Demo

A continuación haremos una exploración de información sobre un muestreo del peso y estatura de 500 personas. La información está dividida en dos fuentes de información distintas.

Recordemos que el primer paso es importar algunas librerías que nos serán útiles

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Además tendremos que importar las fuentes de información.

```
df_data = pd.read_csv('500_Person_Gender_Height_Weight_Index.csv')
df_description = pd.read_csv('/Users/marco/Desktop/Descriptions.csv')
```

Para saber con qué estamos trabajando será necesario ver por lo menos las primeras filas de cada **data frame**.

```
df_data.head()
```

	Gender	Height	Weight	Index
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3

```
df_description.head()
```

	Index	Description
0	0	Extremely Weak
1	1	Weak
2	2	Normal
3	3	Overweight
4	4	Obesity

Con lo que hemos visto arriba existen dos columnas en común, por lo que podemos combinarlas.

```
df = pd.merge(left= df_data, right= df_description, on= 'Index')
```

Verificamos cómo luce nuestro nuevo *data frame*.

```
df.head()
```

	Gender	Height	Weight	Index	Description
0	Male	174	96	4	Obesity
1	Female	185	110	4	Obesity
2	Female	169	103	4	Obesity
3	Female	159	80	4	Obesity
4	Female	169	97	4	Obesity

Podemos ver las columnas que lo componen

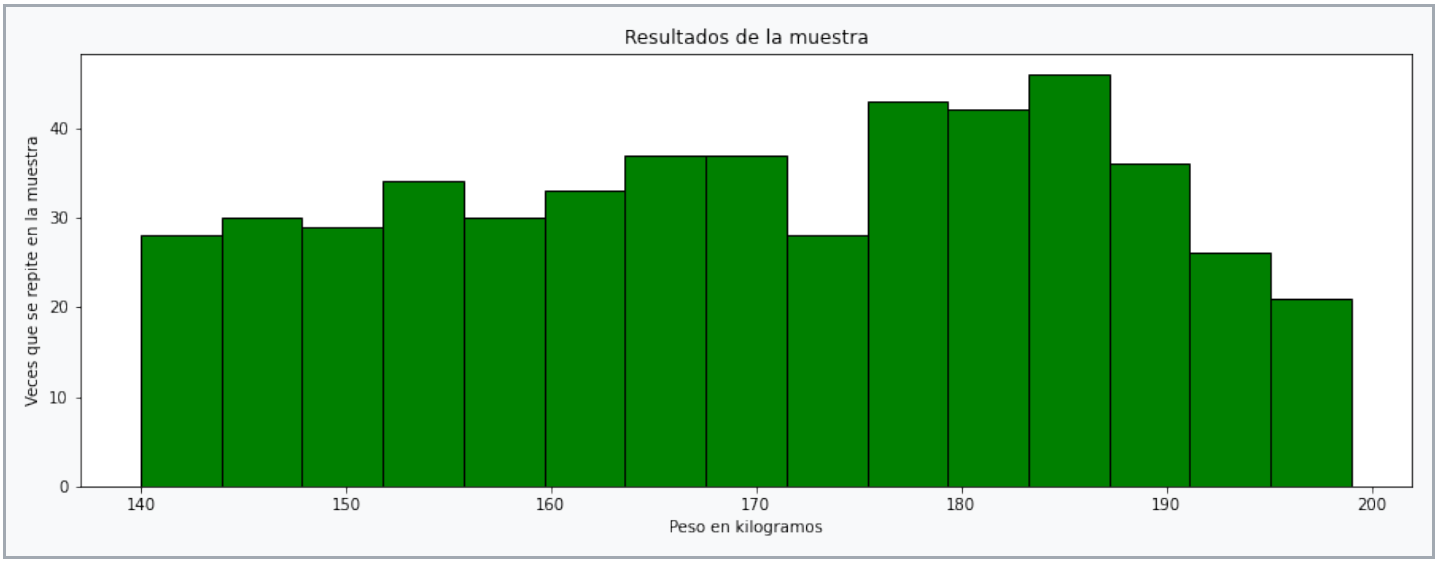
```
df.columns
```

Otra opción para que lo indique como una fila es:

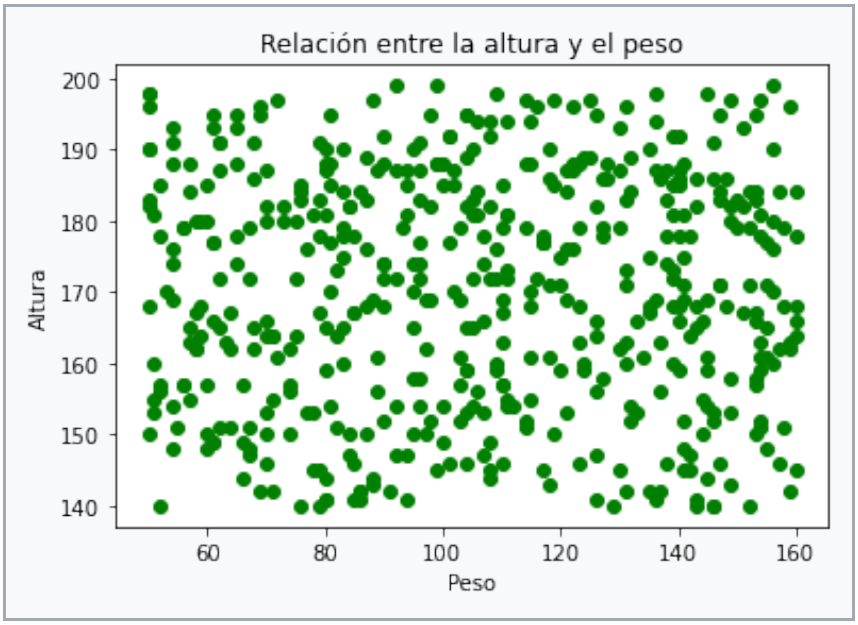
```
list(df.columns)
```

La información antes mostrada nos permitirá hacer unas gráficas, ejemplo de ello son:

```
fig, ax = plt.subplots(figsize= (15,5))
x = df['Height']
ax.hist(x, color= 'green', bins=15, edgecolor= 'black')
ax.set_title('Resultados de la muestra')
ax.set_xlabel('Peso en kilogramos')
ax.set_ylabel('Veces que se repite en la muestra')
plt.show()
```




```
fig, ax = plt.subplots()
x = df['Weight']
y = df['Height']
ax.scatter(x, y, color='green')
ax.set_title('Relación entre la altura y el peso')
ax.set_xlabel('Peso')
ax.set_ylabel('Altura')
plt.show()
```



Es posible que nuestro *data frame* principal lo separemos en uno exclusivos con datos de hombres y otro de mujeres, para ello es necesario filtrar la información.

```
df_hombres = df[df['Gender'] == 'Male']
df_mujeres = df[df['Gender'] == 'Female']
```

También es posible explorar los datos con *groupby*, en los siguientes dos ejemplos los datos muestran promedios.

```
df_hombres.groupby(by= 'Description')['Weight', 'Height'].mean()
```

	Weight	Height
Description		
Extreme Obesity	130.552381	160.457143
Extremely Weak	51.500000	188.666667
Normal	72.250000	178.035714
Obesity	107.813559	173.322034
Overweight	85.937500	174.906250
Weak	59.733333	185.066667

```
print('Hay ' + str(len(df_hombres)) + ' registros de hombres.')
```

Lo anterior imprimirá `Hay 245 registros de hombres.`.

```
df_mujeres.groupby(by= 'Description')['Weight', 'Height'].mean()
```

	Weight	Height
Description		
Extreme Obesity	135.526882	161.569892
Extremely Weak	51.857143	186.571429
Normal	66.926829	171.682927
Obesity	108.070423	174.338028
Overweight	87.722222	176.944444
Weak	58.714286	184.142857

```
print('Hay ' + str(len(df_mujeres)) + ' registros de hombres.')
```

Lo anterior imprimirá `Hay 255 registros de hombres.`.