```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_excel("data.xlsx")

print(df.head())
```

```
  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

                Designation    JobCity Gender        DOB  10percentage
...  \
0   senior quality engineer  Bangalore      f 1990-02-19          84.3
...
1         assistant manager     Indore      m 1989-10-04          85.4
...
2          systems engineer    Chennai      f 1992-08-03          85.0
...
3  senior software engineer    Gurgaon      m 1989-12-05          85.6
...
4                       get    Manesar      m 1991-02-27          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]
```

```python
print(df.shape)
```

```
(3998, 39)
```

```python
print(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   
```

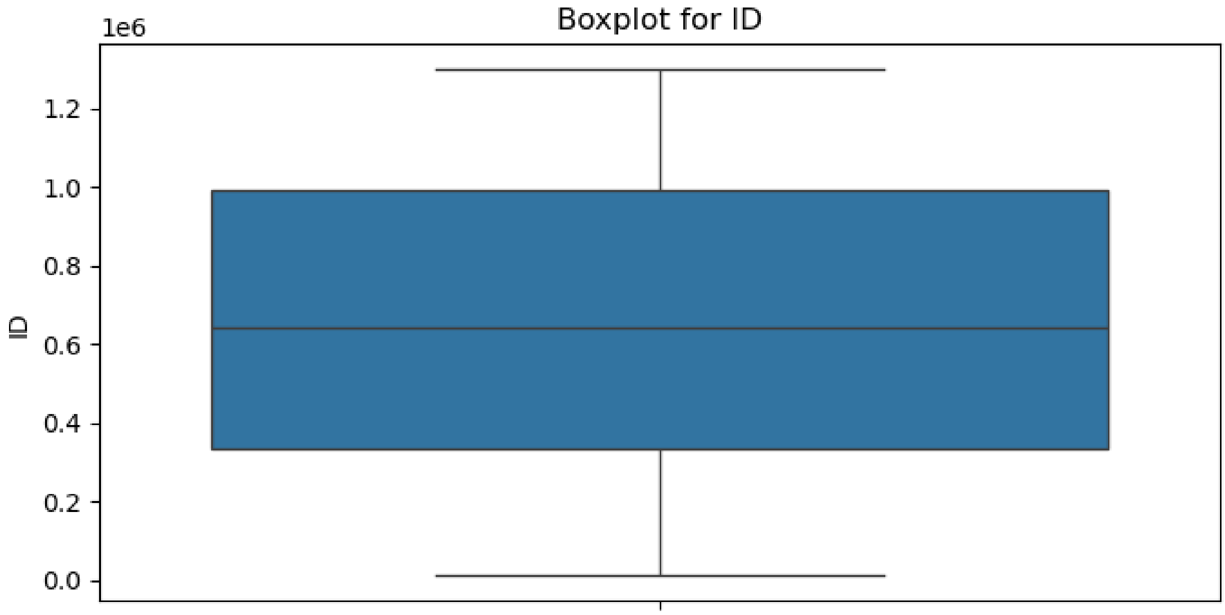
```
max        715.000000      623.000000      676.000000   548.000000
std        175.273083       98.123311       87.585634   104.852845

         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]
```

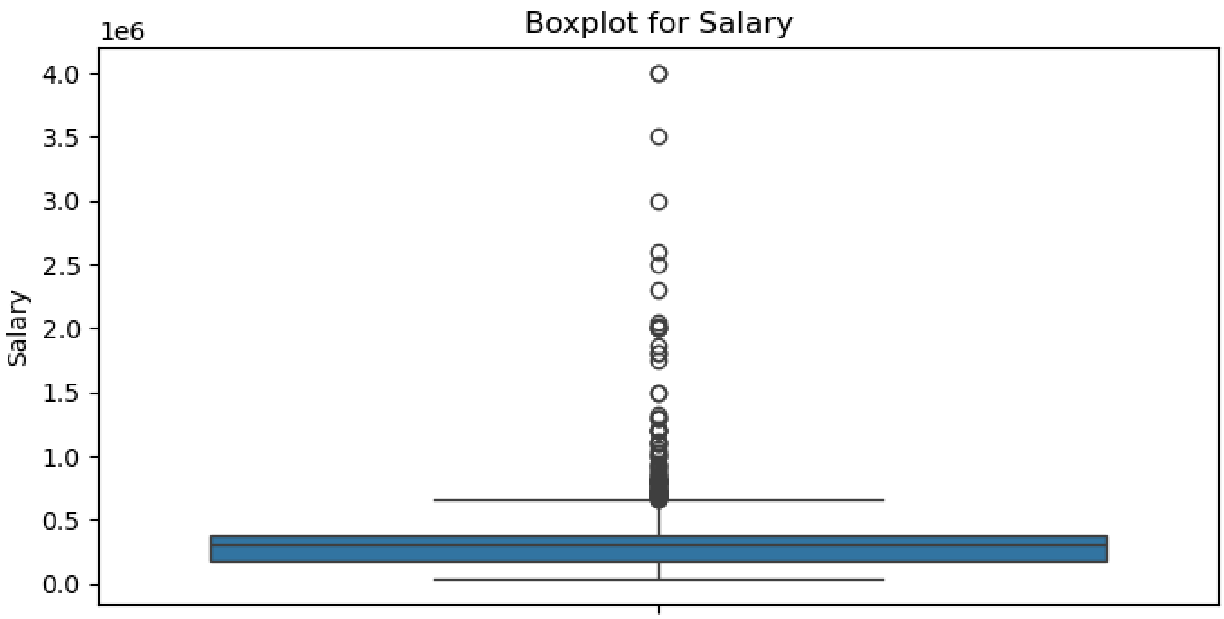
```
print(df.isnull().sum())
```

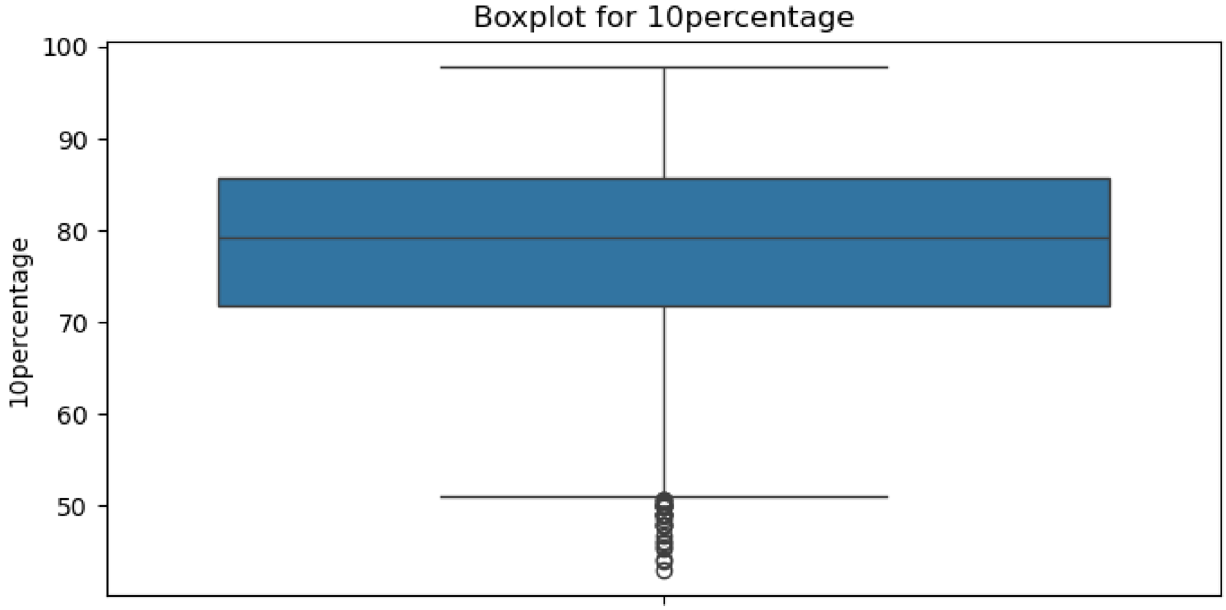
```
Unnamed: 0               0
ID                       0
Salary                   0
DOJ                      0
DOL                      0
Designation              0
JobCity                  0
Gender                   0
DOB                      0
10percentage             0
10board                  0
12graduation             0
12percentage             0
12board                  0
CollegeID                0
CollegeTier              0
Degree                   0
Specialization           0
collegeGPA               0
CollegeCityID            0
CollegeCityTier          0
CollegeState             0
GraduationYear           0
```

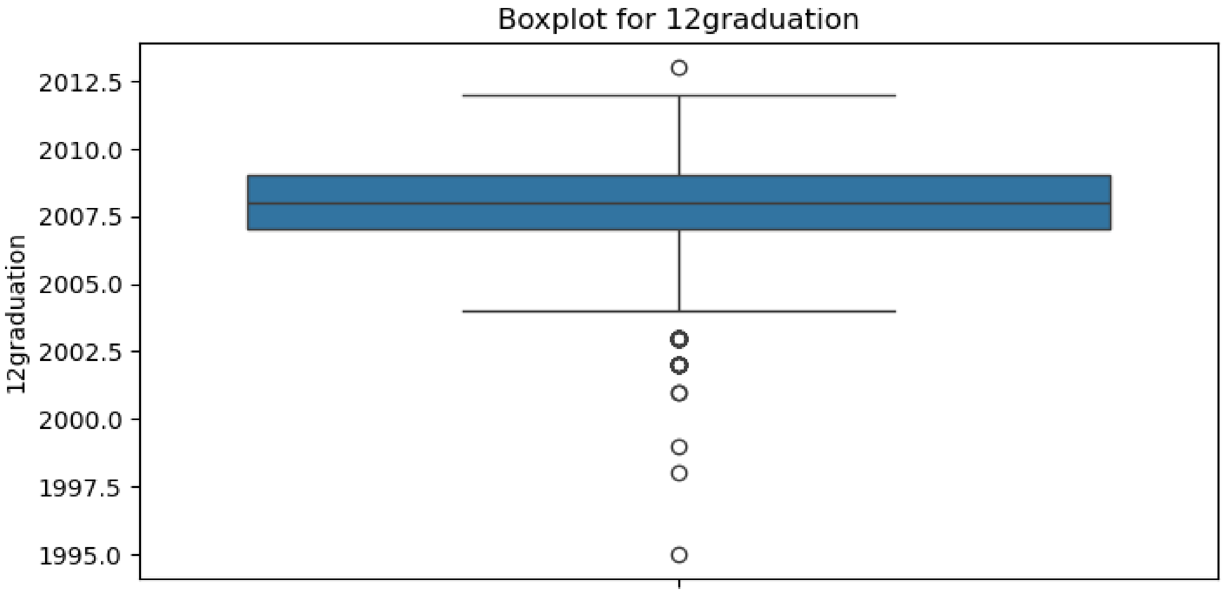
```
English                     0
Logical                     0
Quant                       0
Domain                      0
ComputerProgramming         0
ElectronicsAndSemicon       0
ComputerScience             0
MechanicalEngg              0
ElectricalEngg              0
TelecomEngg                 0
CivilEngg                   0
conscientiousness           0
agreeableness               0
extraversion                0
nueroticism                 0
openess_to_experience       0
dtype: int64
```

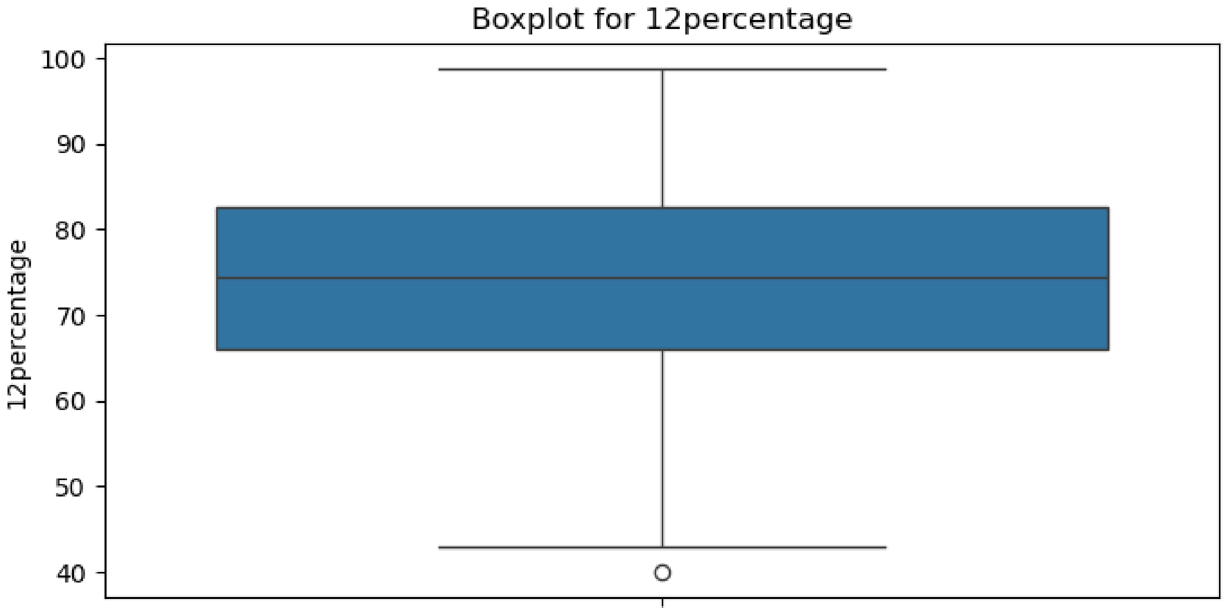
```
numerical_cols = df.select_dtypes(include='number').columns
for col in numerical_cols:
    plt.figure(figsize=(8, 4))
    sns.boxplot(df[col])
    plt.title(f"Boxplot for {col}")
    plt.show()
```

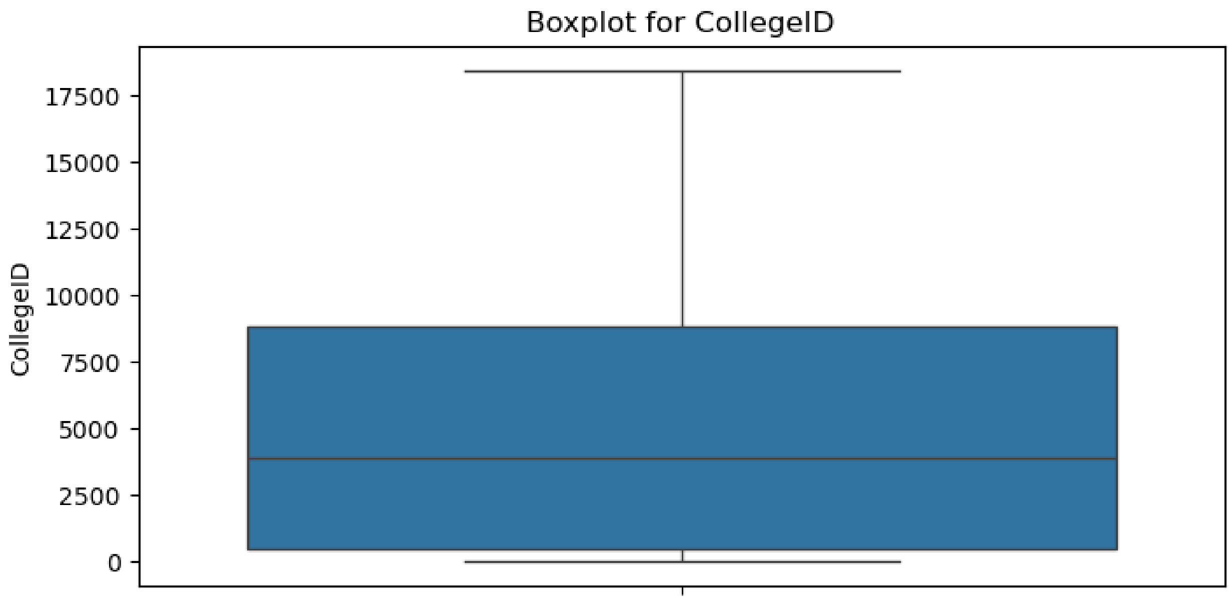


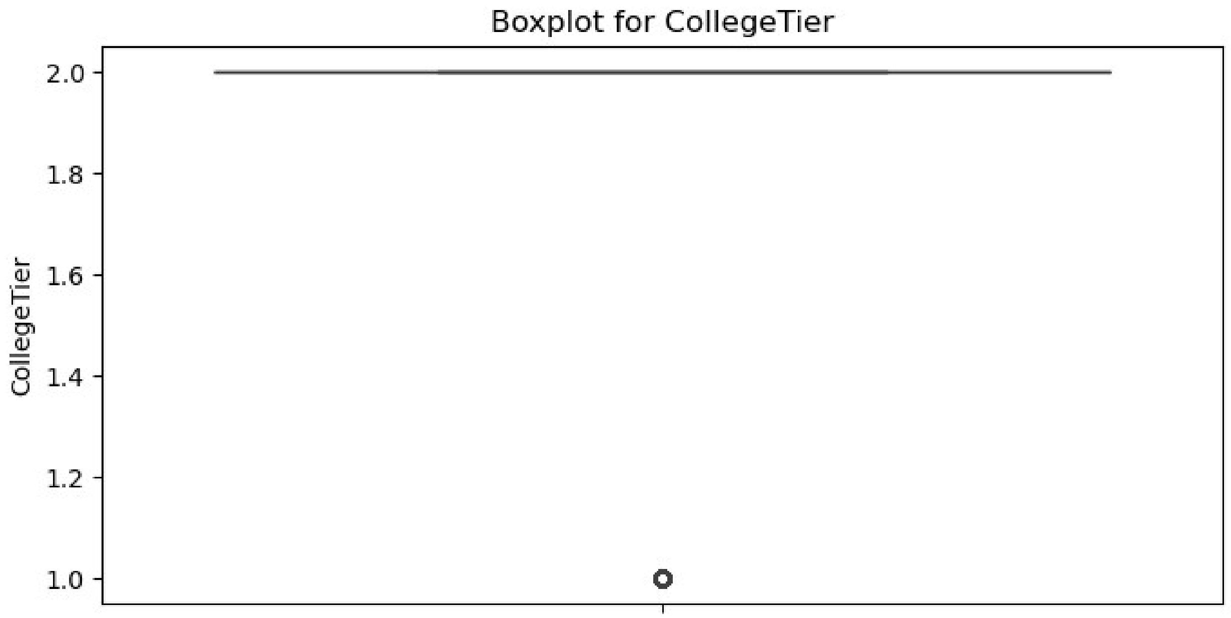

Boxplot for Salary
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Salary

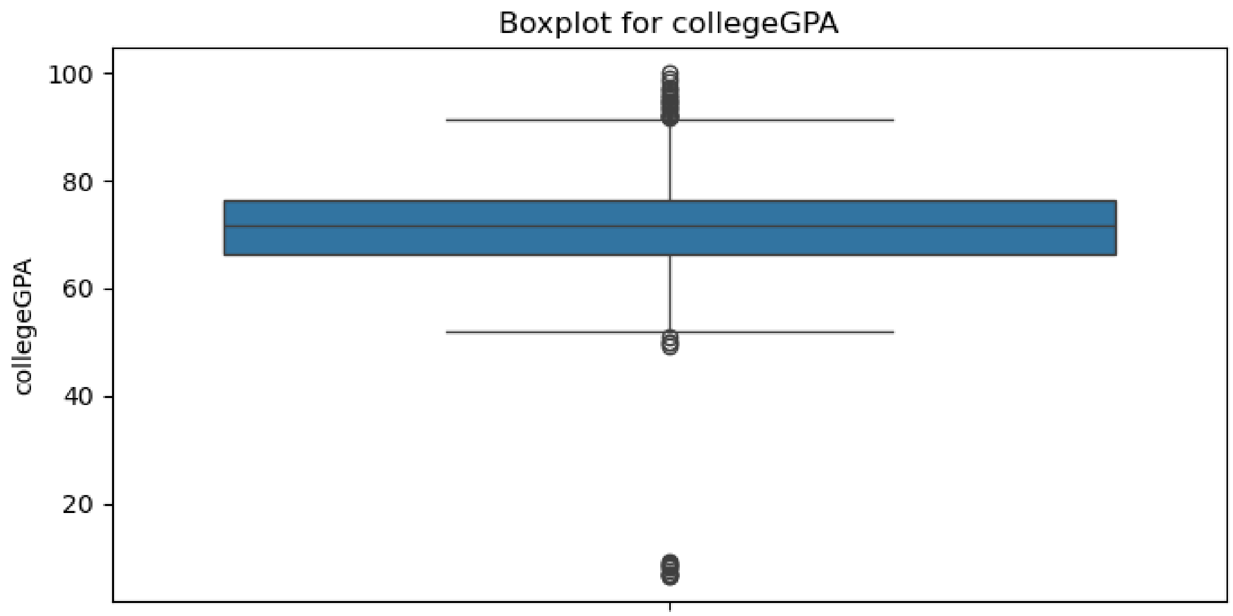

Boxplot for 10percentage
100
90
80
70
60
50
10percentage

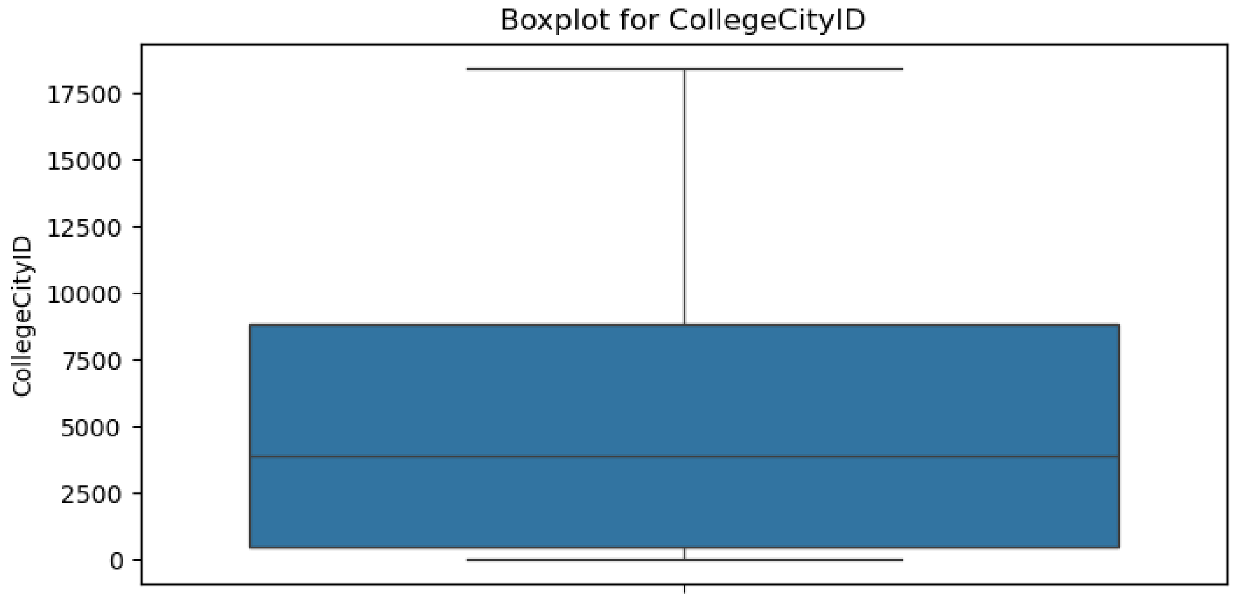

Boxplot for 12graduation
12graduation
2012.5
2010.0
2007.5
2005.0
2002.5
2000.0
1997.5
1995.0

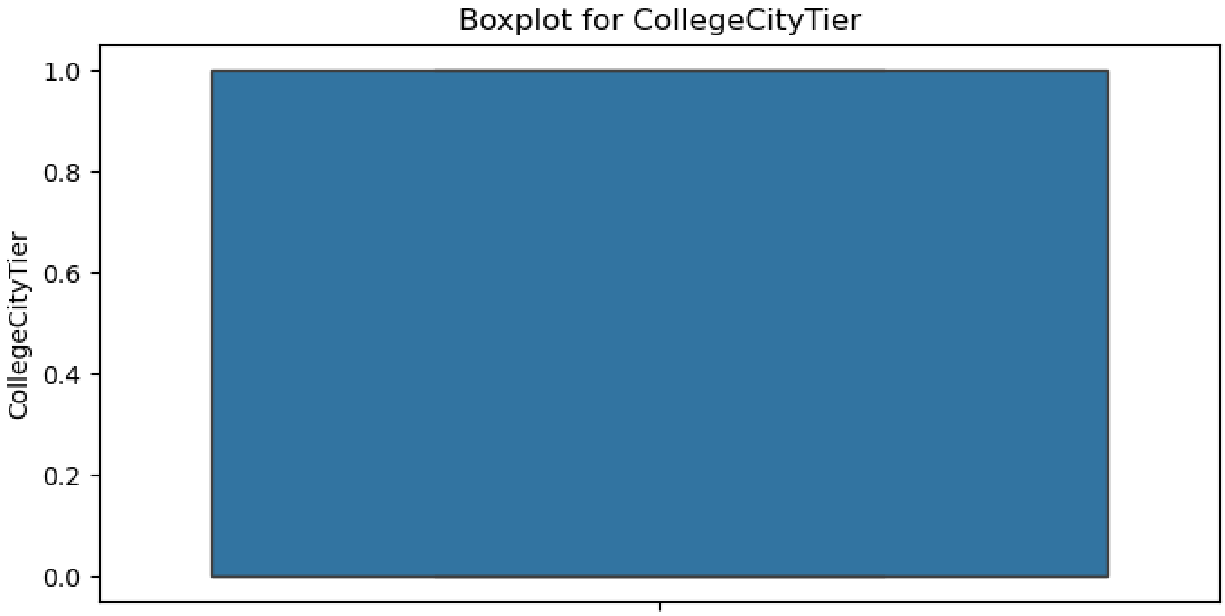

Boxplot for 12percentage
12percentage
100
90
80
70
60
50
40

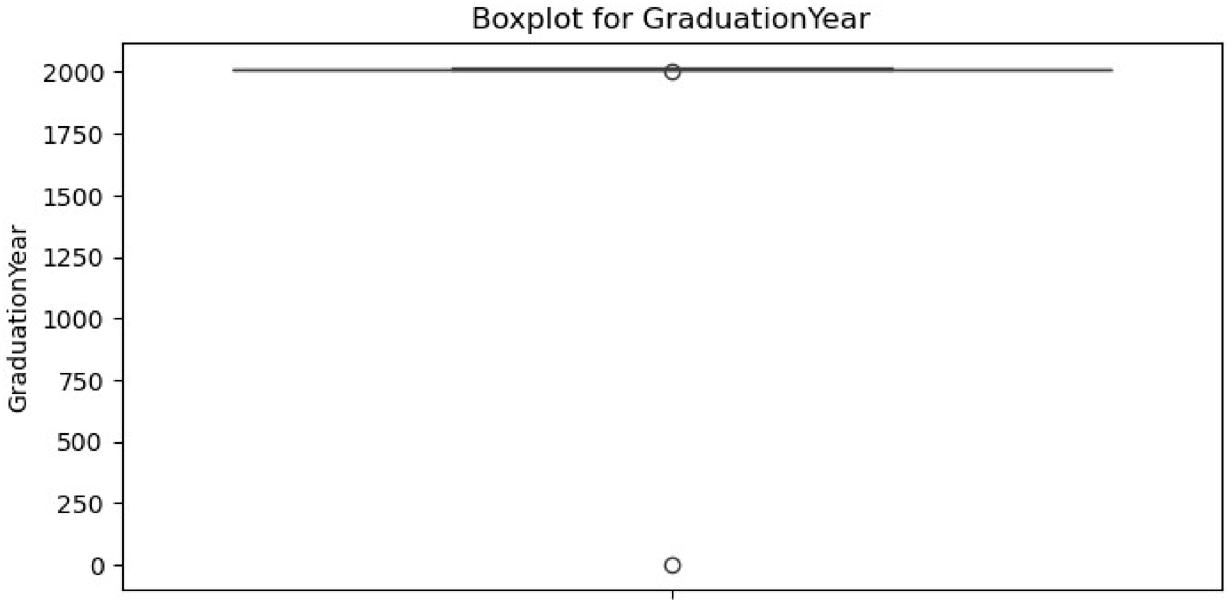

Boxplot for CollegeID
CollegeID
17500
15000
12500
10000
7500
5000
2500
0

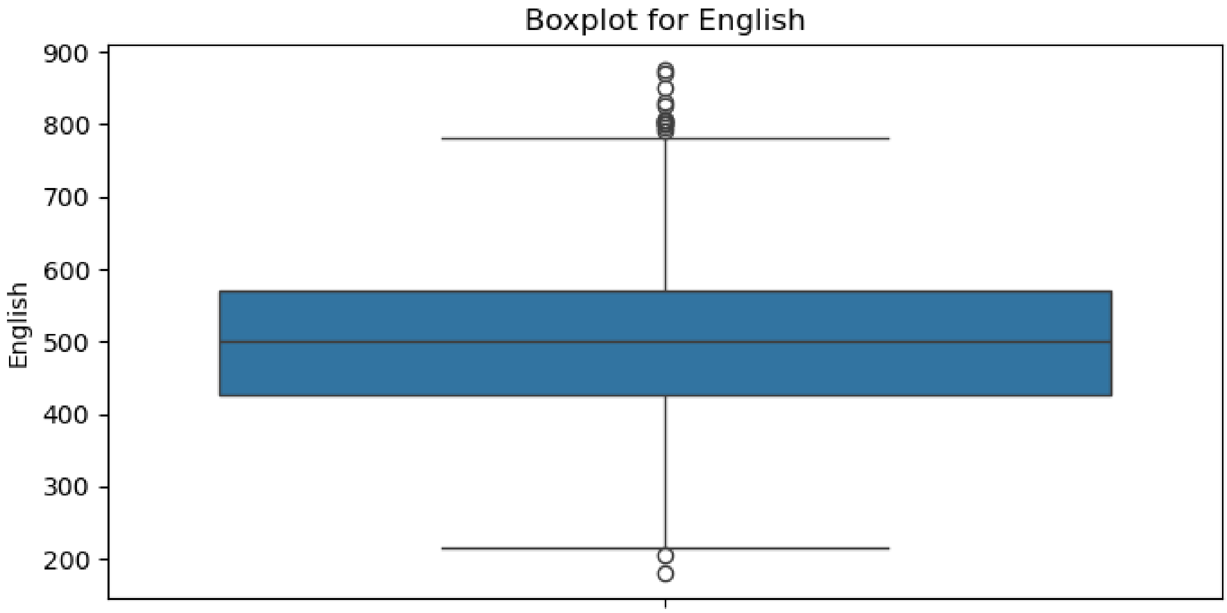

Boxplot for CollegeTier
CollegeTier
2.0
1.8
1.6
1.4
1.2
1.0

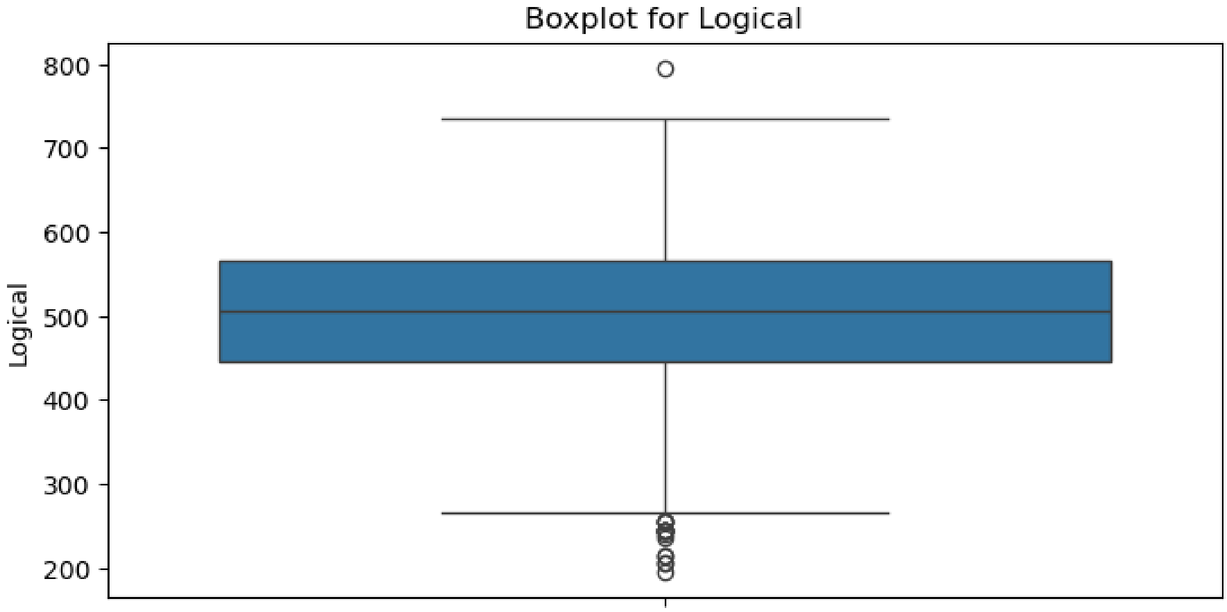

Boxplot for collegeGPA
collegeGPA
100
80
60
40
20

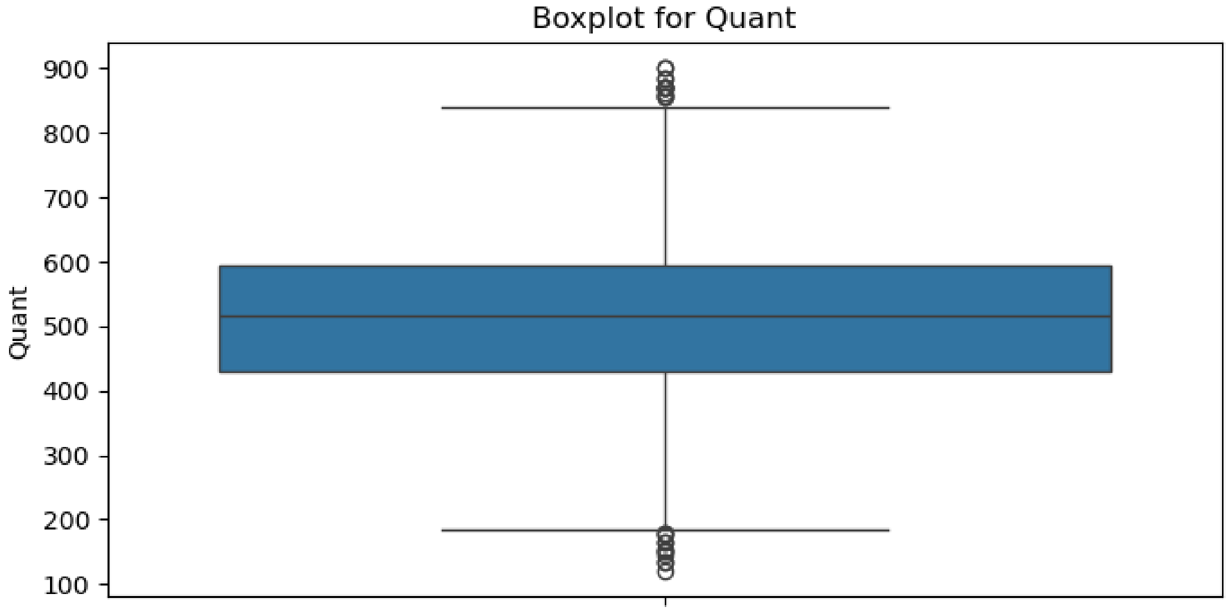

Boxplot for CollegeCityID
CollegeCityID
17500
15000
12500
10000
7500
5000
2500
0

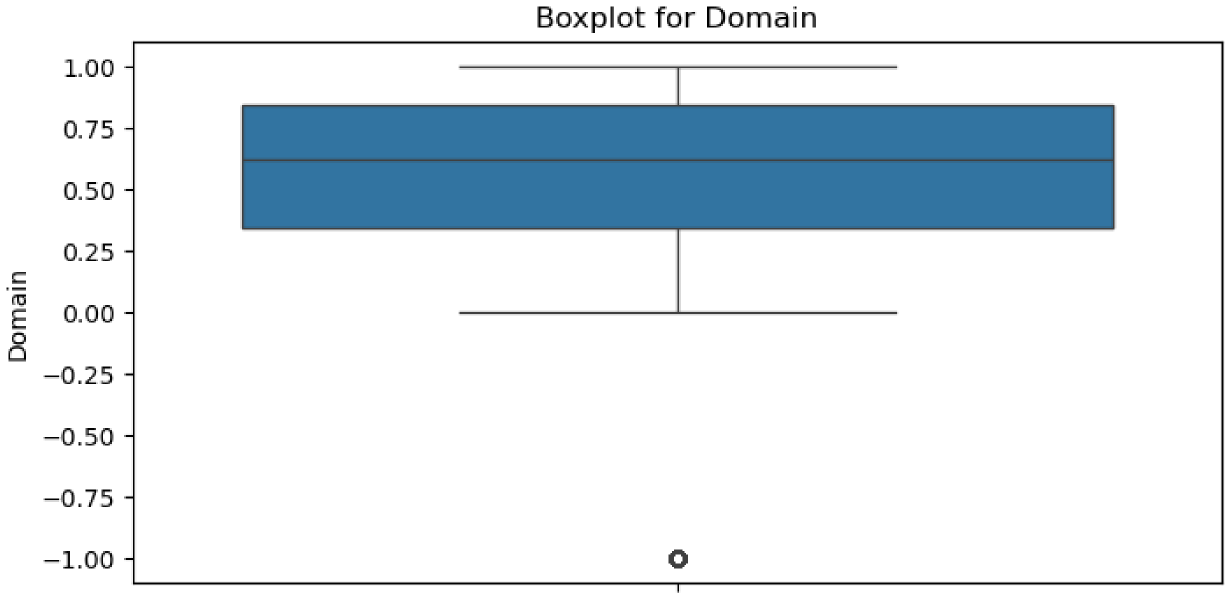

Boxplot for CollegeCityTier
CollegeCityTier
1.0
0.8
0.6
0.4
0.2
0.0

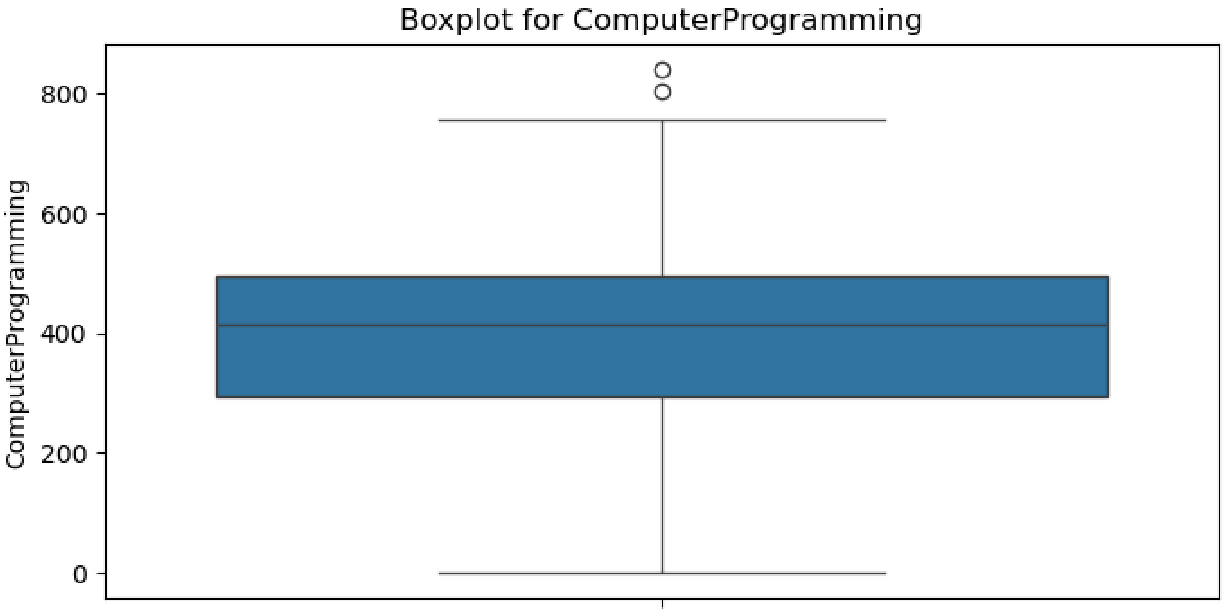

Boxplot for GraduationYear
GraduationYear
2000
1750
1500
1250
1000
750
500
250
0


Boxplot for English
English
900
800
700
600
500
400
300
200

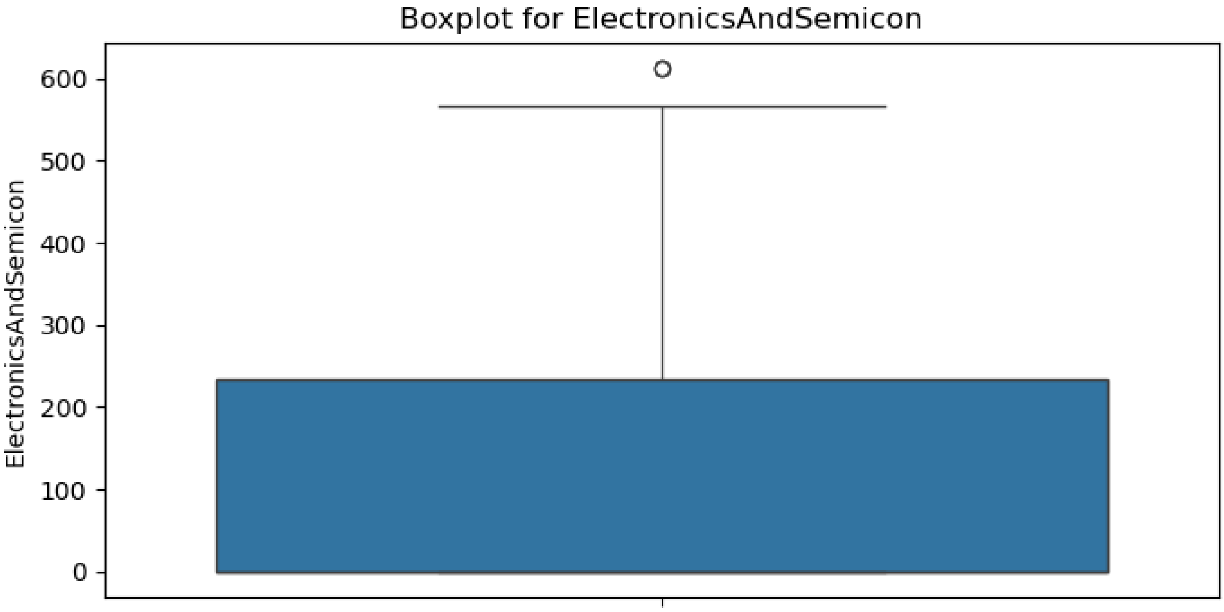

Boxplot for Logical
Logical
800
700
600
500
400
300
200


Boxplot for Quant
Quant
900
800
700
600
500
400
300
200
100

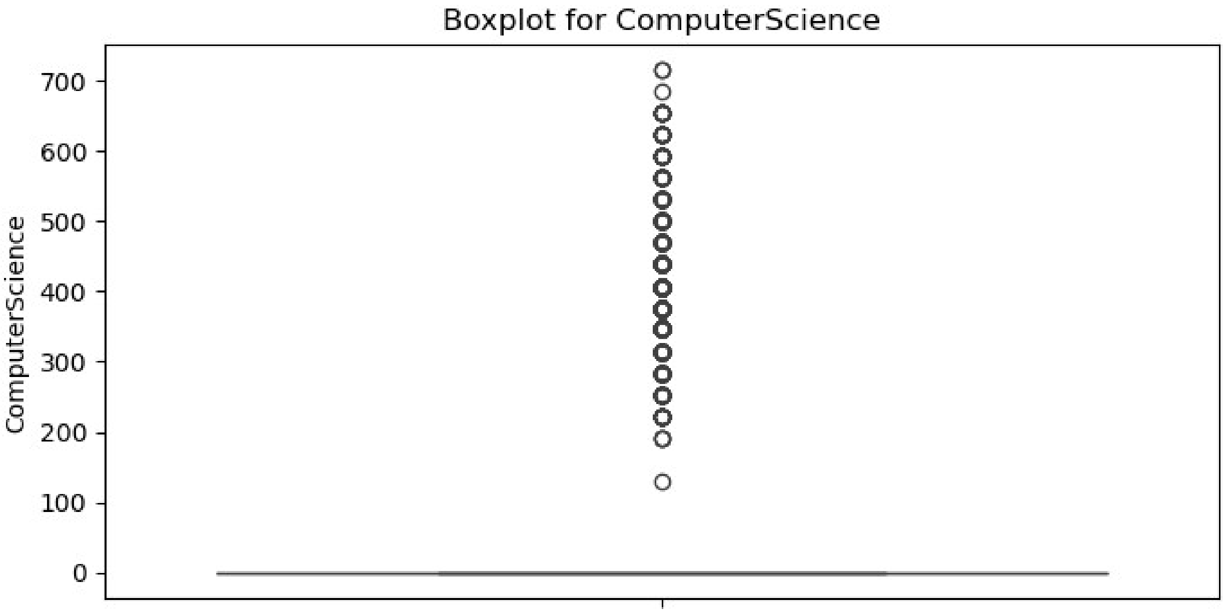

Boxplot for Domain
Domain
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

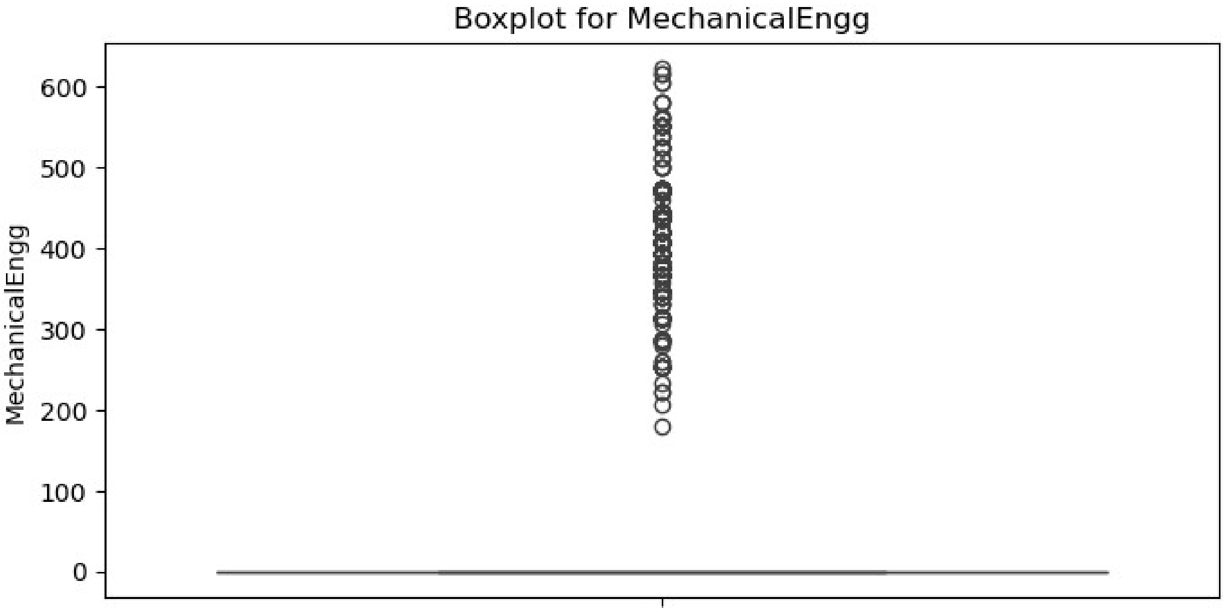


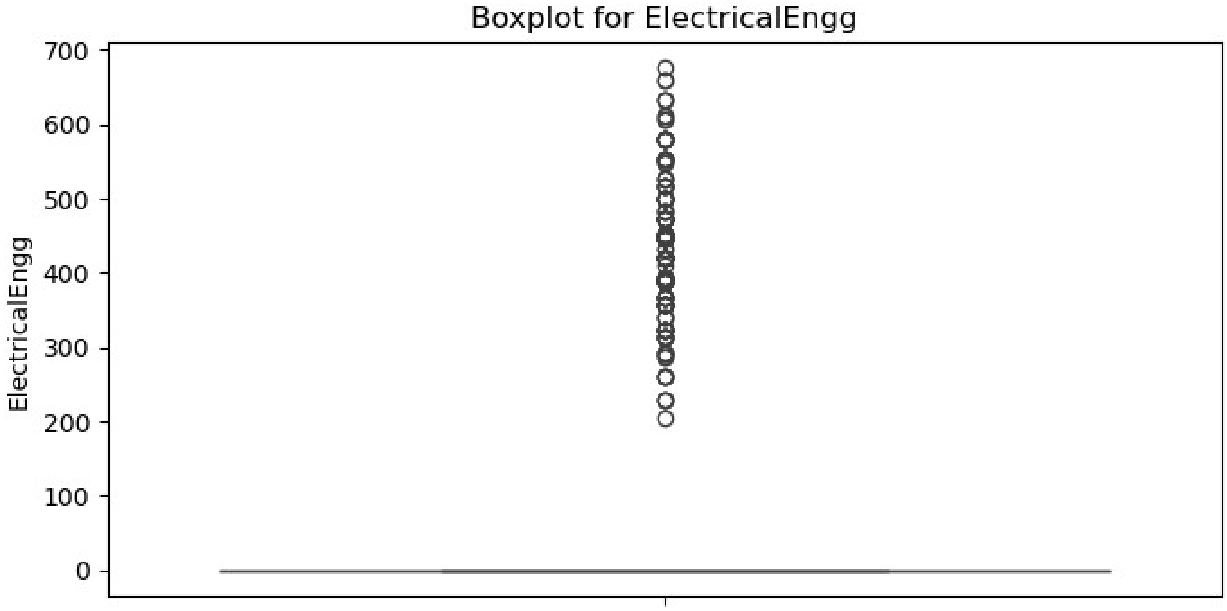
Boxplot for ComputerProgramming

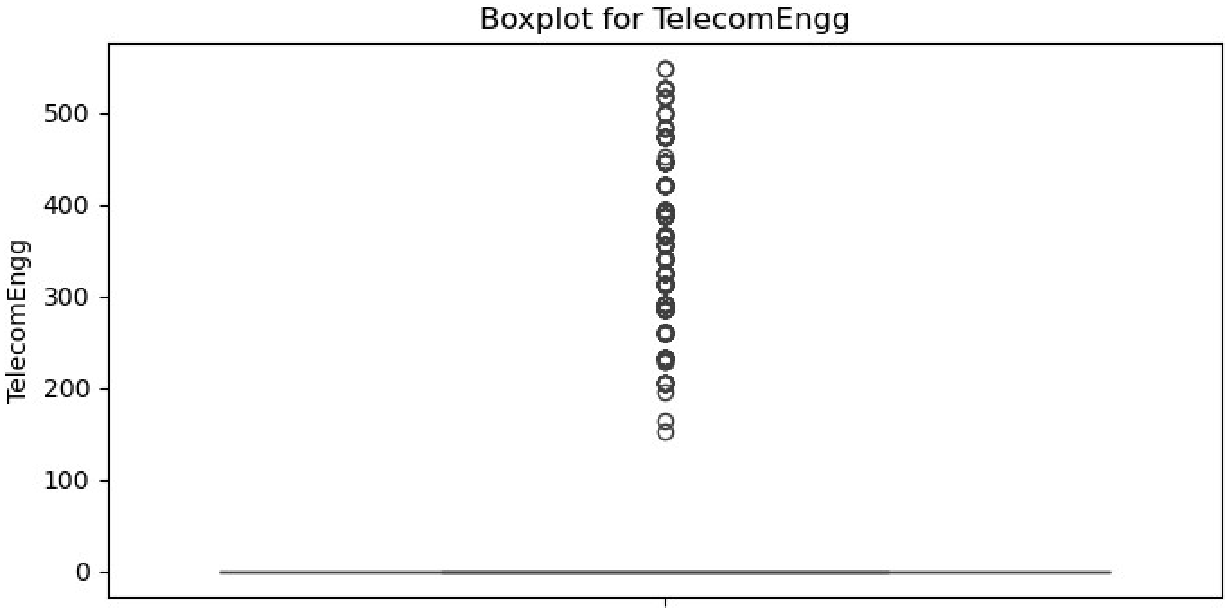


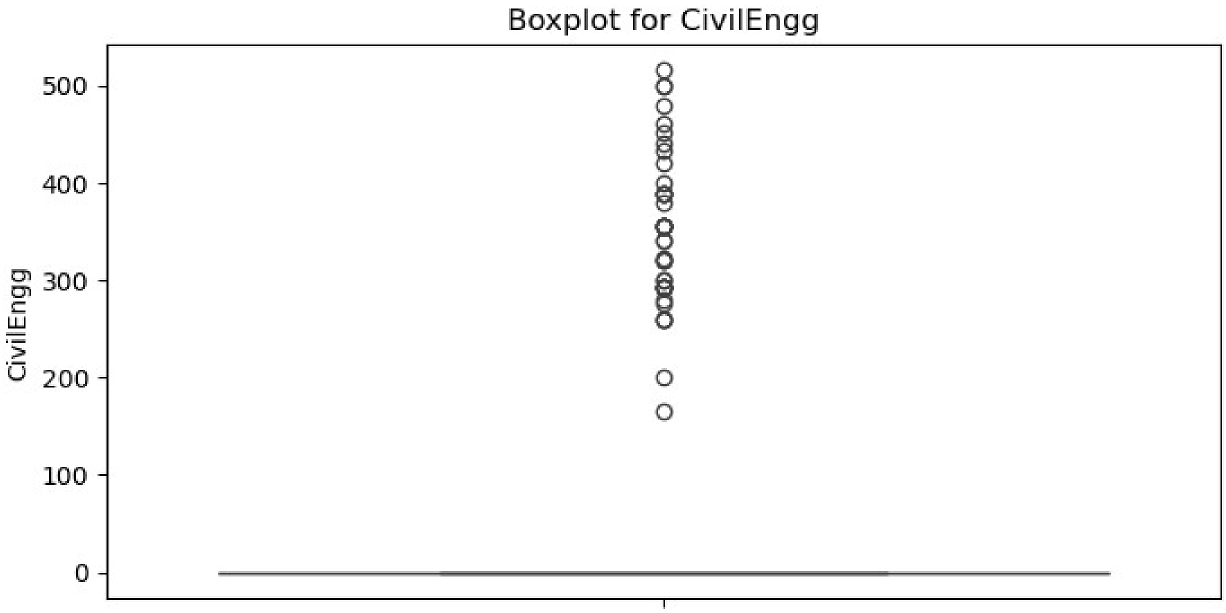
Boxplot for ElectronicsAndSemicon

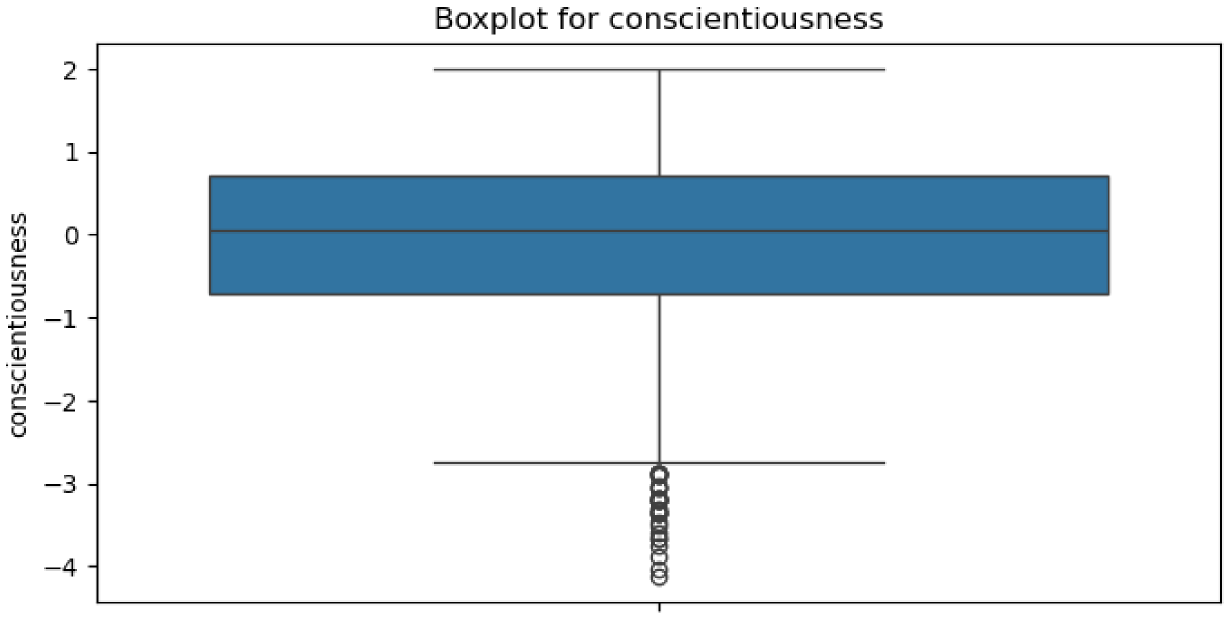

Boxplot for ComputerScience
ComputerScience
700
600
500
400
300
200
100
0

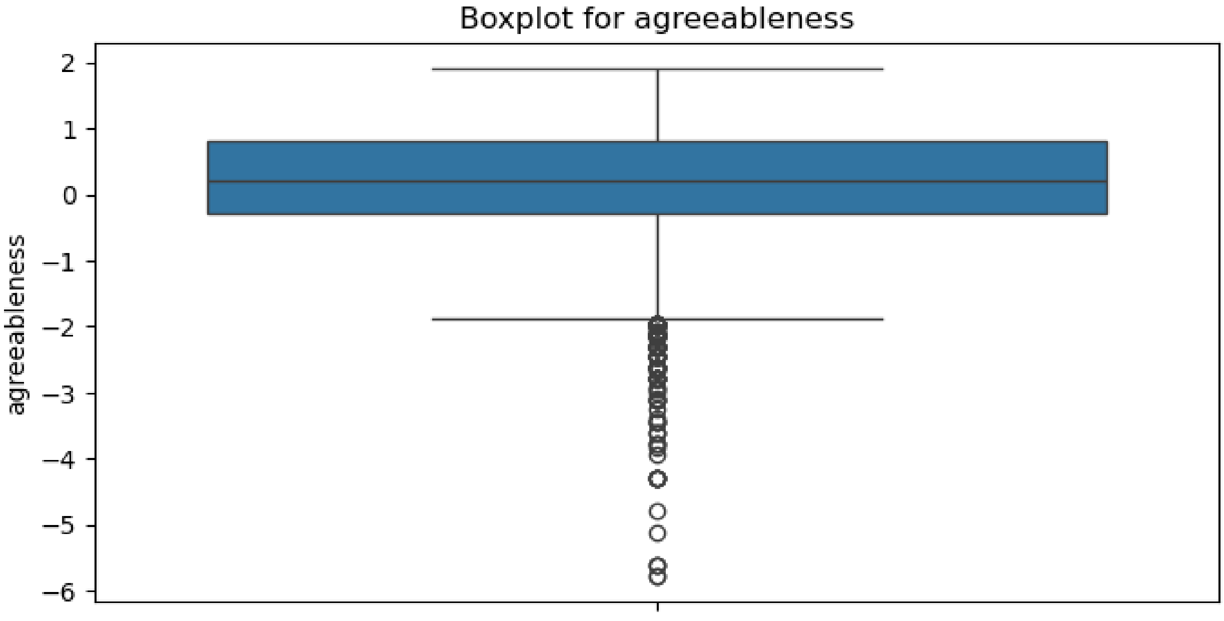

Boxplot for MechanicalEngg
MechanicalEngg
600
500
400
300
200
100
0

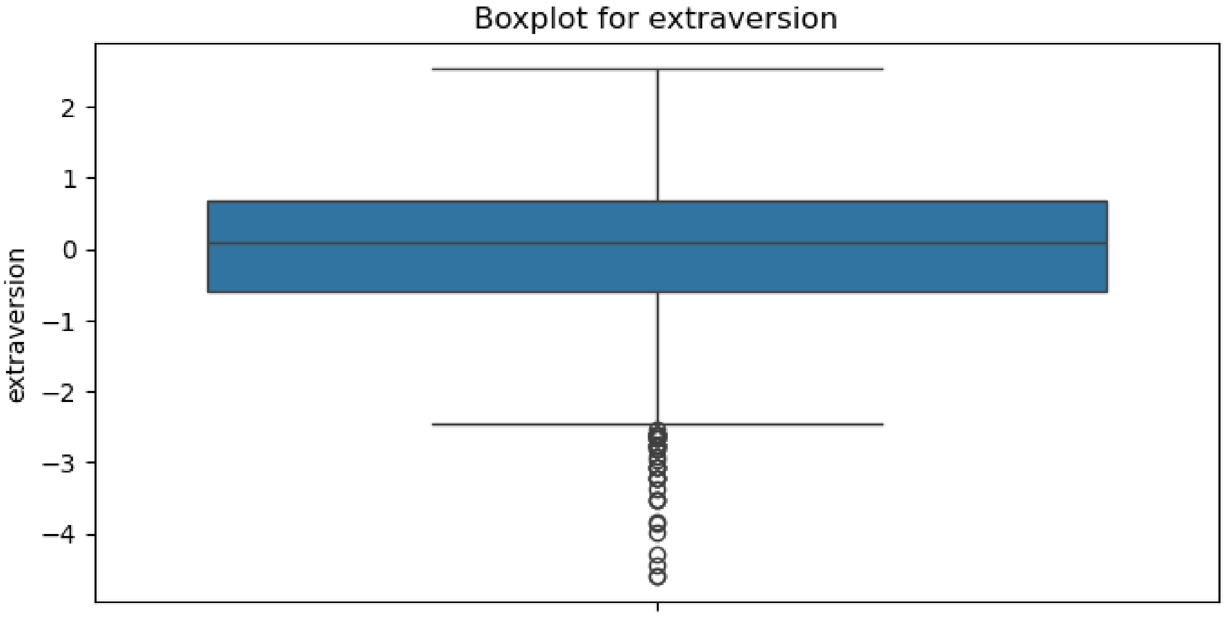

Boxplot for ElectricalEngg
ElectricalEngg
700
600
500
400
300
200
100
0

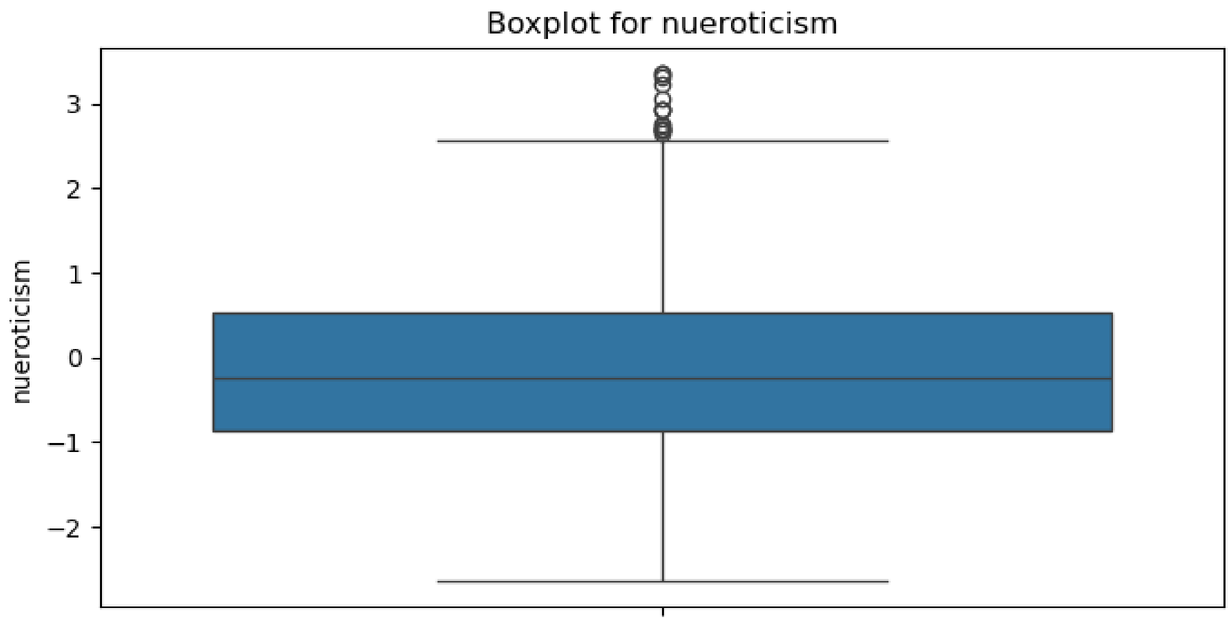

Boxplot for TelecomEngg
TelecomEngg
500
400
300
200
100
0

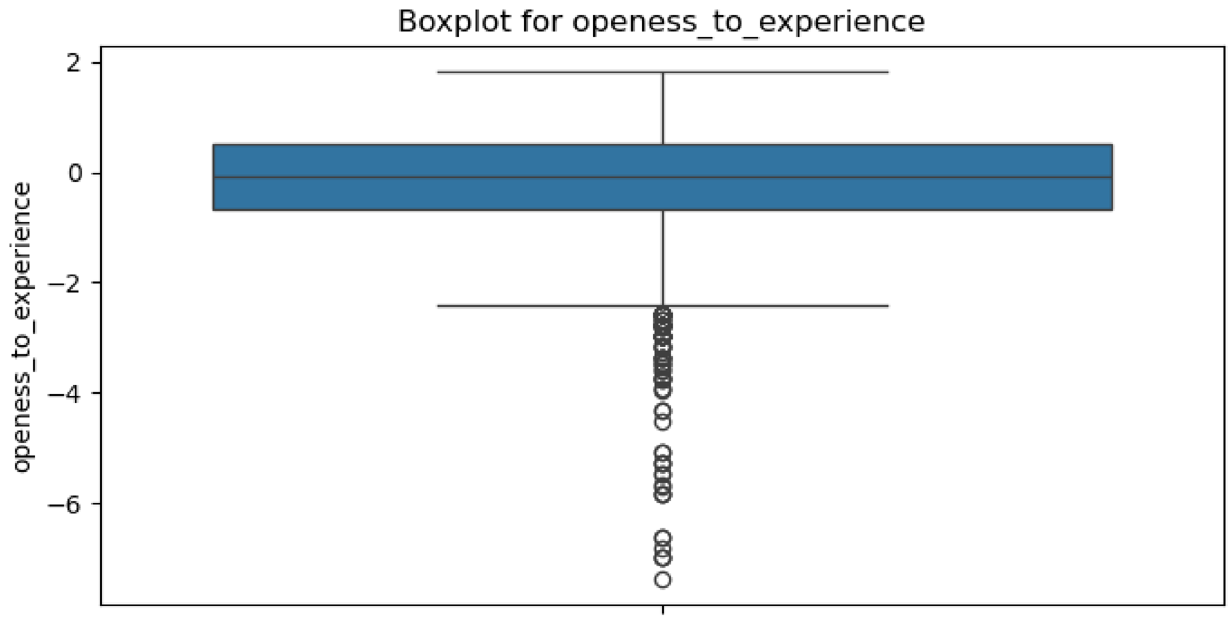

Boxplot for CivilEngg
CivilEngg
500
400
300
200
100
0


Boxplot for conscientiousness
conscientiousness
2
1
0
−1
−2
−3
−4

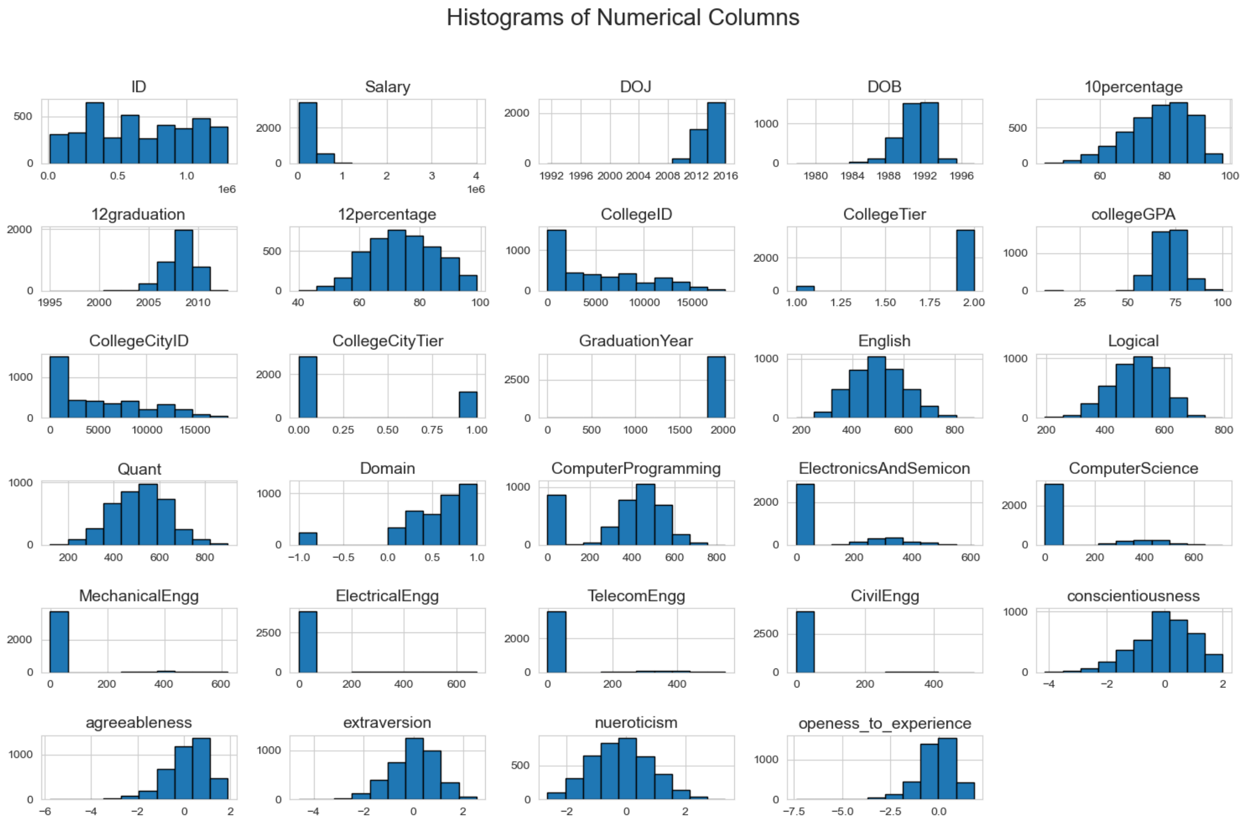

Boxplot for agreeableness


Boxplot for extraversion

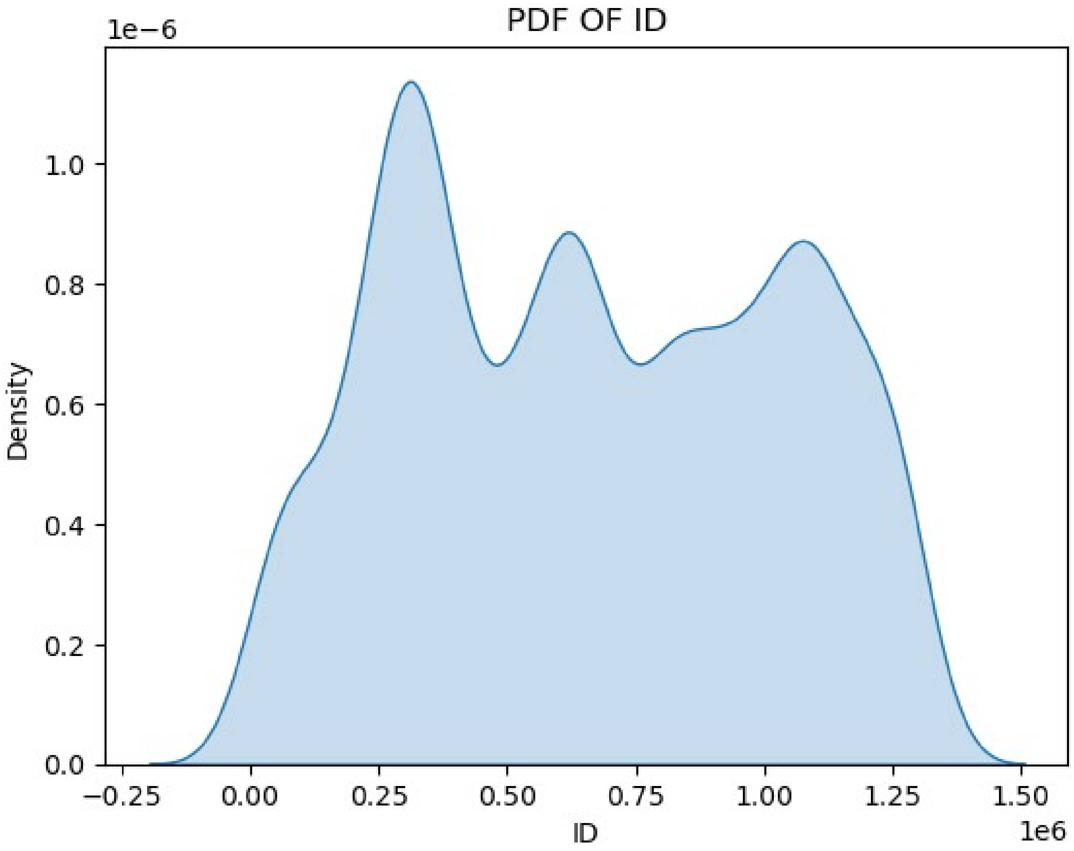


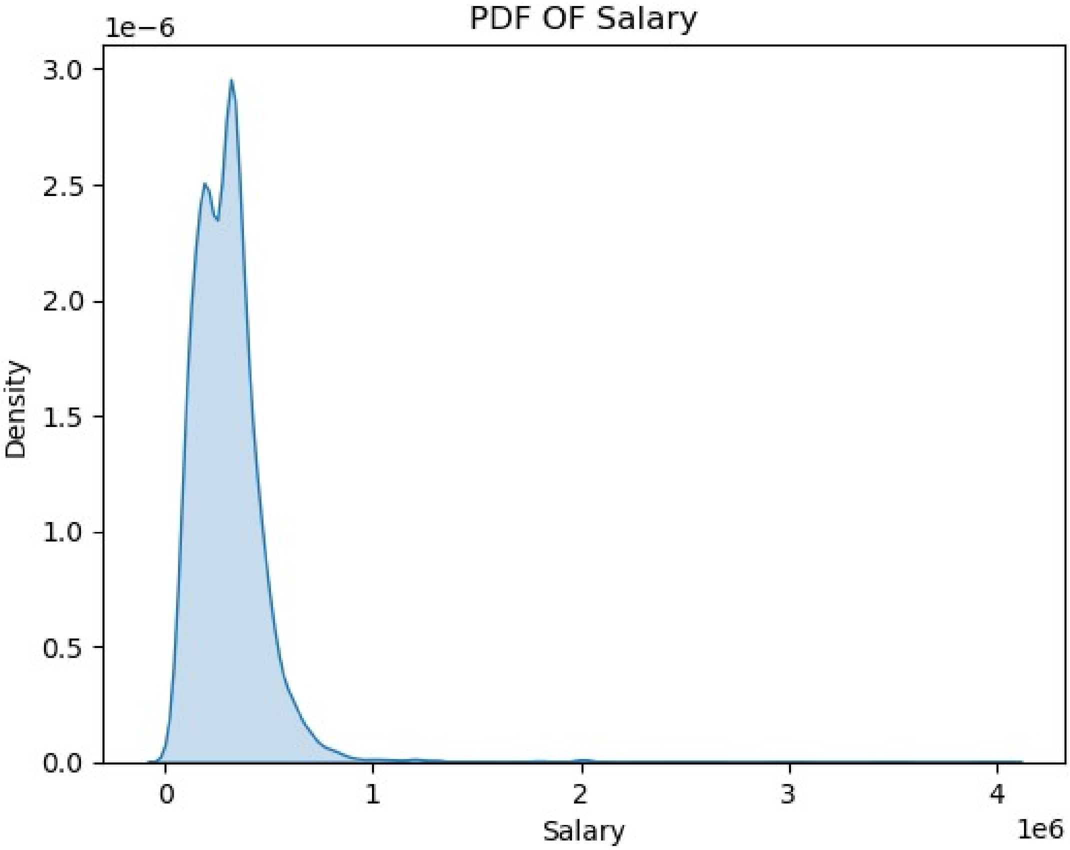


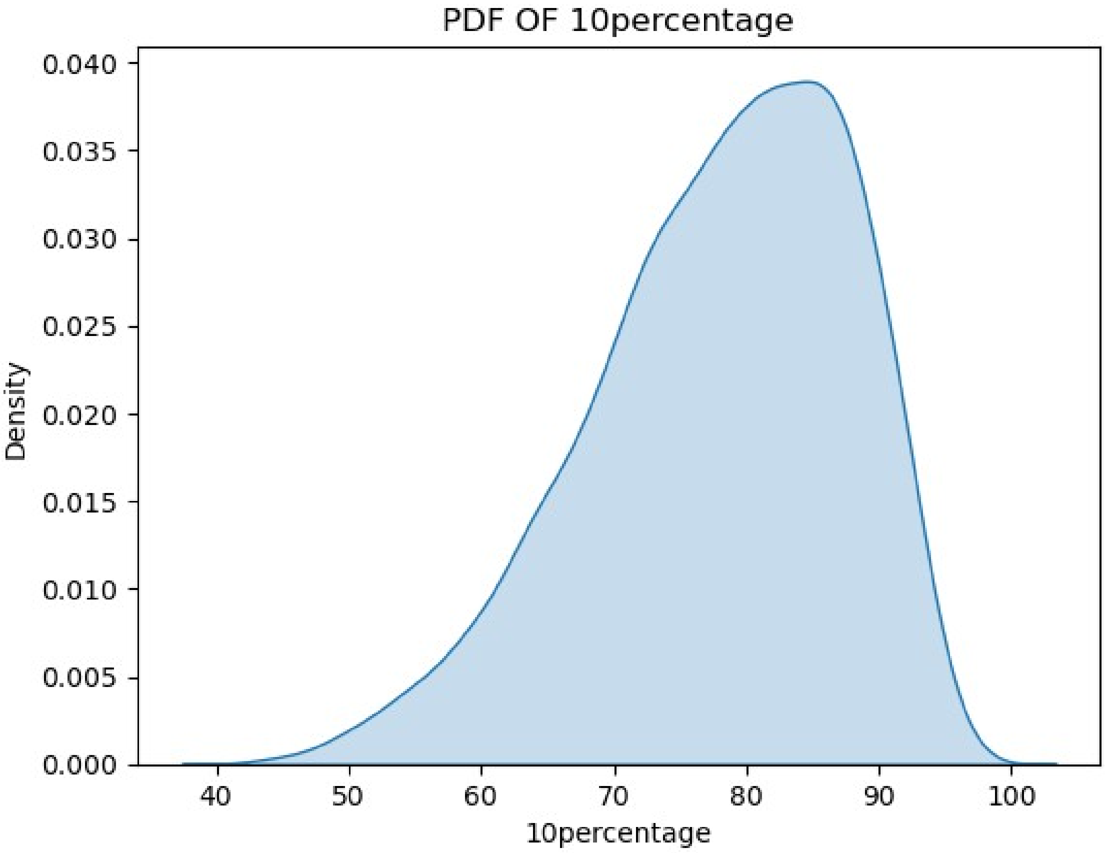
```
df.hist(bins=10, figsize=(15, 10), edgecolor='black')
plt.suptitle('Histograms of Numerical Columns', fontsize=20)
for ax in plt.gcf().axes:
    ax.set_xlabel(ax.get_xlabel(), fontsize=12)
    ax.set_ylabel(ax.get_ylabel(), fontsize=12)
    ax.title.set_size(14)
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()
```

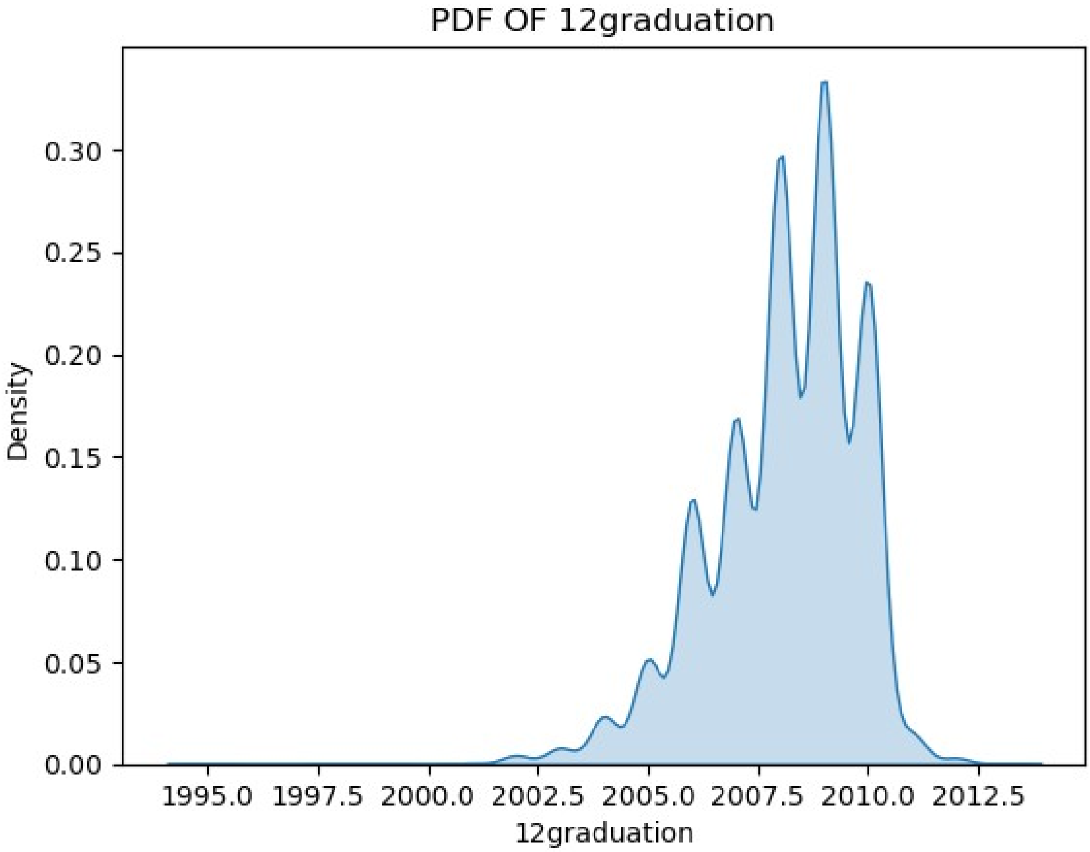


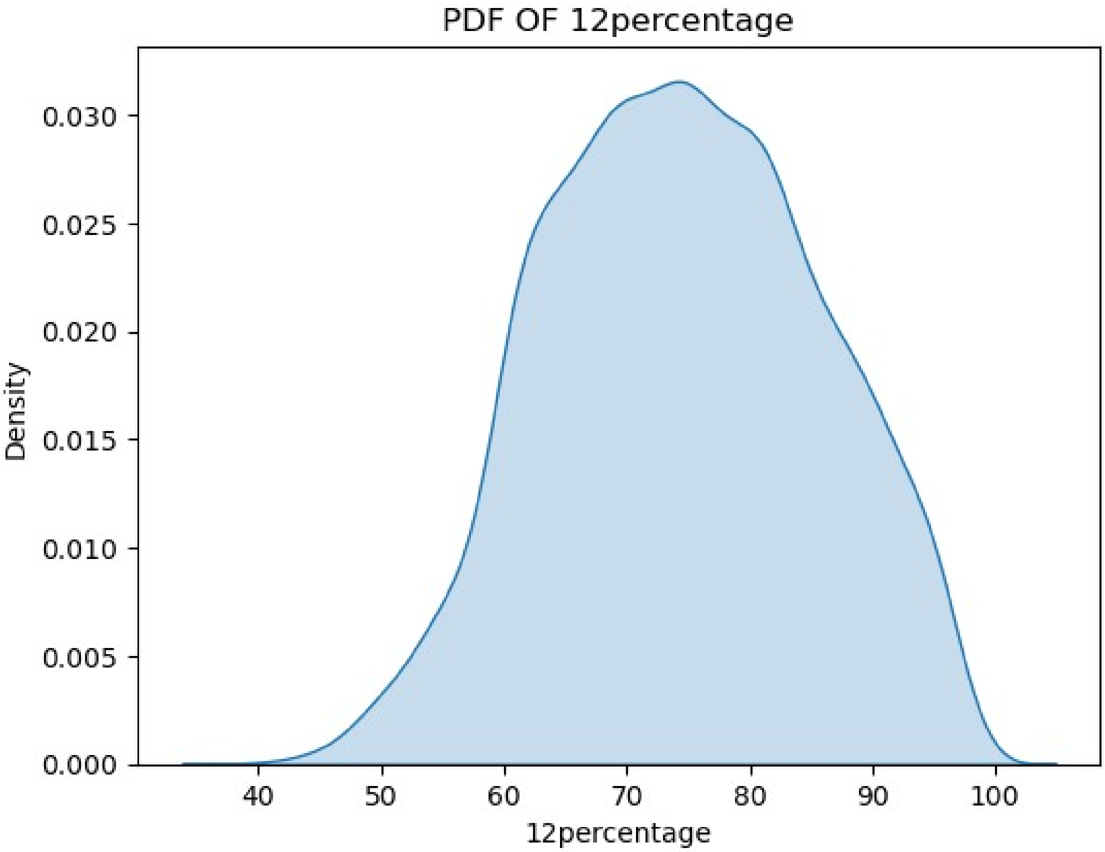
```
for col in df.select_dtypes(include= 'number').columns:
  sns.kdeplot(df[col],fill=True)
  plt.title(f"PDF OF {col}")
  plt.show()
```

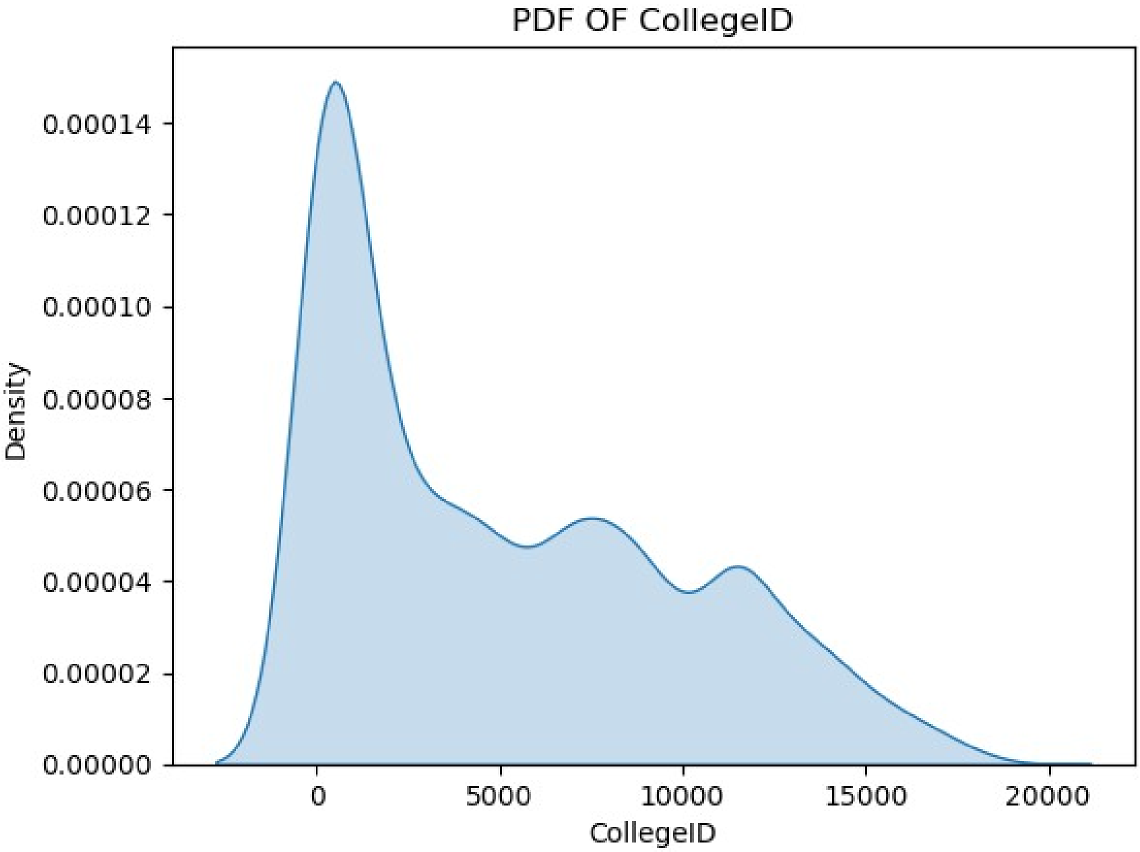

PDF OF ID
1e−6
1.0
0.8
0.6
0.4
0.2
0.0
Density
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e6
ID

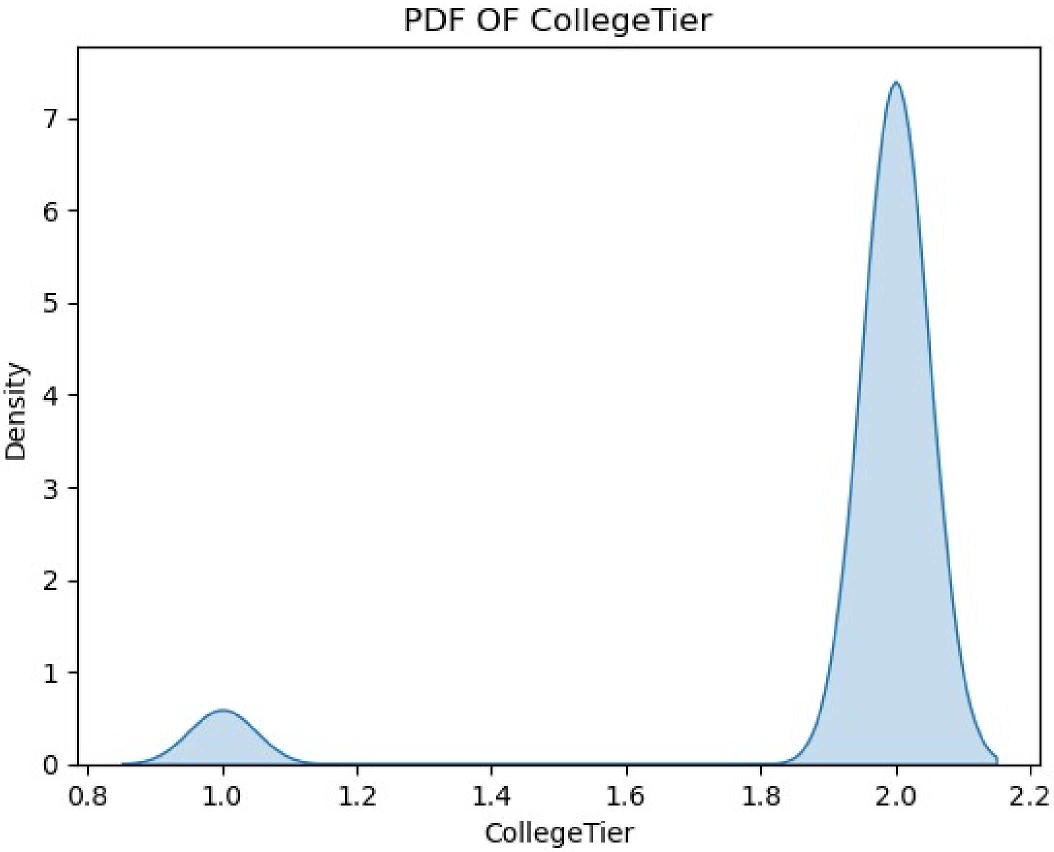

PDF OF Salary
1e−6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
0
1
2
3
4
1e6
Salary

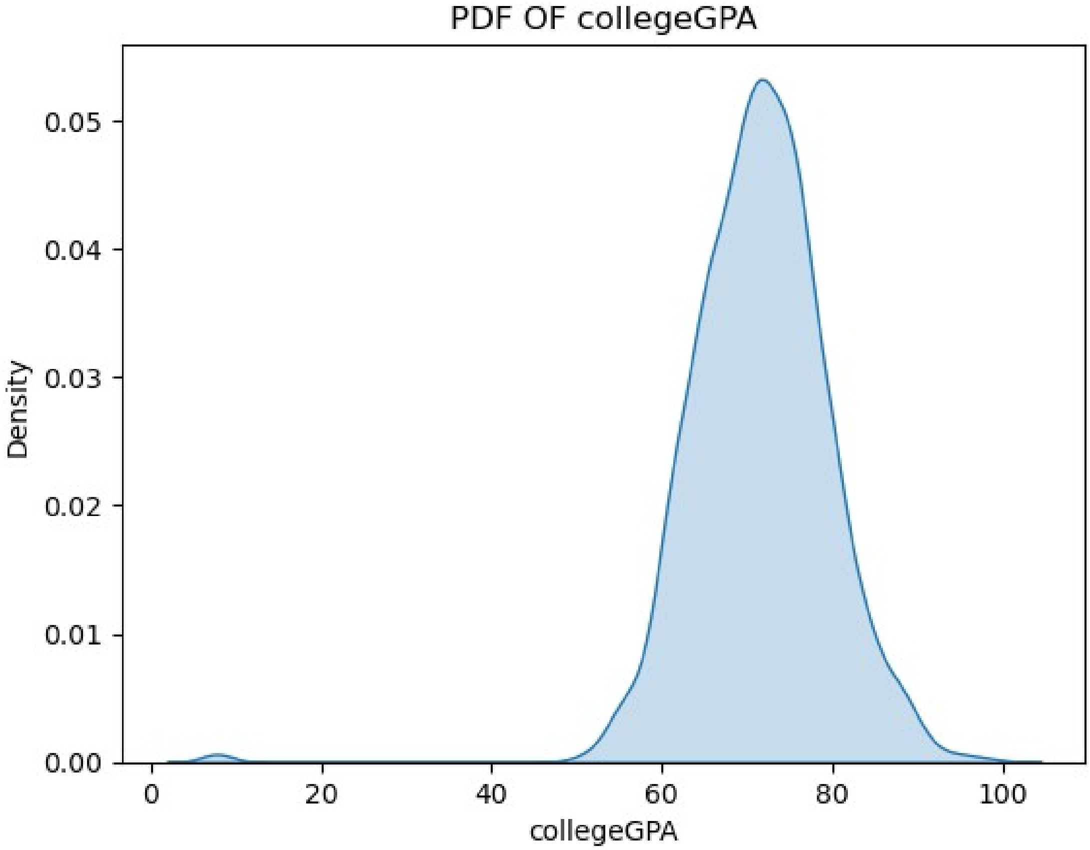

PDF OF 10percentage
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
40
50
60
70
80
90
100
10percentage

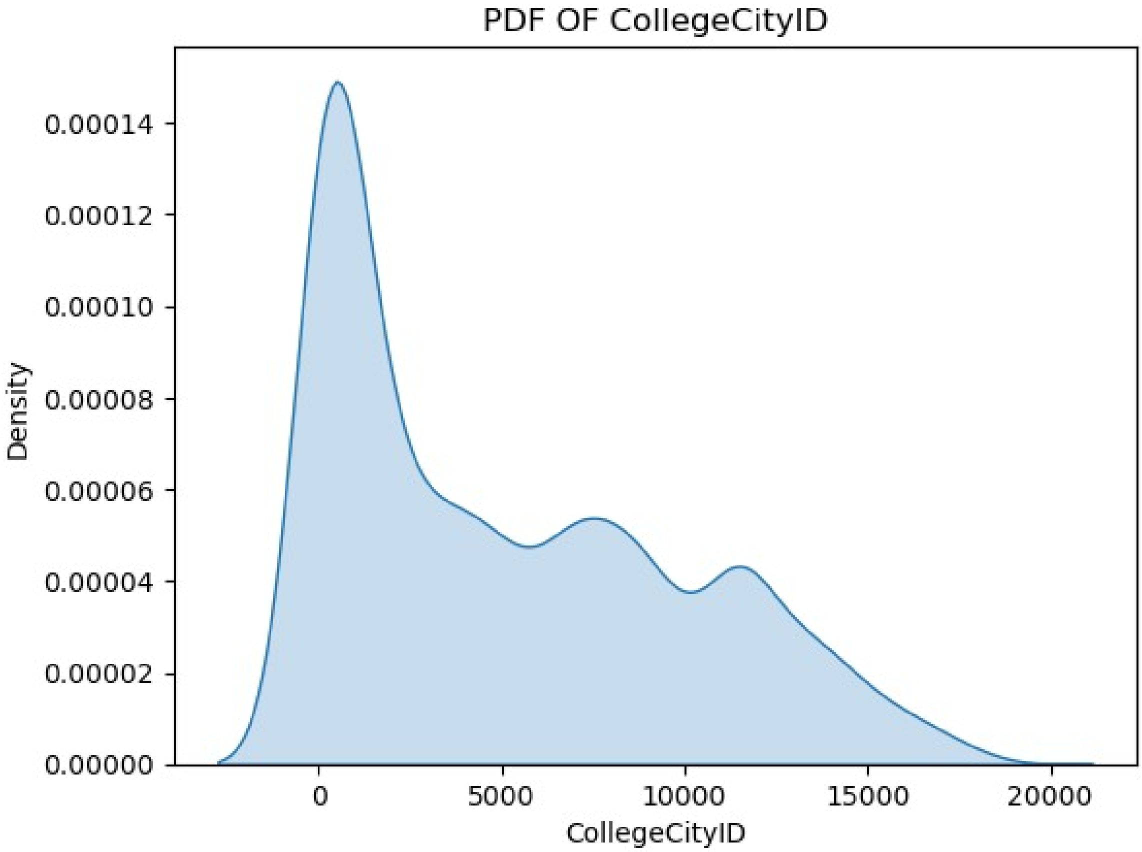

PDF OF 12graduation
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1995.0
1997.5
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
12graduation

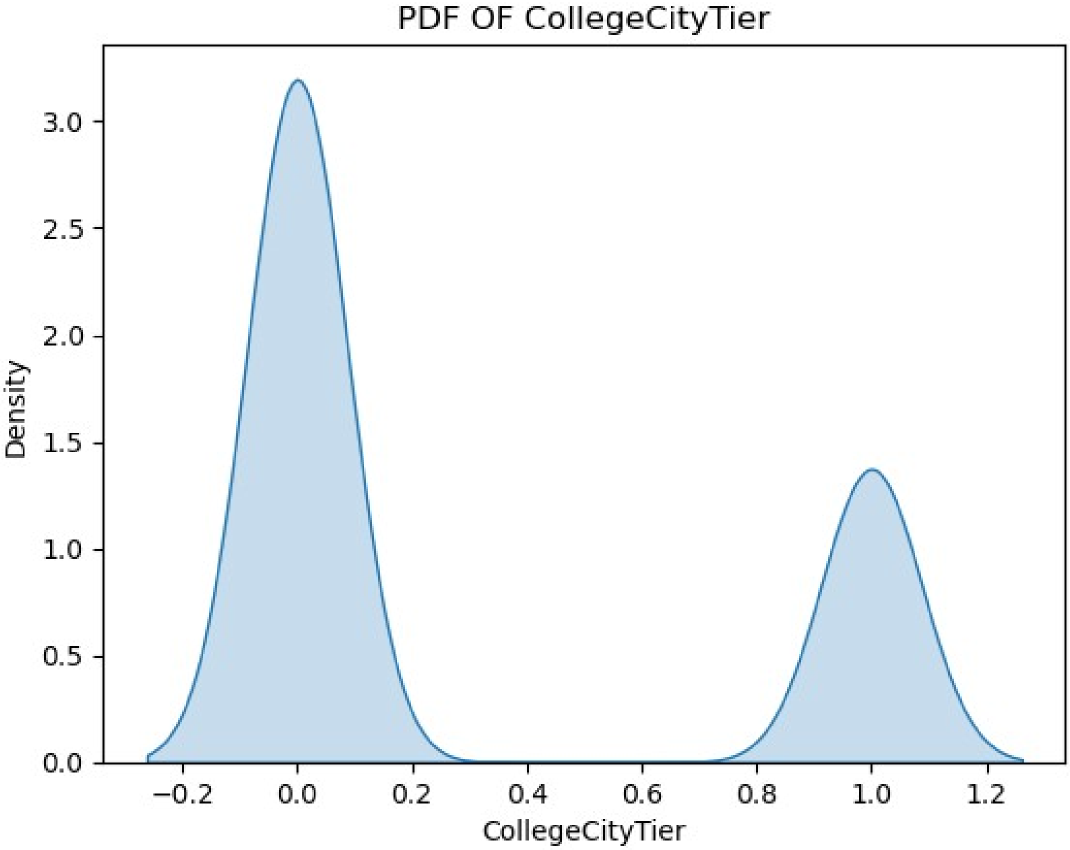

PDF OF 12percentage
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
40
50
60
70
80
90
100
12percentage

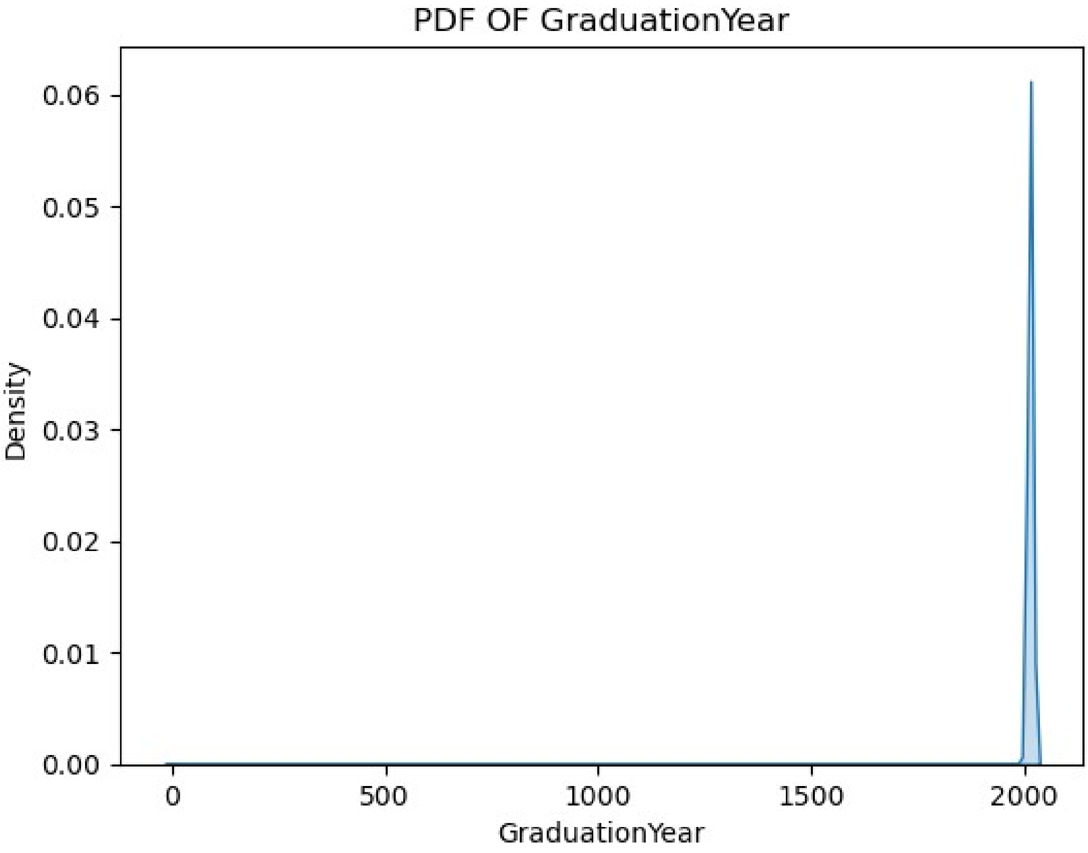

PDF OF CollegeID
Density
0.00000
0.00002
0.00004
0.00006
0.00008
0.00010
0.00012
0.00014
0
5000
10000
15000
20000
CollegeID

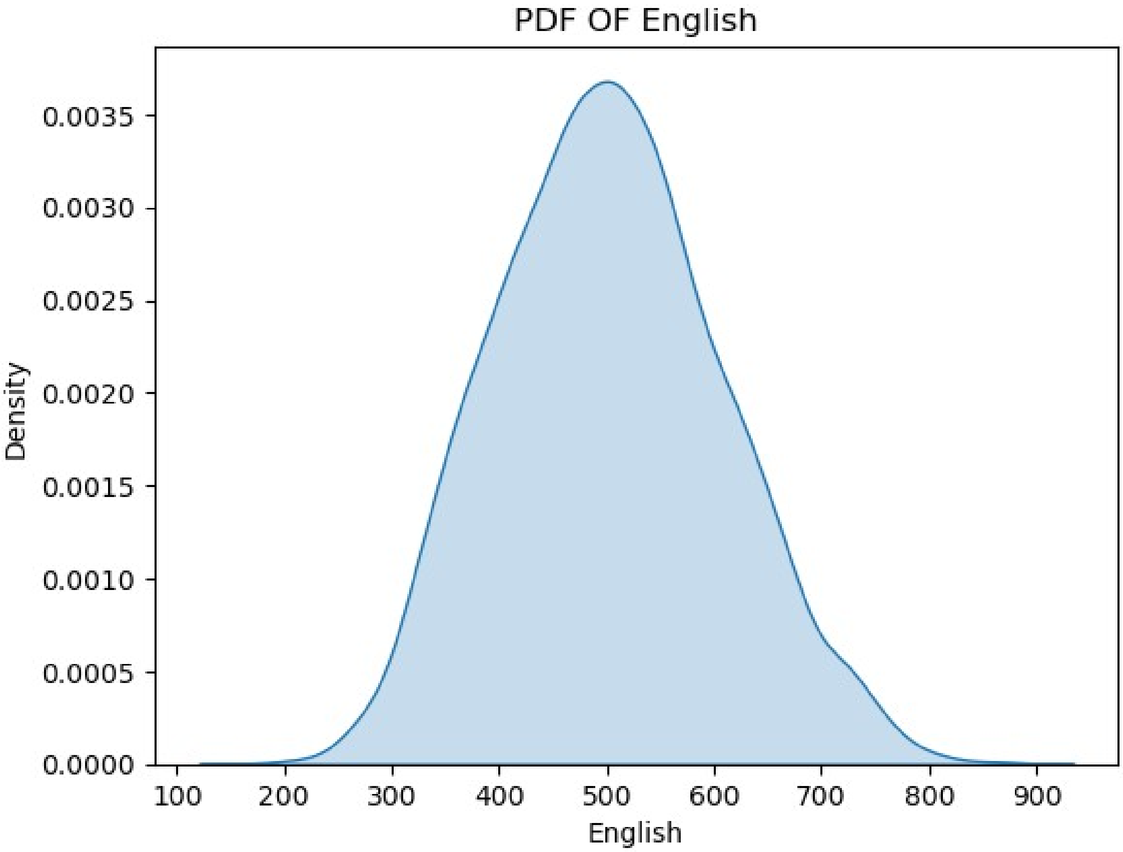

PDF OF CollegeTier
Density
0
1
2
3
4
5
6
7
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
CollegeTier

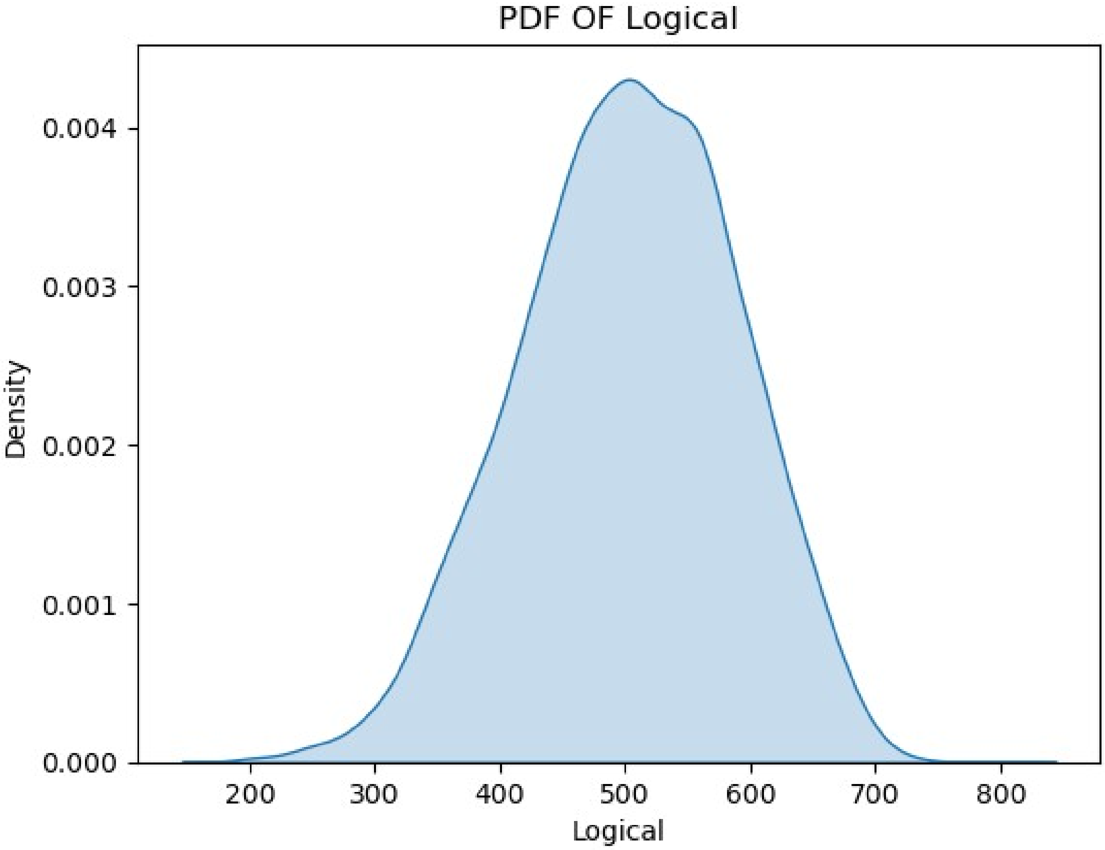

PDF OF collegeGPA
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
100
collegeGPA

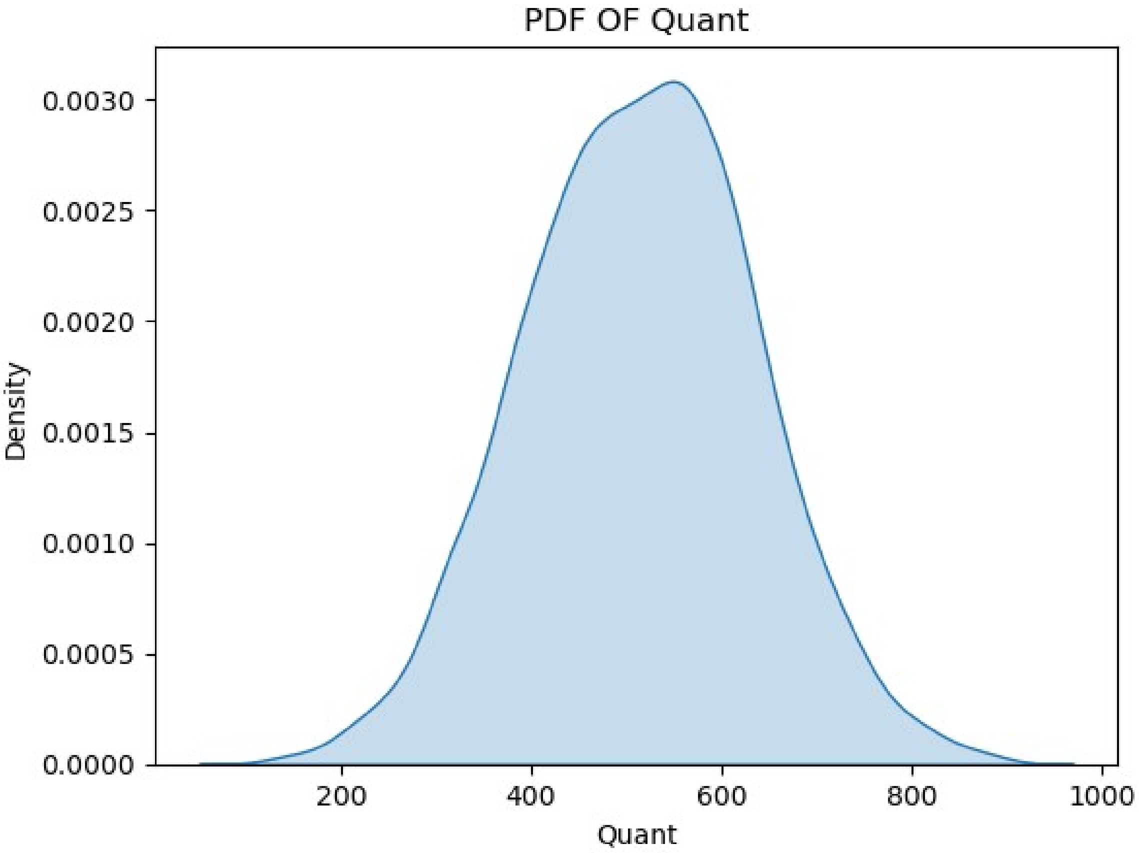

PDF OF CollegeCityID
Density
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
CollegeCityID

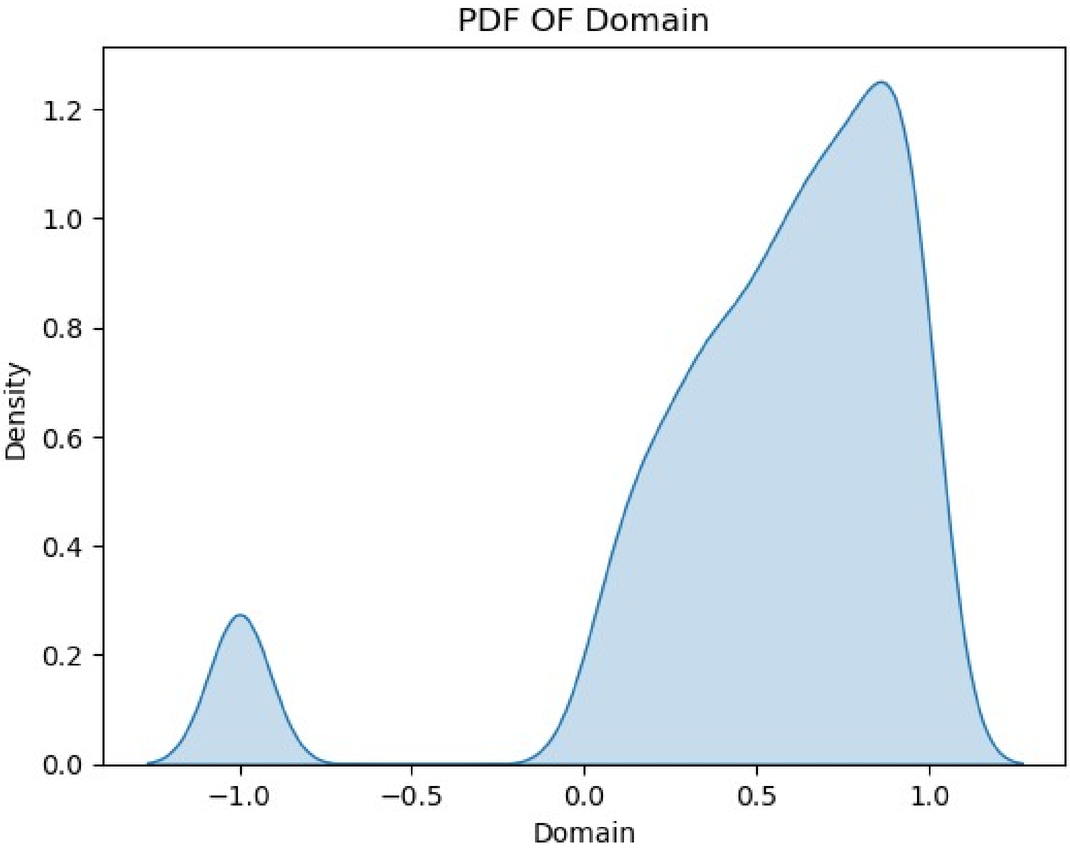

PDF OF CollegeCityTier
Density
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
CollegeCityTier

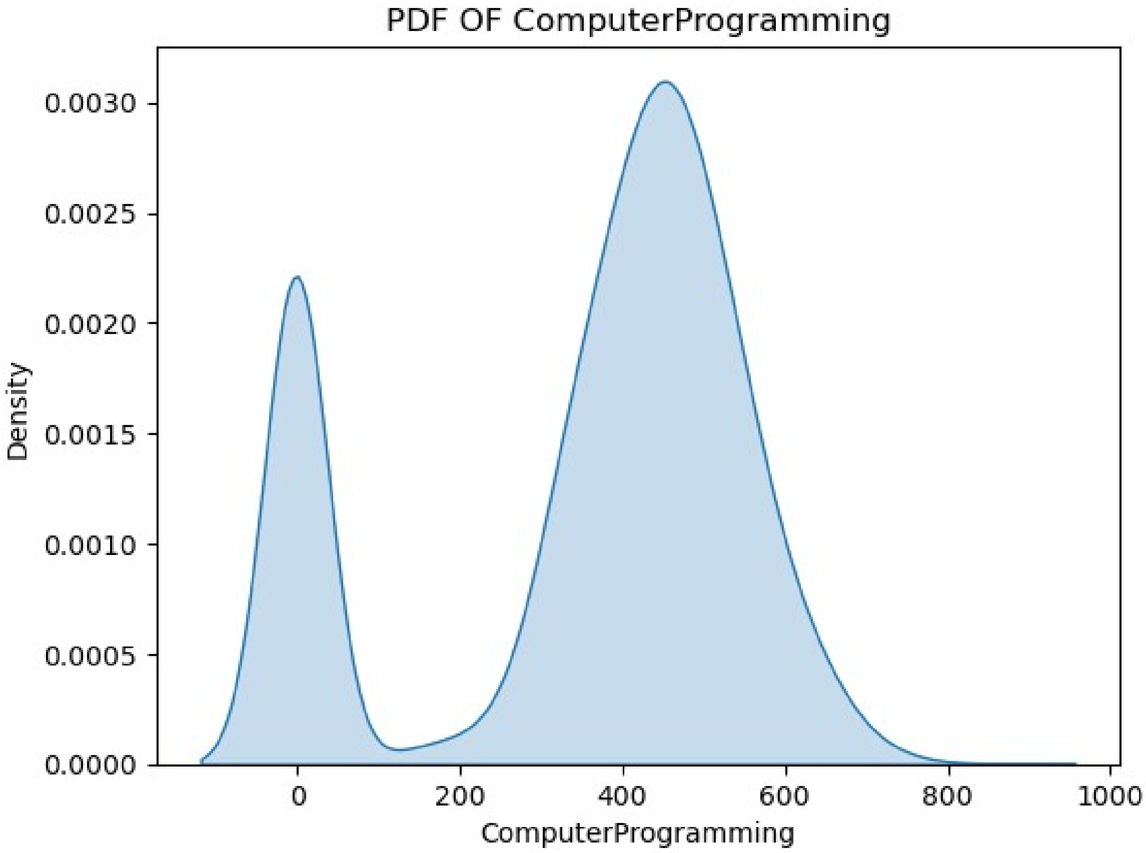

PDF OF GraduationYear
Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
500
1000
1500
2000
GraduationYear

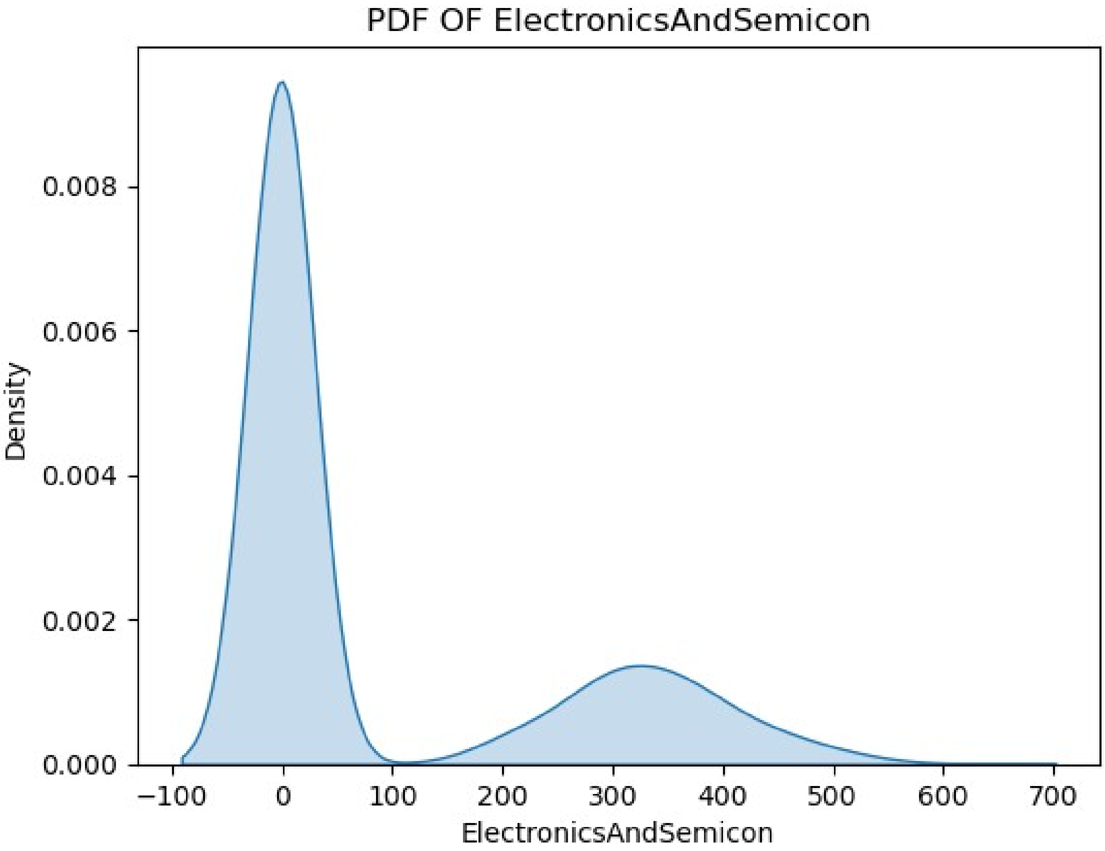

PDF OF English
Density
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
100
200
300
400
500
600
700
800
900
English

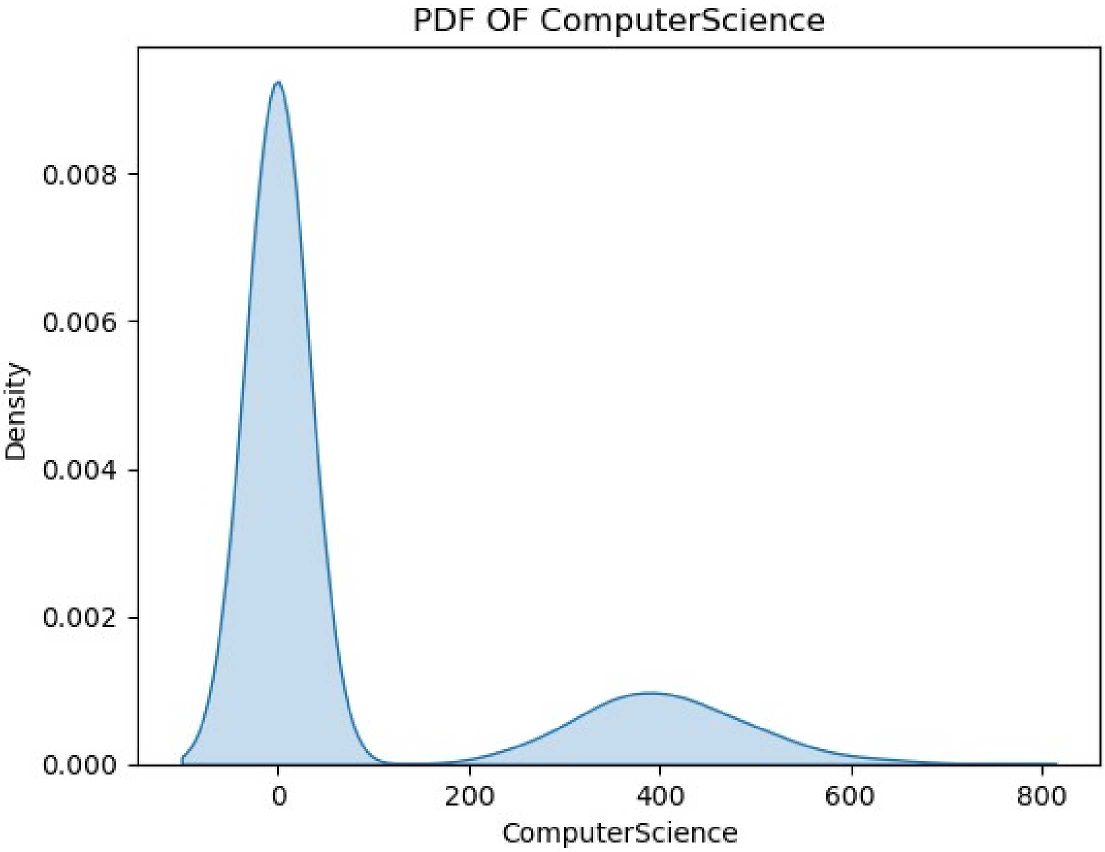

PDF OF Logical
Density
0.000
0.001
0.002
0.003
0.004
200
300
400
500
600
700
800
Logical

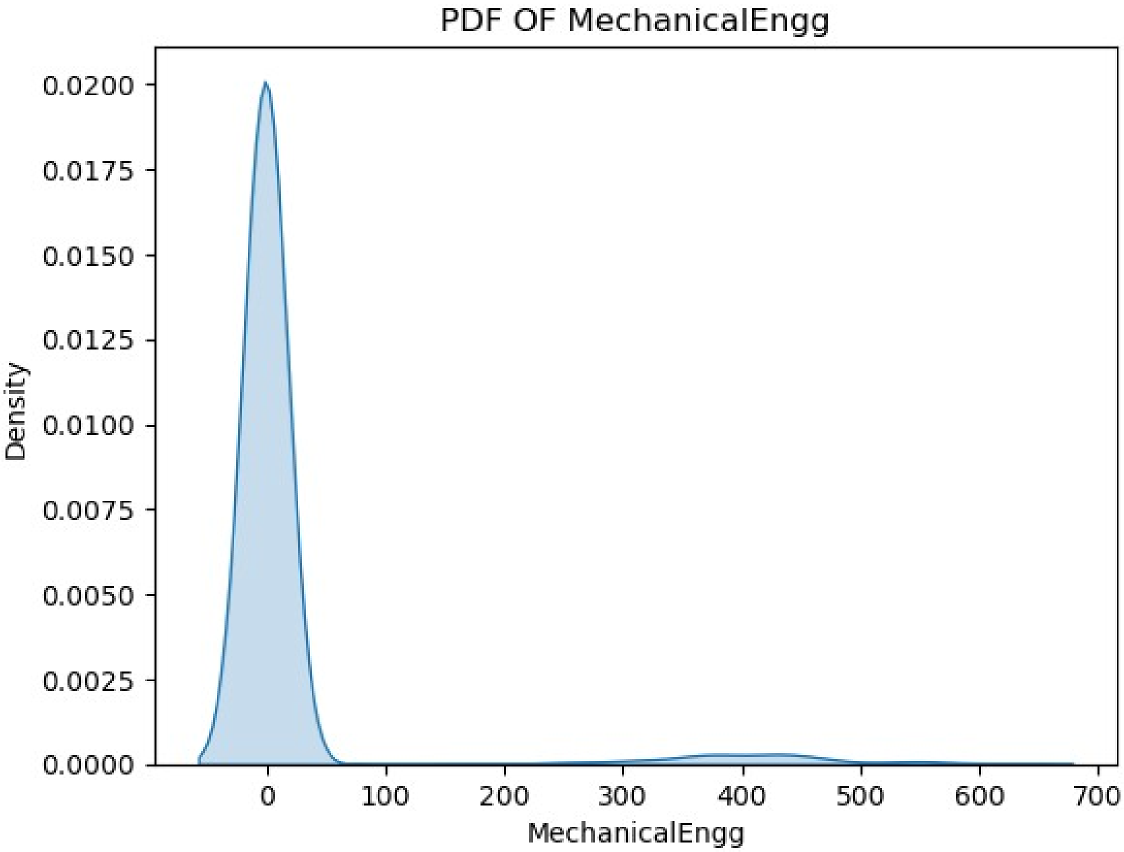

PDF OF Quant
Density
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
200
400
600
800
1000
Quant

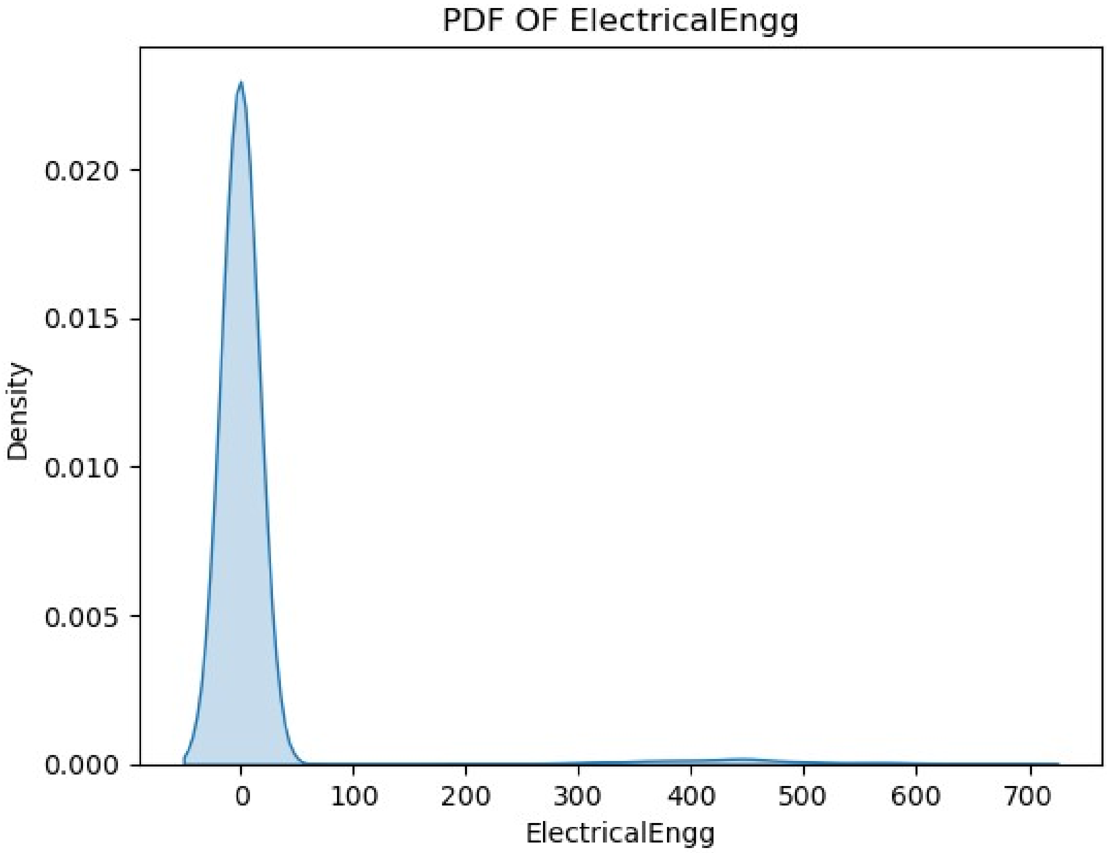

PDF OF Domain
Density
Domain
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−1.0
−0.5
0.0
0.5
1.0

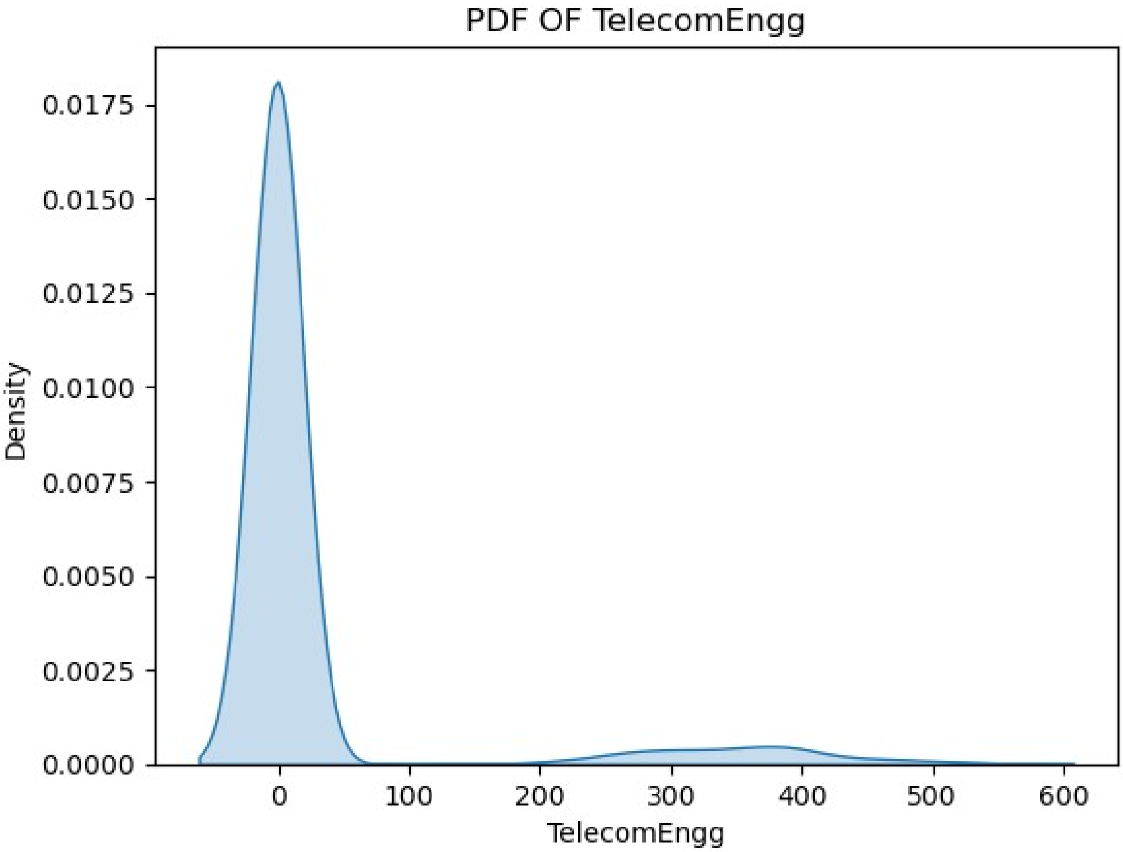

PDF OF ComputerProgramming
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0
200
400
600
800
1000
ComputerProgramming

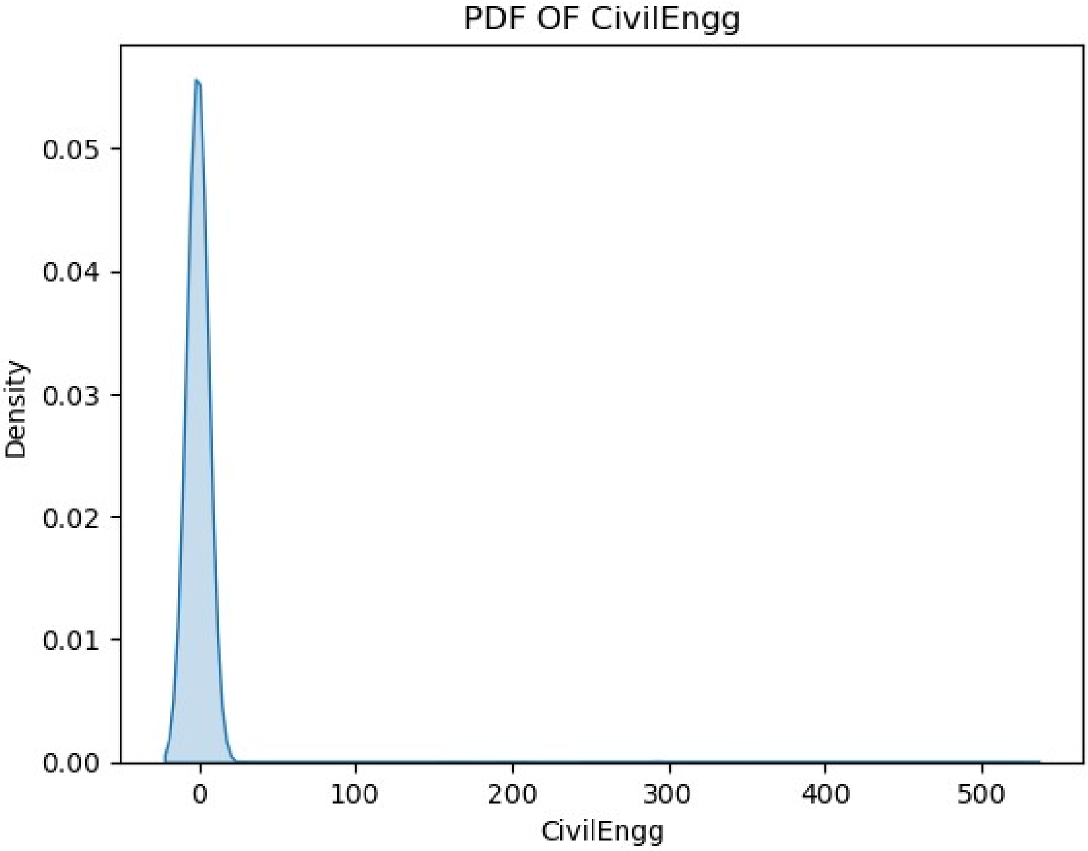

PDF OF ElectronicsAndSemicon
Density
0.000
0.002
0.004
0.006
0.008
−100
0
100
200
300
400
500
600
700
ElectronicsAndSemicon

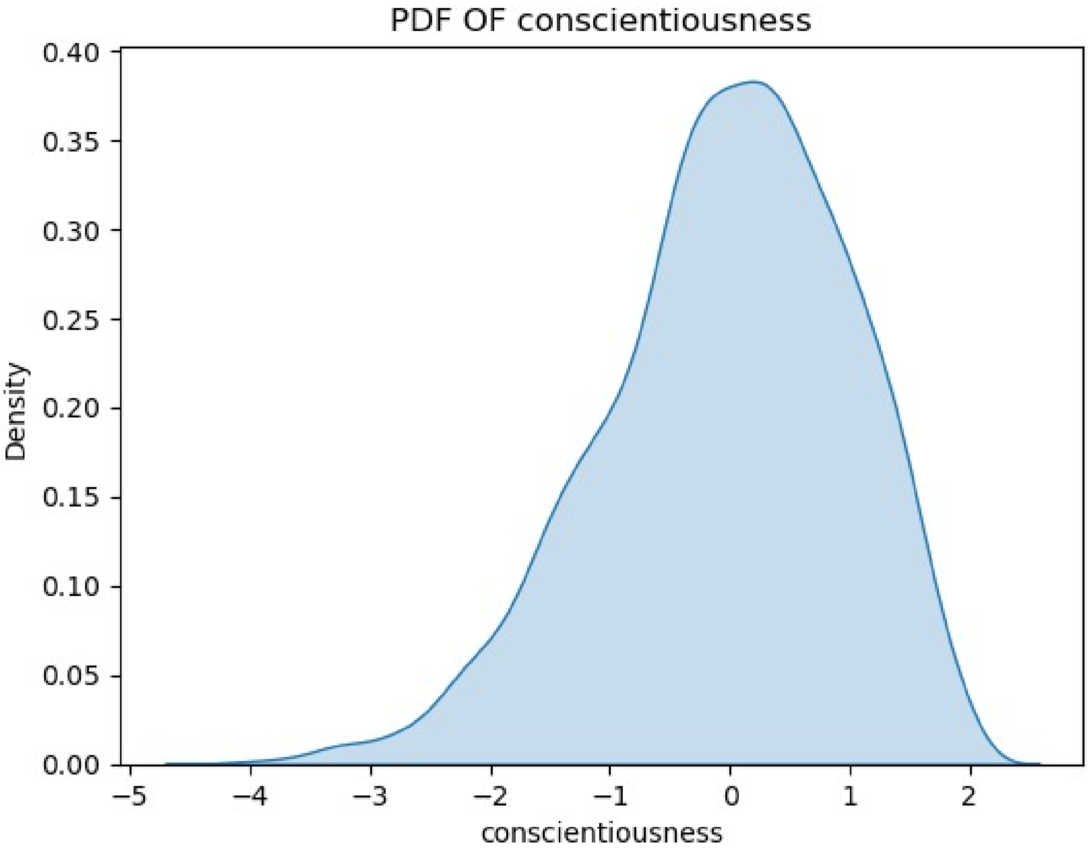

PDF OF ComputerScience
Density
0.000
0.002
0.004
0.006
0.008
0
200
400
600
800
ComputerScience

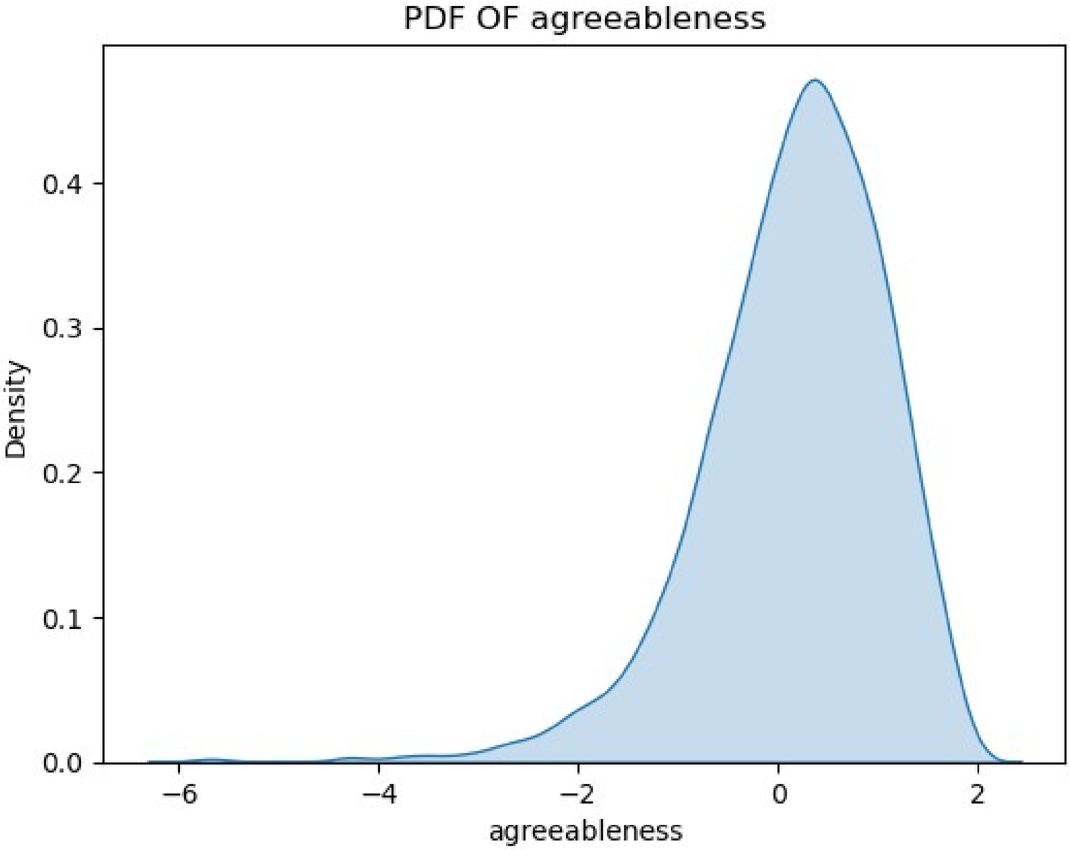

PDF OF MechanicalEngg
Density
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
100
200
300
400
500
600
700
MechanicalEngg

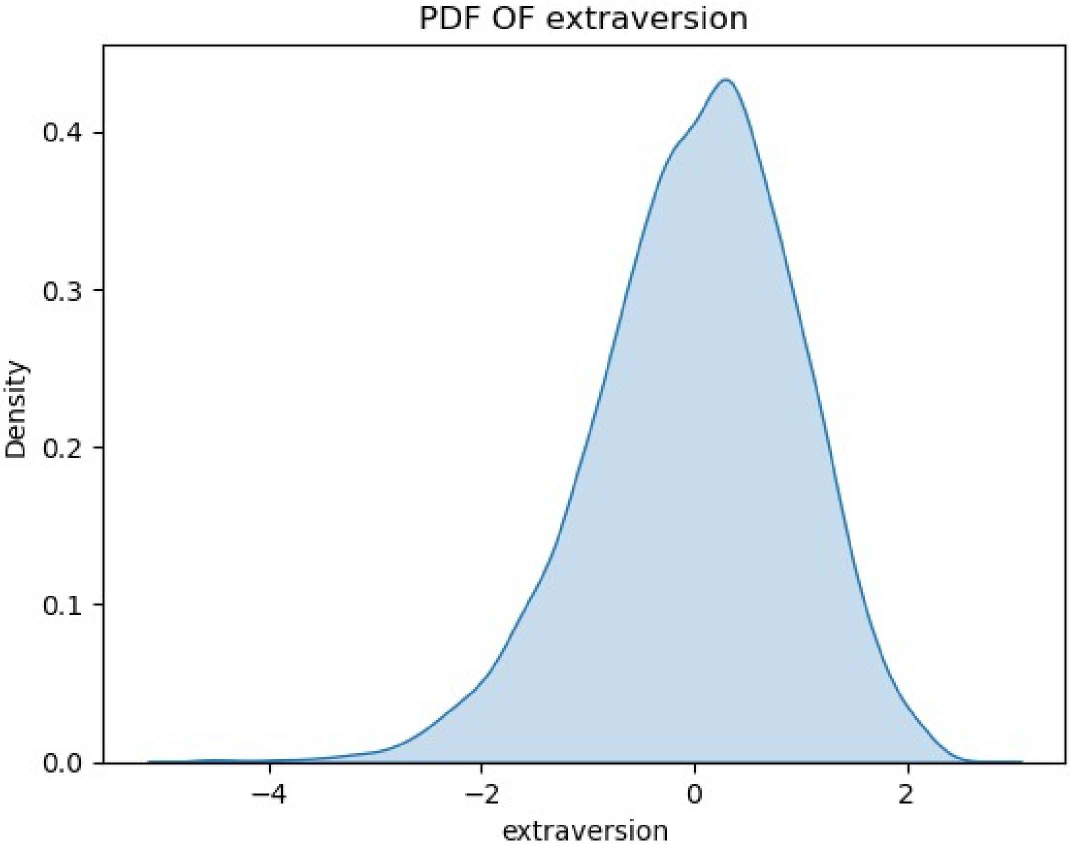

PDF OF ElectricalEngg
Density
0.000
0.005
0.010
0.015
0.020
0
100
200
300
400
500
600
700
ElectricalEngg

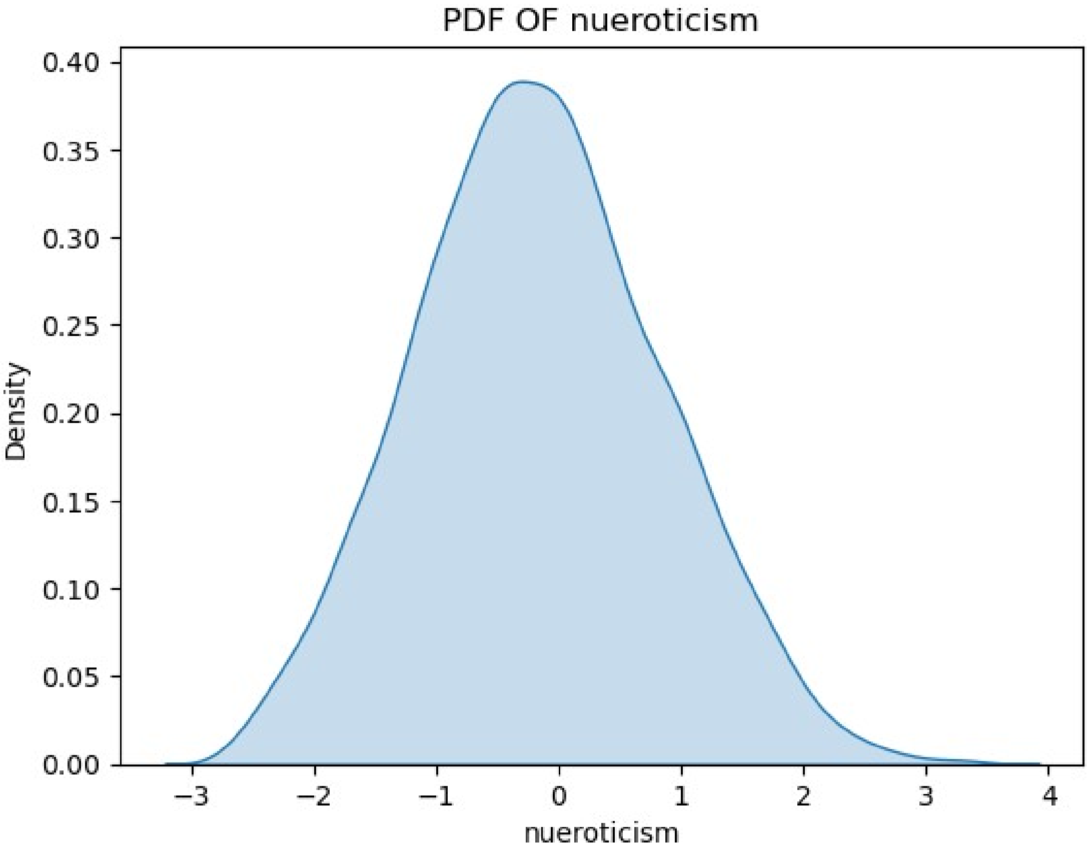

PDF OF TelecomEngg
Density
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
100
200
300
400
500
600
TelecomEngg

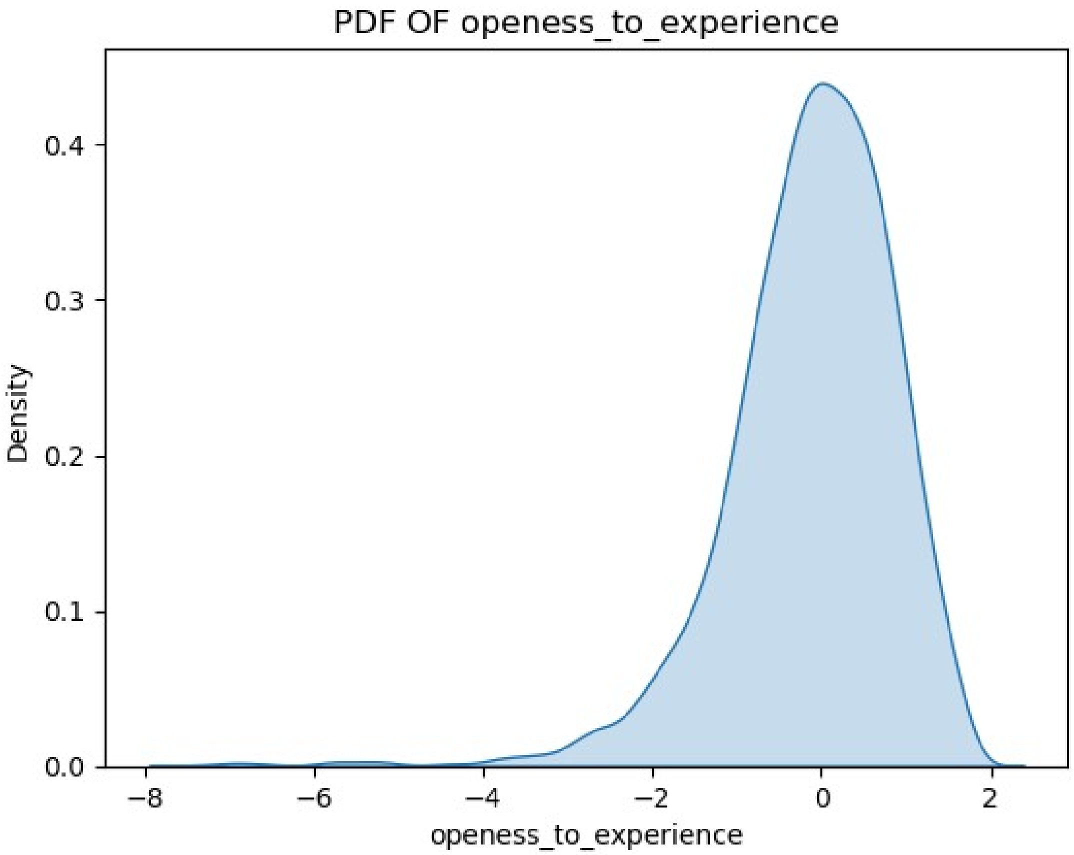

PDF OF CivilEngg
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
CivilEngg


PDF OF conscientiousness
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−5
−4
−3
−2
−1
0
1
2
conscientiousness

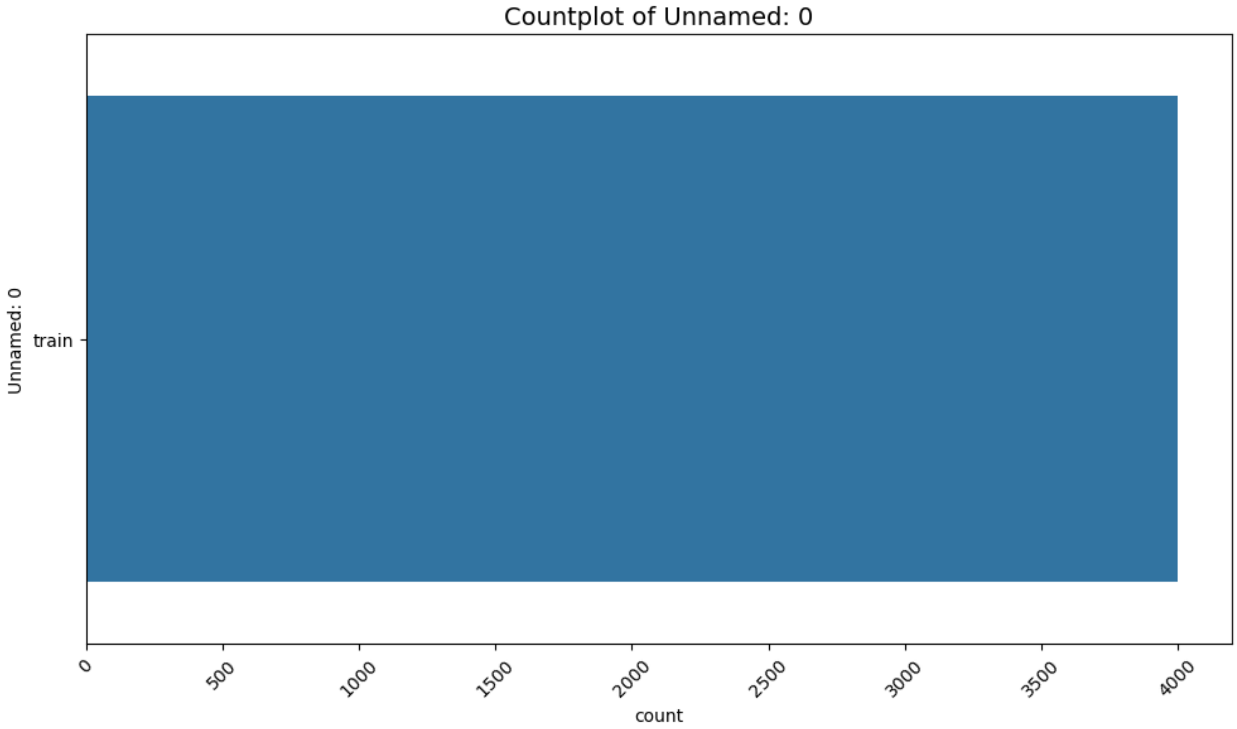

PDF OF agreeableness
Density
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
agreeableness

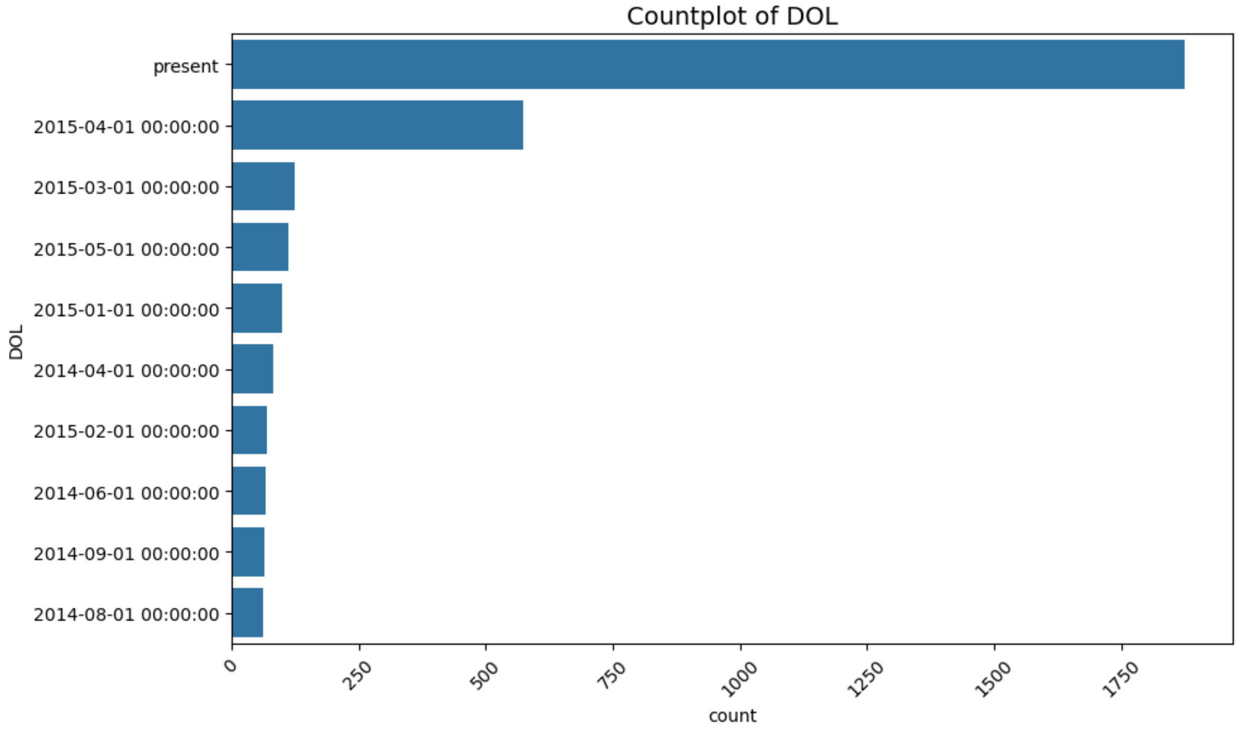

PDF OF extraversion
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
extraversion

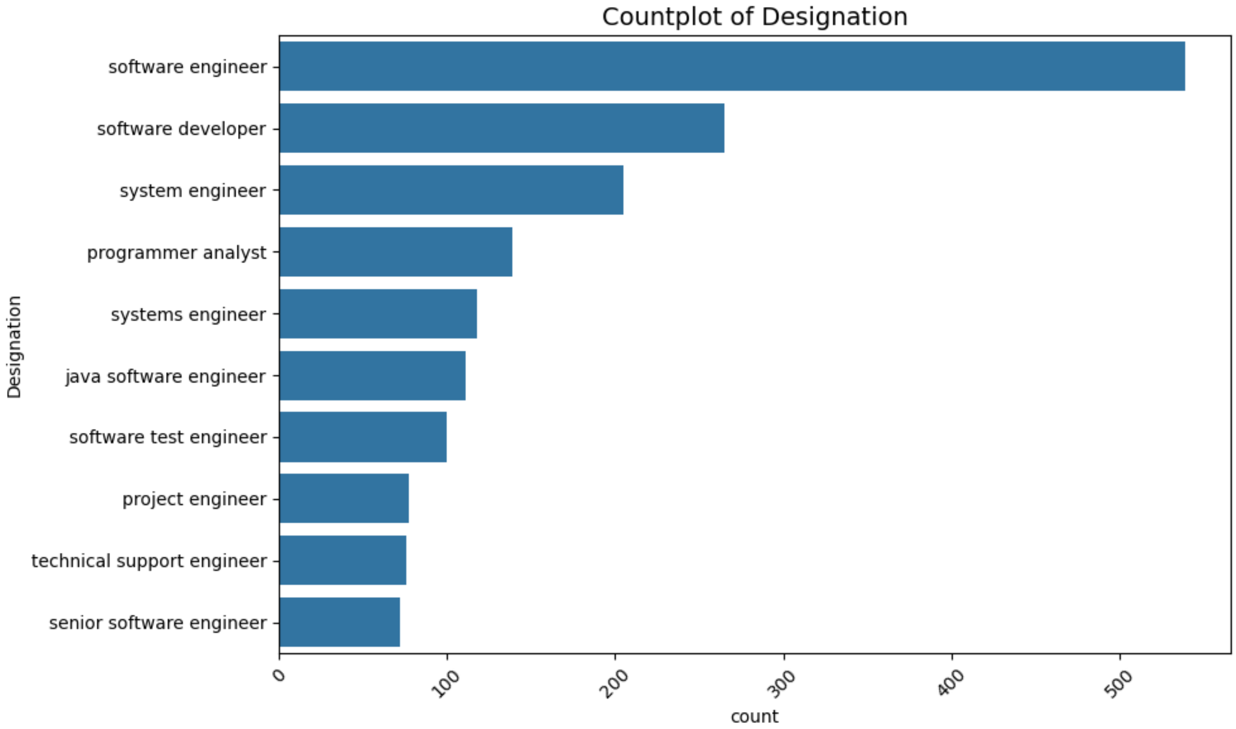

PDF OF nueroticism
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−3
−2
−1
0
1
2
3
4
nueroticism

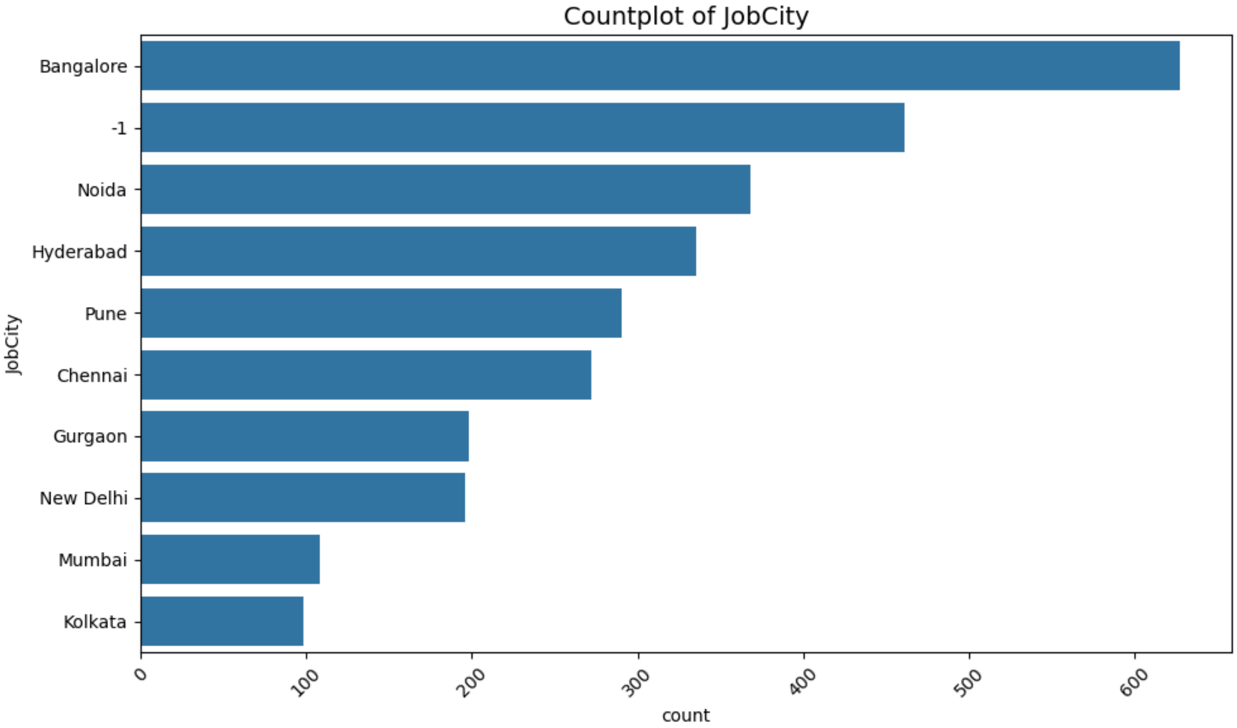


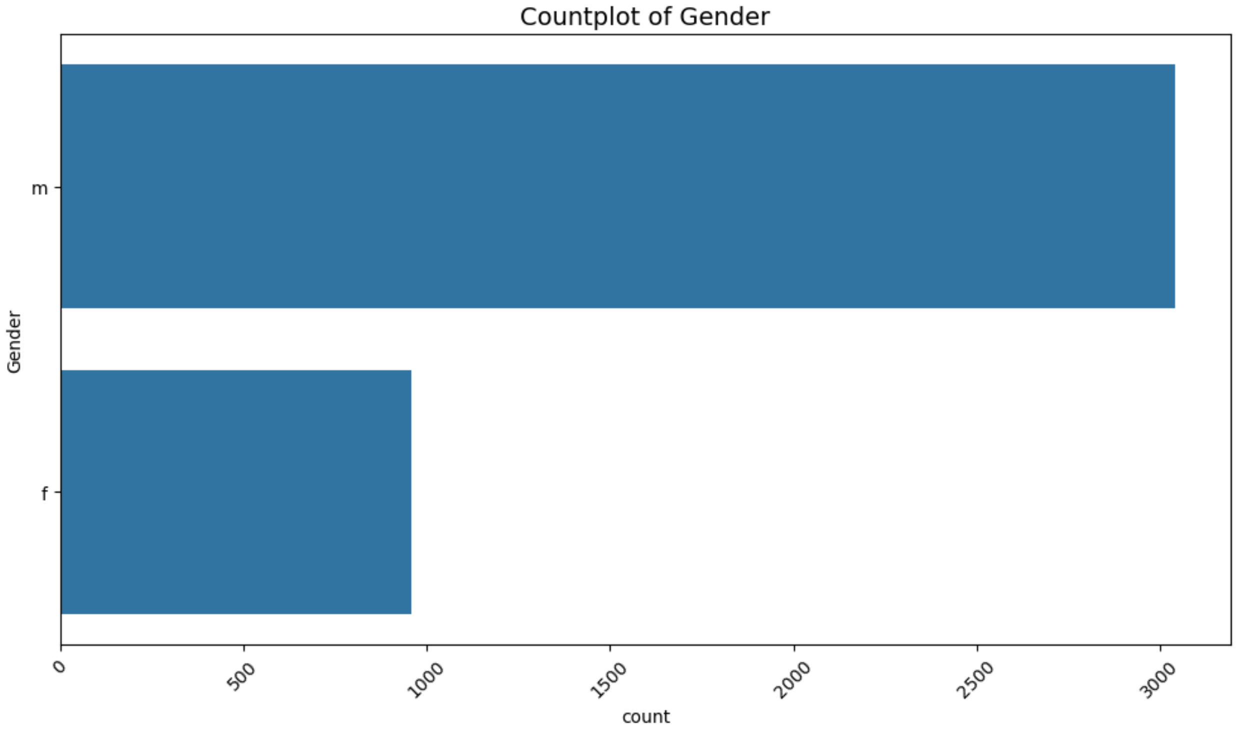
```
for col in df.select_dtypes(include='object').columns:
    plt.figure(figsize=(10, 6))
    sns.countplot(y=df[col], order=df[col].value_counts().index[:10])
    plt.title(f"Countplot of {col}", fontsize=14)
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()
```

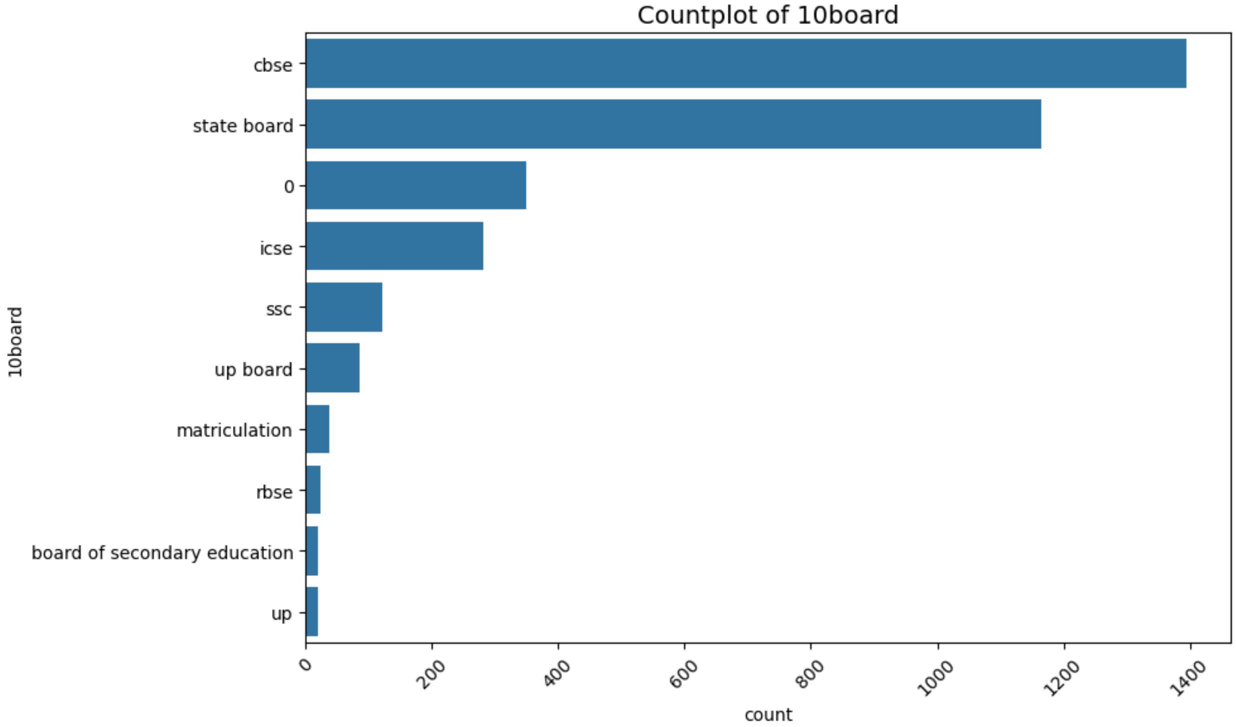

Countplot of Unnamed: 0
Unnamed: 0
train
0
500
1000
1500
2000
2500
3000
3500
4000
count

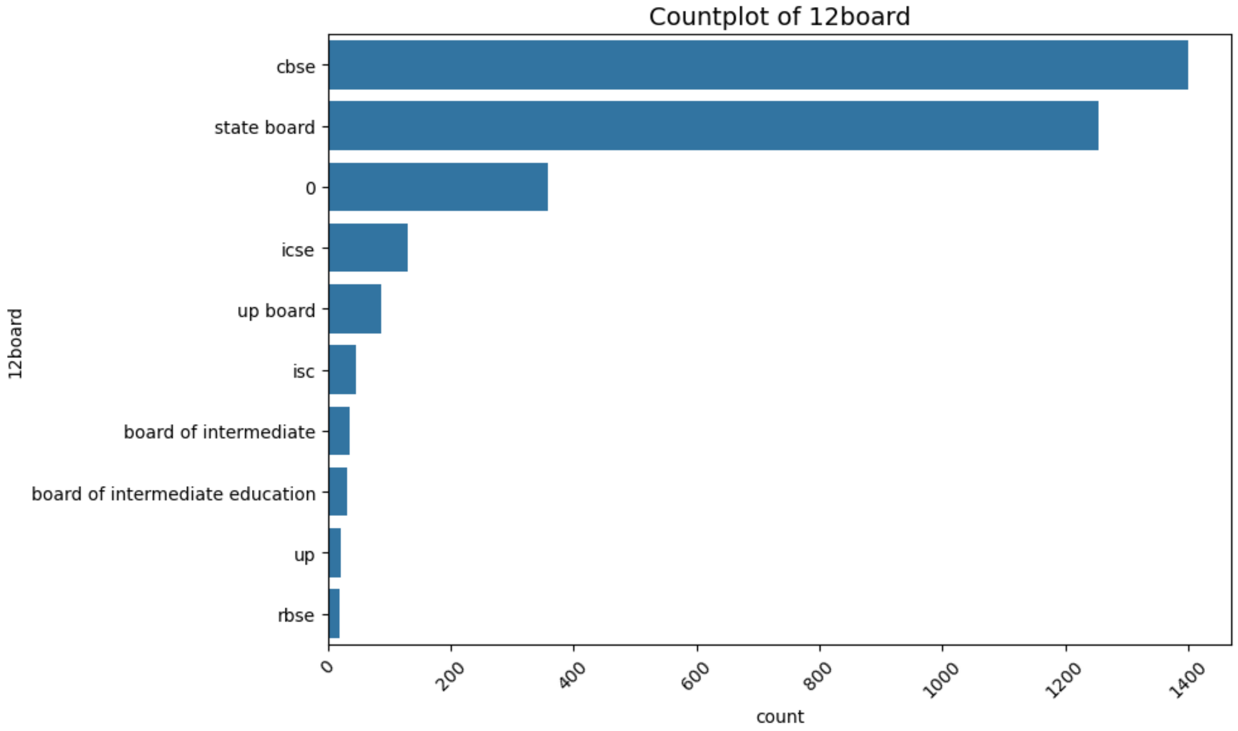

Countplot of DOL
DOL
present
2015-04-01 00:00:00
2015-03-01 00:00:00
2015-05-01 00:00:00
2015-01-01 00:00:00
2014-04-01 00:00:00
2015-02-01 00:00:00
2014-06-01 00:00:00
2014-09-01 00:00:00
2014-08-01 00:00:00
0
250
500
750
1000
1250
1500
1750
count

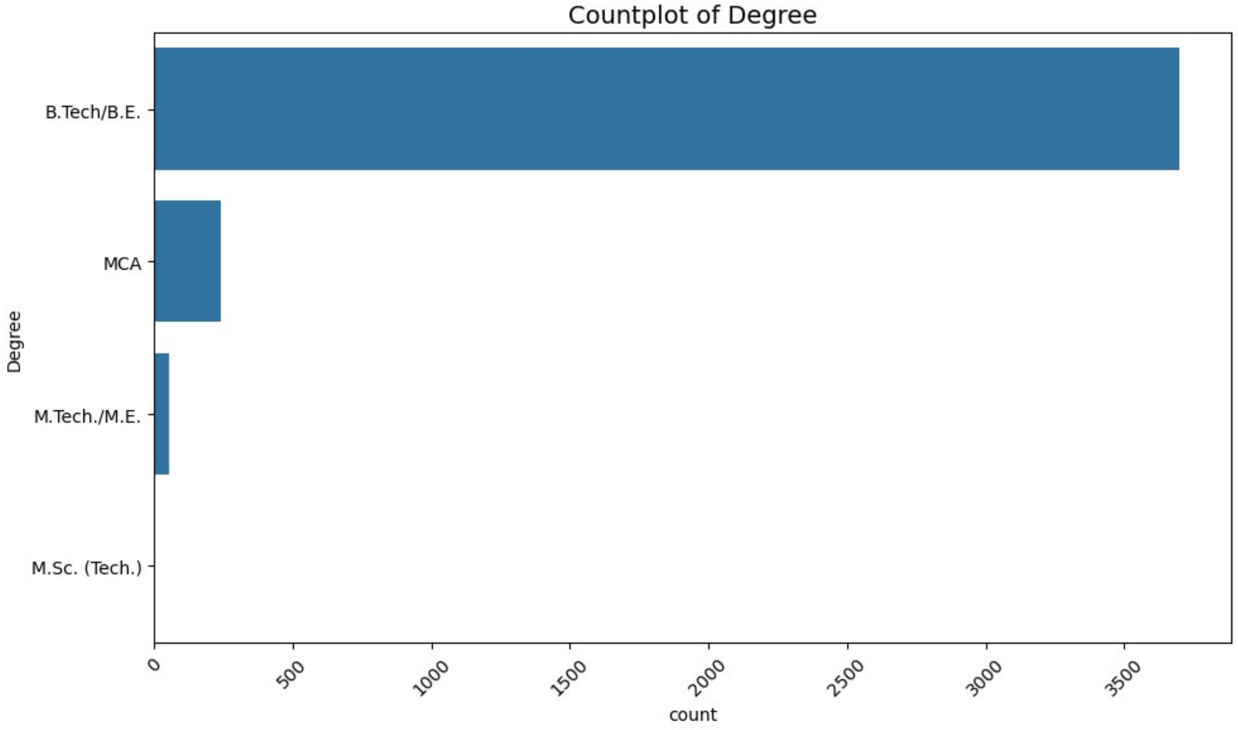

Countplot of Designation
Designation
software engineer
software developer
system engineer
programmer analyst
systems engineer
java software engineer
software test engineer
project engineer
technical support engineer
senior software engineer
0
100
200
300
400
500
count

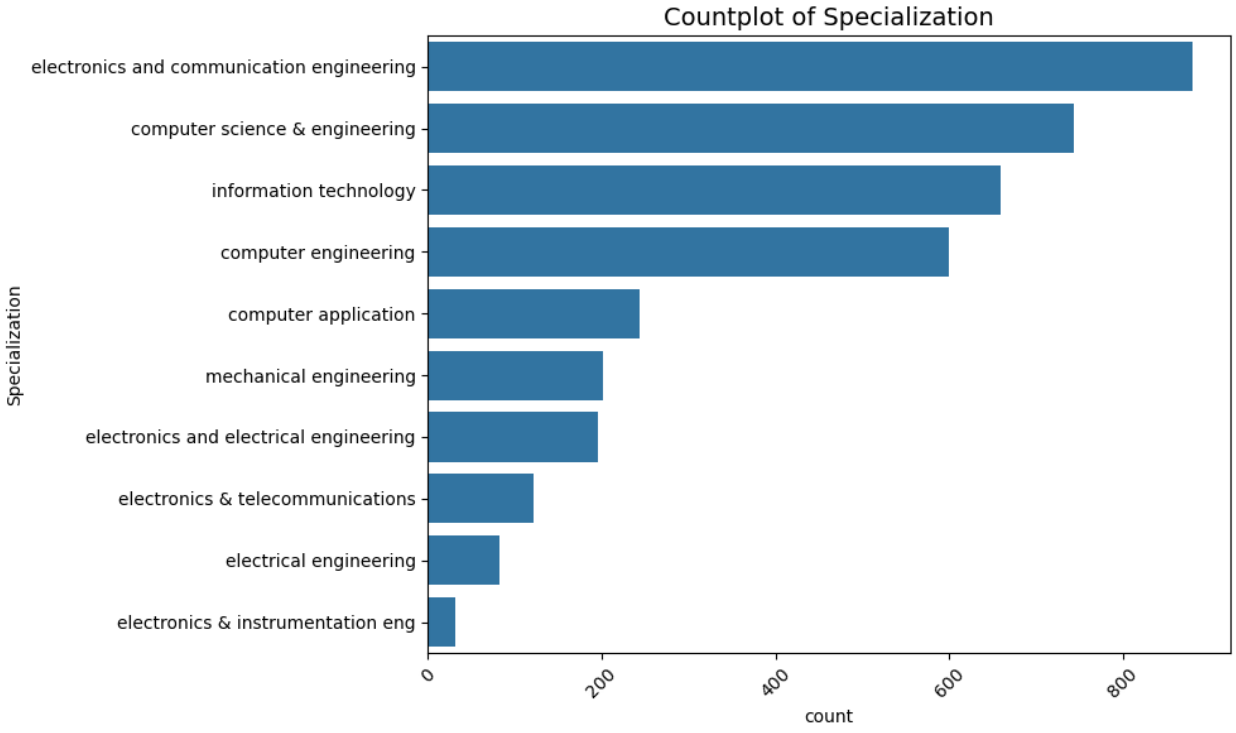

Countplot of JobCity
JobCity
Bangalore
-1
Noida
Hyderabad
Pune
Chennai
Gurgaon
New Delhi
Mumbai
Kolkata
0
100
200
300
400
500
600
count

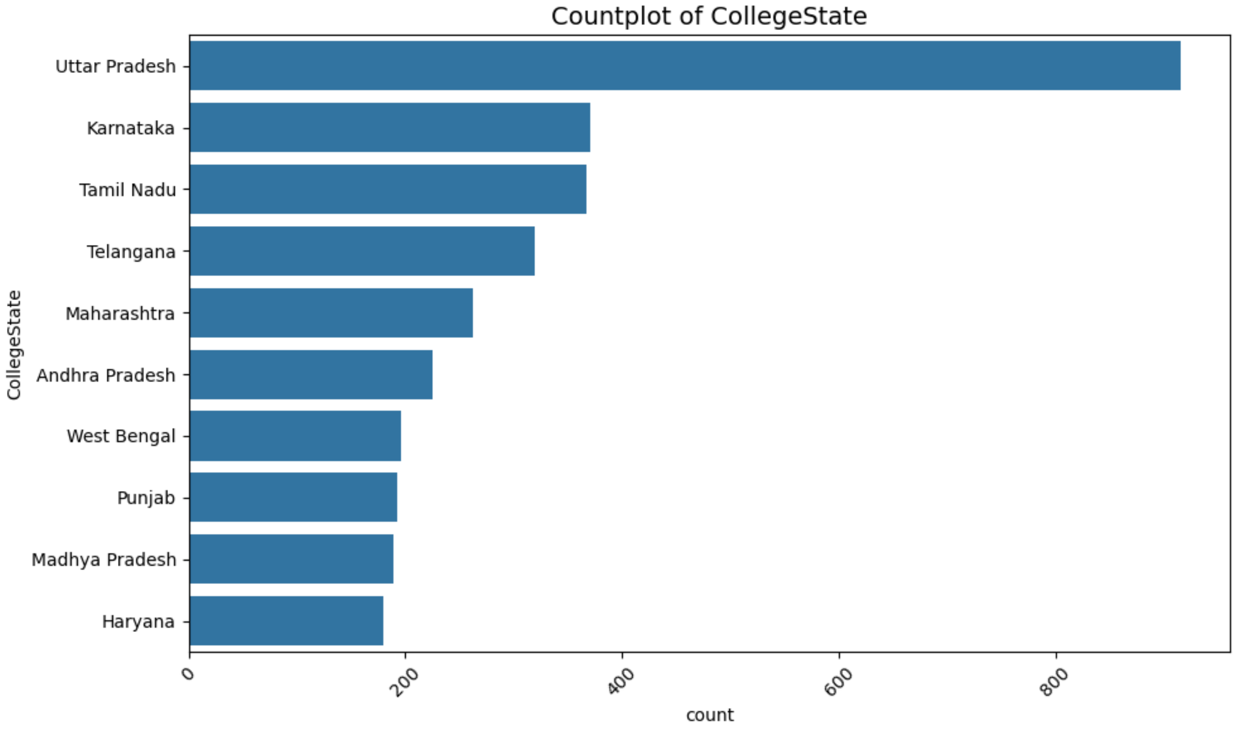

Countplot of Gender
m
f
Gender
0
500
1000
1500
2000
2500
3000
count


Countplot of 10board
cbse
state board
0
icse
ssc
up board
matriculation
rbse
board of secondary education
up
10board
0
200
400
600
800
1000
1200
1400
count

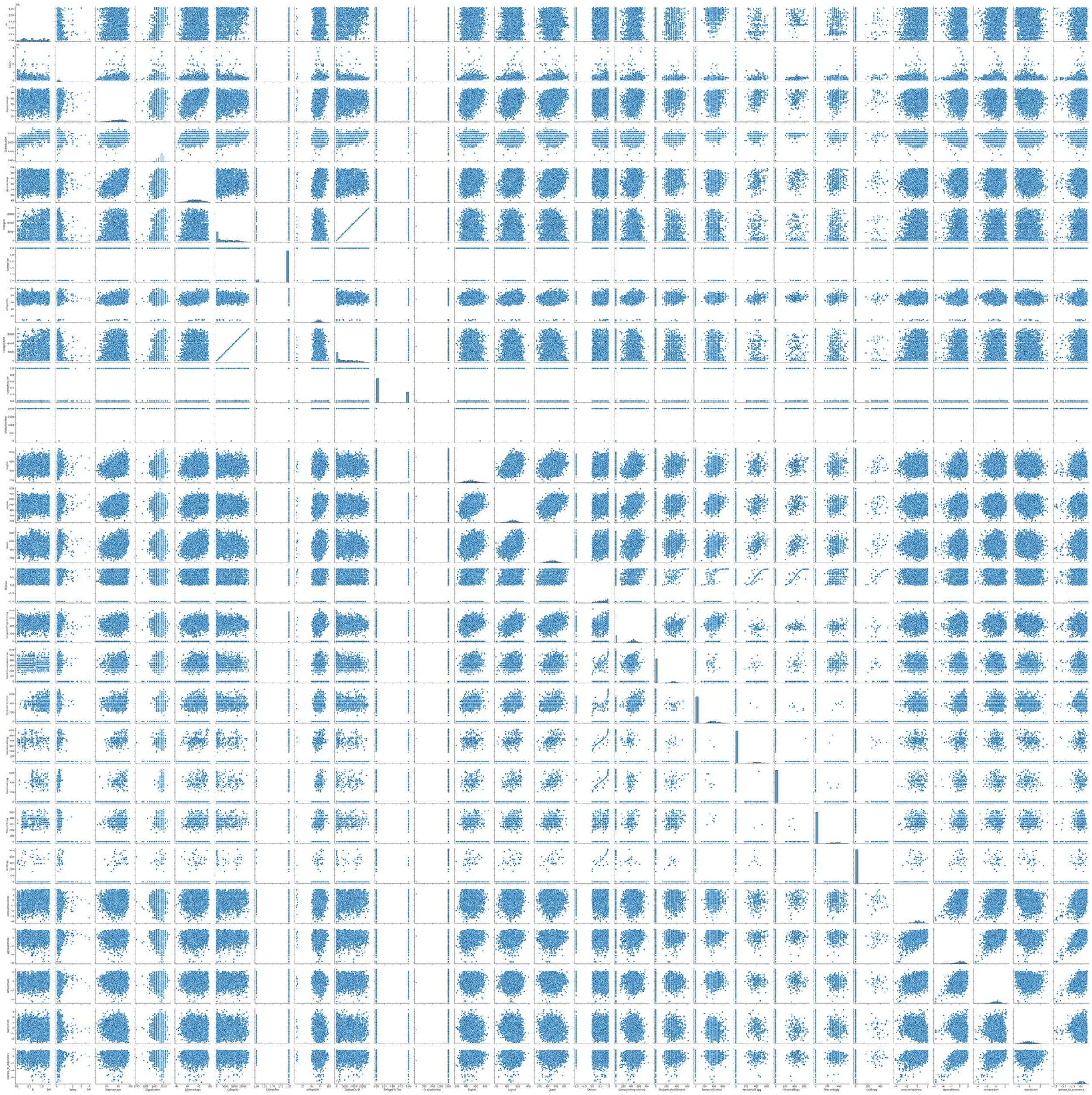

Countplot of 12board
cbse
state board
0
icse
up board
isc
board of intermediate
board of intermediate education
up
rbse
12board
0
200
400
600
800
1000
1200
1400
count


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

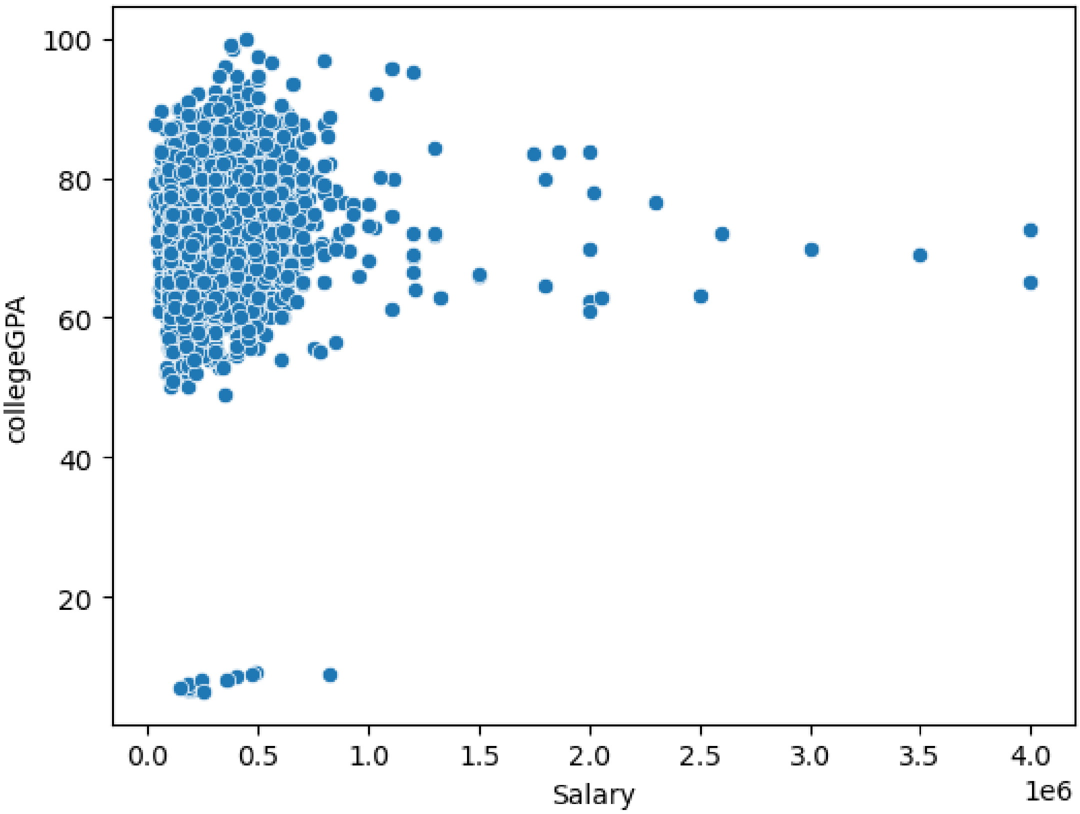




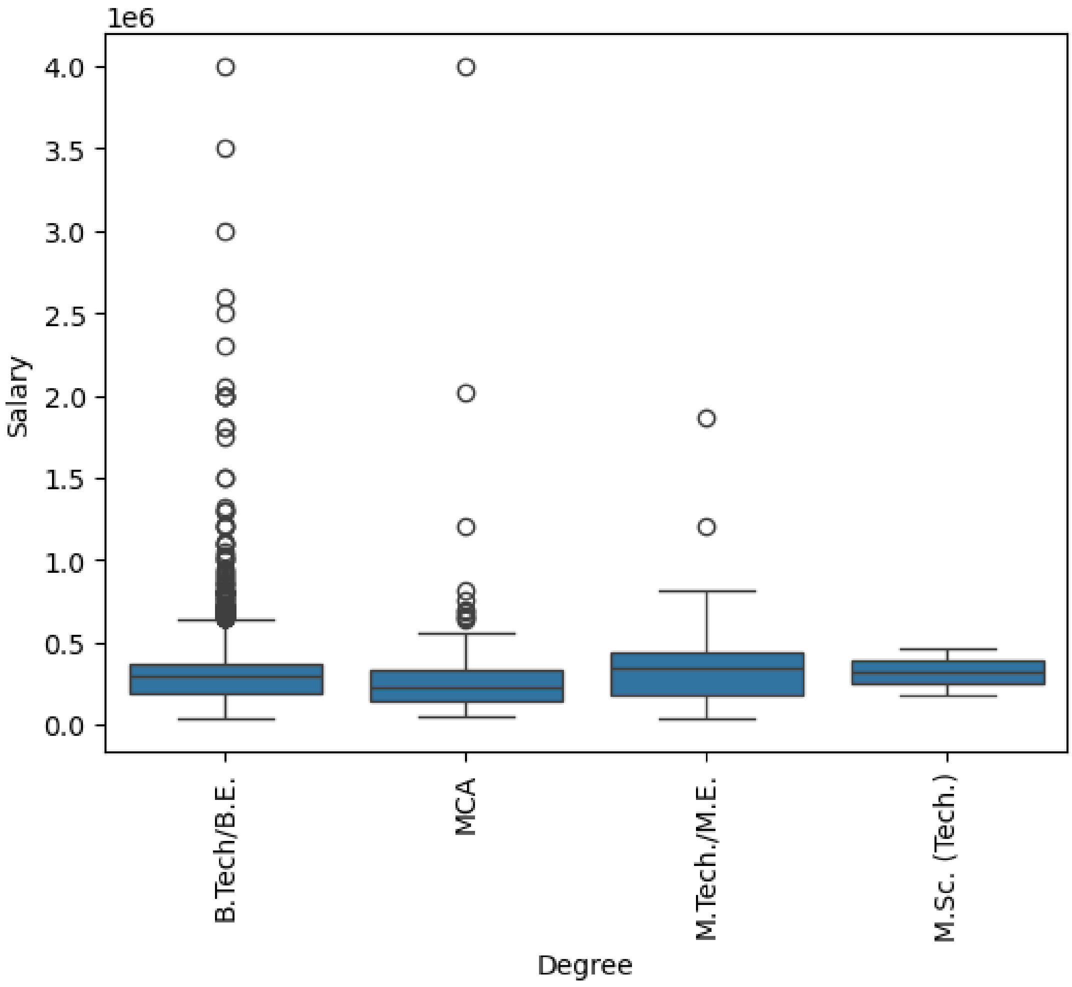
```
sns.pairplot(df[numerical_cols])
plt.show()
```



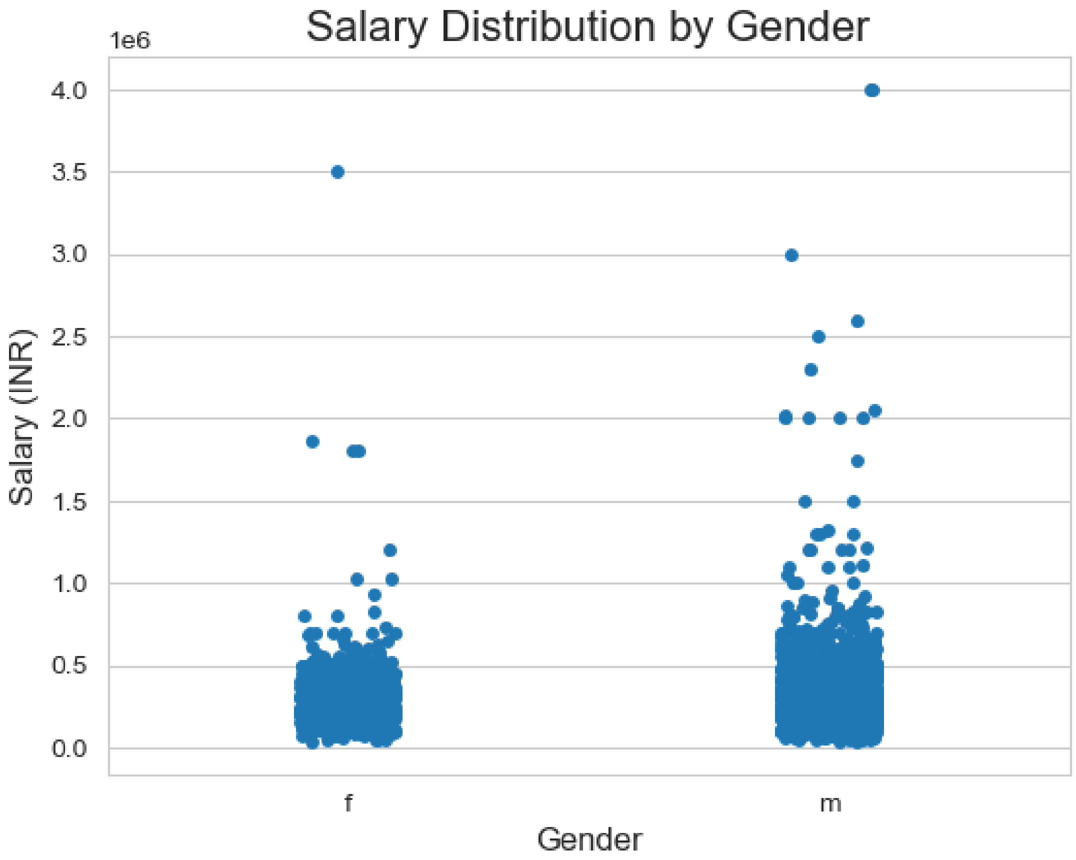
```
sns.scatterplot(x='Salary', y='collegeGPA', data=df)
plt.show()
```



```
sns.boxplot(x='Degree', y='Salary', data=df)
plt.xticks(rotation=90)
plt.show()
```

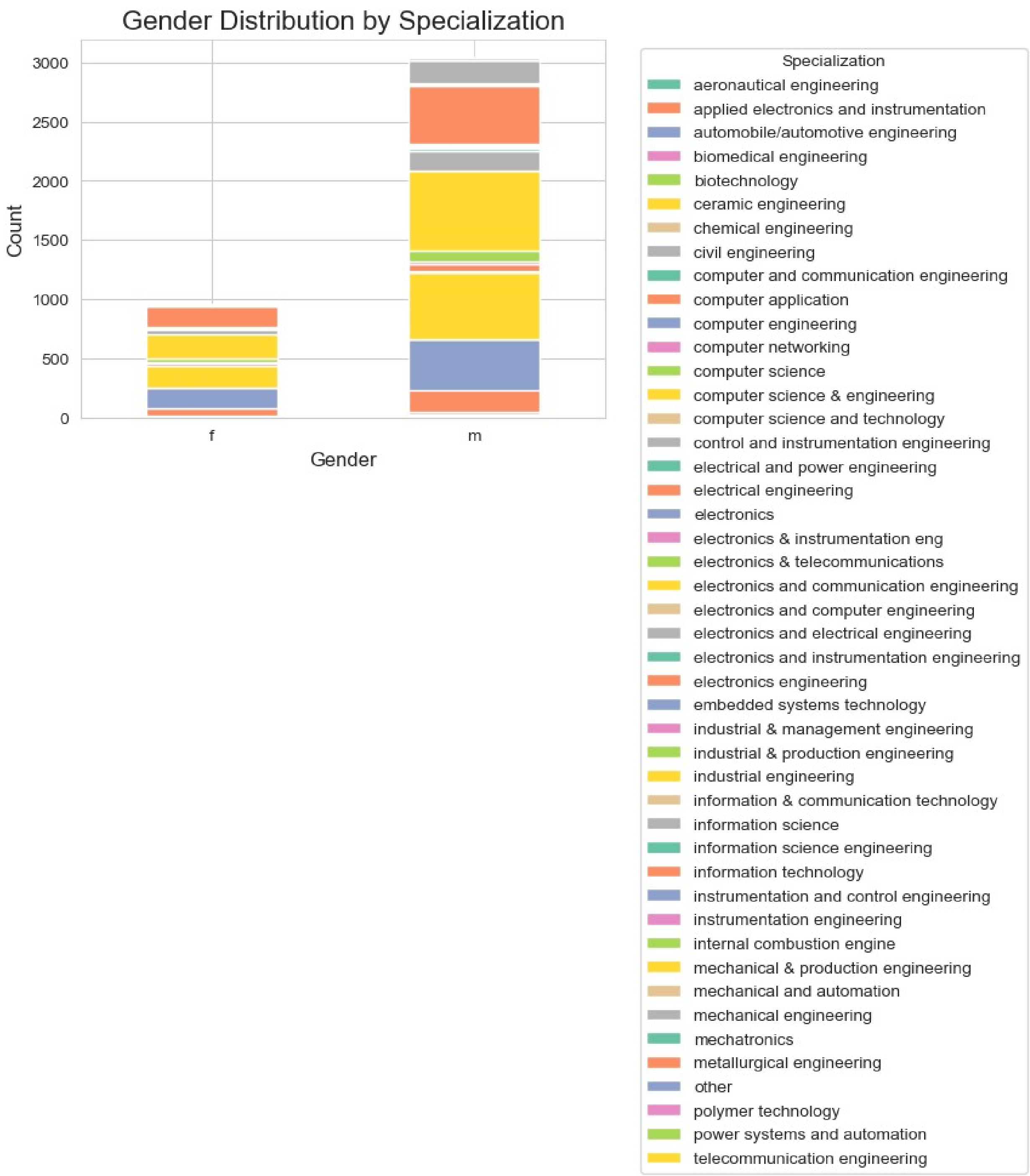


```
sns.stripplot(x='Gender', y='Salary', data=df, jitter=True)
plt.title("Salary Distribution by Gender", fontsize=16)
plt.xlabel("Gender", fontsize=12)
plt.ylabel("Salary (INR)", fontsize=12)
plt.show()
```



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

plt.figure(figsize=(12, 7))
sns.set_style("whitegrid")

cross_tab.plot(kind='bar', stacked=True,
color=sns.color_palette('Set2'))

plt.title("Gender Distribution by Specialization", fontsize=16)
plt.xlabel("Gender", fontsize=12)
plt.ylabel("Count", fontsize=12)

plt.xticks(rotation=0)

plt.legend(title="Specialization", bbox_to_anchor=(1.05, 1),
loc='upper left')

plt.subplots_adjust(bottom=0.2, top=0.85, right=0.8)

plt.show()
```

```
<Figure size 1200x700 with 0 Axes>
```



```python
from scipy import stats

mean_salary_claim = 2.75
actual_mean_salary = df['Salary'].mean()
t_stat, p_value = stats.ttest_1samp(df['Salary'], mean_salary_claim)
print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

```python
if p_value < 0.05:
    print("Reject null hypothesis: There is a significant difference between the actual mean salary and the claim.")
else:
    print("Fail to reject null hypothesis: The actual mean salary aligns with the claim.")
```

```
T-statistic: 91.45358754875822, P-value: 0.0
Reject null hypothesis: There is a significant difference between the actual mean salary and the claim.
```

```python
from scipy.stats import chi2_contingency

gender_spec_ct = pd.crosstab(df['Gender'], df['Specialization'])
chi2_stat, p_val, dof, expected = chi2_contingency(gender_spec_ct)
print(f"Chi2 Stat: {chi2_stat}, P-value: {p_val}")

if p_val < 0.05:
    print("There is a significant relationship between gender and specialization.")
else:
    print("There is no significant relationship between gender and specialization.")
```

```
Chi2 Stat: 104.46891913608455, P-value: 1.2453868176976918e-06
There is a significant relationship between gender and specialization.
```