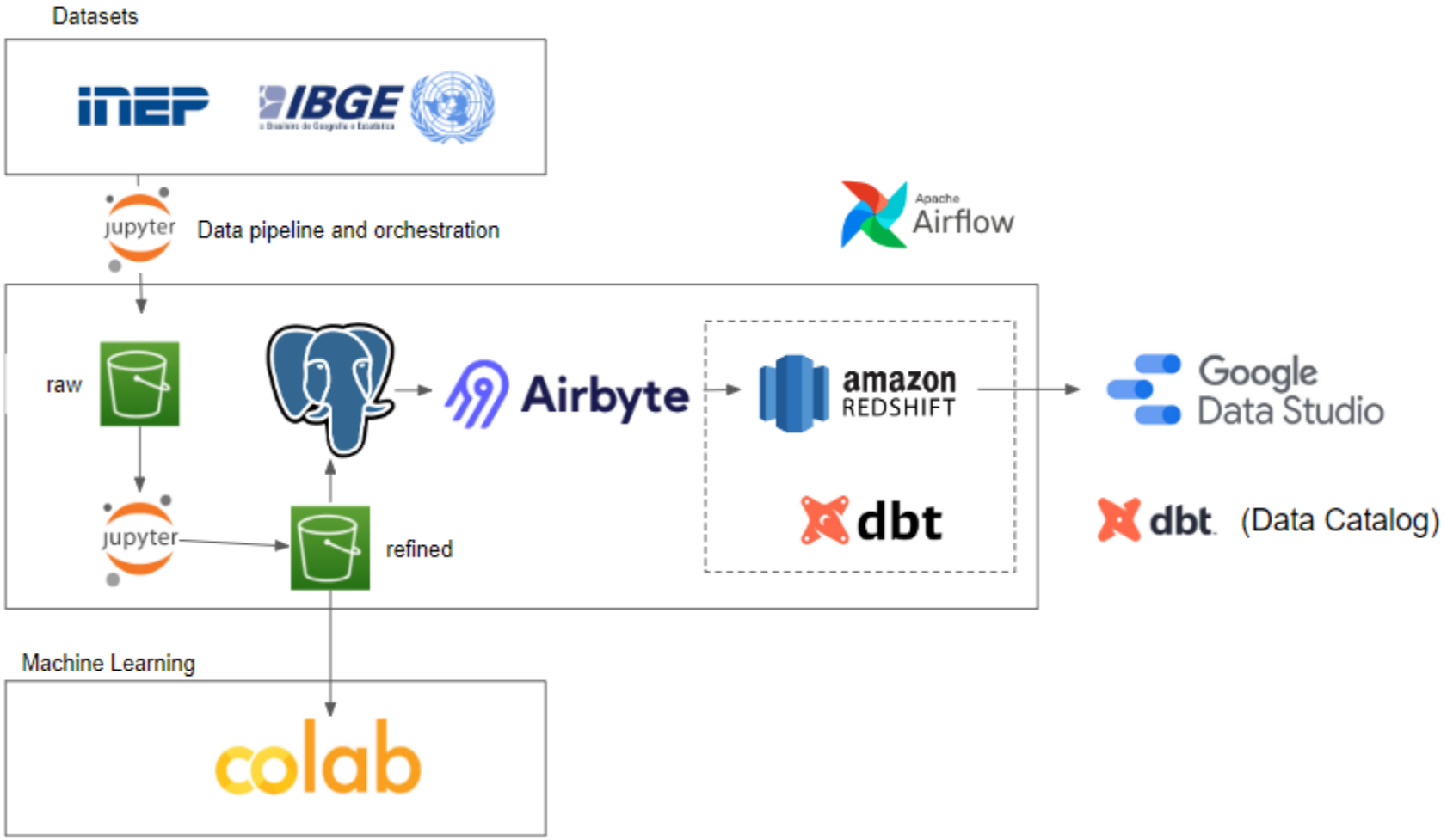
MBA Engenharia de Dados – Fase 5

Jackson Kolling
Marcelo Canabrava

Contexto

A explicação do desafio e resolução detalhada podem ser encontradas no seguinte repositório do GitHub.

Diagrama da Solução



Subindo dados no Postgres

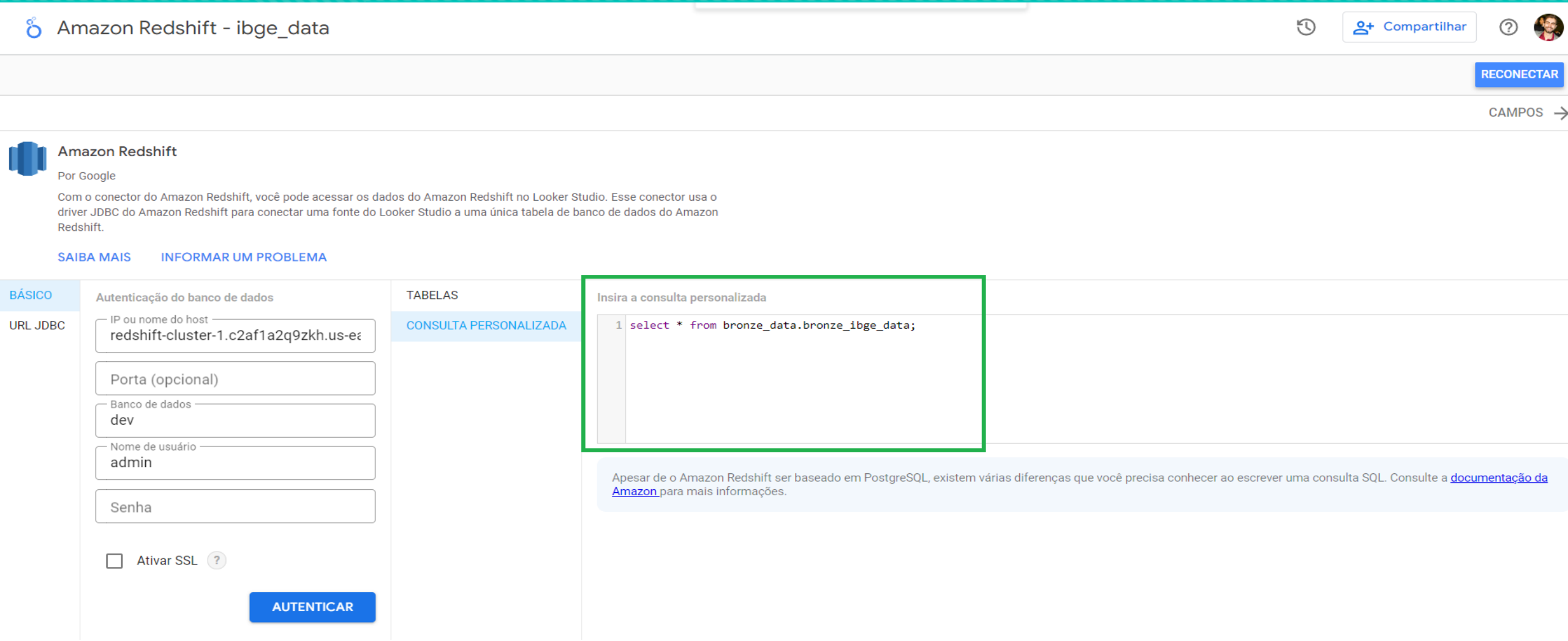
UPLOADING DATA TO POSTGRES

```
[ ]: current_time = datetime.now().time()
print("Starting upload at:", current_time)

# Specify the S3 bucket name and file names
bucket_name = 'inep-cleaned'
file_names = [
    'summary_MICRODADOS_ENEM_2020.csv',
    'summary_MICRODADOS_ENEM_2021.csv',
    'summary_MICRODADOS_ENEM_2022.csv'
]

# Set up a connection to your PostgreSQL RDS instance
host = 'database-2.ckuaogoaistw.us-east-1.rds.amazonaws.com'
port = 5432
database = 'postgres'
user = 'postgres'
password = 'postgres'

# Iterate through the specified file names and upload each DataFrame separately
for file_name in file_names:

    # Construct the public S3 URL
    s3_url = f'https://{bucket_name}.s3.amazonaws.com/{file_name}'

    # Download the CSV file using requests
    response = requests.get(s3_url)

    if response.status_code == 200:
        # Read the CSV data into a pandas DataFrame
        df = pd.read_csv(StringIO(response.text))

        # Set up a SQLAlchemy engine for the current DataFrame
        engine = create_engine(f'postgresql://{user}:{password}@{host}:{port}/{database}')

        table_name='inep_data'

        # Use Pandas to append data to the existing table
        df.to_sql(
            table_name,
            engine,
            if_exists='append',  # Append data to the existing table
            index=False  # Set to False if you don't want to include the DataFrame index as a column
        )
        current_time = datetime.now().time()
        print(f"Data from {file_name} appended to PostgreSQL at {current_time}")
    else:
        print(f"Failed to download {file_name}")

current_time = datetime.now().time()
print("Finishing upload at:", current_time)
```

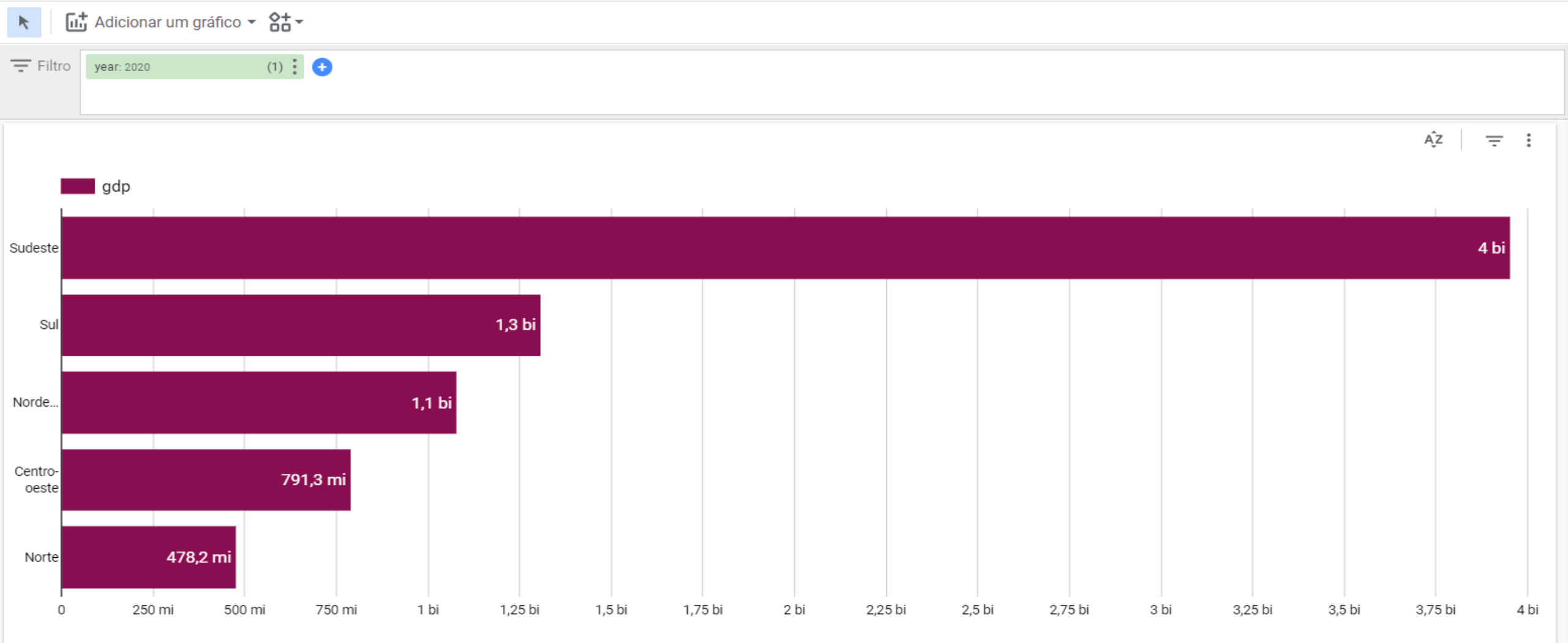
```
Starting upload at: 20:57:38.623277
Data from summary_MICRODADOS_ENEM_2020.csv appended to PostgreSQL
Data from summary_MICRODADOS_ENEM_2021.csv appended to PostgreSQL
```

- ETL para limpeza e refinamento dos dados entre buckets no S3
- Inserção dos dados em banco de dados Postgres
- Oportunidades de otimização do fluxo (tempo de upload de até 2 horas):
    - Redução do número de colunas utilizadas ao eliminar colunas como UF code e metadados dos candidatos
    - Utilização de máquina mais potente para processamento
    - Otimização do código usando parallalel processing

Superando dificuldades técnicas no GDS



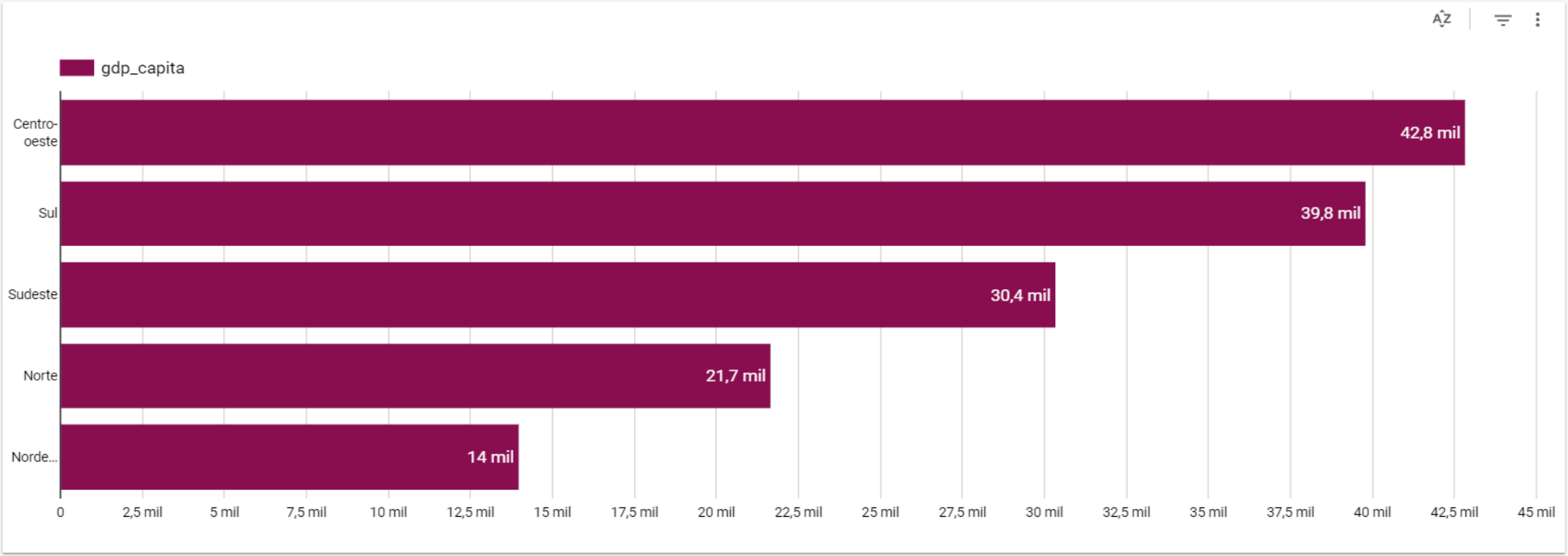
Explorando relatórios no GDS: PIB por Região



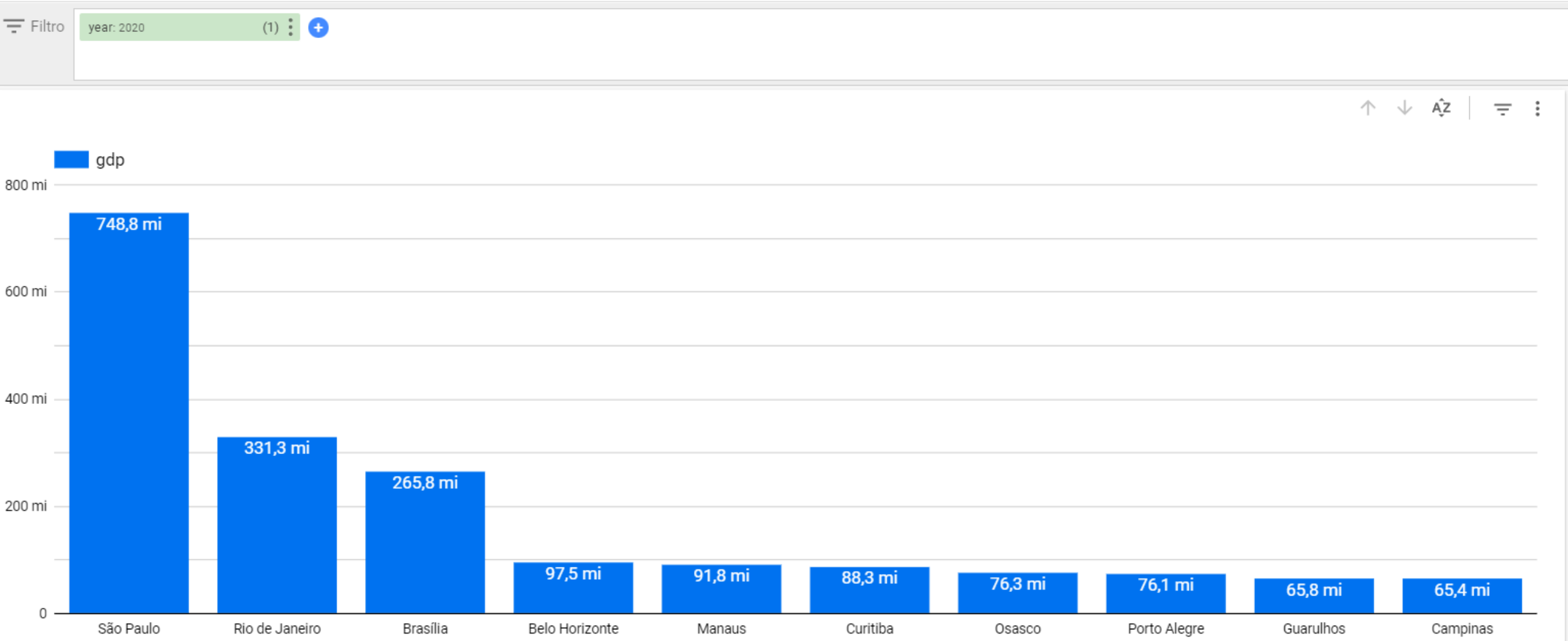
Explorando relatórios no GDS: PIB/capita por Região

Filtro

year: 2020 (1)

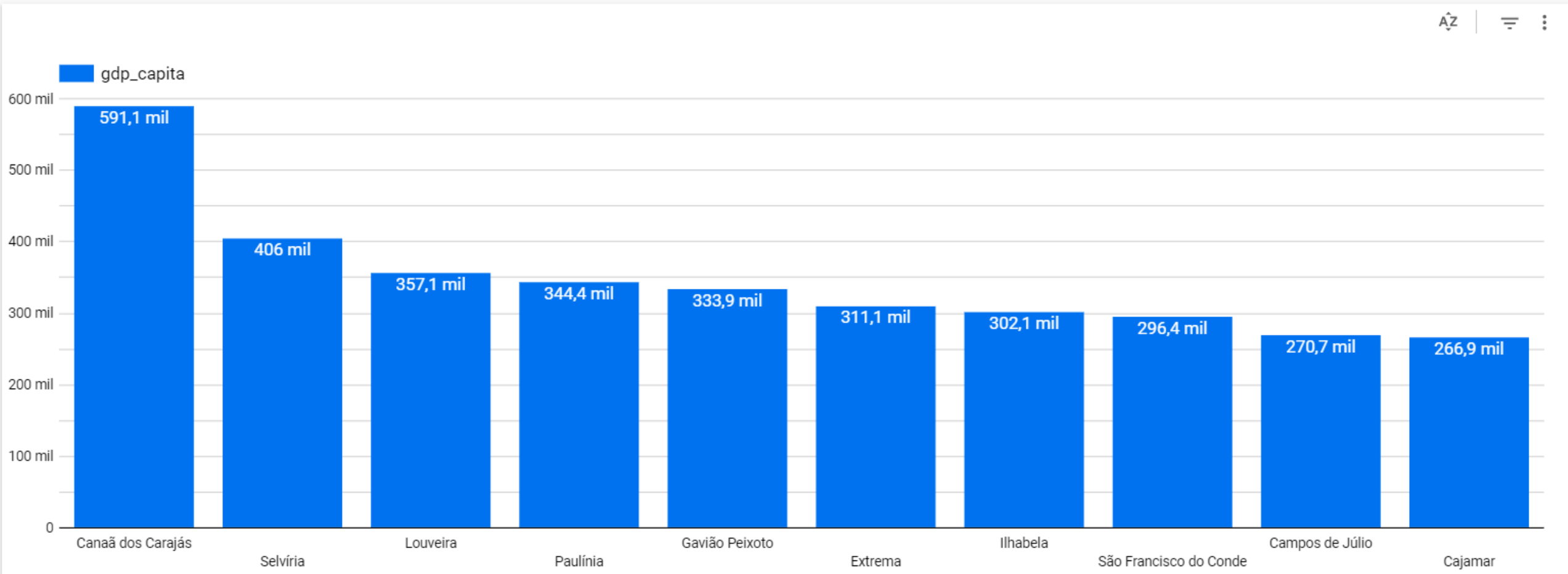


Explorando relatórios no GDS: PIB por cidade

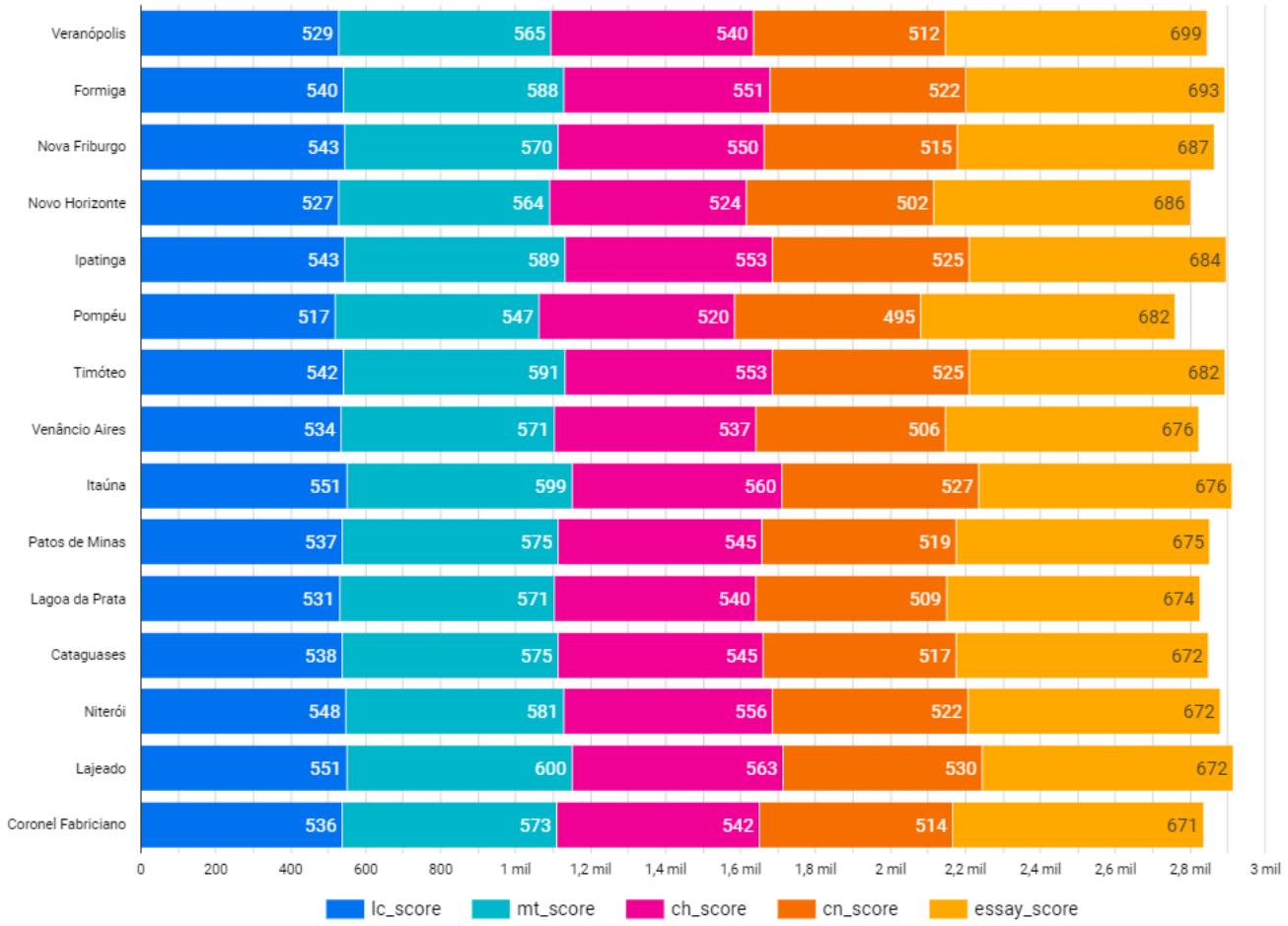


Explorando relatórios no GDS: PIB/capita por cidade

Filtro | year: 2020 (1)



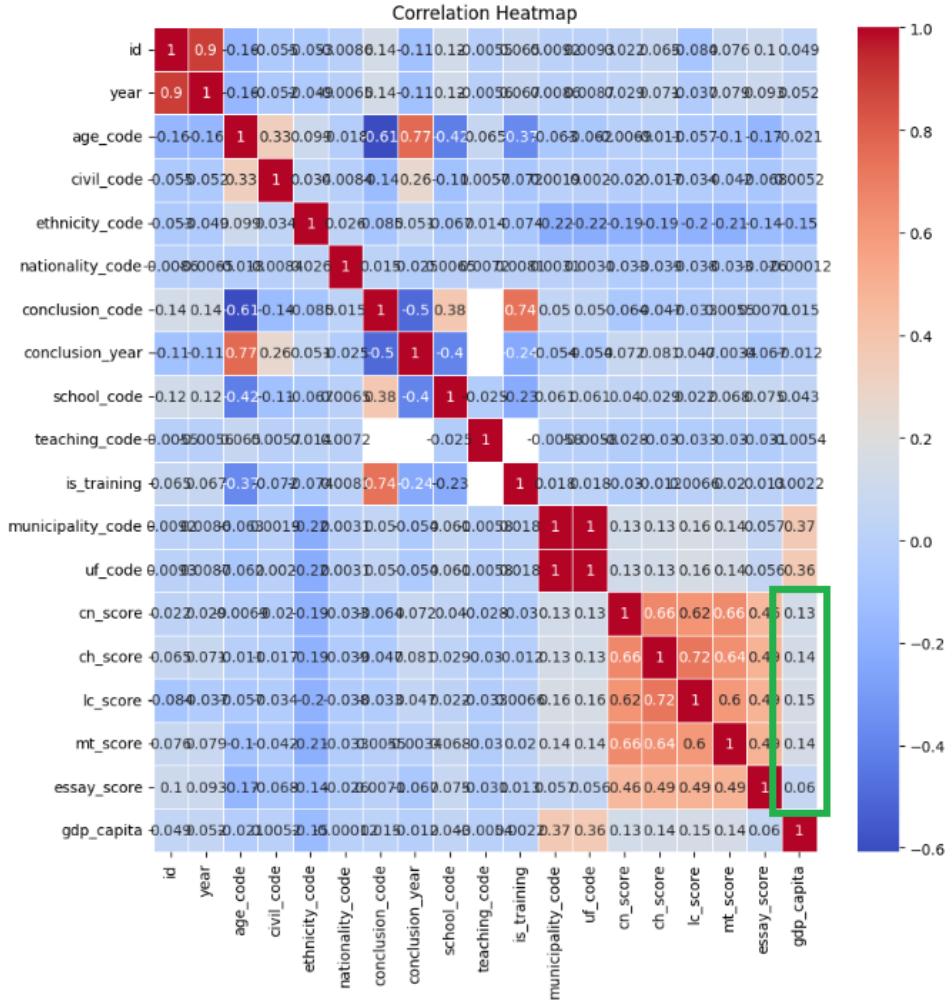
Explorando relatórios no GDS: Cidades com maiores notas de redação



Análise Preliminar

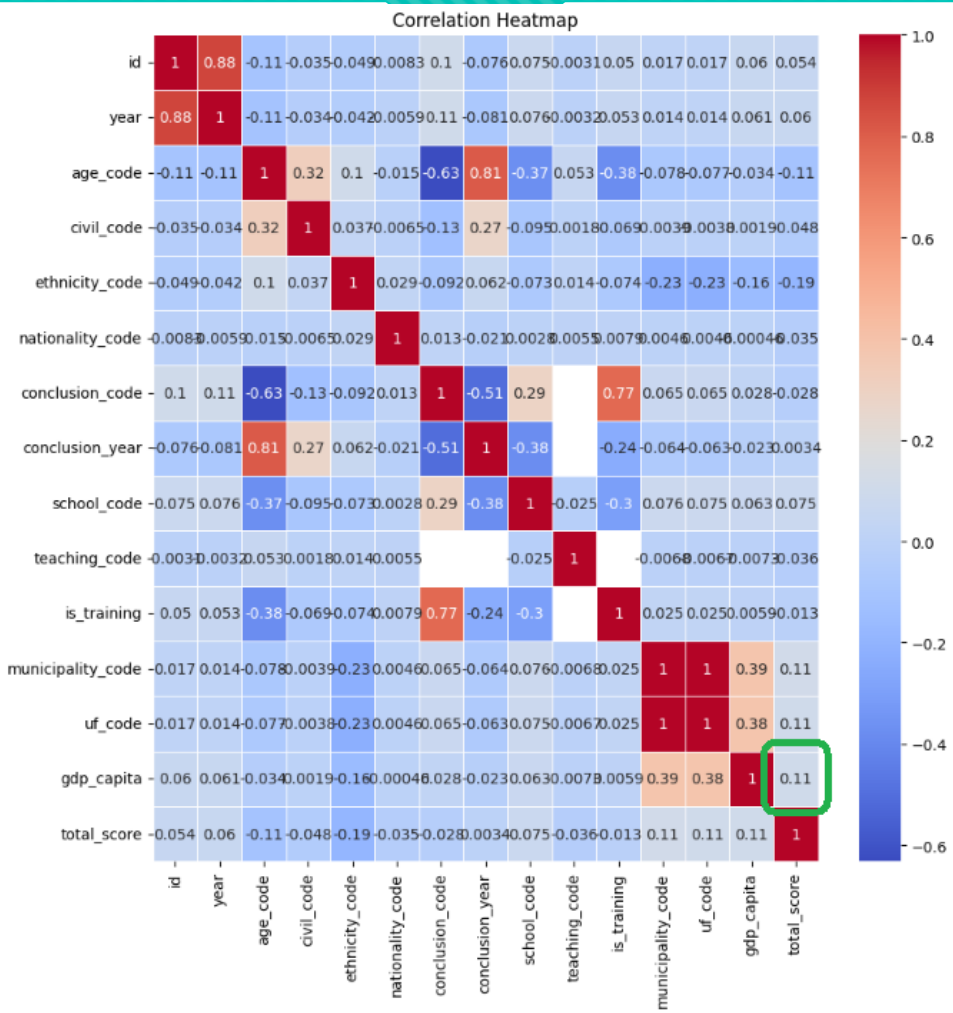
- Nenhuma cidade do maior PIB per capita ou maior PIB aparece no ranking das maiores notas de redação
- De forma análoga, tão pouco é esperado, intuitivamente, que as cidades que apresentaram maior PIB/capita estejam entre as melhores notas do ENEM

Análise estatística



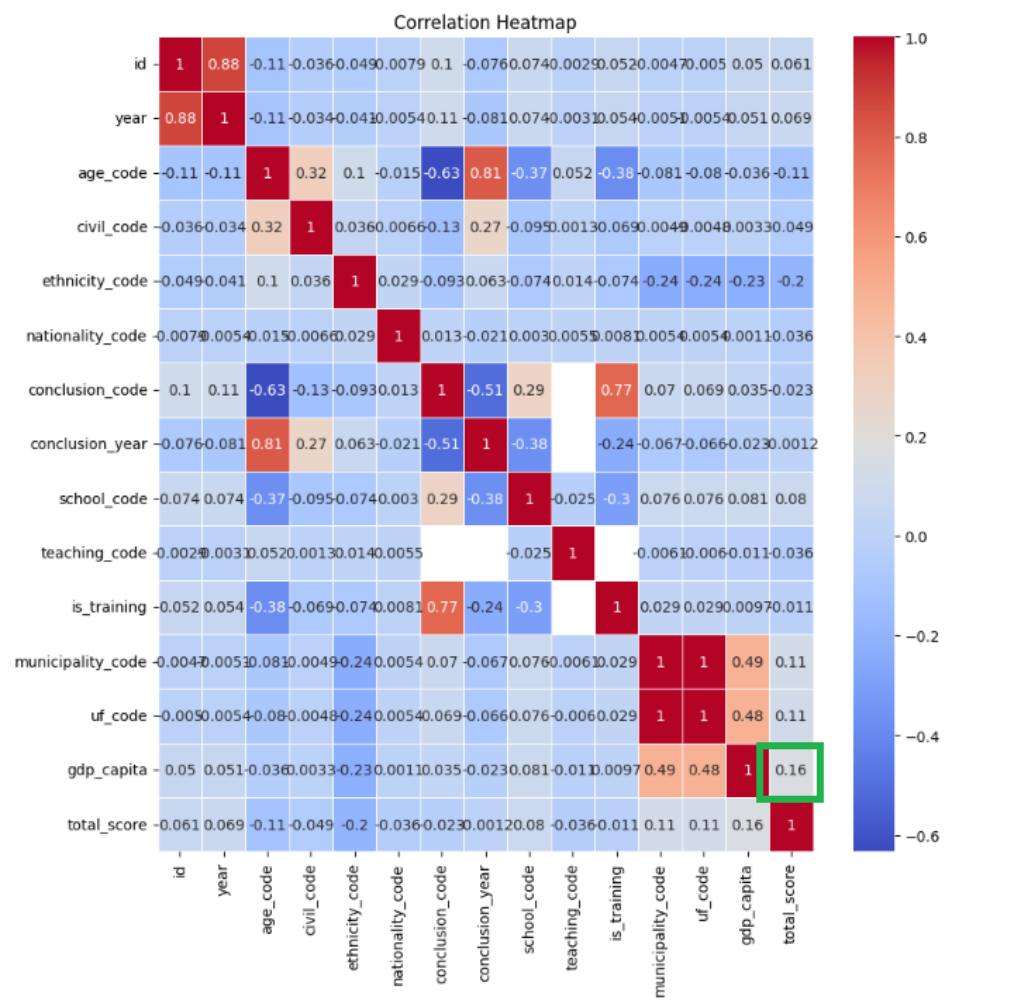
Correlação fraca entre PIB/capita e notas individuais das provas

Análise estatística



Correlação fraca entre PIB/capita e nota total das provas

Análise estatística



Remoção de outliers tem impacto significativo elevando a correlação de 0.11 para 0.16, mas insuficiente para apontar para o PIB/capita como uma variável de forte influência no PIB

Análise estatística

	uf_name	correlation
0	AC	0.130414
1	AL	0.113991
2	AM	0.205076
3	AP	0.140548
4	BA	0.057379
5	CE	0.132005
6	DF	NaN
7	ES	0.090830
8	GO	0.128179
9	MA	0.160313
10	MG	0.104432
11	MS	0.004210
12	MT	0.063293
13	PA	0.068226
14	PB	0.171203
15	PE	0.119964
16	PI	0.087279
17	PR	0.059280
18	RJ	0.102219
19	RN	0.156628
20	RO	0.047718
21	RR	0.117222
22	RS	0.106203
23	SC	0.038034
24	SE	0.127371
25	SP	0.076926
26	TO	0.109453

Estado com correlação mais forte foi o AM (0,20) e o mais fraco SC (0,03) com a maior parte do estados flutuando em torno de 0,10 e indicando que dificilmente algum deles poderia apresentar uma forte relação de PIB/Capita e nota do ENEM

Descobertas vão de encontro ao publicado anteriormente em estudos com a mesma finalidade

FGV EBAPE

RAP
REVISTA DE ADMINISTRAÇÃO PÚBLICA

Impacto das variáveis socioeconômicas no desempenho do Enem: uma análise espacial e sociológica

Rafael Oliveira Melo [1]
Anne Caroline de Freitas [2]
Eduardo de Rezende Francisco [3]
Marcelo Tadeu Motokane [4]

[1] Fundação Getulio Vargas / Educação Executiva, São Paulo / SP – Brasil
[2] Universidade de São Paulo / Programa de Pós-graduação Interunidades em Ensino de Ciências Instituto de Biociências, São Paulo/ SP - Brasil
[3] Fundação Getulio Vargas / Escola de Administração de Empresas de São Paulo, São Paulo / SP – Brasil
[4] Universidade de São Paulo / Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Departamento de Biologia, Ribeirão Preto / SP – Brasil

https://www.scielo.br/j/rap/a/ZHJFnmsrdgGH8cj6xHHwbKg/?lang=pt&format=pdf