INNOMATICS
RESEARCH LABS

INNOVATION. AUTO**MAT**ION. ANALY**TICS**

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

About Me

I am an ambitious and driven individual currently pursuing a Bachelor's degree in Computer Science & Engineering. With a proactive approach and a knack for problem-solving, I am deeply passionate about exploring the intersections of technology and innovation. My journey in computer science & engineering has fueled my interest in leveraging data to drive meaningful insights and foster progress. With experience in full-stack and IoT development, I've gained proficiency in both front-end and back-end development, as well as a strong understanding of Internet of Things applications. Proficient across multiple programming languages, I thrive on navigating intricate challenges and devising inventive solutions.

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.

INNOMATICS
RESEARCH LABS

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 and 40 independent variables, the dataset presents a mix of continuous and categorical data. Moreover, demographic features and unique candidate identifiers enrich the dataset.

INNOMATICS
RESEARCH LABS

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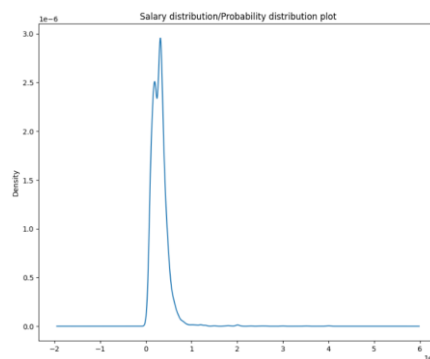
The plot shows the probability density function of salary values, allowing for a clear understanding of how salaries are distributed across different ranges. The x-axis represents salary values, while the y-axis indicates the density of observations at each salary level. This visualization aids in identifying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

INNOMATICS
RESEARCH LABS

2.2 College GPAs Distribution:

A histogram to illustrate the distribution of college GPAs within the dataset. By selecting the "CollegeGPA" column from the DataFrame and plotting it as a histogram, the code visually represents the frequency of different GPA ranges among the dataset's entries.
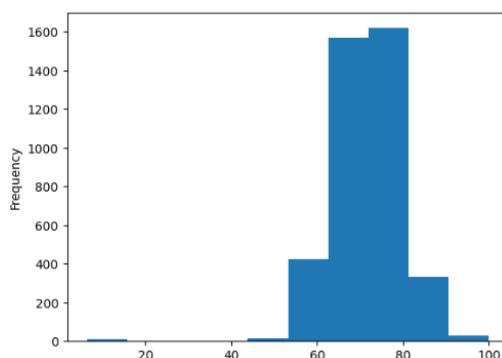
2.3 College GPAs Distribution:

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.
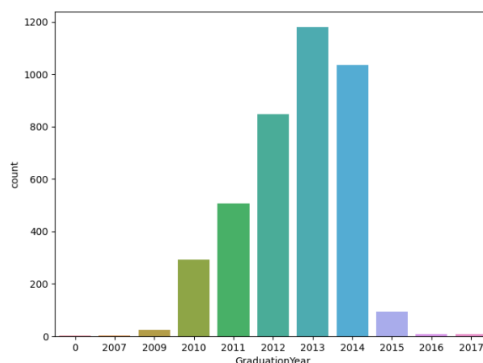

3. Bivariate Analysis

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.
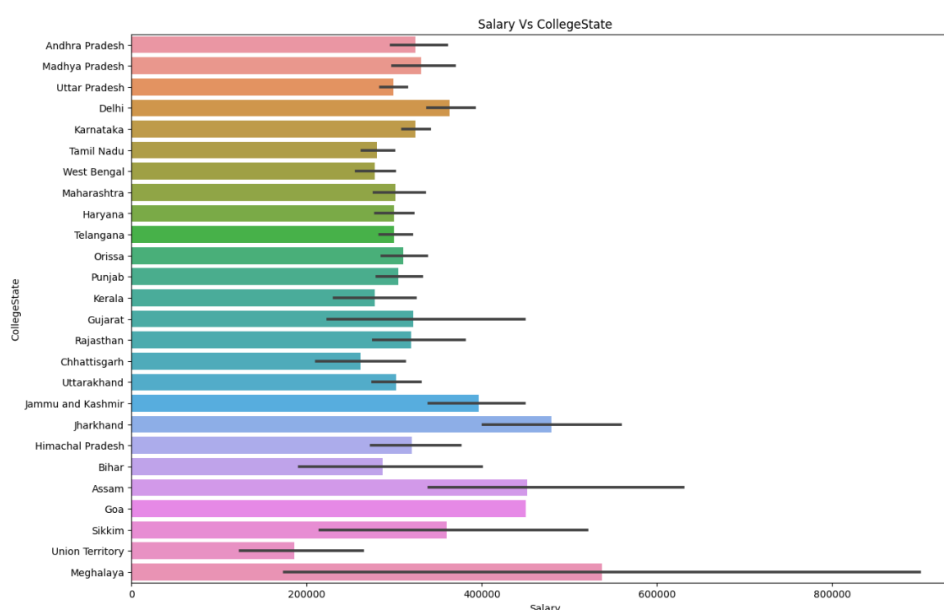
3.1 College Tier and Salary:

a bar plot using Seaborn to compare the average salary based on the college tier from which individuals graduated. By specifying the "Salary" variable on the y-axis and the "CollegeTier" variable on the x-axis, the plot illustrates the average salary for each college tier category.
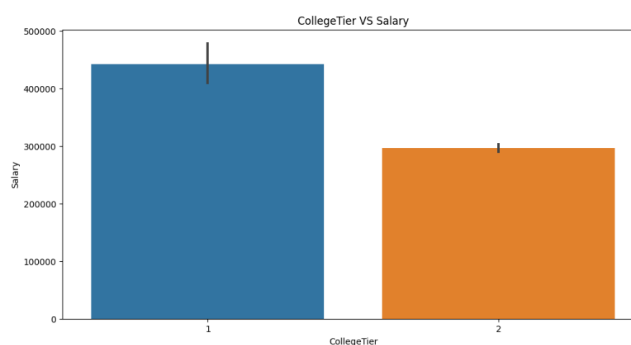
3.2 Salary and College State:

A bar plot using Seaborn to compare the average salary based on the college state where individuals graduated. The y-axis represents the different college states, while the x-axis indicates the corresponding average salary for each state category. The title "Salary Vs College State" summarizes the comparison being made. This visualization facilitates the exploration of potential variations in salary across different states, providing insights into regional disparities in earning potential among graduates.

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 endeavors.

INNOMATICS
RESEARCH LABS