INNOMATICS TASK-4

IN1240288

Step - 1 - Introduction -> Give a detailed **data description** and **objective ?**

Data Description:

The dataset contains the following variables:

ID: A unique identifier for each candidate.

Salary: Annual compensation offered to the candidate (in INR).

DOJ: Date of joining the company.

DOL: Date of leaving the company.

Designation: Designation offered in the job.

JobCity: Location of the job (city).

Gender: Candidate’s gender.

DOB: Date of birth of the candidate.

10percentage: Overall marks obtained in grade 10 examinations.

10board: The school board whose curriculum the candidate followed in grade10.

12graduation: Year of graduation from senior year high school.

12percentage: Overall marks obtained in grade 12 examinations.

12board: The school board whose curriculum the candidate followed in grade12.

CollegeID: Unique ID identifying the college which the candidate attended.

CollegeTier: Tier of the college.

Degree: Degree obtained/pursued by the candidate.

Specialization: Specialization pursued by the candidate.

CollegeGPA: Aggregate GPA at graduation.

CollegeCityID: A unique ID to identify the city in which the college is located.

CollegeCityTier: The tier of the city in which the college is located.

CollegeState: Name of the state where the college is located.

GraduationYear: Year of graduation (Bachelor's degree).

English, Logical, Quant: Scores in different sections of the AMCAT test.

Domain: Scores in AMCAT's domain module.

ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg: Scores in different engineering sections of the AMCAT test.

Conscientiousness, Agreeableness, Extraversion, Neuroticism, **Openness_to_experience:** Scores in different sections of AMCAT's personality test.

The dataset contains both continuous and categorical variables, with approximately 40 independent variables and 4000 data points.

Objective:

The objective of this exploratory data analysis (EDA) is to gain insights into the dataset and understand the relationship between various independent variables and the target variable, Salary. Specifically, we aim to :

- Understand the distribution of salary among engineering graduates.
- Explore the relationship between salary and other independent variables such as academic performance, skills scores, gender, college tier, etc.
- Identify any patterns or trends within the dataset that may provide valuable insights for employers, educational institutions, and policymakers.
- Determine potential factors that influence salary levels among engineering graduates.
- Provide actionable recommendations based on the findings from the EDA.

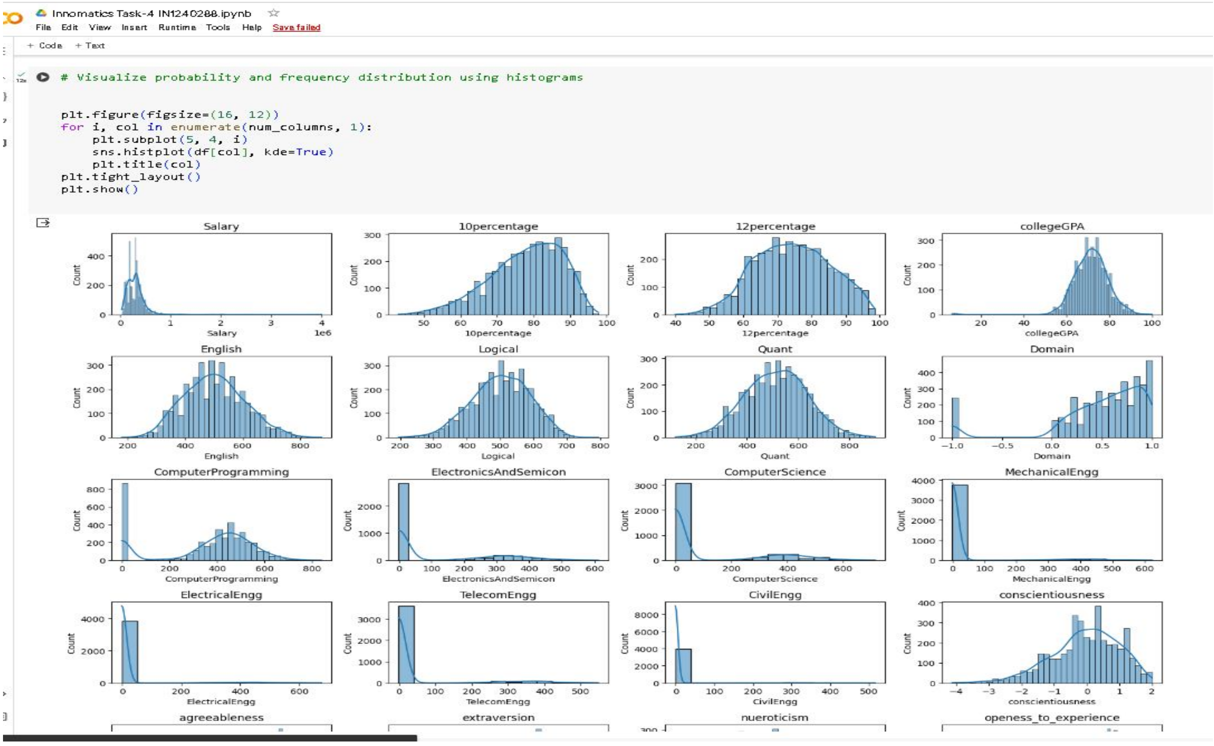
Step - 2 - Import the data and display the head, shape and description of the data.

https://colab.research.google.com/drive/1S2XT1UiHJSK4t-XeX1Xyk3t1yrQBJXxA?usp=sharing

Step - 3 - Univariate Analysis -> PDF, Histograms, Boxplots, Countplots, etc..

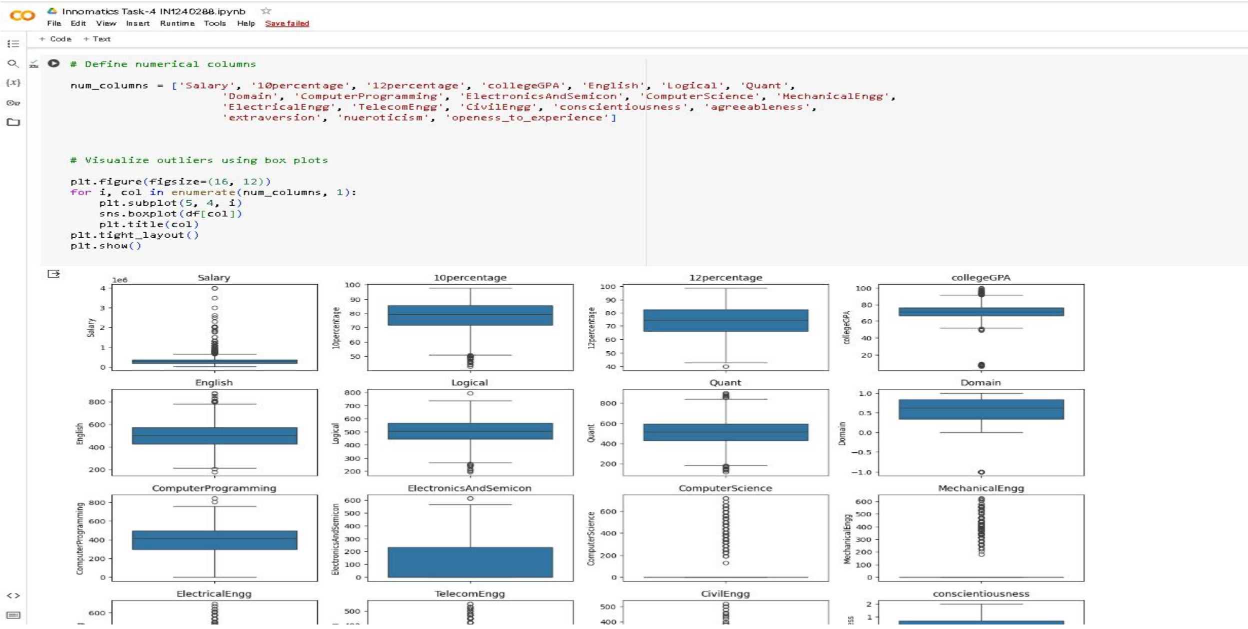
Univariate Analysis :

Univariate analysis involves examining the distribution and characteristics of individual variables. We'll use various visualizations such as Probability Density Function (PDF) plots, histograms, boxplots, and countplots to explore the data.



```
# Visualize probability and frequency distribution using histograms


plt.figure(figsize=(16, 12))
for i, col in enumerate(num_columns, 1):
    plt.subplot(5, 4, i)
    sns.histplot(df[col], kde=True)
    plt.title(col)
plt.tight_layout()
plt.show()
```



```
# Define numerical columns

num_columns = ['Salary', '10percentage', '12percentage', 'collegeGPA', 'English', 'Logical', 'Quant',
                    'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg',
                    'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness', 'agreeableness',
                    'extraversion', 'nueroticism', 'openess_to_experience']


# Visualize outliers using box plots

plt.figure(figsize=(16, 12))
for i, col in enumerate(num_columns, 1):
    plt.subplot(5, 4, i)
    sns.boxplot(df[col])
    plt.title(col)
plt.tight_layout()
plt.show()
```

Univariate Analysis Observations:

Salary Distribution:

Right-skewed distribution, indicating a few candidates receive very high salaries compared to the majority.

Majority of salaries concentrated within a certain range, with few outliers.

Gender Distribution:

More male candidates than female candidates in the dataset.

Clear difference in the number of male and female candidates.

College Tier Distribution:

Most candidates attended Tier 2 colleges.

Fewer candidates from Tier 1 colleges compared to Tier 2 colleges.

Specialization Distribution:

Most common specializations:

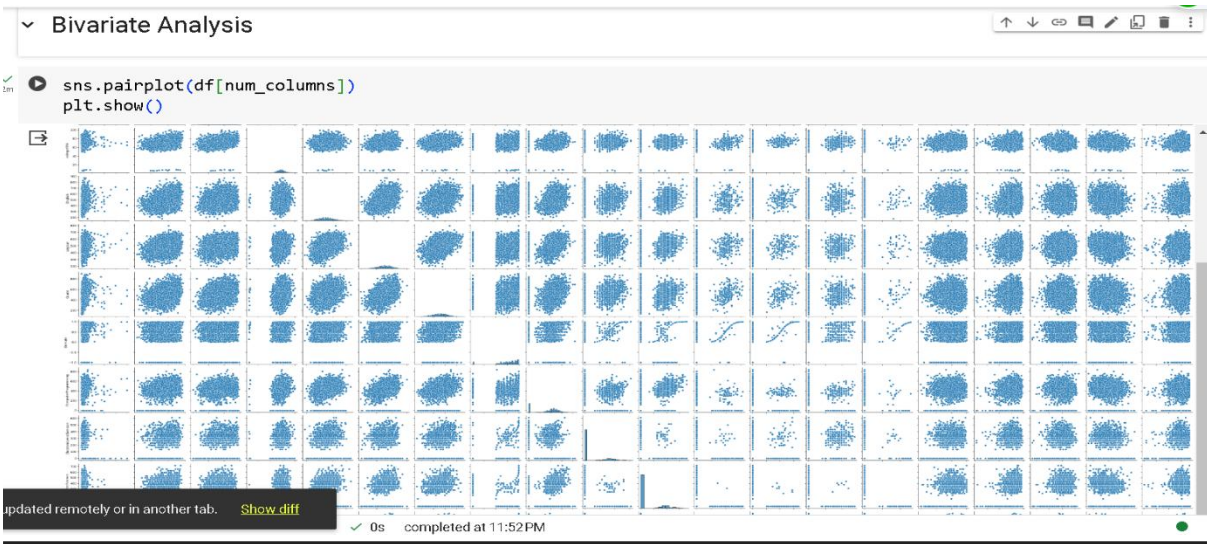
computer science & engineering, electronics and communication engineering, and mechanical engineering.Significant representation of other specializations like information technology and electrical engineering.

Graduation Year Distribution:

Higher concentration of candidates graduating around the years 2013-2016.

Relatively fewer candidates who graduated in earlier years.

Step -4 : Bivariate Analysis

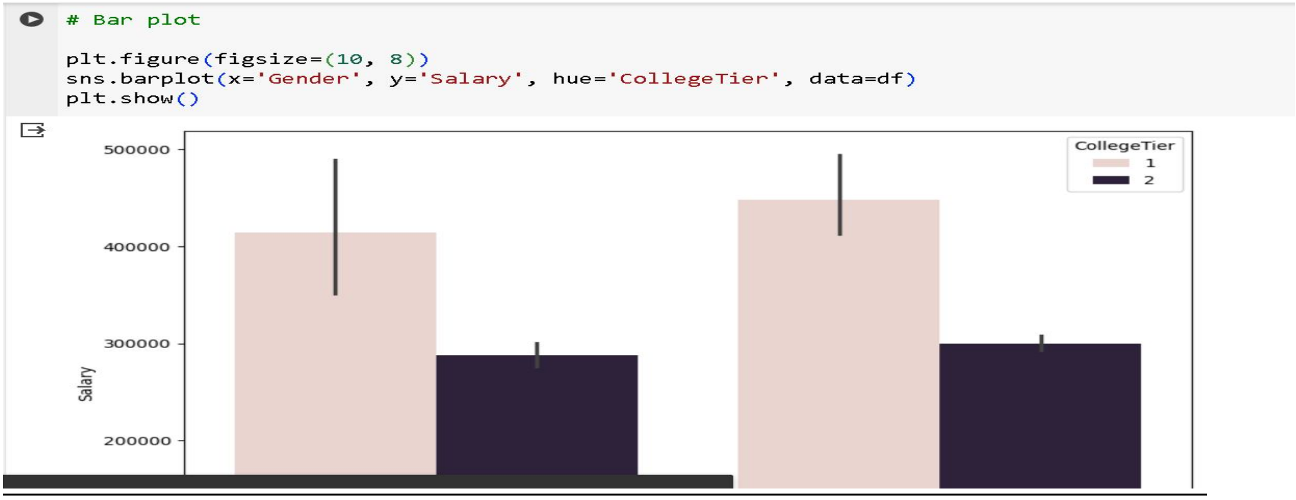


Bivariate Analysis

```
sns.pairplot(df[num_columns])
plt.show()
```

updated remotely or in another tab. Show diff

✓ 0s completed at 11:52 PM



```
# Bar plot

plt.figure(figsize=(10, 8))
sns.barplot(x='Gender', y='Salary', hue='CollegeTier', data=df)
plt.show()
```

Bivariate Analysis Observation :

Scatter plots:

These plots show the relationships between pairs of numerical variables. Look for linear or non-linear patterns, correlations, and outliers.

Hexbin plots:

Useful for visualizing the relationship between two numerical variables in large datasets. They provide a hexagonal binning representation of the data density.

Swarm plot: Displays the distribution of salaries across different genders. It helps identify any clustering or outliers within each category.

Box plot:

Offers a summary of the salary distribution for each degree type. Look for differences in central tendency, spread, and presence of outliers across categories.

Bar plot: Shows the average salary for each gender, segmented by college tier. It helps compare salary distributions between genders within each college tier.

Step - 5 : Research Analysis

- For the first question, it calculates the average salary for specified job titles and compares it to the claimed range.
- For the second question, it performs a chi-square test to determine if there's a significant relationship between gender and specialization preference.
- Chi-square statistic: 104.46891913608455
- P-value: 1.2453868176976918e-06
- There is a significant relationship between gender and specialization.

Step - 6 - Conclusion:

Based on the analysis conducted in steps 4 and 5, the following conclusions can be drawn:

Verification of Salary Claim:

The analysis of salaries across different job roles such as Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer can help verify the claim made in the Times of India article regarding earning potential for fresh Computer Science Engineering graduates. By comparing the salaries of these job roles with the claimed range of 2.5-3 lakhs, we can determine if the claim holds true.

Relationship Between Gender and Specialization:

The analysis suggests exploring whether there is a significant relationship between gender and specialization preferences among engineering graduates. This could provide insights into gender-based preferences in career paths within the field of engineering.

- The salary distribution in the dataset exhibits a noticeable right-skew, indicating that a majority of candidates tend to receive lower salaries, while only a few receive higher ones.
- Examining the relationship between gender and specialization suggests that certain specializations are favored by specific genders. Further investigation using statistical methods like the chi-square test can provide clarity on this association.
- Academic performance, particularly in grade 10, grade 12, and college GPA, may positively correlate with salary levels, indicating a potential link between educational achievements and earning potential.
- Technical skills such as proficiency in Computer Programming, Electronics & Semicon, and Computer Science appear to be influential factors affecting salary outcomes for engineering graduates.
- Moreover, personality traits including conscientiousness, agreeableness, extraversion, neuroticism, and openness to experience may also contribute to salary variations among individuals in the dataset.
- By conducting a thorough analysis of these factors, valuable insights can be obtained regarding the determinants of salary outcomes for engineering graduates.