INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project 1 - Analysis of AMCAT Data

- Santha Lakshmi S

About me

- **Name:** Santha Lakshmi S
- **Education:** B.E. in Electronics and Communication Engineering (ECE), Final Year
- **Institution:** Panimalar Institute of Technology, Chennai

- **Interests:**
  - **Data Science & Analytics:** Passionate about transforming data into actionable insights.
  - **Electronics & Communication:** Keen to explore innovative technologies in IoT and embedded systems.

- **About Me:**
  I am an enthusiastic final-year student with hands-on experience in technology and leadership. My roles in IEEE have sharpened my teamwork, time management, and problem-solving skills. Known for my creative problem-solving and attention to detail, I excel in collaborative environments. I am excited to leverage my technical expertise and leadership experience to contribute to organizations that value creativity and growth.

Why I Want to Learn Data Science

- **Data-Driven Decision Making**:
  - I believe data is the key to making informed decisions in any industry.
  - Understanding data helps in extracting valuable insights for better outcomes.
- **Interdisciplinary Nature**:
  - Data Science merges analytical skills with domain knowledge from ECE.
  - It provides opportunities to apply technical skills in a new and exciting field.
- **Career Opportunities**:
  - Opens up a wide range of career paths in data analysis, machine learning, and AI.
  - Enhances my technical expertise and employability.

Work Experience

Internships:

- **Data Science Intern** at **Tamil Nadu Skill Development Corporation**
  Conducted data analysis projects to enhance training programs, focusing on skill development metrics and outcomes. Developed skills in data cleaning, manipulation, and analysis to derive actionable insights.

Projects:

- Conducted Exploratory Data Analysis (EDA) on various datasets using Python, Pandas, and visualization tools such as Matplotlib and Seaborn.
- Developed hands-on skills in data cleaning, manipulation, and analysis, enhancing my ability to derive insights from complex datasets.

LinkedIn & GitHub Profiles

- LinkedIn: https://www.linkedin.com/in/santha-lakshmi-s/
- GitHub: https://github.com/Santha-Lakshmi-S

INNOMATICS RESEARCH LABS

Agenda

- Business Problem & Objective of the Project
- Objective of the Project
- Data Collection Process
- Summary of the Data
- Data Overview
- Univariate Analysis
- Bivariate Analysis
- Conclusion
- My Experience and Challenges

Business Problem & Objective of the Project

Business Problem

- Understanding factors affecting salaries of AMCAT candidates.
- Identifying trends related to education and demographic variables.

Use Case Domain

- Employment and Recruitment Industry.

Main Objectives

- Analyze salary distribution among candidates.
- Explore relationships between educational background and salary.
- Provide insights that can assist HR and recruitment agencies.

Data Collection Process

Process Steps

- Identifying data sources.
- Scraping the data.
- Storing data in structured format (CSV).

Summary of the Data

- **Data Structure**
  - Total Records: 3998
  - Total Features: 39
- **Key Features**
  - Salary, Designation, Job City, Gender, Education Percentages.

Data Given

```
print(df.head())
```

```
  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]
```

```
[4] print(df.shape)
```

```
(3998, 39)
```

```
[5] print(df.describe())
```

```
                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]
```

INNOMATICS RESEARCH LABS

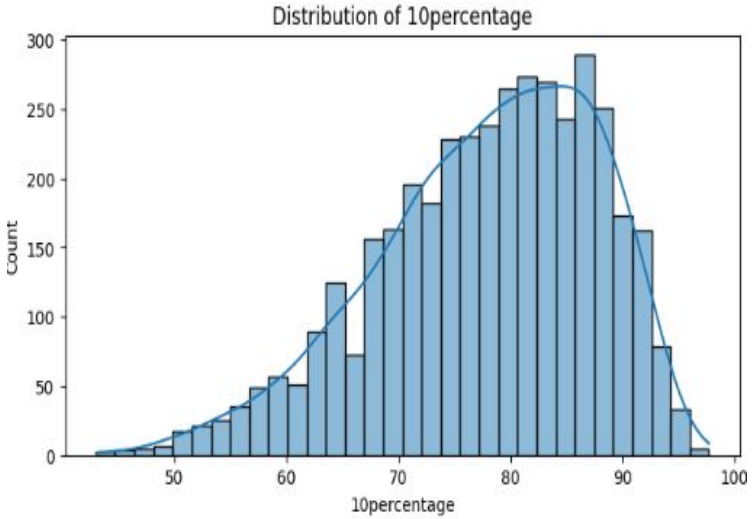
Data Visualization

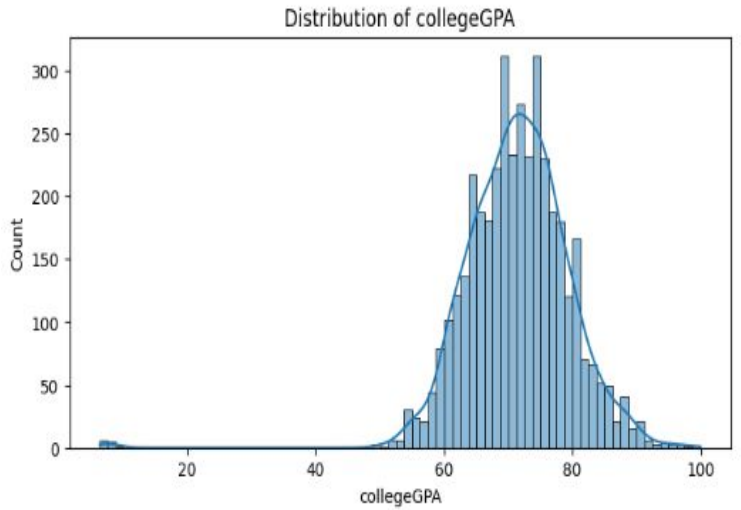
- Univariate Analysis

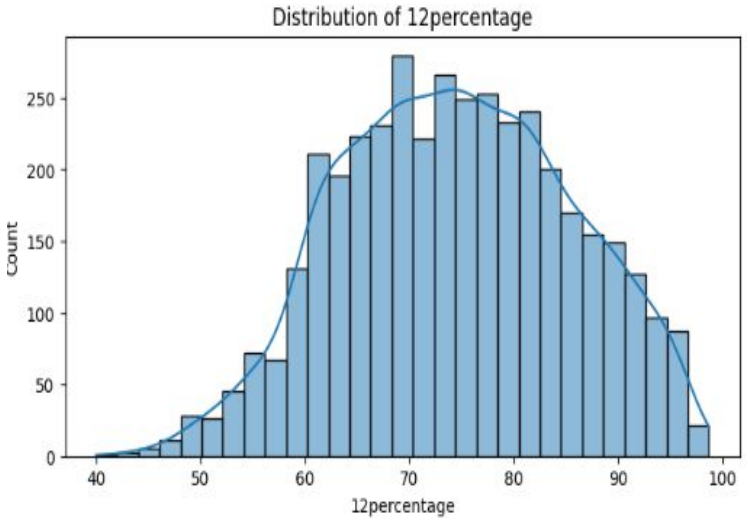
Numerical Variables Distribution

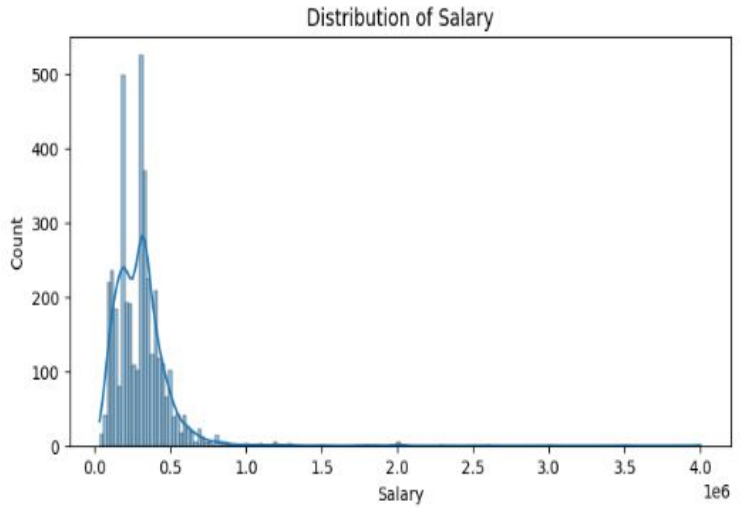
Histograms with KDE for each numerical variable.

Importance of visualizing distribution to understand data patterns.









INNOMATICS RESEARCH LABS

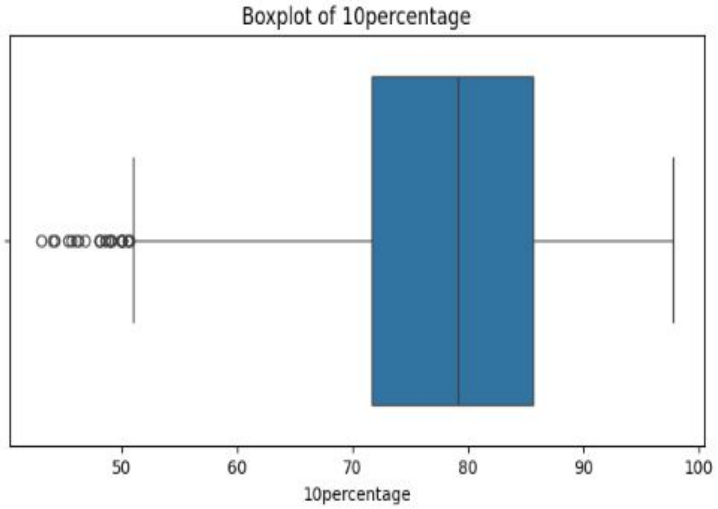
Data Visualization

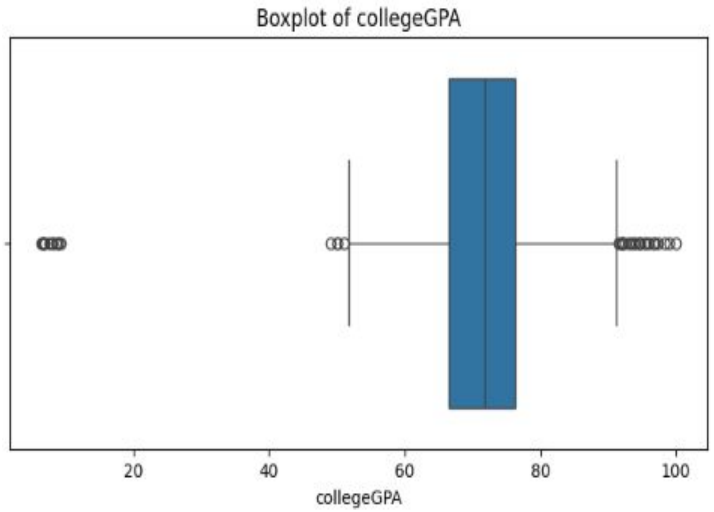
- Univariate Analysis

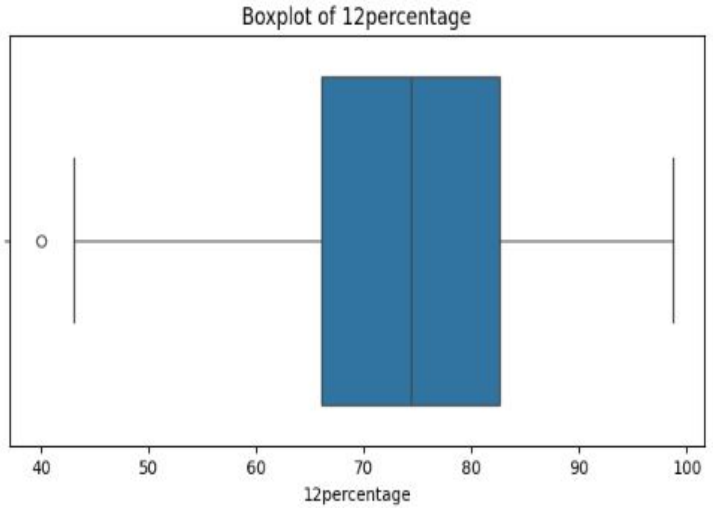
Boxplots of Numerical Variables

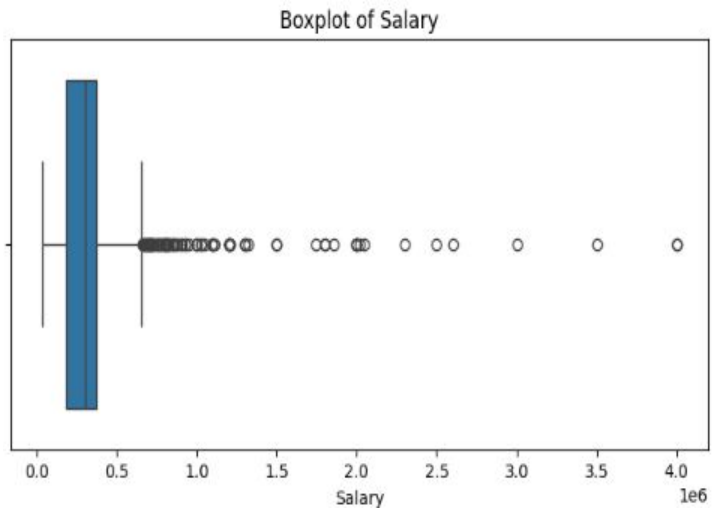
Use boxplots to identify outliers and spread of numerical data.

Importance of boxplots in descriptive statistics.









INNOMATICS RESEARCH LABS

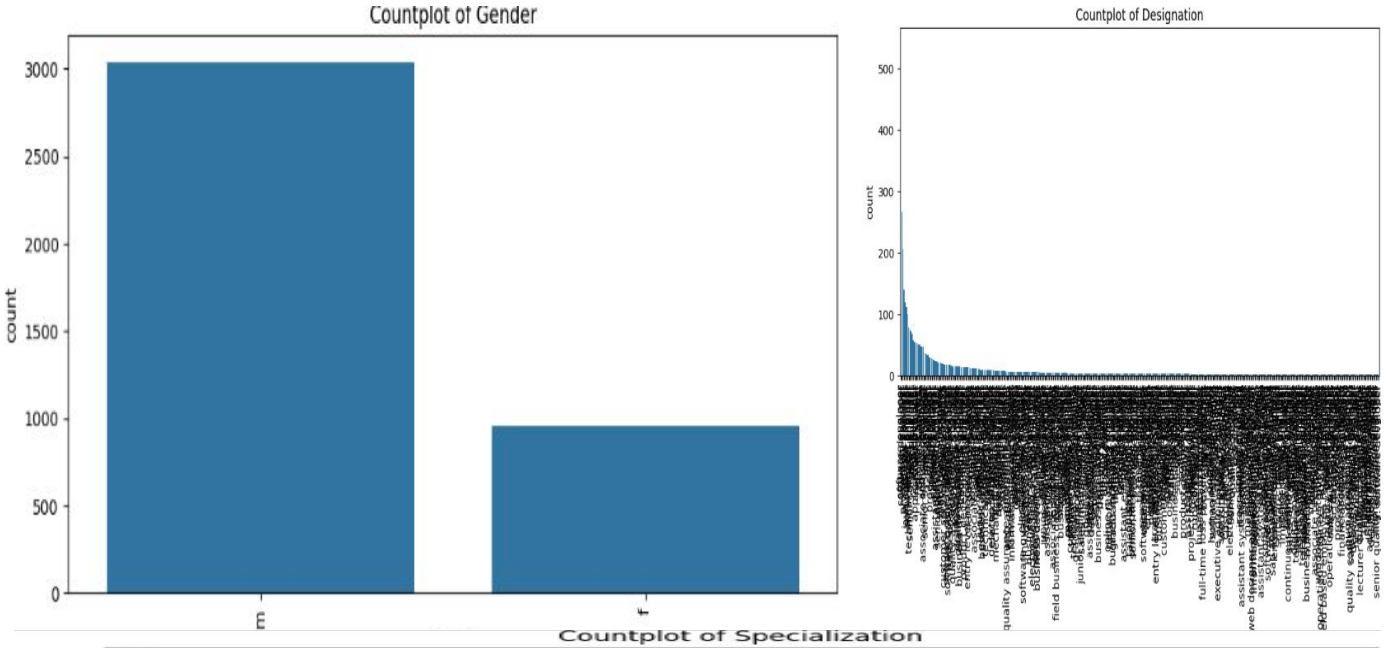
Data Visualization

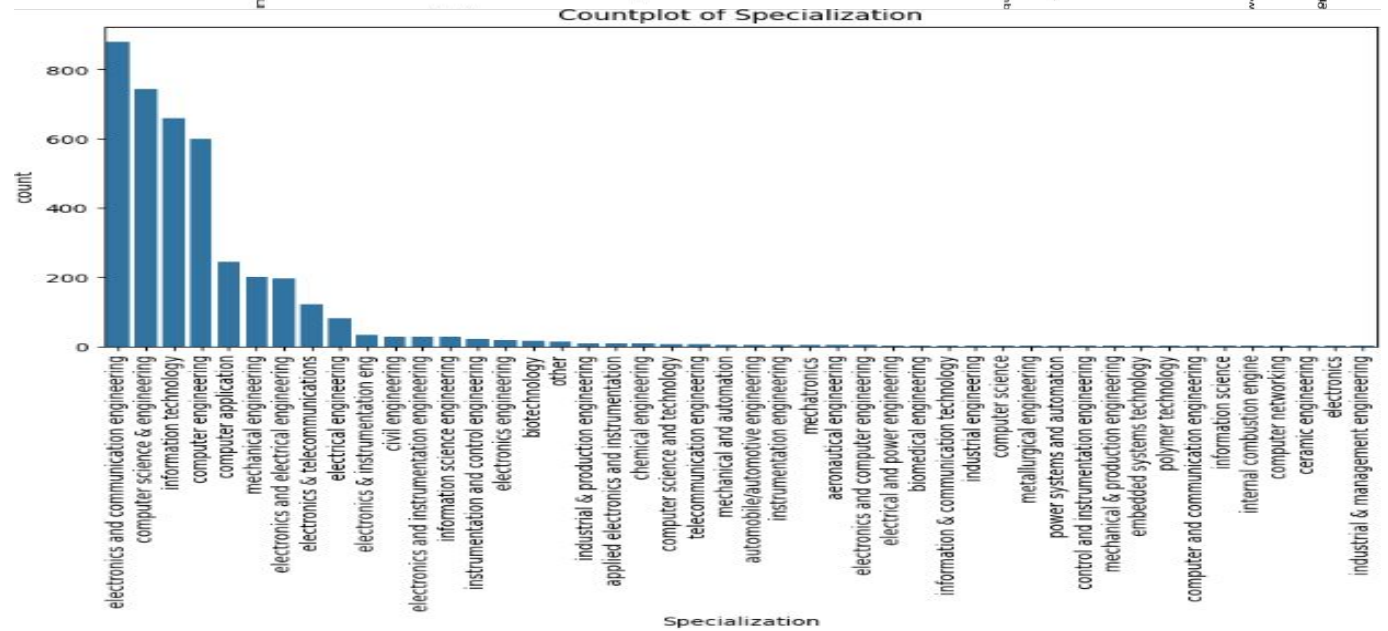
- Univariate Analysis

Counplot of Categorical Variables Analysis

Count Plots show the frequency of categories.

Useful for understanding distribution among categorical variables.





INNOMATICS RESEARCH LABS

Data Cleaning
- Bivariate Analysis

Data Cleaning Steps

Convert numerical columns to numeric data type.

Handle missing values to ensure clean data for analysis.

Bivariate Analysis

```
[13] for col in numerical_columns + ['Salary']:
         df[col] = pd.to_numeric(df[col], errors='coerce')
```

```
[14] df_cleaned = df[numerical_columns + ['Salary']].dropna()
```

```
# Display data types and the first few rows to identify any problematic columns
print(df[numerical_columns + ['Salary']].dtypes)
print(df[numerical_columns + ['Salary']].head())
```

```
10percentage    float64
12percentage    float64
collegeGPA      float64
Salary          float64
English           int64
Logical           int64
Quant             int64
Domain          float64
Salary          float64
dtype: object
   10percentage  12percentage  collegeGPA     Salary  English  Logical  Quant  \
0          84.3          95.8       78.00   420000.0      515      585    525
1          85.4          85.0       70.06   500000.0      695      610    780
2          85.0          68.2       70.00   325000.0      615      545    370
3          85.6          83.6       74.64  1100000.0      635      585    625
4          78.0          76.8       73.90   200000.0      545      625    465

     Domain     Salary
0  0.635979   420000.0
1  0.960603   500000.0
2  0.450877   325000.0
3  0.974396  1100000.0
4  0.124502   200000.0
```

```
[17] # Convert all values to numeric, coercing errors to NaN
     for col in numerical_columns + ['Salary']:
         df[col] = pd.to_numeric(df[col], errors='coerce')

     # Drop rows with NaN values after converting
     df_cleaned = df[numerical_columns + ['Salary']].dropna()
```

```
[18] print(df_cleaned.shape)  # Should be (n, m) where n is the number of samples and m is the number of columns
```

```
(3998, 9)
```

```
[22] # Create the subset of numerical columns including 'Salary'
     df_subset = df[numerical_columns + ['Salary']]
```

```
[23] # Remove duplicate columns in the subset
     df_subset = df_subset.loc[:, ~df_subset.columns.duplicated()]
```

INNOMATICS RESEARCH LABS

Data Visualization

- Bivariate Analysis

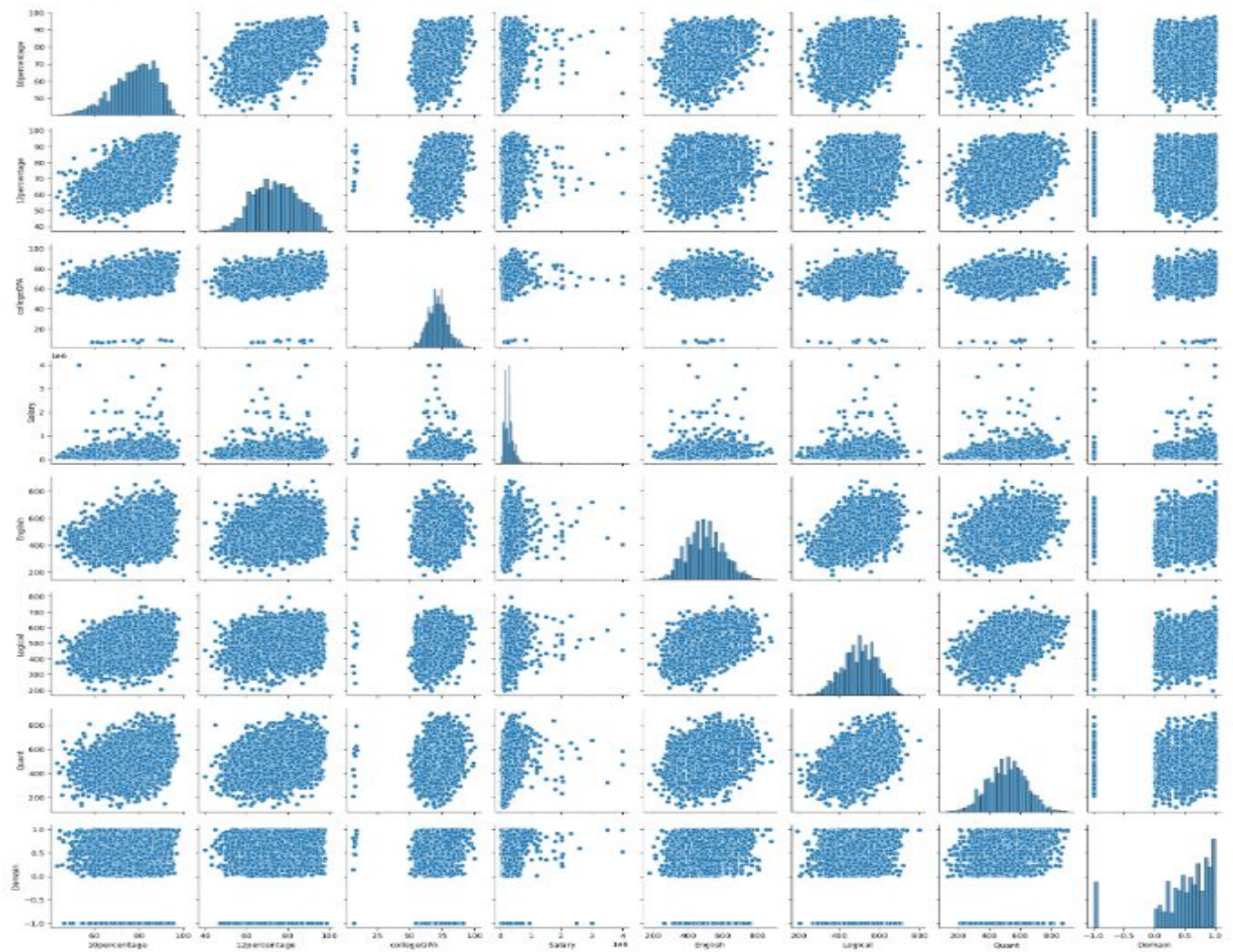
Pairplot Visualization of Numerical Variables

Pairplot visualizes relationships between numerical variables and Salary.

Helps identify potential correlations.

```
import seaborn as sns
import matplotlib.pyplot as plt

# Create the pairplot with the cleaned data
sns.pairplot(df_subset)
plt.show()
```



INNOMATICS RESEARCH LABS

Data Visualization

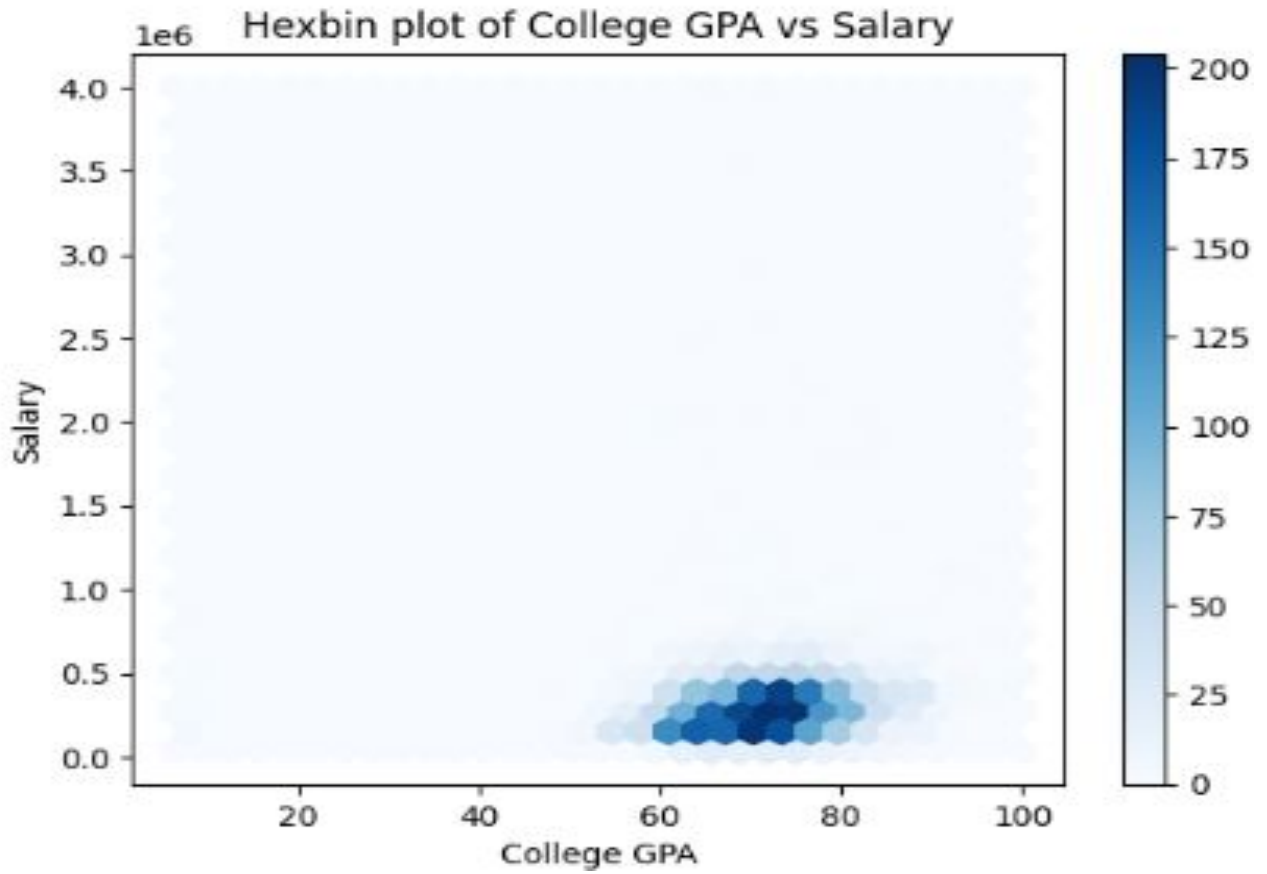
- Bivariate Analysis

Hexbin Plot of College GPA vs Salary

Shows the relationship between collegeGPA and Salary.

Density of data points helps identify trends.

```
plt.hexbin(df['collegeGPA'], df['Salary'], gridsize=30, cmap='Blues')
plt.colorbar()
plt.xlabel('College GPA')
plt.ylabel('Salary')
plt.title('Hexbin plot of College GPA vs Salary')
plt.show()
```



Data Visualization

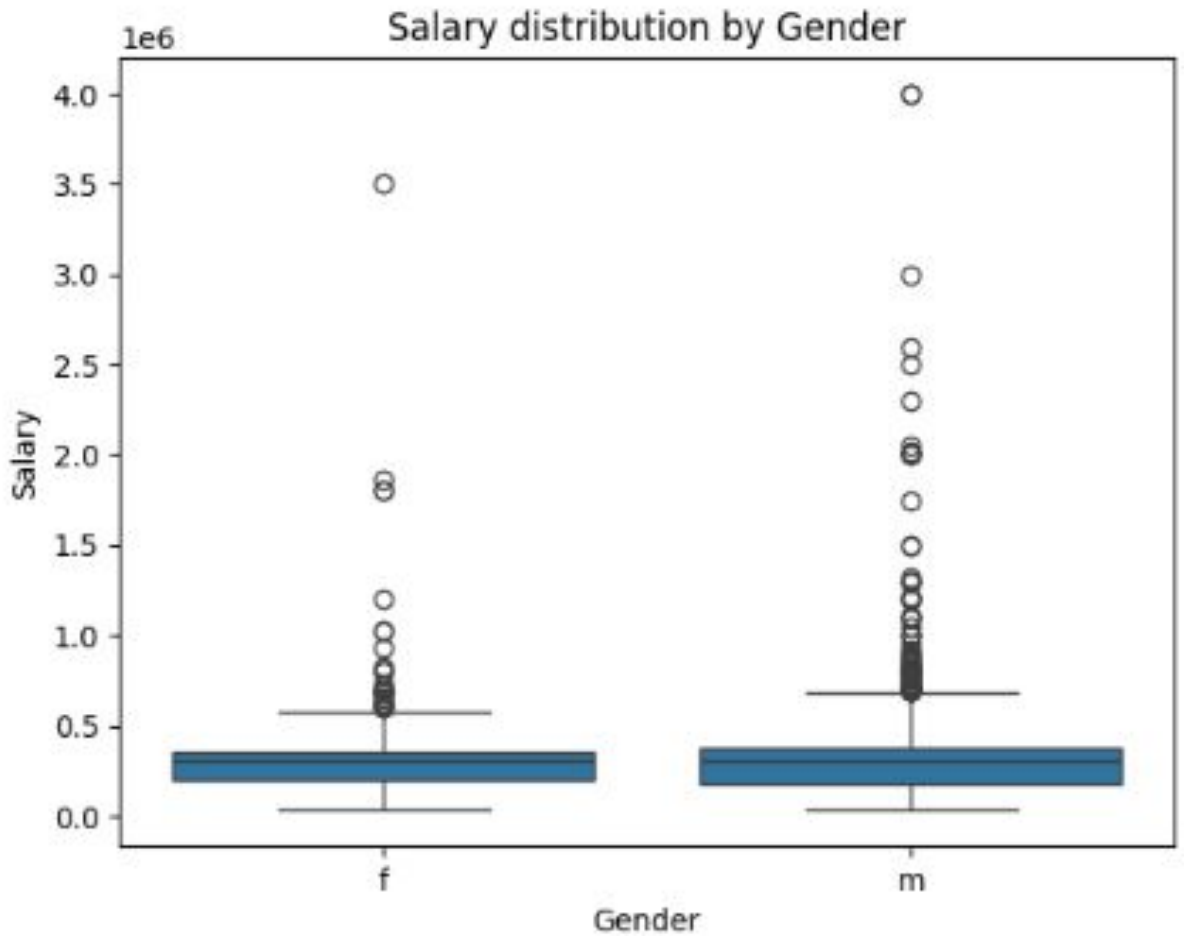
- Bivariate Analysis

Salary Distribution by Gender

Boxplot shows salary distribution across genders.

Helps identify disparities in salary based on gender.

```
sns.boxplot(x='Gender', y='Salary', data=df)
plt.title('Salary distribution by Gender')
plt.show()
```



INNOMATICS RESEARCH LABS

Data Visualization

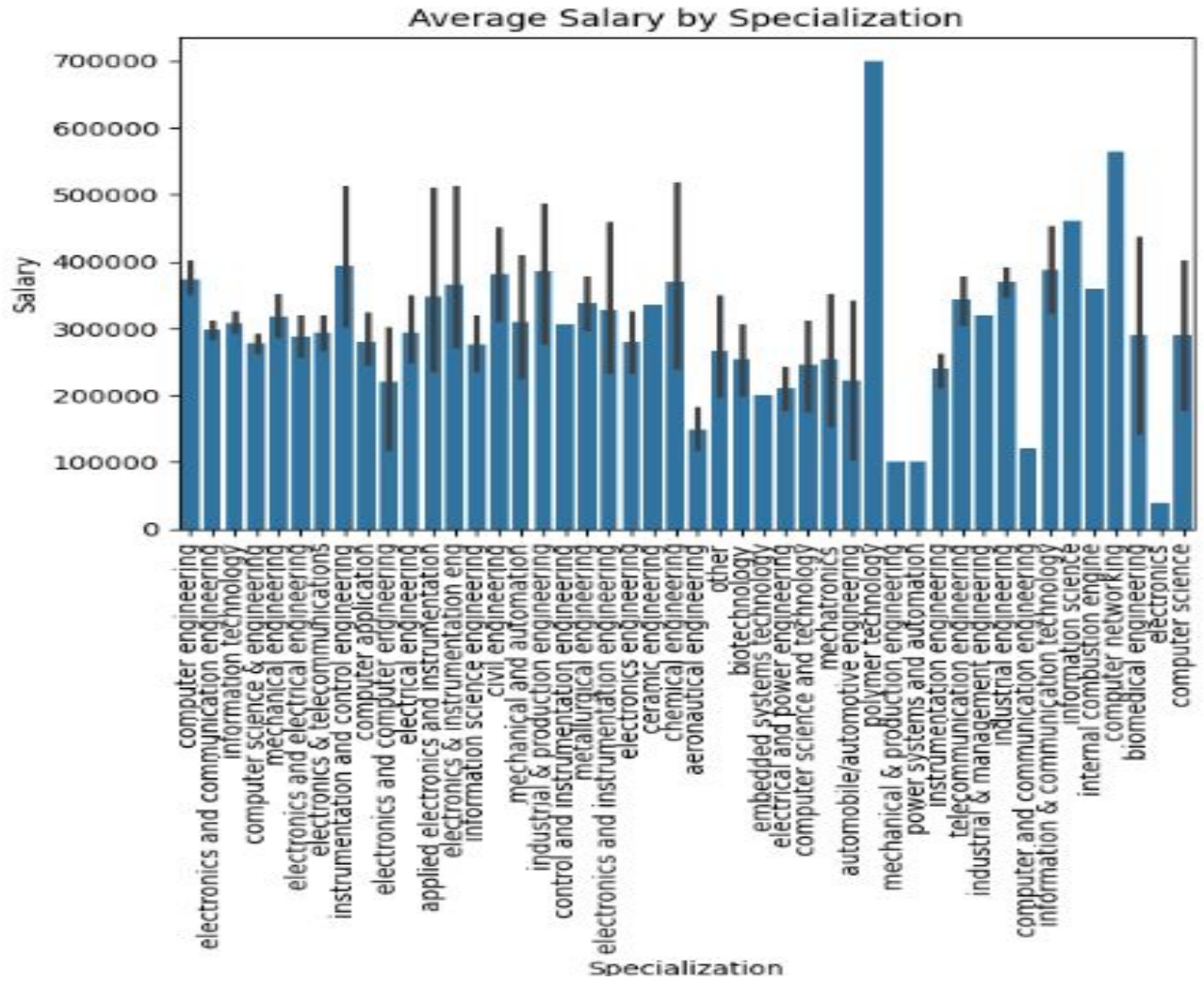
- Bivariate Analysis

Average Salary by Specialization

Barplot shows average salaries based on specialization.

Highlights trends and salary potential in various fields.

```
sns.barplot(x='Specialization', y='Salary', data=df)
plt.xticks(rotation=90)
plt.title('Average Salary by Specialization')
plt.show()
```



INNOMATICS RESEARCH LABS

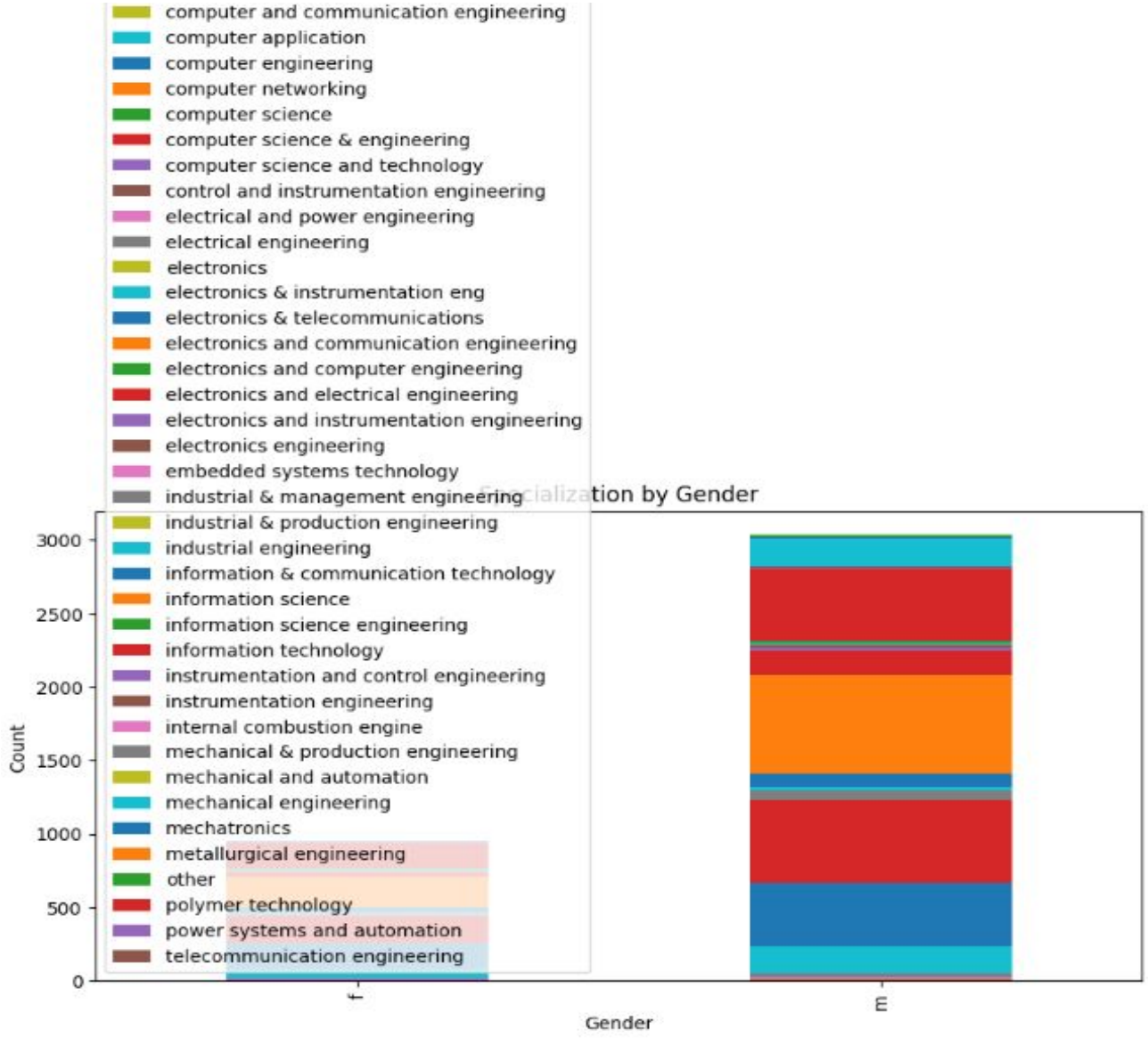
Data Visualization

- Bivariate Analysis

Stacked Bar Plot Analysis

Stacked bar plot visualizes specialization distribution by gender.

Useful for understanding gender representation in specializations.



INNOMATICS RESEARCH LABS

Research Questions

- **Average Salary for Relevant Job Roles.**
  - Filtered dataset to find average salary for computer science-related roles.
  - Average Salary calculated: **302,995.39**
- **Chi-square Test on the relationship between Gender and Specialization.**
  - Explanation of the Chi-square test and its significance.
  - Chi-square Test p-value: **1.25e-06**, indicating a significant relationship.

Research Questions

```
[ ] cs_jobs = df[(df['Specialization'].str.contains('computer science', case=False)) &
                 (df['Designation'].str.contains('analyst|engineer', case=False))]
    avg_salary = cs_jobs['Salary'].mean()
    print(f"Average Salary for relevant job roles: {avg_salary}")
```

```
Average Salary for relevant job roles: 302995.3917050691
```

```
[ ] from scipy.stats import chi2_contingency

    contingency_table = pd.crosstab(df['Gender'], df['Specialization'])
    chi2, p, dof, expected = chi2_contingency(contingency_table)
    print(f"Chi-square Test p-value: {p}")
```

```
Chi-square Test p-value: 1.2453868176976918e-06
```

INNOMATICS RESEARCH LABS

Conclusion

Summary of Key Findings

1. **Univariate Analysis**:
   - Numerical variables such as 10percentage, 12percentage, and collegeGPA exhibit normal distributions, while Salary shows a right skew, indicating high earners.
   - Boxplots revealed outliers in Salary, prompting further investigation into pay disparities.
2. **Bivariate Analysis**:
   - **Correlation**: Higher collegeGPA generally correlates with higher Salary.
   - **Salary by Gender**: Boxplot analysis indicated a disparity in salaries between genders.
   - **Specialization Impact**: Average salaries varied significantly across specializations, with fields like Computer Science offering higher salaries.
   - **Chi-Square Test**: A significant relationship between Gender and Specialization was observed.

My Experience and Challenges

Experience:

- Gained practical skills in data cleaning, visualization, and insight generation using Python.
- Familiarized with the AMCAT dataset, enhancing analytical thinking.
- Collaborated with peers to refine analytical approaches.

Challenges:

- Addressed data quality issues like missing values and outliers.
- Improved skills in creating effective visualizations for clear communication.
- Navigated statistical tests, enhancing understanding of inferential statistics.
- Managed time effectively amidst academic responsibilities.

THANK YOU

INNOMATICS RESEARCH LABS