INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

About Me

I am Vijaykumar Pandurang Mangore. I have completed my graduation through Mechanical Engineering. I have worked with two start-up group and was doing business in partnership in green energy sector. My role and responsibility were checking viability and feasibility of different projects like solar installation, waste management & EV. I have completed my Business Management training in IIT Bombay. I have completed CBAP certified course. I have played role in decision making, problem solving and managing the projects.

I am particularly drawn to the field of Data Science for its unparalleled potential to revolutionize industries and society at large. The ability to extract actionable insights from vast troves of data inspires me, as I believe informed decision-making is the cornerstone of progress. I am excited about the prospect of leveraging cutting-edge tools and techniques to uncover hidden patterns and drive impactful outcomes.

As I continue to evolve in my career, I am committed to embracing new challenges and pushing myself to new heights. My goal is to not only contribute meaningfully to the ever-evolving landscape of technology but also to inspire others to embark on their own journey of exploration and discovery.

1. Business Problem and Use Case Domain Understanding

The business problem for this project revolves around understanding and gaining insights from a dataset related to job applicants or employees. The dataset contains various attributes such as ID, salary, educational qualifications, performance metrics, and personality traits. The use case domain involves human resources management, recruitment, and talent acquisition. The objective is to explore the data, identify patterns, and derive meaningful insights to aid decision-making processes within the organization.

2. Objective of the project

The primary aim of this analysis is to extract insights from the provided dataset, focusing particularly on understanding the relationship between various features and the target variable, which is Salary.

Our specific goals include:

- Comprehensive description of the dataset and its features.
- Identification of patterns or trends within the data.
- Exploration of relationships between independent variables and Salary.
- Detection of outliers or anomalies in the dataset.

3. Summary of the Data

The dataset, titled Aspiring Mind Employment Outcome 2015 (AMEO), curated by Aspiring Minds, centres around employment outcomes for engineering graduates. It encompasses dependent variables such as Salary, Job Titles, and Job Locations, alongside standardized scores in cognitive, technical, and personality skills. With approximately 4000 data points and40 independent variables, the dataset presents a mix of continuous and categorical data. Moreover, demographic features and unique candidate identifiers enrich the dataset.

Exploratory Data Analysis

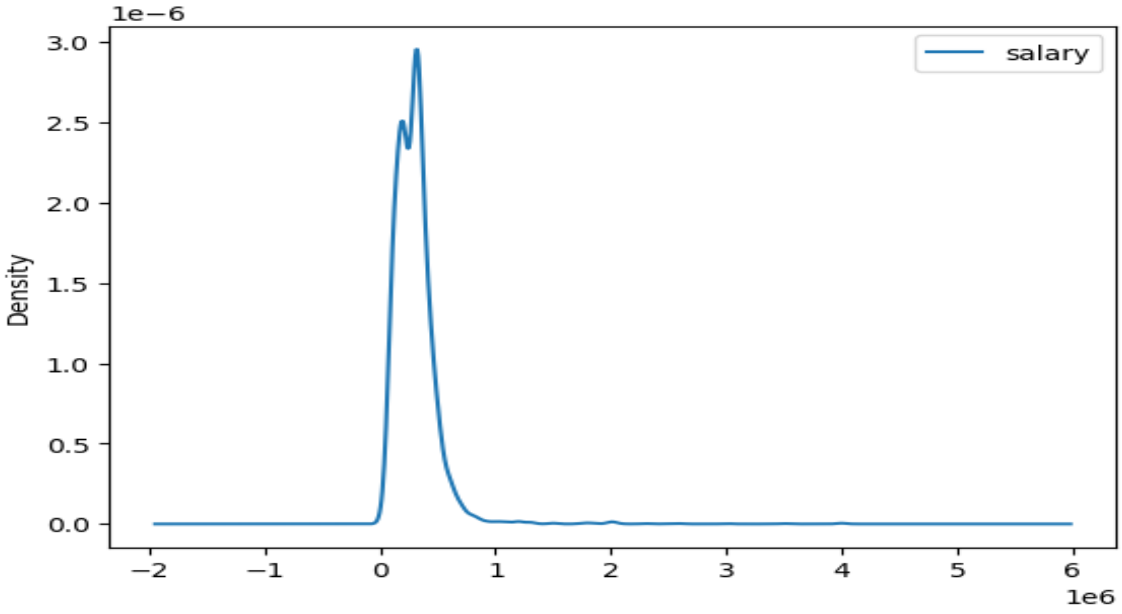
1. Data Cleaning and Pre-processing

- **Datatype Conversion:** The 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) fields were converted to date time objects. 'Present' values in the DOL field were replaced with the end date of the survey (2024-02-17).
- **Aggregating Categories:** The dataset was streamlined to include only the top 10 most frequent categories within specific columns. Additional categories were grouped under 'Other' to simplify analysis.

2. Univariate Analysis

Univariate Analysis is a type of data visualization where we visualize only a single variable at a time. Univariate Analysis helps us to analyze the distribution of the variable present in the data so that we can perform further analysis.

2.1 Salary and Probability Distribution:



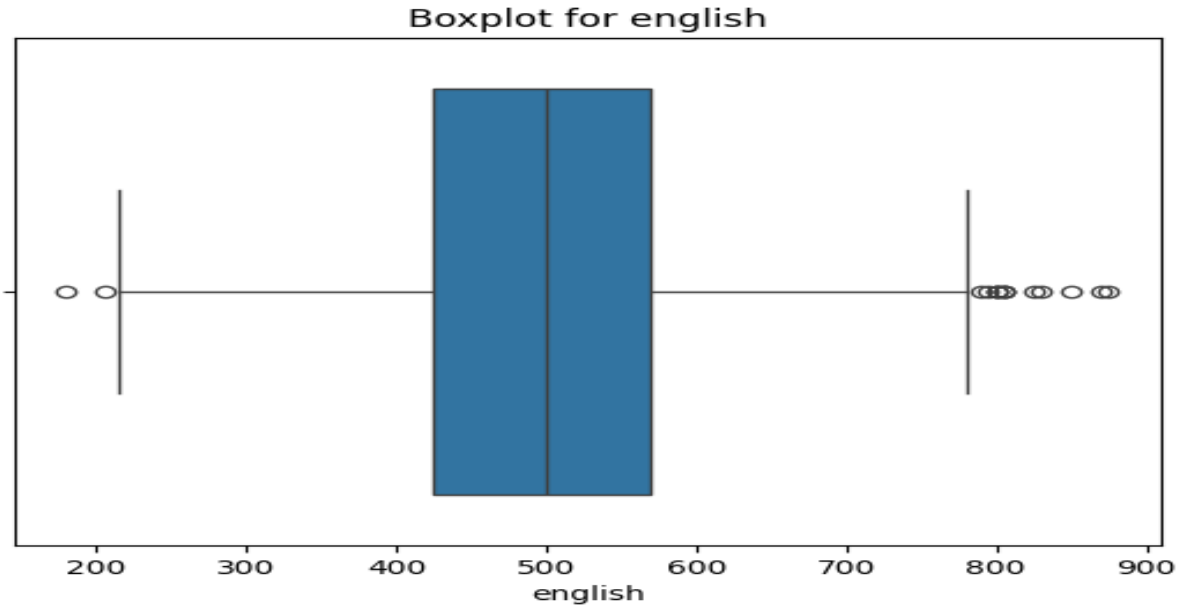
The plot shows the probability density function of salary values, allowing for a clear understanding of how salaries are distributed across different ranges. The x-axis represents salary values, while the y-axis indicates the density of observations at each salary level. This visualization aids in identifying patterns such as the central tendency and variability of salaries

within the dataset, providing valuable insights for further analysis and decision-making.

2.2 Box plot for English

1. **Median and Interquartile Range**: The median score appears to be around 500, with the interquartile range (IQR) falling between approximately 400 and 600. This shows that the central 50% of the data lies within this range.
2. **Outliers**: There are several outliers on both the lower and upper ends of the distribution. These are the data points that fall outside the whiskers, which extend approximately 1.5 times the IQR from the first and third quartiles. Outliers suggest that some students have either very low or very high scores compared to the majority.
3. **Skewness**: The data appears slightly skewed towards higher scores, as indicated by the presence of more outliers on the upper side.

In summary, the box plot reveals that most students have "english" scores between 400 and 600, with a few outliers on both extremes.



2.3 The pair plot

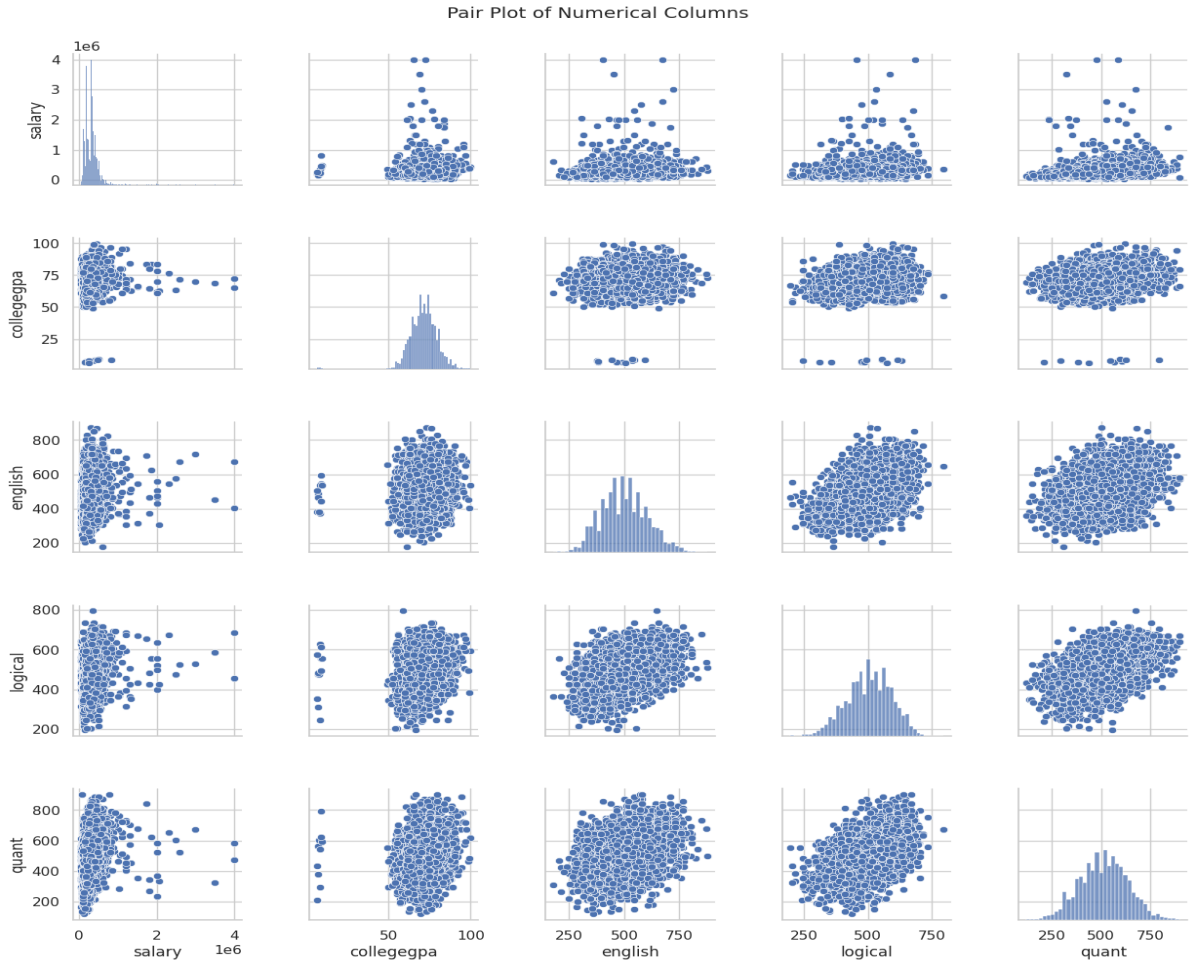
The pair plot shows the relationships between several numerical columns, such as **salary**, **college GPA**, **English**, **logical**, and **quant**. Here's a summary of insights:

1. **Salary**:
    - The salary distribution is right-skewed, with a few high outliers.
    - There doesn't appear to be a clear linear correlation between salary and other variables like GPA, English, logical, or quant scores.

INNOMATICS RESEARCH LABS

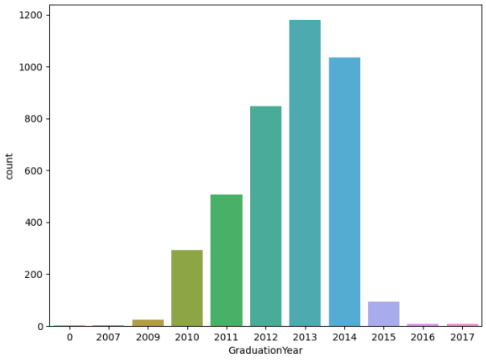
2. **College GPA**:
    - The GPA distribution seems to be fairly normal.
    - There is some moderate positive correlation between GPA and English, logical, and quant scores, as shown by the oval-shaped scatterplots.
3. **English, Logical, and Quant Scores**:
    - English, logical, and quant scores are positively correlated with each other. This is evident from the scatter plots between these variables, which show upward trends.
4. **Overall**:
    - The pair plot suggests a strong relationship between academic performance (GPA) and skills like English, logical, and quantitative reasoning.
    - However, the relationship between salary and these variables is less clear, indicating that other factors may play a larger role in determining salary.

In summary, the plot reveals strong inter-correlations among the skill-based variables (English, logical, quant) but less of a direct connection between these skills and salary.



2.3 College GPAs Distribution:

A count plot using Seaborn to visualize the distribution of graduation years within the dataset. By specifying the "Graduation Year" column from the DataFrame df, the count plot displays the frequency of each graduation year category.

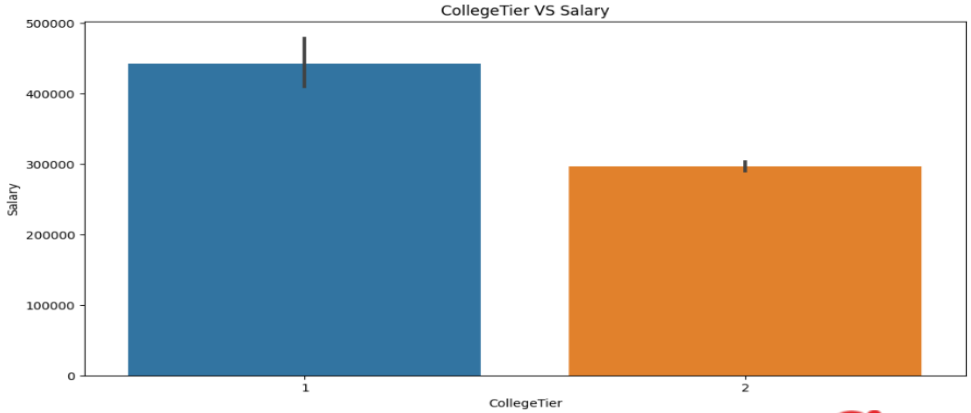


3. Bivariate Analysis

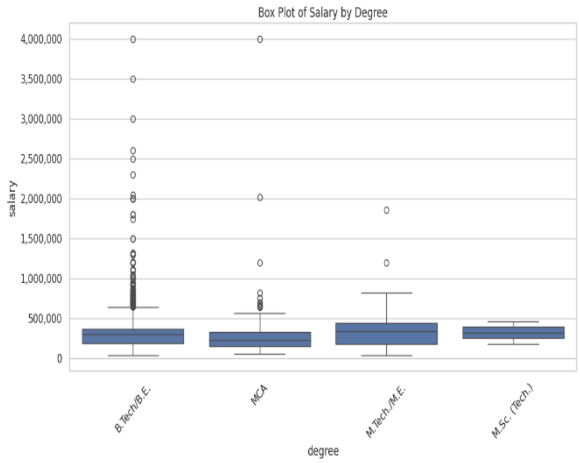
Bivariate analysis is the simultaneous analysis of two variables. It explores the concept of the relationship between two variable whether there exists an association and the strength of this association or whether there are differences between two variables and the significance of these differences.

3.1 College Tier and Salary:

Higher Average Salary for Tier 1 Colleges: Graduates from Tier 1 colleges have a significantly higher average salary (around 450,000) compared to those from Tier 2 colleges (approximately 300,000).
Smaller Variability in Tier 2: The error bar (likely representing the confidence interval or standard deviation) for Tier 2 colleges is smaller, indicating less variability in the salary data for Tier 2 graduates.
Conclusion: Graduates from Tier 1 colleges tend to receive higher salaries than their counterparts from Tier 2, suggesting that the reputation or resources available at Tier 1 institutions could positively influence salary outcomes.



Box Plot of salary by degree



Conclusion: B.Tech/B.E. and M.Tech/M.E. graduates appear to have better salary prospects than MCA and M.Sc. (Tech) graduates, with B.Tech/B.E. having the most variability and potential for high earnings. This suggests that job market demand may favor engineering-related degrees over other technical specializations in terms of salary.

Conclusion

The analysis offers valuable insights into the dataset, uncovering relationships, patterns, and trends. While certain factors like tenure and college tier influence salary, others such as gender and academic scores show minimal correlation. This suggests a complex interplay of factors affecting salaries in the given context. Additionally, the analysis sheds light on gender-based specialization preferences, providing actionable insights for recruitment strategies. This report serves as a foundation for further investigations and decision-making processes within the organization, guiding future analytical eneavors.

•