EXPLORATORY DATA ANALYSIS OF AMCAT PROJECT REPORT

By

Sanhita Thangallapalli

INTRODUCTION

The dataset was released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO). The study is primarily limited only to students with engineering disciplines. The dataset contains the employment outcomes of engineering graduates as dependent variables (Salary, Job Titles, and Job Locations) along with the standardized scores from three different areas – cognitive skills, technical skills and personality skills. The dataset also contains demographic features. The dataset contains around 40 independent variables and 4000 data points. The independent variables are both continuous and categorical in nature. The dataset contains a unique identifier for each candidate. Below mentioned table contains the details for the original dataset.

**OBJECTIVE:**

Uncover valuable insights and patterns from the AMEO 2015 dataset, with a specific focus on understanding the relationship between salary and various independent variables. By analyzing this rich dataset, we aim to gain actionable knowledge that can inform decision-making in the context of entry-level labour markets for engineering graduates. Through exploratory data analysis (EDA), we aim to:

* Describe the dataset's structure and contents.
* Identify outliers and distributions within numerical columns.
* Analyze the frequency distribution of categorical variables.
* Conduct bivariate analysis to uncover relationships between variables.
* Investigate specific research questions related to career prospects and gender-specialization preferences in CSE.

**STEPS INVOLVING IN EXPLORATORY DATA ANALYSIS:**

1. **OVERVIEW OF DATASET:**

**Shape:** Considered dataset has overall shape of 4000 rows with 39 columns.

**Describe function:** It provides a summary of the central tendency, dispersion, and shape of the dataset's distribution, which is crucial for understanding the data at a glance. Here's an explanation of what each statistic represents:

* **Count:** Represents the number of non-null observations in each column i.e. **3998** for all columns
* **Mean:** Indicates the average value of the data. The mean of **10th and 12th percentage** is **77.92 and 74.46.**
* **Std (Standard Deviation):** Measures the spread of data around the mean. A higher standard deviation implies greater variability. The Standard deviation is **10th and 12th percentage** is **9.85 and 10.99**
* **Min:** Represents the minimum value in the dataset. The min of **10th and 12th percentage** is **43 and 40.Year of pass is 1995**
* **25% (First Quartile):** Also known as the lower quartile, it signifies the value below which 25% of the observations fall.**71.68 and 66.00** are the first quartile of **10th and 12th percentage**.
* **50% (Median or Second Quartile):** Represents the middle value of the dataset. It is the value below which 50% of the observations fall.**79.15 and 74.4** are the second quartile of **10th and 12th percentage.**
* **75% (Third Quartile):** Also known as the upper quartile, it signifies the value below which 75% of the observations fall.

**85.67 and 82.60** are the third quartile of **10th and 12th percentage**.

* **Max:** Represents the maximum value in the dataset.
* **Info():** The info() method in pandas provides a concise summary of a DataFrame, including the data types, non-null counts, and memory usage.

**2. DATA CLEANING:**

**1. Next step is looking for any null values and duplicates in the dataset.**

our task involves inspecting the dataset for any instances of null values and duplicates. Utilizing the isnull() and duplicated() functions, we conducted thorough checks, revealing no occurrences of null or duplicated values within the dataset.

#### 2. Here we are transforming the column name by using strip and lower function and as well dropping the unwanted columns

In this phase, we are refining the column names by employing the strip() and lower() functions to eliminate excess spaces and convert them to lowercase. Subsequently, we proceed to discard the redundant columns deemed irrelevant for our analysis.

**3.DataType Conversion:**

Upon inspecting the data types of each variable using the .info() function, it has been observed that while the 'doj' (Date of Joining) and 'dob' (Date of Birth) columns are appropriately interpreted as datetime objects, the 'dol' (Date of Leaving) column is currently identified as an object. To rectify this inconsistency, we are standardizing the 'dol' column to datetime format. Given that the data is from 2015, we are replacing the 'present' entries in the 'dol' column with the current date to ensure consistency and accuracy in our analysis.

**4. Here we are replacing the values with consisting values for better performance of the dataset**

To enhance the performance and consistency of the dataset, several value replacements and transformations are being implemented:

* For the 'college city tier' column: Replacing 0 with 'No' and 1 with 'Yes'.
* For the 'college tier' column: Replacing 1 with 'Tier One' and 2 with 'Tier Two'.
* For the 'college GPA' column: Converting GPA scores to percentages to ensure uniformity.
* For '10board' and '12board' columns: Mapping values to appropriate representations for improved handling.
* For other columns: Replacing -1 values with 0 (null) across all relevant columns.
* For the 'designation' column: Addressing 'get' as a NaN value by replacing it with the most frequent value using mode.
* For the 'specialization' column: Updating value names to stream names for clarity and consistency in analysis.

**5. Ensuring completeness and accuracy in our dataset, we are imputing mode values into categorical columns and median values into numerical columns:**

* Mode Imputation for Categorical Columns:

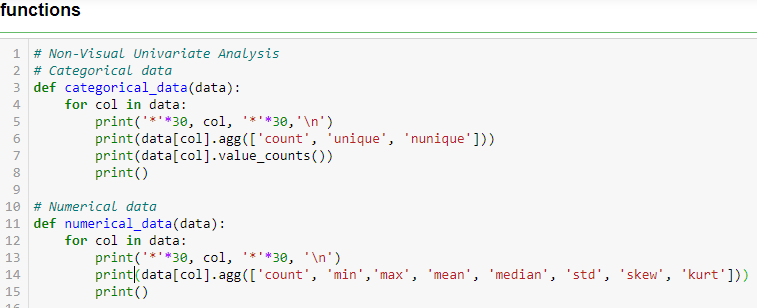
Missing values in categorical columns will be replaced with the mode, representing the most frequently occurring value in each respective column.

* **Median Imputation for Numerical Columns:**

Missing values in numerical columns will be replaced with the median, representing the middle value of sorted numerical datasets for each respective column.

### EXPLORATORY DATA ANALYSIS:

### Non Visualization Analysis:



We have implemented two functions, namely "categorical\_aggregations" and numerical\_aggregations", to perform aggregations on categorical and numerical column, respectively.

**Categorical Aggregations:**

**Count:** Counts the total number of non-null values in each categorical column.

**Unique:** Identifies the unique values present in each categorical column.

**Nunique (Number of Unique):** Counts the number of unique values in each categorical column.

**Numerical Aggregations:**

**Count:** Calculates the total number of non-null values in each numerical column.

**Min:** Finds the minimum value in each numerical column.

**Max:** Finds the maximum value in each numerical column.

**Mean:** Computes the average (mean) value of each numerical column.

**Median:** Determines the median value of each numerical column.

**Standard Deviation (Std):** Measures the dispersion or spread of data around the mean.

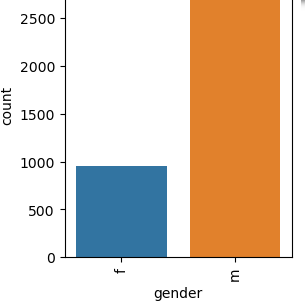
**Variance:** Measures the variability or spread of data points around the mean.

**Skewness:** Measures the asymmetry of the distribution of values in each numerical column. Positive skewness indicates a tail on the right side of the distribution, while negative skewness indicates a tail on the left side.

**Kurtosis:** Measures the peakedness or flatness of the distribution of values in each numerical column. Higher kurtosis indicates a sharper peak and heavier tails compared to a normal distribution.

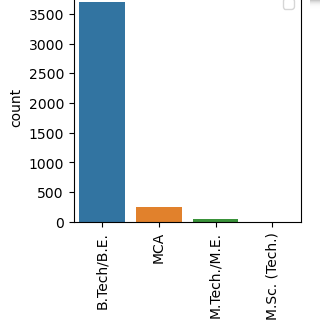
**UNIVARIATE VISUALIZATION:**

**GENDER:**

****

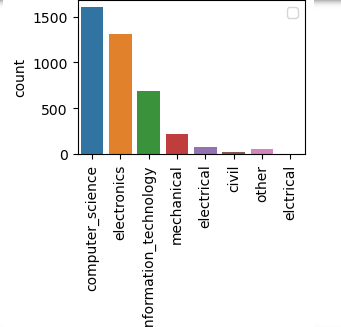
The observation suggests a higher number of male students compared to female students.

**DEGREE:**

****

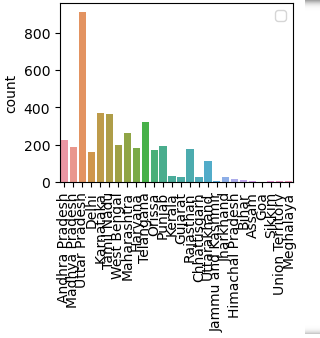
The analysis reveals that a greater proportion of employed individuals hold a B.Tech degree compared to any other educational qualification.

**SPECIALIZATION:**

****

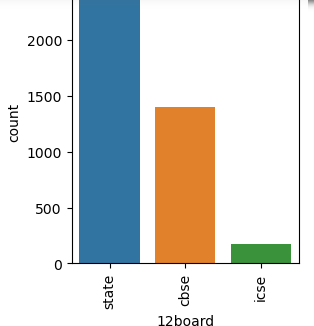
The data indicates that individuals with a background in computer science exhibit the highest rate of employment within the IT domain.

**COLLEGE STATE:**

****

* The findings suggest that Uttar Pradesh emerges as the leading state for college graduates, with a significant portion of graduates also receiving noteworthy salaries.

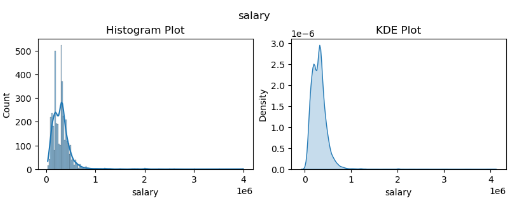
**12 Board:**

****

* The analysis highlights that the majority of students come from state board educational backgrounds, closely followed by students from CBSE-affiliated schools..

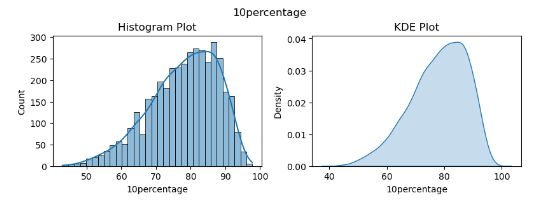
**Numerical Columns:**

**SALARY:**

****

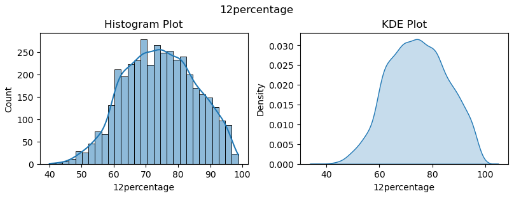
The KDE plot illustrates a bimodal distribution within the data, with a notable concentration around the value of 3, indicating higher data density. Meanwhile, the histogram displays a slight right skewness, suggesting an unequal spread of data.

**10 PERCENTAGES:**

****

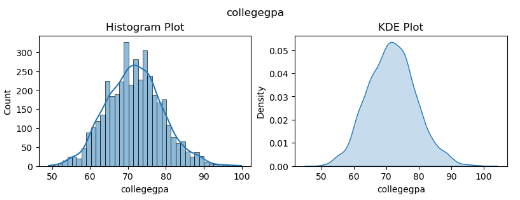
The analysis indicates that the majority of students have a performance spread ranging from 70% to 95%

**12 PERCENTAGE:**



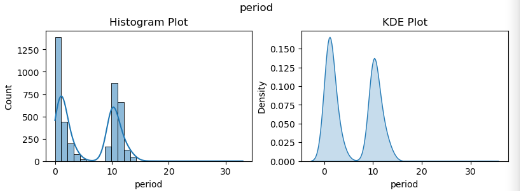
The histogram and KDE plot demonstrate that a significant portion of students exhibit a spread between 80% to 95%.

**COLLEGEGPA:**

****

Based on the histogram and KDE plot presented above, it can be inferred that the majority of students achieved a score of approximately 75% during their college tenure.

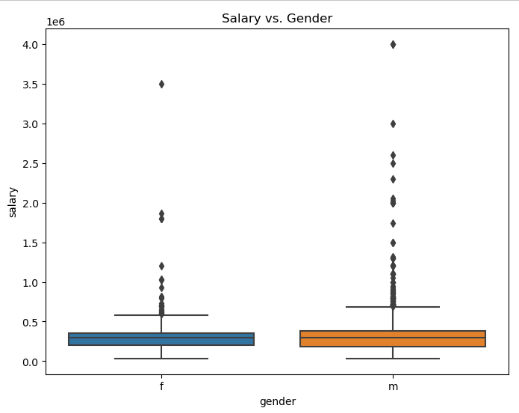
**EXPERIENCE:**

****

The bimodal distribution suggests that individuals in jobs primarily possess experience within the range of 0-4 years and 9-13 years.

### Bi- Variate Analysis

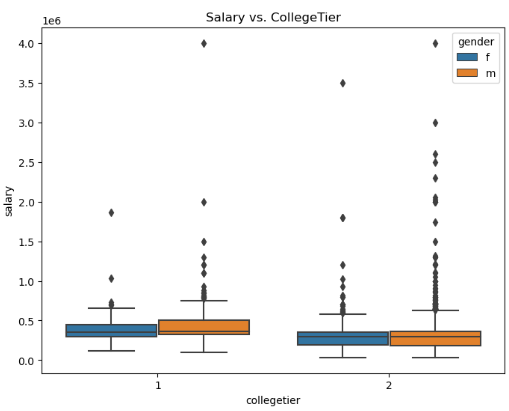
**SALARY VS GENDER:**

****

**Observations:**

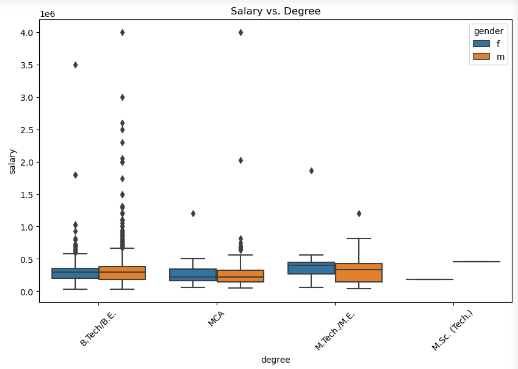
* The salary data contains numerous outliers.
* There is minimal disparity between the median salaries for both genders.
* Males exhibit more outliers, suggesting a higher proportion of individuals receiving higher pays compared to females.

**SALARY VS COLLEGE TIER:**

****

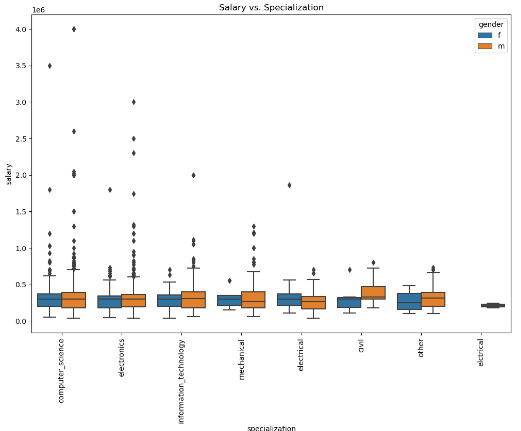
Tier 1 indicates a higher level of performance, typically associated with higher

**SALARY VS DEGREE:**

****

* The analysis reveals that students with a BTech degree secure a greater number of job opportunities, with outliers indicating higher salaries predominantly awarded to them. Conversely, MTech graduates, especially females, tend to receive comparatively lower salaries.

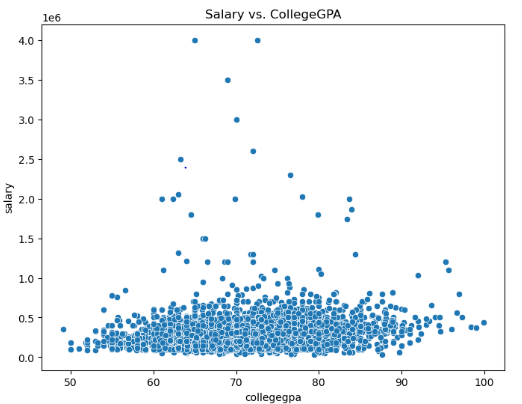
**SALARY VS SPECIALIZATION:**

****

**Observations:**

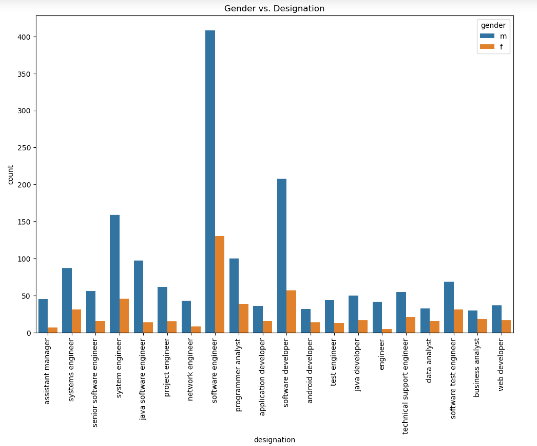
* Within the IT specialization, male professionals tend to command higher salaries compared to other specializations. Additionally, the median salary between male and female professionals remains similar.
* Across all specializations, there is a higher representation of male students compared to female students, with the majority stemming from IT and Electronics streams.

**SALARY VS COLLEGE GPA:**

****

From the scatter plot, it is evident that the majority of students earn up to 3 lakhs per annum.

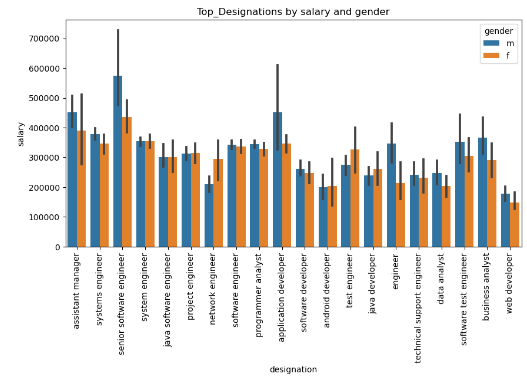
**JOBCITY AND DESIGNATION:**

****

**Observations:**

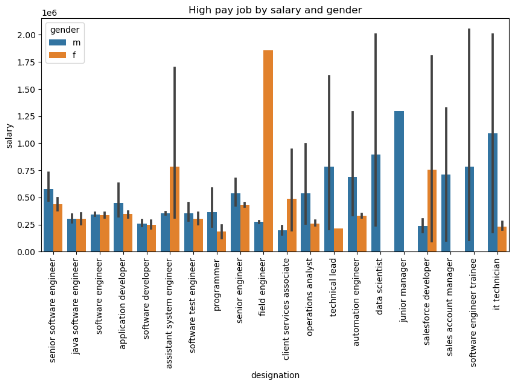
* Among the various designations, "Software Engineer" holds the top position, followed by "Software Developer", "System Engineer", "Java Software Developer", and "Software Test Engineer".
* Both male and female students predominantly aspire to become "Software Engineers".
* A notable trend is that females tend to consider careers in "Programmer Analyst" more frequently compared to other designations, following "Software Engineering".

**TOP DESIGNATION WITH SALARY IN GENDER:**

****

Observations:

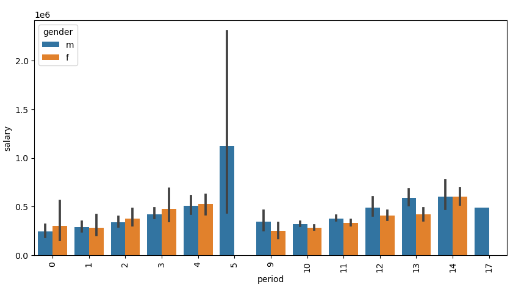
* "Senior Software Engineer" emerges as the top designation in terms of salary, followed by "Assistant Manager", "Application Developer", and "Business Analyst".
* Both female and male professionals predominantly occupy these top designations.

**HIGH PAYING DESIGNATION:**

**Observations:**

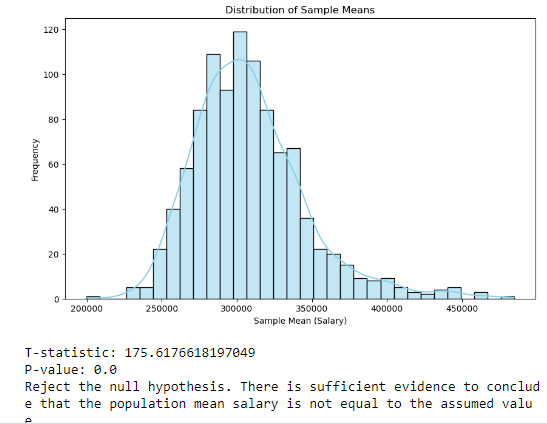
* "Field Engineer" stands out as the highest paying designation, with female students receiving the highest compensation in this role, followed by "Assistant System Engineer".
* Notably, there are no female students reported with salaries in roles such as "Data Science", "Junior Manager", "Sales Account Manager", and "Software Engineer Trainee".

**EXPERIENCE VS HIGH SALARIES**



* Here male with 5 years of experience has high salary. 9-11 years of experience irrespective of gender has less salary.

**HYPOTHESIS TESTING:**

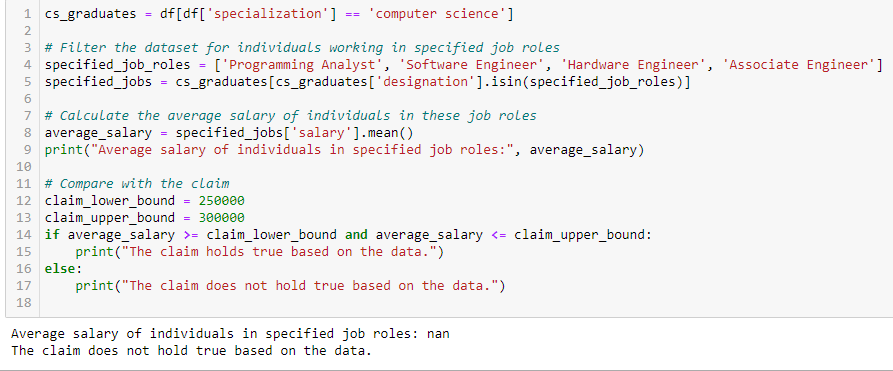


#### Observation:

* Conclusion: Since the P-value is below the significance level (alpha) of 0.05 (commonly used threshold), we reject the null hypothesis. This means that there is sufficient evidence to conclude that the population mean salary is not equal to the assumed value of 100,000 INR.
* In summary, based on the statistical analysis, we have strong evidence to suggest that the population mean salary differs significantly from the assumed value of 100,000 INR.

**5.Research Questions**

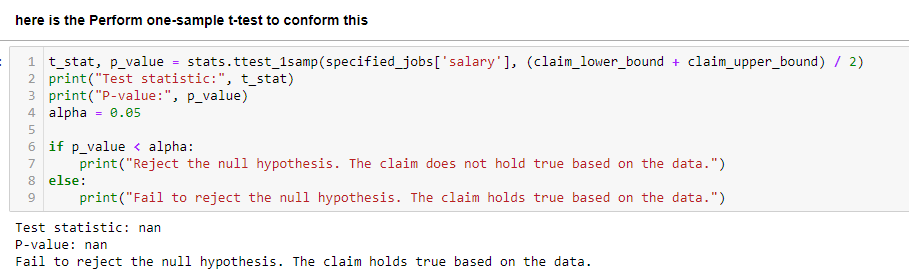
* Times of India article dated Jan 18, 2019 states that “After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate.” Test this claim with the data given to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



**Observasions:**

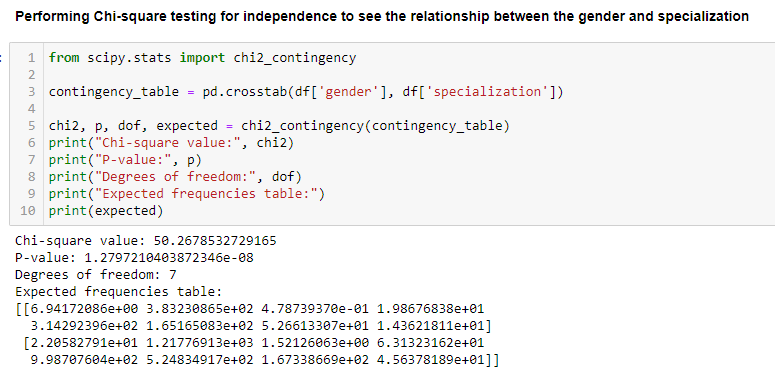
The claim does hold true based on the date.

**ONE-SAMPLE T TEST:**



The claim holds true based on the data.

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

****

**Observasions:**

* The calculated p-value (1.27) exceeds the significance level (e.g., 0.05), indicating a failure to reject the null hypothesis. This implies that there is no significant relationship between gender and specialization within the dataset. Thus, the preference for specialization appears to be independent of gender based on the provided data.
* To validate the claim made in the article using the dataset, a filtering process is required. Specifically, we need to isolate individuals who have completed Computer Science Engineering and are employed in the specified job roles. Subsequently, we can compute the average salary of these individuals and compare it with the claim asserted in the article.