

Programação

Módulo pandas

Fátima Leal

DCT DEPARTAMENTO CIÊNCIA
E TECNOLOGIA

Até agora ...

- Módulos em Python
 - Funções recursivas
 - Funções de ordem superior
 - Geradores e compreensões de listas
 - Escrita e leitura de ficheiros
 - Diferentes tipos de ficheiros
 - Gráficos
-
- Nesta aula:
 - Módulo pandas

Pandas

- Permite manipular tabelas denominadas por DataFrames
- Permite atribuir nomes a linhas/columnas
- Oferece operações de manipulação e transformação
- Vasto suporte para lidar com dados em falta

- Dois tipos
 - Séries
 - DataFrames

A	B	C	D	A
10	50	23	70	34

Índice	Age	Gender	Rating
Steve	32	Male	3.45
Lia	28	Female	4.6
Steve	45	Male	3.9
Katie	38	Female	2.78

Pandas

- Leitura de ficheiros csv.

```
import pandas as pd

df = pd.read_csv("data.csv")
print(df)
```

- Leitura de ficheiros json

```
import pandas as pd

df = pd.read_json("data.json")
print(df)
```

Pandas: Ver dados

- Aceder a Colunas

```
df['Calories']
```

- Linhas e colunas com o loc[] e iloc[]

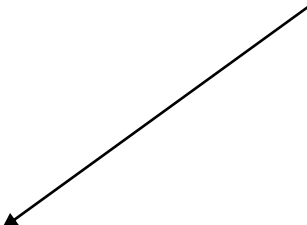
```
df.loc[df['Calories'] >= 1000]
```

```
df.iloc[0:2, 0:2]
```

- Modificar valores mediante condições

```
df.loc[df['Calories'] >= 1000, 'Maxpulse']=1000
```

Coluna a
modificar



Pandas: Análise dos dados

- Primeiras linhas

```
df.head(5)
```

- Últimas linhas

```
df.tail(5)
```

- Mais informação

```
df.info()
```

- Valores Null – Para análise de dados os valores null devem ser limpos. O pandas contém os métodos necessários para manipular valores null

Pandas: Cleaning

- Preencher valores NULL e NaN.

```
df.fillna(130, inplace = True)
```

Inplace= True muda a dataFrame original

```
df["Calories"].fillna(130, inplace = True)
```

- Remover as linhas com valores NULL ou NaN

```
new_df = df.dropna()
```

```
df.dropna(inplace = True)
```

Pandas

- Medidas estatísticas por colunas ou linhas

- Mean
- Median
- Max
- Min
- Std

```
df = pd.read_json("data.json")
mediaCalorias =
df['Calories'].mean()
print(mediaCalorias)
```

- Operações com colunas ou linhas

- Soma
- Subtração
- Divisão
- Etc.

```
df = pd.read_json("data.json")
diifpulso = df['Maxpulse'] - df['Pulse']
print(diifpulso)
```


Pandas: Exercício 1

- Preenche os valores NULL da coluna Calories com a média dessa coluna.
- Calcula a mediana, o máximo e o mínimo da coluna Calories.
 - **Mediana: 321.0**
 - **Maximo: 1860.4**
 - **Minimo: 50.3**
- Qual o Pulse para o máximo das calorias? **R. 137**
- Qual o desvio padrão da duração? **R. 42.3**
- Quantas calorias foram gastadas no máximo da duração? **R. 1500.2**

Pandas: análise estatística

- O pandas permite uma análise estatística direta com o describe

```
df.describe()
```

	Duration	Pulse	Maxpulse	Calories
count	169.000000	169.000000	169.000000	169.000000
mean	63.846154	107.461538	134.047337	375.800000
std	42.299949	14.510259	16.450434	262.383248
min	15.000000	80.000000	100.000000	50.300000
25%	45.000000	100.000000	124.000000	253.300000
50%	60.000000	105.000000	131.000000	321.000000
75%	60.000000	111.000000	141.000000	384.000000
max	300.000000	159.000000	184.000000	1860.400000

Pandas: correlações

- Para analisar o relacionamento entre variáveis, o pandas permite calcular correlações

```
df.corr()
```

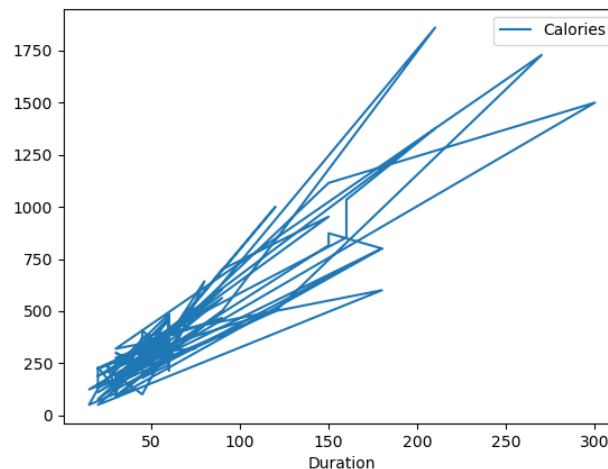
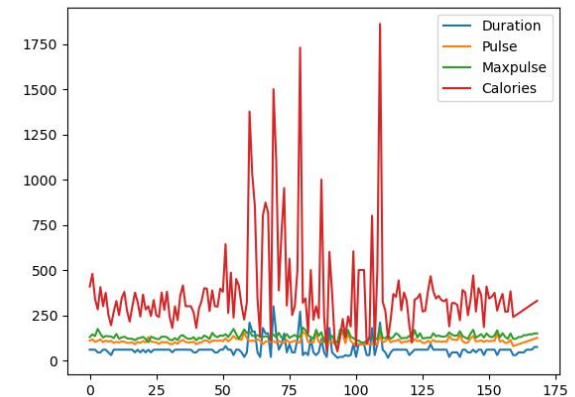
	Duration	Pulse	Maxpulse	Calories
Duration	1.000000	-0.155408	0.009403	0.921539
Pulse	-0.155408	1.000000	0.786535	0.024960
Maxpulse	0.009403	0.786535	1.000000	0.202377
Calories	0.921539	0.024960	0.202377	1.000000

Pandas: gráficos

- O pandas incorpora gráficos que depois podem ser visualizados com recurso ao matplotlib usando o show

```
df.plot()  
plt.show()
```

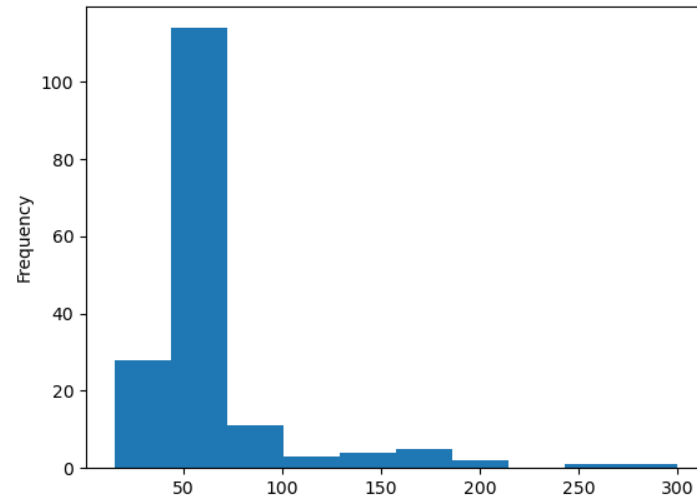
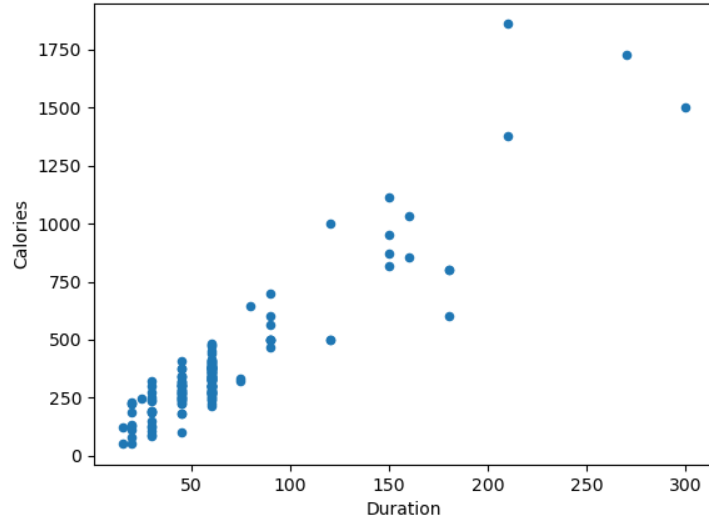
```
df.plot(x = 'Duration', y = 'Calories')  
plt.show()
```



Pandas: gráficos

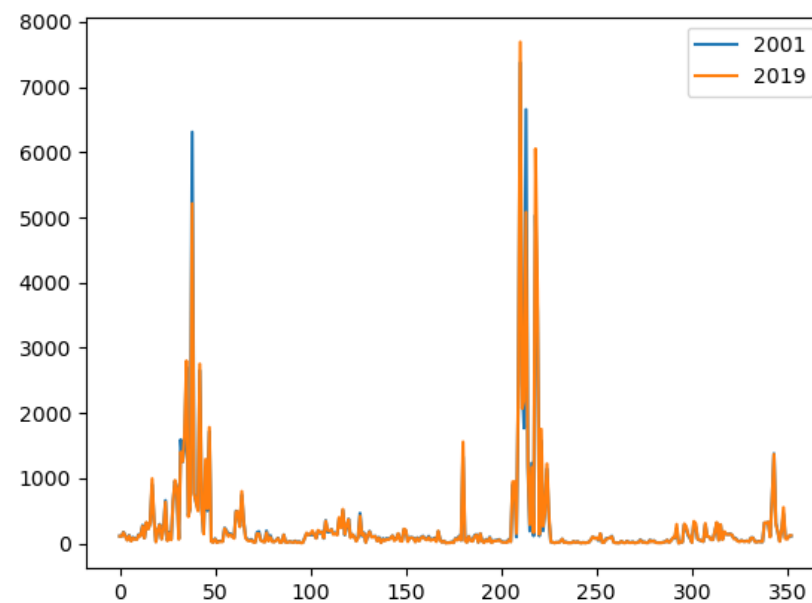
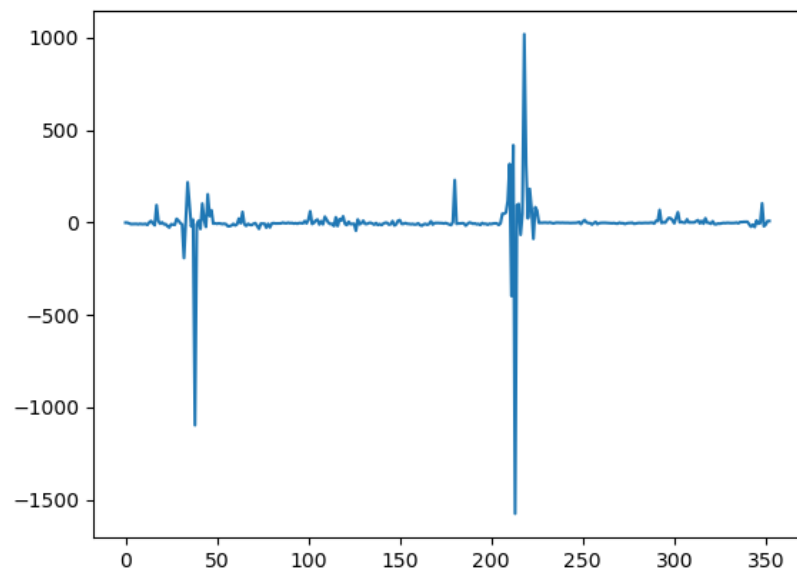
- Para outros tipos de gráfico, acrescentar o kind nos parâmetros: bar, scatter, hist, etc.

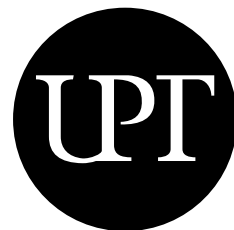
```
df.plot(x = 'Duration', y = 'Calories', kind='bar')
```



Pandas: Exercício 2

- Faz o download do ficheiro densidadepop.csv. Utilizando o pandas constrói uma dataframe.
- Verifica as 5 primeiras linhas e analisa os dados. Tem valores NULL?
- Qual a média das densidades em 2001 e 2019? R. **277 e 276**
- Qual a região com maior densidade em 2001? E em 2019? R. **Amadora**
- Calcula a diferença das densidades em cada região de 2001 para 2019. Faz um plot com essa diferença.
- Utilizando gráficos, analisa as densidades em ambos os anos.





UNIVERSIDADE
PORTUCALENSE

Do conhecimento à prática.