

# EDA

```
In [2]: import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
df=pd.read_csv("Book2.csv")
df
```

Out[2]:

	Column1	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	...	ComputerScience	Mechan
0	train	203097	420000	06-01-2012 00:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	...	-1	
1	train	579905	500000	09-01-2013 00:00	present	assistant manager	Indore	m	10/4/89 0:00	85.40	...	-1	
2	train	810601	325000	06-01-2014 00:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.00	...	-1	
3	train	267447	1100000	07-01-2011 00:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	...	-1	
4	train	343523	200000	03-01-2014 00:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.00	...	-1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
3993	train	47916	280000	10-01-2011 00:00	10/1/12 0:00	software engineer	New Delhi	m	4/15/87 0:00	52.09	...	-1	
3994	train	752781	100000	07-01-2013 00:00	7/1/13 0:00	technical writer	Hyderabad	f	8/27/92 0:00	90.00	...	-1	
3995	train	355888	320000	07-01-2013 00:00	present	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	...	-1	
3996	train	947111	200000	07-01-2014 00:00	1/1/15 0:00	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	...	438	
3997	train	324966	400000	02-01-2013 00:00	present	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	...	-1	

3998 rows × 39 columns

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

Out[3]: (3998, 39)

```
In [4]: #size
df.size
```

Out[4]: 155922

```
In [5]: #top 5 rows i.e head
df.head()
```

Out[5]:	Column1	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	...	ComputerScience	MechanicalEngg
0	train	203097	420000	06-01-2012 00:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	...	-1	-1
1	train	579905	500000	09-01-2013 00:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	...	-1	-1
2	train	810601	325000	06-01-2014 00:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	...	-1	-1
3	train	267447	1100000	07-01-2011 00:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	...	-1	-1
4	train	343523	200000	03-01-2014 00:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.0	...	-1	-1

5 rows × 39 columns

```
In [6]: df.columns=df.columns.str.lower()
```

```
In [7]: #categorical columns
cat_columns=df.select_dtypes(include='object').columns
cat_columns
```

```
Out[7]: Index(['column1', 'doj', 'dol', 'designation', 'jobcity', 'gender', 'dob',
       '10board', '12board', 'degree', 'specialization', 'collegestate'],
       dtype='object')
```

```
In [8]: len(cat_columns)
```

```
Out[8]: 12
```

```
In [9]: #numerical columns
num_columns=df.select_dtypes(exclude='object').columns
num_columns
```

```
Out[9]: Index(['id', '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')
```

```
In [10]: #Univariate analysis
#drop ID column
num_col=df.drop(columns=["id"])
num_col
```

Out[10]:

	column1	salary	doj	dol	designation	jobcity	gender	dob	10percentage	10board	...	computerscience	mecl
0	train	420000	06-01-2012 00:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	board ofsecondary education,ap	...	-1	
1	train	500000	09-01-2013 00:00	present	assistant manager	Indore	m	10/4/89 0:00	85.40	cbse	...	-1	
2	train	325000	06-01-2014 00:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.00	cbse	...	-1	
3	train	1100000	07-01-2011 00:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	cbse	...	-1	
4	train	200000	03-01-2014 00:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.00	cbse	...	-1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
3993	train	280000	10-01-2011 00:00	10/1/12 0:00	software engineer	New Delhi	m	4/15/87 0:00	52.09	cbse	...	-1	
3994	train	100000	07-01-2013 00:00	7/1/13 0:00	technical writer	Hyderabad	f	8/27/92 0:00	90.00	state board	...	-1	
3995	train	320000	07-01-2013 00:00	present	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	bse,odisha	...	-1	
3996	train	200000	07-01-2014 00:00	1/1/15 0:00	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	state board	...	438	
3997	train	400000	02-01-2013 00:00	present	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	cbse	...	-1	

3998 rows × 38 columns

In [11]: num\_col.describe()

Out[11]:

	salary	10percentage	12graduation	12percentage	collegeid	collegetier	collegegpa	collegecityid	collegecitytier	gradua
count	3.998000e+03	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	399
mean	3.076998e+05	77.925443	2008.087544	74.466366	5156.851426	1.925713	71.486171	5156.851426	0.300400	201
std	2.127375e+05	9.850162	1.653599	10.999933	4802.261482	0.262270	8.167338	4802.261482	0.458489	3
min	3.500000e+04	43.000000	1995.000000	40.000000	2.000000	1.000000	6.450000	2.000000	0.000000	
25%	1.800000e+05	71.680000	2007.000000	66.000000	494.000000	2.000000	66.407500	494.000000	0.000000	201
50%	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000	2.000000	71.720000	3879.000000	0.000000	201
75%	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000	2.000000	76.327500	8818.000000	1.000000	201
max	4.000000e+06	97.760000	2013.000000	98.700000	18409.000000	2.000000	99.930000	18409.000000	1.000000	201

8 rows × 26 columns

In [12]: #Skewness  
for i in num\_columns[1:]:  
 skew=round(df[i].skew(),2)  
 print(f"skew of column{i} is '{skew}'")

```

skew of columnsalary is '6.45'
skew of column10percentage is '-0.59'
skew of column12graduation is '-0.96'
skew of column12percentage is '-0.03'
skew of columncollegeid is '0.65'
skew of columncollegetier is '-3.25'
skew of columncollegegpa is '-1.25'
skew of columncollegecityid is '0.65'
skew of columncollegecitytier is '0.87'
skew of columngraduationyear is '-63.07'
skew of columnenglish is '0.19'
skew of columnlogical is '-0.22'
skew of columnquant is '-0.02'
skew of columndomain is '-1.92'
skew of columncomputerprogramming is '-0.78'
skew of columnelectronicsandsemicon is '1.2'
skew of columncomputerscience is '1.53'
skew of columnmechanicalengg is '4.03'
skew of columnelectricalengg is '5.06'
skew of columntelecomengg is '3.04'
skew of columnncivilengg is '10.32'
skew of columnconscientiousness is '-0.53'
skew of columnagreeableness is '-1.2'
skew of columnextraversion is '-0.52'
skew of columnnueroticism is '0.17'
skew of columnopeness_to_experience is '-1.51'

```

```
In [13]: #Kurtosis os each column
for i in num_columns[1:]:
    kurt=round(df[i].kurt(),2)
    print(f"Kurtosis of col{i} is'{kurt}'")
```

```

Kurtosis of colsalary is'80.93'
Kurtosis of col10percentage is'-0.11'
Kurtosis of col12graduation is'1.95'
Kurtosis of col12percentage is'-0.63'
Kurtosis of colcollegeid is'-0.77'
Kurtosis of colcollegetier is'8.55'
Kurtosis of colcollegegpa is'10.23'
Kurtosis of colcollegecityid is'-0.77'
Kurtosis of colcollegecitytier is'-1.24'
Kurtosis of colgraduationyear is'3984.37'
Kurtosis of colenglish is'-0.25'
Kurtosis of collogical is'-0.22'
Kurtosis of colquant is'-0.1'
Kurtosis of coldomain is'3.9'
Kurtosis of colcomputerprogramming is'-0.67'
Kurtosis of colelectronicsandsemicon is'-0.21'
Kurtosis of colcomputerscience is'0.69'
Kurtosis of colmechanicalengg is'15.02'
Kurtosis of colelectricalengg is'24.88'
Kurtosis of coltelecomengg is'7.81'
Kurtosis of colcivilengg is'109.04'
Kurtosis of colconscientiousness is'0.12'
Kurtosis of colagreeableness is'3.39'
Kurtosis of colextaversion is'0.64'
Kurtosis of colnueroticism is'-0.19'
Kurtosis of colopeness_to_experience is'5.79'

```

```
In [14]: cov_matrix = df.select_dtypes(include=[np.number]).cov()
print(cov_matrix)
```

	id	salary	10percentage	\
id	1.319275e+11	-1.910842e+10	159378.051803	
salary	-1.910842e+10	4.525724e+10	371684.456159	
10percentage	1.593781e+05	3.716845e+05	97.025700	
12graduation	4.042768e+05	-5.677157e+04	4.397119	
12percentage	2.824308e+04	3.984122e+05	69.710710	
collegeid	4.963137e+08	-1.212565e+08	997.261052	
collegetier	3.349438e+03	-1.000579e+04	-0.325618	
collegegpa	1.398544e+05	2.260530e+05	25.143547	
collegecityid	4.963137e+08	-1.212565e+08	997.261052	
collegecitytier	-5.991252e+03	1.500495e+03	0.527071	
graduationyear	3.186574e+05	-6.813186e+04	-4.330051	
english	5.164917e+06	3.978683e+06	362.593281	
logical	3.221944e+06	3.309785e+06	270.138109	
quant	-2.449201e+06	6.000532e+06	382.660566	
domain	-2.138752e+04	1.043457e+04	0.362683	
computerprogramming	1.406654e+06	5.053027e+06	108.421537	
electronicsandsemicon	-6.644284e+06	2.240082e+04	132.769360	
computerscience	3.072515e+07	-3.755551e+06	-32.687912	
mechanicalengg	-9.318690e+05	3.856527e+05	48.678119	
electricalengg	3.322958e+06	-8.868903e+05	64.203771	
telecomengg	-1.876479e+06	5.061435e+05	50.998564	
civilengg	-2.379522e+05	2.935306e+05	10.833465	
conscientiousness	6.559340e+04	-1.403798e+04	0.685534	
agreeableness	8.496059e+03	1.150482e+04	1.267615	
extraversion	4.180926e+04	-2.067186e+03	-0.043855	
nueroticism	-5.353756e+04	-1.172182e+04	-1.314998	

openness_to_experience	1.148209e+04	-2.425976e+03	0.364343
id	12graduation	12percentage	collegeid \
salary	404276.838583	28243.078928	4.963137e+08
10percentage	-56771.565543	398412.208861	-1.212565e+08
12graduation	4.397119	69.710710	9.972611e+02
12percentage	2.734391	4.714096	2.017185e+03
collegeid	4.714096	120.998528	1.179912e+03
collegegettier	2017.184639	1179.911862	2.306172e+07
collegegpa	0.012009	-0.290720	8.445431e+01
collegecityid	1.161489	31.096999	6.761679e+02
collegecitytier	2017.184639	1179.911862	2.306172e+07
graduationyear	-0.002287	0.657967	1.707967e+01
english	0.761563	-4.532195	-2.638458e+01
logical	25.669208	245.744098	-1.148619e+04
quant	15.195251	232.515001	-1.962685e+04
domain	0.278868	420.294392	-6.735002e+04
computerprogramming	-0.026476	0.382008	-1.662287e+02
electronicsandsemicon	-16.297969	182.558565	-3.329351e+04
computerscience	-1.541392	203.850411	-1.553103e+04
mechanicalengg	85.047828	-83.932520	8.610934e+04
electricalengg	5.753526	40.621034	-4.378244e+03
telecomengg	17.923053	61.660745	9.645816e+03
civilengg	4.069324	50.980677	1.290070e+04
conscientiousness	-0.286551	2.383094	1.012081e+03
agreeableness	0.175763	0.659664	3.775682e+02
extraversion	0.064135	1.077373	-2.380609e+01
nueroticism	0.097479	-0.078347	2.703415e+01
openness_to_experience	-0.123908	-1.045919	-4.341715e+01
	-0.025119	0.070213	-5.169028e+01
id	collegetier	collegegpa	collegecityid \
salary	3349.438212	139854.366484	4.963137e+08
10percentage	-10005.790486	226052.995923	-1.212565e+08
12graduation	-0.325618	25.143547	9.972611e+02
12percentage	0.012009	1.161489	2.017185e+03
collegeid	84.454309	676.167910	2.306172e+07
collegegettier	0.068786	-0.185889	8.445431e+01
collegegpa	-0.185889	66.705404	6.761679e+02
collegecityid	84.454309	676.167910	2.306172e+07
collegecitytier	-0.012204	0.065421	1.707967e+01
graduationyear	-0.046429	2.265281	-2.638458e+01
english	-5.059851	91.260290	-1.148619e+04
logical	-4.160910	139.354778	-1.962685e+04
quant	-8.054460	217.137676	-6.735002e+04
domain	-0.007552	0.410539	-1.662287e+02
computerprogramming	-3.966342	229.100139	-3.329351e+04
electronicsandsemicon	-1.310348	38.584726	-1.553103e+04
computerscience	0.048407	10.880967	8.610934e+04
mechanicalengg	-0.554542	-25.456707	-4.378244e+03
electricalengg	0.059591	37.382367	9.645816e+03
telecomengg	0.000200	-4.475103	1.290070e+04
civilengg	-0.324218	-5.673773	1.012081e+03
conscientiousness	0.014885	0.584587	3.775682e+02
agreeableness	-0.009400	0.525213	-2.380609e+01
extraversion	0.002488	-0.253988	2.703415e+01
nueroticism	0.006284	-0.616035	-4.341715e+01
openness_to_experience	-0.005071	0.231115	-5.169028e+01
id	collegecitytier	...	computerscience mechanicalengg \
salary	-5991.252230	...	3.072515e+07 -931869.041883
10percentage	1500.495432	...	-3.755551e+06 385652.660706
12graduation	0.527071	...	-3.268791e+01 48.678119
12percentage	-0.002287	...	8.504783e+01 5.753526
collegeid	0.657967	...	-8.393252e+01 40.621034
collegegettier	17.079669	...	8.610934e+04 -4378.244435
collegegpa	-0.012204	...	4.840737e-02 -0.554542
collegecityid	0.065421	...	1.088097e+01 -25.456707
collegecitytier	17.079669	...	8.610934e+04 -4378.244435
graduationyear	0.210212	...	-8.552884e-01 -2.357183
english	0.119072	...	1.345049e+02 -208.949791
logical	2.427936	...	1.094388e+03 -25.504990
quant	0.809818	...	6.765949e+02 -83.975112
domain	0.442781	...	-9.298783e+02 239.207955
computerprogramming	0.001988	...	4.827045e+00 2.229099
electronicsandsemicon	6.051422	...	9.107722e+03 -5740.597853
computerscience	2.980629	...	-7.591367e+03 -1699.196105
mechanicalengg	-0.855288	...	3.072065e+04 -2138.701531
electricalengg	-2.357183	...	-2.138702e+03 9628.184250
telecomengg	0.414069	...	-1.286421e+03 -348.250350
civilengg	2.397733	...	-2.721675e+03 -729.942715
conscientiousness	-0.561244	...	-3.380489e+02 274.099592
agreeableness	0.006963	...	1.625467e+01 -1.095935
extraversion	0.002403	...	6.580696e+00 -2.641643
nueroticism	-0.003578	...	1.703584e+01 -1.657003
openness_to_experience	0.002052	...	-1.989447e+01 3.573884
	-0.007760	...	1.025486e+01 -2.768478

id	electricalengg	telecomengg	civilengg	\
salary	3.322958e+06	-1.876479e+06	-237952.207851	
10percentage	-8.868903e+05	-5.061435e+05	293530.550489	
12graduation	6.420377e+01	5.099856e+01	10.833465	
12percentage	1.792305e+01	4.069324e+00	-0.286551	
collegeid	6.166074e+01	5.098068e+01	2.383094	
collegetier	9.645816e+03	1.290070e+04	1012.081362	
collegegpa	5.959109e-02	1.999373e-04	-0.324218	
collegecityid	3.738237e+01	-4.475103e+00	-5.673773	
collegecitytier	9.645816e+03	1.290070e+04	1012.081362	
graduationyear	4.140691e-01	2.397733e+00	-0.561244	
english	2.378671e+01	1.411507e+01	1.980169	
logical	2.981455e+02	-6.405849e+01	-29.714678	
quant	9.123815e+01	-1.178150e+02	-35.906209	
domain	2.246815e+02	2.742630e+02	2.365782	
computerprogramming	1.759970e+00	1.201130e+00	0.301847	
electronicsandsemicon	-2.486120e+03	-5.345745e+03	-664.341271	
computerscience	5.123694e+02	6.423440e+03	16.605733	
mechanicalengg	-1.286421e+03	-2.721675e+03	-338.048947	
electricalengg	-3.482503e+02	-7.299427e+02	274.099592	
telecomengg	7.671243e+03	-4.726665e+02	-64.405022	
civilengg	-4.726665e+02	1.099412e+04	-121.048815	
conscientiousness	-6.440502e+01	-1.210488e+02	1343.845979	
agreeableness	2.685442e+00	-5.334358e-01	-0.660911	
extraversion	-1.274755e+00	-1.444439e+00	-1.182602	
nueroticism	3.722801e-01	-3.895795e+00	-1.109929	
openness_to_experience	-2.724256e+00	2.180398e+00	0.389861	
	-1.111177e+00	-1.491966e-02	-1.153006	

id	conscientiousness	agreeableness	extraversion	\
salary	65593.400265	8496.058635	41809.257150	
10percentage	-14037.983625	11504.821563	-2067.186195	
12graduation	0.685534	1.267615	-0.043855	
12percentage	0.175763	0.064135	0.097479	
collegeid	0.659664	1.077373	-0.078347	
collegetier	377.568165	-23.806085	27.034147	
collegegpa	0.014885	-0.009400	0.002488	
collegecityid	0.584587	0.525213	-0.253988	
collegecitytier	377.568165	-23.806085	27.034147	
graduationyear	0.006963	0.002403	-0.003578	
english	-0.433725	-0.086324	0.254526	
logical	3.772022	19.270980	1.872663	
quant	2.309967	13.666005	-0.573754	
domain	-0.709410	11.914767	-3.329959	
computerprogramming	-0.019032	0.022927	-0.010991	
electronicsandsemicon	2.717050	14.879090	8.500256	
computerscience	-4.310882	-3.619239	-6.693617	
mechanicalengg	16.254671	6.580696	17.035838	
electricalengg	-1.095935	-2.641643	-1.657003	
telecomengg	2.685442	-1.274755	0.372280	
civilengg	-0.533436	-1.444439	-3.895795	
conscientiousness	-0.660911	-1.182602	-1.109929	
agreeableness	1.058153	0.466777	0.347980	
extraversion	0.466777	0.886954	0.407150	
nueroticism	0.347980	0.407150	0.905298	
openness_to_experience	-0.342356	-0.196882	-0.092504	
	0.410277	0.561601	0.417303	

id	nueroticism	openness_to_experience	
salary	-53537.564434	11482.092235	
10percentage	-11721.817357	-2425.975501	
12graduation	-1.314998	0.364343	
12percentage	-0.123908	-0.025119	
collegeid	-1.045919	0.070213	
collegetier	-43.417151	-51.690285	
collegegpa	0.006284	-0.005071	
collegecityid	-0.616035	0.231115	
collegecitytier	-43.417151	-51.690285	
graduationyear	0.002052	-0.007760	
english	-0.013375	0.541276	
logical	-16.444814	7.191299	
quant	-15.632782	4.236020	
domain	-16.253350	2.512329	
computerprogramming	-0.008466	0.004919	
electronicsandsemicon	-17.451760	8.929076	
computerscience	3.352440	-2.147063	
mechanicalengg	-19.894473	10.254859	
electricalengg	3.573884	-2.768478	
telecomengg	-2.724256	-1.111177	
civilengg	2.180398	-0.014920	
conscientiousness	0.389861	-1.153006	
agreeableness	-0.342356	0.410277	
extraversion	-0.196882	0.561601	
nueroticism	-0.092504	0.417303	
openness_to_experience	1.015217	-0.066829	
	-0.066829	1.016214	

[27 rows x 27 columns]

In [15]:

```
df_numeric = df.apply(pd.to_numeric, errors='coerce')
cov_matrix = df_numeric.cov()
print(cov_matrix)
```

	column1	id	salary	doj	dol	\
column1	NaN	NaN	NaN	NaN	NaN	
id	NaN	1.319275e+11	-1.910842e+10	NaN	NaN	
salary	NaN	-1.910842e+10	4.525724e+10	NaN	NaN	
doj	NaN	NaN	NaN	NaN	NaN	
dol	NaN	NaN	NaN	NaN	NaN	
designation	NaN	NaN	NaN	NaN	NaN	
jobcity	NaN	0.000000e+00	0.000000e+00	NaN	NaN	
gender	NaN	NaN	NaN	NaN	NaN	
dob	NaN	NaN	NaN	NaN	NaN	
10percentage	NaN	1.593781e+05	3.716845e+05	NaN	NaN	
10board	NaN	0.000000e+00	0.000000e+00	NaN	NaN	
12graduation	NaN	4.042768e+05	-5.677157e+04	NaN	NaN	
12percentage	NaN	2.824308e+04	3.984122e+05	NaN	NaN	
12board	NaN	0.000000e+00	0.000000e+00	NaN	NaN	
collegeid	NaN	4.963137e+08	-1.212565e+08	NaN	NaN	
collegetier	NaN	3.349438e+03	-1.000579e+04	NaN	NaN	
degree	NaN	NaN	NaN	NaN	NaN	
specialization	NaN	NaN	NaN	NaN	NaN	
collegegpa	NaN	1.398544e+05	2.260530e+05	NaN	NaN	
collegecityid	NaN	4.963137e+08	-1.212565e+08	NaN	NaN	
collegecitytier	NaN	-5.991252e+03	1.500495e+03	NaN	NaN	
collegestate	NaN	NaN	NaN	NaN	NaN	
graduationyear	NaN	3.186574e+05	-6.813186e+04	NaN	NaN	
english	NaN	5.164917e+06	3.978683e+06	NaN	NaN	
logical	NaN	3.221944e+06	3.309785e+06	NaN	NaN	
quant	NaN	-2.449201e+06	6.000532e+06	NaN	NaN	
domain	NaN	-2.138752e+04	1.043457e+04	NaN	NaN	
computerprogramming	NaN	1.406654e+06	5.053027e+06	NaN	NaN	
electronicsandsemicon	NaN	-6.644284e+06	2.240082e+04	NaN	NaN	
computerscience	NaN	3.072515e+07	-3.755551e+06	NaN	NaN	
mechanicalengg	NaN	-9.318690e+05	3.856527e+05	NaN	NaN	
electricalengg	NaN	3.322958e+06	-8.868903e+05	NaN	NaN	
telecomengg	NaN	-1.876479e+06	-5.061435e+05	NaN	NaN	
civilengg	NaN	-2.379522e+05	2.935306e+05	NaN	NaN	
conscientiousness	NaN	6.559340e+04	-1.403798e+04	NaN	NaN	
agreeableness	NaN	8.496059e+03	1.150482e+04	NaN	NaN	
extraversion	NaN	4.180926e+04	-2.067186e+03	NaN	NaN	
nueroticism	NaN	-5.353756e+04	-1.172182e+04	NaN	NaN	
openness_to_experience	NaN	1.148209e+04	-2.425976e+03	NaN	NaN	

	designation	jobcity	gender	dob	10percentage	...	\
column1	NaN	NaN	NaN	NaN	NaN	...	
id	NaN	0.0	NaN	NaN	159378.051803	...	
salary	NaN	0.0	NaN	NaN	371684.456159	...	
doj	NaN	NaN	NaN	NaN	NaN	...	
dol	NaN	NaN	NaN	NaN	NaN	...	
designation	NaN	NaN	NaN	NaN	NaN	...	
jobcity	NaN	0.0	NaN	NaN	0.000000	...	
gender	NaN	NaN	NaN	NaN	NaN	...	
dob	NaN	NaN	NaN	NaN	NaN	...	
10percentage	NaN	0.0	NaN	NaN	97.025700	...	
10board	NaN	0.0	NaN	NaN	0.000000	...	
12graduation	NaN	0.0	NaN	NaN	4.397119	...	
12percentage	NaN	0.0	NaN	NaN	69.710710	...	
12board	NaN	0.0	NaN	NaN	0.000000	...	
collegeid	NaN	0.0	NaN	NaN	997.261052	...	
collegetier	NaN	0.0	NaN	NaN	-0.325618	...	
degree	NaN	NaN	NaN	NaN	NaN	...	
specialization	NaN	NaN	NaN	NaN	NaN	...	
collegegpa	NaN	0.0	NaN	NaN	25.143547	...	
collegecityid	NaN	0.0	NaN	NaN	997.261052	...	
collegecitytier	NaN	0.0	NaN	NaN	0.527071	...	
collegestate	NaN	NaN	NaN	NaN	NaN	...	
graduationyear	NaN	0.0	NaN	NaN	-4.330051	...	
english	NaN	0.0	NaN	NaN	362.593281	...	
logical	NaN	0.0	NaN	NaN	270.138109	...	
quant	NaN	0.0	NaN	NaN	382.660566	...	
domain	NaN	0.0	NaN	NaN	0.362683	...	
computerprogramming	NaN	0.0	NaN	NaN	108.421537	...	
electronicsandsemicon	NaN	0.0	NaN	NaN	132.769360	...	
computerscience	NaN	0.0	NaN	NaN	-32.687912	...	
mechanicalengg	NaN	0.0	NaN	NaN	48.678119	...	
electricalengg	NaN	0.0	NaN	NaN	64.203771	...	
telecomengg	NaN	0.0	NaN	NaN	50.998564	...	
civilengg	NaN	0.0	NaN	NaN	10.833465	...	
conscientiousness	NaN	0.0	NaN	NaN	0.685534	...	
agreeableness	NaN	0.0	NaN	NaN	1.267615	...	
extraversion	NaN	0.0	NaN	NaN	-0.043855	...	
nueroticism	NaN	0.0	NaN	NaN	-1.314998	...	
openness_to_experience	NaN	0.0	NaN	NaN	0.364343	...	
	computerscience	mechanicalengg	electricalengg	\			
column1		NaN	NaN	NaN			
id	3.072515e+07	-931869.041883	3.322958e+06				

salary	-3.755551e+06	385652.660706	-8.868903e+05
doj	NaN	NaN	NaN
dol	NaN	NaN	NaN
designation	NaN	NaN	NaN
jobcity	0.000000e+00	0.000000	0.000000e+00
gender	NaN	NaN	NaN
dob	NaN	NaN	NaN
10percentage	-3.268791e+01	48.678119	6.420377e+01
10board	0.000000e+00	0.000000	0.000000e+00
12graduation	8.504783e+01	5.753526	1.792305e+01
12percentage	-8.393252e+01	40.621034	6.166074e+01
12board	0.000000e+00	0.000000	0.000000e+00
collegeid	8.610934e+04	-4378.244435	9.645816e+03
collegetier	4.840737e-02	-0.554542	5.959109e-02
degree	NaN	NaN	NaN
specialization	NaN	NaN	NaN
collegegpa	1.088097e+01	-25.456707	3.738237e+01
collegecityid	8.610934e+04	-4378.244435	9.645816e+03
collegecitytier	-8.552884e-01	-2.357183	4.140691e-01
collegestate	NaN	NaN	NaN
graduationyear	1.345049e+02	-208.949791	2.378671e+01
english	1.094388e+03	-25.504990	2.981455e+02
logical	6.765949e+02	-83.975112	9.123815e+01
quant	-9.298783e+02	239.207955	2.246815e+02
domain	4.827045e+00	2.229099	1.759970e+00
computerprogramming	9.107722e+03	-5740.597853	-2.486120e+03
electronicsandsemicon	-7.591367e+03	-1699.196105	5.123694e+02
computerscience	3.072065e+04	-2138.701531	-1.286421e+03
mechanicalengg	-2.138702e+03	9628.184250	-3.482503e+02
electricalengg	-1.286421e+03	-348.250350	7.671243e+03
telecomengg	-2.721675e+03	-729.942715	-4.726665e+02
civilengg	-3.380489e+02	274.099592	-6.440502e+01
conscientiousness	1.625467e+01	-1.095935	2.685442e+00
agreeableness	6.580696e+00	-2.641643	-1.274755e+00
extraversion	1.703584e+01	-1.657003	3.722801e-01
nuerotism	-1.989447e+01	3.573884	-2.724256e+00
openess_to_experience	1.025486e+01	-2.768478	-1.111177e+00

column1	telecomengg	civilengg	conscientiousness	\
id	NaN	NaN	NaN	
salary	-1.876479e+06	-237952.207851	65593.400265	
doj	NaN	NaN	NaN	
dol	NaN	NaN	NaN	
designation	NaN	NaN	NaN	
jobcity	0.000000e+00	0.000000	0.000000	
gender	NaN	NaN	NaN	
dob	NaN	NaN	NaN	
10percentage	5.099856e+01	10.833465	0.685534	
10board	0.000000e+00	0.000000	0.000000	
12graduation	4.069324e+00	-0.286551	0.175763	
12percentage	5.098068e+01	2.383094	0.659664	
12board	0.000000e+00	0.000000	0.000000	
collegeid	1.290070e+04	1012.081362	377.568165	
collegetier	1.999373e-04	-0.324218	0.014885	
degree	NaN	NaN	NaN	
specialization	NaN	NaN	NaN	
collegegpa	-4.475103e+00	-5.673773	0.584587	
collegecityid	1.290070e+04	1012.081362	377.568165	
collegecitytier	2.397733e+00	-0.561244	0.006963	
collegestate	NaN	NaN	NaN	
graduationyear	1.411507e+01	1.980169	-0.433725	
english	-6.405849e+01	-29.714678	3.772022	
logical	-1.178150e+02	-35.906209	2.309967	
quant	2.742630e+02	2.365782	-0.709410	
domain	1.201130e+00	0.301847	-0.019032	
computerprogramming	-5.345745e+03	-664.341271	2.717050	
electronicsandsemicon	6.423440e+03	16.605733	-4.310882	
computerscience	-2.721675e+03	-338.048947	16.254671	
mechanicalengg	-7.299427e+02	274.099592	-1.095935	
electricalengg	-4.726665e+02	-64.405022	2.685442	
telecomengg	1.099412e+04	-121.048815	-0.533436	
civilengg	-1.210488e+02	1343.845979	-0.660911	
conscientiousness	-5.334358e-01	-0.660911	1.058153	
agreeableness	-1.444439e+00	-1.182602	0.466777	
extraversion	-3.895795e+00	-1.109929	0.347980	
nuerotism	2.180398e+00	0.389861	-0.342356	
openess_to_experience	-1.491966e-02	-1.153006	0.410277	

column1	agreeableness	extraversion	nuerotism	\
id	NaN	NaN	NaN	
salary	8496.058635	41809.257150	-53537.564434	
doj	NaN	NaN	NaN	
dol	NaN	NaN	NaN	
designation	NaN	NaN	NaN	
jobcity	0.000000	0.000000	0.000000	
gender	NaN	NaN	NaN	
dob	NaN	NaN	NaN	

10percentage	1.267615	-0.043855	-1.314998
10board	0.000000	0.000000	0.000000
12graduation	0.064135	0.097479	-0.123908
12percentage	1.077373	-0.078347	-1.045919
12board	0.000000	0.000000	0.000000
collegeid	-23.806085	27.034147	-43.417151
collegetier	-0.009400	0.002488	0.006284
degree	NaN	NaN	NaN
specialization	NaN	NaN	NaN
collegegpa	0.525213	-0.253988	-0.616035
collegecityid	-23.806085	27.034147	-43.417151
collegecitytier	0.002403	-0.003578	0.002052
collegestate	NaN	NaN	NaN
graduationyear	-0.086324	0.254526	-0.013375
english	19.270980	1.872663	-16.444814
logical	13.666005	-0.573754	-15.632782
quant	11.914767	-3.329959	-16.253350
domain	0.022927	-0.010991	-0.008466
computerprogramming	14.879090	8.500256	-17.451760
electronicsandsemicon	-3.619239	-6.693617	3.352440
computerscience	6.580696	17.035838	-19.894473
mechanicalengg	-2.641643	-1.657003	3.573884
electricalengg	-1.274755	0.372280	-2.724256
telecomengg	-1.444439	-3.895795	2.180398
civilengg	-1.182602	-1.109929	0.389861
conscientiousness	0.466777	0.347980	-0.342356
agreeableness	0.886954	0.407150	-0.196882
extraversion	0.407150	0.905298	-0.092504
nuerotism	-0.196882	-0.092504	1.015217
openness_to_experience	0.561601	0.417303	-0.066829

	openness_to_experience
column1	NaN
id	11482.092235
salary	-2425.975501
doj	NaN
dol	NaN
designation	NaN
jobcity	0.000000
gender	NaN
dob	NaN
10percentage	0.364343
10board	0.000000
12graduation	-0.025119
12percentage	0.070213
12board	0.000000
collegeid	-51.690285
collegetier	-0.005071
degree	NaN
specialization	NaN
collegegpa	0.231115
collegecityid	-51.690285
collegecitytier	-0.007760
collegestate	NaN
graduationyear	0.541276
english	7.191299
logical	4.236020
quant	2.512329
domain	0.004919
computerprogramming	8.929076
electronicsandsemicon	-2.147063
computerscience	10.254859
mechanicalengg	-2.768478
electricalengg	-1.111177
telecomengg	-0.014920
civilengg	-1.153006
conscientiousness	0.410277
agreeableness	0.561601
extraversion	0.417303
nuerotism	-0.066829
openness_to_experience	1.016214

[39 rows x 39 columns]

```
In [16]: corr_matrix = df.select_dtypes(include=[np.number]).corr()
print(corr_matrix)
```

id	1.000000	-0.247294	0.044547	0.673102	\
salary	-0.247294	1.000000	0.177373	-0.161383	
10percentage	0.044547	0.177373	1.000000	0.269957	
12graduation	0.673102	-0.161383	0.269957	1.000000	
12percentage	0.007069	0.170254	0.643378	0.259166	
collegeid	0.284540	-0.118690	0.021082	0.254021	
collegetier	0.035160	-0.179332	-0.126042	0.027691	
collegegpa	0.047144	0.130103	0.312538	0.086001	
collegecityid	0.284540	-0.118690	0.021082	0.254021	
collegecitytier	-0.035977	0.015384	0.116707	-0.003016	
graduationyear	0.027539	-0.010053	-0.013799	0.014457	

english	0.135505	0.178219	0.350780	0.147925
logical	0.102215	0.179275	0.316014	0.105887
quant	-0.055134	0.230627	0.317640	0.001379
domain	-0.125639	0.104656	0.078563	-0.034163
computerprogramming	0.018859	0.115665	0.053600	-0.047995
electronicsandsemicon	-0.115601	0.000665	0.085179	-0.005891
computerscience	0.482626	-0.100720	-0.018933	0.293439
mechanicalengg	-0.026147	0.018475	0.050364	0.035459
electricalengg	0.104454	-0.047598	0.074419	0.123751
telecomengg	-0.049272	-0.022691	0.049378	0.023470
civilengg	-0.017871	0.037639	0.030002	-0.004727
conscientiousness	0.175557	-0.064148	0.067657	0.103329
agreeableness	0.024837	0.057423	0.136645	0.041182
extraversion	0.120979	-0.010213	-0.004679	0.061956
nueroticism	-0.146289	-0.054685	-0.132496	-0.074369
openness_to_experience	0.031359	-0.011312	0.036692	-0.015069

	12percentage	collegeid	collegetier	collegegpa	\
id	0.007069	0.284540	0.035160	0.047144	
salary	0.170254	-0.118690	-0.179332	0.130103	
10percentage	0.643378	0.021082	-0.126042	0.312538	
12graduation	0.259166	0.254021	0.027691	0.086001	
12percentage	1.000000	0.022336	-0.100771	0.346137	
collegeid	0.022336	1.000000	0.067054	0.017240	
collegetier	-0.100771	0.067054	1.000000	-0.086781	
collegegpa	0.346137	0.017240	-0.086781	1.000000	
collegecityid	0.022336	1.000000	0.067054	0.017240	
collegecitytier	0.130462	0.007757	-0.101494	0.017471	
graduationyear	-0.012933	-0.000172	-0.005557	0.008706	
english	0.212888	-0.022792	-0.183843	0.106478	
logical	0.243571	-0.047094	-0.182811	0.196610	
quant	0.312413	-0.114672	-0.251103	0.217380	
domain	0.074099	-0.073857	-0.061436	0.107252	
computerprogramming	0.080818	-0.033760	-0.073644	0.136596	
electronicsandsemicon	0.117112	-0.020438	-0.031573	0.029855	
computerscience	-0.043534	0.102303	0.001053	0.007601	
mechanicalengg	0.037635	-0.009291	-0.021548	-0.031765	
electricalengg	0.064001	0.022933	0.002594	0.052258	
telecomengg	0.044201	0.025620	0.000007	-0.005226	
civilengg	0.005910	0.005749	-0.033722	-0.018950	
conscientiousness	0.058299	0.076432	0.055174	0.069582	
agreeableness	0.103998	-0.005264	-0.038055	0.068282	
extraversion	-0.007486	0.005917	0.009970	-0.032684	
nueroticism	-0.094369	-0.008973	0.023778	-0.074859	
openness_to_experience	0.006332	-0.010678	-0.019179	0.028071	

	collegecityid	collegecitytier	...	computerscience	\
id	0.284540	-0.035977	...	0.482626	
salary	-0.118690	0.015384	...	-0.100720	
10percentage	0.021082	0.116707	...	-0.018933	
12graduation	0.254021	-0.003016	...	0.293439	
12percentage	0.022336	0.130462	...	-0.043534	
collegeid	1.000000	0.007757	...	0.102303	
collegetier	0.067054	-0.101494	...	0.001053	
collegegpa	0.017240	0.017471	...	0.007601	
collegecityid	1.000000	0.007757	...	0.102303	
collegecitytier	0.007757	1.000000	...	-0.010643	
graduationyear	-0.000172	0.008152	...	0.024089	
english	-0.022792	0.050462	...	0.059500	
logical	-0.047094	0.020353	...	0.044481	
quant	-0.114672	0.007896	...	-0.043379	
domain	-0.073857	0.009250	...	0.058762	
computerprogramming	-0.033760	0.064272	...	0.253039	
electronicsandsemicon	-0.020438	0.041083	...	-0.273707	
computerscience	0.102303	-0.010643	...	1.000000	
mechanicalengg	-0.009291	-0.052395	...	-0.124355	
electricalengg	0.022933	0.010311	...	-0.083798	
telecomengg	0.025620	0.049876	...	-0.148095	
civilengg	0.005749	-0.033392	...	-0.052613	
conscientiousness	0.076432	0.014763	...	0.090155	
agreeableness	-0.005264	0.005565	...	0.039866	
extraversion	0.005917	-0.008203	...	0.102153	
nueroticism	-0.008973	0.004442	...	-0.112652	
openness_to_experience	-0.010678	-0.016790	...	0.058039	

	mechanicalengg	electricalengg	telecomengg	civilengg	\
id	-0.026147	0.104454	-0.049272	-0.017871	
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	\
id	0.175557	0.024837	0.120979	
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		
id	-0.146289	0.031359		
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		

[27 rows x 27 columns]

```
In [17]: #Correlation
df_numeric = df.apply(pd.to_numeric, errors='coerce')
corr_matrix = df_numeric.corr()
print(corr_matrix)
```

	column1	id	salary	doj	dol	designation	\
column1	NaN	NaN	NaN	NaN	NaN	NaN	
id	NaN	1.000000	-0.247294	NaN	NaN	NaN	
salary	NaN	-0.247294	1.000000	NaN	NaN	NaN	
doj	NaN	NaN	NaN	NaN	NaN	NaN	
dol	NaN	NaN	NaN	NaN	NaN	NaN	
designation	NaN	NaN	NaN	NaN	NaN	NaN	
jobcity	NaN	NaN	NaN	NaN	NaN	NaN	
gender	NaN	NaN	NaN	NaN	NaN	NaN	

dob	NaN	NaN	NaN	NaN	NaN	NaN
10percentage	NaN	0.044547	0.177373	NaN	NaN	NaN
10board	NaN	NaN	NaN	NaN	NaN	NaN
12graduation	NaN	0.673102	-0.161383	NaN	NaN	NaN
12percentage	NaN	0.007069	0.170254	NaN	NaN	NaN
12board	NaN	NaN	NaN	NaN	NaN	NaN
collegeid	NaN	0.284540	-0.118690	NaN	NaN	NaN
collegetier	NaN	0.035160	-0.179332	NaN	NaN	NaN
degree	NaN	NaN	NaN	NaN	NaN	NaN
specialization	NaN	NaN	NaN	NaN	NaN	NaN
collegegpa	NaN	0.047144	0.130103	NaN	NaN	NaN
collegecityid	NaN	0.284540	-0.118690	NaN	NaN	NaN
collegecitytier	NaN	-0.035977	0.015384	NaN	NaN	NaN
collegestate	NaN	NaN	NaN	NaN	NaN	NaN
graduationyear	NaN	0.027539	-0.010053	NaN	NaN	NaN
english	NaN	0.135505	0.178219	NaN	NaN	NaN
logical	NaN	0.102215	0.179275	NaN	NaN	NaN
quant	NaN	-0.055134	0.230627	NaN	NaN	NaN
domain	NaN	-0.125639	0.104656	NaN	NaN	NaN
computerprogramming	NaN	0.018859	0.115665	NaN	NaN	NaN
electronicsandsemicon	NaN	-0.115601	0.000665	NaN	NaN	NaN
computerscience	NaN	0.482626	-0.100720	NaN	NaN	NaN
mechanicalengg	NaN	-0.026147	0.018475	NaN	NaN	NaN
electricalengg	NaN	0.104454	-0.047598	NaN	NaN	NaN
telecomengg	NaN	-0.049272	-0.022691	NaN	NaN	NaN
civilengg	NaN	-0.017871	0.037639	NaN	NaN	NaN
conscientiousness	NaN	0.175557	-0.064148	NaN	NaN	NaN
agreeableness	NaN	0.024837	0.057423	NaN	NaN	NaN
extraversion	NaN	0.120979	-0.010213	NaN	NaN	NaN
nueroticism	NaN	-0.146289	-0.054685	NaN	NaN	NaN
openness_to_experience	NaN	0.031359	-0.011312	NaN	NaN	NaN

	jobcity	gender	dob	10percentage	...	\
column1	NaN	NaN	NaN	NaN	...	
id	NaN	NaN	NaN	0.044547	...	
salary	NaN	NaN	NaN	0.177373	...	
doj	NaN	NaN	NaN	NaN	...	
dol	NaN	NaN	NaN	NaN	...	
designation	NaN	NaN	NaN	NaN	...	
jobcity	NaN	NaN	NaN	NaN	...	
gender	NaN	NaN	NaN	NaN	...	
dob	NaN	NaN	NaN	NaN	...	
10percentage	NaN	NaN	NaN	1.000000	...	
10board	NaN	NaN	NaN	NaN	...	
12graduation	NaN	NaN	NaN	0.269957	...	
12percentage	NaN	NaN	NaN	0.643378	...	
12board	NaN	NaN	NaN	NaN	...	
collegeid	NaN	NaN	NaN	0.021082	...	
collegetier	NaN	NaN	NaN	-0.126042	...	
degree	NaN	NaN	NaN	NaN	...	
specialization	NaN	NaN	NaN	NaN	...	
collegegpa	NaN	NaN	NaN	0.312538	...	
collegecityid	NaN	NaN	NaN	0.021082	...	
collegecitytier	NaN	NaN	NaN	0.116707	...	
collegestate	NaN	NaN	NaN	NaN	...	
graduationyear	NaN	NaN	NaN	-0.013799	...	
english	NaN	NaN	NaN	0.350780	...	
logical	NaN	NaN	NaN	0.316014	...	
quant	NaN	NaN	NaN	0.317640	...	
domain	NaN	NaN	NaN	0.078563	...	
computerprogramming	NaN	NaN	NaN	0.053600	...	
electronicsandsemicon	NaN	NaN	NaN	0.085179	...	
computerscience	NaN	NaN	NaN	-0.018933	...	
mechanicalengg	NaN	NaN	NaN	0.050364	...	
electricalengg	NaN	NaN	NaN	0.074419	...	
telecomengg	NaN	NaN	NaN	0.049378	...	
civilengg	NaN	NaN	NaN	0.030002	...	
conscientiousness	NaN	NaN	NaN	0.067657	...	
agreeableness	NaN	NaN	NaN	0.136645	...	
extraversion	NaN	NaN	NaN	-0.004679	...	
nueroticism	NaN	NaN	NaN	-0.132496	...	
openness_to_experience	NaN	NaN	NaN	0.036692	...	

	computerscience	mechanicalengg	electricalengg	\
column1	NaN	NaN	NaN	
id	0.482626	-0.026147	0.104454	
salary	-0.100720	0.018475	-0.047598	
doj	NaN	NaN	NaN	
dol	NaN	NaN	NaN	
designation	NaN	NaN	NaN	
jobcity	NaN	NaN	NaN	
gender	NaN	NaN	NaN	
dob	NaN	NaN	NaN	
10percentage	-0.018933	0.050364	0.074419	
10board	NaN	NaN	NaN	
12graduation	0.293439	0.035459	0.123751	
12percentage	-0.043534	0.037635	0.064001	
12board	NaN	NaN	NaN	
collegeid	0.102303	-0.009291	0.022933	

collegetier	0.001053	-0.021548	0.002594
degree	NaN	NaN	NaN
specialization	NaN	NaN	NaN
collegegpa	0.007601	-0.031765	0.052258
collegecityid	0.102303	-0.009291	0.022933
collegecitytier	-0.010643	-0.052395	0.010311
collegestate	NaN	NaN	NaN
graduationyear	0.024089	-0.066844	0.008525
english	0.059500	-0.002477	0.032438
logical	0.044481	-0.009861	0.012003
quant	-0.043379	0.019933	0.020975
domain	0.058762	0.048472	0.042875
computerprogramming	0.253039	-0.284891	-0.138224
electronicsandsemicon	-0.273707	-0.109434	0.036968
computerscience	1.000000	-0.124355	-0.083798
mechanicalengg	-0.124355	1.000000	-0.040522
electricalengg	-0.083798	-0.040522	1.000000
telecomengg	-0.148095	-0.070947	-0.051469
civilengg	-0.052613	0.076201	-0.020059
conscientiousness	0.090155	-0.010858	0.029806
agreeableness	0.039866	-0.028586	-0.015454
extraversion	0.102153	-0.017748	0.004467
nueroticism	-0.112652	0.036148	-0.030870
openness_to_experience	0.058039	-0.027988	-0.012585

	telecomengg	civilengg	conscientiousness	\
column1	NaN	NaN	NaN	
id	-0.049272	-0.017871	0.175557	
salary	-0.022691	0.037639	-0.064148	
doj	NaN	NaN	NaN	
dol	NaN	NaN	NaN	
designation	NaN	NaN	NaN	
jobcity	NaN	NaN	NaN	
gender	NaN	NaN	NaN	
dob	NaN	NaN	NaN	
10percentage	0.049378	0.030002	0.067657	
10board	NaN	NaN	NaN	
12graduation	0.023470	-0.004727	0.103329	
12percentage	0.044201	0.005910	0.058299	
12board	NaN	NaN	NaN	
collegeid	0.025620	0.005749	0.076432	
collegetier	0.000007	-0.033722	0.055174	
degree	NaN	NaN	NaN	
specialization	NaN	NaN	NaN	
collegegpa	-0.005226	-0.018950	0.069582	
collegecityid	0.025620	0.005749	0.076432	
collegecitytier	0.049876	-0.033392	0.014763	
collegestate	NaN	NaN	NaN	
graduationyear	0.004226	0.001696	-0.013235	
english	-0.005822	-0.007724	0.034943	
logical	-0.012947	-0.011286	0.025876	
quant	0.021387	0.000528	-0.005639	
domain	0.024442	0.017569	-0.039478	
computerprogramming	-0.248269	-0.088249	0.012862	
electronicsandsemicon	0.387140	0.002863	-0.026483	
computerscience	-0.148095	-0.052613	0.090155	
mechanicalengg	-0.070947	0.076201	-0.010858	
electricalengg	-0.051469	-0.020059	0.029806	
telecomengg	1.000000	-0.031492	-0.004946	
civilengg	-0.031492	1.000000	-0.017526	
conscientiousness	-0.004946	-0.017526	1.000000	
agreeableness	-0.014627	-0.034254	0.481820	
extraversion	-0.039050	-0.031822	0.355537	
nueroticism	0.020638	0.010555	-0.330312	
openness_to_experience	-0.000141	-0.031201	0.395649	

	agreeableness	extraversion	nueroticism	\
column1	NaN	NaN	NaN	
id	0.024837	0.120979	-0.146289	
salary	0.057423	-0.010213	-0.054685	
doj	NaN	NaN	NaN	
dol	NaN	NaN	NaN	
designation	NaN	NaN	NaN	
jobcity	NaN	NaN	NaN	
gender	NaN	NaN	NaN	
dob	NaN	NaN	NaN	
10percentage	0.136645	-0.004679	-0.132496	
10board	NaN	NaN	NaN	
12graduation	0.041182	0.061956	-0.074369	
12percentage	0.103998	-0.007486	-0.094369	
12board	NaN	NaN	NaN	
collegeid	-0.005264	0.005917	-0.008973	
collegetier	-0.038055	0.009970	0.023778	
degree	NaN	NaN	NaN	
specialization	NaN	NaN	NaN	
collegegpa	0.068282	-0.032684	-0.074859	
collegecityid	-0.005264	0.005917	-0.008973	
collegecitytier	0.005565	-0.008203	0.004442	
collegestate	NaN	NaN	NaN	

graduationyear	-0.002877	0.008397	-0.000417
english	0.194990	0.018755	-0.155528
logical	0.167207	-0.006949	-0.178781
quant	0.103443	-0.028616	-0.131895
domain	0.051944	-0.024647	-0.017928
computerprogramming	0.076934	0.043504	-0.084344
electronicsandsemicon	-0.024286	-0.044458	0.021026
computerscience	0.039866	0.102153	-0.112652
mechanicalengg	-0.028586	-0.017748	0.036148
electricalengg	-0.015454	0.004467	-0.030870
telecomengg	-0.014627	-0.039050	0.020638
civilengg	-0.034254	-0.031822	0.010555
conscientiousness	0.481820	0.355537	-0.330312
agreeableness	1.000000	0.454369	-0.207480
extraversion	0.454369	1.000000	-0.096491
nueroticism	-0.207480	-0.096491	1.000000
openness_to_experience	0.591541	0.435074	-0.065795

	openness_to_experience
column1	NaN
id	0.031359
salary	-0.011312
doj	NaN
dol	NaN
designation	NaN
jobcity	NaN
gender	NaN
dob	NaN
10percentage	0.036692
10board	NaN
12graduation	-0.015069
12percentage	0.006332
12board	NaN
collegeid	-0.010678
collegetier	-0.019179
degree	NaN
specialization	NaN
collegegpa	0.028071
collegecityid	-0.010678
collegecitytier	-0.016790
collegestate	NaN
graduationyear	0.016855
english	0.067979
logical	0.048420
quant	0.020377
domain	0.010412
computerprogramming	0.043133
electronicsandsemicon	-0.013460
computerscience	0.058039
mechanicalengg	-0.027988
electricalengg	-0.012585
telecomengg	-0.000141
civilengg	-0.031201
conscientiousness	0.395649
agreeableness	0.591541
extraversion	0.435074
nueroticism	-0.065795
openness_to_experience	1.000000

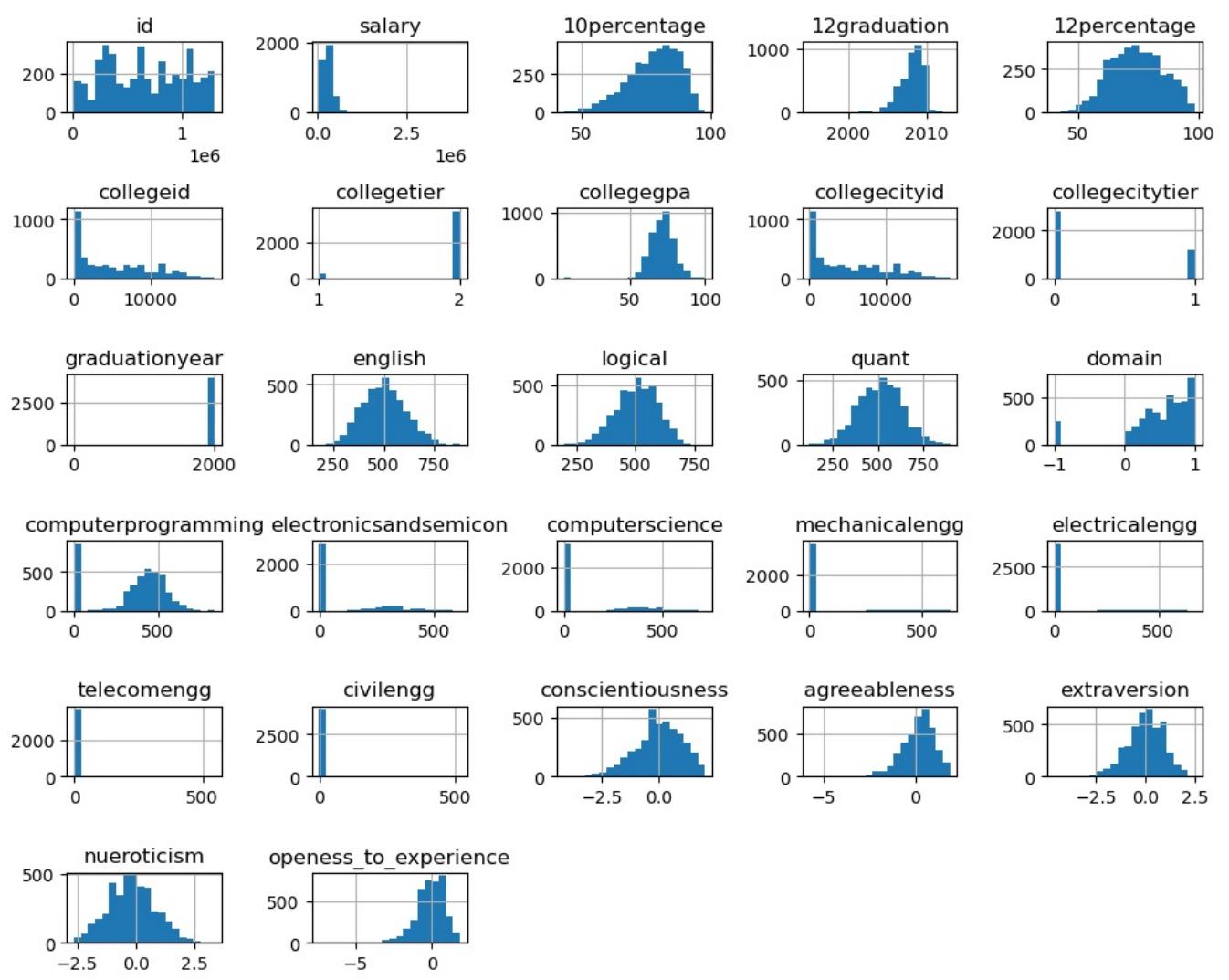
[39 rows x 39 columns]

```
In [18]: #Histogram plot of numerical columns
import pandas as pd
import matplotlib.pyplot as plt

# Assuming df is your DataFrame
df_numeric = df.select_dtypes(include=[np.number])

# Generate histograms for all numeric columns
df_numeric.hist(figsize=(10, 8), bins=20)

# Display the plots
plt.tight_layout()
plt.show()
```



```
In [19]: import pandas as pd
import matplotlib.pyplot as plt

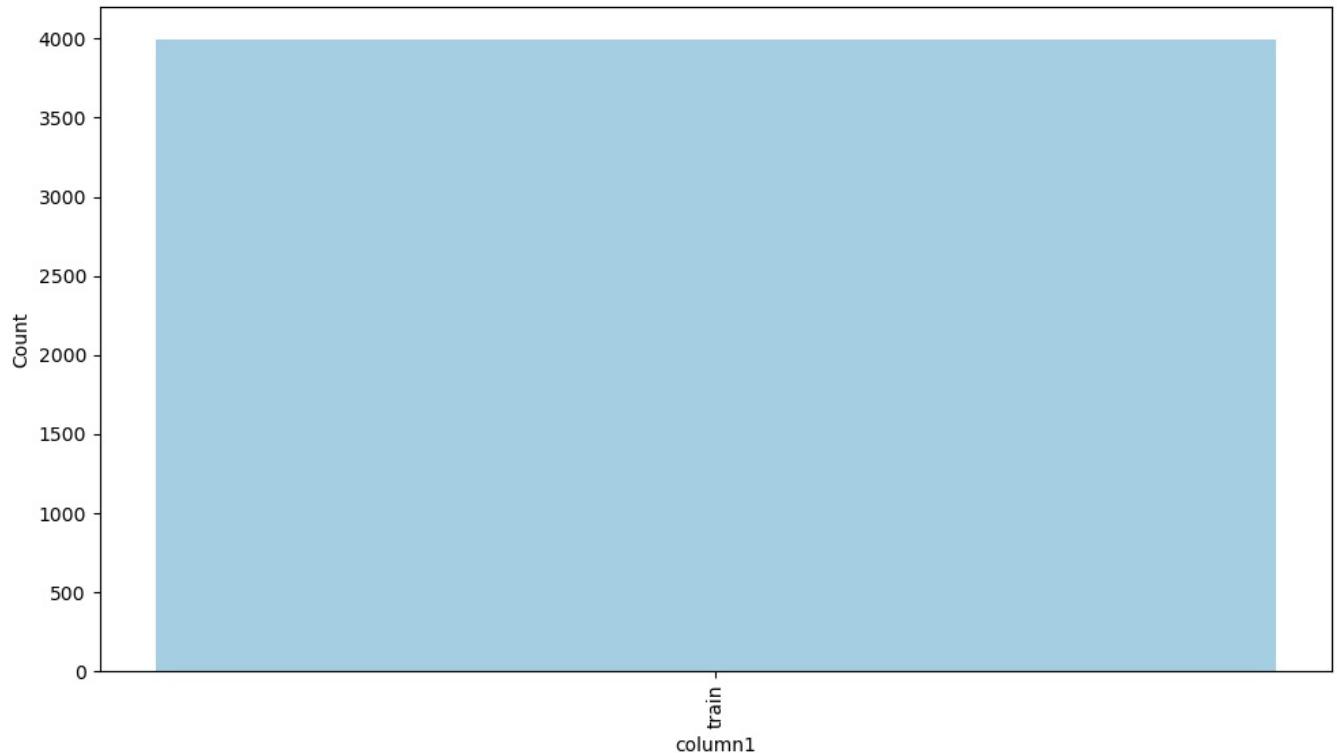
# List of your categorical columns
categorical_columns = ['column1', 'doj', 'dol', 'designation', 'jobcity', 'gender', 'dob',
                      '10board', '12board', 'degree', 'specialization', 'collegestate']

# Loop through each categorical column and plot a bar chart
for column in categorical_columns:
    plt.figure(figsize=(10, 6)) # Adjust figure size as needed
    category_counts = df[column].value_counts()

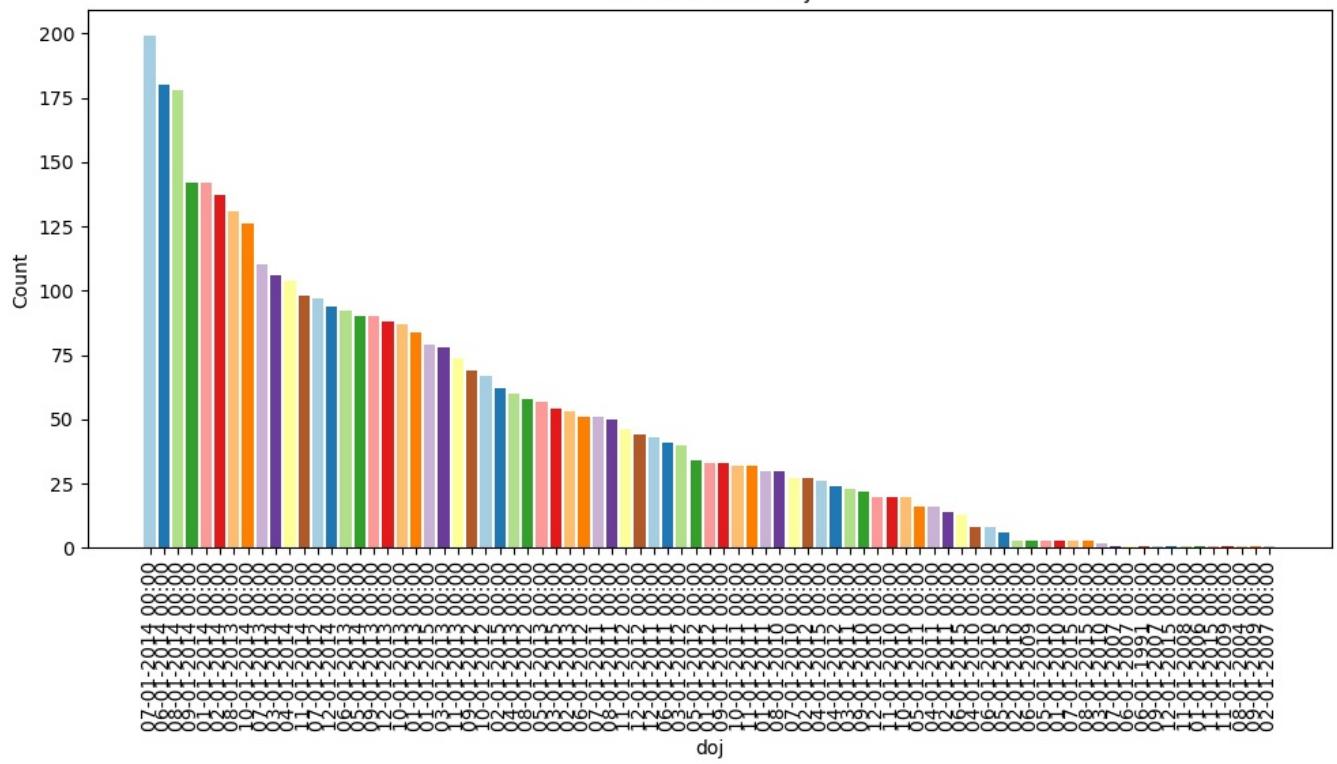
    # Plot the bar chart
    plt.bar(category_counts.index, category_counts.values, color=plt.cm.Paired.colors)
    plt.title(f'Distribution of {column}')
    plt.xlabel(column)
    plt.ylabel('Count')
    plt.xticks(rotation=90) # Rotate the x-axis labels for better readability

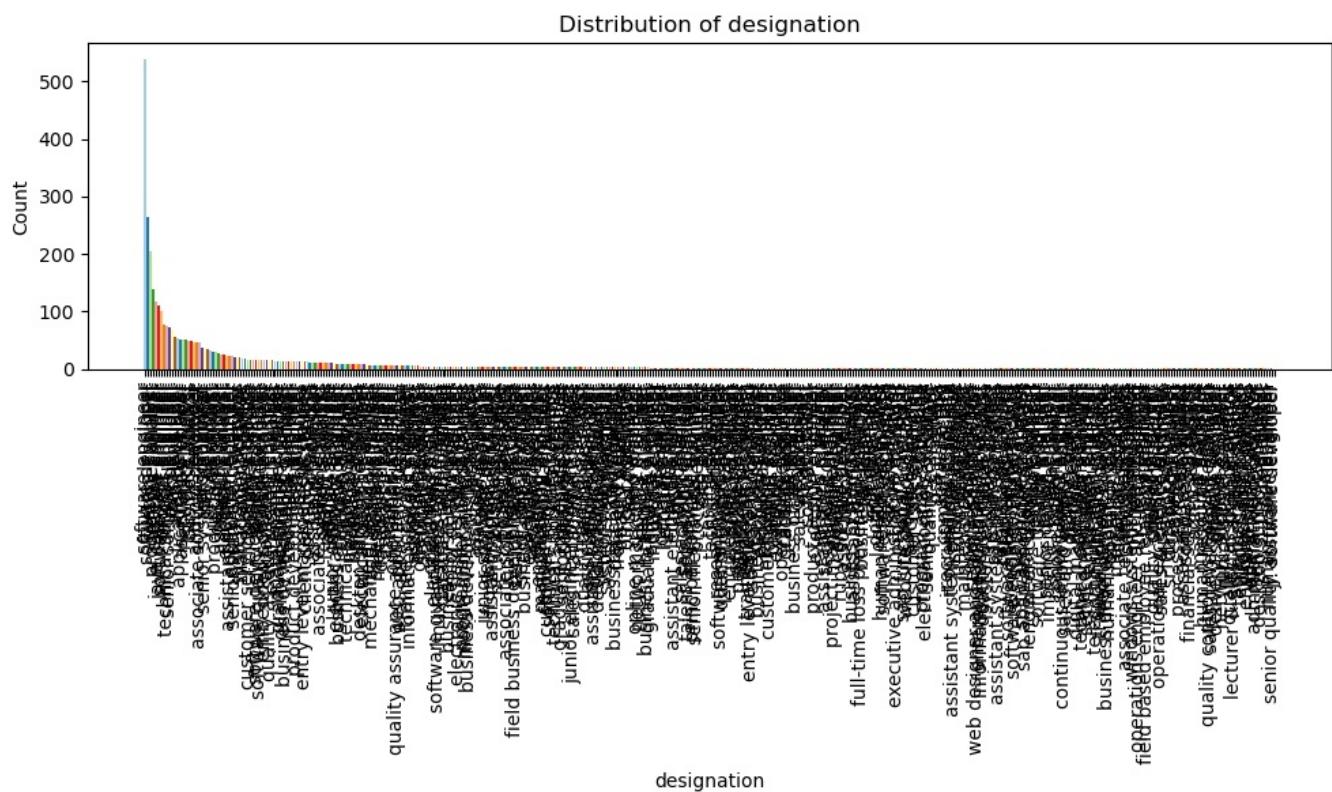
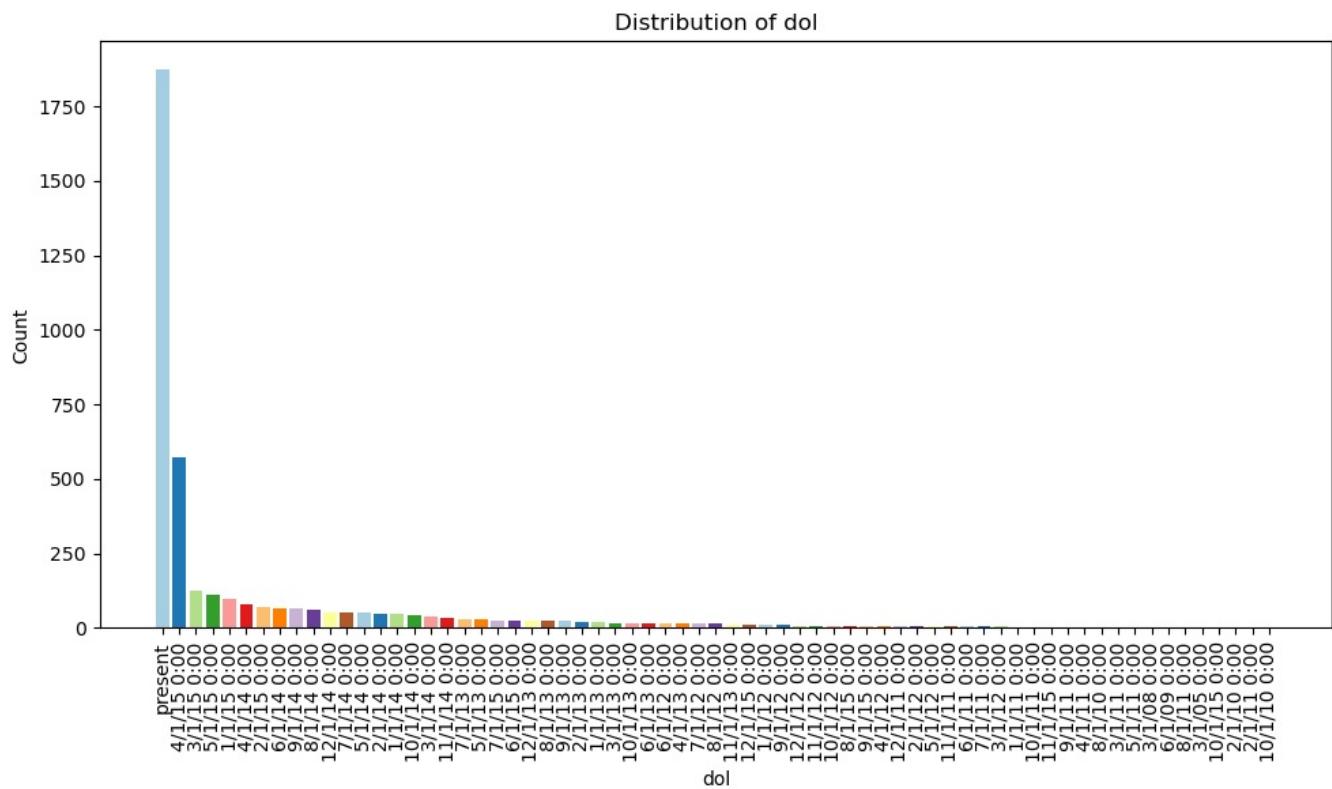
    # Show the plot
    plt.tight_layout()
    plt.show()
```

Distribution of column1

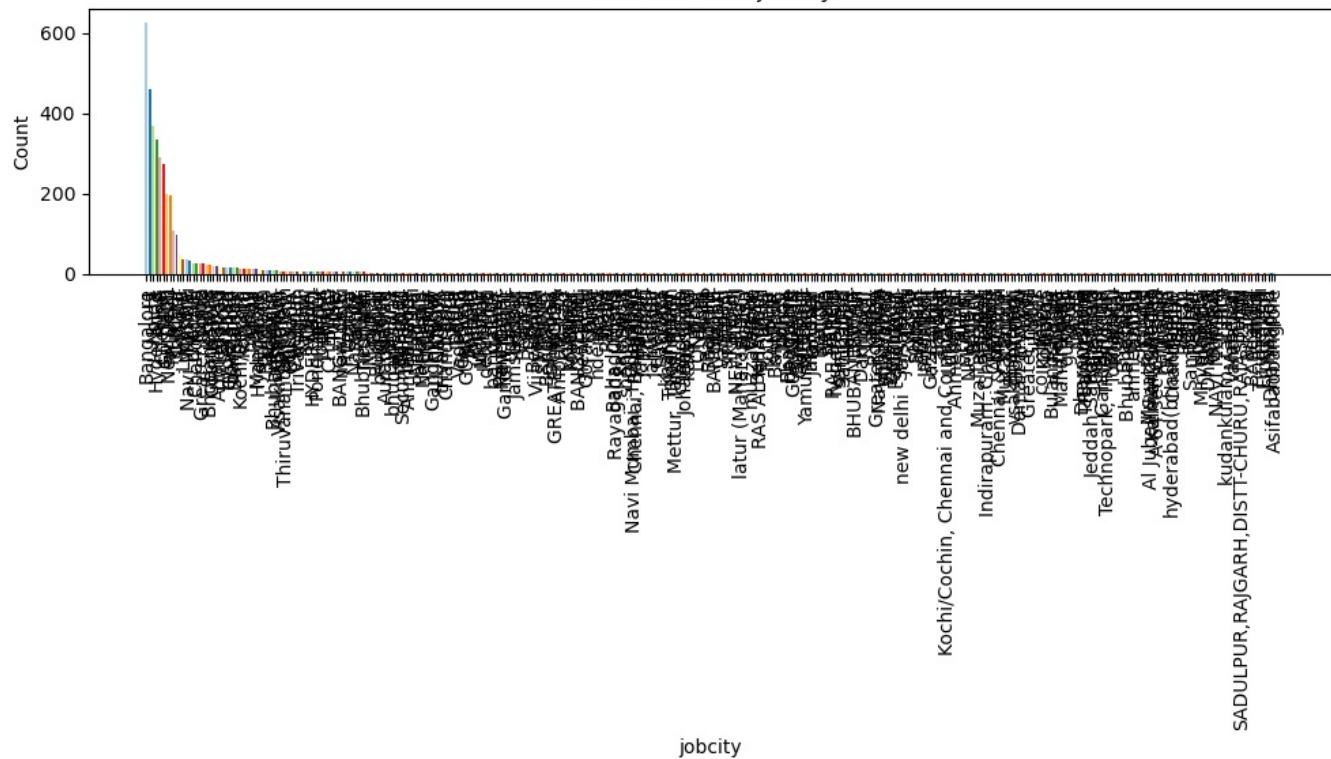


Distribution of doj

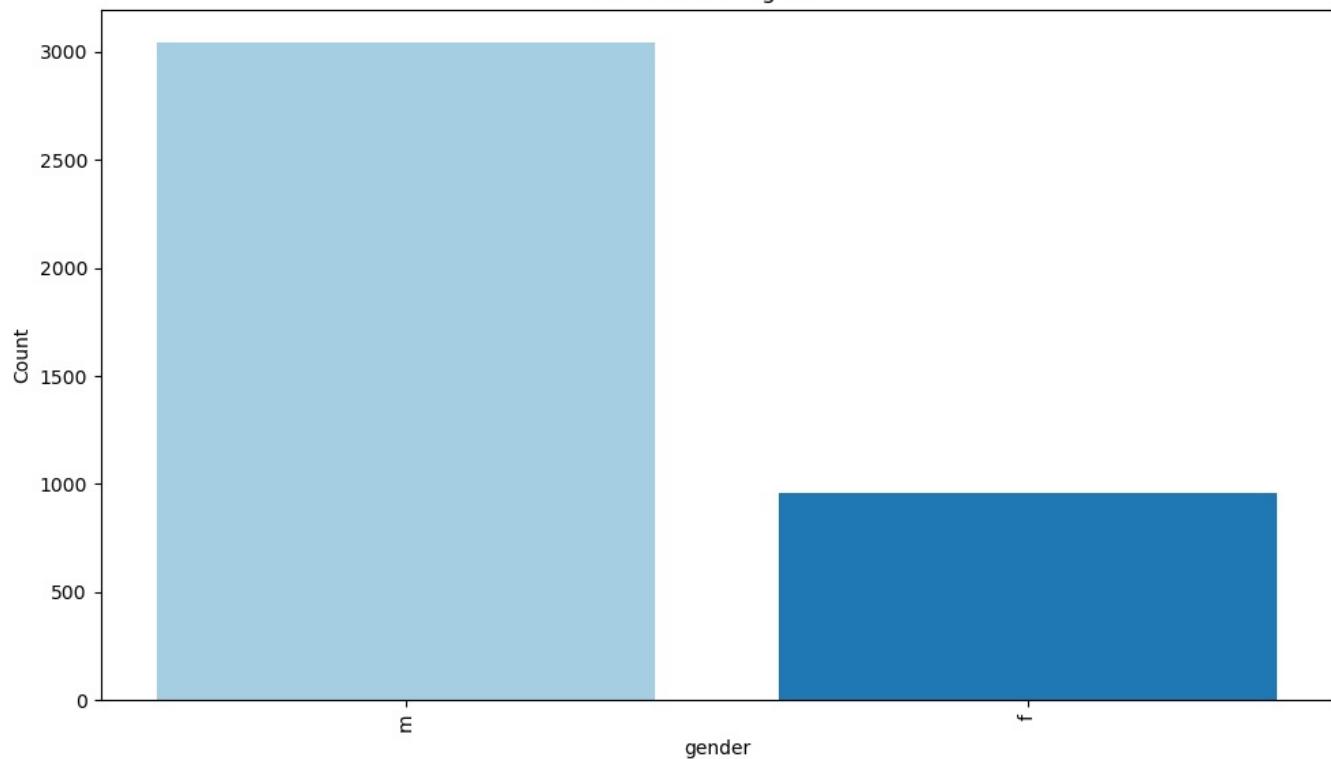




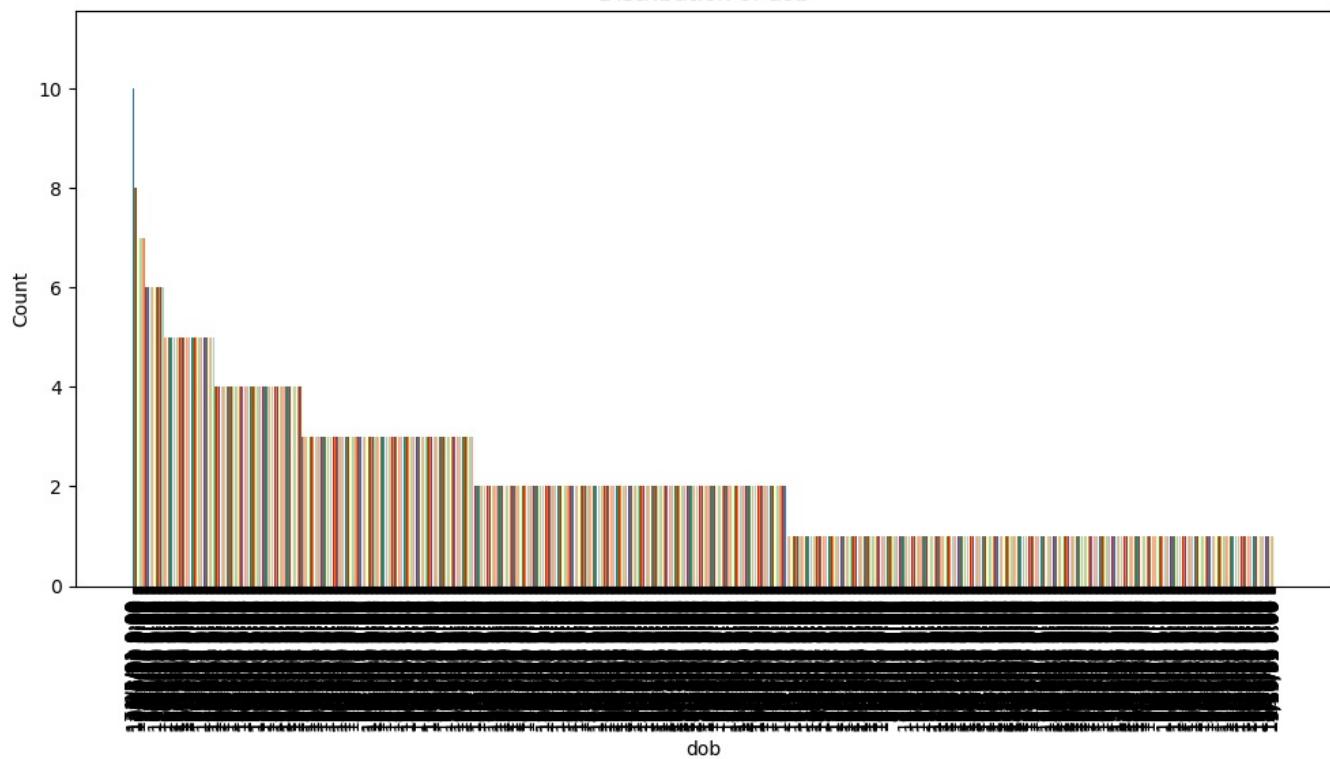
Distribution of jobcity



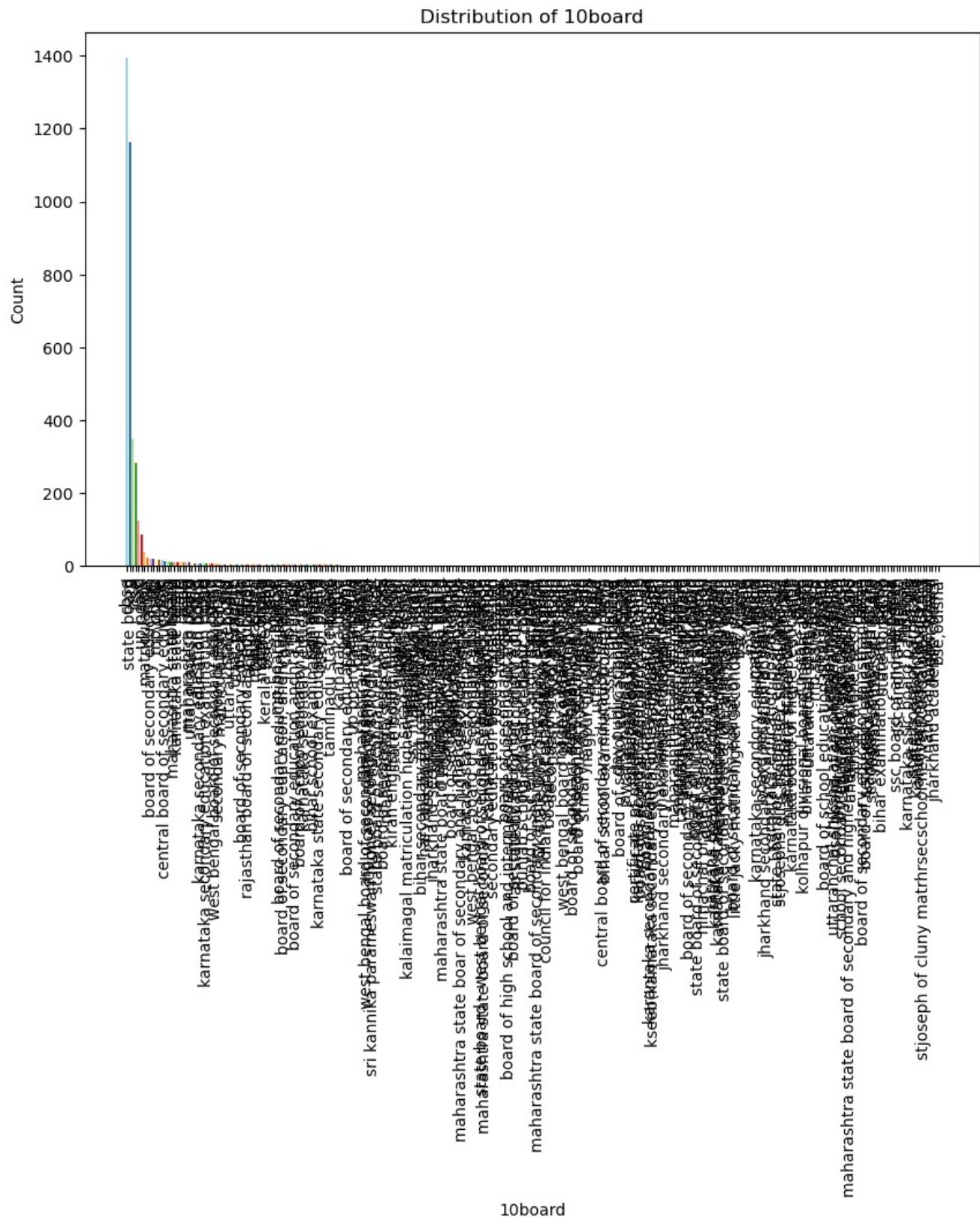
Distribution of gender



Distribution of dob

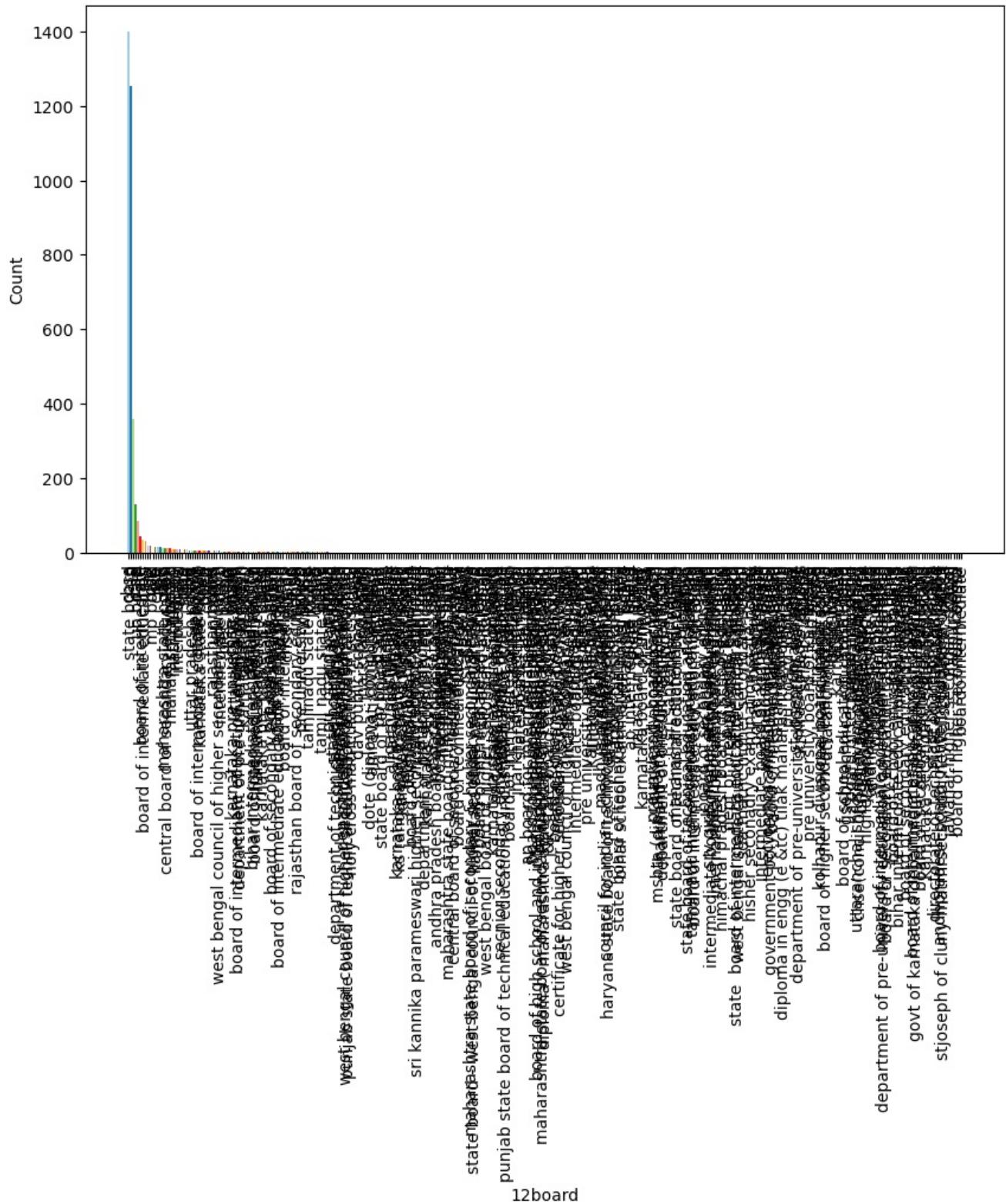


C:\Users\shannm\AppData\Local\Temp\ipykernel\_26380\2272193603.py:21: UserWarning: Tight layout not applied. The bottom and top margins cannot be made large enough to accommodate all axes decorations.  
plt.tight\_layout()

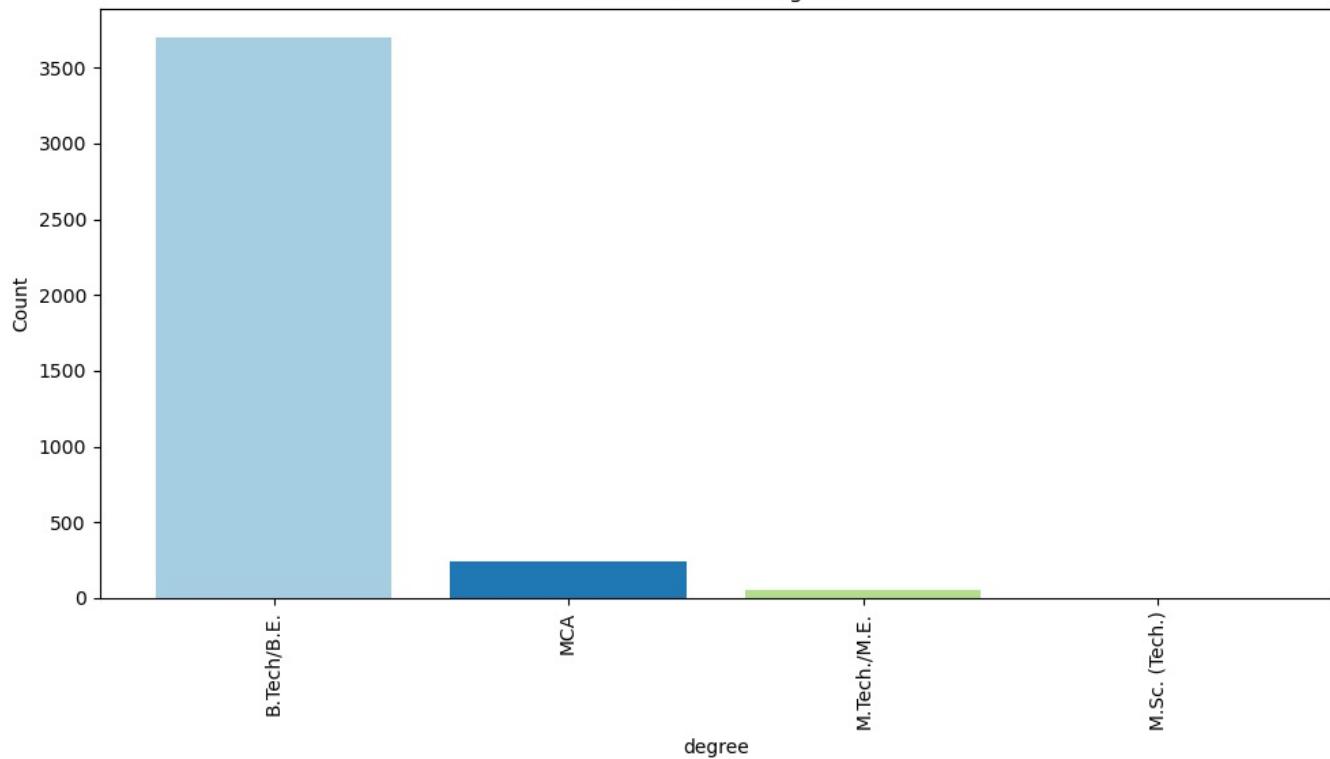


```
C:\Users\shannm\AppData\Local\Temp\ipykernel_26380\2272193603.py:21: UserWarning: Tight layout not applied. The bottom and top margins cannot be made large enough to accommodate all axes decorations.  
    plt.tight_layout()
```

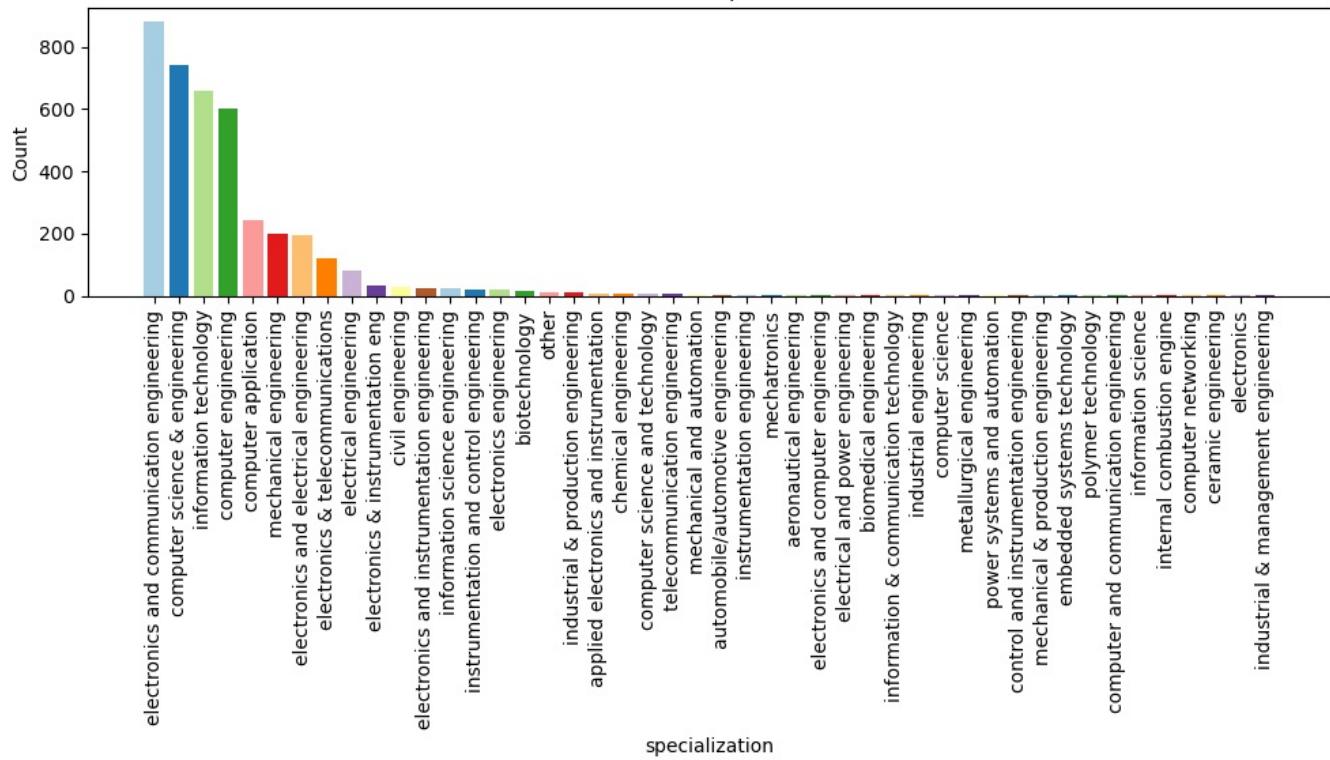
### Distribution of 12board



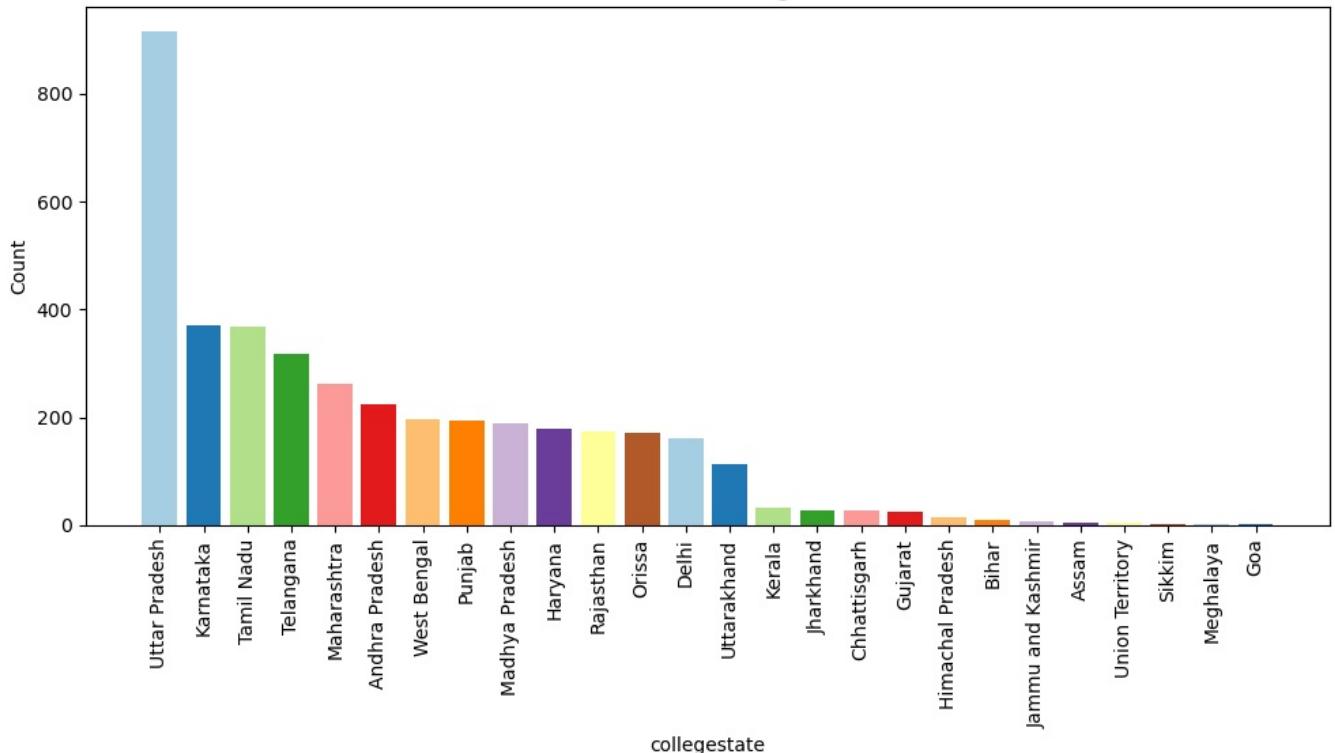
Distribution of degree



Distribution of specialization



### Distribution of collegestate

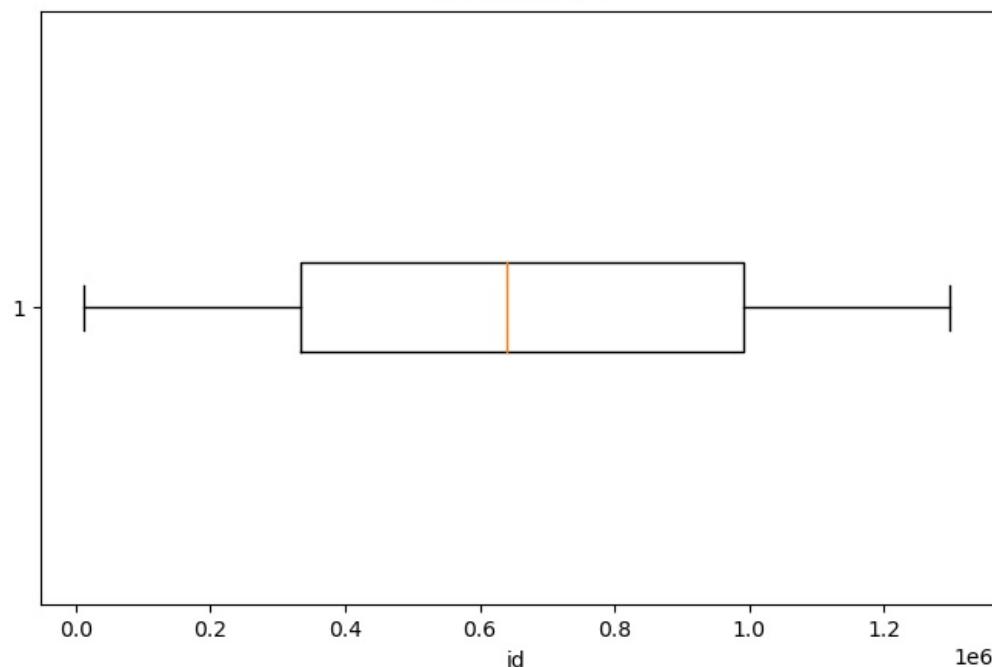


```
In [20]: import pandas as pd
import matplotlib.pyplot as plt

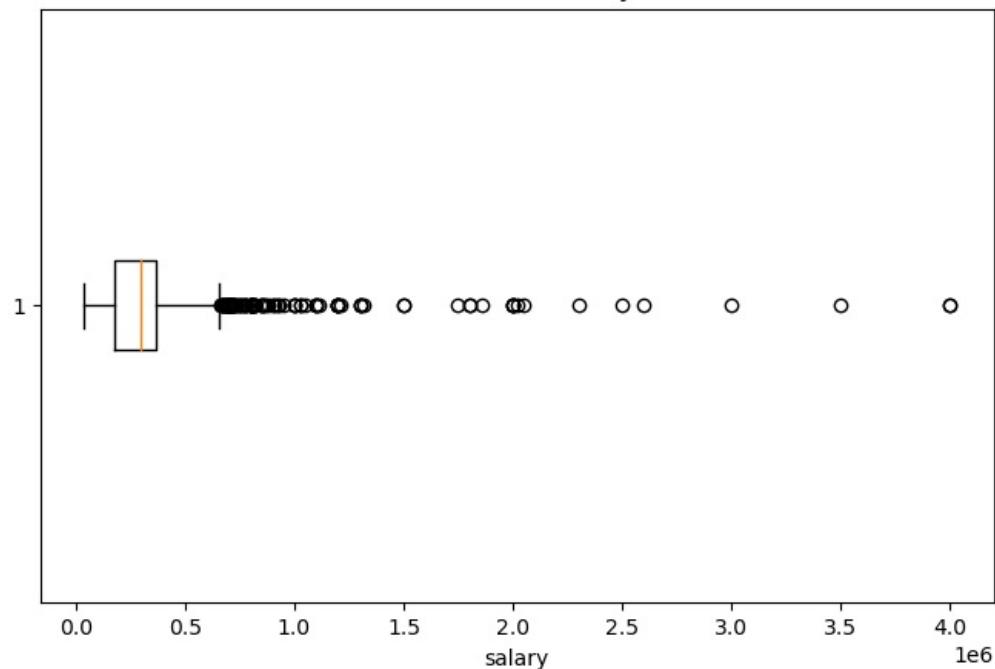
# Assuming df is your DataFrame and num_columns is your list of numerical columns
num_columns = ['id', 'salary', '10percentage', '12graduation', '12percentage',
               'collegeid', 'collegetier', 'collegegpa', 'collegecityid',
               'collegecitytier', 'graduationyear', 'english', 'logical', 'quant',
               'domain', 'computerprogramming', 'electronicsandsemicon',
               'computerscience', 'mechanicalengg', 'electricalengg', 'telecomengg',
               'civilengg', 'conscientiousness', 'agreeableness', 'extraversion',
               'nueroticism', 'openess_to_experience']

# Create box plots for each numerical column
for column in num_columns:
    plt.figure(figsize=(8, 5))
    plt.boxplot(df[column], vert=False) # Horizontal box plot
    plt.xlabel(column)
    plt.title(f'Box Plot for {column}')
    plt.show()
```

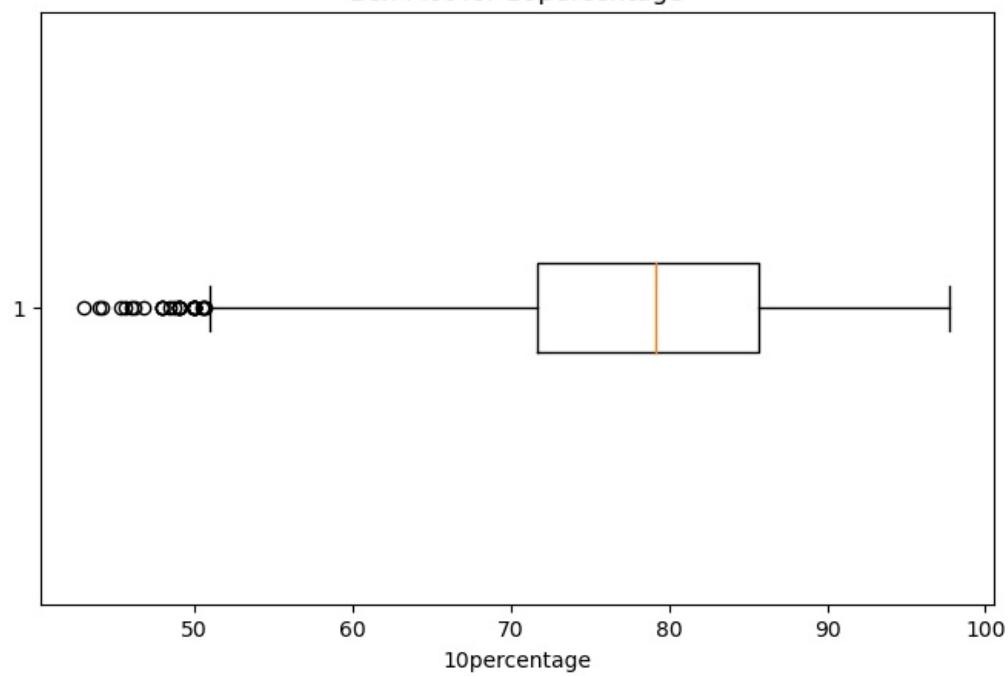
Box Plot for id



Box Plot for salary



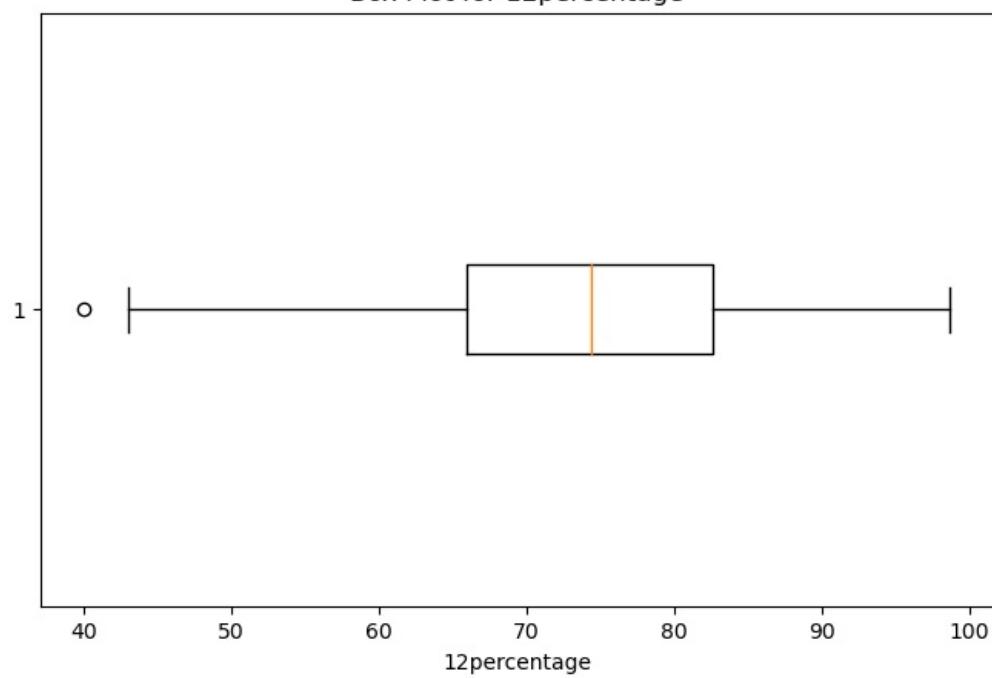
Box Plot for 10percentage



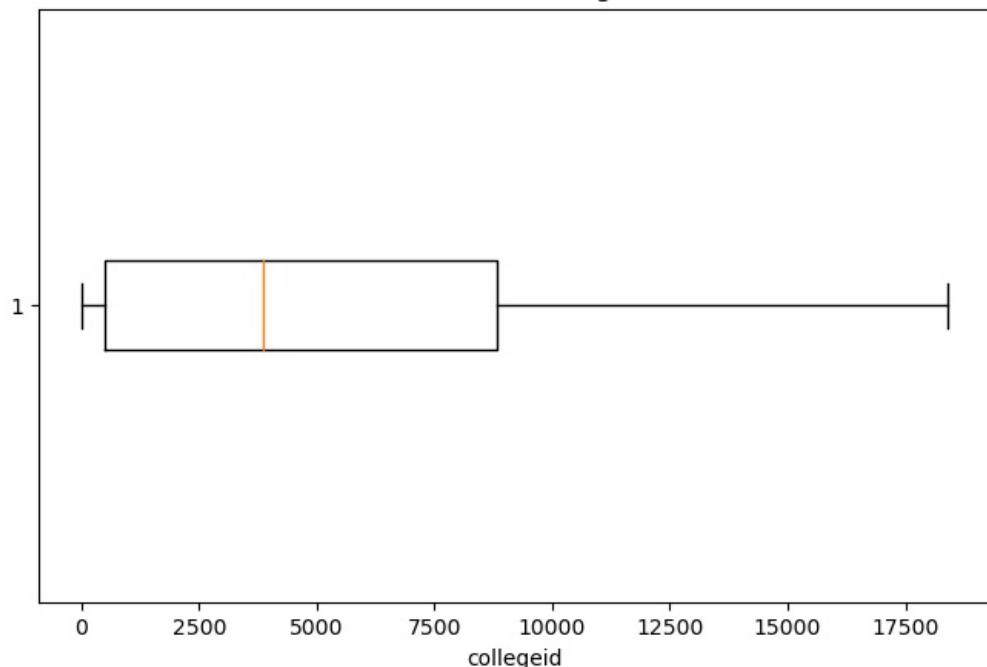
Box Plot for 12graduation



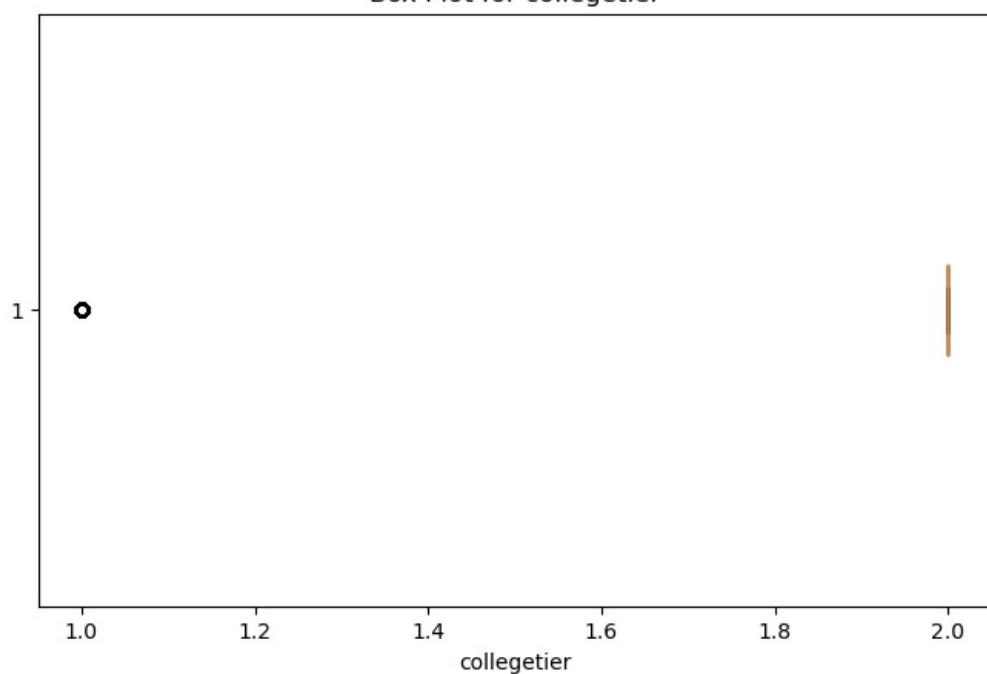
Box Plot for 12percentage



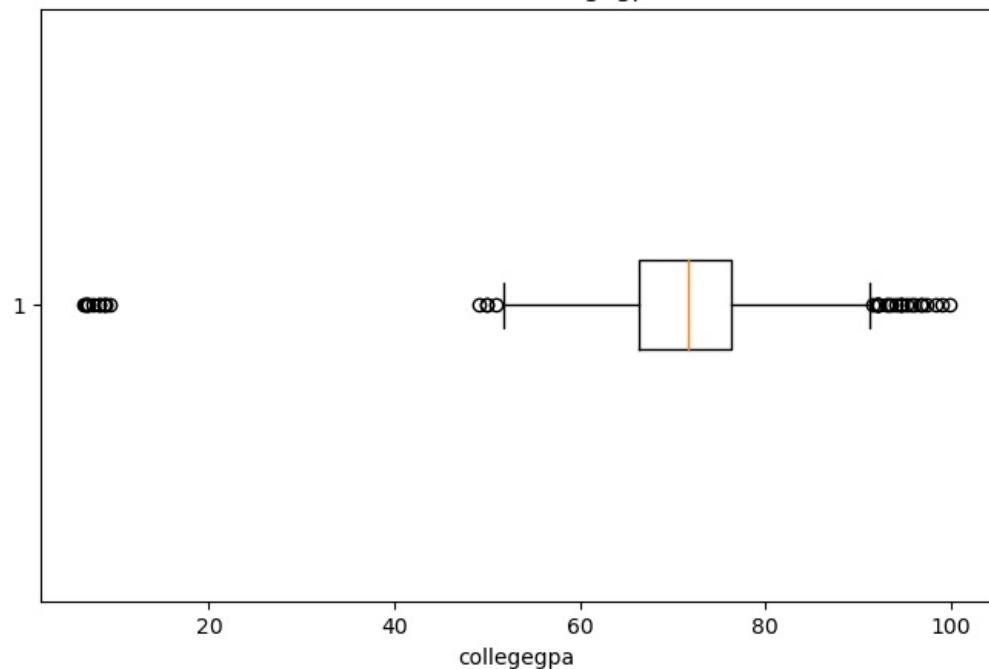
Box Plot for collegeid



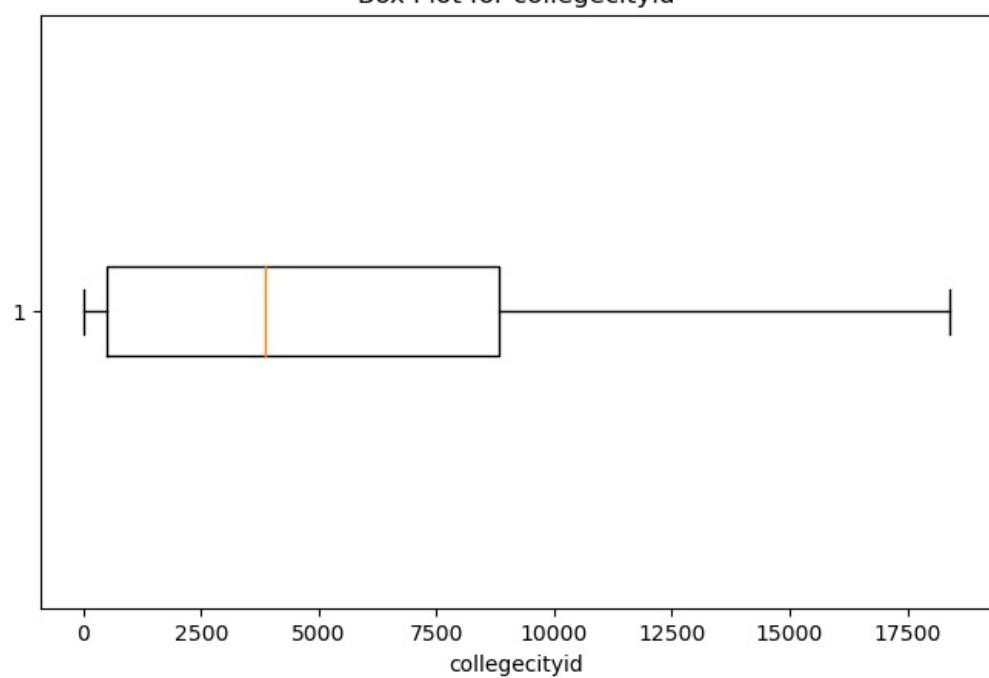
Box Plot for collegetier



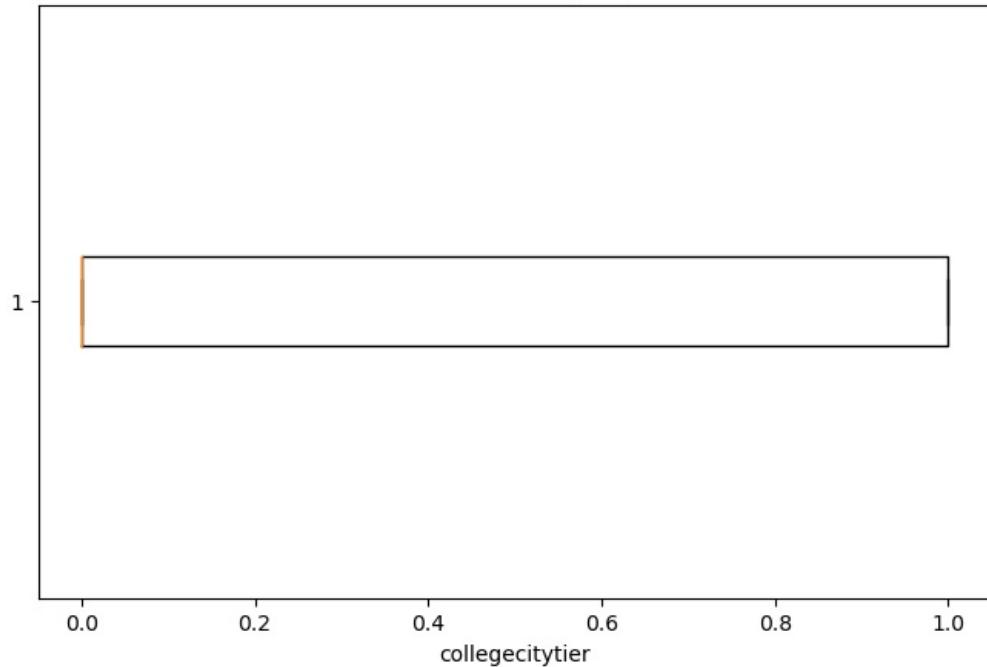
Box Plot for collegegpa



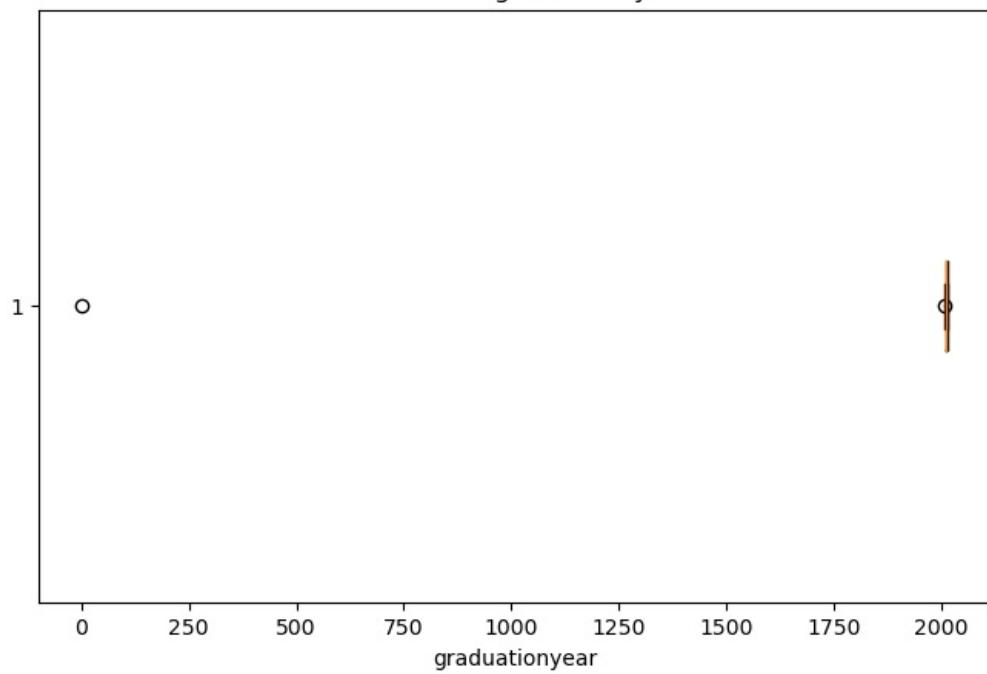
Box Plot for collegecityid



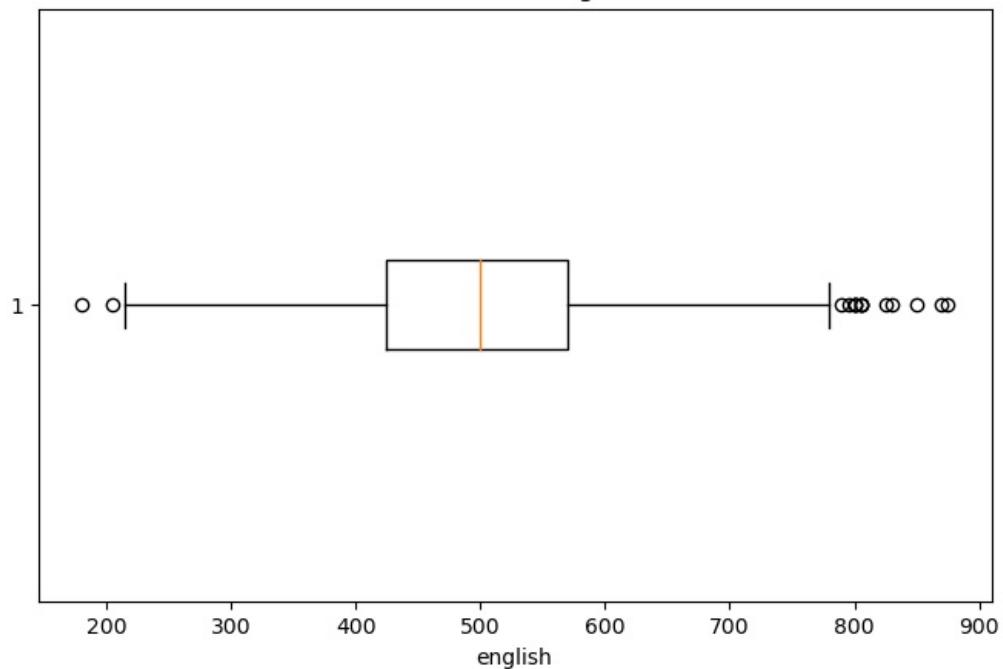
Box Plot for collegecitytier



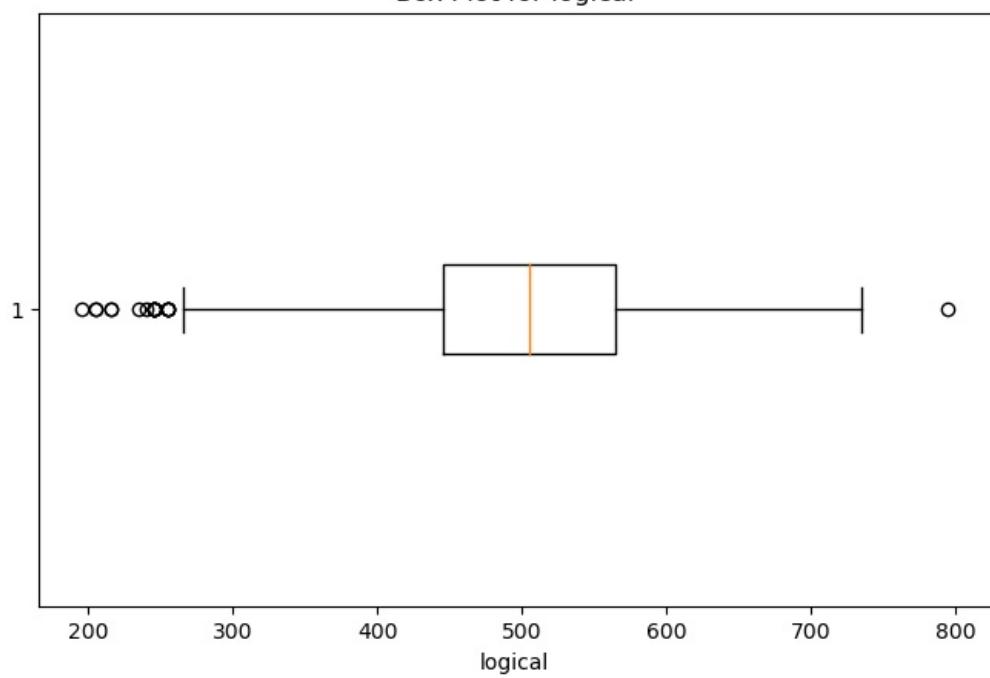
Box Plot for graduationyear



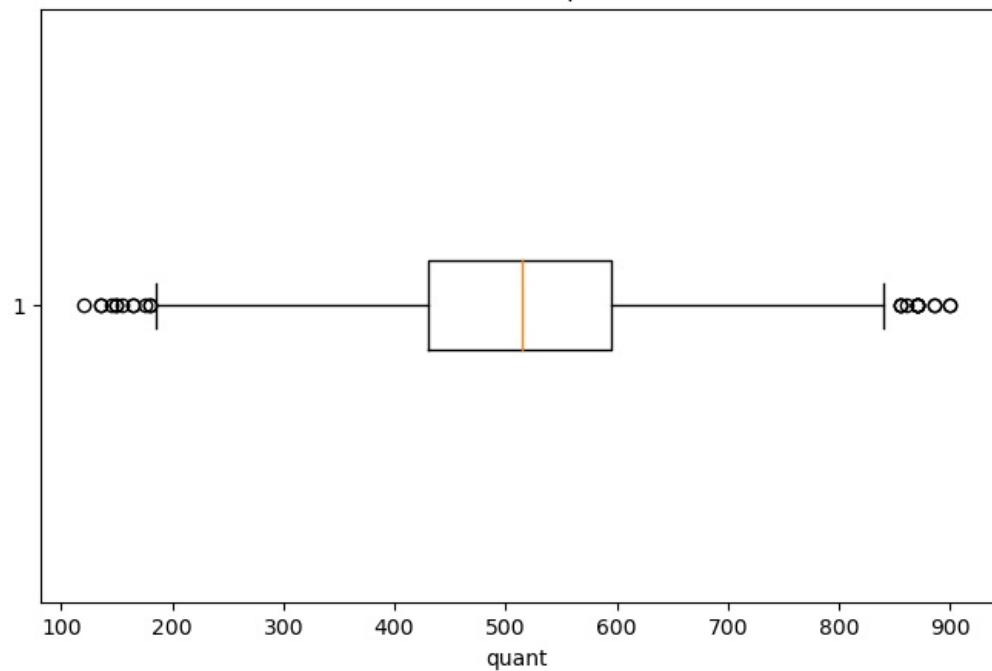
Box Plot for english



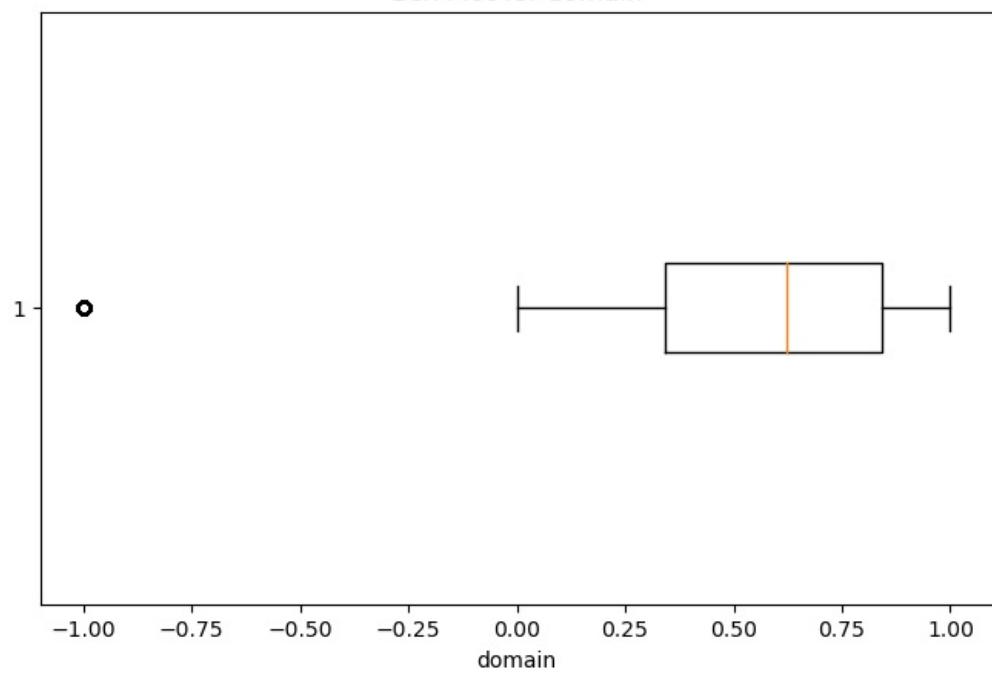
Box Plot for logical



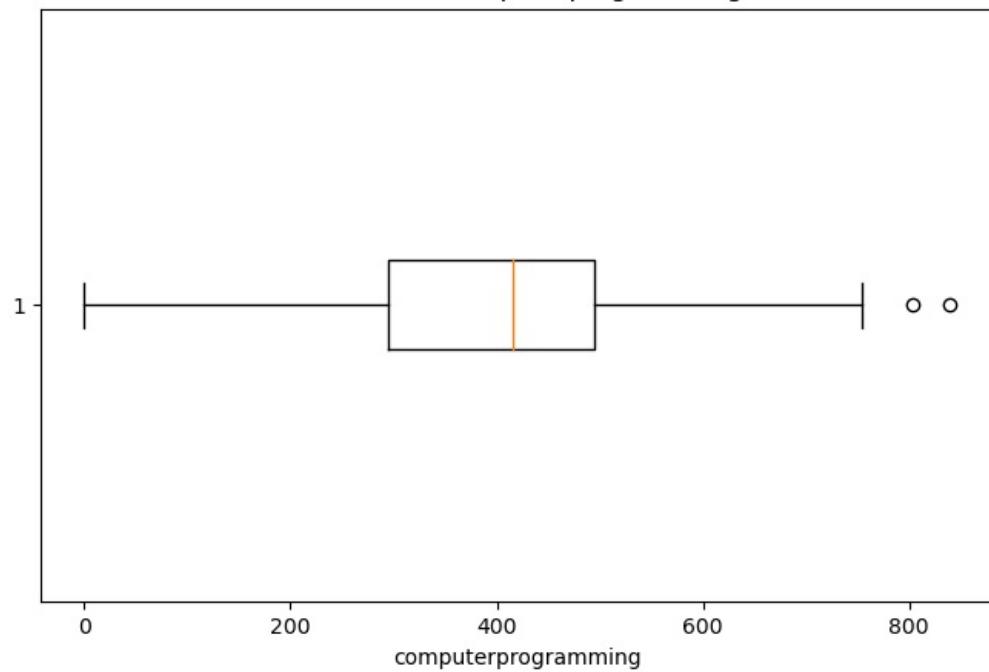
Box Plot for quant



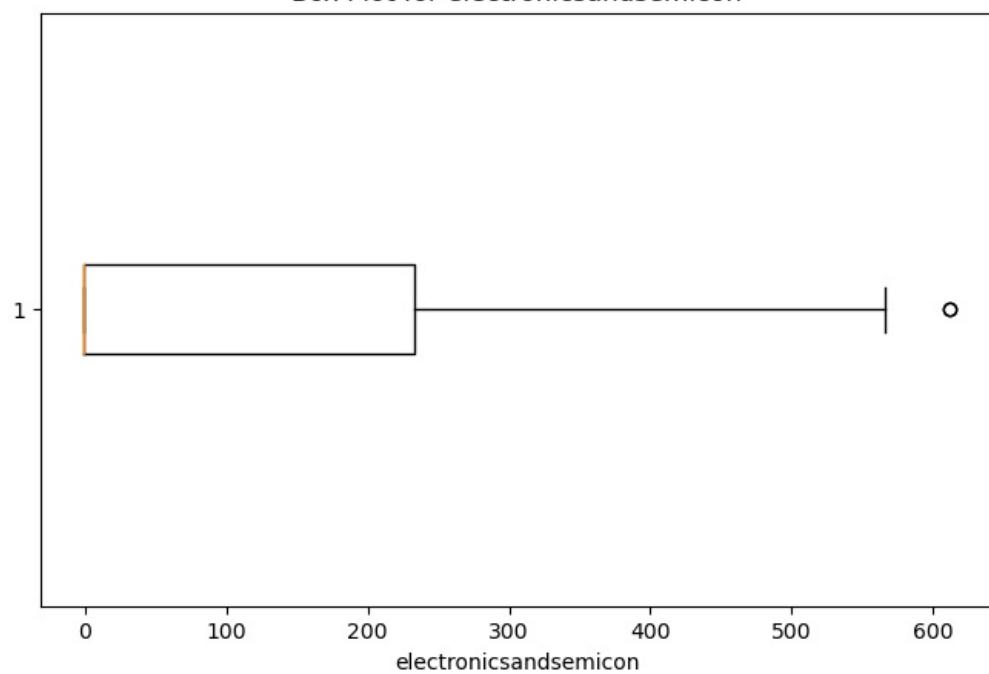
Box Plot for domain



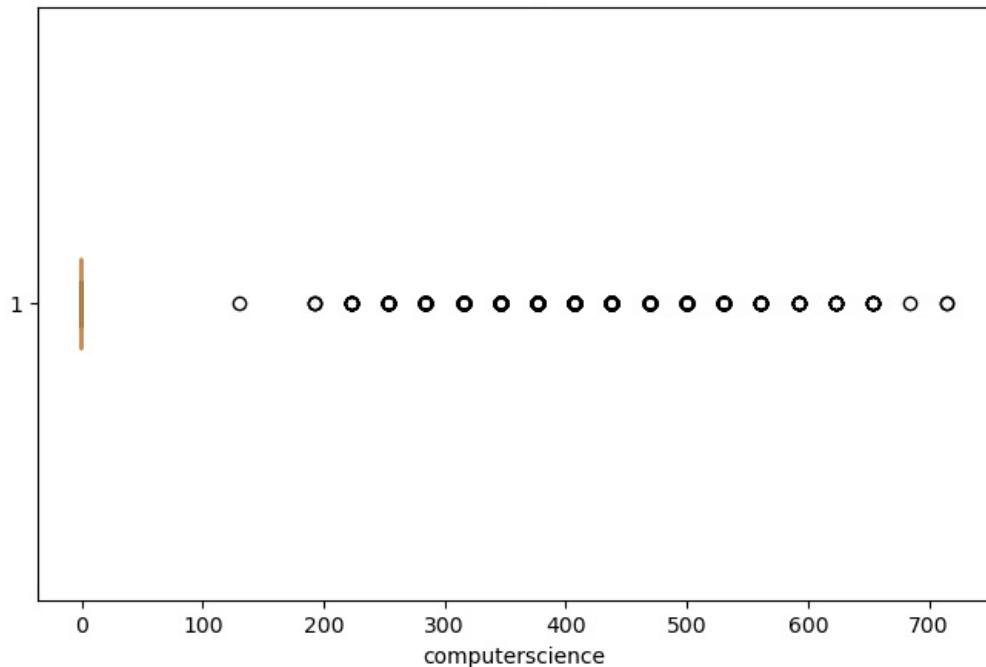
Box Plot for computerprogramming



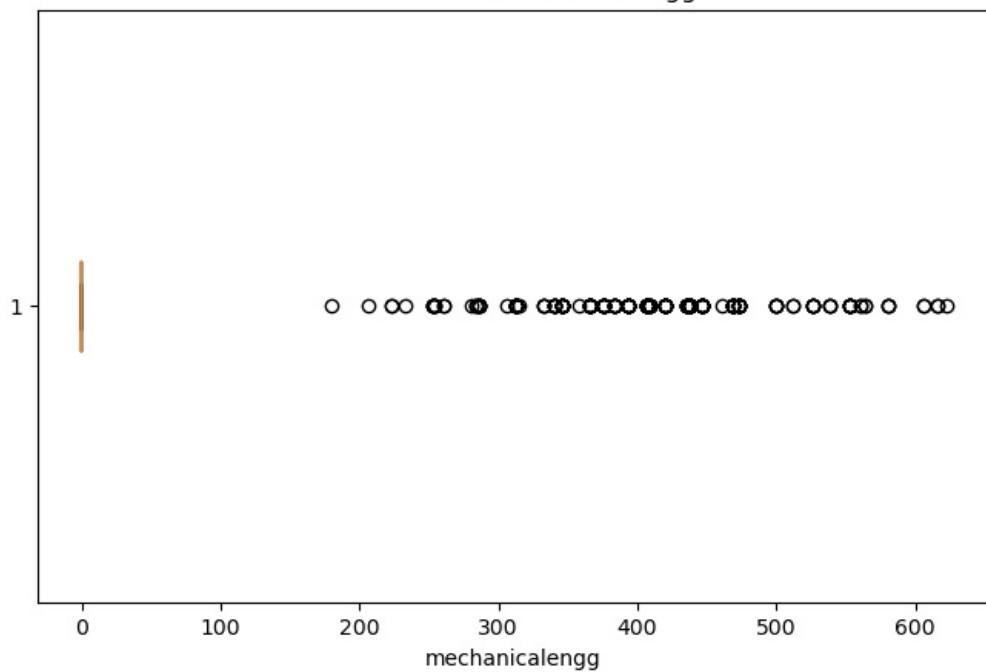
Box Plot for electronicsandsemicon



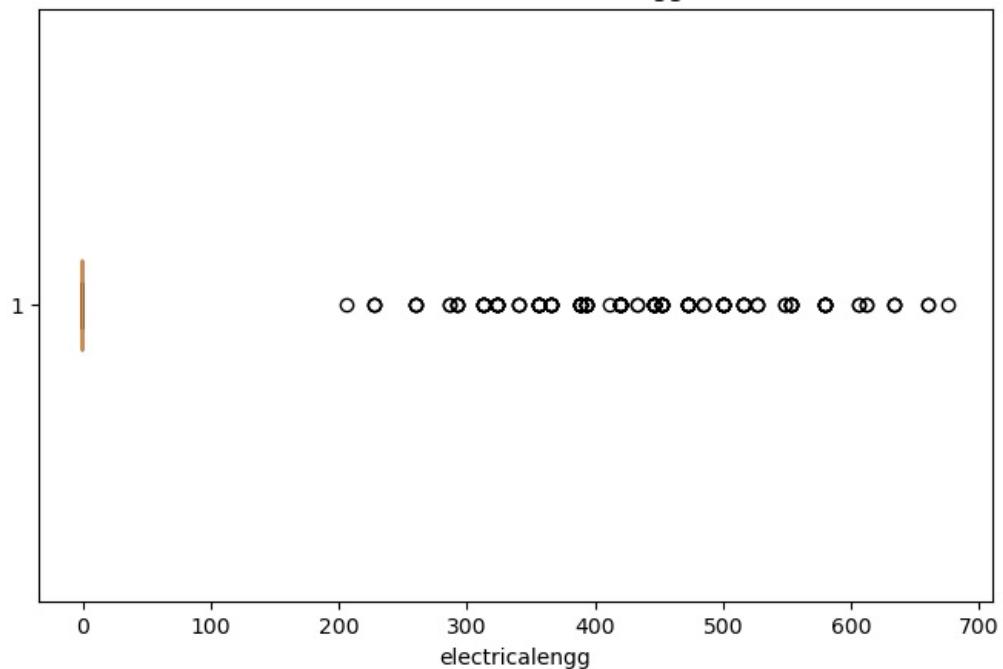
Box Plot for computergroup



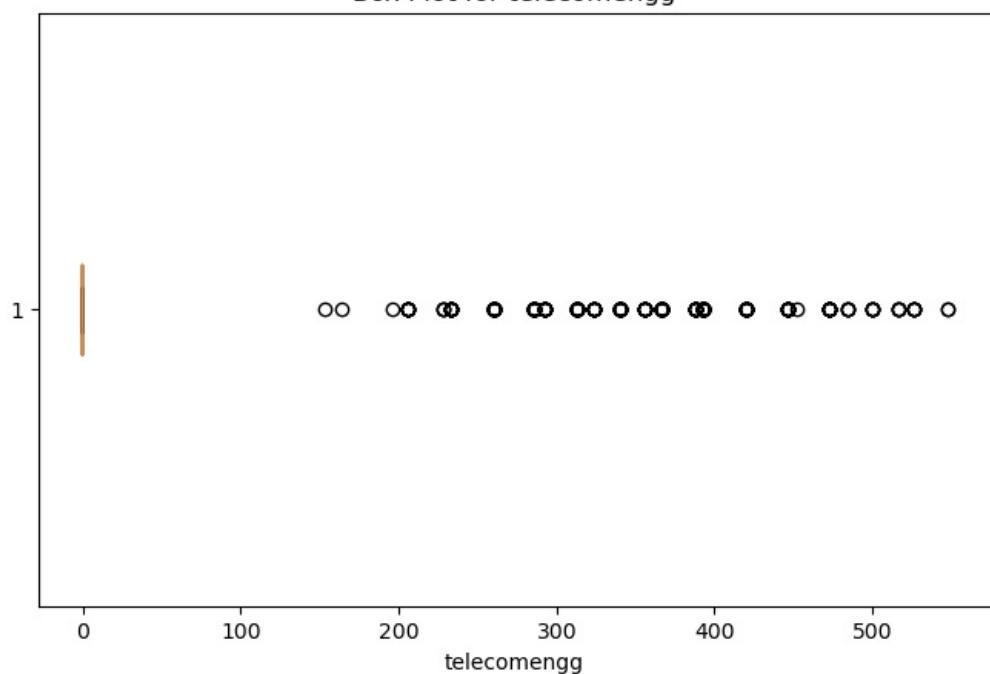
Box Plot for mechanicalengg



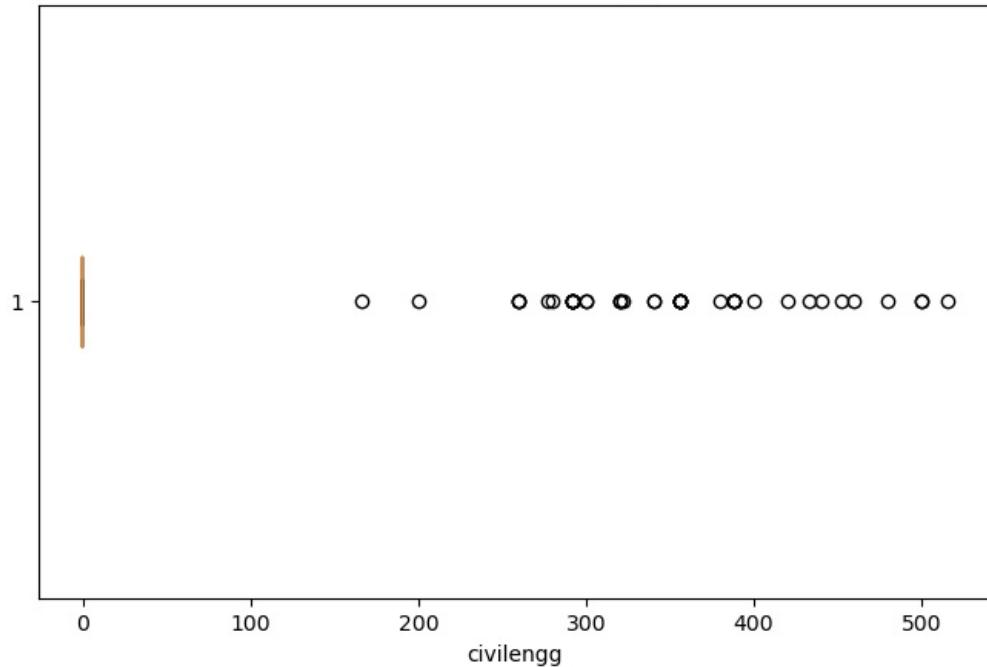
Box Plot for electricalengg



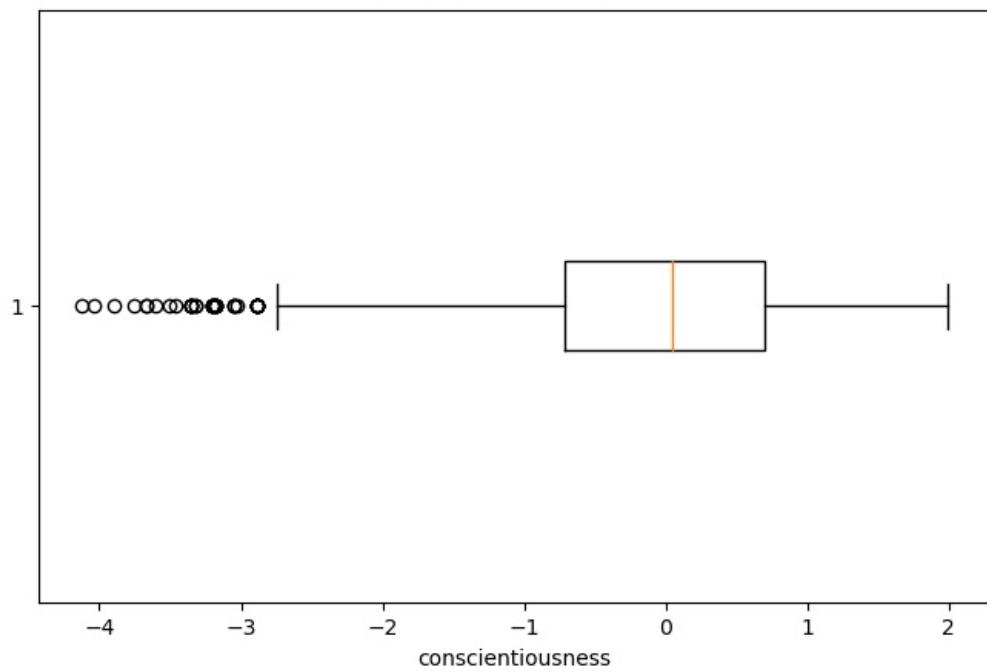
Box Plot for telecomengg



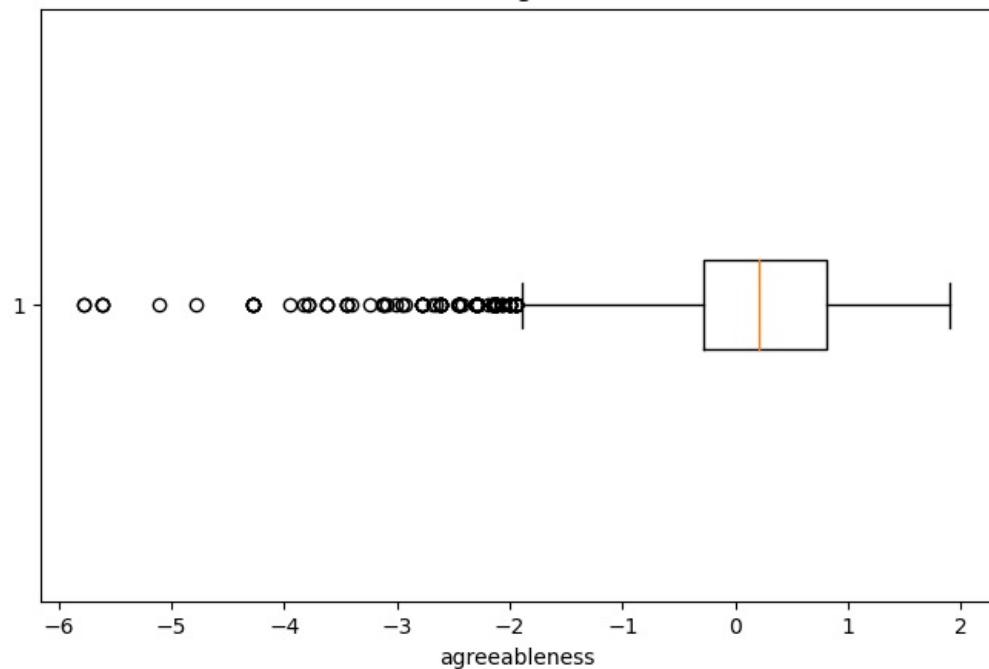
Box Plot for civilengg



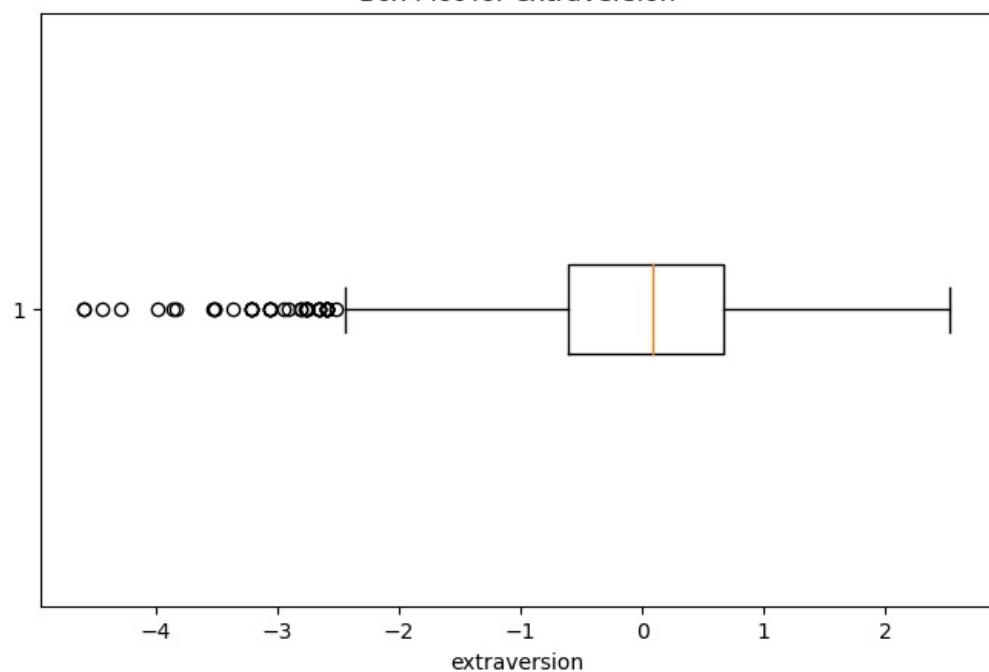
Box Plot for conscientiousness



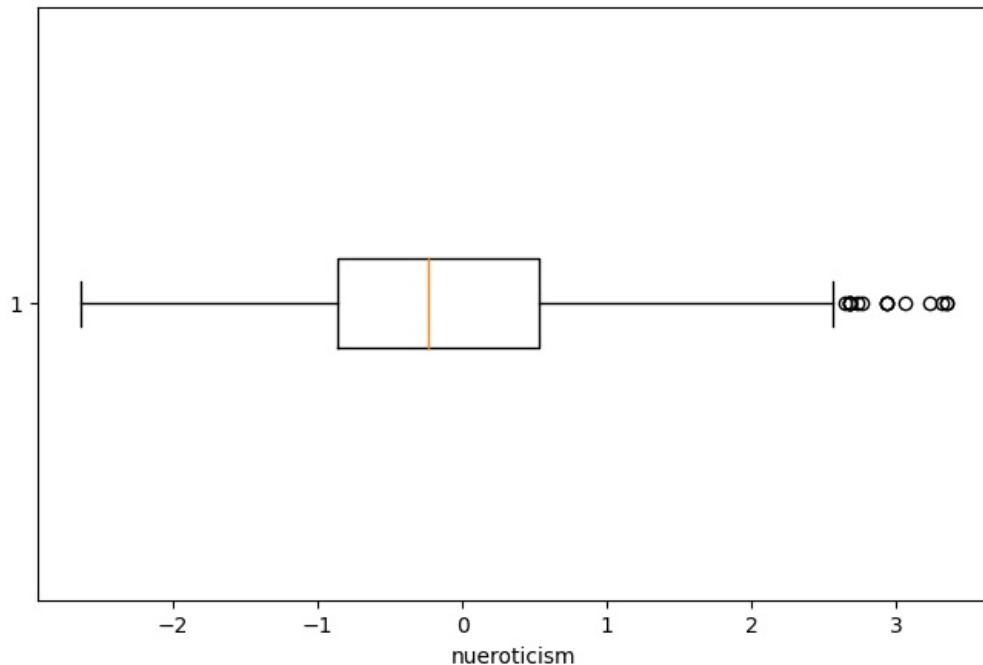
Box Plot for agreeableness



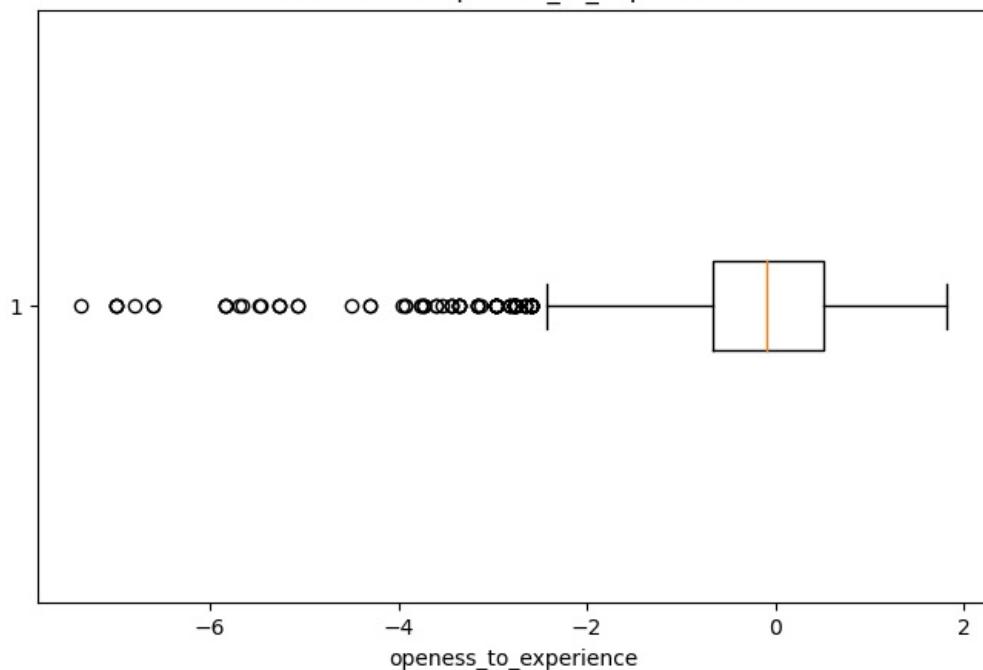
Box Plot for extraversion



Box Plot for nueroticism



Box Plot for openness\_to\_experience



```
In [21]: #The probability and frequency distribution of each numerical column
import warnings
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Assuming df is your DataFrame and num_columns is your list of numerical columns
num_columns =['salary', '10percentage', '12percentage', 'collegegpa','english', 'logical',
'quant', 'computerprogramming', 'computerscience',
'mechanicalengg',
'electricalengg', 'telecomengg', 'civilengg',
'conscientiousness',
'agreeableness', 'extraversion', 'nueroticism',
'openness_to_experience']

warnings.filterwarnings('ignore')

# Loop through each numerical column starting from the second column
for column in num_columns:
    plt.figure(figsize=(12, 6))

    # Use df[column] to access the column data
```

```

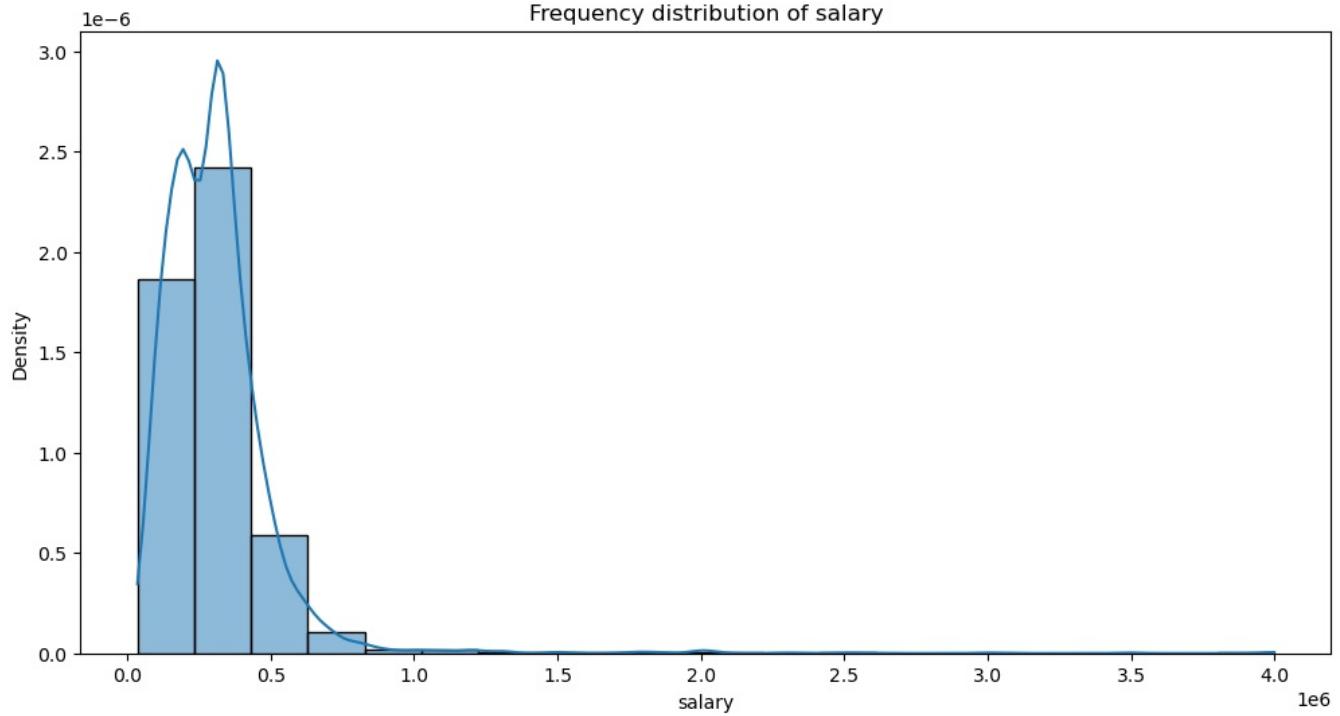
sns.histplot(df[column], bins=20, kde=True, stat='density')

plt.title(f"Frequency distribution of {column}")
plt.xlabel(column)
plt.ylabel('Density')

# Show the plot
plt.show()

# Calculate and print the probability distribution
prob_dist = df[column].value_counts(normalize=True)
print(f"Probability distribution of {column}: \n{prob_dist}\n")

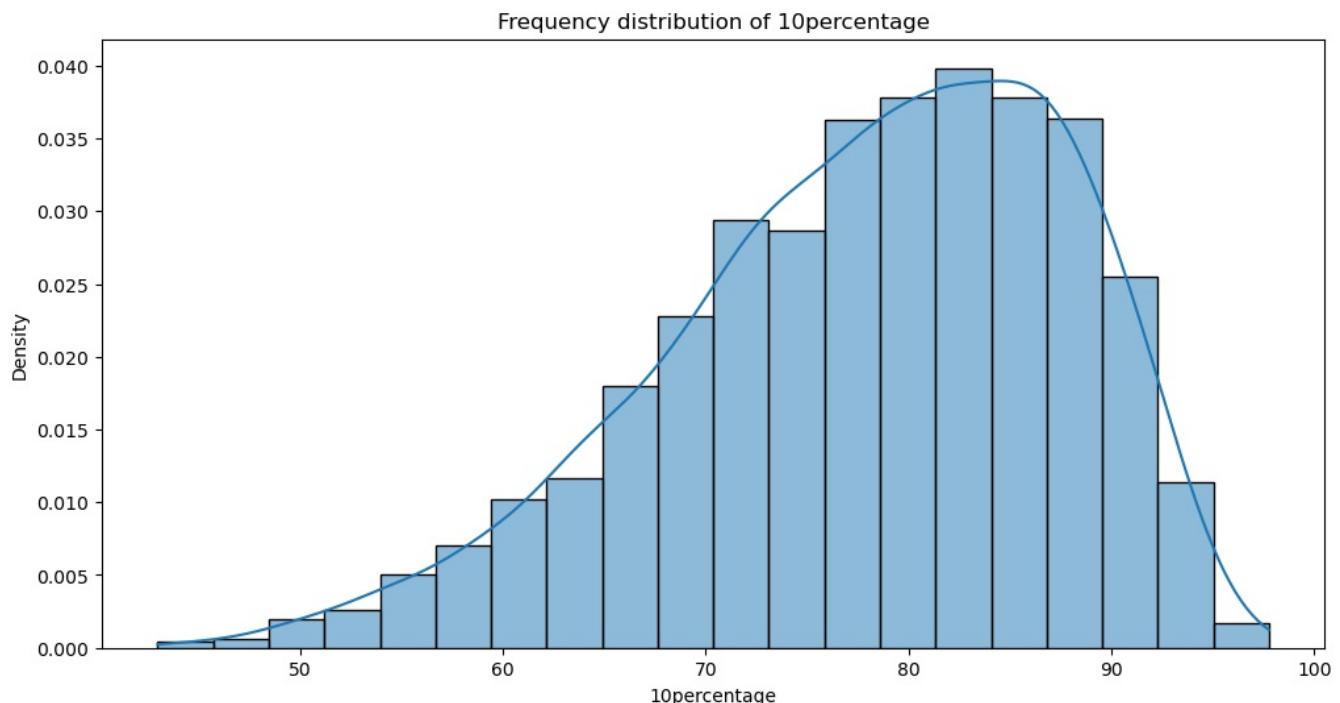
```



```

Probability distribution of salary:
300000    0.073287
180000    0.059780
200000    0.051276
325000    0.047024
120000    0.041271
...
2050000   0.000250
144000    0.000250
1320000   0.000250
755000    0.000250
925000    0.000250
Name: salary, Length: 177, dtype: float64

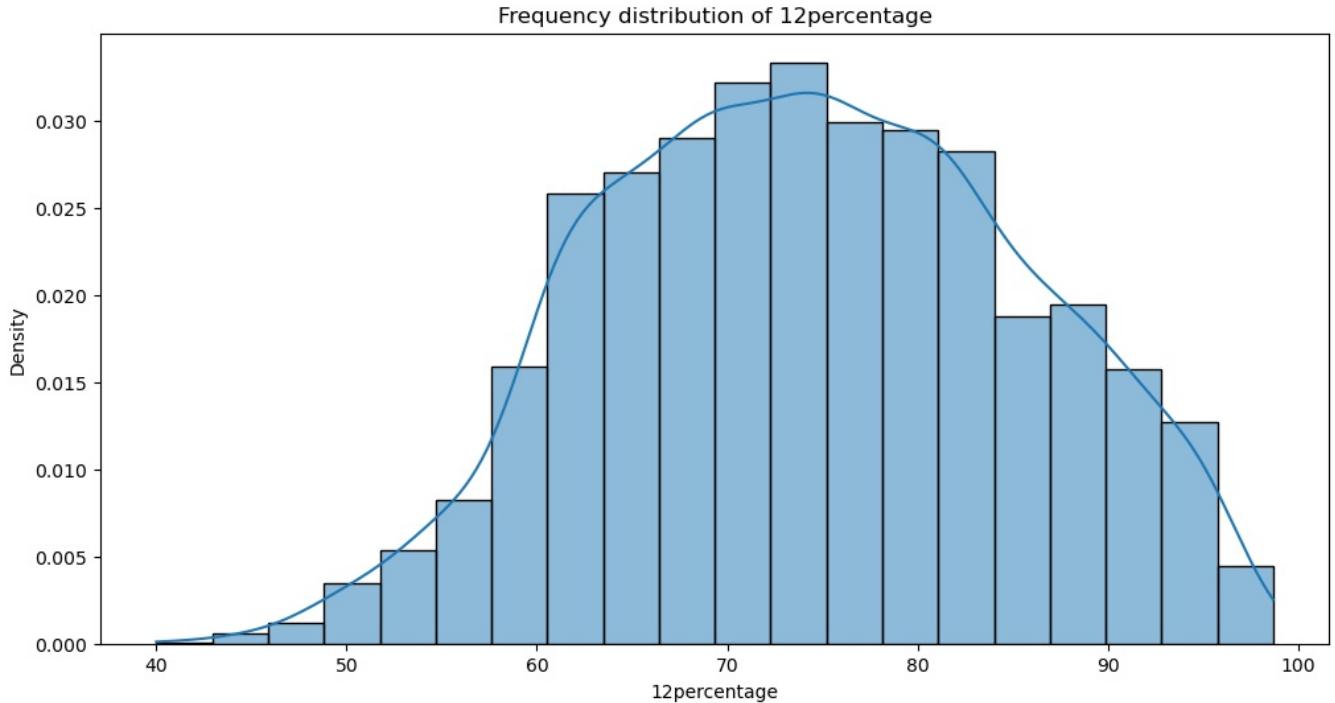
```



Probability distribution of 10percentage:

78.00	0.019010
82.00	0.017759
85.00	0.016758
76.00	0.016508
80.00	0.016258
...	
82.56	0.000250
87.04	0.000250
81.14	0.000250
61.75	0.000250
78.72	0.000250

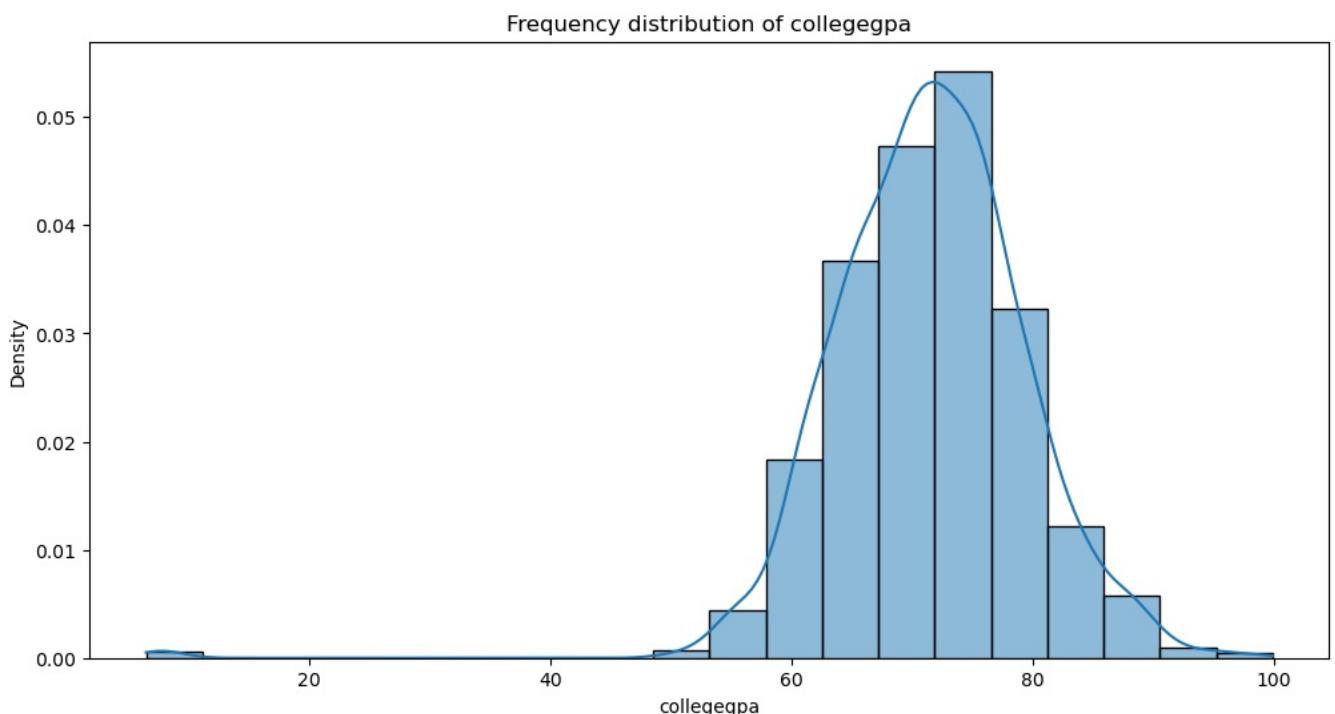
Name: 10percentage, Length: 851, dtype: float64



Probability distribution of 12percentage:

70.00	0.018009
72.00	0.017009
74.00	0.015758
62.00	0.014507
68.00	0.014507
...	
58.50	0.000250
74.45	0.000250
95.41	0.000250
83.58	0.000250
82.55	0.000250

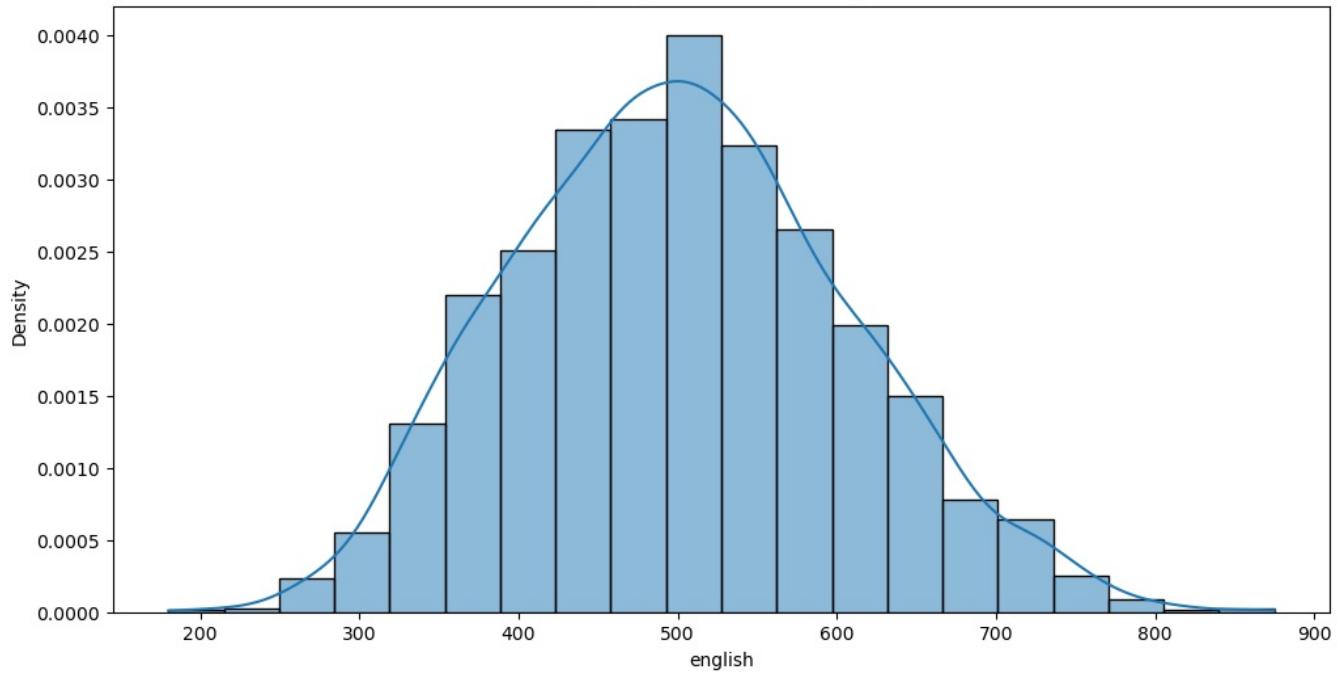
Name: 12percentage, Length: 801, dtype: float64



Probability distribution of collegegpa:

```
70.00    0.028014
72.00    0.024762
75.00    0.020760
65.00    0.019760
71.00    0.018759
...
71.68    0.000250
73.15    0.000250
90.01    0.000250
71.36    0.000250
70.42    0.000250
Name: collegegpa, Length: 1282, dtype: float64
```

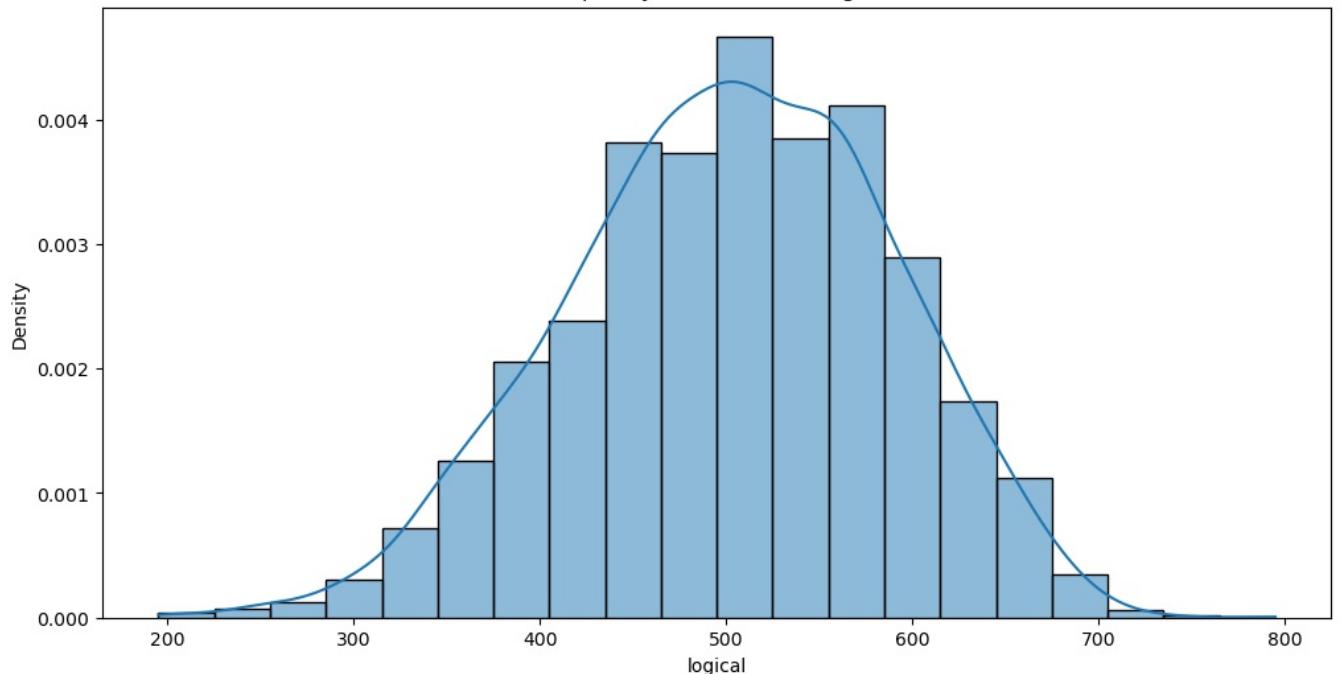
Frequency distribution of english



Probability distribution of english:

```
475     0.040020
545     0.037769
465     0.037519
535     0.034517
405     0.027764
...
180     0.000250
875     0.000250
825     0.000250
870     0.000250
334     0.000250
Name: english, Length: 111, dtype: float64
```

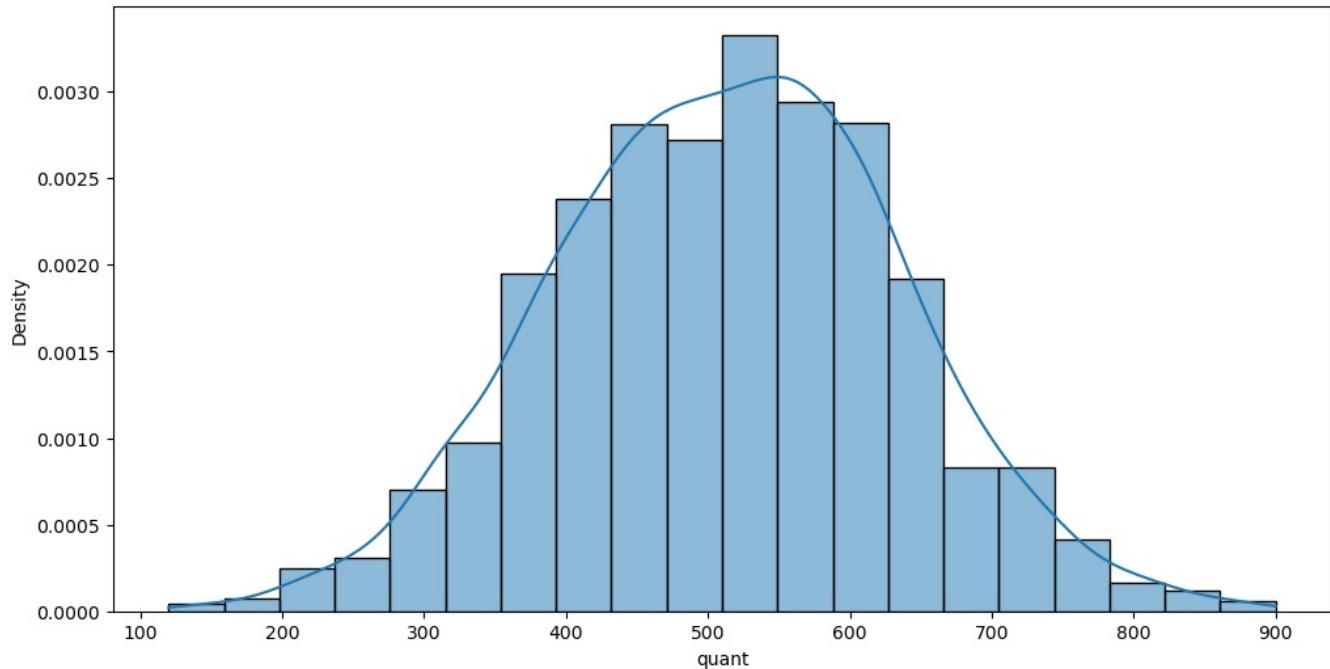
Frequency distribution of logical



Probability distribution of logical:

```
495    0.039520
545    0.037769
555    0.037769
485    0.037769
505    0.029265
...
310    0.000250
795    0.000250
534    0.000250
454    0.000250
660    0.000250
Name: logical, Length: 107, dtype: float64
```

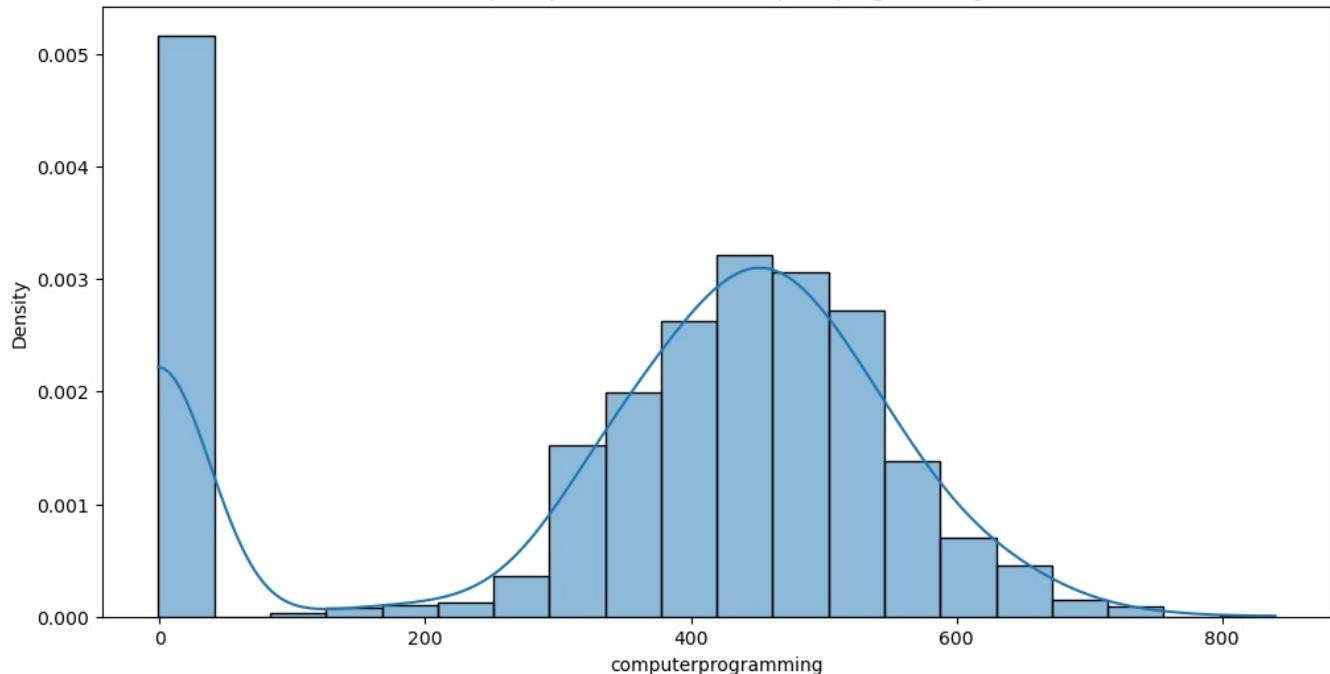
Frequency distribution of quant



Probability distribution of quant:

```
605    0.035768
485    0.032516
545    0.031266
575    0.029015
515    0.024762
...
805    0.000250
175    0.000250
214    0.000250
860    0.000250
394    0.000250
Name: quant, Length: 138, dtype: float64
```

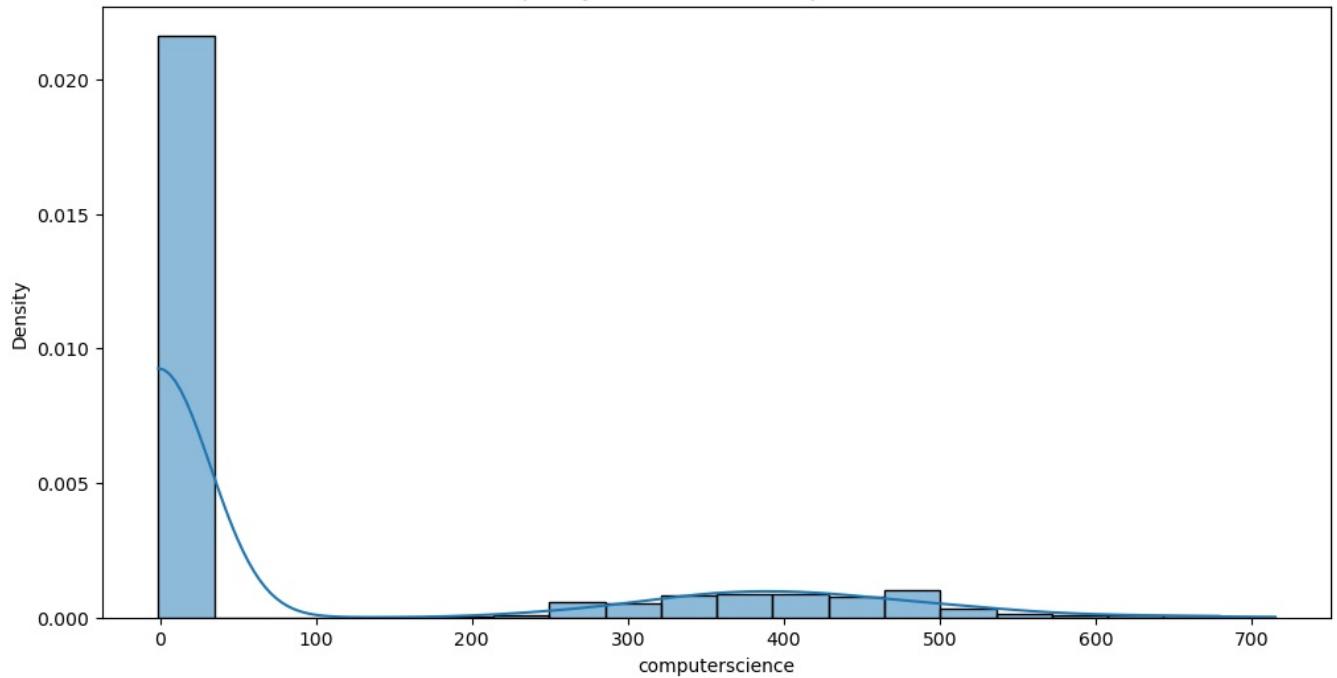
Frequency distribution of computerprogramming



Probability distribution of computerprogramming:

```
-1      0.217109
445     0.037769
435     0.036018
475     0.034767
465     0.033517
...
214     0.000250
494     0.000250
840     0.000250
394     0.000250
554     0.000250
Name: computerprogramming, Length: 79, dtype: float64
```

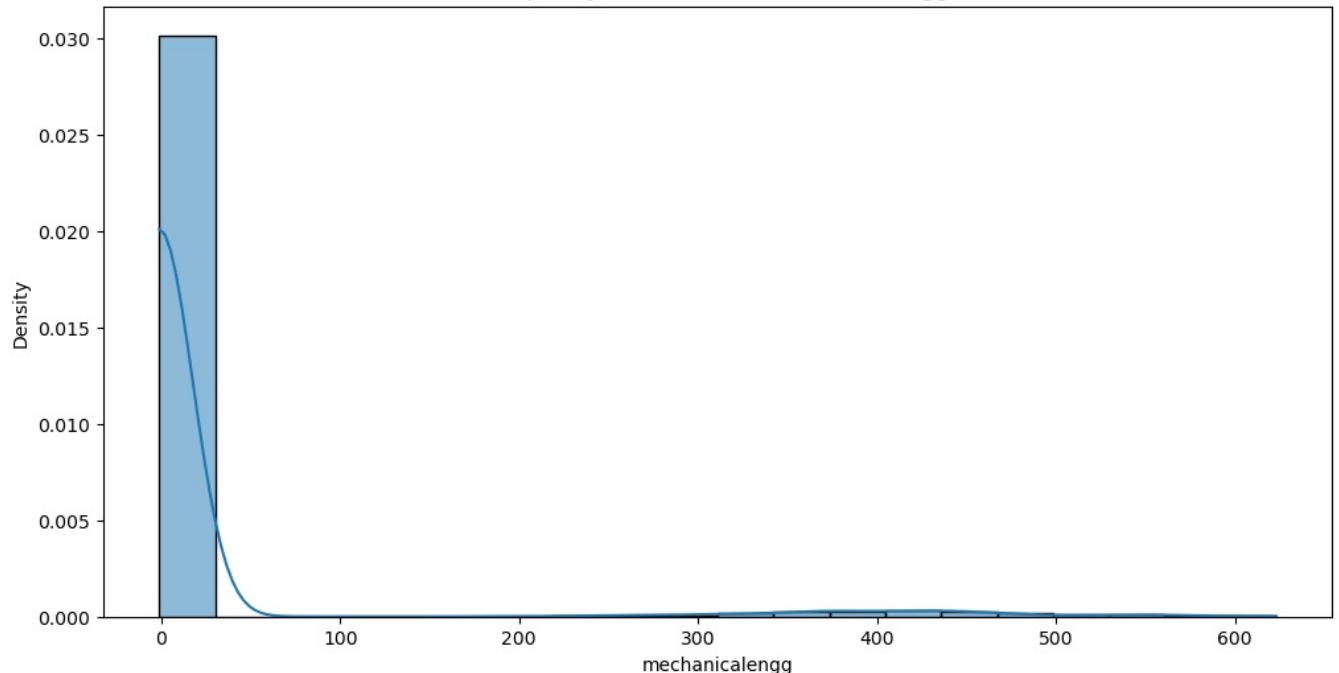
Frequency distribution of computergscience



Probability distribution of computergscience:

```
-1      0.774387
407     0.032016
376     0.030765
346     0.029515
438     0.027764
469     0.020010
315     0.019260
500     0.016008
284     0.012506
530     0.011256
253     0.007504
561     0.005503
223     0.003502
592     0.003502
623     0.002501
653     0.002251
192     0.000750
715     0.000500
684     0.000250
130     0.000250
Name: computergscience, dtype: float64
```

Frequency distribution of mechanicalengg



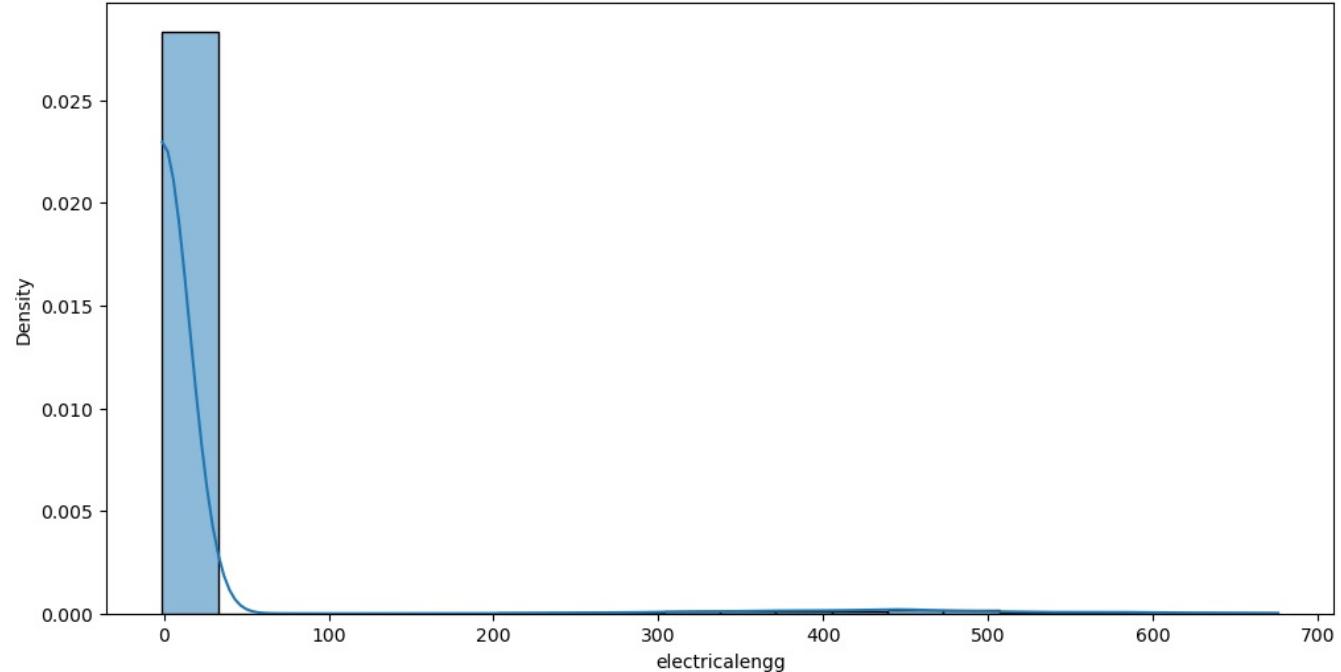
Probability distribution of mechanicalengg:

```
-1      0.941221
366    0.005003
446    0.004002
438    0.003752
420    0.003502
376    0.003252
313    0.003252
393    0.003252
407    0.003002
346    0.002751
469    0.002501
473    0.002501
553    0.002001
435    0.001751
383    0.001501
340    0.001501
526    0.001251
409    0.001251
286    0.001251
500    0.001001
253    0.001001
284    0.000750
332    0.000750
538    0.000750
254    0.000750
580    0.000750
616    0.000500
564    0.000500
606    0.000500
223    0.000500
512    0.000500
561    0.000500
260    0.000500
358    0.000250
280    0.000250
315    0.000250
```

```
233    0.000250
306    0.000250
461    0.000250
180    0.000250
206    0.000250
623    0.000250
```

Name: mechanicalengg, dtype: float64

Frequency distribution of electricalengg



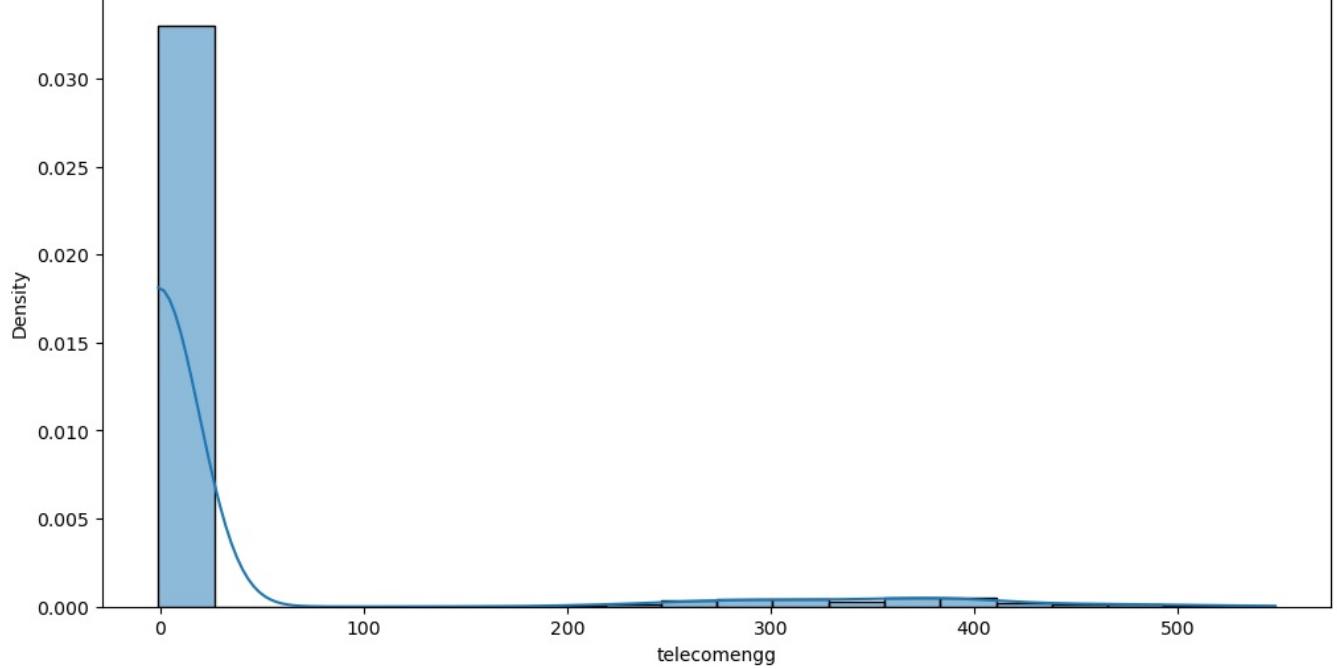
Probability distribution of electricalengg:

```
-1      0.959730
420    0.004002
446    0.003502
388    0.002501
473    0.002501
452    0.002501
356    0.002251
500    0.002001
580    0.002001
366    0.001751
324    0.001751
393    0.001751

553    0.001501
313    0.001501
516    0.001251
260    0.001001
292    0.001001
340    0.000750
228    0.000750
526    0.000750
484    0.000750
633    0.000750
548    0.000500
433    0.000500
606    0.000500
612    0.000500
660    0.000500
286    0.000500
676    0.000250
411    0.000250
206    0.000250
```

Name: electricalengg, dtype: float64

Frequency distribution of telecomengg



Probability distribution of telecomengg:

-1 0.906453

393 0.011256

366 0.010755

260 0.008754

313 0.008504

340 0.008004

286 0.007754

420 0.006503

446 0.004002

388 0.003502

233 0.003502

473 0.003252

292 0.003252

356 0.003002

324 0.002751

206 0.002001

500 0.001251

526 0.001251

516 0.001001

484 0.001001

228 0.000750

548 0.000500

153 0.000250

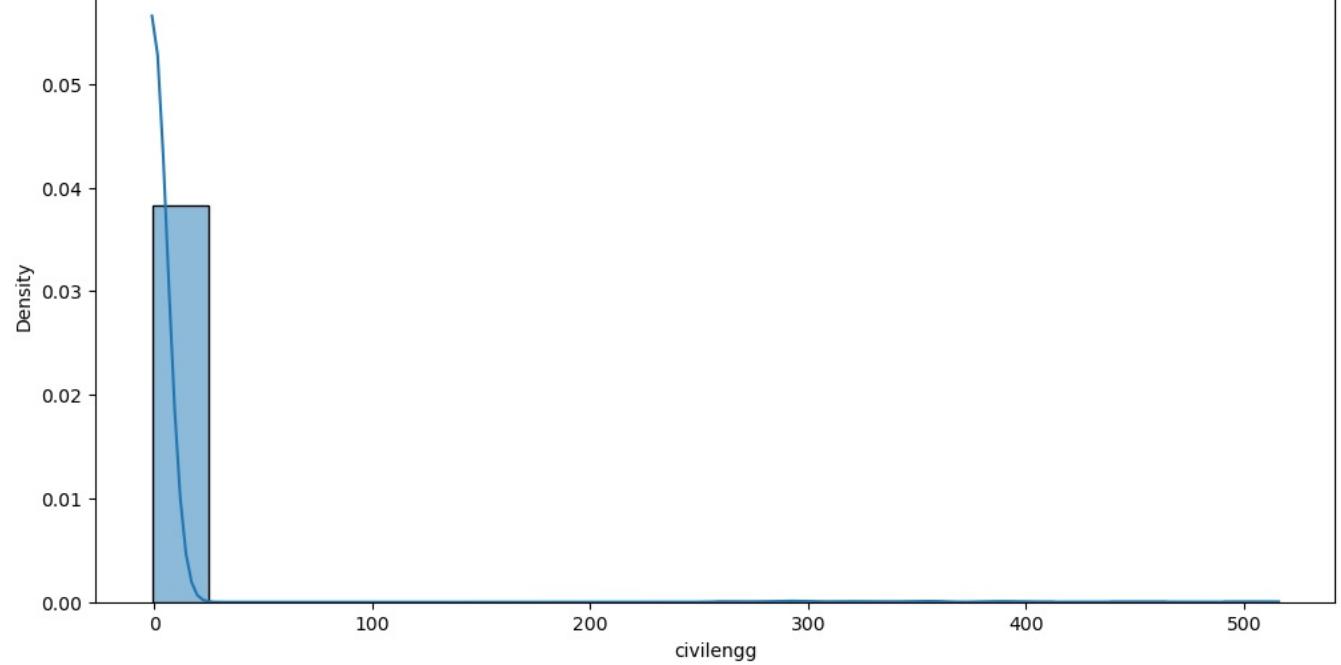
196 0.000250

164 0.000250

452 0.000250

Name: telecomengg, dtype: float64

Frequency distribution of civilengg



Probability distribution of civilengg:

-1 0.989495

356 0.001501

292 0.001501

388 0.001001

260 0.000750

320 0.000750

500 0.000500

300 0.000500

340 0.000500

516 0.000250

460 0.000250

420 0.000250

280 0.000250

433 0.000250

380 0.000250

452 0.000250

277 0.000250

166 0.000250

322 0.000250

200 0.000250

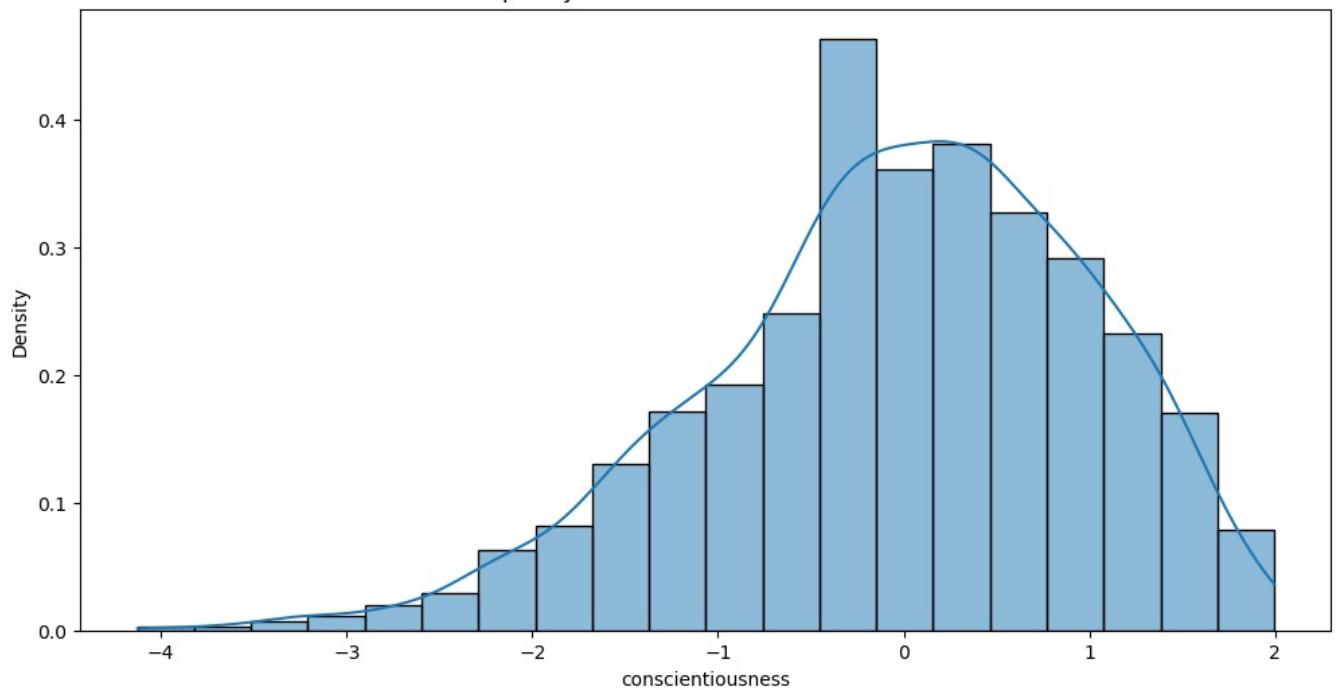
440 0.000250

400 0.000250

480 0.000250

Name: civilengg, dtype: float64

Frequency distribution of conscientiousness

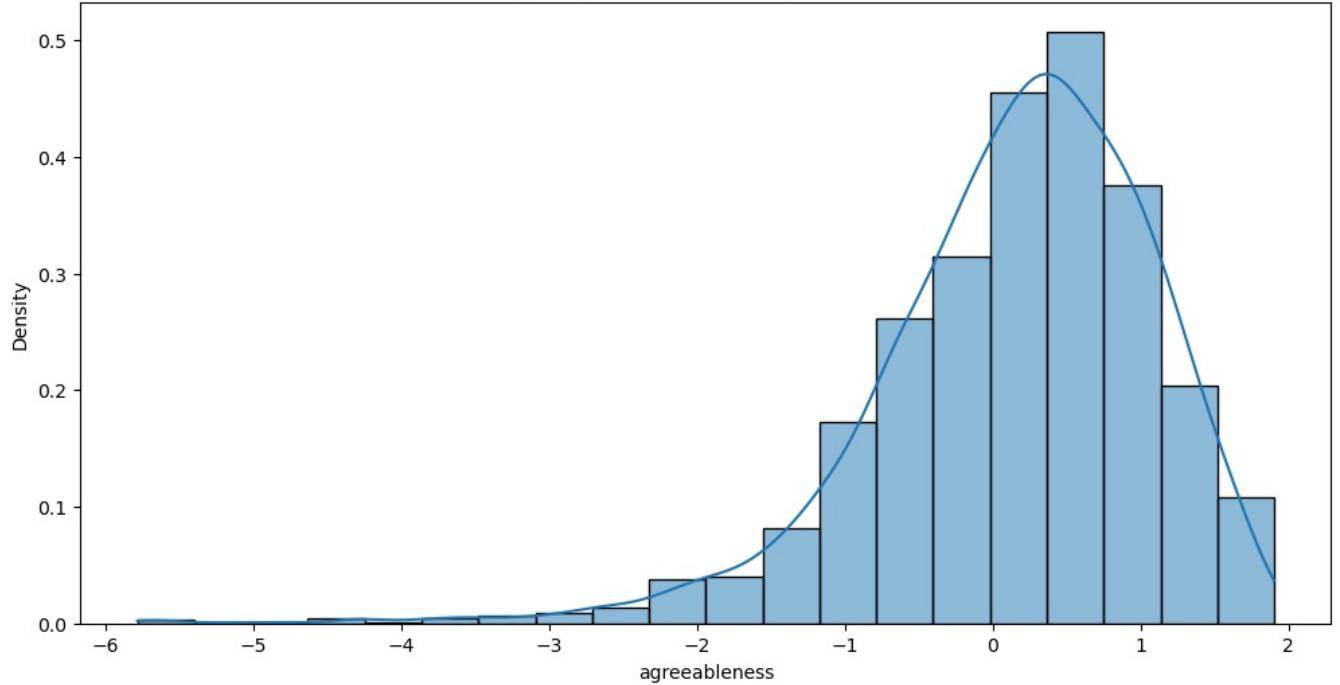


Probability distribution of conscientiousness:

```
0.2718  0.036268
0.1282  0.033517
-0.1590  0.033267
0.4155  0.032766
-0.0154  0.032266
...
-3.4624  0.000250
-1.2950  0.000250
-0.9653  0.000250
-0.4854  0.000250
0.8986  0.000250
```

Name: conscientiousness, Length: 141, dtype: float64

Frequency distribution of agreeableness

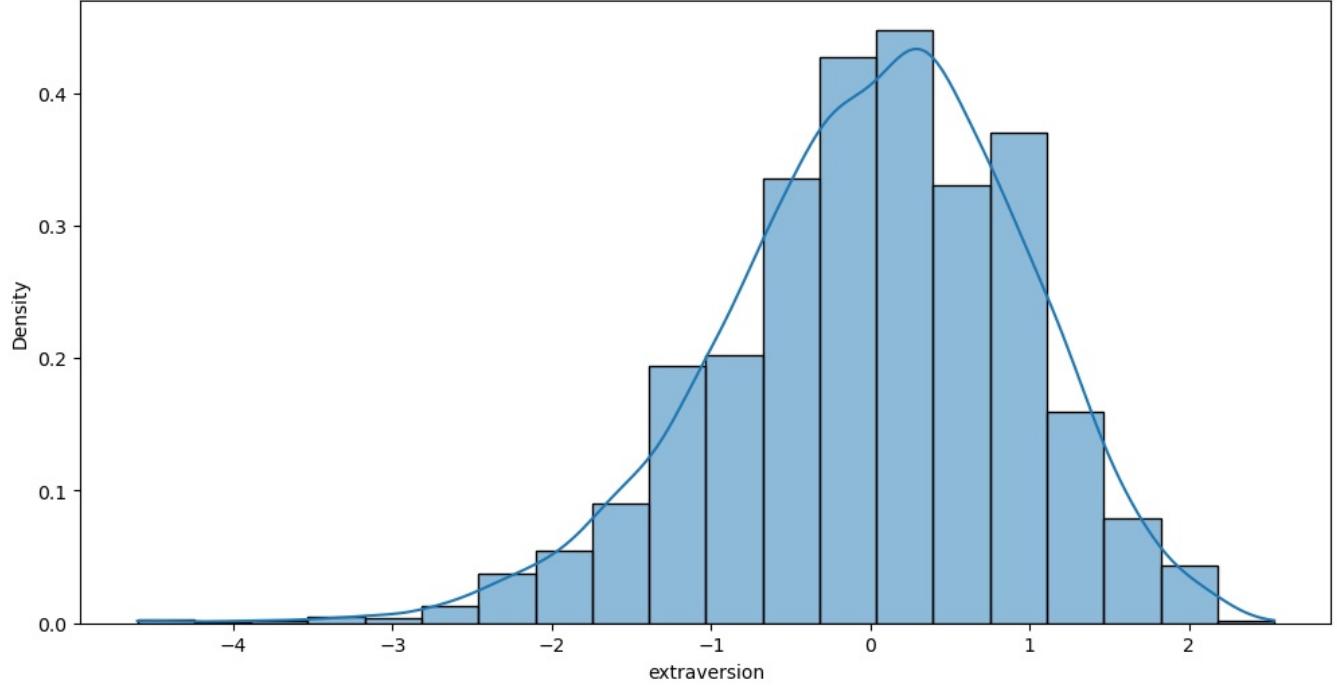


Probability distribution of agreeableness:

```
0.3789  0.048274
0.2124  0.045023
0.5454  0.043772
0.0459  0.041521
0.8784  0.039520
...
-3.1264 0.000250
-3.0094 0.000250
-3.9501 0.000250
-1.7223 0.000250
-0.8320 0.000250
```

Name: agreeableness, Length: 149, dtype: float64

Frequency distribution of extraversion

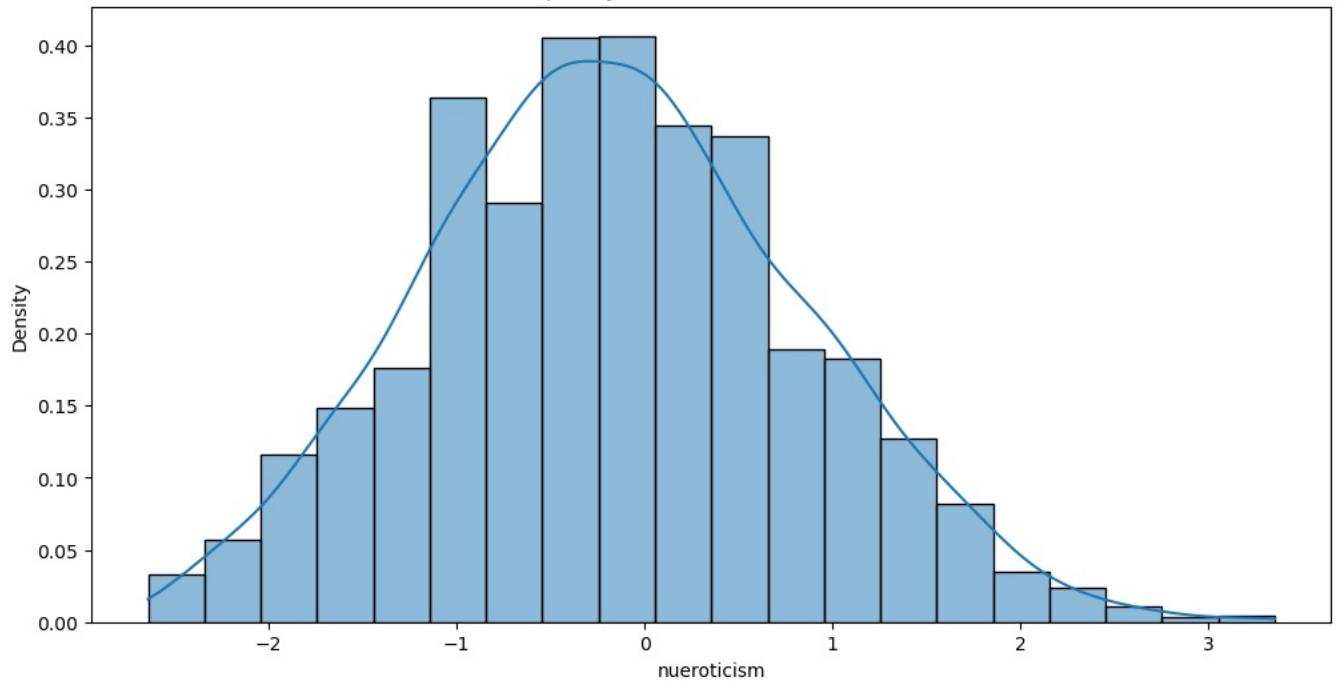


Probability distribution of extraversion:

```
0.4711  0.044772
0.3174  0.044522
0.1637  0.038769
0.7785  0.036518
-0.1437 0.033767
...
-3.5370 0.000250
-0.4226 0.000250
1.5791  0.000250
-0.1408 0.000250
-1.2056 0.000250
```

Name: extraversion, Length: 154, dtype: float64

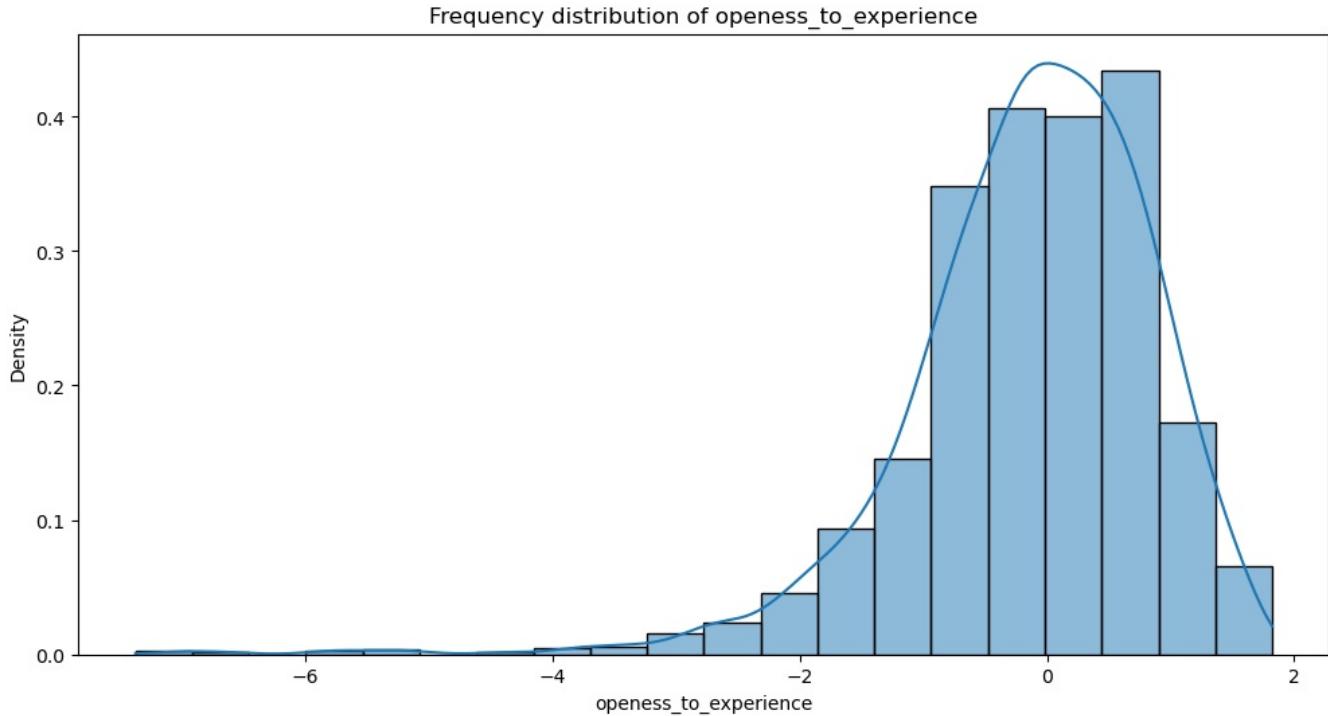
Frequency distribution of nueroticism



```

Probability distribution of nueroticism:
-0.48790    0.031516
-0.74150    0.029515
0.01920    0.028014
-0.61470    0.027264
-0.36120    0.026513
...
1.06113    0.000250
-0.74960    0.000250
2.76500    0.000250
1.82493    0.000250
2.03060    0.000250
Name: nueroticism, Length: 217, dtype: float64

```



```

Probability distribution of openness_to_experience:
0.6721    0.046773
-0.0943    0.045523
0.0973    0.045523
0.4805    0.044272
0.2889    0.043522
...
-6.8009    0.000250
0.1187    0.000250
-5.6860    0.000250
-1.1291    0.000250
-0.4229    0.000250
Name: openness_to_experience, Length: 142, dtype: float64

```

```
In [22]: #Frequency distribution of each categorical Variable/Column
frequency_dis={i: df[i].value_counts() for i in cat_columns}
for i,freq in frequency_dis.items():
    print(f"Frequency distribution for {i}:\n{freq}\n")
```

```

Frequency distribution for column1:
train    3998
Name: column1, dtype: int64

Frequency distribution for doj:
07-01-2014 00:00    199
06-01-2014 00:00    180
08-01-2014 00:00    178
09-01-2014 00:00    142
01-01-2014 00:00    142
...
11-01-2015 00:00    1
11-01-2009 00:00    1
08-01-2004 00:00    1
09-01-2009 00:00    1
02-01-2007 00:00    1
Name: doj, Length: 81, dtype: int64

```

```

Frequency distribution for dol:
present      1875
4/1/15 0:00    573
3/1/15 0:00    124
5/1/15 0:00    112
1/1/15 0:00    99

```

3/1/05 0:00 1  
10/1/15 0:00 1  
2/1/10 0:00 1  
2/1/11 0:00 1  
10/1/10 0:00 1  
Name: dol, Length: 67, dtype: int64

Frequency distribution for designation:  
software engineer 539  
software developer 265  
system engineer 205  
programmer analyst 139  
systems engineer 118  
...  
cad drafter 1  
noc engineer 1  
human resources intern 1  
senior quality assurance engineer 1  
jr. software developer 1  
Name: designation, Length: 419, dtype: int64

Frequency distribution for jobcity:  
Bangalore 627  
-1 461  
Noida 368  
Hyderabad 335  
Pune 290  
...  
Tirunelvelli 1  
Ernakulam 1  
Nanded 1  
Dharmapuri 1  
Asifabadbanglore 1  
Name: jobcity, Length: 339, dtype: int64

Frequency distribution for gender:  
m 3041  
f 957  
Name: gender, dtype: int64

Frequency distribution for dob:  
1/1/91 0:00 11  
7/15/91 0:00 10  
7/5/91 0:00 8  
12/13/91 0:00 8  
6/3/91 0:00 8  
...  
12/30/92 0:00 1  
10/20/86 0:00 1  
11/17/89 0:00 1  
9/30/92 0:00 1  
4/15/87 0:00 1  
Name: dob, Length: 1872, dtype: int64

Frequency distribution for 10board:  
cbse 1395  
state board 1164  
0 350  
icse 281  
ssc 122  
...  
hse,orissa 1  
national public school 1  
nagpur board 1  
jharkhand academic council 1  
bse,odisha 1  
Name: 10board, Length: 275, dtype: int64

Frequency distribution for 12board:  
cbse 1400  
state board 1254  
0 359  
icse 129  
up board 87  
...  
jawahar higher secondary school 1  
nagpur board 1  
bsemp 1  
board of higher secondary orissa 1  
boardofintermediate 1  
Name: 12board, Length: 340, dtype: int64

Frequency distribution for degree:  
B.Tech/B.E. 3700  
MCA 243  
M.Tech./M.E. 53  
M.Sc. (Tech.) 2  
Name: degree, dtype: int64

```
Frequency distribution for specialization:  
electronics and communication engineering    880  
computer science & engineering             744  
information technology                      660  
computer engineering                       600  
computer application                      244  
mechanical engineering                     201  
electronics and electrical engineering      196  
electronics & telecommunications           121  
electrical engineering                     82  
electronics & instrumentation eng         32  
civil engineering                          29  
electronics and instrumentation engineering 27  
information science engineering            27  
instrumentation and control engineering   20  
electronics engineering                   19  
biotechnology                            15  
other                                    13  
industrial & production engineering       10  
applied electronics and instrumentation    9  
chemical engineering                      9  
computer science and technology           6  
telecommunication engineering            6  
mechanical and automation                5  
automobile/automotive engineering          5  
instrumentation engineering              4  
mechatronics                             4  
aeronautical engineering                 3  
electronics and computer engineering      3  
electrical and power engineering          2  
biomedical engineering                   2  
information & communication technology   2  
industrial engineering                  2  
computer science                         2  
metallurgical engineering                2  
power systems and automation             1  
control and instrumentation engineering   1  
mechanical & production engineering     1  
embedded systems technology              1  
polymer technology                      1  
computer and communication engineering   1  
information science                      1  
internal combustion engine               1  
computer networking                     1  
ceramic engineering                     1  
electronics                            1  
industrial & management engineering    1  
Name: specialization, dtype: int64
```

```
Frequency distribution for collegestate:
```

```
Uttar Pradesh        915  
Karnataka          370  
Tamil Nadu          367  
Telangana           319  
Maharashtra         262  
Andhra Pradesh      225  
West Bengal          196  
Punjab              193  
Madhya Pradesh       189  
Haryana              180  
Rajasthan             174  
Orissa                172  
Delhi                  162  
Uttarakhand           113  
Kerala                  33  
Jharkhand                28  
Chhattisgarh            27  
Gujarat                  24  
Himachal Pradesh        16  
Bihar                    10  
Jammu and Kashmir       7  
Assam                      5  
Union Territory           5  
Sikkim                      3  
Meghalaya                  2  
Goa                        1  
Name: collegestate, dtype: int64
```

```
In [23]: df['degree'].unique()
```

```
Out[23]: array(['B.Tech/B.E.', 'MCA', 'M.Tech./M.E.', 'M.Sc. (Tech.)'],  
              dtype=object)
```

```
In [24]: df['degree'].nunique()
```

```
Out[24]: 4
```

```
In [25]: df['jobcity'].value_counts()
```

```
Out[25]:
```

jobcity	count
Bangalore	627
-1	461
Noida	368
Hyderabad	335
Pune	290
...	
Tirunelvelli	1
Ernakulam	1
Nanded	1
Dharmapuri	1
Asifabadbanglore	1

Name: jobcity, Length: 339, dtype: int64

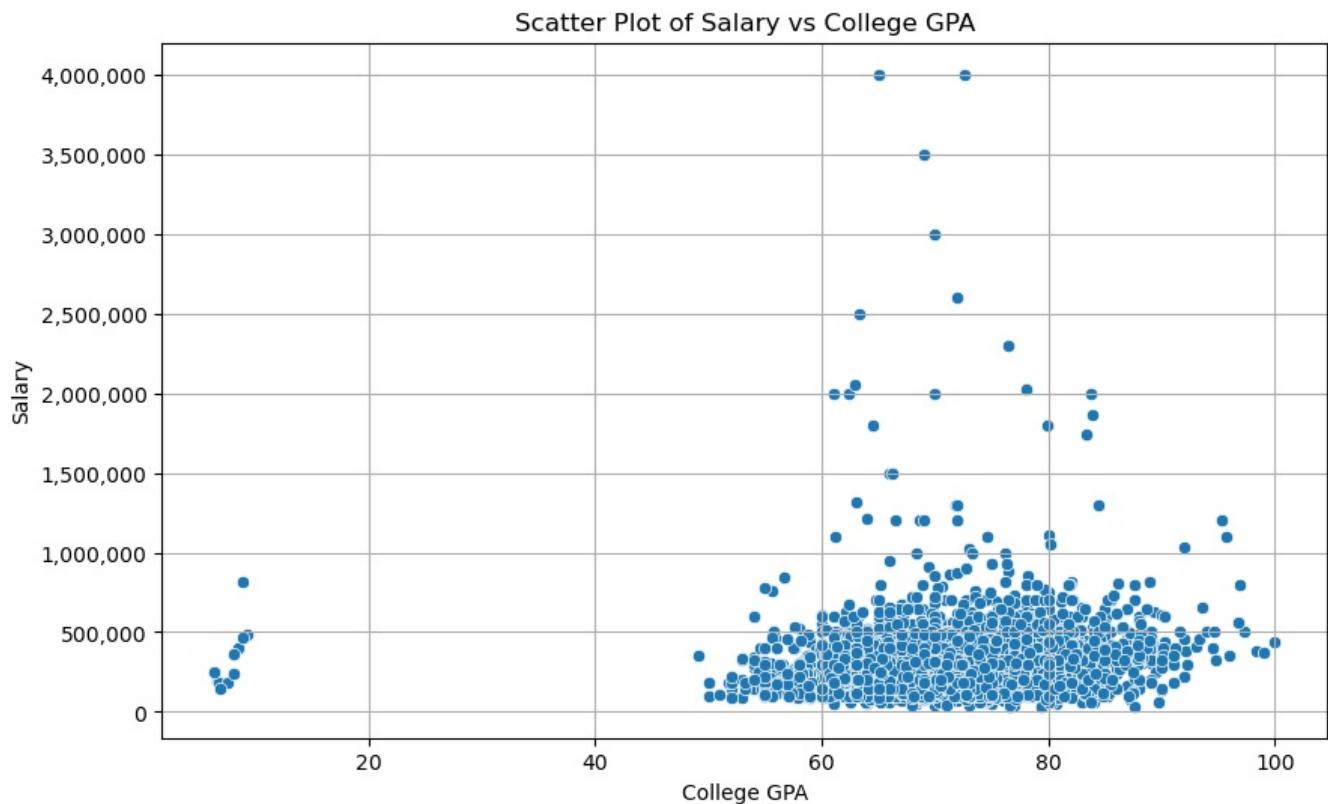
## 2.Bivariate Analysis

```
In [26]: import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib.ticker import FuncFormatter
```

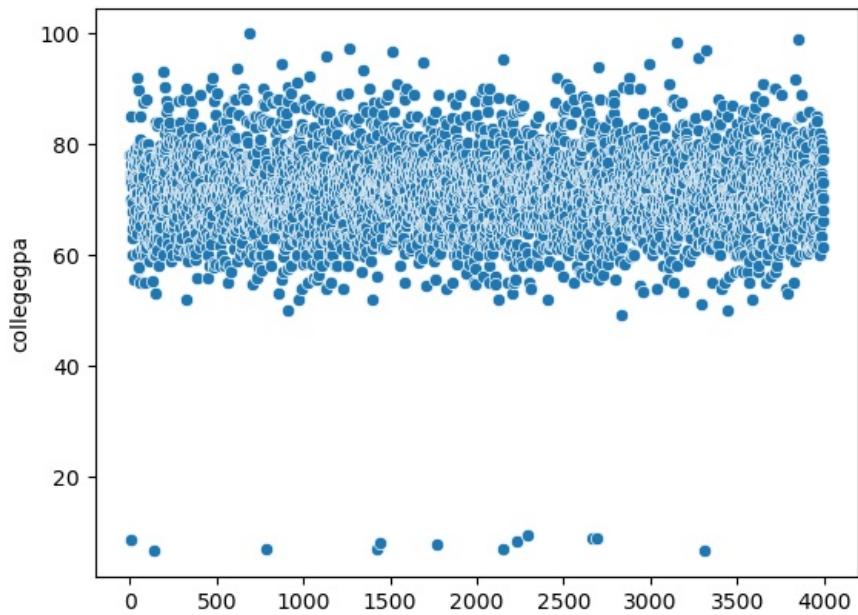
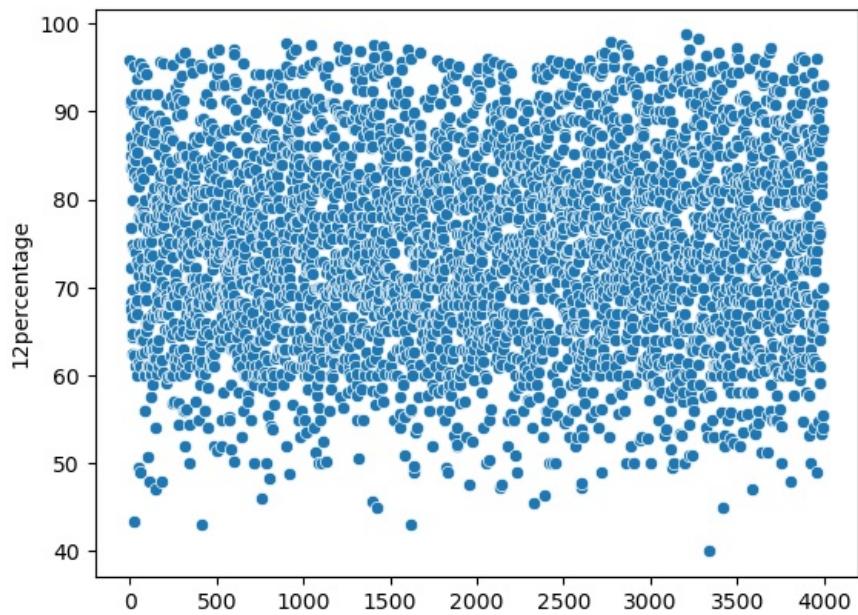
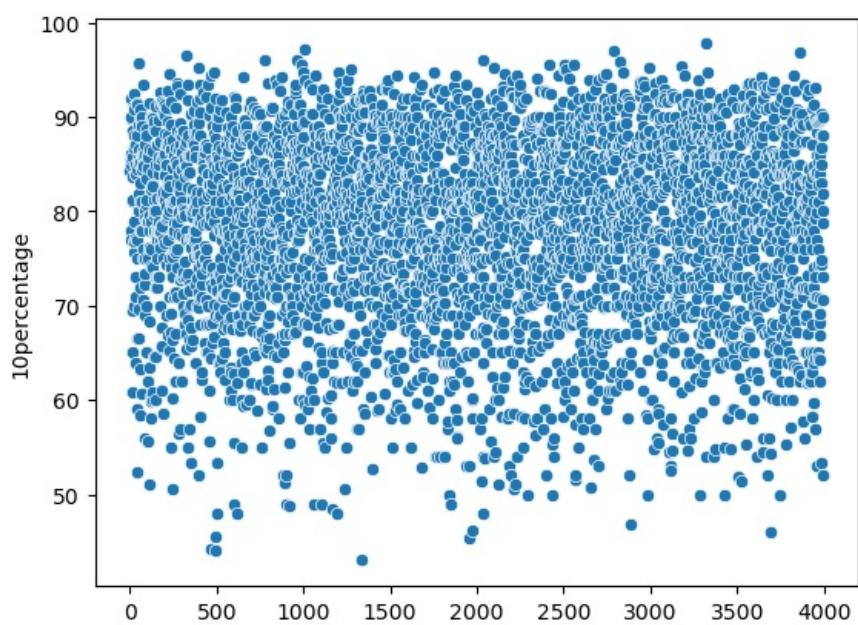
```
def currency(x, _):
    return f'{int(x):,}'

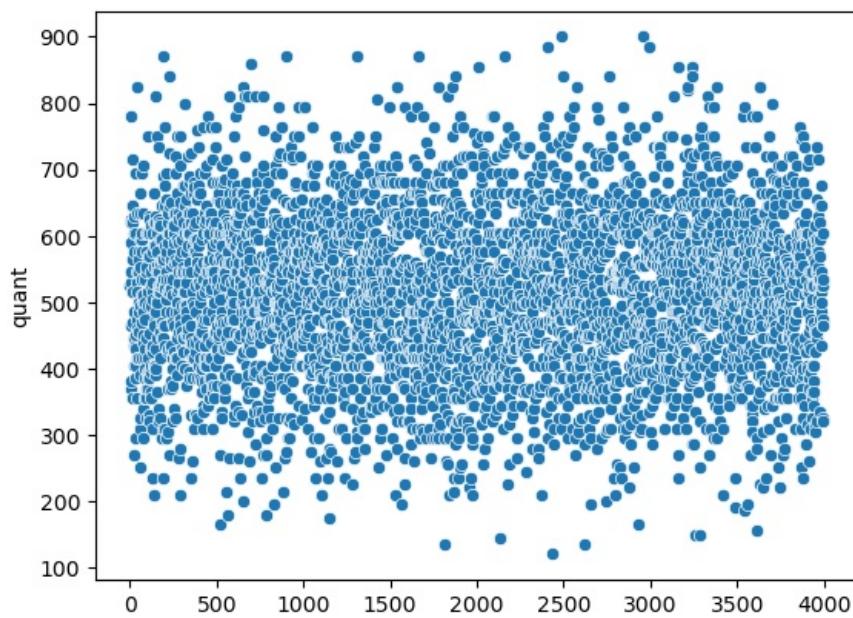
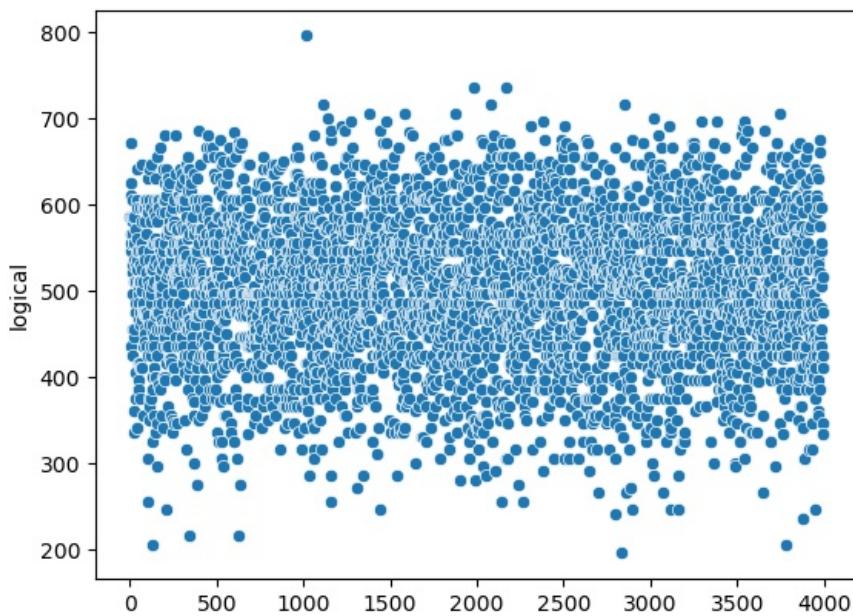
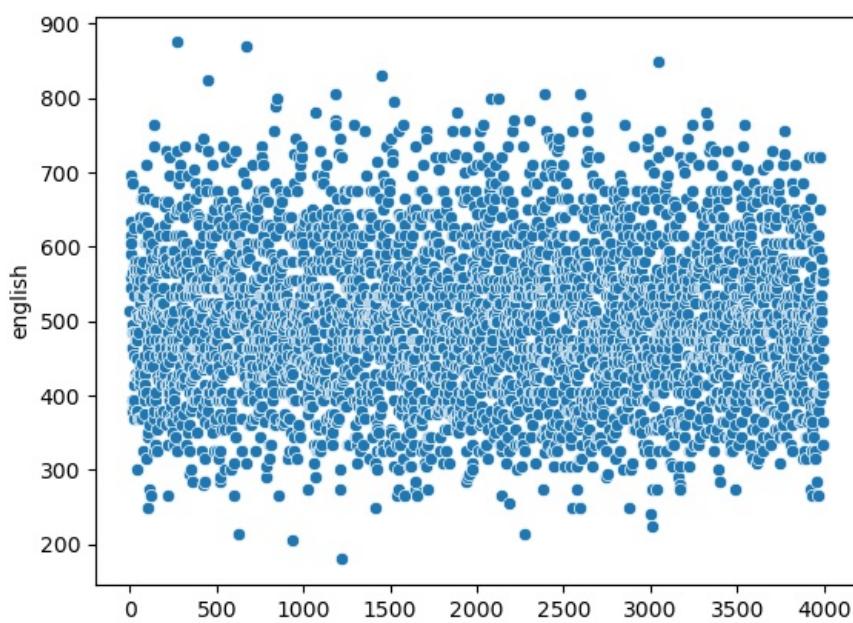
plt.figure(figsize=(10, 6))

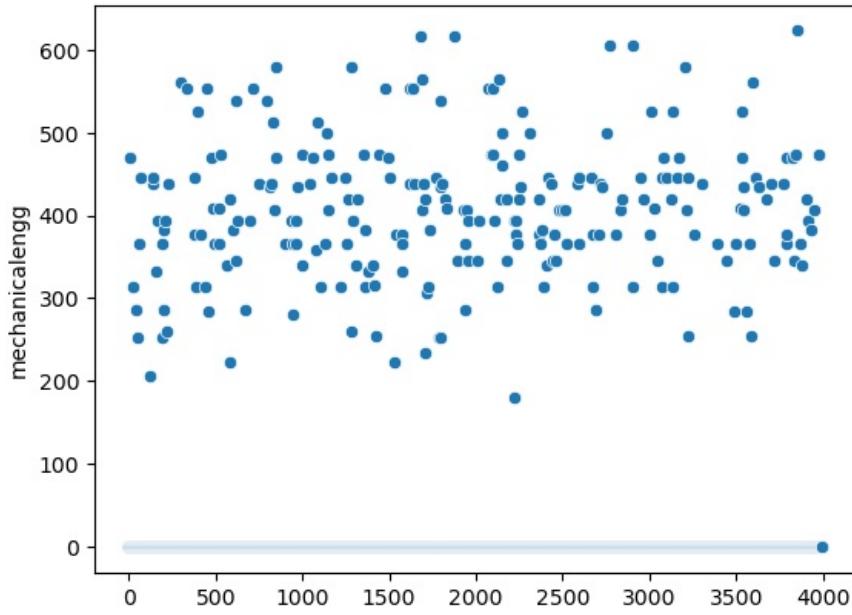
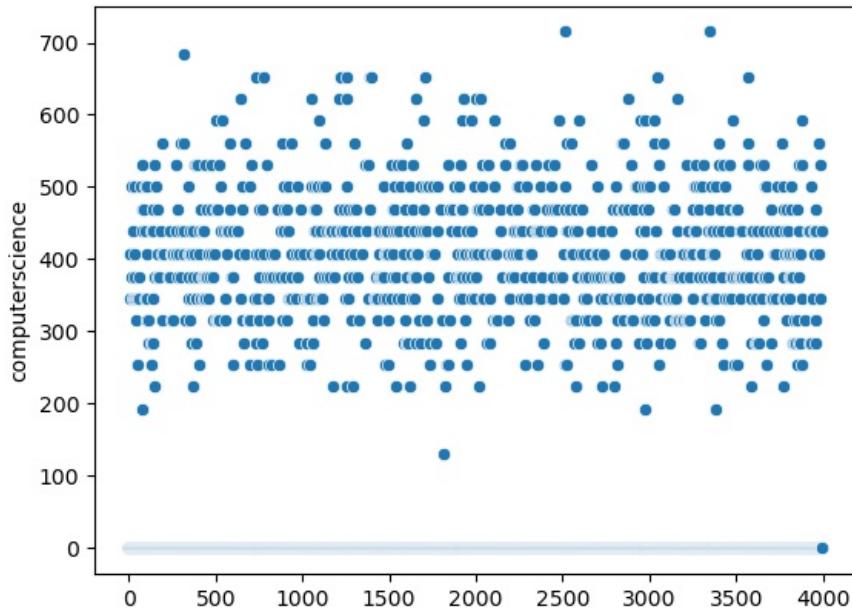
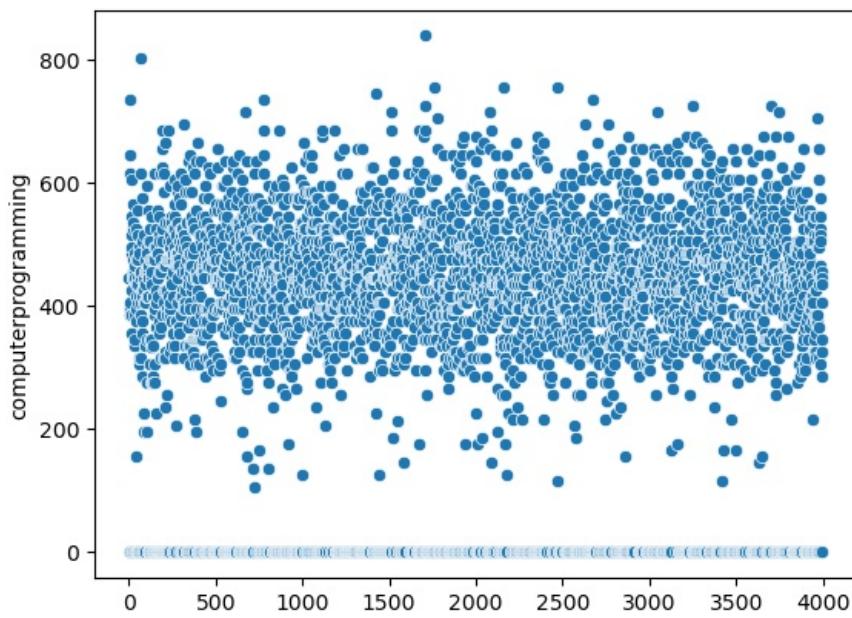
# Scatter plot of Salary vs College GPA
sns.scatterplot(data=df, x='collegegpa', y='salary')
plt.title('Scatter Plot of Salary vs College GPA')
plt.xlabel('College GPA')
plt.ylabel('Salary')
plt.grid(True)
plt.gca().yaxis.set_major_formatter(FuncFormatter(currency))
plt.show()
```

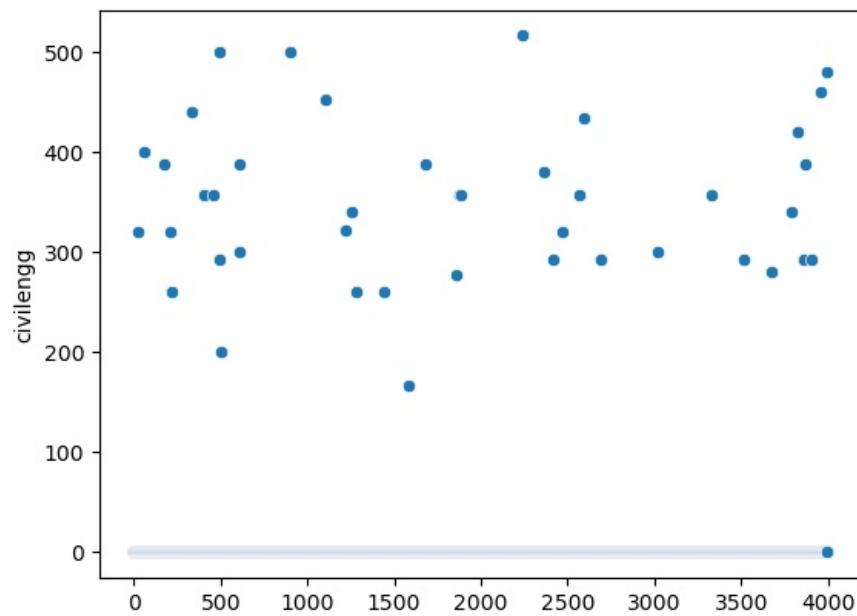
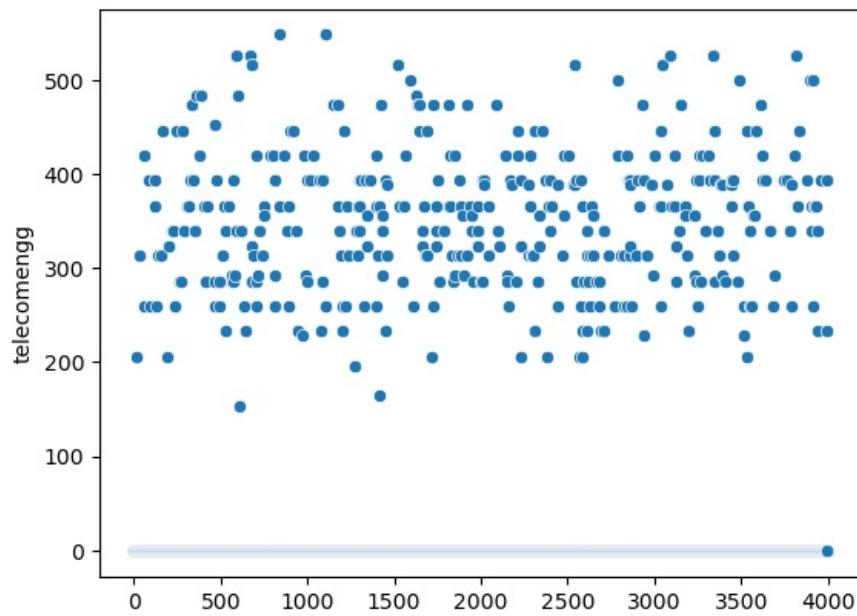
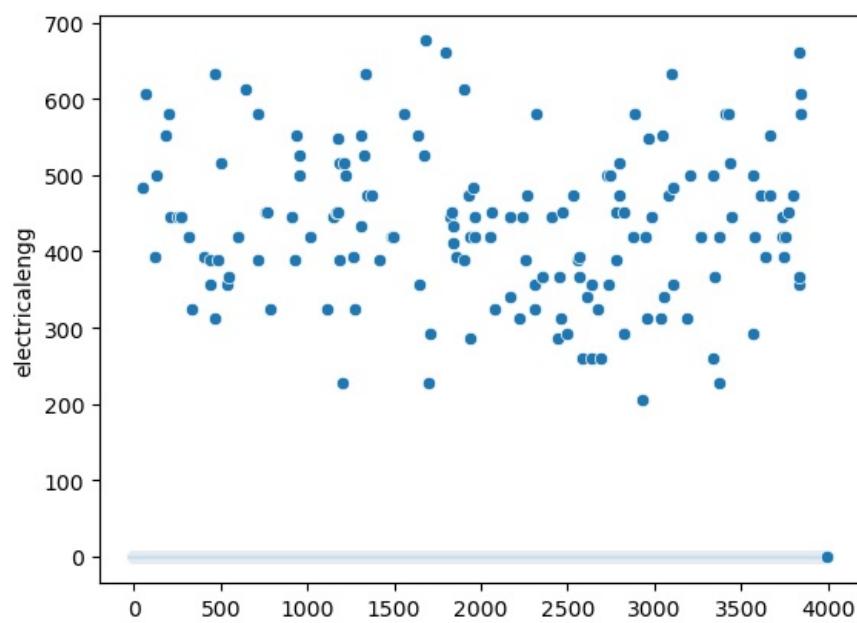


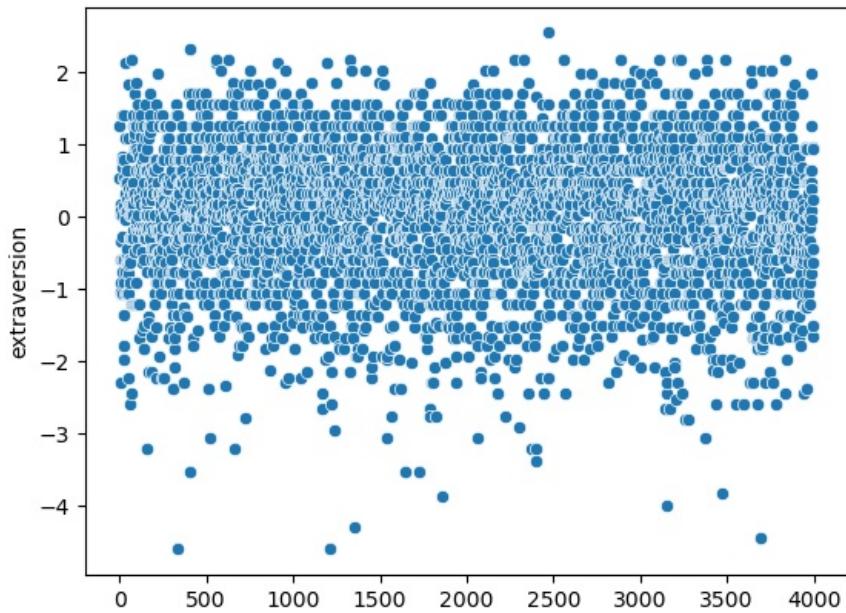
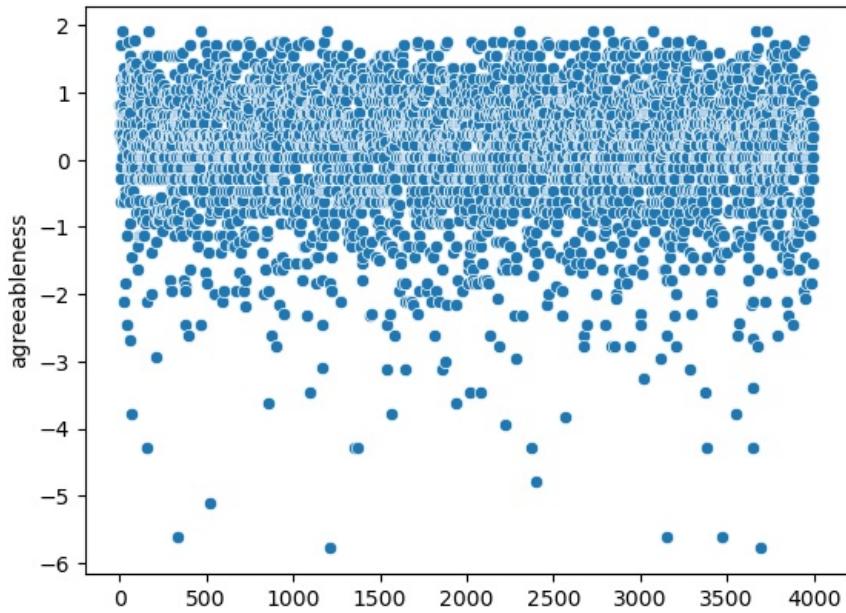
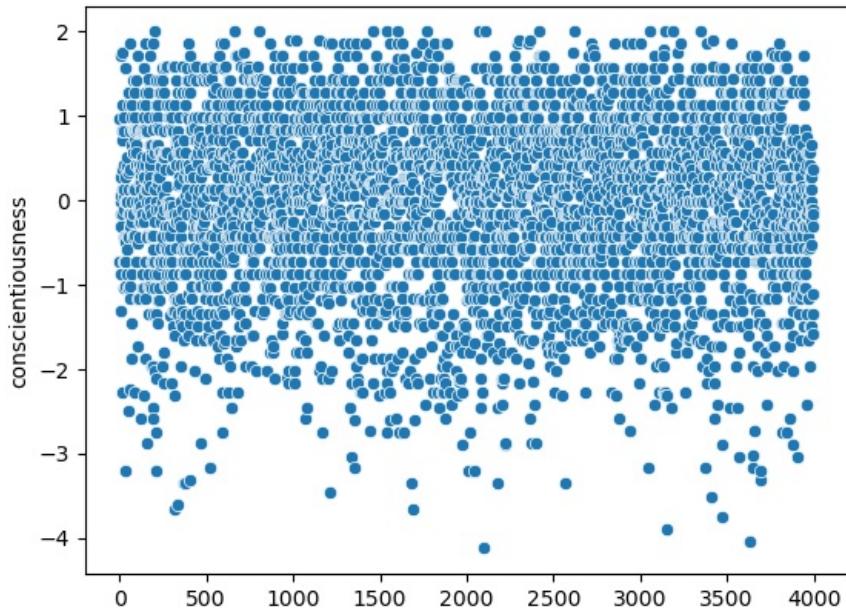
```
In [27]: #Bi-Variate Analysis
for i in num_columns[1:]:
    sns.scatterplot(df[i])
    plt.show()
```

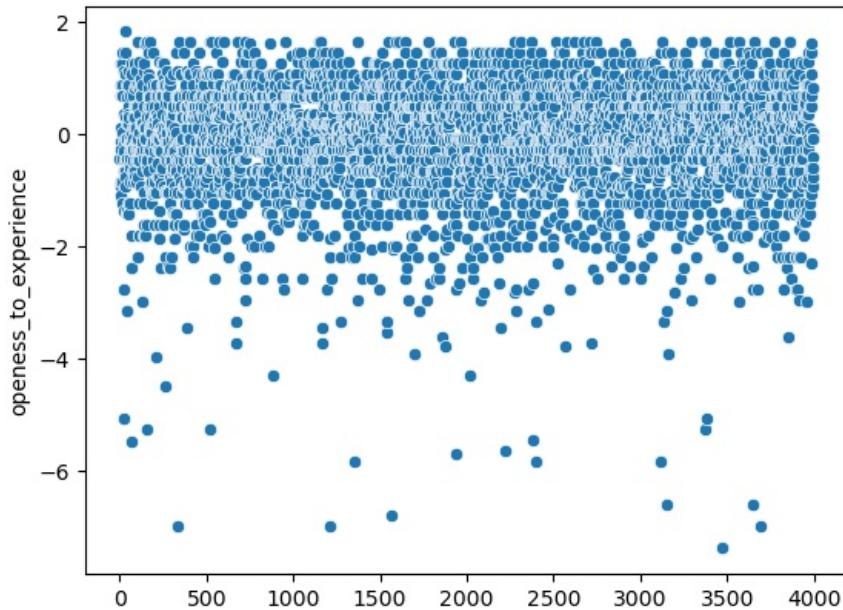
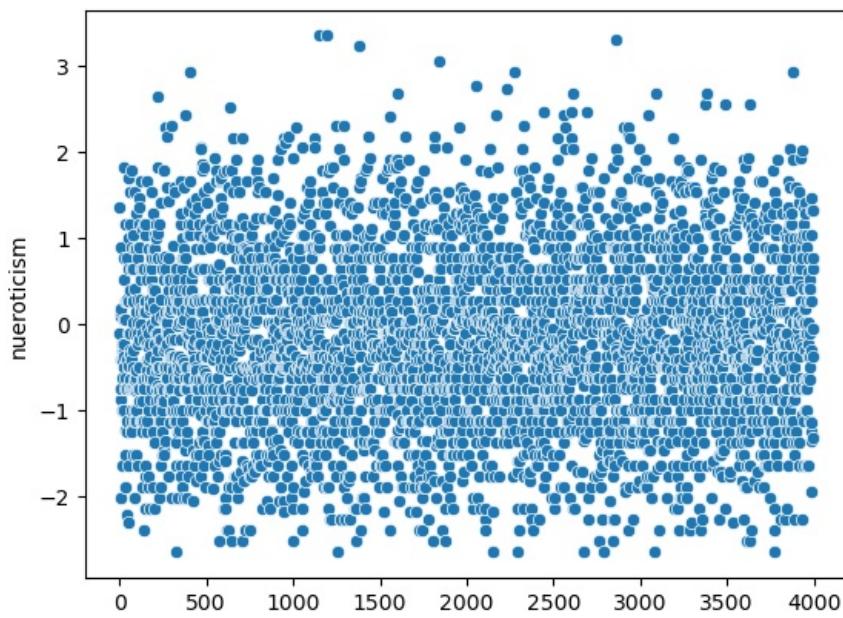






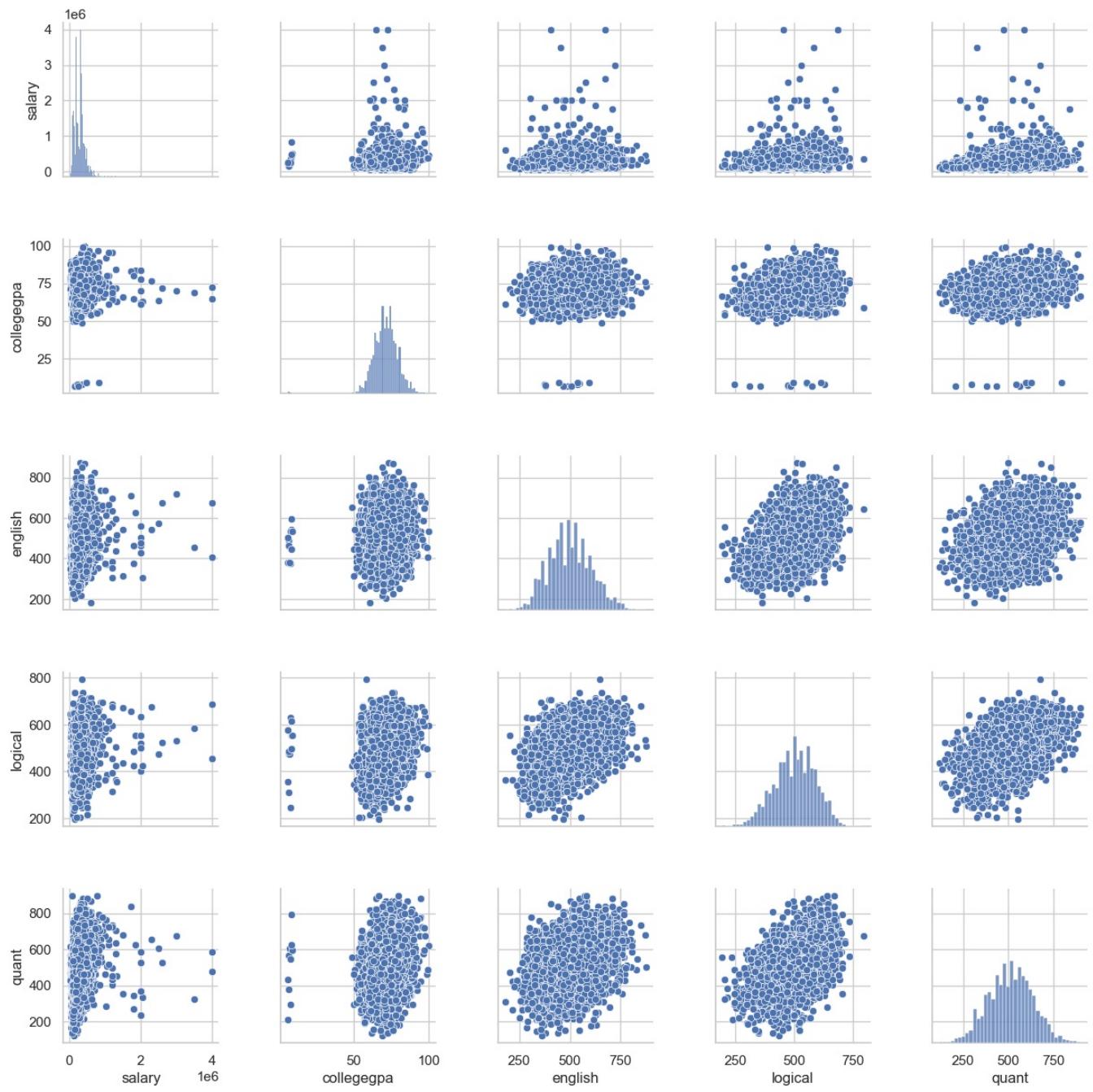




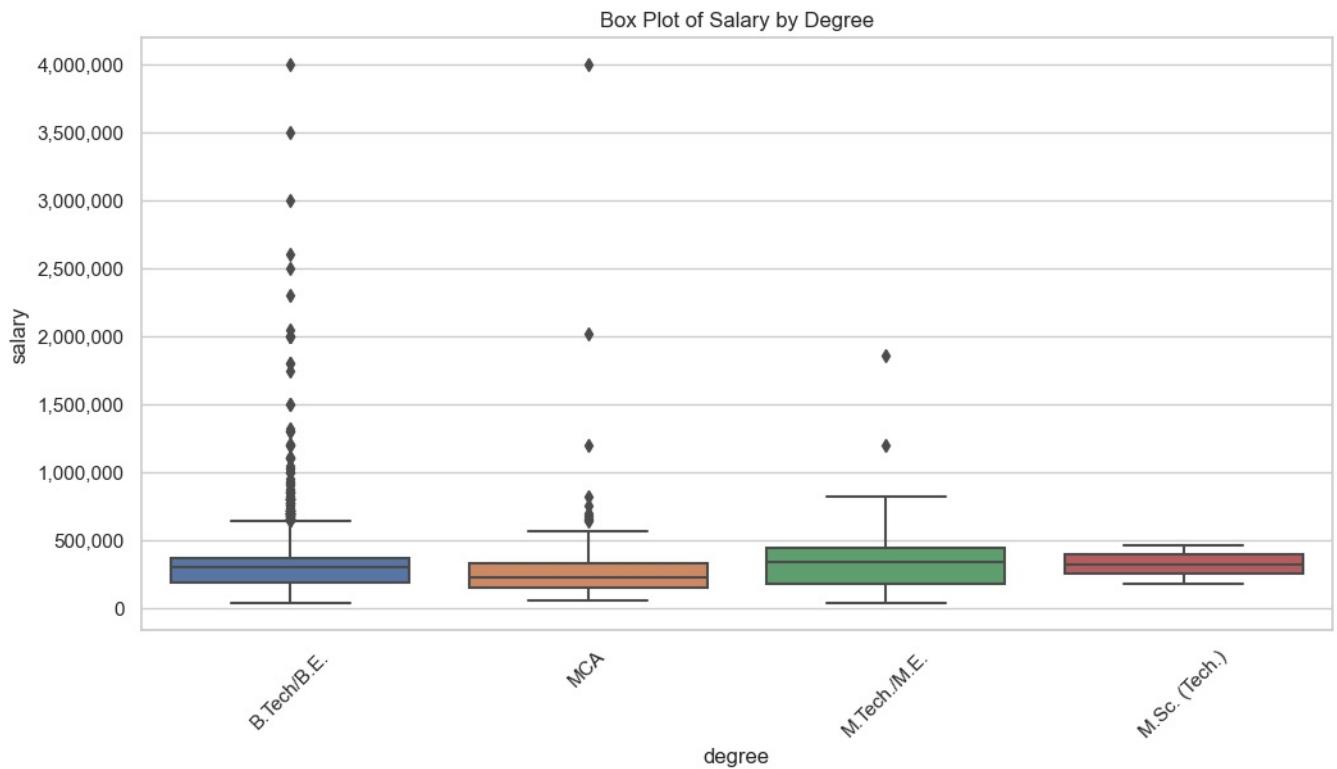


```
In [28]: numerical_columns = ['salary', 'collegegpa', 'english', 'logical', 'quant']
sns.set(style="whitegrid")
pair_plot = sns.pairplot(df[numerical_columns])
plt.suptitle('Pair Plot of Numerical Columns', y=1.02)
plt.subplots_adjust(hspace=0.4, wspace=0.4)
plt.gca().yaxis.set_major_formatter(FuncFormatter(currency))
plt.show()
```

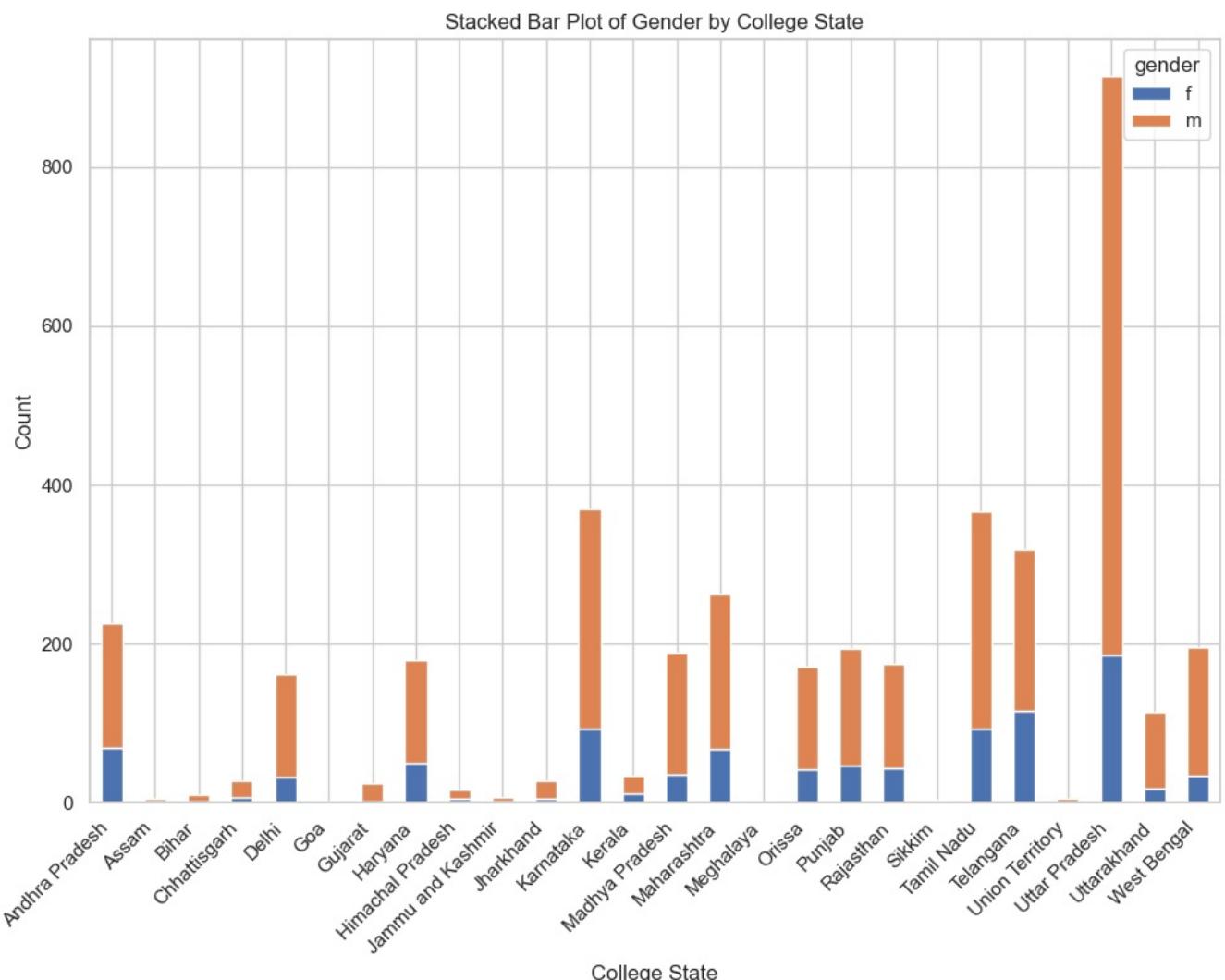
Pair Plot of Numerical Columns



```
In [29]: plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='degree', y='salary')
plt.title('Box Plot of Salary by Degree')
plt.xticks(rotation=45)
plt.gca().yaxis.set_major_formatter(FuncFormatter(currency))
plt.show()
```



```
In [30]: pivot_table = df.pivot_table(index='collegestate', columns='gender', values='salary', aggfunc='count').fillna(0)
pivot_table.plot(kind='bar', stacked=True, figsize=(10, 8))
plt.title('Stacked Bar Plot of Gender by College State')
plt.xlabel('College State')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right') # Adjusted alignment to 'right'
plt.tight_layout() # Adjust layout to prevent clipping
plt.show()
```



```
In [31]: #gpa across different features
g1=df.groupby("specialization")[[ "collegegpa"]].mean().sort_values(by="collegegpa", ascending=False)
```

g1

Out[31]:

	collegegpa
	specialization
embedded systems technology	88.000000
control and instrumentation engineering	82.100000
information science	81.200000
internal combustion engine	80.600000
industrial & management engineering	80.000000
computer science	77.385000
computer and communication engineering	77.260000
power systems and automation	76.000000
other	75.619231
metallurgical engineering	75.550000
information & communication technology	75.500000
instrumentation and control engineering	75.380000
telecommunication engineering	74.776667
mechatronics	74.375000
industrial engineering	73.850000
computer application	73.700779
mechanical and automation	73.530000
biotechnology	73.155333
industrial & production engineering	73.146000
electrical engineering	72.820000
polymer technology	72.790000
civil engineering	72.761034
automobile/automotive engineering	72.690000
electronics & instrumentation eng	72.679063
electronics and communication engineering	72.126170
electronics and electrical engineering	72.097143
ceramic engineering	72.000000
applied electronics and instrumentation	71.888889
computer science & engineering	71.779798
electronics and instrumentation engineering	71.634815
computer engineering	71.046500
electronics	71.000000
information technology	70.510803
chemical engineering	70.138889
computer networking	70.130000
mechanical engineering	70.109154
computer science and technology	69.091667
electronics & telecommunications	69.020413
aeronautical engineering	68.033333
instrumentation engineering	67.547500
information science engineering	67.322593
electronics and computer engineering	67.313333
biomedical engineering	64.650000
electronics engineering	61.318947
mechanical & production engineering	58.000000
electrical and power engineering	35.705000

In [32]:

```
#Does designation affect salary
inc=df.groupby("designation")[[ "salary"]].mean()
inc
```

Out[32] :

designation	salary
.net developer	223382.352941
.net web developer	196250.000000
account executive	287500.000000
account manager	350000.000000
admin assistant	102500.000000
...	...
web designer and seo	200000.000000
web developer	168981.481481
web intern	205000.000000
website developer/tester	200000.000000
windows systems administrator	200000.000000

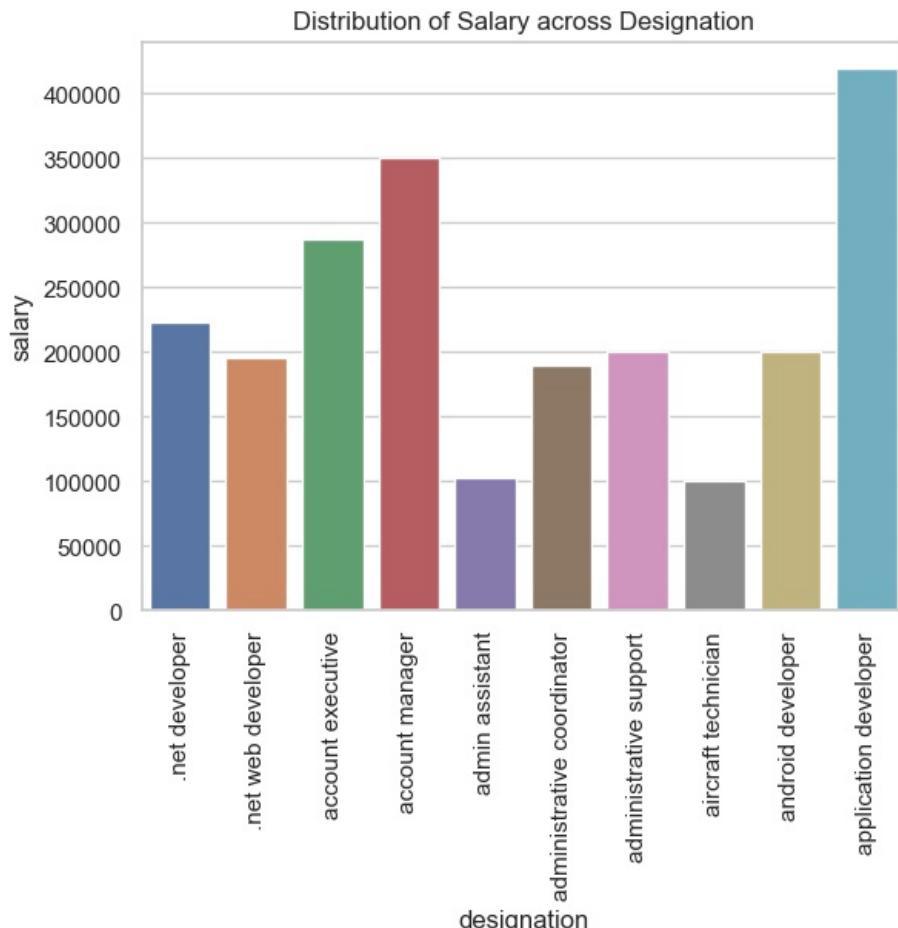
419 rows × 1 columns

```
In [33]: # Create the bar plot for the top 10 designations based on salary
sns.barplot(x=inc.index[:10], y=inc["salary"][:10])

# Rotate the x-axis labels for better readability
plt.xticks(rotation=90)

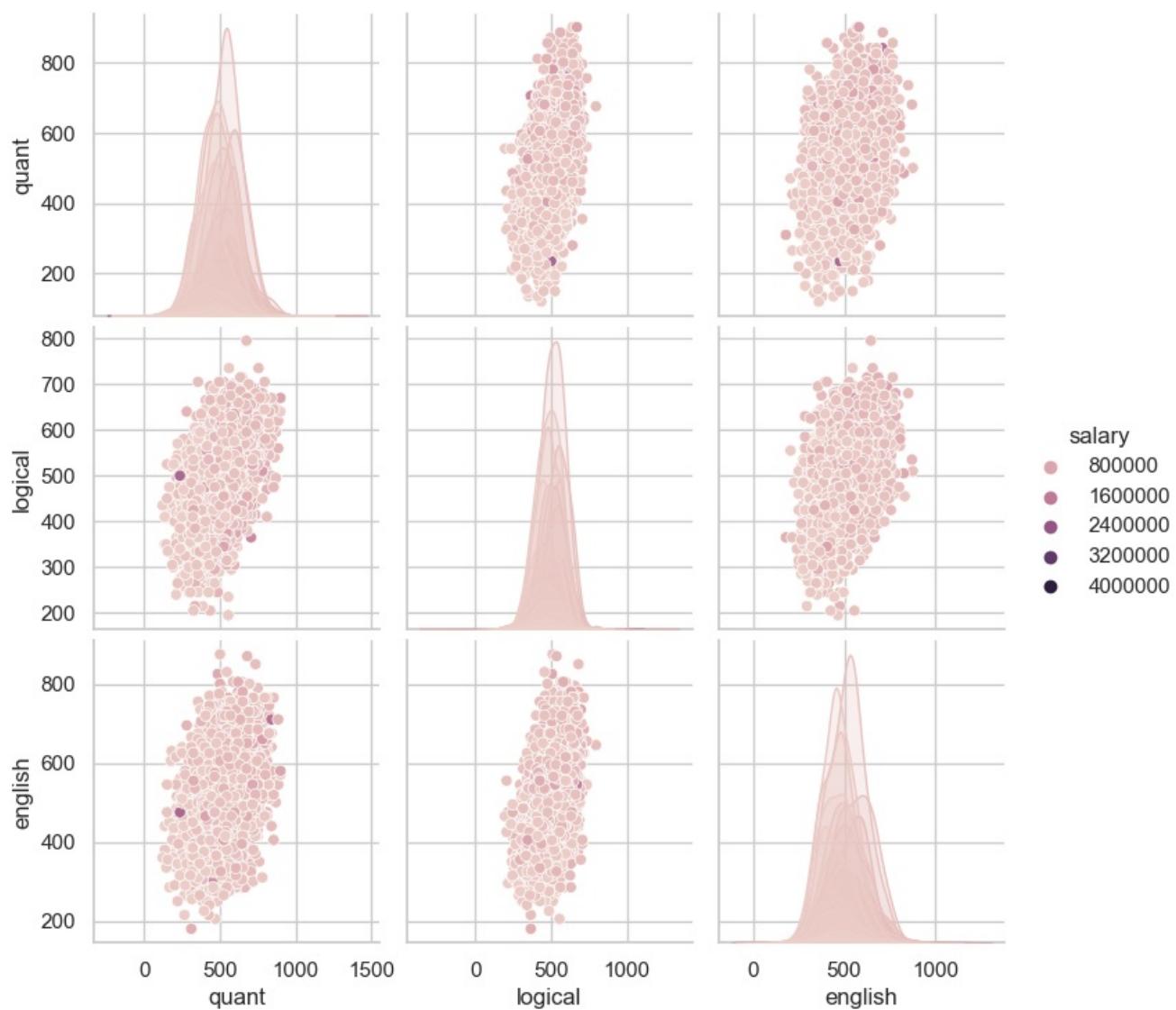
# Set the title
plt.title("Distribution of Salary across Designation")

# Show the plot
plt.show()
```

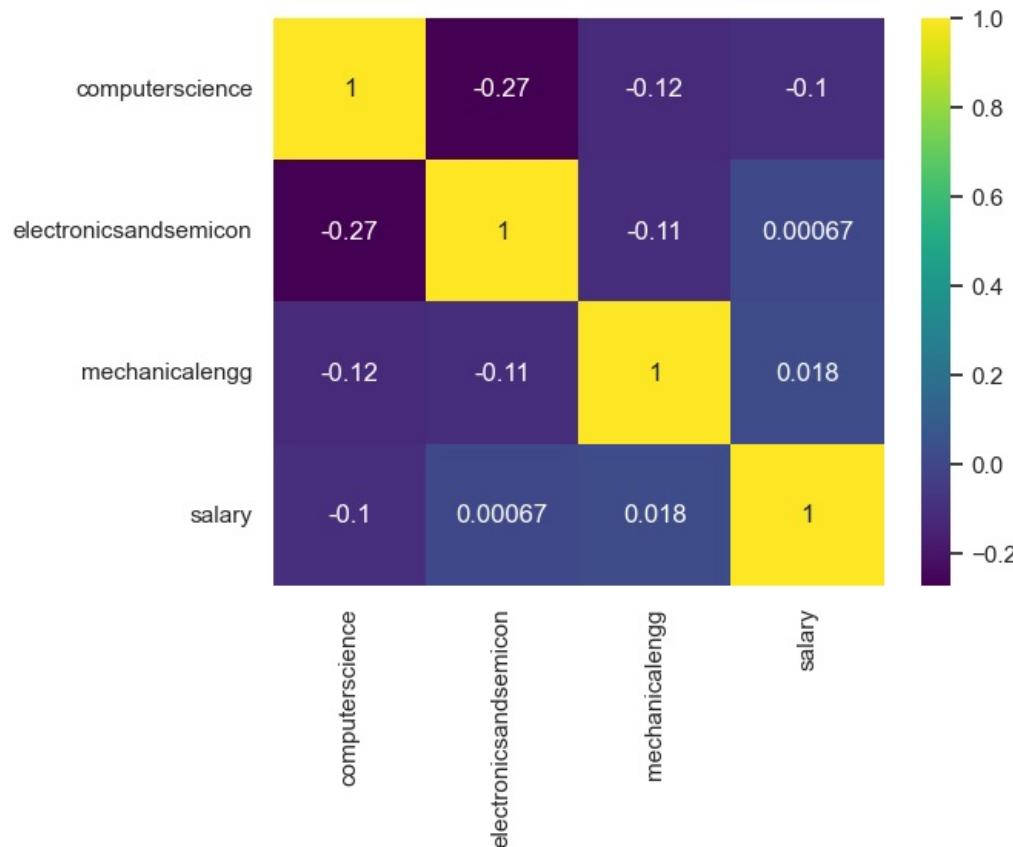


## Multivariate Analysis

```
In [34]: #how does language affect salary
sns.pairplot(df, vars=['quant', 'logical', 'english'], hue='salary')
plt.show()
```



```
In [38]: # Specialization affecting salaries
df[['computerscience', 'electronicsandsemicon', 'mechanicalengg', 'salary']].corr()
sns.heatmap(df[['computerscience', 'electronicsandsemicon', 'mechanicalengg', 'salary']].corr(), annot=True, cmap=
```



## Chi2&ttest

```
In [36]: import pandas as pd
from scipy.stats import chi2_contingency

# Create a contingency table
contingency_table = pd.crosstab(df['designation'], df['jobcity'])

# Perform the Chi-Square test
chi2_stat, p_val, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-Square Statistic: {chi2_stat}")
print(f"P-Value: {p_val}")

if p_val < 0.05:
    print("There is a significant association between Designation and Job City.")
else:
    print("There is no significant association between Designation and Job City.")
```

Chi-Square Statistic: 218968.308460868  
P-Value: 0.0  
There is a significant association between Designation and Job City.

```
In [37]: from scipy.stats import ttest_ind

# Filter the data for two different designations
designation_1 = df[df['designation'] == 'Designation_1']['salary']
designation_2 = df[df['designation'] == 'Designation_2']['salary']

# Perform an independent T-test
t_stat, p_val = ttest_ind(designation_1, designation_2)

print(f"T-Statistic: {t_stat}")
print(f"P-Value: {p_val}")

if p_val < 0.05:
    print("The difference in salary between the two designations is statistically significant.")
else:
    print("There is no significant difference in salary between the two designations.")
```

T-Statistic: nan  
P-Value: nan  
There is no significant difference in salary between the two designations.

## Conclusion:

AMCAT provide insights regarding the jobs, salaries, colleges, and the skillset of the students. Graduates with specializations in Computer Science and IT-related fields have shown a tendency to secure higher salaries, confirming the high demand for these skills in the tech industry.

Loading [MathJax]/extensions/Safe.js