INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project - AMCAT Data Analysis

About me

- **Currently i am working as Project CoOrdinator lead at Tech Mahindra, handling different projects on content Moderation, Data Labelling and Team Management and Client for various domains which help process to improve and grow.**

- **I am passionate about the data analysis which help organizations to solve the problems,**

Name- Amit Mazumdar
Linkin- https://www.linkedin.com/in/amitkmazumdar/
GitHub- https://github.com/Kingsukh?tab=repositories

Objective of the project

The goal of this Exploratory Data Analysis (EDA) is to extensively investigate the provided dataset, with a particular emphasis on understanding the link between various variables and the target variable, Salary.

The key aims of this analysis include:

Providing a detailed explanation of the dataset's features.

Find any observable patterns or trends in the data.

Investigating the relationships between the independent factors and the target variable (salary).

Identify any outliers or abnormalities in the dataset.

Offering practical insights and recommendations based on the analysis.

Exploratory Data Analysis:

Imported Libraries: Pandas, NumPy, Matplotlib, and Seaborn for data manipulation and visualization.

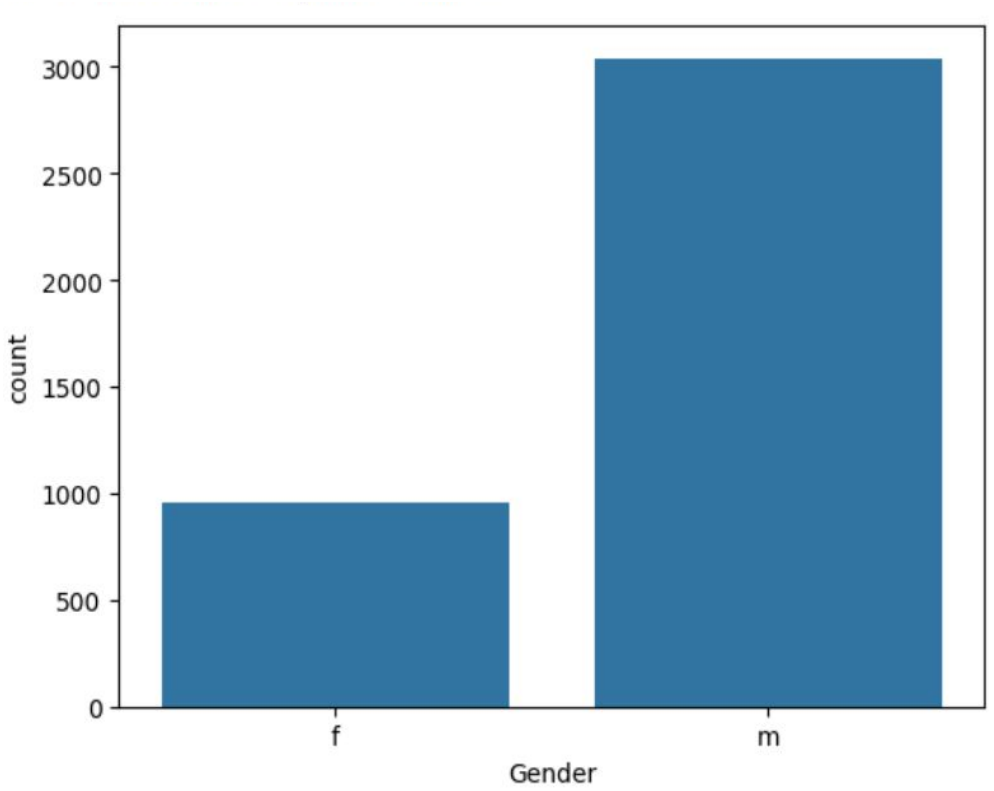
Load Dataset: Read the dataset into a Pandas DataFrame for exploration.

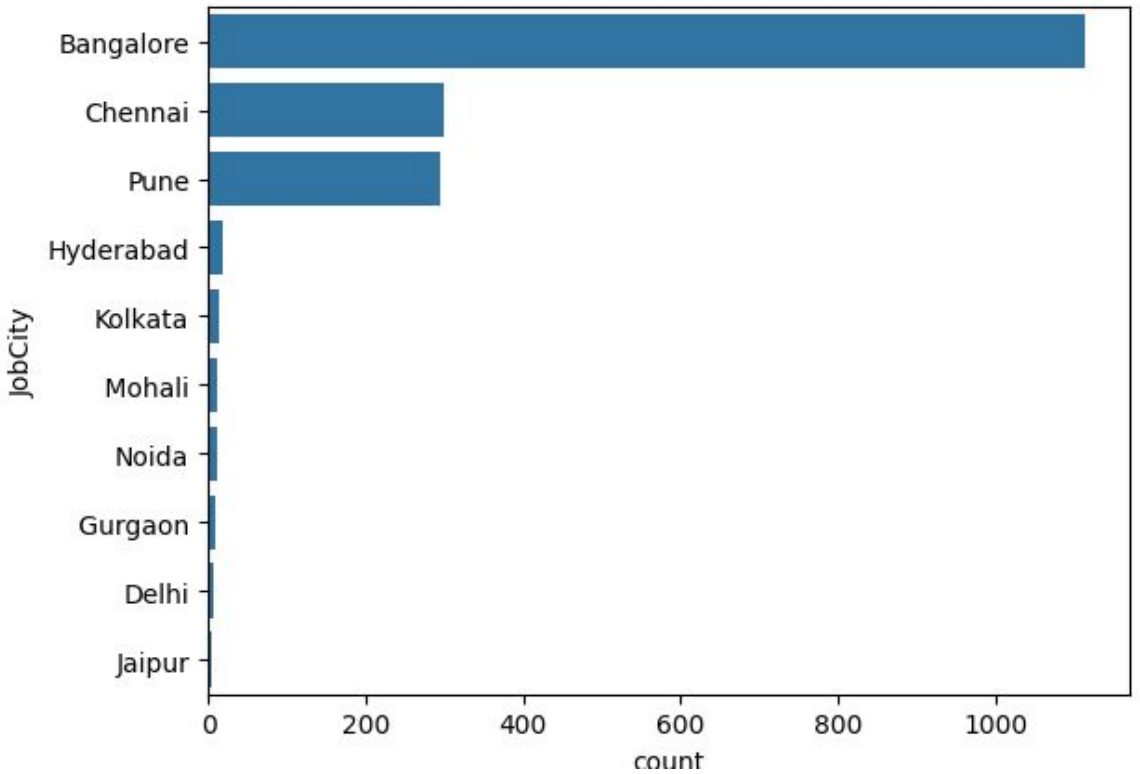
Initial Exploration: Use functions like .shape, .describe(), and .info() to understand the dataset size, data types, and summary statistics.

Data Cleaning Steps:
Handling Missing Data: Identify and address any missing values.
Remove Duplicates: Eliminate any duplicate entries in the dataset.
Correct Inconsistent Values: Fix any inconsistencies in the data. Data cleaning involves rectifying or removing incorrect, corrupted, improperly formatted, duplicate, or incomplete data.

Data Manipulation Steps:
Map Categorical Values:: Mapped Categorical values to proper streams
Removed Unnecessary Columns: Dropped unnecessary columns i.e Unnamed
Filterting :filtered the data as per the requirement for analysis

Univariate Data Analysis:

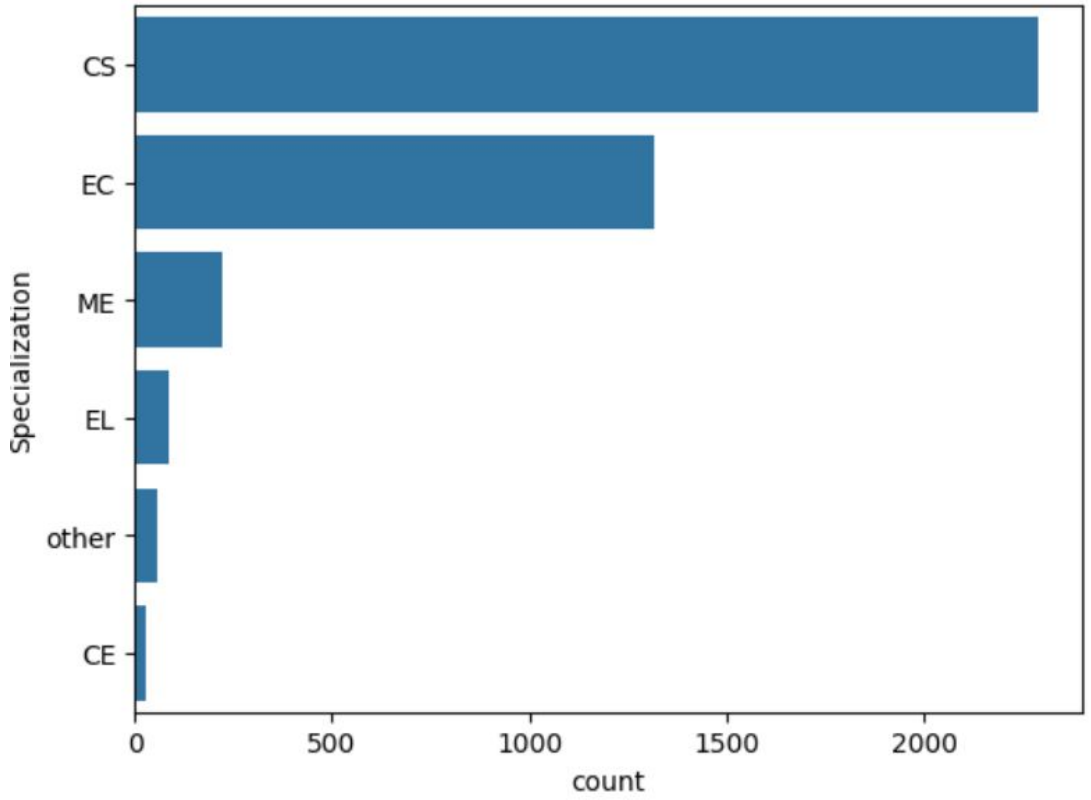


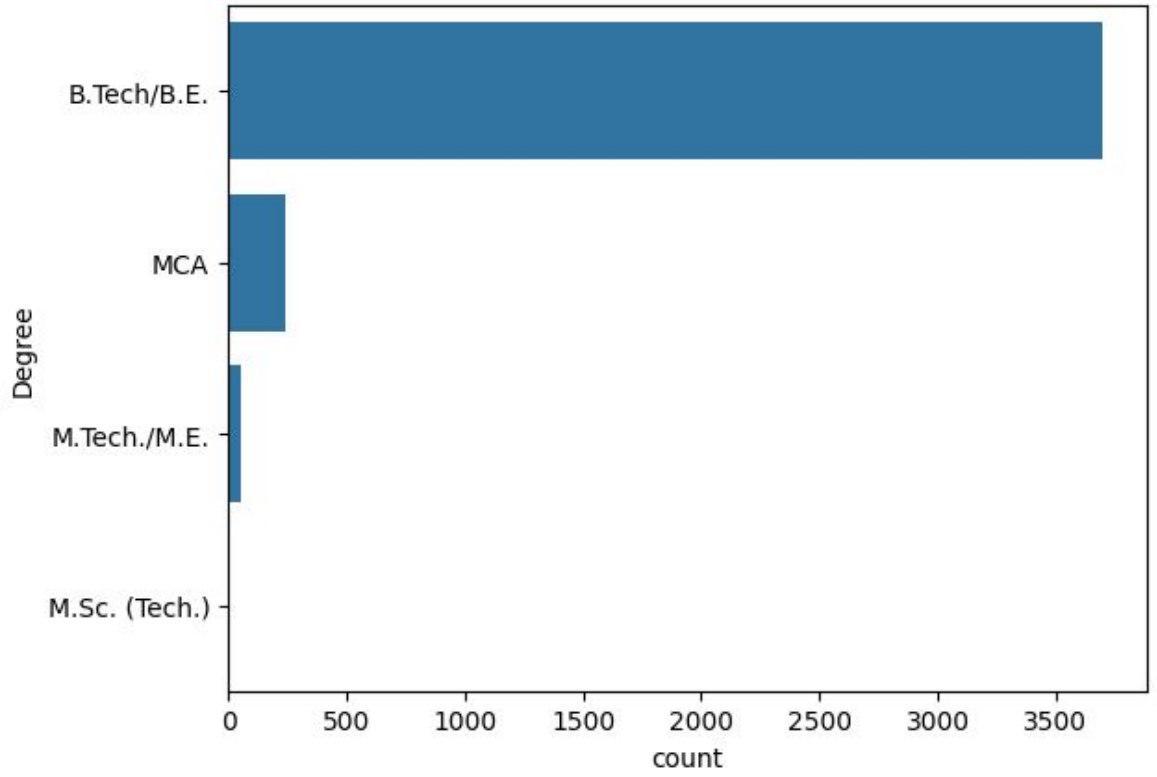


- We have observed the females graduates are pursuing less jobs compare to the male
- Observed the highest number of jobs are available at Bangalore and followed by Chennai

INNOMATICS RESEARCH LABS

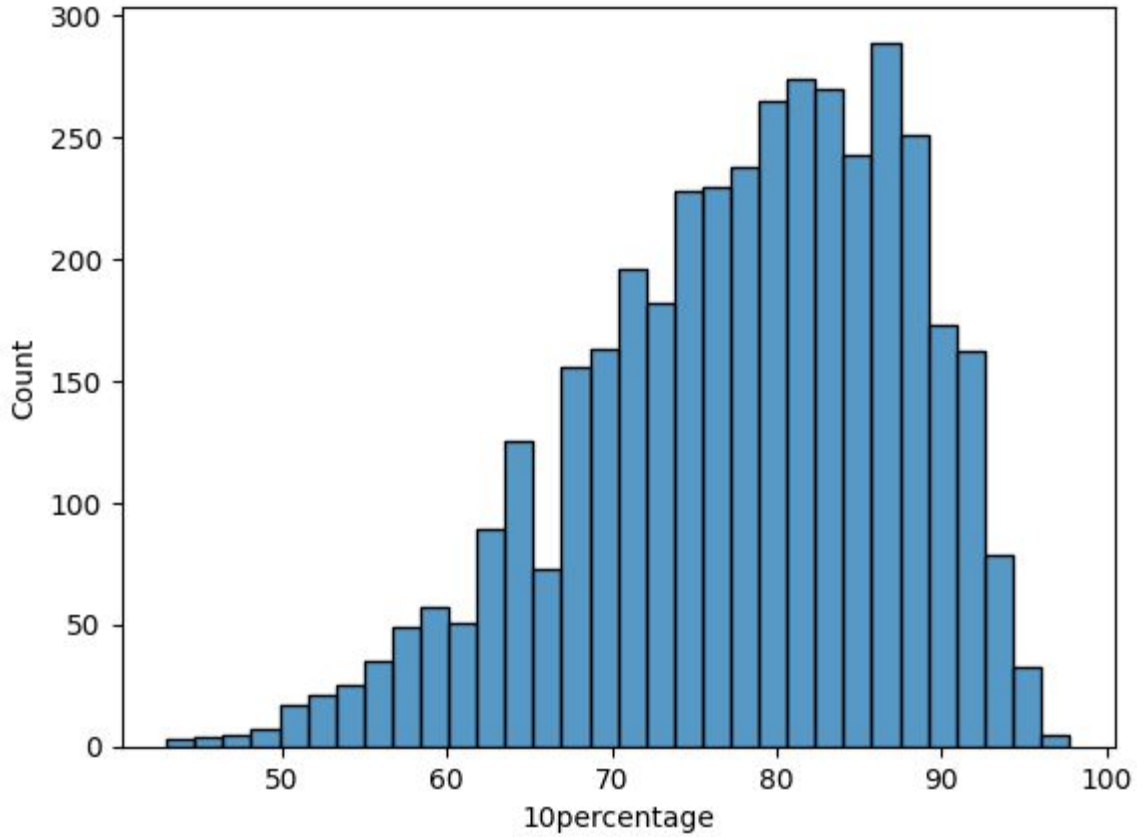
Univariate Data Analysis:



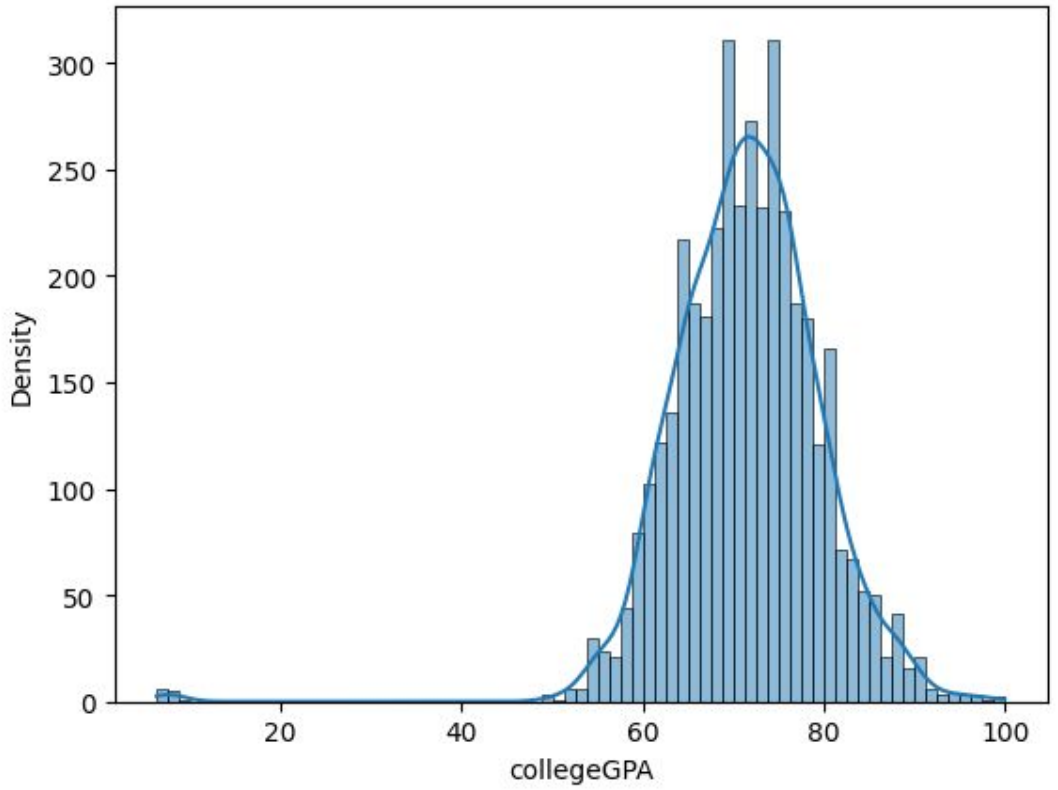


- from the above three plots we can conclude that most of the employees are from BTech graduate
- Observed the highest number of jobs are available at Bangalore and followed by Noida

Univariate Data Analysis:



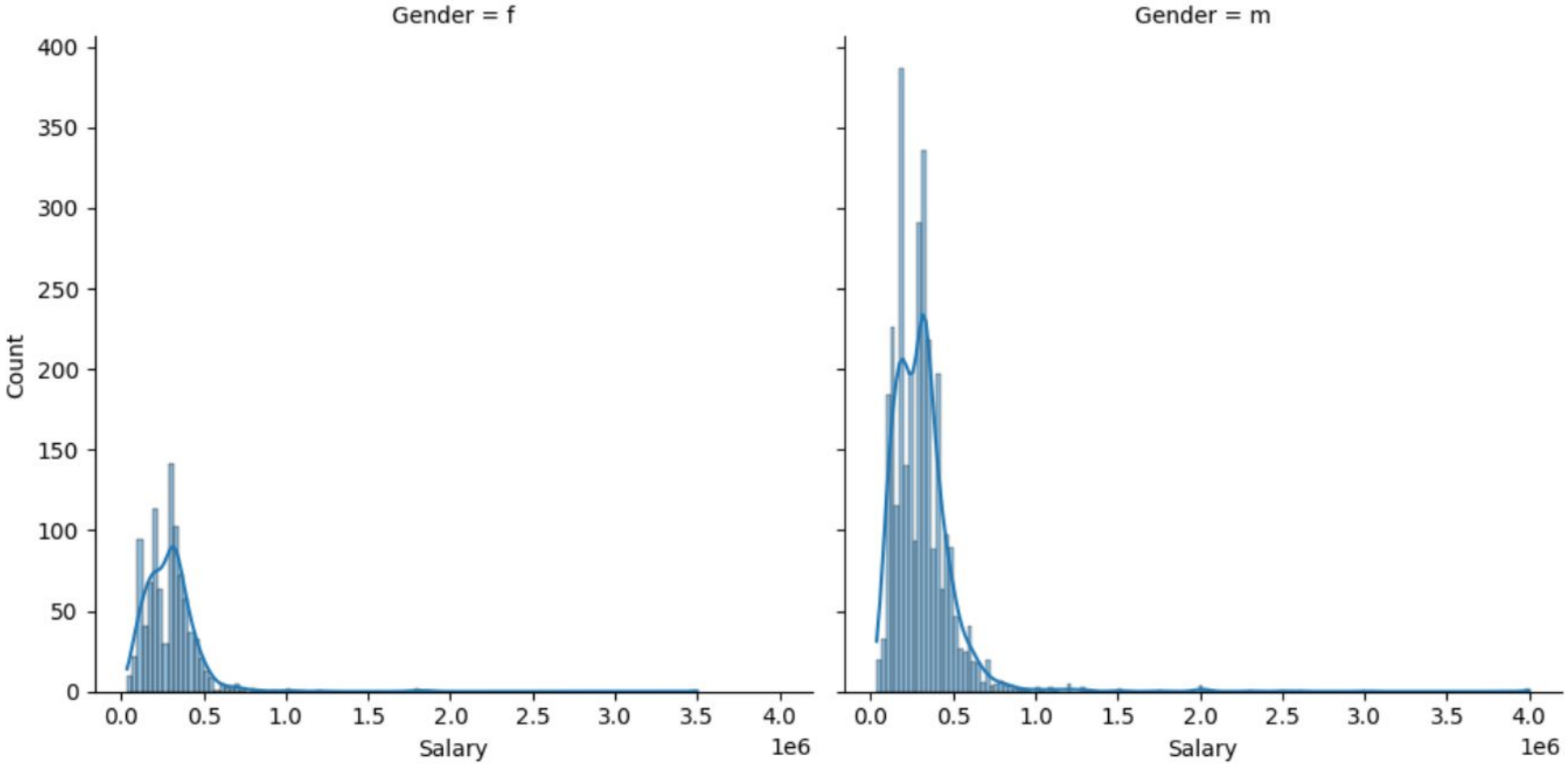
Most of the employees passed with around 88 percentage in 10th standard



College GPA of the employees are around 60 to 75

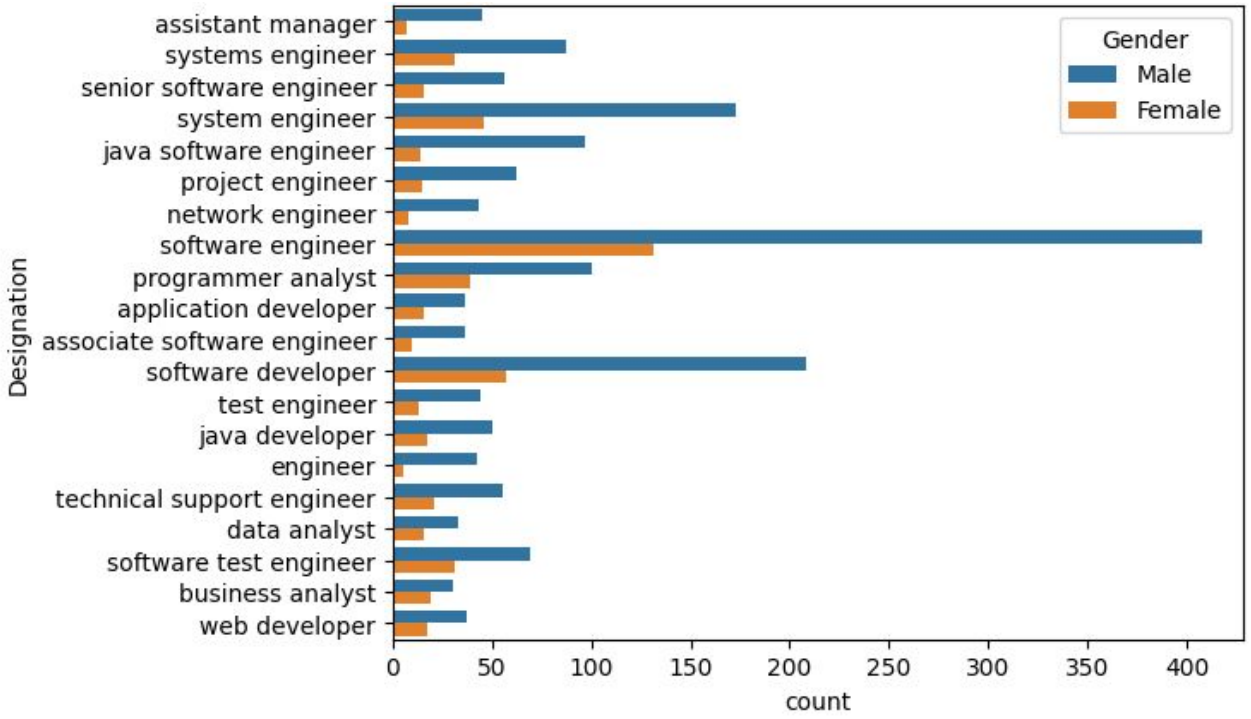
INNOMATICS RESEARCH LABS

Bivariate Data Analysis:

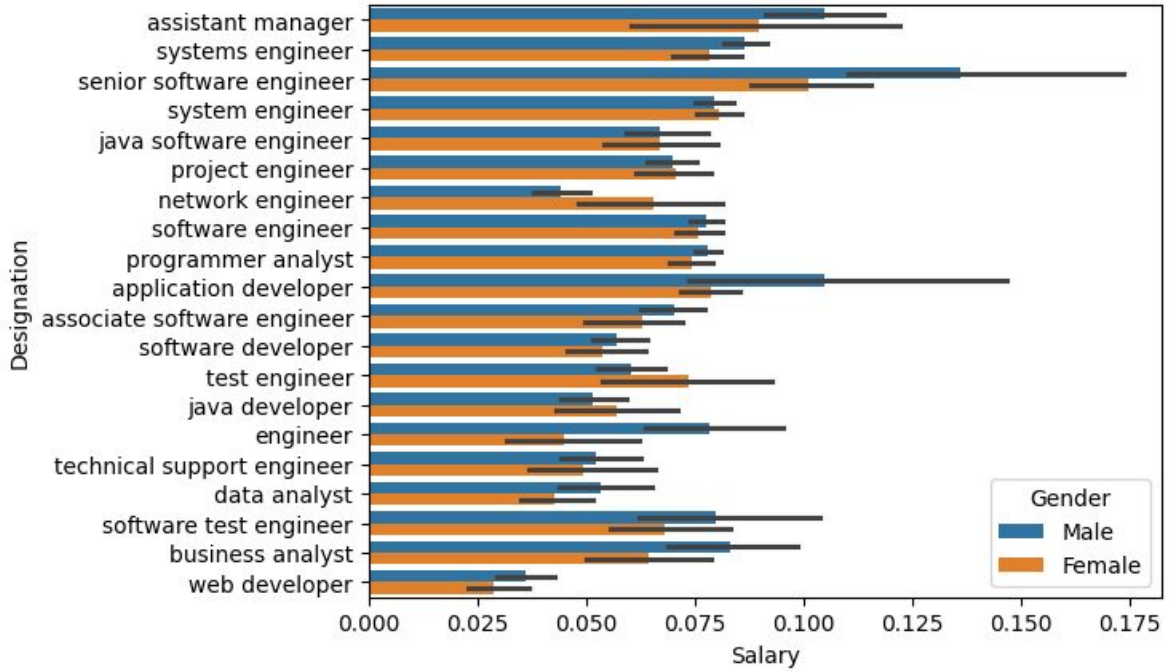


Observed highest salaries are offered to the male compare to female employees

Bivariate Data Analysis:



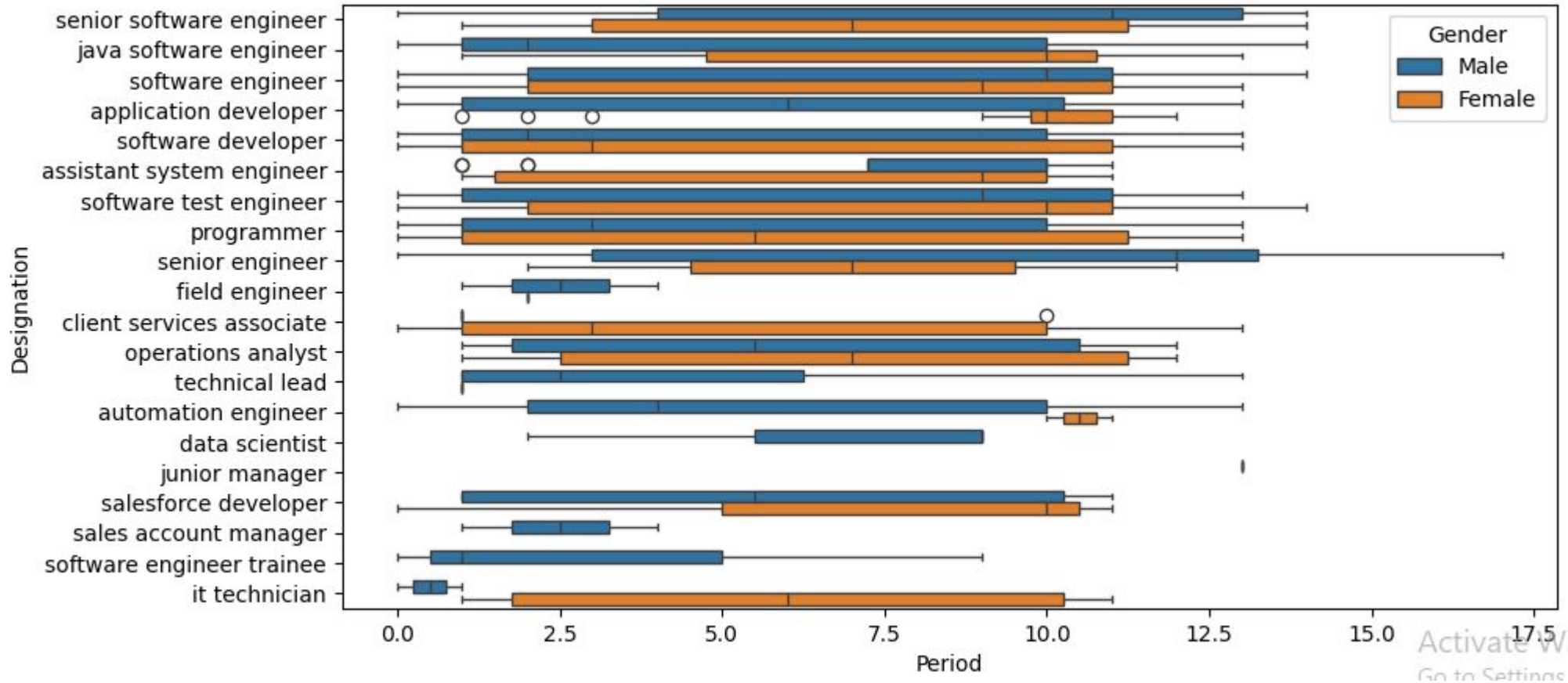
In all the professions the jobs are mainly dominated by Male



From the above barplot we can conclude that in all the designations male are not dominating with respect to salary, female employees are also getting better salary in some designations

INNOMATICS RESEARCH LABS

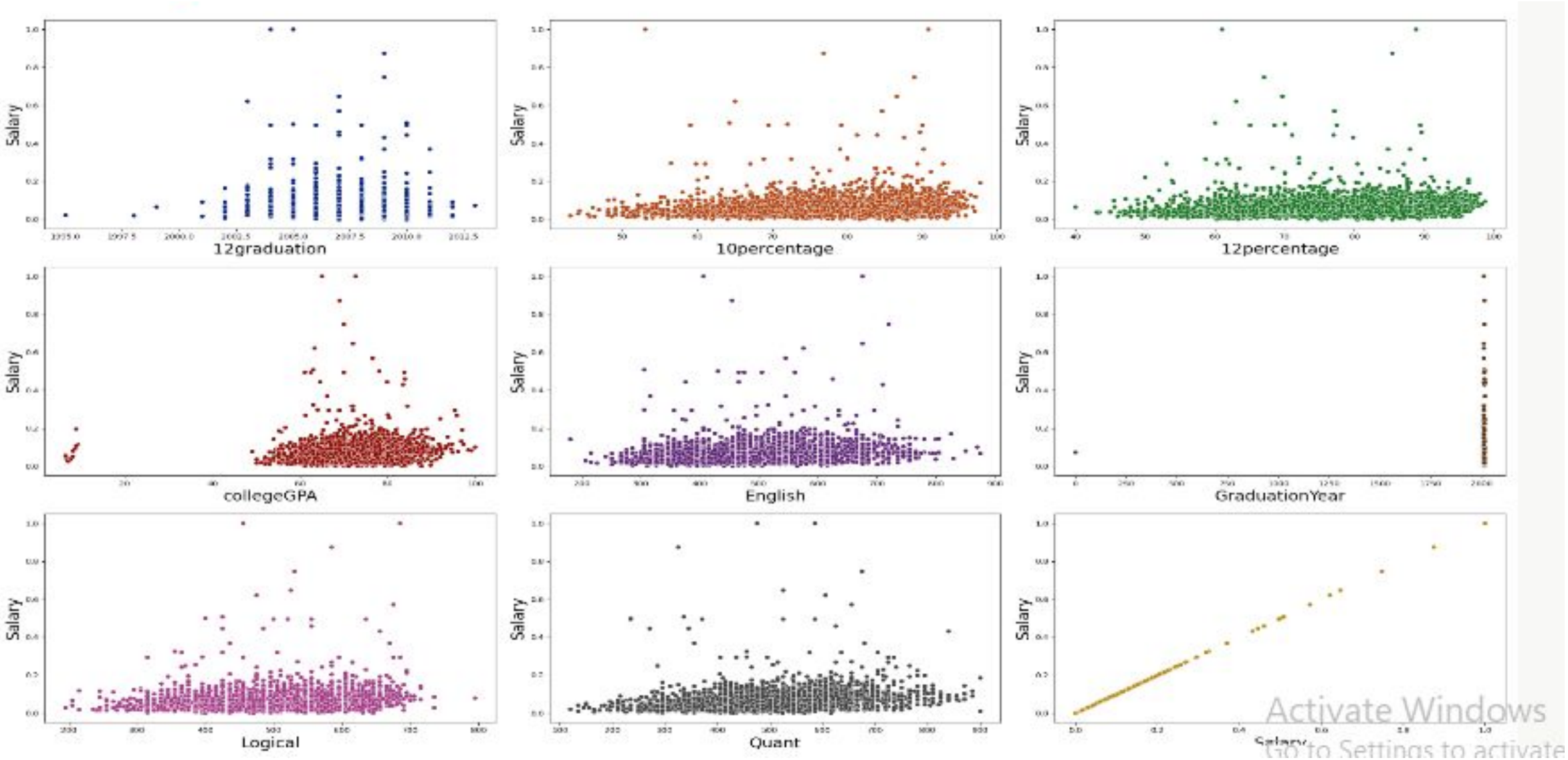
Bivariate Data Analysis:



The analysis reveals that while experience varies among designations, and there are median experience differences between genders, the lack of a strong correlation between experience and salary, alongside instances where women with higher experience earn less than men, suggests that salary disparities cannot be solely attributed to experience levels.

INNOMATICS RESEARCH LABS

Outlier Analysis:



Challenges:

While working on the AMCAT dataset project, one of the primary challenges was managing the complexity and diversity of data features, Ensuring data quality through effective cleaning was crucial, as missing and inconsistent values in variables such as job location, and education presented significant obstacles during the analysis.

Conclusion:

- Based on the analysis, the claim of a salary range between ₹2.5-3 lakhs for Computer Science graduates in specific job roles is not supported by the data.
- There is significant relationship between gender and specialization ,p-value is less than 0.05.
- No strong correlation between salary and GPA and also there is no correlation between tier colleges and salary
- Observed highest salaries are offered to the male compare to female employees
- The dataset includes employment outcomes of engineering graduates (Salary, Job Titles, and Locations) and scores in cognitive, technical, and personality skills. The dataset contains 4000 rows and 40 columns with many duplicates, so it first needs cleaning, including removing unwanted rows and checking for missing values.We performed univariate analysis with plots like PDFs, histograms, boxplots, and countplots to identify outliers and frequency distributions in the data. Additionally, bivariate analysis was done using scatter plots, pair plots, barplots etc to examine relationships between numerical and categorical columns.Through these analyses, we explored relationships between employees' salaries and their backgrounds, such as graduation year, designation, and academic performance (10th and 12th percentages).

THANK YOU

INNOMATICS RESEARCH LABS