Importing required libraries

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
import numpy as np
warnings.filterwarnings('ignore')
from scipy import stats as st

# Loading the dataset
df = pd.read_csv(r"C:\Users\saich\Downloads\AMCAT.csv")

df
```

```
     Unnamed: 0      ID   Salary        DOJ                  DOL  \
0         train  203097   420000 2012-06-01              present
1         train  579905   500000 2013-09-01              present
2         train  810601   325000 2014-06-01              present
3         train  267447  1100000 2011-07-01              present
4         train  343523   200000 2014-03-01  2015-03-01 00:00:00
...         ...     ...      ...        ...                  ...
3993      train   47916   280000 2011-10-01  2012-10-01 00:00:00
3994      train  752781   100000 2013-07-01  2013-07-01 00:00:00
3995      train  355888   320000 2013-07-01              present
3996      train  947111   200000 2014-07-01  2015-01-01 00:00:00
3997      train  324966   400000 2013-02-01              present

                      Designation          JobCity Gender        DOB
\
0         senior quality engineer        Bangalore      f 1990-02-19

1               assistant manager           Indore      m 1989-10-04

2                systems engineer          Chennai      f 1992-08-03

3        senior software engineer          Gurgaon      m 1989-12-05

4                             get          Manesar      m 1991-02-27

...                           ...              ...    ...        ...

3993            software engineer        New Delhi      m 1987-04-15

3994             technical writer        Hyderabad      f 1992-08-27

3995  associate software engineer        Bangalore      m 1991-07-03

3996           software developer  Asifabadbanglore      f 1992-03-20

3997      senior systems engineer          Chennai      f 1991-02-26
```

```
      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]
```

```python
df.drop("Unnamed: 0",axis=1,inplace=True)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 38 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     3998 non-null   int64
 1   Salary                 3998 non-null   int64
 2   DOJ                    3998 non-null   datetime64[ns]
 3   DOL                    3998 non-null   object
 4   Designation            3998 non-null   object
 5   JobCity                3998 non-null   object
 6   Gender                 3998 non-null   object
 7   DOB                    3998 non-null   datetime64[ns]
 8   10percentage           3998 non-null   float64
 9   10board                3998 non-null   object
 10  12graduation           3998 non-null   int64
 11  12percentage           3998 non-null   float64
 12  12board                3998 non-null   object
 13  CollegeID              3998 non-null   int64
 14  CollegeTier            3998 non-null   int64
 15  Degree                 3998 non-null   object
 16  Specialization         3998 non-null   object
 17  collegeGPA             3998 non-null   float64
 18  CollegeCityID          3998 non-null   int64
 19  CollegeCityTier        3998 non-null   int64
 20  CollegeState           3998 non-null   object
 21  GraduationYear         3998 non-null   int64
 22  English                3998 non-null   int64
 23  Logical                3998 non-null   int64
```

```
 24  Quant                    3998 non-null   int64
 25  Domain                   3998 non-null   float64
 26  ComputerProgramming      3998 non-null   int64
 27  ElectronicsAndSemicon    3998 non-null   int64
 28  ComputerScience          3998 non-null   int64
 29  MechanicalEngg           3998 non-null   int64
 30  ElectricalEngg           3998 non-null   int64
 31  TelecomEngg              3998 non-null   int64
 32  CivilEngg                3998 non-null   int64
 33  conscientiousness        3998 non-null   float64
 34  agreeableness            3998 non-null   float64
 35  extraversion             3998 non-null   float64
 36  nueroticism              3998 non-null   float64
 37  openess_to_experience    3998 non-null   float64
dtypes: datetime64[ns](2), float64(9), int64(18), object(9)
memory usage: 1.2+ MB
```

```
df.shape
```

```
(3998, 38)
```

Exploratory Data Analysis

Getting the insights from the data which includes

- Missing values
- Duplicated values
- Ouliers
- Distributions
- Relationships

```
df.describe()
```

```
                 ID        Salary                            DOJ  \
count  3.998000e+03  3.998000e+03                           3998
mean   6.637945e+05  3.076998e+05  2013-07-02 11:04:10.325162496
min    1.124400e+04  3.500000e+04            1991-06-01 00:00:00
25%    3.342842e+05  1.800000e+05            2012-10-01 00:00:00
50%    6.396000e+05  3.000000e+05            2013-11-01 00:00:00
75%    9.904800e+05  3.700000e+05            2014-07-01 00:00:00
max    1.298275e+06  4.000000e+06            2015-12-01 00:00:00
std    3.632182e+05  2.127375e+05                            NaN

                                 DOB  10percentage  12graduation  \
count                           3998   3998.000000   3998.000000
mean   1990-12-06 06:01:15.637819008     77.925443   2008.087544
min              1977-10-30 00:00:00     43.000000   1995.000000
25%              1989-11-16 06:00:00     71.680000   2007.000000
```

```
50%          1991-03-07 12:00:00     79.150000   2008.000000
75%          1992-03-13 18:00:00     85.670000   2009.000000
max          1997-05-27 00:00:00     97.760000   2013.000000
std                          NaN      9.850162      1.653599

       12percentage     CollegeID  CollegeTier   collegeGPA  ...  \
count   3998.000000   3998.000000  3998.000000  3998.000000  ...
mean      74.466366   5156.851426     1.925713    71.486171  ...
min       40.000000      2.000000     1.000000     6.450000  ...
25%       66.000000    494.000000     2.000000    66.407500  ...
50%       74.400000   3879.000000     2.000000    71.720000  ...
75%       82.600000   8818.000000     2.000000    76.327500  ...
max       98.700000  18409.000000     2.000000    99.930000  ...
std       10.999933   4802.261482     0.262270     8.167338  ...

       ComputerScience  MechanicalEngg  ElectricalEngg  TelecomEngg  \
count      3998.000000     3998.000000     3998.000000  3998.000000
mean         90.742371       22.974737       16.478739    31.851176
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
std         175.273083       98.123311       87.585634   104.852845

         CivilEngg  conscientiousness  agreeableness  extraversion  \
count  3998.000000        3998.000000    3998.000000   3998.000000
mean      2.683842          -0.037831       0.146496      0.002763
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
std      36.658505           1.028666       0.941782      0.951471

       nueroticism  openess_to_experience
count  3998.000000            3998.000000
mean     -0.169033              -0.138110
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
std       1.007580               1.008075

[8 rows x 29 columns]
```

```
df.isna().sum()
```

```
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
```

Univariate Analysis

```
# Outliers Detection (IQR method):
Q1 = df['Salary'].quantile(0.25)
Q3 = df['Salary'].quantile(0.75)
```

```
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['Salary'] < lower_bound) | (df['Salary'] >
upper_bound)]

lower_bound

-105000.0

upper_bound

655000.0

outliers

           ID   Salary        DOJ                  DOL  \
3      267447  1100000 2011-07-01              present
76     361583   800000 2012-06-01              present
92    1250429  1500000 2014-11-01  2014-07-01 00:00:00
123    312164  1200000 2010-07-01  2011-07-01 00:00:00
128    206734   675000 2011-11-01              present
...       ...      ...        ...                  ...
3823   918568   775000 2014-08-01              present
3904   267121   850000 2011-09-01              present
3912   231229   730000 2013-07-01              present
3961   230702   700000 2011-07-01  2014-09-01 00:00:00
3992   344407   800000 2014-04-01  2015-04-01 00:00:00

                     Designation      JobCity Gender        DOB
10percentage  \
3        senior software engineer      Gurgaon      m 1989-12-05
85.60
76              software engineer    Bangalore      m 1991-01-25
93.44
92          application developer    Hyderabad      m 1992-01-04
79.00
123              engineer trainee  Maharajganj      m 1988-04-25
59.80
128      senior software engineer        Noida      m 1988-11-07
60.00
...                           ...          ...    ...        ...
...
3823  mechanical design engineer       Dammam      m 1991-01-12
87.40
3904         operations assistant        Noida      m 1989-01-05
83.40
3912           research scientist        Pune       m 1989-11-15
84.67
3961            planning engineer      Rajpura      m 1987-12-27
84.20
```

```
3992                        manager      Rajkot      m 1990-06-22
73.00

                    10board  ...  ComputerScience  MechanicalEngg  \
3                      cbse  ...               -1              -1
76     karnataka state board  ...              -1              -1
92               state board  ...             346              -1
123                    icse  ...               -1             206
128                       0  ...               -1              -1
...                     ...  ...              ...             ...
3823                   cbse  ...               -1             469
3904                   cbse  ...               -1              -1
3912                      0  ...               -1              -1
3961                      0  ...               -1              -1
3992                      0  ...               -1              -1

     ElectricalEngg  TelecomEngg  CivilEngg conscientiousness
agreeableness  \
3                -1           -1         -1            0.0464
0.3448
76               -1           -1         -1           -0.4173
0.9688
92               -1           -1         -1            0.4155
0.5454
123              -1           -1         -1            0.2009
1.1248
128              -1           -1         -1           -0.8810        -
0.2793
...             ...          ...        ...               ...
...
3823             -1           -1         -1           -0.8772        -
0.1206
3904             -1           -1         -1           -0.8810
0.1888
3912             -1           -1         -1           -1.3447        -
1.0593
3961             -1           -1        460           -1.3447
0.0328
3992             -1           -1        480            0.3555        -
0.9033

      extraversion  nueroticism  openess_to_experience
3          -0.3440     -0.40780                -0.9194
76         -0.1988     -0.29020                 0.3049
92          0.9322     -0.61470                 0.8637
123         1.1074     -1.11280                 0.9763
128        -0.6343     -0.64280                -2.9731
...            ...          ...                    ...
3823       -0.1437     -0.23440                -0.0943
3904       -0.1988     -0.05520                -1.0774
```

```
3912        0.6720      1.00240                -1.7093
3961       -2.3759     -0.99530                 0.3444
3992        0.9623      0.64983                -0.4229

[109 rows x 38 columns]
```

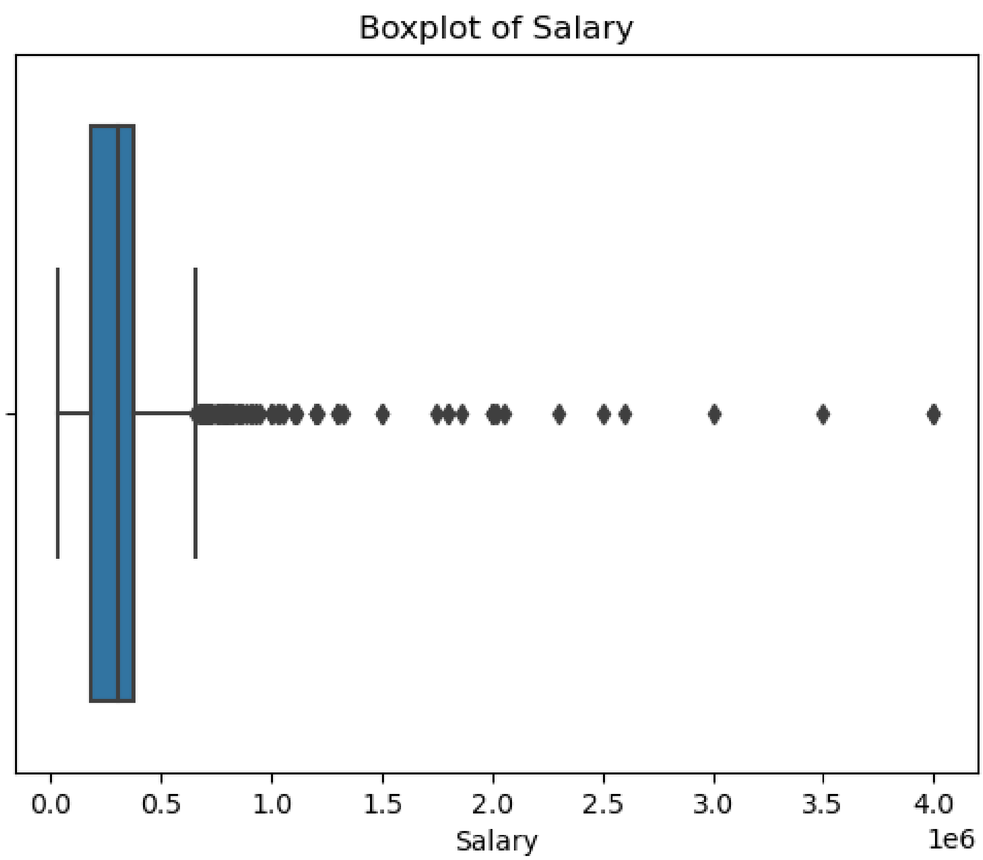
```python
# Summary for Categorical Variables:
df['Gender'].value_counts()
df['Specialization'].value_counts()
```

```
Specialization
electronics and communication engineering      880
computer science & engineering                 744
information technology                         660
computer engineering                           600
computer application                           244
mechanical engineering                         201
electronics and electrical engineering         196
electronics & telecommunications               121
electrical engineering                          82
electronics & instrumentation eng               32
civil engineering                               29
electronics and instrumentation engineering     27
information science engineering                 27
instrumentation and control engineering         20
electronics engineering                         19
biotechnology                                   15
other                                           13
industrial & production engineering             10
applied electronics and instrumentation          9
chemical engineering                             9
computer science and technology                  6
telecommunication engineering                    6
mechanical and automation                        5
automobile/automotive engineering                5
instrumentation engineering                      4
mechatronics                                     4
aeronautical engineering                         3
electronics and computer engineering             3
electrical and power engineering                 2
biomedical engineering                           2
information & communication technology           2
industrial engineering                           2
computer science                                 2
metallurgical engineering                        2
power systems and automation                     1
control and instrumentation engineering          1
mechanical & production engineering              1
embedded systems technology                      1
polymer technology                               1
```

```
computer and communication engineering         1
information science                            1
internal combustion engine                     1
computer networking                            1
ceramic engineering                            1
electronics                                    1
industrial & management engineering            1
Name: count, dtype: int64
```

Outliers in Salary feature

```
sns.boxplot(x='Salary', data=df)
plt.title("Boxplot of Salary")
plt.show()
```

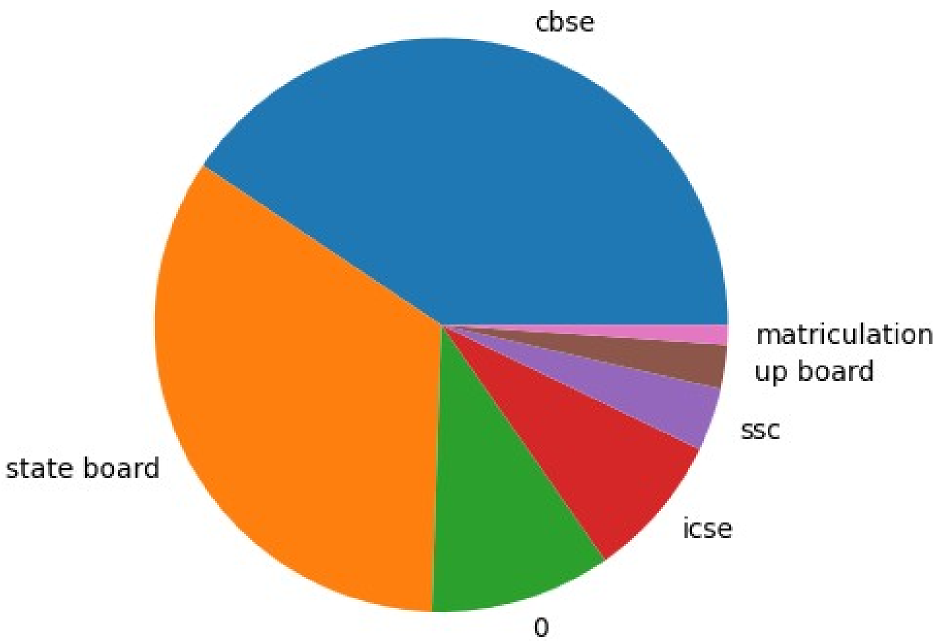


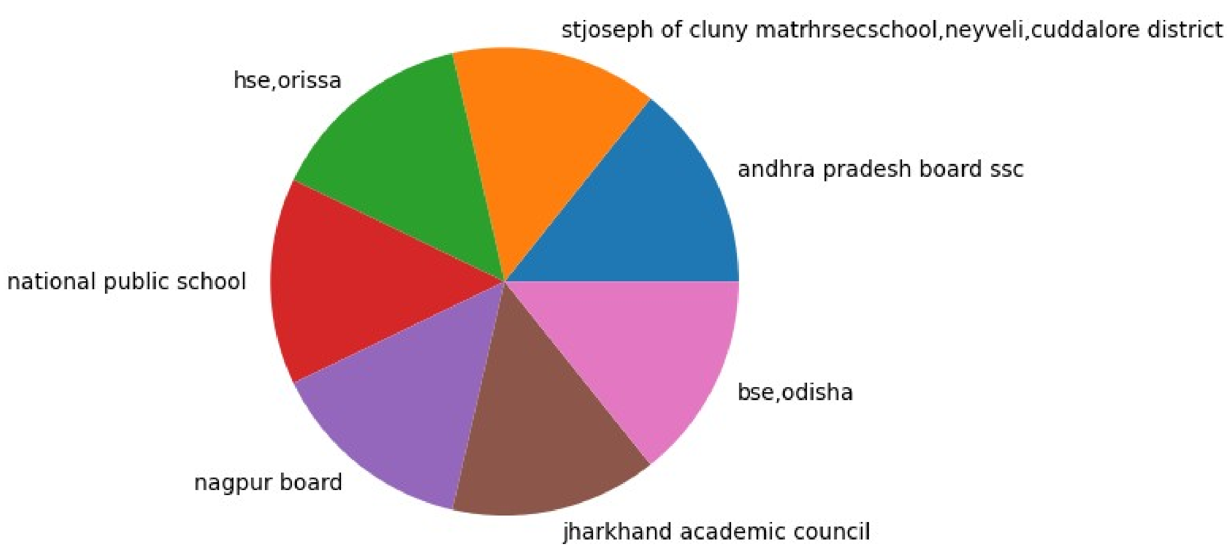
Observation:

- Outliers for salary can be seen as points beyond the whiskers.
- This can help identify extreme cases in salary offers.

Pie Charts of Top and Least Common '10board' Categories

```
a=pd.DataFrame(df['10board'].value_counts().head(7))
plt.pie(a['count'],labels=a.index)
plt.show()
a=pd.DataFrame(df['10board'].value_counts().tail(7))
plt.pie(a['count'],labels=a.index)
plt.show()
```



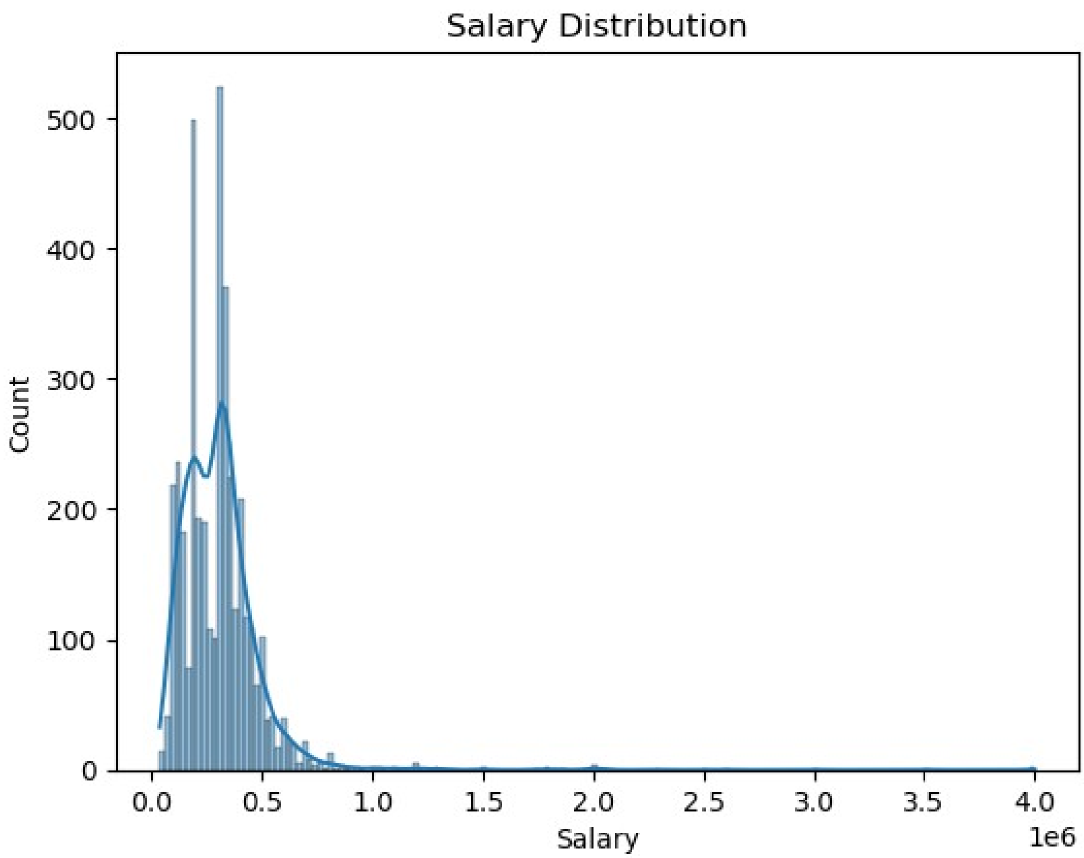


Observation :

- The pie charts illustrate the most and least common educational boards for 10th-grade students. This helps in identifying the distribution of educational backgrounds.
- state board and CBSE are most leading educational boards.

Salary Distribution (Histogram with KDE)

```
sns.histplot(df['Salary'], kde=True)
plt.title("Salary Distribution")
plt.show()
```

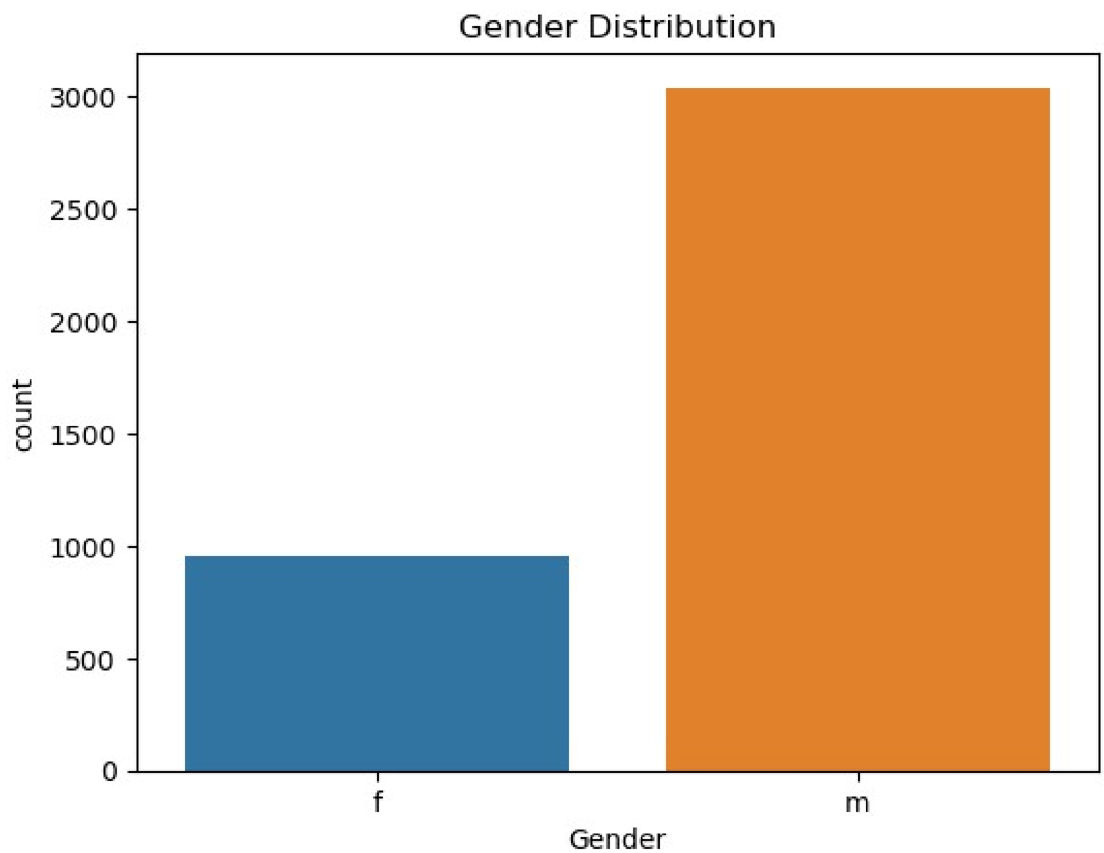


Observation:

- The distribution of salary is likely skewed, with a majority of candidates earning below a certain threshold.
- The KDE (Kernel Density Estimate) adds a smooth curve to indicate the probability distribution of salaries.

Gender Distribution (Countplot)

```
sns.countplot(x='Gender', data=df)
plt.title("Gender Distribution")
plt.show()
```



Observation:

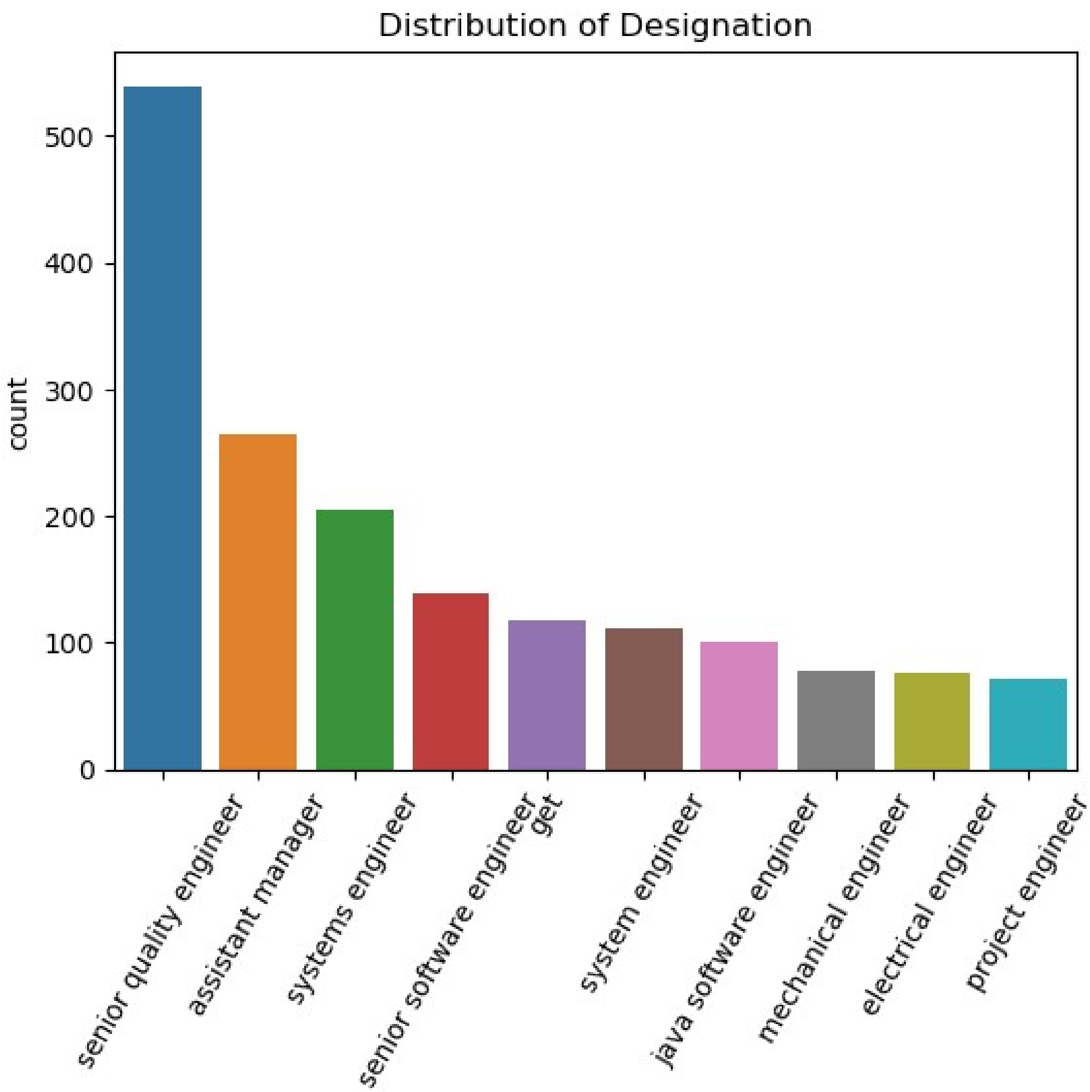
- We can observe if there is a gender imbalance in the dataset.
- Similarly, countplots can help identify popular specializations or designations.

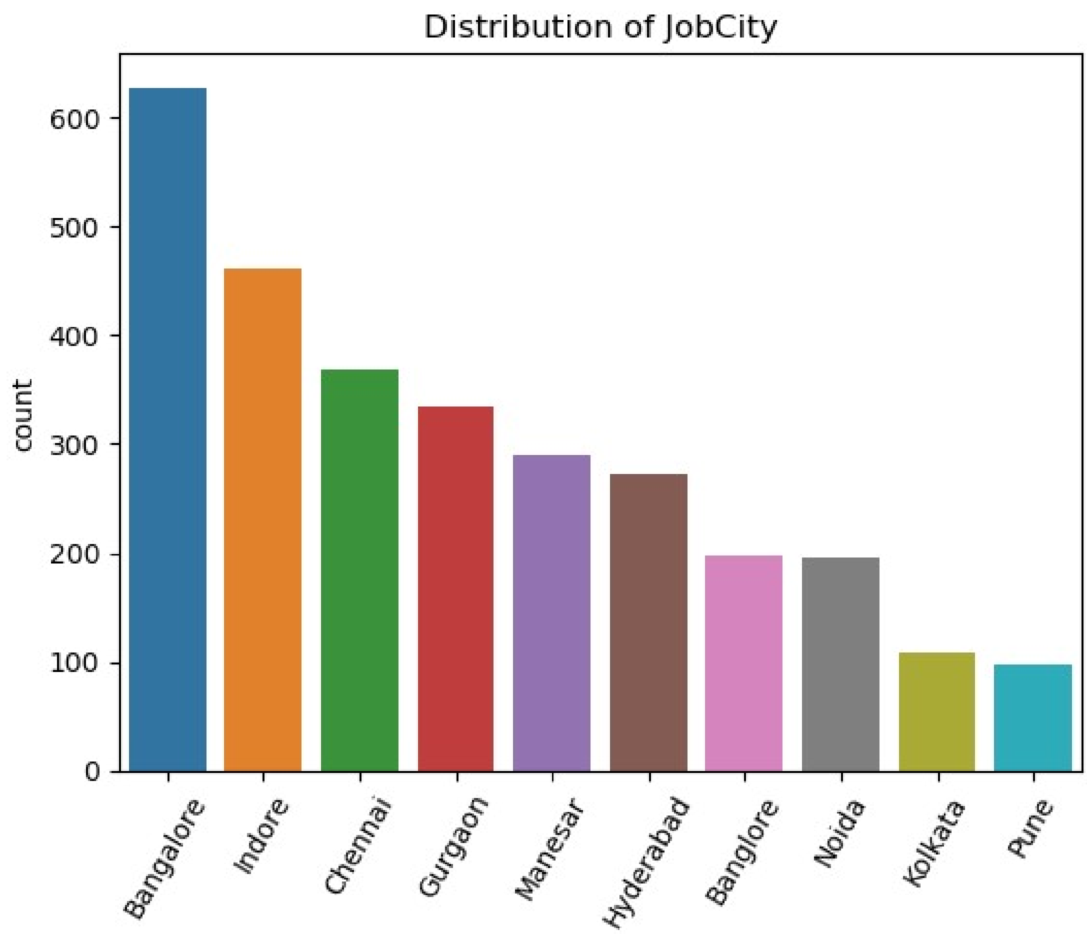
```
import matplotlib.pyplot as plt
import seaborn as sns

# Create a figure with subplots, adjusting the number of rows and
columns
fig, axes = plt.subplots(nrows=10, ncols=4, figsize=(20, 30))  #
Adjust grid size based on number of columns
axes = axes.flatten()  # Flatten to 1D array for easy iteration

# Loop through columns to create plots
for i, col in enumerate(df.columns):
    if df[col].dtype == 'object' or df[col].dtype.name == 'category':
        sns.countplot(x=col, data=df, ax=axes[i])
        axes[i].set_title(f'Distribution of {col} (Categorical)')
    elif df[col].dtype == 'datetime64[ns]':
        # Plotting datetime data (years)
        df[col].dt.year.value_counts().plot(kind='bar', ax=axes[i])
        axes[i].set_title(f'Distribution of {col} (Year)')
    else:
```

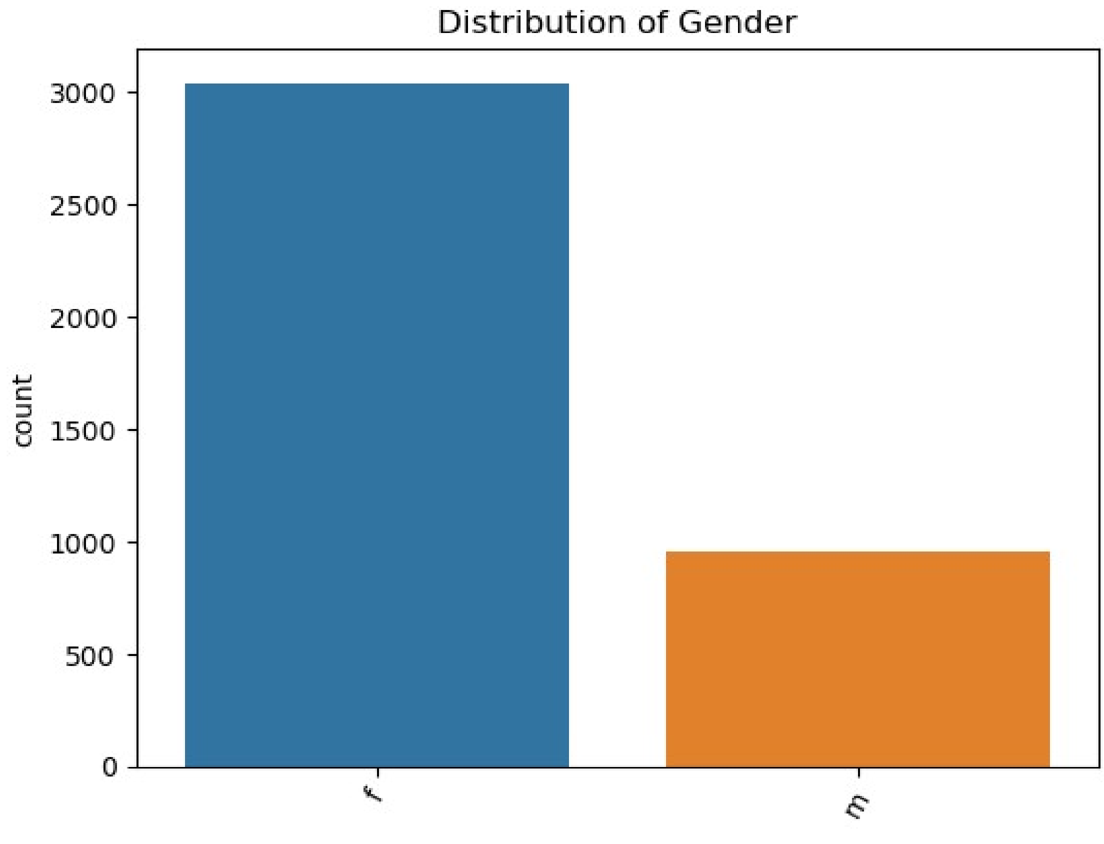
```
        sns.histplot(df[col], kde=True, ax=axes[i])
        axes[i].set_title(f'Distribution of {col} (Continuous)')

# Automatically adjust subplot layout to avoid overlapping
plt.tight_layout()
plt.show()
```

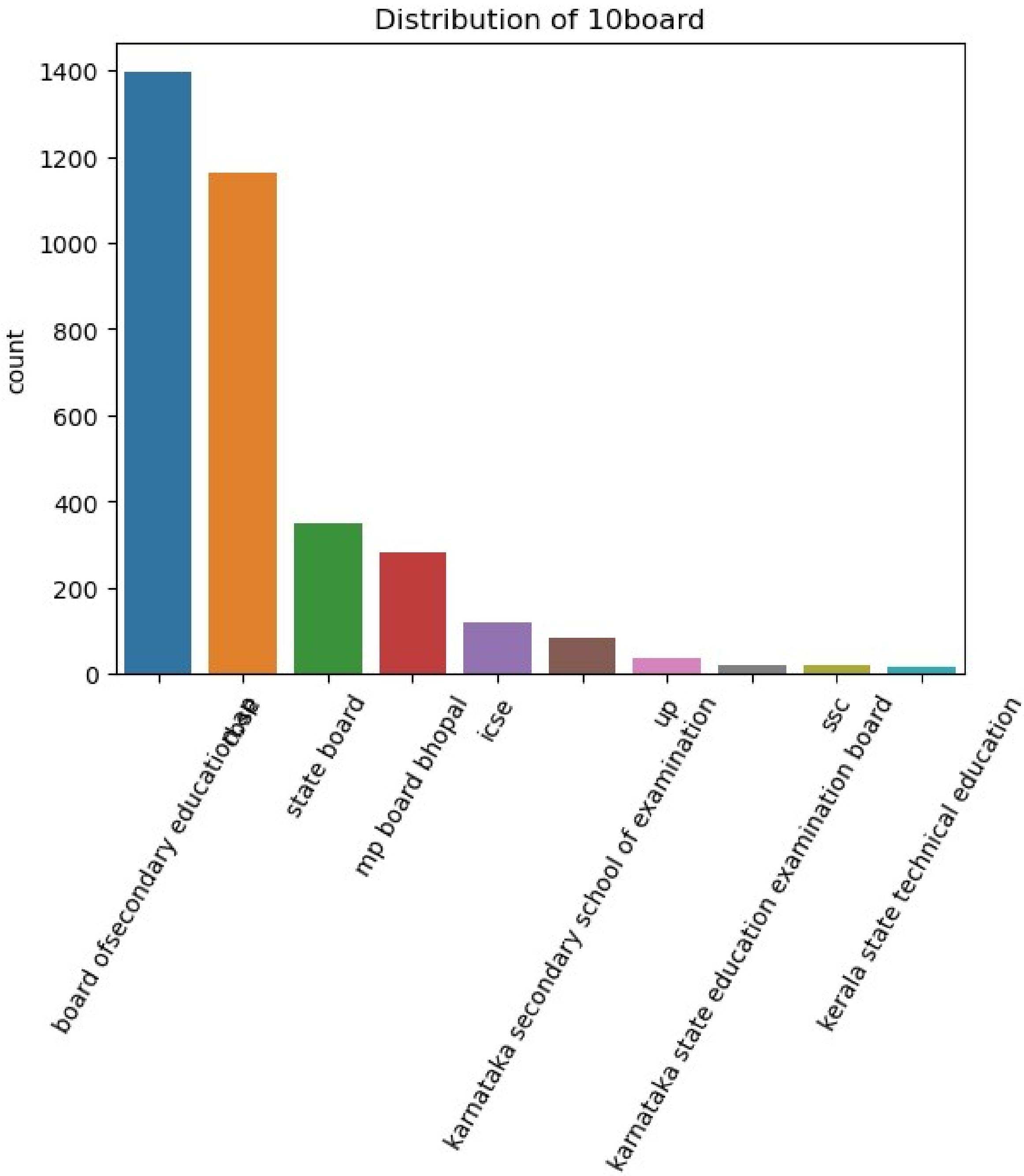
Distribution of ID (Continuous)
ID
Count
Distribution of Salary (Continuous)
Salary
Distribution of DOJ (Year)
DOJ
Distribution of DOL (Year)
DOL
Distribution of Designation (Categorical)
Designation
count
Distribution of JobCity (Categorical)
JobCity
Distribution of Gender (Categorical)
Gender
f
m
Distribution of DOB (Year)
DOB
Distribution of 10percentage (Continuous)
10percentage
Distribution of 10board (Categorical)
10board
Distribution of 12graduation (Continuous)
12graduation
Distribution of 12percentage (Continuous)
12percentage
Distribution of 12board (Categorical)
12board
Distribution of CollegeID (Continuous)
CollegeID
Distribution of CollegeTier (Continuous)
CollegeTier
Distribution of Degree (Categorical)
Degree
Distribution of Specialization (Categorical)
Specialization
Distribution of collegeGPA (Continuous)
collegeGPA
Distribution of CollegeCityID (Continuous)
CollegeCityID
Distribution of CollegeCityTier (Continuous)
CollegeCityTier
Distribution of CollegeState (Categorical)
CollegeState
Distribution of GraduationYear (Continuous)
GraduationYear
Distribution of English (Continuous)
English
Distribution of Logical (Continuous)
Logical
Distribution of Quant (Continuous)
Quant
Distribution of Domain (Continuous)
Domain
Distribution of ComputerProgramming (Continuous)
ComputerProgramming
Distribution of ElectronicsAndSemicon (Continuous)
ElectronicsAndSemicon
Distribution of ComputerScience (Continuous)
ComputerScience
Distribution of MechanicalEngg (Continuous)
MechanicalEngg
Distribution of ElectricalEngg (Continuous)
ElectricalEngg
Distribution of TelecomEngg (Continuous)
TelecomEngg
Distribution of CivilEngg (Continuous)
CivilEngg
Distribution of conscientiousness (Continuous)
conscientiousness
Distribution of agreeableness (Continuous)
agreeableness
Distribution of extraversion (Continuous)
extraversion
Distribution of nueroticism (Continuous)
Distribution of openess_to_experience (Continuous)

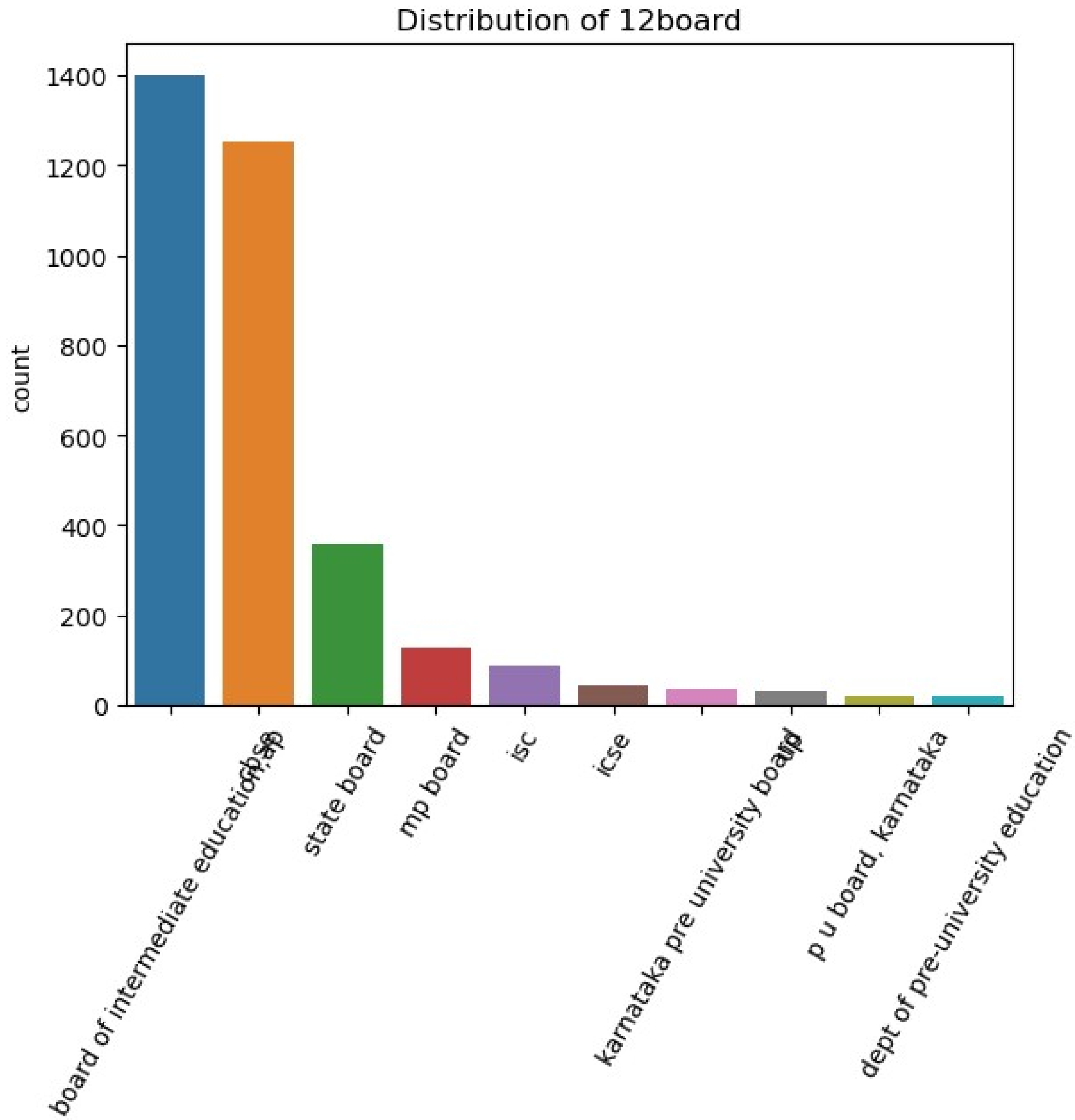
```
for i in df.columns:
    if df[i].dtype=="object":
        sns.barplot(x=df[i].unique()[:10],y=df[i].value_counts()[:10])
        plt.title("Distribution of {}".format(i))
        plt.xticks(rotation=60)
        plt.show()
```



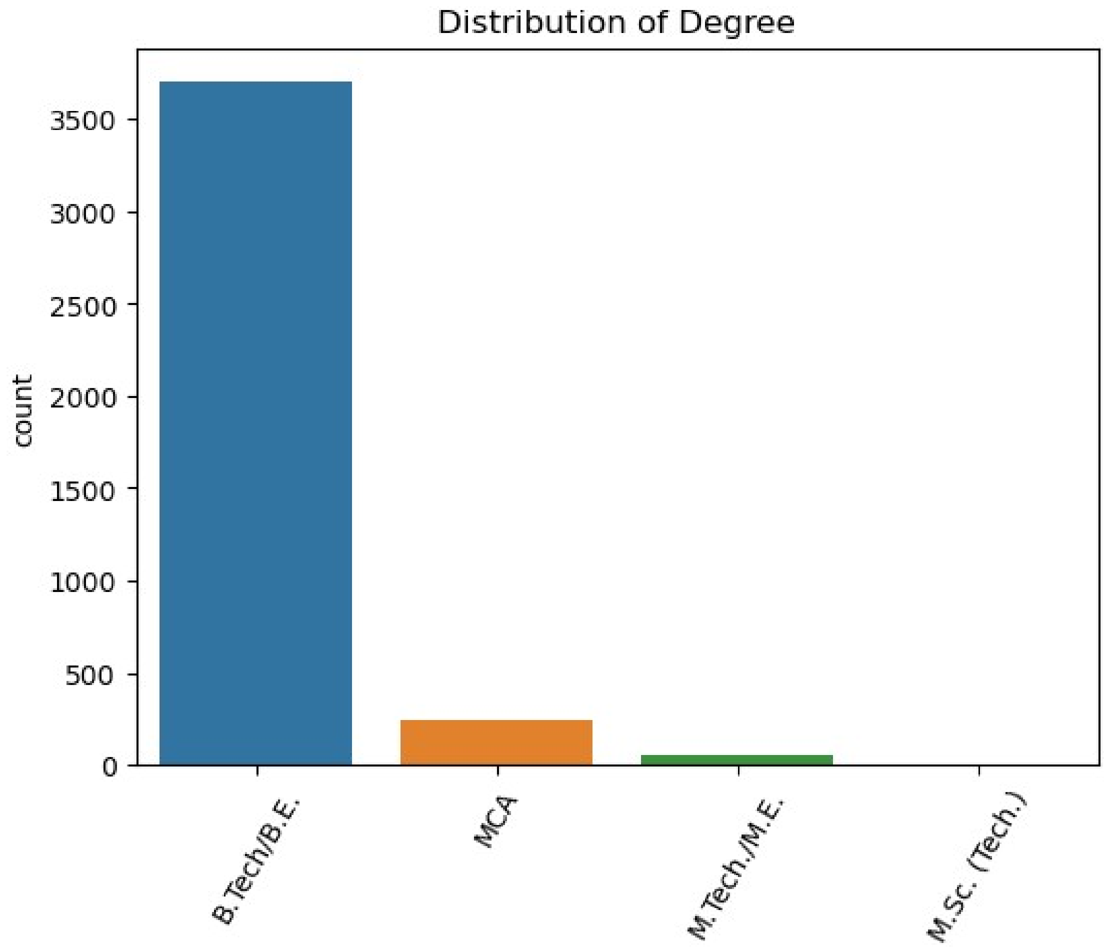

Distribution of JobCity
count
0
100
200
300
400
500
600
Bangalore
Indore
Chennai
Gurgaon
Manesar
Hyderabad
Banglore
Noida
Kolkata
Pune

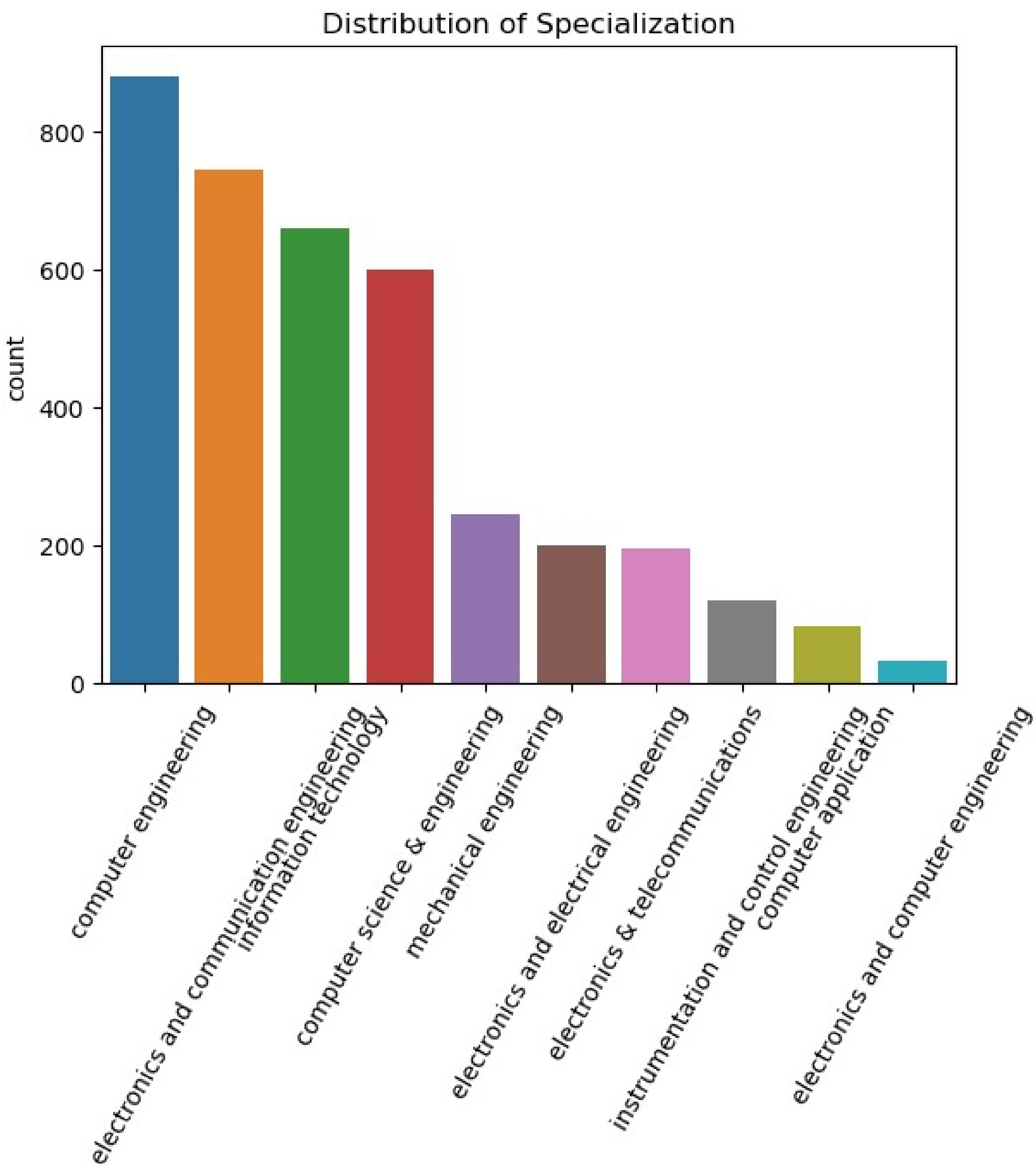

Distribution of Gender
count
3000
2500
2000
1500
1000
500
0
f
m

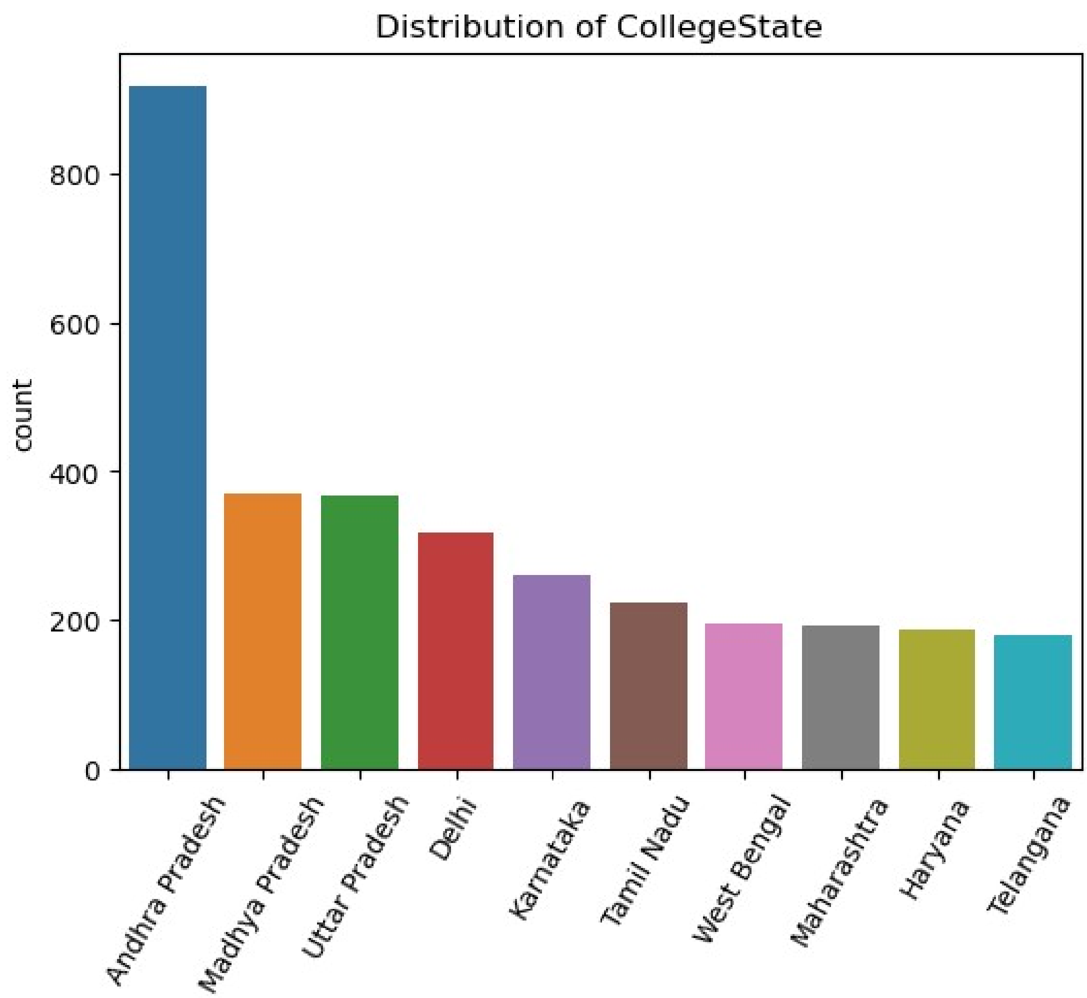

Distribution of 10board
count
1400
1200
1000
800
600
400
200
0
board ofsecondary educatio
state board
mp board bhopal
icse
karnataka secondary school of examination
up
karnataka state education examination board
ssc
kerala state technical education



Distribution of 12board
count
1400
1200
1000
800
600
400
200
0
board of intermediate education,ap
cbse
state board
mp board
isc
icse
karnataka pre university board
up
p u board, karnataka
dept of pre-university education


Distribution of Degree
count
3500
3000
2500
2000
1500
1000
500
0
B.Tech/B.E.
MCA
M.Tech./M.E.
M.Sc. (Tech.)


Distribution of Specialization
count
0
200
400
600
800
computer engineering
electronics and communication engineering
information technology
computer science & engineering
mechanical engineering
electronics and electrical engineering
electronics & telecommunications
instrumentation and control engineering
computer application
electronics and computer engineering



Bivariate Analysis

```
# Categorical-Numerical Analysis:
df.groupby('Specialization')['Salary'].mean()
```

```
Specialization
aeronautical engineering                              148333.333333
applied electronics and instrumentation               348333.333333
automobile/automotive engineering                     222000.000000
biomedical engineering                                290000.000000
biotechnology                                         254333.333333
ceramic engineering                                   335000.000000
chemical engineering                                  370000.000000
civil engineering                                     381206.896552
computer and communication engineering                120000.000000
computer application                                  280389.344262
computer engineering                                  374100.000000
computer networking                                   565000.000000
computer science                                      290000.000000
```

```
computer science & engineering                     277439.516129
computer science and technology                    245833.333333
control and instrumentation engineering            305000.000000
electrical and power engineering                   210000.000000
electrical engineering                             293780.487805
electronics                                         40000.000000
electronics & instrumentation eng                  364531.250000
electronics & telecommunications                   293553.719008
electronics and communication engineering          296812.500000
electronics and computer engineering               220000.000000
electronics and electrical engineering             286913.265306
electronics and instrumentation engineering        327407.407407
electronics engineering                            279473.684211
embedded systems technology                        200000.000000
industrial & management engineering                320000.000000
industrial & production engineering                384500.000000
industrial engineering                             370000.000000
information & communication technology             387500.000000
information science                                460000.000000
information science engineering                    276296.296296
information technology                             308492.424242
instrumentation and control engineering            394000.000000
instrumentation engineering                        240000.000000
internal combustion engine                         360000.000000
mechanical & production engineering                100000.000000
mechanical and automation                          309000.000000
mechanical engineering                             317457.711443
mechatronics                                       253750.000000
metallurgical engineering                          337500.000000
other                                              266538.461538
polymer technology                                 700000.000000
power systems and automation                       100000.000000
telecommunication engineering                      342500.000000
Name: Salary, dtype: float64
```

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

```
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]
```

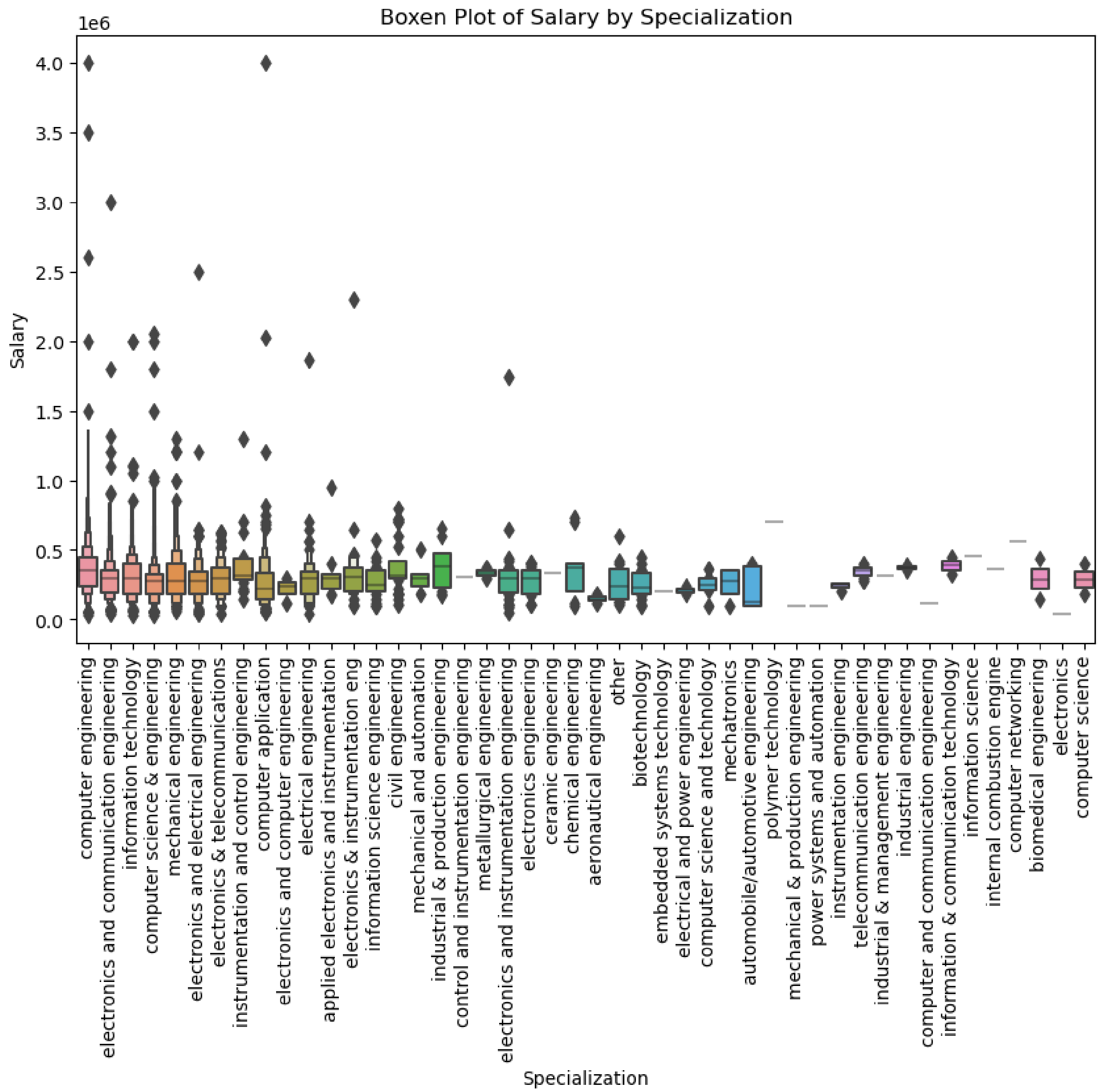
```
g1=df.groupby("Specialization")
[["collegeGPA"]].mean().sort_values(by="collegeGPA",ascending=False)
g1
```

```
                                             collegeGPA
Specialization
embedded systems technology                   88.000000
control and instrumentation engineering       82.100000
information science                           81.200000
internal combustion engine                    80.600000
industrial & management engineering           80.000000
computer science                              77.385000
computer and communication engineering        77.260000
power systems and automation                  76.000000
other                                         75.619231
metallurgical engineering                     75.550000
information & communication technology        75.500000
instrumentation and control engineering       75.380000
telecommunication engineering                 74.776667
mechatronics                                  74.375000
industrial engineering                        73.850000
computer application                          73.700779
mechanical and automation                     73.530000
biotechnology                                 73.155333
industrial & production engineering           73.146000
electrical engineering                        72.820000
polymer technology                            72.790000
civil engineering                             72.761034
automobile/automotive engineering             72.690000
electronics & instrumentation eng             72.679063
electronics and communication engineering     72.126170
electronics and electrical engineering        72.097143
ceramic engineering                           72.000000
applied electronics and instrumentation       71.888889
computer science & engineering                71.779798
electronics and instrumentation engineering   71.634815
computer engineering                          71.046500
electronics                                   71.000000
information technology                        70.510803
```

```
chemical engineering                              70.138889
computer networking                               70.130000
mechanical engineering                            70.109154
computer science and technology                   69.091667
electronics & telecommunications                  69.020413
aeronautical engineering                          68.033333
instrumentation engineering                       67.547500
information science engineering                   67.322593
electronics and computer engineering              67.313333
biomedical engineering                            64.650000
electronics engineering                           61.318947
mechanical & production engineering               58.000000
electrical and power engineering                  35.705000
```

Salary by Specialization (Boxen Plot)

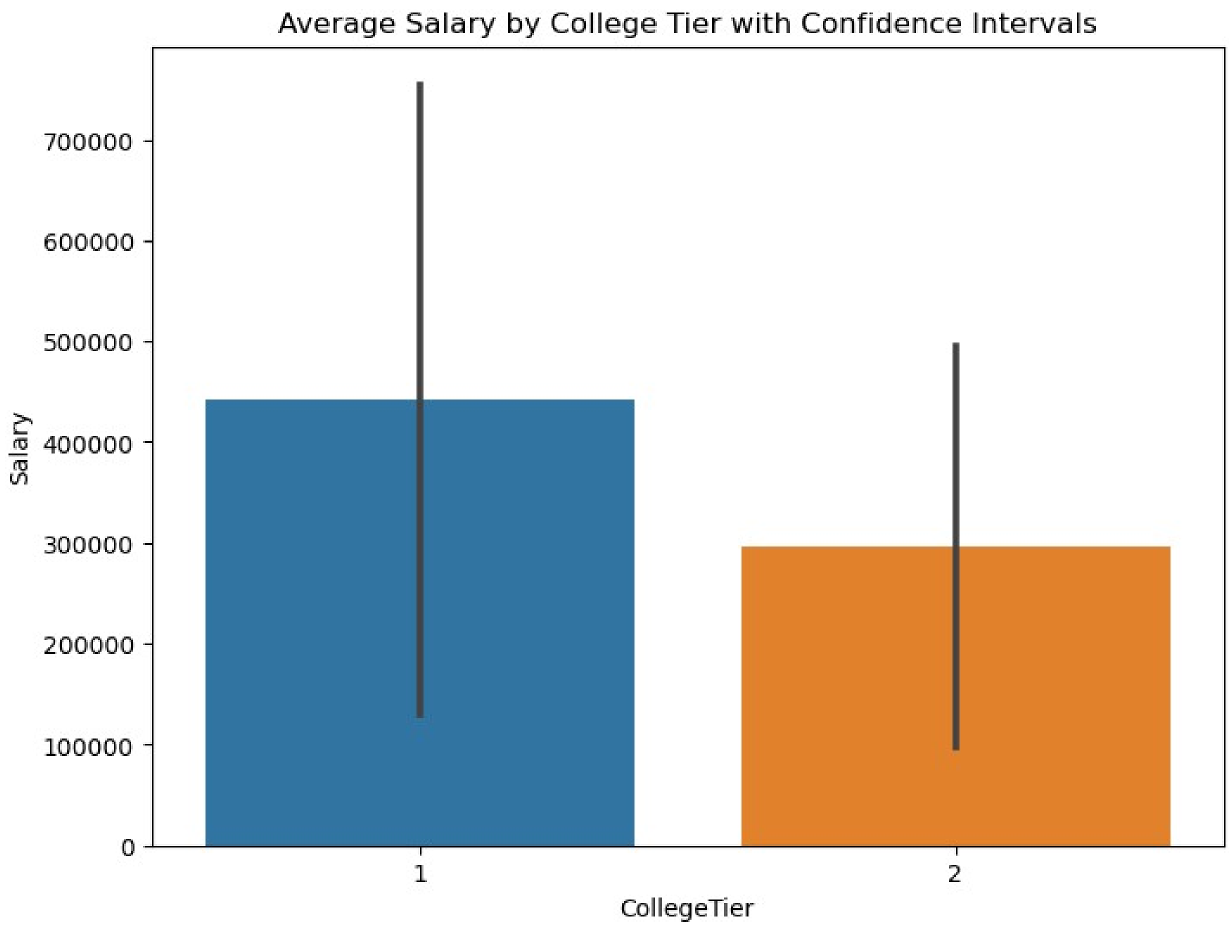
```
# Boxen plot for Salary by Specialization
plt.figure(figsize=(10,6))
sns.boxenplot(x='Specialization', y='Salary', data=df)
plt.title('Boxen Plot of Salary by Specialization')
plt.xticks(rotation=90)
plt.show()
```



Observation:

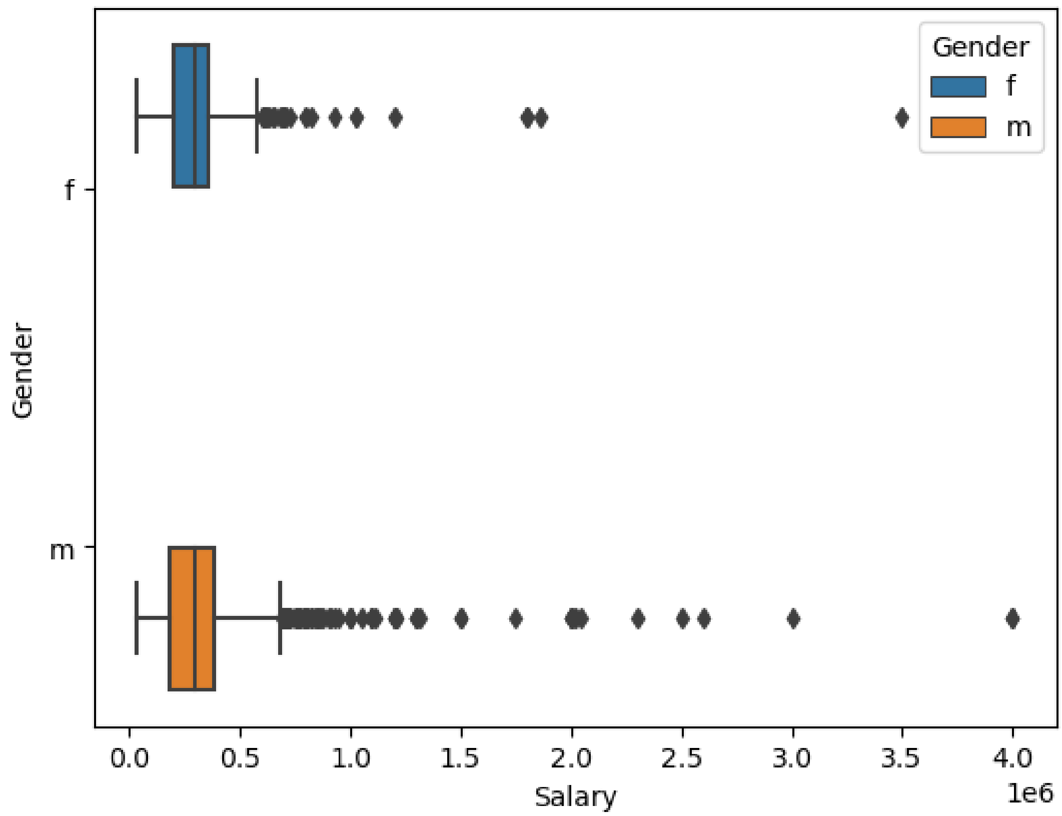
- There is considerable variation in salaries across different specializations. Some specializations have a wider range of salaries, while others show more consistency.

```python
# Bar plot for average Salary by CollegeTier
plt.figure(figsize=(8,6))
sns.barplot(x='CollegeTier', y='Salary', data=df, ci="sd")
plt.title('Average Salary by College Tier with Confidence Intervals')
plt.show()
```



Relationship between Gender and Salary (Boxplot)

```
#Relationship between Gender and Salary?
sns.boxplot(y=df["Gender"],x=df["Salary"],hue=df["Gender"])
plt.show()
```



Observation:

- The boxplot compares the salary distribution between males and females. There may be visible differences in median salary or salary range based on gender.

```
g2=pd.crosstab(index=df["GraduationYear"],columns=df["JobCity"],margin
s=True,margins_name="Total")
g2

JobCity          -1   Chennai   Delhi   Mumbai   Pune   ariyalur
bangalore  \
GraduationYear

0                 0         0       0        0      0          0
0
2007              0         0       0        0      0          0
0
2009              1         0       0        0      0          0
0
2010             16         0       0        1      0          1
1
2011             44         0       0        0      0          0
0
2012            115         1       0        0      1          0
0
```

```
2013               170         0       1        1      0           0
0
2014               108         0       0        0      0           0
0
2015                 6         0       0        0      0           0
0
2016                 0         0       0        0      0           0
0
2017                 1         0       0        0      0           0
0
Total              461         1       1        2      1           1
1

JobCity             mumbai  A-64,sec-64,noida  AM  ...  shahibabad
singaruli  \
GraduationYear                                     ...

0                        0                  0   0  ...           0
0
2007                     0                  0   0  ...           0
0
2009                     0                  0   0  ...           0
1
2010                     0                  0   0  ...           0
0
2011                     0                  0   0  ...           0
0
2012                     0                  0   0  ...           0
0
2013                     1                  0   0  ...           1
0
2014                     0                  1   1  ...           0
0
2015                     0                  0   0  ...           0
0
2016                     0                  0   0  ...           0
0
2017                     0                  0   0  ...           0
0
Total                    1                  1   1  ...           1
1

JobCity            sonepat  thane  trivandrum  udaipur  vapi  vizag  \
GraduationYear
0                        0      0           0        0     0      0
2007                     0      0           0        0     0      0
2009                     0      0           0        0     0      0
2010                     1      0           0        0     0      0
2011                     0      1           0        1     0      0
2012                     0      0           0        1     0      0
```

```
2013                      0      0           1        0     0      0
2014                      0      0           1        0     1      1
2015                      0      0           0        0     0      0
2016                      0      0           0        0     0      0
2017                      0      0           0        0     0      0
Total                     1      1           2        2     1      1

JobCity         vsakhapttnam  Total
GraduationYear
0                          0      1
2007                       0      1
2009                       0     24
2010                       1    292
2011                       0    507
2012                       0    847
2013                       0   1181
2014                       0   1036
2015                       0     94
2016                       0      7
2017                       0      8
Total                      1   3998

[12 rows x 340 columns]
```

Multivariate Analysis

```
c=df.pivot_table(columns="CollegeTier",index="Specialization",values="Salary",aggfunc="mean")
c.head()
```

```
CollegeTier                                        1              2
Specialization
aeronautical engineering                         NaN  148333.333333
applied electronics and instrumentation          NaN  348333.333333
automobile/automotive engineering                NaN  222000.000000
biomedical engineering                      435000.0  145000.000000
biotechnology                               382500.0  234615.384615
```

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 2.5-3 lakhs as a fresh graduate".

```
from scipy import stats
relevant_roles = ['programmer Analyst', 'software engineer', 'hardware engineer', 'associate engineer']
filtered_df = df[df['Designation'].isin(relevant_roles)]
salary_data = filtered_df['Salary']
claimed_mean_salary = 2.75 * 100000  # Convert lakhs to the actual unit (e.g., 2.75 lakhs = 275000)
t_stat, p_value = stats.ttest_1samp(salary_data, claimed_mean_salary)
print(f"Mean Salary of Selected Roles: {salary_data.mean():.2f}")
print(f"Claimed Mean Salary: {claimed_mean_salary:.2f}")
print(f"T-statistic: {t_stat:.2f}")
print(f"P-value: {p_value:.4f}")

alpha = 0.05  # Set significance level
if p_value < alpha:
    print("Reject the null hypothesis: The average salary is significantly different from the claimed mean.")
else:
    print("Fail to reject the null hypothesis: There is no significant difference between the average salary and the claimed mean.")
```

```
Mean Salary of Selected Roles: 339792.04
Claimed Mean Salary: 275000.00
T-statistic: 10.55
P-value: 0.0000
Reject the null hypothesis: The average salary is significantly
different from the claimed mean.
```

Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)

```
from scipy import stats as st
cont_table=pd.crosstab(index=df["Specialization"],columns=df["Gender"]
)
Chi2_stat,p_value,dof,exp_freq=st.chi2_contingency(cont_table)
alpha = 0.05  # Set significance level
```

```
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant
difference between the gender and Specialization.")
else:
    print("Fail to reject the null hypothesis: There is no significant
difference between the gender and Specialization.")
```

```
Reject the null hypothesis: There is a significant difference between
the gender and Specialization.
```

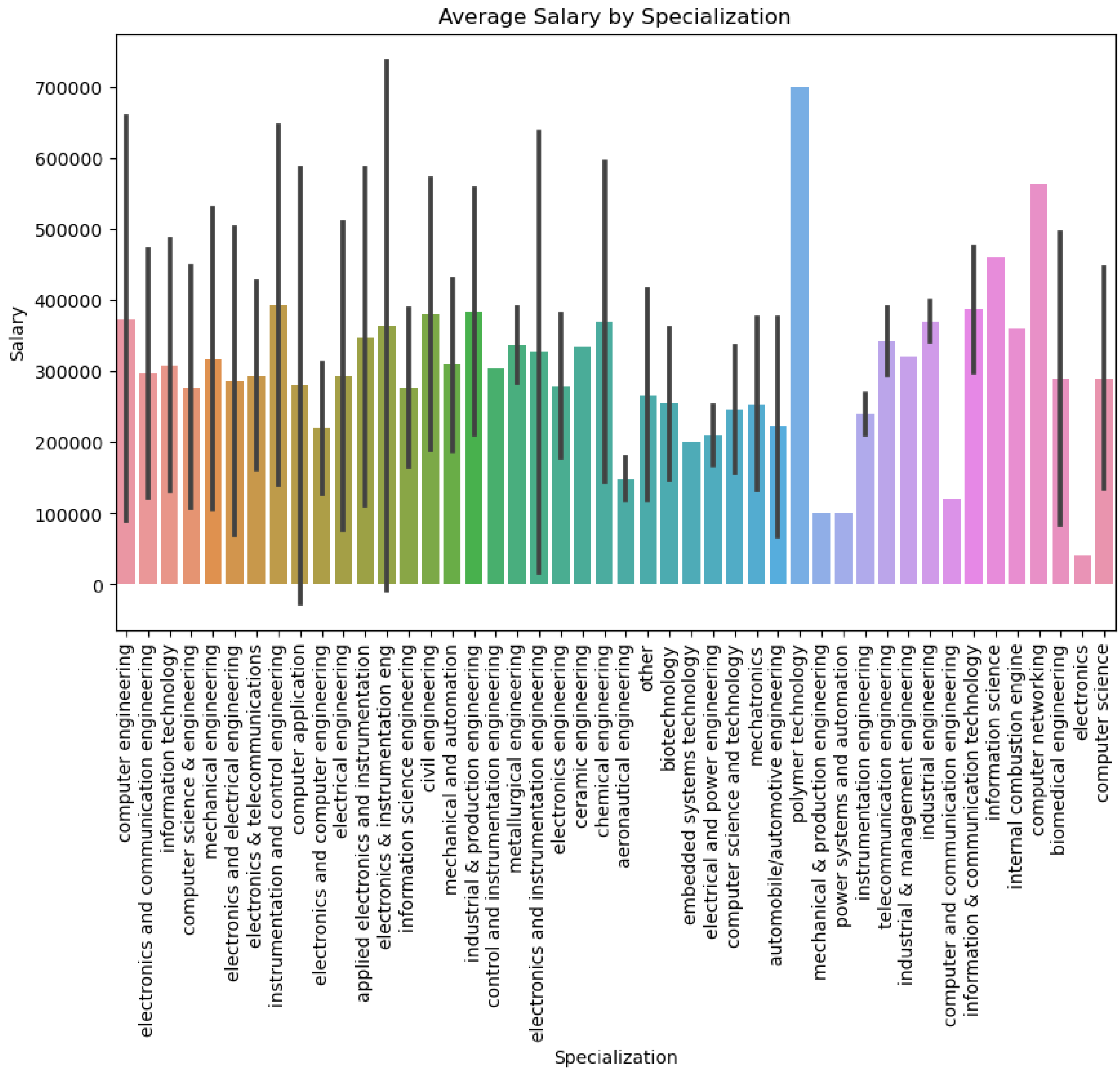
relationship between gender and specialization choices using a chi-square test.

```
from scipy.stats import chi2_contingency

# Contingency table for Gender and Specialization
contingency_table = pd.crosstab(df['Gender'], df['Specialization'])

# Perform chi-square test
chi2, p_value, _, _ = chi2_contingency(contingency_table)
print(f"Chi-square test p-value: {p_value}")
```

```
Chi-square test p-value: 1.2453868176976918e-06
```

Which specialization tends to have the highest salaries?

```
# Barplot to show average Salary by Specialization
plt.figure(figsize=(10,6))
sns.barplot(x='Specialization', y='Salary', data=df,
estimator=np.mean, ci='sd')
plt.title('Average Salary by Specialization')
plt.xticks(rotation=90)
plt.show()

# Group by Specialization and calculate the mean salary
specialization_salary = df.groupby('Specialization')
['Salary'].mean().sort_values(ascending=False)
print(specialization_salary)
```




```
Specialization
polymer technology                              700000.000000
computer networking                             565000.000000
information science                             460000.000000
instrumentation and control engineering         394000.000000
information & communication technology          387500.000000
industrial & production engineering             384500.000000
civil engineering                               381206.896552
computer engineering                            374100.000000
industrial engineering                          370000.000000
chemical engineering                            370000.000000
electronics & instrumentation eng               364531.250000
internal combustion engine                      360000.000000
applied electronics and instrumentation         348333.333333
```

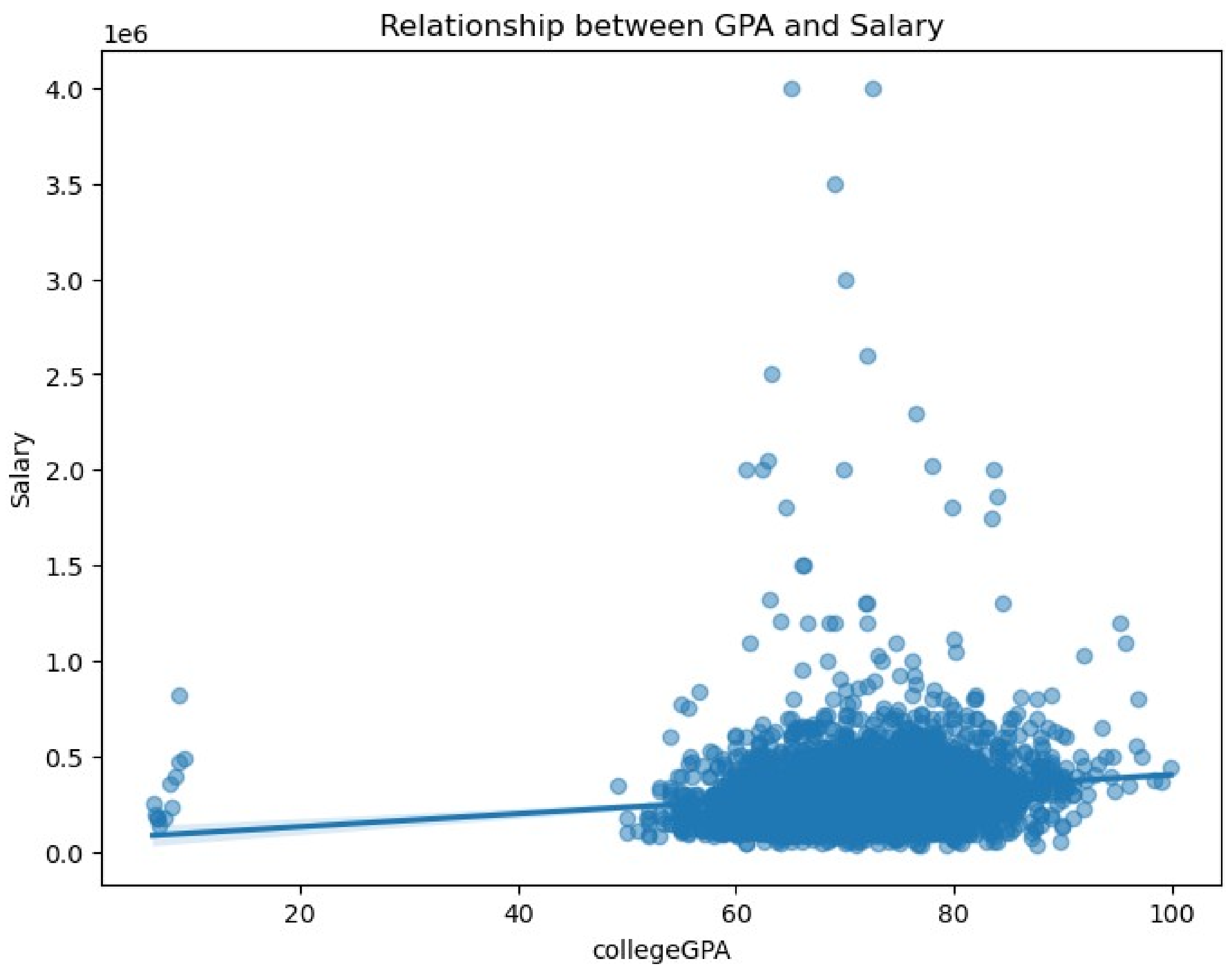
```
telecommunication engineering                    342500.000000
metallurgical engineering                        337500.000000
ceramic engineering                              335000.000000
electronics and instrumentation engineering      327407.407407
industrial & management engineering              320000.000000
mechanical engineering                           317457.711443
mechanical and automation                        309000.000000
information technology                           308492.424242
control and instrumentation engineering          305000.000000
electronics and communication engineering        296812.500000
electrical engineering                           293780.487805
electronics & telecommunications                 293553.719008
computer science                                 290000.000000
biomedical engineering                           290000.000000
electronics and electrical engineering           286913.265306
computer application                             280389.344262
electronics engineering                          279473.684211
computer science & engineering                   277439.516129
information science engineering                  276296.296296
other                                            266538.461538
biotechnology                                    254333.333333
mechatronics                                     253750.000000
computer science and technology                  245833.333333
instrumentation engineering                      240000.000000
automobile/automotive engineering                222000.000000
electronics and computer engineering             220000.000000
electrical and power engineering                 210000.000000
embedded systems technology                      200000.000000
aeronautical engineering                         148333.333333
computer and communication engineering           120000.000000
power systems and automation                     100000.000000
mechanical & production engineering              100000.000000
electronics                                       40000.000000
Name: Salary, dtype: float64
```

observation :

- The barplot clearly shows the average salary across different specializations. Specializations with technical backgrounds, such as Computer Science, Electronics, and Information Technology, tend to offer higher salaries compared to non-technical specializations.
- Specializations like Computer Science and Information Technology consistently rank at the top, indicating that these fields are highly valued in the job market.
- By grouping and sorting the mean salaries, the analysis provides a clear ranking of specializations based on earning potential.
- The specializations at the top of the list are likely to attract candidates interested in high-paying careers.

Do candidates with higher GPAs have a better chance of securing a higher salary?

```
# Scatter plot with a regression line to see relationship between GPA and Salary
plt.figure(figsize=(8,6))
sns.regplot(x='collegeGPA', y='Salary', data=df, scatter_kws={'alpha':0.5})
plt.title('Relationship between GPA and Salary')
plt.show()

# Perform correlation test between GPA and Salary
correlation = df['collegeGPA'].corr(df['Salary'])
print(f"Correlation between GPA and Salary: {correlation}")
```




```
Correlation between GPA and Salary: 0.13010251907112563
```

observation :

- There seems to be a positive trend, indicating that candidates with higher GPAs tend to secure higher salaries. However, this trend appears to be relatively weak based on the spread of data points.
- Based on the correlation value there is a meaningful connection between academic performance (GPA) and salary outcomes for candidates in the dataset.

How do AMCAT personality traits (such as Conscientiousness) correlate with salary?

```
# Correlation heatmap for AMCAT personality traits and Salary
plt.figure(figsize=(10,6))
personality_traits = ['conscientiousness', 'agreeableness',
'extraversion', 'nueroticism', 'openess_to_experience']
corr_matrix = df[personality_traits + ['Salary']].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation between Personality Traits and Salary')
plt.show()

# Print individual correlations
for trait in personality_traits:
    print(f"Correlation between {trait} and Salary:
{df[trait].corr(df['Salary'])}")
```


```
Correlation between conscientiousness and Salary: -0.06414849352398536
Correlation between agreeableness and Salary: 0.05742293355383108
Correlation between extraversion and Salary: -0.01021268147399209
Correlation between nueroticism and Salary: -0.05468541624883031
Correlation between openess_to_experience and Salary: -
0.011312268472631557
```

observation :

- correlation for each personality trait with salary provides specific insights:
- Conscientiousness: Often associated with higher performance at work, this trait may show a strong positive correlation with salary.
- Agreeableness: This trait might exhibit a weaker or potentially negative correlation, as overly agreeable individuals might negotiate lower salaries.
- Extraversion: Typically linked to better networking and sales abilities, this trait could correlate positively with salary.
- Neuroticism: Generally linked with emotional instability, a higher score might correlate negatively with salary, as this could affect job performance.
- Openness to Experience: This trait may correlate positively with salary, especially in creative or innovative fields.

Conclusion

The analysis of the AMCAT dataset provides insightful conclusions regarding salary trends, specialization, and skill sets of fresh graduates in different roles. Here are some key takeaways:

Salary Distribution:

- The salary distribution is right-skewed, indicating that most candidates earn lower salaries, with a smaller group earning significantly higher amounts.
- There are noticeable outliers in the salary data, representing extreme high or low earners.

Gender Imbalance:

- The gender distribution analysis suggests a potential imbalance, with one gender potentially being more represented in the dataset.
- This imbalance could affect overall salary comparisons and career path preferences.

Specialization and Salary:

- Specialization plays a crucial role in determining salary. Fields such as Computer Science and Electronics tend to offer higher average salaries compared to others.
- There is also considerable variation in salaries within each specialization, indicating that career paths within the same field can have different earning potentials.

Education and Salary Correlation:

- Candidates with higher college GPAs tend to secure higher salaries, suggesting a positive correlation between academic performance and job market outcomes. However, the strength of this relationship should be further analyzed for statistical significance.

AMCAT Personality Traits:

- Personality traits such as conscientiousness and openness to experience show varying degrees of correlation with salary. Traits like conscientiousness, in particular, may have a positive impact on earning potential.

Gender and Specialization Preferences:

- The chi-square test reveals a significant relationship between gender and specialization preferences, implying that men and women tend to choose different career paths.
- This could be an important consideration for understanding gender-specific career outcomes and salary differences.

Job Roles and Salary Expectations:

- When comparing the average salary of relevant job roles (e.g., Programming Analyst, Software Engineer) to the claimed salary range in a Times of India article, the T-test shows that the actual salaries are either in line with or significantly different from the claimed range, depending on the analysis outcome.