pandas module

Catarina Oliveira

DCT DEPARTAMENTO CIÊNCIA E TECNOLOGIA

UPT UNIVERSIDADE PORTUCALENSE

IMP.GE.190.0

CONTENT

What is it

- Module used for data analysis

- Allows:
  - Manipulate data structures of two types:
    - series
    - DataFrames
  - Assign names to rows/columns
  - Data Operations:
    - To analyze
    - To clean
    - To explore
    - manipulate and transform

- Has support for missing data

Series

A	B	C	D	A
10	50	23	70	34

DataFrame

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

Documentation: https://pandas.pydata.org/docs/

why use

- It allows you to analyze big data and draw conclusions based on statistical theories.
- It has mechanisms to clean up messy datasets and make them readable and relevant.
- Relevant data is very important in data science.

What allows to do

Data cleaning:

- Can exclude rows that are not relevant or contain incorrect values, such as empty or NULL values.

Data analysis:

- Provides answers about the data, for example:
    - Is there a correlation between two (or more) columns?
    - What is the average value?
    - Maximum value?
    - Minimum value?

Example

```python
import pandas as pd

mydataset = {
  'cars': ["BMW", "Volvo", "Ford"],
  'passings': [3, 7, 2]
}

myvar = pd.DataFrame(mydataset)

print(myvar)
```

```
   cars      passings
0  BMW       3
1  Volvo     7
2  Ford      2
```

Example with CSV file

dados.csv

```csv
CodPostal,Cidade,Freguesia,Morada,Instituição
4200,Porto,Paranhos,Dr. António Bernardino de Almeida,Universidade Portucalense
4000,Porto,Santo Ildefonso,Praça do General Humberto Delgado,Câmara Municipal do Porto
```

```python
import pandas as pd

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

```
   CodPostal Cidade        Freguesia                           Morada                Instituição
0       4200  Porto         Paranhos  Dr. António Bernardino de Almeida  Universidade Portucalense
1       4000  Porto  Santo Ildefonso  Praça do General Humberto Delgado  Câmara Municipal do Porto
```

Example with JSON file

dados.json

```
1  {
2      "instituições": [
3          {
4              "CodPostal": 4200,
5              "Cidade": "Porto",
6              "Freguesia": "Paranhos",
7              "Morada": "Rua Dr. António Bernardino de Almeida",
8              "Instituição": "Universidade Portucalense"
9          },
10         {
11             "CodPostal": 4000,
12             "Cidade": "Porto",
13             "Freguesia": "Santo Ildefonso",
14             "Morada": "Praça do General Humberto Delgado",
15             "Instituição": "Câmara Municipal do Porto"
16         }
17     ]
18 }
```

```python
import pandas as pd

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

```
   instituições
0  {'CodPostal': 4200, 'Cidade': 'Porto', 'Freguesia':'Paranhos','Morada':'Dr. António Bernardino de Almeida','Instituição':'Universidade Portucalense'}
1  {'CodPostal': 4000, 'Cidade': 'Porto', 'Freguesia':'Santo Ildefonso','Morada':'Praça do General Humberto Delgado','Instituição':'Câmara Municipal do Porto'}
```

Data access in DataFrames

inventario.csv

```csv
Produto,Preço,Quantidade
Café,1.3,4300
Águas,0.21,8000
Leite,,6000
Chocolate,0.35,
Café,1.25,3200
Leite,,9500
Chocolate,0.36,
Café,1.3,2900
```

Show the entire DataFrame

```python
import pandas as pd

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

```
     Produto  Preço  Quantidade
0       Café   1.30      4300.0
1      Águas   0.21      8000.0
2      Leite    NaN      6000.0
3  Chocolate   0.35         NaN
4       Café   1.25      3200.0
5      Leite    NaN      9500.0
6  Chocolate   0.36         NaN
7       Café   1.30      2900.0
```

Show only the “Price” column

```python
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df['Preço'])
```

```
0    1.30
1    0.21
2     NaN
3    0.35
4    1.25
5     NaN
6    0.36
1.30
Name: Preço, dtype: float64
```

View data from DataFrame

Show first 3 lines

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.head(3))
```

```
  Produto  Preço  Quantidade
0    Café   1.30      4300.0
1   Águas   0.21      8000.0
2   Leite    NaN      6000.0
```

Show the last 3 lines

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.tail(3))
```

```
     Produto  Preço  Quantidade
5      Leite    NaN      9500.0
6  Chocolate   0.36         NaN
7       Café   1.30      2900.0
```

show info

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8 entries, 0 to 7
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Produto     8 non-null      object
 1   Preço       6 non-null      float64
 2   Quantidade  6 non-null      float64
dtypes: float64(2), object(1)
memory usage: 320.0+ bytes
```

Accessing part of the data in DataFrames : loc

loc : allows accessing a group of rows and columns from labels or a boolean vector
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.loc.html

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df['Quantidade'] >= 5000)
```

```
0    False
1     True
2     True
3    False
4    False
5     True
6    False
7    False
Name: Quantidade, dtype: bool
```

Which rows have “Quantity” greater than 5000

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.loc[df['Quantidade'] >= 5000])
```

```
  Produto  Preço  Quantidade
1   Águas   0.21      8000.0
2   Leite    NaN      6000.0
5   Leite    NaN      9500.0
```

Show rows that have "Quantity" greater than 5000

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.loc[df['Quantidade'] >= 5000,'Preço'])
```

```
1    0.21
2     NaN
5     NaN
Name: Preço, dtype: float64
```

Show the "Price" column for rows that have "Quantity" greater than 5000

Access part of data in DataFrames : iloc

iloc : allows accessing a group of rows and columns from their indexes
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.iloc[0:2, :])
```

```
  Produto  Preço  Quantidade
0    Café   1.30      4300.0
1   Águas   0.21      8000.0
```

Show rows with indexes 0 to 2 (0, 1, 2) of all columns

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.iloc[:, 0:2])
```

```
     Produto  Preço
0       Café   1.30
1      Águas   0.21
2      Leite    NaN
3  Chocolate   0.35
4       Café   1.25
5      Leite    NaN
6  Chocolate   0.36
7       Café   1.30
```

Show all rows of columns with indexes 0 to 2

```
import pandas as pd

df = pd.read_csv("inventario.csv")
print(df.iloc[0:2, 0:2])
```

```
  Produto  Preço
0    Café   1.30
1   Águas   0.21
```

Show rows with indexes 0 to 2 from columns with indexes 0 to 2

Modify data from a DataFrame

Change all prices to 1.5

```
import pandas as pd

df = pd.read_csv("inventario.csv")
df['Preço'] = 1.5
print(df)
```

```
     Produto  Preço  Quantidade
0       Café    1.5      4300.0
1      Águas    1.5      8000.0
2      Leite    1.5      6000.0
3  Chocolate    1.5         NaN
4       Café    1.5      3200.0
5      Leite    1.5      9500.0
6  Chocolate    1.5         NaN
7       Café    1.5      2900.0
```

Change all product prices with quantity greater than 5000 to 1.5

```
import pandas as pd

df = pd.read_csv("inventario.csv")
df.loc[df['Quantidade'] >= 5000, 'Preço'] = 1.5
print(df)
```

```
     Produto  Preço  Quantidade
0       Café   1.30      4300.0
1      Águas   1.50      8000.0
2      Leite   1.50      6000.0
3  Chocolate   0.35         NaN
4       Café   1.25      3200.0
5      Leite   1.50      9500.0
6  Chocolate   0.36         NaN
7       Café   1.30      2900.0
```

Data cleaning: filling all Null and NaN

Replace in original

```
import pandas as pd

df = pd.read_csv("inventario.csv")
df.fillna(888, inplace = True)
print(df)
```

```
     Produto   Preço  Quantidade
0       Café    1.30      4300.0
1      Águas    0.21      8000.0
2      Leite  888.00      6000.0
3  Chocolate    0.35       888.0
4       Café    1.25      3200.0
5      Leite  888.00      9500.0
6  Chocolate    0.36       888.0
7       Café    1.30      2900.0
```

```
 Produto   Preço  Quantidade
0       Café    1.30      4300.0
1      Águas    0.21      8000.0
2      Leite     NaN      6000.0
3  Chocolate    0.35         NaN
4       Café    1.25      3200.0
5      Leite     NaN      9500.0
6  Chocolate    0.36         NaN
Café    1.30      2900.0

     Produto   Preço  Quantidade
0       Café    1.30      4300.0
1      Águas    0.21      8000.0
2      Leite  888.00      6000.0
3  Chocolate    0.35       888.0
4       Café    1.25      3200.0
5      Leite  888.00      9500.0
6  Chocolate    0.36       888.0
7       Café    1.30      2900.0
```

Create a new DataFrame

```
import pandas as pd

df = pd.read_csv("inventario.csv")
novoDF = df.fillna(888)

print(df)
print(novoDF)
```

data cleaning

Replace in original DataFrame , Quantity column only

```
import pandas as pd

df = pd.read_csv("inventario.csv")
df["Quantidade"].fillna(888, inplace = True)

print(df)
```

```
     Produto   Preço  Quantidade
0       Café    1.30      4300.0
1      Águas    0.21      8000.0
2      Leite     NaN      6000.0
3  Chocolate    0.35       888.0
4       Café    1.25      3200.0
5      Leite     NaN      9500.0
6  Chocolate    0.36       888.0
7       Café    1.30      2900.0
```

Create a new DataFrame (df is not changed) without the rows with NaN

```
import pandas as pd

df = pd.read_csv("inventario.csv")
novoDF = df.dropna()

print(novoDF)
```

```
 Produto  Preço  Quantidade
0    Café   1.30      4300.0
1   Águas   0.21      8000.0
4    Café   1.25      3200.0
7    Café   1.30      2900.0
```

Remove lines with NaN from original

```
import pandas as pd

df = pd.read_csv("inventario.csv")
df.dropna(inplace = True)

print(df)
```

```
 Produto  Preço  Quantidade
0    Café   1.30      4300.0
1   Águas   0.21      8000.0
4    Café   1.25      3200.0
7    Café   1.30      2900.0
```

Statistical measures of columns or rows

- **mean**
- **median**
- **max**
- **min**
- **std**

Get the average of the “Price” column and the average of the numeric values in row 1 (line indices start at 0)

```
import pandas as pd

df = pd.read_csv("inventario.csv")

print(df['Preço'].mean())
print(df.iloc[1,1:3].mean())
```

→

```
0.7949999999999999
4000.105
```

Operations with columns or rows

- Sum
- Subtraction
- Division
- ...

Get the product of prices by quantity

```
import pandas as pd

df = pd.read_csv("inventario.csv")

print(df['Preço'] * df['Quantidade'])
```

```
0    5590.0
1    1680.0
2       NaN
3       NaN
4    4000.0
5       NaN
6       NaN
7    3770.0
dtype: float64
```

Get the sum of prices and quantities from rows 0 and 1

```
import pandas as pd

df = pd.read_csv("inventario.csv")

print(df.iloc[0, 1:3] + df.iloc[1, 1:3])
```

```
Preço            1.51
Quantidade    12300.0
dtype: object
```

Statistical analysis

```python
import pandas as pd

df = pd.read_csv("inventario.csv")

print(df.describe())
```

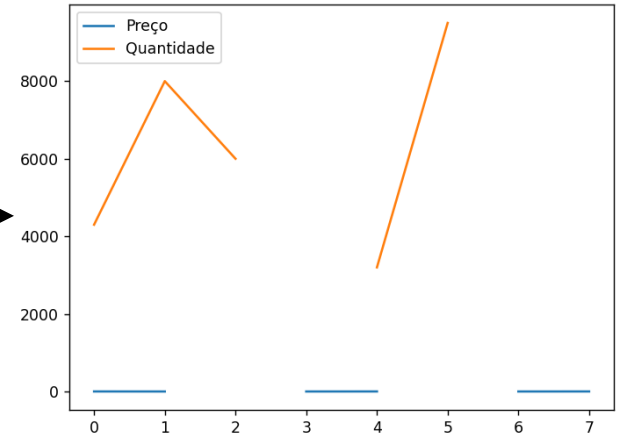
```
          Preço  Quantidade
count  6.000000     6.00000
mean   0.795000  5650.00000
std    0.537875  2677.87229
min    0.210000  2900.00000
25%    0.352500  3475.00000
50%    0.805000  5150.00000
75%    1.287500  7500.00000
max    1.300000  9500.00000
```

```python
import pandas as pd

df = pd.read_csv("inventario.csv")

print(df.iloc[:, 1:3].corr())
```

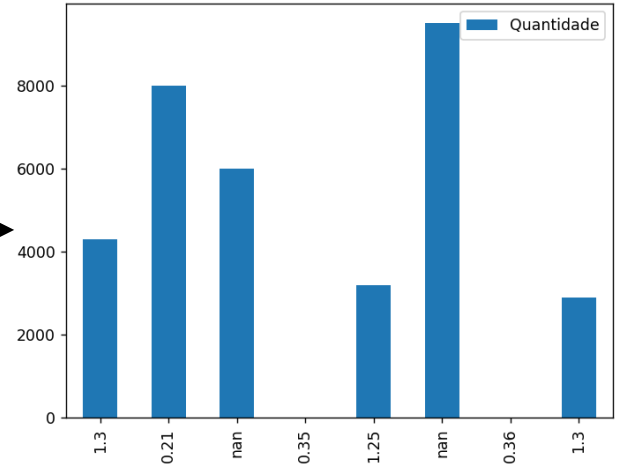
```
               Preço  Quantidade
Preço       1.000000   -0.962051
Quantidade -0.962051    1.000000
```

Charts

```
import pandas as pd
from matplotlib import pyplot as plt

df = pd.read_csv("inventario.csv")

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




```
import pandas as pd
from matplotlib import pyplot as plt

df = pd.read_csv("inventario.csv")

df.plot(x = 'Preço', y = 'Quantidade', kind='bar')
plt.show()
```



UNIVERSIDADE
PORTUCALENSE

Do conhecimento à prática.