INNOMATİCS
RESEARCH LABS

PROJECT REPORT ON

AMCAT DATA ANALYSIS

Submitted by:

Payal Kumari

INNOMATICS RESEARCH LABS

Project Objective

- The objective of this project is to understand the factors influencing salary prediction based on various attributes such as demographic data, educational background, job details, and psychological traits.
- We aim to clean the data, perform exploratory data analysis (EDA), and generate insights that can help in building a predictive model for salary estimation.

Project Roles:

- **Data Engineer/Scientist:** Responsible for data cleaning, transformation, and exploratory analysis.
- **Visualization Specialist:** Provides different types of visualizations to represent the data insights.

Data Cleaning:

1. **Missing Values:** Identify and handle missing or erroneous data.
2. **Data Formatting:** Ensure all columns, such as Salary, DOJ, DOL, etc., are in the correct format.
3. **Outliers:** Detect and handle any outliers in the Salary column.
4. **Categorical Data:** Convert categorical data into appropriate numerical forms for analysis.

ANALYSIS:

Univariate Analysis:

This involves analysing individual features to understand their distribution and summary statistics.

Bivariate Analysis:

This explores relationships between two variables (e.g., Salary vs JobCity, Gender, etc.) to identify correlations and patterns.

Observations:

1. **Salary Distribution (Univariate Analysis):** The salary distribution appears right-skewed, indicating that most candidates earn in the lower to mid-range of salaries, with fewer individuals earning significantly higher amounts.
2. **Salary vs. JobCity (Bivariate Analysis):** There are significant variations in salary based on job cities. Some cities show a wider salary range, suggesting they might offer more opportunities or have larger firms that pay higher salaries.
3. **Salary vs. Gender (Bivariate Analysis):** There seems to be a noticeable difference in the median salary between males and females, with males appearing to earn more on average in this dataset.
4. **Salary vs. College Tier (Bivariate Analysis):** Candidates from Tier 1 colleges tend to earn more compared to those from Tier 2, suggesting that college reputation plays a role in salary levels.

Visualizations:

1. **Count of Candidates by Gender**: Shows that there are more male candidates compared to female candidates in the dataset.
2. **Distribution of 10th Grade Percentage**: The majority of candidates scored between 70-90% in their 10th-grade exams, with fewer candidates scoring outside this range.
3. **Boxplot of Salary**: Highlights the salary range, with some candidates earning significantly higher salaries.
4. **Salary by Degree**: Reveals that different degrees show varied salary ranges, with certain degrees offering higher earning potential.
5. **Correlation Heatmap**: Shows how various features like Salary, 10percentage, and psychological traits such as conscientiousness correlate with each other.

amcat-data-analysis

October 2, 2024

1 AMCAT DATA ANALYSIS

[2]:
```
import numpy as np
import pandas as pd
import matplotlib as plt
```

2 Load Data

[4]:
```
data_description = pd.read_excel('F:/Innomatics/data_description_doc.xlsx')
results = pd.read_excel('F:/Innomatics/results.xlsx')
test_data = pd.read_excel('F:/Innomatics/test.xlsx')
train_data = pd.read_excel('F:/Innomatics/train.xlsx')
```

[7]:
```
data_description.head() , results.head()
```

[7]:
```
(   Unnamed: 0  \
 0         NaN
 1         NaN
 2         NaN
 3         NaN
 4         NaN

   *This dataset contains self-reported information. May contain
meaningless/misspelled/mistyped entries.  \
 0                                             Input
 1                                        Train/Test
 2                                                ID
 3                                DEPENDENT VARIABLES
 4                                            Salary

                                        Unnamed: 2 Unnamed: 3
 0                                      Description   Comments
 1  Whether the data belongs to the train set or t…        NaN
 2                A unique ID to identify a candidate        NaN
 3                                               NaN        NaN
 4       Annual CTC offered to the candidate (in INR)        NaN  ,
         ID  Salary
```

```
0   664736     NaN
1  1123290     NaN
2  1062444     NaN
3  1072028     NaN
4   267259     NaN)
```

[8]:
```
test_data.head(), train_data.head()
```

[8]:
```
(  Unnamed: 0       ID Salary DOJ DOL Designation JobCity Gender        DOB  \
0       test   664736      ?   ?   ?           ?       ?      m 1992-01-16
1       test  1123290      ?   ?   ?           ?       ?      m 1992-06-05
2       test  1062444      ?   ?   ?           ?       ?      f 1992-11-22
3       test  1072028      ?   ?   ?           ?       ?      f 1990-10-17
4       test   267259      ?   ?   ?           ?       ?      m 1990-03-20

   10percentage  … ComputerScience  MechanicalEngg  ElectricalEngg  \
0          75.0  …              -1              -1              -1
1          83.0  …             253              -1              -1
2          85.2  …              -1              -1              -1
3          81.8  …             469              -1              -1
4          78.0  …              -1              -1              -1

  TelecomEngg  CivilEngg  conscientiousness agreeableness extraversion  \
0          -1         -1             0.2718       -0.2871       0.4711
1          -1         -1             0.7027        0.2124       1.2396
2          -1         -1             0.1282        1.0449      -0.6048
3          -1         -1             0.4155        1.0449      -0.6048
4          -1         -1             0.0464        0.0328      -0.0537

   nueroticism  openess_to_experience
0      -0.7415               -0.4776
1      -0.8682                1.0554
2      -1.6289               -0.8608
3       1.5404                1.0554
4       0.0623                0.6603

[5 rows x 39 columns],
  Unnamed: 0      ID   Salary        DOJ                DOL  \
0      train  203097   420000 2012-06-01            present
1      train  579905   500000 2013-09-01            present
2      train  810601   325000 2014-06-01            present
3      train  267447  1100000 2011-07-01            present
4      train  343523   200000 2014-03-01  2015-03-01 00:00:00

                Designation    JobCity Gender        DOB  10percentage  …  \
0   senior quality engineer  Bangalore      f 1990-02-19          84.3  …
1         assistant manager     Indore      m 1989-10-04          85.4  …
```

```
2          systems engineer    Chennai      f 1992-08-03          85.0  …
3  senior software engineer    Gurgaon      m 1989-12-05          85.6  …
4                       get    Manesar      m 1991-02-27          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])
```

[11]:
```
data_description.head()
```

[11]:
```
   Unnamed: 0  \
0         NaN
1         NaN
2         NaN
3         NaN
4         NaN

  *This dataset contains self-reported information. May contain
meaningless/misspelled/mistyped entries.  \
0                                          Input
1                                     Train/Test
2                                             ID
3                              DEPENDENT VARIABLES
4                                         Salary

                                     Unnamed: 2 Unnamed: 3
0                                   Description   Comments
1  Whether the data belongs to the train set or t…        NaN
```

```
2                A unique ID to identify a candidate        NaN
3                                                NaN        NaN
4       Annual CTC offered to the candidate (in INR)        NaN
```

[12]:
```
results.head()
```

[12]:
```
        ID  Salary
0   664736     NaN
1  1123290     NaN
2  1062444     NaN
3  1072028     NaN
4   267259     NaN
```

[10]:
```
test_data.head()
```

[10]:
```
  Unnamed: 0       ID Salary DOJ DOL Designation JobCity Gender        DOB  \
0       test   664736      ?   ?   ?           ?       ?      m 1992-01-16
1       test  1123290      ?   ?   ?           ?       ?      m 1992-06-05
2       test  1062444      ?   ?   ?           ?       ?      f 1992-11-22
3       test  1072028      ?   ?   ?           ?       ?      f 1990-10-17
4       test   267259      ?   ?   ?           ?       ?      m 1990-03-20

   10percentage  … ComputerScience  MechanicalEngg  ElectricalEngg  \
0          75.0  …               -1              -1              -1
1          83.0  …              253              -1              -1
2          85.2  …               -1              -1              -1
3          81.8  …              469              -1              -1
4          78.0  …               -1              -1              -1

  TelecomEngg  CivilEngg  conscientiousness agreeableness extraversion  \
0          -1         -1             0.2718       -0.2871       0.4711
1          -1         -1             0.7027        0.2124       1.2396
2          -1         -1             0.1282        1.0449      -0.6048
3          -1         -1             0.4155        1.0449      -0.6048
4          -1         -1             0.0464        0.0328      -0.0537

   nueroticism  openess_to_experience
0      -0.7415               -0.4776
1      -0.8682                1.0554
2      -1.6289               -0.8608
3       1.5404                1.0554
4       0.0623                0.6603

[5 rows x 39 columns]
```

[9]:
```
train_data.head()
```

[9]:
```
  Unnamed: 0      ID   Salary        DOJ                 DOL  \
0      train  203097   420000 2012-06-01             present
1      train  579905   500000 2013-09-01             present
2      train  810601   325000 2014-06-01             present
3      train  267447  1100000 2011-07-01             present
4      train  343523   200000 2014-03-01  2015-03-01 00:00:00

                Designation    JobCity Gender        DOB  10percentage  …  \
0   senior quality engineer  Bangalore      f 1990-02-19          84.3  …
1         assistant manager     Indore      m 1989-10-04          85.4  …
2          systems engineer    Chennai      f 1992-08-03          85.0  …
3  senior software engineer    Gurgaon      m 1989-12-05          85.6  …
4                       get    Manesar      m 1991-02-27          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]
```

3 Data Cleaning Summary:

[13]:
```
# 1. Missing Values: Both the train and test datasets contain missing values in various columns. We will handle these by imputing or dropping them based on the context.
# 2. Data Formatting: Certain columns such as DOJ, DOL, and Salary will require formatting adjustments.
# 3. Erroneous Data: The test dataset contains placeholder values like ?, which will need to be addressed.
```

[16]:
```
pip install ace_tools
```

```
Collecting ace_toolsNote: you may need to restart the kernel to use updated
packages.

  Obtaining dependency information for ace_tools from https://files.pythonhosted
.org/packages/27/c4/402d3ae2ecbfe72fbdcb2769f55580f1c54a3ca110c44e1efc034516a499
/ace_tools-0.0-py3-none-any.whl.metadata
  Downloading ace_tools-0.0-py3-none-any.whl.metadata (300 bytes)
Downloading ace_tools-0.0-py3-none-any.whl (1.1 kB)
Installing collected packages: ace_tools
Successfully installed ace_tools-0.0


[notice] A new release of pip is available: 23.2.1 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

[18]:
```
import sys
sys.path.append('/path/to/ace_tools_directory')
```

[23]:
```
# Checking for missing values in the training dataset
missing_values_train = train_data.isnull().sum()

# Checking basic statistics of the train dataset
train_stats = train_data.describe()

# Display missing values in the console
print("Missing Values in Training Data:\n", missing_values_train)

# Display basic statistics of the train dataset
print("\nBasic Statistics of the Training Data:\n", train_stats)

pd.DataFrame(missing_values_train, columns=["Missing Values"])

# Display basic statistics
train_stats
```

```
Missing Values in Training Data:
 Unnamed: 0               0
ID                        0
Salary                    0
DOJ                       0
DOL                       0
Designation               0
JobCity                   0
Gender                    0
DOB                       0
10percentage              0
```

```
10board                   0
12graduation              0
12percentage              0
12board                   0
CollegeID                 0
CollegeTier               0
Degree                    0
Specialization            0
collegeGPA                0
CollegeCityID             0
CollegeCityTier           0
CollegeState              0
GraduationYear            0
English                   0
Logical                   0
Quant                     0
Domain                    0
ComputerProgramming       0
ElectronicsAndSemicon     0
ComputerScience           0
MechanicalEngg            0
ElectricalEngg            0
TelecomEngg               0
CivilEngg                 0
conscientiousness         0
agreeableness             0
extraversion              0
nueroticism               0
openess_to_experience     0
dtype: int64

Basic Statistics of the Training Data:
                  ID        Salary                             DOJ  \
count  3.998000e+03  3.998000e+03                             3998
mean   6.637945e+05  3.076998e+05  2013-07-02 11:04:10.325162496
min    1.124400e+04  3.500000e+04            1991-06-01 00:00:00
25%    3.342842e+05  1.800000e+05            2012-10-01 00:00:00
50%    6.396000e+05  3.000000e+05            2013-11-01 00:00:00
75%    9.904800e+05  3.700000e+05            2014-07-01 00:00:00
max    1.298275e+06  4.000000e+06            2015-12-01 00:00:00
std    3.632182e+05  2.127375e+05                            NaN

                                DOB  10percentage  12graduation  \
count                          3998   3998.000000   3998.000000
mean   1990-12-06 06:01:15.637819008     77.925443   2008.087544
min              1977-10-30 00:00:00     43.000000   1995.000000
25%              1989-11-16 06:00:00     71.680000   2007.000000
50%              1991-03-07 12:00:00     79.150000   2008.000000
```

```
75%              1992-03-13 18:00:00     85.670000   2009.000000
max              1997-05-27 00:00:00     97.760000   2013.000000
std                              NaN      9.850162      1.653599

       12percentage     CollegeID  CollegeTier   collegeGPA  …  \
count   3998.000000   3998.000000  3998.000000  3998.000000  …
mean      74.466366   5156.851426     1.925713    71.486171  …
min       40.000000      2.000000     1.000000     6.450000  …
25%       66.000000    494.000000     2.000000    66.407500  …
50%       74.400000   3879.000000     2.000000    71.720000  …
75%       82.600000   8818.000000     2.000000    76.327500  …
max       98.700000  18409.000000     2.000000    99.930000  …
std       10.999933   4802.261482     0.262270     8.167338  …

       ComputerScience  MechanicalEngg  ElectricalEngg  TelecomEngg  \
count      3998.000000     3998.000000     3998.000000  3998.000000
mean         90.742371       22.974737       16.478739    31.851176
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
std         175.273083       98.123311       87.585634   104.852845

         CivilEngg  conscientiousness  agreeableness  extraversion  \
count  3998.000000        3998.000000    3998.000000   3998.000000
mean      2.683842          -0.037831       0.146496      0.002763
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
std      36.658505           1.028666       0.941782      0.951471

       nueroticism  openess_to_experience
count  3998.000000            3998.000000
mean     -0.169033              -0.138110
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
std       1.007580               1.008075

[8 rows x 29 columns]
```

[23]:

	ID	Salary	DOJ	\
count	3.998000e+03	3.998000e+03	3998	
mean	6.637945e+05	3.076998e+05	2013-07-02 11:04:10.325162496	
min	1.124400e+04	3.500000e+04	1991-06-01 00:00:00	
25%	3.342842e+05	1.800000e+05	2012-10-01 00:00:00	
50%	6.396000e+05	3.000000e+05	2013-11-01 00:00:00	
75%	9.904800e+05	3.700000e+05	2014-07-01 00:00:00	
max	1.298275e+06	4.000000e+06	2015-12-01 00:00:00	
std	3.632182e+05	2.127375e+05	NaN	

	DOB	10percentage	12graduation	\
count	3998	3998.000000	3998.000000	
mean	1990-12-06 06:01:15.637819008	77.925443	2008.087544	
min	1977-10-30 00:00:00	43.000000	1995.000000	
25%	1989-11-16 06:00:00	71.680000	2007.000000	
50%	1991-03-07 12:00:00	79.150000	2008.000000	
75%	1992-03-13 18:00:00	85.670000	2009.000000	
max	1997-05-27 00:00:00	97.760000	2013.000000	
std	NaN	9.850162	1.653599	

	12percentage	CollegeID	CollegeTier	collegeGPA	…	\
count	3998.000000	3998.000000	3998.000000	3998.000000	…	
mean	74.466366	5156.851426	1.925713	71.486171	…	
min	40.000000	2.000000	1.000000	6.450000	…	
25%	66.000000	494.000000	2.000000	66.407500	…	
50%	74.400000	3879.000000	2.000000	71.720000	…	
75%	82.600000	8818.000000	2.000000	76.327500	…	
max	98.700000	18409.000000	2.000000	99.930000	…	
std	10.999933	4802.261482	0.262270	8.167338	…	

	ComputerScience	MechanicalEngg	ElectricalEngg	TelecomEngg	\
count	3998.000000	3998.000000	3998.000000	3998.000000	
mean	90.742371	22.974737	16.478739	31.851176	
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	
std	175.273083	98.123311	87.585634	104.852845	

	CivilEngg	conscientiousness	agreeableness	extraversion	\
count	3998.000000	3998.000000	3998.000000	3998.000000	
mean	2.683842	-0.037831	0.146496	0.002763	
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
std      36.658505           1.028666       0.941782       0.951471

       nueroticism  openess_to_experience
count  3998.000000            3998.000000
mean     -0.169033              -0.138110
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
std       1.007580               1.008075

[8 rows x 29 columns]
```

[24]:
```python
# Checking for missing values in the test dataset
missing_values_test = test_data.isnull().sum()

# Checking basic statistics of the test dataset
test_stats = test_data.describe()

# Display missing values in the console or Jupyter Notebook
print("Missing Values in Test Data:\n", missing_values_test)

# Display basic statistics of the test dataset
print("\nBasic Statistics of the Test Data:\n", test_stats)

# Display missing values as a DataFrame
pd.DataFrame(missing_values_test, columns=["Missing Values"])

# Display basic statistics for the test dataset
test_stats
```

```
Missing Values in Test Data:
 Unnamed: 0               0
ID                        0
Salary                    0
DOJ                       0
DOL                       0
Designation               0
JobCity                   0
Gender                    0
DOB                       0
10percentage              0
10board                   0
12graduation              0
12percentage              0
```

```
12board                   0
CollegeID                 0
CollegeTier               0
Degree                    0
Specialization            0
collegeGPA                0
CollegeCityID             0
CollegeCityTier           0
CollegeState              0
GraduationYear            0
English                   0
Logical                   0
Quant                     0
Domain                    0
ComputerProgramming       0
ElectronicsAndSemicon     0
ComputerScience           0
MechanicalEngg            0
ElectricalEngg            0
TelecomEngg               0
CivilEngg                 0
conscientiousness         0
agreeableness             0
extraversion              0
nueroticism               0
openess_to_experience     0
dtype: int64

Basic Statistics of the Test Data:
                  ID                          DOB  10percentage  12graduation  \
count  1.500000e+03                         1500   1500.000000   1500.000000
mean   6.652863e+05  1990-12-20 01:04:19.200000     78.384553   2008.122000
min    7.474000e+03         1984-01-20 00:00:00     43.080000   2000.000000
25%    3.350625e+05         1989-12-04 18:00:00     72.000000   2007.000000
50%    6.419305e+05         1991-02-11 00:00:00     80.000000   2008.000000
75%    9.858025e+05         1992-03-18 18:00:00     85.600000   2009.000000
max    1.298259e+06         1995-12-28 00:00:00     97.600000   2013.000000
std    3.605324e+05                         NaN      9.565983      1.588542

       12percentage     CollegeID  CollegeTier   collegeGPA  CollegeCityID  \
count   1500.000000   1500.000000  1500.000000  1500.000000    1500.000000
mean      74.947040   5202.454667     1.926667    71.615147    5202.454667
min       40.830000      2.000000     1.000000     7.000000       2.000000
25%       67.000000    830.000000     2.000000    67.147500     830.000000
50%       75.000000   3879.000000     2.000000    72.000000    3879.000000
75%       83.200000   8784.750000     2.000000    76.662500    8784.750000
max       98.000000  17293.000000     2.000000    95.000000   17293.000000
std       10.632432   4750.131676     0.260770     8.747405    4750.131676
```

```
       CollegeCityTier  …  ComputerScience  MechanicalEngg  ElectricalEngg  \
count      1500.000000  …      1500.000000     1500.000000     1500.000000
mean          0.278000  …        84.992000       22.992667       20.673333
min           0.000000  …        -1.000000       -1.000000       -1.000000
25%           0.000000  …        -1.000000       -1.000000       -1.000000
50%           0.000000  …        -1.000000       -1.000000       -1.000000
75%           1.000000  …        -1.000000       -1.000000       -1.000000
max           1.000000  …       746.000000      653.000000      676.000000
std           0.448163  …       171.721189       99.572364       98.467198

       TelecomEngg    CivilEngg  conscientiousness  agreeableness  \
count  1500.000000  1500.000000        1500.000000    1500.000000
mean     34.525333     3.997333          -0.038361       0.183612
min      -1.000000    -1.000000          -3.508500      -4.283100
25%      -1.000000    -1.000000          -0.733500      -0.287100
50%      -1.000000    -1.000000           0.046400       0.344800
75%      -1.000000    -1.000000           0.702700       0.812800
max     553.000000   548.000000           1.995300       1.904800
std     106.704076    43.220335           1.021743       0.858094

       extraversion  nueroticism  openess_to_experience
count   1500.000000  1500.000000            1500.000000
mean       0.048418    -0.091588              -0.102384
min       -3.371300    -2.643000              -5.842800
25%       -0.598000    -0.868200              -0.669200
50%        0.091400    -0.107600              -0.046350
75%        0.672000     0.532330               0.502400
max        2.315400     3.061700               1.630200
std        0.919562     1.010601               0.908886

[8 rows x 27 columns]
```

[24]:

```
                ID                          DOB  10percentage  12graduation  \
count  1.500000e+03                         1500   1500.000000   1500.000000
mean   6.652863e+05  1990-12-20 01:04:19.200000     78.384553   2008.122000
min    7.474000e+03         1984-01-20 00:00:00     43.080000   2000.000000
25%    3.350625e+05         1989-12-04 18:00:00     72.000000   2007.000000
50%    6.419305e+05         1991-02-11 00:00:00     80.000000   2008.000000
75%    9.858025e+05         1992-03-18 18:00:00     85.600000   2009.000000
max    1.298259e+06         1995-12-28 00:00:00     97.600000   2013.000000
std    3.605324e+05                         NaN      9.565983      1.588542

       12percentage     CollegeID  CollegeTier   collegeGPA  CollegeCityID  \
count   1500.000000   1500.000000  1500.000000  1500.000000    1500.000000
mean      74.947040   5202.454667     1.926667    71.615147    5202.454667
min       40.830000      2.000000     1.000000     7.000000       2.000000
```

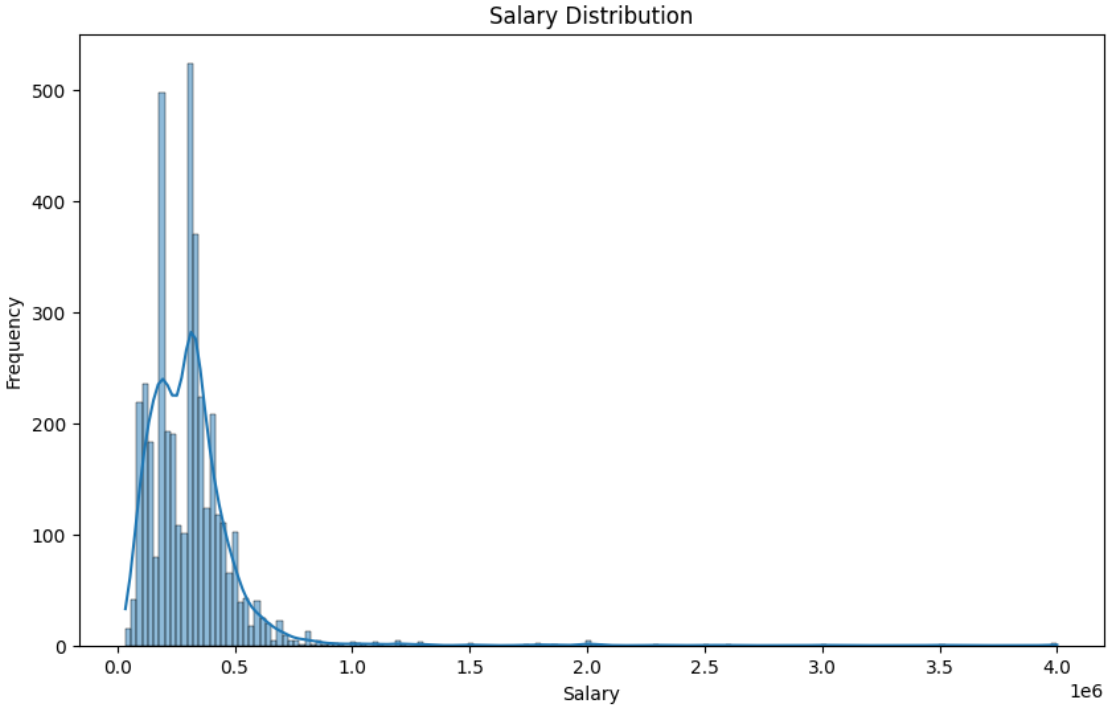
```
25%       67.000000    830.000000      2.000000    67.147500     830.000000
50%       75.000000   3879.000000      2.000000    72.000000    3879.000000
75%       83.200000   8784.750000      2.000000    76.662500    8784.750000
max       98.000000  17293.000000      2.000000    95.000000   17293.000000
std       10.632432   4750.131676      0.260770     8.747405    4750.131676

       CollegeCityTier  …  ComputerScience  MechanicalEngg  ElectricalEngg  \
count      1500.000000  …      1500.000000     1500.000000     1500.000000
mean          0.278000  …        84.992000       22.992667       20.673333
min           0.000000  …        -1.000000       -1.000000       -1.000000
25%           0.000000  …        -1.000000       -1.000000       -1.000000
50%           0.000000  …        -1.000000       -1.000000       -1.000000
75%           1.000000  …        -1.000000       -1.000000       -1.000000
max           1.000000  …       746.000000      653.000000      676.000000
std           0.448163  …       171.721189       99.572364       98.467198

       TelecomEngg    CivilEngg  conscientiousness  agreeableness  \
count  1500.000000  1500.000000        1500.000000    1500.000000
mean     34.525333     3.997333          -0.038361       0.183612
min      -1.000000    -1.000000          -3.508500      -4.283100
25%      -1.000000    -1.000000          -0.733500      -0.287100
50%      -1.000000    -1.000000           0.046400       0.344800
75%      -1.000000    -1.000000           0.702700       0.812800
max     553.000000   548.000000           1.995300       1.904800
std     106.704076    43.220335           1.021743       0.858094

       extraversion  nueroticism  openess_to_experience
count   1500.000000  1500.000000            1500.000000
mean       0.048418    -0.091588              -0.102384
min       -3.371300    -2.643000              -5.842800
25%       -0.598000    -0.868200              -0.669200
50%        0.091400    -0.107600              -0.046350
75%        0.672000     0.532330               0.502400
max        2.315400     3.061700               1.630200
std        0.919562     1.010601               0.908886

[8 rows x 27 columns]
```

[27]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Data Cleaning: Replace '?' with NaN in test data and convert salary to␣
 ↪numeric in train data
test_data.replace('?', pd.NA, inplace=True)
train_data['Salary'] = pd.to_numeric(train_data['Salary'], errors='coerce')

# Dropping rows with NaN in Salary for the sake of univariate analysis
```

```
clean_train_data = train_data.dropna(subset=['Salary'])
```

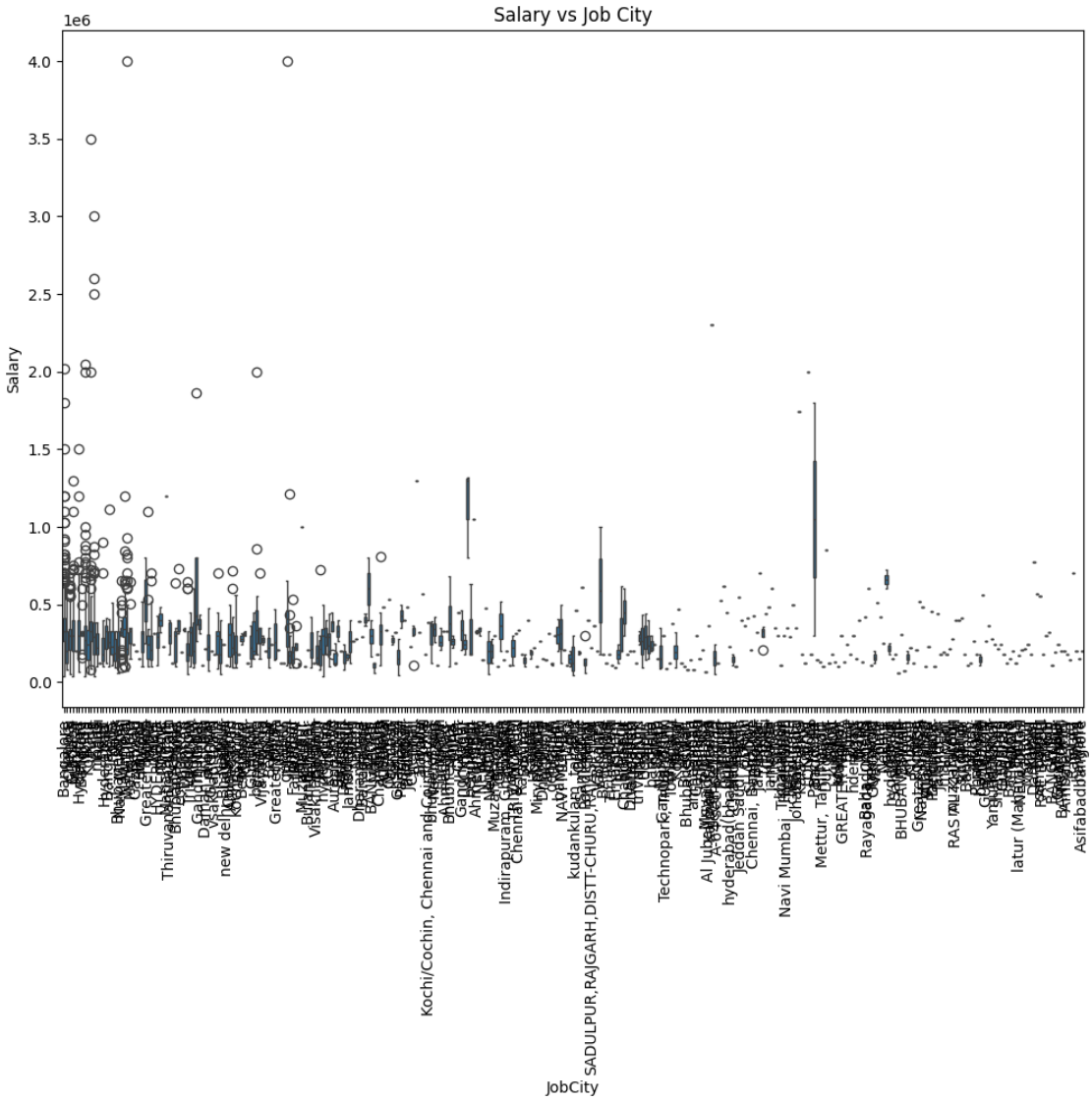
4 Univariate Analysis

[28]:
```
# Univariate Analysis: Distribution of Salary
plt.figure(figsize=(10, 6))
sns.histplot(clean_train_data['Salary'], kde=True)
plt.title('Salary Distribution')
plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.show()
```

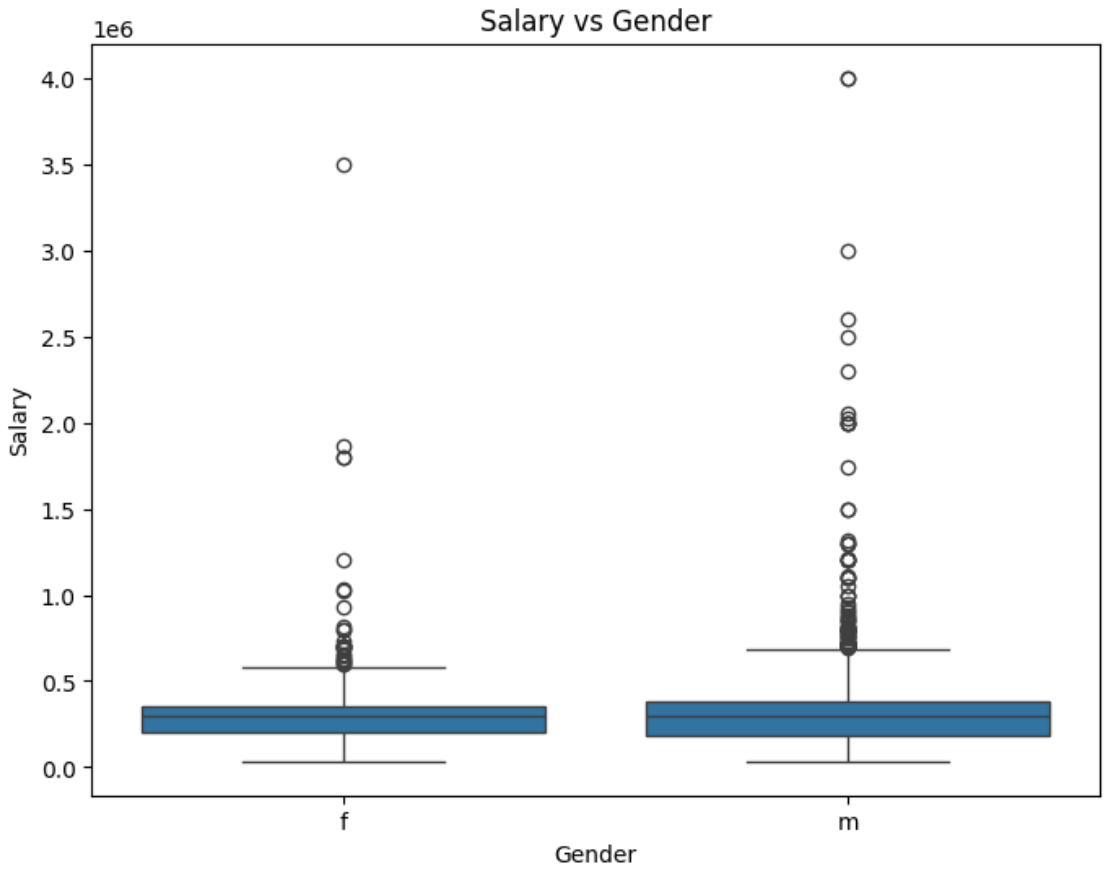


5 Bivariate Analysis

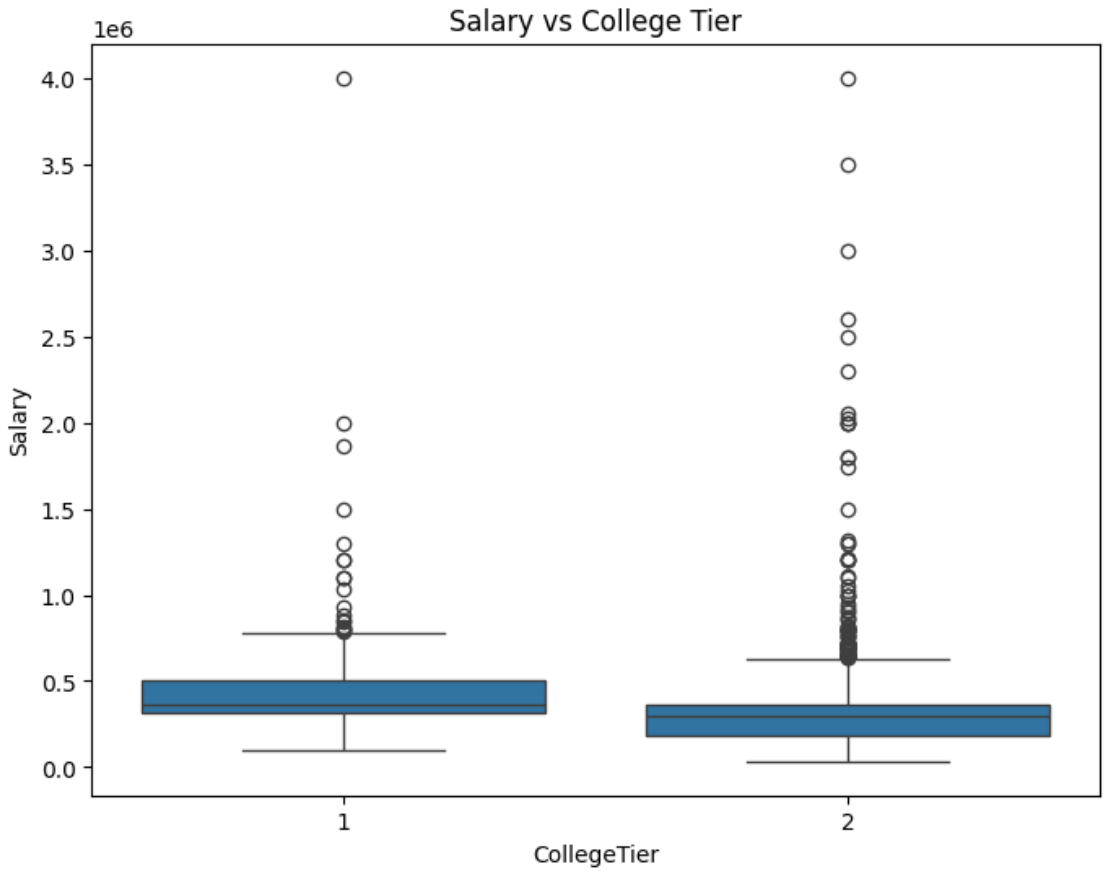
[29]:
```
# Bivariate Analysis: Salary vs JobCity
plt.figure(figsize=(12, 8))
sns.boxplot(x='JobCity', y='Salary', data=clean_train_data)
plt.title('Salary vs Job City')
plt.xticks(rotation=90)
plt.show()
```



[30]:
```
# Bivariate Analysis: Salary vs Gender
plt.figure(figsize=(8, 6))
sns.boxplot(x='Gender', y='Salary', data=clean_train_data)
plt.title('Salary vs Gender')
plt.show()
```

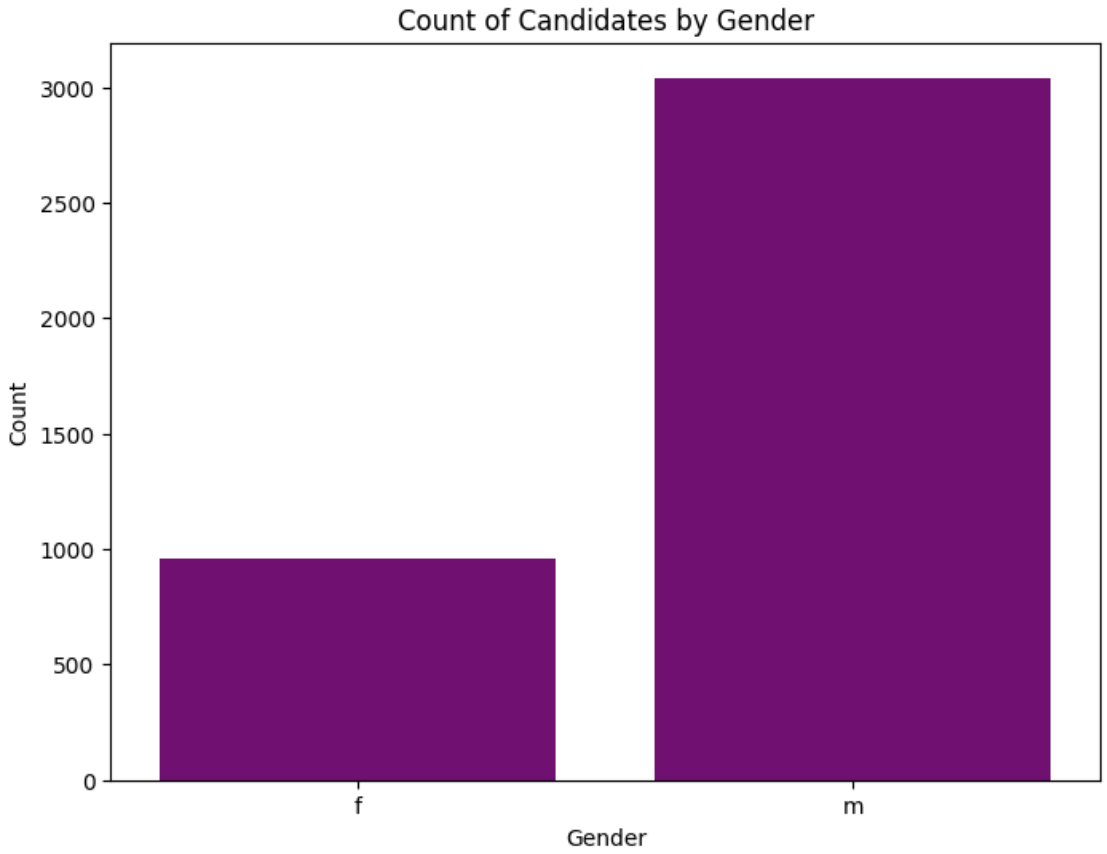


[31]:
```
# Bivariate Analysis: Salary vs CollegeTier
plt.figure(figsize=(8, 6))
sns.boxplot(x='CollegeTier', y='Salary', data=clean_train_data)
plt.title('Salary vs College Tier')
plt.show()
```

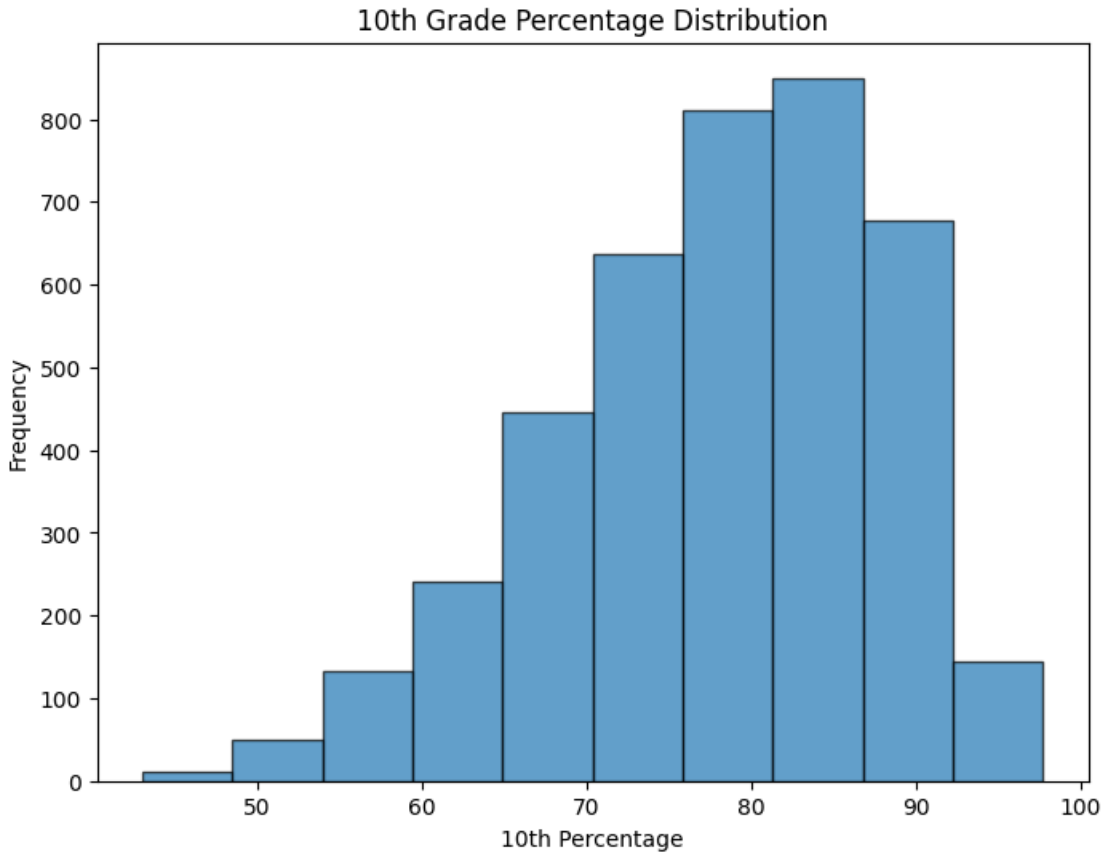


6 Different Types of Visualization

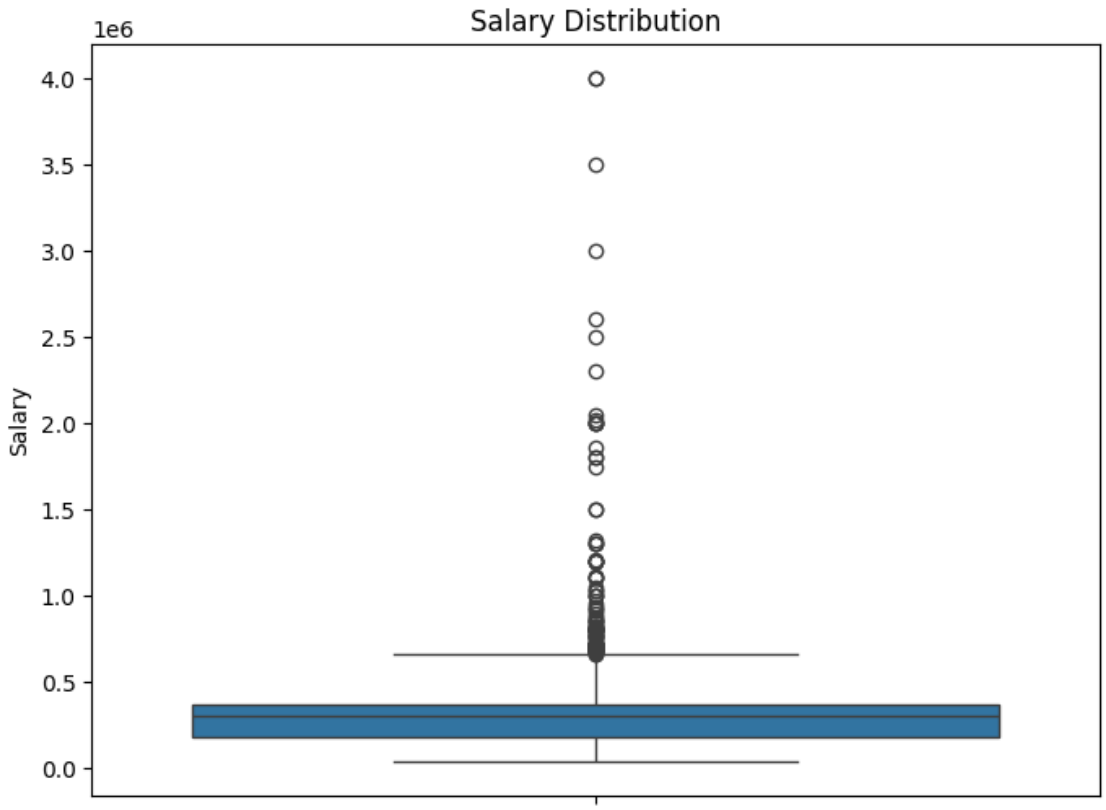
[32]:
```
# Gender distribution
plt.figure(figsize=(8,6))
sns.countplot(x='Gender', data=clean_train_data, color='purple')
plt.title('Count of Candidates by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



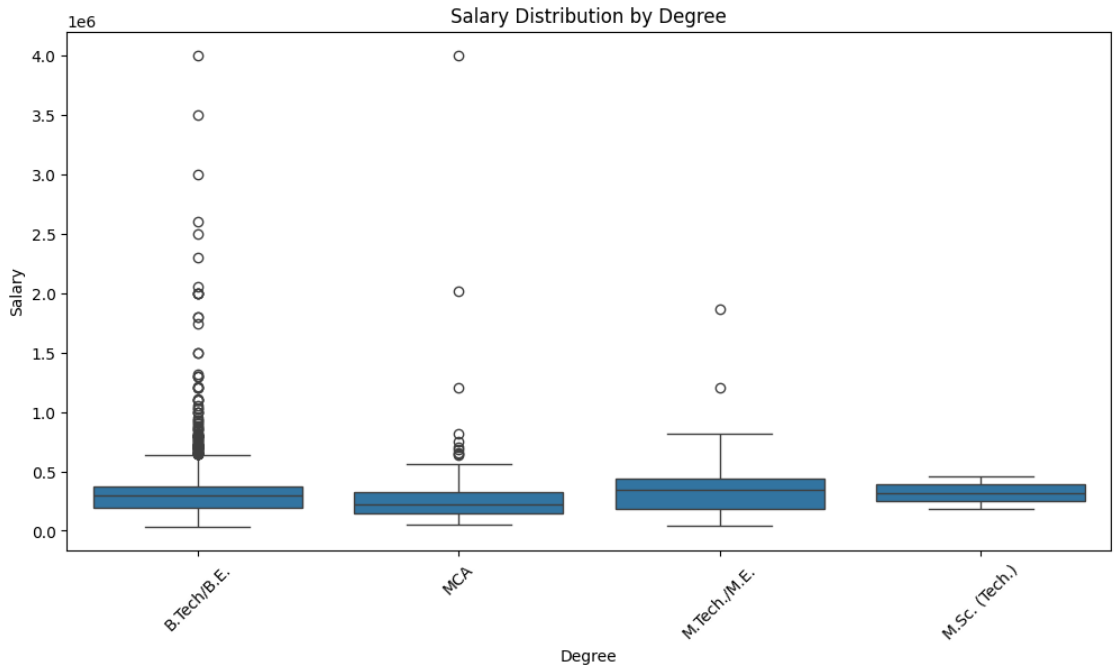
[33]:
```
# Distribution of 10th Grade Percentage
plt.figure(figsize=(8,6))
plt.hist(clean_train_data['10percentage'], bins=10, edgecolor='k', alpha=0.7)
plt.title('10th Grade Percentage Distribution')
plt.xlabel('10th Percentage')
plt.ylabel('Frequency')
plt.show()
```



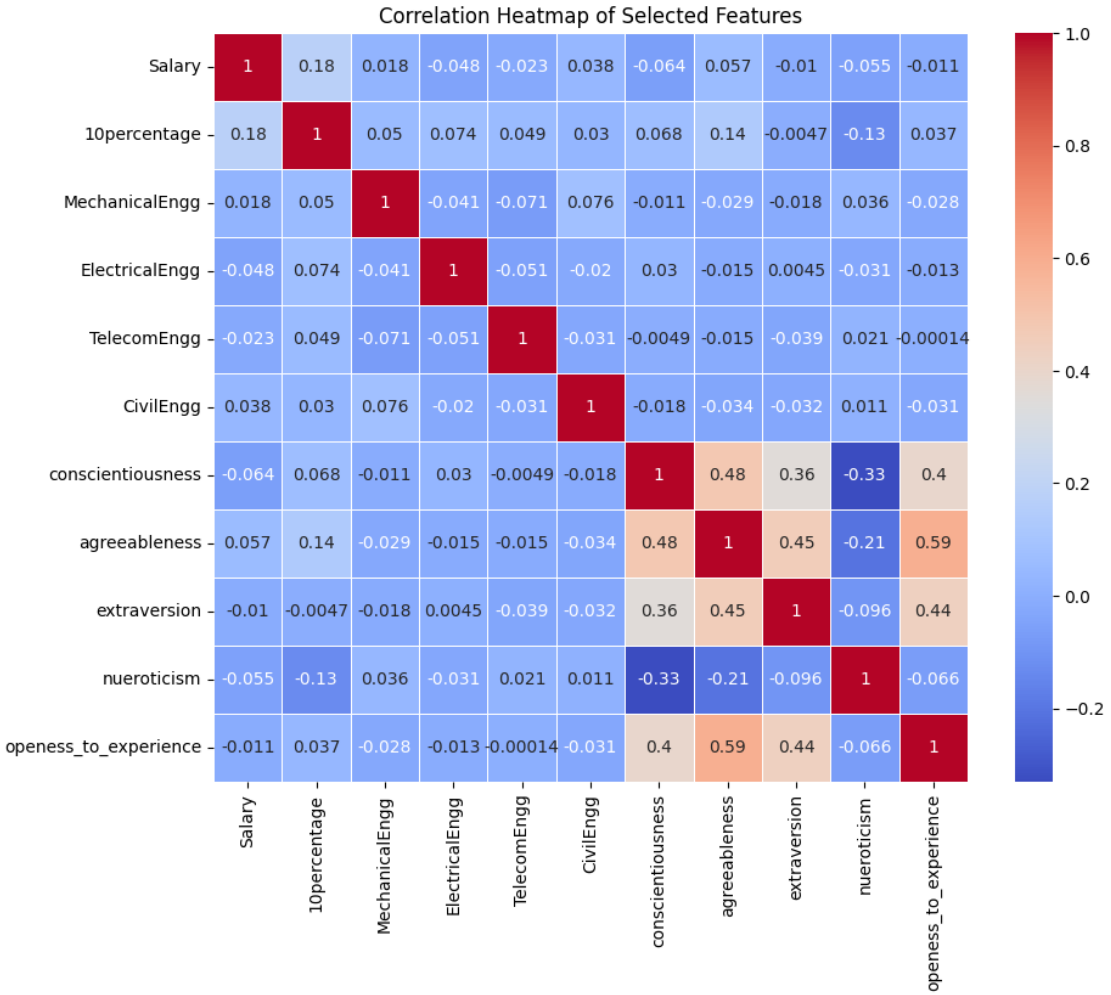
[34]:
```
# Boxplot for Salary Distribution
plt.figure(figsize=(8,6))
sns.boxplot(clean_train_data['Salary'])
plt.title('Salary Distribution')
plt.show()
```



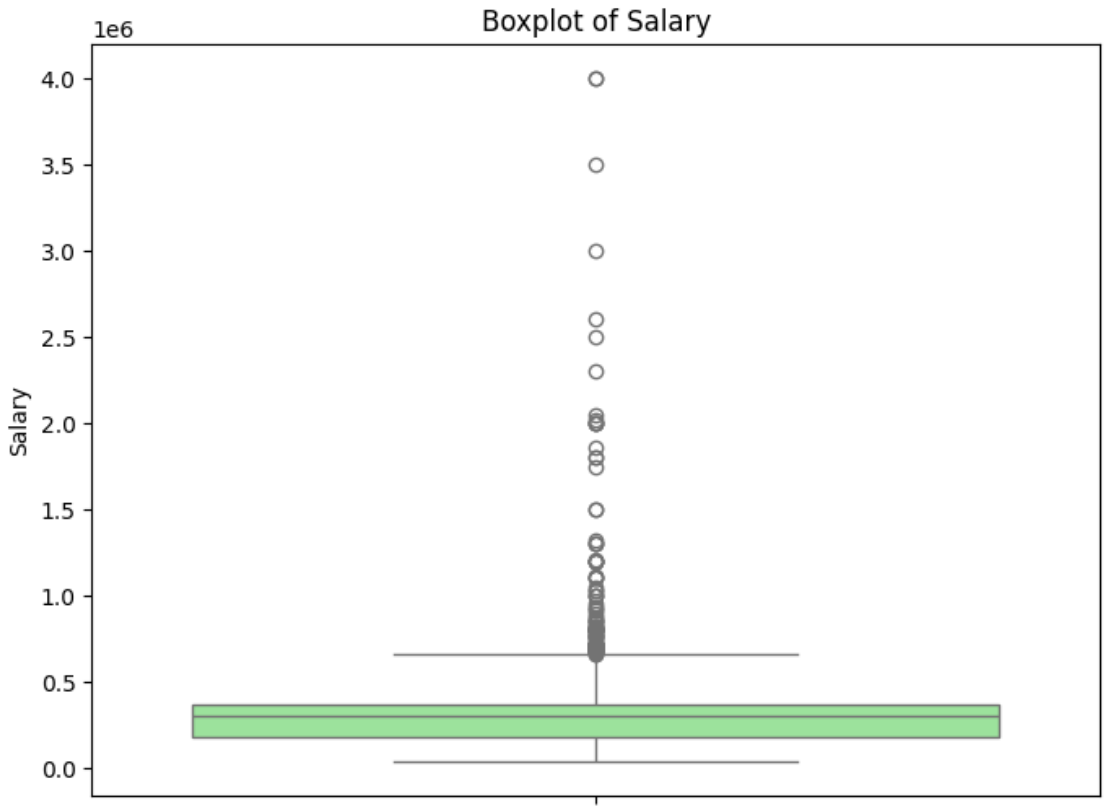
[35]:
```
# Salary by Degree
plt.figure(figsize=(12,6))
sns.boxplot(data=clean_train_data, x="Degree", y='Salary')
plt.title("Salary Distribution by Degree")
plt.xticks(rotation=45)
plt.show()
```



[36]:
```
# Correlation Heatmap for Selected Features
numerical_columns = ['Salary', '10percentage', 'MechanicalEngg',␣
 ↪'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness',␣
 ↪'agreeableness', 'extraversion', 'nueroticism', 'openess_to_experience']
corr_matrix = clean_train_data[numerical_columns].corr()

plt.figure(figsize=(10,8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap of Selected Features')
plt.show()
```

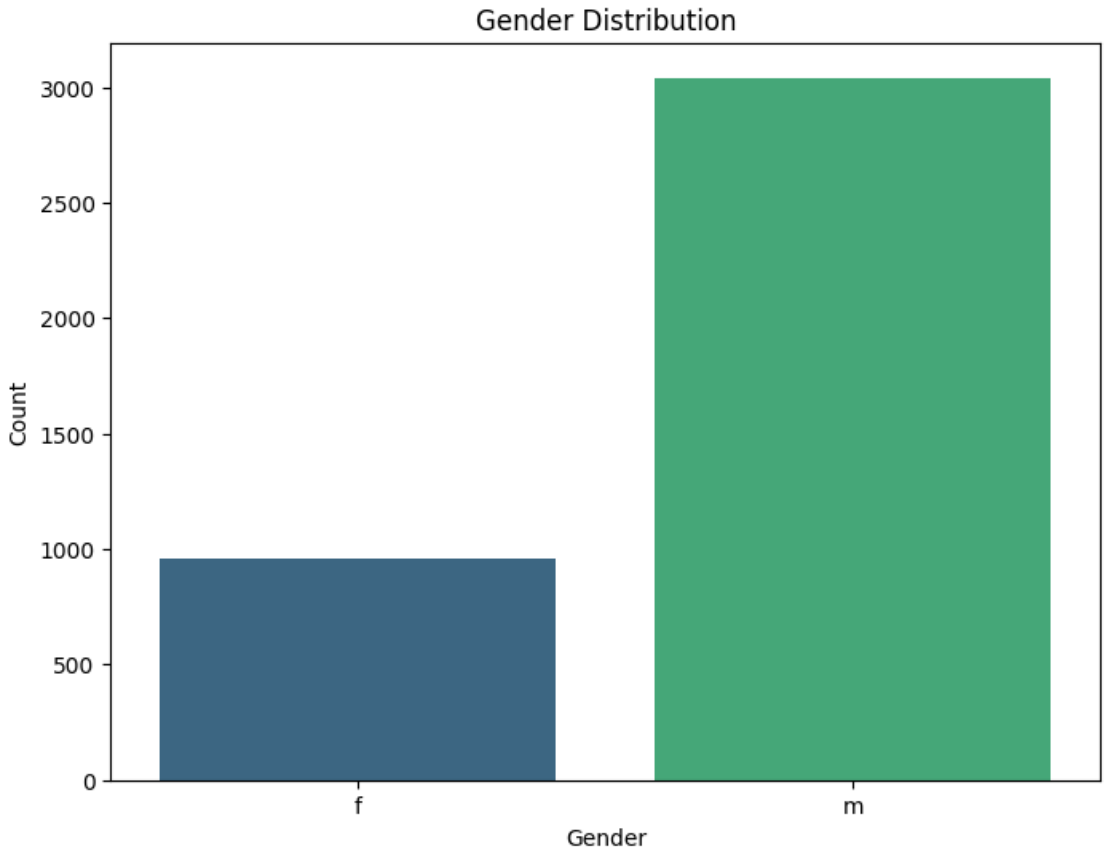


[38]:
```
#  Boxplot of Salary
plt.figure(figsize=(8, 6))
sns.boxplot(clean_train_data['Salary'], color='lightgreen')
plt.title('Boxplot of Salary')
plt.show()
```

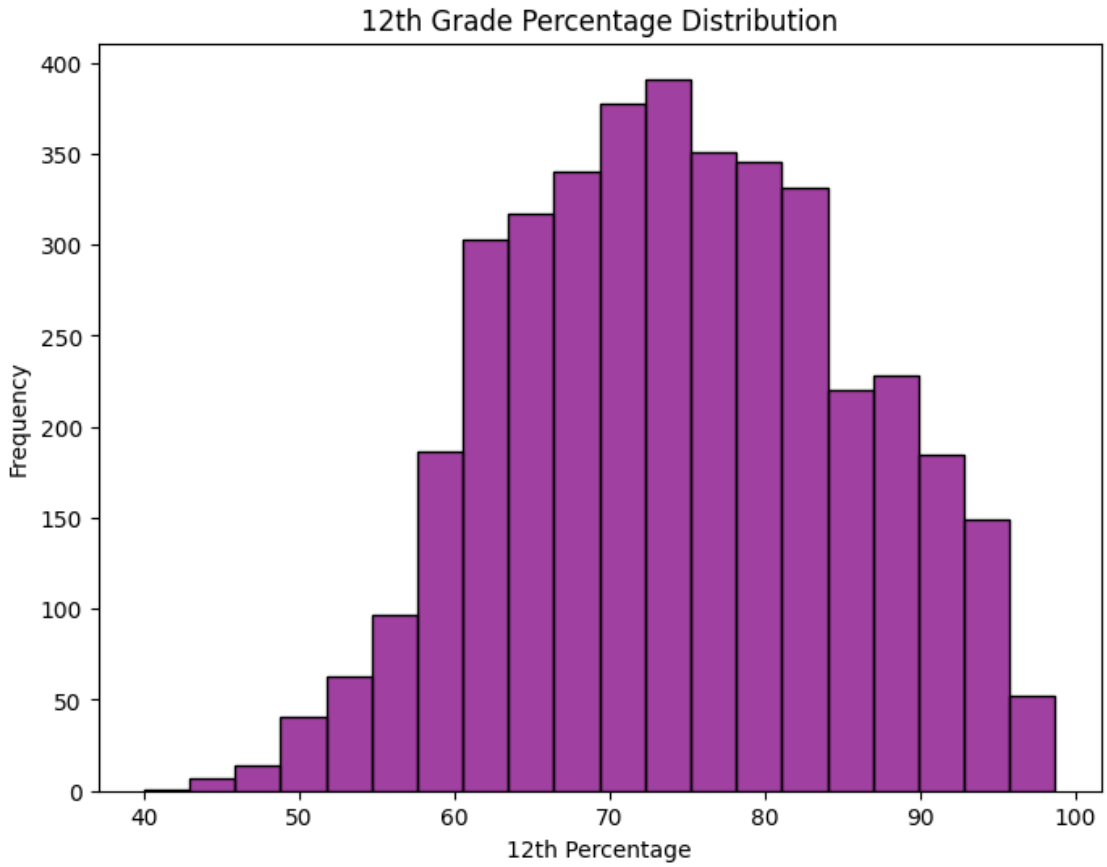


[39]:
```
#  Gender Distribution (Countplot)
plt.figure(figsize=(8, 6))
sns.countplot(x='Gender', data=clean_train_data, palette='viridis')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

```
C:\Users\payal\AppData\Local\Temp\ipykernel_2116\4058931214.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(x='Gender', data=clean_train_data, palette='viridis')
```

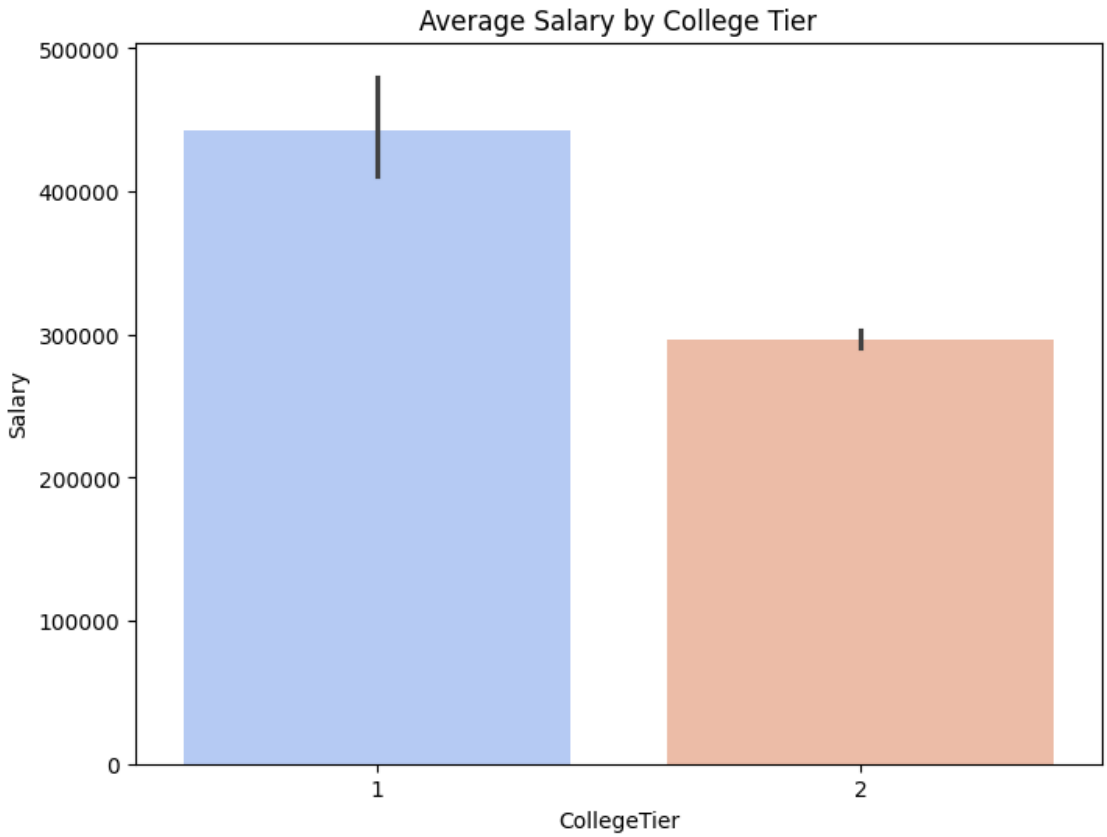


[40]:
```
# 12th Grade Percentage Distribution (Histogram)
plt.figure(figsize=(8, 6))
sns.histplot(clean_train_data['12percentage'], bins=20, color='purple')
plt.title('12th Grade Percentage Distribution')
plt.xlabel('12th Percentage')
plt.ylabel('Frequency')
plt.show()
```



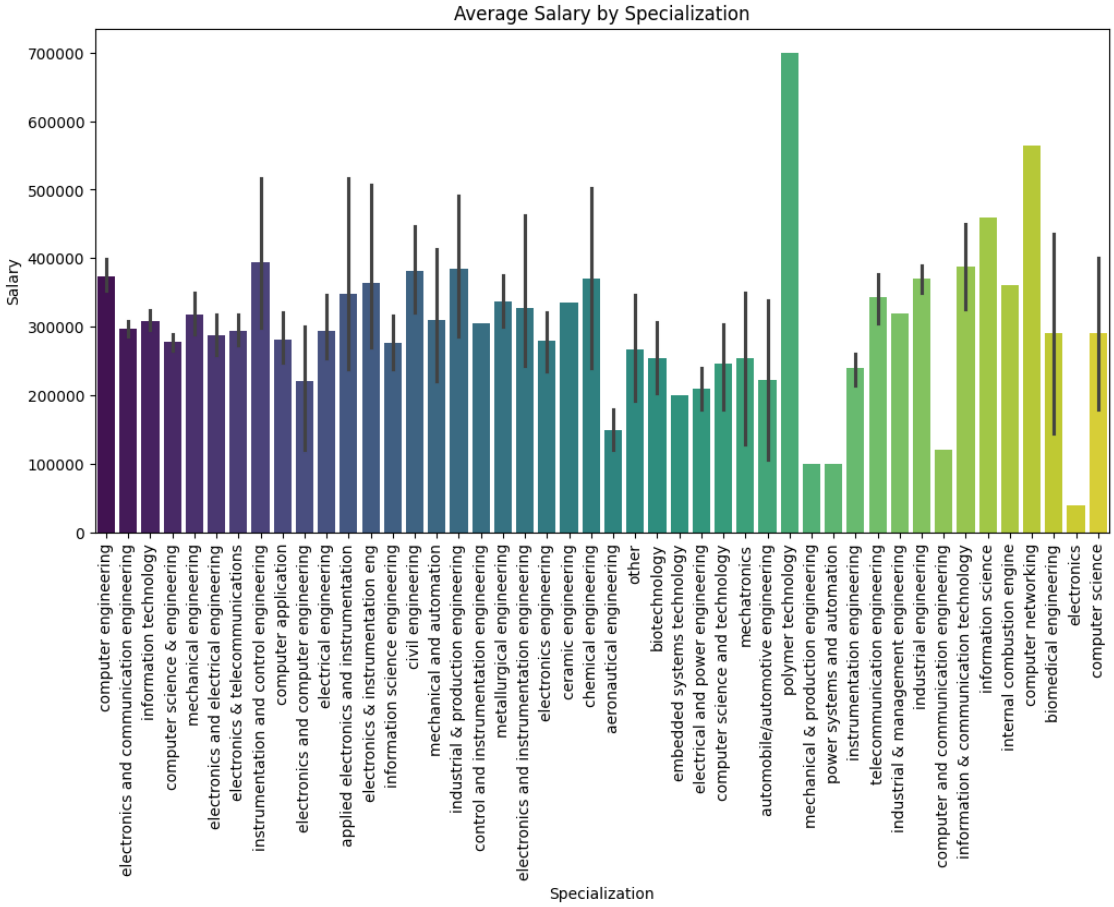
[41]:
```
#  Salary vs College Tier (Barplot)
plt.figure(figsize=(8, 6))
sns.barplot(data=clean_train_data, x="CollegeTier", y='Salary', estimator=np.
 ↪mean, palette='coolwarm')
plt.title('Average Salary by College Tier')
plt.show()
```

```
C:\Users\payal\AppData\Local\Temp\ipykernel_2116\1311747816.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(data=clean_train_data, x="CollegeTier", y='Salary',
estimator=np.mean, palette='coolwarm')
```

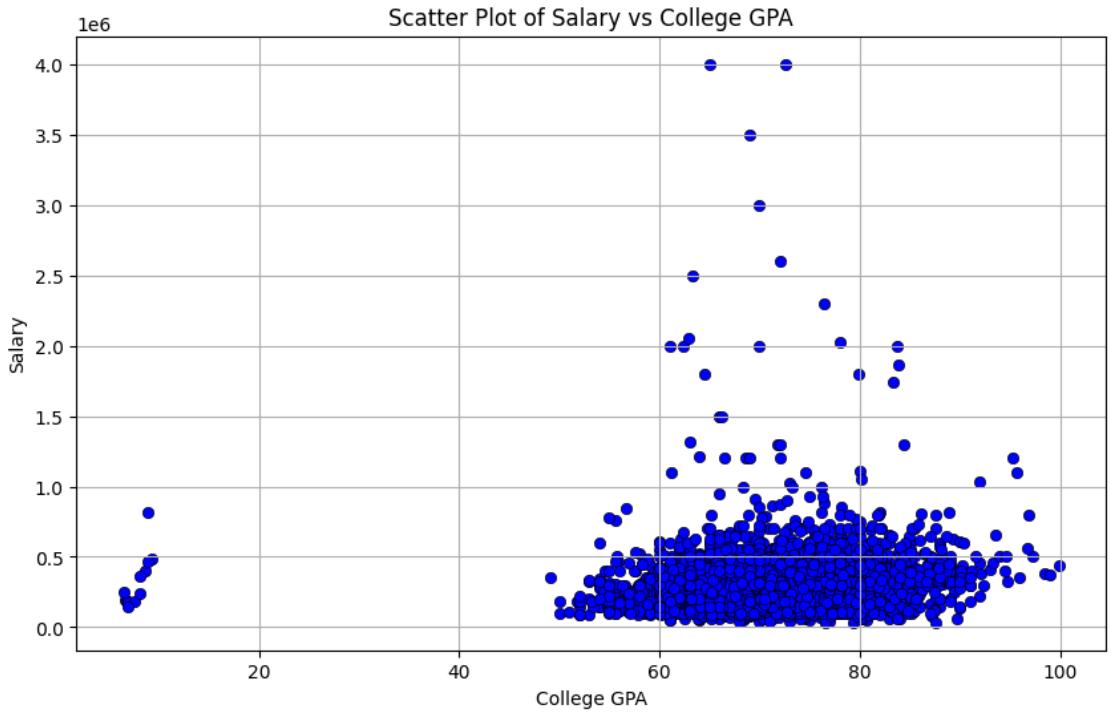


[44]:
```
#  Average Salary by Specialization (Barplot)
plt.figure(figsize=(12, 6))
sns.barplot(x='Specialization', y='Salary', data=clean_train_data, estimator=np.
 ↪mean, palette='viridis')
plt.title('Average Salary by Specialization')
plt.xticks(rotation=90)
plt.show()
```

```
C:\Users\payal\AppData\Local\Temp\ipykernel_2116\850905258.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(x='Specialization', y='Salary', data=clean_train_data,
estimator=np.mean, palette='viridis')
```

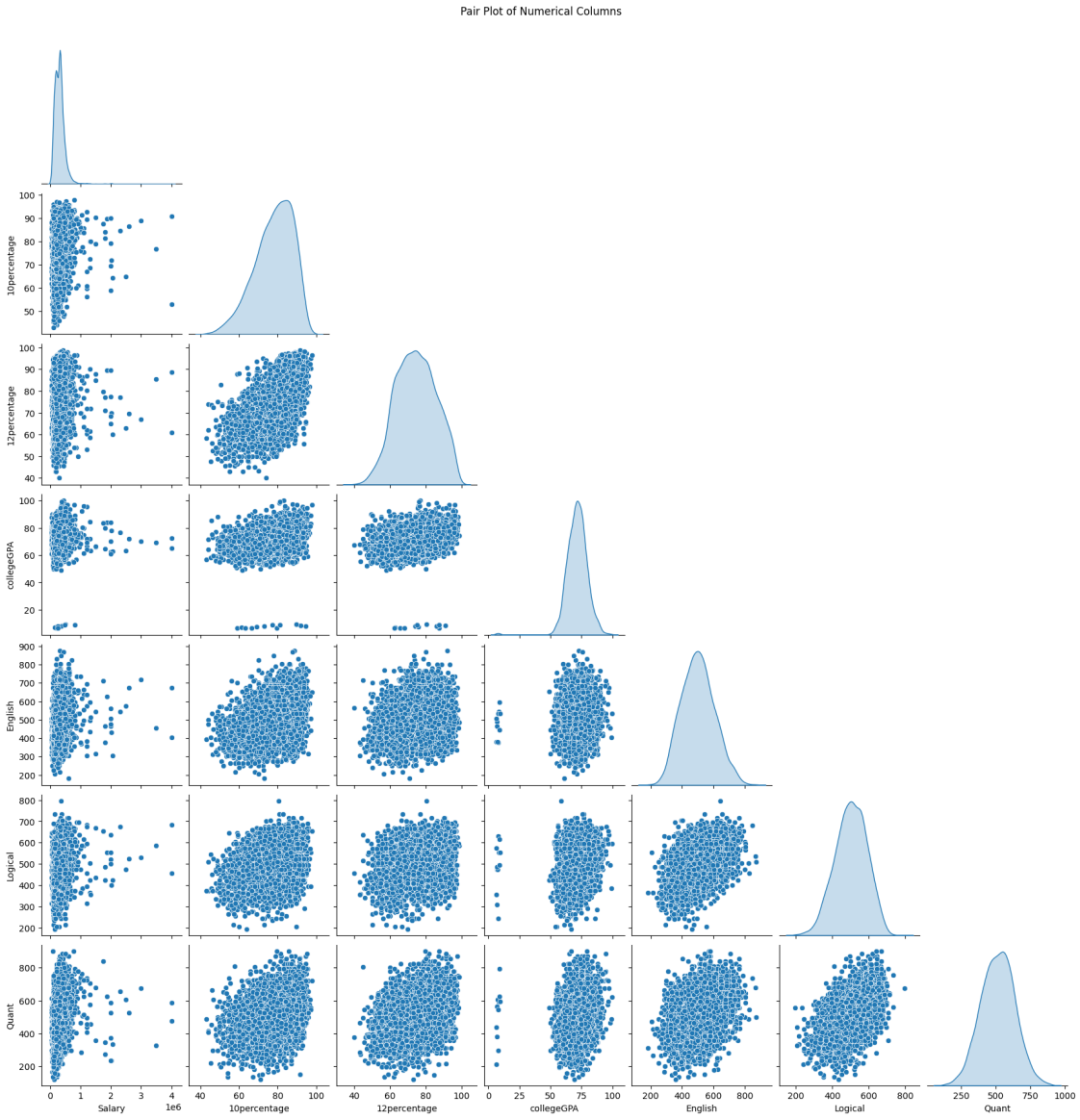


[45]:
```
# Scatter Plot of Salary vs College GPA
plt.figure(figsize=(10, 6))
sns.scatterplot(data=clean_train_data, x='collegeGPA', y='Salary', color='b',␣
↪edgecolor='black')
plt.title('Scatter Plot of Salary vs College GPA')
plt.xlabel('College GPA')
plt.ylabel('Salary')
plt.grid(True)
plt.show()
```

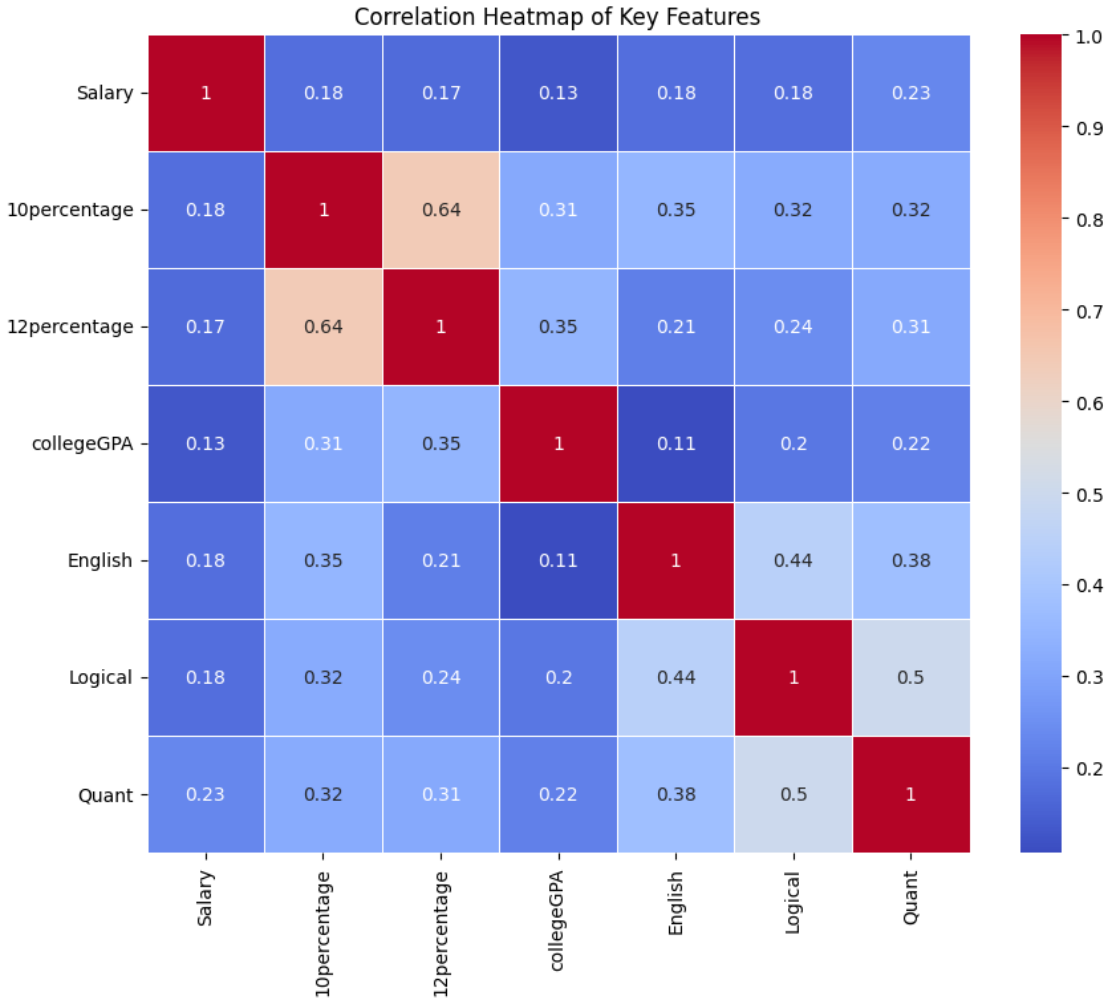


[46]:
```
#  Hexbin Plot of Salary vs College GPA
plt.figure(figsize=(10, 6))
plt.hexbin(clean_train_data['collegeGPA'], clean_train_data['Salary'],␣
 ↪gridsize=30, cmap='Blues')
plt.colorbar(label='Count in Bin')
plt.title('Hexbin Plot of Salary vs College GPA')
plt.xlabel('College GPA')
plt.ylabel('Salary')
plt.show()
```

[47]:
```
# Pair Plot of Numerical Columns
numerical_columns = ['Salary', '10percentage', '12percentage', 'collegeGPA',␣
 ↪'English', 'Logical', 'Quant']
sns.pairplot(clean_train_data[numerical_columns], diag_kind='kde', corner=True)
plt.suptitle('Pair Plot of Numerical Columns', y=1.02)
plt.show()
```



[48]:
```
# Correlation Heatmap of Key Features
plt.figure(figsize=(10, 8))
sns.heatmap(clean_train_data[numerical_columns].corr(), annot=True,␣
 ↪cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap of Key Features')
plt.show()
```



[51]:
```
# Salary vs Designation (Swarm Plot)
plt.figure(figsize=(12, 6))
sns.swarmplot(x='Designation', y='Salary', data=clean_train_data,␣
 ↪palette='tab10')
plt.title('Salary vs Designation (Swarm Plot)')
plt.xticks(rotation=90)
plt.show()
```

```
C:\Users\payal\AppData\Local\Temp\ipykernel_2116\975783360.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  sns.swarmplot(x='Designation', y='Salary', data=clean_train_data,
palette='tab10')
```

```
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 20.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 88.5% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 94.9% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 88.9% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 64.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 98.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 93.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 57.1% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 73.9% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 96.1% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 50.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 88.6% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

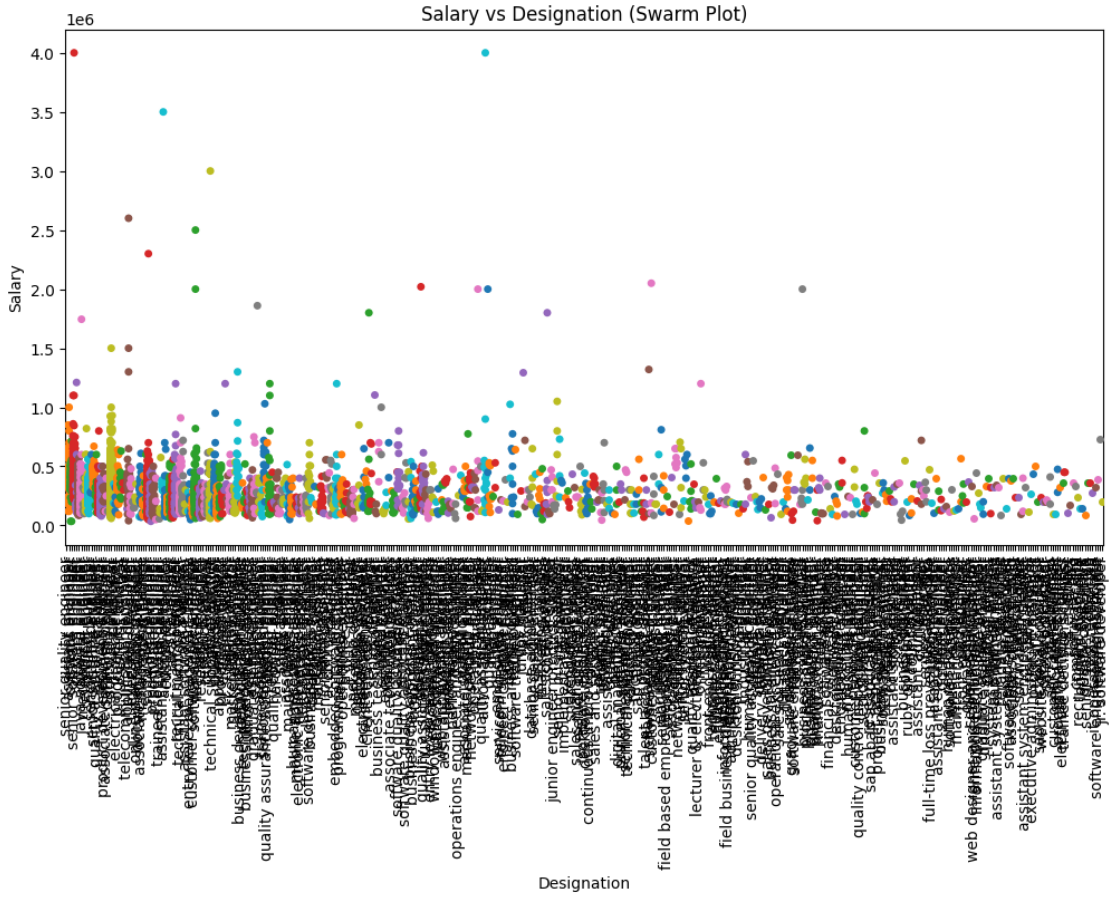
```
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 78.6% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 82.8% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 92.2% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 69.2% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 25.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 66.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 98.1% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 95.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 63.6% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 80.8% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 61.5% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 75.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

```
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 33.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 89.1% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 84.2% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 97.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 88.2% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 79.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 60.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 91.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 87.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 94.1% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 91.2% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 94.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

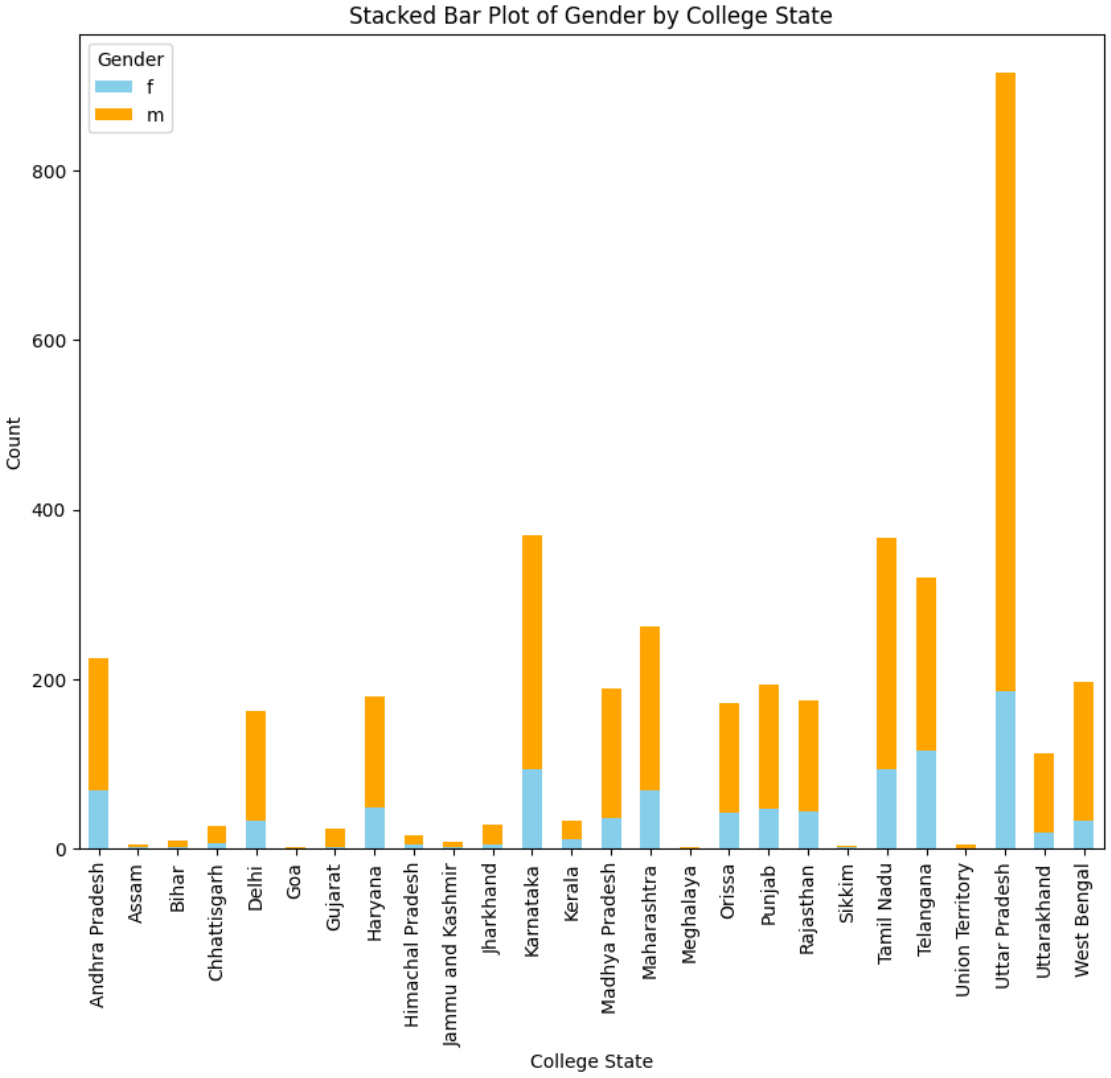
```
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 83.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 90.8% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 89.8% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 93.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 87.9% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 86.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 58.8% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 80.6% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 85.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 96.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 77.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 62.5% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

```
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 80.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 40.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 55.6% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 76.9% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 71.4% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 83.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 77.8% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 81.8% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 87.5% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 84.6% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 82.4% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 56.2% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

```
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 42.9% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 90.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 64.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 72.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 53.3% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 70.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 28.6% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 16.7% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\payal\AppData\Local\Programs\Python\Python311\Lib\site-
packages\seaborn\categorical.py:3399: UserWarning: 92.0% of the points cannot be
placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```



[49]:
```
#  Stacked Bar Plot of Gender by College State
pivot_table = clean_train_data.pivot_table(index='CollegeState',␣
 ↪columns='Gender', values='Salary', aggfunc='count').fillna(0)
pivot_table.plot(kind='bar', stacked=True, figsize=(10, 8), color=['skyblue',␣
 ↪'orange'])
plt.title('Stacked Bar Plot of Gender by College State')
plt.xlabel('College State')
plt.ylabel('Count')
plt.show()
```


Stacked Bar Plot of Gender by College State
Gender
f
m
Count
0
200
400
600
800
Andhra Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Meghalaya
Orissa
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Union Territory
Uttar Pradesh
Uttarakhand
West Bengal
College State

[ ]: