raman-project

February 23, 2024

[8]:
```
!pip install pandas
```

```
Requirement already satisfied: pandas in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (2.2.0)
Requirement already satisfied: numpy<2,>=1.23.2 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from pandas)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
pandas) (2023.4)
Requirement already satisfied: six>=1.5 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from python-
dateutil>=2.8.2->pandas) (1.16.0)
```

[9]:
```
!pip install matplotlib
```

```
Requirement already satisfied: matplotlib in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (3.8.2)
Requirement already satisfied: contourpy>=1.0.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib) (4.48.1)
Requirement already satisfied: kiwisolver>=1.3.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib) (1.4.5)
Requirement already satisfied: numpy<2,>=1.21 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
```

```
matplotlib) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from matplotlib)
(23.1)
Requirement already satisfied: pillow>=8 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from matplotlib)
(2.8.2)
Requirement already satisfied: six>=1.5 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from python-
dateutil>=2.7->matplotlib) (1.16.0)
```

[10]:
```
!pip install openpyxl
```

```
Requirement already satisfied: openpyxl in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (3.1.2)
Requirement already satisfied: et-xmlfile in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
openpyxl) (1.1.0)
```

[11]:
```
import pandas as pd
```

```
C:\Users\raman\AppData\Local\Temp\ipykernel_11060\4080736814.py:1:
DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of
pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better
interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

  import pandas as pd
```

1 step:-2 Import the data and display the head, shape and description of the data.

[12]:
```
dataset=pd.read_excel("data.xlsx")
```

[13]:
```
dataset.head()
```

[13]:
```
  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]
```

[14]:
```
dataset.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   int64
```

```
 3   DOJ                    3998 non-null   datetime64[ns]
 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   datetime64[ns]
 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: datetime64[ns](2), float64(9), int64(18), object(10)
memory usage: 1.2+ MB
```

[15]:
```
dataset.shape
```

[15]:
```
(3998, 39)
```

[16]:
```
dataset.describe()
```

[16]:
```
                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]
```

2 step:-1 Introduction -> Give a detailed data description and objective

[17]:
```
df = pd.read_excel("data.xlsx")

# Display basic information about the dataset
print("Data Description:")
print(df.info())

# Display summary statistics for numerical columns
print("\nSummary Statistics:")
print(df.describe())

# Display unique values and their counts for categorical columns
print("\nUnique Values for Categorical Columns:")
for column in df.select_dtypes(include=['object']):
    print(f"\n{column}:")
    print(df[column].value_counts())
```

```
Data Description:
<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   int64
 3   DOJ                    3998 non-null   datetime64[ns]
 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   datetime64[ns]
 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: datetime64[ns](2), float64(9), int64(18), object(10)
memory usage: 1.2+ MB
None

Summary Statistics:
                  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]

Unique Values for Categorical Columns:

Unnamed: 0:
Unnamed: 0
train    3998
Name: count, dtype: int64

DOL:
DOL
present                1875
2015-04-01 00:00:00     573
2015-03-01 00:00:00     124
2015-05-01 00:00:00     112
2015-01-01 00:00:00      99
                        ...
2005-03-01 00:00:00       1
2015-10-01 00:00:00       1
2010-02-01 00:00:00       1
2011-02-01 00:00:00       1
2010-10-01 00:00:00       1
Name: count, Length: 67, dtype: int64

Designation:
Designation
software engineer                    539
software developer                   265
system engineer                      205
programmer analyst                   139
systems engineer                     118
                                    ...
cad drafter                            1
noc engineer                           1
human resources intern                 1
senior quality assurance engineer      1
jr. software developer                 1
Name: count, Length: 419, dtype: int64

JobCity:
JobCity
Bangalore           627
-1                  461
Noida               368
Hyderabad           335
Pune                290
                   ...
Tirunelvelli          1
```

```
Ernakulam             1
Nanded                1
Dharmapuri            1
Asifabadbanglore      1
Name: count, Length: 339, dtype: int64

Gender:
Gender
m    3041
f     957
Name: count, dtype: int64

10board:
10board
cbse                          1395
state board                   1164
0                              350
icse                           281
ssc                            122
                               …
hse,orissa                       1
national public school           1
nagpur board                     1
jharkhand academic council       1
bse,odisha                       1
Name: count, Length: 275, dtype: int64

12board:
12board
cbse                                1400
state board                         1254
0                                    359
icse                                 129
up board                              87
                                      …
jawahar higher secondary school        1
nagpur board                           1
bsemp                                  1
board of higher secondary orissa       1
boardofintermediate                    1
Name: count, Length: 340, dtype: int64

Degree:
Degree
B.Tech/B.E.      3700
MCA               243
M.Tech./M.E.       53
M.Sc. (Tech.)       2
```

```
Name: count, dtype: int64

Specialization:
Specialization
electronics and communication engineering      880
computer science & engineering                 744
information technology                         660
computer engineering                           600
computer application                           244
mechanical engineering                         201
electronics and electrical engineering         196
electronics & telecommunications               121
electrical engineering                          82
electronics & instrumentation eng               32
civil engineering                               29
electronics and instrumentation engineering     27
information science engineering                 27
instrumentation and control engineering         20
electronics engineering                         19
biotechnology                                   15
other                                           13
industrial & production engineering             10
applied electronics and instrumentation          9
chemical engineering                             9
computer science and technology                  6
telecommunication engineering                    6
mechanical and automation                        5
automobile/automotive engineering                5
instrumentation engineering                      4
mechatronics                                     4
aeronautical engineering                         3
electronics and computer engineering             3
electrical and power engineering                 2
biomedical engineering                           2
information & communication technology           2
industrial engineering                           2
computer science                                 2
metallurgical engineering                        2
power systems and automation                     1
control and instrumentation engineering          1
mechanical & production engineering              1
embedded systems technology                      1
polymer technology                               1
computer and communication engineering           1
information science                              1
internal combustion engine                       1
computer networking                              1
ceramic engineering                              1
```

```
electronics                                    1
industrial & management engineering            1
Name: count, dtype: int64

CollegeState:
CollegeState
Uttar Pradesh        915
Karnataka            370
Tamil Nadu           367
Telangana            319
Maharashtra          262
Andhra Pradesh       225
West Bengal          196
Punjab               193
Madhya Pradesh       189
Haryana              180
Rajasthan            174
Orissa               172
Delhi                162
Uttarakhand          113
Kerala                33
Jharkhand             28
Chhattisgarh          27
Gujarat               24
Himachal Pradesh      16
Bihar                 10
Jammu and Kashmir      7
Assam                  5
Union Territory        5
Sikkim                 3
Meghalaya              2
Goa                    1
Name: count, dtype: int64
```

3 Step :- 3 Univariate Analysis -> PDF, Histograms, Boxplots, Countplots,

[18]:
```
!pip install seaborn
```

```
Requirement already satisfied: seaborn in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages
(0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
seaborn) (1.26.4)
Requirement already satisfied: pandas>=1.2 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
```

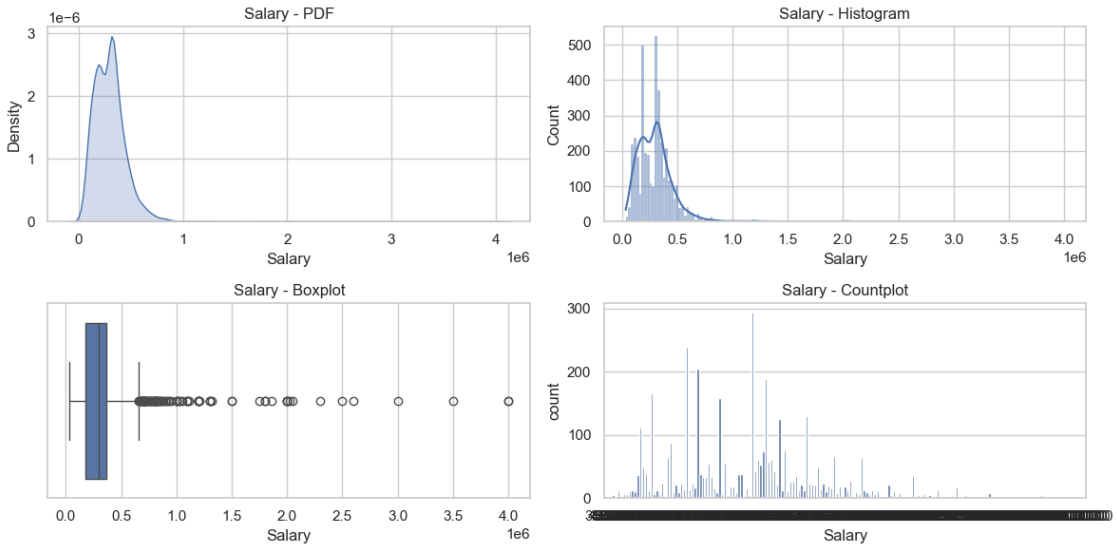
```
seaborn) (2.2.0)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
seaborn) (3.8.2)
Requirement already satisfied: contourpy>=1.0.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (1.2.0)
Requirement already satisfied: cycler>=0.10 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (4.48.1)
Requirement already satisfied: kiwisolver>=1.3.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (1.4.5)
Requirement already satisfied: packaging>=20.0 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (23.1)
Requirement already satisfied: pillow>=8 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
pandas>=1.2->seaborn) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in
c:\users\raman\appdata\local\programs\python\python311\lib\site-packages (from
pandas>=1.2->seaborn) (2023.4)
Requirement already satisfied: six>=1.5 in
c:\users\raman\appdata\roaming\python\python311\site-packages (from python-
dateutil>=2.7->matplotlib!=3.6.1,>=3.4->seaborn) (1.16.0)
```

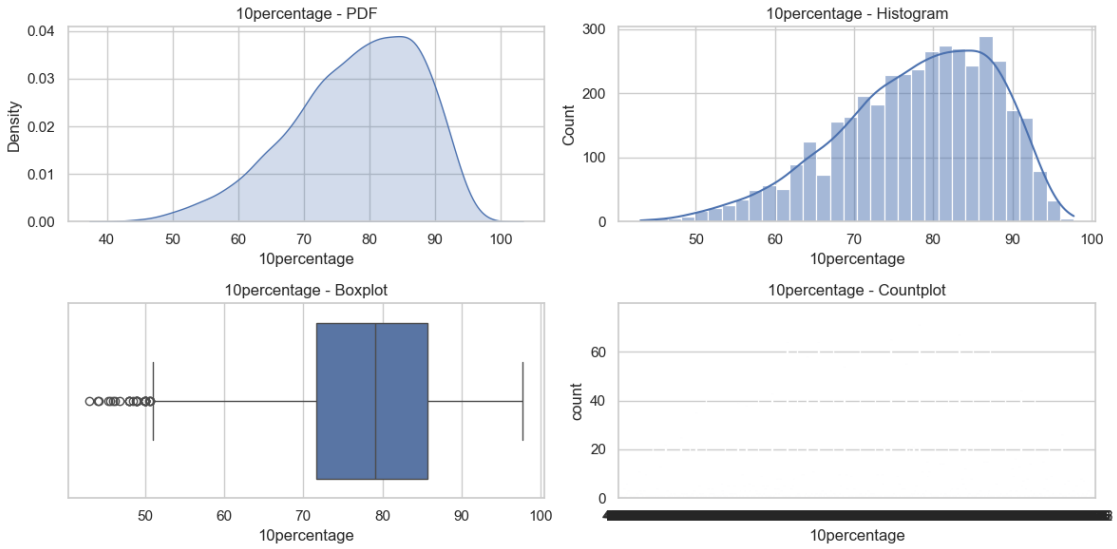
[19]:
```
import matplotlib.pyplot as plt
```

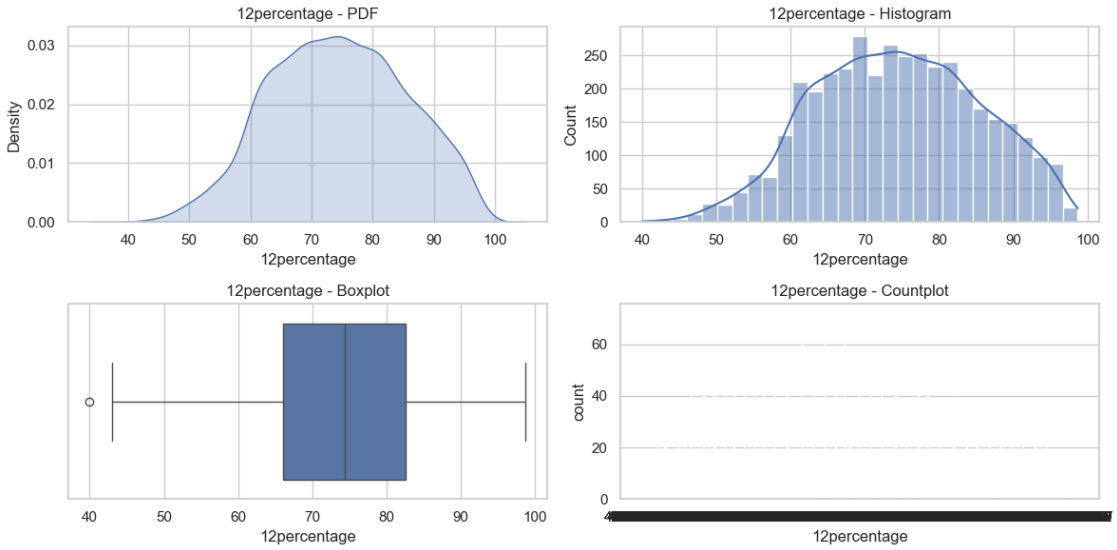
[20]:
```
import seaborn as sns
```

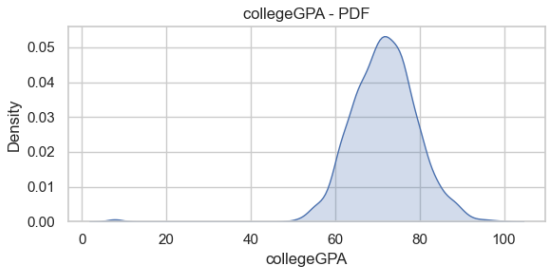
[21]:
```
sns.set(style="whitegrid")

# Define the columns for univariate analysis
columns_for_analysis = ['Salary', '10percentage', '12percentage', 'collegeGPA',␣
 ↪'conscientiousness', 'agreeableness',
```

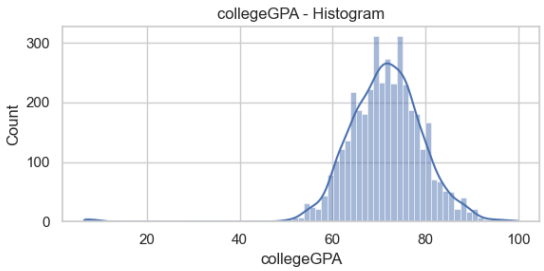
```
                        'extraversion', 'nueroticism', 'openess_to_experience']

for column in columns_for_analysis:
    plt.figure(figsize=(12, 6))

    # Probability Density Function (PDF)
    plt.subplot(2, 2, 1)
    sns.kdeplot(df[column], fill=True)
    plt.title(f'{column} - PDF')

    # Histogram
    plt.subplot(2, 2, 2)
    sns.histplot(df[column], kde=True)
    plt.title(f'{column} - Histogram')

    # Boxplot
    plt.subplot(2, 2, 3)
    sns.boxplot(x=df[column])
    plt.title(f'{column} - Boxplot')

    # Countplot
    plt.subplot(2, 2, 4)
    sns.countplot(x=df[column])
    plt.title(f'{column} - Countplot')

    plt.tight_layout()
    plt.show()
```

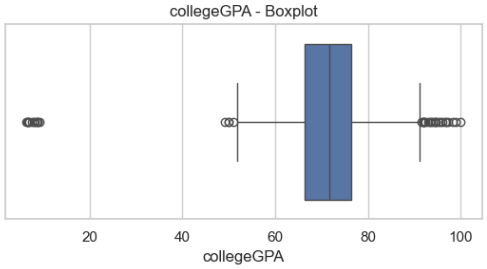


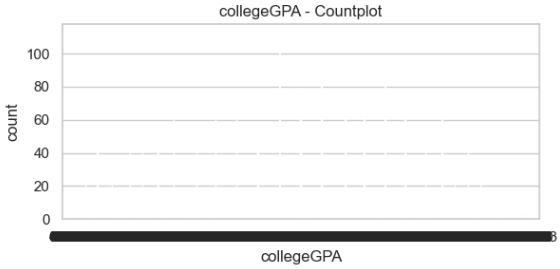

10percentage - PDF
Density
10percentage
10percentage - Histogram
Count
10percentage
10percentage - Boxplot
10percentage
10percentage - Countplot
count
10percentage

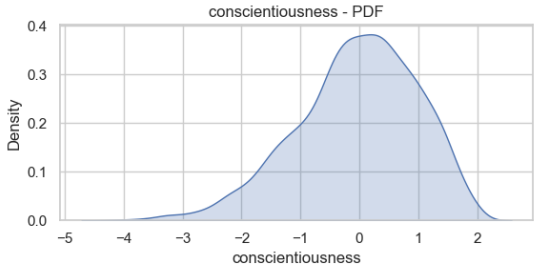

12percentage - PDF
Density
12percentage
12percentage - Histogram
Count
12percentage
12percentage - Boxplot
12percentage
12percentage - Countplot
count
12percentage

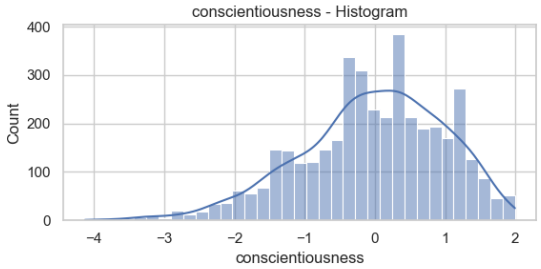

collegeGPA - PDF
Density
collegeGPA

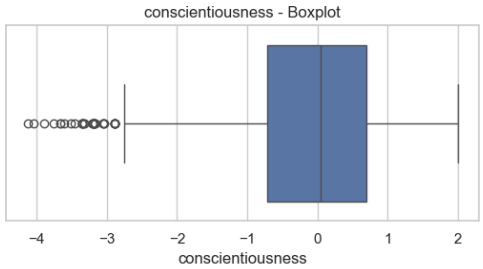

collegeGPA - Histogram
Count
collegeGPA

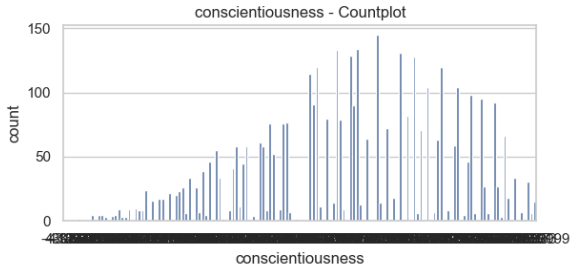

collegeGPA - Boxplot
collegeGPA

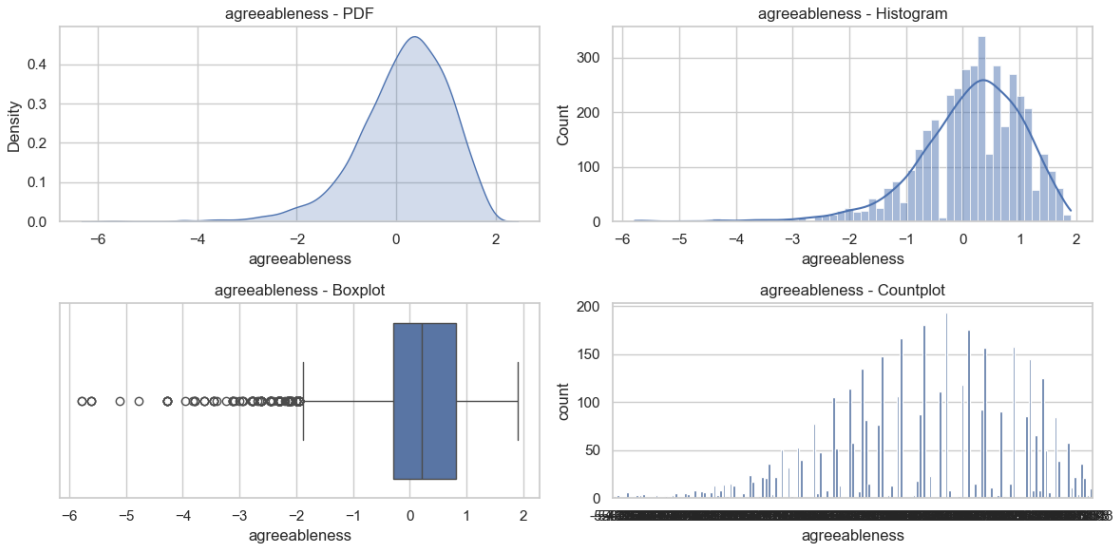

collegeGPA - Countplot
count
collegeGPA

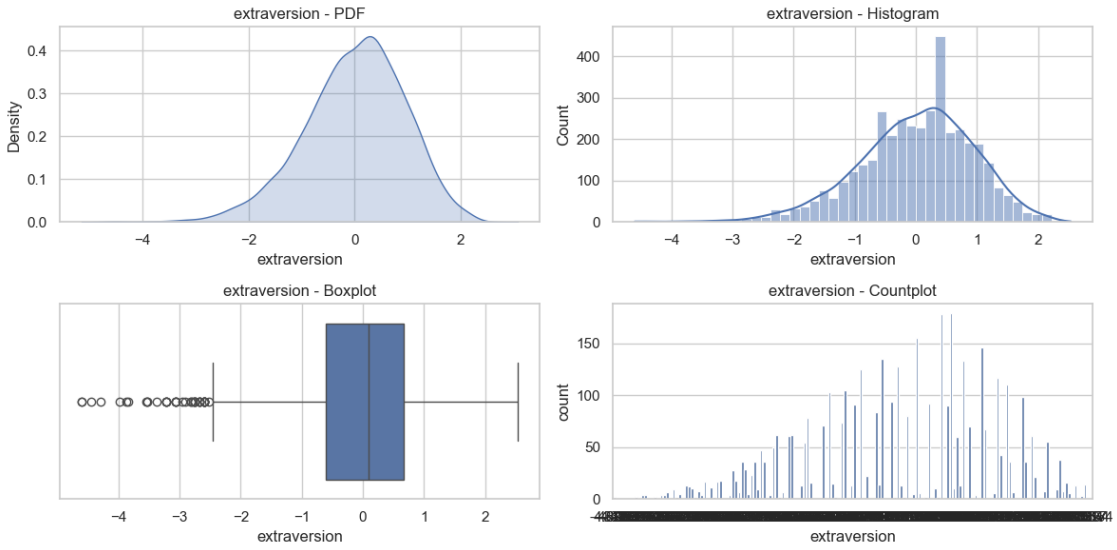

conscientiousness - PDF
Density
conscientiousness

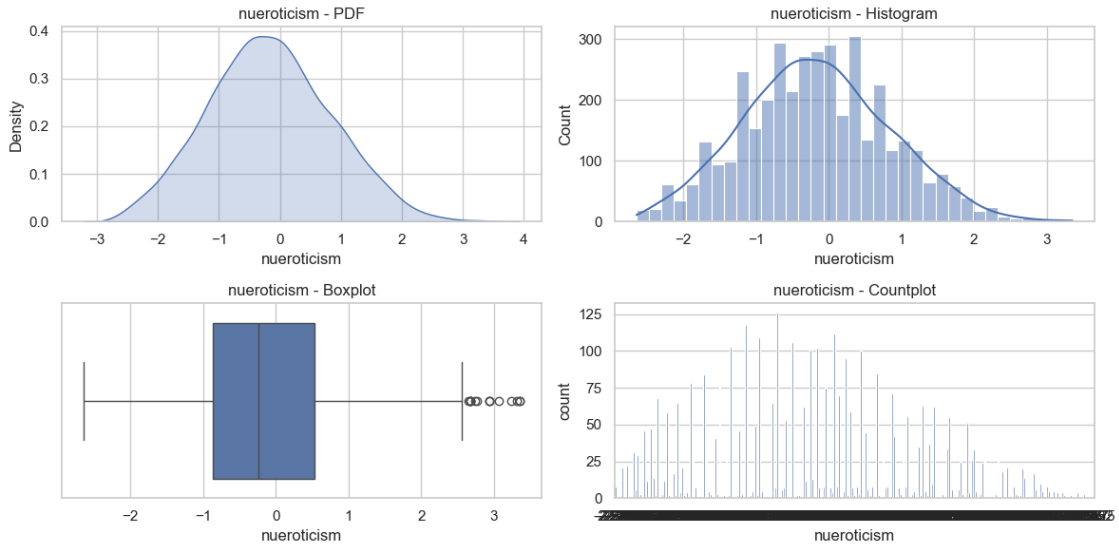

conscientiousness - Histogram
Count
conscientiousness

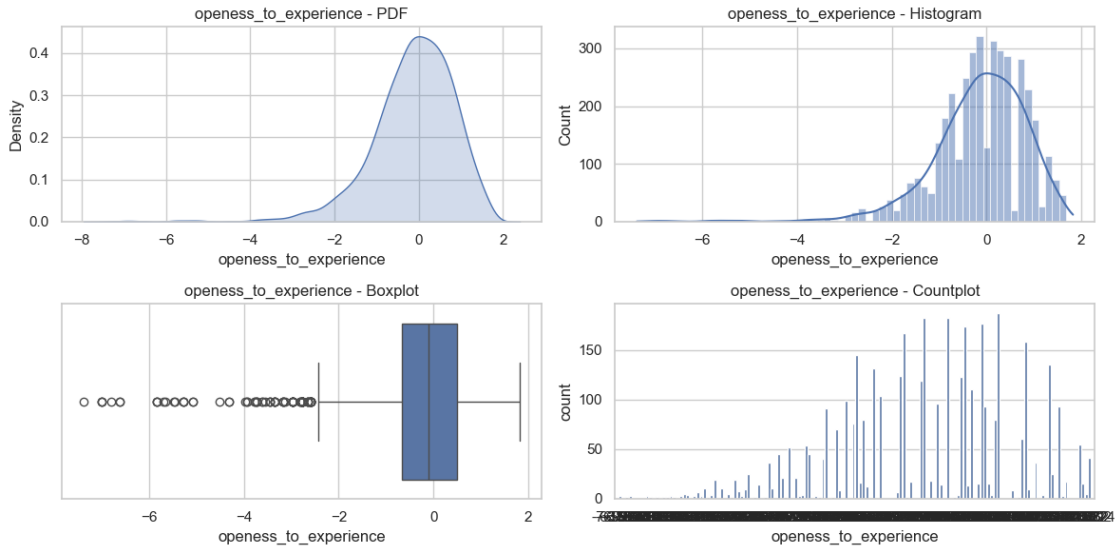

conscientiousness - Boxplot
conscientiousness


conscientiousness - Countplot
count
conscientiousness


agreeableness - PDF
Density
agreeableness
agreeableness - Histogram
Count
agreeableness
agreeableness - Boxplot
agreeableness
agreeableness - Countplot
count
agreeableness


extraversion - PDF
Density
extraversion
extraversion - Histogram
Count
extraversion
extraversion - Boxplot
extraversion
extraversion - Countplot
count
extraversion





[22]:
```python
# Get a list of numerical columns
numerical_columns = df.select_dtypes(exclude=['object']).columns.tolist()

# Print the list of numerical columns
print("Numerical Columns:")
print(numerical_columns)

# Get the number of numerical columns
num_numerical_columns = len(numerical_columns)
print("\nNumber of Numerical Columns:", num_numerical_columns)
```

```
Numerical Columns:
['ID', 'Salary', 'DOJ', 'DOB', '10percentage', '12graduation', '12percentage',
'CollegeID', 'CollegeTier', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier',
'GraduationYear', 'English', 'Logical', 'Quant', 'Domain',
'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience',
'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg',
'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism',
'openess_to_experience']

Number of Numerical Columns: 29
```

3.1 find the outliers in the columns

[23]:
```python
def find_outliers_iqr(data):
    Q1 = data.quantile(0.25)
    Q3 = data.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = (data < lower_bound) | (data > upper_bound)
    return outliers

# Define numerical columns for outlier detection
numerical_columns = ['Salary', '10percentage', '12percentage', 'collegeGPA',␣
 ↪'conscientiousness', 'agreeableness',
                     'extraversion', 'nueroticism', 'openess_to_experience']

# Find outliers in each numerical column
outliers_dict = {}
for column in numerical_columns:
    outliers_dict[column] = df[find_outliers_iqr(df[column])][column]

# Display outliers for each numerical column
for column, outliers in outliers_dict.items():
    print(f'Outliers in {column}:')
    print(outliers)
    print('\n')
```

```
Outliers in Salary:
3       1100000
76       800000
92      1500000
123     1200000
128      675000
         …
3823     775000
3904     850000
3912     730000
```

```
3961     700000
3992     800000
Name: Salary, Length: 109, dtype: int64


Outliers in 10percentage:
245     50.60
466     44.16
490     44.00
491     45.60
502     48.00
600     49.00
613     48.00
898     49.00
919     48.80
1064    49.00
1102    49.00
1169    48.50
1193    48.00
1235    50.60
1334    43.00
1838    50.00
1845    49.00
1955    45.33
1976    46.24
2037    48.00
2215    50.50
2292    50.00
2432    50.00
2655    50.66
2885    46.80
2982    50.00
3284    50.00
3425    50.00
3690    46.00
3743    49.90
Name: 10percentage, dtype: float64


Outliers in 12percentage:
3337    40.0
Name: 12percentage, dtype: float64


Outliers in collegeGPA:
7        8.58
44      92.10
138      6.63
```

```
187     93.00
477     92.00
614     93.60
690     99.93
788      6.80
874     94.50
907     50.00
1029    92.30
1134    96.00
1264    97.30
1345    93.30
1419     6.85
1439     8.07
1510    96.70
1685    94.70
1767     7.56
2151     6.95
2152    95.30
2229     8.13
2293     9.30
2463    92.00
2662     8.88
2691     8.89
2703    94.00
2836    49.07
2880    92.00
2988    94.60
3151    98.40
3276    95.70
3293    51.00
3308     6.45
3323    96.90
3448    50.00
3833    91.60
3850    99.00
Name: collegeGPA, dtype: float64


Outliers in conscientiousness:
29     -3.1994
159    -2.8879
210    -3.1994
315    -3.6631
335    -3.6060
373    -3.3539
382    -3.3539
408    -3.3188
468    -2.8879
```

```
523    -3.1752
1211   -3.4624
1337   -3.0448
1353   -3.1752
1684   -3.3539
1687   -3.6631
1972   -2.8903
2005   -3.1994
2046   -3.1994
2101   -4.1267
2182   -3.3539
2223   -2.8903
2224   -2.8879
2377   -2.8879
2396   -2.8879
2569   -3.3539
3047   -3.1752
3150   -3.8933
3372   -3.1752
3407   -3.5085
3468   -2.8903
3473   -3.7496
3569   -3.0448
3629   -4.0369
3646   -3.1752
3650   -3.0315
3694   -3.3188
3697   -3.1994
3876   -2.8903
3910   -3.0448
Name: conscientiousness, dtype: float64


Outliers in agreeableness:
23     -2.1186
43     -2.4516
63     -2.6847
67     -3.7836
157    -2.1186
         …
3843   -2.1186
3855   -2.3073
3878   -2.4516
3939   -1.9521
3953   -2.0733
Name: agreeableness, Length: 123, dtype: float64
```

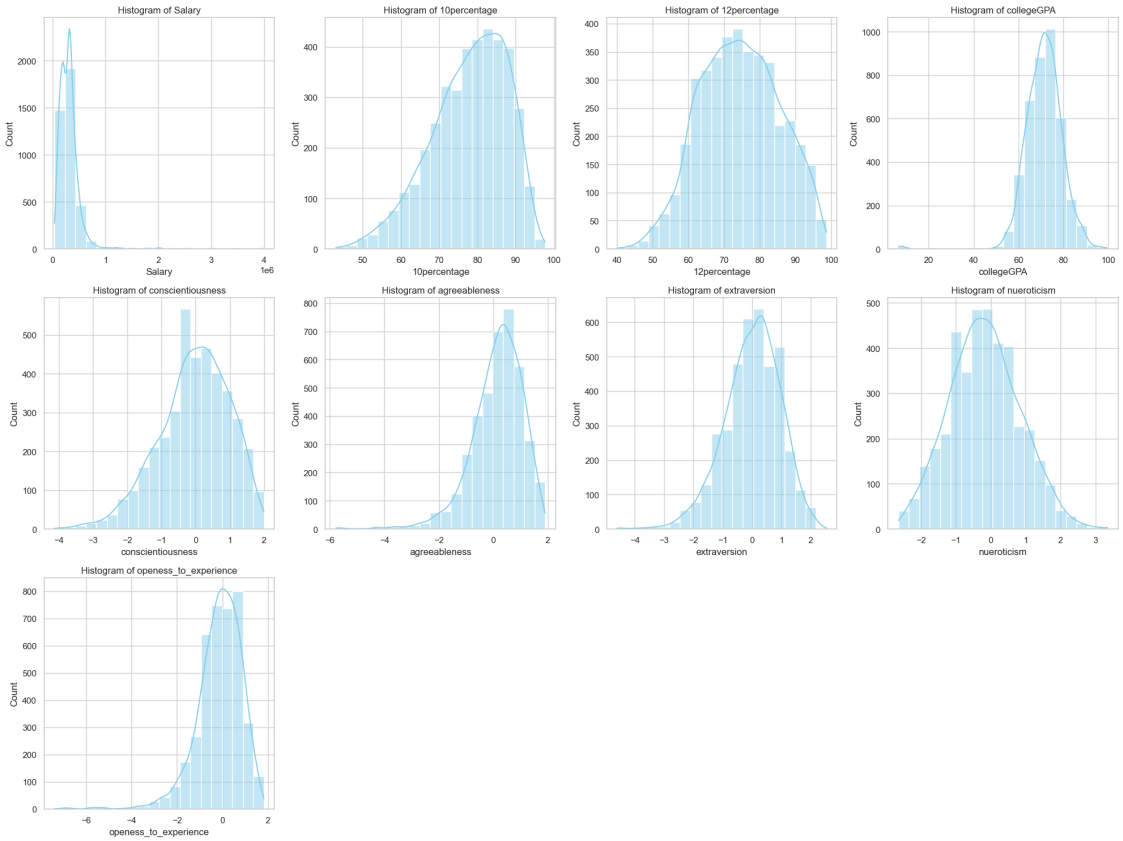
```
Outliers in extraversion:
63     -2.6028
159    -3.2176
335    -4.6009
408    -3.5250
523    -3.0639
666    -3.2176
726    -2.7750
1169   -2.6662
1211   -4.6009
1217   -2.6028
1242   -2.9565
1353   -4.2935
1538   -3.0639
1566   -2.7565
1649   -3.5370
1728   -3.5250
1785   -2.6662
1789   -2.7565
1822   -2.7565
1860   -3.8636
2060   -3.0639
2224   -2.7565
2305   -2.9102
2377   -3.2176
2396   -3.2176
2403   -3.3713
3141   -2.6662
3150   -3.9861
3174   -2.6662
3202   -2.5210
3263   -2.8113
3273   -2.8113
3372   -3.0639
3434   -2.6028
3473   -3.8324
3548   -2.6028
3595   -2.6028
3674   -2.6028
3694   -4.4472
3778   -2.6028
Name: extraversion, dtype: float64


Outliers in nueroticism:
222     2.6475
405     2.9349
1151    3.3525
```

```
1191    3.3525
1383    3.2350
1602    2.6814
1843    3.0617
2054    2.7650
2234    2.7356
2275    2.9349
2608    2.6814
2859    3.3152
3089    2.6814
3384    2.6814
3880    2.9349
Name: nueroticism, dtype: float64


Outliers in openess_to_experience:
22     -2.7769
23     -5.0763
43     -3.1602
63     -5.4770
128    -2.9731
         …
3868   -2.5853
3892   -2.7769
3901   -2.7769
3918   -2.9686
3957   -2.9731
Name: openess_to_experience, Length: 95, dtype: float64
```

3.2 Understand the probability and frequency distribution of each numerical column To understand the probability and frequency distribution of each numerical column,we can visualize histograms for each column

[24]:
```
import math
num_cols = len(numerical_columns)
num_subplot_rows = math.ceil(num_cols / 4)  # Assuming 4 columns per row
plt.figure(figsize=(20, num_subplot_rows * 5))  # Adjust the height based on␣
 ↪the number of rows

# Iterate through numerical columns
for i, column in enumerate(numerical_columns):
    # Create a subplot for each column
    plt.subplot(num_subplot_rows, 4, i + 1)  # Adjust the number of columns per␣
 ↪row
```

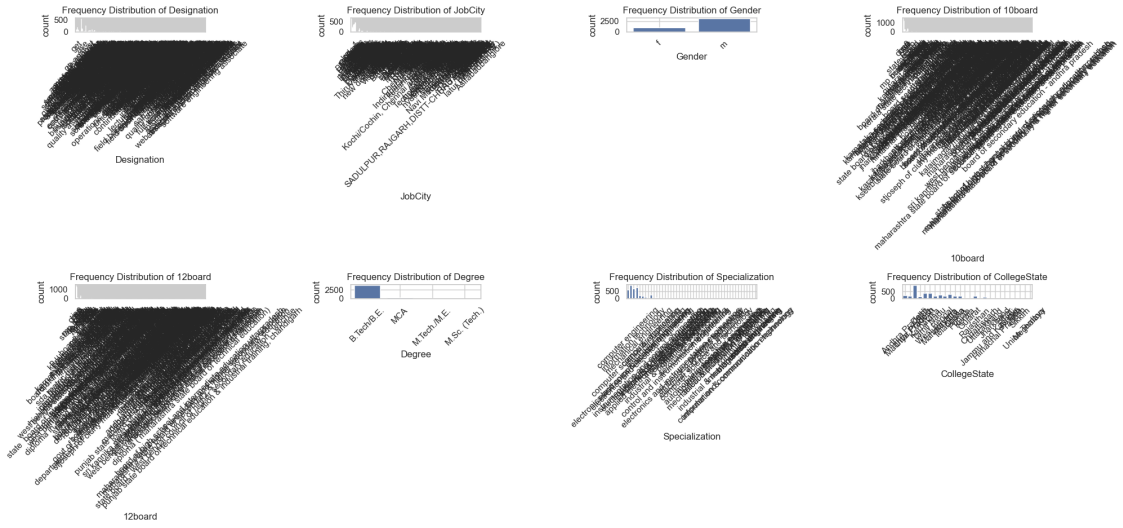
```
    # Plot histogram
    sns.histplot(df[column], kde=True, color='skyblue', bins=20)

    # Set title
    plt.title(f'Histogram of {column}')

    # Add grid for better visualization
    plt.grid(True)

plt.tight_layout()

plt.show()
```



3.3 To understand the frequency distribution of each categorical column, we can create bar plots showing the counts of unique values in each column.

[25]:
```
categorical_columns = ['Designation', 'JobCity', 'Gender', '10board',␣
↪'12board', 'Degree', 'Specialization', 'CollegeState']

plt.figure(figsize=(20, 15))  # Increase the height of the figure
```

```
# Iterate through categorical columns
for i, column in enumerate(categorical_columns):
    # Create a subplot for each column
    plt.subplot(4, 4, i + 1)

    # Plot bar plot
    sns.countplot(data=df, x=column)

    # Rotate x-axis labels for better readability
    plt.xticks(rotation=45)

    # Set title
    plt.title(f'Frequency Distribution of {column}')

    # Add grid for better visualization
    plt.grid(True)

plt.tight_layout()

plt.show()
```

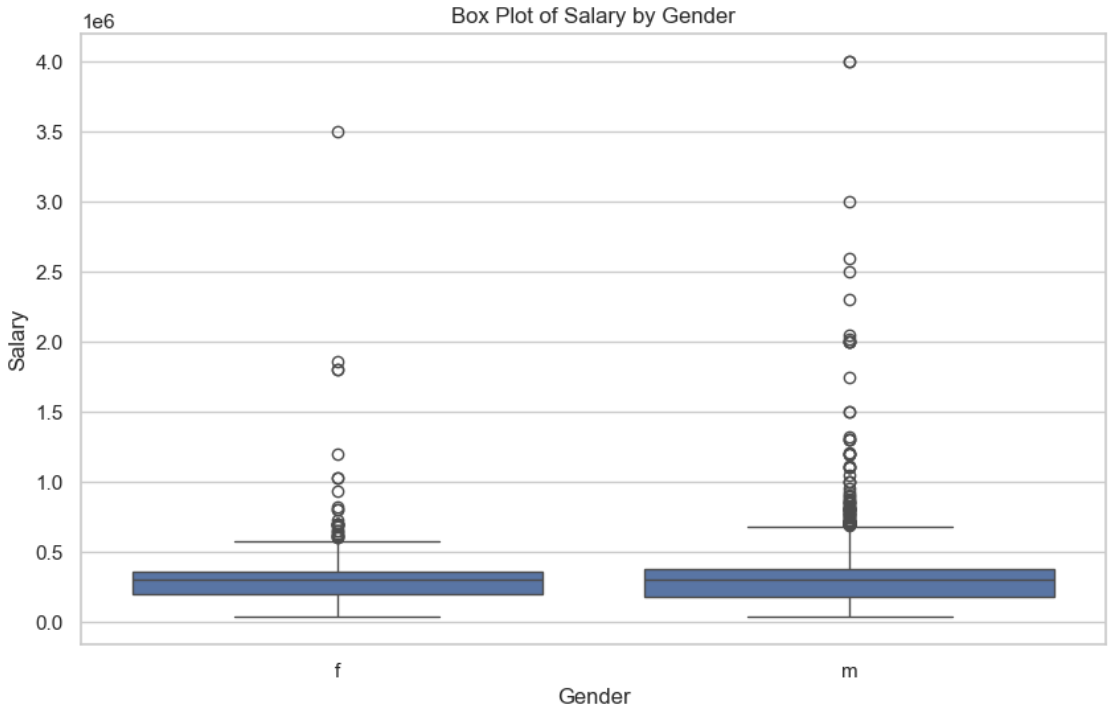


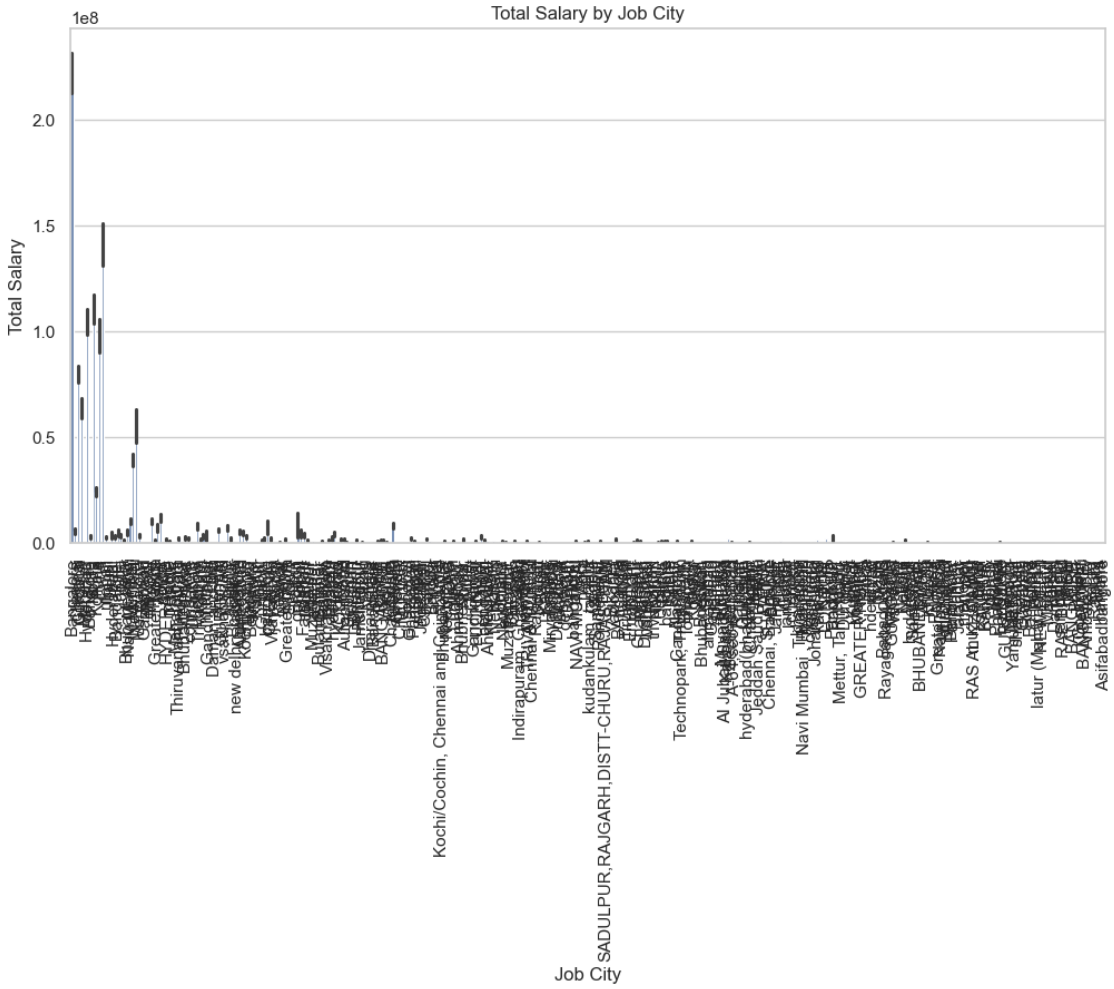
4 Step:- 4 Bivariate Analysis

[26]:
```
import matplotlib.pyplot as plt
```

4.1 2 Identify the patterns between categorical and numerical columns using swarmplot, boxplot, barplot

[27]:
```
df = pd.read_excel('data.xlsx')  # Replace 'data.xlsx' with the actual file path

# Define categorical and numerical columns
categorical_columns = ['Designation', 'JobCity', 'Gender', '10board',␣
 ↪'12board', 'Degree', 'Specialization', 'CollegeState']
numerical_columns = ['Salary', '10percentage', '12percentage', 'collegeGPA',␣
 ↪'Domain', 'ComputerProgramming',
                     'ElectronicsAndSemicon', 'ComputerScience',␣
 ↪'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg',
                     'CivilEngg', 'conscientiousness', 'agreeableness',␣
 ↪'extraversion', 'nueroticism',
                     'openess_to_experience']

# Box plot for categorical and numerical columns
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='Salary', data=df)
plt.xlabel('Gender')
plt.ylabel('Salary')
plt.title('Box Plot of Salary by Gender')
plt.show()

# Bar plot for categorical and numerical columns
plt.figure(figsize=(12, 6))
sns.barplot(x='JobCity', y='Salary', data=df, estimator=sum)
plt.xlabel('Job City')
plt.ylabel('Total Salary')
plt.title('Total Salary by Job City')
plt.xticks(rotation=90)
plt.show()
```


Box Plot of Salary by Gender
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Salary
f
m
Gender


Total Salary by Job City
1e8
2.0
1.5
1.0
0.5
0.0
Total Salary
Job City

5 Conclusion

5.0.1 The analysis of the dataset revealed valuable insights into the distribution and characteristics of various attributes related to education, job details, gender, and specialization. Through univariate analysis, outliers were identified in certain numerical columns, while probability density functions, histograms, and countplots provided a comprehensive understanding of the data's frequency distribution. Bivariate analysis further explored relationships between numerical and categorical variables, uncovering potential correlations and patterns. Notably, stacked bar plots highlighted relationships between categorical variables, such as gender and specialization. Moreover, the analysis rigorously tested the salary claim for fresh graduates with a Computer Science Engineering degree and examined the relationship between gender and specialization.

[ ]: