



Introducción al análisis de Datos

Programación Estadística con Python

Sesión 6

Data Transformation.

Intro to Google Colab

Alberto Sanz, Ph.D

alberto.sanz@bigwaveanalytics.es
www.linkedin.com/in/alberto-sanz-4b6bb5106

MASTER EN DATA ANALYTICS PARA LA EMPRESA

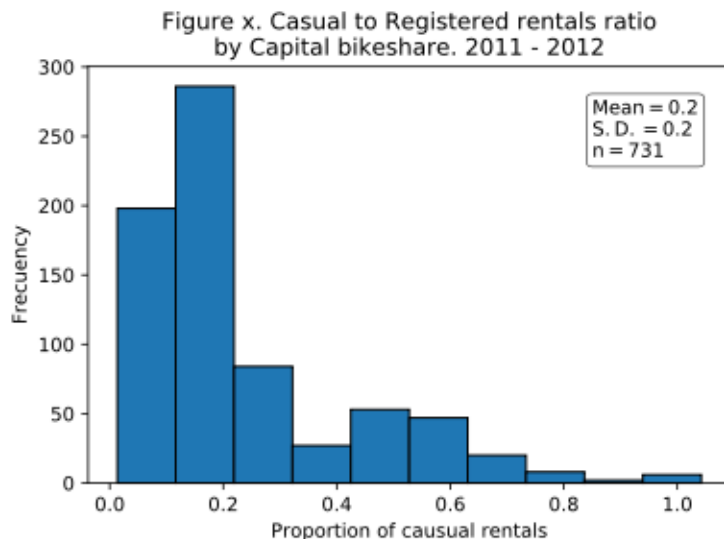
Computing new variables

(adding new columns to dataframe)

```
##### 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 as a conditional transformation

Recoding I

```
# 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)
```

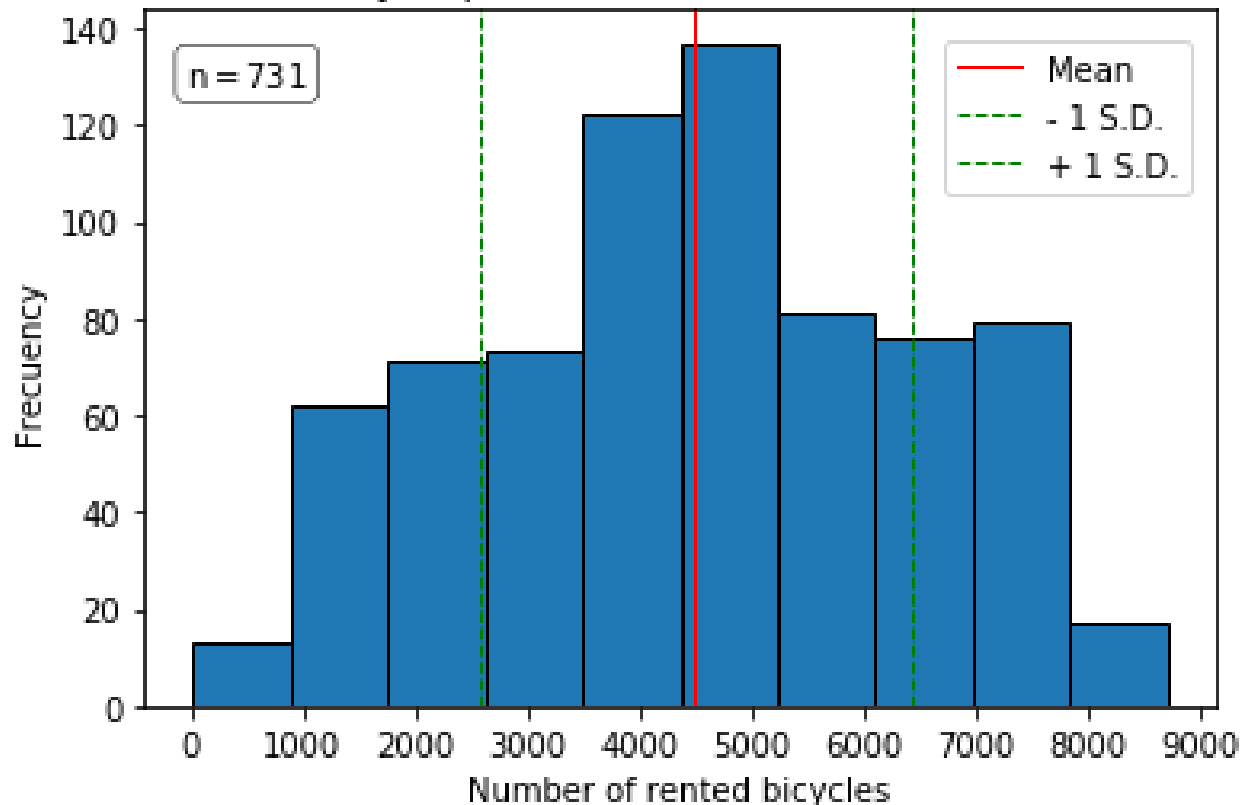
season_cat	Autum	Spring	Summer	Winter
season				
1	0	0	0	181
2	0	184	0	0
3	0	0	188	0
4	178	0	0	0



QC OK

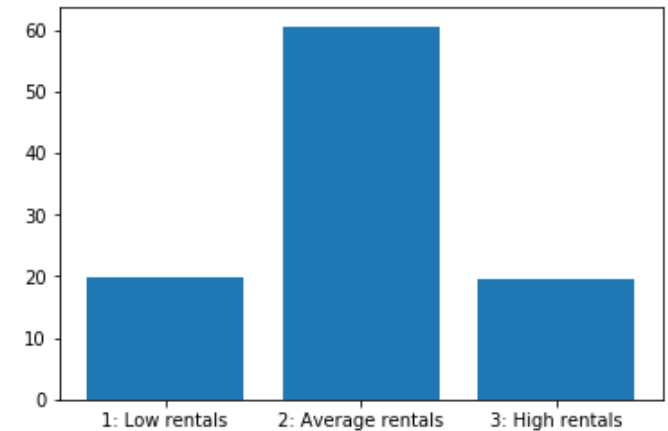
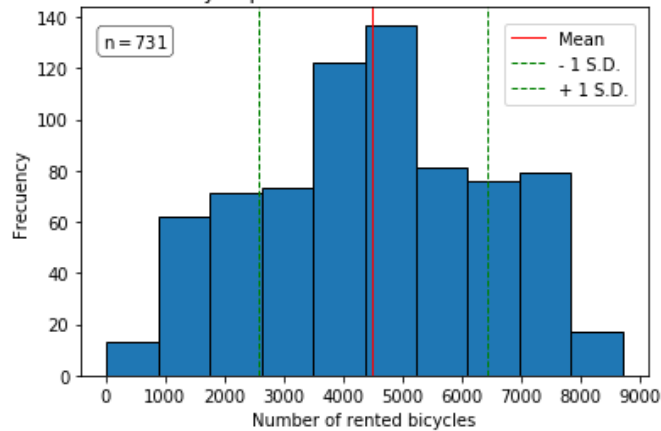
Recoding (II)

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



Recoding (II)

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



Recoding (II)

7

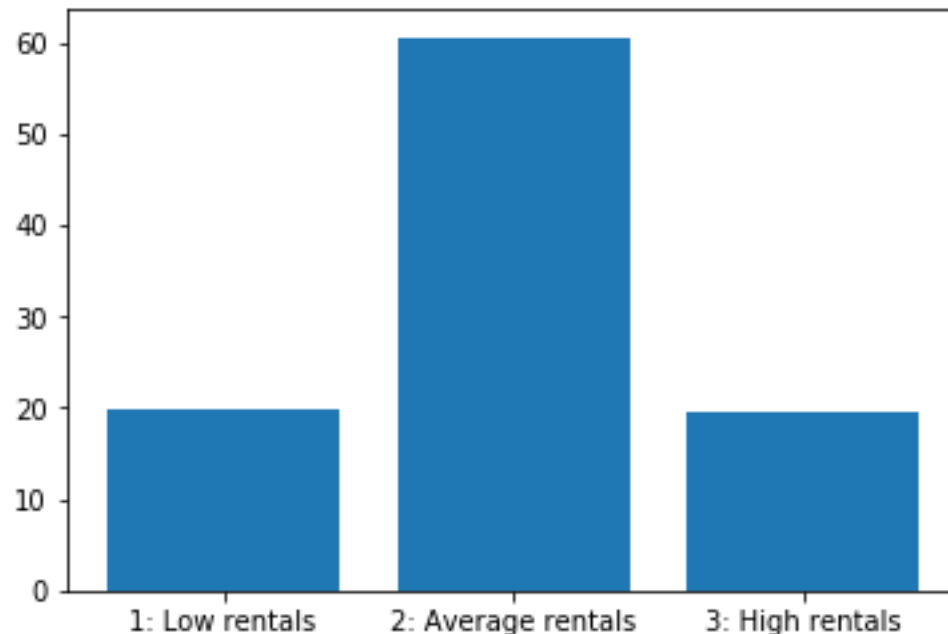
```
# Recode the number of rentals in Three Groups
```

```
### 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"
```



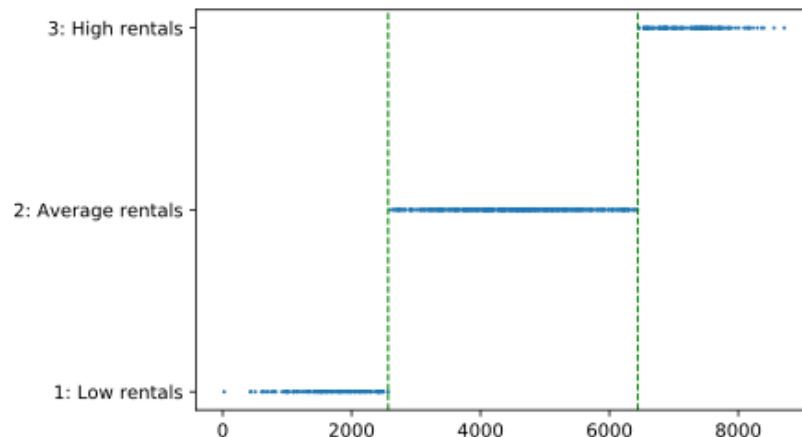
Recoding (II)

8

```
# 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)

9

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

#Compute & store the cutting points

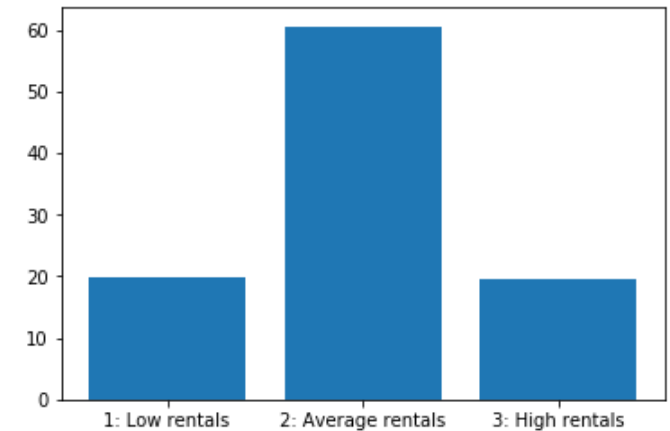
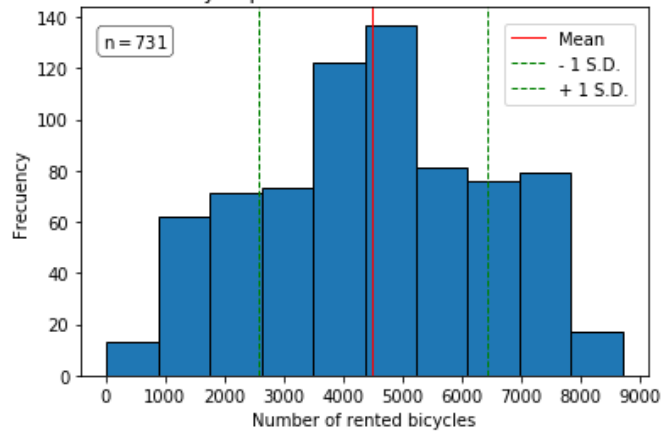
res = wbr['cnt'].describe()

# 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

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



Recoding into ordinal categories

(Data preparation)

11

```
# 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

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)

13

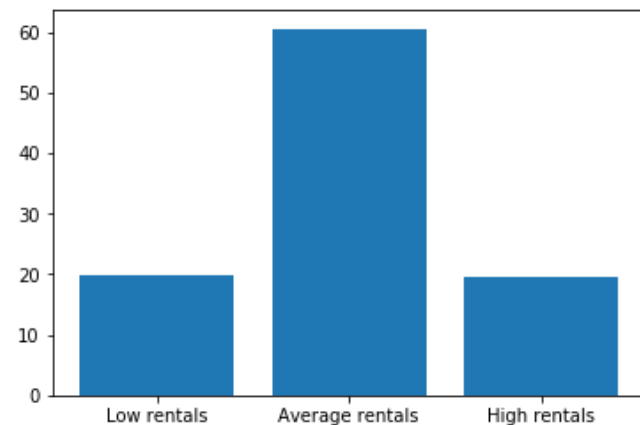
```
# Import specific functionality
from pandas.api.types import CategoricalDtype

# 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
```



Preparing the Colab/Python session

Preparing the Colab/Python session

15

```
#First we import and load all basic libraries
```

```
from google.colab import drive      # For linking colab to Google Drive
import pandas as pd                 # For dataframe handling
import numpy as np                  # For matrix and list computations
import matplotlib.pyplot as plt     # For advanced graphs
```

Loading data

```
drive.mount('mydrive')      #Bridge to Google Drive
```

```
wbr= pd.read_excel ('Path to file in Google drive') #Load data  
wbr          #Visualise dataframe
```

Note: In this example we named our dataframe with the name `wbr`
But you can give any desired and meaningful name to your dataframe

Questions?

Thank you !

Alberto Sanz

alberto.sanz@bigwaveanalytics.es

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