EDEM
Centro Universitario

Programación Estadística con Python

Session 6
Data Transformation

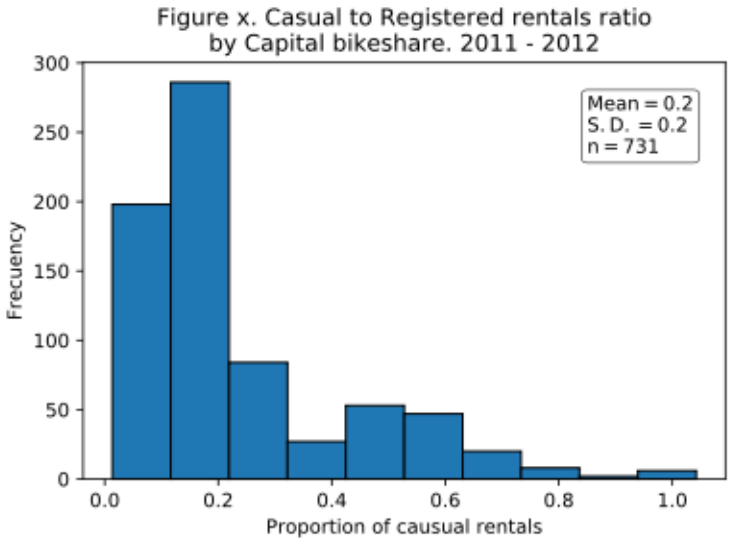
Alberto Sanz, Ph.D
asanz@edem.es

MASTER EN DATA ANALYTICS PARA LA EMPRESA

Computing new variables
(adding new columns to dataframe)

EDEM
Centro Universitario

```
##### Computing new columns
# let's compute the casual to registered rentals rati

wbr['cs_ratio']=(wbr.casual)/(wbr.registered)
wbr.cs_ratio.describe()


#Note that for creation of new columns we use "robust" column specification
with ["",] not attribute (.)call
```



Recoding

EDEM
Centro Universitario

- Recoding as a conditional transformation

Recoding I

EDEM
Centro Universitario

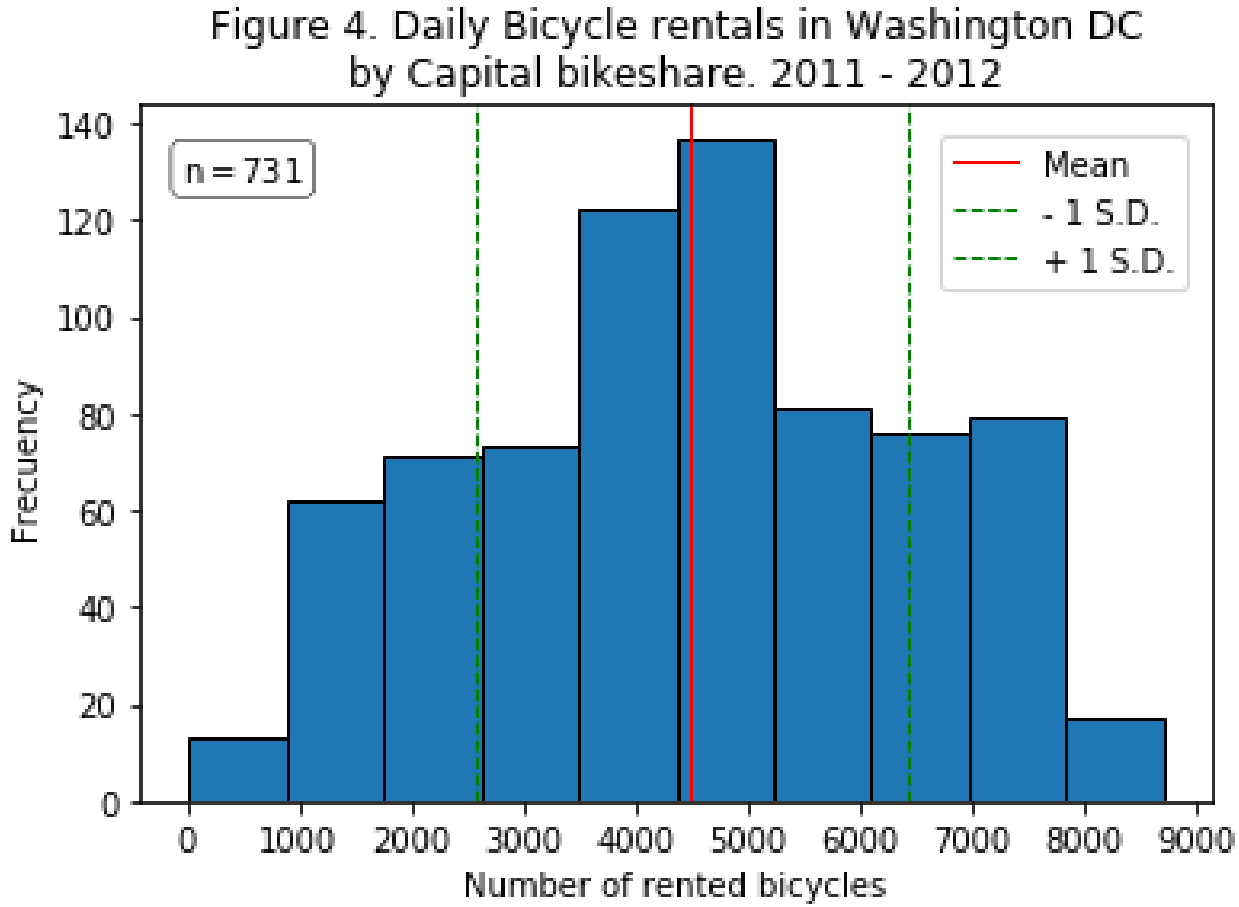
```
# Recoding season into a string variable (season_cat)
wbr.loc[(wbr['season']==1),"season_cat"]= "Winter"
wbr.loc[(wbr['season']==2),"season_cat"]= "Spring"
wbr.loc[(wbr['season']==3),"season_cat"]= "Summer"
wbr.loc[ wbr['season']==4),"season_cat"]= "Autum"

# Quality control
pd.crosstab(wbr.season, wbr.season_cat)
```

season_cat	Autum	Spring	Summer	Winter
season				
1	0	0	0	181
2	0	184	0	0
3	0	0	188	0
4	178	0	0	0

QC OK

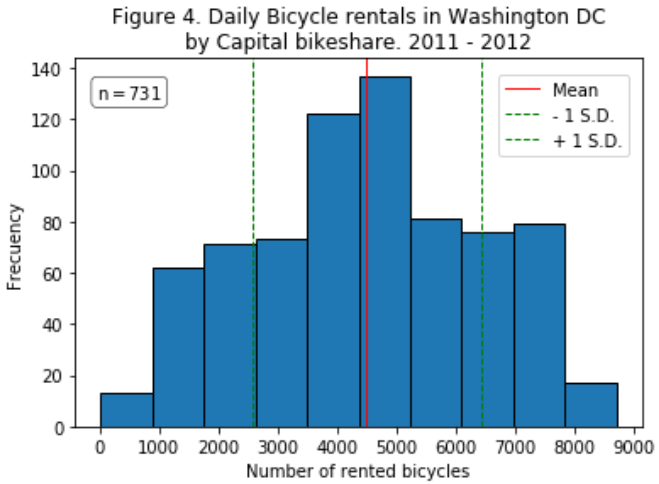
Recoding (II)

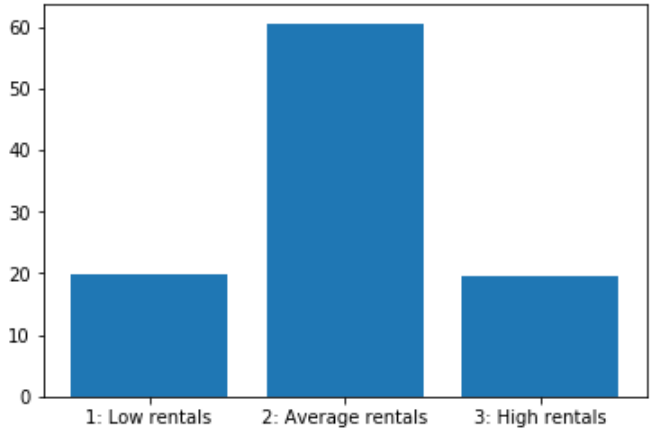
EDEM
Centro Universitario



Recoding (II)

EDEM
Centro Universitario

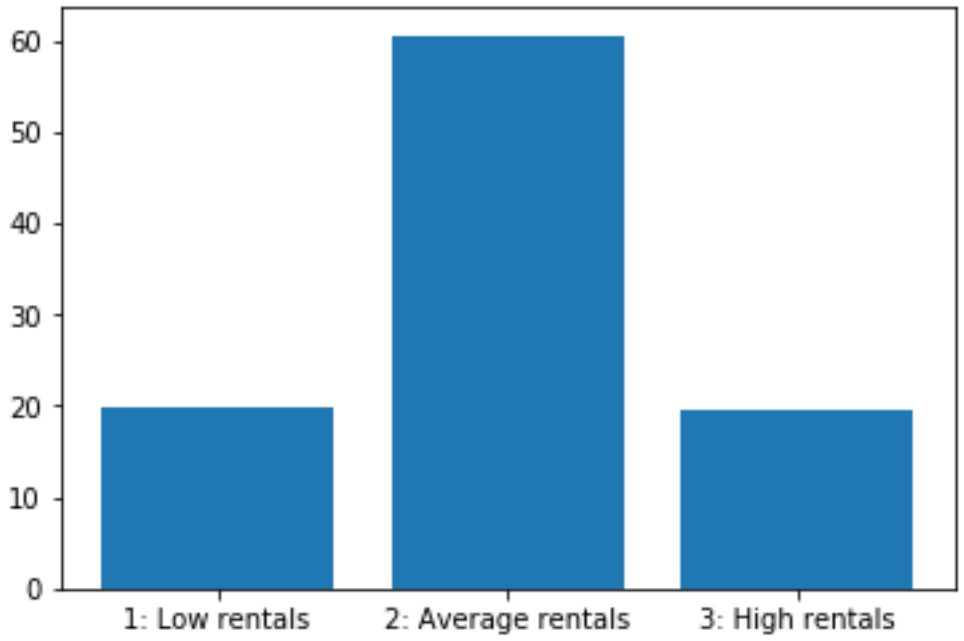




Recoding (II)

EDEM
Centro Universitario

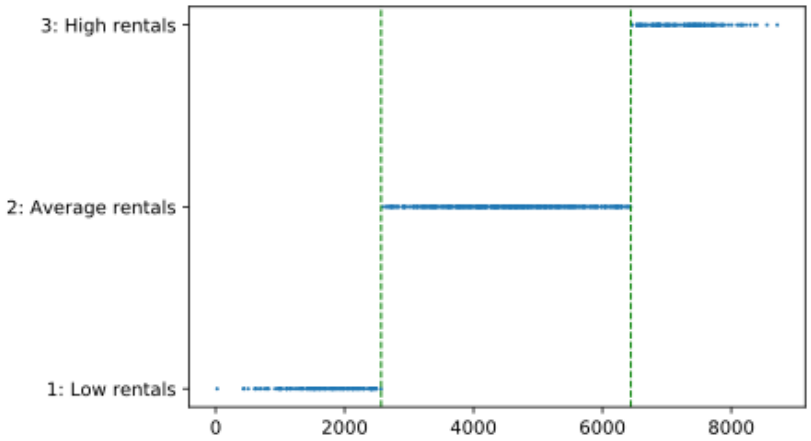

```
# Recode the number of rentals in Three Groups

### Recode 1
wbr.loc[ (wbr['cnt']<2567.1) ,"cnt_cat2"]= "1: Low rentals"
wbr.loc[ ((wbr['cnt']>2567.1) & (wbr['cnt']<6441.6)) ,"cnt_cat2"]= "2: Average rentals"
wbr.loc[ (wbr['cnt']>6441.6) ,"cnt_cat2"]= "3: High rentals"
```



Recoding (II)

EDEM
Centro Universitario


```
# Recode the number of rentals in three Groups

### Recode 1
wbr.loc[ (wbr['cnt']<2567.1) ,"cnt_cat2"]= "1: Low rentals"
wbr.loc[ ((wbr['cnt']>2567.1) & (wbr['cnt']<6441.6)) ,"cnt_cat2"]= "2: Average
rentals"
wbr.loc[ (wbr['cnt']>6441.6) ,"cnt_cat2"]= "3: High rentals"

##### Quality control?
plt.scatter(wbr.cnt, wbr.cnt_cat2, s=1)
```



QC OK

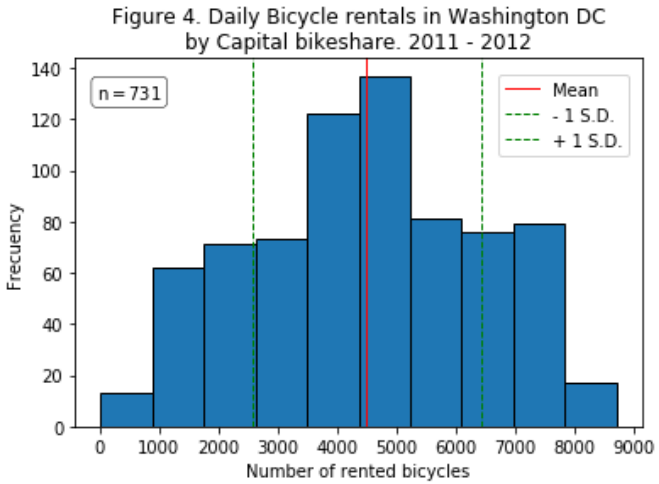
Recoding (II)

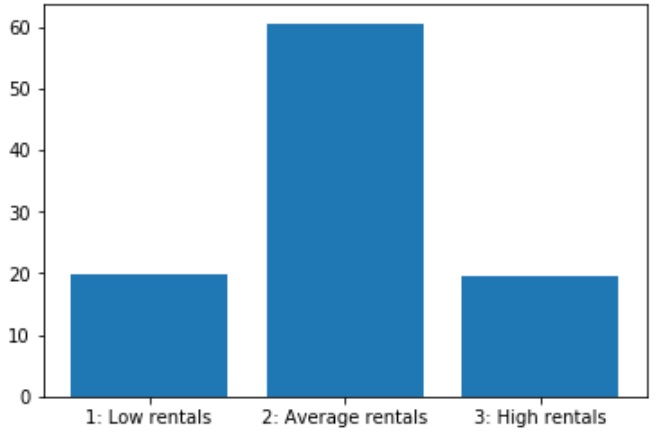
EDEM
Centro Universitario


```
# Recode the number of rentals in Three Groups

#Compute & store the cutting points

res = wbr['cnt'].describe()

# Store parameters as numbers
m  = res[1]
sd = res[2]
n  = res[0]


### Recode 2
wbr.loc[  (wbr['cnt']<(m-sd)) ,"cnt_cat2"]= "1 Low rentals"
wbr.loc[ ((wbr['cnt']>(m-sd)) & (wbr['cnt']<(m+sd))) ,"cnt_cat2"]= "2 Average rentals"
wbr.loc[  (wbr['cnt']>(m+sd)) ,"cnt_cat2"]= "3 High rentals"
```

Recoding

EDEM
Centro Universitario





Recoding into ordinal categories (Data preparation)

EDEM Centro Universitario


```
# Recode the number of rentals in Three Groups

#Compute & store the cutting points

res = wbr['cnt'].describe()

# Store parameters as numbers
m  = res[1]
sd = res[2]
n  = res[0]


### Recode 2
wbr.loc[  (wbr['cnt']<(m-sd)) ,"cnt_cat2"]= "Low rentals"
wbr.loc[ ((wbr['cnt']>(m-sd)) & (wbr['cnt']<(m+sd))) ,"cnt_cat2"]= "Average rentals"
wbr.loc[  (wbr['cnt']>(m+sd)) ,"cnt_cat2"]= "High rentals"
```

Note that now there are no numbers in the labels.

Excursus on PANDAS data types

EDEM
Centro Universitario

Pandas `dtype` mapping

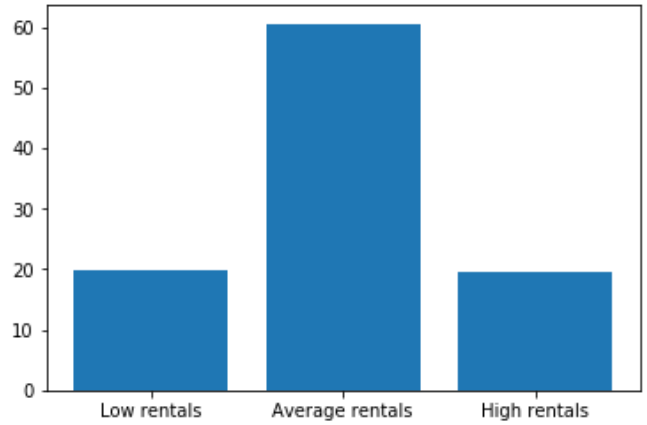
Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

Source: https://pbpython.com/pandas_dtypes.html

Recoding into ordinal categories (Method II)

EDEM
Centro Universitario


```
# Import specific functionality
from pandas.api.types import CategoricalDtype

# First define a specific categorical data type specific for us!!! (in two sub-steps)
# Step 1: declare the ordered categories

my_categories=["Low rentals", "Average rentals", "High rentals"]
#Step 2: Define new data type
my_rentals_type = CategoricalDtype(categories=my_categories, ordered=True)

# Second create a new categorical_ordered variable using our specific data type
wbr["cnt_cat5"] = wbr.cnt_cat2.astype(my_rentals_type)

#Then when you plot the variable or include it in further analyses, the categories will show up
# in your desired order
```



EDEM
Centro Universitario

Questions?

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

Thank you !

Alberto Sanz

asanz@edem.es