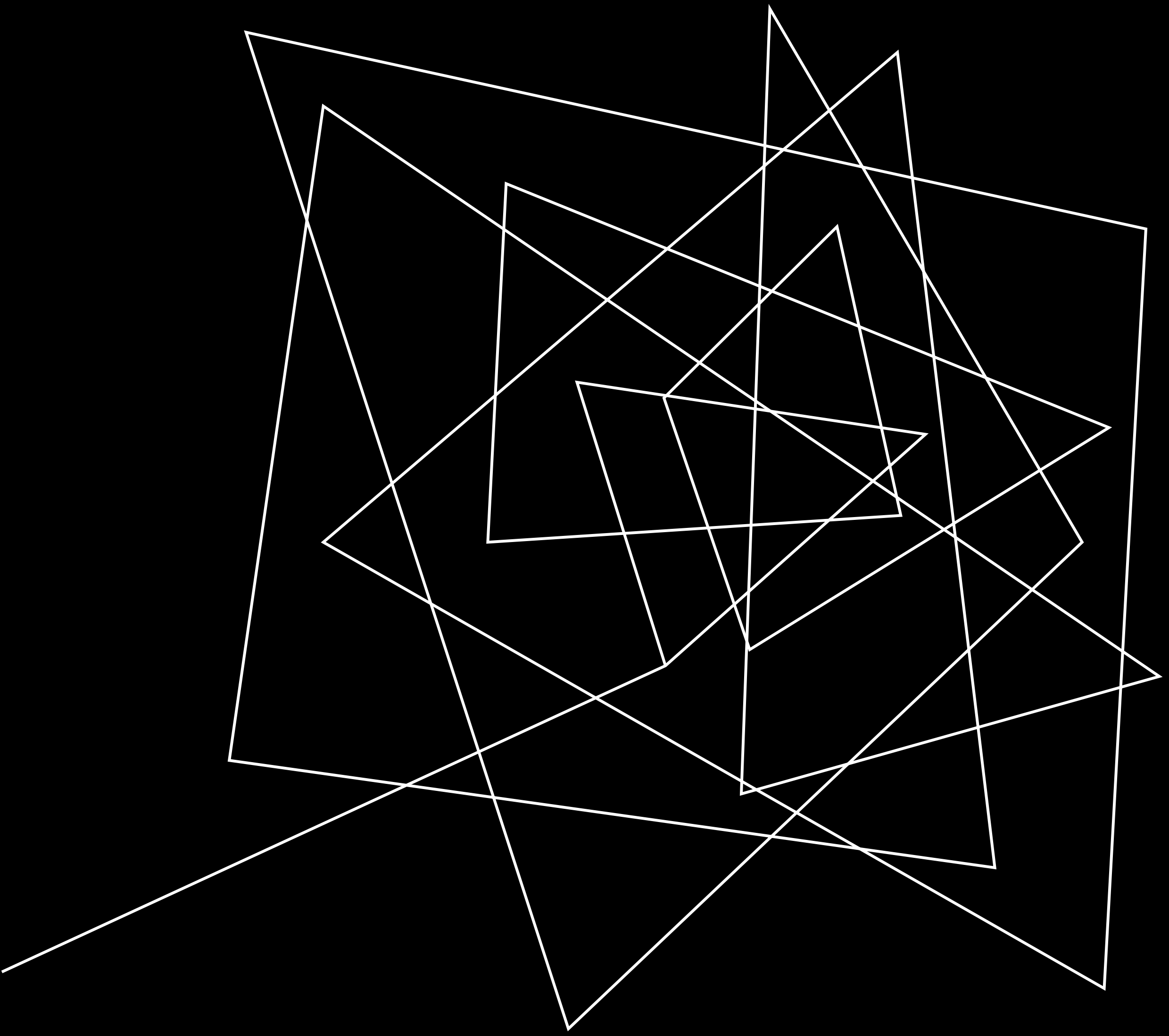
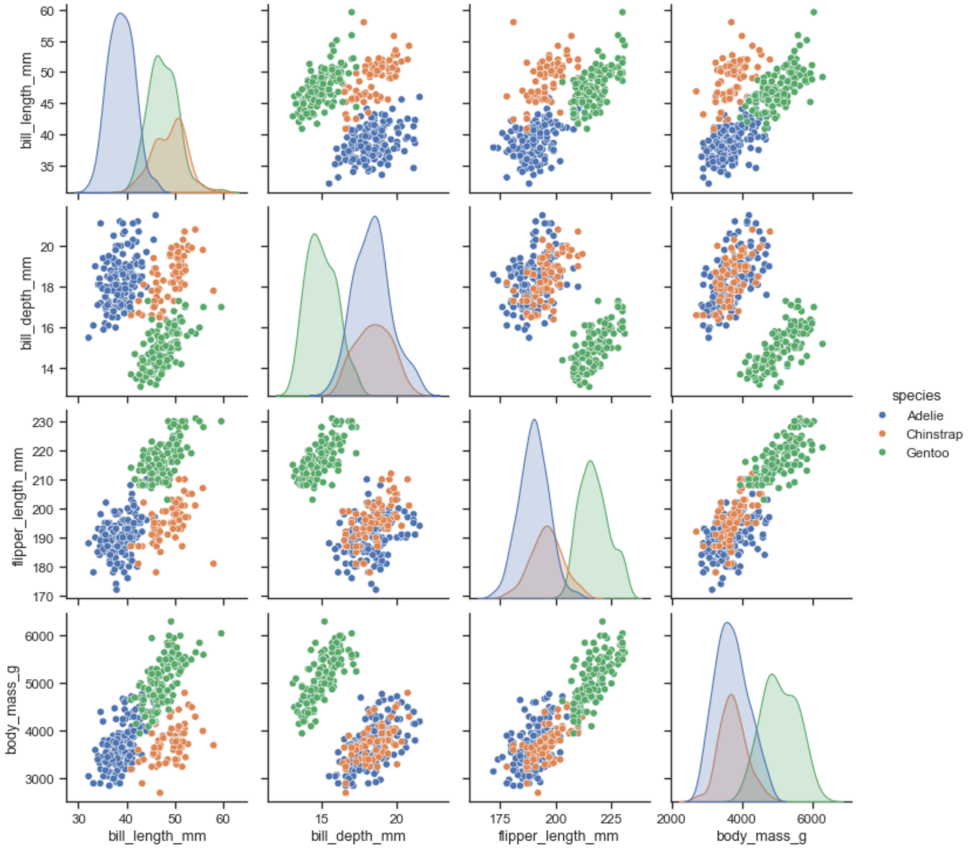


SESSION 2: DATA VISUALIZATION AND EDA

Understand the importance of data visualization and learn exploratory data analysis techniques.

Data Visualization and EDA

Data visualization and **exploratory data analysis (EDA)** play crucial roles in the field of data science, providing essential tools for understanding and extracting insights from complex datasets.



Data Visualization and EDA

Data Visualization:

Data visualization involves representing data graphically to gain insights. Visualization is a powerful tool for communication and exploration. Common types of visualizations include:

- **Scatter plots:** Show the relationship between two continuous variables.
- **Histograms:** Display the distribution of a single variable.
- **Box plots:** Summarize the distribution of a variable and identify outliers.
- **Line charts:** Show trends or patterns over time.
- **Heatmaps:** Visualize patterns in a matrix of data.
- **Bar charts:** Display the distribution of a categorical variable.
- **Pie charts:** Show the proportion of different categories in a whole.

Data Visualization and EDA:
<Importance of Data Visualization>

1. Understanding Patterns and Trends

Data visualization allows for the graphical representation of data, making it easier to identify patterns, trends, and relationships. Visualizations can reveal insights that may be challenging to discern from raw data alone.

2. Communication of Complex Information

Visualizations provide an effective means of communicating complex information to a diverse audience. Charts, graphs, and dashboards can convey insights quickly and clearly, enabling stakeholders to make informed decisions.

3. Identification of Outliers and Anomalies

Visualizations help in the identification of outliers or anomalies in the data. Anomalies may not be apparent in tabular data but can be easily spotted in visual representations, aiding in data cleaning and preprocessing.

Data Visualization and EDA:
<Importance of Data Visualization>

4. Storytelling with Data

Visualization facilitates storytelling with data, allowing data scientists to create compelling narratives that highlight key findings and support data-driven conclusions. This is particularly valuable in conveying the significance of the results to non-technical stakeholders.

5. Comparisons and Benchmarking

Visualizations enable effective comparisons between different variables or datasets. Whether it's comparing performance over time or benchmarking against a standard, visual representations enhance the understanding of data relationships.

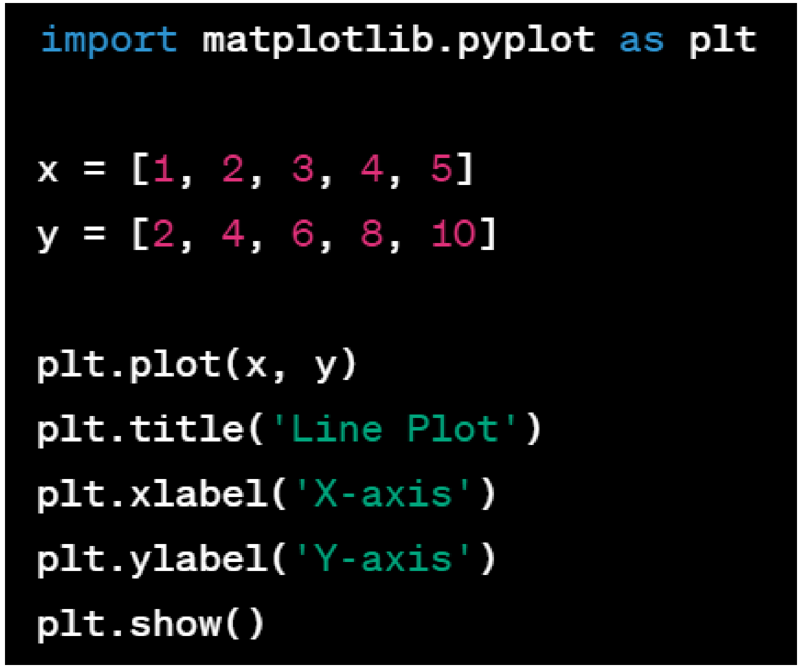
6. Enhanced Decision-Making

Decision-makers often rely on visualizations to grasp the implications of various choices. Visual data exploration supports evidence-based decision-making by providing a clear and intuitive understanding of the data.

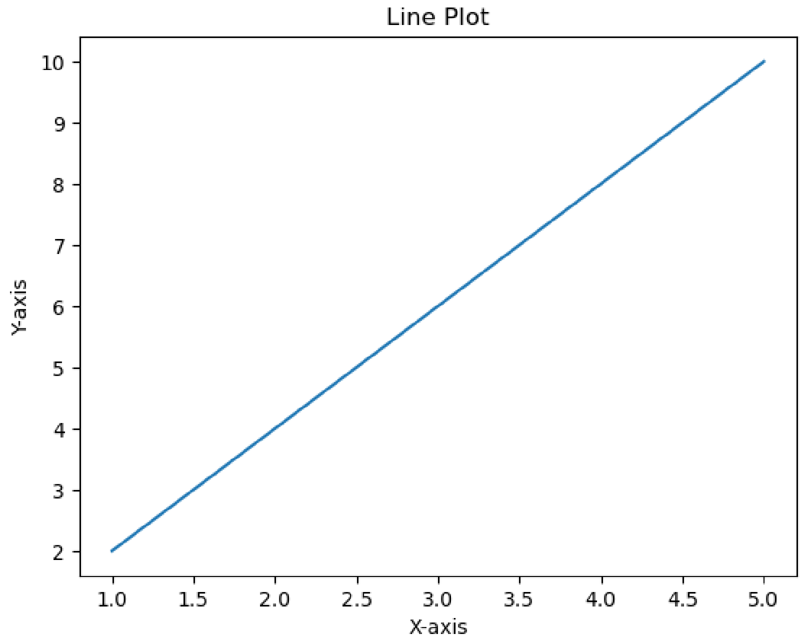
Data Visualization: Matplotlib and Seaborn Libraries

- **Matplotlib** is a library in Python that enables users to generate visualizations like histograms, scatter plots, bar charts, pie charts and much more.
- **Seaborn** is a visualization library that is built on top of Matplotlib. It provides data visualizations that are typically more aesthetic and statistically sophisticated.
- **Matplotlib** and **Seaborn** are two popular Python libraries for data visualization, each serving different purposes while complementing each other.

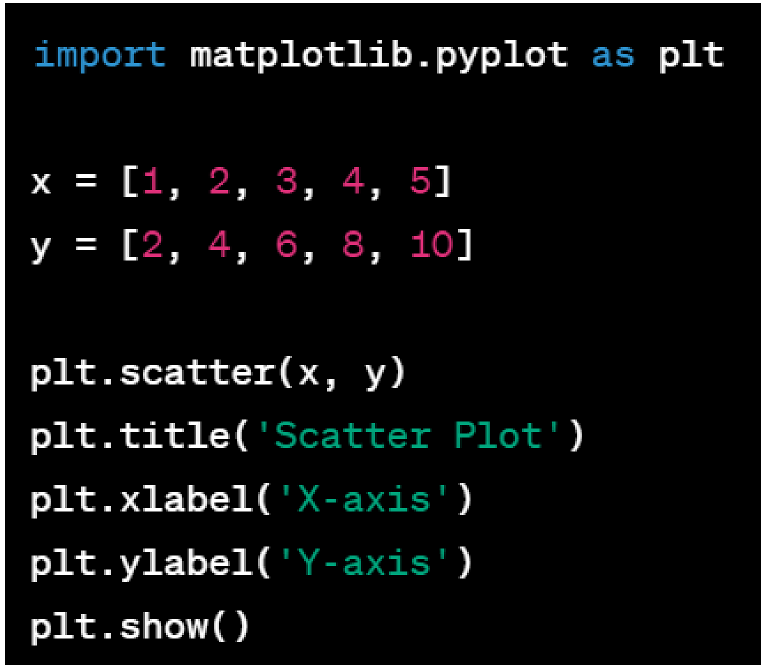
Data Visualization: Matplotlib Plots <Line Plot>



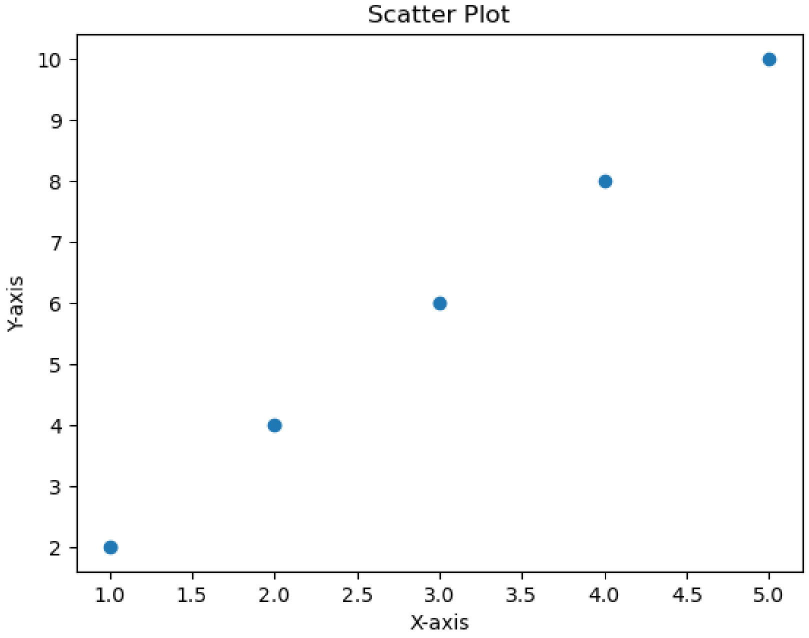
```python
import matplotlib.pyplot as plt

x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

plt.plot(x, y)
plt.title('Line Plot')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.show()
```



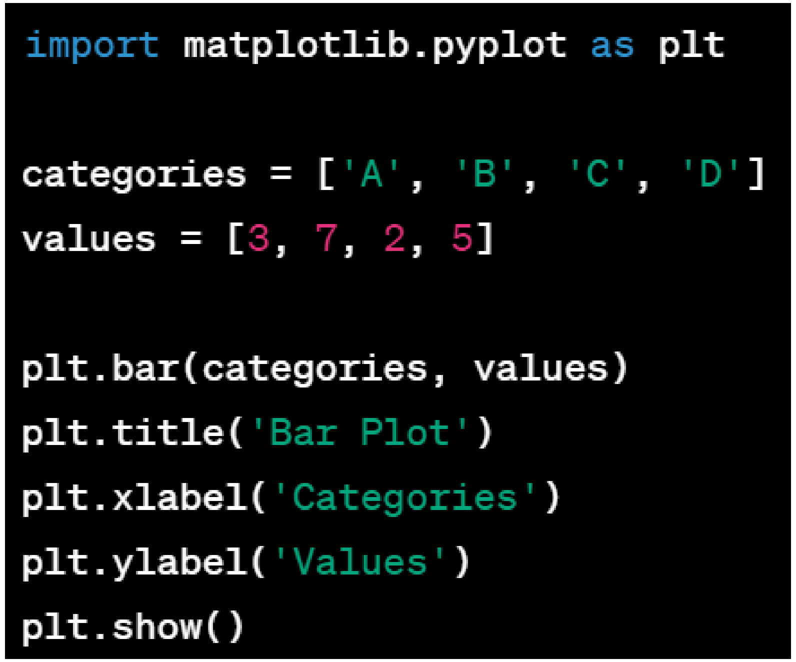
Data Visualization: Matplotlib Plots <Scatter Plot>



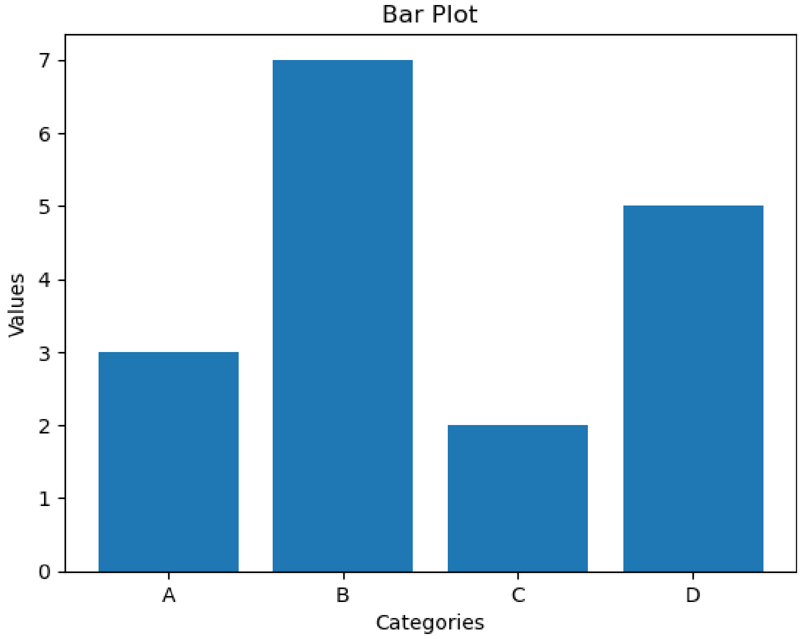
```python
import matplotlib.pyplot as plt

x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

plt.scatter(x, y)
plt.title('Scatter Plot')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.show()
```



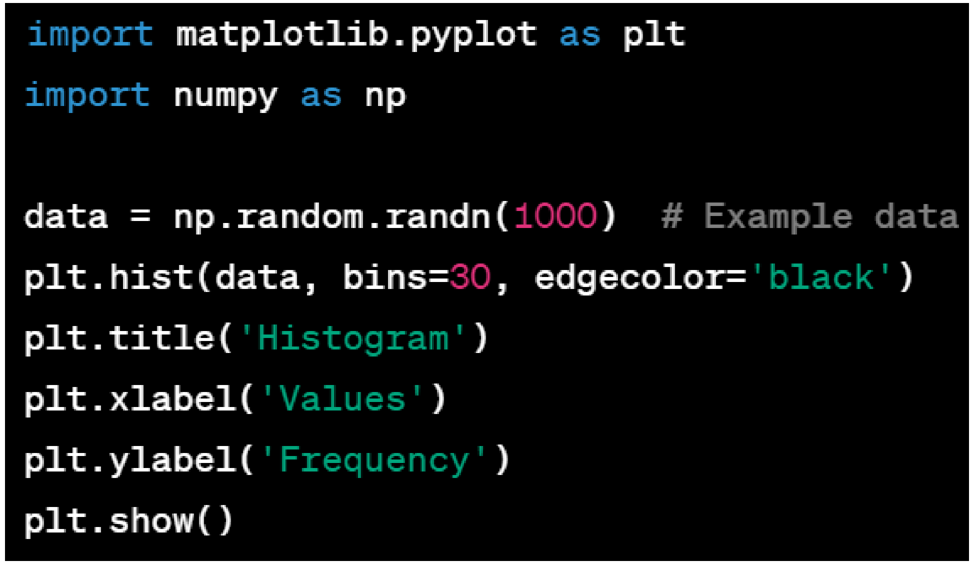
Data Visualization: Matplotlib Plots
<Bar Plot>



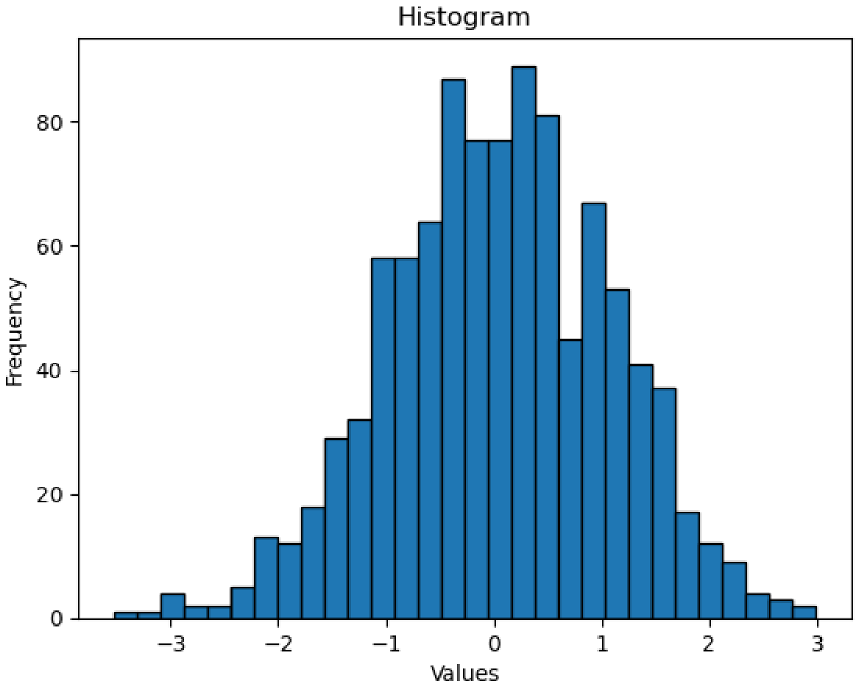
```python
import matplotlib.pyplot as plt

categories = ['A', 'B', 'C', 'D']
values = [3, 7, 2, 5]

plt.bar(categories, values)
plt.title('Bar Plot')
plt.xlabel('Categories')
plt.ylabel('Values')
plt.show()
```



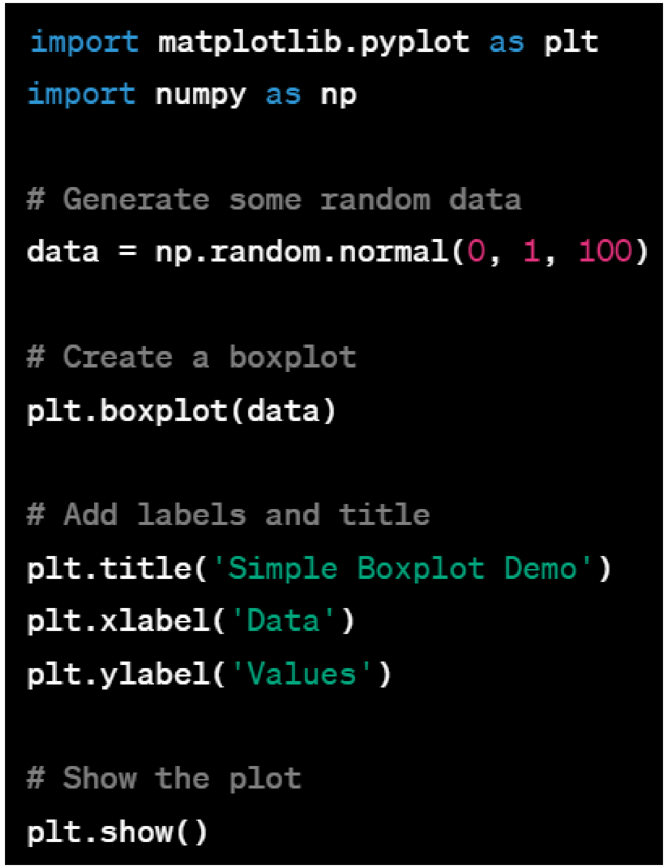
Data Visualization: Matplotlib Plots <Histogram>



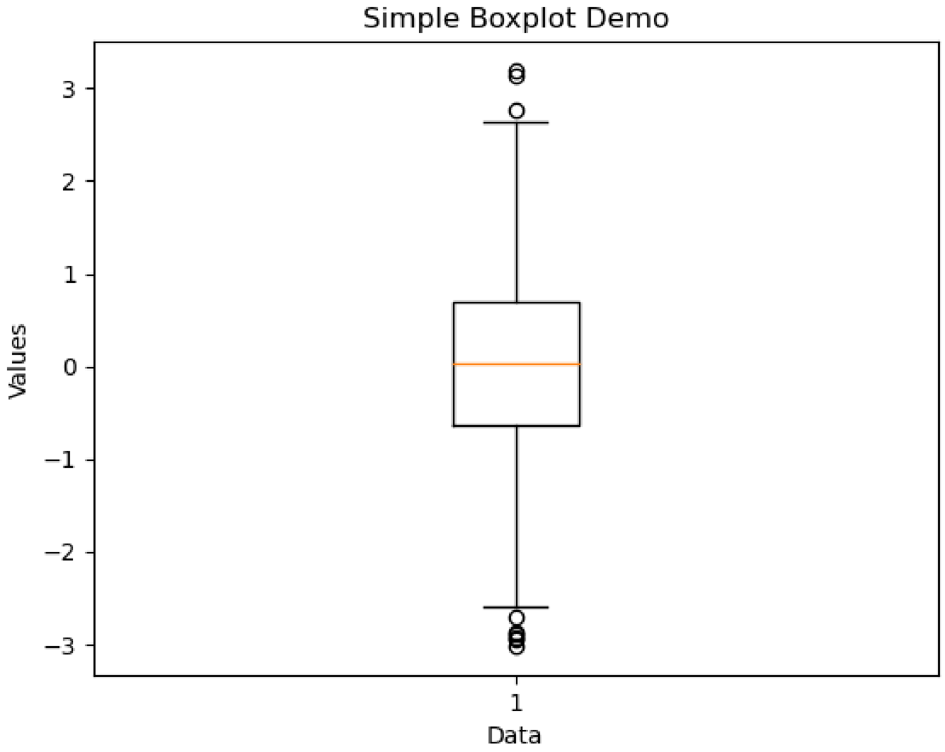
```
import matplotlib.pyplot as plt
import numpy as np

data = np.random.randn(1000)  # Example data
plt.hist(data, bins=30, edgecolor='black')
plt.title('Histogram')
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.show()
```



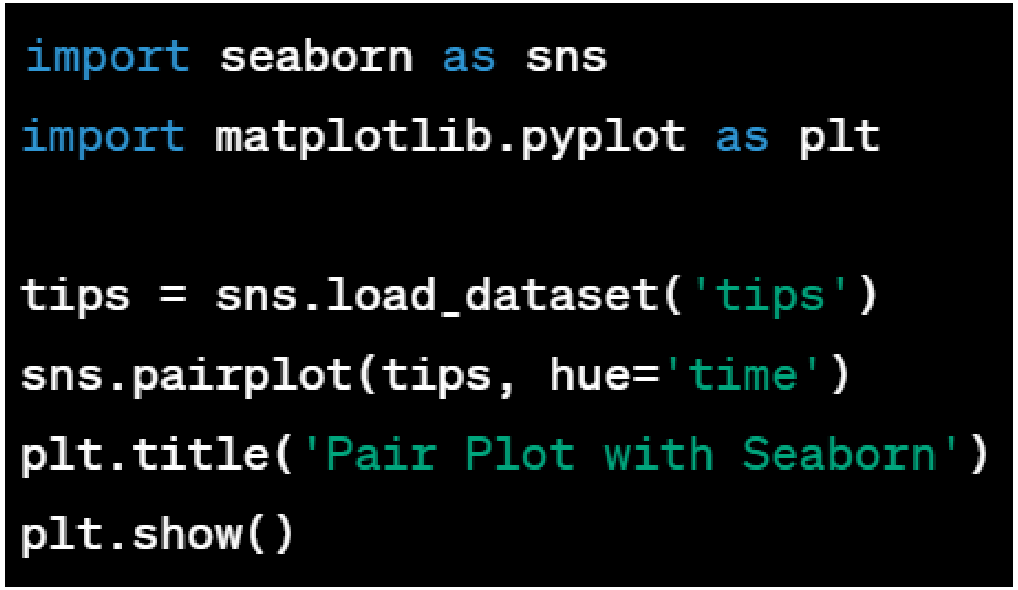
Data Visualization: Matplotlib Plots <Box Plot>



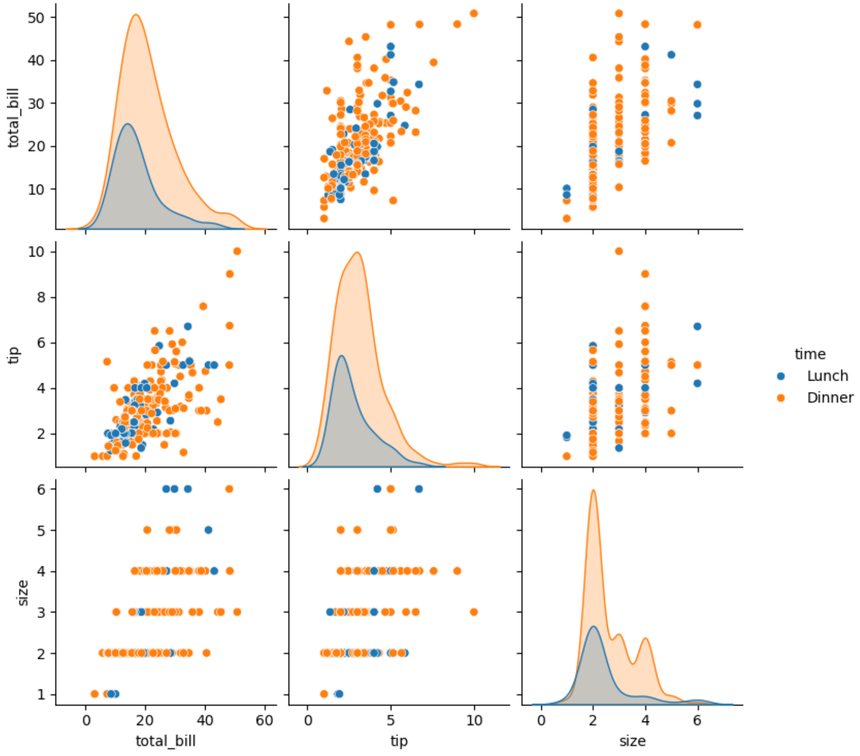
```
import matplotlib.pyplot as plt
import numpy as np

# Generate some random data
data = np.random.normal(0, 1, 100)

# Create a boxplot
plt.boxplot(data)

# Add labels and title
plt.title('Simple Boxplot Demo')
plt.xlabel('Data')
plt.ylabel('Values')

# Show the plot
plt.show()
```



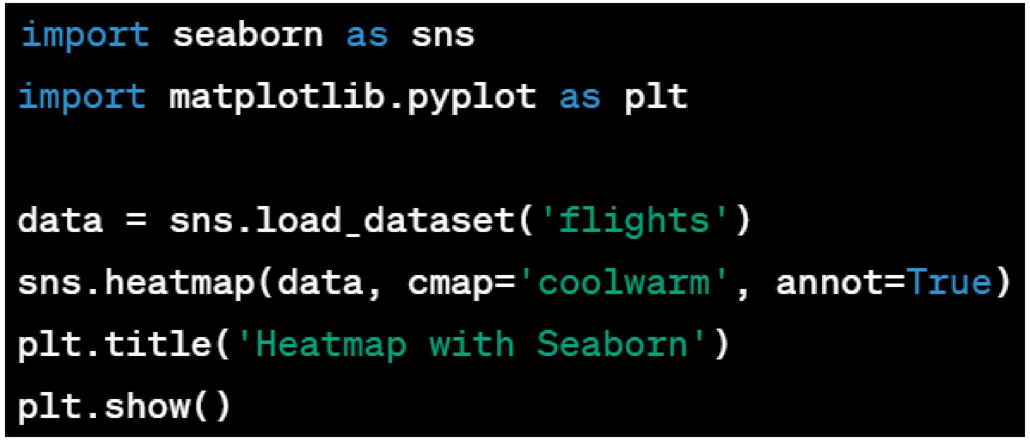
Data Visualization: Seaborn Plots <Pair Plot>



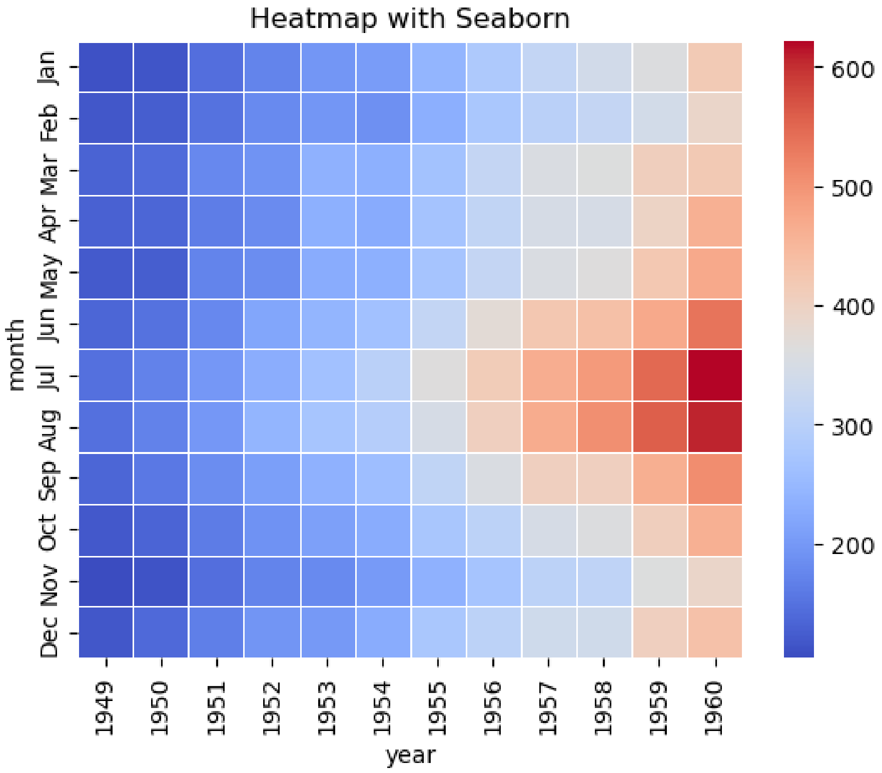
```
import seaborn as sns
import matplotlib.pyplot as plt

tips = sns.load_dataset('tips')
sns.pairplot(tips, hue='time')
plt.title('Pair Plot with Seaborn')
plt.show()
```



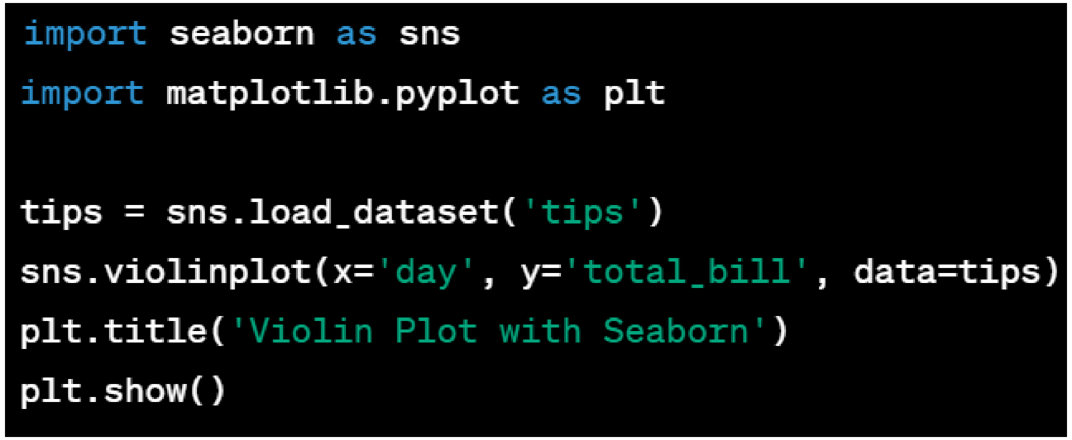
Data Visualization: Seaborn Plots <Heatmap>



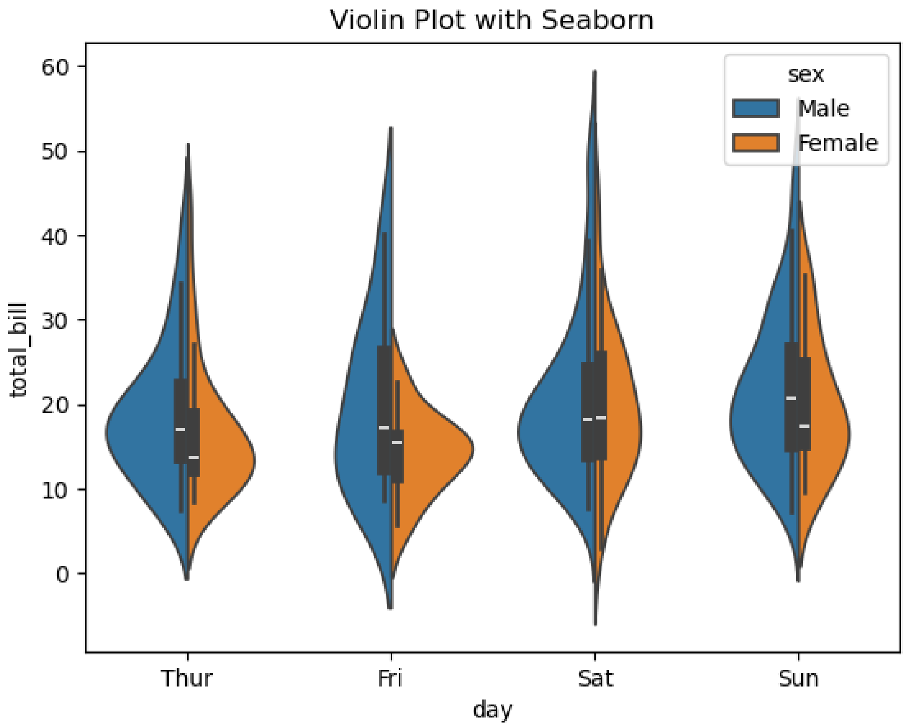
```
import seaborn as sns
import matplotlib.pyplot as plt

data = sns.load_dataset('flights')
sns.heatmap(data, cmap='coolwarm', annot=True)
plt.title('Heatmap with Seaborn')
plt.show()
```



Data Visualization: Seaborn Plots <Violin Plot>

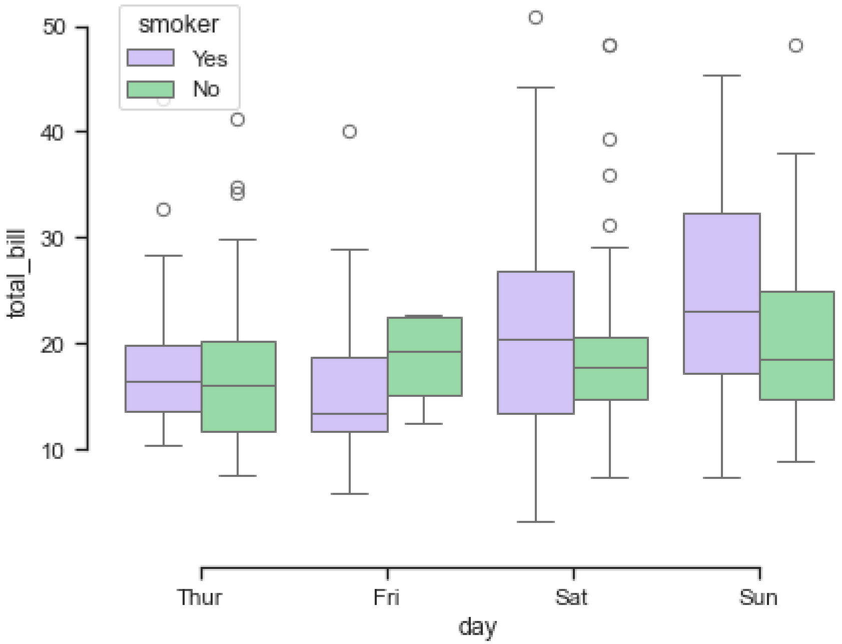


```
import seaborn as sns
import matplotlib.pyplot as plt

tips = sns.load_dataset('tips')
sns.violinplot(x='day', y='total_bill', data=tips)
plt.title('Violin Plot with Seaborn')
plt.show()
```



Data Visualization: Matplotlib and Seaborn Plots

- For more Matplotlib plots:
  - https://matplotlib.org/
- For more Seaborn plots:
  - https://seaborn.pydata.org/

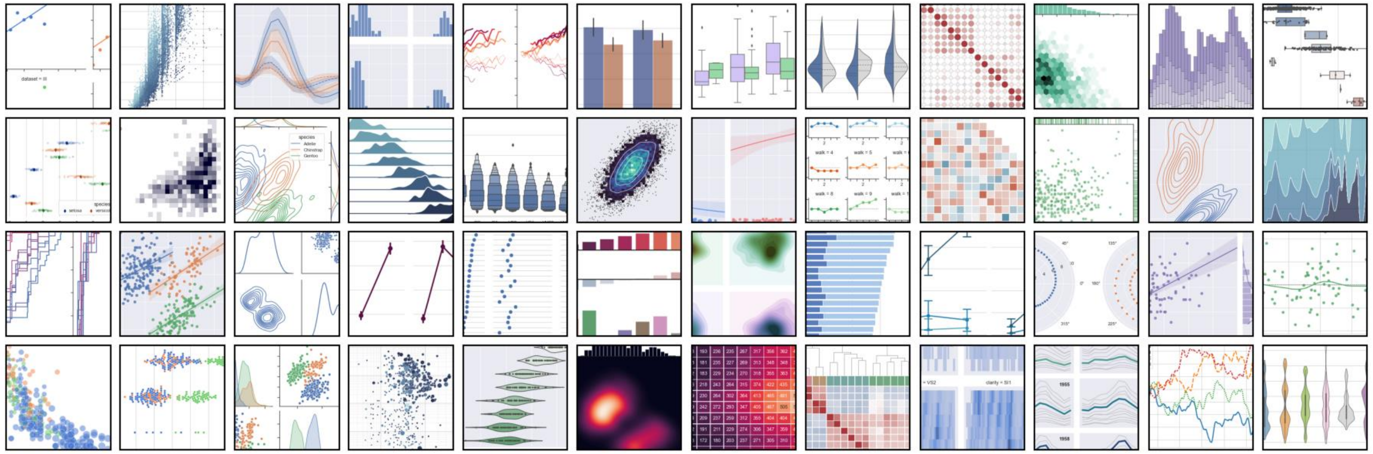


Data Visualization: Matplotlib and Seaborn Plots <DEMO>

Demo: Session 2 – Data Visualization and EDA

Data Visualization and EDA

Exploratory Data Analysis (**EDA**) is a **crucial** step in the data analysis process, allowing you to understand the main characteristics of your dataset. EDA involves examining and visualizing the data to identify patterns, relationships, and potential outliers. Univariate, bivariate, and multivariate analyses are key components of EDA, and they can be performed using both graphical and non-graphical methods.



Data Visualization and EDA: <Importance of EDA>

1. Data Understanding and Familiarity

EDA is a critical initial step in the data science workflow. It involves summarizing and familiarizing oneself with the main characteristics of the dataset, including its structure, distributions, and key statistics.

2. Data Cleaning and Preprocessing

Through EDA, data scientists can identify and address issues such as missing values, outliers, and inconsistencies in the dataset. This is crucial for ensuring the quality and reliability of the data before moving on to more advanced analyses.

3. Feature Engineering and Selection

EDA aids in the identification of relevant features for modeling. Data scientists can uncover patterns and relationships that inform decisions about which features to include or exclude in the model-building process.

Data Visualization and EDA: <Importance of EDA>

4. Hypothesis Generation

EDA helps in generating hypotheses about the relationships within the data. By visualizing and exploring data distributions and correlations, data scientists can form initial hypotheses that guide further analysis.

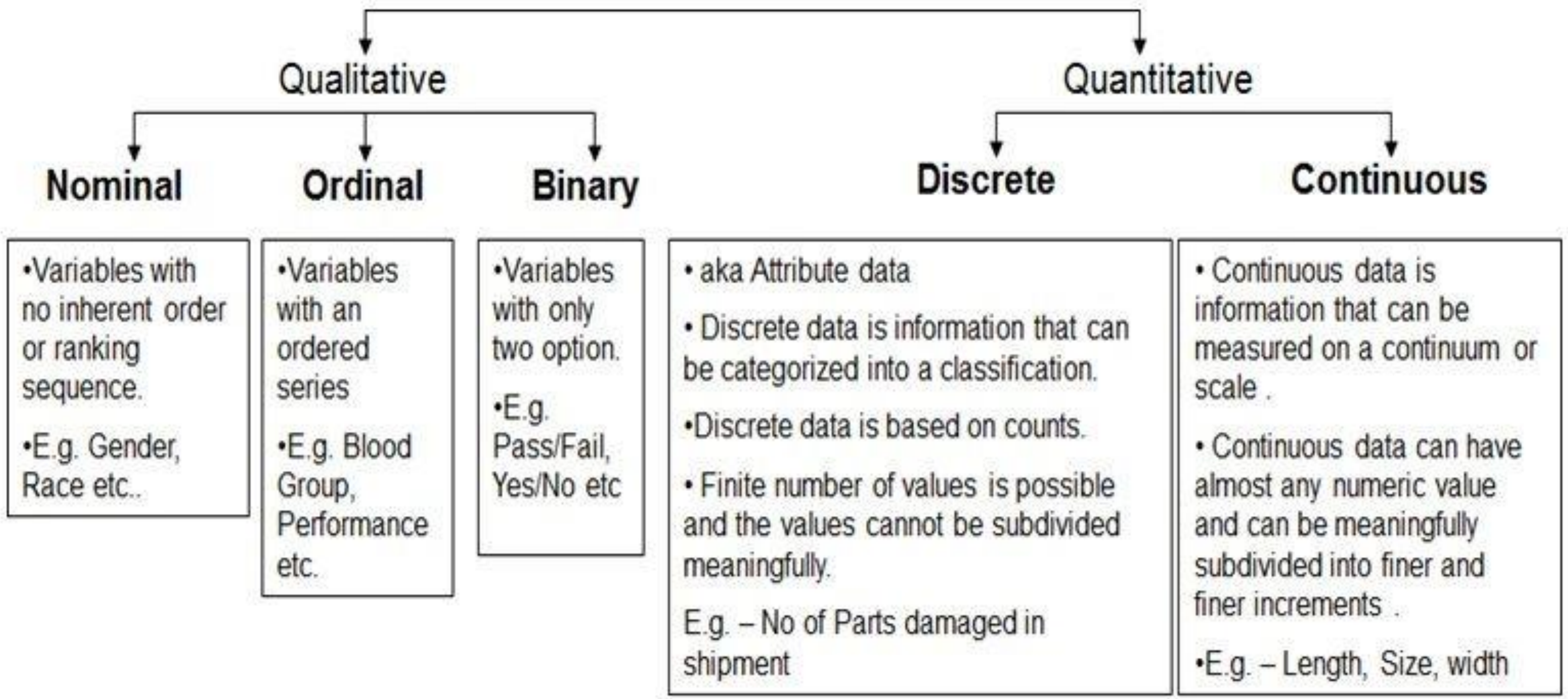
5. Model Assumptions and Validations

EDA assists in validating assumptions made during the modeling process. Understanding the underlying assumptions of statistical models and verifying their applicability to the dataset ensures the reliability of the analysis.

6. Communication with Stakeholders

EDA provides insights that can be effectively communicated to stakeholders. Whether it's presenting initial findings or explaining the nuances of the dataset, EDA supports clear and transparent communication.

Data Visualization and EDA: <Numerical vs. Categorical Data>



Qualitative

Quantitative

Nominal

- Variables with no inherent order or ranking sequence.
- E.g. Gender, Race etc..

Ordinal

- Variables with an ordered series
- E.g. Blood Group, Performance etc.

Binary

- Variables with only two option.
- E.g. Pass/Fail, Yes/No etc

Discrete

- aka Attribute data
- Discrete data is information that can be categorized into a classification.
- Discrete data is based on counts.
- Finite number of values is possible and the values cannot be subdivided meaningfully.

E.g. – No of Parts damaged in shipment

Continuous

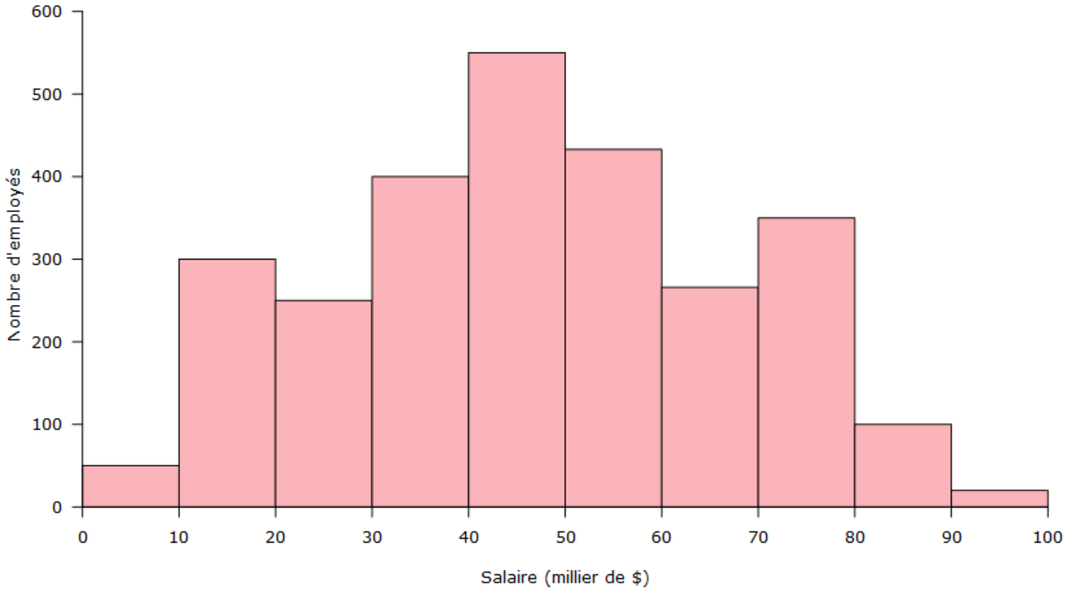
- Continuous data is information that can be measured on a continuum or scale .
- Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
- E.g. – Length, Size, width

Data Visualization and EDA: <Numerical vs. Categorical Data>

In **data science**, data can be broadly categorized into **numerical** and **categorical** types based on their nature and characteristics:

- **Numerical Data:**
  - **Definition:** Numerical data consists of numbers and represents measurable quantities. It can be further categorized into discrete and continuous data.
  - **Examples:**
    - **Discrete Numerical Data:** Count of items, number of people, etc.
    - **Continuous Numerical Data:** Height, weight, temperature, etc.
  - **Analysis Techniques:**
    - **Descriptive statistics:** Mean, median, mode, range, etc.
    - **Inferential statistics:** Regression analysis, hypothesis testing, etc.
    - **Visualization:** Histograms, box plots, scatter plots, etc.

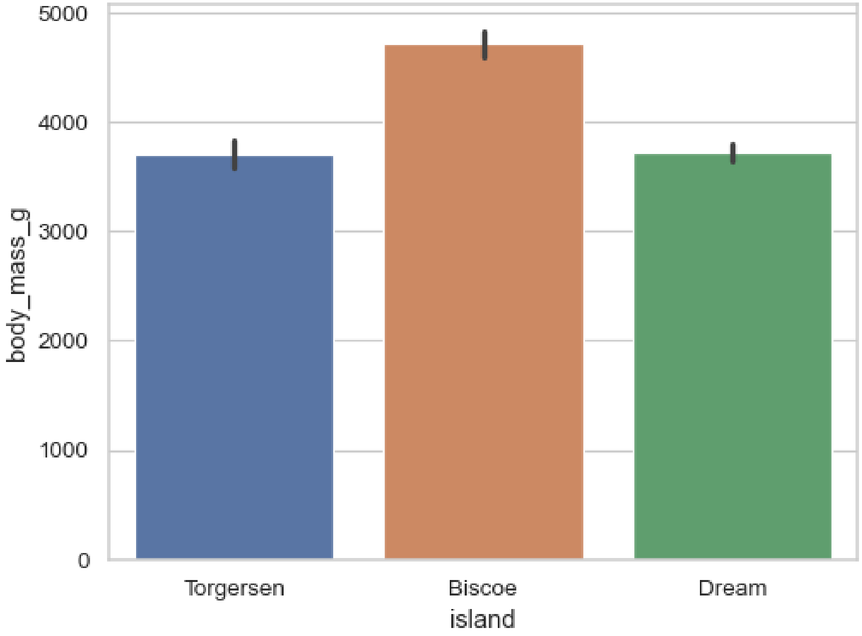
Graphique 5.7.1
Distribution des salaires des employés de la societe ABC



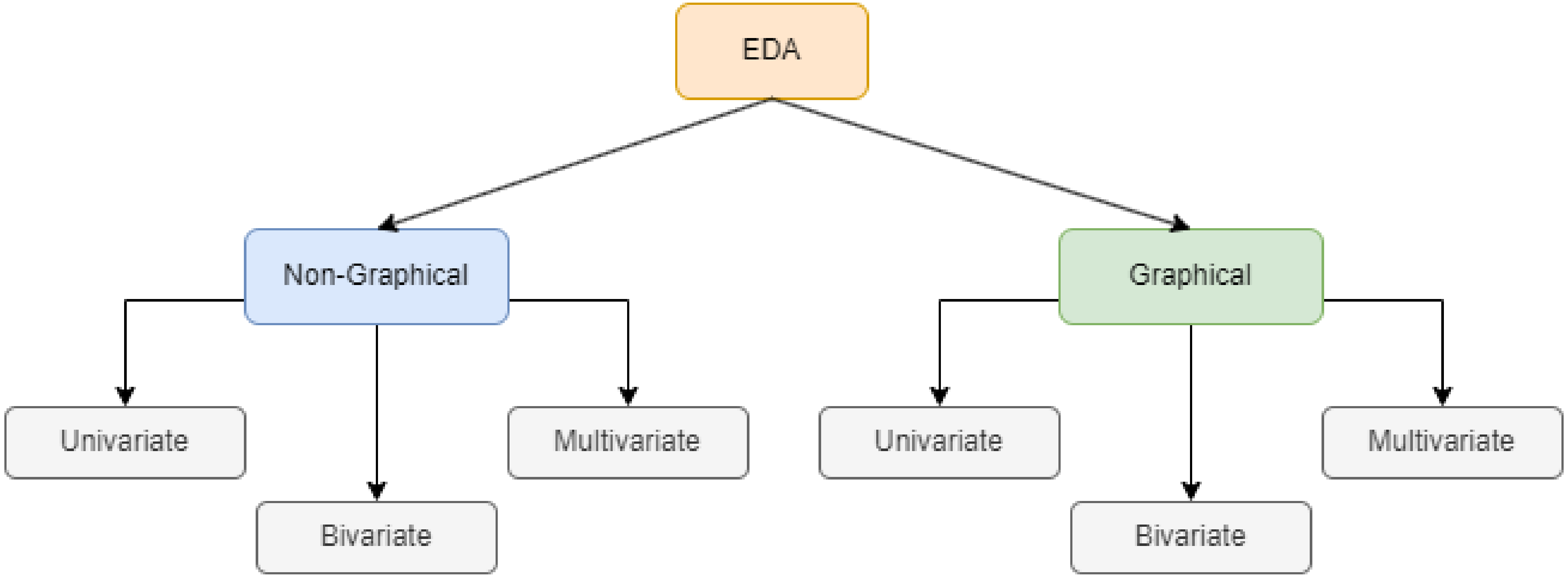
Data Visualization and EDA: <Numerical vs. Categorical Data>

In **data science**, data can be broadly categorized into **numerical** and **categorical** types based on their nature and characteristics:

- **Categorical Data:**
  - **Definition:** Categorical data represents categories or labels and cannot be measured in a numerical sense. It can be nominal or ordinal.
  - **Examples:**
    - **Nominal Categorical Data:** Colors, gender, country, etc.
    - **Ordinal Categorical Data:** Education levels, customer satisfaction ratings, etc.
  - **Analysis Techniques:**
    - Frequency counts and proportions.
    - Cross-tabulations and contingency tables.
    - **Visualization:** Bar charts, pie charts, stacked bar charts, etc.

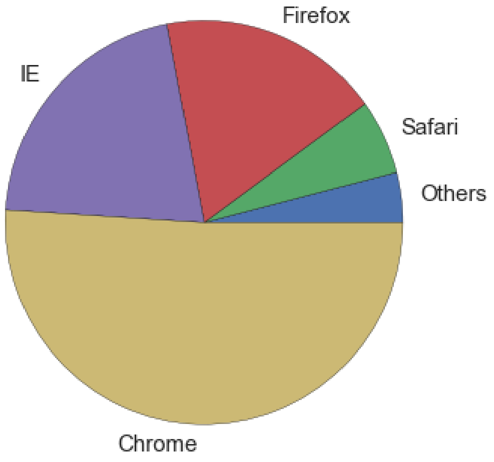


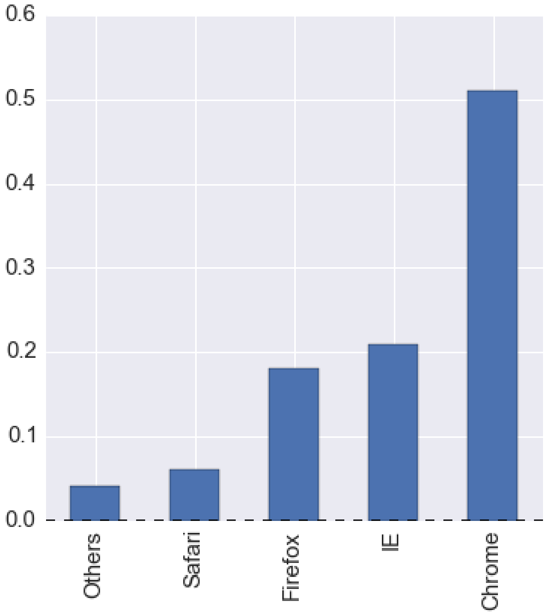
Data Visualization and EDA: Univariate, Bivariate, and Multivariate Analysis



Data Visualization and EDA: Univariate, Bivariate, and Multivariate Analysis <Categorical Data>

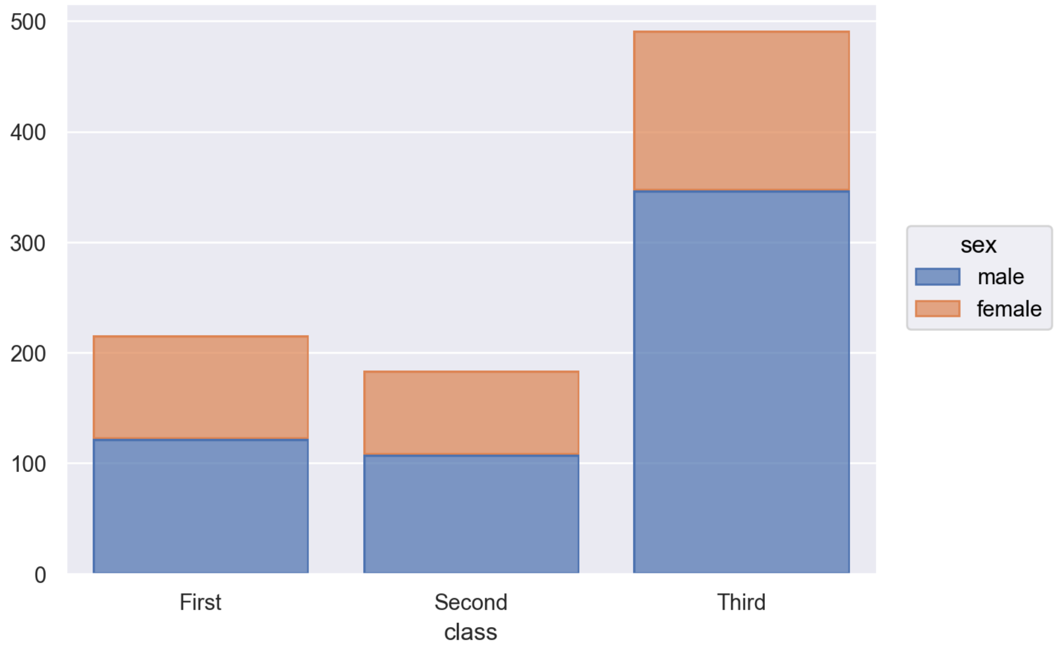
- **Univariate Analysis:**
  - **Graphical Methods:**
    - **Bar charts:** Display the frequency or proportion of each category.
    - **Pie charts:** Show the relative contribution of each category to the whole.
  - **Non-graphical Methods:**
    - **Frequency tables:** Summarize the count of each category.
    - **Mode:** Identify the most frequently occurring category.





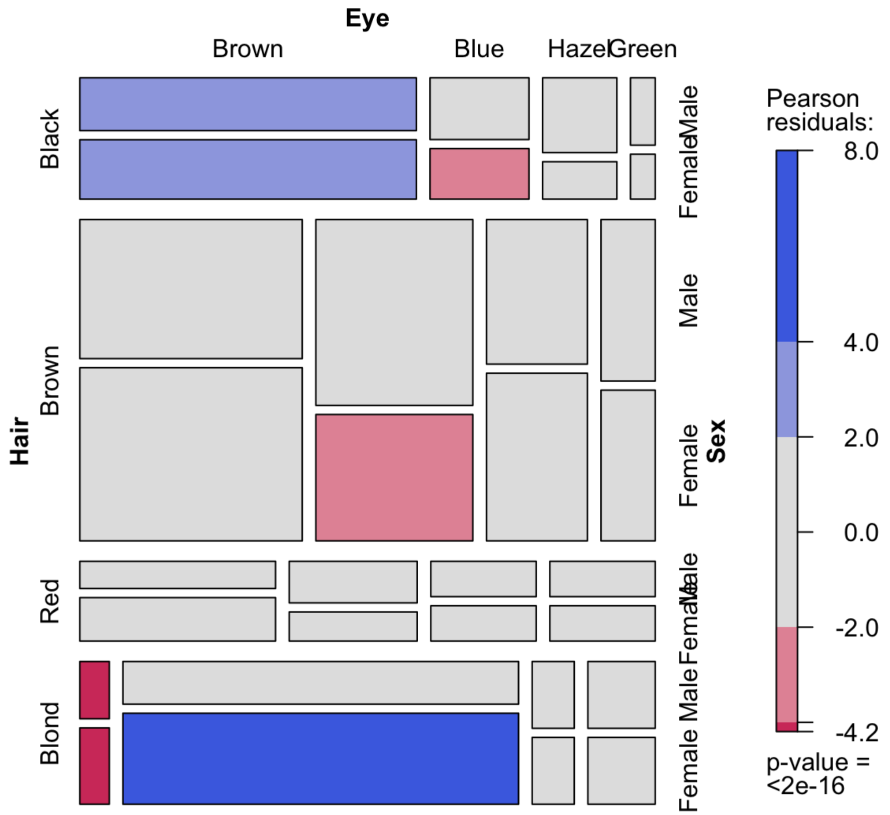
Data Visualization and EDA: Univariate, Bivariate, and Multivariate Analysis <Categorical Data>

- **Bivariate Analysis:**
  - **Graphical Methods:**
    - **Clustered bar charts:** Compare the distribution of one categorical variable across different levels of another categorical variable.
    - **Stacked bar charts:** Illustrate the composition of one categorical variable relative to another.
  - **Non-graphical Methods:**
    - **Chi-square test:** Assess the independence between two categorical variables.



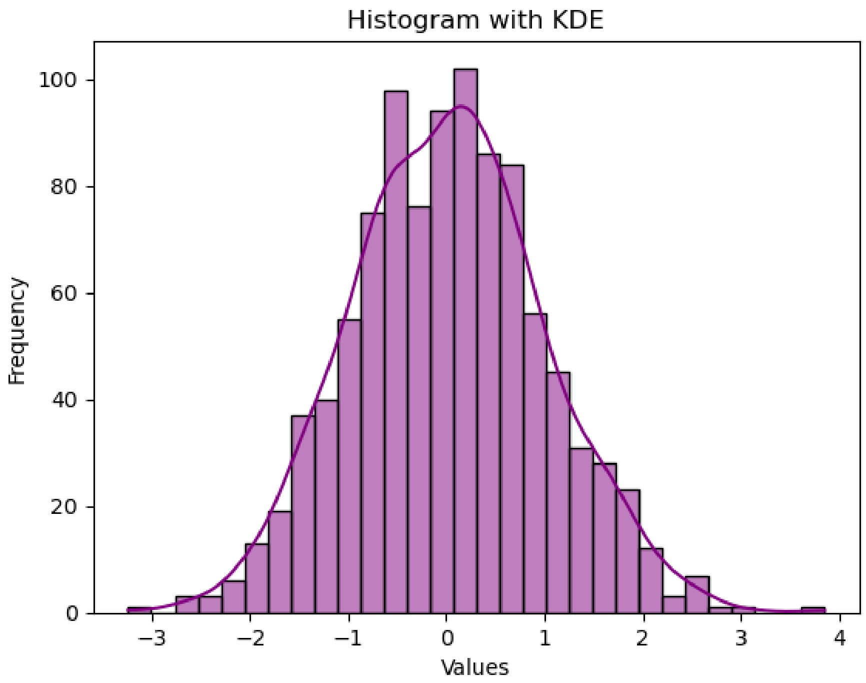
Data Visualization and EDA: Univariate, Bivariate, and Multivariate Analysis <Categorical Data>

- **Multivariate Analysis:**
  - **Graphical Methods:**
    - **Mosaic plots:** Visualize the relationship between three categorical variables.
    - **Clustered stacked bar charts:** Represent multiple categorical variables simultaneously.
  - **Non-graphical Methods:**
    - **Multinomial logistic regression:** Examine the relationship between multiple categorical variables.



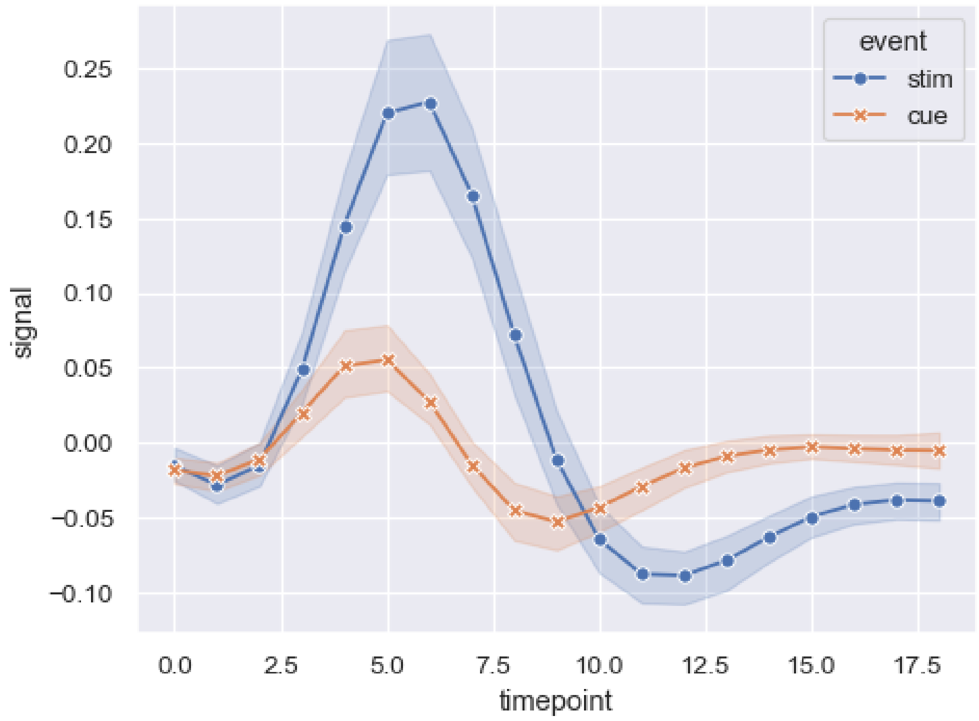
Data Visualization and EDA: Univariate, Bivariate, and Multivariate Analysis <Numerical Data>

- **Univariate Analysis:**
  - ○ **Graphical Methods:**
    - Histograms: Display the distribution of numerical values.
    - Box plots: Show summary statistics and identify outliers.
    - Kernel Density Plots: Estimate the probability density function
  - ○ **Non-graphical Methods:**
    - Descriptive statistics: Mean, median, mode, variance, standard deviation.
    - Percentiles and quantiles: Identify values below which a given percentage of observations fall.
    - Skewness and kurtosis: Measure of asymmetry and tailedness.



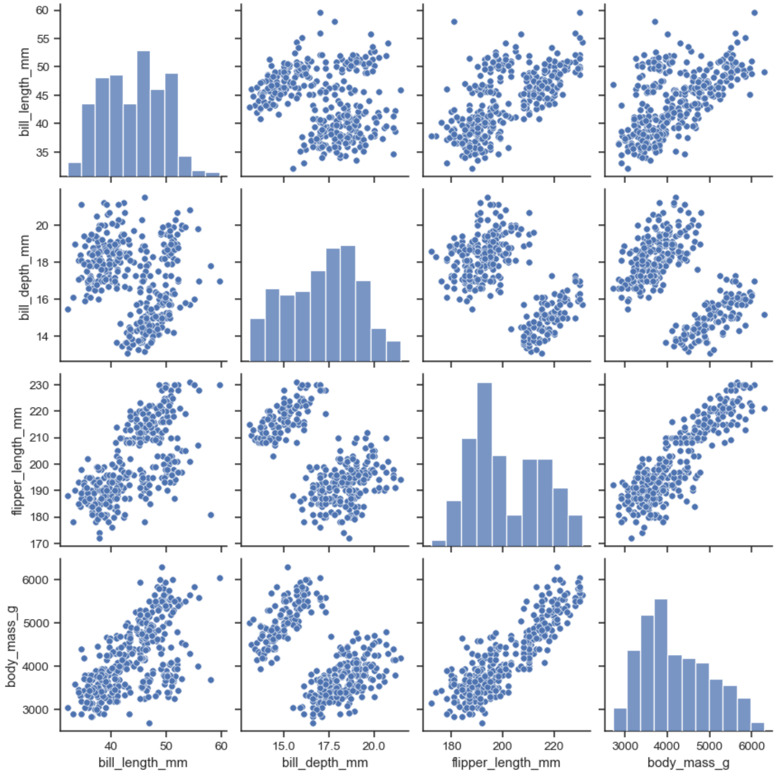
Data Visualization and EDA: Univariate, Bivariate, and Multivariate Analysis <Numerical Data>

- **Bivariate Analysis:**
  - **Graphical Methods:**
    - Scatter plots: Explore the relationship between two numerical variables.
    - Line plots: Show trends over time or across a variable.
    - Heatmaps: Visualize the correlation matrix between two numerical variables.
  - **Non-graphical Methods:**
    - Correlation coefficient: Quantify the strength and direction of a linear relationship.
    - Covariance: Measure the joint variability of two numerical variables.



Data Visualization and EDA: Univariate, Bivariate, and Multivariate Analysis <Numerical Data>

- **Multivariate Analysis:**
  - **Graphical Methods:**
    - 3D Scatter plots: Extend scatter plots to three dimensions for multiple numerical variables.
    - Pair plots: Combine scatter plots and histograms for multiple numerical variables.
  - **Non-graphical Methods:**
    - Multiple regression analysis: Assess the relationship between one dependent numerical variable and multiple independent numerical variables.
    - Principal Component Analysis (PCA): Reduce dimensionality while preserving as much variance as possible.



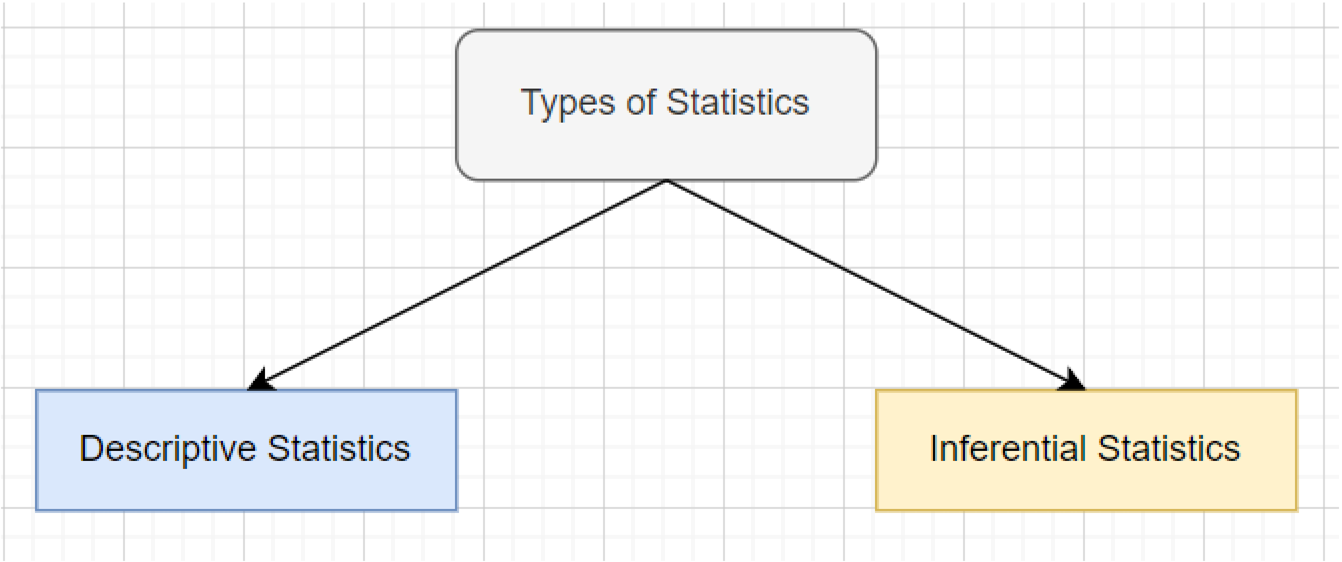
Data Visualization and EDA: General Tips

- **Data Cleaning:** You can perform EDA before and after cleaning the data by handling missing values, outliers, and ensuring data consistency.
- **Visualization Libraries:** Use tools like Matplotlib, Seaborn, and Plotly for graphical analysis in Python. R has ggplot2, lattice, and other packages.
- **Interpretation:** Always interpret the results of your analyses, considering the context of the data and potential implications for further analysis or decision-making.
- **Note:** Remember, EDA is an iterative process, and the insights gained can guide further analysis or modeling efforts.

Data Visualization and EDA: Statistics Fundamentals

- **Definition 1:** Statistics is a branch of mathematics that involves collecting, analyzing, interpreting, presenting, and organizing data. It provides methods for making inferences and decisions in the face of uncertainty. The goal of statistics is to extract meaningful patterns and insights from data, helping to understand and describe complex phenomena.
- **Definition 2:** Statistics is the science of learning from data. It involves collecting, analyzing, interpreting, presenting, and organizing data. It provides methods for drawing reliable conclusions and making informed decisions in the face of uncertainty.

Data Visualization and EDA: Statistics Fundamentals <Key Components>



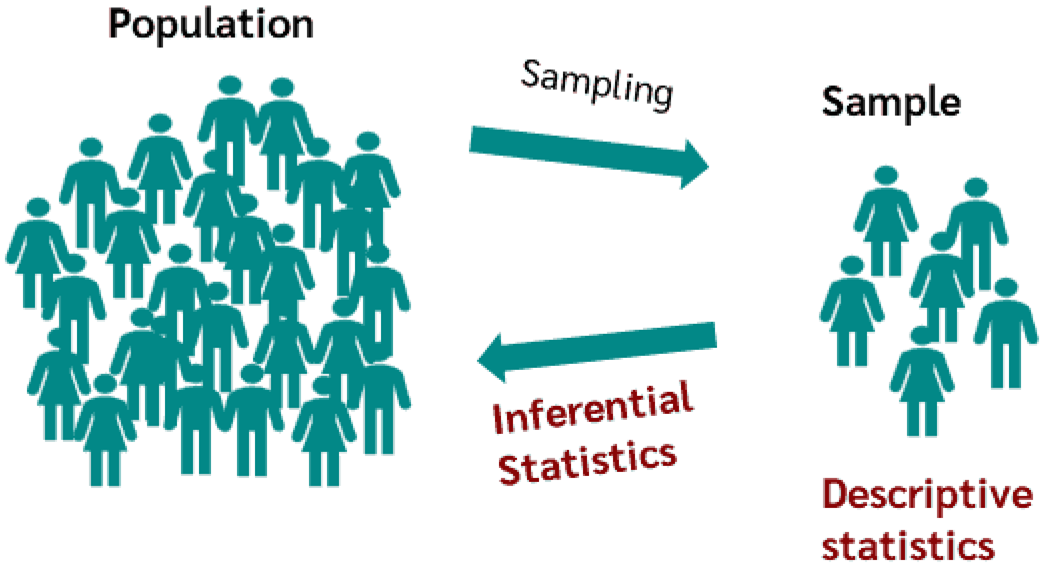
Types of Statistics

Descriptive Statistics

Inferential Statistics

Data Visualization and EDA: Statistics Fundamentals <Key Components>

Descriptive statistics and **inferential** statistics, along with exploratory statistics, are the main areas of statistics. **Descriptive** statistics provides tools to **describe** a sample. Starting from the sample, **inferential** statistics can now be used to make a statement about the population.



Source: Datalab.net

Data Visualization and EDA: Statistics Fundamentals <Key Components>

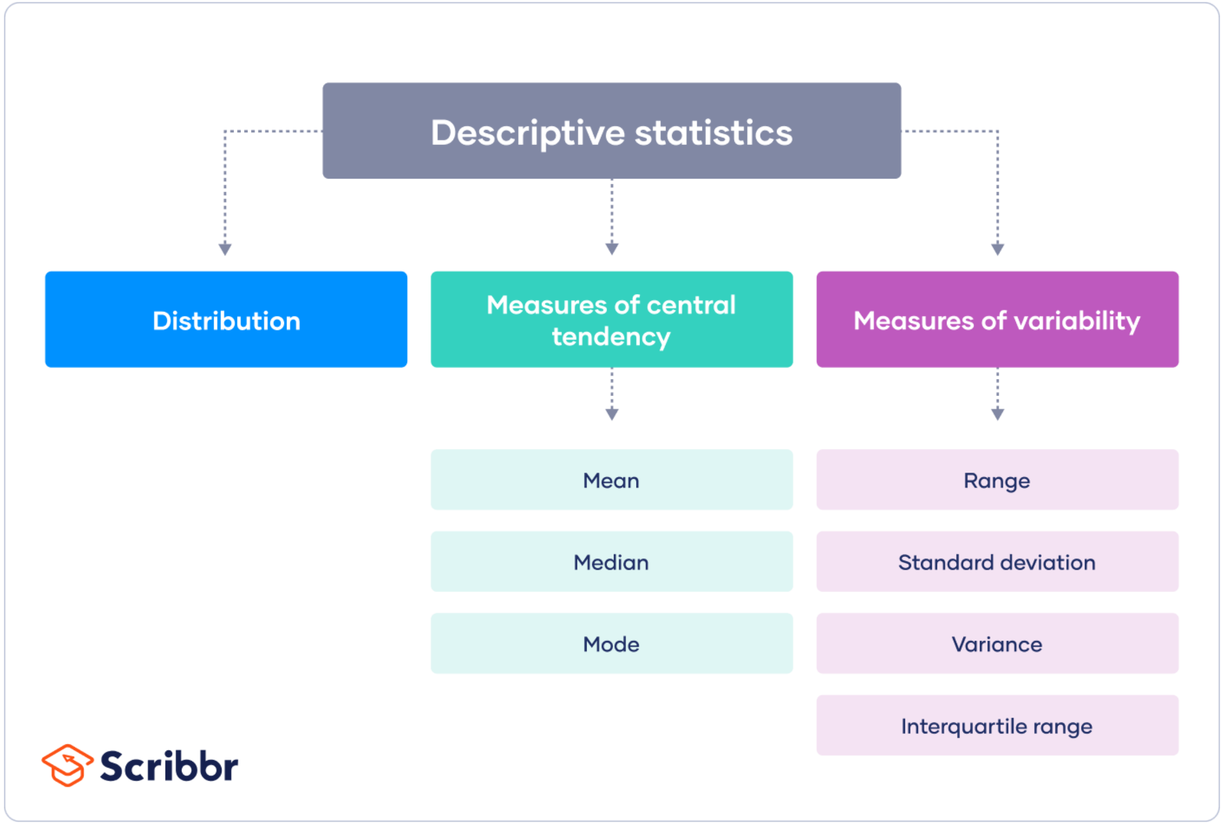
- **Descriptive Statistics:** This branch involves summarizing and describing the main features of a dataset. Descriptive statistics include measures such as mean, median, mode, standard deviation, and various graphical representations like histograms and scatter plots. These methods help in organizing and simplifying large amounts of data to make it more understandable.
- **Definition 2:** The term descriptive statistics covers statistical methods for describing data using statistical characteristics, charts, graphics or tables.

centralelille

ufr3s faculté d'ingénierie et management de la santé

Université de Lille

Data Visualization and EDA: Statistics Fundamentals <Key Components>



Descriptive statistics

Distribution

Measures of central tendency

Measures of variability

Mean

Median

Mode

Range

Standard deviation

Variance

Interquartile range

Source: Scribbr.com

Data Visualization and EDA: Statistics Fundamentals <Key Components>

- **Inferential Statistics:** This branch uses statistical methods to make predictions or inferences about a population based on a sample of data taken from that population. Inferential statistics include hypothesis testing, confidence intervals, regression analysis, and more. These methods help researchers draw conclusions from a limited set of observations.
- **Definition 2:** What's inferential statistics? In contrast to descriptive statistics, inferential statistics want to make a statement about the population. However, since it is almost impossible in most cases to survey the entire population, a sample is used, i.e. a small data set originating from the population. With this sample a statement about the population can be made. An example would be if a sample of 1,000 citizens is taken from the population of all Canadian citizens.

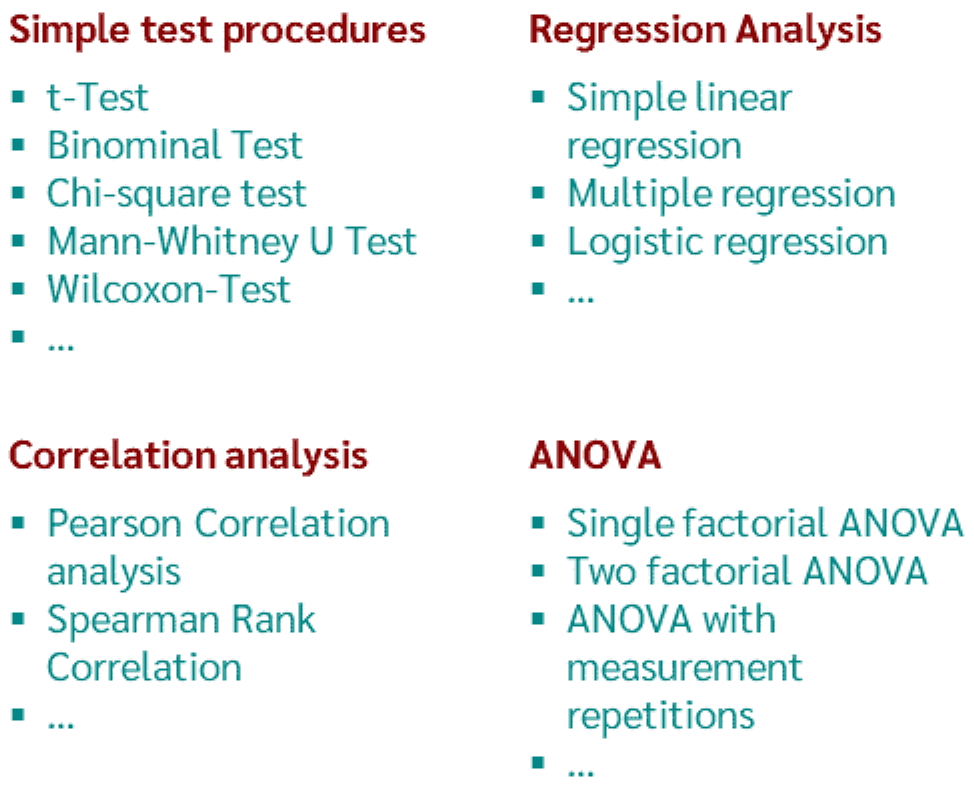
Inferential statistics:

Testing statements about the population on the basis of sample characteristics.

Source: Datalab.net

Data Visualization and EDA: Statistics Fundamentals <Key Components>

Depending on which statement is to be made about the population or which question is to be answered about the population, different statistical methods or hypothesis tests are used. The best known are the hypothesis tests with which a group difference can be tested, such as the t-test, the chi-square test or the analysis of variance. Then there are the hypothesis tests with which a correlation of variables can be tested, such as correlation analysis and regression.



Simple test procedures

- t-Test
- Binominal Test
- Chi-square test
- Mann-Whitney U Test
- Wilcoxon-Test
- ...

Regression Analysis

- Simple linear regression
- Multiple regression
- Logistic regression
- ...

Correlation analysis

- Pearson Correlation analysis
- Spearman Rank Correlation
- ...

ANOVA

- Single factorial ANOVA
- Two factorial ANOVA
- ANOVA with measurement repetitions
- ...

Source: Datalab.net

Data Visualization and EDA: Statistics Fundamentals <Key Components>

Difference Between Descriptive and Inferential Statistics

Descriptive statistics provide a **summary** of the features or attributes of a **dataset**, while **inferential** statistics enable **hypothesis testing** and evaluation of the applicability of the data to a larger population.

Data Visualization and EDA: Statistics Fundamentals <Key Components>

Difference Between Descriptive and Inferential Statistics

	Descriptive Statistics	Inferential Statistics
Purpose	Describe and Summarize data	Make inferences and draw conclusions about a population based on sample data
Data Analysis	Analyzes and interprets the characteristics of a dataset	Uses sample data to make generalizations or predictions about a larger population
Population vs. Sample	Focuses on the entire population or a dataset	Focuses on a subset of the population (sample) to draw conclusions about the entire population
Measurements	Provides measures of central tendency and dispersion	Estimates parameters, tests hypotheses, and determines the level of confidence or significance in the results
Examples	Mean, median, mode, standard deviation, range, frequency tables	Hypothesis testing, confidence intervals, regression analysis, ANOVA (analysis of variance), chi-square tests, t-tests, etc.

Source: simplilearn.com

Data Visualization and EDA: Statistics Fundamentals <Key Components>

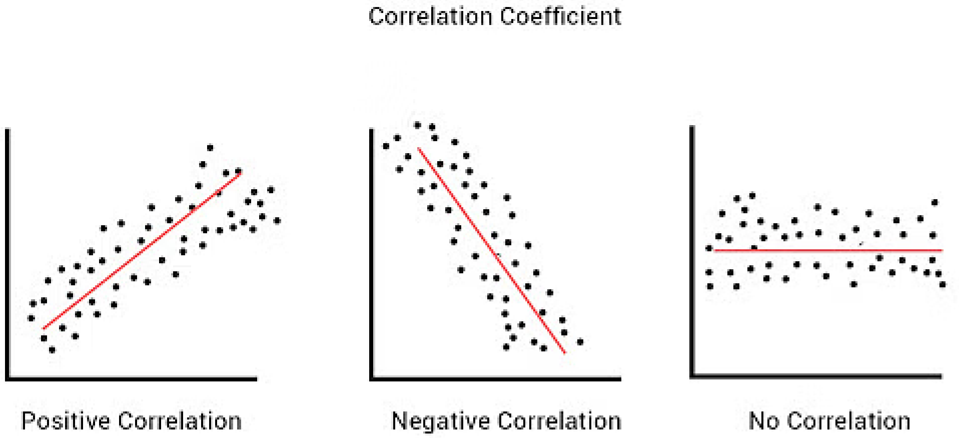
Difference Between Descriptive and Inferential Statistics

	Descriptive Statistics	Inferential Statistics
Goal	Summarize organize and present data	Generalize findings to a larger population, make predictions, test hypotheses, evaluate relationships, and support decision-making
Population Parameters	Not typically estimated	Estimated using sample statistics (e.g., sample mean as an estimate of population mean)
Sample Representativeness	Not required	Crucial; the sample should be representative of the population to ensure accurate inferences

Source: simplilearn.com

Data Visualization and EDA: Statistics Fundamentals <Correlation>

Correlation in statistics refers to the statistical relationship or association between two or more variables. It measures the degree to which changes in one variable are associated with changes in another. In other words, correlation helps to quantify the strength and direction of a linear relationship between two variables. The two most common measures of correlation are the Pearson correlation coefficient and the Spearman rank correlation coefficient.

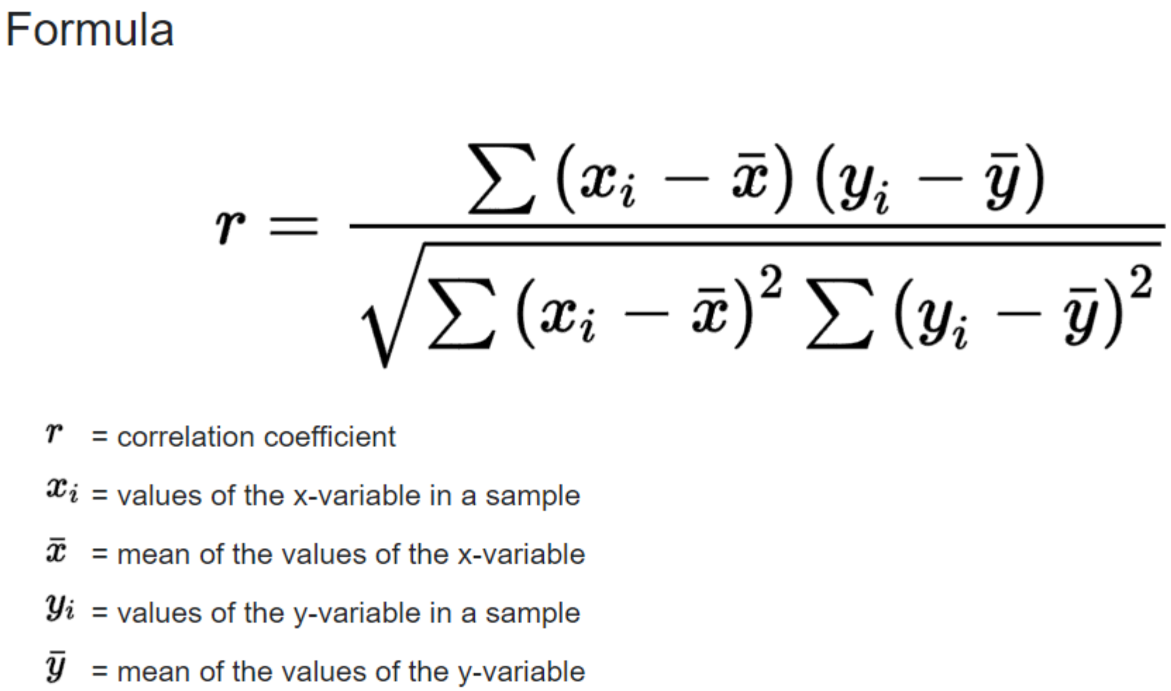


Data Visualization and EDA: Statistics Fundamentals <Pearson Correlation Coefficient>

Pearson Correlation Coefficient:

- Denoted by r, the Pearson correlation coefficient ranges from -1 to 1.
- r=1 indicates a perfect positive linear relationship.
- r=−1 indicates a perfect negative linear relationship.
- r=0 indicates no linear relationship.
- Positive values of r signify a positive association, while negative values indicate a negative association.

Note: It cannot determine the nonlinear relationships between variables.



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

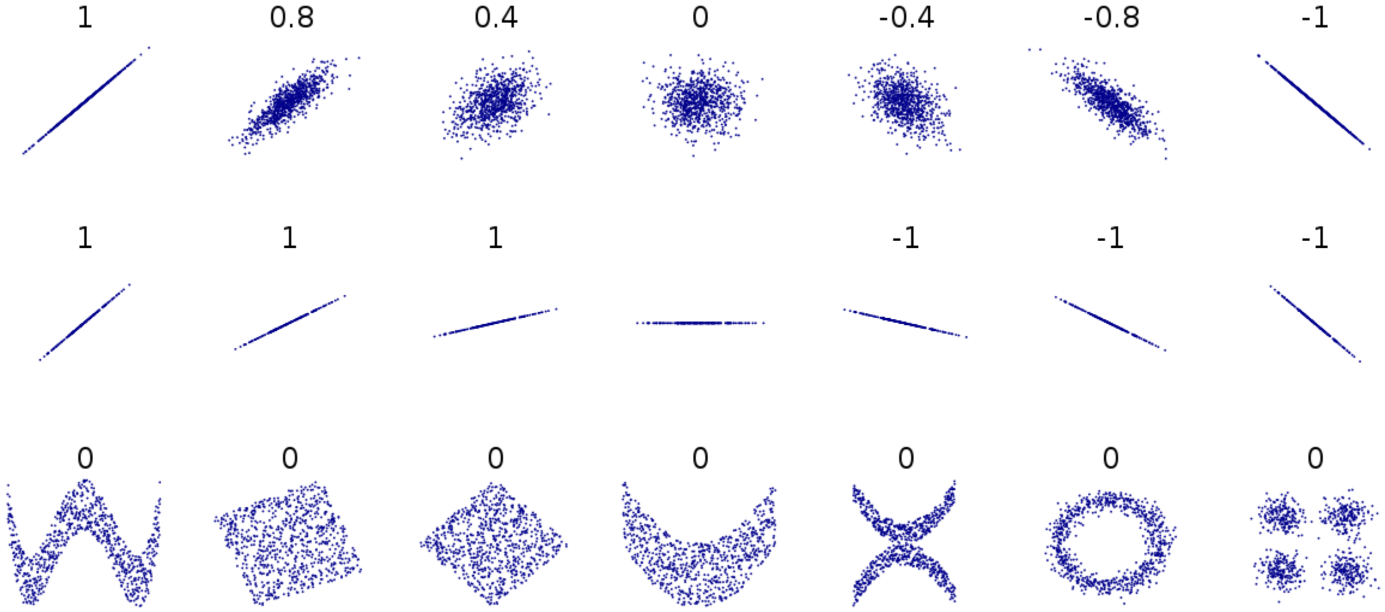
$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

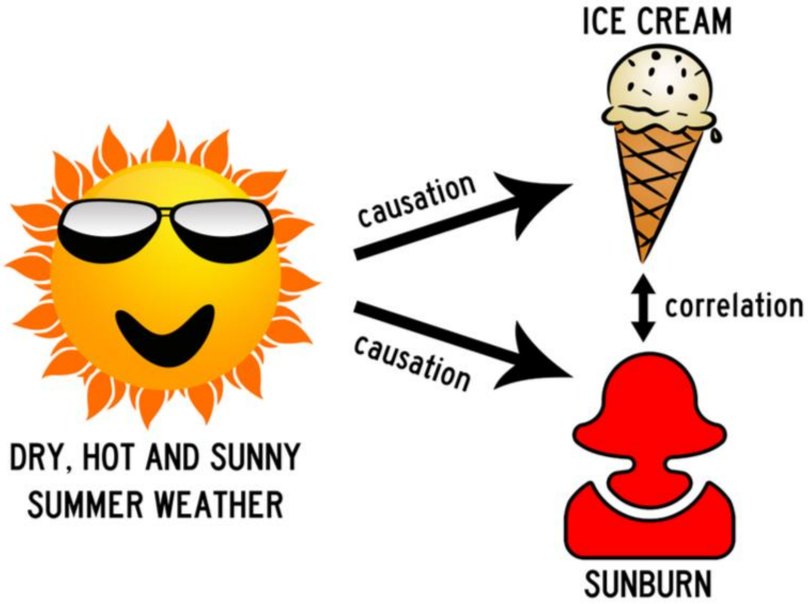
Data Visualization and EDA: Statistics Fundamentals <Pearson Correlation Coefficient>

Note: It cannot determine the nonlinear relationships between variables.



Data Visualization and EDA: Statistics Fundamentals <Correlation != Causation>

Correlation tests for a relationship between two variables. However, seeing two variables moving together does not necessarily mean we know whether one variable causes the other to occur. This is why we commonly say “correlation does not imply causation.”



Data Visualization and EDA

Demo: [Session 2 – Data Visualization and EDA]