INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA PROJECT – AMCAT ANALYSIS

RANIYA ARIF
INTERN ID – IN1242093

About me

Holding a Bachelor of Technology in Computer Science, I've developed a robust foundation in analytical problem-solving and technical skills.

My passion for data-driven decision making propelled me to pursue a Postgraduate Program in Data Science and Business Analytics at the University of Texas at Austin. This program has deepened my expertise in machine learning and Python, equipping me with essential tools for sophisticated data analysis.

My academic projects during this program have allowed me to apply these skills in practical scenarios, utilizing predictive modeling to address real-world business challenges. This experience has honed my ability to transform theoretical knowledge into actionable insights.

LinkedIn : http://www.linkedin.com/in/raniya-arif

Business Problem and Use case Domain

The AMCAT Data Analysis project focuses on understanding the employment outcomes of engineering graduates based on a dataset released by Aspiring Minds, a leader in employment assessment. AMCAT, or Aspiring Minds' Computer Adaptive Test, is widely used by employers to assess the employability of graduates across various domains. The dataset contains detailed information on students' standardized scores in cognitive, technical, and personality skills, as well as their demographics and employment outcomes like salary, job titles, and locations.

The business problem centers around the disparities in salaries among engineering graduates and aims to identify key factors that affect their earning potential. By analyzing this data, we aim to uncover insights into how different skills, academic specializations, and personal characteristics influence salary levels.

Objective of the Project

The objective of the AMCAT Data Analysis project is to systematically analyze the data collected by Aspiring Minds on engineering graduates to identify and quantify the factors that influence their initial salary levels after graduation. By focusing on variables such as academic performance, specialization fields, and demographic factors, this analysis seeks to provide actionable insights into how these elements correlate with employment outcomes.

Summary of the Data

Dataset Overview: The dataset used for analysis is the Aspiring Minds Employment Outcomes 2015 (AMEO), provided by Aspiring Minds. This dataset includes comprehensive information on the employment outcomes of engineering graduates in India. It spans demographics, educational backgrounds, test scores, and job placements, offering a granular view into the factors influencing graduate employability.

Data Composition:

- **Total Entries:** The dataset comprises data from 3,998 candidates.
- **Variables:** There are 39 distinct variables, including candidate demographics, educational qualifications, standardized test scores, and job placement details.

Key Statistics:

- **Salary Information:**
  - **Mean Salary:** ₹307,699.8
  - **Standard Deviation:** ₹212,737.5
  - **Range:** Minimum salary is ₹35,000, and the maximum is ₹4,000,000.
- **Academic Performance:**
  - **10th Percentage:** Average 77.93% (SD = 9.85, Min = 43, Max = 97.76).
  - **12th Percentage:** Average 71.49% (SD = 8.17, Min = 6.45, Max = 99.93).
  - **College GPA:** Mean GPA is 1.93 in a Tier-based college system indicating most participants come from Tier 2 colleges.

Column Composition:

The dataset contains 39 diverse columns, including:

Identifiers: ID, Date of Birth (DOB)
Employment Specifications: Salary, Designation, Job Location
Academic Credentials: Marks in 10th and 12th grades, College GPA
Assessment Scores: English, Logical Reasoning, Quantitative Analysis
Academic Specializations: Degree Type, Field of Specialization
Personality Metrics: Conscientiousness, Agreeableness, Extraversion, Neuroticism, Openness to Experience

Data Typology: The dataset incorporates a variety of data types, including float64 (10), int64 (12), datetime (2) and categorical (object type) (12). None of the columns contained missing values.

Integrity of Data: The dataset maintains a high standard of data cleanliness, with no missing values across all columns, facilitating immediate analytical processing.

Exploratory Data Analysis

a. Data Cleaning and Manipulation

IrrevelantColumns Elimination: Columns such as ID, CollegeID, CollegeCityID, and an unspecified column named ‘Unknown’ were identified as non-contributory towards the analysis and hence removed from the dataset.

Data Type Standardization: The Date of Birth (DOB) and Date of Joining (DOJ) fields, initially stored as object types, were converted to Date type to facilitate analysis.

Error Correction in Key Fields:

Location and Specialization Cleanup: The JobCity and Specialization columns displayed numerous inconsistencies, such as spelling errors and incorrect values. Particularly, the JobCity field had over 300 unique entries, including erroneous labels such as '-1'. Application of the fuzzywuzzy library refined these entries, reducing the count to 214 valid categories.

Educational Background Standardization: Erroneous entries in academic columns (10th, 12th grades, and CollegeGPA) like '-1' and '0', which indicated missing data, were corrected to ensure data integrity.

Examination Board Rationalization: Multiple redundant or inconsistent values in the Board of Examination column were corrected.

Handling Missing and Faulty Entries: Columns such as GraduationYear and JobCity included invalid entries like '0' and '-1', respectively. These were identified and corrected to preserve the dataset's factual accuracy. Domains such as MechanicalEngg, ElectricalEngg, TelecomEngg, and CivilEngg had significant data absence, exceeding 94%. Given the potential bias in estimating these missing values, these columns were excluded from the dataset.

Duplicate Records: A check for duplicates ensured the dataset was free of redundant entries, affirming the uniqueness of each data point.

Domain Data Rectification: The data within the Domain column was revised, assigning accurate abbreviations to corresponding domains, thus enhancing the dataset's applicability in specialized analytical contexts.

Age Calculation: A new column named 'Age' was derived from the Date of Birth (DOB) to facilitate analyses that consider the age factor of the candidates, thereby allowing a deeper exploration into age-related trends in employment outcomes.

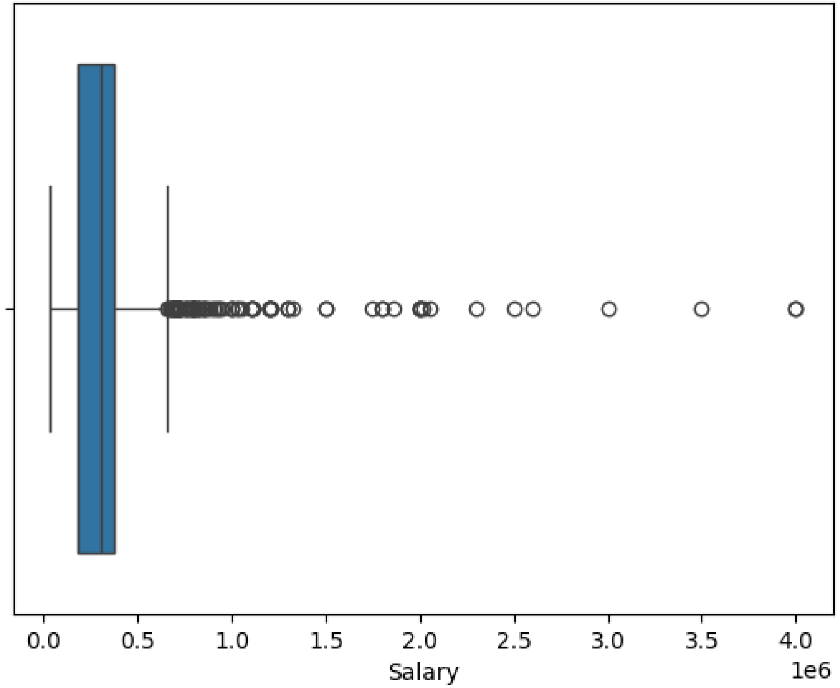
Text Data Standardization: Categorical text-based columns underwent further cleaning to remove any extraneous leading or trailing spaces, ensuring uniformity. Additionally, the case of alphabetic characters was standardized to maintain consistency across the dataset, enhancing the reliability of text-based queries and analyses.

b. Univariate Analysis

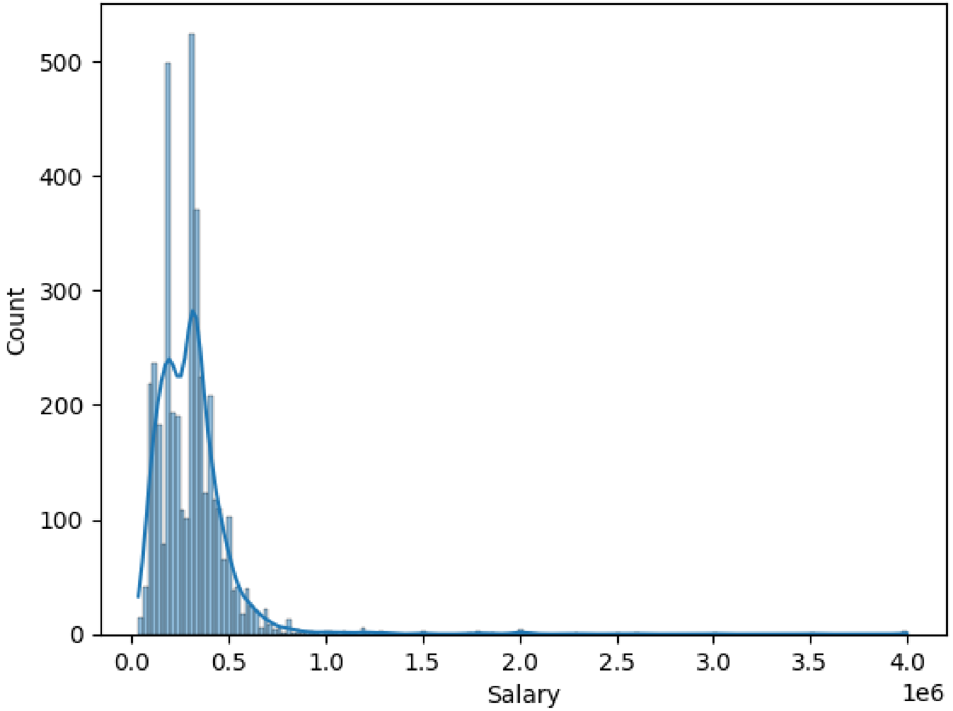
1. **Salary**

- Multi-modal Distribution: The histogram shows multiple peaks, suggesting that the 'Salary' variable may represent several different groups or populations within the data.
- Extremely right skewed: The distribution is right-skewed, with a tail extending towards higher salary values. Most of the data is concentrated in the lower salary range, with fewer instances of very high salaries. With one outlier being as high salary as 4000000 rupees.
- High Frequency at Lower Range: There is a significant spike at the lower end of the salary range, which could indicate a large number of entries at a minimum or starting salary level.

Boxplot:Salary

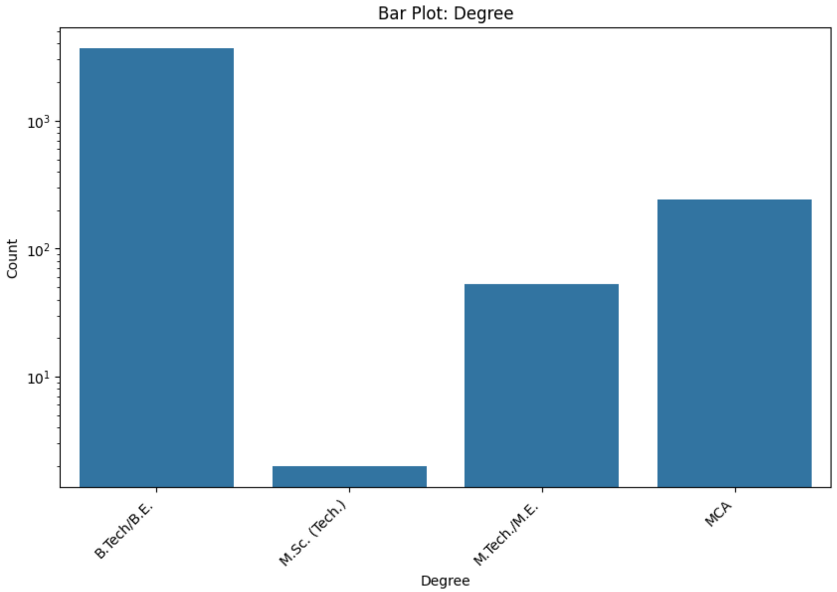


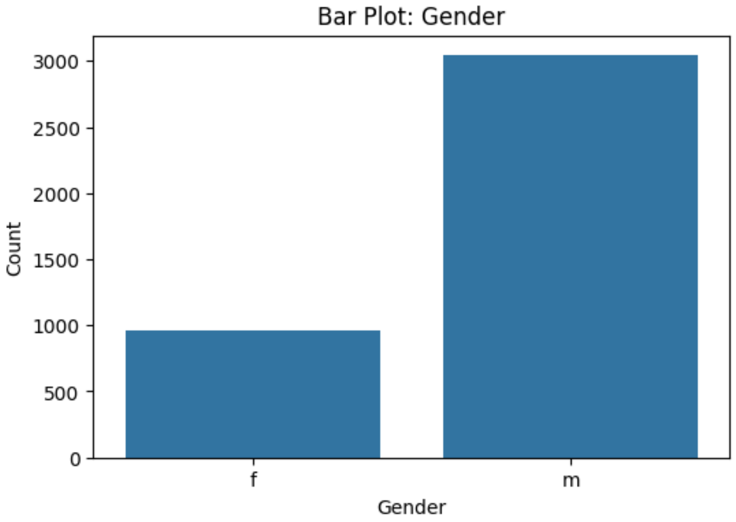
Histogram:Salary



2. Degree

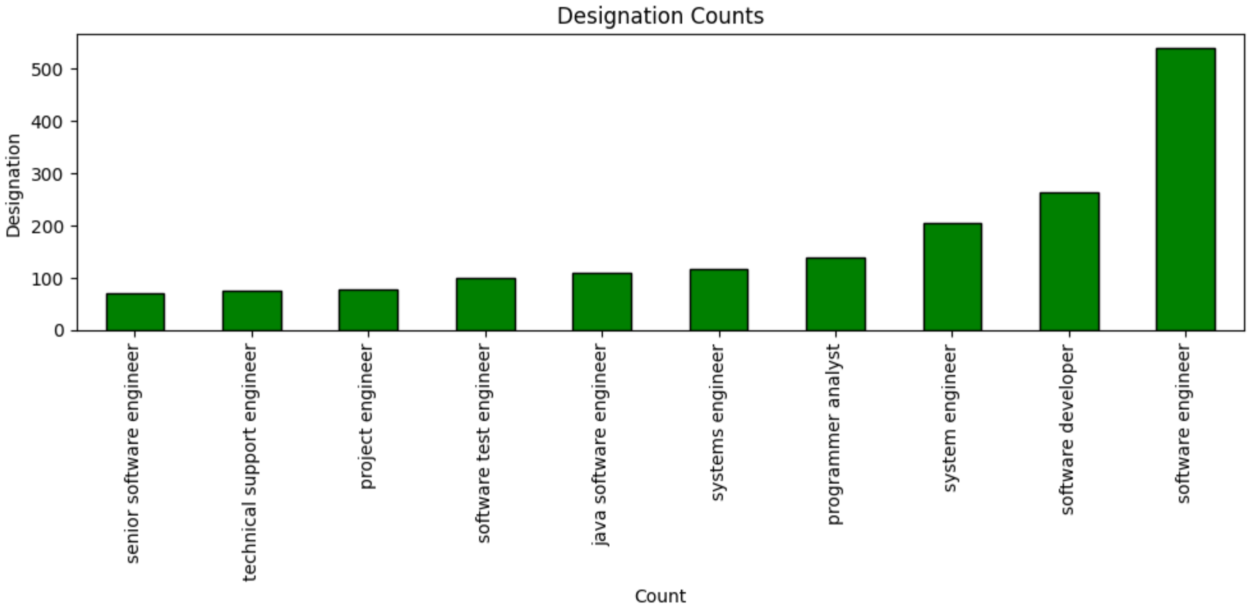
Few students hold an M.Sc. in technology, with the majority of graduates having completed their B.Tech degrees, followed by MCA graduates.





3. Designations

Common Roles: Dominated by titles like 'Software Engineer' and 'Software Developer'.

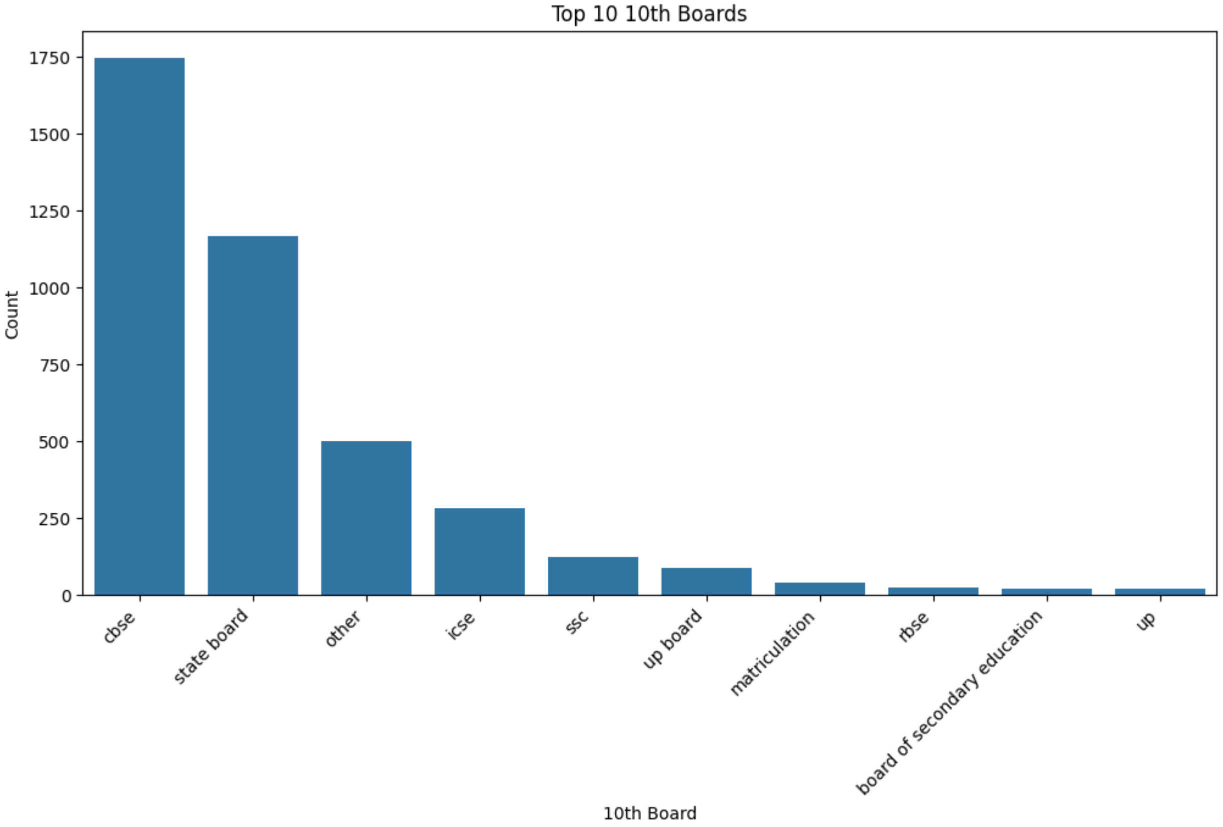


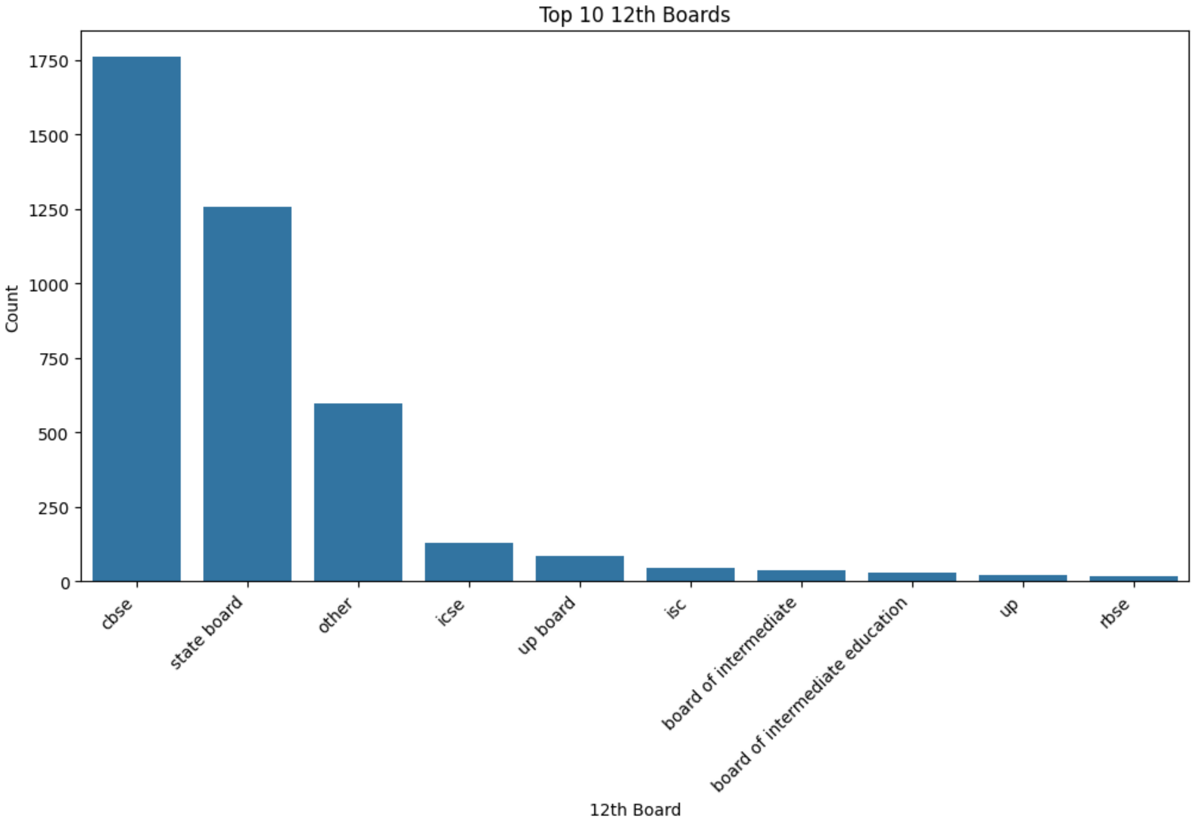
4. Gender

Male Dominance: A significant majority of the data points are male.

5. Board

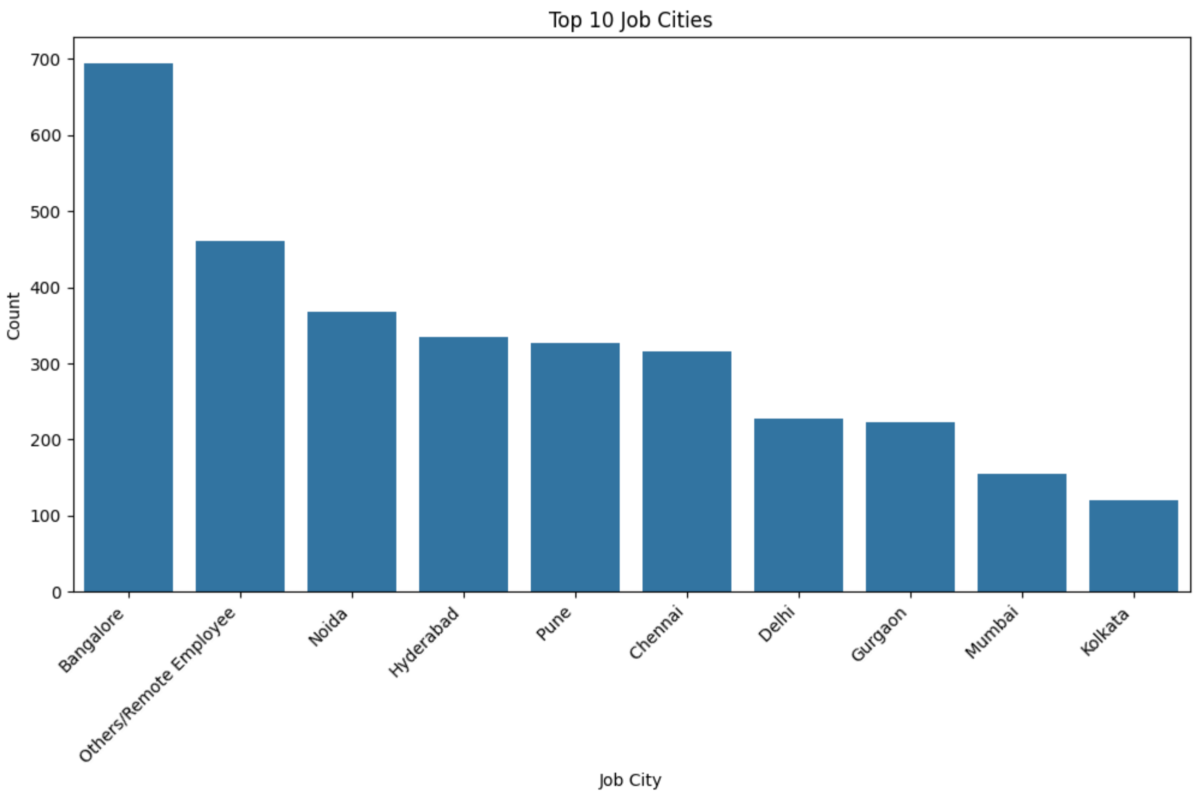
CBSE Prevalence: Majority attended CBSE for both 10th and 12th grades.





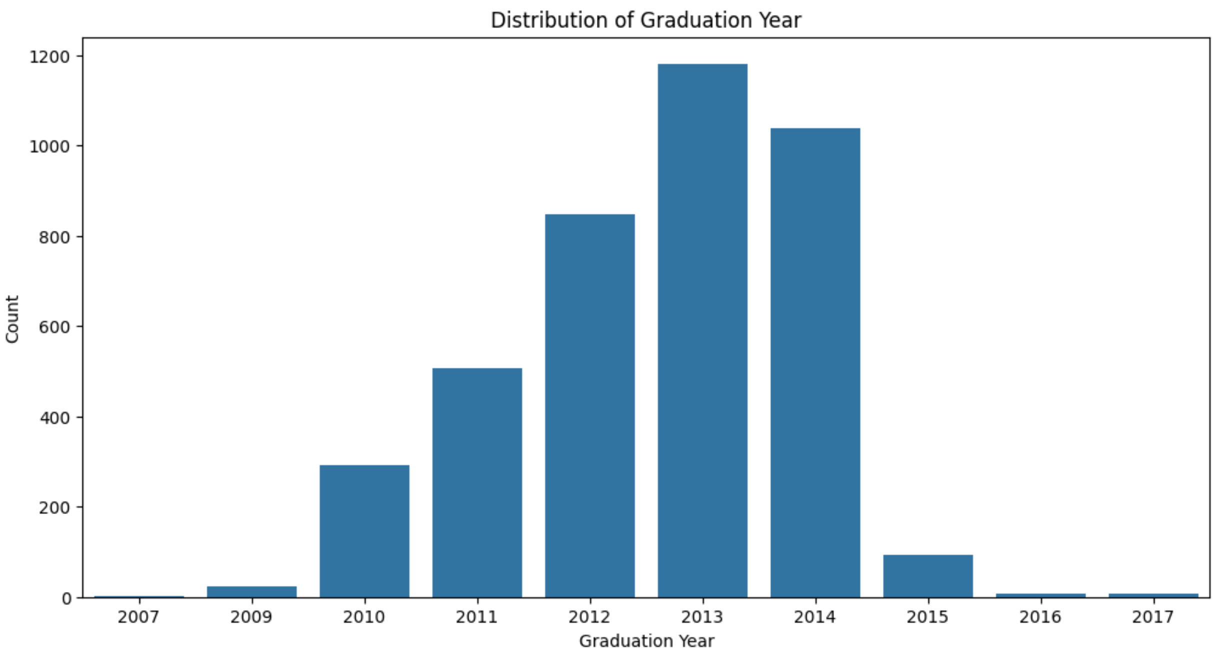
6. JobCity

Placement Concentration: Most graduates placed in Bangalore, least in Kolkata.



7. Graduation Year:

Recent Graduates: Most graduated around 2013-2014, with a decline before and after.

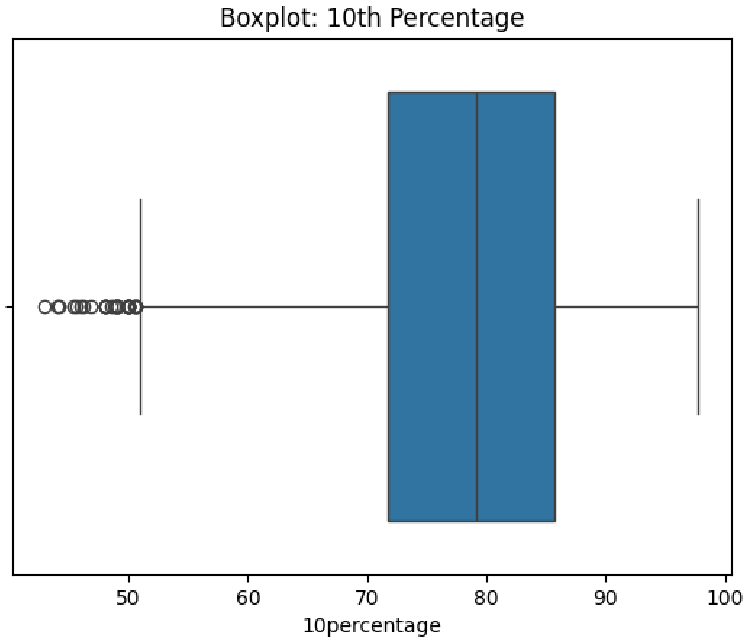


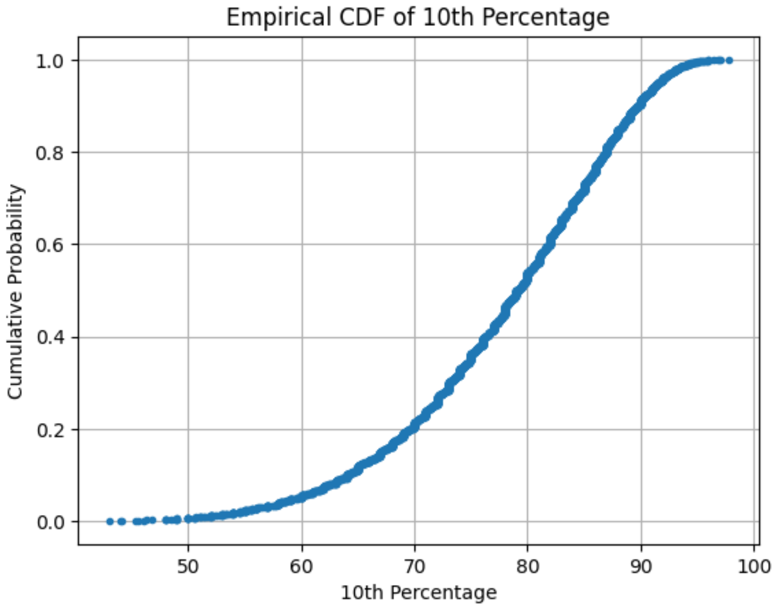
INNOMATICS
RESEARCH LABS

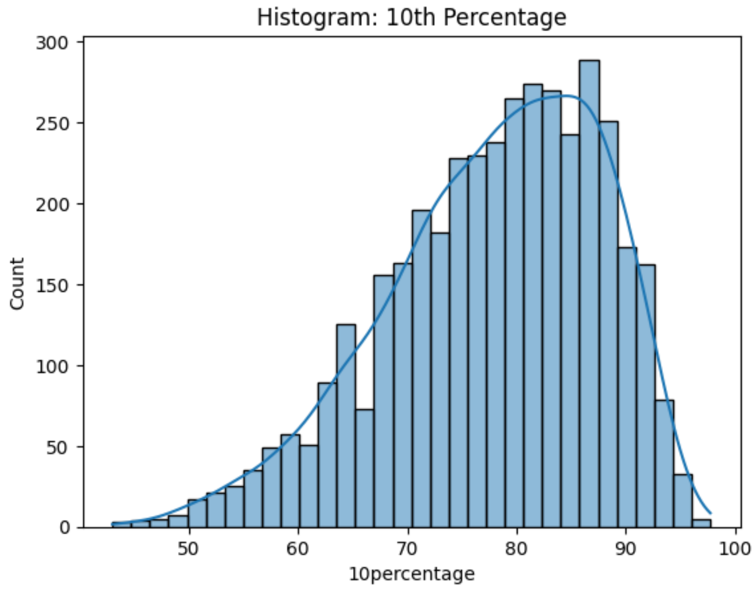
8. 10percentage

Distribution Spread: Most of the students' scores are spread between 50 and 95, with the bulk concentrated above 70.
High Achievers: The sharp rise towards the end of the curve indicates a significant number of high achievers, as a large proportion of students have scores near the maximum of 100.

Lower Scores Are Rare: The initial flat region of the CDF indicates that lower scores (below 50) are relatively rare among the students.





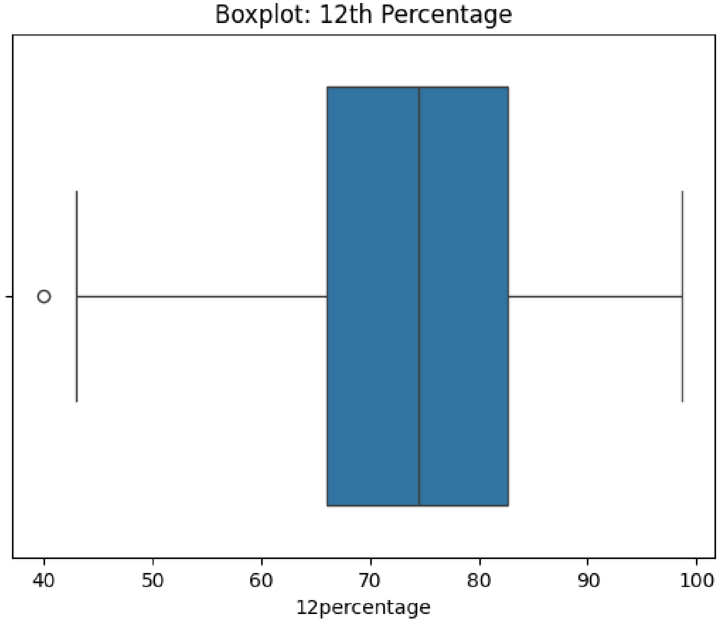


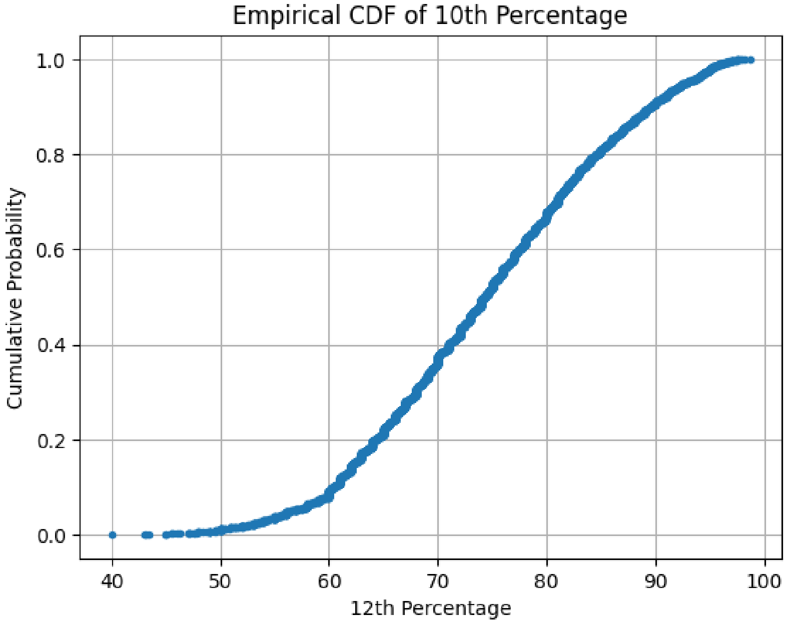
9. 12percentage

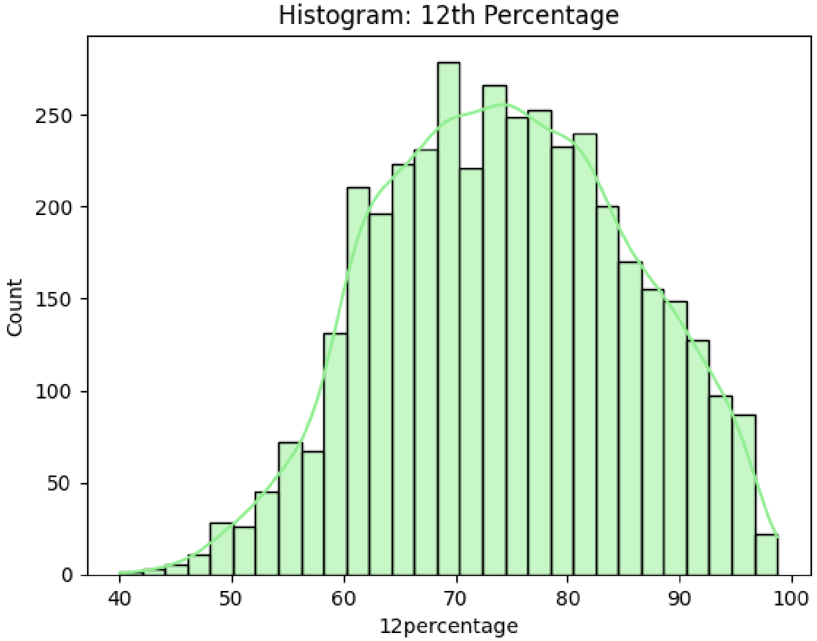
Middle Range Concentration: Most students scored between 50% and 90%, with a significant concentration in the middle of this range.
High Achievers Prevalence: The steeper section towards the end of the curve signifies a higher concentration of students scoring between 80% and 95%, indicating a prevalence of high achievers.

Low Scores Uncommon: The initial flat section of the curve indicates that lower scores (below 50%) are quite uncommon, suggesting a general trend of higher academic achievement among the students in 12th grade.

Distribution Characteristics: The shape of the curve suggests a normal-like distribution, albeit slightly skewed towards higher scores, which is typical for academic performance distributions where failing scores are less common.


Boxplot: 12th Percentage


Empirical CDF of 10th Percentage


Histogram: 12th Percentage

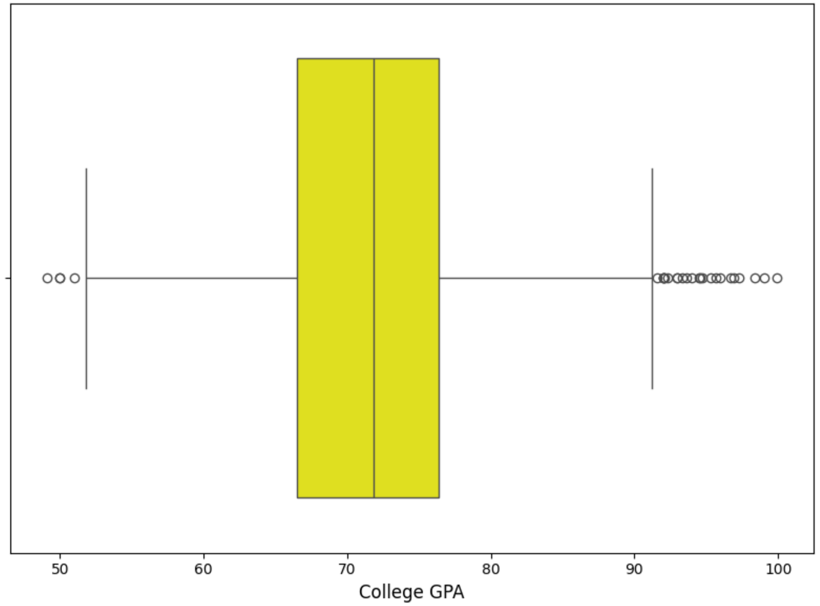
10. College GPA:

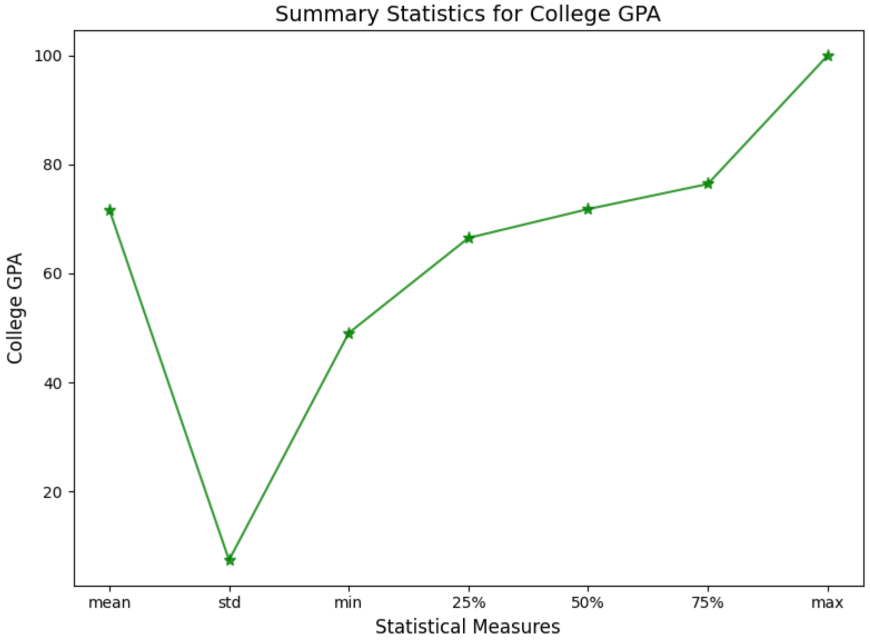
Distribution Spread: The high standard deviation relative to the mean indicates a wide spread in the GPAs, implying diverse academic outcomes among the students.

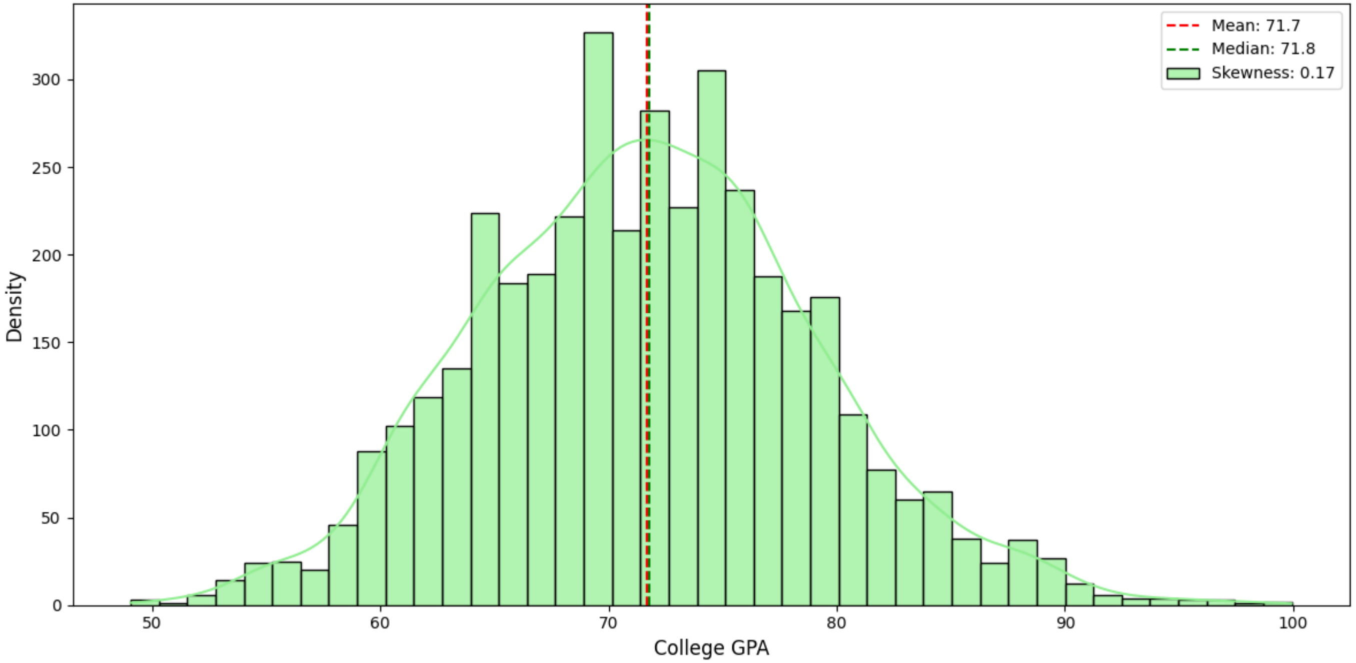
Performance Levels: The significant difference between the minimum and the lower quartile suggests that while most students perform moderately to well, there are outliers with very low GPAs.

Symmetry in Distribution: The close values of the mean and median suggest that the GPA distribution might be roughly symmetric, but the presence of outliers, especially on the lower end, could be skewing the distribution slightly.

High Achievers: The high maximum value shows that there are students performing exceptionally well, though they are not numerous enough to shift the overall average significantly higher.





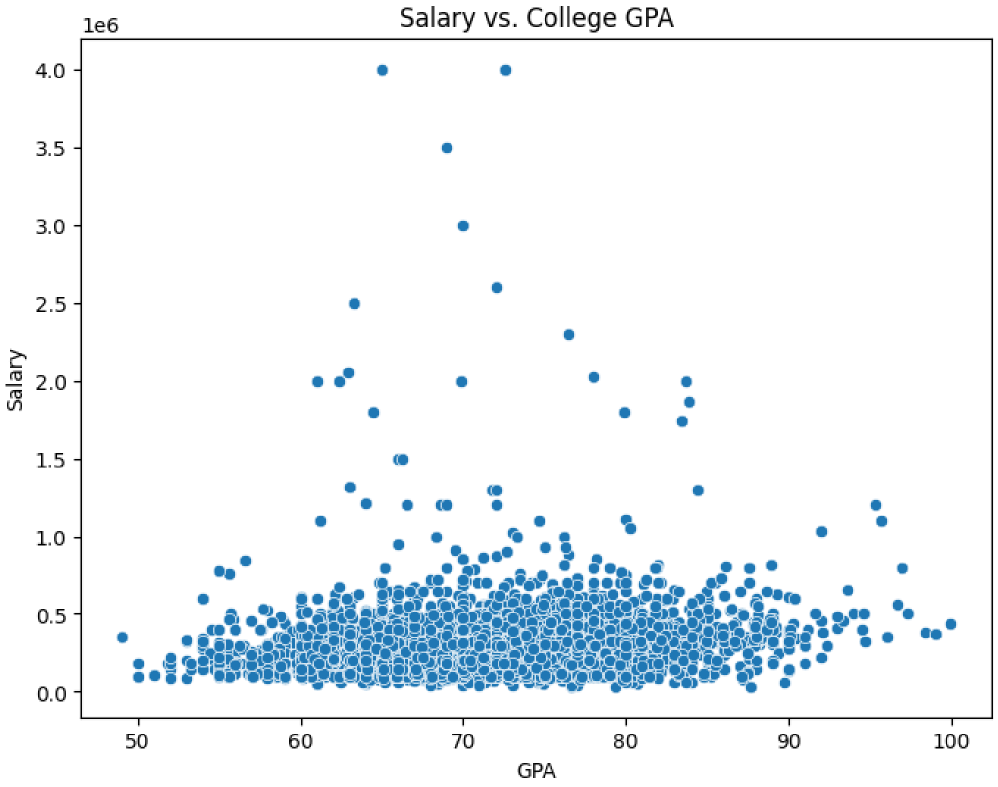


INNOMATICS
RESEARCH LABS

c. Bivariate Analysis

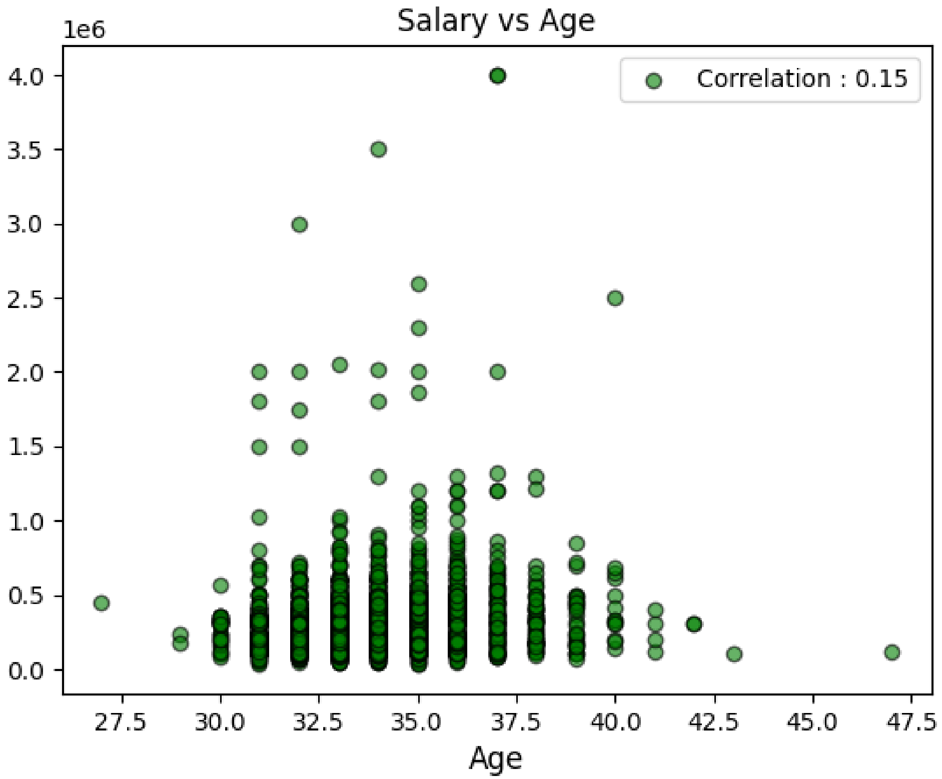
Salary vs CollegeGPA

College GPA does not impact the salary earned by the employees



Salary vs Age

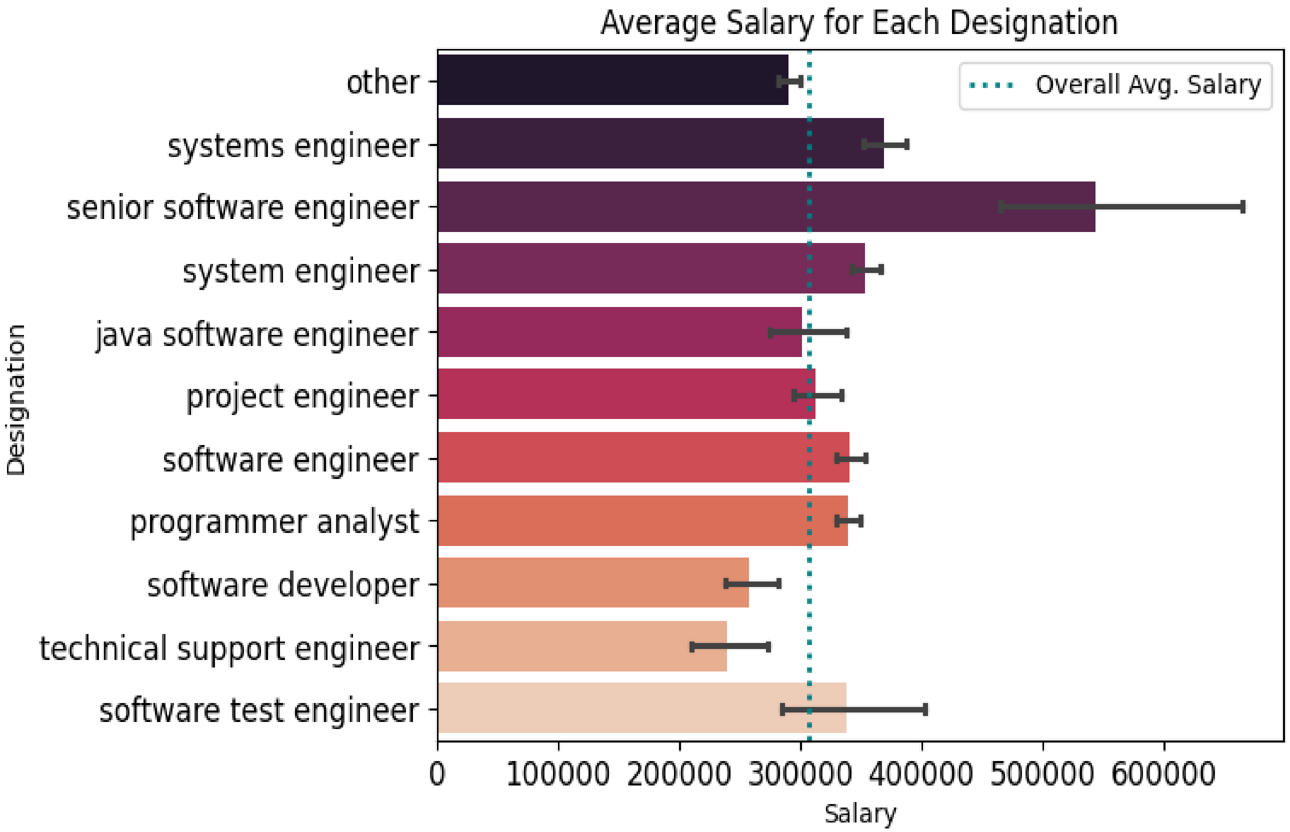
There is no relation between Age of the Employees and the compensation they get



INNOMATICS RESEARCH LABS

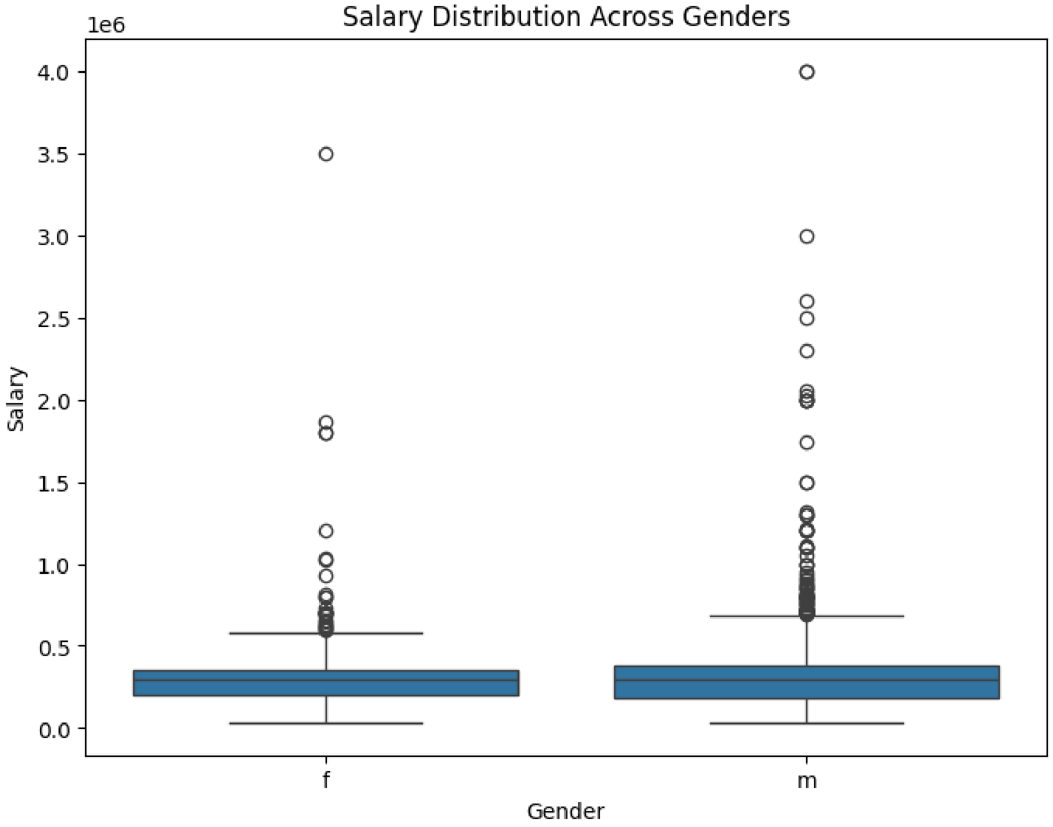
Salary vs Designation

Senior Software Engineers earn the highest followed by System Engineers



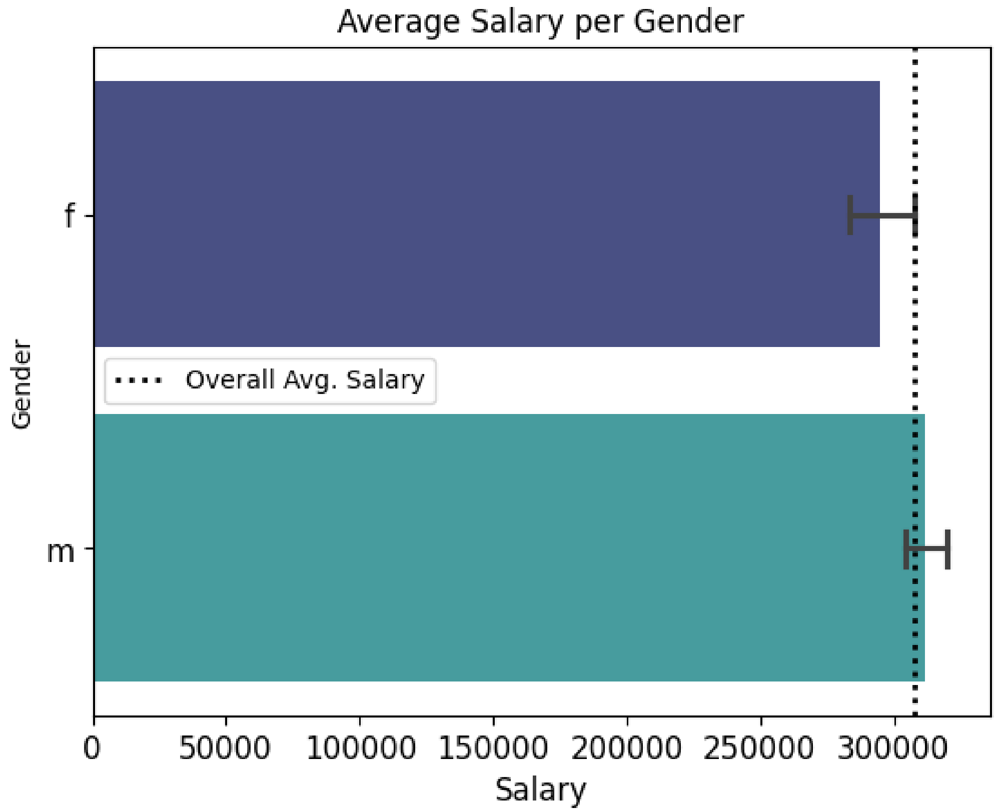
Salary vs Gender

There is a slight difference between the wages earned by the women and men with the spread of Men's salary being higher.



Average Salary vs Gender

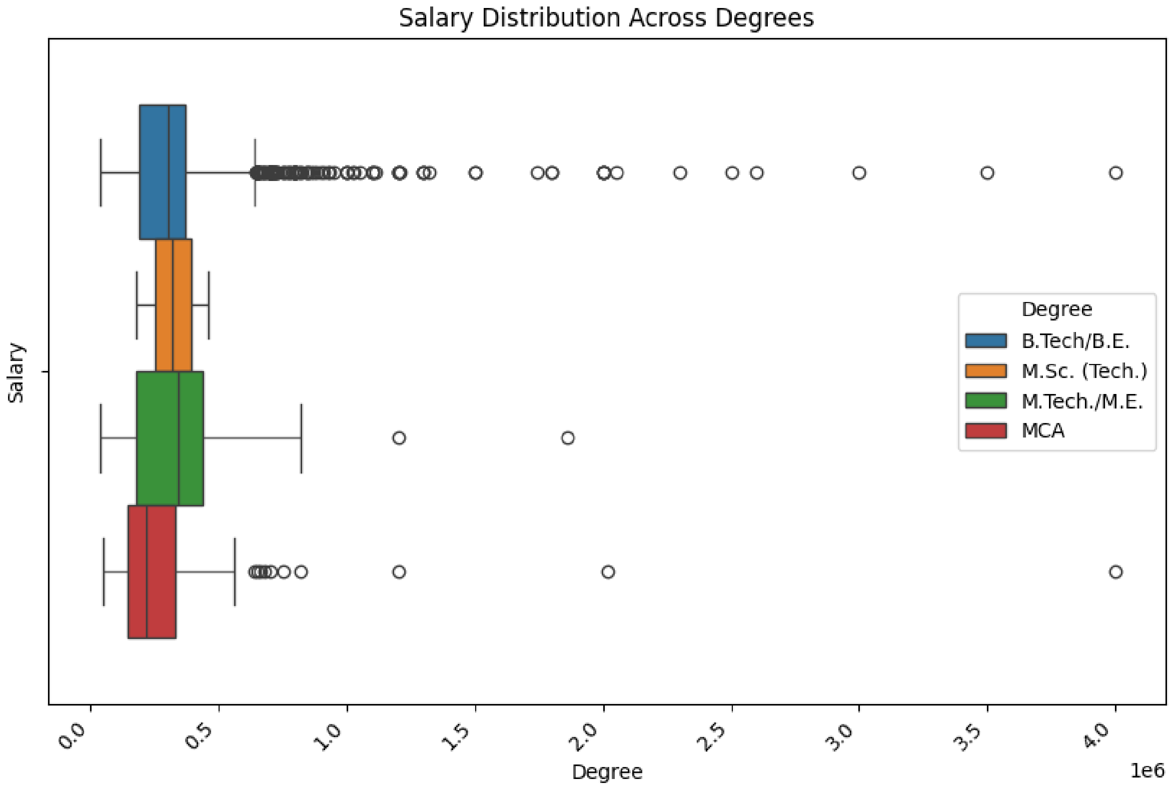
The overall average salary for Female workers is less when compared to the male employees



Salary vs Degree

1.The employees who hold a degree in Bachelor of Engineering or Bachelor of Technology earn higher than the others.

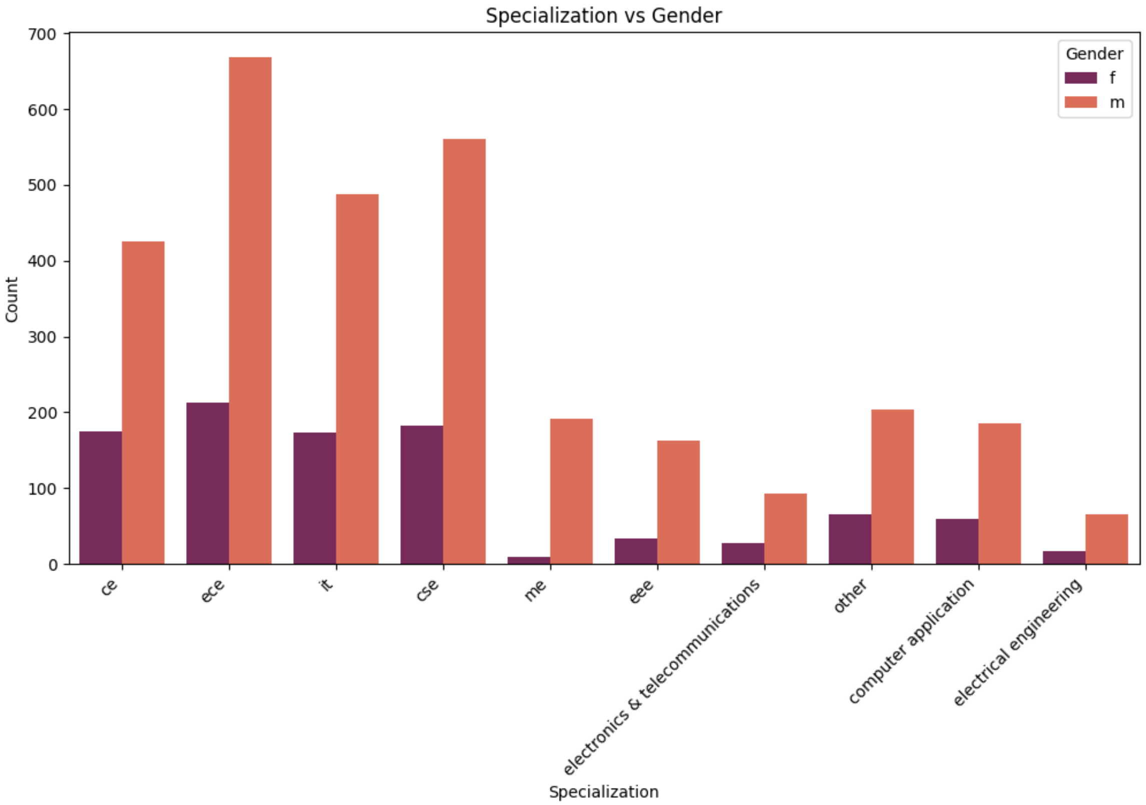
2.The second high paid degree is MCA



INNOMATICS RESEARCH LABS

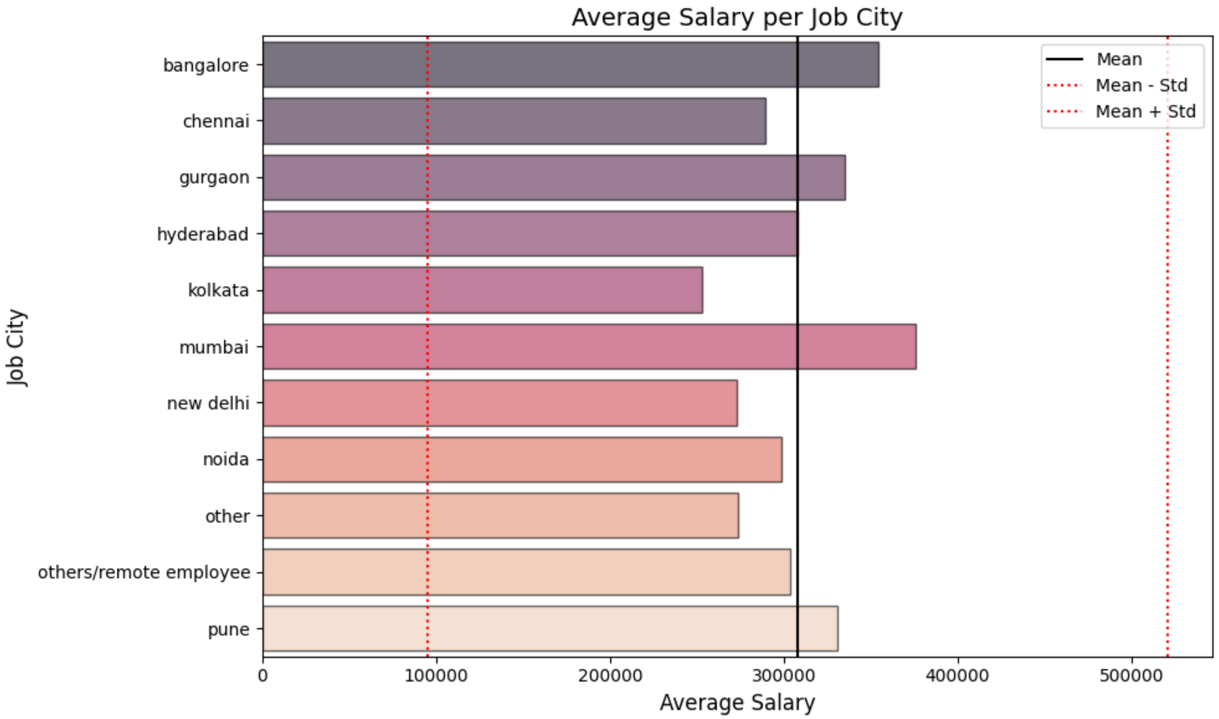
Gender vs Specialization

1.The amount of male employees is significantly higher than the female workers.
2.For both the genders, ECE is the most popular Specialization.
3.Mechanical Engineering has the least number of female employees.



Salary vs JobCity

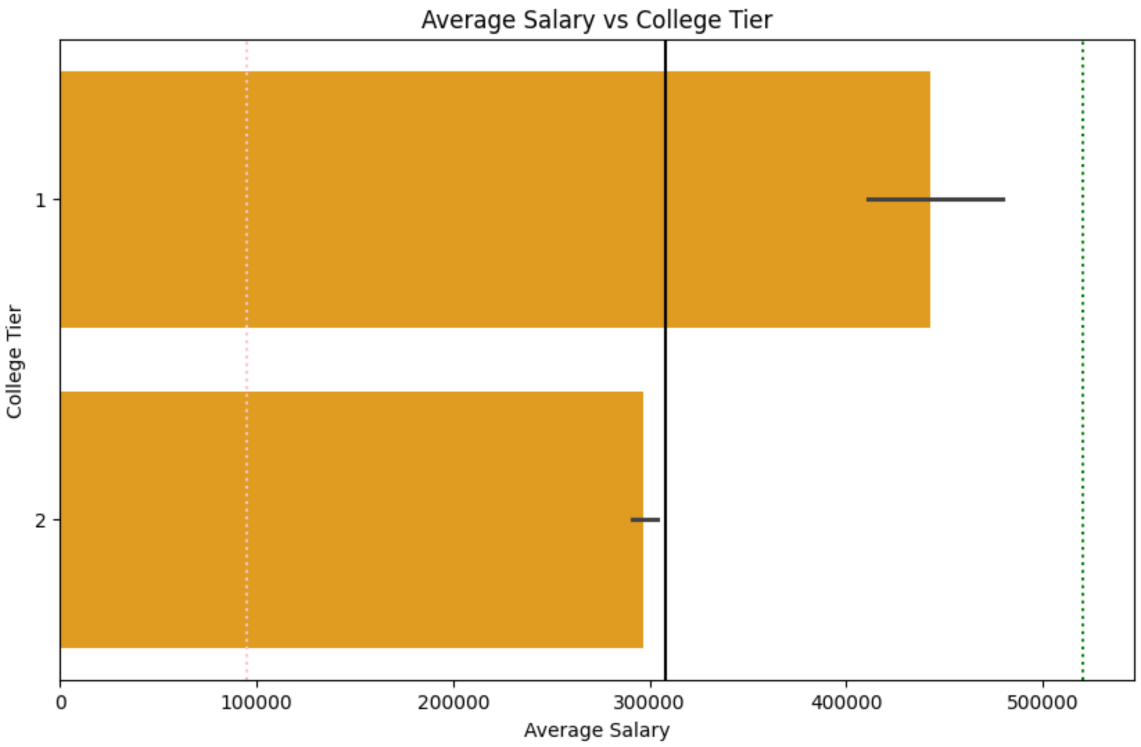
1.The highest average salary offered is in Mumbai followed by Bangalore.
2.Kolkata has least average salary, less number of jobs available in the city can be contributing factor.



INNOMATICS RESEARCH LABS

Salary vs CollegeTier

Employees who graduated from Tier 1 earn significantly higher than those from Tier 2



Research Questions

Q1. Times of India article dated Jan 18, 2019 states that “After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate.” Test this claim with the data given to you.

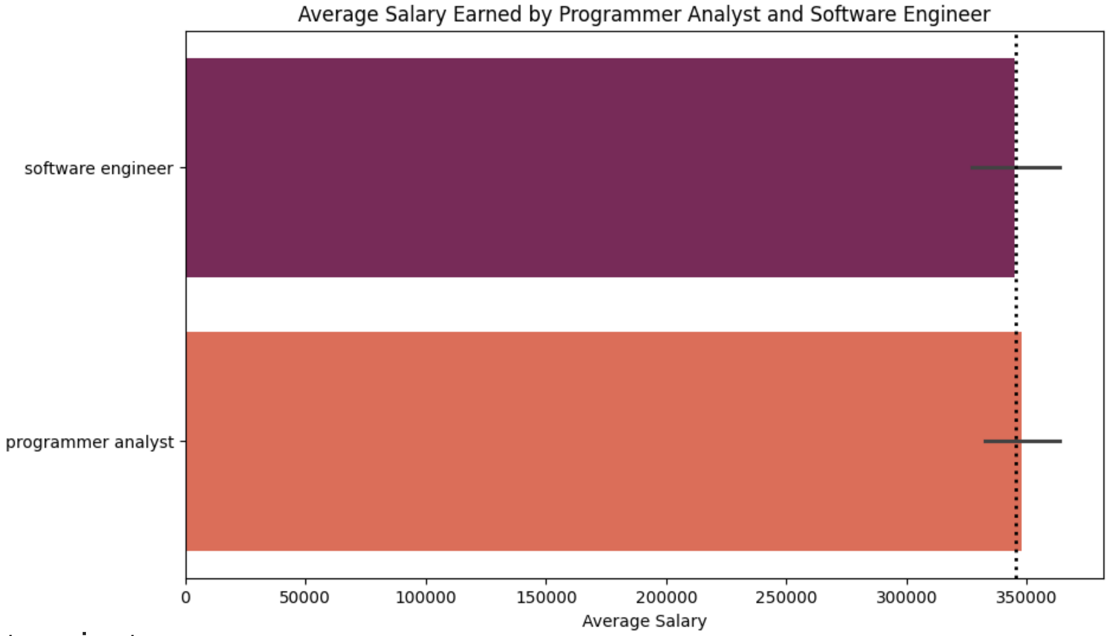
•Null Hypothesis (H0): The average salary of fresh graduates in Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer roles is ₹2.75 lakhs.

•Alternative Hypothesis (H1): The average salary of fresh graduates in these roles is significantly different from ₹2.75 lakhs.

This is a two-tailed test since we must check whether the actual salary is significantly different from ₹2.75 lakhs, either higher or lower.

- Output:

```
t-statistic: 2.763615869342021, p-value: 0.008030843099690977
```


Average Salary Earned by Programmer Analyst and Software Engineer

- With a t-statistic of 2.84 and a p-value of 0.0066, we have strong evidence to reject the null hypothesis that the average salary for fresh graduates in roles such as Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer is ₹2.75 lakhs (within the range of ₹2.5 to ₹3 lakhs).

- **Conclusion:** Both Programmer Analyst and Software Engineer positions show significantly higher salaries compared to the expected lower bound of the salary range (2.5-3 lakhs)

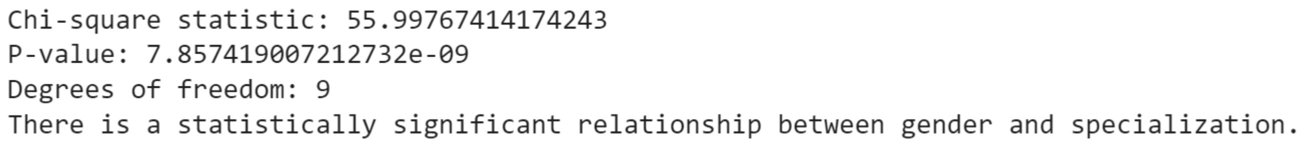
Q2. Is there a relationship between gender and specialization? (i.e. Does the preference of Specialization depend on the Gender?)

- Null Hypothesis (Ho): There is no association between the variables. The variables are independent.

- Alternative Hypothesis (H1): There is an association between the variables. The variables are not independent.

- Chi Square Test can be used to determine whether or not there exists a relation between the Gender and Specialization.

- Output:



```
Chi-square statistic: 55.99767414174243
P-value: 7.857419007212732e-09
Degrees of freedom: 9
There is a statistically significant relationship between gender and specialization.
```

Conclusion: The p-value is much less than the typical significance level (0.05 or 5%), indicating that we should reject the null hypothesis. This suggests that there is a statistically significant relationship between gender and specialization.

THANK YOU