INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

"Exploratory Data Analysis on AMEO Dataset"

Name : Sakshi Yogendra Yadav
Date : 22nd February 2024

1 INTRODUCTION

Every year, 1.5 million engineers are produced in India. A relevant question is what determines the salary and the jobs these engineers are offered right after graduation. Aspiring Minds produced the **Aspiring Minds Employment Outcome** (AMEO) dataset, which provides a thorough compilation of data about engineering graduates' employment results. This dataset, which was created especially to examine the job environment for engineering specialties, is an invaluable tool for learning about the variables affecting graduates' prospects for employment and other career outcomes.

We hope to find patterns, trends, and linkages in the information we gather by using statistical analytic methods and data visualization software. This analysis will serve as a basis for well-informed decision-making in academics, industry, and policy-making, in addition to offering insightful information about the factors influencing job outcomes.

The report is set up to methodically examine several facets of the dataset, beginning with preparing the data to guarantee its integrity and quality. Subsequently, we delve into exploratory analyses of demographic features, academic metrics, job-related variables, and personality traits. Through visualizations and statistical summaries, we aim to provide a comprehensive understanding of the factors influencing engineering graduates' career trajectories.

2 OBJECTIVE

To obtain a thorough understanding of the factors impacting pay offers for engineering graduates, **the Aspiring Minds Employment Outcome (AMEO)** dataset with **Salary** as the goal variable is being analyzed. This analysis aims to gain insights and understanding from the provided dataset, particularly focusing on the relationship between various features and the target variable, which is **Salary**.

Specifically, the goals of this analysis include:

- **Describing** the dataset and its features comprehensively.
- **Identifying** any **patterns** or **trends** present in the data.
- Exploring the **relationships** between independent and target variables (Salary).
- Identifying any **outliers** or anomalies in the data.

3 OVERVIEW OF DATA

Origin: Aspiring Minds performed the Aspiring Minds Employment Outcome 2015 (AMEO) study, which is where the dataset originated.

Range: The dataset, which mostly focuses on students with engineering credentials, records many aspects of their job search, such as compensation offers, job titles, locations, and standardized test results for cognitive, technical, and personality qualities.

Size: The dataset provides a broad and diversified set of data for research, with about 4000 data points and about 40 independent variables.

3.1 DEPENDENT VARIABLES:

Salary: Reflecting the annual Cost to Company (CTC) offered to candidates.

Job Titles: Detailing the 'Designations' offered in job placements.

Job Locations: Identifying the geographical locations (cities & states) of job placements.

3.2 INDEPENDENT VARIABLES:

Educational Background: Including grades, board curriculums, college GPA, graduation year, college tier, degree pursued, and specialization.

Demographic Information: Gender and date of birth.

Standardized Test Scores: Covering areas such as English, logical reasoning, quantitative aptitude, computer programming, and various engineering disciplines.

Personality Traits: Assessing traits such as conscientiousness, agreeableness, extraversion, neuroticism, and openness to experience.

4 EXPLORATORY DATA ANALYSIS

The process of evaluating data sets to highlight their key features is known as exploratory data analysis (EDA), and it frequently uses visual aids. It is the process of looking into data to find trends, abnormalities, connections, and insights; summary statistics and visualization are the main tools used in this process. EDA seeks to reveal a dataset's essential characteristics and comprehend its underlying structure.

The data analysis process requires the use of EDA since it enables analysts to:

- Recognize the features and organization of the data.
- Determine outliers, trends, and patterns.
- Create theories and put presumptions to the test.
- Provide new perspectives to direct future research and decision-making.

4.1 DATA CLEANING AND PREPROCESSING

4.1.1 Data Conversion

Several transformations were used to improve the Aspiring Minds Employment Outcome (AMEO) dataset's usability and analytical potential during the data conversion and pretreatment stages of the analysis process. First, datetime format was applied to the Date of Birth (DOB) and Date of Joining (DOJ) variables. Time-based analytics and trend identification are made easier by this conversion, which allows the dataset to accurately capture temporal information. The dataset is made more structured and suitable for time series analysis by converting these variables into a standardized datetime format. This allows researchers to investigate connections between employment outcomes and temporal aspects like tenure and age at joining. Furthermore, the Date of Leaving (DOL) variable was adjusted to include a present value, indicating that individuals who have not left their current positions have a DOL value replaced by the current date (today_date()).

4.1.2 Case Transformation

Executed a transformation on all categorical feature variables by converting them to lowercase. This comprehensive adjustment aimed to standardize the representation of categorical data throughout the dataset, thereby enhancing its consistency and facilitating more robust analyses.

4.1.3 Collapsing Categories

Categorical variables that had values 0 or -1, which mean they were incomplete or missing data, were combined into one category called "others". These numbers frequently indicate situations in which respondents omitted information or for which data was unavailable, which, if ignored, could distort studies.

4.2 FEATURE ENGINEERING

4.2.1 Age Calculation:

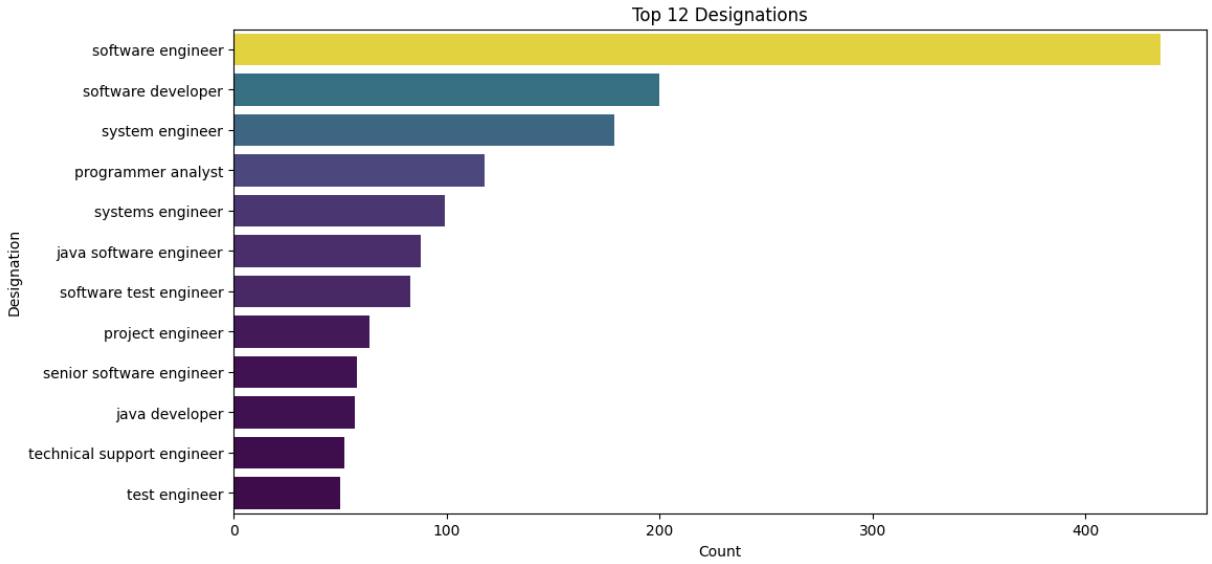
By deducting the year of birth (DOB) from 2024, an extra column denoting age has been added to the dataset, indicating the individual's age as of that year.

4.3 UNIVARIATE ANALYSIS

4.3.1 CATEGORICAL FEATURES:

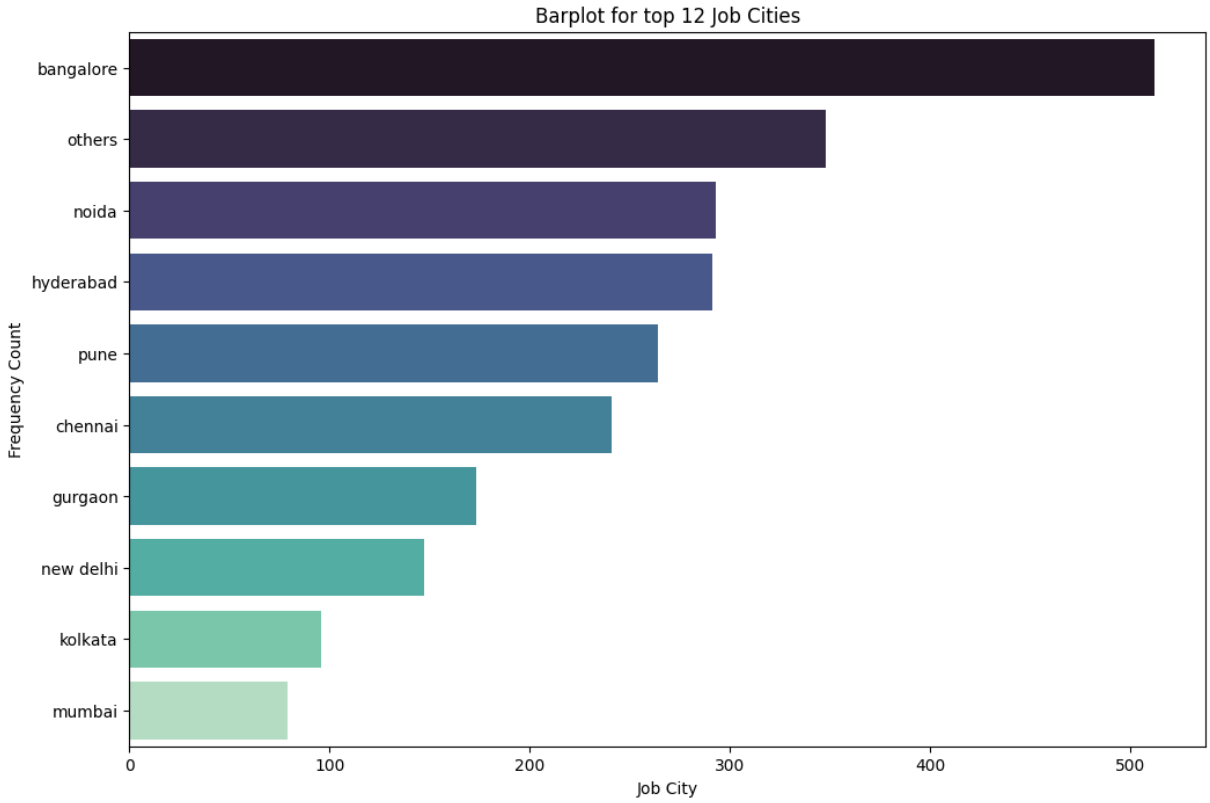
4.3.1.1 Designation

The AMEO dataset's horizontal bar plot of the "Designation" variable, which has 373 distinct values, offered useful information regarding the distribution of career designations among engineering graduates. The title that is most frequently seen is "Software Engineer," closely followed by "System Engineer" and "Software Developer."



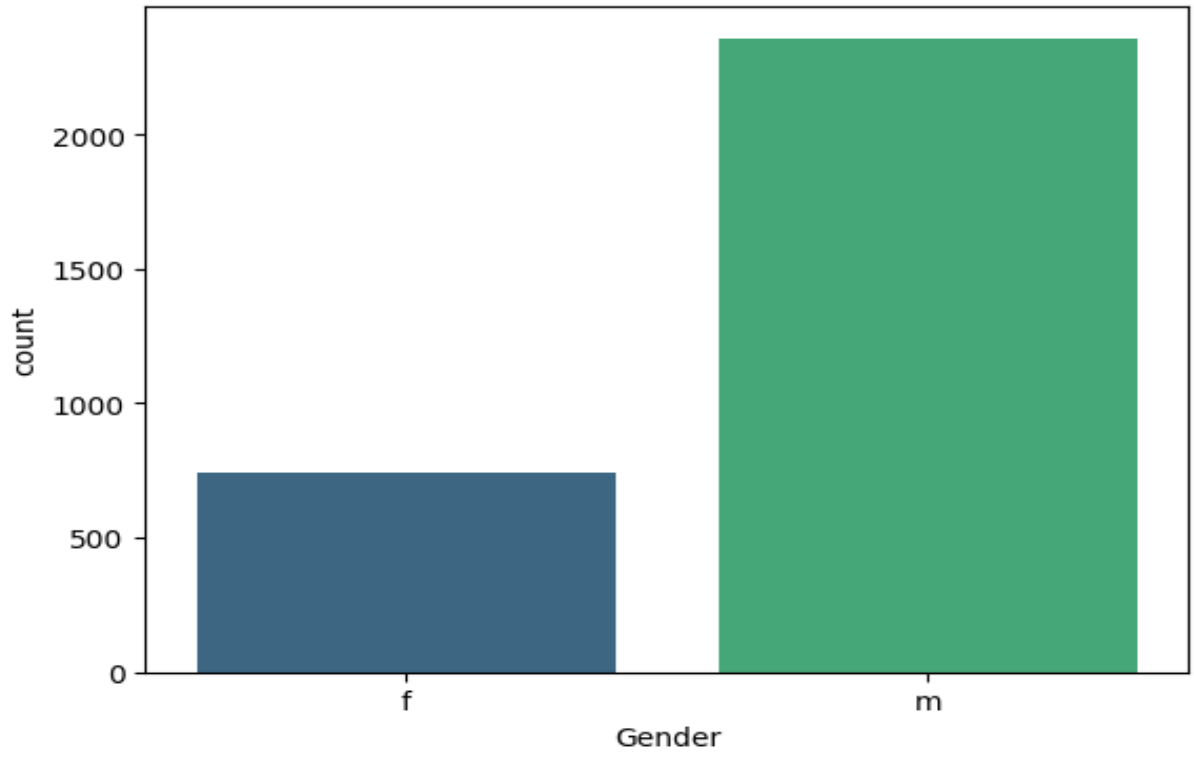
4.3.1.2 JobCity

We conclude that there is a significant variation in the job locations of engineering graduates in the AMEO dataset, with 216 distinct values in the "JobCity" variable. But after some study, we determined which cities were best for finding jobs; Bangalore came out on top, followed by Hyderabad and Noida. Note: Since the other category received no responses, we disregarded it.



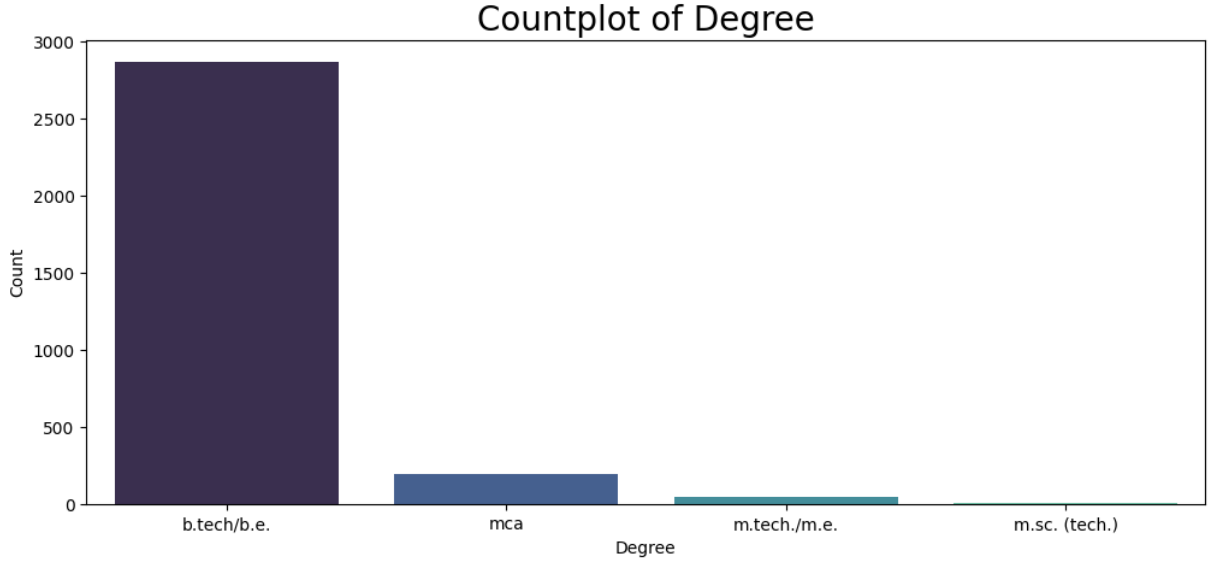
4.3.1.3 **Gender**

The gender distribution of candidates in the AMEO dataset reveals a considerable class imbalance, with a significantly higher number of male applicants than female candidates. In particular, there are 744 female candidates and 2358 male candidates in the sample. Determining that the male category is actually bigger than the female one.



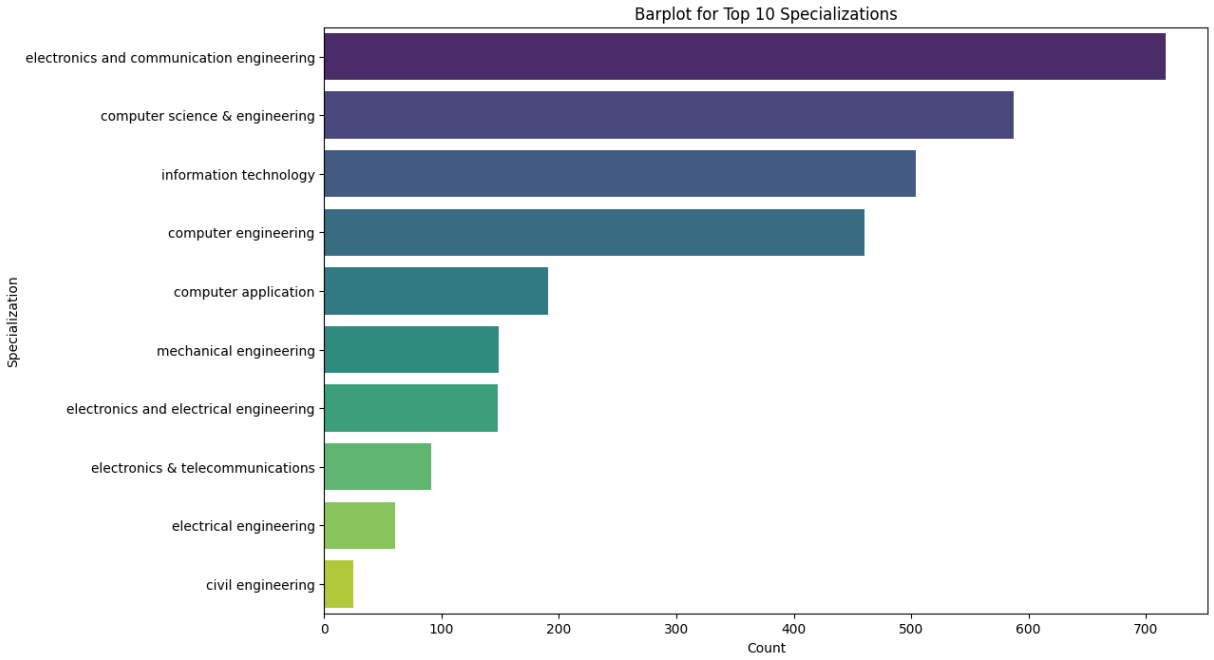
4.3.1.4 **Degree**

The distribution of educational backgrounds in the AMEO dataset indicates that B.Tech/B.E. degrees are overwhelmingly common, accounting for 2868 of the total. Students from MCA, M.Tech/M.E, and M.Sc. (Tech.) degrees, on the other hand, are significantly underrepresented—1961, 41, and only 2 people, respectively.



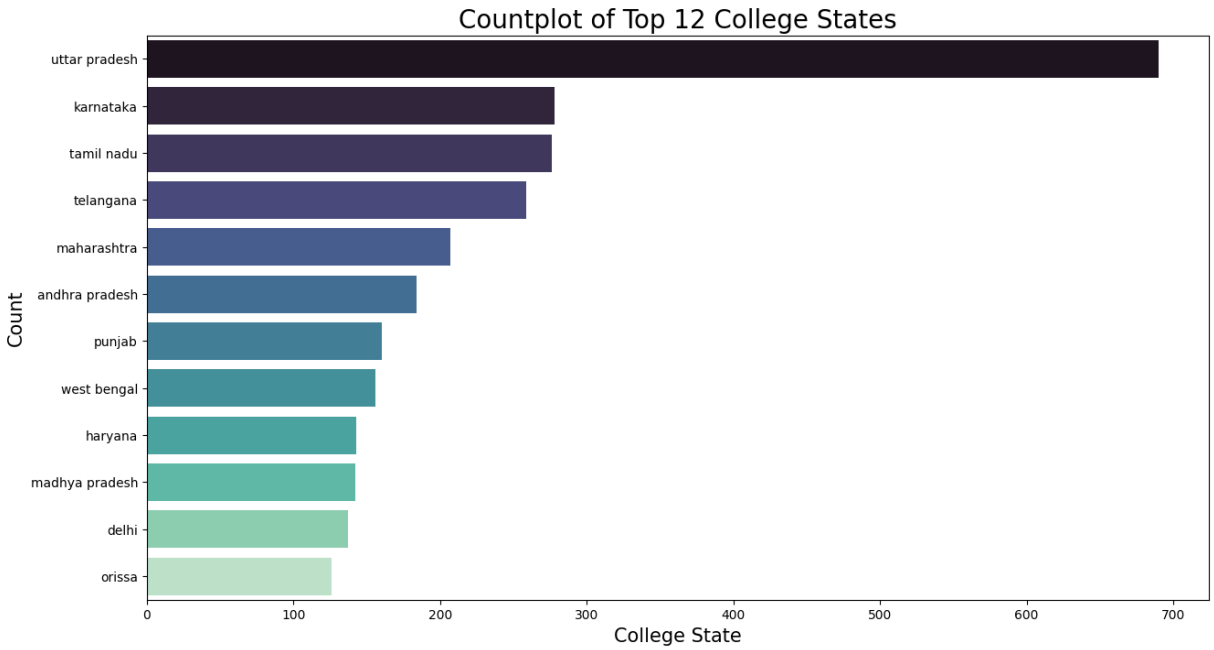
4.3.1.5 **Specialization**

"Electronics and Communication Engineering" is the most common engineering specialization in the AMEO dataset, with 717 persons belonging to this category. This specialization covers topics including wireless communication, signal processing, telecommunications, analog and digital electronics, and the design, development, and application of electronic devices and communication systems. "Computer Science & Engineering," with 587 people, is just behind. The study of computer systems, software development, algorithms, and programming languages are at the center of this specialization. With 504 people, "Information Technology" is yet another noteworthy specialization in the dataset. Information technology pertains to the administration, setup, and upkeep of IT infrastructure and systems in businesses.



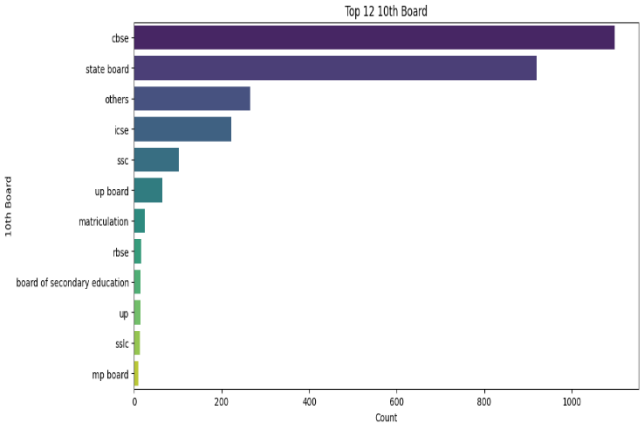
4.3.1.6 **CollegeState**

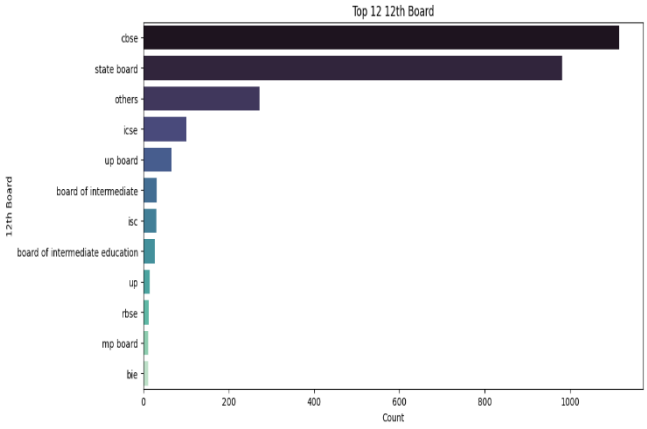
The AMEO dataset contains 26 distinct values for the "CollegeState" variable, which suggests a heterogeneous distribution of colleges among various states. With 690 people having finished their schooling there, Uttar Pradesh is the most common state among them. Karnataka, Tamil Nadu, and Telangana are noteworthy states with 278, 276, and 259 people, respectively, after Uttar Pradesh. This state-by-state breakdown of colleges offers information about the geographic distribution of academic establishments that support the engineering labor force.



4.3.1.7 **10thBoard & 12thBoard**

In the AMEO dataset, there are 221 unique values for the "10board" variable and 266 unique values for the "12board" variable. Of these, state board and CBSE are the most common boards for both 10th and 12th grades; the other boards make up a lower percentage of the dataset.

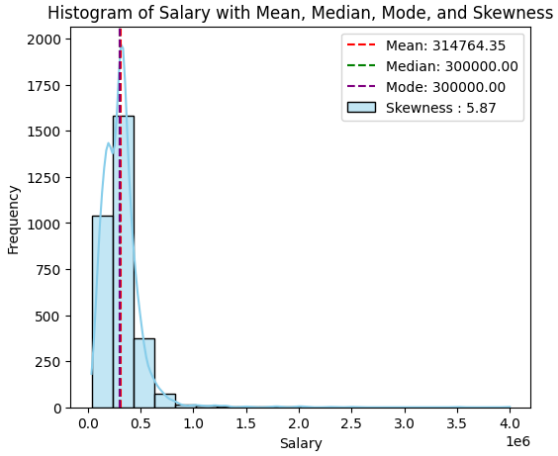




4.3.2 NUMERICAL FEATURES:

4.3.2.1 Salary

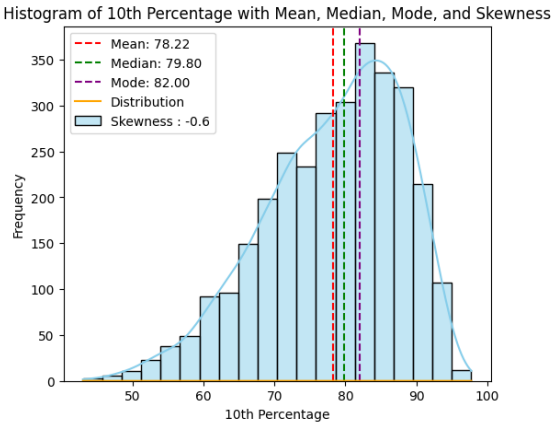
The AMEO dataset's salary summary statistics offer important information about the range of compensation packages offered to engineering graduates. The dataset includes 3,102 records that show variations in compensation offers across the sample. The mean income is around 314,764.30 INR, and the standard deviation is roughly 199,058.90 INR. The Salary data illustrates the wide range of compensation packages available to engineering graduates, with minimum income of 35,000 INR and maximum salary of 4,000,000 INR. Furthermore, the median 50% of wage offers fall within the interquartile range (IQR), which extends from the 25th percentile value of 200,000 INR to the 75th percentile value of 380,000 INR. The 50th percentile, or median pay, is 300,000 INR. This means that half of the salary offers in the dataset are below this figure and the other half are beyond it. This statistic offers a reliable way to quantify central tendency, particularly for datasets with skewed distributions.



With a skewness value of roughly 6, the data shows strong positive skewness, which deviates from a normal distribution. The mean, median, and mode—the three central tendency measurements—are about equal.

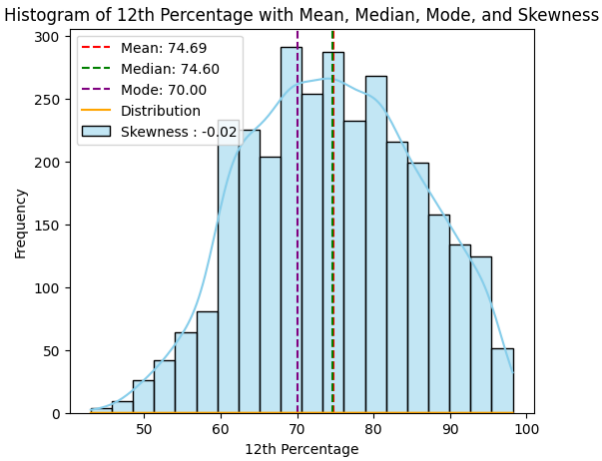
4.3.3 10percentage

The average percentage obtained in grade 10 exams stands at approximately 78.22%, with a standard deviation of around 9.73%. The minimum score recorded is 43%, while the maximum score is 97.76%. The 25th percentile falls at 72%, and the 75th percentile at 86%.



4.3.4 12percentage

Exam results for grade 12 typically yield an average percentage of 74.69%, with a standard deviation of roughly 11.0 percent. 43% is the lowest recorded score, and 98.2% is the highest. At 66.4% and 83%, respectively, are the 25th and 75th percentiles.



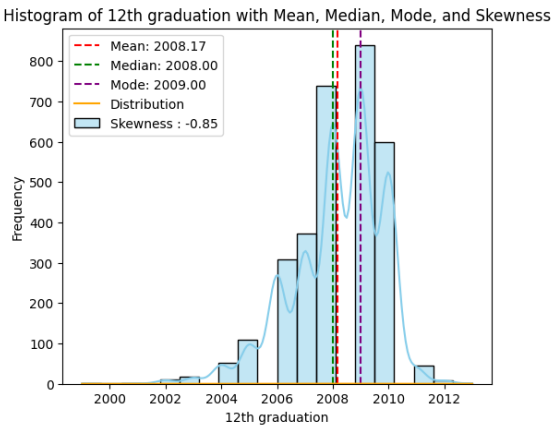
4.3.5 CollegeTier

The CollegeTier variable, representing the tier of the college attended by graduates, has an average value of approximately 1.93, indicating that a majority of graduates attended colleges classified as Tier 2. The minimum tier recorded is 1, representing Tier 1 colleges, while the maximum tier recorded is 2, indicating Tier 2 colleges.

4.3.6 12graduation

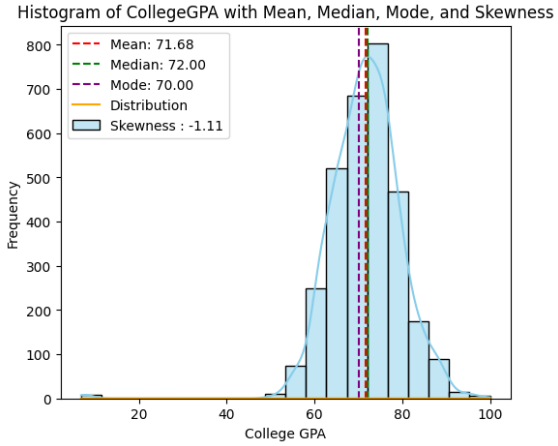
The average year of graduation from grade 12 is approximately 2008.17, with a standard deviation of approximately 1.61.

The earliest graduation year recorded is 1999, while the most recent graduation year is 2013. The 25th percentile graduation year is 2007, and the 75th percentile is 2009. It shows a negative skew.



4.3.7 CollegeGPA

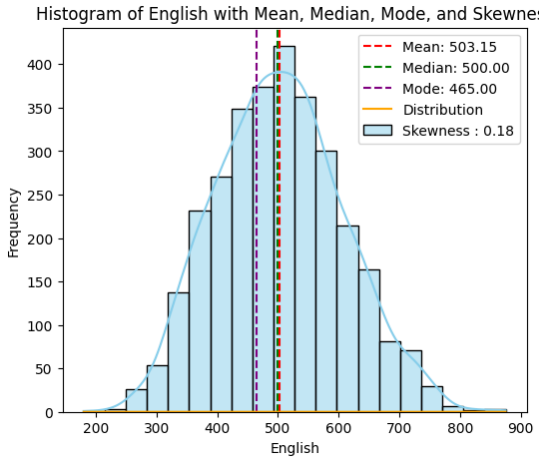
The average GPA stands at approximately 71.68, with a standard deviation of around 8.04. The minimum GPA recorded is 6.80, while the maximum GPA is 99.93. The 25th percentile falls at 66.69, and the 75th percentile at 76.50.

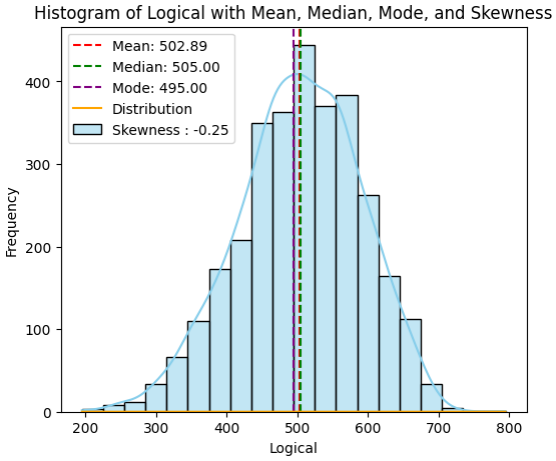


4.3.8 English & Logical Scores

The average scores for the English and Logical sections of the AMCAT test stand at approximately 503.15 and 502.89, respectively. The standard deviations are approximately 104.63 and 86.64, respectively. The minimum scores recorded for both sections

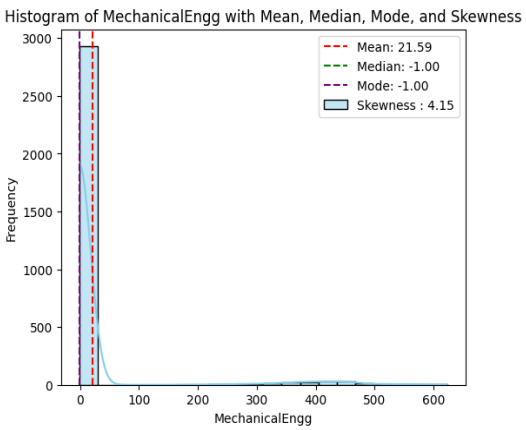
are 180 and 195, respectively, while the maximum scores are 875 and 795, respectively.

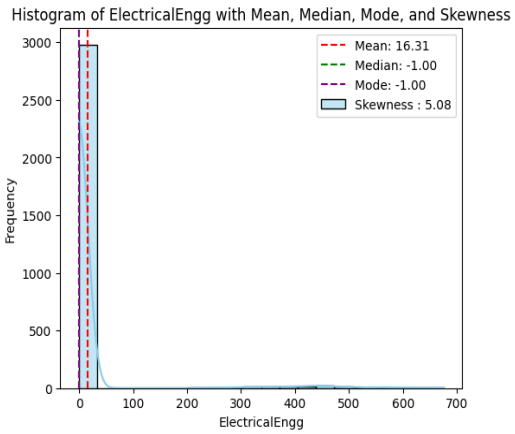


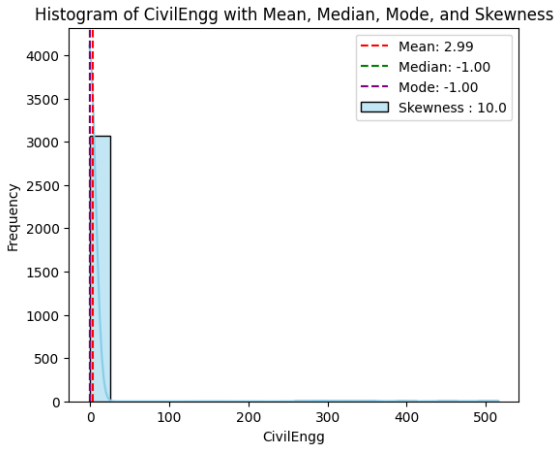


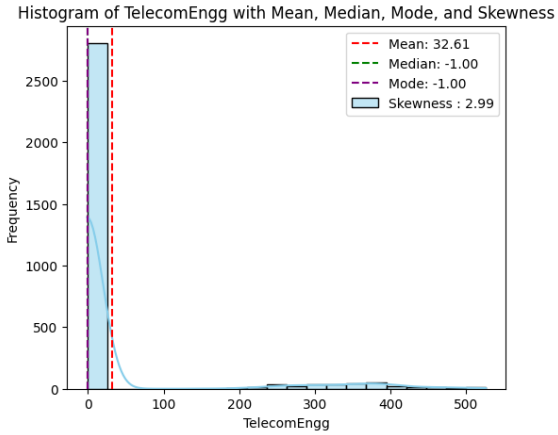
4.3.9 MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg

These variables represent scores in different engineering disciplines within the AMCAT test. The mean scores are approximately 21.59, 16.31, 32.61, and 2.99, respectively. However, the data suggests that the majority of individuals did not take these specific sections, as indicated by the prevalence of -1 values.



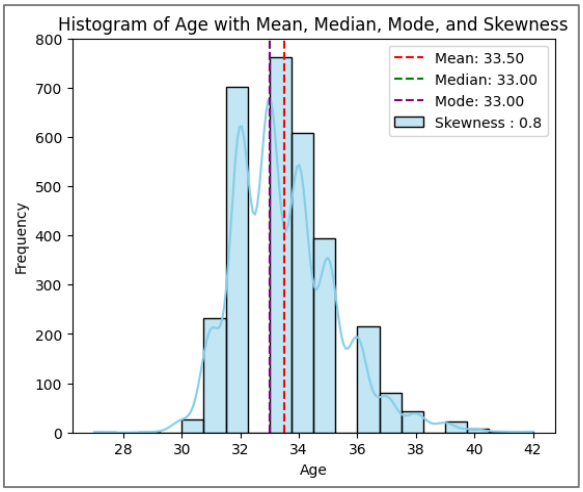






4.3.10 Age

The "Age" column in the dataset offers insights into the ages of individuals, likely representing their ages at the time of data collection or assessment. With a total of 3102 entries, the average age of individuals stands at approximately 33.50 years, showcasing the central tendency of the age distribution. The

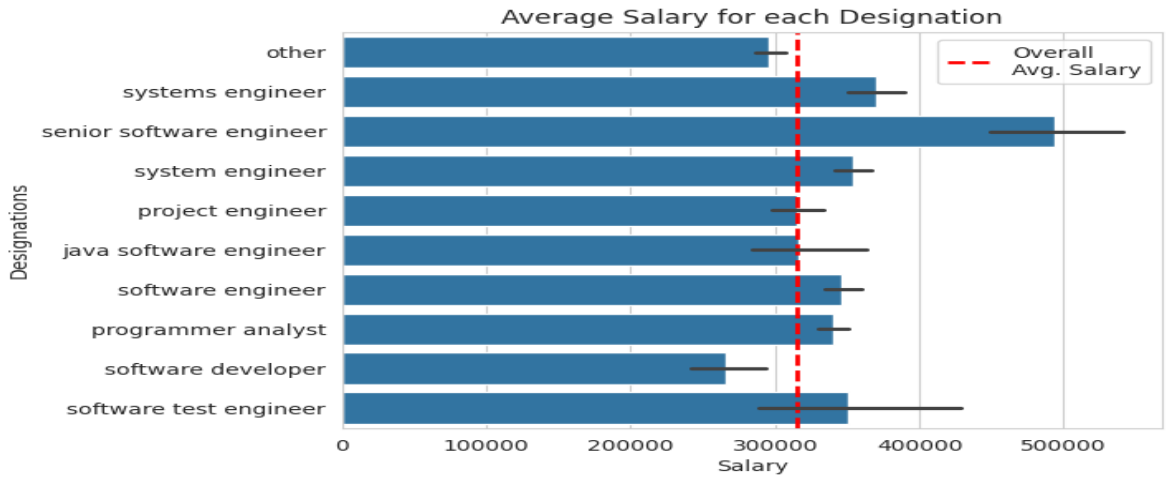


standard deviation of approximately 1.72 years indicates the degree of spread or dispersion of ages around the mean, with ages ranging from a minimum of 27 years to a maximum of 42 years. The quartile values further delineate the age distribution: 25% of individuals are aged 32 years or younger (25th percentile), while 75% are aged 34 years or younger (75th percentile). The median age, representing the middle value of the dataset, is 33 years, indicating that half of the individuals are aged 33 years or younger.

A skewness value of 5.87 indicates that the distribution of the data in the "Age" column is highly positively skewed. Positive skewness means that the tail of the distribution is longer on the right side

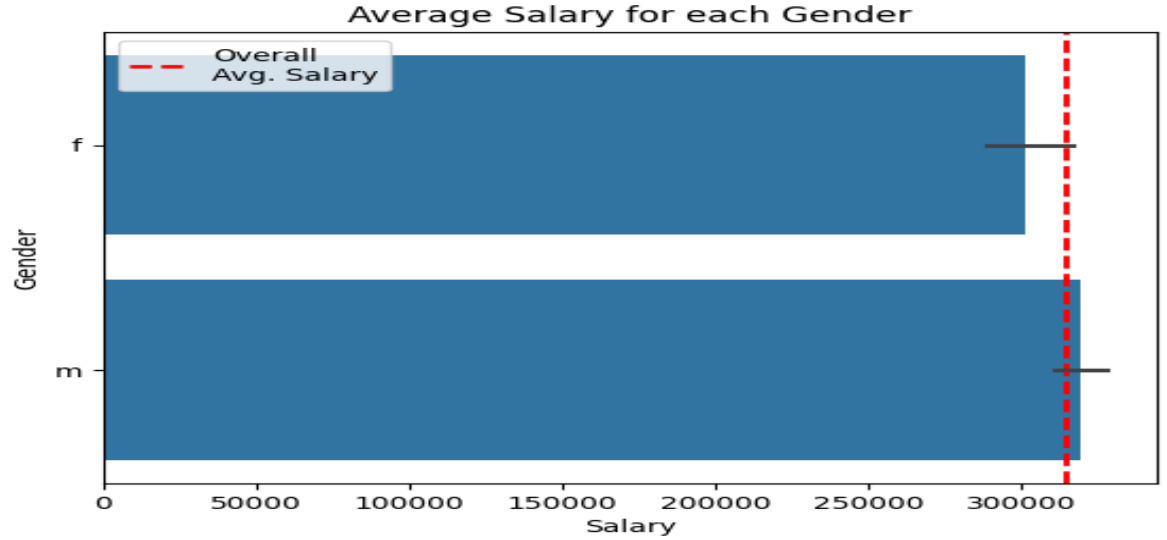
4.4 BIVARIATE *ANALYSIS*

4.4.1 Salary & Designation



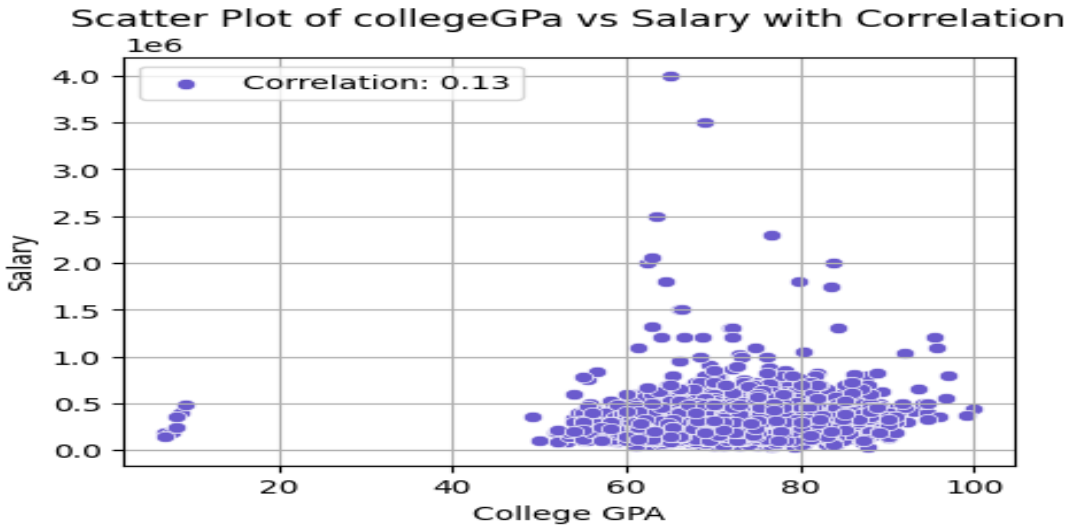
Senior Software Engineers have the highest salary, but also the highest standard deviation. Software Developers and Technical Support Engineers have salaries below the average.

4.4.2 Gender & Salary



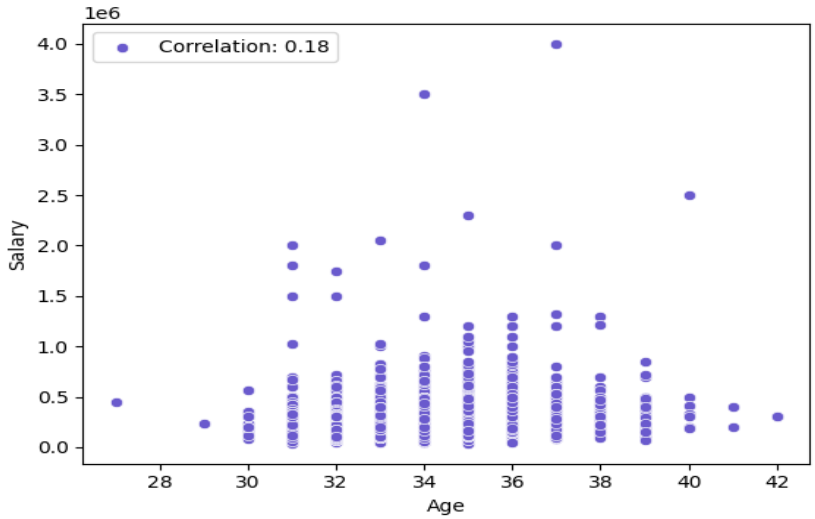
Both male and female salaries are approximately equal on average, suggesting no gender bias overall, though females tend to receive salaries below the overall average.

4.4.3 College GPA & Salary



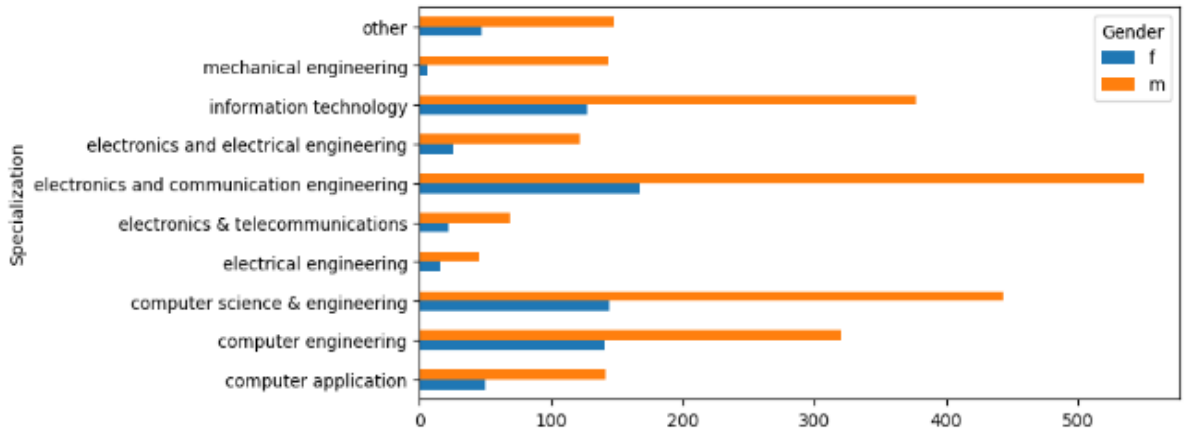
There is no significant correlation between salary and score in College GPA.

4.4.4 Age & Salary



There's no apparent relationship between age and salary after removing outliers.

4.4.5 Gender & Specialization



Male participation is approximately double that of female across all specializations, with fewer females opting for mechanical and electronics.

5 RESEARCH QUESTIONS & OUTCOMES

5.1.1 "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."

Outcome:

Designation	t-value	p -value	Result
Programmer Analyst	9.75	2.65118e-10	Reject Null Hypothesis
Software Engineer	7.58	6.09466e-12	Reject Null Hypothesis
Hardware Engineer	NaN	NaN	Not Enough Evidence
Associate Engineer	NaN	NaN	Not Enough Evidence

The analysis begins by grouping the dataset by job designation, calculating the mean and standard deviation of salaries for each job role. This provides insights into salary distribution across different designations. Notably, Software Engineers have the highest mean salary and standard deviation, indicating both higher earnings and variability in pay compared to Programmer Analysts and Associate Engineers.

Following this, a one-sample t-test is conducted for each job designation to compare their average salary against an expected range. For Programmer Analysts and Software Engineers, the test results show sufficient evidence to reject the null hypothesis, suggesting that their salaries significantly differ from the expected range.
However, for Hardware Engineers and Associate Engineers, there is not enough evidence to reject the null hypothesis, indicating that their salaries may not significantly deviate from the expected range.

Overall, these analyses provide valuable insights into salary distributions among different job roles and help in understanding the significance of salary differences within the dataset.

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

Outcome:

Test	Value
chi2_critical	16.918977604620448
chi2_statistic	47.625462024510526
chi2_p_value	3.000409236234154e-07

The analysis conducted using a Chi-Square test examined the relationship between gender and specialization preferences. The test revealed a statistically significant relationship between the two variables, indicating that specialization preferences are dependent on gender.

The calculated chi2 statistic exceeded the critical value, and the p-value was significantly less than the chosen significance level, leading to the rejection of the null hypothesis. Therefore, there is sufficient evidence to conclude that gender and specialization are related, suggesting that certain fields may be more preferred or accessible to individuals of particular genders. This finding underscores the importance of considering gender diversity and inclusivity in various fields and highlights potential barriers or biases that may exist in certain specializations.

6 CONCLUSION

To sum up, the Aspiring Mind Employment Outcome 2015 (AMEO) dataset's exploratory data analysis (EDA) has given researchers a thorough grasp of the academic achievement and employment outcomes of engineering graduates. Several significant discoveries have been made after a thorough analysis of a number of variables, including personality traits, job-related variables, academic metrics, and demographic data.

The investigation showed that there is a wide range of work options for engineering graduates, with different regional distributions, job titles, and salary ranges. Even though the dataset showed some similarities, like the popularity of software engineering professions, it also emphasized how different every career path is. Metrics of academic performance such college GPA, standardized test scores, and percentages of students in 10th and 12th grades were important in determining job placement and pay results. Higher incomes were typically demanded by graduates from elite universities and those with specialized degrees, highlighting the significance of educational qualifications.

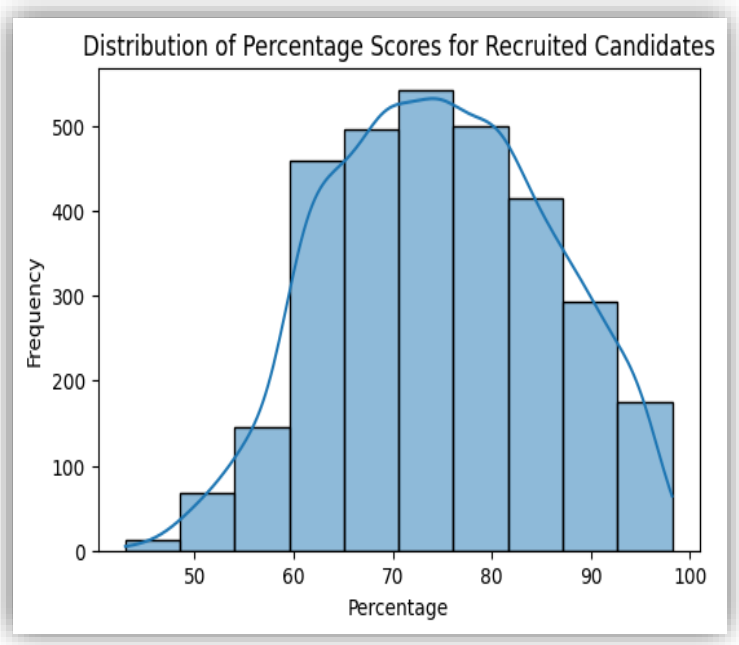
Additionally, personality qualities like agreeableness and conscientiousness were shown to be more common among engineering graduates, despite personality traits showing fewer significant relationships with wage outcomes. This emphasizes the need of soft skills in the workplace.

In general, this research offers a comprehensive overview of the determinants impacting the career trajectories of engineering graduates, which is essential information for academics, industry, and policy-making. It provides the groundwork for strategic planning and well-informed decision-making with the goal of improving future graduates' professional prospects and possibilities.

7 ADDITIONAL RESEARCH QUESTION

7.1 "In a comparative study of recruitment practices among leading companies, does AMEO's hiring policy of recruiting candidates with a minimum percentage of 70% and maintaining an average percentage of 80% hold true?"

Outcome:



The analysis reveals that in order to improve the caliber of candidates it hires, MEO may need to reevaluate its minimum percentage requirement or put in place extra screening measures. It's possible that the existing standards won't successfully identify applicants who meet the required requirements or have the potential to succeed in the company. MEO may be able to increase the quality of its hiring process and draw in applicants of a higher grade by changing the recruiting criteria or adding additional screening procedures.