Módulo | Análise de Dados: Data Wrangling II

Caderno de **Exercícios**
Professor André Perez

Tópicos

1. Agregação e Ordenação;
2. Combinação;
3. Técnicas Avançadas.

Exercícios

Neste exercícios, vamos trabalhar com dados geográficos, demográficos e econômicos do Brasil. Vamos manipular e combinar dados de duas frentes distintas para poder responder perguntas de negócios.

1. Data Wrangling

1.1. Estados

O arquivo `estados-bruto.xml` contém informações sobre estados (nome, sigla e região). **Carregue-o na máquina virtual do Google Colab**. Um arquivo do tipo XML é similar a um arquivo do tipo HTML, exemplo do estado do Acre:

```
<ESTADO>
  <ID>1</ID>
  <NOME>ACRE</NOME>
  <IDCAPITAL>16</IDCAPITAL>
  <SIGLA>AC</SIGLA>
  <REGIAO>NORTE</REGIAO>
</ESTADO>
```

Utilize o pacote Python `beautifulsoup4` para extrair os dados do arquivo `estados-bruto.xml` providenciado. Salve os dados extraidos no arquivo `estados-limpo.csv` separado por `;`. Exemplo das três primeiras linhas mais o cabeçalho:

```
estado;sigla;regiao
ACRE;AC;NORTE
ALAGOAS;AL;NORDESTE
AMAPA;AP;NORTE
```

Dica: Utilize o parser de xml chamado `lxml` do `beautifulsoup4`.

In [ ]:
```
# ler o arquivo estados-bruto.xml, utilize o xml parser chamado lxml

from bs4 import BeautifulSoup

NOME_ARQUIVO_FONTE = 'estados-bruto.xml'

fonte = BeautifulSoup(..., 'lxml')
```

In [ ]:
```
# visualize os resultados

fonte
```

In [ ]:
```
# manipule os dados

# continue o codigo aqui
```

In [ ]:
```
# escrever o conteudo extraido no arquivo estados-limpo.csv separados por ;

NOME_ARQUIVO_DESTINO = 'estados-limpo.csv'

# continue o codigo aqui
```

1.2. Cidades

O arquivo `cidades-bruto.csv` contém informações demográficas e socioeconomicas das cidades do Brasil. **Carregue-o na máquina virtual do Google Colab**. Utilize o pacote Python `pandas` para extrair os dados do arquivo `cidades-bruto.xml` providenciado. Seguindo as seguintes especificações:

1. Apenas dados do censo de 2010;
2. Apenas as colunas UF, Nome, PIB, Pop_est_2009 e PIB_percapita.

Salve os dados extraidos no arquivo `cidades-limpo.csv` separado por `;`. Exemplo das três primeiras linhas mais o cabeçalho:

```
estado;cidade;populacao;pib;pib_percapita
BAHIA;TREMEDAL;18433;57883.9921875;3140.23999023
RIO GRANDE DO SUL;TURUÇU;4000;45723875;11430.96972656
ESPIRITO SANTO;VITÓRIA;320156;19782628;61790.58984375
```

In [ ]:
```
# ler o arquivo cidades-bruto.csv

import pandas as pd

NOME_ARQUIVO_FONTE = 'cidades-bruto.csv'

fonte = ... # continue o codigo aqui
```

In [ ]:
```
# visualize os resultados

# continue o codigo aqui
```

In [ ]:
```
# manipule os dados

# continue o codigo aqui
```

In [ ]:
```
# escrever o conteudo extraido no arquivo cidades-limpo.csv separados por ;

NOME_ARQUIVO_DESTINO = 'cidades-limpo.csv'

# continue o codigo aqui
```

1.3. Brasil

Utilize o pacote Python `pandas` para combinar os dados do arquivo `estados-bruto.csv` com os dados do arquivo `cidades-bruto.csv` em um único dataframe. Escolha a coluna e o método de combinação de tal forma que **não haja perda de dados** no processo (não produzirá valores nulos `NaN`). Salve os dados do dataframe no arquivo `brasil.csv`

In [ ]:
```
# solução do exercício 1.3
```

---

2. Data Analytics

2.1. DataFrame

Utilize o pacote Python `pandas` para carregar o arquivo `brasil.csv` no dataframe `brasil_df`.

In [ ]:
```
# solução do exercício 2.1
```

2.2. Analise

Utilize o dataframe `brasil_df` para responder as seguintes perguntas de negócio:

- Quais são as 10 cidades mais populosas do Brasil?

In [ ]:
```
# código para responder a pergunta
```

- Quais são as 5 cidades com a menor PIB per capita da região nordeste?

In [ ]:
```
# código para responder a pergunta
```

- Quais são as 15 cidades com maior PIB do do estado de São Paulo?

In [ ]:
```

```

```
# código para responder a pergunta
```

- Qual é o PIB do estado de Santa Catarina?

In [ ]:
```
# código para responder a pergunta
```

- Qual é o população da região sul?

In [ ]:
```
# código para responder a pergunta
```

- Qual é o PIB per capito médio das cidades do Mato Grosso do Sul?

In [ ]:
```
# código para responder a pergunta
```

- Qual é a população do Brasil?

In [ ]:
```
# código para responder a pergunta
```

2.3. Visualização

Utilize o dataframe `brasil_df` para gerar as seguintes visualizações.

- Gere um gráfico de barras com as 10 cidades menos populosas do Brasil.

In [ ]:
```
# código para gerar a visualização
```

- Gere um gráfico de pizza com a proporção da população do Brasil por região.

In [ ]:
```
# código para gerar a visualização
```