EDEM
Centro Universitario

Introducción al análisis de Datos
Programación Estadística con Python

Sesión 8

Percentage comparisons.
Descriptive analytics, hypothesis testing & Graphic methods

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

MASTER EN DATA ANALYTICS PARA LA EMPRESA

Goals

EDEM
Centro Universitario

- Hypothesis testing over the relatioship of two nominal (categorical) variables.
  - Numeric methods:
    - Cross tabulations (Descriptive, sample level) +
    - Chi2 (Hypothesis testing, at the **population** level) +
    - Cramer's V (Strength of the association, at the **pop.** level) (To be developed)
  - Graphic methods:
    - Grouped barplots.
    - Mosaic plots. (To be developed)

Our Dependent Variable

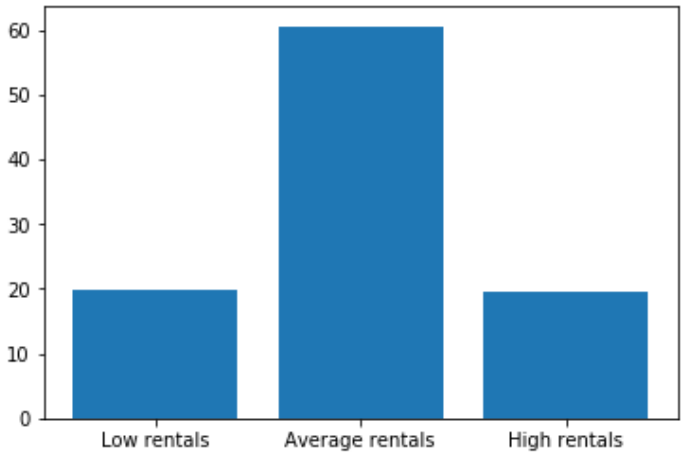
EDEM
Centro Universitario


```
#Recoding DV for analysis
res = wbr.cnt.describe()
m  = res[1]
sd = res[2]
n  = res[0]

### Recode cnt to string
wbr.loc[(wbr['cnt']<(m-sd)),"cnt_str"]= "Low rentals"
wbr.loc[((wbr['cnt']>(m-
sd))&(wbr['cnt']<(m+sd))),"cnt_str"]="Average rentals"
wbr.loc[  (wbr['cnt']>(m+sd)),"cnt_str"]="High rentals"

### Recode cnt to ordinal
my_categories=["Low rentals", "Average rentals", "High
rentals"]
my_rentals_type =
CategoricalDtype(categories=my_categories, ordered=True)
wbr["cnt_cat"] = wbr.cnt_str.astype(my_rentals_type)
wbr.info()

#Percentage table & barchart
mytable = pd.crosstab(wbr.cnt_cat, columns="count",
normalize='columns')*100

plt.bar(mytable.index, mytable['count'])
```

Table 1. Percentage of days with different rentals in Washington D.C.

Low rentals	19,8
Average rentals	60,6
High rentals	19,56
TOTAL	100,0
(n)=731	

Source: Own analyses over Fanaee, Hadi and Gama (2013) data



Percentage comparison (*n* groups)

EDEM Centro Universitario

1. **Describe the two variables involved in the hypothesis separately. Special atention to be paid at the distribution of the DV***
2. **Describe the DV, by the levels in the IV**
   (Cross tabulation of DV by IV)
3. **Perform the numeric test for inference: Chi^2 test**
4. **Graphic representation: Combined barplot**
5. **When posible, combine:**
   1. Crosstabs + inference test (as footnote)
   2. Combined barplot + inference test (as text insert)

* DV stands for Dependent variable. IV stands for Independent Variable

alberto.sanz @bigwaveanalytics.es

Research Question

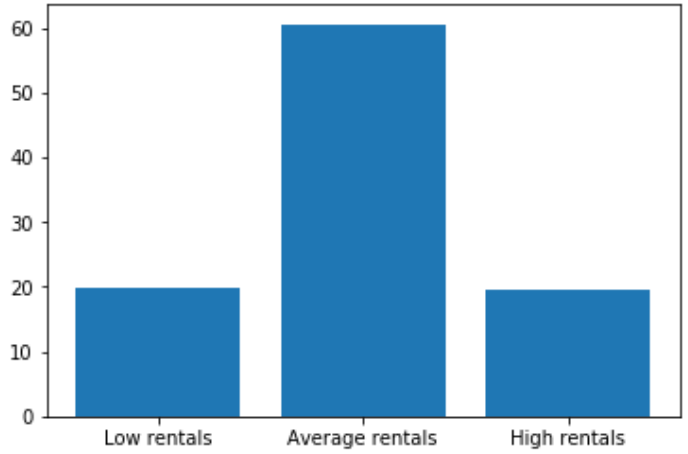
EDEM
Centro Universitario

Why some days are rent *more* bikes?

Table 1. Percentage of days with different rentals in Washington D.C.

Low rentals	19,8
Average rentals	60,6
High rentals	19,56
TOTAL	100,0
(n)=731	

Source: Own analyses over Fanaee, Hadi and Gama (2013) data



- H0.: Percentage of days with *low/average/high* rentals is the same in working days vs. not working days.

- H1.: Percentage of days with *low/average/high* rentals differes in working days vs. not working days.

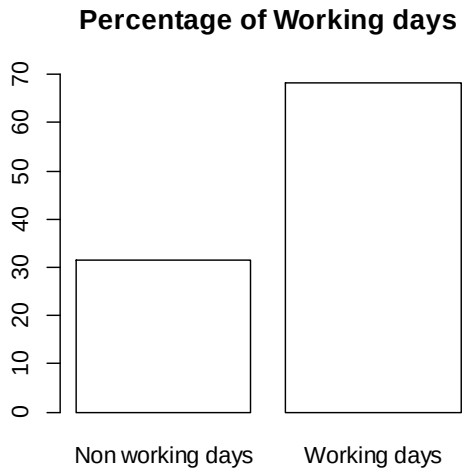
Percentage comparison (*n* groups)

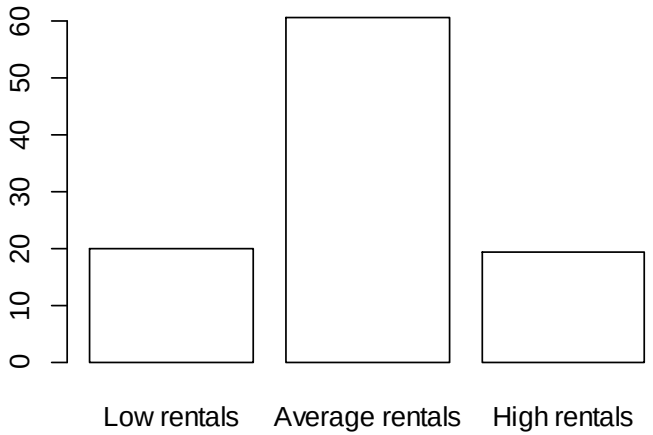
EDEM Centro Universitario

1. Describe the two variables involved in hypothesis

Working days

Rentals





Percentage comparison (*n* groups)

EDEM Centro Universitario

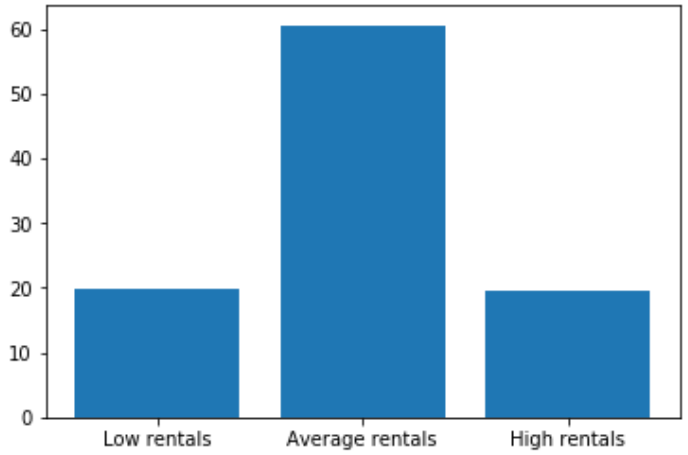
1. Special atention to the distribition of Dependent Variable

Rentals

Table 1. Percentage of days with different rentals in Washington D.C.

Low rentals	19,8
Average rentals	60,6
High rentals	19,56
TOTAL	100,0
(n)=731	

Source: Own analyses over Fanaee, Hadi and Gama (2013) data



Percentage comparison (*n* groups)

EDEM Centro Universitario

Use DV as reference

Table 2. Rental levels in Washington, by type of day. (In percentage points)

	Non working days	Working days	All days
Low rentals	24,7	17,6	19,9
Average rentals	57,1	62,3	60,7
High rentals	18,2	20,0	19,5
TOTAL	100	100	100
(n)=731			

Source: Own analyses over Fanaee, Hadi and Gama (2013) data.

Percentage comparison (*n* groups)

EDEM Centro Universitario

Equal or different?

Use DV as reference

Table 2. Rental levels in Washington, by type of day. (In percentage points)

	Non working days	Working days	All days
Low rentals	24,7	17,6	19,9
Average rentals	57,1	62,3	60,7
High rentals	18,2	20,0	19,5
TOTAL	100	100	100
(n)=731			

Source: Own analyses over Fanaee, Hadi and Gama (2013) data.

Percentage comparison (*n* groups)

EDEM Centro Universitario

Equal or different?

Table 2. Rental levels in Washington, by type of day. (In percentage points)

	Non working days	Working days	All days
Low rentals	24,7	17,6	19,9
Average rentals	57,1	62,3	60,7
High rentals	18,2	20,0	19,5
TOTAL	100	100	100
(n)=731			

Source: Own analyses over Fanaee, Hadi and Gama (2013) data.

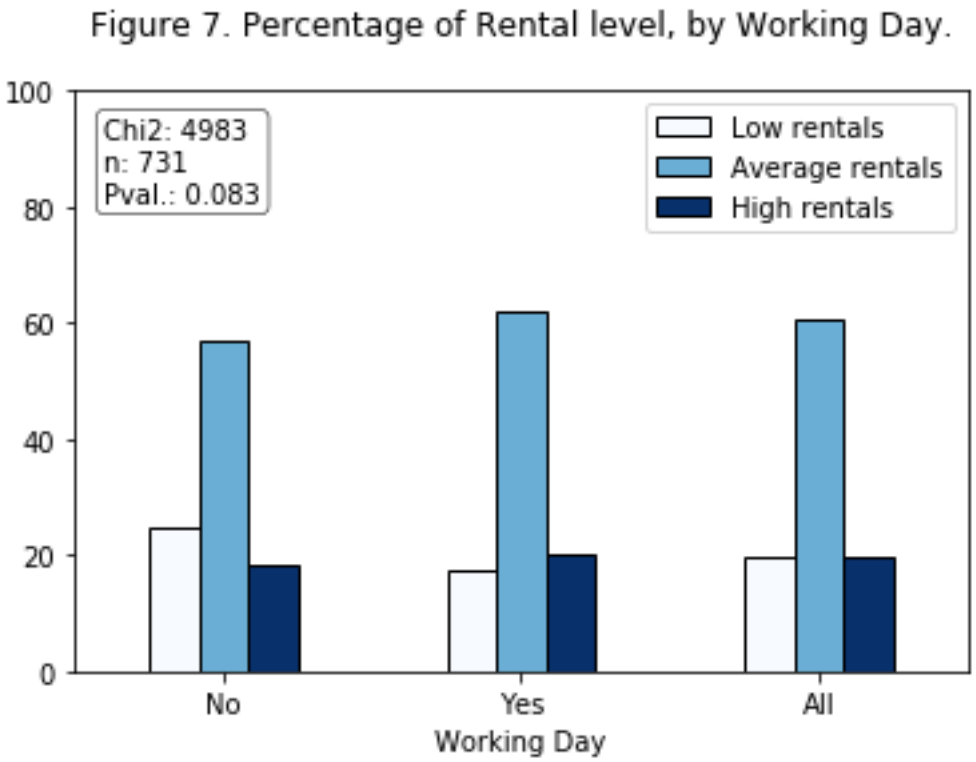
Answer: In the **sample** of 731 days, yes. In non working days it seems that we rent more bikes. BUT,…. what about in the **population**? Answer: Still don't know.

alberto.sanz@bigwaveanalytics.es

Percentage comparison (*n* groups)

EDEM
Centro Universitario

Graphic representation: Combined barplot



alberto.sanz@bigwaveanalytics.es

Percentage comparison (*n* groups)

EDEM Centro Universitario

2. Describe the DV, by the factor levels in the IV

(Cross tabulation of DV by IV)

```
pd.crosstab(wbr.cnt_cat, wbr.wd_cat, normalize='columns', margins=True)*100
```

	Non working days	Working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Percentage comparison (*n* groups)

EDEM Centro Universitario

2. Describe the DV, by the factor levels in the IV

(Cross tabulation of DV by IV)

Use DV as reference

	Non working days	Working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Percentage comparison (*n* groups)

EDEM Centro Universitario

2. Describe the DV, by the factor levels in the IV

(Cross tabulation of DV by IV)

Equal or different?

Use DV as reference

	Non working days	working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Percentage comparison (*n* groups)

EDEM Centro Universitario

2. Describe the DV, by the factor levels in the IV

(Cross tabulation of DV by IV)

Equal or different?

	Non working days	working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Answer: In the **sample** of 731 days, yes. In non working days it seems that we rent more bikes. BUT,…. what about in the **population**? Answer: Still don't know.

Percentage comparison (*n* groups)

EDEM Centro Universitario

3. Perform the numeric test for inference: Chi^2 test

	Non working days	Working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

```
# We apply the stats.chi2_contingency()
over the original
crosstab containing  FREQUENCIES

ct= pd.crosstab(wbr.cnt_cat, wbr.wd_cat)
stats.chi2_contingency(ct)

Output:
(4.9833225686178624,
 0.082772343895498146,
  2,
```

This is the P. Value

CONCLUSION:

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

In other words:

Percentage days with high/mid/low rentals do not significantly differ in Working days and Non working days.

alberto.sanz@bigwaveanalytics.es

Percentage comparison (*n* groups)

EDEM Centro Universitario

Table 2. Rental levels in Washington, by type of day. (In percentage points)

	Non working days	Working days	All days
Low rentals	24,7	17,6	19,9
Average rentals	57,1	62,3	60,7
High rentals	18,2	20,0	19,5
TOTAL	100	100	100
(n)=731			

Chi^2=4.983, p-value = 0.083. *Source:* Own analyses over Fanaee, Hadi and Gama (2013) data.

Conclussion: As P. Value > 0.05 Do not reject H0.

- ✓ H0.: Percentage of days with *low/average/high* rentals is the same in working days vs. not working days.
- ✗ H1.: Percentage of days with *low/average/high* rentals differes in working days vs. not working days.

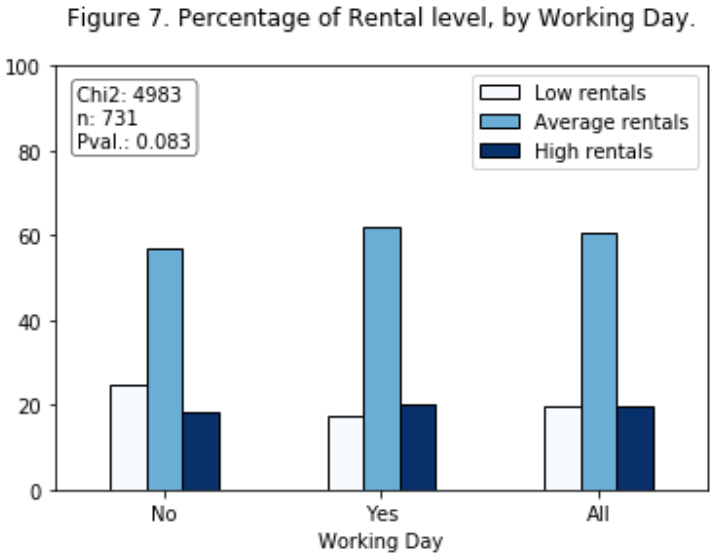
alberto.sanz@bigwaveanalytics.es

Percentage comparison (*n* groups)

EDEM Centro Universitario

4. Graphic representation: Combined barplot

```
#Transpose and plot
my_ct2=my_ct.transpose()

my_ct2.plot(kind="bar",  edgecolor = "black", colormap='Blues')
props = dict(boxstyle='round', facecolor='white', lw=0.5)
plt.text(-0.4, 81, 'Chi2: 4983''\n''n: 731' '\n' 'Pval.: 0.083',    bbox=props)
plt.xlabel('Working Day')
plt.title('Figure 7. Percentage of Rental level, by Working Day.''\n')
plt.legend(['Low rentals','Average rentals','High rentals'])
plt.ylim(0,100)
plt.xticks(rotation='horizontal')
```



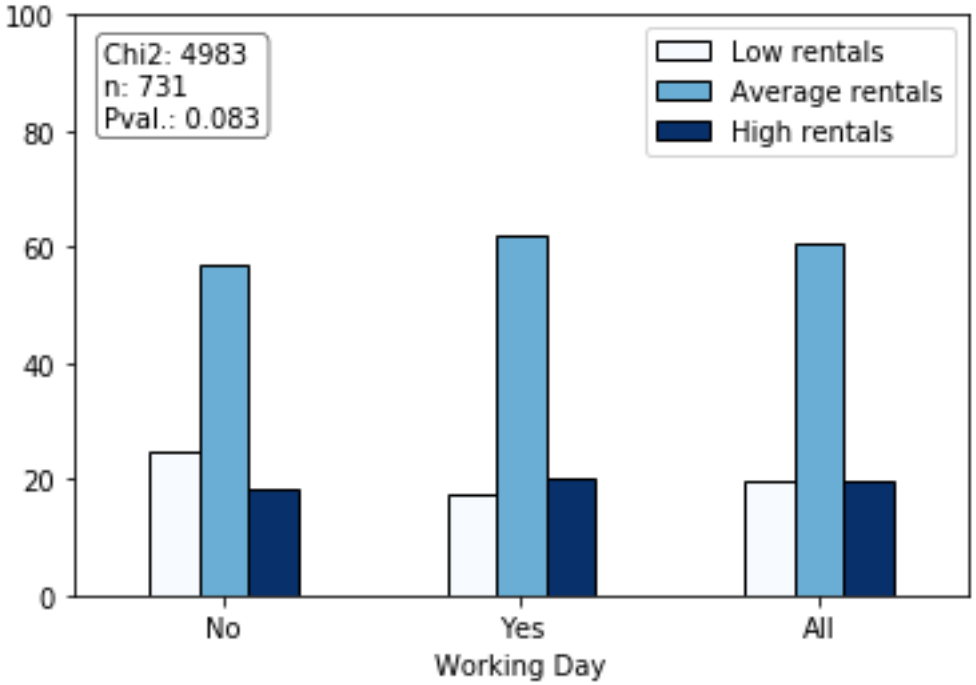
alberto.sanz@bigwaveanalytics.es

Percentage comparison (*n* groups)

EDEM Centro Universitario

4. Graphic representation: Combined barplot

Figure 7. Percentage of Rental level, by Working Day.



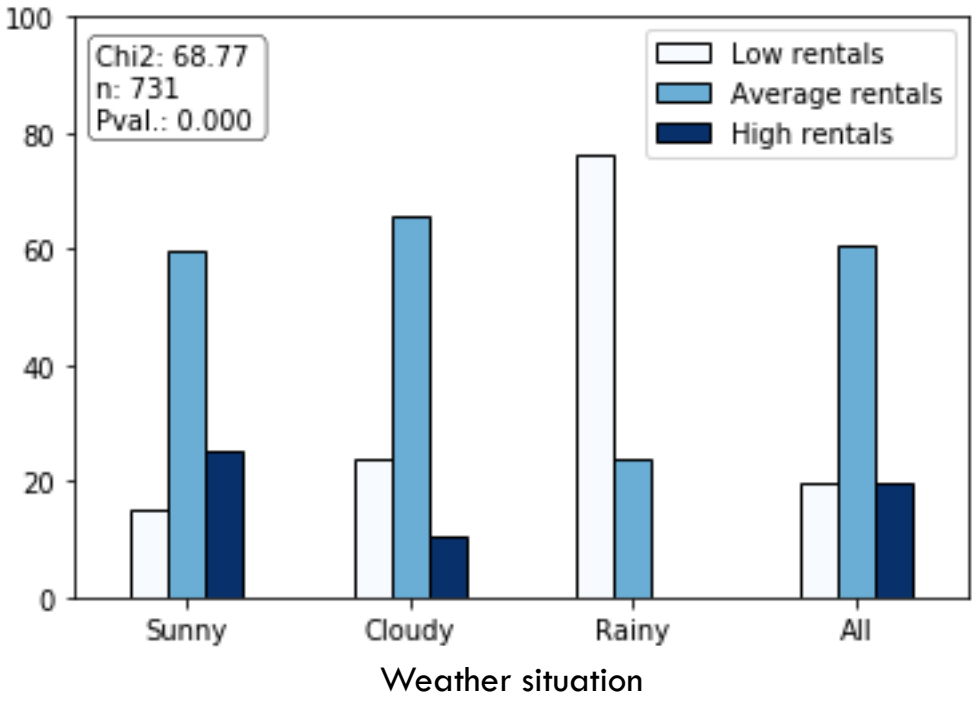
alberto.sanz@bigwaveanalytics.es

Percentage comparison (*n* groups)

EDEM
Centro Universitario

4. Graphic representation: Combined barplot (ex. II)

Figure 8. Percentage of Rental level,
by Weather situation



alberto.sanz@bigwaveanalytics.es

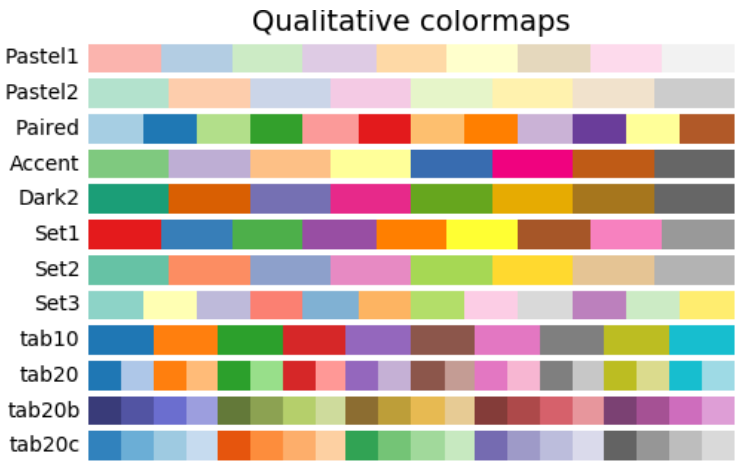
Tricks of the Trade: Color in Python

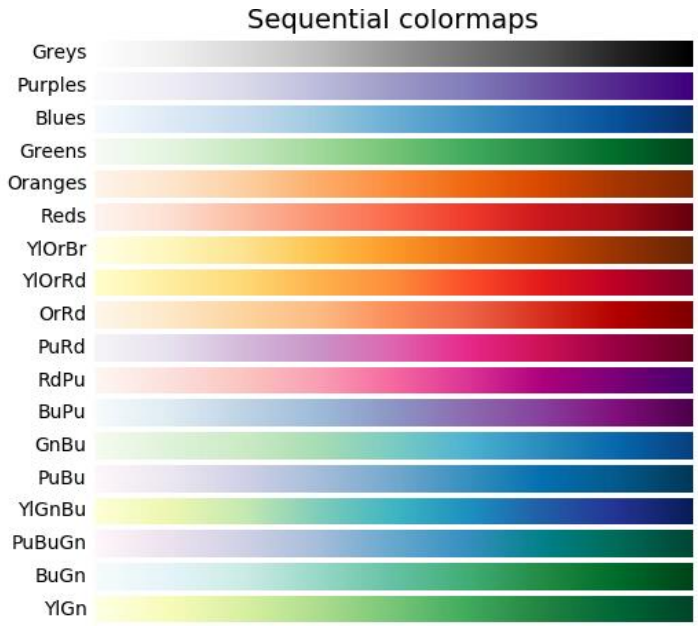
EDEM Centro Universitario

You may want to explore:

- The matplotlib Colormaps

  https://matplotlib.org/2.0.2/examples/color/colormaps_reference.html





alberto.sanz@bigwaveanalytics.es

Percentage Comparison Summing UP

EDEM Centro Universitario

- General Remainder:
  - Allways **describe/explore your data** (numerically + graphically) prior to perform any statistical analysis.
- Main Numeric Procedure:
  - Crosstabulation with Column percentages
  - Chi2: test
- Main Graphic Procedure:
  - Combined Barplot

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

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