


DATA SCIENCE

PPGIa/PUCPR

Prof. Jean Paul Barddal



1

EXPLORATORY DATA ANALYSIS

2

Definition

- Task conducted when we find a dataset we know nothing or very little about
- Examples:
 - Dataset with the shots made by a basketball player
 - Dataset about wines (white/red)
- Can we extract any insights about these dataset?

3

How to?

- There is no recipe on how to conduct an exploratory data analysis
- It is much more about talent and resiliency rather than bits and bytes
- Yet, there are some tools and steps that can help us

4

UNIVARIATE DATA ANALYSIS

5

Univariate analysis

- The sum of
 - Descriptive analysis
 - Distribution plots
 - Thinking
- At this point, it is important for us to recall skewness (symmetry) and kurtosis

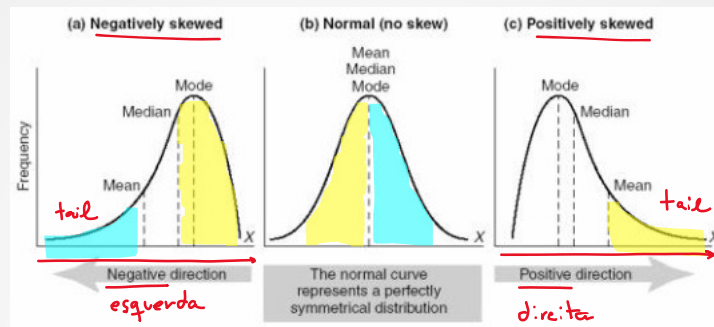
6

Asymmetry (Skewness)

- Evaluates a data distribution to a gaussian distribution
- When the mean, median, and mode are the same, then the asymmetry coefficient is zero
- When the mean is larger than the median and mode, we have positive asymmetry
- When the mean is smaller than the median and mode, we have negative asymmetry

7

Skewness



Left (negative) skew
 $\text{mean} < \text{median} < \text{mode}$

No skew (symmetric)
 $\text{mode} = \text{median} = \text{mean}$

Right (positive) skew
 $\text{mode} < \text{median} < \text{mean}$

Hint: think about the tail of the curve!

8

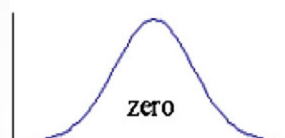
Kurtosis

- Measures how "tailedness" a distribution is
- A distribution with zero kurtosis is called mesokurtic
- A distribution with positive kurtosis is called leptokurtic
- A distribution with negative kurtosis is called platykurtic

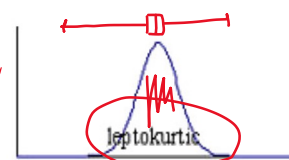
9

Skewness and Kurtosis

Also called a **right**-tailed distribution



Also called a **left**-tailed distribution



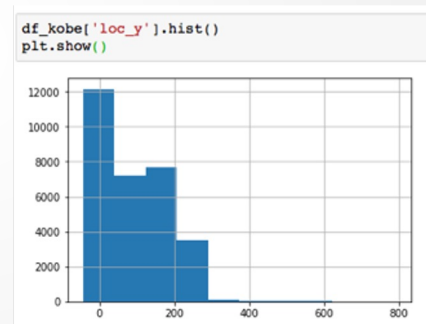
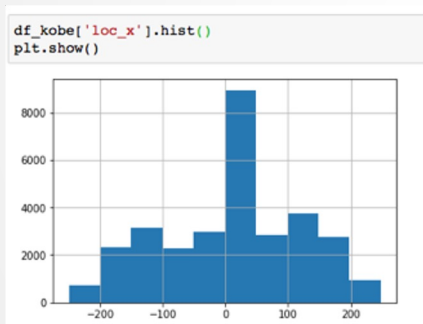
10

HISTOGRAM

11

Histogram

The easiest way to check the distribution of a variable is to plot a histogram



12

Questions

- From the plots in the previous slide, what kind of skewness and kurtosis we observe in **loc_x** and **loc_y**?
- Do you see any outliers in this data?
- How do we compute the skewness and the kurtosis from this data?
- Hint: `skew()` and `kurtosis()` from `scipy`

13

BOX-PLOT

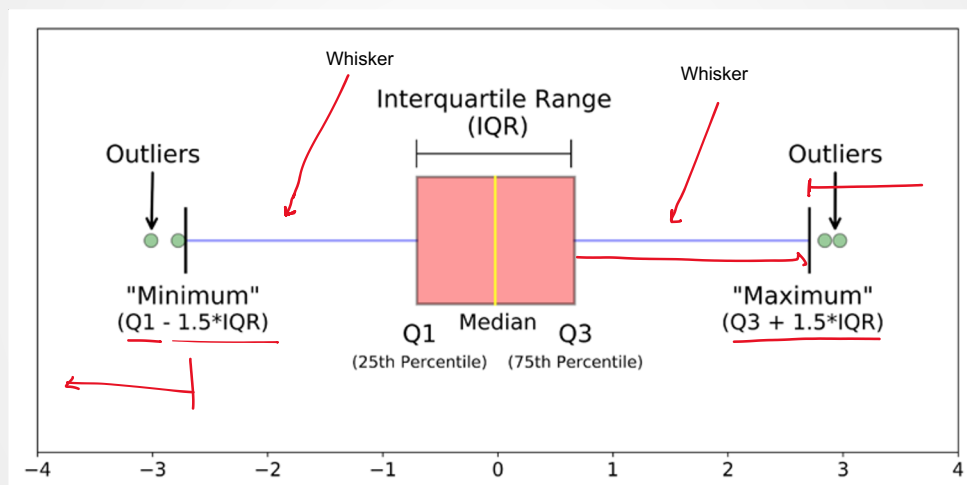
14

Box-plot

- Another handy way to check the distribution of a variable is to use box-plots
- Box-plots are a visual approach to visualize descriptive metrics from a data distribution
- **sns.boxplot()**

15

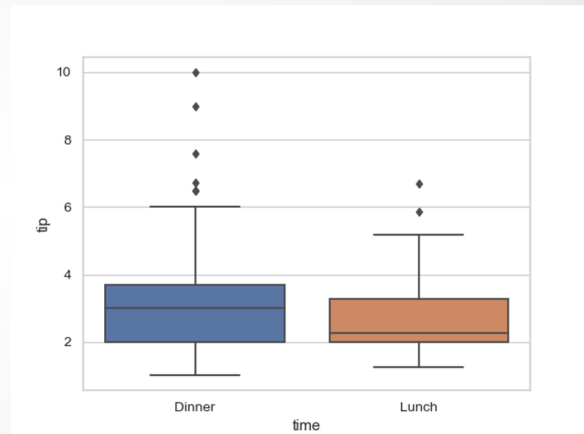
Box-plot



16

Box-plot

- Box-plots are specially useful when we need to analyze the behavior of a numeric variable with changes in a categorical variable
- Note that this plot is, in practice, a bivariate plot



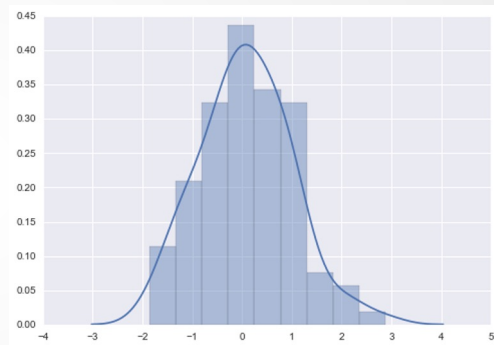
17

KDE Plots

18

Kernel Density Estimate

- A Kernel Density Estimate (KDE) allows us to estimate the distribution of a variable given its sample (the data we have)
- `sns.kdeplot()`



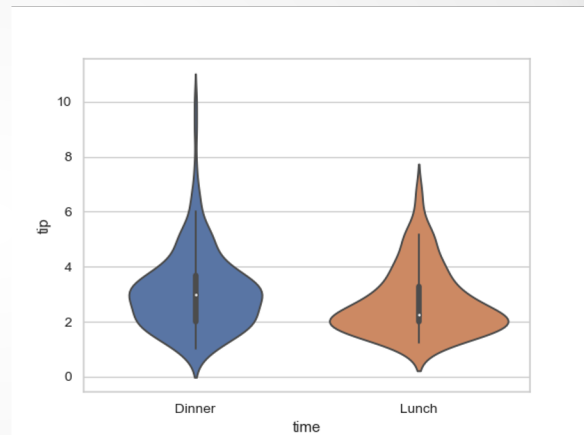
19

VIOLIN PLOT

20

Violin Plot

- A violin plot is quite similar to a box-plot, yet, instead of plotting a box, KDEs are plotted
- This gives us a better idea on how the data is distributed
- `sns.violinplot()`



Bear in mind that this example is a bivariate analysis!

21

CODE

22

Time to code

- Let's code the aforementioned topics using Python