





Project Report

On

**Exploratory Data Analysis of AMEO Dataset**

Prepared By – Shahbaz Akhtar

**About Me**

I am an MBA graduate with highly passionate in learning new techniques evolving around. I have worked in corporate as well as business sector.

And with my gained experience I have learnt about the inconsistency in the fields and how to deal around with it. With the development in the AI world, it really motivated me to pursue the Data Science course. With the tons of information available, it is helpful in generating the insights for better decision making. The world is dependent on data driven approach, and its shaping the society. I have worked in so many projects with Python enabled, Sql, and Tableau. I am dedicated to give my contributions towards this field. commendable to embrace it.



**Objective**

The aim is to gain insights and understanding from the dataset, particularly focusing on the relationship between the features and the target variable, which is Salary. The goal of this analysis includes:

• Describing the dataset and its features comprehensively.

• Identifying any patterns or trends present in the data.

• Exploring the relationships between independent and target variables (Salary).

• Identifying any outliers or anomalies in the data

**Summary Of Data**

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses on employment outcomes for engineering graduates. It includes dependent variables such as Salary, Job Titles, and Job Locations, along with standardized scores in cognitive skills, technical skills, and personality skills.

With around 40 independent variables and 4000 data points, these variables encompass both continuous and categorical data. The dataset also includes demographic features and unique identifiers for each candidate.

**Data Cleaning And Preprocessing**

1. **Datatype Conversion**

To ensure the accuracy and consistency of our analysis, we converted the data types of the 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) fields from their original format to datetime objects. Given that the survey was conducted in 2015, the assumption was made that respondents who indicated their status as 'present' for DOL had left the company by the latest survey date, which was recorded as 2024-02-17. Therefore, we replaced the 'present' values in the DOL field with this end date.

The data handling process has been successfully completed. Firstly, the columns '10board', '12board', 'GraduationYear', 'JobCity', and 'Domain' were processed to replace null values represented by 0 or -1. Following this, columns with over 80% -1 values, including 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', and 'CivilEngg', were removed from further analysis. Lastly, for the remaining optional subject columns, 'ElectronicsAndSemicon' and 'ComputerScience', -1 values were replaced with 0, indicating that the subjects were not pursued.

1. **Collapsing Categories**

Under this process, the dataset has been refined to encompass solely the top 10 most frequent categories within specific columns. Any categories beyond this selection have been categorized as other. This approach aims to streamline the dataset, focusing solely on the most prevalent categories for subsequent analysis.



**FEATURE ENGINEERING**

**Tenure Calculation:**

Another new feature, 'tenure', has been introduced by subtracting the 'Date of Leaving' (DOL) from the 'Date of Joining' (DOJ). This indicates the duration of an individual's employment within the company.

**Graduation Year Filtering:**

Rows where the graduation year is greater than or equal to the date of joining have been removed. This ensures data integrity by excluding instances where the graduation year suggests a date after the individual's employment start date.

**EXPLORATORY DATA ANALYSIS**

**Univariate analysis**

A Univariate Analysis has been done on the columns individually with visualizations of Box plot, Histograms, Pdf.

**Bivariate analysis**

**B**ivariate Analysis has been done between the two columns like Salary and Tenure, Salary and Designation

**Research Outcomes**

The analysis commences by categorizing the dataset according to job titles and computing the mean and standard deviation of salaries for each role. This examination offers insights into the distribution of salaries across various job designations. Notably, Software Engineers exhibit the highest mean salary and standard deviation, implying both higher earnings and greater variability in pay in comparison to Programmer Analysts and Associate Engineers. Subsequently, a one-sample t-test is executed for each job designation to compare their average salary with an anticipated range. Results indicate that for Programmer Analysts and Software Engineers, there is substantial evidence to refute the null hypothesis, indicating significant discrepancies in their salaries from the anticipated range. Conversely, for Hardware Engineers and Associate Engineers, insufficient evidence is found to reject the null hypothesis, suggesting that their salaries may not significantly diverge from the expected range. In sum, these analyses furnish valuable insights into the salary distributions among different job roles and facilitate an understanding of the significance of salary differentials within the dataset.

Top of Form





Conclusion

By updating the Flask route to handle GET requests and updating the empty form action with action and method. form data for POST requests, the bug was been successfully resolved, and the application now functions as intended. The improvements has contributed to an efficient user experience