EDEM
Centro Universitario

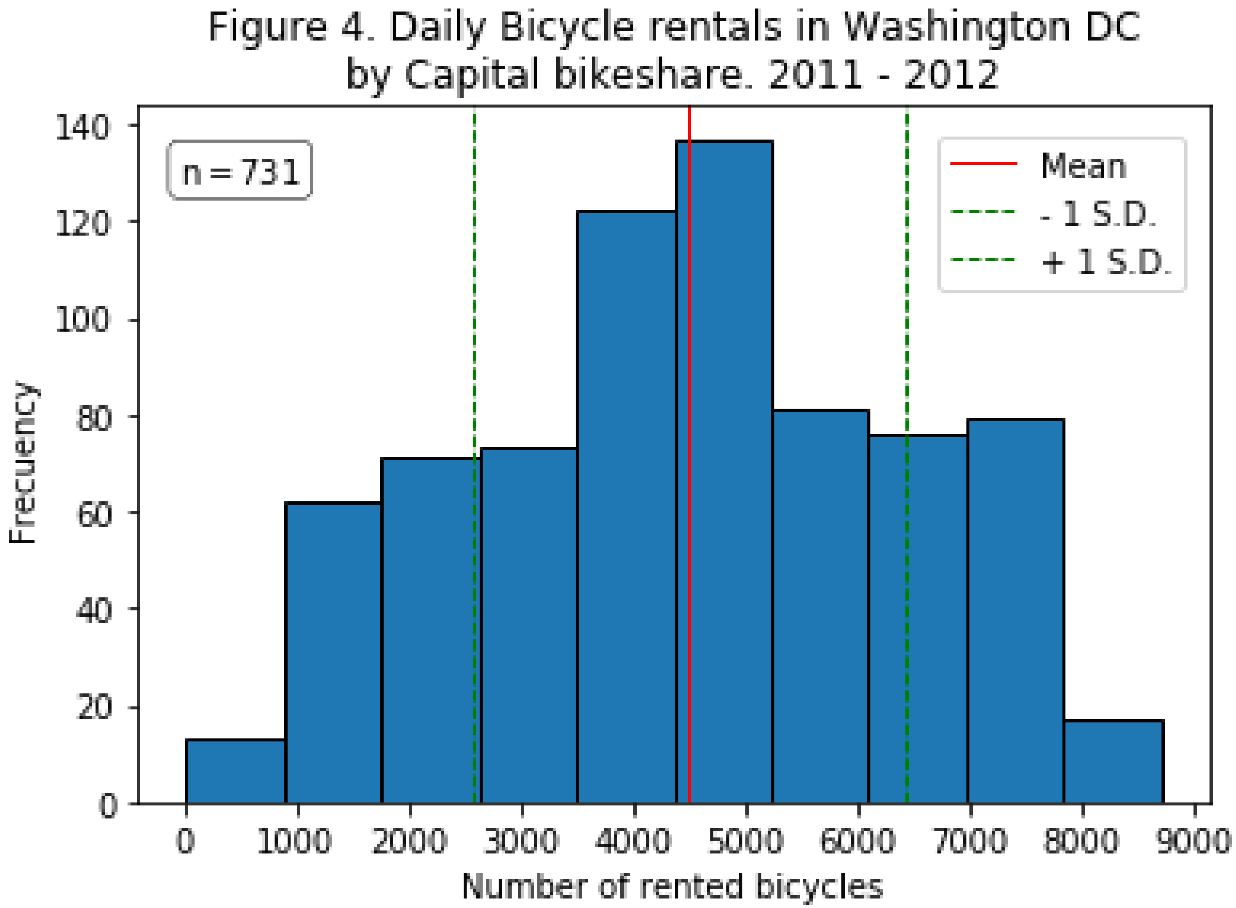
Introducción al análisis de Datos
Programación Estadística con Python

Sesión 7
Mean comparisons

Alberto Sanz, Ph.D
alberto.sanz@bigwaveanalytics.es
www.linkedin.com/in/alberto-sanz-4b6bb5106

MASTER EN DATA ANALYTICS PARA LA EMPRESA

Describing quantitative variables



Research Question

EDEM
Centro Universitario

Why some days are rent *more* bikes than other days in Washington D.C.?

□ **working days** → ?

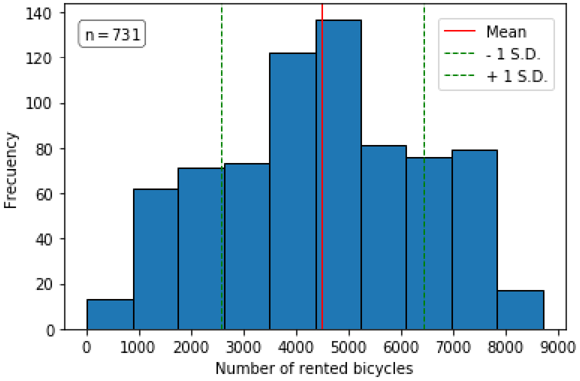


Figure 4. Daily Bicycle rentals in Washington DC by Capital bikeshare. 2011 - 2012

□ H0.: *μ* rentals in **working days** = *μ* rentals in **holidays**

□ H1.: *μ* rentals in **working days** ≠ *μ* rentals in **holidays**

Mean comparison (2 groups)

- □ H0.: μ rentals in **working days** = μ rentals in **holidays**
- □ H1.: μ rentals in **working days** ≠ μ rentals in **holidays**

- **Numeric Procedure** ⇨ **t test for independent samples**
- **Graphic procedure** ⇨ **confidence interval plot**

Mean comparison (2 groups)

1. **Describe the two variables involved in the hypothesis**

2. **Perform the numeric test:** t.test

3. **Perform the graphic test:** plot of the means

4. **When posible:** combine both numeric and graphic in same plot

Mean comparison (2 groups)

1. Describe the two variables involved in hypothesis

Rentals

```
wbr.cnt.describe()
plt.hist(wbr.cnt)
```

Working days

```
mytable = pd.crosstab(index=wbr["wd_cat"], 
columns="count")
n=mytable.sum()
mytable2 = (mytable/n)*100
plt.bar(mytable2.index, mytable2['count'])
```

Figure 4. Daily Bicycle rentals in Washington DC by Capital bikeshare. 2011 - 2012

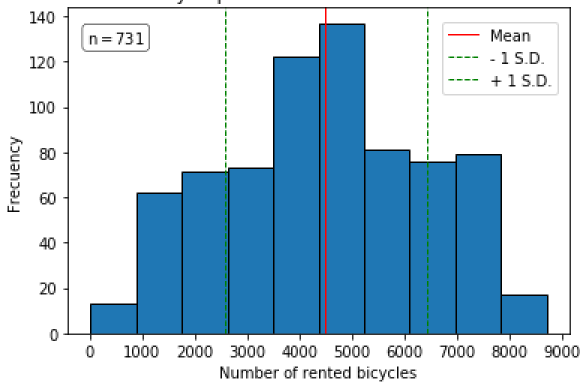
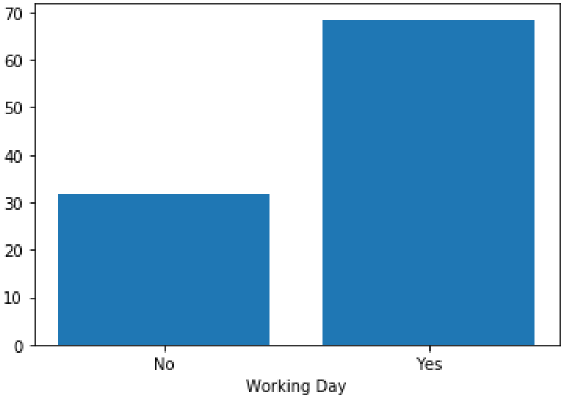


Figure 5. Percentage of Working days



Mean comparison (2 groups)

2. Perform the numeric test: t.test

```
#Descriptive comparison:
wbr.groupby('wd_cat').cnt.mean()

#Statistical comparison:
#Extract the two sub samples and store them in two objects
cnt_wd=wbr.loc[wbr.wd_cat=='Yes', "cnt"]
cnt_nwd=wbr.loc[wbr.wd_cat=='No', "cnt"]

#Perform a t test for mean comparison
#import scipy.stats as stats
stats.ttest_ind(cnt_wd, cnt_nwd, equal_var = False)
```

```
Output:
wd_cat
No      4330.168831
Yes     4584.820000

Ttest_indResult(statistic= 1.60137, pvalue = 0.1105)
```

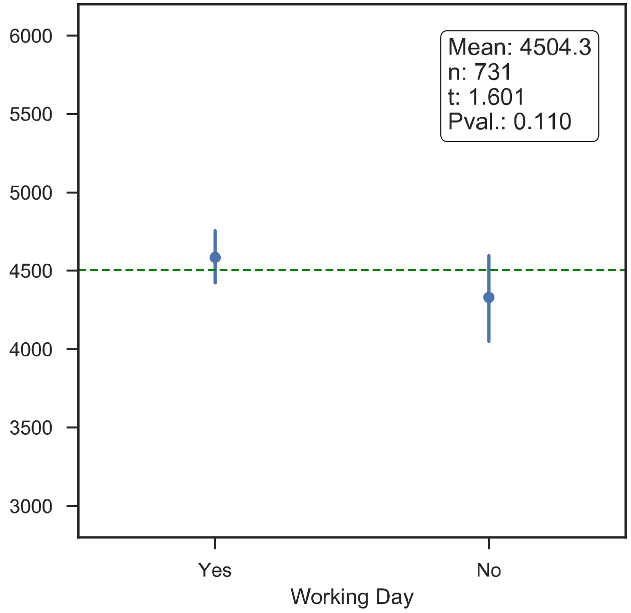
Mean comparison (2 groups)

3. Perform the mean comparison graphic test (I)

3.1. Define parameters & plot

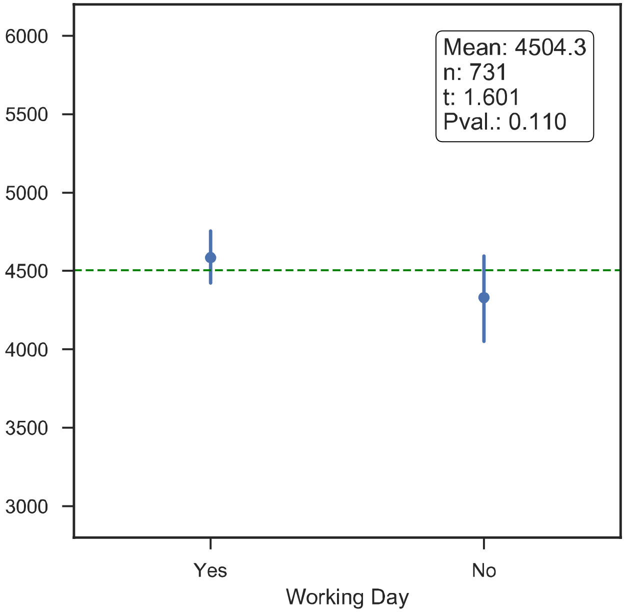
```
#CI meanplot
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(5,5))
ax = sns.pointplot(x="wd_cat", y="cnt",
                   data=wbr,ci=95, join=0)
plt.yticks(np.arange(3000, 7000, step=500))
plt.ylim(2800,6200)
plt.axhline(y=wbr.cnt.mean(),
            linewidth=1,
            linestyle= 'dashed',
            color="green")
props = dict(boxstyle='round’,
             facecolor='white', lw=0.5)
plt.text(0.85,5400,'Mean:4504.3''\n''n:731' '\n' 't:1.601' '\n' 'Pval.:0.110',    bbox=props)
plt.xlabel('Working Day')
plt.title('Figure 6. Average rentals by Working Day.''\n')
```

Figure 6. Average rentals by Working Day.



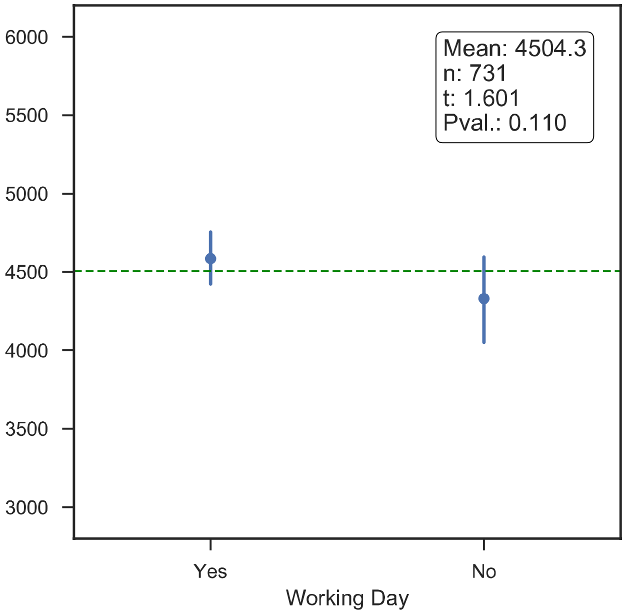
Mean comparison (2 groups)

Figure 6. Average rentals by Working Day.



Mean comparison (2 groups)

Figure 6. Average rentals by Working Day.



✓ H0.: *μ* rentals in **work days** = *μ* rentals in **holidays**

✗ H1.: *μ* rentals in **work days** ≠ *μ* rentals in **holidays**

⇨

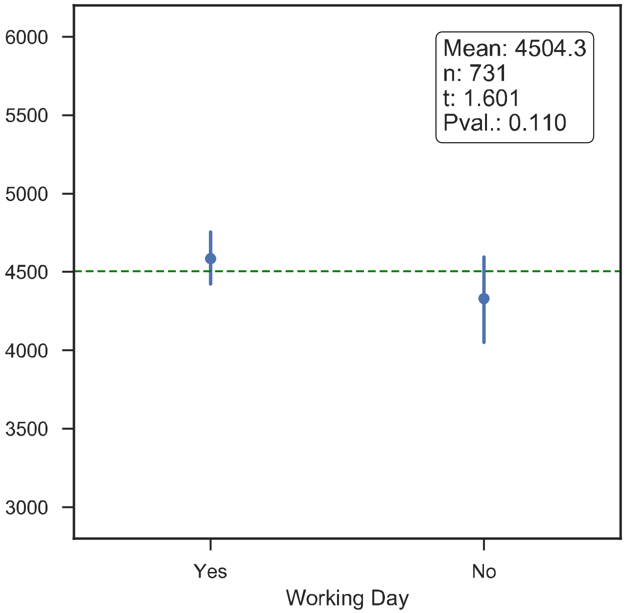
⇨

CONCLUSION:
As P. Val > 0.05, we do NOT REJECT H0.:

In other words:
Average rentals do not significantly differ in Working days and Non working days.

Mean comparison (2 groups)

Figure 6. Average rentals by Working Day.



CONCLUSION:
As P. Val > 0.05

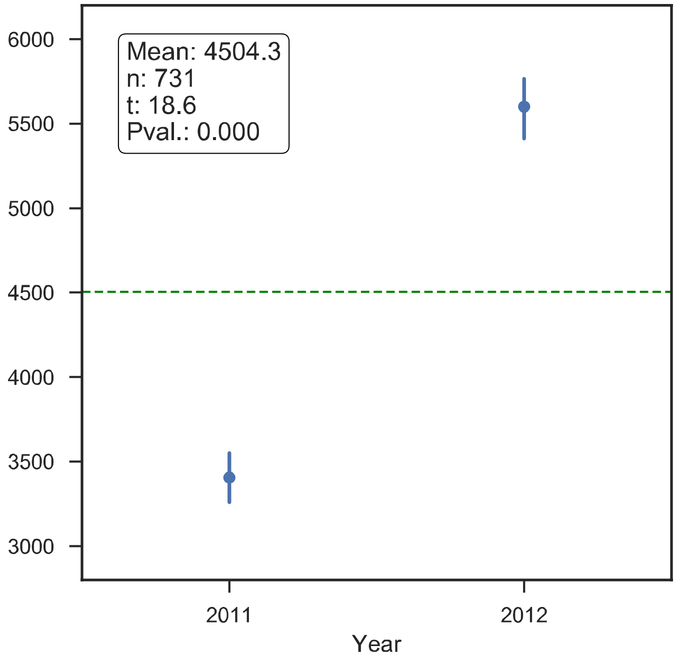
Average rentals do not significantly differ in Working days and Non working days.

Mean comparison (2 gr.) Example #2

- ✗ □ H0.: μ rentals in 2011= μ rentals in 2012
- ✓ □ H1.: μ rentals in 2011≠ μ rentals in 2012

```
#Plotmeans
plt.figure(figsize=(5,5))
ax=sns.pointplot(x="yr",y="cnt",data=wbr,ci=95,join=0)
ax.set_ylabel('')
plt.yticks(np.arange(3000, 7000, step=500))
plt.ylim(2800,6200)
plt.axhline(y=wbr.cnt.mean(),
            linewidth=1,
            linestyle= 'dashed',
            color="green")
props = dict(boxstyle='round', facecolor='white', lw=0.5)
plt.xticks((0,1), ("2011", "2012"))
plt.xlabel('Year')
plt.title('Figure 7. Average rentals by Year.''\n')
```

Figure 7. Average rentals by Year.



```
plt.text(-0.35,5400,'Mean:4504.3''\n''n:731' '\n' 't:18.6' '\n' 'Pval.: 0.000',bbox=props)
```

A Panel of results:

Figure 6. Average rentals by Working Day.

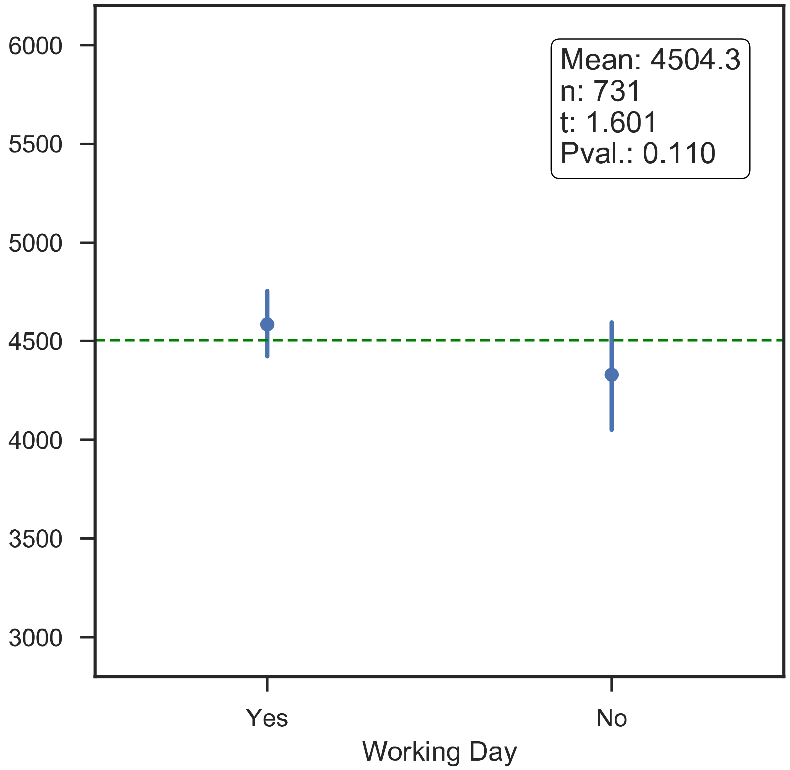
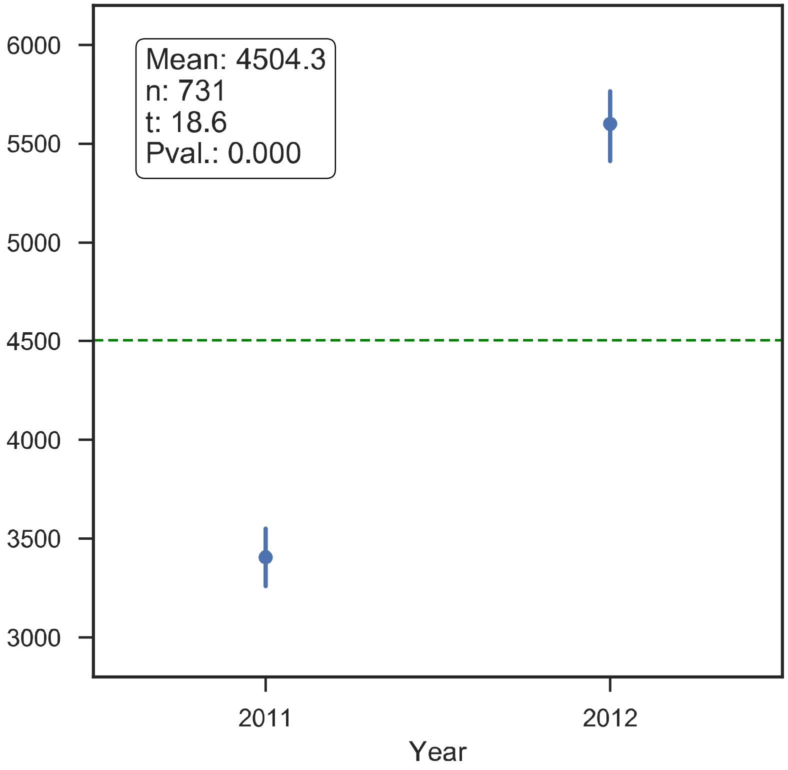


Figure 7. Average rentals by Year.



A Panel of results:

Figure 6. Average rentals by Working Day.

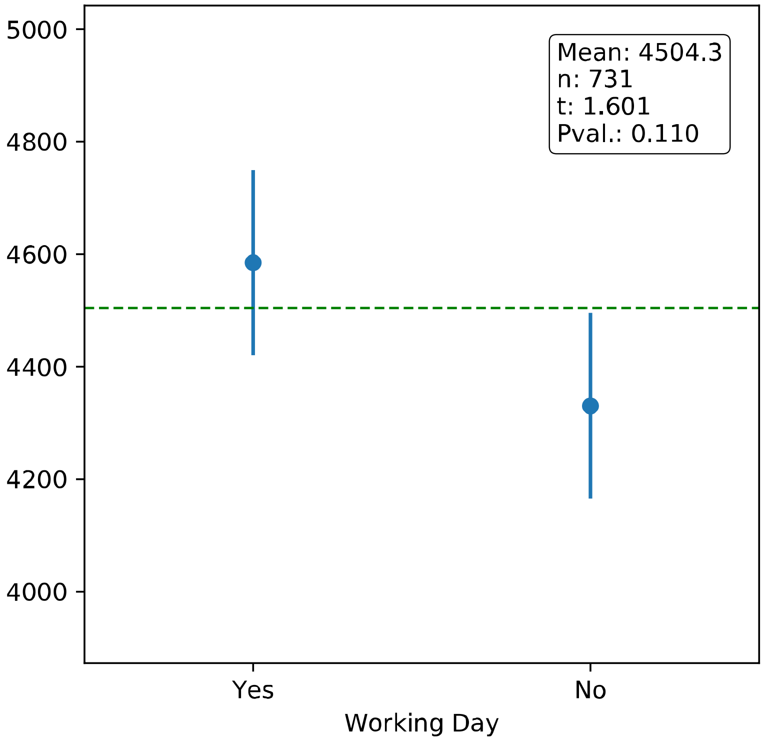
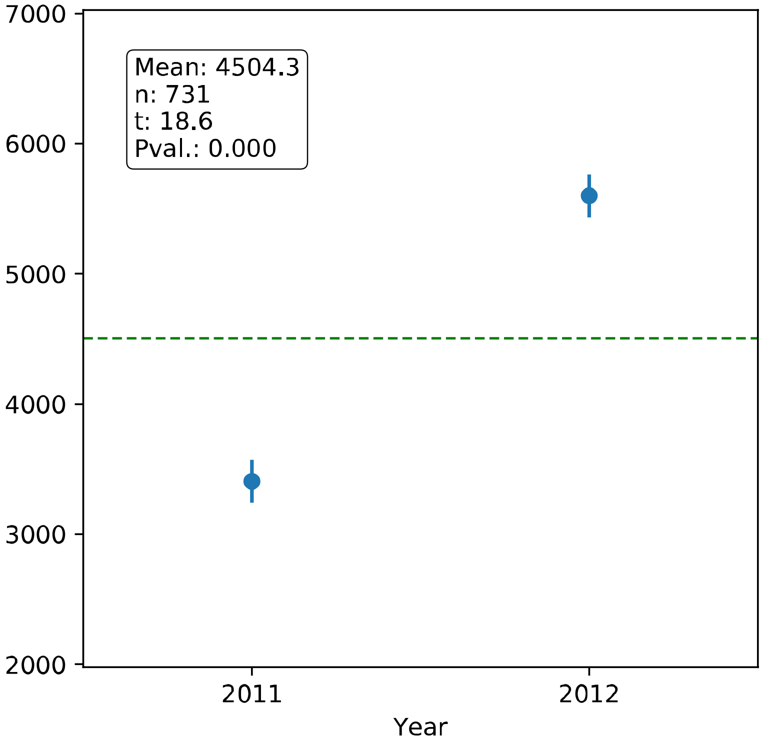
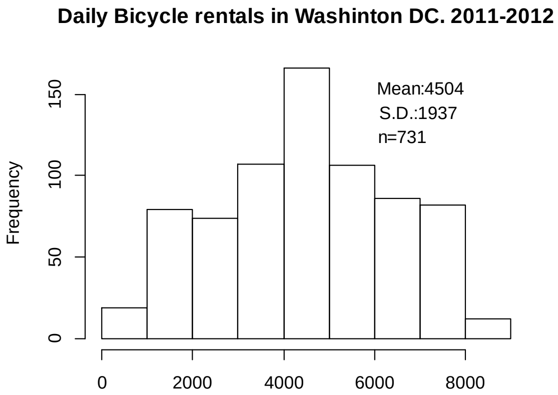


Figure 7. Average rentals by Year.



Research Question

Why some days are rent *more* bikes than other days in Washington D.C.?



- □ H0.: μ rentals **sunny** = μ rentals **cloudy**= μ rentals **stormy.**
- □ H1.: μ rentals differ in **at least** 2 of the 3 groups compared.

Mean comparison (> 2 groups)

- □ H0.: *μ* rentals **sunny** = *μ* rentals **cloudy**= *μ* rentals **stormy.**
- □ H1.: *μ* rentals differ in **at least** 2 of the 3 groups compared

- **Numeric Procedure** ⇨ **One-Way ANOVA**
- **Graphic procedure** ⇨ **Confidence interval plot**

Mean comparison (> 2 groups)

1. **Describe the two variables involved in the hypothesis**

2. **Perform the numeric test:** One-Way ANOVA

3. **Perform the graphic test:** plot of the means

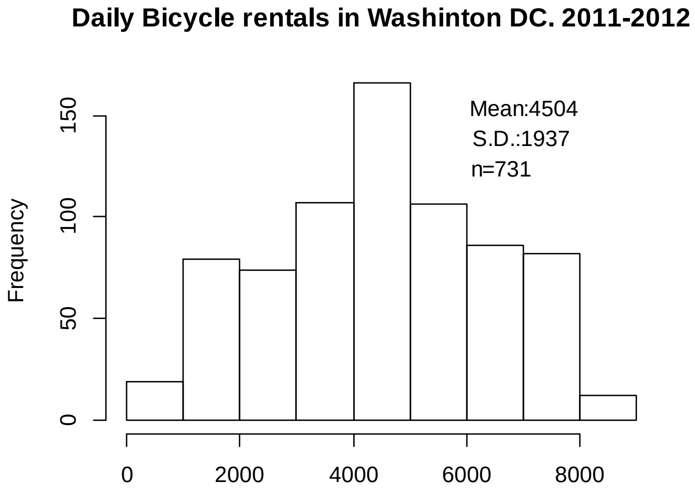
4. **When posible:** combine both numeric and graphic in same plot.

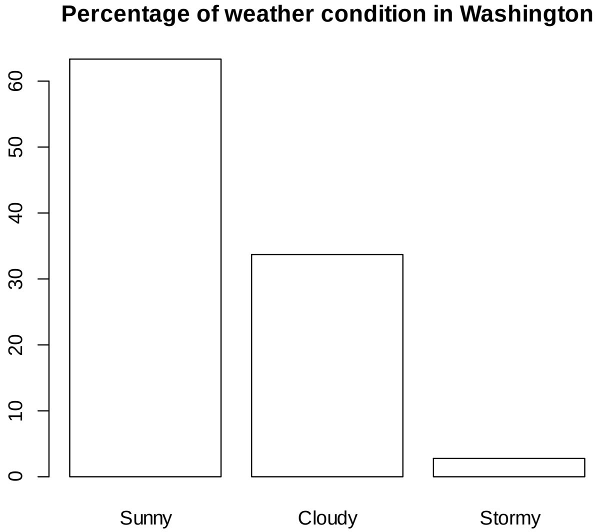
Mean comparison (> 2 groups)

1. Describe the two variables involved in hypothesis

Rentals

Weather condition





Mean comparison (> 2 groups)

2. **Perform the numeric test:** One-Way ANOVA

```
##Descriptive comparison
wbr.groupby('ws_cat').cnt.mean()

#Statistical comparison
cnt_sunny=wbr.loc[wbr.ws_cat=='Sunny', "cnt"]
cnt_cloudy=wbr.loc[wbr.ws_cat=='Cloudy', "cnt"]
cnt_rainy=wbr.loc[wbr.ws_cat=='Rainy', "cnt"]

stats.f_oneway(cnt_sunny, cnt_cloudy,cnt_rainy )
```

```
OUTPUT:
Sunny     4876.786177
Cloudy    4035.862348
Rainy     1803.285714

F_onewayResult(statistic=40.0660, pvalue=3.10631e-17)
```

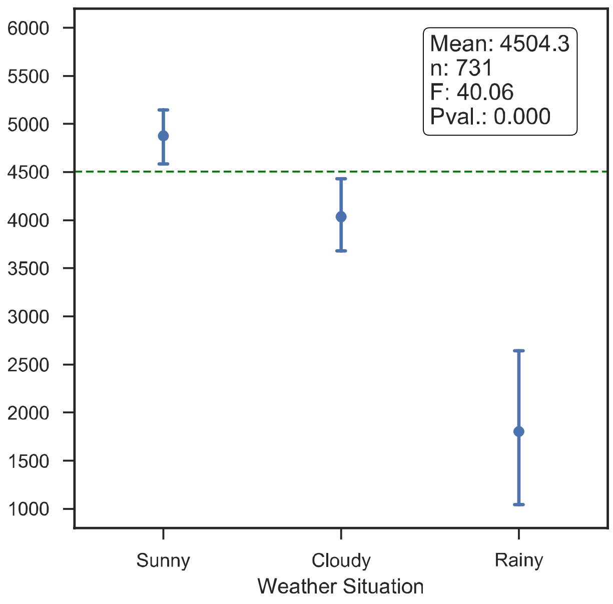
Interpretation.
As P.Value < 0.05: **REJECT** the H0 about equality of the means in all groups.
In other words: **at leats two groups differ** in average bicycle rentals

Mean comparison (> 2 groups)

3. **Perform the graphic test:** plot of the means

```
#Graphic comparison: confidence intervals for the means
plt.figure(figsize=(5,5))
ax = sns.pointplot(x="ws_cat", y="cnt", data=wbr, capsize=0.05,
ci=99.9, join=0)
ax.set_ylabel('')
plt.yticks(np.arange(1000, 7000, step=500))
plt.ylim(800,6200)
plt.axhline(y=wbr.cnt.mean(),
            linewidth=1,
            linestyle= 'dashed',
            color="green")
props = dict(boxstyle='round', facecolor='white', lw=0.5)
plt.text(1.5, 5000, 'Mean: 4504.3''\n''n: 731' '\n' 'F: 40.06'
'\n' 'Pval.: 0.000',    bbox=props)
plt.xlabel('Weather Situation')
plt.title('Figure 8. Average rentals by Weather Situation.''\n')
```

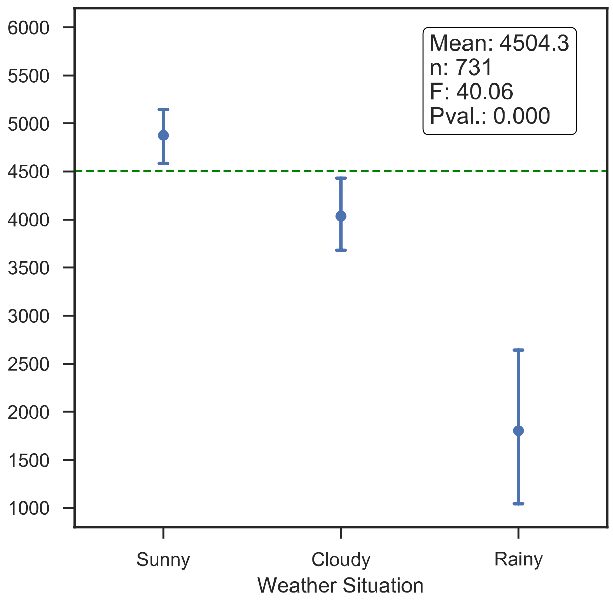
Figure 8. Average rentals by Weather Situation.



Mean comparison (> 2 groups)

4. Combine graphic & numeric tests

Figure 8. Average rentals by Weather Situation.



✗ H0.: μ rentals **sunny** = μ rentals **cloudy**= μ rentals **stormy.**

✓ H1.: μ rentals differ in **at least** 2 of the 3 groups compared

CONCLUSION:

As P. Value < 0.05*, we do REJECT H0.:

In other words:

Different weather conditions are significantly associated to **diffefrnt average in rentals.**

* Note: In this specific case, as p.value is indeed < 0.01, we reject H0 with a confidence level larger tan 99 percent.

Mean Comparison Summing UP

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

- Main Graphic Procedure:
  - Confidence interval plot

- Main Numeric Procedures:
  - 2 Groups : t test
  - >2 Groups: One-way ANOVA

Statistical Programming with Python

Questions?

Thank you !

Alberto Sanz

alberto.sanz@bigwaveanalytics.es

www.linkedin.com/in/alberto-sanz-4b6bb5106