INNOMATİCS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project - AMCAT Data Analysis

By
Pramod Mahendarkar

About me

I currently work at Infosys as System engineer, where I'm gaining valuable experience in Mainframes Technology but right now, I'm enhancing my skills through an exciting internship opportunity. Currently, I'm taking an exciting Internship in data science, a field that's revolutionizing industries through the power of data-driven insights. Passionate about Generative AI and Machine Learning, Data Engineering. I'm always eager to learn and adapt in this ever- evolving industry. I'm passionate about harnessing the potential of data to solve real-world problems, and this hands-on experience is sharpening my skills in programming languages like Python, working with big data, and utilizing cutting-edge tools like TensorFlow and Hadoop.

https://www.linkedin.com/in/pramod-mahendarkar/

https://github.com/PramodMahendarkar

INNOMATICS RESEARCH LABS

Agenda (This should be the PPT flow)

- **Business Problem and Use case domain understanding(If Required)**
- **Objective of the Project**
- **Web Scraping – Details (Websites, Processor you followed)**
- **Summary of the Data**

- **Exploratory Data Analysis:**

a. ***Data Cleaning Steps***
b. ***Data Manipulation Steps***
c. ***Univariate Analysis Steps***
d. ***Bivariate Analysis Steps***

- **Key Business Question**
- **Conclusion (Key finding overall)**
- **Q&A Slide**
- **Your Experience/Challenges working on Web Scraping – Data Analysis Project.**

INNOMATICS RESEARCH LABS

Dataset Description

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses on employment outcomes for engineering graduates. It includes dependent variables such as Salary, Job Titles, and Job Locations, along with standardized scores in cognitive skills, technical skills, and personality skills. With around 40 independent variables and 4000 data points, these variables encompass both continuous and categorical data. The dataset also includes demographic features and unique identifiers for each candidate.

Objective

The goal of this Exploratory Data Analysis (EDA) is to extensively investigate the provided dataset, with a particular emphasis on understanding the link between various variables and the target variable, Salary.
The key aims of this analysis include:

- Providing a detailed explanation of the dataset's features.
- Find any observable patterns or trends in the data.
- Investigating the relationships between the independent factors and the target variable (salary).

Summary of Dataset

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses on employment outcomes for engineering graduates. It includes dependent variables such as Salary, Job Titles, and Job Locations, along with standardized scores in cognitive skills, technical skills, and personality skills. With around 40 independent variables and 4000 data points, these
variables encompass both continuous and categorical data. The dataset also includes demographic features and unique identifiers for each candidate

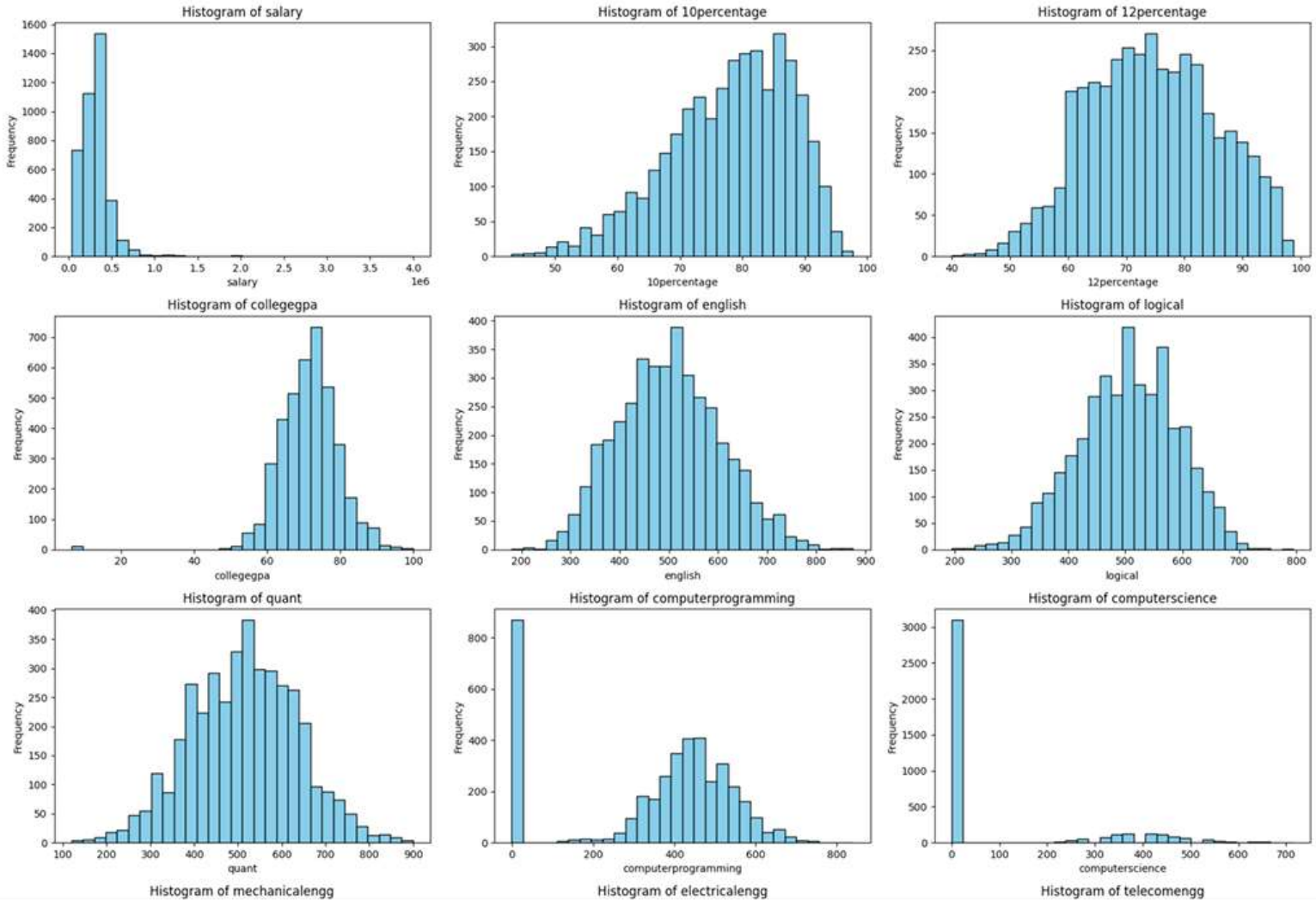
Problem Statement

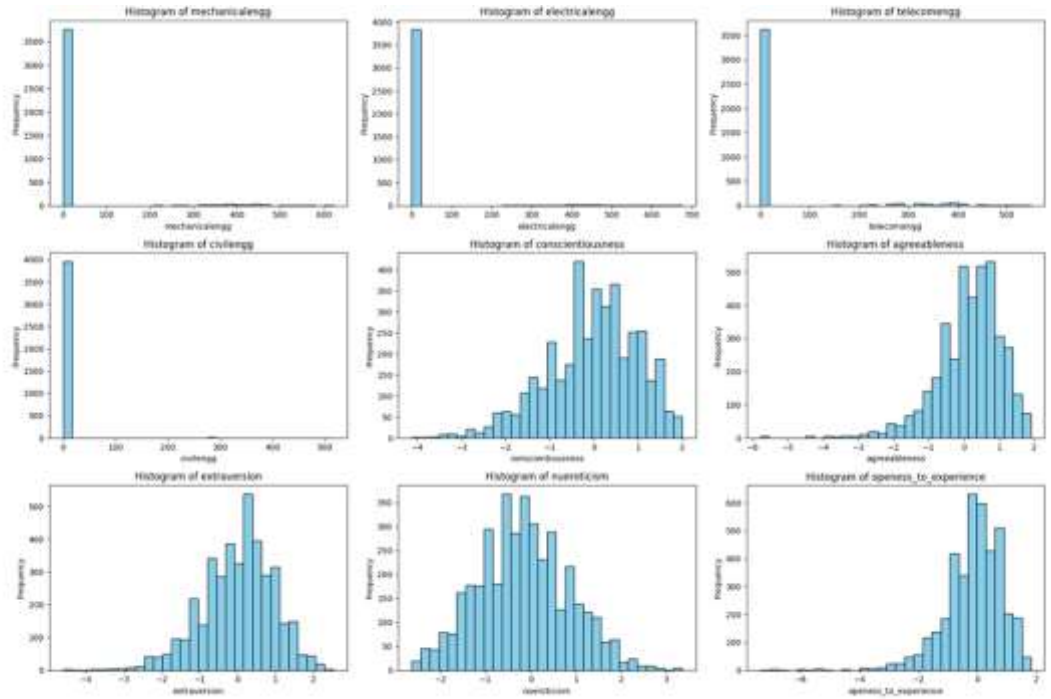
Exploring the relationship between various factors and salary in a dataset comprising academic information of individuals. Investigate how factors such as gender, education, specialization, location, and other attributes correlate with salary levels. Identify patterns, trends, and potential predictors of salary to provide insights for recruitment strategies, career planning, and salary negotiations.

Steps Involved to solve

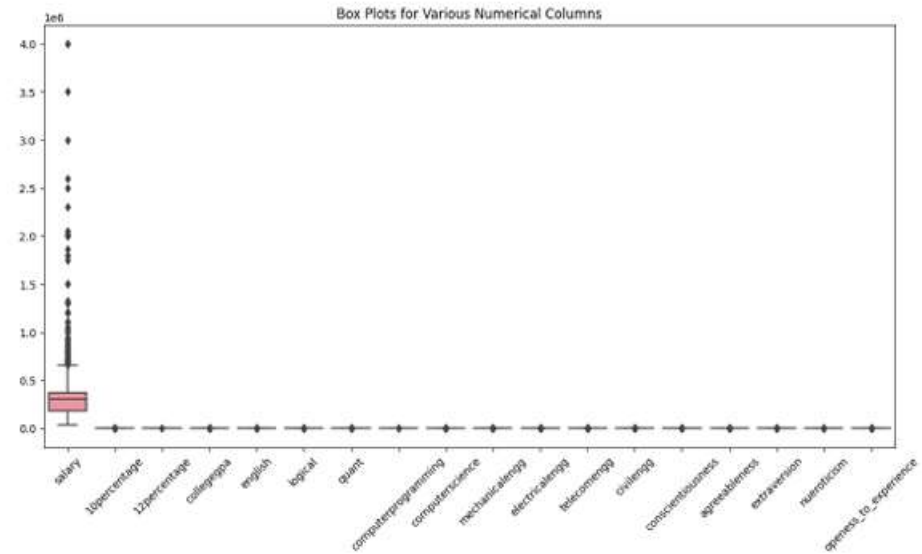
- Importing the Libraries
- Loading the EDA dataset
- Finding the shape, describe, Information of the dataset
- Data cleaning methods like finding out null values , value counts , duplicate values , unique values .Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset .
- Data Manipulation steps likes merging and joining (used for DOB column) , mapping function for 10board , 12board .
- Univariate Analysis - for categorical values and numerical values - Visual representations plotting graphs Box plot, histogram , PDF , Count plot and Non-visual representations - summary statistic (mean,standard deviation , min, max, 25 percentage etc)
- Bivariate Analysis - for numerical vs numerical columns , numerical vs categorical column , categorical vs categorical columns - Visual representations scatter plot, bar plot, hexbin plot, box plot, stacked plot and Non-visual representations correlation matrix .
- Conclusions based upon the Univariate and Bi variate analysis .

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





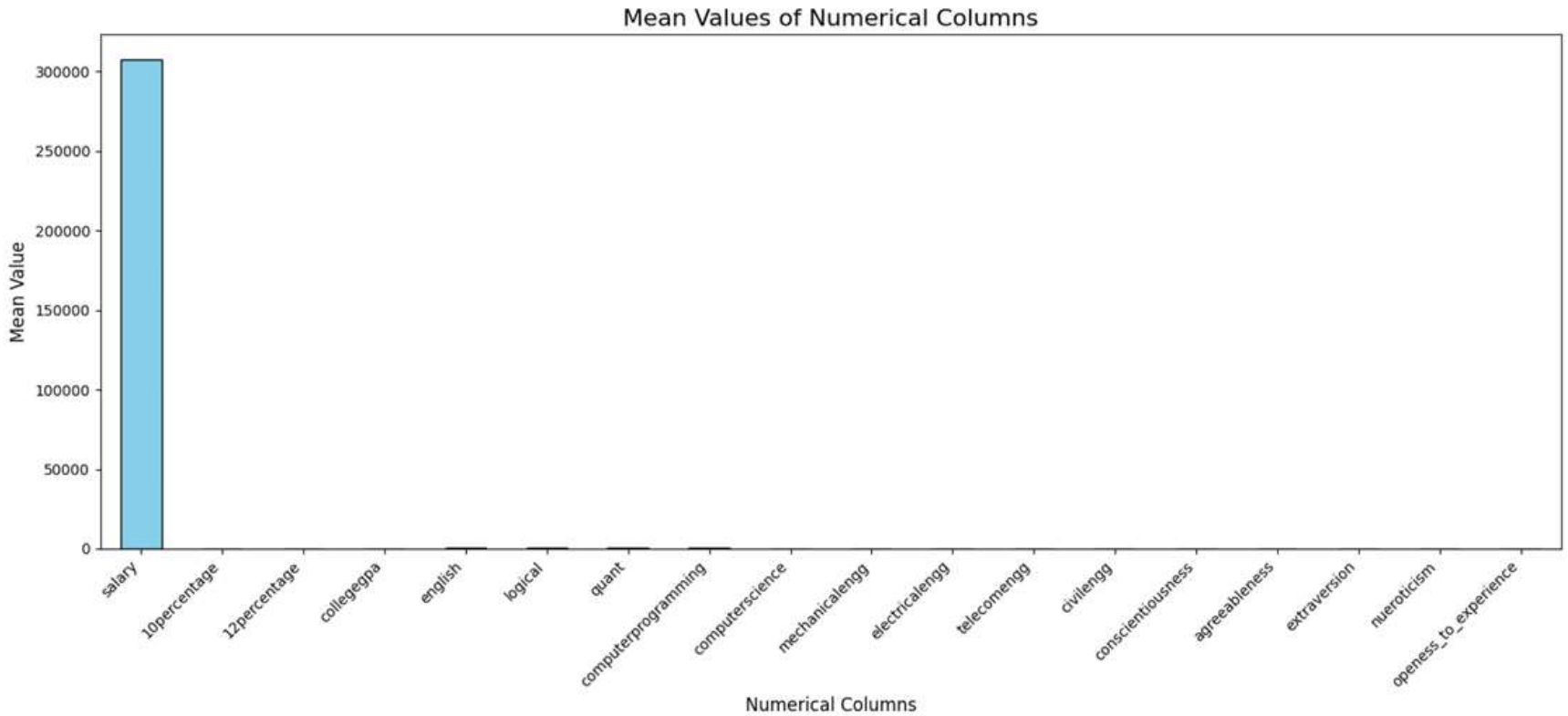
[30]:
```
# Correct list of columns to plot (only numerical columns)
columns_to_plot = ['salary', '10percentage', '12percentage', 'collegegpa',
                   'english', 'logical', 'quant', 'computerprogramming',
                   'computerscience', 'mechanicalengg', 'electricalengg',
                   'telecomengg', 'civilengg', 'conscientiousness',
                   'agreeableness', 'extraversion', 'nueroticism',
                   'openess_to_experience']

# Plot the box plot with valid columns
plt.figure(figsize=(14, 7))
sns.boxplot(data=df[columns_to_plot])
plt.title('Box Plots for Various Numerical Columns')
plt.xticks(rotation=45)
plt.show()
```



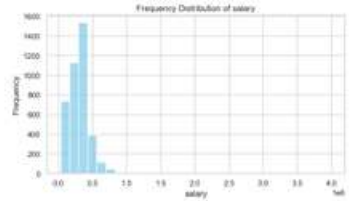
[31]:
```
import matplotlib.pyplot as plt

# Select only numerical columns
columns_to_plot = ['salary', '10percentage', '12percentage', 'collegegpa',␣
↪'english', 'logical',
                   'quant', 'computerprogramming', 'computerscience',␣
↪'mechanicalengg',
```

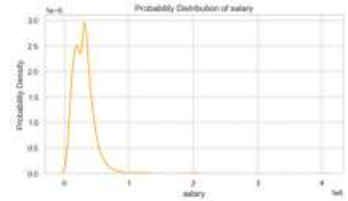


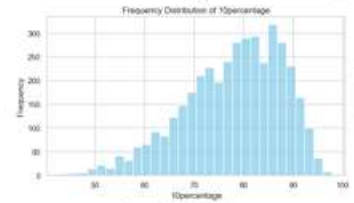
```
                   'electricalengg', 'telecomengg', 'civilengg',␣
↪'conscientiousness',
                   'agreeableness', 'extraversion', 'nueroticism',␣
↪'openess_to_experience']

# Calculate the mean of each numerical column
mean_values = df[columns_to_plot].mean()

# Create the bar plot
plt.figure(figsize=(15, 7))  # Set the figure size
mean_values.plot(kind='bar', color='skyblue', edgecolor='black')

# Customize the plot
plt.title('Mean Values of Numerical Columns', fontsize=16)
plt.xlabel('Numerical Columns', fontsize=12)
plt.ylabel('Mean Value', fontsize=12)
plt.xticks(rotation=45, ha='right')  # Rotate x labels for better visibility

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



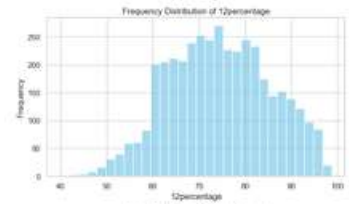
```
# Loop through each numerical column to plot
for i, column in enumerate(columns_to_plot):
    # Frequency Distribution
    sns.histplot(df[column], ax=axes[i, 0], bins=30, kde=False, color='skyblue')
    axes[i, 0].set_title(f'Frequency Distribution of {column}', fontsize=12)
    axes[i, 0].set_xlabel(column)
    axes[i, 0].set_ylabel('Frequency')

    # Probability Distribution (KDE)
    sns.kdeplot(df[column], ax=axes[i, 1], color='orange')
    axes[i, 1].set_title(f'Probability Distribution of {column}', fontsize=12)
    axes[i, 1].set_xlabel(column)
    axes[i, 1].set_ylabel('Probability Density')

# Adjust layout
plt.tight_layout()
plt.show()
```

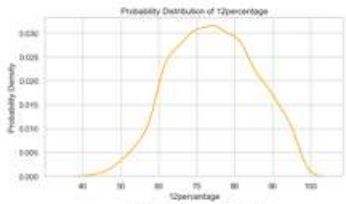

Frequency Distribution of salary
Frequency
salary

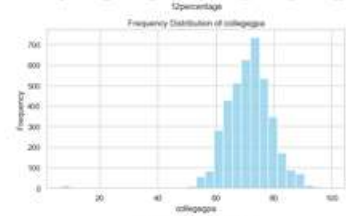

Probability Distribution of salary
1e−6
Probability Density
salary

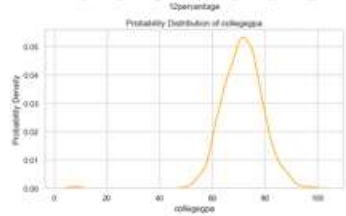

Frequency Distribution of 10percentage
Frequency
10percentage

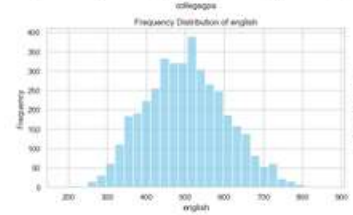

Probability Distribution of 10percentage
Probability Density
10percentage

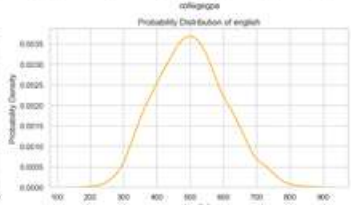

Frequency Distribution of 12percentage
Frequency
12percentage

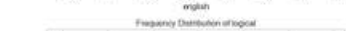

Probability Distribution of 12percentage
Probability Density
12percentage


Frequency Distribution of collegegpa
Frequency
collegegpa


Probability Distribution of collegegpa
Probability Density
collegegpa


Frequency Distribution of english
Frequency
english

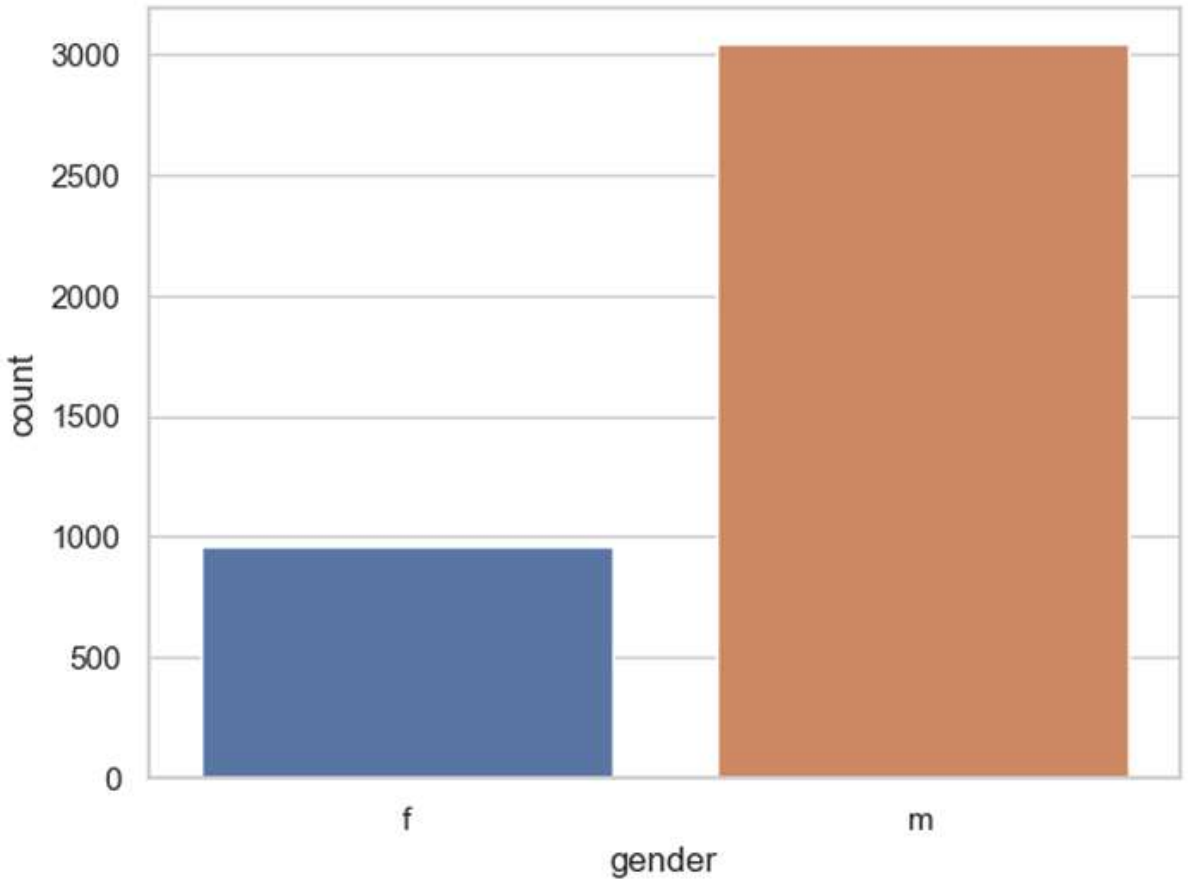

Probability Distribution of english
Probability Density
english


Frequency Distribution of logical


Probability Distribution of logical

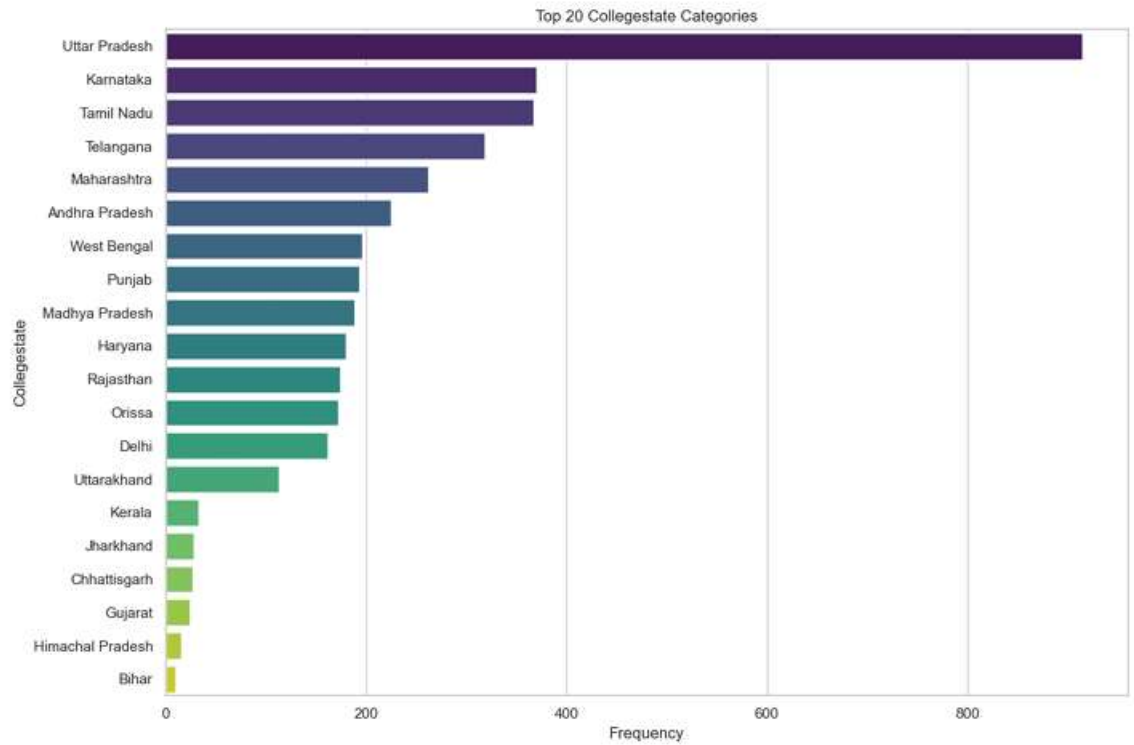
[33]:
```
sns.countplot(x=df['gender'])
```

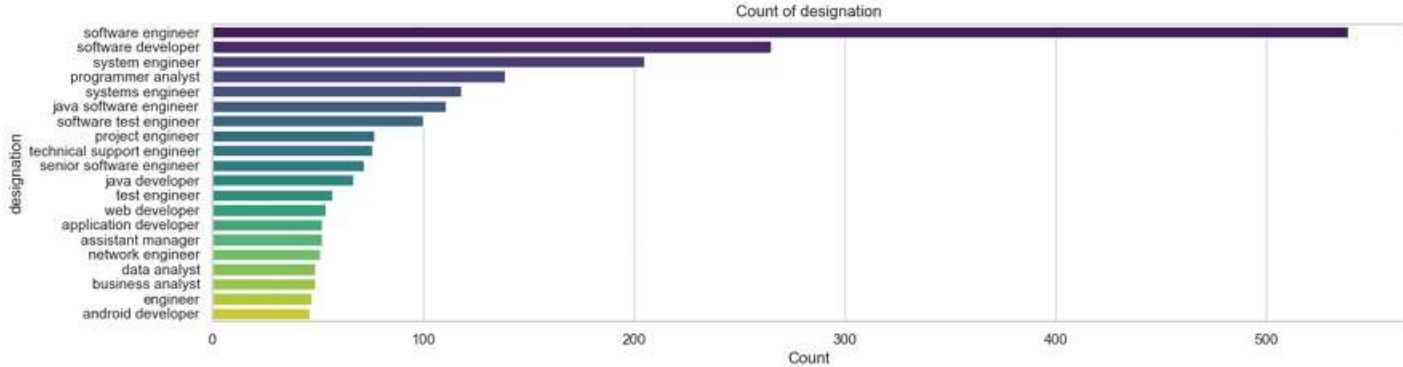
[33]: <Axes: xlabel='gender', ylabel='count'>

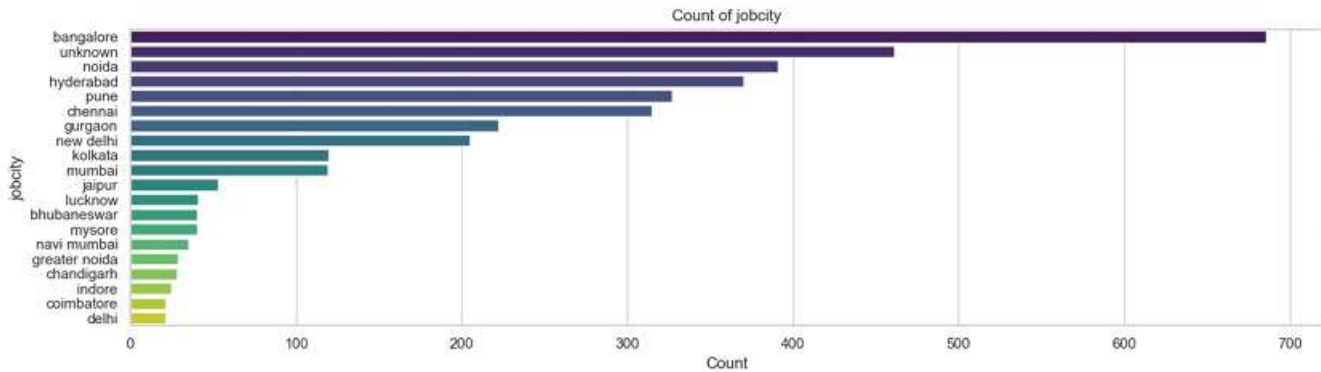


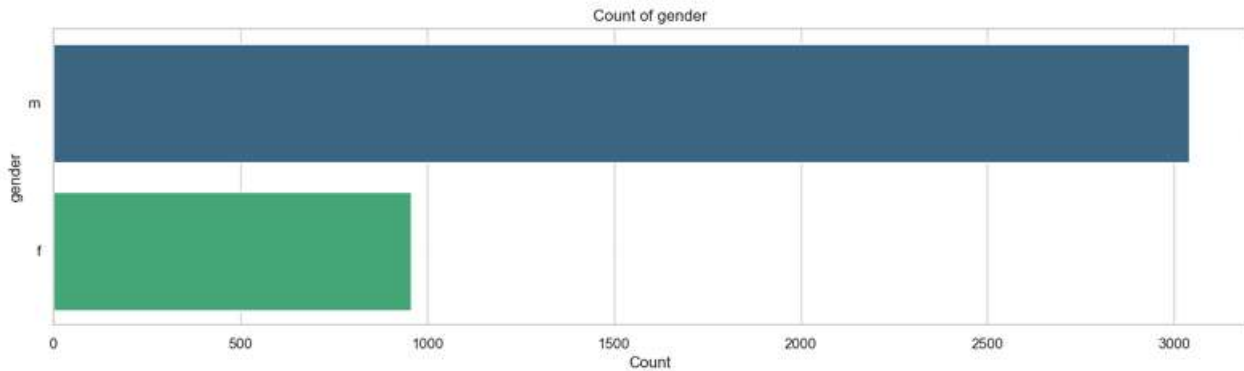
INNOMATICS
RESEARCH LABS

[35]:
```
top_collegestates = df['collegestate'].value_counts().nlargest(20)
plt.figure(figsize=(12, 8))
sns.countplot(y='collegestate', data=df[df['collegestate'].
 ↪isin(top_collegestates.index)],
              palette='viridis', order=top_collegestates.index)
plt.title('Top 20 Collegestate Categories')
plt.xlabel('Frequency')
plt.ylabel('Collegestate')
plt.tight_layout()
plt.show()
```

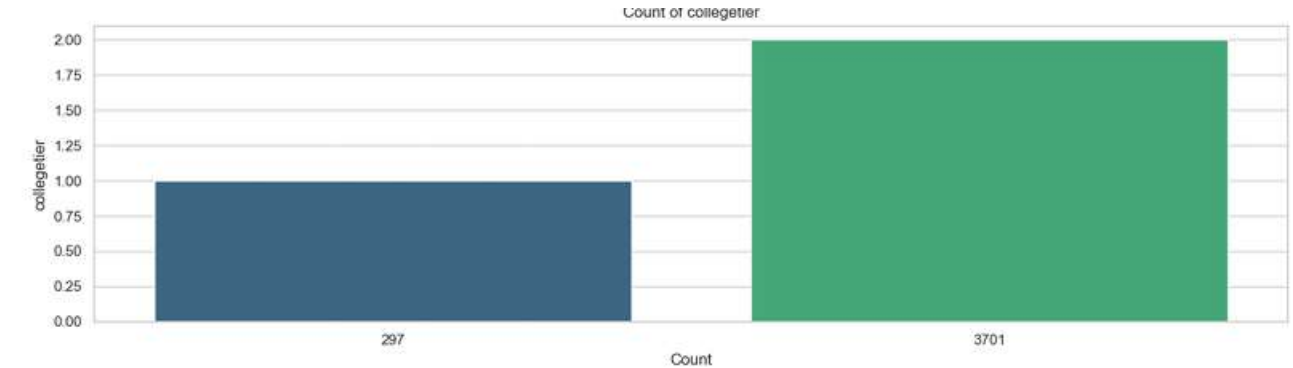


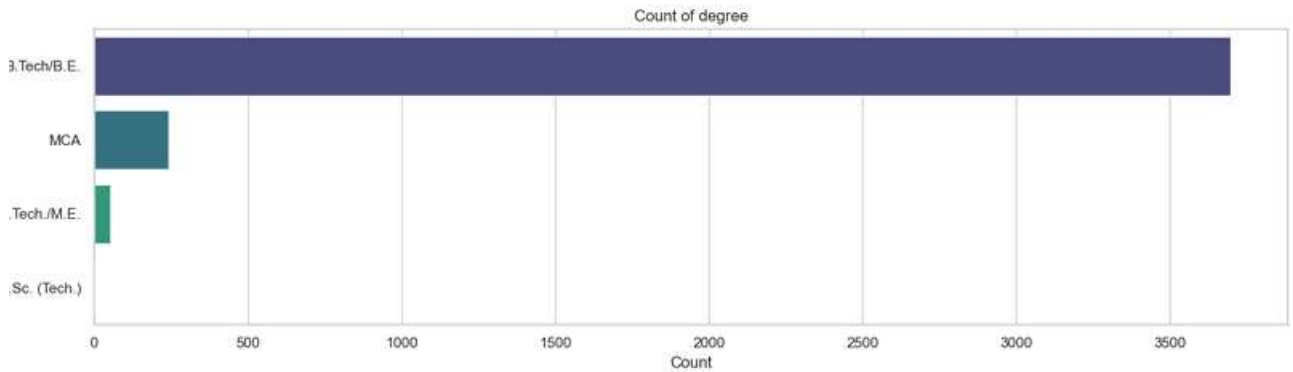






INNOMATICS
RESEARCH LABS

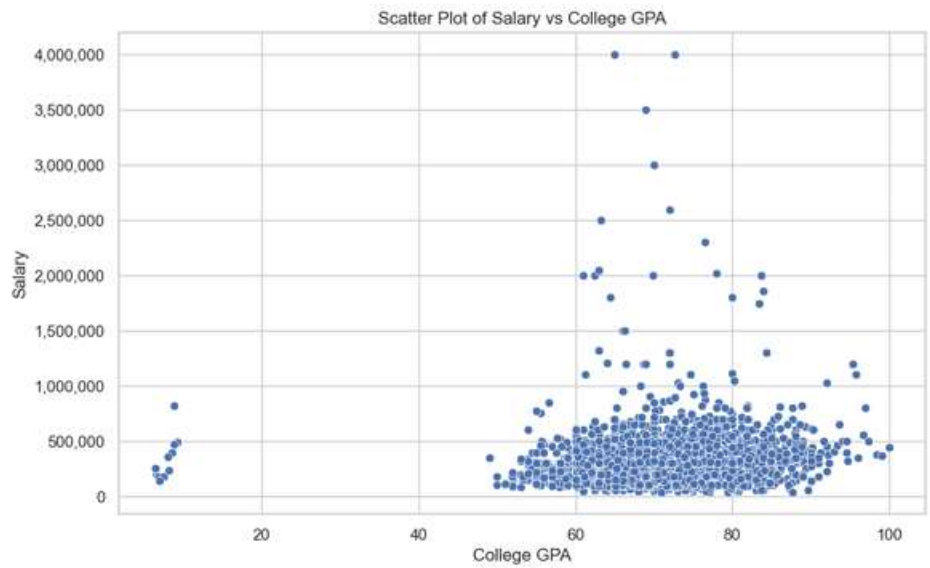


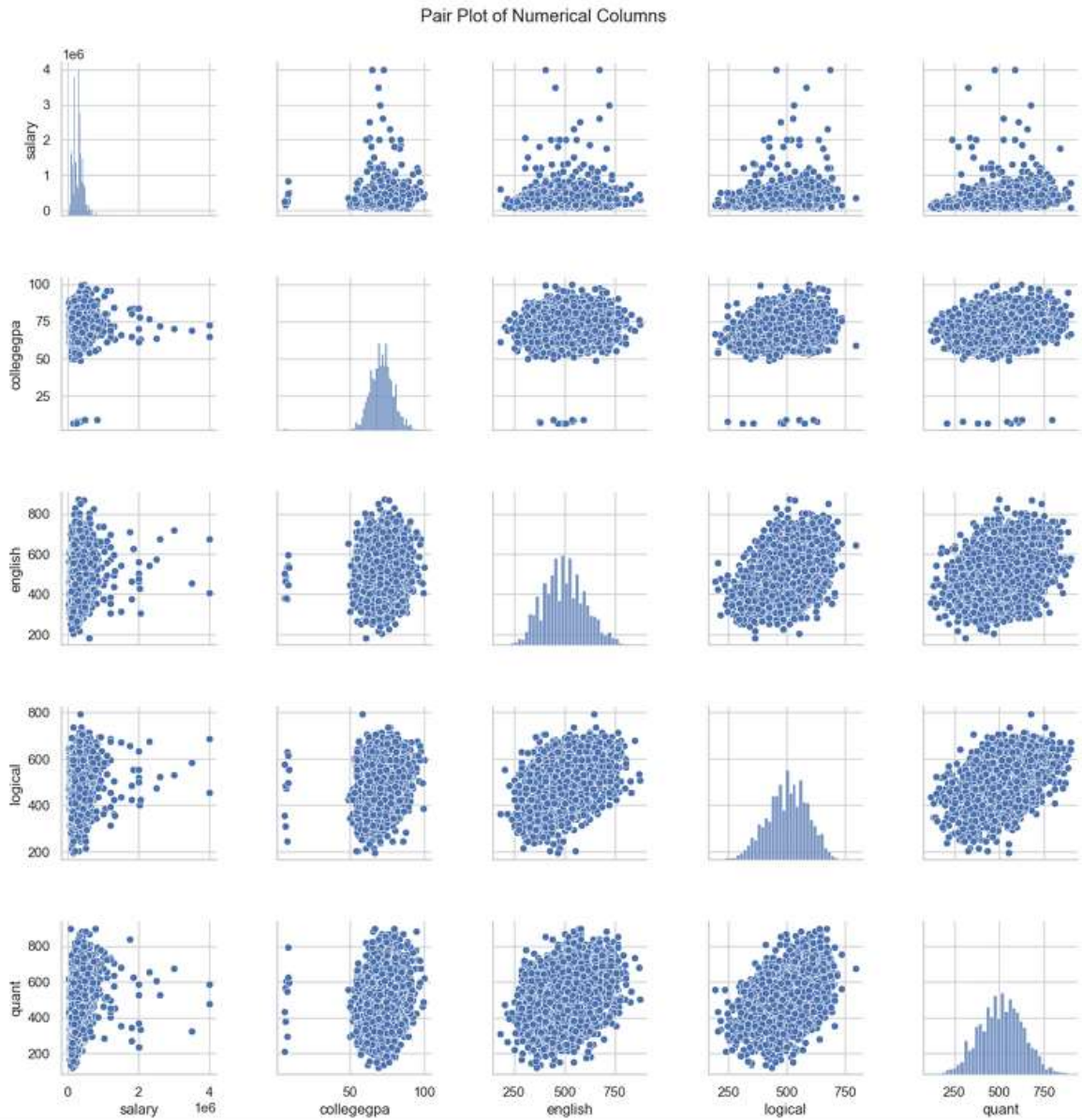


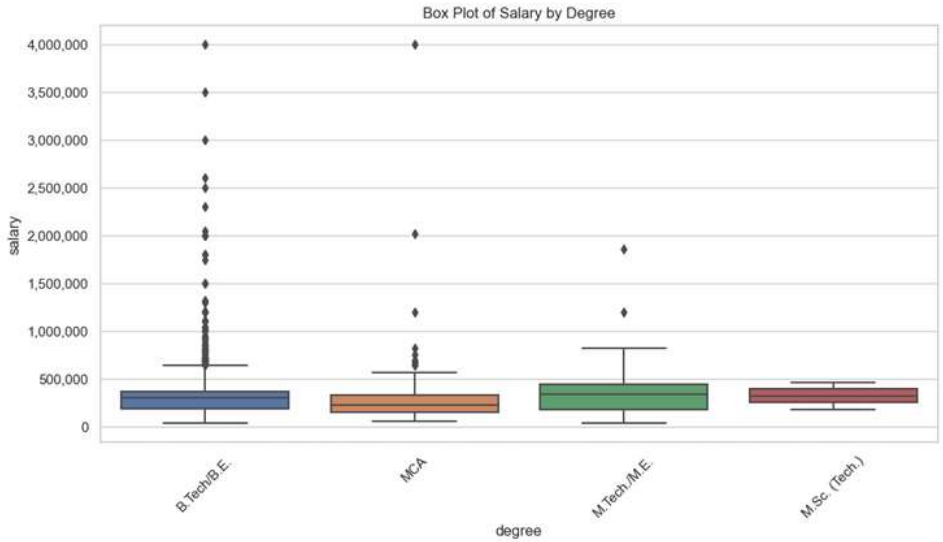
INNOMATICS
RESEARCH LABS

2 Bivariate Analysis

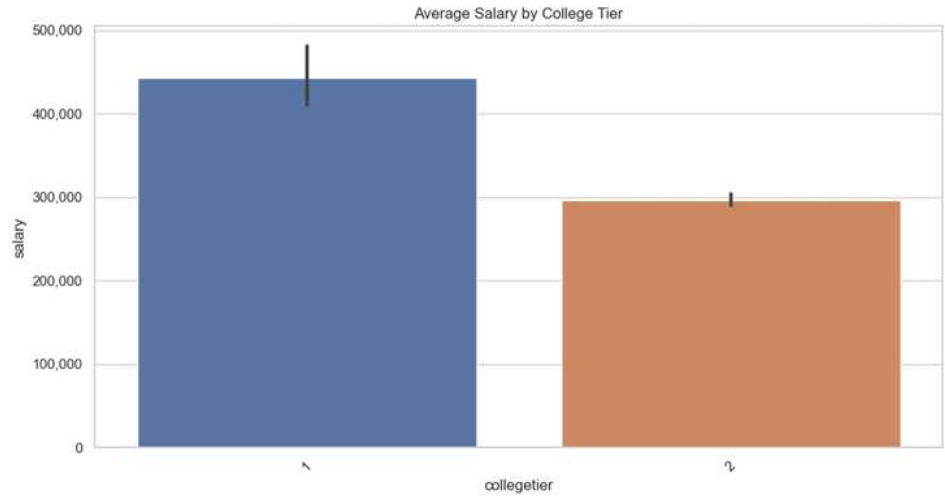
[37]:
```
from matplotlib.ticker import FuncFormatter

# Function to format y-axis labels
def currency(x, _):
    return f'{int(x):,}'  # Format as integer with commas

plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='collegegpa', y='salary')
plt.title('Scatter Plot of Salary vs College GPA')
plt.xlabel('College GPA')
plt.ylabel('Salary')
plt.grid(True)

# Apply the formatter to the y-axis
plt.gca().yaxis.set_major_formatter(FuncFormatter(currency))

plt.show()
```





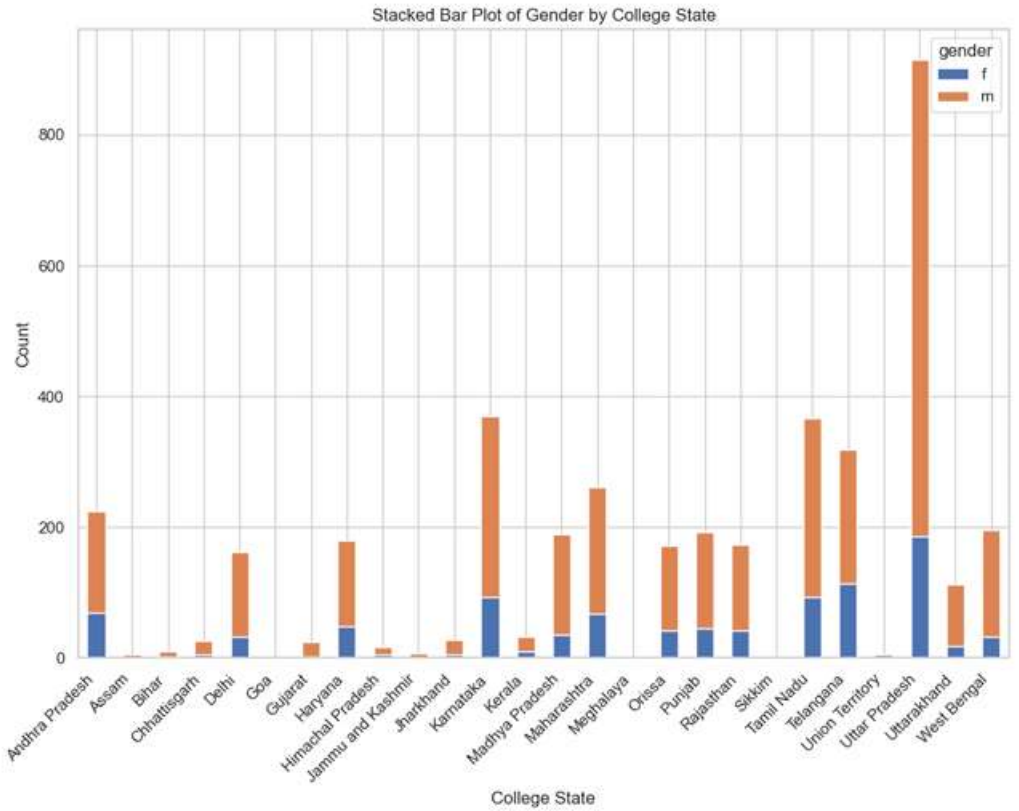


[41]:
```
plt.figure(figsize=(12, 6))
sns.barplot(data=df, x='collegetier', y='salary', estimator=np.mean)
plt.title('Average Salary by College Tier')
plt.xticks(rotation=45)
plt.gca().yaxis.set_major_formatter(FuncFormatter(currency))
plt.show()
```



INNOMATICS RESEARCH LABS

[42]:
```
pivot_table = df.pivot_table(index='collegestate', columns='gender',␣
 ↪values='salary', aggfunc='count').fillna(0)
pivot_table.plot(kind='bar', stacked=True, figsize=(10, 8))
plt.title('Stacked Bar Plot of Gender by College State')
plt.xlabel('College State')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')  # Adjusted alignment to 'right'
plt.tight_layout()  # Adjust layout to prevent clipping
plt.show()
```



Conclusion

The analysis of the provided dataset uncovers several significant insights regarding salary distribution, gender pay gap, and specialization trends within the workforce. Firstly, it reveals a right-skewed distribution of salaries, indicating a presence of high-salary outliers and suggesting that a considerable portion of candidates earn above-average salaries. Moreover, there exists a notable discrepancy in salaries across different designations, with Software Engineers commanding the highest average salary while Data Analysts earn the lowest.

Furthermore, the dataset highlights a concerning gender pay gap, where men consistently earn higher salaries than women on average. This finding underscores the need for further exploration into the underlying causes contributing to this disparity. Additionally, the analysis reveals a positive correlation between age and salary, implying that older candidates tend to earn higher salaries compared to their younger counterparts.

Moreover, there is a noticeable gender imbalance within the top specializations, with males predominating in most areas. This observation raises questions about the factors driving this gender disparity within specific fields. Overall, these insights provide valuable information for identifying areas of concern and potential avenues for promoting fairness and inclusivity in the workplace. Further investigation into the root causes of these trends could lead to informed strategies aimed at addressing salary discrepancies, promoting gender equality, and fostering a more inclusive work environment.

THANK YOU

INNOMATICS RESEARCH LABS