EDEM
Centro Universitario

Correlation:

Hypothesis testing & Graphic Methods

Session 9

Programación Estadística con Python

Alberto Sanz, Ph.D
asanz@edem.es

MASTER EN DATA ANALYTICS PARA LA EMPRESA

Goals (class session + extra topics)

EDEM
Centro Universitario

- Hypothesis testing over the relatioship of two quantitative variables.
  - Numeric methods:
    - Pearson's r linear correlation coefficient +
    - Significante tests +
    - Kendall's Tau-b / Spearmann Rho (for ordinal variables) (To be developed)
  - Graphic methods:
    - Scatterplots

Correlation

EDEM
Centro Universitario

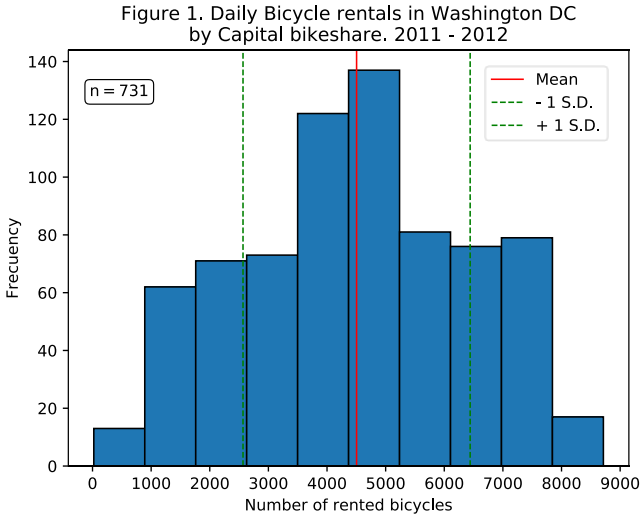
1. **Describe** the two variables involved in the hypothesis separately. Check and validate the integrity of the data prior to any analysis.
2. **Graphic representation** of bivariate relation: **Scatterplot**
3. **Numeric representation** of bivariate relation: **Pearson's r**
4. **Inference test: p.values.**
5. **When posible, combine:**
   - Scatterplot +
   - Pearson's r +
   - Inference test

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*.

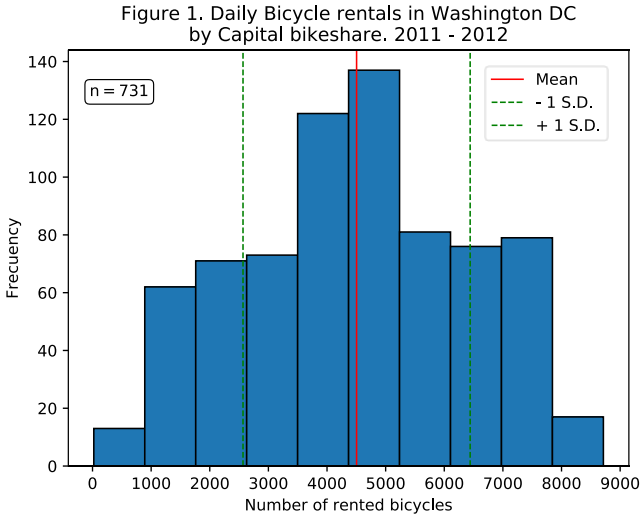
- H1.: There is a linear association ($r\neq0$) between the *number of rentals* and the *temperature*.

Alberto Sanz, Ph.D. asanz@edem.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')
```



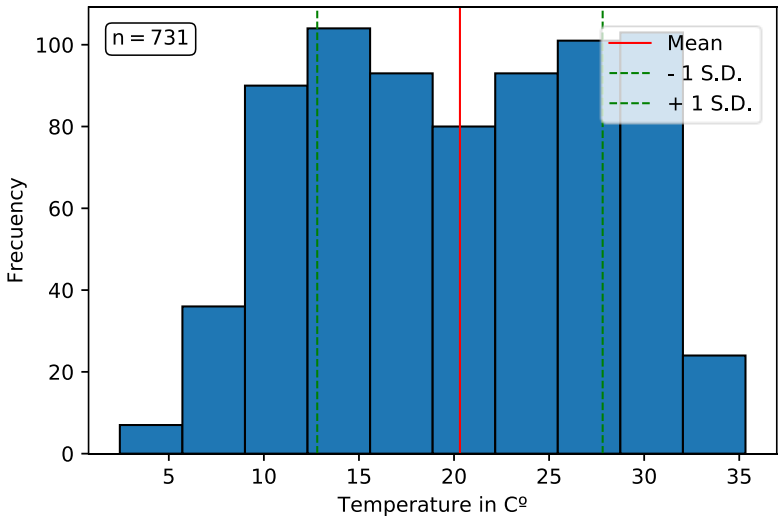

```
props = dict(boxstyle='round',
facecolor='white', lw=0.5)
textstr = '$\mathrm{n}=%.0f$'%(n)
plt.text (-50,128, textstr ,
bbox=props)
```

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



Correlation

EDEM
Centro Universitario

1. Describe the two variables involved in hypothesis

Temperature

Rentals

Figure 5. Temperature in Celsius

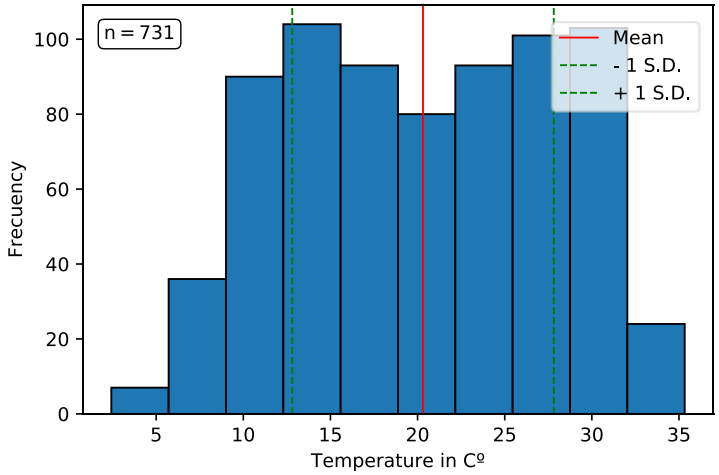
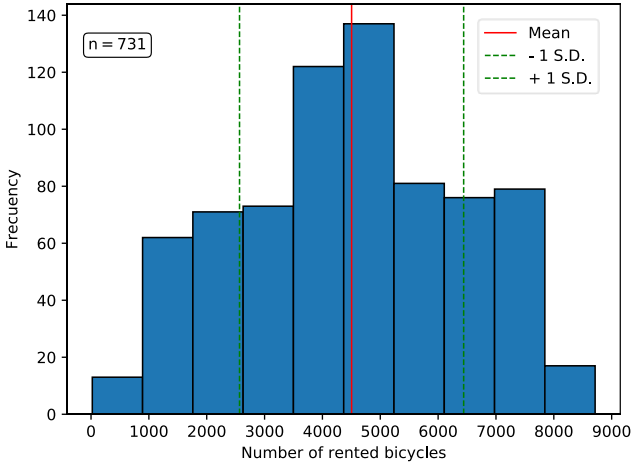


Figure 1. Daily Bicycle rentals in Washington DC by Capital bikeshare. 2011 - 2012



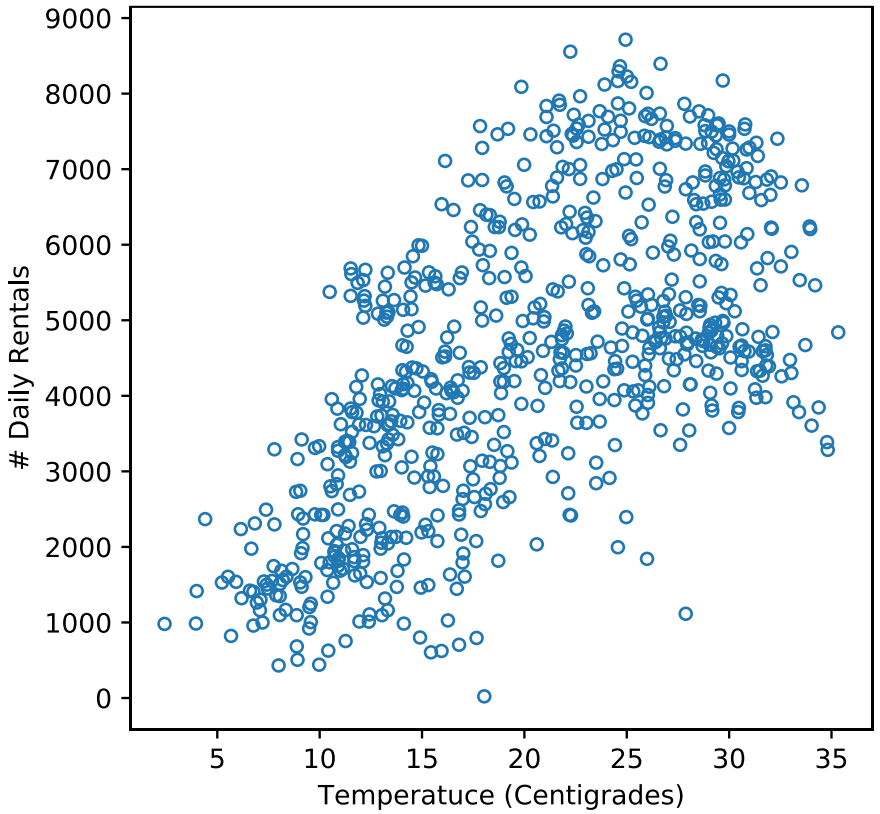
Correlation

EDEM
Centro Universitario

2. Scatterplot

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

Figure 9. Daily bicycle rentals, by temperature.



Correlation

3. Pearson's r

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

[1] (0.6274940090334919 5, 2.8106223975901415e-81)
```

This is
Perason's r

This is
The P.Value

#correlation

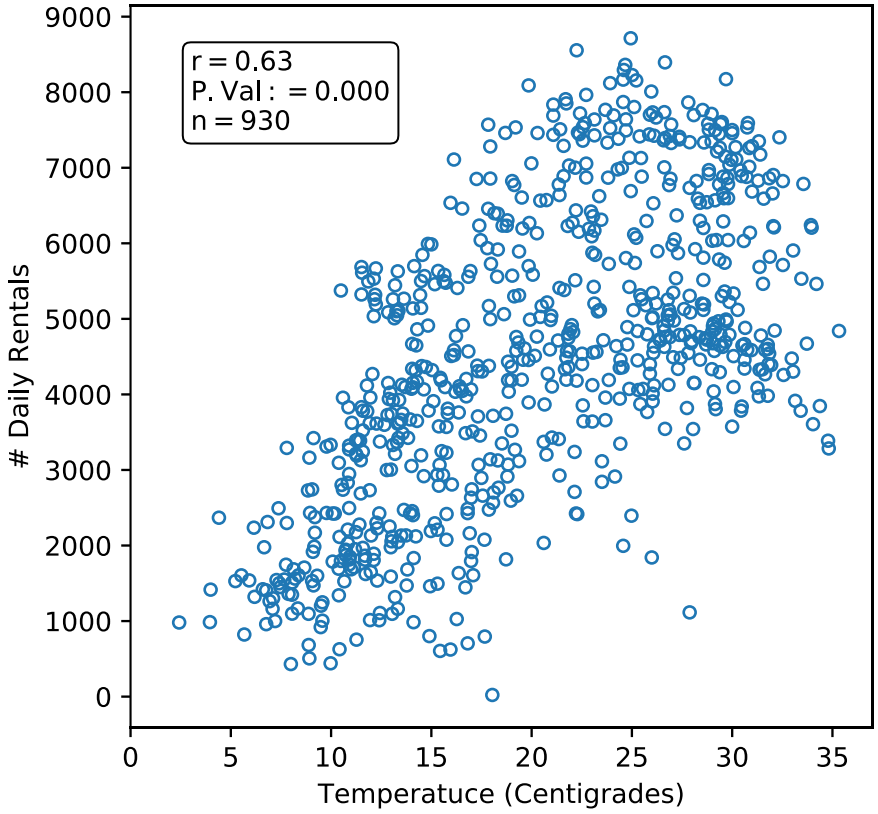
EDEM
Centro Universitario

```
#correlation
from scipy.stats.stats import pearsonr
x=wbr.temp_celsius # Select the variable to plot
y=wbr.cnt # Select the variable to plot
res = pearsonr(x, y)
r  = res[0]
p_val  = res[1]
n = len (wbr.cnt)
```

Scatterplot + Pearson's r + test

EDEM
Centro Universitario

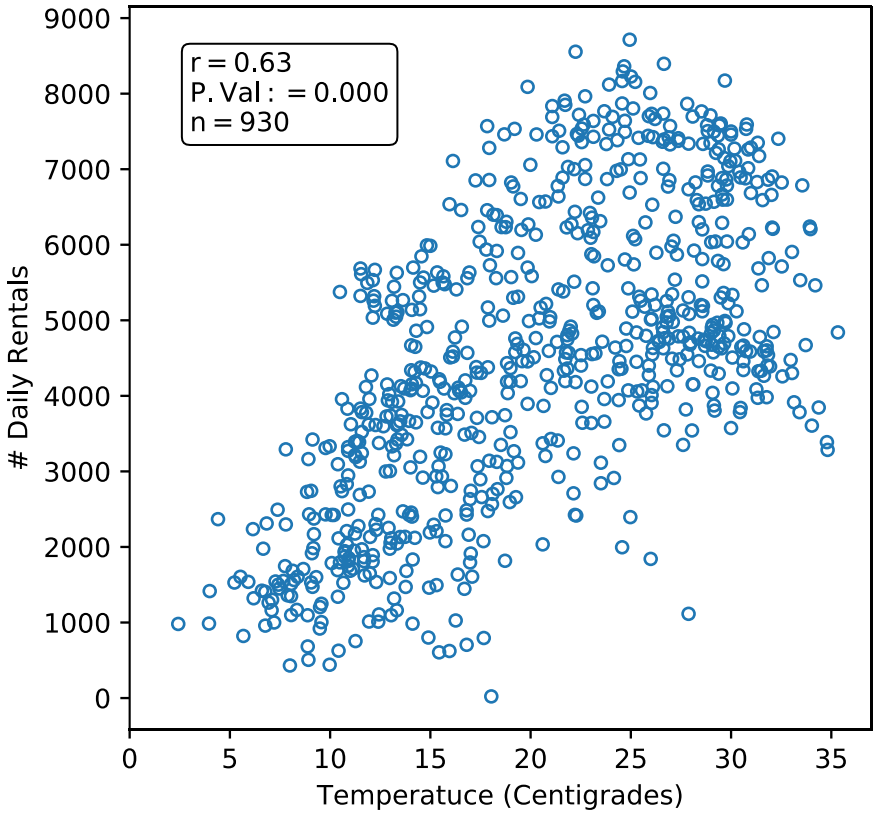
Figure 9. Daily bicycle rentals, by temperature.



Scatterplot + Pearson's r + test

EDEM
Centro Universitario

Figure 9. Daily bicycle rentals, by temperature.



Alberto Sanz, Ph.D. asanz@edem.es

Conclussion

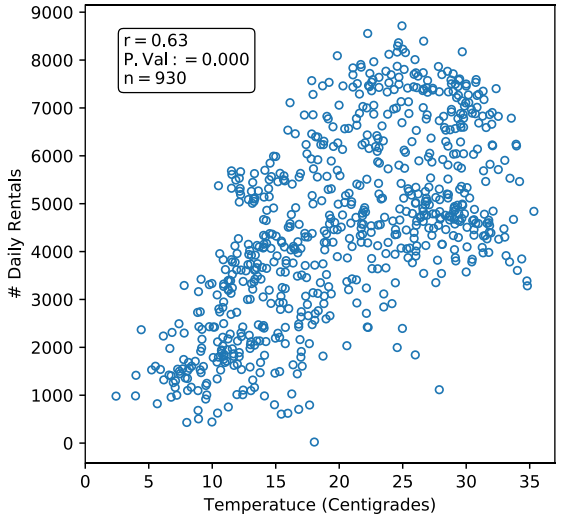
EDEM
Centro Universitario

Conclussion:

As P. Value < 0.000

We can reject H0 with a confidence higer tan 99.9

Figure 9. Daily bicycle rentals, by temperature.



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

✓ H1.: There is a linear association (r≠0) between the *number of rentals* and the *temperature*.

Tricks of the Trade: Split by year…

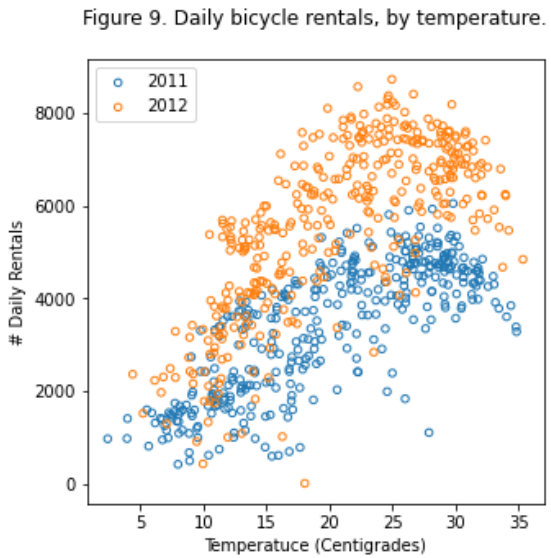
EDEM Centro Universitario


```
#Extra topic
plt.scatter ('temp_celsius', 'cnt',  data=wbr, c='yr')
```

Tricks of the Trade: Split by year

EDEM
Centro Universitario

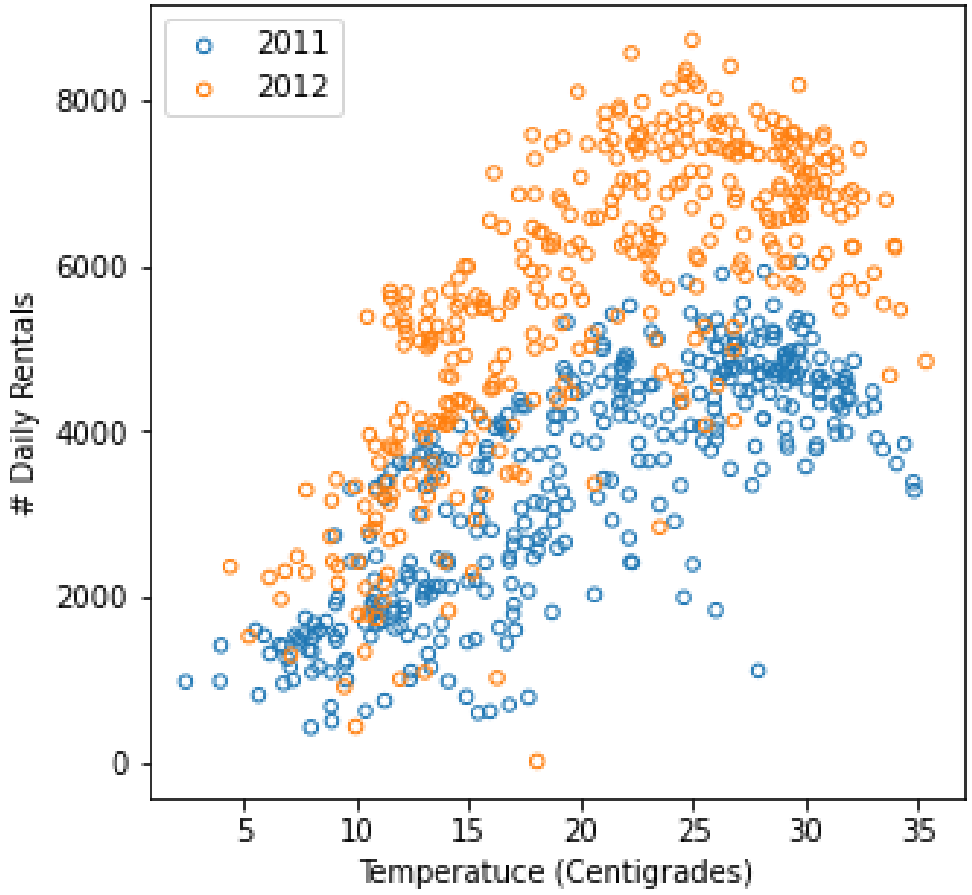

```
plt.figure(figsize=(5, 5))
plt.scatter (wbr.temp_celsius[wbr.yr==0],wbr.cnt[wbr.yr==0], s=20,
             marker="s", facecolors='none', edgecolors='C0', label="2011")
plt.scatter (wbr.temp_celsius[wbr.yr==1],wbr.cnt[wbr.yr==1], s=20,
             facecolors='none', edgecolors='C1', label="2012")
plt.legend(loc="upper right")
props = dict(boxstyle='round', facecolor='white', lw=0.5)
textstr = '$\mathrm{r}=%.2f$\n$\mathrm{P.Val:}=%.3f$\n$\mathrm{n}=%.0f$'%(r,
p_val, n)
plt.text (3,7500, textstr ,   bbox=props)
plt.title('Figure 9. Daily bicycle rentals, by temperature.''\n' )
plt.ylabel('# Daily Rentals')
plt.xlabel('Temperatuce (Centigrades)')
plt.show()
```



Tricks of the Trade: Split by year

EDEM
Centro Universitario

Figure 9. Daily bicycle rentals, by temperature.

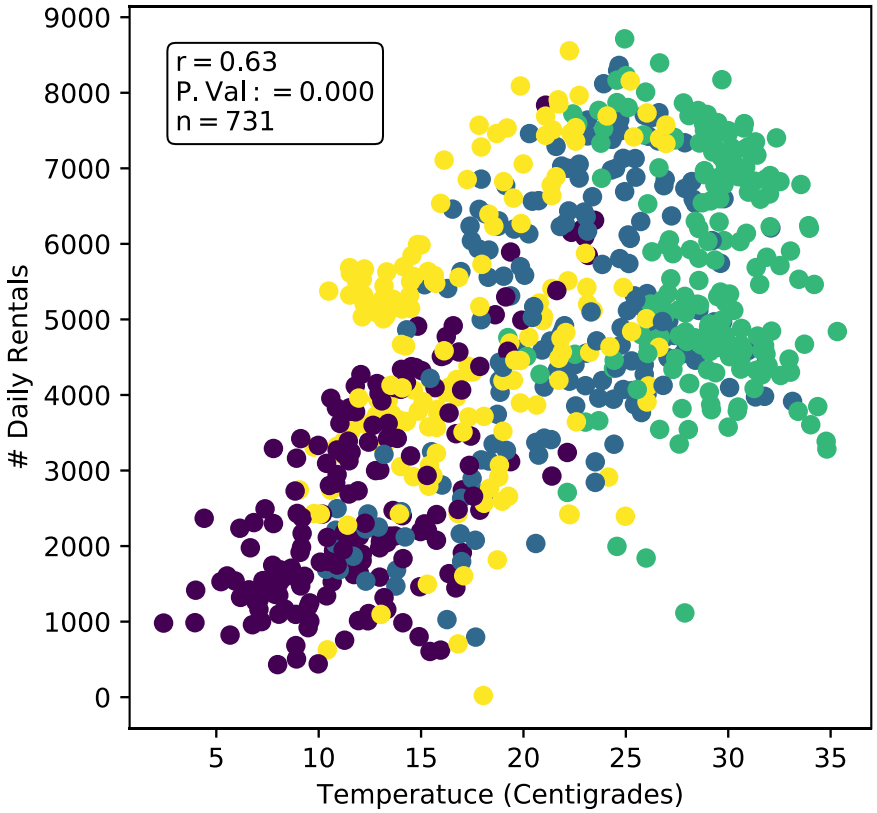


Tricks of the Trade:

EDEM
Centro Universitario

- Split by Season

Figure 9. Daily bicycle rentals, by temperature.



Correlation. Summing UP

EDEM
Centro Universitario

- General Remainder:
  - Allways **describe/explore your data** (numerically + graphically) prior to perform any statistical analysis.

- Main Numeric Procedure:
  - Pearson's Correlation
  - P.Value

- Main Graphic Procedure:
  - Scatterplot

Statistical Programming with Python

EDEM
Centro Universitario

Questions?

EDEM
Centro Universitario

Thank you !

Alberto Sanz

asanz@edem.es