INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

AMCAT ANALYSIS

OBJECTIVE

- The AMCAT (Aspiring Minds Computer Adaptive Test) Analysis dataset is designed to evaluate and analyze the skills and competencies of job seekers.

- Its primary objective includes:
- **Skill Assessment**: Measure candidates' proficiency in various domains
- **Job Matching**: Help employers match candidates with suitable job roles based on their skill sets and performance.
- **Talent Insights**: Provide insights into the overall skill landscape of job seekers, identifying trends and areas for improvement.

- **Educational Background**: Offer feedback to educational institutions about the preparedness of their graduates for the job market.

- **Enhancing Recruitment Processes**: Streamline the recruitment process for companies by using data-driven insights to make informed hiring decisions.

ABOUT THE DATASET

- The dataset is sourced from the Aspiring Minds Employment Outcome (AMEO 2015) study.
- The dataset includes various demographic attributes and comprises approximately 40 independent variables alongside 4,000 data points. These independent variables are a mix of continuous and categorical types, and each candidate is assigned a unique identifier.
- Below is the head of the dataset and a few columns to get the idea of the data:

Summary of the dataset:

- Candidate Overview with a unique ID. - IN9240176
- Salary and Employment Dates: Includes annual salary offered, dates of joining (DOJ), and dates of leaving (DOL).

- Job Details: Captures job title (Designation) and job location (JobCity).

- **Background –**

  Pursuing a Master of Computer Applications (MCA), and you already hold a Bachelor of Computer Applications (BCA) degree.

- **Why you want to learn Data Science - Interest in Data Science, developed an interest in data analysis and machine learning, and you have knowledge of popular Python libraries such as Pandas, NumPy, and other data analysis libraries. You are continuously gaining hands-on experience in this field through various projects and self-study.**
- **Share your linkedin : https://www.linkedin.com/in/sarthak-hirpurkar2863/**
- **github profile urls : https://github.com/sarthak2863**

[9]:
```
import pandas as pd
df = pd.read_csv("C:\\Users\\sarth\\Downloads\\AMCAT.csv")
print(df)
```

```
     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
...         ...     ...        ...           ...           ...
3993      train   47916   280000.0  10/1/11 0:00  10/1/12 0:00
3994      train  752781   100000.0   7/1/13 0:00   7/1/13 0:00
3995      train  355888   320000.0   7/1/13 0:00       present
3996      train  947111   200000.0   7/1/14 0:00   1/1/15 0:00
3997      train  324966   400000.0   2/1/13 0:00       present

                     Designation           JobCity Gender           DOB  \
0         senior quality engineer         Bangalore      f  2/19/90 0:00
1               assistant manager            Indore      m  10/4/89 0:00
2                systems engineer           Chennai      f   8/3/92 0:00
3        senior software engineer           Gurgaon      m  12/5/89 0:00
4                             get           Manesar      m  2/27/91 0:00
...                           ...               ...    ...           ...
3993            software engineer        New Delhi       m  4/15/87 0:00
3994             technical writer         Hyderabad      f  8/27/92 0:00
3995  associate software engineer         Bangalore      m   7/3/91 0:00
```

[9]:
```
import pandas as pd
df = pd.read_csv("C:\\Users\\sarth\\Downloads\\AMCAT.csv")
print(df)
```

```
     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
...         ...     ...        ...           ...           ...
3993      train   47916   280000.0  10/1/11 0:00  10/1/12 0:00
3994      train  752781   100000.0   7/1/13 0:00   7/1/13 0:00
3995      train  355888   320000.0   7/1/13 0:00       present
3996      train  947111   200000.0   7/1/14 0:00   1/1/15 0:00
3997      train  324966   400000.0   2/1/13 0:00       present

                    Designation           JobCity Gender           DOB  \
0         senior quality engineer         Bangalore      f  2/19/90 0:00
1               assistant manager            Indore      m  10/4/89 0:00
2                systems engineer           Chennai      f   8/3/92 0:00
3        senior software engineer           Gurgaon      m  12/5/89 0:00
4                             get           Manesar      m  2/27/91 0:00
...                           ...               ...    ...           ...
3993            software engineer        New Delhi       m  4/15/87 0:00
3994             technical writer         Hyderabad      f  8/27/92 0:00
3995  associate software engineer         Bangalore      m   7/3/91 0:00
3996           software developer  Asifabadbanglore      f  3/20/92 0:00
3997      senior systems engineer           Chennai      f  2/26/91 0:00

      10percentage  ... ComputerScience  MechanicalEngg  ElectricalEngg  \
0            84.30  ...              -1              -1              -1
1            85.40  ...              -1              -1              -1
2            85.00  ...              -1              -1              -1
3            85.60  ...              -1              -1              -1
4            78.00  ...              -1              -1              -1
...            ...  ...             ...             ...             ...
3993         52.09  ...              -1              -1              -1
3994         90.00  ...              -1              -1              -1
3995         81.86  ...              -1              -1              -1
3996         78.72  ...             438              -1              -1
3997         70.60  ...              -1              -1              -1
```

[13]:
```
df.describe()
```

[13]:

	ID	Salary	10percentage	12graduation	12percentage	CollegeID	CollegeTier	collegeGPA	CollegeCityID	CollegeCityTier	...	ComputerSc
count	3.998000e+03	3.998000e+03	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	...	3998.0(
mean	6.637945e+05	3.076998e+05	77.925443	2008.087544	74.466366	5156.851426	1.925713	71.486171	5156.851426	0.300400	...	90.7
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933	4802.261482	0.262270	8.167338	4802.261482	0.458489	...	175.2
min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000	2.000000	1.000000	6.450000	2.000000	0.000000	...	-1.0(
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000	494.000000	2.000000	66.407500	494.000000	0.000000	...	-1.0(
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000	2.000000	71.720000	3879.000000	0.000000	...	-1.0(
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000	2.000000	76.327500	8818.000000	1.000000	...	-1.0(
max	1.298275e+06	4.000000e+06	97.760000	2013.000000	98.700000	18409.000000	2.000000	99.930000	18409.000000	1.000000	...	715.0(

8 rows × 27 columns

[15]:
```
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
```

```
---  ------                  --------------  -----
 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
```

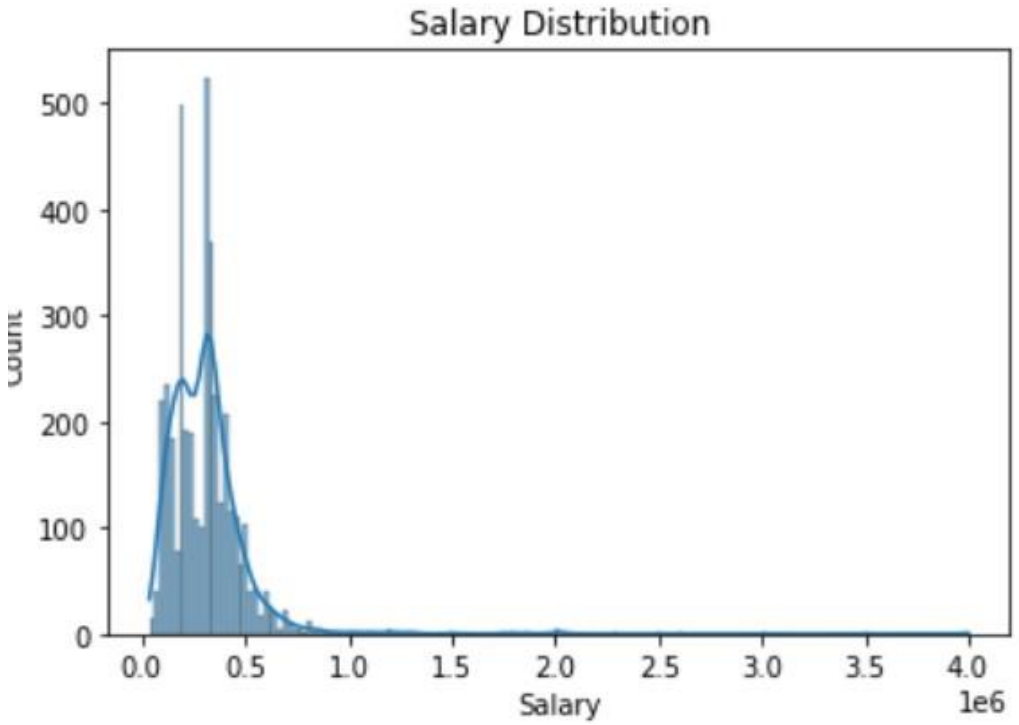
[ ]:

STEPS OF EDA

1. **Introduction:** A detailed description of the dataset and objective outline to analyze the salary distribution and its influencing factors.
2. **Data Import:** Import the dataset and display the first few rows and present the shape of the dataset and summary statistics to grasp its contents.
3. **Univariate Analysis:** Generated histograms, box plots, and count plots for numerical and categorical variables, identifying outliers and examining distributions.
4. **Bivariate Analysis:** Analyze relationships between numerical variables through scatter plots and assess categorical and numerical relationships with box plots and bar plots.
5. **Research Questions:** Test the claim regarding salaries for specific job roles and explore the relationship between 'gender' and 'specialization' using the dataset.
6. **Conclusion:** Summary of key findings from the analysis, highlighting significant trends and insights regarding salary and job roles.
7. **Bonus Insights:** Additional observations and research questions that emerged during the analysis.

UNIVARIATE ANALYSIS

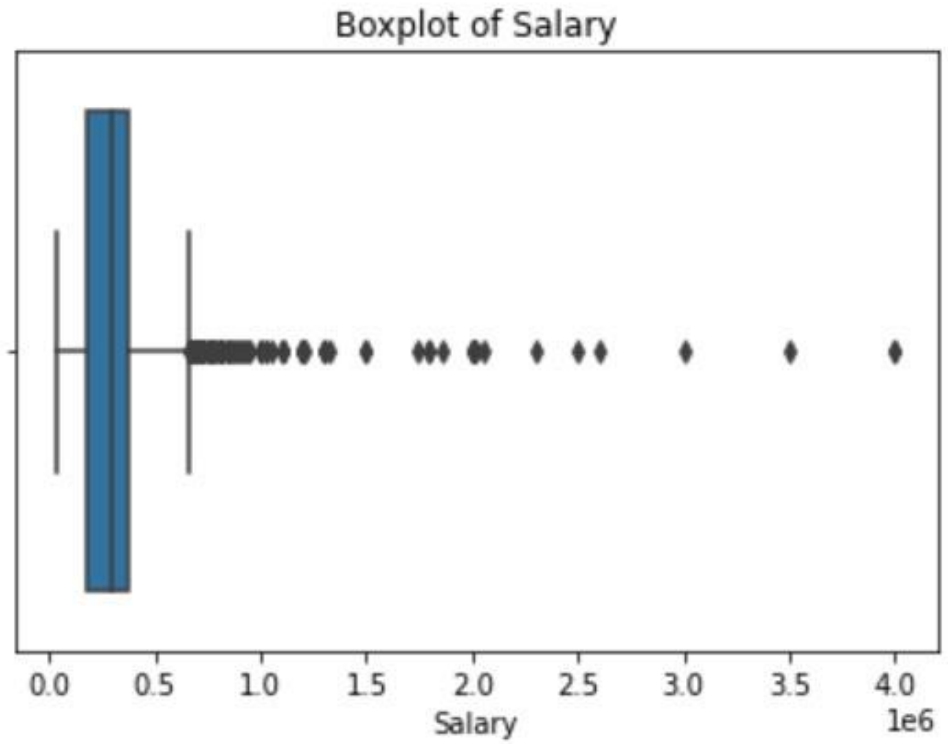
NUMERICAL COLUMNS



T

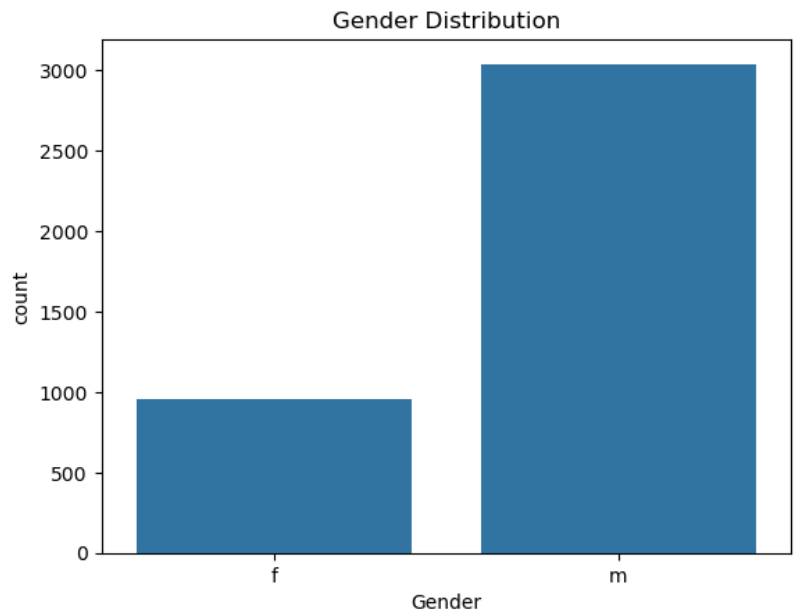
his plot illustrates the distribution of salaries within our dataset, revealing the frequency of various salary ranges and offering insights into the overall shape of the salary distribution.

INNOMATICS RESEARCH LABS



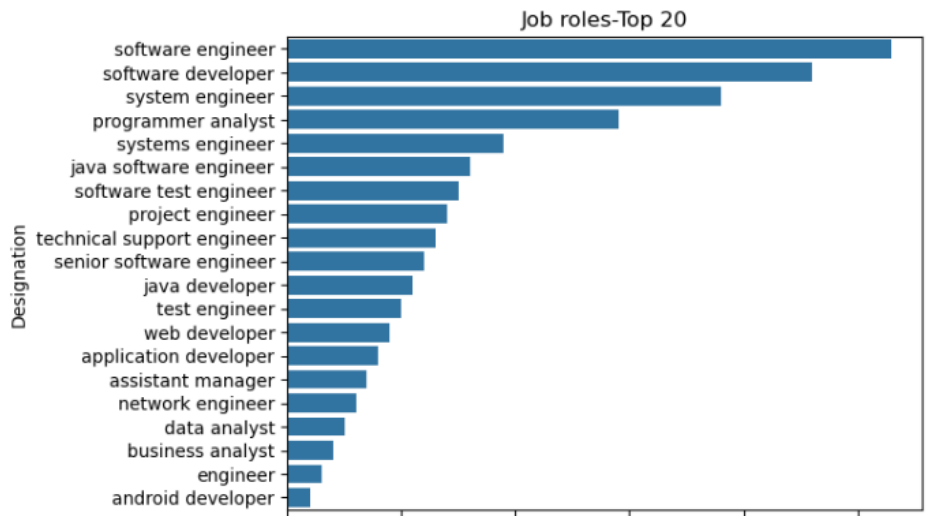
This box plot highlights outliers in the salary data and offers insights into the spread and central tendency of salary values.

[33]:
```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load your dataset
data = pd.read_csv("C:\\Users\\sarth\\Downloads\\AMCAT.csv")  # Make sure the path is correct

# Create a count plot for the 'Gender' column
sns.countplot(x='Gender', data=data)

# Add a title to the plot
plt.title('Gender Distribution')

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



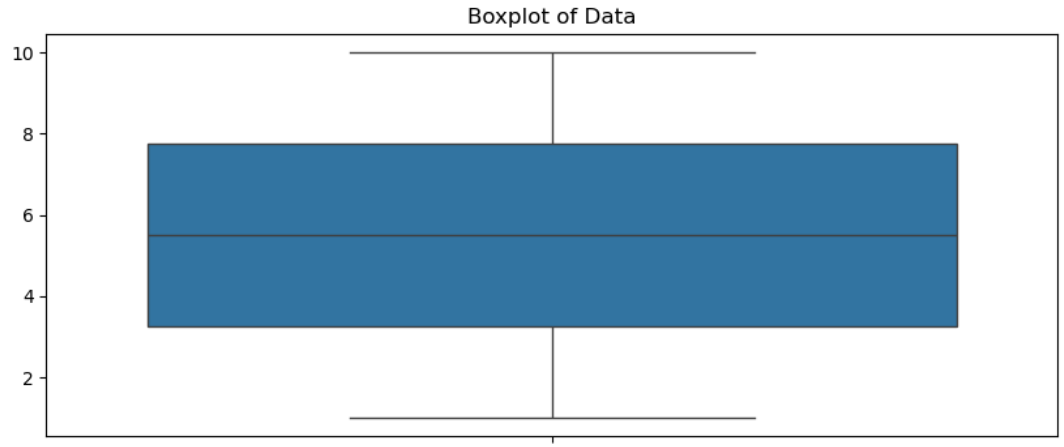
[37]:

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

# Sample data
data = {'Designation': ['software engineer', 'software developer', 'system engineer', 'programmer analyst', 'systems engineer',
                        'java software engineer', 'software test engineer', 'project engineer', 'technical support engineer',
                        'senior software engineer', 'java developer', 'test engineer', 'web developer', 'application developer',
                        'assistant manager', 'network engineer', 'data analyst', 'business analyst', 'engineer',
                        'android developer'],
        'count': [530, 460, 380, 290, 190, 160, 150, 140, 130, 120, 110, 100, 90, 80, 70, 60, 50, 40, 30, 20]}

df = pd.DataFrame(data)

# Create the bar plot
sns.barplot(x='count', y='Designation', data=df, orient='h')

# Set the title of the plot
plt.title('Job roles-Top 20')

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



This count plot effectively summarizes and visualizes the This visualisation helps in understanding hiring trends, popular job roles, and distribution of gender within the data potentially guiding decisions related to recruitment or talent acquisition strategies.

[43]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Sample data
data = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

# Create a boxplot
plt.figure(figsize=(10, 4))
sns.boxplot(data)

# Set title
plt.title('Boxplot of Data')

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



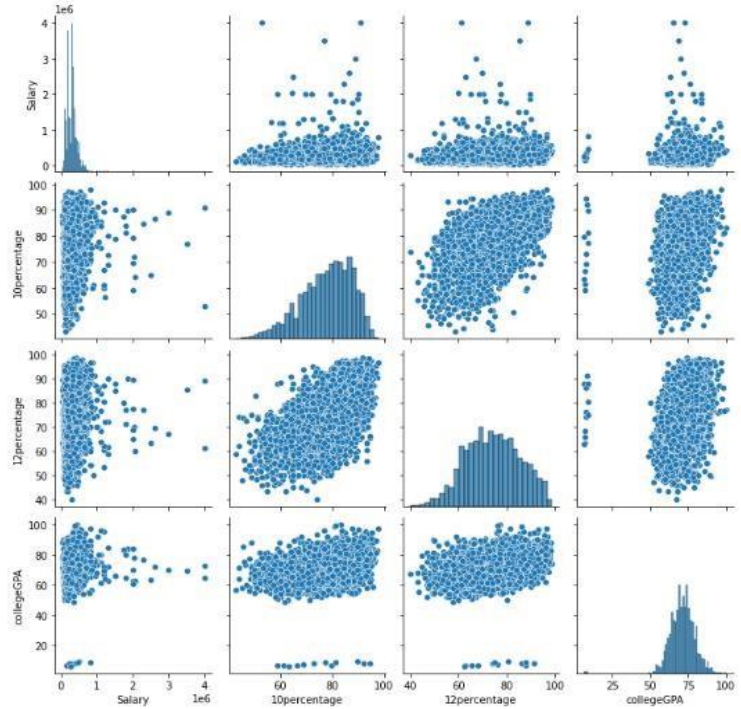
BIVARIATE ANALYSIS

NUMERICAL-NUMERICAL COLUMNS

```
[51]: import matplotlib.pyplot as plt
      import numpy as np

      # Generate some random data for age and salary
      np.random.seed(0)
      age = np.random.uniform(27.5, 47.5, 100)
      salary = np.random.uniform(0.0, 4.0, 100)

      # Create the scatter plot
      plt.scatter(age, salary)

      # Set title and labels
      plt.title('Age vs Salary')
      plt.xlabel('Age')
      plt.ylabel('Salary')

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

The scatter plot visualizes the relationship between a candidate's age and their salary, allowing you to observe trends,

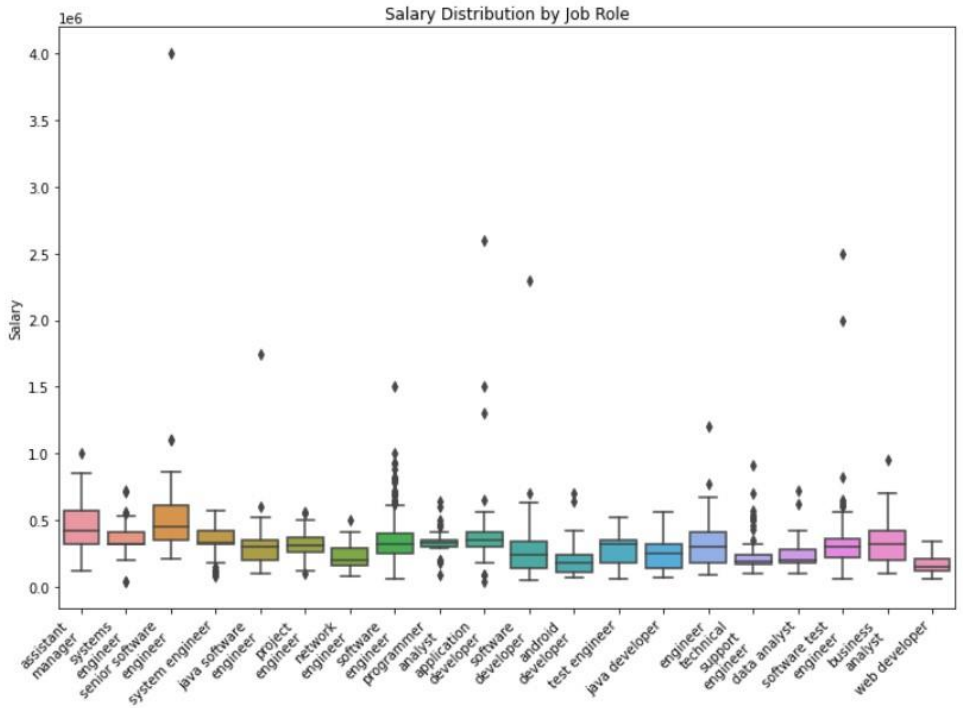
```
In [30]: sns.pairplot(data[['Salary', '10percentage', '12percentage', 'collegeGPA']])
         plt.show()
```



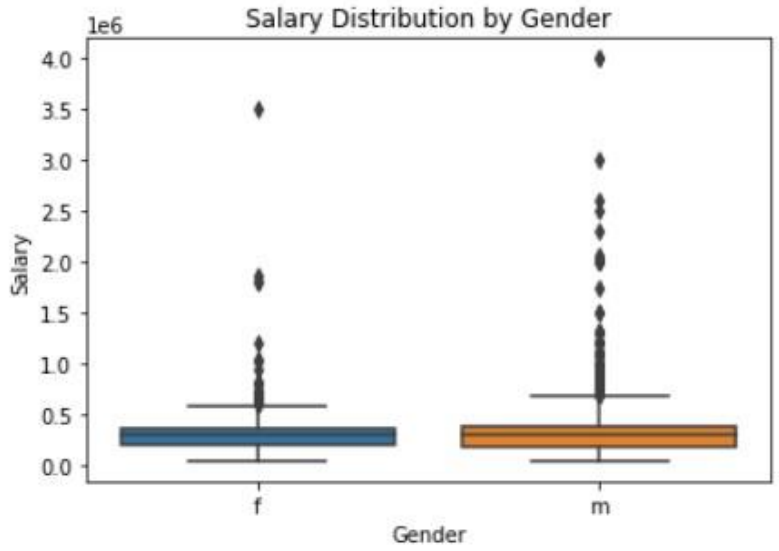
The pair plot allows for an exploratory analysis of how the Salary correlates with academic scores (10th percentage, 12th percentage, and college GPA)

CATEGORICAL-NUMERICAL COLUMNS

INNOMATICS RESEARCH LABS

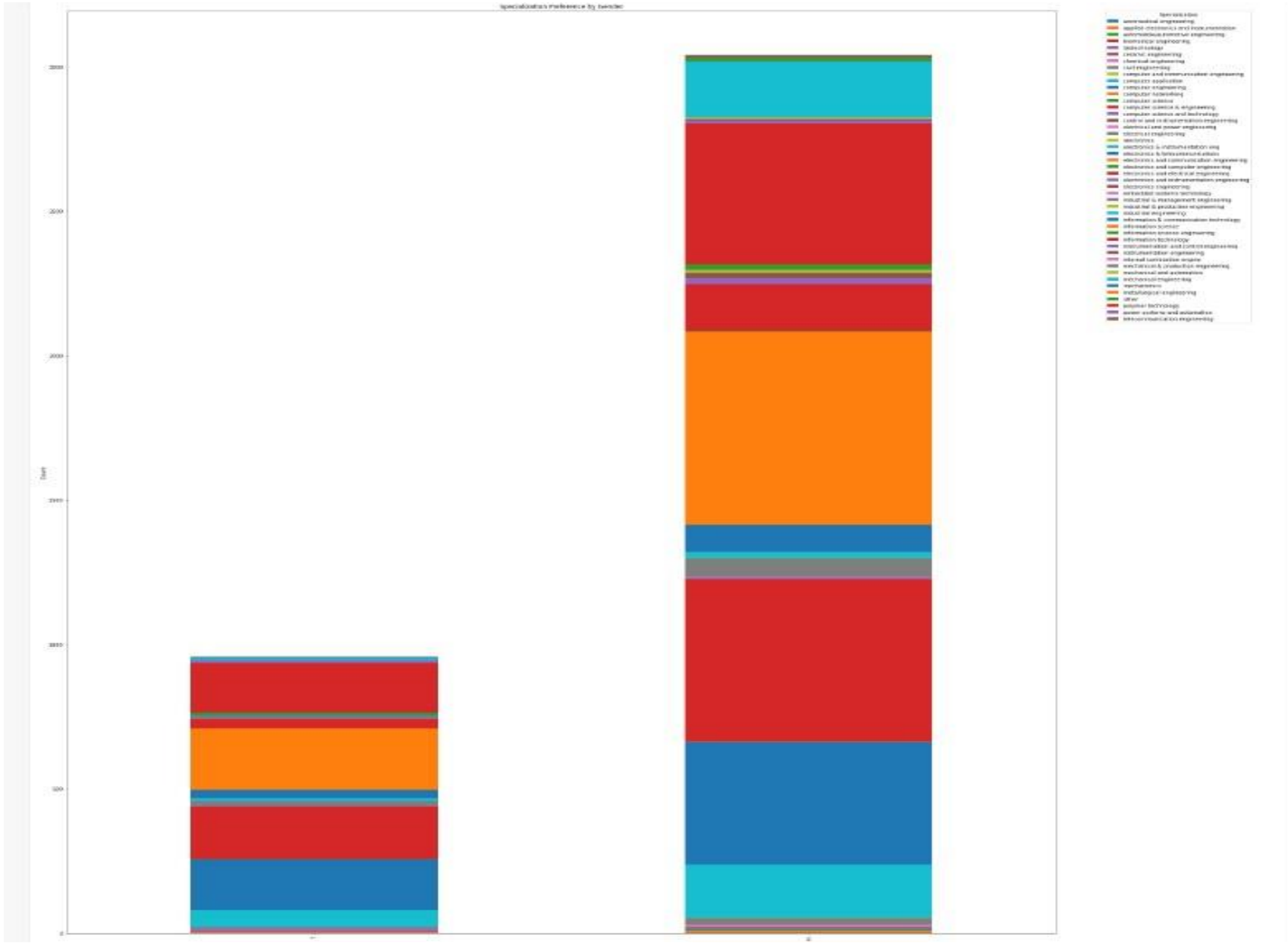



```
In [42]: sns.boxplot(x='Gender', y='Salary', data=data)
         plt.title('Salary Distribution by Gender')
         plt.show()
```



The box plot allows for an easy comparison of salary distributions across The box plot allows for quick insights into the variability of salaries different job designations, enabling insights into which roles tend to within each gender group, revealing if one gender has a wider range of offer higher or lower salaries on average. salaries compared to the other.

CATEGORICAL- CATEGORICAL



INNOMATICS
RESEARCH LABS

RESEARCH QUESTIONS

In [55]:

```
comp_roles = data[data['Designation'].isin(['programming analyst', 'software engineer', 'hardware engineer',
                                            'associate engineer'])]


fresh_grad_salary = comp_roles['Salary'].mean()
print(f"Average Salary for fresh graduates in Computer Science roles: {fresh_grad_salary}")
```

```
Average Salary for fresh graduates in Computer Science roles: 339304.1606886657
```

The claim is not supported by the data. The average salary for fresh graduates in these Computer Science roles is higher than the range suggested in the article. This indicates that fresh graduates may expect to earn more than the stated amount when entering these positions.

In [58]:
```
#gender specialization
crosstab_gender_specialization = pd.crosstab(data['Gender'], data['Specialization'])

crosstab_gender_specialization
```

Out[58]:

Specialization	aeronautical engineering	applied electronics and instrumentation	automobile/automotive engineering	biomedical engineering	biotechnology	ceramic engineering	chemical engineering	civil engineering	computer and communication engineering	compu applicati
Gender										
f	1	2	0	2	9	0	1	6	0	
m	2	7	5	0	6	1	8	23	1	1

2 rows × 46 columns

In [60]:
```
#statistical testing
from scipy.stats import chi2_contingency

# Chi-Square test
chi2, p, dof, expected = chi2_contingency(crosstab_gender_specialization)

print(f'Chi-Square Test Statistic: {chi2:.2f}')
print(f'p-value: {p:.4f}')

if p < 0.05:
    print("There is a significant relationship between Gender and Specialization.")
else:
    print("There is no significant relationship between Gender and Specialization.")
```

```
Chi-Square Test Statistic: 104.47
p-value: 0.0000
There is a significant relationship between Gender and Specialization.
```

The claim supports that there is a significant relationship between Gender and Specialization

ADDITIONAL INSIGHTS

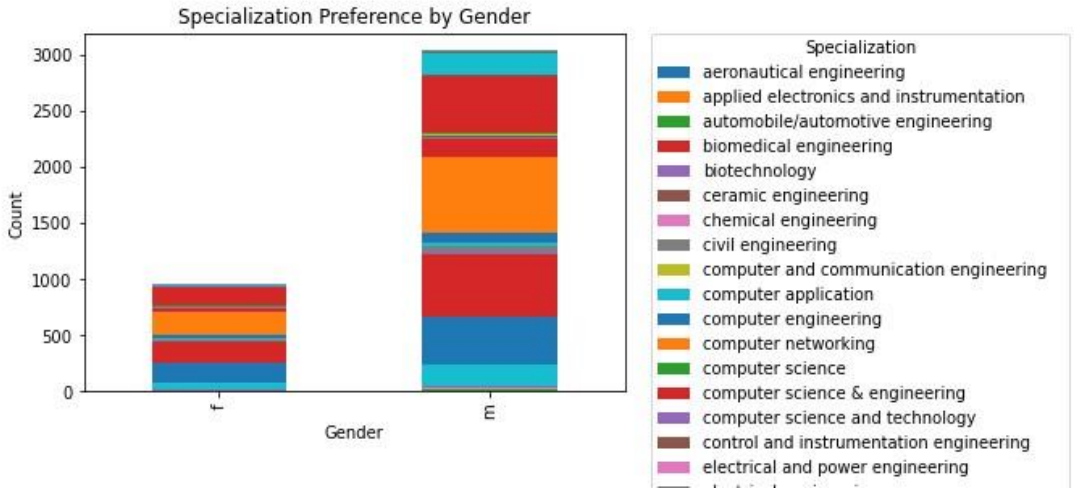
In [61]:
```
# Average salaries vs specialization
# Calculate average salary by specialization
average_salary_specialization = data.groupby('Specialization')['Salary'].mean().sort_values()

# Plotting
plt.figure(figsize=(12, 6))
average_salary_specialization.plot(kind='bar', color='skyblue')
plt.title('Average Salary by Specialization')
plt.xlabel('Specialization')
plt.ylabel('Average Salary (in Lakhs)')
plt.xticks(rotation=45)
plt.show()
```

INNOMATICS RESEARCH LABS

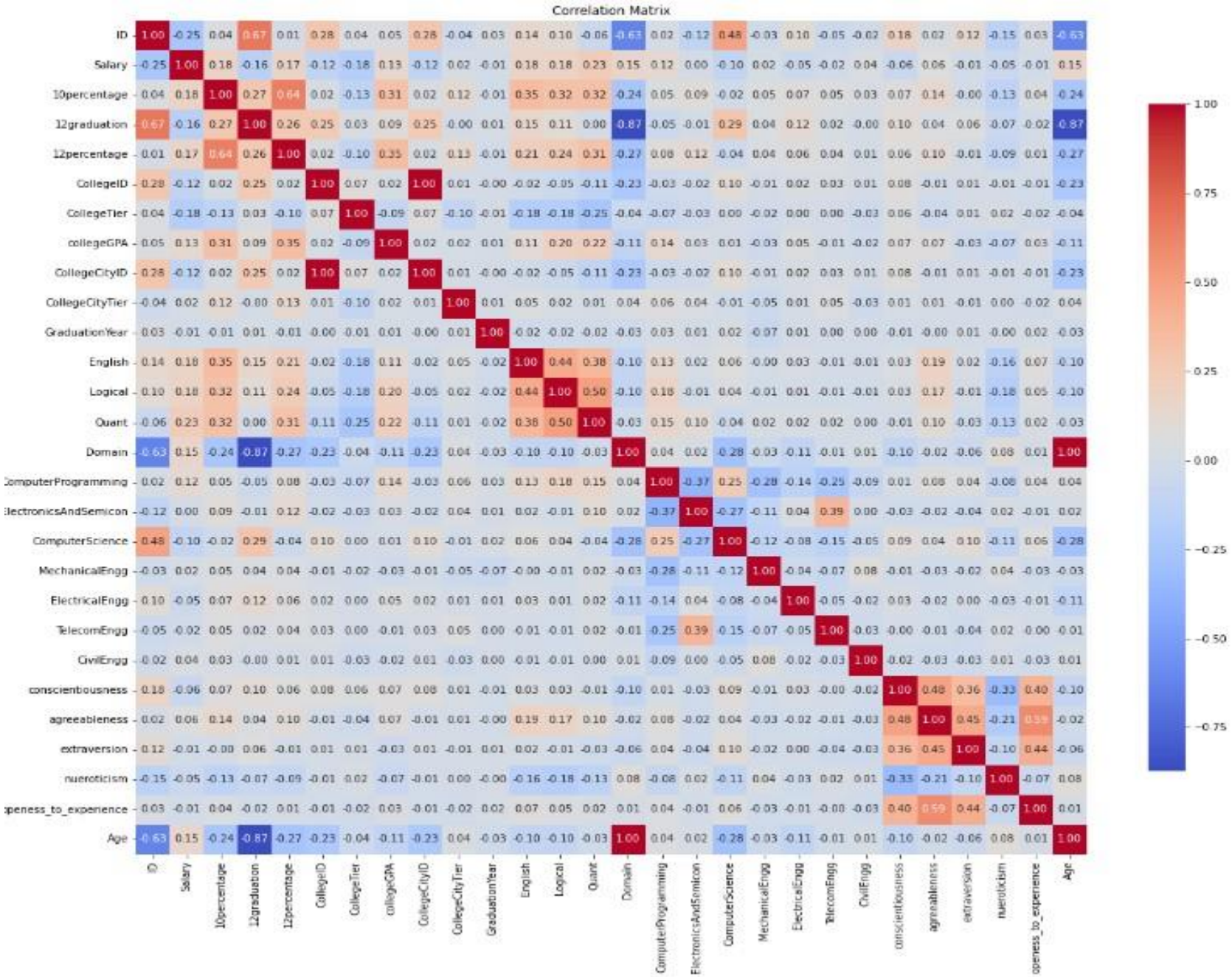
In [62]:
```
# crosstab for Gender and Specialization
crosstab_gender_specialization = pd.crosstab(data['Gender'], data['Specialization'])

# Plot stacked bar chart
plt.figure(figsize=(10, 6))
crosstab_gender_specialization.plot(kind='bar', stacked=True)
plt.title('Specialization Preference by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.legend(title='Specialization', bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.tight_layout()
plt.show()
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_6560\3810993246.py:11: UserWarning: Tight layout not applied. The bottom and top m
argins cannot be made large enough to accommodate all axes decorations.
  plt.tight_layout()
```

```
<Figure size 720x432 with 0 Axes>
```



This image displays the crosstab for Gender and Specialization, illustrating the preferences for specialization based on gender.



sights into how different factors influence

INNOMATICS RESEARCH LABS

one another

CONCLUSION

- Salary Distribution Insights: Analyzed salary ranges, revealing the overall shape and frequency of salaries in the dataset. Identified key trends and variations across different roles.
- Outlier Detection: Box plot analysis highlighted significant outliers in salary data, indicating potential areas for further investigation.
- Gender Distribution: A count plot visualized the distribution of gender, uncovering potential imbalances in the dataset.
- Specialization Preferences: Crosstab analysis demonstrated how specialization choices vary by gender, revealing important trends in educational pathways.
- Key Findings: Notable differences in salary based on specialization and gender.
- Visualizations: Included histograms, box plots, count plots, and crosstabs to effectively convey insights and support findings.

THANK YOU

SARTHAK HIRPURKAR

INNOMATICS RESEARCH LABS