INNOMATİCS
RESEARCH LABS

# PROJECT ON

## Exploratory Data Analysis on AMEO Dataset

-Amit Dhamale

# About me

I am Dhamale Amit Madhukar, pursuing my B-Tech in Mechanical Engineering from IIT BHU, Varanasi. I am really enthusiastic in field of Data Science and it’s applications not only in corporate market but also in day to day life.

In this project, my role was to perform exploratory data analysis including univariate and bivariate analysis on the dataset that was released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO).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.

Current wealth of any profession as I believe is data. I'm drawn to the field of data science because of its incredible capacity to extract meaningful insights from data, guiding well-informed decision-making processes. In today's data-centric landscape, mastering data science is highly desirable, presenting avenues to tackle intricate challenges and effect positive change. What captivates me most is the interdisciplinary essence of data science, melding together statistics, computer science, and specialized domain knowledge. Above all, I'm enthusiastic about the boundless opportunities to innovate and generate substantial value through the art of data analysis.

# Objective of the Project

- Perform an Exploratory Data Analysis (EDA) on the AMEO dataset.
- Describe the dataset comprehensively, including its features and attributes.
- Perform Univariate and Bivariate analysis for gaining insights.
- Identify patterns and trends present in the data.
- Explore relationships between independent variables and the target variable (Salary).
- Detect outliers or anomalies within the dataset.
- Gain insights into the dataset, focusing on understanding the relationship between various features and the target variable.
- Use insights to make informed decisions and drive innovation based on data patterns, trends, and correlations.

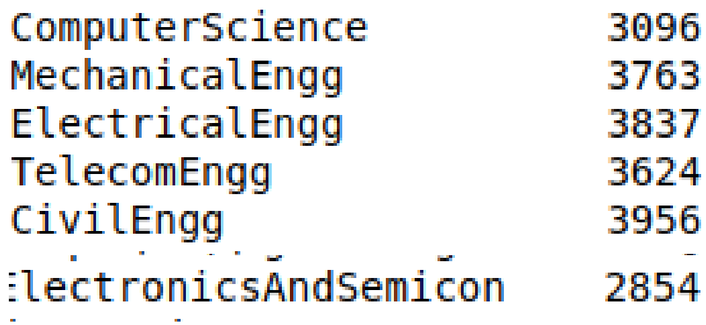
# Summary of Data

The Aspiring Minds Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses exclusively on the employment outcomes of engineering graduates, capturing crucial variables such as salaries, job titles, and locations. It not only includes demographic details but also showcases standardized scores across cognitive abilities, technical skills, and personality traits. With about 40 independent variables, both continuous and categorical, and 4000 data points, the dataset provides a rich basis for analyzing the professional landscape faced by these individuals.

This dataset is an invaluable tool for stakeholders in education and employment, offering insights that could bridge the gap between engineering education and market needs. It facilitates a deep dive into how various factors influence career outcomes, supporting research aimed at enhancing employability and informing policy and curriculum development. Each record includes a unique identifier, ensuring privacy while allowing detailed study, making the AMEO dataset pivotal for improving the alignment of engineering education with industry demands.

# Analysis of Raw data, it’s Cleaning and Manipulation

- Columns ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg , CivilEngg have high number of null values(in the form of -1) therefore they are removed.



| | |
|---|---|
| ComputerScience | 3096 |
| MechanicalEngg | 3763 |
| ElectricalEngg | 3837 |
| TelecomEngg | 3624 |
| CivilEngg | 3956 |
| ElectronicsAndSemicon | 2854 |

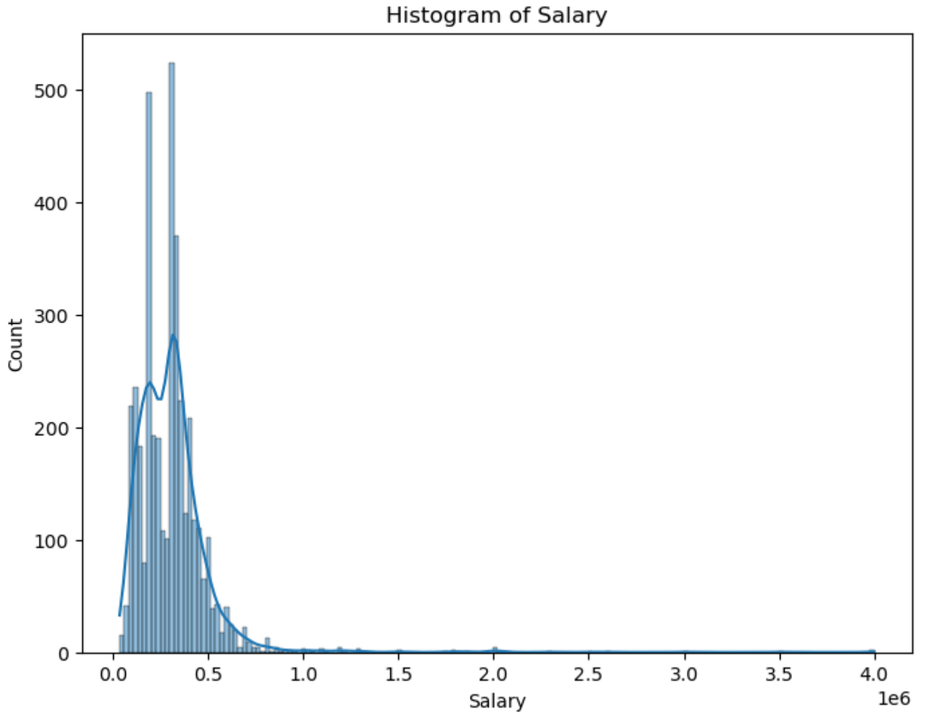
- Given Date of Joining(DOJ) and Date of Leaving(DOL)[replacing it with current date if it is present), we can include an extra column that may effect Salary, i.e. Tenure of an Employee(in years)
- Given Date of Birth we can include an extra column of Age as a variable correlating with Salary.
- Remove the Outliers from numerical columns Salary, English Score, Quant Score, Logical Score, 12th Percentage and 10th Percentage, using IQR Method.
- For Categorical columns like JobCity and Designation, carry out Data Analysis for top 10 Largest w.r.t count/frequency.

# Exploratory Data Analysis

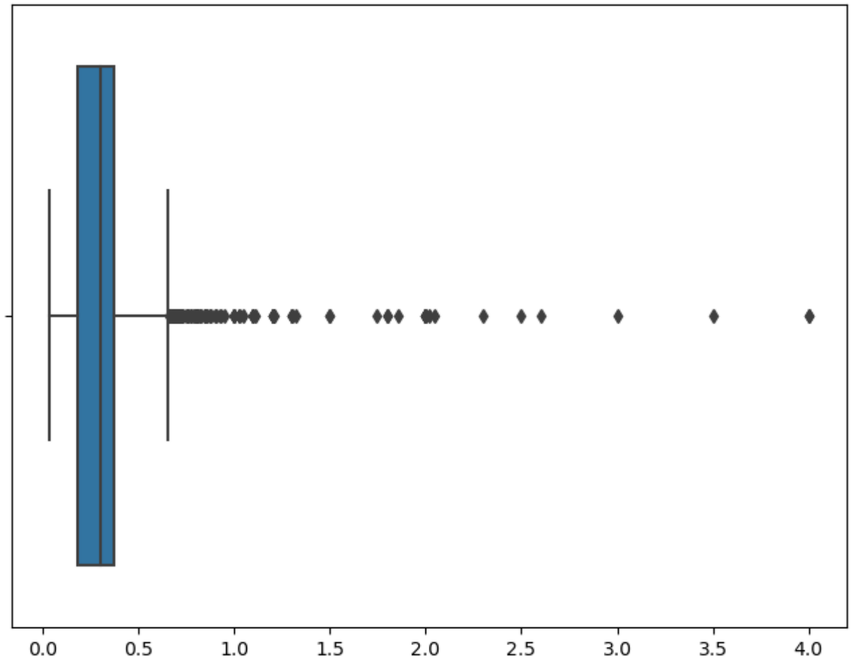
## Univariate Analysis:

## Salary Analysis:

The histogram displaying salary distribution clearly shows that the highest salary offered is 40 LPA (Lakhs per Annum), while the lowest salary hovers around 35K (Thousand). It's evident from the data that the majority of individuals receive a salary within the 2.5 LPA to 5 LPA range, indicating a common salary bracket for most employees captured in this analysis.



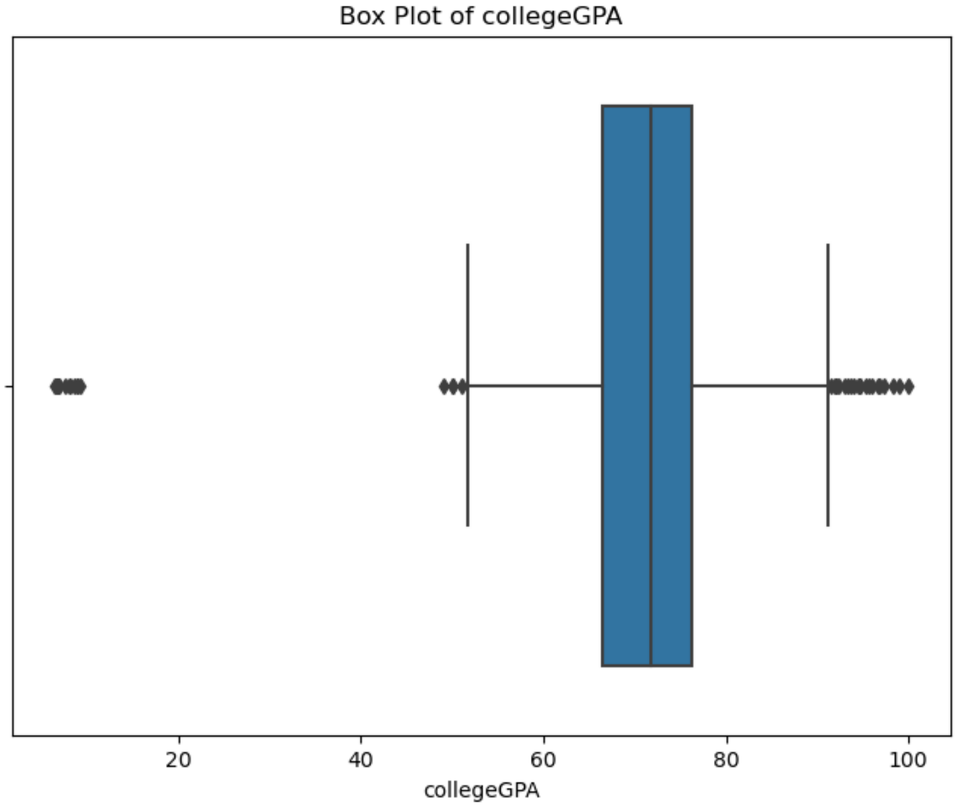

The box plot is describing the statistical aspect of data on salary:
Mean salary: 3,07,699
Min salary: 35,000
Max salary: 40,00,000

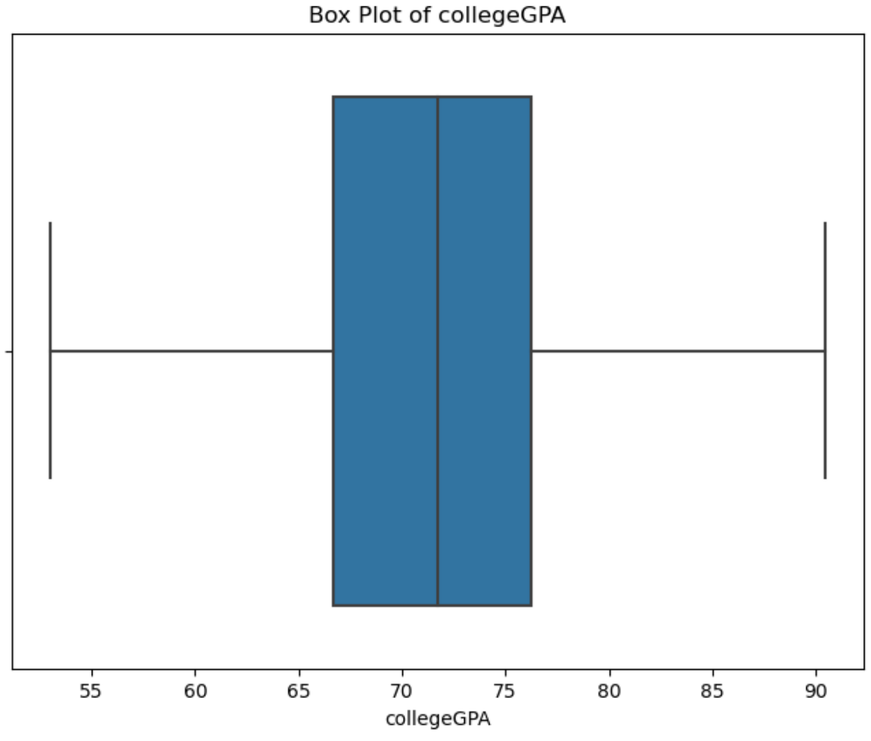



Removing the outliers using IQR Method will give the range as 35KPA-6.3LPA

# College GPA Analysis:

The box plot reveals that the majority of employees' college GPAs fall within the 60-80 range. Notably, there are no employees with a college GPA between 20 and 40. Despite this, a small number of outliers exist, with some employees having a GPA below 20. These outliers can be effectively removed using the Interquartile Range (IQR) method to ensure a more accurate analysis.



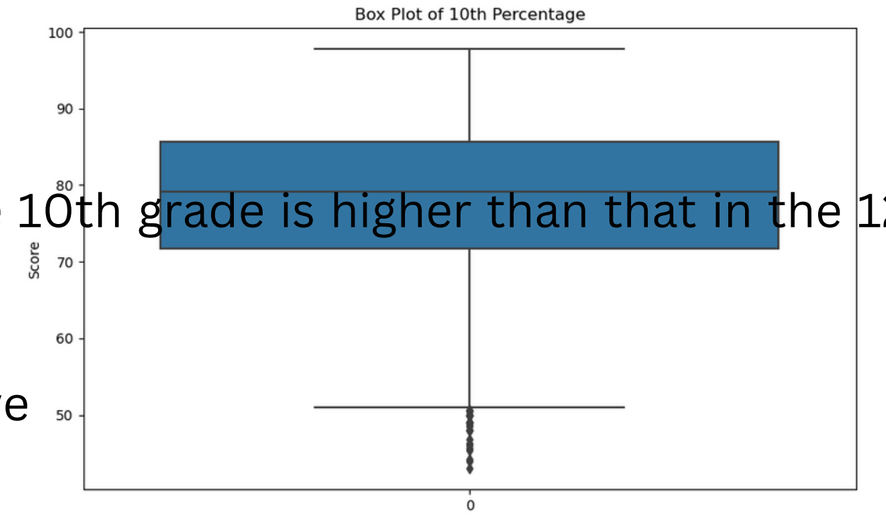

| | |
|---|---|
| count | 3664.000000 |
| mean | 71.604192 |
| std | 7.045845 |
| min | 53.000000 |
| 25% | 66.660000 |
| 50% | 71.730000 |
| 75% | 76.252500 |
| max | 90.440000 |

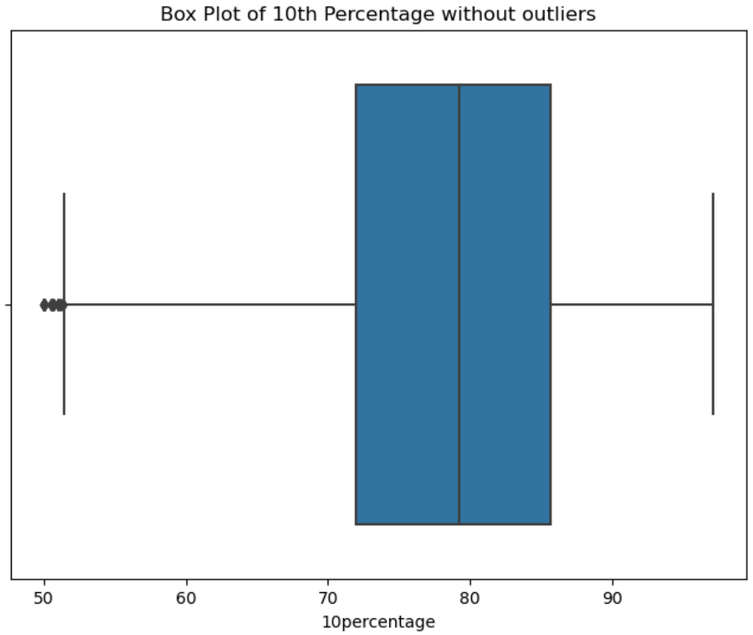
# 10th and 12th percentage Analysis:

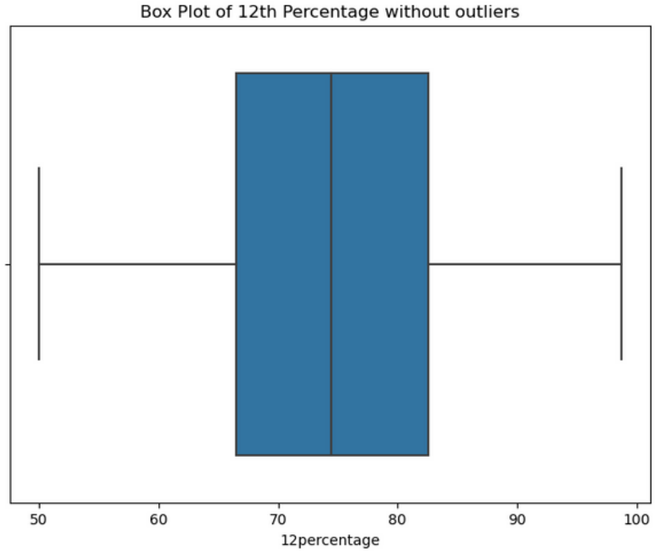
Shown in left Box plots are 10th and 12th percentage distribution:

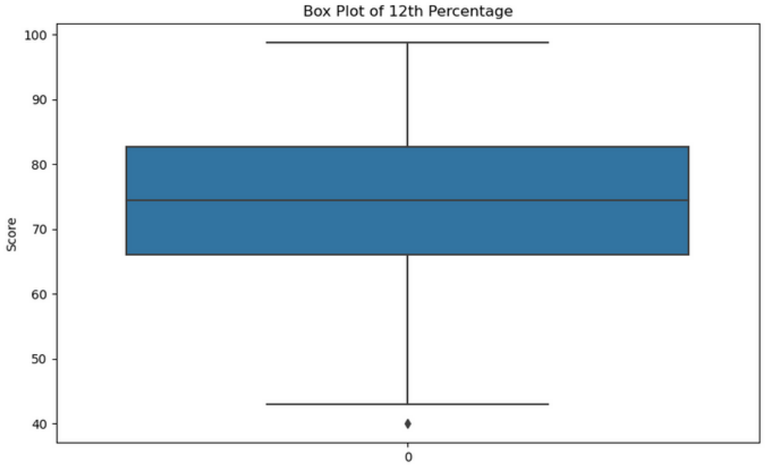
The boxplots indicate that the average percentage obtained in the 10th grade is higher than that in the 12th grade.

There are considerable amount of outliers present, removing will give following box plots of distributions:



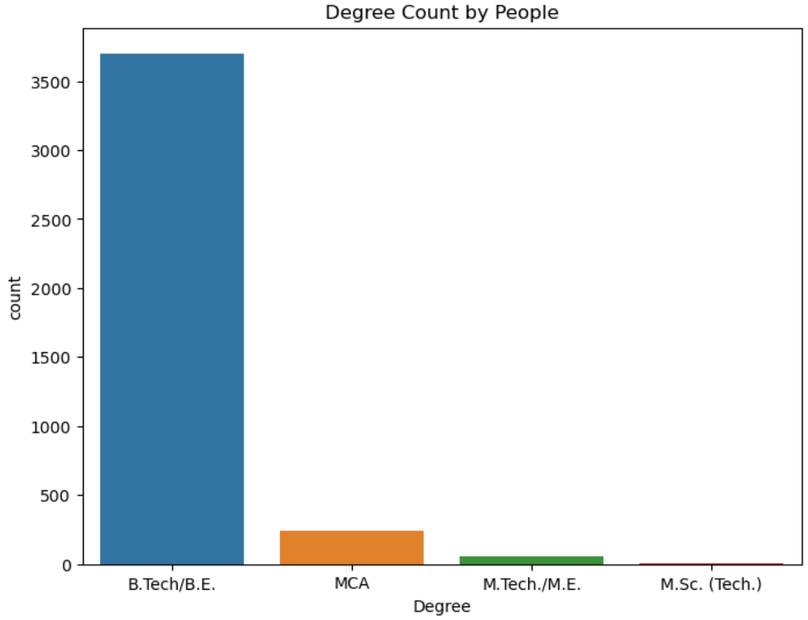



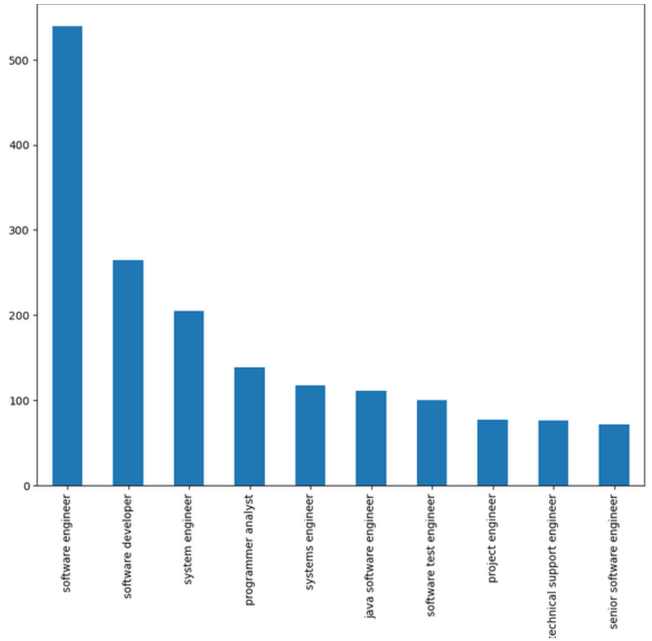
# Degrees vs there count

It is clear that most of the people had pursued the degree of B.Tech approximately around 3500 then the second most preferable degree is MCA ,followed by M.Tech and then M.Sc.



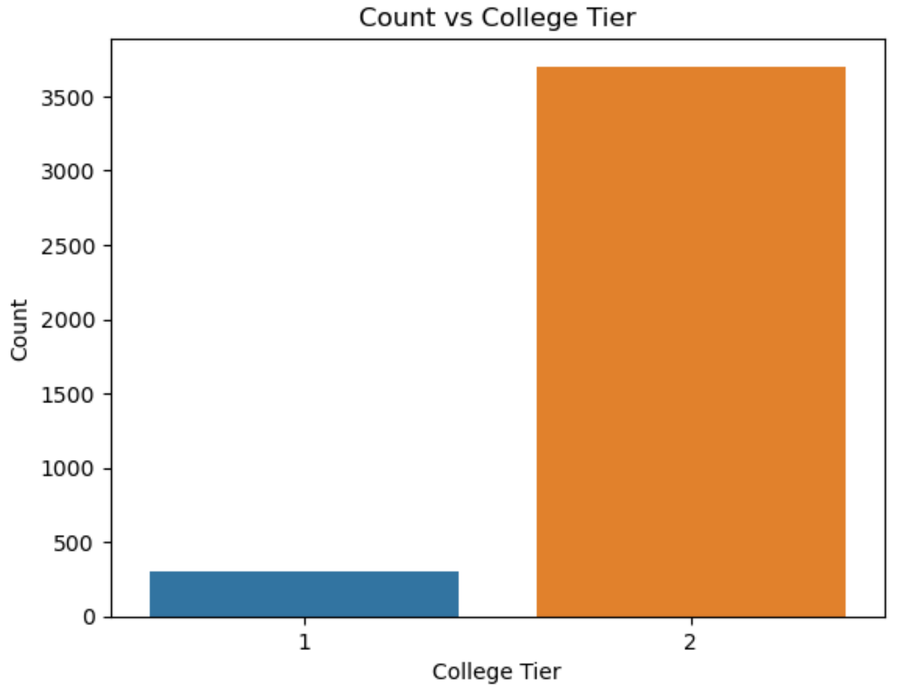

# Top 10 Deisgnation w.r.t count

The Bar plot reveals that the most preferred job designation is Software Engineer, followed by Software Developer. As the plot progresses upwards, the preference for designations decreases, positioning the most favored roles at the bottom and the less favored ones higher up in the chart.



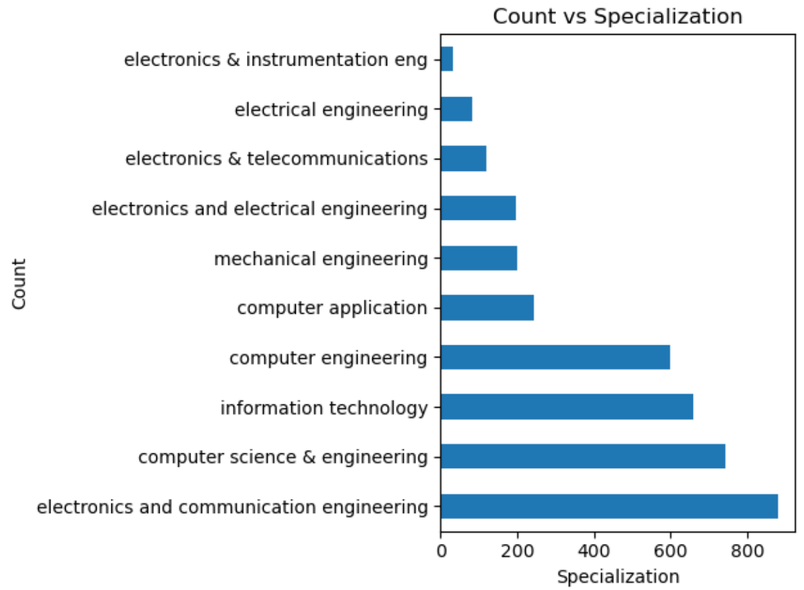
# College Tier vs Count

It's evident from the data that the majority of employees come from Tier 2 colleges, numbering around 3500, while the remaining employees hail from Tier 1 institutions.



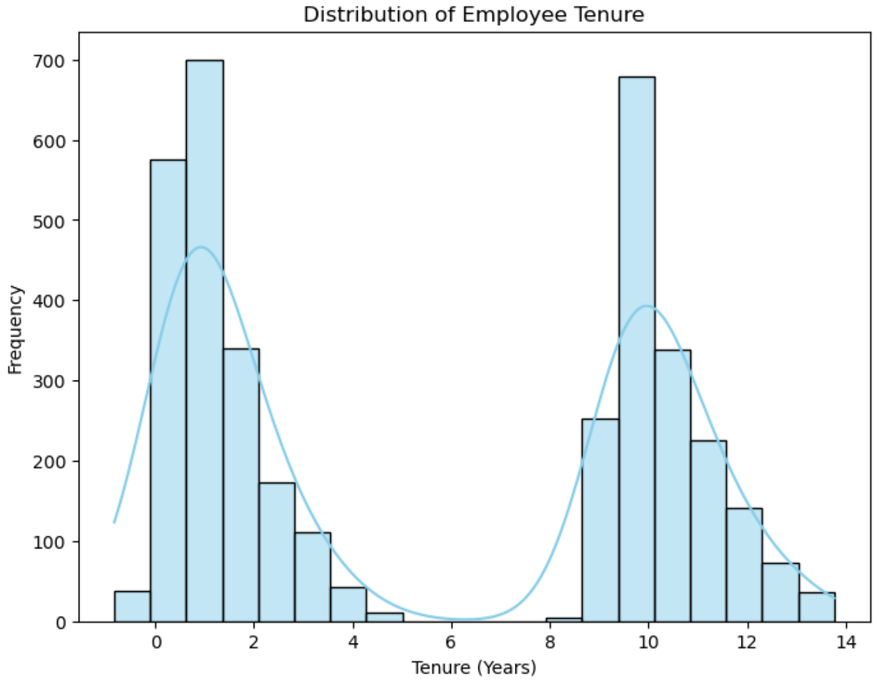

# Top 10 Specialization w.r.t. to count

The bar graph clearly represents that Electronics and Communication Engineering is most frequent specialization of Employees followed by Computer Science then by Information Technology.



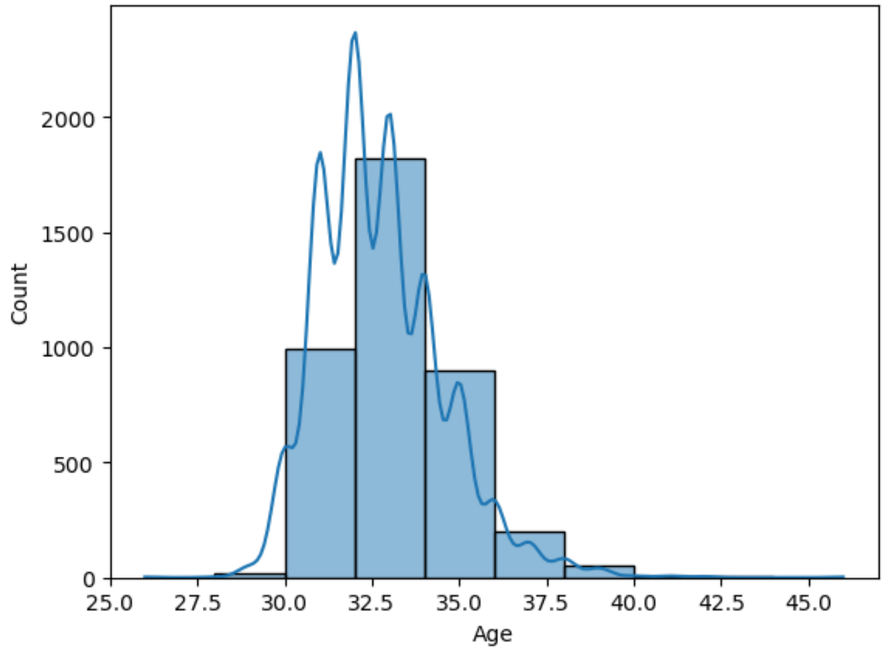
# Tenure vs Count

From Histogram and KDE plot shown, A distinctive feature observed here is the absence of employees with 6-7 years of tenure. The majority of employees fall within the tenure periods of 1-5 years and 8-16 years, showcasing a concentration in these ranges.




# Age vs Count

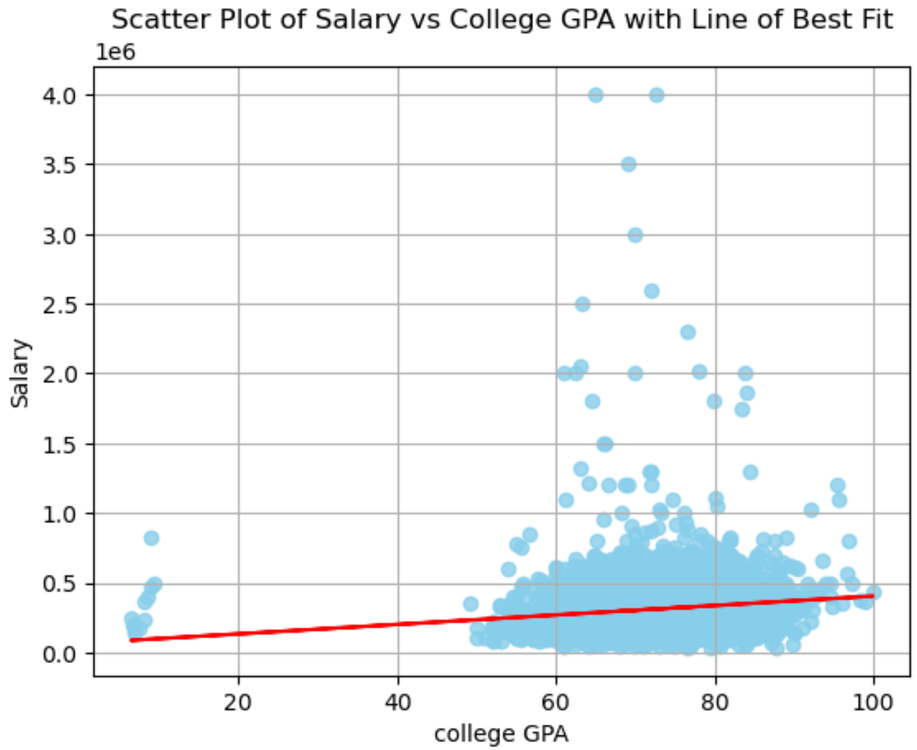
Analyzing the histogram plot, it is evident that the majority of individuals fall within the 30-36 age group, with the most concentrated segment being 32.5-35. The maximum age recorded for an employee is 40 years.



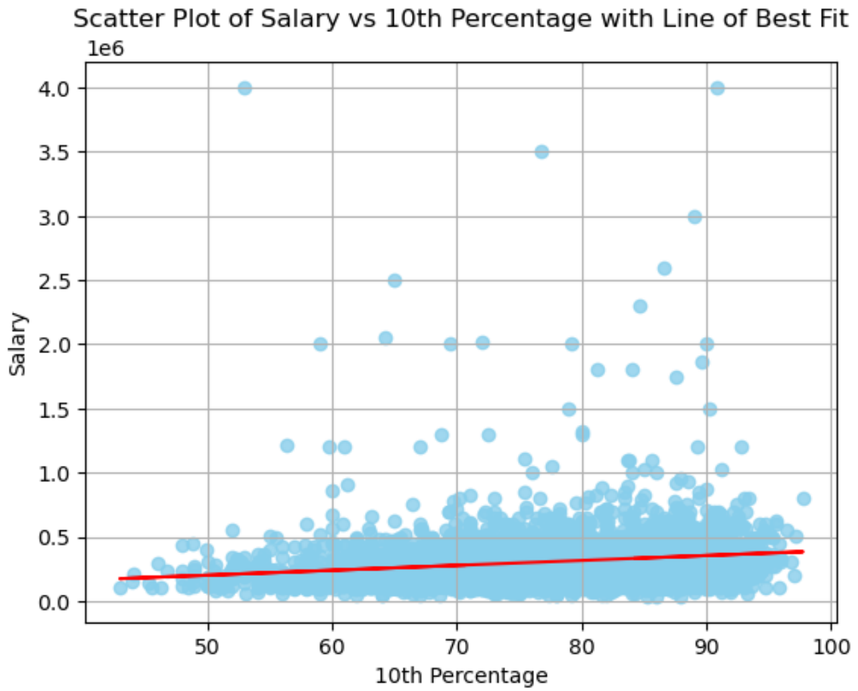
## Bivariate Analysis:

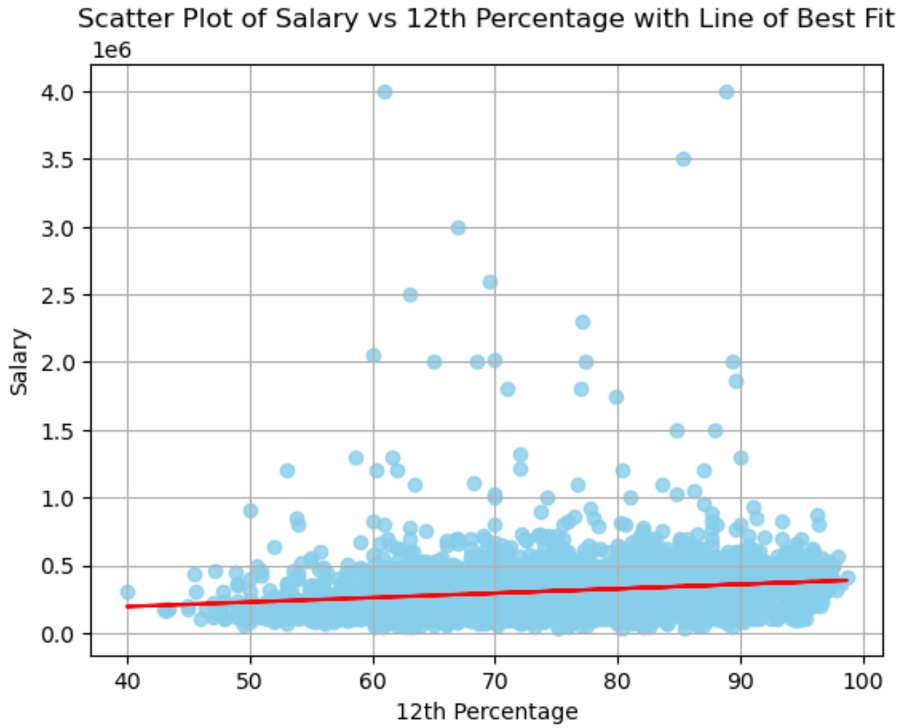
# Salary vs College GPA

The mean GPA value is approximately 71, indicating a moderately high average. A significant portion of students possesses GPAs within the 60-80 range. Interestingly, a few individuals with GPAs in the 0-15 range have secured salary packages of up to 10 LPA. Moreover, individuals with the highest salary packages typically exhibit GPAs within the 60-80 range, suggesting a positive correlation between GPA and salary attainment.



# Salary vs 10th Percentage and 12th Percentage

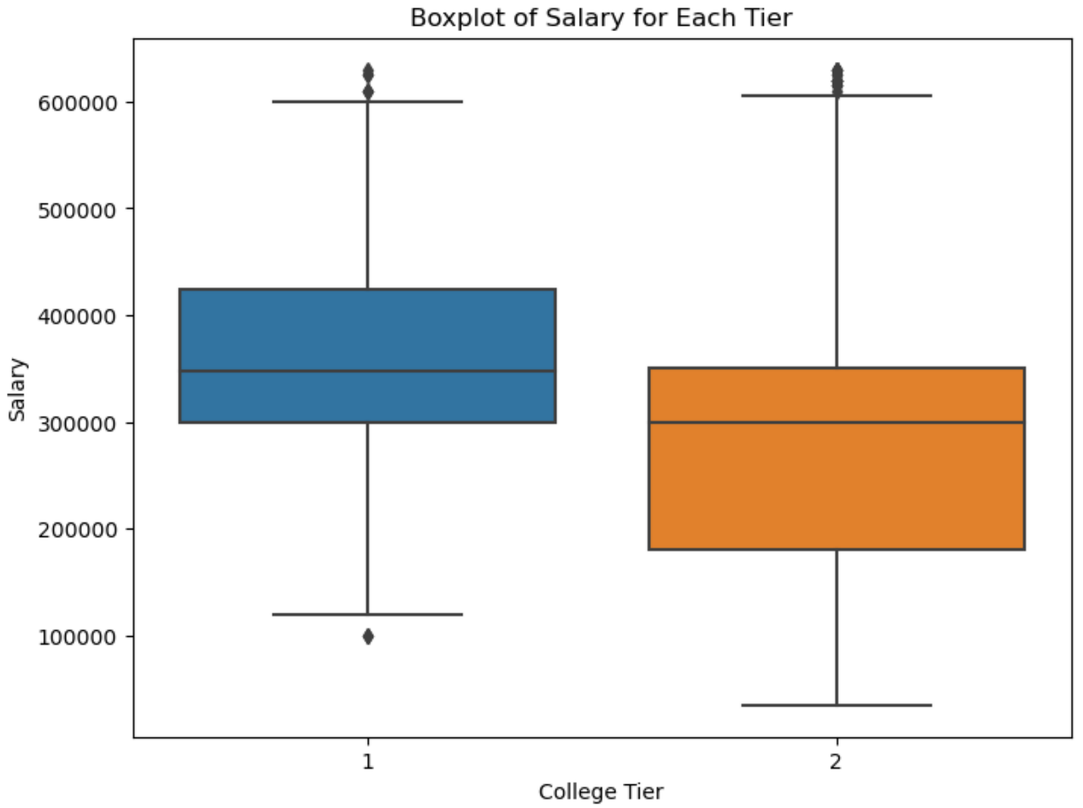



The scatter plot, along with the best-fit line, illustrates that higher scores in both the 10th and 12th grades correlate with higher salaries. However, inconsistencies arise in the dataset, such as individuals with 10th percentages around 55 and 12th percentages of 60 receiving a salary of 40 LPA. The majority of individuals achieve scores above 70% in both their 10th and 12th grades, as evidenced by the scatter plot's distribution.

# Salary vs College Tier

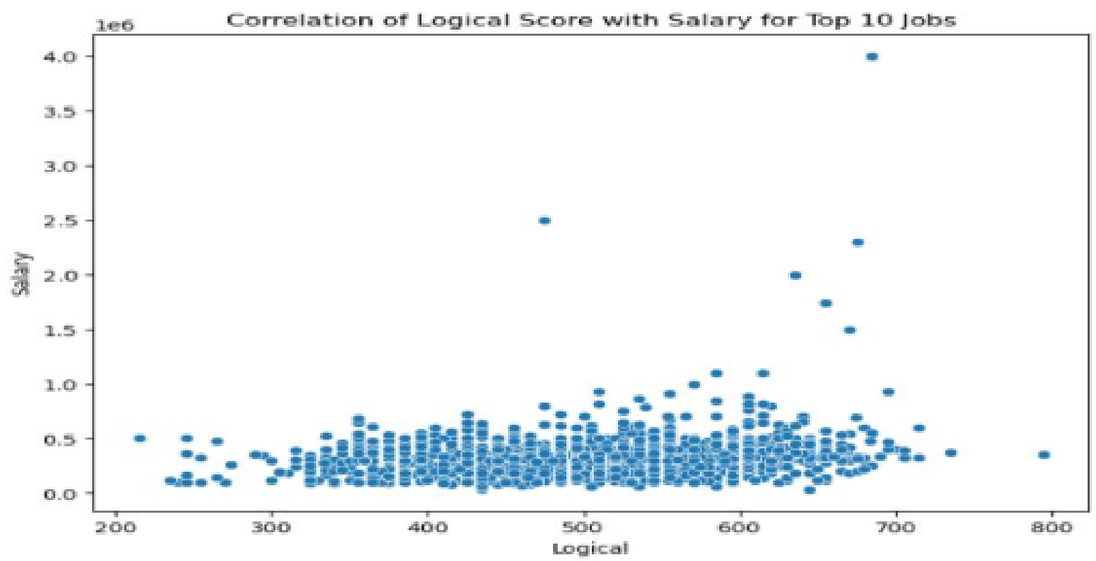
The data indicates that, on average, employees from Tier 1 colleges earn higher salaries than those from Tier 2 colleges. Notably, the employee with the lowest salary comes from a Tier 2 college. However, it's worth noting that employees with the highest salaries are almost evenly distributed between Tier 1 and Tier 2 colleges.

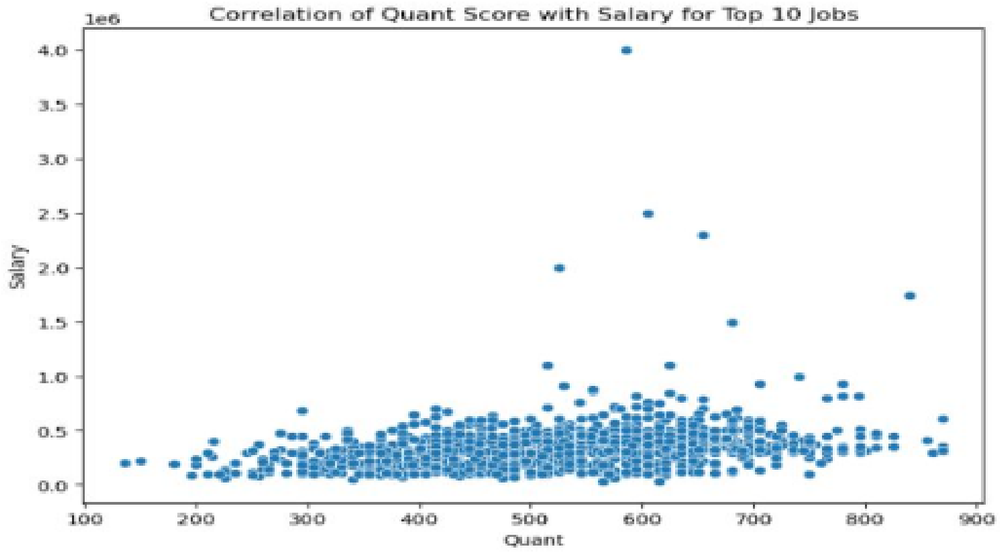



# Salary vs English Score


The given correlation between Score in English and salary is showing the need of english skill for certain salary expectation.
The plot is showing the need of atleast moderate english skill for a
decent salary.
High salary category need the english score greater than 500.
Most of the people has the score in the range of 400-600.

# Salary vs Logical and Quant Score



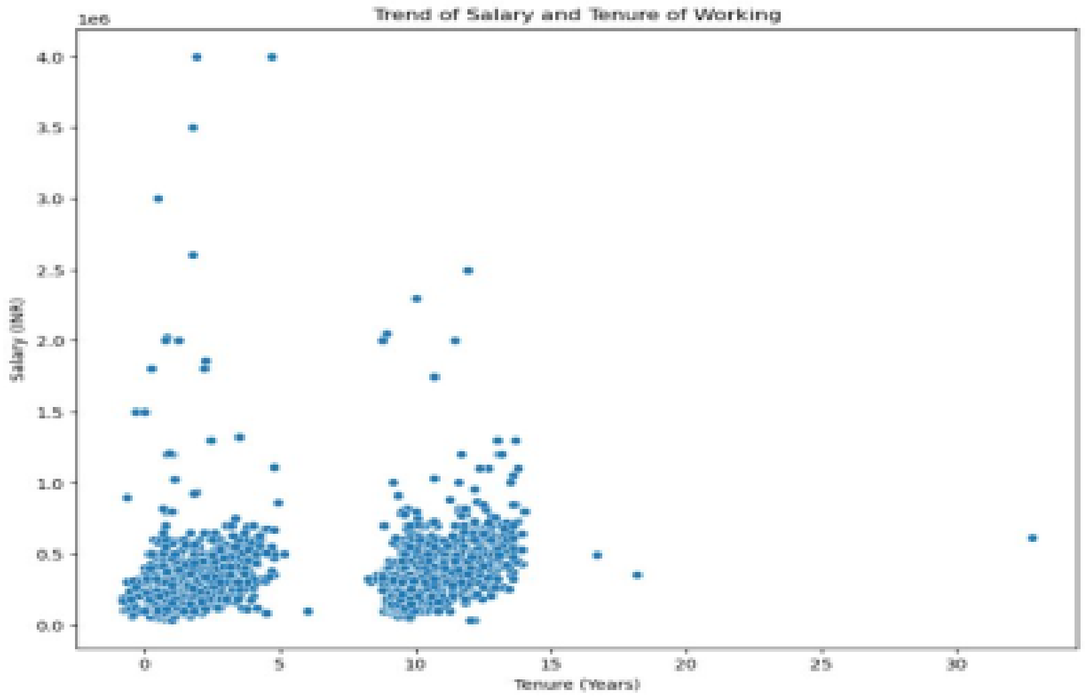

High salary jobs need the logical score of approximately greater than 600 and Quant score greater than 500.

Most of the people has quant and logical scores in the range of 400-600.

Individual with maximum salary has the logical score of approximately 700.

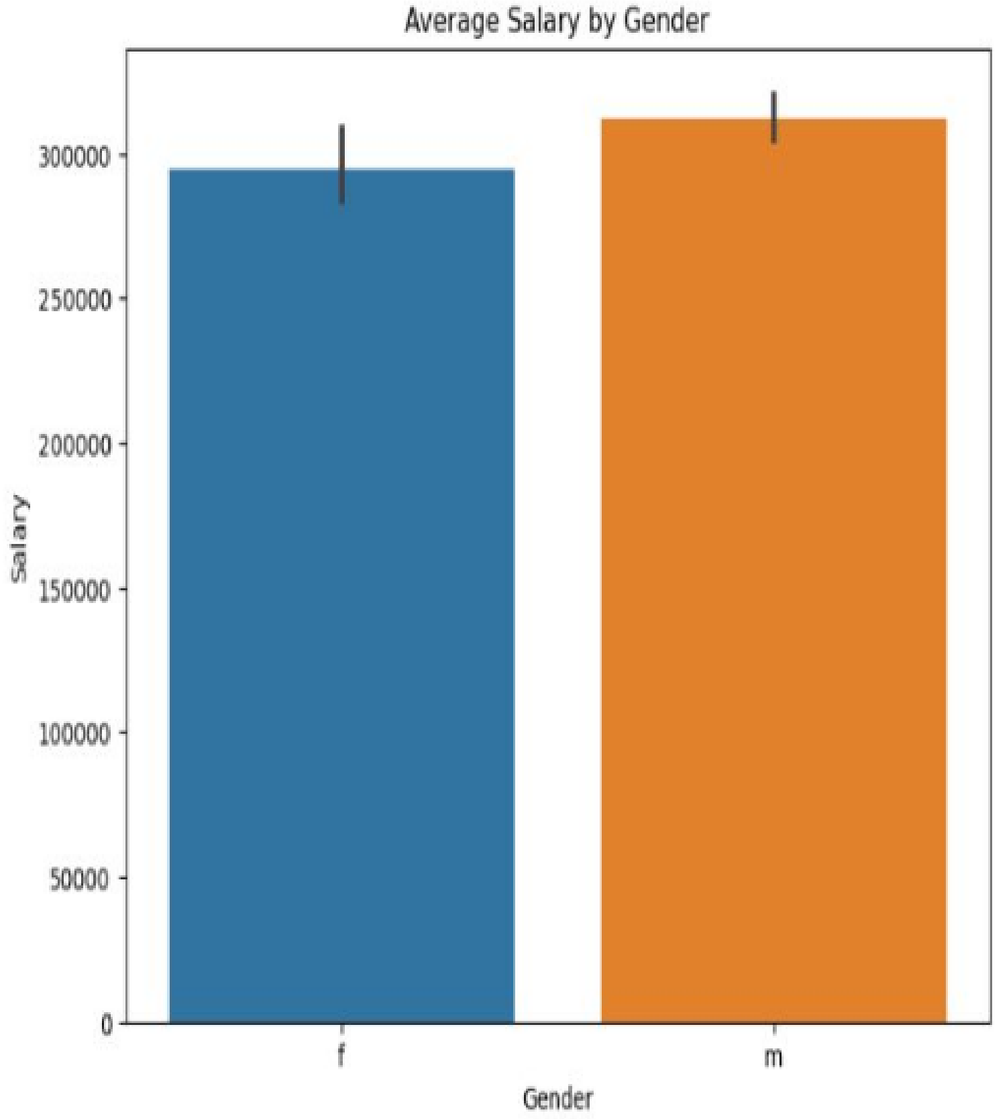
Similarly, individual with maximum salary has Quant score of almost 600.

# Salary vs Tenure




Given plot is showing how the salary is changing with respect to the job period.
It is clear from the plot that as the tenure is increasing the salary of individual also increases.

# Average Salary Vs Gender




The histogram illustrates the average salaries of males and females, indicating that males generally earn more than females. However, the disparity in average salaries between males and females is relatively small.

# Conclusions:

- Salaries in the dataset are influenced by various factors including tenure, college level, and job designation. Senior Software Engineers tend to earn the highest incomes compared to other job designations
- The project delved into a comprehensive dataset of engineering graduates' employment outcomes, with particular attention to the variable "Salary." It employed a range of data manipulation and visualization techniques to analyze and interpret the data effectively.
- Region-specific insights highlighted salary trends in major cities and identified specific job roles with competitive average salaries.
- Academic performance indicators, such as 10th, 12th, and college GPA scores, do not exhibit a clear correlation with pay levels.
- Further analysis, potentially incorporating machine learning, was proposed to gain deeper insights into salary influencers and inform future decision-making.
- In summary, the project lays a crucial groundwork for comprehending the employment dynamics among engineering graduates and offers valuable insights for organizations and policymakers to improve their employment practices.

# THANK YOU