eda-innomatics-project

February 20, 2024

[1]:
```
# importing the necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import warnings
warnings.filterwarnings("ignore")
%matplotlib inline
```

[2]:
```
# loading the dataset
df=pd.read_csv("data.xlsx - Sheet1.csv")
```

1 description of dataset

'ID': Identification number of the individuals.

'Salary': Salary of the individuals.

'DOJ': Date of joining.

'DOL': Date of leaving (if applicable).

'Designation': Job designation or title.

'JobCity': City where the job is located.

'Gender': Gender of the individuals.

'DOB': Date of birth.

'10percentage': Percentage obtained in 10th grade.

'10board': Board of education for 10th grade.

‘12graduation’: Year of graduation from 12th grade.

‘12percentage’: Percentage obtained in 12th grade.

‘12board’: Board of education for 12th grade.

‘CollegeID’: Identification number of the college.

‘CollegeTier’: Tier of the college.

‘Degree’: Type of degree obtained.

‘Specialization’: Field of specialization.

‘collegeGPA’: GPA obtained during college.

‘CollegeCityID’: Identification number of the college city.

‘CollegeCityTier’: Tier of the college city.

‘CollegeState’: State where the college is located.

‘GraduationYear’: Year of graduation from college.

‘English’: Score in English proficiency.

‘Logical’: Score in logical reasoning.

‘Quant’: Score in quantitative ability.

‘Domain’: Domain knowledge score.

‘ComputerProgramming’: Score in computer programming.

‘ElectronicsAndSemicon’: Score in electronics and semiconductor knowledge.

‘ComputerScience’: Score in computer science knowledge.

‘MechanicalEngg’: Score in mechanical engineering knowledge.

‘ElectricalEngg’: Score in electrical engineering knowledge.

‘TelecomEngg’: Score in telecom engineering knowledge.

‘CivilEngg’: Score in civil engineering knowledge.

'conscientiousness': Level of conscientiousness trait.

'agreeableness': Level of agreeableness trait.

'extraversion': Level of extraversion trait.

'nueroticism': Level of neuroticism trait.

'openess_to_experience': Level of openness to experience trait

[4]:
```
#head of dataset
df.head()
```

[4]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	…	ComputerScience	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	conscientiousness	agreeableness	extraversion	nueroticism	openess_to_experience
0	train	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	…	-1	-1	-1	-1	-1	0.9737	0.8128	0.5269	1.35490	-0.4455
1	train	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	…	-1	-1	-1	-1	-1	-0.7335	0.3789	1.2396	-0.10760	0.8637
2	train	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	…	-1	-1	-1	-1	-1	0.2718	1.7109	0.1637	-0.86820	0.6721
3	train	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	…	-1	-1	-1	-1	-1	0.0464	0.3448	-0.3440	-0.40780	-0.9194
4	train	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.0	…	-1	-1	-1	-1	-1	-0.8810	-0.2793	-1.0697	0.09163	-0.1295

```
[5 rows x 39 columns]
```

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

[5]:
```
     Unnamed: 0      ID    Salary           DOJ           DOL  \
3993      train   47916  280000.0  10/1/11 0:00  10/1/12 0:00
3994      train  752781  100000.0   7/1/13 0:00   7/1/13 0:00
3995      train  355888  320000.0   7/1/13 0:00       present
3996      train  947111  200000.0   7/1/14 0:00   1/1/15 0:00
3997      train  324966  400000.0   2/1/13 0:00       present

                     Designation           JobCity Gender           DOB  \
3993            software engineer        New Delhi       m  4/15/87 0:00
3994             technical writer         Hyderabad      f  8/27/92 0:00
3995  associate software engineer         Bangalore      m   7/3/91 0:00
3996          software developer  Asifabadbanglore      f  3/20/92 0:00
3997     senior systems engineer           Chennai      f  2/26/91 0:00

      10percentage  … ComputerScience  MechanicalEngg  ElectricalEngg  \
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  \
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
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

[5 rows x 39 columns]
```

[6]:
```
df.describe()
```

[6]:
```
                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]

[7]:
```
# as some of the columns are missing
df.describe(include="all")
```

[7]:

	Unnamed: 0	ID	Salary	DOJ	DOL	\
count	3998	3.998000e+03	3.998000e+03	3998	3998	
unique	1	NaN	NaN	81	67	
top	train	NaN	NaN	7/1/14 0:00	present	
freq	3998	NaN	NaN	199	1875	
mean	NaN	6.637945e+05	3.076998e+05	NaN	NaN	
std	NaN	3.632182e+05	2.127375e+05	NaN	NaN	
min	NaN	1.124400e+04	3.500000e+04	NaN	NaN	
25%	NaN	3.342842e+05	1.800000e+05	NaN	NaN	
50%	NaN	6.396000e+05	3.000000e+05	NaN	NaN	
75%	NaN	9.904800e+05	3.700000e+05	NaN	NaN	
max	NaN	1.298275e+06	4.000000e+06	NaN	NaN	

	Designation	JobCity	Gender	DOB	10percentage	…	\
count	3998	3998	3998	3998	3998.000000	…	
unique	419	339	2	1872	NaN	…	
top	software engineer	Bangalore	m	1/1/91 0:00	NaN	…	
freq	539	627	3041	11	NaN	…	
mean	NaN	NaN	NaN	NaN	77.925443	…	
std	NaN	NaN	NaN	NaN	9.850162	…	
min	NaN	NaN	NaN	NaN	43.000000	…	
25%	NaN	NaN	NaN	NaN	71.680000	…	
50%	NaN	NaN	NaN	NaN	79.150000	…	
75%	NaN	NaN	NaN	NaN	85.670000	…	
max	NaN	NaN	NaN	NaN	97.760000	…	

	ComputerScience	MechanicalEngg	ElectricalEngg	TelecomEngg	\
count	3998.000000	3998.000000	3998.000000	3998.000000	
unique	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	
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	
unique	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	

```
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
unique          NaN                    NaN
top             NaN                    NaN
freq            NaN                    NaN
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

[11 rows x 39 columns]
```

[8]:
```
df.shape
```

[8]:
```
(3998, 39)
```

1.0.1 that is in the give dataset we have 3998 rows and 39 features/columns

[9]:
```
# checking the types of data that are present in the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 39 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             3998 non-null   object
 1   ID                     3998 non-null   int64
 2   Salary                 3998 non-null   float64
 3   DOJ                    3998 non-null   object
 4   DOL                    3998 non-null   object
 5   Designation            3998 non-null   object
 6   JobCity                3998 non-null   object
 7   Gender                 3998 non-null   object
 8   DOB                    3998 non-null   object
 9   10percentage           3998 non-null   float64
```

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

[10]:
```
df.columns
```

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

from the dataset we can see that some columns has unique values which are not useful for our analysis

so let us remove that columns

[3]:
```
df.drop(["Unnamed: 0",'ID'],axis=1,inplace=True)
```

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

[4]:
```
      Salary          DOJ          DOL                Designation    JobCity  \
0   420000.0  6/1/12 0:00      present   senior quality engineer  Bangalore
1   500000.0  9/1/13 0:00      present         assistant manager     Indore
2   325000.0  6/1/14 0:00      present          systems engineer    Chennai
3  1100000.0  7/1/11 0:00      present  senior software engineer    Gurgaon
4   200000.0  3/1/14 0:00  3/1/15 0:00                       get    Manesar

  Gender           DOB  10percentage                         10board  \
0      f  2/19/90 0:00          84.3  board ofsecondary education,ap
1      m  10/4/89 0:00          85.4                            cbse
2      f   8/3/92 0:00          85.0                            cbse
3      m  12/5/89 0:00          85.6                            cbse
4      m  2/27/91 0:00          78.0                            cbse

   12graduation  …  ComputerScience  MechanicalEngg  ElectricalEngg  \
0          2007  …               -1              -1              -1
1          2007  …               -1              -1              -1
2          2010  …               -1              -1              -1
3          2007  …               -1              -1              -1
4          2008  …               -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

[5 rows x 37 columns]
```

[64]:
```
# check for missing values
df.isnull().sum()
```

[64]:
```
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
```

1.0.2 There are no null values in the dataset

2 Univariate Analysis

2.0.1 seperating the data into numerical and categorical columns

[5]:
```
num_feat=df.select_dtypes(include=["float64","int64"]).columns
```

[6]:
```
num_feat
```

[6]:
```
Index(['Salary', '10percentage', '12graduation', '12percentage', 'CollegeID',
       'CollegeTier', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier',
       'GraduationYear', 'English', 'Logical', 'Quant', 'Domain',
       'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience',
       'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg',
       'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism',
       'openess_to_experience'],
      dtype='object')
```

[11]:
```
def describe(num_data):
    for i in num_data:
        print("**********************")
        print(i)
        print(num_data[i].
 ↪agg(["min","max","mean","median","std","skew","kurt"]))
        print()

describe(df[num_feat])
```

```
**********************
Salary
min       3.500000e+04
max       4.000000e+06
mean      3.076998e+05
median    3.000000e+05
std       2.127375e+05
skew      6.451081e+00
kurt      8.093000e+01
Name: Salary, dtype: float64

**********************
10percentage
min       43.000000
max       97.760000
mean      77.925443
median    79.150000
std        9.850162
skew      -0.591019
kurt      -0.110284
Name: 10percentage, dtype: float64

**********************
12graduation
min       1995.000000
max       2013.000000
mean      2008.087544
median    2008.000000
```

```
std          1.653599
skew        -0.964090
kurt         1.951164
Name: 12graduation, dtype: float64

**********
12percentage
min       40.000000
max       98.700000
mean      74.466366
median    74.400000
std       10.999933
skew      -0.032607
kurt      -0.630737
Name: 12percentage, dtype: float64

**********
CollegeID
min           2.000000
max       18409.000000
mean       5156.851426
median     3879.000000
std        4802.261482
skew          0.649176
kurt         -0.767441
Name: CollegeID, dtype: float64

**********
CollegeTier
min       1.000000
max       2.000000
mean      1.925713
median    2.000000
std       0.262270
skew     -3.247991
kurt      8.553722
Name: CollegeTier, dtype: float64

**********
collegeGPA
min        6.450000
max       99.930000
mean      71.486171
median    71.720000
std        8.167338
skew      -1.249209
kurt      10.234244
Name: collegeGPA, dtype: float64
```

```
**********
CollegeCityID
min           2.000000
max       18409.000000
mean       5156.851426
median     3879.000000
std        4802.261482
skew          0.649176
kurt         -0.767441
Name: CollegeCityID, dtype: float64

**********
CollegeCityTier
min       0.000000
max       1.000000
mean      0.300400
median    0.000000
std       0.458489
skew      0.871120
kurt     -1.241771
Name: CollegeCityTier, dtype: float64

**********
GraduationYear
min          0.000000
max       2017.000000
mean      2012.105803
median    2013.000000
std         31.857271
skew       -63.068064
kurt      3984.369696
Name: GraduationYear, dtype: float64

**********
English
min       180.000000
max       875.000000
mean      501.649075
median    500.000000
std       104.940021
skew        0.191997
kurt       -0.254133
Name: English, dtype: float64

**********
Logical
min       195.000000
```

```
max       795.000000
mean      501.598799
median    505.000000
std        86.783297
skew       -0.216602
kurt       -0.224761
Name: Logical, dtype: float64

**********
Quant
min       120.000000
max       900.000000
mean      513.378189
median    515.000000
std       122.302332
skew       -0.019399
kurt       -0.102472
Name: Quant, dtype: float64

**********
Domain
min      -1.000000
max       0.999910
mean      0.510490
median    0.622643
std       0.468671
skew     -1.922146
kurt      3.895951
Name: Domain, dtype: float64

**********
ComputerProgramming
min        -1.000000
max       840.000000
mean      353.102801
median    415.000000
std       205.355519
skew       -0.778106
kurt       -0.666352
Name: ComputerProgramming, dtype: float64

**********
ElectronicsAndSemicon
min        -1.000000
max       612.000000
mean       95.328414
median     -1.000000
std       158.241218
```

```
skew        1.195975
kurt       -0.210374
Name: ElectronicsAndSemicon, dtype: float64

**********
ComputerScience
min        -1.000000
max       715.000000
mean       90.742371
median     -1.000000
std       175.273083
skew        1.529521
kurt        0.692641
Name: ComputerScience, dtype: float64

**********
MechanicalEngg
min        -1.000000
max       623.000000
mean       22.974737
median     -1.000000
std        98.123311
skew        4.029563
kurt       15.018957
Name: MechanicalEngg, dtype: float64

**********
ElectricalEngg
min        -1.000000
max       676.000000
mean       16.478739
median     -1.000000
std        87.585634
skew        5.060407
kurt       24.878194
Name: ElectricalEngg, dtype: float64

**********
TelecomEngg
min        -1.000000
max       548.000000
mean       31.851176
median     -1.000000
std       104.852845
skew        3.041261
kurt        7.810221
Name: TelecomEngg, dtype: float64
```

```
**********
CivilEngg
min        -1.000000
max       516.000000
mean        2.683842
median     -1.000000
std        36.658505
skew       10.315681
kurt      109.041349
Name: CivilEngg, dtype: float64

**********
conscientiousness
min      -4.126700
max       1.995300
mean     -0.037831
median    0.046400
std       1.028666
skew     -0.527003
kurt      0.122596
Name: conscientiousness, dtype: float64

**********
agreeableness
min      -5.781600
max       1.904800
mean      0.146496
median    0.212400
std       0.941782
skew     -1.204915
kurt      3.391242
Name: agreeableness, dtype: float64

**********
extraversion
min      -4.600900
max       2.535400
mean      0.002763
median    0.091400
std       0.951471
skew     -0.523267
kurt      0.643969
Name: extraversion, dtype: float64

**********
nueroticism
min      -2.643000
max       3.352500
```

```
mean     -0.169033
median   -0.234400
std       1.007580
skew      0.165710
kurt     -0.191539
Name: nueroticism, dtype: float64

**********
openess_to_experience
min      -7.375700
max       1.822400
mean     -0.138110
median   -0.094300
std       1.008075
skew     -1.506962
kurt      5.788327
Name: openess_to_experience, dtype: float64
```

[12]:
```
cat_feat=df.select_dtypes(include=["object"]).columns
```

[13]:
```
cat_feat
```

[13]:
```
Index(['DOJ', 'DOL', 'Designation', 'JobCity', 'Gender', 'DOB', '10board',
       '12board', 'Degree', 'Specialization', 'CollegeState'],
      dtype='object')
```

[15]:
```
def dis(cat_feat):
    for i in cat_feat:
        print("**************************")
        print(cat_feat[i].agg(["count","nunique","unique"]))
        print("valuecounts\n",cat_feat[i].value_counts(normalize=True))
        print()
dis(df[cat_feat])
```

```
*************************
count                                                 3998
nunique                                                 81
unique     [6/1/12 0:00, 9/1/13 0:00, 6/1/14 0:00, 7/1/11…
Name: DOJ, dtype: object
valuecounts
 7/1/14 0:00     0.049775
6/1/14 0:00      0.045023
8/1/14 0:00      0.044522
9/1/14 0:00      0.035518
1/1/14 0:00      0.035518
                   …
11/1/15 0:00     0.000250
```

```
11/1/09 0:00    0.000250
8/1/04 0:00     0.000250
9/1/09 0:00     0.000250
2/1/07 0:00     0.000250
Name: DOJ, Length: 81, dtype: float64

*************************
count                                                  3998
nunique                                                  67
unique     [present, 3/1/15 0:00, 5/1/15 0:00, 7/1/15 0:0…
Name: DOL, dtype: object
valuecounts
 present         0.468984
4/1/15 0:00     0.143322
3/1/15 0:00     0.031016
5/1/15 0:00     0.028014
1/1/15 0:00     0.024762
                  …
3/1/05 0:00     0.000250
10/1/15 0:00    0.000250
2/1/10 0:00     0.000250
2/1/11 0:00     0.000250
10/1/10 0:00    0.000250
Name: DOL, Length: 67, dtype: float64

*************************
count                                                  3998
nunique                                                 419
unique     [senior quality engineer, assistant manager, s…
Name: Designation, dtype: object
valuecounts
 software engineer                    0.134817
software developer                   0.066283
system engineer                      0.051276
programmer analyst                   0.034767
systems engineer                     0.029515
                                       …
cad drafter                          0.000250
noc engineer                         0.000250
human resources intern               0.000250
senior quality assurance engineer    0.000250
jr. software developer               0.000250
Name: Designation, Length: 419, dtype: float64

*************************
count                                                  3998
nunique                                                 339
unique     [Bangalore, Indore, Chennai, Gurgaon, Manesar,…
```

```
Name: JobCity, dtype: object
valuecounts
 Bangalore           0.156828
-1                   0.115308
Noida                0.092046
Hyderabad            0.083792
Pune                 0.072536
                       …
Tirunelvelli         0.000250
Ernakulam            0.000250
Nanded               0.000250
Dharmapuri           0.000250
Asifabadbanglore     0.000250
Name: JobCity, Length: 339, dtype: float64

**************************
count         3998
nunique          2
unique      [f, m]
Name: Gender, dtype: object
valuecounts
 m     0.76063
f     0.23937
Name: Gender, dtype: float64

**************************
count                                                    3998
nunique                                                  1872
unique      [2/19/90 0:00, 10/4/89 0:00, 8/3/92 0:00, 12/5…
Name: DOB, dtype: object
valuecounts
 1/1/91 0:00       0.002751
7/15/91 0:00      0.002501
7/5/91 0:00       0.002001
12/13/91 0:00     0.002001
6/3/91 0:00       0.002001
                    …
12/30/92 0:00     0.000250
10/20/86 0:00     0.000250
11/17/89 0:00     0.000250
9/30/92 0:00      0.000250
4/15/87 0:00      0.000250
Name: DOB, Length: 1872, dtype: float64

**************************
count                                                    3998
nunique                                                   275
unique      [board ofsecondary education,ap, cbse, state b…
```

```
Name: 10board, dtype: object
valuecounts
 cbse                          0.348924
state board                   0.291146
0                             0.087544
icse                          0.070285
ssc                           0.030515
                                 …
hse,orissa                    0.000250
national public school        0.000250
nagpur board                  0.000250
jharkhand academic council    0.000250
bse,odisha                    0.000250
Name: 10board, Length: 275, dtype: float64

*************************
count                                                    3998
nunique                                                   340
unique     [board of intermediate education,ap, cbse, sta…
Name: 12board, dtype: object
valuecounts
 cbse                                0.350175
state board                         0.313657
0                                   0.089795
icse                                0.032266
up board                            0.021761
                                       …
jawahar higher secondary school     0.000250
nagpur board                        0.000250
bsemp                               0.000250
board of higher secondary orissa    0.000250
boardofintermediate                 0.000250
Name: 12board, Length: 340, dtype: float64

*************************
count                                                  3998
nunique                                                   4
unique     [B.Tech/B.E., MCA, M.Tech./M.E., M.Sc. (Tech.)]
Name: Degree, dtype: object
valuecounts
 B.Tech/B.E.      0.925463
MCA               0.060780
M.Tech./M.E.      0.013257
M.Sc. (Tech.)     0.000500
Name: Degree, dtype: float64

*************************
count                                                  3998
```

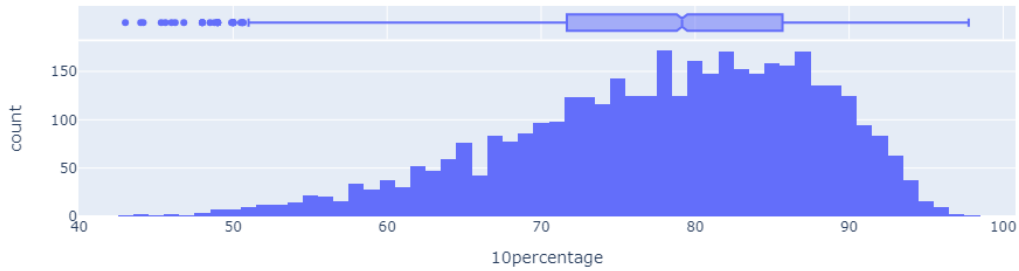
```
nunique                                                    46
unique     [computer engineering, electronics and communi…
Name: Specialization, dtype: object
valuecounts
 electronics and communication engineering      0.220110
computer science & engineering                  0.186093
information technology                          0.165083
computer engineering                            0.150075
computer application                            0.061031
mechanical engineering                          0.050275
electronics and electrical engineering          0.049025
electronics & telecommunications                0.030265
electrical engineering                          0.020510
electronics & instrumentation eng               0.008004
civil engineering                               0.007254
electronics and instrumentation engineering     0.006753
information science engineering                 0.006753
instrumentation and control engineering         0.005003
electronics engineering                         0.004752
biotechnology                                   0.003752
other                                           0.003252
industrial & production engineering             0.002501
applied electronics and instrumentation         0.002251
chemical engineering                            0.002251
computer science and technology                 0.001501
telecommunication engineering                   0.001501
mechanical and automation                       0.001251
automobile/automotive engineering               0.001251
instrumentation engineering                     0.001001
mechatronics                                    0.001001
aeronautical engineering                        0.000750
electronics and computer engineering            0.000750
electrical and power engineering                0.000500
biomedical engineering                          0.000500
information & communication technology          0.000500
industrial engineering                          0.000500
computer science                                0.000500
metallurgical engineering                       0.000500
power systems and automation                    0.000250
control and instrumentation engineering         0.000250
mechanical & production engineering             0.000250
embedded systems technology                     0.000250
polymer technology                              0.000250
computer and communication engineering          0.000250
information science                             0.000250
internal combustion engine                      0.000250
computer networking                             0.000250
ceramic engineering                             0.000250
```

```
electronics                                        0.000250
industrial & management engineering                0.000250
Name: Specialization, dtype: float64

**************************
count                                                3998
nunique                                                26
unique     [Andhra Pradesh, Madhya Pradesh, Uttar Pradesh…
Name: CollegeState, dtype: object
valuecounts
 Uttar Pradesh        0.228864
Karnataka             0.092546
Tamil Nadu            0.091796
Telangana             0.079790
Maharashtra           0.065533
Andhra Pradesh        0.056278
West Bengal           0.049025
Punjab                0.048274
Madhya Pradesh        0.047274
Haryana               0.045023
Rajasthan             0.043522
Orissa                0.043022
Delhi                 0.040520
Uttarakhand           0.028264
Kerala                0.008254
Jharkhand             0.007004
Chhattisgarh          0.006753
Gujarat               0.006003
Himachal Pradesh      0.004002
Bihar                 0.002501
Jammu and Kashmir     0.001751
Assam                 0.001251
Union Territory       0.001251
Sikkim                0.000750
Meghalaya             0.000500
Goa                   0.000250
Name: CollegeState, dtype: float64
```

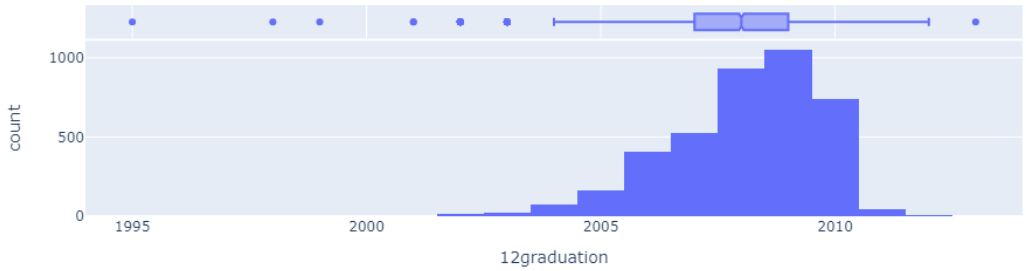
[26]:
```
for column in num_feat:
    fig = px.histogram(df, x=column, marginal='box', title=f'Histogram and␣
 ↪Boxplot for {column}')
    fig.show()
```

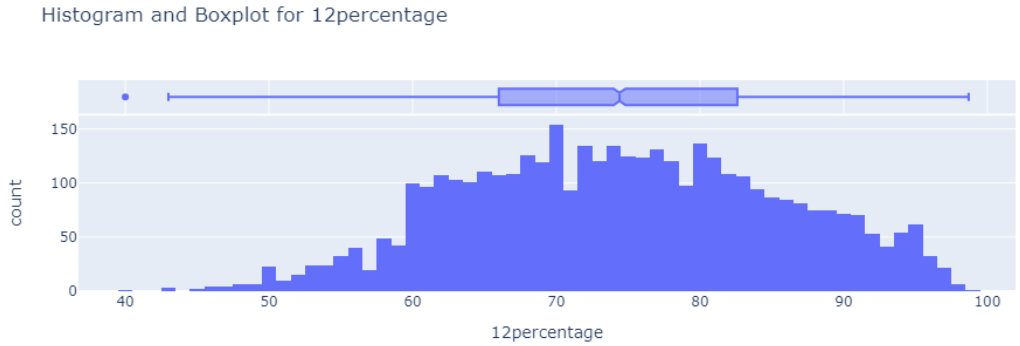
Histogram and Boxplot for Salary

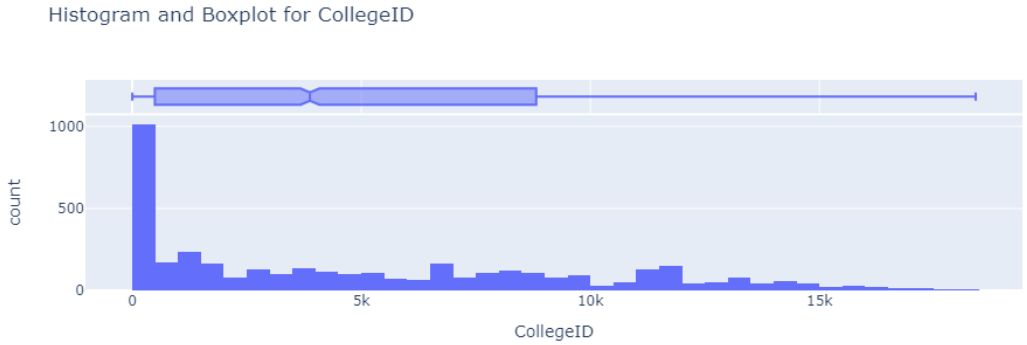
Histogram and Boxplot for 10percentage

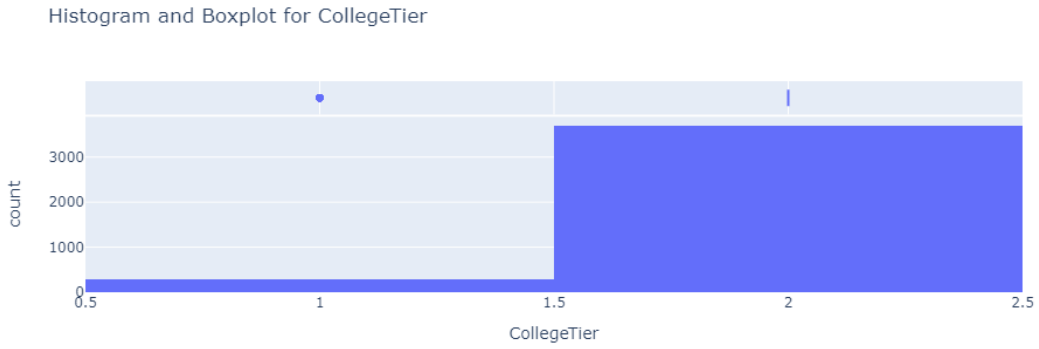


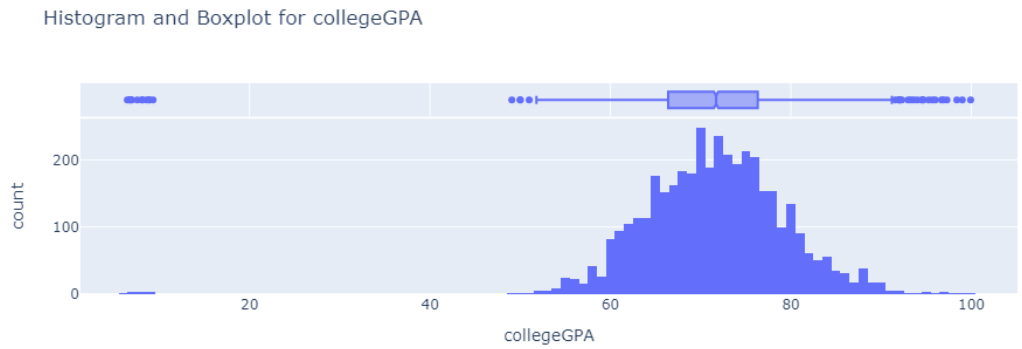
Histogram and Boxplot for 12graduation

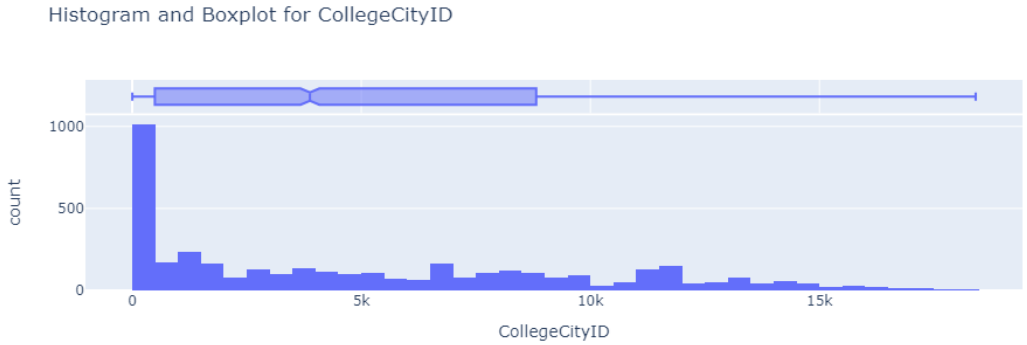


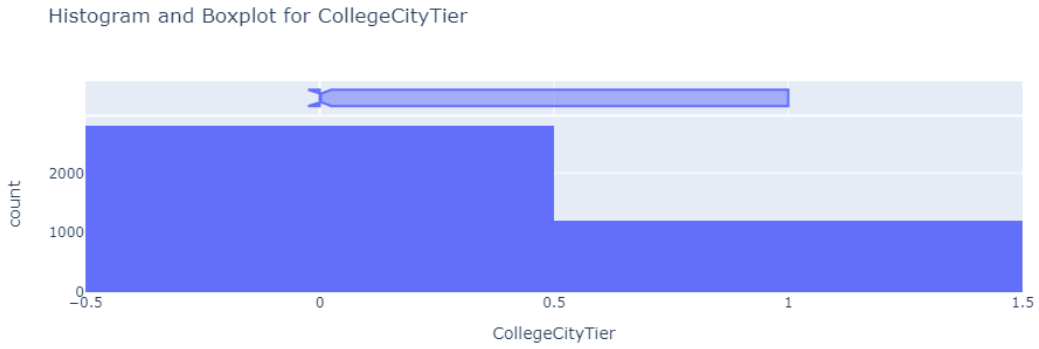

Histogram and Boxplot for 12percentage
count
150
100
50
0
40
50
60
70
80
90
100
12percentage


Histogram and Boxplot for CollegeID
count
1000
500
0
0
5k
10k
15k
CollegeID

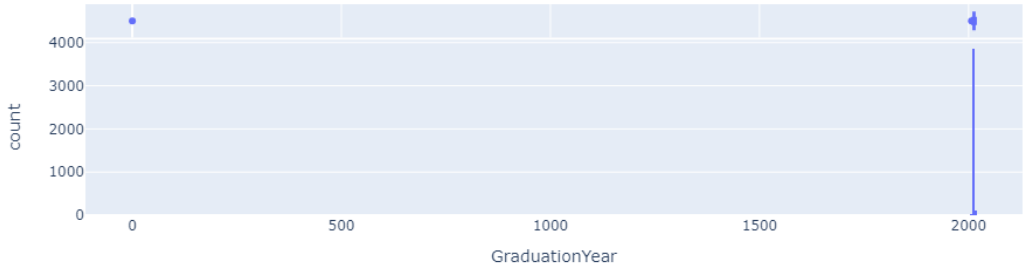

Histogram and Boxplot for CollegeTier
count
3000
2000
1000
0
0.5
1
1.5
2
2.5
CollegeTier


Histogram and Boxplot for collegeGPA
count
200
100
0
20
40
60
80
100
collegeGPA

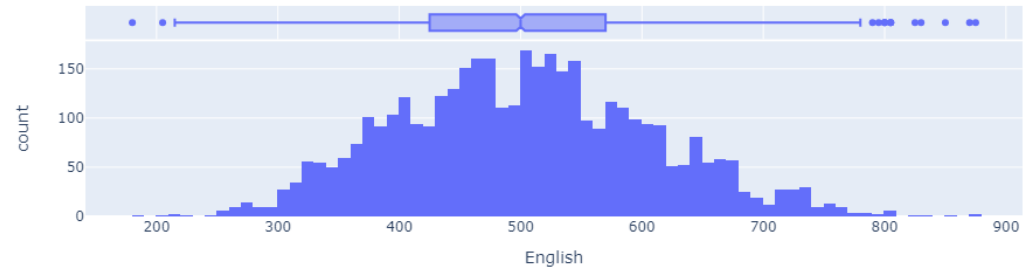

Histogram and Boxplot for CollegeCityID
count
1000
500
0
0
5k
10k
15k
CollegeCityID


Histogram and Boxplot for CollegeCityTier
count
2000
1000
0
−0.5
0
0.5
1
1.5
CollegeCityTier

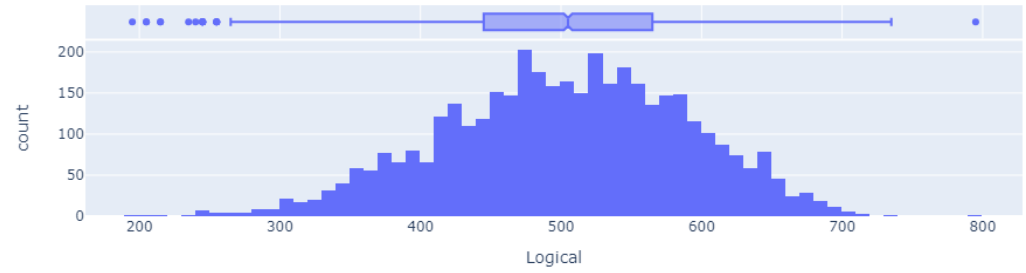
Histogram and Boxplot for GraduationYear



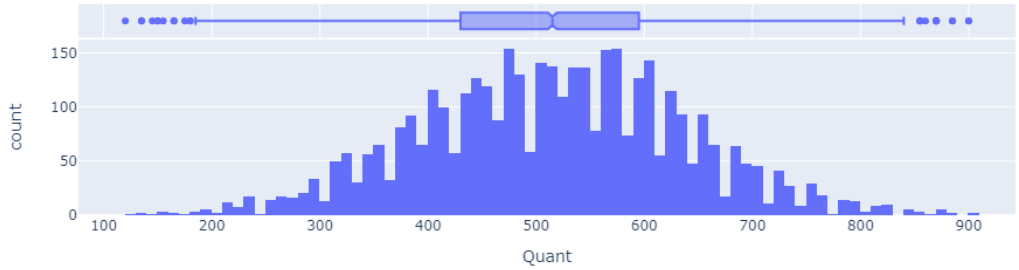
Histogram and Boxplot for English



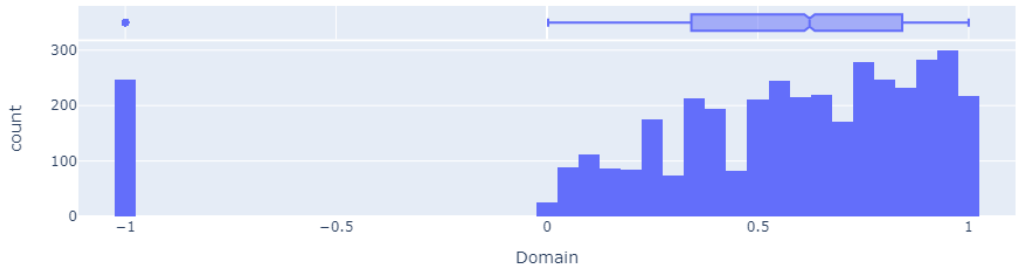
Histogram and Boxplot for Logical



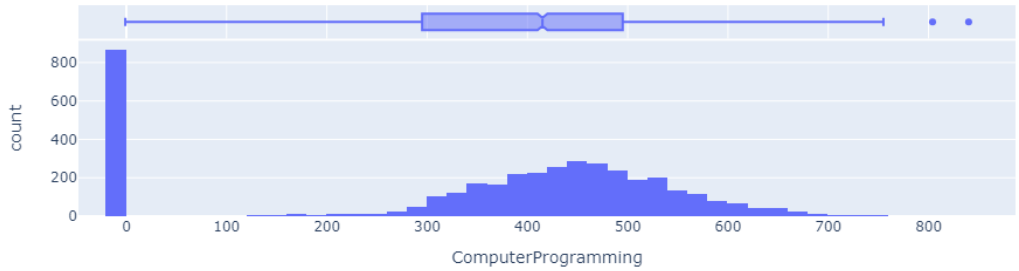
Histogram and Boxplot for Quant

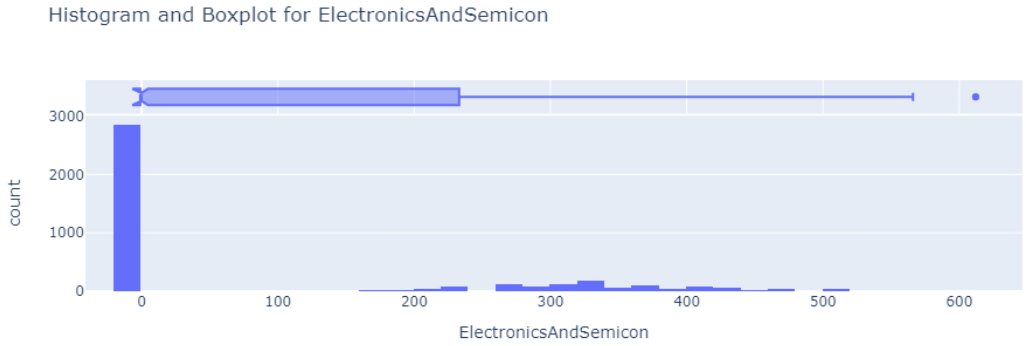


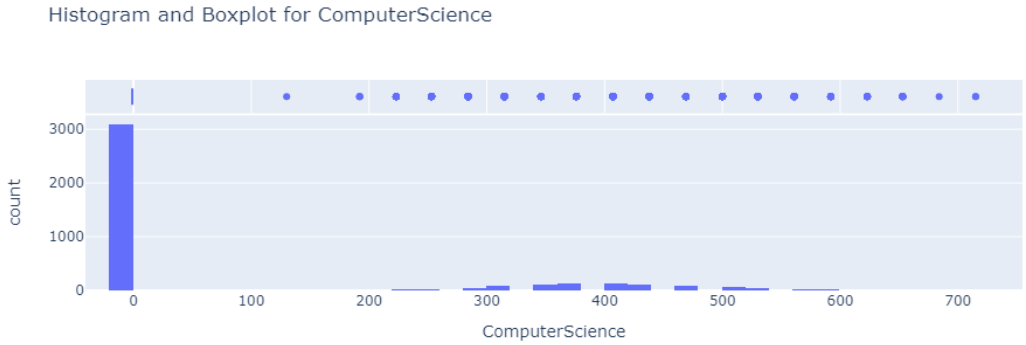
Histogram and Boxplot for Domain

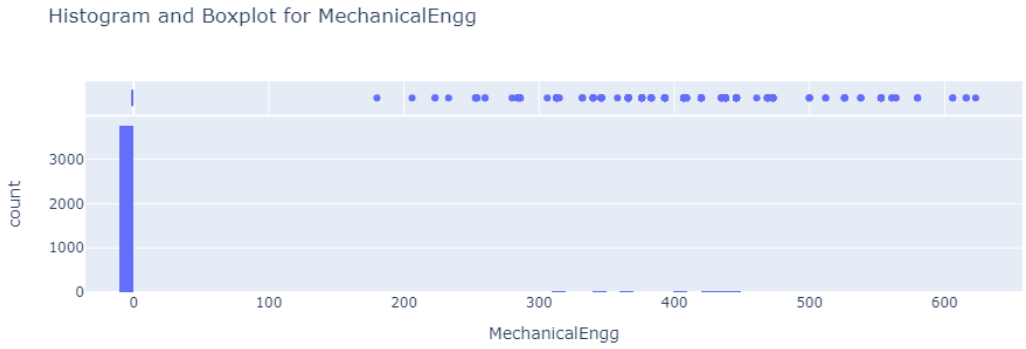


Histogram and Boxplot for ComputerProgramming

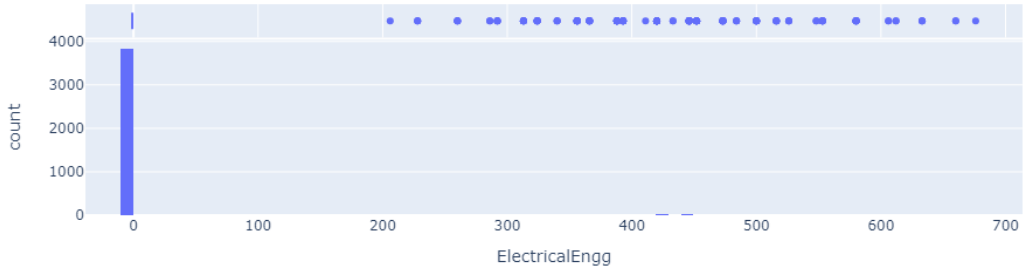



Histogram and Boxplot for ElectronicsAndSemicon
count
3000
2000
1000
0
0
100
200
300
400
500
600
ElectronicsAndSemicon

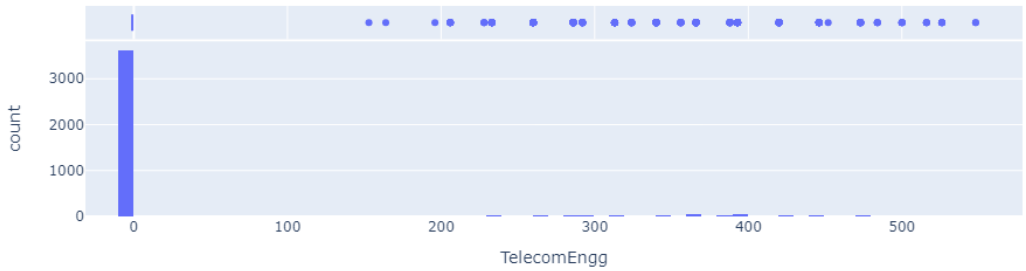

Histogram and Boxplot for ComputerScience
count
3000
2000
1000
0
0
100
200
300
400
500
600
700
ComputerScience


Histogram and Boxplot for MechanicalEngg
count
3000
2000
1000
0
0
100
200
300
400
500
600
MechanicalEngg

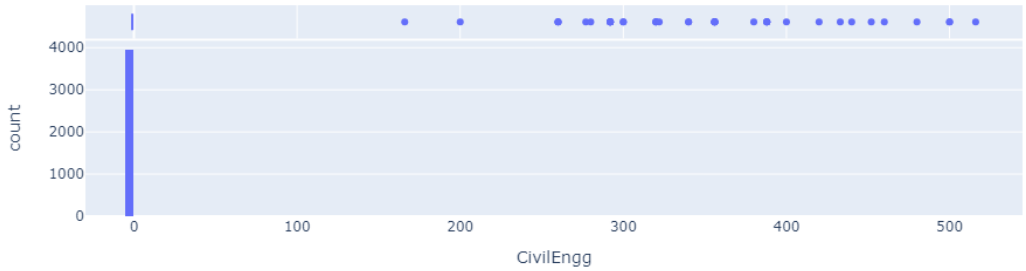
Histogram and Boxplot for ElectricalEngg

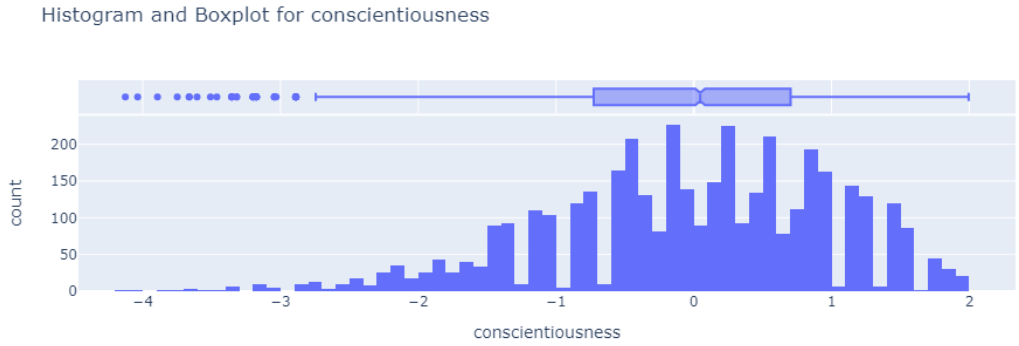


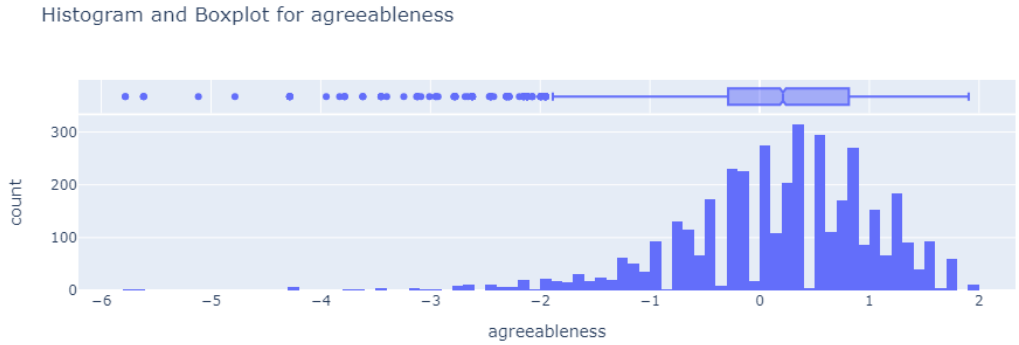
Histogram and Boxplot for TelecomEngg

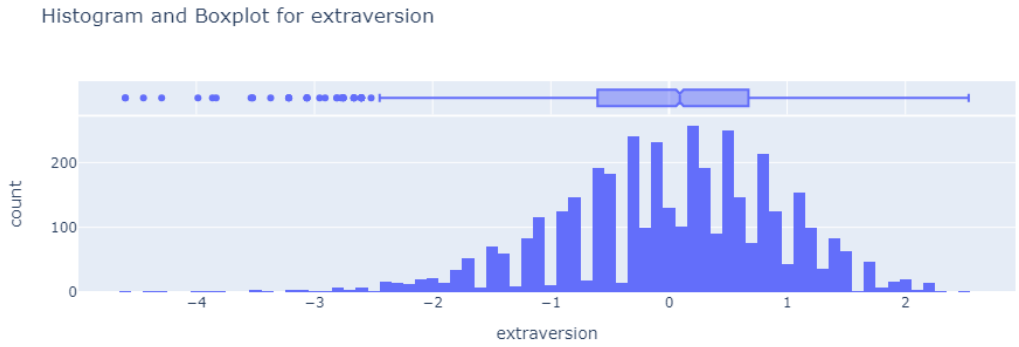


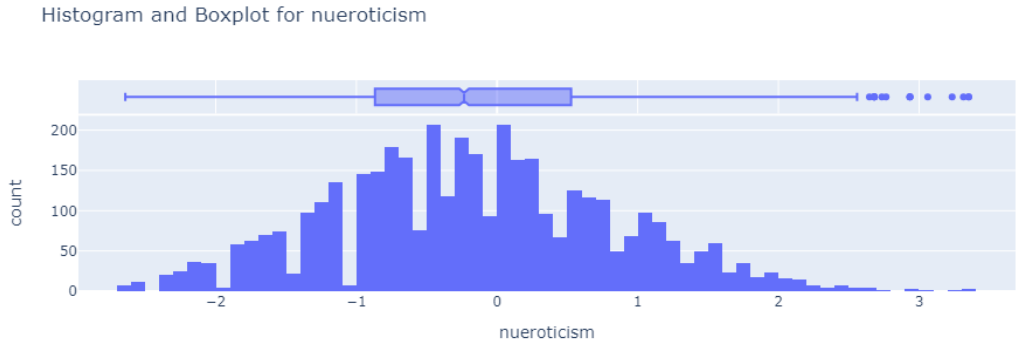
Histogram and Boxplot for CivilEngg

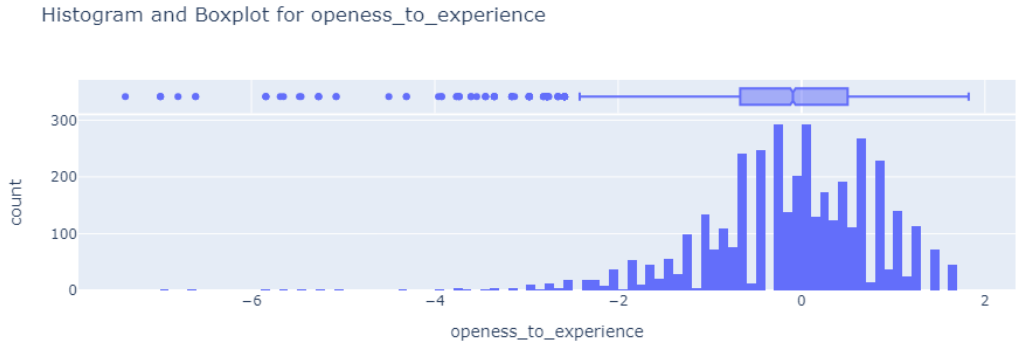



Histogram and Boxplot for conscientiousness
count
200
150
100
50
0
−4
−3
−2
−1
0
1
2
conscientiousness


Histogram and Boxplot for agreeableness
count
300
200
100
0
−6
−5
−4
−3
−2
−1
0
1
2
agreeableness


Histogram and Boxplot for extraversion
count
200
100
0
−4
−3
−2
−1
0
1
2
extraversion


Histogram and Boxplot for nueroticism
count
200
150
100
50
0
−2
−1
0
1
2
3
nueroticism


Histogram and Boxplot for openess_to_experience
count
300
200
100
0
−6
−4
−2
0
2
openess_to_experience

2.1 Summary from the above plots

2.1.1 1. In salary feature we have most outliers which has minimum value of 35k and maximum value of 4million

2.1.2 The average salary of most of them is 300k

2.1.3 2. In 10th percentage feature we have some outliers

2.1.4 The minimum percentage is 43 and the maximum percentage is 97.76 and the average percentage of all of them is 79.15

2.1.5 3. In 12th graduation feature we have very less no.of outliers

2.1.6 maximum number of students were graduated in the year of 2008

2.1.7 4. In 12th percent feature we can see that the highest percentage secured by a student is 98.7

2.1.8 and the average % is 74.4

2.1.9 5. we can see that majority of the college students are from tier 2 colleges

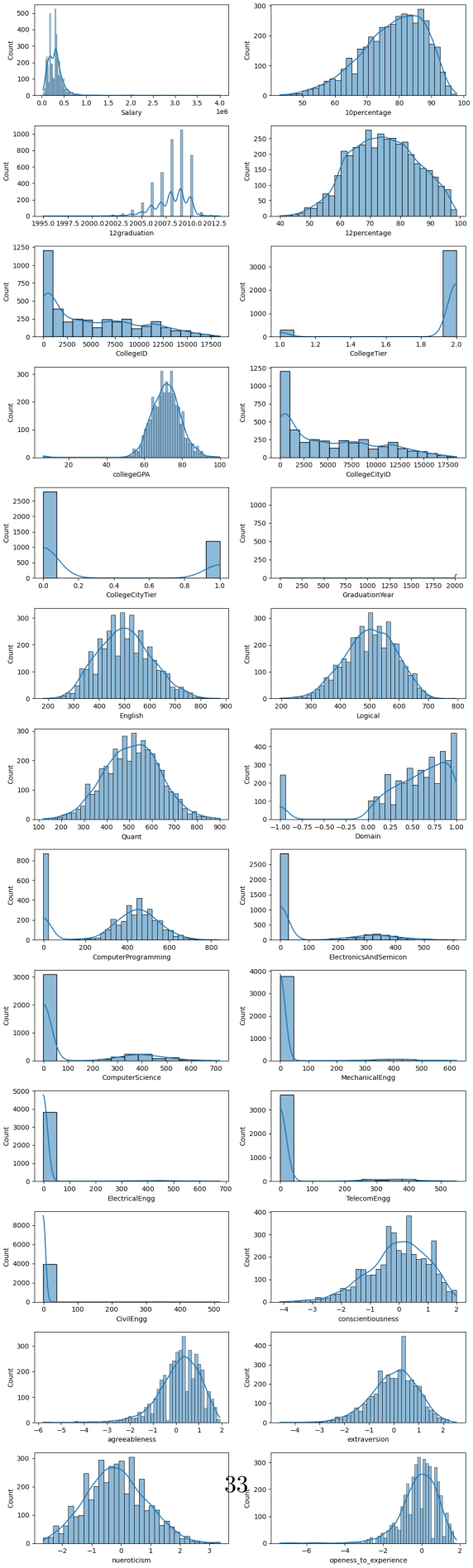
2.1.10 6. The average percentage in college is 71.72

2.1.11 7. we can see that all the students are graduated in the year 2000's

2.1.12 8. we can see that in average all the students are good in english,logical ans quant

2.1.13 9. we can see that when compare to all less than 50% of students are comfortable in computer programming

[52]:
```
# now let us understand the distributions each numerical feature
plt.figure(figsize=(10,50))
for i in range (0,len(num_feat)):
    plt.subplot(20,2,i+1)
    sns.histplot(data=df,x=df[num_feat[i]],kde=True)
plt.tight_layout()
```


Count
Salary
1e6
10percentage
12graduation
12percentage
CollegeID
CollegeTier
collegeGPA
CollegeCityID
CollegeCityTier
GraduationYear
English
Logical
Quant
Domain
ComputerProgramming
ElectronicsAndSemicon
ComputerScience
MechanicalEngg
ElectricalEngg
TelecomEngg
CivilEngg
conscientiousness
agreeableness
extraversion
nueroticism
openess_to_experience

3 Understanding the distributions followed by numerical features

3.0.1 1. the salary feature is right skewed

3.0.2 2. The 10th percentage is slightly similar to normal distribution but it is slightly left skewed

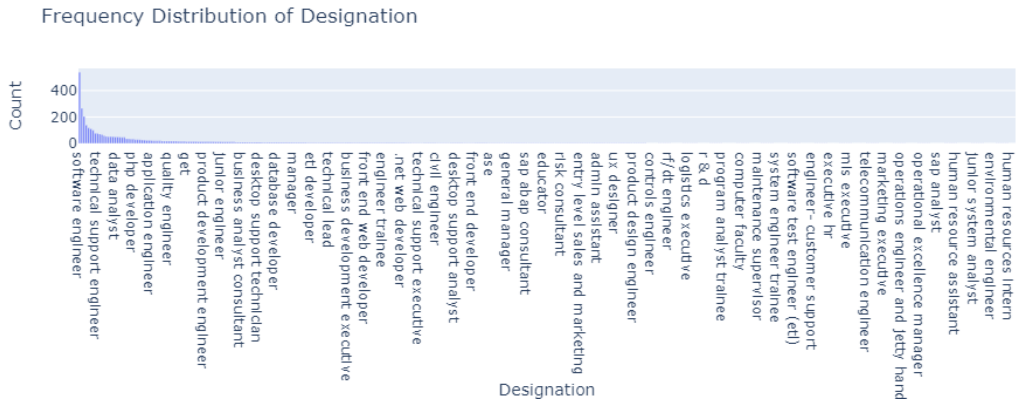
3.0.3 3. The feature 12 graduation is mostly left skewed

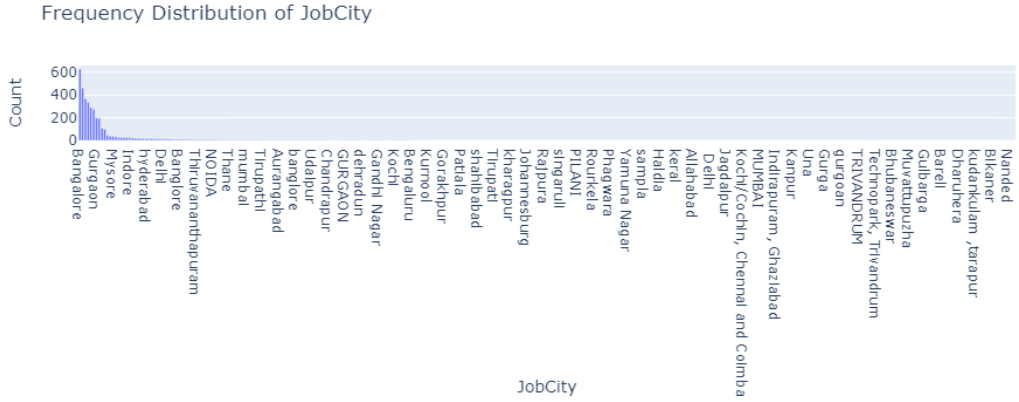
3.0.4 4. the feature 12 percentage is very near to normal distrinution with slight left skewness

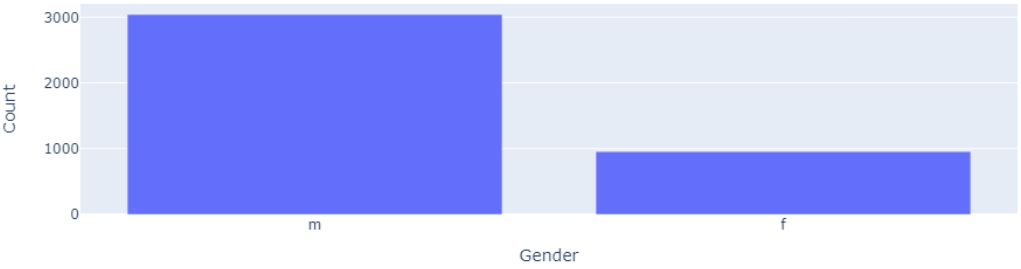
3.0.5 5. The features English,Quant,logical,nueroticism follows normal distribution

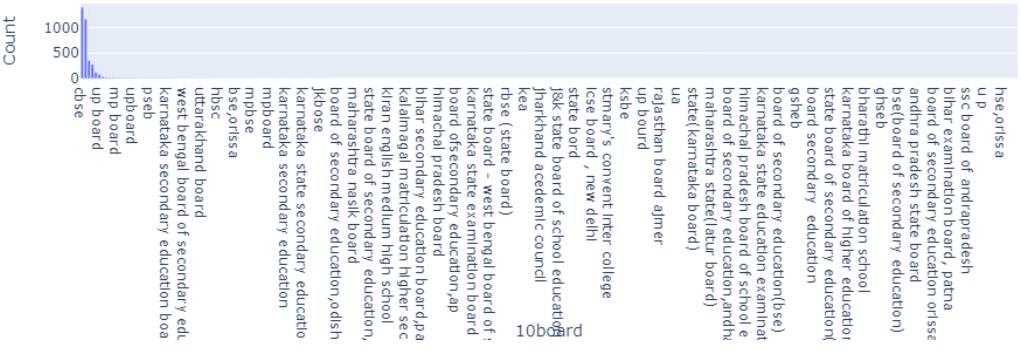
3.0.6 6. The features openess_to_experience,agreeableness,extraversion are left skewed

[62]:
```
for column in cat_feat:
    fig = px.bar(df[column].value_counts().reset_index(), x='index', y=column,␣
 ↪title=f'Frequency Distribution of {column}')
    fig.update_xaxes(title=column)
    fig.update_yaxes(title='Count')
    fig.show()
```

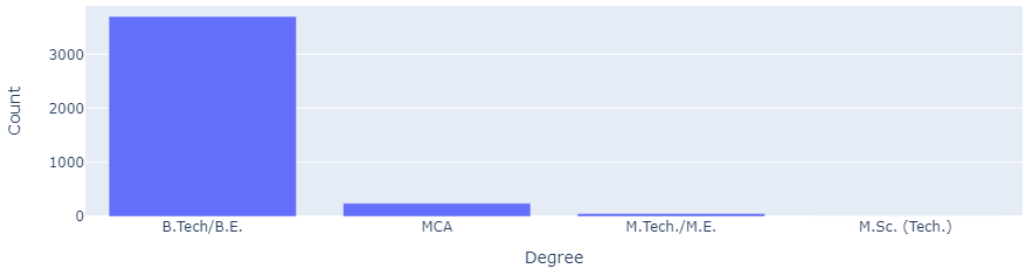



Frequency Distribution of JobCity
Count
600
400
200
0
Bangalore
Gurgaon
Mysore
Indore
Hyderabad
Delhi
Banglore
Thiruvananthapuram
NOIDA
Thane
mumbai
Tirupathi
Aurangabad
banglore
Udaipur
Chandrapur
GURGAON
dehradun
Gandhi Nagar
Kochi
Bengaluru
Kurnool
Tirupati
khargpur
Johannesburg
shahibabad
Patiala
Gorakhpur
Rajpura
singarull
PILANI
Rourkela
Phagwara
Yamuna Nagar
sampla
Haldia
keral
Allahabad
Delhi
Jagdalpur
Kochi/Cochin, Chennai and Coimba
MUMBAI
Indirapuram, Ghaziabad
Kampur
Una
Gurga
gurgoan
TRIVANDRUM
Technopark, Trivandrum
Bhubaneswar
Muvattupuzha
Gulbarga
Bareli
Dharuhera
kudankulam ,tarapur
Bikaner
Nanded
JobCity

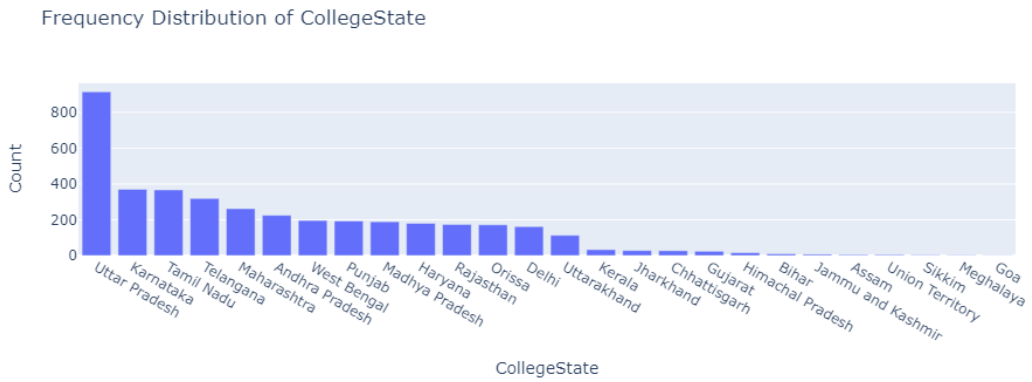

Frequency Distribution of Gender
Count
3000
2000
1000
0
m
f
Gender


Frequency Distribution of 10board
Count
1000
500
0
cbse
up board
mp board
upboard
pseb
karnataka secondary education boa
west bengal board of secondary edu
uttarakhand board
hbsc
bse,orissa
mpbse
mpboard
karnataka secondary education
karnataka state secondary educatio
jkbose
board of secondary education,odish
maharashtra nasik board
state board of secondary education,
kiran english medium high school
kalaimagal matriculation higher sec
bihar secondary education board,pa
himachal pradesh board
board ofsecondary education,ap
karnataka state examination board
state board - west bengal board of
rbse (state board)
kea
jharkhand academic council
J&K state board of school education
state bord
icse board , new delhi
stmary's convent inter college
ksbe
up bourd
rajasthan board ajmer
ua
state(karnataka board)
maharashtra state(latur board)
board of secondary education,andhr
himachal pradesh board of school e
karnataka state education examinat
board secondary education
gsheb
board of secondary education(bse)
bharathi matriculation school
karnataka board of higher educatio
state board of secondary education(
andhra pradesh state board
board of secondary education orissa
bihar examination board, patna
ssc board of andrapradesh
u p
hse,orissa
10board

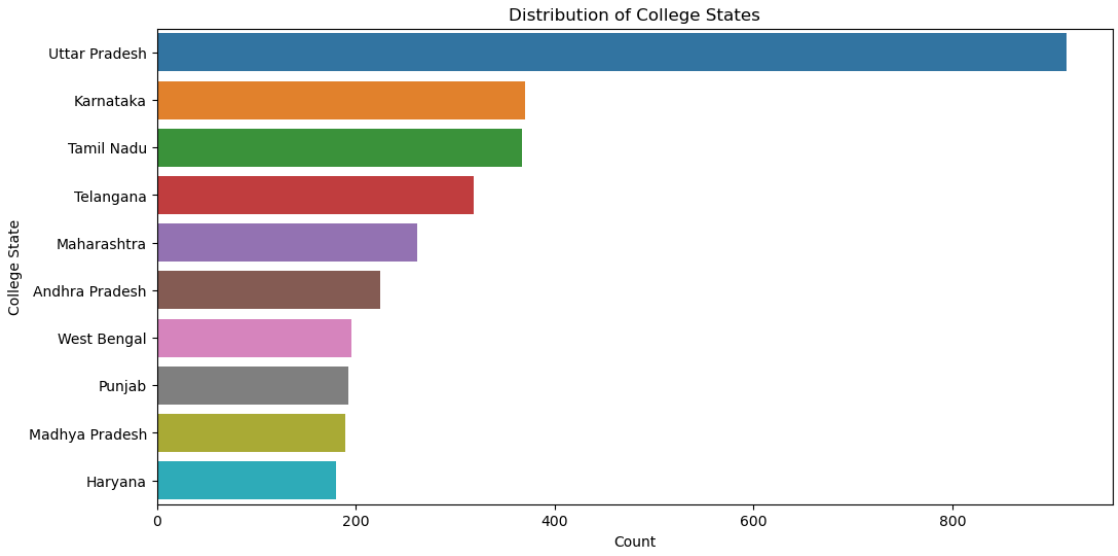

Frequency Distribution of 12board
Count
1000
500
0
cbse
board of intermediate
cbse
maharashtra board
hbse
karnataka state board
u p board
cbse,orissa
karnataka pu board
kerala state board
sbtet
biec
wbbhse
aplbie
punjab state board of technical edu
state board (jac, ranchi)
jharkhand board
bseb, patna
narayana junior college
himachal pradesh board
intermediate council patna
andhra pradesh board of secondary
nagpur board,nagpur
state syllabus
board of intmediate education,ap
hsic (tamil nadu state board)
board of intmediate education ap
rbse (state board)
board of intermidiate education,ajm
board of intermeadiate education
madhya pradesh board
ssm srsecschool
aurangabad board
mp-bse
cgbse raipur
karnataka board of university
msbte (diploma in computer techno
chse, odisha
state board of technical education a
dte
international baccalaureate (ib) dipl
kerala state hse board
chsc
intermediate board of andhra parde
andhra board
kerala
science college
kolhapur divisional board, maharash
bice
karnatak pu board
board of school education harayana
jyoti nivas
pu board ,karnataka
matric board
nagpur board
12board


Frequency Distribution of Degree
Count
3000
2000
1000
0
B.Tech/B.E.
MCA
M.Tech./M.E.
M.Sc. (Tech.)
Degree


Frequency Distribution of Specialization
Count
800
600
400
200
0
electronics and communication engi
computer science & engineering
information technology
computer engineering
computer application
electronics and electrical enginee
mechanical engineering
electronics & telecommunications
electrical engineering
electronics & instrumentation eng
civil engineering
electronics and instrumentation eng
instrumentation and control enginee
information science engineering
electronics engineering
biotechnology
other
industrial & production engineering
applied electronics and instrumenta
chemical engineering
computer science and technology
telecommunication engineering
mechanical and automation
automobile/automotive engineering
instrumentation engineering
mechatronics
aeronautical engineering
electronics and computer engineeri
electrical and power engineering
biomedical engineering
information & communication techn
industrial engineering
computer science
metallurgical engineering
power systems and automation
control and instrumentation enginee
mechanical & production engineerin
embedded systems technology
polymer technology
computer and communication engir
information science
internal combustion engine
computer networking
ceramic engineering
electronics
industrial & management engineerir
Specialization

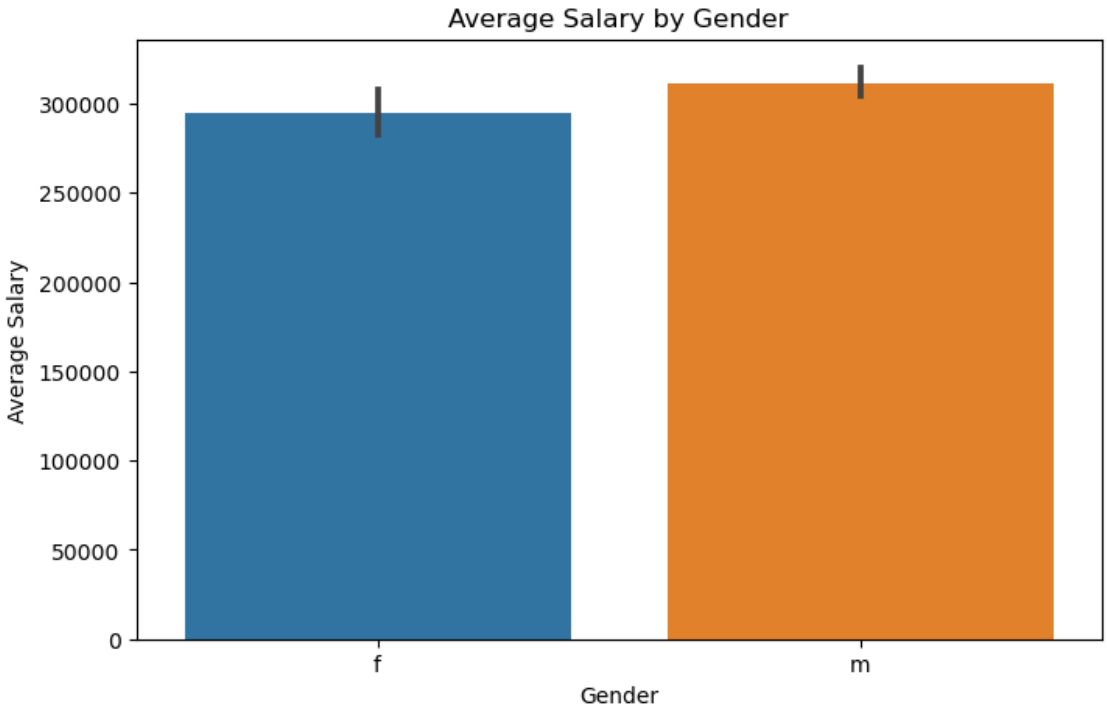


[16]:
```
# Distribution of college states
plt.figure(figsize=(12, 6))
sns.countplot(y='CollegeState', data=df, order=df['CollegeState'].
 ↪value_counts().index[:10])
plt.title('Distribution of College States')
plt.xlabel('Count')
plt.ylabel('College State')
plt.show()
```



[17]:
```
plt.figure(figsize=(8, 5))
sns.barplot(x='Gender', y='Salary', data=df, estimator=lambda x: sum(x)/len(x))
plt.title('Average Salary by Gender')
plt.xlabel('Gender')
plt.ylabel('Average Salary')
```

```
plt.show()
```



3.0.7 In general male's are getting more salary on average than female's

4 Observations from above graphs

4.0.1 1. Majority of them are working in the role of Software engineer

4.0.2 2. Most of them are living in Banglore for there job

4.0.3 3. In the given dataset we have more number of males when compare to females

4.0.4 4. Most of them have completed there tenth and 12th in CBSE board

4.0.5 5. Majority of them belongs to B.Tech/B.E degree

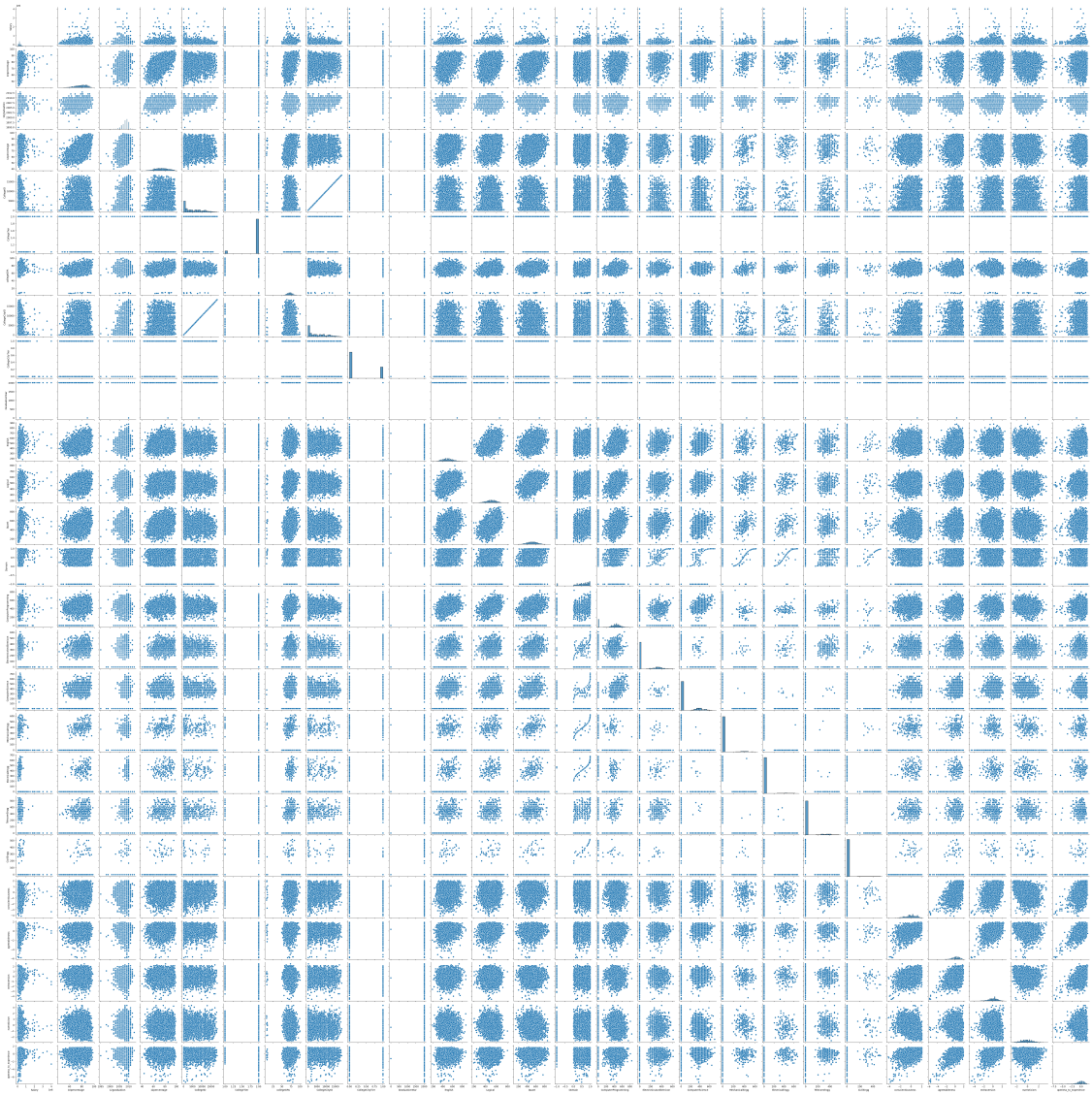
4.0.6 6. Majority are from Electronics and communication engineering as there domain

4.0.7 7. Majority of the colleges where the students studied are present in Uttar Pradesh

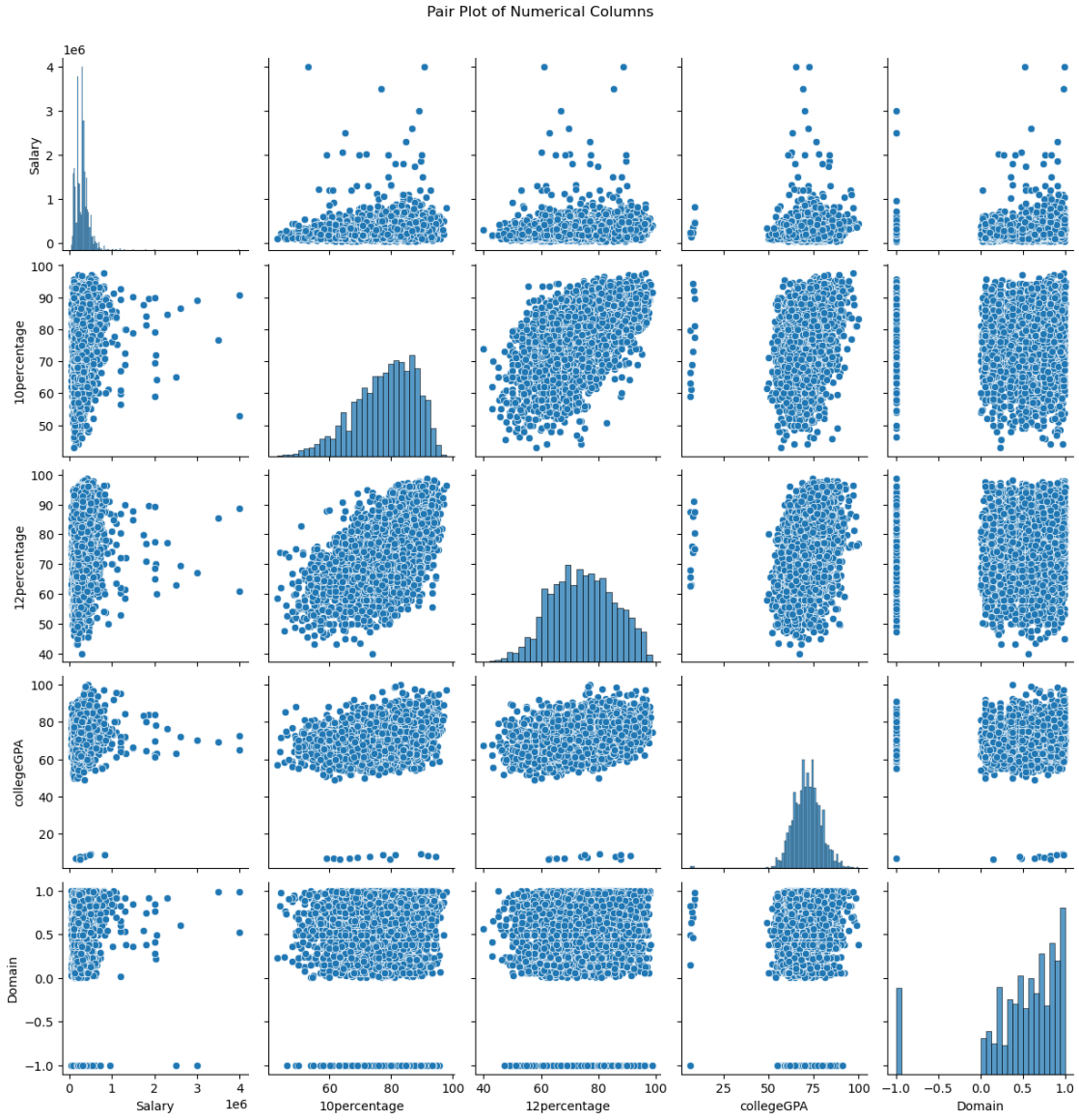
5 Bivariate Analysis

[69]:
```
sns.pairplot(data=df)
```

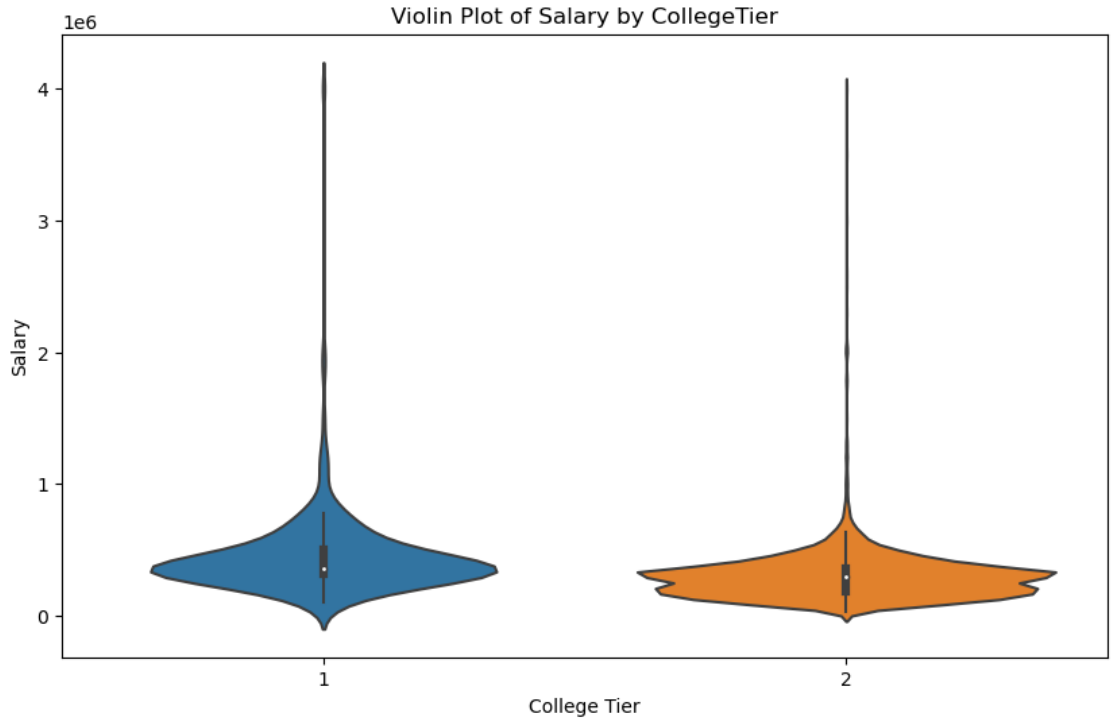
[69]: <seaborn.axisgrid.PairGrid at 0x26cc4654190>



[70]:
```
sns.pairplot(df[['Salary', '10percentage', '12percentage', 'collegeGPA',␣
 ↪'Domain']])
plt.suptitle('Pair Plot of Numerical Columns', y=1.02)
plt.show()
```



[22]:
```
plt.figure(figsize=(10, 6))
sns.violinplot(x='CollegeTier', y='Salary', data=df)
plt.title('Violin Plot of Salary by CollegeTier')
plt.xlabel('College Tier')
plt.ylabel('Salary')
plt.show()
```



5.1 observations

5.1.1 1. As the 10th percentage increases the 12th percentage and college GPA also increases

[72]:
```
corr=df.corr()
```

[74]:
```
corr
```

[74]:

	Salary	10percentage	12graduation	12percentage	\
Salary	1.000000	0.177373	-0.161383	0.170254	
10percentage	0.177373	1.000000	0.269957	0.643378	
12graduation	-0.161383	0.269957	1.000000	0.259166	
12percentage	0.170254	0.643378	0.259166	1.000000	
CollegeID	-0.118690	0.021082	0.254021	0.022336	
CollegeTier	-0.179332	-0.126042	0.027691	-0.100771	
collegeGPA	0.130103	0.312538	0.086001	0.346137	
CollegeCityID	-0.118690	0.021082	0.254021	0.022336	
CollegeCityTier	0.015384	0.116707	-0.003016	0.130462	
GraduationYear	-0.010053	-0.013799	0.014457	-0.012933	
English	0.178219	0.350780	0.147925	0.212888	
Logical	0.179275	0.316014	0.105887	0.243571	
Quant	0.230627	0.317640	0.001379	0.312413	
Domain	0.104656	0.078563	-0.034163	0.074099	

ComputerProgramming	0.115665	0.053600	-0.047995	0.080818
ElectronicsAndSemicon	0.000665	0.085179	-0.005891	0.117112
ComputerScience	-0.100720	-0.018933	0.293439	-0.043534
MechanicalEngg	0.018475	0.050364	0.035459	0.037635
ElectricalEngg	-0.047598	0.074419	0.123751	0.064001
TelecomEngg	-0.022691	0.049378	0.023470	0.044201
CivilEngg	0.037639	0.030002	-0.004727	0.005910
conscientiousness	-0.064148	0.067657	0.103329	0.058299
agreeableness	0.057423	0.136645	0.041182	0.103998
extraversion	-0.010213	-0.004679	0.061956	-0.007486
nueroticism	-0.054685	-0.132496	-0.074369	-0.094369
openess_to_experience	-0.011312	0.036692	-0.015069	0.006332

	CollegeID	CollegeTier	collegeGPA	CollegeCityID	\
Salary	-0.118690	-0.179332	0.130103	-0.118690	
10percentage	0.021082	-0.126042	0.312538	0.021082	
12graduation	0.254021	0.027691	0.086001	0.254021	
12percentage	0.022336	-0.100771	0.346137	0.022336	
CollegeID	1.000000	0.067054	0.017240	1.000000	
CollegeTier	0.067054	1.000000	-0.086781	0.067054	
collegeGPA	0.017240	-0.086781	1.000000	0.017240	
CollegeCityID	1.000000	0.067054	0.017240	1.000000	
CollegeCityTier	0.007757	-0.101494	0.017471	0.007757	
GraduationYear	-0.000172	-0.005557	0.008706	-0.000172	
English	-0.022792	-0.183843	0.106478	-0.022792	
Logical	-0.047094	-0.182811	0.196610	-0.047094	
Quant	-0.114672	-0.251103	0.217380	-0.114672	
Domain	-0.073857	-0.061436	0.107252	-0.073857	
ComputerProgramming	-0.033760	-0.073644	0.136596	-0.033760	
ElectronicsAndSemicon	-0.020438	-0.031573	0.029855	-0.020438	
ComputerScience	0.102303	0.001053	0.007601	0.102303	
MechanicalEngg	-0.009291	-0.021548	-0.031765	-0.009291	
ElectricalEngg	0.022933	0.002594	0.052258	0.022933	
TelecomEngg	0.025620	0.000007	-0.005226	0.025620	
CivilEngg	0.005749	-0.033722	-0.018950	0.005749	
conscientiousness	0.076432	0.055174	0.069582	0.076432	
agreeableness	-0.005264	-0.038055	0.068282	-0.005264	
extraversion	0.005917	0.009970	-0.032684	0.005917	
nueroticism	-0.008973	0.023778	-0.074859	-0.008973	
openess_to_experience	-0.010678	-0.019179	0.028071	-0.010678	

	CollegeCityTier	GraduationYear	…	ComputerScience	\
Salary	0.015384	-0.010053	…	-0.100720	
10percentage	0.116707	-0.013799	…	-0.018933	
12graduation	-0.003016	0.014457	…	0.293439	
12percentage	0.130462	-0.012933	…	-0.043534	
CollegeID	0.007757	-0.000172	…	0.102303	

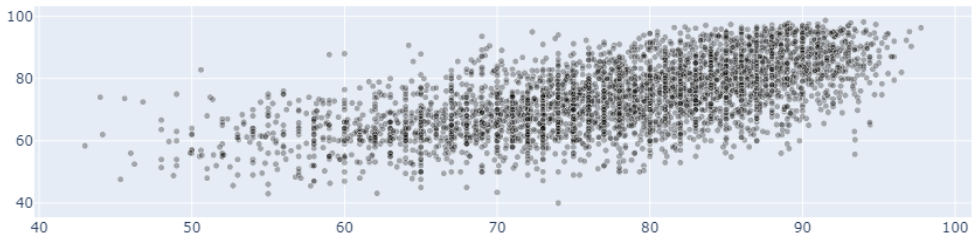
CollegeTier	-0.101494	-0.005557	…	0.001053
collegeGPA	0.017471	0.008706	…	0.007601
CollegeCityID	0.007757	-0.000172	…	0.102303
CollegeCityTier	1.000000	0.008152	…	-0.010643
GraduationYear	0.008152	1.000000	…	0.024089
English	0.050462	-0.024089	…	0.059500
Logical	0.020353	-0.024018	…	0.044481
Quant	0.007896	-0.021781	…	-0.043379
Domain	0.009250	-0.009741	…	0.058762
ComputerProgramming	0.064272	0.026688	…	0.253039
ElectronicsAndSemicon	0.041083	0.006179	…	-0.273707
ComputerScience	-0.010643	0.024089	…	1.000000
MechanicalEngg	-0.052395	-0.066844	…	-0.124355
ElectricalEngg	0.010311	0.008525	…	-0.083798
TelecomEngg	0.049876	0.004226	…	-0.148095
CivilEngg	-0.033392	0.001696	…	-0.052613
conscientiousness	0.014763	-0.013235	…	0.090155
agreeableness	0.005565	-0.002877	…	0.039866
extraversion	-0.008203	0.008397	…	0.102153
nueroticism	0.004442	-0.000417	…	-0.112652
openess_to_experience	-0.016790	0.016855	…	0.058039

	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	\
Salary	0.018475	-0.047598	-0.022691	0.037639	
10percentage	0.050364	0.074419	0.049378	0.030002	
12graduation	0.035459	0.123751	0.023470	-0.004727	
12percentage	0.037635	0.064001	0.044201	0.005910	
CollegeID	-0.009291	0.022933	0.025620	0.005749	
CollegeTier	-0.021548	0.002594	0.000007	-0.033722	
collegeGPA	-0.031765	0.052258	-0.005226	-0.018950	
CollegeCityID	-0.009291	0.022933	0.025620	0.005749	
CollegeCityTier	-0.052395	0.010311	0.049876	-0.033392	
GraduationYear	-0.066844	0.008525	0.004226	0.001696	
English	-0.002477	0.032438	-0.005822	-0.007724	
Logical	-0.009861	0.012003	-0.012947	-0.011286	
Quant	0.019933	0.020975	0.021387	0.000528	
Domain	0.048472	0.042875	0.024442	0.017569	
ComputerProgramming	-0.284891	-0.138224	-0.248269	-0.088249	
ElectronicsAndSemicon	-0.109434	0.036968	0.387140	0.002863	
ComputerScience	-0.124355	-0.083798	-0.148095	-0.052613	
MechanicalEngg	1.000000	-0.040522	-0.070947	0.076201	
ElectricalEngg	-0.040522	1.000000	-0.051469	-0.020059	
TelecomEngg	-0.070947	-0.051469	1.000000	-0.031492	
CivilEngg	0.076201	-0.020059	-0.031492	1.000000	
conscientiousness	-0.010858	0.029806	-0.004946	-0.017526	
agreeableness	-0.028586	-0.015454	-0.014627	-0.034254	
extraversion	-0.017748	0.004467	-0.039050	-0.031822	

```
nueroticism                0.036148       -0.030870     0.020638   0.010555
openess_to_experience     -0.027988       -0.012585    -0.000141  -0.031201

                       conscientiousness  agreeableness  extraversion  \
Salary                         -0.064148       0.057423     -0.010213
10percentage                    0.067657       0.136645     -0.004679
12graduation                    0.103329       0.041182      0.061956
12percentage                    0.058299       0.103998     -0.007486
CollegeID                       0.076432      -0.005264      0.005917
CollegeTier                     0.055174      -0.038055      0.009970
collegeGPA                      0.069582       0.068282     -0.032684
CollegeCityID                   0.076432      -0.005264      0.005917
CollegeCityTier                 0.014763       0.005565     -0.008203
GraduationYear                 -0.013235      -0.002877      0.008397
English                         0.034943       0.194990      0.018755
Logical                         0.025876       0.167207     -0.006949
Quant                          -0.005639       0.103443     -0.028616
Domain                         -0.039478       0.051944     -0.024647
ComputerProgramming             0.012862       0.076934      0.043504
ElectronicsAndSemicon          -0.026483      -0.024286     -0.044458
ComputerScience                 0.090155       0.039866      0.102153
MechanicalEngg                 -0.010858      -0.028586     -0.017748
ElectricalEngg                  0.029806      -0.015454      0.004467
TelecomEngg                    -0.004946      -0.014627     -0.039050
CivilEngg                      -0.017526      -0.034254     -0.031822
conscientiousness               1.000000       0.481820      0.355537
agreeableness                   0.481820       1.000000      0.454369
extraversion                    0.355537       0.454369      1.000000
nueroticism                    -0.330312      -0.207480     -0.096491
openess_to_experience           0.395649       0.591541      0.435074

                       nueroticism  openess_to_experience
Salary                   -0.054685              -0.011312
10percentage             -0.132496               0.036692
12graduation             -0.074369              -0.015069
12percentage             -0.094369               0.006332
CollegeID                -0.008973              -0.010678
CollegeTier               0.023778              -0.019179
collegeGPA               -0.074859               0.028071
CollegeCityID            -0.008973              -0.010678
CollegeCityTier           0.004442              -0.016790
GraduationYear           -0.000417               0.016855
English                  -0.155528               0.067979
Logical                  -0.178781               0.048420
Quant                    -0.131895               0.020377
Domain                   -0.017928               0.010412
ComputerProgramming      -0.084344               0.043133
```

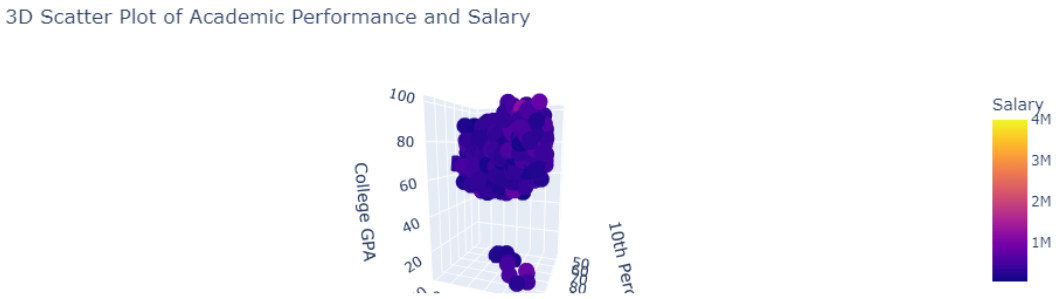
```
ElectronicsAndSemicon     0.021026              -0.013460
ComputerScience          -0.112652               0.058039
MechanicalEngg            0.036148              -0.027988
ElectricalEngg           -0.030870              -0.012585
TelecomEngg               0.020638              -0.000141
CivilEngg                 0.010555              -0.031201
conscientiousness        -0.330312               0.395649
agreeableness            -0.207480               0.591541
extraversion             -0.096491               0.435074
nueroticism               1.000000              -0.065795
openess_to_experience    -0.065795               1.000000

[26 rows x 26 columns]
```

[97]:
```
fig = go.Figure(go.Scattergl(
    x = df['10percentage'],
    y = df['12percentage'],
    mode = 'markers',
    marker=dict(color='rgba(0, 0, 0, 0.3)'),
    text=df['Salary'],
    hoverinfo='text',
))

fig.update_traces(marker=dict(size=5, line=dict(width=0.5, color='white')),␣
 ↪selector=dict(mode='markers'))



fig.show()
```

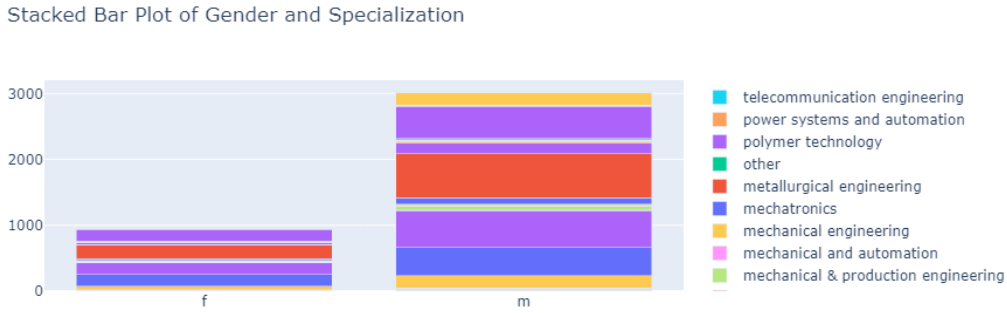


[80]:
```
px.box(x=df['Gender'], y=df['Salary'])
```

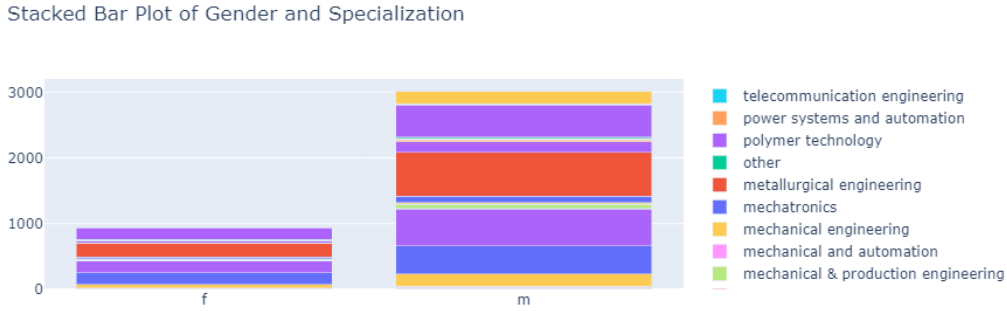
[91]:
```
fig = px.scatter_3d(df, x='10percentage', y='12percentage', z='collegeGPA',␣
↪color='Salary',
                    labels={'10percentage': '10th Percentage', '12percentage':␣
↪'12th Percentage', 'collegeGPA': 'College GPA', 'Salary': 'Salary'},
                    title='3D Scatter Plot of Academic Performance and Salary')
fig.show()
```

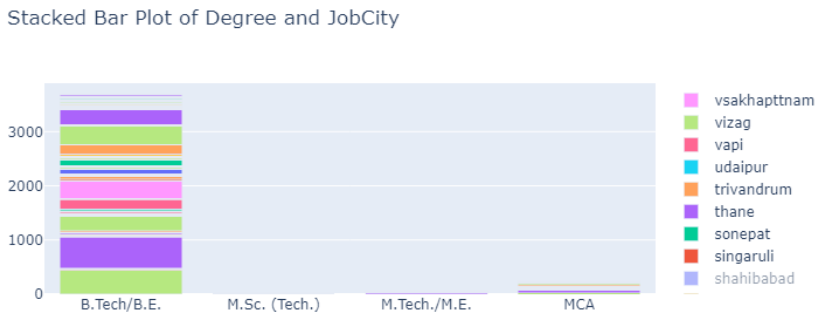


[92]:
```
contingency_table = pd.crosstab(df['Gender'], df['Specialization'])
fig = go.Figure()

for i, specialization in enumerate(contingency_table.columns):
    fig.add_trace(go.Bar(x=contingency_table.index,␣
↪y=contingency_table[specialization], name=specialization))

fig.update_layout(barmode='stack', title='Stacked Bar Plot of Gender and␣
↪Specialization')
fig.show()
```

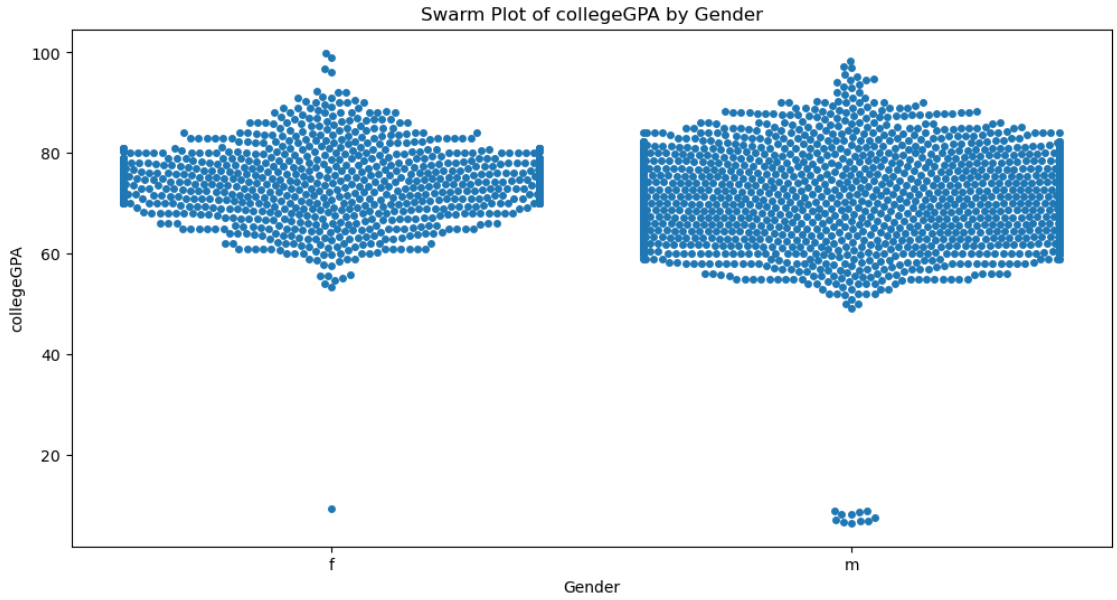


[93]:
```
categorical_pairs = [('Gender', 'Specialization'), ('Degree', 'JobCity')]

for pair in categorical_pairs:
    col1, col2 = pair
    contingency_table = pd.crosstab(df[col1], df[col2])

    fig = go.Figure()
    for i, category in enumerate(contingency_table.columns):
        fig.add_trace(go.Bar(x=contingency_table.index,␣
 ↪y=contingency_table[category], name=category))

    fig.update_layout(barmode='stack', title=f'Stacked Bar Plot of {col1} and␣
 ↪{col2}')
    fig.show()
```





[76]:
```
plt.figure(figsize=(12, 6))
sns.swarmplot(x='Gender', y='collegeGPA', data=df)
plt.title('Swarm Plot of collegeGPA by Gender')
plt.show()
```



[90]:
```
"""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."""
```

[90]:
```
'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.'

[81]:
```
cse_graduates = df[df['Specialization'] == 'computer science & engineering']

# Filter for specified job roles
specified_jobs = cse_graduates[cse_graduates['Designation'].isin(['Programming␣
 ↪Analyst', 'Software Engineer', 'Hardware Engineer', 'Associate Engineer'])]

# Calculate the salary statistics
salary_statistics = specified_jobs['Salary'].describe()

print("Salary statistics for specified job roles:")
print(salary_statistics)

# Test the claim
lower_bound = 250000
upper_bound = 300000
claim_test_result = (salary_statistics['mean'] >= lower_bound) and␣
 ↪(salary_statistics['mean'] <= upper_bound)

print("\nClaim Test Result:")
if claim_test_result:
    print("The claim is supported by the data.")
else:
    print("The claim is not supported by the data.")
```

```
Salary statistics for specified job roles:
count    0.0
mean     NaN
std      NaN
min      NaN
25%      NaN
50%      NaN
75%      NaN
max      NaN
Name: Salary, dtype: float64

Claim Test Result:
The claim is not supported by the data.
```

5.2 The claim is not supported by the data.

[82]:
```
from scipy.stats import chi2_contingency
contingency_table = pd.crosstab(df['Gender'], df['Specialization'])
chi2, p, dof, expected = chi2_contingency(contingency_table)
print("Chi-squared statistic:", chi2)
```

```
print("p-value:", p)
alpha = 0.05
print("\nSignificance level:", alpha)
if p < alpha:
    print("Reject the null hypothesis: There is a relationship between gender␣
 ↪and specialization.")
else:
    print("Fail to reject the null hypothesis: There is no relationship between␣
 ↪gender and specialization.")
```

```
Chi-squared statistic: 104.46891913608455
p-value: 1.2453868176976918e-06

Significance level: 0.05
Reject the null hypothesis: There is a relationship between gender and
specialization.
```

5.3 There is a relationship between gender and specialization.

[83]:
```
import scipy.stats as stats
male_salary = df[df['Gender'] == 'Male']['Salary']
female_salary = df[df['Gender'] == 'Female']['Salary']

t_stat, p_val = stats.ttest_ind(male_salary, female_salary, equal_var=False)
if p_val < 0.05:
    print("There is a significant difference in salary between genders.")
else:
    print("There is no significant difference in salary between genders.")
```

```
There is no significant difference in salary between genders.
```

5.3.1 There is no significant difference in salary between genders

[85]:
```
job_opportunities_by_college_tier = df.groupby('CollegeTier').size()
print(job_opportunities_by_college_tier)
```

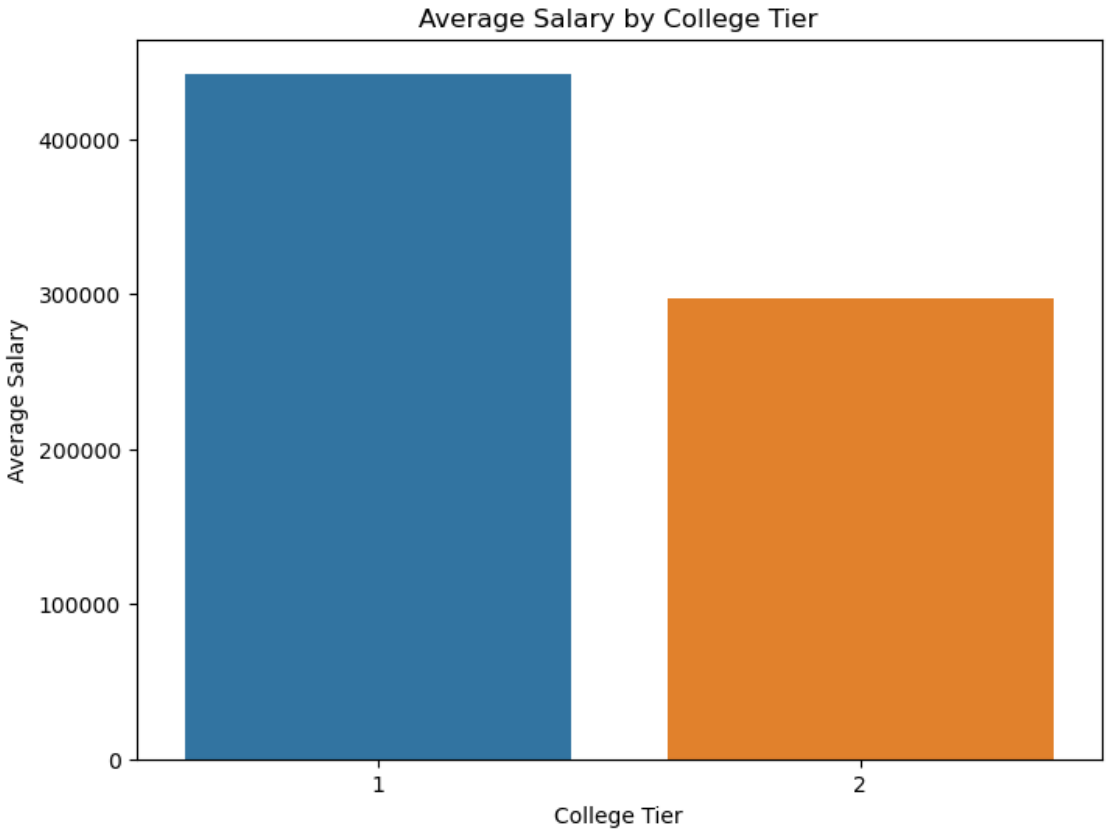
```
CollegeTier
1     297
2    3701
dtype: int64
```

5.4 Job opportunities are effected by tier to which the college belongs

[87]:
```
correlation_matrix = df[['10percentage', '12percentage', 'collegeGPA',␣
 ↪'Salary']].corr()
print(correlation_matrix)
```

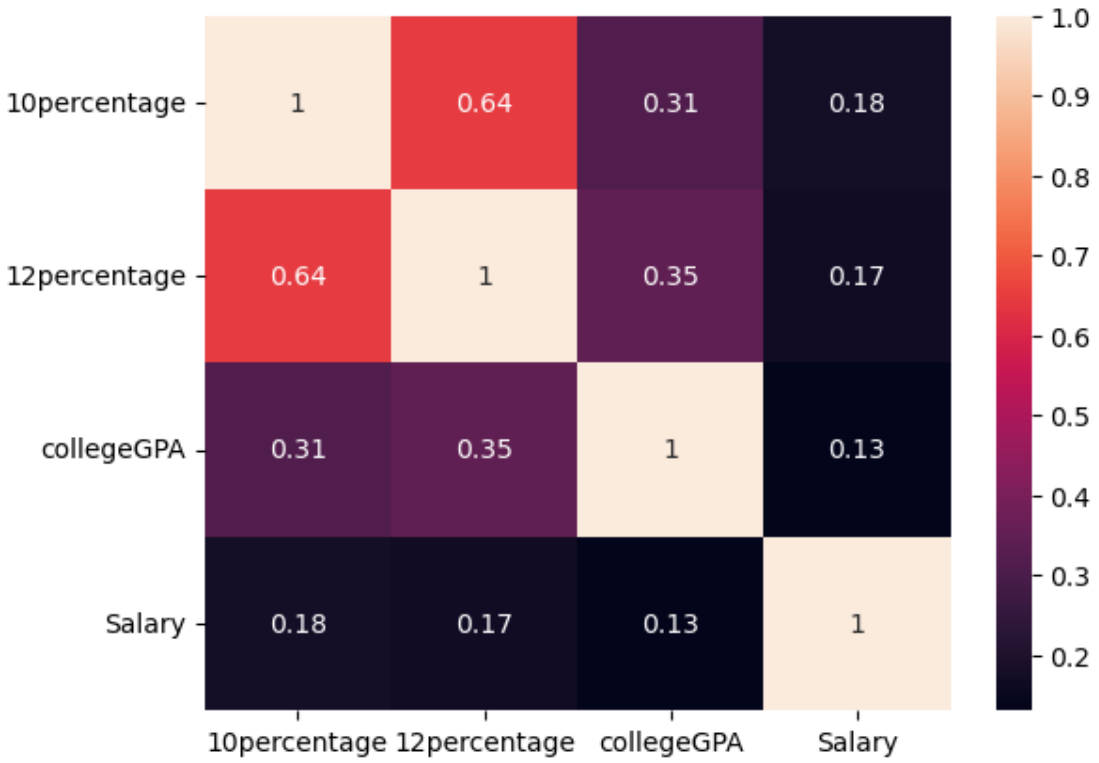
	10percentage	12percentage	collegeGPA	Salary
10percentage	1.000000	0.643378	0.312538	0.177373
12percentage	0.643378	1.000000	0.346137	0.170254
collegeGPA	0.312538	0.346137	1.000000	0.130103
Salary	0.177373	0.170254	0.130103	1.000000

[24]:
```
avg_salary_by_tier = df.groupby('CollegeTier')['Salary'].mean().reset_index()

# Create a bar plot
plt.figure(figsize=(8, 6))
sns.barplot(x='CollegeTier', y='Salary', data=avg_salary_by_tier)
plt.title('Average Salary by College Tier')
plt.xlabel('College Tier')
plt.ylabel('Average Salary')
plt.show()
```



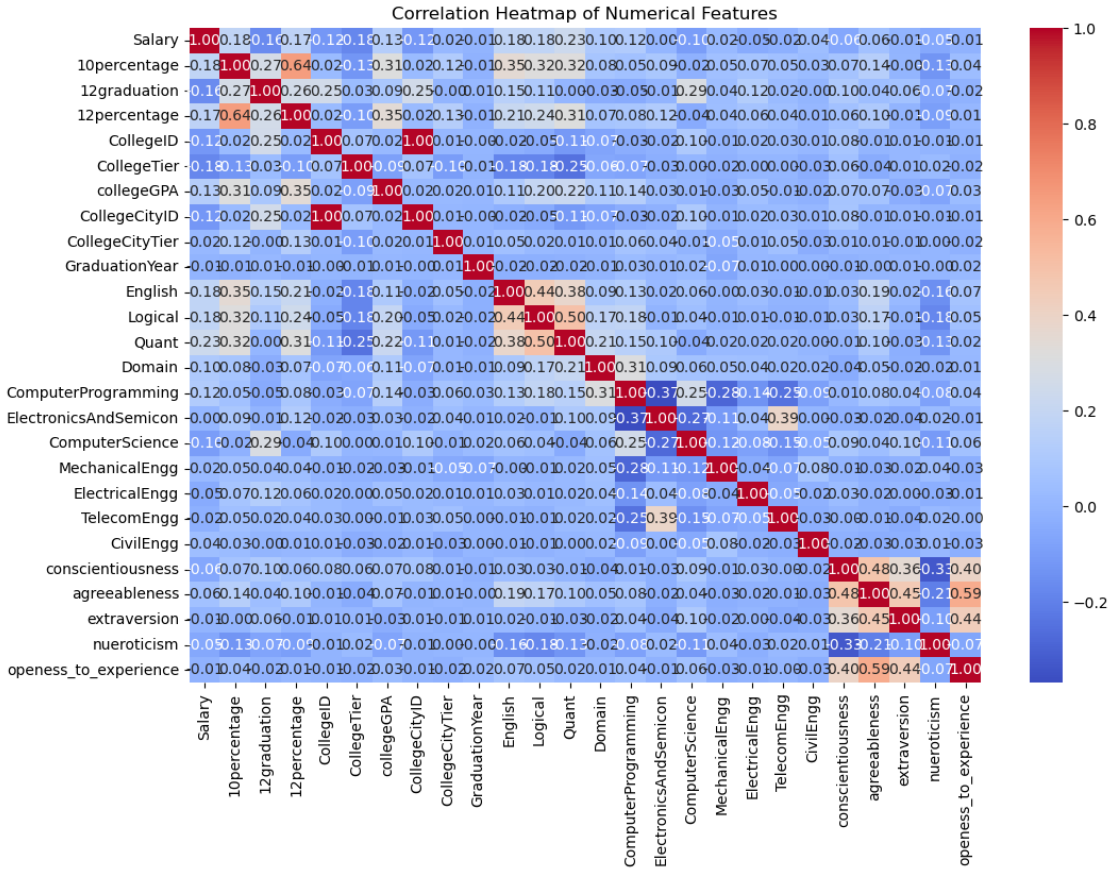
[89]:
```
sns.heatmap(correlation_matrix,annot=True)
```

[89]: <Axes: >



5.5 Salary is not dependent on percentages of 10th,12th and college GPA

[23]:
```
plt.figure(figsize=(12, 8))
sns.heatmap(df[num_feat].corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap of Numerical Features')
plt.show()
```


Correlation Heatmap of Numerical Features
Salary
10percentage
12graduation
12percentage
CollegeID
CollegeTier
collegeGPA
CollegeCityID
CollegeCityTier
GraduationYear
English
Logical
Quant
Domain
ComputerProgramming
ElectronicsAndSemicon
ComputerScience
MechanicalEngg
ElectricalEngg
TelecomEngg
CivilEngg
conscientiousness
agreeableness
extraversion
nueroticism
openess_to_experience
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

[ ]: