INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis-On AMCAT data

About me

- Hello, my name is Prashanth, a dedicated data enthusiast who has recently completed a comprehensive Data Science course and is currently engaged in an enriching internship. While I may be considered a newcomer to the field, my fervent passion for unlocking insights from data knows no bounds.
- With a solid foundation in Data Science principles and practical experience gained through my internship, I stand ready to leverage my skills to extract meaningful insights from complex datasets. My aim is to contribute positively to our collective understanding of data-driven decision-making.
- I invite you to connect with me as I embark on this professional journey. Together, we can explore the vast landscape of data analysis and harness its potential to drive innovation and informed decision-making.

www.linkedin.com/in/prashantheleshala

OBJECTIVE:

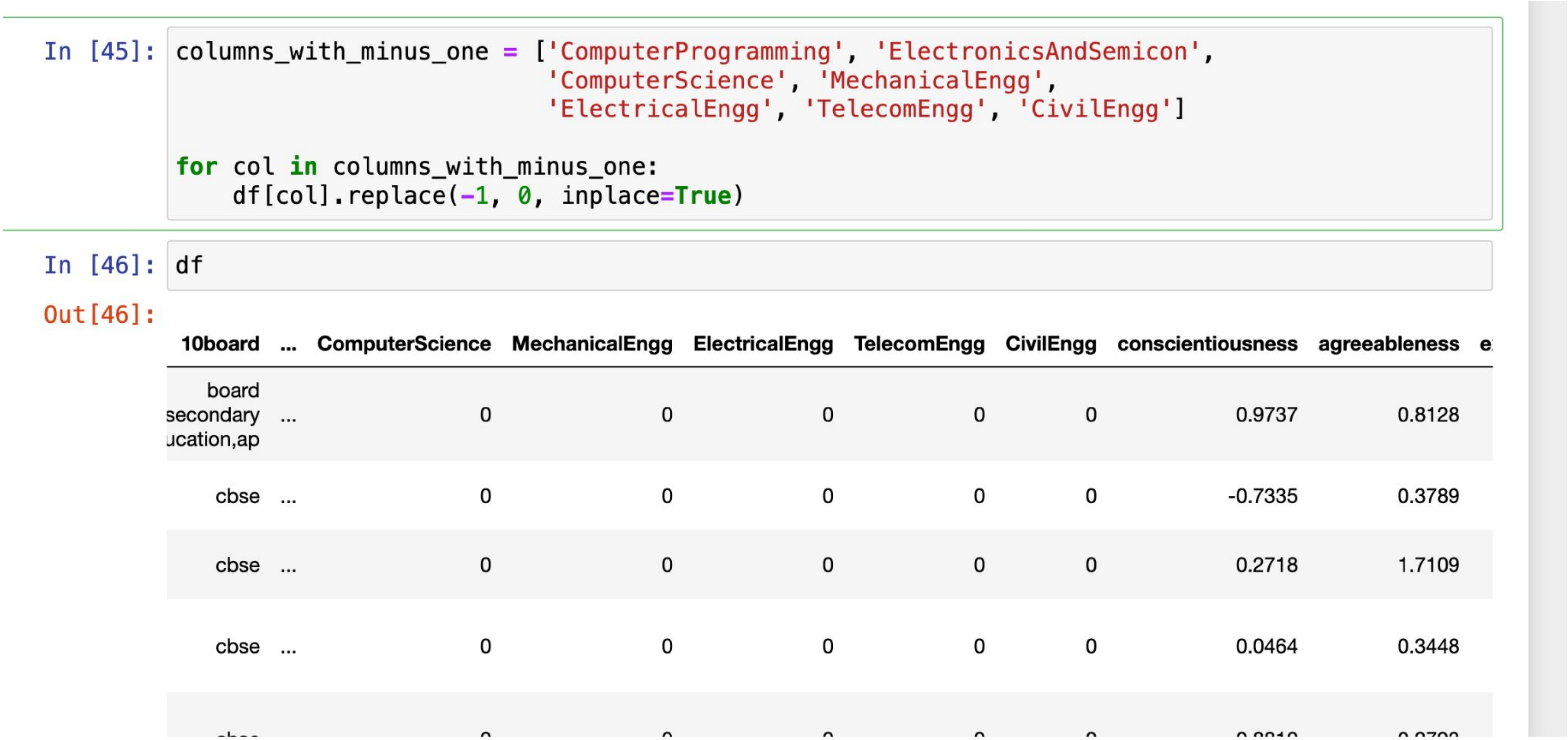
- Provide a thorough description of the dataset and its characteristics.
- Investigate the connections between independent variables and the target variable (Salary).
- Gender Disparities Exploration: Analyze gender gaps in salaries, roles, and specializations.
- Specialization Comparison: Compare salaries, roles, and across engineering specializations.
- Interesting Observations Generation: Identify and present intriguing insights from the data.

DATA DESCRIPTION:

- 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 PREPROCESSING

- Removing Unwanted Columns
- Data Type Conversion
- Describing the data with .describe function and getting information about the data with .info function
- Thus, removing the unwanted columns,adding new column i.e Age and converting data type of DOL.



```
In [45]: columns_with_minus_one = ['ComputerProgramming', 'ElectronicsAndSemicon',
                                   'ComputerScience', 'MechanicalEngg',
                                   'ElectricalEngg', 'TelecomEngg', 'CivilEngg']

         for col in columns_with_minus_one:
             df[col].replace(-1, 0, inplace=True)
```

```
In [46]: df
```

Out[46]:

10board	...	ComputerScience	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	conscientiousness	agreeableness	e
board secondary ucation,ap	...	0	0	0	0	0	0.9737	0.8128	
cbse	...	0	0	0	0	0	-0.7335	0.3789	
cbse	...	0	0	0	0	0	0.2718	1.7109	
cbse	...	0	0	0	0	0	0.0464	0.3448	
cbse		0	0	0	0	0	[illegible]	[illegible]	

preprocessing step to handle missing or undefined data represented by -1, converting them to a neutral value of 0.

Univariate Analysis:

- Univariate analysis conducted encompasses both categorical and numerical variables, providing a comprehensive overview of the dataset's characteristics and distributions.
- Both categorical and numerical univariate analyses focus on understanding the individual characteristics and distributions of variables within a dataset.

Observations Summary:

Designation:

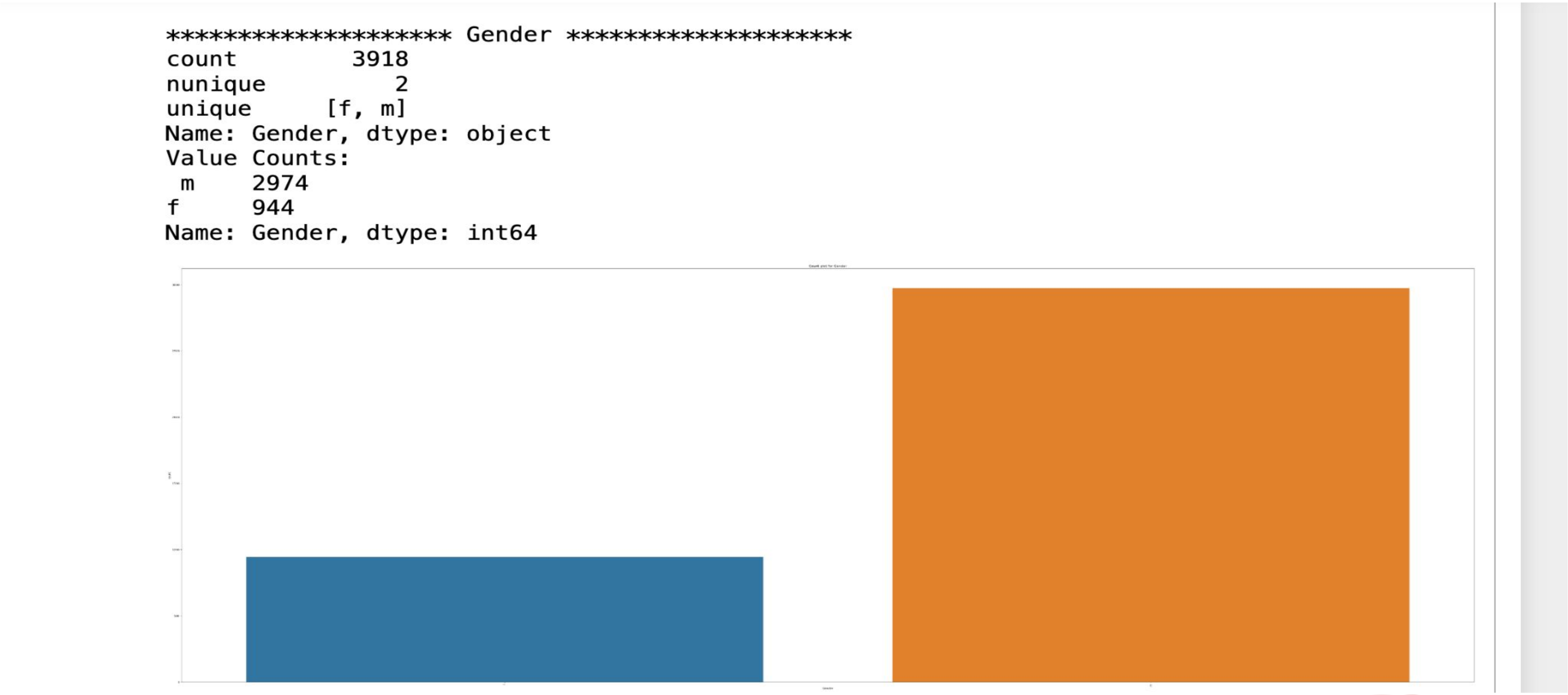
- There are 416 unique designations.
- The most common designation is "software engineer" with 529 occurrences.
- The count plot shows the distribution of various designations.

JobCity:

- There are 175 unique job cities.
- "BANGALORE" is the most common job city with 671 occurrences.
- The count plot illustrates the distribution of job cities.

Gender:

- There are two unique gender categories: 'm' (male) and 'f' (female).
- There are 2974 males and 944 females.
- The count plot visualizes the distribution of genders.



```
******************** Gender ********************
count        3918
nunique         2
unique     [f, m]
Name: Gender, dtype: object
Value Counts:
 m     2974
f     944
Name: Gender, dtype: int64
```

10 board:

- There are 271 unique 10th board types.
- "cbse" is the most common 10th board with 1362 occurrences.
- The count plot displays the distribution of different 10th boards.

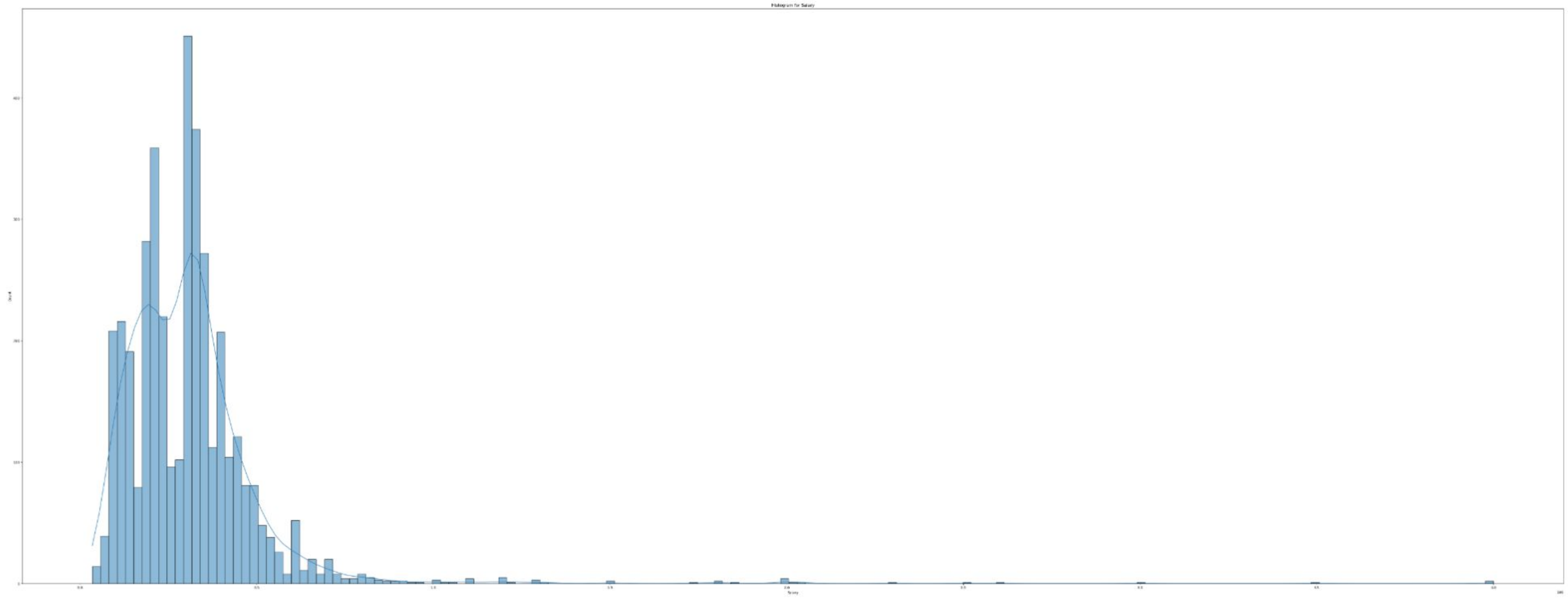
12 board:

- There are 335 unique 12th board types.
- "cbse" is the most common 12th board with 1365 occurrences.
- The count plot illustrates the distribution of different 12th boards.

Numerical Analysis:

- Investigated numerical variables such as 'ID', 'Salary', '10percentage', '12graduation', '12percentage', 'CollegeID', 'CollegeTier', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier', 'GraduationYear', 'English', 'Logical', 'Quant', 'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion', 'neuroticism', 'openess_to_experience', and 'Age'.
- Computed descriptive statistics including minimum, maximum, mean, median, and standard deviation for each numerical variable, providing insights into the central tendency and spread of the data.
- Visualized the distribution of numerical data using histograms, allowing for a visual understanding of the data distribution.

```
******************** Salary ********************
min       3.500000e+04
max       4.000000e+06
mean      3.088333e+05
median    3.000000e+05
std       2.135426e+05
Name: Salary, dtype: float64
```



Bivariate Analysis:

- Bivariate analyses provide valuable insights into the relationships between gender and specialization choices, as well as the influence of college tier and degree type on salary levels.

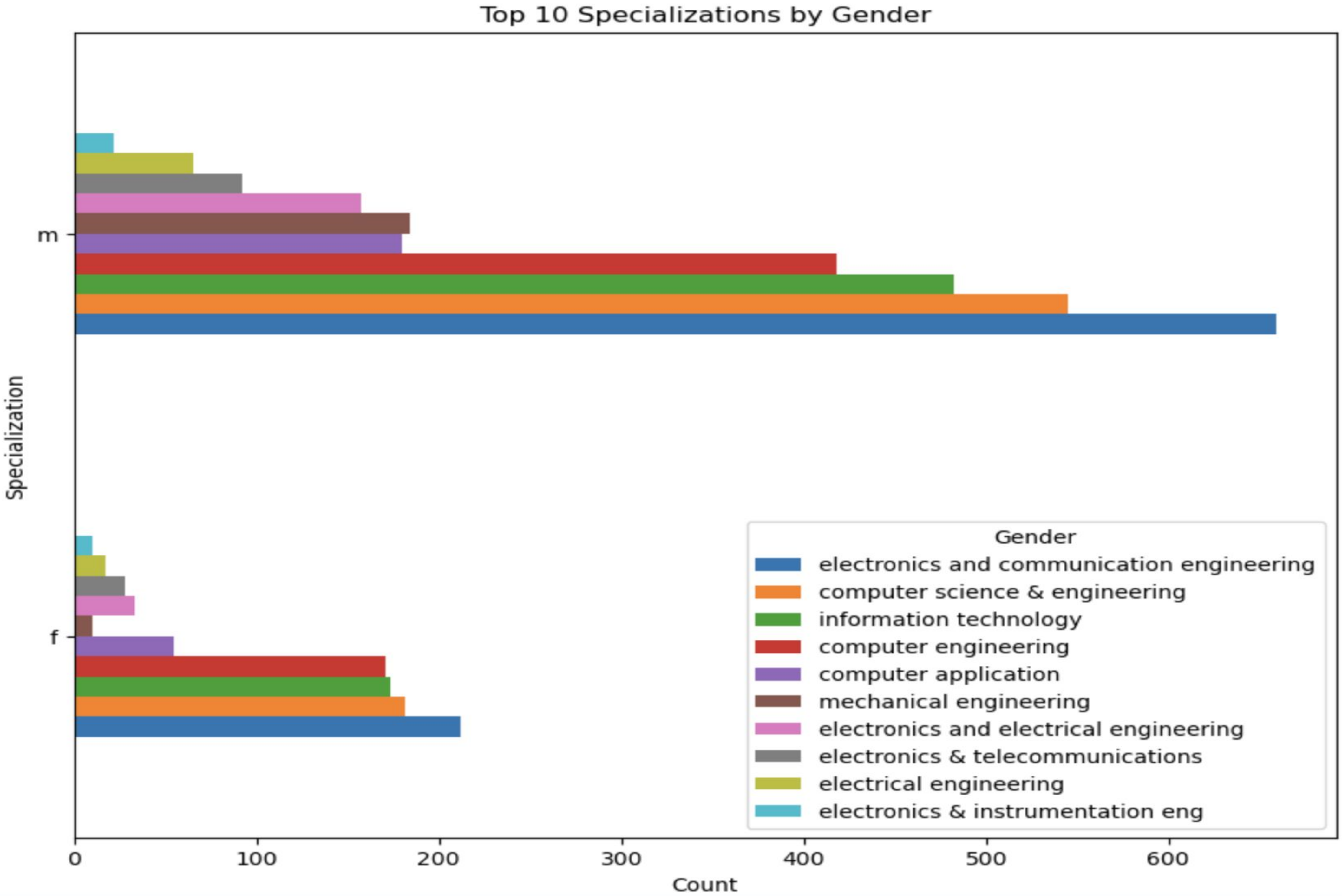
Observations Summary:

Gender vs. Specialization Crosstab:

- The crosstabulation between the 'Gender' and 'Specialization' columns is created, showing the count of individuals for each combination of gender and specialization.
- The data is sorted based on the sum of values in each specialization, and the top 10 specializations are selected and plotted in a horizontal bar chart.

Observations:

- The visualization shows the distribution of top specializations by gender.
- It helps identify gender-based trends in specialization choices.

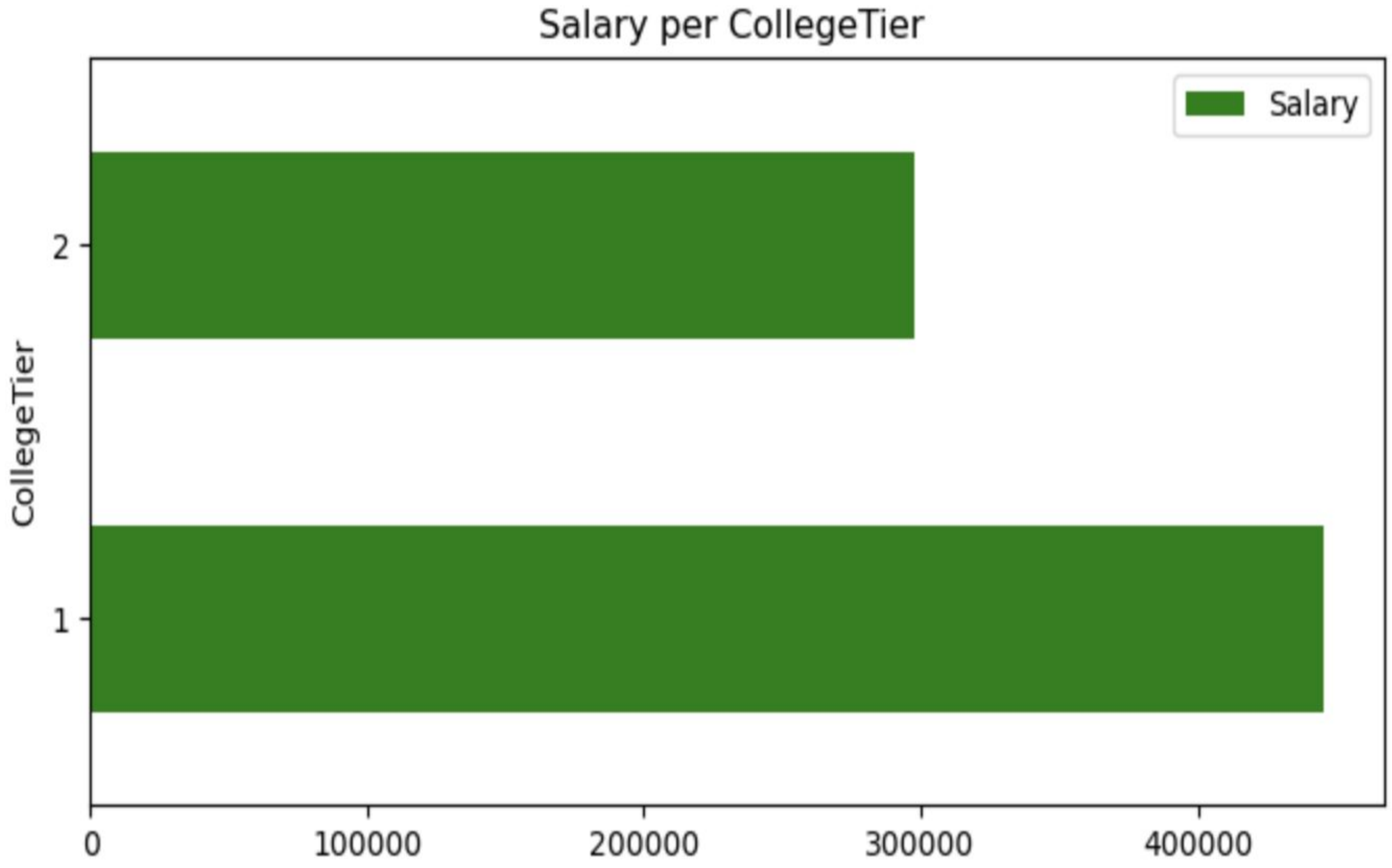

Top 10 Specializations by Gender
Specialization
m
f
Count
0
100
200
300
400
500
600
Gender
electronics and communication engineering
computer science & engineering
information technology
computer engineering
computer application
mechanical engineering
electronics and electrical engineering
electronics & telecommunications
electrical engineering
electronics & instrumentation eng

Salary per College Tier:

- A pivot table is created where the index is the 'CollegeTier', and the values are the 'Salary'.
- The average salary for each college tier is calculated and plotted in a horizontal bar chart.

Observations:

- It provides insights into the average salary based on the college tier.
- Helps in understanding if there's a significant difference in salaries between different college tiers.
- Tier Disparity: Graduates from Tier 1 colleges tend to receive higher salaries compared to those from Tier 2 colleges.

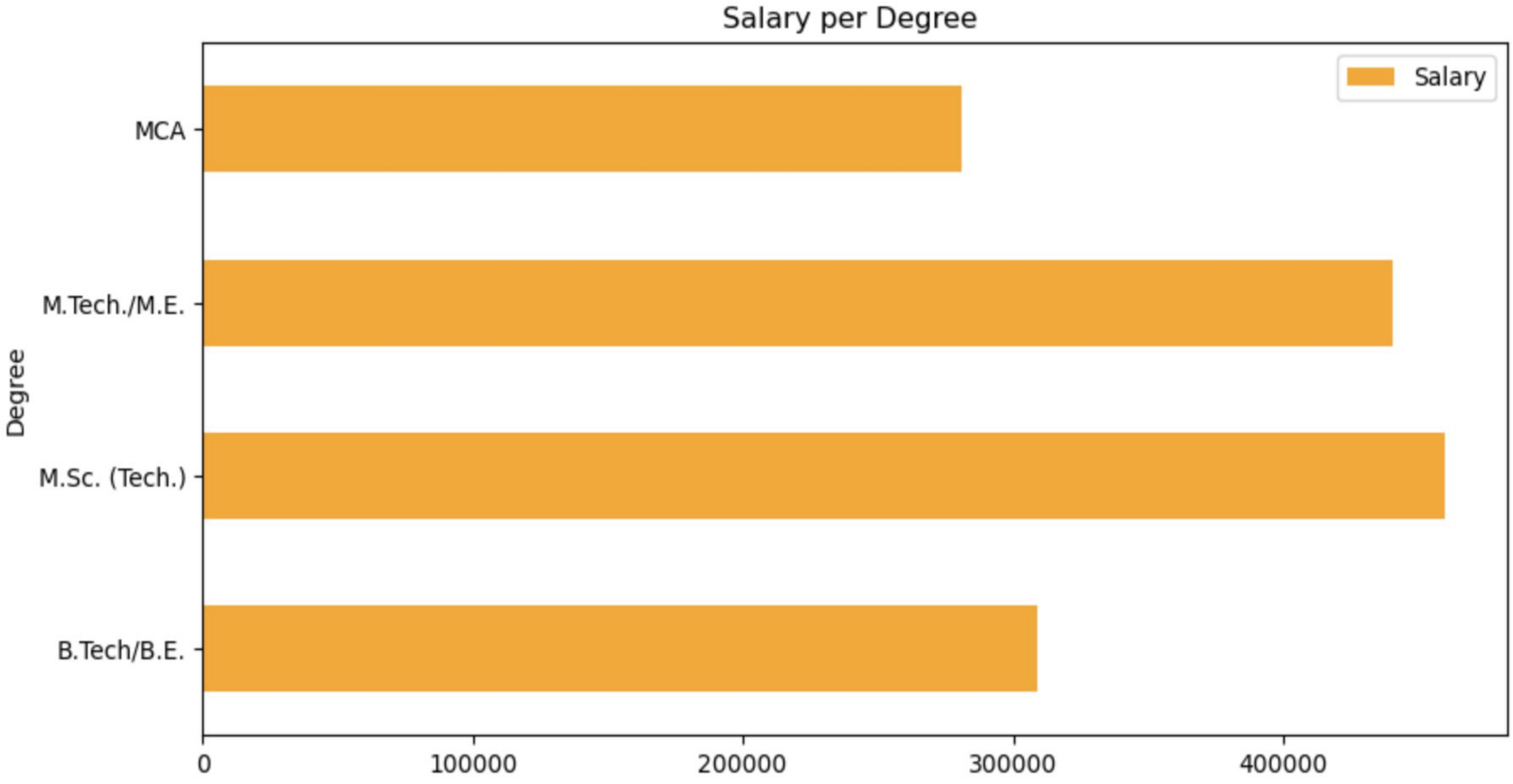

Salary per CollegeTier
🟩 Salary
CollegeTier
2
1
0
100000
200000
300000
400000

Salary per Degree:

- Another pivot table is created with the index as 'Degree' and the values as 'Salary'.
- The average salary for each degree type is calculated and plotted in a horizontal bar chart.

Observations:

- It helps in understanding the average salary differences among various degree types.
- Offers insights into which degree might be more lucrative in terms of salary.
- Though the Btech/B.E is done by majority the higher salaries are drawn out from further education
- Point to note: The dataset exhibits a notable scarcity of data points for individuals with master's degrees compared to other educational levels.


Salary per Degree
Salary
Degree
MCA
M.Tech./M.E.
M.Sc. (Tech.)
B.Tech/B.E.
0
100000
200000
300000
400000

RESEARCH QUESTION:

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.” Test this claim with the data given to you.

- The confidence interval calculated indicates that with 95.0% confidence, the true population mean salary for Computer Science Engineering graduates in the specified roles lies within the range of 287978.74 lakhs to 357621.26 lakhs.

- Therefore, the claim made in the Times of India article regarding the salary range of 2.5-3 lakhs for fresh Computer Science Engineering graduates appears to be inaccurate. The calculated confidence interval suggests that the actual salary range is significantly higher than the claim provided in the article.

- In summary, based on the analysis, there is strong evidence to suggest that the salary range stated in the Times of India article may not be representative of the actual salary range for fresh Computer Science Engineering graduates in the specified roles.

THANK YOU

INNOMATICS RESEARCH LABS