Exploratory Data Analysis

Introduction to Data Analysis with Python - Spring Semester 2026

Cecilia Graiff

April 3, 2026

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

Table of Contents

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!

Exploratory Data Analysis

Summary Statistics

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.

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

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

$$\text{Mode} = 10$$

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

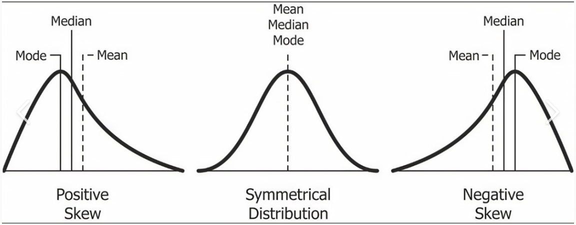


Figure 1: Example of left- and right-skewness.

Source: Analytics Vidhya

Kurtosis

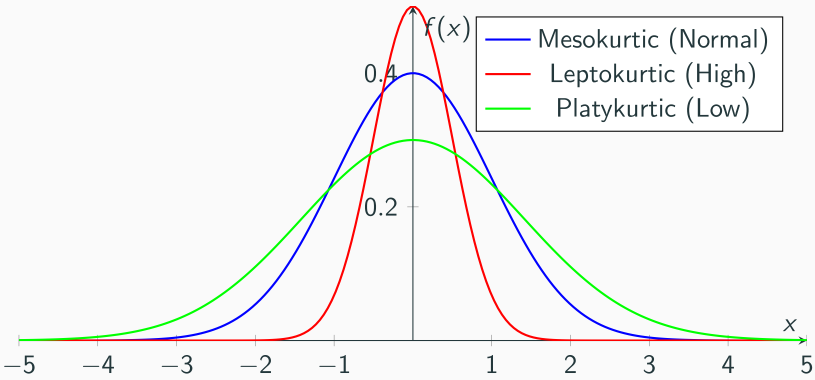
$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

- 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

Comparison of Different Kurtosis Levels



Graphical Exploratory Data Analysis

Graphical EDA

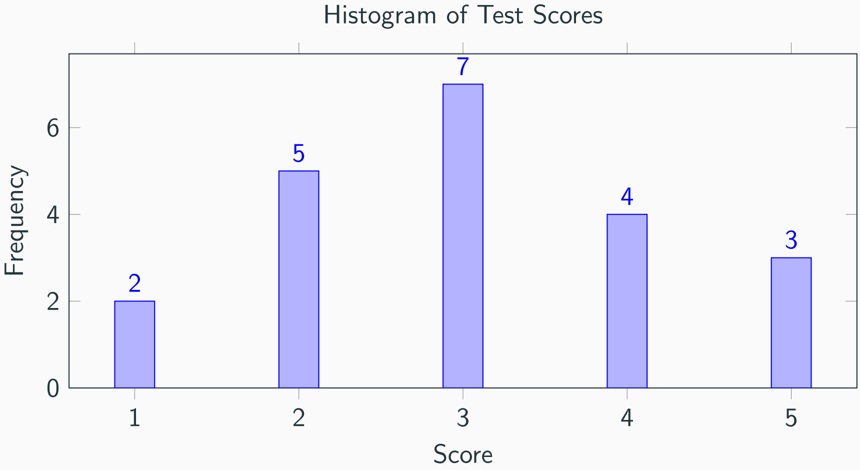
- Last lecture: different types of diagrams, and how to plot them
- Focus of this lecture:
    - How the learned visualization techniques relate to your data
    - What they tell you about your data
    - When and how to choose the right visualization techniques

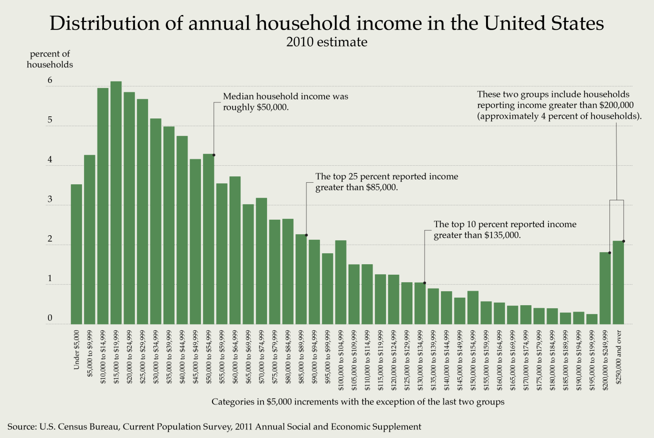
Sample Dataset

Data	Frequency
1	2
2	5
3	7
4	4
5	3

Table 1: Data and Frequency Table

Histograms





Source: Wikimedia Commons.

Boxplots

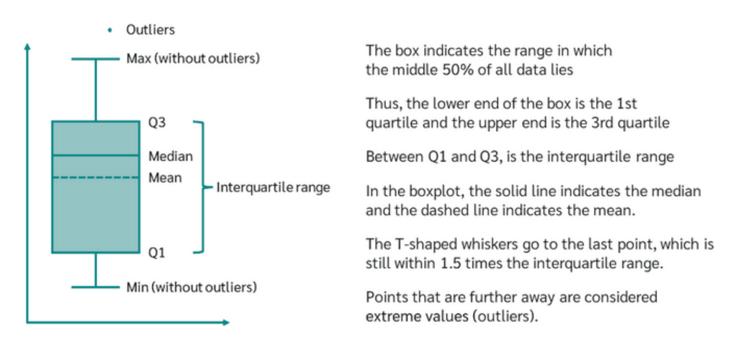


Figure 2: Explanation of boxplots.

Source: Datatab.

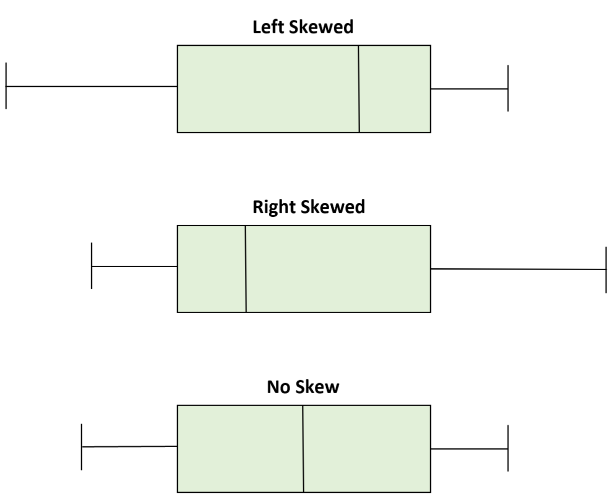


Figure 3: How to spot skewness in boxplots.

Source: Statology.

Quantile-Normal Plots

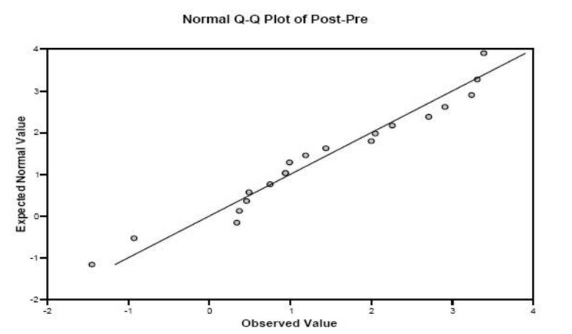


Figure 4: Example of quantile-normal plot where the data approximately follows normal distribution.

Source: Stanford lectures on EDA.

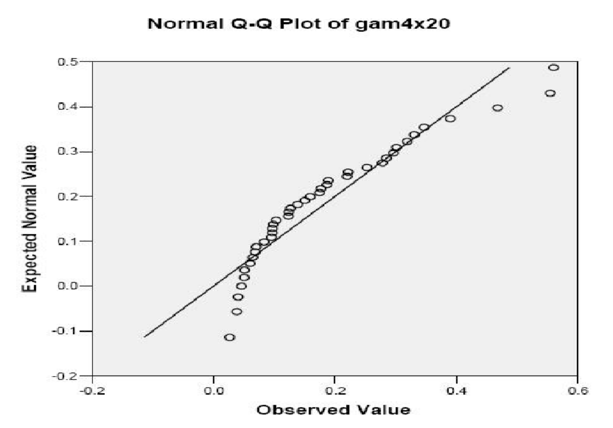


Figure 5: Example of quantile-normal plot displaying right skew.

Source: Stanford lectures on EDA.

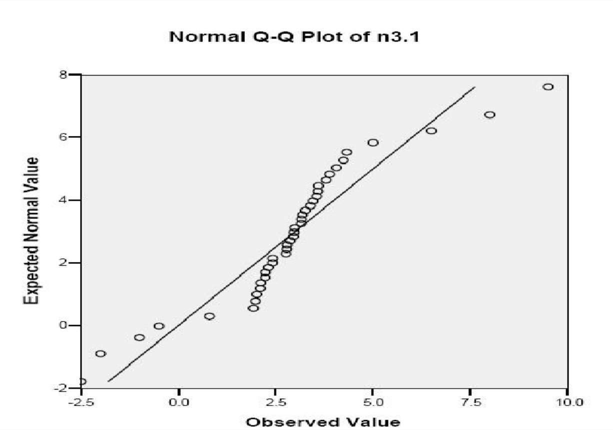


Figure 6: Example of quantile-normal plot with fat tails (positive kurtosis).

Source: Stanford lectures on EDA.

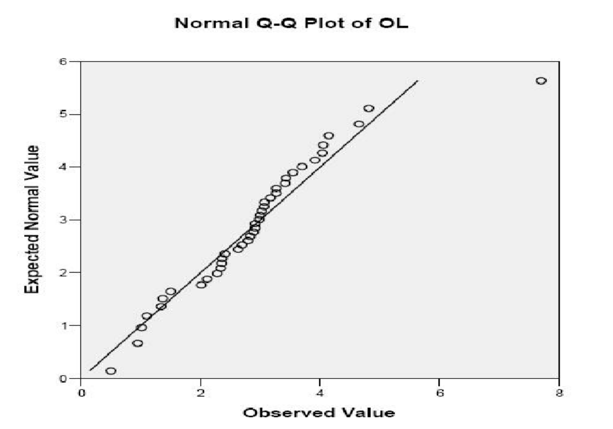


Figure 7: Example of quantile-normal plot with an outlier.

Source: Stanford lectures on EDA.

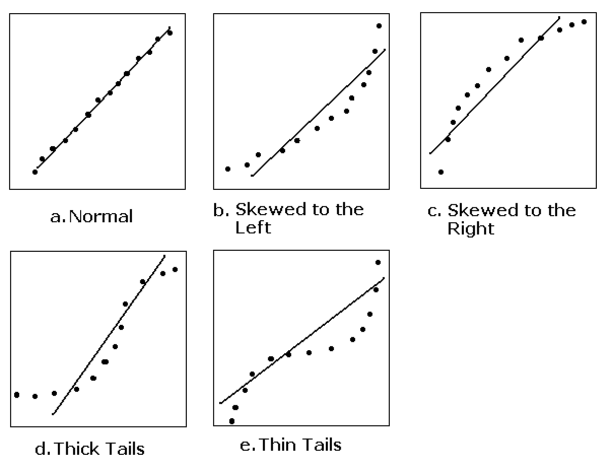


Figure 8: Examples of interpretations of quantile-normal plots.

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:

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

- 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:

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

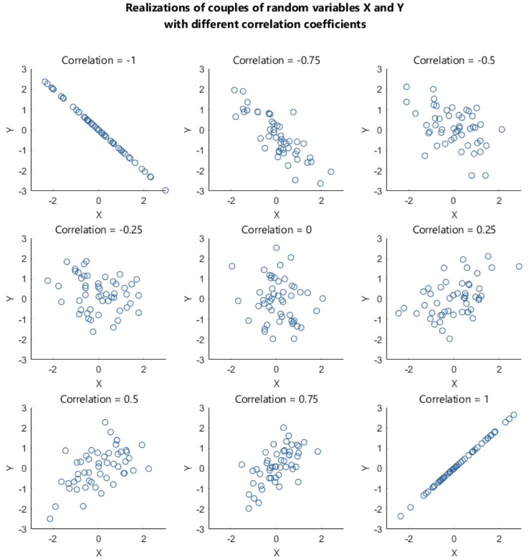
where σ_X and σ_Y are standard deviations of X and Y.

- 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 Exploratory Data Analysis

Side by Side Boxplots

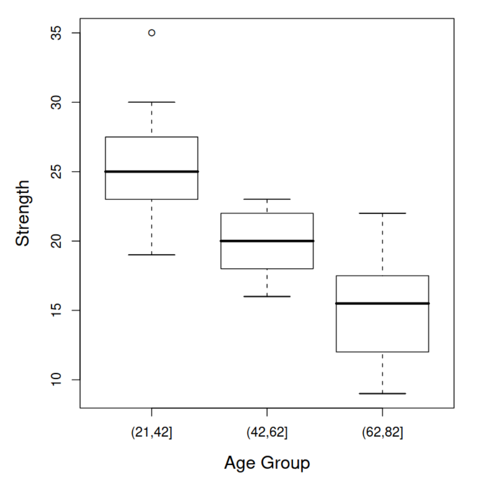


Figure 9: Example of side-by-side boxplot.

Source: Stanford lectures on EDA.

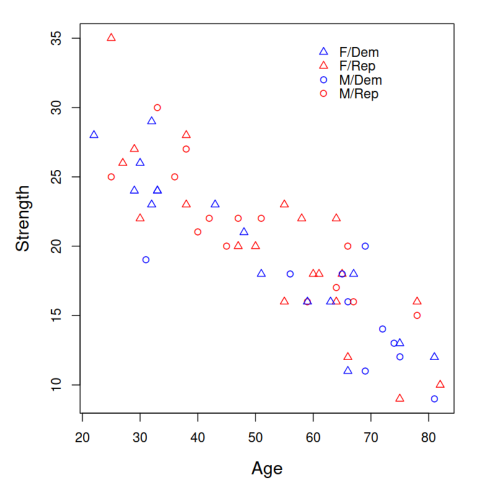


Figure 10: Example of scatterplot.

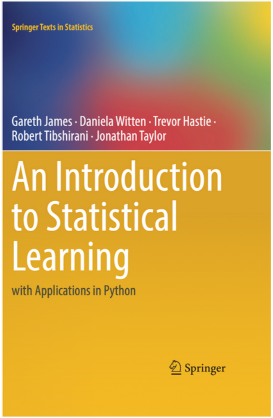
Source: Stanford lectures on EDA.

Conclusion

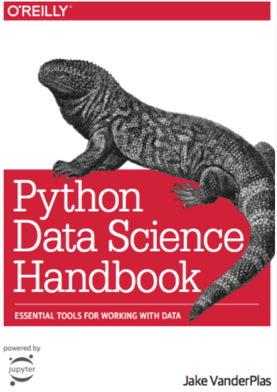
Take-home messages

- 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