eda-project-amcat-data-analysis

October 4, 2024

0.1 Badveli Chandrika

0.2 ID : : IN9240298

0.3 Step 1

0.3.1 ANALYSIS OF AMCAT DATA

The dataset originates from the Aspiring Minds Employment Outcome 2015 (AMEO) and focuses on the employment outcomes of engineering graduates. It includes a mix of demographic information, educational details, standardized test scores in cognitive and technical skills, and personality traits, across approximately 4000 data points. Key features include:

Personal and Demographic Information: Includes the candidate's ID, gender, date of birth, job designation, job city, and salary.

Educational Background: Covers high school and college academic performances, the tier of the college, specialization, degree, and graduation year.

Technical and Cognitive Skills: Scores from AMCAT tests in areas such as English, logical reasoning, quantitative ability, computer programming, and various engineering disciplines.

Personality Traits: Scores in conscientiousness, agreeableness, extraversion, neuroticism, and openness to experience.

0.3.2 Objective:

The primary aim is to analyze the relationship between the educational background, skillset, and personality traits of engineering graduates and their employment outcomes, such as job roles and salaries. This includes validating industry claims about salary expectations for specific roles and exploring the influence of gender on specialization preferences.

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from datetime import datetime, timedelta
import seaborn as sns
```

```
C:\ProgramData\Anaconda3\lib\site-packages\scipy\_init_.py:155: UserWarning: A
NumPy version >=1.18.5 and <1.25.0 is required for this version of SciPy
(detected version 1.26.4
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

[2]:
```python
df = pd.read_csv("AMCAT.csv")
```

0.4 Step 2

[3]:
```python
df.head()
```

[3]:
```
  Unnamed: 0      ID     Salary               DOJ               DOL  \
0      train  203097   420000.0  01-06-2012 00:00           present
1      train  579905   500000.0  01-09-2013 00:00           present
2      train  810601   325000.0  01-06-2014 00:00           present
3      train  267447  1100000.0  01-07-2011 00:00           present
4      train  343523   200000.0  01-03-2014 00:00  01-03-2015 00:00

               Designation    JobCity Gender               DOB  10percentage  \
0  senior quality engineer  Bangalore      f  19-02-1990 00:00          84.3
1        assistant manager     Indore      m  04-10-1989 00:00          85.4
2         systems engineer    Chennai      f  03-08-1992 00:00          85.0
3 senior software engineer    Gurgaon      m  05-12-1989 00:00          85.6
4                      get    Manesar      m  27-02-1991 00:00          78.0

  ... ComputerScience  MechanicalEngg ElectricalEngg TelecomEngg  CivilEngg  \
0 ...              -1              -1             -1          -1         -1
1 ...              -1              -1             -1          -1         -1
2 ...              -1              -1             -1          -1         -1
3 ...              -1              -1             -1          -1         -1
4 ...              -1              -1             -1          -1         -1

  conscientiousness agreeableness extraversion  nueroticism  \
0            0.9737        0.8128       0.5269      1.35490
1           -0.7335        0.3789       1.2396     -0.10760
2            0.2718        1.7109       0.1637     -0.86820
3            0.0464        0.3448      -0.3440     -0.40780
4           -0.8810       -0.2793      -1.0697      0.09163

  openess_to_experience
0               -0.4455
1                0.8637
2                0.6721
3               -0.9194
4               -0.1295

[5 rows x 39 columns]
```

[4]:
```python
df.shape
```

[4]: (3998, 39)

[5]:
```
df.columns
```

[5]:
```
Index(['Unnamed: 0', 'ID', 'Salary', 'DOJ', 'DOL', 'Designation', 'JobCity',
       'Gender', 'DOB', '10percentage', '10board', '12graduation',
       '12percentage', '12board', 'CollegeID', 'CollegeTier', 'Degree',
       'Specialization', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier',
       'CollegeState', 'GraduationYear', 'English', 'Logical', 'Quant',
       'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon',
       'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg',
       'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion',
       'nueroticism', 'openess_to_experience'],
      dtype='object')
```

[6]:
```
df.describe()
```

[6]:
```
                 ID        Salary  10percentage  12graduation  12percentage  \
count  3.998000e+03  3.998000e+03   3998.000000   3998.000000   3998.000000
mean   6.637945e+05  3.076998e+05     77.925443   2008.087544     74.466366
std    3.632182e+05  2.127375e+05      9.850162      1.653599     10.999933
min    1.124400e+04  3.500000e+04     43.000000   1995.000000     40.000000
25%    3.342842e+05  1.800000e+05     71.680000   2007.000000     66.000000
50%    6.396000e+05  3.000000e+05     79.150000   2008.000000     74.400000
75%    9.904800e+05  3.700000e+05     85.670000   2009.000000     82.600000
max    1.298275e+06  4.000000e+06     97.760000   2013.000000     98.700000

          CollegeID  CollegeTier   collegeGPA  CollegeCityID  CollegeCityTier  \
count   3998.000000  3998.000000  3998.000000    3998.000000      3998.000000
mean    5156.851426     1.925713    71.486171    5156.851426         0.300400
std     4802.261482     0.262270     8.167338    4802.261482         0.458489
min        2.000000     1.000000     6.450000       2.000000         0.000000
25%      494.000000     2.000000    66.407500     494.000000         0.000000
50%     3879.000000     2.000000    71.720000    3879.000000         0.000000
75%     8818.000000     2.000000    76.327500    8818.000000         1.000000
max    18409.000000     2.000000    99.930000   18409.000000         1.000000

       ...  ComputerScience  MechanicalEngg  ElectricalEngg  TelecomEngg  \
count  ...      3998.000000     3998.000000     3998.000000  3998.000000
mean   ...        90.742371       22.974737       16.478739    31.851176
std    ...       175.273083       98.123311       87.585634   104.852845
min    ...        -1.000000       -1.000000       -1.000000    -1.000000
25%    ...        -1.000000       -1.000000       -1.000000    -1.000000
50%    ...        -1.000000       -1.000000       -1.000000    -1.000000
75%    ...        -1.000000       -1.000000       -1.000000    -1.000000
max    ...       715.000000      623.000000      676.000000   548.000000

         CivilEngg  conscientiousness  agreeableness  extraversion  \
count  3998.000000        3998.000000    3998.000000   3998.000000
```

```
mean      2.683842           -0.037831      0.146496      0.002763
std      36.658505            1.028666      0.941782      0.951471
min      -1.000000           -4.126700     -5.781600     -4.600900
25%      -1.000000           -0.713525     -0.287100     -0.604800
50%      -1.000000            0.046400      0.212400      0.091400
75%      -1.000000            0.702700      0.812800      0.672000
max     516.000000            1.995300      1.904800      2.535400

       nueroticism  openess_to_experience
count  3998.000000            3998.000000
mean     -0.169033              -0.138110
std       1.007580               1.008075
min      -2.643000              -7.375700
25%      -0.868200              -0.669200
50%      -0.234400              -0.094300
75%       0.526200               0.502400
max       3.352500               1.822400

[8 rows x 27 columns]
```

[7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 39 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             3998 non-null   object
 1   ID                     3998 non-null   int64
 2   Salary                 3998 non-null   float64
 3   DOJ                    3998 non-null   object
 4   DOL                    3998 non-null   object
 5   Designation            3998 non-null   object
 6   JobCity                3998 non-null   object
 7   Gender                 3998 non-null   object
 8   DOB                    3998 non-null   object
 9   10percentage           3998 non-null   float64
 10  10board                3998 non-null   object
 11  12graduation           3998 non-null   int64
 12  12percentage           3998 non-null   float64
 13  12board                3998 non-null   object
 14  CollegeID              3998 non-null   int64
 15  CollegeTier            3998 non-null   int64
 16  Degree                 3998 non-null   object
 17  Specialization         3998 non-null   object
 18  collegeGPA             3998 non-null   float64
 19  CollegeCityID          3998 non-null   int64
```

```
 20  CollegeCityTier        3998 non-null   int64
 21  CollegeState           3998 non-null   object
 22  GraduationYear         3998 non-null   int64
 23  English                3998 non-null   int64
 24  Logical                3998 non-null   int64
 25  Quant                  3998 non-null   int64
 26  Domain                 3998 non-null   float64
 27  ComputerProgramming    3998 non-null   int64
 28  ElectronicsAndSemicon  3998 non-null   int64
 29  ComputerScience        3998 non-null   int64
 30  MechanicalEngg         3998 non-null   int64
 31  ElectricalEngg         3998 non-null   int64
 32  TelecomEngg            3998 non-null   int64
 33  CivilEngg              3998 non-null   int64
 34  conscientiousness      3998 non-null   float64
 35  agreeableness          3998 non-null   float64
 36  extraversion           3998 non-null   float64
 37  nueroticism            3998 non-null   float64
 38  openess_to_experience  3998 non-null   float64
dtypes: float64(10), int64(17), object(12)
memory usage: 1.2+ MB
```

[8]:
```python
date_columns = ["DOJ","DOB"]
for col in date_columns:
    df[col] = pd.to_datetime(df[col], errors='ignore', format="%m/%d/%y %H:%M")
```

[9]:
```python
today_date = datetime.today().strftime("%Y-%m-%d")
df["DOL"]=df["DOL"].replace('present',today_date)
df["DOL"] = pd.to_datetime(df["DOL"], dayfirst=True)
```

[10]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 39 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             3998 non-null   object
 1   ID                     3998 non-null   int64
 2   Salary                 3998 non-null   float64
 3   DOJ                    3998 non-null   object
 4   DOL                    3998 non-null   datetime64[ns]
 5   Designation            3998 non-null   object
 6   JobCity                3998 non-null   object
 7   Gender                 3998 non-null   object
 8   DOB                    3998 non-null   object
 9   10percentage           3998 non-null   float64
 10  10board                3998 non-null   object
```

```
 11  12graduation          3998 non-null   int64
 12  12percentage          3998 non-null   float64
 13  12board               3998 non-null   object
 14  CollegeID             3998 non-null   int64
 15  CollegeTier           3998 non-null   int64
 16  Degree                3998 non-null   object
 17  Specialization        3998 non-null   object
 18  collegeGPA            3998 non-null   float64
 19  CollegeCityID         3998 non-null   int64
 20  CollegeCityTier       3998 non-null   int64
 21  CollegeState          3998 non-null   object
 22  GraduationYear        3998 non-null   int64
 23  English               3998 non-null   int64
 24  Logical               3998 non-null   int64
 25  Quant                 3998 non-null   int64
 26  Domain                3998 non-null   float64
 27  ComputerProgramming   3998 non-null   int64
 28  ElectronicsAndSemicon 3998 non-null   int64
 29  ComputerScience       3998 non-null   int64
 30  MechanicalEngg        3998 non-null   int64
 31  ElectricalEngg        3998 non-null   int64
 32  TelecomEngg           3998 non-null   int64
 33  CivilEngg             3998 non-null   int64
 34  conscientiousness     3998 non-null   float64
 35  agreeableness         3998 non-null   float64
 36  extraversion          3998 non-null   float64
 37  nueroticism           3998 non-null   float64
 38  openess_to_experience 3998 non-null   float64
dtypes: datetime64[ns](1), float64(10), int64(17), object(11)
memory usage: 1.2+ MB
```

[11]:
```
print(df.isnull().sum())
```

```
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
```

[41]:
```
desig = df['Designation'].unique()

desig.sort()
```

[42]:
```
desig
```

[42]:
```
array(['.net developer', '.net web developer', 'account executive',
       'account manager', 'admin assistant', 'administrative coordinator',
       'administrative support', 'aircraft technician',
       'android developer', 'application developer',
       'application engineer', 'apprentice', 'ase', 'asp.net developer',
       'assistant administrator', 'assistant electrical engineer',
       'assistant engineer', 'assistant manager', 'assistant professor',
       'assistant programmer', 'assistant software engineer',
       'assistant store manager', 'assistant system engineer',
       'assistant system engineer - trainee',
       'assistant system engineer trainee', 'assistant systems engineer',
       'associate developer', 'associate engineer',
       'associate software developer', 'associate software engg',
       'associate software engineer', 'associate system engineer',
       'associate test engineer', 'automation engineer', 'branch manager',
```

'bss engineer', 'business analyst', 'business analyst consultant', 'business consultant', 'business development executive', 'business development manager', 'business development managerde', 'business intelligence analyst', 'business office manager', 'business system analyst', 'business systems analyst', 'business systems consultant', 'business technology analyst', 'c# developer', 'cad drafter', 'catalog associate', 'civil engineer', 'clerical', 'clerical assistant', 'client services associate', 'cloud engineer', 'computer faculty', 'controls engineer', 'customer service', 'customer service representative', 'customer support engineer', 'data analyst', 'data entry operator', 'data scientist', 'database administrator', 'database developer', 'db2 dba', 'dcs engineer', 'delivery software engineer', 'design engineer', 'designer', 'desktop support analyst', 'desktop support engineer', 'desktop support technician', 'developer', 'digital marketing specialist', 'documentation specialist', 'dotnet developer', 'educator', 'electrical controls engineer', 'electrical design engineer', 'electrical engineer', 'electrical field engineer', 'electrical project engineer', 'electronic field service engineer', 'embedded engineer', 'embedded software engineer', 'engineer', 'engineer trainee', 'engineering manager', 'enterprise solutions developer', 'entry level management trainee', 'etl developer', 'executive assistant', 'executive engg', 'executive hr', 'faculty', 'field business development associate', 'field engineer', 'field service engineer', 'financial analyst', 'firmware engineer', 'front end developer', 'front end web developer', 'full stack developer', 'full-time loss prevention associate', 'game developer', 'general manager', 'get', 'gis/cad engineer', 'graduate apprentice trainee', 'graduate engineer trainee', 'graduate trainee engineer', 'graphic designer', 'hardware engineer', 'help desk analyst', 'help desk technician', 'hr assistant', 'hr generalist', 'hr manager', 'hr recruiter', 'html developer', 'human resource assistant', 'human resources analyst', 'human resources associate', 'human resources intern', 'industrial engineer', 'information security analyst', 'information technology specialist', 'ios developer', 'it analyst', 'it assistant', 'it business analyst', 'it engineer', 'it executive', 'it recruiter', 'it support specialist', 'it technician', 'java developer', 'java software engineer', 'java trainee', 'javascript developer', 'jr. software developer', 'jr. software engineer', 'junior .net developer', 'junior engineer', 'junior engineer product support', 'junior manager', 'junior research fellow', 'junior software developer', 'junior software engineer',

'junior system analyst', 'lead engineer', 'lecturer',
'linux systems administrator', 'logistics executive',
'maintenance engineer', 'management trainee', 'manager',
'manual tester', 'marketing analyst', 'marketing assistant',
'marketing coordinator', 'marketing executive',
'marketing manager', 'mis executive',
'mobile application developer', 'network administrator',
'network engineer', 'network security engineer',
'network support engineer', 'noc engineer', 'office coordinator',
'online marketing manager', 'operation executive',
'operational executive', 'operations', 'operations analyst',
'operations assistant', 'operations executive',
'operations manager', 'oracle dba', 'performance engineer',
'phone banking officer', 'php developer', 'planning engineer',
'portfolio analyst', 'principal software engineer',
'process advisor', 'process associate', 'process control engineer',
'process engineer', 'process executive', 'product design engineer',
'product development engineer', 'product engineer',
'product manager', 'production engineer',
'program analyst trainee', 'program manager', 'programmer',
'programmer analyst', 'programmer analyst trainee',
'project assistant', 'project coordinator', 'project engineer',
'project management officer', 'project manager',
'python developer', 'qa analyst', 'qa engineer', 'quality analyst',
'quality associate', 'quality assurance',
'quality assurance automation engineer',
'quality assurance engineer', 'quality assurance test engineer',
'quality assurance tester', 'quality controller',
'quality engineer', 'r & d', 'r&d engineer',
'recruitment coordinator', 'research analyst',
'research associate', 'research engineer', 'research staff member',
'rf engineer', 'rf/dt engineer', 'risk consultant',
'risk investigator', 'ruby on rails developer', 'sales associate',
'sales coordinator', 'sales development manager', 'sales engineer',
'sales executive', 'sales management trainee', 'sales trainer',
'salesforce developer', 'sap abap consultant', 'sap consultant',
'sap functional consultant', 'senior .net developer',
'senior business analyst', 'senior developer', 'senior engineer',
'senior java developer', 'senior network engineer',
'senior php developer', 'senior programmer',
'senior project engineer', 'senior quality assurance engineer',
'senior quality engineer', 'senior research fellow',
'senior risk consultant', 'senior sales executive',
'senior software developer', 'senior software engineer',
'senior systems engineer', 'senior test engineer',
'senior web developer', 'seo', 'seo analyst', 'seo engineer',
'seo executive', 'service and sales engineer',

```
 'service coordinator', 'service engineer', 'site engineer',
 'site manager', 'software analyst', 'software architect',
 'software designer', 'software developer',
 'software development engineer', 'software devloper',
 'software engg', 'software engineer', 'software engineer analyst',
 'software engineer associate', 'software engineer trainee',
 'software engineere', 'software enginner', 'software executive',
 'software programmer', 'software quality assurance analyst',
 'software quality assurance tester', 'software test engineer',
 'software test engineer (etl)', 'software trainee',
 'software trainee engineer', 'sql dba', 'sql developer',
 'sr. engineer', 'staffing recruiter', 'support engineer',
 'system administrator', 'system engineer',
 'system engineer trainee', 'systems administrator',
 'systems analyst', 'systems engineer',
 'talent acquisition specialist', 'team lead', 'team leader',
 'technical analyst', 'technical assistant', 'technical consultant',
 'technical engineer', 'technical lead',
 'technical operations analyst', 'technical recruiter',
 'technical support engineer', 'technical support executive',
 'technical support specialist', 'technical writer',
 'technology analyst', 'technology lead', 'telecom engineer',
 'teradata dba', 'teradata developer', 'test engineer',
 'test technician', 'testing engineer', 'trainee engineer',
 'trainee software developer', 'trainee software engineer',
 'training specialist', 'ui developer', 'ux designer',
 'visiting faculty', 'web application developer', 'web designer',
 'web designer and seo', 'web developer', 'web intern',
 'website developer/tester', 'windows systems administrator'],
dtype=object)
```

[43]:
```
def feature_cleaning(input_val, input_list):
  if type(input_val) == str:
    for item in [i for i in input_list if len(i.split()) > 1]:
      if all([x in input_val for x in item.split()]):
        return item.title()

    for item in [i for i in input_list if len(i.split()) == 1]:
      if item in input_val:
        return item.title()
    if "engineer" in input_val:
      return "Hardware Engineer"
    try:
      matched_item = get_close_matches(input_val, input_list)[0]
      return matched_item.title()
    except:
      return "Other"
```

```
    else:
        return np.nan
```

[44]:
```
roles_list = ["software engineer", "system engineer", "developer", "analyst",␣
↪"test engineer", "dba",
             "administrator", "customer service", "quality engineer", "quality",␣
↪"automation engineer",
             "network engineer", "support", "it engineer", "manager",␣
↪"management", "programmer",
             "tester", "qa engineer", "design"]
```

[45]:
```
df["Job_Role"] = df["Designation"].apply(lambda x: feature_cleaning(x,␣
↪roles_list))
jr_sorted = df["Job_Role"].unique()
jr_sorted.sort()
jr_sorted
```

[45]:
```
array(['Administrator', 'Analyst', 'Automation Engineer',
       'Customer Service', 'Dba', 'Design', 'Developer',
       'Hardware Engineer', 'It Engineer', 'Management', 'Manager',
       'Network Engineer', 'Other', 'Programmer', 'Qa Engineer',
       'Quality', 'Quality Engineer', 'Software Engineer', 'Support',
       'System Engineer', 'Test Engineer', 'Tester'], dtype=object)
```

[47]:
```
df["Job_Role"] = df["Job_Role"].replace({"It Engineer": "Software Engineer",␣
↪"Network Engineer": "System Engineer", "Dba": "System Engineer",
                                         "Support": "Administrator", "Customer␣
↪Service": "Administrator",
                                         "Tester": "Test Engineer", "Qa Engineer":␣
↪"Test Engineer", "Quality": "Test Engineer",
                                         "Quality Engineer": "Test Engineer",␣
↪"Automation Engineer": "Test Engineer",
                                         "Programmer": "Developer", "Management":␣
↪"Manager", "Design": "Other"})
```

[48]:
```
df["Job_Role"].value_counts(dropna=False)
```

[48]:
```
Software Engineer    710
Developer            599
System Engineer      333
Analyst              302
Other                235
Hardware Engineer    220
Administrator        124
Test Engineer        118
```

```
Manager                68
Name: Job_Role, dtype: int64
```

[33]:
```
df['Specialization'].unique()
```

[33]:
```
array(['computer engineering',
       'electronics and communication engineering',
       'information technology', 'computer science & engineering',
       'electronics and electrical engineering', 'computer application',
       'electronics and computer engineering',
       'applied electronics and instrumentation',
       'instrumentation and control engineering',
       'electrical engineering', 'electronics & instrumentation eng',
       'electronics & telecommunications', 'civil engineering',
       'mechanical engineering', 'metallurgical engineering',
       'electronics and instrumentation engineering',
       'information science engineering', 'chemical engineering',
       'electronics engineering', 'computer science and technology',
       'mechatronics', 'biotechnology', 'instrumentation engineering',
       'information & communication technology', 'computer science',
       'telecommunication engineering'], dtype=object)
```

[34]:
```
specialization_mapping = {'electronics and communication engineering' : 'ECE',
 'computer science & engineering' : 'CSE',
 'information technology' : 'CSE' ,
 'computer engineering' : 'CSE',
 'computer application' : 'CSE',
 'mechanical engineering' : 'MECH',
 'electronics and electrical engineering' : 'ECE',
 'electronics & telecommunications' : 'ECE',
 'electrical engineering' : 'EEE',
 'electronics & instrumentation eng' : 'ECE',
 'civil engineering' : 'CE',
 'electronics and instrumentation engineering' : 'ECE',
 'information science engineering' : 'CSE',
 'instrumentation and control engineering' : 'ECE',
 'electronics engineering' : 'ECE',
 'biotechnology' : 'other',
 'other' : 'other',
 'industrial & production engineering' : 'other',
 'chemical engineering' : 'other',
 'applied electronics and instrumentation' : 'ECE',
 'computer science and technology' : 'CSE',
 'telecommunication engineering' : 'ECE',
 'mechanical and automation' : 'MECH',
 'automobile/automotive engineering' : 'MECH',
 'instrumentation engineering' : 'ECE',
```

```
'mechatronics' : 'MECH',
'electronics and computer engineering' : 'CSE',
'aeronautical engineering' : 'MECH',
'computer science' : 'CSE',
'metallurgical engineering' : 'other',
'biomedical engineering' : 'other',
'industrial engineering' : 'other',
'information & communication technology' : 'ECE',
'electrical and power engineering' : 'EEE',
'industrial & management engineering' : 'other',
'computer networking' : 'CSE',
'embedded systems technology' : 'ECE',
'power systems and automation' : 'EEE',
'computer and communication engineering' : 'CSE',
'information science' : 'CSE',
'internal combustion engine' : 'MECH',
'ceramic engineering' : 'other',
'mechanical & production engineering' : 'MECH',
'control and instrumentation engineering' : 'ECE',
'polymer technology' : 'other',
'electronics' : 'ECE'}

for old, new in specialization_mapping.items():
    df['Specialization'] = df['Specialization'].replace(old, new)
```

[35]:
```
df['Specialization'].unique()
```

[35]: array(['CSE', 'ECE', 'EEE', 'CE', 'MECH', 'other'], dtype=object)

0.5 Step 3 - Univariate Analysis

0.6 Non Visual Analysis

[12]:
```
discrete_df = df.select_dtypes(include=['object'])

numerical_df = df.select_dtypes(include=['int64', 'float64'])
```

[13]:
```
def discrete_univariate_analysis(discrete_data):
    for col_name in discrete_data:
        print("*"*10, col_name, "*"*10)
        print(discrete_data[col_name].agg(['count', 'nunique', 'unique']))
        print('Value Counts: \n', discrete_data[col_name].value_counts())
        print()
```

[14]:
```
discrete_univariate_analysis(discrete_df)
```

```
********** Unnamed: 0 **********
```

```
count         3998
nunique          1
unique     [train]
Name: Unnamed: 0, dtype: object
Value Counts:
 train    3998
Name: Unnamed: 0, dtype: int64

********** DOJ **********
count                                                   3998
nunique                                                   81
unique     [01-06-2012 00:00, 01-09-2013 00:00, 01-06-201...
Name: DOJ, dtype: object
Value Counts:
 01-07-2014 00:00    199
01-06-2014 00:00    180
01-08-2014 00:00    178
01-09-2014 00:00    142
01-01-2014 00:00    142
                    ...
01-11-2015 00:00      1
01-11-2009 00:00      1
01-08-2004 00:00      1
01-09-2009 00:00      1
01-02-2007 00:00      1
Name: DOJ, Length: 81, dtype: int64

********** Designation **********
count                                                   3998
nunique                                                  419
unique     [senior quality engineer, assistant manager, s...
Name: Designation, dtype: object
Value Counts:
 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: Designation, Length: 419, dtype: int64

********** JobCity **********
count                                                   3998
```

```
nunique                                                339
unique     [Bangalore, Indore, Chennai, Gurgaon, Manesar,...
Name: JobCity, dtype: object
Value Counts:
 Bangalore          627
-1                  461
Noida               368
Hyderabad           335
Pune                290
                   ...
Tirunelvelli          1
Ernakulam             1
Nanded                1
Dharmapuri            1
Asifabadbanglore      1
Name: JobCity, Length: 339, dtype: int64

********** Gender **********
count       3998
nunique        2
unique    [f, m]
Name: Gender, dtype: object
Value Counts:
 m    3041
f     957
Name: Gender, dtype: int64

********** DOB **********
count                                                   3998
nunique                                                 1872
unique     [19-02-1990 00:00, 04-10-1989 00:00, 03-08-199...
Name: DOB, dtype: object
Value Counts:
 01-01-1991 00:00    11
15-07-1991 00:00    10
05-07-1991 00:00     8
13-12-1991 00:00     8
03-06-1991 00:00     8
                    ..
30-12-1992 00:00     1
20-10-1986 00:00     1
17-11-1989 00:00     1
30-09-1992 00:00     1
15-04-1987 00:00     1
Name: DOB, Length: 1872, dtype: int64

********** 10board **********
count                                                   3998
```

```
nunique                                              275
unique      [board ofsecondary education,ap, cbse, state b...
Name: 10board, dtype: object
Value Counts:
 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: 10board, Length: 275, dtype: int64

********** 12board **********
count                                                3998
nunique                                               340
unique      [board of intermediate education,ap, cbse, sta...
Name: 12board, dtype: object
Value Counts:
 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: 12board, Length: 340, dtype: int64

********** Degree **********
count                                               3998
nunique                                                4
unique      [B.Tech/B.E., MCA, M.Tech./M.E., M.Sc. (Tech.)]
Name: Degree, dtype: object
Value Counts:
 B.Tech/B.E.       3700
MCA                243
M.Tech./M.E.        53
M.Sc. (Tech.)        2
Name: Degree, dtype: int64
```

```
********** Specialization **********
count                                                     3998
nunique                                                     46
unique     [computer engineering, electronics and communi...
Name: Specialization, dtype: object
Value Counts:
 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: Specialization, dtype: int64

********** CollegeState **********
count                                                   3998
nunique                                                   26
unique     [Andhra Pradesh, Madhya Pradesh, Uttar Pradesh...
Name: CollegeState, dtype: object
Value Counts:
 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: CollegeState, dtype: int64
```

[15]:
```python
def numerical_univariate_analysis(numerical_data):
    for col_name in numerical_data:
        print("*"*10, col_name, "*"*10)
        print(numerical_data[col_name].agg(['min', 'max', 'mean', 'median',␣
↪'std']))
        print()
```

[16]:

```
numerical_univariate_analysis(numerical_df)
```

```
********** ID **********
min       1.124400e+04
max       1.298275e+06
mean      6.637945e+05
median    6.396000e+05
std       3.632182e+05
Name: ID, dtype: float64

********** Salary **********
min       3.500000e+04
max       4.000000e+06
mean      3.076998e+05
median    3.000000e+05
std       2.127375e+05
Name: Salary, dtype: float64

********** 10percentage **********
min       43.000000
max       97.760000
mean      77.925443
median    79.150000
std        9.850162
Name: 10percentage, dtype: float64

********** 12graduation **********
min       1995.000000
max       2013.000000
mean      2008.087544
median    2008.000000
std          1.653599
Name: 12graduation, dtype: float64

********** 12percentage **********
min       40.000000
max       98.700000
mean      74.466366
median    74.400000
std       10.999933
Name: 12percentage, dtype: float64

********** CollegeID **********
min           2.000000
max       18409.000000
mean       5156.851426
median     3879.000000
```

```
std        4802.261482
Name: CollegeID, dtype: float64

********** CollegeTier **********
min       1.000000
max       2.000000
mean      1.925713
median    2.000000
std       0.262270
Name: CollegeTier, dtype: float64

********** collegeGPA **********
min        6.450000
max       99.930000
mean      71.486171
median    71.720000
std        8.167338
Name: collegeGPA, dtype: float64

********** CollegeCityID **********
min           2.000000
max       18409.000000
mean       5156.851426
median     3879.000000
std        4802.261482
Name: CollegeCityID, dtype: float64

********** CollegeCityTier **********
min       0.000000
max       1.000000
mean      0.300400
median    0.000000
std       0.458489
Name: CollegeCityTier, dtype: float64

********** GraduationYear **********
min          0.000000
max       2017.000000
mean      2012.105803
median    2013.000000
std         31.857271
Name: GraduationYear, dtype: float64

********** English **********
min       180.000000
max       875.000000
mean      501.649075
median    500.000000
```

```
std        104.940021
Name: English, dtype: float64

********** Logical **********
min        195.000000
max        795.000000
mean       501.598799
median     505.000000
std         86.783297
Name: Logical, dtype: float64

**********  Quant  **********
min        120.000000
max        900.000000
mean       513.378189
median     515.000000
std        122.302332
Name: Quant, dtype: float64

**********  Domain  **********
min       -1.000000
max        0.999910
mean       0.510490
median     0.622643
std        0.468671
Name: Domain, dtype: float64

**********  ComputerProgramming  **********
min         -1.000000
max        840.000000
mean       353.102801
median     415.000000
std        205.355519
Name: ComputerProgramming, dtype: float64

**********  ElectronicsAndSemicon  **********
min         -1.000000
max        612.000000
mean        95.328414
median      -1.000000
std        158.241218
Name: ElectronicsAndSemicon, dtype: float64

**********  ComputerScience  **********
min         -1.000000
max        715.000000
mean        90.742371
median      -1.000000
```

```
std        175.273083
Name: ComputerScience, dtype: float64

********** MechanicalEngg **********
min         -1.000000
max        623.000000
mean        22.974737
median      -1.000000
std         98.123311
Name: MechanicalEngg, dtype: float64

********** ElectricalEngg **********
min         -1.000000
max        676.000000
mean        16.478739
median      -1.000000
std         87.585634
Name: ElectricalEngg, dtype: float64

********** TelecomEngg **********
min         -1.000000
max        548.000000
mean        31.851176
median      -1.000000
std        104.852845
Name: TelecomEngg, dtype: float64

********** CivilEngg **********
min         -1.000000
max        516.000000
mean         2.683842
median      -1.000000
std         36.658505
Name: CivilEngg, dtype: float64

********** conscientiousness **********
min       -4.126700
max        1.995300
mean      -0.037831
median     0.046400
std        1.028666
Name: conscientiousness, dtype: float64

********** agreeableness **********
min       -5.781600
max        1.904800
mean       0.146496
median     0.212400
```

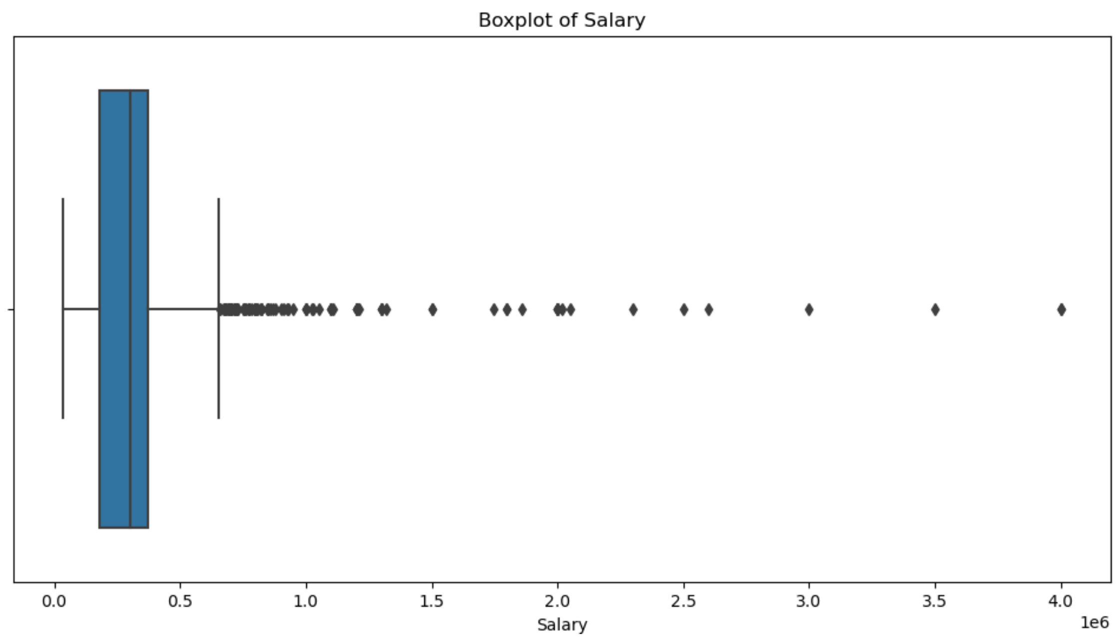
```
std         0.941782
Name: agreeableness, dtype: float64

********** extraversion **********
min        -4.600900
max         2.535400
mean        0.002763
median      0.091400
std         0.951471
Name: extraversion, dtype: float64

********** nueroticism **********
min        -2.643000
max         3.352500
mean       -0.169033
median     -0.234400
std         1.007580
Name: nueroticism, dtype: float64

********** openess_to_experience **********
min        -7.375700
max         1.822400
mean       -0.138110
median     -0.094300
std         1.008075
Name: openess_to_experience, dtype: float64
```

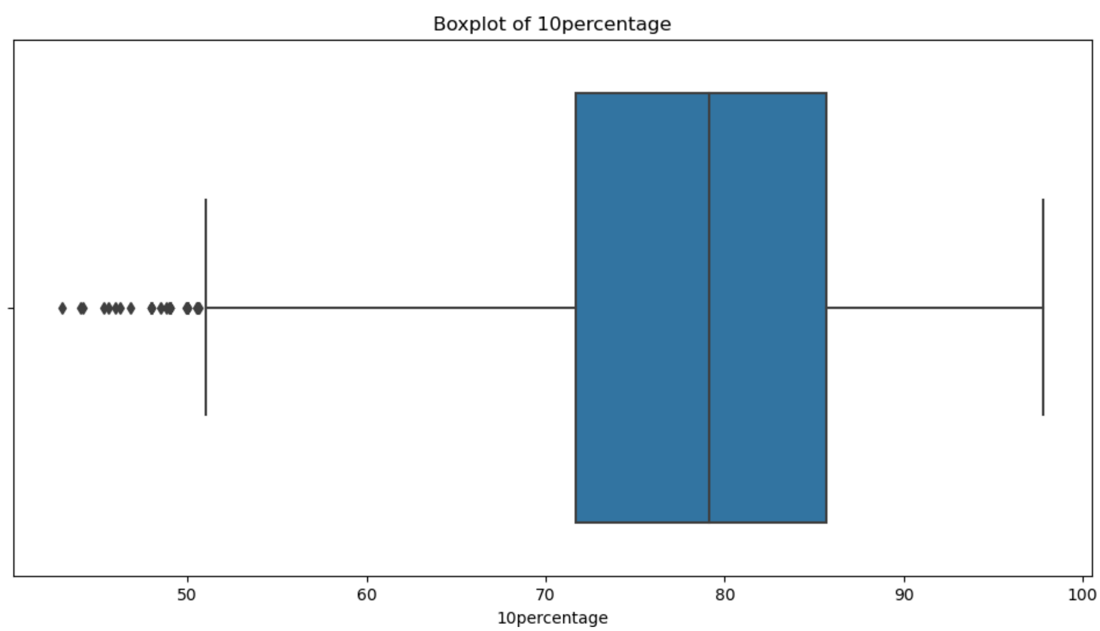
0.7 Univariate - Visual Analysis

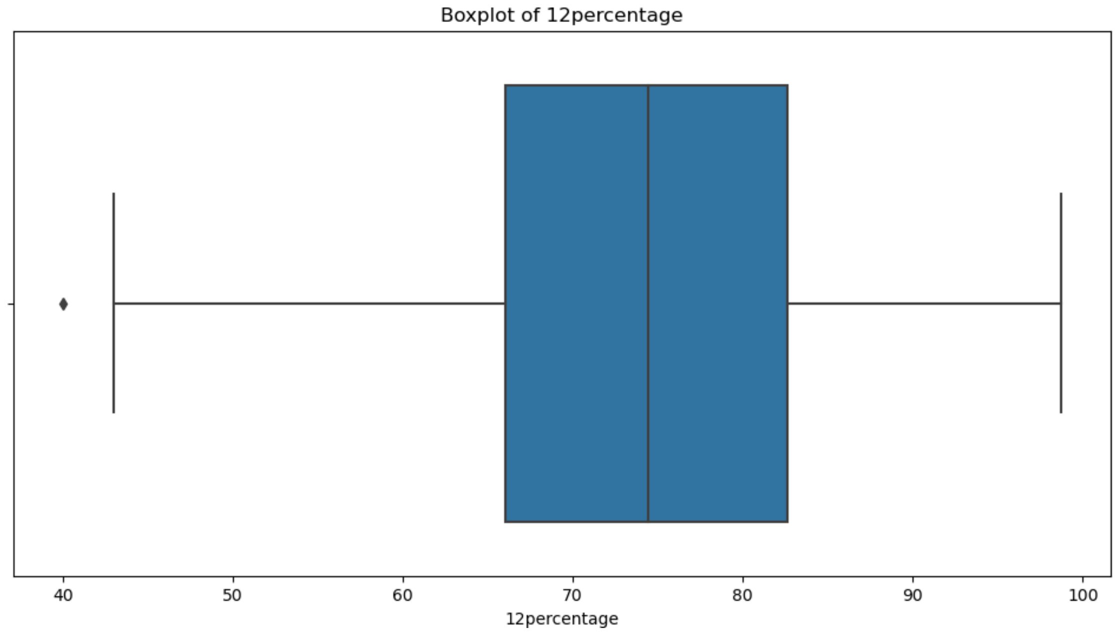
0.7.1 Outlier Detection

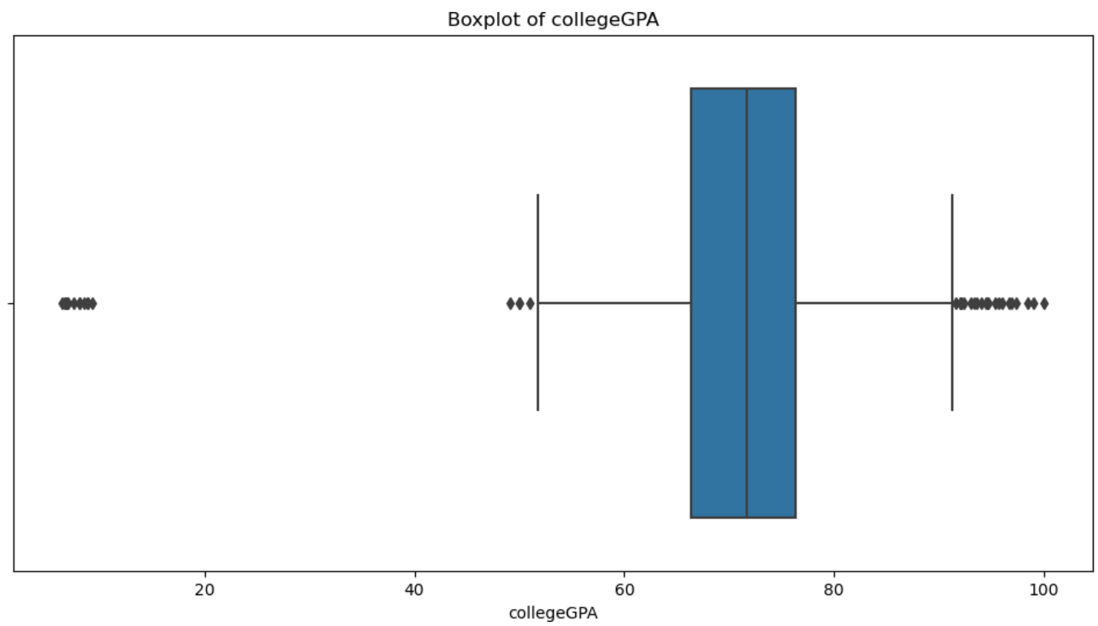
[17]:
```
# Univariate Analysis - Numerical Variables
numerical_cols = ['Salary', '10percentage', '12percentage', 'collegeGPA',␣
 ↪'English', 'Logical', 'Quant', 'Domain',
                  'ComputerProgramming', 'ElectronicsAndSemicon',␣
 ↪'ComputerScience', 'MechanicalEngg', 'ElectricalEngg',
                  'TelecomEngg', 'CivilEngg', 'conscientiousness',␣
 ↪'agreeableness', 'extraversion', 'nueroticism',
                  'openess_to_experience']
```

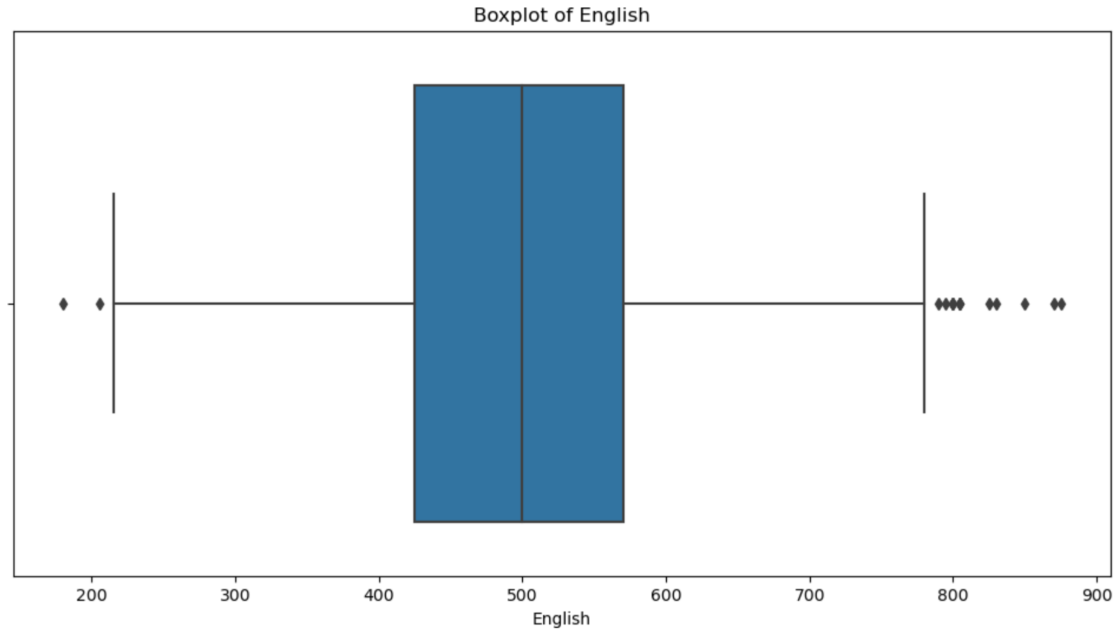
[18]:
```
# Plotting boxplots to detect outliers

for column in numerical_cols:
    plt.figure(figsize=(12, 6))
    sns.boxplot(x = df[column])
    plt.title(f'Boxplot of {column}')
    plt.show()
```

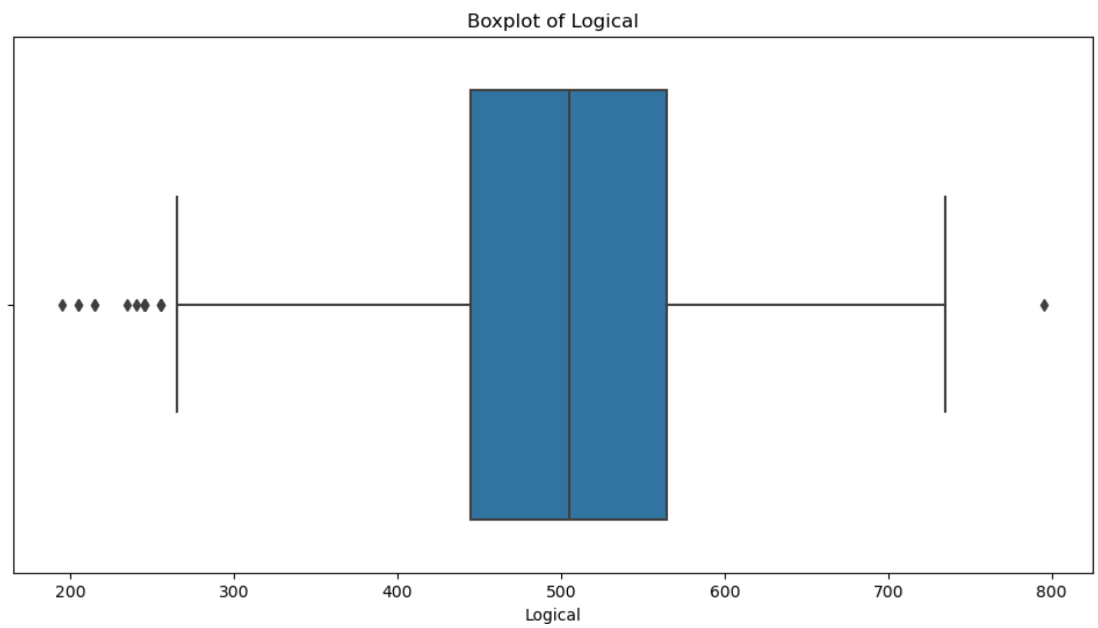

Boxplot of Salary
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
Salary

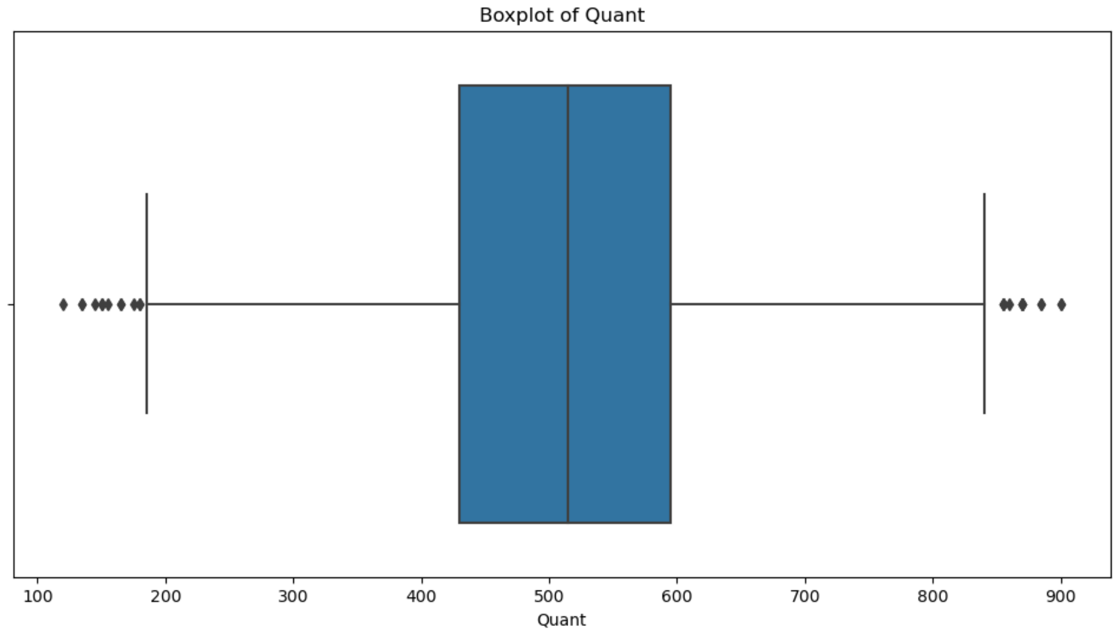

Boxplot of 10percentage
50
60
70
80
90
100
10percentage

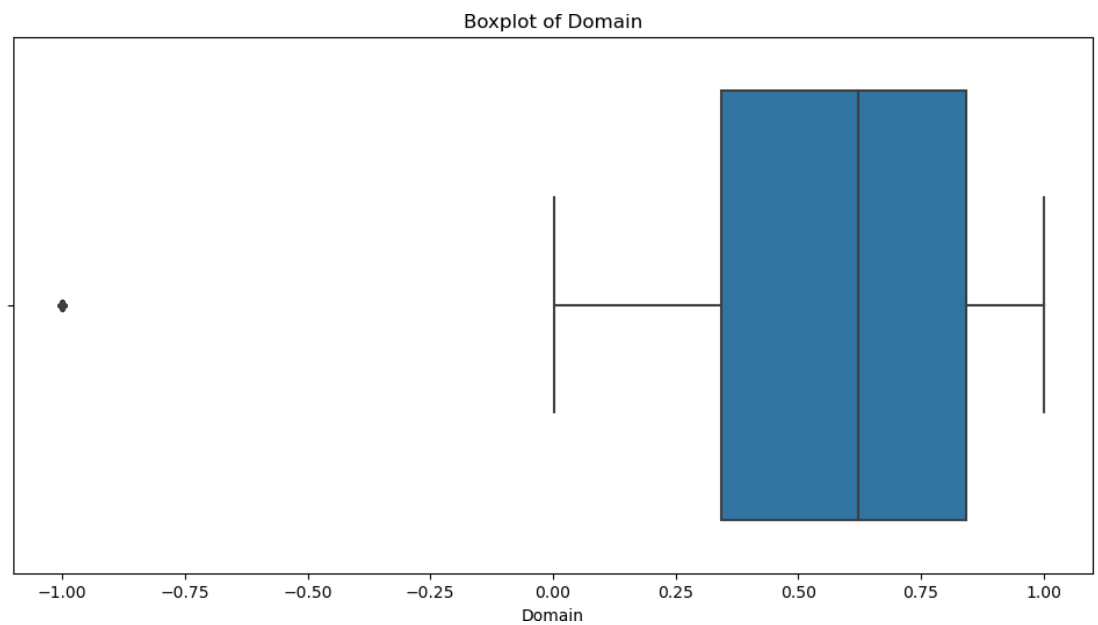

Boxplot of 12percentage
40
50
60
70
80
90
100
12percentage

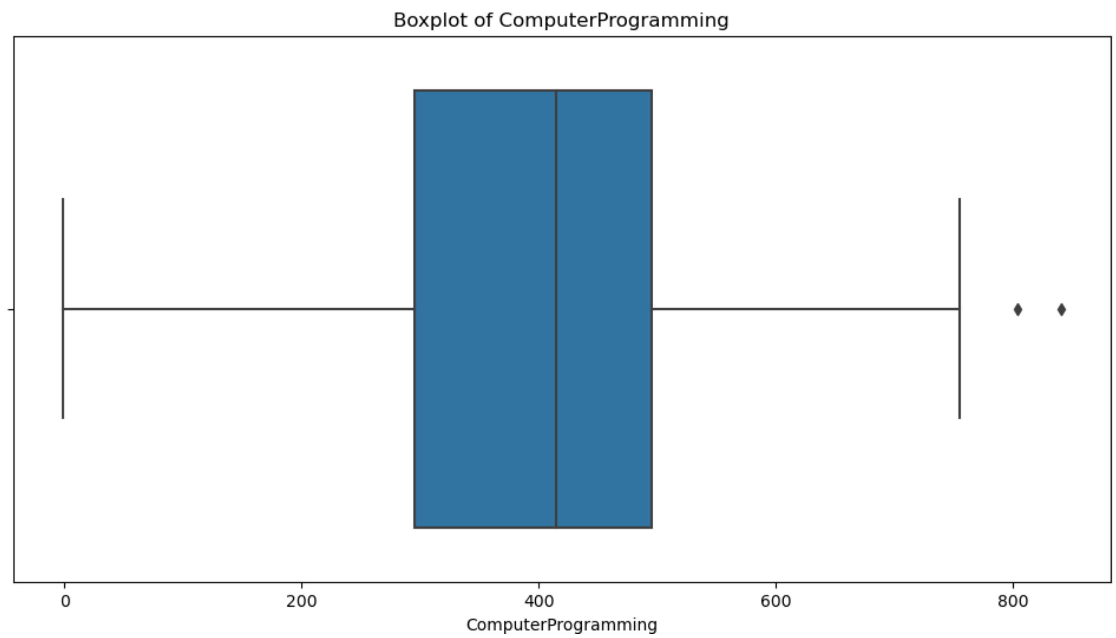

Boxplot of collegeGPA
20
40
60
80
100
collegeGPA

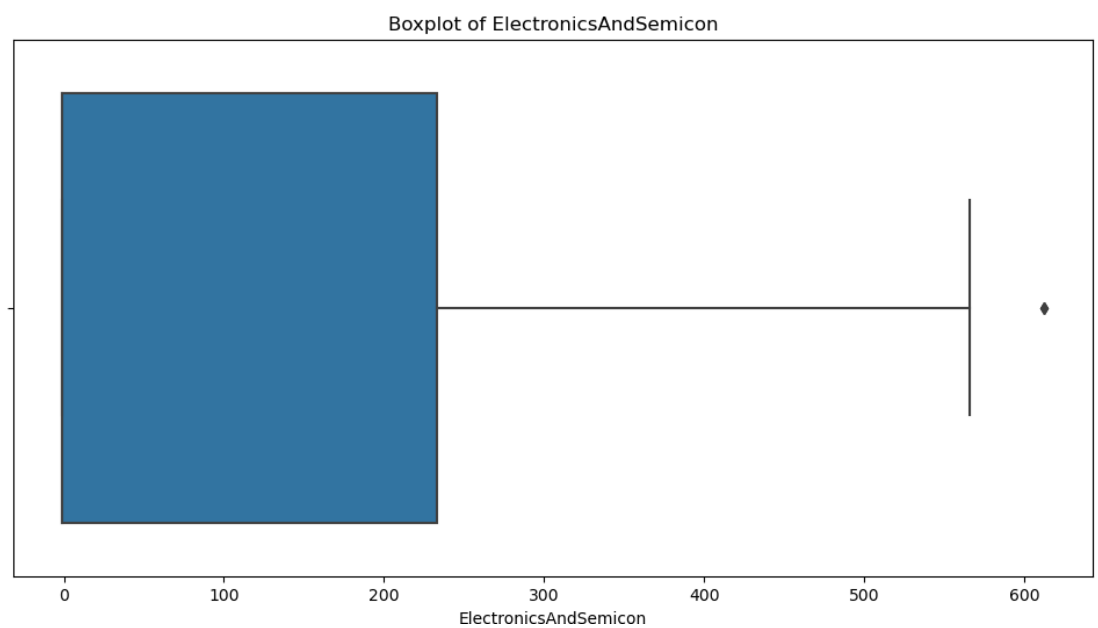

Boxplot of English
200
300
400
500
600
700
800
900
English

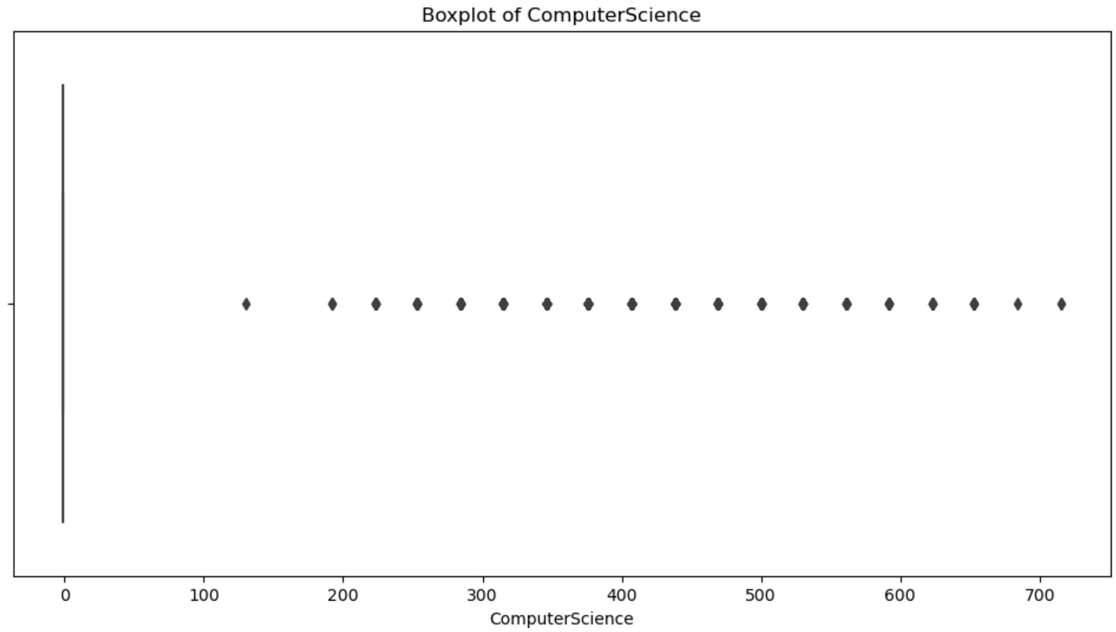

Boxplot of Logical
200
300
400
500
600
700
800
Logical


Boxplot of Quant
100
200
300
400
500
600
700
800
900
Quant


Boxplot of Domain
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Domain

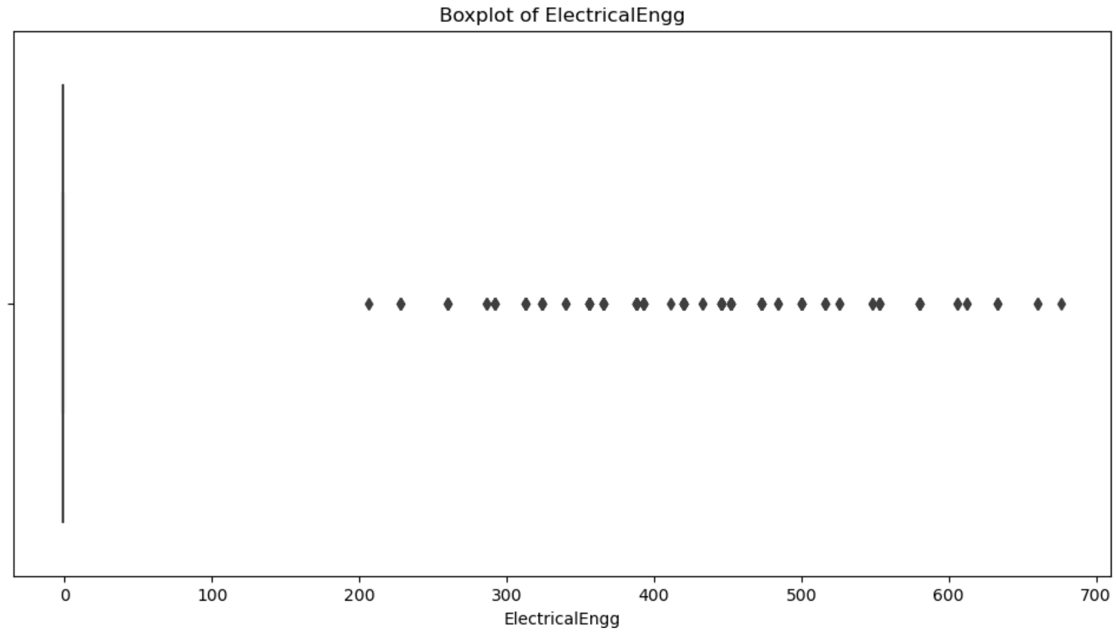

Boxplot of ComputerProgramming
0
200
400
600
800
ComputerProgramming

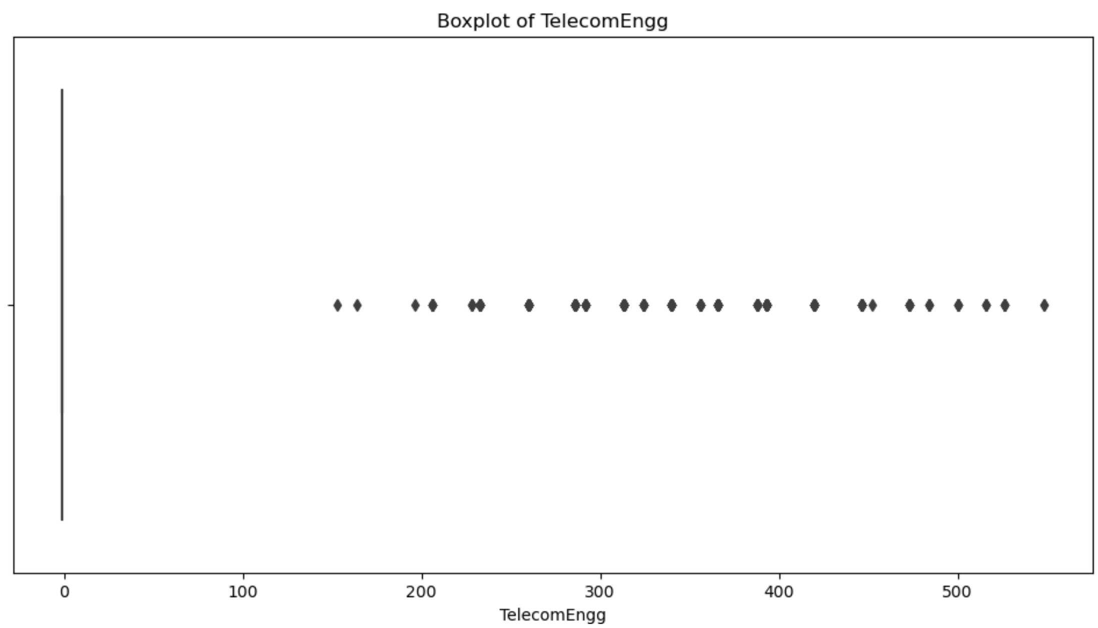

Boxplot of ElectronicsAndSemicon
0
100
200
300
400
500
600
ElectronicsAndSemicon

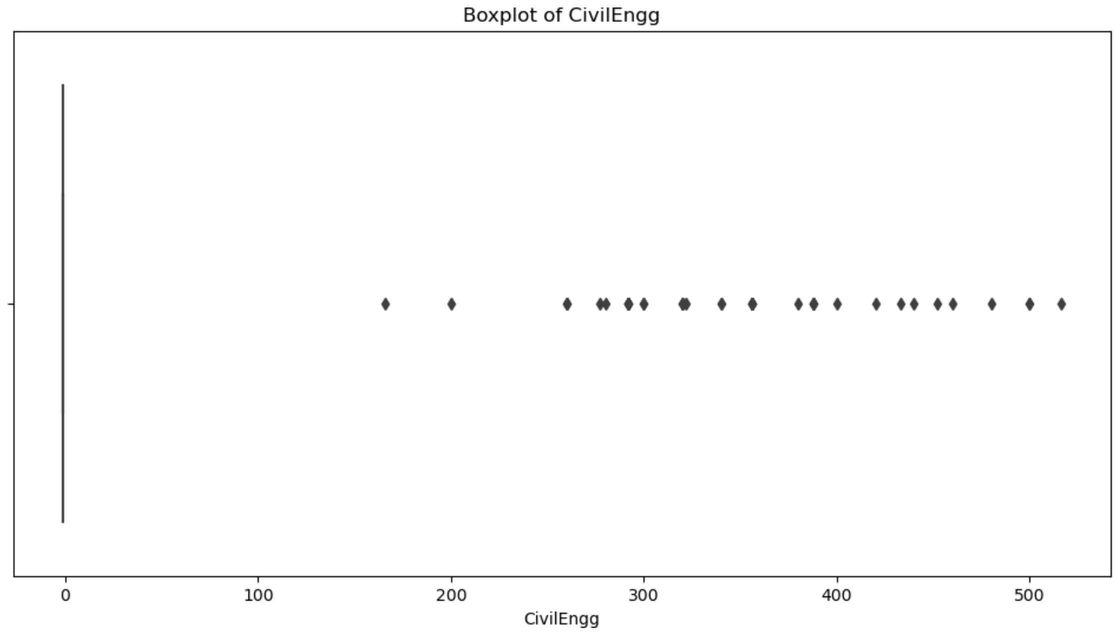


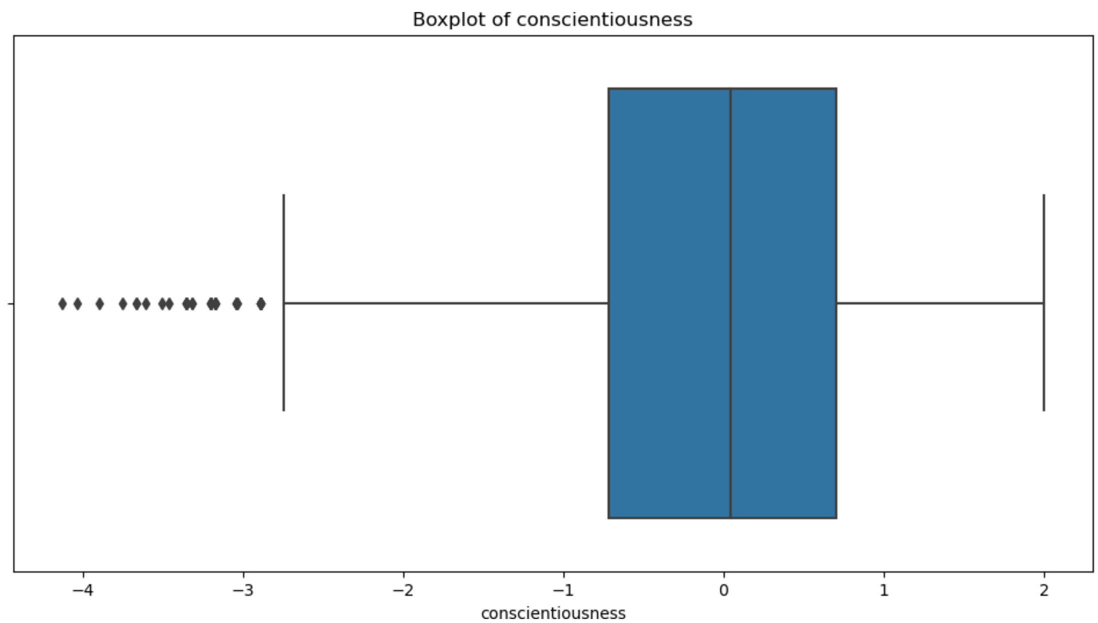
Boxplot of ComputerScience
0
100
200
300
400
500
600
700
ComputerScience

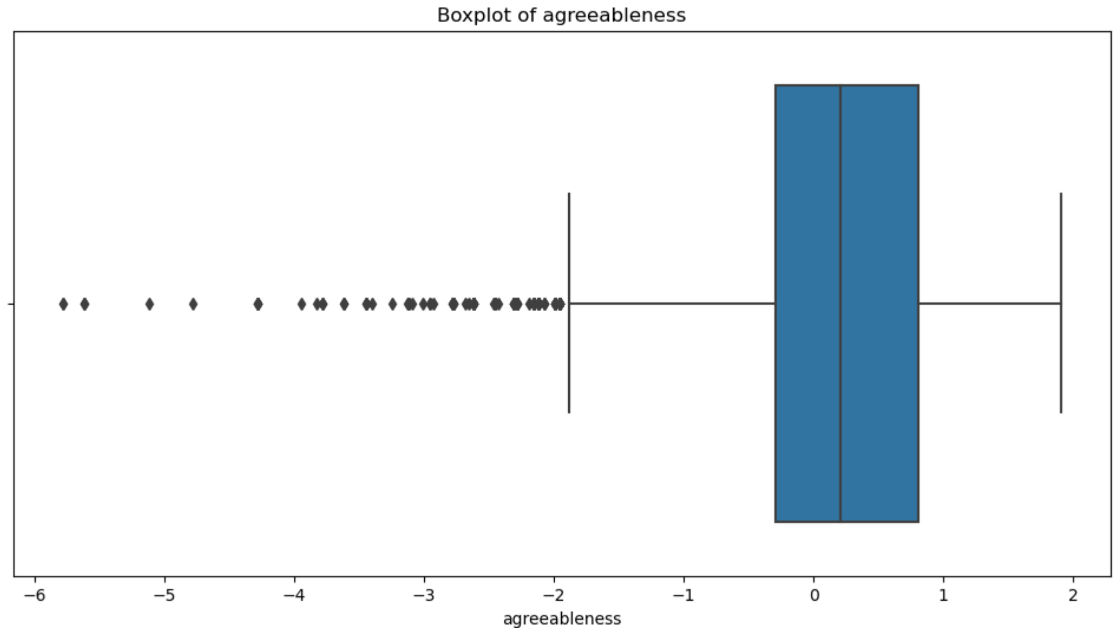
Boxplot of MechanicalEngg
0
100
200
300
400
500
600
MechanicalEngg

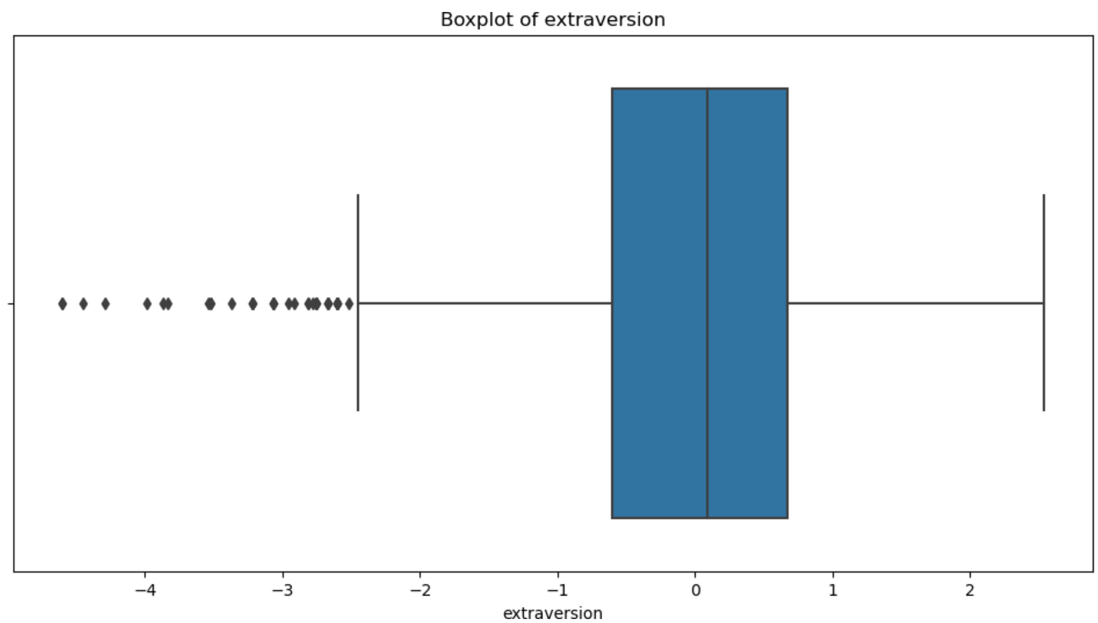

Boxplot of ElectricalEngg
0
100
200
300
400
500
600
700
ElectricalEngg

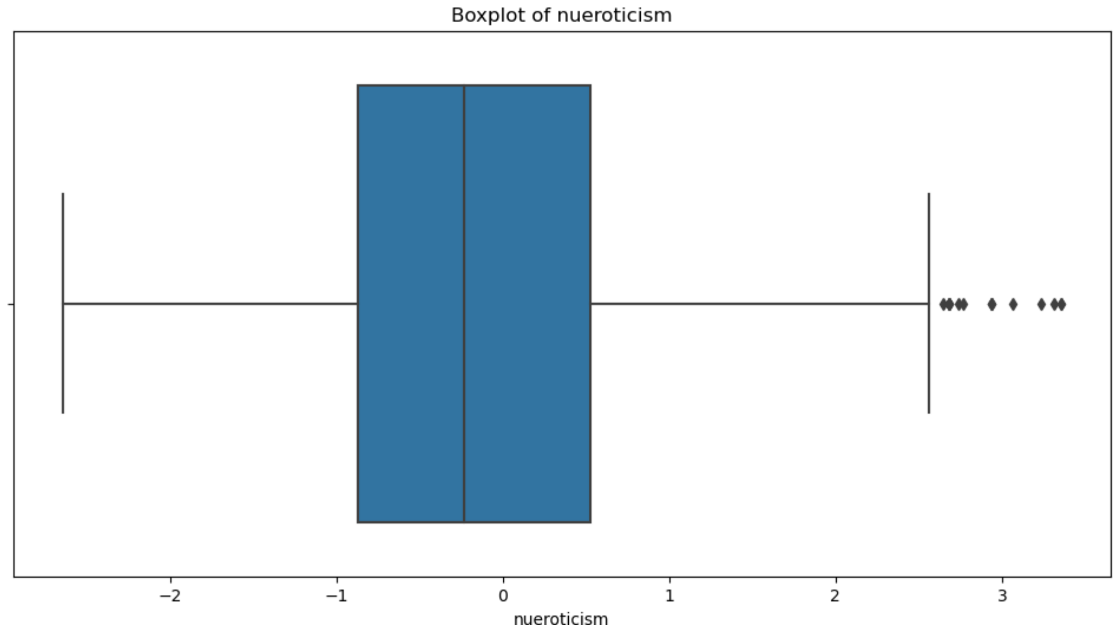

Boxplot of TelecomEngg
0
100
200
300
400
500
TelecomEngg

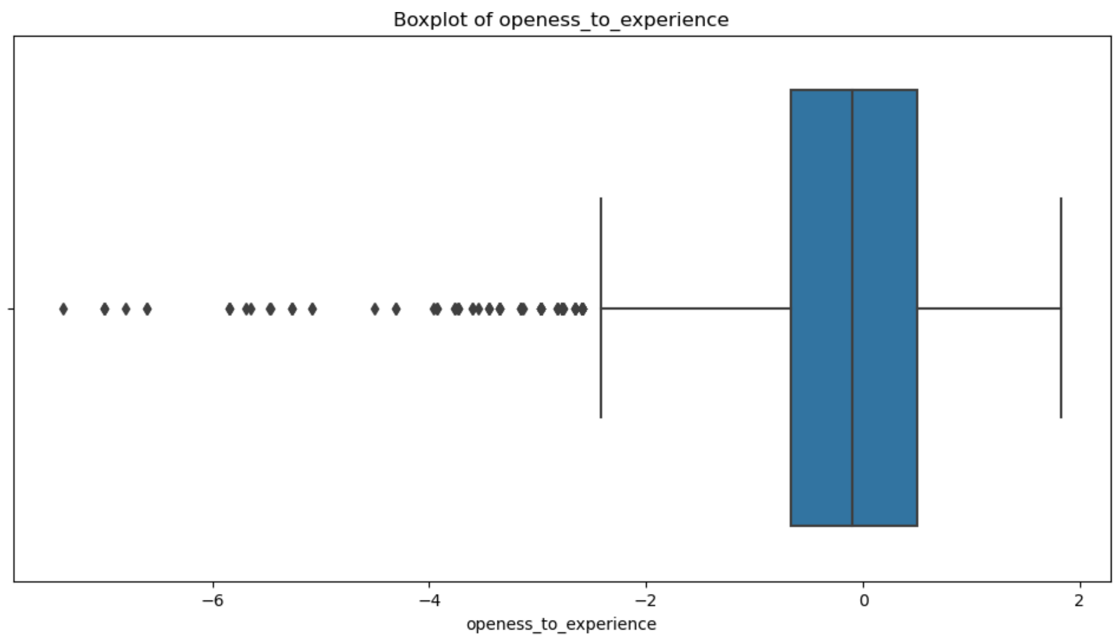

Boxplot of CivilEngg
0
100
200
300
400
500
CivilEngg


Boxplot of conscientiousness
−4
−3
−2
−1
0
1
2
conscientiousness


Boxplot of agreeableness
−6
−5
−4
−3
−2
−1
0
1
2
agreeableness


Boxplot of extraversion
−4
−3
−2
−1
0
1
2
extraversion





[19]:
```
# Outlier Detection
for col in numerical_cols:
    q1 = df[col].quantile(0.25)
    q3 = df[col].quantile(0.75)
    iqr = q3 - q1
```

```
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    outliers = df[(df[col] < lower_bound) | (df[col] > upper_bound)]
    print(f"Outliers in {col}: {len(outliers)}")
```

```
Outliers in Salary: 109
Outliers in 10percentage: 30
Outliers in 12percentage: 1
Outliers in collegeGPA: 38
Outliers in English: 15
Outliers in Logical: 18
Outliers in Quant: 25
Outliers in Domain: 246
Outliers in ComputerProgramming: 2
Outliers in ElectronicsAndSemicon: 2
Outliers in ComputerScience: 902
Outliers in MechanicalEngg: 235
Outliers in ElectricalEngg: 161
Outliers in TelecomEngg: 374
Outliers in CivilEngg: 42
Outliers in conscientiousness: 39
Outliers in agreeableness: 123
Outliers in extraversion: 40
Outliers in nueroticism: 15
Outliers in openess_to_experience: 95
```

0.7.2 Outlier Treatment

Filtering the data so that there would be consistency in the data

[20]:
```
df=df.loc[(df["Domain"]>-1)]
df.shape
```

[20]: (3752, 39)

[21]:
```
df=df.loc[(df["MechanicalEngg"]< 200)]
df.shape
```

[21]: (3521, 39)

[22]:
```
df=df.loc[(df["ElectricalEngg"]< 200)]
df.shape
```

[22]: (3363, 39)

[23]:
```
df=df.loc[(df["TelecomEngg"]< 100)]
df.shape
```

[23]: (2995, 39)

[24]:
```
df=df.loc[(df["agreeableness"]> -1.5)]
df.shape
```

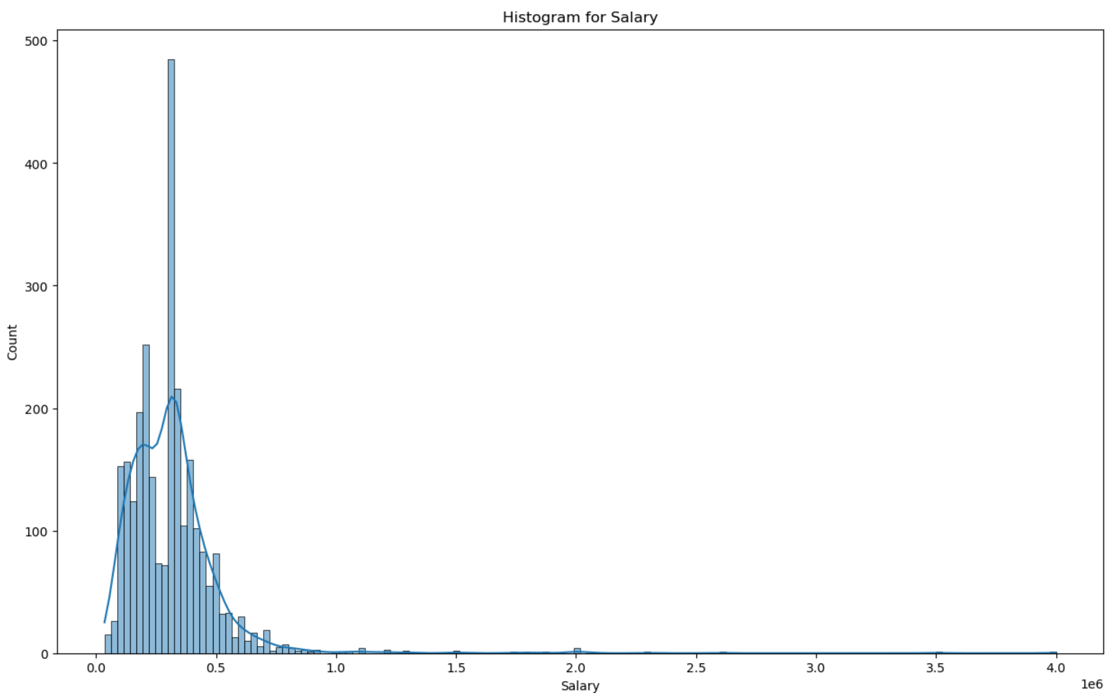
[24]: (2853, 39)

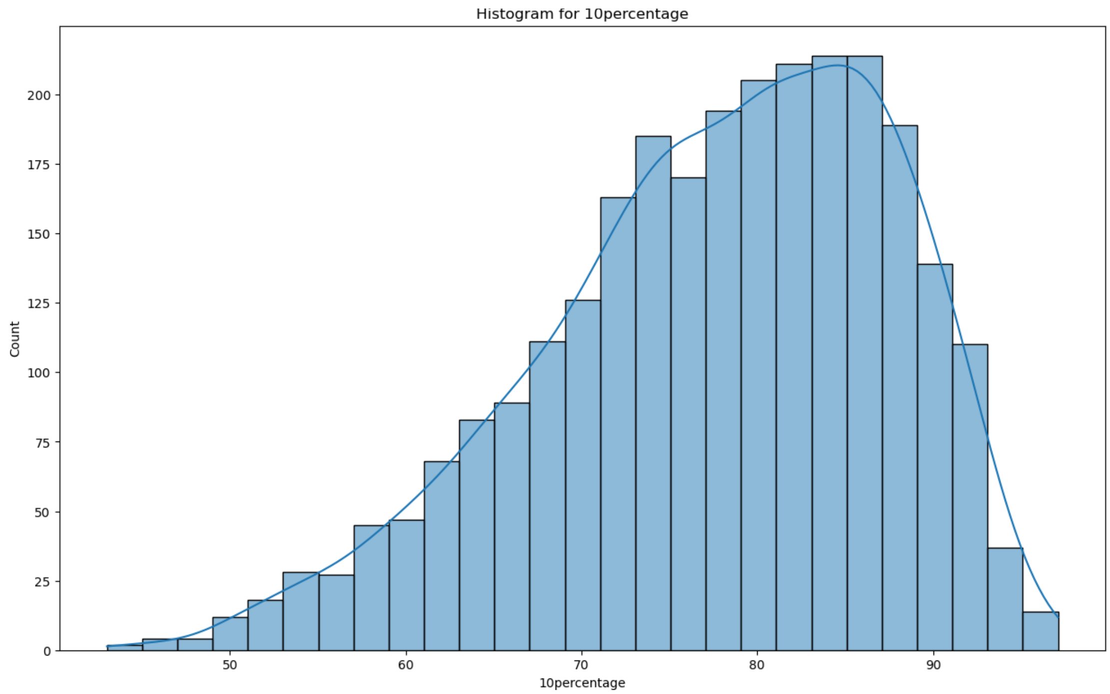
[25]:
```
df=df.loc[(df["openess_to_experience"]> -1.5)]
df.shape
```

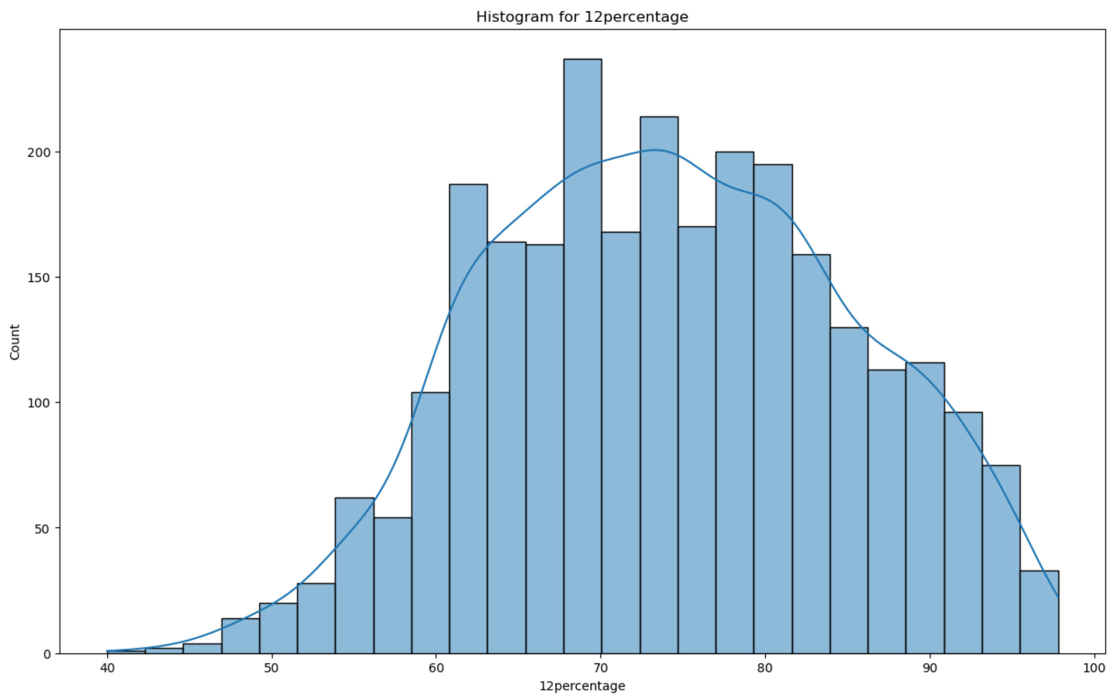
[25]: (2709, 39)

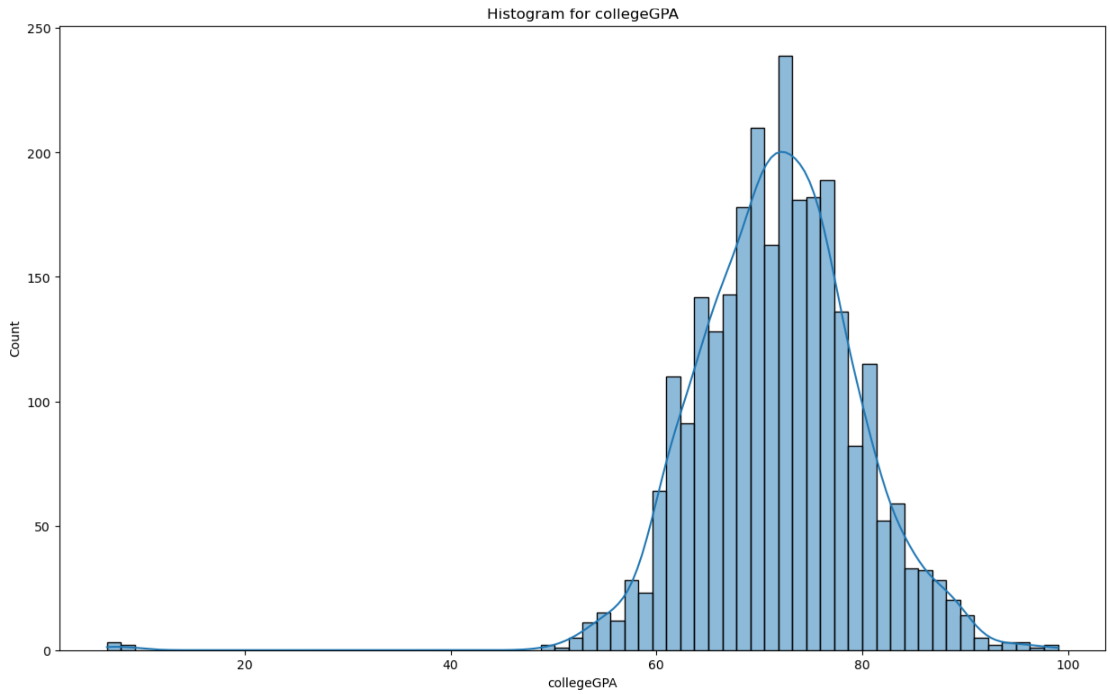
0.7.3 Frequency Distribution

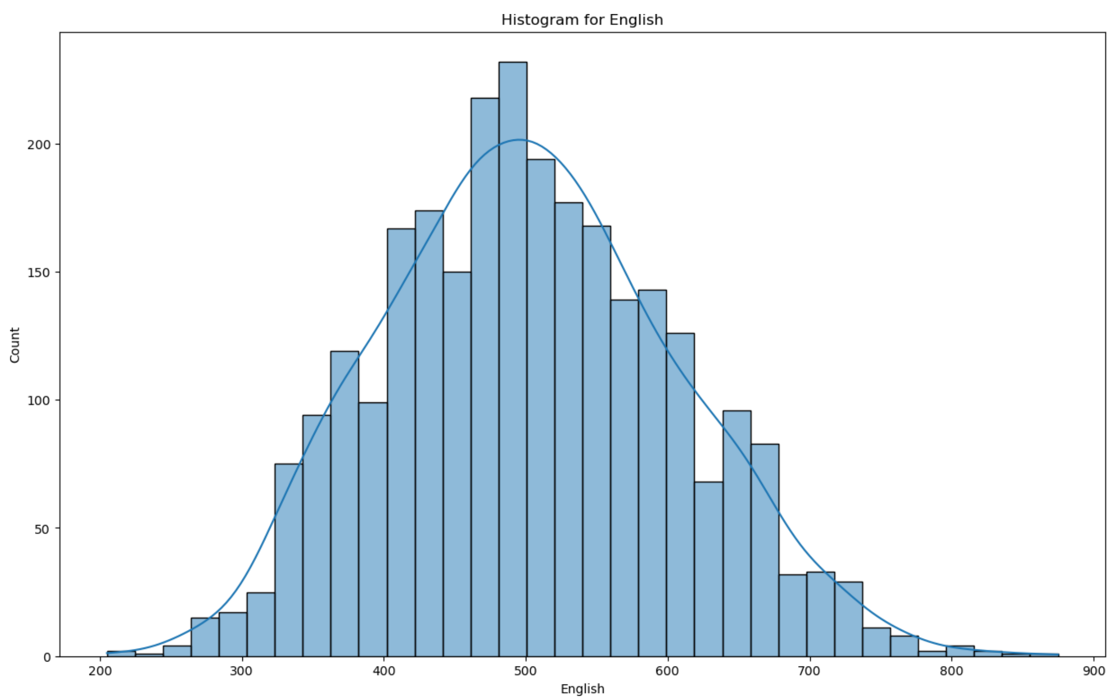
[26]:
```
for column in numerical_cols:

    plt.figure(figsize=(15,9))
    sns.histplot(df[column], kde=True)
    plt.title(f"Histogram for {column}")
    plt.show()
```

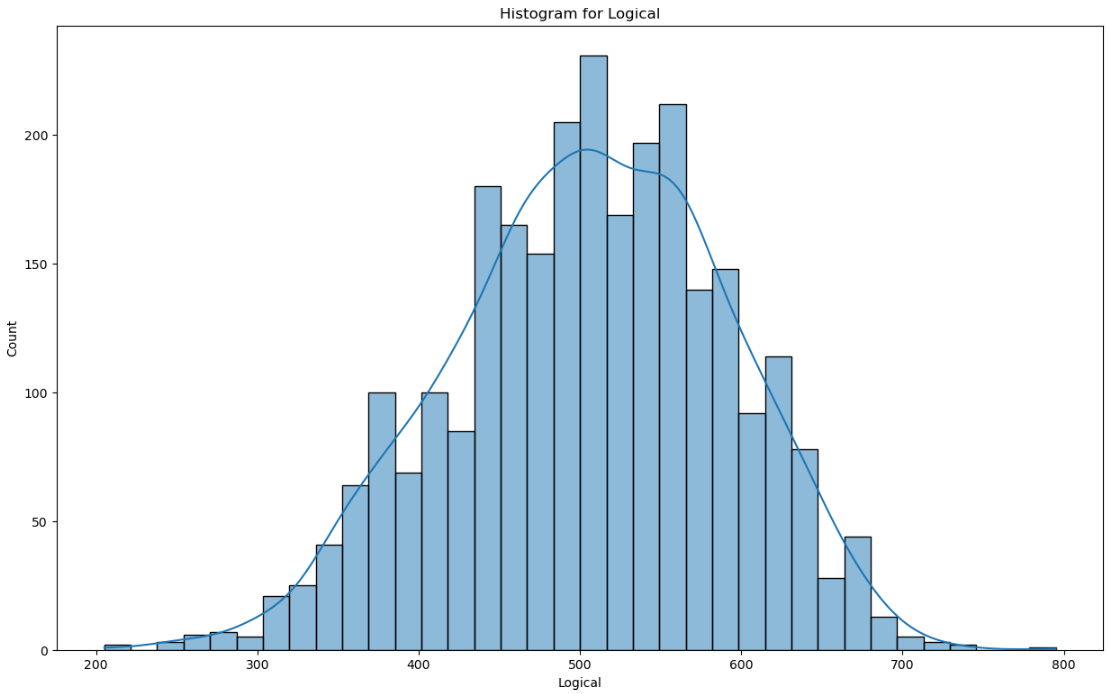


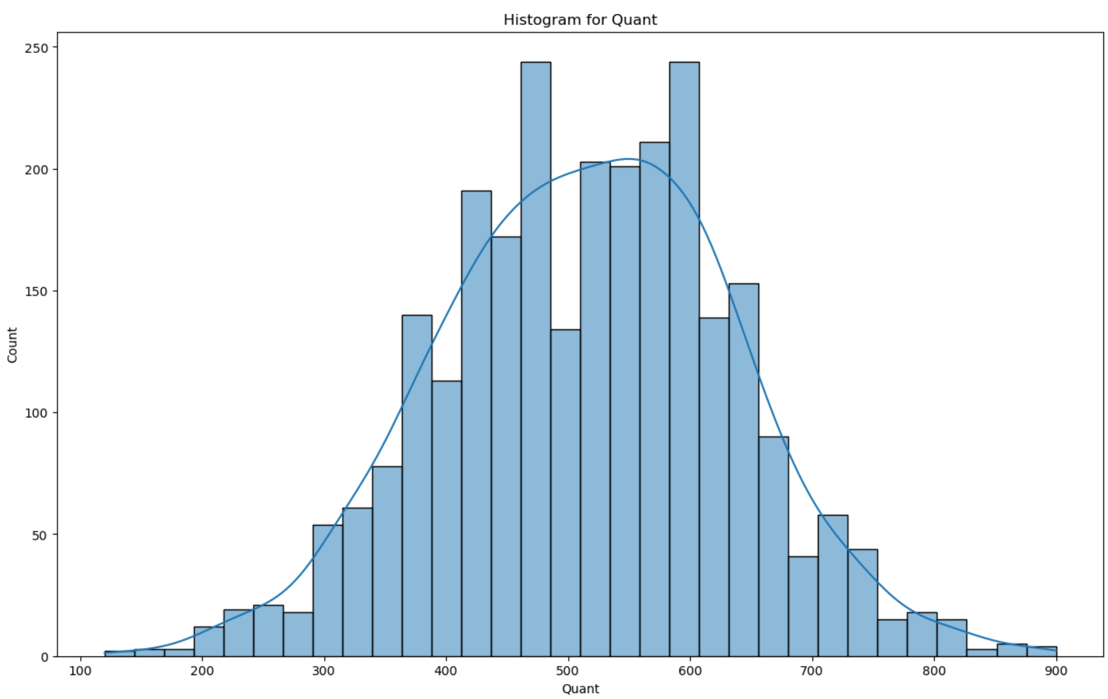

Histogram for 10percentage
Count
0
25
50
75
100
125
150
175
200
50
60
70
80
90
10percentage

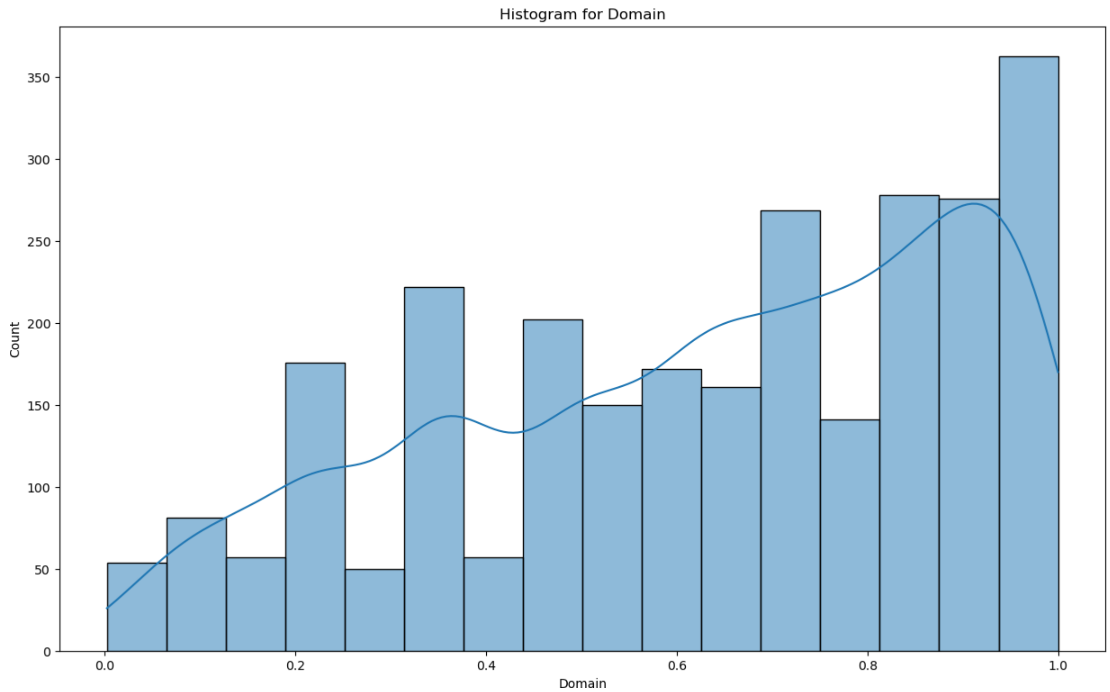

Histogram for 12percentage
Count
0
50
100
150
200
40
50
60
70
80
90
100
12percentage

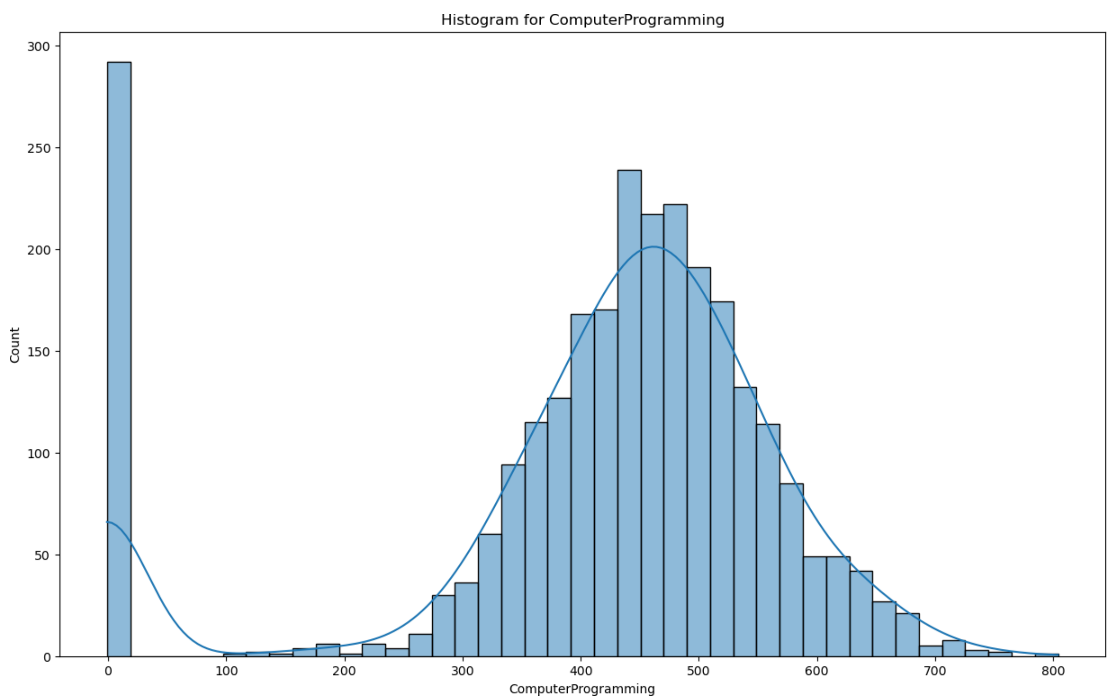

Histogram for collegeGPA
Count
collegeGPA
0
50
100
150
200
250
20
40
60
80
100

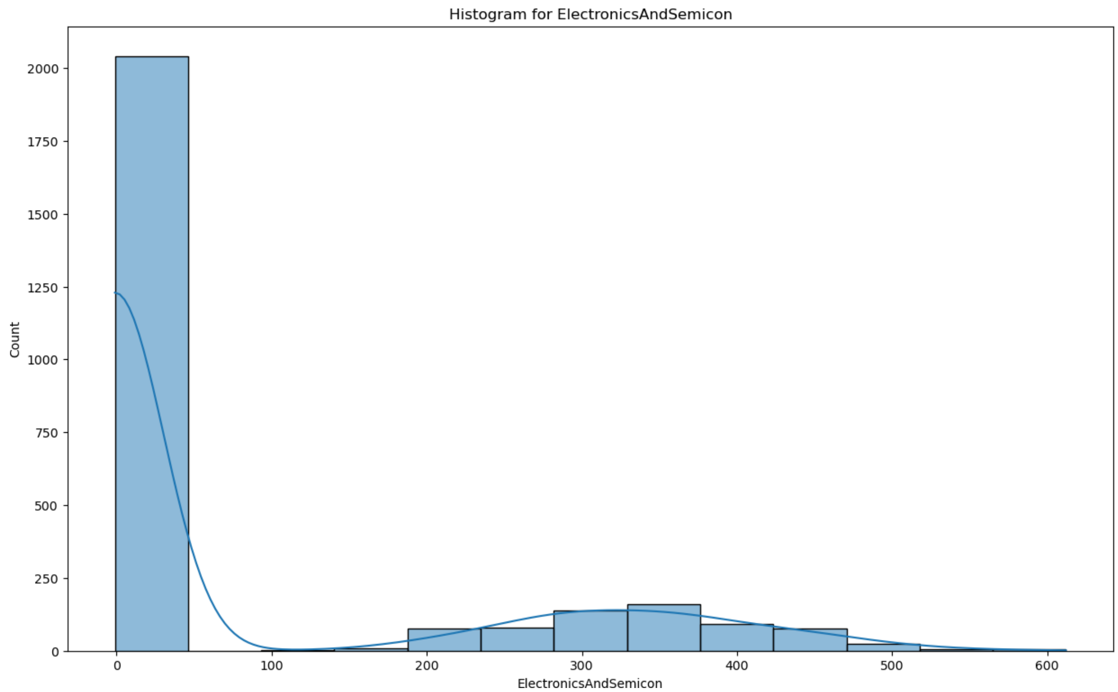

Histogram for English
Count
English
0
50
100
150
200
200
300
400
500
600
700
800
900

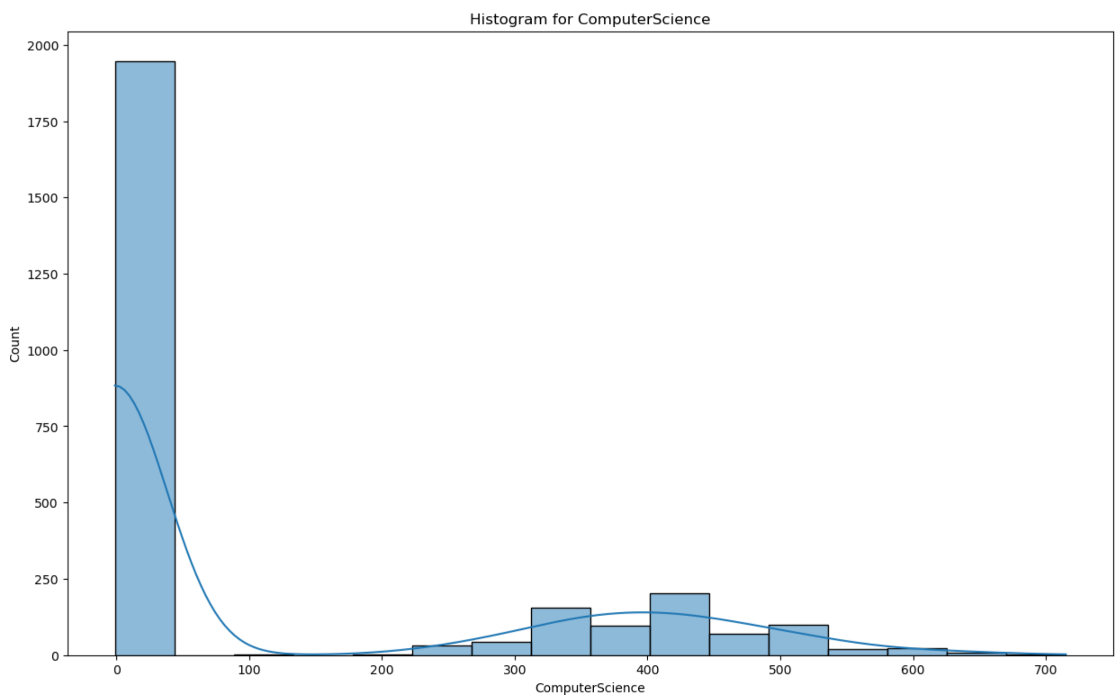

Histogram for Logical
Count
Logical
0
50
100
150
200
200
300
400
500
600
700
800

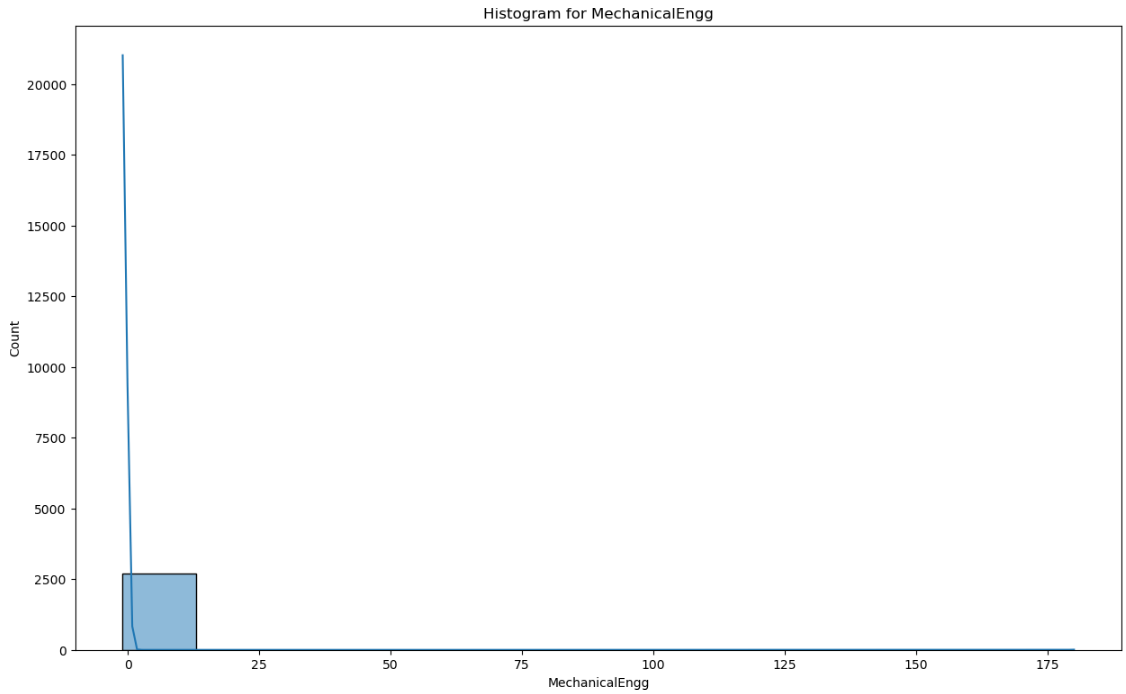

Histogram for Quant
Count
Quant
0
50
100
150
200
250
100
200
300
400
500
600
700
800
900

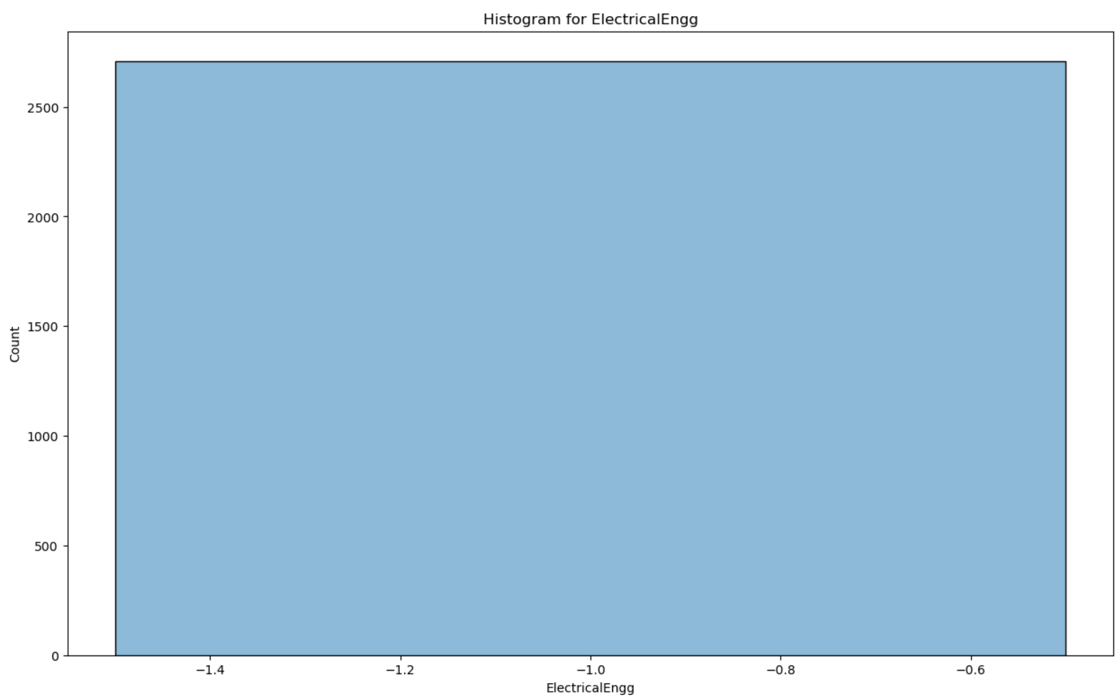

Histogram for Domain
Count
Domain
0
50
100
150
200
250
300
350
0.0
0.2
0.4
0.6
0.8
1.0

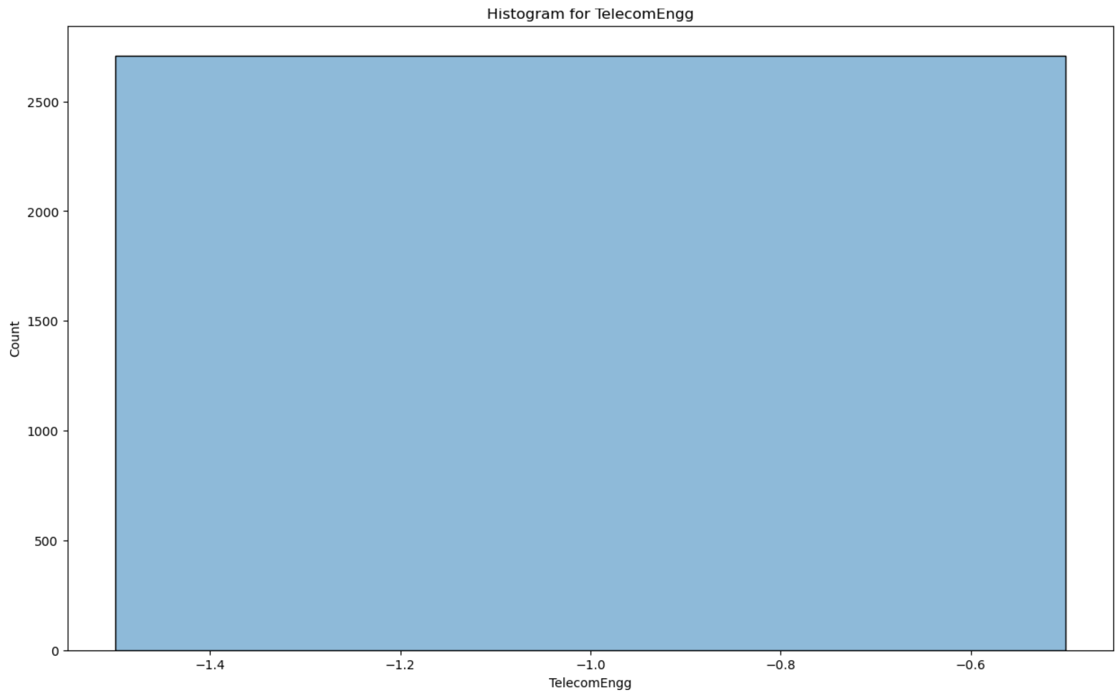

Histogram for ComputerProgramming
Count
ComputerProgramming
0
50
100
150
200
250
300
0
100
200
300
400
500
600
700
800

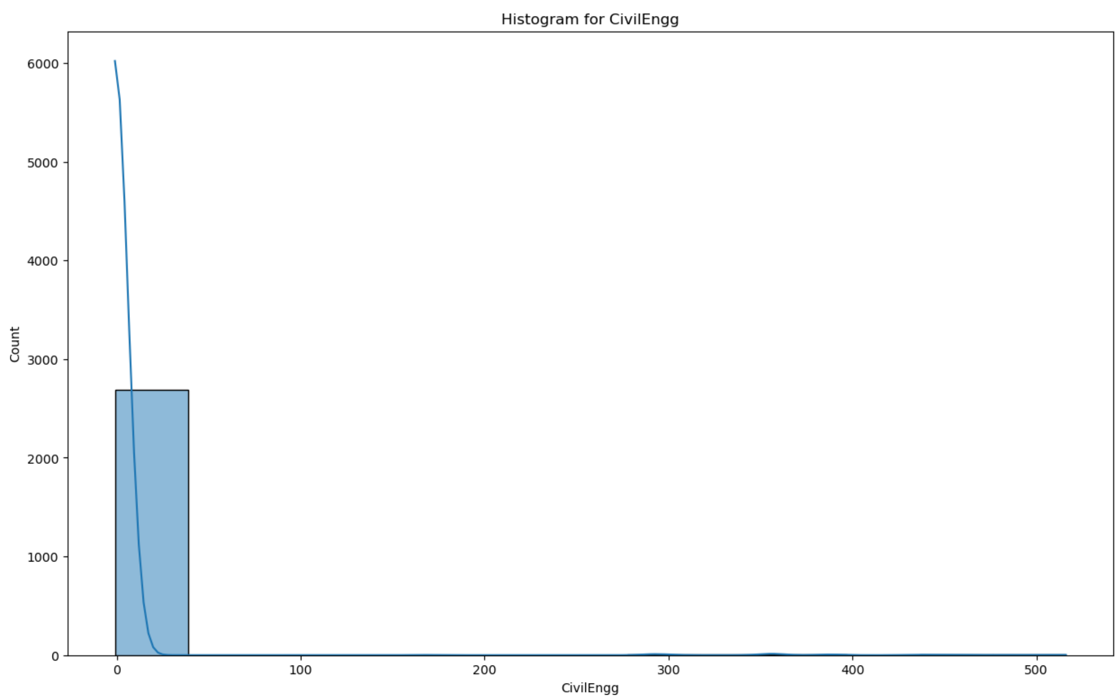

Histogram for ElectronicsAndSemicon
Count
0
250
500
750
1000
1250
1500
1750
2000
0
100
200
300
400
500
600
ElectronicsAndSemicon

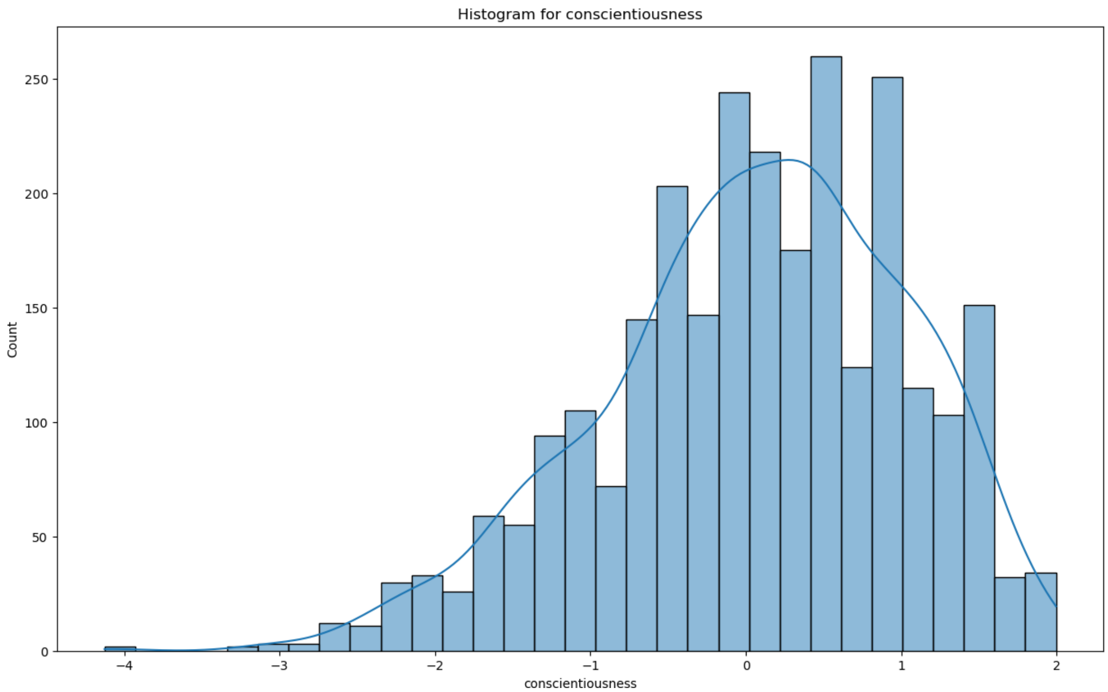

Histogram for ComputerScience
Count
0
250
500
750
1000
1250
1500
1750
2000
0
100
200
300
400
500
600
700
ComputerScience

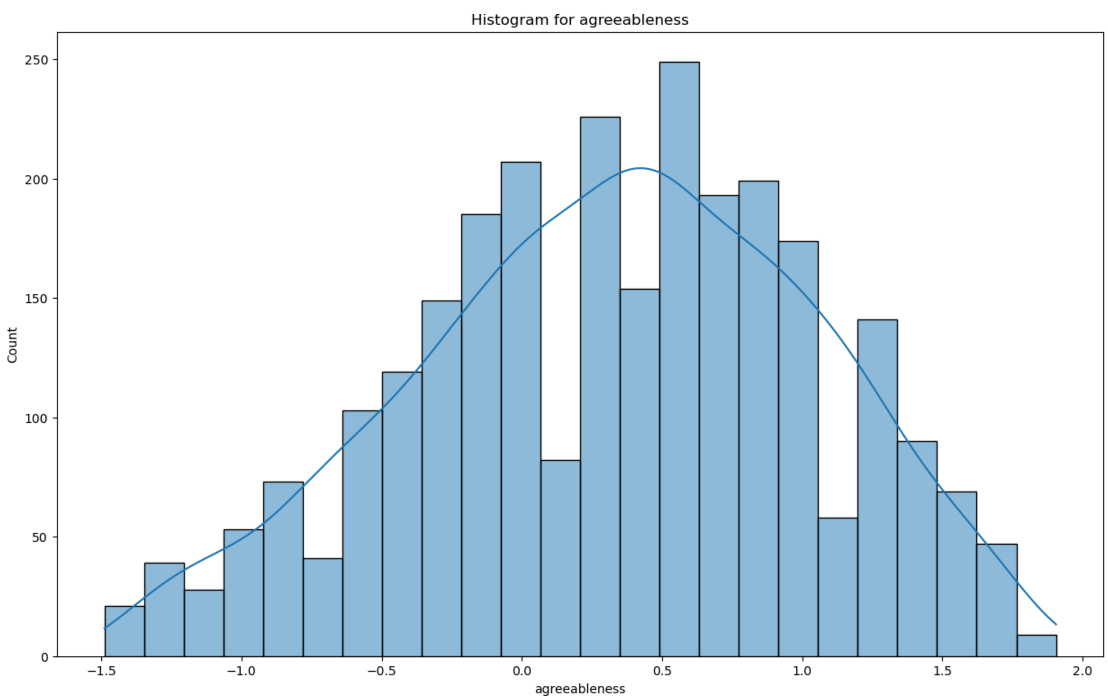

Histogram for MechanicalEngg
Count
MechanicalEngg
20000
17500
15000
12500
10000
7500
5000
2500
0
0
25
50
75
100
125
150
175

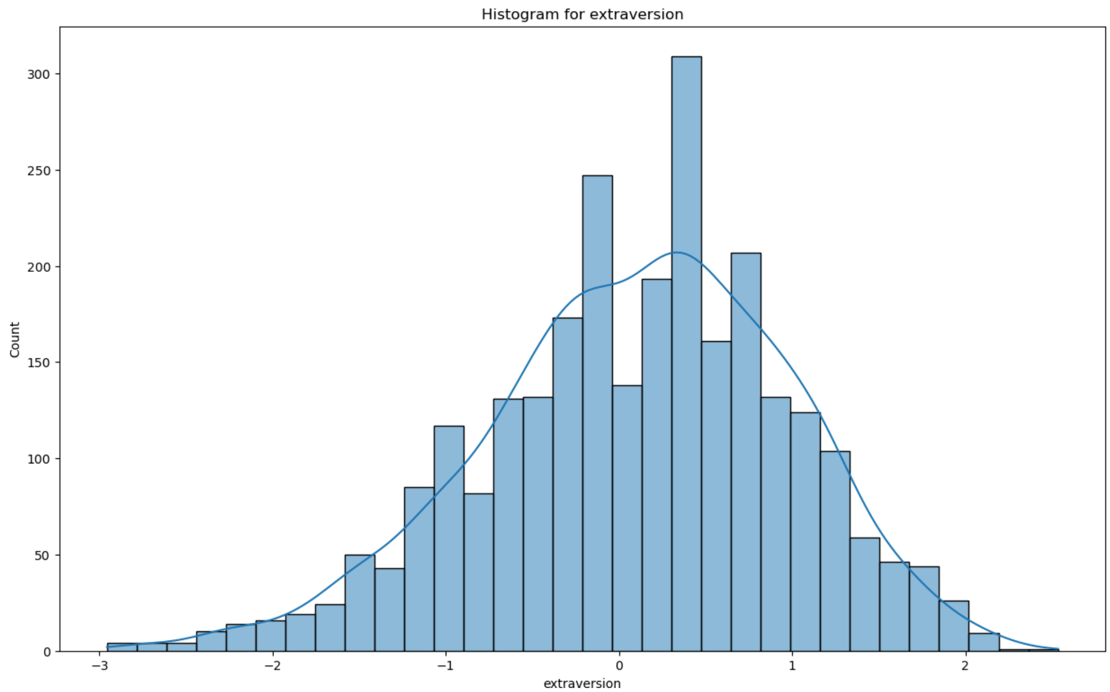

Histogram for ElectricalEngg
Count
ElectricalEngg
2500
2000
1500
1000
500
0
−1.4
−1.2
−1.0
−0.8
−0.6

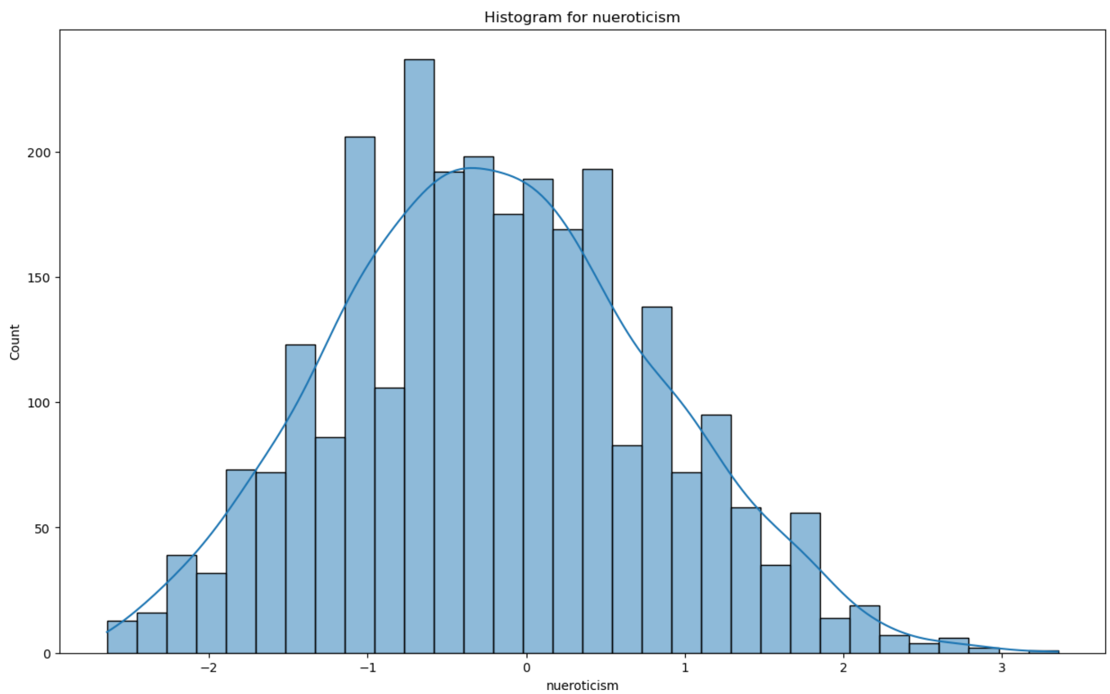

Histogram for TelecomEngg
Count
TelecomEngg
0
500
1000
1500
2000
2500
−1.4
−1.2
−1.0
−0.8
−0.6

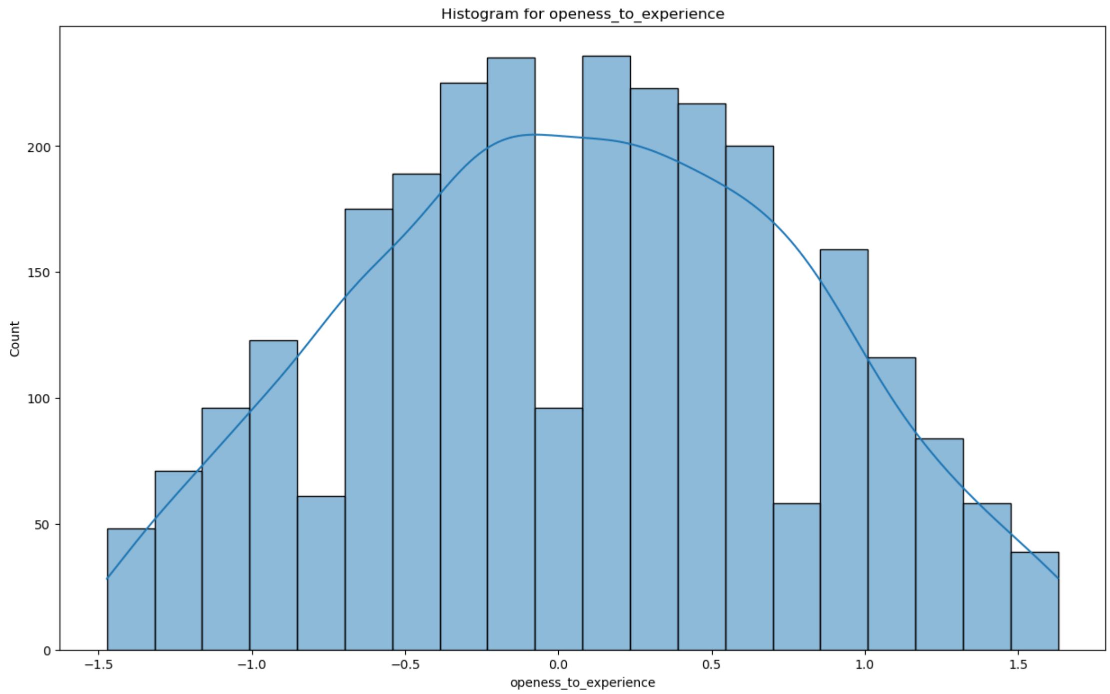

Histogram for CivilEngg
Count
CivilEngg
0
1000
2000
3000
4000
5000
6000
0
100
200
300
400
500


Histogram for conscientiousness
Count
conscientiousness
250
200
150
100
50
0
−4
−3
−2
−1
0
1
2


Histogram for agreeableness
Count
agreeableness
250
200
150
100
50
0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Histogram for extraversion
Count
extraversion
300
250
200
150
100
50
0
−3
−2
−1
0
1
2


Histogram for nueroticism
Count
nueroticism
200
150
100
50
0
−2
−1
0
1
2
3



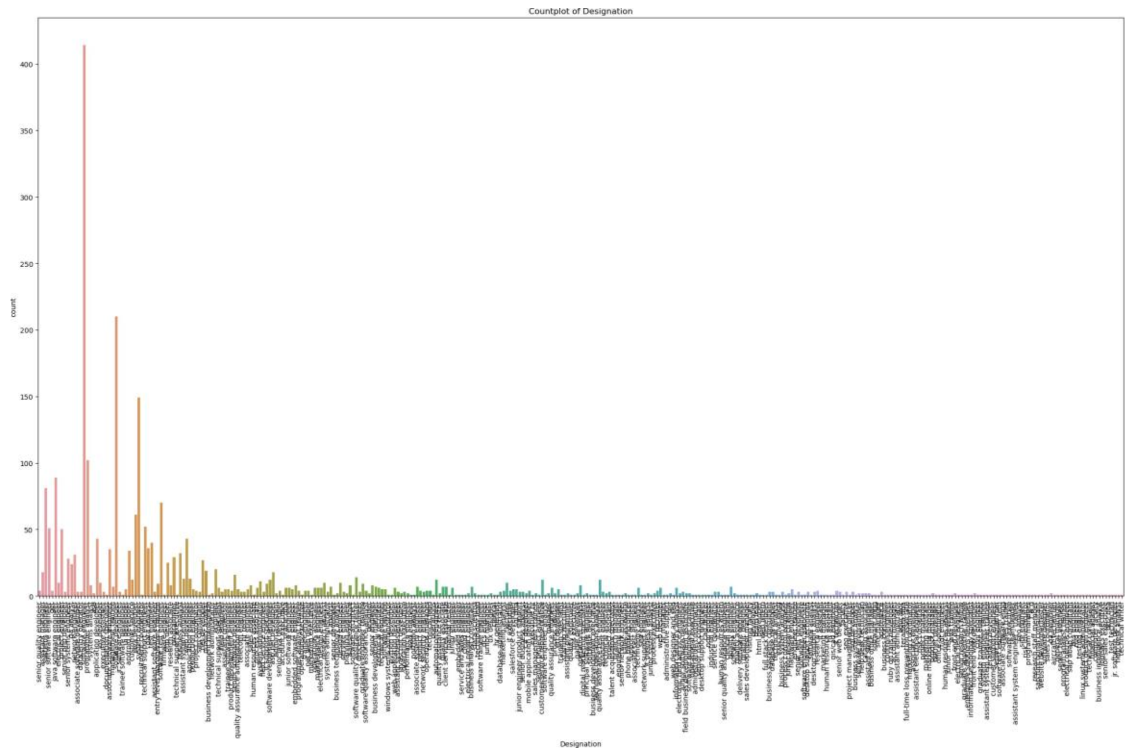
0.7.4 From these visualisations

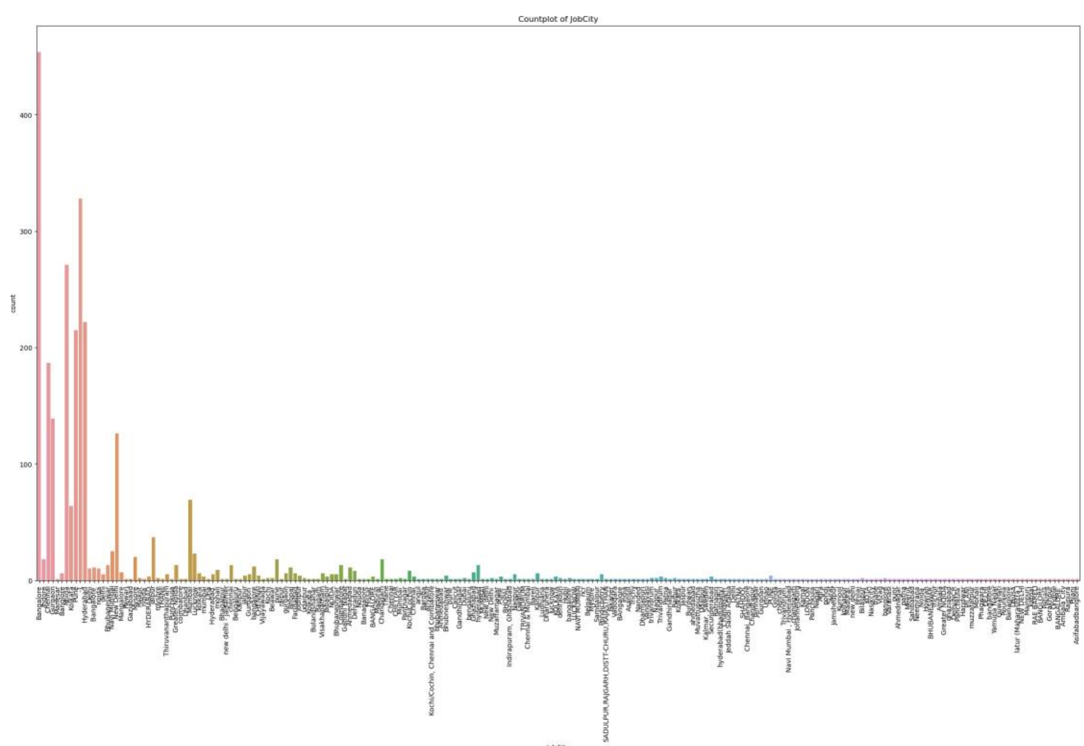
- Most of the salaries are between 100000 and 1000000.
- Most of the persons have around 90%. (left skewed distribution)
- Most number of persons are graduate 12th in between 2007 and 2010
- The histogram plot of 12percentage is slightly leftskewed (very slight). Most of the person have 70% on their 12th.
- Most of the students are from tier 2 colleges.
- Most of the students 70-80 CGPA on their college and they graduated in around 2000s.

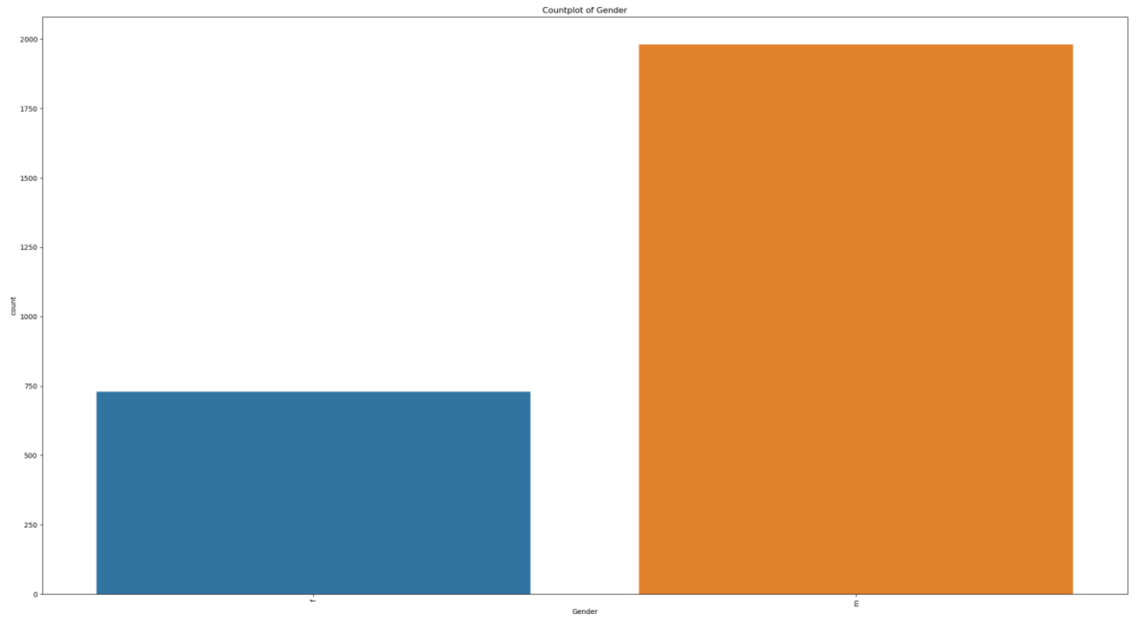
0.8 Categorical Variables

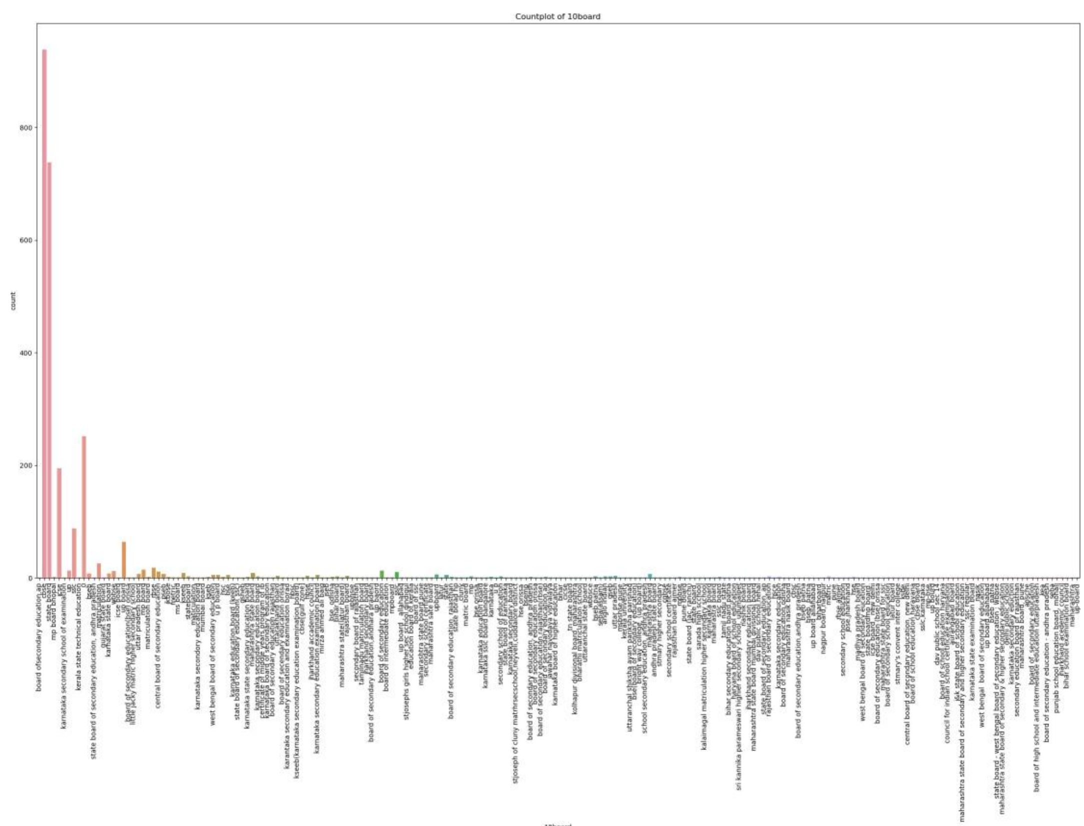
[27]:
```
# Frequency Distribution for Categorical Variables
categorical_cols = ['Designation', 'JobCity', 'Gender', '10board', '12board',␣
↪'CollegeTier', 'Degree', 'Specialization',
                    'CollegeCityTier', 'CollegeState']

for col in categorical_cols:
    plt.figure(figsize=(28,15))

    sns.countplot(x=df[col])
    plt.title(f"Countplot of {col}")
    plt.xticks(rotation=90)
plt.tight_layout()
```

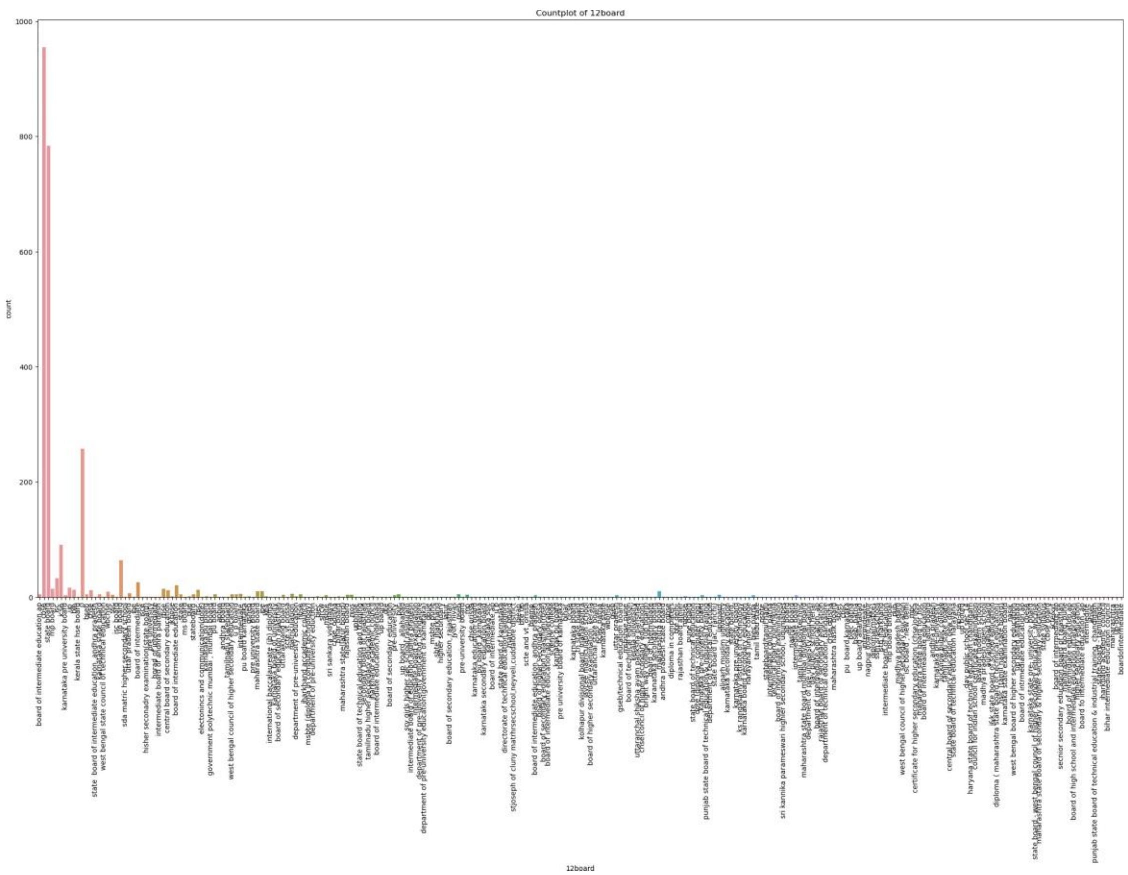
```
plt.show()
```

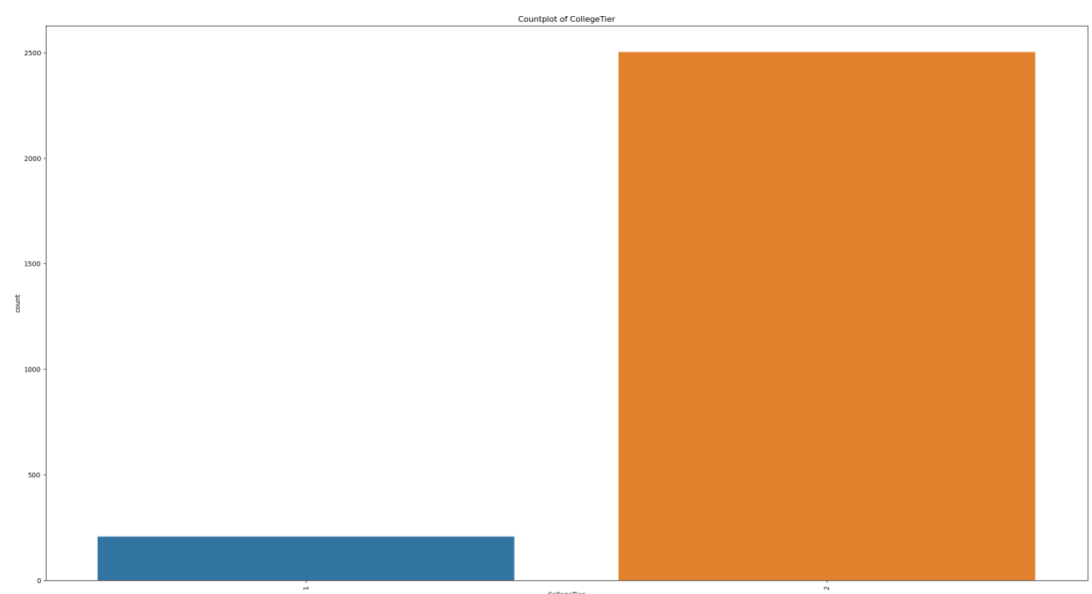

Countplot of Designation
count
Designation

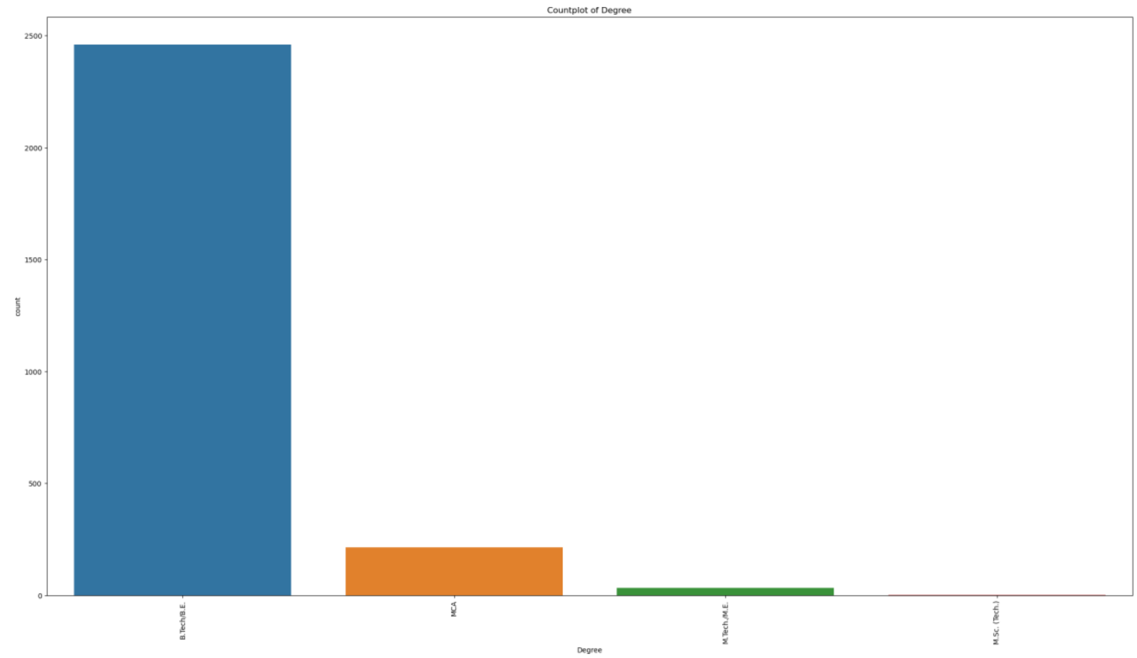

Countplot of JobCity
count
JobCity

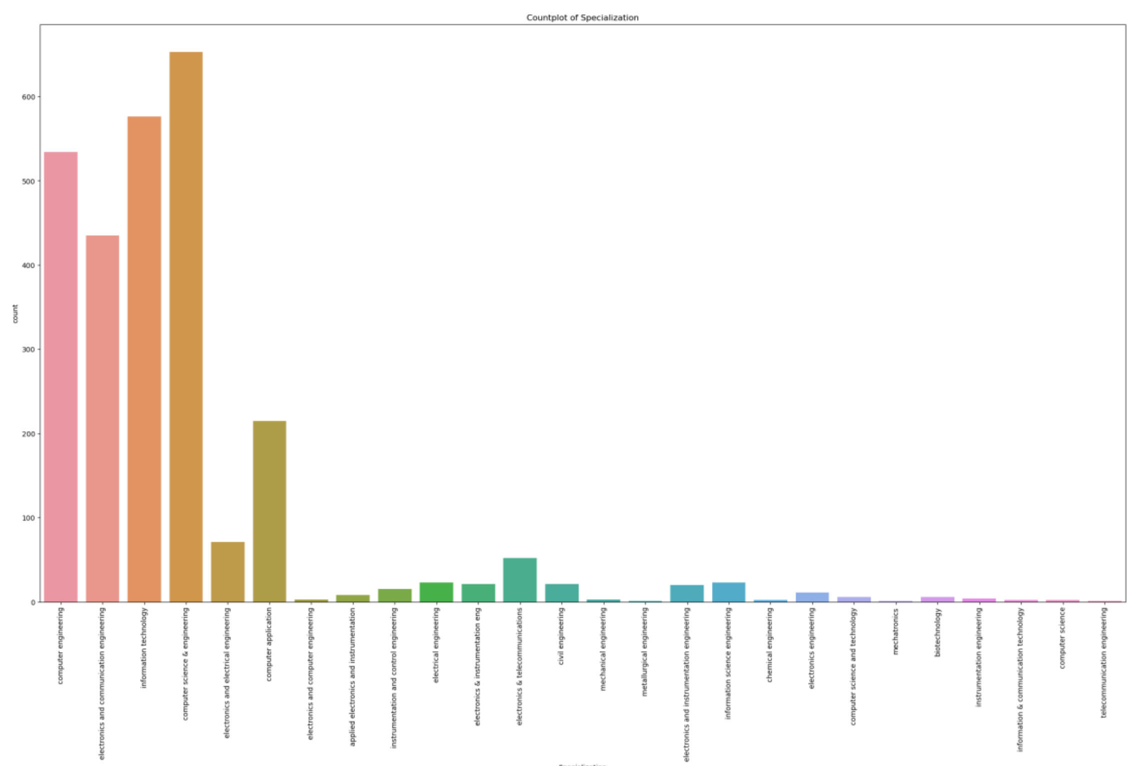

Countplot of Gender
count
Gender
0
250
500
750
1000
1250
1500
1750
2000
f
m

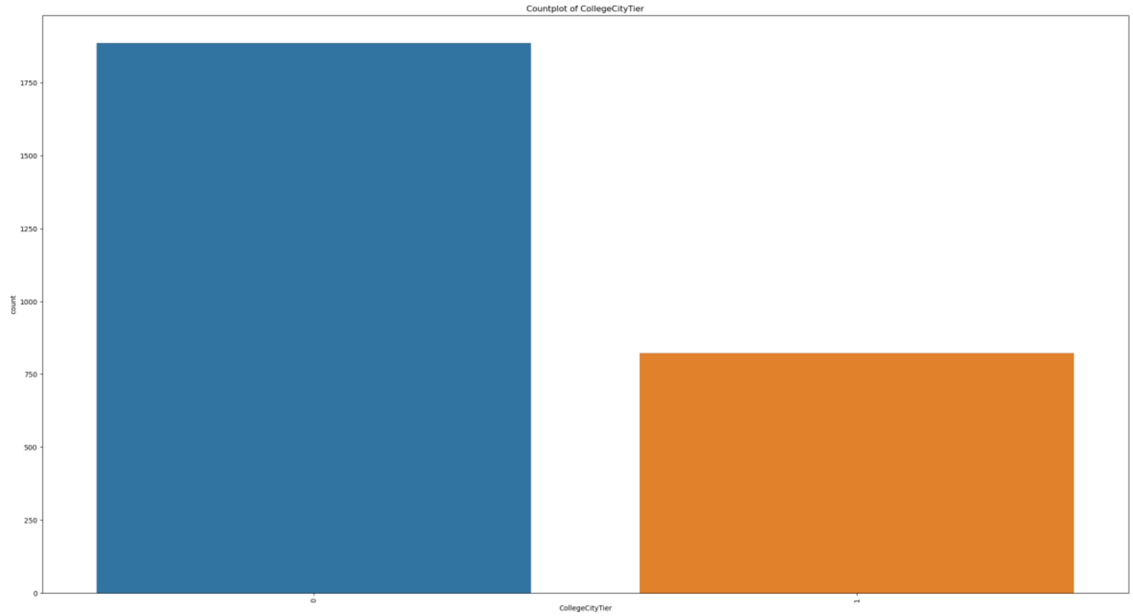

Countplot of 10board
count
10board
0
200
400
600
800

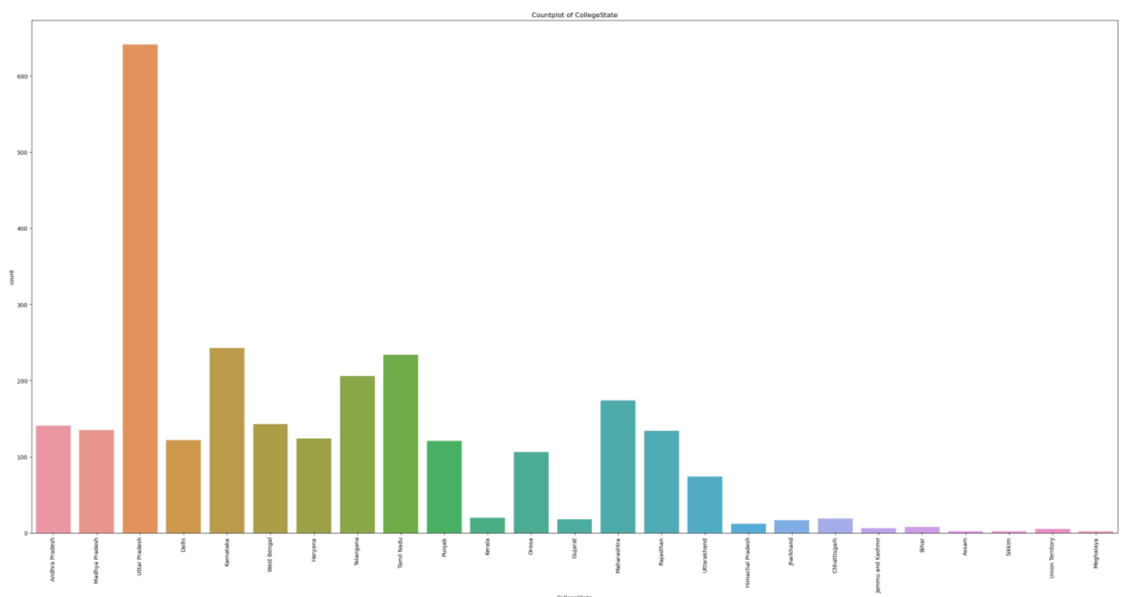

Countplot of 12board
count
0
200
400
600
800
1000
12board


Countplot of CollegeTier
count
0
500
1000
1500
2000
2500
1
2
CollegeTier


Countplot of Degree
count
2500
2000
1500
1000
500
0
B.Tech/B.E.
MCA
M.Tech./M.E.
M.Sc. (Tech.)
Degree


Countplot of Specialization
count
600
500
400
300
200
100
0
computer engineering
electronics and communication engineering
information technology
computer science & engineering
electronics and electrical engineering
computer application
electronics and computer engineering
applied electronics and instrumentation
instrumentation and control engineering
electrical engineering
electronics & instrumentation eng
electronics & telecommunications
civil engineering
mechanical engineering
metallurgical engineering
electronics and instrumentation engineering
information science engineering
chemical engineering
electronics engineering
computer science and technology
mechatronics
biotechnology
instrumentation engineering
information & communication technology
computer science
telecommunication engineering
Specialization





0.9 Step 4 - Bivariate Visual and Non Visual Analysis

[28]:
```
df.columns
```

[28]:
```
Index(['Unnamed: 0', 'ID', 'Salary', 'DOJ', 'DOL', 'Designation', 'JobCity',
       'Gender', 'DOB', '10percentage', '10board', '12graduation',
       '12percentage', '12board', 'CollegeID', 'CollegeTier', 'Degree',
```

```
       'Specialization', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier',
       'CollegeState', 'GraduationYear', 'English', 'Logical', 'Quant',
       'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon',
       'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg',
       'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion',
       'nueroticism', 'openess_to_experience'],
      dtype='object')
```

[29]:
```
df.corr()
```

[29]:
```
                             ID    Salary  10percentage  12graduation  \
ID                     1.000000 -0.253513      0.023843      0.686332
Salary                -0.253513  1.000000      0.209723     -0.143079
10percentage           0.023843  0.209723      1.000000      0.263105
12graduation           0.686332 -0.143079      0.263105      1.000000
12percentage          -0.011916  0.210189      0.643323      0.247061
CollegeID              0.276407 -0.100161      0.035372      0.265697
CollegeTier            0.035974 -0.191846     -0.119124      0.031316
collegeGPA             0.041150  0.146688      0.319736      0.072646
CollegeCityID          0.276407 -0.100161      0.035372      0.265697
CollegeCityTier       -0.045305  0.031335      0.112246     -0.012582
GraduationYear         0.826515 -0.211138      0.083448      0.796481
English                0.114377  0.191779      0.343932      0.151548
Logical                0.075074  0.204790      0.324946      0.099572
Quant                 -0.066181  0.239366      0.314038     -0.020797
Domain                -0.042281  0.191677      0.161276     -0.038077
ComputerProgramming    0.039246  0.125277      0.083267     -0.016384
ElectronicsAndSemicon -0.068386  0.014616      0.099278      0.008108
ComputerScience        0.575251 -0.125329     -0.002791      0.377201
MechanicalEngg        -0.031074  0.007895      0.008875     -0.022683
ElectricalEngg              NaN       NaN           NaN           NaN
TelecomEngg                 NaN       NaN           NaN           NaN
CivilEngg              0.025354  0.045341      0.037666      0.046299
conscientiousness      0.196506 -0.075857      0.030128      0.110904
agreeableness          0.045804  0.061069      0.127151      0.077190
extraversion           0.161519 -0.035436     -0.038216      0.083115
nueroticism           -0.148510 -0.048994     -0.136929     -0.100481
openess_to_experience  0.091721 -0.039208     -0.011832      0.021565

                       12percentage  CollegeID  CollegeTier  collegeGPA  \
ID                        -0.011916   0.276407     0.035974    0.041150
Salary                     0.210189  -0.100161    -0.191846    0.146688
10percentage               0.643323   0.035372    -0.119124    0.319736
12graduation               0.247061   0.265697     0.031316    0.072646
12percentage               1.000000   0.029934    -0.102323    0.346490
CollegeID                  0.029934   1.000000     0.068761    0.032171
CollegeTier               -0.102323   0.068761     1.000000   -0.085842
```

collegeGPA	0.346490	0.032171	−0.085842	1.000000
CollegeCityID	0.029934	1.000000	0.068761	0.032171
CollegeCityTier	0.114692	0.011273	−0.103069	−0.001765
GraduationYear	0.050178	0.260039	−0.019372	0.090769
English	0.201549	−0.030402	−0.160695	0.089569
Logical	0.234033	−0.057360	−0.192000	0.188207
Quant	0.304095	−0.124671	−0.241471	0.205683
Domain	0.166567	−0.096676	−0.128843	0.184999
ComputerProgramming	0.101064	−0.023530	−0.085559	0.142678
ElectronicsAndSemicon	0.158497	−0.034412	−0.048185	0.050898
ComputerScience	−0.042151	0.133429	0.005795	0.005567
MechanicalEngg	0.011206	−0.018655	0.005527	−0.026402
ElectricalEngg	NaN	NaN	NaN	NaN
TelecomEngg	NaN	NaN	NaN	NaN
CivilEngg	0.003490	0.019282	−0.071117	0.006362
conscientiousness	0.021221	0.083662	0.086754	0.061387
agreeableness	0.098764	0.022440	−0.027778	0.057475
extraversion	−0.026008	0.034994	0.015684	−0.039635
nueroticism	−0.098781	0.001412	0.018323	−0.065426
openess_to_experience	−0.040206	0.036020	0.010418	−0.004528

	CollegeCityID	CollegeCityTier	...	ComputerScience	\
ID	0.276407	−0.045305	...	0.575251	
Salary	−0.100161	0.031335	...	−0.125329	
10percentage	0.035372	0.112246	...	−0.002791	
12graduation	0.265697	−0.012582	...	0.377201	
12percentage	0.029934	0.114692	...	−0.042151	
CollegeID	1.000000	0.011273	...	0.133429	
CollegeTier	0.068761	−0.103069	...	0.005795	
collegeGPA	0.032171	−0.001765	...	0.005567	
CollegeCityID	1.000000	0.011273	...	0.133429	
CollegeCityTier	0.011273	1.000000	...	−0.025438	
GraduationYear	0.260039	−0.067982	...	0.483505	
English	−0.030402	0.051114	...	0.067863	
Logical	−0.057360	0.013836	...	0.039324	
Quant	−0.124671	0.000704	...	−0.056632	
Domain	−0.096676	−0.002201	...	0.052974	
ComputerProgramming	−0.023530	0.038281	...	0.169312	
ElectronicsAndSemicon	−0.034412	0.015265	...	−0.280969	
ComputerScience	0.133429	−0.025438	...	1.000000	
MechanicalEngg	−0.018655	0.029090	...	−0.011633	
ElectricalEngg	NaN	NaN	...	NaN	
TelecomEngg	NaN	NaN	...	NaN	
CivilEngg	0.019282	−0.035639	...	−0.053510	
conscientiousness	0.083662	−0.009524	...	0.114154	
agreeableness	0.022440	−0.013297	...	0.033534	
extraversion	0.034994	−0.024983	...	0.123327	

nueroticism	0.001412	0.015892	...	-0.123003
openess_to_experience	0.036020	-0.050870	...	0.079165

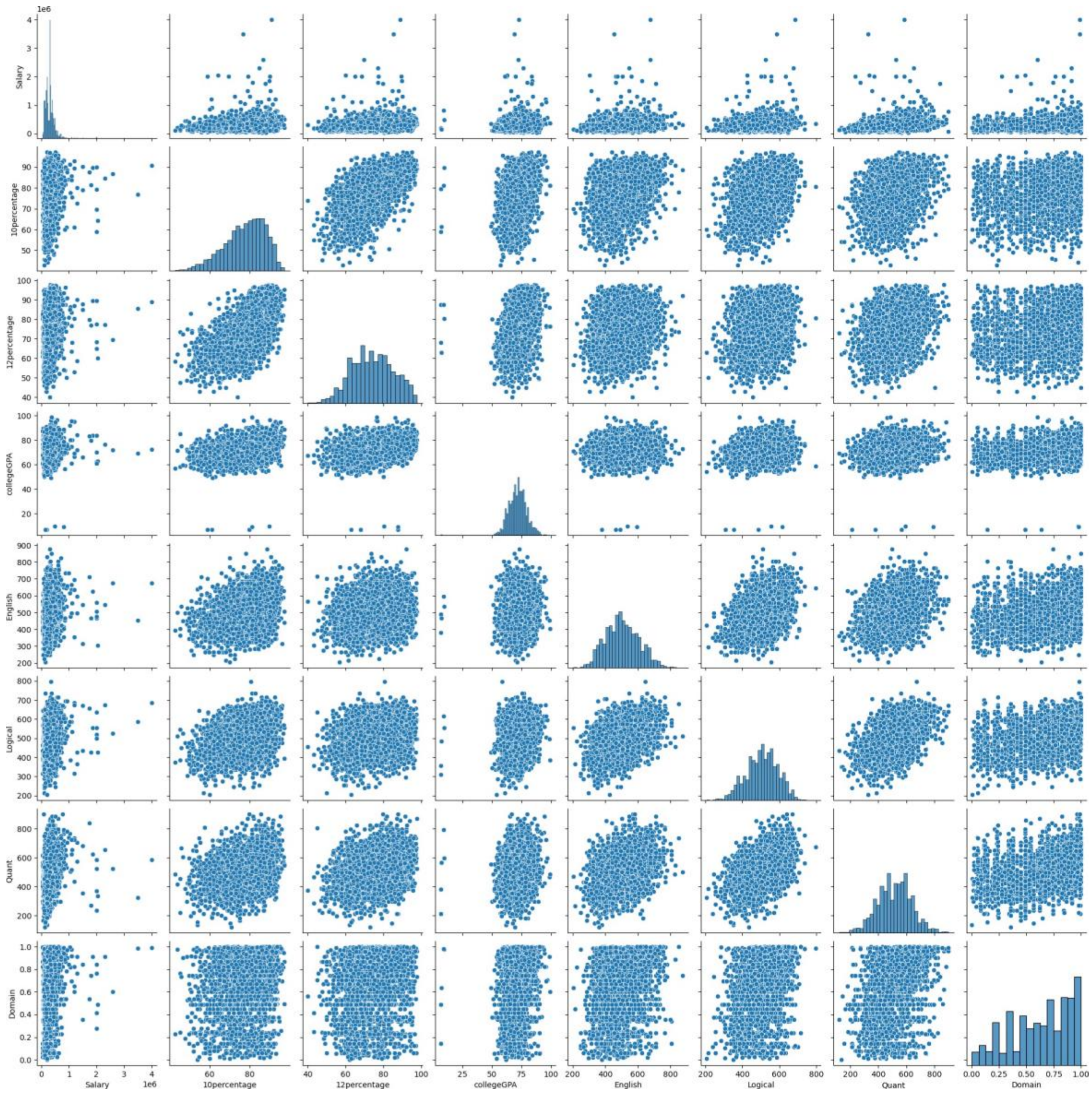
	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg \
ID	-0.031074	NaN	NaN	0.025354
Salary	0.007895	NaN	NaN	0.045341
10percentage	0.008875	NaN	NaN	0.037666
12graduation	-0.022683	NaN	NaN	0.046299
12percentage	0.011206	NaN	NaN	0.003490
CollegeID	-0.018655	NaN	NaN	0.019282
CollegeTier	0.005527	NaN	NaN	-0.071117
collegeGPA	-0.026402	NaN	NaN	0.006362
CollegeCityID	-0.018655	NaN	NaN	0.019282
CollegeCityTier	0.029090	NaN	NaN	-0.035639
GraduationYear	-0.036577	NaN	NaN	0.048997
English	-0.001444	NaN	NaN	0.009335
Logical	-0.009101	NaN	NaN	0.037641
Quant	0.009388	NaN	NaN	0.032211
Domain	-0.036125	NaN	NaN	0.007451
ComputerProgramming	-0.015778	NaN	NaN	-0.143122
ElectronicsAndSemicon	0.019037	NaN	NaN	-0.039709
ComputerScience	-0.011633	NaN	NaN	-0.053510
MechanicalEngg	1.000000	NaN	NaN	-0.001699
ElectricalEngg	NaN	NaN	NaN	NaN
TelecomEngg	NaN	NaN	NaN	NaN
CivilEngg	-0.001699	NaN	NaN	1.000000
conscientiousness	-0.009090	NaN	NaN	-0.013034
agreeableness	-0.028972	NaN	NaN	-0.012668
extraversion	-0.003405	NaN	NaN	-0.018528
nueroticism	0.009519	NaN	NaN	-0.015358
openess_to_experience	-0.001241	NaN	NaN	-0.004765

	conscientiousness	agreeableness	extraversion \
ID	0.196506	0.045804	0.161519
Salary	-0.075857	0.061069	-0.035436
10percentage	0.030128	0.127151	-0.038216
12graduation	0.110904	0.077190	0.083115
12percentage	0.021221	0.098764	-0.026008
CollegeID	0.083662	0.022440	0.034994
CollegeTier	0.086754	-0.027778	0.015684
collegeGPA	0.061387	0.057475	-0.039635
CollegeCityID	0.083662	0.022440	0.034994
CollegeCityTier	-0.009524	-0.013297	-0.024983
GraduationYear	0.137882	0.049936	0.122741
English	-0.008814	0.192459	-0.017309
Logical	-0.040995	0.116998	-0.053061
Quant	-0.064322	0.071734	-0.051329

```
Domain                 -0.048119  0.064033  -0.067426
ComputerProgramming    -0.002157  0.076376   0.008772
ElectronicsAndSemicon  -0.030535 -0.037518  -0.034174
ComputerScience         0.114154  0.033534   0.123327
MechanicalEngg         -0.009090 -0.028972  -0.003405
ElectricalEngg               NaN        NaN        NaN
TelecomEngg                  NaN        NaN        NaN
CivilEngg              -0.013034 -0.012668  -0.018528
conscientiousness       1.000000  0.390280   0.276662
agreeableness           0.390280  1.000000   0.341837
extraversion            0.276662  0.341837   1.000000
nueroticism            -0.355232 -0.229158  -0.108542
openess_to_experience   0.278304  0.372215   0.298506

                       nueroticism  openess_to_experience
ID                       -0.148510               0.091721
Salary                   -0.048994              -0.039208
10percentage             -0.136929              -0.011832
12graduation             -0.100481               0.021565
12percentage             -0.098781              -0.040206
CollegeID                 0.001412               0.036020
CollegeTier               0.018323               0.010418
collegeGPA               -0.065426              -0.004528
CollegeCityID             0.001412               0.036020
CollegeCityTier           0.015892              -0.050870
GraduationYear           -0.098999               0.039004
English                  -0.147969               0.027620
Logical                  -0.171760              -0.025763
Quant                    -0.117478              -0.026928
Domain                   -0.109648              -0.048364
ComputerProgramming      -0.095920               0.020141
ElectronicsAndSemicon     0.009627              -0.025960
ComputerScience          -0.123003               0.079165
MechanicalEngg            0.009519              -0.001241
ElectricalEngg                 NaN                    NaN
TelecomEngg                    NaN                    NaN
CivilEngg                -0.015358              -0.004765
conscientiousness        -0.355232               0.278304
agreeableness            -0.229158               0.372215
extraversion             -0.108542               0.298506
nueroticism               1.000000              -0.076209
openess_to_experience    -0.076209               1.000000

[27 rows x 27 columns]
```

[30]:
```
# Scatter plot between Salary and other numerical columns
```

```
sns.pairplot(df, vars=['Salary', '10percentage', '12percentage', 'collegeGPA',␣
 ↪'English', 'Logical', 'Quant', 'Domain'])
plt.show()
```



0.10 Salary vs Job

[49]:
```
df.groupby('Job_Role')['Salary'].describe().round(2).sort_values('mean')
```

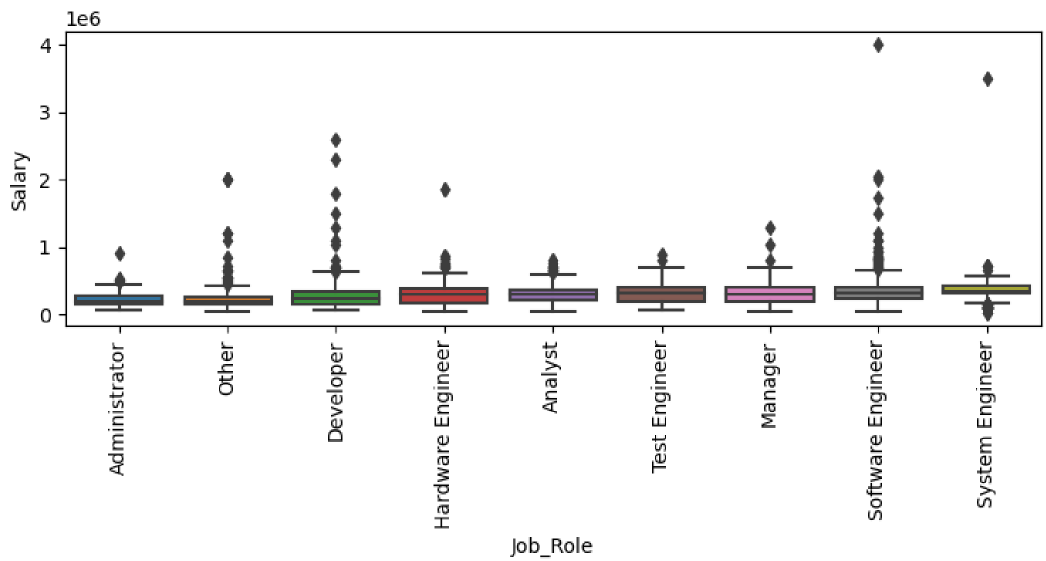
[49]:

	count	mean	std	min	25%	50% \
Job_Role						
Administrator	124.0	232177.42	117028.32	80000.0	150000.0	200000.0
Other	235.0	258170.21	256590.59	45000.0	145000.0	200000.0

```
Developer          599.0  269098.50  211345.08  60000.0  145000.0  240000.0
Hardware Engineer  220.0  306568.18  182966.85  50000.0  183750.0  295000.0
Analyst            302.0  318907.28  135441.19  50000.0  210000.0  312500.0
Test Engineer      118.0  331610.17  158412.10  60000.0  200000.0  325000.0
Manager             68.0  342279.41  216204.43  50000.0  205000.0  300000.0
Software Engineer  710.0  354957.75  233538.42  50000.0  240000.0  320000.0
System Engineer    333.0  362417.42  202256.69  35000.0  320000.0  335000.0

                        75%        max
Job_Role
Administrator      287500.0   910000.0
Other              267500.0  2000000.0
Developer          340000.0  2600000.0
Hardware Engineer  381250.0  1860000.0
Analyst            368750.0   800000.0
Test Engineer      415000.0   900000.0
Manager            403750.0  1300000.0
Software Engineer  413750.0  4000000.0
System Engineer    420000.0  3500000.0
```

[50]:
```
order = df.groupby("Job_Role")["Salary"].mean().sort_values().index
```

[51]:
```
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(8,8))
sns.barplot(x="Job_Role", y="Salary", data=df, order=order, ax=ax1)
sns.boxplot(x="Job_Role", y="Salary", data=df, order=order, ax=ax2)
ax1.tick_params("x", labelrotation=90)
ax2.tick_params("x", labelrotation=90)
plt.tight_layout()
plt.suptitle("Salary")
plt.show()
```



0.10.1 Observation:

- By the above graph Managers are Earning More than others.
- The second Most Earner from the plot is System Engineer

0.11 Salary vs CollegeTier

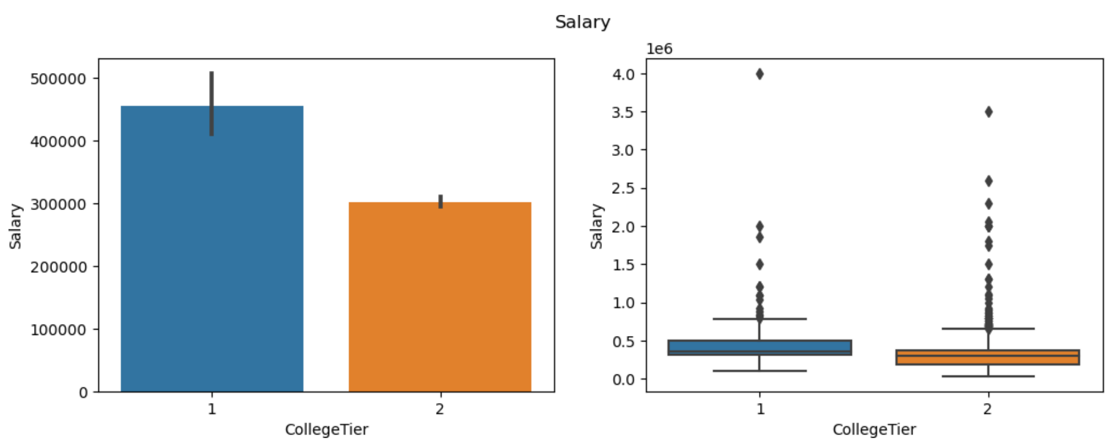
[31]:
```
df.groupby('CollegeTier')['Salary'].describe()
```

[31]:

CollegeTier	count	mean	std	min	25%	50% \
1	207.0	453864.73430	355333.55185	100000.0	310000.0	360000.0
2	2502.0	301984.41247	189070.38349	35000.0	180000.0	300000.0

```
                   75%        max
CollegeTier
1             500000.0  4000000.0
2             370000.0  3500000.0
```

[32]:
```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12,4))
sns.barplot(x='CollegeTier', y='Salary', data=df, ax=ax1)
sns.boxplot(x='CollegeTier', y='Salary', data=df, ax=ax2)
plt.suptitle('Salary')
plt.show()
```



0.11.1 Observation:

The people who are from Tier-1 college are Earning More as compared to Tire-2

0.12 Salary vs Specialization

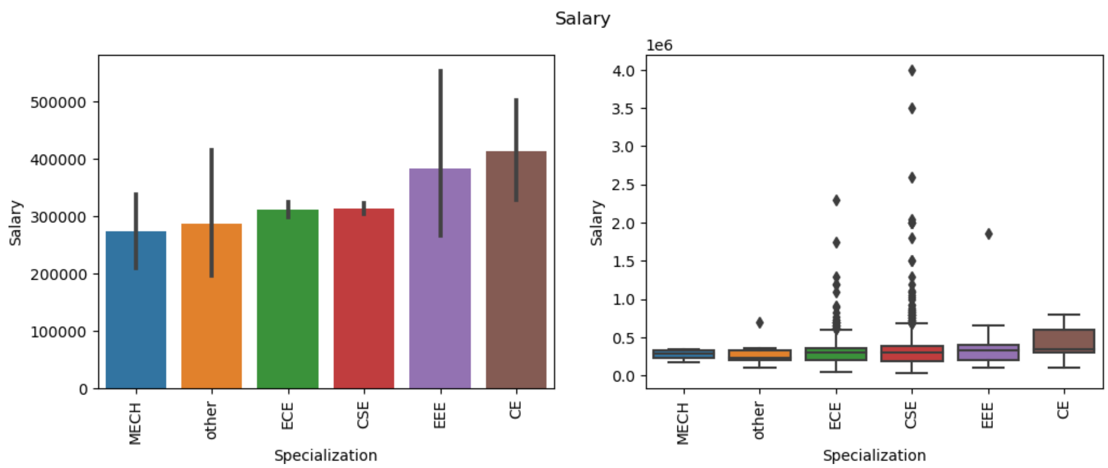
[36]:
```
df.groupby('Specialization')['Salary'].describe().round(1).sort_values('mean')
```

[36]:
```
                 count      mean       std       min       25%       50%  \
Specialization
MECH               4.0  273750.0   78249.1  180000.0  225000.0  282500.0
other              9.0  287222.2  174393.8  100000.0  200000.0  235000.0
ECE              640.0  311312.5  181752.2   45000.0  200000.0  300000.0
CSE             2012.0  312676.4  216744.0   35000.0  185000.0  300000.0
EEE               23.0  382826.1  351980.8  110000.0  205000.0  335000.0
CE                21.0  413571.4  214302.0  110000.0  295000.0  345000.0

                      75%        max
Specialization
MECH             331250.0   350000.0
```

```
other           325000.0   700000.0
ECE             361250.0  2300000.0
CSE             385000.0  4000000.0
EEE             407500.0  1860000.0
CE              600000.0   800000.0
```

[37]:
```python
order = df.groupby('Specialization')['Salary'].mean().sort_values().index
```

[38]:
```python
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12,4))
sns.barplot(x='Specialization', y='Salary', data=df, order=order, ax=ax1)
sns.boxplot(x='Specialization', y='Salary', data=df, order=order, ax=ax2)
ax1.tick_params('x', labelrotation=90)
ax2.tick_params('x', labelrotation=90)
plt.suptitle('Salary')
plt.show()
```



0.12.1 Observation:

CSE people are earning more as compared to other students

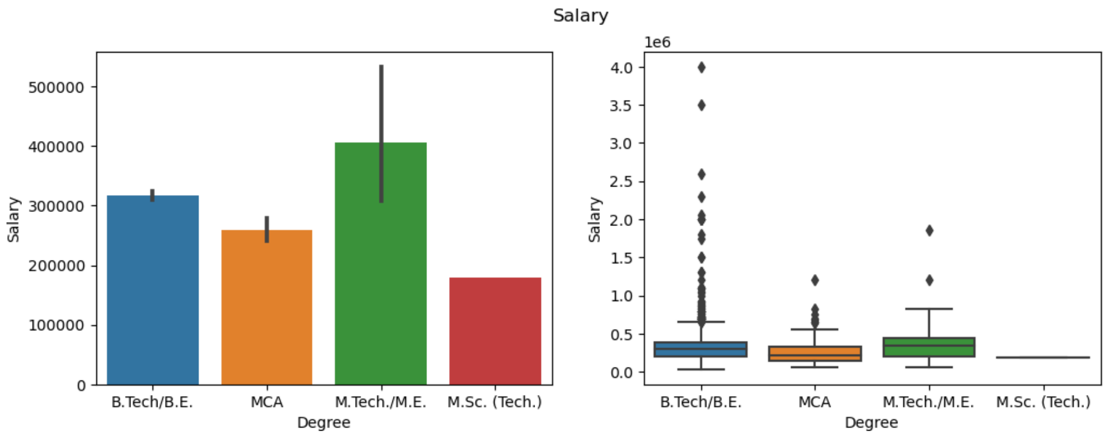
0.13 Salary vs Degree

[39]:
```python
df.groupby('Degree')['Salary'].describe()
```

[39]:
```
                 count           mean            std      min       25%  \
Degree
B.Tech/B.E.     2460.0  317081.300813  211143.976154  35000.0  200000.0
M.Sc. (Tech.)      1.0  180000.000000            NaN 180000.0  180000.0
M.Tech./M.E.      34.0  406470.588235  347705.747706  65000.0  200000.0
MCA              214.0  259322.429907  156805.353943  60000.0  145000.0
```

```
                   50%       75%        max
Degree
B.Tech/B.E.     300000.0  381250.0  4000000.0
M.Sc. (Tech.)   180000.0  180000.0   180000.0
M.Tech./M.E.    345000.0  448750.0  1860000.0
MCA             217500.0  325000.0  1200000.0
```

[40]:
```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12,4))
sns.barplot(x='Degree', y='Salary', data=df, ax=ax1)
sns.boxplot(x='Degree', y='Salary', data=df, ax=ax2)
plt.suptitle('Salary')
plt.show()
```

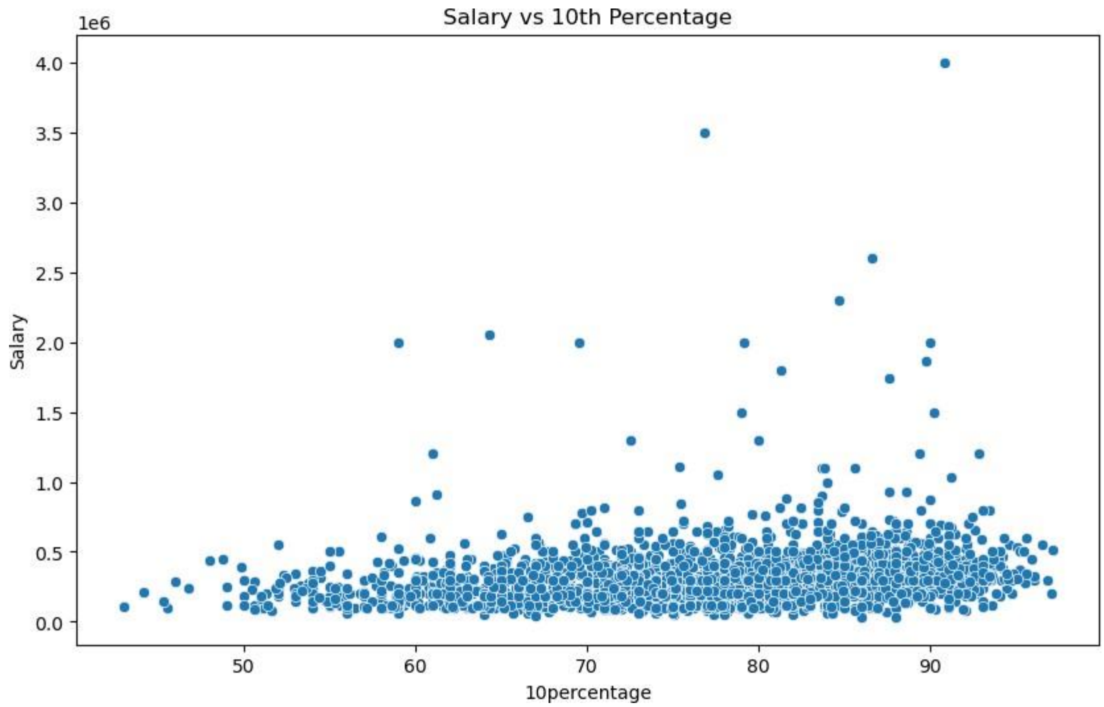


0.13.1 Observation:

M.Tech/M.E students are earning More than others, but B.Tech/B.E Students having more chances to earn better than M.Tech Students.

0.13.2 Numerical vs. Numerical Relationships

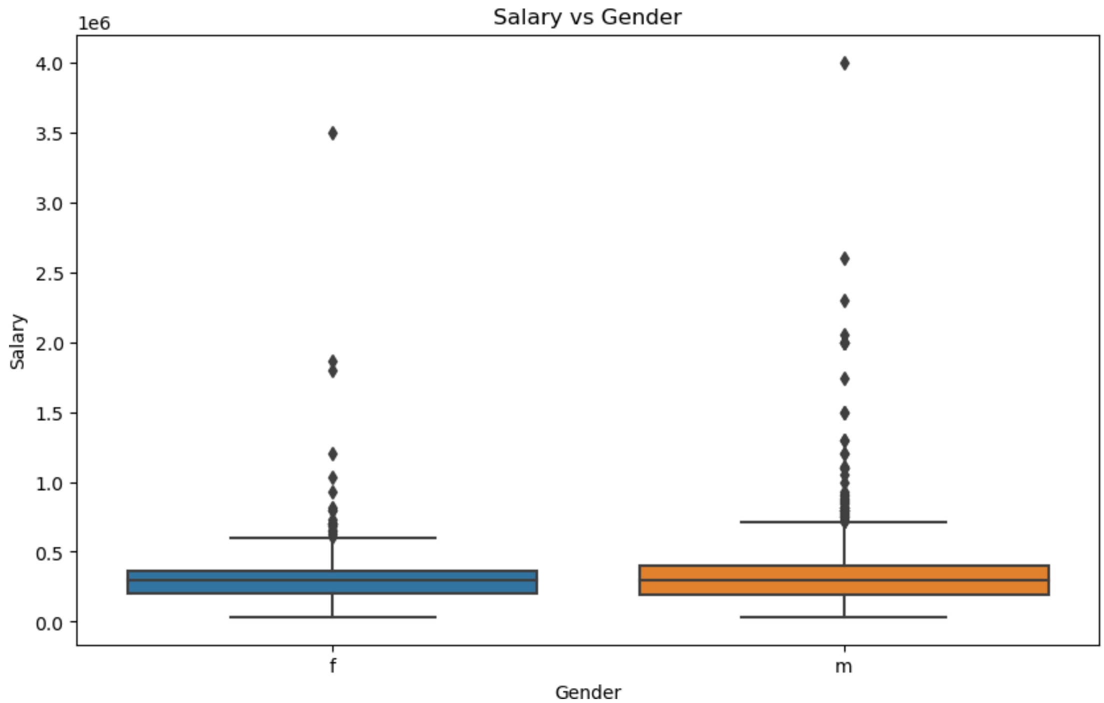
[54]:
```
# Scatter Plot for Salary vs Other Numerical Columns
plt.figure(figsize=(10, 6))
sns.scatterplot(x='10percentage', y='Salary', data=df)
plt.title('Salary vs 10th Percentage')
plt.show()
```



[56]:
```
# Hexbin Plot for Salary vs 10percentage
plt.figure(figsize=(10, 6))
plt.hexbin(df['10percentage'], df['Salary'], gridsize=50, cmap='Blues')
plt.colorbar()
plt.title('Hexbin Plot of Salary vs 10th Percentage')
plt.xlabel('10th Percentage')
plt.ylabel('Salary')
plt.show()
```

0.13.3 Categorical vs. Numerical Relationships

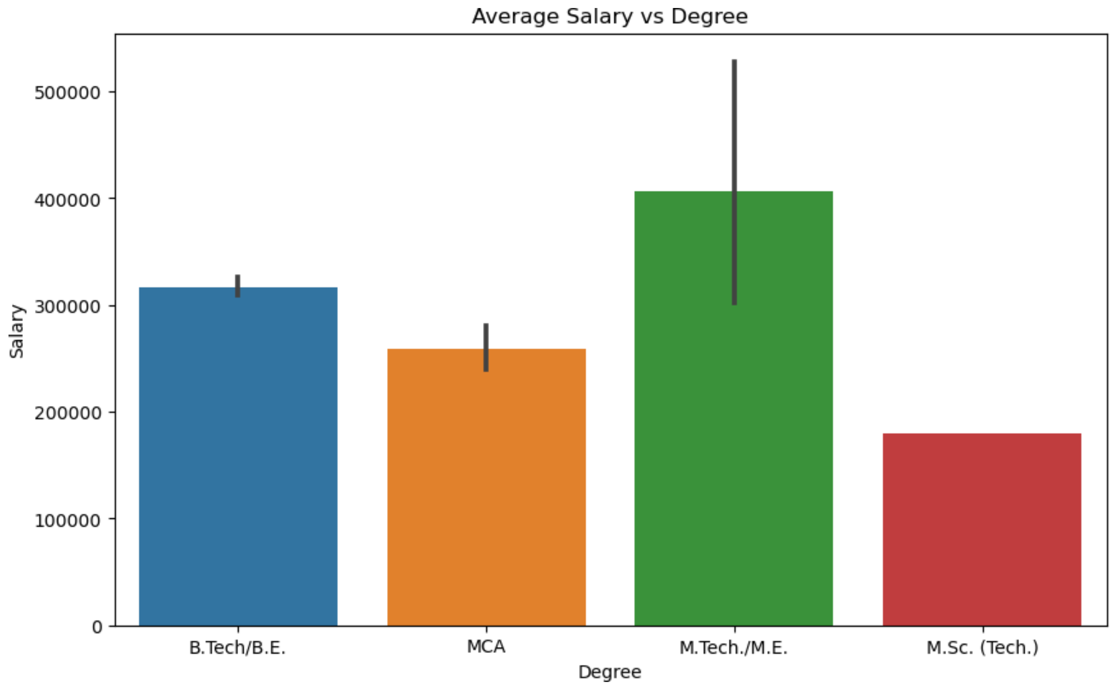
[57]:
```
# Boxplot to compare Salary across different Gender
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='Salary', data=df)
plt.title('Salary vs Gender')
plt.show()
```



[58]:
```
# Swarmplot for Salary vs Degree
plt.figure(figsize=(10, 6))
sns.swarmplot(x="Degree", y="Salary", data=df)
plt.title("Salary vs Degree")
plt.show()
```

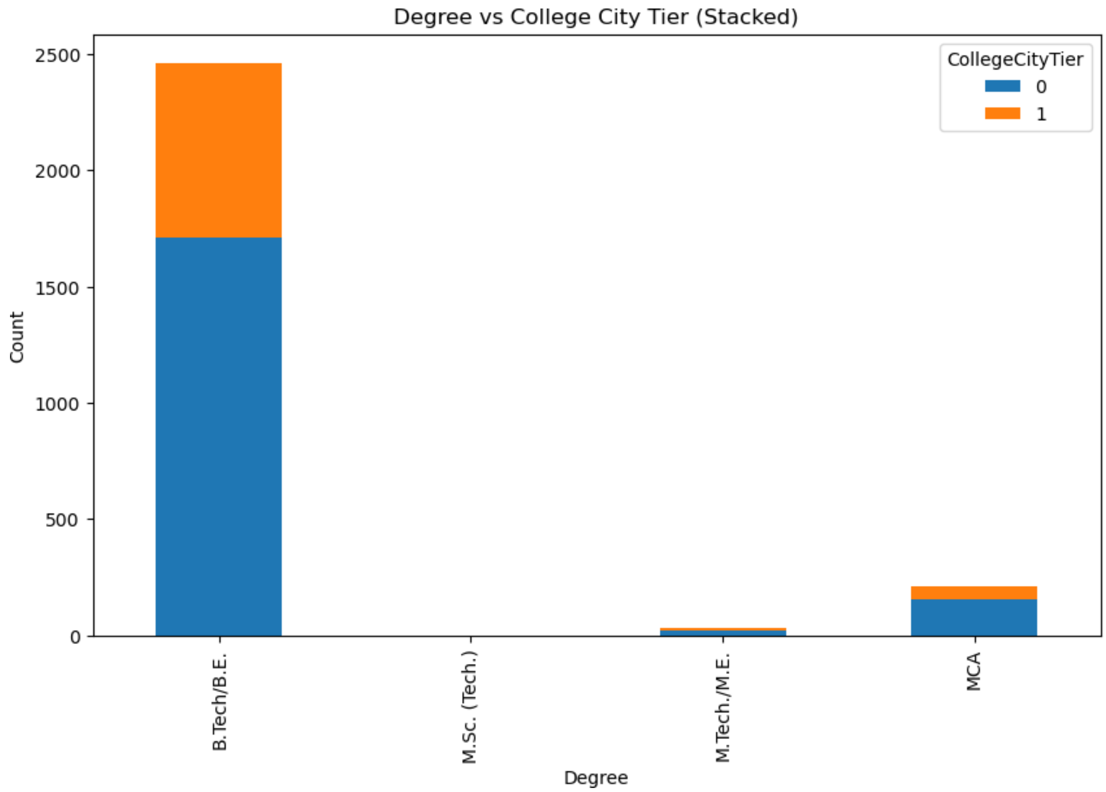
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1296:
UserWarning: 88.4% of the points cannot be placed; you may want to decrease the
size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1296:
UserWarning: 36.9% of the points cannot be placed; you may want to decrease the
size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

[59]:
```
# Barplot for Salary vs Degree
plt.figure(figsize=(10, 6))
sns.barplot(x='Degree', y='Salary', data=df)
plt.title('Average Salary vs Degree')
plt.show()
```



0.13.4 Categorical vs. Categorical Relationships

[61]:
```
# Stacked Bar Plot for Degree and CollegeCityTier
cross_tab = pd.crosstab(df['Degree'], df['CollegeCityTier'])
cross_tab.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Degree vs College City Tier (Stacked)')
plt.xlabel('Degree')
plt.ylabel('Count')
plt.show()
```

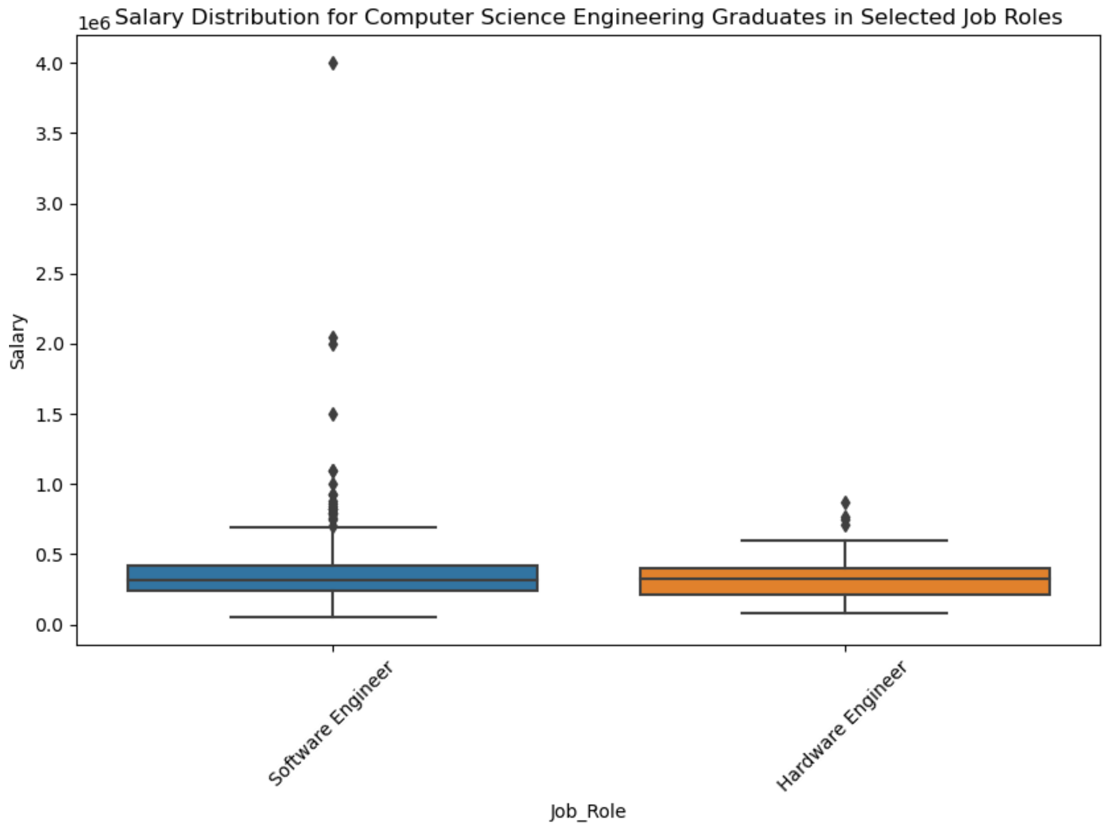


0.14 Step - 5 - Research Questions

[68]:

```
# Filter data for Computer Science Engineering graduates
cse_graduates = df[df['Specialization'] == 'CSE']

# List of job roles to consider
roles_of_interest = ['Programming Analyst', 'Software Engineer', 'Hardware␣
↪Engineer', 'Associate Engineer']

# Filter data to only include these roles
role_data = cse_graduates[cse_graduates['Job_Role'].isin(roles_of_interest)]

# Show the salary distribution for these roles
plt.figure(figsize=(10, 6))
sns.boxplot(x='Job_Role', y='Salary', data=role_data)
plt.title('Salary Distribution for Computer Science Engineering Graduates in␣
↪Selected Job Roles')
plt.xticks(rotation=45)
plt.show()

# You can also check the specific salary range in these roles
```

```python
salary_range = role_data['Salary'].describe()
print(salary_range)
```

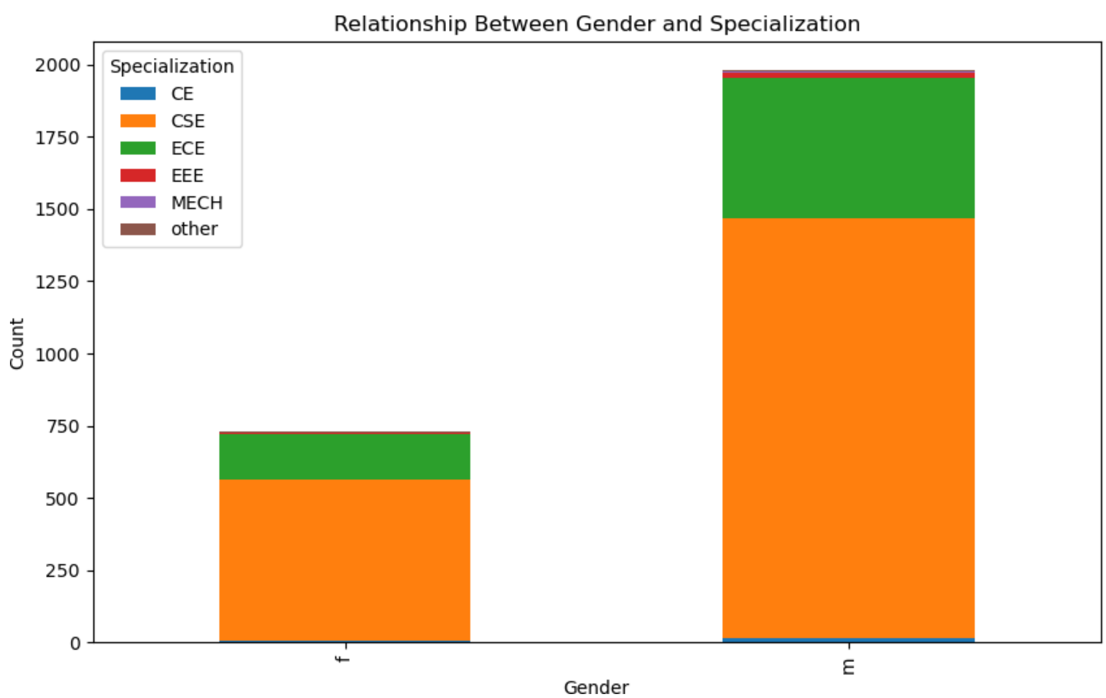


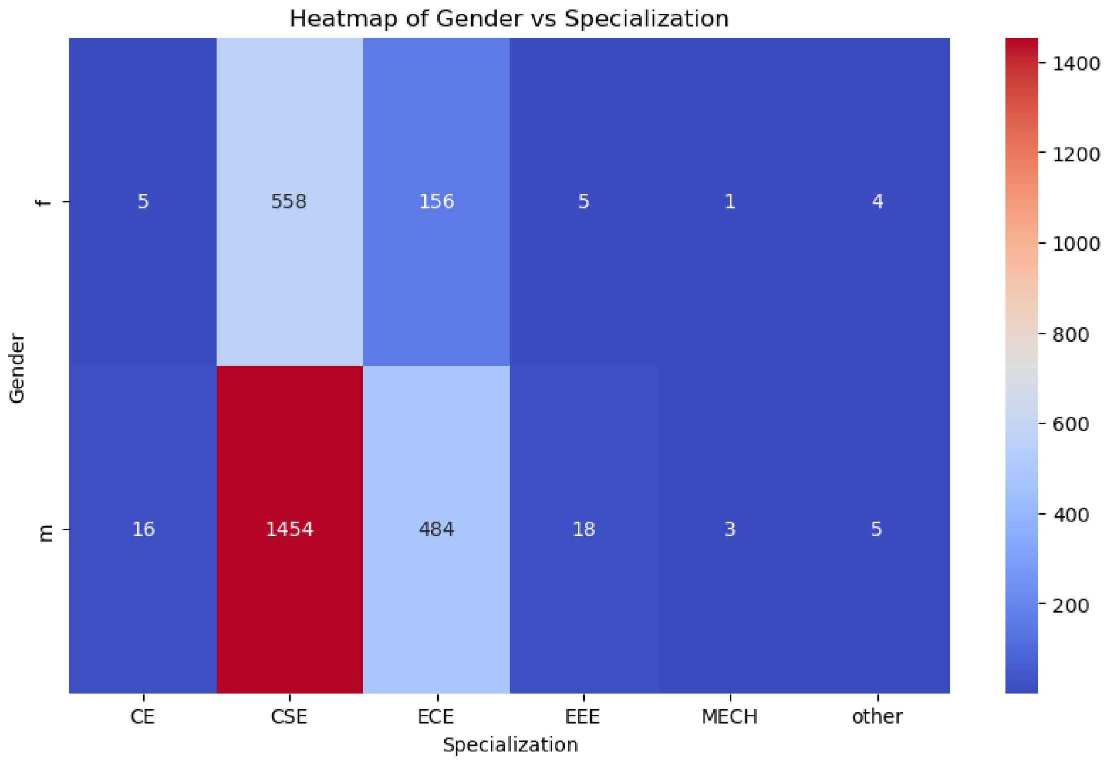

```
count    6.850000e+02
mean     3.510365e+05
std      2.304520e+05
min      5.000000e+04
25%      2.400000e+05
50%      3.200000e+05
75%      4.150000e+05
max      4.000000e+06
Name: Salary, dtype: float64
```

0.14.1 Observations

0.14.2 Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)

[69]:
```
# Create a contingency table to see the relationship between Gender and␣
↪Specialization
gender_specialization = pd.crosstab(df['Gender'], df['Specialization'])

# Plot a stacked bar plot
gender_specialization.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Relationship Between Gender and Specialization')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

# Alternatively, use a heatmap to visualize the distribution
plt.figure(figsize=(10, 6))
sns.heatmap(gender_specialization, annot=True, cmap='coolwarm', fmt='d')
plt.title('Heatmap of Gender vs Specialization')
plt.show()
```





0.14.3 Observation

- The analysis shows that while both genders show a preference for CSE, the male students dominate in terms of number. The other specializations (like ECE, EEE) are also selected by both genders, but CSE remains the most popular overall, especially among male students.

0.15 Step - 6 - Conclusion

- Technical expertise is crucial: The prevalence of Bachelor of Technology/Engineering graduates reflects the high demand for technical skills in the job market.
- Earnings by Role: Managerial and technical positions are the highest-earning roles, emphasizing the value placed on leadership and technical expertise.
- Impact of College Tier: Graduates from Tier-1 colleges consistently earn higher salaries than those from other tiers.
- Gender-Based Salary Differences: While there are some salary disparities between genders, the results warrant further investigation to understand the exact factors contributing to this.
- No Support for Claim on Fresh Graduate Earnings: The data does not support the claim of 2.5-3 lakh earnings for Computer Science graduates, suggesting that salaries may not align with the general assumptions.
- Gender and Specialization Preference: No significant relationship exists between gender and specialization preferences, challenging common assumptions about the correlation.

- Salary Insights:
  - Computer Science & Engineering (CSE) specialization has the highest median salary.
  - On average, females earn 203,648.65, while males earn 194,105.26, with males being slightly under this average.
  - The highest average salary is associated with CSE at 209,166.67 per year.
  - Dominant Roles: The Software Engineer domain employs the largest number of graduates, showcasing the demand for this role in the market.
- Specialization Choices:
  - CSE graduates are the most likely to pursue specialization courses related to their degree.
  - Females tend to opt for Information Technology (IT), while males are more likely to choose Computer Science as their specialization.
  - Average Graduate Salary: Graduates with a B.Tech/B.E. degree generally expect an average salary of 200,000 annually.

```
[ ]:
```