# Assignment 5

## Problem statement

**Visualize the data using R/Python by plotting the graphs for assignment no. 1 and 2. Consider a suitable data set.**
**a) Use Scatter plot, bar plot, Box plot, and Histogram OR**
**b) Perform the data visualization operations using Tableau for the given dataset.**

## S/W Packages and H/W apparatus used:

**OS**: **Ubuntu/Windows, Google Colab**
**Packages**: **Numpy, Pandas, Matplotlib and Seaborn**

## Theory :

## 1. Scatter Plot:

- Representation of Relationship: Shows the relationship between two numerical variables.
- Pattern Identification: Helps identify patterns, trends, and correlations in the data.
- Correlation Assessment: Enables assessment of the strength and direction of the relationship between variables.
- Exploratory Analysis: Useful for exploratory data analysis before building predictive models.
- **Importance and Advantages:**

  - Facilitates data exploration and pattern recognition.

  - Provides visual insights into complex relationships that may not be evident from raw data.

- **Application**: - Used in scientific research, social sciences, finance, and machine learning for understanding correlations and making predictions.

## 2. Bar Plot:

- Categorical Data Representation: Displays the frequency or count of categorical variables as bars.
- Comparison of Categories: Facilitates comparison of values across different categories.
- Most Frequent Categories: Identifies the most frequent or dominant categories within the dataset.
- Descriptive Statistics: Widely used in descriptive statistics to summarize categorical data.
- **Importance and Advantages**:

  - Provides a clear visual representation of categorical data distributions.

  - Simplifies complex data for easy interpretation.

- **Application**:

  - Used in market research, opinion polls, and surveys for analyzing categorical data and presenting findings.

## 3. Box Plot:

- Distribution Summary: Provides a visual summary of the distribution of numerical data through quartiles.
- Spread and Skewness Identification: Helps visualize the spread, skewness, and presence of outliers in the data distribution.

- Central Tendency: Displays the median and interquartile range, providing insights into the central tendency and variability of the data.
- Comparison Across Groups: Useful for comparing the distribution of numerical variables across different groups or categories.
- **Importance and Advantages**:

  - Efficiently summarizes the distribution of data and detects outliers.

  - Provides a visual summary of key statistics such as median, quartiles, and range.

- **Application:**

  - Used in quality control, finance, and healthcare for identifying anomalies and comparing data distributions.

## 4. Histogram:

- Frequency Distribution: Represents the frequency distribution of a single numerical variable.
- Shape and Spread Visualization: Visualizes the shape, center, and spread of the data distribution.
- Bin-Based Representation: Data is grouped into bins, and the height of each bar represents the frequency of observations within that bin.
- Pattern Identification: Identifies patterns, detects outliers, and assesses the overall distributional characteristics of numerical variables.
- **Importance and Advantages:**

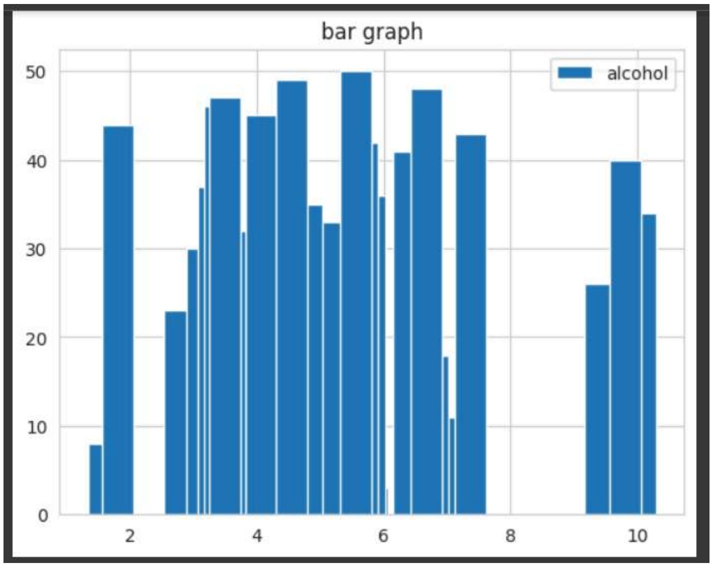
  - Provides a detailed view of data distribution and helps identify underlying patterns.

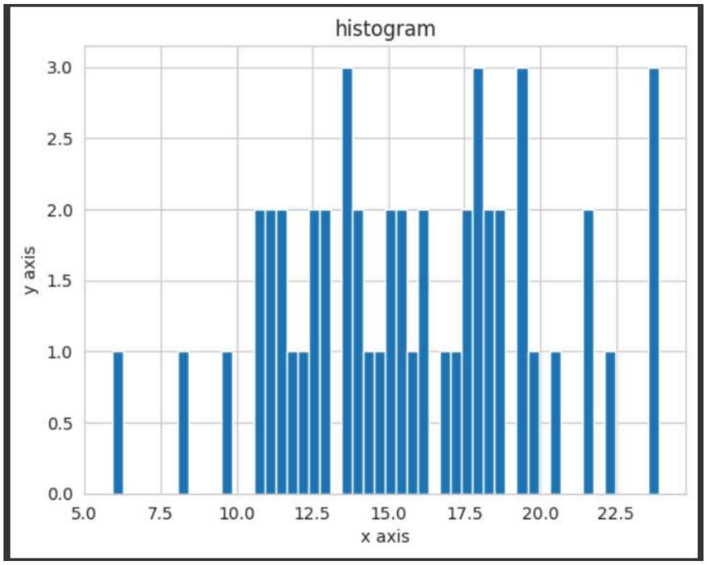
- Facilitates data preprocessing by identifying outliers and understanding data distributions.

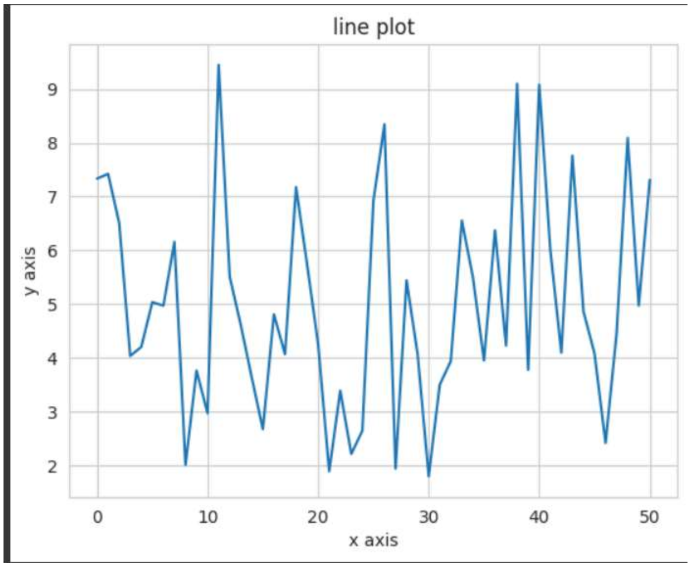
- **Application**:

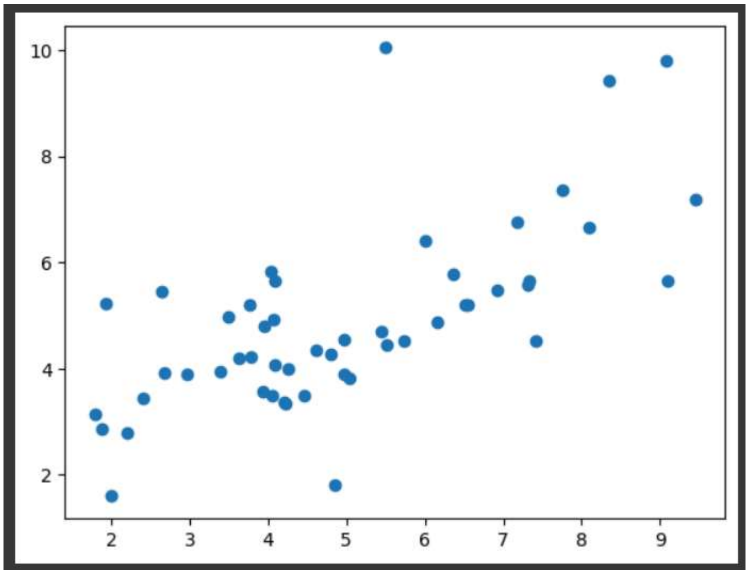
  - Used in data analysis, machine learning, and statistical modeling for exploring data distributions and understanding data characteristics.

## Diagram:









## Conclusion

Data visualization plays a crucial role in data analysis, providing valuable insights into complex datasets and facilitating decision-making processes. Techniques such as scatter plots, bar plots, box plots, and histograms offer effective ways to explore, summarize, and communicate patterns and trends within data. By visually representing relationships, distributions, and comparisons, data visualization enhances understanding, enables discovery, and supports evidence-based decision-making across diverse domains. Incorporating data visualization techniques into data analysis workflows empowers analysts and stakeholders to extract actionable insights, identify opportunities, and address challenges effectively. As an indispensable tool in the data analysis toolkit, data visualization continues to drive innovation, inform strategies, and unlock the full potential of data-driven approaches in research, business, and society.