INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA AMCAT ANALYSIS

About me

- My name is Bejugam Divya & I Graduated from B Tech(CSE)
- Data science has the potential to improve the way we live and work, and it can empower others to make better decisions, solve problems, discover new advancements, and address some of the world's most pressing issues. And it help’s in business and industry strategies
- No any experience
- Linkedin: https://www.linkedin.com/in/divya-bejugam/
- github : https://github.com/DivyaBejugam

INNOMATICS RESEARCH LABS

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.

1.

THANK YOU

INNOMATICS
RESEARCH LABS

Introduction:
The Aspiring Minds Employment Outcome (AMEO) 2015 dataset captures the employment results of engineering graduates across India, offering a deep dive into how different skills and demographics affect career outcomes.

Problem Statement:
The primary goal is to analyze the factors influencing the employment outcomes of engineering graduates based on their cognitive, technical, and personality skills, as well as demographic factors. By studying these variables, we aim to determine key predictors for job titles, job locations, and salary, thereby providing insights into how skills and demographics impact career success in the engineering field.

Data Description:

- **Data Points:** 4000 records
- **Variables:** 40 independent variables (continuous and categorical)
- **Key Variables:** Cognitive, technical, and personality skills, demographic details, and employment outcomes (Salary, Job Titles, Job Locations)

Objective:
To analyze and identify the influence of various skills and demographic factors on the employment outcomes of engineering graduates, focusing on salary, job roles, and job locations.

Importing the dataset

```python
import pandas as pd
df = pd.read_csv(r"C:\Users\admin\Downloads\data.xlsx - Sheet1.csv" )
df
```

```
     Unnamed: 0      ID     Salary            DOJ           DOL  \
0         train  203097   420000.0   6/1/12 0:00        present
1         train  579905   500000.0   9/1/13 0:00        present
2         train  810601   325000.0   6/1/14 0:00        present
3         train  267447  1100000.0   7/1/11 0:00        present
4         train  343523   200000.0   3/1/14 0:00    3/1/15 0:00
...         ...     ...        ...           ...           ...
3993      train   47916   280000.0  10/1/11 0:00   10/1/12 0:00
3994      train  752781   100000.0   7/1/13 0:00    7/1/13 0:00
3995      train  355888   320000.0   7/1/13 0:00        present
3996      train  947111   200000.0   7/1/14 0:00    1/1/15 0:00
3997      train  324966   400000.0   2/1/13 0:00        present

                      Designation           JobCity Gender
DOB  \
```

```
0         senior quality engineer         Bangalore      f  2/19/90
0:00
1               assistant manager            Indore      m  10/4/89
0:00
2                systems engineer           Chennai      f   8/3/92
0:00
3        senior software engineer           Gurgaon      m  12/5/89
0:00
4                             get           Manesar      m  2/27/91
0:00
...                           ...               ...    ...           .
..
3993            software engineer        New Delhi       m  4/15/87
0:00
3994             technical writer         Hyderabad      f  8/27/92
0:00
3995  associate software engineer         Bangalore      m   7/3/91
0:00
3996           software developer  Asifabadbanglore      f  3/20/92
0:00
3997       senior systems engineer          Chennai      f  2/26/91
0:00

      10percentage  ... ComputerScience  MechanicalEngg
ElectricalEngg  \
0            84.30  ...              -1              -1              -
1
1            85.40  ...              -1              -1              -
1
2            85.00  ...              -1              -1              -
1
3            85.60  ...              -1              -1              -
1
4            78.00  ...              -1              -1              -
1
...            ...  ...             ...             ...             ..
.
3993         52.09  ...              -1              -1              -
1
3994         90.00  ...              -1              -1              -
1
3995         81.86  ...              -1              -1              -
1
3996         78.72  ...             438              -1              -
1
3997         70.60  ...              -1              -1              -
1

     TelecomEngg  CivilEngg  conscientiousness agreeableness
```

```
extraversion  \
0              -1         -1              0.9737        0.8128
0.5269
1              -1         -1             -0.7335        0.3789
1.2396
2              -1         -1              0.2718        1.7109
0.1637
3              -1         -1              0.0464        0.3448      -
0.3440
4              -1         -1             -0.8810       -0.2793      -
1.0697
...           ...        ...                 ...           ...
...
3993           -1         -1             -0.1082        0.3448
0.2366
3994           -1         -1             -0.3027        0.8784
0.9322
3995           -1         -1             -1.5765       -1.5273      -
1.5051
3996           -1         -1             -0.1590        0.0459      -
0.4511
3997           -1         -1             -1.1128       -0.2793      -
0.6343

      nueroticism  openess_to_experience
0         1.35490                -0.4455
1        -0.10760                 0.8637
2        -0.86820                 0.6721
3        -0.40780                -0.9194
4         0.09163                -0.1295
...           ...                    ...
3993      0.64980                -0.9194
3994      0.77980                -0.0943
3995     -1.31840                -0.7615
3996     -0.36120                -0.0943
3997      1.32553                -0.6035

[3998 rows x 39 columns]

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 39 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             3998 non-null   object
 1   ID                     3998 non-null   int64
 2   Salary                 3998 non-null   float64
 3   DOJ                    3998 non-null   object
```

```
 4   DOL                    3998 non-null   object
 5   Designation            3998 non-null   object
 6   JobCity                3998 non-null   object
 7   Gender                 3998 non-null   object
 8   DOB                    3998 non-null   object
 9   10percentage           3998 non-null   float64
 10  10board                3998 non-null   object
 11  12graduation           3998 non-null   int64
 12  12percentage           3998 non-null   float64
 13  12board                3998 non-null   object
 14  CollegeID              3998 non-null   int64
 15  CollegeTier            3998 non-null   int64
 16  Degree                 3998 non-null   object
 17  Specialization         3998 non-null   object
 18  collegeGPA             3998 non-null   float64
 19  CollegeCityID          3998 non-null   int64
 20  CollegeCityTier        3998 non-null   int64
 21  CollegeState           3998 non-null   object
 22  GraduationYear         3998 non-null   int64
 23  English                3998 non-null   int64
 24  Logical                3998 non-null   int64
 25  Quant                  3998 non-null   int64
 26  Domain                 3998 non-null   float64
 27  ComputerProgramming    3998 non-null   int64
 28  ElectronicsAndSemicon  3998 non-null   int64
 29  ComputerScience        3998 non-null   int64
 30  MechanicalEngg         3998 non-null   int64
 31  ElectricalEngg         3998 non-null   int64
 32  TelecomEngg            3998 non-null   int64
 33  CivilEngg              3998 non-null   int64
 34  conscientiousness      3998 non-null   float64
 35  agreeableness          3998 non-null   float64
 36  extraversion           3998 non-null   float64
 37  nueroticism            3998 non-null   float64
 38  openess_to_experience  3998 non-null   float64
dtypes: float64(10), int64(17), object(12)
memory usage: 1.2+ MB
```

```
df.isnull().sum()
```

```
Unnamed: 0               0
ID                       0
Salary                   0
DOJ                      0
DOL                      0
Designation              0
JobCity                  0
Gender                   0
DOB                      0
10percentage             0
```

```
10board                        0
12graduation                   0
12percentage                   0
12board                        0
CollegeID                      0
CollegeTier                    0
Degree                         0
Specialization                 0
collegeGPA                     0
CollegeCityID                  0
CollegeCityTier                0
CollegeState                   0
GraduationYear                 0
English                        0
Logical                        0
Quant                          0
Domain                         0
ComputerProgramming            0
ElectronicsAndSemicon          0
ComputerScience                0
MechanicalEngg                 0
ElectricalEngg                 0
TelecomEngg                    0
CivilEngg                      0
conscientiousness              0
agreeableness                  0
extraversion                   0
nueroticism                    0
openess_to_experience          0
dtype: int64
```

```
df.duplicated().sum()
```

```
0
```

```
df.dtypes
```

```
Unnamed: 0                  object
ID                           int64
Salary                     float64
DOJ                         object
DOL                         object
Designation                 object
JobCity                     object
Gender                      object
DOB                         object
10percentage               float64
10board                     object
12graduation                 int64
12percentage               float64
```

```
12board                     object
CollegeID                    int64
CollegeTier                  int64
Degree                      object
Specialization              object
collegeGPA                 float64
CollegeCityID                int64
CollegeCityTier              int64
CollegeState                object
GraduationYear               int64
English                      int64
Logical                      int64
Quant                        int64
Domain                     float64
ComputerProgramming          int64
ElectronicsAndSemicon        int64
ComputerScience              int64
MechanicalEngg               int64
ElectricalEngg               int64
TelecomEngg                  int64
CivilEngg                    int64
conscientiousness          float64
agreeableness              float64
extraversion               float64
nueroticism                float64
openess_to_experience      float64
dtype: object
```

```
df.columns
```

```
Index(['Unnamed: 0', 'ID', 'Salary', 'DOJ', 'DOL', 'Designation',
'JobCity',
       'Gender', 'DOB', '10percentage', '10board', '12graduation',
       '12percentage', '12board', 'CollegeID', 'CollegeTier',
'Degree',
       'Specialization', 'collegeGPA', 'CollegeCityID',
'CollegeCityTier',
       'CollegeState', 'GraduationYear', 'English', 'Logical',
'Quant',
       'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon',
       'ComputerScience', 'MechanicalEngg', 'ElectricalEngg',
'TelecomEngg',
       'CivilEngg', 'conscientiousness', 'agreeableness',
'extraversion',
       'nueroticism', 'openess_to_experience'],
      dtype='object')
```

```
df.describe()
```

```
                 ID        Salary  10percentage  12graduation
12percentage  \
count  3.998000e+03  3.998000e+03   3998.000000   3998.000000
3998.000000
mean   6.637945e+05  3.076998e+05     77.925443   2008.087544
74.466366
std    3.632182e+05  2.127375e+05      9.850162      1.653599
10.999933
min    1.124400e+04  3.500000e+04     43.000000   1995.000000
40.000000
25%    3.342842e+05  1.800000e+05     71.680000   2007.000000
66.000000
50%    6.396000e+05  3.000000e+05     79.150000   2008.000000
74.400000
75%    9.904800e+05  3.700000e+05     85.670000   2009.000000
82.600000
max    1.298275e+06  4.000000e+06     97.760000   2013.000000
98.700000

          CollegeID  CollegeTier   collegeGPA  CollegeCityID
CollegeCityTier  \
count   3998.000000  3998.000000  3998.000000    3998.000000
3998.000000
mean    5156.851426     1.925713    71.486171    5156.851426
0.300400
std     4802.261482     0.262270     8.167338    4802.261482
0.458489
min        2.000000     1.000000     6.450000       2.000000
0.000000
25%      494.000000     2.000000    66.407500     494.000000
0.000000
50%     3879.000000     2.000000    71.720000    3879.000000
0.000000
75%     8818.000000     2.000000    76.327500    8818.000000
1.000000
max    18409.000000     2.000000    99.930000   18409.000000
1.000000

       ...  ComputerScience  MechanicalEngg  ElectricalEngg
TelecomEngg  \
count  ...      3998.000000     3998.000000     3998.000000
3998.000000
mean   ...        90.742371       22.974737       16.478739
31.851176
std    ...       175.273083       98.123311       87.585634
104.852845
min    ...        -1.000000       -1.000000       -1.000000    -
1.000000
25%    ...        -1.000000       -1.000000       -1.000000    -
1.000000
```

```
50%    ...        -1.000000       -1.000000       -1.000000    -
1.000000
75%    ...        -1.000000       -1.000000       -1.000000    -
1.000000
max    ...       715.000000      623.000000      676.000000
548.000000

         CivilEngg  conscientiousness  agreeableness  extraversion  \
count  3998.000000        3998.000000    3998.000000   3998.000000
mean      2.683842          -0.037831       0.146496      0.002763
std      36.658505           1.028666       0.941782      0.951471
min      -1.000000          -4.126700      -5.781600     -4.600900
25%      -1.000000          -0.713525      -0.287100     -0.604800
50%      -1.000000           0.046400       0.212400      0.091400
75%      -1.000000           0.702700       0.812800      0.672000
max     516.000000           1.995300       1.904800      2.535400

       nueroticism  openess_to_experience
count  3998.000000            3998.000000
mean     -0.169033              -0.138110
std       1.007580               1.008075
min      -2.643000              -7.375700
25%      -0.868200              -0.669200
50%      -0.234400              -0.094300
75%       0.526200               0.502400
max       3.352500               1.822400

[8 rows x 27 columns]
```

```python
df.describe(include='object')
```

```
       Unnamed: 0          DOJ      DOL         Designation    JobCity
Gender  \
count        3998         3998     3998                3998       3998
3998
unique          1           81       67                 419        339
2
top         train  7/1/14 0:00  present  software engineer  Bangalore
m
freq         3998          199     1875                 539        627
3041

                DOB 10board 12board       Degree  \
count          3998    3998    3998         3998
unique         1872     275     340            4
top     1/1/91 0:00    cbse    cbse  B.Tech/B.E.
freq             11    1395    1400         3700

                                   Specialization   CollegeState
count                                        3998           3998
```

```
unique                                              46            26
top     electronics and communication engineering  Uttar Pradesh
freq                                               880           915
```

```
df.head()
```

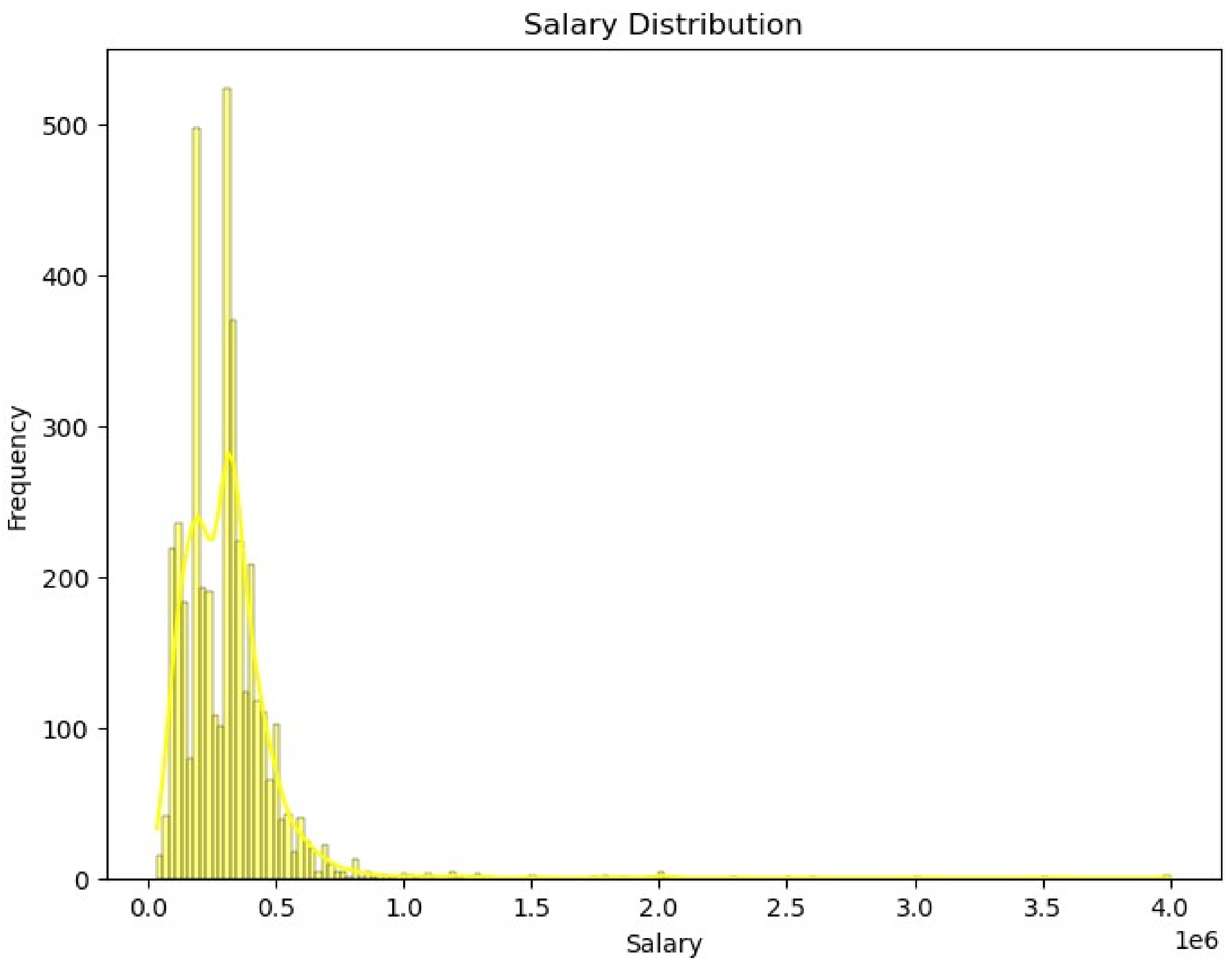
```
  Unnamed: 0      ID     Salary          DOJ          DOL  \
0      train  203097   420000.0  6/1/12 0:00      present
1      train  579905   500000.0  9/1/13 0:00      present
2      train  810601   325000.0  6/1/14 0:00      present
3      train  267447  1100000.0  7/1/11 0:00      present
4      train  343523   200000.0  3/1/14 0:00  3/1/15 0:00

                Designation    JobCity Gender           DOB
10percentage  \
0   senior quality engineer  Bangalore      f  2/19/90 0:00
84.3
1         assistant manager     Indore      m  10/4/89 0:00
85.4
2          systems engineer    Chennai      f   8/3/92 0:00
85.0
3  senior software engineer    Gurgaon      m  12/5/89 0:00
85.6
4                       get    Manesar      m  2/27/91 0:00
78.0

   ... ComputerScience  MechanicalEngg  ElectricalEngg TelecomEngg
CivilEngg  \
0  ...              -1              -1              -1          -1
-1
1  ...              -1              -1              -1          -1
-1
2  ...              -1              -1              -1          -1
-1
3  ...              -1              -1              -1          -1
-1
4  ...              -1              -1              -1          -1
-1

   conscientiousness agreeableness extraversion  nueroticism  \
0             0.9737        0.8128       0.5269      1.35490
1            -0.7335        0.3789       1.2396     -0.10760
2             0.2718        1.7109       0.1637     -0.86820
3             0.0464        0.3448      -0.3440     -0.40780
4            -0.8810       -0.2793      -1.0697      0.09163

   openess_to_experience
0                -0.4455
1                 0.8637
2                 0.6721
```

```
3                  -0.9194
4                  -0.1295

[5 rows x 39 columns]

df.shape

(3998, 39)

df.size

155922
```

Univariate Analysis: Numerical Variables

1. Salary Distribution

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8,6))
sns.histplot(df['Salary'],kde=True,color='yellow')
plt.title('Salary Distribution')
plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.show()
```

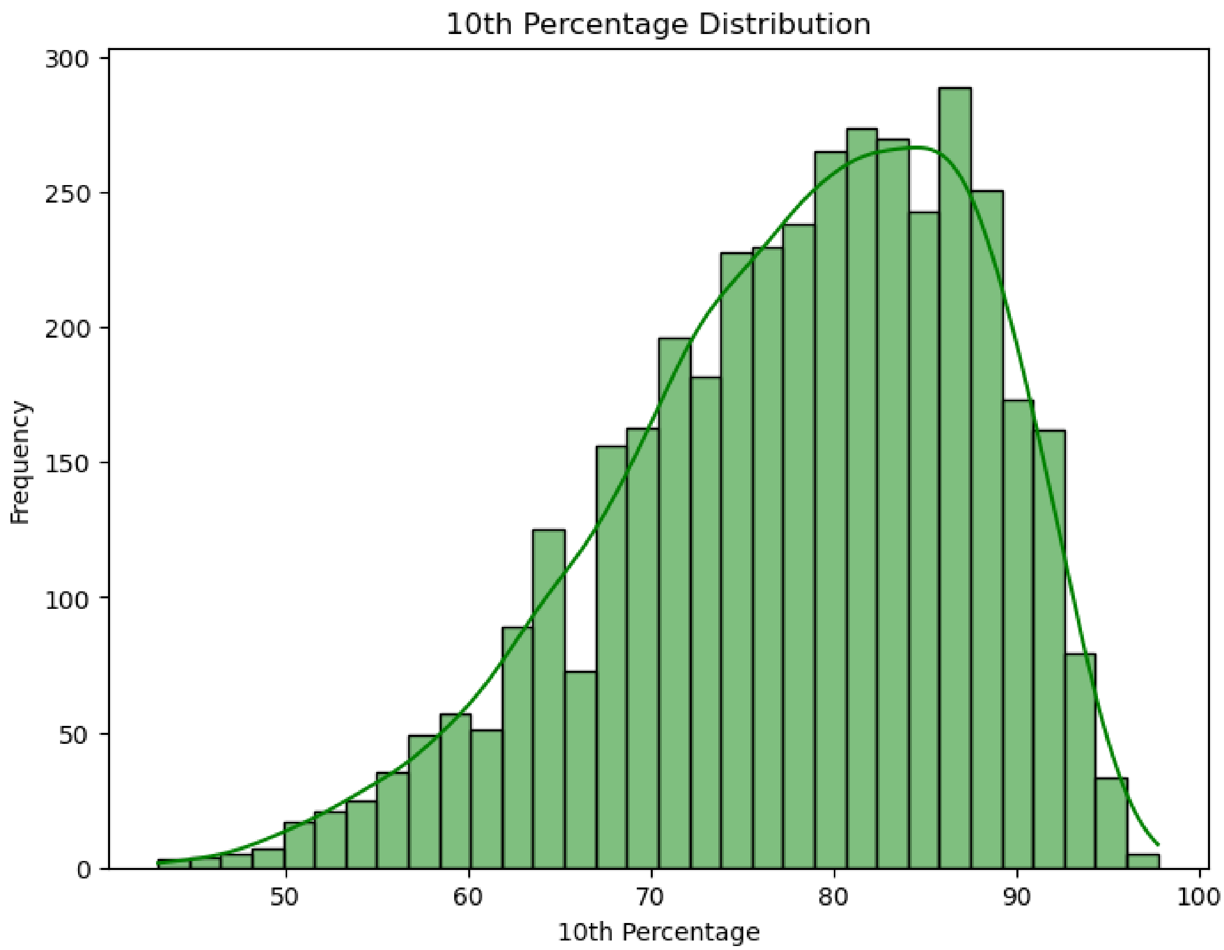


Observation:

The salary distribution is skewed to the right, meaning most of the engineers are earning a salary below the mean, while a few are earning much higher salaries.

2. 10th Percentage Distribution

```
plt.figure(figsize=(8, 6))
sns.histplot(df['10percentage'], kde=True, color='green')
plt.title('10th Percentage Distribution')
plt.xlabel('10th Percentage')
plt.ylabel('Frequency')
plt.show()
```

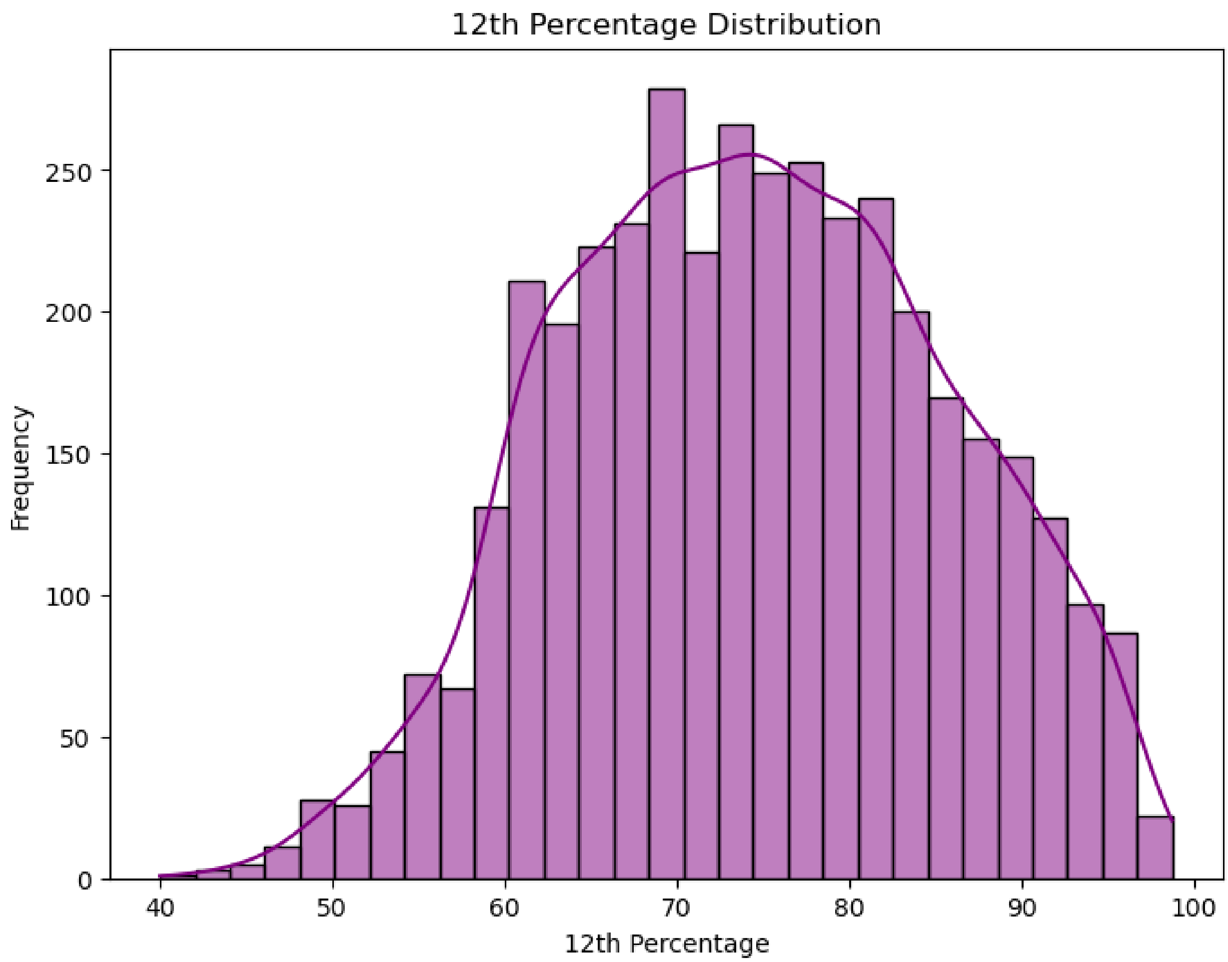


Observation:

Most students scored between 60% and 80% in their 10th-grade exams, with a small number of students achieving higher or lower percentages.

3. 12th Percentage Distribution

```
plt.figure(figsize=(8, 6))
sns.histplot(df['12percentage'], kde=True, color='purple')
plt.title('12th Percentage Distribution')
plt.xlabel('12th Percentage')
plt.ylabel('Frequency')
plt.show()
```

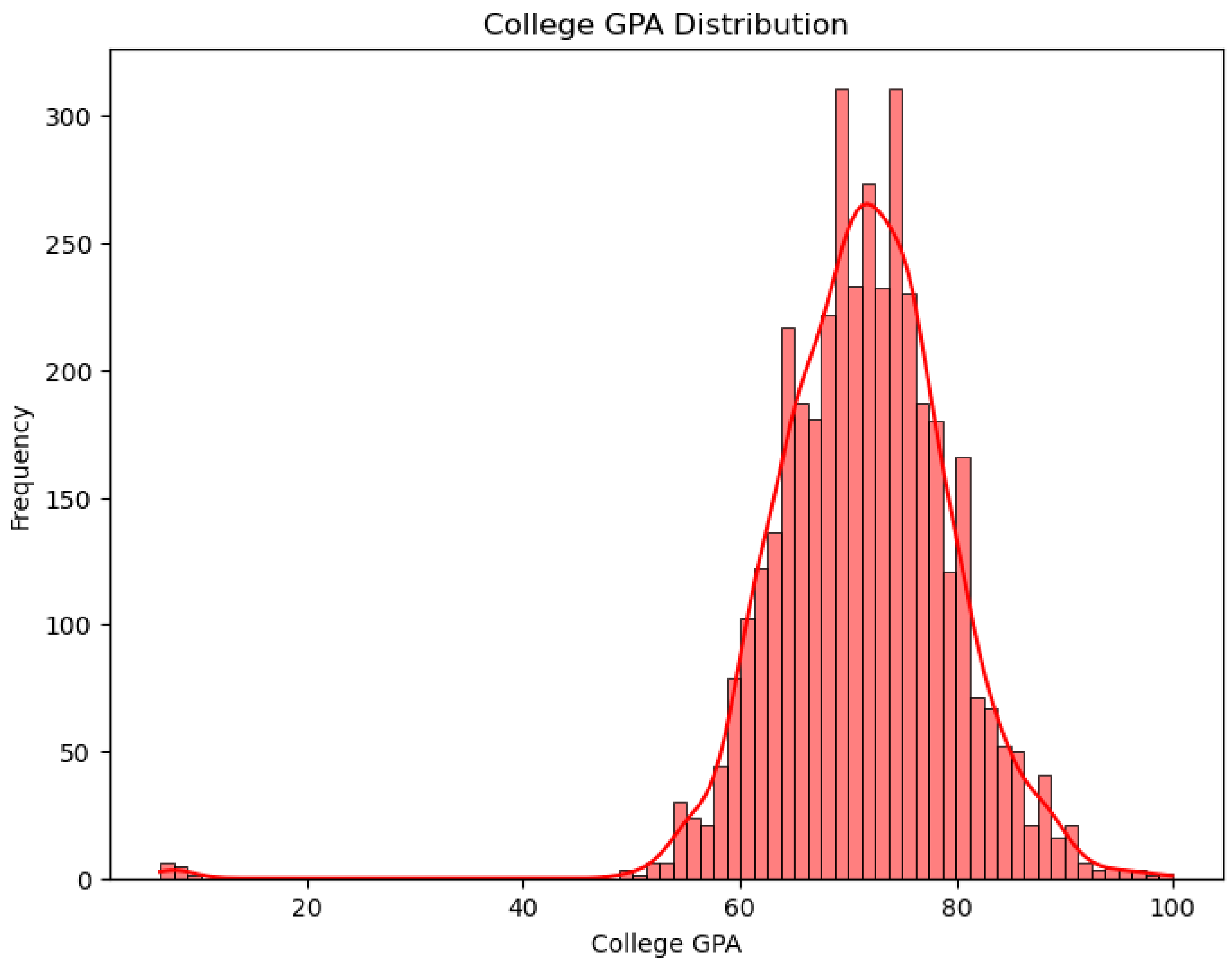


Observation:

The distribution of 12th-grade percentages follows a similar trend to the 10th percentage, with the majority of students scoring between 60% and 80%.

4. College GPA Distribution

```
plt.figure(figsize=(8, 6))
sns.histplot(df['collegeGPA'], kde=True, color='red')
plt.title('College GPA Distribution')
plt.xlabel('College GPA')
plt.ylabel('Frequency')
plt.show()
```

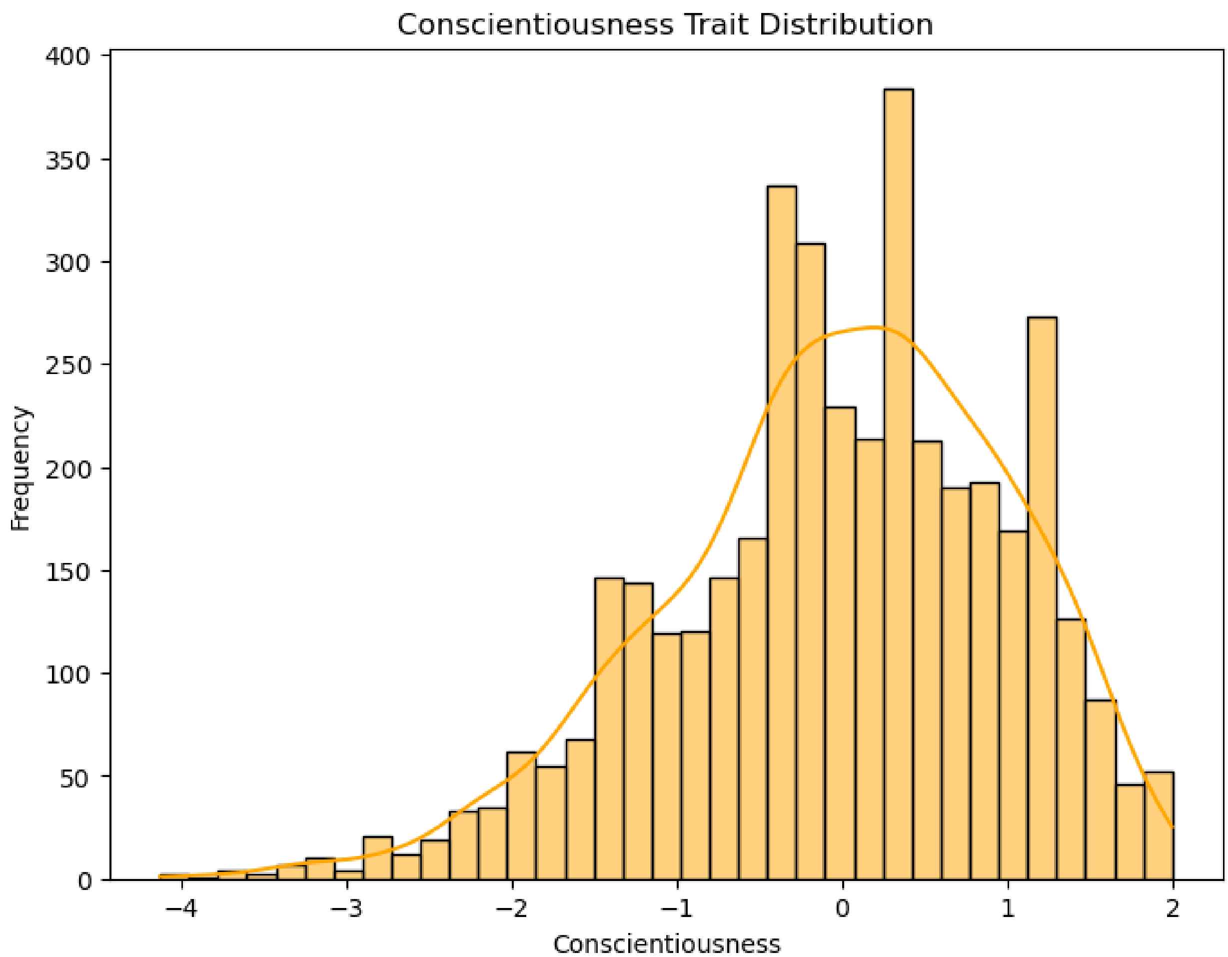


Observation:

College GPA is mostly concentrated between 6.0 and 8.0, indicating that the majority of students have decent academic performance during their college years.

5. Personality Traits: Conscientiousness

```
plt.figure(figsize=(8, 6))
sns.histplot(df['conscientiousness'], kde=True, color='orange')
plt.title('Conscientiousness Trait Distribution')
plt.xlabel('Conscientiousness')
plt.ylabel('Frequency')
plt.show()
```



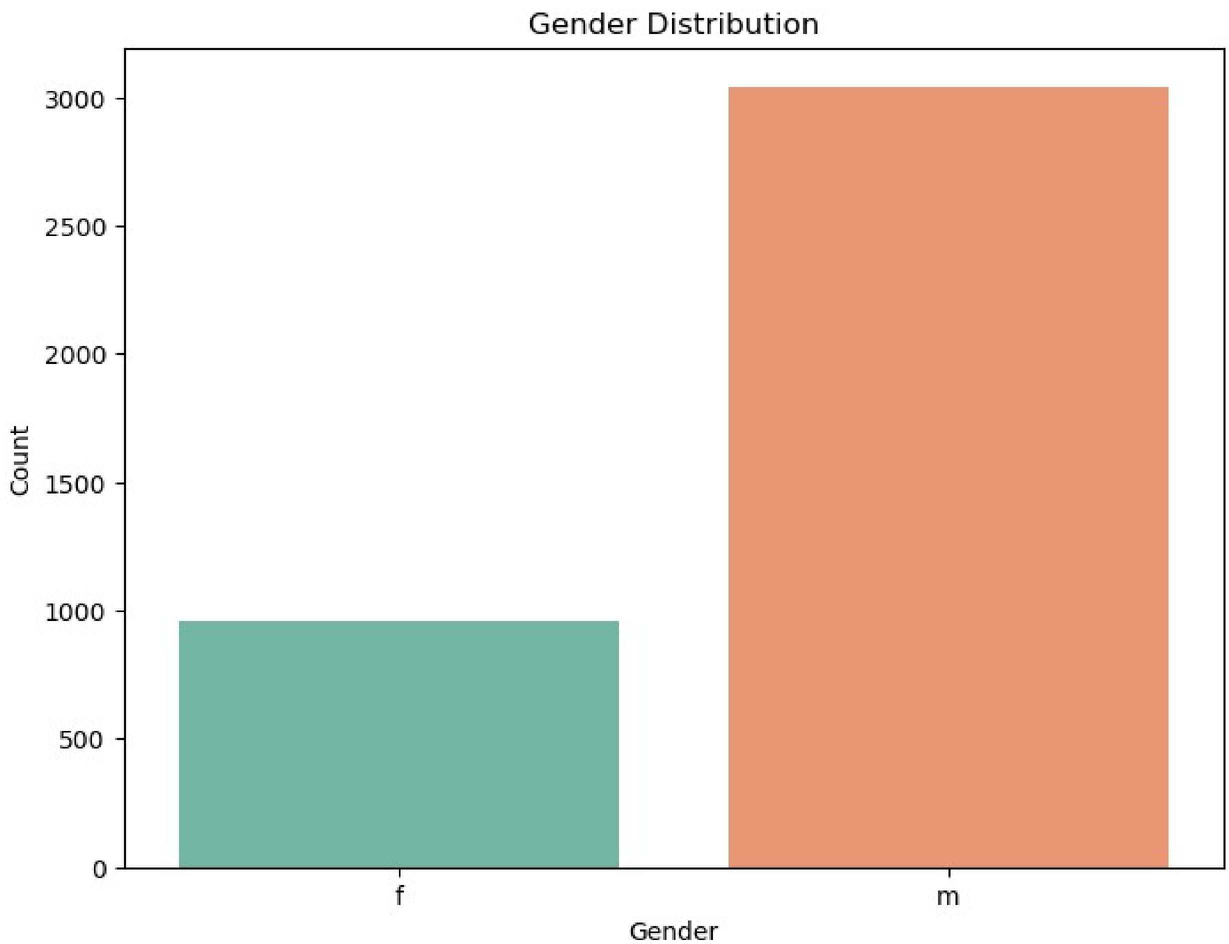
Observation:

The conscientiousness trait is fairly normally distributed, with most students scoring around the mid-level for this personality trait

Univariate Analysis: Categorical Variables

6. Gender Distribution

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Gender', data=df, palette='Set2')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

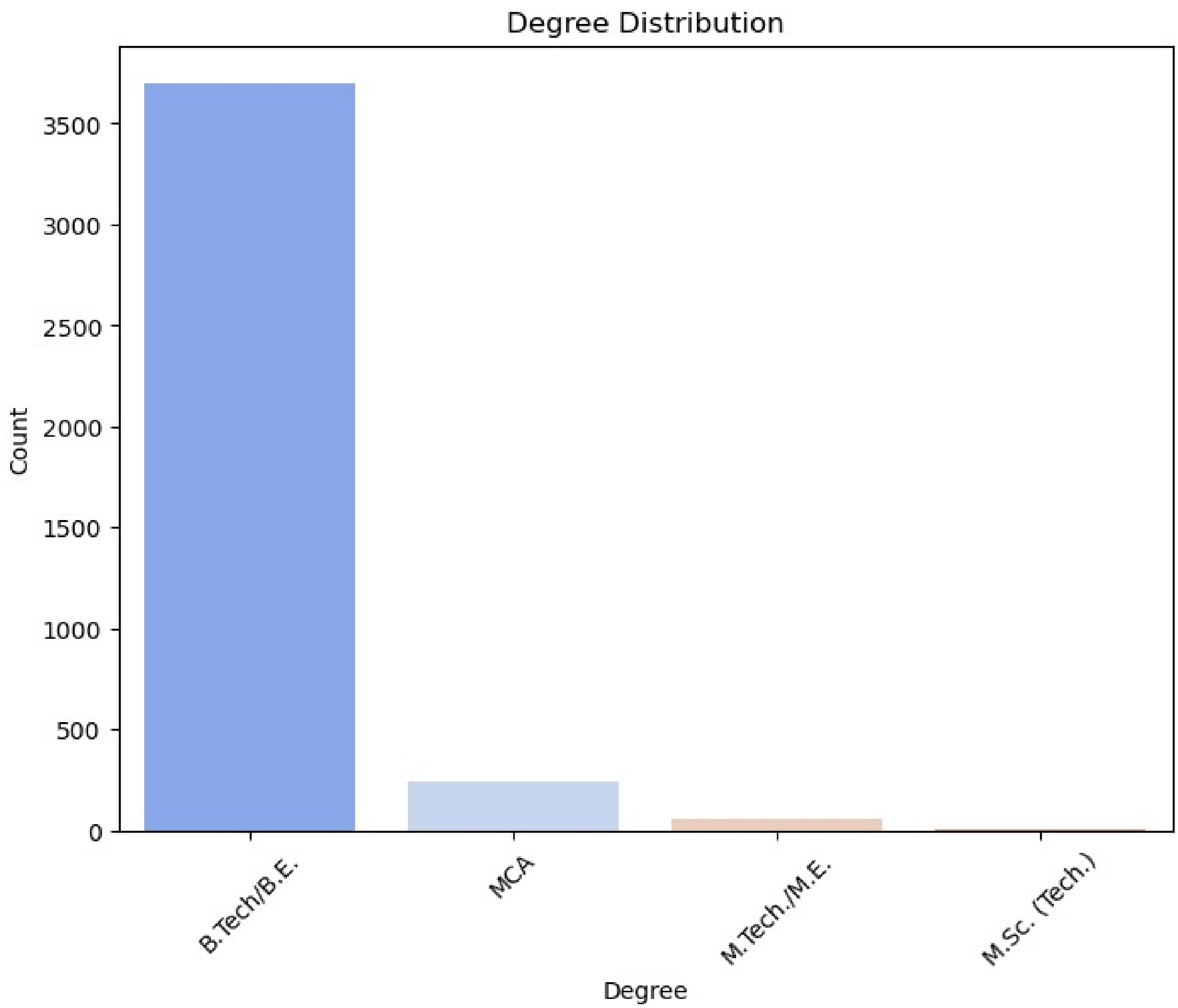


Observation:

There are more male students than female students in the dataset, indicating a gender imbalance in engineering disciplines.

7. Degree Distribution

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Degree', data=df, palette='coolwarm')
plt.title('Degree Distribution')
plt.xticks(rotation=45)
plt.xlabel('Degree')
plt.ylabel('Count')
plt.show()
```

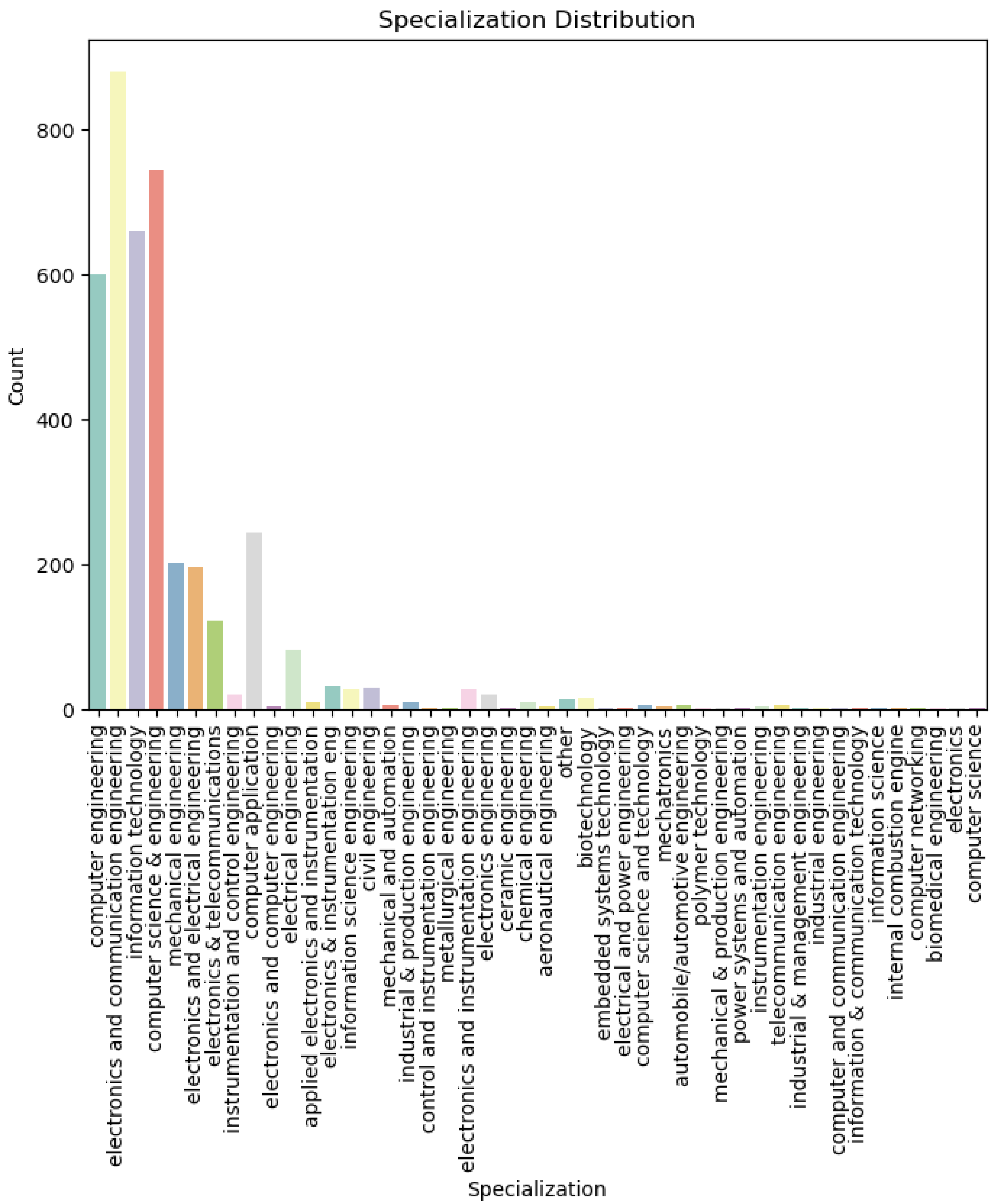


Observation:

B.Tech (Bachelor of Technology) is the most common degree among the students in the dataset.

8. Specialization Distribution

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Specialization', data=df, palette='Set3')
plt.title('Specialization Distribution')
plt.xticks(rotation=90)
plt.xlabel('Specialization')
plt.ylabel('Count')
plt.show()
```



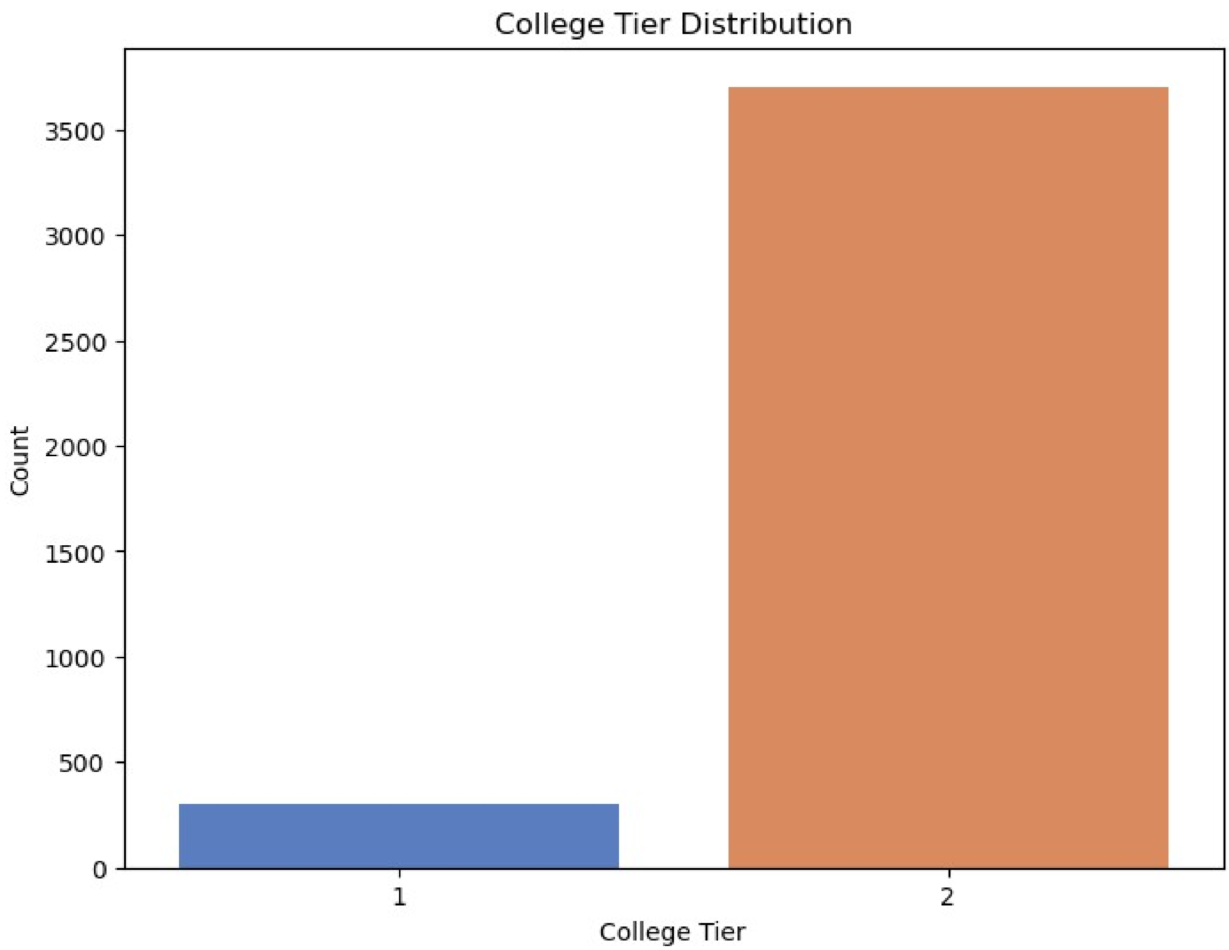
Observation:

Computer Science Engineering (CSE) is the most popular specialization, followed by Mechanical Engineering and Electronics and Communication Engineering.

Univariate Analysis: Other Variables

9. College Tier Distribution

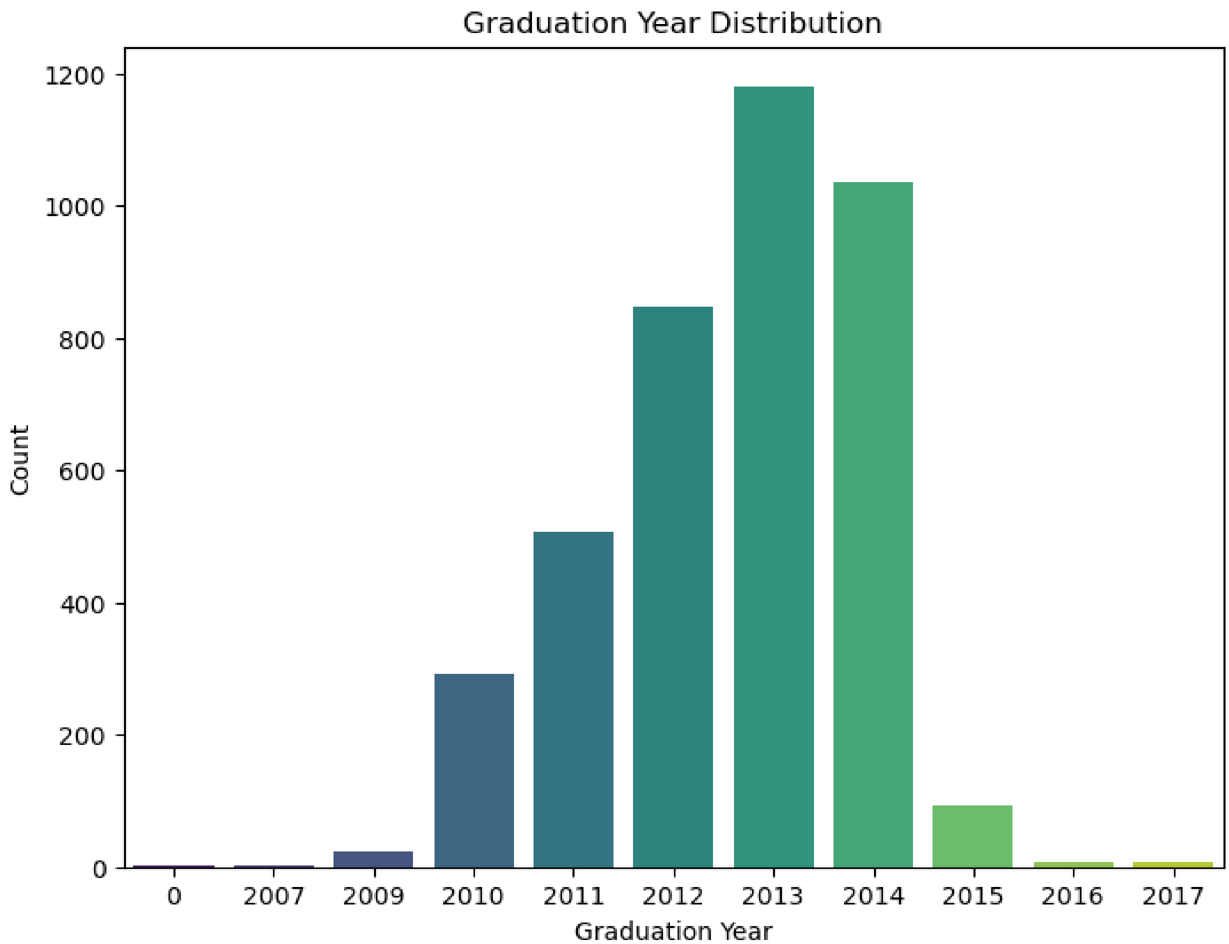
```
plt.figure(figsize=(8, 6))
sns.countplot(x='CollegeTier', data=df, palette='muted')
plt.title('College Tier Distribution')
plt.xlabel('College Tier')
plt.ylabel('Count')
plt.show()
```



10. Graduation Year Distribution

```
plt.figure(figsize=(8, 6))
sns.countplot(x='GraduationYear', data=df, palette='viridis')
plt.title('Graduation Year Distribution')
```

```
plt.xlabel('Graduation Year')
plt.ylabel('Count')
plt.show()
```



Observation:

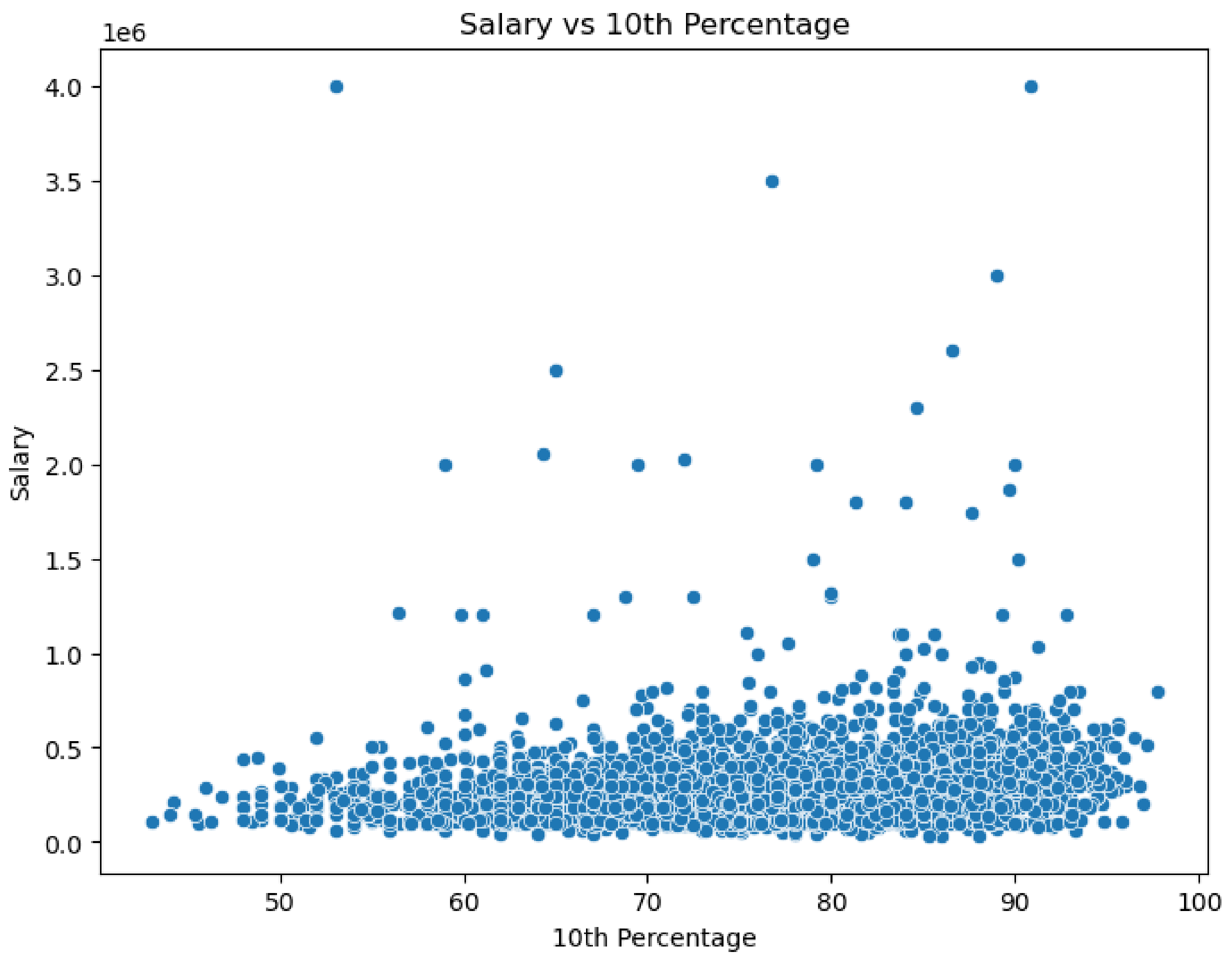
The majority of students in the dataset graduated in recent years, which could reflect a trend of increased engineering enrollments.

Bivariate Analysis: Numerical vs Numerical

1. Salary vs 10th Percentage

```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='10percentage', y='Salary', data=df)
```

```
plt.title('Salary vs 10th Percentage')
plt.xlabel('10th Percentage')
plt.ylabel('Salary')
plt.show()
```

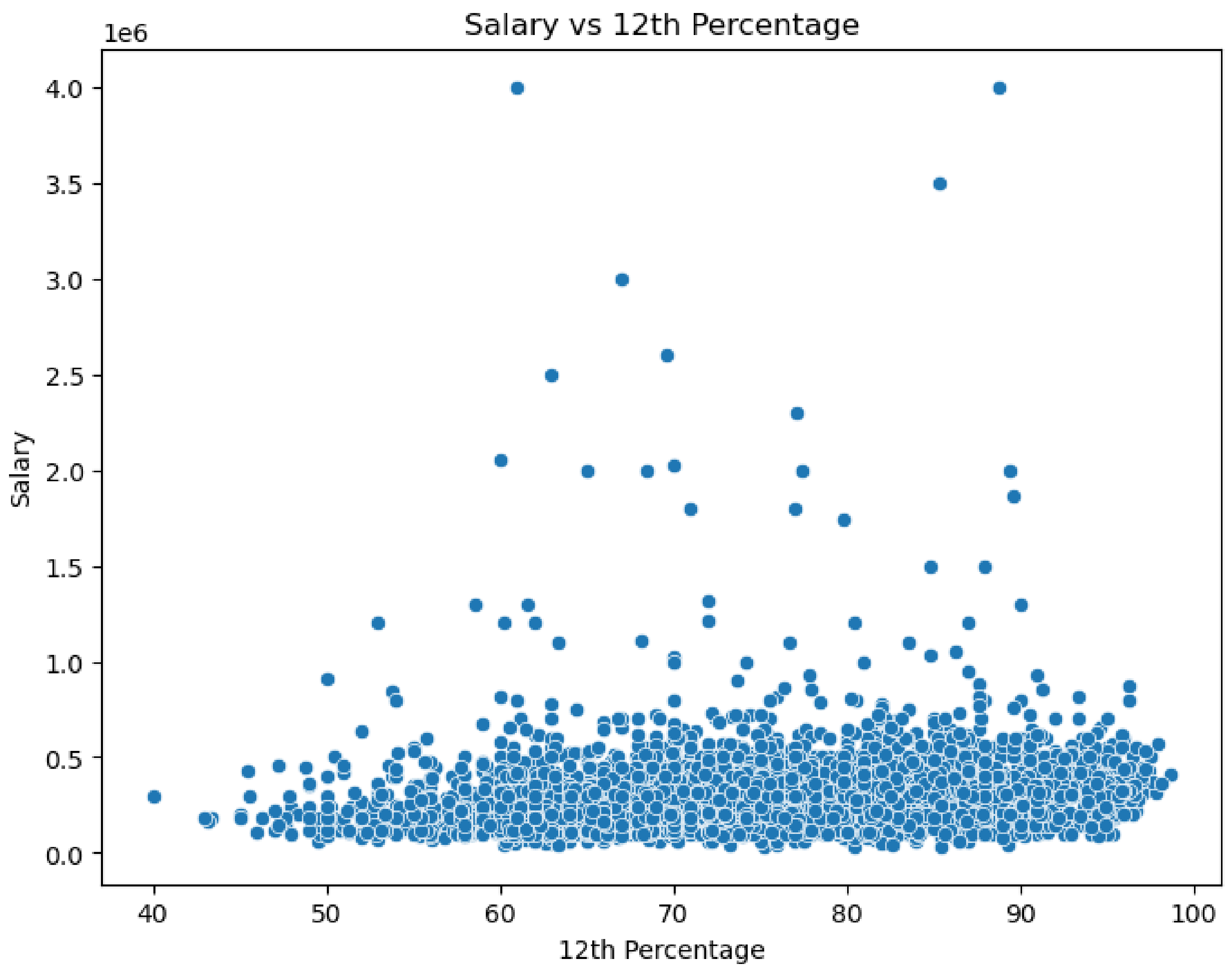


Observation:

There is no clear linear relationship between 10th-grade percentage and salary. Some students with lower percentages in 10th grade seem to earn higher salaries, indicating that early academic performance might not be strongly correlated with job outcomes.

2. Salary vs 12th Percentage

```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='12percentage', y='Salary', data=df)
plt.title('Salary vs 12th Percentage')
plt.xlabel('12th Percentage')
plt.ylabel('Salary')
plt.show()
```

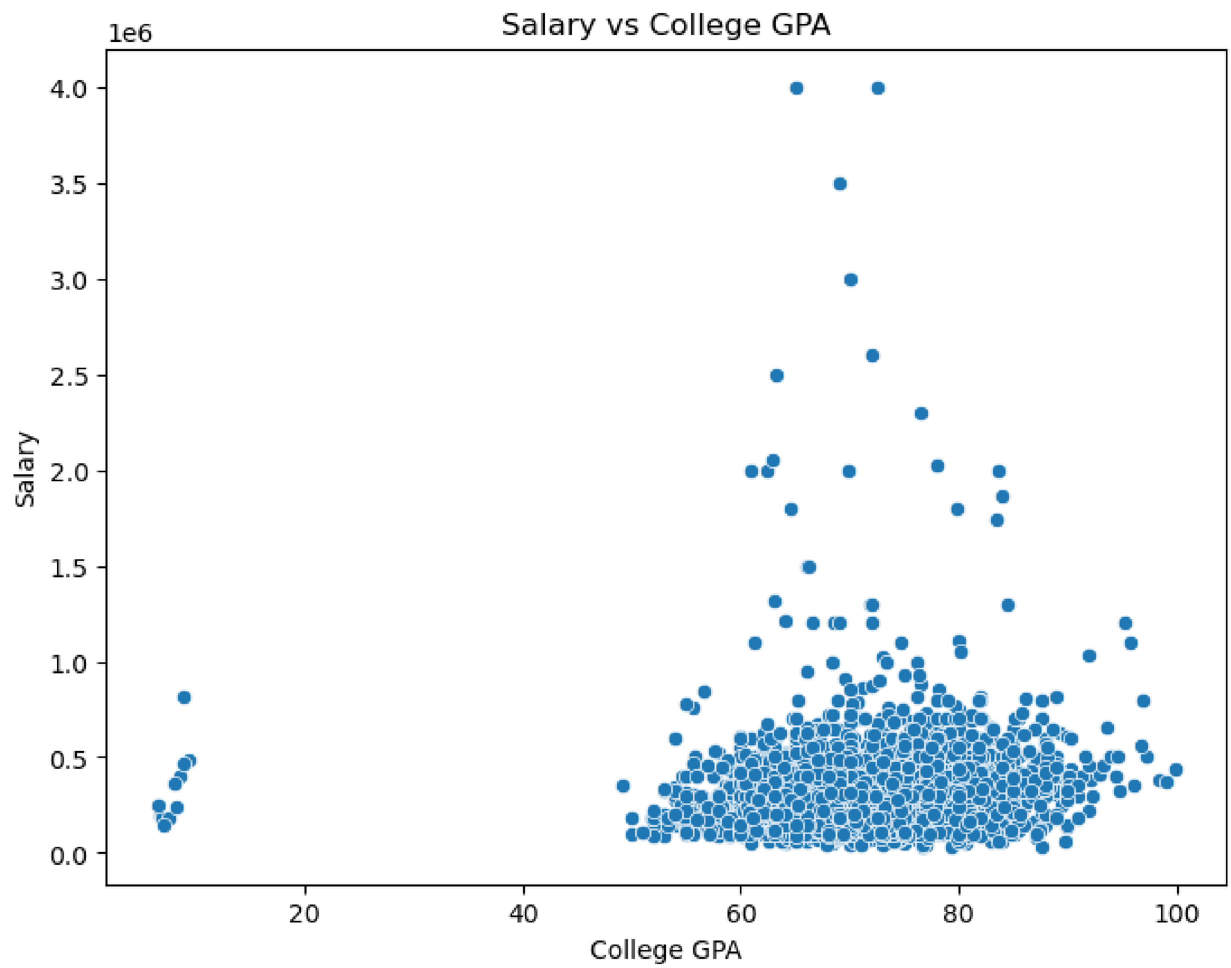


Observation:

Similar to the previous observation, the salary does not appear to be strongly correlated with the 12th-grade percentage. This suggests that high school academic performance may not have a direct influence on salary outcomes.

3. Salary vs College GPA

```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='collegeGPA', y='Salary', data=df)
plt.title('Salary vs College GPA')
plt.xlabel('College GPA')
plt.ylabel('Salary')
plt.show()
```



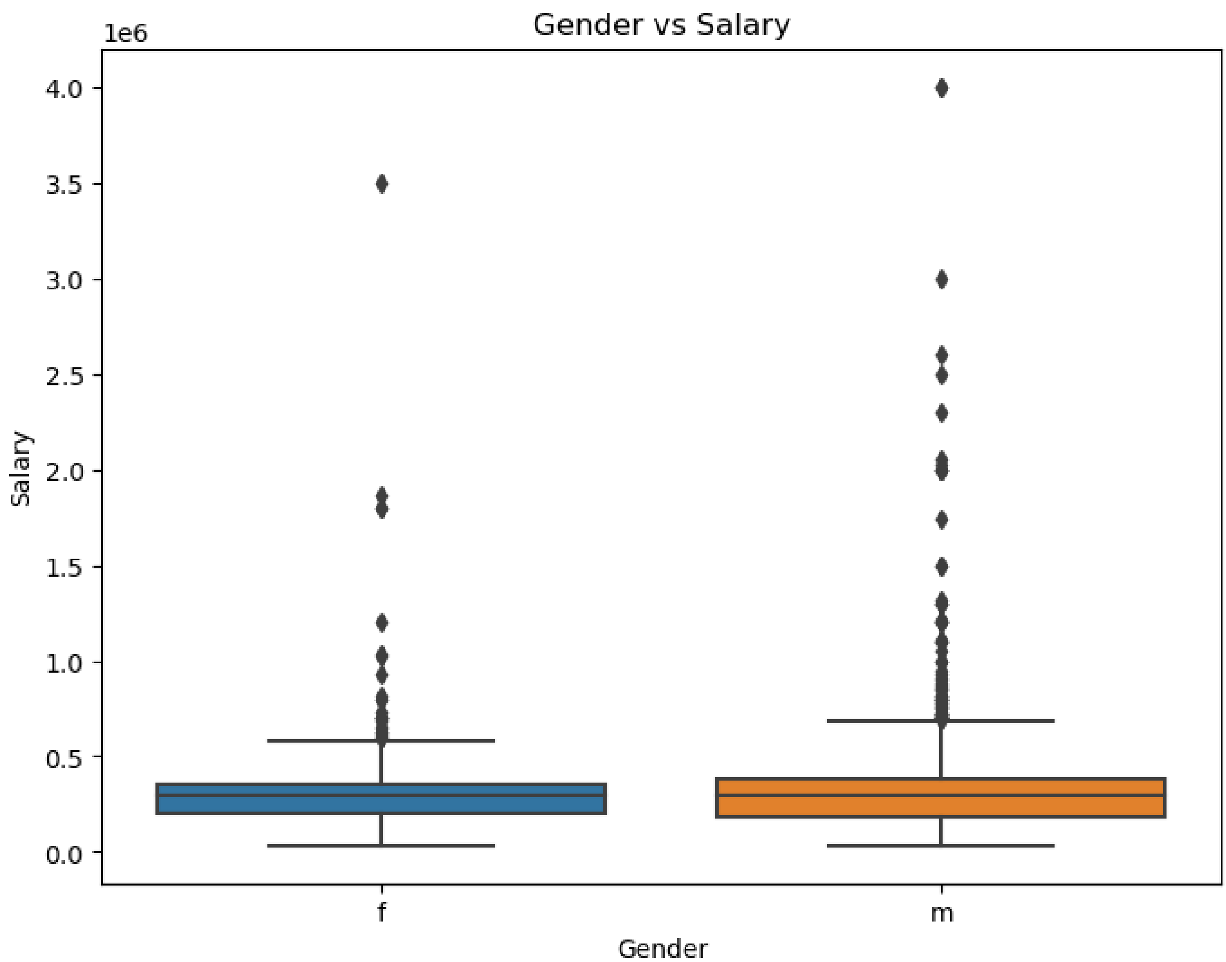
Observation:

There is a slight positive trend between college GPA and salary. Higher GPAs seem to lead to slightly higher salaries, though this relationship is not very strong.

Bivariate Analysis: Categorical vs Numerical

4. Gender vs Salary

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Gender', y='Salary', data=df)
plt.title('Gender vs Salary')
plt.xlabel('Gender')
plt.ylabel('Salary')
plt.show()
```

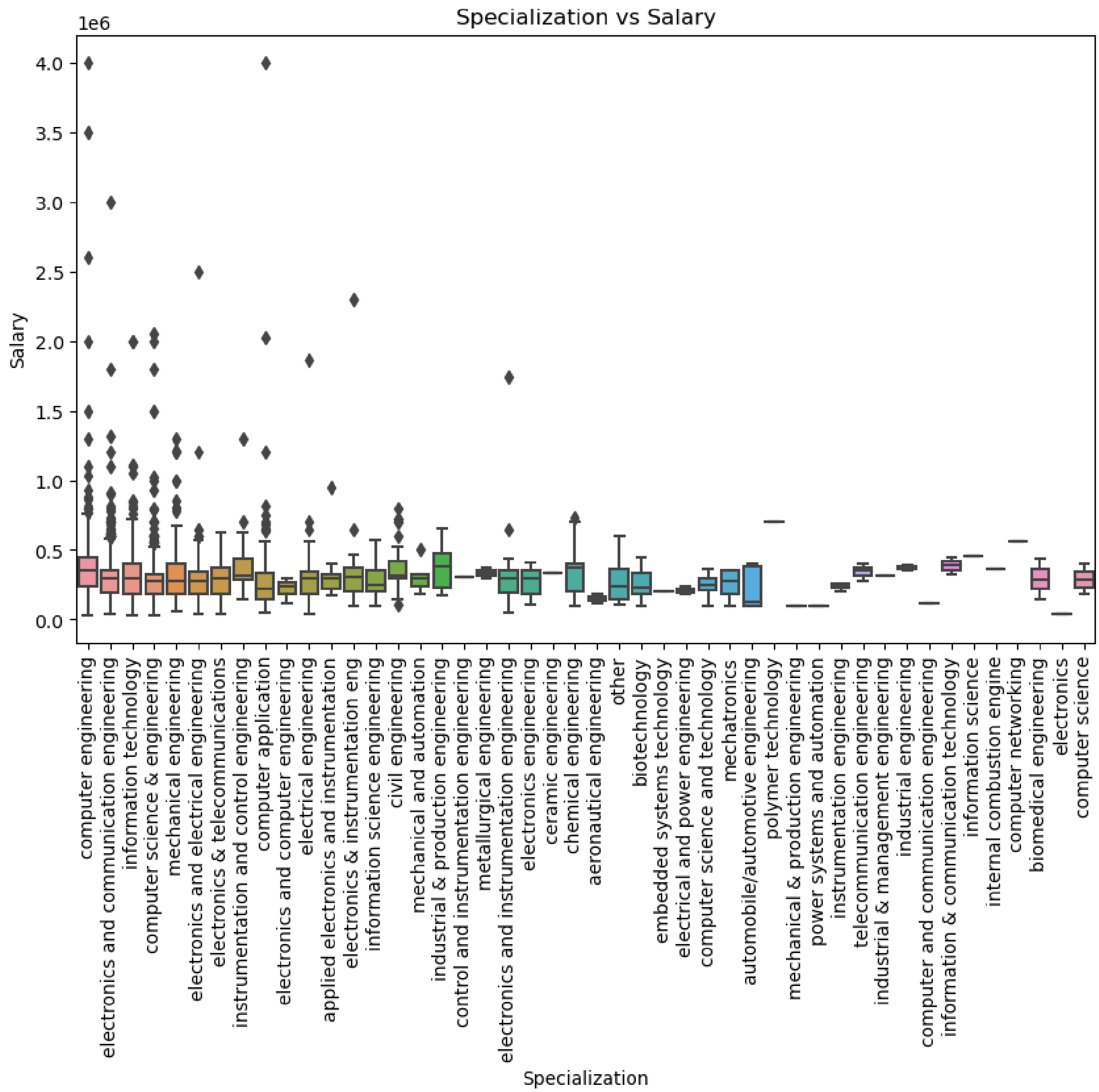


Observation:

There is a noticeable difference in salary distribution between genders. On average, males earn higher salaries than females, though there are outliers in both categories.

5. Specialization vs Salary

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Specialization', y='Salary', data=df)
plt.xticks(rotation=90)
plt.title('Specialization vs Salary')
plt.xlabel('Specialization')
plt.ylabel('Salary')
plt.show()
```



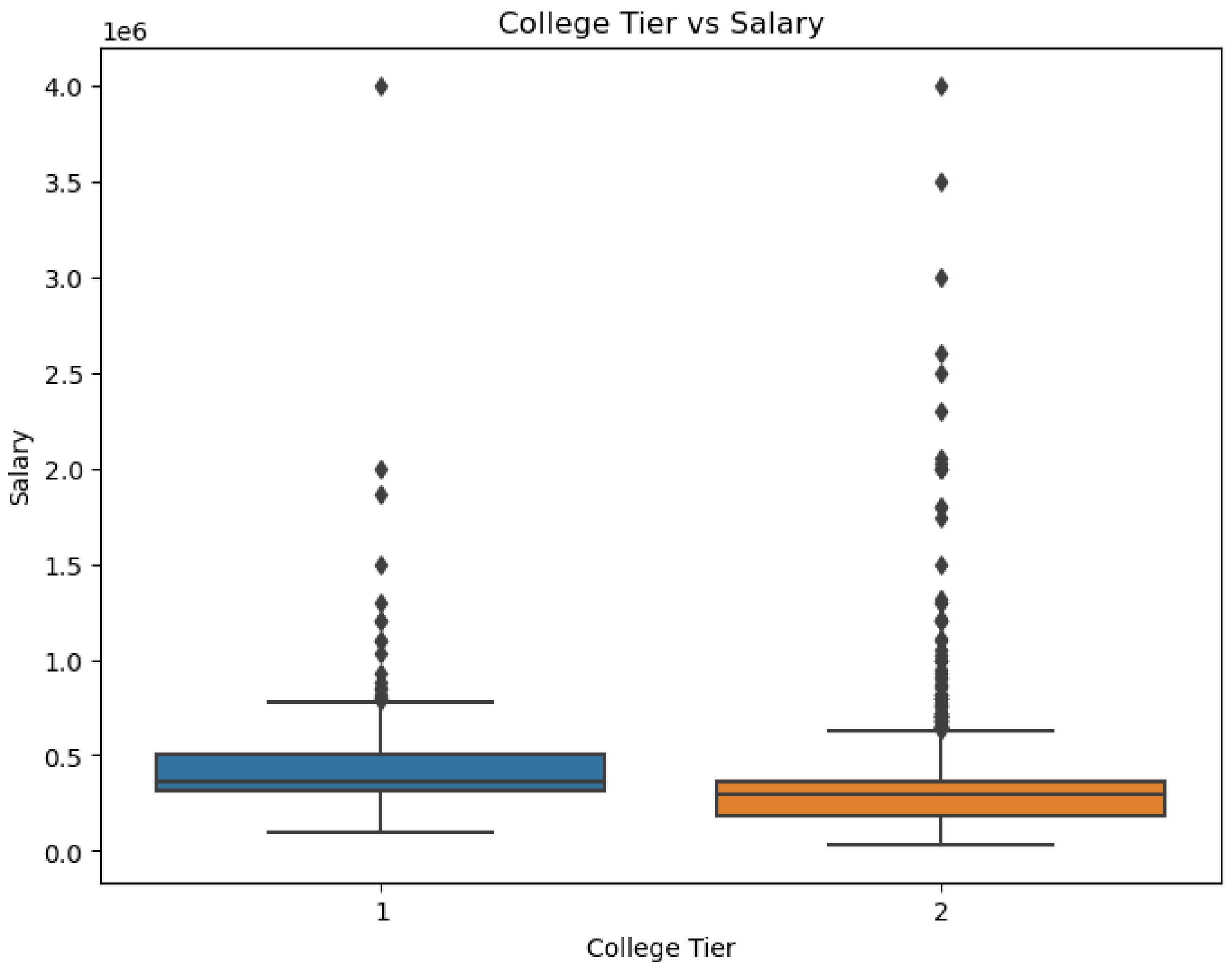
Observation:

Students with specializations in Computer Science and Electronics generally earn higher salaries compared to students in Civil or Mechanical Engineering. There are some high-salary outliers in the Computer Science specialization.

6. College Tier vs Salary

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='CollegeTier', y='Salary', data=df)
plt.title('College Tier vs Salary')
plt.xlabel('College Tier')
```

```
plt.ylabel('Salary')
plt.show()
```



Observation:

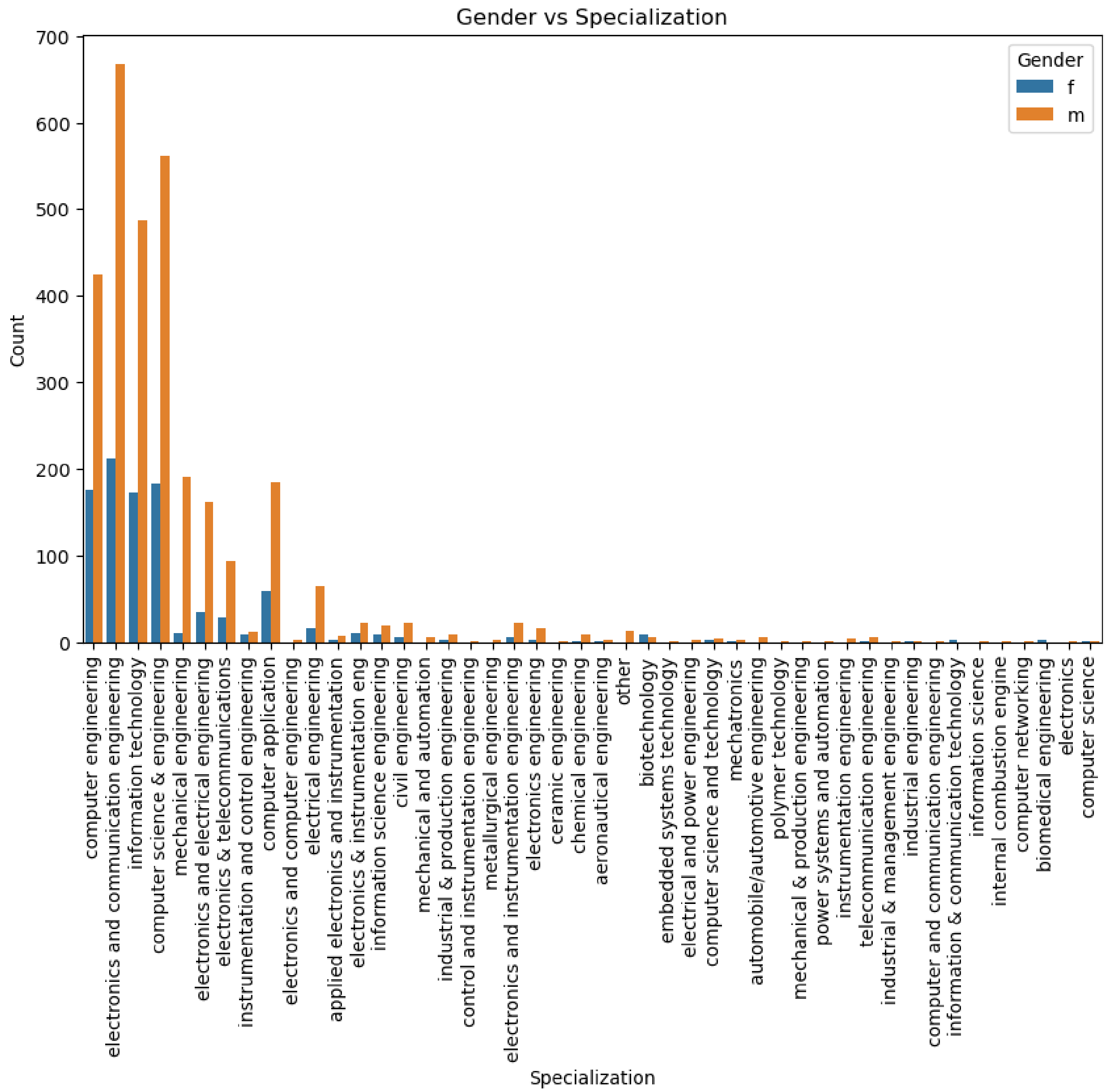
Students from tier 1 colleges tend to have higher salaries compared to those from tier 2 and tier 3 colleges. This suggests that college reputation or tier might have an impact on salary outcomes.

Bivariate Analysis: Categorical vs Categorical

7. Gender vs Specialization

```
plt.figure(figsize=(10, 6))
sns.countplot(x='Specialization', hue='Gender', data=df)
plt.xticks(rotation=90)
```

```
plt.title('Gender vs Specialization')
plt.xlabel('Specialization')
plt.ylabel('Count')
plt.show()
```

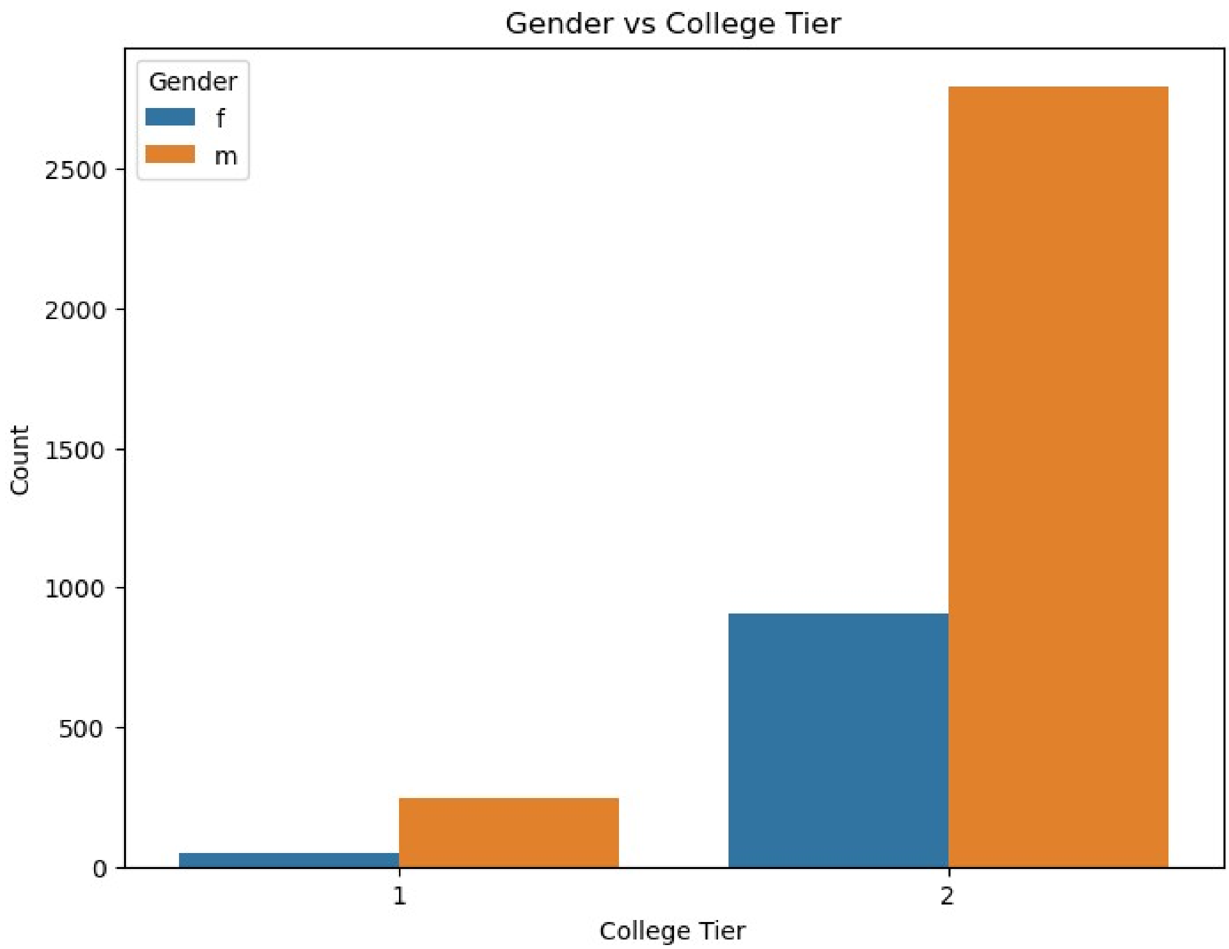


Observation:

In most specializations, males outnumber females. However, in some fields such as Computer Science, there is a more balanced distribution between genders compared to Mechanical and Electrical Engineering, where males dominate.

8. College Tier vs Gender

```
plt.figure(figsize=(8, 6))
sns.countplot(x='CollegeTier', hue='Gender', data=df)
plt.title('Gender vs College Tier')
plt.xlabel('College Tier')
plt.ylabel('Count')
plt.show()
```



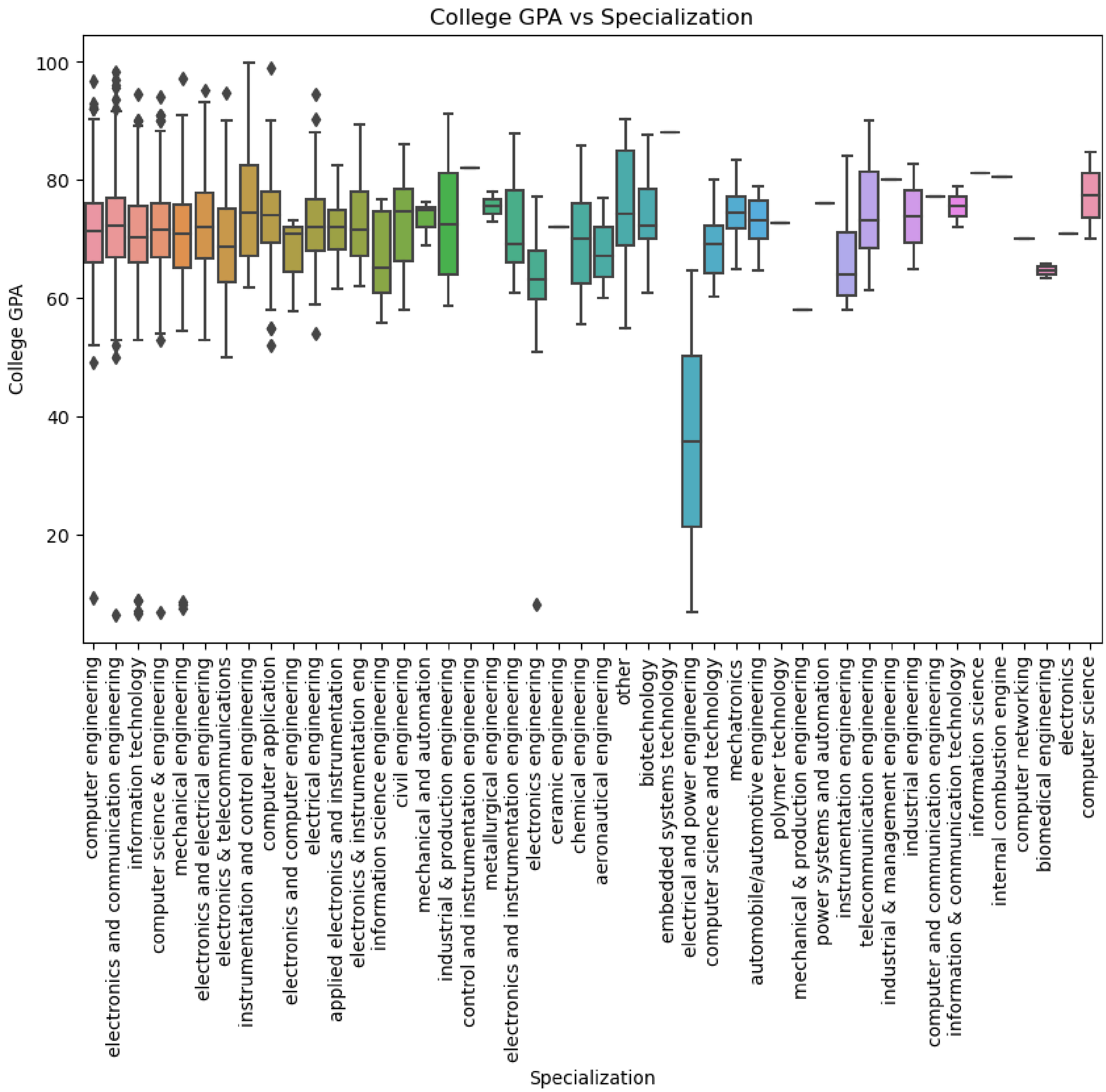
Observation:

Males are more prevalent across all college tiers, but the distribution is relatively consistent across different college tiers. Bivariate Analysis: Numerical vs Categorical

9. College GPA vs Specialization

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Specialization', y='collegeGPA', data=df)
plt.xticks(rotation=90)
```

```
plt.title('College GPA vs Specialization')
plt.xlabel('Specialization')
plt.ylabel('College GPA')
plt.show()
```



Observation:

College GPA tends to be higher for students in Computer Science and Electronics compared to Mechanical, Electrical, and Civil Engineering. There is less variation in GPA for students specializing in Mechanical and Civil Engineering.

Bivariate Analysis: Cognitive Skills vs Salary

10. Logical Scores vs Salary

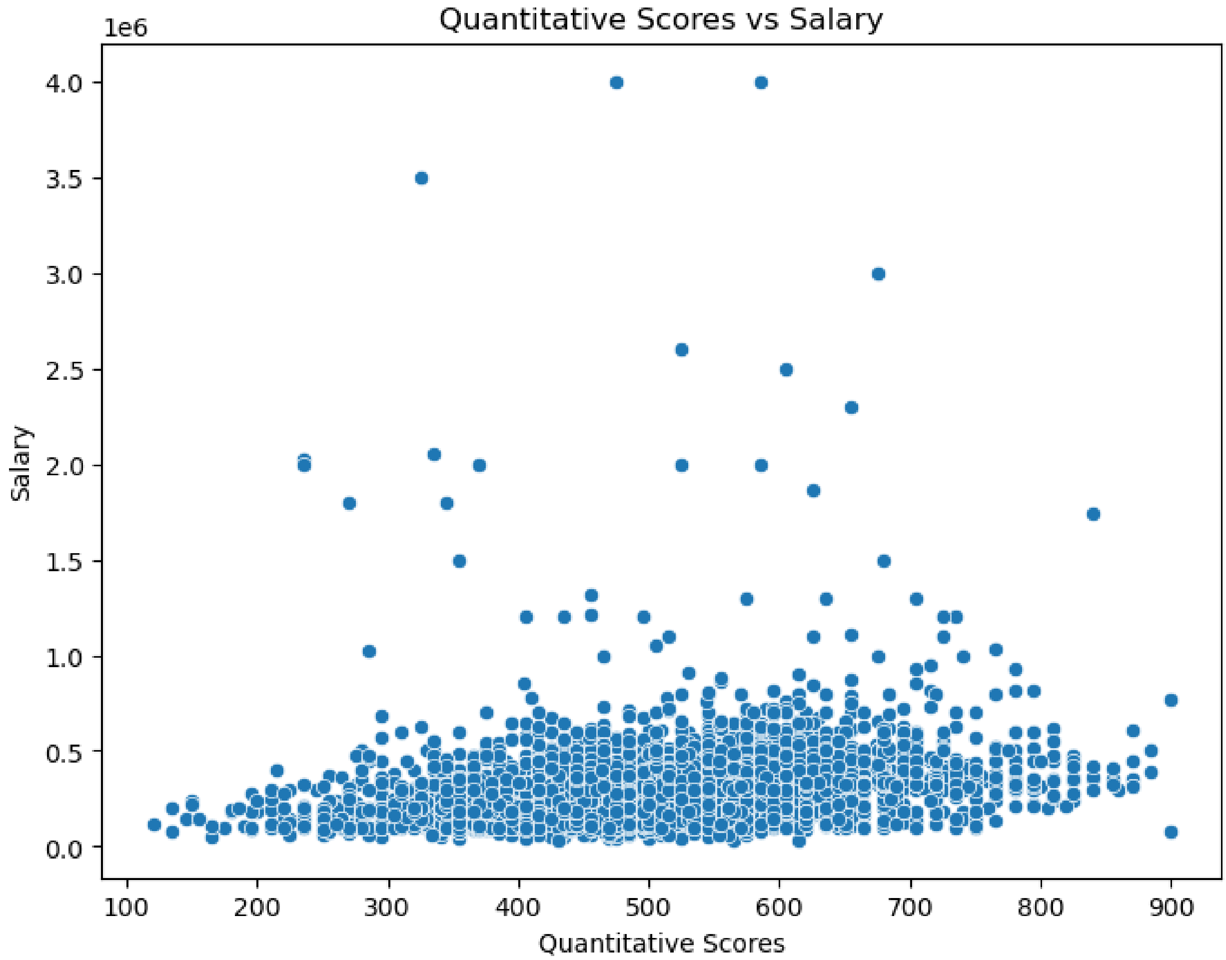
```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Logical', y='Salary', data=df)
plt.title('Logical Scores vs Salary')
plt.xlabel('Logical Scores')
plt.ylabel('Salary')
plt.show()
```

Observation:

There is no strong relationship between logical reasoning scores and salary. This indicates that cognitive skills alone may not significantly influence salary outcomes.

11. Quantitative Scores vs Salary

```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Quant', y='Salary', data=df)
plt.title('Quantitative Scores vs Salary')
plt.xlabel('Quantitative Scores')
plt.ylabel('Salary')
plt.show()
```



Observation:

Similar to logical scores, there is no clear correlation between quantitative scores and salary. High scores in this category do not necessarily lead to higher salaries.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 39 columns):
 #   Column                    Non-Null Count  Dtype
```

```
---  ------                 --------------  -----
 0   Unnamed: 0             3998 non-null   object
 1   ID                     3998 non-null   int64
 2   Salary                 3998 non-null   float64
 3   DOJ                    3998 non-null   object
 4   DOL                    3998 non-null   object
 5   Designation            3998 non-null   object
 6   JobCity                3998 non-null   object
 7   Gender                 3998 non-null   object
 8   DOB                    3998 non-null   object
 9   10percentage           3998 non-null   float64
 10  10board                3998 non-null   object
 11  12graduation           3998 non-null   int64
 12  12percentage           3998 non-null   float64
 13  12board                3998 non-null   object
 14  CollegeID              3998 non-null   int64
 15  CollegeTier            3998 non-null   int64
 16  Degree                 3998 non-null   object
 17  Specialization         3998 non-null   object
 18  collegeGPA             3998 non-null   float64
 19  CollegeCityID          3998 non-null   int64
 20  CollegeCityTier        3998 non-null   int64
 21  CollegeState           3998 non-null   object
 22  GraduationYear         3998 non-null   int64
 23  English                3998 non-null   int64
 24  Logical                3998 non-null   int64
 25  Quant                  3998 non-null   int64
 26  Domain                 3998 non-null   float64
 27  ComputerProgramming    3998 non-null   int64
 28  ElectronicsAndSemicon  3998 non-null   int64
 29  ComputerScience        3998 non-null   int64
 30  MechanicalEngg         3998 non-null   int64
 31  ElectricalEngg         3998 non-null   int64
 32  TelecomEngg            3998 non-null   int64
 33  CivilEngg              3998 non-null   int64
 34  conscientiousness      3998 non-null   float64
 35  agreeableness          3998 non-null   float64
 36  extraversion           3998 non-null   float64
 37  nueroticism            3998 non-null   float64
 38  openess_to_experience  3998 non-null   float64
dtypes: float64(10), int64(17), object(12)
memory usage: 1.2+ MB
```

1. Testing the Claim from Times of India

```
cs_graduates = df[df['Specialization'] == 'computer science']
cs_graduates
```

```
     Unnamed: 0       ID    Salary           DOJ           DOL  \
3256      train  1250504  400000.0  9/1/14 0:00  2/1/15 0:00
3505      train   455860  180000.0  4/1/13 0:00  7/1/13 0:00

                  Designation    JobCity Gender            DOB
10percentage  \
3256  associate software engg  Hyderabad      m   2/25/90 0:00
69.5
3505               programmer   Phagwara      f  12/27/89 0:00
73.0

      ... ComputerScience  MechanicalEngg  ElectricalEngg TelecomEngg
\
3256  ...             500              -1              -1          -1

3505  ...             469              -1              -1          -1


      CivilEngg  conscientiousness agreeableness extraversion
nueroticism  \
3256         -1             0.9900       -0.2871       0.7785      -
1.6289
3505         -1            -0.0696        0.5008       0.8171
0.4442

      openess_to_experience
3256                -0.8608
3505                 0.0284

[2 rows x 39 columns]
```

```python
cs_graduates_roles =
cs_graduates[cs_graduates['Designation'].isin(['programmer', 'software
engineer', 'hardware engineer', 'associate software engg'])]
cs_graduates_roles
```

```
     Unnamed: 0       ID    Salary           DOJ           DOL  \
3256      train  1250504  400000.0  9/1/14 0:00  2/1/15 0:00
3505      train   455860  180000.0  4/1/13 0:00  7/1/13 0:00

                  Designation    JobCity Gender            DOB
10percentage  \
3256  associate software engg  Hyderabad      m   2/25/90 0:00
69.5
3505               programmer   Phagwara      f  12/27/89 0:00
73.0

      ... ComputerScience  MechanicalEngg  ElectricalEngg TelecomEngg
\
3256  ...             500              -1              -1          -1
```

```
3505  ...              469              -1              -1          -1


      CivilEngg  conscientiousness agreeableness extraversion
nueroticism  \
3256         -1             0.9900       -0.2871       0.7785      -
1.6289
3505         -1            -0.0696        0.5008       0.8171
0.4442

      openess_to_experience
3256                -0.8608
3505                 0.0284

[2 rows x 39 columns]
```

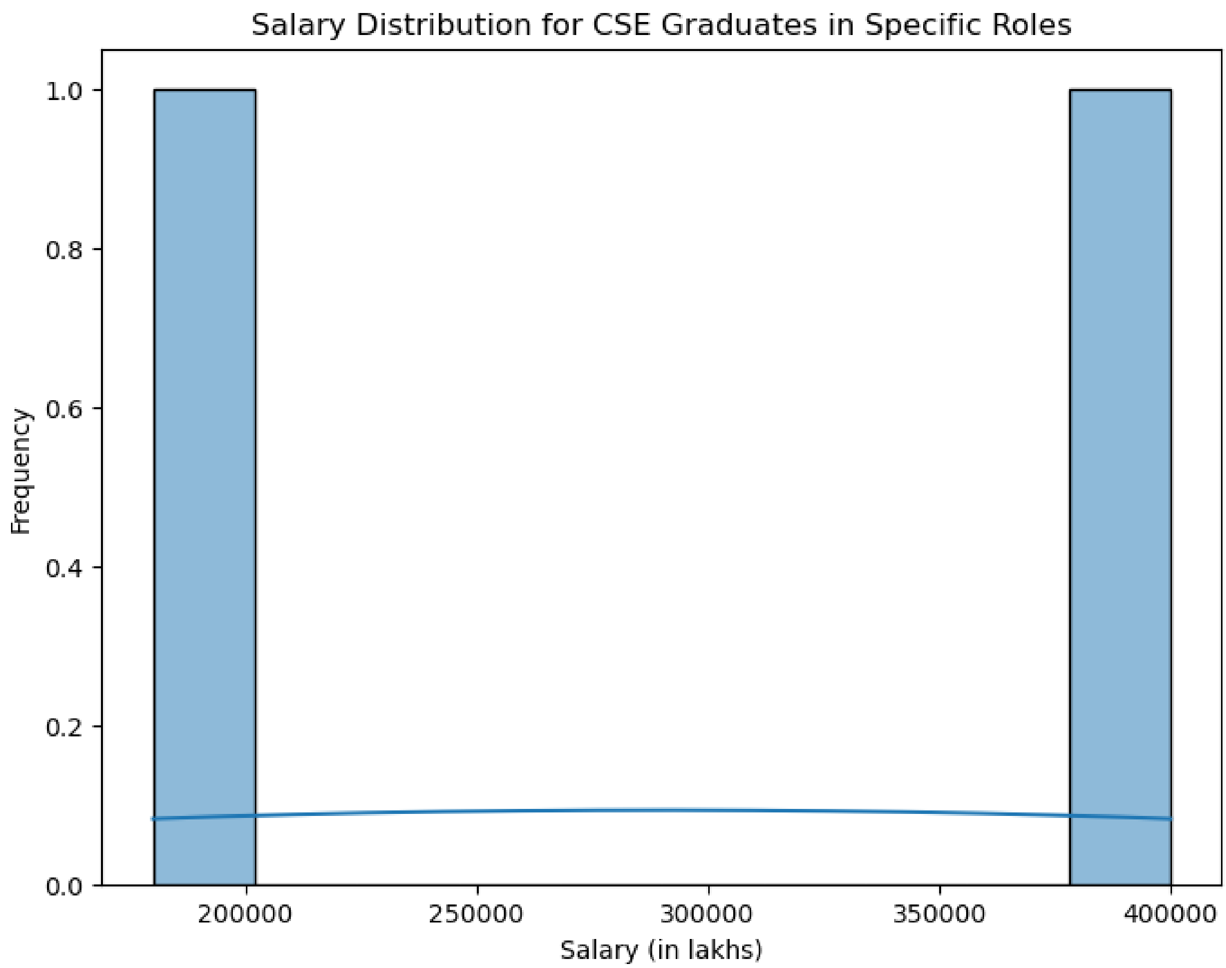
```
cs_graduates_roles_salary = cs_graduates_roles['Salary']
cs_graduates_roles_salary
```

```
3256    400000.0
3505    180000.0
Name: Salary, dtype: float64
```

```
cs_graduates_roles_salary.describe()
```

```
count         2.000000
mean     290000.000000
std      155563.491861
min      180000.000000
25%      235000.000000
50%      290000.000000
75%      345000.000000
max      400000.000000
Name: Salary, dtype: float64
```

```
plt.figure(figsize=(8, 6))
sns.histplot(cs_graduates_roles_salary, bins=10, kde=True)
plt.title('Salary Distribution for CSE Graduates in Specific Roles')
plt.xlabel('Salary (in lakhs)')
plt.ylabel('Frequency')
plt.show()
```



Observation:

The summary statistics (mean, median) will show whether the average salary of graduates in these roles falls within the range of 2 to 4 lakhs, as mentioned in the article. The histogram will help visually assess the salary distribution for these roles.

2. The relationship between gender and specialization.

```
contingency_table = pd.crosstab(df['Gender'], df['Specialization'])
contingency_table

Specialization  aeronautical engineering  \
Gender
f                                      1
m                                      2

Specialization  applied electronics and instrumentation  \
```

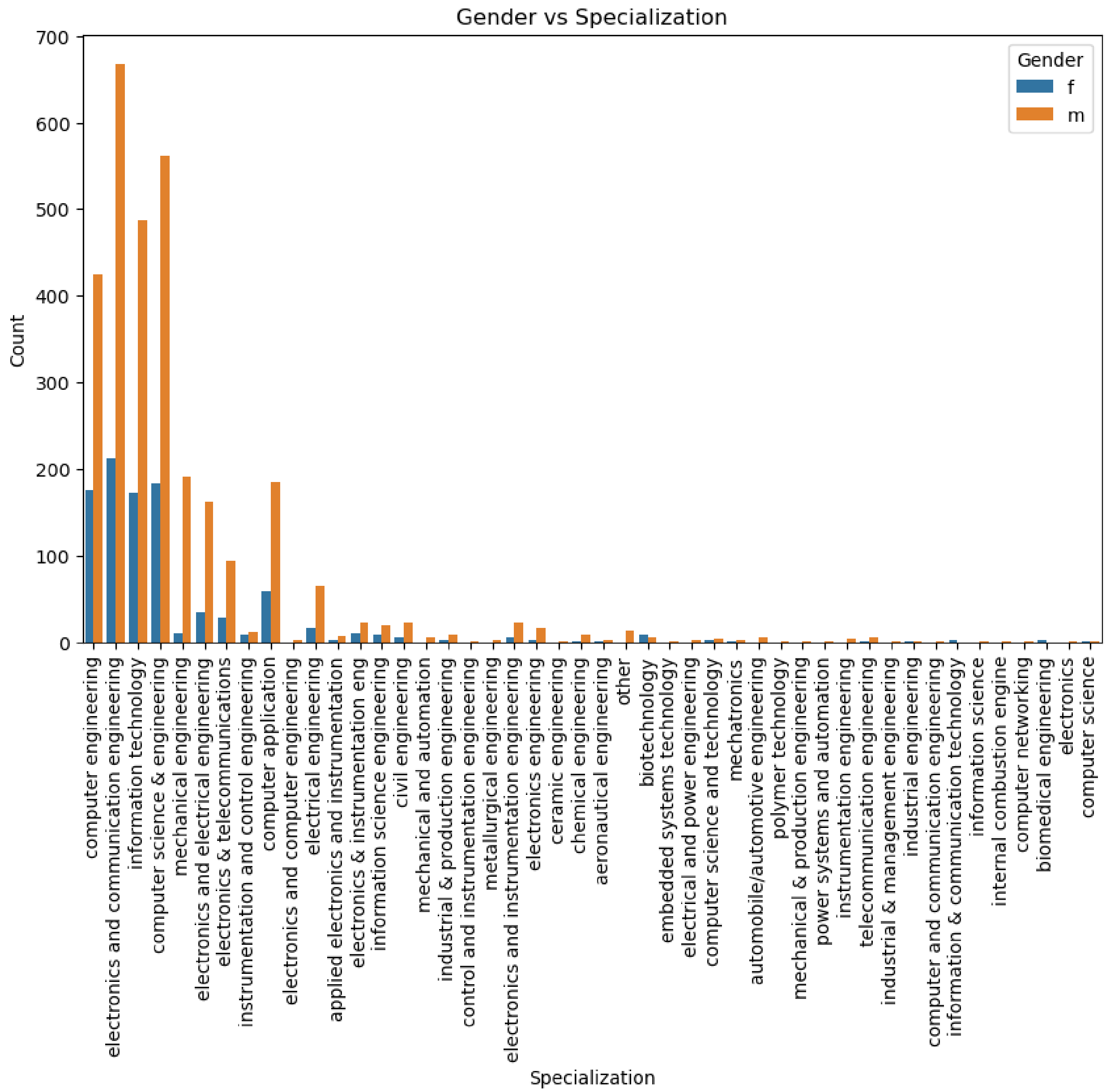
```
Gender
f                                                      2
m                                                      7

Specialization  automobile/automotive engineering  biomedical
engineering  \
Gender

f                                                0
2
m                                                5
0

Specialization  biotechnology  ceramic engineering  chemical
engineering  \
Gender

f                           9                    0
1
m                           6                    1
8

Specialization  civil engineering  computer and communication
engineering  \
Gender

f                               6
0
m                              23
1

Specialization  computer application  ...  internal combustion engine
\
Gender                                ...

f                                 59  ...                           0

m                                185  ...                           1


Specialization  mechanical & production engineering  \
Gender
f                                                  0
m                                                  1

Specialization  mechanical and automation  mechanical engineering  \
Gender
f                                       0                      10
m                                       5                     191
```

```
Specialization  mechatronics  metallurgical engineering  other  \
Gender
f                          1                          0      0
m                          3                          2     13

Specialization  polymer technology  power systems and automation  \
Gender
f                                0                             0
m                                1                             1

Specialization  telecommunication engineering
Gender
f                                           1
m                                           5

[2 rows x 46 columns]
```

```
from scipy.stats import chi2_contingency

chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-square statistic: {chi2}")
print(f"P-value: {p_value}")
```

```
Chi-square statistic: 104.46891913608455
P-value: 1.2453868176976918e-06
```

```
plt.figure(figsize=(10, 6))
sns.countplot(x='Specialization', hue='Gender', data=df)
plt.xticks(rotation=90)
plt.title('Gender vs Specialization')
plt.xlabel('Specialization')
plt.ylabel('Count')
plt.show()
```



Observation: If the p-value is less than 0.05, we can reject the null hypothesis and conclude that there is a relationship between gender and specialization. If the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that gender does not have a significant impact on the choice of specialization.

Conclusion:

The dataset provides valuable insights into the employment outcomes of engineering graduates, revealing a diverse range of salaries, job locations, and specializations. While technical and cognitive skills play a significant role in determining job outcomes, the analysis also highlights that demographic factors like gender and specialization are not strongly related. This suggests that job opportunities in engineering are largely merit-based, with skills and qualifications being the key determinants for career progression.