EDEM
Centro Universitario

Correlation

Hypothesis testing+ Graphic Methods

Session 9

Programación Estadística con Python

Alberto Sanz, Ph.D
albertosanz@bigwaveanalytics.es

MASTER EN DATA ANALYTICS PARA LA EMPRESA

Goals

EDEM
Centro Universitario

- Hypothesis testing over the relatioship of two quantitative variables by the means of regression.
  - Graphic approach (Scatterplot)
  - Numeric approach (r coefficient & p.values)

- Beyond the numbers and plots:
  - Reflections on correlation and non linearity
  - The False Discovery Rate experiment

Correlation

EDEM
Centro Universitario

- Always **DESCRIBE** the two variables involved in the correlation.
  - Check and validate the integrity of the data prior to any analysis.

- **EXPLORE** of bivariate relation:
  - Graphically: **Scatterplot**
  - Numerically: Pearson's **r** & **p.value**

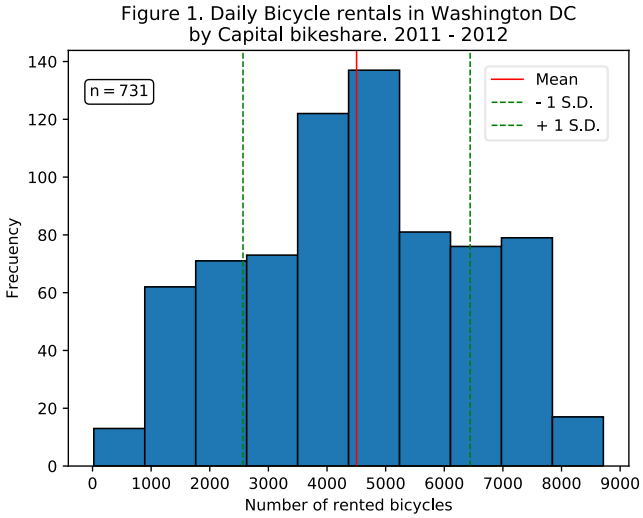
alberto.sanz@bigwaveanalytics.es

Research Question

EDEM
Centro Universitario

Why some days are rent *more* bikes?

- Temperature ?



- H0.: There is no linear association ($r=0$) between the *number of rentals* and the *temperature*.

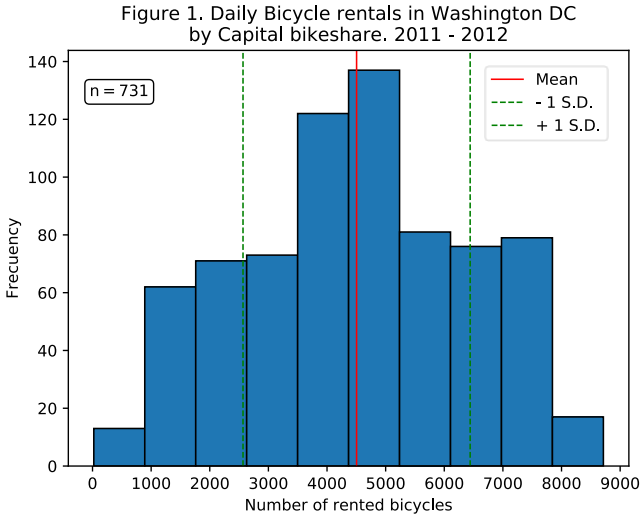
- H1.: There is a linear association ($r\neq0$) between the *number of rentals* and the *temperature*.

alberto.sanz@bigwaveanalytics.es

Describing quantitative variables

EDEM Centro Universitario


```
x=wbr['cnt']
plt.hist(x, bins=10,
edgecolor='black')
plt.xticks(np.arange(0, 10000,
step=1000))
plt.title('Figure 4. Daily Bicycle
rentals in Washington DC'
          '\n'
          'by Capital bikeshare.
2011 - 2012')
plt.ylabel('Frecuency')
plt.xlabel('Number of rented
bicycles')
```



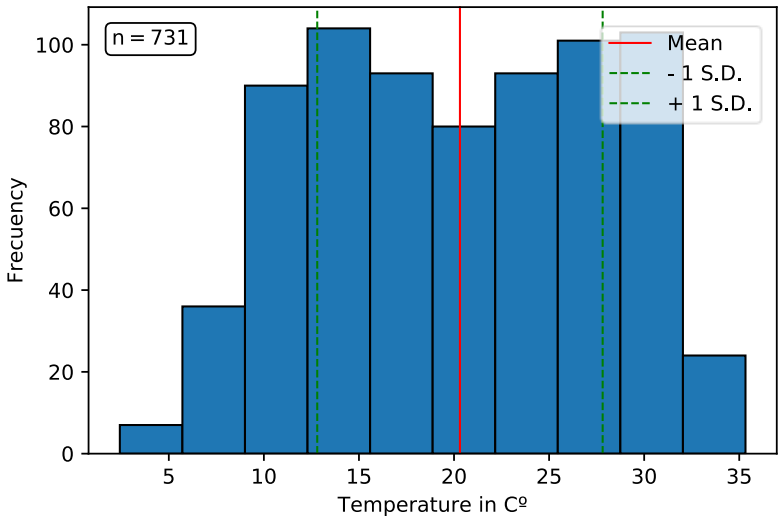

```
props = dict(boxstyle='round',
facecolor='white', lw=0.5)
textstr = '$\mathrm{n}=%.0f$'%(n)
plt.text (-50,128, textstr ,
bbox=props)
```

Describing quantitative variables

EDEM Centro Universitario


```
##histogram ver4
x=wbr['temp_celsius']
plt.hist(x, bins=10,
edgecolor='black')
#plt.xticks(np.arange(0, 10000,
step=1000))
plt.title('Figure 5. Temperature in
Celsius'
          '\n')
plt.ylabel('Frecuency')
plt.xlabel('Temperature in C°')
props = dict(boxstyle='round',
facecolor='white', lw=0.5)
textstr = '$\mathrm{n}=%.0f$'%(n)
plt.text (2,100, textstr ,
bbox=props)
```

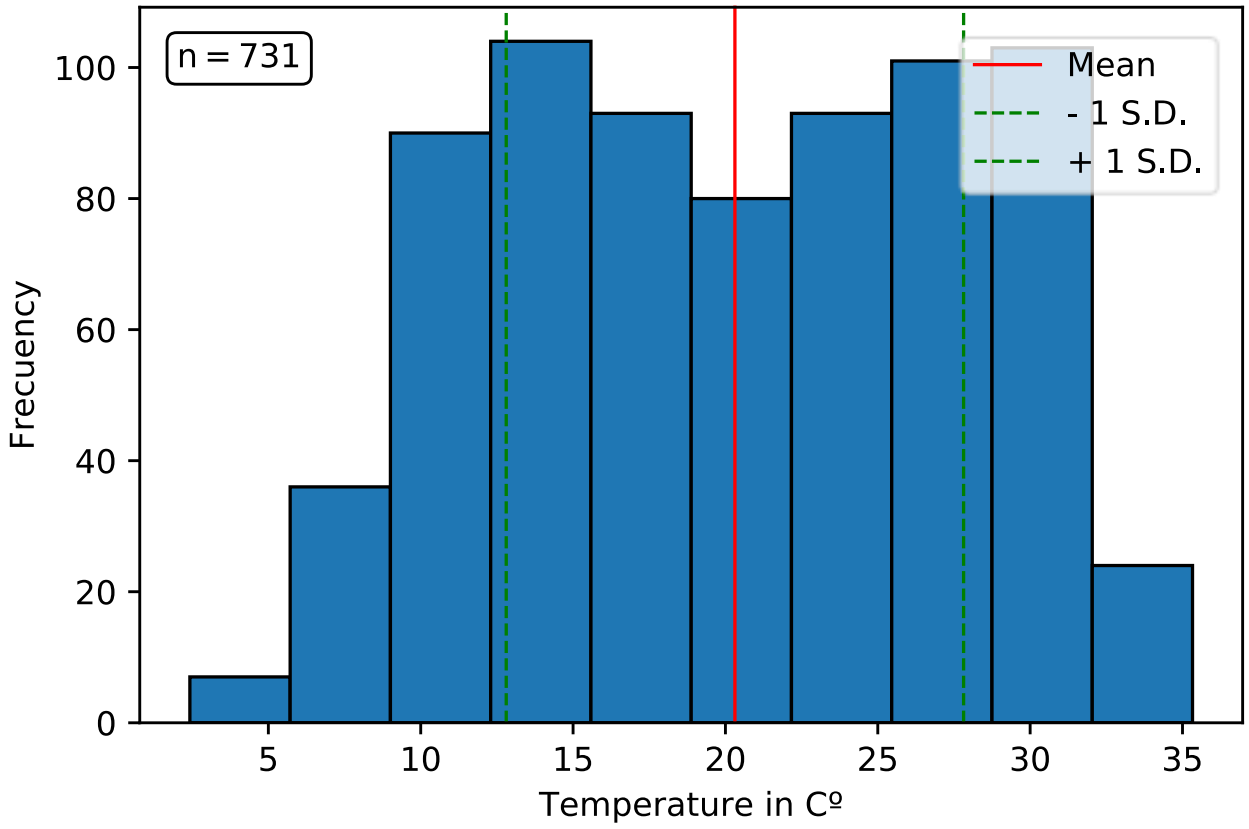
Figure 5. Temperature in Celsius



Describing quantitative variables

EDEM
Centro Universitario

Figure 5. Temperature in Celsius



alberto.sanz@bigwaveanalytics.es

Regression

EDEM
Centro Universitario

1. Describe the two variables involved in hypothesis

Temperature

Rentals

Figure 5. Temperature in Celsius

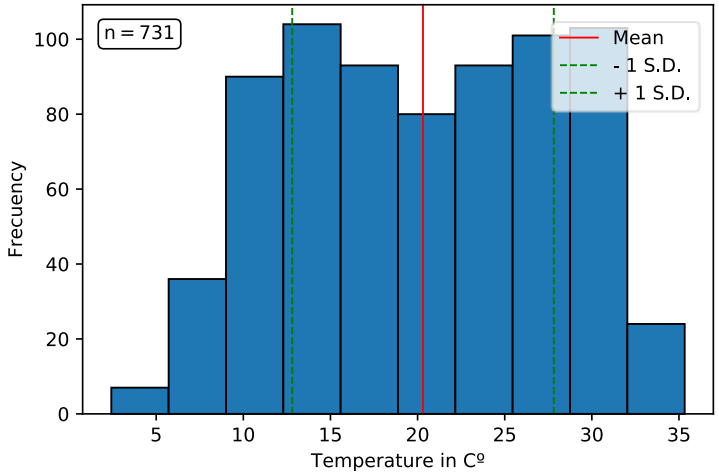
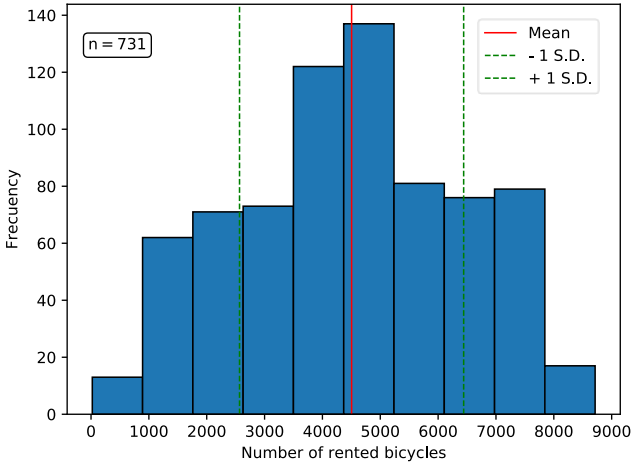


Figure 1. Daily Bicycle rentals in Washington DC
by Capital bikeshare. 2011 - 2012



alberto.sanz@bigwaveanalytics.es

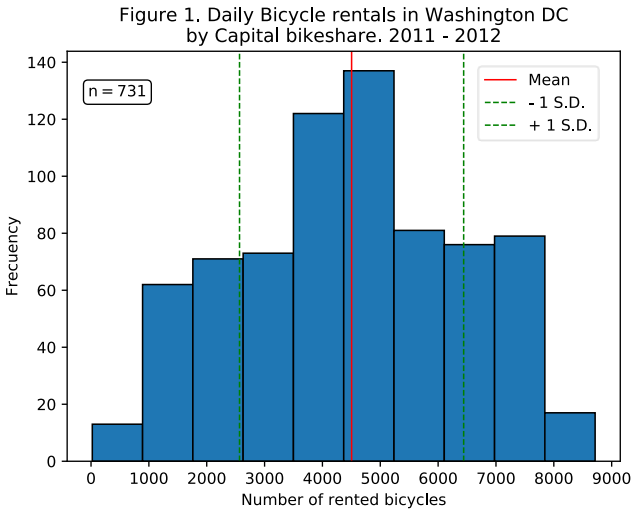
Correlation

EDEM
Centro Universitario

1. Describe the two variables involved in hypothesis

Windspeed

Rentals



alberto.sanz@bigwaveanalytics.es

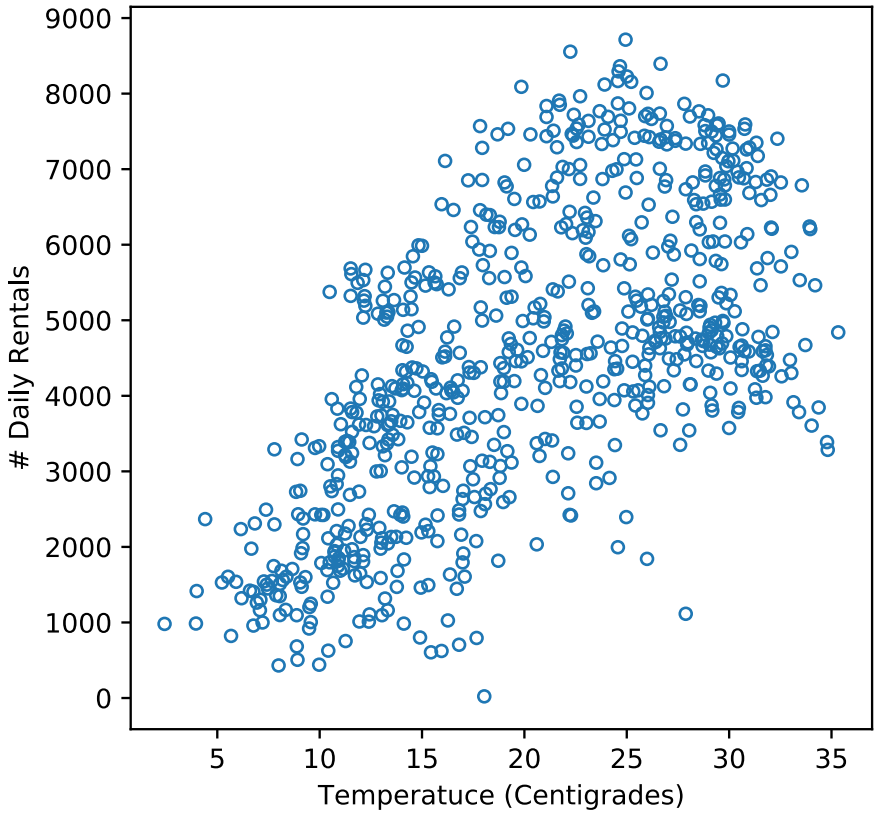
Regression

EDEM
Centro Universitario

2. Scatterplot

```
x=wbr.temp_Celsius
y=wbr.cnt
plt.scatter (x,y)
```

Figure 9. Daily bicycle rentals, by temperature.

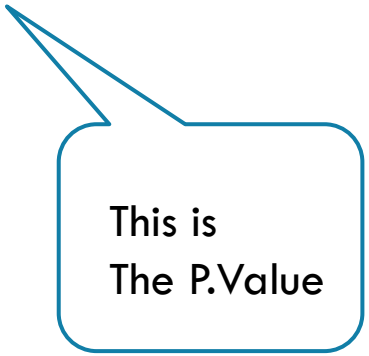


Regression

EDEM
Centro Universitario

3. Pearson's r

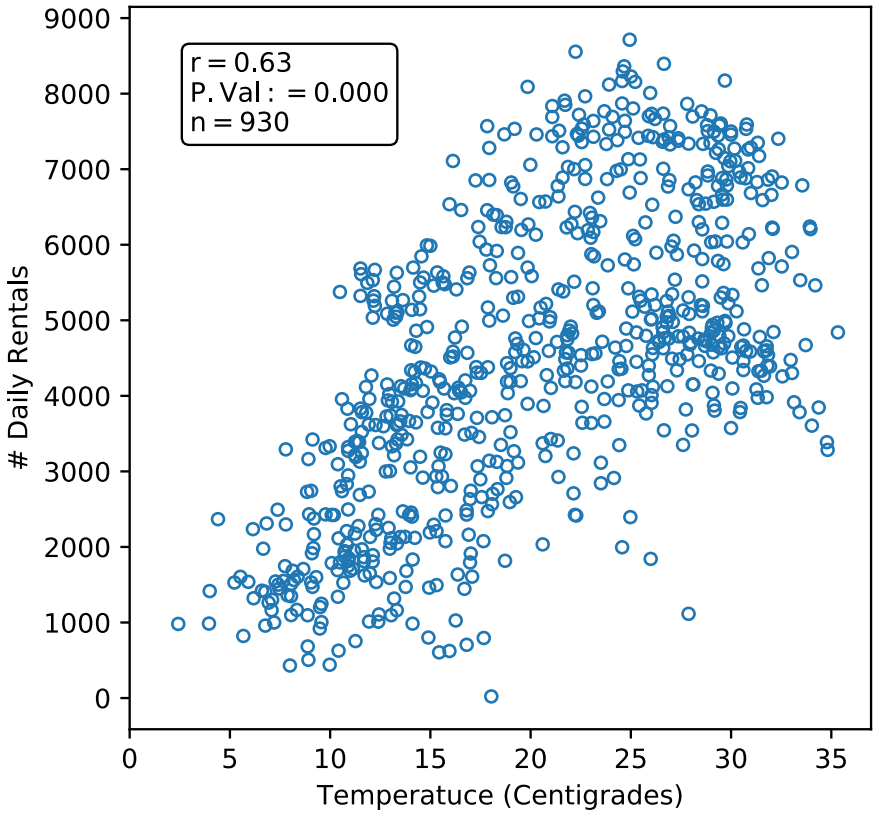
```
from scipy.stats.stats import pearsonr
res = pearsonr(x, y)
print (res)

[1] (0.62749400903349195, 2.8106223975901415e-81)
```



Scatterplot + Pearson's r + test

EDEM
Centro Universitario

Figure 9. Daily bicycle rentals, by temperature.

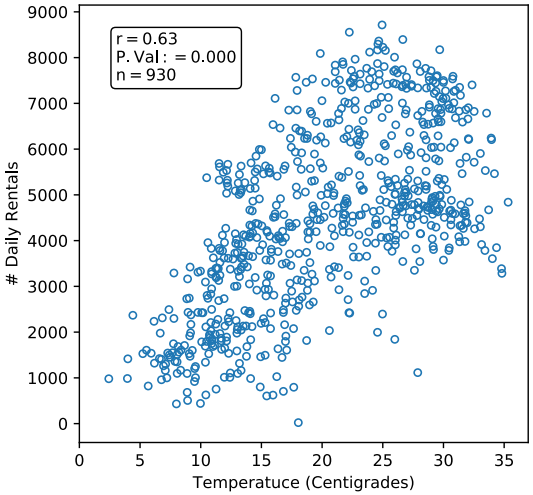


alberto.sanz@bigwaveanalytics.es

Conclussion

EDEM
Centro Universitario

Figure 9. Daily bicycle rentals, by temperature.



Conclussion:

As P. Value < 0.000

We can reject H0 with a confidence higer tan 99.9

✗ H0.: There is no linear association between the *number of rentals* and the *temperature*.

✓ H1.:There is a linear association between the *number of rentals* and the *temperature*.

alberto.sanz@bigwaveanalytics.es

Statistical Programming with Python

EDEM
Centro Universitario

Questions?

Statistical Programming with Python

EDEM
Centro Universitario

Thank you !

Alberto Sanz

alberto.sanz@bigwaveanalytics.es