

Manipulação de dataframe - TAREFA 2 - Parte III

1-) Leitura dos dados

A base de dados [Churn\\_Modelling](#) é utilizada para prever (classificar) quando um cliente irá parar de utilizar o serviço de um banco.

Conjunto de dados (características):

- Número de Instâncias: 10.000
- Número de Atributos: 13 atributos (incluindo o target)
- Informações dos Atributos:
  - RowNumber
  - CustomerId
  - Surname
  - CreditScore
  - Geography
  - Gender
  - Age
  - Tenure
  - Balance
  - NumOfProducts
  - HasCrCard
  - IsActiveMember
  - EstimatedSalary
  - Exited (variável de destino - target)

Clone o repositório de dados da disciplina (<https://github.com/malegopc/DSBD>), faça a leitura do arquivo **Churn\_Modelling.csv** disponível na pasta [Churn\\_Modelling](#) e armazene-o numa variável do tipo dataframe. Basta utilizar a função [pd.read\\_csv](#) da biblioteca pandas.

```
import pandas as pd

!git clone "https://github.com/malegopc/DSBD"

data = pd.read_csv('DSBD/Datasets/Churn/Churn_Modelling.csv')

print(type(data))
```

 fatal: destination path 'DSBD' already exists and is not an empty directory.  
<class 'pandas.core.frame.DataFrame'>

2-) Visualização do dataframe

Mostre as 5 primeiras e as 5 últimas linhas do dataframe.

```
data.head()
```

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Bal
0	1	15634602	Hargrave	619	France	Female	42	2	
1	2	15647311	Hill	608	Spain	Female	41	1	8381
2	3	15619304	Onio	502	France	Female	42	8	15961
3	4	15701354	Boni	699	France	Female	39	1	
4	5	15737888	Mitchell	850	Spain	Female	43	2	1255

Próximas etapas:

Gerar código com data


☒ Ver gráficos recomendados

New interactive sheet

3-) Eliminação manual de variáveis

Elimine as variáveis (colunas): 'RowNumber', 'CustomerId' e 'Surname'.

```
data.drop(['RowNumber', 'CustomerId', 'Surname'], axis = 1)
```




	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCar
0	619	France	Female	42	2	0.00	1	
1	608	Spain	Female	41	1	83807.86	1	
2	502	France	Female	42	8	159660.80	3	
3	699	France	Female	39	1	0.00	2	
4	850	Spain	Female	43	2	125510.82	1	
...	...	...	...	...	...	...	...	...
9995	771	France	Male	39	5	0.00	2	
9996	516	France	Male	35	10	57369.61	1	
9997	709	France	Female	36	7	0.00	1	
9998	772	Germany	Male	42	3	75075.31	2	
9999	792	France	Female	28	4	130142.79	1	

10000 rows × 11 columns

4-) Dados ausentes

Verifique se há dados ausentes no dataframe.

```
print('Dados null(vazios):\n')
data.isnull().sum()
```



```
Dados null(vazios):
```

	0
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

dtype: int64

5-) Separa as variáveis

Utilizando a função [iloc](#), copie para uma variável "X" os valores de todas as variáveis (colunas) do dataframe, exceto os valores da variável "Exited".

E copie para uma variável "y" somente os valores da variável "Exited".

```
df = pd.DataFrame(data)

print("\nVariável X:\n")
x = df.iloc[:, 0 : 13]
print(x)

print('\n-----')
print("\nVariável Y:\n")
y = df.iloc[:, 13]
```

print(y)



	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	\
0	1	15634602	Hargrave	619	France	Female	42	
1	2	15647311	Hill	608	Spain	Female	41	
2	3	15619304	Onio	502	France	Female	42	
3	4	15701354	Boni	699	France	Female	39	
4	5	15737888	Mitchell	850	Spain	Female	43	
...	...	...	...	...	...	...	...	...
9995	9996	15606229	Obijiaku	771	France	Male	39	
9996	9997	15569892	Johnstone	516	France	Male	35	
9997	9998	15584532	Liu	709	France	Female	36	
9998	9999	15682355	Sabbatini	772	Germany	Male	42	
9999	10000	15628319	Walker	792	France	Female	28	

	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	\
0	2	0.00	1	1	1	
1	1	83807.86	1	0	1	
2	8	159660.80	3	1	0	
3	1	0.00	2	0	0	
4	2	125510.82	1	1	1	
...	...	...	...	...	...	...
9995	5	0.00	2	1	0	
9996	10	57369.61	1	1	1	
9997	7	0.00	1	0	1	
9998	3	75075.31	2	1	0	
9999	4	130142.79	1	1	0	

	EstimatedSalary
0	101348.88
1	112542.58
2	113931.57
3	93826.63
4	79084.10
...	...
9995	96270.64
9996	101699.77
9997	42085.58
9998	92888.52
9999	38190.78

[10000 rows x 13 columns]

-----

Variável Y:

0	1
1	0
2	1
3	0
4	0
...	..
9995	0
9996	0
9997	1
9998	1
9999	0

Name: Exited, Length: 10000, dtype: int64