Tutorial 7 - Group By (Split Apply Combine)

In [1]:
```
import pandas as pd
df = pd.read_csv('sample_data_tutorial_07.csv')
df
```

Out[1]:

	day	city	temperature	windspeed	event
0	1/1/2017	new york	32	6	Rain
1	1/2/2017	new york	36	7	Sunny
2	1/3/2017	new york	28	12	Snow
3	1/4/2017	new york	33	7	Sunny
4	1/1/2017	mumbai	90	5	Sunny
5	1/2/2017	mumbai	85	12	Fog
6	1/3/2017	mumbai	87	15	Fog
7	1/4/2017	mumbai	92	5	Rain
8	1/1/2017	paris	45	20	Sunny
9	1/2/2017	paris	50	13	Cloudy
10	1/3/2017	paris	54	8	Cloudy
11	1/4/2017	paris	42	10	Cloudy

In [2]:
```
# Vamos agrupar este DataFrame pelas cidades:
g = df.groupby('city')
g
```

Out[2]:
```
<pandas.core.groupby.groupby.DataFrameGroupBy object at 0x0000023981518978>
```

In [3]:
```
# O comando anterior agrupará as cidades como "key" e o resto como valores
for city, city_df in g:
    print(city)
    print (city_df)
```

```
mumbai
        day    city  temperature  windspeed  event
4  1/1/2017  mumbai           90          5  Sunny
5  1/2/2017  mumbai           85         12    Fog
6  1/3/2017  mumbai           87         15    Fog
7  1/4/2017  mumbai           92          5   Rain
new york
        day      city  temperature  windspeed  event
0  1/1/2017  new york           32          6   Rain
1  1/2/2017  new york           36          7  Sunny
2  1/3/2017  new york           28         12   Snow
3  1/4/2017  new york           33          7  Sunny
paris
         day   city  temperature  windspeed   event
8   1/1/2017  paris           45         20   Sunny
9   1/2/2017  paris           50         13  Cloudy
10  1/3/2017  paris           54          8  Cloudy
11  1/4/2017  paris           42         10  Cloudy
```

In [4]:
```
# Especificando um grupo:
g.get_group('paris')
```

Out[4]:

	day	city	temperature	windspeed	event
8	1/1/2017	paris	45	20	Sunny
9	1/2/2017	paris	50	13	Cloudy
10	1/3/2017	paris	54	8	Cloudy
11	1/4/2017	paris	42	10	Cloudy

In [5]:
```
# Pegando o valor máximo nos grupos
g.max()
```

Out[5]:

	day	temperature	windspeed	event
city				
mumbai	1/4/2017	92	15	Sunny
new york	1/4/2017	36	12	Sunny
paris	1/4/2017	54	20	Sunny

In [6]:
```
# Pegando o resumo estatístico
g.describe()
```

Out[6]:

	temperature								windspeed				
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25
city													
mumbai	4.0	88.50	3.109126	85.0	86.50	88.5	90.50	92.0	4.0	9.25	5.057997	5.0	5.(
new york	4.0	32.25	3.304038	28.0	31.00	32.5	33.75	36.0	4.0	8.00	2.708013	6.0	6.1
paris	4.0	47.75	5.315073	42.0	44.25	47.5	51.00	54.0	4.0	12.75	5.251984	8.0	9.5

In [7]:
```
%matplotlib inline
g.plot()
# O primeiro comando é para rodar o Matplotlib no Jupyter Notebook
# Procure na documentação dos pandas "Groupby" para ver mais detalhes!!!
```

Out[7]:
```
city
mumbai      AxesSubplot(0.125,0.125;0.775x0.755)
new york    AxesSubplot(0.125,0.125;0.775x0.755)
paris       AxesSubplot(0.125,0.125;0.775x0.755)
dtype: object
```