INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

Yogapriya P

About Me

I am Yogapriya, a recent master's graduate specializing in Bioinformatics. My passion lies in leveraging data-driven insights to make a significant impact. Eager to embark on an exciting journey in the field of Data Science, I am actively seeking an entry-level position that offers challenges and continuous learning opportunities.

My motivation originates from a strong desire to contribute to a company's mission and goals, aspiring to become a valuable asset by consistently refining my skills and capabilities. Proficient in Microsoft Excel, Tableau, SQL Server, and other essential data analysis tools, I am dedicated to achieving excellence in data-driven decision-making.

With a robust academic foundation and an unwavering commitment to skill development, I am a dynamic candidate prepared to contribute to the ever-evolving landscape of Bioinformatics and Data Science.

Link to Project Repo: https://github.com/YogapriyaPeethambaram/Exploratory-Data-Analysis-on-AMEO-Dataset

A. OBJECTIVE

This analysis endeavors to extract insights and comprehension from the given dataset, with a primary focus on unraveling the intricate relationships between diverse features and the target variable—Salary. The overarching objectives of this exploration encompass:

1. Comprehensive Dataset Description:

Provide a thorough and nuanced description of the dataset, elucidating the nature and characteristics of its features.

2. Pattern and Trend Identification:

Unearth and articulate any discernible patterns, trends, or recurring motifs embedded within the dataset.

3. Relationship Exploration:

Investigate and delineate the relationships between the independent variables and the target variable, Salary. This involves scrutinizing how various features interplay and potentially influence salary outcomes.

4. Outlier Detection:

Employ robust methodologies to identify and characterize outliers within the dataset. Outliers, if present, can significantly impact the overall analysis and subsequent insights.

B. SUMMARY

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, curated by Aspiring Minds, provides a comprehensive exploration of employment outcomes for engineering graduates. The dataset comprises essential dependent variables, including Salary, Job Titles, and Job Locations, alongside standardized scores gauging cognitive skills, technical proficiency, and personality traits. Boasting approximately 40 independent variables and encapsulating around 4000 data points, this dataset encapsulates a rich variety of information. The variables exhibit a diverse range, encompassing both continuous and categorical data, facilitating a nuanced analysis of the factors influencing employment outcomes.

C. DATA CLEANING

pd.to_datetime() to convert the 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) columns to DateTime objects.

Ensuring the accuracy and consistency of our analysis, we performed essential data type conversions on the 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) fields. Both were transformed from their original formats to datetime objects, facilitating standardized date calculations and comparisons.

In handling the 'Date of Leaving' (DOL) field, we encountered instances where respondents indicated their status as 'present.' To align with our analysis and account for this, we made an assumption based on the survey date of 2015. We considered that individuals marked as 'present' in the DOL field had left the company by the latest survey data, recorded as 2024-02-17. Consequently, we replaced these 'present' values in the DOL field with this designated end date, ensuring uniformity in our datRemoves the specified columns from the DataFrame. (`columns_to_drop` containing the names of the columns you want to remove: 'Unnamed: 0', 'ID', 'CollegeID', 'CollegeCityID'.)

Checking missing values and duplicates.

D. FEATURE ENGINEERING

Replace 'get' with the mode of the 'Designation' column

An additional column representing age has been incorporated into the dataset by subtracting the year of birth (DOB) from 2015, reflecting the individual's age as of 2015.

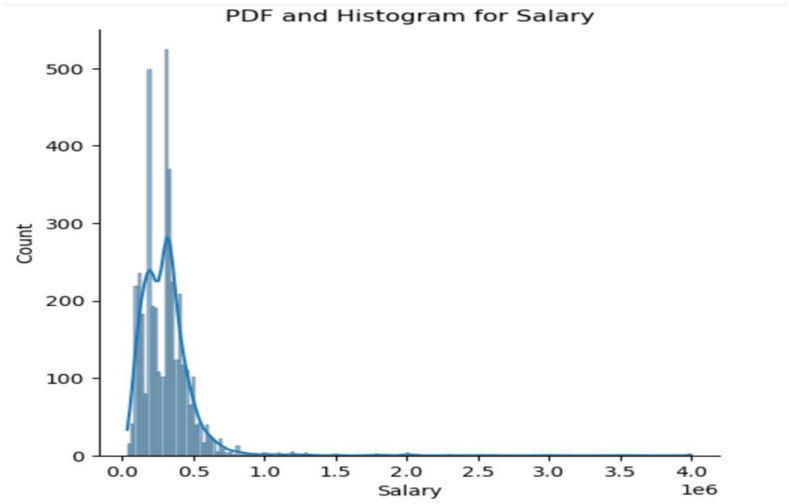
E. EXPLORATORY DATA ANALYSIS

1. UNIVARIATE ANALYSIS

1.1 CONTINUOS FEATURE

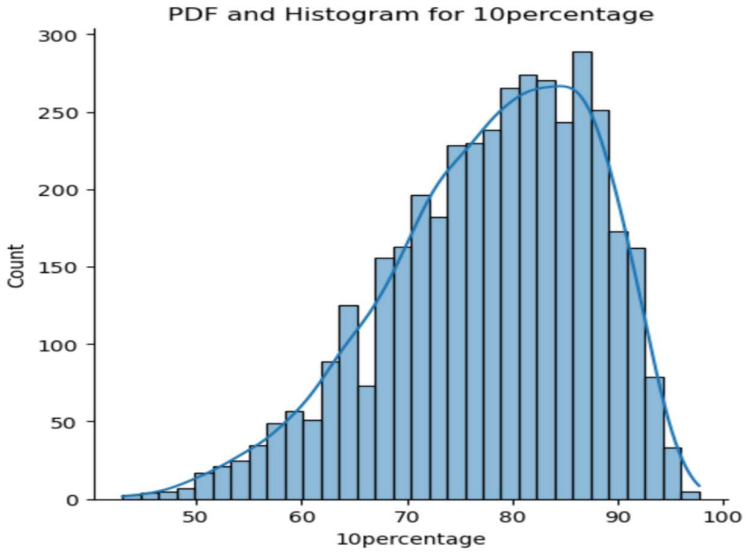
1.1.1 SALARY

The Probability Density Function (PDF) plot shows the distribution of salaries. The majority of salaries seem to be concentrated in the lower range, with a peak around a specific value. The Histogram confirms the concentration of salaries in certain ranges, and it highlights the count of individuals within each salary bin.



1.1.2 10 PERCENTAGE

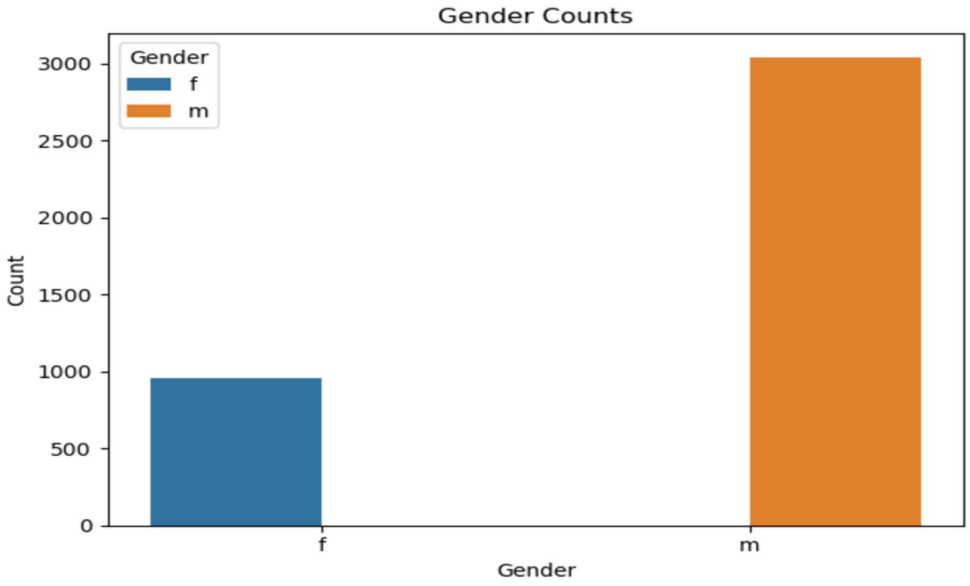
The Probability Density Function (PDF) plot shows the distribution of 10th percentages. The Histogram provides a visual representation of the count of individuals within different 10th percentage ranges.



1.2 CATEGORICAL FEATURES

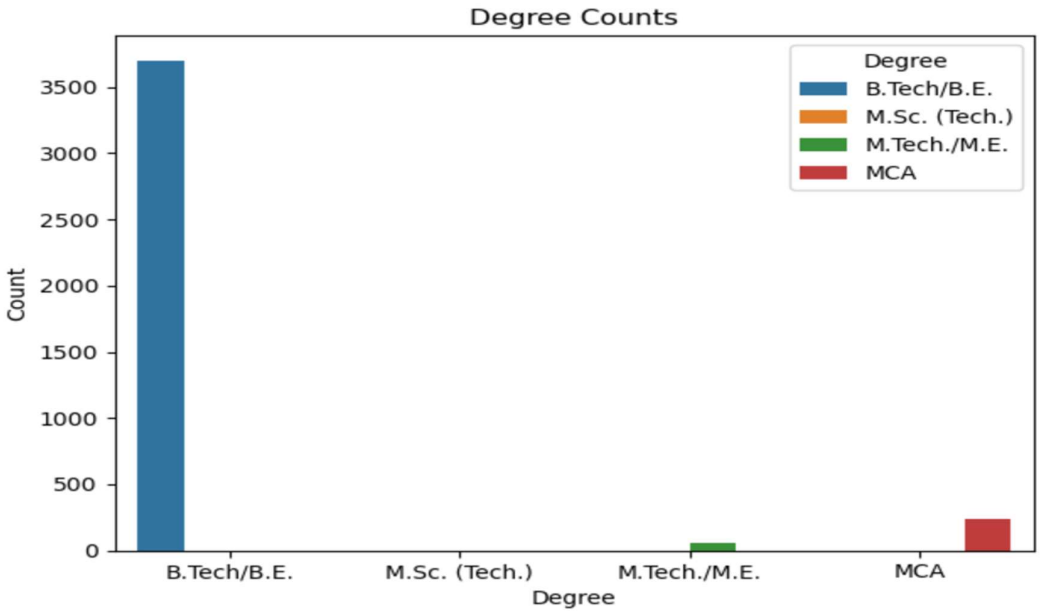
1.2.1 GENDER

The countplot illustrates the distribution of genders within the dataset.The dataset consists of two main gender categories: 'M'(Male) and 'F'(Female).The count of 'Male' individuals appears to be higher than the count of 'Female' individuals.



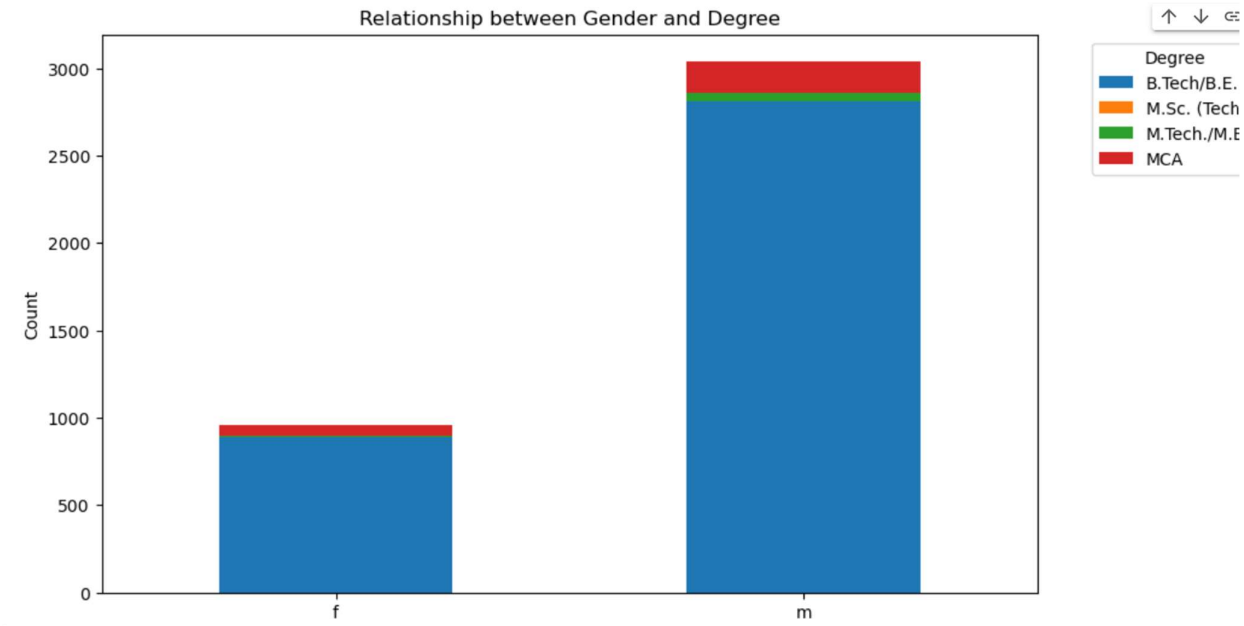
1.2.2 DEGREE

The count plot illustrates the distribution of different degrees within the dataset. The dataset encompasses various degree categories, including 'B.Tech/B.E,' 'MCA,' 'M.Tech/M.E,' and 'Ph.D.'The most predominant degree appears to be 'B.Tech/B.E,' which has the highest count. Other degrees such as 'MCA' and 'M.Tech/M.E' also contribute to the dataset. The count of individuals with a 'Ph.D.' is noticeably lower compared to other degree categories. This suggests that the dataset is primarily composed of individuals with undergraduate and postgraduate degrees, with a smaller representation of Ph.D. holders.



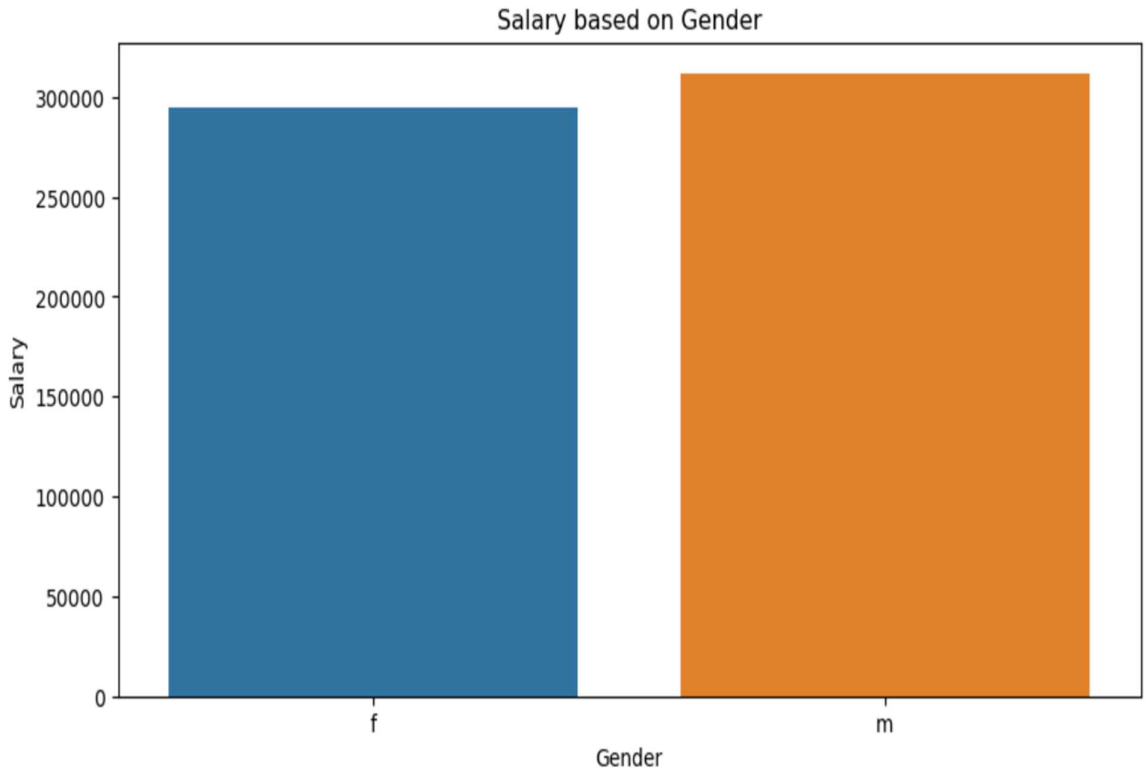
2. BIVARIATE ANALYSIS
2.1 GENDER AND AGE

The chart breaks down the distribution of different degrees for both genders. 'B.Tech/B.E' is the most prevalent degree for both males and females.'MCA' and 'M.Tech/M.E' also have substantial representation for both genders. The chart highlights potential gender disparities in certain degree categories. For example, there are more males with 'B.Tech/B.E' and 'M.Tech/M.E' degrees compared to females. The 'MCA' category shows a more balanced distribution between genders.



2.2 SALARY AND GENDER

The plot shows the average salary for both male (m) and female (f) individuals. On average, males appear to have higher salaries compared to females in the dataset. The salary range for both genders varies, indicating diversity in income levels. Males have a wider salary distribution, suggesting a broader range of salaries within the male group.



2.3 SALARY AND AGE

The data points are scattered across various age groups, showcasing the diversity in ages within the dataset. Most individuals are concentrated in the age range between 20 and 40. Salaries vary widely across different age groups, indicating that age alone may not be the sole determinant of salary. Some individuals in the younger age group receive high salaries, while some older individuals have comparatively lower salaries. There isn't a clear linear trend between age and salary, suggesting that factors other than age significantly contribute to variations in salary.

F. CONCLUSION

Data Understanding: You've provided an overview of the dataset, focusing on employment outcomes and standardized scores. Mentioned the initial dimensions of the dataset (3998 rows, 39 columns). Noted the presence of duplicate values and initiated data manipulation.

Data Manipulation: Highlighted the removal of redundant rows and columns. Acknowledged the need for handling missing values (NaN).

Data Visualization: Conducted univariate analysis with PDFs, histograms, box plots, and count plots. Executed bivariate analysis using scatterplots and bar plots.

THANK YOU