Trabalho da A2 de Banco de Dados

Integrantes:

- Juan Belieni de Castro Araújo
- Matheus Medeiros Carvalho da Fonseca

Tema: evolução e estado atual do PIB brasileiro.

Escolhas dos dados e criação do banco de dados

GeoData BR

Para possibilitar a visualização dos dados geográficos, utilizamos o projeto do Geodata BR para obter os dados de geometria das cidades brasileiras. Esses dados são disponibilizados no formato GeoJson, que é permite uma maior portabilidade e compatibilidade com diferentes bibliotecas. No caso, iremos utilizar o GeoPandas para manipular as informações.

No entanto, os dados geográficos necessitam sofrer uma conversão para serem inseridos nos MySQL. Para isso, nós convertemos a coluna `geometry` que existe no GeoDataframe para um formato de texto, que consegue ser lido na hora da inserção no banco de dados.

Conversão dos dados:

```
data = (row["id"], row["name"], row["geometry"].wkt,
row["state_code"])
```

Inserção no banco de dados:

```
INSERT INTO cities (id, name, geom, state_code)
VALUES (%s, %s, ST_GeomFromText(%s), %s)
```

PIB brasileiro

O dataset escolhido para realizar as análises foi o Produto Interno Bruto do Brasil, disponibilizado pelo projeto da Base dos Dados. Foi escolhido essa base pela qualidade dos dados existentes nela, por já está todo padronizado e documentado. Além disso, é utilizado nesse dataset o mesmo número de identificação utilizado pelo projeto do Geodata BR, o ID do IBGE de 7 dígitos, o que facilita na hora de juntar as duas base de dados.

Por estar disponibilizado no formato CSV, a tarefa de inserir no banco criado foi trivial. Para essa finalidade, foi criado uma nova tabela que referencia a tabela que abarca os dados geoespaciais.

Análises

1. Tamanho do PIB em cada cidade brasileira
2. Participação de cada setor no PIB em dois anos distintos no Brasil
3. Quais cidades mais cobram imposto em relação ao pib? (> imposto / pib)

In [1]:
```
# Importando os pacotes necessários
import geopandas as gpd
import pandas as pd
import mysql.connector
import matplotlib.pyplot as plt
import matplotlib as mpl
```

In [2]:
```
# Conectando ao banco de dados
cnx = mysql.connector.connect(user="root", password="docker", database="brazi
```

Análise 1: tamanho do PIB em cada cidade brasileira

Análise 1.1: Brasil.

In [3]:
```
# Query para buscar os dados
query = f"""
select cities.name as cidade,
ano,
ST_AsText(geom) as geom,
pib
from pib
inner join cities
on pib.id_municipio = cities.id
where ano in (2002, 2018)
and id_municipio in (
select id_municipio
from pib
where ano in (2002, 2018)
group by id_municipio
having count(*) = 2
)
"""

# Executando a query com o Pandas
df1 = pd.read_sql(query, cnx)

# Convertendo o DataFrame para um GeoDataFrame
df1['geom'] = gpd.GeoSeries.from_wkt(df1['geom'])
df1 = gpd.GeoDataFrame(df1, geometry='geom')

df1.head()
```

Out[3]:

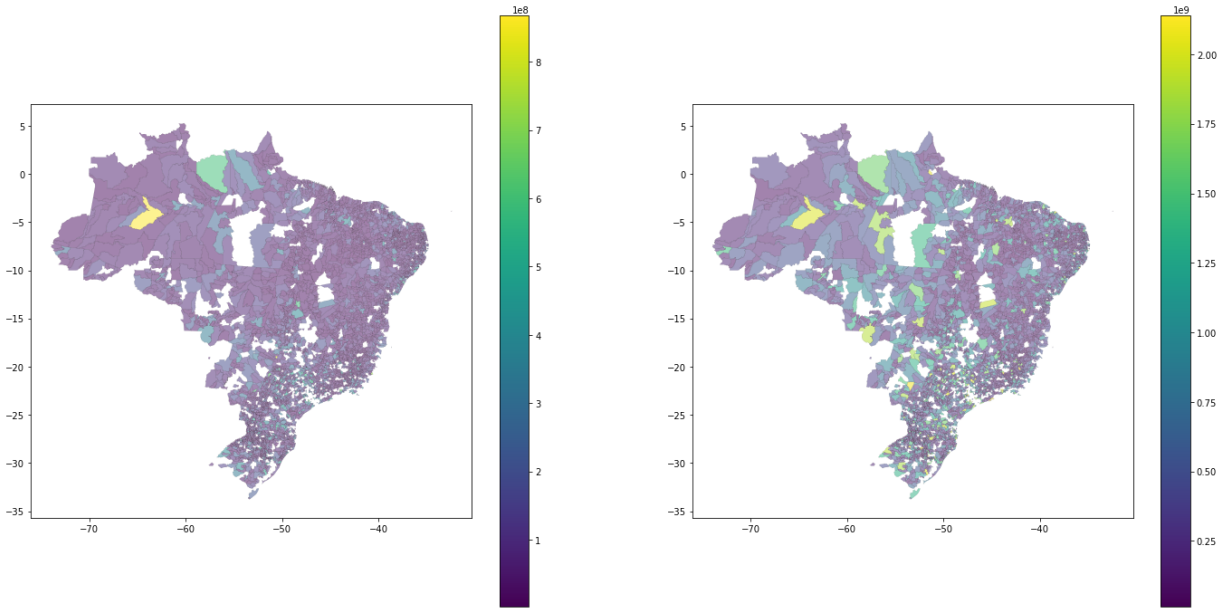
	cidade	ano	geom	pib
0	Alta Floresta D'Oeste	2002	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	111290995
1	Alta Floresta D'Oeste	2018	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	499305982
2	Cabixi	2002	POLYGON ((-60.39940 -13.45584, -60.40195 -13.4...	31767520
3	Cabixi	2018	POLYGON ((-60.39940 -13.45584, -60.40195 -13.4...	140502269
4	Cerejeiras	2002	POLYGON ((-61.50047 -13.00392, -61.47901 -13.0...	79173614

In [4]:
```
df1_2002 = df1[df1['ano'] == 2002]
pib = df1_2002['pib']

df1_2018 = df1[df1['ano'] == 2018]
pib = df1_2018['pib']

# Plotando 2 mapas com o PIB
```

```
fig, ax = plt.subplots(1, 2, figsize=(24, 12))

df1_2002.plot(

    column='pib',
    legend=True,
    ax=ax[0],
    edgecolor='black',
    linewidth=0.1,
    alpha=0.5
)

df1_2018.plot(
    column='pib',
    legend=True,
    ax=ax[1],
    edgecolor='black',
    linewidth=0.1,
    alpha=0.5
)
```

Out[4]:

```
<AxesSubplot:>
```



Análise 1.2: Rio Grande do Sul

In [5]:

```
# Query para buscar os dados
query = f"""
select cities.name as cidade,
ano,
ST_AsText(geom) as geom,
pib
from pib
inner join cities
on pib.id_municipio = cities.id
where ano in (2002, 2018)
and state_code = 'RS'
and id_municipio in (
select id_municipio
from pib
where ano in (2002, 2018)
group by id_municipio
having count(*) = 2
)
```

```
"""
# Executando a query com o Pandas
df1 = pd.read_sql(query, cnx)

# Convertendo o DataFrame para um GeoDataFrame
df1['geom'] = gpd.GeoSeries.from_wkt(df1['geom'])
df1 = gpd.GeoDataFrame(df1, geometry='geom')

df1.head()
```

Out[5]:

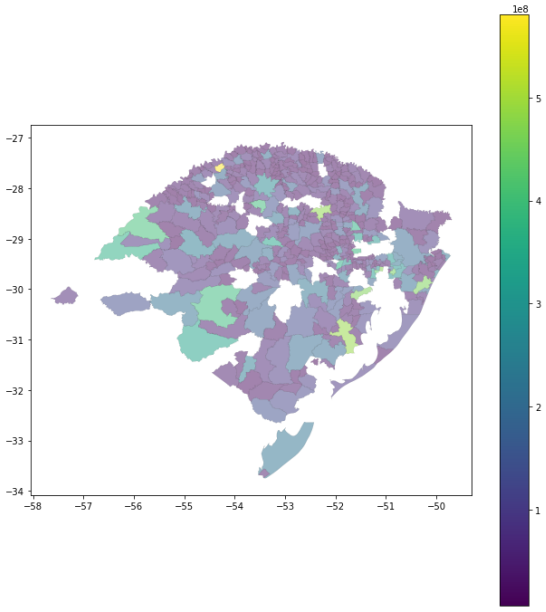
	cidade	ano	geom	pib
0	Aceguá	2002	POLYGON ((-54.11276 -31.42931, -54.09821 -31.4...	43719950
1	Aceguá	2018	POLYGON ((-54.11276 -31.42931, -54.09821 -31.4...	248093560
2	Água Santa	2002	POLYGON ((-52.04263 -28.11703, -52.03489 -28.1...	36196634
3	Água Santa	2018	POLYGON ((-52.04263 -28.11703, -52.03489 -28.1...	330831762
4	Agudo	2002	POLYGON ((-53.25560 -29.44736, -53.23478 -29.4...	126006848

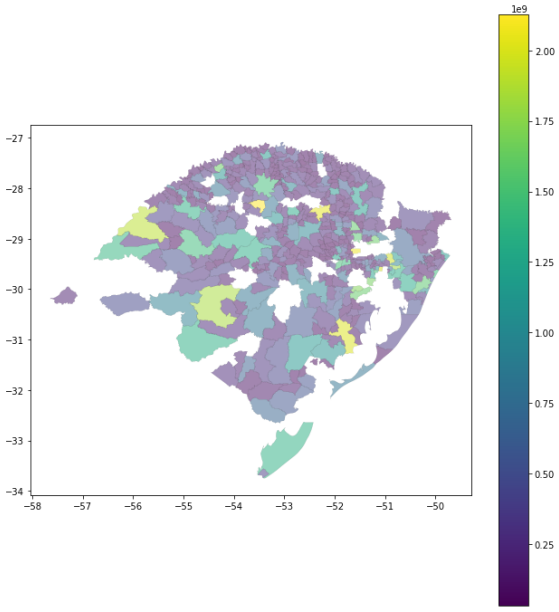
In [7]:

```
df1_2002 = df1[df1['ano'] == 2002]

df1_2018 = df1[df1['ano'] == 2018]

# Plotando 2 mapas com o PIB
fig, ax = plt.subplots(1, 2, figsize=(24, 12))

df1_2002.plot(
    column='pib',
    legend=True,
    ax=ax[0],
    edgecolor='black',
    linewidth=0.1,
    alpha=0.5
)

df1_2018.plot(
    column='pib',
    legend=True,
    ax=ax[1],
    edgecolor='black',
    linewidth=0.1,
    alpha=0.5
)
```

Out[7]:

```
<AxesSubplot:>
```





Análise 2: participação de cada setor no PIB em dois anos distintos no Brasil

In [8]:

```
# Query para buscar os dados
query = f"""
select cities.name as cidade,
ano,
ST_AsText(geom) as geom,
(va_agropecuaria / va) * 100 as agropecuaria,
(va_industria / va) * 100 as industria,
(va_servicos / va) * 100 as servicos
from pib
inner join cities
on pib.id_municipio = cities.id
where ano in (2002, 2018)
and id_municipio in (
select id_municipio
from pib
where ano in (2002, 2018)
group by id_municipio
having count(*) = 2
)
"""

# Executando a query com o Pandas
df2 = pd.read_sql(query, cnx)

# Convertendo o DataFrame para um GeoDataFrame
df2['geom'] = gpd.GeoSeries.from_wkt(df2['geom'])
df2 = gpd.GeoDataFrame(df2, geometry='geom')

df2.head()
```

Out[8]:

	cidade	ano	geom	agropecuaria	industria	servicos
0	Alta Floresta D'Oeste	2002	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	26.0389	9.0387	23.7620
1	Alta Floresta D'Oeste	2018	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	35.2179	5.5980	26.2210
2	Cabixi	2002	POLYGON ((-60.39940 -13.45584, -60.40195 -13.4...	36.4012	6.5617	17.1537

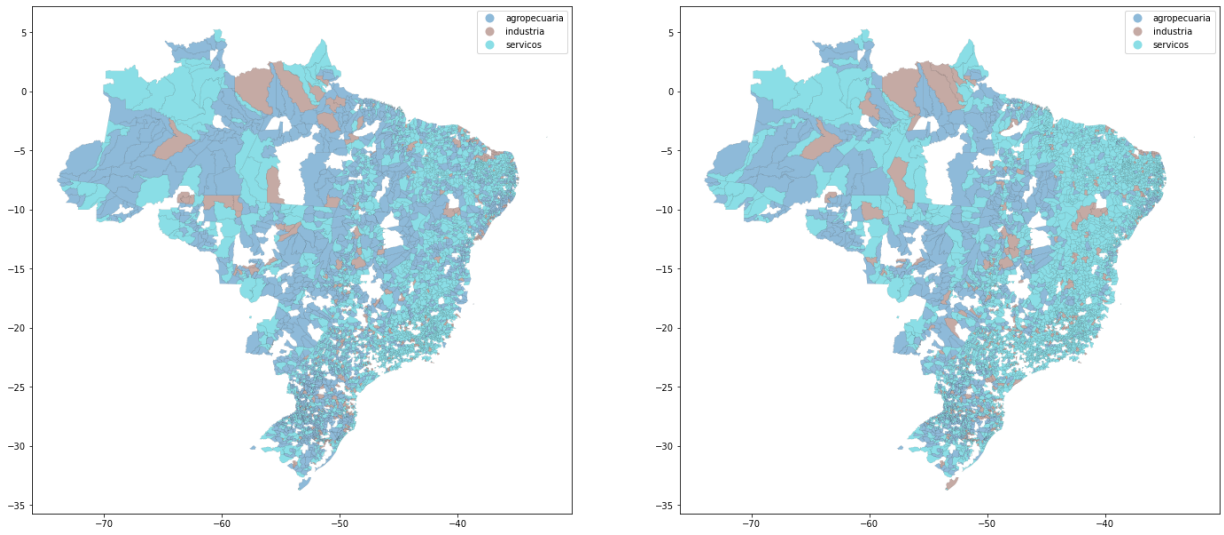
	cidade	ano	geom	agropecuaria	industria	servicos
3	Cabixi	2018	POLYGON ((-60.39940 -13.45584, -60.40195 -13.4...	46.2270	3.8405	18.6655
4	Cerejeiras	2002	POLYGON ((-61.50047 -13.00392, -61.47901 -13.0...	15.8988	9.3719	35.6560

In [9]:

```
# Criando uma coluna nova para o setor com mais participação
df2['maior_setor'] = df2[['agropecuaria', 'industria', 'servicos']].idxmax(ax

# Plotando 2 mapas com a maior setor em cada ano
fig, ax = plt.subplots(1, 2, figsize=(24, 12))

df2[df2['ano'] == 2002].plot(
    column='maior_setor',
    legend=True,
    ax=ax[0],
    edgecolor='black',
    linewidth=0.1,
    alpha=0.5
)

df2[df2['ano'] == 2018].plot(
    column='maior_setor',
    legend=True,
    ax=ax[1],
    edgecolor='black',
    linewidth=0.1,
    alpha=0.5
)
```

Out[9]:

```
<AxesSubplot:>
```



Análise 3: quais cidades mais cobram imposto em relação ao pib? (> imposto / pib)

In [10]:

```
# Query para buscar os dados
query = f"""
select cities.name as cidade,
ano,
ST_AsText(geom) as geom,
(impostos_liquidos / pib) * 100 as impostos,
impostos_liquidos
```

```
from pib
inner join cities
on pib.id_municipio = cities.id
"""

# Executando a query com o Pandas
df3 = pd.read_sql(query, cnx)

# Convertendo o DataFrame para um GeoDataFrame
df3['geom'] = gpd.GeoSeries.from_wkt(df3['geom'])
df3 = gpd.GeoDataFrame(df3, geometry='geom')

df3.head()
```

Out[10]:

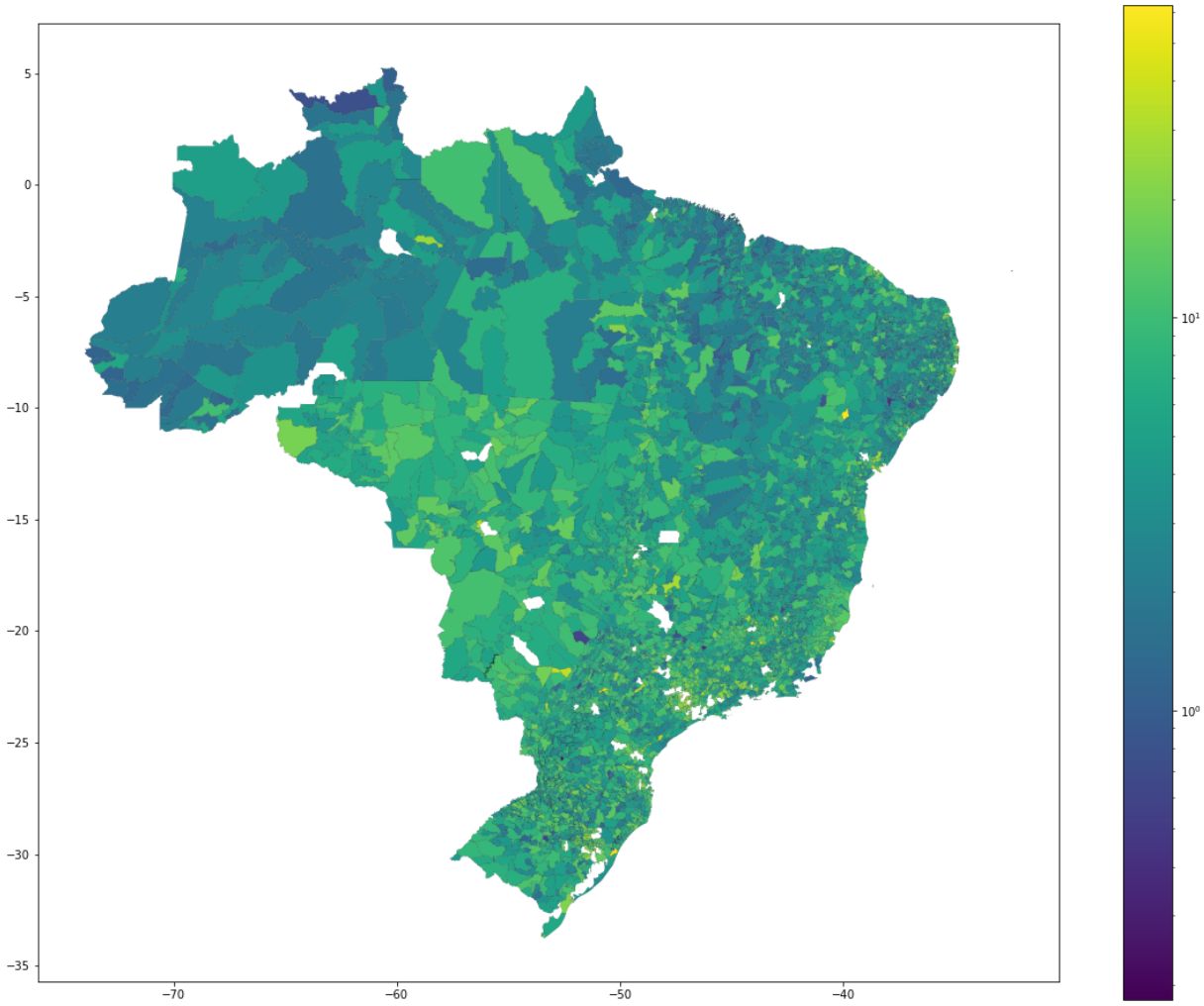
	cidade	ano	geom	impostos	impostos_liquidos
0	Alta Floresta D'Oeste	2002	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	6.7834	7549266
1	Alta Floresta D'Oeste	2003	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	7.3394	10511613
2	Alta Floresta D'Oeste	2004	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	7.0228	12219047
3	Alta Floresta D'Oeste	2005	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	7.7389	12933774
4	Alta Floresta D'Oeste	2006	POLYGON ((-62.18209 -11.86686, -62.16230 -11.8...	8.0974	13668803

In [11]:

```
# Plotando o mapa com a porcentagem de impostos
df3_2002 = df3[df3['ano'] == 2002]
impostos = df3_2002['impostos']

df3[df3['ano'] == 2002].plot(
    column='impostos',
    legend=True,
    figsize=(20, 16),
    edgecolor='black',
    linewidth=0.1,
    norm=mpl.colors.LogNorm(vmin=impostos.min(), vmax=impostos.max())
)
```

Out[11]:

```
<AxesSubplot:>
```



Conclusão sobre as análises

Com o projeto Geodata BR e com o dataset Produto Interno Bruto do Brasil da Base dos Dados, foi possivel realizar as análises demonstradas acima. Podemos inferir pela analise dos 3 gráficos que o Brasil de 2002 para 2018 teve um significativo aumento no PIB; se tornou um pais mais voltado para o 3° setor; e que a relação do imposto sobre o PIB em 2002 foi, na maior parte do território parecida, pois o país está em sua maior parte em uma grande escala de verde.

Analisando especificamente os gráficos, no 1° percebemos que as regiões Centro-Oeste e Sul tiveram uma grande evolução do PIB de 2002 para 2018, em especial Rio Grande do Sul (RS), com muitos municípios se desenvolvendo nesse período.

Se analisarmos com mais cuidado o 2° gráfico podemos perceber que a ideia de que o Brasil é um país focado majoritariamente em Agropecuaria não condiz com a absoluta realidade, uma vez que a agropecuaria está cada vez mais concentrada nas regiões Centro-Oeste e Norte. Observa-se no gráfico também que o setor de Serviços expande-se cada vez mais para todo o Brasil, mais especificamente no Nordeste do país.

Observando o 3° gráfico percebemos que existem duas cores predominantes, um verde mais vívido (concentrado no Centro-Oeste e em parte no Sudeste) e um verde levemente azulado (concentrado no Norte e Nordeste), indicando que o PIB nas Regiões Norte e Nordeste tem maior tributação em relação ao PIB.

Conclusão geral

Nosso trabalho foi dividido em três partes importantes:

1. Coleta e manipulação dos dados geoespaciais do Brasil
2. Coleta e manipulação de dados sobre o PIB brasileiro
3. Análise geral dos dados encontrados

Na primeira parte, tivemos mais dificuldade em encontrar uma forma limpa e simples de transformar e colocar os dados geoespaciais em um banco de dados MySQL. Como foi mostrado, utilizamos funções nativas do próprio MySQL e GeoPandas para essa tarefa. No fim, construímos duas tabelas: uma dos estados brasileiros e outro com as cidades, que foi populada com os dados extraídos do GeoJson por meio do GeoPandas.

Na segunda parte, tinhamos o banco de dados já criado com os dados geoespaciais. Para não misturar as informações já existentes com os dados que iríamos inserir, criamos uma nova tabela que iria armazenar o PIB de cada cidade em cada ano. Aqui, não houve tantas dificuldades.

Na última etapa, realizamos as análises com o objetivo de observar a evolução do PIB brasileiro, e como outras informações se relacionavam com ele. Dessa maneira, construímos algumas queries SQL para atingir o objetivo almejado. Utilizamos então o Pandas para resgatar os dados, e depois o GeoPandas para lidar com a coluna dos dados geoespaciais. Terminamos então com vários GeoDataframes que serviram de base para a construção de diversas visualizações.

Concluindo, tanto o GeoPandas quanto o conhecimento da linguagem SQL foram essenciais para o desenvolvimento do trabalho.