AMCAT EDA

October 4, 2024

[1]:
```
# Importing Required Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import chi2_contingency
```

[2]:
```
# Loading Data
df = pd.read_excel(r"C:\Users\upscv\Downloads\data.xlsx")
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)
```

[3]:
```
df.head()
```

[3]:
```
  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

                          10board  12graduation  12percentage  \
0  board ofsecondary education,ap          2007          95.8
1                            cbse          2007          85.0
2                            cbse          2010          68.2
3                            cbse          2007          83.6
4                            cbse          2008          76.8

                           12board  CollegeID  CollegeTier       Degree  \
```

```
0  board of intermediate education,ap       1141            2  B.Tech/B.E.
1                                cbse       5807            2  B.Tech/B.E.
2                                cbse         64            2  B.Tech/B.E.
3                                cbse       6920            1  B.Tech/B.E.
4                                cbse      11368            2  B.Tech/B.E.

                           Specialization  collegeGPA  CollegeCityID  \
0                     computer engineering       78.00           1141
1  electronics and communication engineering       70.06           5807
2                   information technology       70.00             64
3                     computer engineering       74.64           6920
4  electronics and communication engineering       73.90          11368

   CollegeCityTier    CollegeState  GraduationYear  English  Logical  Quant  \
0                0  Andhra Pradesh            2011      515      585    525
1                0  Madhya Pradesh            2012      695      610    780
2                0   Uttar Pradesh            2014      615      545    370
3                1           Delhi            2011      635      585    625
4                0   Uttar Pradesh            2012      545      625    465

     Domain  ComputerProgramming  ElectronicsAndSemicon  ComputerScience  \
0  0.635979                  445                     -1               -1
1  0.960603                   -1                    466               -1
2  0.450877                  395                     -1               -1
3  0.974396                  615                     -1               -1
4  0.124502                   -1                    233               -1

   MechanicalEngg  ElectricalEngg  TelecomEngg  CivilEngg  conscientiousness  \
0              -1              -1           -1         -1             0.9737
1              -1              -1           -1         -1            -0.7335
2              -1              -1           -1         -1             0.2718
3              -1              -1           -1         -1             0.0464
4              -1              -1           -1         -1            -0.8810

   agreeableness  extraversion  nueroticism  openess_to_experience
0         0.8128        0.5269      1.35490                -0.4455
1         0.3789        1.2396     -0.10760                 0.8637
2         1.7109        0.1637     -0.86820                 0.6721
3         0.3448       -0.3440     -0.40780                -0.9194
4        -0.2793       -1.0697      0.09163                -0.1295
```

[4]:
```
# EDA
```

[5]:
```
df.shape
```

[5]: (3998, 39)

[6]:
```
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   int64
 3   DOJ                    3998 non-null   datetime64[ns]
 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   datetime64[ns]
 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: datetime64[ns](2), float64(9), int64(18), object(10)
memory usage: 1.2+ MB
```

[7]:
```
df.describe()
```

[7]:

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

	CollegeCityTier	GraduationYear	English	Logical	Quant
count	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000
mean	0.300400	2012.105803	501.649075	501.598799	513.378189
min	0.000000	0.000000	180.000000	195.000000	120.000000
25%	0.000000	2012.000000	425.000000	445.000000	430.000000
50%	0.000000	2013.000000	500.000000	505.000000	515.000000
75%	1.000000	2014.000000	570.000000	565.000000	595.000000
max	1.000000	2017.000000	875.000000	795.000000	900.000000
std	0.458489	31.857271	104.940021	86.783297	122.302332

	Domain	ComputerProgramming	ElectronicsAndSemicon
count	3998.000000	3998.000000	3998.000000
mean	0.510490	353.102801	95.328414
min	-1.000000	-1.000000	-1.000000
25%	0.342315	295.000000	-1.000000

```
50%       0.622643           415.000000             -1.000000
75%       0.842248           495.000000            233.000000
max       0.999910           840.000000            612.000000
std       0.468671           205.355519            158.241218

       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]:
```
# checking null values
```

[9]:
```
df.isna().sum()
```

[9]:
```
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
```

[10]:
```
df.duplicated().sum()
```

[10]:
```
0
```

0.0.1 Univariate Analysis

[11]:
```
df.drop('Unnamed: 0',inplace=True,axis=1)
```

[12]:
```
df.head()
```

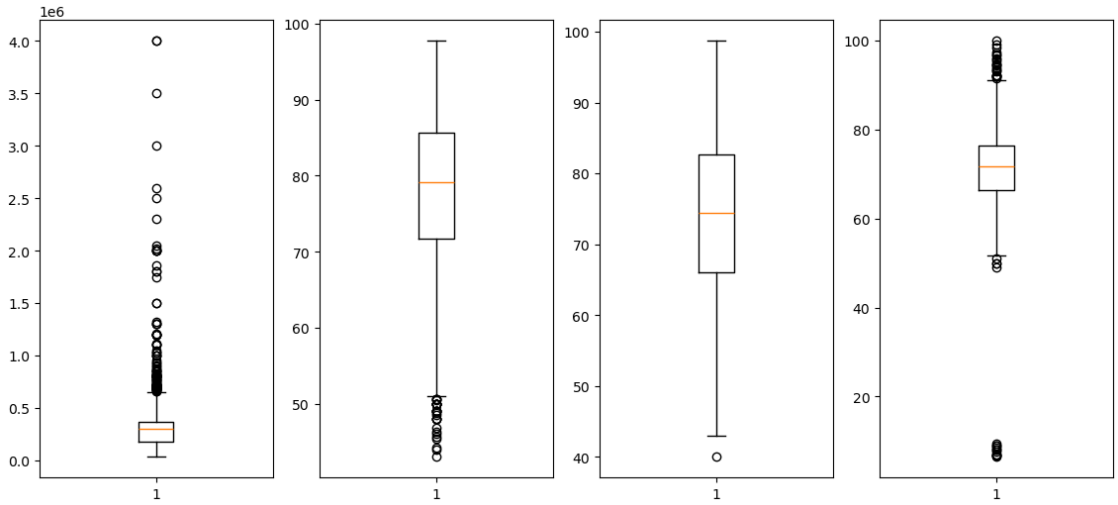
[12]:

	ID	Salary	DOJ	DOL	Designation	\
0	203097	420000	2012-06-01	present	senior quality engineer	
1	579905	500000	2013-09-01	present	assistant manager	
2	810601	325000	2014-06-01	present	systems engineer	

```
3  267447  1100000 2011-07-01              present  senior software engineer
4  343523   200000 2014-03-01  2015-03-01 00:00:00                       get

     JobCity Gender        DOB  10percentage                          10board  \
0  Bangalore      f 1990-02-19          84.3  board ofsecondary education,ap
1     Indore      m 1989-10-04          85.4                              cbse
2    Chennai      f 1992-08-03          85.0                              cbse
3    Gurgaon      m 1989-12-05          85.6                              cbse
4    Manesar      m 1991-02-27          78.0                              cbse

   12graduation  12percentage                              12board  CollegeID  \
0          2007          95.8  board of intermediate education,ap       1141
1          2007          85.0                                cbse       5807
2          2010          68.2                                cbse         64
3          2007          83.6                                cbse       6920
4          2008          76.8                                cbse      11368

   CollegeTier       Degree                              Specialization  \
0            2  B.Tech/B.E.                        computer engineering
1            2  B.Tech/B.E.  electronics and communication engineering
2            2  B.Tech/B.E.                      information technology
3            1  B.Tech/B.E.                        computer engineering
4            2  B.Tech/B.E.  electronics and communication engineering

   collegeGPA  CollegeCityID  CollegeCityTier    CollegeState  GraduationYear  \
0       78.00           1141                0  Andhra Pradesh            2011
1       70.06           5807                0  Madhya Pradesh            2012
2       70.00             64                0   Uttar Pradesh            2014
3       74.64           6920                1           Delhi            2011
4       73.90          11368                0   Uttar Pradesh            2012

   English  Logical  Quant    Domain  ComputerProgramming  \
0      515      585    525  0.635979                  445
1      695      610    780  0.960603                   -1
2      615      545    370  0.450877                  395
3      635      585    625  0.974396                  615
4      545      625    465  0.124502                   -1

   ElectronicsAndSemicon  ComputerScience  MechanicalEngg  ElectricalEngg  \
0                     -1               -1              -1              -1
1                    466               -1              -1              -1
2                     -1               -1              -1              -1
3                     -1               -1              -1              -1
4                    233               -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

   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
```

[13]:
```
fig,(x1,x2,x3,x4) = plt.subplots(1,4,figsize=(14,6))

# salary
x1.set_label('salary')
x1.boxplot(df['Salary'])

# X class percentage
x2.set_label('X class percentage')
x2.boxplot(df['10percentage'])

# XII class Percentage
x3.set_label(' XII class Percentage')
x3.boxplot(df['12percentage'])

# B-Tech CGPA
x4.set_label('B-Tech CGPA')
x4.boxplot(df['collegeGPA'])

plt.show()
```

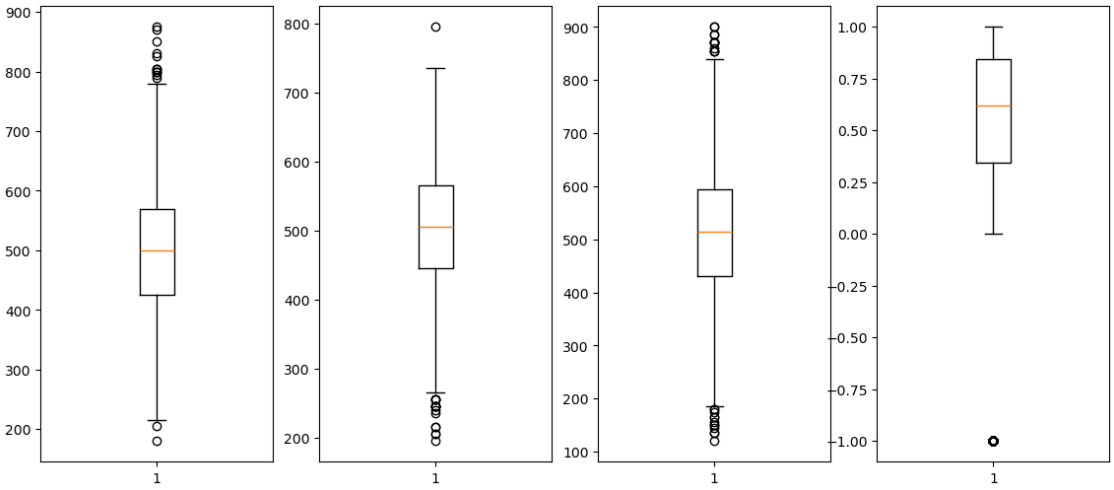


[14]:
```
df.dtypes
```

[14]:
```
ID                              int64
Salary                          int64
DOJ                    datetime64[ns]
DOL                            object
Designation                    object
JobCity                        object
Gender                         object
DOB                    datetime64[ns]
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
```

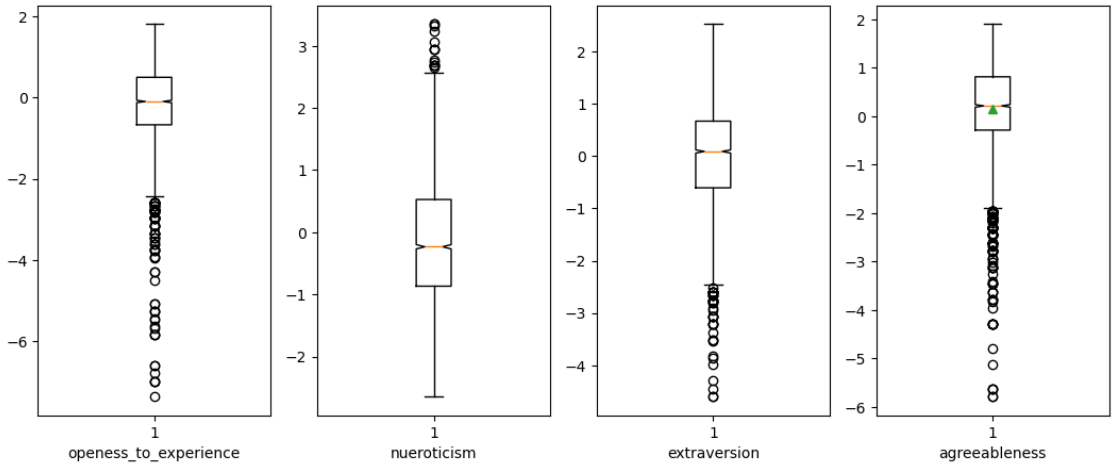
[15]:
```
fig,(x5,x6,x7,x8) = plt.subplots(1,4,figsize=(14,6))

# English
x5.set_label('English')
```

```
x5.boxplot(df['English'])

# Logical
x6.set_label('Logical')
x6.boxplot(df['Logical'])

# Quant
x7.set_label('Quant')
x7.boxplot(df['Quant'])

# Domain
x8.set_label('Domain')
x8.boxplot(df['Domain'])

plt.show()
```

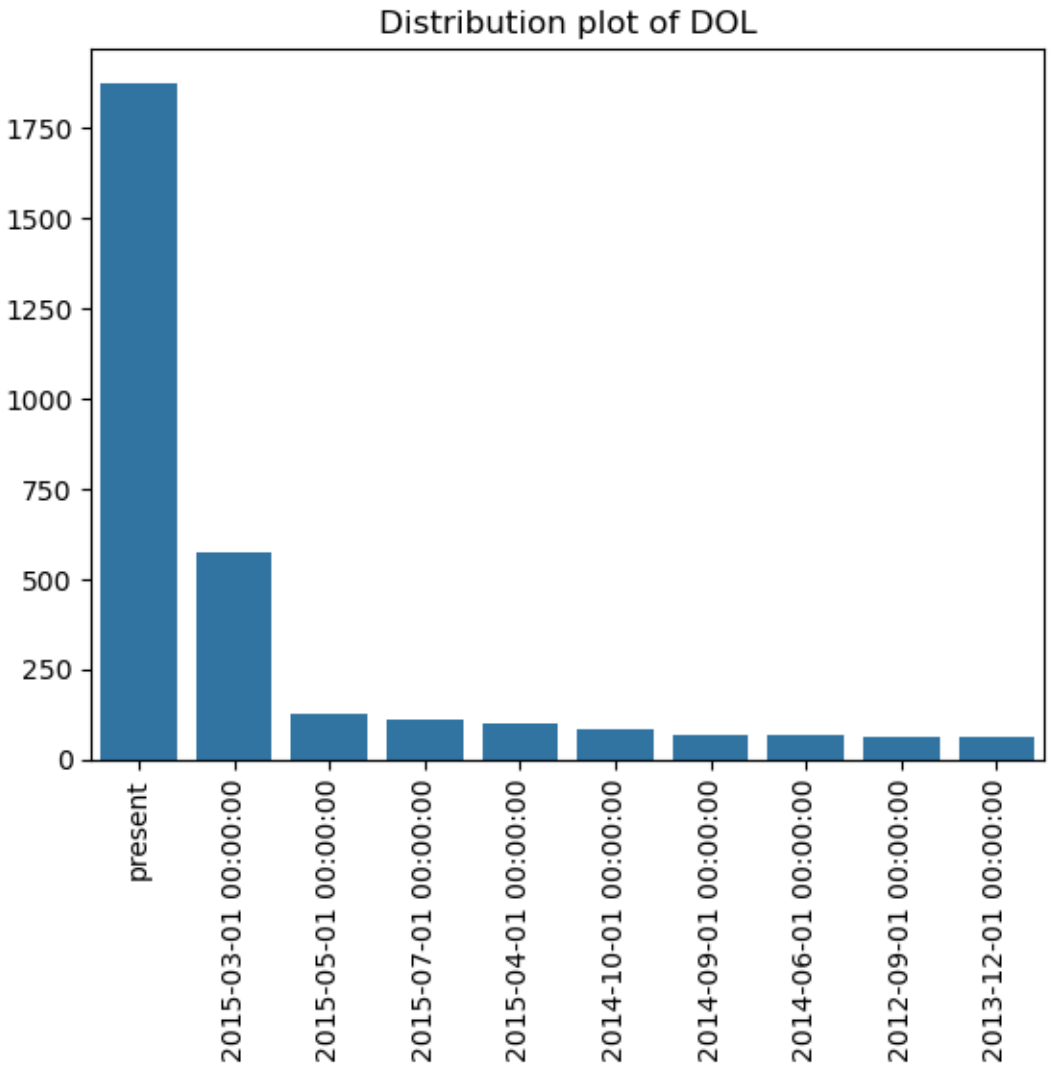


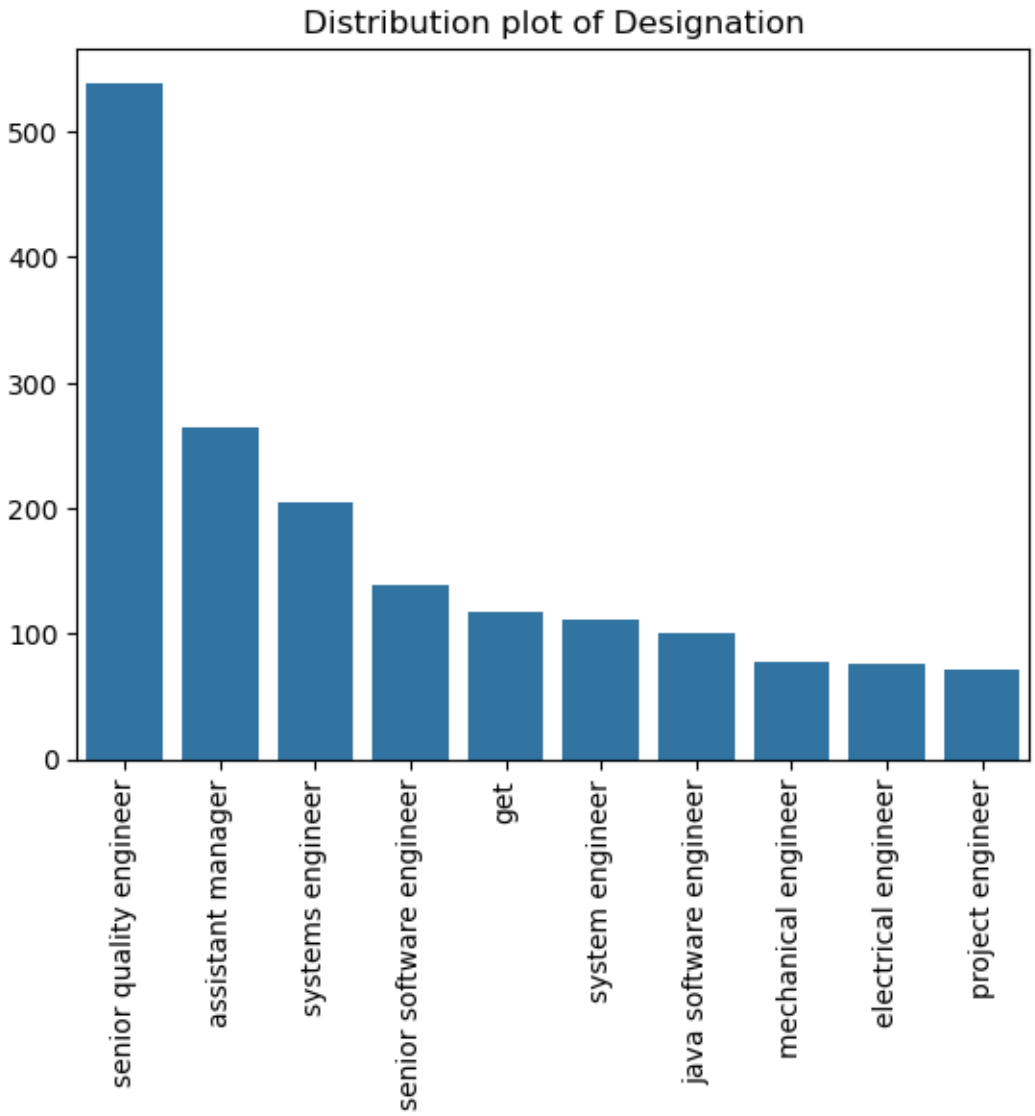
[16]:
```
fig,(x9,x10,x11,x12) = plt.subplots(1,4,figsize=(13,5))

x9.set_xlabel("openess_to_experience")
x9.boxplot(df["openess_to_experience"],notch=True)


x10.set_xlabel("nueroticism")
x10.boxplot(df["nueroticism"],notch=True)


x11.set_xlabel("extraversion")
x11.boxplot(df["extraversion"],notch=True)
```

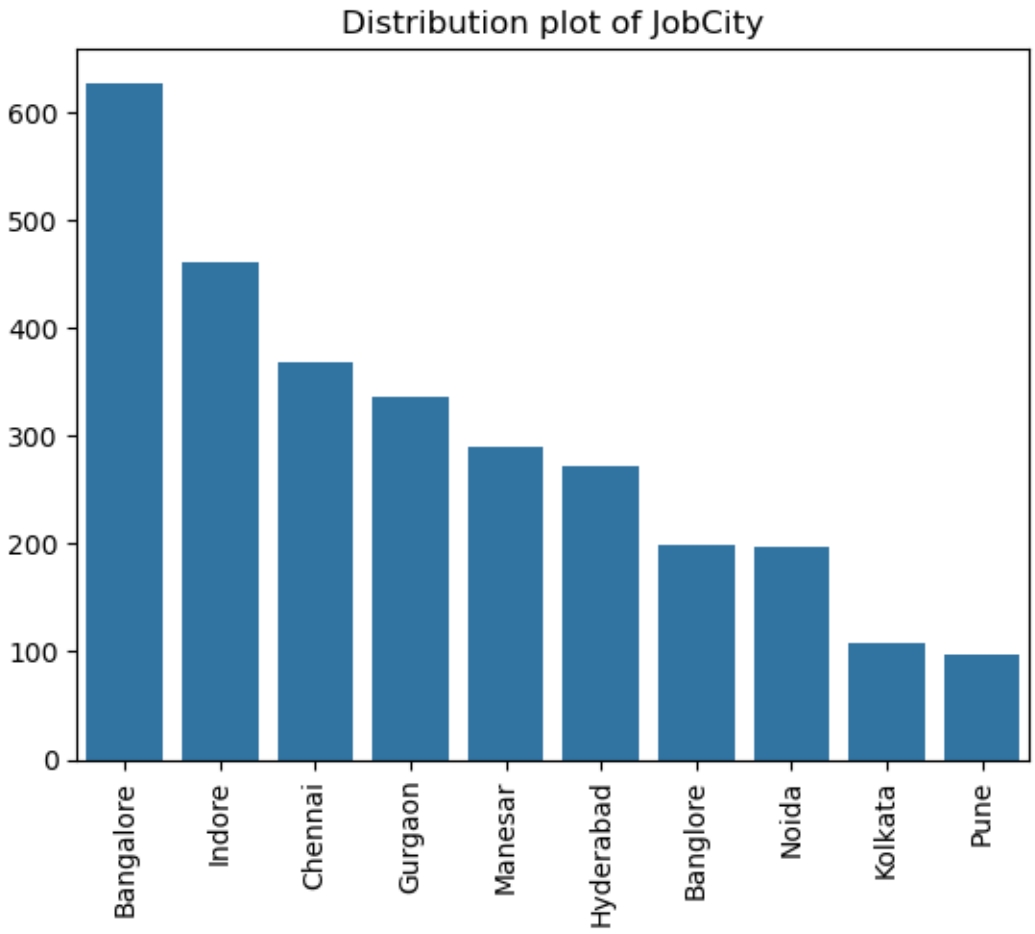
```
x12.set_xlabel("agreeableness")
x12.boxplot(df["agreeableness"],notch=True,showmeans=True)

plt.show()
```

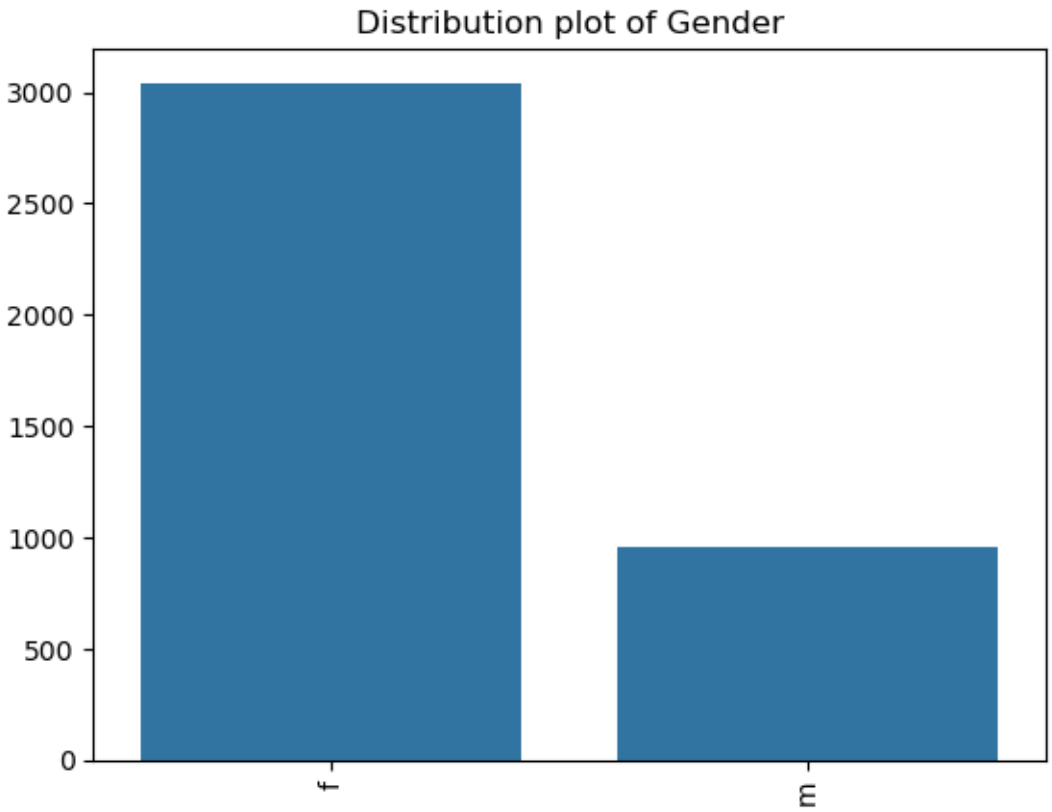


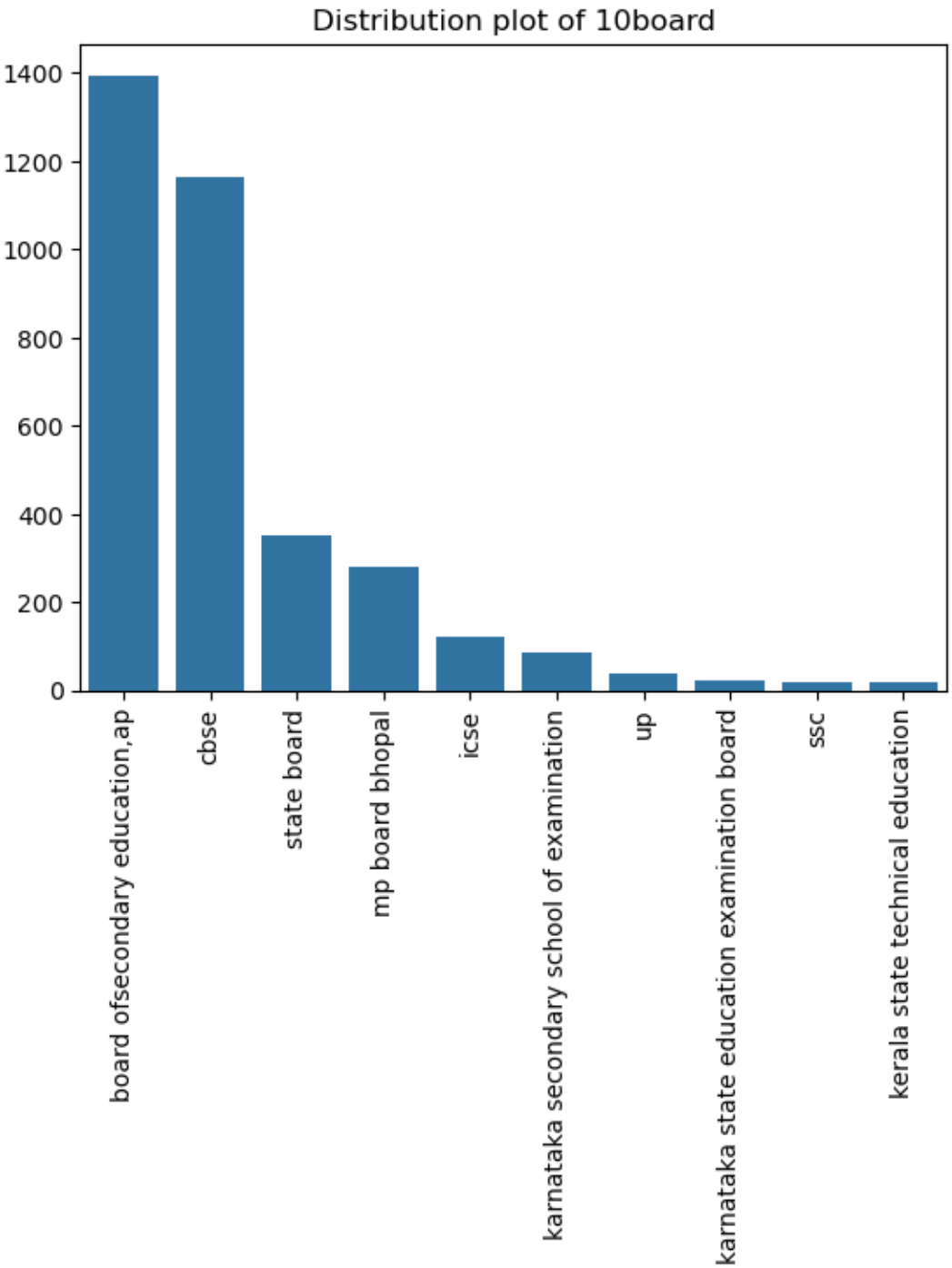
[17]:
```
for i in df.columns:

    if df[i].dtype == "object" :

        plt.title(f"Distribution plot of {i}")
        sns.barplot(x=df[i].unique()[:10],y=np.array(df[i].value_counts()[:
↪10]),)
        plt.xticks(rotation=90)
        plt.show()
```

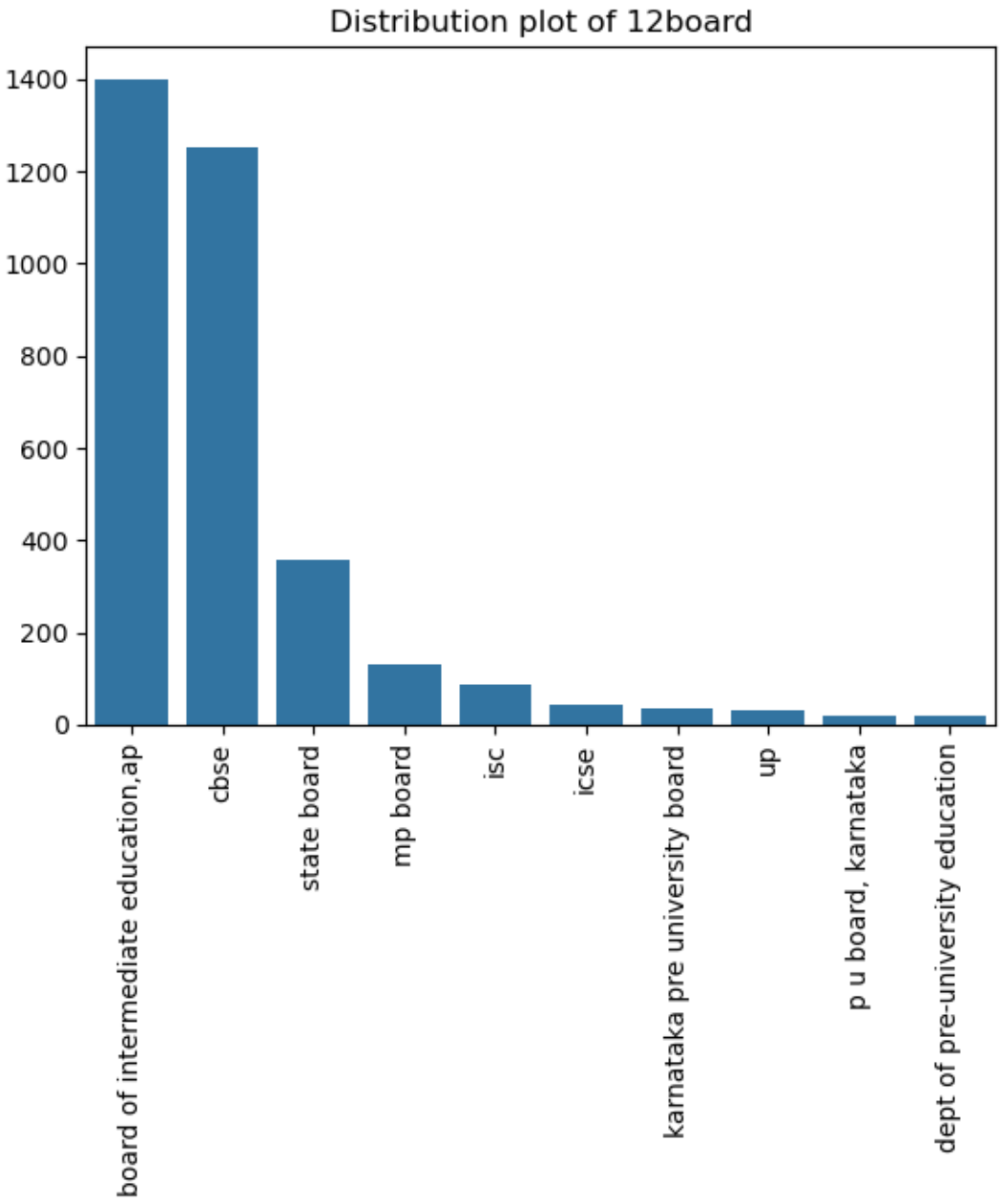

Distribution plot of DOL
1750
1500
1250
1000
750
500
250
0
present
2015-03-01 00:00:00
2015-05-01 00:00:00
2015-07-01 00:00:00
2015-04-01 00:00:00
2014-10-01 00:00:00
2014-09-01 00:00:00
2014-06-01 00:00:00
2012-09-01 00:00:00
2013-12-01 00:00:00

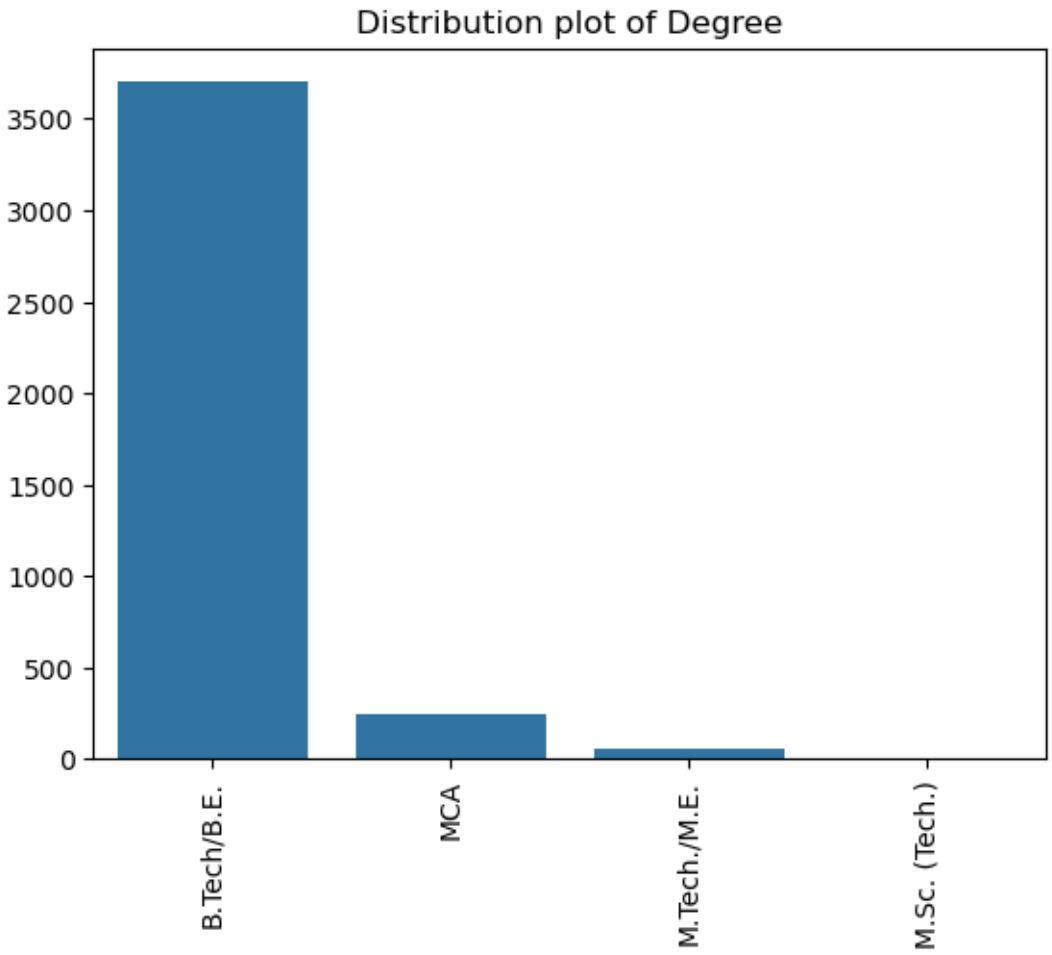

Distribution plot of Designation
500
400
300
200
100
0
senior quality engineer
assistant manager
systems engineer
senior software engineer
get
system engineer
java software engineer
mechanical engineer
electrical engineer
project engineer

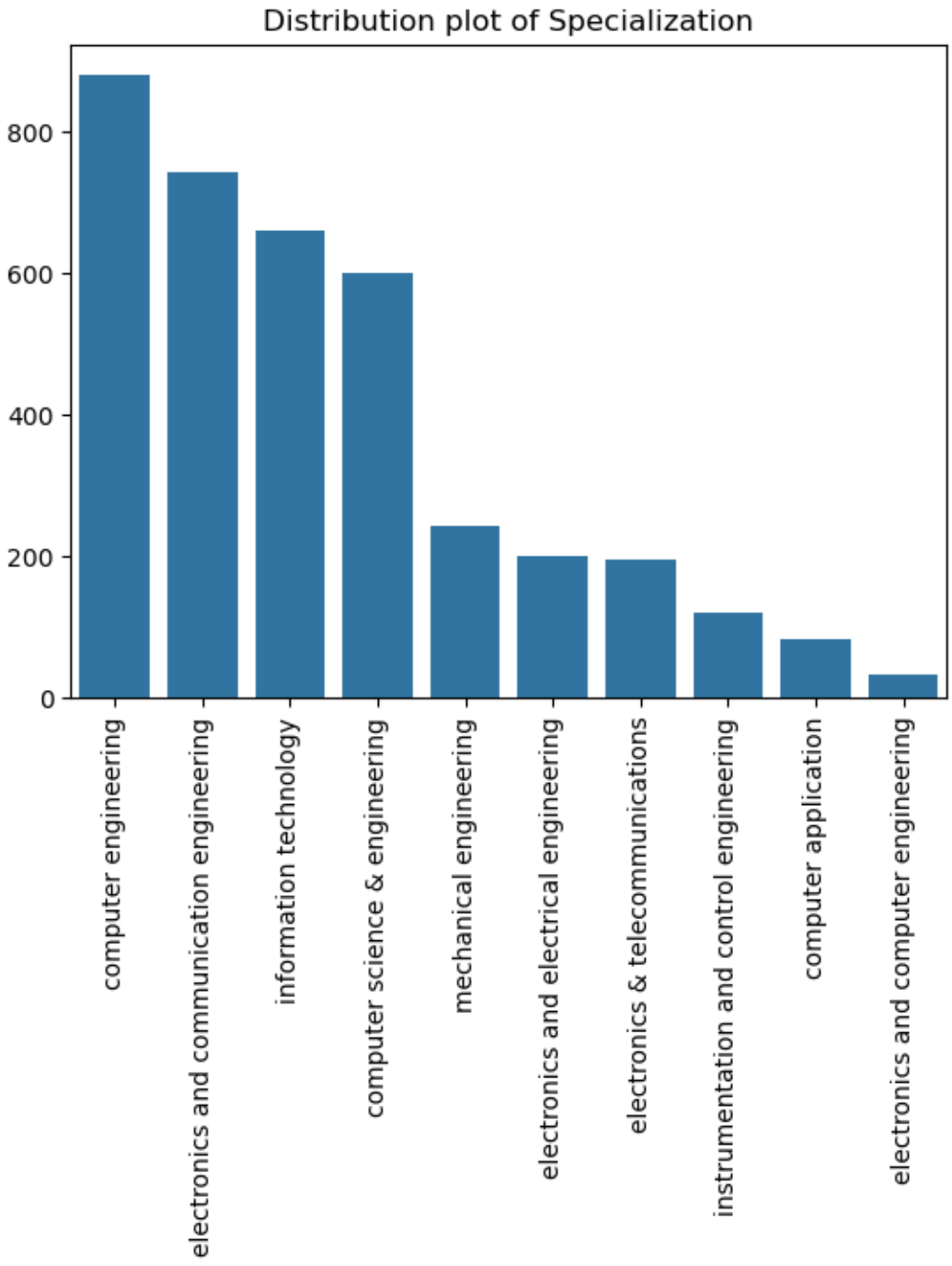

Distribution plot of JobCity
600
500
400
300
200
100
0
Bangalore
Indore
Chennai
Gurgaon
Manesar
Hyderabad
Banglore
Noida
Kolkata
Pune

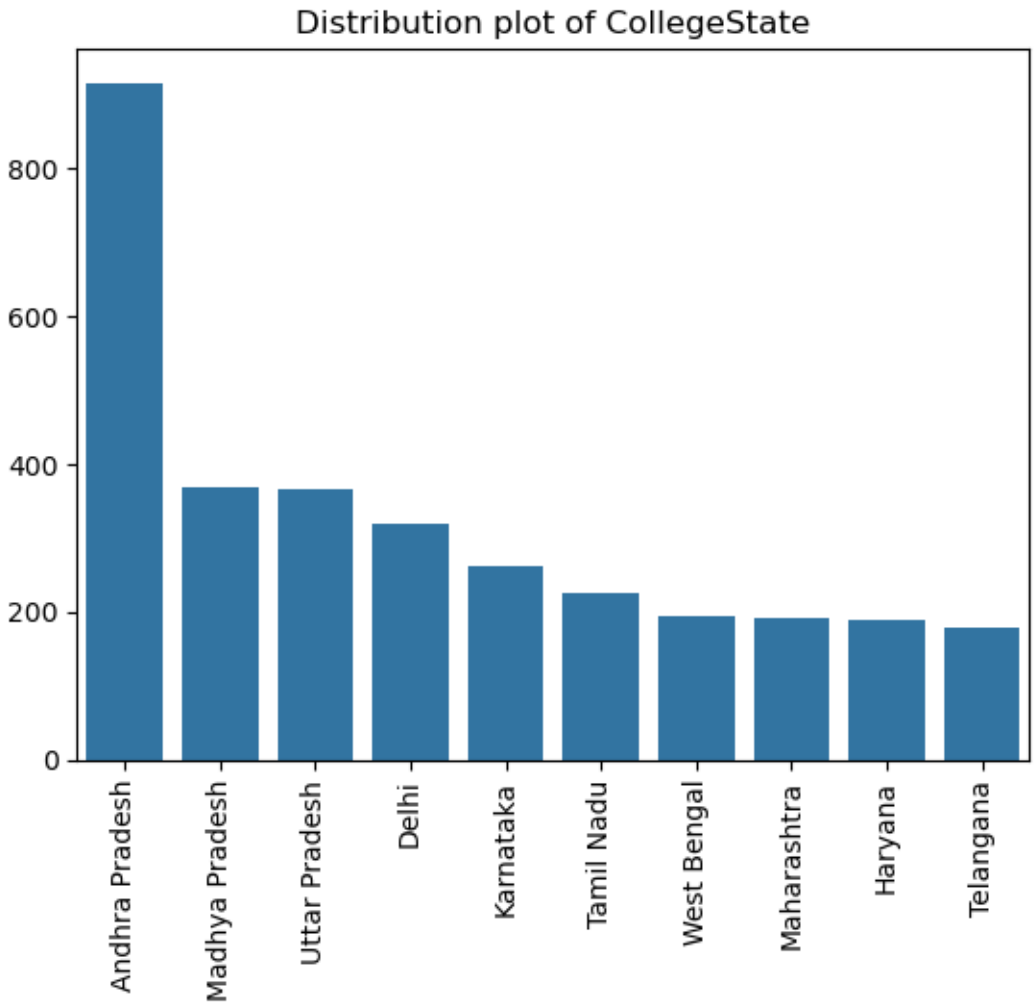

Distribution plot of Gender
3000
2500
2000
1500
1000
500
0
f
m


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


Distribution plot of 12board
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 plot of Degree
3500
3000
2500
2000
1500
1000
500
0
B.Tech/B.E.
MCA
M.Tech./M.E.
M.Sc. (Tech.)

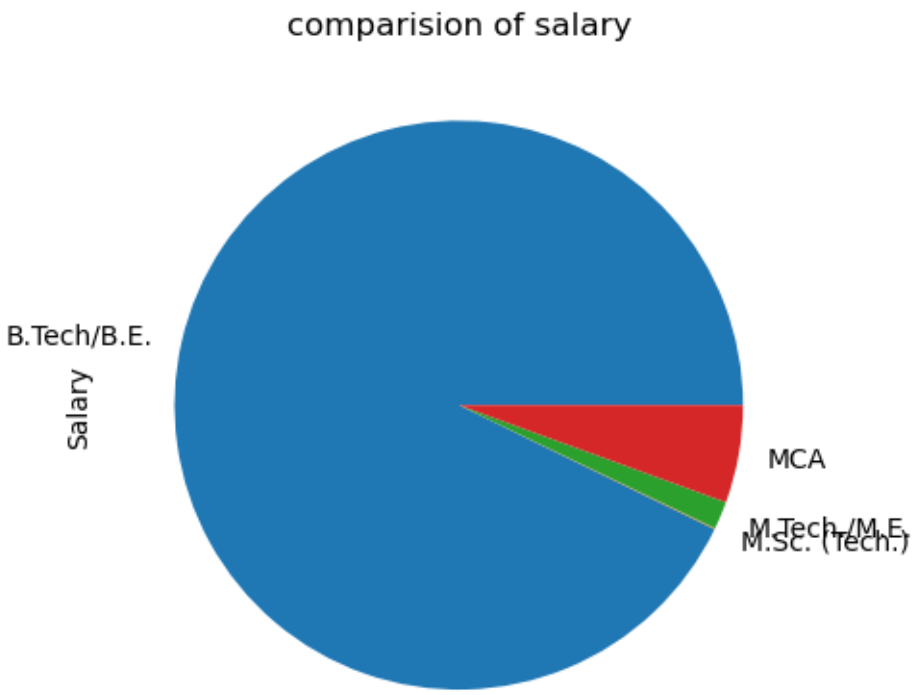

Distribution plot of Specialization
800
600
400
200
0
computer engineering
electronics and communication engineering
information technology
computer science & engineering
mechanical engineering
electronics and electrical engineering
electronics & telecommunications
instrumentation and control engineering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



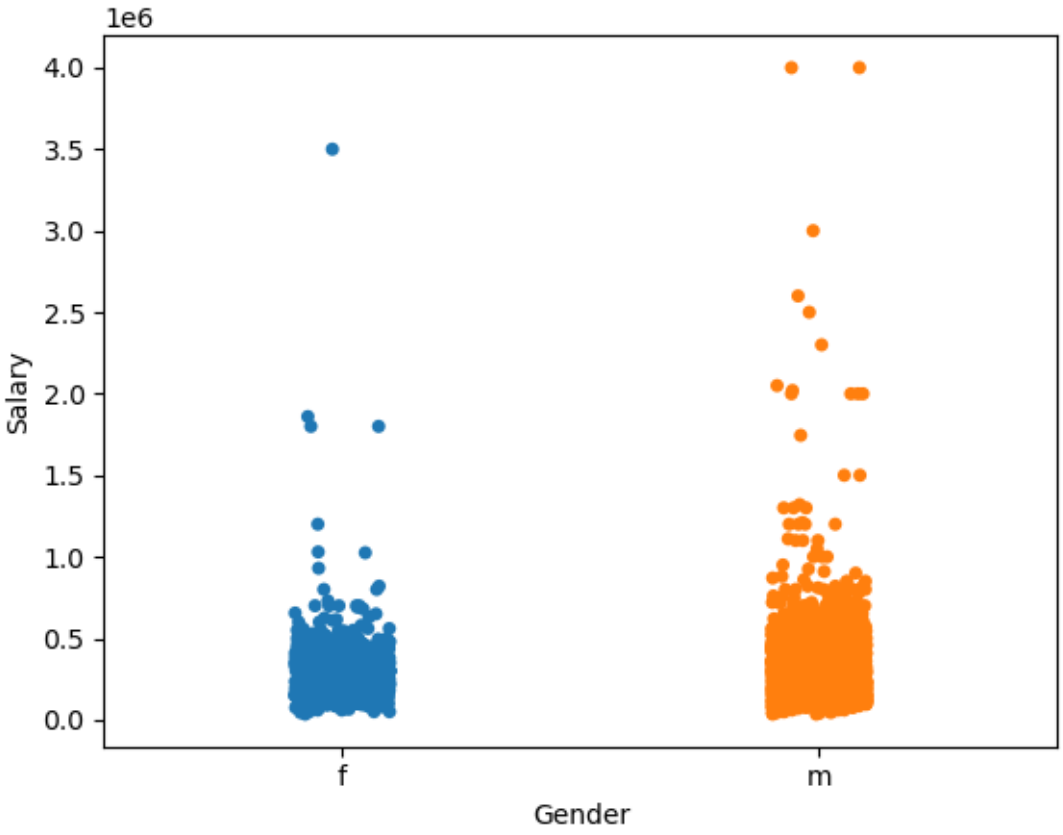
[ ]:

0.0.2 Bivariate Ananlysis

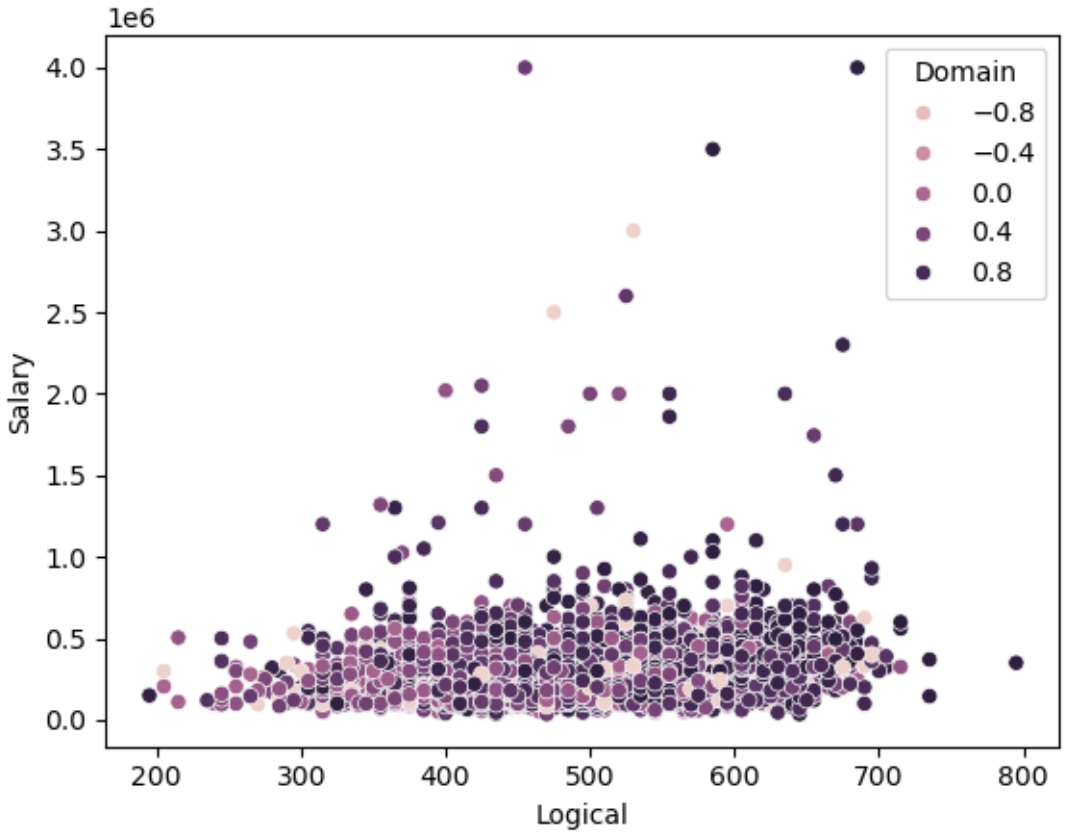
[18]:
```
deg_sal = df.groupby("Degree")[["Salary"]].sum()
deg_sal.plot(kind="pie",y="Salary",legend=False,title="comparision of salary",)

plt.show()
```



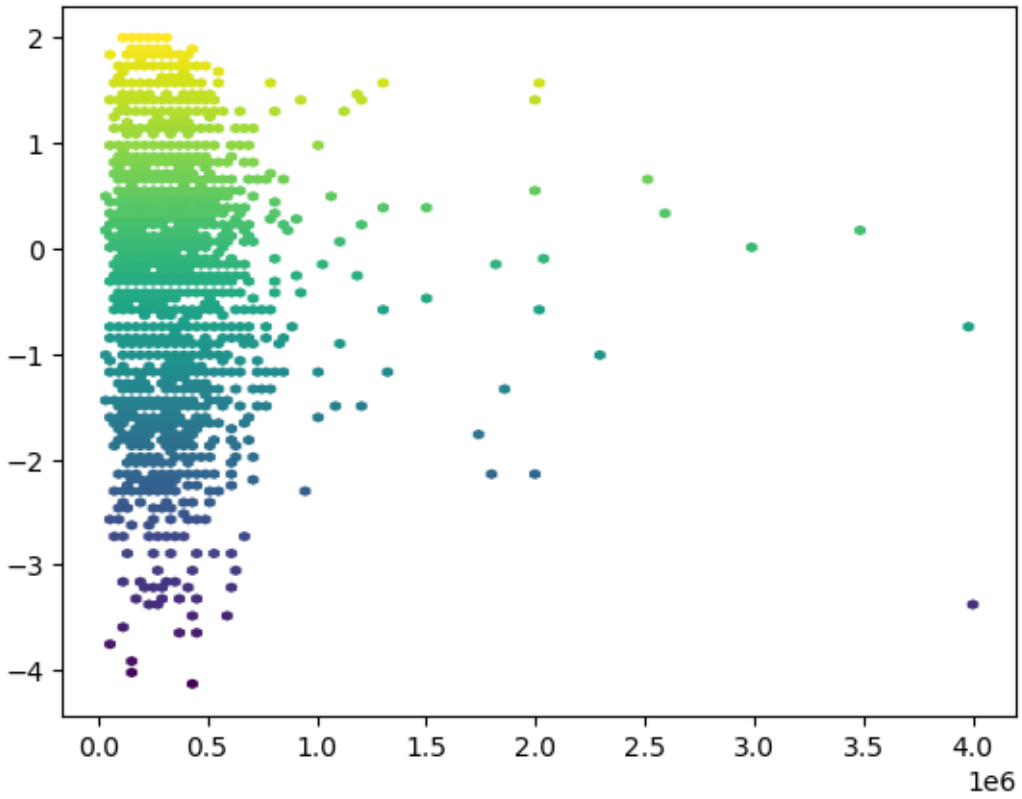
[19]:
```
# relationship between gender and salary
sns.stripplot(x=df.Gender,y=df.Salary,hue=df.Gender)
plt.show()
```



[20]:
```
sns.scatterplot(x=df.Logical,y = df.Salary,hue=df.Domain)
plt.show()
```



[21]:
```
plt.hexbin(x=df.Salary,y=df.conscientiousness,C=df.conscientiousness)
plt.show()
```



[22]:
```
df.info()
```

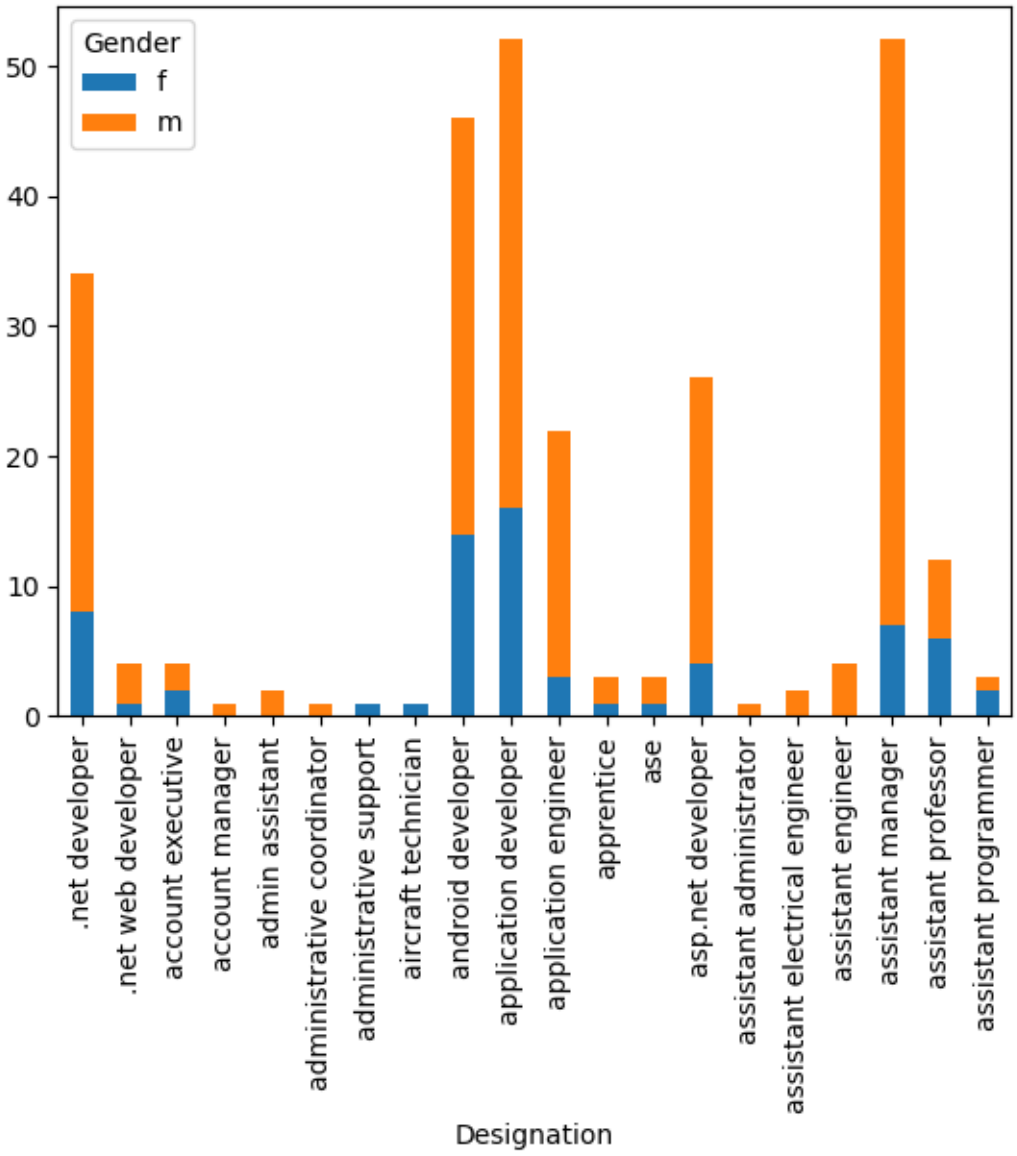
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 38 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     3998 non-null   int64
 1   Salary                 3998 non-null   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
```

1 categorical - categorical

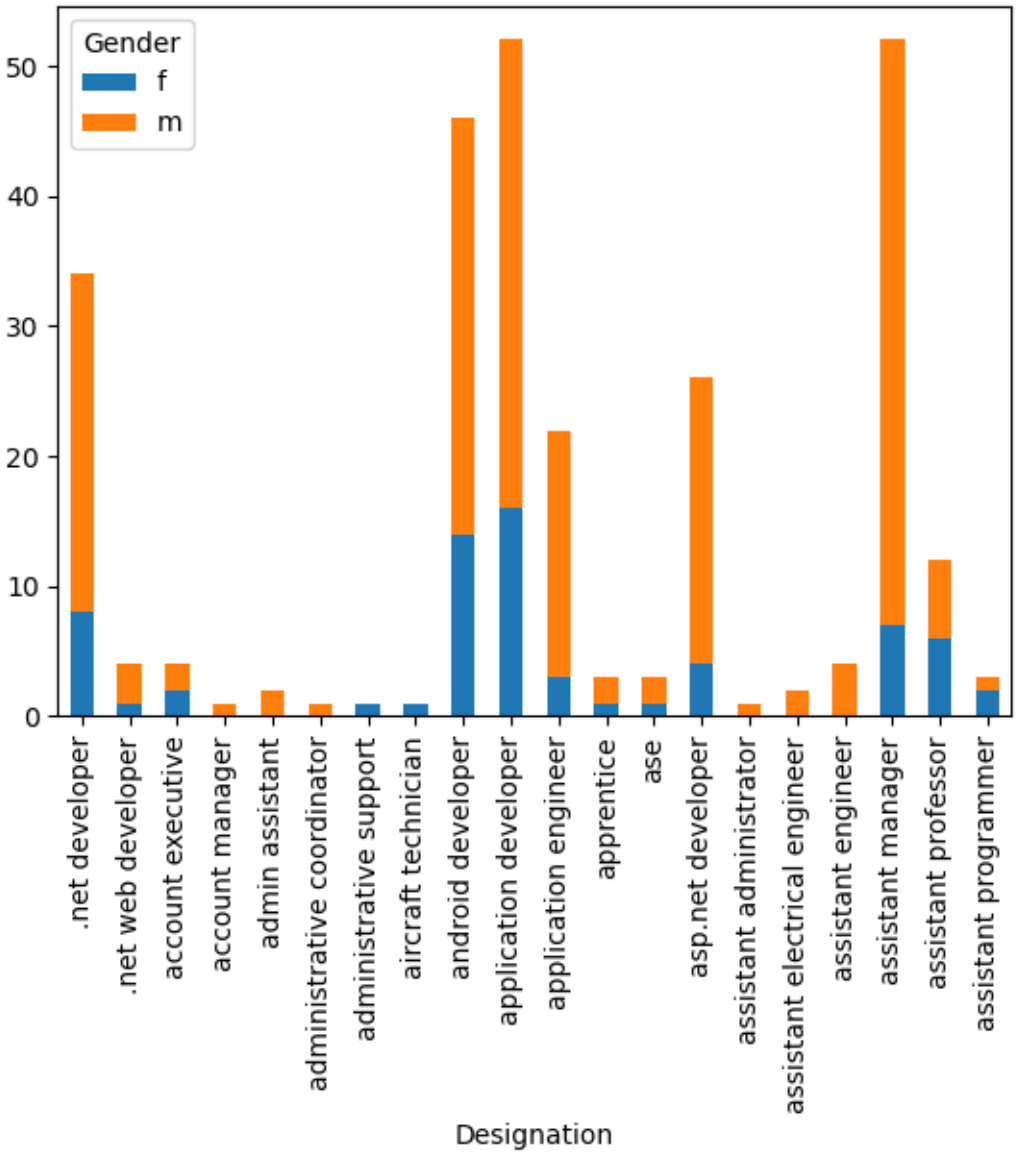
[23]:
```
pvt = df.
  ↪pivot_table(index="Designation",columns='Gender',values="Salary",aggfunc='count').
  ↪fillna(0)[:20]
pvt.plot(kind="bar",stacked=True,)
```

[23]:
```
<Axes: xlabel='Designation'>
```



[24]:
```
pvt = df.
  ↪pivot_table(index="Designation",columns='Gender',values="Salary",aggfunc='count').
  ↪fillna(0)[:20]
pvt.plot(kind="bar",stacked=True)
```

[24]: <Axes: xlabel='Designation'>



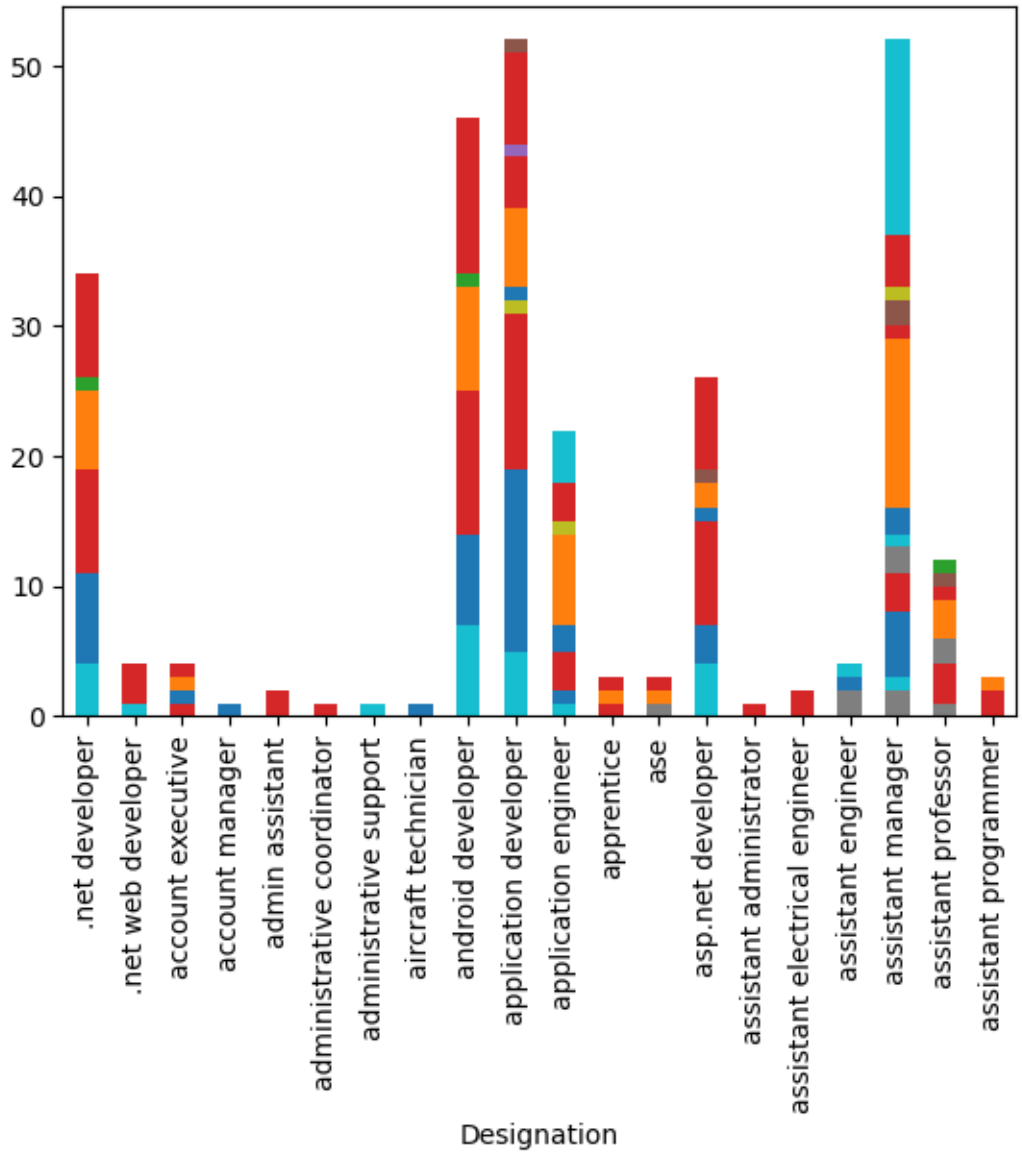
[25]:
```
df.dtypes
```

[25]:
```
ID                          int64
Salary                      int64
DOJ                datetime64[ns]
DOL                        object
Designation                object
JobCity                    object
Gender                     object
DOB                datetime64[ns]
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
```

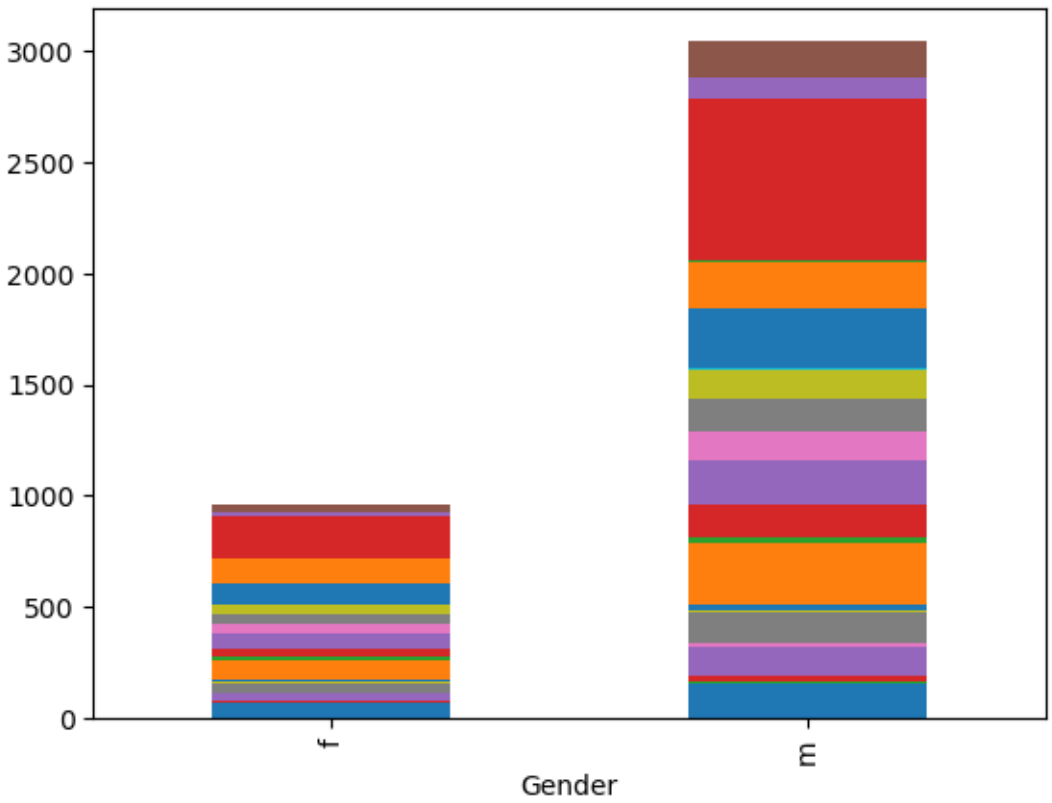
[26]:
```
pvt = df.
  ↪pivot_table(index="Designation",columns='Specialization',values="Salary",aggfunc='count').
  ↪fillna(0)[:20]
pvt.plot(kind="bar",stacked=True,legend=False)
```

[26]:
```
<Axes: xlabel='Designation'>
```



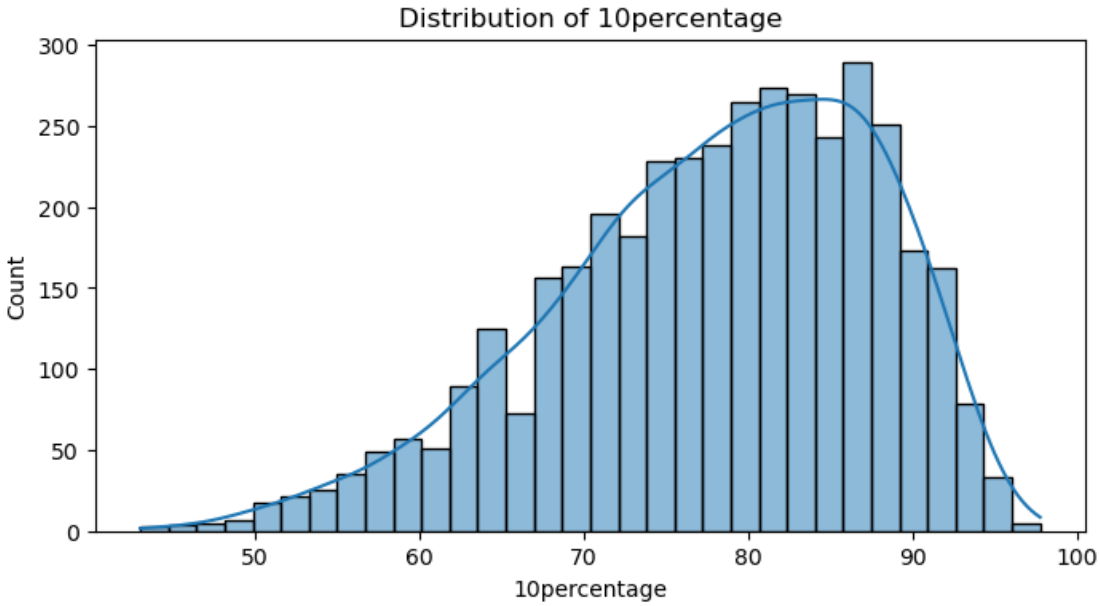
[27]:
```
pvt = df.
  ↪pivot_table(index="Gender",columns='CollegeState',values="Salary",aggfunc='count').
  ↪fillna(0)[:20]
pvt.plot(kind="bar",stacked=True,legend=False)
```

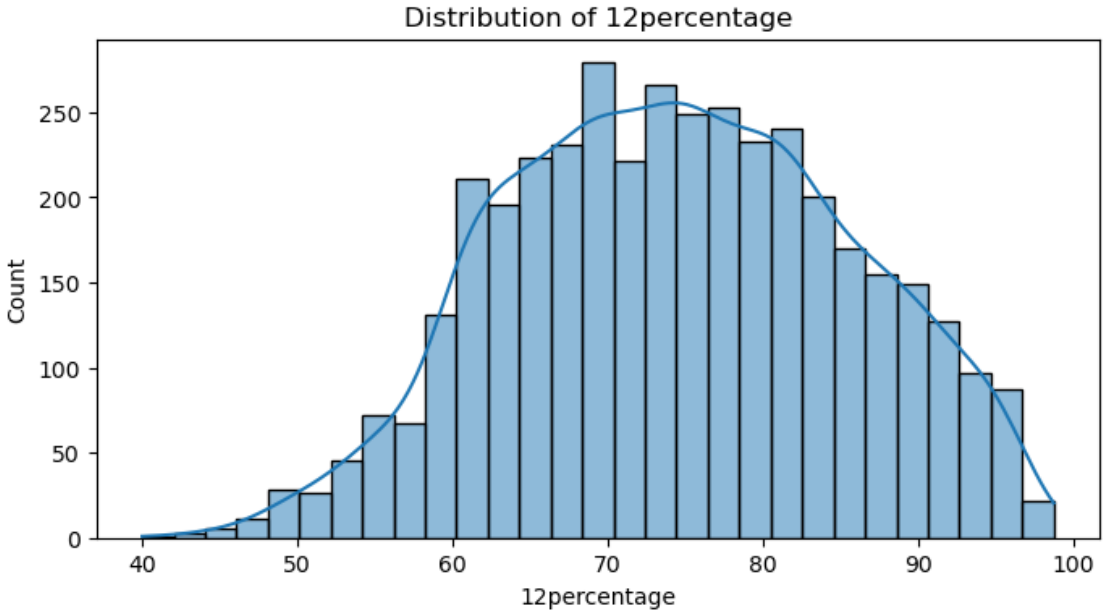
[27]: <Axes: xlabel='Gender'>

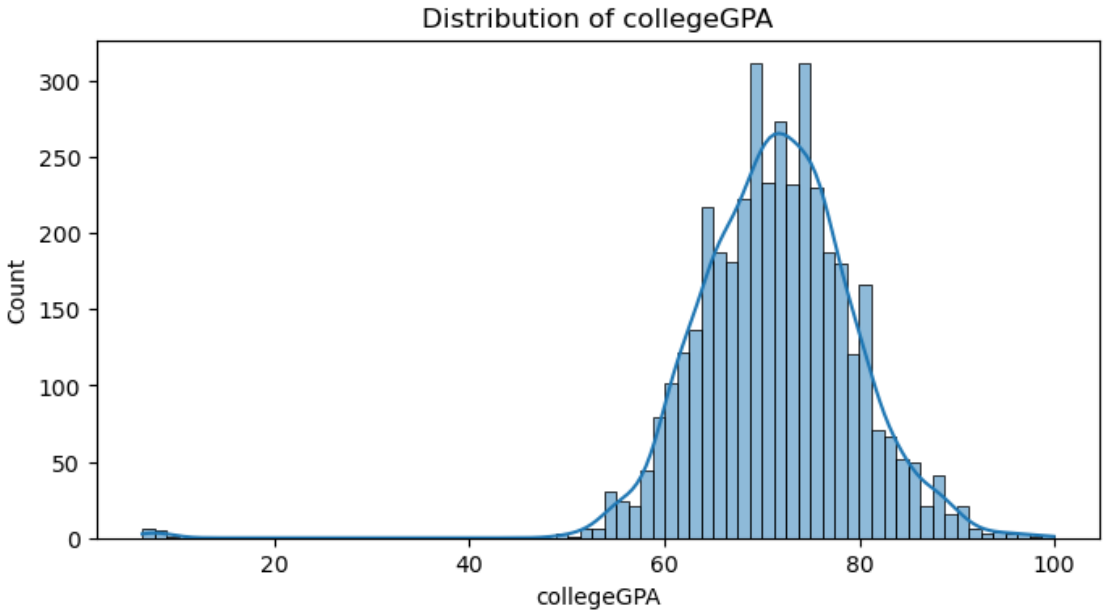


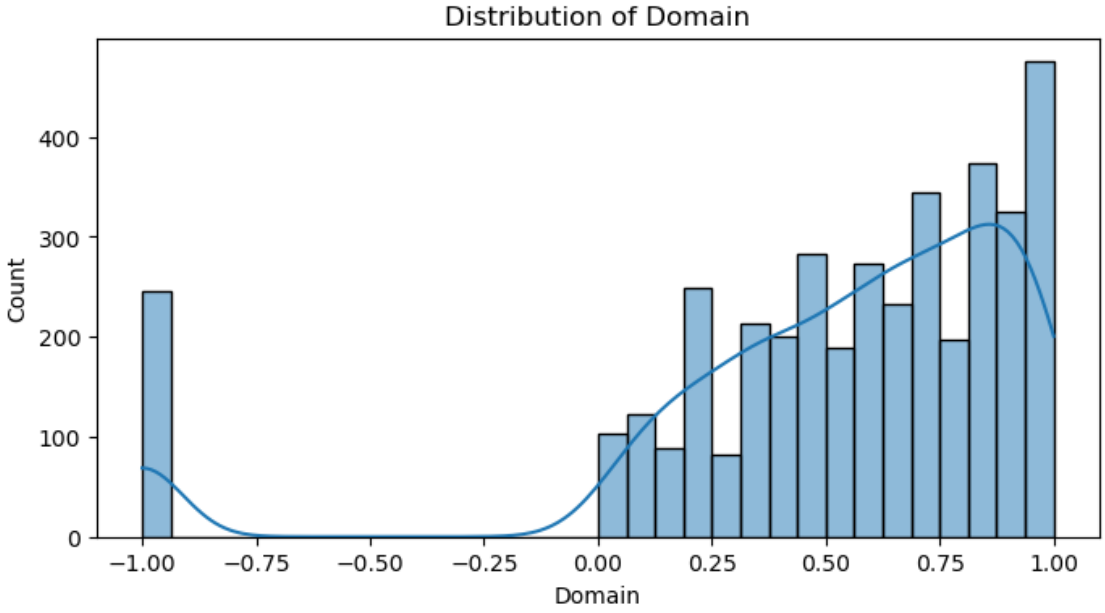
[28]:
```
categorical_columns = df.select_dtypes(include=['object']).columns
numerical_columns = df.select_dtypes(include=['float64']).columns
```

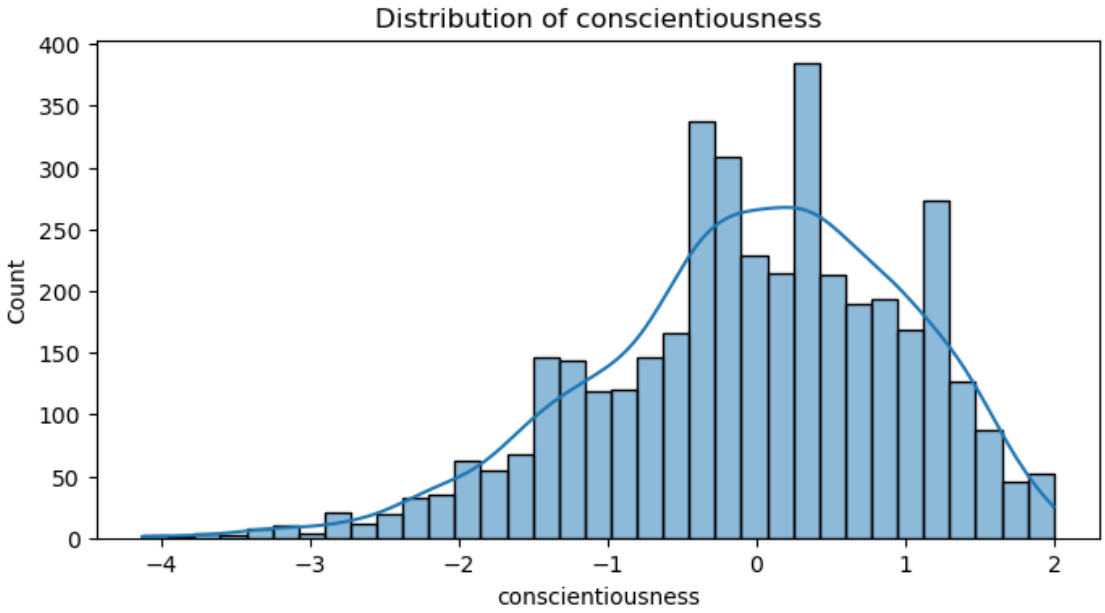
[31]:
```
# Distributions
for col in numerical_columns:
    plt.figure(figsize=(8, 4))
    sns.histplot(df[col], kde=True)
    plt.title(f"Distribution of {col}")
    plt.show()
```

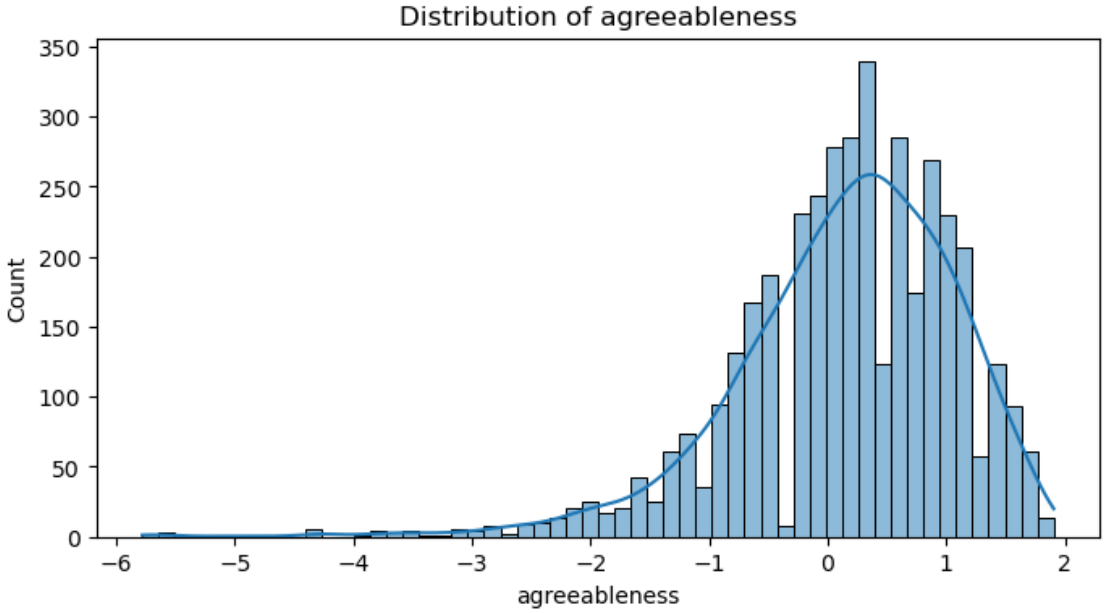

Distribution of 10percentage
Count
10percentage
0
50
100
150
200
250
300
50
60
70
80
90
100

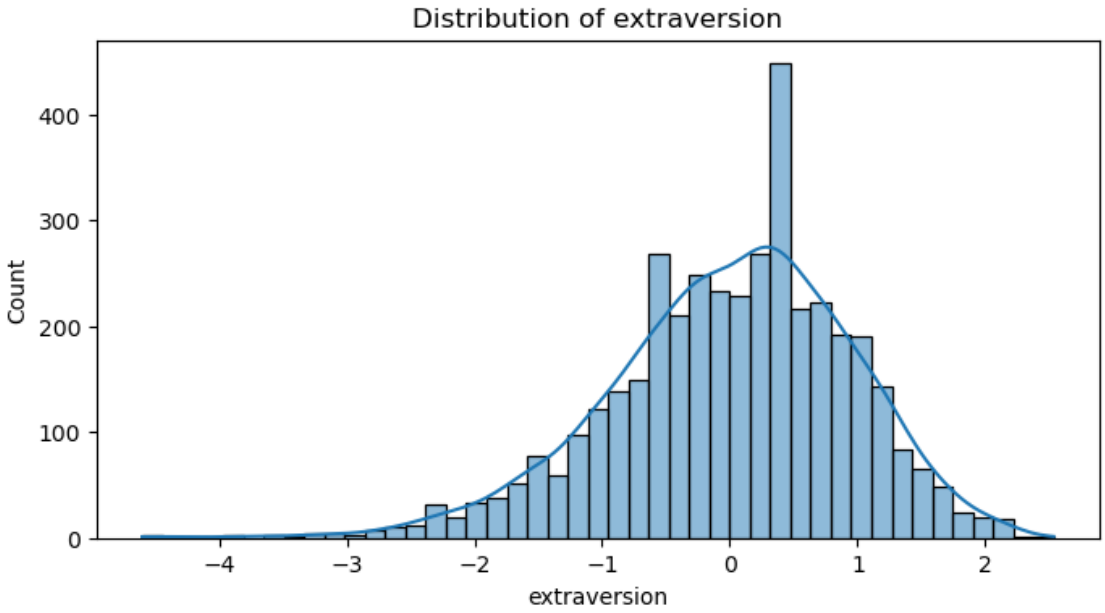

Distribution of 12percentage
Count
12percentage
0
50
100
150
200
250
40
50
60
70
80
90
100

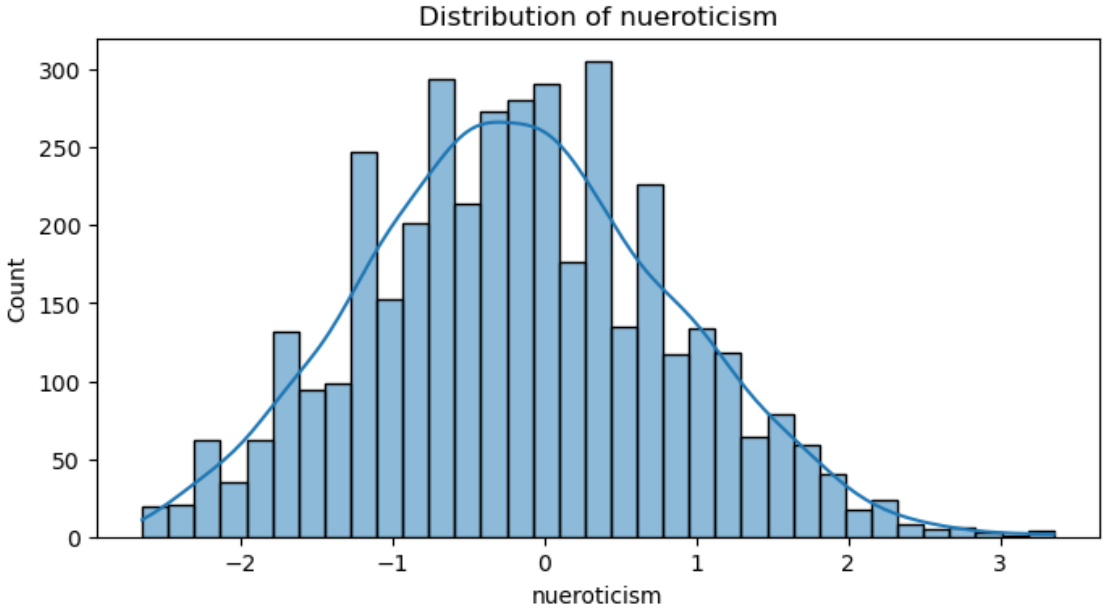

Distribution of collegeGPA
Count
collegeGPA
0
50
100
150
200
250
300
20
40
60
80
100

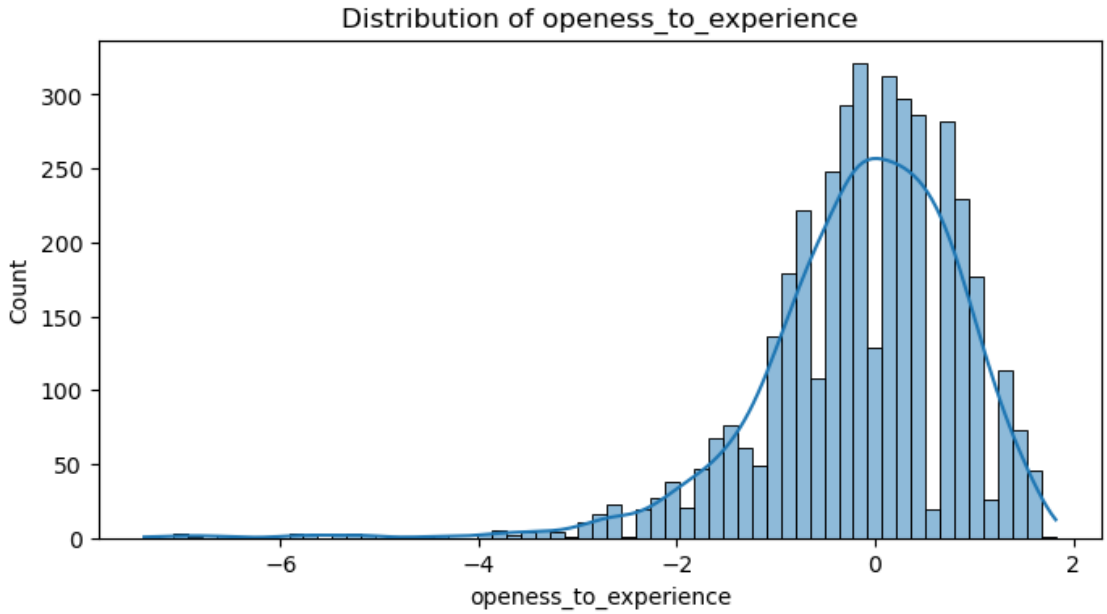

Distribution of Domain
Count
Domain
0
100
200
300
400
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Distribution of conscientiousness
Count
conscientiousness
400
350
300
250
200
150
100
50
0
−4
−3
−2
−1
0
1
2


Distribution of agreeableness
Count
agreeableness
350
300
250
200
150
100
50
0
−6
−5
−4
−3
−2
−1
0
1
2


Distribution of extraversion
Count
extraversion
400
300
200
100
0
−4
−3
−2
−1
0
1
2


Distribution of nueroticism
Count
nueroticism
300
250
200
150
100
50
0
−2
−1
0
1
2
3



[ ]:
```
df[]
```

[ ]:

2 Insights

- Most employees are currently working, showing a stable workforce.
- There are more women than men in the company.
- A large number of employees are based in Bangalore, the main job location.
- The most common job title is Senior Quality Engineer.
- Many employees completed secondary school in Andhra Pradesh, mostly through the CBSE board.
- After high school, most employees pursued B.Tech/B.E. degrees, while some continued to MCA or M.Tech.
- Computer Engineering is the most popular field of study.
- Salaries vary widely by job title and location, with senior positions and Bangalore generally offering higher pay.
- A few employees earn significantly high salaries, indicating specialized roles may pay more.
- More years of experience usually lead to higher salaries.Salaries vary widely by job title and location, with - - – senior positions and Bangalore generally offering higher pay.
- A few employees earn significantly high salaries, indicating specialized roles may pay more.
- More years of experience usually lead to higher salaries.
- Employees in senior roles typically have more experience, which helps them earn higher salaries.

[ ]:

3 Question

Times of India article dated Jan 18, 2019 states that "After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate."

[32]:
```
specialization_columns = ['ComputerScience', 'MechanicalEngg',␣
 ↪'ElectricalEngg', 'TelecomEngg', 'CivilEngg']
data_filtered = df[['Gender'] + specialization_columns]

data_filtered = data_filtered.replace(-1, pd.NA)

summary_table = data_filtered.melt(id_vars='Gender',␣
 ↪value_vars=specialization_columns, var_name='Specialization',␣
 ↪value_name='Value')
summary_table = summary_table.dropna(subset=['Value'])
gender_specialization_count = summary_table.groupby(['Gender',␣
 ↪'Specialization']).size().unstack(fill_value=0)

chi2, p, dof, expected = chi2_contingency(gender_specialization_count)
gender_specialization_count, p
```

[32]:
```
(Specialization  CivilEngg  ComputerScience  ElectricalEngg  MechanicalEngg  \
 Gender
 f                     7              221              30              17
 m                    35              681             131             218

 Specialization  TelecomEngg
 Gender
 f                      94
 m                     280  ,
 1.486278445299714e-07)
```

[33]:
```
pwd
```

[33]:
```
'C:\\Users\\upscv'
```

[ ]:
```

```