INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

Himanshu Agarwal

19th February 2023

About Me

I am a proactive, responsible, and results-oriented professional currently pursuing a bachelor's degree in computer engineering. My interests lie in solving technical issues, conducting research, and innovating new technologies. I thrive in team environments and enjoy connecting with new individuals. With a kind and outgoing nature, I am also a quick learner. Moreover, I excel in working under pressure and possess excellent stress management skills.

I am keen to dive more into the field of Data Science due to its transformative potential across various industries. Firstly, I am intrigued by the prospect of extracting valuable insights from vast amounts of data, which can drive informed decision-making and innovation. Data Science offers a powerful toolkit to uncover patterns, trends, and correlations that can significantly impact businesses and society.

At TCET - Open Source, I've held key roles driving organizational success. As Co-Founder & CEO, I lead project development, manage open-source internships, resolve conflicts, and foster student engagement. In my previous stint as Documentation Team Lead, I provided strategic direction, nurtured team growth, and improved knowledge accessibility.

Link to Project Repo: [EDA on AMEO Dataset]

I. OBJECTIVE OF THE PROJECT

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.

II. SUMMARY OF DATA

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses on employment outcomes for engineering graduates. It includes dependent variables such as **Salary**, **Job Titles**, and **Job Locations**, along with standardized scores in **cognitive skills**, **technical skills**, and **personality skills**. With around 40 independent variables and 4000 data points, these variables encompass both continuous and categorical data. The dataset also includes demographic features and unique identifiers for each candidate.

III. DATA CLEANING AND PREPROCESSING

A. Datatype Conversion

To ensure the accuracy and consistency of our analysis, we converted the data types of the 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) fields from their original format to datetime objects. Given that the survey was conducted in 2015, the assumption was made that respondents who indicated their status as 'present' for DOL had left the company by the latest survey date, which was recorded as 2024-02-17. Therefore, we replaced the 'present' values in the DOL field with this end date.

B. Validating 0 or -1

Sr. No.	Column Name	Null Score
1	Electronics&Semicon	71.392341
2	ComputerScience	77.605884
3	MechanicalEngg	94.040071
4	ElectricalEngg	96.094344
5	TelecomEngg	90.565559
6	CivilEngg	98.934821

The data handling process has been successfully completed. Firstly, the columns '**10board**', '**12board**', '**GraduationYear**', '**JobCity**', and '**Domain**' were processed to replace null values represented by 0 or -1.

Following this, columns with over 80% -1 values, including '**MechanicalEngg**', '**ElectricalEngg**', '**TelecomEngg**', and '**CivilEngg**', were removed from further analysis.

Lastly, for the remaining optional subject columns, '**ElectronicsAndSemicon**' and '**ComputerScience**', -1 values were replaced with 0, indicating that the subjects were not pursued.

C. Collapsing Categories

Under this process, the dataset has been refined to encompass solely the top 10 most frequent categories within specific columns. Any categories beyond this selection have been categorized as other. This approach aims to streamline the dataset, focusing solely on the most prevalent categories for subsequent analysis.

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.

4. Cumulative Distribution Function (CDF) Function:

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.

```
def cdf(data):
    x = np.sort(data)
    y = np.arange(1,
len(x)+1)/len(x)
    return x, y
```

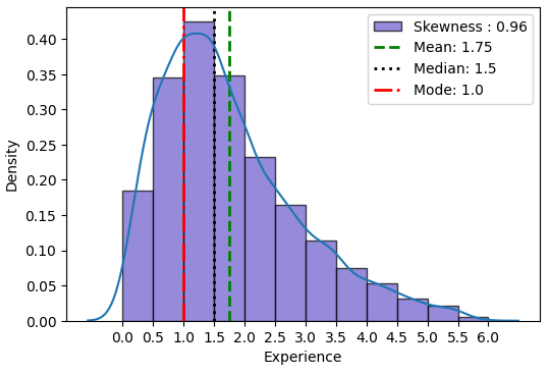
V. EXPLORATORY DATA ANALYSIS

A. Univariate analysis

1. Continuous Features:

1.1. Tenure

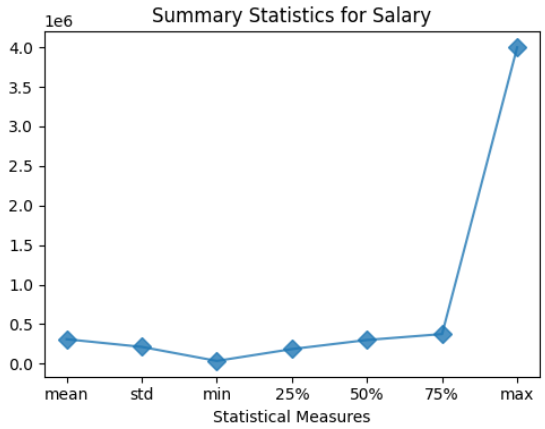
Summary plots showcased a 4-year experience range. Histograms displayed a positively skewed distribution with a median tenure of 1.5 years and outliers signifying longer tenures. Box plots further emphasized these outliers. Additionally, the Cumulative Distribution Function (CDF) highlighted the non-normal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



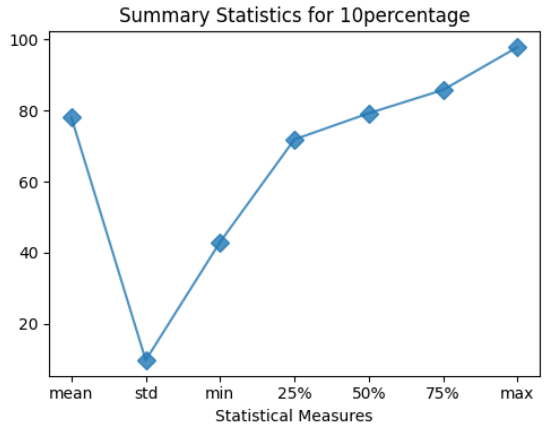
1.2 Salary

The summary plot indicates considerable variation, while the histogram reveals significant positive skewness, suggesting departure from

normal distribution. Box plots emphasize a concentration of high salaries. Furthermore, the cumulative distribution function (CDF) underscores the data's skewness, deviating notably from a normal distribution pattern.



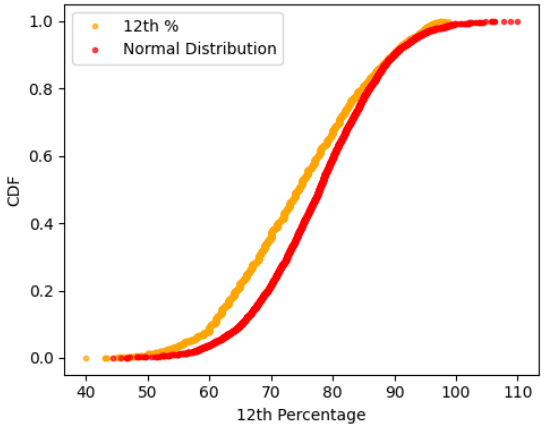
1.3 10th Percentage



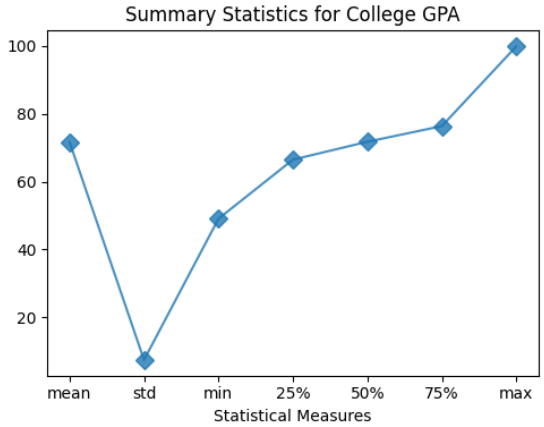
Around half of the students achieved scores of approximately 80% or lower, as depicted by the summary plot. The histogram illustrates a scarcity of students with low percentages, with the majority falling within the 75% to 90% range, peaking at 78%. Extreme outliers are evident from the box plot, indicating some irregularities in the data distribution. Moreover, the cumulative distribution function (CDF) highlights skewness in the data, deviating from a normal distribution pattern.

1.4 12th Percentage

The dataset analysis shows that approximately half of the students scored around 78% or lower, with a scarcity of low scores. Most students scored between 69% and 84%, peaking at 70%. An outlier with an extremely low score is evident. The data deviates from a normal distribution pattern, as shown by the cumulative distribution function (CDF).



1.5 College GPA

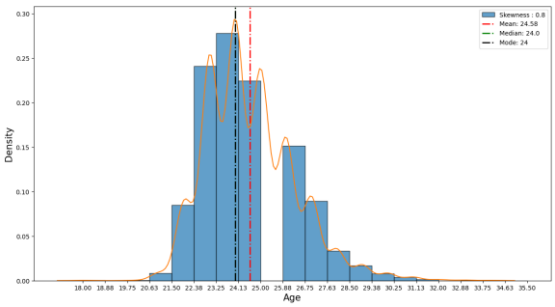


The analysis of student GPAs yields valuable insights. Approximately 75% of students had GPAs around 80% or lower, according to the summary plot. The histogram shows that most students

had GPAs between 63% and 78%, peaking at 70%, with an average GPA of 74%. Both low and high extreme values are apparent in the dataset, as indicated by the box plot. Interestingly, the cumulative distribution function (CDF) suggests that the data is sufficiently normally distributed, contributing to its reliability for further analysis.

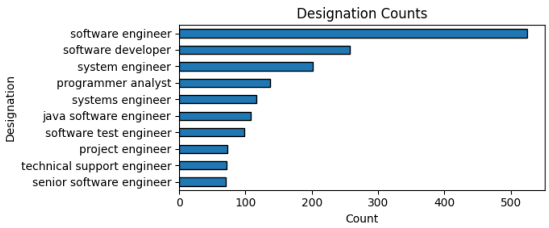
1.6 English, Logical, Quant, Computer Programming, Electronics & Semiconductors, Age

The dataset analysis across various subjects reveals distinct patterns. In English exams, approximately half of the students scored below 500, with scores predominantly ranging from 389 to 545 and a noticeable presence of extreme values. Similarly, in Logical exams, a significant portion of students scored below 500, with scores concentrated between 454 to 584, displaying both lower and higher extreme values. In Quants, a majority of students scored below 600, with scores spanning from 425 to 608, showing a mix of low and high extreme values. Conversely, in Computer Programming, around 50% of students scored below 500, with scores clustering between 416 to 459 and a notable presence of extreme values. Electronics and Semiconductors saw about 75% of students scoring less than 250, with scores mainly falling between 0 to 79 and a non-normal distribution evident. Lastly, the age distribution indicates that approximately 75% of students are under 26 years old, with the majority aged between 22 to 25 and notable outliers at both ends of the age spectrum.

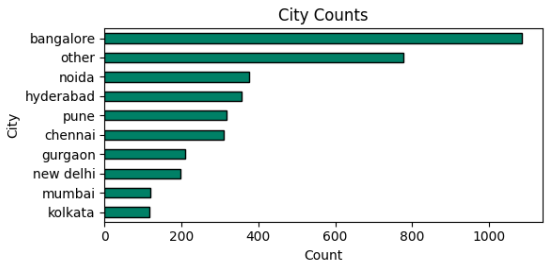


2. Categorical Features:

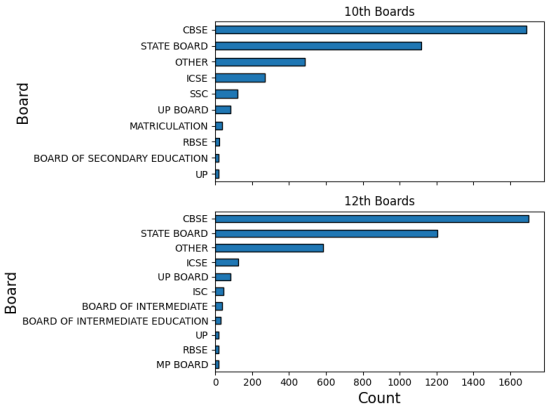
The dataset observations across various categories provide valuable insights into the demographics and educational backgrounds of individuals. In terms of Designation, Software Engineer emerges as the most prevalent designation, followed by System Engineer and Software Developer, with an "OTHER" category also present.



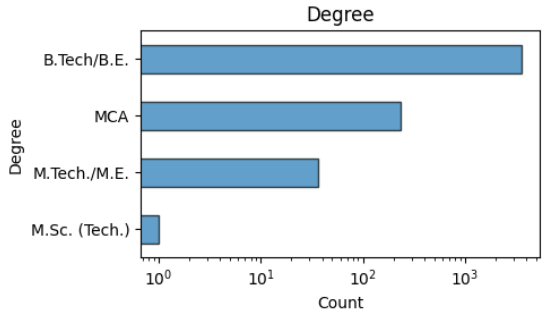
Regarding Job City preferences, Bangalore stands out as the most favorable city for job placements, followed by Noida, Hyderabad, and Pune, while Mumbai and Kolkata are less preferred.



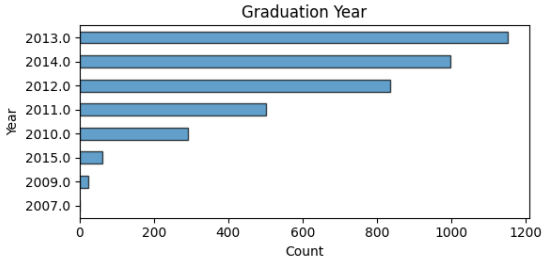
Gender distribution highlights an imbalance, with a significantly larger male population compared to females.



CBSE emerges as the most common school board for both 10th and 12th grades, while College Tier analysis indicates a dominance of Tier 1 colleges. Most students have pursued a B.Tech degree, with minimal representation from M.Sc(Tech) graduates.

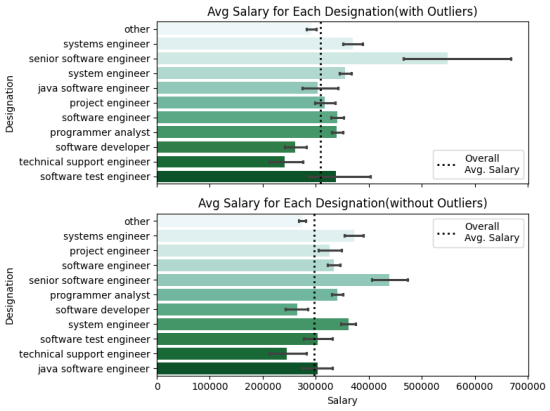


The majority of colleges are located in Tier 0 cities, and 2013 saw the highest number of graduations, followed by 2014 and 2012. These observations collectively offer valuable insights into the educational and professional landscape captured by the dataset.



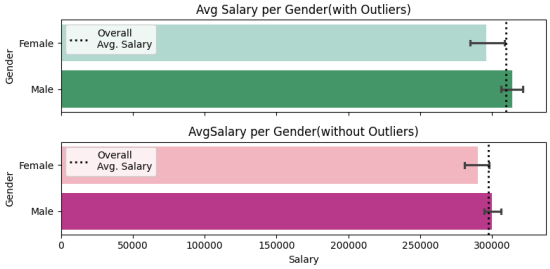
B. Bivariate analysis

1. Designations & Salary



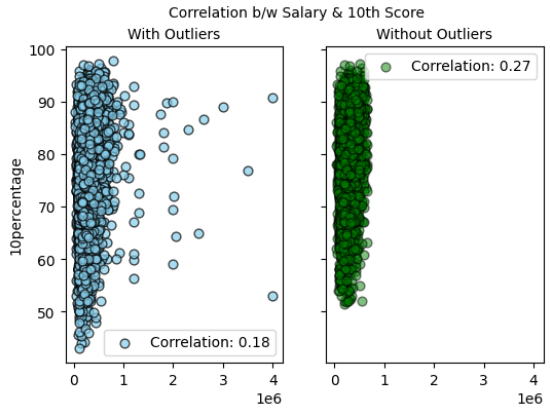
Senior Software Engineers have the highest salary, but also the highest standard deviation. Software Developers and Technical Support Engineers have salaries below the average.

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.

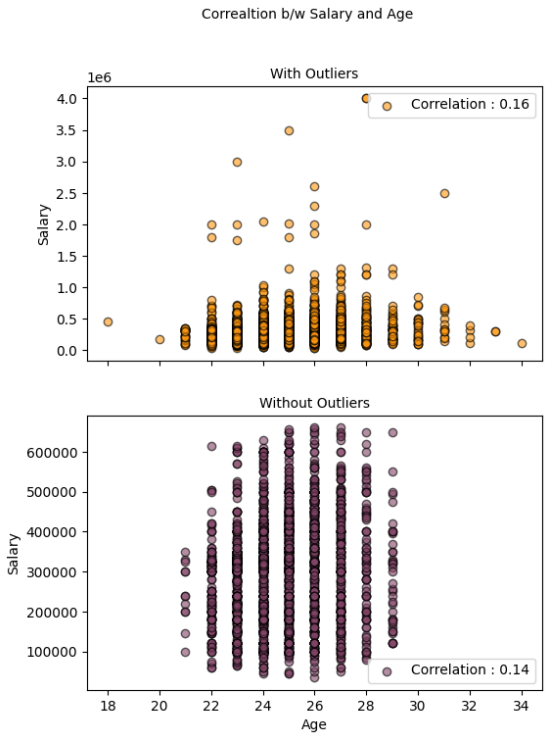
3. Academic Scores & Salary



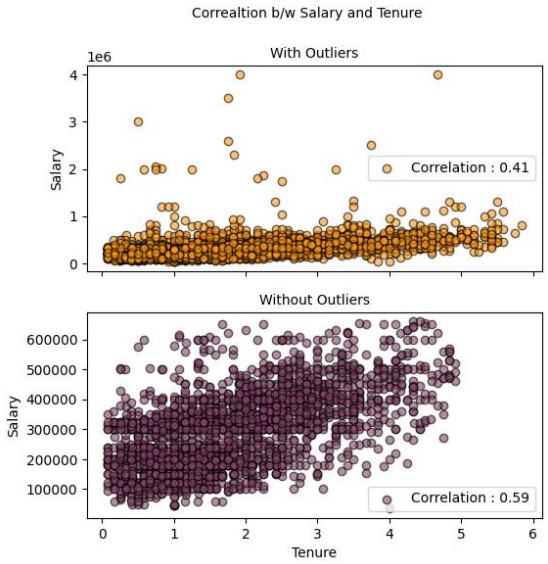
There is no significant correlation between salary and scores in 10th, 12th, or College GPA.

4. Age & Salary

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



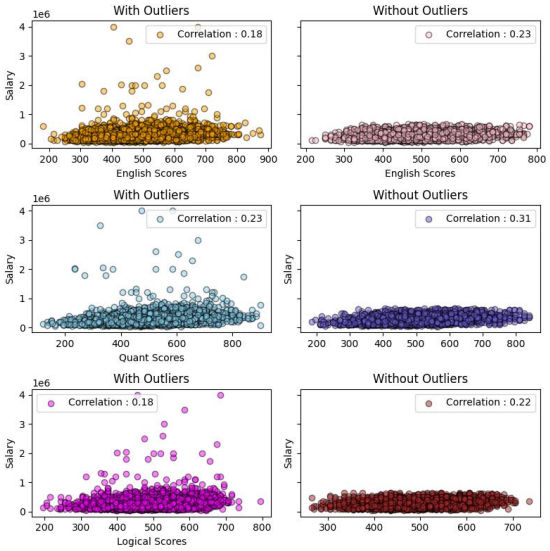
5. Tenure & Salary



There's a positive correlation between tenure and salary, with approximately a 50% salary increase with tenure, suggesting experience plays a role.

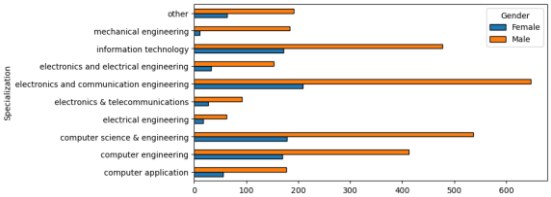
6. Skills & Salary

There's no apparent effect of English, Quants, or Logical scores on salary.



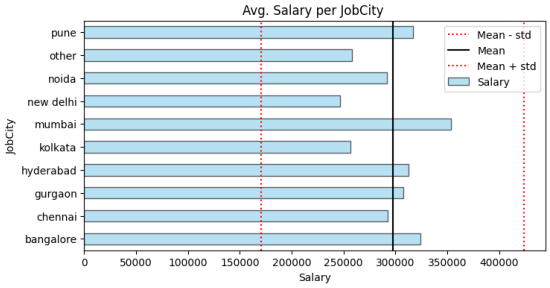
7. Gender & Specialization

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



8. College Factors & Salary

Tier 1 colleges offer higher salaries compared to Tier 2 colleges, and cities in Tier 1 and Tier 2 offer similar salaries to students.



In conclusion, while certain factors like tenure and college tier influence salary, others such as gender and academic scores show little to no correlation. Outliers in age were removed, suggesting age alone doesn't dictate salary. These observations imply a complex interplay of factors influencing salaries in the studied context.

VI. RESEARCH OUTCOMES

"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."

Designation	t	p	Result
Programmer Analyst	12.77	2.20e-18	Reject Null Hypothesis
Software Engineer	10.21	5.82e-21	Reject Null Hypothesis
Hardware Engineer	NaN	NaN	Not Enough Evidence
Associate Engineer	0.61	0.30	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.

Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation 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 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.

VII. CONCLUSION

The extensive data analysis yields several notable discoveries about the factors impacting pay levels in the dataset. While certain criteria, such as tenure and college level, have a strong link with compensation, others, such as gender and academic performance, have no relationship.

Senior Software Engineers demand the greatest incomes, but with greater unpredictability, while Software Developers and Technical Support Engineers make less than the average. Gender does not appear to play a large impact in income determination on average, yet females do receive less than the total average salary. Academic performance, as measured by 10th, 12th, and college GPA scores, does not clearly correlate with pay levels. After removing outliers, age does not appear to be a determining factor in compensation.