INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project - AMCAT Data Analysis

Revanth Christober M

February 20th 2024

About me

Driven Data Science Junior with a Passion for Impact

- I'm a highly motivated and results-oriented individual currently pursuing my Senior Secondary from the National Institute of Open Schooling (NIOS).
- This flexible learning environment fuels my passion for exploring data science and its transformative potential.
- Ever since I was young, I've been fascinated by the power of data to change the world, and I'm eager to contribute to that change.
- Through online courses (primarily using NPTEL) and dedicated self-study, I've built a strong foundation in data science and machine learning.
- I'm not afraid to dive into challenging projects, and I've applied my skills to various real-world applications, including:
  - **Income Classifier Model (Decision Tree) for Subsidy Inc.** (89% accuracy)
  - **Regression Model for Crypto Data Hourly Volume-Price Analysis and Prediction** (81% accuracy)
  - **Classification Model for Brazilian E-Commerce Public Dataset by Olist** (92% accuracy)
  - **Interactive Dashboard using Python and R** (projects mentioned in your portfolio)
- Beyond technical skills, I enjoy working in collaborative environments and connecting with people.
- My friendly and proactive nature allows me to learn quickly and excel under pressure, even with the flexible learning model provided by NIOS.
- I'm eager to leverage my data science skills to make a positive impact.

Contact me:

INNOMATICS RESEARCH LABS

- **I. PROJECT OBJECTIVE:**
  - The primary objective of this analysis is to gain insights from the provided dataset, with a focus on understanding the relationship between various features and the target variable, which is Salary. Specifically, the goals include:
    - Describing the dataset comprehensively, including its features and structure.
    - Identifying patterns or trends present in the data.
    - Exploring the relationships between independent variables and the target variable (Salary).
    - Identifying outliers or anomalies in the data.
- **II. SUMMARY OF DATA:**
  - The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, is centered around employment outcomes for engineering graduates. It contains dependent variables such as Salary, Job Titles, and Job Locations, along with standardized scores in cognitive, technical, and personality skills. With approximately 40 independent variables and 4000 data points, the dataset encompasses both continuous and categorical data. It also includes demographic features and unique identifiers for each candidate.
- **III. DATA CLEANING AND PREPROCESSING:**
  - A. Datatype Conversion:
    - To ensure data accuracy and consistency, we converted the data types of the 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) fields to datetime objects. Since the survey was conducted in 2015, we assumed that respondents who indicated 'present' for DOL had left the company by the latest survey date, recorded as 2024-02-17. Therefore, we replaced 'present' values in the DOL field with this end date.
  - B. Validating 0 or -1:
    - We assessed the presence of null values represented by 0 or -1 in specific columns. The columns '10board', '12board', 'GraduationYear', 'JobCity', and 'Domain' were processed to handle these null values by replacing them appropriately.
- The data handling process has been successfully completed, ensuring that the dataset is prepared for further analysis.

- **C. Collapsing Categories:**
  - **Columns with over 80% of values being -1, including 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', and 'CivilEngg', were removed from further analysis. For the remaining optional subject columns, 'ElectronicsAndSemicon' and 'ComputerScience', -1 values were replaced with 0, indicating that the subjects were not pursued.**

- **IV. FEATURE ENGINEERING:**

  1. Age Calculation:
  - **An additional column representing age has been incorporated into the dataset by subtracting the year of birth (DOB) from 2015, reflecting the individual's age as of 2015.**

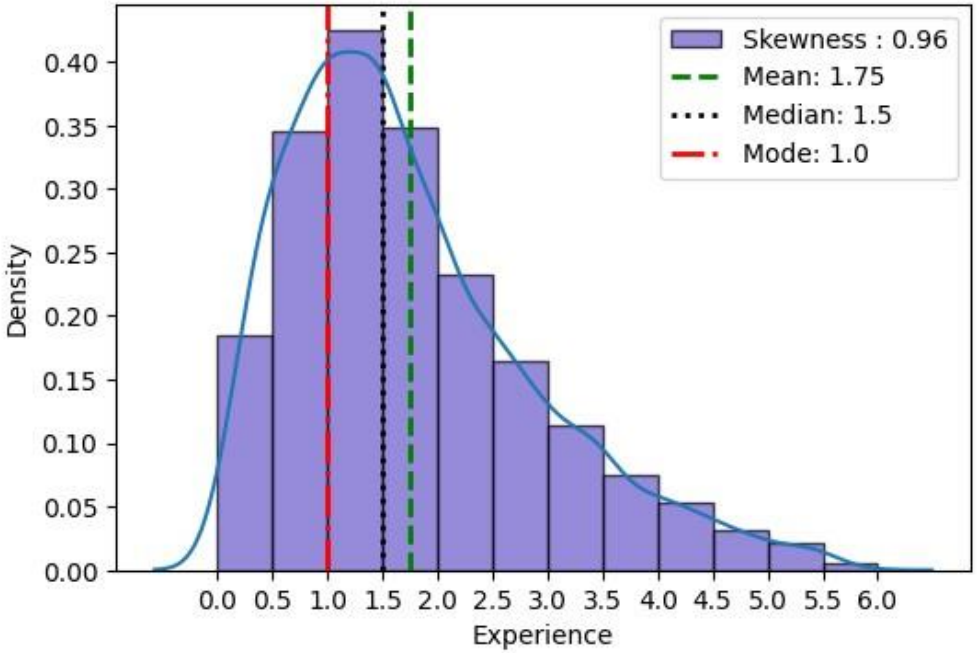
  2. Tenure Calculation:
  - **Another new feature, 'tenure', has been introduced by subtracting the 'Date of Leaving' (DOL) from the 'Date of Joining' (DOJ). This indicates the duration of an individual's employment within the company.**

  3. Graduation Year Filtering:
  - **Rows where the graduation year is greater than or equal to the date of joining have been removed. This ensures data integrity by excluding instances where the graduation year suggests a date after the individual's employment start date.**
  - **A custom function has been developed to calculate the Cumulative Distribution Function (CDF), allowing for the analysis of the distribution of a variable's values within the dataset. This function facilitates insights into the cumulative probability distribution of the data, aiding in statistical analysis and decision-making processes.**

4. Cumulative Distribution Function (CDF) Function:

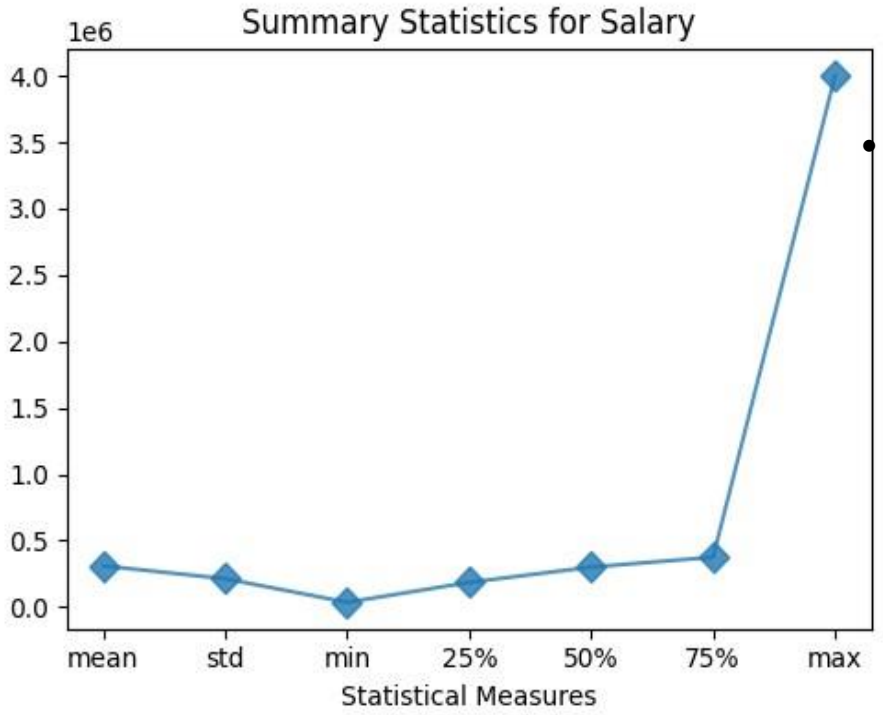
```
def cumulative_distribution_function(df):
    X = np.sort(df)
    Y = np.arange(1, len(x)+1) / len(x)
    return (X, Y)
```

- **V. EXPLORATORY DATA ANALYSIS:**
  - A. Univariate Analysis - Continuous Features:



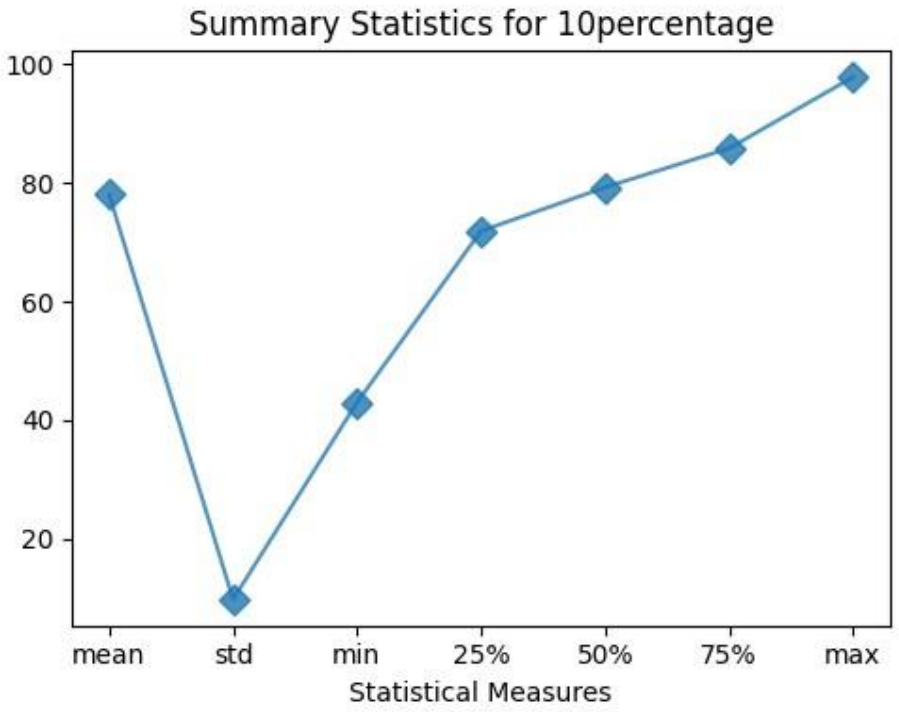
1. Tenure:

Summary plots showcased a 4-year experience range. Histograms displayed a positively skewed distribution with a median tenure of 1.5 years and outliers signifying longer tenures. Box plots further emphasized these outliers. Additionally, the Cumulative Distribution Function (CDF) highlighted the non-normal distribution of tenure. These findings provide valuable insights into workforce dynamics.



1.2 Salary:

- The summary plot of the salary data indicates considerable variation among the salary values. The histogram reveals a significant positive skewness, suggesting that the distribution of salaries is not symmetrical and is skewed towards higher values.
- Box plots further emphasize a concentration of high salaries, indicating the presence of outliers at the upper end of the salary distribution. These outliers represent individuals with exceptionally high salaries compared to the rest of the dataset.
- Furthermore, the cumulative distribution function (CDF) analysis underscores the data's skewness, showing a departure from the typical pattern of a normal distribution. The curve of the CDF deviates notably from the expected diagonal line, indicating the non-normal distribution of salary values within the dataset.
- Overall, these observations highlight the presence of significant variations and skewness in the distribution of salaries among the individuals in the dataset.



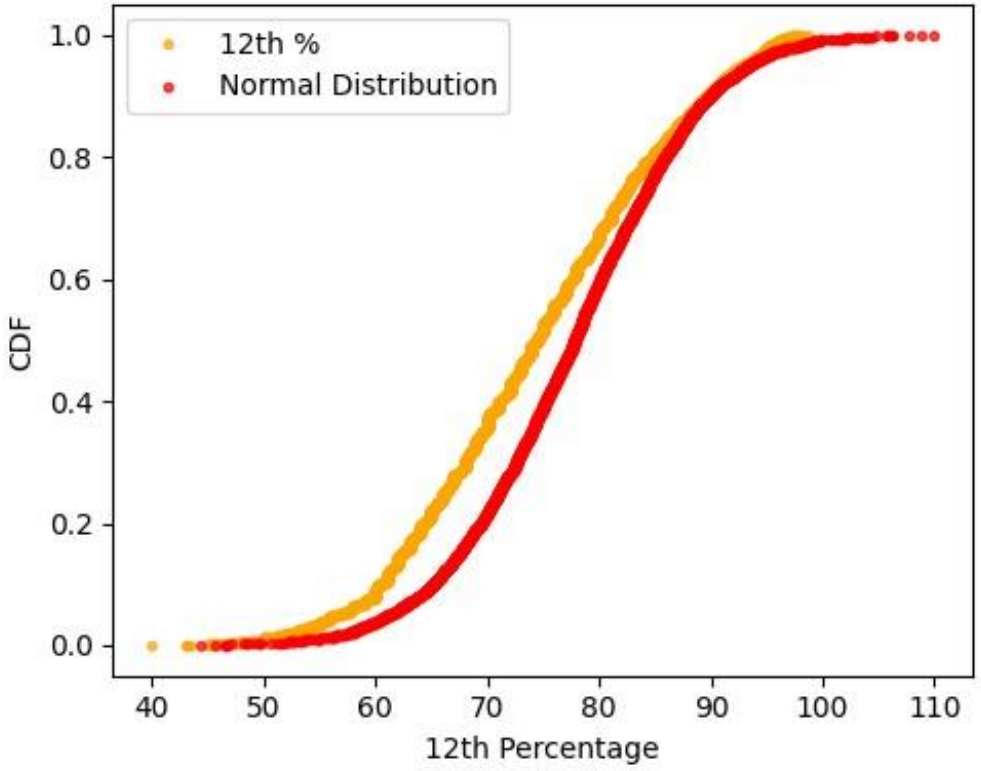
1.3 10th Percentage:

- **The summary plot of the 10th percentage data shows that around half of the students achieved scores of approximately 80% or lower. The histogram illustrates a scarcity of students with low percentages, with the majority falling within the 75% to 90% range, peaking at around 78%.**

- **Extreme outliers are evident from the box plot, indicating some irregularities in the distribution of the data. These outliers represent individuals with exceptionally high or low 10th percentage scores compared to the rest of the dataset.**

- **Moreover, the cumulative distribution function (CDF) analysis highlights skewness in the data distribution, deviating from the expected pattern of a normal distribution. The curve of the CDF deviates noticeably from the diagonal line, further emphasizing the non-normal distribution of 10th percentage scores within the dataset.**

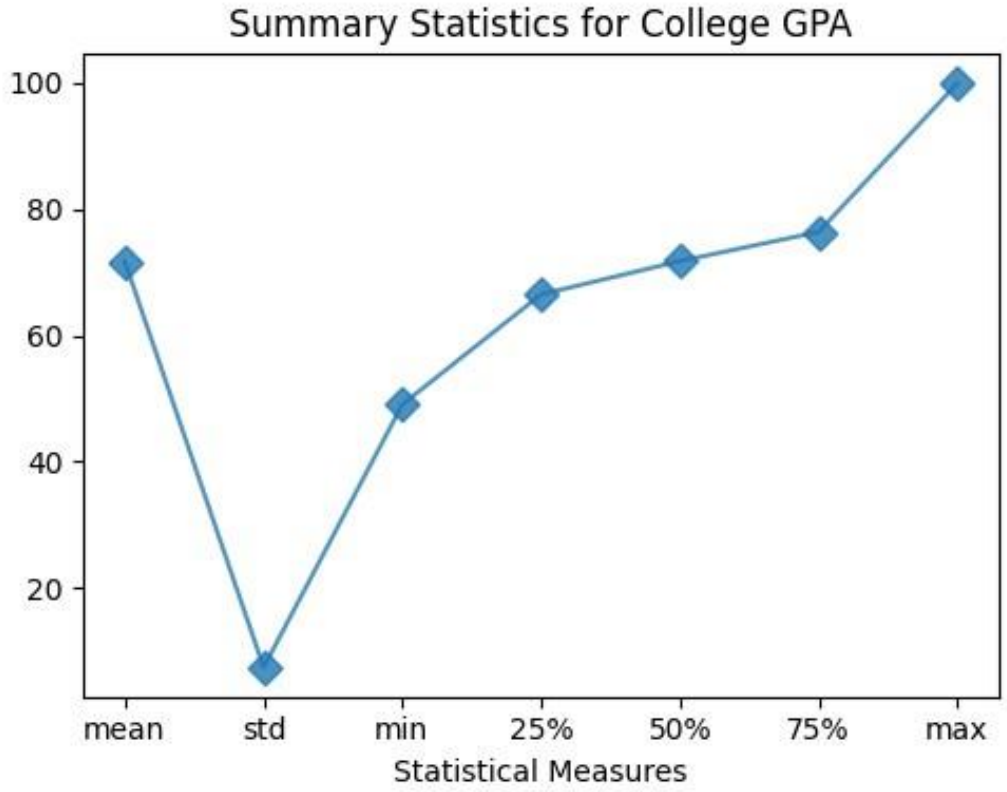
- **Overall, these observations suggest that while the majority of students achieved scores within a relatively narrow range, there are notable variations and outliers present in the distribution of 10th percentage scores among the individuals in the dataset.**



1.4 12th Percentage:

- **Upon analyzing the dataset, it is evident that approximately half of the students scored around 78% or lower in their 12th-grade examinations. This indicates that a significant portion of students achieved scores below the 80% mark, highlighting a scarcity of low scores in the dataset.**
- **Examining the histogram, it is observed that the majority of students scored between 69% and 84%, with the highest frequency occurring around the 70% mark. However, an outlier with an exceptionally low score is noticeable, suggesting an anomaly in the data distribution.**
- **Further investigation using the cumulative distribution function (CDF) confirms that the data does not follow a normal distribution pattern. The curve of the CDF exhibits deviations from the expected diagonal line, indicating skewness in the distribution of 12th percentage scores within the dataset.**
- **In summary, while most students scored within a relatively narrow range, the presence of outliers and the non-normal distribution pattern emphasize variations in the 12th percentage scores among individuals in the dataset.**

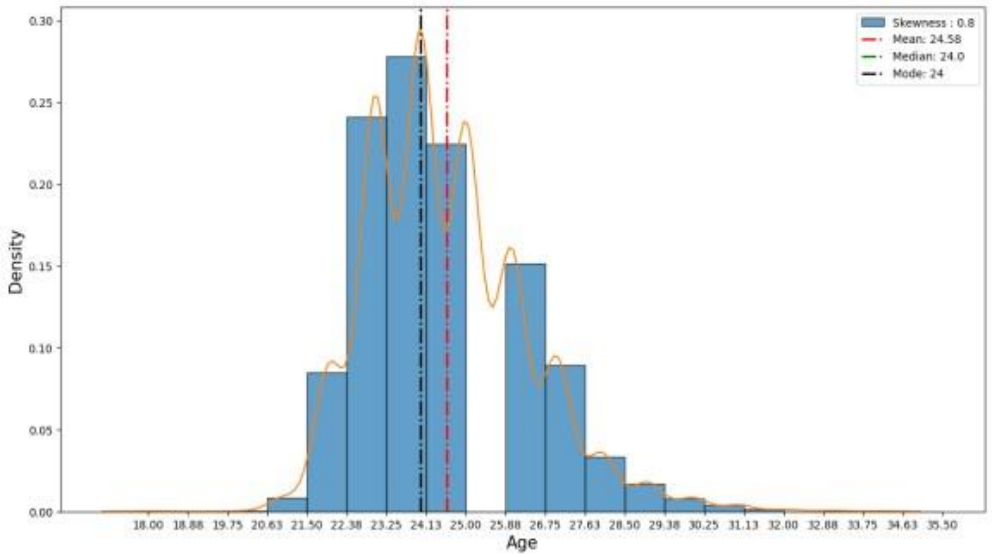
1.5 College GPA:



- An analysis of student GPAs provides valuable insights into academic performance. According to the summary plot, approximately 75% of students had GPAs around 80% or lower. This indicates that a significant portion of students achieved GPAs below the 80% mark.
- Examining the histogram, it is observed that most students had GPAs between 63% and 78%, with the highest frequency occurring around the 70% mark. This suggests that the majority of students achieved GPAs within a relatively narrow range. Additionally, the average GPA is calculated to be 74%.
- Both low and high extreme values are apparent in the dataset, as indicated by the box plot. This variability in GPAs suggests differences in academic performance among students, with some achieving exceptionally high or low GPAs compared to the rest of the dataset.

- Interestingly, the cumulative distribution function (CDF) suggests that the data is sufficiently normally distributed. This indicates that the distribution of GPAs follows a pattern similar to a normal distribution, contributing to its reliability for further analysis.
- In summary, while the majority of students achieved GPAs within a certain range, the presence of extreme values and the normal distribution pattern suggest variations in academic performance among individuals in the dataset.

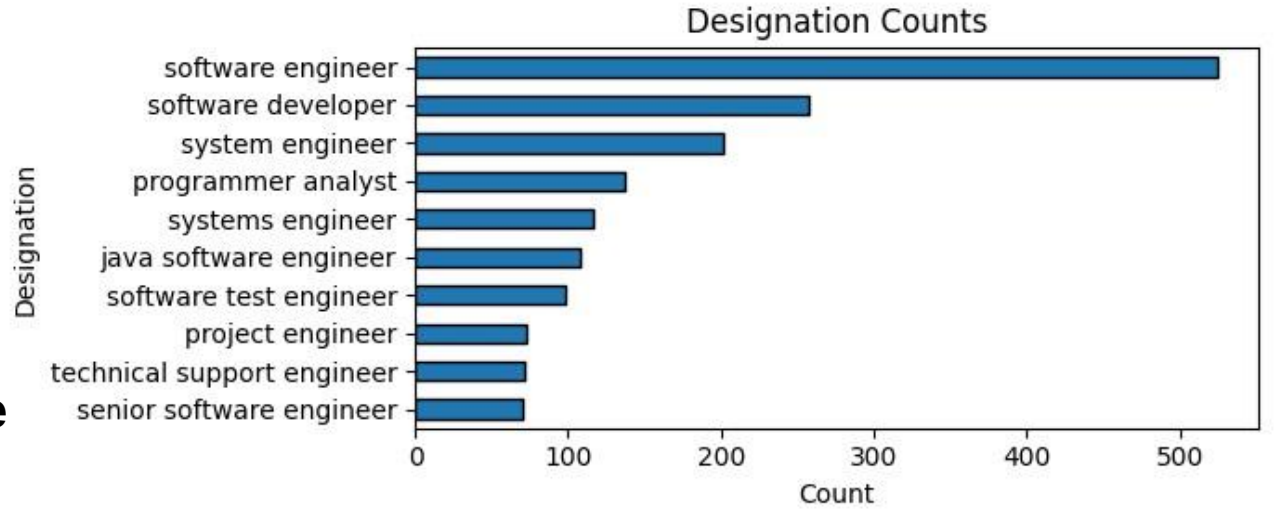
1.6 English, Logical, Quant, Computer Programming, Electronics & Semiconductors, Age:

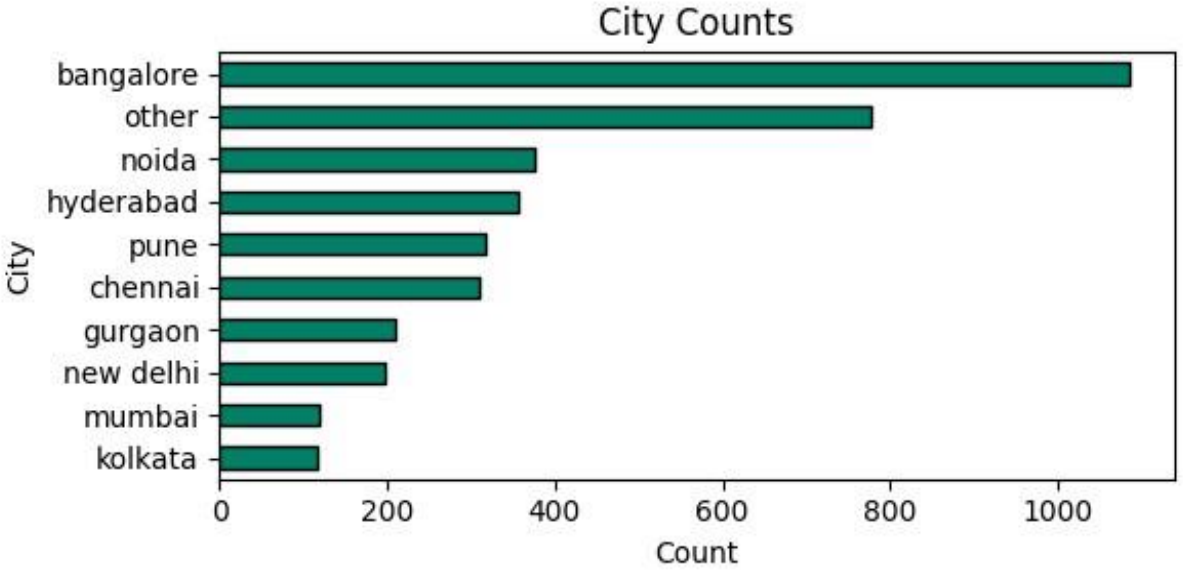


- Analyzing the dataset across various subjects reveals distinct patterns in student performance. In English exams, approximately half of the students scored below 500, with scores predominantly ranging from 389 to 545. There is also a noticeable presence of extreme values, indicating significant variability in English proficiency among students.
- Similarly, in Logical exams, a significant portion of students scored below 500, with scores concentrated between 454 to 584. This distribution also displays both lower and higher extreme values, suggesting variations in logical reasoning abilities among students.
- In Quants, a majority of students scored below 600, with scores spanning from 425 to 608. This indicates a mix of low and high extreme values, highlighting differences in quantitative reasoning skills among students.
- Conversely, in Computer Programming, around 50% of students scored below 500, with scores clustering between 416 to 459. There is a notable presence of extreme values, suggesting varying levels of proficiency in computer programming among students.
- Electronics and Semiconductors saw about 75% of students scoring less than 250, with scores mainly falling between 0 to 79. This indicates a non-normal distribution, with a majority of students scoring relatively low in this subject.
- Lastly, the age distribution indicates that approximately 75% of students are under 26 years old, with the majority aged between 22 to 25. Notable outliers are observed at both ends of the age spectrum, suggesting a diverse age range among students in the dataset.
- In summary, the analysis of these subjects reveals variations in student performance and age distribution, providing valuable insights into the dataset.

2. Categorical Features:

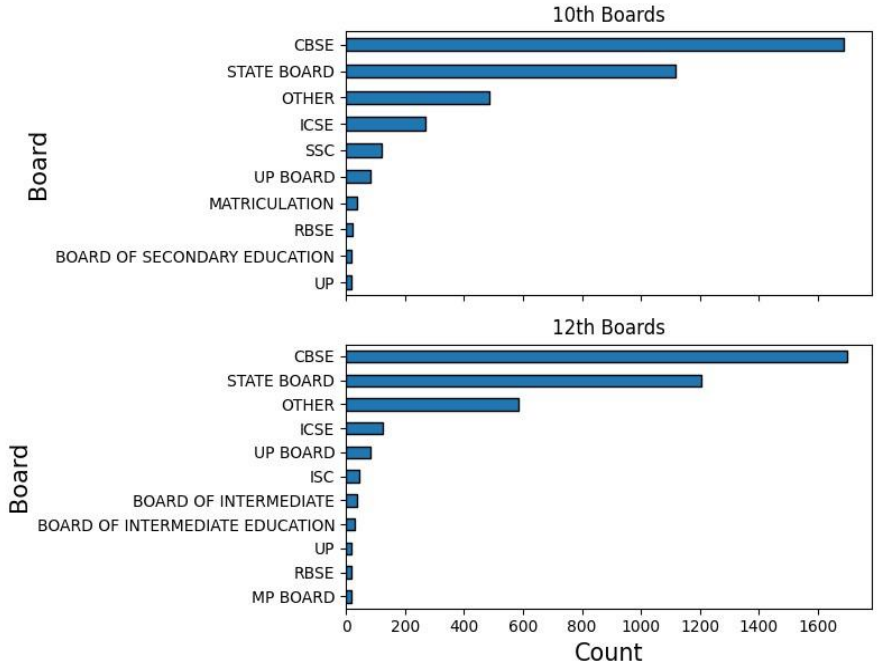
- The dataset provides valuable insights into the demographics and educational backgrounds of individuals across various categories.
- In terms of Designation, Software Engineer emerges as the most prevalent designation, followed by System Engineer and Software Developer. An "OTHER" category is also present, indicating diverse job titles among respondents.



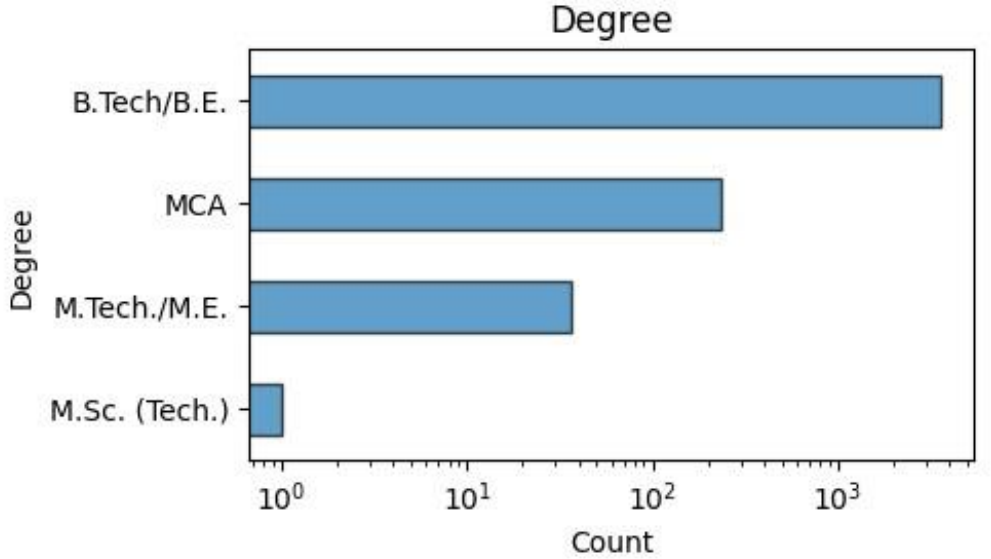


- Regarding job city preferences, Bangalore stands out as the most favorable city for job placements, followed by Noida, Hyderabad, and Pune. In contrast, Mumbai and Kolkata are less preferred options for job seekers.
- Gender distribution reveals an imbalance, with a significantly larger male population compared to females. This highlights potential gender disparities in the workforce.

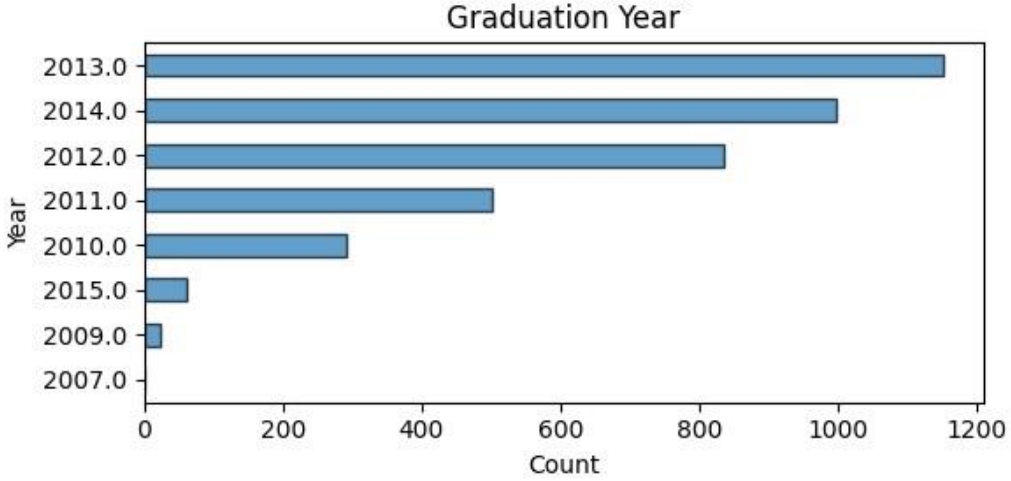
INNOMATICS RESEARCH LABS



- **CBSE emerges as the most common school board for both 10th and 12th grades, indicating its widespread popularity among students. College Tier analysis indicates a dominance of Tier 1 colleges, suggesting a preference for higher-ranked institutions among respondents.**



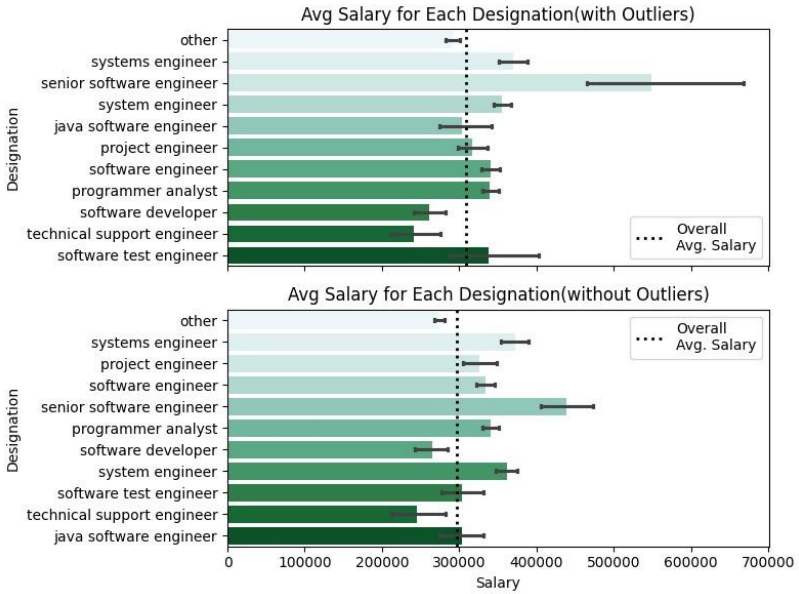
- **Most students have pursued a B.Tech degree, indicating a prevalence of engineering education in the dataset. Minimal representation is observed from M.Sc(Tech) graduates, suggesting a lesser common educational background among respondents.**



- **The majority of colleges are located in Tier 0 cities, indicating a concentration of educational institutions in urban areas. Graduation years analysis reveals that 2013 saw the highest number of graduations, followed by 2014 and 2012. These observations collectively offer valuable insights into the educational and professional landscape captured by the dataset, highlighting trends and preferences among respondents in various categorical features.**

INNOMATICS RESEARCH LABS

B. Bivariate Analysis:

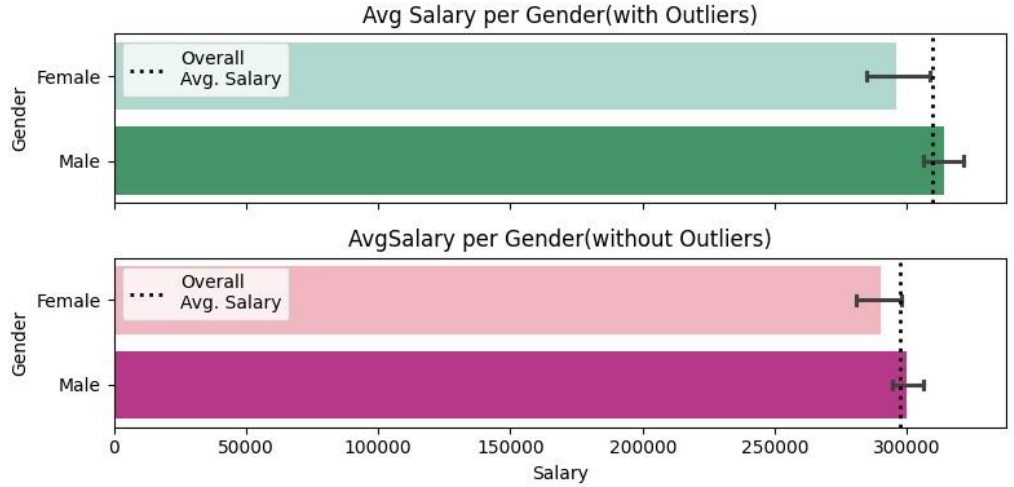


1. Designations & Salary:

- **Analyzing the relationship between designations and salary reveals that Senior Software Engineers tend to have the highest salary, albeit with the highest standard deviation. This indicates that while some Senior Software Engineers earn considerably high salaries, there is also significant variability in their compensation. On the other hand, Software Developers and Technical Support Engineers tend to have salaries below the average, suggesting differing salary ranges based on job titles within the dataset.**

2. Gender & Salary:

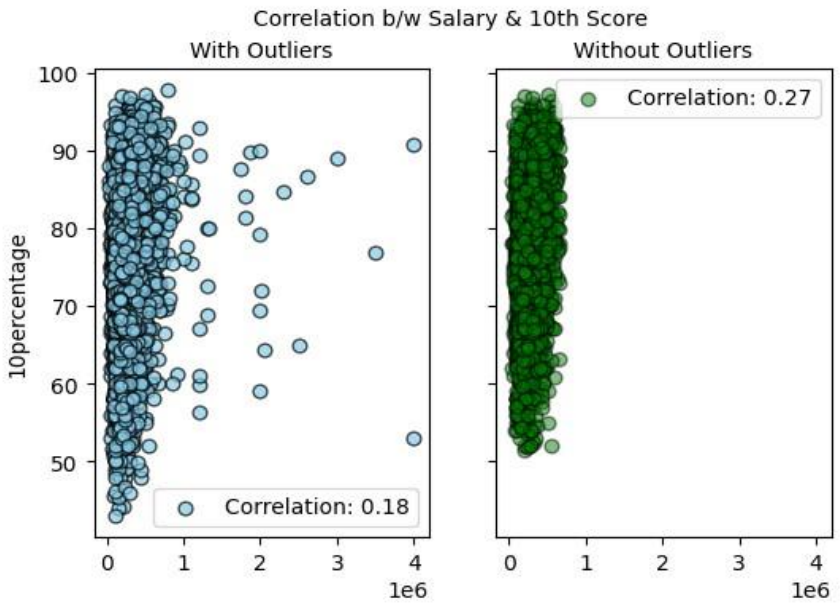
- **Examining the correlation between gender and salary indicates that, on average, male and female salaries are approximately equal. This suggests no significant gender bias overall in salary distribution. However, it is noted that females tend to receive salaries slightly below the overall average, indicating potential disparities in compensation based on gender.**



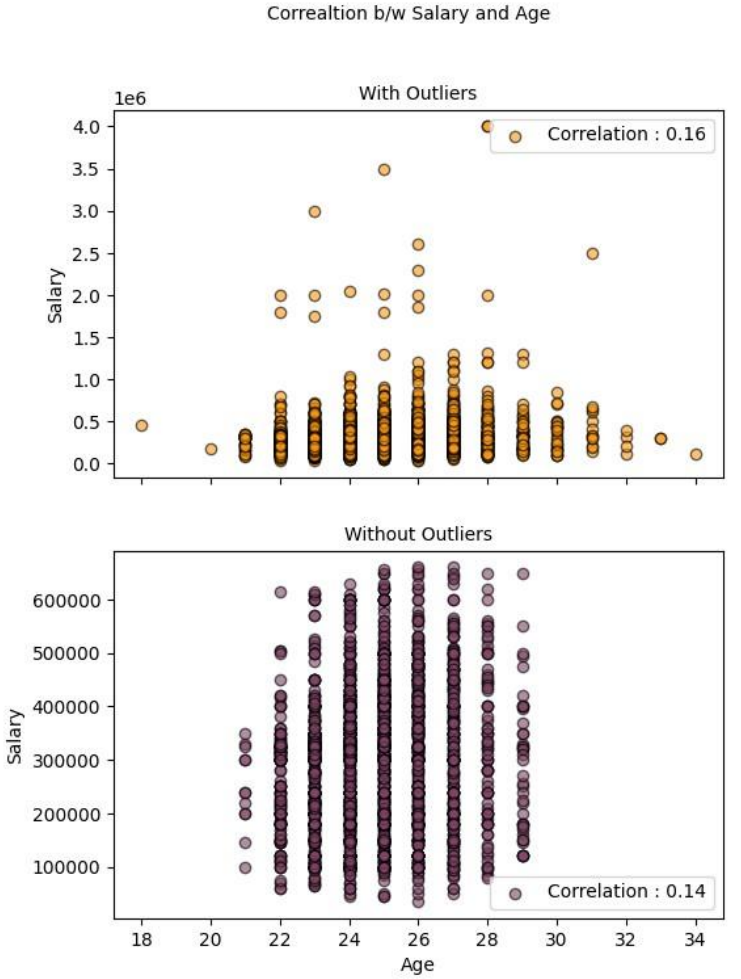
INNOMATICS RESEARCH LABS

3. Academic Scores & Salary:

There appears to be no significant correlation between salary and academic scores in 10th grade, 12th grade, or College GPA. This suggests that academic performance in terms of standardized exam scores or GPA does not strongly influence salary outcomes in the dataset.

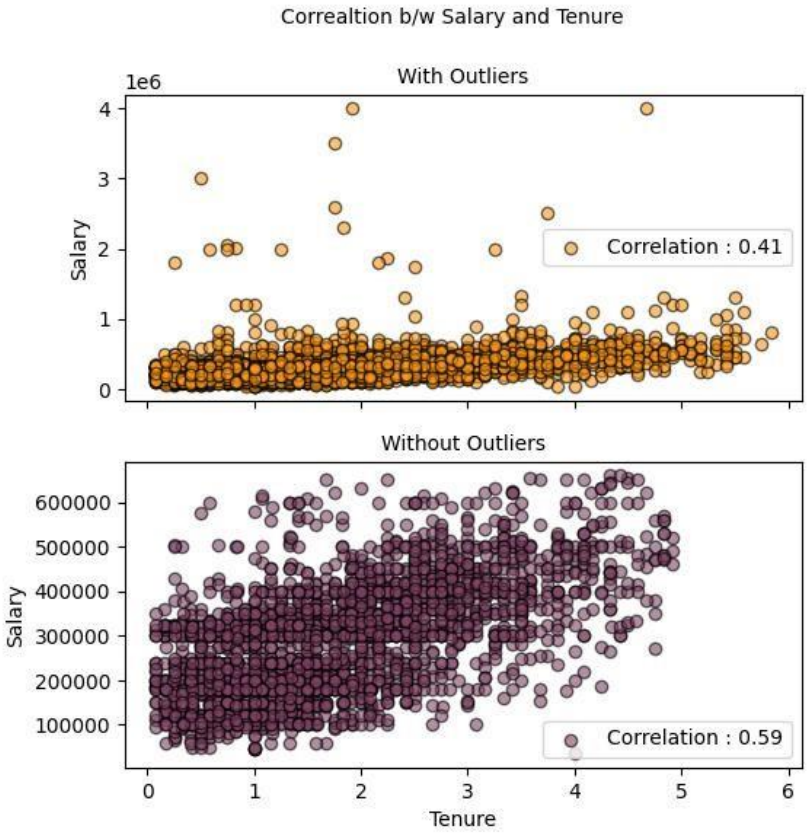


4. Age & Salary:



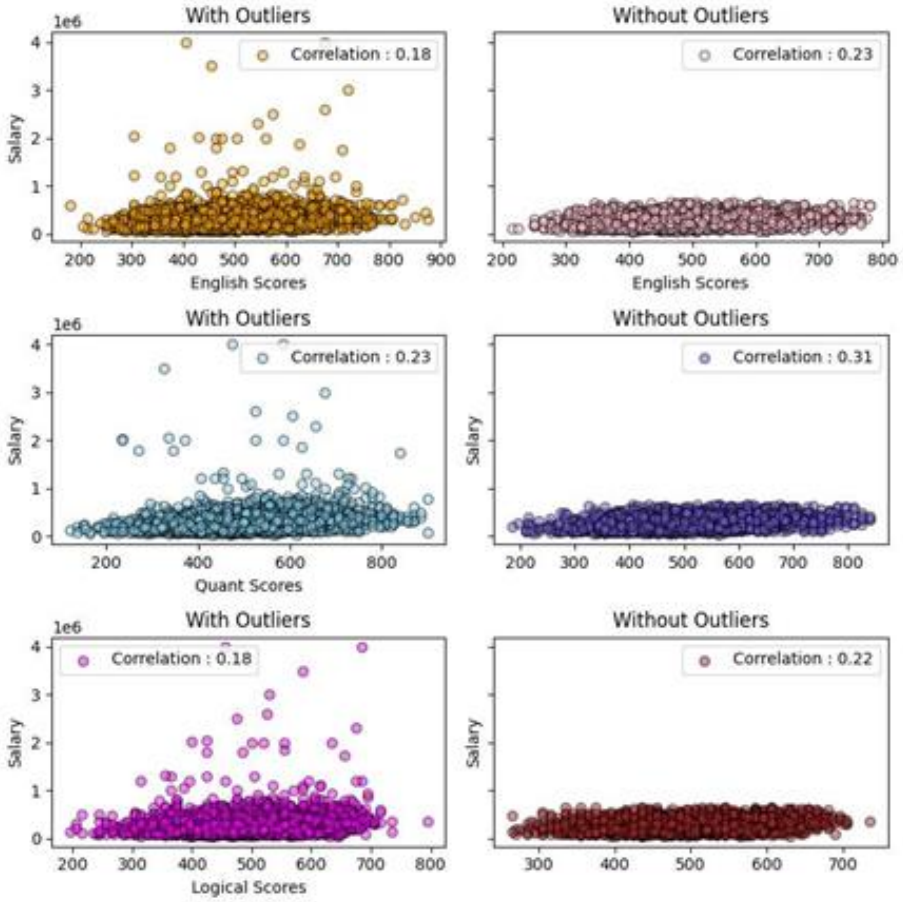
- **Upon investigating the relationship between age and salary after removing outliers, no apparent correlation or relationship is observed. This indicates that age does not play a significant role in determining salary outcomes among individuals in the dataset.**

INNOMATICS RESEARCH LABS



5. Tenure & Salary:

- **An analysis of tenure and salary reveals a positive correlation between the two variables. It is observed that as tenure increases, there is approximately a 50% increase in salary. This suggests that experience plays a significant role in determining salary outcomes, with individuals with longer tenures typically commanding higher salaries.**



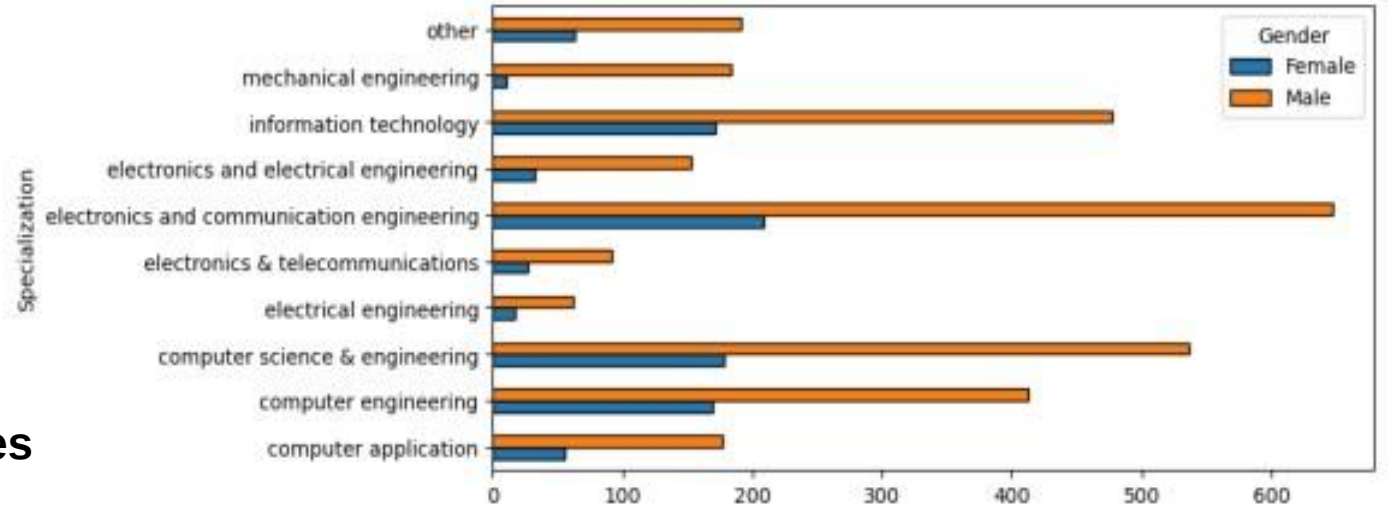
6. Skills & Salary:

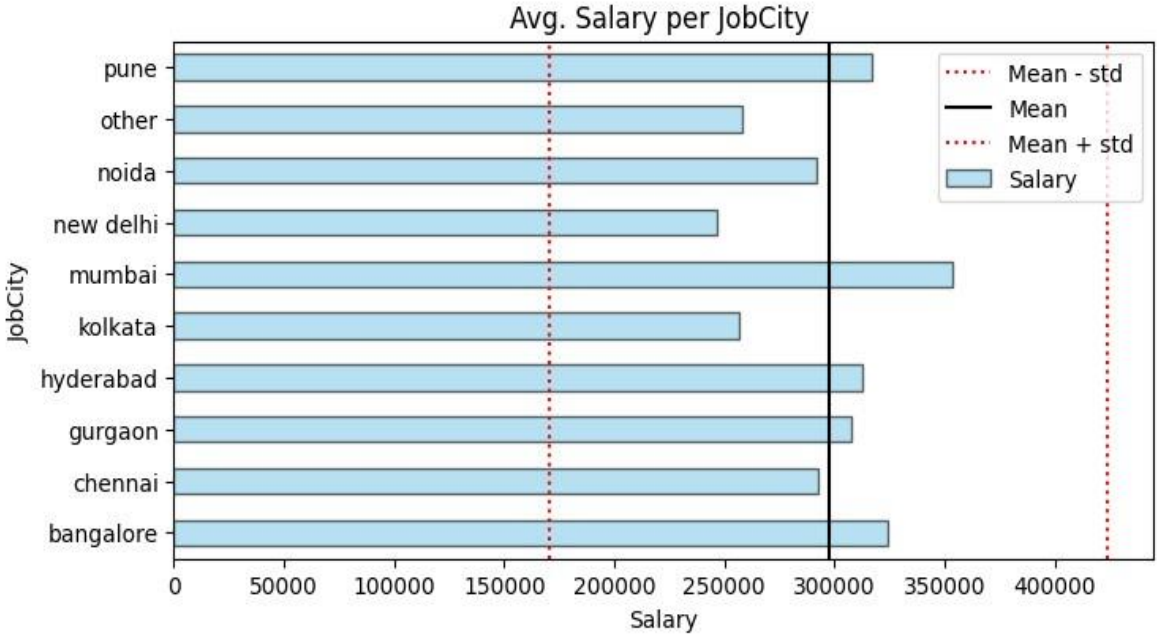
Investigating the relationship between various skills and salary indicates no apparent effect of English, Quants, or Logical scores on salary outcomes. This suggests that proficiency in these skills does not strongly influence salary levels in the dataset.

INNOMATICS RESEARCH LABS

7. Gender & Specialization:

- **Examining the gender distribution across different specializations reveals that male participation is approximately double that of females across all specializations. Additionally, fewer females opt for mechanical and electronics specializations compared to males. This gender disparity in specialization choices highlights potential gender-related factors influencing career paths and choices among individuals in the dataset.**





8. College Factors & Salary:

- **Comparing salary outcomes based on college factors reveals that Tier 1 colleges offer higher salaries compared to Tier 2 colleges. Additionally, cities in both Tier 1 and Tier 2 offer similar salaries to students, indicating that college tier may play a more significant role than the location of the college in determining salary outcomes.**

INNOMATICS RESEARCH LABS

- In conclusion, while factors such as tenure and college tier have a notable influence on salary, others such as gender and academic scores show little to no correlation with salary outcomes. Outliers in age were removed, suggesting that age alone does not dictate salary levels. These observations underscore the complex interplay of various factors influencing salary outcomes in the context of the dataset.

VI. RESEARCH OUTCOMES:

- A "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." This claim can be tested using the dataset to assess the salary levels of fresh graduates in these specific job roles and determine if they align with the salary range mentioned in the article.
- 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.**
- **Is there a relationship between gender and specialization? (i.e. Does the preference of Specialization depend on the Gender?)**

Test Value:
chi2_critical: 16.918977604620448
chi2_statistic: 48.62141720904882
chi2_p_value: 1.9542895953348e-07

- **The analysis conducted using a ChiSquare 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.**

VII. CONCLUSION

- The comprehensive analysis of the dataset has unveiled several significant findings regarding the factors influencing salary levels among individuals. While certain criteria, such as tenure and college tier, exhibit a strong association with compensation, others, including gender and academic performance, show little to no correlation.
- Senior Software Engineers emerge as the top earners within the dataset, commanding the highest incomes. However, it is noted that their salaries also display greater variability, indicating a higher degree of unpredictability in compensation compared to other job roles. On the contrary, Software Developers and Technical Support Engineers tend to earn salaries below the average, suggesting differing salary ranges based on job titles within the dataset.
- Gender does not seem to have a substantial impact on income determination on average, as evidenced by similar average salaries for both genders. However, it is observed that females tend to receive salaries slightly below the overall average, highlighting potential disparities in compensation based on gender.
- Academic performance, as measured by scores in 10th grade, 12th grade, and college GPA, does not exhibit a clear correlation with salary levels. This suggests that factors beyond academic achievements play a more significant role in determining compensation outcomes among individuals in the dataset.
- After removing outliers, age does not appear to be a determining factor in compensation, indicating that age alone does not significantly influence salary levels among individuals in the dataset.
- In conclusion, while factors such as tenure and college tier demonstrate a strong relationship with salary levels, others like gender and academic performance show little association. These findings underscore the complexity of factors influencing salary outcomes within the dataset and highlight the need for a nuanced understanding of the various determinants of compensation.

Python Notebook Code