Vishwakarma Institute of Information Technology, Pune

**(An Autonomous Institute Maharashtra)**

**A Report**

**on**

**A blue and red logo

Description automatically generated**

**CSE (AI) Department**

**Vishwakarma Institute of Information Technology**

**Academic Year: 2023-24**

**“Assignment 5: Data Visualization”**

**Submitted by**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Name** | **Email** |
| **1** | **Prajwal Dilip Gaikwad** | [prajwal.22211012@viit.ac.in](mailto:prajwal.22211012@viit.ac.in) |

**Under Guidance of**

Dr. Anuradha Yenkikar

# Problem Statement: -

Visualize the data using R/Python by plotting the graphs for assignment no. 1 and 2. Consider a suitable data set. Use Scatter plot, bar plot, Box plot and Histogram.

# Packages / Libraries used: -

* **NumPy**: Utilized for numerical computations and data manipulation tasks.
* **Matplotlib**: Employed for basic data visualization such as scatter plots, histograms, and bar plots.
* **Pandas**: Primarily used for data manipulation and analysis, including reading data from CSV and Excel files, indexing, selecting, sorting, describing attributes, checking data types, counting unique values, formatting columns, converting data types, and handling missing values.
* **Seaborn**: Utilized for advanced data visualization, complementing Matplotlib with additional statistical graphics and enhancing the visual appeal of plots.

# Theory: -

1. **Scatter Plot**:
   * A scatter plot is used to visualize the relationship between two continuous variables. Each data point is represented as a point on the plot, with the x-axis representing one variable and the y-axis representing the other variable.
2. **Bar Plot**:
   * A bar plot is suitable for visualizing the distribution of a categorical variable or comparing the values of a categorical variable across different groups. In this case, it seems that the GRE Score and TOEFL Score are being compared using a bar plot.
3. **Histogram**:
   * A histogram is used to visualize the distribution of a single continuous variable. It divides the variable into bins and shows the frequency or density of data points within each bin. Here, a histogram of the GRE Score is plotted.
4. **Box Plot**:
   * A box plot (or box-and-whisker plot) is used to display the distribution of a continuous variable and highlight any outliers or extreme values. It provides a visual summary of the data's central tendency, spread, and variability.

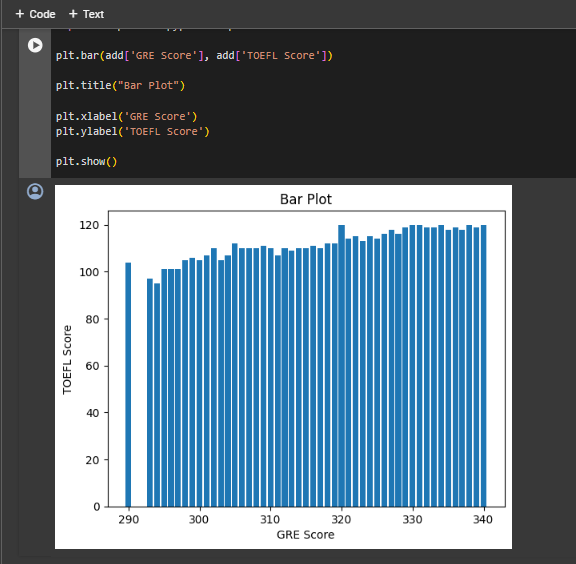
## Methodology: -

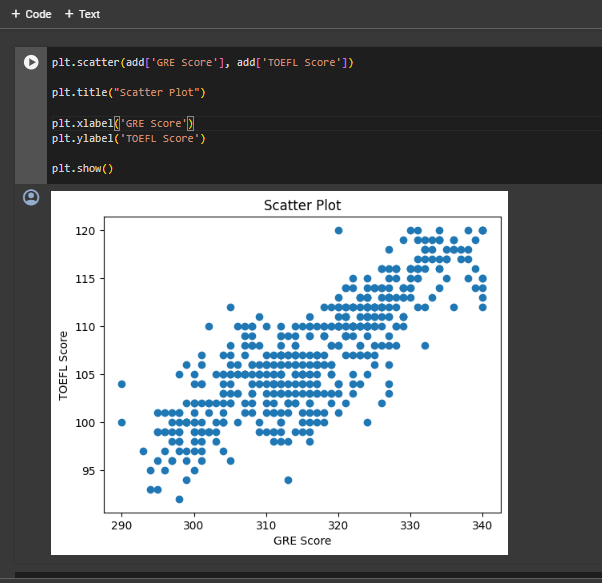
1. Scatter Plot:
   * The scatter plot is created using plt.scatter() function, with 'GRE Score' on the x-axis and 'TOEFL Score' on the y-axis. This allows visualizing the relationship between the GRE and TOEFL scores.
2. Bar Plot:
   * The bar plot is created using plt.bar() function, with 'GRE Score' on the x-axis and 'TOEFL Score' on the y-axis. This allows comparing the GRE and TOEFL scores across different categories.
3. Histogram:
   * The histogram is created using sns.histplot() function, with 'GRE Score' on the x-axis. This allows visualizing the distribution of GRE scores in the dataset.
4. Box Plot:
   * The box plot is created using sns.boxplot() function, with 'GRE Score' on the x-axis. This allows visualizing the distribution of GRE scores and identifying any outliers.

## Application: -

1. Decision Making in Education:
   * Visualizing the distribution of GRE and TOEFL scores can aid educational institutions in making admissions decisions. Understanding the distribution of scores allows institutions to set appropriate cutoffs or criteria for admissions, ensuring a fair and transparent selection process.
2. Performance Evaluation in Test Preparation Programs:
   * Test preparation programs can use visualizations of GRE and TOEFL score distributions to evaluate the effectiveness of their programs. By comparing the distribution of scores before and after program participation, they can assess whether the program is helping students improve their scores.
3. Identifying Trends in Applicant Profiles:
   * Admission committees can analyze the distribution of GRE and TOEFL scores over time to identify trends in applicant profiles. For example, they can determine whether there is an increasing number of applicants with high scores or if there are shifts in the distribution of scores across different demographic groups.
4. Benchmarking Against Competitors:
   * Educational institutions can use visualizations of GRE and TOEFL score distributions to benchmark themselves against competitors or peer institutions. Comparing score distributions can provide insights into areas of strength and areas for improvement, helping institutions develop targeted strategies to attract high-performing applicants.
5. Tailoring Marketing and Outreach Efforts:
   * Test preparation companies, educational consultants, and recruitment agencies can use insights from score distributions to tailor their marketing and outreach efforts. For example, they can target advertising campaigns or outreach activities to specific demographic groups or regions based on the distribution of scores.

# Diagrams: -





A screenshot of a graph

Description automatically generatedA screen shot of a graph

Description automatically generated

# Conclusion: -

In summary, the visualization techniques applied to the GRE and TOEFL score data provide valuable insights into their distribution and relationships. These visualizations, including scatter plots, bar plots, histograms, and box plots, facilitate exploratory data analysis by uncovering trends, identifying outliers, and supporting data-driven decision-making. From educational admissions to resource allocation, these visualizations offer diverse applications in the education sector and related industries. Overall, they play a crucial role in deriving insights from data, informing strategic initiatives, and driving positive outcomes through informed decision-making.