INNOMATİCS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

- Vaibhav Saini

About me

I am Vaibhav Saini, pursuing my Dual Degree(BTech + MTech) in Electrical Engineering from IIT BHU, Varanasi. Throughout my academic and professional journey, I have developed a keen interest in Data Science particularly in the field of Data Analysis.

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.

I want to learn data science because I'm intrigued by its ability to uncover insights from data, driving informed decisions. It's a sought-after skill in today's data-driven world, offering opportunities to solve complex problems and make a positive impact. The interdisciplinary nature of data science, combining statistics, computer science, and domain expertise, is particularly appealing. Overall, I'm excited about the potential to innovate and create value through data analysis.

In my journey through data science, I gained valuable hands-on experience during my internship at Nexus Software, where I delved into data analytics. This opportunity equipped me with top-notch skills and practical insights, enhancing my expertise in the field.

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 the Data

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.

Data cleaning and preparation

In the column name JobCity some of the values are -1 which doesn't make any sense so either consider it in others category or just simply ignore it to prevent data inconsistency.

In DOL column some of the cells have “present” in it ,so replace “present” with today’s date.

Preparing the data by converting data in DOJ and DOL columns in datetime format and using it to calculate tenure.

Exploratory Data Analysis

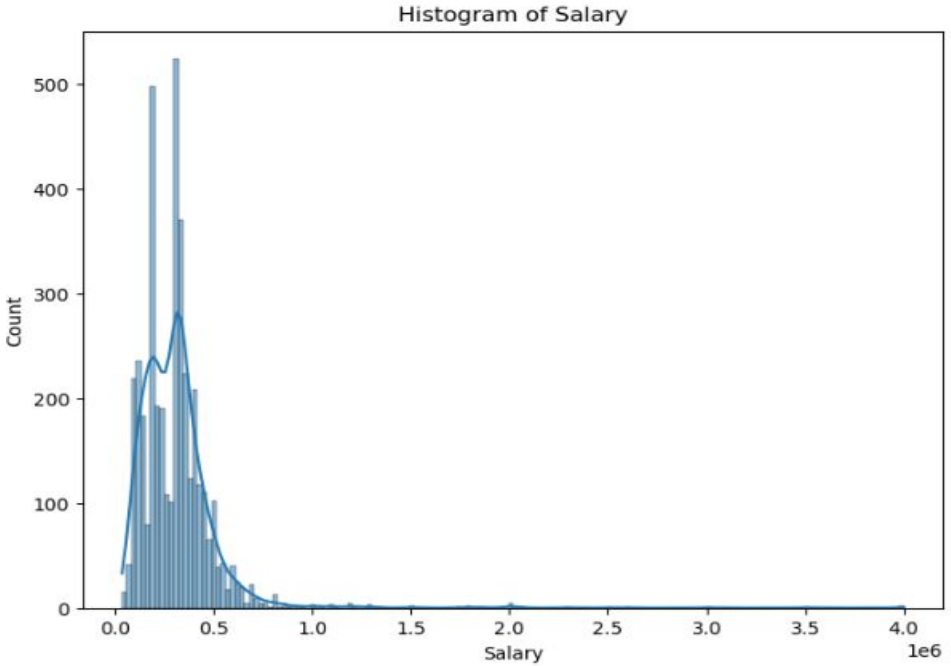
Univariate Analysis

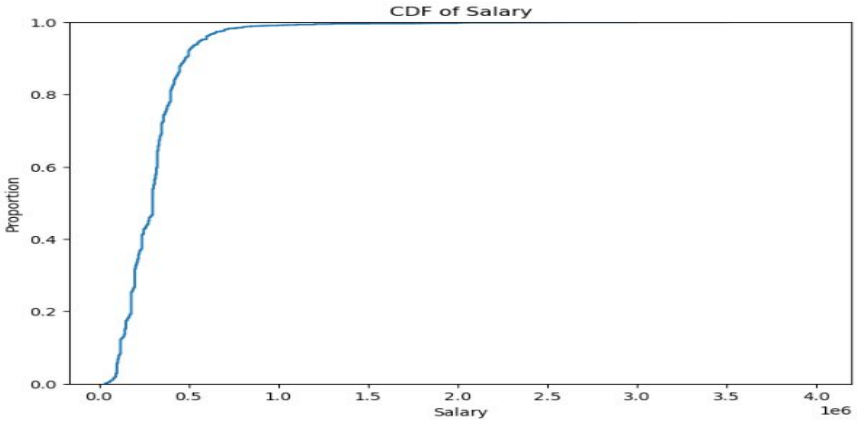
Salary Analysis

Histogram representing the **count of salaries**, it is clearly observable here that maximum salary is 40 lpa and minimum salary is about 35K.

Most of the people are having salary in the range of 2.5 lpa - 5 lpa.

The trend line over the histogram is clearly showing how the people count changes with the salary value.





The above shown plot of CDF is showing how the distribution of the salary data is converging to 1 on 20 Lpa showing most the people salaries are under the range of 20 lpa.

The box plot is describing the statistical aspect of data on salary:

Mean salary: 3,07,699
Min salary: 35,000
Max salary: 40,00000

Though 40,0000 is outlier in our data ,besides this there are many outliers in the data as shown in box plot.

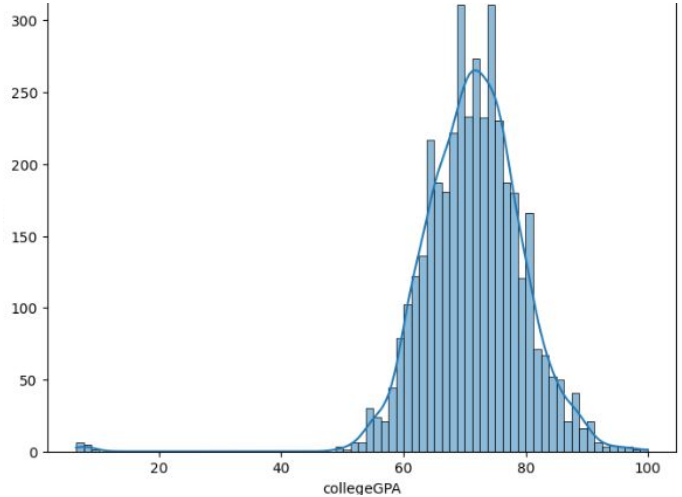
Without outliers range of salary will be approximately: [35,000 - 5.5 lpa].

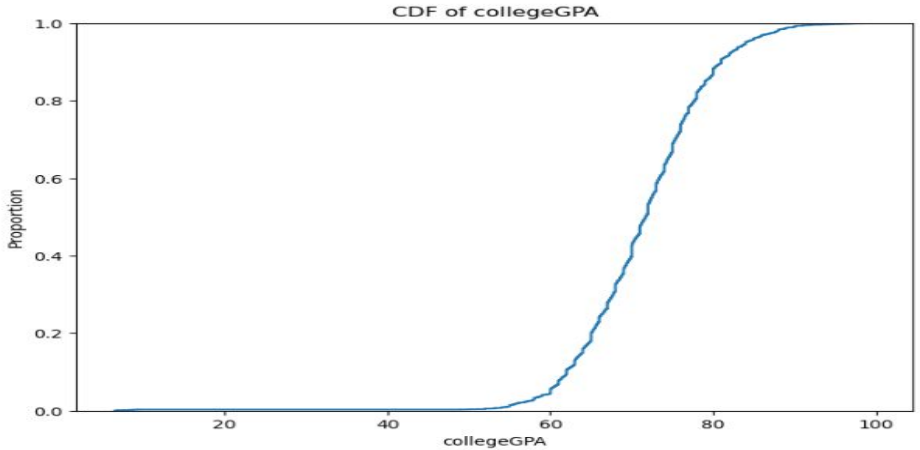
College GPA analysis:

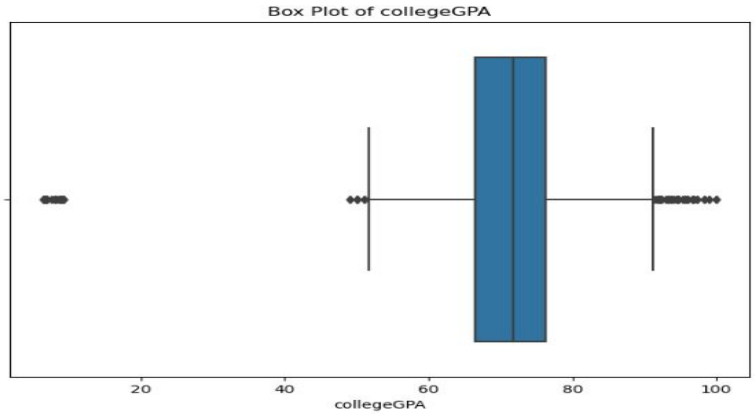
The histogram is showing the people count with certain range of GPA in college. The histogram is showing most of the students has the college GPA in the range of 60-80.

The trend line in the histogram is showing how the count of people are changing wrt college GPA.

There are no student with GPA especially in the range of 20-40.





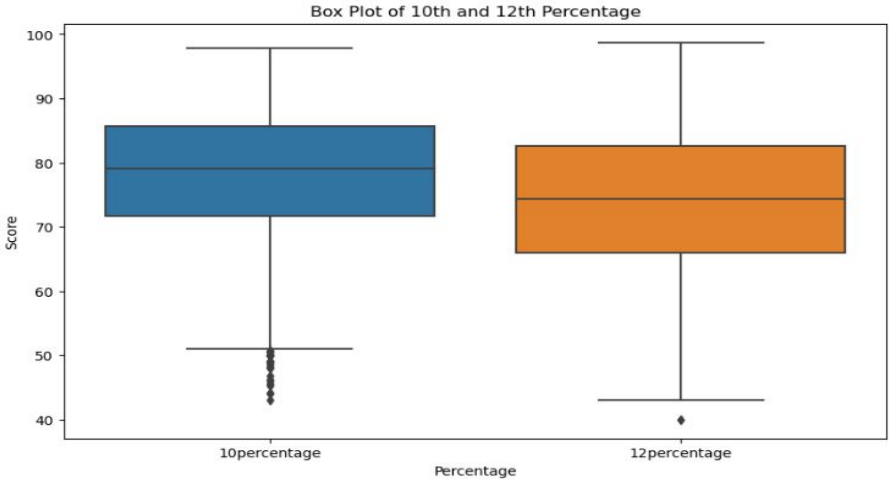


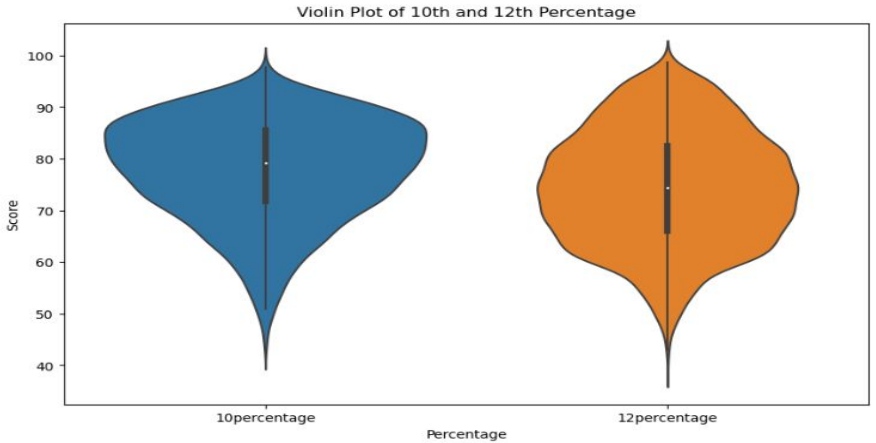
The above given CDF plot is showing how the curve for college GPA is converging near to 100 (not exactly 100),showing that no student has achieved the 100 GPA .

The box plot is clearly depicting the statistical measure of data on college GPA :
Mean GPA: 71.486171
Max GPA: 99.930000
Min GPA: 6.450000

There are some outliers as shown in box plot by dots ,specifically in the low GPA region indicating few student score GPA of 0-15.

Analysis of 10th and 12th percentage





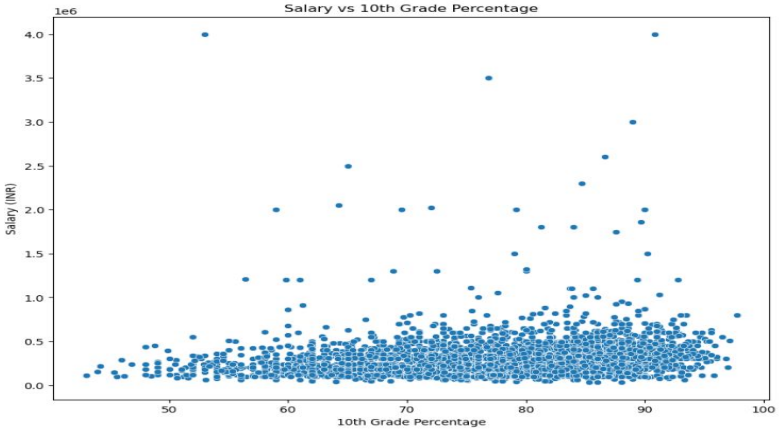
The above shown Box Plot and Violin plot is representing the distribution of 10th and 12th percentage.

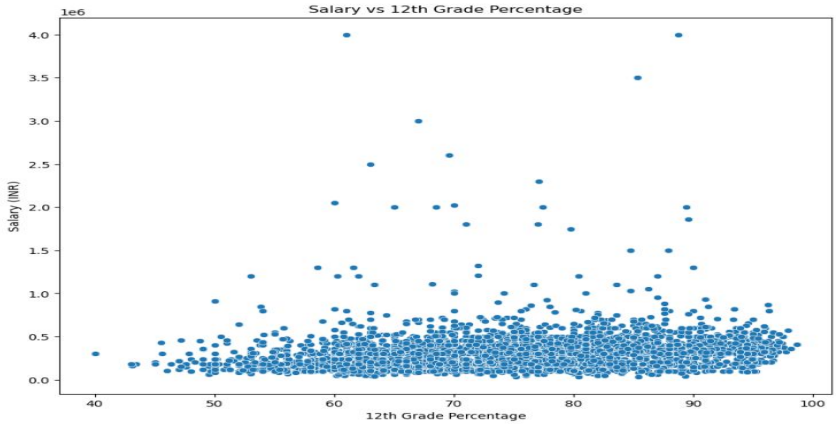
We can gain statistical insight from these graph as:

	10th	12th
Mean:	77.925443	74.466366
Max:	97.760000	98.700000
Min:	43.000000	40.000000

Outliers are present in the data and are shown in the box plot by the dots.

Effect of 10th and 12th scores on salary





It can be depicted from the scatter plots that more the individual scores in 10th and 12th the more will be his/her salary.

There are some inconsistencies in the data as it can be seen that the individual with 10th percentage around 55 and individual with 12th percentage as 60 is getting the salary of 40 lpa.

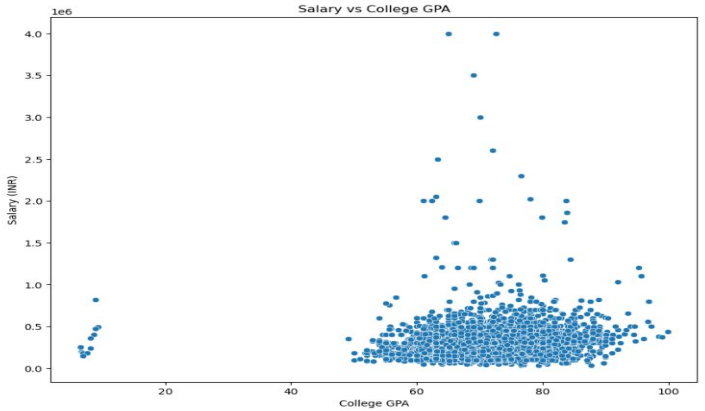
Most individuals score above 70% in 10th and 12th as can be seen by the distribution of scatter plot.

It can be seen that mean GPA value is around 71.

Most of the students are having the GPA in the range if 60-80.

There are few individuals having GPA in the range if 0-15 and able to grab the salary package if upto 10 lpa.

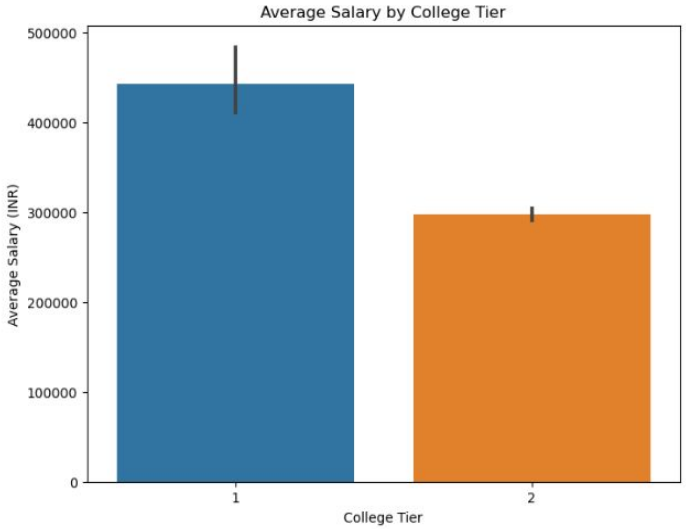
Individuals with maximum salary package are having GPA in the range of 60 -80.



College Tier and Average salary

Tier 1 college is having the more average salary as compared to Tier 2 college .

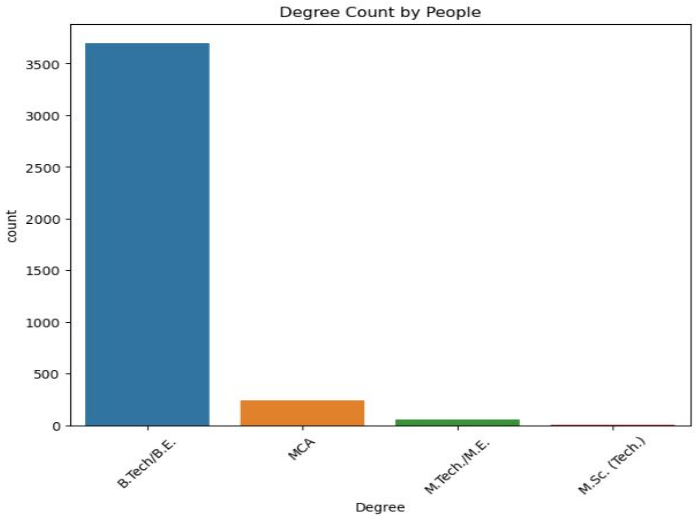
And as per the analysis college with city tier 0 and 1 are having almost same average salary



Degree Count by People

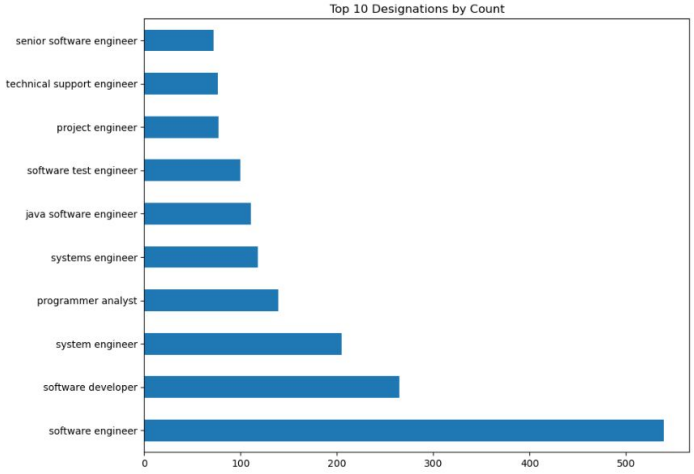
Given histogram is showing the count of people with the specific given degree-

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.



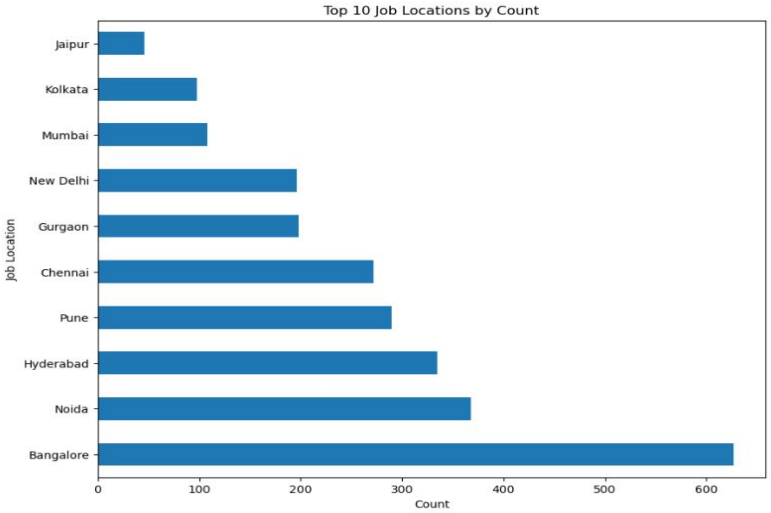
Top 10 preferable designations

Most preferred designation is software engineer, then software developer followed by as shown in the histogram plot with most preferable designation at the bottom and less preferable as going up.



Top 10 preferable job locations

The histogram is clearly showing the order of preferred location for doing job as Bangalore is the most preferred location followed by noida, then hyderabad and so on.



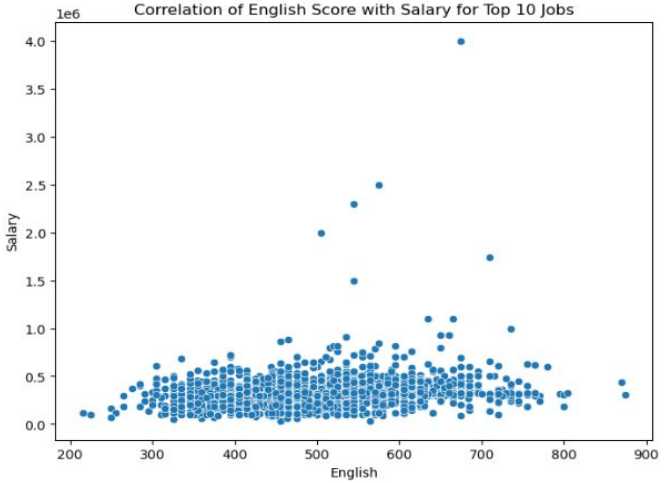
Correlation of English score with salary

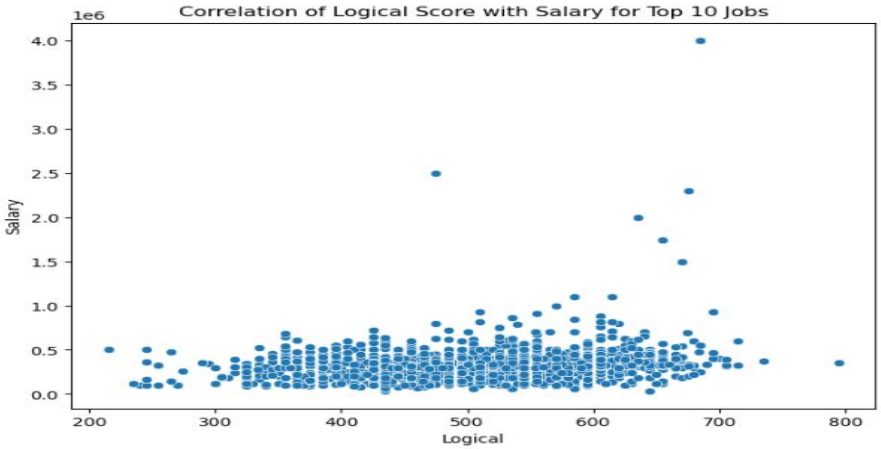
The given correlation between Score in English and salary is showing the need of english skill for certain salary expectation.

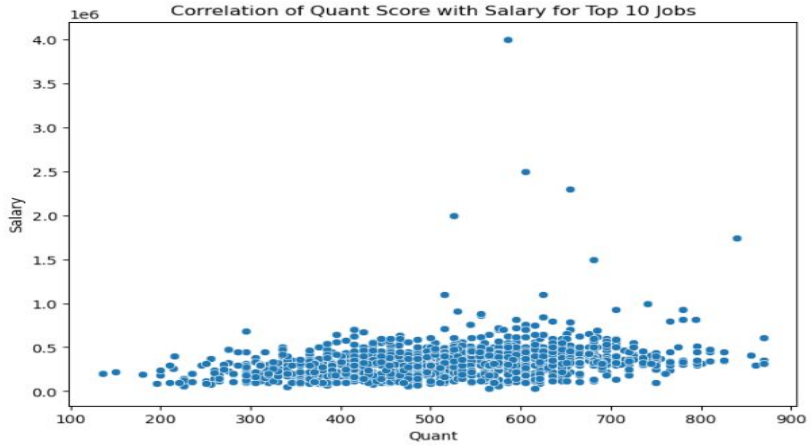
The plot is showing the need of atleast moderate english skill for a decent salary.

High salary category need the english score greater than 500.

Most of the people has the score in the range of 400-600.







Correlation between Logical and Quant with salary

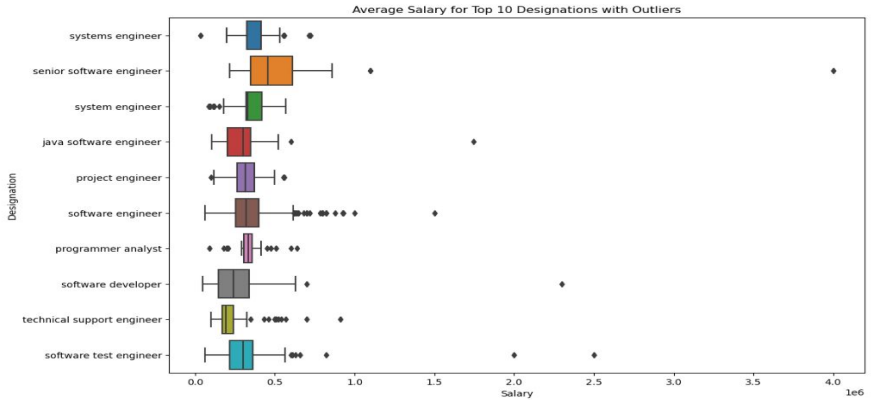
High salary jobs need the logical score of approximately greater than 600 and Quant score greater than 500.

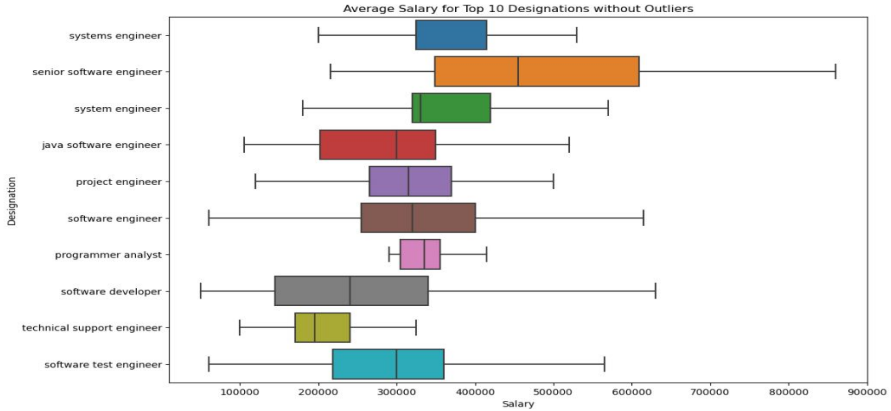
Most of the people has quant and logical scores in the range of 400-600.

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

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

Bivariate Analysis





Box plot for average salary of top 10 designation

The box plot is showing the statistical analysis of salaries in each of the top 10 designation .

Senior software engineer is having the wide range of salary from 2- 9 lpa on the other hand programmer analyst has the small range but moderate salary range of 3-4 lpa.

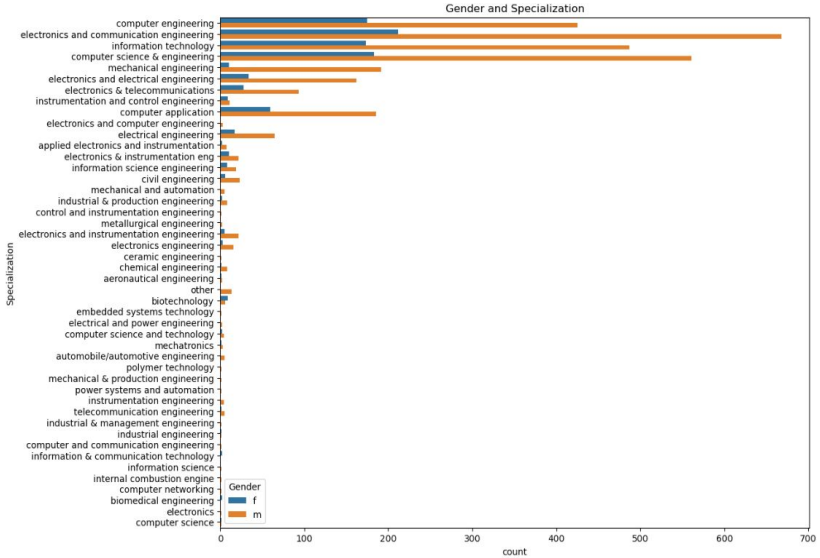
Software engineer salaries has maximum outliers and senior software engineering is having the maximum salary of 40 lpa and as of minimum salary is in the field of system engineer.

Overall analysis , reveal senior software engineer as high paying job and technical support engineer as low paying job.

Correlation of tenure and salary

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.



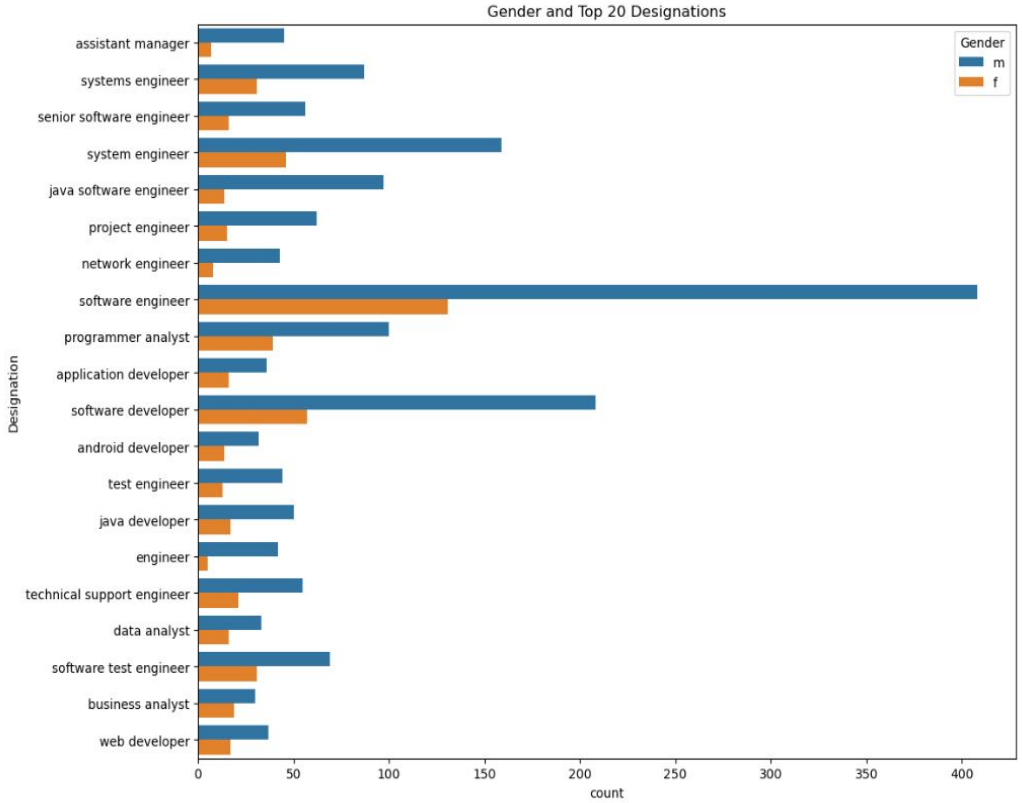
Gender and specialization

The given histogram is showing the count of male and female from a certain engineering field.

Electronics and communication is the most preferred discipline followed by computer science and technology, Information technology, Computer engineering and so on.

In almost all the discipline the no. of males are more than females.

But biotechnology is the branch which is more preferred by females than males.



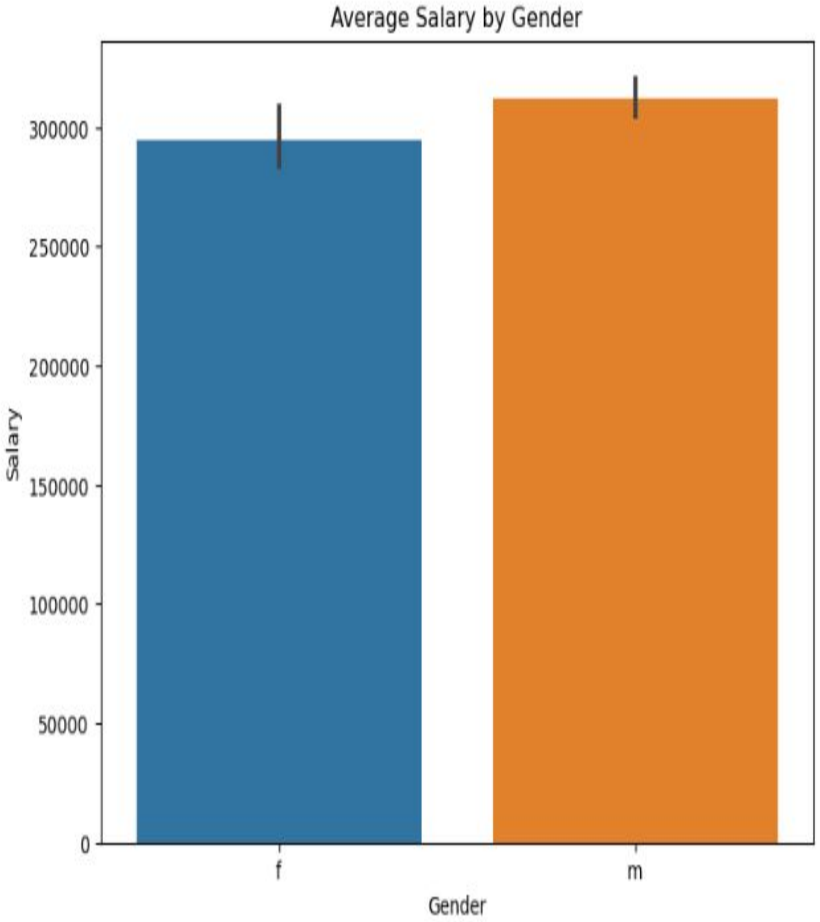
Gender by Top 10 designations

The given histogram is giving insight about the count of male and female in a certain designation.

As can be seen, that software engineer is the most preferred designation among both male and female.

Business analyst is having the most balanced male is to female ratio among other designations.

There are more assistant male manager as compare to no. of female assistant manager.



Average salary by Gender

The histogram is giving the insight about the average salary of male and female , showing the average salary of male is more than female.

But difference in average salary of male and female is not too large.

CONCLUSION

- Salary levels in the dataset are impacted by factors such as tenure, college level, and job designation, with Senior Software Engineers commanding the highest incomes.
- Gender appears to have a limited impact on average income determination, but females tend to receive lower salaries than the overall average.
- Academic performance indicators, including 10th, 12th, and college GPA scores, do not show a clear correlation with pay levels.
- The project analyzed a comprehensive dataset of engineering graduates' employment outcomes, focusing on the variable "Salary" and employing various data manipulation and visualization techniques.
- Region-specific insights highlighted salary trends in major cities and identified specific job roles with competitive average salaries.
- The project emphasized the importance of addressing gender pay gaps and understanding the relationship between education and salary for more equitable employment practices.
- Further analysis, potentially incorporating machine learning, was proposed to gain deeper insights into salary influencers and inform future decision-making.
- Overall, the project serves as a valuable foundation for understanding employment dynamics among engineering graduates and provides insights for organizations and policymakers to enhance employment practices.

THANK YOU