INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

AMCAT Data Analysis

About me

Name : Vishal Kumar

Degress : Msc Computer Science(Specialization in Machine Intelligence)

Why Data Science

I want to learn data science because I find it fascinating to play with data and uncover hidden patterns that can help solve problems. Plus, I'm curious about how data science can be used in different areas like healthcare, finance, and technology to make smarter decisions and improve processes.

Work Experience

Data Science Intern at Innomatics Research Labs

Profile URLs

LinkedIn : https://www.linkedin.com/in/vishaldeoprasad/

Github: https://github.com/VishalDeoPrasad

INNOMATICS RESEARCH LABS

Agenda

Exploratory Data Analysis:

1. *Business Objective*
2. *Introduction about dataset*
3. *Data Cleaning Steps*
4. *Data Manipulation Steps*
5. *Univariate Analysis Steps*
6. *Bivariate Analysis Steps*
7. *Research Question 1*
8. *Research Question 2*
9. *Conclusions*

1. ***Business Objective:***

Objective 1: Examine how salary is influenced by age, gender, academic performance (10th and 12th grades), college tier, specialization, college GPA, and various components of the AMCAT scores (English, logical reasoning, quantitative skills, etc.), aiming to understand their impact on compensation structures.

Objective 2: 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.

Objective 3: Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)

2. Introduction about dataset

The dataset provided contains information about candidates and various attributes related to their academic background, employment, and scores in tests and personality assessments.

Below is a detailed description of the dataset columns:

1. **ID(UID)** : A unique ID to identify a candidate
2. **Salary(Continuous)** : Annual CTC offered to the candidate (in INR)
3. **DOJ(Date)** : Date of joining the company
4. **DOL(Date)** : Date of leaving the company
5. **Designation(Categorical)** :Designation offered in the job
6. **JobCity(Categorical)** : Location of the job (city)
7. **Gender(Categorical)** : Candidate's gender
8. **DOB(Date)** : Date of birth of candidate
9. **10percentage(Continuous)** : Overall marks obtained in grade 10 examinations
10. **10board(Continuous)** : The school board whose curriculum the candidate followed in grade 10
11. **12graduation(Date)** : Year of graduation - senior year high school
12. **12percentage(Continuous)** : Overall marks obtained in grade 12 examinations
13. **12board(Date)** : The school board whose curriculum the candidate followed in grade 12
14. **CollegeID(NA/ID)** : Unique ID identifying the college which the candidate attended
15. **CollegeTier(Categorical)** : Tier of college
16. **Degree(Categorical)** : Degree obtained/pursued by the candidate
17. **Specialization(Categorical)** : Specialization pursued by the candidate
18. **CollegeGPA(Continuous)** : Aggregate GPA at graduation
19. **CollegeCityID(NA/ID)** : A unique ID to identify the city in which the college is located in
20. **CollegeCityTier(Categorical)** : The tier of the city in which the college is located
21. **CollegeState(Categorical)** : Name of States
22. **GraduationYear(Date)** : Year of graduation (Bachelor's degree)
23. **English(Continuous)** : Scores in AMCAT English section
24. **Logical(Continuous)** : Scores in AMCAT Logical section
25. **Quant(Continuous)** : Scores in AMCAT Quantitative section
26. **Domain(Continuous/Standardized)** : Scores in AMCAT's domain module
27. **ComputerProgramming(Continuous)** : Score in AMCAT's Computer programming section
28. **ElectronicsAndSemicon(Continuous)** : Score in AMCAT's Electronics & Semiconductor Engineering section
29. **ComputerScienc(Continuous)** : Score in AMCAT's Computer Science section
30. **MechanicalEngg(Continuous)** : Score in AMCAT's Mechanical Engineering section
31. **ElectricalEngg(Continuous)** : Score in AMCAT's Electrical Engineering section
32. **TelecomEngg(Continuous)** : Score in AMCAT's Telecommunication Engineering section
33. **CivilEngg(Continuous)** : Score in AMCAT's Civil Engineering section
34. **conscientiousness(Continuous/ Standardized)** : Scores in one of the sections of AMCAT's personality test
35. **agreeableness(Continuous/Standardized)** : Scores in one of the sections of AMCAT's personality test
36. **extraversion(Continuous/ Standardized)** : Scores in one of the sections of AMCAT's personality test
37. **neuroticism(Continuous/ Standardized)** : Scores in one of the sections of AMCAT's personality test
38. **openess_to_experience(Continuous/ Standardized)** : Scores in one of the sections of AMCAT's personality test

INNOMATICS RESEARCH LABS

Dataset Overview

- A total of 38 columns are utilized to evaluate the individual impacts on salary.
- There are 3998 data points in all, and each one corresponds to a single engineering graduate.
- The analysis incorporates 4000 data points, facilitating comprehensive insights with all necessary information.
- There are 10 variables of type float, 17 variables of type int, and 12 variables of type object

3. Data Cleaning Steps

Steps for Cleaning Data

1. '**Unnamed: 0**' and other unwanted columns were eliminated from the dataset
2. **Null Values:** after it was checked for null values and duplicate entries.
3. **Datetime format:** 'DOJ' (Date of Joining), 'DOB' (Date of Birth), and 'DOL' (Date of Leaving) datetime columns were converted to the correct datetime format.
4. **Missing values:**'Designation', 'JobCity', '10board', '12board', and 'Specialization' columns have had their Missing values.

Following the completion of the data cleaning procedure, the dataset looks like this:

```
data.head()
```

Python

	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	12board	CollegeID	CollegeTier	Degree	Specialization	collegeGPA	CollegeCity
0	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	board ofsecondary education,ap	2007	95.8	board of intermediate education,ap	1141	2	B.Tech/B.E.	computer engineering	78.00	11
1	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	cbse	2007	85.0	cbse	5807	2	B.Tech/B.E.	electronics and communication engineering	70.06	58
2	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	cbse	2010	68.2	cbse	64	2	B.Tech/B.E.	information technology	70.00	
3	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	cbse	2007	83.6	cbse	6920	1	B.Tech/B.E.	computer engineering	74.64	69
4	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.0	cbse	2008	76.8	cbse	11368	2	B.Tech/B.E.	electronics and communication engineering	73.90	113

4. Data Manipulation Steps

Steps for Manipulation Data

1. **Datetime format:** 'DOJ' (Date of Joining), 'DOB' (Date of Birth), and 'DOL' (Date of Leaving) datetime columns were converted to the correct datetime format.
2. **Missing values:** 'Designation', 'JobCity', '10board', '12board', and 'Specialization' columns have had their Missing values.
3. **Transform Features:** 'collegeGPA', 'English', 'Logical', 'Quant' and 'Salary' transformed columns to increase accuracy and consistency of the data.
4. **Add Features:** Find out 'Age' of the student form 'DOB' columns

```
data.head()
```

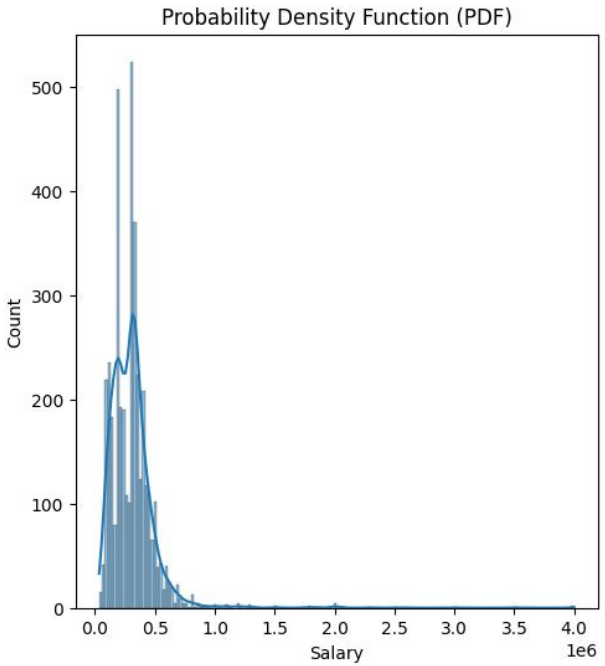
Python

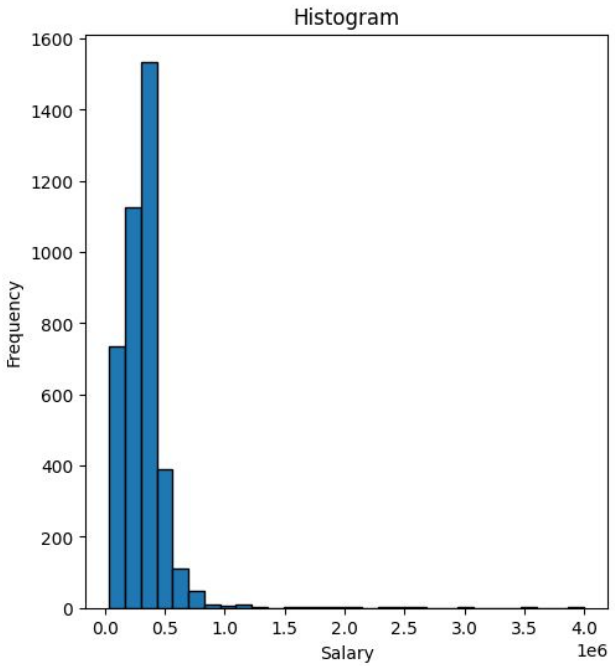
	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	12board	CollegeID	CollegeTier	Degree	Specialization	collegeGPA	CollegeCityI
0	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	board ofsecondary education,ap	2007	95.8	board of intermediate education,ap	1141	2	B.Tech/B.E.	computer engineering	78.00	11
1	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	cbse	2007	85.0	cbse	5807	2	B.Tech/B.E.	electronics and communication engineering	70.06	58
2	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	cbse	2010	68.2	cbse	64	2	B.Tech/B.E.	information technology	70.00	
3	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	cbse	2007	83.6	cbse	6920	1	B.Tech/B.E.	computer engineering	74.64	69
4	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.0	cbse	2008	76.8	cbse	11368	2	B.Tech/B.E.	electronics and communication engineering	73.90	113

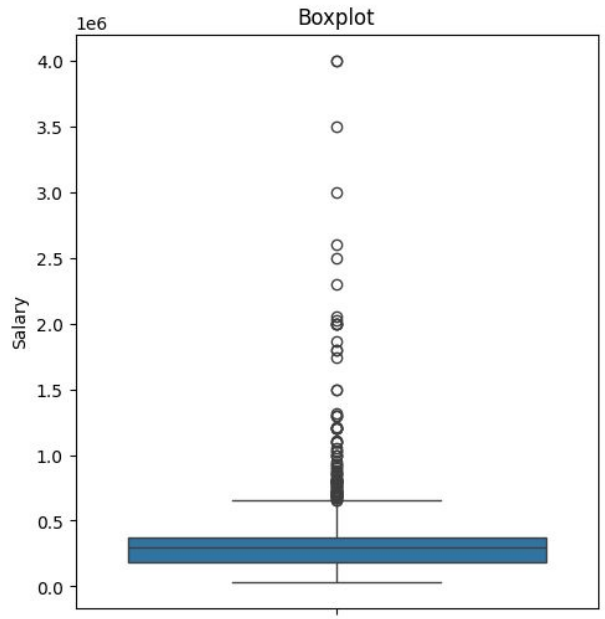
INNOMATICS RESEARCH LABS

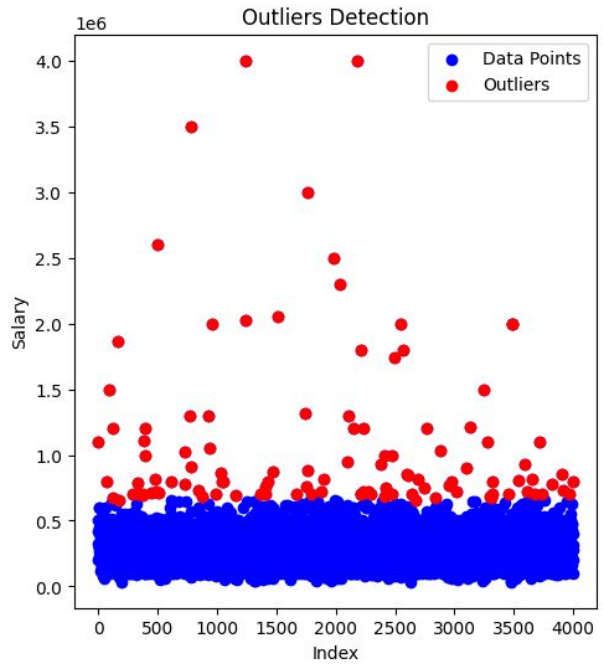
5. Univariate Analysis Steps

1. **Salary:**
   a. **The salary distribution reveals that the predominant range lies between 2LPA to 5LPA, indicating a common income bracket among respondents.**
   b. **The minimum recorded salary is 35000, showcasing the lower end of the salary spectrum within the dataset.**
   c. **On the higher end, the maximum reported salary peaks at 40LPA, indicating substantial earning potential among certain individuals in the dataset.**







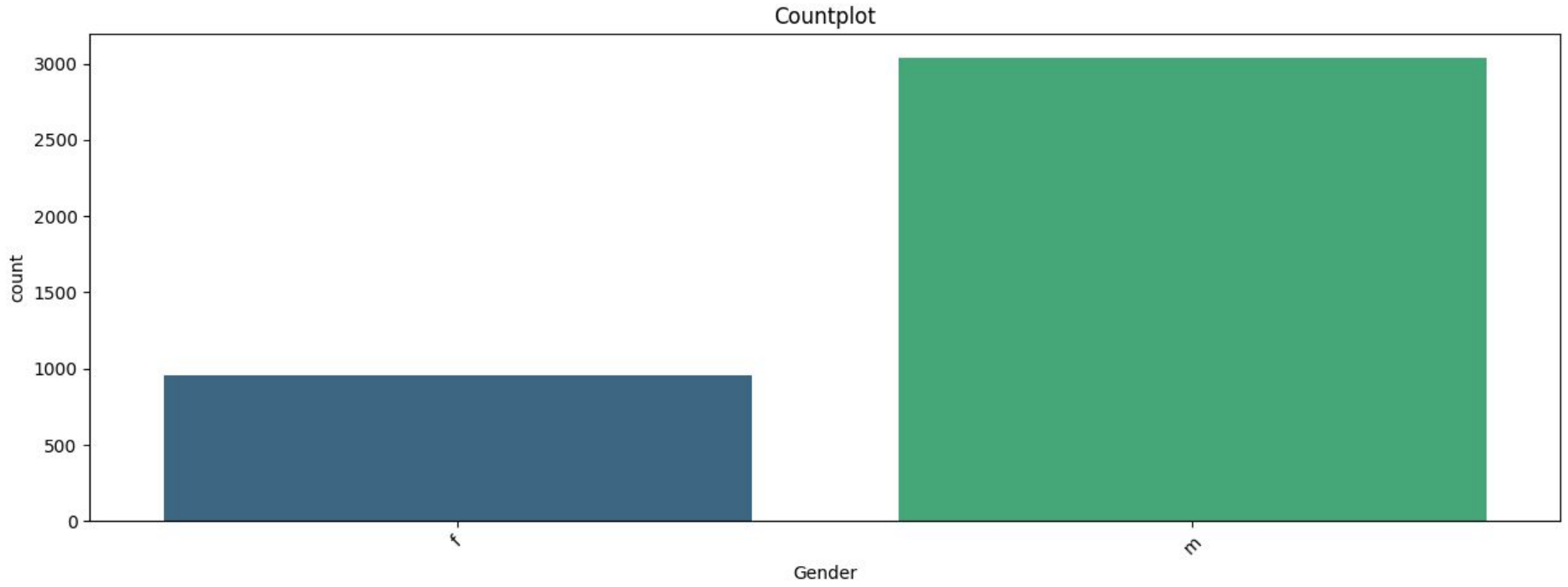


INNOMATICS RESEARCH LABS

5. Univariate Analysis Steps

2. Gender:

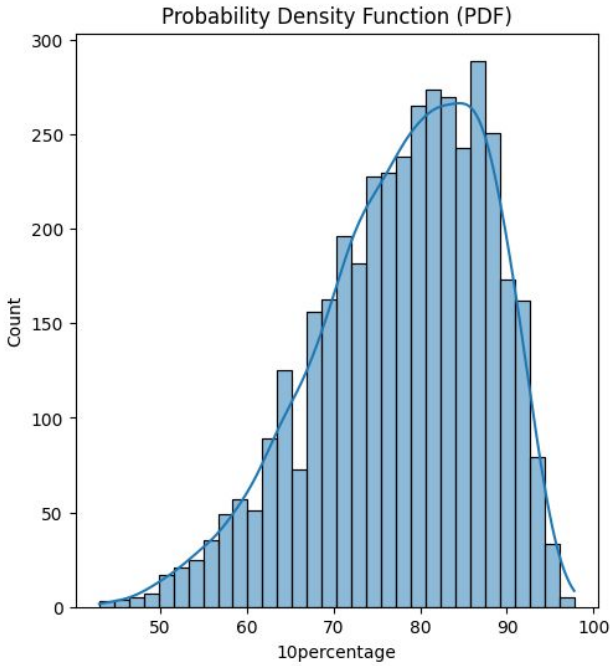
a. **The dataset comprises a larger representation of male students compared to female students, indicating a gender disparity in the sample.**
b. **Specifically, there are 3041 male students, significantly outnumbering the 957 female students recorded in the dataset.**

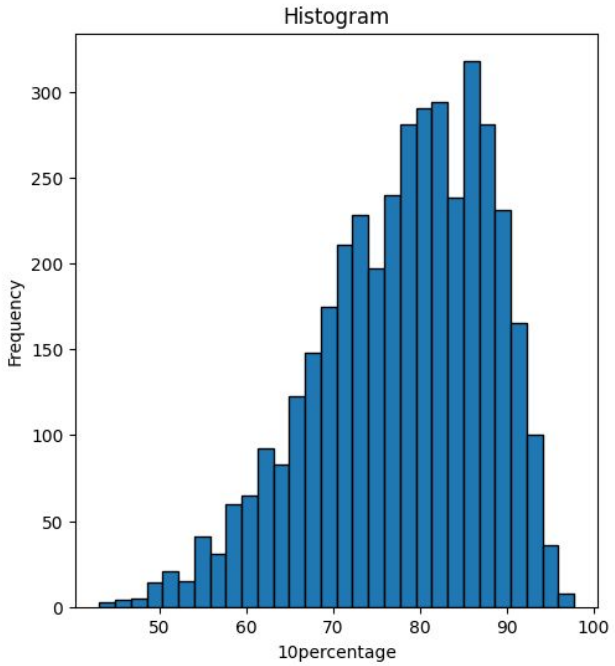


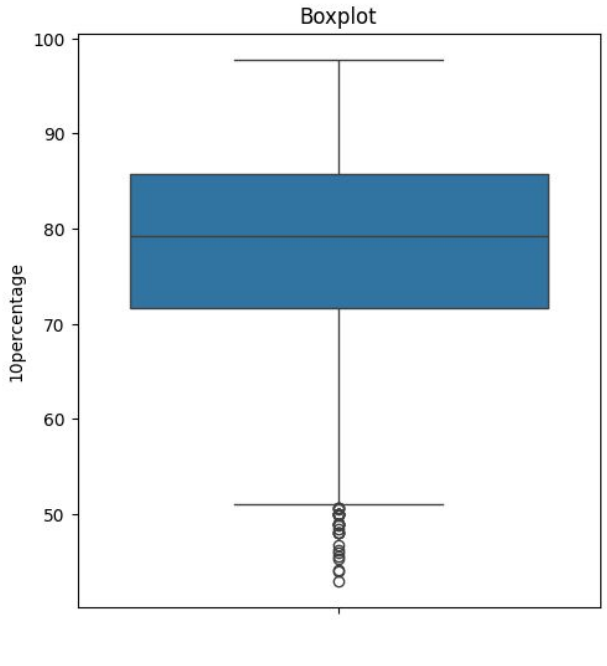
5. Univariate Analysis Steps

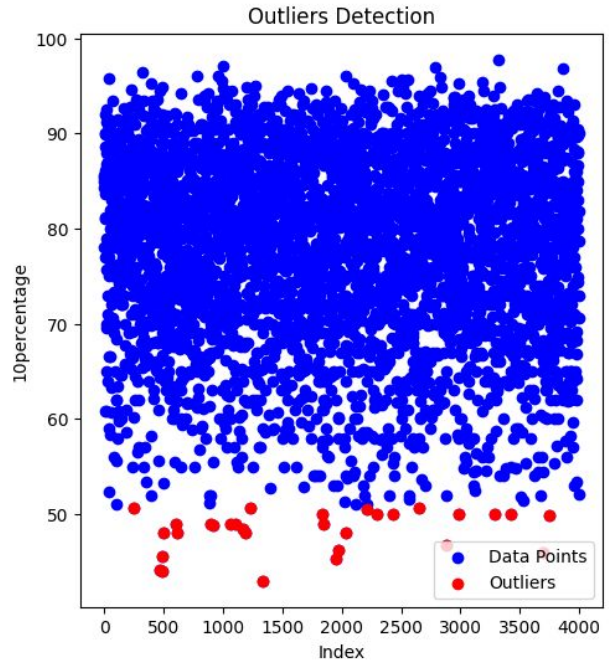
3. **10percentage**:

a. The minimum 10th percentage attained by students is 43.0, indicating the lowest academic performance within the dataset.
b. On the higher end, the maximum 10th percentage achieved reaches 97.76, showcasing exceptional academic achievement by some individuals.
c. The average 10th percentage among students stands at 77.93%, representing the typical level of academic performance in this cohort.
d. With a standard deviation of approximately 9.85, there exists a notable degree of variability in 10th percentage scores, highlighting differences in academic performance among students.





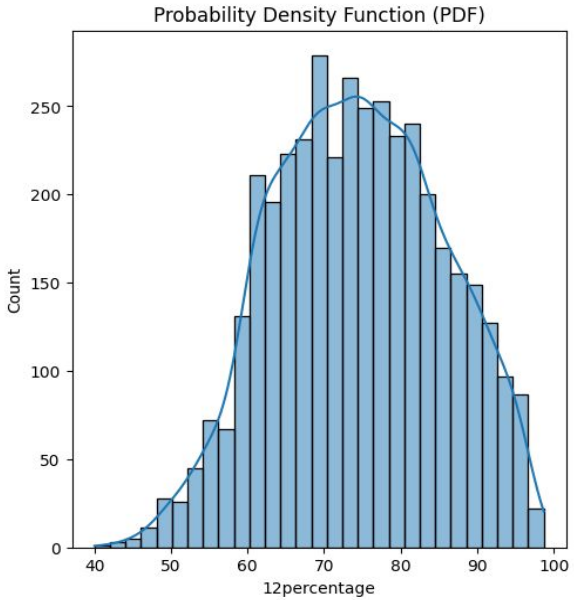


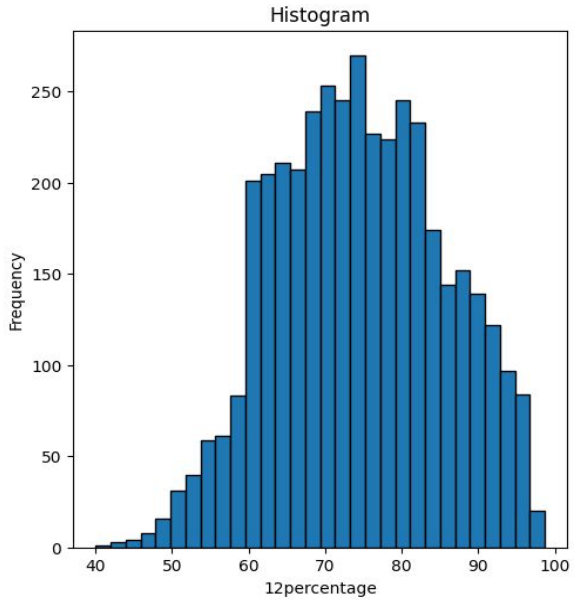


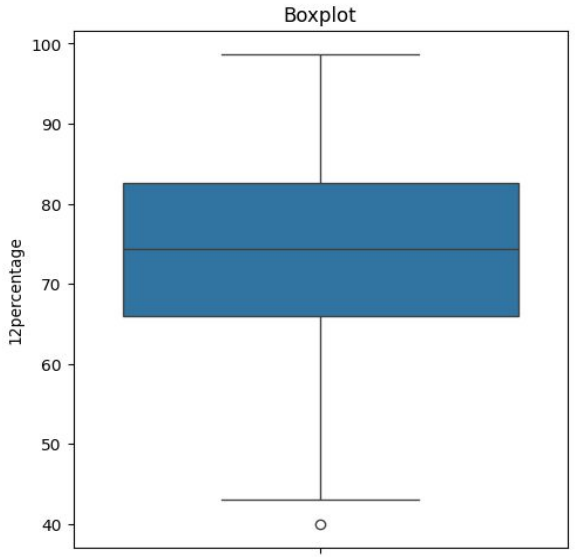
5. Univariate Analysis Steps

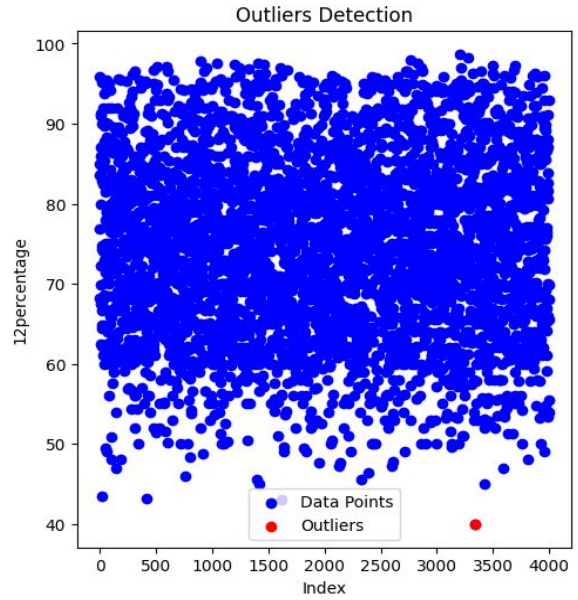
4. **12percentage**:

a. The mean 12th percentage among students is 74.47, reflecting the average level of academic achievement in this dataset.
b. With a standard deviation of approximately 11, there is a notable degree of variability in 12th percentage scores, indicating differences in academic performance among students.
c. lowest recorded 12th percentage is 40.00, suggesting some students achieved lower academic results.
d. Conversely, the highest 12th percentage attained is 98.70, showcasing exceptional academic performance by certain individuals.
e. Additionally, 75% of students have a 12th percentage equal to or below 82.60, indicating the distribution of scores within the dataset.







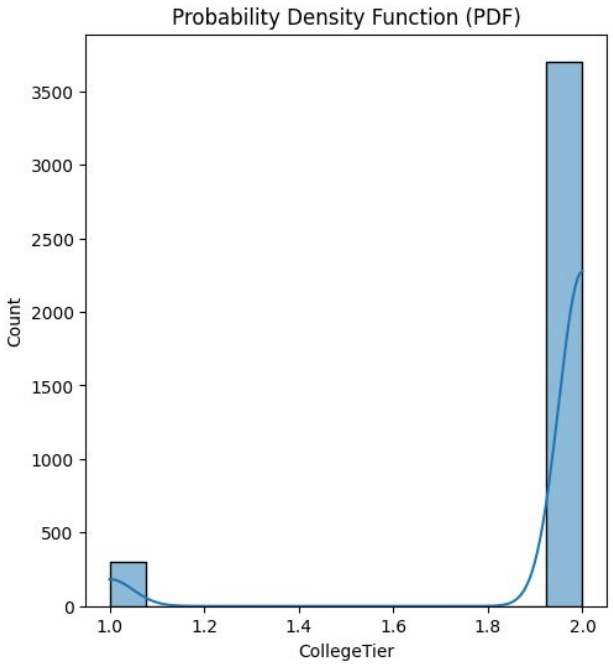


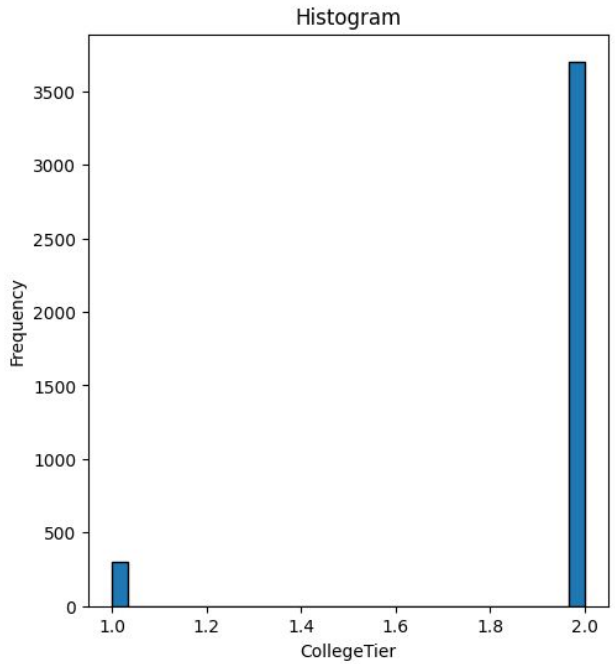
5. Univariate Analysis Steps

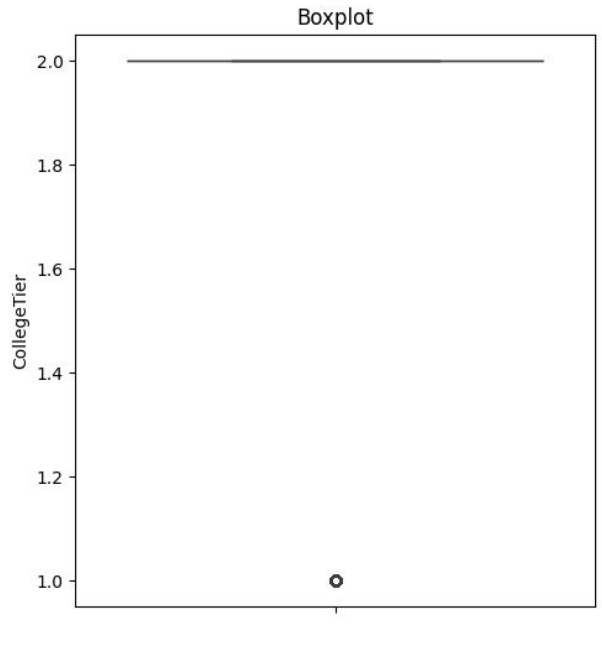
5. **collegeTier**:

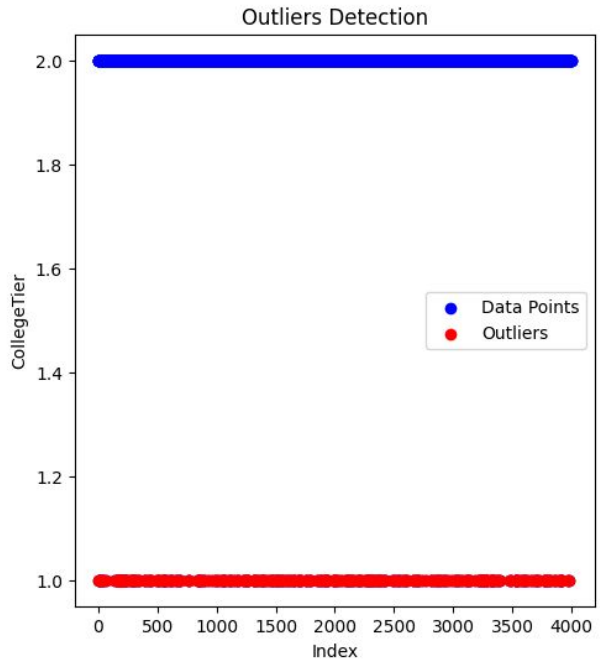
a. The majority of students in the dataset originate from Tier 2 colleges, with a total of 3701 students belonging to this category.

b. In comparison, the number of students hailing from Tier 1 colleges is significantly lower, with a total of 297 students.









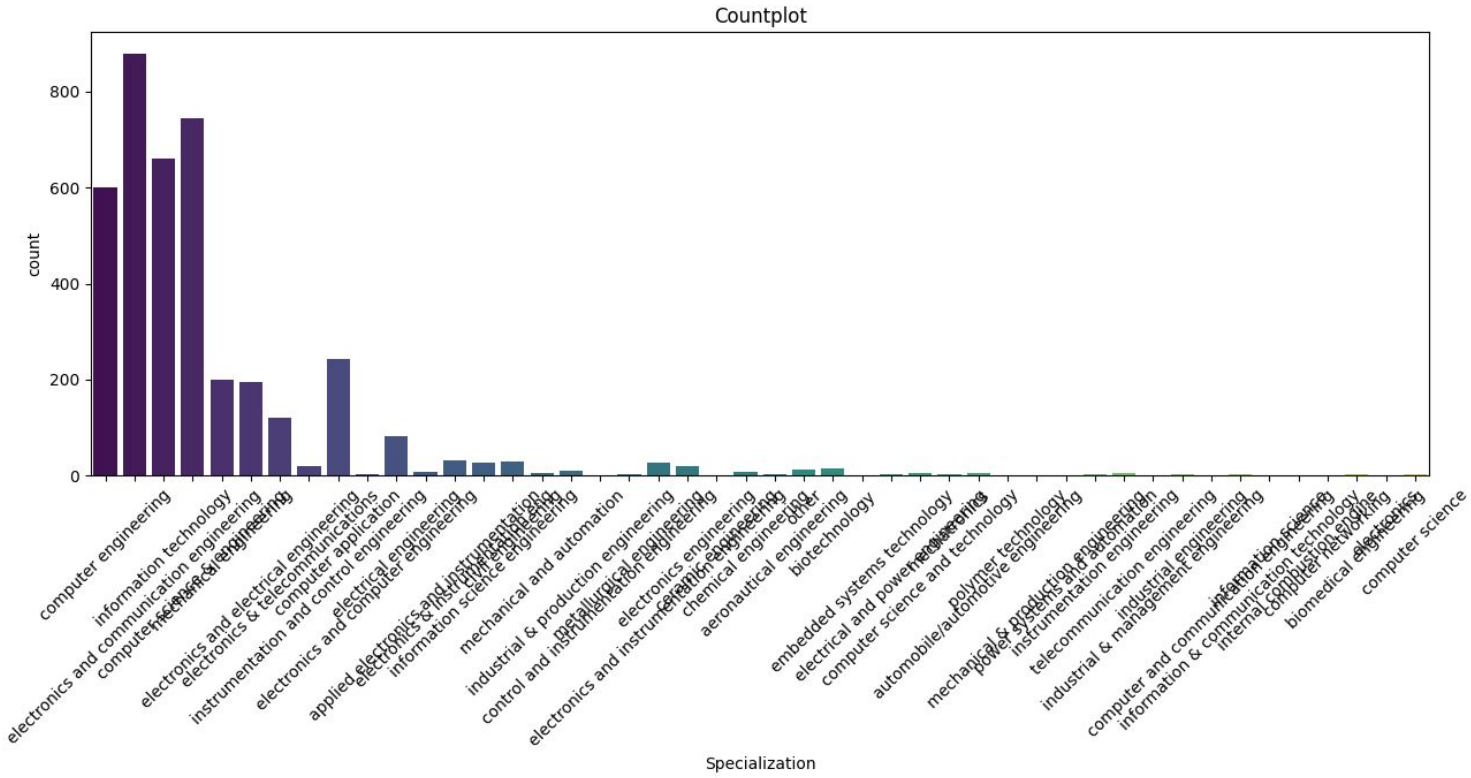
INNOMATICS RESEARCH LABS

5. Univariate Analysis Steps

6. **Specialization**:

The dataset reveals the top 5 chosen specializations among students, based on the frequency of occurrence:

a. Electronics and Communication Engineering: 880 students
b. Computer Science & Engineering: 744 students
c. Information Technology: 660 students
d. Computer Engineering: 600 studer
e. Computer Application: 244 studen

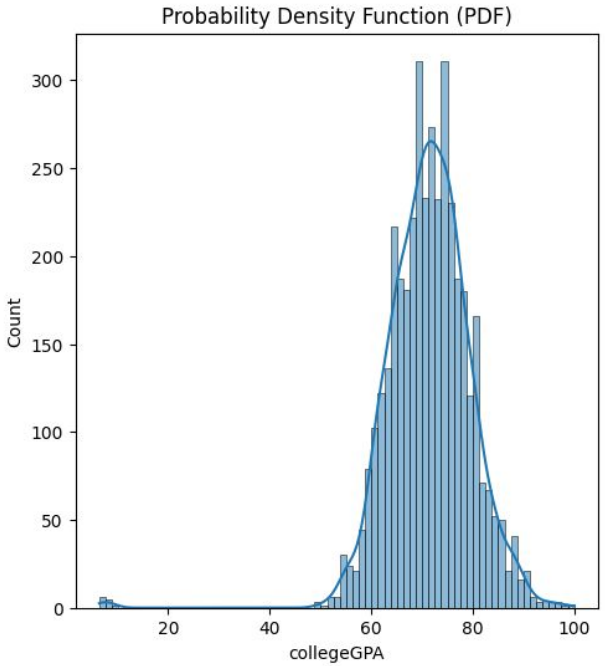


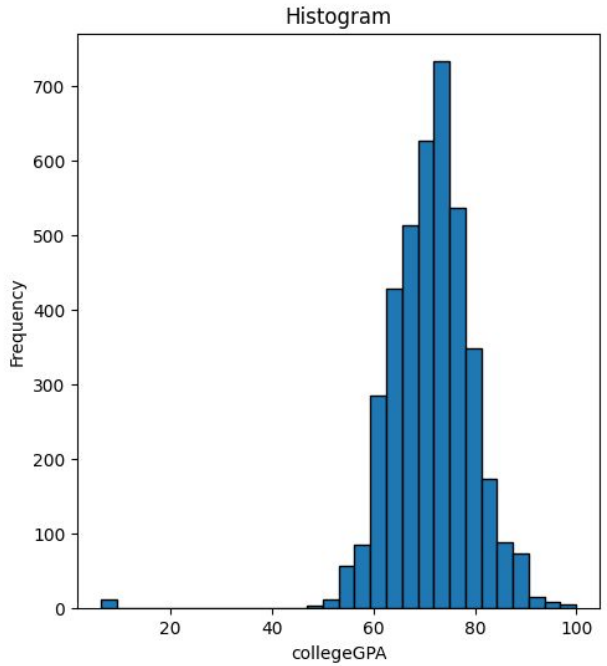
INNOMATICS RESEARCH LABS

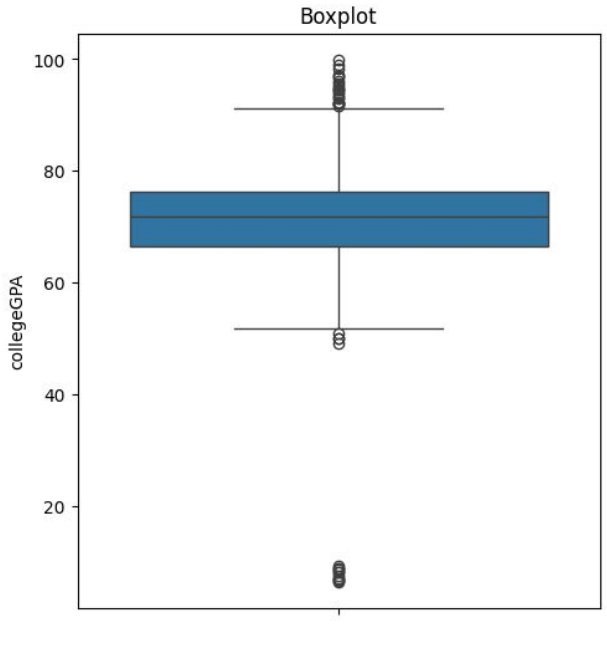
5. Univariate Analysis Steps

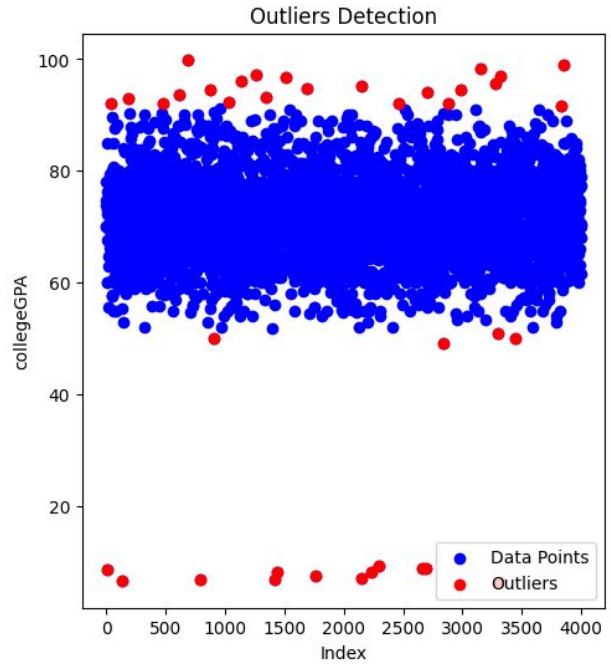
7. **collegeGPA**:

a. The mean college GPA among students is 71.49, indicating the average academic performance in terms of GPA within the dataset.
b. With a standard deviation of approximately 8.17, there is variability in college GPA scores, reflecting differences in academic achievement among students.
c. The lowest recorded college GPA is 6.45, suggesting some students have struggled academically.
d. On the other end of the spectrum, the highest college GPA achieved is 99.93, showcasing exceptional academic performance by certain individuals.



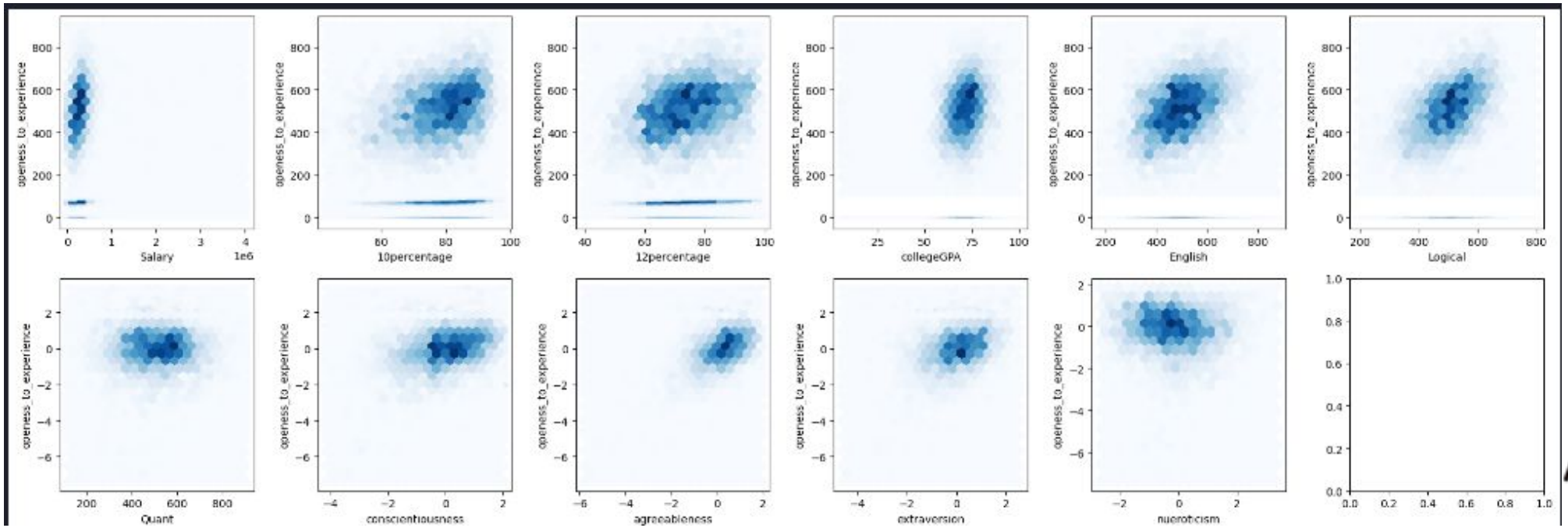




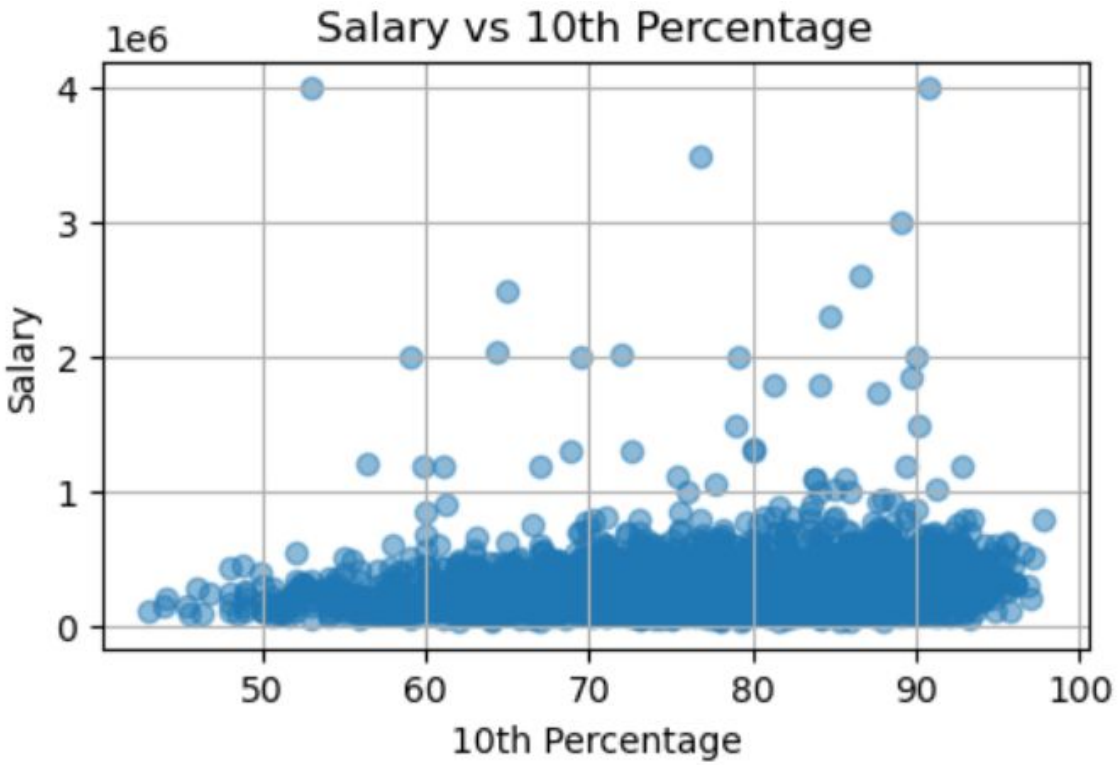


6. Bivariate Analysis Steps

1. Objective: The grid of hexbin plots serves as a visual exploration tool to analyze relationships between pairs of numerical features within the dataset.
2. Grid Layout: The grid comprises a 3 by 6 arrangement, with each subplot representing a distinct combination of two numerical columns.
3. Data Representation: Hexbin plots are employed to visually represent the distribution and density of data points for the selected pairs of features.
4. Insight Generation: By examining the hexbin plots, potential patterns, correlations, or clusters between different numerical variables can be identified.
5. Enhanced Understanding: This approach enhances comprehension of the dataset's characteristics, facilitating deeper insights into the interplay between numerical features.



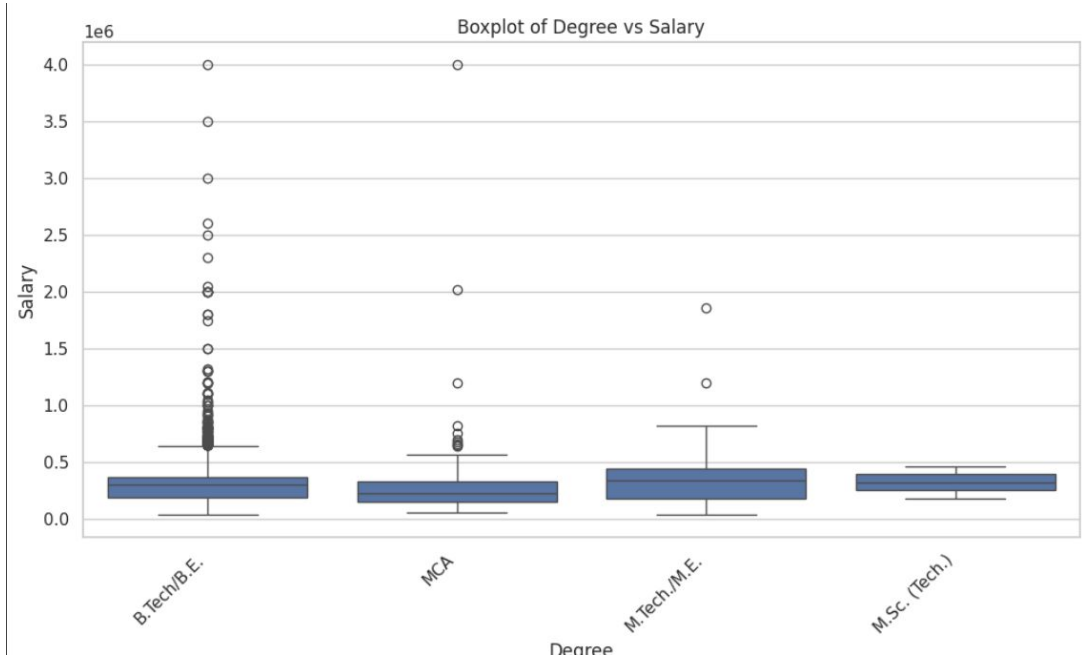
6. Bivariate Analysis Steps

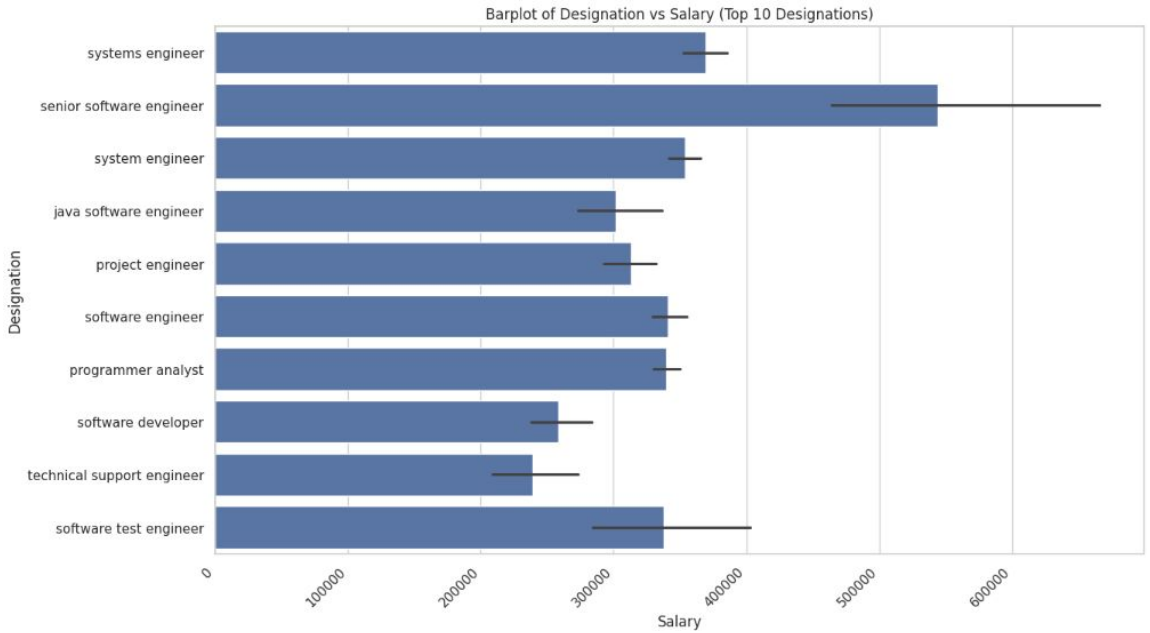


- Here, we have used the salary and 10th % columns from the provided dataframe to do a bivariate analysis.
- The scatter plot demonstrates that the pay and proportion of Indians in the 10th grade are positively correlated. Every dot on the graph reflects the data of one individual. The person's 10th-grade percentage is displayed on the x-axis, while their pay is displayed on the y-axis.
- The data clearly shows an increasing tendency, meaning that those with greater percentages in the 10th grade often earn better wages.

- The data does, however, also show a great deal of variety. There are some who earn great wages while having low 10th-grade percentages and those who earn poor salaries despite having high percentages.
- This implies that criteria like education, experience, and abilities have an impact on compensation in addition to the percentage of students who pass the 10th grade.

6. Bivariate Analysis Steps





- The M.Tech/M.E. degree has a somewhat higher pay than the other degrees in the degree vs. salary bar plot.
- There are undoubtedly more outliers for B.Tech/B.E. degrees, indicating that in extreme circumstances, there may be a potential for a greater wage.

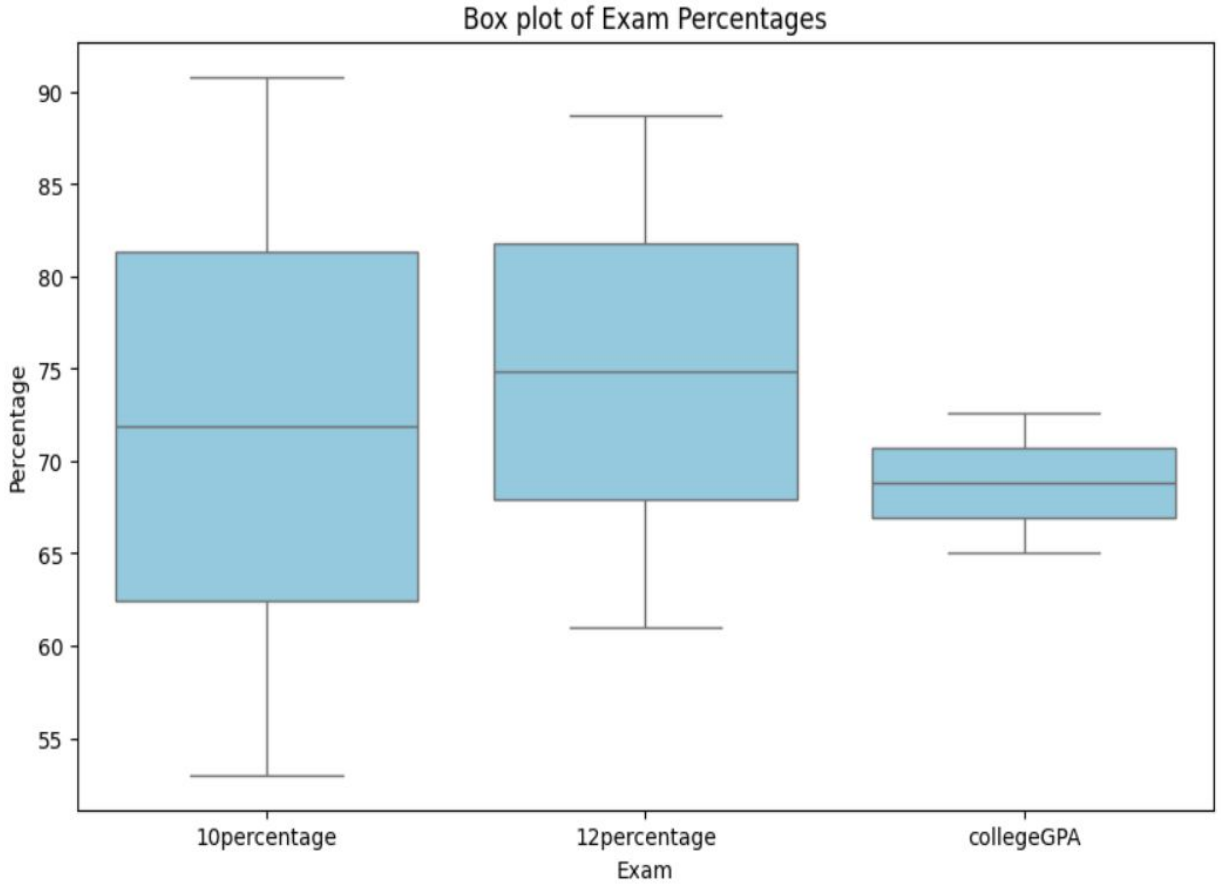
- It is evident from the bar plot that senior software engineers often earn higher salaries than those with other designations.

- Of the top ten highest paid positions, technical support engineers make the least money.

INNOMATICS RESEARCH LABS

6. Bivariate Analysis Steps

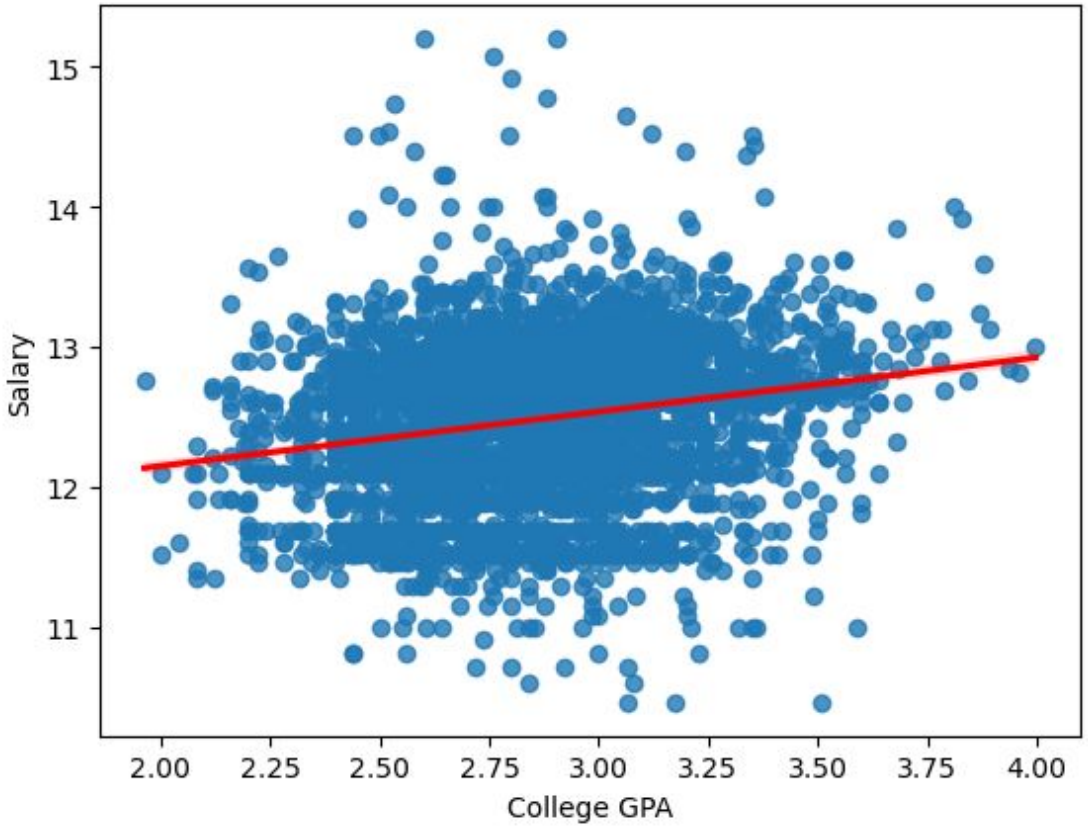
box plot shows that the exam scores were fairly evenly distributed, with a slight skew towards the higher end. There is a small number of outliers, but the majority of scores fall within the IQR.



- The box plot shows the distribution of exam percentages.
- The median score is 75%.
- The first quartile is 65%, meaning that 25% of students scored lower than 65%.
- The third quartile is 85%, meaning that 75% of students scored lower than 85%.
- The interquartile range (IQR) is 20, meaning that the middle 50% of scores fall within a range of 20 percentage points.
- There are outliers at 55% and 90%

6. Bivariate Analysis Steps

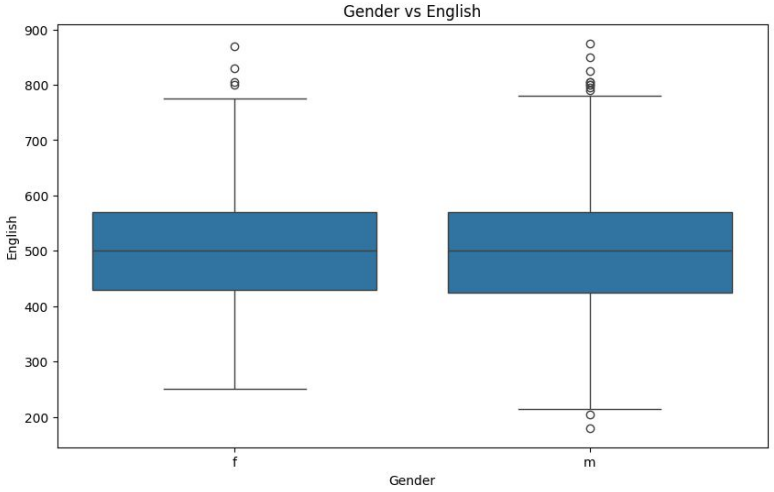
Relationship between Salary and College GPA

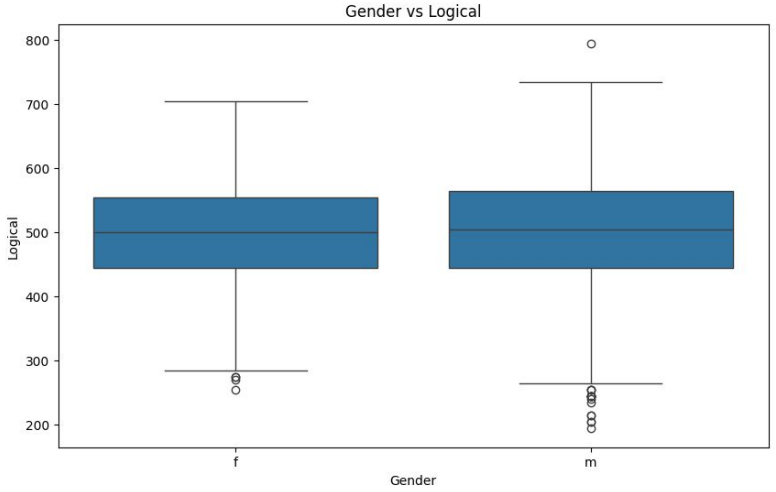


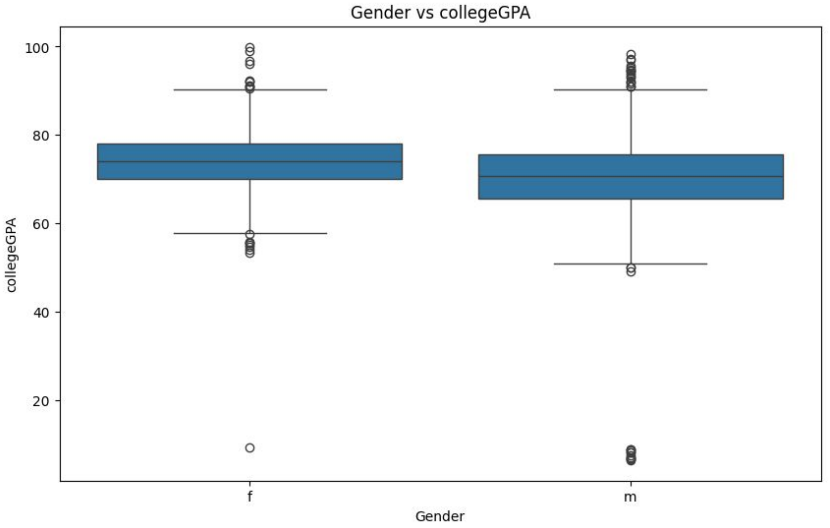
- With scatter plots, you may see how two numerical variables relate to one another and look for trends or connections.
- Analysis of correlation: calculated correlation coefficients to measure the direction and intensity of a linear relationship between numerical variables, such as Pearson correlation

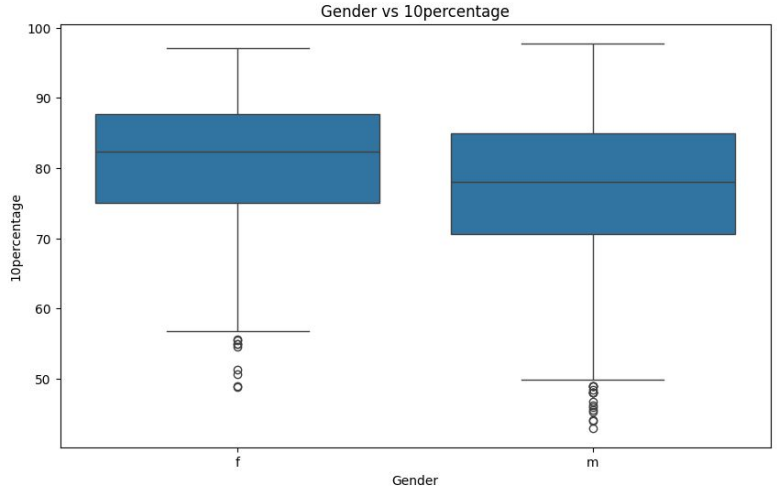
6. Bivariate Analysis Steps

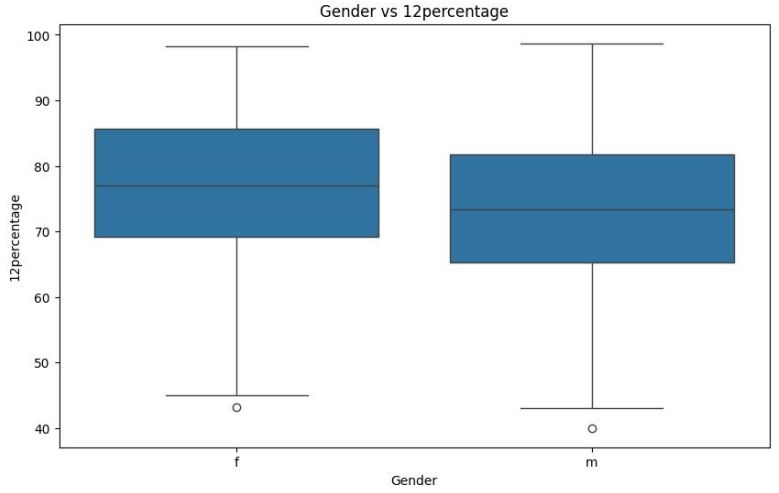
Variation in grades between males and females

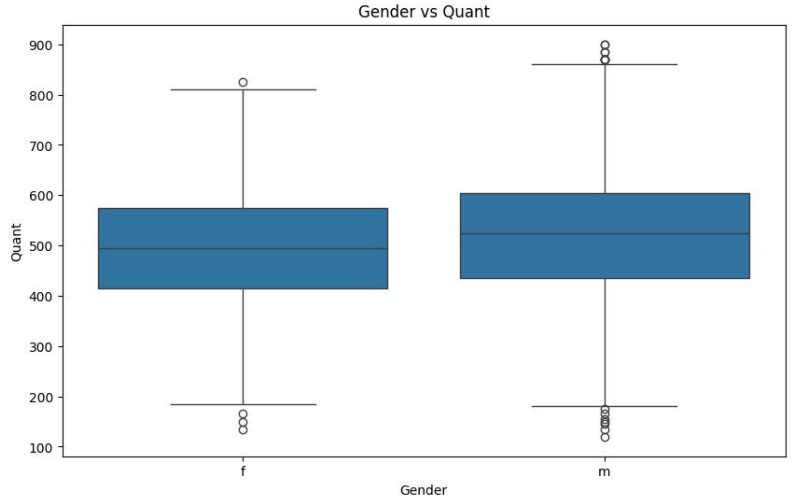








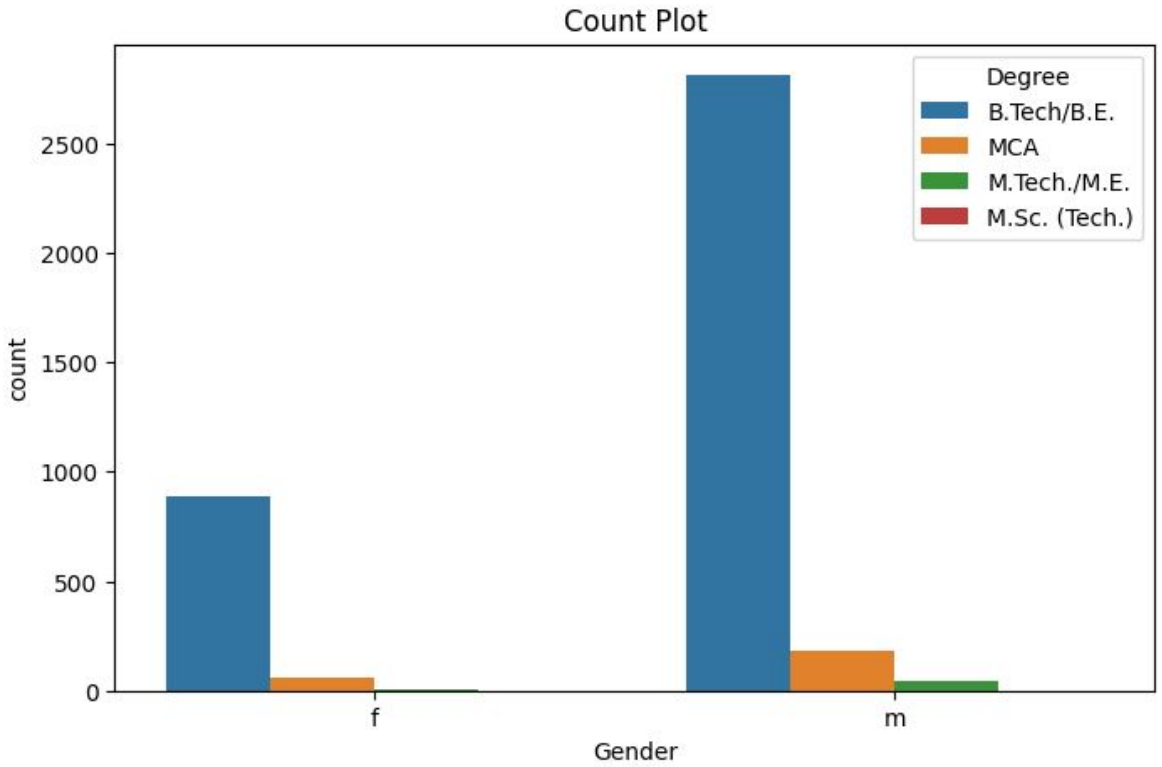


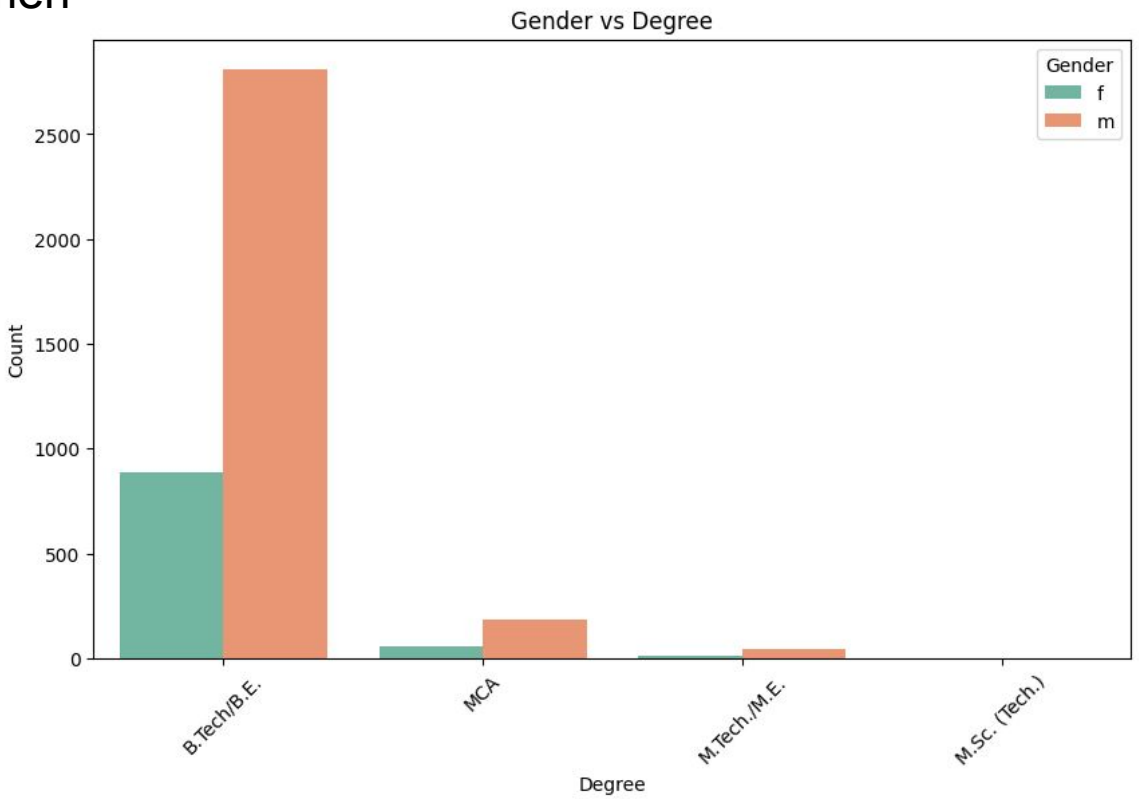


INNOMATICS RESEARCH LABS

6. Bivariate Analysis Steps

- There are more men with a bachelor's degree than women for all four levels of education listed.
- The biggest difference in the number of men and women with a bachelor's degree is for B.Tech/B.E., where there are almost twice as many men as women.
- The smallest difference in the number of men and women with a bachelor's degree is for M.Sc. (Tech.), where there are only about 50 more men than women

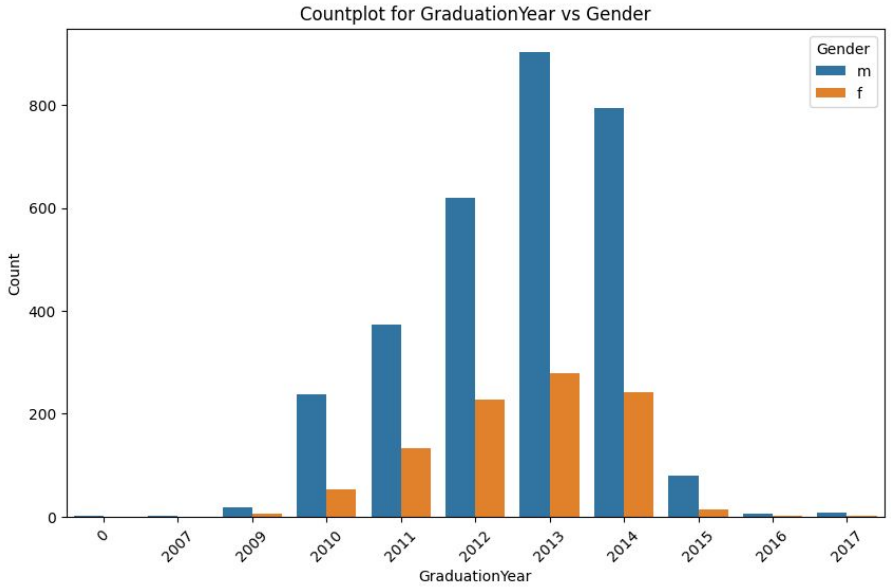


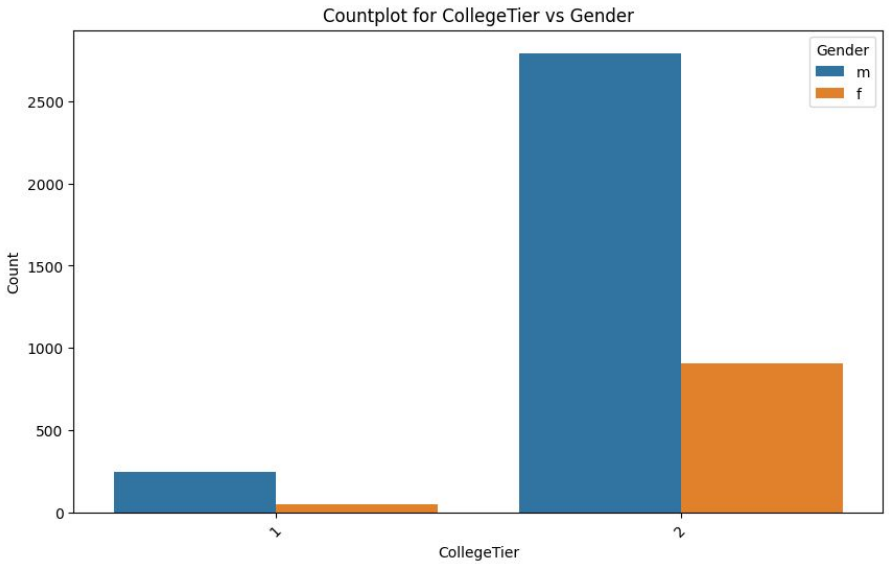


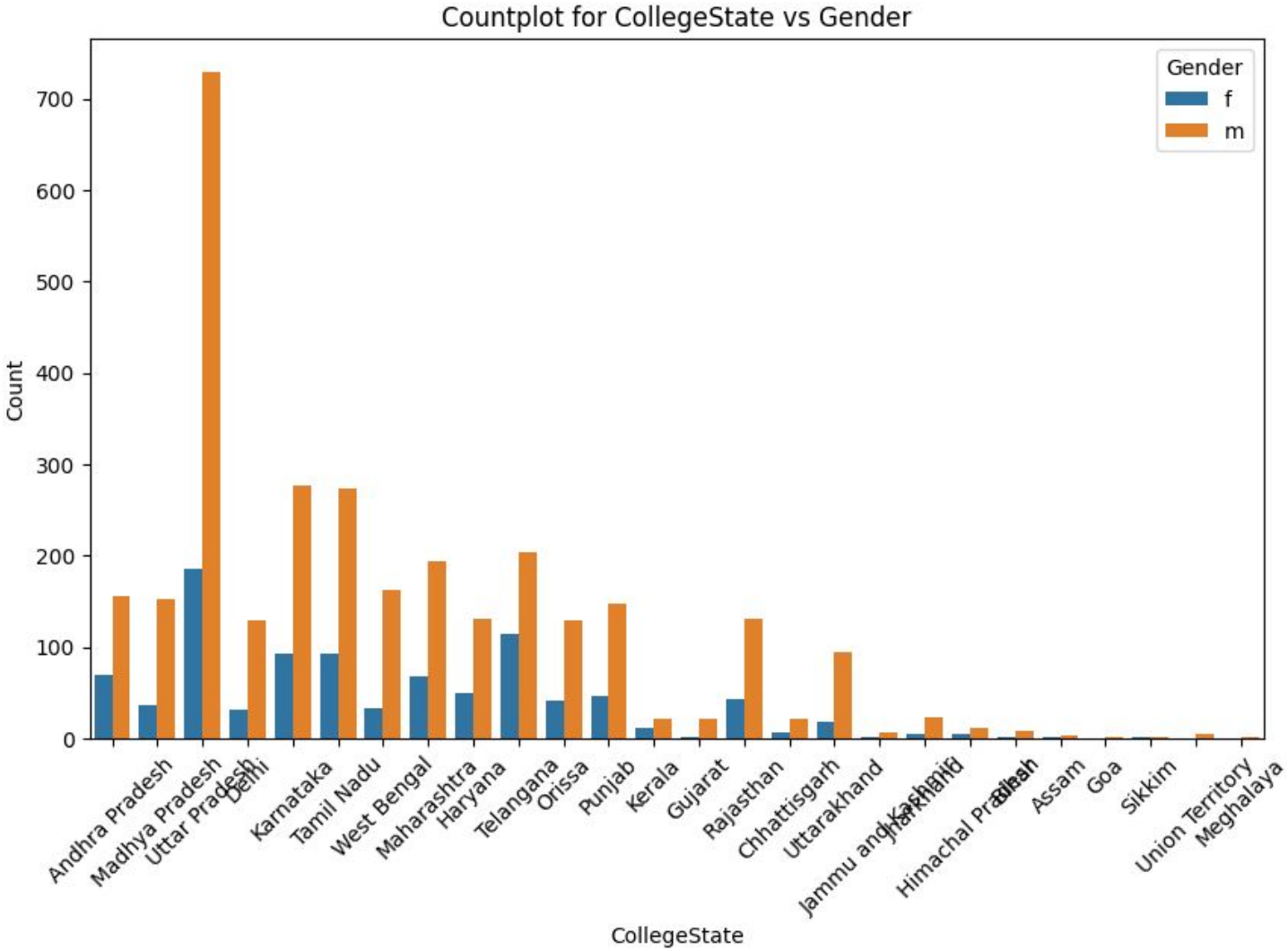
Gender Vs Degree

INNOMATICS RESEARCH LABS

6. Bivariate Analysis Steps







7. Research Question 1

Question 1: 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.

Observation

The average salary of fresh graduates in the specified job roles is higher than the upper limit of the claimed range (2.5-3 lakhs). This suggests that individuals with a background in Computer Science Engineering and employed in roles like Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer may earn more on average than what was stated in the Times of India article.

INNOMATICS RESEARCH LABS

7. Research Question 2

Question 1: Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)

```
from scipy.stats import chi2_contingency
# Step 1: Create Contingency Table
contingency_table = pd.crosstab(data['Gender'], data['Specialization'])

# Step 2: Perform Chi-Squared Test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

# Step 3: Interpret Results
alpha = 0.05  # significance level
print("Chi-Squared Test Results:")
print("Chi-Squared Statistic:", chi2)
print("p-value:", p_value)

if p_value < alpha:
    print("There is a significant relationship between gender and specialization.")
else:
    print("There is no significant relationship between gender and specialization.")
```

```
Chi-Squared Test Results:
Chi-Squared Statistic: 104.46891913608455
p-value: 1.2453868176976918e-06
There is a significant relationship between gender and specialization.
```

Observation

1. **The chi-squared** test results indicate a significant relationship between gender and specialization. This suggests that there are notable differences in specialization preferences between genders within the dataset.

2. **Chi-squared** shows that men are more specialized than women in several areas.

3. Less specialization is preferred by women

INNOMATICS RESEARCH LABS

8. Conclusions

- **The conclusion holds true**: various factors such as academic performance, college tier, specialization, and gender significantly influence salary offers.
- Second, MCA and M.tech graduates receive average salaries among B.tech/B.E graduates, but AMCAT graduates are mostly B.tech/B.E graduates and receive high salaries.
- **AMCAT scores** play a crucial role in determining salary levels.
- **Male** individuals tend to receive higher salaries.
- **Graduates** from top-tier colleges are more likely to secure higher salary packages.
- Experience correlates positively with salary, leading to higher earnings.
- **Impact of College Tier, GPA, and Specialization on Career Prospects:**

  The analysis of college tier, GPA, and specialization reveals their significant impact on salary offers or job placements, shedding light on the influence of college ranking on career prospects.

THANK YOU

INNOMATICS RESEARCH LABS