INNOMATICS® RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Data

M SAI CHARAN

About me

Background:

I am M Sai Charan holding with Bachelor of Science in computer Science and Statistics.

Motivation for Data Science:

Following graduation, I found myself drawn to the world of Artificial Intelligence. Intrigued, I embarked on a journey of exploration, seeking the right path to align with my interests. Through research and introspection, I discovered that Data Science resonates deeply with me, offering a perfect fit for my aspirations and skills.

Work Experience:

I am Currently interning at Innomatics Research Labs, transitioning from a Bsc background to data science.

LinkedIn: https://www.linkedin.com/in/sai-charan-mora/
GitHub: https://github.com/saicharanmora

Agenda (This should be the PPT flow)

- Business Problem and Use case domain understanding(If Required)
- Objective of the Project
- Web Scraping – Details (Websites, Processor you followed)
- Summary of the Data

- Exploratory Data Analysis:

a. *Data Cleaning Steps*
b. *Data Manipulation Steps*
c. *Univariate Analysis Steps*
d. *Bivariate Analysis Steps*

- Key Business Question
- Conclusion (Key finding overall)
- Q&A Slide
- Your Experience/Challenges working on Web Scraping – Data Analysis Project.

Objective:

- Our mission in conducting Exploratory Data Analysis (EDA) on the Aspiring Mind Employment Outcome 2015 (AMEO) dataset is to unlock a comprehensive understanding of employment outcomes for engineering graduates.
- Through meticulous analysis, we seek to uncover nuanced patterns, unveil correlations, and reveal emerging trends within the dataset. By scrutinizing factors like salary, job titles, locations, and AMCAT exam performance, our aim is to illuminate the intricate web of influences shaping candidates' employment journeys, providing invaluable insights into this dynamic landscape.

Description:

- The dataset contains information on 3,998 individuals, spanning across 39 columns. Each row represents a unique individual, while each column provides specific details about their employment and educational background.
- **Key columns include:**
  - **ID**: Unique identifier for each individual.
  - **Salary**: The salary earned by the individual.
  - **DOJ**: Date of joining the organization.
  - **DOL**: Date of leaving the organization.
  - **Designation**: Job title or position held.
  - **Job City**: City where the job is located.
  - **Gender**: Gender of the individual.
  - **DOB**: Date of birth.
  - **10percentage**: Percentage obtained in 10th grade.
  - **10board**: Board of education for 10th grade.
  - **12graduation**: Year of graduation from 12th grade.
  - **12percentage**: Percentage obtained in 12th grade.

Description:

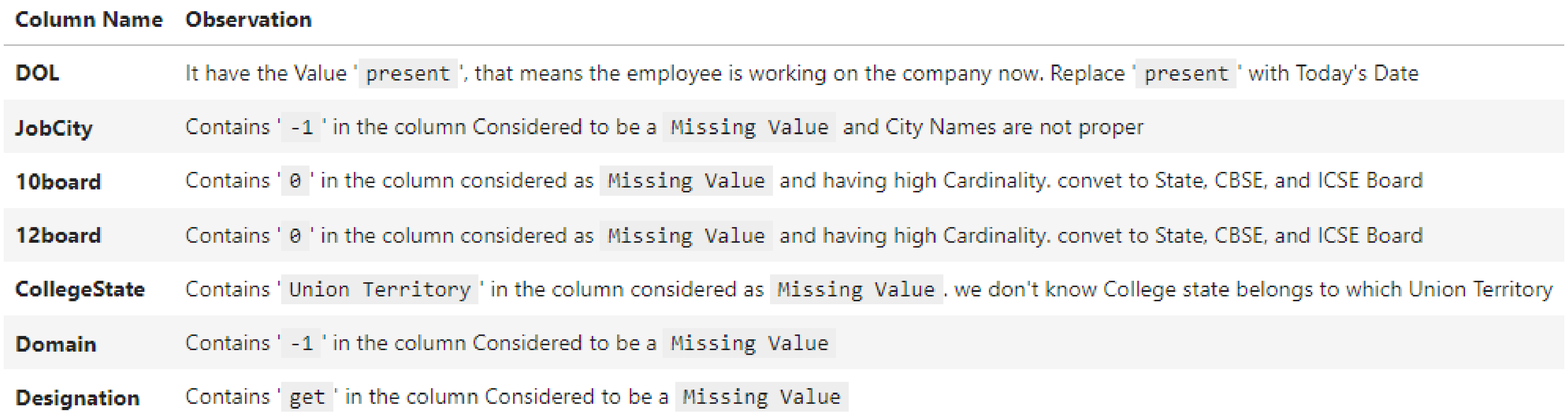
- **12board**: Board of education for 12th grade.
- **CollegeID**: Unique identifier for the college.
- **CollegeTier**: Tier of the college.
- **Degree**: Degree obtained.
- **Specialization**: Field of specialization.
- **collegeGPA**: GPA obtained in college.
- **CollegeCityID**: Unique identifier for the college city.
- **CollegeCityTier**: Tier of the college city.
- **CollegeState**: State where the college is located.
- **GraduationYear**: Year of graduation from college.
- **English**, **Logical**, **Quant**: Scores obtained in respective subjects.
- **Domain**: Domain knowledge score.
- **ComputerProgramming**, **ElectronicsAndSemicon**, **ComputerScience**, **MechanicalEngg**, **ElectricalEngg**, **TelecomEngg**, **CivilEngg**: Scores or qualifications in various engineering disciplines.
- **conscientiousness**, **agreeableness**, **extraversion**, **nueroticism**, **openess_to_experience**: Personality trait scores.

- The dataset offers a rich source of information for exploring employment outcomes, educational backgrounds, and personality traits among engineering graduates. With a diverse range of variables, it provides ample opportunities for indepth analysis and insights into factors influencing career trajectories.

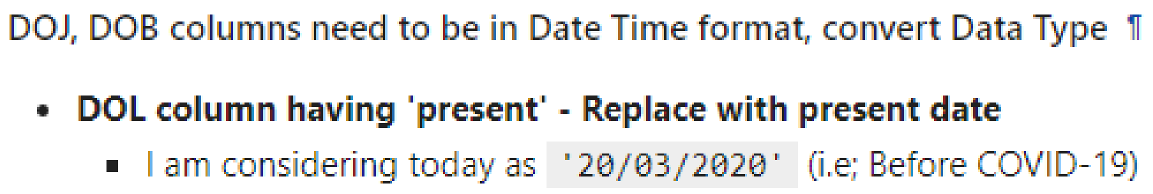
Exploratory Data Analysis (EDA)

Data Cleaning:

- After conducting preliminary assessments on the provided data, it has come to my attention that there are some irregularities present within the dataset.



Column Name	Observation
DOL	It have the Value '`present`', that means the employee is working on the company now. Replace '`present`' with Today's Date
JobCity	Contains '`-1`' in the column Considered to be a `Missing Value` and City Names are not proper
10board	Contains '`0`' in the column considered as `Missing Value` and having high Cardinality. convet to State, CBSE, and ICSE Board
12board	Contains '`0`' in the column considered as `Missing Value` and having high Cardinality. convet to State, CBSE, and ICSE Board
CollegeState	Contains '`Union Territory`' in the column considered as `Missing Value`. we don't know College state belongs to which Union Territory
Domain	Contains '`-1`' in the column Considered to be a `Missing Value`
Designation	Contains '`get`' in the column Considered to be a `Missing Value`

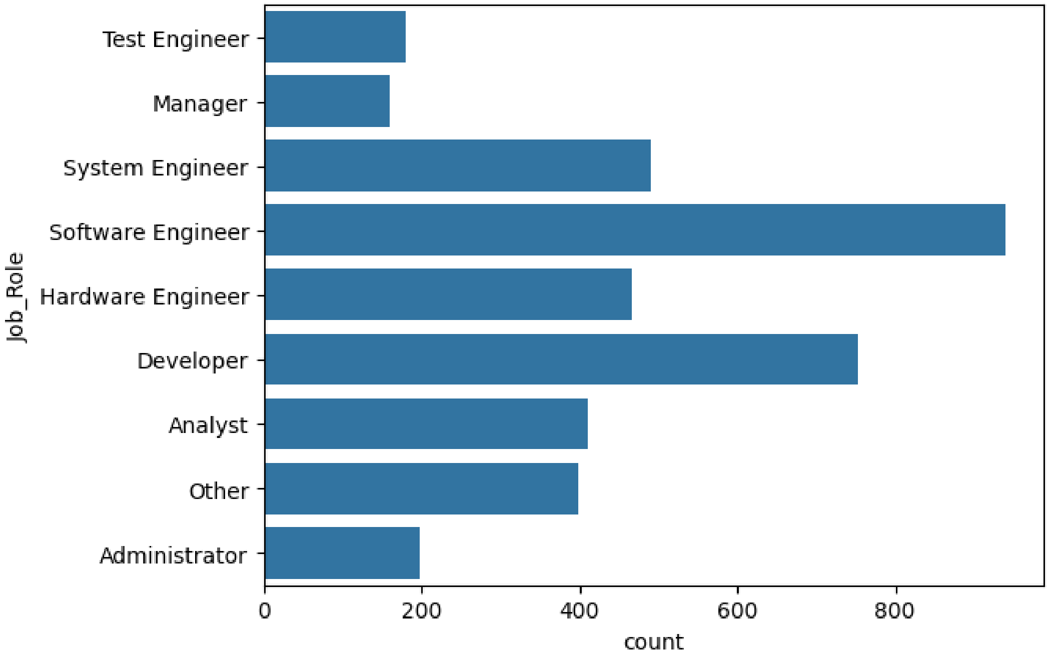


DOJ, DOB columns need to be in Date Time format, convert Data Type ¶

- **DOL column having 'present' - Replace with present date**
  - I am considering today as `'20/03/2020'` (i.e; Before COVID-19)

Univariate Analysis:

Moving forward, we'll use the **Job_Role** column derived from **Designation** column for deeper exploration and informed decisionmaking.

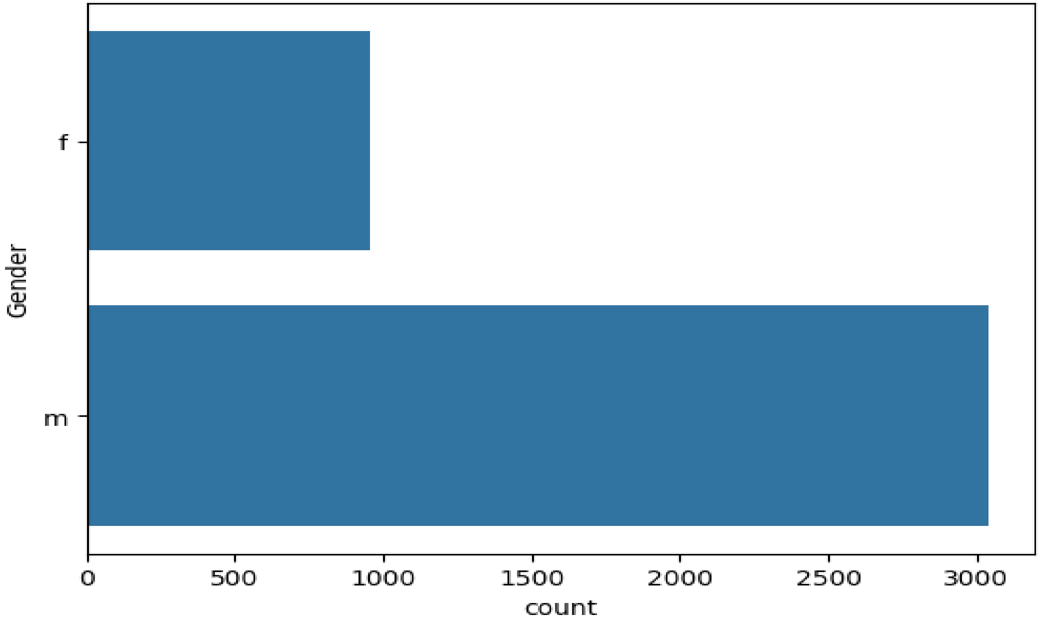


Following an analysis of the **Designation** column, I observed high cardinality, making analysis challenging.

Thus, I reclassified it into broader categories for easier interpretation.

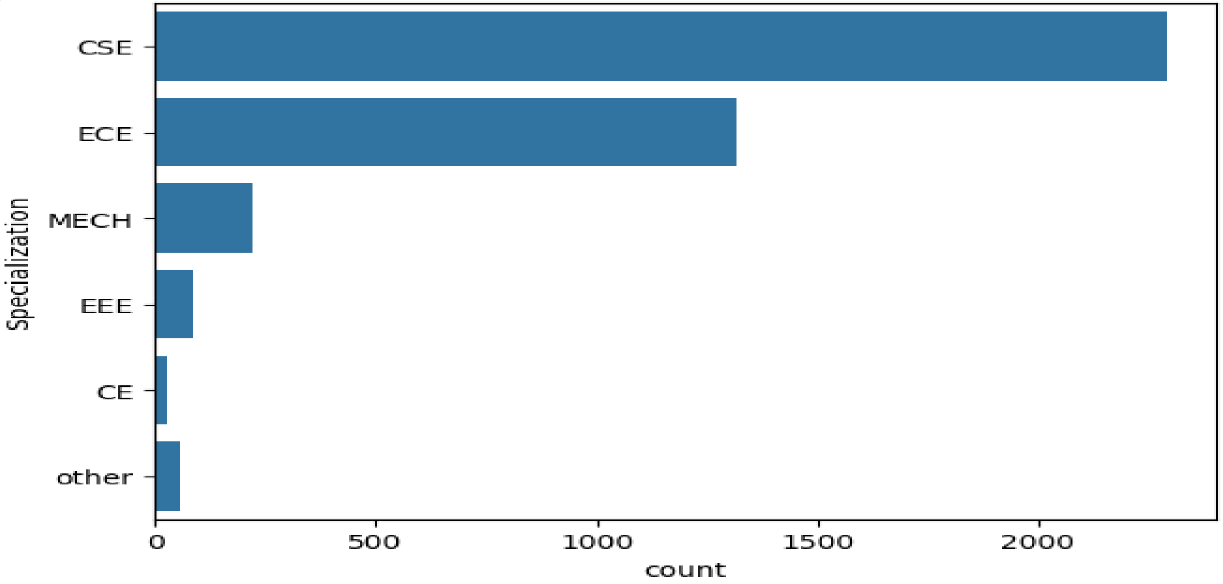
The **most prevalent roles are Software Engineer and Developer**.

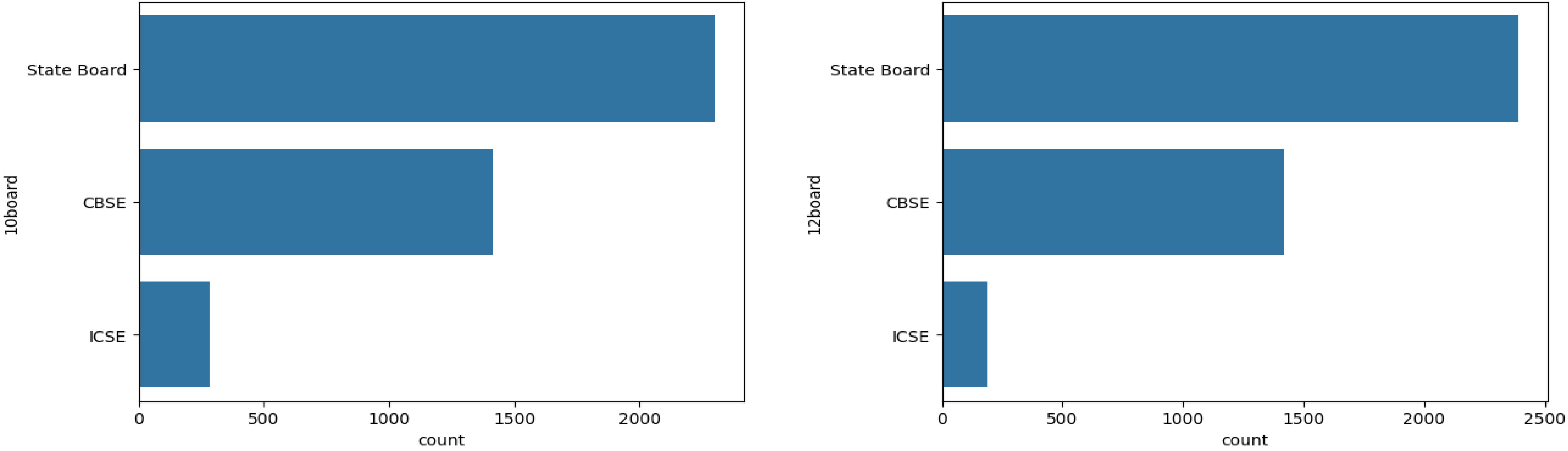
This adjustment allows for more structured analysis, providing insights into job role distributions and their impacts on various aspects like salary and performance.



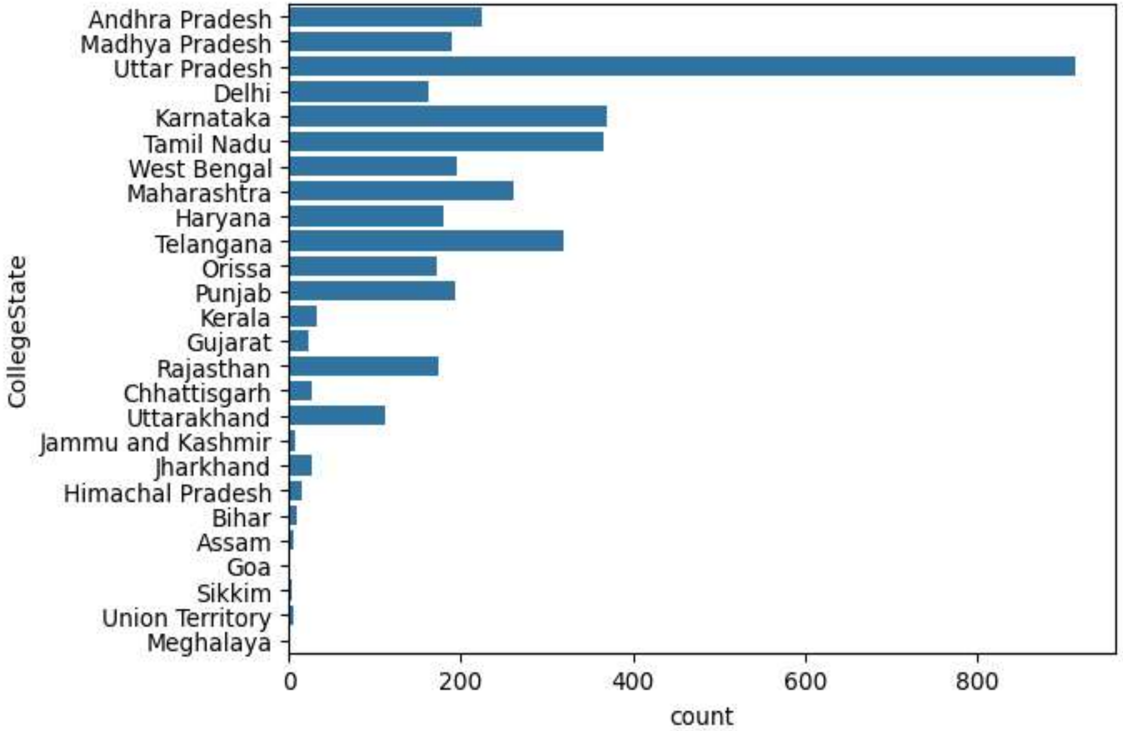
- Analysis of the Gender column reveals a substantial gender gap, with a proportion of approximately 1 female for every 3 males.
- This highlights the need for initiatives promoting gender diversity and inclusion in the workplace to ensure equal opportunities and representation.

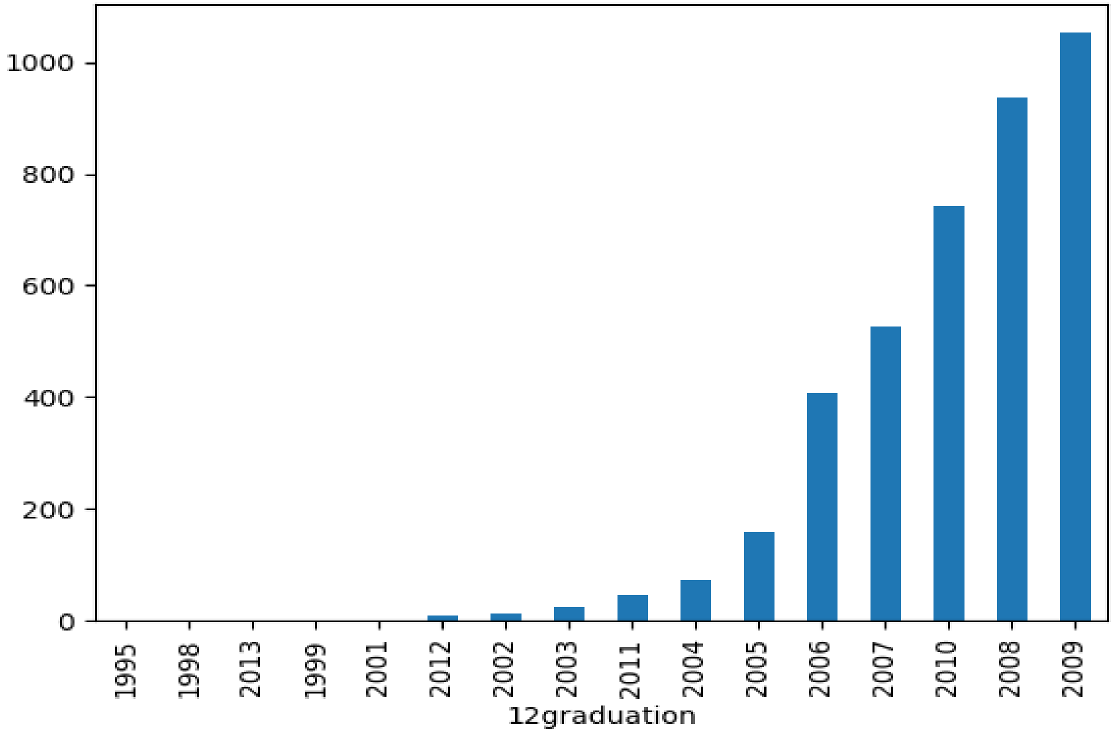
- Data suggests a prevalence in Computer Science and Engineering (CSE) specialization, followed by Electronics and Communication Engineering (ECE), Mechanical Engineering (MECH), Electrical and Electronics Engineering (EEE), and Civil Engineering (CE). These trends shape recruitment and curriculum strategies in the engineering sector.



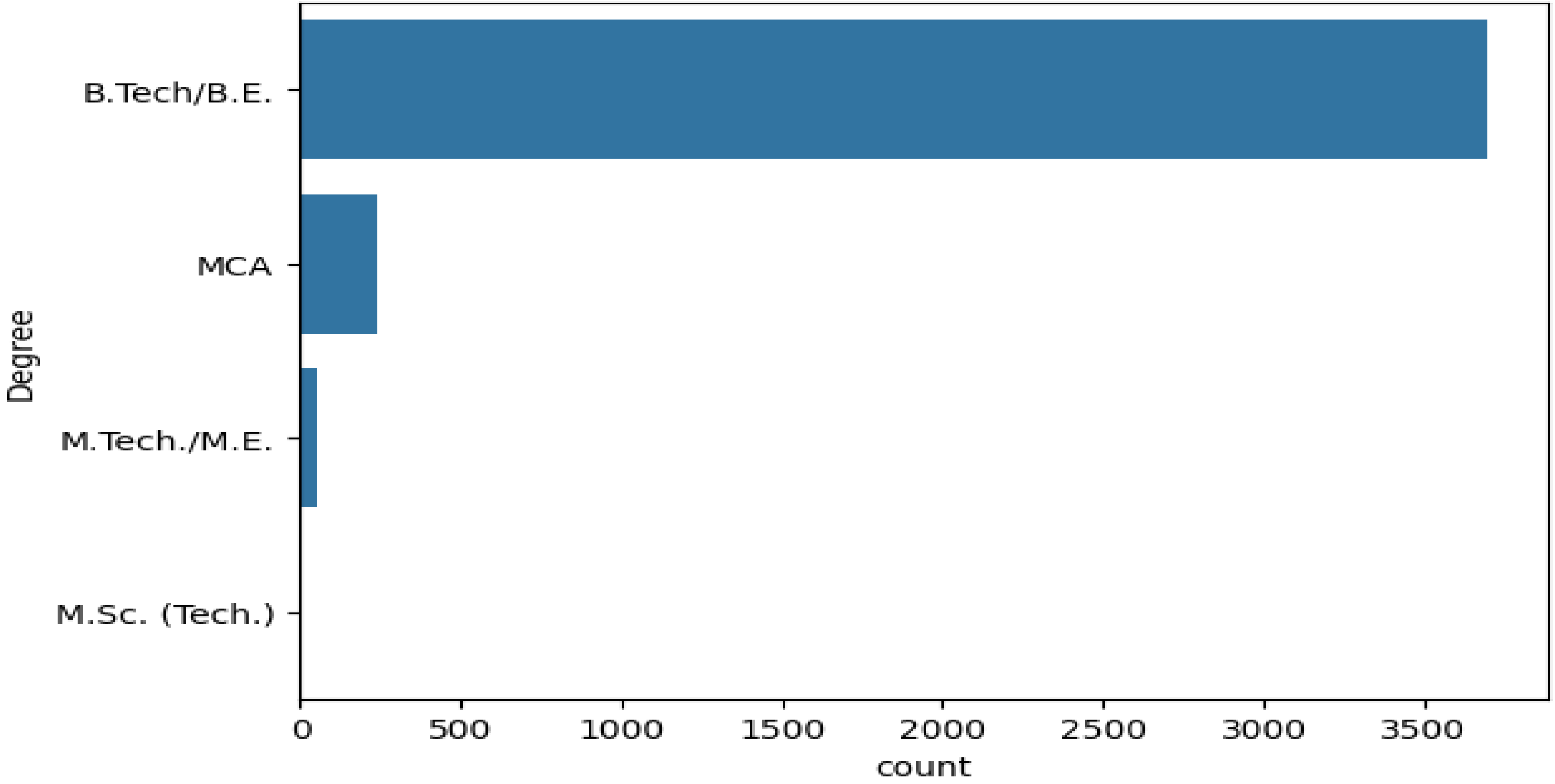


- State Boards dominate as the preferred examination boards for both 10th and 12th grades than , CBSE and ICSE Boards.
- These trends reflect a widespread preference among individuals surveyed.
- Insights into board preferences inform educational policies and curriculum development to better serve diverse student backgrounds.



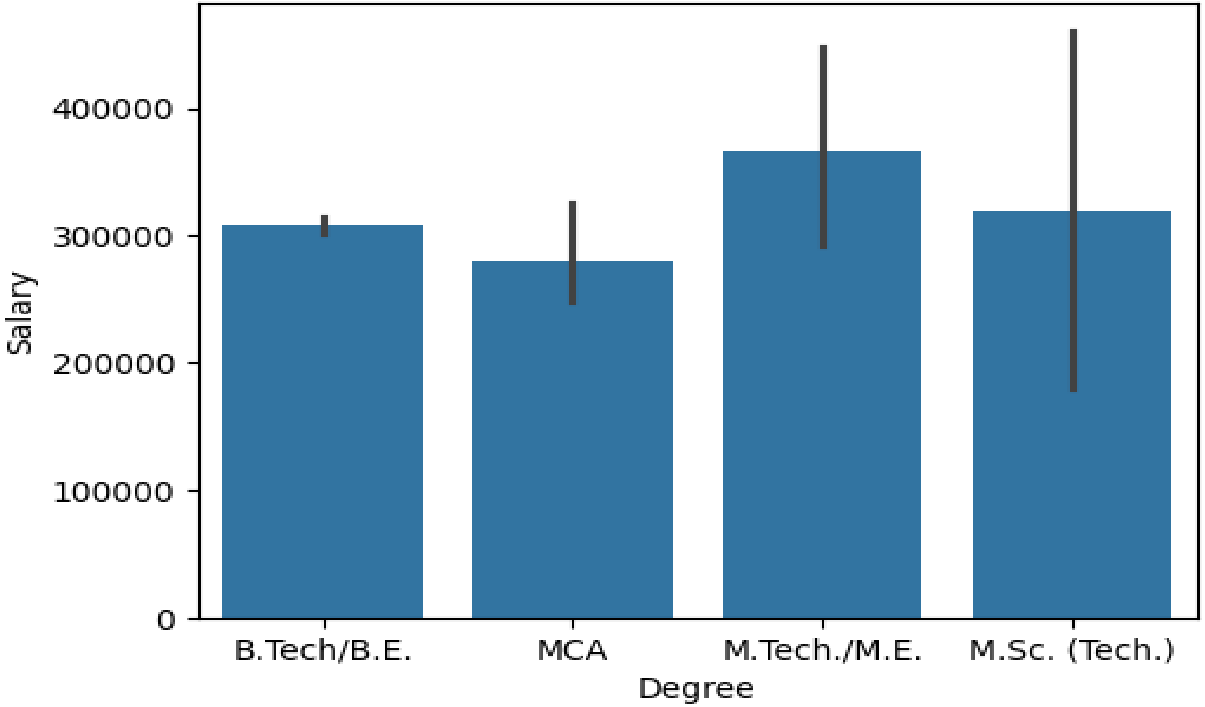


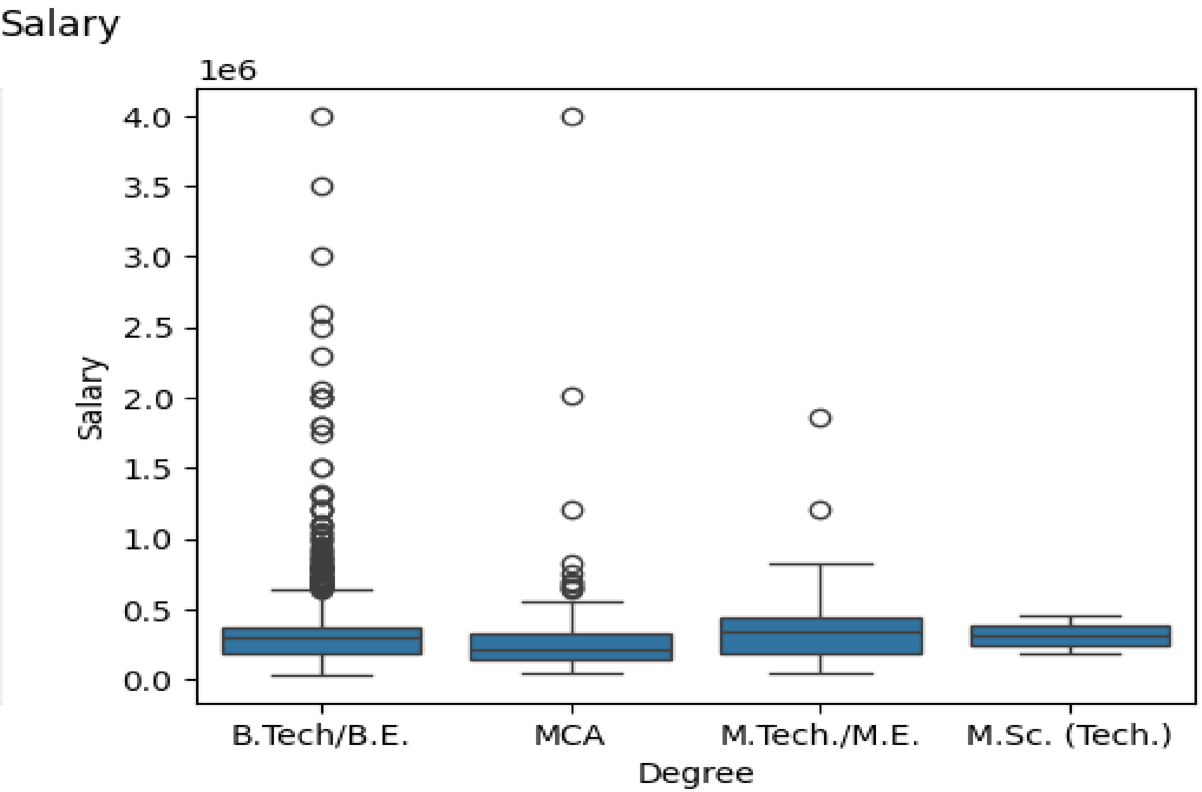
- A significant majority, approximately **70%**, completed their graduation between **2006** and **2009**.
- Colleges from **Uttar Pradesh** dominate, indicating a strong presence of graduates from this region.
- Following **Uttar Pradesh, Karnataka** and **Tamil Nadu** are notable for their college representation.
- Conversely, **Meghalaya** and **Goa** have fewer graduates, suggesting lower college participation rates.
- These insights provide valuable guidance for educational planning and resource allocation across diverse regions.



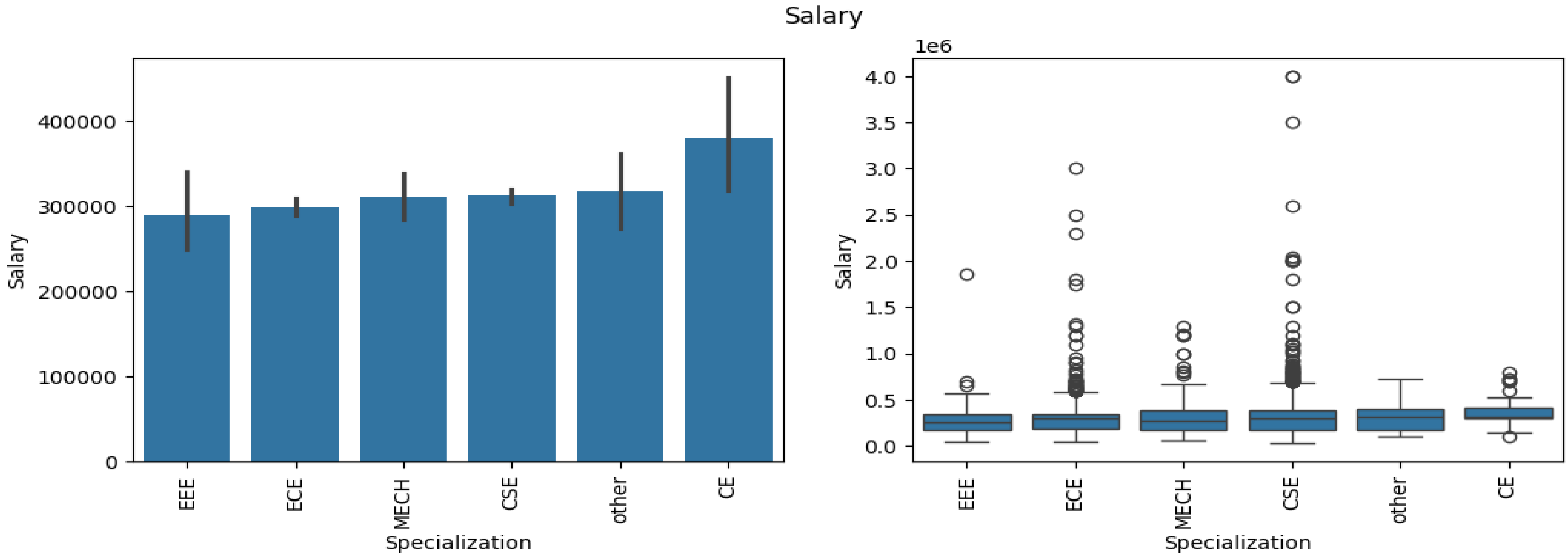
- The predominant qualification among students is **Bachelor of Technology/Engineering (B.Tech/B.E).**
- Following closely, **Master of Computer Applications (MCA)** emerges as the second most prevalent qualification.
- These findings indicate a strong inclination towards technical fields among the surveyed individuals.
- Insights into qualification preferences inform educational institutions and employers in tailoring programs and career opportunities to align with student aspirations.

- Bivariate Analysis:

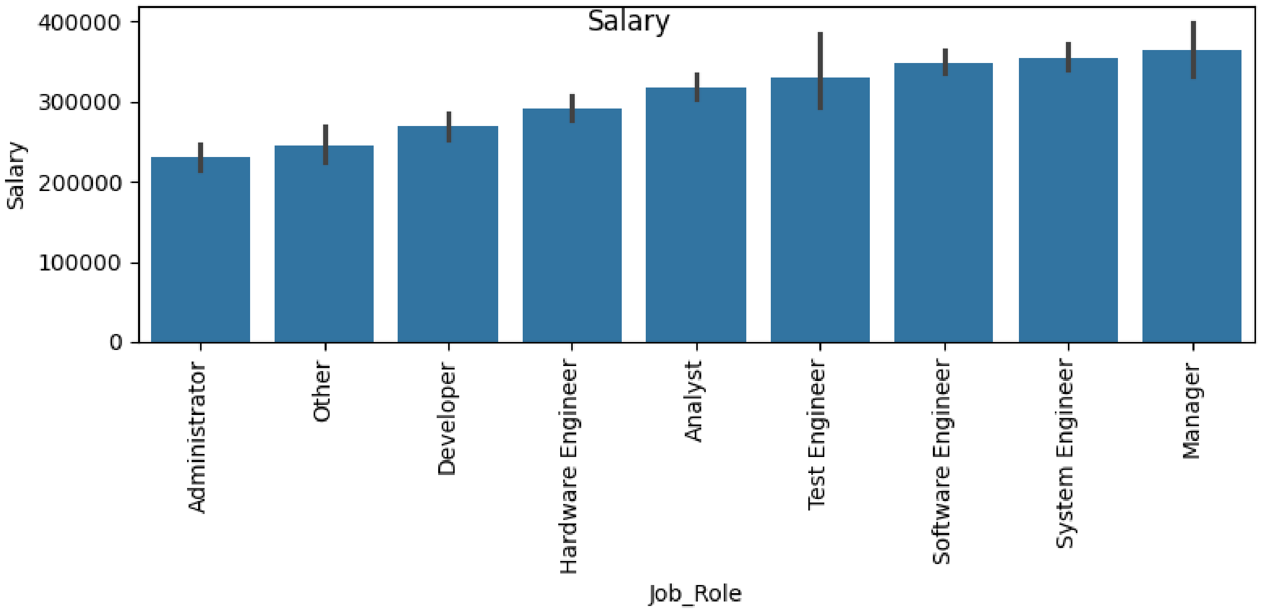


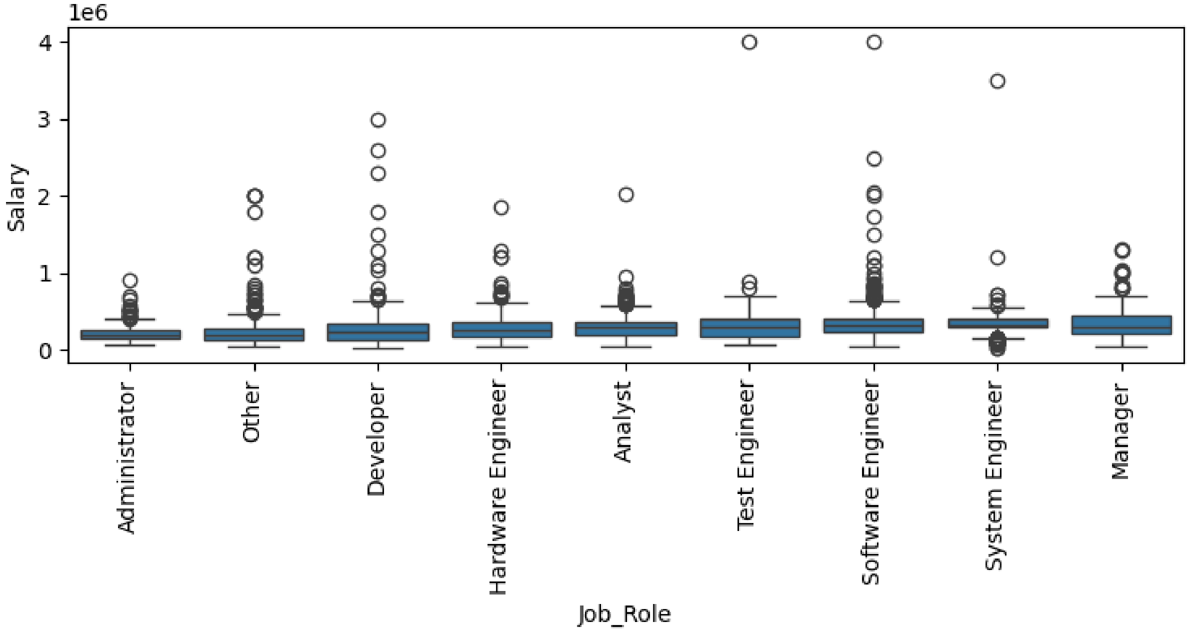


- **M.Tech/M.E** graduates generally earn higher average salaries compared to others.
- Despite this, **B.Tech/B.E** graduates have a greater likelihood of earning better than **M.Tech/M.E** graduates overall.
- This suggests that while **M.Tech/M.E** qualifications may lead to higherpaying roles in some cases, **B.Tech/B.E** graduates enjoy a broader range of earning opportunities in the job market.

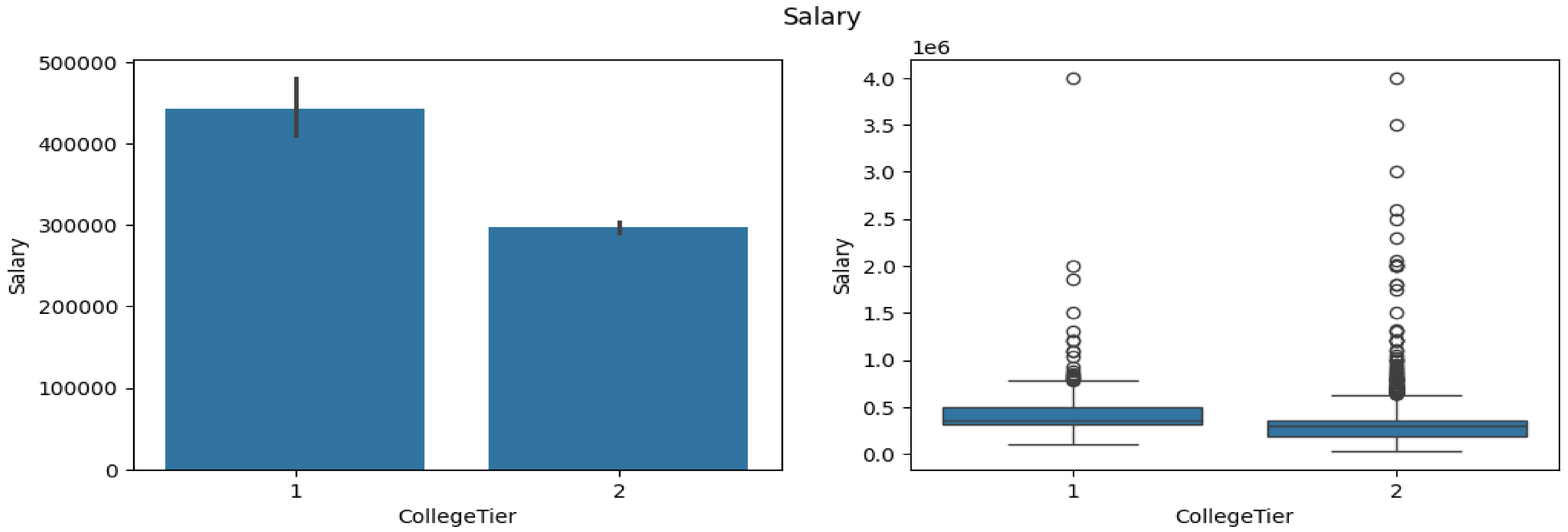


- **CSE** graduates typically command **higher salaries** compared to their counterparts from other disciplines, indicating a strong demand for their skill set in the job market.
- The majority of students surveyed are pursuing Bachelor of Technology/Engineering (**B.Tech/B.E**) degrees, with Master of Computer Applications (**MCA**) as the second most prevalent qualification.
- These findings underscore the importance of technical expertise, particularly in computer science, for maximizing earning potential and career opportunities in today's job market.



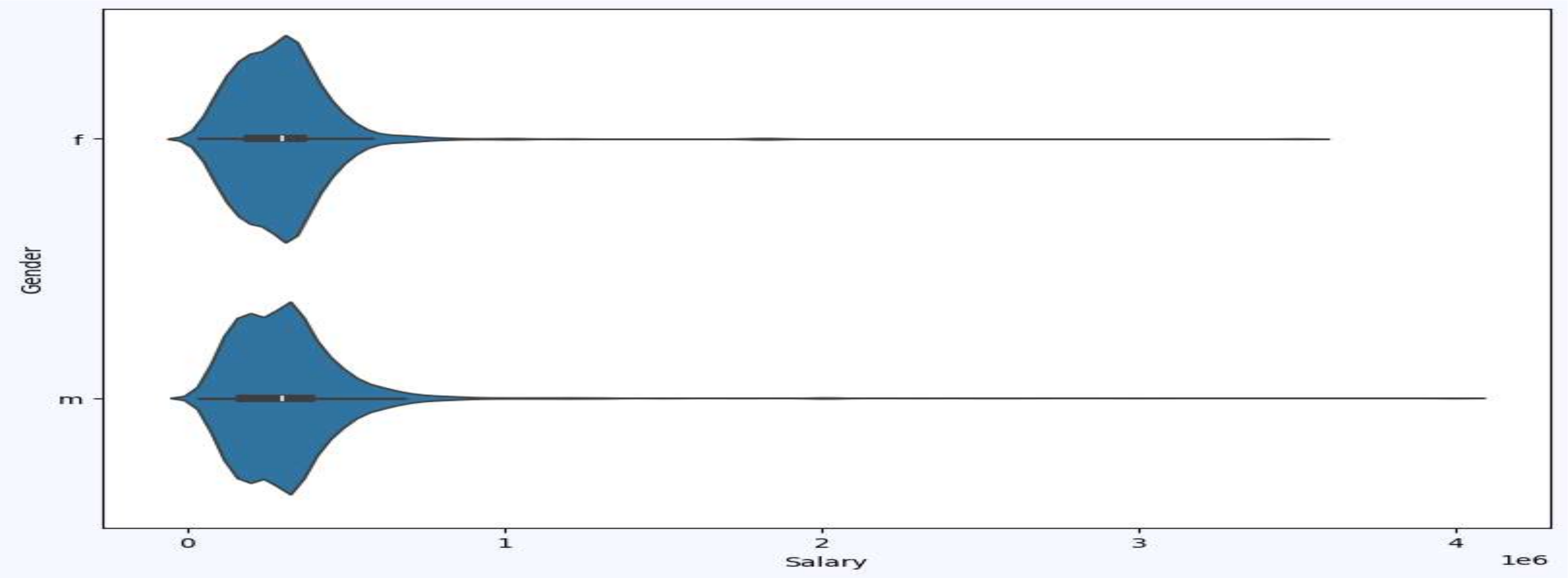


- **Managers** emerge as the **highest earners** according to the graph, indicating the lucrative nature of managerial positions within the dataset.
- Following closely, **System Engineers** represent the second highest earners, underscoring the significant earning potential associated with technical roles.
- These observations emphasize the importance of both managerial and technical skills in achieving higher earning potential within the workforce.

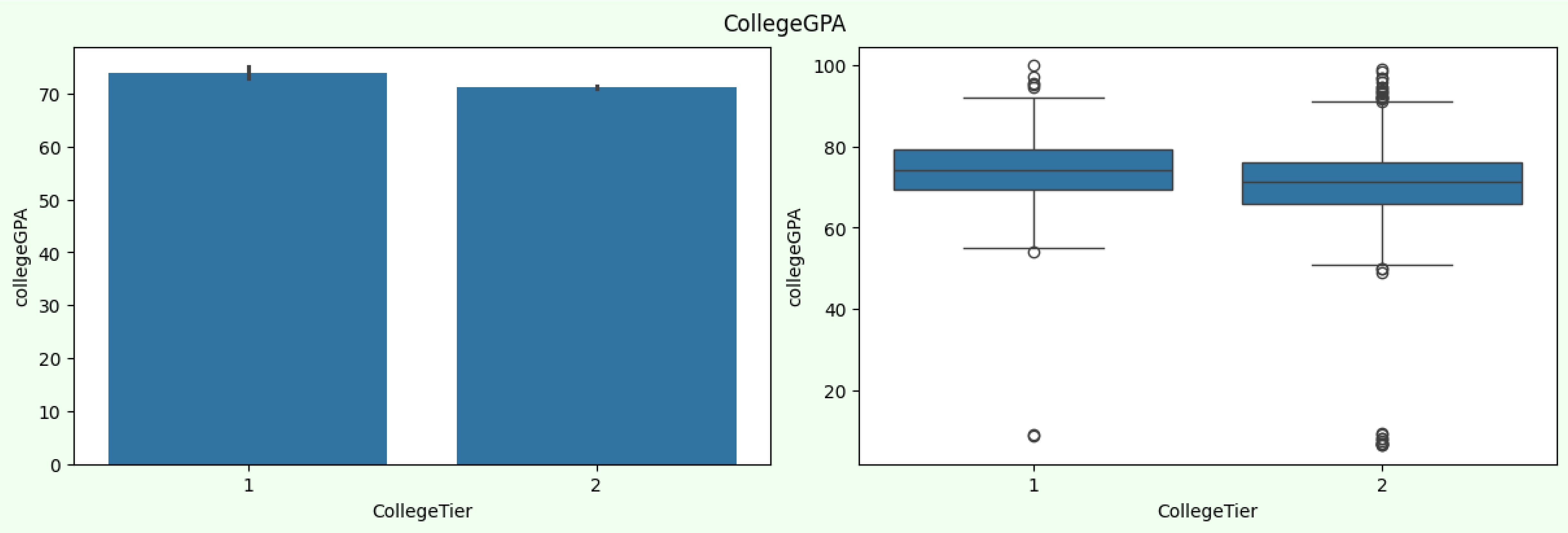


- Individuals from **Tier1** colleges exhibit **higher earnings** compared to their counterparts from Tier2 institutions, reflecting the perceived value and prestige associated with Tier1 educational institutions.
- The data underscores the significant impact of college tier on earnings potential, highlighting the advantages afforded to graduates from Tier1 colleges in terms of career advancement and salary prospects.

INNOMATICS RESEARCH LABS



- The observation reveals a subtle variance in median salary between **female** and **male** individuals, suggesting a potential gender disparity in earnings, although the extent of this difference remains uncertain.
- While the difference in **median salary appears minor**, further analysis is required to determine the significance of this gap and to address any underlying factors contributing to potential genderbased discrepancies in earnings.

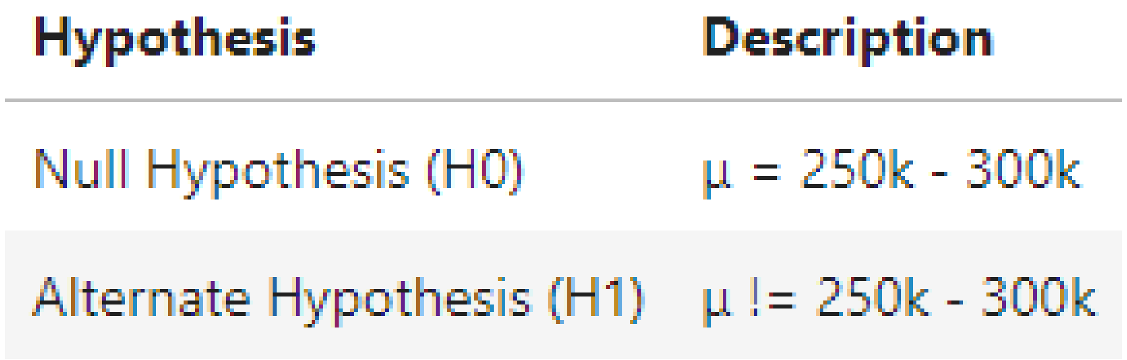


- Students from **Tier1** colleges exhibit **slightly higher** performance levels compared to those from **Tier2** institutions, indicating a potential correlation between college tier and academic achievement.
- This observation suggests that attending a Tier1 college may provide students with **additional resources or opportunities** that contribute to enhanced academic performance, highlighting the potential advantages associated with highertier educational institutions.

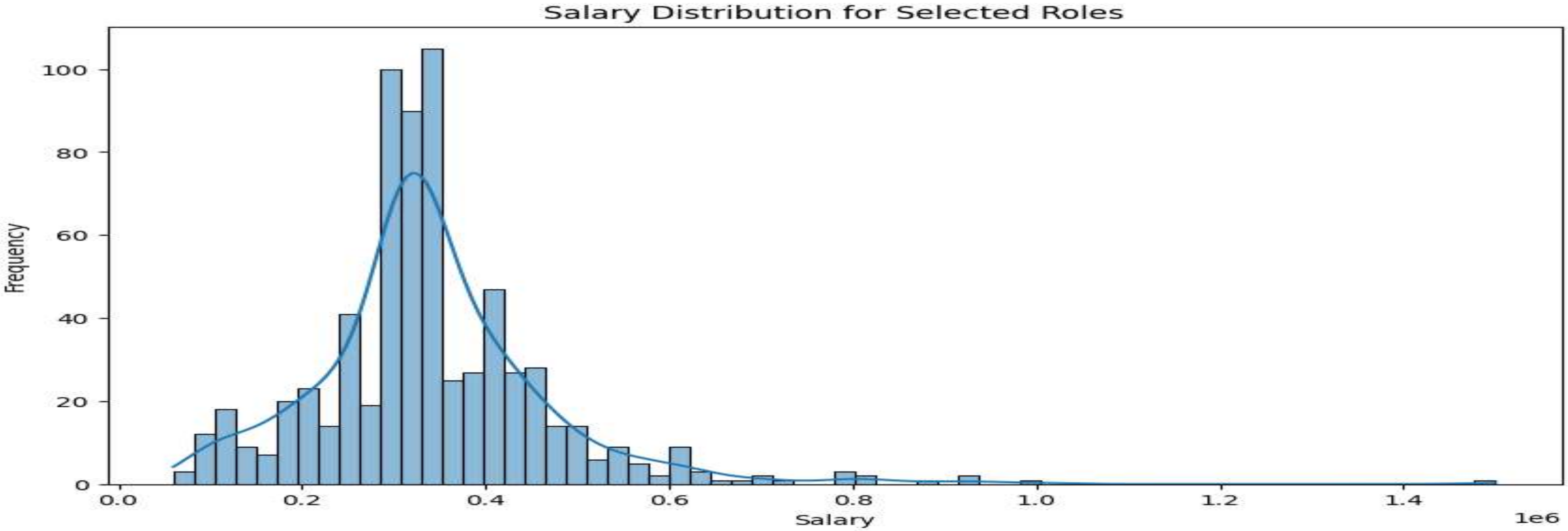
INNOMATICS RESEARCH LABS

Research Questions

- 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 2.53 lakhs as a fresh graduate.” Test this claim with the data given to you.
- Is there a relationship between gender and specialization? (i.e. Does the preference of Specialization depend on the Gender?)
- **Let's Verify the claim**
- **Defining hypothesis**



Hypothesis	Description
Null Hypothesis (H0)	µ = 250k - 300k
Alternate Hypothesis (H1)	µ != 250k - 300k



- **The data does not support the claim made in the Times of India article** that fresh graduates in Computer Science Engineering can earn between **2.5 3 lakhs** in roles such as **Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer.**
- There is no significant relationship between gender and specialization preference observed in the data, as the claim that gender influences specialization preference is not supported by the analysis.

Conclusions:

- The data reveals a gender imbalance, indicating a need for diversity efforts in the workforce.
- Technical skills, particularly in **Computer Science and Engineering**, are in high demand based on the prevalence of related degrees.
- Job roles vary widely, with **Software Engineer** being the most common, followed by **Developer**.
- Educational board preferences influence policies, with a preference for **State Boards**, **CBSE**, and **ICSE**.
- Technical expertise is crucial, as evidenced by the prevalence of **Bachelor of Technology/Engineering** graduates.
- **Managerial** and **technical positions** are the highest earning roles.
- College tier impacts earnings, with **Tier1** graduates earning more.
- Gender based salary differences exist, though further analysis is needed for clarity.
- The claim about recent graduates' earnings in **Computer Science Engineering** was **not supported by the data**.
- There's **no significant link between gender and specialization preference**, challenging assumptions about their correlation.

THANK YOU !