INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project - AMCAT Data Analysis

Done by :

Farheen Shaik

About me :

- **Name :** Farheen Shaik
- **Qualification :** BTech (Electronic Computer Engineering)
- **Linkedin :** https://www.linkedin.com/in/shaikfarheen
- **Github :** https://github.com/Farheen2809

INNOMATICS RESEARCH LABS

Objective :

- The dataset was released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO). The study is primarily limited only to students with engineering disciplines. The dataset contains the employment outcomes of engineering graduates as dependent variables (Salary, Job Titles, and Job Locations) along with the standardized scores from three different areas – cognitive skills, technical skills and personality skills. The dataset also contains demographic features.
- The dataset contains around 40 independent variables and 4000 data points. The independent variables are both continuous and categorical in nature.
- The dataset contains a unique identifier for each candidate.

INNOMATICS RESEARCH LABS

Raw Data:

	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	84.30	...	-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	85.40	...	-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	85.00	...	-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	85.60	...	-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	78.00	...	-1	-1	-1	-1	-1	-0.8810	-0.2793	-1.0697	0.09163	-0.1295
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3993	train	47916	280000.0	10/1/11 0:00	10/1/12 0:00	software engineer	New Delhi	m	4/15/87 0:00	52.09	...	-1	-1	-1	-1	-1	-0.1082	0.3448	0.2366	0.64980	-0.9194
3994	train	752781	100000.0	7/1/13 0:00	7/1/13 0:00	technical writer	Hyderabad	f	8/27/92 0:00	90.00	...	-1	-1	-1	-1	-1	-0.3027	0.8784	0.9322	0.77980	-0.0943
3995	train	355888	320000.0	7/1/13 0:00	present	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	...	-1	-1	-1	-1	-1	-1.5765	-1.5273	-1.5051	-1.31840	-0.7615
3996	train	947111	200000.0	7/1/14 0:00	1/1/15 0:00	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	...	438	-1	-1	-1	-1	-0.1590	0.0459	-0.4511	-0.36120	-0.0943
3997	train	324966	400000.0	2/1/13 0:00	present	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	...	-1	-1	-1	-1	-1	-1.1128	-0.2793	-0.6343	1.32553	-0.6035

3998 rows × 39 columns

INNOMATICS RESEARCH LABS

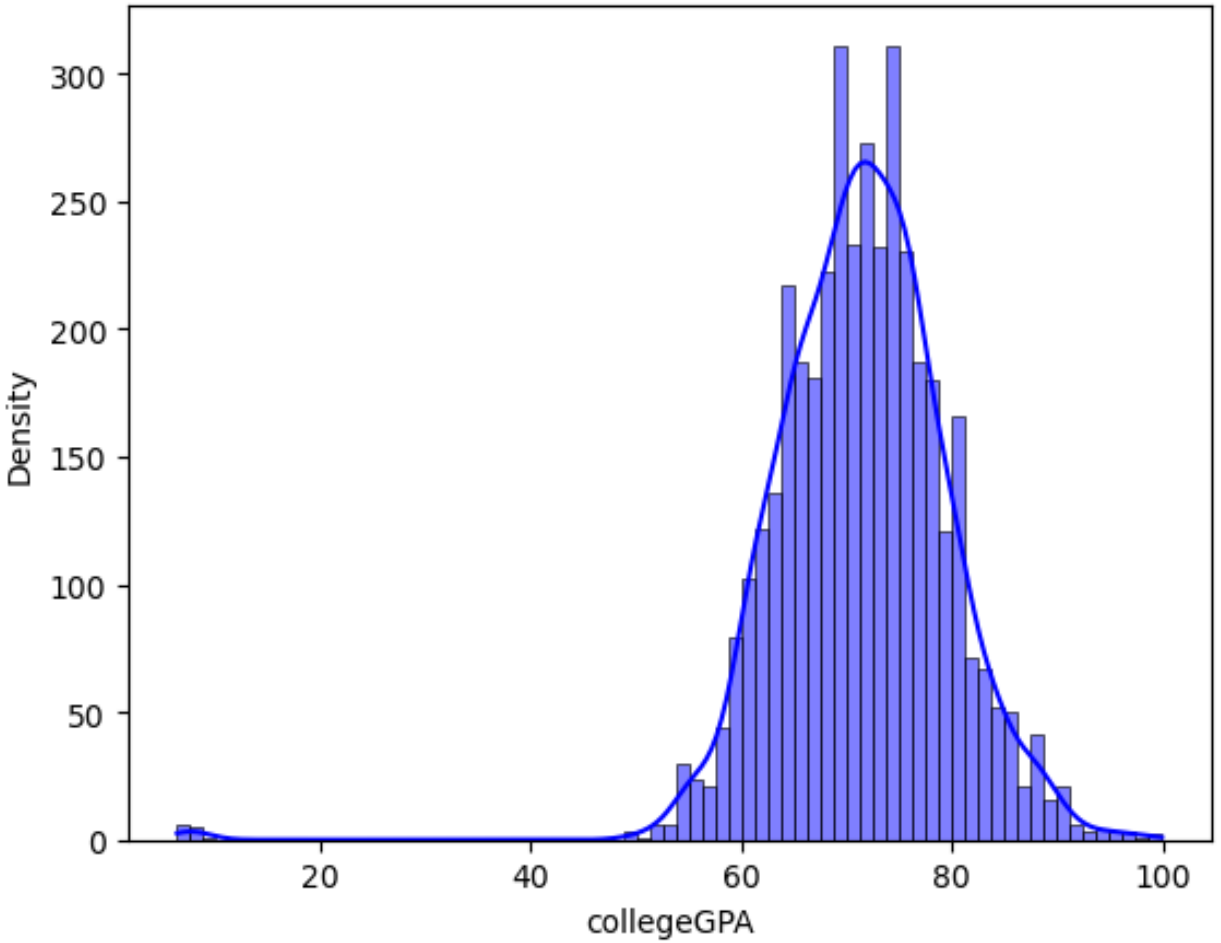
Data Cleaning :

```
<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
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3998 entries, 0 to 3997
Data columns (total 26 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     3998 non-null   int64
 1   Salary                 3998 non-null   float64
 2   Designation            3998 non-null   object
 3   JobCity                3998 non-null   object
 4   Gender                 3998 non-null   object
 5   DOB                    3998 non-null   object
 6   10percentage           3998 non-null   float64
 7   10board                3998 non-null   object
 8   12graduation           3998 non-null   int64
 9   12percentage           3998 non-null   float64
 10  12board                3998 non-null   object
 11  CollegeID              3998 non-null   int64
 12  CollegeTier            3998 non-null   int64
 13  Degree                 3998 non-null   object
 14  Specialization         3998 non-null   object
 15  collegeGPA             3998 non-null   float64
 16  CollegeState           3998 non-null   object
 17  GraduationYear         3998 non-null   int64
 18  English                3998 non-null   int64
 19  Logical                3998 non-null   int64
 20  Quant                  3998 non-null   int64
 21  conscientiousness      3998 non-null   float64
 22  agreeableness          3998 non-null   float64
 23  extraversion           3998 non-null   float64
 24  nueroticism            3998 non-null   float64
 25  openess_to_experience  3998 non-null   float64
dtypes: float64(9), int64(8), object(9)
memory usage: 843.3+ KB
```

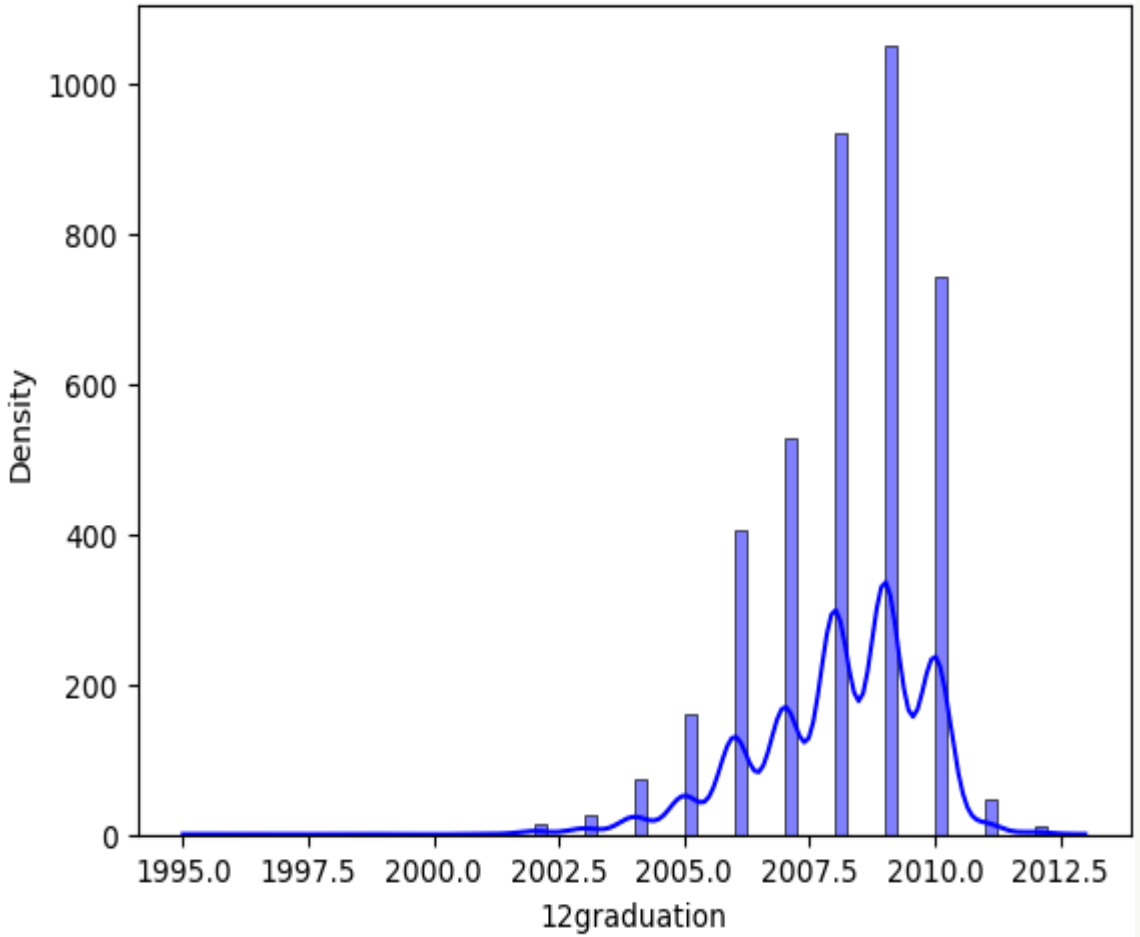
INNOMATICS RESEARCH LABS

Analysis on collegeGPA :



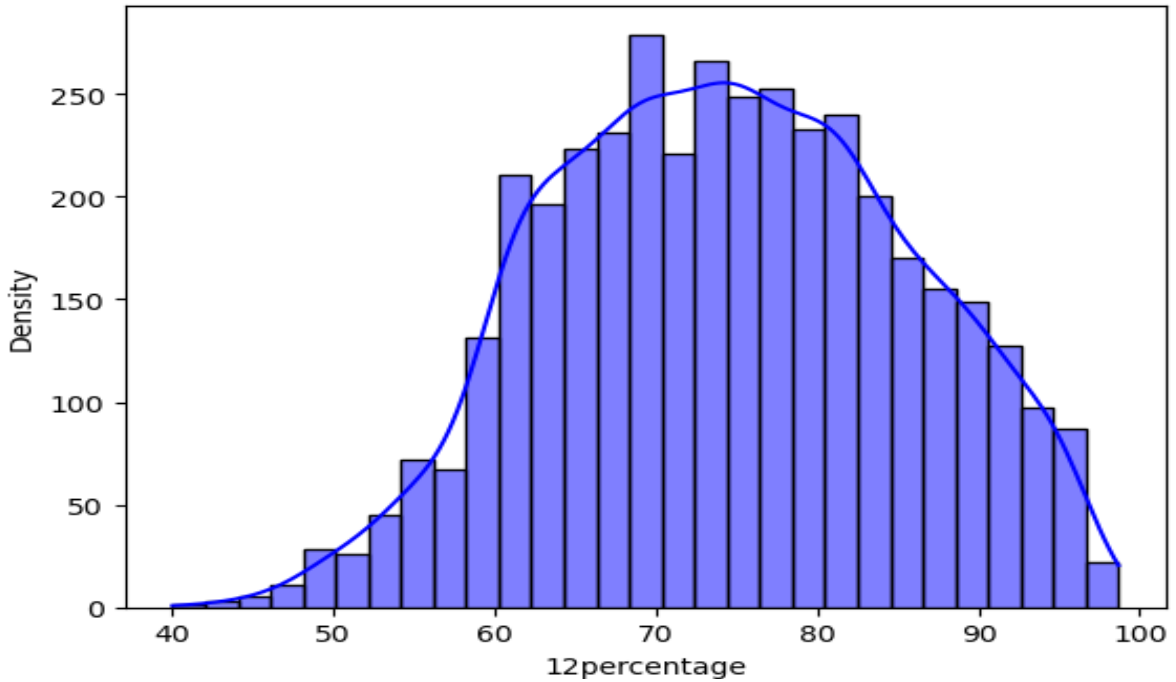
- The probability Density Function Describes(PDF) the density of collegeGPA of all employees. Most of the employees are between 65 to 75.

Analysis on 12graduation :



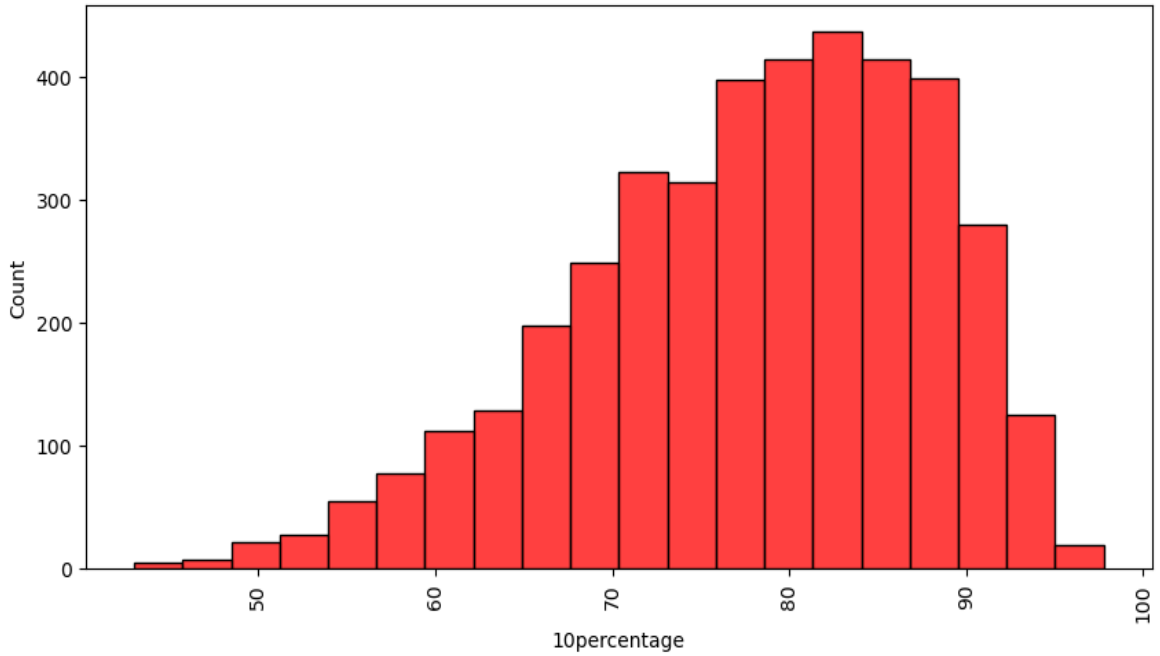
- This probability density function(PDF) describes the density of graduation of all employees are passed out in the year 2009.

INNOMATICS RESEARCH LABS



Analysis on 12percentage

- This Probability density function(PDF) describes the density of 12percentage all employees and the most of the employees are passed in between 70 percentage.

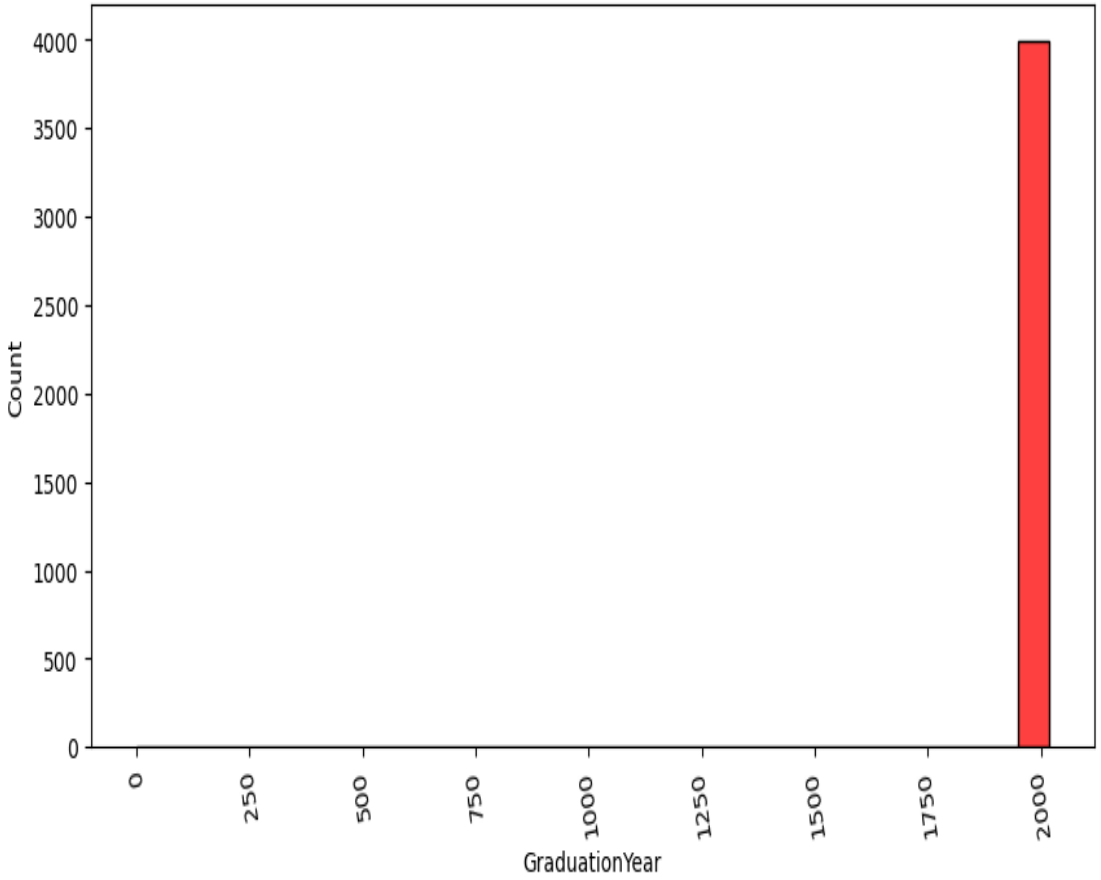


Analysis on 10percentage

- This Histogram plot describes the den of 10percentage of all employees and most of the employees are passed in approximately 83percentage.

INNOMATICS RESEARCH LABS

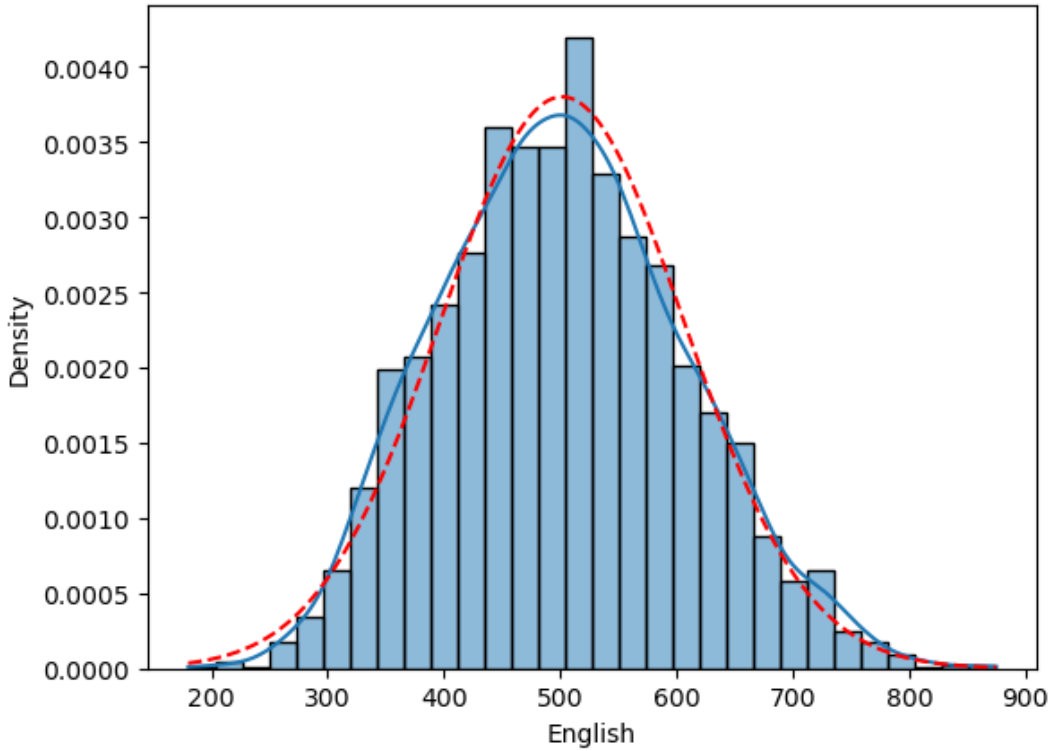
Analysis on GraduationYear :



- Here the Histogram describes the GraduationYear of employees. so, the most of the emplyees are completed their graduation in the year of 2000.

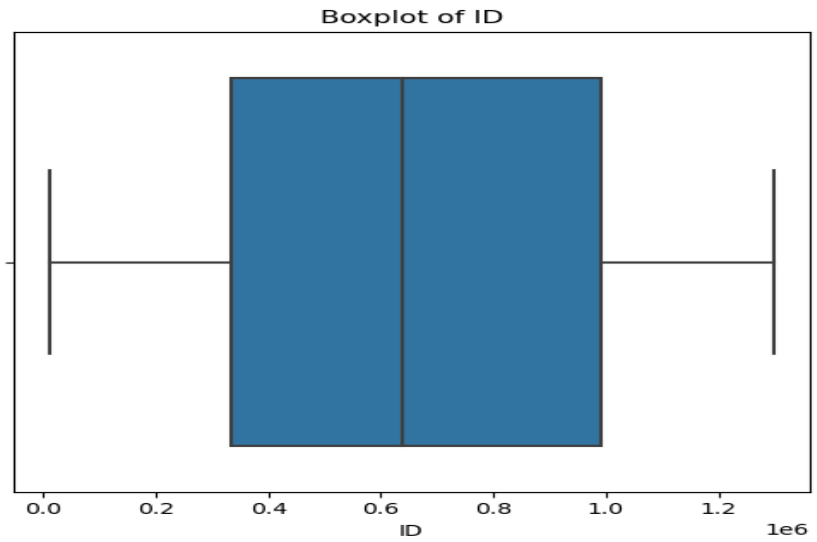
Analysis on English variable :

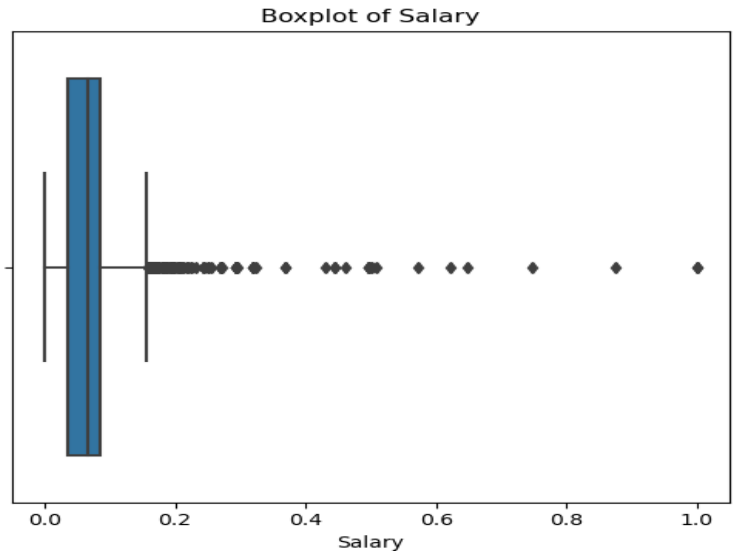
- Here the plot describes the density of English. Most of the employees has given preference to English from 440 to 520.

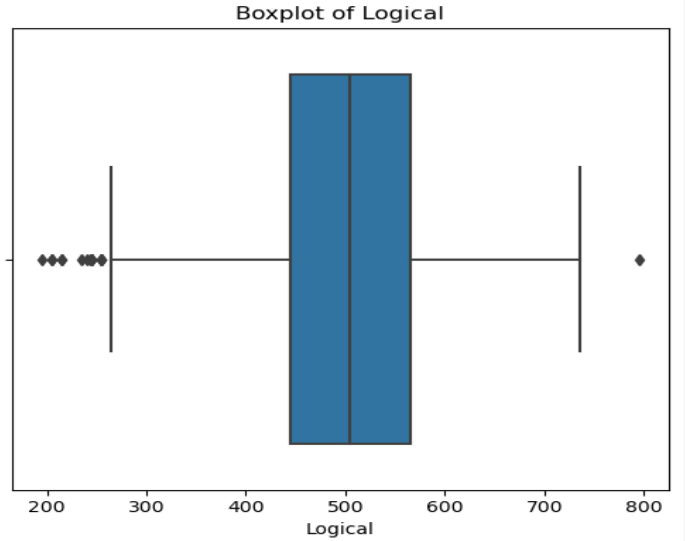


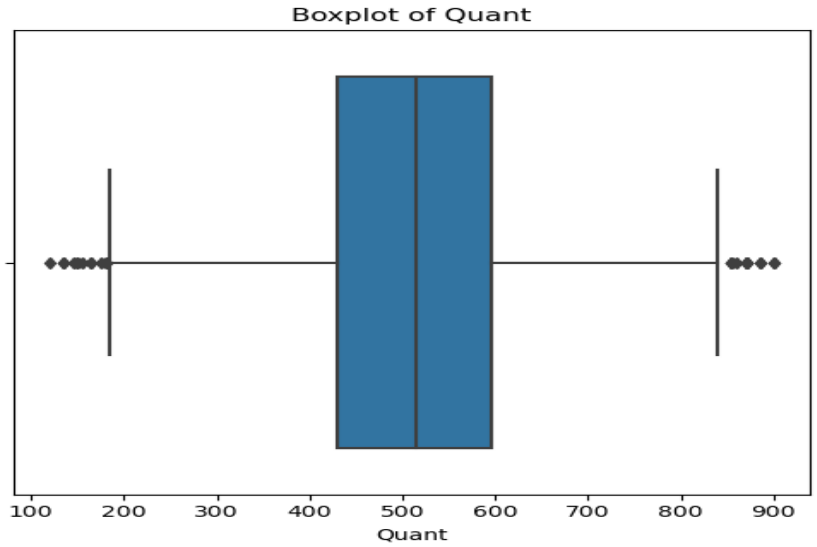
INNOMATICS RESEARCH LABS

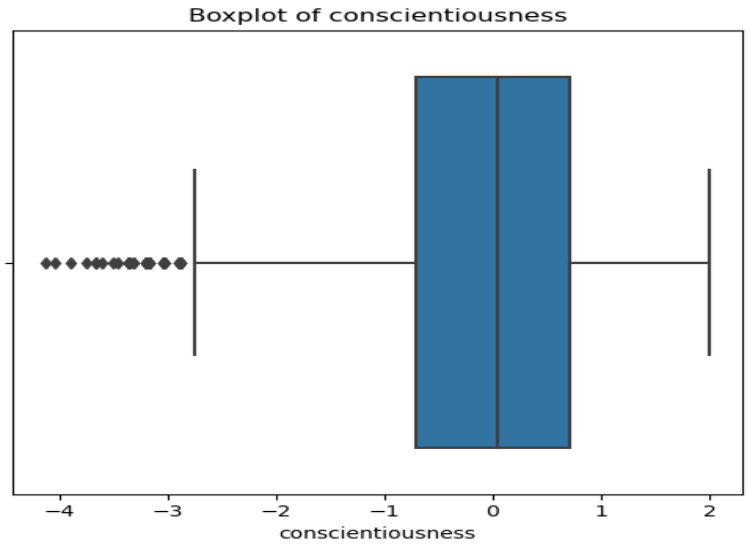
Univariate Analysis of Boxplot :







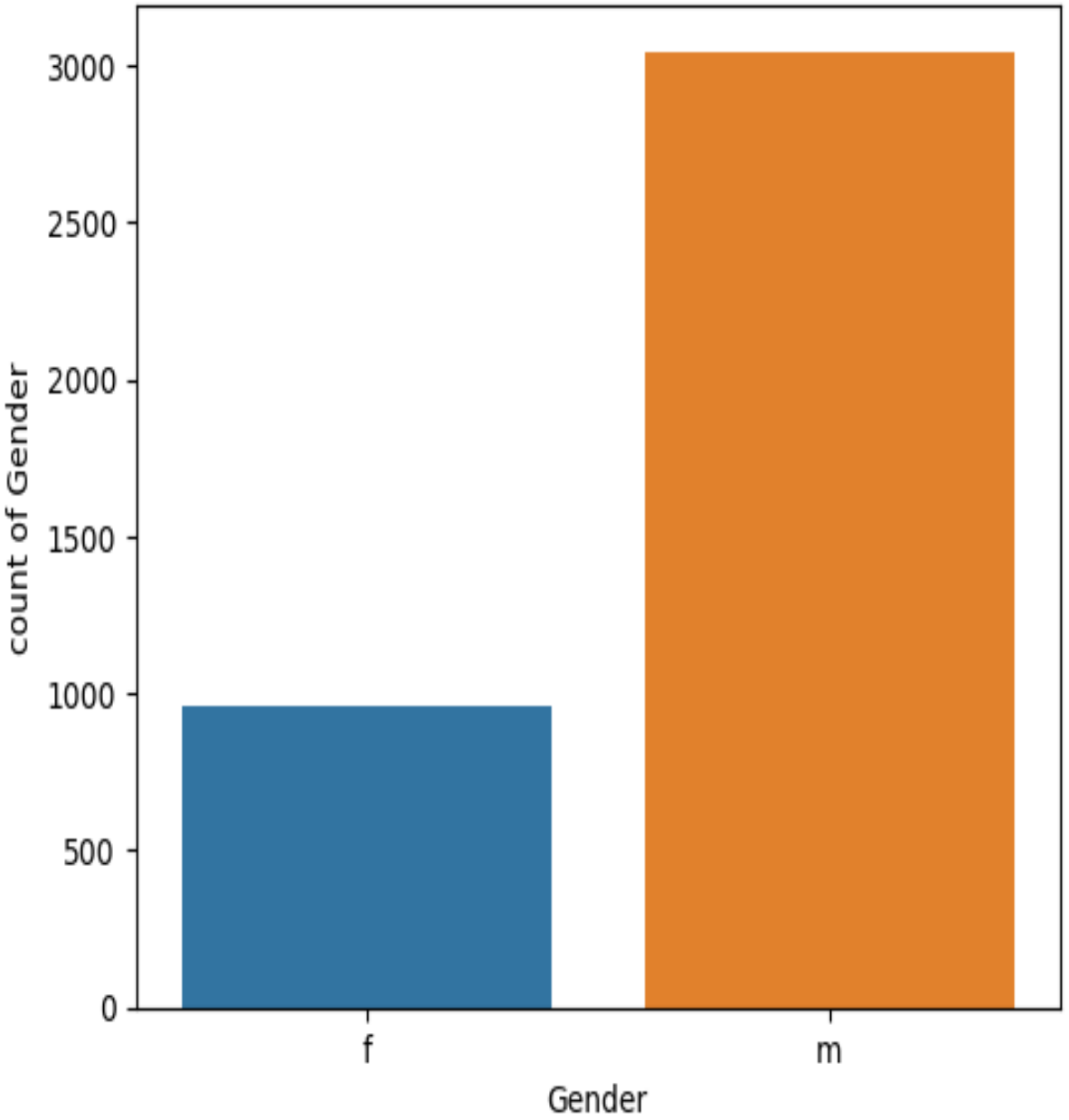




INNOMATICS RESEARCH LABS

Observations of Outliers:

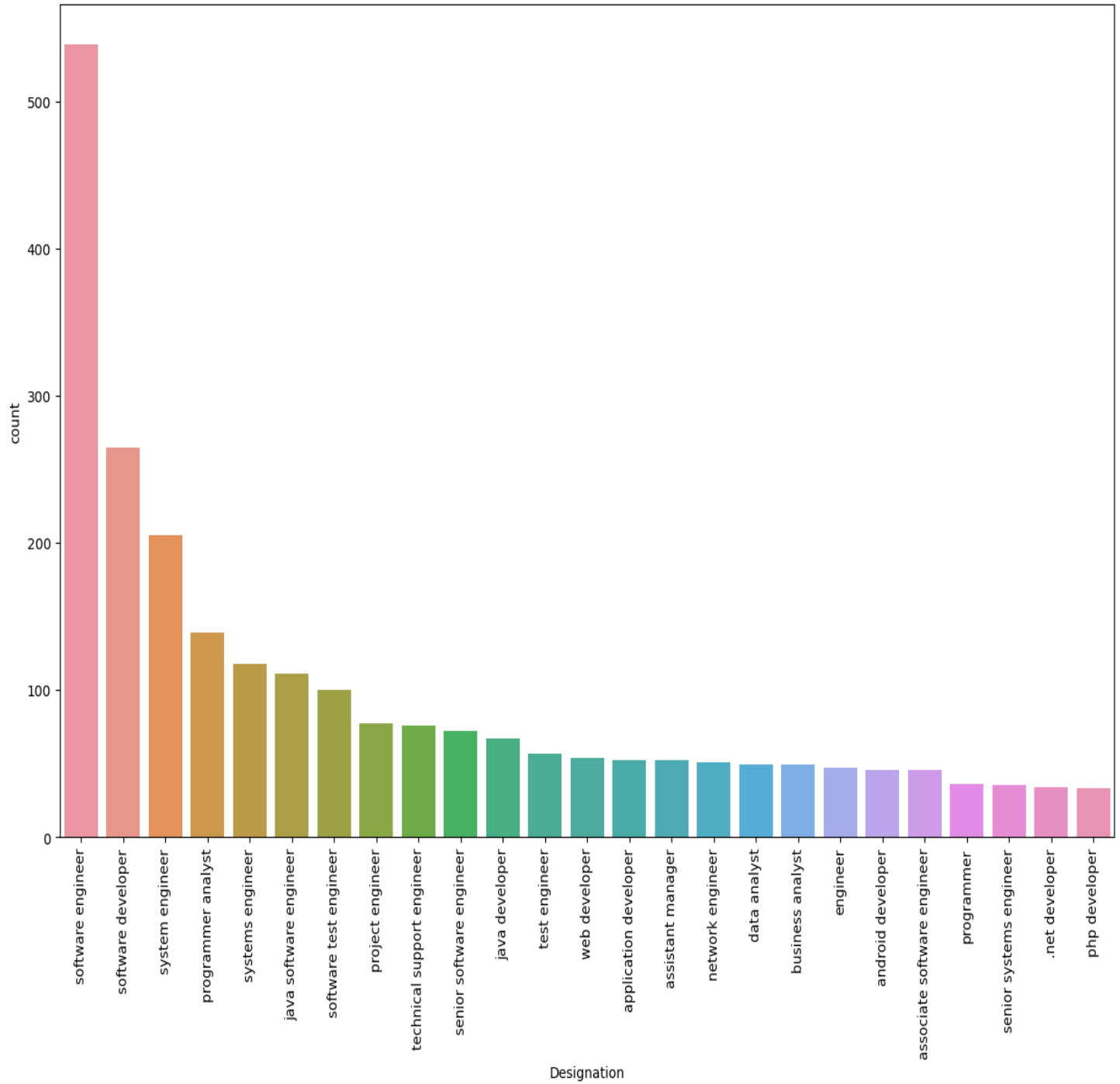
- Here the plots are describing the outliers of variables. After visualizing the data using a boxplot for the "ID", there are no outliers present in the distribution. The boxplot displays a symmetric and compact box, indicating that the majority of the data points fall within the interquartile range (IQR) with no extreme values or influential outliers. The absence of outliers enhances the reliability of measures of central tendency, such as the median, providing a more accurate representation of the typical value for "ID".
- In the boxplot analysis of the "Salary" it shows the presence of right-side outliers, the potential extreme values in the upper range of the data. The distribution of the data is positively skewed, it has a long right tail, and the outliers are pulling the mean and the right end of the boxplot upward.
- The boxplot analysis of the "Logical" variable reveals an asymmetric distribution with a notable difference in the number of outliers on the left and right sides.The "Logical" variable exhibits a median value of 505, an interquartile range (IQR) of 120, and a spread of values within 195 min, 795max.
- The boxplot analysis of the "Quant" reveals a symmetric distribution with outliers present on both the right and left sides. The "Quant" variable exhibits a median value of 515, an interquartile range (IQR) of 165, and a spread of values within 120 min, 900 max. The boxplot for "Quant" displays a central box representing the interquartile range, with a median line dividing the box. Notably, outliers are observed on both the right and left sides, indicating a balanced skewness in both directions.
- In the boxplot analysis of the "conscientiousness" it shows the presence of left-side outliers, the potential extreme values in the lower range of the data. The distribution of the data is negatively skewed, it has a long left tail, and the outliers are pulling the mean and the left end of the boxplot downward.



Analysis on Gender :

- The countplot analysis of the “Gender” variable. Here the highest number of employees are males and lowest is females.

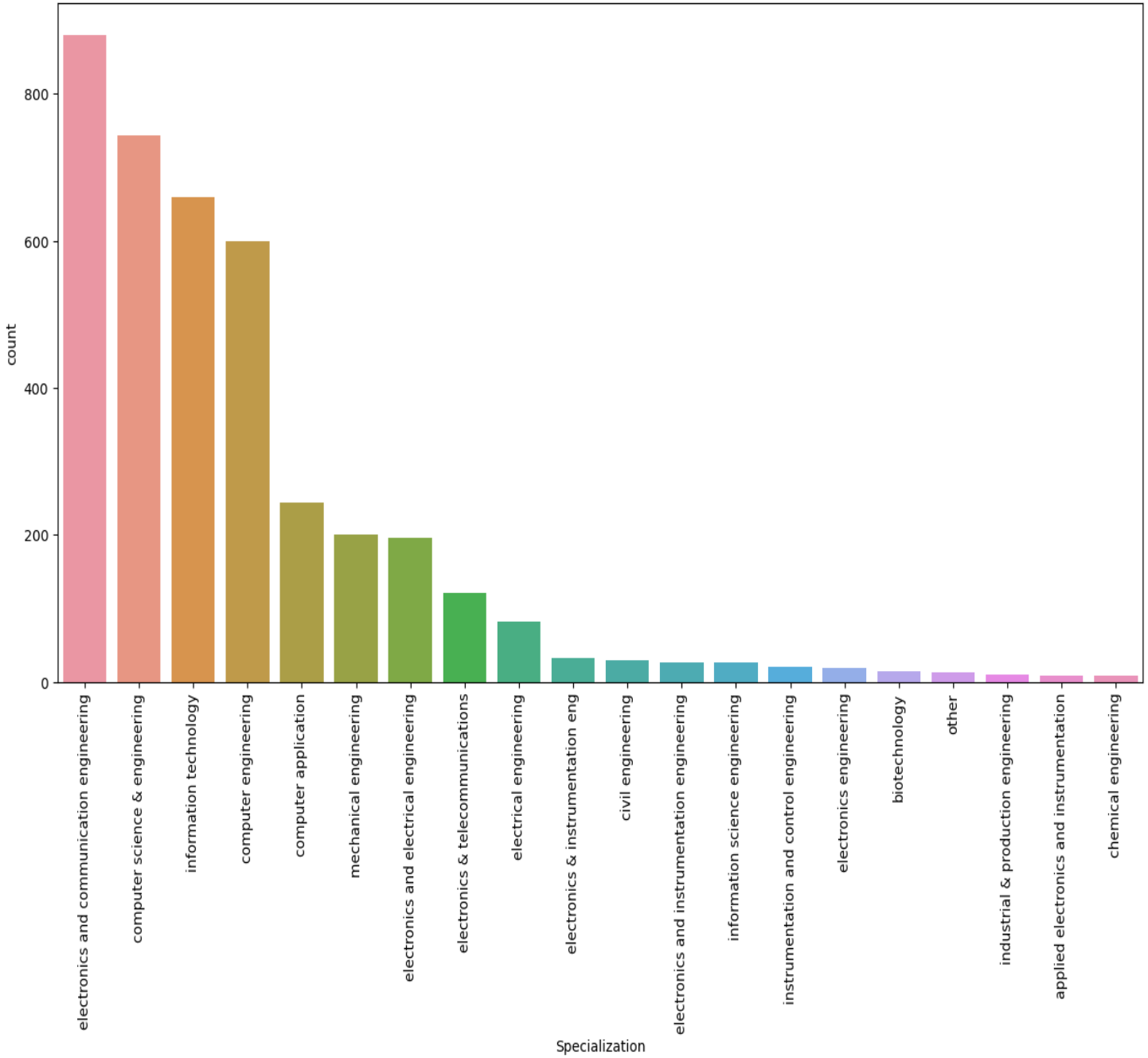
INNOMATICS RESEARCH LABS



Analysis on Designation:

- In this plot analysis of "Designation". Here most of the employees Designation is software engineer and lowest number of Designation is php developer.

INNOMATICS RESEARCH LABS

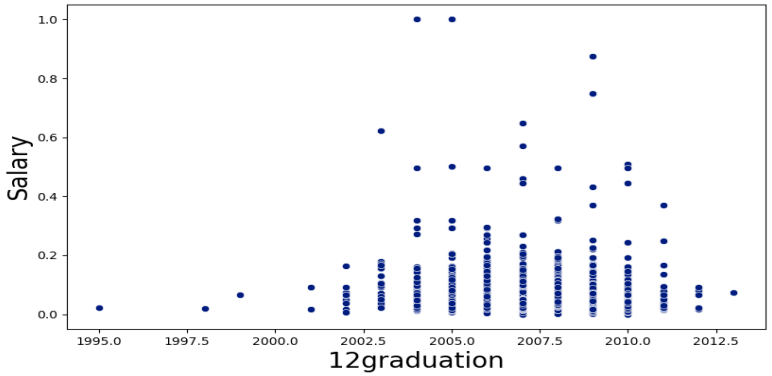


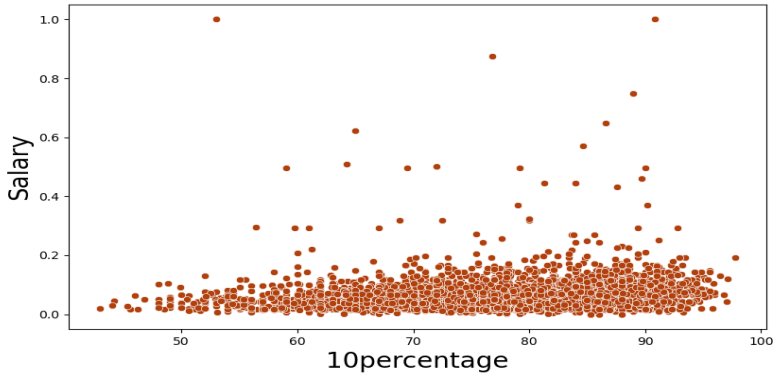
Analysis on Specialization:

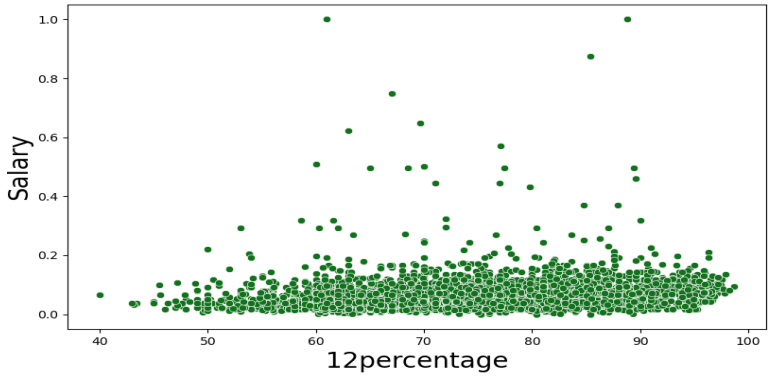
- In this plot analysis of "Specialization". Here most of the employees "Specialization" is electronics and communication engineering and lowest number of employees "Specialization" is chemical engineering.

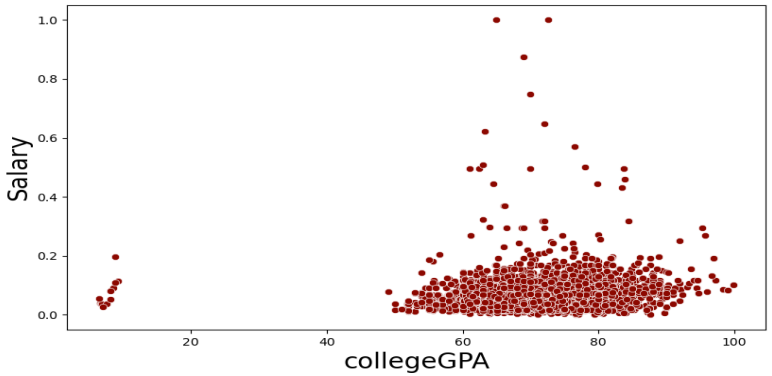
INNOMATICS RESEARCH LABS

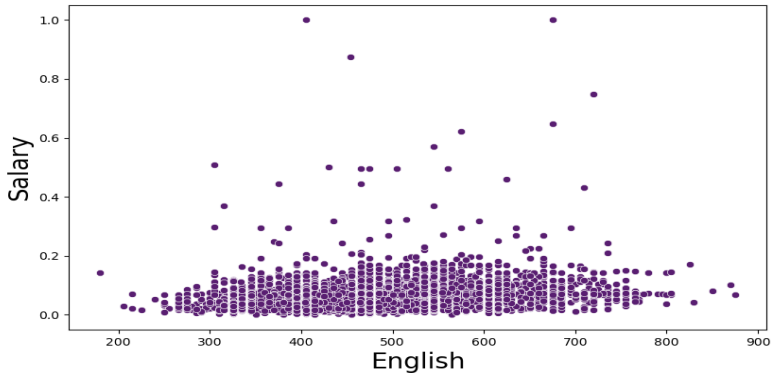
Analysis on Outliers :

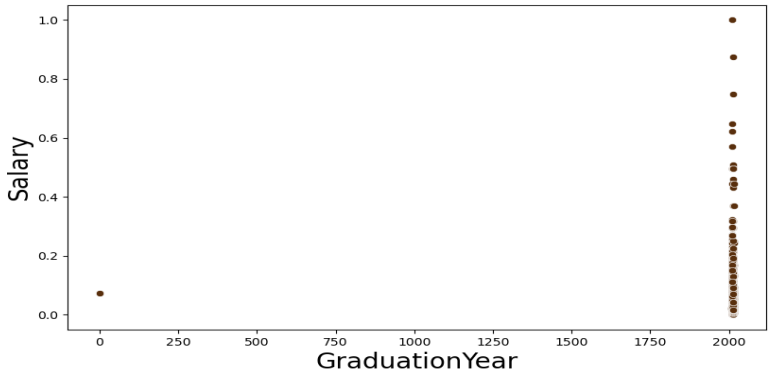


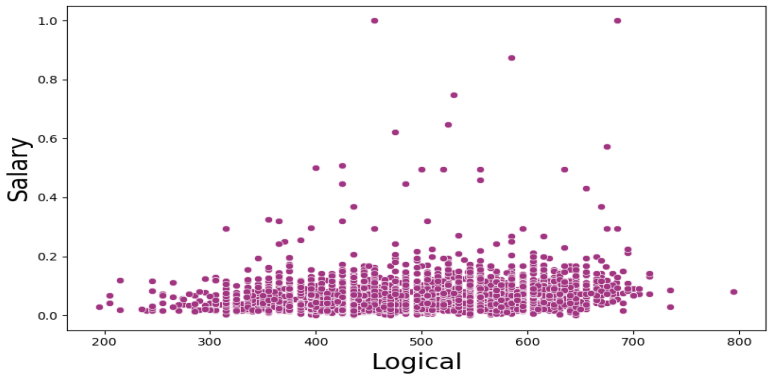


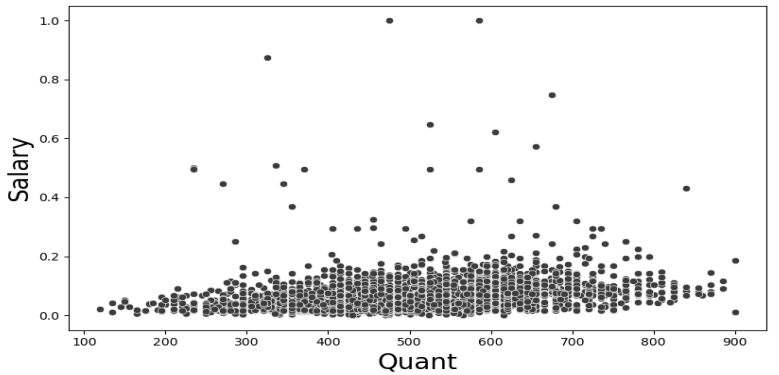


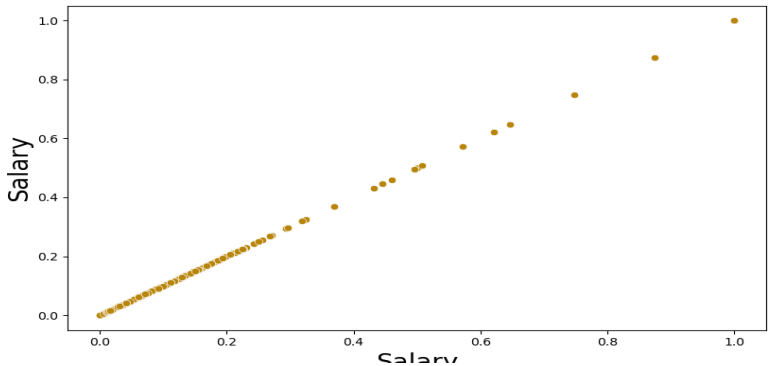






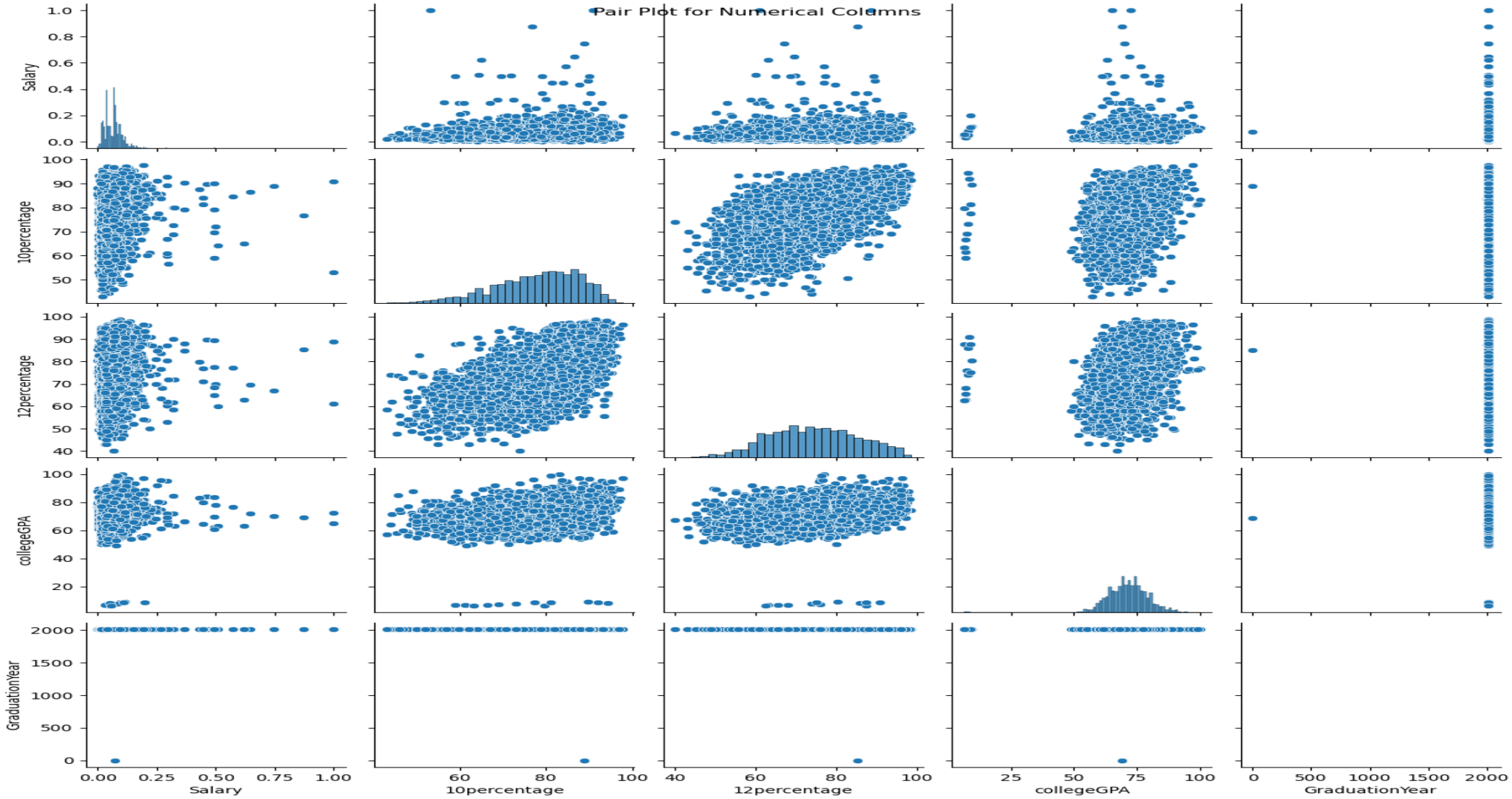






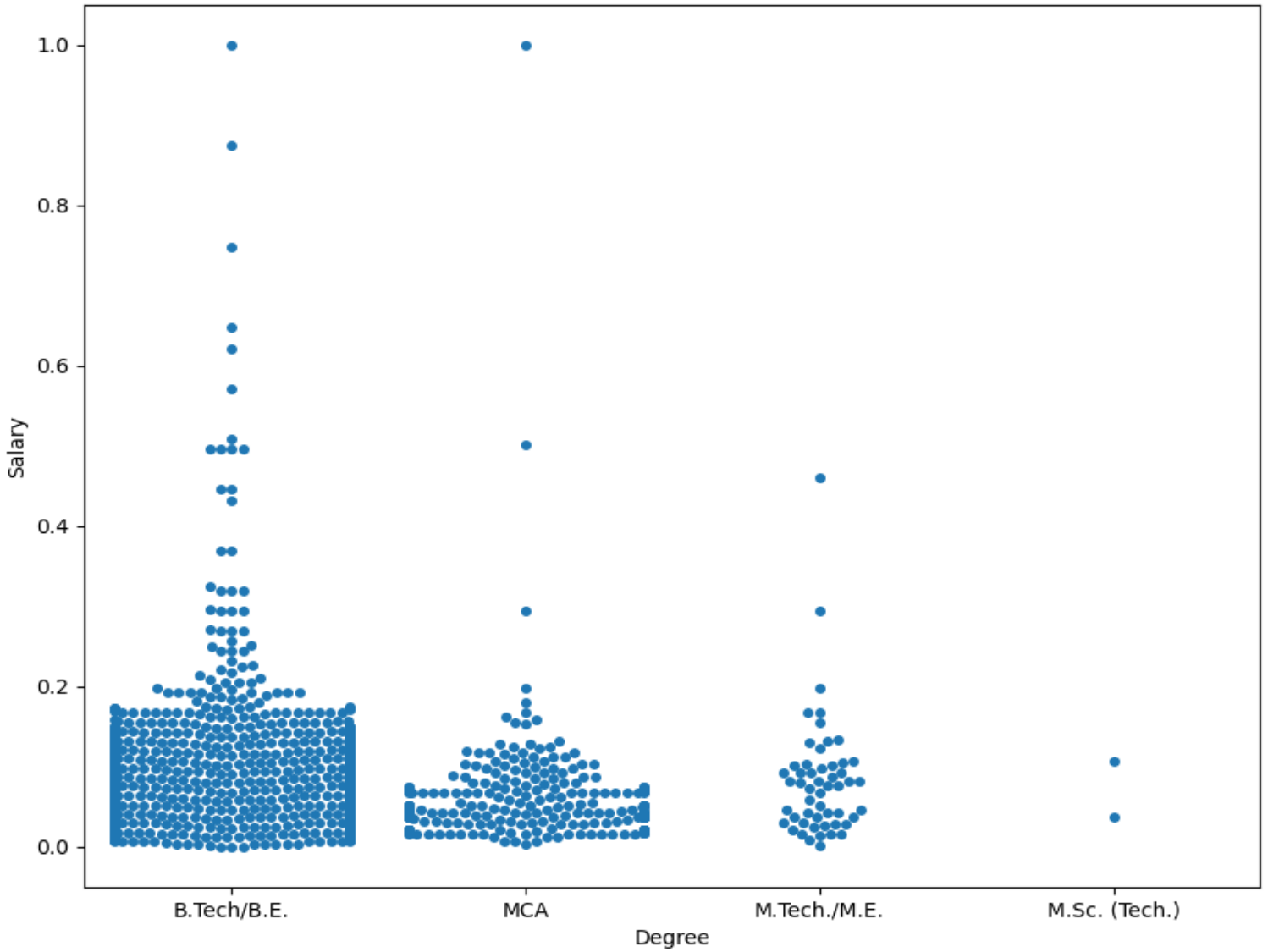
INNOMATICS RESEARCH LABS

Analysis on Pair plots of numerical columns :



INNOMATICS RESEARCH LABS

Relationship between Degree & Salary:

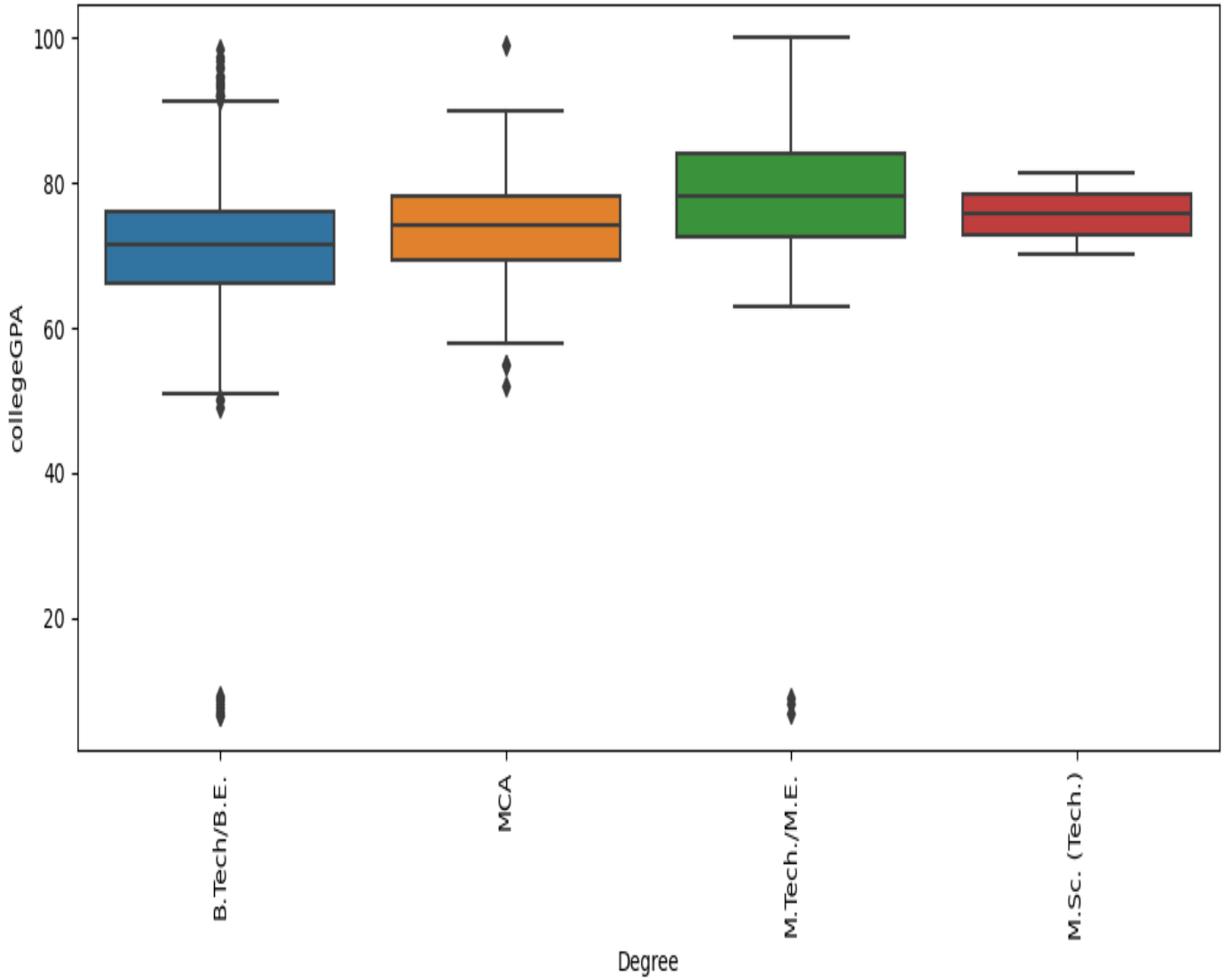


Observations :

- This swarmplot analysis the relationship between Degree and Salary. Here the b.Tech/B.E backround employees have highest salaries and M.sc.(Tech) employees have lowest salaries.

INNOMATICS RESEARCH LABS

Relationship between Degree & collegeGPA:

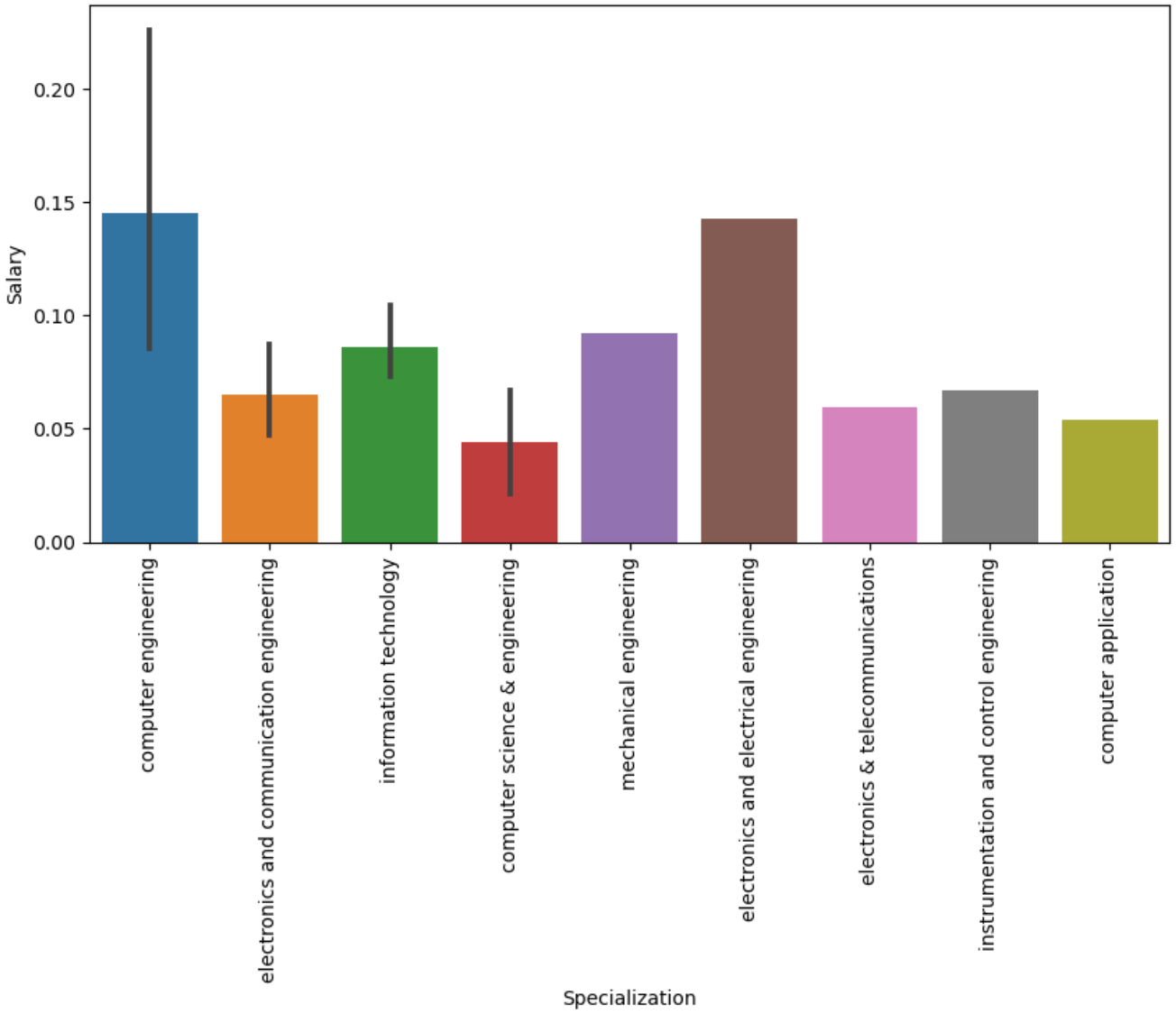


Observations :

- This Box plot analysis the relationship between Degree and collegeGPA variables. Here the B.Tech/B.E Degree have more outliers in upward direction and downward direction comparing with other Degrees. so, the B.tech/B.E background employees have highest collegeGPA.

INNOMATICS RESEARCH LABS

Relationship between Specialization and Salary:



Observation:

This bar plot represent an estimate of central tendency between categorical and numerical variable with the height of each rectangle. This is the relationship between Specialization and Salary variables of Top 20 values.

INNOMATICS RESEARCH LABS

Research Questions:

- 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. Test this claim with the data given to you.

  - **Is there a relationship between gender and specialization? (i.e. Does the preference of Specialization depend on the Gender)?**

```
cse_data=df[df['Degree']=='B.Tech/B.E.']
job_roles=[ 'Programming Analyst', 'Software Engineer', 'Hardware Engineer' ,'Associate Engineer']
req_data=cse_data[cse_data['Designation'].isin(job_roles)]
salary_range=(250000,300000)
avg_sal=req_data['Salary'].mean()
if salary_range[0]<=avg_sal<=salary_range[1]:
    print('The claim is supported by the Data')
else:
    print('The claim is not supported by the Data')
```

```
The claim is not supported by the Data
```

INNOMATICS RESEARCH LABS

Relationship between endGer & Specialization:

```
from scipy import stats
contingency_table=pd.crosstab(df['Gender'],df['Designation'])
chi2_stats,p_val,dof,expected=stats.chi2_contingency(contingency_table)
alpha=0.05
if p_val< alpha:
    print("There is a significant relationship between gender and specialization preference.")
else:
    print("There is no significant relationship between gender and specialization preference.")
```

```
There is a significant relationship between gender and specialization preference.
```

Observation:

- Based on the analysis, the claim suggesting a salary range of 2.5-3 lakhs for Computer Science Engineering graduates in specific job roles is not supported by the data. Additionally, the statistical test results indicate that there is no significant relationship between gender and specialization preference among graduates, as per the AMCAT dataset.

Conclusion:

- The dataset contains the employement outcomes of engineering graduates as dependent variable (Salary, Job Titles and Job locations)along with the standardized scores from three different areas – cognitive skill, technical skill, personality skill.
- Here we have observe the dataset contains 4000 rows and 40 columns and this dataset have so many duplicated values and first we have to manipulate the dataset and remove the unwanted rows and columns after that check the nan values are there or not after that we have to take a cleaned dataset visualizations.
- Here we used univariate analysis and many plot to analyse the dataset like PDF, Histograms, Boxplots, Countplots in this analysis we found outliers in each numerical column, probability and frequency distribution of each numerical column, e frequency distribution of each categorical Variable/Column and we Mention observations after each plot.
- And along with univariate analysis we use bivariate analysis, In this bivariate analysis we analysis many plots and we find the relationships between numerical columns using Scatter plots, hexbin plots, pair plots, and also here we analyse the patterns between categorical and numerical columns using swarmplot, boxplot, barplot.
- In this project By using these plots we analyse the relationships between employees salary, Graduationyear, Designation, 12percentage, 10percentage, 12graduation, 10board like this we find each and every employees background. Just like background verification.


thank
you

INTRODUCTION

The dataset was released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO). The study is primarily limited only to students with engineering disciplines. The dataset contains the employment outcomes of engineering graduates as dependent variables (Salary, Job Titles, and Job Locations) along with the standardized scores from three different areas – cognitive skills, technical skills and personality skills. The dataset also contains demographic features. The dataset contains around 40 independent variables and 4000 data points. The independent variables are both continuous and categorical in nature. The dataset contains a unique identifier for each candidate.

Import the data and display the head, shape and description of the data.

In [1]:
```
import pandas as pd
import numpy as np
```

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

In [3]:
```
df
```

Out[3]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	...	ComputerScience	Me
0	train	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	...	-1	
1	train	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.40	...	-1	
2	train	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.00	...	-1	
3	train	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	...	-1	
4	train	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.00	...	-1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
3993	train	47916	280000.0	10/1/11 0:00	10/1/12 0:00	software engineer	New Delhi	m	4/15/87 0:00	52.09	...	-1	
3994	train	752781	100000.0	7/1/13 0:00	7/1/13 0:00	technical writer	Hyderabad	f	8/27/92 0:00	90.00	...	-1	
3995	train	355888	320000.0	7/1/13 0:00	present	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	...	-1	
3996	train	947111	200000.0	7/1/14 0:00	1/1/15 0:00	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	...	438	
3997	train	324966	400000.0	2/1/13 0:00	present	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	...	-1	

3998 rows × 39 columns

In [4]:
```
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
```

In [5]:
```
df.head()
```

Out[5]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	...	ComputerScience	MechanicalEng
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	train	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	...	-1	
2	train	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	...	-1	
3	train	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	...	-1	
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	

5 rows × 39 columns

In [6]:
```
df.shape
```

Out[6]:
```
(3998, 39)
```

In [7]:
```
description = df.describe()
```

In [8]:
```
description
```

Out[8]:

	ID	Salary	10percentage	12graduation	12percentage	CollegeID	CollegeTier	collegeGPA	CollegeCityID	Colleg
count	3.998000e+03	3.998000e+03	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	399
mean	6.637945e+05	3.076998e+05	77.925443	2008.087544	74.466366	5156.851426	1.925713	71.486171	5156.851426	
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933	4802.261482	0.262270	8.167338	4802.261482	
min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000	2.000000	1.000000	6.450000	2.000000	
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000	494.000000	2.000000	66.407500	494.000000	
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000	2.000000	71.720000	3879.000000	
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000	2.000000	76.327500	8818.000000	
max	1.298275e+06	4.000000e+06	97.760000	2013.000000	98.700000	18409.000000	2.000000	99.930000	18409.000000	

8 rows × 27 columns

In [9]:
```
df.columns = df.columns.str.strip()
```

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

Out[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')
```

Data Cleaning

Dropping unwanted columns

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

In [12]:
```
df
```

Out[12]:

	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	10board	...	ComputerScience	M
0	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	board ofsecondary education,ap	...	-1	
1	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.40	cbse	...	-1	
2	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.00	cbse	...	-1	
3	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	cbse	...	-1	
4	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.00	cbse	...	-1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
3993	47916	280000.0	10/1/11 0:00	10/1/12 0:00	software engineer	New Delhi	m	4/15/87 0:00	52.09	cbse	...	-1	
3994	752781	100000.0	7/1/13 0:00	7/1/13 0:00	technical writer	Hyderabad	f	8/27/92 0:00	90.00	state board	...	-1	
3995	355888	320000.0	7/1/13 0:00	present	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	bse,odisha	...	-1	
3996	947111	200000.0	7/1/14 0:00	1/1/15 0:00	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	state board	...	438	
3997	324966	400000.0	2/1/13 0:00	present	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	cbse	...	-1	

3998 rows × 38 columns

In [13]:
```
df.drop(['DOJ', 'DOL', 'CollegeCityID', 'CollegeCityTier',
       'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon',
       'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg',
       'CivilEngg'], axis=1,inplace = True)
```

In [14]:
```
df
```

Out[14]:

	ID	Salary	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	...	CollegeS
0	203097	420000.0	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	board ofsecondary education,ap	2007	95.80	...	An Prac
1	579905	500000.0	assistant manager	Indore	m	10/4/89 0:00	85.40	cbse	2007	85.00	...	Mac Prac
2	810601	325000.0	systems engineer	Chennai	f	8/3/92 0:00	85.00	cbse	2010	68.20	...	U Prac
3	267447	1100000.0	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	cbse	2007	83.60	...	D
4	343523	200000.0	get	Manesar	m	2/27/91 0:00	78.00	cbse	2008	76.80	...	U Prac
...	...	...	...	...	...	...	...	...	...	...	...	
3993	47916	280000.0	software engineer	New Delhi	m	4/15/87 0:00	52.09	cbse	2006	55.50	...	Hary
3994	752781	100000.0	technical writer	Hyderabad	f	8/27/92 0:00	90.00	state board	2009	93.00	...	Telang
3995	355888	320000.0	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	bse,odisha	2008	65.50	...	Or
3996	947111	200000.0	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	state board	2010	69.88	...	Karna
3997	324966	400000.0	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	cbse	2008	68.00	...	Tamil N

3998 rows × 26 columns

In [15]:
```
df.isnull().sum()
```

Out[15]:
```
ID                      0
Salary                  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
CollegeState            0
GraduationYear          0
English                 0
Logical                 0
Quant                   0
conscientiousness       0
agreeableness           0
extraversion            0
nueroticism             0
openess_to_experience   0
dtype: int64
```

In [16]:
```
df.duplicated().sum()
```

Out[16]:
```
0
```

In [17]:
```
df.columns
```

Out[17]:
```
Index(['ID', 'Salary', 'Designation', 'JobCity', 'Gender', 'DOB',
       '10percentage', '10board', '12graduation', '12percentage', '12board',
       'CollegeID', 'CollegeTier', 'Degree', 'Specialization', 'collegeGPA',
       'CollegeState', 'GraduationYear', 'English', 'Logical', 'Quant',
       'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism',
       'openess_to_experience'],
      dtype='object')
```

In [18]:
```
df["Specialization"].value_counts()
```

Out[18]:
```
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
```

In [19]:
```
df["10board"].value_counts()
```

Out[19]:
```
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
```

In [20]:
```
df["Salary"].value_counts()
```

Out[20]:
```
300000.0     293
180000.0     239
200000.0     205
325000.0     188
120000.0     165
            ...
2050000.0      1
144000.0       1
1320000.0      1
755000.0       1
925000.0       1
Name: Salary, Length: 177, dtype: int64
```

In [21]:
```
df["Salary"].unique()
```

Out[21]:
```
array([ 420000.,  500000.,  325000., 1100000.,  200000.,  300000.,
        400000.,  600000.,  230000.,  450000.,  270000.,  350000.,
        250000.,  120000.,  320000.,  190000.,  180000.,  335000.,
        435000.,  345000.,  145000.,  220000.,  530000.,  340000.,
        360000.,  215000.,   80000.,  330000.,  380000.,  110000.,
        205000.,   95000.,  390000.,   60000.,  240000.,  525000.,
        305000.,  150000.,  310000.,  455000.,  800000.,  100000.,
        280000.,  445000.,  315000.,  370000.,  275000., 1500000.,
        425000.,  470000.,  460000.,  510000.,  480000.,  170000.,
        640000.,  225000.,  440000., 1200000.,  675000.,  105000.,
        195000.,  385000.,  235000.,  615000.,  290000.,  140000.,
        405000., 1860000.,  375000.,  430000.,  660000.,   70000.,
        410000.,  550000.,   35000.,  115000.,  415000.,  265000.,
        285000.,  245000.,  395000.,  560000.,  700000.,  185000.,
        160000.,  625000.,   85000.,  135000.,  785000.,  210000.,
        155000.,  355000.,  535000.,  690000.,  260000., 1110000.,
       1000000.,  505000.,  475000.,  715000.,  820000.,   90000.,
        720000., 2600000.,  515000.,   55000.,  495000.,   65000.,
        655000.,  545000.,  520000.,  645000., 1025000.,  775000.,
        490000., 1300000., 3500000.,  910000.,  570000.,  255000.,
        130000.,  175000.,  730000.,  555000.,  465000.,  680000.,
        165000.,  630000.,  365000., 1050000., 2000000.,  860000.,
        125000.,   50000.,  580000.,  485000., 4000000., 2020000.,
        650000.,   45000.,  610000.,  760000.,  585000.,  620000.,
        870000., 2050000.,  540000.,  144000.,  605000., 1320000.,
        755000.,  880000., 3000000.,   75000.,  295000.,   40000.,
        575000.,  565000., 2500000., 2300000.,  590000.,  950000.,
       1800000.,  725000.,  930000.,  750000.,  705000., 1745000.,
        850000.,  845000.,  670000., 1030000.,  770000.,  900000.,
       1210000.,  810000.,  925000.])
```

In [22]:
```
## Identify Missing Values

missing_values = df.isnull().sum()
print("Missing Values:\n", missing_values)
```

```
Missing Values:
 ID                       0
Salary                   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
CollegeState             0
GraduationYear           0
English                  0
Logical                  0
Quant                    0
conscientiousness        0
agreeableness            0
extraversion             0
nueroticism              0
openess_to_experience    0
dtype: int64
```

In [23]:
```
## Handling Outliers
# Identify and handle outliers using,Z-score method

from scipy.stats import zscore

z_scores = np.abs(zscore(df.select_dtypes(include=np.number)))
df_no_outliers = df[(z_scores < 3).all(axis=1)]
```

In [24]:
```
z_scores
```

Out[24]:

	ID	Salary	10percentage	12graduation	12percentage	CollegeID	CollegeTier	collegeGPA	GraduationYear	English	Logical
0	1.268535	0.527947	0.647233	0.657765	1.939676	0.836346	0.283282	0.797646	0.034716	0.127240	0.961148
1	0.230991	0.904045	0.758921	0.657765	0.957729	0.135401	0.283282	0.174641	0.003322	1.842720	1.249258
2	0.404233	0.081332	0.718307	1.156686	0.569744	1.060644	0.283282	0.181988	0.059466	1.080285	0.500173
3	1.091347	3.724775	0.779227	0.657765	0.830439	0.367196	3.530054	0.386200	0.034716	1.270894	0.961148
4	0.881871	0.506320	0.007570	0.052948	0.212176	1.293542	0.283282	0.295584	0.003322	0.413154	1.422124
...	...	...	...	...	...	...	...	...	...	...	...
3993	1.695828	0.130223	2.623172	1.262582	1.724441	0.231409	0.283282	1.222849	0.066109	1.302326	1.931475
3994	0.245025	0.976442	1.225977	0.551869	1.685097	0.057033	0.283282	0.711928	0.028072	0.825804	1.055621
3995	0.847824	0.057826	0.399491	0.052948	0.815231	0.964072	0.283282	0.181988	0.003322	0.253978	0.306535
3996	0.780115	0.506320	0.080674	1.156686	0.416997	0.870085	0.283282	0.130557	0.059466	0.492239	1.055621
3997	0.932968	0.433923	0.743781	0.052948	0.587929	0.302426	0.283282	0.426896	0.003322	0.603763	0.154441

3998 rows × 17 columns

In [25]:
```
## handling missing values

df = df.dropna()
```

In [26]:
```
df
```

Out[26]:

	ID	Salary	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	...	CollegeS
0	203097	420000.0	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	board ofsecondary education,ap	2007	95.80	...	An Prac
1	579905	500000.0	assistant manager	Indore	m	10/4/89 0:00	85.40	cbse	2007	85.00	...	Mac Prac
2	810601	325000.0	systems engineer	Chennai	f	8/3/92 0:00	85.00	cbse	2010	68.20	...	U Prac
3	267447	1100000.0	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	cbse	2007	83.60	...	D
4	343523	200000.0	get	Manesar	m	2/27/91 0:00	78.00	cbse	2008	76.80	...	U Prac
...	...	...	...	...	...	...	...	...	...	...	...	
3993	47916	280000.0	software engineer	New Delhi	m	4/15/87 0:00	52.09	cbse	2006	55.50	...	Hary
3994	752781	100000.0	technical writer	Hyderabad	f	8/27/92 0:00	90.00	state board	2009	93.00	...	Telang
3995	355888	320000.0	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	bse,odisha	2008	65.50	...	Or
3996	947111	200000.0	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	state board	2010	69.88	...	Karna
3997	324966	400000.0	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	cbse	2008	68.00	...	Tamil N

3998 rows × 26 columns

In [27]:
```
## Handling Duplicate Data

df = df.drop_duplicates()
df
```

Out[27]:

	ID	Salary	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	...	CollegeS
0	203097	420000.0	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	board ofsecondary education,ap	2007	95.80	...	An Prac
1	579905	500000.0	assistant manager	Indore	m	10/4/89 0:00	85.40	cbse	2007	85.00	...	Mac Prac
2	810601	325000.0	systems engineer	Chennai	f	8/3/92 0:00	85.00	cbse	2010	68.20	...	U Prac
3	267447	1100000.0	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	cbse	2007	83.60	...	D
4	343523	200000.0	get	Manesar	m	2/27/91 0:00	78.00	cbse	2008	76.80	...	U Prac
...	...	...	...	...	...	...	...	...	...	...	...	
3993	47916	280000.0	software engineer	New Delhi	m	4/15/87 0:00	52.09	cbse	2006	55.50	...	Har
3994	752781	100000.0	technical writer	Hyderabad	f	8/27/92 0:00	90.00	state board	2009	93.00	...	Telang
3995	355888	320000.0	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	bse,odisha	2008	65.50	...	Or
3996	947111	200000.0	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	state board	2010	69.88	...	Karna
3997	324966	400000.0	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	cbse	2008	68.00	...	Tamil N

3998 rows × 26 columns

In [28]:
```
## Standardize/Normalize Data

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
df[['Salary']] = scaler.fit_transform(df[['Salary']])
```

In [29]:
```
scaler
```

Out[29]:
```
MinMaxScaler()
```

In [30]:
```
## Address Inconsistent or Erroneous Data
# Manually correct errors or inconsistencies

df['12graduation'].replace({'erroneous_value': 'correct_value'}, inplace=True)
```

Data Visualization

In [31]:
```
df
```

Out[31]:

	ID	Salary	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	...	CollegeSt
0	203097	0.097100	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	board ofsecondary education,ap	2007	95.80	...	Andl Prade
1	579905	0.117276	assistant manager	Indore	m	10/4/89 0:00	85.40	cbse	2007	85.00	...	Madh Prade
2	810601	0.073140	systems engineer	Chennai	f	8/3/92 0:00	85.00	cbse	2010	68.20	...	U Prade
3	267447	0.268600	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	cbse	2007	83.60	...	De
4	343523	0.041614	get	Manesar	m	2/27/91 0:00	78.00	cbse	2008	76.80	...	U Prade
...	...	...	...	...	...	...	...	...	...	...	...	
3993	47916	0.061791	software engineer	New Delhi	m	4/15/87 0:00	52.09	cbse	2006	55.50	...	Harya
3994	752781	0.016393	technical writer	Hyderabad	f	8/27/92 0:00	90.00	state board	2009	93.00	...	Telanga
3995	355888	0.071879	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	bse,odisha	2008	65.50	...	Oris
3996	947111	0.041614	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	state board	2010	69.88	...	Karnata
3997	324966	0.092055	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	cbse	2008	68.00	...	Tamil Na

3998 rows × 26 columns

In [32]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3998 entries, 0 to 3997
Data columns (total 26 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   ID                     3998 non-null   int64  
 1   Salary                 3998 non-null   float64
 2   Designation            3998 non-null   object 
 3   JobCity                3998 non-null   object 
 4   Gender                 3998 non-null   object 
 5   DOB                    3998 non-null   object 
 6   10percentage           3998 non-null   float64
 7   10board                3998 non-null   object 
 8   12graduation           3998 non-null   int64  
 9   12percentage           3998 non-null   float64
 10  12board                3998 non-null   object 
 11  CollegeID              3998 non-null   int64  
 12  CollegeTier            3998 non-null   int64  
 13  Degree                 3998 non-null   object 
 14  Specialization         3998 non-null   object 
 15  collegeGPA             3998 non-null   float64
 16  CollegeState           3998 non-null   object 
 17  GraduationYear         3998 non-null   int64  
 18  English                3998 non-null   int64  
 19  Logical                3998 non-null   int64  
 20  Quant                  3998 non-null   int64  
 21  conscientiousness      3998 non-null   float64
 22  agreeableness          3998 non-null   float64
 23  extraversion           3998 non-null   float64
 24  nueroticism            3998 non-null   float64
 25  openess_to_experience  3998 non-null   float64
dtypes: float64(9), int64(8), object(9)
memory usage: 843.3+ KB
```

In [33]:
```
df.head()
```

Out[33]:

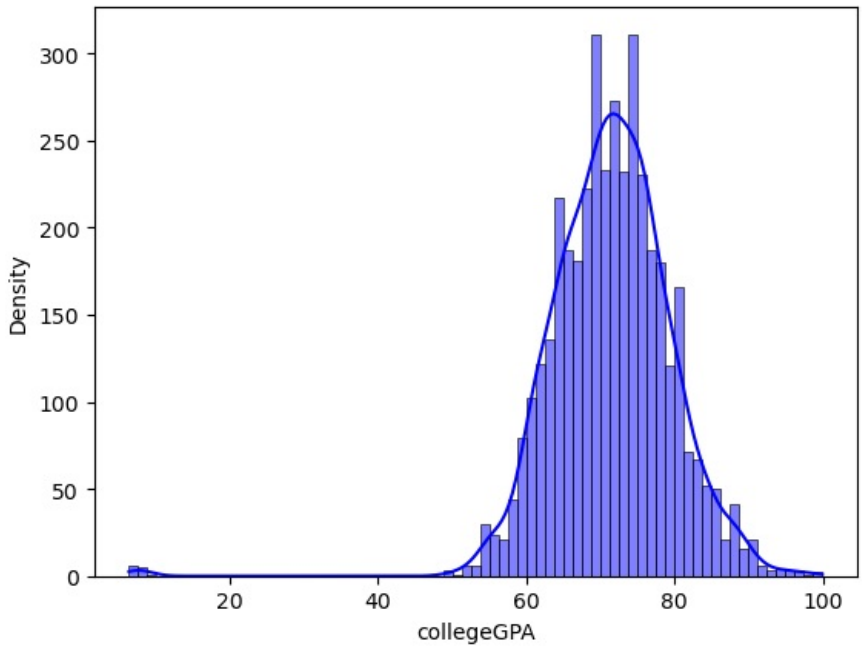
	ID	Salary	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	...	CollegeState	Gradu
0	203097	0.097100	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	board ofsecondary education,ap	2007	95.8	...	Andhra Pradesh	
1	579905	0.117276	assistant manager	Indore	m	10/4/89 0:00	85.4	cbse	2007	85.0	...	Madhya Pradesh	
2	810601	0.073140	systems engineer	Chennai	f	8/3/92 0:00	85.0	cbse	2010	68.2	...	Uttar Pradesh	
3	267447	0.268600	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	cbse	2007	83.6	...	Delhi	
4	343523	0.041614	get	Manesar	m	2/27/91 0:00	78.0	cbse	2008	76.8	...	Uttar Pradesh	

5 rows × 26 columns

Univariate Analysis of Numerical and categorical columns

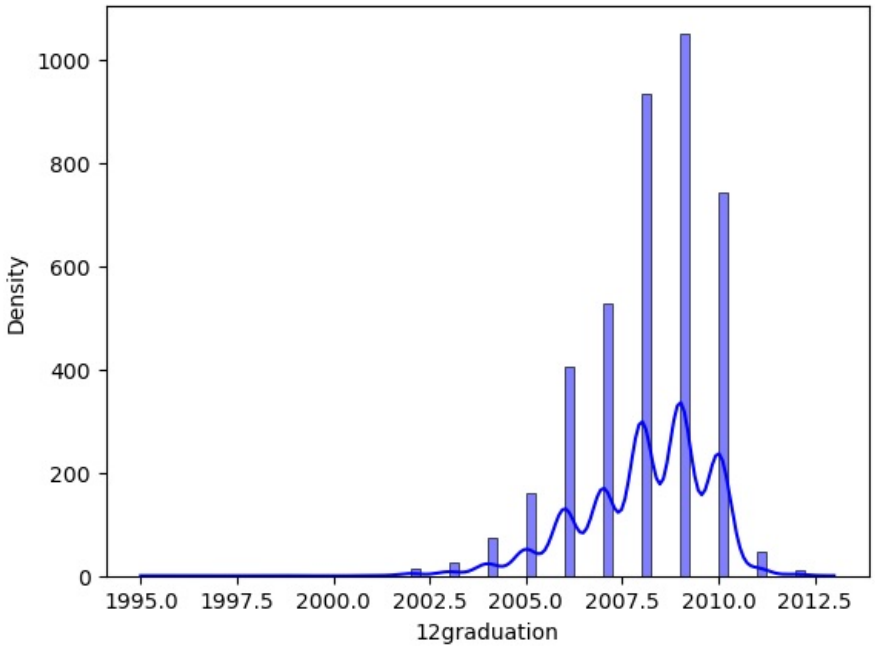
Probability Density Function (PDF) plot

In [34]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.histplot(df['collegeGPA'], kde=True, color = 'blue')
plt.xlabel('collegeGPA')
plt.ylabel('Density')
plt.show()
```



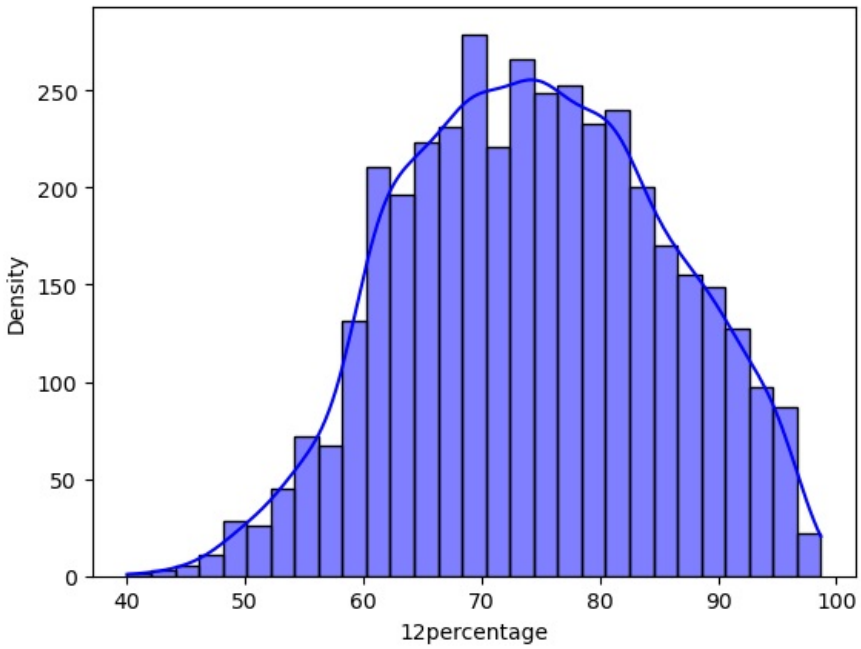
This Probability density function(PDF) decribes the density of collegeGPA of all employees. Most of the employees are between 65 to 75.

In [35]:
```
sns.histplot(df['12graduation'], kde=True, color = 'blue')
plt.xlabel('12graduation')
plt.ylabel('Density')
plt.show()
```



This Probability density function(PDF) describes the density of 12graduation of all employees. most of the employees are passedout in the year 2009.

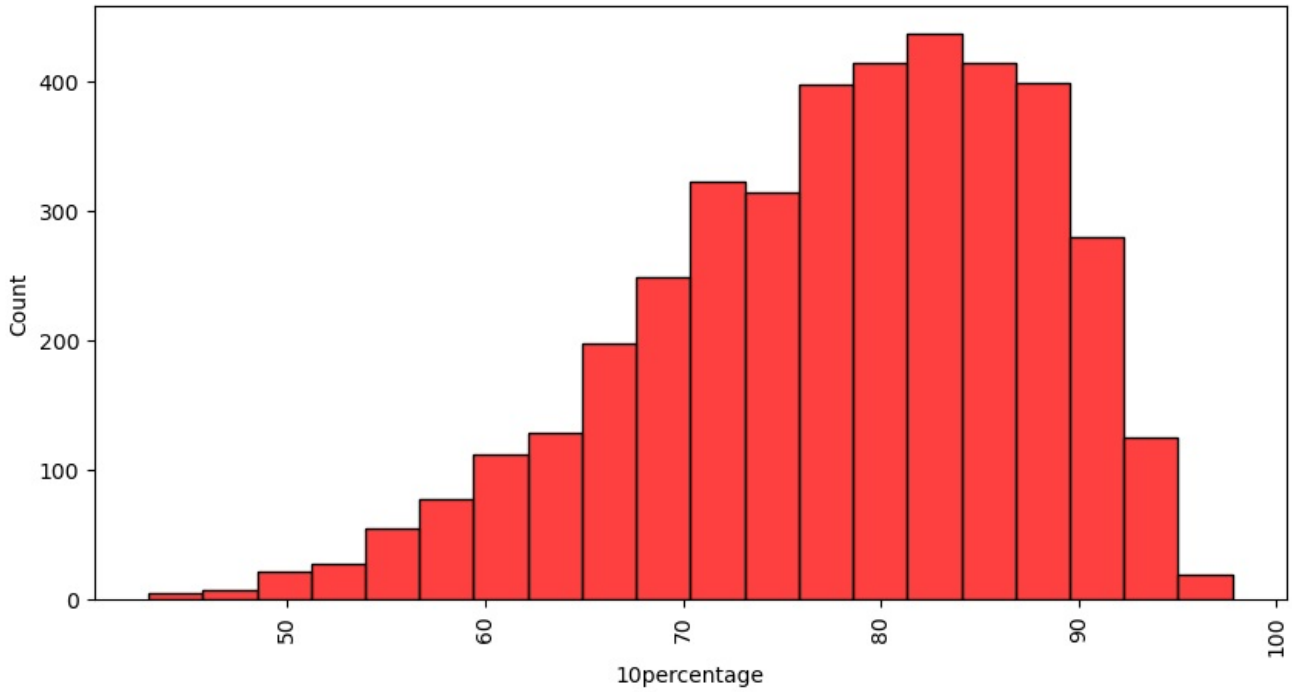
In [36]:
```
sns.histplot(df['12percentage'], kde=True, color = 'blue')
plt.xlabel('12percentage')
plt.ylabel('Density')
plt.show()
```



This Probability density function(PDF) describes the density of 12percentage of all employees and the most of the employees are passed in between 70 to 73 percentage.

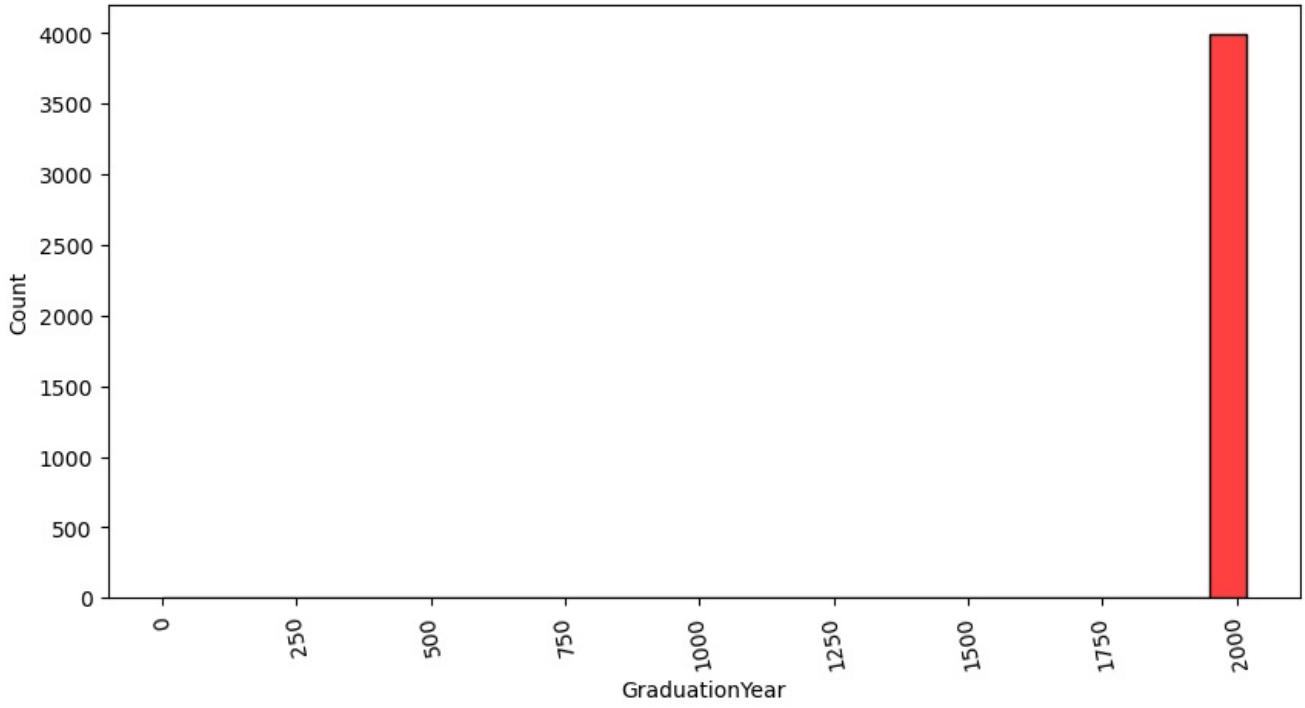
Histogram

In [37]:
```
fig, ax = plt.subplots(figsize=(10, 5))
plt.xticks(rotation=90)
sns.histplot(data=df, x="10percentage", color="r", bins=20, ax=ax)
plt.show()
```



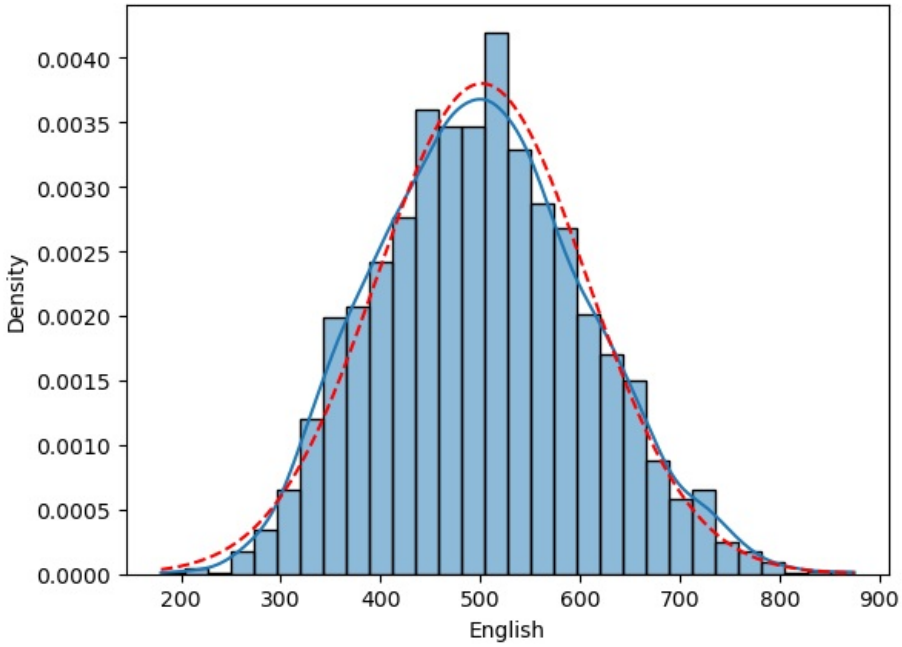
This Histogram plot describes the density of 10percentage of all employees and the most of the employees are passed in approximately 83percentage.

In [38]:
```
fig, ax = plt.subplots(figsize=(10, 5))
plt.xticks(rotation=100)
sns.histplot(data=df, x="GraduationYear", color="r", bins=30, ax=ax)
plt.show()
```



Here the Histogram describes the GraduationYear of employees. so, the most of the emplyees are completed their graduation in the year of 2000.

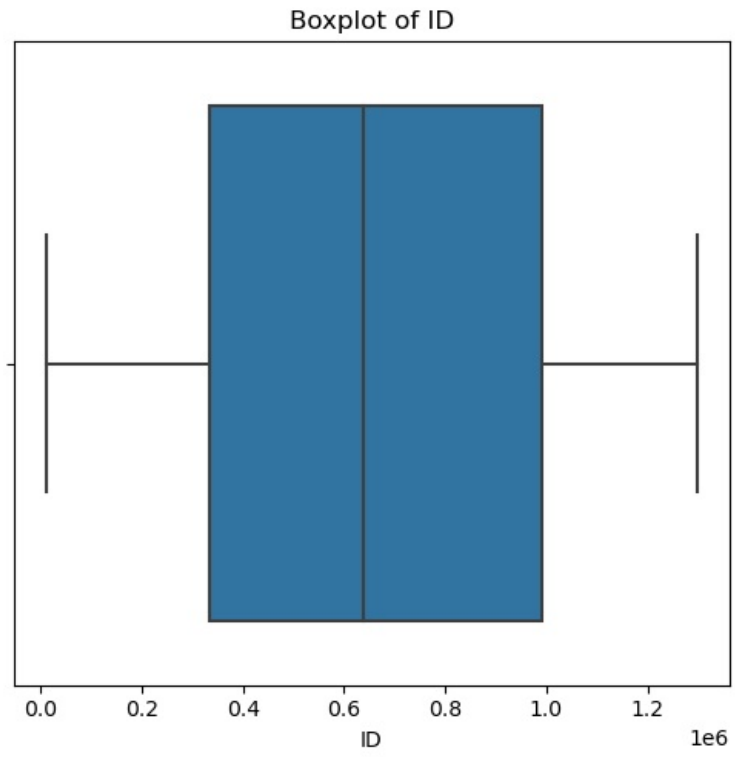
In [39]:
```
import scipy.stats as stats

sns.histplot(data=df['English'], stat='density', bins=30, kde=True)
x = np.linspace(df['English'].min(), df['English'].max(), 1000)
plt.plot(x, stats.norm.pdf(x, df['English'].mean(), df['English'].std()), color='red', linestyle='dashed')
plt.show()
```

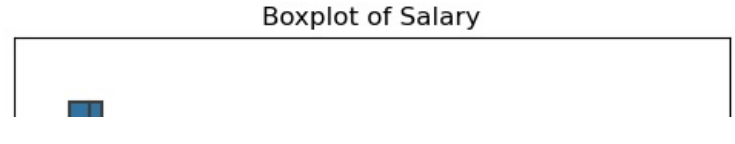


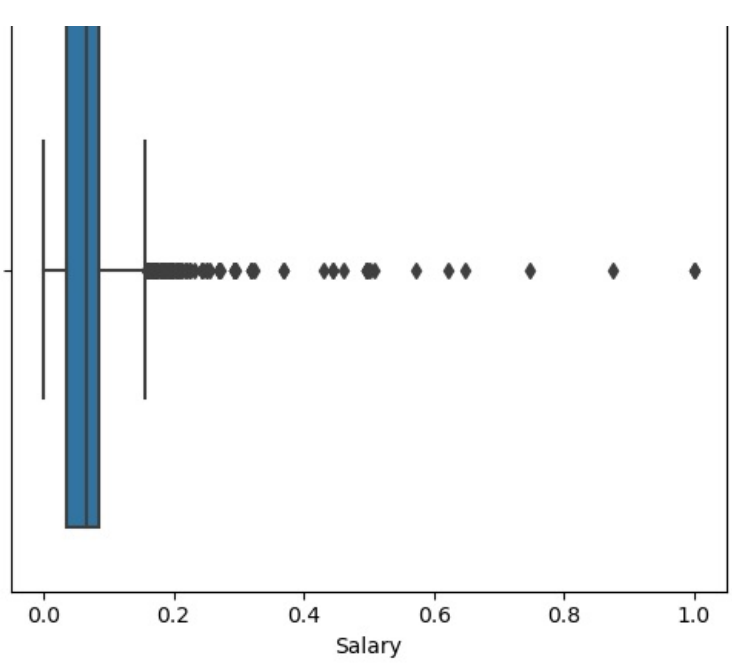
Here the plot describes the density of English. Most of the employees has given preference to english from 440 to 520

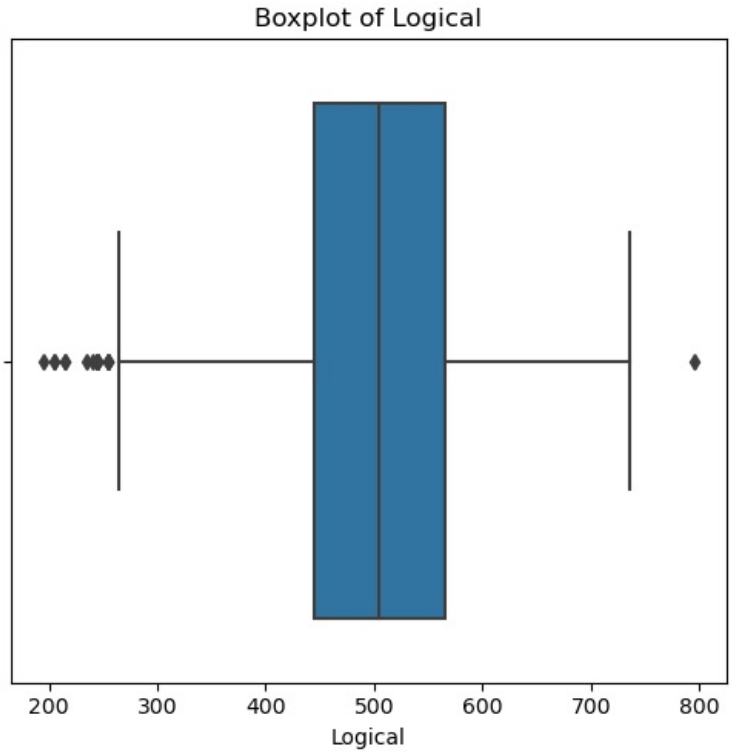
Boxplot

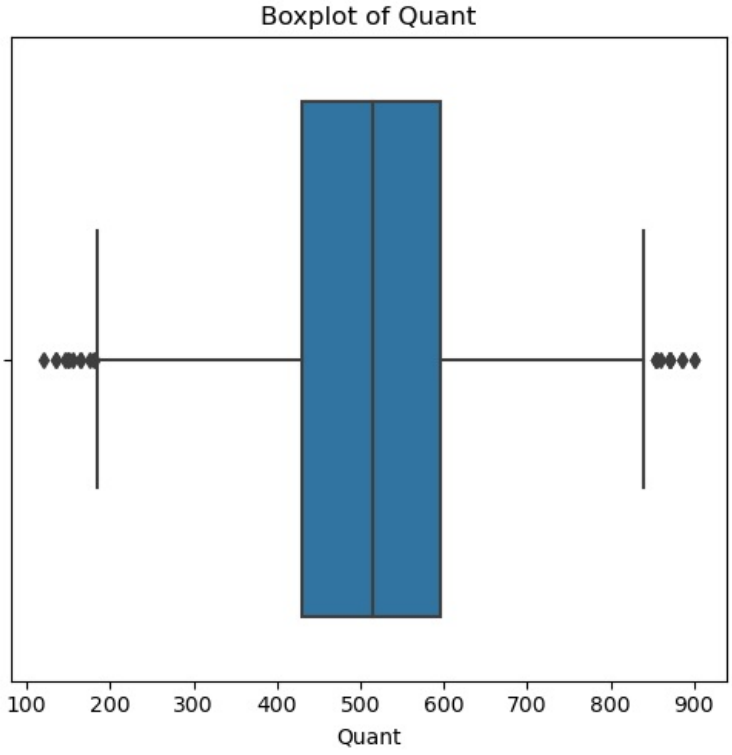
In [40]:
```
import seaborn as sns
import matplotlib.pyplot as plt

selected_columns = ['ID','Salary','Logical', 'Quant',
       'conscientiousness']

fig, axes = plt.subplots(nrows=len(selected_columns), ncols=1, figsize=(5, 5 * len(selected_columns)))

for i, column in enumerate(selected_columns):
    ax = axes[i]
    sns.boxplot(x=df[column], ax=ax)
    ax.set_title(f'Boxplot of {column}')

plt.tight_layout()

plt.show()
```



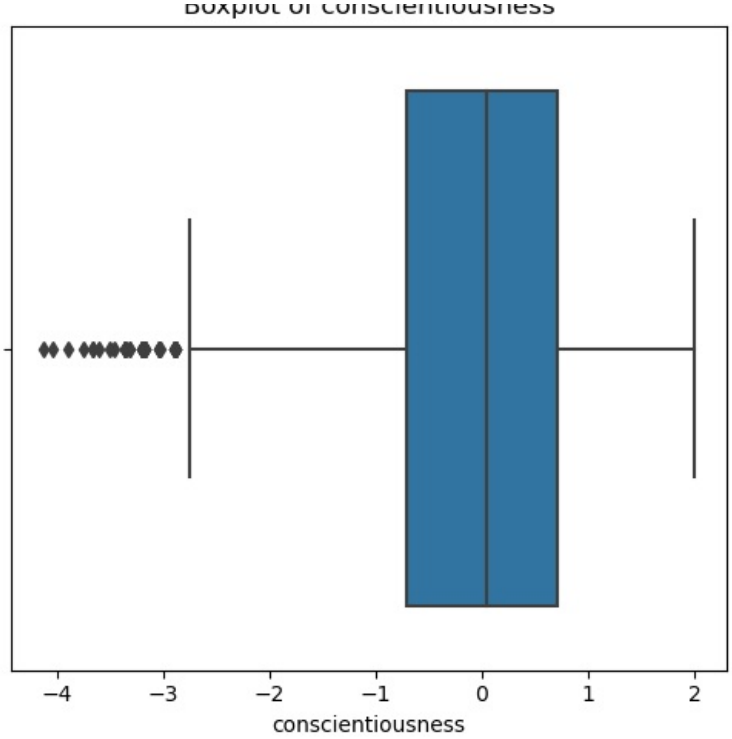








Boxplot of conscientiousness



Observations of outliers:

Here the plots are describing the outliers of variables. After visualizing the data using a boxplot for the "ID", there are no outliers present in the distribution. The boxplot displays a symmetric and compact box, indicating that the majority of the data points fall within the interquartile range (IQR) with no extreme values or influential outliers. The absence of outliers enhances the reliability of measures of central tendency, such as the median, providing a more accurate representation of the typical value for "ID".

In the boxplot analysis of the "Salary" it shows the presence of right-side outliers, the potential extreme values in the upper range of the data. The distribution of the data is positively skewed, it has a long right tail, and the outliers are pulling the mean and the right end of the boxplot upward.

The boxplot analysis of the"Logical" variable reveals an asymmetric distribution with a notable difference in the number of outliers on the left and right sides.The "Logical" variable exhibits a median value of 505, an interquartile range (IQR) of 120, and a spread of values within 195 min, 795max.

The boxplot analysis of the "Quant" reveals a symmetric distribution with outliers present on both the right and left sides.The "Quant" variable exhibits a median value of 515, an interquartile range (IQR) of 165, and a spread of values within 120 min, 900 max.The boxplot for "Quant" displays a central box representing the interquartile range, with a median line dividing the box. Notably, outliers are observed on both the right and left sides, indicating a balanced skewness in both directions.

In the boxplot analysis of the "conscientiousness" it shows the presence of left-side outliers, the potential extreme values in the lower range of the data. The distribution of the data is negatively skewed, it has a long left tail, and the outliers are pulling the mean and the left end of the boxplot downward.

In [41]:
```
column_median = df['Logical'].median()

print(f"The median of 'Logical' is: {column_median}")
```

```
The median of 'Logical' is: 505.0
```

In [42]:
```
column_median = df['Quant'].median()

print(f"The median of 'Quant' is: {column_median}")
```

```
The median of 'Quant' is: 515.0
```

In [43]:
```
column_min = df['Logical'].min()
column_max = df['Logical'].max()

print(f"The minimum value of 'Logical' is: {column_min}")
```

```
print(f"The maximum value of 'Logical' is: {column_max}")
```

```
The minimum value of 'Logical' is: 195
The maximum value of 'Logical' is: 795
```

In [44]:
```
column_min = df['Quant'].min()
column_max = df['Quant'].max()

print(f"The minimum value of 'Quant' is: {column_min}")
print(f"The maximum value of 'Quant' is: {column_max}")
```

```
The minimum value of 'Quant' is: 120
The maximum value of 'Quant' is: 900
```

In [45]:
```
column_data = df['Logical']
Q1 = column_data.quantile(0.25)
Q3 = column_data.quantile(0.75)
column_iqr = Q3 - Q1

print(f"The Interquartile Range (IQR) of 'Logical' is: {column_iqr}")
```

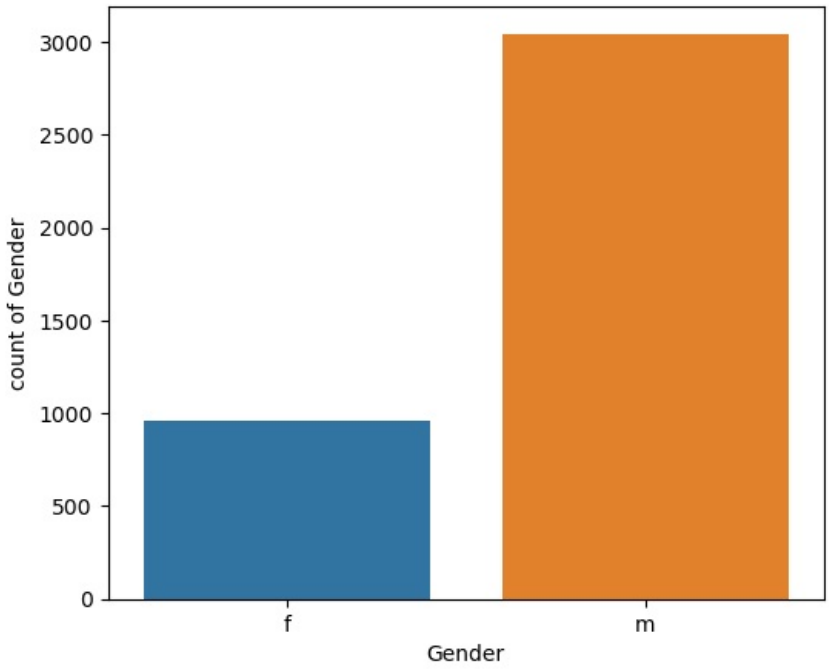
```
The Interquartile Range (IQR) of 'Logical' is: 120.0
```

In [46]:
```
column_data = df['Quant']
Q1 = column_data.quantile(0.25)
Q3 = column_data.quantile(0.75)
column_iqr = Q3 - Q1

print(f"The Interquartile Range (IQR) of 'Quant' is: {column_iqr}")
```

```
The Interquartile Range (IQR) of 'Quant' is: 165.0
```

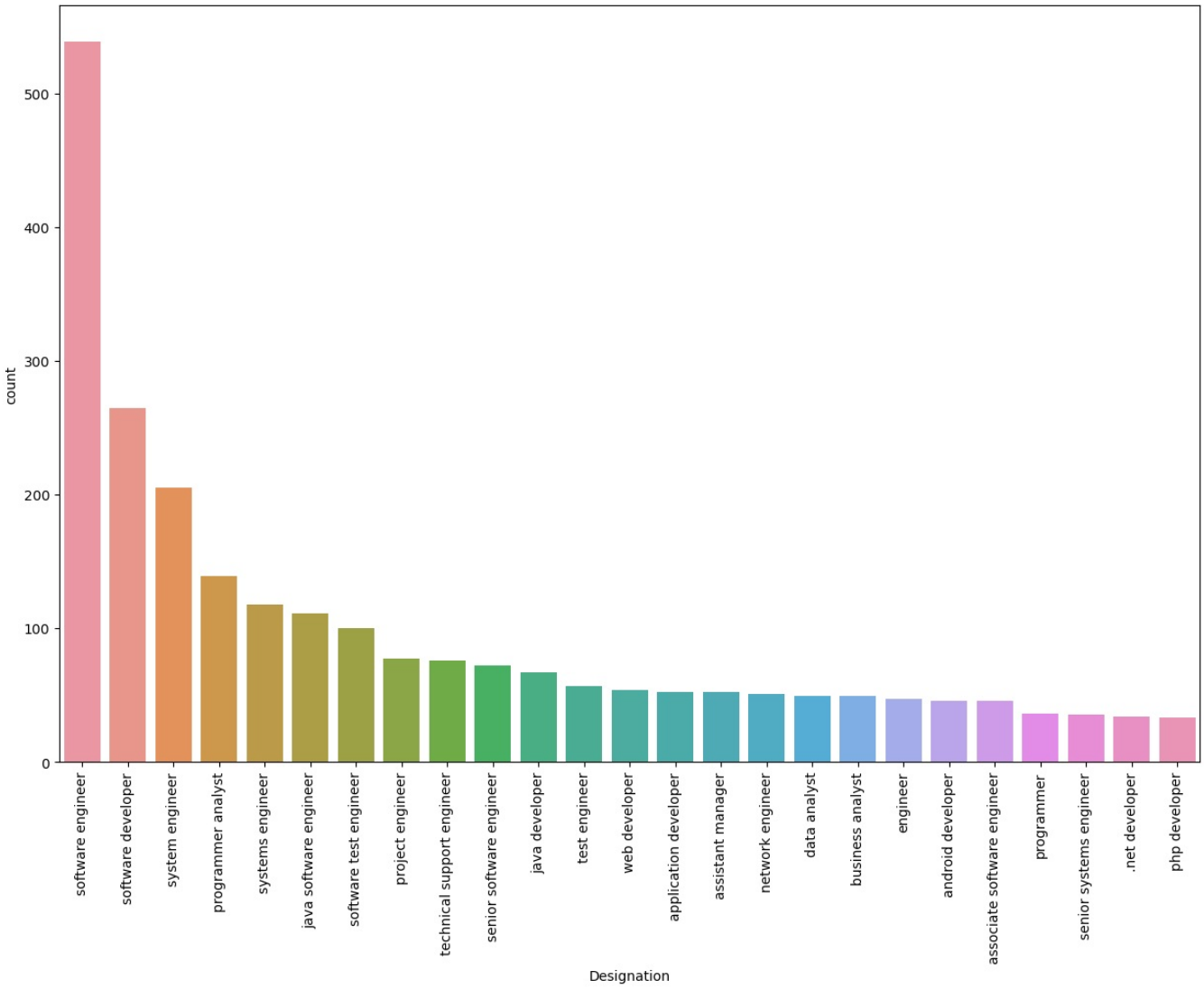
Countplot

In [47]:
```
import seaborn as sns

plt.figure(figsize = (6,5))

sns.countplot(x = 'Gender', data= df)
plt.ylabel('count of Gender')
plt.xlabel('Gender')

plt.xticks(rotation=0)
plt.show()
```



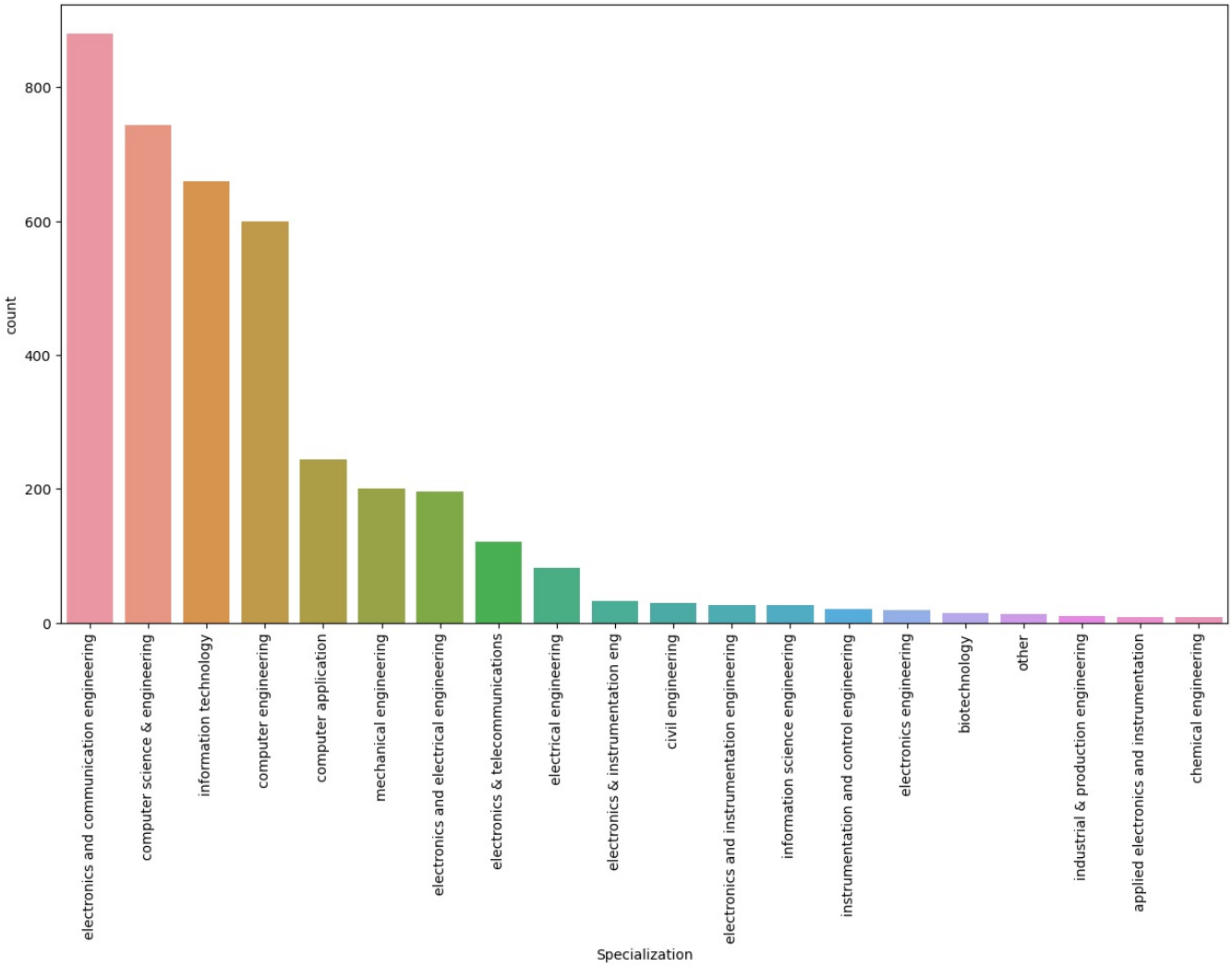
The countplot analysis of the "Gender"variable. Here the highest no.of employees are males and lowest is females.

In [48]:
```
plt.figure(figsize=(15, 10))
plt.xticks(rotation=90)
sns.countplot(data=df, x='Designation', order=df['Designation'].value_counts().index[:25])
plt.show()
```



In this plot analysis of "Designation". Here most of the employees Designation is software engineer and lowest no.of Designation is php developer

In [49]:
```
plt.figure(figsize=(15, 8))
plt.xticks(rotation=90)
sns.countplot(data=df, x='Specialization', order=df['Specialization'].value_counts().index[:20])
plt.show()
```



In this plot analysis of "Specialization". Here most of the employees "Specialization" is electronics and communication engineering and lowest no.of employees "Specialization" is chemical engineering.

Find the outliers in each numerical column

In [50]:
```
def detect_outliers(column):
    q1 = column.quantile(0.25)
    q3 = column.quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    outliers = (column < lower_bound) | (column > upper_bound)
    return outliers

outliers_dict = {}
for col in df.select_dtypes(include=np.number).columns:
    outliers_dict[col] = detect_outliers(df[col])

print("Outliers:")
print(outliers_dict)
```

```
Outliers:
{'ID': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: ID, Length: 3998, dtype: bool, 'Salary': 0       False
1       False
2       False
3        True
4       False
        ...
3993    False
```

```
3994    False
3995    False
3996    False
3997    False
Name: Salary, Length: 3998, dtype: bool, '10percentage': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: 10percentage, Length: 3998, dtype: bool, '12graduation': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: 12graduation, Length: 3998, dtype: bool, '12percentage': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: 12percentage, Length: 3998, dtype: bool, 'CollegeID': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: CollegeID, Length: 3998, dtype: bool, 'CollegeTier': 0       False
1       False
2       False
3        True
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: CollegeTier, Length: 3998, dtype: bool, 'collegeGPA': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: collegeGPA, Length: 3998, dtype: bool, 'GraduationYear': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: GraduationYear, Length: 3998, dtype: bool, 'English': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
```

```
3995    False
3996    False
3997    False
Name: English, Length: 3998, dtype: bool, 'Logical': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: Logical, Length: 3998, dtype: bool, 'Quant': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: Quant, Length: 3998, dtype: bool, 'conscientiousness': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: conscientiousness, Length: 3998, dtype: bool, 'agreeableness': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: agreeableness, Length: 3998, dtype: bool, 'extraversion': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: extraversion, Length: 3998, dtype: bool, 'nueroticism': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: nueroticism, Length: 3998, dtype: bool, 'openess_to_experience': 0       False
1       False
2       False
3       False
4       False
        ...
3993    False
3994    False
3995    False
3996    False
3997    False
Name: openess_to_experience, Length: 3998, dtype: bool}
```

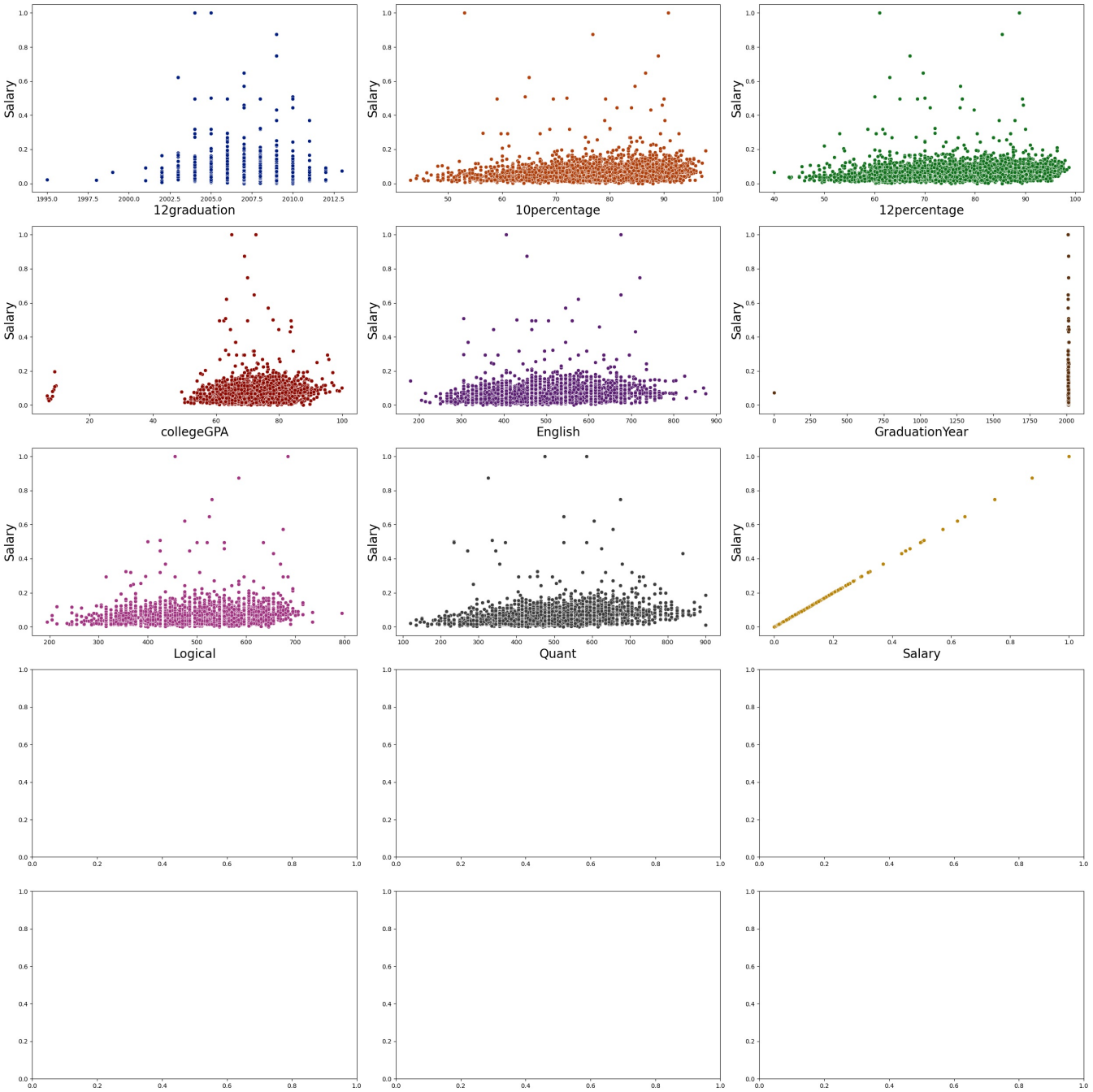
Bivariate Analysis

Scatter plot

relationships between numerical columns

Analysis of numerical columns

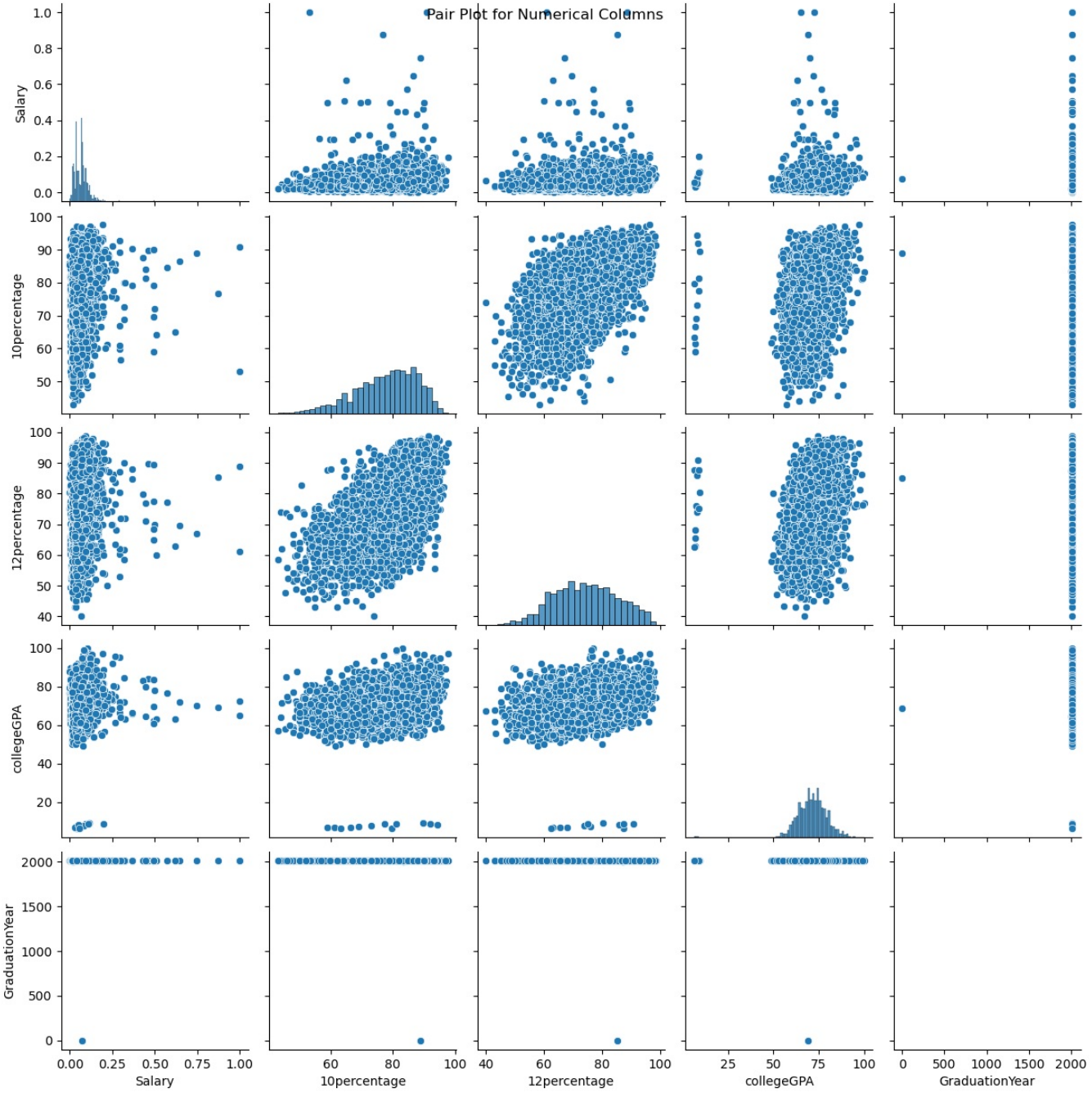
In [51]:
```
num_cols_3 = ['12graduation','10percentage','12percentage','collegeGPA','English',
       'GraduationYear','Logical', 'Quant','Salary' ]
num_rows = len(num_cols_3) // 2 + len(num_cols_3) % 2
num_cols = 3

fig, axes = plt.subplots(num_rows, num_cols, figsize=(25, 25))

if num_rows == 1:
    axes = [axes]

colors = sns.color_palette('dark', len(num_cols_3))

for i, col_name in enumerate(num_cols_3):
    row = i // num_cols
    col = i % num_cols
    sns.scatterplot(x=col_name, y='Salary', data=df, ax=axes[row][col], color=colors[i])
    axes[row][col].set_xlabel(col_name,fontsize=20)
    axes[row][col].set_ylabel('Salary',fontsize=20)


plt.tight_layout()
plt.show()
```



Pair plots of numerical columns

In [52]:
```
numerical_columns = ['Salary' '10percentage' '12percentage' 'collegeGPA' 'GraduationYear']
```

In [52]:
```
numerical_columns = ['Salary', '10percentage', '12percentage', 'collegeGPA', 'GraduationYear']
sns.pairplot(df, vars=numerical_columns)
plt.suptitle('Pair Plot for Numerical Columns')
plt.show()
```

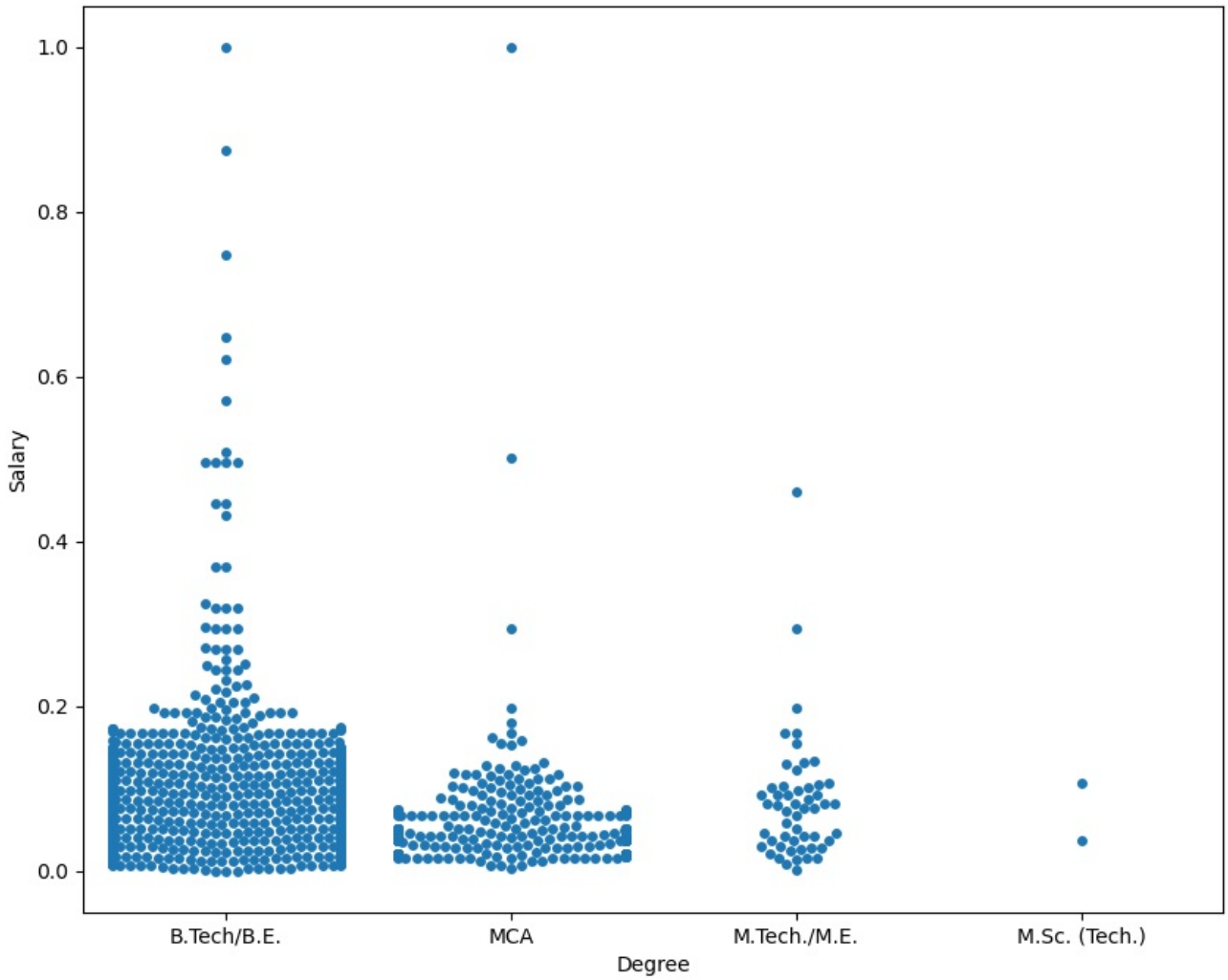


patterns between categorical and numerical columns

swarmplot

In [53]:
```
import warnings
warnings.filterwarnings('ignore')
fig = plt.subplots(figsize = (10,8))
sns.swarmplot(data = df, x= 'Degree', y = 'Salary')
```

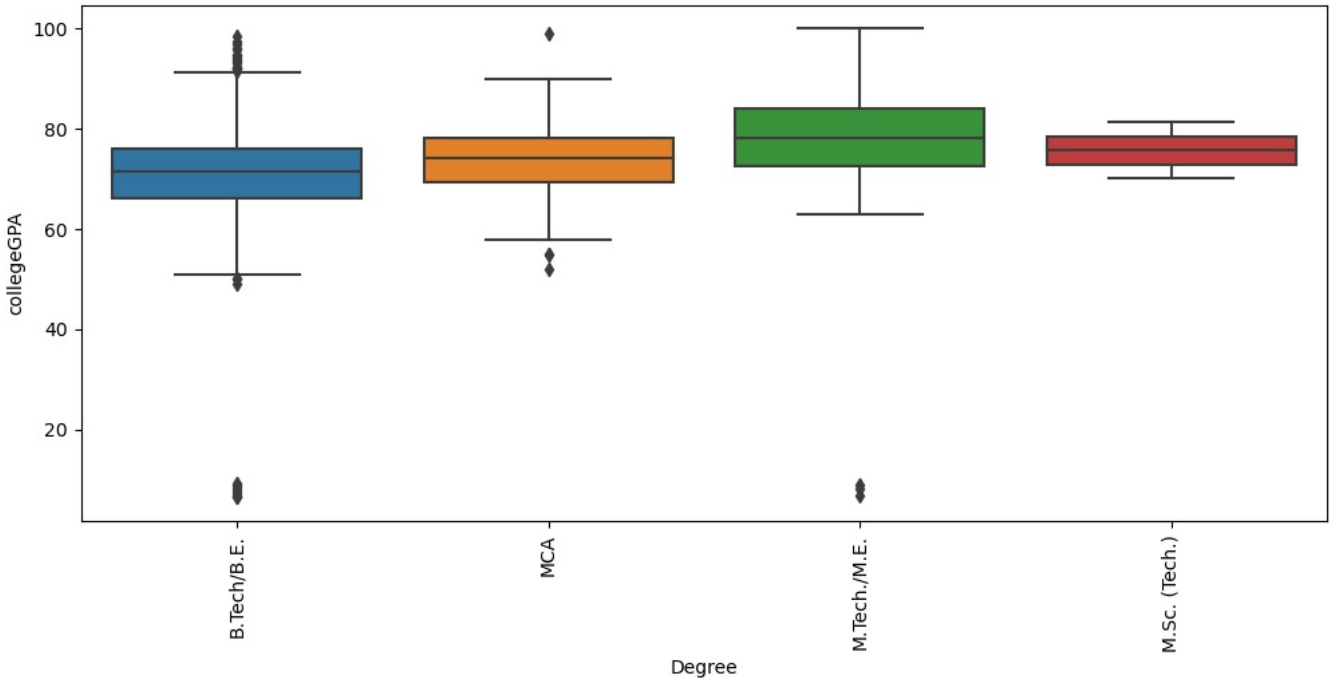
Out[53]:
```
<Axes: xlabel='Degree', ylabel='Salary'>
```



This swarmplot analysis the relationship between Degree and Salary. Here the b.Tech/B.E backround employees have highest salaries and M.sc.(Tech) employees have lowest salaries.

Boxplot

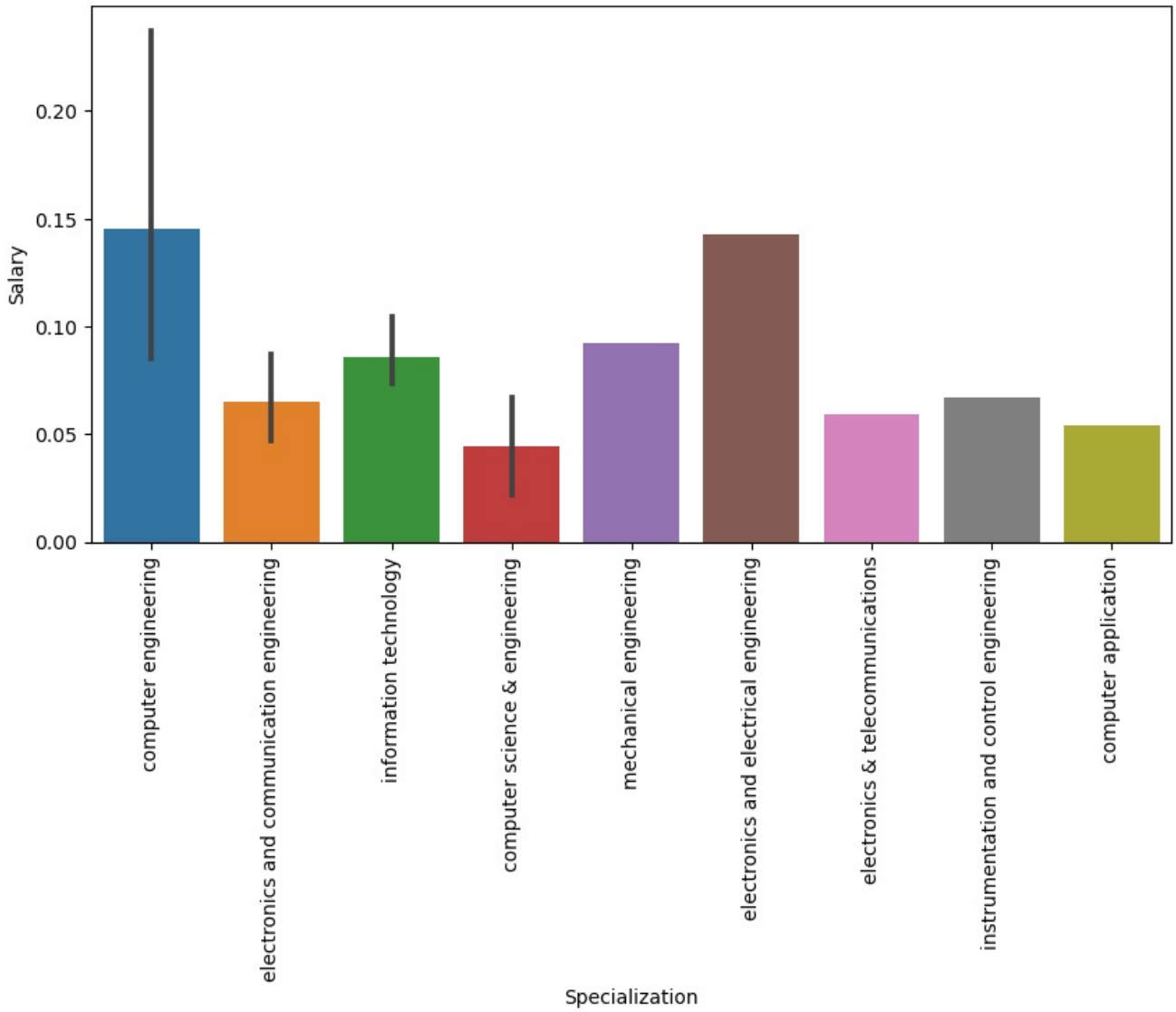
In [54]:
```
fig = plt.subplots(figsize = (12,5))
sns.boxplot(data = df, x= 'Degree', y = 'collegeGPA')
plt.xticks(rotation=90)
plt.show()
```



This Box plot analysis the relationship between Degree and collegeGPA variables. Here the B.Tech/B.E Degree have more outliers in upward direction and downward direction comparing with other Degrees. so,the B.tech/B.E background employees have heighest collegeGPA.

Barplot

In [55]:
```
plt.figure(figsize=(10, 5))
sns.barplot(data = df.head(20), x= 'Specialization', y = 'Salary')
plt.xticks(rotation=90)
plt.show()
```



This bar plot represent an estimate of central tendency between categorical and numerical variable with the height of each rectangle. This is the relationship between Specialization and Salary varialbles of Top 20 values.

Research Questions

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.”Test this claim with the data given to you.

In [56]:
```
df
```

Out[56]:

	ID	Salary	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	...	CollegeSt
0	203097	0.097100	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	board ofsecondary education,ap	2007	95.80	...	Andl Prade
1	579905	0.117276	assistant manager	Indore	m	10/4/89 0:00	85.40	cbse	2007	85.00	...	Madh Prade
2	810601	0.073140	systems engineer	Chennai	f	8/3/92 0:00	85.00	cbse	2010	68.20	...	U Prade
3	267447	0.268600	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	cbse	2007	83.60	...	De
4	343523	0.041614	get	Manesar	m	2/27/91 0:00	78.00	cbse	2008	76.80	...	U Prade
...	...	...	...	...	...	...	...	...	...	...	...	
3993	47916	0.061791	software engineer	New Delhi	m	4/15/87 0:00	52.09	cbse	2006	55.50	...	Harya
3994	752781	0.016393	technical writer	Hyderabad	f	8/27/92 0:00	90.00	state board	2009	93.00	...	Telanga
3995	355888	0.071879	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	bse,odisha	2008	65.50	...	Oris
3996	947111	0.041614	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	state board	2010	69.88	...	Karnata
3997	324966	0.092055	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	cbse	2008	68.00	...	Tamil Na

3998 rows × 26 columns

In [57]:
```
df['Specialization'].unique()
```

Out[57]:
```
array(['computer engineering',
       'electronics and communication engineering',
       'information technology', 'computer science & engineering',
       'mechanical engineering', 'electronics and electrical engineering',
       'electronics & telecommunications',
       'instrumentation and control engineering', 'computer application',
       'electronics and computer engineering', 'electrical engineering',
       'applied electronics and instrumentation',
       'electronics & instrumentation eng',
       'information science engineering', 'civil engineering',
       'mechanical and automation', 'industrial & production engineering',
       'control and instrumentation engineering',
       'metallurgical engineering',
       'electronics and instrumentation engineering',
       'electronics engineering', 'ceramic engineering',
       'chemical engineering', 'aeronautical engineering', 'other',
       'biotechnology', 'embedded systems technology',
       'electrical and power engineering',
       'computer science and technology', 'mechatronics',
       'automobile/automotive engineering', 'polymer technology',
       'mechanical & production engineering',
       'power systems and automation', 'instrumentation engineering',
       'telecommunication engineering',
       'industrial & management engineering', 'industrial engineering',
       'computer and communication engineering',
       'information & communication technology', 'information science',
       'internal combustion engine', 'computer networking',
       'biomedical engineering', 'electronics', 'computer science'],
      dtype=object)
```

In [58]:
```
df['Designation'].unique()
```

Out[58]:
```
array(['senior quality engineer', 'assistant manager', 'systems engineer',
       'senior software engineer', 'get', 'system engineer',
       'java software engineer', 'mechanical engineer',
       'electrical engineer', 'project engineer', 'senior php developer',
       'senior systems engineer', 'quality assurance engineer',
       'qa analyst', 'network engineer', 'product development engineer',
       'associate software developer', 'data entry operator',
       'software engineer', 'developer', 'electrical project engineer',
       'programmer analyst', 'systems analyst', 'ase',
       'telecommunication engineer', 'application developer',
       'ios developer', 'executive assistant', 'online marketing manager',
       'documentation specialist', 'associate software engineer',
       'management trainee', 'site manager', 'software developer',
       '.net developer', 'production engineer', 'jr. software engineer',
       'trainee software developer', 'ui developer',
       'assistant system engineer', 'android developer',
       'customer service', 'test engineer', 'java developer', 'engineer',
       'recruitment coordinator', 'technical support engineer',
```

```
'data analyst', 'assistant software engineer', 'faculty',
'entry level management trainee',
'customer service representative', 'software test engineer',
'firmware engineer', 'php developer', 'research associate',
'research analyst', 'quality engineer', 'programmer',
'technical support executive', 'business analyst', 'web developer',
'application engineer', 'project coordinator', 'engineer trainee',
'sap consultant', 'quality analyst', 'marketing coordinator',
'system administrator', 'senior engineer',
'business development managerde', 'network administrator',
'technical support specialist', 'business development executive',
'junior software engineer', 'asp.net developer',
'graduate engineer trainee', 'field engineer',
'assistant professor', 'trainee software engineer',
'senior software developer',
'quality assurance automation engineer', 'design engineer',
'telecom engineer', 'quality control engineer',
'hardware engineer', 'hr recruiter', 'sales associate',
'junior engineer', 'associate engineer', 'maintenance engineer',
'sales engineer', 'human resources associate',
'mobile application developer',
'electronic field service engineer', 'process associate',
'field service engineer', 'it support specialist',
'software development engineer', 'business process analyst',
'operation engineer', 'electrical designer', 'marketing assistant',
'sales executive', 'admin assistant', 'senior java developer',
'account executive', 'oracle dba', 'rf engineer',
'embedded software engineer', 'programmer analyst trainee',
'technical engineer', 'operations executive', 'trainee engineer',
'recruiter', 'lecturer', '.net web developer',
'marketing executive', 'operations assistant', 'associate manager',
'electrical design engineer', 'systems administrator',
'client services associate', 'it analyst', 'senior developer',
'cad designer', 'business technology analyst', 'asst. manager',
'service engineer', 'executive recruiter', 'planning engineer',
'associate technical operations', 'web designer',
'software architect', 'software quality assurance tester',
'seo trainee', 'process engineer',
'software quality assurance analyst', 'designer',
'business systems consultant', 'business development manager',
'junior research fellow', 'technical recruiter',
'operations analyst', 'quality assurance test engineer',
'linux systems administrator', 'software trainee',
'entry level sales and marketing', 'electrical field engineer',
'windows systems administrator', 'junior software developer',
'python developer', 'web application developer',
'assistant systems engineer', 'javascript developer',
'operation executive', 'performance engineer', 'technical writer',
'operations engineer and jetty handling', 'lead engineer',
'portfolio analyst', 'associate system engineer',
'mechanical design engineer', 'product engineer',
'network security engineer', 'operations manager',
'technical lead', 'operations', 'quality assurance tester',
'automation engineer', 'data scientist', 'quality associate',
'manual tester', 'sr. engineer', 'embedded engineer',
'service and sales engineer', 'telecom support engineer',
'engineer- customer support', 'cloud engineer', 'branch manager',
'business analyst consultant', 'technology lead',
'software trainee engineer', 'dcs engineer', 'junior manager',
'ux designer', 'clerical', 'hr generalist',
'database administrator', 'senior design engineer', 'seo',
'assistant engineer', 'marketing analyst', 'it executive',
'salesforce developer', 'software tester', 'sql dba',
'junior engineer product support', 'manager',
'senior business analyst', 'c# developer',
'implementation engineer', 'executive hr', 'executive engineer',
'sharepoint developer', 'system analyst',
'sales management trainee', 'senior project engineer',
'it recruiter', 'software engineer analyst',
'desktop support technician', 'continuous improvement engineer',
'process advisor', 'etl developer', 'sales and service engineer',
'project manager', 'training specialist', 'product manager',
'staffing recruiter', 'assistant programmer', 'quality controller',
'mis executive', 'game developer', 'digital marketing specialist',
'principal software engineer', 'software devloper',
'senior mechanical engineer', 'technical operations analyst',
'service coordinator', 'testing engineer', 'technical assistant',
'sap abap consultant', 'seo engineer', 'project assistant',
'talent acquisition specialist', 'sales account manager',
'software engineer trainee', 'customer service manager',
'help desk analyst', 'general manager', 'engineering manager',
'senior network engineer',
'field based employee relations manager', 'phone banking officer',
'support engineer', 'associate test engineer',
'technology analyst', 'network support engineer',
'it business analyst', 'junior system analyst',
'senior .net developer', 'secretary', 'research engineer',
'quality assurance auditor', 'process executive',
'lecturer & electrical maintenance', 'office coordinator',
```

```
'hr manager', 'html developer', 'sales support',
'front end web developer', 'administrative support',
'territory sales manager', 'project administrator',
'environmental engineer', 'web designer and seo',
'information security analyst',
'field business development associate', 'operational executive',
'administrative coordinator', 'senior risk consultant',
'desktop support engineer', 'cad drafter', 'noc engineer',
'industrial engineer', 'it engineer', 'human resources intern',
'senior quality assurance engineer', 'clerical assistant',
'software enginner', 'quality assurance',
'delivery software engineer', 'graphic designer',
'sales development manager', 'visiting faculty',
'business intelligence analyst', 'team lead',
'operational excellence manager', 'sales & service engineer',
'web intern', 'full stack developer', 'database developer',
'sr. database engineer', 'graduate apprentice trainee',
'software engineer associate', 'technical analyst',
'executive engg', 'it technician', 'business system analyst',
'process control engineer', 'technical consultant',
'business office manager', 'quality control inspector',
'product design engineer', 'manufacturing engineer',
'seo executive', 'sap analyst', 'software engineere',
'financial service consultant', 'co faculty', 'software analyst',
'desktop support analyst', 'graduate engineer',
'engineering technician', 'it assistant', 'marketing manager',
'human resource assistant', 'hr assistant', 'product developer',
'customer support engineer',
'quality control inspection technician', 'gis/cad engineer',
'senior web developer', 'sql developer', 'research staff member',
'sap abap associate consultant', 'associate qa',
'corporate recruiter', 'project management officer',
'business systems analyst', 'software programmer',
'help desk technician', 'sales manager', 'catalog associate',
'assistant store manager', 'software engg', 'it developer',
'apprentice', 'business consultant', 'controls engineer',
'ruby on rails developer', 'risk consultant', 'account manager',
'professor', 'assistant administrator', 'civil engineer',
'educator', 'service manager', 'teradata dba',
'full-time loss prevention associate', 'junior recruiter',
'associate developer', 'assistant electrical engineer',
'shift engineer', 'dotnet developer', 'rf/dt engineer',
'human resources analyst', 'software test engineerte',
'junior .net developer', 'java trainee', 'maintenance supervisor',
'r&d engineer', 'front end developer', 'engineer-hws',
'operations engineer', 'senior research fellow',
'web designer and joomla administrator',
'enterprise solutions developer',
'information technology specialist', 'site engineer',
'graduate trainee engineer', 'quality assurance analyst',
'cnc programmer', 'financial analyst', 'system engineer trainee',
'sap mm consultant', 'assistant system engineer trainee',
'qa trainee', 'teradata developer', 'hr executive',
'senior programmer', 'software test engineer (etl)',
'associate software engg', 'supply chain analyst', 'sales trainer',
'software executive', 'team leader',
'assistant system engineer - trainee', 'seo analyst',
'risk investigator', 'executive administrative assistant',
'program manager', 'r & d', 'sap functional consultant',
'website developer/tester', 'software designer',
'sales coordinator', 'qa engineer', 'aircraft technician',
'customer care executive', 'senior test engineer',
'program analyst trainee', 'electrical controls engineer',
'trainee decision scientist', 'editor', 'bss engineer', 'dba',
'software eng', 'computer faculty', 'recruitment associate',
'logistics executive', 'quality consultant',
'senior sales executive', 'db2 dba', 'test technician',
'it operations associate', 'software engineering associate',
'research scientist', 'jr. software developer'], dtype=object)
```

Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)

In [60]:
```
cse_data=df[df['Degree']=='B.Tech/B.E.']
job_roles=[ 'Programming Analyst', 'Software Engineer', 'Hardware Engineer' ,'Associate Engineer']
req_data=cse_data[cse_data['Designation'].isin(job_roles)]
salary_range=(250000,300000)
avg_sal=req_data['Salary'].mean()
if salary_range[0]<=avg_sal<=salary_range[1]:
    print('The claim is supported by the Data')
else:
    print('The claim is not supported by the Data')
```

```
The claim is not supported by the Data
```

In [61]:
```
from scipy import stats
contingency_table=pd.crosstab(df['Gender'],df['Designation'])
```

```
chi2_stats,p_val,dof,expected=stats.chi2_contingency(contingency_table)
alpha=0.05
if p_val< alpha:
    print("There is a significant relationship between gender and specialization preference.")
else:
     print("There is no significant relationship between gender and specialization preference.")
```

```
There is a significant relationship between gender and specialization preference.
```

Observation

Based on the analysis, the claim suggesting a salary range of 2.5-3 lakhs for Computer Science Engineering graduates in specific job roles is not supported by the data. Additionally, the statistical test results indicate that there is no significant relationship between gender and specialization preference among graduates, as per the AMCAT dataset.

Conclusion

The dataset contains the employement outcomes of engineering graduates as dependent variable(Salary,Job Titles and Job loctions)along with the standardized scores from three different areas - cognitive skill,technical skill,personality skill.

Here we have observe the dataset contains 4000 rows and 40 columns and this dataset have so many duplicated values and first we have to manipulate the dataset and remove the unwantes rows and columns after that check the nan values are there or not after that we have to take a cleaned dataset vizualizaions.

Here we used univariate analysis and many plot to analyse the dataset like PDF, Histograms, Boxplots, Countplots in this analysis we found outliers in each numerical column, probability and frequency distribution of each numerical column, e frequency distribution of each categorical Variable/Column and we Mention observations after each plot.

And along with univariate analysis we use bivariate analysis, In this bivariate analysis we analysis many plots and we find the relationships between numerical columns using Scatter plots, hexbin plots, pair plots, and also here we analyse the patterns between categorical and numerical columns using swarmplot, boxplot, barplot.

In this project By using these plots we analyse the relationships between employees salary, Graduationyear, Designation, 12percentage, 10percentage, 12graduation, 10board like this we find each and every empolyees background. Just like background verification.

In [ ]:

Loading [MathJax]/extensions/Safe.js