**EXPLORATORY DATA ANALYSIS (EDA) REPORT**

**INTRODUCTION**

The objective of this Exploratory Data Analysis (EDA) report is to analyze the provided dataset containing information about individuals' employment and educational background. Through this analysis, we aim to gain insights into various aspects of the dataset, identify patterns, detect outliers, and address specific research questions.

The dataset provided for analysis encapsulates a wealth of information concerning individuals' professional and educational backgrounds. With its multitude of attributes ranging from salary and tenure to academic achievements and specialization, this dataset presents an opportunity to glean valuable insights into various aspects of workforce dynamics and educational trends. Through this Exploratory Data Analysis (EDA), our objective is to delve into the dataset's intricacies, uncover patterns, detect outliers, and address specific research inquiries.

By employing a combination of statistical techniques and data visualization, we aim to elucidate the relationships between different variables, validate claims, and provide actionable insights for stakeholders in human resources, education, and related fields. This analysis holds the promise of uncovering nuanced trends, facilitating evidence-based decision-making, and informing future research endeavors in the realms of employment dynamics and educational trajectories.

**DATASET OVERVIEW**

The dataset comprises rows and columns, with each row representing an individual and each column representing specific attributes such as ID, salary, date of joining (DOJ), date of leaving (DOL), designation, job city, gender, date of birth (DOB), educational qualifications, and other relevant information.

1. Unnamed: 0': An unnamed or index column, possibly representing row numbers.
2. 'ID': Unique identifier for each individual or candidate.
3. 'Salary': The salary of the individual.
4. 'DOJ': Date of joining the job.
5. 'DOL': Date of leaving the job.
6. 'Designation': Job title or position of the individual.
7. 'JobCity': The city where the job is located.
8. 'Gender': Gender of the individual.
9. 'DOB': Date of birth of the individual.
10. '10percentage': Percentage obtained in 10th grade.
11. '10board': Board of education for 10th grade.
12. '12graduation': Year of graduation from 12th grade.
13. '12percentage': Percentage obtained in 12th grade.
14. '12board': Board of education for 12th grade.
15. 'CollegeID': Unique identifier for the college.
16. 'CollegeTier': Tier of the college.
17. 'Degree': Degree obtained by the individual (e.g., B.Tech/B.E., MCA, M.Tech./M.E.).
18. 'Specialization': Specialization in the degree.
19. 'collegeGPA': Grade Point Average obtained in college.
20. 'CollegeCityID': Unique identifier for the city where the college is located.
21. 'CollegeCityTier': Tier of the city where the college is located.
22. 'CollegeState': State where the college is located.
23. 'GraduationYear': Year of graduation.
24. 'English', 'Logical', 'Quant': Scores in different aptitude tests.
25. 'Domain': A measure of skills relevant to the job.
26. 'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg': Scores or information related to specific engineering disciplines.

Numerical columns are unnamed:0, ID, Salary, 10Percentage, 12graduation, 12percentage, CollegeID, collegeGPA, CollegeCityID, GraduationYear, English, logical, quant, domain, computerProgramming, electronicsandsemicon, computerscience,MechanicalEngg, ElectricalEngg, TelecomEngg, Civilengg, Consientiousness, agreeableness, extraversion, neuroticism, openness\_to\_experience.

Categorical columns are Designation, JobCity, Gender, 10board, 12board, Degree, Specialization, CollegeTier, CollegeCityTier, CollegeState

**OBJECTIVE**

The primary objective is to analyze and derive insights from the dataset to gain a better understanding of the factors influencing salary, job performance, and other relevant aspects for job candidates.

Potential tasks and analyses could include:

Exploratory Data Analysis (EDA): Understanding the distribution of key variables, identifying trends, and exploring relationships between different features.

Salary Prediction: Developing a model to predict salary based on various attributes.

Factors Influencing Job Performance: Investigating correlations between academic performance, skills, and personality traits with job performance or salary.

Demographic Analysis: Examining patterns based on gender, location, and other demographic factors.

Feature Importance: Determining the most significant features impacting job-related outcomes.

By addressing these objectives, we can provide valuable insights for recruitment processes, talent management, and overall understanding of the factors contributing to success in a professional environment.

**UNIVARIATE ANALYSIS**

Univariate analysis is a statistical method used to analyze and summarize data with a single variable. It focuses on examining the distribution, central tendency, and dispersion of the data within that variable. The main goal is to describe and understand the characteristics of one variable at a time.

Common techniques for univariate analysis include:

1. Descriptive Statistics: Measures such as mean, median, mode, and standard deviation provide a summary of the central tendency and spread of the variable.
2. Histograms and Frequency Distributions: These graphical representations illustrate the distribution of values in a variable, allowing you to identify patterns, trends, or outliers.
3. Boxplots: Boxplots provide a visual summary of the distribution's central tendency and spread, highlighting outliers and the interquartile range.
4. Probability Density Function (PDF): The PDF represents the likelihood of different outcomes occurring in a continuous variable, giving insights into the probability distribution.
5. Cumulative Distribution Function (CDF): The CDF shows the cumulative probability of observing values less than or equal to a given point in a variable.

**Observations:**

* The income of the most people lies inbetween 10,000 and 70,000.
* 10th percentage of most people is inbetween 70 to 90 perecentage.
* 12th percentage of people lies inbetween 60 to 90 percentage.
* College GPA of the people is almost in a skew manner.
* Each person has different College ID's.
* most of the people completed their 12th grade between 2007 and 2010.
* Score of people in English is in a skew manner.
* Score In their respective subjects is very poor.
* Salary column contains outliers because most of the people's earnings lies between 20,200 to 70,000 but very few people earn more than that.
* Very few people scored less than 50% in 10th class.
* Most of the people scored more than 40% in 12th grade.
* Almost all people graduated after year 2000.
* Many people scored average marks in English.
* Many people scored average marks in Logical subject.
* Many people scored average marks in Light blue subject.
* Almost all people are good at their respective domains.
* Many people scored average marks in computer programming.
* Many people are bad at Electronics and semicon.
* Many people are bad at Computerscience.
* Many people are bad at Mechamical Engineering.
* Only some people are good at Electrical engineering.
* Only some people are good at Telecom engineering.
* Majority of the people in our data are males.
* Majority of the people are from B.Tech background.
* Majority of the people studied Electronics and Communication Engineering.
* Majority of the people are from Uttar Pradesh.

**BIVARIATE ANALYSIS**

Bivariate analysis involves the simultaneous analysis of two variables to understand the relationships between them. The primary goal is to explore how changes in one variable may be associated with changes in another. This type of analysis is crucial for uncovering patterns, correlations, and dependencies in data. Here are some common techniques used in bivariate analysis:

1. **Scatter Plots:** A scatter plot is a graphical representation of the relationship between two continuous variables. Each point on the plot represents a pair of values, allowing you to visually identify patterns and trends.
2. **Correlation Coefficient:** The correlation coefficient (e.g., Pearson's correlation) quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).
3. **Categorical vs. Continuous Variable Analysis:** Techniques like box plots, violin plots, or bar charts are used when comparing the distribution of a continuous variable across different categories of a categorical variable.
4. **Crosstabulation (Contingency Tables):** This method is suitable for analyzing the relationship between two categorical variables. It helps to understand the frequency distribution and associations between categories.
5. **Correlation Matrix:** For datasets with multiple variables, a correlation matrix provides a comprehensive view of the relationships between all pairs of variables.
6. **Regression Analysis:** Regression models are employed to quantify the relationship between a dependent variable and one or more independent variables. Simple linear regression involves one independent variable, while multiple linear regression deals with more than one.

**Observations:**

* In the plot between salary and computerprogramming, we can observe that the people who are good at programming are getting decent salary. and also we can clearly say that from the graph that majority of the people who have average programming skills are getting average salary.
* In the plot salary and English, we can observe that the people who are good at English are getting decent salary. and also we can clearly say that from the graph that majority of the people are speaking intermediate English who are getting average salary.
* In the graph between gender and salary, we can clearly observe that males and females are equally paid. Even though the males are more in number, that doesn't reflect here in the salary.
* I don't think salary depends upon Specialization. People can get paid based on their knowledge and degree in their respective domain. We can clearly see that in the above graph.
* By observing the plot between salary and degree, we can clearly say that the people who did M.SC. are paying highest salaries compared to others and M.Tech./M.E. is at secong place and then comes B.Tech./B.E. and MCA.
* If we keep Gender aside, then most people choose Electronics and communication Engineering and then Computer Science and Engineering is at second place and then the number is gradually decreasing for other courses irrespective of gender. If we see instrumentation and control engineering, the number of males and females are almost equal, if we see biotechnology, females are more than males and in some cases like chemical engineering, mechanical and automation, there are no female students.
* Compared to all Degrees, many people choose B.Tech/B.E. and the second one is MCA. In that also males are majority people compared to females. No female person competed M.Tech./M.E. or M.Sc.
* The plot between collegetier and gender suggests that the majority of the entries in the "CollegeTier" column have a value of 2, with a smaller number having a value of 1 and ofcourse males are in more number than females.

**RESEARCH**

* As Times Of India stated, if a person do their specialization in Computer Science and Engineering and take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer then they can get upto 2.5 to 3 lakhs or more than that.
* I don't think there is any relationship between gender and specialization because in the provided dataset the number of females are very less. Although the females are are very few in number, they opted almost all the branches provided as same as males. So I would say there is no relation between gender and specialization.
* There is relation between salary and educational background like if a person stops his education after B.Tech, then he will get less salary compared to the person who did M.Tech/M.Sc. Salary doesn't depend upon the specialiation you have choosen, it only depends upon the knowledge you have in that specific domain.
* Majority of the employees are working in Bangalore. And the salary of the employee is based on the designation and not on the city where they are working and ofcourse cost of living changes from city to city.
* Studying in different colleges impact the Jobs of the people and salaries. Students studying at top colleges are paying good salaries and people studying in normal colleges are paying average salaries. It doesn't occur most of the times.
* In the given dataset, there is no correlation between personality traits like 'conscientiousness', 'agreeableness', 'extraversion', 'neuroticism', 'openness\_to\_experience' and job performance metrics like 'English', 'Logical', 'Quant', 'Domain', etc..

**CONCLUSION**

The dataset includes diverse information, ranging from personal details such as gender and date of birth to academic and professional metrics like educational qualifications, salary, and job-related attributes. The majority of individuals in the dataset have engineering degrees, and there's a mix of specializations across various engineering disciplines. The salary column exhibits a wide range of values, suggesting diversity in income levels among the individuals in the dataset. There are some instances of extreme values or potential outliers, as indicated by the box plot. Academic performance metrics, such as 10th and 12th percentages, vary among individuals. The distribution provides insights into the educational background of the dataset.

Preliminary correlation analysis reveals potential relationships between certain variables, such as positive correlations between academic performance metrics and certain job-related skills. The dataset includes personality trait scores (conscientiousness, agreeableness, extraversion, neuroticism, openness\_to\_experience), providing a psychological dimension to the analysis. Further analysis could delve into specific research questions, such as evaluating the impact of specialization on career trajectory, assessing the correlation between personality traits and job performance, and investigating the relationship between city and salary.

In conclusion, this dataset offers a rich source of information about individuals' educational and professional backgrounds, providing a foundation for exploring various aspects of career paths, academic achievements, and industry preferences. Further analyses and domain-specific questions can help derive more nuanced insights.