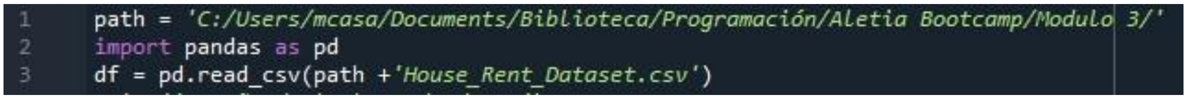
Data Analyst formation – Challenge 3.

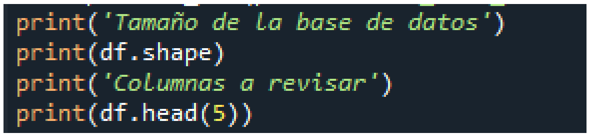
Kaggle dataset right here.

1. Examine the shape and the first five rows of the dataset.
    - First, we import the dataset into Python using the Spyder platform.

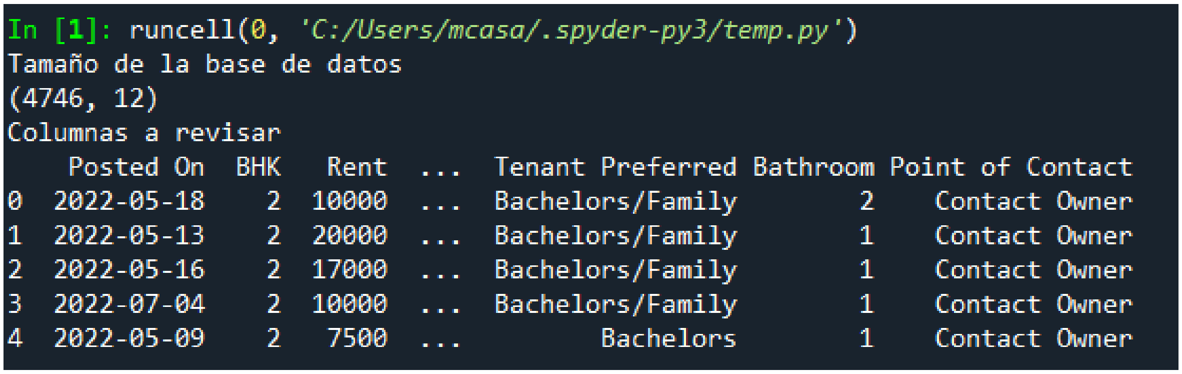


```
path = 'C:/Users/mcasa/Documents/Biblioteca/Programación/Aletia Bootcamp/Modulo 3/'
import pandas as pd
df = pd.read_csv(path +'House_Rent_Dataset.csv')
```

    - Then, we enter the following code to obtain the size of the dataset and the corresponding rows, taking a sample of 5.

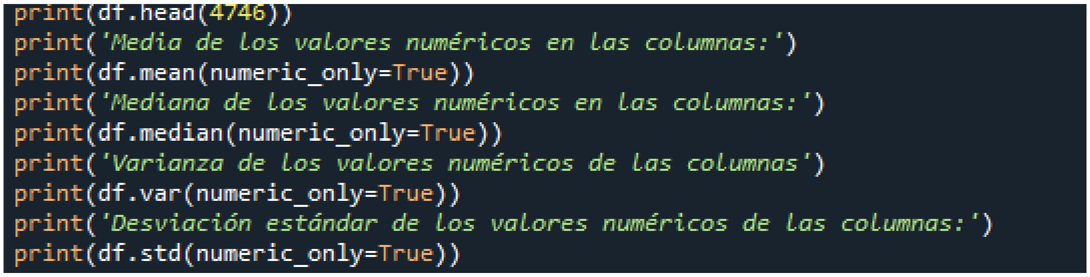


```
print('Tamaño de la base de datos')
print(df.shape)
print('Columnas a revisar')
print(df.head(5))
```

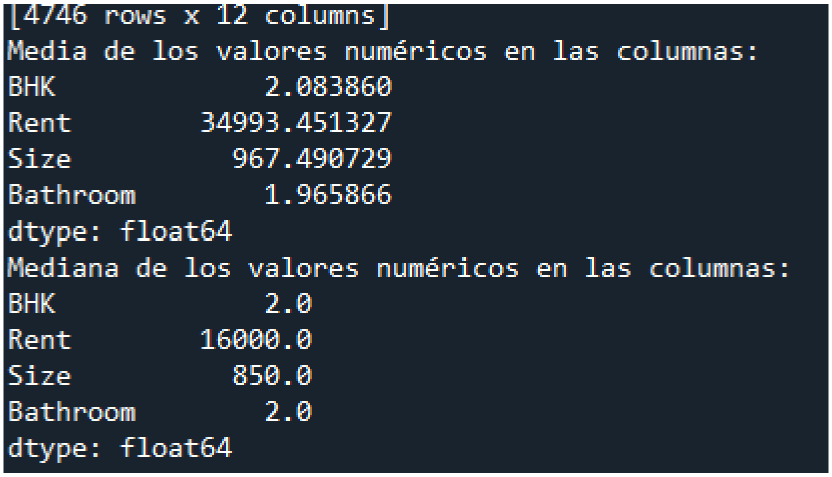


```
In [1]: runcell(0, 'C:/Users/mcasa/.spyder-py3/temp.py')
Tamaño de la base de datos
(4746, 12)
Columnas a revisar
    Posted On  BHK   Rent  ...  Tenant Preferred Bathroom Point of Contact
0  2022-05-18    2  10000  ...  Bachelors/Family        2    Contact Owner
1  2022-05-13    2  20000  ...  Bachelors/Family        1    Contact Owner
2  2022-05-16    2  17000  ...  Bachelors/Family        1    Contact Owner
3  2022-07-04    2  10000  ...  Bachelors/Family        1    Contact Owner
4  2022-05-09    2   7500  ...         Bachelors        1    Contact Owner
```

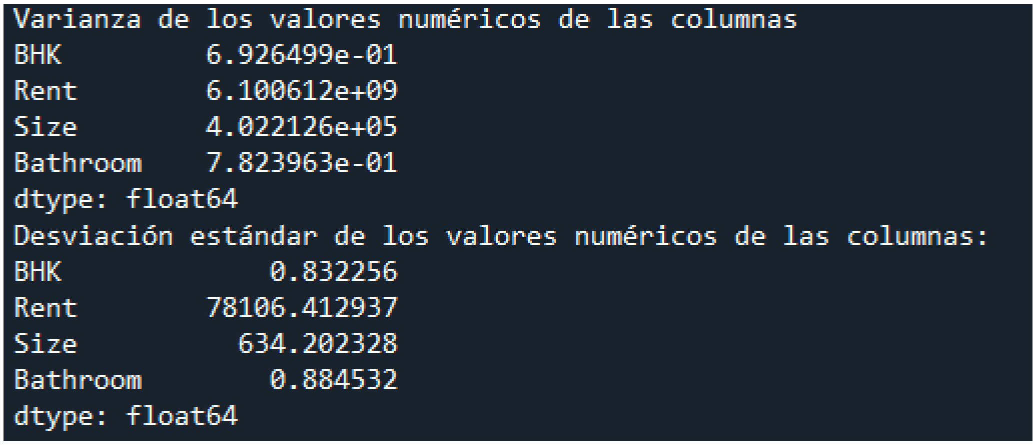
2. Calculate some measures of central tendency and dispersion:
Next, we request from the program different measures of central tendency and dispersion considering the entire dataset and only the numerical values in it.



```
print(df.head(4746))
print('Media de los valores numéricos en las columnas:')
print(df.mean(numeric_only=True))
print('Mediana de los valores numéricos en las columnas:')
print(df.median(numeric_only=True))
print('Varianza de los valores numéricos de las columnas')
print(df.var(numeric_only=True))
print('Desviación estándar de los valores numéricos de las columnas:')
print(df.std(numeric_only=True))
```

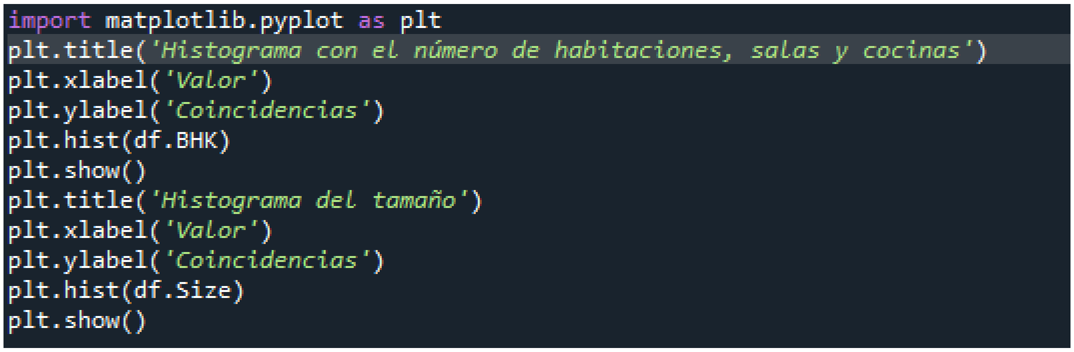


```
[4746 rows x 12 columns]
Media de los valores numéricos en las columnas:
BHK             2.083860
Rent        34993.451327
Size          967.490729
Bathroom        1.965866
dtype: float64
Mediana de los valores numéricos en las columnas:
BHK             2.0
Rent        16000.0
Size          850.0
Bathroom        2.0
dtype: float64
```



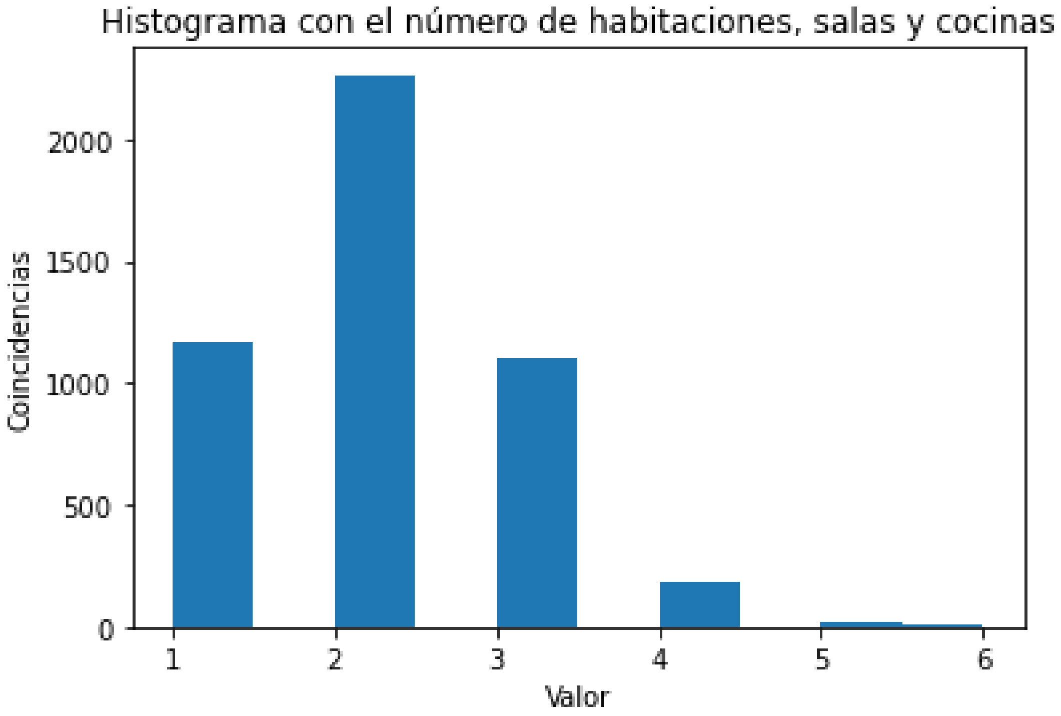
```
Varianza de los valores numéricos de las columnas
BHK         6.926499e-01
Rent        6.100612e+09
Size        4.022126e+05
Bathroom    7.823963e-01
dtype: float64
Desviación estándar de los valores numéricos de las columnas:
BHK             0.832256
Rent        78106.412937
Size          634.202328
Bathroom        0.884532
dtype: float64
```

3. Now we can proceed to make an hist plot for the number of the rooms, and the correlation of the value, and a hist plot about the value and the coincidences of the values for the dataset.

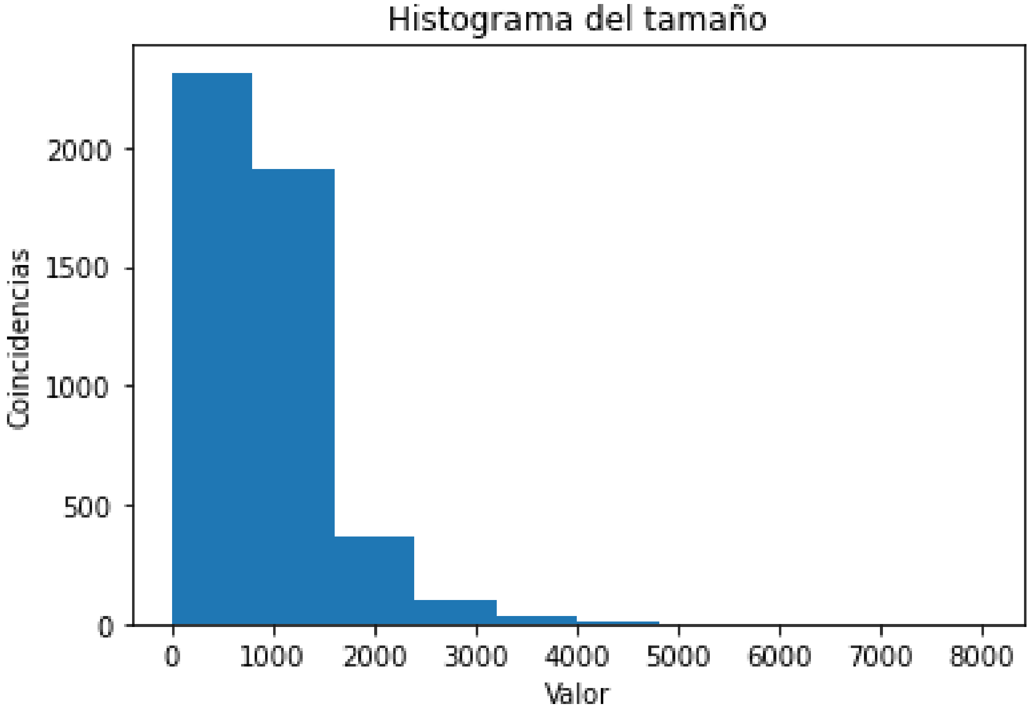


```python
import matplotlib.pyplot as plt
plt.title('Histograma con el número de habitaciones, salas y cocinas')
plt.xlabel('Valor')
plt.ylabel('Coincidencias')
plt.hist(df.BHK)
plt.show()
plt.title('Histograma del tamaño')
plt.xlabel('Valor')
plt.ylabel('Coincidencias')
plt.hist(df.Size)
plt.show()
```

4. As we can conclude of the plots, the most common value is of 2, with +2000 coincidences, as the average home has that features. Also, the most common values oscilates between 0 and 1500.



Histograma con el número de habitaciones, salas y cocinas



Histograma del tamaño