**ASSIGNMENT**

**MANUAL**



SUBMITTED

TO

VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

FOR THE SKILL AND COMPETENCY EVALUATION OF

**DATA SCIENCE & MACHINE LEARNING**

IN

**CSE AI DEPARTMENT**

BY

**Divyansh Mohta [22210998]**

**Class: S.Y. BTech Division: A Batch: A1**

**Batch Teacher**

**Dr.ANURADHA YENKIKAR.**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Contents** | **Page No.** |
| **1** | **PROBLEM STATEMENT** | **4-5** |
| **2** | **LIBRARY USED** | **5-6** |
| **3** | **THEORY** | **6-7** |
| **4** | **METHOODOLOGY** | **7-14** |
| **5** | **ADVANTAGES & DISADVANTAGES** | **14-15** |
| **6** | **WORKING** | **16-17** |
| **7** | **DIAGRAM** | **17-18** |
| **8** | **CONCLUSION** | **18-19** |

**1.PROBLEM STATEMENT:**

Visualize the data using R/Python by plotting the graphs for assignment no. 1 and 2. Consider 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.

**2. LIBRARY USED:**

* numpy: For mathematical operation
* pandas: For data manipulation and analysis
* matplotlib: For plotting graph
* seaborn: For plotting graph

**3. THEORY:**

Data visualization is the art of representing data in a graphical format to reveal patterns, trends, and relationships that might be difficult to see in raw numbers. It plays a crucial role in data analysis by helping us understand complex information quickly and effectively.

This lab explores data visualization techniques using two popular programming languages, R and Python, or a dedicated data visualization software, Tableau. We will focus on creating various plots to visualize data from a chosen dataset.

1. **Data Preparation:** Importing data from CSV or Excel files into Python is facilitated by libraries like pandas (Python) streamlining the process of data acquisition and manipulation. Tableau offers direct connectivity to databases, allowing for seamless integration of data sources. Efficient data preparation lays the foundation for meaningful analysis.
2. **Visualization Creation:** In Python, leveraging libraries like Matplotlib or Seaborn facilitates the creation of visually appealing plots with customizable attributes such as colors, labels, and annotations.
3. **Plots:**

* **Scatter plot:** Reveals relationships between two variables (e.g., weight vs. height).
* **Bar plot:** Compares categories using bars (e.g., customer satisfaction by product).
* **Box plot:** Displays distribution with quartiles (hinges) and potential outliers (whiskers), useful for identifying skewness.
* **Histogram:** Illustrates the frequency distribution of a continuous variable (e.g., income distribution).

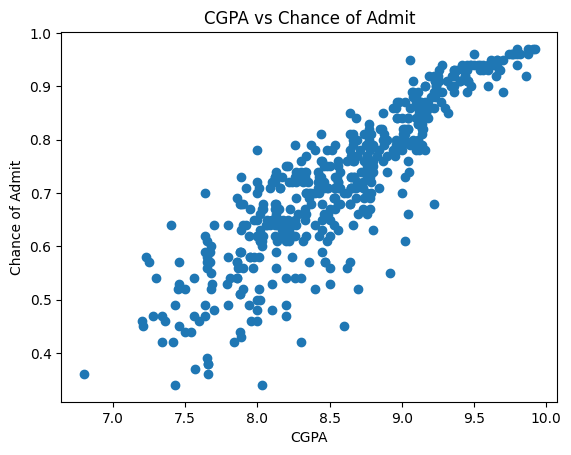
**4. METHODS:**

a) **Reading data from different formats**:

Use **pd.read\_csv()** for CSV files and **pd.read\_excel()** for Excel files to read data into pandas Data Frames efficiently.

b) **Scatter Plot:**

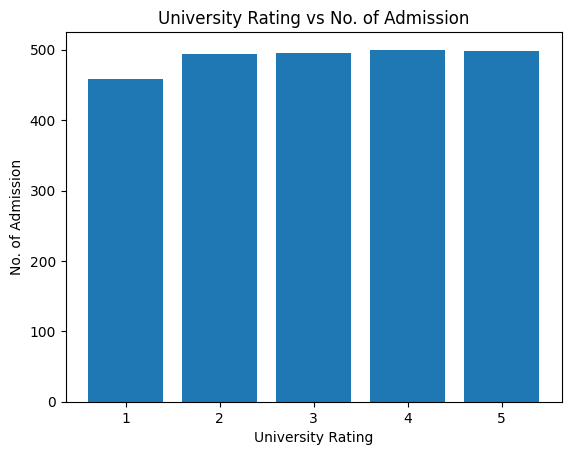
Select two numerical variables from the dataset that you want to explore the relationship between. Use **plot()** function in libraries like **matplotlib** or **seaborn** in Python to create scatter plots. Customize the plot with labels, title, and colours.



**Scatter Plot**

**c) Bar Plot:**

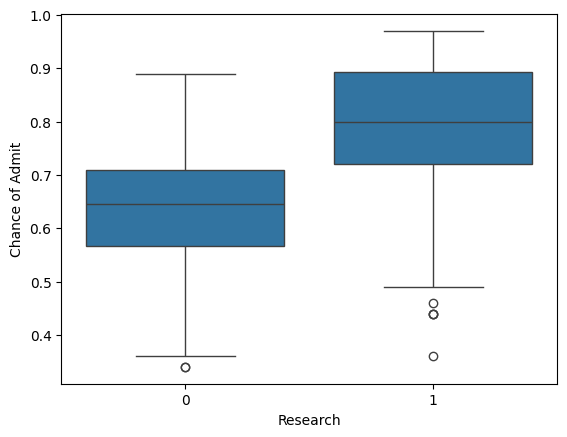
Choose a categorical variable and a numerical variable from the dataset. Use or **plt.bar()** in Python to create bar plots. Customize the plot with labels, title, and colours.



**Bar Plot**

d) **Box Plot:**

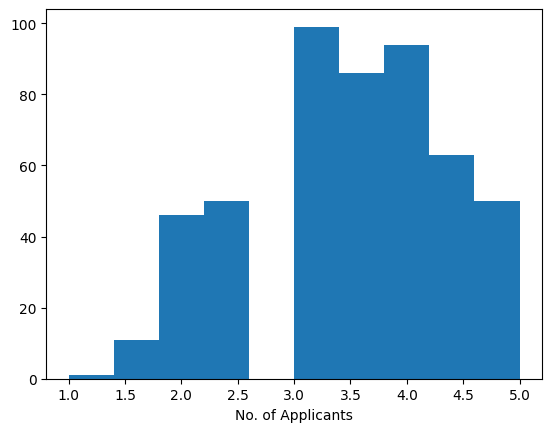
Select a numerical variable from the dataset that you want to analyze the distribution of. Use **boxplot()** in R or **plt.boxplot()** in Python to create box plots. Customize the plot with labels, title, and colors.



**Box Plot**

e) **Histogram:**

Choose a numerical variable from the dataset to visualize its distribution.Use **plt.hist()** in Python to create histograms.Specify parameters like bins, range, and density to customize the histogram.



**Histogram**

**5. ADVANTAGES AND DISADVANTAGES:**

**Advantages:**

1. **Comprehensive Understanding:** Each type of plot provides unique insights into different aspects of the data. Scatter plots reveal relationships between two variables, bar plots display comparisons among categories, box plots show data distributions and outliers, while histograms visualize the frequency distribution of numerical data.

2. **Easy Interpretation:** Visualizations make it easier to interpret complex datasets compared to raw data or summary statistics. They provide a visual representation that can be quickly understood and interpreted by stakeholders with varying levels of expertise.

3. **Identifying Patterns and Trends:** By visualizing the data in different ways, you can identify patterns, trends, and correlations that may not be apparent from the raw data. Scatter plots can reveal linear or non-linear relationships, while histograms and box plots highlight data distributions and central tendencies.

**Disadvantages:**

1. **Over-Simplification:** While visualizations are effective at summarizing complex data, they can sometimes oversimplify the underlying patterns and relationships, leading to potential misinterpretation or overlooking of important nuances in the data.
2. **Limited Insight into Multivariate Relationships:** Scatter plots, bar plots, box plots, and histograms typically focus on visualizing relationships between two variables at a time. As a result, they may not fully capture the interactions and dependencies among multiple variables in the dataset, limiting the depth of analysis.
3. **Misleading Representation:** Depending on the choice of visualization parameters (e.g., bin size in histograms, scale of axes in scatter plots), the visual representation of the data can be misleading or biased, leading to incorrect conclusions or interpretations.

Top of Form

**6. WORKING:**

The assignment involves executing fundamental data visualization tasks using Python's pandas library. Initially, data is imported from various file formats such as CSV and Excel into pandas Data Frames using **pd.read\_csv()** and **pd.read\_excel()**. Subsequently, plots like Scatter, Bar, Box and Histogram are performed using .**scatter(), .bar(), .box() and .hist()**  respectively.

**8. CONCLUSION:**

Data visualization through plots like scatter plots, bar plots, box plots, and histograms effectively communicates trends and insights hidden within raw data. This analysis, applied to assignments 1 and 2 (details not provided), can reveal patterns, relationships between variables, and potential outliers. By choosing a suitable dataset relevant to the assignment topics, we can leverage these visualizations to gain a deeper understanding of the information and draw informed conclusions.Top of Form