DATA SCIENCE

PPGIa/PUCPR

Prof. Jean Paul Barddal

SCIENTIA, VITA ET FIDES

PUCPR

GRUPO MARISTA

EXPLORATORY DATA ANALYSIS

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?

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

UNIVARIATE DATA ANALYSIS

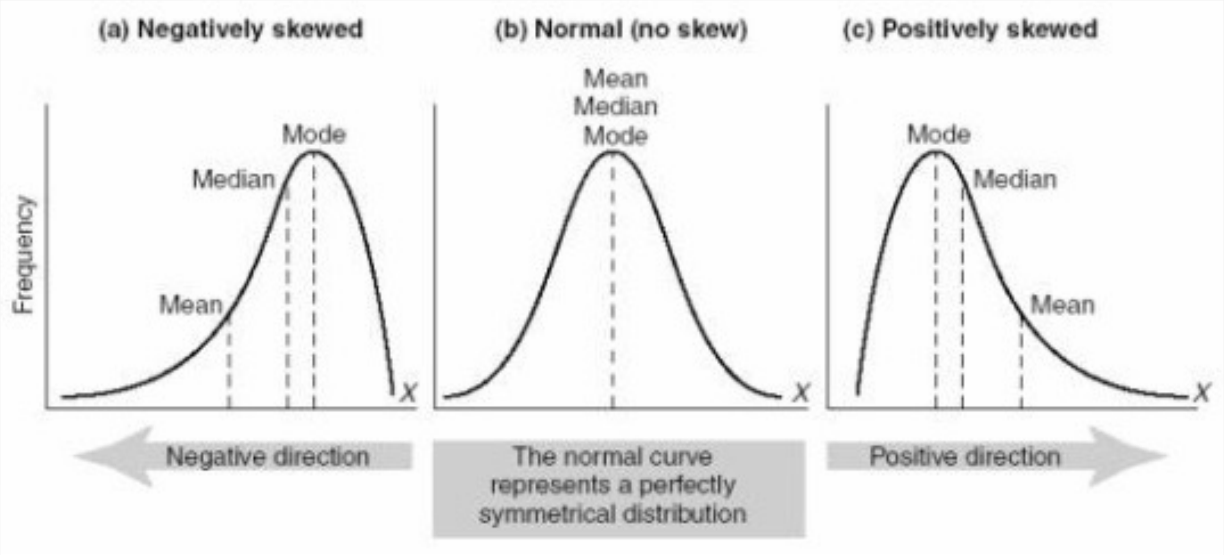
Univariate analysis

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

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

Skewness



Left (negative) skew
mean < median < mode

Right (positive) skew
mode < median < mean

No skew (symmetric)
mode = median = mean

Hint: think about the tail of the curve!

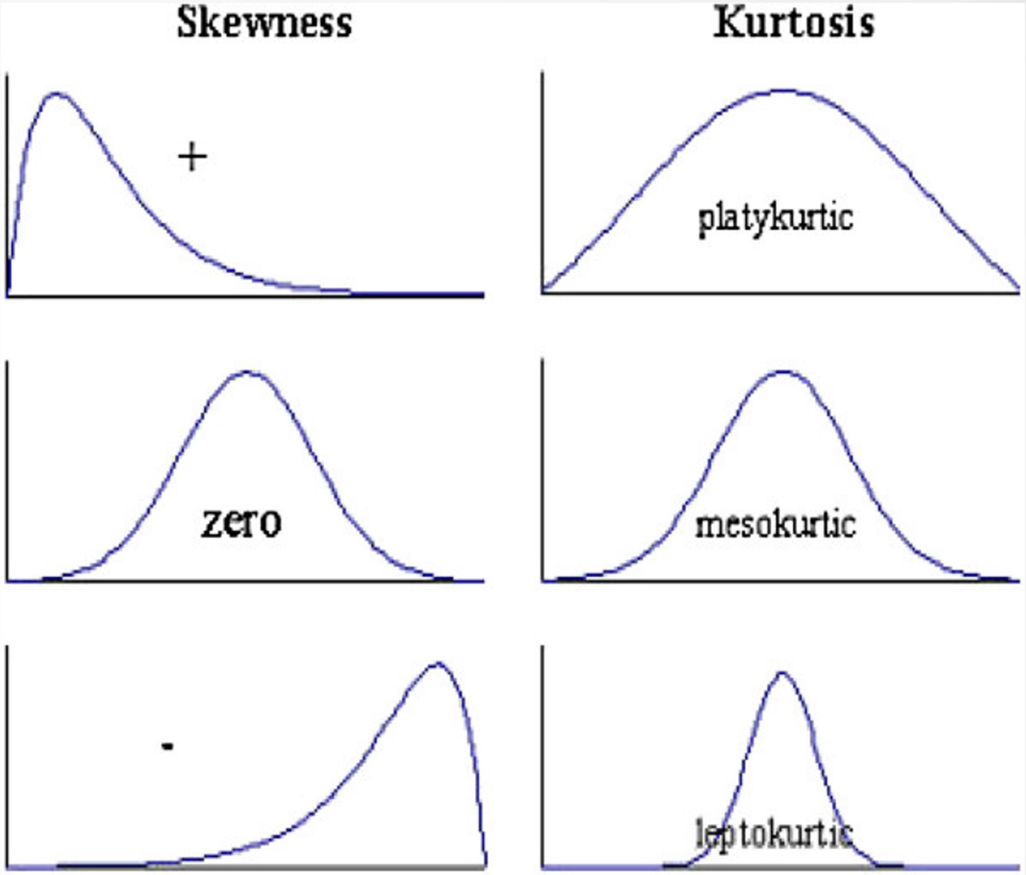
Kurtosis

- 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

Skewness and Kurtosis

Also called a **right**-tailed distribution

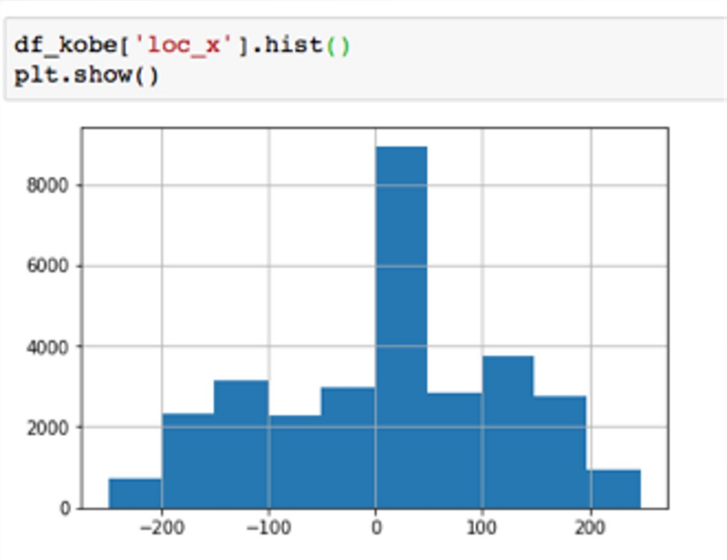
Also called a **left**-tailed distribution



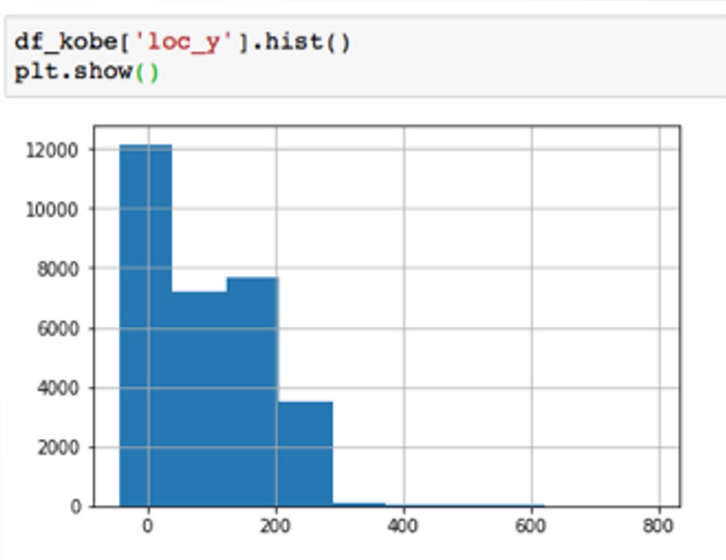
HISTOGRAM

Histogram

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



```
df_kobe['loc_x'].hist()
plt.show()
```



```
df_kobe['loc_y'].hist()
plt.show()
```

Questions

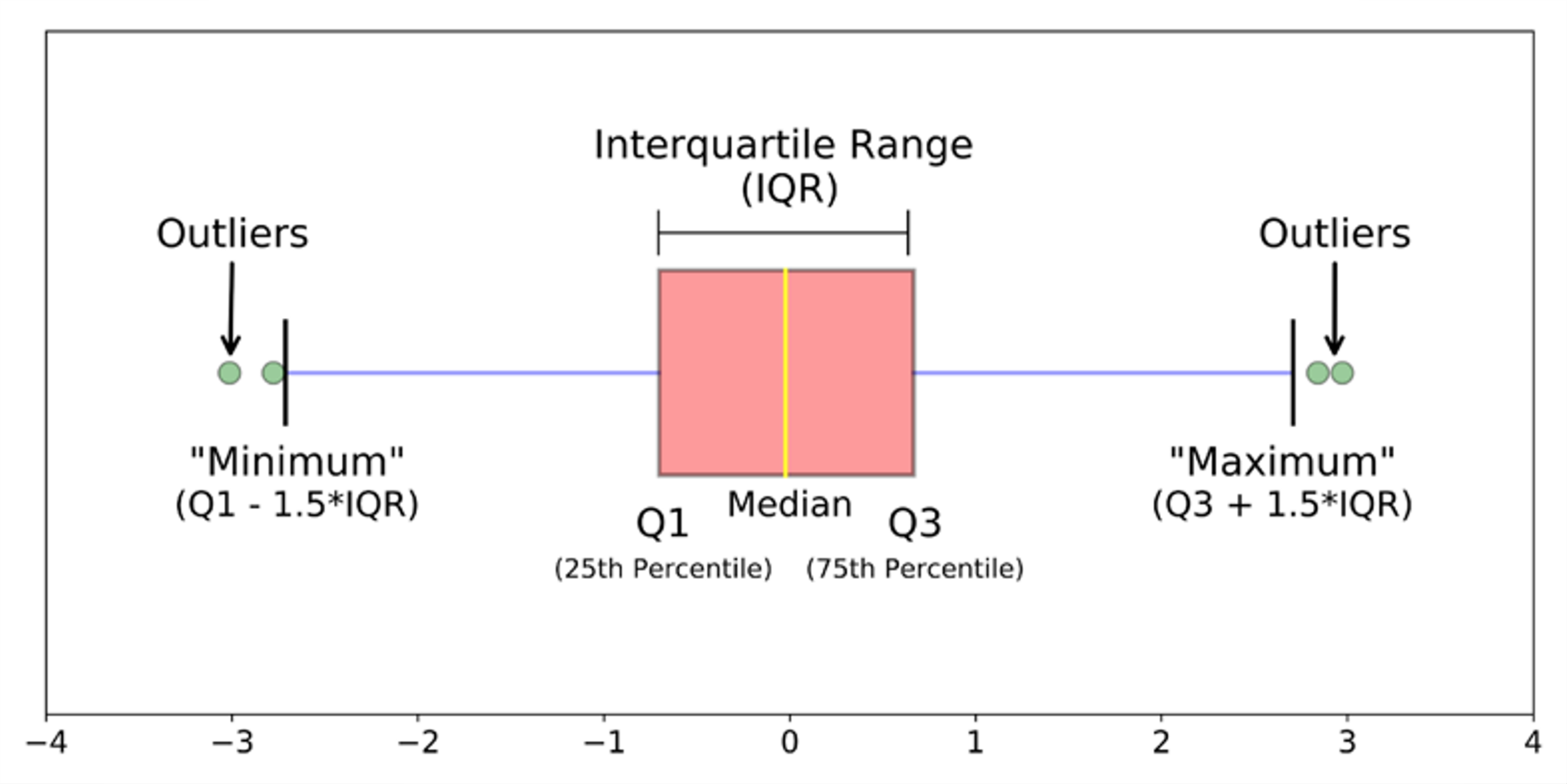
- From the plots in the previous slide, what kind of skewness and kurtosis we observe in **loc_x** and **loc_y**?
- Do you see any outliers in this data?
- How do we compute the skewness and the kurtosis from this data?
- Hint: skew() and kurtosis() from scipy

BOX-PLOT

Box-plot

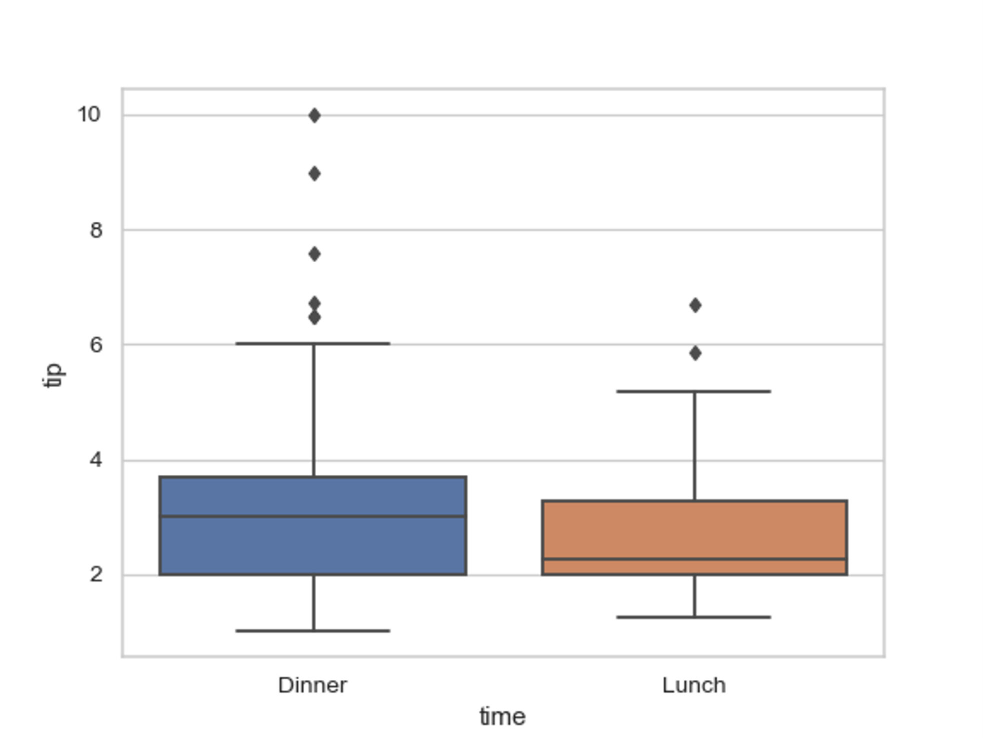
- Another handy way to check the distribution of a variable is to use box-plots
- Box-plots are a visual approach to visualize descriptive metrics from a data distribution
- **sns.boxplot()**

Box-plot


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Box-plot

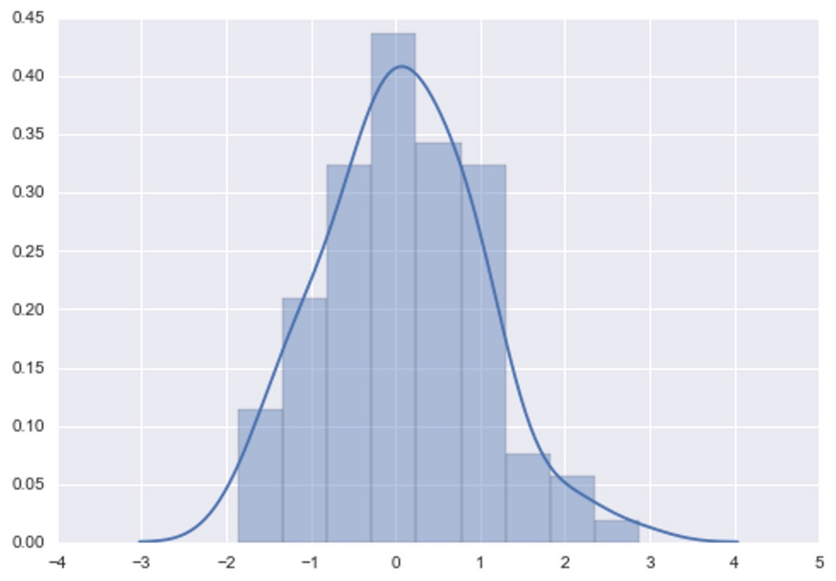
- Box-plots are specially useful when we need to analyze the behavior of a numeric variable with changes in a categorical variable
- Note that this plot is, in practice, a bivariate plot



KDE Plots

Kernel Density Estimate

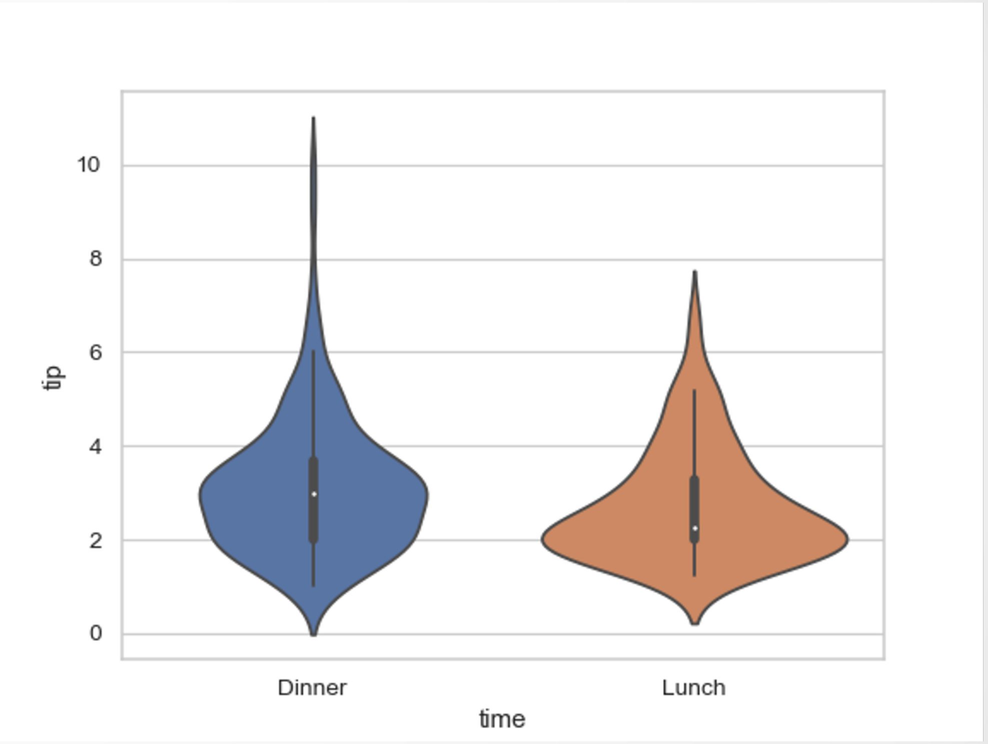
- A Kernel Density Estimate (KDE) allows us to estimate the distribution of a variable given its sample (the data we have)
- sns.kdeplot()



VIOLIN PLOT

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!

CODE

Time to code

- Let’s code the aforementioned topics using Python

WHY IS THIS IMPORTANT?

Data skewness, Data analysis, and ML

- Data skewness is relevant as it affects different types of data analysis, statistics, and machine learning tools
- For instance, decision trees are invariant to data skewness, but:
  - Correlation analysis may be incorrect if data is skewed
  - Neural networks tend to converge faster when data is not skewed
  - Clustering techniques are unlikely to cluster data correctly if data is skewed

Power Transformation

- A useful tool for converting non-gaussian data into a gaussian-like distribution is to perform a power transformation
- Scikit-learn has the PowerTransformer class
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html
- Let us work on an quick example