Project Report: Exploratory Data Analysis (EDA) on AMCAT Data

1. Introduction

The AMCAT (Aspiring Minds Computer Adaptive Test) is a widely used employability assessment test that measures a candidate's skills in various domains such as quantitative aptitude, logical reasoning, and verbal ability. This report focuses on performing Exploratory Data Analysis (EDA) on AMCAT data to gain insights into the performance and employability skills of candidates. The purpose of this project is to identify trends, patterns, and key factors that contribute to employability.

2. Objectives

- To understand the distribution of scores in various sections of the AMCAT test.
- To identify correlations between different skill sets.
- To find patterns between demographic features and test performance.
- To identify factors contributing to employability based on the test.

3. Data Description

The AMCAT dataset consists of the following columns:

- **Candidate ID**: Unique identifier for each candidate.
- **Gender**: Male or Female.
- **Location**: Geographical location of the candidate.
- **Quantitative Score**: Score obtained in the Quantitative Ability section.
- **Logical Score**: Score obtained in the Logical Reasoning section.
- **Verbal Score**: Score obtained in the Verbal Ability section.
- **Overall Score**: Combined score based on performance across all sections.
- **Employability Status**: Whether the candidate is considered employable based on AMCAT criteria (Yes/No).
- **Educational Qualification**: Highest education level of the candidate (e.g., Graduate, Postgraduate).

4. Steps Involved in EDA

4.1 Data Cleaning

- **Missing Values**: Checked for missing values in the dataset and handled them using mean, median, or mode where appropriate.
- **Outlier Detection**: Identified and treated outliers in the score columns using boxplots and Z-score analysis.
- **Data Type Conversion**: Ensured that categorical variables like Gender and Location were correctly encoded.

4.2 Univariate Analysis

- **Quantitative Score Distribution**: Plotted histograms to observe the distribution of scores in Quantitative Ability.

- **Logical and Verbal Scores**: Similar histograms and boxplots were used for these sections to assess their distribution and variability.
- **Gender Proportion**: Visualized the proportion of male and female candidates using pie charts or bar charts.

4.3 Bivariate Analysis

- **Correlation Analysis**: Investigated the relationship between Quantitative, Logical, and Verbal scores using a heatmap and Pearson correlation coefficient.
- **Score Comparison by Gender**: Analyzed the performance of male and female candidates across different sections using bar plots and boxplots.
- **Employability vs. Scores**: Explored the relationship between employability status and scores in each section through bar graphs and scatter plots.

4.4 Multivariate Analysis

- **Employability Prediction**: Used pairplots to visualize the interaction between multiple features like Quantitative Score, Logical Score, and Verbal Score in relation to employability status.
- **Education Level Impact**: Investigated how educational qualification affects scores and employability status using stacked bar charts.

5. Findings and Insights

- **Score Distribution**: Most candidates scored in the mid-range (40-70 percentile) for all sections, indicating an average performance in the dataset.
- **Gender-Based Performance**: Males and females performed similarly in the Logical and Verbal sections, but a slight male advantage was observed in the Quantitative section.
- **Correlation of Scores**: Moderate positive correlation between Logical and Quantitative scores (~0.4), indicating candidates who are strong in one area tend to perform better in the other.
- **Employability Factors**: Higher scores in all sections significantly improved the likelihood of employability, with Quantitative ability being a strong predictor.
- **Education Level Impact**: Graduates performed better in the Verbal section, while Postgraduates excelled in Quantitative reasoning.

6. Conclusion

The EDA revealed several important trends in the AMCAT data. Employability is closely linked to performance across all sections, with the Quantitative and Logical scores being particularly significant. Educational qualifications and gender were also notable factors in influencing performance. This analysis provides a solid foundation for deeper machine learning models or predictive analytics to forecast employability based on test scores and demographic factors.

7. Future Work

- Perform predictive modeling using the AMCAT data to predict employability status.

- Explore deeper insights using clustering techniques to segment candidates based on performance and employability potential.
- Investigate time-series patterns if data over multiple test attempts is available.

8. Tools Used

- **Python Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.
- **Jupyter Notebook** for performing the analysis and visualization.

9. References

- AMCAT official documentation for score interpretation.
- Research papers on employability prediction using AMCAT data.