EDEM
Centro Universitario

# Programación Estadística con Python

**Sesiones 4 y 5**
**Subsetting data & avoiding artifacts**

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

**MASTER EN DATA ANALYTICS PARA LA EMPRESA**

# Subsetting (I) Selecting cases

EDEM
Centro Universitario



```
# Select a subsample from our data

# Select cases only from 2011
# Create a new dataframe  containing observations from 2011

#Explore years
mytable = wbr.groupby(['yr']).size()
print(mytable)

#Excursus to Operators

# Subset year 0
wbr_2011 = wbr[wbr.yr == 0]

# Subset year 1
wbr_2012 = wbr[wbr.yr == 1]
```

# *Excursus*: Basic operators in Python

EDEM Centro Universitario



## Logic Operators

| Operator | Description |
|---|---|
| < | less than |
| <= | less than or equal to |
| > | greater than |
| >= | greater than or equal to |
| == | exactly equal to |
| != | not equal to |
| not x | Not x |
| x \| y | x OR y |
| x & y | x AND y |

## Arithmetic Operators

| Operator | Description |
|---|---|
| + | addition |
| - | subtraction |
| * | multiplication |
| / | division |
| ** | exponentiation |
| x % y | modulus (x mod y) |
| x //y | integer division |

# Subsetting (II) Selecting variables

EDEM Centro Universitario



```
# Select variables, by column name
#Define a list with the subset of variables I want to extract
#e.g. create a dataframe with the number of rentals (cnt) and the
temperatura only

my_vars=['temp_celsius','cnt']

#Extract those varibles and save them into wbr_minimal

wbr_minimal= wbr[my_vars]
wbr_minimal.shape
# QC OK
```

# Subsetting (I) Selecting cases

EDEM Centro Universitario



```
# Select a subsample from our data

# Select cases only from 2011
# Create a new dataframe  containing observations from 2011

#Explore years
mytable = wbr.groupby(['yr']).size()
print(mytable)

#Excursus to Operators

# Subset year 0
wbr_2011 = wbr[wbr.yr == 0]

# Subset year 1
wbr_2012 = wbr[wbr.yr == 1]
```

# Exercise # 1a

EDEM
Centro Universitario

- Make a histogram of the Bike rentals in Washington on the Winter of 2012
  - 1 Subset
  - 2 Describe
    - Graphically
    - Numerically

# Exercise # 1b

EDEM
Centro Universitario

- Make a histogram of the Bike rentals in Washington during the Winter AND the Fall
  - 1 Subset
  - 2 Describe
    - Graphically
    - Numerically

# Export data from Python

EDEM Centro Universitario



```
#### Export data
# to CSV

wbr.to_csv('wbr_edem2019.csv', sep=';', decimal=',')

## CAUTION ## The parameters for sep and decimal will depend very much of
the language of your Operative System. A typical alternative to the
example avobe would be sep = ",", dec = "."
```

```
#######################################################################
## Additional topic
##You can export directly to Excel, if desired

# Save dataframe to Excel
wbr.to_excel("wbr.xlsx")

#######################################################################
```

# BREAK

# Exercise #2

EDEM
Centro Universitario



- Exercise #2 (wbr_ue.csv)

Compute the average temperuture and the standard deviation in Washington

# Avoiding artefacts (I): Detecting non valid codes

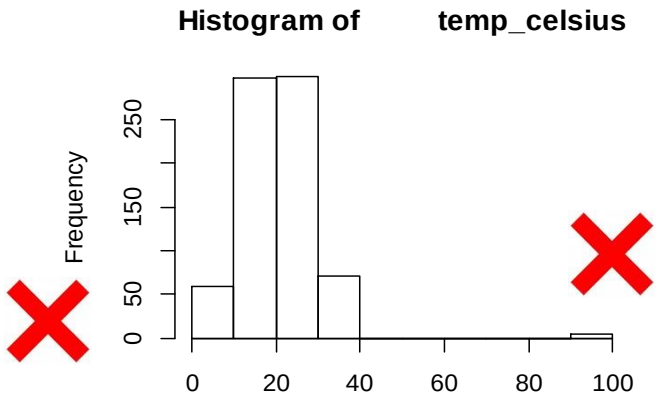
EDEM Centro Universitario



```
## ALWAYS PLOT YOUR DATA

plt.hist(wbr_ue.temp_celsius)
wbr_ue.temp_celsius.describe()[1]
#plt.boxplot(wbr_ue.temp_celsius)
```

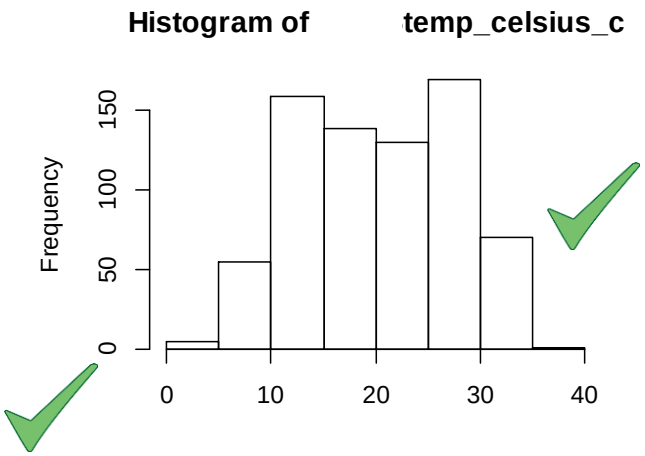
Mean: 20.9




```
# Let's clean the temp_Celsius variable
wbr_ue['temp_celsius_c']=wbr_ue.temp_celsius.replace(99,np.nan)
wbr_ue.temp_celsius_c.describe()[1]

# Hist or boxplot will not work if a pandas series
#contain nan
# We need to drop the nan before plotting
plt.hist(wbr_ue.temp_celsius_c.dropna())
```

Mean: 20.3




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

# Avoiding artefacts (II):
## Removing cases that have nan in any variable

EDEM Centro Universitario



```
# Remove cases with nan in any variable

# Create a new dataframe where the observations containing nan in
any of the variables are removed

wbr_ue2 = wbr_ue.dropna()

print(wbr_ue.shape)    (732,18)
print(wbr_ue2.shape)   (724,18)
```

# Summing UP (I)

EDEM
Centro Universitario

- Introduced Python & Spyder environment.
- Introduced some popular Python libraries:
  - Os
  - Pandas
  - Numpy
  - Matplot lib
- Introduced the notion of metrics:
  - Nominal, ordinal & Quantitative variables
- Description
  - Nominal variables:
    - Percentages & Bar plots
  - Quantitative variables:
    - Mean, standard deviation & Histograms

# But over all….

EDEM
Centro Universitario

- Allways plot you data!!!!!

# Summing UP (II)

EDEM
Centro Universitario

- Detecting non valid values
- Replacing non valid values by nan
- Removing cases with nan

EDEM
Centro Universitario

# Questions?

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

# Thank you !

Alberto Sanz

asanz@edem.es