Análise dos dados de demanda dos cursos de Ciência da Computação e Engenharia da Computação

Lucas Brum, Leonardo Bilhalva, Artur Turatti

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

1. Introdução

No primeiro semestre de 2023, na Universidade Federal do Rio Grande do Sul (UFRGS), observamos um aumento notável no número de estudantes matriculados na instituição, o que despertou nossa curiosidade. Diante desse cenário, decidimos conduzir uma análise abrangente desse aumento, utilizando dados dos processos seletivos da UFRGS e informações de matrículas dos alunos da universidade, a fim de, possivelmente, estabelecer uma relação entre eles. Nosso objetivo foi desenvolver uma aplicação que apresente de forma clara e acessível as conclusões dessa análise. A aplicação pode vir a ser valiosa para diversos fins, incluindo pesquisa demográfica, elaboração de censos, estudos relacionados à dinâmica da população em geral e análises voltadas para a comunidade acadêmica.

Utilizamos o PostgreSQL como sistema de gerenciamento de banco de dados para armazenar os dados relevantes. Além disso, empregamos o PowerBI para realizar análises visuais e a biblioteca pandas do Python para conduzir análises adicionais, conforme necessário. Plotamos gráficos diretamente da base de dados utilizando a biblioteca *matplotlib,* para melhor entender a relação entre os dados obtidos.

O restante deste documento estará disposto nas seguintes seções:
2. Descrição da Base de Dados: Nesta seção, apresentaremos uma descrição detalhada dos dados que utilizamos, incluindo dicionários de dados que fornecem informações sobre as diferentes variáveis.
3. Ambiente de Desenvolvimento: Aqui, descreveremos o ambiente de desenvolvimento que estamos utilizando, detalhando os equipamentos, as linguagens de programação e o Sistema de Gerenciamento de Banco de Dados que escolhemos para armazenar e processar nossos dados.
4. Metodologia de Trabalho: Nesta seção, explicaremos como foi planejado o trabalho ao longo do tempo até a conclusão do projeto.
5. Funcionalidades Propostas: Descreveremos as funcionalidades que propomos para nossa aplicação, incluindo as consultas e as visualizações que permitirão uma análise eficaz dos dados, desde as básicas até as mais avançadas.
6. Referências: Finalmente, forneceremos uma lista das fontes e referências que temos utilizado até o momento para embasar nosso projeto.

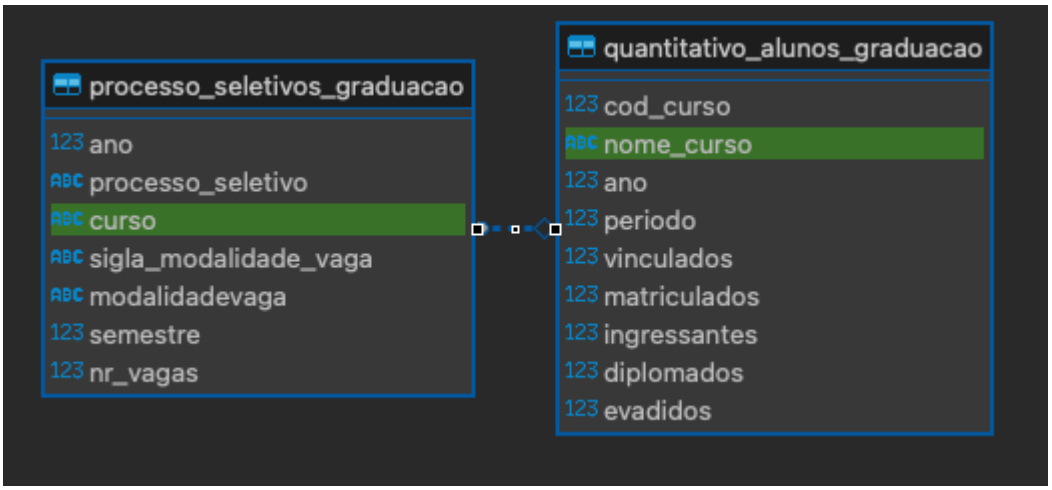
Este documento serve como um guia abrangente para nosso trabalho de análise de dados e desenvolvimento de aplicações, visando entender e comunicar o número de alunos da UFRGS durante o semestre de 2023/1.

2. Descrição da base de dados

Nesta atividade, empregamos duas fontes de dados distintas: uma referente aos Processos Seletivos da Graduação, contendo informações quantitativas relativas aos processos de admissão dos cursos de graduação na UFRGS, e outra referente ao Quantitativo de Alunos de Graduação, que também inclui dados demográficos dos estudantes da instituição. Ambas as fontes de dados estão disponíveis na seção de dados abertos da UFRGS, acessível através do seguinte link: https://dados.ufrgs.br/.

Para uma melhor visão, decidimos tratar os dados de entrada para que tivéssemos correspondências entre essas tabelas e pudéssemos agregá-las (JOIN). Como os nomes dos cursos se encontravam diferentes e não havia códigos comuns, inserimos as entradas na tabela utilizando uma lista de nomes de cursos e um algoritmo chamado *longest prefix matching* que fez com que os nomes ficassem iguais. Então, temos o nome do curso como foreign key.

O diagrama ER:



3. Ambiente de desenvolvimento

Para o tratamento e importação de dados, utilizamos a linguagem de programação Python 3.10, juntamente com a biblioteca *psycopg2* que faz a conexão entre a linguagem e o SGBD PostgreSQL (utilizado para armazenar os dados tratados). Na parte de visualização e interpretação dos dados, utilizamos o PowerBI, mas principalmente Python com sua biblioteca *matplotlib* para plotar os gráficos, sempre utilizando os dados providos do PostgreSQL.

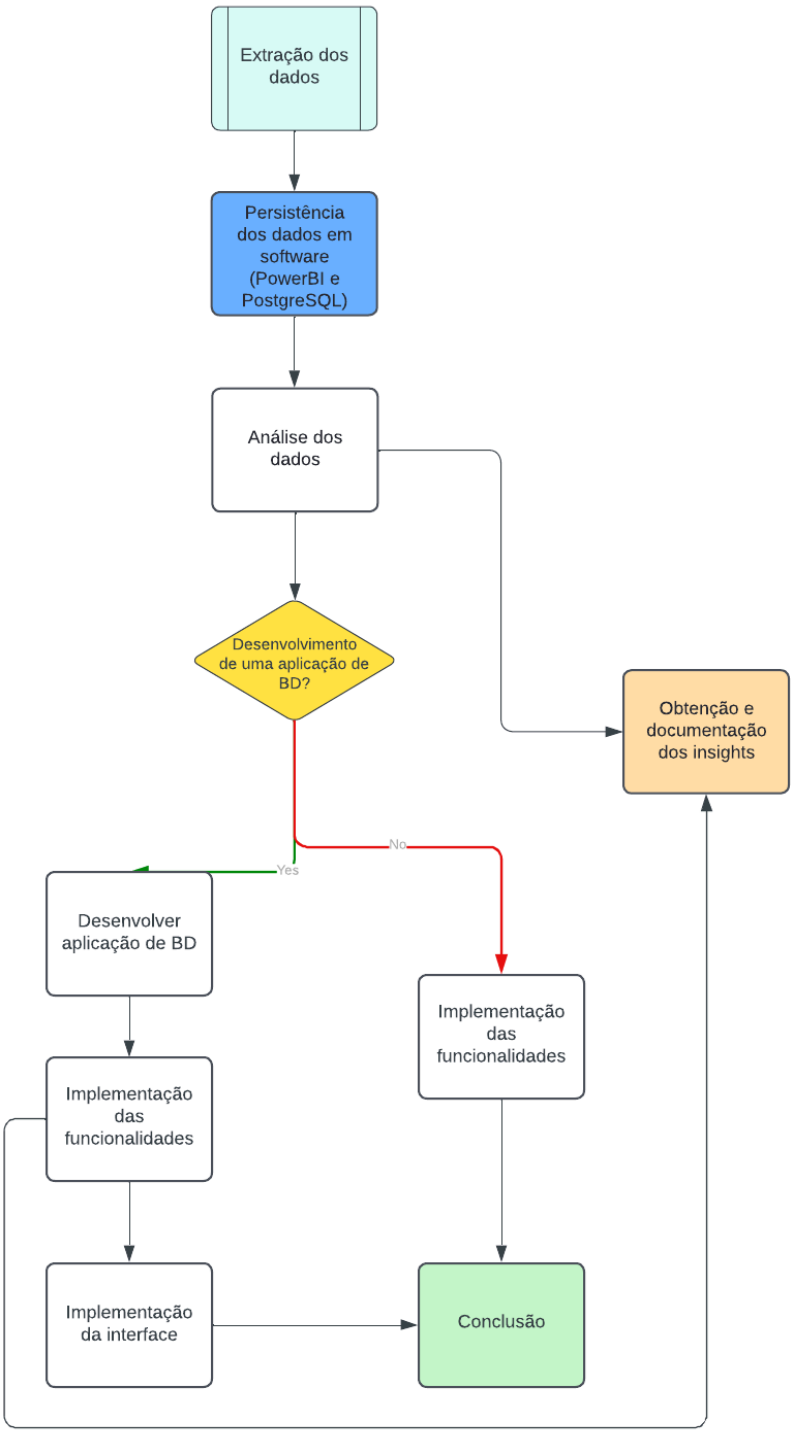
Equipamento 1

- PC MEGAWARE UPD/LX MEGACORP MOD. M7SERIES
- Placa mãe MEGAWARE MW-H61HD-MA
- Processador Intel i7-3770
- Memória 16GB DDR3 SDRAM
- HD Seagate ST1000DM010-2EP102 1000GB
- Sistema operacional Microsoft Windows 10 Home (x64) Build 19045.3208 (22H2)

Equipamento 2

- Placa mãe ASRock Steel Legend B450M mATX DDR4
- Processador AMD Ryzen 5 3600X 6-core 3.8GHz
- Memória 2x8GB DDR4 3000MHz
- Sistema operacional Microsoft Windows 11 Pro 64 bits (10.0, build 22621)
- M.2 NVMe Kingston KC3000 1TB

4. Metodologia



Obtivemos os dados do portal de Dados Abertos da UFRGS, em formato CSV:

1. (https://dados.ufrgs.br/dataset/processos-seletivos-graduacao)
2. (https://dados.ufrgs.br/dataset/quantitativo-de-alunos-de-graduacao)

Os dados foram tratados utilizando a linguagem Python, juntamente com as bibliotecas CSV (para a leitura dos arquivos CSV) e psycopg2 (para a conexão com o BD para realizar os inserts). Além disso, utilizamos um algoritmo de *longest prefix matching* para criar uma lista

comum de nomes de cursos que fossem comuns entre as duas tabelas, inserindo nomes diferentes do que no CSV para que pudessem ser usados para agregar as tabelas.
Também implementamos uma aplicação em Python utilizando a biblioteca *matplotlib* que pega dinamicamente nomes de cursos e campos diferentes para termos uma visão mais abrangente e confirmarmos teses que tivemos durante insights (como por exemplo: Engenharia da Computação tem mais evasão que Ciência da Computação pois tem disciplinas com maior grau de reprovação, como por exemplo Física).

Problemas no caminho

Nosso principal objetivo ao tratar os dados e deixar o nome dos cursos corretos correspondentes era utilizar a densidade ou notas de corte para tentar relacionar isso a taxa de evasão, porém, não conseguimos achar nada sobre densidades em tabelas prontas
Achamos então, na página da UFRGS as densidades e pensamos em puxar os dados do próprio site
https://www.ufrgs.br/vestibular/cv2023/densidade/

Administração - Diurno	127	56	2,27
Acesso Universal	87	28	
Ensino Público independentemente da renda familiar	19	3	
Ensino Público independentemente da renda familiar e Autodeclarado Preto/Pardo/Índio	4	3	
Ensino Público com Renda Igual ou Inferior a 1,5 Salários Mínimos	14	3	
Ensino Público com Renda Igual ou Inferior a 1,5 Salários Mínimos e Autodeclarado Preto/Pardo/Índio	2	3	
Ensino Público independentemente da renda familiar PcD		4	
Ensino Público independentemente da renda familiar e Autodeclarado Preto/Pardo/Índio PcD		4	
Ensino Público com Renda Igual ou Inferior a 1,5 Salários Mínimos PcD	1	4	
Ensino Público com Renda Igual ou Inferior a 1,5 Salários Mínimos e Autodeclarado Preto/Pardo/Índio PcD		4	

No final a ideia foi abandonada pois a densidade era apenas do vestibular e não tínhamos os dados suficientes para completar a query proposta e responder essas perguntas

5. Funcionalidades

As funcionalidades são no formato de cinco consultas, três básicas e duas avançadas. Algumas são feitas em SQL, outras visualizações. Adicionamos os insights iniciais que nos levaram a montar as perguntas e o que buscamos com elas.

5.1 Back-end e Front-end

Para nosso back-end utilizamos o SGBD PostgreSQL 14.9, que serviu para armazenar os dados e fazer as consultas utilizando cláusulas e agrupando linhas de nosso interesse

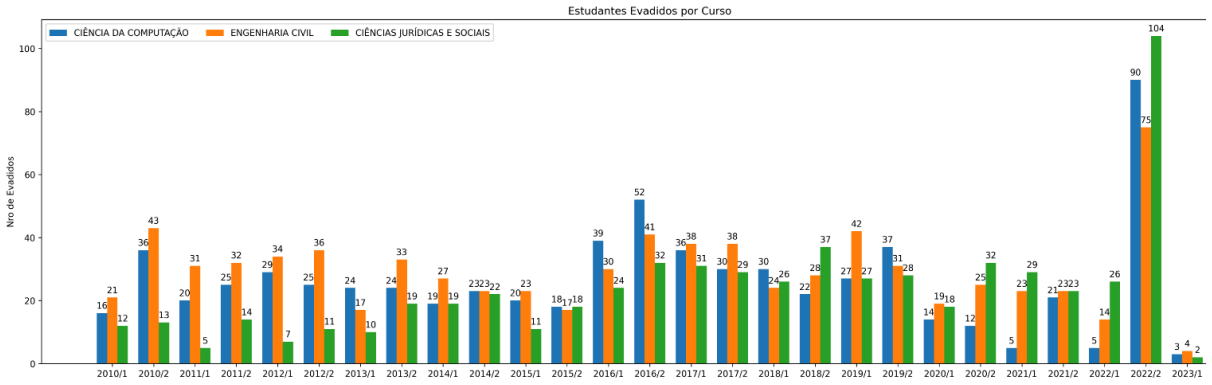
```
psql (14.9 (Ubuntu 14.9-0ubuntu0.22.04.1))
Type "help" for help.

projetobd=# SELECT ano,periodo,sum(evadidos) AS evadidos FROM quantitativo_alunos_graduacao WHERE nome_curso like 'CIÊNCIA DA COMPUTAÇÃO' GROUP BY nome_curso,ano,periodo ORDER BY ano,periodo
```

ano	periodo	evadidos
2010	1	16
2010	2	36
2011	1	20
2011	2	25
2012	1	29
2012	2	25
2013	1	24
2013	2	24
2014	1	19
2014	2	23
2015	1	20
2015	2	18

Essa consulta por exemplo, mostra todos os evadidos do curso de Ciência da Computação divididos por (ano/semestre)

No front-end, a principal ferramenta foram gráficos plotados com Python e matplotlib, como por exemplo esse, que compara os alunos evadidos em toda base de dados entre os cursos de CiC, Engenharia Civil e Direito:



5.2. Funcionalidades

1. Filtra os conceitos iniciais para a pesquisa

SELECT ano, processo_seletivo, curso, nr_vagas **FROM** processo_seletivos_graduacao;

ano	processo_seletivo	curso	nr_vagas
2018	Vestibular	Administração - Diurno	14
2018	Vestibular	Administração - Diurno	14
2018	Vestibular	Administração - Diurno	2
2018	Vestibular	Administração - Diurno	1
2018	Vestibular	Administração - Diurno	2
2018	Vestibular	Administração - Diurno	2
2018	Vestibular	Administração - Diurno	2

Permite ter uma visão ampla dos dados

2. Achar os cursos com maior número de evasão na base de dados (nesse caso, os 5 maiores)

```
SELECT nome_curso, SUM(evadidos) AS total_evadidos
FROM quantitativo_alunos_graduacao
GROUP_BY nome_curso
ORDER BY total_evadidos DESC
LIMIT 5;
```

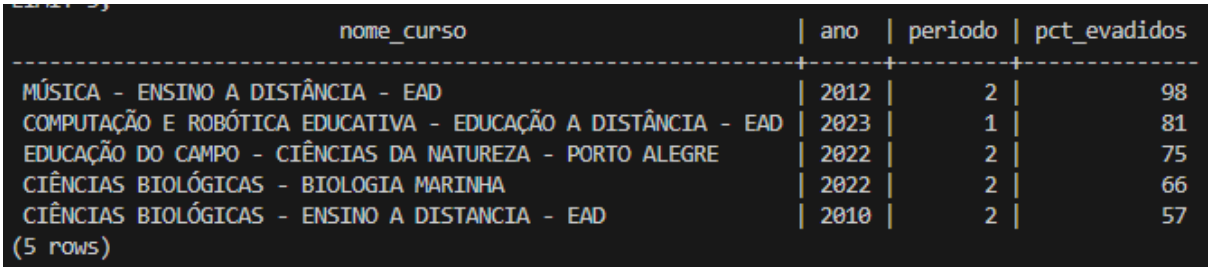
nome_curso	total_evadidos
CIÊNCIAS SOCIAIS	1892
LETRAS	1750
MATEMÁTICA	1542
ADMINISTRAÇÃO	1486
FÍSICA	1485

(5 rows)

Foi listado os 5 cursos que mais tiveram alunos evadindo

3. Achar os cursos com maiores TAXAS (%) de alunos evadidos por vinculados, que mostrará com mais precisão os cursos que mais tem alunos evadindo, pois existem cursos com muitos mais alunos que outros, e, naturalmente eles teriam uma evasão maior proporcional.

```
SELECT nome_curso, ano, periodo, ((evadidos*100/vinculados)) AS
pct_evadidos
FROM quantitativo_alunos_graduacao
WHERE ((evadidos*100/vinculados)) < 100
ORDER BY pct_evadidos DESC
LIMIT 5;
```



nome_curso	ano	periodo	pct_evadidos
MÚSICA - ENSINO A DISTÂNCIA - EAD	2012	2	98
COMPUTAÇÃO E ROBÓTICA EDUCATIVA - EDUCAÇÃO A DISTÂNCIA - EAD	2023	1	81
EDUCAÇÃO DO CAMPO - CIÊNCIAS DA NATUREZA - PORTO ALEGRE	2022	2	75
CIÊNCIAS BIOLÓGICAS - BIOLOGIA MARINHA	2022	2	66
CIÊNCIAS BIOLÓGICAS - ENSINO A DISTANCIA - EAD	2010	2	57

(5 rows)

Mostra os cursos que tiveram a maior taxa de evasão e o período em que ocorreu (ano/semestre)

Código completo disponível em:
https://github.com/leonardobilhalva/DatabaseProjectFinalAssignment

Demonstração disponível em:

https://www.youtube.com/watch?v=V4s4ot9yPso

6. Conclusão

Para uma melhor compreensão dos movimentos de alunos dentro de cursos na UFRGS, tivemos a ideia de analisar os dados para tentar entender por que acontecem, principalmente se aliando a Bancos de Dados, objetivo principal da cadeira. Para isso foram desenvolvidas queries específicas para cada um dos insights que os participantes do grupo iam tendo, com a intenção que afirmassem o insight ou refutassem o mesmo.

Os principais aprendizados que obtivemos foi a ideia de aliar ideias (insights) que obtivemos a bancos de dados e utilizar isso para tentar enxergar melhor se faz sentido ou não. Apesar do pouco volume de dados disponível por parte da Universidade, conseguimos obter diversos pontos fortes de por que em alguns cursos ocorrem coisas que em outros.

E o mais importante, a principal pergunta que tinhamos: por que existe tanta evasão? - isso pode ser melhor explicado após entendermos quais áreas tem maior taxa, e, principalmente, tentar melhorar e evitar que estudantes evadam cursos ou até que entrem em cursos que desistam posteriormente.

7. Referências bibliográficas

Ribeiro, R. (2023). Uma proposta de extração, transformação, carga e visualização para os dados do Censo Escolar, https://lume.ufrgs.br/handle/10183/259957

Perrone, S. P. (2023). Uma ferramenta web para a automatização de relatórios da Sociedade Brasileira da Computação sobre dados referentes ao ensino nacional de tecnologia, https://lume.ufrgs.br/handle/10183/261792

Site oficial PostgreSQL https://www.postgresql.org/

Site oficial PowerBI https://powerbi.microsoft.com/

Portal de Dados Abertos da UFRGS: Processos Seletivos da Graduação https://dados.ufrgs.br/dataset/processos-seletivos-graduacao

Portal de Dados Abertos da UFRGS: Quantitativo de Alunos da Graduação

https://dados.ufrgs.br/dataset/quantitativo-de-alunos-de-graduacao

GZH. Com aumento de quase 45% no número de candidatos, UFRGS divulga a densidade do vestibular 2023, https://gauchazh.clicrbs.com.br/educacao-e-emprego/noticia/2022/12/com-aumento-de-quase-45-no-numero-de-candidatos-ufrgs-divulga-densidade-do-vestibular-2023-clbs0gm5p001x018fhys2dsd9.html