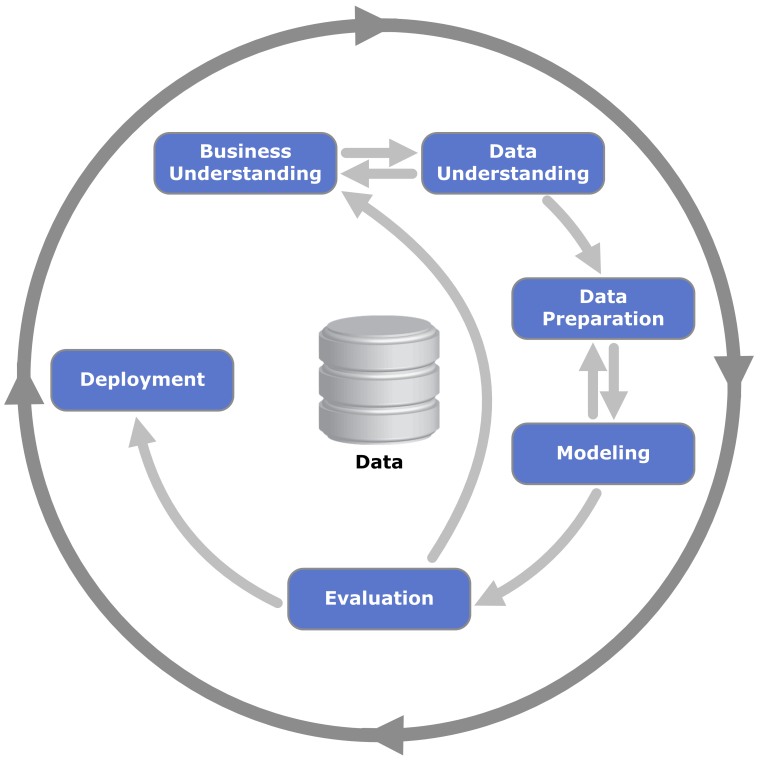
Machine learning

Managing Data



1. Problem Understanding lub Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Normal distribution

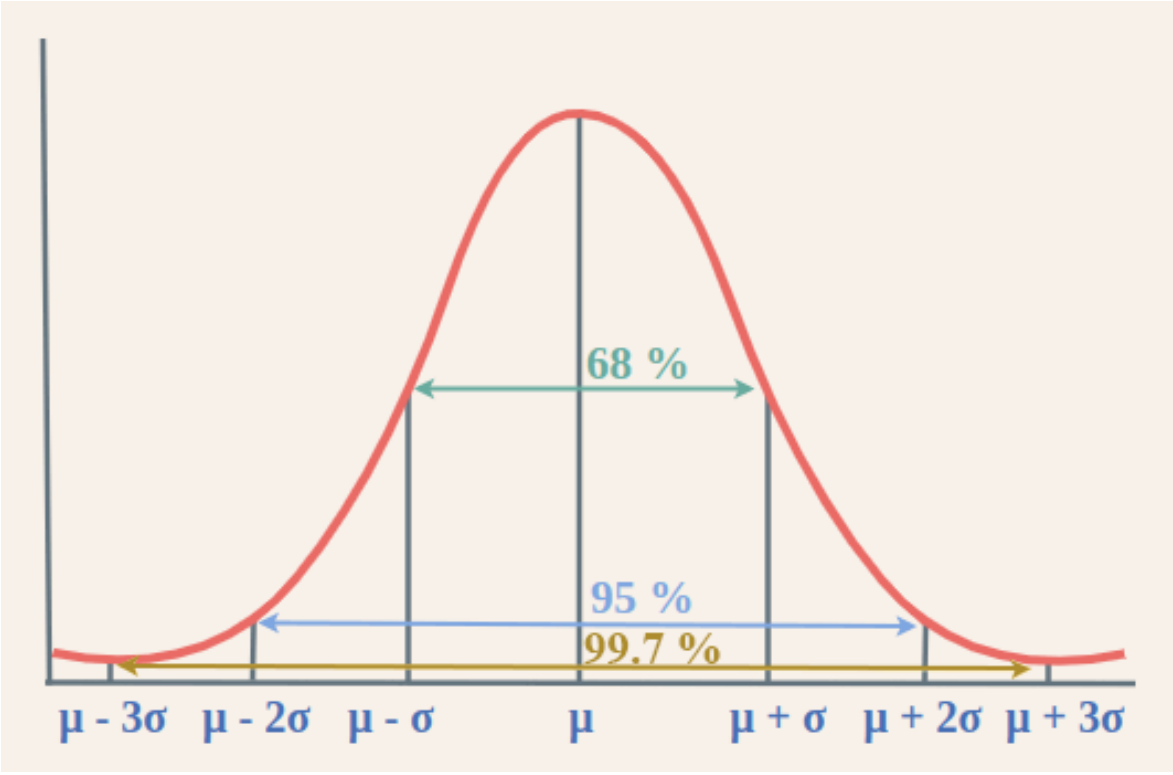
- In machine learning, the Gaussian distribution, is also known as the normal distribution.
- It is a continuous probability distribution function that is symmetrical at the mean, and the majority of data falls within one standard deviation of the mean.
- It is characterized by its bell-shaped curve.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where

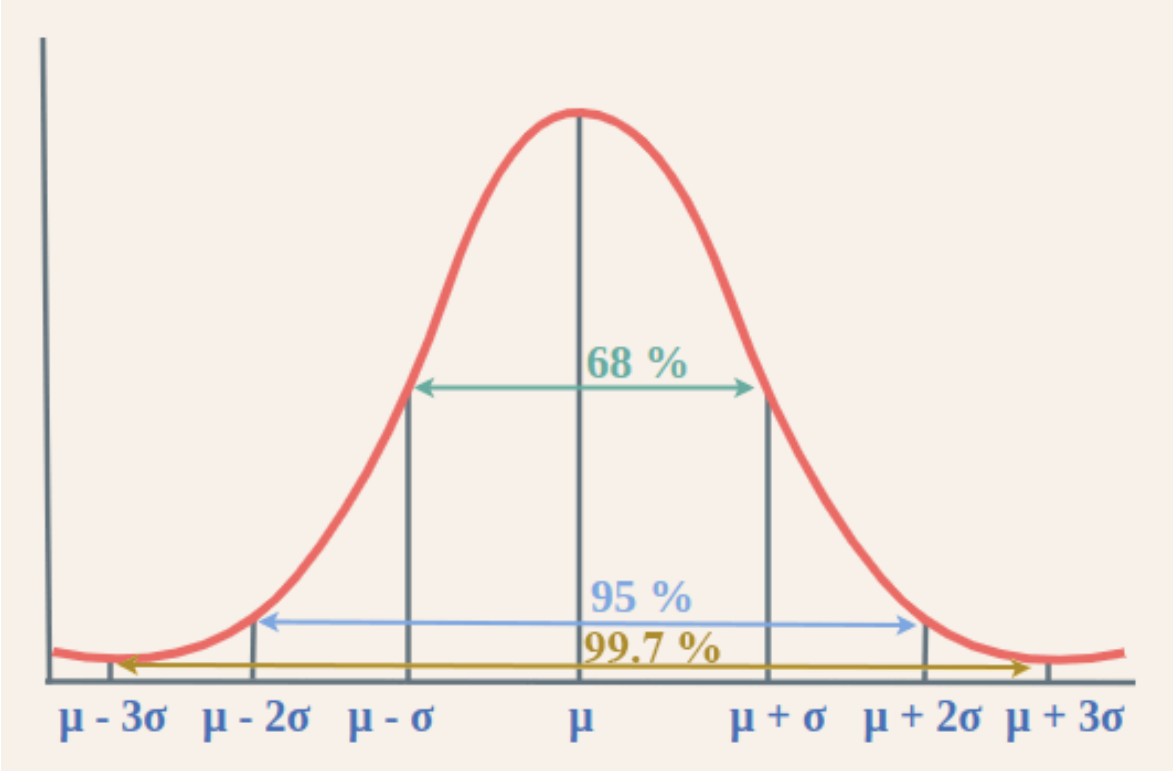
- x represents the Variable
- μ represents the Mean
- σ represents the Standard Deviation
- e represents the base of the Natural Logarithm.

Normal distribution



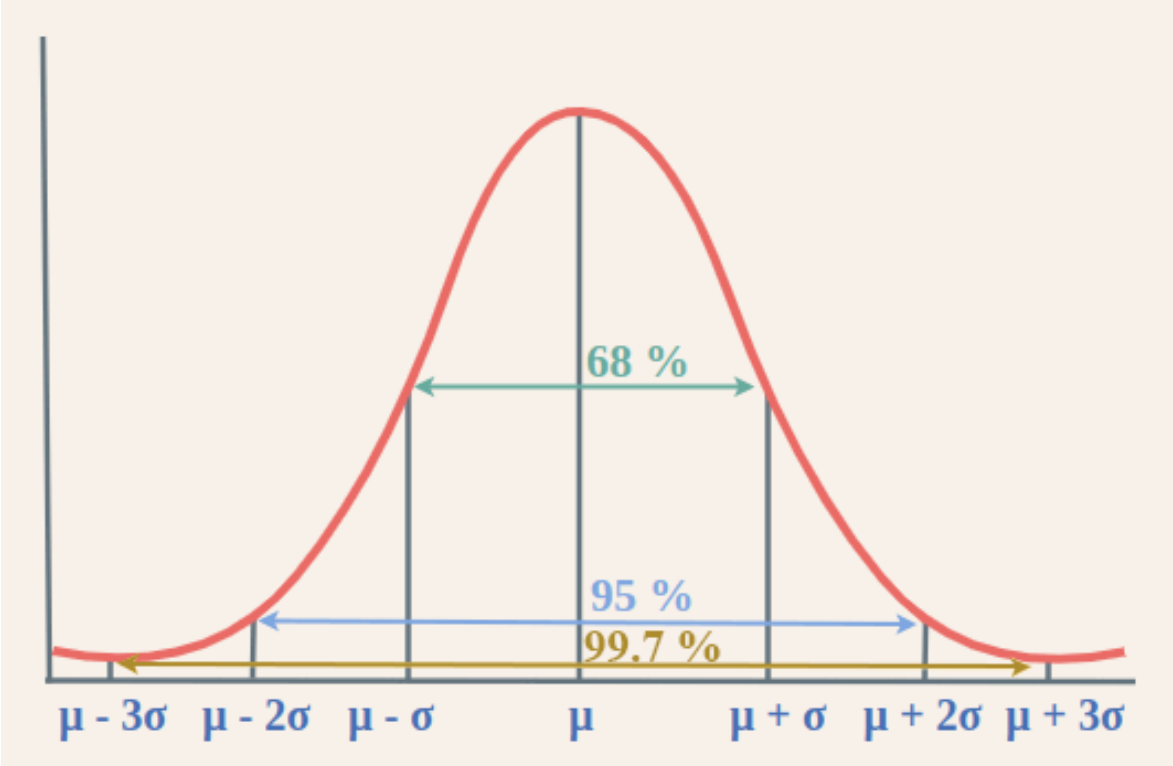
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Normal distribution



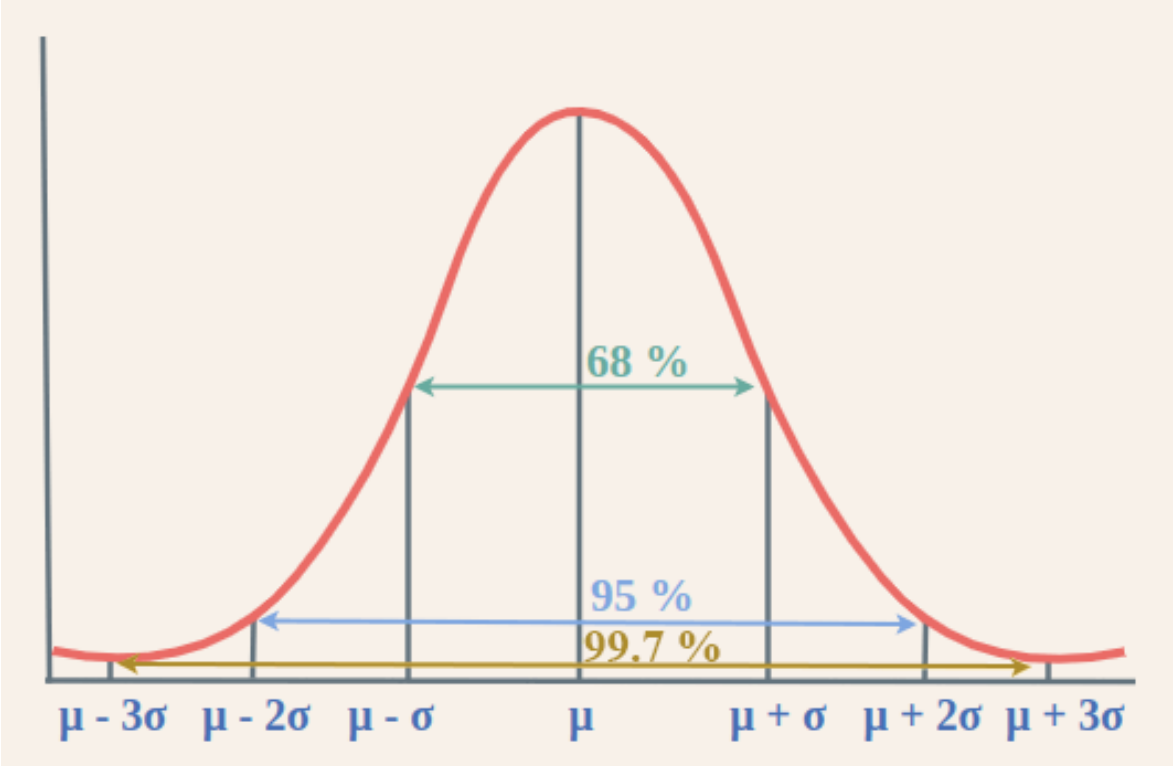
The standard deviations are used to subdivide the area under the normal curve. Each subdivided section **defines the percentage of data, which falls into the specific region of a graph**.

Normal distribution



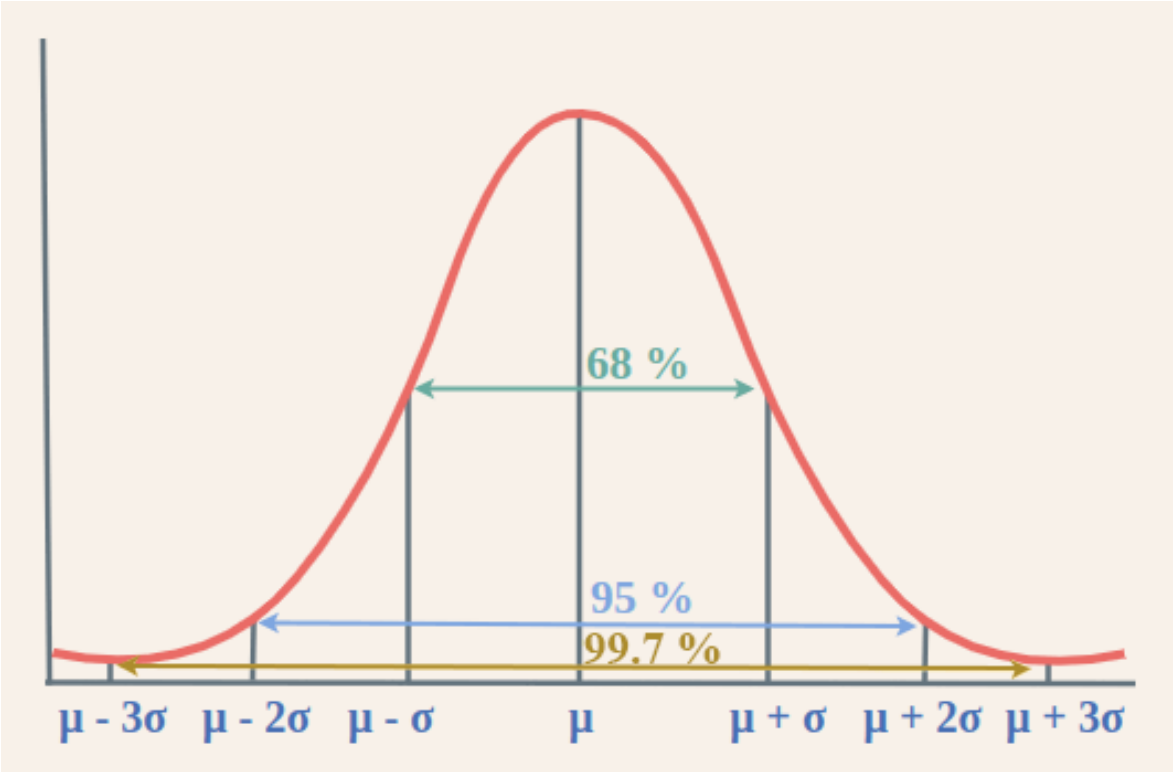
The Empirical Rule, also known as the **68-95-99.7** rule, quantifies the proportion of data falling within certain intervals around the mean in a normal distribution. It provides a quick way to estimate the spread of data without performing detailed calculations.

Normal distribution



A smaller standard deviation results in a narrower and taller bell curve, indicating that data points are clustered closely around the mean.
Conversely, a **larger standard deviation leads to a wider and shorter bell curve, suggesting that data points are more spread out from the mean.**

Normal distribution



The standard normal distribution has a mean(central value) $= 0$ and a standard deviation $= 1$.

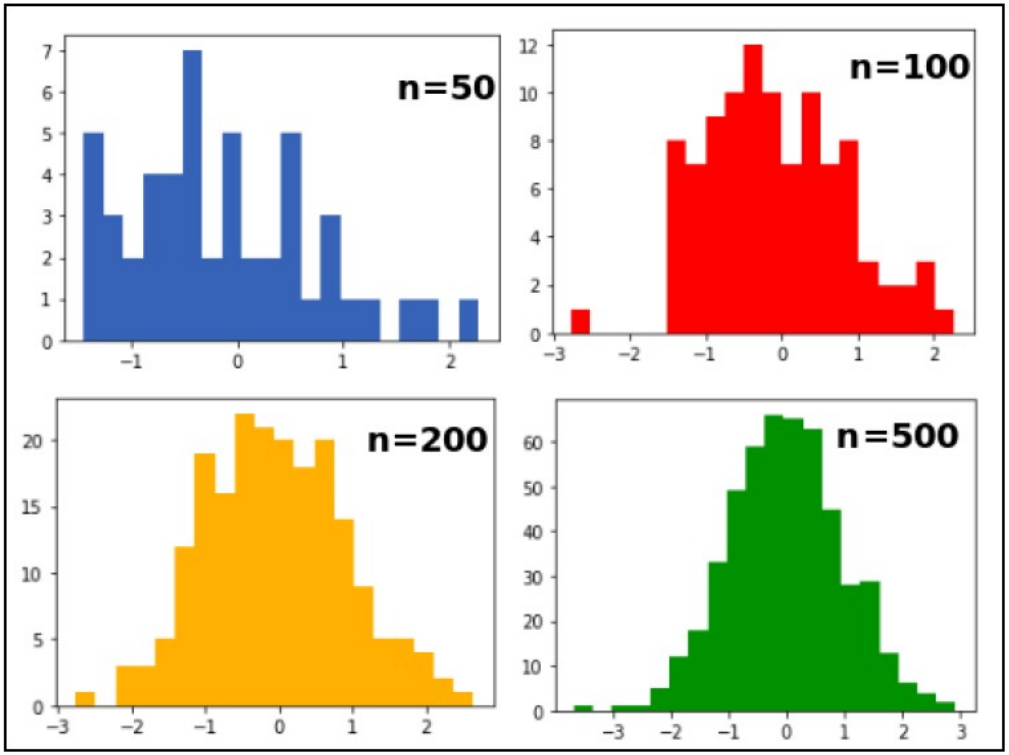
Machine Learning Methods that uses Gaussian Distribution

- In algorithms, such as **linear regression, logistic regression, and Gaussian mixture models,** it is often assumed that the observed data is generated from a Gaussian distribution. It simplifies the model and allows for efficient parameter estimation.
- In **Bayesian machine learning**, the Gaussian distribution is commonly used as a prior distribution over model parameters. This prior distribution reflects about the parameters before observing any data and is updated to a posterior distribution using Bayes' theorem.
- **Anomaly Detection** where the goal is to identify rare events or outliers in the data. Anomalies are detected based on the likelihood of the data under the Gaussian distribution.
- **Dimensionality Reduction** - Principal Component Analysis (PCA), it finds the directions of maximum variance in the data, which correspond to the principal components.
- **Kernel Methods** - Gaussian kernel is commonly used in kernelized machine learning algorithms, such as Support Vector Machines (SVMs).

Central limit theorem

- Data analysis methods involve hypothesis testing and deciding confidence intervals. All statistical tests assume that the population is normally distributed.
- **The central limit theorem is the core of hypothesis testing.**
- According to this theorem, **the sampling distribution approaches a normal distribution with an increase in the sample size**.
- Also, the mean of the sample gets closer to the population means and the standard deviation of the sample gets reduced.
- This theorem is essential for working with inferential statistics, helping data analysts figure out how samples can be useful in getting insights about the population.

Central limit theorem

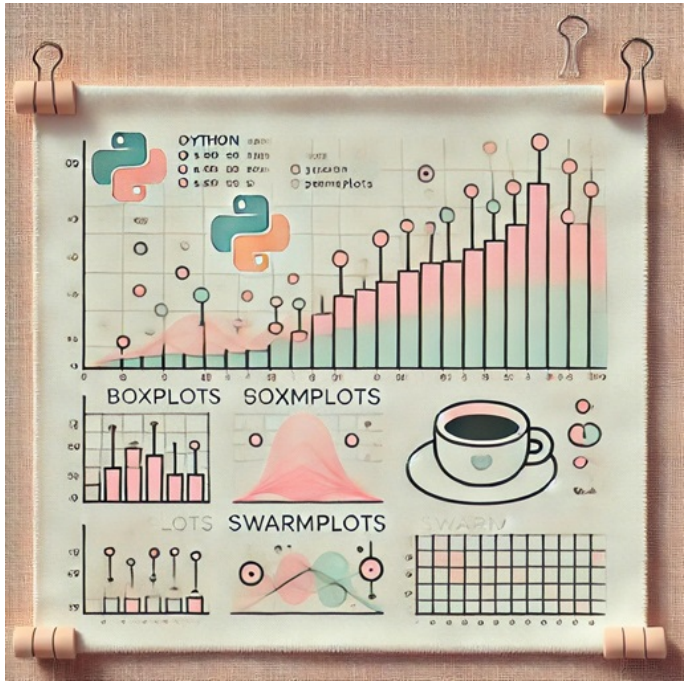


In the preceding diagram, you can see four histograms for different-different sample sizes 50, 100, 200, and 500. If you observe here, as the sample size increases, the histogram approaches a normal curve.

Collecting samples

- A sample is a small set of the population used for data analysis purposes.
- Sampling is a method or process of collecting sample data from various sources.
- **It is the most crucial part of data collection.** The success of an experiment depends upon how well the data is collected. If anything goes wrong with sampling, it will hugely affect the final interpretations.
- Also, it is impossible to collect data for the whole population.
- Sampling helps researchers to infer the population from the sample and reduces the survey cost and workload to collect and manage data.

Data Visualization



Data visualization is the initial move in the data analysis system toward easily understanding and communicating information.

It helps analysts to understand patterns, trends, outliers, distributions, and relationships.

Data visualization represents information and data in graphical form using visual elements such as charts, graphs, plots, and maps.

Data Visualization with Python

Python offers various libraries for data visualization, such as

- Matplotlib - was the first Python data visualization library, many other libraries are built on top of it or designed to work in tandem with it during analysis. Some libraries like pandas and Seaborn are "wrappers" over matplotlib. They allow you to access a number of matplotlib's methods with less code;
- Seaborn;
- Bokeh - is based on The Grammar of Graphics;

> The **Grammar of graphics** - any data graphic can be created by combining data with layers of plot components such as axes, tickmarks, gridlines, dots, bars, and lines.

- Plotly - interactive plots, also some charts you won't find in most libraries, like contour plots, dendrograms, and 3D charts;
- geoplotlib is a toolbox for creating maps and plotting geographical data;
- missingno - a visual summary of dataset;

Data Visualization - comparison

- A comparison visualization is used to illustrate the difference between two or more items at a given point in time or over a period of time.
- A commonly used comparison chart is the **boxplot.**
- Boxplots are typically used to compare the distribution of a continuous feature against the values of a categorical feature.

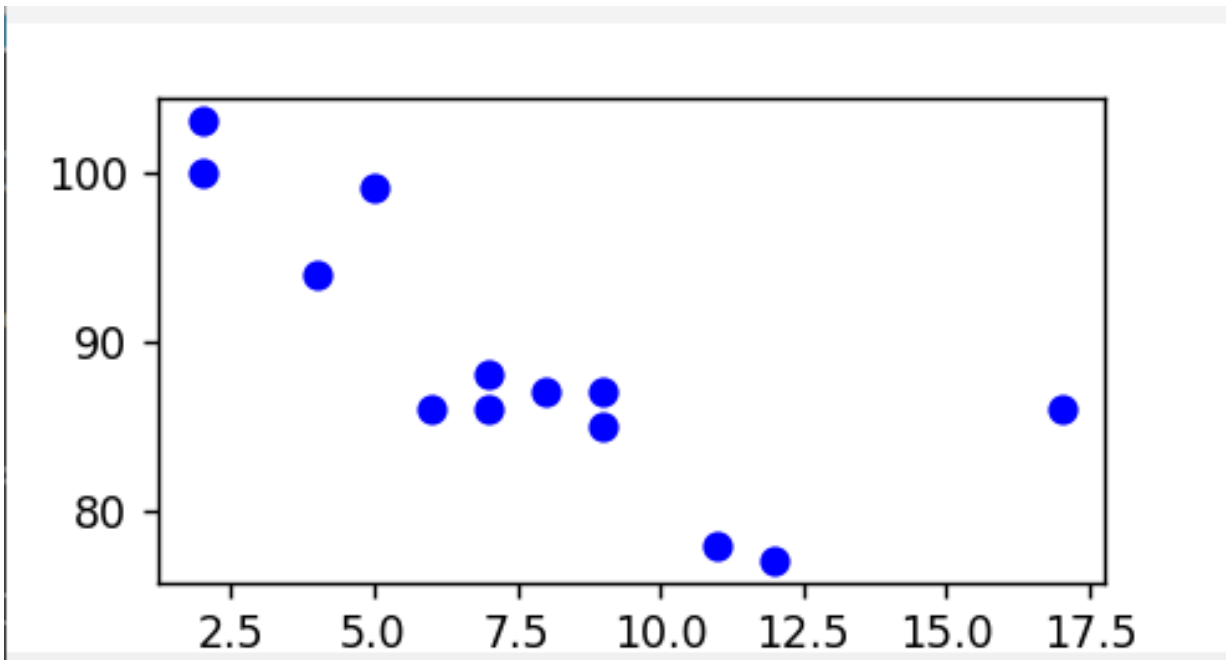
Boxplot visualizes the five summary statistics (minimum, first quartile, median, third quartile, and maximum) and all outlying points individually.

Visualizing numeric features – boxplots

A common visualization of the five-number summary is a **boxplot**, also known as a **box-and-whisker** plot. The boxplot displays the center and spread of a numeric variable in a format that allows you to quickly obtain a sense of its range and compare it to other features.

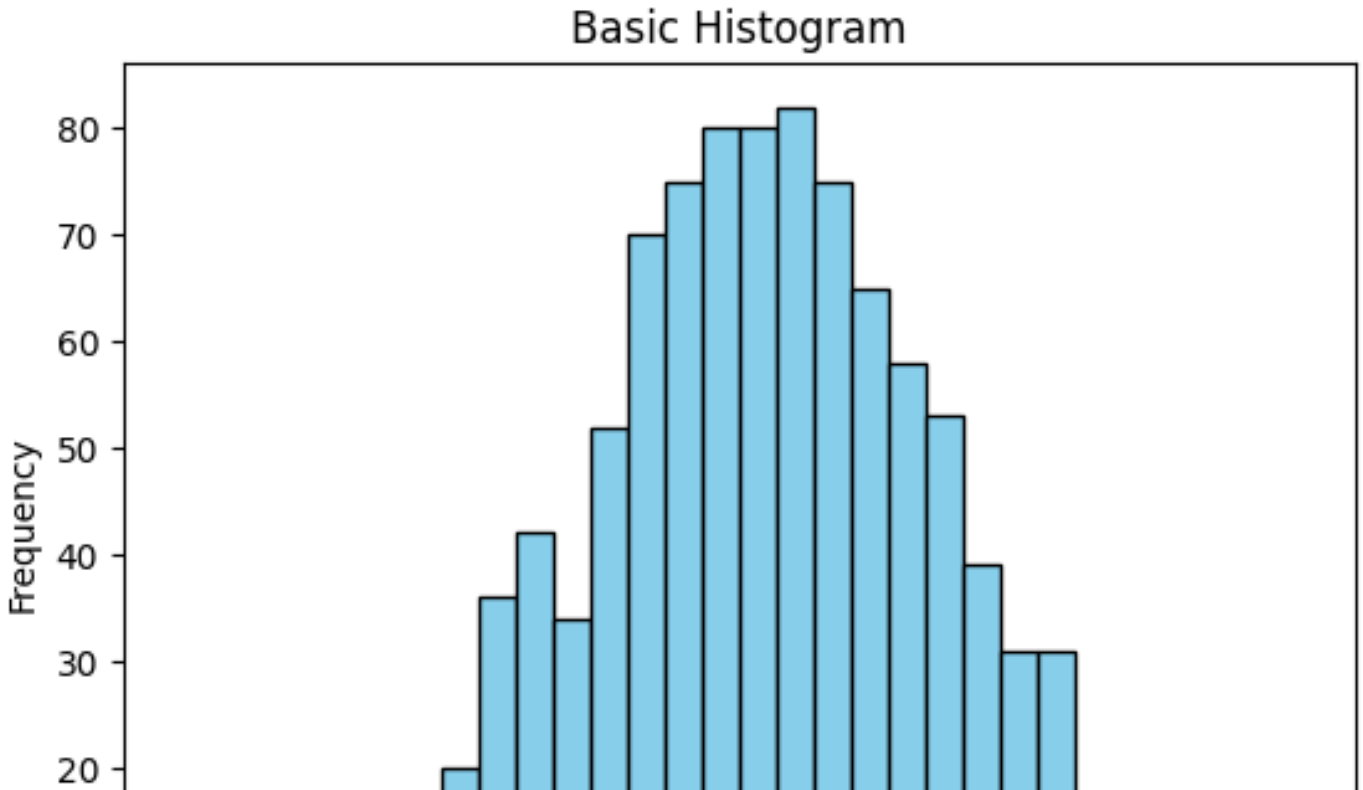
Data Visualization - relationship

- Relationship visualizations are used to illustrate the correlation between two or more variables.
- **Scatterplots** are one of the most commonly used relationship visualizations.
- These are typically for both continuous features.



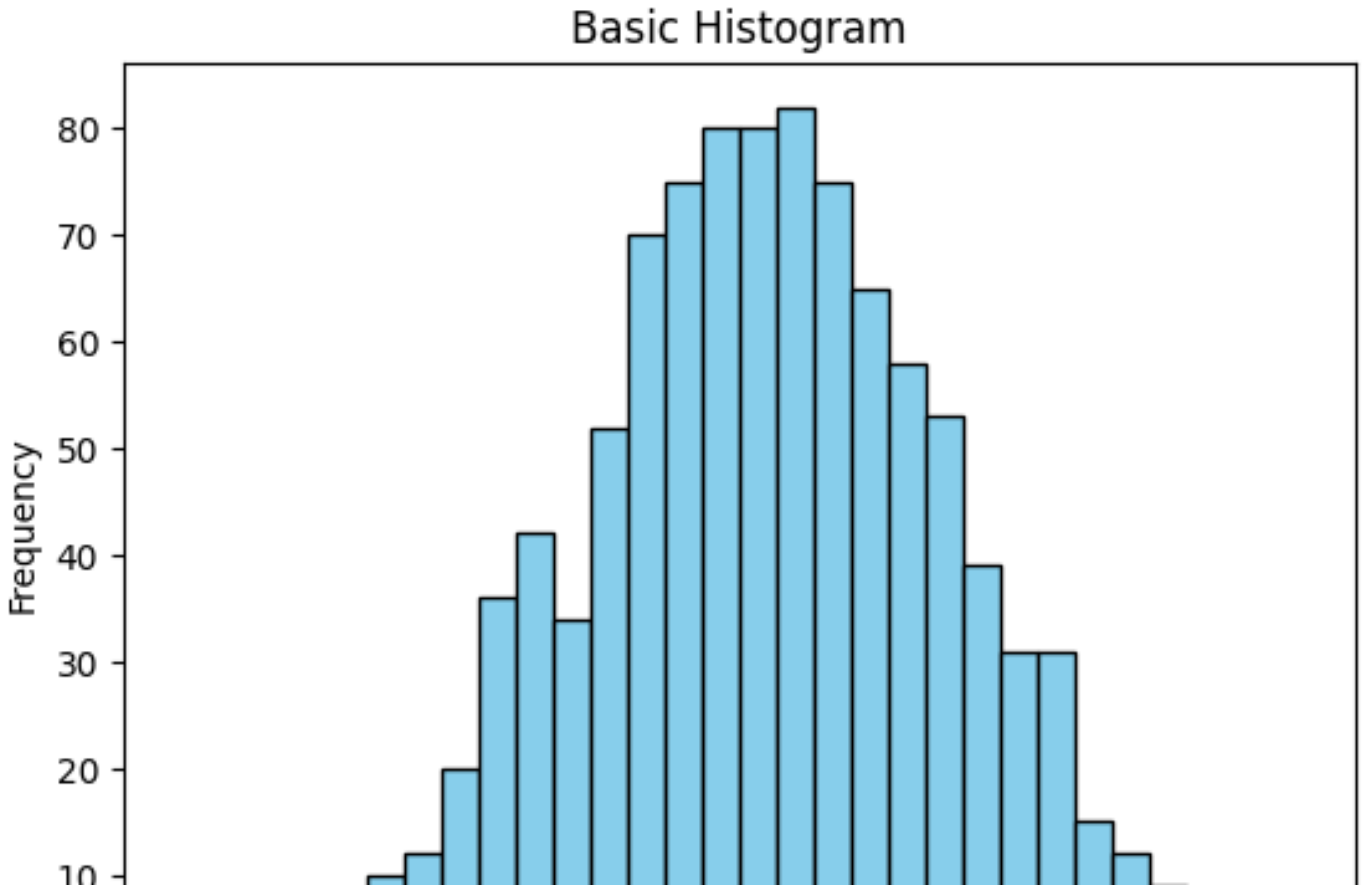
Data Visualization - distribution

- Distribution visualizations show the statistical distribution of the values of a feature.
- One of the most commonly used distribution visualizations is the histogram.
- With a histogram you can show the **spread** and **skewness** of data for a particular feature



Data Visualization - distribution

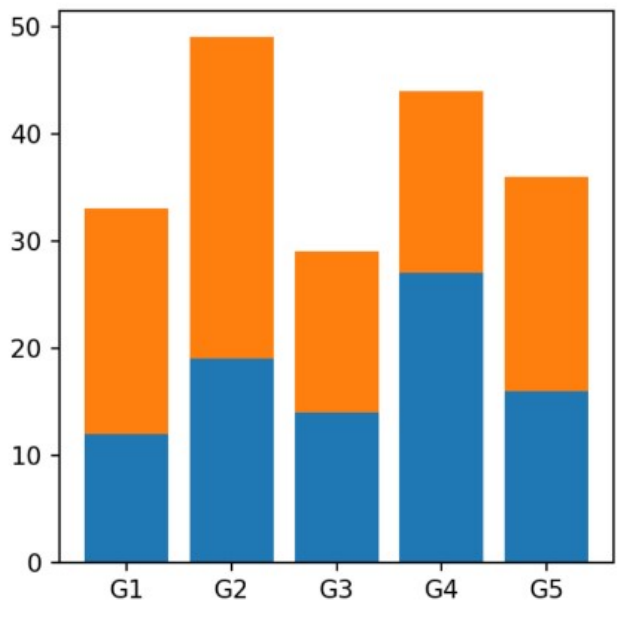
- Histogram divides the feature values into a predefined number of portions or bins that act as containers for values, with the same range.
- Histogram is composed of a series of bars with heights indicating the count, or frequency, of values falling within each of the equal-width bins partitioning the values.



Data Visualization - composition

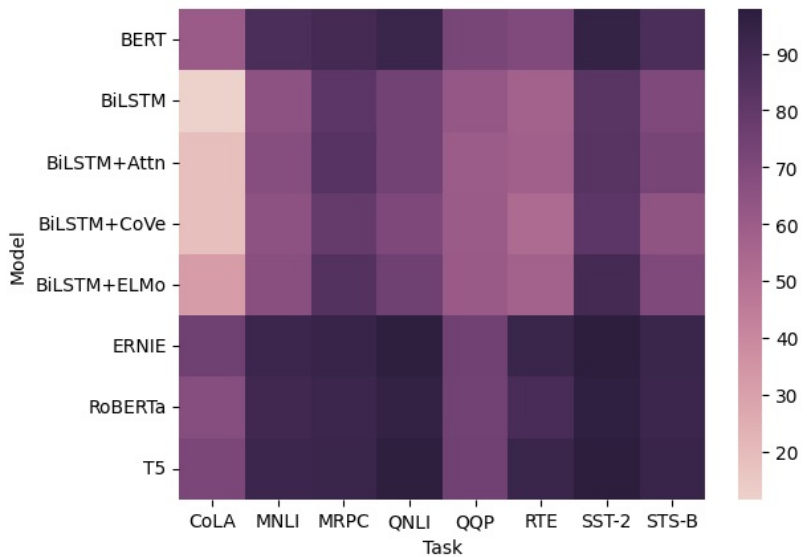
- A composition visualization shows the component makeup of the data.
- Stacked bar charts and pie charts are two of the most commonly used composition visualizations.
- With a stacked bar chart, you can show how a total value can be divided into parts or highlight the significance of each part relative to the total value.

```
import matplotlib.pyplot as plt

# Data
groups = ['G1', 'G2', 'G3', 'G4', 'G5']
values1 = [12, 19, 14, 27, 16]
values2 = [21, 30, 15, 17, 20]

fig, ax = plt.subplots()

# Stacked bar chart
ax.bar(groups, values1)
ax.bar(groups, values2, bottom = values1)

# plt.show()
```

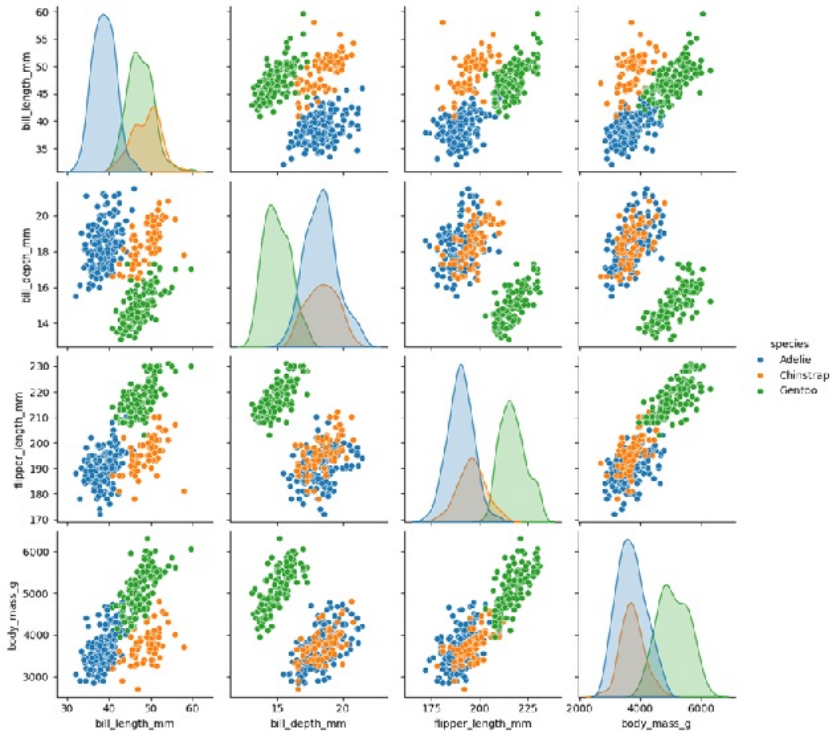


Data Visualization - heatmap

Heatmap is an intelligent, analytical tool that uses a color system to graphically represent various values.



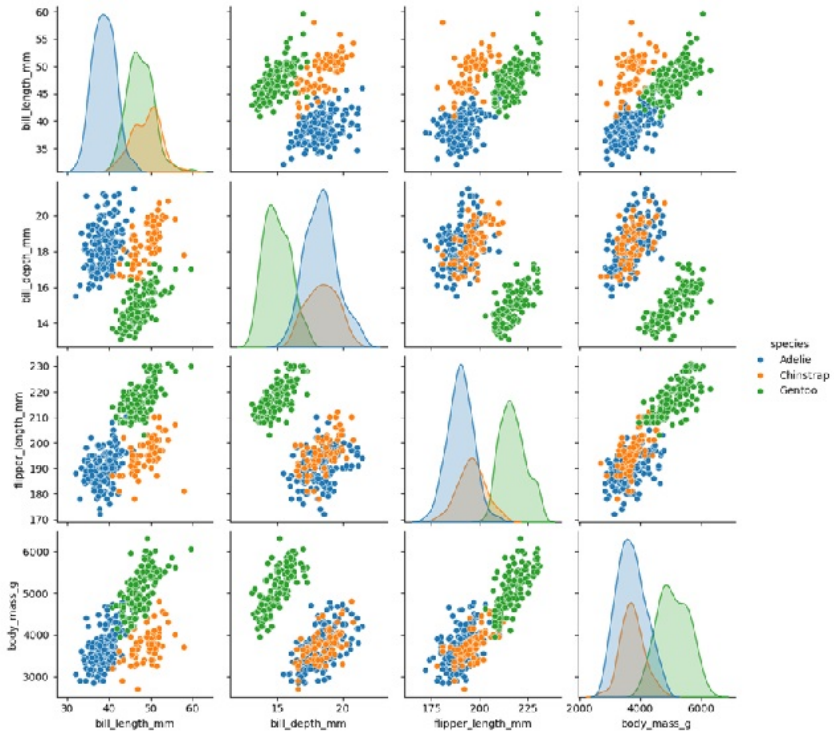
Data Visualization - pair plot



A pair plot, also known as a scatterplot matrix, is a matrix of graphs that enables the visualization of the relationship between each pair of variables in a dataset.

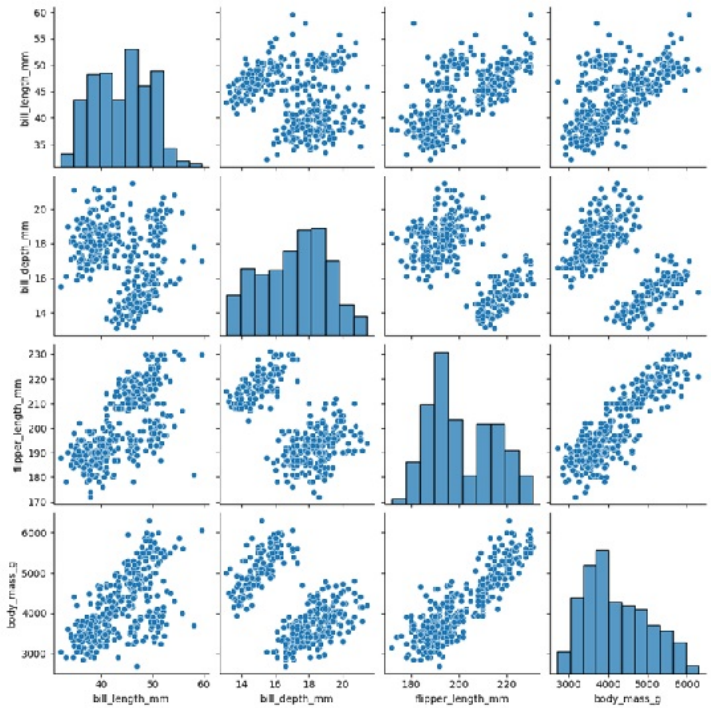
It combines both histogram and scatter plots, providing a unique overview of the dataset's distributions and correlations.

Data Visualization - pair plot



The primary purpose of a pair plot is to simplify the initial stages of data analysis by offering a comprehensive snapshot of potential relationships within the data.

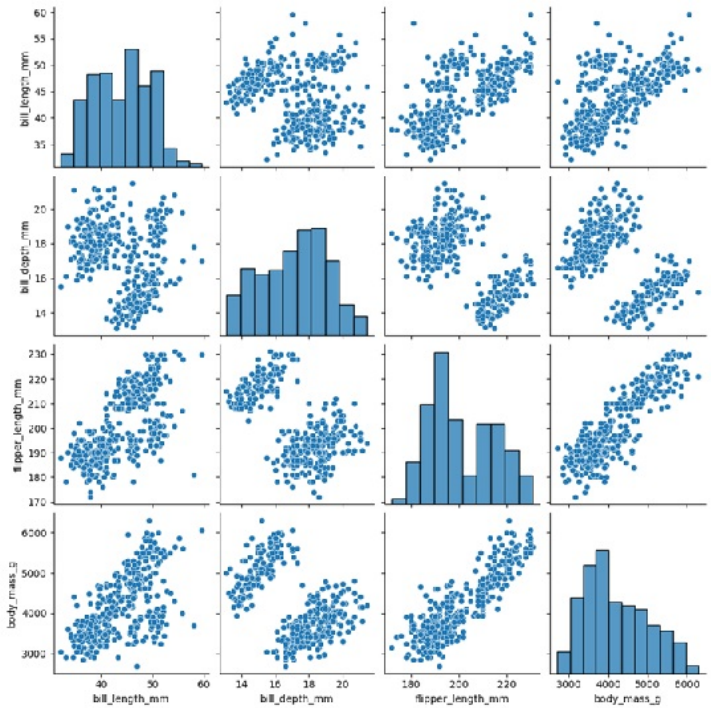
Data Visualization - pair plot



Pair plots enable data scientists to:

- visualize distributions - understand the distribution of single variables.
- identify relationships - observe linear or nonlinear relationships between variables.
- detect anomalies - spot outliers that may indicate errors or unique insights.

Data Visualization - pair plot



Pair plots enable data scientists to:

- find trends - linear or nonlinear relationships that suggest predictability.
- find clusters - groups of data points that share similar characteristics, hinting at subpopulations within the dataset.
- find correlations - the strength and direction of relationships between variables.

Data Visualization

Thank you for your attention!!!