escola britânica de artes criativas & tecnologia

# **Módulo** | Análise de Dados: Data Wrangling I

Caderno de **Exercícios**
Professor André Perez

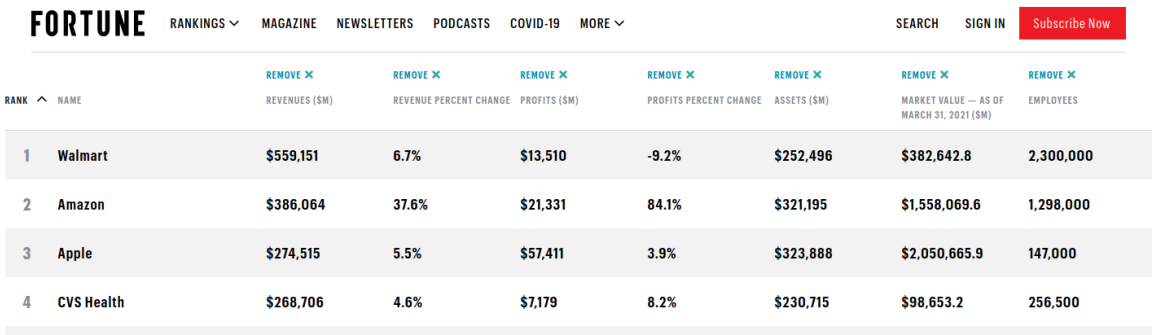
# Tópicos

1. DataFrame Pandas;
2. Selecão e Filtros;
3. Inserção, Deleção e Atualização.

# Exercícios

## 1. Fortune 500

O Fortune 500 é uma ranking anual compilado pela revista Fortune das 500 maiores empresas dos EUA.




| RANK | NAME | REVENUES ($M) | REVENUE PERCENT CHANGE | PROFITS ($M) | PROFITS PERCENT CHANGE | ASSETS ($M) | MARKET VALUE — AS OF MARCH 31, 2021 ($M) | EMPLOYEES |
|---|---|---|---|---|---|---|---|---|
| 1 | Walmart | $559,151 | 6.7% | $13,510 | -9.2% | $252,496 | $382,642.8 | 2,300,000 |
| 2 | Amazon | $386,064 | 37.6% | $21,331 | 84.1% | $321,195 | $1,558,069.6 | 1,298,000 |
| 3 | Apple | $274,515 | 5.5% | $57,411 | 3.9% | $323,888 | $2,050,665.9 | 147,000 |
| 4 | CVS Health | $268,706 | 4.6% | $7,179 | 8.2% | $230,715 | $98,653.2 | 256,500 |

O arquivo `fortune.html` contem o código fonte em HTML da página com as 100 primeiras empresas do ranking. **Carregue-o na máquina virtual do Google Colab**.

Utilize o pacote Python `beautifulsoup4` para extrair todas as 100 empresas do arquivo `fortune.html` providenciado. Salve os dados extraidos no arquivo `fortune.csv` separado por `;`. Exemplo das três primeiras linhas (sem o cabeçalho):

```
1;Walmart;$559,151;6.7%;$13,510;-9.2%;$252,496;$382,642.8;2,300,000
```

```
2;Amazon;$386,064;37.6%;$21,331;84.1%;$321,195;$1,558,069.6;1,298,000

3;Apple;$274,515;5.5%;$57,411;3.9%;$323,888;$2,050,665.9;147,000
```

**Dica**: Utilize os código abaixo para ajudar na extração dos dados.

**Dica**: Você não precisa extrair o cabeçalho da tabela, utilize o nome das colunas armazenados na variável `header` abaixo.

In [ ]:
```
# ler o arquivo fortune.html

from bs4 import BeautifulSoup

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

In [ ]:
```
# extrair as linhas da tabela

tabela = pagina.find('div', {'class': 'rt-table'})
linhas = tabela.find('div', {'class': 'rt-tbody'})
```

In [ ]:
```
# extrair o conteudo das linhas da tabela

for linha in linhas:
  colunas = linha.find('div', {'role': 'row'})
  ... # continue o codigo aqui
```

In [ ]:
```
# escrever o conteudo extraido no arquivo fortune.csv
# utilize a variavel header para construir a o cabecalho do arquivo csv

header = [
  'rank',
  'name',
  'revenues',
  'revenues-percent-change',
  'profits',
  'profits-percent-change',
  'assets',
  'market-value',
  'employees'
]

... # continue o codigo aqui
```

---

# 2. Data Wrangling

## 2.1. Criando o DataFrame

Crie o dataframe Pandas na variável `fortune_df` através da leitura do arquivo `fortune.csv`

In [ ]:
```
fortune_df = ... # continue o codigo aqui
```

## 2.2. Explorando o DataFrame

Utilizando os métodos vistos em aula, explore o dataframe.

- Liste as 10 primeiras linhas do dataframe:

In [ ]:
```
... # continue o codigo aqui
```

- Liste os tipos de dados armazenados na coluna do dataframe:

In [ ]:
```
... # continue o codigo aqui
```

- Liste o numero de linhas e colunas do dataframe:

In [ ]:
```
... # continue o codigo aqui
```

## 2.3. Limpando o DataFrame

Grande parte das colunas numéricas (exceto a coluna `ranking` e `employees`) possuem o caracter `$` ou `%` que as classificam com o tipo `object` (ou `str` do Python) ao invés do tipo correto como `int` ou `float`. Utilizando os métodos de atualizam, remova os caracteres das linhas das colunas numéricas.

In [ ]:
```
... # continue o codigo aqui
```

**Dica**: Você pode utilizar qualquer método de atualização, eu recomendo o uso do método `apply`.

## 2.4. Salvando o DataFrame

Utilize o método `to_csv` para salvar o dataframe `fortune_df` no arquivo `fortune-limpo.csv`.

In [ ]:
```
... # continue o codigo aqui
```

**Dica**: Confira a documentação oficial do método `to_csv` neste link.

---