Advanced Data Analysis with Python

Cecilia Graiff

September 4, 2025

ALMAnaCH, Inria Paris - École d'Affaires Publiques, SciencesPo

Course Overview

Course Overview

- **Week 1:** Exploratory Data Analysis
- **Week 2:** Statistical Modeling Foundations
- **Week 3:** Linear Models
- **Week 4:** Causal Inference Methods
- **Week 5:** Time Series & Forecasting
- **Week 6:** Machine Learning

What is expected from you

Attend Lectures:

In case of **more than two** unjustified absences, your course cannot be validated. You are also invited to participate actively and bring all of your questions to the class :)

Complete the assignment:

At the end of the course, you will have to submit your project and a paper that documents it thoroughly. You will also have to submit all the used materials and the code.

Assignment

- Each project should be submitted by groups of 2-3 people;
- Project ideas are due by the end of October:
  - Individuate the dataset you want to work on
  - Individuate 2-3 research questions
  - Summarize the approach that you want to follow
- You will have to design a complete pipeline of data analysis. Do not worry: this will become more clear during the course!
- The project is due after the end of the course. More precise dates will be communicated.

Homework

- **Not mandatory!**
- Homework will be uploaded on GitHub before or after every class
- They will be corrected together at the beginning of each class
- **You are strongly encouraged to complete your homework, because it will help you a lot!**

Datasets

- Our first experiments will be performed on the Housing Prices Dataset, freely available on Kaggle.
- To complete the course, you will have to **submit your own project**. It is important that this project is based on **your interests and domain of specialization**.
- A more list of dataset will be presented in future lectures, after evaluation of your skills. If you want to use a different dataset, this is possible, but **you will need to justify it to me**.
- However, you can have a look at Datasets of the EU to individuate possible topics and research questions that are interesting to you.

Class delegate election!

Please raise your hands if you would like to be class delegate.

What is Data Analysis

Definition of Advanced Data Analysis

Definition

Advanced data analysis refers to a collection of techniques and tools that are used to analyze large volumes of data, uncover hidden patterns and provide actionable insights.

Adapted from IBM Research.

What Makes Data Analysis Advanced?

- **From descriptive to predictive:**
  - Not just "What happened?" but "What will happen?" and "What if we change something?"
  - E.g., predicting GDP growth, election outcomes, or patient survival.
- **Stronger coding & math foundations:**
  - Implement models from scratch (linear models, causal inference, ML algorithms).
  - Use linear algebra, probability, optimization.
- **Handling more and richer data:**
  - Complex data types: text (political speeches), spatial data (disease spread), networks.
- **Model evaluation & generalization:**
  - Cross-validation, out-of-sample testing, uncertainty quantification.

Adapted from IBM Research.

Example: Regression

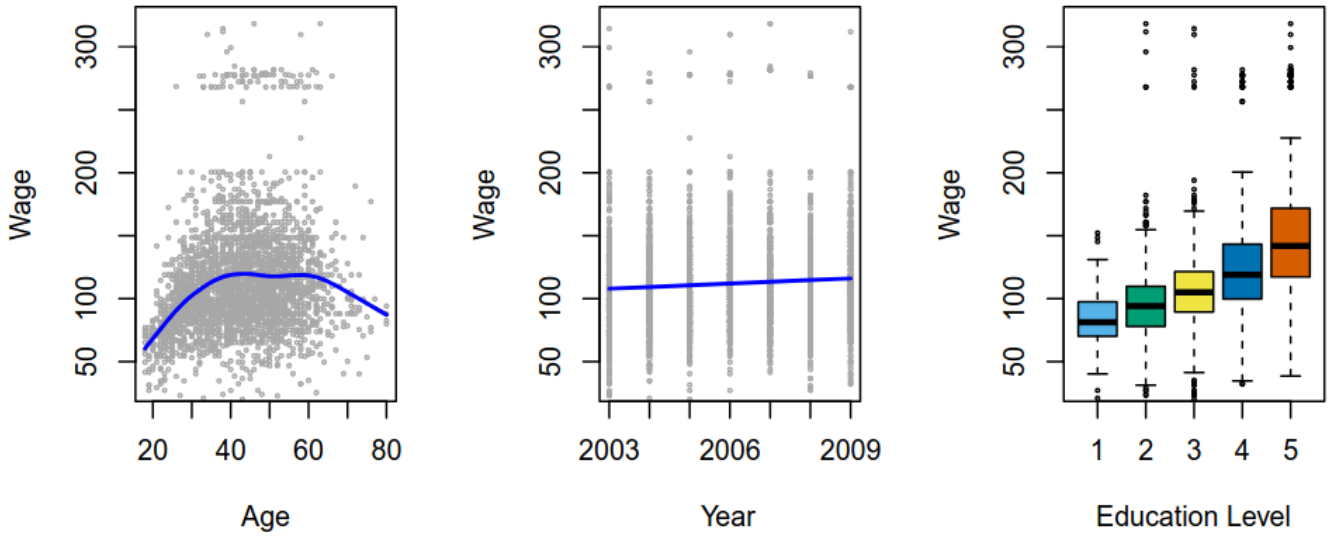


FIGURE 1.1. `Wage` *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009. Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Classification

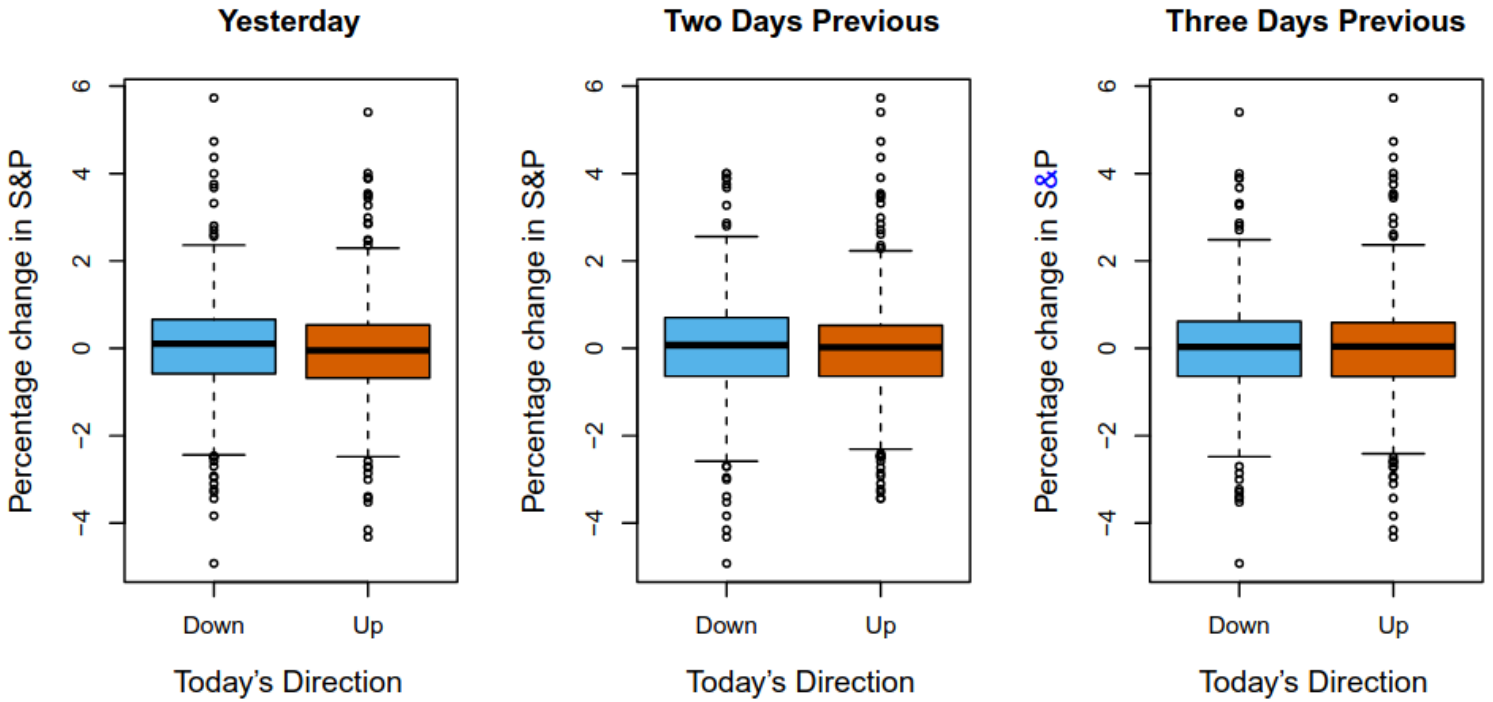


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Classification

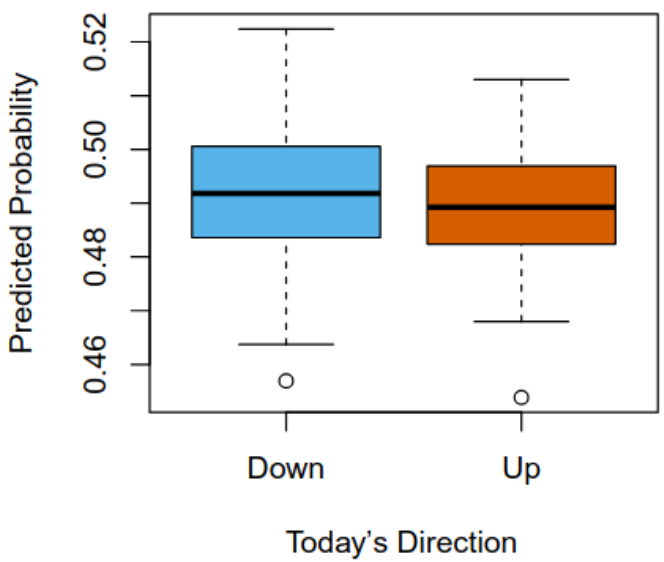


FIGURE 1.3. *We fit a quadratic discriminant analysis model to the subset of the* `Smarket` *data corresponding to the 2001–2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market 60% of the time.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Clustering

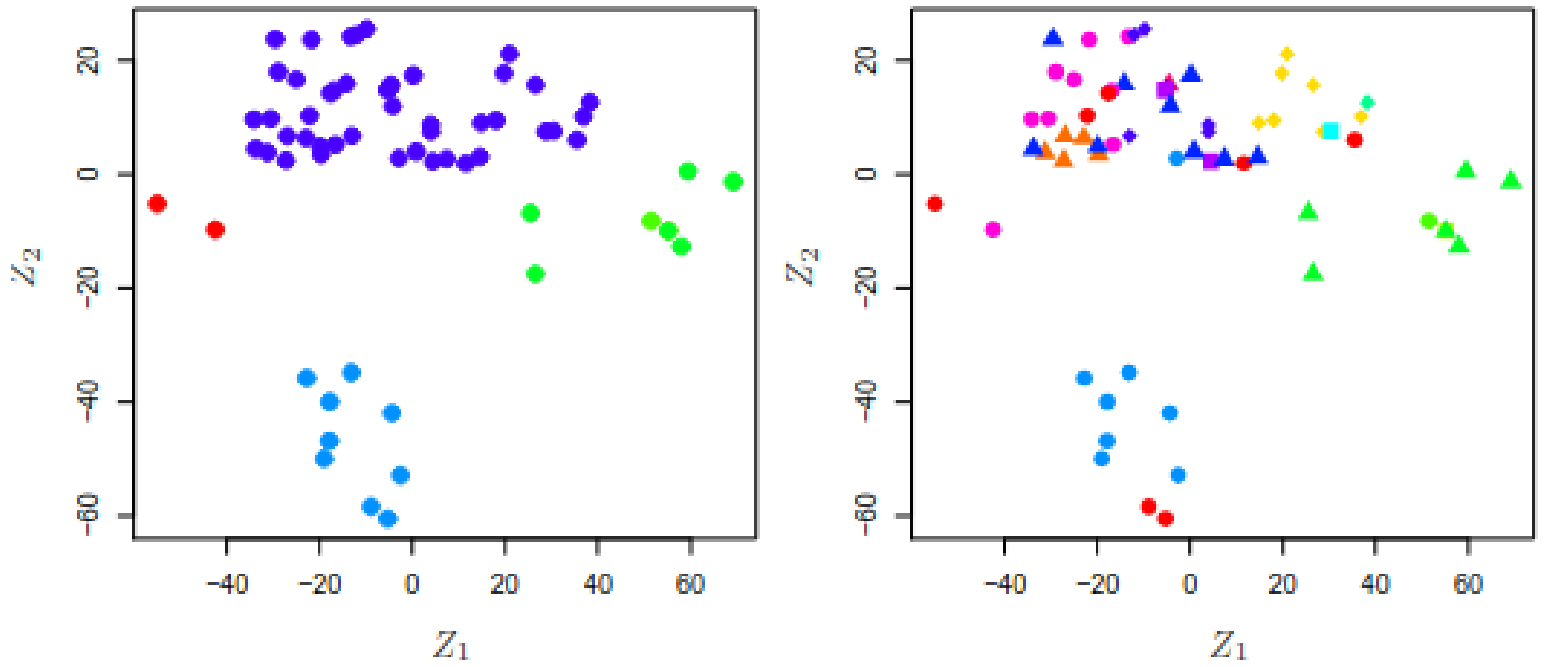


FIGURE 1.4. Left: *Representation of the* NCI60 *gene expression data set in a two-dimensional space,* Z_1 *and* Z_2*. Each point corresponds to one of the* 64 *cell lines. There appear to be four groups of cell lines, which we have represented using different colors.* Right: *Same as left panel except that we have represented each of the* 14 *different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.*

Variable types

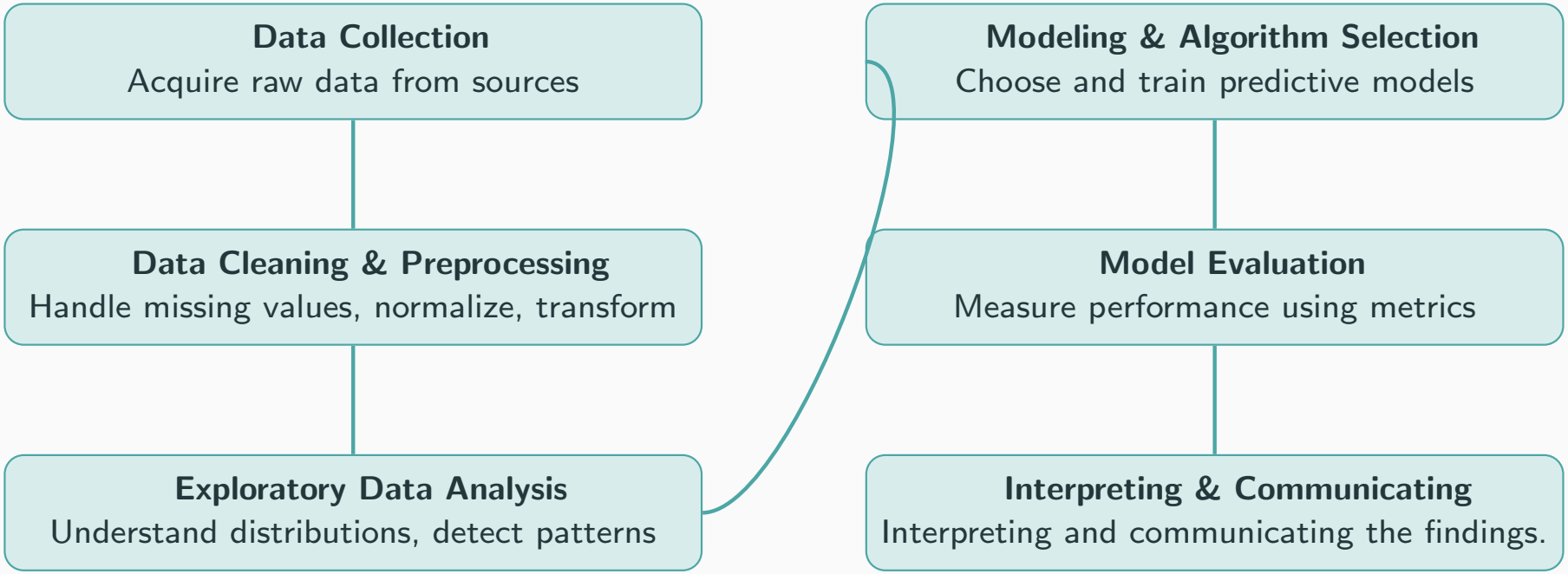
- **Quantitative data**: In the first example, we predict a **numerical value**.

Variable types

- **Quantitative data**: In the first example, we predict a **numerical value**.
- **Qualitative data**: In the second example, we predict a **label**. Qualitative variables are also referred to as **categorical**.

Disclaimer: For the purpose of this class, we will only deal with these two types of variables.

Data Analysis Pipeline



Exploratory Data Analysis

What is Exploratory Data Analysis?

Definition

Exploratory Data Analysis (EDA) is the process of visually and statistically examining datasets to uncover patterns, spot anomalies, test hypotheses, and check assumptions using summary statistics and graphical representations.

Goals of EDA:

- Understand data structure and distributions, thus uncovering relationships between variables
- Detect outliers and missing values
- Formulate and test hypotheses
- Guide the choice of applicable models

Types of EDA

There are 4 types of EDA:

- **Non-graphical**: it consists of **summary statistics**.
- **Graphical**: summarizing data properties in a **diagrammatic or pictorial** way.
- **Univariate**: takes only **one variable** into account.
- **Multivariate**: takes into account **multiple variables and their relation**.

This part of the presentation greatly follows the contents of the Stanford lecture about EDA, which I recommend reading.

Population vs Sample – Overview

Key Idea:

- **Population:** Entire set of data or subjects you want to study.
- **Sample:** A subset of the population used to make estimates.

The values for the **population** are usually unavailable, hence the necessity of analyzing it through the **samples**.

Quick Tip: Population = everyone, Sample = some of them

Attention!

Always distinguish between population and sample. Using the wrong formulas can lead to incorrect analysis results!

Central Tendency

Definition

The **central tendency** measures attempt to describe a datasets with a value that represents the centre of its distribution.

- Common measures of central tendency:
  - **Mean** – Arithmetic average
  - **Median** – Middle value
  - **Mode** – Most frequent value

Mean (Arithmetic Average)

- Represents the central value of the dataset.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- $\bar{x}$: Sample mean
- x_i: Individual value
- n: Total number of observations
- Sum of all of the data values divided by the number of values

Mean: Example

Dataset:

$$x = [2, 4, 4, 6, 8, 10, 10, 10]$$

Mean:

$$\text{Mean} = \frac{\sum x_i}{n} = \frac{2 + 4 + 4 + 6 + 8 + 10 + 10 + 10}{8} = \frac{54}{8} = 6.75$$

Median

- The middle value when data is sorted.
- Not sensitive to outliers.

$$\text{Median} = \begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & \text{if } n \text{ is even} \end{cases}$$

Median: Example

Dataset:

$$x = [2, 4, 4, 6, 8, 10, 10, 10]$$

Median: - Sorted data: [2, 4, 4, 6, 8, 10, 10, 10] - Number of points $n = 8$ (even), median = average of 4th and 5th values:

$$\text{Median} = \frac{6 + 8}{2} = 7$$

Mode

- The **mode** is the value that appears most frequently in the dataset.
- A dataset can have one mode (unimodal), more than one (multimodal), or none.

Mode: Example

Dataset:

$$x = [2, 4, 4, 6, 8, 10, 10, 10]$$

Mode: - Most frequent value(s): 10

$$\text{Mode} = 10$$

Mean, Median, Mode

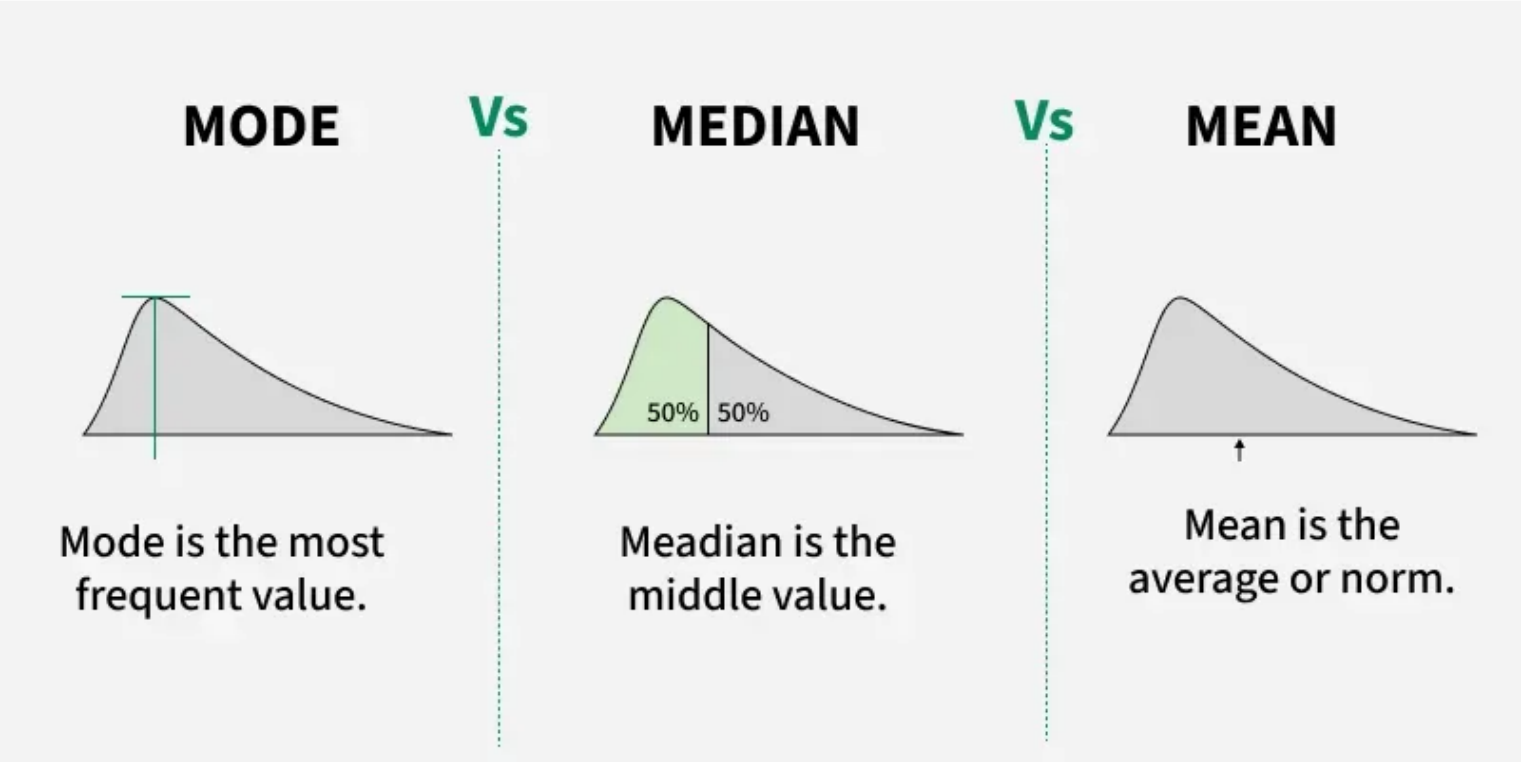


Figure 1: Visual representation of mean, median, and mode.

Source: GeekForGeeks

Spread

Definition

The **spread** is an indicator of how far away from the center we are still likely to find data values.

- Common measures of spread:
  - **Variance**
  - **Standard Deviation**
  - **Interquartile range**

Unbiasedness

Unbiasedness: when calculated for many different random samples from the same population, the average should match the corresponding population quantity. Therefore, when calculating spread metrics for a **sample** and not for the whole population, the denominator will be different.

- $n - 1$ instead of n for **variance** and **standard deviation**
- For other metrics, unbiasing can be more complicated - in any case, **we do not delve into the mathematical details of unbiasedness in this class**!

Variance

Variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Measures **average squared distance** of data from the mean.

Standard Deviation

Standard Deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Square root of the variance.
- It expresses **distance from the mean**, but it is in the same units as the data.
- Commonly used to express data dispersion, and useful to detect **outliers**.

Interquartile Range (IQR)

Range:

$$\text{Range} = \max(x) - \min(x)$$

Interquartile Range:

$$\text{IQR} = Q_3 - Q_1$$

- Range of the middle 50% of the data.
- Q_1: 25th percentile, Q_3: 75th percentile
- Because the values in the top and bottom quarter of the data can be moved without influencing IQR, IQR is very **robust** against outliers. This is not true for range, which can **drastically change** if the analyzed sample changes!
- Useful for detecting outliers.

Skewness

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

- You start by the **mean** ($\bar{x}$) and **standard deviation** (s) of your dataset; the computed values are divided by the **population size**.
- Measures **asymmetry** of the data distribution.
- In simplified terms, you can think of it as a measure that tells you if the data is **"leaning"** towards one side.
- Positive: right-skewed, Negative: left-skewed

The above formula refers to **population**. As the purpose of this class is for you to understand the **theory** behind summary statistics, we will not delve into the details of unbiasedness.

Skewness: Example

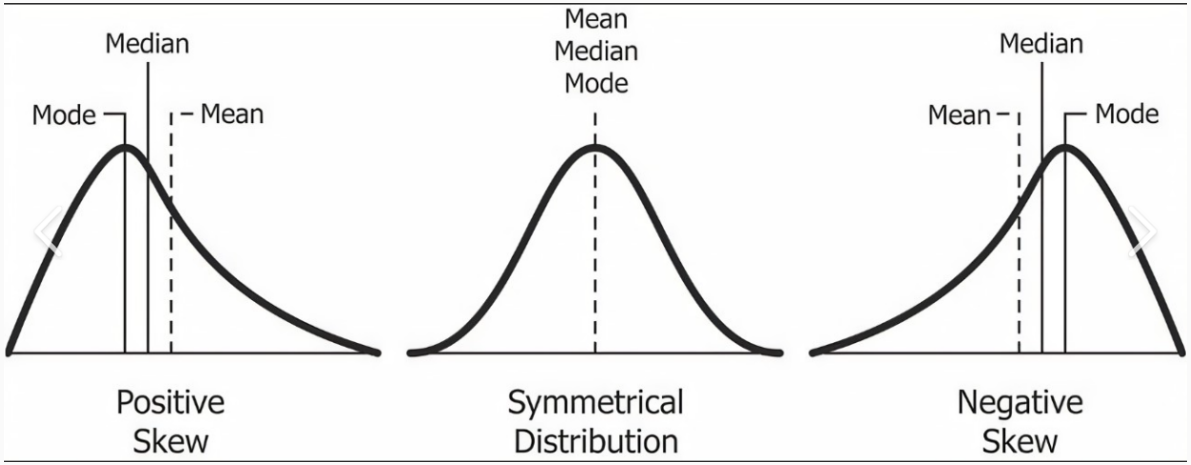


Figure 2: Example of left- and right-skewness.

Source: Analytics Vidhya

Kurtosis

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

- You start by the **mean** ($\bar{x}$) and **standard deviation** (s) of your dataset; the computed values are divided by the **population size**.
- Measures **"peakedness"** of the distribution.
- High kurtosis = heavy tails
- Kurtosis can tell you whether the dataset contains more **extreme values** (= high tails)

The above formula refers to **population**. As the purpose of this class is for you to understand the **theory** behind summary statistics, we will not delve into the details of unbiasedness.

Kurtosis: Example

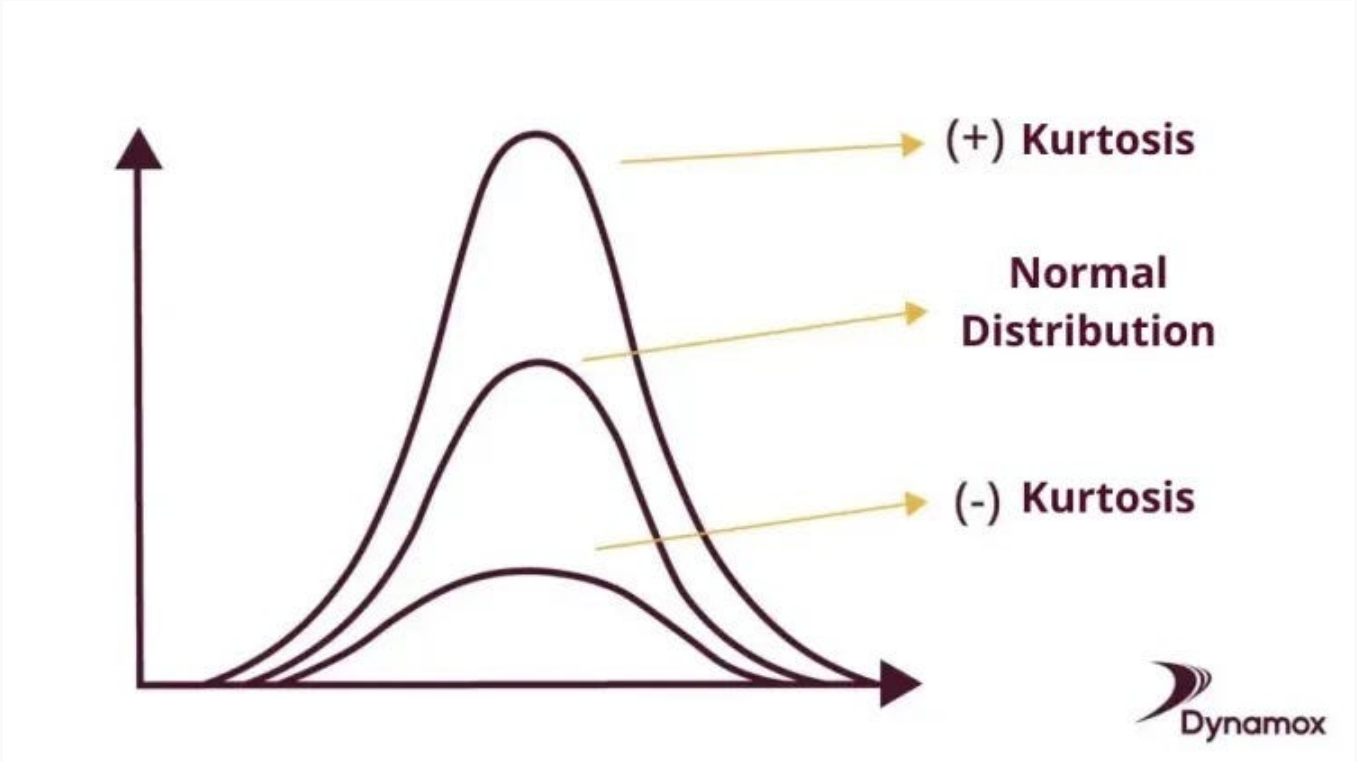


Figure 3: Explanation of kurtosis.

Source: Medium

Graphical EDA

Sample Dataset

Data	Frequency
1	6
2	2
3	9
4	6
5	11
6	3
7	4
8	1
9	2

Table 1: Data and Frequency Table

Histograms

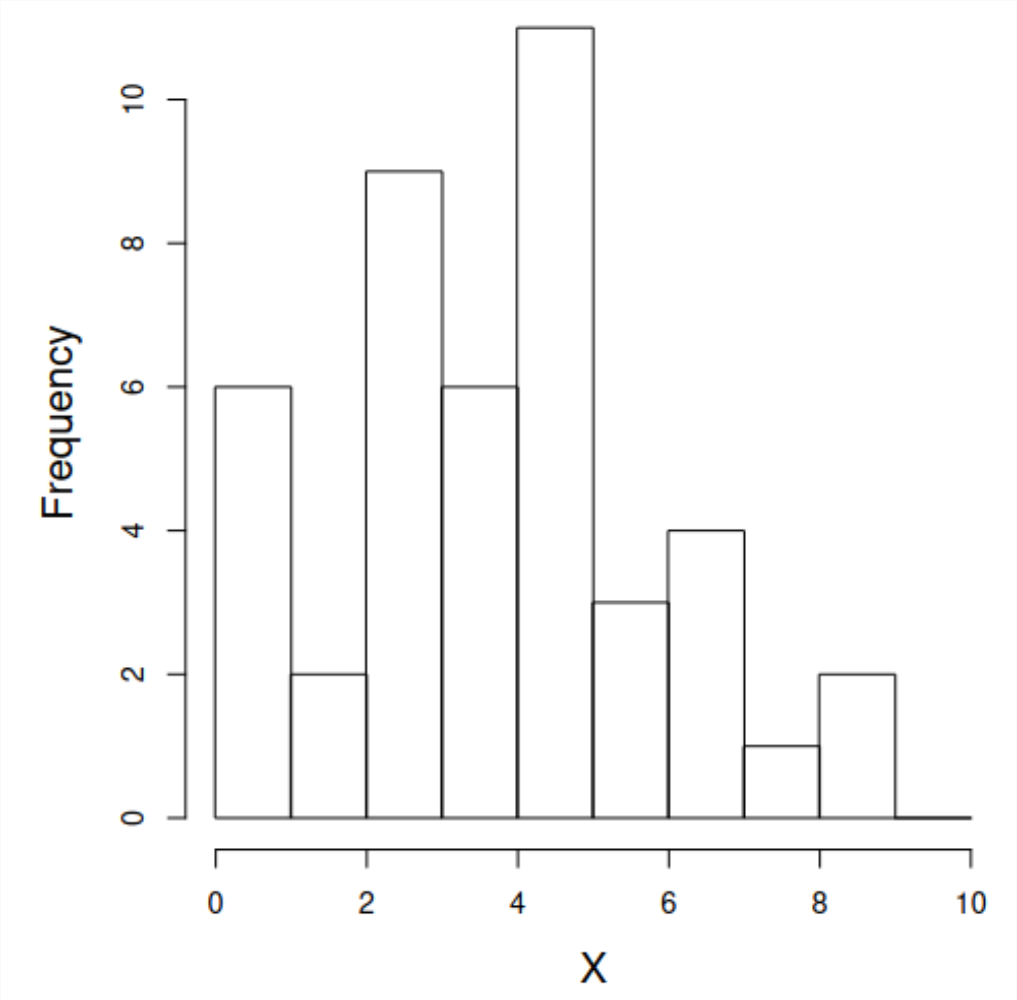


Figure 4: Example histogram of Table 1.

Source: Stanford lectures on EDA.

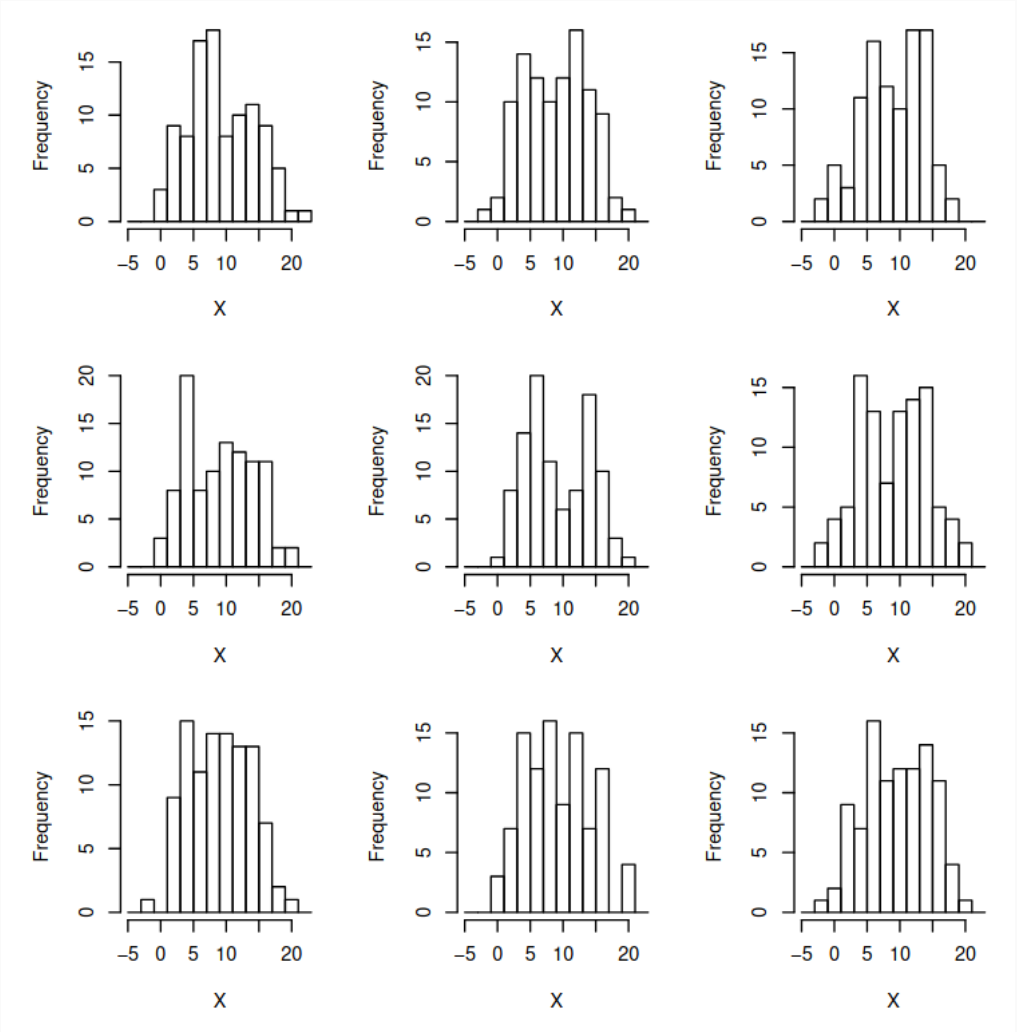
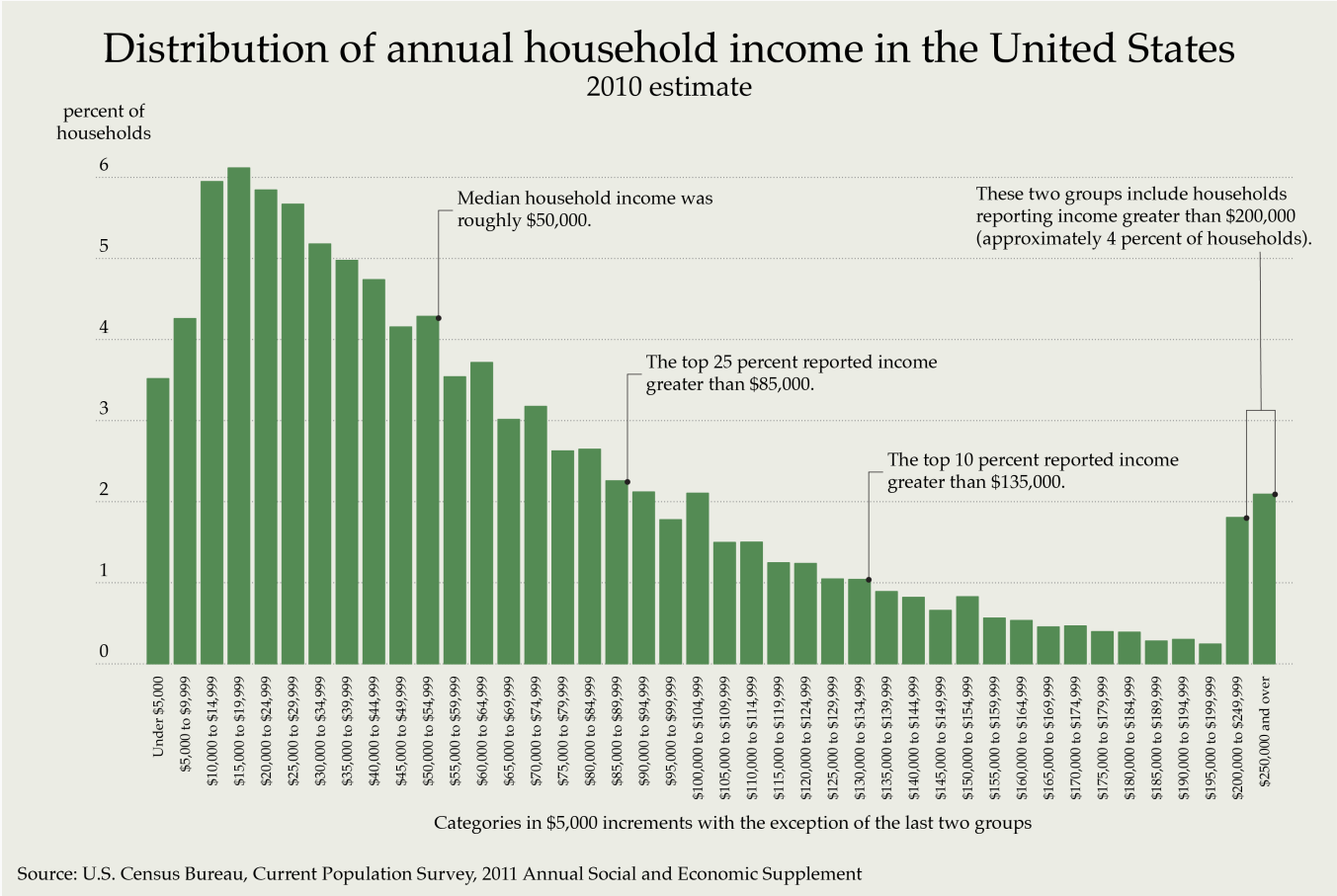


Figure 5: Example histograms of several samples of size 100.

Source: Stanford lectures on EDA.



Source: Wikimedia Commons.

Stem and Leaf Plots

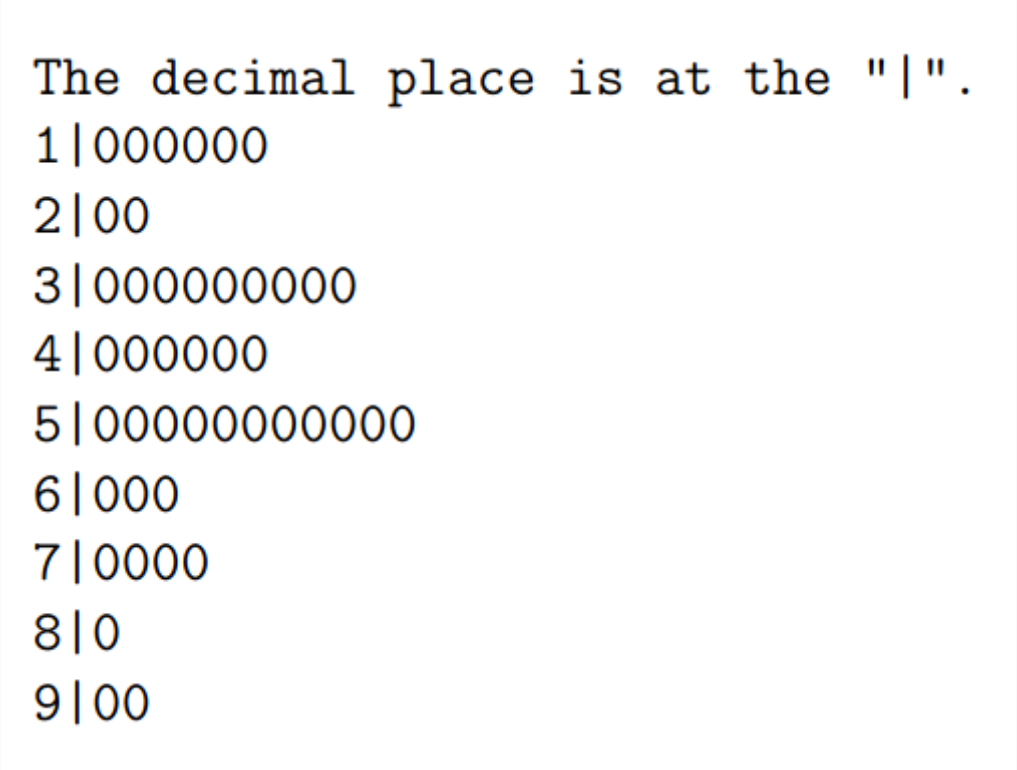
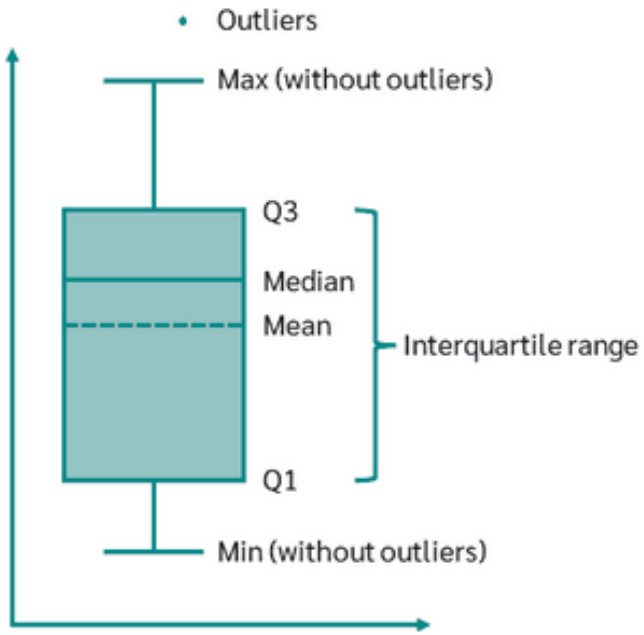


Figure 6: Example stem-and-leaf plot for the data from Table 1.

Source: Stanford lectures on EDA.

Boxplots



The box indicates the range in which the middle 50% of all data lies

Thus, the lower end of the box is the 1st quartile and the upper end is the 3rd quartile

Between Q1 and Q3, is the interquartile range

In the boxplot, the solid line indicates the median and the dashed line indicates the mean.

The T-shaped whiskers go to the last point, which is still within 1.5 times the interquartile range.

Points that are further away are considered extreme values (outliers).

Figure 7: Explanation of boxplots.

Source: Datatab.

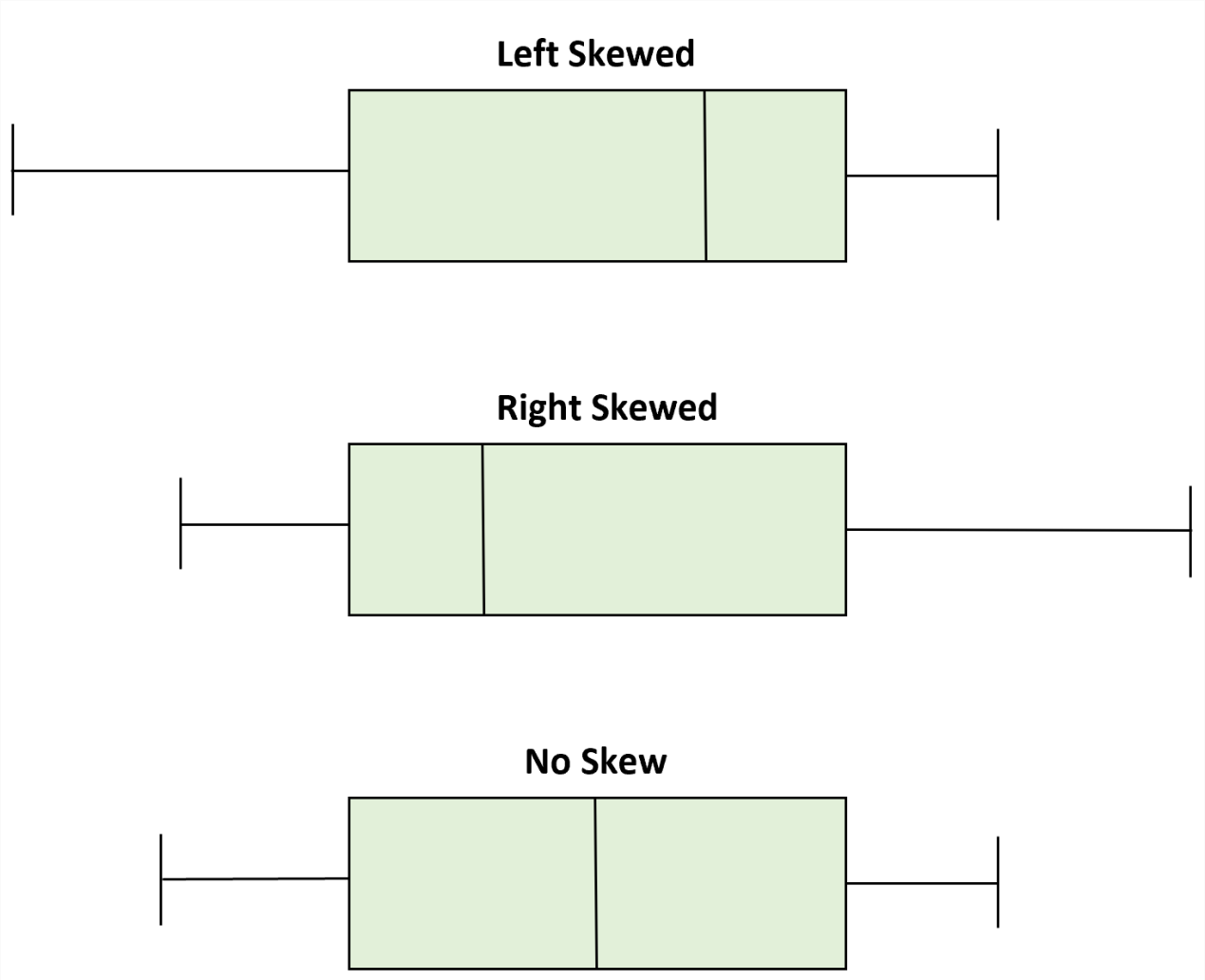


Figure 8: How to spot skewness in boxplots.

Source: Statology.

Quantile-Normal Plots

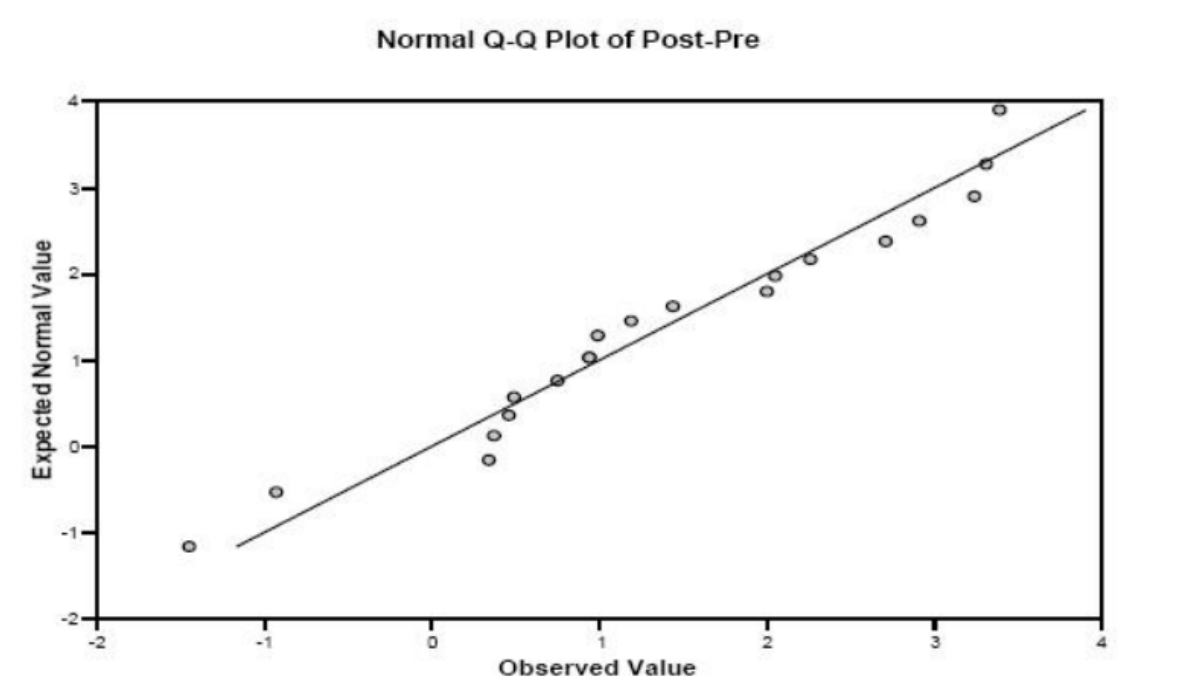


Figure 9: Example of quantile-normal plot where the data approximately follows normal distribution.

Source: Stanford lectures on EDA.

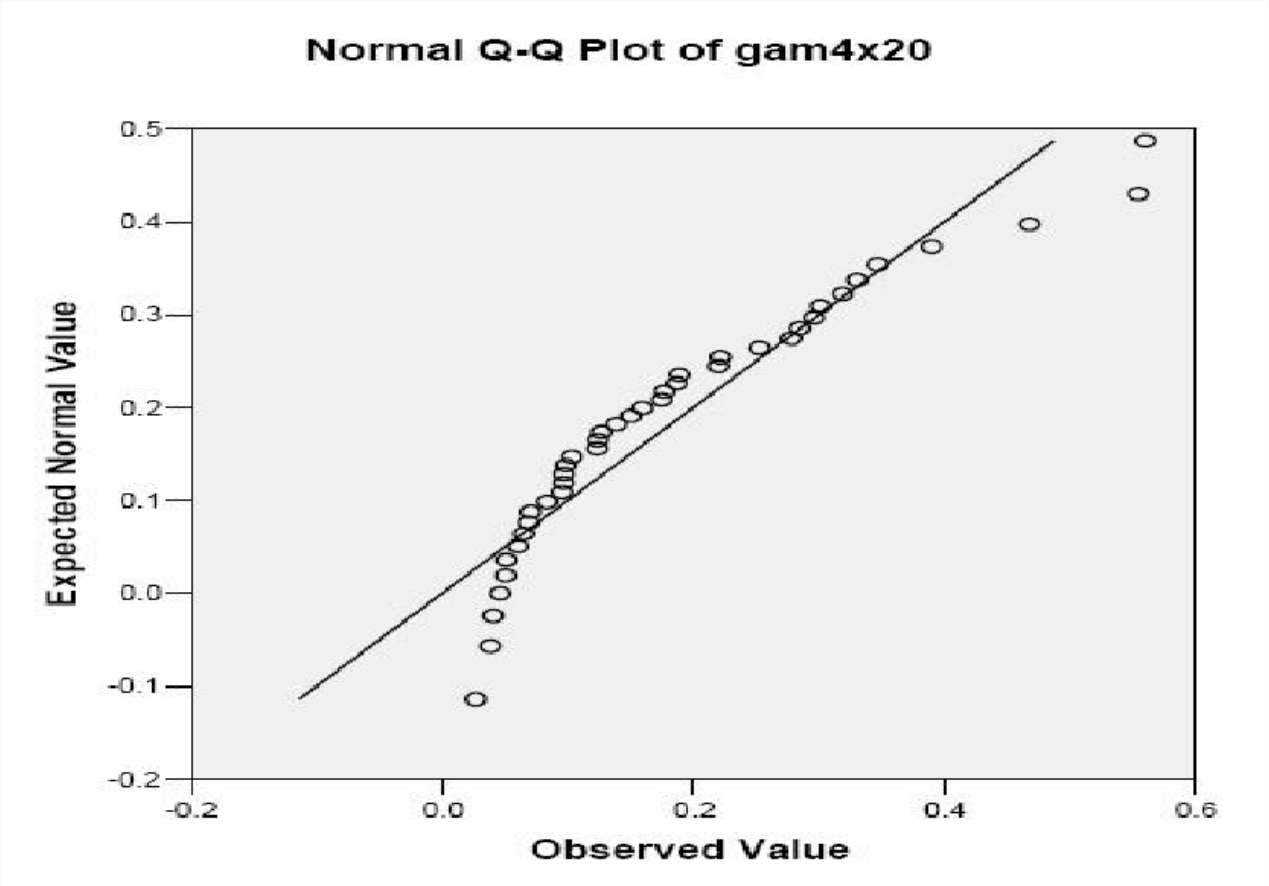


Figure 10: Example of quantile-normal plot displaying right skew.

Source: Stanford lectures on EDA.

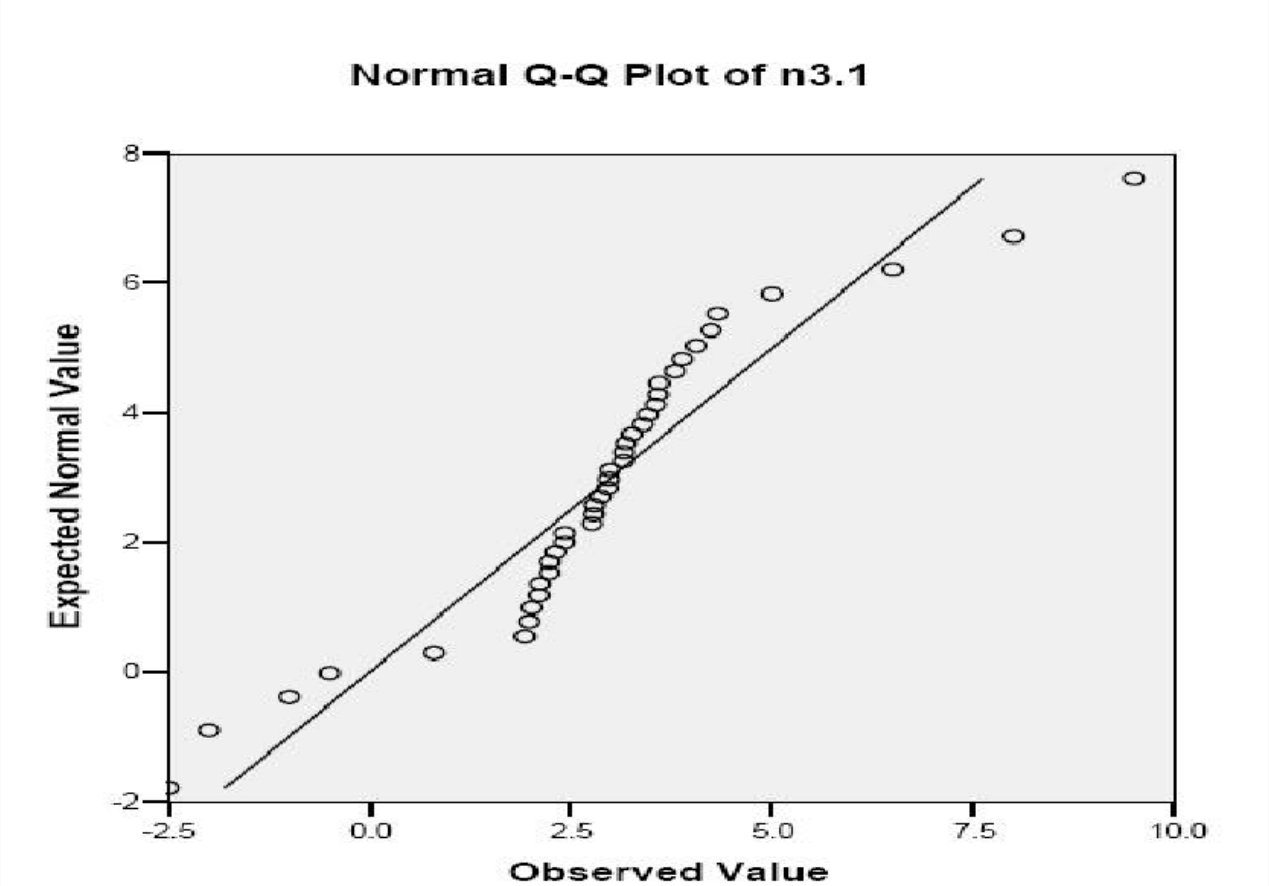


Figure 11: Example of quantile-normal plot with fat tails (positive kurtosis).

Source: Stanford lectures on EDA.

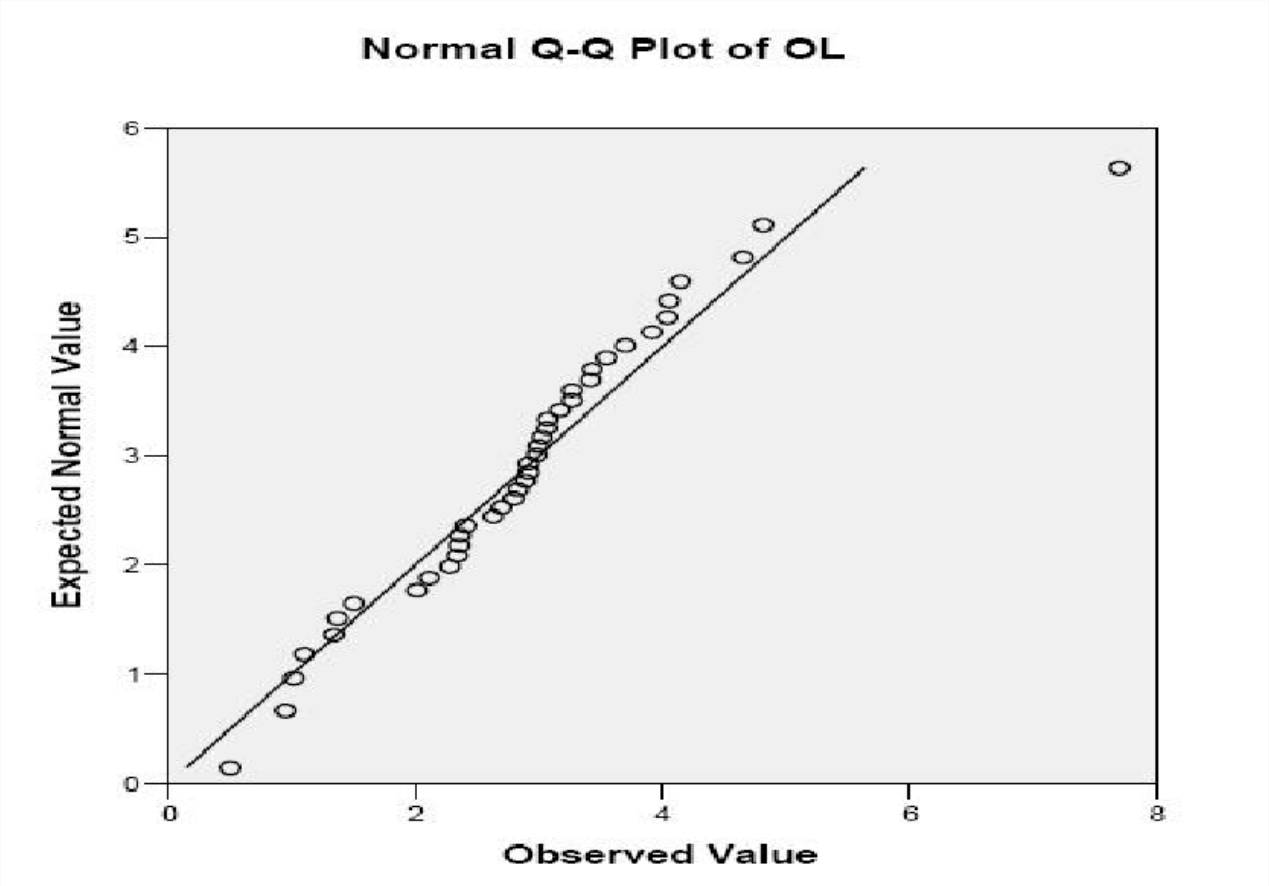


Figure 12: Example of quantile-normal plot with an outlier.

Source: Stanford lectures on EDA.

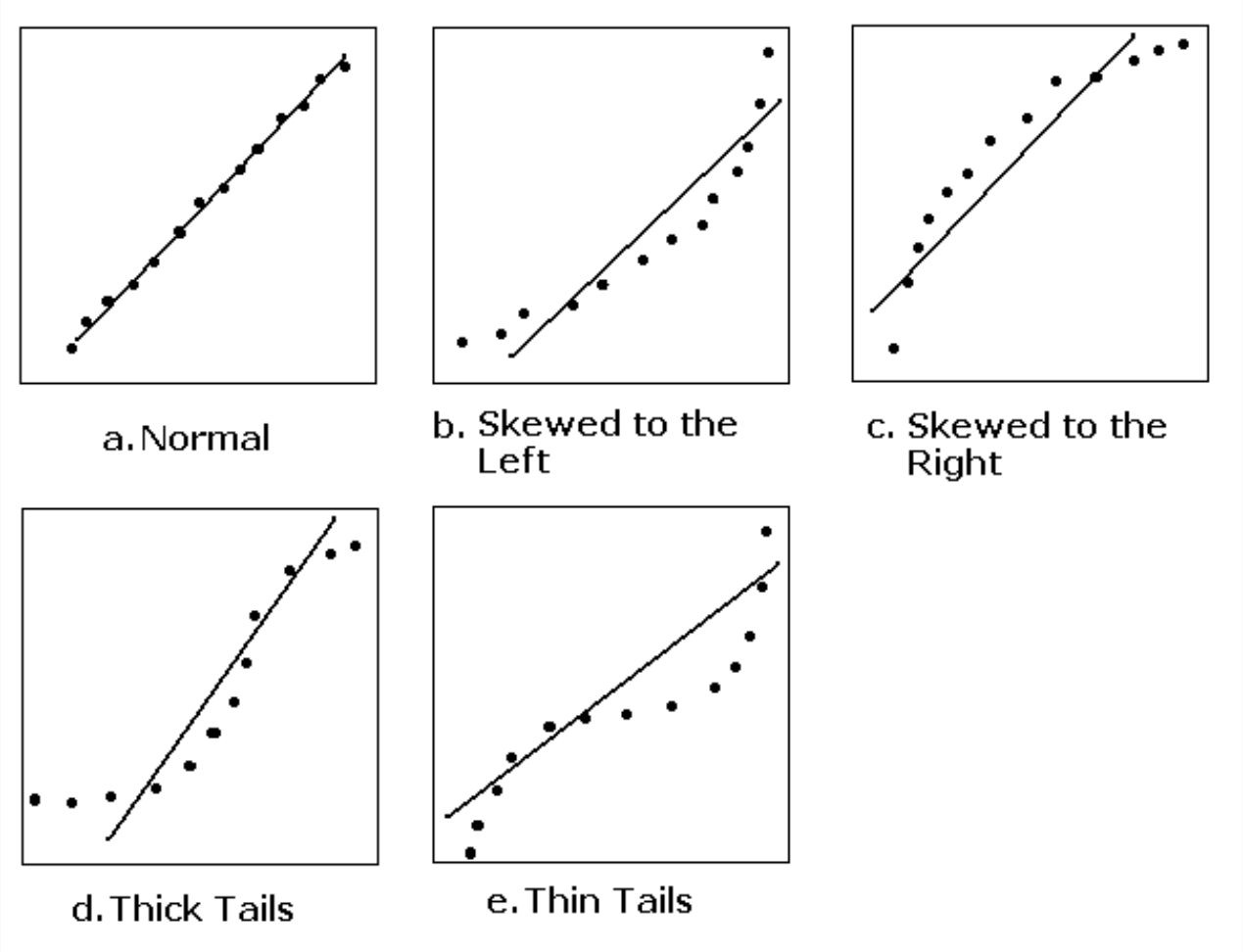


Figure 13: Examples of interpretations of quantile-normal plots.

Source: Normal Plot

Multivariate Non-Graphical Data Analysis

Cross-Tabulation

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Age and sex data for a population sample.

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Cross-Tabulation of the original data.

What Are Covariance and Correlation?

- Both measure the relationship between two variables.
- **Covariance:** Indicates the strength and direction of the relationship: *How much does one variable change if the other one changes?*
- **Correlation:** Indicates the strength and direction, **standardized**.

Here, we will present **sample** covariance and correlation (remember the definition of unbiasedness). To compute **population** correlation and covariance, the denominator will be n instead of $n - 1$.

Sample Covariance

Definition:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Positive: Variables move in the same direction.
- Negative: Variables move in opposite directions.
- Covariance near 0: Variables vary independently.
- Magnitude depends on units (not standardized).

Sample Correlation

Definition:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

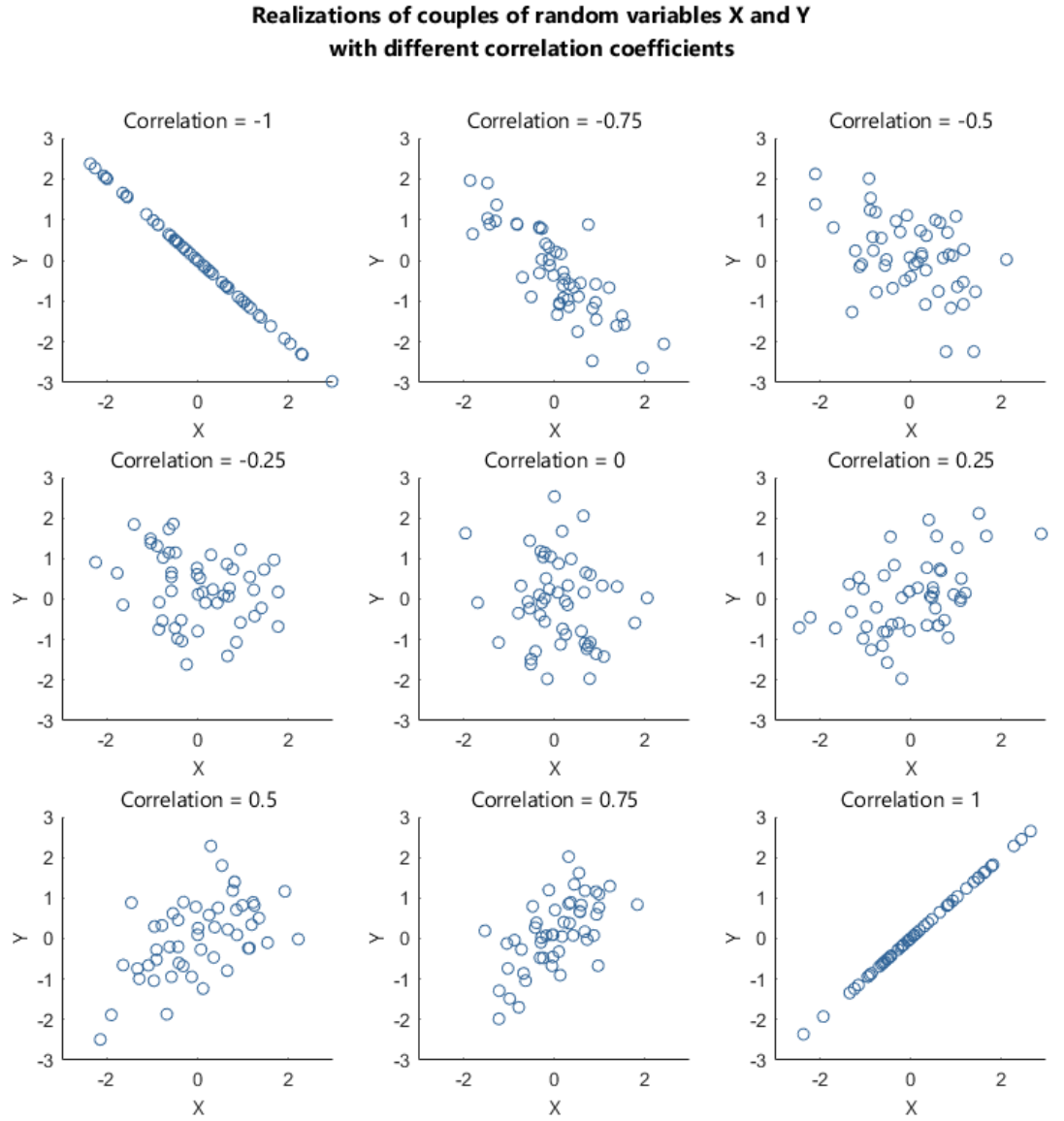
where σ_X and σ_Y are standard deviations of X and Y.

- Ranges from -1 to $+1$.
- $r = 1$: Perfect positive relationship.
- $r = -1$: Perfect negative relationship.
- $r = 0$: No linear relationship.

Relationship Between Covariance and Correlation

- Covariance measures direction, correlation standardizes it.
- Correlation removes the effect of units.
- Formula: $\rho = \dfrac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$

Visualization of Correlation



- Strong positive correlation: points close to upward line.
- Strong negative correlation: points close to downward line.
- No correlation: points scattered randomly.

Multivariate Graphical EDA

Side by Side Boxplots

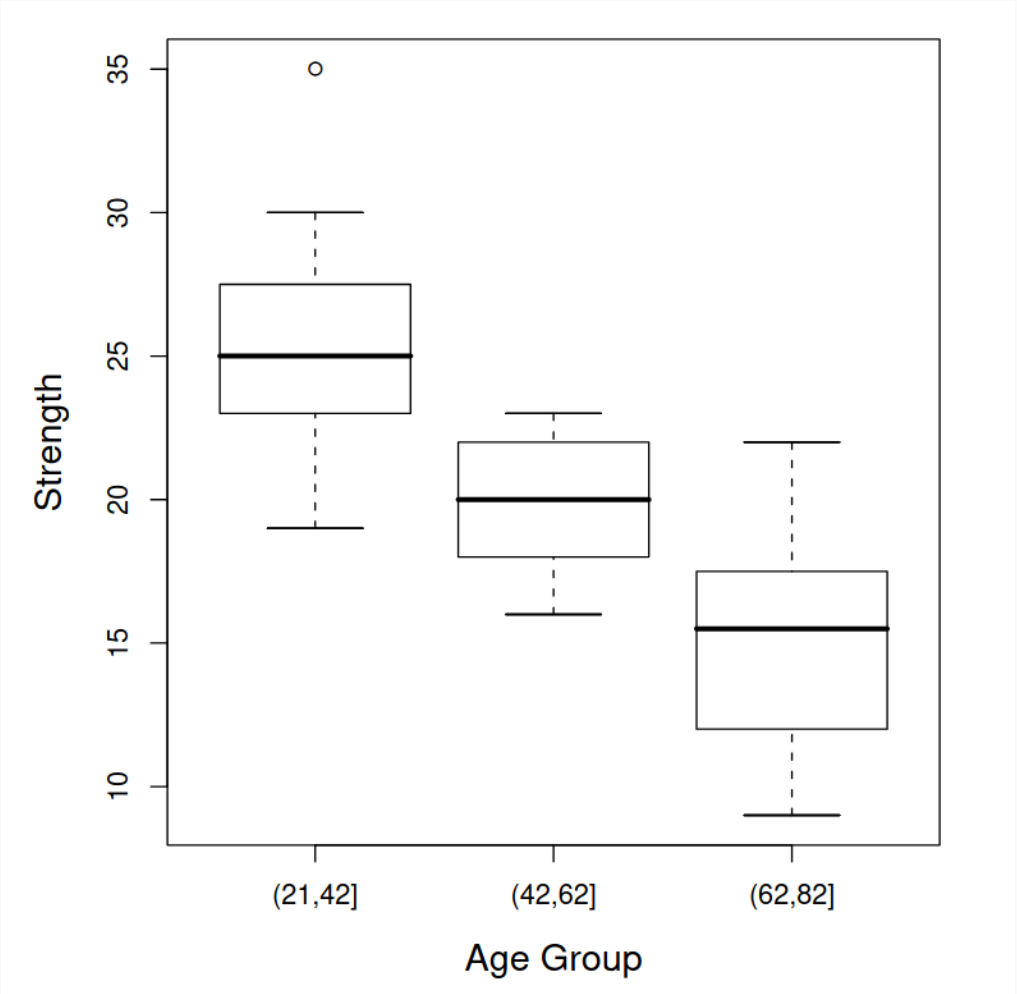


Figure 14: Example of side-by-side boxplot.

Source: Stanford lectures on EDA.

Scatterplots

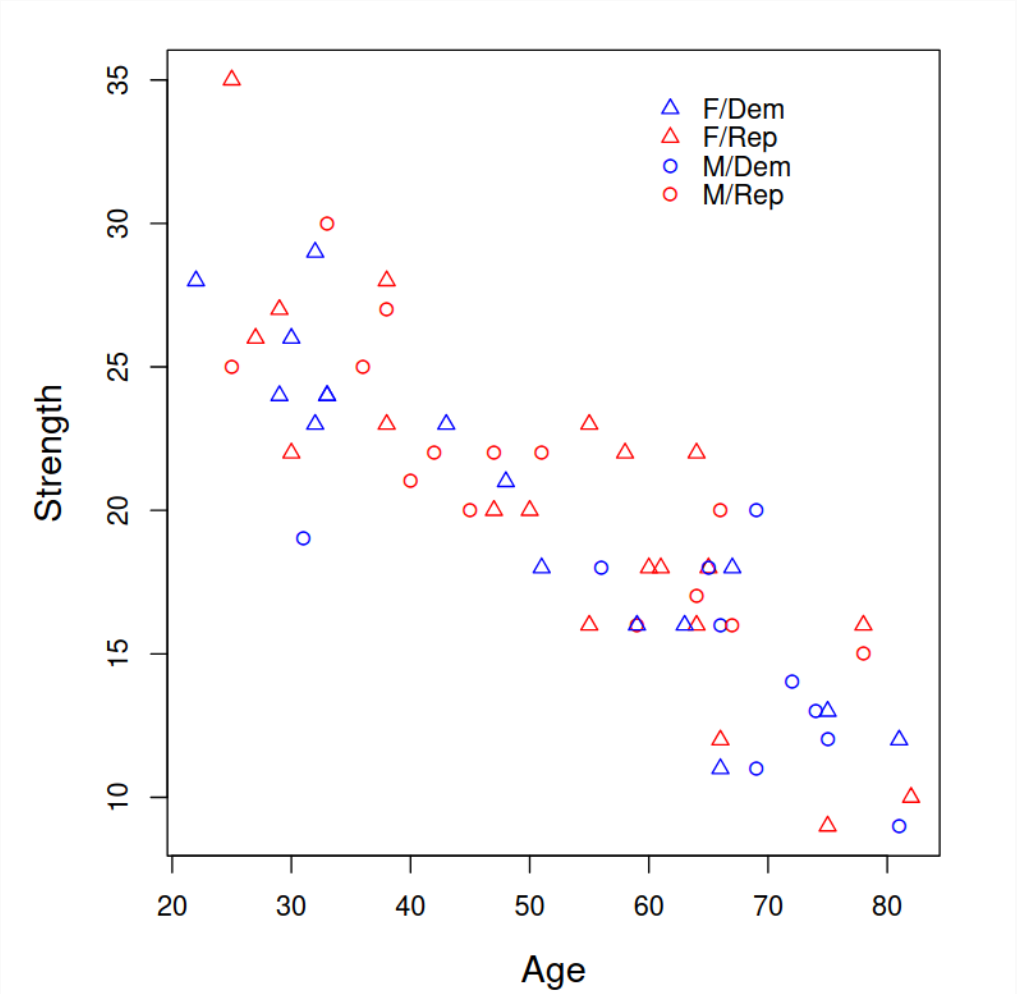


Figure 15: Example of scatterplot.

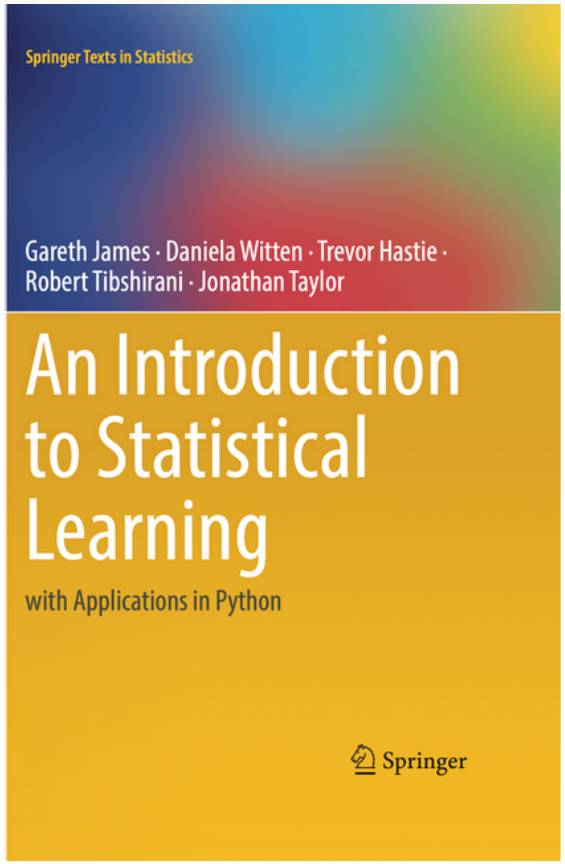
Source: Stanford lectures on EDA.

Conclusion

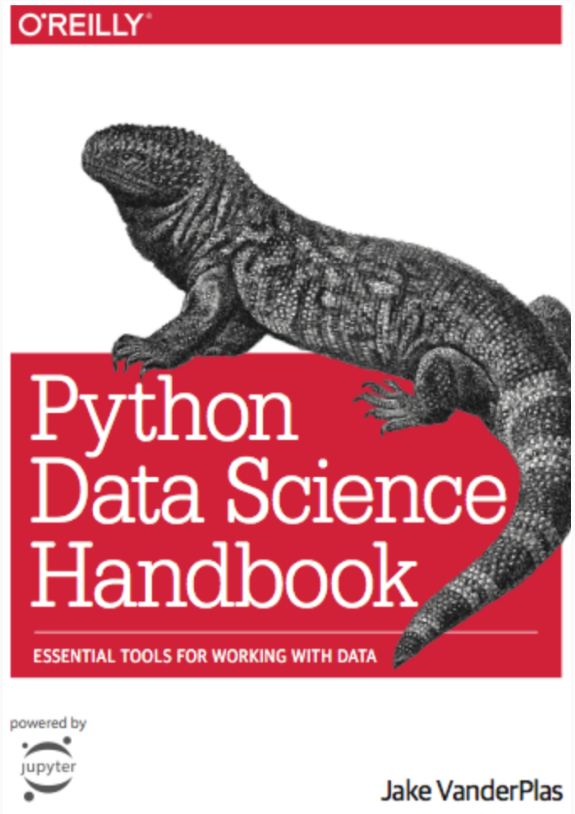
Take-home messages

- Reviewed and extended basic statistical concepts
- Applications of EDA to real-world data
- Revising of Pandas and Seaborn
- **ToDo**: exercising implementation of EDA in Python (homework to come on Moodle - stay tuned!)
- **Next**: Statistical modeling foundations.

Recommended Readings



An Introduction to Statistical Learning



Python Data Science Handbook