UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

RODRIGO PASSOS - 115196299

UFRJ

TRABALHO DE DW SUP. A DECISÃO
RIO DE JANEIRO
2021

Link do Github: https://github.com/Rodrigoes08/trabalho_dw

Questão 1

- *Onde esses dados podem ser coletados:*

Esses dados estão armazenados no site do INEP e foi possível obter eles por meio de uma biblioteca do python chamada requests, onde ele manda a requisição para um link específico e consegue extrair os arquivos. A Figura 1 abaixo exemplifica como é feito os downloads desses dados, de qual link e como extrair os arquivos dessa pasta zipada:

1. Primeira Questão - Extraindo Arquivos

```
os.makedirs("./enade",exist_ok=True) # Criando uma pasta com os arquivos
# URLs com os arquivos gerado
url2019 = "http://download.inep.gov.br/microdados/Enade_Microdados/microdados_enade_2019.zip"
url2018 = "http://download.inep.gov.br/microdados/Enade_Microdados/microdados_enade_2018.zip"
url2017 = "http://download.inep.gov.br/microdados/Enade_Microdados/microdados_Enade_2017_portal_2018.10.09.zip"
```

```
# Extraindo dados de 2019
filebytes = BytesIO(requests.get(url2019).content)
myzip = zipfile.ZipFile(filebytes)
myzip.extractall("./enade/2019")

# Extraindo dados de 2018
filebytes2 = BytesIO(requests.get(url2018).content)
myzip2 = zipfile.ZipFile(filebytes2)
myzip2.extractall("./enade")

# Extraindo dados de 2017
filebytes3 = BytesIO(requests.get(url2017).content)
myzip3 = zipfile.ZipFile(filebytes3)
myzip3.extractall("./enade/2017")
```

Figura 1 - Exemplo de Código

- *Que arquivos são obtidos ?*

São obtidos três pastas, como mostra a Figura 2 abaixo:

1.LEIA-ME	27/11/2021 00:17	Pasta de arquivos
2.INPUTS	27/11/2021 00:17	Pasta de arquivos
3.DADOS	27/11/2021 00:17	Pasta de arquivos

Figura 2 - Pastas geradas

A primeira pasta tem os dicionários de variáveis, onde podemos encontrar todas as correlações entre as variáveis e também podemos encontrar o significado de cada uma delas e avaliar as especificações de cada coluna. A segunda pasta tem alguns arquivos de input, onde eu realmente não consegui identificar muito bem do que se tratava e como usar ela para alguma coisa no decorrer do processo. E a última pasta é onde estão os dados de cada aluno propriamente dito, mas sem identificação dos mesmos.

- *Tipo de Informação:*

A informação nela é armazenada em arquivos em excel com mais de 100 colunas, onde cada uma delas descreve as questões apresentadas na prova e o estilo do aluno. Por exemplo, características como raça, estado civil, ano de ingresso na faculdade e dentre outras coisas. A Figura 3 abaixo, mostra o tamanho de cada uma das três bases extraídas:

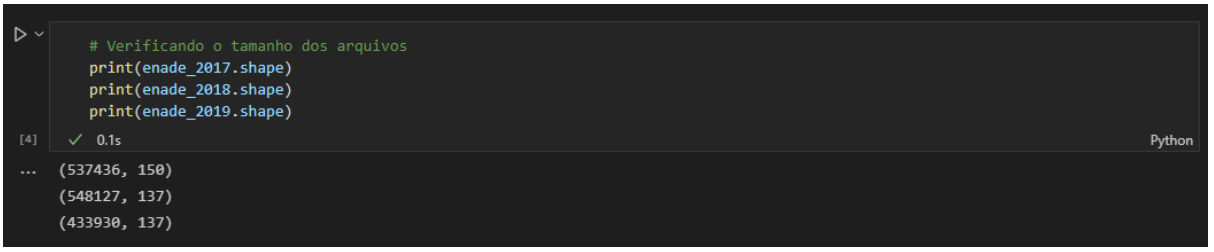

Figura 3 - Tamanho e Quantidade de colunas nas bases

Questão 2

O modelo dimensional (estrela) foi construído com 12 dimensões no geral e construído por meio da ferramenta SQL Power Architect e a Figura 4 abaixo apresenta esse modelo:

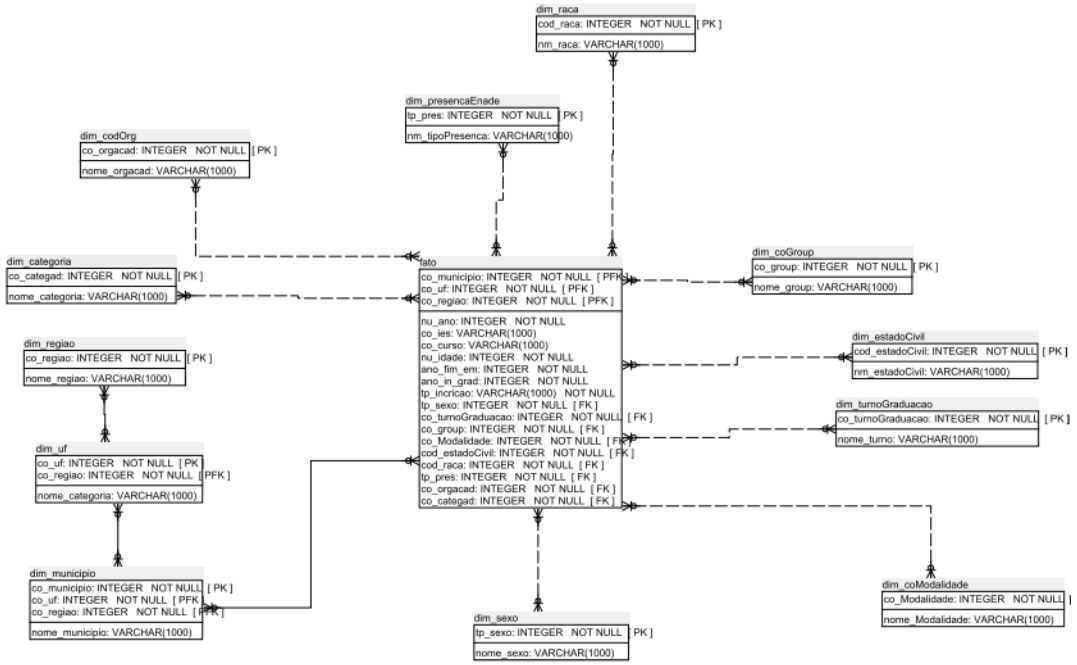

Figura 4 - Modelo dimensional

Para análise do modelo, foi utilizado o Excel, nos dicionários de dados, para poder entender quais eram as dimensões mais interessantes a serem utilizadas no

gráfico. E por fim, fechamos o modelo com uma tabela fato e 12 dimensões e as dimensões são as seguintes:

1. ***dim_codOrg***: Código da organização acadêmica da IES
2. ***dim_categoria***: Código da categoria administrativa da IES
3. ***dim_regiao***: Código da região de funcionamento do curso
4. ***dim_uf***: Código da UF de funcionamento do curso
5. ***dim_municipio***: Código do município de funcionamento do curso
6. ***dim_sexo***: Sexo
7. ***dim_coModalidade***: Código da Modalidade de Ensino
8. ***dim_turnoGraduacao***: Código do turno de graduação
9. ***dim_estadoCivil***: Estado Civil do Candidato
10. ***dim_coGroup***: Código da Área de enquadramento do curso no Enade
11. ***dim_raca***: Cor ou raça do candidato
12. ***dim_presencaEnade***: Tipo de presença do Enade

Questão 3

Foi utilizado o MySQL workbench de forma local para poder simular um banco de dados e criar um database via Python com a biblioteca SQLAlchemy para poder construir um banco relacional. A figura 5 abaixo exemplifica uma parte do código e o resto pode ser encontrado no Github.

Figura 5 - Parte do código

Questão 4

- *Fluxo:*

O primeiro passo é adicionar os dados nas dimensões do nosso datawarehouse construído. Então a primeira coisa a ser feita é ler os dados do excel com o dicionário de dados e pegar cada opção disponível para cada dimensão, a Figura 6 abaixo exemplifica uma dimensão e mostra como é feita cada dimensão.

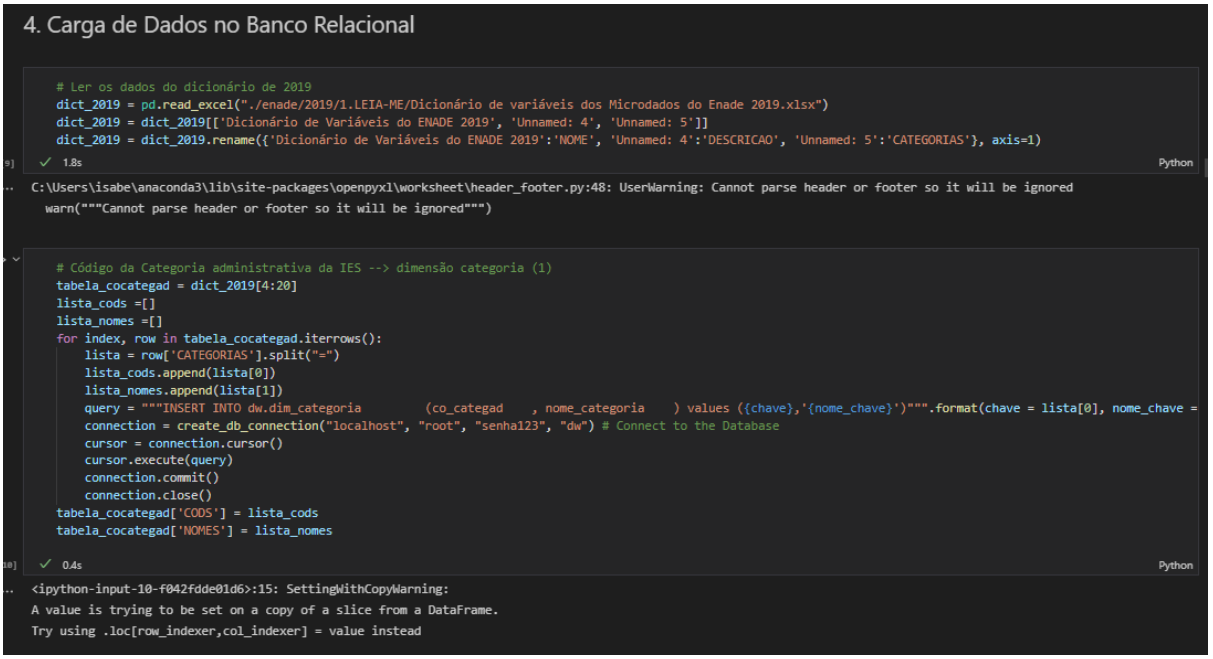

Figura 6 - Exemplo de dimensão

Após as dimensões serem inseridas uma por uma, no código é possível verificar como é feito, pois cada célula é uma dimensão e apenas a dimensão de lugar, tem o município, uf e região agrupados, mas os demais estão em uma única. Então a tabela fato é preenchida com os dados propriamente dito, pois as dimensões já estão com as correspondências da foreign key. Abaixo, segue a Figura 7 exemplificando a adição na tabela fato dos dados do enade de 2019, mas o mesmo é feito com os outros dois anos.

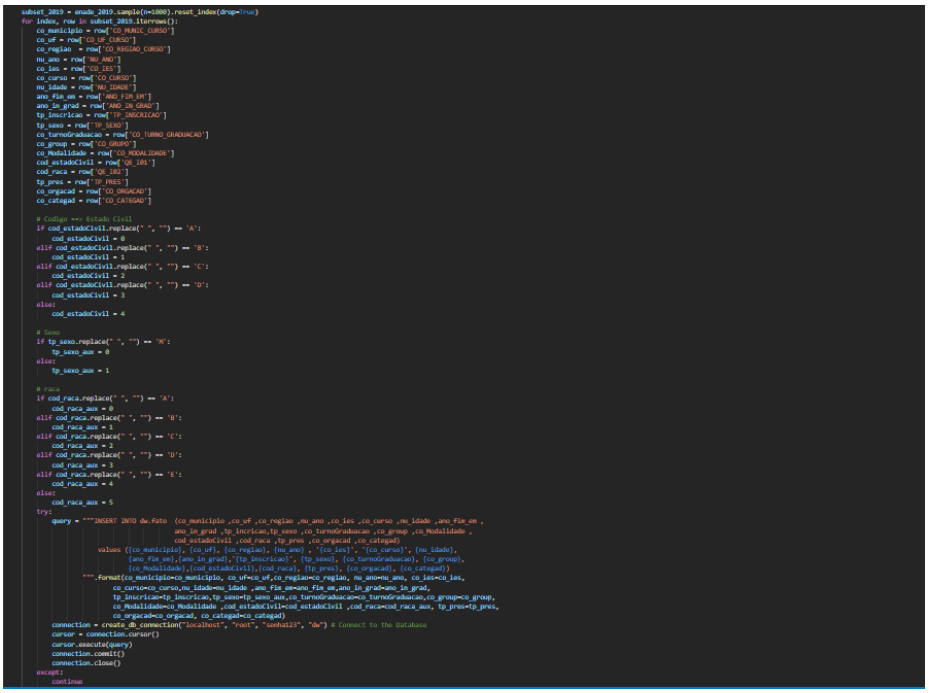

Figura 7 - Exemplo de adição de dados na tabela fato

- *Diferença entre 3 bases*:

Acho que a diferença maior entre as três bases é o número de colunas, pois a de 2017 é mais antiga e possui um número de colunas maior que as duas mais recentes. Além disso, nas bases de 2018 e 2017, tem alguns valores nulos. Além de que os dados de 2017, tem um dicionário estruturado com basicamente a mesma correlação, mas formatado de forma diferente.

Questão 5

No pandas, existe uma lib que se chama “pandas-profiling”, ela busca realizar uma análise de dados, encontrando a correlação entre todas as colunas e verificando as proporções de cada uma delas. No código é possível olhar essa análise detalhadamente e tirar várias conclusões em cima disso. Sendo assim, a partir disso, tirei as seguintes perguntas para serem analisadas:

1. Qual a distribuição de idade das pessoas que participaram durante os três anos de enade estudados ?

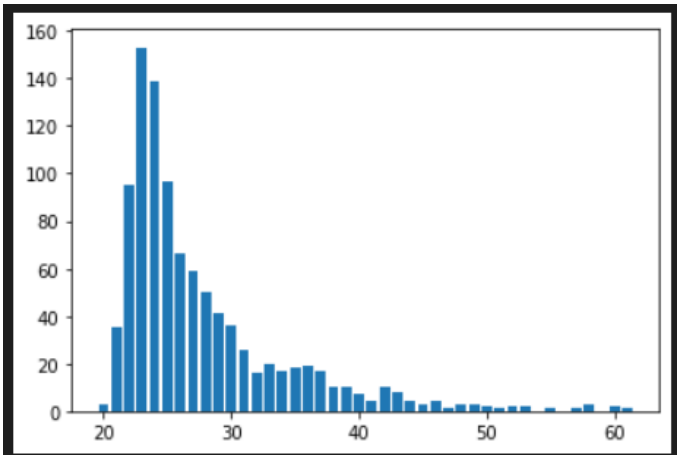

Figura 7 - Distribuição de idade

Nela podemos ver que o maior percentual de pessoas está entre 20 e 30 anos, como o que já é esperado de uma universidade. Sendo a menor parte do público, as pessoas maiores de 40 anos.

2. Qual o ano, onde a maior parte das pessoas que estão fazendo a prova ingressou na faculdade?

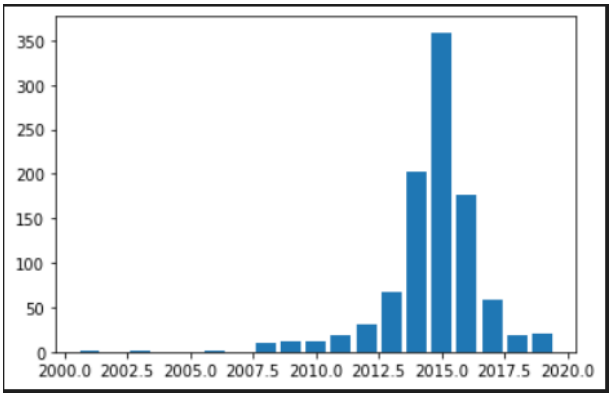

Figura 8 - Ano de entrada na faculdade

3. Qual o ano, onde a maior parte das pessoas que estão fazendo a prova concluiu?

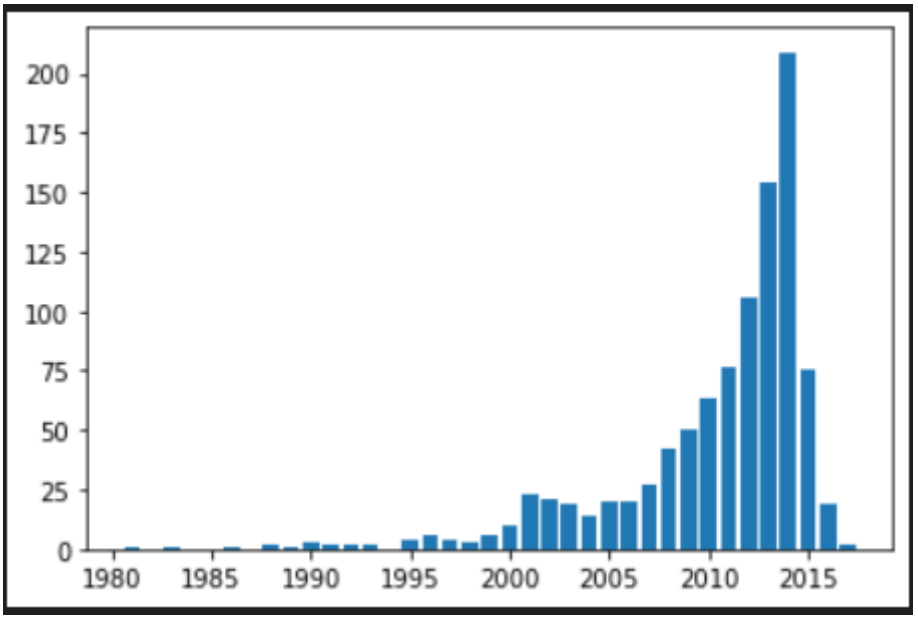

Figura 9 - Ano de conclusão

4. Qual a distribuição de pessoas por raça ?

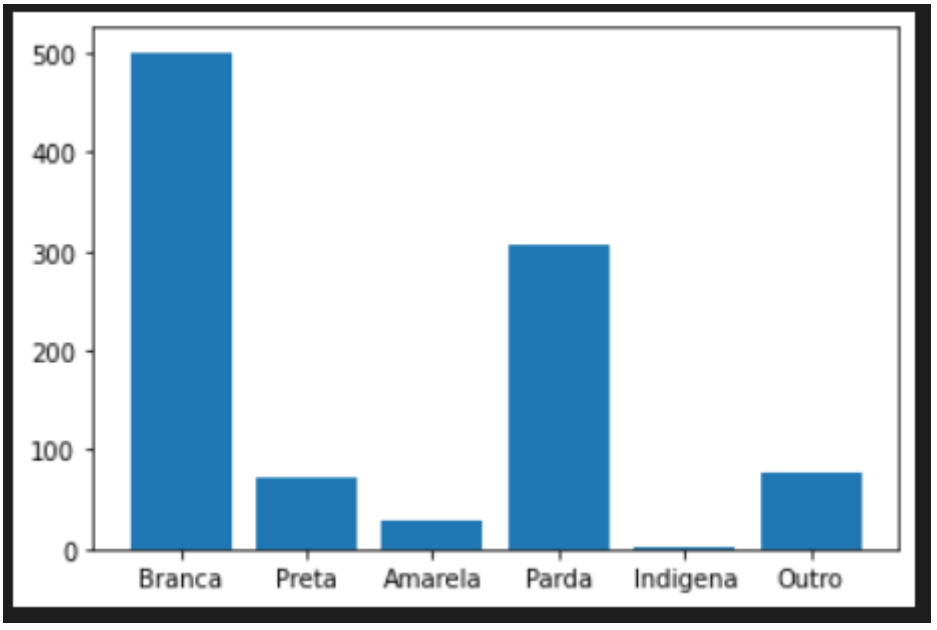

Figura 10 - Distribuição por raça

A distribuição por raça, mostra nitidamente que a maior parte da população que realiza o ENADE se identifica como Branca e Parda no geral.

5. Qual a distribuição de pessoas por estado civil ?

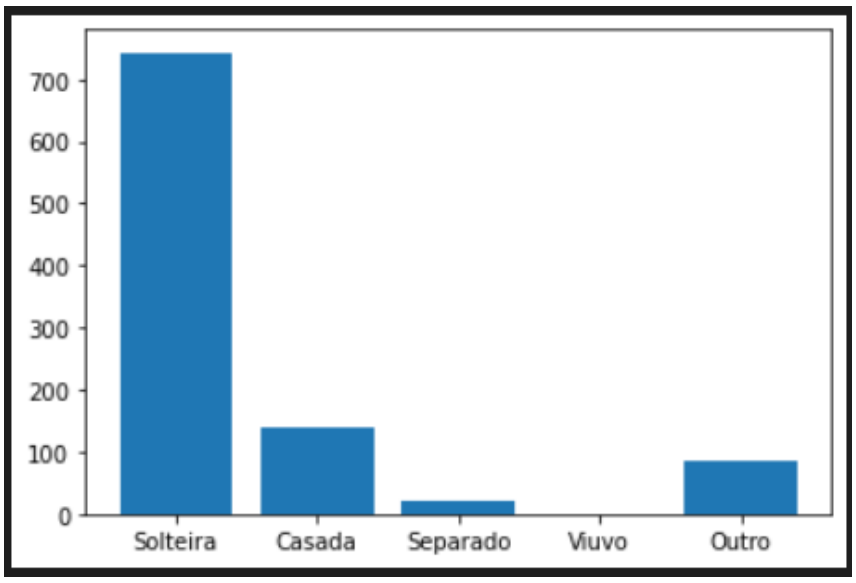

Figura 11 - Distribuição por Estado Civil

Como o previsto, que se correlaciona nitidamente com a idade, a maior parte dos estudantes, mais que a metade, é composto por pessoas solteiras.

6. Qual a distribuição de pessoas por região do pais ?

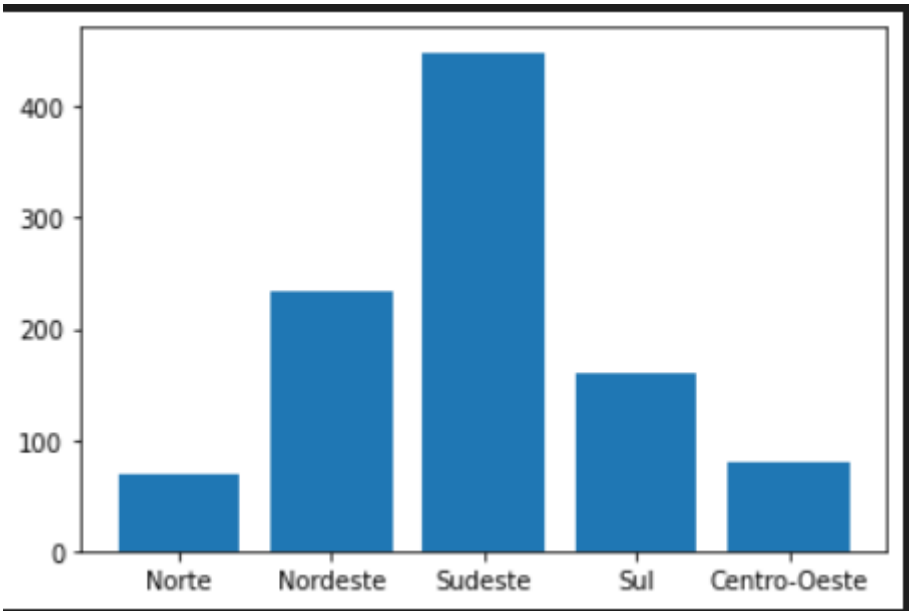

Figura 12 - Distribuição por região do país

As regiões com maior concentração populacional do país, tendem a ter uma quantidade de pessoas maior que as demais.

7. Qual a relação entre a quantidade de pessoas de determinada raça e a sua região ?

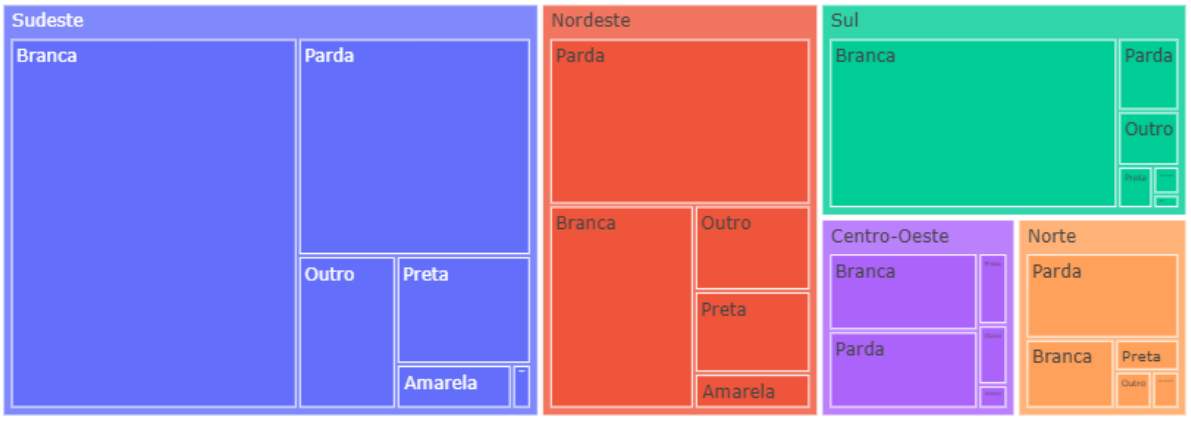

Figura 13 - TreeMap de Região e por Raça

Se pensamos em um lado histórico cultural, podemos entender um pouco melhor essa distribuição de raça por região. E a partir daí vem a pergunta de como o sexo se correlaciona com esses dois.

8. E quando olhamos o Sexo, como isso é distribuido ?

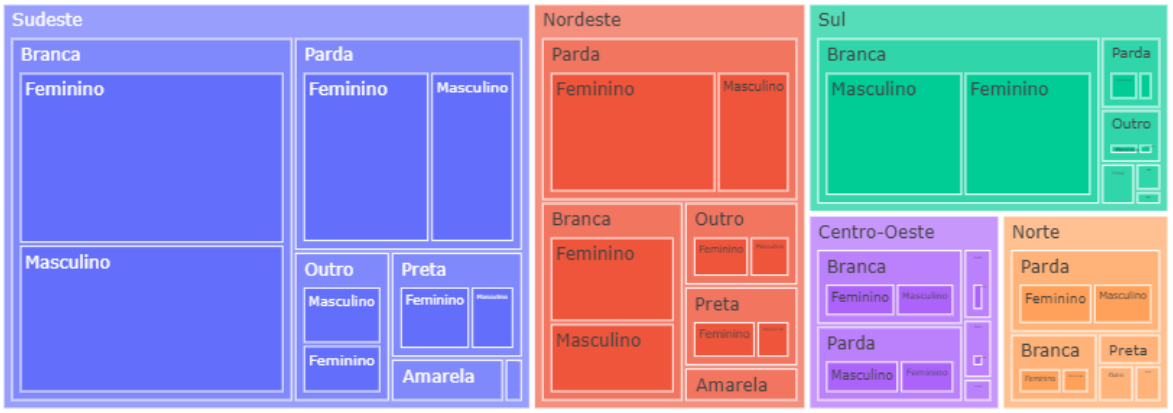

Figura 14 - TreeMap em três niveis

9. Qual o percentual que cada um dos gêneros corresponde em cada região ?

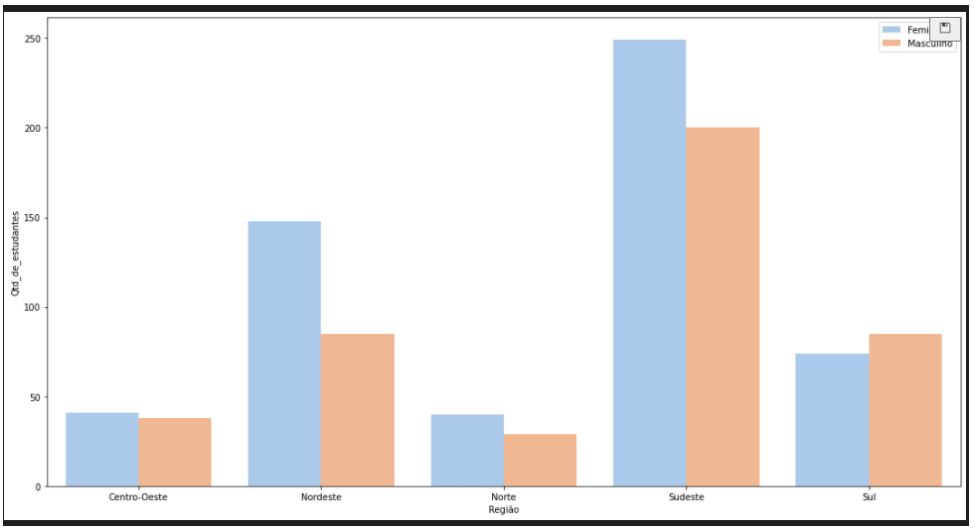

Figura 15 - Distribuição por região e sexo

Questão 6

Durante o estudo realizado, foi aplicado um modelo de classificação para dois parâmetros diferentes da tabela, ou seja, foram escolhidas duas colunas target para serem usadas no nosso classificador e verificar qual o resultado obtido.

O primeiro target foi o target do estado civil, nele foi usado 37% do dataset para teste e o resto para treino. E no final, foi possível obter uma acurácia no modelo de 75%.

O segundo target utilizado, foi a raça e nele também foi usado 37% do dataset para teste e o resto para treino. No final, foram feitos alguns testes de percentual de treino para aumento da acurácia e o resultado máximo obtido foi de 57%, não conseguindo fazer o modelo aprender o mínimo de 60% do que foi acordado.

Questão 7

SQL Power Architect: ferramenta usada para desenho da modelagem dimensional e construção das relações entre as tabelas no nosso modelo estrela.
link: http://www.bestofbi.com/page/architect_download_os

Python: linguagem de programação usada para desenvolvimento dos códigos necessários durante todo o processo.
link: https://www.python.org/

Jupyter: uma espécie de IDE para rodar os códigos em Python e para estruturar a prova realizada.
link: https://jupyter.org/

MySQL: banco de dados utilizado para inserção de dados.
link: https://www.mysql.com/

MySQL Workbench: plataforma utilizada para visualização de tabelas e entendimento de funcionamento das tabelas do banco de dados.
link: https://dev.mysql.com/downloads/workbench/

Google Docs: ferramenta utilizada para inscrição do relatório.
link:https://workspace.google.com/intl/pt-BR/products/docs/?utm_source=google&utm_medium=cpc&utm_campaign=latam-BR-all-pt-dr-bkws-all-all-trial-e-dr-1009897-LUAC0011906&utm_content=text-ad-none-any-DEV_c-CRE_470571214209-ADGP_Hybrid%20%7C%20BKWS%20-%20MIX%20%7C%20Txt%20~%20Docs-KWID_43700057676888777-kwd-4379564344&utm_term=KW_google%20docs-ST_google%20docs&gclid=Cj0KCQiA7oyNBhDiARIsADtGRZZQnT5Ha8A1NKn0PBUSwOvcj8zKG1-VcOfavgMpLsNB4fKspPsXlWwaAnKuEALw_wcB&gclsrc=aw.ds

Excel: ferramenta utilizada para análise de colunas.
link: https://www.microsoft.com/pt-br/microsoft-365/excel