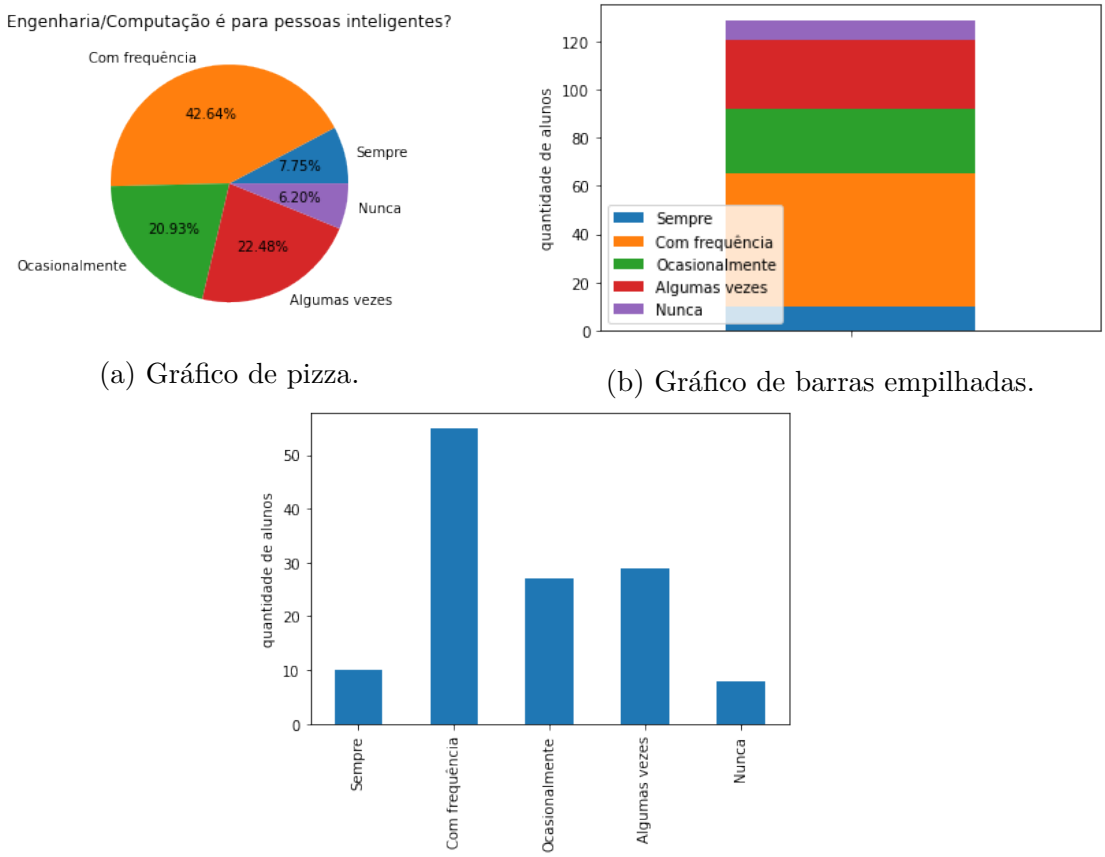
Figura 4.14: Visualizações da Pergunta 2 do domínio de informações de percepção.

Pergunta 3: Os alunos costumam escutar que Engenharia/Computação é para pessoas super inteligentes?

A terceira pergunta utiliza uma das questões de percepção do questionário - "Com que frequência você ouve 'Engenharia/Computação é para pessoas super inteligentes' ". A resposta consiste de valores que expressam categorias, além de seguir uma ordem, logo os algoritmos candidatos são o gráfico de barras e de barras empilhadas, além do gráfico de pizza. A partir das visualizações apresentadas nas Figuras 4.15a, 4.15b e 4.15c, conclui-se que a maior parte dos alunos tomam conhecimento da frase com frequência, enquanto que são poucos os que nunca escutaram.

Pergunta 4: Os alunos costumam escutar que Engenharia/Computação é para homens?

A quarta pergunta utiliza as respostas para a questão do questionário: "Com que frequência você ouve 'Engenharia/Computação é para homens' ". Assim como a pergunta anterior, as respostas consistem de valores que expressam categorias, além de seguir uma ordem e, por essa razão, os gráfico candidatos são o gráfico de pizza, de barras e barras empilhadas, que são ilustrados nas Figuras 4.16a, 4.16b e 4.16c. A maior parte dos que responderam o questionário disseram que nunca ouviram essa frase, diferentemente da pergunta anterior, em que os alunos disseram que costumam ouvir que o curso é para pessoas inteligentes.



(a) Gráfico de pizza.

(b) Gráfico de barras empilhadas.

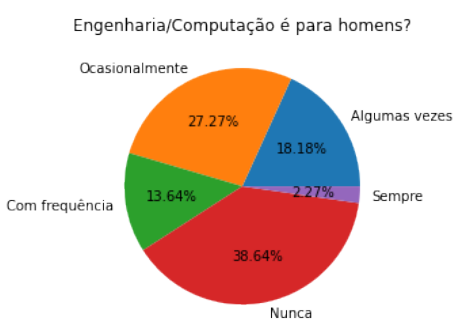
(c) Gráfico de barras.

Figura 4.15: Visualizações da Pergunta 3 do domínio de informações de percepção.

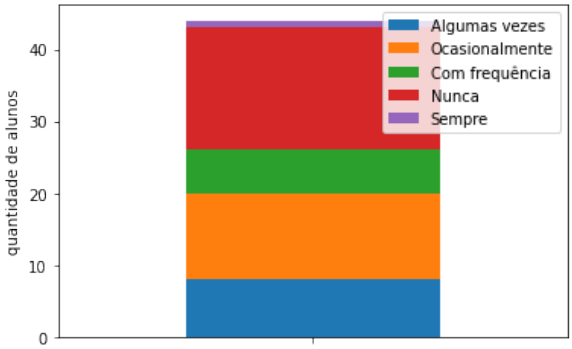
Pergunta 5: Qual a relação entre área de trabalho dos pais com o nível educacional dos pais?

Para visualizar a relação entre a área de trabalho dos pais dos alunos com o nível educacional dos mesmos, foram utilizadas três questões do questionário: "Em que área os seus pais e irmãos trabalham?", "Seu pai e/ou sua mãe terminaram algum curso superior?" e "Você entrou pelo sistema de cotas da UnB?". As respostas para as perguntas são categóricas e, por essa razão, foram selecionados os seguintes algoritmos candidatos: gráfico de barras, diagrama de sankey e coordenadas paralelas.

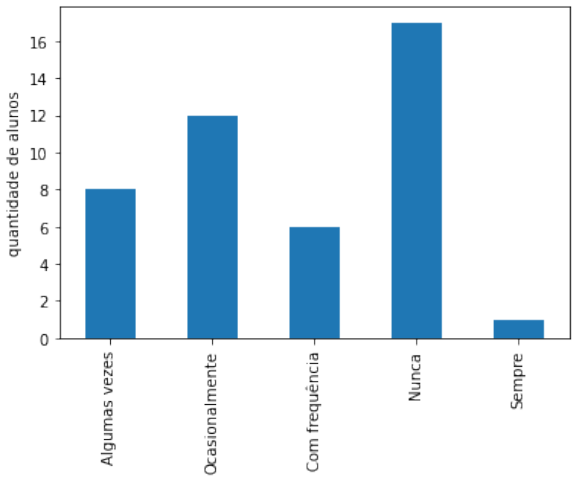
O gráfico de coordenadas paralelas permite a comparação entre categorias distintas, assim como o diagrama de sankey enfatiza o fluxo entre as categorias e permite a comparação de mais do que dois grupos. Entretanto, esses gráficos são pouco conhecidos e podem confundir o usuário. O gráfico de barras não permite a comparação das três perguntas, entretanto, ao agrupar em formas distintas, diferentes informações são enfatizadas. As visualizações são apresentadas nas Figuras 4.17a, 4.17b, 4.17c e 4.17d. Assim, é possível analisar que a maioria dos pais não possuem educação superior e trabalham na área da



(a) Gráfico de pizza.



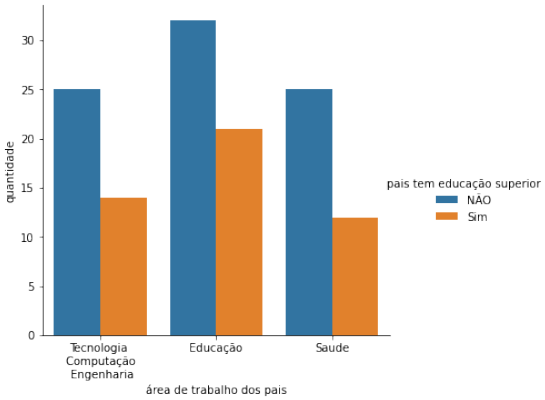
(b) Gráfico de barras empilhadas.



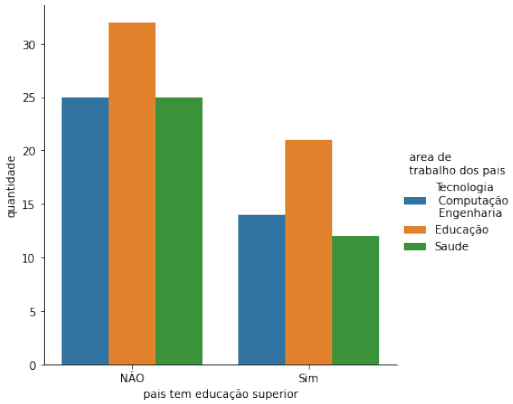
(c) Gráfico de barras.

Figura 4.16: Visualizações da Pergunta 4 do domínio de informações de percepção.

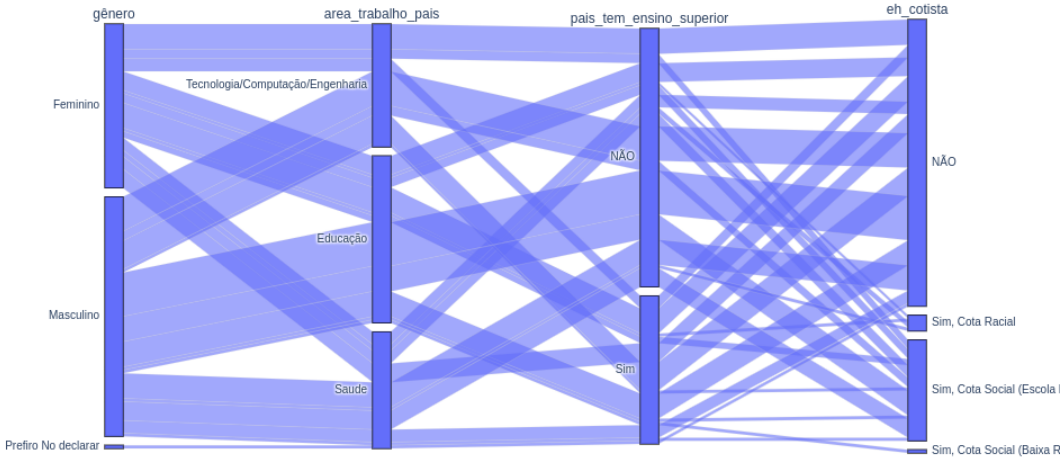
educação. Além disso, em relação às cotas, os pais dos alunos que entram na faculdade por meio do sistema de cotas de escola pública não costumam ter ensino superior e trabalham na área da Saúde.



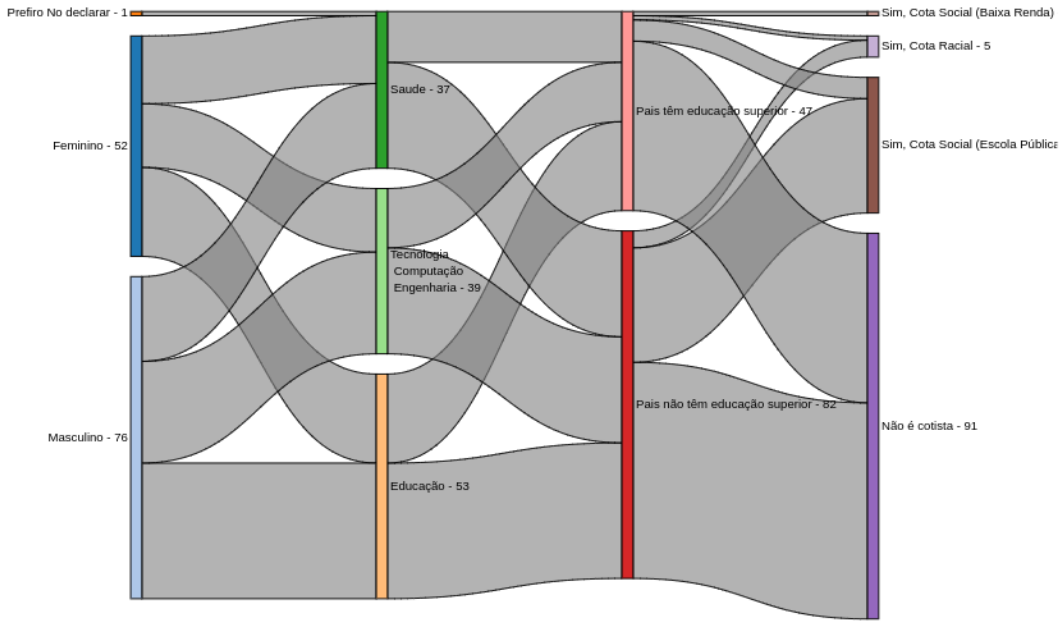
(a) Gráfico de barras agrupado pela área de trabalho dos país.



(b) Gráfico de barras agrupado pelo nível superior dos pais.



(c) Coordenadas paralelas.



(d) Diagrama de sankey.

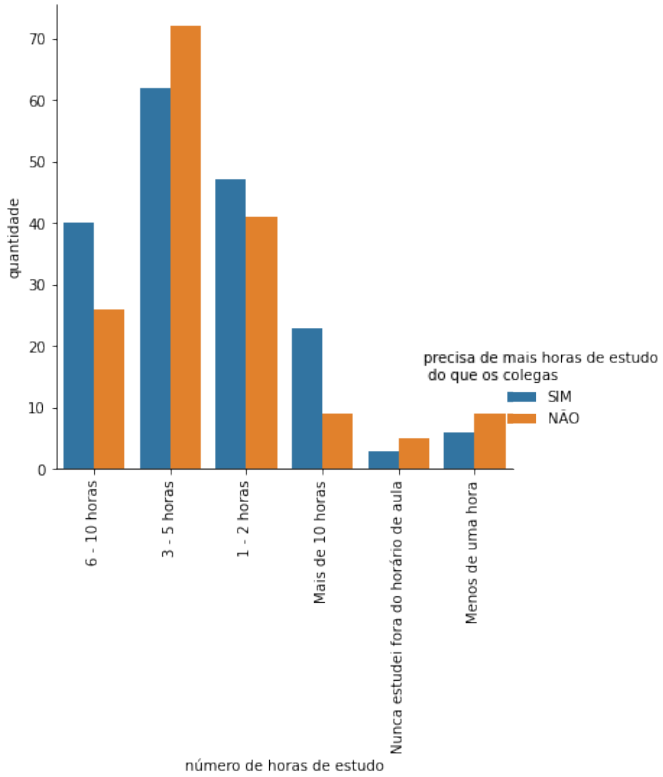
Figura 4.17: Visualizações da Pergunta 5 do domínio de informações de percepção.

Pergunta 6: Qual a relação entre a quantidade de horas estudadas com a percepção de que os alunos precisam ou não de mais horas de estudos do que os colegas?

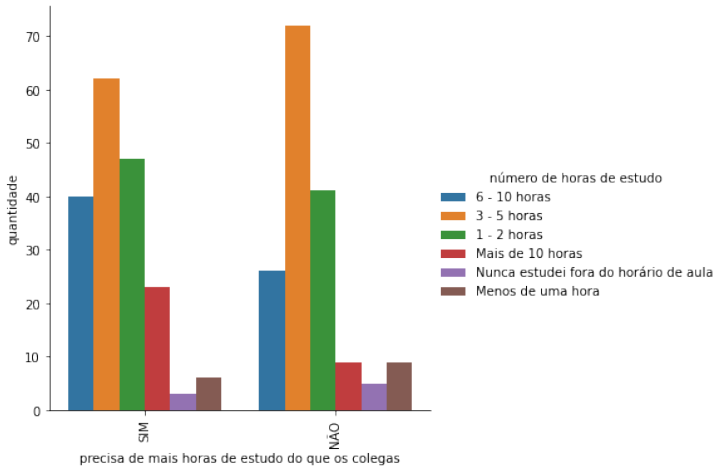
Com o intuito de visualizar a relação entre a quantidade de horas estudadas pelos alunos fora da sala de aula com a percepção dos mesmos de que precisam, ou não, de mais

horas de estudo que os seus colegas, foram utilizadas as respostas de duas questões do questionário: “Durante o semestre, quantas horas você passou fazendo os deveres de casa e estudando fora da sala de aula por semana para essa disciplina” e “Você acha que precisa de mais horas de estudos nesta disciplina do que seus amigos de sala?”. Para visualizar as informações, foram selecionados os seguintes algoritmos candidatos: gráfico de barras, diagrama de sankey e coordenadas paralelas.

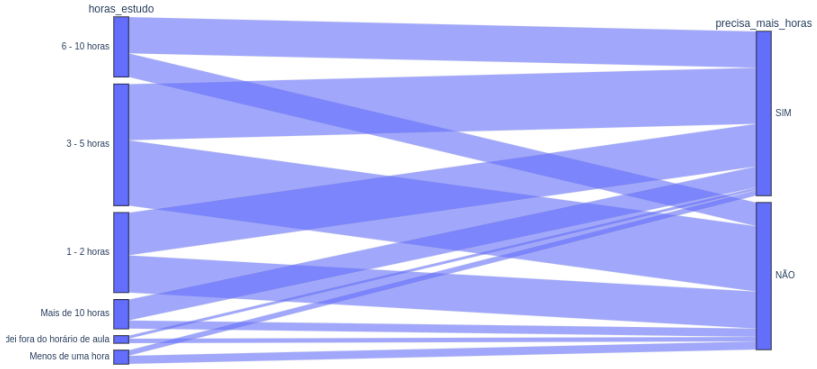
De forma a comparar apenas duas perguntas, o gráfico de barras é capaz de transmitir a informação desejada, assim como o gráfico de coordenadas paralelas que permite a comparação entre categorias distintas e como o diagrama de sankey que enfatiza o fluxo e permite a comparação entre os grupos. As visualizações são apresentadas na Figuras 4.18a, 4.18b, 4.18c e 4.18d, em que a maioria dos alunos estudaram entre 3 a 5 horas fora da sala de aula. Entretanto, os alunos que estudaram mais do que 10 horas fora da sala de aula acreditam que precisam de mais horas de estudo que os colegas, enquanto que os alunos que nunca estudaram fora do horário de aula não acreditam que precisam de mais horas de estudo que os amigos.



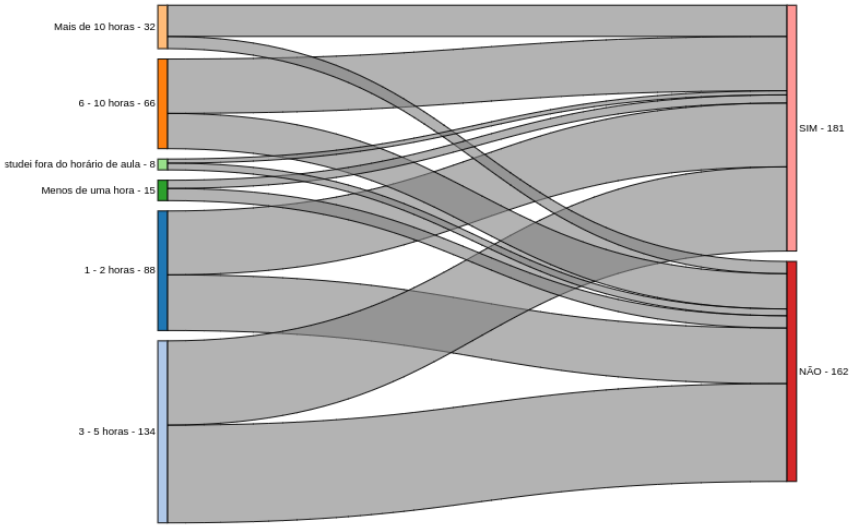
(a) Gráfico de barras agrupado pela quantidade de horas de estudo.



(b) Gráfico de barras agrupado pela necessidade de mais horas de estudo.



(c) Coordenadas paralelas.



(d) Diagrama de sankey.

Figura 4.18: Visualizações da Pergunta 6 do domínio de informações de percepção.

4.4 Realização de Experimentos

Na revisão de literatura apresentada na Seção 3.2 foram explicitadas diferentes formas de avaliação de sistemas de visualização (questionário e caso de uso). De maneira geral, os trabalhos na literatura propõem a condução de pesquisas com o usuário, de modo a objetivar a avaliação da qualidade da informação que é apresentada nas representações gráficas [50] [51]. Desta forma, foi elaborado um questionário a ser aplicado aos professores e gestores.

O questionário possui dois tipos de perguntas: questões iniciais que visam a analisar o conhecimento do questionado sobre as visualizações; e questões elaboradas, cujo objetivo é determinar se as visualizações candidatas comunicam a informação correta e de forma clara e prática. Utilizou-se como fundamento o trabalho de Valiati [52], em que as questões iniciais pertencem ao questionário para avaliação de perfil do usuário utilizado nos ensaios de interação. As questões elaboradas foram desenvolvidas com base nas questões/tarefas utilizadas nos ensaios de interação.

A avaliação das visualizações consistiu na análise das respostas das questões elaboradas, levando em consideração: o tempo em que o questionado leva para identificar a informação e responder e a corretude da resposta. Estas foram classificadas como: "corretas", caso a resposta seja a esperada; "errada", caso a resposta seja diferente da esperada; e "não sabe", quando o questionado preferiu não responder por não conseguir identificar a informação. Em relação ao tempo de solução, foram definidos os seguintes intervalos: 1 a 4 segundos, 5 a 9 segundos, 10 a 15 segundos e acima de 15 segundos. As respostas "não sei" não foram contabilizadas para o critério relacionado com o tempo de resposta. Os resultados da aplicação do questionário são apresentados no próximo capítulo.

4.5 Considerações finais

A metodologia que integra este trabalho contou com o estudo de duas bases de dados (SIGRA e as respostas de um questionário aplicado às turmas de disciplinas introdutórias de programação da UnB) e de 19 algoritmos de visualização distintos selecionados a partir dos algoritmos mais utilizados na literatura na área de análise visual de dados educacionais. Na etapa seguinte, três domínios de informações foram identificados (informações gerais, informações acadêmicas e informações de percepção), em que perguntas foram compostas para guiar a elaboração das visualizações, bem como a aplicação do caso de uso. Por fim, foi aplicado um questionário para professores e gestores educacionais com o objetivo de avaliar as visualizações geradas. No próximo capítulo são apresentados os resultados do questionário.

Capítulo 5

Questionário

Para avaliar a comunicação das informações por meio das visualizações, foram elaborados três questionários distintos, um para cada domínio (informações acadêmicas, gerais e de percepção), com duração de 30 minutos, conforme especificado na Seção 4.4. O questionário de informações acadêmicas foi realizado com sete usuários distintos, ao passo que os questionários de informações gerais e de informações de percepção foram aplicados, cada um, a oito pessoas. Isto posto, um total de 23 professores/gestores foram questionados.

Neste capítulo são relatados os experimentos realizados para validação dos algoritmos candidatos por meio do questionário, avaliando as respostas das questões iniciais e das questões elaboradas. As análises são apresentadas nas Seções 5.1 e 5.2, respectivamente e os resultados são concluídos na Seção 5.3.

5.1 Questões Iniciais

A partir das perguntas iniciais, foi possível avaliar o nível inicial de conhecimento dos questionados sobre as visualizações utilizadas [52]. As questões são apresentadas abaixo.

1. Qual a sua área de atuação/atividade atual?
2. Qual a sua formação no nível de graduação?
3. Qual a sua formação no nível de pós-graduação?
4. Qual o seu nível de conhecimento sobre técnicas de visualização de informações?
5. Qual a sua experiência no uso de técnicas de visualização de informações?
6. Quais dessas técnicas você utilizou?

A partir das respostas a essas perguntas, observou-se que o gráfico menos conhecido é o gráfico de violino, seguido pelo diagrama de sankey e pelo gráfico de coordenadas

paralelas. Porém, todos os professores e gestores educacionais já tinham conhecimento prévio do gráfico de barras, de barras empilhadas, de pizza e do gráfico de caixas. A Tabela 5.1 apresenta a quantidade de professores/gestores que conheciam, ou não, cada uma das visualizações utilizadas nos questionários.

Tabela 5.1: Nível de conhecimento dos questionados sobre as visualizações.

Gráfico	**Conhece**	**Não conhece**	**Total**
Pizza	16	0	16
Barras	23	0	23
Radar	4	4	8
Linhas	14	1	15
Treemap	3	5	8
Coordenadas Paralelas	7	16	23
Caixas	15	0	15
Violino	2	13	15
Densidade	13	2	15
Histograma	13	2	15
Área	5	2	7
Bolhas	3	4	7
Sankey	2	6	8
Barras Empilhadas	8	0	8

5.2 Questões Elaboradas

As questões elaboradas, diferentemente das questões iniciais, variam a cada domínio. Isto posto, esta seção apresenta os resultados do tempo de resposta e da corretude de cada questão em cada domínio.

5.2.1 Domínio de Informações Gerais

No domínio de informações gerais foram elaboradas 4 perguntas especificadas na Seção 4.3.1 e selecionadas 15 visualizações candidatas para responder estas perguntas. Nesta seção, é detalhada a avaliação dos gráficos selecionados para este domínio.

Pergunta 1: Qual a quantidade de aprovações/reprovações nas disciplinas?

O gráfico de pizza e o gráfico de barras foram escolhidos como candidatos para apresentar a informação relativa à primeira pergunta do domínio de informações gerais. De forma a

avaliar as visualizações, foi realizado o seguinte questionamento: “Qual a diferença entre a quantidade de alunos que passaram e que reprovaram a disciplina?”. Em vista disso, dos oito professores/gestores abordados para responder ao questionário do domínio das informações gerais, todos acertaram a resposta ao utilizar o gráfico de barras, enquanto que dois erraram ao empregar o gráfico de pizza. Entretanto, em ambas as visualizações, a maior parte dos questionados demorou mais do que 15 segundos para identificar a informação e responder. O resultado relacionado a esta questão é apresentado na Tabela 5.2.

Tabela 5.2: Resultado dos algoritmos da Pergunta 1 do domínio de informações gerais.

	Tempo de resposta (segundos)				**Resposta**		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Barras (4.4a)		3	1	4	6	2	
Pizza (4.4b)		3	1	4	8		

C - resposta correta E - resposta errada NS - resposta ‘não sei’

Pergunta 2: Em qual semestre os alunos cursam a disciplina?

Na segunda pergunta, o gráfico de barras e de densidade, o treemap e o histograma, foram selecionados como candidatos. De forma a avaliar os algoritmos, foi elaborada a seguinte questão: “Em qual semestre os alunos costumam cursar a disciplina?”. Para tanto, das 8 pessoas que responderam o questionário, quatro erraram a resposta durante a visualização do histograma. Além disso, os questionados responderam mais rapidamente no gráfico de barras, seguido pelo treemap, gráfico de densidade e, por último, no histograma. O resultado da aplicação do questionário referente a esta pergunta é apresentado na Tabela 5.3.

Tabela 5.3: Resultado dos algoritmos da Pergunta 2 do domínio de informações gerais.

	Tempo de resposta (segundos)				**Resposta**		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Barras (4.5a)	6			2	8		
Densidade (4.5b)	4	1	2	1	8		
Histograma (4.5c)	3		3	2	4	4	
Treemap (4.5d)	5			2	7		1

C - resposta correta E - resposta errada NS - resposta ‘não sei’

Pergunta 3: Quantas vezes os alunos cursam a disciplina?

A seguinte questão foi elaborada para avaliar os algoritmos da terceira pergunta: "A maioria dos alunos tendem a cursar as disciplinas quantas vezes?". Todos os 8 professores/gestores acertaram a resposta nos quatro gráficos candidatos. Entretanto, os usuários responderam mais rápido ao visualizar os dados utilizando o treemap, seguido pelo gráfico de barras, gráfico de densidade e, por último, no histograma. A Tabela 5.4 evidencia os resultados.

Tabela 5.4: Resultado dos algoritmos da Pergunta 3 do domínio de informações gerais.

	Tempo de resposta (segundos)				**Resposta**		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Barras (4.6a)	6	2			8		
Densidade (4.6b)	5	3			8		
Histograma (4.6c)	4	1	2	1	8		
Treemap (4.6d)	8				8		

C - resposta correta E - resposta errada NS - resposta 'não sei'

Pergunta 4: Quais os cursos dos alunos que fizeram pela primeira vez a disciplina em comparação com os que fazem novamente?

Para a quarta pergunta do domínio de informações gerais foram utilizados o gráfico de barras, agrupado de duas formas distintas, o gráfico de caixas, de violino e de coordenadas paralelas, além de utilizar o gráfico de linhas combinado com o gráfico de barras. Três questões foram elaboradas para esta pergunta, em que os resultados são apresentados na Tabela 5.5.

A primeira questão elaborada (Q1) abrange os cursos e a quantidade de alunos - "Os alunos que mais frequentam a disciplina são de qual curso?". Essa questão foi abordada nos dois gráficos de barras, no gráfico de coordenadas paralelas e na visualização do gráfico de barras associado ao gráfico de linhas. Isto posto, dois questionados não souberam responder no gráfico de coordenadas paralelas e dois erraram a resposta no gráfico de barras agrupado pelo curso. Todos acertaram a questão ao visualizar o gráfico de linhas com o de barras e responderam no intervalo de 5 a 9 segundos, apesar de algumas respostas erradas terem sido registradas.

A outra questão elaborada (Q2), "Os alunos de qual curso possuem a menor média das notas?", foi indagada no gráfico de caixas, no gráfico de violino, de coordenadas paralelas e no de linhas associado ao de barras. Das 8 pessoas que responderam ao questionário, 5 não souberam responder essa questão no gráfico de coordenadas paralelas e 4 no gráfico

de violino. Ao observar o gráfico de linhas e barras, a maioria dos questionados acertou a questão. De forma curiosa, o gráfico de violino foi aquele em que os usuários responderam de forma mais rápida, entre 5 a 9 segundos.

O terceiro questionamento (Q3), "Qual curso onde os alunos que fazem a disciplina mais de uma vez possuem a maior média?", foi aplicado nos gráficos de caixa, de violino, de coordenadas paralelas e de linhas associado ao de barras. Nenhum questionado acertou a questão ao observar o gráfico de coordenadas paralelas ou de violino. Todavia, a maioria acertou ao observar o gráfico de caixas, assim como foram mais rápidos, respondendo entre 5 a 9 segundos.

Tabela 5.5: Resultado dos algoritmos da Pergunta 4 do domínio de informações gerais.

		Tempo de resposta (segundos)				**Resposta**		
	Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Q1	**Linhas com Barras (4.7c)**		5		3	8		
	Barras (4.7d)		3	3	2	6	2	
	Barras (4.7e)		1	2	5	7	1	
	Coordenadas Paralelas (4.7f)	1		1	4	6		2
Q2	**Caixas (4.7a)**		1	2	3	3	3	2
	Violino (4.7b)		2	1	1	3	1	4
	Linhas com Barras (4.7c)			2	5	4	3	1
	Coordenadas Paralelas (4.7f)				3	3		5
Q3	**Caixas (4.7a)**		6	2		7	1	
	Violino (4.7b)		2	2			4	4
	Linhas com Barras (4.7c)		4	3		6	1	1
	Coordenadas Paralelas (4.7f)				4		4	4

C - resposta correta E - resposta errada NS - resposta 'não sei'

5.2.2 Domínio de Informações Acadêmicas

No domínio de informações acadêmicas foram elaboradas 5 perguntas especificadas na Seção 4.3.2 e selecionadas 21 visualizações candidatas para responder estas perguntas. Nesta seção, é detalhada a avaliação dos gráficos selecionados para este domínio.

Pergunta 1: Qual a evolução das notas dos alunos?

O gráfico de linhas, de área, de caixas e de barras foram escolhidos como candidatos para apresentar a informação relativa à primeira pergunta do domínio de informações acadêmicas. De forma a avaliar as visualizações, a questão elaborada relacionada à primeira pergunta foi: "Qual o semestre teve a maior média de notas?". Em vista disso, todos os questionados acertaram a resposta ao utilizar o gráfico de linhas e de barras, enquanto que a maioria errou ao utilizar o gráfico de caixas. O resultado relacionado a esta questão é apresentado na Tabela 5.6.

Tabela 5.6: Resultado dos algoritmos da Pergunta 1 do domínio de informações acadêmicas.

	Tempo de resposta (segundos)				**Resposta**		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Área (4.8a)	2	4			5	1	1
Linhas (4.8b)	4	3			7		
Barras (4.8c)	4	3			7		
Caixas (4.8d)		3	2	1	2	4	1

C - resposta correta E - resposta errada NS - resposta 'não sei'

Pergunta 2: Qual a relação entre a quantidade de créditos cursados no semestre pelos alunos com o número de aprovações?

Duas questões do questionário foram elaboradas para a Pergunta 2 do domínio de informações acadêmicas, em que os resultados são apresentados na Tabela 5.7. A primeira questão (Q1) foi: "Em qual semestre os alunos tiveram a maior média de notas?" Pode-se observar que apenas um questionado errou a resposta em cada gráfico, mas que o gráfico de barras associado ao gráfico de linhas obteve respostas mais rápidas.

Em relação à segunda questão (Q2) - "Em qual o semestre os alunos que frequentaram a disciplina tiveram o maior percentual de créditos aprovados?" - várias pessoas erraram a resposta ao utilizar o gráfico de área associado com o gráfico de linhas. Já o gráfico de barras associado ao gráfico de linhas foi a visualização com a maior quantidade de respostas corretas.

Pergunta 3: Quais são as métricas das notas dos alunos?

O gráfico de pizza, de barras, de radar, de densidade, de caixas, de violino e o histograma foram utilizados como candidatos para responder a terceira pergunta. Quatro questões

Tabela 5.7: Resultado dos algoritmos da Pergunta 2 do domínio de informações acadêmicas.

		Tempo de resposta (segundos)				Resposta		
	Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Q1	**Área com Linhas (4.9a)**	1	5		1	6	1	
	Barras com Linhas (4.9b)	2	5			6	1	
Q2	**Área com Linhas (4.9a)**		3	3	1	1	6	
	Barras com Linhas (4.9b)	4	3			4	3	

C - resposta correta E - resposta errada NS - resposta 'não sei'

foram elaboradas para a esta pergunta do domínio de informações acadêmicas, em que os resultados são apresentados na Tabela 5.8.

A primeira (Q1) consiste em: "Qual a nota mais tirada pelos alunos?". Dessa forma, aplicou-se apenas nos gráficos de pizza, de barras e de radar. Para responder a esta questão, a maior parte dos professores/gestores abordados demoraram mais para identificar a resposta no gráfico de pizza, gastando entre 5 a 9 segundos, enquanto que no gráfico de barras e de radar, a maioria demorou de 1 a 4 segundos para responder. Das 7 pessoas que responderam ao questionário de informações gerais, todos acertaram a resposta da pergunta em todas as visualizações.

A segunda questão elaborada (Q2) - "Qual a média das notas dos alunos?" - foi empregada ao gráfico de densidade, de caixas, e ao histograma. Mais questionados acertaram a resposta ao utilizar o gráfico de caixas, seguido pelo gráfico de violino, histograma e gráfico de densidade.

Além disso, a terceira questão (Q3) que consiste em "Qual o limite superior das notas dos alunos?" foi aplicada nos gráficos de densidade, de caixas, de violino e no histograma. Ao utilizar o histograma a maioria dos questionados demorou de 1 a 4 segundos, como também foi a visualização em que mais pessoas responderam de forma correta, em conjunto com o gráfico de caixas.

Assim como nas duas anteriores, a questão (Q4) "Existem pontos discrepantes?" foi aplicada aos gráficos de densidade, de caixas, de violino e ao histograma. Nesta questão, mais professores/gestores educacionais acertaram a resposta ao utilizar o histograma seguido pelo gráfico de caixas, violino e, por último, o gráfico de densidade, em que apenas 1 pessoa acertou a resposta.

Tabela 5.8: Resultado das questões elaboradas da Pergunta 3 do domínio de informações acadêmicas.

		Tempo de resposta (segundos)				Resposta		
	Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Q1	**Pizza (4.10e)**	2	4		1	7		
	Radar (4.10f)	5			2	7		
	Barras (4.10g)	5	2			7		
Q2	**Caixas (4.10a)**	4	3			7		
	Violino (4.10b)	3	3			6		1
	Densidade (4.10c)			4	1	4	1	2
	Histograma (4.10d)	1	2	1	2	5	1	1
Q3	**Caixas (4.10a)**	1	4	1	1	5	2	
	Violino (4.10b)	3		2		3	2	2
	Densidade (4.10c)	3		2		1	4	2
	Histograma (4.10d)	4	1	1		5	1	1
Q4	**Caixas (4.10a)**	4	2			5	1	1
	Violino (4.10b)	1	2	2		2	3	2
	Densidade (4.10c)		4	1		1	4	2
	Histograma (4.10d)	5		2		6	1	

C - resposta correta E - resposta errada NS - resposta 'não sei'

Pergunta 4: Qual a evolução da quantidade de alunos aprovados comparando com o total?

Para a quarta pergunta, o gráfico de linhas, de área e de barras foram escolhidos como candidatos. A questão elaborada foi: "Qual o semestre que teve o maior percentual de alunos aprovados?", na qual mais questionados acertaram ao utilizar o gráfico de barras, enquanto que as respostas mais rápidas ocorreram no gráfico de linhas. Os resultados são expostos na Tabela 5.9.

Tabela 5.9: Resultado dos algoritmos da Pergunta 4 do domínio de informações acadêmicas.

	Tempo de resposta (segundos)				Resposta		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Área (4.11a)	4		1	1	3	3	1
Barras (4.11b)	1	1	2	3	3	4	
Linhas (4.11c)	4	3			1	6	

C - resposta correta E - resposta errada NS - resposta 'não sei'

Pergunta 5: Qual a relação entre as notas e os créditos dos alunos que fizeram pela primeira vez a disciplina com as dos alunos que fazem novamente?

O gráfico de linhas, de linhas associado ao de área, de linhas associado com o de barras, de bolhas e de coordenadas paralelas foram escolhidos como candidatos para apresentar a informação relativa à quarta pergunta do domínio de informações acadêmicas. De forma a avaliar as visualizações, foram elaboradas duas questões: "Quais alunos tendem a tirar as maiores notas, os que fazem pela primeira vez, ou os que fazem mais de uma vez a disciplina?" (Q1) e "Existe diferença na quantidade de créditos cursados entre os alunos que fazem a primeira vez a disciplina ou aqueles que a fazem novamente?" (Q2). Os resultados são apresentados na Tabela 5.10.

Na primeira questão (Q1), o gráfico de linhas teve uma maior quantidade de acertos, porém, foi o gráfico onde os questionados demoraram mais para conseguir identificar a resposta. No gráfico de linhas associado ao gráfico de barras, 4 pessoas acertaram as perguntas e responderam no intervalo de 1 a 4 segundos.

Em relação à segunda questão elaborada (Q2), foi identificado uma alta quantidade de respostas "não sei", quando o questionado não consegue encontrar a informação, ou seja, são gráficos mais difíceis de serem interpretados. No gráfico de bolhas, das pessoas que responderam, todas acertaram e em um tempo menor do que o gráfico de coordenadas paralelas.

Tabela 5.10: Resultado dos algoritmos da Pergunta 5 do domínio de informações acadêmicas.

		Tempo de resposta (segundos)				**Resposta**		
	Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Q1	**Linhas (4.12a)**	1	1	3	2	6	1	
	Área com Linhas (4.12b)	3	3	1		4	3	
	Barras com Linhas (4.12c)	1	1	2	3	3	4	
Q2	**Bolhas (4.12d)**		3	2		5		2
	Coordenadas Pralelas (4.12e)			3	2	4	1	2

C - resposta correta E - resposta errada NS - resposta 'não sei'

5.2.3 Domínio de Informações de Percepção

No domínio de informações de percepção foram elaboradas 6 perguntas especificadas na Seção 4.3.3 e selecionadas 23 visualizações candidatas para responder estas perguntas.

Nesta seção é detalhada a avaliação dos gráficos selecionados para este domínio. Desta forma, os experimentos referentes a cada pergunta são apresentados abaixo.

Pergunta 1: Os alunos já entram na disciplina com contato prévio com programação?

Na primeira pergunta do domínio de informações de percepção foram selecionados dois gráficos candidatos, o gráfico de pizza e o gráfico de barras. A questão elaborada para os dois gráficos foi: "Há mais alunos que tiveram contato prévio com programação antes da disciplina, ou alunos que nunca tiveram contato?". Em vista disso, dos 8 professores/gestores abordados para responder ao questionário do domínio das informações de percepção, todos acertaram a resposta ao utilizar ambos os gráficos. No gráfico de pizza, a maioria dos usuários demorou entre 5 a 9 segundos para identificar a informação e responder, enquanto que no gráfico de barras os usuários demoraram de 1 a 4 segundos. O resultado é apresentado na Tabela 5.11.

Tabela 5.11: Resultado dos algoritmos da Pergunta 1 do domínio de informações de percepção.

	Tempo de resposta (segundos)				**Resposta**		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Pizza (4.13a)	2	6			8		
Barras (4.13b)	7	1			8		

C - resposta correta E - resposta errada NS - resposta 'não sei'

Pergunta 2: Com qual linguagem os alunos já tiveram contato?

De forma a responder a segunda pergunta, o gráfico de pizza e o gráfico de barras foram selecionados como candidatos. A questão elaborada na segunda pergunta foi: "Com qual a linguagem de programação os alunos mais tiveram contato?". Novamente todos os questionados acertaram a resposta nas duas visualizações. Ao analisar o gráfico de pizza os usuários levaram de 1 a 4 segundos para identificar a informação e responder, enquanto que no gráfico de barras demoraram mais, conforme apresentado na Tabela 5.12.

Pergunta 3: Os alunos costumam escutar que Engenharia/Computação é para pessoas super inteligentes?

A terceira pergunta contou com o gráfico de pizza, de barras e de barras empilhadas como candidatos. A questão elaborada foi: "Os alunos costumam escutar que engenharia/computação é para pessoas inteligentes?". Apenas 1 questionado errou a resposta no

Tabela 5.12: Resultado dos algoritmos da Pergunta 2 do domínio de informações de percepção.

	Tempo de resposta (segundos)				Resposta		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Pizza (4.14a)	8				8		
Barras (4.14b)	2	3	3		8		

C - resposta correta E - resposta errada NS - resposta 'não sei'

gráfico de pizza, mas a maioria demorou entre 5 a 9 segundos para localizar a informação e responder tanto no gráfico de pizza quanto no gráfico de barras. A Tabela 5.13 expõe os resultados.

Tabela 5.13: Resultado dos algoritmos da Pergunta 3 do domínio de informações de percepção.

	Tempo de resposta (segundos)				Resposta		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Pizza (4.15a)	1	5	1	1	7	1	
Barras Empilhadas (4.15b)	2	3	3		8		
Barras (4.15c)	3	5			8		

C - resposta correta E - resposta errada NS - resposta 'não sei'

Pergunta 4: Os alunos costumam escutar que Engenharia/Computação é para homens?

A quarta pergunta contou com o gráfico de pizza, de barras e de barras empilhadas como candidatos. A seguinte questão foi elaborada de forma a avaliar os algoritmos candidatos: "Os alunos costumam escutar que engenharia/computação é para homens?". Apenas 1 questionado errou a resposta em cada um dos três gráficos, mas o tempo para localizar a informação foi menor no gráfico de pizza, levando de 1 a 4 segundos. Diferentemente do gráfico de barras e de barras empilhadas, em que a maioria dos entrevistados demoraram de 5 a 9 segundos. Os resultados são evidenciados na Tabela 5.14.

Pergunta 5: Qual a relação entre área de trabalho dos pais com o nível educacional dos pais?

Para a quinta pergunta foram utilizados, como candidatos, o diagrama de sankey, de coordenadas paralelas e o gráfico de barras agrupado de duas formas distintas. Isto posto, para avaliar as informações dos gráficos, foi elaborada a questão: "Qual a área

Tabela 5.14: Resultado dos algoritmos da Pergunta 4 do domínio de informações de percepção.

	Tempo de resposta (segundos)				Resposta		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Pizza (4.16a)	5	2		1	7	1	
Barras Empilhadas (4.16b)		5		3	7	1	
Barras (4.16c)	2	5		1	7	1	

C - resposta correta E - resposta errada NS - resposta 'não sei'

de atuação dos pais que menos têm ensino superior completo?", e aplicada nos quatro gráficos. Novamente, quatro questionados não souberam responder ao visualizar o gráfico de coordenadas paralelas. Entretanto, este foi o gráfico com o maior número de acertos, em conjunto com o gráfico de barras agrupado pelo nível educacional dos pais, seguido pelo diagrama de sankey e pelo gráfico de barras agrupado pela área de atuação dos pais. Os resultados da questão elaborada são apresentados na Tabela 5.15.

Tabela 5.15: Resultado dos algoritmos da Pergunta 5 do domínio de informações de percepção.

	Tempo de resposta (segundos)				Resposta		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Barras (4.17a)	1	1	4	2	1	7	
Barras (4.17b)		4	2	2	3	5	
Coordenadas Paralelas (4.17c)	1		1	2	3	1	4
Sankey (4.17d)		2	1	3	2	4	2

C - resposta correta E - resposta errada NS - resposta 'não sei'

Pergunta 6: Qual a relação entre a quantidade de horas estudadas com a percepção de que os alunos precisam ou não de mais horas de estudos do que os colegas?

O gráfico de coordenadas paralelas, de sankey e dois gráficos de barra foram utilizados como candidatos para responder a última pergunta. A questão elaborada para avaliar os algoritmos candidatos foi: "Os alunos que nunca estudaram fora do horário de sala de aula acreditam que precisam de mais horas de estudo que seus amigos?". Para tanto, 7 usuários acertaram as respostas nos dois gráficos de barras e no diagrama de sankey. Já o gráfico de coordenadas paralelas contou com 6 acertos. A Tabela 5.16 apresenta os resultados.

Tabela 5.16: Resultado dos algoritmos da Pergunta 6 do domínio de informações de percepção.

	Tempo de resposta (segundos)				Resposta		
Gráfico	**1 - 4**	**5 - 9**	**10 - 15**	**>15**	**C**	**E**	**NS**
Barras (4.18a)	3	2		3	7	1	
Barras (4.18b)	2	3	2	1	7	1	
Coordenadas Paralelas (4.18c)	5	1	1		6	1	1
Sankey (4.18d)	4	2		1	7		1

C - resposta correta E - resposta errada NS - resposta 'não sei'

5.3 Resultados

A partir do estudo dos algoritmos de visualização, observou-se que mais de um algoritmo pode representar a mesma informação. Embora algoritmos mais complexos apresentam mais informações em uma forma compacta, eles não são tão conhecidos e são mais difíceis de interpretar. Nesse sentido, foi aplicado o questionário de forma a validar os algoritmos e selecionar aqueles mais apropriados. Isto posto, esta seção discute e compara o resultado do estudo dos algoritmos de visualização com o resultado do questionário.

5.3.1 Informações Gerais

Em relação à Pergunta 1, o gráfico de pizza permite uma visualização mais rápida em relação à porcentagem, enquanto que a visualização do gráfico de barras é focada em comparações diretas. De acordo com o estudo realizado dos algoritmos, visualizar a porcentagem pode levar à perda de informação, entretanto, por se tratar de uma pergunta em que o importante é a comparação entre as categorias, é possível extrair conhecimento a partir da análise de ambos os gráficos. Segundo o resultado da aplicação do questionário, o tempo de resposta dos gráficos candidatos foi o mesmo, todavia mais pessoas acertaram a questão elaborada ao utilizar o gráfico de pizza.

Nas Perguntas 2 e 3, foram utilizados os gráficos de barras, de densidade, o histograma e o treemap. Cada um dos gráficos fornecem informações distintas, o gráfico de barras realiza uma comparação direta dos valores, o gráfico de densidade foca na frequência, o histograma na densidade e o treemap permite a visualização da proporção em relação à área no gráfico. No questionário da Pergunta 2, o gráfico de barras foi a visualização em que mais participantes do questionário acertaram a resposta em menor tempo. Já na Pergunta 3, todos os usuários responderam de forma correta a questão, porém, as informações eram obtidas em menos tempo ao utilizarem o treemap.

Por fim, dentre os algoritmos candidatos da Pergunta 4, o gráfico de violino é a visualização que mais gera conhecimento, entretanto, este gráfico foi o mais confuso e difícil de interpretar, segundo o resultado do questionário. Como segunda opção, o gráfico de barras associado ao gráfico de linhas possibilita descobertas de conhecimento semelhantes ao gráfico de violino, sendo bem mais amigável aos professores e gestores, que acertaram mais respostas ao utilizar esta visualização.

5.3.2 Informações Acadêmicas

Em relação aos gráficos selecionados para a Pergunta 1 (gráfico de área, de linhas, de barras e de caixas), de acordo com o estudo, o gráfico de caixas é o que apresenta mais informações, porém o excesso deixa a visualização comprometida. No resultado do questionário, o gráfico de linhas e o de barras levaram os professores e gestores a responderem as questões de maneira mais rápida e assertiva, entretanto, dentre os dois, o gráfico de linhas é o que melhor representa a evolução no tempo.

Na Pergunta 2, o gráfico de linhas foi associado ao gráfico de barras, que enfatiza o total dos dados, e ao gráfico de área, que enfatiza a magnitude da alteração ao longo do tempo. Porém, segundo o resultado do questionário, mais participantes acertaram em menos tempo as questões ao utilizar o gráfico de linhas associado ao gráfico de barras.

No tocante à Pergunta 3, de forma a identificar a quantidade de notas, foram utilizados os gráficos de pizza, de barras e de radar, em que todos possibilitaram que os questionados respondessem de forma correta. Já em relação às métricas de forma geral, o gráfico de caixas e o histograma foram os gráficos com maior quantidade de acertos nas três questões elaboradas. Porém, o gráfico de violino representa as informações do gráfico de caixas e do histograma juntos, mas não é um gráfico conhecido pelos usuários e, por isso, foram verificadas várias respostas “não sei” nessa visualização.

Foram selecionados os gráficos de linhas, de área e de barras para representar a Pergunta 4 em razão de apresentarem informações categóricas ao longo do tempo. O gráfico de barras é utilizado na comparação direta dos valores e o gráfico de área para enfatizar a magnitude da evolução temporal. Em vista disso, o gráfico de área foi a visualização, segundo o resultado do questionário, em que mais pessoas identificaram a resposta correta no menor tempo.

A Pergunta 5 utiliza os atributos de tempo, vez em que cursou a disciplina, quantidade de créditos totais e aprovados, além das notas dos alunos. Desta forma, os gráficos de bolha e de coordenadas paralelas são utilizados para evidenciar a relação entre valores numéricos de diferentes categorias. Além disso, o gráfico de linhas, de linhas associado ao de barras e ao de área podem ser adaptados para representar os atributos. No gráfico de linhas, quanto mais categorias, mais confuso ele pode se tornar, assim como o de linha

associado ao de área. Apesar disso, essas foram as visualizações com maior quantidade de respostas corretas, mas para permitir ao usuário distinguir as informações com mais facilidade, foi selecionado o gráfico de barras associado ao de linhas.

5.3.3 Informações de Percepção

As Perguntas 1 e 2 tinham como candidatos os gráficos de barras e de pizza, sendo que a Pergunta 2 possui mais categorias de respostas. O gráfico de pizza não é indicado para comparações de muitos grupos, porém, mais pessoas responderam de forma mais rápida e assertiva ao utilizar este gráfico na Pergunta 2. Por outro lado, o gráfico de barras apresentou melhores resultados na Pergunta 1.

Em relação às Perguntas 3 e 4 foram selecionados como candidatos os gráficos de pizza, de barras e de barras empilhadas. Analisando as respostas do questionário, foi identificado que os professores e gestores educacionais utilizaram o gráfico de barras para responder à Pergunta 3, enquanto que na Pergunta 4, o gráfico de pizza apresentou melhores resultados. Pelo fato do gráfico de barras empilhadas enfatizar o total das categorias, os questionados apresentaram maior taxa de erro nas respostas em relação a ambas as perguntas.

Por fim, as Perguntas 5 e 6 utilizaram os gráficos de barras, de coordenadas paralelas e o diagrama de sankey. Em relação à Pergunta 5, tanto o gráfico de coordenadas paralelas quanto o diagrama de sankey tiveram várias respostas "não sei", pois ambos os gráficos eram pouco conhecidos pelos participantes. Todavia, ao responder a Pergunta 6 utilizando o diagrama de sankey, foram apresentadas respostas mais rápidas e corretas. Assim, conclui-se que o diagrama de sankey não é uma visualização popular entre os usuários, o que provoca um nível de confusão durante o primeiro contato, mas fornece informações precisas e rápidas quando já conhecido.

5.4 Considerações Finais

Neste capítulo foram apresentados os resultados dos três questionários aplicados a 23 professores e gestores educacionais. As questões iniciais foram utilizadas para avaliar o conhecimento inicial dos participantes em relação às visualizações, enquanto que as questões elaboradas foram utilizadas para avaliar como os gráficos eram analisados pelos usuários. Para cada questão dos domínios de informações gerais, acadêmicas e de percepção, foram selecionados os melhores gráficos para representar a informação a partir do resultado do questionário e do estudo das visualizações. Com o objetivo de extrair conhecimento, no próximo capítulo as visualizações selecionadas acima foram utilizadas em dados de gênero, cotas e ao comparar as disciplinas de ICC e APC.

Capítulo 6

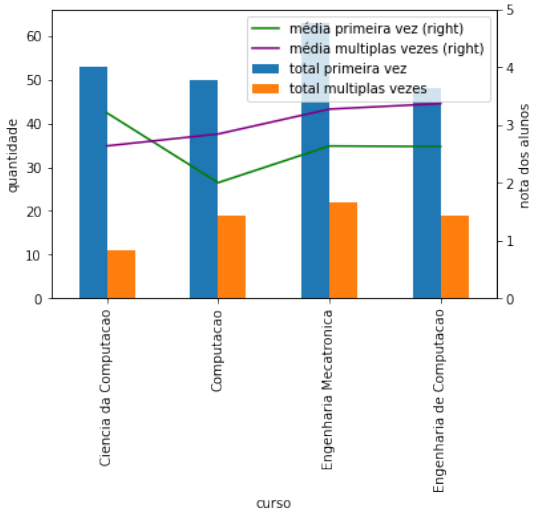
Aplicação do Resultado

No capítulo anterior foram selecionados os algoritmos que melhor representam cada uma das perguntas dos domínios de informações gerais, acadêmicas e de percepção e, com base nos resultados, é possível extrair conhecimento útil por meio das visualizações selecionadas. Neste capítulo são abordados três assuntos distintos, um em cada domínio. A questão de gênero nas disciplinas iniciais de programação foi analisada no domínio de informações gerais, destacando o curso com maior e menor quantidade de alunas. No domínio de informações acadêmicas foi analisado o tema das cotas com enfoque no período antes da implementação da Lei Federal nº 12.711/2012 (conhecida como a Lei das Cotas), o período de transição e após a lei ser efetivada. Por fim, no domínio de informações de percepção foram considerados os dados dos alunos de cada disciplina em que o questionário foi aplicado.

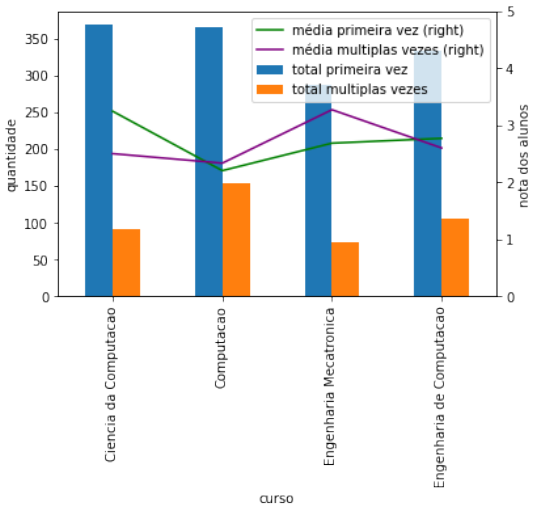
6.1 Domínio de Informações Gerais

Com o intuito de descobrir conhecimento no domínio de informações gerais, foram considerados os dados referentes ao gênero, analisando, como mostrado nas Figuras 6.1, as médias das notas e a quantidade de alunos que fizeram as disciplinas APC e ICC, comparando o desempenho dos estudantes que fizeram pela primeira vez e os que repetiram. Observa-se que a média das notas dos alunos, tanto do sexo masculino, apresentado na Figura 6.1d, quanto do sexo feminino, ilustrado na Figura 6.1c, é maior quando a disciplina de ICC é cursada pela primeira vez. Diferentemente do que acontece na disciplina APC, segundo a Figura 6.1a, em que a nota das mulheres é maior quando fazem a disciplina de novo.

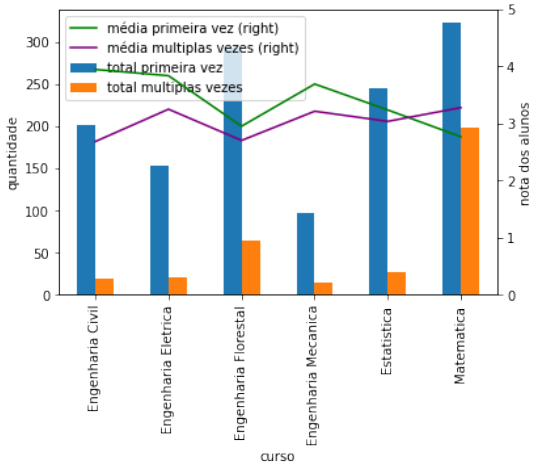
Para as próximas análises foram destacados dois cursos distintos - "Computação", que possui apenas 11,8% de mulheres, e "Engenharia Florestal", que possui 54% de alunas. Isto posto, observa-se nas Figuras 6.2a e 6.2c que as mulheres, de ambos os cursos, ao



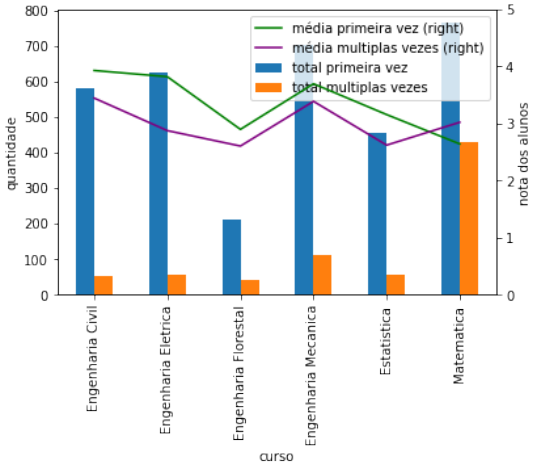
(a) Dados dos alunos do sexo feminino da disciplina APC.



(b) Dados dos alunos do sexo masculino da disciplina APC.



(c) Dados dos alunos do sexo feminino da disciplina ICC.

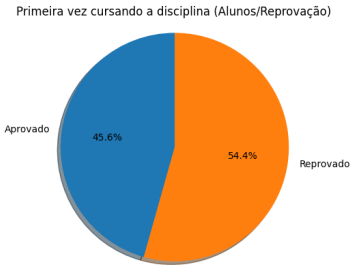


(d) Dados dos alunos do sexo masculino da disciplina ICC.

Figura 6.1: Comparação entre a quantidade de alunos por curso e as respectivas notas

fazerem a disciplina pela primeira vez, reprovam mais do que os homens, cujos dados são apresentados nas Figuras 6.2b e 6.2d. Porém, a quantidade de reprovações é maior para Computação do que Engenharia Florestal.

(a) Mulheres do curso de Ciência da Computação.



(b) Homens do curso de Ciência da Computação.

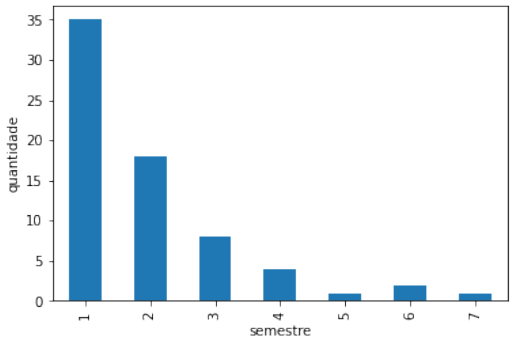
(c) Mulheres do curso de Engenharia Florestal.

(d) Homens do curso de Engenharia Florestal.

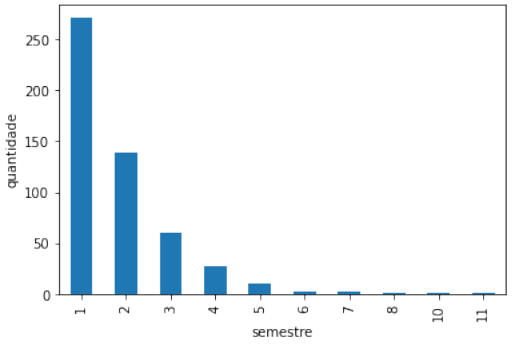
Figura 6.2: Quantidade de aprovações e reprovações.

A Figura 6.3 apresenta o semestre em que os alunos cursam a disciplina, tal qual na Figura 6.3a são apresentados os dados das alunas de Ciência da Computação, enquanto que na Figura 6.3b são as informações referentes aos homens deste curso. Em relação à Engenharia Florestal as Figuras 6.3c e 6.3d ilustram os dados dos alunos do sexo feminino e masculino, respectivamente. A partir das visualizações é possível verificar que, pelas disciplinas de APC e ICC serem introdutórias em programação, estas costumam ser cursadas no primeiro semestre dos alunos. Ademais, ao comparar as visualizações das Figuras 6.4a e 6.4b, em que é representado a quantidade de vezes em que os alunos de Computação do sexo feminino e masculino cursam a disciplina, com as visualizações das Figuras 6.4c e 6.4d, que correspondem aos dados dos estudantes de Engenharia Florestal, pode-se inferir que a quantidade de vezes que os alunos fazem a disciplina é igual ao comparar homens e mulheres.

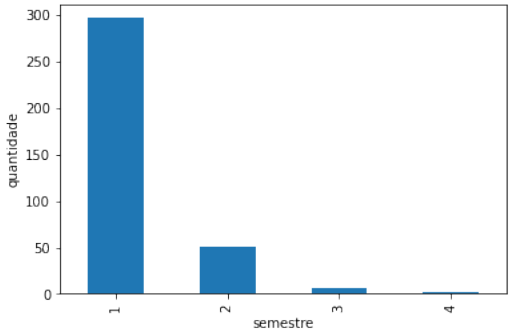
A partir das visualizações geradas no domínio de informações gerais ao comparar os dados relacionados à gênero dos estudantes do curso de Computação e Engenharia



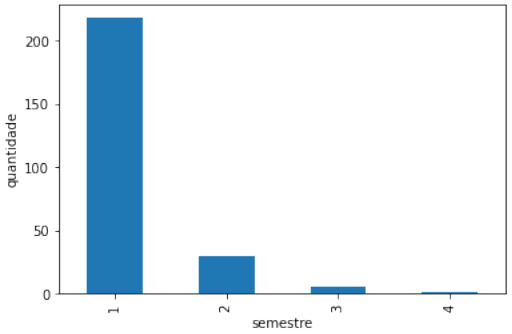
(a) Dados dos alunos do sexo feminino do curso de Ciência da Computação.



(b) Dados dos alunos do sexo masculino do curso de Ciência da Computação.



(c) Dados dos alunos do sexo feminino do curso de Engenharia Florestal.



(d) Dados dos alunos do sexo masculino do curso de Engenharia Florestal.

Figura 6.3: Semestre em que os alunos cursam a disciplina.

(a) Dados dos alunos do sexo feminino do curso de Ciência da Computação.

(b) Dados dos alunos do sexo masculino do curso de Ciência da Computação.

(c) Dados dos alunos do sexo feminino do curso de Engenharia Florestal.

(d) Dados dos alunos do sexo masculino do curso de Engenharia Florestal.

Figura 6.4: Quantidade de vezes que os alunos cursam a disciplina.

Florestal, conclui-se que os alunos do sexo feminino, ao serem reprovados tem melhor desempenho ao repetir a disciplina.

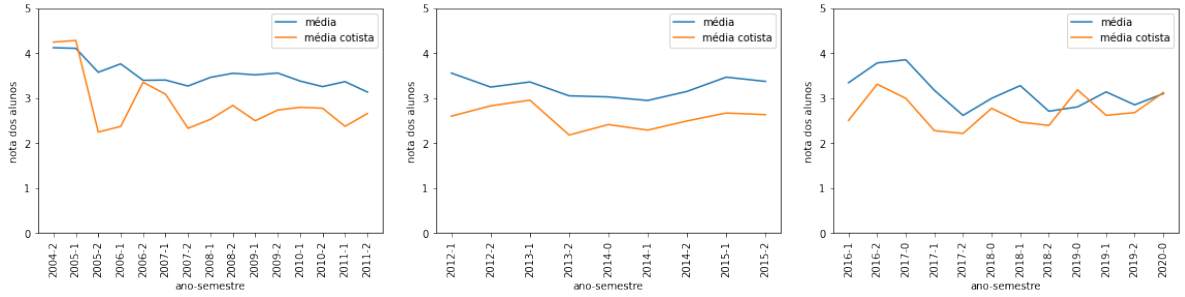
6.2 Domínio de Informações Acadêmicas

Em agosto de 2012, foi promulgada no Brasil a Lei Federal nº 12.711/2012, conhecida como a Lei das Cotas, onde foi definida a obrigatoriedade de reserva de vagas em todas as Universidades e Institutos Federais de Educação, Ciência e Tecnologia do país para estudantes que cursaram todo o ensino médio em escolas públicas. Desta forma, 50% das vagas são reservadas para estudantes de escolas públicas, em que, dessas vagas, 25% são para estudantes com renda familiar bruta per capita igual ou inferior a um salário e meio. Além disso, dentro da divisão das cotas de escolas públicas, uma parte das vagas é reservada aos alunos que se declaram pardos, pretos ou indígenas. Esta porcentagem é definida com base na soma total da população que integra esses grupos em cada unidade da Federação, conforme o último censo do IBGE [53]. A UnB também implementou, em 2004, uma cota para estudantes que se consideram negros, de ambos os tipos de escolaridade, em que mais 5% das vagas são reservadas para estes alunos.

Nesse sentido, para o domínio de informações acadêmicas foram analisados os dados das disciplinas de APC, CB e ICC em três períodos de tempo relacionados ao histórico das cotas: de 2004 a 2011, período antes da Lei das Cotas ser decretada; de 2012 até 2015, período de transição para implantação da lei; a partir de 2016, período em que finalizou o tempo de adaptação da lei.

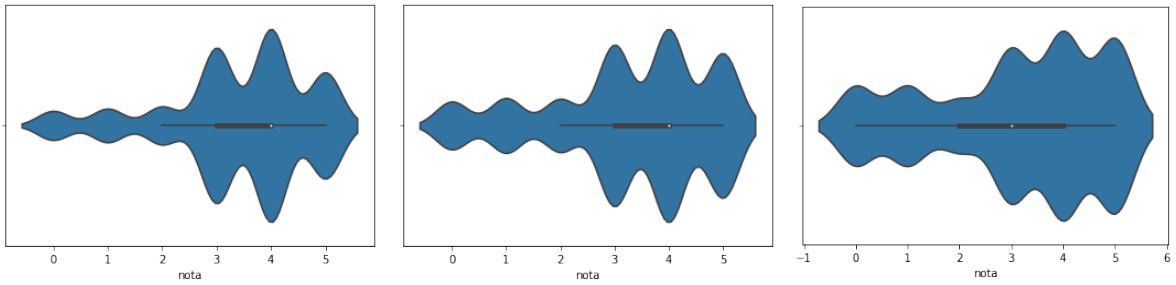
Em relação às notas dos alunos, a Figura 6.5 apresenta a evolução das médias nos períodos definidos, comparando os alunos que não ingressaram por meio de cotas com os dados dos alunos cotistas. Já a Figura 6.6 complementa as informações de notas apresentando as métricas dos alunos não-cotistas, bem como a Figura 6.7 fornece informações dos estudantes que entraram pelo sistema de cotas. Ao visualizar os gráficos é possível concluir que as notas dos alunos que não entraram pelo sistema de cotas são maiores do que a segunda categoria, embora as notas do grupo não-cotista tenham decrescido no último período.

A evolução da quantidade de alunos aprovados comparando com o total, em relação a cada período do histórico das cotas são apresentados na Figura 6.8, em que apresenta os dados dos não-cotistas, e na Figura 6.9, em que são representados os cotistas. Observa-se que os estudantes que não entraram por cotas eram mais numerosos do que os estudantes cotistas, entretanto após a implantação da Lei em 2016, a quantidade ficou equivalente. Além disso, as visualizações também mostram que, no último período, os alunos cotistas tendem a reprovar mais do que os alunos não-cotistas. Somado a isso, ao comparar a



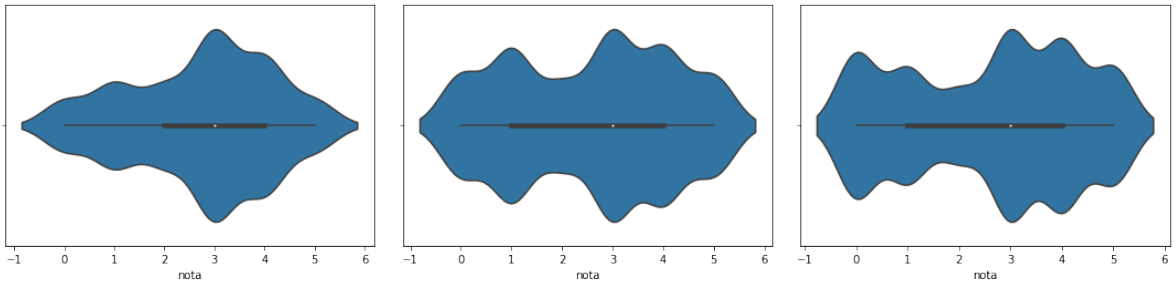
(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

Figura 6.5: Evolução das notas dos alunos.



(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

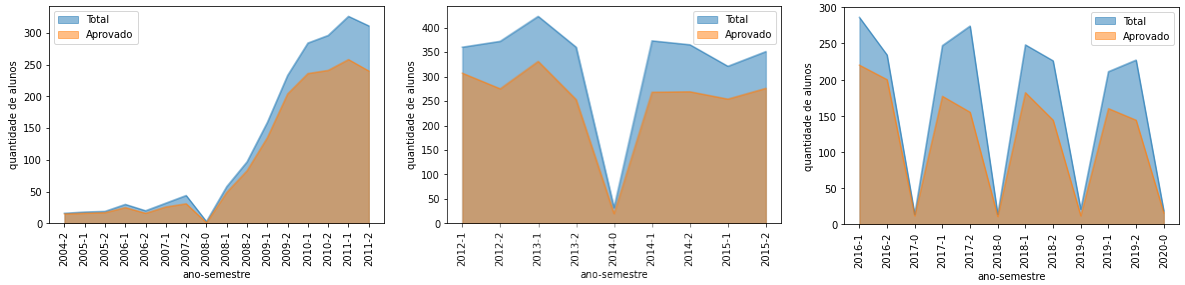
Figura 6.6: Métricas das notas dos alunos que não ingressaram na universidade por meio do sistema de cotas.



(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

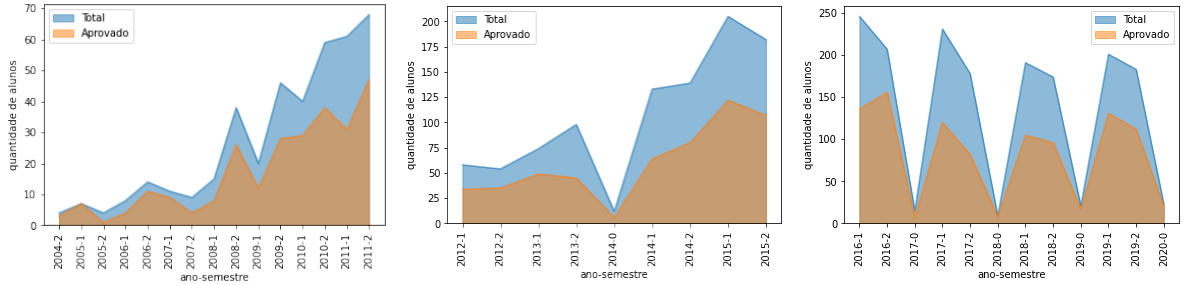
Figura 6.7: Métricas das notas dos alunos que ingressaram na universidade por meio do sistema de cotas.

quantidade de créditos aprovados pelos alunos cotistas da Figura 6.11 e não cotistas da Figura 6.10, verifica-se que ambas as categorias pegam a mesma quantidade de créditos, porém a quantidade de aprovações é menor para os alunos cotistas. Ou seja, ao fazerem as disciplinas introdutórias de programação, é comum os estudantes cotistas reprovarem em outras disciplinas cursadas no mesmo período.



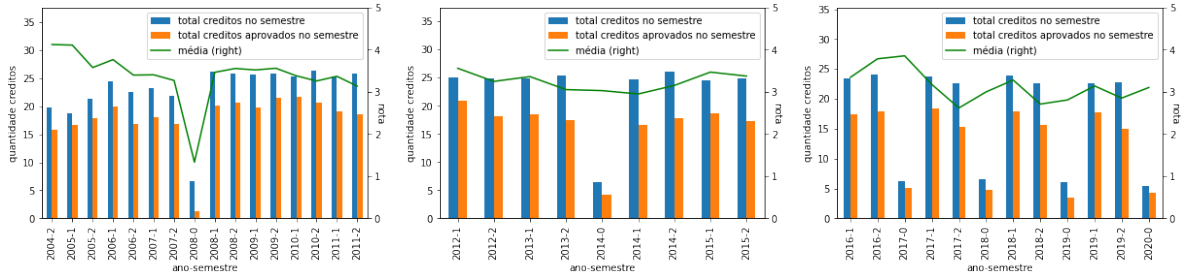
(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

Figura 6.8: Evolução da quantidade de aprovação dos alunos que não ingressaram na universidade por meio do sistema de cotas.



(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

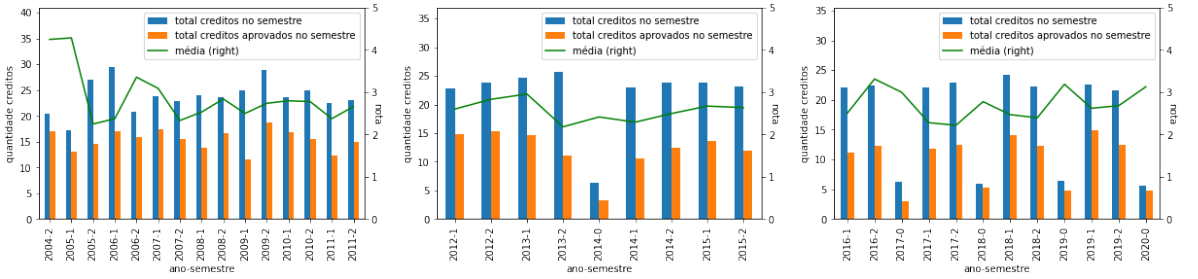
Figura 6.9: Evolução da quantidade de aprovação dos alunos que ingressaram na universidade por meio do sistema de cotas.



(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

Figura 6.10: Evolução da quantidade de créditos dos alunos que não ingressaram na universidade por meio do sistema de cotas.

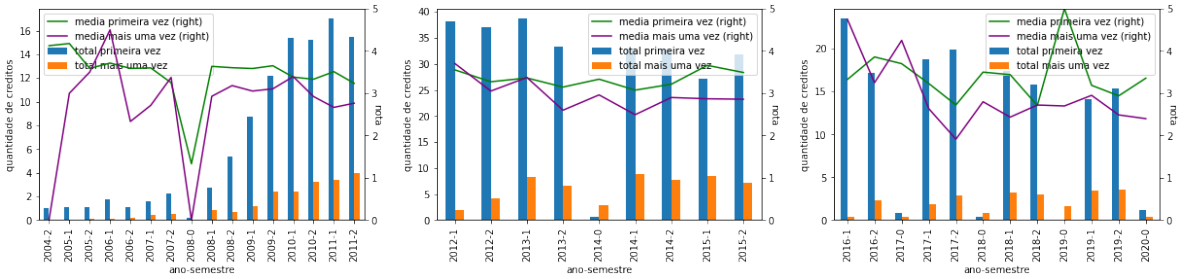
Os alunos que fazem novamente a disciplina tendem a pegar menos créditos no semestre do que os alunos que estão fazendo a disciplina pela primeira vez. Este é um comportamento observado nas visualizações de ambas as categorias de alunos, cotistas (Figura 6.13) e não-cotistas (Figura 6.12), e que se mantém desde 2004. Porém, nos semestres de verão de 2014, 2018 e 2019 (representado pelos valores 2014-0, 2018-0 e 2019-0),



(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

Figura 6.11: Evolução da quantidade de créditos dos alunos que ingressaram na universidade por meio do sistema de cotas.

ambas as categorias de alunos que estavam cursando novamente a disciplina pegaram mais créditos do que os alunos que estavam matriculados pela primeira vez. Outro resultado observado é que a nota dos alunos não cotistas que concluem a disciplina pela primeira vez é superior à dos que estão cursando novamente. Porém, na categoria dos cotistas as notas tendem a se manter semelhantes. Pode-se concluir com este estudo que os alunos que entraram por cotas apresentam mais dificuldade na disciplina, o que demandaria uma ação por parte dos professores para auxiliá-los.

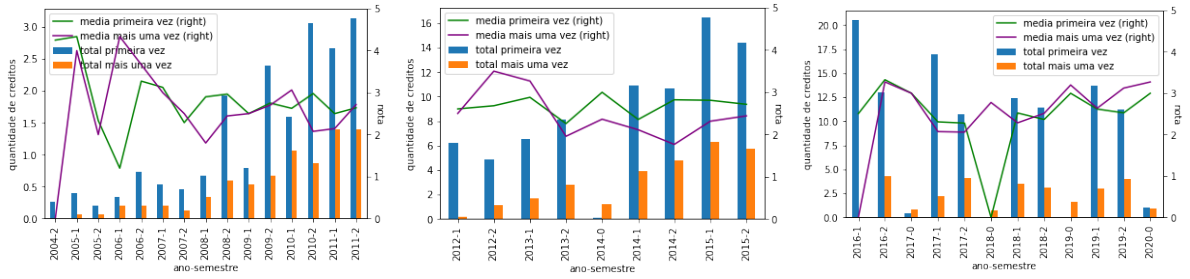


(a) Período anterior à Lei das Cotas. (b) Período de adaptação à Lei das Cotas. (c) Período após a implantação da Lei das Cotas.

Figura 6.12: Comparação entre os alunos que não ingressaram na universidade por meio do sistema de cotas e que fizeram a disciplina pela primeira vez com os que cursaram novamente.

6.3 Domínio de Informações de Percepção

Com o objetivo de conhecer a percepção dos alunos que cursam as disciplinas de programação, foram utilizados os dados do questionário aplicado às turmas de APC e ICC. A disciplina APC é ministrada aos alunos do Departamento de Computação, ao passo que ICC é para estudantes de outros departamentos.



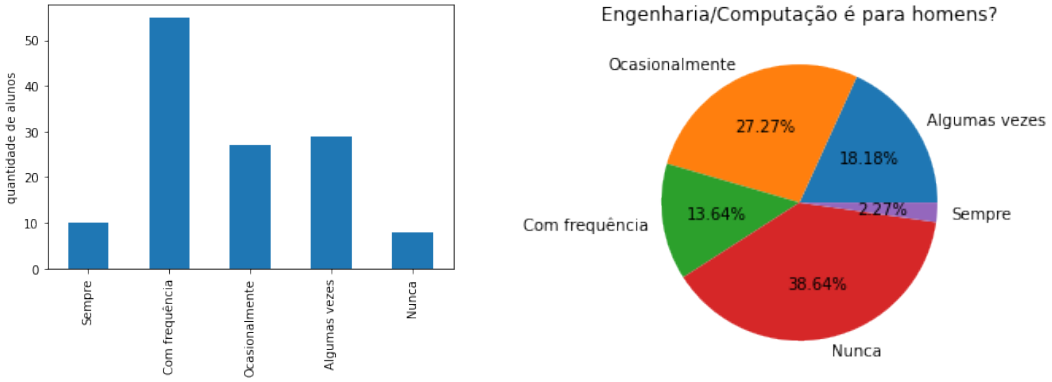
(a) Período anterior à Lei das Cotas.

(b) Período de adaptação à Lei das Cotas.

(c) Período após a implantação da Lei das Cotas.

Figura 6.13: Comparação entre os alunos que ingressaram na universidade por meio do sistema de cotas e que fizeram a disciplina pela primeira vez com os que cursaram novamente.

No questionário aplicado aos alunos de APC são abordados temas sobre o sentimento de pertencimento ao curso. Desta forma foram abordadas as duas perguntas: "Com que frequência você ouve que Engenharia/Computação é para pessoas super inteligentes", cujo resultado é apresentado na Figura 6.14a e "Com que frequência você ouve que Engenharia/Computação é para homens", que está representado na visualização da Figura 6.14b. A partir do resultado do questionário é possível perceber que os alunos costumam escutar com mais frequência que os cursos de computação e engenharia são para pessoas inteligentes do que escutam que são para homens. Os cursos de engenharia e de computação são conhecidos por terem muita matemática aplicada, o que fomenta o pensamento de que pessoas inteligentes são boas em matemática e que pode levar a certa frustração ao decorrerem as primeiras dificuldades.



(a) Frequência com que os alunos escutam que Engenharia/Computação é para pessoas super inteligentes".

(b) Frequência com que os alunos escutam que Engenharia/Computação é para homens".

Figura 6.14: Questões sobre pertencimento do questionário aplicado aos alunos de APC.

Além disso, para o questionário nas turmas de APC, também foi verificado a área

de atuação dos pais e o nível educacional. Na Figura 6.15 possível visualizar que mais estudantes, tanto do sexo feminino quanto do sexo masculino, em que a família trabalha com educação, optaram por fazer um curso do Departamento de Ciência da Computação, já que APC só é ministrado para alunos desses cursos. Além disso, a maioria dos familiares que trabalham na área de educação e tecnologia têm nível superior, ao passo que a maioria dos pais dos alunos cotistas não tem ensino superior.

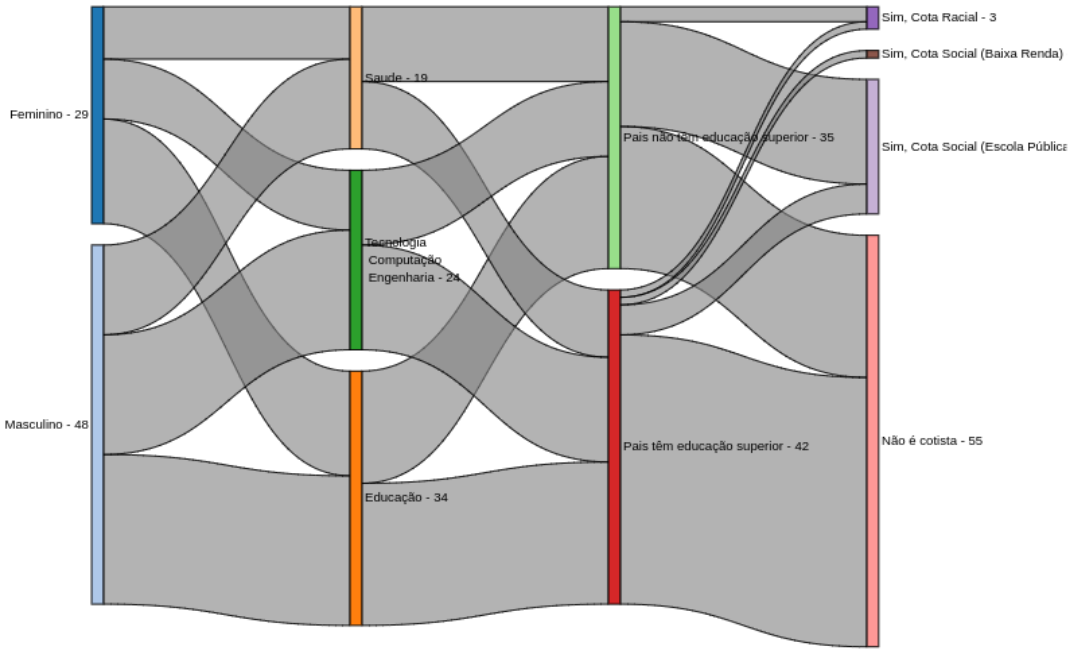
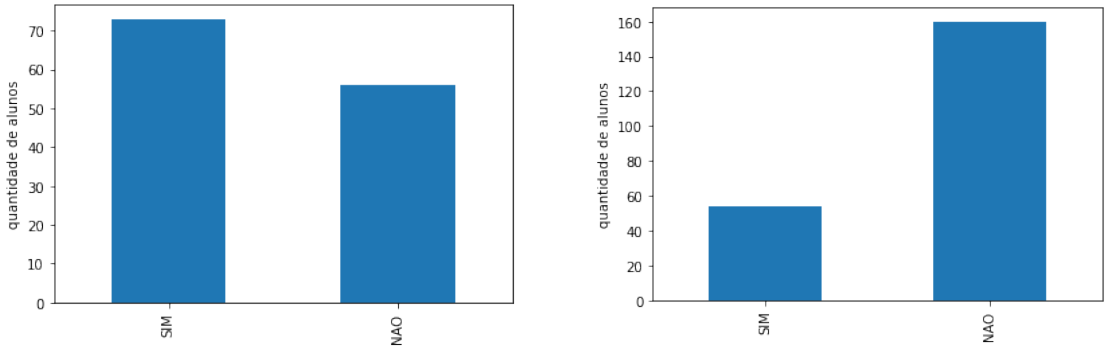


Figura 6.15: Relação entre a área de trabalho dos pais com o nível de ensino dos mesmos.

Além do tópico de pertencimento foram comparadas as respostas em relação à experiência em programação dos alunos de APC, apresentado na Figura 6.16a, e de ICC, ilustrado na Figura 6.16b. Além disso, as Figuras 6.17a, 6.17b explicitam as linguagens que são mais conhecidas pelos alunos. Esses dados apontam que, mesmo sendo uma disciplina introdutória, a maioria dos alunos de APC já entra com conhecimento prévio em programação, principalmente nas linguagens Python e C/C++. Ao contrário dos alunos que fizeram a disciplina de ICC, em que a maioria não teve contato com programação.

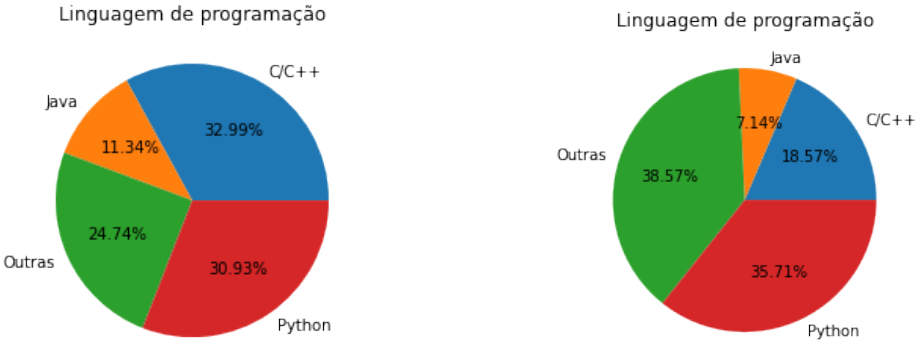
Ademais, as Figuras 6.18a e 6.18b comparam a quantidade de horas estudadas pelos alunos com a percepção de que precisam ou não de mais horas de estudos que os colegas. Observa-se que na disciplina APC são os alunos que mais estudam fora do horário de sala de aula (mais do que 6 horas por dia) que consideram precisar de mais horas de estudos do que os colegas. Já em ICC, a maioria dos estudantes acreditam que precisam se dedicar mais do que os colegas. Somado essas informações com as análises anteriores, em que os estudantes de ICC não entram na disciplina sabendo programar, conclui-se que os alunos



(a) Dados dos alunos que cursaram a disciplina APC.

(b) Dados dos alunos que cursaram a disciplina ICC.

Figura 6.16: Informação sobre se os alunos tiveram contato prévio com programação antes da disciplina.

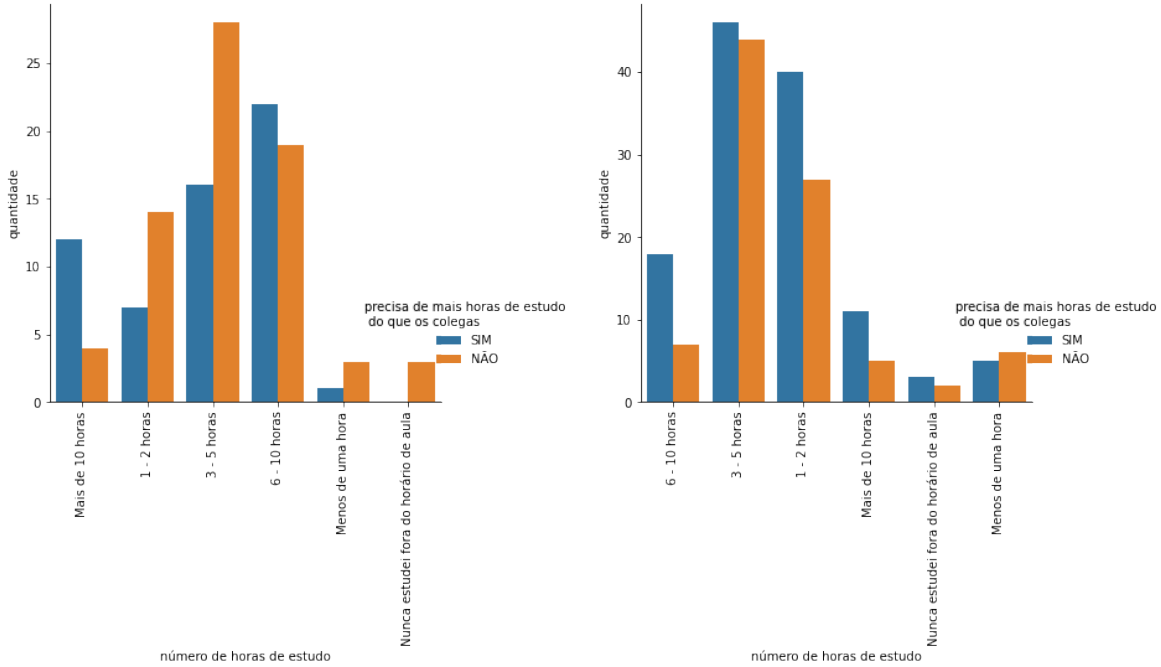


(a) Dados dos alunos que cursaram a disciplina APC.

(b) Dados dos alunos que cursaram a disciplina ICC.

Figura 6.17: Quais as linguagens de programação mais conhecidas pelos alunos.

que não tiveram contato prévio com programação antes da disciplina têm a sensação de que os colegas têm mais facilidade, ao passo que na disciplina APC, são os alunos que mais estudam que têm essa percepção.



(a) Dados dos alunos que cursaram a disciplina APC.

(b) Dados dos alunos que cursaram a disciplina ICC.

Figura 6.18: Relação entre quantidade de horas de estudo fora da disciplina com a percepção de que precisam de mais horas de estudos que os colegas.

Capítulo 7

Conclusão

Considerando a grande quantidade de informações geradas reiteradamente, a aquisição de conhecimento e a visualização de dados se tornaram um desafio nos dias atuais. Nesse sentido, é importante entender os propósitos das técnicas de visualização que apresentam graficamente conjuntos de dados para que a representação visual gerada possa explorar a capacidade de percepção humana. A partir da utilização de técnicas de visualização, especialistas de domínio e usuários de várias áreas de conhecimento podem interpretar e compreender as relações espaciais dos dados exibidos, levando à obtenção de conhecimento implícito e potencialmente útil.

Nesse contexto, este trabalho consistiu no estudo e na avaliação de técnicas de visualização aplicadas no contexto educacional, com enfoque nas disciplinas iniciais de programação da Universidade de Brasília. Para tanto, foi realizado um estudo de caso em três disciplinas (ICC, CB e APC), levando em consideração fatores sociais e acadêmicos dos alunos.

Foram analisados os algoritmos de visualização pertencentes aos grupos: visualização de dados, de informação e algoritmos não categorizados por Lengler & Eppler em [33]. Os algoritmos foram aplicados em duas bases de dados distintas, o SIGRA e um questionário aplicado no início e no final dos semestres nas disciplinas de APC e ICC. A partir dessas informações, foram identificados três domínios para a elaboração das visualizações: informações acadêmicas, cujo intuito é informar sobre o desempenho dos alunos; informações gerais que permite a visualização das disciplinas como um todo; e informações de percepção que explicita os dados sociais dos alunos. A base de dados do SIGRA foi utilizada nos domínios de informações acadêmicas e gerais, enquanto que o questionário foi empregado no domínio de informações de percepção.

Para guiar os estudos nos domínios especificados, foram elaboradas perguntas em que as visualizações candidatas e os dados foram estabelecidos. Com o objetivo de determinar se os algoritmos selecionados comunicam a informação de forma clara e coesa foi

aplicado um questionário a 23 professores e coordenadores, composto por perguntas que visam a analisar o conhecimento do questionado sobre as visualizações, além de questões elaboradas para cada conjunto de algoritmos, de forma a avaliá-los.

No domínio de informações gerais, os seguintes gráficos foram os favoritos por parte dos questionados para representar as informações referentes às perguntas: gráfico de pizza, de barras, de linhas associadas ao gráfico de barras e o treemap. Já no domínio de informações acadêmicas, o gráfico de linhas, de violino, de área, e de barras associado ao de linhas permitiram que mais questionados interpretassem os dados de forma correta e rápida. Por fim, o domínio de informações de percepção incluiu adequadamente os seguintes algoritmos: gráfico de barras, gráfico de pizza e o diagrama de sankey.

A partir do resultado do estudo, observou-se que os questionados se sentiram mais seguros ao utilizar algoritmos já conhecidos, como o gráfico de pizza e o gráfico de barras. Dentre os selecionados, o diagrama de sankey, o treemap e o gráfico de violino eram os menos conhecidos pelos questionados. Ademais, o gráfico de barras foi o algoritmo que teve melhores resultados quando aplicado. Nos gráficos selecionados foram aplicados os dados com o objetivo de extrair conhecimento, sendo possível analisar a questão de gênero referente às disciplinas iniciais de programação, o desenvolvimento dos alunos cotistas nestas disciplinas e a diferença de percepção entre os alunos de ICC e APC.

A partir do estudo percebeu-se que os estudantes do sexo feminino possuem mais dificuldade no primeiro contato com programação, se comparado com os alunos do sexo masculino, o que aumenta o nível de reprovação. Entretanto, ao cursar novamente a disciplina as mulheres apresentam melhora no desempenho. Em relação aos alunos que entraram na universidade por meio do sistema de cotas, foi possível visualizar que estes possuem um desempenho pior do que os alunos não-cotistas, entretanto as notas deste último grupo caíram no último período. Por fim, observou-se que os alunos que cursam a disciplina APC já entram sabendo programação, ao contrário dos estudantes de ICC. Além disso, este último grupo tem a percepção de que necessita de mais horas de estudos do que os colegas.

Nesse sentido, o presente trabalho proporcionou um estudo detalhado de diversas técnicas de visualização comumente encontradas na literatura. Entretanto, por terem sido selecionados diversos algoritmos, apenas oito professores e gestores avaliaram cada conjunto de gráficos, restringindo assim a avaliação. Além disso, as análises das notas foram limitadas às menções e aos dados acadêmicos obtidos do SIGRA, sem levar em consideração questões como metodologia de ensino, mudanças de linguagem ao longo do tempo e professores da disciplina.

Para dar continuidade a este trabalho, é proposto o estudo de técnicas de visualização a serem aplicadas em outras disciplinas do Departamento de Ciência da Computação, com

foco no auxílio de tomada de decisão por parte dos professores e gestores educacionais. Como também é proposta a criação de um módulo virtual, a ser implantado no Moodle, com as visualizações selecionadas a partir deste trabalho.

Referências

[1] Castells, Manuel e Rita Espanha: *A era da informação: economia, sociedade e cultura*, volume 1. Fundação Calouste Gulbenkian. Serviço de Educação e Bolsas, 2007. 1

[2] Kinnunen, Päivi e Lauri Malmi: *Why students drop out cs1 course?* Proceedings of the Second International Workshop on Computing Education Research, páginas 97–108, 2006. 1, 20

[3] Guzdial, Mark e Elliot Soloway: *Teaching the nintendo generation to program.* Communications of the ACM, 45(4):17–21, 2002. 1, 20

[4] Bennedsen, Jens e Michael E Caspersen: *Failure rates in introductory programming.* ACM SIGCSE Bulletin, 39(2):32–36, 2007. 1, 20

[5] Rodrigues, Rodrigo Lins, Jorge Luis Cavalcanti Ramos, João Carlos Sedraz Silva e Alex Sandro Gomes: *A literatura brasileira sobre mineração de dados educacionais.* Anais dos Workshops do Congresso Brasileiro de Informática na Educação, 3(1):621, 2014. 1

[6] Dragon, Toby e Carrie Lindeman: *Automated assessment of students' conceptual understanding: Supporting students and teachers using data from an interactive textbook.* International Symposium on Multimedia (ISM), páginas 567–572, 2017. 1

[7] De Amo, Sandra: *Técnicas de mineração de dados.* Jornada de Atualização em Informática, 2004. 1

[8] Camilo, Cássio Oliveira e João Carlos da Silva: *Mineração de dados: Conceitos, tarefas, métodos e ferramentas.* Universidade Federal de Goiás (UFC), páginas 1–29, 2009. 1

[9] Freitas, Carla Maria Dal Sasso, Olinda Mioka Chubachi, Paulo Roberto Gomes Luzzardi e Ricardo Andrade Cava: *Introdução à visualização de informações.* Revista de Informática Teórica e Aplicada. Porto Alegre. Vol. 8, n. 2 (out. 2001), p. 143-158, 2001. 2

[10] Costa, Jean Carlos Araújo: *Vis-Scholar: uma metodologia de visualização e análise de dados na educação.* Tese de Doutoramento, Universidade do Vale do Rio dos Sinos, 2016. 2

[11] Nascimento, Hugo Alexandre Dantas do e Cristiane Bastos Rocha Ferreira: *Uma introdução à visualização de informações.* Visualidades, 2011. 2

[12] Estivalet, Luiz Fernando: *O Uso de Ícones na visualização de Informações.* Tese de Doutoramento, Universidade Federal do Rio Grande do Sul, 2000. 2

[13] Do Nascimento, Hugo AD e Cristiane BR Ferreira: *Visualização de informações–uma abordagem prática.* Em *XXV Congresso da Sociedade Brasileira de Computação, XXIV Jornada de Atualização em Informatica. UNISINOS, S. Leopoldo–RS*, 2005. 2

[14] Ackoff, Russell L: *From data to wisdom.* Journal of Applied Systems Analysis, 16(1):3–9, 1989. 4

[15] INEP: *Conheça o INEP Disponível em: http://portal.inep.gov.br/conheca-o-inep.* 4, 5

[16] Nascimento, Rafaella Leandra Souza do, Geraldo Gomes da Cruz Junior e Roberta Andrade de Araújo Fagundes: *Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep.* RENOTE-Revista Novas Tecnologias na Educação, 16(1), 2018. 4

[17] Silberschatz, Abraham, Henry F Korth, Shashank Sudarshan *et al.*: *Database system concepts*, volume 5. McGraw-Hill New York, 1997. 4

[18] Komorowski, Jan e Jan Zytkow: *Principles of Data Mining and Knowledge Discovery: First European Symposium, PKDD'97, Trondheim, Norway, June 24-27, 1997 Proceedings*, volume 1. Springer Science & Business Media, 1997. 5

[19] Bandeira, Judson, Thiago Ávila, Williams Alcantara, Armando Sobrinho, Ig Ibert Bittencourt e Seiji Isotani: *Dados abertos conectados para a educação.* Jornada de Atualização em Informática na Educação, 4(1):47–69, 2015. 5

[20] Júnior, Tércio Sampaio Ferraz: *Sigilo de dados: o direito à privacidade e os limites à função fiscalizadora do estado.* Revista da Faculdade de Direito, Universidade de São Paulo, 88:439–459, 1993. 5

[21] Cechinel, Cristian e Sandro da Silva Camargo: *Mineração de dados educacionais: avaliação e interpretação de modelos de classificação.* Jaques, Patrícia Augustin; Siqueira; Sean; Bittencourt, Ig; Pimentel, Mariano.(Org.) Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa. Porto Alegre: SBC, 2020. 5

[22] Spiegel, Murray R, John J Schiller, R Alu Srinivasan e Mike LeVan: *Probability and statistics*, volume 2. Mcgraw-hill New York, 2009. 5

[23] Rice, John A.: *Mathematical statistics and data analysis.* Wadsworth & Brooks/Cole, 2007. 6

[24] Morettin, Pedro Alberto e Wilton Oliveira Bussab: *Estatística básica.* Editora Saraiva, 2017. 6

[25] Sedgwick, Philip: *Pearson's correlation coefficient.* British Medical Journal Publishing Group, 345, 2012. 7

[26] Ward, Matthew O, Georges Grinstein e Daniel Keim: *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010. 7

[27] Khan, Muzammil e Sarwar Shah Khan: *Data and information visualization methods, and interactive mechanisms: A survey*. International Journal of Computer Applications, 34(1):1–14, 2011. 7

[28] Dias, Mateus Pereira e José Oscar Fontanini de Carvalho: *A visualização da informação e a sua contribuição para a ciência da informação*. Revista de Ciência da Informação, 8(5):01–16, 2007. 7

[29] Tufte, Edward e P Graves-Morris: *The visual display of quantitative information.; 1983*, 2014. 7

[30] Card, Mackinlay: *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. 7

[31] Liu, Shixia, Weiwei Cui, Yingcai Wu e Mengchen Liu: *A survey on information visualization: recent advances and challenges*. The Visual Computer, 30(12):1373–1393, 2014. 7

[32] Butler, David M, James C Almond, R Daniel Bergeron, Ken W Brodlie e Robert B Haber: *Visualization reference models*. Proceedings of the 4th Conference on Visualization'93, páginas 337–342, 1993. 7

[33] Lengler, Ralph e Martin J Eppler: *Towards a periodic table of visualization methods for management*. Em *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA*, 2007. 7, 8, 31, 84

[34] Akoka, Jacky, Isabelle Comyn-Wattiau e Nabil Laoufi: *Research on big data–a systematic mapping study*. Computer Standards & Interfaces, 54:105–115, 2017. 16

[35] Caiado, Rodrigo, Luiz Alberto Rangel, Osvaldo Quelhas e Daniel Nascimento: *Metodologia de revisão sistemática da literatura com aplicação do método de apoio multicritério à decisão smarter*. Congresso Nacional de Excelência em Gestão e III Inovarse – Responsabilidade Social e Aplicada, 12:1–20, 2016. 16

[36] VOSviewer: *VOSviewer - visualizing scientifics landscape. Disponível em: https://www.vosviewer.com/*. 16

[37] Hansen, Luiza, Vinicius RP Borges e Maristela Holanda: *A literature study of visual analysis in an educational context*. Em *2020 IEEE Frontiers in Education Conference (FIE)*, páginas 1–8. IEEE, 2020. 18, 31

[38] Essa, Alfred e Hanan Ayad: *Improving student success using predictive models and data visualisations*. Research in Learning Technology, 20, 2012. 18

[39] Li, Mingran, Wenjie Wu, Junhan Zhao, Keyuan Zhou, David Perkis, Timothy N Bond, Kevin Mumford, David Hummels, Yingjie Victor Chen e Mike Potel: *Careervis: hierarchical visualization of career pathway data*. IEEE Computer Graphics and Applications, 38(6):96–105, 2018. 19

[40] Zlatarov, P, G Ivanova e D Baeva: *A web-based system for personalized learning path tracking of doctoral students*. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), páginas 773–778, 2019. 19

[41] Menezes, Douglas Afonso Tenório de, Diogo Lima Florêncio, Renato Ely Domingues Silva, Isabel Dillmann Nunes, Ulrich Schiel e Marcus Salerno de Aquino: *David—a model of data visualization for the instructional design*. Em *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, páginas 281–285, 2017. 19

[42] Venant, Rémi, Philippe Vidal e Julien Broisin: *Evaluation of learner performance during practical activities: an experimentation in computer education*. Em *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, páginas 237–241, 2016. 19

[43] Hernández-García, Ángel, Inés González-González, Ana Isabel Jiménez-Zarco e Julián Chaparro-Peláez: *Visualizations of online course interactions for social network learning analytics*. International Journal of Emerging Technologies in Learning (iJET), 11(07):6–15, 2016. 19

[44] Jankowski, Krzysztof, Antti Knutas, Jouni Ikonen e Jari Porras: *Automated social network analysis of online student collaboration activity*. Em *Proceedings of the 16th International Conference on Computer Systems and Technologies*, páginas 326–333, 2015. 19

[45] Dobashi, Konomu: *Automatic data integration from moodle course logs to pivot tables for time series cross section analysis*. Procedia computer science, 112:1835–1844, 2017. 19

[46] Dragon, Toby e Carrie Lindeman: *Automated assessment of students' conceptual understanding: Supporting students and teachers using data from an interactive textbook*. Em *2017 IEEE International Symposium on Multimedia (ISM)*, páginas 567–572, 2017. 19

[47] Barria-Pineda, Jordan, Julio Guerra-Hollstein e Peter Brusilovsky: *A fine-grained open learner model for an introductory programming course*. Em *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, páginas 53–61, 2018. 19

[48] Culligan, Natalie, Keith Quille e Susan Bergin: *Veap: A visualisation engine and analyzer for press#*. Proceedings of the 16th Koli Calling International Conference on Computing Education Research, páginas 130–134, 2016. 20

[49] Menezes, Douglas Afonso Tenório de, Diogo Lima Florêncio, Renato Ely Domingues Silva, Isabel Dillmann Nunes, Ulrich Schiel e Marcus Salerno de Aquino: *David—a model of data visualization for the instructional design*. 17th International Conference on Advanced Learning Technologies (ICALT), páginas 281–285, 2017. 20

[50] Silva, Gabriel Lenon Barros, Janaína Alexandre de Carvalho e Alexandre Magno Andrade Maciel: *Desenvolvimento de um learning analytics dashboard a partir de modelos de mineração de dados educacionais*. Revista de Engenharia e Pesquisa Aplicada, 6(3):59–69, 2021. 20, 55

[51] Luzzardi, Paulo Roberto Gomes: *Critérios de Avaliação de Técnicas de Visualização de Informações Hierárquicas. 2003*. Tese de Doutoramento, Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre., 2003. 20, 55

[52] Valiati, Eliane Regina de Almeida: *Avaliação de usabilidade de técnicas de visualização de informações multidimensionais*. 2008. 55, 56

[53] Oliva, Luciana Menezes da Cunha Rêgo: *Sistema de cotas na universidade pública brasileira: avaliação da experiência da unb após a lei 12.711/12*. 2020. 76

Apêndice A

Informações Detalhadas dos Dados do Formulário

- Turma - informa a disciplina do aluno que está respondendo o questionário, em que as opções são APC ou ICC;
- Período - informa o ano e o semestre em que o questionário foi respondido. É um campo numérico, onde os quatro primeiros dígitos são o ano de saída da opção e o último representa o semestre;
- Escolha do curso - Desscreve quais fatores contribuíram para a escolha do curso pelo aluno. As opções são: Alta probabilidade de arrumar emprego, Tem bons salários, Conselho da Família, Conselho do Professor, Vocação e Outros;
- Área de trabalho dos pais - Em que área os seus pais e irmãos trabalham? - representa as áreas em que a família do aluno trabalha, podendo ser: Educação, Saúde, Tecnologia/Computação/Engenharia, Nenhuma das opções;
- Turma APC - Qual a sua Turma de APC? - identifica, dos formulários da matéria APC, quais eram as turmas dos alunos que responderam. Os valores nulos são provenientes das respostas dos formulários de alunos que não possuem essa pergunta, incluindo as turmas de ICC;
- Pais têm ensino superior - Você é a primeira pessoa da família (entre seu pai e mãe) a fazer curso superior? - informa se é a primeira geração da família do aluno a cursar o nível superior, a resposta pode ser sim ou não;
- Matemática é para pessoas inteligentes - Com que frequência você ouve “Engenharia/Computação é para pessoas super inteligentes” - essa pergunta é para analisar a percepção do aluno sobre os cursos de Engenharia/Computação e se escuta fre-

quentemente que esses cursos são para pessoas inteligentes. As opções de resposta são: Algumas vezes, Com frequência, Nunca, Ocasionalmente e Sempre;

- Computação é para homens - Com que frequência você ouve "Engenharia/Computação é para homens" - essa pergunta é para analisar a percepção do aluno sobre os cursos de Engenharia/Computação e se escuta frequentemente que esses cursos são para homens. As opções de resposta são: Algumas vezes, Com frequência, Nunca, Ocasionalmente e Sempre;
- Idade - Qual a sua idade? - identifica a faixa etária dos alunos que responderam o questionário, onde as opções são: 18 - 20 anos, 20 -23 anos, Mais de 23 anos e Menos que 17 anos;
- Gênero - Qual o seu sexo? - informa o gênero do aluno que respondeu o questionário, podendo ser Masculino, Feminino ou Prefiro não declarar;
- É cotista - Você entrou pelo sistema de cota da UnB? - informa se o aluno entrou por cotas na UnB e qual, sendo as respostas: Não,-; Sim, Cota Racial; Sim, Cota Social (Baixa Renda); Sim, Cota Social (Escola Pública); ou Sim. Cota Racial;
- Tipo de escola - Você fez o seu Ensino em Médio em: - informa se o aluno estudou o ensino médio na escola pública ou particular;
- Curso - Qual o seu curso de graduação? - informa o curso do aluno. As opções são: Ciência da Computação, Computação (Noturno), Engenharia Civil, Engenharia da Computação, Engenharia de Produção, Engenharia Elétrica, Engenharia Florestal, Engenharia Mecânica, Engenharia Mecatrônica, Estatística, Matemática ou Outro;
- Gostava de matemática - Matemática era a matéria que você mais gostava no ensino médio? - informa se o aluno tinha preferência, ou não, por matemática no ensino médio;
- Experiência em programação - Você já tinha experiência em programação? - informa se o aluno já tinha, ou não, experiência em programação.
- Onde teve experiência - Se você respondeu sim na questão anterior, informe onde aprendeu a programar - informa, caso o aluno tenha experiência em programação, onde se deu essa experiência. As opções de resposta são: Em cursos em escolas de programação, Já fiz essa disciplina anteriormente, Em cursos online, Na minha escola do ensino médio, Outros;
- Liguagem de experiência - Se você tinha experiência em programação antes dessa disciplina, informe em quais linguagens - informa, caso o aluno tenha experiência

em programação, em qual linguagem de programação se deu essa experiência. As opções são: C/C++, Python, Java, Outras.

- Horas estudadas - Durante o semestre, quantas horas você passou fazendo os deveres de casa e estudando fora da sala de aula por semana para essa disciplina - informa a faixa de horas estudadas pelo aluno, podendo ser: 1 - 2 horas, 3 - 5 horas, 6 - 10 horas, Mais de 10 horas, Menos de uma hora e Nunca estudei fora do horário de aula;
- Precisa de mais horas de estudo - Você acha que precisa de mais horas de estudos nessa disciplina do que seus amigos de sala? - é uma pergunta sobre a percepção do aluno sobre si mesmo e os colegas, de forma a considerar se este precisa de mais horas de estudos, ou não;
- Nota consistente - Você acha que seu esforço na disciplina foi recompensado (sua nota está coerente com os seus estudos)? - essa pergunta é para analisar a percepção do aluno sobre as notas e se acredita que foram consistentes com os estudos;
- Gosta de programar - Quanto você gosta de programar? - informa se o aluno gosta, gosta mais ou menos, gosta muito, gosta um pouco ou não gosta de programar;
- É inteligente - Você acha que seus amigos de sala pensam que você não é inteligente o suficiente para fazer um curso de Computação/Engenharia? - essa pergunta é para analisar a percepção do aluno sobre o que acredita que os colegas pensam dele;
- Tem idéias ignoradas - Com que frequência você sente que teve boas ideias porém seus amigos de sala as ignoraram? - informa se o aluno considera que suas boas ideias são ignoradas algumas vezes, com frequência, nunca, ocasionalmente, ou sempre;
- Sugestão (*suggestions*) - Que sugestão você daria para melhorarmos a disciplina? - é um campo textual para o aluno fazer alguma sugestão para a matéria.