INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

VANITA DESHMUKH

About me

- **Civil Engineering (BE)**
- **I want to learn Data Science because I'm interested in using information to solve problems and make things better.**
- **Any work experience (Yes, I have four years of experience in the construction industry.**
- **Share your linkedin and github profile urls**

linkedin
https://www.linkedin.com/in/vanitadeshmukh121/

github
https://github.com/VanitaDeshmukh

INNOMATICS
RESEARCH LABS

Agenda (This should be the PPT flow)

- **Business Problem and Use case domain understanding(If Required)**
- **Objective of the Project**
- **Web Scraping – Details (Websites, Processor you followed)**
- **Summary of the Data**

- **Exploratory Data Analysis:**

a. Data Cleaning Steps

b. Data Manipulation Steps

c. Univariate Analysis Steps

d. Bivariate Analysis Steps

- **Key Business Question**
- **Conclusion (Key finding overall)**
- **Q&A Slide**
- **Your Experience/Challenges working on Web Scraping – Data Analysis Project.**

Business Problem Statement

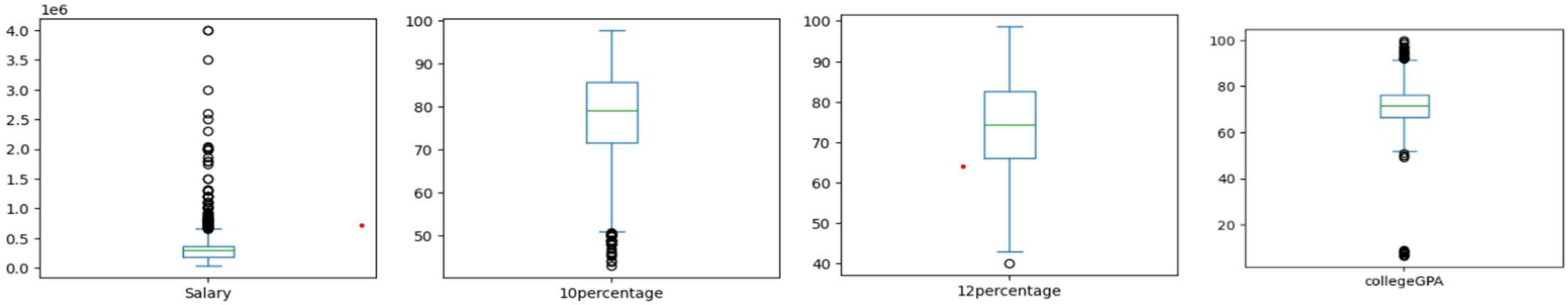
- **Exploratory Data Analysis (EDA):**
  - Analyzing distributions of variables.
  - Investigating relationships between variables.
- **Research Questions:**
  - Testing salary claim.
  - Exploring gender-specialization relationship.
- **Conclusion:**
  - Summarizing key findings.
  - Providing recommendations.

Objective of the Project:

The objective of the project is to conduct exploratory data analysis (EDA) on the Aspiring Minds Employment Outcome 2015 (AMEO) dataset to:

- Gain insights into the employment outcomes of engineering graduates.
- Understand the factors influencing salary expectations and specialization preferences among graduates.
- Provide recommendations for optimizing recruitment strategies and improving graduate outcomes based on data-driven insights.

Univariate Numerical Analysis



Salary:

- Salary has a wide range of values, with the minimum salary being $35,000 and the maximum being $4,000,000.
- The mean salary is approximately $307,700, with a median of $300,000, indicating a positively skewed distribution.

10th Percentage:

- 10th percentage ranges from 43% to 97.76%, with a mean of approximately 77.93% and a median of 79.15%.
- Overall, the distribution seems relatively symmetric and centered around the mean.

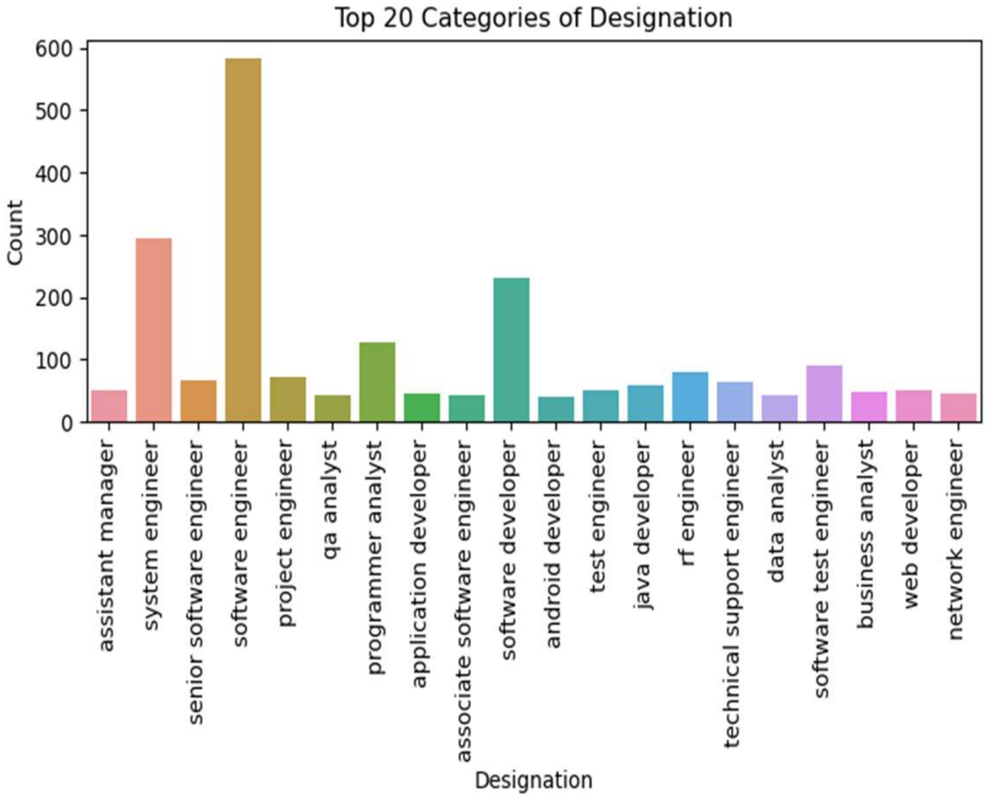
12th Percentage:

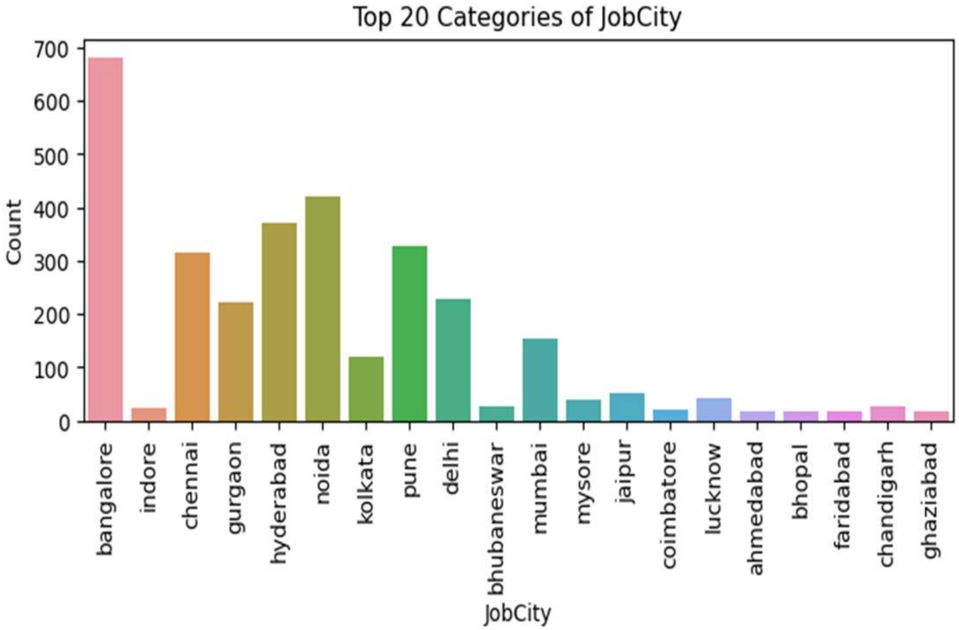
- 12th percentage varies from 40% to 98.7%, with a mean of approximately 74.47% and a median of 74.4%.
- The distribution appears to be approximately symmetric (skew = -0.03) with minimal kurtosis (kurt = -0.63).

College GPA:

- College GPA ranges from 6.45 to 99.93, with a mean GPA of approximately 71.49 and a median of 71.72.
- The distribution is negatively skewed (skew = -1.25) and exhibits high positive kurtosis (kurt = 10.23), indicating a significant departure from normality.

Univariate Categorical Analysis

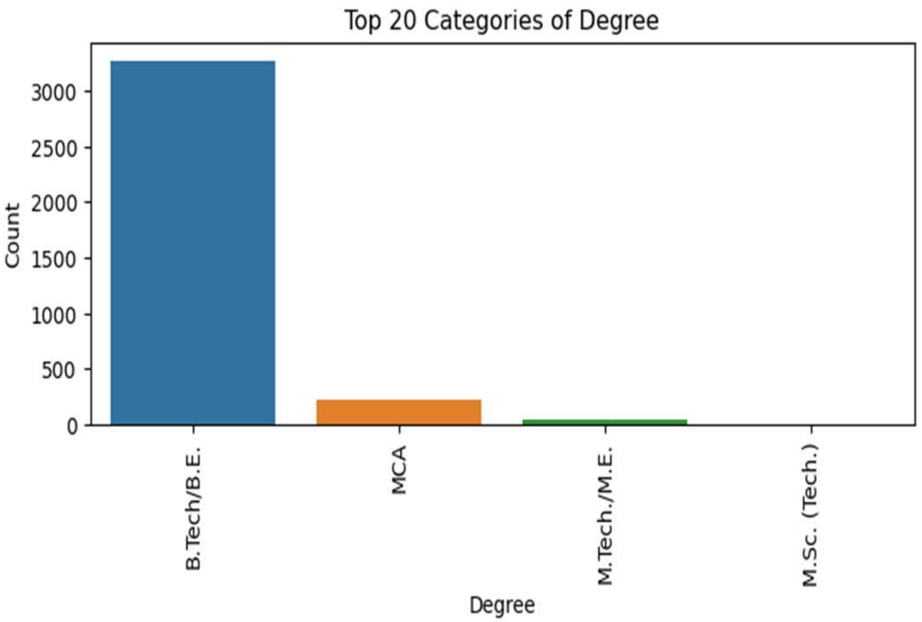


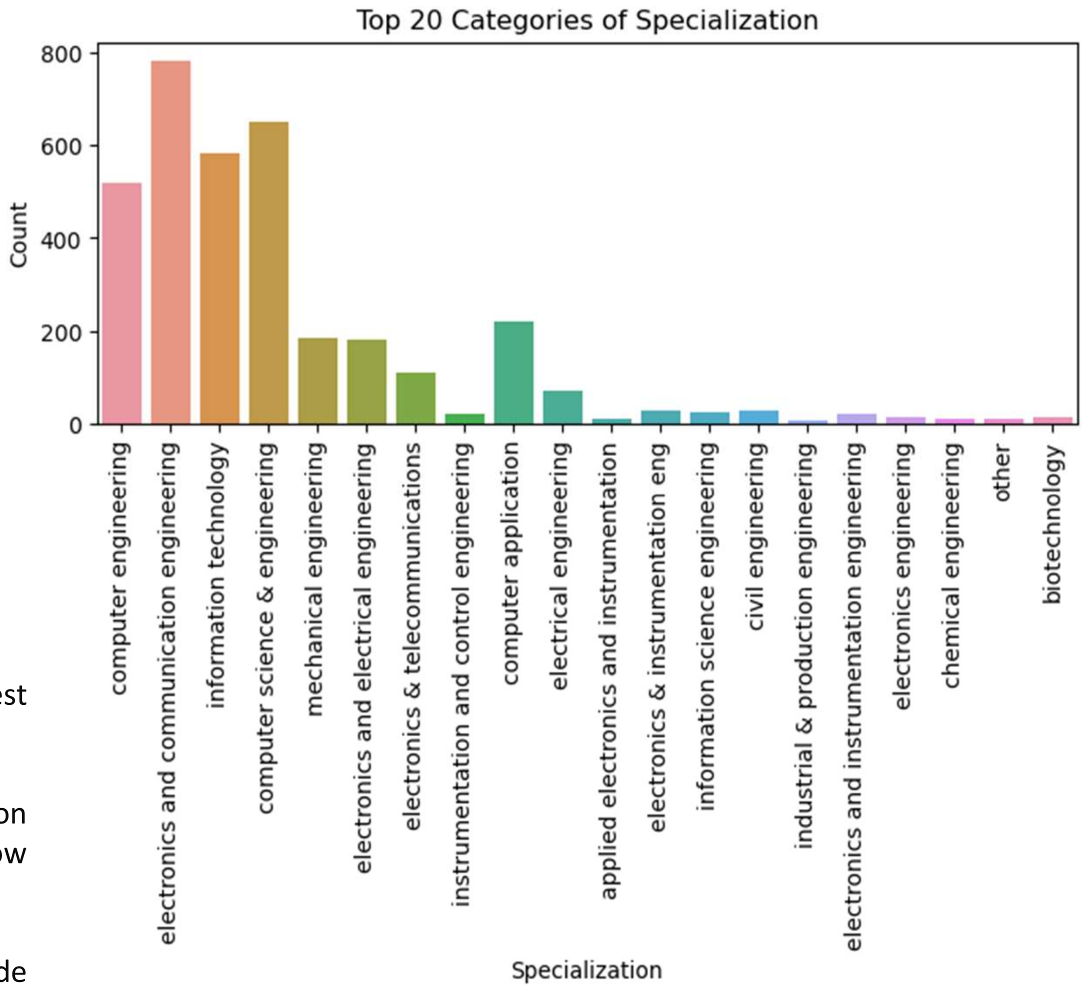


- Among the dataset's job titles or roles, the most frequent designation is 'Software Engineer', with a count ranging from 500 to 600. Conversely, ‘android developer' and ‘data analyst have the lowest counts compared to other designations.

- Regarding job locations, the highest count of jobs is observed in Bangalore, ranging from approximately 600 to 700. Conversely, Bhopal, gaziabad, farindabad, Ahmedabad & coimbatore has the lowest count of jobs compared to other cities.

INNOMATICS RESEARCH LABS



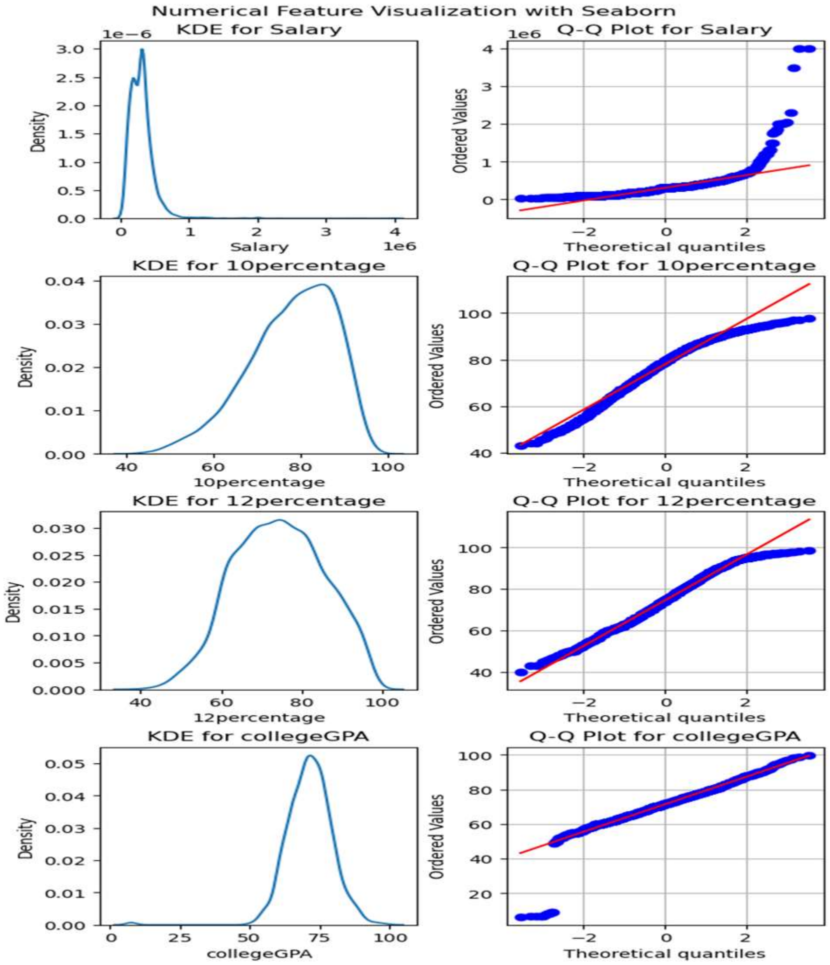


- The top 20 degree plot reveals that B.Tech/BE boasts the highest count, while M.Sc. (Tech) exhibits the lowest count.

- Among specializations, electronics and communication engineering, along with computer science & engineering, show the highest counts.

- Conversely, the specializations with the lowest counts include chemical engineering, industrial & production engineering, and other disciplines.

Univariate Numerical Analysis



Salary:

•Shapiro-Wilk test indicates a significant departure from normality ($p < 0.05$), suggesting the distribution is likely not Gaussian.

•The visualization reveals potential outliers with salaries significantly higher than the median, indicating possible high-income earners.

10percentage:

•Shapiro-Wilk test shows a significant departure from normality ($p < 0.05$), indicating a non-Gaussian distribution.

•Histogram suggests a skewed distribution, potentially indicating variations in academic performance among students.

12percentage:

•Shapiro-Wilk test demonstrates a significant deviation from normality ($p < 0.05$), implying a non-Gaussian distribution.

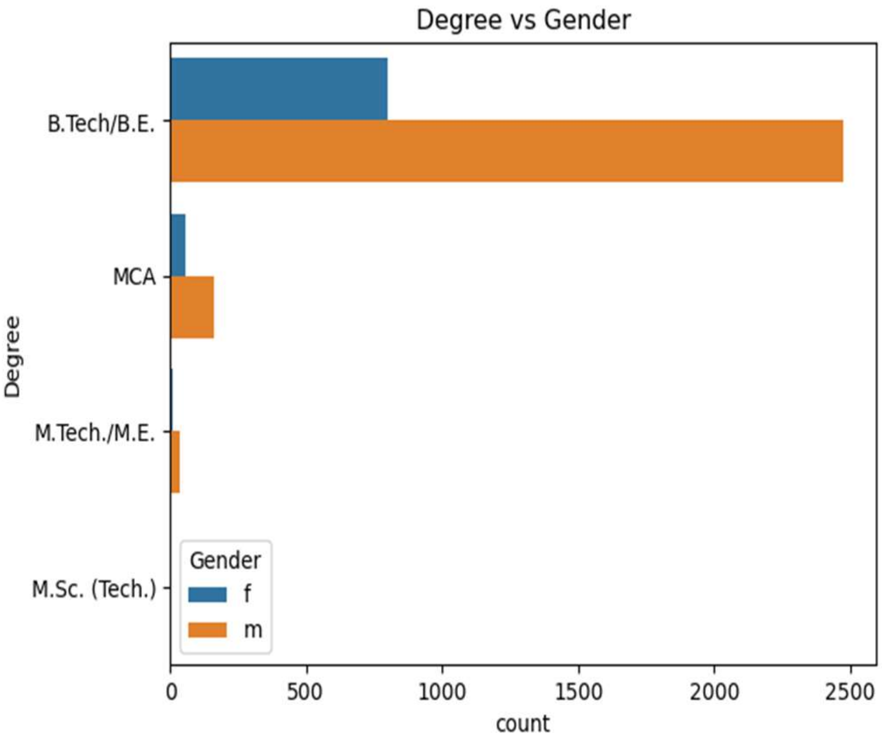
•Bimodal distribution observed in the histogram indicates the presence of two distinct peaks, suggesting differences in academic performance among students.

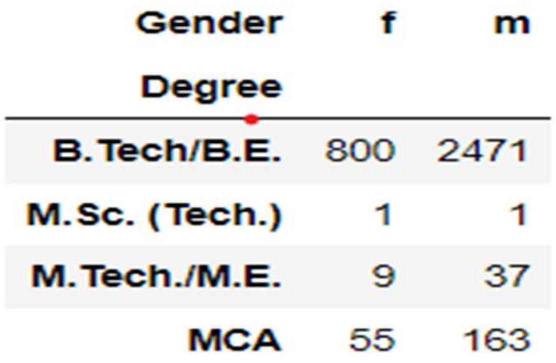
collegeGPA:

•Shapiro-Wilk test reveals a significant departure from normality ($p < 0.05$), indicating a non-Gaussian distribution.

•Probability plot deviations from the diagonal line suggest departures from normality, implying the distribution may not follow a perfect normal distribution.

Bivariate Analysis - Categorical vs Categorical





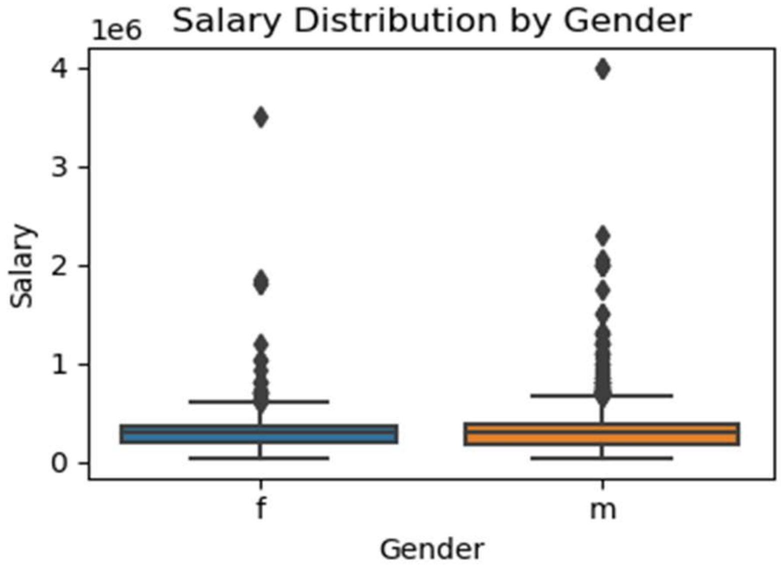
Gender	f	m
Degree		
B.Tech/B.E.	800	2471
M.Sc. (Tech.)	1	1
M.Tech./M.E.	9	37
MCA	55	163

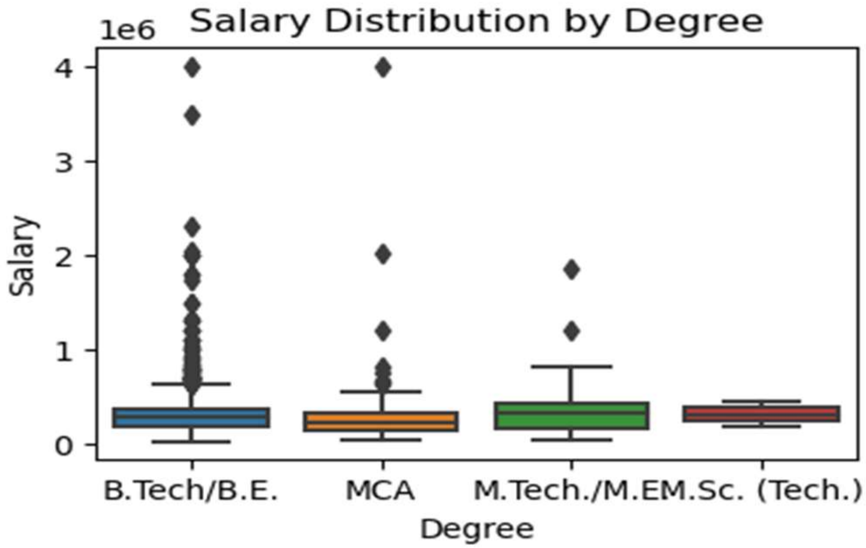
- The highest number of males and females hold a degree in B.Tech/BE compared to other degrees.

- The chi-square statistic is 2.172 with a p-value of 0.538. At the 0.05 significance level, the null hypothesis (H0) is not rejected.

- Thus, there is insufficient evidence to suggest a relationship between individuals' degree and gender.

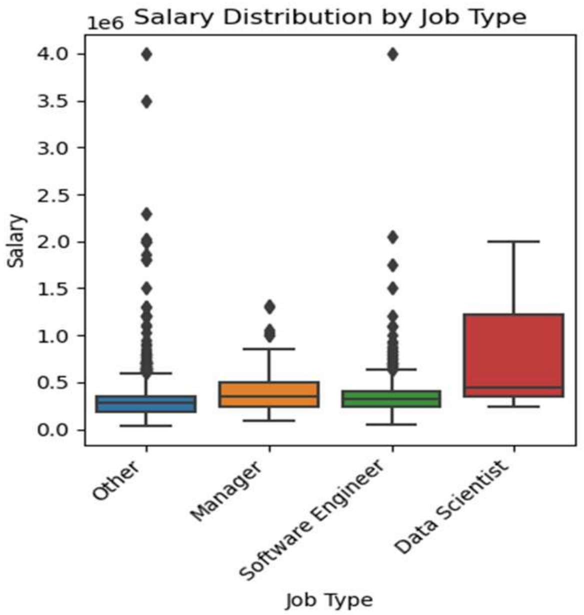
Bivariate Analysis - Categorical vs Categorical

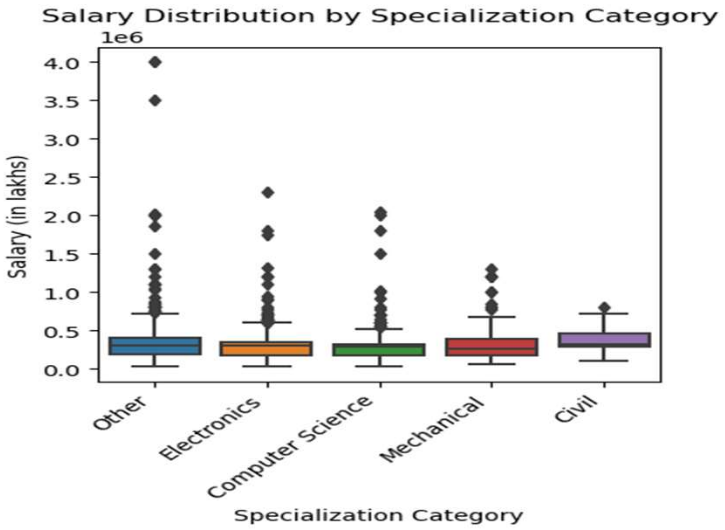


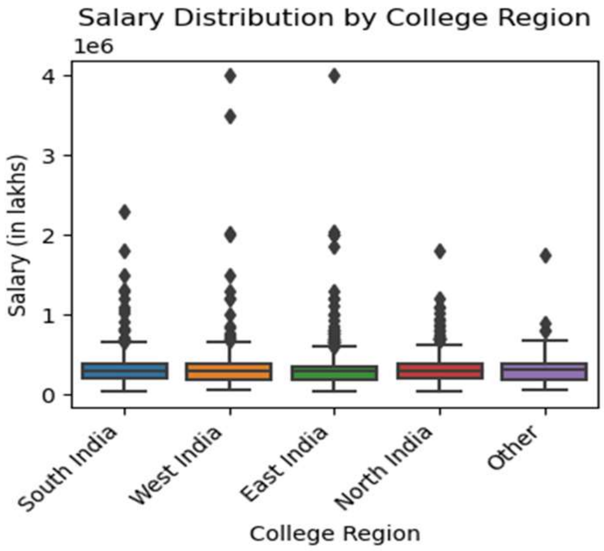


- The salary distribution by degree plot indicates that the highest amount of salary is observed for individuals with B.Tech/BE and MCA degrees, with a predominant range of 400,000/Annum, compared to other degrees.
- The 25th, 50th (median), and 75th percentiles of the salary distribution across all degrees show approximately equal values, ranging from 0.3 to 0.5.
- The salary distribution by gender reveals that the highest salaries are found among males, with a peak in the 400,000 range, compared to females.
- The mean salary is approximately equal for both males and females.

INNOMATICS RESEARCH LABS







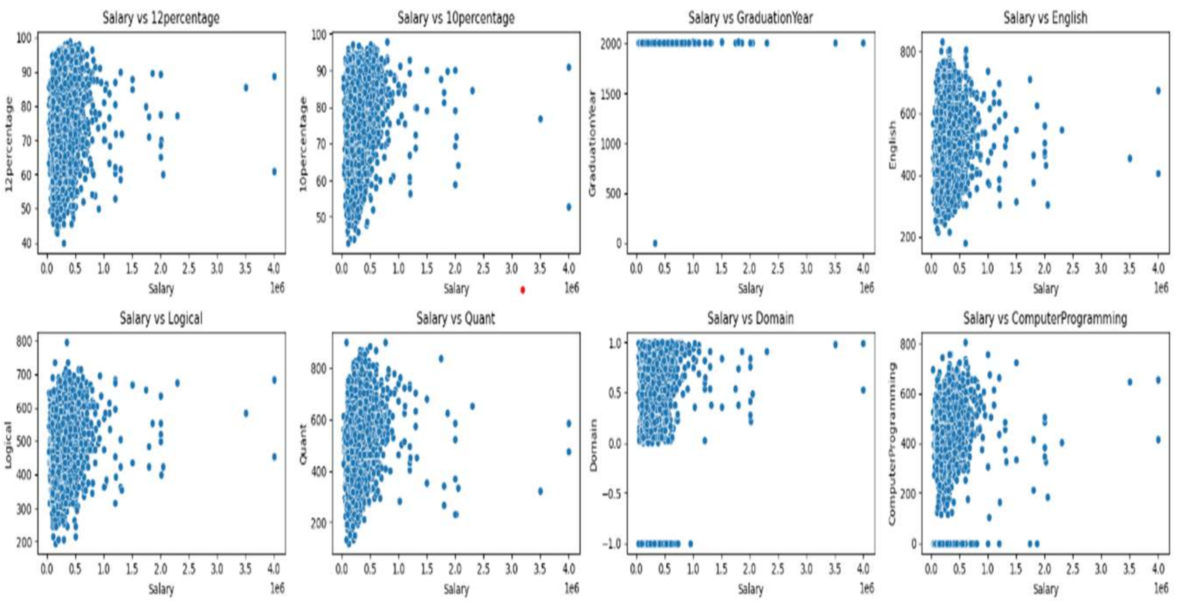
- **Salary Distribution by Job Type:Data Scientists:**
  - Highest Salary Range: 2.0
  - Mean Salary: 0.5
  - 75th Percentile: 1.0
  - No outliers observed

- **Other Job Types & Software Engineers:**
  - Highest Salary Range: Up to 4.0 (with outliers)
  - Mean Salary: Typically in the range of 0 to 0.5
  - 75th Percentile: Around 0.4

- **Salary Distribution by Specialization:**
  - Mean salaries across specializations are approximately equal.
  - The maximum salary is observed in the "Other" specialization, surpassing those in Electronics, Computer Science, Mechanical, and Civil.
- **Salary Distribution by College Region:**
  - West India and East India exhibit the highest salaries compared to other regions.
  - Mean salaries remain consistent across all regions.

INNOMATICS
RESEARCH LABS

Bivariate Analysis - Numerical vs Numerical



Positive Linear Relationships:

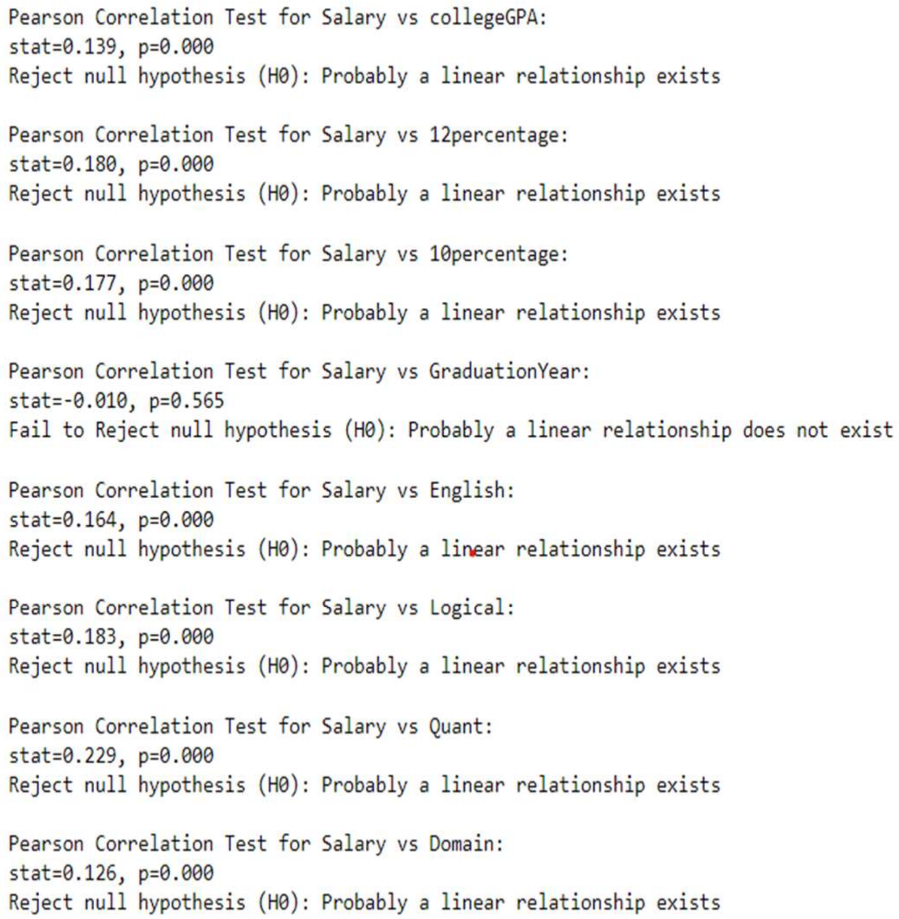
- 'Salary' exhibits statistically significant positive linear relationships with:
- 'college GPA', '12percentage', '10percentage', 'English', 'Logical', 'Quant', 'Domain', 'Computer Programming', 'Civil Engg', 'conscientiousness', 'agreeableness', and 'nueroticism'.
- Higher values in these columns tend to correspond to higher salaries.

No Significant Linear Relationships:

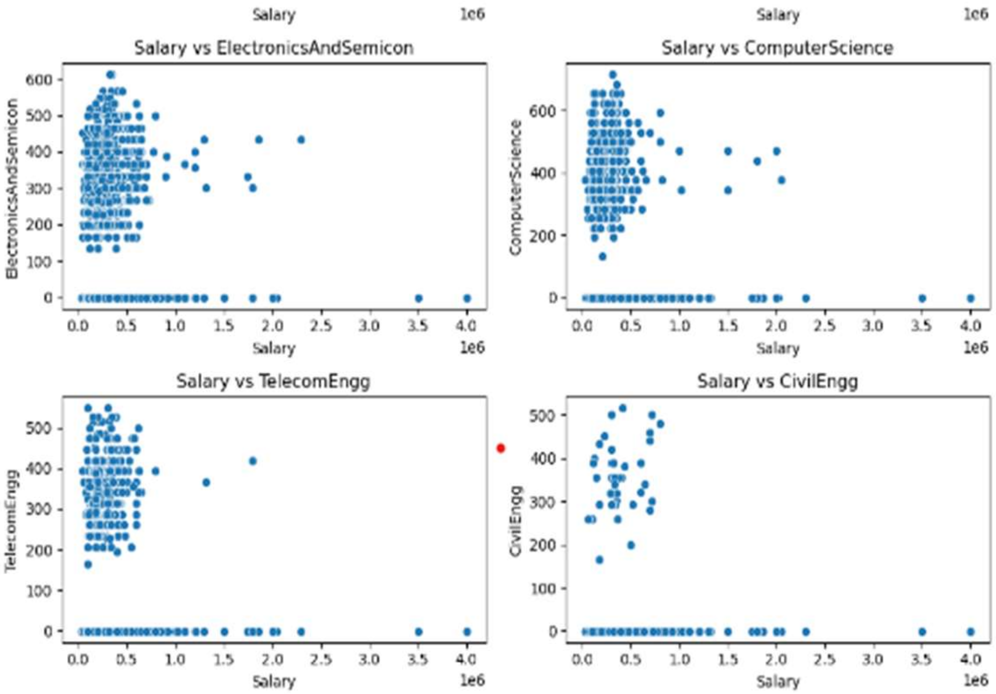
- 'Salary' does not show statistically significant linear relationships with:
- 'Graduation Year’, 'Electronics And Semicon', 'Mechanical Engg', 'Telecom Engg', 'extraversion', and 'openness to experience'.
- Salary does not notably vary based on these factors.

Negative Linear Relationships:

- 'Salary' exhibits statistically significant negative linear relationships with:
- 'Electrical Engg'.
- Higher values in this column tend to correspond to lower salaries.



```
Pearson Correlation Test for Salary vs collegeGPA:
stat=0.139, p=0.000
Reject null hypothesis (H0): Probably a linear relationship exists

Pearson Correlation Test for Salary vs 12percentage:
stat=0.180, p=0.000
Reject null hypothesis (H0): Probably a linear relationship exists

Pearson Correlation Test for Salary vs 10percentage:
stat=0.177, p=0.000
Reject null hypothesis (H0): Probably a linear relationship exists

Pearson Correlation Test for Salary vs GraduationYear:
stat=-0.010, p=0.565
Fail to Reject null hypothesis (H0): Probably a linear relationship does not exist

Pearson Correlation Test for Salary vs English:
stat=0.164, p=0.000
Reject null hypothesis (H0): Probably a linear relationship exists

Pearson Correlation Test for Salary vs Logical:
stat=0.183, p=0.000
Reject null hypothesis (H0): Probably a linear relationship exists

Pearson Correlation Test for Salary vs Quant:
stat=0.229, p=0.000
Reject null hypothesis (H0): Probably a linear relationship exists

Pearson Correlation Test for Salary vs Domain:
stat=0.126, p=0.000
Reject null hypothesis (H0): Probably a linear relationship exists
```



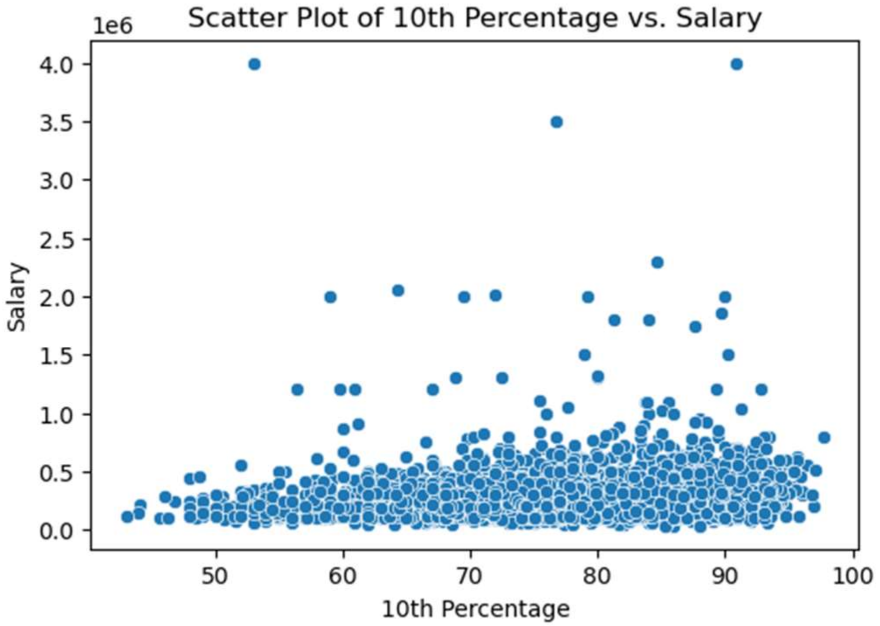
Research Question

Research Question 1: Correlation between academic performance and salary

Correlation between academic performance and salary:

- The scatter plot of 10th percentage versus salary indicates a potential positive correlation between academic performance in the 10th grade and salary levels.

- As the 10th percentage increases, there seems to be a trend of higher salaries, suggesting that individuals with better performance in their 10th grade examinations may earn higher salaries.

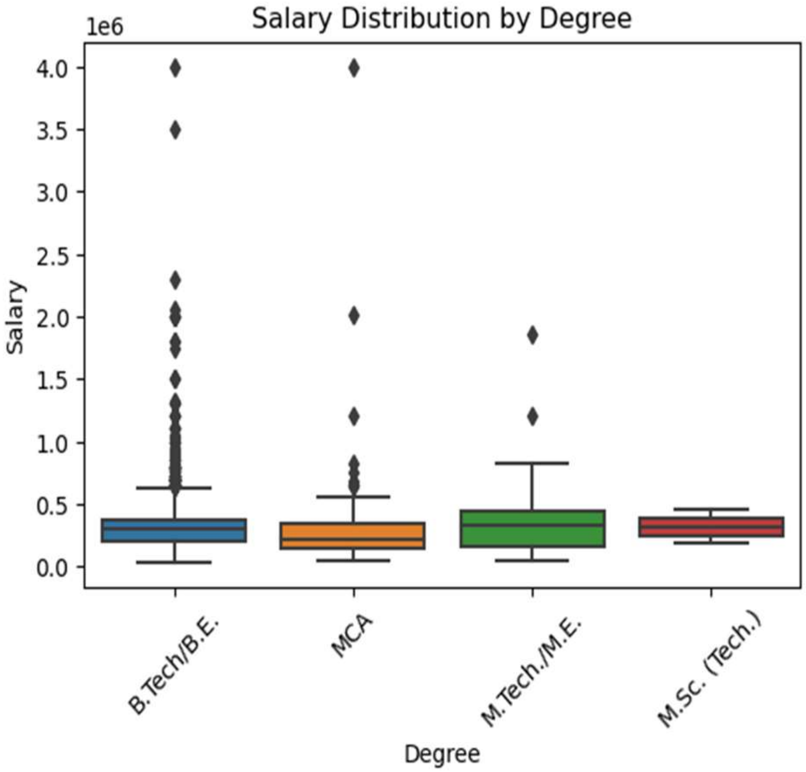


Research Question 2: Salary distributions across different educational degrees

OBSERVATION OF **Salary Distribution by Degree:**

- The plot shows that individuals with degrees in B.Tech/BE and MCA tend to have higher salaries compared to other degrees.

- It indicates a potential correlation between educational qualification and salary level, with B.Tech/BE and MCA leading in terms of salary distribution.

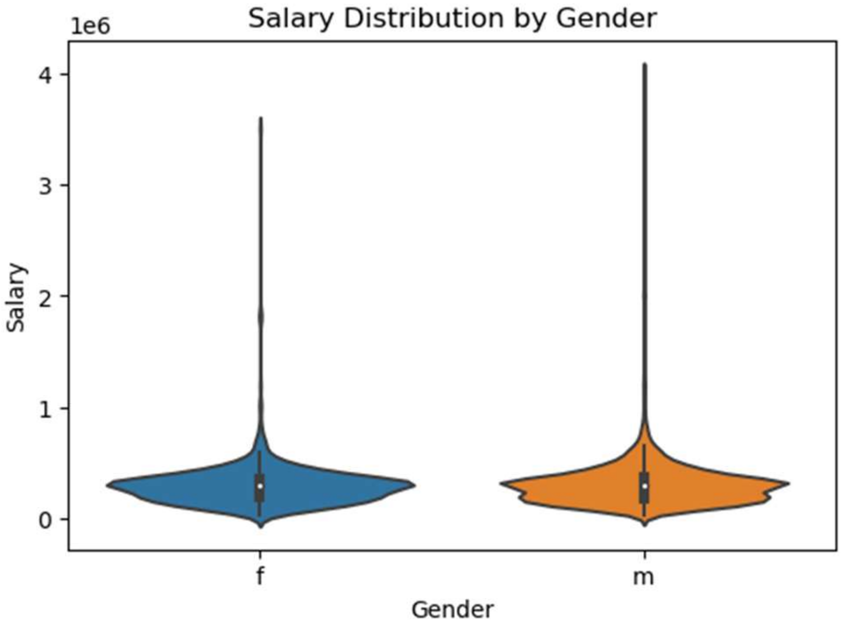


Research Question 3: Impact of gender on salary levels

OBSERVATION OF **Salary Distribution by Gender:**

- The plot suggests that males generally have higher salaries compared to females, as the highest salary range is more prominent for males.

- This highlights potential gender disparities in salary levels, warranting further investigation into factors contributing to this gap.

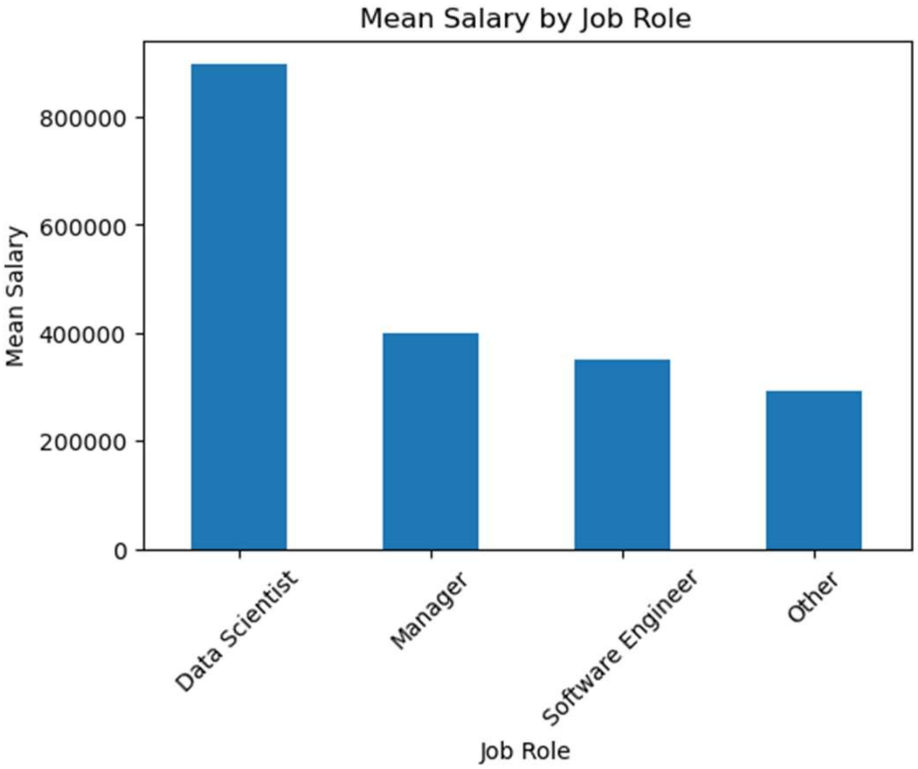


Research Question 4: Salary distributions across different job roles

OBSERVATION OF **Salary Distribution by Job Type**:

- Data Scientists appear to have the highest salary range, indicating the demand and premium associated with this role.

- The presence of outliers in other job types suggests variability in salary levels within those roles, possibly due to factors like experience, skills, or industry.



Conclusion

- **Degree and Gender Distribution:**
  - B.Tech/BE is the most common degree for both males and females.
  - The chi-square test suggests no significant relationship between degree and gender.
- **Salary Distribution:**
  - The highest salaries are typically observed for B.Tech/BE and MCA graduates, with outliers indicating exceptionally high earners.
  - Salary distributions by gender show comparable median salaries, but males tend to have higher maximum salaries.
  - Data scientists command the highest salaries, with other job types also exhibiting high maximum salaries but with outliers.
- **Academic Performance:**
  - Academic performance metrics such as 10th, 12th percentages, and college GPA demonstrate non-Gaussian distributions, with significant departures from normality.
  - Variations in academic performance are evident, indicated by skewed distributions and bimodal patterns.
- **Specialization and College Region Impact:**
  - While mean salaries are consistent across different specializations and regions, outliers contribute to variations in maximum salaries.
  - West India and East India regions exhibit higher salaries compared to others, suggesting regional impacts on salary trends.
- **Overall Implications:**
  - The analysis provides insights into the distribution of salaries, academic performance, and their relationships with degrees, gender, specializations, and regions.
  - These findings can inform strategic decisions for individuals, educational institutions, and employers in understanding and navigating the job market landscape.

THANK YOU

INNOMATICS
RESEARCH LABS