Assignment 3

Name: Muskan
Roll No: 281028
Batch: A2
Prn: 22310735

Statement

In this assignment, we aim to:
a) Perform data visualization using Matplotlib and Seaborn.
b) Identify patterns and relationships between features using scatter plots, bar plots, and box plots.
c) Conduct exploratory data analysis (EDA) on structured data.
d) Process and clean data for improved visualization.

Objective

1. Utilize Pandas, Matplotlib, and Seaborn for data analysis and visualization.
2. Develop skills in exploratory data analysis (EDA) through effective data representation.
3. Identify key insights by visualizing relationships between different features in a dataset.

Resources Used

- **Software:** VS Code
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn

Introduction to Data Visualization

Data visualization is a crucial aspect of data analysis, helping interpret large datasets efficiently. By using **Matplotlib** and **Seaborn**, we can create insightful visual representations of data to identify trends, correlations, and outliers.

Key Functionalities Used:

1. **Data Handling with Pandas**

- pd.read_csv(): Reads data from a CSV file into a DataFrame.
- isnull().sum(): Identifies missing values in the dataset.
- describe(): Provides summary statistics for numerical columns.

2. **Data Visualization with Matplotlib and Seaborn**
   - sns.scatterplot(): Creates scatter plots to analyze relationships between variables.
   - sns.barplot(): Generates bar plots to compare categorical values.
   - sns.boxplot(): Produces box plots to examine data distributions and outliers.
   - plt.show(): Displays the plotted graphs.

Methodology

1. Data Collection and Preprocessing

- **Dataset Used:** *admission.csv*
- **Features:** GRE Score, CGPA, University Rating, Research, and Chance of Admit.
- **Initial Steps:**
  - Loaded the dataset using Pandas.
  - Checked for missing values.
  - Renamed columns to remove unnecessary spaces.

2. Data Visualization

- **Scatter Plot:** *GRE Score vs. Chance of Admit*
  - Visualized the relationship between GRE scores and admission chances, using color coding for research experience.
  - Helped identify trends in student admissions.
- **Bar Plot:** *Average CGPA by University Rating*
  - Compared the average CGPA of students across different university ratings.
  - Provided insights into the correlation between university prestige and student performance.
- **Box Plot:** *Distribution of SOP Scores by Research*

- Showed the spread of Statement of Purpose (SOP) scores for students with and without research experience.
- Highlighted variations in SOP scores across different categories.

Advantages of Data Visualization

1. Helps in identifying trends, correlations, and patterns.
2. Makes complex datasets more interpretable.
3. Enhances decision-making by providing visual insights.

Disadvantages

1. Can be misleading if not properly scaled or labeled.
2. May oversimplify complex relationships in the data.

Conclusion

This assignment focused on **exploratory data analysis (EDA)** and **data visualization** using Pandas, Matplotlib, and Seaborn. We created scatter plots, bar plots, and box plots to uncover meaningful patterns in the dataset. By applying these techniques, we gained a better understanding of the relationships between various student admission factors. These visualization skills will be essential for future data analysis and machine learning tasks.