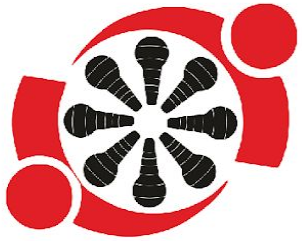


INNOMATİCS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

About me

I'm currently pursuing my B.Tech degree at IIT Jodhpur.
I love solving technical problems, doing research, and coming up with new ideas in technology. I'm great at working with teams and meeting new people because I'm friendly and outgoing. I learn new things quickly and can handle stress well.
I'm interested in Data Science because it can change how different industries work. I'm curious about finding important information in big sets of data and using it to make smart decisions and create new things. Data Science is like having a special toolbox to find patterns and connections in information
My internship experience at 7-Eleven GSC Bangalore exposed me to the practical applications of analytics and development, fueling my passion further. Learning data science isn't just a career choice; it's a personal journey to grasp the power of information. Through this pursuit, I aim to contribute meaningfully to solving real-world problems.

Linkedin URL: https://www.linkedin.com/in/shruthika-polkam
GitHub URL: https://github.com/Shruthika08

INNOMATICS RESEARCH LABS

Objective and dataset description:

Introduction:
Exploratory Data Analysis (EDA) on the AMEO dataset focusing on Salary as the target variable.

Dataset Overview:
The dataset, released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO), is centered on employment outcomes of engineering graduates. It includes variables such as Salary, Job Titles, and Job Locations, alongside standardized scores in cognitive, technical, and personality skills. Demographic features are also included. With approximately 40 independent variables and 4000 data points, it encompasses both continuous and categorical variables, each candidate being assigned a unique identifier.

Objectives:
1. Provide a comprehensive understanding of the dataset.
2. Determine the relationship between various variables and Salary.
3. Identify and address outliers.
4. Uncover patterns in the data through univariate analysis.

Exploratory Data Analysis

1. **Understanding the dataset**:
   We first visualise the dataset, by looking at the head, null values, datatypes, duplicates and unique values in each column and their respective column names.

2. **Data cleaning and manipulation:**

- Drop the unnecessary columns, change the dtype of DOB, DOJ, DOJ.
- Calculate the age and tenure by subtracting the required columns. {creating new and useful features. }
- Drop the rows with tenure -1, which indicates that the DOL<DOJ (which maybe a normal human error.)
- Check if people entered cgpa instead of percentage and convert it accordingly.
- Check wherever 0's or -1's are present in different columns and then replace it with nan since they are unknown.

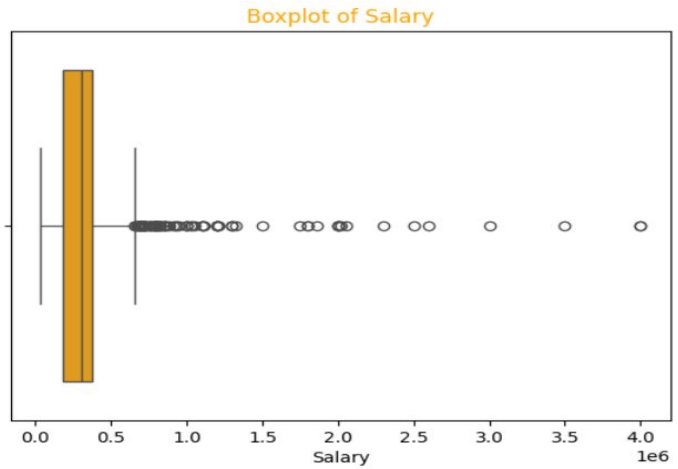
Domain 6.11
ComputerProgramming 21.80
ElectronicsAndSemicon 71.45
ComputerScience 77.49
MechanicalEngg 94.06
ElectricalEngg 96.06
TelecomEngg 90.60
CivilEngg 98.94

INNOMATICS RESEARCH LABS

- These are the % null values after imputing the nan values, here we can drop the columns with % greater than 90% since most of the rows in that column are not defined.
- Now for categorical columns replace the nan values with mode and with median for the numerical columns.
- Convert all the content of df to lowercase to maintain uniformity.
- Now we have to separate the top 10 columns with more occurrences and then group everything else into ‘other’ category, this will make our analysis simpler.

3. Univariate analysis:

- Numerical features: (Plot PDF, Histogram, box_plot and CDF)
  - **Salary**:
    - Mean and median are nearly identical, indicating that the data is relatively symmetrically distributed around the center.
    - The histogram displays positive skewness, suggesting that the majority of salaries are clustered towards the lower end, with a few higher salaries skewing the distribution to the right.
    - The boxplot reveals the presence of outliers, particularly those with higher salaries compared to the rest of the data.
    - The Cumulative Distribution Function (CDF) demonstrates slight deviation from the normal distribution, indicating that the data may not perfectly follow a normal distribution but is fairly close to it.



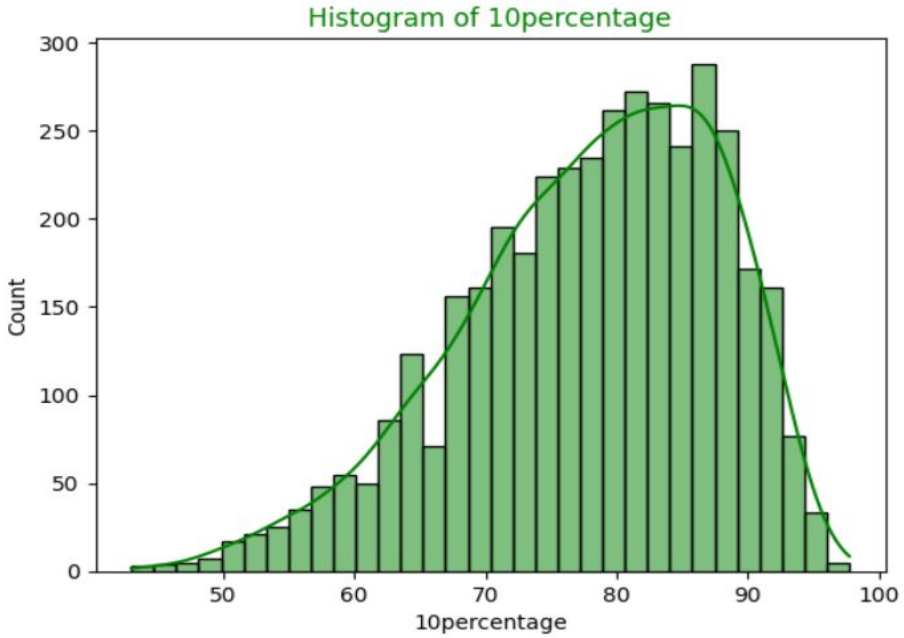
10 Percentage Observations:

Descriptive statistics reveal that approximately 50% of students scored less than 80%.

The histogram illustrates that the majority of students scored between 70-90%, with fewer students achieving lower percentages.

The box plot indicates the presence of outliers among students who scored very high percentages.

The Cumulative Distribution Function (CDF) shows some skewness and doesn't precisely adhere to a normal distribution.



12 Percentage Observations:

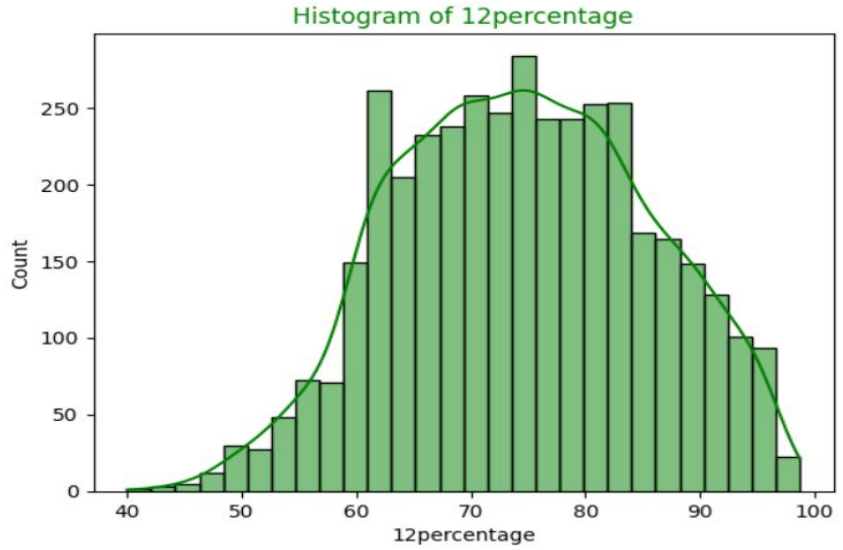
Descriptive statistics indicate that around 50% of students scored less than 75%.

Mean and median exhibit minimal difference, suggesting relatively symmetrical distribution.

The histogram illustrates that the majority of students scored between 60-85%.

The box plot reveals only one student considered as an outlier, scoring around 40%.

The Cumulative Distribution Function (CDF) demonstrates the data closely follows a normal distribution.



INNOMATICS RESEARCH LABS

College GPA Observations:

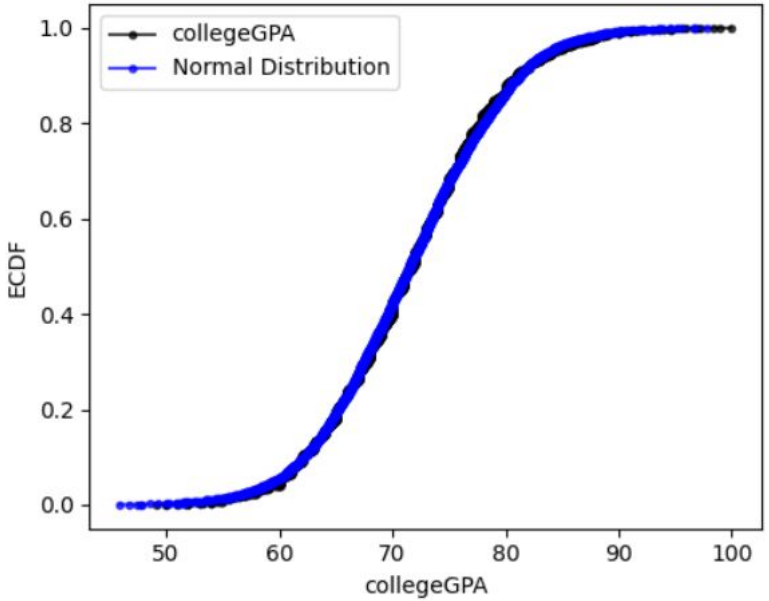
Descriptive statistics reveal that 75% of students scored less than 75%.

Mean and median are identical, indicating a symmetrical distribution.

The histogram displays values ranging from 65-80%.

The box plot indicates the presence of considerable outliers at both low and high extremes.

The Cumulative Distribution Function (CDF) shows that the data is normally distributed.



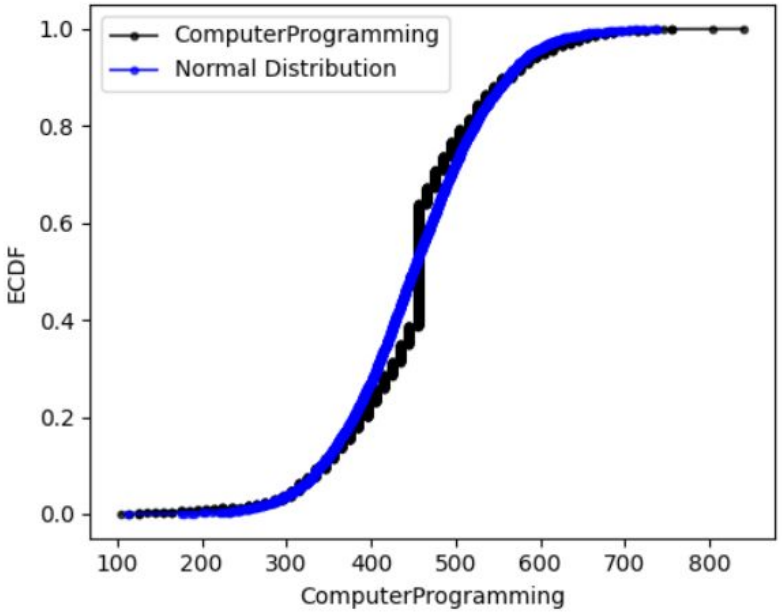
Computer Programming Observations:

Mean and median exhibit minimal difference, indicating a balanced distribution.

The histogram illustrates that the majority of people scored around 450.

The box plot indicates sufficient representation of people at both higher and lower extremities.

The Cumulative Distribution Function (CDF) suggests the data follows a distribution close to normal.



INNOMATICS RESEARCH LABS

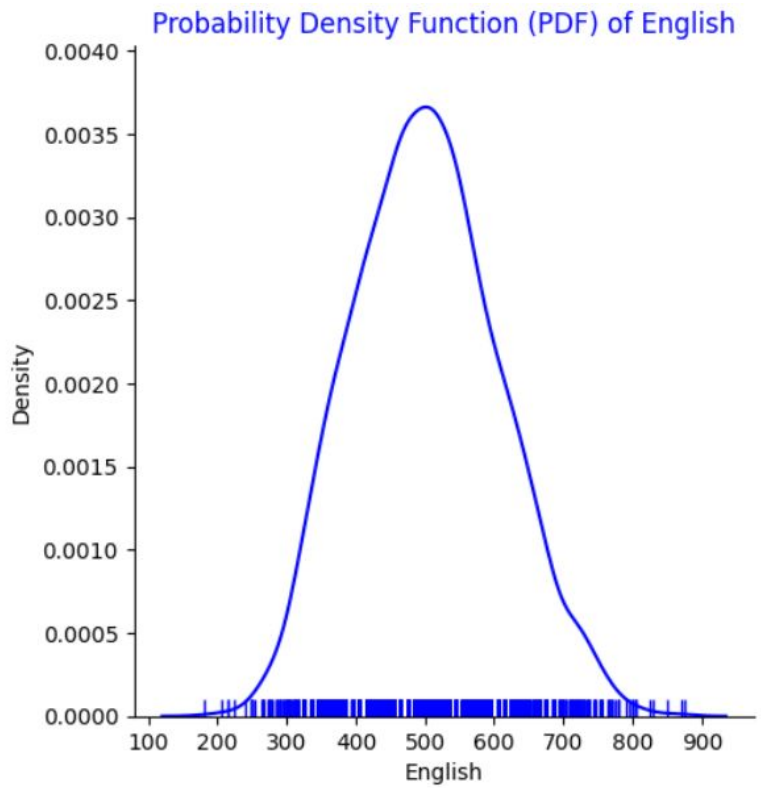
English Observations:

Mean and median are nearly identical.

Scores fall within the range of 390 to 550.

The box plot indicates the presence of both higher and lower extreme values.

The Cumulative Distribution Function (CDF) follows a normal distribution.



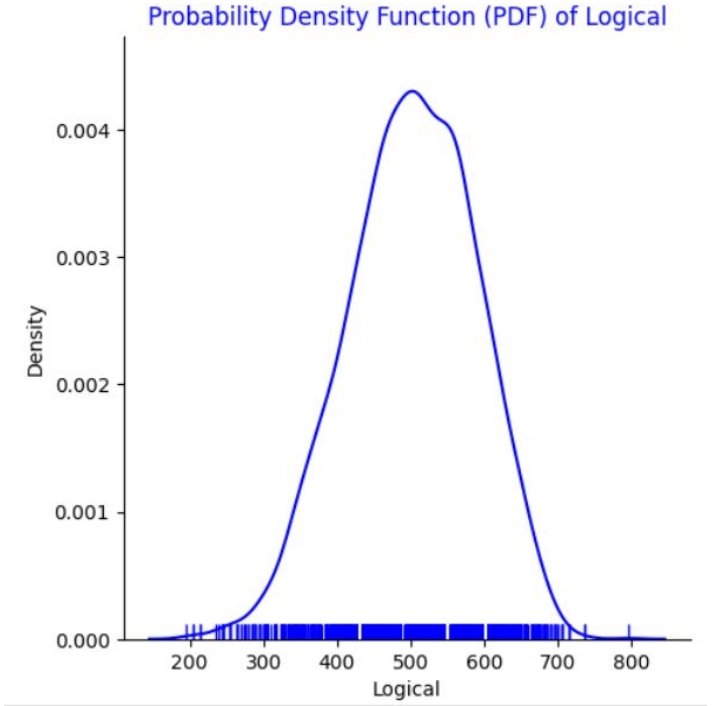
Logical Observations:

Mean and median exhibit minimal difference.

Scores range from 450 to 590.

Lower extreme values are present, with only one higher extreme value.

The Cumulative Distribution Function (CDF) conforms to a normal distribution.



INNOMATICS RESEARCH LABS

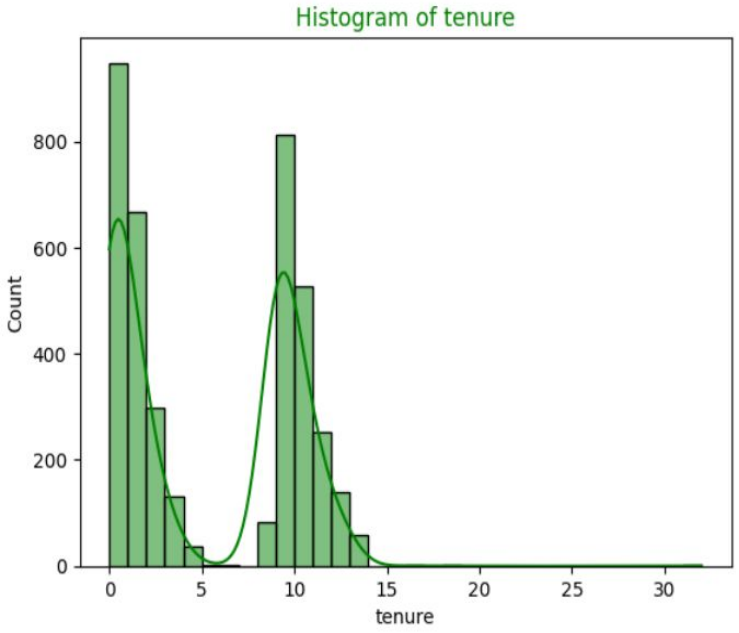
Quant Observations:

Mean and median are nearly identical.

Scores range from 425 to 610.

Both lower and higher extreme values are present.

The Cumulative Distribution Function (CDF) follows a normal distribution.



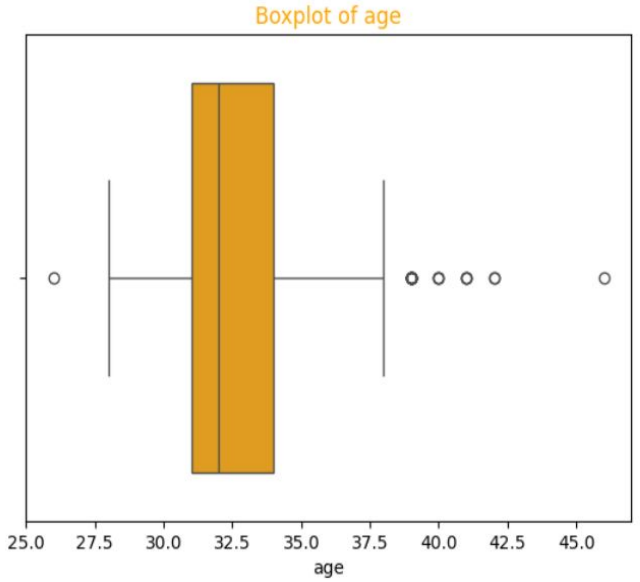
Age Observations:

75% of students are below 35 years of age.

The majority lie in the range of 31-34.

The box plot indicates the presence of a few outliers, representing individuals working at very young and old ages respectively.

The Cumulative Distribution Function (CDF) does not follow a normal distribution.



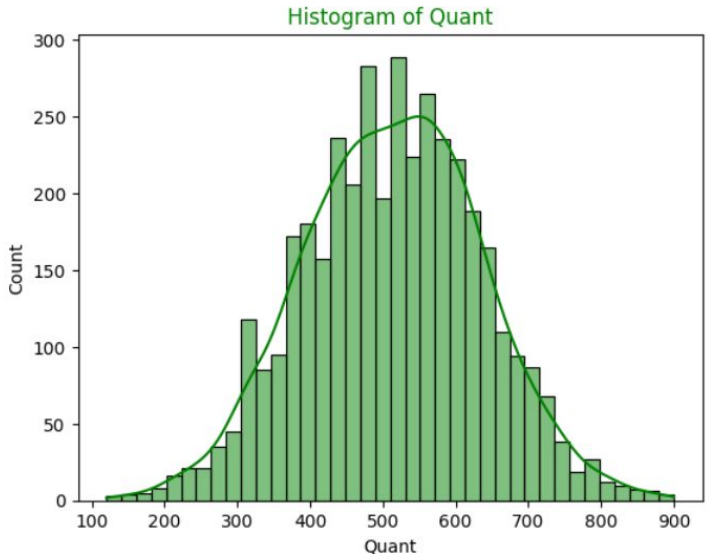
Tenure Observations:

The range is around 4 years.

The histogram shows that most people are either freshers or experienced.

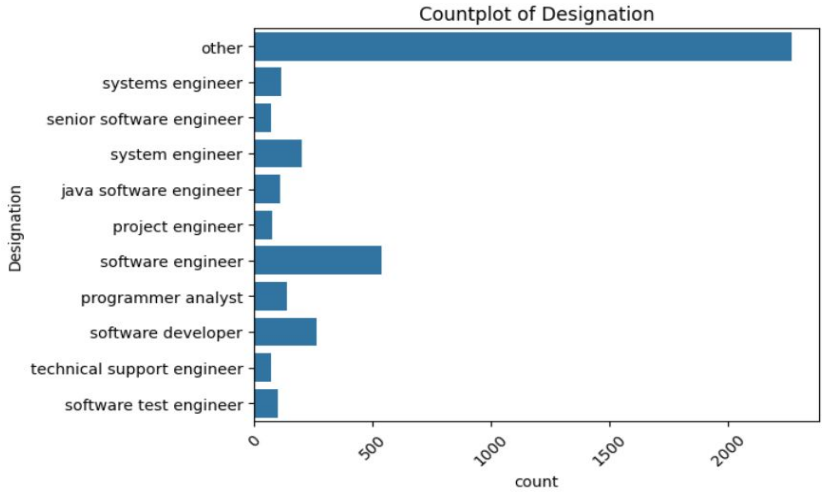
The box plot displays only one outlier with maximum experience.

The Cumulative Distribution Function (CDF) indicates that the data is not normally distributed.

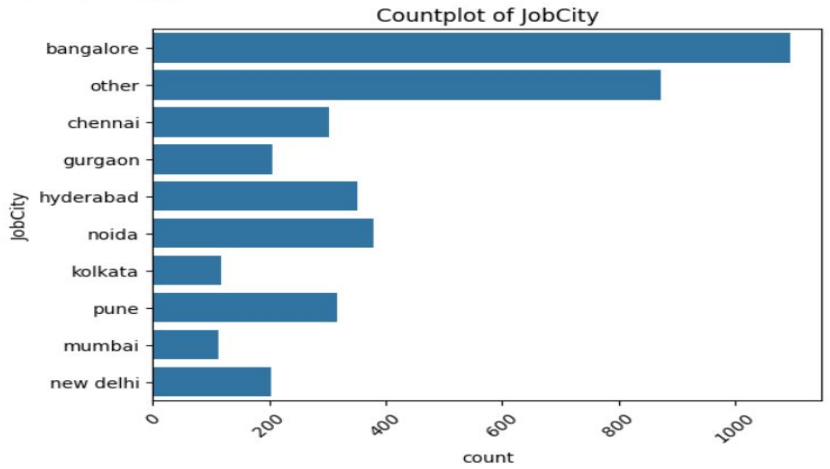


INNOMATICS RESEARCH LABS

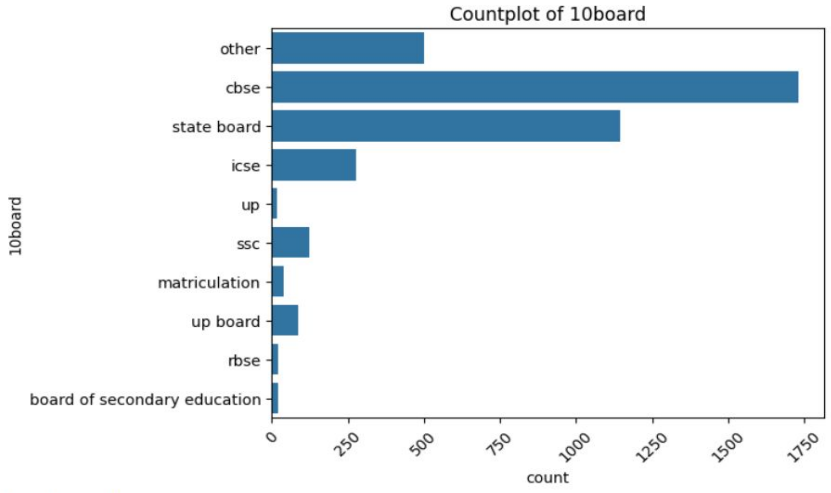
Categorical features: (count plot)

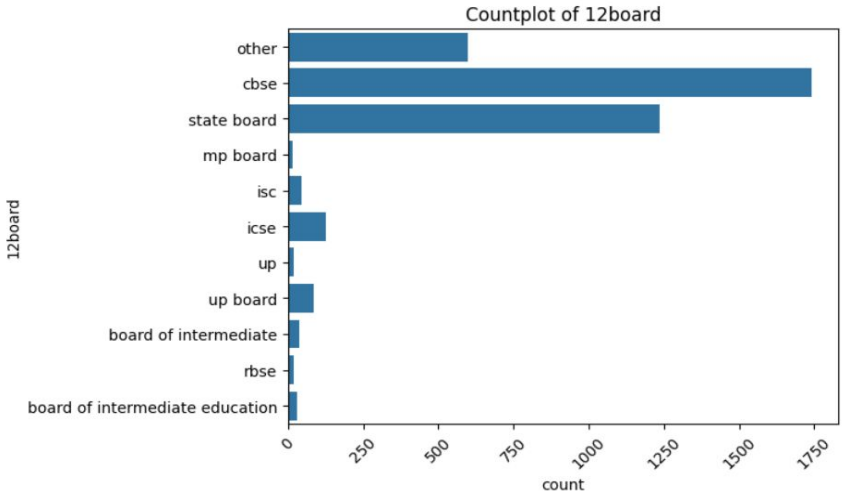


No.of software engineers are the highest.



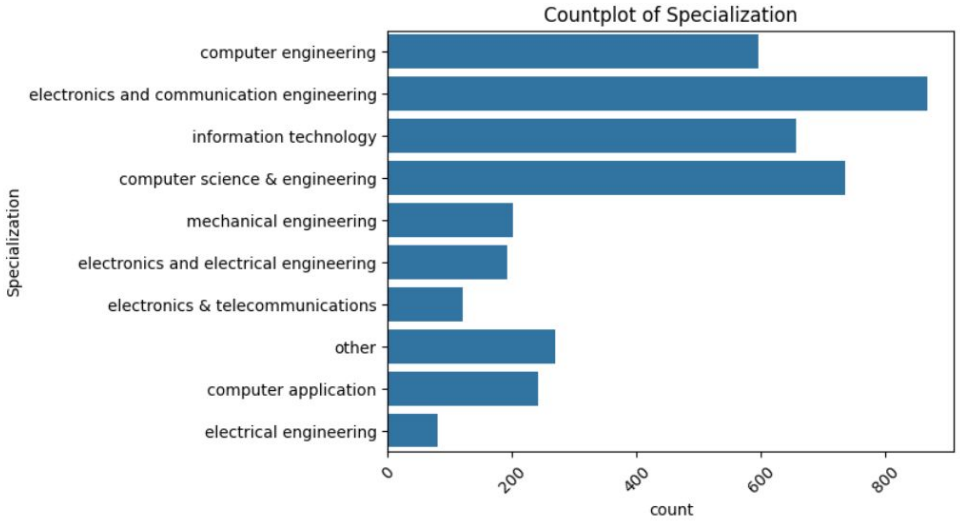
Bangalore is the most preferred job city followed by Noida and Hyd.



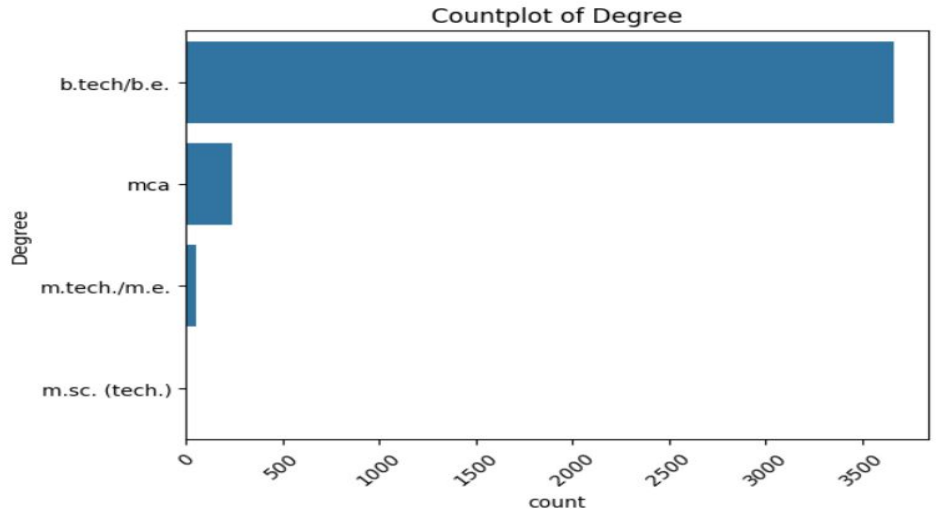


CBSE is the most common board for 10th and 12th grades whereas the college tier analysis shows a dominance of tier 1 colleges.

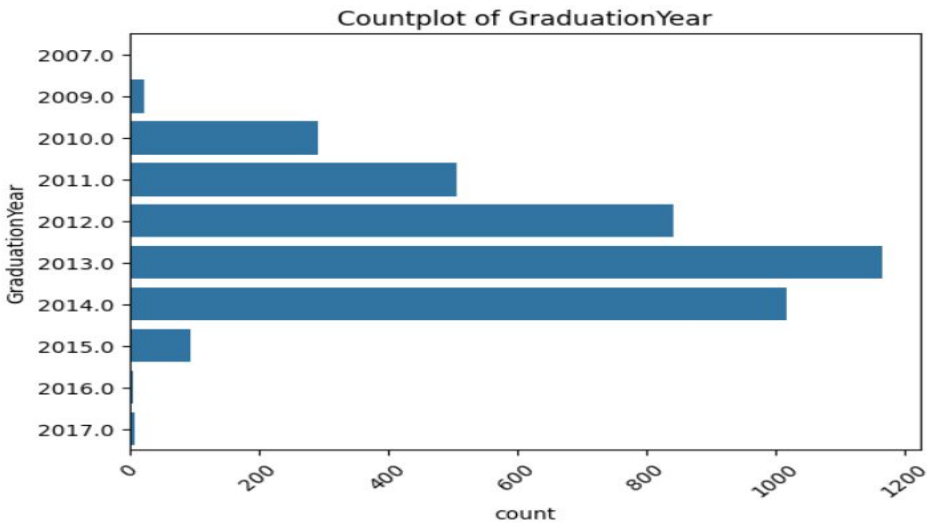
INNOMATICS RESEARCH LABS



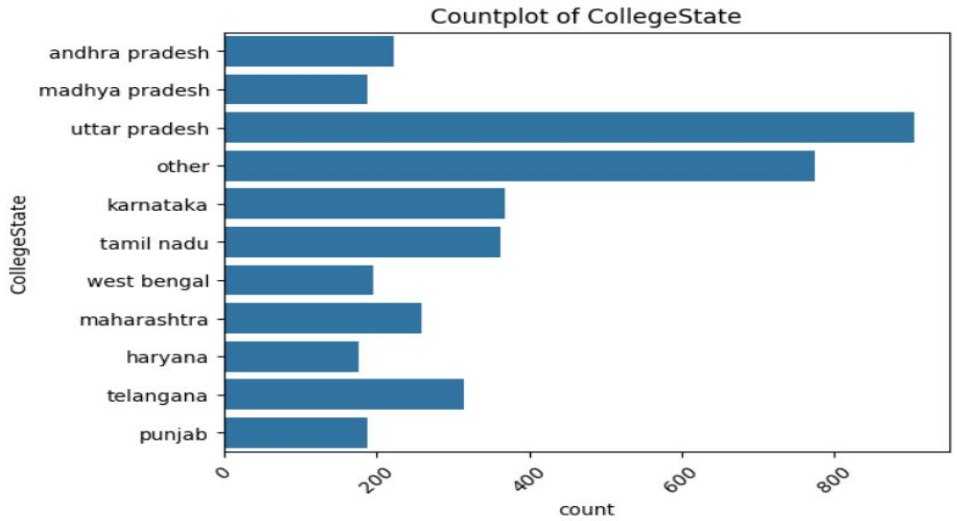
People pursuing ECE are the highest.



Most of the students prefer Btech degree over others.



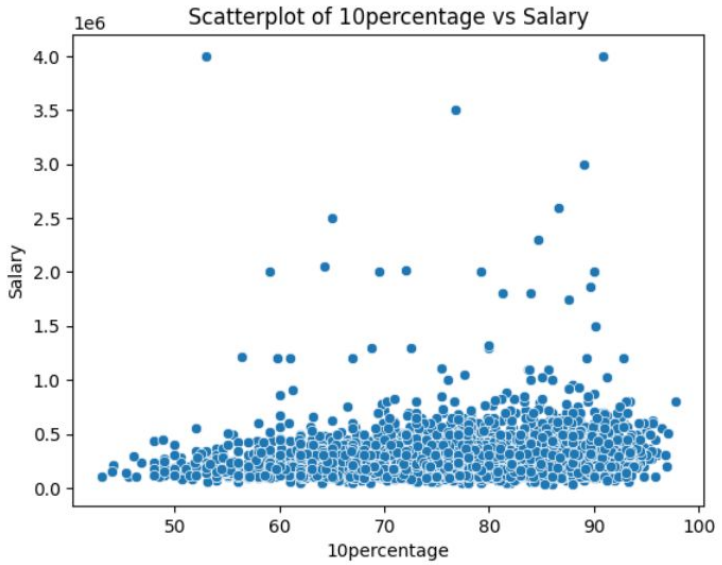
Most of the graduates are from the year 2013

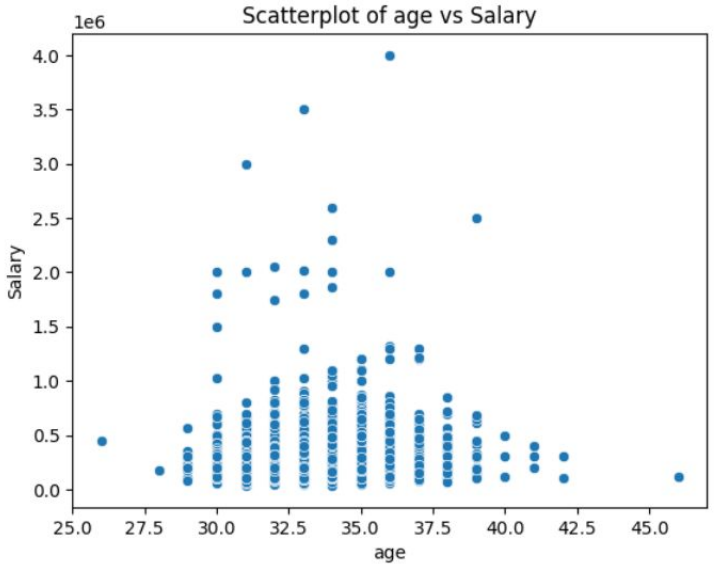


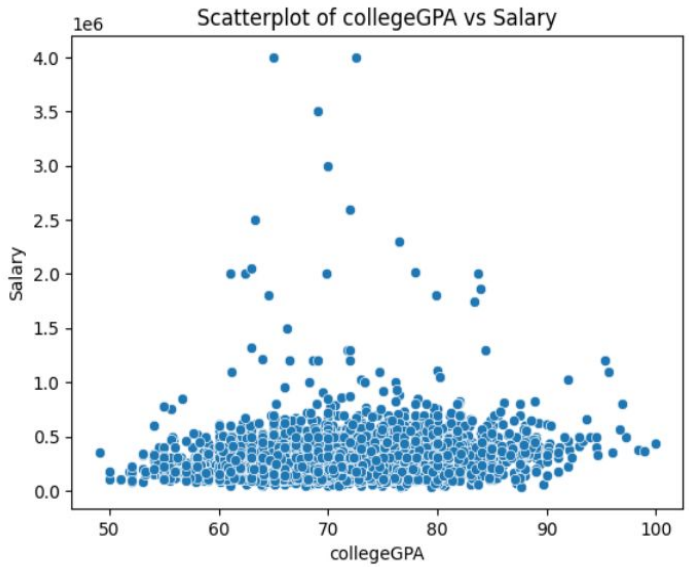
Most of the students are from UP

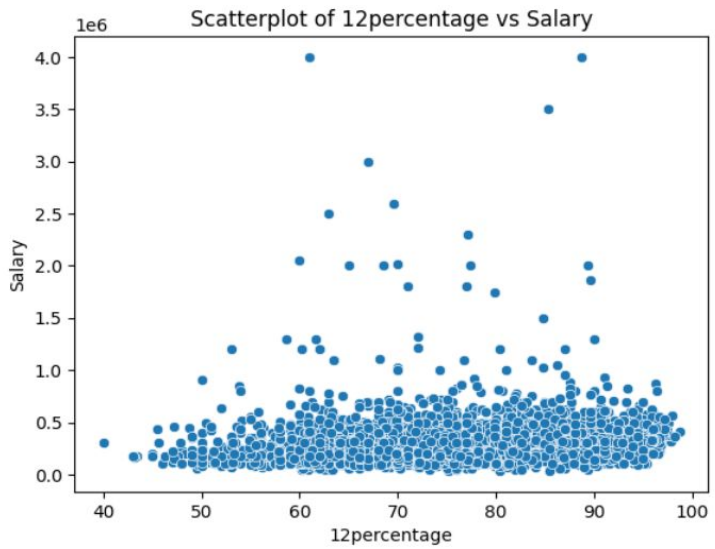
INNOMATICS RESEARCH LABS

4. Bivariate analysis: **a) NUM-NUM**

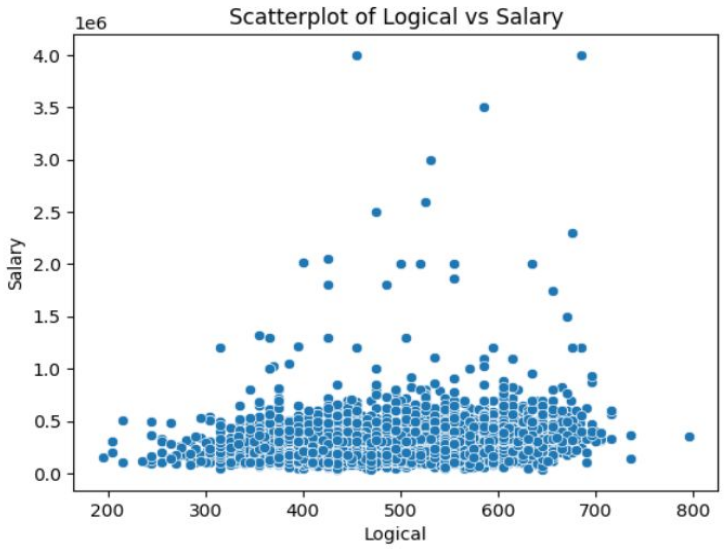


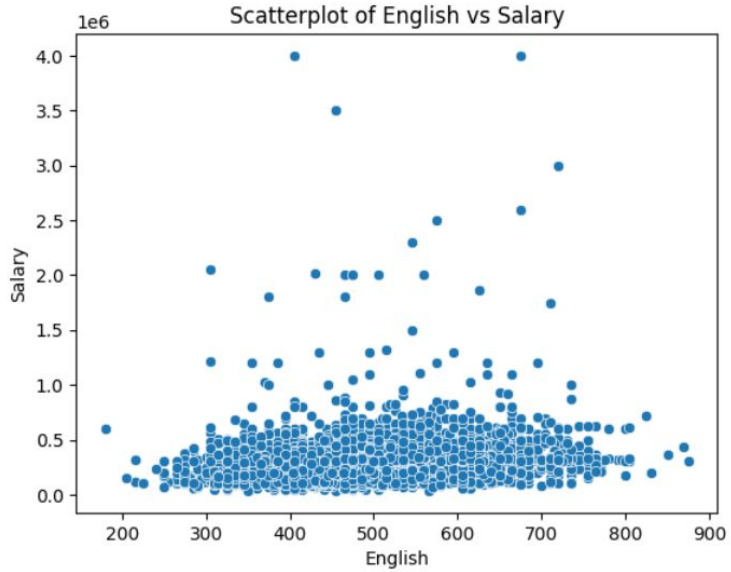


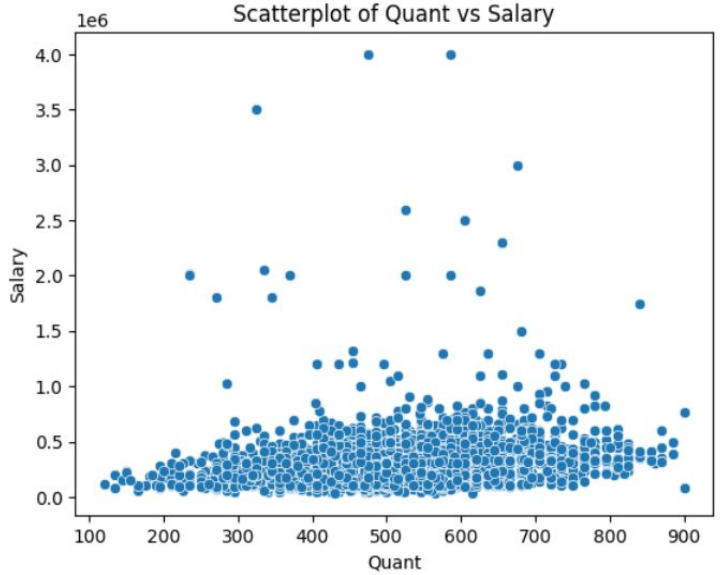




INNOMATICS RESEARCH LABS

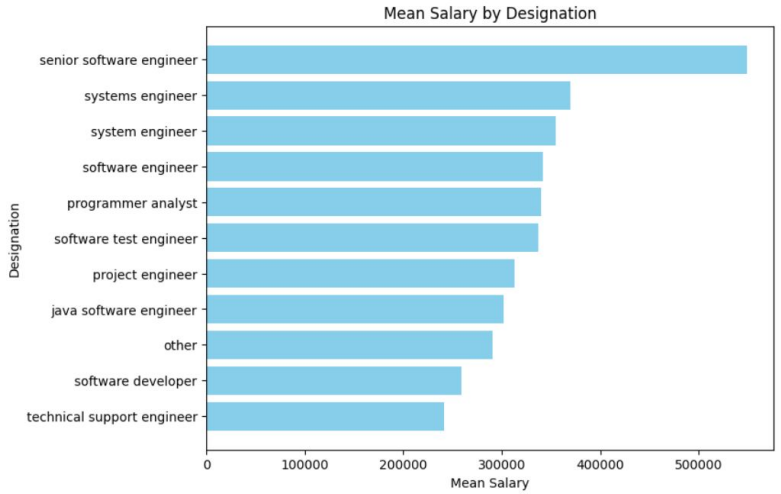




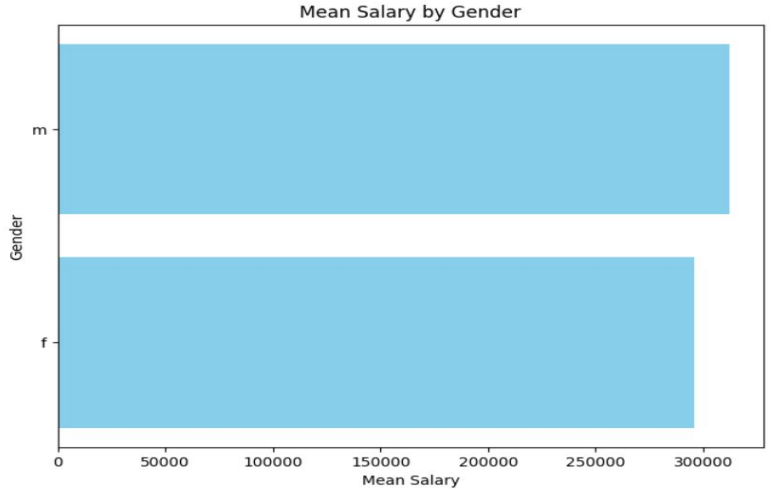


These are some of the bivariate plots between num-num type features.
From the plots we can infer that there is no direct correlation of these features with our target variable i.e ‘Salary’.

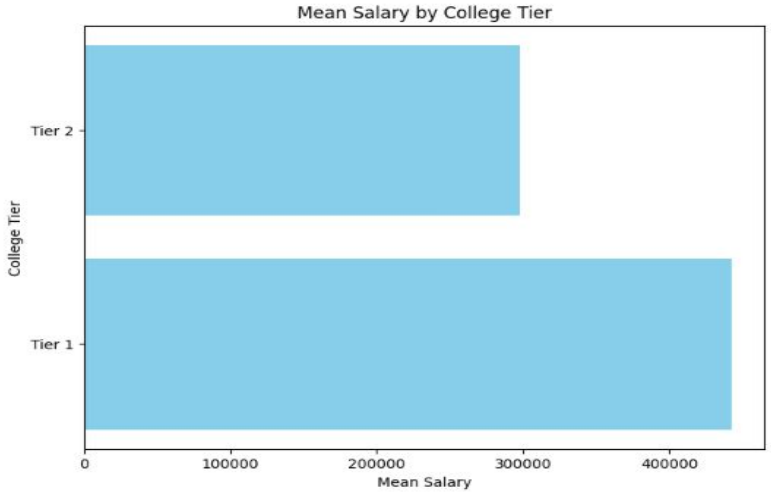
B. NUM-CAT



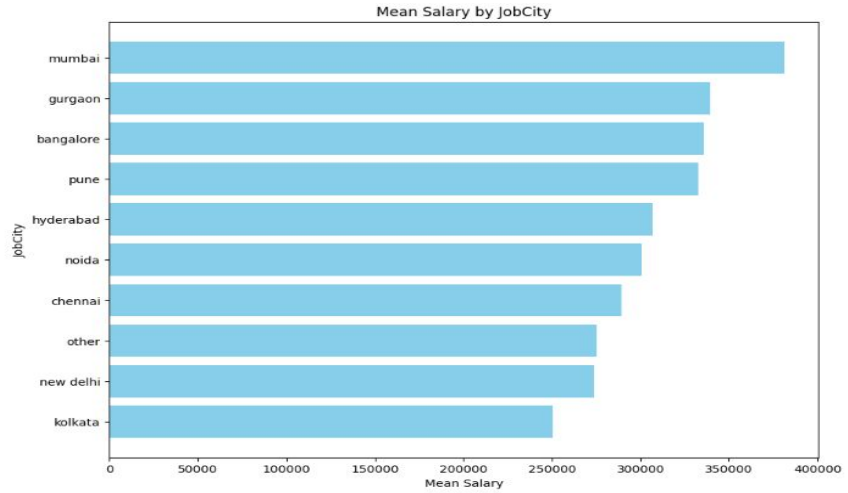
We can infer that the senior software engineers are among the highest paying.



Both male and female are being paid equally.



The mean salary of people from the tier 1 colleges is higher than that of the others.

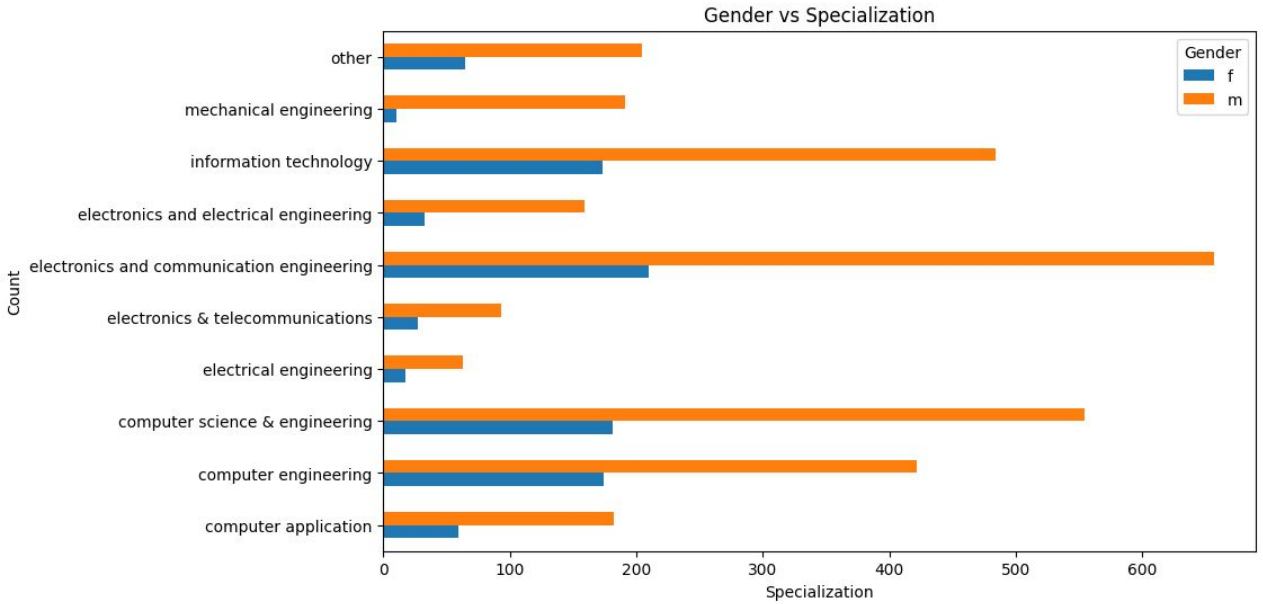


People working in mumbai are paid the highest compared to other locations

INNOMATICS RESEARCH LABS

Relation between Gender and specialisation

- The contingency table created from the data shows the count of individuals categorized by their specialization and gender.
- Upon performing a chi-square test of independence, we found a statistically significant relationship between gender and specialization (Chi-square statistic = 56.28, p-value < 0.05).
- This indicates that gender and specialization are not independent of each other, suggesting that there is some association between the two variables.
- Now the relation between the gender and specialization is plotted below which also concludes that the male population is almost double than that of the female population and very few females have opted for mechanical specialisation.



INNOMATICS RESEARCH LABS

Conclusion

The comprehensive analysis of the dataset uncovers several significant findings regarding the factors influencing salary levels. While certain factors such as tenure and college tier exhibit a strong correlation with compensation, others like gender and academic performance show no discernible relationship. Senior Software Engineers command the highest salaries, albeit with greater variability, while Software Developers and Technical Support Engineers typically earn less than the average. Although gender does not seem to significantly affect overall income determination, females tend to receive lower salaries on average. Additionally, academic performance metrics such as 10th, 12th, and college GPA scores do not exhibit a clear correlation with salary levels.

THANK YOU

Presented by:
Shruthika Polkam
IN1241067

INNOMATICS RESEARCH LABS