UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

LUCAS MATHEUS DIAS BRUM

Análise da demanda de discentes pós - Ensino Remoto Emergencial: uma abordagem filosófica

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Profa. Dra. Renata Galante

Porto Alegre
2024

“ Tudo no mundo está dando respostas, o que demora é o tempo das perguntas.“

— JOSÉ SARAMAGO (MEMORIAL DO CONVENTO)

AGRADECIMENTOS

Queria agradecer primeiramente à minha mãe, meu alicerce desde sempre, me apoiou desde o início nos momentos bons e ruins, não permitiu que eu desistisse, me deu tudo o que eu precisava e muito mais, eu estou aqui em grande parte por causa dela, te amo mãe, é um orgulho ser seu filho, não sei o que seria de mim sem você. Obrigado à professora Renata Galante pela orientação feita que me conduziu durante todo esse trabalho. Obrigado Deus por permitir que tudo fluísse até chegar onde estou hoje. Obrigado UFRGS pela oportunidade, desafios, momentos alegres, me apresentar pessoas incríveis e que levarei pra vida toda, tanto quanto a formação. Por fim, obrigado a quem, por minimamente que seja, ajudou de alguma forma para a realização deste trabalho.

RESUMO

A análise de dados deste tipo envolve a coleta, processamento e interpretação dos dados numéricos com o objetivo de extrair informações úteis e obter *insights*. No contexto educacional, esta análise permite avaliar diversos aspectos , como taxas de matrícula, evasão, desempenho acadêmico e engajamento dos alunos. Com a adoção do ERE, essas métricas podem ter sofrido alterações significativas, refletindo os desafios e as adaptações enfrentadas por alunos e instituições. Este trabalho tem como objetivo tentar explicar a procura pelos cursos de ensino superior, sumamente importante para a formação do ser humano como trabalhador e pensante, principalmente pelos cursos de Ciência da Computação e Engenharia da Computação na Universidade Federal do Rio Grande do Sul, onde notou se um número elevado de alunos tanto no Instituto de Informática quanto no Restaurante Universitário próximo a ele, ambos localizados no Campus do Vale, em Porto Alegre. Sendo assim, uma análise exploratória de dados intensa foi feita, por meio de consultas, gráficos e *dashboard* que buscam ilustrar os padrões que se mostram nos dados, que por sua vez podem ou não responder nossas perguntas e também fortemente discutidas, tal qual como em um trabalho de pesquisa.

Palavras-chave: Ciência de Dados. Estatística Descritiva. Inteligência de Negócio. Desempenho Acadêmico.

Analysis of Student Demand Post Emergency Remote Teaching: A Philosophical Approach

ABSTRACT

The analysis of this type of data involves the collection, processing, and interpretation of numerical data with the goal of extracting useful information and gaining insights. In the educational context, this analysis allows for the evaluation of various aspects such as enrollment rates, dropout rates, academic performance, and student engagement. With the adoption of remote education (ERE), these metrics may have undergone significant changes, reflecting the challenges and adaptations faced by students and institutions. This work aims to explain the demand for higher education courses, which is extremely important for human development as both workers and thinkers, especially for the Computer Science and Computer Engineering courses at the Federal University of Rio Grande do Sul, where a high number of students were observed both at the Institute of Informatics and the nearby University Restaurant, both located at the Campus do Vale in Porto Alegre. Therefore, an intense exploratory data analysis was conducted through queries, charts, and a dashboard that seek to illustrate the patterns present in the data, which may or may not answer our questions and will also be thoroughly discussed, as in any research work.

Keywords: Data Science, Descriptive Statistics, Business Intelligence, Academic Performance.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

UFRGS - Universidade Federal do Rio Grande do Sul

CIC - Ciência da Computação

CPD - Centro de Processamento de Dados

ECP - Engenharia da Computação

INF - Instituto de Informática

SQL - Structured Query Language - Linguagem de Consulta Estruturada

ABNT - Associação Brasileira de Normas Técnicas

ERE - Ensino Remoto Emergencial

EAD - Ensino A Distância

SGBD - Sistema de Gerenciamento de Banco de Dados

ICOMP - Instituto de Computação da Universidade Federal do Amazonas

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, a pandemia do Sars-Cov-2, vírus que causa a COVID-19, que surgiu em 2020 (OPAS, 2024) trouxe mudanças significativas para diversos setores da sociedade. Uma das principais medidas adotadas para conter a disseminação da covid-19 foi o *lockdown* ou confinamento, fazendo com que o país e o mundo se recolhessem em suas casas a fim de evitar o contágio com um vírus até então totalmente novo para a medicina da época, fechando o comércio, escolas, universidades e outros por tempo indeterminado. Durante esse período, imagens de grandes centros urbanos com ruas desertas circularam pelo mundo, e em muitos desses lugares, as saídas de casa eram permitidas apenas para comprar medicamentos, ir ao hospital ou ao mercado para adquirir itens essenciais (GUITARRARA, 2024). Com isso, a situação da época fez com que muitas instituições educacionais migrassem para o Ensino Remoto Emergencial, para atender aos seus estudantes durante o período em que essas instituições precisaram ficar fechadas e seguindo o confinamento em suas casas (BEHAR, 2020). Este período se mostrou um desafio, tanto para alunos quanto para as instituições de ensino superior pelo mundo. Diante deste cenário, a análise de dados quantitativos emergiu como uma ferramenta essencial para entender as dinâmicas e os impactos causados por essas mudanças na trajetória acadêmica dos estudantes.

A análise de dados deste tipo envolve a coleta, processamento e interpretação dos dados numéricos com o objetivo de extrair informações úteis e obter *insights*. No contexto educacional, esta análise permite avaliar diversos aspectos , como taxas de matrícula, evasão, desempenho acadêmico e engajamento dos alunos. Com a adoção do ERE, essas métricas podem ter sofrido alterações significativas, refletindo os desafios e as adaptações enfrentadas por alunos e instituições.

Este trabalho de conclusão de curso tem como objetivo explorar essas questões através de uma análise detalhada e crítica dos dados de alunos da Universidade Federal do Rio Grande do Sul, compreendidos entre os anos de 2010 a 2022, buscando fornecer respostas baseadas em evidências quantitativas. Através de consultas, gráficos e *dashboards*, serão ilustrados os padrões e tendências emergentes dos dados coletados, que podem ou não confirmar nossas hipóteses iniciais, a começar pela verificação se houve real aumento de discentes pós-ERE, quando os discentes retornaram ao modo presencial em 2022. Tal qual em um trabalho de pesquisa, a análise de dados quantitativos nos permitirá explorar profundamente os impactos do Ensino Remoto Emergencial e a pandemia

do coronavírus também, contribuindo para o entendimento e melhoramento das práticas educacionais no futuro.

Em suma, a combinação da análise exploratória de dados quantitativos e da ciência de dados, bem como da investigação de dados, apresenta um método robusto para explorar e compreender as modificações que ocorreram no cenário educacional durante e após a pandemia. O objetivo deste estudo busca não apenas abordar as nossas questões de investigação, mas também equipar outros investigadores com uma base sólida para futuras investigações ou intervenções no campo da educação.

O objetivo deste trabalho é analisar como um fenômeno biológico tão grande como a pandemia mundial do coronavírus afetou a vida acadêmica de tantos alunos na UFRGS e refletir sobre isso, seja pelos alunos que entraram, seja pelos que saíram ou pelos que permanecem. Esta análise tenta, além de verificar se realmente houve um aumento de alunos após o período de exceção na universidade, estudar os efeitos deste nos dados quantitativos da universidade, trazendo uma possível ampliação na visão que temos do aluno no ensino superior e também fora dele.

O restante do texto está organizado da seguinte forma: no Capítulo 2 estão os conceitos teóricos e tecnológicos acerca do trabalho, ilustrando as mais relevantes definições e contextualizações necessárias para o entendimento do mesmo. No capítulo 3, estão relatados os trabalhos que estão relacionados com a proposta deste de alguma forma. No capítulo 4, está explicada a metodologia empregada durante o trabalho, explicando de forma introdutória desde o entendimento do problema até a análise exploratória feita. No capítulo 5, estão expostas as análises quantitativa e exploratória, na análise quantitativa as consultas feitas na base de dados a fim de responder questões que, por sua vez, podem auxiliar na busca pela resposta de se realmente houve um impacto aumentativo nos dados de alunos que ingressaram depois do período de exceção e a análise exploratória em si, onde são mostrados alguns gráficos relevantes que podem responder à questão de pesquisa. Por fim, o capítulo 6 traz as considerações finais sobre os resultados obtidos, encerrando a monografia.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, enfatiza-se os conceitos e ferramentas utilizados para a realização deste trabalho, como bibliotecas, linguagem de programação e outros *softwares*. Na seção 2.1, explicam-se os conceitos enquanto que, na seção 2.2 explica-se as tecnologias e ferramentas utilizadas.

2.1 Conceitos

Esta seção apresenta os conceitos teóricos utilizados durante a realização deste trabalho, estão explicados os conceitos de estatística utilizados para a execução e justificativa dos mesmos para este trabalho. A seção 2.1.1 discorre sobre a análise exploratória de dados.

2.1.1 Análise exploratória de dados

Como (MEDRI, 2011), desde a antiguidade, diversos povos registravam dados sobre população, nascimentos, óbitos, riquezas e impostos utilizando métodos que hoje chamamos de Estatística. Derivada do latim *status*, que em português significa "estado", a Estatística inicialmente servia para administração estatal. Hoje, é definida como um conjunto e métodos quantitativos para estudar fenômenos coletivos, relacionando-se e comportando-se de acordo com a ciência.

A Estatística vai muito além da construção de gráficos e cálculo de médias, envolvendo planejamento de experimentos, coleta, organização, resumo, análise, interpretação de dados e tomada de decisões. Estas informações são amplamente apresentadas em meios de comunicação, como dados sobre nascimentos, mortes e previsão do tempo. Apesar das técnicas clássicas da Estatística serem desenvolvidas sob rigorosas suposições, a experiência mostrou que essas técnicas podem falhar em situações práticas não ideais. Métodos exploratórios robustos estão sendo desenvolvidos para melhorar a análise estatística.

Os profissionais capacitados de Estatística examinam os dados detalhadamente antes de fazer suposições e testes de hipóteses. No entanto, o uso indiscriminado de pacotes estatísticos sem cuidadosa análise pode levar a resultados errôneos. A análise exploratório-

ria de dados é fundamental para extrair o máximo de informação dos dados, indicando modelos plausíveis para uma posterior análise confirmatória ou inferência estatística.

Se entendemos estatística como a Ciência de Dados, o domínio de seu conhecimento será extremamente importante. Como ponto de partida, podemos dividir a Estatística em duas áreas principais: descritiva e inferencial (ou indutiva).

2.1.1.1 Estatística descritiva

A estatística descritiva é empregada na organização, apresentação e síntese dos dados, usam de gráficos, tabelas e medidas descritivas como ferramentas. É aplicada na fase inicial da análise para obter informações (e até mesmo *insights*) que sugerem possíveis modelos a serem empregados posteriormente, na fase conhecida como Inferência Estatística.

2.1.1.2 Estatística inferencial

A estatística inferencial emprega técnicas para utilizar dados de uma amostra e fazer generalizações sobre a população. Essas técnicas incluem a determinação do tamanho da amostra, o método de seleção das unidades observacionais, o cálculo das medidas estatísticas, a determinação da confiança nas estimativas, a significância dos testes estatísticos e a precisão das estimativas. Essa generalização é baseada na estimação das medidas estatísticas, antecipando um grau de certeza de que a amostra representa a população. A probabilidade é utilizada para avaliar a confiabilidade de cada inferência baseada na amostra. Antes de iniciar o estudo dos métodos estatísticos para a análise de dados, tanto qualitativos quanto quantitativos, é fundamental introduzir alguns conceitos preliminares. Isto não apenas nomeará os instrumentos, mas também padronizará a terminologia que será utilizada ao longo do curso. Na Estatística, o termo *população* se refere ao grande conjunto de dados que contém uma determinada característica de interesse. Este conjunto pode ser qualquer alvo de estudo, como todos os habitantes de uma cidade ou todas as lâmpadas produzidas por uma fábrica em um determinado período. Embora seja possível estudar toda uma população em certos casos, muitas vezes isso não é viável devido à restrições econômicas, por exemplo, éticas ou práticas. Por exemplo, uma empresa pode não possuir recursos suficientes para pesquisar todos os consumidores ou pode ser impossível verificar todas as lampadas produzidas para determinar seu tempo de funcionamento/ciclo de vida. Tendo em vista todas essas dificuldades provindas de várias naturezas em obser-

var todos os elementos de uma população, selecionamos um subconjunto menor para se estudar. Este subconjunto da população em geral com menor dimensão, chamaremos de *amostra*.

2.1.2 População e Amostra

A seguir, nas seções 2.1.2.1 e 2.1.2.2 estão descritos os conceitos da Estatística mais importantes para o ponto de partida de uma pesquisa estatística, *população* e *amostra*.

2.1.2.1 População

População nada mais é que o conjunto de todos os indivíduos que compartilham pelo menos uma característica em comum, cujo comportamento é de interesse para análise (inferência). O objetivo das generalizações estatísticas é determinar se algo pode ser afirmado sobre várias características da população estudada, com base em dados conhecidos.

2.1.2.2 Amostra

Uma amostra é um subconjunto selecionado das observações acerca da população, usado para fazer inferências sobre as características da população. É crucial que uma amostra seja representativa e sua seleção e manipulação exigem cuidados especiais para evitar distorções nos resultados. A escolha da amostra pode variar conforme o conhecimento que temos da população, os recurso disponíveis, entre outros fatores.

2.1.3 Variáveis

Quando se executa um estudo estatístico de um determinado fato ou grupo, deve-se considerar o tipo da variável (ou variáveis) que será objeto de estudo. Podem existir variáveis *quantitativas* ou variáveis *qualitativas*.

2.1.3.1 Variáveis qualitativas

Variáveis qualitativas são variáveis que expressam valores em forma de categorias, classes ou rótulos. Tratam-se, portanto de dados não-numéricos. Apesar de possuírem de baixo nível de mensuração, estas variáveis tem uma ampla aplicabilidade nas ciências sociais e comportamentais. Descrevem características individuais das unidade de análise, como sexo, estado civil, naturalidade, raça, nível educacional, entre outras, permitindo a estratificação das unidades para análise conjunta com outras variáveis.

2.1.3.2 Variáveis quantitativas

Variáveis quantitativas são aquelas expressas com nível de mensuração intervalar ou de razão. Em outras palavras, essas variáveis assumem valores em uma escala métrica definida por uma origem e uma unidade, como idade, salário e peso, por exemplo. As variáveis qualitativas, por outro lado, podem ser classificadas como nominal ou ordinal. As quantitativas, por sua vez, podem ser classificadas como discretas, quando assumem um número finito de valores, ou contínuas, quando assumem um número infinito de valores, geralmente em intervalos.

2.1.3.3 Variáveis aleatórias

Uma variável aleatória é uma função que associa a cada ponto pertencente a um espaço amostral Ω (SERMARINI, 2016). São divididas em dois tipos: variável aleatória discreta e variável aleatória contínua.

- Uma quantidade X, associada a cada possível resultado do espaço amostral , que assume valores dentro de um conjunto enumerável, cada um com uma certa probabilidade é denominada variável aleatória discreta.

- Uma quantidade X associada a cada possível resultado do espaço amostral, cujo conjunto de valores é qualquer intervalo dos números reais, formando assim um conjunto não enumerável é denominada variável aleatória contínua.

2.1.4 Correlação

Conforme (FILHO; JÚNIOR, 2009), verificar se existe relação entre X e Y é uma pergunta fundamental na vida de qualquer pesquisador. Por exemplo, para afirmar que a taxa de suicídio entre protestantes é maior do que a de católicos, Durkheim sugere uma correlação entre denominação religiosa e a propensão ao suicídio. Do mesmo modo, ao postular que o sistema eleitoral majoritário tende a produzir sistemas bipartidários, a Lei de Duverger sugere a existência de uma correlação entre o tipo de regra eleitoral (majoritária ou proporcional) e a quantidade de partidos. Mas o que significa dizer que duas variáveis estão correlacionadas?

Na verdade, as correlações variam com respeito à sua força. Podemos visualizar diferenças quanto à força de correlação por meio de um diagrama de dispersão, que trata-se de um gráfico capaz de mostrar a maneira pela qual os valores de duas variáveis X e Y distribuem-se ao longo da faixa dos possíveis resultados.

Podemos dizer que a força de uma correlação entre X e Y aumenta à medida que os pontos, no diagrama de dispersão, mais compactamente se agrupam em torno de uma linha reta *imaginária*.

O sentido da correlação também podem descrever coisas sobre as variáveis que estão sendo cruzadas. O sentido pode ser *positivo* ou *negativo*. Uma *correlação positiva* pode indicar que, conforme X cresce, Y também cresce de forma semelhante, de forma recíproca, se Y descresce, X tambem tende a decrescer. Quando há um crescimento em X e em contrapartida um decrescimento na variável Y, existe o que se chama *correlação negativa*. Reciprocamente, também ocorre a mesma quando X cresce e Y decresce e vice-versa.

Conforme (ANDERSON, 2023), também existe o que chamamos de *coeficiente de correlação*, considerando a definição dos diferentes tipos de correlação que podemos ter, podemos ver como este valor estatístico é calculado, da seguinte forma: o coeficiente da correlação, também chamado de coeficiente de correlação linear ou coeficiente de correlação de Pearson, é o valor da correlação entre duas variáveis.

O coeficiente de correlação entre duas variáveis estatísticas é definido como o quociente da covariância das variáveis pela raiz quadrada do produto das variâncias de cada variável. Assim, a fórmula para calcular o coeficiente de correlação é a seguinte:

Quando calcula-se o coeficiente de correlação em uma população, o símbolo de correlação é a letra grega rô. Porém quando o coeficiente é calculado em relação a uma

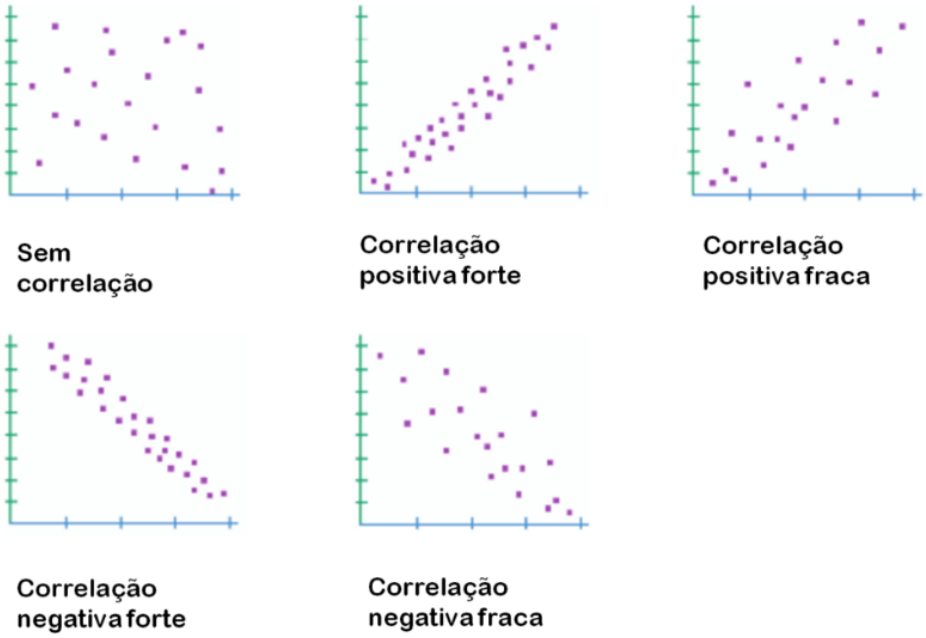
$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

amostra, a letra r é geralmente usada como símbolo. O valor do índice de correlação pode estar entre -1 e +1, estes mesmos inclusos. Vejamos a seguir como o valor do coeficiente de correlação é interpretado.

- **r = -1**: Existe uma correlação perfeitamente negativa entre as duas variáveis, o que significa que todos os pontos estão dispostos ao longo de uma linha com inclinação negativa.
- **-1 < r < 0**: Existe uma correlação negativa entre as duas variáveis, portanto, quando uma aumenta, a outra diminui. Quanto mais próximo o valor estiver de -1, mais negativamente correlacionadas as variáveis estão.
- **r = 0**: Existe uma correlação potencialmente fraca, na verdade, a relação linear entre elas é zero, porém, isso não significa que sejam independentes, pois poderiam ter uma relação não linear.
- **0 < r < 1**: Existe uma correlação positiva entre as variáveis, e quanto m,ais próximo o valor estiver de +1, mais forte será a relação entre elas. Neste caso, uma variável tende a aumentar quando a outra também aumenta.
- **r = 1**: Existe uma correlação positiva perfeita entre as duas variáveis, isto é, possuem uma relação linear positiva.

O comportamento da correlação está graficamente representado na Figura 2.1.

Figura 2.1: Gráficos que ilustram como variáveis que se correlacionam podem se comportar dependendo de seu valor r.

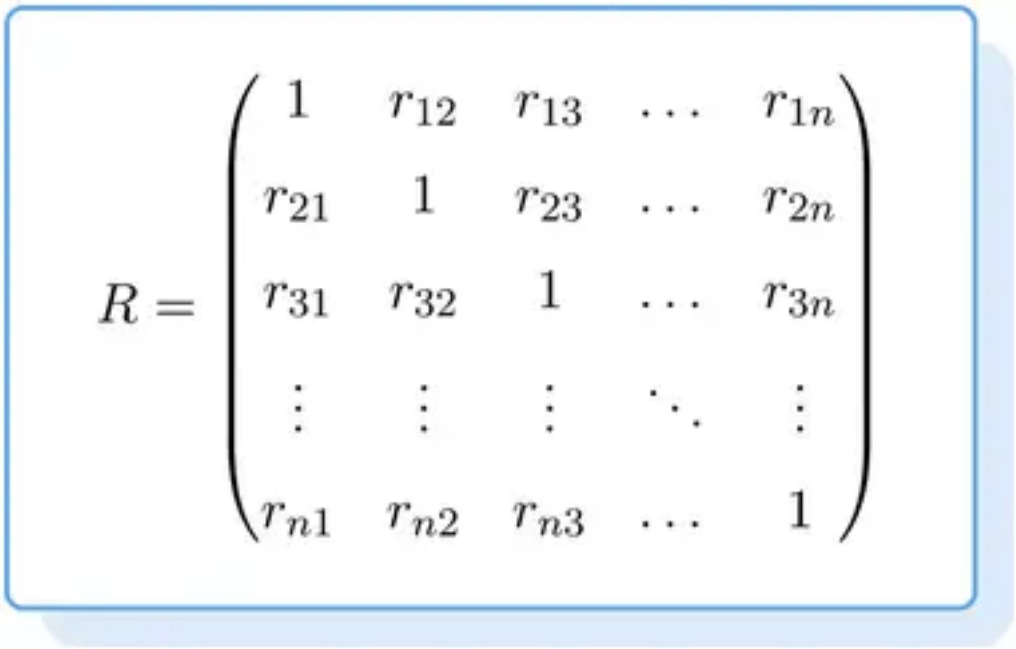


Fonte: (SILVA, 2022)

2.1.5 Matriz de correlação

Ainda como cita (ANDERSON, 2023), a matriz de correlação é uma matriz onde o coeficiente de correlação entre as variáveis i e j é representado na célula localizada na interseção da linha i e coluna j. Assim, essa matriz é quadrada, com valores de um na diagonal principal, e cada elemento fora da diagonal principal indica o coeficiente de correlação entre as variáveis correspondentes. Sendo assim, a fórmula da matriz de correlação é apresentada na Figura 2.2, onde rij é o coeficiente de correlação entre as variáveis i e j. A matriz é muito útil para resumir resultados e comparar a correlação entre variáveis ao mesmo tempo. Isso permite que rapidamente se visualize as relações que são mais fortes, bem como as mais fracas também.

Figura 2.2: Fórmula da Matriz de correlação



Fonte: Benjamim Anderson (ANDERSON, 2023).

2.1.6 Média aritmética

A média aritmética é fundamental no ramo da matemática, Estatística e da ciência experimental, muito usada no cotidiano e na vida escolar. Muitas disciplinas referenciam-na dentro de conceitos expressos em termos de média ou soma. Isso ocorre devido ao fato de que a média é capaz de proporcionar um indicador, que pode ser interpretado como uma espécie de escore / pontuação típica que, por sua vez, representa um conjunto de dados (CAZORLA, 2003).

A formulação matemática da média simples envolve somar todos os valores de uma variável e dividir pelo número total de observações. A média ponderada, uma variação da média aritmética, é usada quando há valores repetidos, diferentes pesos atribuídos aos valores, ou quando os valores estão agrupados em intervalos de classe. Nesses casos, a soma total é obtida a partir das somas parciais, multiplicando o valor da variável (ou ponto médio no caso de intervalos de classe) pela sua frequência. A média ponderada é frequentemente utilizada em escolas para avaliar o desempenho geral dos alunos durante um semestre ou ano acadêmico, especialmente quando provas ou disciplinas possuem pesos diferentes. Conforme Strauss e Bichler (1988), a média aritmética possui sete propriedades:

1. A média está situada entre os valores extremos *(mínimo $\leq$ média $\leq$ máximo)*.
2. A soma dos desvios em relação à média é zero (Σ (X_i – média) = 0).
3. A média é afetada por cada valor e por todos os valores (média = Σ X_i / n).
4. A média não necessariamente precisa coincidir com qualquer um dos valores.
5. A média pode ser uma fração que não corresponde a uma realidade física (por exemplo, o número médio de filhos por mulher pode ser 2,3).
6. O cálculo da média leva em consideração todos os valores, incluindo os nulos e os negativos, e, por fim,
7. A média é um valor representativo dos dados dos quais foi calculada. Em termos espaciais, a média é o valor mais próximo de todos os outros valores.

Ainda segundo os mesmos autores, as três primeiras propriedades dizem respeito ao aspecto estatístico da média e são fundamentais como função matemática. As três seguintes tratam do aspecto abstrato, permitindo valores não observados. A última propriedade aborda o aspecto representativo de um grupo de valores individuais, que é o aspecto central da média. (CAZORLA, 2003)

2.1.7 Desvio padrão

Na Estatística, o desvio padrão refere-se a uma medida que expressa o grau de dispersão de um conjunto de dados, isto é, indica o quanto um conjunto de dados é uniforme. Quanto mais próximo de zero for o desvio padrão, mais homogêneo é o conjunto dos dados. O desvio padrão é calculado usando a seguinte fórmula:

Figura 2.3: Fórmula do desvio padrão.

$$s = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}}$$

Fonte: Alícia Soares (SOARES, 2021).

onde:

- **S**: desvio padrão
- **Xi**: números do conjunto de dados (seja $i = 1, 2, 3, \ldots$)
- **X**: média aritmética
- **n**: quantidade de números no conjunto de dados

O desvio padrão é calculado como a raiz quadrada do resultado da soma da diferença de suas variáveis pela média aritmética, dividida pelo número de variáveis. Desta forma, quanto menor for o valor do resultado, mais homogêneo e constante esse conjunto de dados será, por outro lado, quanto maior for o resultado, mais heterogêneo o conjunto de dados será (SOARES, 2021).

2.1.8 Teste de hipótese

Os testes de hipótese são uma técnica de inferência estatística. Hipóteses nada mais são do que declarações preliminares sobre parâmetros populacionais e são testadas para verificar se são consideradas verdadeiras ou falsas (LOPES; LEINIOSKI; CECCON, 2015).

2.1.8.1 Teste t de student

O teste t de Student vem a ser um teste de hipótese que utiliza de conceitos estatísticos para rejeitar ou não uma hipótese nula quanto a estatística de teste (t) segue uma distribuição chamada t de Student. O teste t pode ser conduzido para:

- Comparar uma amostra com uma população
- Comparar duas amostras pareadas (ou independentes)

Neste trabalho estamos comparando duas amostras independentes. Este teste se aplica a planos amostrais onde se deseja comparar dois grupos distintos, cujos mesmos podem ter sido formados de duas maneiras

a) Extraiu-se uma amostra da população X e outra amostra da população Y.

b) Indivíduos da mesma população foram alocados aleatoriamente a um dos dois tratamentos que se quer estudar

Quando queremos comparar duas populações levando em conta uma variável quantitativa, é muito comum que os pesquisadores não conheçam os parâmetros de nenhuma delas, isto é, sejam desconhecidas as médias μ e também os desvios padrão σ. Assim, muitos estudos biológicos são realizados com duas amostras independentes de indivíduos, denominadas grupo experimental e grupo controle, respectivamente.

Suponha duas populações distintas, compostas por um número elevado de indivíduos. Seja X uma variável aleatória (contínua) de interesse e duas populações 1 e 2. Na população 1, a média de X é μ_1 , enquanto que na população 2, a média de X é μ_2 . Nosso objetivo é comparar as médias populacionais μ_1 e μ_2 e, por sua vez, executar o teste de hipótese:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Como (SPERANDEI, 2007), o teste t é um teste que serve para comparação de médias, onde deseja-se observar a probabilidade de ocorrência do resultado encontrado para a média calculada a partir da hipótese nula que é a de igualdade entre as médias populacionais das amostras comparadas. Em nosso trabalho a hipótese nula é a inexistência da diferença no número de ingressantes e evadidos no âmbito estatístico a fim de verificar a discrepância de ambos, principalmente durante o período de exceção entre 2020 e 2022.

2.2 Tecnologias utilizadas

2.2.1 PostgreSQL

O PostgreSQL (STONEBRAKER, 1986) é um sistema de banco de dados objeto-relacional de código aberto que utiliza e também estende a linguagem SQL. Conta com muitos recursos que auxiliam no armazenamento e dimensionamento com segurança das cargas de trabalho de dados mais complicadas. O *software* possui origem no ano de 1986 como parte de um projeto chamado POSTGRES na Universidade de Berkeley, na Califórnia.(POSTGRESQL, 2024) Foi desenvolvido em 1986 como um acompanhamento do INGRES, que é um projeto de banco de dados relacional SQL de código aberto iniciado no começo dos anos 1970, ideia de Michael Stonebraker, um professor do curso de Ciência da Computação em Berkeley (IBM, 2024).

O *software* ganhou notoriedade por sua arquitetura comprovada, confiabilidade, integridade de dados, conjunto robusto de recursos, extensibilidade e o compromisso da comunidade de código aberto que apoia o software em fornecer soluções inovadoras e de alto desempenho de forma consistente. O PostgreSQL é compatível com todos os principais sistemas operacionais, possui conformidade com ACID desde 2001 e oferece complementos poderosos, como o popular extensor de banco de dados geoespacial PostGIS (POSTGRESQL, 2024). Dado tudo isso, o PostgreSQL foi o programa escolhido para persistir os dados da análise.

2.2.2 Python

Python[1] é uma linguagem de alto nível, orientada a objeto, de tipagem dinâmica e forte. Possui uma sintaxe clara e concisa que facilita a legibilidade do código-fonte. Possui diversas estruturas de alto nível, como dicionários, listas, data/hora, entre outras e uma grande coleção de módulos prontos para uso, além de frameworks de terceiros que podem ser adicionados ao código. Por ser multiparadigma, a linguagem suporta programação modular e funcional, além de ser orientada a objeto, como já mencionado. Até mesmo os tipos básicos de Python são objetos.

A linguagem também conta com um instalador de pacotes chamado *pip*, que oferece uma poderosa capacidade de ampliar o código ao utilizar um vasto conjunto de bi-

[1] Página oficial: https://www.python.org/

bliotecas disponíveis na Internet. Estas bibliotecas são hospedadas no repositório oficial Pypi ou em repositórios não-oficiais, como o Anaconda. Além disso, a linguagem possui diversas bibliotecas e *frameworks* úteis para diversas aplicações, tais como desenvolvimento web, mineração de dados, aprendizado de máquina e aplicações orientadas a objeto. Neste trabalho, utilizamos a linguagem em *notebooks* para fácil compreensão e rapidez nos retornos da execução do código em blocos para as análises e emulação das consultas, com o auxílio de algumas bibliotecas importantes da linguagem, que serão descritas a seguir.

2.2.2.1 Pandas

O Pandas (PANDAS, 2024) é uma biblioteca muito importante e poderosa da linguagem Python, que contém estruturas de dados muito importantes para análise de dados, séries temporais e estatística. Duas destas características são muito importantes para este trabalho, e por esta razão, foi escolhida para executar o mesmo. Rápido, flexível e fácil de usar, usa muito bem da manipulação de dados de código aberto. Nesta biblioteca, utiliza-se muito uma entidade chamada *dataframe*, que é uma estrutura de dados que organiza os dados em forma de tabela com linhas e colunas como uma espécie de planilha e/ou tabela. São uma das estruturas de dados mais comuns e utilizadas na análise de dados moderna, pois trata-se de uma maneira flexível e intuitiva de armazenar e trabalhar com dados (DATABRICKS, 2024). Além disso, Pandas permite a integração com diversas outras bibliotecas, algumas delas foram utilizadas nesse trabalho, como *Seaborn*, *Matplotlib* e *Numpy*, sendo descritas a seguir.

2.2.2.2 Seaborn

A biblioteca *Seaborn* é uma biblioteca muito útil para visualização de dados, baseada em *Matplotlib* que fornece uma interface de alto nível para a exibição de gráficos estatísticos informativos e atraentes. As funções da biblioteca *seaborn* descrevem uma API declarativa e orientada a conjuntos de dados, que favorece a tradução de questões sobre os gráficos que podem respondê-las. Dados um conjunto de dados e uma especificação do gráfico que se deseja exibir, são mapeados para os valores atributos visuais como cor, tamanho e estilo. É uma biblioteca projetada para participar do ciclo de um projeto científico por completo. Produzindo gráficos completos com uma única chamada de função e poucos argumentos, o *seaborn* facilita a prototipagem rápida e a análise ex-

ploratória de dados. Além disso, ao proporcionar extensas opções de personalização e acesso aos objetos subjacentes do *matplotlib*, ele permite a criação de Figuras refinadas e de qualidade para publicação (WASKOM, 2021).

2.2.2.3 Matplotlib

A *Matplotlib* é uma biblioteca útil para criar visualizações estáticas, interativas e animadas na linguagem Python. Produz visualizações de qualidade para publicações em uma variedade de formatos impressos e ambientes interativos em diversas plataformas, podendo ser usado em scripts Python, shells Python/IPython, servidores de aplicações web e diversas ferramentas de interface gráfica do usuário (PYPI, 2024).

2.2.2.4 NumPy

NumPy é uma biblioteca de código aberto que permite a computação numérica utilizando Python. Foi criado em 2005 baseado em outras duas bibliotecas de Python Numeric e Numarray (NUMPY, 2005). NumPy é a principal biblioteca de programação de arrays para a linguagem Python e desempenha um papel crucial em pipelines de análise de pesquisa em áreas como física, química, astronomia, geociência, biologia, psicologia, ciência dos materiais, engenharia, finanças e economia. Em astronomia, por exemplo, Numpy foi fundamental na descoberta de ondas gravitacionais e até mesmo de um buraco negro. NumPy é a base sobre a qual o ecossistema científico Python é construído. Sua influência é tão grande que vários projetos, voltados para públicos com necessidades especializadas, desenvolveram suas próprias interfaces e objetos de array semelhantes ao NumPy. Devido à sua posição central no ecossistema, o NumPy cada vez mais atua como uma camada de interoperabilidade entre essas bibliotecas de computação de arrays e, juntamente com sua interface de programação de aplicativos (API), oferece uma estrutura flexível para sustentar a próxima década de análise científica e industrial (HARRIS et al., 2020).

2.2.3 PowerBI

PowerBI [2] é um serviço de análise de negócios e análise de dados desenvolvido pela Microsoft em 2015. Funciona como uma coleção de serviços de software e aplicati-

[2]Página oficial: https://www.microsoft.com/pt-br/power-platform/products/power-bi

vos que juntos transformam uma coleção de dados em informações coerentes, visualmente envolventes e interativas. Estas fontes de dados podem estar em uma simples planilha do Excel, na nuvem ou até mesmo em data warehouses híbridos locais, podendo utilizar a plataforma ou por um navegador web, ou por aplicativo, seja no Windows em Desktop ou até mesmo no celular, para Android ou iOS. Neste trabalho foi utilizada a versão web e o aplicativo para Desktop para montar as visualizações que compõem um *dashboard* interativo feito em outro trabalho anterior.

2.2.4 Controle de versionamento com Git e GitHub

O controle de versionamento é um sistema que gerencia arquivos bem como suas alterações ao longo do tempo, sendo possível até restaurar versões anteriores (CHACON; STRAUB, 2014). O controle de versionamento é feito através da ferramenta Git e gerenciado pelo GitHub. O Git foi criado em 2005 por Linus Torvalds, o mesmo criador do kernel do Linux, visando criar uma melhor performance com metas como: velocidade, suporte para desenvolvimento não-linear, distribuído e ideal para projetos grandes (BERNARDO, 2022). A principal diferença entre o Git e qualquer outro sistema de controle de versão é a forma como o Git processa seus dados. Conceitualmente, a maioria dos outros sistemas armazena informações como uma lista de mudanças baseadas em arquivos. Esses outros sistemas (CVS, Subversion, Perforce, etc.) veem as informações que armazenam como um conjunto de arquivos e as alterações feitas em cada arquivo ao longo do tempo (isso é comumente descrito como controle de versão baseado em deltas). O Git não processa ou armazena seus dados dessa maneira. Em vez disso, o Git vê seus dados como uma série de *snapshots* de um sistema de arquivos em miniatura. Com o Git, toda vez que você faz um *commit*, ou salva o estado do seu projeto, o Git basicamente 'tira uma foto' de como todos os seus arquivos estão naquele momento e armazena uma referência a esse *snapshot*. Para ser eficiente, se os arquivos não mudaram, o Git não armazena o arquivo novamente, apenas um *link* para o arquivo idêntico que já foi armazenado. O Git visualiza os dados mais como um fluxo de *snapshots*. Essa é uma distinção importante entre o Git e quase todos os outros sistemas de controle de versão, pois faz com que o Git reconsidere quase todos os aspectos do controle de versão que a maioria dos outros sistemas copiou da geração anterior. Dessa forma, o Git se assemelha mais a um sistema de arquivos em miniatura com algumas ferramentas incrivelmente poderosas construídas sobre ele, em

vez de ser apenas um sistema de controle de versão. O GitHub [3] é um sistema que guarda repositórios Git de forma online, criado em 19 de outubro de 2007 (GITHUB, 2021). O Github funciona como uma espécie de rede social, permitindo o acesso e contribuição universal de usuários aos projetos neles contidos, integrando até mesmo mais de um usuário dentro de um mesmo projeto, ou um usuário integrado a diversos outros projetos, criados por ele ou em parcerias. No presente trabalho, para melhor transporte do projeto entre as trocas de equipamentos/ambientes, o GitHub foi utilizado para hospedar o projeto como um todo.

[3] Página oficial: https://github.com/

3 TRABALHOS RELACIONADOS

Neste capítulo estão descritos os trabalhos anteriormente realizados que se relacionam de alguma forma com o presente trabalho, através de contextos como banco de dados, análise exploratória, criação de painéis de dados e até mesmo avaliação de desempenho acadêmico na UFRGS. Cada proposta é discutida em resumo e são discutidas suas diferenças e semelhanças com este trabalho.

3.1 Visualização de dados institucionais da UFRGS

A monografia em forma de Trabalho de Conclusão de Curso apresentada por Marcos André Bontatto (BONATTO, 2018) propõe e executa a criação do *Painel de Dados*, uma ferramenta *web* para expor os dados institucionais da Universidade Federal do Rio Grande do Sul, com técnicas de visualização de informações fáceis e flexíveis ao usuário e com atualização periódica dos dados. Este trabalho envolveu a elaboração de painéis de dados com filtros e gráficos assim como o presente trabalho, trazendo, além dos dados referente aos alunos também dados referente aos servidores da universidade.

3.2 Uma análise sobre reprovação no curso de Ciência da Computação na UFRGS sob a ótica dos alunos

Neste trabalho, o autor Vilmar Neto (NETO, 2021) propõe delinear um perfil demográfico e produtivo dos alunos do curso de Ciências da Computação na UFRGS, aplicando um questionário tanto para alunos quanto para egressos, obtendo os seguintes dados: 64% dos alunos são reprovados ao longo do curso, 70% dos discentes estão atrasados na graduação, e a taxa de aproveitamento por matrícula é de 84%., e, por fim, sugerindo melhorias como a atualização da formação pedagógica dos professores, a reestruturação do curso, a oferta de cursos complementares, a criação de uma opção noturna e a disponibilização de apoio pedagógico aos alunos. De um ponto de vista filosófico, o trabalho se assemelha ao presente, buscando saber como auxiliar alunos diante da reprovação e atrasos na graduação de alunos desse curso enquanto que este trabalho busca saber como a pandemia afetou esses alunos, além dos oriundos do curso de Engenharia da Computação no âmbito estatístico.

3.3 Visualização de dados quantitativos como apoio à análise de desempenho de alunos de graduação da UFRGS

O trabalho de Daniela Cavalheiro (CAVALHEIRO, 2018) visou criar um painel de dados que mostram dados dos alunos e dos cursos para uso interno da universidade como um auxílio para melhor avaliar o desempenho dos alunos e a gestão da tomada de decisões quanto a isso. Assim como o presente trabalho, a autora iniciou seu projeto a partir de dados obtidos da própria universidade, e, com isso, criou painéis de dados como no presente trabalho, porém os da autora são voltados para turmas e inclusive para alunos específicos, com classificação de dados com filtragem, gráficos do tipo *pie chart*, de barras e de linhas, esses dois últimos também incorporados a este trabalho. A proposta da autora tenta auxiliar internamente a universidade a monitorar o desempenho acadêmico dos alunos utilizando uma ferramenta que apresenta visuais que auxiliam nisto.

3.4 Um estudo sobre o impacto das disciplinas na evasão acadêmica sobre diferentes tipos de alunos no ICOMP

O artigo dos autores Alexandre Uchôa e Leandro Carvalho (UCHÔA; CARVALHO, 2024) iniciaram um estudo referente ao impacto da evasão acadêmica no Instituto de Computação da Universidade Federal do Amazonas analisando dados disponibilizados pelo próprio instituto através de um histórico escolar analítico de todos os alunos dos cursos do instituto, sendo eles Ciência da Computação e Engenharia de Software, onde eles expõem um gráfico que mostra a progressão acumulativa da evasão (categorizada através de uma variável chamada taxa de desistência acumulada) por períodos dos cursos. Posteriormente, o artigo apresenta uma tabela que mostra a mesma taxa de desistência sobre as disciplinas levando em conta ampla concorrência e cotistas por renda. O trabalho proposto pelo artigo se assemelha com o presente ao fazer uma análise exploratória com os dados referentes a evasão acadêmica em um instituto da Universidade Federal do Amazonas, estudando como se aplica até mesmo em modalidades de ingresso à mesma, embora o presente seja um pouco mais aprofundado e exploratório.

3.5 Uma proposta de extração, transformação, carga e visualização para os dados do Censo Escolar

O trabalho de conclusão de curso de Rafael Ribeiro (RIBEIRO, 2023) propõe, além de um projeto que utiliza mineração de dados em dados abertos do Censo Escolar como uma ferramenta que visa o auxílio a educação básica brasileira, contribuindo com o desenvolvimento da mesma, executando atividades como extração e transformação dos dados, desenvolvimento de um aplicativo *web*, apresentando gráficos, filtragens e muitas outras informações a partir de microdados e indicadores do Censo Escolar entre 2016 e 2022 por localidade, dependência administrativa, localidade geográfica, incluindo até mesmo mapas. O autor, assim como neste trabalho, parte da premissa movida pela educação, buscando auxiliar o ensino básico, bem como este trabalho, porém, neste, aborda-se o ensino superior como *stakeholder* do presente projeto. Também como o presente trabalho, o autor buscou expor muitas visualizações que ilustram e trazem *insights* importantes e que podem nortear futuras decisões quanto a educação básica brasileira, com ambos passando pelos procedimentos padrões ao se trabalhar com banco de dados, como extração e transformação de dados, entre outras etapas sumamente importantes para um projeto orientado à dados.

3.6 Uma ferramenta web para a automatização de relatórios da Sociedade Brasileira de Computação sobre dados referentes ao ensino nacional de tecnologia

O trabalho de Pedro Perrone (PERRONE, 2023) propôs a criação de uma ferramenta que automatizasse a criação de relatórios com dados do Ensino Superior de tecnologia no Brasil dentro da Sociedade Brasileira de Computação, motivado pela constatação de que a SBC executava esse trabalho manualmente, proporcionando assim redução do custo operacional deste processo e uma maior flexibilidade na criação desses relatórios. O autor criou uma aplicação *web* utilizando linguagens como Elixir, Javascript e PostgreSQL que, ao receber planilhas como entrada, alimenta o banco de dados e os exibe conforme o usuário desejar. Assim como o presente trabalho, o projeto do autor passou por modelagem de dados, processamento das planilhas oriundas do Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) de cursos e instituições de ensino para esses relatórios com muitas colunas, onde o autor apenas seleciona para o projeto quatro delas: identificação da instituição de ensino, nome da instituição, sigla do estado onde a

instituição se baseia e nome da região administrativa na qual a instituição de ensino se encontra. Implementando consultas, filtros e gráficos, o trabalho propõe uma ferramenta importante de bastante auxílio ao SBC, principalmente na mobilidade e na redução do custo operacional na elaboração de um relatório pela mesma.

3.7 Tabela comparativa entre os trabalhos

Nesta seção estão descritas as características dos trabalhos em paralelo juntamente com as do presente trabalho a fim de estabelecer uma comparação mais direta, relacionando técnicas, tipo de projeto e origem das bases de dados.

Tabela 3.1: Tabela comparativa entre os trabalhos relacionados a este trabalho

Trabalhos por autor	**Tipo de projeto**	**Origem dos dados**	**Técnicas de avaliação**
(BONATTO, 2018)	Ferramenta web	Dados internos da UFRGS	Visualização de dados; análise exploratória estatística
(NETO, 2021)	Análise/estudo	Pesquisas	Análise exploratória estatística
(CAVALHEIRO, 2018)	Ferramenta web	Dados internos privados da UFRGS	Visualização de dados; análise exploratória estatística
(UCHÔA; CARVALHO, 2024)	Análise/estudo	Dados do ICOMP	Análise exploratória
(RIBEIRO, 2023)	Ferramenta web	Dados institucionais do INEP	Análise exploratória
(PERRONE, 2023)	Ferramenta web	Dados do SBC	Visualização de dados; análise exploratória estatística
Este trabalho	Análise/estudo	Dados abertos da UFRGS	Visualização de dados; análise exploratória estatística; experimentos

4 METODOLOGIA

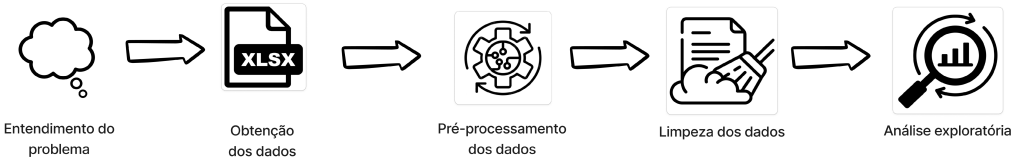
Neste capítulo, está detalhado a metodologia do trabalho de análise exploratória e experimentos referente aos dados quantitativos disponibilizados pela UFRGS. Desde a implementação do projeto de banco de dados (extração e obtenção dos dados, limpeza de dados e modelagem conceitual) até a execução dos testes e experimentos, fornecendo informações detalhadas e essenciais para quem desejar replicar e/ou validar os experimentos. Além disso, está também detalhada a metodologia utilizada na elaboração dos experimentos, incluindo uma análise crítica sobre as razões e as técnicas usadas em cada etapa da implementação (análogo para os resultados dos procedimentos) dos procedimentos e objetivos desta investigação.

A Seção 4.1 apresenta uma visão geral das etapas de pesquisa através de um fluxograma, que expõe todo o processo do início ao fim, que começa com o entendimento do problema e termina com a análise exploratória. Na sequência cada etapa é explicada e comentada, argumentando as técnicas e experimentos utilizados. Na Seção 4.2 é explicado o entendimento do problema, na seção 4.3 é descrito como foi realizada a obtenção dos dados brutos. Na seção 4.4 é discutido o pré-processamento dos dados, na seção 4.5 foi detalhado o processo de limpeza dos dados, e, por fim, na seção 4.6 são apresentadas e discutidas as visualizações empregadas para a pesquisa.

4.1 Visão geral

A Figura 4.1 ilustra o fluxograma da metodologia utilizada para este trabalho. A metodologia empregada nesse trabalho foi baseada no fluxo normalmente empregado em um projeto de banco de dados, que finaliza com uma análise exploratória de dados para responder as questões referentes ao aumento de discentes no retorno pós-ERE. De início, deve-se entender o que se está buscando e então, obter os dados brutos que passam pela limpeza de dados, criação da modelagem conceitual e por fim a análise exploratória dos dados, na qual foram feitas as consultas, gráficos e demais experimentos. Como neste trabalho estão sendo analisados os cursos e anos em que mais houveram demandas de alunos da UFRGS, optou-se por executar análises pertinentes a dados quantitativos, já que sua grande maioria são dados numéricos.

Figura 4.1: Fluxograma da metodologia do trabalho



O projeto está todo contido em um repositório no GitHub, sistema de controle de versionamento escolhido para armazenar o projeto, favorecendo a troca fácil e organizada de diferentes ambientes para o desenvolvimento do trabalho.

4.2 Entendimento do problema

A motivação deste trabalho se dá ao fato de ter sido observado um grande retorno em massa, tanto de veteranos como de calouros logo após o período de exceção que iniciou em 19 de agosto de 2020 (BEHAR, 2020), na modalidade 100% remota, cujo retorno foi feito de forma gradual, começando com apenas algumas cadeiras fazendo aulas presenciais e/ou provas presenciais, até, por fim, retornar totalmente em 13 de junho de 2022, mediante comprovação vacinal contra o vírus Sars-Covid-19, que causou a pandemia global em 2020. Sendo assim, busca-se entender o problema a fim de formular um contexto e ser capaz de trabalhar com as fontes de dados disponíveis para responder essa questão que permeia o trabalho. Neste trabalho, a proposta é verificar quais os cursos que estão demandando mais alunos dentro da Universidade Federal do Rio Grande do Sul, principalmente, verificar se são de fato Ciência da Computação e/ou Engenharia da Computação os cursos que estão entre os preferidos depois do retorno ao modo presencial, no qual podemos elencar possíveis interessados, como o Centro de Processamento de Dados da UFRGS.

4.3 Obtenção dos dados

Depois de bem definido o problema e a pesquisa a ser feita, buscamos os dados nos quais faremos a análise. Por vias de facilidade e acessibilidade, utilizamos os dados disponíveis no site Dados UFRGS. Nessa seção, também analisamos o dicionário de dados o qual geralmente está disponível no mesmo site para *download* juntamente com os dados [1]. Os dados são referentes ao quantitativo de alunos na UFRGS entre os anos 2010

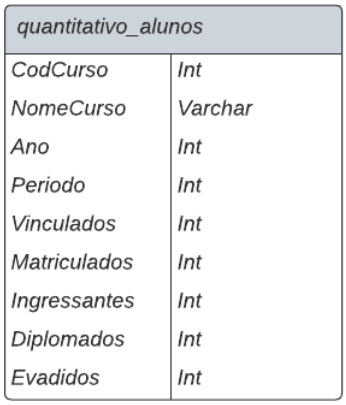
[1] Disponível em: https://dados.ufrgs.br/dataset/quantitativo-de-alunos-de-graduacao

Tabela 4.1: Dicionário dos dados brutos.

Campo	**Tipo**	**Descrição**
CodCurso	Inteiro	Código do curso
NomeCurso	Texto	Nome do curso
Ano	Inteiro	Ano do período letivo de referência
Periodo	Inteiro	Semestre do período letivo de referência
Vinculados	Inteiro	Número de alunos que possuem vínculo ativo no curso de graduação no período letivo de referência. O aluno pode estar matriculado, em trancamento, em licença, etc.
Matriculados	Inteiro	Número de alunos que estão matriculados em pelo menos uma atividade de ensino do curso de graduação, inclusive em mobilidade acadêmica, no período letivo de referência
Ingressantes	Inteiro	Número de alunos ingressantes em processos seletivos (vestibular, transferência interna, ingresso diplomado, etc.) que efetivaram matrícula por curso no período letivo de referência
Diplomados	Inteiro	Número de alunos que, após a conclusão de todos os créditos acadêmicos, apresentaram registro de diplomação por curso no período letivo de referência
Evadidos	Inteiro	Número de alunos que se desligaram por abandono, desistência de vaga, falecimento, transferência interna ou outras formas de saída por curso de graduação no período letivo de referência

e 2023. A base de dados foi obtida no formato xlsx, no qual cada linha fornece dados quantitativos sobre os alunos da universidade em cada ano e periodo descritos no início de cada linha. Apesar de nem todos os atributos serem usados para a análise, por fins de completude, estão expressos aqui todos os atributos dos dados (Tabela 4.1).

Figura 4.2: Diagrama ER da base de dados do trabalho



Para a modelagem conceitual, este trabalho consiste de um modelo Entidade-Relacionamento bastante simples (vide Tabela 4.2), pois possui apenas uma entidade, a *quantitativo_alunos*, como veio sendo tratada no PostgreSQL, SGBD onde está armazenado o *dataset* trabalhado, não se relacionando com nenhuma outra. Possuindo a maioria de seus dados numéricos, apenas *NomeCurso* que é do tipo *Varchar* que descreve o nome dos cursos oferecidos pela universidade.

4.4 Pré-processamento

A etapa de pré-processamento é fundamental para preparar e organizar os dados de forma adequada para ser armazenados e usados de maneira eficiente nas análises e experimentos. Neste trabalho, o pré-processamento começa com a importação dos dados em *dataframes*. Para melhor manipulação dos dados, o arquivo do *dataset* foi convertido de xslx para csv antes da importação. Esta etapa foi feita majoritariamente utilizando a linguagem de programação Python utilizando a biblioteca Pandas. Todos os dados foram importados utilizando *dataframes* para a execução dos gráficos, visualizações e experimentos.

4.5 Limpeza

A etapa de limpeza foi necessária devido ao fato de que possuíamos dados parciais referentes ao ano de 2023, isto é, enquanto os outros anos possuíam dados dos períodos 1 e 2, 2023 continha dados referentes ao seu ano somente para o período 1, certamente devido ao dicionário (este mesmo concedido pelo próprio site) no qual consta a data em que foi atualizado pela ultima vez o *dataset* era 05 de outubro de 2022. Logo, optamos por remover todos os dados referentes ao ano de 2023.

4.6 Análise exploratória

Nesta etapa é onde se observa o comportamento das variáveis, compreender e resumir as principais características do conjunto de dados, identificar padrões, detectar anomalias, formular hipóteses e preparar os dados para análises mais avançadas. Para uma análise efetiva foram feitos vários gráficos que verificam a relação entre as variáveis que mais são relevantes para a pesquisa como o número de ingressantes relacionados com o de evadidos, que por sua vez podem ser relacionados com o número dos já vinculados a universidade ou o número de diplomados. Para executar as análises, os dados depois de preparados foram importados em um *notebook Python*, utilizando a biblioteca Pandas para manipular os dataframes, a biblioteca Seaborn e Matplotlib para a criação dos gráficos e a bibliteca scipy.stats para os experimentos estatísticos que serão explicados posteriormente.

5 EXPERIMENTOS E ANÁLISE DE DADOS

Neste capítulo estão apresentados como foi feita a análise dos resultados através de consultas ao banco de dados, experiências e também visualizações de dados que ilustrem os *insights* obtidos durante a análise exploratória. Esta análise foi dividida em duas grandes partes: análise quantitativa, que ilustra consultas em banco de dados e análise exploratória, com os gráficos, *dashboard* e experimentos. Todos os resultados foram cuidadosamente analisados e discutidos.

5.1 Visão geral

A questão inicial dessa análise e que motivou o trabalho foi avaliar se as demandas de discentes - principalmente dos cursos de Ciência da Computação e Engenharia da Computação - realmente se mostraram expressivas, especificamente no retorno do Ensino Remoto Emergencial através de uma abordagem *top-bottom*, começando pelos cursos mais evadidos, analisando as taxas em relação aos vinculados e, posteriormente analisa-se os anos que mostraram mais ingressantes e evadidos por ano e período em todos os cursos até limitar as análises para o período de exceção entre 2020 e 2022. Todas as consultas foram feitas no PostgreSQL, onde os dados estão armazenados. Por fim, foi feita uma análise estatística através do teste t de Student para verificar se no período de exceção da UFRGS houve diferença estatisticamente significativa entre o numero de ingressantes e evadidos, utilizando a linguagem Python. A análise está dividida em dois grandes aspectos: quantitativa e exploratória, nas seções 5.2 e 5.3, respectivamente.

5.2 Análise Quantitativa

A análise quantitativa é uma metodologia muito empregada para coleta e interpretação dos dados numéricos para obter informações e *insights* relevantes (ESCOLA, 2024). É uma abordagem fundamental para a tomada de decisões baseadas em evidências, pesquisas científicas e estudos gerais. O processamento de consultas SQL feitas para a análise e respondendo várias perguntas pertinente à pesquisa estão descritas nas próximas seções. A análise está detalhada em oito questionamentos principais:

5.2.1 Dez cursos com maior evasão da UFRGS

O objetivo da Consulta (5.1) é ilustrar os cursos com maior evasão de alunos no período de 2010 a 2022, período compreendido pelo *dataset* a fim de averiguar se os cursos de Ciência e/ou Engenharia da Computação apareciam na consulta.

Porém, a consulta retornou que os cursos mais evadidos nos dados compreendidos pelo *dataset* foram os cursos de Ciências Sociais, Letras, Matemática, Física, Administração (estes dois últimos curiosamente com o mesmo valor de discentes evadidos), entre outros.

O retorno da consulta não mostrou nem Ciência da Computação nem Engenharia da Computação ocupando posição alguma nesse *ranking*. Vemos isso ilustrado na Figura 5.2.

Figura 5.1: Consulta que retorna os cursos mais evadidos da UFRGS.

```sql
SELECT nomecurso, SUM(evadidos) AS total_evadidos
FROM quantitativo_alunos
GROUP BY nomecurso
ORDER BY total_evadidos DESC
LIMIT 10;
```

Figura 5.2: Resultado da consulta em 5.1.

	nomecurso	total_evadidos
1	CIÊNCIAS SOCIAIS	1.889
2	LETRAS	1.746
3	MATEMÁTICA	1.541
4	FÍSICA	1.483
5	ADMINISTRAÇÃO	1.483
6	QUÍMICA	1.304
7	HISTÓRIA	1.199
8	EDUCAÇÃO FÍSICA	1.170
9	CIÊNCIAS ECONÔMICAS	1.043
10	FARMÁCIA	971

Figura 5.3: Consulta em SQL que retorna os dez cursos com as maiores taxas de evasão para vinculados na UFRGS.

```sql
SELECT nomecurso, ano, periodo, ((evadidos*100/vinculados)) AS pct_evadidos
FROM quantitativo_alunos
WHERE ((evadidos*100/vinculados)) < 100
ORDER BY pct_evadidos DESC
LIMIT 10;
```

Figura 5.4: Resultado da consulta em 5.4.

	nomecurso	ano	periodo	pct_evadidos
1	MÚSICA - ENSINO A DISTÂNCIA - EAD	2.012	2	98
2	EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - POA	2.022	2	75
3	CIÊNCIAS BIOLÓGICAS - BIOLOGIA MARINHA	2.022	2	66
4	CIÊNCIAS BIOLÓGICAS - ENSINO A DISTANCIA - EAD	2.010	2	57
5	MATEMÁTICA - ENSINO A DISTÂNCIA - EAD	2.011	1	53
6	CIÊNCIAS DA NATUREZA PARA OS ANOS FINAIS DO ENSINO FUNDAMENTAL - EAD	2.018	2	49
7	GEOGRAFIA - ENSINO A DISTÂNCIA - EAD	2.019	2	43
8	CIÊNCIAS SOCIAIS - ENSINO A DISTÂNCIA - EAD	2.022	1	42
9	CIÊNCIAS SOCIAIS - ENSINO A DISTÂNCIA - EAD	2.019	1	40
10	BACHARELADO EM DESENVOLVIMENTO RURAL - EAD	2.019	1	40

5.2.2 Dez cursos com maior taxa de alunos evadidos pra vinculados por ano e período na UFRGS

Nesta consulta, busca-se saber quais cursos foram os mais evadidos em relação ao número de vinculados e se Ciência da Computação e/ou Engenharia da Computação estão nesse *ranking*.

Na Figura 5.3, está ilustrada a consulta que retorna os dez cursos com a porcentagem de taxa de evadidos por vinculados por curso em ordem decrescente. Essa consulta é executada visando a taxa devido ao fato de que existem cursos com muitos mais alunos que outros, e, naturalmente, eles teriam uma evasão maior, sendo desproporcional. Já na Figura 5.4, mostra-se os cursos que são retornados na execução dessa consulta.

Como visualizado na Figura 5.4, os cursos com maior taxa de evasão para vínculo são Música EAD, Computação e Robótica Educativa EAD, Educação do Campo, Matemática - Ensino a distância - EAD, e mais, não constando, infelizmente, os cursos de Ciência da Computação nem Engenharia da Computação. Nota-se uma maior preferência por cursos EAD e maior ingresso ao período 2 (6 dos 10 cursos).

5.2.3 Dez cursos com maior taxa de ingressantes para vinculados na UFRGS

Já nesta consulta, busca-se ilustrar os cursos com maior taxa de ingressantes para vinculados na Universidade no período geral de 2010 a 2022 para verificar se, por outro lado, não houve um volume expressivo de discentes ingressantes de um modo geral nos cursos de Ciência da Computação e Engenharia da Computação.

Aqui são apresentados a consulta (Figura 5.5) e o resultado da mesma (Figura

Figura 5.5: Consulta em SQL que retorna os dez cursos com maior taxa de ingressantes para evadidos na UFRGS

```
SELECT nomecurso , ano, periodo, MAX((ingressantes*100/vinculados)) AS max_pct_ingressantes
FROM quantitativo_alunos qda
WHERE (ingressantes*100/vinculados) < 100
GROUP BY nomecurso, ano, periodo
ORDER BY max_pct_ingressantes desc
LIMIT 10;
```

Figura 5.6: Resultado da consulta em 5.5.

	nomecurso	ano	periodo	max_pct_ingressantes
1	BACHARELADO EM DESENVOLVIMENTO RURAL - EAD	2.018	1	97
2	ENGENHARIA DE SERVIÇOS - LITORAL NORTE	2.018	1	85
3	DESENVOLVIMENTO REGIONAL - LITORAL NORTE	2.018	1	75
4	EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - POA	2.015	2	70
5	EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - LITORAL	2.015	2	58
6	ENGENHARIA DE GESTÃO DE ENERGIA - LITORAL NORTE	2.018	1	57
7	EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - POA	2.016	2	57
8	ENGENHARIA DE ENERGIA	2.011	1	55
9	PUBLICIDADE E PROPAGANDA	2.019	2	54
10	ADMINISTRAÇÃO PÚBLICA E SOCIAL	2.015	2	53

5.6), que mostra os cinco cursos com maiores taxas de alunos ingressantes por vinculados, bem como ano e período, que mostrará com mais precisão os cursos que mais tem alunos ingressando, pois existem cursos com muitos mais alunos que outros, e, naturalmente eles teriam uma ingressão maior proporcional.

Como mostrado na Figura 5.6, os cursos que mais obtiveram taxa de ingressantes para vinculados foram Bacharelado em Desenvolvimento Rural (EAD), Engenharia de Serviços (Litoral Norte), Desenvolvimento Regional (Litoral Norte), Educação do Campo - Ciências da Natureza (POA), Educação do Campo - Ciência da Natureza (Litoral) entre outros, mas Ciência da Computação e Engenharia da Computação ficaram fora desse *ranking* de um modo geral.

5.2.4 Anos com maior taxa de ingressantes para vinculados em todos os cursos da UFRGS

Nesta consulta, busca-se saber os anos em que as maiores taxas de ingressantes aparece em todos os cursos da UFRGS a fim de averiguar se os anos 2020, 2021, e/ou 2022 aparecem nessa consulta aplicada aos dados gerais.

Como apresentado na Figura 5.8, os anos que mais tem taxa de alunos ingressantes para vinculados são de 2010 até 2020, dos anos que interessa para a pesquisa, apenas 2020 aparece na consulta feita na Figura 5.7.

Figura 5.7: Consulta em SQL que retorna os anos que contém as maiores taxas de ingressantes para vinculados na UFRGS.

```sql
SELECT ano, periodo, MAX((ingressantes*100/vinculados)) AS max_pct_ingressantes
FROM quantitativo_alunos qda
WHERE (ingressantes*100/vinculados) < 100
GROUP BY ano, periodo
ORDER BY max_pct_ingressantes desc
LIMIT 10;
```

Figura 5.8: Retorno da consulta em 5.7.

	123 ano	123 periodo	123 max_pct_ingressantes
1	2.018	1	97
2	2.015	2	70
3	2.016	2	57
4	2.011	1	55
5	2.019	2	54
6	2.013	2	52
7	2.020	1	50
8	2.010	1	50
9	2.010	2	50
10	2.012	2	48

5.2.5 Anos com maior taxa de evadidos para vinculados em todos os cursos da UFRGS

Nesta consulta, busca-se saber os anos em que mais evasão ocorreu na UFRGS em todos os cursos para averiguar se ocorreu evasão expressiva em relação ao número de vinculados entre os anos 2020 e 2022.

Nota-se que o ano 2022 figura entre os anos mais evadidos, no segundo lugar inclusive (Figura 5.10). Como se observa, o ano em que houve mais evasão foi no ano de

2012 e em segundo lugar, o ano de 2022, como já foi dito. Felizmente, um dos anos mais importantes para a pesquisa.

Figura 5.9: Consulta em SQL que retorna os anos com maior taxa de evadidos para vinculados na UFRGS.

```sql
SELECT ano, periodo, MAX((evadidos*100/vinculados)) AS max_pct_evadidos
FROM quantitativo_alunos qda
WHERE (evadidos*100/vinculados) < 100
GROUP BY ano, periodo
ORDER BY max_pct_evadidos desc
LIMIT 10;
```

Figura 5.10: Retorno da consulta feita em 5.9.

	123 ano	123 periodo	123 max_pct_evadidos
1	2.012	2	98
2	2.022	2	75
3	2.010	2	57
4	2.011	1	53
5	2.018	2	49
6	2.019	2	43
7	2.022	1	42
8	2.019	1	40
9	2.010	1	36
10	2.015	1	30

5.2.6 Cursos com maior taxa de ingressantes para vinculados em todos os cursos da UFRGS entre 2020 e 2022

Nesta consulta (Figura 5.11) procura-se saber quais os cursos que mais tiveram ingressantes entre os anos de 2020 e 2022, anos em que houve a pandemia do coronavírus bem como o período de exceção na universidade a fim de verificar se os cursos de Ciência da Computação e Engenharia da Computação Figuram entre esses cursos. Como visualizado, porém, esses cursos não aparecem na consulta, alguns dos cursos mais ingressados são Desenvolvimento Regional, Ciências Biológicas, Engenharia de Gestão de Energia, Engenharia de Serviços, Educação do Campo e mais.

Figura 5.11: Consulta em SQL que retorna os cursos com maior taxa de ingressantes para vinculados na UFRGS entre 2020 e 2022.

```sql
SELECT nomecurso, ano, periodo, ((ingressantes*100/vinculados)) AS pct_ingressantes
FROM quantitativo_alunos
WHERE ((ingressantes*100/vinculados)) < 100
AND ano BETWEEN 2020 AND 2022
ORDER BY pct_ingressantes DESC
LIMIT 10;
```

Figura 5.12: Retorno da consulta feita em 5.11.

	nomecurso	ano	periodo	pct_ingressantes
1	DESENVOLVIMENTO REGIONAL - LITORAL NORTE	2.020	1	50
2	CIÊNCIAS BIOLÓGICAS - LICENCIATURA	2.020	1	48
3	ENGENHARIA DE GESTÃO DE ENERGIA - LITORAL NORTE	2.022	2	42
4	ENGENHARIA DE SERVIÇOS - LITORAL NORTE	2.020	2	40
5	EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - POA	2.021	1	39
6	EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - LITORAL	2.020	2	39
7	BIOTECNOLOGIA	2.021	1	30
8	BIOTECNOLOGIA	2.020	1	30
9	PUBLICIDADE E PROPAGANDA	2.020	1	30
10	GEOGRAFIA - LITORAL NORTE	2.022	1	29

5.2.7 Cursos com maior taxa de evadidos para vinculados em todos os cursos da UFRGS entre 2020 e 2022

Na consulta seguinte, verifica-se quais cursos mais tiveram evadidos entre os anos 2020 e 2022 a verificar se os cursos de Ciência da Computação e Engenharia da Computação aparecem nessa consulta.

Nenhum dos dois cursos aparecem no resultado da consulta na Figura 5.14. Aqui os cursos que se apresentam são Educação do Campo, Ciências Biológicas, Ciências Sociais, Ciências Biológicas, Engenharia Metalúrgica e mais.

Figura 5.13: Consulta em SQL que retorna os cursos com maior taxa de evadidos para vinculados na UFRGS entre 2020 e 2022.

```sql
SELECT nomecurso, ano, periodo, ((evadidos*100/vinculados)) AS pct_evadidos
FROM quantitativo_alunos
WHERE ((evadidos*100/vinculados)) < 100
AND ano BETWEEN 2020 AND 2022
ORDER BY pct_evadidos DESC
LIMIT 10;
```

Figura 5.14: Retorno da consulta em 5.13.

	nomecurso	ano	periodo	pct_evadidos
1	EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - POA	2.022	2	75
2	CIÊNCIAS BIOLÓGICAS - BIOLOGIA MARINHA	2.022	2	66
3	CIÊNCIAS SOCIAIS - ENSINO A DISTÂNCIA - EAD	2.022	1	42
4	CIÊNCIAS BIOLÓGICAS - CECLIMAR	2.022	2	35
5	ENGENHARIA METALÚRGICA	2.022	2	29
6	INTERDISCIPLINAR EM CIÊNCIA E TECNOLOGIA - LITORAL NORTE	2.022	2	28
7	ENGENHARIA HÍDRICA	2.022	2	25
8	GEOGRAFIA - ENSINO A DISTÂNCIA - EAD	2.022	1	24
9	SAÚDE COLETIVA	2.022	2	24
10	EDUCAÇÃO FÍSICA	2.022	2	22

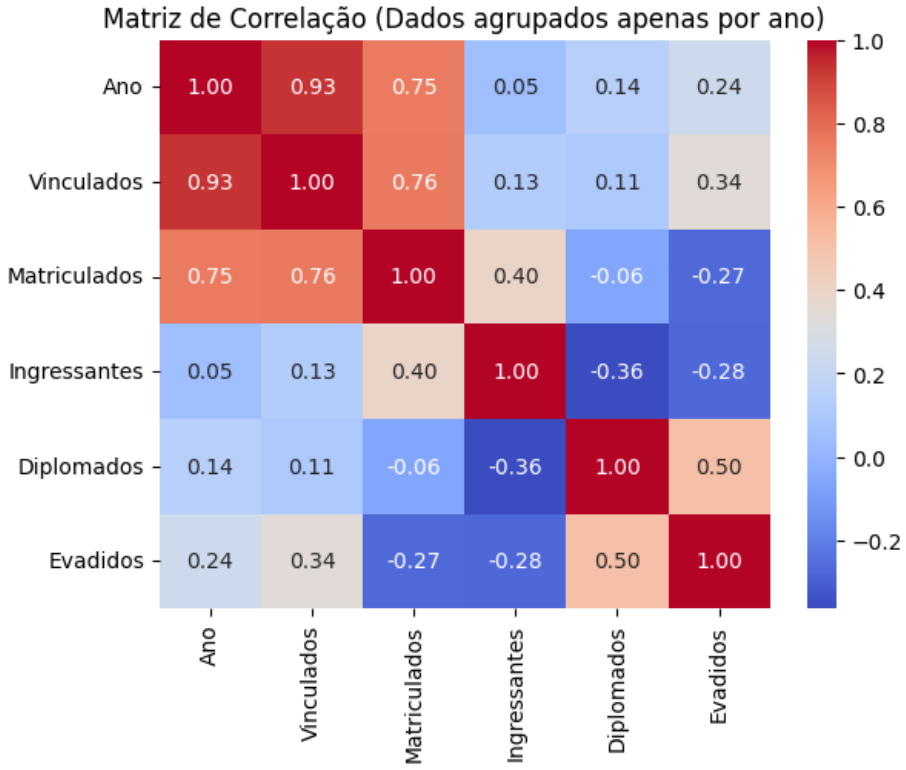
5.3 Análise exploratória

Nesta seção, fazemos uma análise *top-bottom*, isto é, começando a análise pelos dados mais gerais de todos os cursos oferecidos pela universidade e em todos os anos compreendidos pelo *dataset*. Posteriormente, executamos um 'afunilamento' dos dados a fim de verificar os dados que competem aos cursos de Ciência da Computação e Engenharia da Computação e, também, por entre o período do Ensino Remoto Emergencial, que compreende os anos 2020 e 2022, onde também fazemos o teste t de Student para os dados. Foram feitos diversos gráficos e os mesmos estão explicados na sequência.

Os dados estão agrupados por curso, código do curso, ano e período no *dataset* importado. Cria-se um *dataframe* que agrupa apenas os dados de interesse (ingressantes, evadidos, vinculados, diplomados e matriculados) todos agrupados por ano para melhor organização Primeiramente, parece promissor observar como os dados quantitativos se correlacionam entre si. Cria-se duas matrizes de correlação: uma para o *dataset* que leva os dados brutos por completo desde a importação e outra baseada no *dataframe* que leva os dados agrupados por ano.

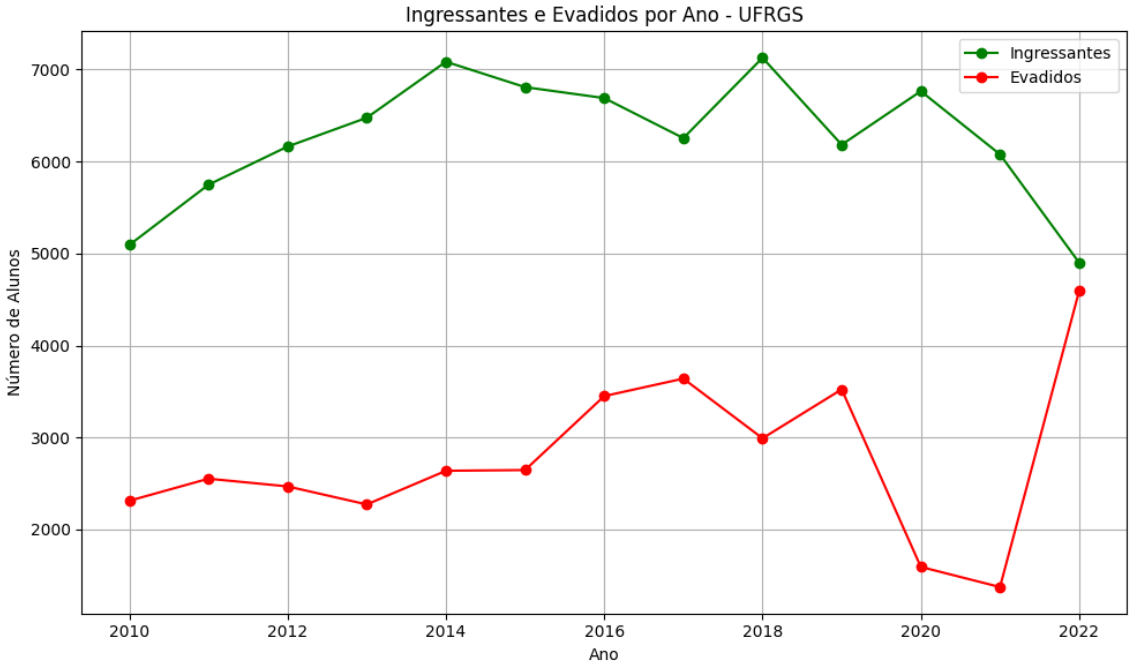
Nos dados agrupados por ano na matriz de correlação, podemos observar na Figura 5.15 que os números de matriculados estão fortemente correlacionados com o número de vinculados, indicando que, lógicamente, ao passo que aumentam as matriculas na universidade, aumentam os vinculos. O número de evadidos, por sua vez, tem uma correlação muito mais fraca com o número de matriculados, o que é muito bom. O numero de diplomados se correlaciona medianamente com o número de evadidos, o que pode explicar que o curso ao passo que diploma muitos discentes, faz com que os discentes desistam dos cursos também de certa forma.

Figura 5.15: Matriz de correlação dos dados agrupados por ano



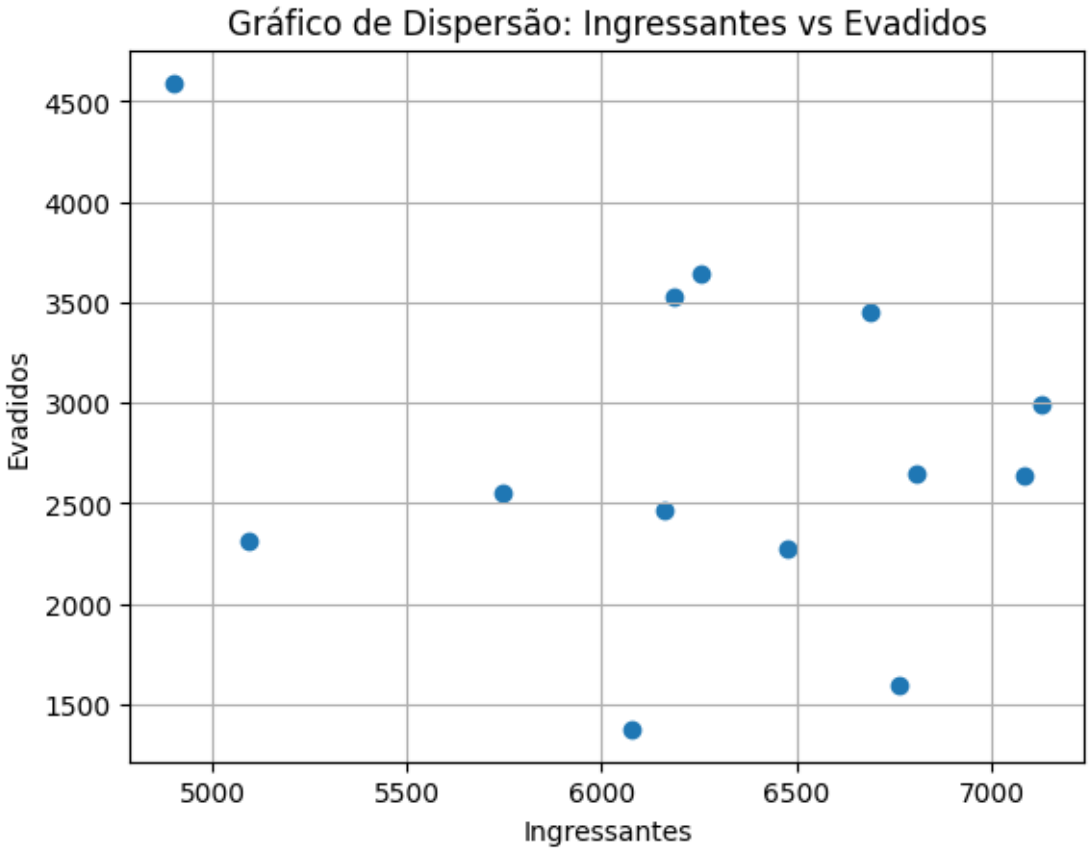
Atentar a correlação substancialmente fraca (-0.28) entre ingressantes e evadidos foi uma ideia inicial adotada, pois reflete diretamente as demandas de discentes na universidade, e, por conseguinte, os dados que mais sofrem alterações ao passar dos anos e períodos. Naturalmente, observa-se que o número de ingressantes não acompanha necessariamente o número de evadidos, mas podemos analisar como eles se comportam com o passar dos anos. No gráfico da Figura 5.16, observamos que o pico de ingressantes foi no ano de 2018 somente (7128 alunos ingressantes), porém, em 2020, o primeiro ano do período de exceção na universidade, também obtivemos um valor considerável de alunos ingressantes (6765). Por outro lado, para os evadidos, o maior valor ocorre em 2022 (4591 alunos), último ano do período de exceção.

Figura 5.16: Gráfico que ilustra o paralelo entre ingressantes e evadidos na universidade ao longo dos anos.



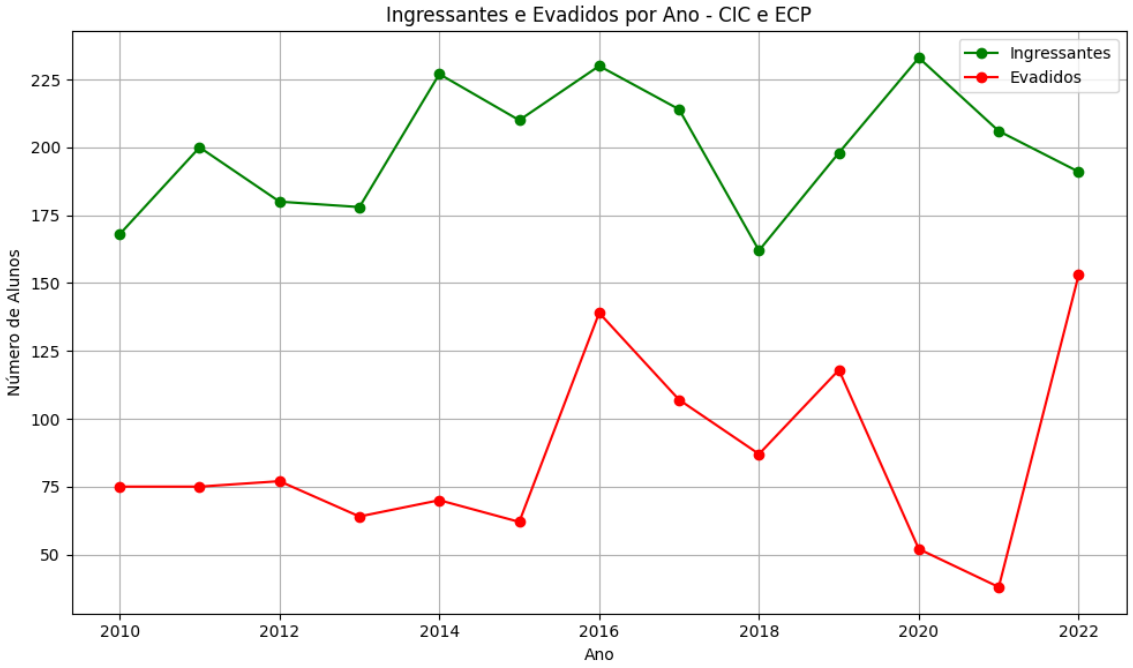
A correlação entre essas duas variáveis é uma correlação negativa débil (-0.27678), indicando que as duas variáveis crescem inversamente uma a outra, o que é útil para nossa pesquisa. De acordo com o gráfico de dispersão dos dados de ingressantes e evadidos, os dados são realmente muito dispersos no plano, mostrando que os dados não se acompanham de forma mútua em nenhum momento.

Figura 5.17: Gráfico de dispersão entre ingressantes e evadidos na universidade



Nos limitemos agora ao escopo que compreende aos cursos de Ciência da Computação e Engenharia da Computação, cursos oferecidos pelo Instituto de Informática onde foi notada a alta população. Filtramos os dados em um novo *dataframe* que agrupam apenas os dados que mencionam 'CIÊNCIA DA COMPUTAÇÃO' e 'ENGENHARIA DE COMPUTAÇÃO' e posteriormente agrupamos eles por ano como feito com os dados referentes a todos os cursos. Feito isso, traçamos o paralelo novamente entre ingressantes e evadidos na universidade apenas nos cursos de Computação.

Figura 5.18: Gráfico que ilustra o paralelo entre ingressantes e evadidos na universidade nos cursos de CIC e ECP ao longo dos anos.



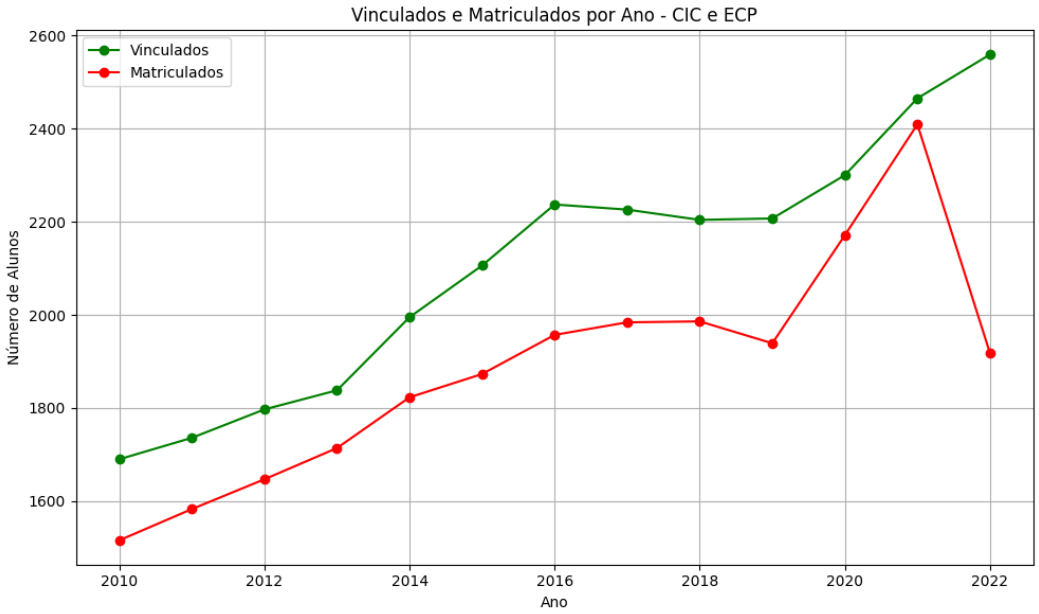
Os dados do gráfico ilustrado na Figura 5.18 se comportam de forma bastante semelhante ao 5.16, nota-se um elevado número de evadidos no ano de 2016 (139 alunos) e no ano de 2022 também temos a maior evasão nos cursos selecionados (153). Já para os ingressantes, os anos de 2014, 2016 e 2020 foram o que mais tiveram ingressados nos cursos da Computação (217,230 e 233 novos alunos, respectivamente). O aumento no número de alunos ingressantes no início do período de exceção pode explicar o aumento dos discentes no retorno ao modo presencial que já estavam ingressados na universidade.

Nos anos de 2021 e 2022 o número de diplomados foi bastante expressivo, o que pode indicar que o período de aulas remotas foi eficiente para adiantamento de cadeiras e catalisou de certa forma a aprovação dos alunos, mas também indo contra a hipótese inicial de que estava realmente havendo um aumento significativo no número de calouros e veteranos depois do ensino remoto. O comportamento do gráfico de diplomados está expresso na Figura 5.19.

Figura 5.19: Gráfico de diplomados nos cursos de CIC e ECP

Visualizando esse paralelo, percebe-se que o aumento de vinculados é mais próximo da resposta da pesquisa, mostrando que não necessariamente alunos matriculados na universidade produziram volume pós-ERE e sim, somente discentes que possuem algum vínculo com a universidade (conforme dicionário na Tabela 4.1).

Figura 5.20: Gráfico correspondente aos vinculados e matriculados por ano nos cursos CIC e ECP



Quanto à matriz de correlação limitada à CIC e ECP (Figura 5.21), logicamente matriculados e vinculados possuem a maior correlação (0.86) (desconsiderando Ano e Vinculados, preservados pela totalidade do *dataset*), logo após, vemos que diplomados e vinculados possuem correlação média (0.55), mostrando que ambas variáveis crescem de forma semelhante - alunos são diplomados ao passo que alunos obtêm vínculo junto a universidade, o que é muito positivo estatisticamente. Ingressantes e matriculados também possuem correlação média (0.45), positiva no que diz ao crescimento de ambas as variáveis. Quanto às correlações fracas, observamos matriculados x evadidos (-0.11) e ingressantes x evadidos (0.01), esta apesar de positiva, se mostra uma correlação muito baixa, o que também é positivo, pois não indicam que ao passo que alunos entram, alunos também saem necessariamente.

Figura 5.21: Matriz de correlação dos dados referentes aos cursos de CIC e ECP

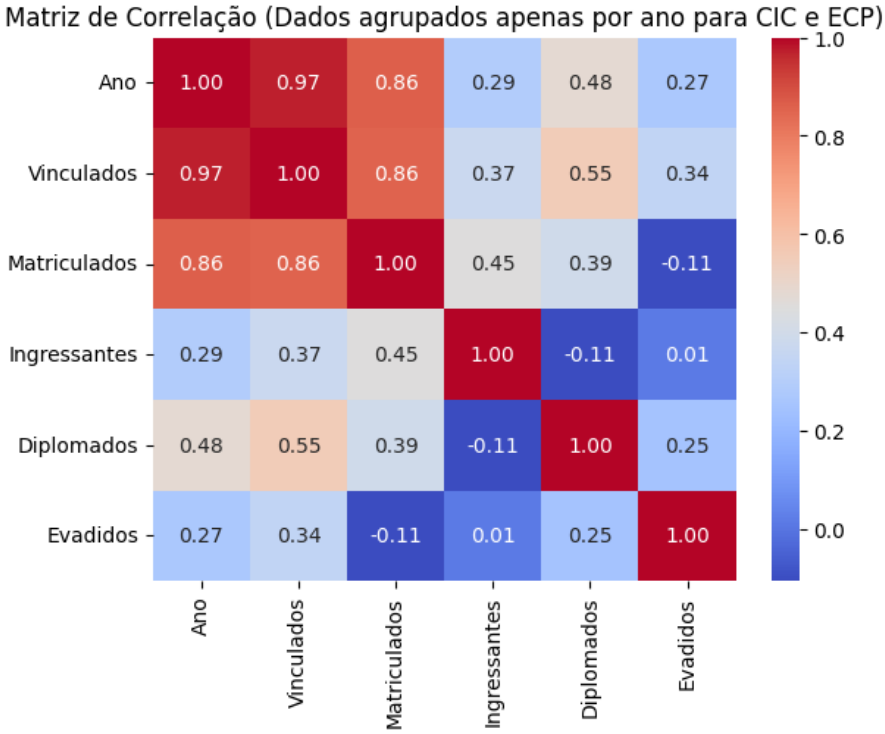


Figura 5.22: Descritivo do *dataframe* que corresponde aos dados referentes aos cursos de CIC e ECP.

	Ano	Vinculados	Matriculados	Ingressantes	Diplomados	Evadidos
count	13.00000	13.000000	13.000000	13.000000	13.000000	13.000000
mean	2016.00000	2104.615385	1886.153846	199.769231	86.230769	85.923077
std	3.89444	275.779060	242.002702	23.249207	9.257014	33.988874
min	2010.00000	1690.000000	1516.000000	162.000000	73.000000	38.000000
25%	2013.00000	1838.000000	1714.000000	180.000000	78.000000	64.000000
50%	2016.00000	2204.000000	1918.000000	200.000000	86.000000	75.000000
75%	2019.00000	2237.000000	1984.000000	214.000000	90.000000	107.000000
max	2022.00000	2559.000000	2409.000000	233.000000	104.000000	153.000000

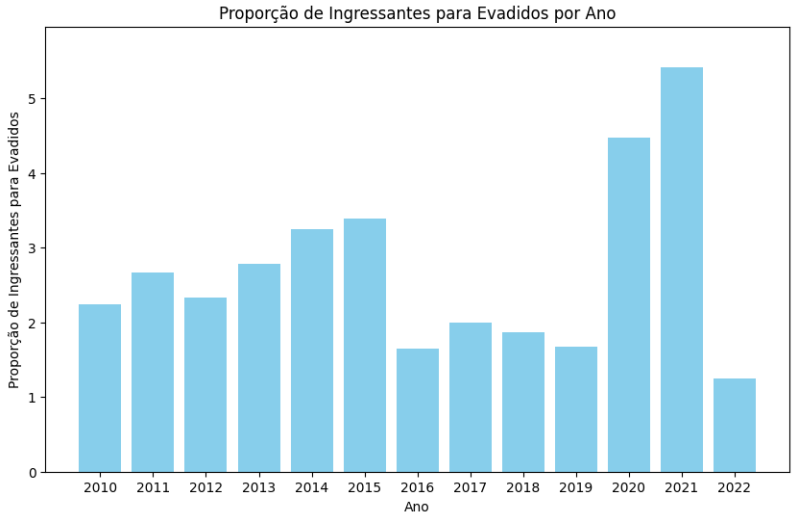
O *describe* do *dataframe* de densidades por ano para os cursos de CIC e ECP (vinculados, matriculados, ingressantes, diplomados e evadidos) mostra algumas informações: contagem de dados, media, desvio padrão, menor valor, quartis e maior valor. Médias gerais interessantes sobre as mesmas, como a semelhança forte entre os diplomados e evadidos. Logicamente, a maior de todas as médias é a de vinculados e logo após vem a de matriculados, e a menor, felizmente, é de evadidos (Figura 5.22).

Utilizando os comandos *apply*, *lambda* e *loc* foi feito um *dataframe* que indica quando as variáveis numéricas com exceção do ano superaram suas respectivas médias (S/N). Verificamos que em 2020, vinculados, matriculados e ingressantes superaram suas médias, em 2021, o único número que não superou a média foi o de evadidos e, por fim,

em 2022 todos exceto pelo número de ingressantes superaram suas médias, mas o que chama atenção aqui é o fato de os evadidos superarem sua média somente em 2022 no período de interesse/exceção, certamente pelo fato de que seu ápice de alunos evadidos foi apenas no ano de 2022. Esta análise é muito extensa, por tanto, está disponível no código contido no GitHub.

Para entender melhor os dados que mais oscilam (ingressantes e evadidos), olhamos mais de perto a relação entre eles calculando a proporção dos que ingressam para os que saem da universidade. Como vemos em 5.23, os dados referentes às proporções começa a aumentar em 2020 e em 2021 atinge seu ápice (5.421053), para depois cair em 2022, o que curiosamente atinge o menor valor de todos os anos (1.248366), mostrando que existe uma forte influencia pós-período de exceção na universidade também em suas proporções devidamente calculadas.

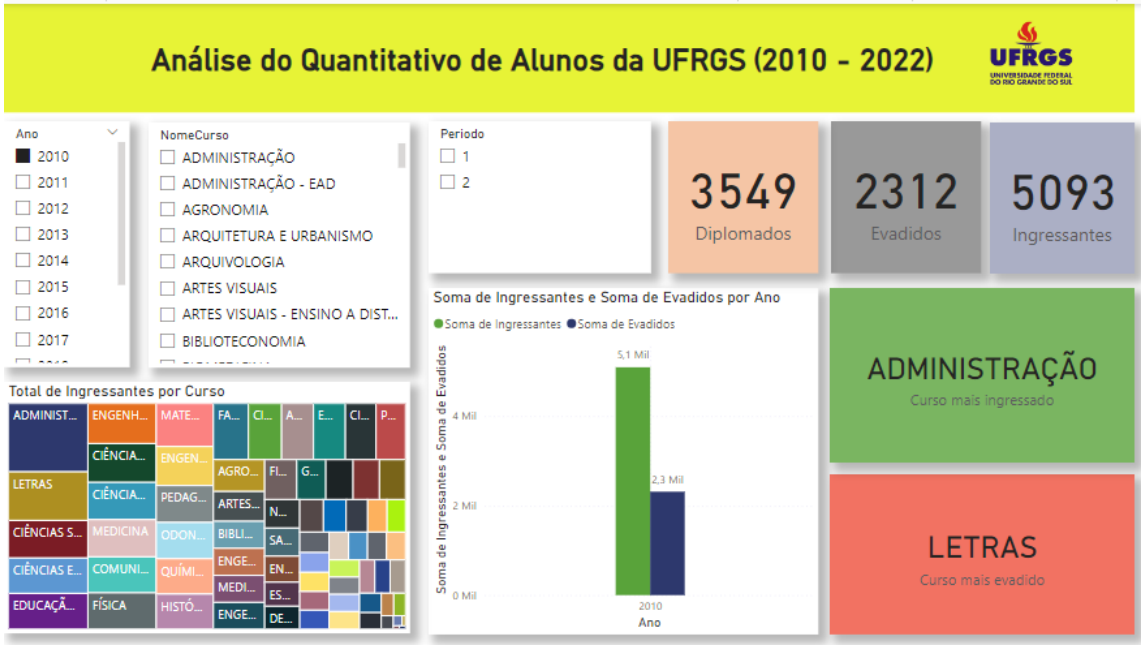
Figura 5.23: Gráfico de barras que ilustra a proporção de ingressantes para evadidos por ano na Universidade.



5.3.1 *Dashboard* dos dados com PowerBI

Os gráficos explicados utilizando as bibliotecas Matplotlib e Seaborn na linguagem Python foram úteis, porém são pouco interativos. Para aumentar a interatividade, dinamicidade e explorabilidade, foi desenvolvido um *dashboard*, um painel de dados interativo com a base de dados utilizando o programa PowerBI, ferramenta referência em análise e ciência de dados e de *Business Intelligence*. Este painel combina os gráficos expostos nesse trabalho e muitos mais, graças ao poder dos filtros que podem ser utilizados pelo usuário, como ano e curso por exemplo. No primeiro momento, foi feita a limpeza de dados que remove os dados correspondentes ao ano de 2023 como explicado no capítulo 4.5. Posteriormente, foram ajustados alguns visuais do painel, como os botões de período e gráfico de soma de ingressantes e soma de evadidos. O *dashboard* reúne de forma efetiva e organizada algumas das análises gráficas feitas e descritas neste trabalho, com visuais que resumem o número total de diplomados, evadidos, ingressantes através de cartões, um heatmap que ilustra os cursos mais ingressados, um paralelo entre ingressantes e evadidos com gráfico de barras e o curso mais ingressado e o mais evadido, todos coordenados através de um *checkbox* que controla a opção de ano desejado para consulta do usuário, bem como o curso e período, que são interativos.

Figura 5.24: Dashboard interativo criado com os dados quantitativos da UFRGS.



5.3.2 Teste t de student com ingressantes e evadidos no período de 2020 a 2022

Nesta seção está explicado foi feita uma análise com teste t de Student, que avalia se duas amostras tem diferença estatisticamente significativa, no caso, o número de ingressantes e o número de evadidos dentro do período que compreende os anos 2020 e 2022 que ocorreram o período de exceção da universidade e também o retorno ao presencial. Testes de Shapiro-Wilk (distribuição normal dos valores) e de Levene (se as amostras possuem variâncias aproximadamente iguais) foram feitos a fim de averiguar se era possível fazer o teste t de Student nas duas amostras (além do fato de ambas serem independentes), ambos resultaram em p-valor maior que 0.05, sendo assim, podendo ser feito o teste t Student de forma efetiva. Nas Figuras a seguir, estão os dados sobre os testes preliminares para o teste t Student. Como mostrado em 5.25, o teste resultou em um p valor de 9.40×10^{-12}. Este valor indica a significância estatística da diferença observada entre as duas amostras testadas. Se o p-value for menor que um nível de significância pré-definido (geralmente 0,05), rejeita-se a hipótese nula e conclui-se que há uma diferença significativa entre as médias das duas amostras (ingressantes e evadidos), o que acontece de fato.

Figura 5.25: Trecho de código em Python que utiliza o teste t de Student para ingressantes e evadidos no período de 2020 e 2022.

```
from scipy.stats import ttest_ind

# Vamos supor que você tenha duas séries de dados: ingressantes e evadidos
ingressantes = df_filtrado['Ingressantes']  # dados de ingressantes
evadidos = df_filtrado['Evadidos']  # dados de evadidos

# Realizar o teste t de Student independente
t_statistic, p_value = ttest_ind(ingressantes, evadidos)

# Exibir os resultados
print("Estatística do teste t:", t_statistic)
print("Valor-p:", p_value)
```

```
Estatística do teste t: 7.69273913866411
Valor-p: 9.40163713024816e-12
```

6 CONCLUSÃO

Com o objetivo de visualizar de forma mais aprofundada os dados referentes aos quantitativos de alunos da UFRGS, principalmente no período de exceção entre 2020 e 2022, este trabalho apresentou a metodologia e o decorrer do mesmo com o uso de gráficos, painel de dados, experimentos e consultas, que corroboraram com a pesquisa, respondendo algumas questões e elucidando outras que surgiram durante a análise exploratória executada e descrita neste texto. Fazendo consultas no banco de dados, percebeu-se que os cursos de Ciência da Computação e Engenharia da Computação não foram os cursos mais ingressados nem os mais evadidos, porém, em 2020, primeiro ano do período de exceção, foi o ano que mais houveram alunos desses cursos ingressando na universidade; ingressantes e evadidos durante o período de exceção na universidade são especialmente discrepantes e, além disso, no final do período de exceção, os cursos tiveram seu maior número de evadidos, muito possivelmente pela pandemia do coronavírus ter levado ao falecimento de muitos alunos; o ano de 2022 teve seu maior valor de alunos vinculados à universidade nos cursos CIC e ECP, o que pode explicar o aumento de pessoas na universidade, no Instituto de Informática e no Restaurante Universitário 6; a proporção ingressantes X evadidos na universidade foi a maior no ano de 2021. Os resultados aqui descritos tem por objetivo compor uma pesquisa definida com problema de pesquisa, análise exploratória e um olhar mais filosófico sobre o efeito da pandemia e quarentena no meio acadêmico. Fica como perspectiva de futuro elaborar, aprimorar e documentar melhor toda a análise exploratória, já que nesta monografia estão documentadas apenas as visualizações e experimentos que melhor resumem toda a pesquisa. Filosoficamente, fica a importância, a relevância de tentar compreender melhor como situações de excepcionalidade afetam a vida das pessoas, não só no campo acadêmico mas no cotidiano e no social.

REFERÊNCIAS

ANDERSON, D. B. **Correlação**. 2023. Acessado em: 21 de junho de 2024. Disponível na Internet: <https://statorials.org/pt/correlacao/>.

BEHAR, P. A. **Artigo: O Ensino Remoto Emergencial e a Educação a Distância**. 2020. Acessado em 10 de julho de 2024. Disponível na Internet: <https://www.ufrgs.br/coronavirus/base/artigo-o-ensino-remoto-emergencial-e-a-educacao-a-distancia/>.

BERNARDO, F. **Git: o que é, para que serve e principais comandos Git!** 2022. Acessado em 01 de agosto de 2024. Disponível na Internet: <https://blog.betrybe.com/git/#1>.

BONATTO, M. A. **Visualização de Dados Institucionais da UFRGS**. Monografia (Monografia) — Faculdade de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

CAVALHEIRO, D. M. **Visualização de dados quantitativos como apoio à análise de desempenho de alunos de graduação da UFRGS**. Monografia (Monografia) — Faculdade de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

CAZORLA, I. M. Média aritmética: um conceito prosaico e complexo. **Anais do IX Seminário de Estatística Aplicada**, p. 1–14, 2003.

CHACON, S.; STRAUB, B. **Pro Git: Everything You Need To Know About It**. 2nd. ed. Apress, 2014. Disponível na Internet: <https://git-scm.com/book/en/v2>.

DATABRICKS. **O que é um DataFrame?** 2024. Acessado em 30 de junho de 2024. Disponível na Internet: <https://www.databricks.com/br/glossary/what-are-dataframes>.

ESCOLA, S. **O que é : Análise Quantitativa**. 2024. Acessado em 07 de julho de 2024. Disponível na Internet: <https://www.soescola.com/glossario/o-que-e-analise-quantitativa>.

FILHO, D. B. F.; JÚNIOR, J. A. S. Desvendando os mistérios do coeficiente de correlação de pearson (r). **Revista Política Hoje**, v. 18, n. 1, p. 115–146, 2009.

GITHUB. **Hey, this is us!** 2021. Acessado em 05 de agosto de 2024. Disponível na Internet: <https://github.com/github>.

GUITARRARA, P. **Pandemia de covid-19**. 2024. Acessado em 20 de julho de 2024. Disponível na Internet: <https://brasilescola.uol.com.br/geografia/pandemia-de-covid-19.htm>.

HARRIS, C. R. et al. Array programming with numpy. **Nature**, Nature Publishing Group UK London, v. 585, n. 7825, p. 357–362, 2020.

IBM. **O que é PostgreSQL?** 2024. Acessado em 29 de junho de 2024. Disponível na Internet: <https://www.ibm.com/br-pt/topics/postgresql>.

LOPES, A. C. B.; LEINIOSKI, A. d. C.; CECCON, L. Testes t para comparação de médias de dois grupos independentes. **Universidade Federal do Paraná–UFPR**, 2015.

MEDRI, W. Análise exploratória de dados. **Londrina: Universidade Estadual de Londrina**, v. 8, n. 8, p. 151–170, 2011.

NETO, V. D. A. **Uma analise sobre reprovação no curso de Ciência da Computação na UFRGS sob a otica dos alunos**. Monografia (Monografia) — Faculdade de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021.

NUMPY. **About Us**. 2005. Acessado em 02 de julho de 2024. Disponível na Internet: <https://numpy.org/about/>.

OPAS. **Histórico da pandemia de COVID-19**. 2024. Acessado em 20 de julho de 2024. Disponível na Internet: <https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19#:~:text=Em%2011%20de%20mar%C3%A7o%20de%202020%2C%20a%20COVID-19,COVID-19%20em%20v%C3%A1rios%20pa%C3%ADses%20e%20regi%C3%B5es%20do%20mundo.>

PANDAS. **pandas - Python Data Analysis Library**. 2024. Acessado em 30 de junho de 2024. Disponível na Internet: <https://pandas.pydata.org/>.

PERRONE, P. S. **Uma ferramenta web para a automatização de relatorios da Sociedade Brasileira de Computação sobre dados referentes ao ensino nacional de tecnologia**. Monografia (Monografia) — Faculdade de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

POSTGRESQL. **PostgreSQL: About**. 2024. Acessado em: 26 de junho de 2024. Disponível na Internet: <https://www.postgresql.org/about/>.

PYPI. **matplotlib 3.9.0**. 2024. Acessado em 01 de julho de 2024. Disponível na Internet: <https://pypi.org/project/matplotlib/>.

RIBEIRO, R. J. **Uma proposta de extração, transformação, carga e visualização para os dados do Censo Escolar**. Monografia (Monografia) — Faculdade de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

SERMARINI, R. A. Variáveis aleatórias. 2016.

SILVA, A. **Calculadora Coeficiente de Correlação Online**. 2022. Acessado em 18 de agosto de 2024. Disponível na Internet: <https://calculareconverter.com.br/calculadora-coeficiente-de-correlacao/>.

SOARES, A. **O que é desvio padrão e como calculá-lo? Saiba mais sobre essa variante estatística!** 2021. Acessado em: 22 de junho de 2024. Disponível na Internet: <https://voitto.com.br/blog/artigo/o-que-e-desvio-padrao>.

SPERANDEI, S. L. M. Testes t de student e bayesianos aplicados a microarranjos: Impacto dos métodos de transformação e do tamanho da amostra. **Universidade Federal do Rio de Janeiro**, 2007.

STONEBRAKER, L. A. R. M. The design of postgres. **SIGMOD: Vol 15 No.2**, Association for Computing MachineryNew YorkNYUnited States, p. 340–355, 1986.

UCHÔA, A. M.; CARVALHO, L. S. G. ao de. Um estudo sobre o impacto das disciplinas na evasão acadêmica sobre diferentes tipos de alunos no icomp. **Anais Estendidos do IV Simpoio Brasileiro de Educação em Computação**, 2024.

WASKOM, M. L. seaborn: statistical data visualization. **Journal of Open Source Software**, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível na Internet: <https://doi.org/10.21105/joss.03021>.