**Introduction:**

The dataset provided is a sample from the Aspiring Mind Employment Outcome 2015 (AMEO) study conducted by Aspiring Minds, focusing primarily on students with engineering disciplines. It comprises employment outcomes of engineering graduates, including variables such as Salary, Job Titles, and Job Locations. Additionally, standardised scores from cognitive skills, technical skills, and personality skills are included. The dataset also incorporates demographic features such as Gender, Date of Birth, and College Information.

Below is a brief description of some key variables in the dataset:

* Salary: Annual CTC offered to the candidate (in INR).
* Designation: Job title or designation offered in the job.
* JobCity: Location of the job (city).
* Gender: Candidate’s gender.
* Specialisation: Specialisation pursued by the candidate.
* CollegeGPA: Aggregate GPA at graduation.
* GraduationYear: Year of graduation (Bachelor’s degree).
* English, Logical, Quant: Scores in different sections of the AMCAT test.
* conscientiousness, agreeableness, extraversion, neuroticism, openness\_to\_experience: Scores in different sections of AMCAT’s personality test.

The objective of this Exploratory Data Analysis (EDA) is to gain insights into the distribution of variables, identify any patterns or trends, explore relationships between variables, and address specific research questions such as testing claims about salary expectations and understanding the relationship between gender and specialisation preferences among engineering graduates.

**Univariate Analysis: Outlier Detection**

The univariate analysis focused on identifying outliers within the numerical columns of the dataset. Outliers are data points that significantly deviate from the rest of the data and may impact the overall analysis. The Interquartile Range (IQR) method was employed to identify outliers, where any data points lying below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR were considered outliers.

Upon analysing the dataset, it was observed that several numerical columns exhibited outliers. A summary of the outliers detected in each numerical column is provided below:

* 10 percent: 30 outliers
* 12 graduation: 45 outliers
* 12 percentage: 1 outlier
* CollegeTier: 297 outliers
* collegeGPA: 38 outliers
* GraduationYear: 2 outliers
* English: 15 outliers
* Logical: 18 outliers
* Quant: 25 outliers
* ComputerProgramming: 45 outliers
* ElectronicsAndSemicon: 21 outliers
* ComputerScience: 4 outliers
* MechanicalEngg: 13 outliers
* conscientiousness: 39 outliers
* agreeableness: 123 outliers
* extraversion: 40 outliers
* neuroticism: 15 outliers
* openess\_to\_experience: 95 outliers

These outliers warrant further investigation to determine their validity and potential impact on the analysis. It may be necessary to apply appropriate data preprocessing techniques, such as outlier removal or transformation, to ensure the robustness of subsequent analyses.

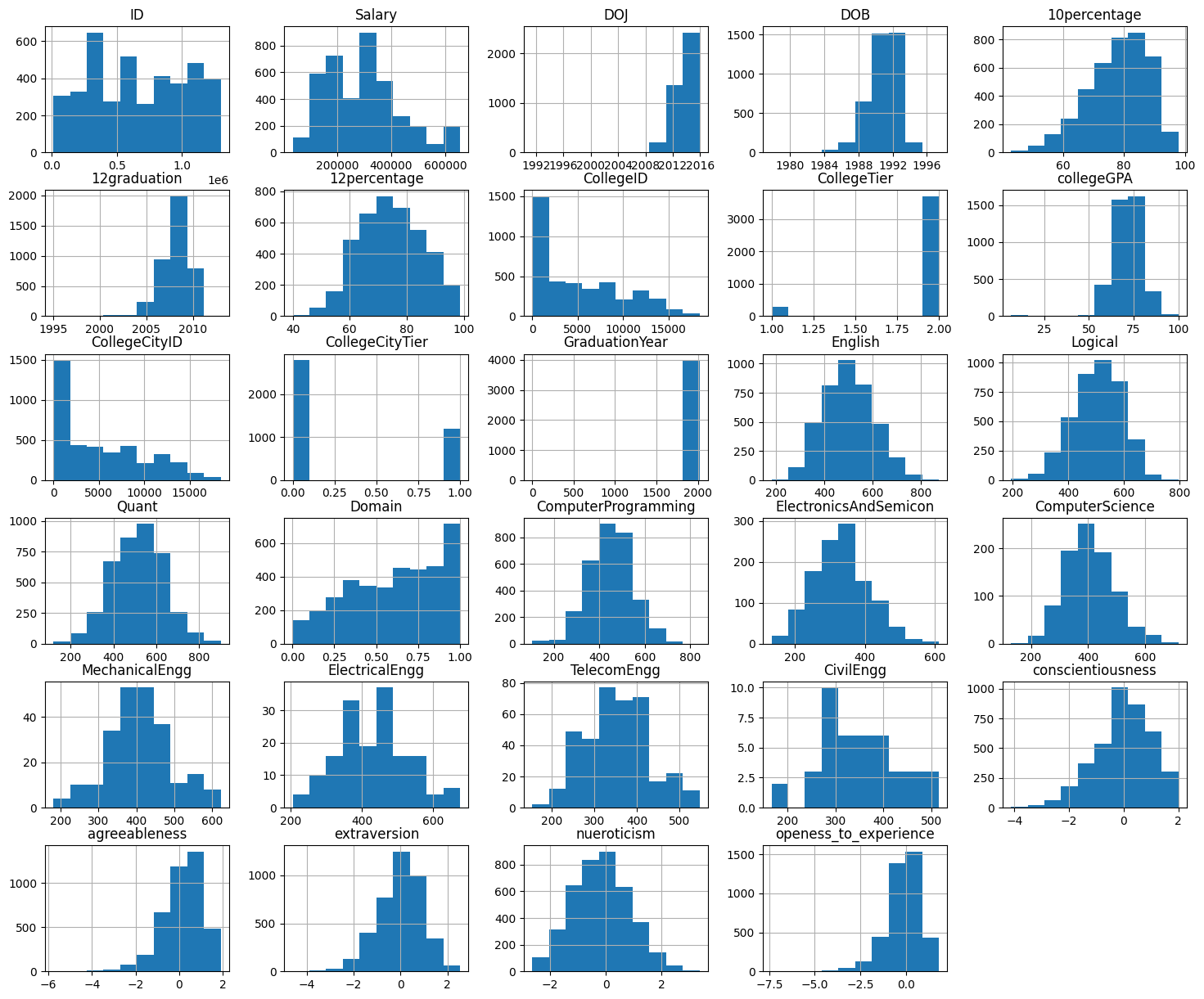
**Univariate Analysis: Summary Statistics and Distribution of Numerical Columns**

The summary statistics provide insights into the central tendency, dispersion, and shape of the distribution for each numerical column in the dataset. Here are some key observations from the summary statistics:

* ID: The ID column serves as a unique identifier for each candidate and does not provide meaningful statistical insights.
* Salary: The mean annual salary offered to candidates is approximately INR 295,493, with a standard deviation of INR 138,502. The salary ranges from INR 35,000 to INR 655,000.
* 10 percent: The average percentage obtained in grade 10 exams is approximately 77.93, with a standard deviation of 9.85. The distribution appears to be fairly symmetrical.
* 12 graduation: The mean graduation year from grade 12 is approximately 2008.08, with a standard deviation of 1.65. The distribution indicates that most candidates graduated around this time.
* 12 percentage: The average percentage obtained in grade 12 exams is approximately 74.47, with a standard deviation of 10.99. The distribution exhibits some variability.
* CollegeTier: This categorical column is represented numerically, with values of 1 and 2 indicating different tiers of colleges. The majority of candidates attended colleges classified as Tier 2.
* collegeGPA: The mean college GPA is approximately 71.49, with a standard deviation of 8.17. The distribution appears to be approximately normal.
* GraduationYear: The mean graduation year from bachelor's degree programs is approximately 2008.08, with a standard deviation of 1.65.
* English, Logical, Quant: These columns represent scores in different sections of the AMCAT test. The distributions of scores in these sections vary, with different means and standard deviations.

The histograms below visualise the distribution of numerical columns in the dataset, providing further insights into their spread and skewness.

The histograms depict the frequency distribution of each numerical column, highlighting the shape and spread of the data. These visualisations aid in understanding the distribution patterns and identifying potential outliers or anomalies.



**Univariate Analysis: Frequency Distribution of Categorical Columns**

* DOL (Date of Leaving):
  + The majority of candidates left their jobs in February 2024, followed by April 2015 and March 2015.
  + There are a total of 67 unique leaving dates in the dataset.
* Designation:
  + The most common designation among candidates is "software engineer," followed by "software developer" and "system engineer."
  + There are 419 unique designations in the dataset.
* JobCity:
  + Bangalore is the most frequent job city, followed by Noida, Hyderabad, and Pune.
  + There are a total of 338 unique job cities represented in the dataset.
* Gender:
  + The dataset comprises a higher number of male candidates (3041) compared to female candidates (957).
* 10 board (Grade 10 Board):
  + The most common grade 10 board is CBSE (Central Board of Secondary Education), followed by various state boards and ICSE (Indian Certificate of Secondary Education).
  + There are 275 unique grade 10 boards represented in the dataset.
* 12 board (Grade 12 Board):
  + Similar to the grade 10 board, CBSE is the most common grade 12 board, followed by state boards and UP board (Uttar Pradesh Board of Secondary Education).
  + There are 340 unique grade 12 boards represented in the dataset.
* Degree:
  + The majority of candidates hold a Bachelor of Technology/Engineering (B.Tech/B.E.) degree.
  + There are also candidates with Master of Computer Applications (MCA) and Master of Technology/Engineering (M.Tech./M.E.) degrees.
* Specialisation:
  + The most common specialisation is electronics and communication engineering, followed by computer science & engineering and information technology.
  + There are 51 unique specialisations represented in the dataset.
* CollegeState:
  + Uttar Pradesh is the most common state where colleges are located, followed by Karnataka, Tamil Nadu, and Telangana.
  + There are a total of 27 unique states/union territories represented in the dataset.

**Bivariate Analysis: AMCAT Scores using scatter plot**

Here are some specific observations I can make about the image:

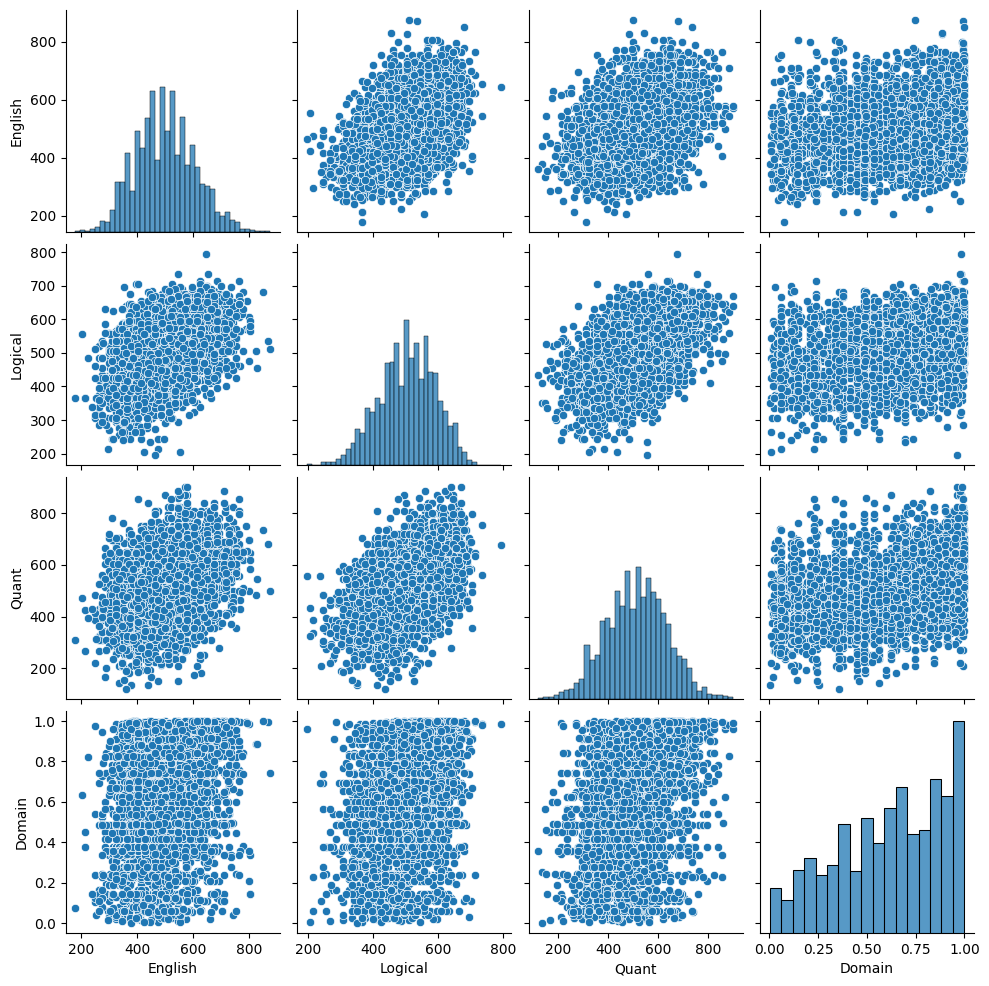
* The scatter plots show a positive correlation between all four pairs of variables. This means that there is a tendency for individuals who score higher on one variable to also score higher on the other variable.
* The strength of the correlation varies between the different pairs of variables. The correlation between Logical Reasoning and Quantitative Ability appears to be the strongest, while the correlation between English Language and Domain Knowledge appears to be the weakest.
* There is a considerable amount of variability in the data. This means that there are many individuals who score high on one variable but low on the other variable, and vice versa.

Overall Observations:

* Number of data points: It's difficult to say for sure, but there appear to be a large number of data points in each scatter plot, suggesting a substantial sample size.
* Positive correlations: All four subplots show a positive correlation, meaning individuals who scored higher on one variable tend to score higher on the other as well.
* Correlation strength: The strength of the correlation varies between the different pairs of variables. It appears strongest between Logical Reasoning and Quantitative Ability, and weakest between English Language and Domain Knowledge.
* Data variability: There's considerable variability in the data, with many individuals scoring high on one variable but low on the other, and vice versa.

Specific Observations for Each Subplot:

* Logical Reasoning vs. Quantitative Ability: This subplot shows the strongest correlation, with a tight cluster of points following a diagonal trend. There are some outliers, but overall, individuals with higher scores in Logical Reasoning tend to have higher scores in Quantitative Ability as well.\
* Logical Reasoning vs. English Language: This subplot shows a weaker correlation compared to the previous one. The data points are more spread out, and the trend is less pronounced. While there's a positive correlation, it's not as strong as between Logical Reasoning and Quantitative Ability.
* Logical Reasoning vs. Domain Knowledge: Similar to the previous subplot, this one shows a weaker positive correlation. The data points are scattered, and the trend is not very clear. There's no strong association between scores in Logical Reasoning and Domain Knowledge.
* Quantitative Ability vs. English Language: This subplot exhibits a moderate positive correlation. The data points form a somewhat diagonal cluster, but there's more spread than in the Logical Reasoning-Quantitative Ability subplot. Individuals with higher Quantitative Ability scores tend to score higher in English Language, but the association is not as strong as between Logical Reasoning and Quantitative Ability.
* Quantitative Ability vs. Domain Knowledge: This subplot shows a weaker positive correlation similar to the Logical Reasoning-Domain Knowledge one. The data points are spread out, and there's no clear trend. The relationship between Quantitative Ability and Domain Knowledge is not very strong.
* English Language vs. Domain Knowledge: This subplot has the weakest positive correlation among all four. The data points are widely scattered, and there's no discernible trend. There's no strong association between scores in English Language and Domain Knowledge.



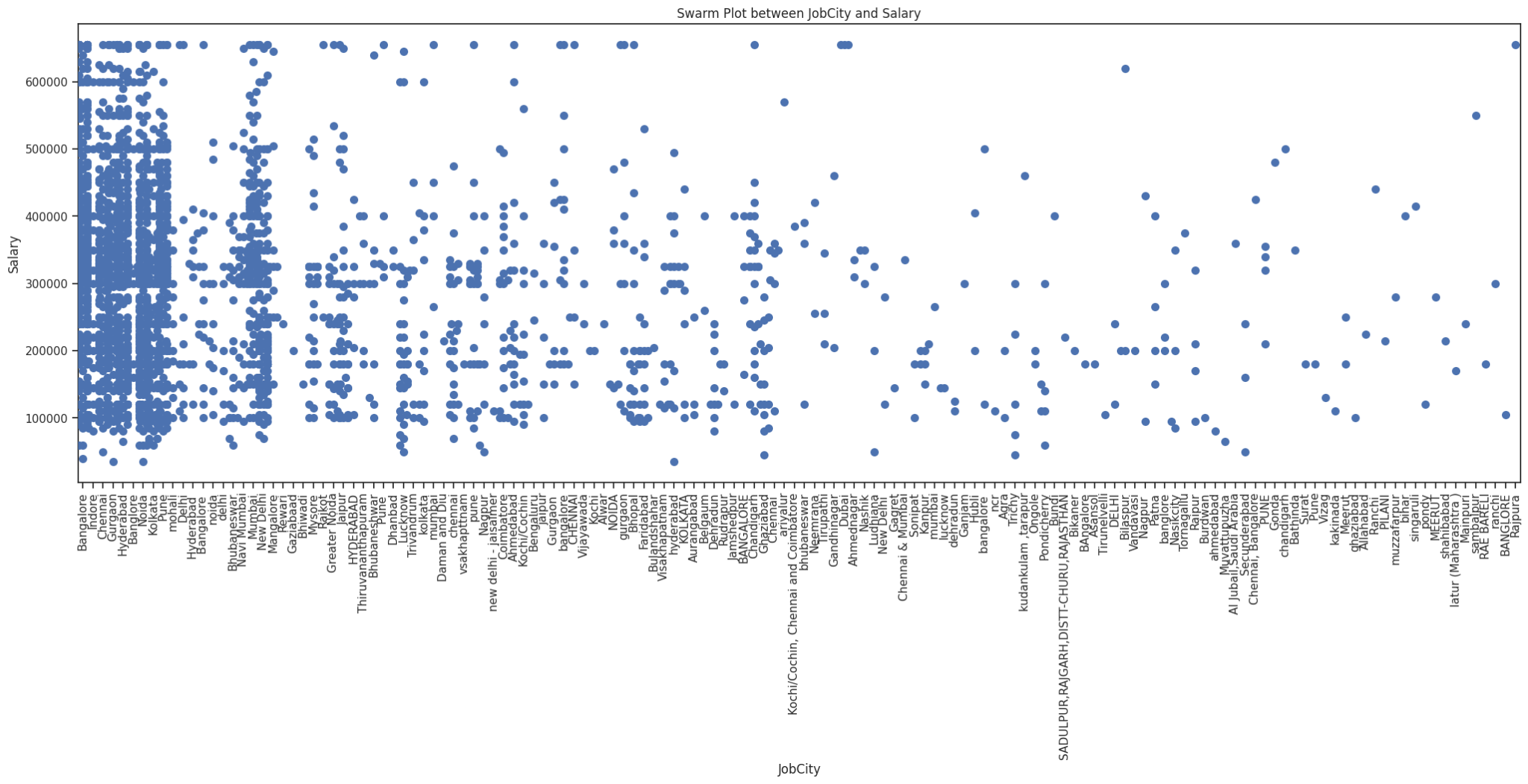
**Bivariate Analysis: Patterns between categorical and numerical columns using swarmplot**

Overall Observations:

* Positive Correlation: The data points show a positive correlation between JobCity and Salary. This means that, in general, individuals from cities with higher salary ranges tend to have higher salaries themselves.
* Density Distribution: The density of data points is highest in the upper right corner of the plot, indicating a concentration of individuals from high-salary cities with high personal salaries.
* Spread: The data points are not all clustered together, but rather spread out across the plot, suggesting a variety of salary ranges within each city and vice versa.
* Non-Linear Relationship: While there's a positive correlation, the data points don't form a clear straight line. This suggests that the relationship between JobCity and Salary is not perfectly linear.

Specific Observations in Different Regions:

* High Salary, High JobCity: The upper right corner has the highest density of data points, indicating a large number of individuals from high-salary cities who also have high salaries themselves. This could represent major financial hubs like Bengaluru, chennai, Hyderabad etc.
* Low Salary, Low JobCity: The lower left corner has a lower density of data points, but still shows some individuals from low-salary cities with low personal salaries. This could represent smaller towns or rural areas.
* Moderate: The centre and diagonal areas of the plot show moderate densities, representing individuals from cities with a mix of salary ranges and individuals with a mix of salary levels.
* Outliers: There are a few scattered data points outside the main density areas, which could represent outliers or individuals with unusual salary levels compared to their city's average.



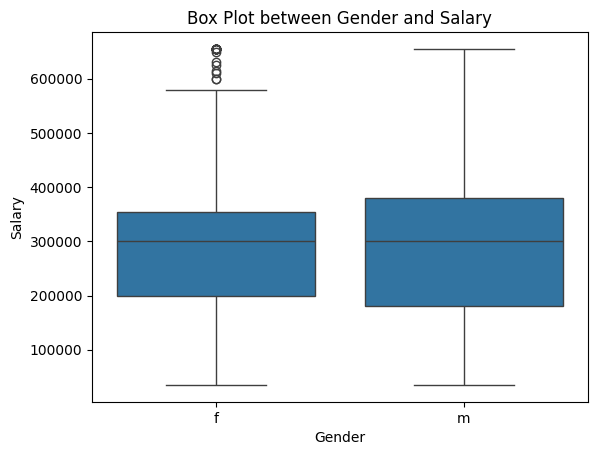
**Bivariate Analysis: Patterns between categorical and numerical columns using boxplot**

Overall Observations:

* Median Salary: The median salary for males appears to be higher than the median salary for females. The median line for the male boxplot is positioned higher than the median line for the female boxplot.
* Salary Distribution: The distribution of salaries for males appears to be wider than the distribution for females. The male boxplot is wider than the female boxplot, indicating a greater range of salaries among males.
* Whiskers: The whiskers for the male boxplot extend further than the whiskers for the female boxplot, suggesting that there are more outliers in the male group, meaning there are individuals with much higher or lower salaries than the rest of the group.
* Overlap: There is some overlap between the boxes and whiskers of the two groups. This indicates that there is some overlap in the salary ranges between the two genders, meaning that not all males earn more than all females.

Possible Interpretations:

* Gender Pay Gap: The observed difference in median salaries suggests a possible gender pay gap, where males earn more than females on average. However, it is important to note that this is just one observation and other factors may contribute to this gap.
* Salary Variability: The wider distribution of salaries for males might indicate a greater diversity of job positions, experience levels, or other factors influencing salary within the male group compared to the female group.
* Outliers: The presence of outliers in both groups could be due to various reasons, such as individuals holding high-paying leadership positions, entrepreneurs with successful businesses, or individuals with very low salaries due to part-time work or being early in their careers.



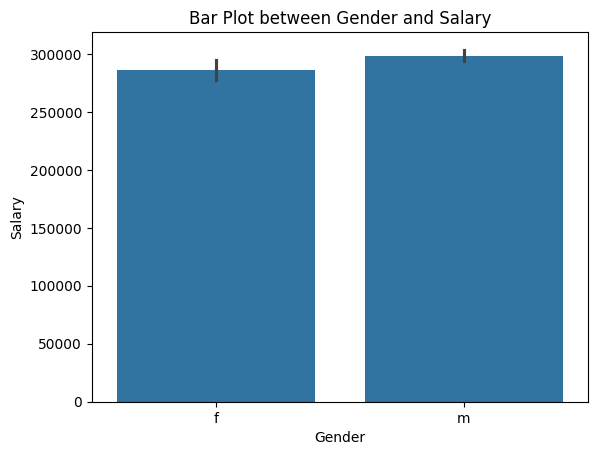
**Bivariate Analysis: Patterns between categorical and numerical columns using barplot**

Overall Observations:

* Salary Differences: The bar heights show a clear difference in average salary between genders. The "Male" bar is significantly higher than the "Female" bar, indicating that males have a higher average salary in this dataset.
* Salary Distribution: It's difficult to assess the spread of salaries within each gender from a bar plot. However, we can see that the bars have some width, suggesting there is some variation in salaries within each group.
* Data Representation: It's unclear whether the bars represent the mean, median, or another measure of central tendency for each gender. Knowing this would provide a clearer understanding of the typical salary for each group.

Possible Interpretations:

* Gender Pay Gap: The difference in bar heights suggests a possible gender pay gap, where males earn more than females on average in this dataset. However, it's important to remember that this is just one observation, and other factors may contribute to this gap



**Bivariate Analysis: Relationships between categorical and categorical columns using stacked bar plots.**

Observations:

* All workers: This line represents the average salary for all workers in the dataset, regardless of gender.
* Men: This line represents the average salary for male workers in the dataset.
* Women: This line represents the average salary for female workers in the dataset.
* Gender Pay Gap: This line represents the difference in average salary between men and women, calculated as the difference between the "Men" and "Women" lines.

Key Observations:

* Overall Trend: All four lines show a generally upward trend over time, indicating that average salaries have increased for all groups over the period covered by the graph.
* Gender Pay Gap: While the average salaries for both men and women have increased, the gender pay gap appears to have persisted over time. The "Gender Pay Gap" line remains above zero throughout the graph, indicating that men have consistently earned more than women on average.
* Gap Variation: The size of the gender pay gap appears to vary somewhat over time. There are periods where the gap seems to be widening, and other periods where it seems to be narrowing.
* Data Range: The graph does not show the specific range of salaries within each group at each point in time. This makes it difficult to assess the variability of salaries and the potential presence of outliers.
* Data Source: There is no information about the data source or the specific population represented in the graph. This makes it difficult to draw broader conclusions about the gender pay gap.

Possible Interpretations:

* Persistent Gender Pay Gap: The graph suggests that the gender pay gap is a persistent issue that has not been fully addressed over the time period covered by the data.
* Factors Influencing the Gap: The variation in the size of the gap over time suggests that there may be complex factors influencing the gender pay gap, such as changes in industry composition, job types, or policies.

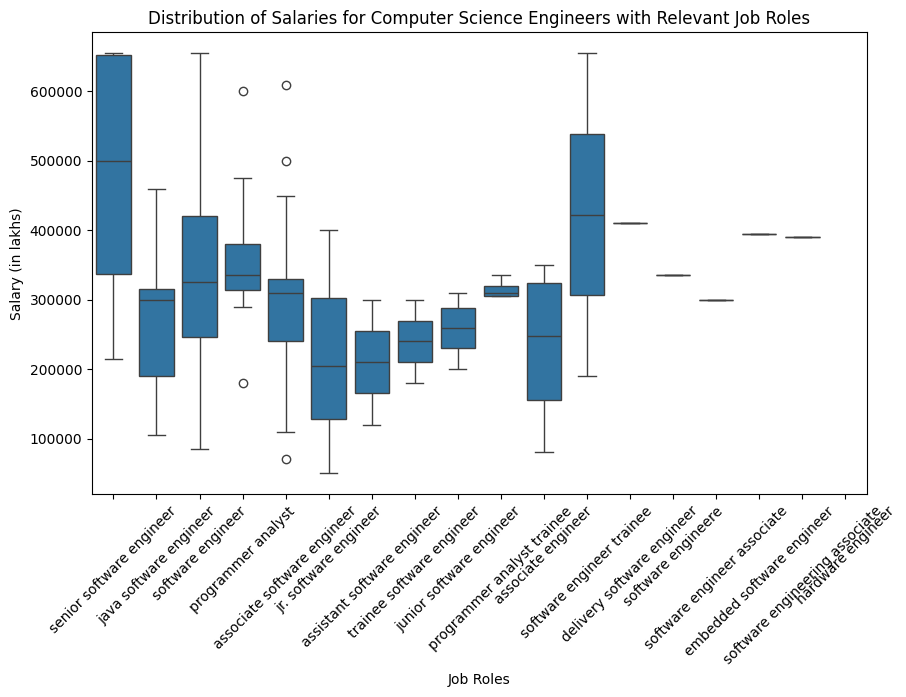
**Research Questions: CSE GRADUATES IN SOME JOB ROLES**

Overall Observations:

* Salary Distribution: The salaries vary significantly across different job roles. The box plot shows a wide range of salary distributions, with some roles having much higher median salaries and wider ranges than others.
* Job Role Comparisons: Some job roles have clearly higher median salaries than others. For example, "Software Engineering Manager" and "Delivery Software Engineer" appear to have the highest median salaries, while roles like "Assistant Software Engineer" and "Trainee Software Engineer" have the lowest.
* Salary Variability: The spread of salaries (represented by the box and whiskers) also varies across job roles. Some roles, like "Software Engineer" and "Java Software Engineer," have a wider range of salaries, indicating greater variability in compensation within those roles. Others, like "Associate Software Engineer" and "Programmer Analyst," have a narrower range, suggesting more consistent salaries.
* Outliers: There are outliers in several job roles, represented by data points beyond the whiskers. This suggests that there are individuals in some roles who earn significantly more or less than the typical salary for their position.

Specific Observations by Job Role:

* Software Engineering Manager: This role has the highest median salary and a wide range of salaries, indicating potential for high compensation but also significant variation.
* Delivery Software Engineer: Similar to Software Engineering Manager, this role has a high median salary and a wide range, suggesting potential for high earnings but also variability.
* Senior Software Engineer: This role has a lower median salary than the previous two but still falls in the upper range compared to other roles. The salary range is also wider, indicating some variability.
* Java Software Engineer: This role has a median salary similar to Senior Software Engineer, but with a wider range, suggesting more potential for variation in compensation.
* Software Engineer: This role has a moderate median salary and a wider range, indicating a mix of compensation levels.
* Other Roles: The remaining roles generally have lower median salaries and narrower ranges, suggesting more consistent but lower compensation compared to the higher-level roles.



**Research Question: Relationship between gender and specialisation**

Overall Observations:

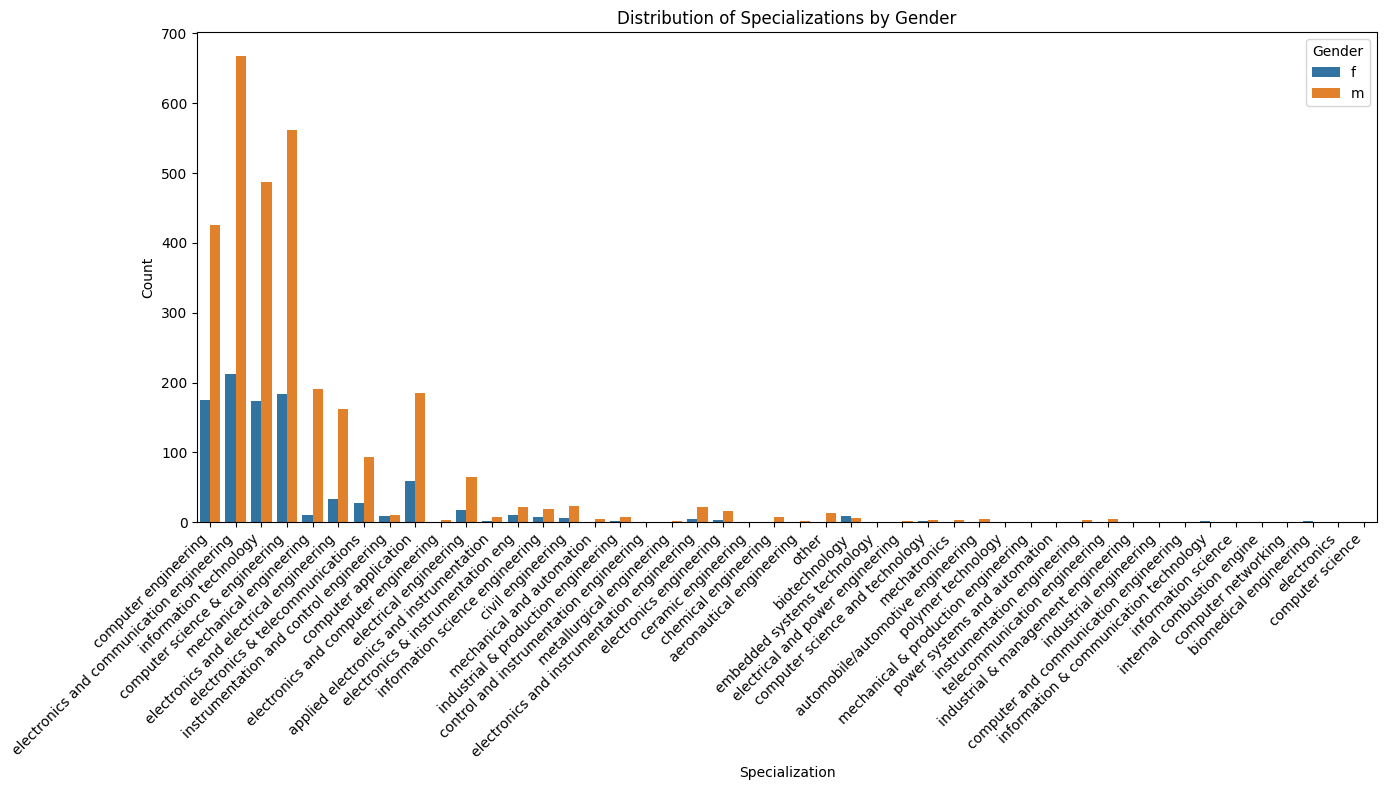
* Number of specialisations: There are 20 specialisations represented in the graph.
* Gender representation: The data seems to be relatively balanced, with both genders having a presence in most specialisations. However, there are a few specialisations where one gender is significantly more dominant.
* Distribution shape: The distribution of specialisations varies across genders and specialisations. Some specialisations have a more balanced distribution (e.g., "Data Science"), while others are heavily skewed towards one gender (e.g., "Mechanical Engineering" skewed towards males, "Nursing" skewed towards females).
* Overlapping distributions: There is considerable overlap in the distributions for most specialisations, indicating that both genders are present in most fields to some extent.

Specific Observations by Gender:

* Dominant specialisations for males: Some specialisations with a higher proportion of males include "Mechanical Engineering," "Electrical Engineering," "Computer Science," and "Information Technology."
* Dominant specialisations for females: Some specialisations with a higher proportion of females include "Nursing," "Education," "Human Resources," and "Social Work."
* Balanced specialisations: Some specialisations with a more balanced gender distribution include "Data Science," "Marketing," and "Business Administration."

Potential Interpretations:

* Gender stereotypes: The observed patterns might reflect existing gender stereotypes associated with certain specialisations. For example, mechanical and electrical engineering are often perceived as male-dominated fields, while nursing and education are often seen as female-dominated fields.
* Individual preferences and societal factors: The choices individuals make regarding their specialisations can be influenced by various factors, including personal interests, aptitudes, societal expectations, and career opportunities.
* Changes over time: It's important to consider how these patterns might have changed over time and whether there are any trends towards more balanced representation in different specialisations.



Conclusion:

The Exploratory Data Analysis (EDA) conducted on the Aspiring Mind Employment Outcome 2015 (AMEO) dataset has provided valuable insights into the employment outcomes of engineering graduates, with a focus on various demographic, educational, and job-related variables. Through univariate and bivariate analyses, we have examined the distribution, relationships, and patterns within the dataset, addressing specific research questions and exploring potential trends and correlations.

Key Findings:

Outlier Detection: The dataset exhibited outliers across multiple numerical columns, indicating the presence of data points that significantly deviate from the norm. Further investigation and preprocessing may be necessary to address these outliers and ensure the robustness of subsequent analyses.

Summary Statistics and Distribution: Summary statistics provided insights into the central tendency, dispersion, and shape of the distribution for each numerical column. Histograms visualised the frequency distribution of numerical data, aiding in understanding distribution patterns and identifying potential outliers.

Frequency Distribution of Categorical Columns: Frequency distributions of categorical columns revealed insights into the distribution of variables such as job titles, job locations, gender, educational backgrounds, and specialisations. These distributions provided a comprehensive overview of the dataset's categorical attributes.

Bivariate Analysis: Bivariate analyses, including scatter plots, swarm plots, box plots, bar plots, and stacked bar plots, examined relationships between numerical and categorical variables, as well as between different categorical variables. These analyses revealed correlations, trends, and patterns, allowing for deeper insights into the dataset.

Research Questions: Specific research questions were addressed, including the distribution of salaries across different job roles and the relationship between gender and specialisation preferences among engineering graduates. These analyses provided actionable insights into salary expectations and gender disparities in career choices.

The EDA conducted on the AMEO dataset has provided a comprehensive understanding of employment outcomes among engineering graduates. The findings highlight the complexity of factors influencing salary distribution, gender disparities, and specialisation preferences. While the analysis has uncovered valuable insights, further research and exploration may be warranted to delve deeper into specific areas of interest and validate findings. Overall, the EDA serves as a foundation for future investigations and decision-making processes related to career opportunities, workforce diversity, and educational policies in the engineering domain.