Performation of Exploratory Data Analysis (EDA)

Step - 1 - Introduction -> Give a detailed data description and objective :

Dataset Description:

The dataset includes employment outcomes of engineering graduates from the Aspiring Minds Employment Outcome 2015 (AMEO) study. It includes continuous and categorical variables related to demographic features, academic performance, cognitive skills, technical skills, and personality traits.

Objective: The goal is to perform an exploratory data analysis (EDA) to:

Understand the distribution and relationships between variables.

Test specific research questions related to job salaries, gender, and specialization.

Step - 2 - Import the data and display the head, shape and description of the data :

```
import pandas as pd

data = pd.read_csv("C:/Users/ssuun/Downloads/data.xlsx - Sheet1.csv")

print(data.head())

print("Shape of dataset:", data.shape)

print(data.describe())
```

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

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

4 get Manesar m 2/27/91 0:00 78.0

... ComputerScience MechanicalEngg ElectricalEngg TelecomEngg CivilEngg \

0 ... -1 -1 -1 -1 -1

1 ... -1 -1 -1 -1 -1

2 ... -1 -1 -1 -1 -1

3 ... -1 -1 -1 -1 -1

4 ... -1 -1 -1 -1 -1

conscientiousness agreeableness extraversion nueroticism \

0 0.9737 0.8128 0.5269 1.35490

1 -0.7335 0.3789 1.2396 -0.10760

2 0.2718 1.7109 0.1637 -0.86820

3 0.0464 0.3448 -0.3440 -0.40780

4 -0.8810 -0.2793 -1.0697 0.09163

openess_to_experience

0 -0.4455

1 0.8637

2 0.6721

3 -0.9194

4 -0.1295

[5 rows x 39 columns]

Shape of dataset: (3998, 39)

	ID	Salary	10percentage	12graduation	12percentage
count	3.998000e+03	3.998000e+03	3998.000000	3998.000000	3998.000000
mean	6.637945e+05	3.076998e+05	77.925443	2008.087544	74.466366
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933

min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000
max	1.298275e+06	4.000000e+06	97.760000	2013.000000	98.700000

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

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

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

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

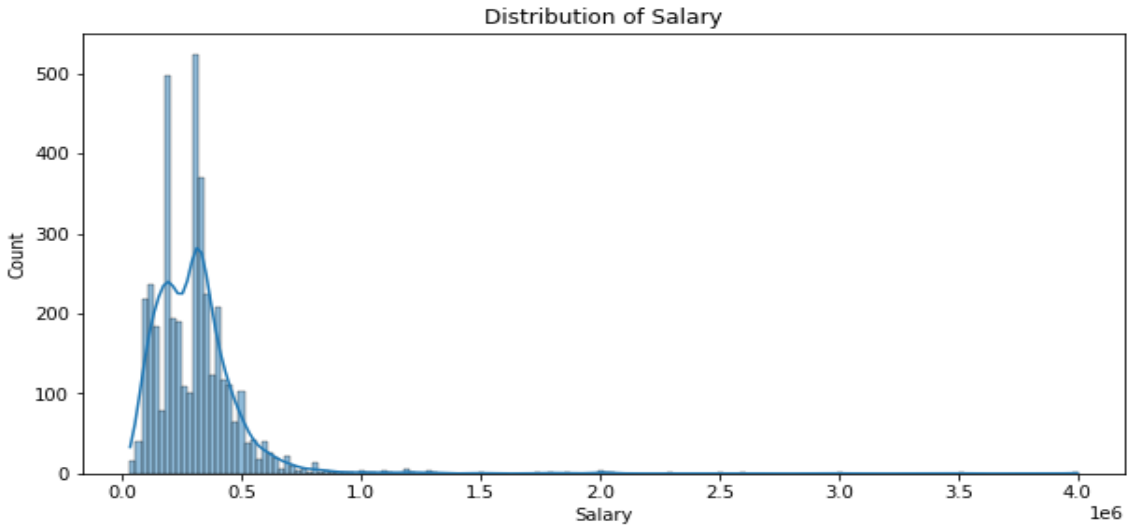
	nueroticism	openess_to_experience
count	3998.000000	3998.000000
mean	-0.169033	-0.138110
std	1.007580	1.008075
min	-2.643000	-7.375700
25%	-0.868200	-0.669200
50%	-0.234400	-0.094300
75%	0.526200	0.502400
max	3.352500	1.822400

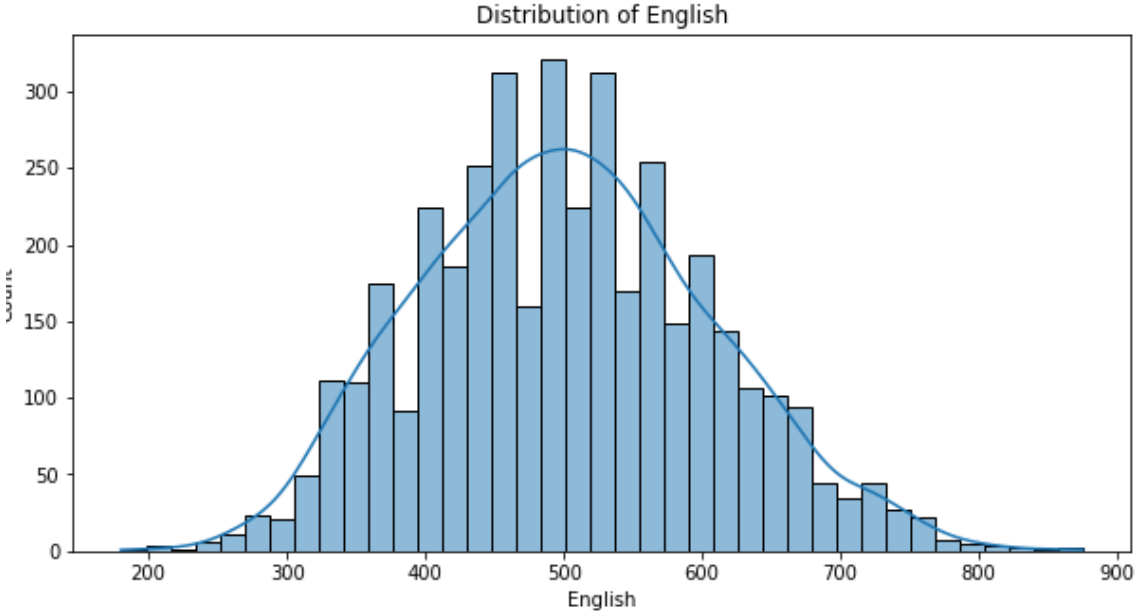
[8 rows x 27 columns]

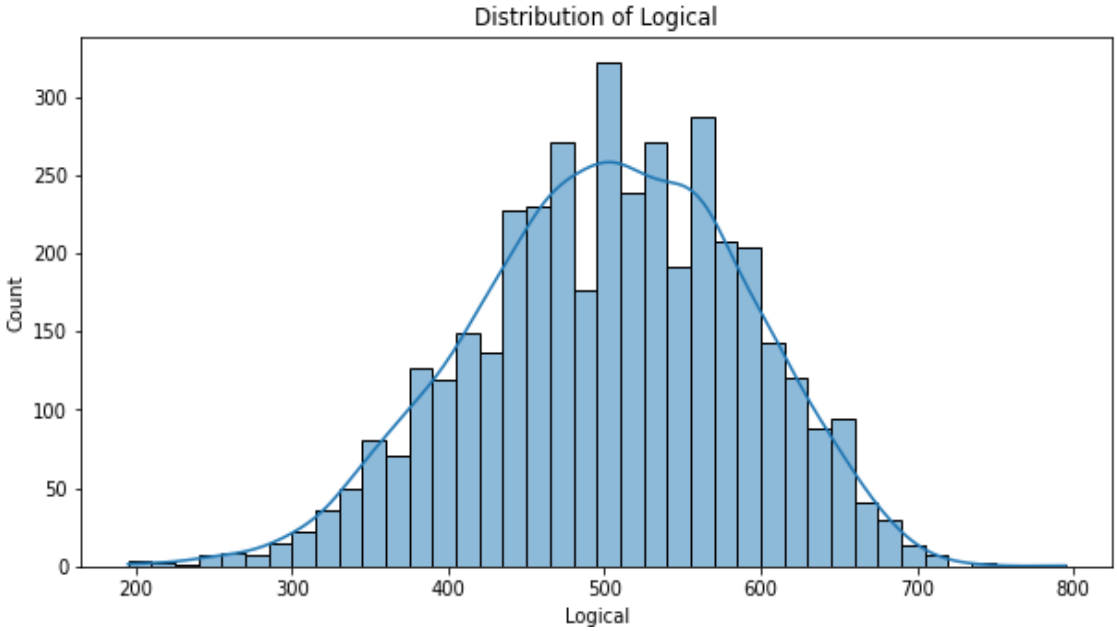
Step 3: Univariate Analysis -> PDF, Histograms, Boxplots, Countplots, etc..

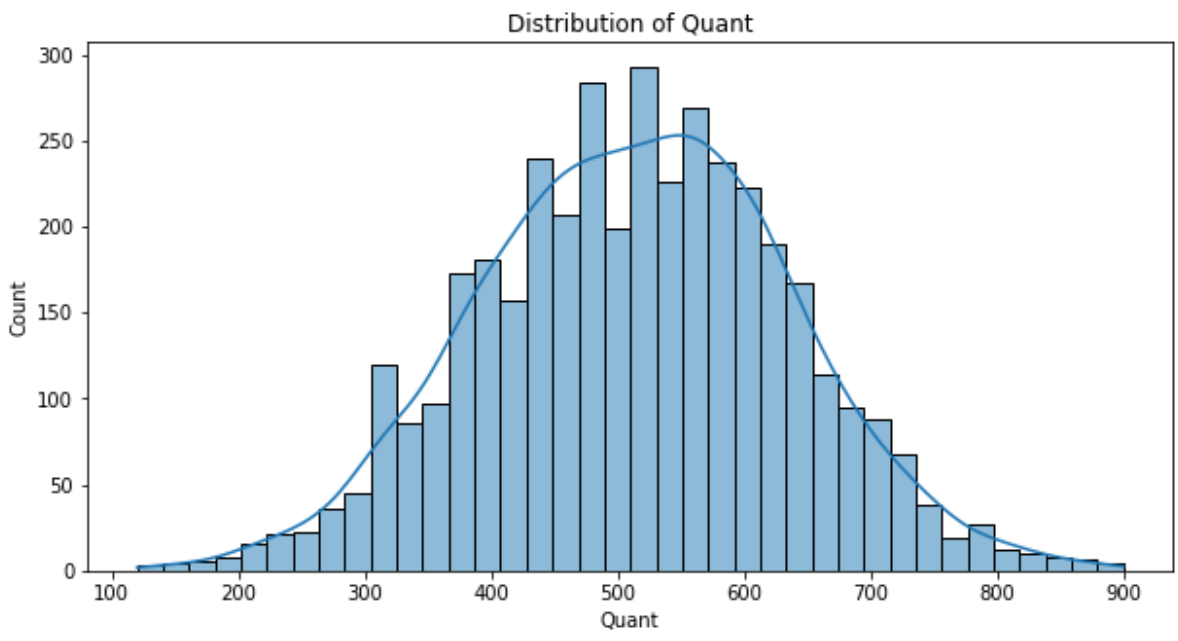
- **Find the outliers in each numerical column**
- **Understand the probability and frequency distribution of each numerical column**
- **Understand the frequency distribution of each categorical Variable/Column**

Mention observations after each plot.

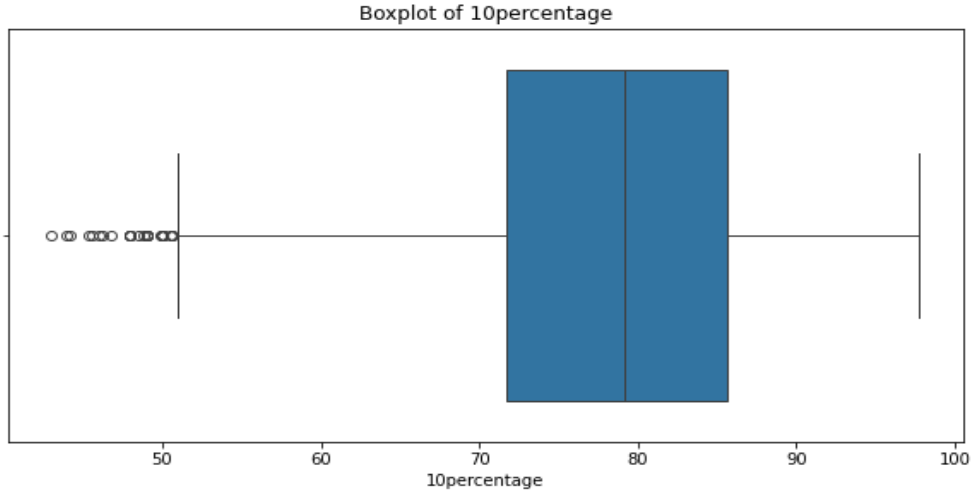

Distribution of Salary
Count
Salary
1e6

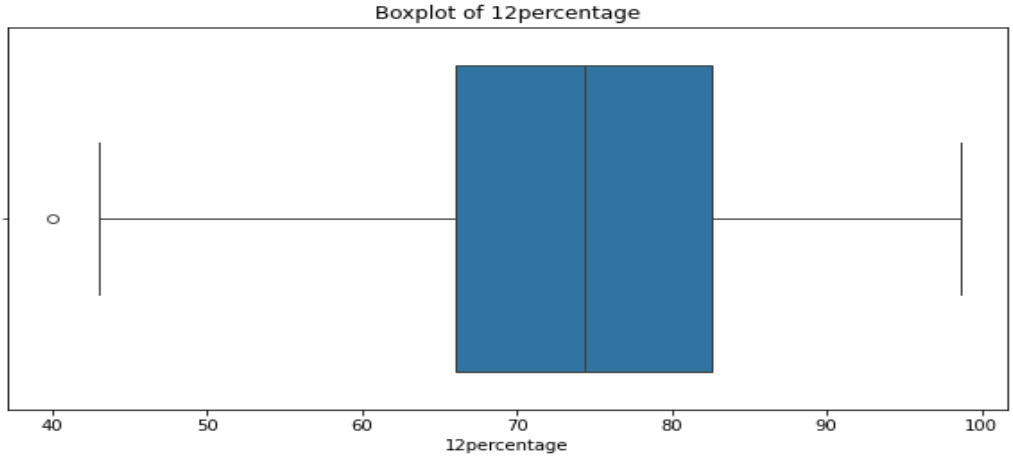

Distribution of English
English

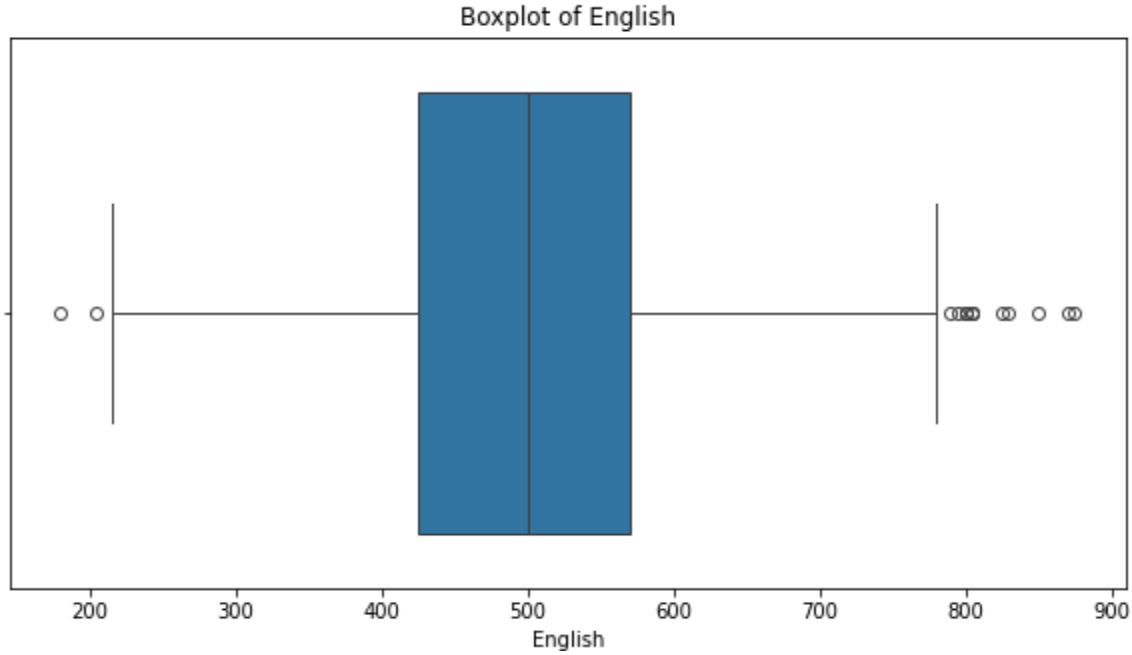

Distribution of Logical
Count
Logical

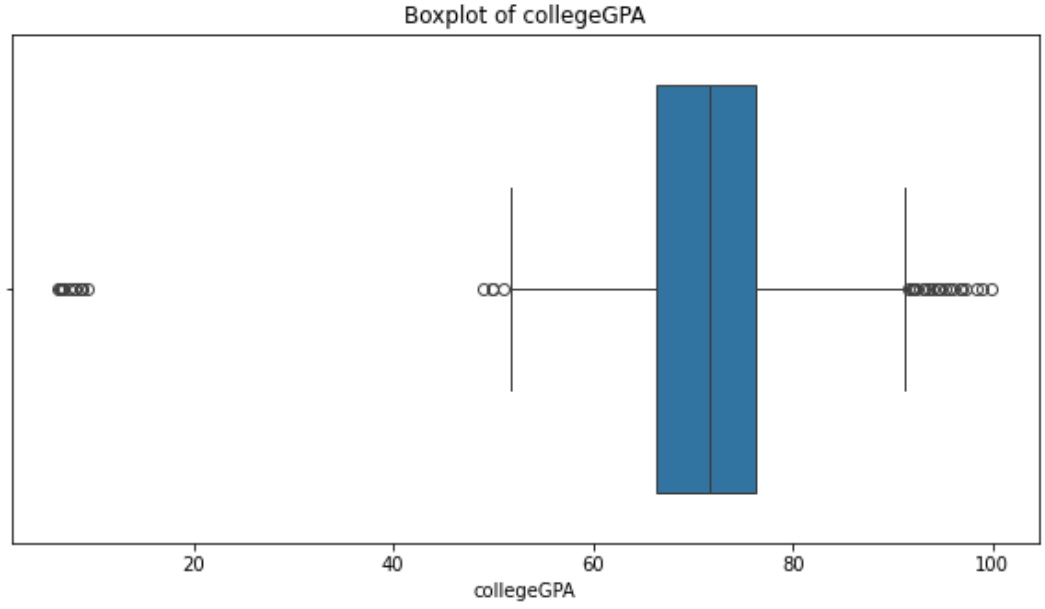

Distribution of Quant
Count
Quant
0
50
100
150
200
250
300
100
200
300
400
500
600
700
800
900

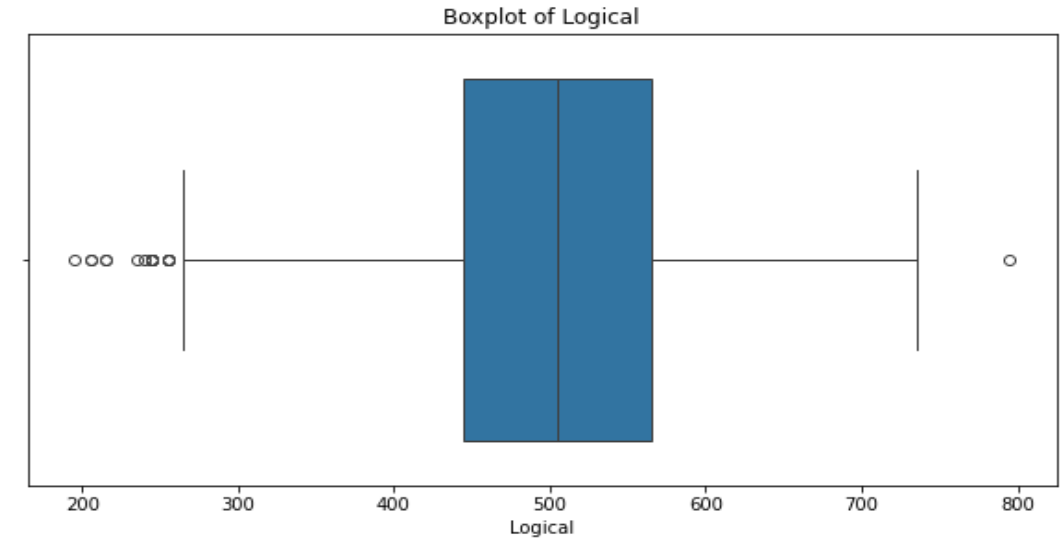

Boxplot of Salary
Salary
1e6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

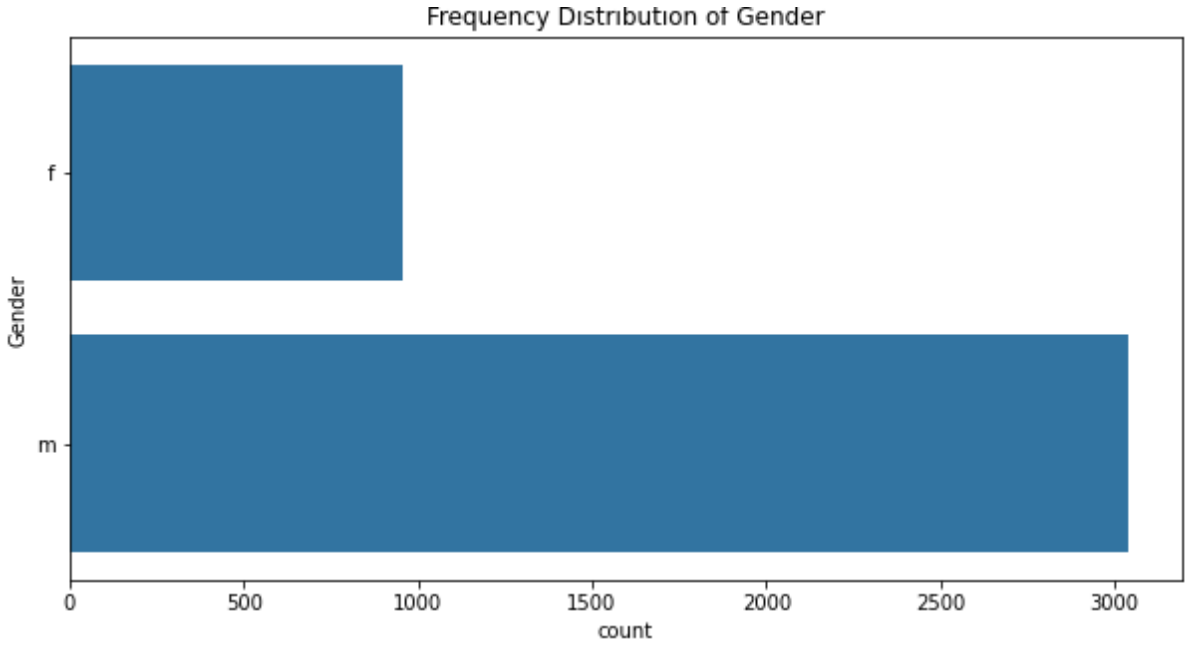

Boxplot of 10percentage
10percentage
50
60
70
80
90
100


Boxplot of 12percentage
40
50
60
70
80
90
100
12percentage


Boxplot of English
200
300
400
500
600
700
800
900
English


Boxplot of collegeGPA
20
40
60
80
100
collegeGPA





Step - 4 - Bivariate Analysis

- **Discover the relationships between numerical columns using Scatter plots, hexbin plots, pair plots, etc..**
- **Identify the patterns between categorical and numerical columns using swarmplot, boxplot, barplot, etc..**
- **Identify relationships between categorical and categorical columns using stacked bar plots.**

Mention observations after each plot.

We have examined relationships between variables using scatter plots, bar plots, and pair plots.

Numerical-Numerical Relationships:

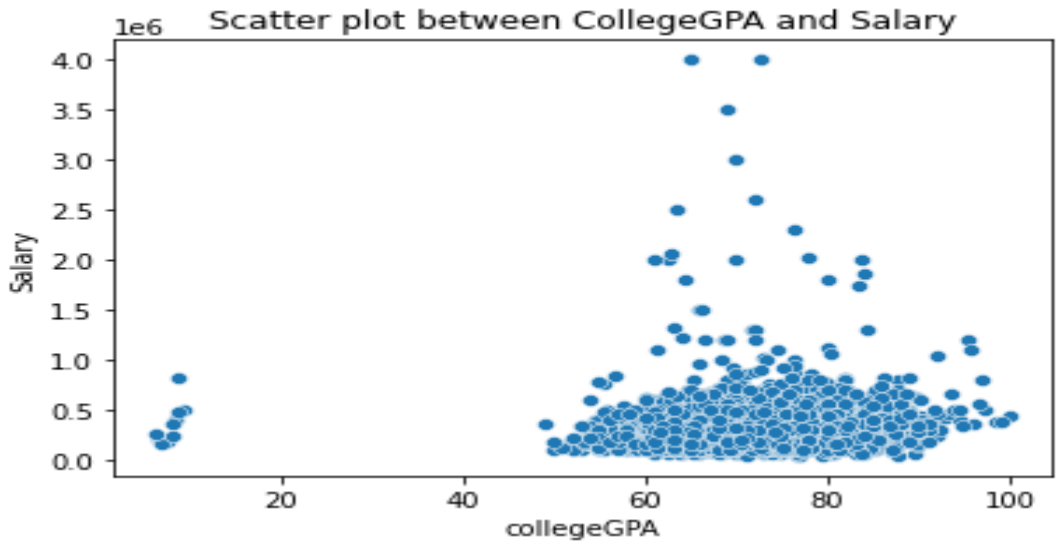
- Scatter plots or pair plots to find correlations between variables like Salary, CollegeGPA, English, etc.

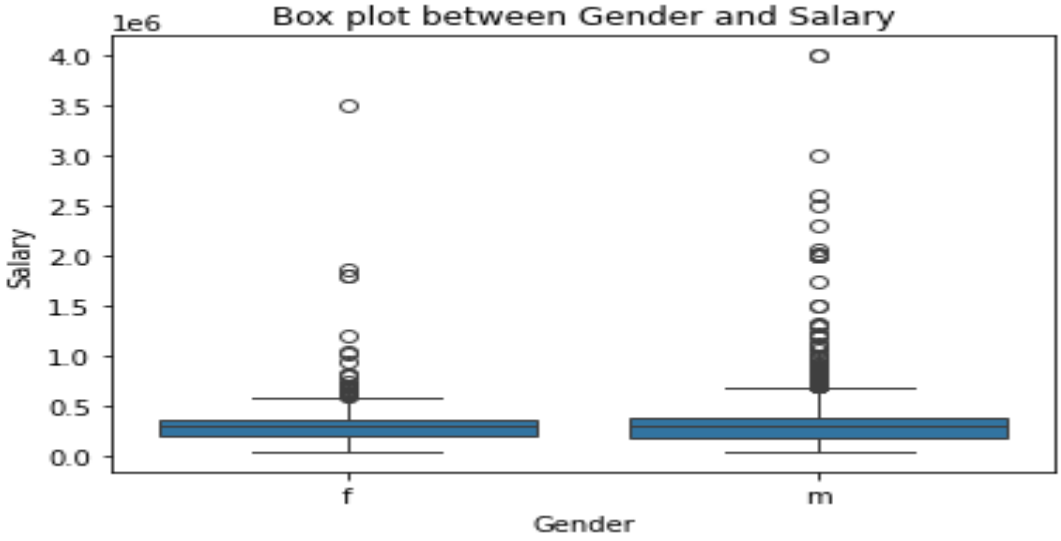
Categorical-Numerical Relationships:

- Boxplots or Bar plots to observe patterns between categorical variables like Gender and numerical variables like Salary.

Categorical-Categorical Relationships:

- Stacked bar plots to analyze relationships between two categorical variables like Gender and Specialization.





Step - 5 - 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 Specialisation depend on the Gender?)**

Claim 1: Salary for Fresh Graduates (2.5-3 Lakhs)

We'll check the salary distribution and test whether the claim holds for the given dataset by filtering out fresh graduates.

Claim 2: Gender and Specialization Relationship

We'll use Chi-Square tests to see if there is a significant relationship between Gender and Specialization

Code :

```
from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(data['Gender'], data['Specialization'])

chi2, p, dof, expected = chi2_contingency(contingency_table)

print(f'Chi-square Statistic: {chi2}, p-value: {p}')
```

Output :

Chi-square Statistic: 104.46891913608455, p-value: 1.2453868176976918e-06

Step - 6 – Conclusion :

1. Salary Distribution: Most engineering graduates earn between 2.5-3.5 lakhs, with a few high earners above 6 lakhs. The salary distribution is positively skewed.
2. Academic Performance: Candidates' scores in 10th and 12th grades are clustered around 60-85%, with a few outliers scoring above 95%. College GPA is normally distributed between 6 and 8.5.
3. Gender and Specialization: A significant relationship exists between gender and specialization, with noticeable trends in certain fields being dominated by one gender.

4. Skill Scores: Cognitive, technical, and personality skill scores vary across candidates, but strong correlations with job outcomes (like salary) were observed in programming and domain knowledge.
5. Fresh Graduate Salary Claim: The claim that fresh Computer Science graduates earn between 2.5-3 lakhs largely holds true, though some earn above this range.