Huda Maniyar

AMCAT ENGINEERING EMPLOYMENT ANALYSIS

1. DATA DESCRIPTION

The AMCAT dataset offers a comprehensive overview of the educational backgrounds, skill sets, and employment profiles of engineering graduates, consisting of 39 variables and 3,998 records, gathered as part of the Aspiring Mind Employment Outcome (AMEO) 2015 study. This dataset covers several key areas:

- **Demographics**: Data includes individuals' gender, birth date, and geographic location, offering foundational context for analysis.
- **Academic Background**: Information on high school grades, college GPA, graduation year, and specialization gives insights into academic performance and areas of expertise.
- **Employment Details**: The dataset records salary, job designations, job cities, and employment timelines (date of joining and leaving), helping track career progression and tenure.
- **Skill Evaluation**: Scores in English, logical reasoning, quantitative aptitude, and domain-specific knowledge evaluate the candidates' job-relevant competencies.
- **Personality Traits**: Measures of conscientiousness, agreeableness, extraversion, neuroticism, and openness provide a psychometric assessment that can influence job performance.

2. OBJECTIVES

- Analysing Employment Patterns: Examine the correlation between education, skills, and employment outcomes, including salary and job placements.
- Conduct Exploratory Data Analysis (EDA): Perform in-depth EDA using univariate, and bivariate techniques to uncover factors affecting salary and employment.
- Creating Data Visualizations: Develop charts and graphs to visualize trends in demographic, educational, and employment-related variables.

3. EXPLORATORY DATA ANALYSIS

This section will present an overview of the data and key observations, followed by data cleaning, univariate and bivariate analyses, and a discussion of the research questions.

3.1. Data Overview (Initial Inspection)

1. Data Preview - $head()$
2. Data Dimension - $shape$
3. Data Summary - $info()$
4. Summary Statistics - $describe()$
5. Data Columns - $columns$

Observation

- The dataset comprises of 39 variables and 3998 records.
- Columns to be **Dropped** (not required for the purpose of this project):
    1. **Unnamed:0**
    2. **CollegeID**
    3. **CollegeCityID**
- Fix Column **DataTypes**:
    1. **ID** column to be changed from int to object datatype.
    2. **Salary** column to be change from int to float datatype.
    3. **CollegeTier** and **CollegeCityTier** to be changed from int to object type.
    4. **12Graduation**, **GraduationYear** columns to be changed from int to date datatype.

5. **DOL** column to be further investigated and changed from object to datetime datatype.

- Missing Values:
    1. **10Board**, **12Board**, **GraduationYear** columns contain values like **0**, indicating missing values and need to further be dealt with.
    2. **JobCity** column contains **-1** which implies that these could be missing values.
    3. Numerical Columns like **Domain**, **Skill Assessment** also contain **-1** which need to be further investigated.
    4. **CollegeState** Column contains **Union Territory** values which needs to be further investigated.
- Standardize Column Values:
    1. **10Board** & **12Board** - [State, CBSE, ICSE]
    2. **JobCity**
    3. **Specialisation** - ['CS', 'EC', 'ME', 'EL', 'CE', 'Other']
- Add Columns:
    1. Age
    2. Experience
- Standardize Column Names and values by stripping any extra space and converting them to title case.

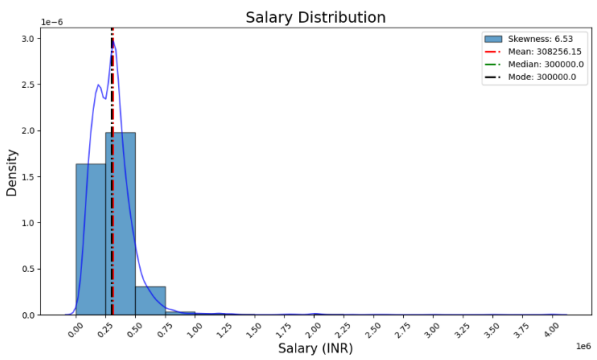
3.2. Data Cleaning

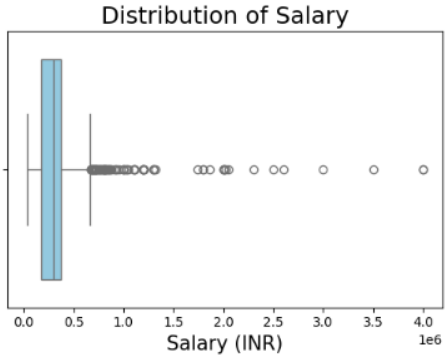
1. Removing Columns
2. Standardizing Column Names
3. Dealing with Missing Values
4. Fixing Column Data Types
5. Standardizing Column Values
6. Adding Columns

3.3 Univariate Analysis

Numerical Variables - Generate statistical summaries for each numerical column individually to understand their distributions and key statistics.

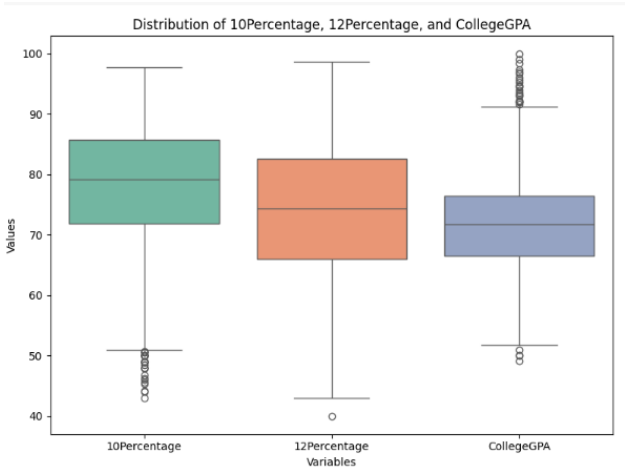
1. Salary





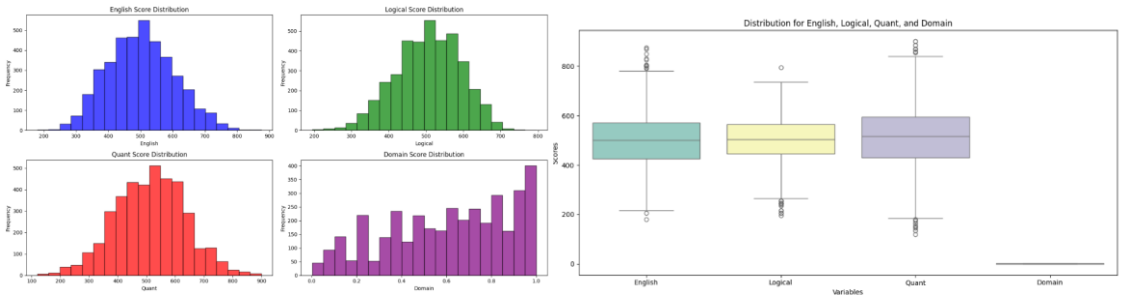
- The mean salary of ₹308,256 is slightly higher than the median of ₹300,000, indicating a right-skewed distribution with a range from ₹35,000 to ₹4,000,000.
- A standard deviation of ₹211,763 highlights significant variation around the mean, with 50% of individuals earning between ₹180,000 (25th percentile) and ₹372,500 (75th percentile).
- The box plot reveals a high number of outliers, suggesting considerable disparities in salaries among individuals.

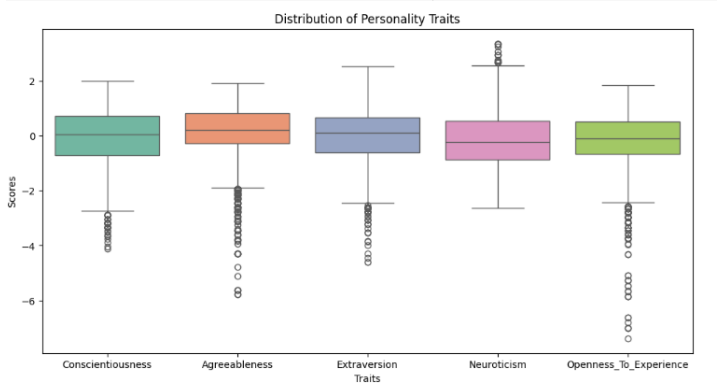
2. 10Percentage, 12Percentage, CollegeGPA



- **10Percentage** and **12Percentage** show similar distribution patterns, though students performed slightly better in 10th grade than in 12th grade on average.
- **CollegeGPA** has a slightly lower mean compared to the high school percentages, but it shows less variability, indicating a more consistent academic performance during college.

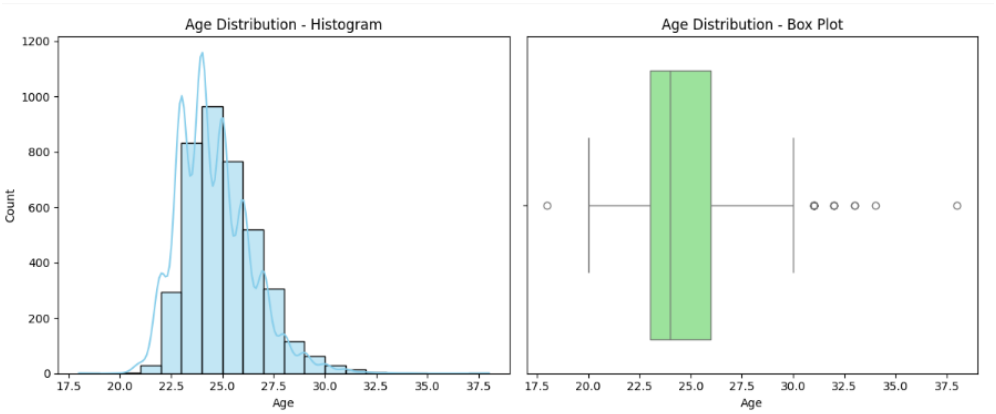
3. AMCAT Scores





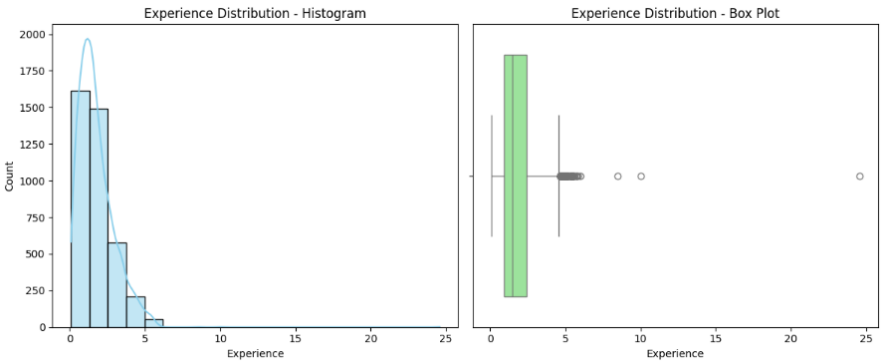
- The average scores for **English** (501.79), **Logical** (501.61), and **Quant** (513.31) reflect significant variability in student proficiency across these core subjects, with Quant demonstrating the highest average performance.
- The average **Domain** score of 0.61 reflects varying participation and performance levels due to students selecting different fields based on their specializations and interests.
- The engineering-related variables exhibit fewer data points and high proportions of missing values (over 70%), indicating that visual analysis may not yield meaningful insights due to insufficient data.
- **Personality traits** exhibit significant variability, with conscientiousness and neuroticism showing low average scores, likely reflecting the diverse backgrounds and experiences of respondents.

4. Age



- The average age is approximately **24** years, indicating that most individuals in the dataset are in their **mid-twenties**.
- The standard deviation of 1.77 indicates a relatively tight spread, meaning most individuals are close in age. About 68% of the ages fall between 22 and 26 years.
- The minimum age is 18, and the maximum is 38, which suggests there may be a few outliers above the age of 30.

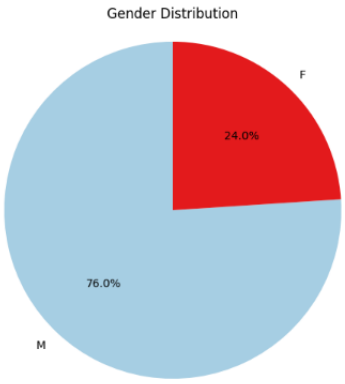
5. Experience



- The mean experience is 1.28 years, while the median is 1 year, indicating most individuals have around **1 year** of experience.
- A standard deviation of 1.23 years shows a moderate spread, with most individuals having relatively **low work experience**.
- The interquartile range (IQR) is between 0 and 2 years, suggesting many are **fresh graduates** or **early** in their careers.
- The maximum experience is 24 years, an extreme outlier, as the 75th percentile is only 2 years, showing a few individuals with significantly higher experience.

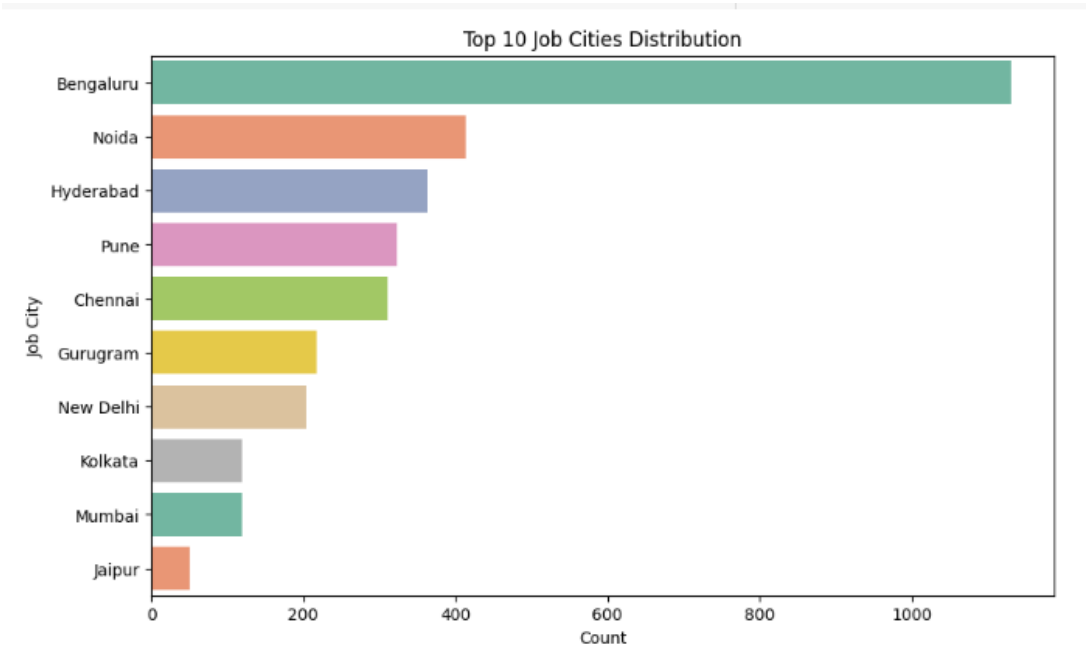
Categorical Variables - Generate frequency distributions for each categorical column to understand the distribution of categories.

1. Gender



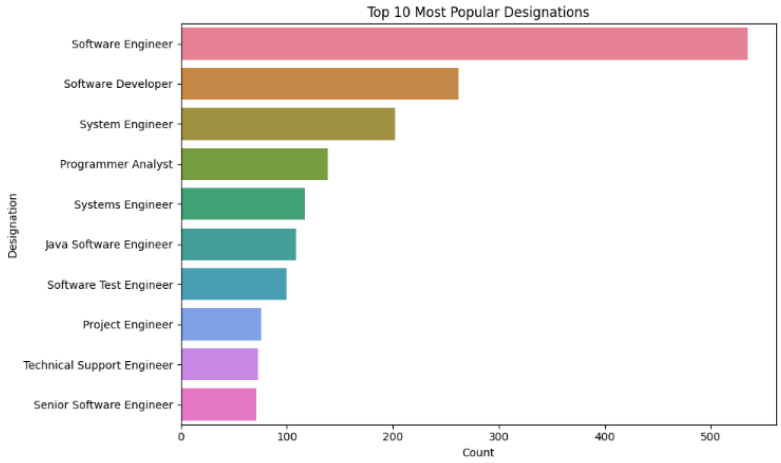
- **76%** of the candidates are **male** (2998 candidates), while **24**% of the candidates are **female** (945 candidates).
- This indicates a significant **gender imbalance**, with nearly three times more male candidates than female candidates in the dataset.

2. JobCity



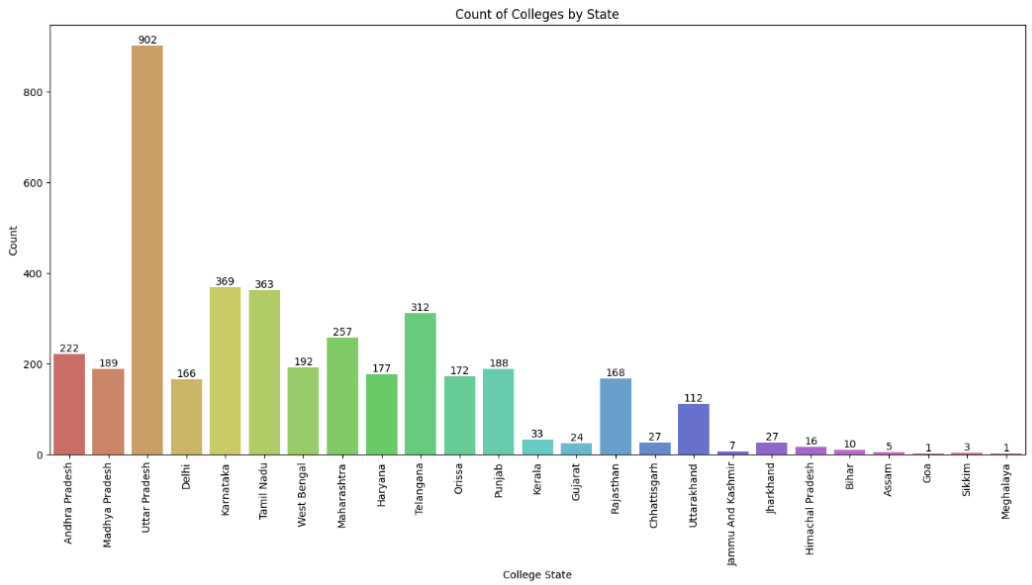
- The bar chart illustrates the distribution of job locations across various cities.
- A key insight from the plot is that **Bangalore** stands out as the most popular job location, with over 1,000 employees, making it the top provider of jobs among the cities displayed.
- Following Bangalore, Noida and Hyderabad rank as the second and third most common job locations, respectively, while Jaipur ranks lowest among the top 10 cities.

3. Designation



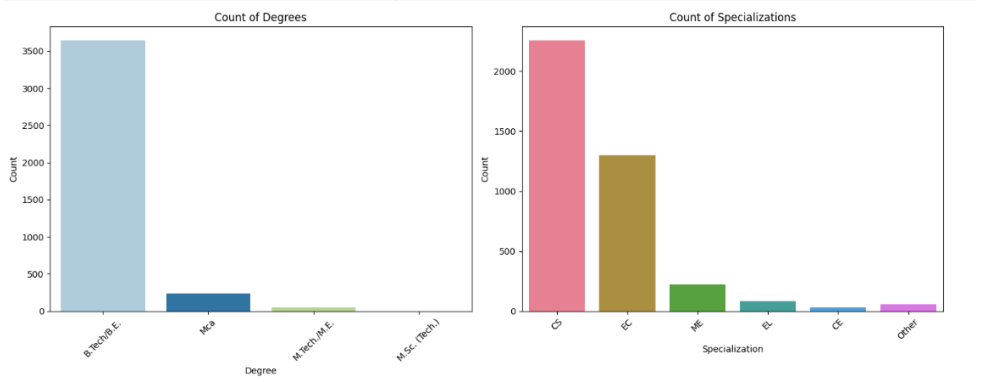
- Since there are numerous designations in the dataset, only the top 10 or most popular designations were analysed.
- The results show that **Software Engineer** is the most common designation among candidates.

4. CollegeStat



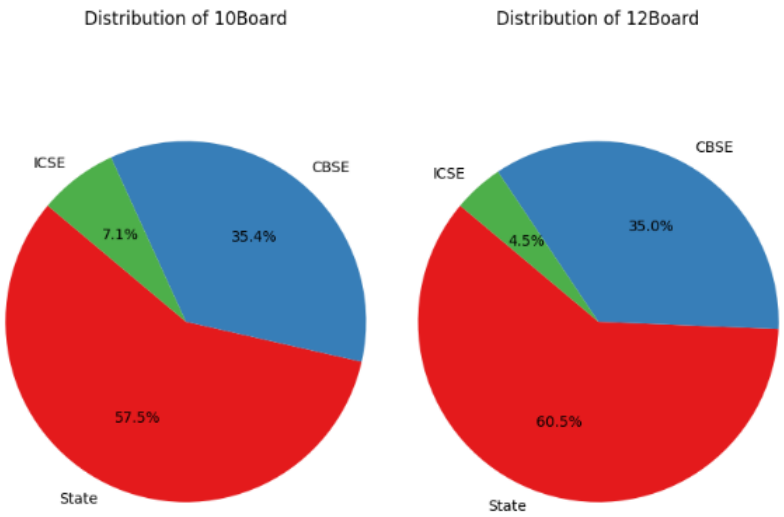
- A large number of employees in the dataset graduated from colleges in **Uttar Pradesh**.

5. Degree & Specialisation



- From the bar chart, **B.Tech/B.E** is the dominant degree with over 3,500 occurrences, while MCA follows with significantly fewer entries.
- M.Tech/M.E and M.Sc.(Tech.) have minimal representation, emphasizing the overwhelming prevalence of B.Tech/B.E degrees.
- In terms of specialization, **Computer Science** (CS) is the most popular, followed by Electronics and Communication (EC), reflecting a higher interest or demand in these fields.

6. 10Board & 12Board

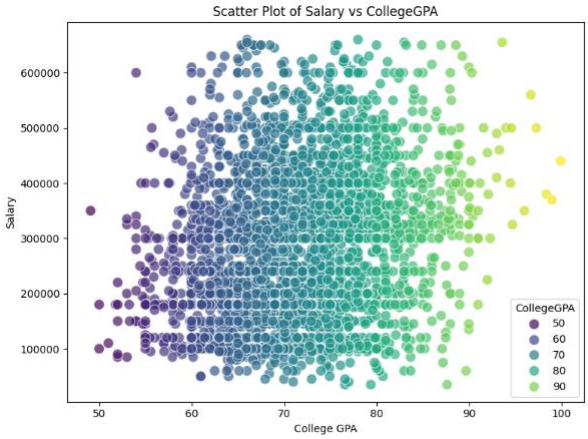


- After categorizing 10th and 12th board exams into State, CBSE, and ICSE boards, the majority of students come from **State** boards in both 10th and 12th grades.

3.4. Bivariate Analysis

Numerical vs Numerical

1. Salary vs CollegeGPA

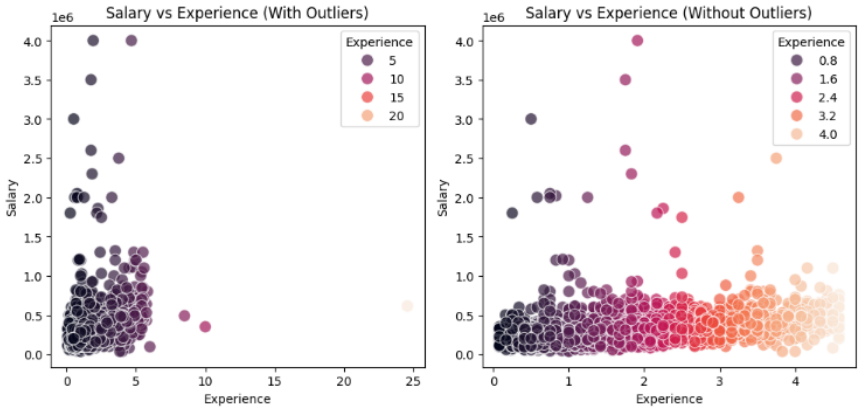


- Salaries vary widely across all GPA levels, with most salaries falling below 400,000, showing an even spread regardless of GPA.
- High salaries are observed across different GPA levels, suggesting that **College GPA alone may not strongly determine Salary**, though some high-GPA individuals earn higher salaries.
- Most data points are concentrated in the **mid-range of GPAs (60-80)** and salaries up to 400,000, with only a few outliers at both extremes.

2. Salary vs Age

- Salaries vary widely across all age groups, with most data points clustered between ages 22 and 30, indicating salary distribution is not solely dependent on age.
- While younger individuals tend to have lower salaries, some outliers earn high salaries at younger ages. Overall, salary levels within each age group vary, showing no clear linear relationship between age and salary.

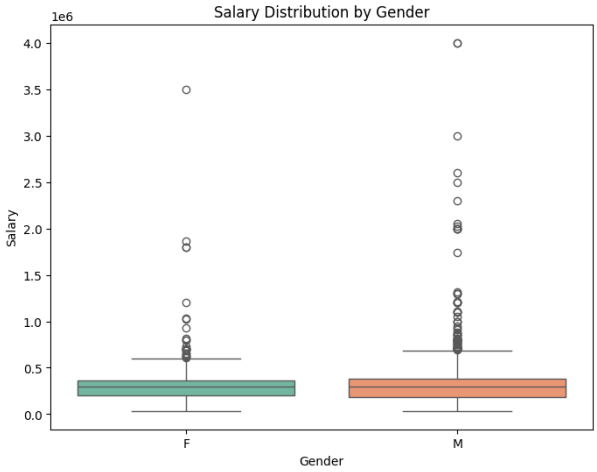
3. Salary vs Experience



- **With Outliers**: The left plot, including outliers, shows a skewed salary range for individuals with low experience (0-5 years), making it difficult to identify trends among the majority.
- **Without Outliers**: The right plot, which excludes outliers, reveals a clearer and more coherent pattern of salary growth with experience. Most salaries are concentrated under 1 million, highlighting a more realistic distribution for those with fewer years of experience.

Categorical vs Numerical

1. Salary vs Gender

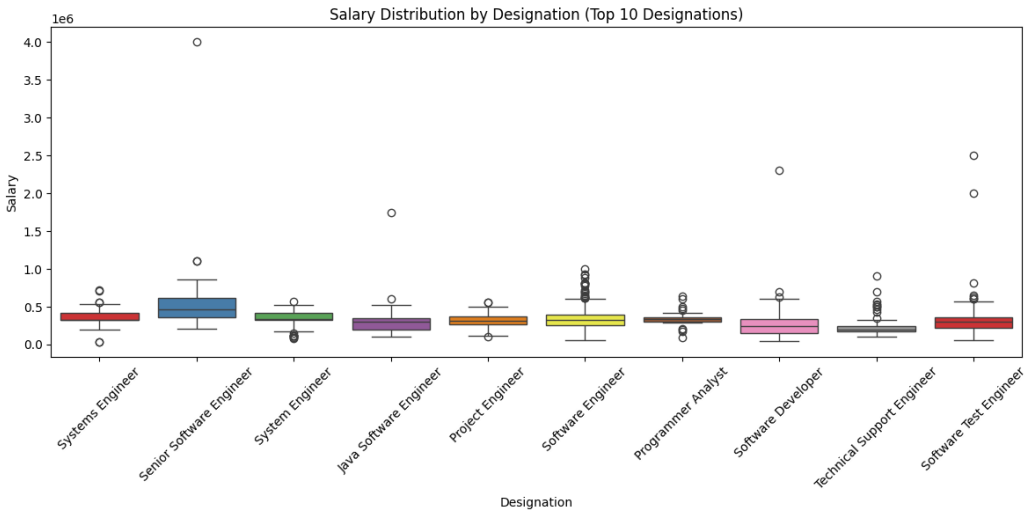


- Both genders have a similar median salary, as indicated by the central line inside the boxes, with little difference between male and female salaries at the median level.
- The interquartile range (IQR) is slightly wider for males, indicating more variability in salary for males compared to females. The whiskers for males extend further, showing a wider overall salary range.
- Both genders exhibit outliers, but there are more extreme high outliers among males, with a few notable cases showing significantly higher salaries for males compared to females.

2. Salary vs JobCity

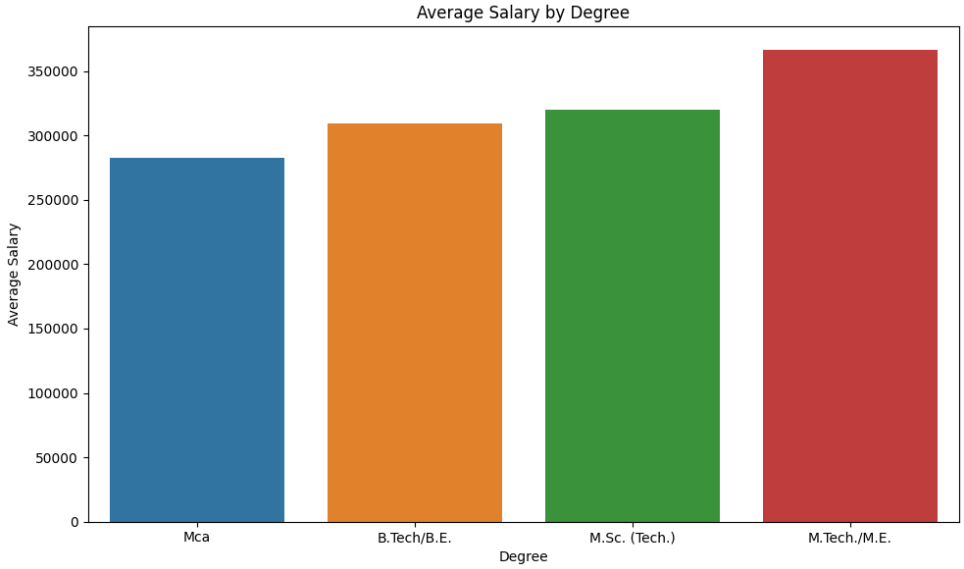
- The median salary across the top 10 cities is fairly consistent, with Bengaluru and Pune showing slightly higher medians compared to the rest of the cities.
- The range of salary variation (spread of the box and whiskers) differs by city, with cities like Bengaluru and Pune exhibiting wider salary ranges, while cities like Kolkata and Jaipur show relatively less variation.
- Most cities have outliers, but Bengaluru and Mumbai feature the most extreme high salaries, with Bengaluru having a notable outlier close to 4 million, indicating higher earning potential in Bengaluru.

3. Salary vs Designation



- The median salary differs significantly across the top 10 designations, with Senior Software Engineer having a noticeably higher median compared to other roles.
- Most designations show a wide salary range, with many outliers present, especially for Senior Software Engineer, Software Engineer, and Programmer Analyst, which have higher salaries and more extreme values.
- System Engineer has a relatively smaller salary range, with salaries clustered around lower values, while Java Software Engineer shows fewer extreme outliers but still exhibits a broad salary distribution.

4. Salary vs Degree



- Individuals with an M.Tech./M.E. degree have the highest average salary compared to other degrees, leading the chart.
- Degrees like B.Tech./B.E. and M.Sc. (Tech.) also result in relatively high average salaries, highlighting the value of technical education.
- MCA degree holders, while earning competitive salaries, have the lowest average salary among the degree groups presented.

5. Salary vs Specialisation

- Computer Science (CS) and other specializations have the highest median salary and show broader salary distribution ranges compared to others.
- Mechanical Engineering (ME) and Electrical Engineering (EL) have lower median salaries than CS but still display a significant range of salary variation.
- Civil Engineering (CE) shows a smaller salary range and a relatively lower median salary, with outliers present across most specializations, indicating wide salary variation depending on the role or company.

3.4. Research Questions

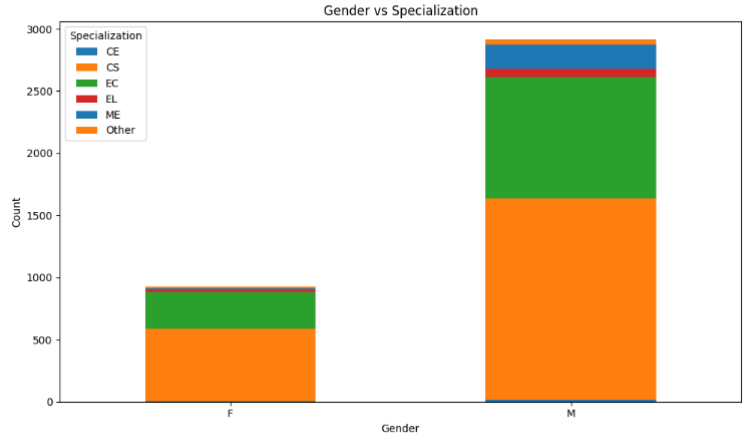
1. Validate Industry Salary Claims for Fresh Graduates

Average Salary for selected roles: ₹327042.91

The average salary does not fall within the claimed range.

2. Evaluate Gender-Specialization Relationship

```
Specialization  CE    CS   EC  EL  ME  Other
Gender
F                5   582  297  16  12     14
M               19  1617  978  64  197    38
```




```
Chi-squared Test Statistic: 46.77, p-value: 0.0000
There is a significant relationship between Gender and Specialization.
```

3. Are Top Cities Driving Higher Salaries?

Cities like Mumbai, Pune and Bengaluru provide higher-paying opportunities due to the concentration of high-tech industries, while others have more modest salary ranges.

4. Conclusion

- The salary distribution is right-skewed, with a mean salary of ₹308,256, indicating higher variability in salaries, as shown by numerous outliers and significant salary range.
- There is no strong correlation between College GPA and Salary, with high salaries observed across various GPA levels.
- Age and salary are not directly related, but younger individuals tend to have lower salaries, with a few outliers earning significantly higher salaries at younger ages.
- Salary increases with experience, although extreme outliers distort this trend. Without outliers, a clearer pattern of salary growth with experience is observed.
- M.Tech./M.E. degree holders earn the highest average salaries, followed by B.Tech./B.E. and M.Sc. (Tech.), while MCA graduates have the lowest average salaries.
- Computer Science graduates have the highest median salaries, followed by Mechanical and Electrical Engineering graduates, with Civil Engineering showing the lowest median salaries and salary range.
- Gender does not significantly impact the median salary, though males exhibit higher variability and a few extreme salary outliers.
- Job location plays a role in salary distribution, with Bengaluru and Pune offering higher median salaries and wider salary ranges compared to other cities.
- Software Engineer is the most common designation, with Senior Software Engineers earning the highest median salaries and wider salary ranges compared to other roles.