Tutorial 9 - Merge Dataframes

In [1]:
```
import pandas as pd
```

In [2]:
```
df1 = pd.DataFrame({
    "city": ["new york","chicago","orlando"],
    "temperature": [21,14,35],
})
df1
```

Out[2]:

	city	temperature
0	new york	21
1	chicago	14
2	orlando	35

In [3]:
```
df2 = pd.DataFrame({
    "city": ["chicago","new york","orlando"],
    "humidity": [68,65,75]
})
df2
```

Out[3]:

	city	humidity
0	chicago	68
1	new york	65
2	orlando	75

In [4]:
```
# Para juntar os 2 DataFrame fazendo o concatenamento com a coluna "city":
df3 = pd.merge(df1,df2,on="city")
df3
```

Out[4]:

	city	temperature	humidity
0	new york	21	65
1	chicago	14	68
2	orlando	35	75

In [5]:
```
# Vamos criar em um DataFrame uma cidade a mais numa variável:
df4 = pd.DataFrame({
    "city": ["new york","chicago","orlando","baltimore"],
    "temperature": [21,14,35,32],
})
df4
```

Out[5]:

	city	temperature
0	new york	21
1	chicago	14
2	orlando	35
3	baltimore	32

In [6]:
```
df5 = pd.DataFrame({
    "city": ["chicago","new york","san francisco"],
    "humidity": [68,65,71]
})
df5
```

Out[6]:

	city	humidity
0	chicago	68
1	new york	65
2	san francisco	71

In [7]:
```
# Observe que os dados das cidades que não interseção ("inner join") de dados não e
ntra no novo DataFrame!
df6 = pd.merge(df4,df5,on="city")
df6
```

Out[7]:

	city	temperature	humidity
0	new york	21	65
1	chicago	14	68

In [8]:
```
# Fazendo um merge com a união ("outer join")
# Outras opções do how="": "left" (df4 + intersection df4 e df5) e "right" (df5 + i
ntersection df4 e df5)
df7 = pd.merge(df4,df5, on="city", how="outer")
df7
```

Out[8]:

	city	temperature	humidity
0	new york	21.0	65.0
1	chicago	14.0	68.0
2	orlando	35.0	NaN
3	baltimore	32.0	NaN
4	san francisco	NaN	71.0

In [9]:
```
# Fazendo um merge com a união ("outer join") e ligando a indicação de onde vem os
 dados:
df7 = pd.merge(df4,df5, on="city", how="outer", indicator="True")
df7
```

Out[9]:

	city	temperature	humidity	True
0	new york	21.0	65.0	both
1	chicago	14.0	68.0	both
2	orlando	35.0	NaN	left_only
3	baltimore	32.0	NaN	left_only
4	san francisco	NaN	71.0	right_only

In [10]:
```
# Vamos agora pegar dados com diferentes dados para cada cidade:
df1 = pd.DataFrame({
    "city": ["new york","chicago","orlando","baltimore"],
    "temperature": [21,14,35,38],
    "humidity": [68,65,71,75]
})
df1
```

Out[10]:

	city	temperature	humidity
0	new york	21	68
1	chicago	14	65
2	orlando	35	71
3	baltimore	38	75

In [11]:
```
# Observe que os números são iguais mas as cidades foram invertidas!
df2 = pd.DataFrame({
    "city": ["chicago","new york","san diego"],
    "temperature": [21,14,35],
    "humidity": [68,65,71]
})
df2
```

Out[11]:

	city	temperature	humidity
0	chicago	21	68
1	new york	14	65
2	san diego	35	71

In [12]:
```
df3 = pd.merge(df1,df2,on="city",how="outer")
df3
```

Out[12]:

	city	temperature_x	humidity_x	temperature_y	humidity_y
0	new york	21.0	68.0	14.0	65.0
1	chicago	14.0	65.0	21.0	68.0
2	orlando	35.0	71.0	NaN	NaN
3	baltimore	38.0	75.0	NaN	NaN
4	san diego	NaN	NaN	35.0	71.0

In [13]:
```
df3 = pd.merge(df1,df2,on="city",how="outer",suffixes=('_LH','_RH'))
df3
```

Out[13]:

	city	temperature_LH	humidity_LH	temperature_RH	humidity_RH
0	new york	21.0	68.0	14.0	65.0
1	chicago	14.0	65.0	21.0	68.0
2	orlando	35.0	71.0	NaN	NaN
3	baltimore	38.0	75.0	NaN	NaN
4	san diego	NaN	NaN	35.0	71.0