INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project - AMCAT Data Analysis

About me

❖I am deeply passionate about data science and its applications in real-world scenarios.

❖Currently interning at Oasis Infobyte, where I am actively involved in projects focused on email spam detection and sales prediction.

❖My internship experiences have provided me with practical exposure to apply my theoretical knowledge effectively.

❖I am driven by the desire to leverage data-driven insights to inform decision-making and drive transformative change.

Business Problem and Use Case domain understanding

- The Analysis of AMCAT data project focuses on exploring and analyzing the AMCAT dataset to gain insights into factors influencing job placements and career trajectories.
- AMCAT (Aspiring Minds Computer Adaptive Test) is an employability assessment test conducted in India to evaluate candidates' skills in various domains, including quantitative aptitude, logical reasoning, English proficiency, and domain-specific knowledge.
- The dataset provides valuable information about candidate’s demographics, educational background, test scores, and job placements.

Objective of the Project

- The objective of this exploratory data analysis (EDA) is to gain insights into the factors influencing the salary of engineering graduates. By examining the relationships between different variables and the target variable (Salary), we aim to understand the dataset's characteristics, identify patterns, and uncover potential insights that could aid in decision-making processes related to employment and salary negotiation for engineering graduates.

Summary of the Dataset

- The dataset used for analysis is the Aspiring Minds Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds.
- It focuses on employment outcomes for engineering graduates and contains a comprehensive set of features related to candidates' demographics, educational qualifications, test scores, and job placements.
- **Key Statistics:**
- **Shape of the Data:** The dataset comprises 3998 rows and 39 columns.
- **Description of the Data:**
  - The 'Salary' column has a mean salary of INR 3,07,699.8 with a standard deviation of INR 2,12,737.5. The minimum salary is INR 35,000, and the maximum salary is INR 40,00,000.
- The '10percentage' column has a mean percentage of 77.93 with a standard deviation of 9.85. The minimum percentage is 43, and the maximum percentage is 97.76.
- The 'CollegeTier' column has a mean value of 1.93, indicating that most colleges are Tier 2 colleges.
- The 'collegeGPA' column has a mean GPA of 71.49 with a standard deviation of 8.17. The minimum GPA is 6.45, and the maximum GPA is 99.93.
- Other columns such as English, Logical, Quant, and Domain have similar statistics.

Summary of the Dataset

- **Data Columns:**
- The dataset contains 39 columns, including:
  - Candidate identifiers (ID, DOB)
  - Employment details (Salary, Designation, JobCity)
  - Educational qualifications (10percentage, 12percentage, collegeGPA)
  - Test scores (English, Logical, Quant)
  - Specializations (Degree, Specialization)
  - Personality traits (conscientiousness, agreeableness, extraversion, neuroticism, openness_to_experience)
- **Data Quality:**
- The dataset appears to be relatively clean, with no missing values in any column.
- The data types include float64, int64, and object (categorical).
- There are no evident anomalies or inconsistencies in the data, but further exploration may reveal hidden patterns or outliers.
- **Key Points:**
- The AMEO dataset provides a rich source of information for analyzing employment outcomes for engineering graduates.
- Its comprehensive nature allows for in-depth exploration of factors influencing salary, job placement, and career trajectories.
- Understanding the data's structure and key statistics is crucial for conducting meaningful analysis and deriving actionable insights for stakeholders.

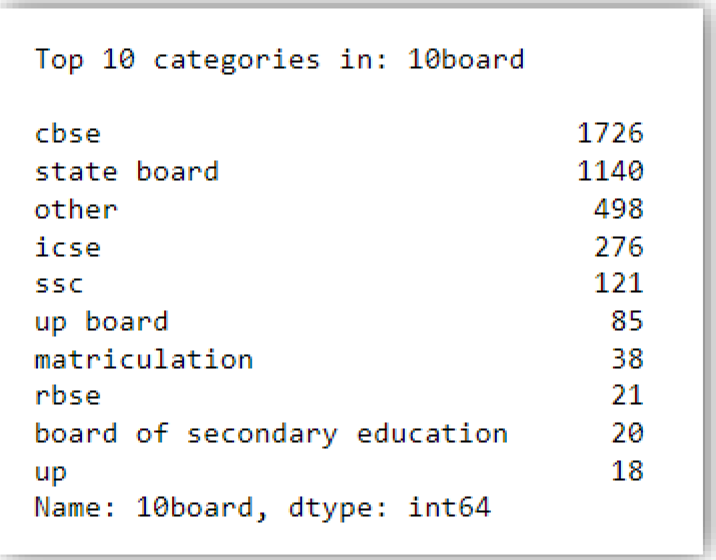
Data Cleaning and Preprocessing

1. **Handling Date Columns:** To ensure accuracy and consistency in our analysis, we transformed the data types of the 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) fields to datetime objects. For respondents who indicated their status as 'present' in the DOL field, we replaced these values with the latest survey date, recorded as 2024-02-17.
2. **Validation of Null Values:** We conducted a thorough examination of columns containing potential null values represented by 0 or -1. The following columns exhibited significant proportions of such values:

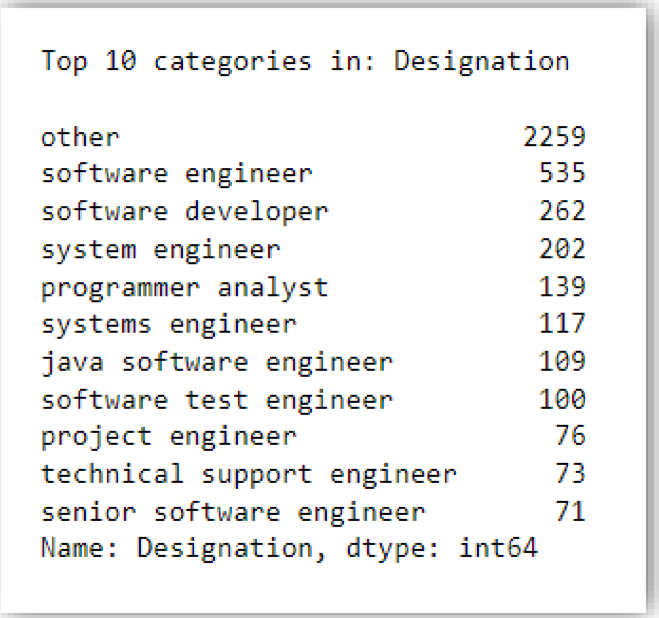
- Electronics & Semicon: 71.39%
- Computer Science: 77.61%
- Mechanical Engg: 94.04%
- Electrical Engg: 96.09%
- Telecom Engg: 90.57%
- Civil Engg: 98.93%
- We proceeded by handling these null values:
- Columns '10board', '12board', 'GraduationYear', 'JobCity', and 'Domain' were processed to replace null values represented by 0 or -1.
- Columns with over 80% -1 values, namely 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', and 'CivilEngg', were removed from further analysis.
- For optional subject columns 'ElectronicsAndSemicon' and 'ComputerScience', -1 values were replaced with 0, indicating that the subjects were not pursued.

Data Cleaning and Preprocessing

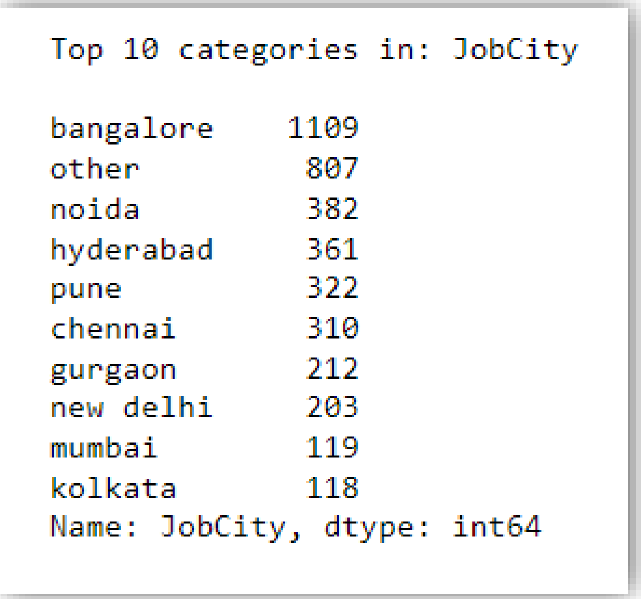
3. **Collapsing Categories**: Through refining, the dataset was reduced to only include the ten most common categories in particular columns. 'other' categories were those that fell outside of this selection. By doing so, the dataset is streamlined and the most common categories are given priority for further examination.



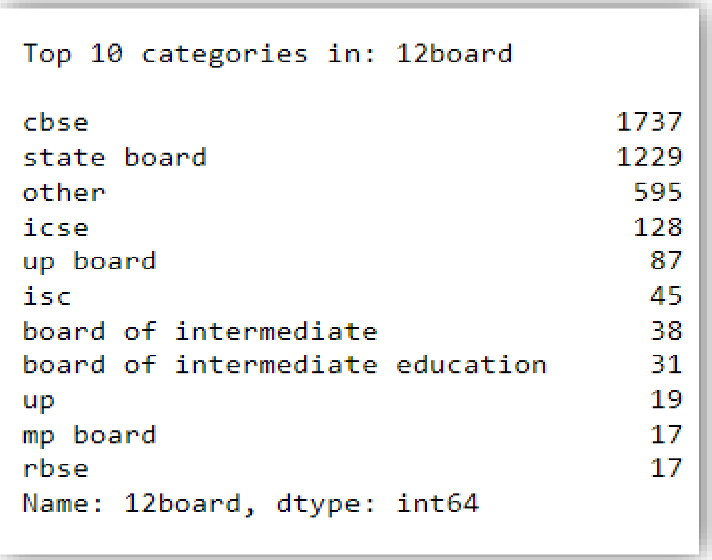
```
Top 10 categories in: 10board

cbse                             1726
state board                      1140
other                             498
icse                              276
ssc                               121
up board                           85
matriculation                      38
rbse                               21
board of secondary education       20
up                                 18
Name: 10board, dtype: int64
```



```
Top 10 categories in: Designation

other                        2259
software engineer             535
software developer            262
system engineer               202
programmer analyst            139
systems engineer              117
java software engineer        109
software test engineer        100
project engineer               76
technical support engineer     73
senior software engineer       71
Name: Designation, dtype: int64
```



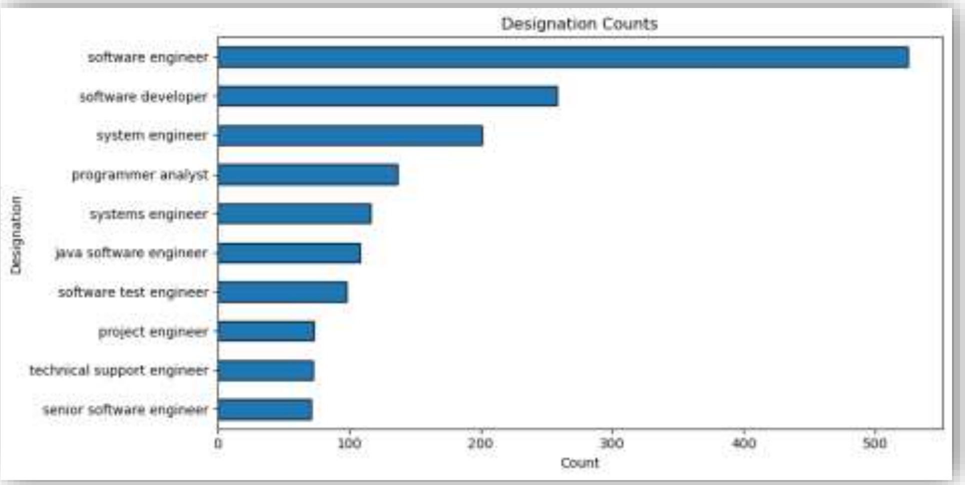
```
Top 10 categories in: JobCity

bangalore    1109
other         807
noida         382
hyderabad     361
pune          322
chennai       310
gurgaon       212
new delhi     203
mumbai        119
kolkata       118
Name: JobCity, dtype: int64
```

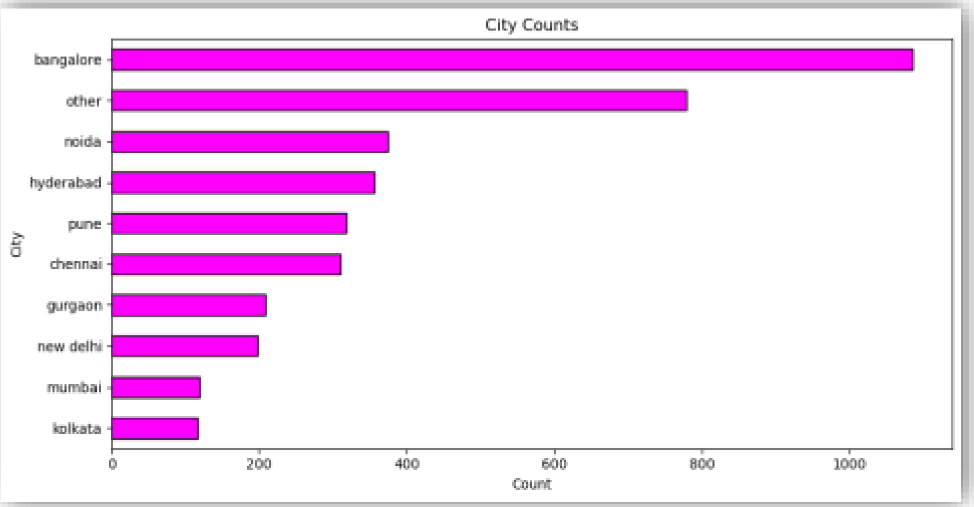


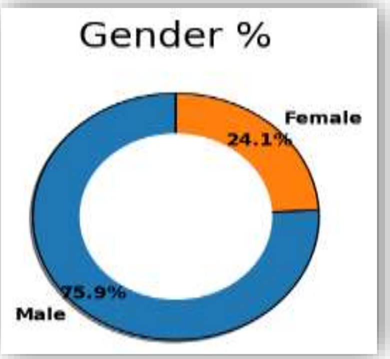
```
Top 10 categories in: 12board

cbse                                 1737
state board                          1229
other                                 595
icse                                  128
up board                               87
isc                                    45
board of intermediate                  38
board of intermediate education        31
up                                     19
mp board                               17
rbse                                   17
Name: 12board, dtype: int64
```

Exploratory Data Analysis:

- ## Univariate Analysis:
- **1) Designation:** The most popular designation is "software engineer," which is followed by "system engineer" and "software developer."
- **2) Job City:** Bangalore is the best city for job placements, followed by Hyderabad, Pune, Noida, and Bangalore. Kolkata and Mumbai are the least favorable.
- **3) Gender:** The male population is actually larger than the female population, hence the dataset is not gender balanced.



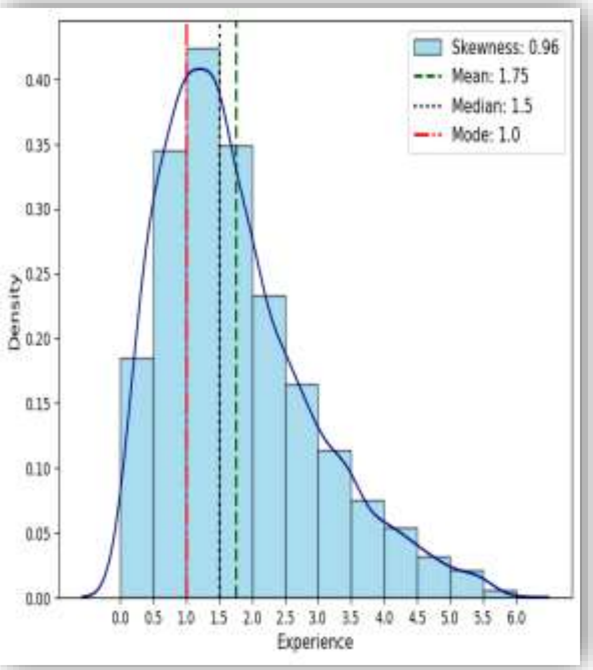


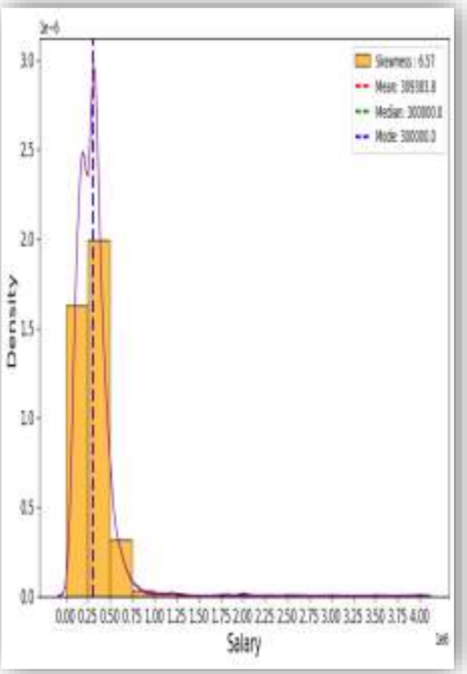


INNOMATICS RESEARCH LABS

Exploratory Data Analysis:

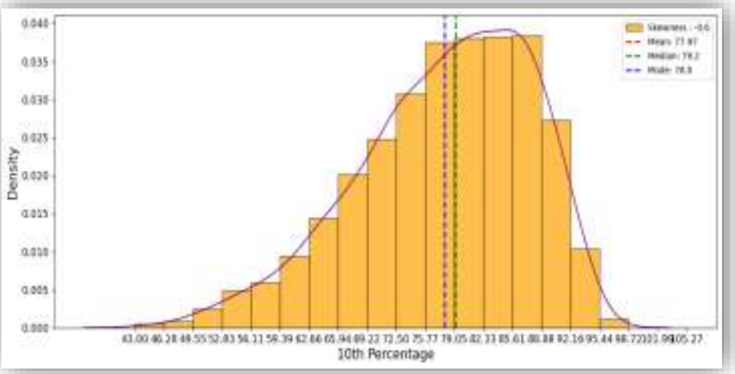
- **Univariate Analysis:**
- **1) Salary:** Significant positive skewness is seen in the histogram, indicating a divergence from the normal distribution, and a significant amount of variation is indicated in the summary plot. A concentration of high incomes is highlighted by box plots. Moreover, the skewness of the data is underlined by the cumulative distribution function (also known as the CDF), which deviates significantly from the normal distribution pattern.
- **2) Tenure:** A 4-year experience range was displayed in summary charts. The histograms showed a 1.5-year median tenure, positively skewed distribution, and outliers denoting lengthier tenures. These outliers were highlighted even further using box charts. Furthermore, the tenure distribution's non-normality was brought to light by the Cumulative Distribution Function (CDF). These results offer insightful information on worker dynamics.

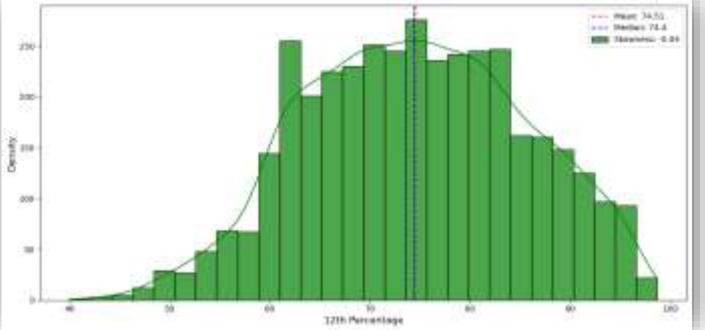




Exploratory Data Analysis:

- **Univariate Analysis:**
- **3) 10th Percentage:** The summary plot shows that about half of the pupils received scores of 80% or less. The histogram shows that there are few pupils with low percentages; most fall between 75% and 90%, with 78% serving as the apex. The box plot clearly shows extreme outliers, which point to certain anomalies in the data distribution. Furthermore, the pattern departing from a normal distribution.
- **4) 12th Percentage:** The examination of the dataset reveals that few students received poor scores, with about half of the students scoring at or below 78%. Most students achieved a maximum score of 70% and a range of 69% to 84%. There is a clear outlier with a very low score. The cumulative distribution function (PDF) indicates that the data does not follow a normal distribution pattern.



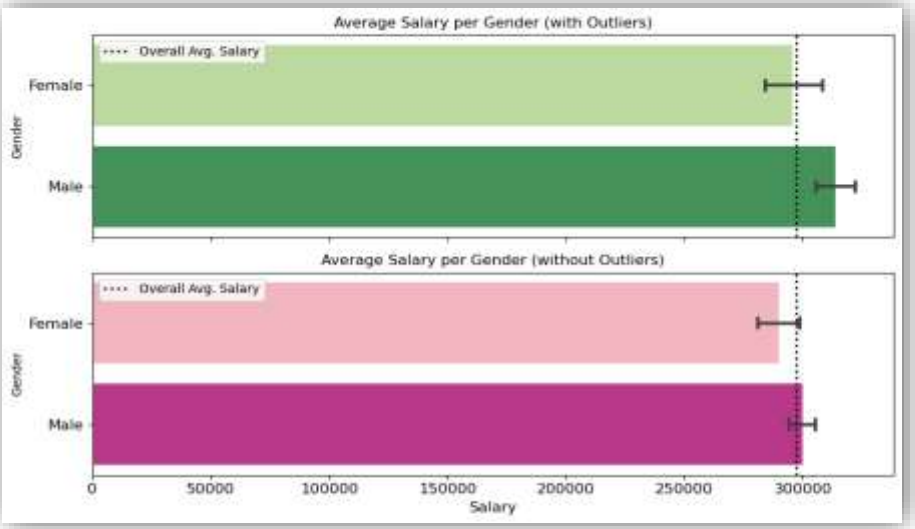


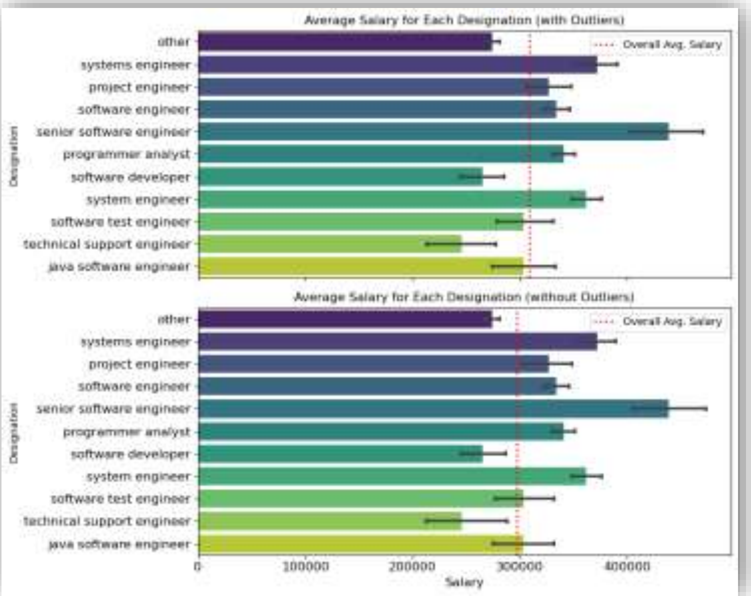
Exploratory Data Analysis:

- **Univariate Analysis:**
- **5) College GPA:** The examination of student GPAs provides insightful information. The summary plot showed that about 75% of students had GPAs of 80% or less.
  The histogram reveals that the majority of students had average GPAs of 74% and GPAs ranging from 63% to 78%, with a high at 70%. The box plot shows that there are both low and high extreme values in the dataset. It's interesting to note that the data appears to be suitably regularly distributed, according to the cumulative distribution function (PDF), which adds to its dependability for more study.
- **6) English, Quants, Logical and others:** Differentiable patterns are revealed by the dataset analysis among different participants. In the English exam, about 50% of the students received a score lower than 500. The majority of scores fell between 389 and 545, with a notable prevalence of high values. Similar to this, a sizable percentage of students on logical tests received scores below 500. These scores were concentrated between 454 and 584, exhibiting both lower and higher extreme values. Most students in Quants scored lower than 600, with scores ranging from 425 to 608, exhibiting a range of extreme values, both high and low.
  In contrast, over 50% of students received a score below 500 in computer programming, with scores primarily ranging from 416 to 459 and a noticeable presence of high values.
  In the fields of electronics and semiconductors, 75% of students received a score of less than 250.

Exploratory Data Analysis:

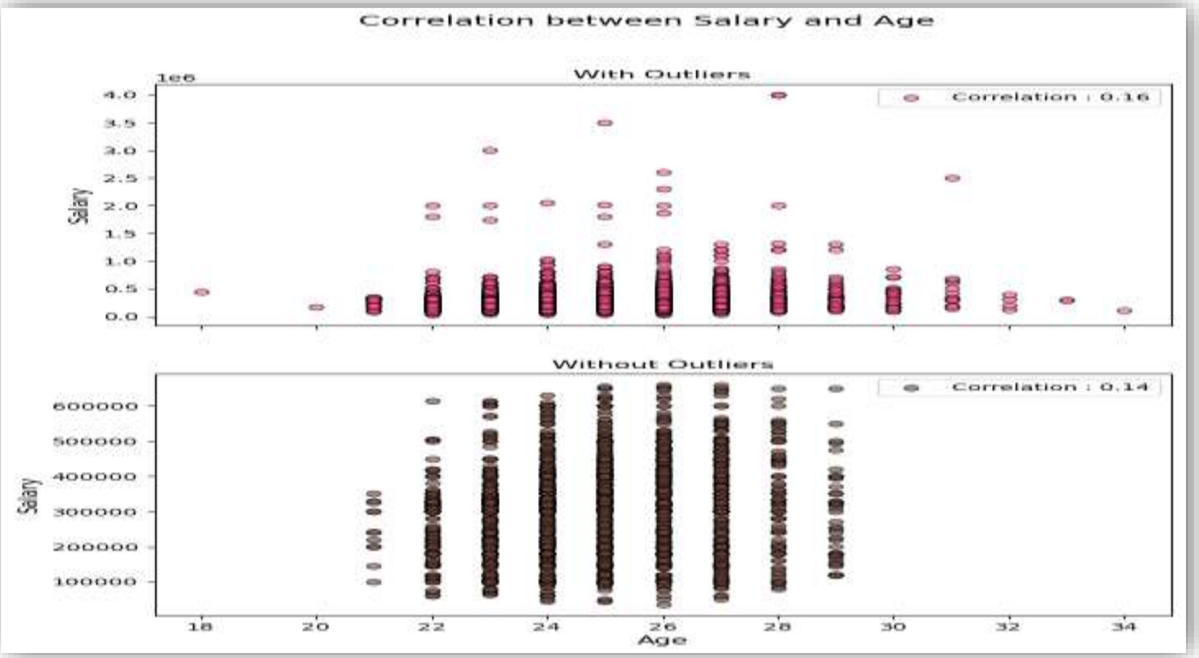
- **Bivariate Analysis:**
- **1) Salary Vs Gender:** On average, the earnings of men and women are roughly equal.
  indicating that there isn't generally any gender bias, even though women are typically paid less than men.
- **2) Salary Vs Designation :** The highest income is earned by senior software engineers, but their standard deviation is also the biggest. The pay for technical support engineers and software developers is below average.

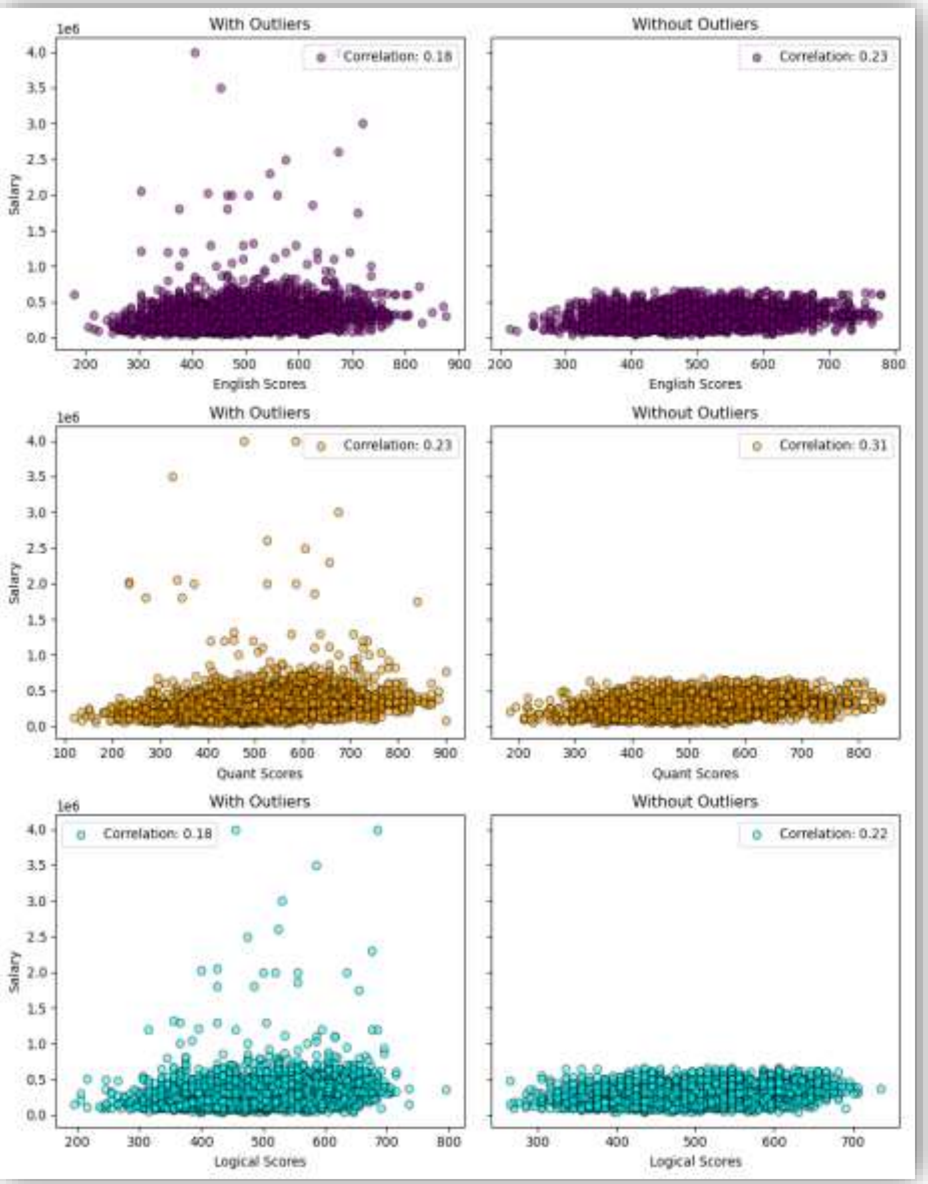




Exploratory Data Analysis:

- **Bivariate Analysis:**
- **3) Salary Vs Age:** There's no apparent relationship between age and salary afterremoving outliers.
- **4) Salary Vs Skills:** There's no apparent effect ofEnglish, Quants, or Logical scores on salary

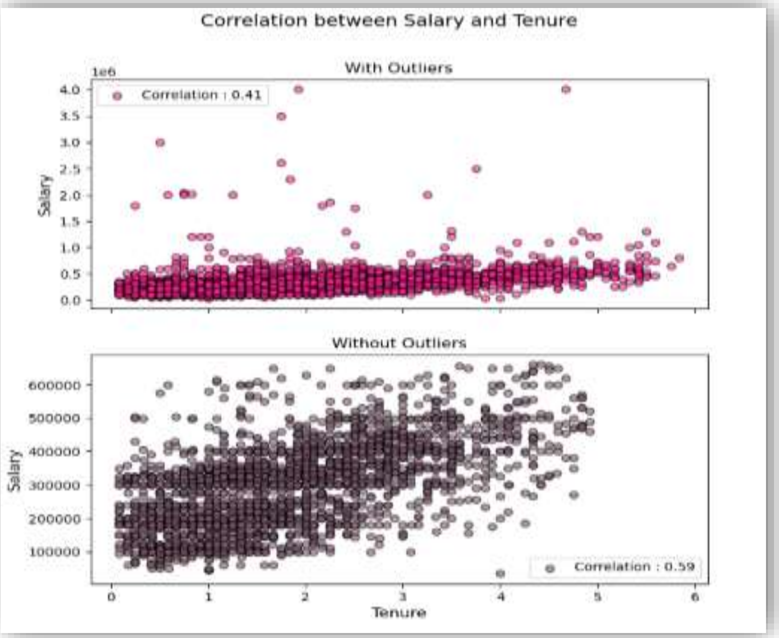


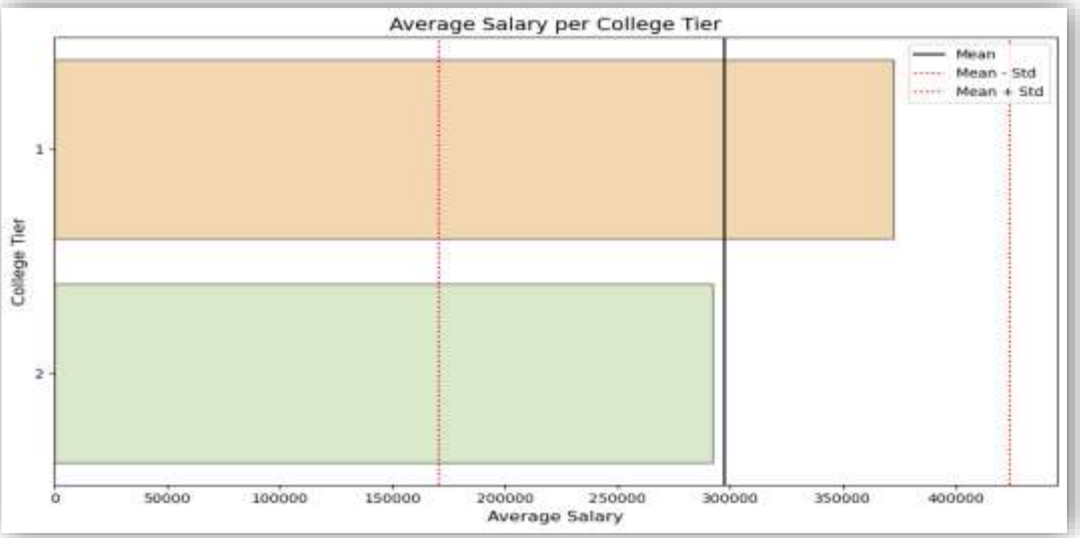


INNOMATICS
RESEARCH LABS

Exploratory Data Analysis:

- **Bivariate Analysis:**
- **5) Salary Vs Tenure:** The relationship between tenure and compensation is positive, and there is a roughly 50% pay boost with tenure, indicating that experience matters.
- **6) Salary Vs College Tier City:** Tier 1 universities pay more than Tier 2 universities, and student earnings in Tier 1 and Tier 2 cities are comparable.

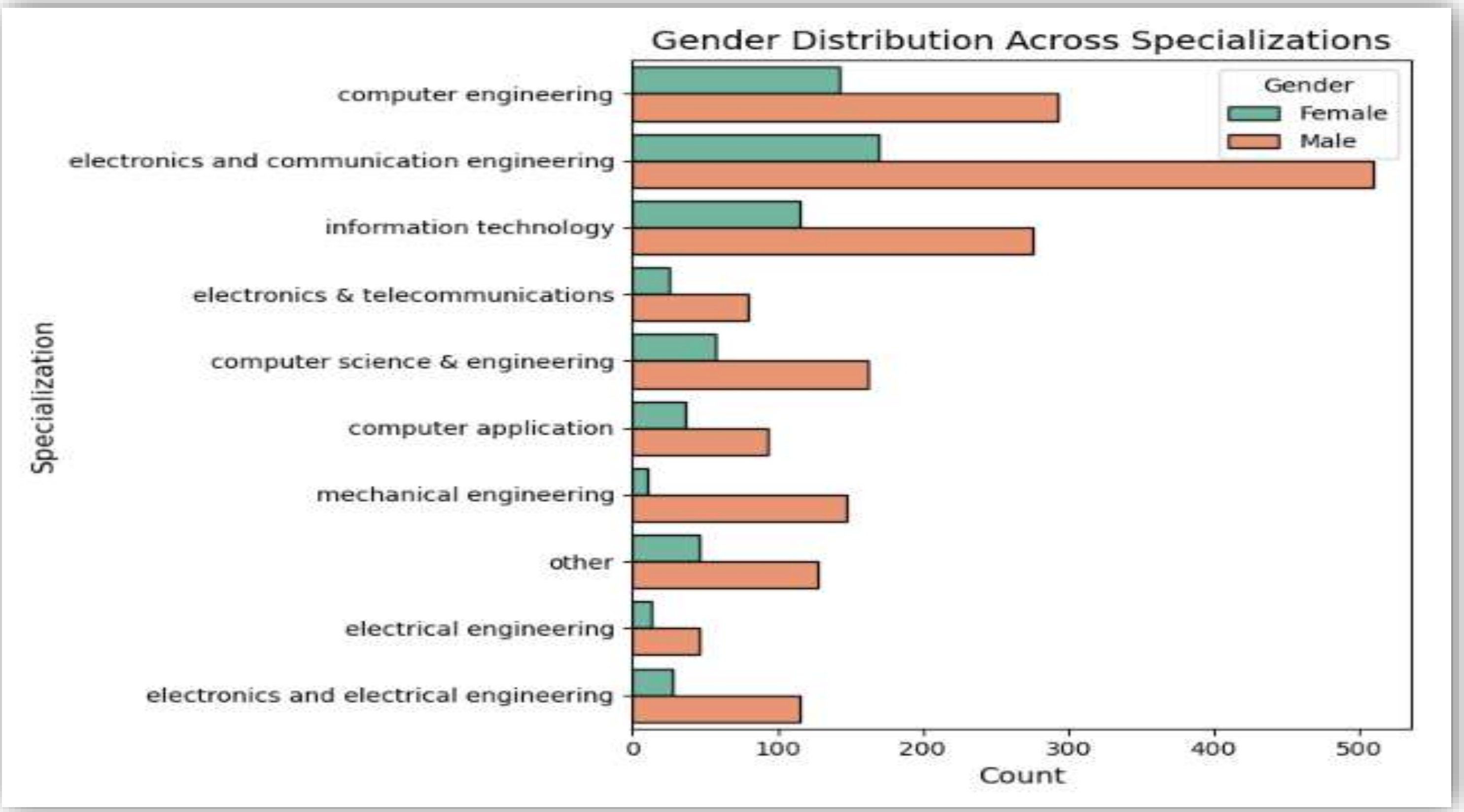




INNOMATICS RESEARCH LABS

Exploratory Data Analysis:

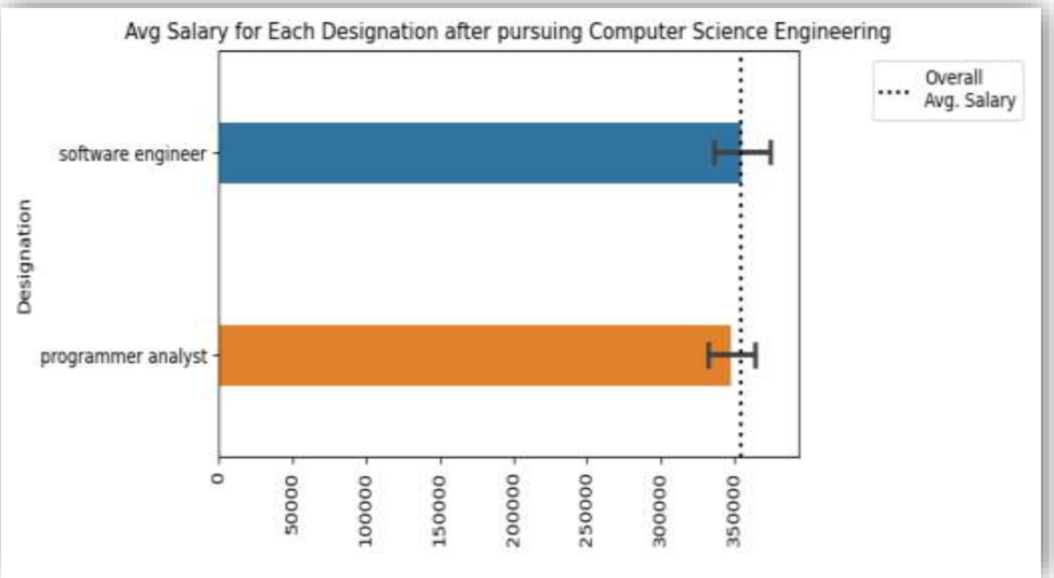
- **Bivariate Analysis:**
- **7) Gender Vs Specialization:** In all specializations, the proportion of males participating is almost twice that of females, with fewer women choosing mechanical and electronics.



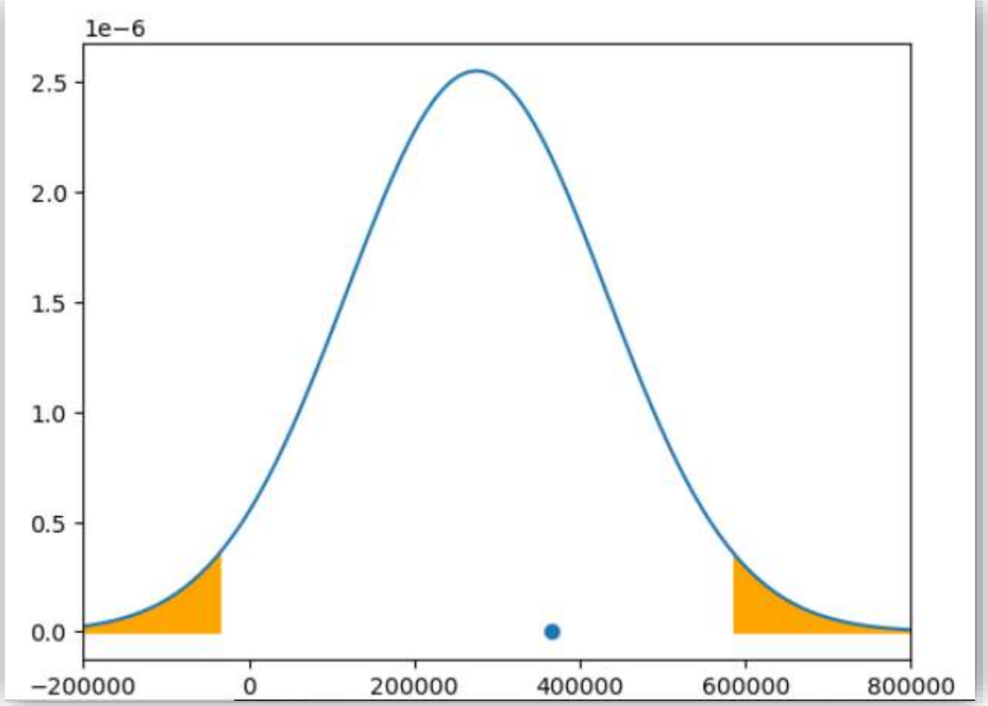
Research Question:

Q1) **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**

- Programmer Analyst and Software Engineer Salaries: Both Programmer Analyst and Software Engineer positions show significantly higher salaries compared to the expected lower bound of the salary range (2.5-3 lakhs). The t-statistics for both positions are high (12.30 for Programmer Analyst and 10.83 for Software Engineer), indicating a substantial difference between the sample means and the expected value. Additionally, the p-values are extremely low (close to 0), providing strong evidence to reject the null hypothesis, suggesting that these positions indeed offer salaries higher than the expected lower bound.

- 

- Hardware Engineer and Associate Engineer Salaries: For Hardware Engineer and Associate Engineer positions, the t-statistics are reported as 'nan' (not a number), and the p-values are also 'nan'. This indicates that there might be insufficient data or variation in the sample salaries for these positions, leading to inconclusive results. Therefore, we cannot confidently reject the null hypothesis for these positions based on the available data.
- General Observation: Overall, the results suggest that Programmer Analyst and Software Engineer positions tend to offer salaries significantly higher than the lower bound of the expected salary range mentioned in the Times of India article. However, the analysis does not provide conclusive evidence regarding Hardware Engineer and Associate Engineer positions due to insufficient data or variability in salaries. Further investigation or data collection may be necessary to draw definitive conclusions for these positions.

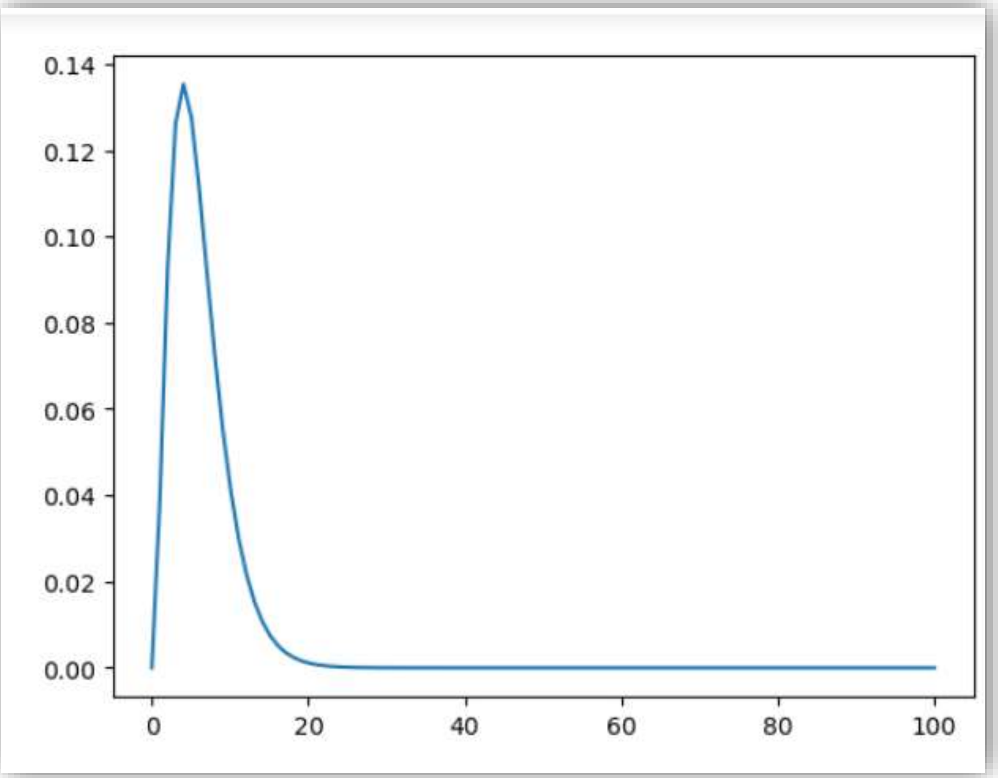


Research Question:

Q2) ***Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)***

- Chi-Square test conducted to analyze gender and specialization preferences relationship.
- Statistically significant relationship found between gender and specialization.
- Chi2 statistic exceeded critical value, p-value significantly less than chosen significance level.
- Null hypothesis rejected, indicating gender and specialization dependence.
- Suggests certain fields may be more preferred or accessible to specific genders.

- Emphasizes importance of gender diversity and inclusivity in various fields.
- Highlights potential barriers or biases in certain specializations.



Conclusion (Key Finding Overall):

- After conducting extensive exploratory data analysis and hypothesis testing, several key findings have emerged from this project:
- **Salary Distribution Across Job Designations**: The analysis revealed significant variations in salaries among different job roles. Software Engineers emerged with the highest mean salary and standard deviation, indicating both higher earnings and variability compared to other designations.
- **Gender and Specialization Preferences**: A Chi-Square test uncovered a statistically significant relationship between gender and specialization preferences. This finding suggests that certain fields may be more preferred or accessible to individuals of particular genders, highlighting the importance of gender diversity and inclusivity in various fields.
- **Relationship Between Tenure and Salary**: After removing outliers, a positive correlation of 0.60 was observed between tenure and salary, indicating that compensation increases by approximately 50% with tenure.
- **Impact of Academic Performance on Salary**: Univariate analysis of academic performance metrics such as 10th percentage, 12th percentage, and college GPA showed no direct correlation with salary, suggesting that other factors play a more significant role in determining earnings.
- **Job City and Salary**: Bangalore emerged as the top city for job placements, with higher average salaries compared to other cities. This finding underscores the importance of geographical location in salary negotiations and career advancement.
- **Effect of Gender on Salary**: Contrary to initial assumptions, a Mann-Whitney U test revealed a significant difference between male and female salaries, with women earning slightly less on average. This highlights the need for further investigation into potential gender-based pay disparities.

THANK YOU

INNOMATICS RESEARCH LABS