

AMCAT_EDA_PROJECT

October 15, 2024

```
[1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
[2]: df=pd.read_csv(r'C:\Users\premc\Downloads\data.xlsx - Sheet1.csv')
```

```
[3]: df.shape
```

```
[3]: (3998, 39)
```

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

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

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

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

```
      conscientiousness  agreeableness  extraversion  nueroticism \
0      0.9737      0.8128      0.5269      1.35490
1     -0.7335      0.3789      1.2396     -0.10760
2      0.2718      1.7109      0.1637     -0.86820
```

3	0.0464	0.3448	-0.3440	-0.40780
4	-0.8810	-0.2793	-1.0697	0.09163

openess_to_experience	
0	-0.4455
1	0.8637
2	0.6721
3	-0.9194
4	-0.1295

[5 rows x 39 columns]

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

```
[5]:
```

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

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

	ComputerScience	MechanicalEngg	ElectricalEngg	TelecomEngg \
count	3998.000000	3998.000000	3998.000000	3998.000000
mean	90.742371	22.974737	16.478739	31.851176
std	175.273083	98.123311	87.585634	104.852845
min	-1.000000	-1.000000	-1.000000	-1.000000
25%	-1.000000	-1.000000	-1.000000	-1.000000
50%	-1.000000	-1.000000	-1.000000	-1.000000
75%	-1.000000	-1.000000	-1.000000	-1.000000
max	715.000000	623.000000	676.000000	548.000000

	CivilEngg	conscientiousness	agreeableness	extraversion \
count	3998.000000	3998.000000	3998.000000	3998.000000
mean	2.683842	-0.037831	0.146496	0.002763

std	36.658505	1.028666	0.941782	0.951471
min	-1.000000	-4.126700	-5.781600	-4.600900
25%	-1.000000	-0.713525	-0.287100	-0.604800
50%	-1.000000	0.046400	0.212400	0.091400
75%	-1.000000	0.702700	0.812800	0.672000
max	516.000000	1.995300	1.904800	2.535400

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

[8 rows x 27 columns]

```
[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   float64
3   DOJ                  3998 non-null   object
4   DOL                   3998 non-null   object
5   Designation           3998 non-null   object
6   JobCity               3998 non-null   object
7   Gender                3998 non-null   object
8   DOB                   3998 non-null   object
9   10percentage          3998 non-null   float64
10  10board                3998 non-null   object
11  12graduation           3998 non-null   int64
12  12percentage           3998 non-null   float64
13  12board                3998 non-null   object
14  CollegeID              3998 non-null   int64
15  CollegeTier            3998 non-null   int64
16  Degree                 3998 non-null   object
17  Specialization         3998 non-null   object
18  collegeGPA             3998 non-null   float64
19  CollegeCityID          3998 non-null   int64
20  CollegeCityTier        3998 non-null   int64
21  CollegeState           3998 non-null   object
```

```

22 GraduationYear      3998 non-null   int64
23 English             3998 non-null   int64
24 Logical             3998 non-null   int64
25 Quant               3998 non-null   int64
26 Domain              3998 non-null   float64
27 ComputerProgramming 3998 non-null   int64
28 ElectronicsAndSemicon 3998 non-null   int64
29 ComputerScience     3998 non-null   int64
30 MechanicalEngg      3998 non-null   int64
31 ElectricalEngg      3998 non-null   int64
32 TelecomEngg         3998 non-null   int64
33 CivilEngg           3998 non-null   int64
34 conscientiousness   3998 non-null   float64
35 agreeableness       3998 non-null   float64
36 extraversion        3998 non-null   float64
37 nueroticism          3998 non-null   float64
38 openness_to_experience 3998 non-null   float64
dtypes: float64(10), int64(17), object(12)
memory usage: 1.2+ MB

```

```
[7]: df.columns=df.columns.str.lower()
```

```
[8]: numerical_features=list(df.select_dtypes(include=['number']).columns)
categorical_features=list(df.select_dtypes(include=['object']).columns)
```

```
[9]: print(numerical_features)
```

```

['id', 'salary', '10percentage', '12graduation', '12percentage', 'collegeid',
'collegetier', 'collegegpa', 'collegacityid', 'collegacitytier',
'graduationyear', 'english', 'logical', 'quant', 'domain',
'computerprogramming', 'electronicsandsemicon', 'computerscience',
'mechanicalengg', 'electricalengg', 'telecomengg', 'civilengg',
'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism',
'openess_to_experience']

```

```
[10]: categorical_features.remove('unnamed: 0')
```

```
[11]: print(categorical_features)
```

```

['doj', 'dol', 'designation', 'jobcity', 'gender', 'dob', '10board', '12board',
'degree', 'specialization', 'collegestate']

```

```
[12]: num_continuos_features=['salary','10percentage','12percentage','collegegpa','domain','conscien
```

```
[13]: numerical_features=[feature for feature in numerical_features if feature not in
    ↪in num_continuos_features]
```

```
[14]: numerical_features
```

```
[14]: ['id',
      '12graduation',
      'collegeid',
      'collegetier',
      'collegecityid',
      'collegecitytier',
      'graduationyear',
      'english',
      'logical',
      'quant',
      'computerprogramming',
      'electronicsandsemicon',
      'computerscience',
      'mechanicalengg',
      'electricalengg',
      'telecomengg',
      'civilengg']
```

```
[15]: ### mean salary
      print(df['salary'].mean())
```

```
307699.8499249625
```

```
[16]: df['salary'].median()
```

```
[16]: 300000.0
```

```
[66]: data=df['salary']
      data=pd.DataFrame(data)
      q1=df.salary.quantile(0.25)
      q3=df.salary.quantile(0.75)
      iqr=q3-q1
```

```
[67]: lower_bound=q1-1.5*iqr
      upper_bound=q3+1.5*iqr
```

```
[68]: outliers1=df[(df['salary']<lower_bound) | (df['salary']>upper_bound)]
      outliers1
```

```
[68]:
```

	id	salary	doj	dol	\
3	267447	1100000.0	7/1/11 0:00	present	
76	361583	800000.0	6/1/12 0:00	present	
92	1250429	1500000.0	11/1/14 0:00	7/1/14 0:00	
123	312164	1200000.0	7/1/10 0:00	7/1/11 0:00	
128	206734	675000.0	11/1/11 0:00	present	
...	
3823	918568	775000.0	8/1/14 0:00	present	
3904	267121	850000.0	9/1/11 0:00	present	

3912	231229	730000.0	7/1/13 0:00	present
3961	230702	700000.0	7/1/11 0:00	9/1/14 0:00
3992	344407	800000.0	4/1/14 0:00	4/1/15 0:00

	designation	jobcity	gender	dob	\
3	senior software engineer	Gurgaon	m	12/5/89 0:00	
76	software engineer	Bangalore	m	1/25/91 0:00	
92	application developer	Hyderabad	m	1/4/92 0:00	
123	engineer trainee	Maharajganj	m	4/25/88 0:00	
128	senior software engineer	Noida	m	11/7/88 0:00	
...	
3823	mechanical design engineer	Dammam	m	1/12/91 0:00	
3904	operations assistant	Noida	m	1/5/89 0:00	
3912	research scientist	Pune	m	11/15/89 0:00	
3961	planning engineer	Rajpura	m	12/27/87 0:00	
3992	manager	Rajkot	m	6/22/90 0:00	

	10percentage	10board	...	computerscience	\
3	85.60	cbse	...	-1	
76	93.44	karnataka state board	...	-1	
92	79.00	state board	...	346	
123	59.80	icse	...	-1	
128	60.00	0	...	-1	
...	
3823	87.40	cbse	...	-1	
3904	83.40	cbse	...	-1	
3912	84.67	0	...	-1	
3961	84.20	0	...	-1	
3992	73.00	0	...	-1	

	mechanicalengg	electricalengg	telecomengg	civilengg	conscientiousness	\
3	-1	-1	-1	-1	0.0464	
76	-1	-1	-1	-1	-0.4173	
92	-1	-1	-1	-1	0.4155	
123	206	-1	-1	-1	0.2009	
128	-1	-1	-1	-1	-0.8810	
...	
3823	469	-1	-1	-1	-0.8772	
3904	-1	-1	-1	-1	-0.8810	
3912	-1	-1	-1	-1	-1.3447	
3961	-1	-1	-1	460	-1.3447	
3992	-1	-1	-1	480	0.3555	

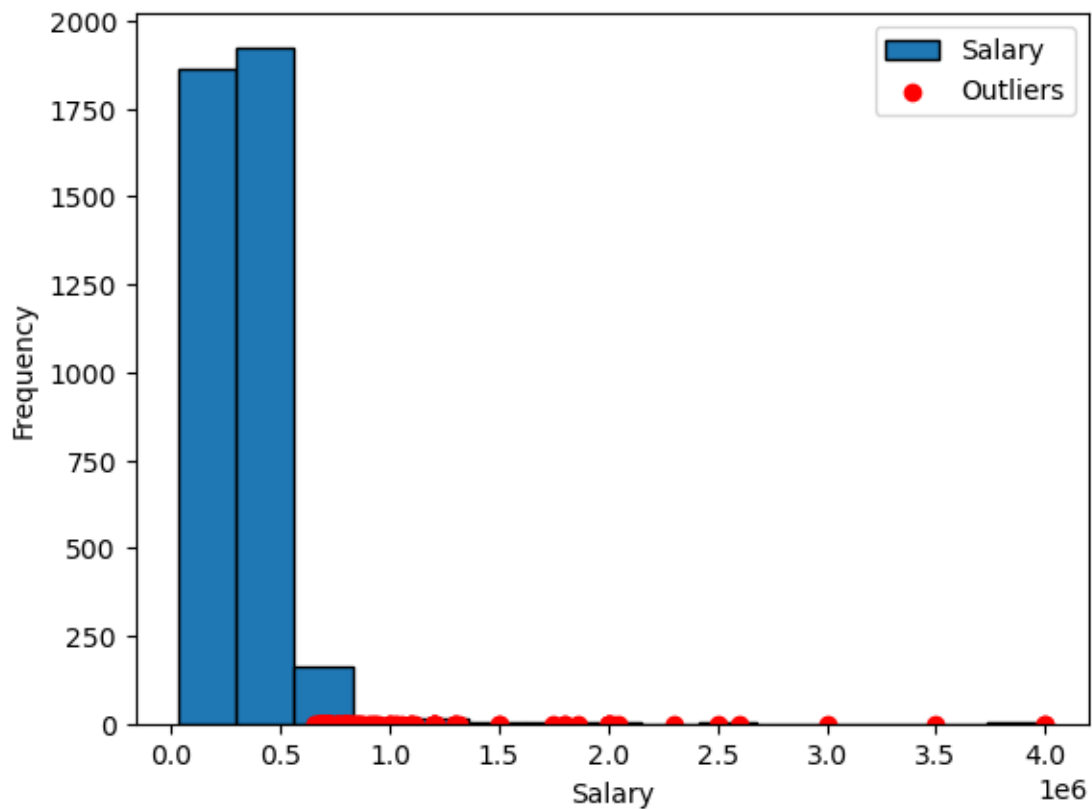
	agreeableness	extraversion	nueroticism	openess_to_experience
3	0.3448	-0.3440	-0.40780	-0.9194
76	0.9688	-0.1988	-0.29020	0.3049
92	0.5454	0.9322	-0.61470	0.8637

123	1.1248	1.1074	-1.11280	0.9763
128	-0.2793	-0.6343	-0.64280	-2.9731
...
3823	-0.1206	-0.1437	-0.23440	-0.0943
3904	0.1888	-0.1988	-0.05520	-1.0774
3912	-1.0593	0.6720	1.00240	-1.7093
3961	0.0328	-2.3759	-0.99530	0.3444
3992	-0.9033	0.9623	0.64983	-0.4229

[109 rows x 38 columns]

```
[20]: outliers1=outliers1['salary']
```

```
[21]: plt.hist(df['salary'],bins=15,edgecolor='black',label='Salary')
plt.scatter(outliers1,np.zeros_like(outliers1),color='red',label='Outliers')
plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.legend()
plt.show()
```



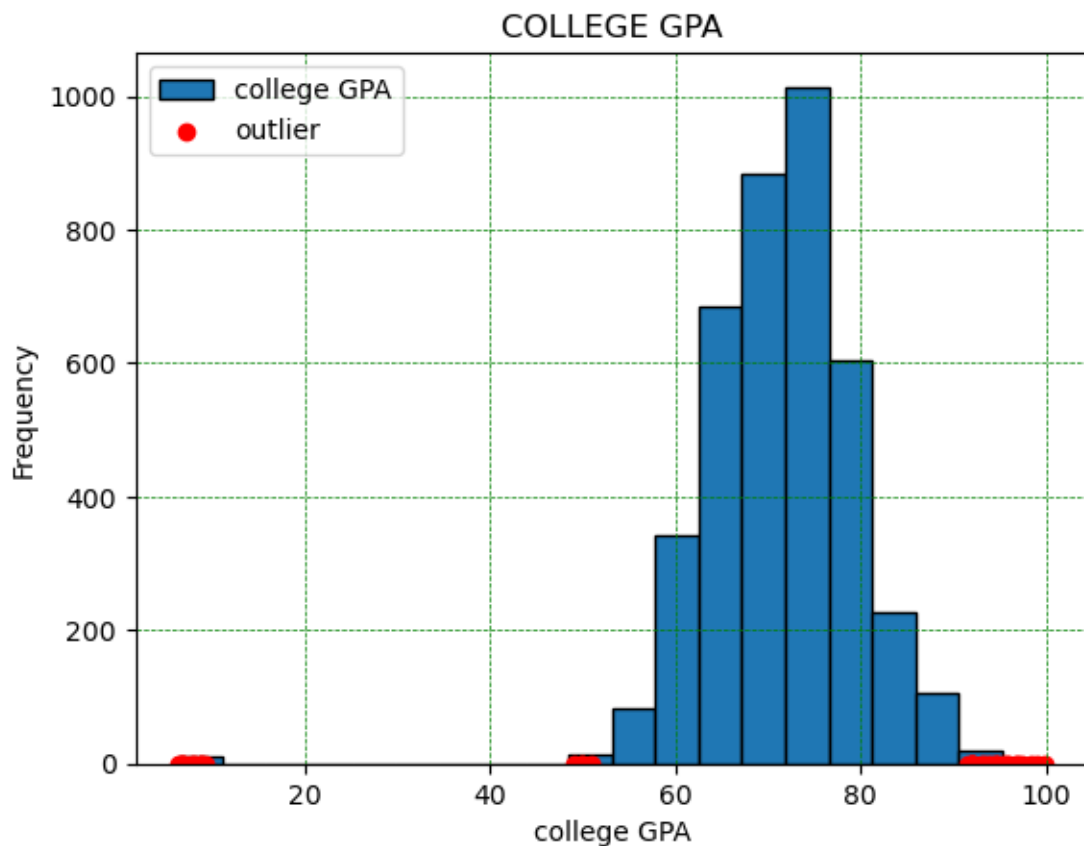
```
[22]: q1=df['collegegpa'].quantile(0.25)
      q3=df['collegegpa'].quantile(0.75)
      iqr=q3-q1
      l=q1-1.5*iqr
      u=q3+1.5*iqr
      print(l,u)
```

```
51.527499999999996 91.207500000000001
```

```
[23]: outliers=df[(df['collegegpa']<l) | (df['collegegpa']>u)]
```

```
[24]: outliers=outliers['collegegpa']
```

```
[25]: plt.hist(df['collegegpa'],bins=20,edgecolor='black',label='college GPA')
      plt.scatter(outliers,np.zeros_like(outliers),color='red',label='outlier')
      plt.xlabel('college GPA')
      plt.ylabel('Frequency')
      plt.title('COLLEGE GPA')
      plt.legend()
      plt.grid(color='green',linestyle='--',linewidth=0.5)
      plt.show()
```



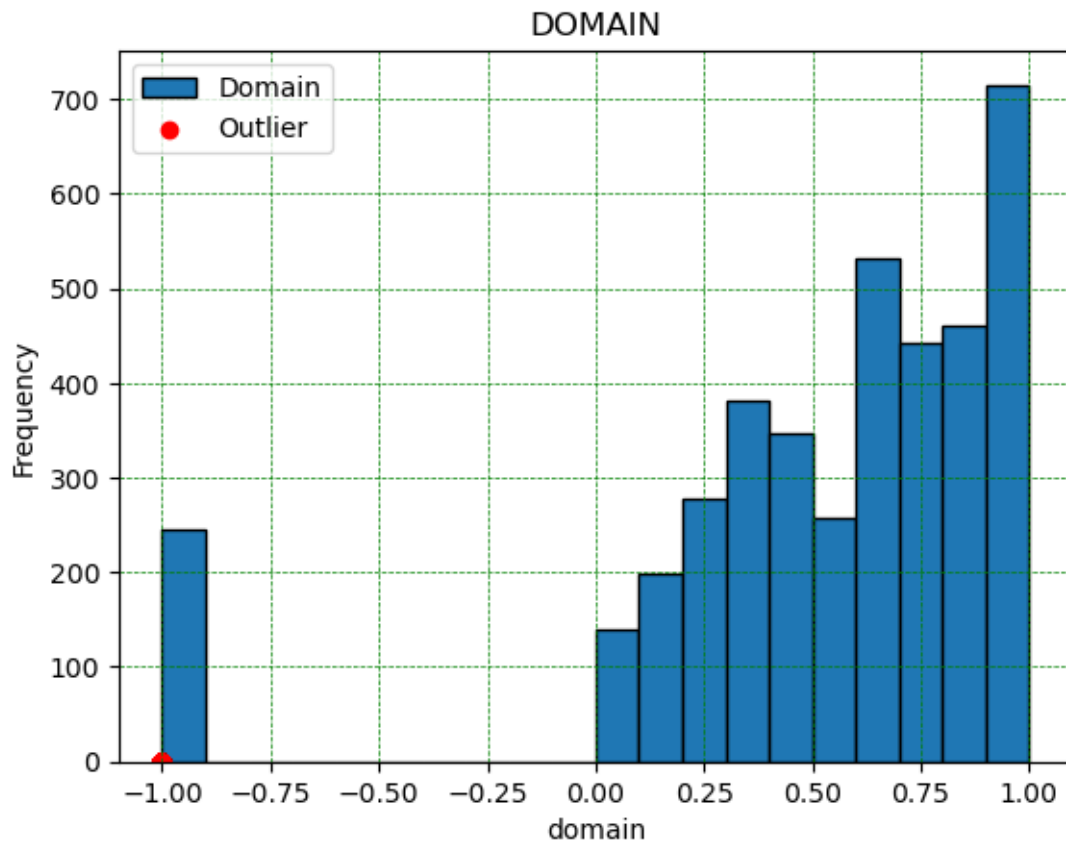

```
[26]: df['colleg GPA'].min(),df['colleg GPA'].max()
```

```
[26]: (6.45, 99.93)
```

```
[27]: q1=df['domain'].quantile(0.25)
q3=df['domain'].quantile(0.75)
iqr=q3-q1
l=q1-1.5*iqr
u=q3+1.5*iqr
```

```
[28]: domain_o=df[(df['domain']<l) | (df['domain']>u)]
domain_o=domain_o['domain']
```

```
[29]: plt.hist(df['domain'],bins=20,edgecolor='black',label='Domain')
plt.scatter(domain_o,np.zeros_like(domain_o),color='red',label='Outlier')
plt.xlabel('domain')
plt.ylabel('Frequency')
plt.title('DOMAIN')
plt.legend()
plt.grid(color='green',linestyle='--',linewidth=0.5)
plt.show()
```



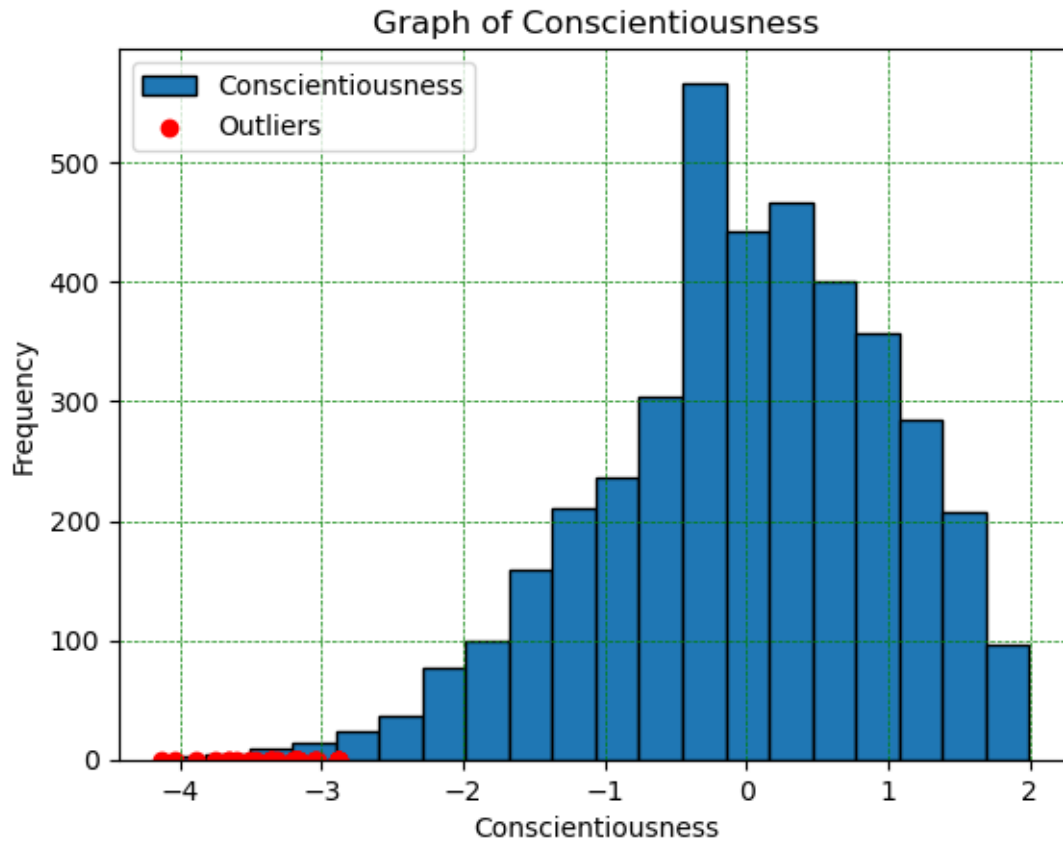
0.0.1 PERSONAL TRAITS

```
[30]: q1=df['conscientiousness'].quantile(0.25)
      q3=df['conscientiousness'].quantile(0.75)
      iqr=q3-q1
      lower_bound=q1-1.5*iqr
      upper_bound=q3+1.5*iqr

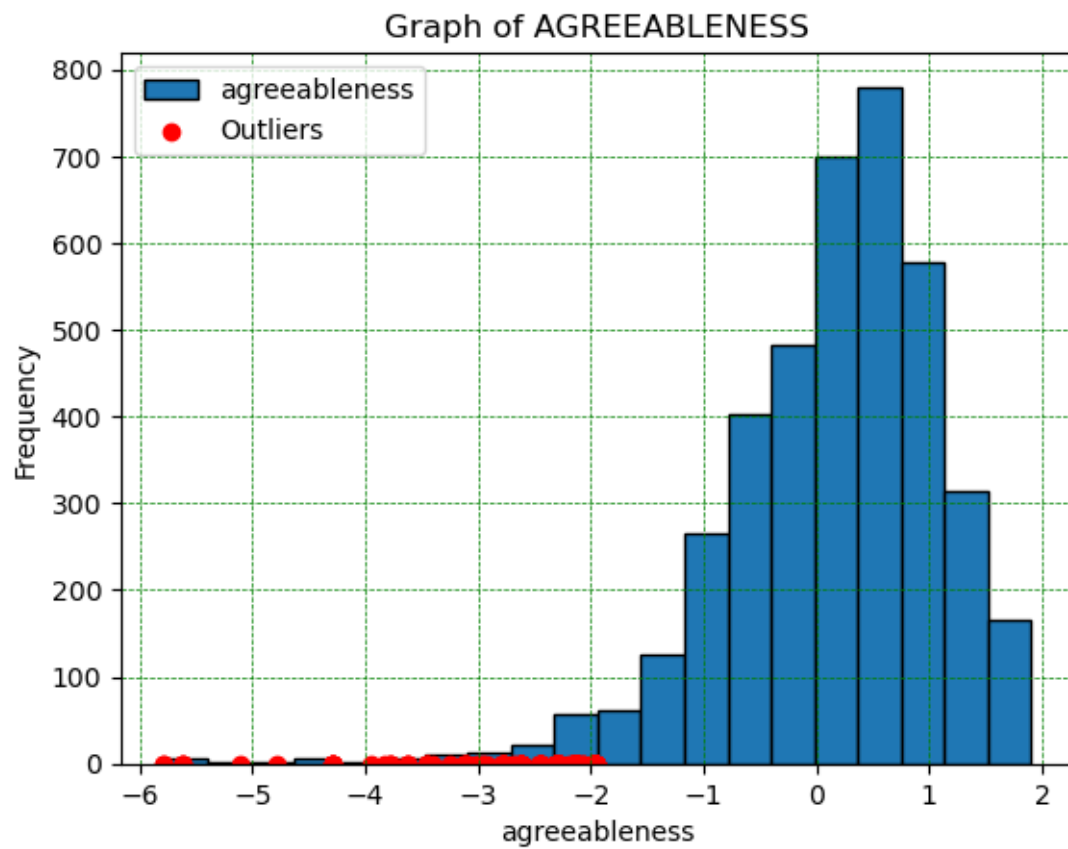
[31]: outliers=df[(df['conscientiousness']<lower_bound)|
      ↪(df['conscientiousness']>upper_bound)]

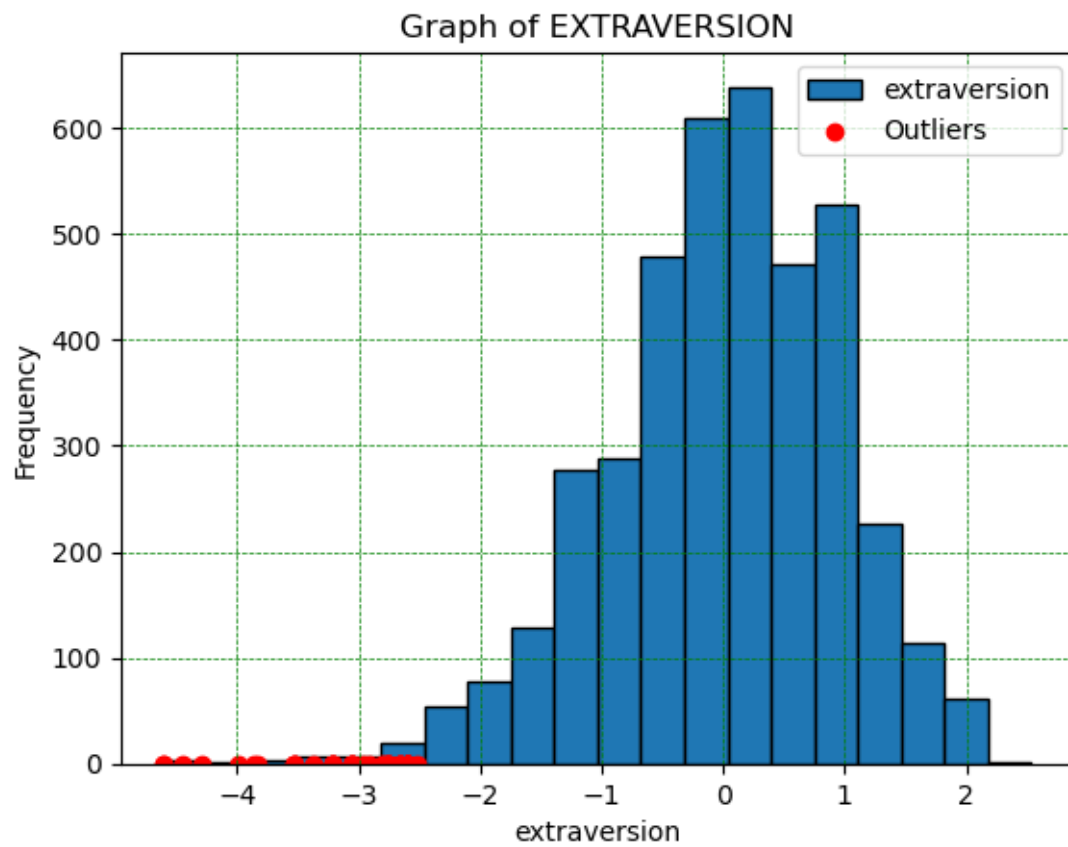
[32]: outliers=outliers['conscientiousness']

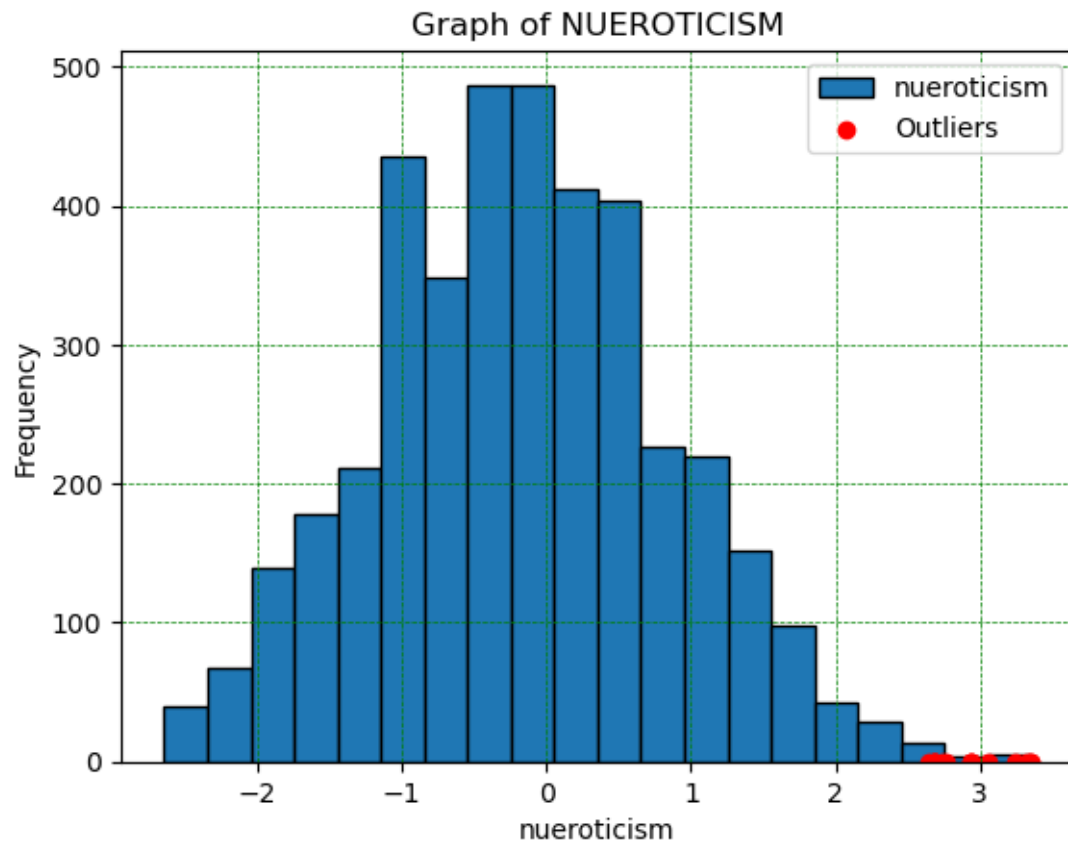
[33]: plt.
      ↪hist(df['conscientiousness'],bins=20,edgecolor='black',label='Conscientiousness')
      plt.scatter(outliers,np.zeros_like(outliers),color='red',label='Outliers')
      plt.xlabel('Conscientiousness')
      plt.ylabel('Frequency')
      plt.legend()
      plt.grid(color='green',linestyle='--',linewidth=0.5)
      plt.title('Graph of Conscientiousness')
      plt.show()
```

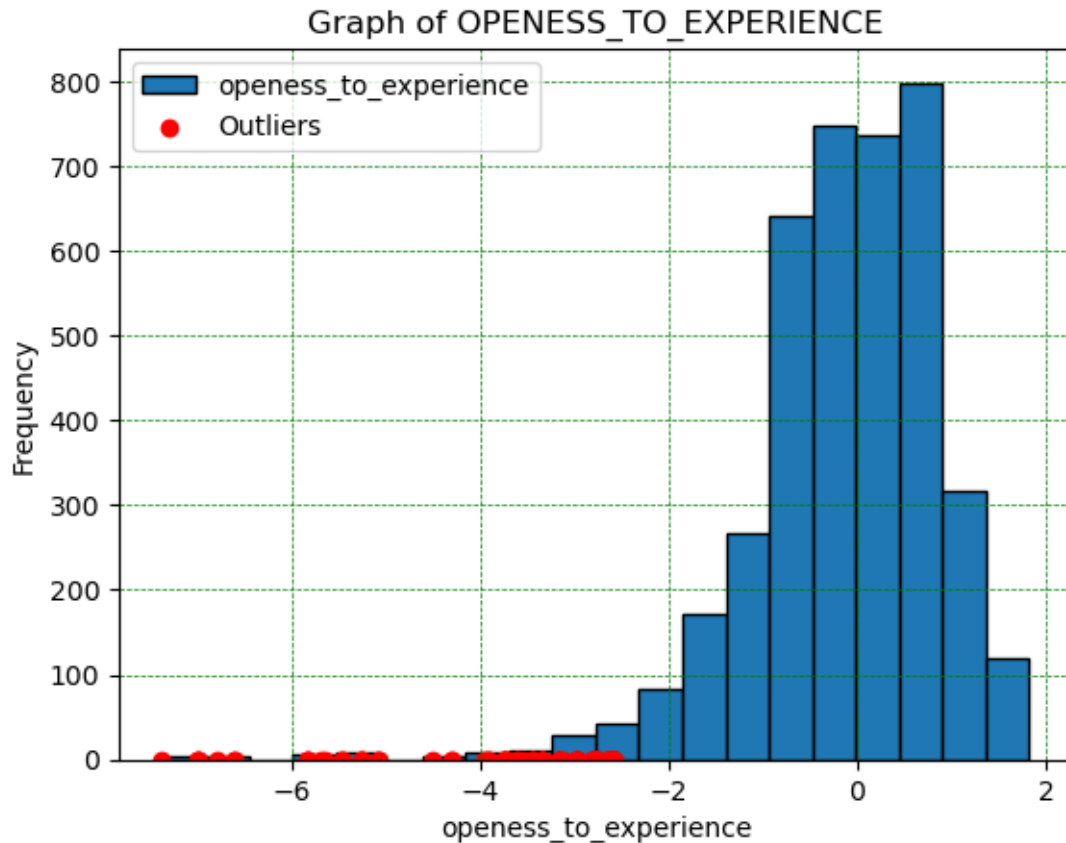


```
[34]: a=['agreeableness','extraversion','nueroticism','openess_to_experience']
for i in a:
    q1=df[i].quantile(0.25)
    q3=df[i].quantile(0.75)
    iqr=q3-q1
    lower_bound=q1-1.5*iqr
    upper_bound=q3+1.5*iqr
    outlier=df[(df[i]<lower_bound) | (df[i]>upper_bound)]
    outlier=outlier[i]
    plt.hist(df[i],bins=20,edgecolor='black',label=i)
    plt.scatter(outlier,np.zeros_like(outlier),color='red',label='Outliers')
    plt.grid(color='green',linestyle='--',linewidth=0.5)
    plt.xlabel(i)
    plt.ylabel('Frequency')
    plt.legend()
    plt.title(f""""Graph of {i.upper()}""")
    plt.show()
```

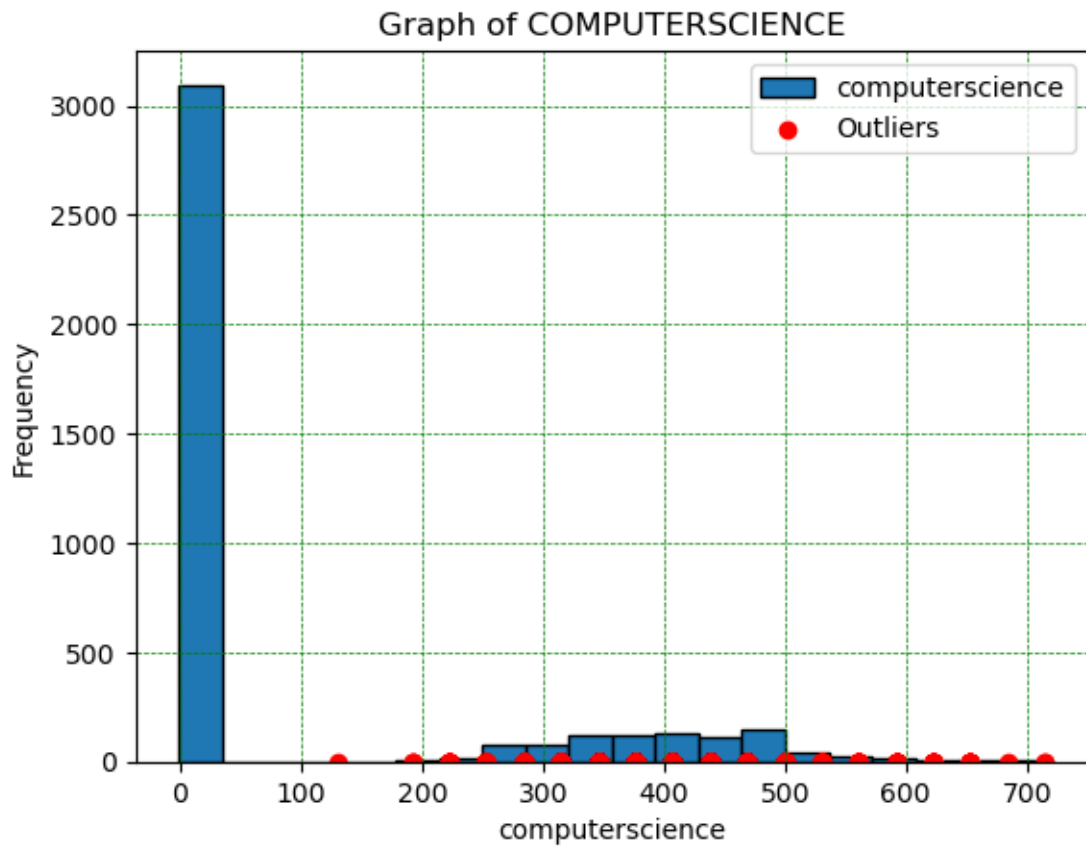


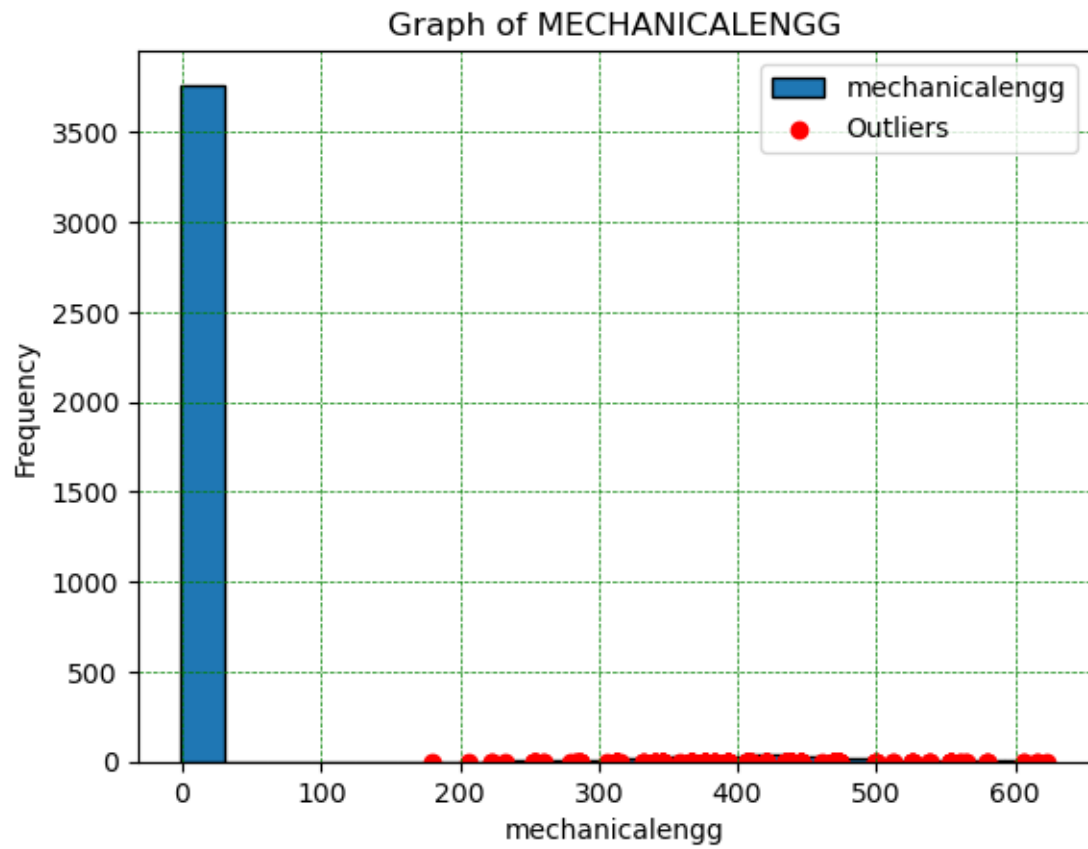


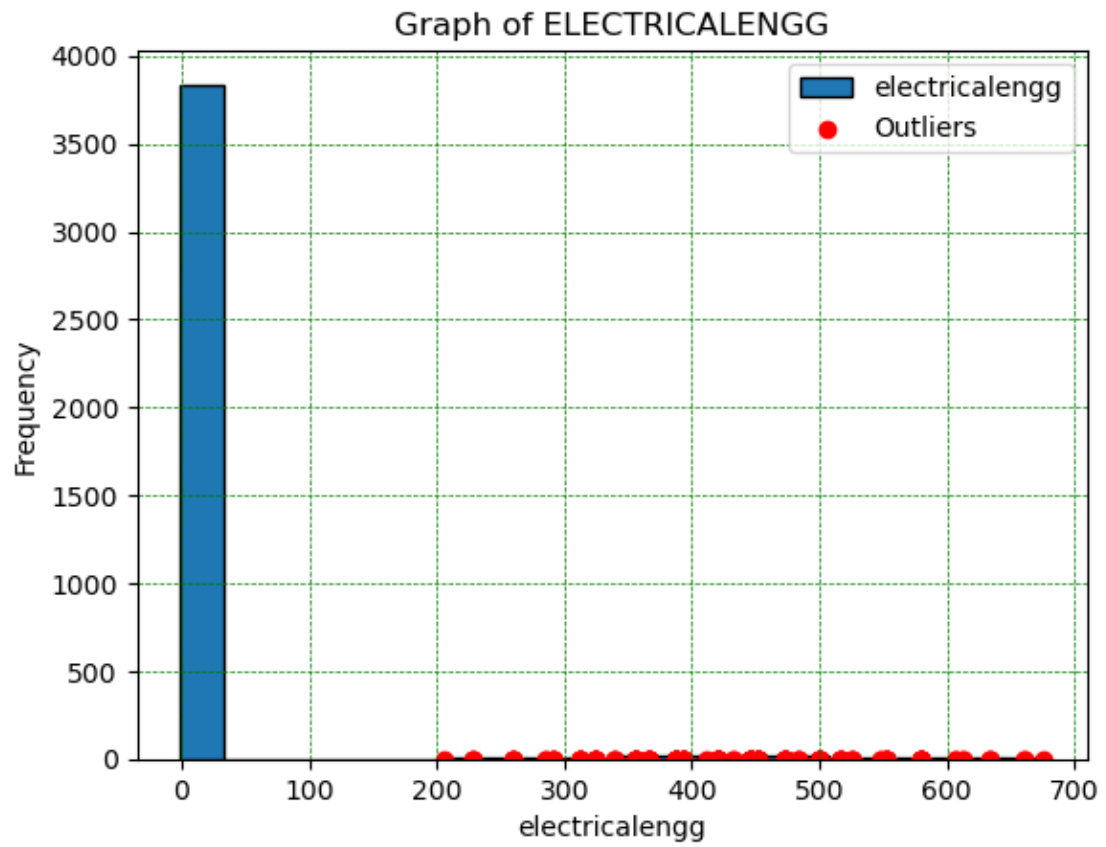


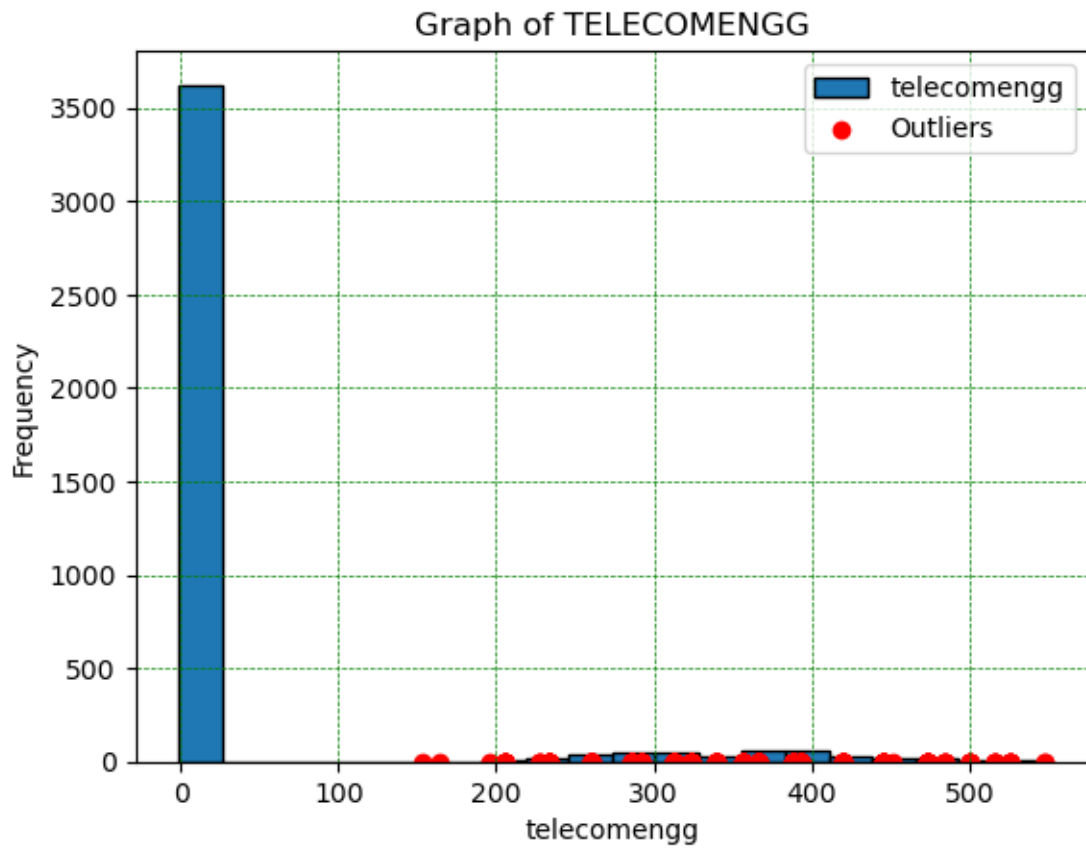


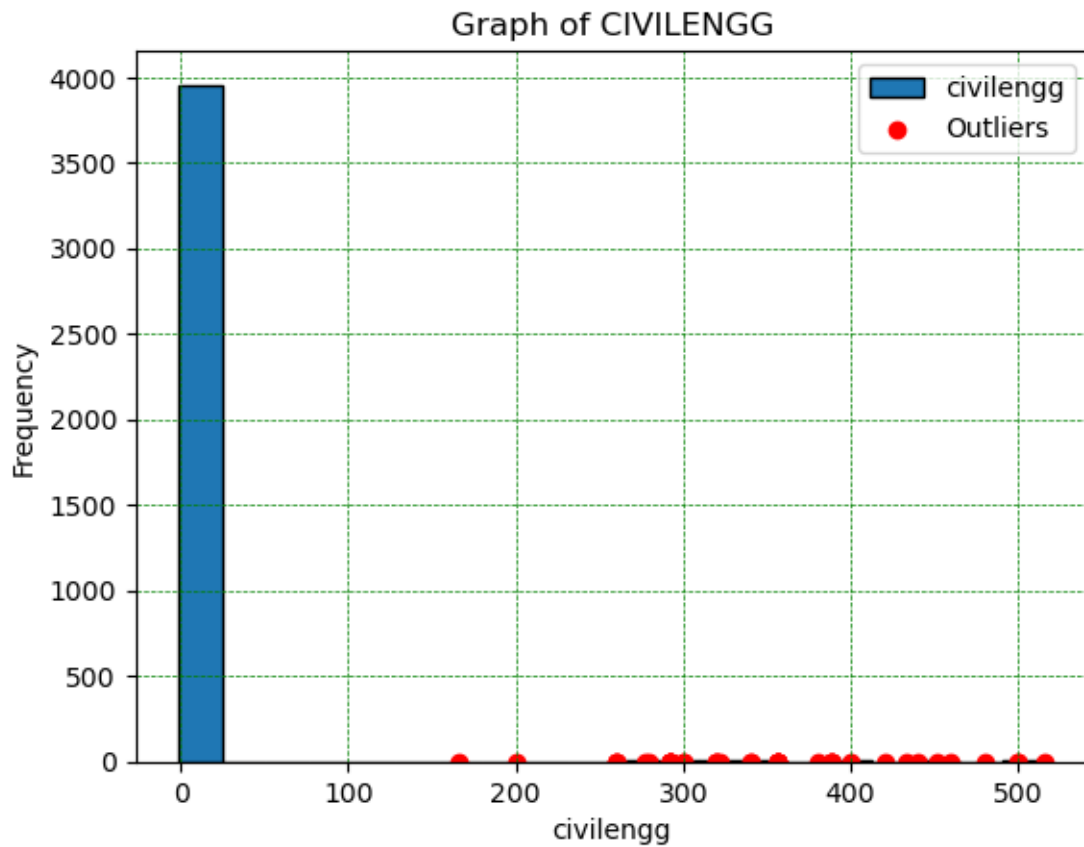
```
[35]: a=['computerscience','mechanicalengg','electricalengg','telecomengg','civilengg']
for i in a:
    q1=df[i].quantile(0.25)
    q3=df[i].quantile(0.75)
    iqr=q3-q1
    lower_bound=q1-1.5*iqr
    upper_bound=q3+1.5*iqr
    outlier=df[(df[i]<lower_bound) | (df[i]>upper_bound)]
    outlier=outlier[i]
    plt.hist(df[i],bins=20,edgecolor='black',label=i)
    plt.scatter(outlier,np.zeros_like(outlier),color='red',label='Outliers')
    plt.grid(color='green',linestyle='--',linewidth=0.5)
    plt.xlabel(i)
    plt.ylabel('Frequency')
    plt.legend()
    plt.title(f"Graph of {i.upper()}")
    plt.show()
```





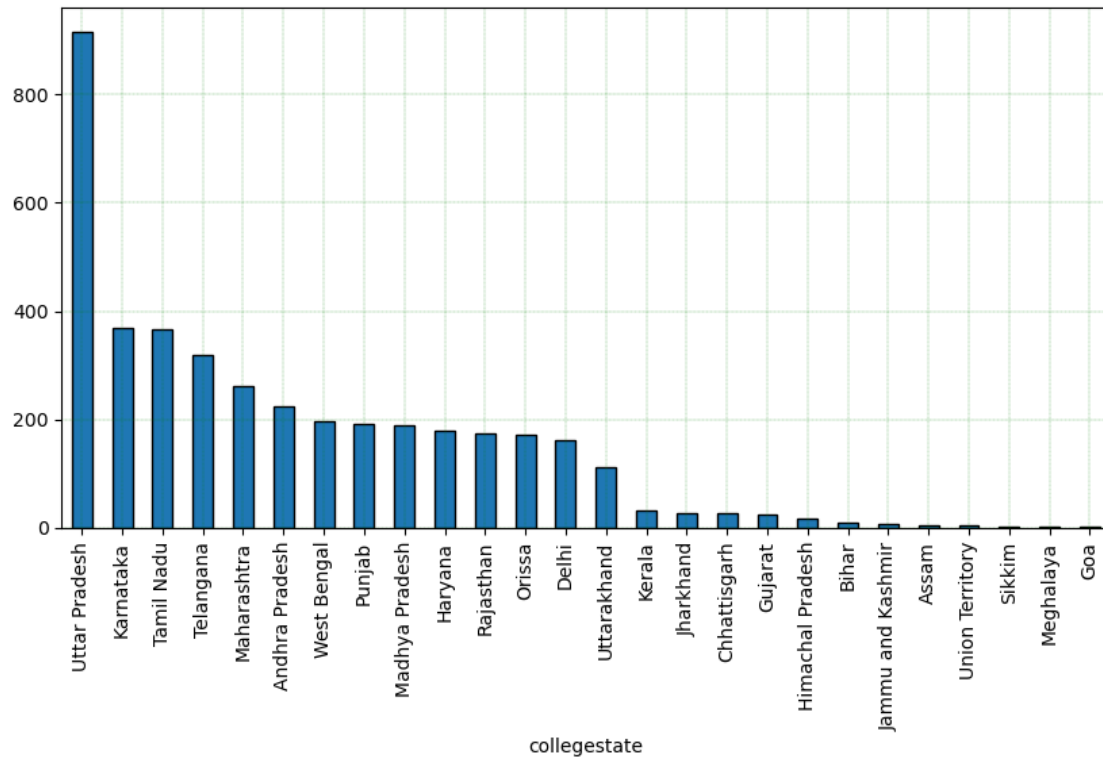






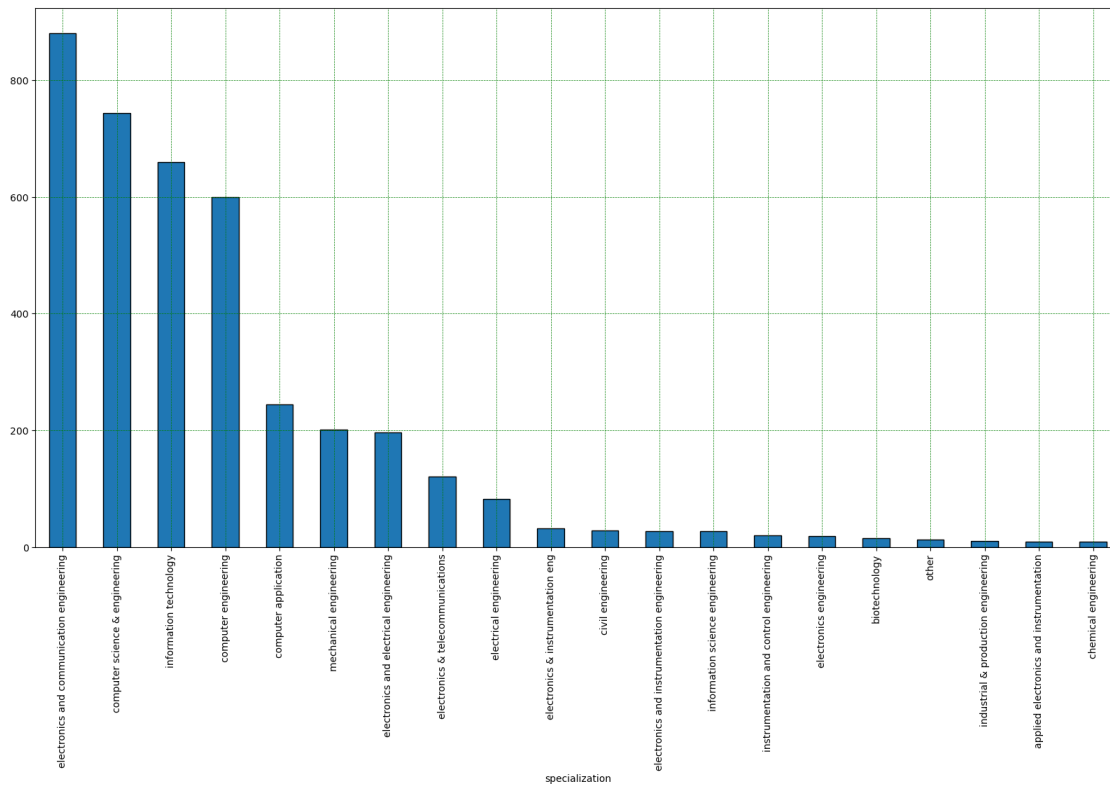
```
[36]: x=df['collegestate'].value_counts()
```

```
[37]: plt.figure(figsize=(10,5))
x.plot(kind='bar',edgecolor='black')
plt.grid(color='green',linestyle='--',linewidth=0.2)
plt.show()
```



```
[38]: y=df['specialization'].value_counts()[:20]
```

```
[39]: plt.figure(figsize=(20,10))
      y.plot(kind='bar',edgecolor='black')
      plt.grid(color='green',linestyle='--',linewidth=0.5)
      plt.show()
```



```
[40]: job_city=df['jobcity'].value_counts()[:20]
      job_city
```

```
[40]: jobcity
Bangalore      627
-1             461
Noida          368
Hyderabad      335
Pune           290
Chennai        272
Gurgaon        198
New Delhi      196
Mumbai         108
Kolkata         98
Jaipur          46
Lucknow         36
Mysore          36
Navi Mumbai    32
chennai         27
Chandigarh     26
pune            26
Greater Noida  26
```

```

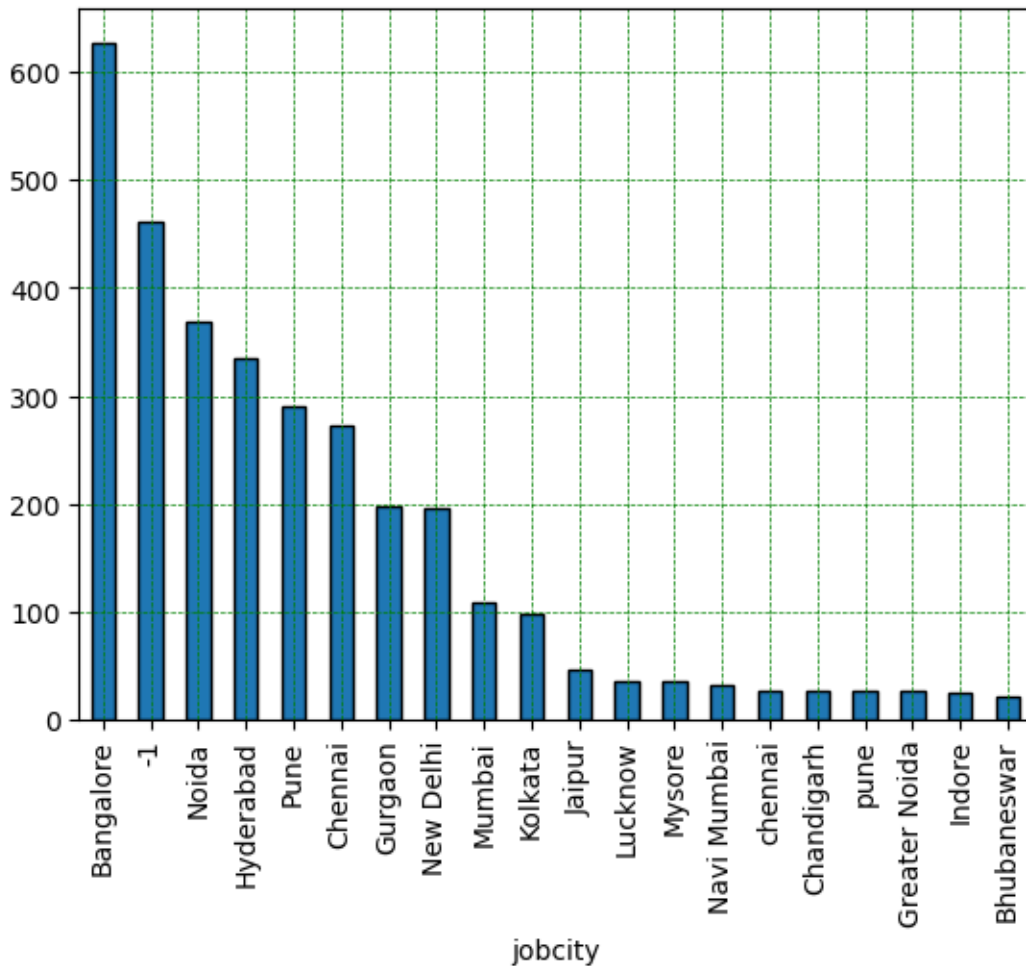
Indore          24
Bhubaneswar     22
Name: count, dtype: int64

```

```

[41]: job_city.plot(kind='bar',edgecolor='black')
plt.grid(color='green',linestyle='--',linewidth=0.5)
plt.show()

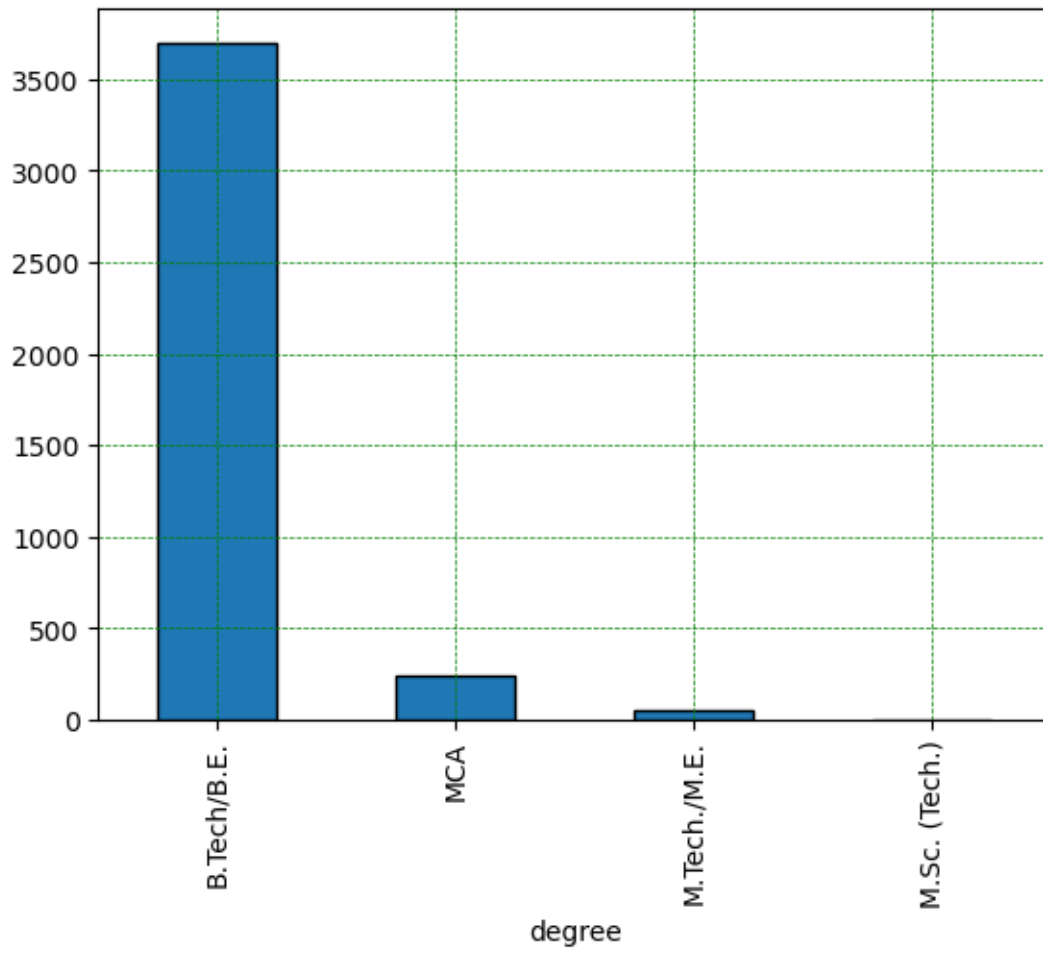
```

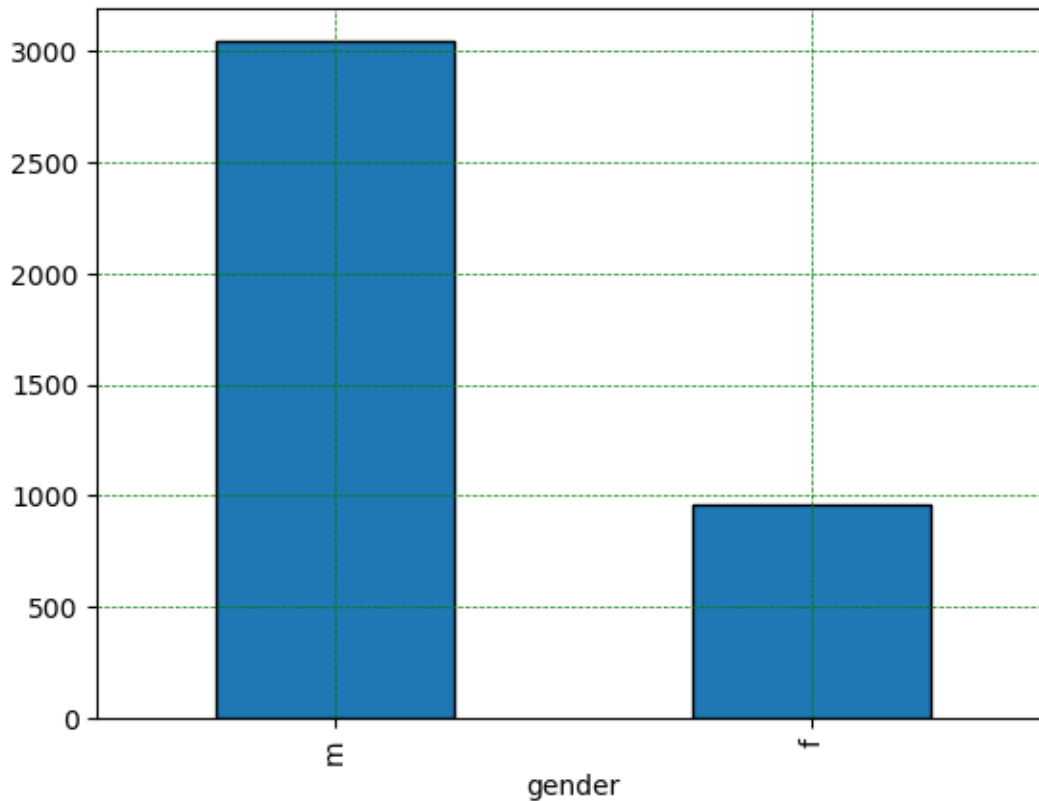


```

[42]: board=['degree','gender']
for i in board:
    x=df[i].value_counts()
    x.plot(kind='bar',edgecolor='black')
    plt.grid(color='green',linestyle='--',linewidth=0.5)
    plt.show()

```





```
[49]: df.drop(columns='unnamed: 0',inplace=True)
```

```
[70]: data=df[(df['salary']>lower_bound) & (df['salary']<upper_bound)]
```

```
[78]: sns.jointplot(x="10percentage",y="salary",data=data,kind='hex')
plt.gca().yaxis.set_major_formatter(plt.FuncFormatter(lambda x,pos:f"{x:,.0f}"))
```

C:\ProgramData\anaconda3\Lib\site-packages\seaborn_oldcore.py:1119:

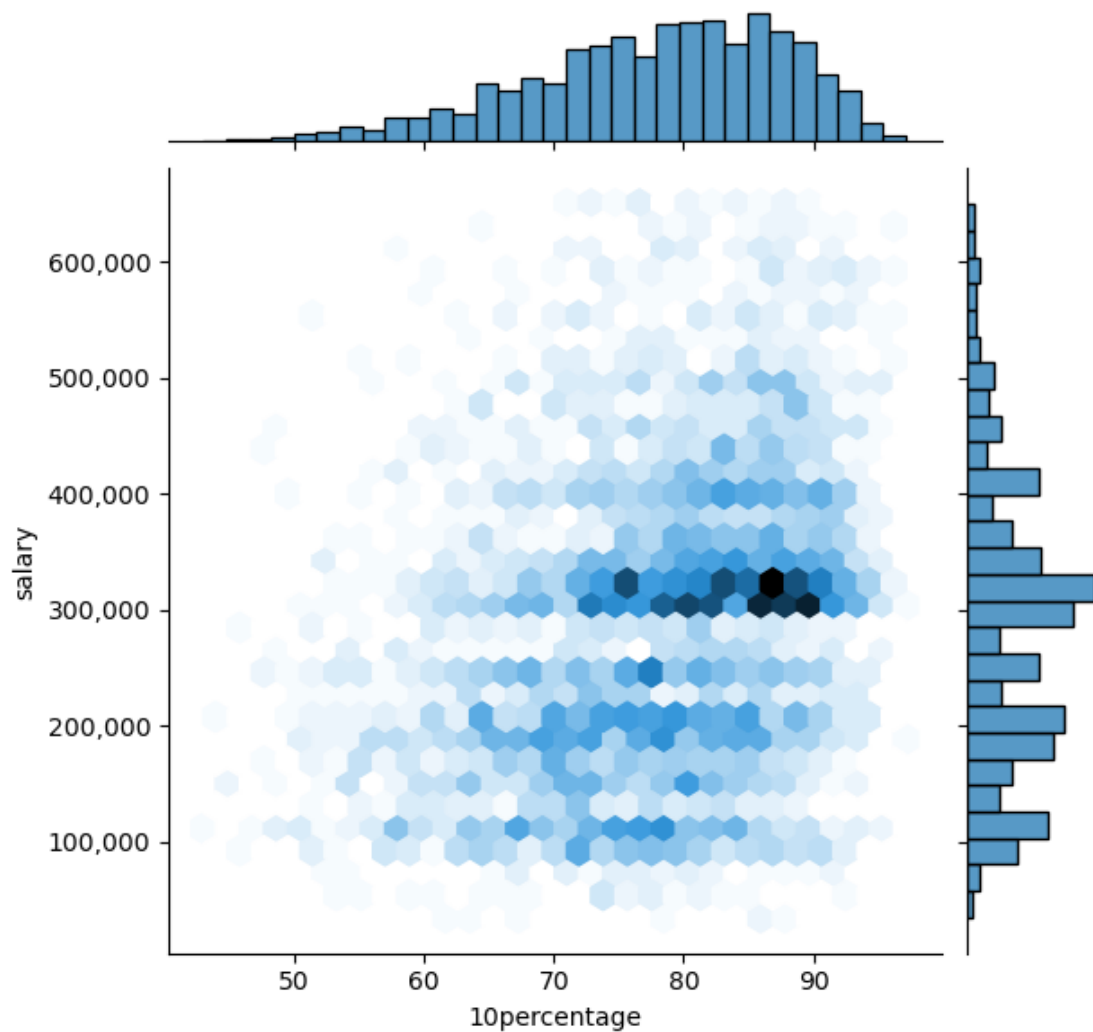
FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option_context('mode.use_inf_as_na', True):

C:\ProgramData\anaconda3\Lib\site-packages\seaborn_oldcore.py:1119:

FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

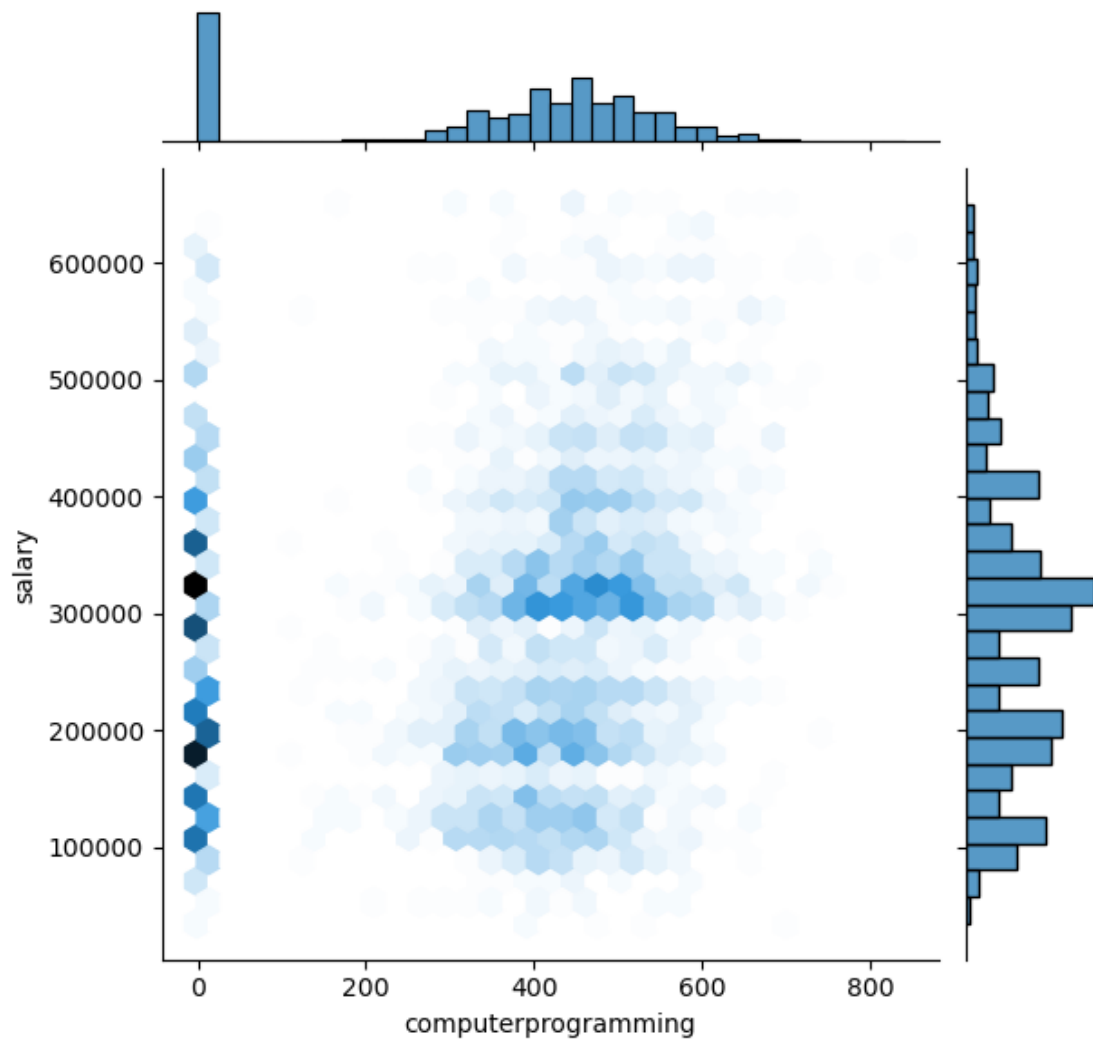
with pd.option_context('mode.use_inf_as_na', True):



```
[81]: sns.jointplot(x='computerprogramming',y='salary',data=data,kind='hex')
```

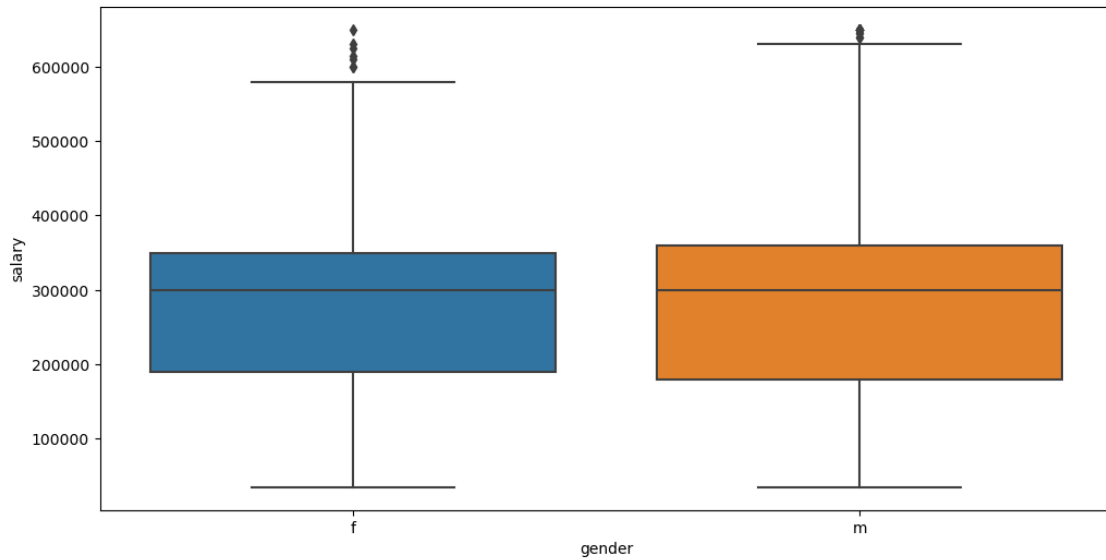
```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

```
[81]: <seaborn.axisgrid.JointGrid at 0x1fad2d20290>
```



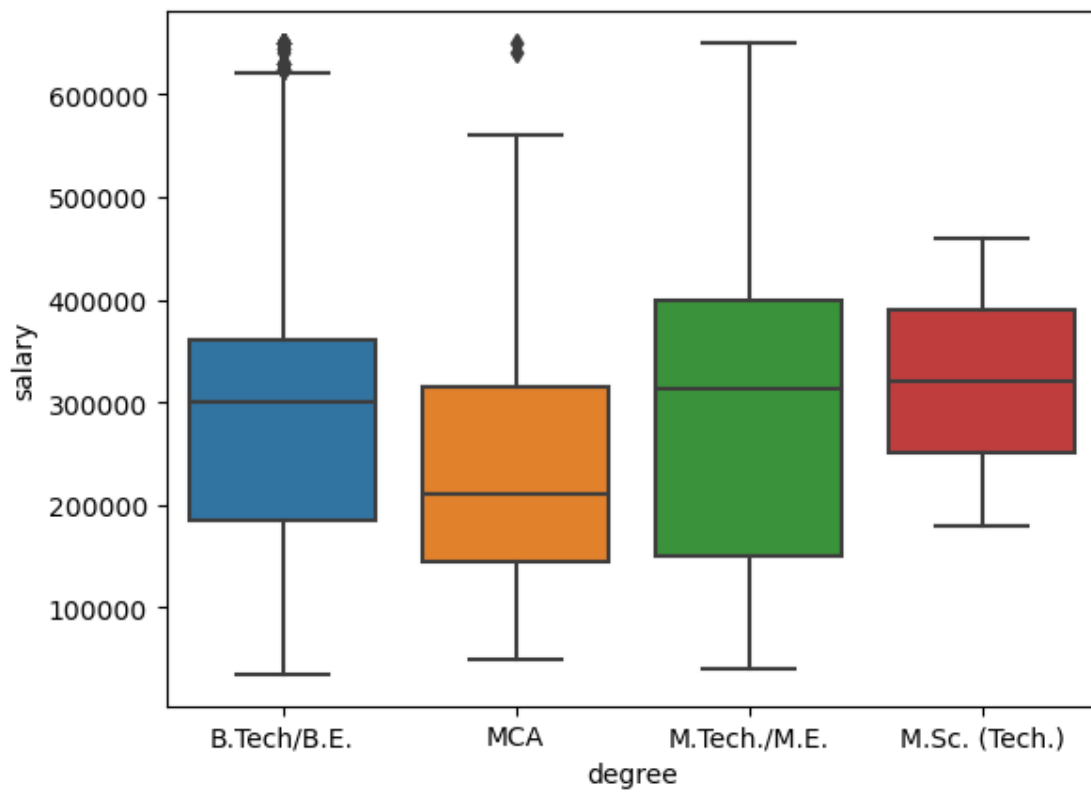
```
[102]: plt.figure(figsize=(12,6))
sns.boxplot(data=data,x='gender',y='salary')
```

```
[102]: <Axes: xlabel='gender', ylabel='salary'>
```



```
[104]: sns.boxplot(data=data,x='degree',y='salary')
```

```
[104]: <Axes: xlabel='degree', ylabel='salary'>
```



```
[105]: data.head()
```

```
[105]:      id      salary      doj      dol      designation \
0  203097  420000.0  6/1/12 0:00    present  senior quality engineer
1  579905  500000.0  9/1/13 0:00    present      assistant manager
2  810601  325000.0  6/1/14 0:00    present      systems engineer
4  343523  200000.0  3/1/14 0:00  3/1/15 0:00      get
5  1027655  300000.0  6/1/14 0:00    present      system engineer

      jobcity gender      dob  10percentage \
0  Bangalore      f  2/19/90 0:00      84.30
1    Indore      m  10/4/89 0:00      85.40
2   Chennai      f   8/3/92 0:00      85.00
4   Manesar      m  2/27/91 0:00      78.00
5  Hyderabad      m   7/2/92 0:00      89.92

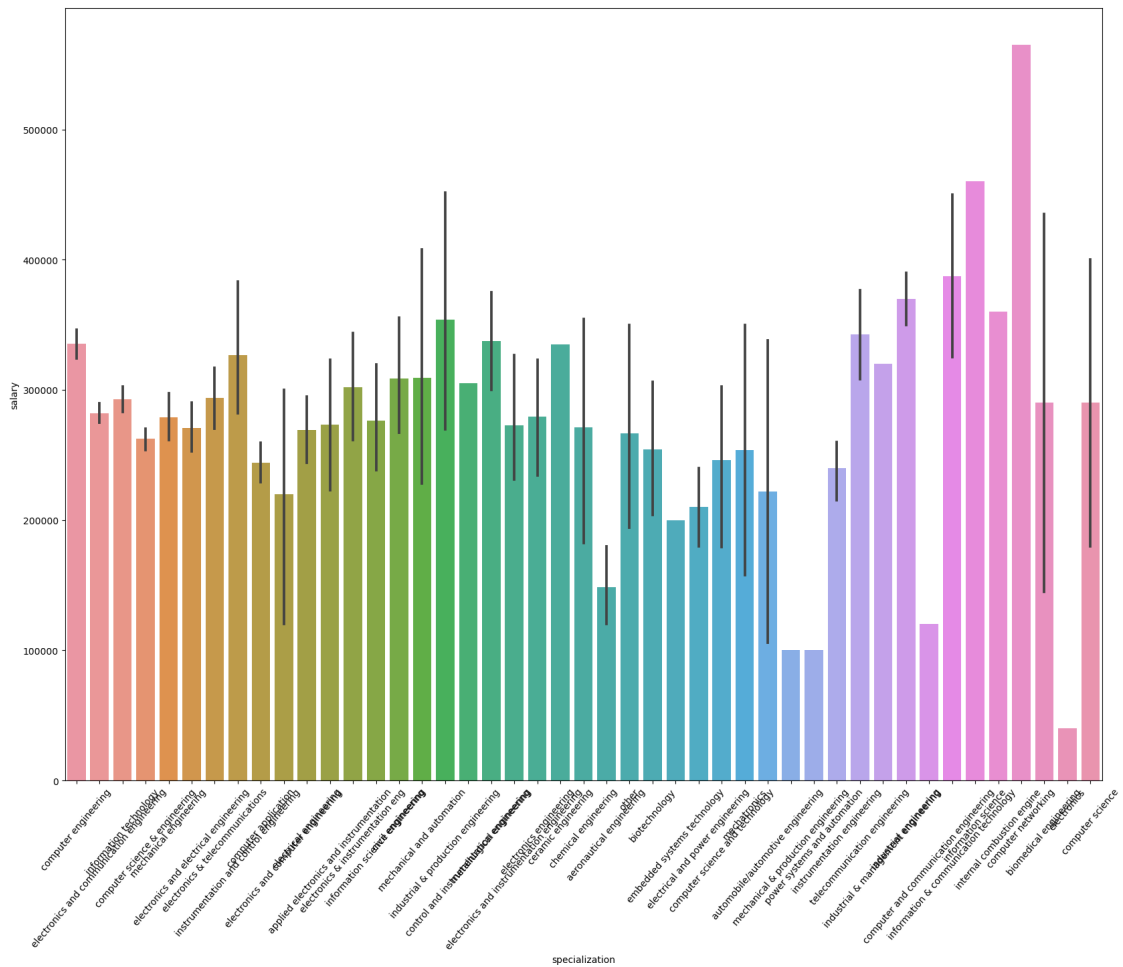
      10board ...  computerscience  mechanicalengg \
0  board ofsecondary education,ap ...      -1      -1
1      cbse ...      -1      -1
2      cbse ...      -1      -1
4      cbse ...      -1      -1
5      state board ...      407      -1

      electricalengg  telecomengg  civilengg  conscientiousness  agreeableness \
0      -1      -1      -1      0.9737      0.8128
1      -1      -1      -1     -0.7335      0.3789
2      -1      -1      -1      0.2718      1.7109
4      -1      -1      -1     -0.8810     -0.2793
5      -1      -1      -1     -0.3027     -0.6201

      extraversion  nueroticism  openness_to_experience
0      0.5269      1.35490      -0.4455
1      1.2396     -0.10760      0.8637
2      0.1637     -0.86820      0.6721
4     -1.0697      0.09163     -0.1295
5     -2.2954     -0.74150     -0.8608
```

```
[5 rows x 38 columns]
```

```
[122]: plt.figure(figsize=(20,15))
sns.barplot(data=data,x='specialization',y='salary')
plt.xticks(rotation=50)
plt.show()
```



[]:

[]: