# An introduction to the dataset.

- The dataset pertains to Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO). The study is primarily focused on students within engineering disciplines. It encompasses the employment outcomes of engineering graduates, encompassing dependent variables such as Salary, Job Titles, and Job Locations. Additionally, the dataset incorporates standardized scores from three distinct areas – cognitive skills, technical skills, and personality skills. Demographic features are also included in the dataset. Comprising around 40 independent variables and 4000 data points, these variables exhibit both continuous and categorical nature. Each candidate is uniquely identified within the dataset. -

# Importing the required libraries.

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

# Importing data

In [176…
```
data=pd.read_excel(r"C:\Users\arsha\Downloads\data.xlsx")
```

In [177…
```
data
```

Out[177]:

| | Unnamed: 0 | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | ... | ComputerScience | Mecha |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | train | 203097 | 420000 | 2012-06-01 | present | senior quality engineer | Bangalore | f | 1990-02-19 | 84.30 | ... | -1 | |
| **1** | train | 579905 | 500000 | 2013-09-01 | present | assistant manager | Indore | m | 1989-10-04 | 85.40 | ... | -1 | |
| **2** | train | 810601 | 325000 | 2014-06-01 | present | systems engineer | Chennai | f | 1992-08-03 | 85.00 | ... | -1 | |
| **3** | train | 267447 | 1100000 | 2011-07-01 | present | senior software engineer | Gurgaon | m | 1989-12-05 | 85.60 | ... | -1 | |
| **4** | train | 343523 | 200000 | 2014-03-01 | 2015-03-01 00:00:00 | get | Manesar | m | 1991-02-27 | 78.00 | ... | -1 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **3993** | train | 47916 | 280000 | 2011-10-01 | 2012-10-01 00:00:00 | software engineer | New Delhi | m | 1987-04-15 | 52.09 | ... | -1 | |
| **3994** | train | 752781 | 100000 | 2013-07-01 | 2013-07-01 00:00:00 | technical writer | Hyderabad | f | 1992-08-27 | 90.00 | ... | -1 | |
| **3995** | train | 355888 | 320000 | 2013-07-01 | present | associate software engineer | Bangalore | m | 1991-07-03 | 81.86 | ... | -1 | |
| **3996** | train | 947111 | 200000 | 2014-07-01 | 2015-01-01 00:00:00 | software developer | Asifabadbanglore | f | 1992-03-20 | 78.72 | ... | 438 | |
| **3997** | train | 324966 | 400000 | 2013-02-01 | present | senior systems engineer | Chennai | f | 1991-02-26 | 70.60 | ... | -1 | |

3998 rows × 39 columns

In [178…
```
data.head()
```

Out[178]:

| | Unnamed: 0 | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | ... | ComputerScience | MechanicalEngg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | train | 203097 | 420000 | 2012-06-01 | present | senior quality engineer | Bangalore | f | 1990-02-19 | 84.3 | ... | -1 | -1 |
| **1** | train | 579905 | 500000 | 2013-09-01 | present | assistant manager | Indore | m | 1989-10-04 | 85.4 | ... | -1 | -1 |
| **2** | train | 810601 | 325000 | 2014-06-01 | present | systems engineer | Chennai | f | 1992-08-03 | 85.0 | ... | -1 | -1 |
| **3** | train | 267447 | 1100000 | 2011-07-01 | present | senior software engineer | Gurgaon | m | 1989-12-05 | 85.6 | ... | -1 | -1 |
| **4** | train | 343523 | 200000 | 2014-03-01 | 2015-03-01 00:00:00 | get | Manesar | m | 1991-02-27 | 78.0 | ... | -1 | -1 |

5 rows × 39 columns

In [179…
```
data.shape
```

Out[179]:
```
(3998, 39)
```

In [180…
```
data.describe()
```

Out[180]:

| | ID | Salary | 10percentage | 12graduation | 12percentage | CollegeID | CollegeTier | collegeGPA | CollegeCityID | Colleg |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 3.998000e+03 | 3.998000e+03 | 3998.000000 | 3998.000000 | 3998.000000 | 3998.000000 | 3998.000000 | 3998.000000 | 3998.000000 | 39 |
| **mean** | 6.637945e+05 | 3.076998e+05 | 77.925443 | 2008.087544 | 74.466366 | 5156.851426 | 1.925713 | 71.486171 | 5156.851426 | |
| **std** | 3.632182e+05 | 2.127375e+05 | 9.850162 | 1.653599 | 10.999933 | 4802.261482 | 0.262270 | 8.167338 | 4802.261482 | |
| **min** | 1.124400e+04 | 3.500000e+04 | 43.000000 | 1995.000000 | 40.000000 | 2.000000 | 1.000000 | 6.450000 | 2.000000 | |
| **25%** | 3.342842e+05 | 1.800000e+05 | 71.680000 | 2007.000000 | 66.000000 | 494.000000 | 2.000000 | 66.407500 | 494.000000 | |
| **50%** | 6.396000e+05 | 3.000000e+05 | 79.150000 | 2008.000000 | 74.400000 | 3879.000000 | 2.000000 | 71.720000 | 3879.000000 | |
| **75%** | 9.904800e+05 | 3.700000e+05 | 85.670000 | 2009.000000 | 82.600000 | 8818.000000 | 2.000000 | 76.327500 | 8818.000000 | |
| **max** | 1.298275e+06 | 4.000000e+06 | 97.760000 | 2013.000000 | 98.700000 | 18409.000000 | 2.000000 | 99.930000 | 18409.000000 | |

8 rows × 27 columns

In [181…
```
data.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
```

In [182…

```
# finding unique vlaues of all columns
data.nunique()
```

Out[182]:
```
Unnamed: 0                1
ID                     3998
Salary                  177
DOJ                      81
DOL                      67
Designation             419
JobCity                 339
Gender                    2
DOB                    1872
10percentage            851
10board                 275
12graduation             16
12percentage            801
12board                 340
CollegeID              1350
CollegeTier               2
Degree                    4
Specialization           46
collegeGPA             1282
CollegeCityID          1350
CollegeCityTier           2
CollegeState             26
GraduationYear           11
English                 111
Logical                 107
Quant                   138
Domain                  243
ComputerProgramming      79
ElectronicsAndSemicon    29
ComputerScience          20
MechanicalEngg           42
ElectricalEngg           31
TelecomEngg              26
CivilEngg                23
conscientiousness       141
agreeableness           149
extraversion            154
nueroticism             217
openess_to_experience   142
dtype: int64
```

# Data Cleaning

In [183…
```
# removed a first column which is unnamed
data.drop("Unnamed: 0",axis=1,inplace=True)
data
```

Out[183]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | Mec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 203097 | 420000 | 2012-06-01 | present | senior quality engineer | Bangalore | f | 1990-02-19 | 84.30 | board ofsecondary education,ap | ... | -1 | |
| **1** | 579905 | 500000 | 2013-09-01 | present | assistant manager | Indore | m | 1989-10-04 | 85.40 | cbse | ... | -1 | |
| **2** | 810601 | 325000 | 2014-06-01 | present | systems engineer | Chennai | f | 1992-08-03 | 85.00 | cbse | ... | -1 | |
| **3** | 267447 | 1100000 | 2011-07-01 | present | senior software engineer | Gurgaon | m | 1989-12-05 | 85.60 | cbse | ... | -1 | |
| **4** | 343523 | 200000 | 2014-03-01 | 2015-03-01 00:00:00 | get | Manesar | m | 1991-02-27 | 78.00 | cbse | ... | -1 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **3993** | 47916 | 280000 | 2011-10-01 | 2012-10-01 00:00:00 | software engineer | New Delhi | m | 1987-04-15 | 52.09 | cbse | ... | -1 | |
| **3994** | 752781 | 100000 | 2013-07-01 | 2013-07-01 00:00:00 | technical writer | Hyderabad | f | 1992-08-27 | 90.00 | state board | ... | -1 | |
| **3995** | 355888 | 320000 | 2013-07-01 | present | associate software engineer | Bangalore | m | 1991-07-03 | 81.86 | bse,odisha | ... | -1 | |
| **3996** | 947111 | 200000 | 2014-07-01 | 2015-01-01 00:00:00 | software developer | Asifabadbanglore | f | 1992-03-20 | 78.72 | state board | ... | 438 | |
| **3997** | 324966 | 400000 | 2013-02-01 | present | senior systems engineer | Chennai | f | 1991-02-26 | 70.60 | cbse | ... | -1 | |

3998 rows × 38 columns

*DOL*

DOL

In [184…
```
data['DOL'].unique()
```

Out[184]:
```
array(['present', datetime.datetime(2015, 3, 1, 0, 0),
       datetime.datetime(2015, 5, 1, 0, 0),
       datetime.datetime(2015, 7, 1, 0, 0),
       datetime.datetime(2015, 4, 1, 0, 0),
       datetime.datetime(2014, 10, 1, 0, 0),
       datetime.datetime(2014, 9, 1, 0, 0),
       datetime.datetime(2014, 6, 1, 0, 0),
       datetime.datetime(2012, 9, 1, 0, 0),
       datetime.datetime(2013, 12, 1, 0, 0),
       datetime.datetime(2015, 6, 1, 0, 0),
       datetime.datetime(2013, 10, 1, 0, 0),
       datetime.datetime(2015, 1, 1, 0, 0),
       datetime.datetime(2014, 4, 1, 0, 0),
       datetime.datetime(2013, 6, 1, 0, 0),
       datetime.datetime(2012, 3, 1, 0, 0),
       datetime.datetime(2014, 7, 1, 0, 0),
       datetime.datetime(2013, 2, 1, 0, 0),
       datetime.datetime(2014, 1, 1, 0, 0),
       datetime.datetime(2013, 4, 1, 0, 0),
       datetime.datetime(2012, 7, 1, 0, 0),
       datetime.datetime(2014, 5, 1, 0, 0),
       datetime.datetime(2013, 9, 1, 0, 0),
       datetime.datetime(2015, 2, 1, 0, 0),
       datetime.datetime(2012, 1, 1, 0, 0),
       datetime.datetime(2015, 8, 1, 0, 0),
       datetime.datetime(2014, 8, 1, 0, 0),
       datetime.datetime(2015, 12, 1, 0, 0),
       datetime.datetime(2014, 12, 1, 0, 0),
       datetime.datetime(2012, 5, 1, 0, 0),
       datetime.datetime(2011, 3, 1, 0, 0),
       datetime.datetime(2011, 7, 1, 0, 0),
       datetime.datetime(2014, 2, 1, 0, 0),
       datetime.datetime(2011, 12, 1, 0, 0),
       datetime.datetime(2015, 10, 1, 0, 0),
       datetime.datetime(2014, 11, 1, 0, 0),
       datetime.datetime(2014, 3, 1, 0, 0),
       datetime.datetime(2011, 11, 1, 0, 0),
       datetime.datetime(2013, 5, 1, 0, 0),
       datetime.datetime(2013, 7, 1, 0, 0),
       datetime.datetime(2013, 11, 1, 0, 0),
       datetime.datetime(2011, 1, 1, 0, 0),
       datetime.datetime(2011, 5, 1, 0, 0),
       datetime.datetime(2012, 2, 1, 0, 0),
       datetime.datetime(2012, 11, 1, 0, 0),
       datetime.datetime(2012, 6, 1, 0, 0),
       datetime.datetime(2013, 8, 1, 0, 0),
       datetime.datetime(2005, 3, 1, 0, 0),
       datetime.datetime(2013, 3, 1, 0, 0),
       datetime.datetime(2012, 10, 1, 0, 0),
       datetime.datetime(2011, 2, 1, 0, 0),
       datetime.datetime(2010, 2, 1, 0, 0),
       datetime.datetime(2013, 1, 1, 0, 0),
       datetime.datetime(2011, 6, 1, 0, 0),
       datetime.datetime(2015, 9, 1, 0, 0),
       datetime.datetime(2012, 4, 1, 0, 0),
       datetime.datetime(2012, 8, 1, 0, 0),
       datetime.datetime(2011, 4, 1, 0, 0),
       datetime.datetime(2011, 10, 1, 0, 0),
       datetime.datetime(2015, 11, 1, 0, 0),
       datetime.datetime(2012, 12, 1, 0, 0),
       datetime.datetime(2011, 9, 1, 0, 0),
       datetime.datetime(2010, 8, 1, 0, 0),
       datetime.datetime(2011, 8, 1, 0, 0),
       datetime.datetime(2009, 6, 1, 0, 0),
       datetime.datetime(2008, 3, 1, 0, 0),
       datetime.datetime(2010, 10, 1, 0, 0)], dtype=object)
```

In [185…
```
from datetime import datetime
```

In [186…
```
#replacing 'present' with the current date and time using the datetime.now()
data['DOL'] = data['DOL'].replace(to_replace='present',value=datetime.now())
data['DOL']
```

Out[186]:
```
0      2024-02-22 22:01:18.975275
1      2024-02-22 22:01:18.975275
2      2024-02-22 22:01:18.975275
3      2024-02-22 22:01:18.975275
4      2015-03-01 00:00:00.000000
                  ...
3993   2012-10-01 00:00:00.000000
3994   2013-07-01 00:00:00.000000
3995   2024-02-22 22:01:18.975275
3996   2015-01-01 00:00:00.000000
3997   2024-02-22 22:01:18.975275
Name: DOL, Length: 3998, dtype: datetime64[ns]
```

*Designation*

In [13]:
```
data['Designation'].unique()
```

Out[13]:
```
array(['senior quality engineer', 'assistant manager', 'systems engineer',
       'senior software engineer', 'get', 'system engineer',
       'java software engineer', 'mechanical engineer',
       'electrical engineer', 'project engineer', 'senior php developer',
       'senior systems engineer', 'quality assurance engineer',
       'qa analyst', 'network engineer', 'product development engineer',
       'associate software developer', 'data entry operator',
       'software engineer', 'developer', 'electrical project engineer',
       'programmer analyst', 'systems analyst', 'ase',
       'telecommunication engineer', 'application developer',
       'ios developer', 'executive assistant', 'online marketing manager',
       'documentation specialist', 'associate software engineer',
       'management trainee', 'site manager', 'software developer',
       '.net developer', 'production engineer', 'jr. software engineer',
       'trainee software developer', 'ui developer',
       'assistant system engineer', 'android developer',
       'customer service', 'test engineer', 'java developer', 'engineer',
       'recruitment coordinator', 'technical support engineer',
       'data analyst', 'assistant software engineer', 'faculty',
       'entry level management trainee',
       'customer service representative', 'software test engineer',
       'firmware engineer', 'php developer', 'research associate',
       'research analyst', 'quality engineer', 'programmer',
       'technical support executive', 'business analyst', 'web developer',
       'application engineer', 'project coordinator', 'engineer trainee',
       'sap consultant', 'quality analyst', 'marketing coordinator',
       'system administrator', 'senior engineer',
       'business development managerde', 'network administrator',
       'technical support specialist', 'business development executive',
       'junior software engineer', 'asp.net developer',
       'graduate engineer trainee', 'field engineer',
       'assistant professor', 'trainee software engineer',
       'senior software developer',
       'quality assurance automation engineer', 'design engineer',
       'telecom engineer', 'quality control engineer',
       'hardware engineer', 'hr recruiter', 'sales associate',
       'junior engineer', 'associate engineer', 'maintenance engineer',
       'sales engineer', 'human resources associate',
       'mobile application developer',
       'electronic field service engineer', 'process associate',
       'field service engineer', 'it support specialist',
       'software development engineer', 'business process analyst',
       'operation engineer', 'electrical designer', 'marketing assistant',
       'sales executive', 'admin assistant', 'senior java developer',
       'account executive', 'oracle dba', 'rf engineer',
       'embedded software engineer', 'programmer analyst trainee',
       'technical engineer', 'operations executive', 'trainee engineer',
       'recruiter', 'lecturer', '.net web developer',
       'marketing executive', 'operations assistant', 'associate manager',
       'electrical design engineer', 'systems administrator',
       'client services associate', 'it analyst', 'senior developer',
       'cad designer', 'business technology analyst', 'asst. manager',
       'service engineer', 'executive recruiter', 'planning engineer',
       'associate technical operations', 'web designer',
       'software architect', 'software quality assurance tester',
       'seo trainee', 'process engineer',
       'software quality assurance analyst', 'designer',
       'business systems consultant', 'business development manager',
       'junior research fellow', 'technical recruiter',
       'operations analyst', 'quality assurance test engineer',
       'linux systems administrator', 'software trainee',
       'entry level sales and marketing', 'electrical field engineer',
       'windows systems administrator', 'junior software developer',
       'python developer', 'web application developer',
       'assistant systems engineer', 'javascript developer',
       'operation executive', 'performance engineer', 'technical writer',
       'operations engineer and jetty handling', 'lead engineer',
       'portfolio analyst', 'associate system engineer',
       'mechanical design engineer', 'product engineer',
       'network security engineer', 'operations manager',
       'technical lead', 'operations', 'quality assurance tester',
       'automation engineer', 'data scientist', 'quality associate',
       'manual tester', 'sr. engineer', 'embedded engineer',
       'service and sales engineer', 'telecom support engineer',
       'engineer- customer support', 'cloud engineer', 'branch manager',
       'business analyst consultant', 'technology lead',
       'software trainee engineer', 'dcs engineer', 'junior manager',
       'ux designer', 'clerical', 'hr generalist',
       'database administrator', 'senior design engineer', 'seo',
       'assistant engineer', 'marketing analyst', 'it executive',
       'salesforce developer', 'software tester', 'sql dba',
       'junior engineer product support', 'manager',
       'senior business analyst', 'c# developer',
       'implementation engineer', 'executive hr', 'executive engineer',
```

```
'sharepoint developer', 'system analyst',
'sales management trainee', 'senior project engineer',
'it recruiter', 'software engineer analyst',
'desktop support technician', 'continuous improvement engineer',
'process advisor', 'etl developer', 'sales and service engineer',
'project manager', 'training specialist', 'product manager',
'staffing recruiter', 'assistant programmer', 'quality controller',
'mis executive', 'game developer', 'digital marketing specialist',
'principal software engineer', 'software devloper',
'senior mechanical engineer', 'technical operations analyst',
'service coordinator', 'testing engineer', 'technical assistant',
'sap abap consultant', 'seo engineer', 'project assistant',
'talent acquisition specialist', 'sales account manager',
'software engineer trainee', 'customer service manager',
'help desk analyst', 'general manager', 'engineering manager',
'senior network engineer',
'field based employee relations manager', 'phone banking officer',
'support engineer', 'associate test engineer',
'technology analyst', 'network support engineer',
'it business analyst', 'junior system analyst',
'senior .net developer', 'secretary', 'research engineer',
'quality assurance auditor', 'process executive',
'lecturer & electrical maintenance', 'office coordinator',
'hr manager', 'html developer', 'sales support',
'front end web developer', 'administrative support',
'territory sales manager', 'project administrator',
'environmental engineer', 'web designer and seo',
'information security analyst',
'field business development associate', 'operational executive',
'administrative coordinator', 'senior risk consultant',
'desktop support engineer', 'cad drafter', 'noc engineer',
'industrial engineer', 'it engineer', 'human resources intern',
'senior quality assurance engineer', 'clerical assistant',
'software enginner', 'quality assurance',
'delivery software engineer', 'graphic designer',
'sales development manager', 'visiting faculty',
'business intelligence analyst', 'team lead',
'operational excellence manager', 'sales & service engineer',
'web intern', 'full stack developer', 'database developer',
'sr. database engineer', 'graduate apprentice trainee',
'software engineer associate', 'technical analyst',
'executive engg', 'it technician', 'business system analyst',
'process control engineer', 'technical consultant',
'business office manager', 'quality control inspector',
'product design engineer', 'manufacturing engineer',
'seo executive', 'sap analyst', 'software engineere',
'financial service consultant', 'co faculty', 'software analyst',
'desktop support analyst', 'graduate engineer',
'engineering technician', 'it assistant', 'marketing manager',
'human resource assistant', 'hr assistant', 'product developer',
'customer support engineer',
'quality control inspection technician', 'gis/cad engineer',
'senior web developer', 'sql developer', 'research staff member',
'sap abap associate consultant', 'associate qa',
'corporate recruiter', 'project management officer',
'business systems analyst', 'software programmer',
'help desk technician', 'sales manager', 'catalog associate',
'assistant store manager', 'software engg', 'it developer',
'apprentice', 'business consultant', 'controls engineer',
'ruby on rails developer', 'risk consultant', 'account manager',
'professor', 'assistant administrator', 'civil engineer',
'educator', 'service manager', 'teradata dba',
'full-time loss prevention associate', 'junior recruiter',
'associate developer', 'assistant electrical engineer',
'shift engineer', 'dotnet developer', 'rf/dt engineer',
'human resources analyst', 'software test engineerte',
'junior .net developer', 'java trainee', 'maintenance supervisor',
'r&d engineer', 'front end developer', 'engineer-hws',
'operations engineer', 'senior research fellow',
'web designer and joomla administrator',
'enterprise solutions developer',
'information technology specialist', 'site engineer',
'graduate trainee engineer', 'quality assurance analyst',
'cnc programmer', 'financial analyst', 'system engineer trainee',
'sap mm consultant', 'assistant system engineer trainee',
'qa trainee', 'teradata developer', 'hr executive',
'senior programmer', 'software test engineer (etl)',
'associate software engg', 'supply chain analyst', 'sales trainer',
'software executive', 'team leader',
'assistant system engineer - trainee', 'seo analyst',
'risk investigator', 'executive administrative assistant',
'program manager', 'r & d', 'sap functional consultant',
'website developer/tester', 'software designer',
'sales coordinator', 'qa engineer', 'aircraft technician',
'customer care executive', 'senior test engineer',
'program analyst trainee', 'electrical controls engineer',
'trainee decision scientist', 'editor', 'bss engineer', 'dba',
'software eng', 'computer faculty', 'recruitment associate',
'logistics executive', 'quality consultant',
```

```
       'senior sales executive', 'db2 dba', 'test technician',
       'it operations associate', 'software engineering associate',
       'research scientist', 'jr. software developer'], dtype=object)
```

In [14]:
```
data['Designation']
```

Out[14]:
```
0           senior quality engineer
1                 assistant manager
2                  systems engineer
3           senior software engineer
4                               get
                    ...
3993               software engineer
3994                technical writer
3995    associate software engineer
3996              software developer
3997         senior systems engineer
Name: Designation, Length: 3998, dtype: object
```

In [15]:
```
data[data.Designation=="get"]
```

Out[15]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | Mecha |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4** | 343523 | 200000 | 2014-03-01 | 2015-03-01 00:00:00.000000 | get | Manesar | m | 1991-02-27 | 78.00 | cbse | ... | -1 | |
| **332** | 597966 | 180000 | 2013-08-01 | 2014-03-01 00:00:00.000000 | get | -1 | m | 1990-08-02 | 91.80 | cbse | ... | -1 | |
| **350** | 38162 | 340000 | 2010-07-01 | 2011-05-01 00:00:00.000000 | get | Faridabad | m | 1988-08-13 | 67.67 | up board | ... | -1 | |
| **1717** | 330551 | 145000 | 2012-10-01 | 2013-01-01 00:00:00.000000 | get | Hyderabad | m | 1991-07-29 | 80.00 | ssc | ... | -1 | |
| **1897** | 1064862 | 175000 | 2014-05-01 | 2024-02-22 17:48:19.511385 | get | Hyderabad | m | 1991-02-04 | 87.00 | state board | ... | -1 | |
| **2140** | 796296 | 600000 | 2014-06-01 | 2024-02-22 17:48:19.511385 | get | Indore | m | 1992-01-05 | 91.20 | cbse | ... | -1 | |
| **2318** | 1094242 | 220000 | 2014-07-01 | 2024-02-22 17:48:19.511385 | get | Lucknow | m | 1992-02-06 | 80.20 | cbse | ... | -1 | |
| **2443** | 1259589 | 110000 | 2014-10-01 | 2015-04-01 00:00:00.000000 | get | kharagpur | m | 1993-03-18 | 80.00 | icse | ... | -1 | |
| **2666** | 110817 | 200000 | 2010-03-01 | 2011-03-01 00:00:00.000000 | get | Hyderabad | m | 1989-05-15 | 80.00 | ssc | ... | -1 | |
| **2757** | 1083682 | 350000 | 2015-01-01 | 2024-02-22 17:48:19.511385 | get | Nashik | m | 1994-01-17 | 90.60 | cbse | ... | -1 | |
| **3045** | 35694 | 180000 | 2013-01-01 | 2013-12-01 00:00:00.000000 | get | Sahibabad | m | 1989-06-18 | 84.80 | cbse | ... | -1 | |
| **3126** | 87319 | 1210000 | 2010-10-01 | 2011-09-01 00:00:00.000000 | get | Bhopal | m | 1986-10-27 | 56.40 | 0 | ... | -1 | |
| **3594** | 967009 | 280000 | 2014-04-01 | 2024-02-22 17:48:19.511385 | get | MEERUT | m | 1991-10-15 | 84.16 | cbse | ... | -1 | |
| **3980** | 197796 | 150000 | 2011-07-01 | 2012-07-01 00:00:00.000000 | get | haryana | m | 1986-08-05 | 84.00 | cbse | ... | -1 | |

14 rows × 38 columns

In [16]:
```
data[data.Designation=="ase"]
```

Out[16]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | Mecha |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **27** | 810458 | 300000 | 2014-09-01 | 2015-01-01 00:00:00.000000 | ase | Bangalore | m | 1993-02-01 | 76.80 | state board | ... | -1 | |
| **2792** | 503183 | 360000 | 2013-07-01 | 2024-02-22 17:48:19.511385 | ase | Pune | m | 1991-08-10 | 86.33 | state board | ... | -1 | |
| **3839** | 644828 | 360000 | 2013-09-01 | 2024-02-22 17:48:19.511385 | ase | bhubaneswar | f | 1992-03-11 | 90.80 | state board | ... | -1 | |

3 rows × 38 columns

In [17]:
```
data[data.Designation=="qa analyst"]
```

Out[17]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | Mech |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **13** | 431800 | 200000 | 2014-01-01 | 2024-02-22 17:48:19.511385 | qa analyst | Gurgaon | m | 1990-10-23 | 60.80 | cbse | ... | -1 | |
| **49** | 305559 | 270000 | 2012-05-01 | 2014-07-01 00:00:00.000000 | qa analyst | Bangalore | m | 1989-03-27 | 81.92 | state board | ... | -1 | |
| **103** | 682615 | 300000 | 2014-09-01 | 2024-02-22 17:48:19.511385 | qa analyst | -1 | f | 1991-03-06 | 62.00 | cbse | ... | -1 | |
| **558** | 597624 | 185000 | 2015-03-01 | 2015-04-01 00:00:00.000000 | qa analyst | Noida | f | 1990-06-09 | 71.30 | cbse | ... | -1 | |
| **602** | 1231995 | 200000 | 2013-10-01 | 2024-02-22 17:48:19.511385 | qa analyst | Hyderabad | f | 1992-06-19 | 87.00 | state board | ... | -1 | |
| **612** | 216484 | 500000 | 2013-04-01 | 2024-02-22 17:48:19.511385 | qa analyst | -1 | m | 1989-05-29 | 82.00 | icse | ... | -1 | |
| **649** | 1018051 | 145000 | 2014-01-01 | 2015-04-01 00:00:00.000000 | qa analyst | Indore | m | 1992-03-24 | 63.00 | cbse | ... | 500 | |
| **865** | 1189528 | 200000 | 2014-12-01 | 2014-04-01 00:00:00.000000 | qa analyst | Hyderabad | m | 1992-11-06 | 92.83 | state board | ... | -1 | |
| **1038** | 47767 | 800000 | 2010-03-01 | 2024-02-22 17:48:19.511385 | qa analyst | New Delhi | m | 1987-12-07 | 70.20 | 0 | ... | -1 | |
| **1895** | 421635 | 120000 | 2013-01-01 | 2014-06-01 00:00:00.000000 | qa analyst | Jaspur | m | 1991-07-25 | 67.10 | uttaranchal shiksha avam pariksha parishad | ... | -1 | |
| **1998** | 628645 | 215000 | 2013-07-01 | 2024-02-22 17:48:19.511385 | qa analyst | Chennai | f | 1991-09-25 | 77.70 | state board | ... | -1 | |
| **2065** | 1192794 | 265000 | 2014-09-01 | 2015-04-01 00:00:00.000000 | qa analyst | Pune | f | 1992-08-06 | 85.83 | state board | ... | 284 | |
| **2091** | 914034 | 480000 | 2014-03-01 | 2024-02-22 17:48:19.511385 | qa analyst | Noida | m | 1991-09-28 | 72.00 | cbse | ... | -1 | |
| **2095** | 1099501 | 110000 | 2014-10-01 | 2015-04-01 00:00:00.000000 | qa analyst | Gurgaon | m | 1993-07-10 | 54.00 | state board | ... | 315 | |
| **2171** | 729525 | 190000 | 2013-11-01 | 2014-04-01 00:00:00.000000 | qa analyst | Hyderabad | m | 1991-04-22 | 82.00 | state board | ... | -1 | |
| **2200** | 457565 | 500000 | 2013-01-01 | 2015-04-01 00:00:00.000000 | qa analyst | -1 | f | 1990-01-19 | 94.00 | 0 | ... | -1 | |
| **2540** | 266098 | 310000 | 2012-05-01 | 2024-02-22 17:48:19.511385 | qa analyst | Hyderabad | f | 1990-06-02 | 76.80 | ssc | ... | -1 | |
| **2557** | 1219113 | 340000 | 2014-09-01 | 2024-02-22 17:48:19.511385 | qa analyst | Kolkata | m | 1992-01-05 | 73.37 | state board | ... | 315 | |
| **2685** | 1156350 | 100000 | 2013-09-01 | 2014-06-01 00:00:00.000000 | qa analyst | Jaipur | f | 1992-10-23 | 78.30 | cbse | ... | -1 | |
| **2713** | 361994 | 330000 | 2013-10-01 | 2024-02-22 17:48:19.511385 | qa analyst | Noida | f | 1990-06-03 | 65.00 | cbse | ... | -1 | |
| **2716** | 1077374 | 140000 | 2013-12-01 | 2014-03-01 00:00:00.000000 | qa analyst | Bangalore | m | 1987-12-31 | 73.00 | cbse | ... | 376 | |
| **2818** | 1271694 | 120000 | 2015-03-01 | 2024-02-22 17:48:19.511385 | qa analyst | Noida | m | 1991-06-22 | 76.00 | cbse | ... | 469 | |
| **2826** | 944290 | 200000 | 2014-01-01 | 2015-04-01 00:00:00.000000 | qa analyst | Chennai | f | 1991-07-23 | 83.00 | state board | ... | 346 | |
| **2866** | 982556 | 350000 | 2014-07-01 | 2024-02-22 17:48:19.511385 | qa analyst | Noida | m | 1992-03-24 | 88.00 | cbse | ... | 346 | |
| **2991** | 274453 | 175000 | 2011-06-01 | 2012-08-01 00:00:00.000000 | qa analyst | Hyderabad | m | 1990-09-01 | 88.50 | ssc | ... | -1 | |
| **3248** | 914312 | 370000 | 2014-05-01 | 2024-02-22 17:48:19.511385 | qa analyst | Gurgaon | m | 1992-02-27 | 79.00 | icse | ... | -1 | |
| **3599** | 305011 | 240000 | 2011-08-01 | 2012-06-01 00:00:00.000000 | qa analyst | Yamuna Nagar | m | 1989-07-24 | 60.00 | hbse | ... | -1 | |
| **3707** | 283980 | 250000 | 2012-08-01 | 2014-06-01 00:00:00.000000 | qa analyst | Gurgaon | m | 1989-07-22 | 74.00 | cbse | ... | -1 | |
| **3772** | 998012 | 190000 | 2014-02-01 | 2014-07-01 00:00:00.000000 | qa analyst | Hyderabad | f | 1991-01-26 | 86.50 | state board | ... | 223 | |

29 rows × 38 columns

In [18]:

```
# as there are many short forms in designation column converting them into their orignal form

data['Designation']=data['Designation'].replace(to_replace='get', value='graduate engineer trainee')
data['Designation']=data['Designation'].replace(to_replace='software eng', value='software engineer')
data['Designation']=data['Designation'].replace(to_replace='associate software engg', value='associate software
data['Designation']=data['Designation'].replace(to_replace='qa', value='quality assurance')
data['Designation']=data['Designation'].replace(to_replace='seo', value='search engine optimization')
```

```
data['Designation']=data['Designation'].replace(to_replace='ase', value='automotive service excellence')
data['Designation']=data['Designation'].replace(to_replace='systems engineer', value='system engineer')
data['Designation']=data['Designation'].replace(to_replace='dotnet developer', value='.netdeveloper')
data['Designation']=data['Designation'].replace(to_replace='dotnet developer', value='.netdeveloper')
data['Designation']=data['Designation'].replace(to_replace='programmer analyst trainee', value='programmer anal
```

In [19]:
```
data[data.Designation=="seo"]
```

Out[19]:

| ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | MechanicalEngg | ElectricalEngg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

0 rows × 38 columns

In [20]:
```
data[data.Designation=="systems engineer"]
```

Out[20]:

| ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | MechanicalEngg | ElectricalEngg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

0 rows × 38 columns

In [21]:
```
data[data.Designation=="system engineer"]
```

Out[21]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | Mecha |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 810601 | 325000 | 2014-06-01 | 2024-02-22 17:48:19.511385 | system engineer | Chennai | f | 1992-08-03 | 85.00 | cbse | ... | -1 | |
| **5** | 1027655 | 300000 | 2014-06-01 | 2024-02-22 17:48:19.511385 | system engineer | Hyderabad | m | 1992-07-02 | 89.92 | state board | ... | 407 | |
| **30** | 192703 | 530000 | 2011-12-01 | 2024-02-22 17:48:19.511385 | system engineer | Hyderabad | m | 1989-10-04 | 84.00 | ssc | ... | -1 | |
| **65** | 1044453 | 310000 | 2014-02-01 | 2014-05-01 00:00:00.000000 | system engineer | Mysore | m | 1990-12-02 | 89.37 | state board | ... | -1 | |
| **66** | 125529 | 455000 | 2011-01-01 | 2013-09-01 00:00:00.000000 | system engineer | Bangalore | m | 1989-08-29 | 90.00 | state board of secondary education, andhra pra... | ... | -1 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **3866** | 63284 | 530000 | 2011-03-01 | 2024-02-22 17:48:19.511385 | system engineer | Kolkata | m | 1987-06-03 | 78.00 | state board | ... | -1 | |
| **3932** | 531480 | 365000 | 2013-10-01 | 2024-02-22 17:48:19.511385 | system engineer | Hyderabad | m | 1991-12-30 | 91.00 | board of ssc education andhra pradesh | ... | -1 | |
| **3933** | 116358 | 490000 | 2011-02-01 | 2024-02-22 17:48:19.511385 | system engineer | Chennai | m | 1989-01-21 | 87.00 | ssc | ... | -1 | |
| **3966** | 224873 | 450000 | 2012-10-01 | 2024-02-22 17:48:19.511385 | system engineer | mumbai | m | 1990-03-15 | 89.60 | karnataka state board | ... | -1 | |
| **3985** | 358393 | 430000 | 2013-03-01 | 2024-02-22 17:48:19.511385 | system engineer | Gurgaon | f | 1990-05-14 | 90.40 | cbse | ... | -1 | |

323 rows × 38 columns

In [22]:
```
data['Designation']
```

Out[22]:
```
0           senior quality engineer
1                 assistant manager
2                   system engineer
3            senior software engineer
4           graduate engineer trainee
                     ...
3993                software engineer
3994                 technical writer
3995      associate software engineer
3996                software developer
3997          senior systems engineer
Name: Designation, Length: 3998, dtype: object
```

*JobCity*

In [254…
```
data['JobCity'].unique()
```

Out[254]:
```
array(['Bangalore', 'Indore', 'Chennai', 'Gurgaon', 'Manesar',
       'Hyderabad', 'Banglore', 'Noida', 'Kolkata', 'Pune', -1, 'mohali',
       'Jhansi', 'Delhi', 'Hyderabad ', 'Bangalore ', 'noida', 'delhi',
       'Bhubaneswar', 'Navi Mumbai', 'Mumbai', 'New Delhi', 'Mangalore',
       'Rewari', 'Gaziabaad', 'Bhiwadi', 'Mysore', 'Rajkot',
       'Greater Noida', 'Jaipur', 'noida ', 'HYDERABAD', 'mysore',
       'THANE', 'Maharajganj', 'Thiruvananthapuram', 'Punchkula',
       'Bhubaneshwar', 'Pune ', 'coimbatore', 'Dhanbad', 'Lucknow',
       'Trivandrum', 'kolkata', 'mumbai', 'Gandhi Nagar', 'Una',
       'Daman and Diu', 'chennai', 'GURGOAN', 'vsakhapttnam', 'pune',
       'Nagpur', 'Bhagalpur', 'new delhi - jaisalmer', 'Coimbatore',
       'Ahmedabad', 'Kochi/Cochin', 'Bankura', 'Bengaluru', 'Mysore ',
       'Kanpur ', 'jaipur', 'Gurgaon ', 'bangalore', 'CHENNAI',
       'Vijayawada', 'Kochi', 'Beawar', 'Alwar', 'NOIDA', 'Greater noida',
       'Siliguri ', 'raipur', 'gurgaon', 'Bhopal', 'Faridabad', 'Jodhpur',
       'udaipur', 'Muzaffarpur', 'Kolkata`', 'Bulandshahar', 'Haridwar',
       'Raigarh', 'Visakhapatnam', 'Jabalpur', 'hyderabad', 'Unnao',
       'KOLKATA', 'Thane', 'Aurangabad', 'Belgaum', 'gurgoan', 'Dehradun',
       'Rudrapur', 'Jamshedpur', 'vizag', 'Nouda', 'Dharamshala',
       'Banagalore', 'Hissar', 'Ranchi', 'BANGALORE', 'Madurai', 'Gurga',
       'Chandigarh', 'Australia', ' Chennai', 'CHEYYAR', 'Mumbai ',
       'sonepat', 'Ghaziabad', 'Pantnagar', 'Siliguri', 'mumbai ',
       'Jagdalpur', 'Chennai ', 'angul', 'Baroda', ' ariyalur', 'Jowai',
       'Kochi/Cochin, Chennai and Coimbatore', 'bhubaneswar', 'Neemrana',
       'VIZAG', 'Tirupathi', 'Lucknow ', 'Ahmedabad ', 'Bhubneshwar',
       'Noida ', 'pune ', 'Calicut', 'Gandhinagar', 'LUCKNOW', 'Dubai',
       'bengaluru', 'MUMBAI', 'Ahmednagar', 'Nashik', 'New delhi',
       'Bellary', 'Ludhiana', 'New Delhi ', 'Muzaffarnagar', 'BHOPAL',
       'Gurgoan', 'Gagret', 'Indirapuram, Ghaziabad', 'Gwalior',
       'new delhi', 'TRIVANDRUM', 'Chennai & Mumbai', 'Rajasthan',
       'Sonipat', 'Bareli', 'Kanpur', 'Hospete', 'Miryalaguda', ' mumbai',
       'Dharuhera', 'lucknow', 'meerut', 'dehradun', 'Ganjam', 'Hubli',
       'bangalore ', 'NAVI MUMBAI', 'ncr', 'Agra', 'Trichy',
       'kudankulam ,tarapur', 'Ongole', 'Sambalpur', 'Pondicherry',
       'Bundi', 'SADULPUR,RAJGARH,DISTT-CHURU,RAJASTHAN', 'AM', 'Bikaner',
       'Vadodara', 'BAngalore', 'india', 'Asansol', 'Tirunelvelli',
       'Ernakulam', 'DELHI', 'Bilaspur', 'Chandrapur', 'Nanded',
       'Dharmapuri', 'Vandavasi', 'Rohtak', 'trivandrum', 'Nagpur ',
       'Udaipur', 'Patna', 'banglore', 'indore', 'Salem', 'Nasikcity',
       'Gandhinagar ', 'Technopark, Trivandrum', 'Bharuch', 'Tornagallu',
       'Raipur', 'Kolkata ', 'Jaspur', 'Burdwan', 'Bhubaneswar ',
       'Shimla', 'ahmedabad', 'Gajiabaad', 'Jammu', 'Shahdol',
       'Muvattupuzha', 'Al Jubail,Saudi Arabia', 'Kalmar, Sweden',
       'Secunderabad', 'A-64,sec-64,noida', 'Ratnagiri', 'Jhajjar',
       'Gulbarga', 'hyderabad(bhadurpally)', 'Nalagarh', 'Chandigarh ',
       'Jaipur ', 'Jeddah Saudi Arabia', ' Delhi', 'PATNA', 'SHAHDOL',
       'Chennai, Bangalore', 'Bhopal ', 'Jamnagar', 'PUNE', 'Tirupati',
       'Gonda', 'jamnagar', 'chennai ', 'orissa', 'kharagpur',
       'Trivandrum ', 'Navi Mumbai , Hyderabad', 'Joshimath',
       'chandigarh', 'Bathinda', 'Johannesburg', 'kala amb ', 'Karnal',
       'LONDON', 'Kota', 'Panchkula', 'Baddi HP', 'Nagari',
       'Mettur, Tamil Nadu ', 'Durgapur', 'pondi', 'Surat', 'Kurnool',
       'kolhapur', 'Madurai ', 'GREATER NOIDA', 'Bhilai', ' Pune',
       'hderabad', 'KOTA', 'thane', 'Vizag', 'Bahadurgarh',
       'Rayagada, Odisha', 'kakinada', 'GURGAON', 'Varanasi', 'punr',
       'Nellore', 'patna', 'Meerut', 'hyderabad ', 'Sahibabad', 'Howrah',
       'BHUBANESWAR', 'Trichur', 'Ambala', 'Khopoli', 'keral', 'Roorkee',
       'Greater NOIDA', 'Navi mumbai', 'ghaziabad', 'Allahabad',
       'Delhi/NCR', 'Panchkula ', 'Ranchi ', 'Jalandhar', 'manesar',
       'vapi', 'PILANI', 'muzzafarpur', 'RAS AL KHAIMAH', 'bihar',
       'singaruli', 'KANPUR', 'Banglore ', 'pondy', 'Mohali', 'Phagwara',
       ' Mumbai', ' bangalore', 'GURAGAON', 'Baripada', 'MEERUT',
       'Yamuna Nagar', 'shahibabad', 'sampla', 'Guwahati', 'Rourkela',
       'Banaglore', 'Vellore', 'Dausa', 'latur (Maharashtra )',
       'NEW DELHI', 'kanpur', 'Mainpuri', 'karnal', 'Dammam', 'Haldia',
       'sambalpur', 'RAE BARELI', 'ranchi', 'jAipur', 'BANGLORE',
       'Patiala', 'Gorakhpur', 'new dehli', 'BANGALORE ', 'Ambala City',
       'Karad', 'Rajpura', 'Pilani', 'haryana', 'Asifabadbanglore'],
      dtype=object)
```

In [255…
```
data['JobCity'] = data['JobCity'].astype(str)
```

In [256…
```
# given states name different forms replacing them with acutal name

data['JobCity'] = data['JobCity'].replace(to_replace=['mumbai', 'Mumbai ', 'mumbai ', 'MUMBAI', ' mumbai', ' Mu
data['JobCity'] = data['JobCity'].replace(to_replace=['KOTA', 'Gajiabaad', 'Banglore', 'jaipur', 'Lucknow ', 'L
data['JobCity'] = data['JobCity'].replace(to_replace=['Kanpur ', 'KANPUR', 'kanpur'], value='Various Cities')
data['JobCity'] = data['JobCity'].replace(to_replace=['Bhubaneshwar', 'bhubaneswar', 'Bhubneshwar', 'Bhubaneswa
data['JobCity'] = data['JobCity'].replace(to_replace=['Asifabadbanglore', 'Bangalore ', 'Bengaluru', 'bangalore
data['JobCity'] = data['JobCity'].replace(to_replace='chandigarh', value='Chandigarh')
data['JobCity'] = data['JobCity'].replace(to_replace=['delhi', 'DELHI', ' Delhi', 'Delhi/NCR', 'New Delhi', 'ne
data['JobCity'] = data['JobCity'].replace(to_replace='dehradun', value='Dehradun')
data['JobCity'] = data['JobCity'].replace(to_replace=['GURGOAN', 'Gurgaon ', 'gurgaon', 'gurgoan', 'Gurgoan', '
data['JobCity'] = data['JobCity'].replace(to_replace=['Hyderabad ', 'HYDERABAD', 'hyderabad', 'hyderabad(bhadur
data['JobCity'] = data['JobCity'].replace(to_replace=['jamnagar', 'Bhopal', 'BHOPAL', 'Bhopal '], value='Variou
data['JobCity'] = data['JobCity'].replace(to_replace=['vsakhapttnam', 'vizag', 'vizag', 'Vizag'], value='Visakh
data['JobCity'] = data['JobCity'].replace(to_replace='raipur', value='Raipur')
data['JobCity'] = data['JobCity'].replace(to_replace=['Ranchi ', 'ranchi'], value='Ranchi')
```

```
data['JobCity'] = data['JobCity'].replace(to_replace=['TRIVANDRUM', 'trivandrum', 'Trivandrum '], value='Trivan
data['JobCity'] = data['JobCity'].replace(to_replace=['Gandhinagar', 'Gandhinagar '], value='Gandhi Nagar')
data['JobCity'] = data['JobCity'].replace(to_replace='chennai', value='Chennai')
data['JobCity'] = data['JobCity'].replace(to_replace=['Kochi/Cochin', 'Kochi/Cochin, Chennai and Coimbatore'],
data['JobCity'] = data['JobCity'].replace(to_replace=['Pune ', 'pune', 'pune ', 'PUNE', ' Pune', 'punr'], value
data['JobCity'] = data['JobCity'].replace(to_replace=['Punchkula', 'Panchkula '], value='Panchkula')
data['JobCity'] = data['JobCity'].replace(to_replace='PATNA', value='Patna')
data['JobCity'] = data['JobCity'].replace(to_replace='manesar', value='Manesar')
data['JobCity'] = data['JobCity'].replace(to_replace=['meerut', 'MEERUT'], value='Meerut')
data['JobCity'] = data['JobCity'].replace(to_replace=['Greater Noida', 'Greater noida', 'GREATER NOIDA', 'Great
data['JobCity'] = data['JobCity'].replace(to_replace=['Lucknow ', 'LUCKNOW', 'lucknow'], value='Lucknow')
data['JobCity'] = data['JobCity'].replace(to_replace='mysore', value='Mysore')
data['JobCity'] = data['JobCity'].replace(to_replace='mohali', value='Mohali')
data['JobCity'] = data['JobCity'].replace(to_replace='Ambala City', value='Ambala')
data['JobCity'] = data['JobCity'].replace(to_replace=['Gandhinagar', 'Gandhinagar '], value='Gandhi Nagar')
data['JobCity'] = data['JobCity'].replace(to_replace='jaipur', value='Jaipur')
data['JobCity']=  data['JobCity'].replace(to_replace=['kolkata','Kolkata`','KOLKATA','Kolkata '], value='Kolkat
data['JobCity']=  data['JobCity'].replace(to_replace=['kolkata','Kolkata`','KOLKATA','Kolkata '], value='Kolkat
```

In [257…
```
# filling -1 with nan and o with mode
data['JobCity'].replace({'-1': np.nan}, inplace=True)
data['JobCity'].fillna(data['JobCity'].mode()[0], inplace=True)
```

In [258…
```
data["JobCity"]
```

Out[258]:
```
0       Bangalore
1          Indore
2         Chennai
3         Gurgaon
4         Manesar
           ...
3993        Delhi
3994    Hyderabad
3995    Bangalore
3996    Bangalore
3997      Chennai
Name: JobCity, Length: 3998, dtype: object
```

*Gender*

In [26]:
```
data["Gender"].unique()
```

Out[26]:
```
array(['f', 'm'], dtype=object)
```

*DOB*

In [27]:
```
data["DOB"]
```

Out[27]:
```
0      1990-02-19
1      1989-10-04
2      1992-08-03
3      1989-12-05
4      1991-02-27
          ...
3993   1987-04-15
3994   1992-08-27
3995   1991-07-03
3996   1992-03-20
3997   1991-02-26
Name: DOB, Length: 3998, dtype: datetime64[ns]
```

*10board*

In [28]:
```
data['10board'].unique()
```

Out[28]:
```
array(['board ofsecondary education,ap', 'cbse', 'state board',
       'mp board bhopal', 'icse',
       'karnataka secondary school of examination', 'up',
       'karnataka state education examination board', 'ssc',
       'kerala state technical education', 0, 'bseb',
       'state board of secondary education, andhra pradesh',
       'matriculation', 'gujarat state board', 'karnataka state board',
       'wbbse', 'maharashtra state board', 'icse board', 'up board',
       'board of secondary education(bse) orissa',
       'little jacky matric higher secondary school',
       'uttar pradesh board', 'bsc,orissa', 'mp board', 'upboard',
       'matriculation board', 'j & k bord', 'rbse',
       'central board of secondary education', 'pseb', 'jkbose',
       'haryana board of school education,(hbse)', 'metric', 'ms board',
       'kseeb', 'stateboard', 'maticulation',
       'karnataka secondory education board', 'mumbai board', 'sslc',
       'kseb', 'board secondary  education', 'matric board',
       'board of secondary education',
       'west bengal board of secondary education',
       'jharkhand secondary examination board,ranchi', 'u p board',
       'bseb,patna', 'hsc', 'bse', 'sss pune',
       'karnataka education board (keeb)', 'kerala',
```

'state board of secondary education( ssc)', 'gsheb',
'up(allahabad)', 'nagpur', 'don bosco maatriculation school',
'karnataka state secondary education board', 'maharashtra',
'karnataka secondary education board',
'himachal pradesh board of school education',
'certificate of middle years program of ib',
'karnataka board of secondary education',
'board of secondary education rajasthan', 'uttarakhand board',
'ua', 'board of secendary education orissa',
'karantaka secondary education and examination borad', 'hbsc',
'kseeb(karnataka secondary education examination board)',
'cbse[gulf zone]', 'hbse', 'state(karnataka board)',
'jharkhand accademic council',
'jharkhand secondary examination board (ranchi)',
'karnataka secondary education examination board', 'delhi board',
'mirza ahmed ali baig', 'jseb', 'bse, odisha', 'bihar board',
'maharashtra state(latur board)', 'rajasthan board', 'mpboard',
'upbhsie', 'secondary board of rajasthan',
'tamilnadu matriculation board', 'jharkhand secondary board',
'board of secondary education,andhara pradesh', 'up baord',
'state', 'board of intermediate education',
'state board of secondary education,andhra pradesh',
'up board , allahabad',
'stjosephs girls higher sec school,dindigul', 'maharashtra board',
'education board of kerala', 'board of ssc',
'maharashtra state board pune',
'board of school education harayana',
'secondary school cerficate', 'maharashtra sate board', 'ksseb',
'bihar examination board, patna', 'latur',
'board of secondary education, rajasthan', 'state borad hp',
'cluny', 'bsepatna', 'up borad', 'ssc board of andrapradesh',
'matric', 'bse,orissa', 'ssc-andhra pradesh', 'mp',
'karnataka education board', 'mhsbse',
'karnataka sslc board bangalore', 'karnataka', 'u p',
'secondary school of education', 'state board of karnataka',
'karnataka secondary board', 'andhra pradesh board ssc',
'stjoseph of cluny matrhrsecschool,neyveli,cuddalore district',
'hse,orissa', 'national public school', 'nagpur board',
'jharkhand academic council', 'bsemp',
'board of secondary education, andhra pradesh',
'board of secondary education orissa',
'board of secondary education,rajasthan(rbse)',
'board of secondary education,ap',
'board of secondary education,andhra pradesh',
'jawahar navodaya vidyalaya', 'aisse',
'karnataka board of higher education', 'bihar',
'kerala state board', 'cicse', 'tn state board',
'kolhapur divisional board, maharashtra',
'bharathi matriculation school', 'uttaranchal state board',
'wbbsce', 'mp state board', 'seba(assam)', 'anglo indian', 'gseb',
'uttar pradesh', 'ghseb', 'board of school education uttarakhand',
'msbshse,pune', 'tamilnadu state board', 'kerala university',
'uttaranchal shiksha avam pariksha parishad',
'bse(board of secondary education)',
'bright way college, (up board)',
'school secondary education, andhra pradesh',
'secondary state certificate',
'maharashtra state board of secondary and higher secondary education,pune',
'andhra pradesh state board', 'stmary higher secondary', 'cgbse',
'secondary school certificate', 'rajasthan board ajmer', 'mpbse',
'pune board', 'cbse ', 'board of secondary education,orissa',
'maharashtra state board,pune', 'up bord',
'kiran english medium high school', 'state board (jac, ranchi)',
'gujarat board', 'state board ', 'sarada high scchool',
'kalaimagal matriculation higher secondary school',
'karnataka board', 'maharastra board', 'sslc board',
'ssc maharashtra board', 'tamil nadu state', 'uttrakhand board',
'bihar secondary education board,patna',
'haryana board of school education',
'sri kannika parameswari highier secondary school, udumalpet',
'ksseb(karnataka state board)', 'nashik board',
'jharkhand secondary education board', 'himachal pradesh board',
'maharashtra satate board',
'maharashtra state board mumbai divisional board',
'dav public school,hehal',
'state board of secondary education, ap',
'rajasthan board of secondary education', 'hsce',
'karnataka secondary education',
'board of secondary education,odisha', 'maharashtra nasik board',
'west bengal board of secondary examination (wbbse)',
'holy cross matriculation hr sec school', 'cbsc', 'apssc',
'bseb patna', 'kolhapur', 'bseb, patna', 'up board allahabad',
'biharboard', 'nagpur board,nagpur', 'pune', 'gyan bharati school',
'rbse,ajmer', 'board of secondaray education',
'secondary school education', 'state bord', 'jbse,jharkhand',
'hse', 'madhya pradesh board', 'bihar school examination board',
'west bengal board of secondary eucation', 'state boardmp board ',
'icse board , new delhi',
'board of secondary education (bse) orissa',

```
       'maharashtra state board for ssc',
       'board of secondary school education', 'latur board',
       "stmary's convent inter college", 'nagpur divisional board',
       'ap state board', 'cgbse raipur', 'uttranchal board', 'ksbe',
       'central board of secondary education, new delhi',
       'bihar school examination board patna', 'cbse board',
       'sslc,karnataka', 'mp-bse', 'up bourd', 'dav public school sec 14',
       'board of school education haryana',
       'council for indian school certificate examination',
       'aurangabad board', 'j&k state board of school education',
       'maharashtra state board of secondary and higher secondary education',
       'maharashtra state boar of secondary and higher secondary education',
       'ssc regular', 'karnataka state examination board', 'nasik',
       'west bengal  board of secondary education', 'up board,allahabad',
       'bseb ,patna',
       'state board - west bengal board of secondary education : wbbse',
       'maharashtra state board of secondary & higher secondary education',
       'delhi public school', 'karnataka secondary eduction',
       'secondary education board of rajasthan',
       'maharashtra board, pune', 'rbse (state board)', 'apsche',
       'board of  secondary education',
       'board of high school and intermediate education uttarpradesh',
       'kea', 'board of secondary education - andhra pradesh',
       'ap state board for secondary education', 'seba',
       'punjab school education board, mohali',
       'jharkhand acedemic council', 'hse,board',
       'board of ssc education andhra pradesh', 'up-board', 'bse,odisha'],
      dtype=object)
```

In [29]:
```
data["10board"].value_counts()
```

Out[29]:
```
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
```

In [30]:
```
# as there are some values with '0' replacing them with Unknown
# converting it datatype object  object to string

data['10board'] = data['10board'].astype(str)
data['10board'] = data['10board'].replace("0", "Unknown")
```

In [31]:
```
data["10board"].value_counts()
```

Out[31]:
```
cbse                          1395
state board                   1164
Unknown                        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*

In [32]:
```
data["12board"].value_counts()
```

Out[32]:
```
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
```

In [117…
```
# as there are some values with '0' replacing them with Unknown
# converting it datatype object  object to string

data['12board'] = data['12board'].astype(str)
data['12board'] = data['12board'].replace("0", "Unknown")
```

*GraduationYear*

In [44]:
```
# datatype changed

data['GraduationYear'] = pd.to_datetime(data['GraduationYear'], errors='coerce')
data['GraduationYear'].value_counts()
```

Out[44]:
```
2013-01-01    1181
2014-01-01    1036
2012-01-01     847
2011-01-01     507
2010-01-01     292
2015-01-01      94
2009-01-01      24
2017-01-01       8
2016-01-01       7
2007-01-01       1
Name: GraduationYear, dtype: int64
```

In [45]:
```
# their is value '0' replaced with unknown
data['GraduationYear'] = data['GraduationYear'].replace("0", "Unknown")
```

In [48]:
```
data['GraduationYear'].value_counts()
```

Out[48]:
```
2013-01-01    1181
2014-01-01    1036
2012-01-01     847
2011-01-01     507
2010-01-01     292
2015-01-01      94
2009-01-01      24
2017-01-01       8
2016-01-01       7
2007-01-01       1
Name: GraduationYear, dtype: int64
```

12*graduation*

In [37]:
```
data['12graduation']
```

Out[37]:
```
0       2007
1       2007
2       2010
3       2007
4       2008
        ...
3993    2006
3994    2009
3995    2008
3996    2010
3997    2008
Name: 12graduation, Length: 3998, dtype: int64
```

In [38]:
```
# the column 12Graduation dtype need to be changed from int to date

data['12graduation']=pd.to_datetime(data['12graduation'], format="%Y")
```

In [243…
```
# Removing unwanted columns
data.drop(columns=['CollegeTier', 'CollegeCityTier'], errors='ignore', inplace=True)
```

In [244…
```
data.drop(["CollegeCityID","Domain"],axis=1)
```

Out[244]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScienc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 203097 | 420000 | 2012-06-01 | 2024-02-22 22:01:18.975275 | senior quality engineer | Bangalore | f | 1990-02-19 | 84.30 | board ofsecondary education,ap | ... | - |
| **1** | 579905 | 500000 | 2013-09-01 | 2024-02-22 22:01:18.975275 | assistant manager | Indore | m | 1989-10-04 | 85.40 | cbse | ... | - |
| **2** | 810601 | 325000 | 2014-06-01 | 2024-02-22 22:01:18.975275 | systems engineer | Chennai | f | 1992-08-03 | 85.00 | cbse | ... | - |
| **3** | 267447 | 1100000 | 2011-07-01 | 2024-02-22 22:01:18.975275 | senior software engineer | Gurgaon | m | 1989-12-05 | 85.60 | cbse | ... | - |
| **4** | 343523 | 200000 | 2014-03-01 | 2015-03-01 00:00:00.000000 | get | Manesar | m | 1991-02-27 | 78.00 | cbse | ... | - |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | . |
| **3993** | 47916 | 280000 | 2011-10-01 | 2012-10-01 00:00:00.000000 | software engineer | New Delhi | m | 1987-04-15 | 52.09 | cbse | ... | - |
| **3994** | 752781 | 100000 | 2013-07-01 | 2013-07-01 00:00:00.000000 | technical writer | Hyderabad | f | 1992-08-27 | 90.00 | state board | ... | - |
| **3995** | 355888 | 320000 | 2013-07-01 | 2024-02-22 22:01:18.975275 | associate software engineer | Bangalore | m | 1991-07-03 | 81.86 | bse,odisha | ... | - |
| **3996** | 947111 | 200000 | 2014-07-01 | 2015-01-01 00:00:00.000000 | software developer | Asifabadbanglore | f | 1992-03-20 | 78.72 | state board | ... | 43 |
| **3997** | 324966 | 400000 | 2013-02-01 | 2024-02-22 22:01:18.975275 | senior systems engineer | Chennai | f | 1991-02-26 | 70.60 | cbse | ... | - |

3998 rows × 34 columns

In [245…

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 36 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     3998 non-null   int64
 1   Salary                 3998 non-null   int64
 2   DOJ                    3998 non-null   datetime64[ns]
 3   DOL                    3998 non-null   datetime64[ns]
 4   Designation            3998 non-null   object
 5   JobCity                3998 non-null   object
 6   Gender                 3998 non-null   object
 7   DOB                    3998 non-null   datetime64[ns]
 8   10percentage           3998 non-null   float64
 9   10board                3998 non-null   object
 10  12graduation           3998 non-null   int64
 11  12percentage           3998 non-null   float64
 12  12board                3998 non-null   object
 13  CollegeID              3998 non-null   int64
 14  Degree                 3998 non-null   object
 15  Specialization         3998 non-null   object
 16  collegeGPA             3998 non-null   float64
 17  CollegeCityID          3998 non-null   int64
 18  CollegeState           3998 non-null   object
 19  GraduationYear         3998 non-null   int64
 20  English                3998 non-null   int64
 21  Logical                3998 non-null   int64
 22  Quant                  3998 non-null   int64
 23  Domain                 3998 non-null   float64
 24  ComputerProgramming    3998 non-null   int64
 25  ElectronicsAndSemicon  3998 non-null   int64
 26  ComputerScience        3998 non-null   int64
 27  MechanicalEngg         3998 non-null   int64
 28  ElectricalEngg         3998 non-null   int64
 29  TelecomEngg            3998 non-null   int64
 30  CivilEngg              3998 non-null   int64
 31  conscientiousness      3998 non-null   float64
 32  agreeableness          3998 non-null   float64
 33  extraversion           3998 non-null   float64
 34  nueroticism            3998 non-null   float64
 35  openess_to_experience  3998 non-null   float64
dtypes: datetime64[ns](3), float64(9), int64(16), object(8)
memory usage: 1.1+ MB
```

*CleanedData*

In [54]:

```
data
```

Out[54]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | CollegeState | Graduati |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 203097 | 420000 | 2012-06-01 | 2024-02-22 17:48:19.511385 | senior quality engineer | Bangalore | f | 1990-02-19 | 84.30 | board ofsecondary education,ap | ... | Andhra Pradesh | 201 |
| **1** | 579905 | 500000 | 2013-09-01 | 2024-02-22 17:48:19.511385 | assistant manager | Indore | m | 1989-10-04 | 85.40 | cbse | ... | Madhya Pradesh | 201 |
| **2** | 810601 | 325000 | 2014-06-01 | 2024-02-22 17:48:19.511385 | system engineer | Chennai | f | 1992-08-03 | 85.00 | cbse | ... | Uttar Pradesh | 201 |
| **3** | 267447 | 1100000 | 2011-07-01 | 2024-02-22 17:48:19.511385 | senior software engineer | Gurgaon | m | 1989-12-05 | 85.60 | cbse | ... | Delhi | 201 |
| **4** | 343523 | 200000 | 2014-03-01 | 2015-03-01 00:00:00.000000 | graduate engineer trainee | Manesar | m | 1991-02-27 | 78.00 | cbse | ... | Uttar Pradesh | 201 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **3993** | 47916 | 280000 | 2011-10-01 | 2012-10-01 00:00:00.000000 | software engineer | Delhi | m | 1987-04-15 | 52.09 | cbse | ... | Haryana | 201 |
| **3994** | 752781 | 100000 | 2013-07-01 | 2013-07-01 00:00:00.000000 | technical writer | Hyderabad | f | 1992-08-27 | 90.00 | state board | ... | Telangana | 201 |
| **3995** | 355888 | 320000 | 2013-07-01 | 2024-02-22 17:48:19.511385 | associate software engineer | Bangalore | m | 1991-07-03 | 81.86 | bse,odisha | ... | Orissa | 201 |
| **3996** | 947111 | 200000 | 2014-07-01 | 2015-01-01 00:00:00.000000 | software developer | Bangalore | f | 1992-03-20 | 78.72 | state board | ... | Karnataka | 201 |
| **3997** | 324966 | 400000 | 2013-02-01 | 2024-02-22 17:48:19.511385 | senior systems engineer | Chennai | f | 1991-02-26 | 70.60 | cbse | ... | Tamil Nadu | 201 |

3998 rows × 27 columns

# Data Visualization

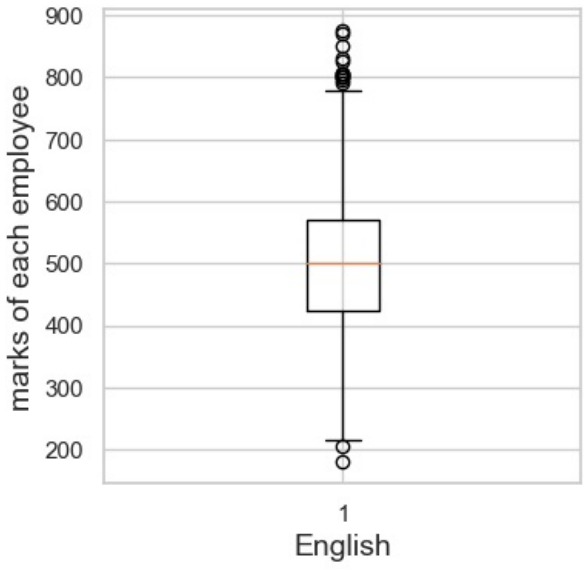
## Univariate Analysis

### Box plot for numerical column

This Box plot tells abouts the English coloumn it shows the marks of each employee and this coloumn have the many outliers like high extream outliers and low extream outliers¶

In [249…

```
plot=plt.subplots(figsize=(4,4))
plt.boxplot(data['English'])
plt.xlabel('English', fontsize=14)
plt.ylabel('marks of each employee', fontsize=14)
plt.show()
```

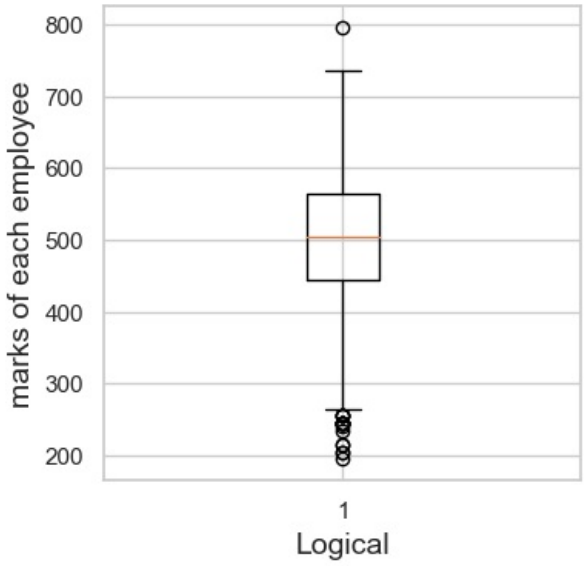



This Box plot tells abouts the Logical coloumn it shows the marks of each employee and this coloumn have the many outliers like high extream outliers and low extream outliers

In [250…

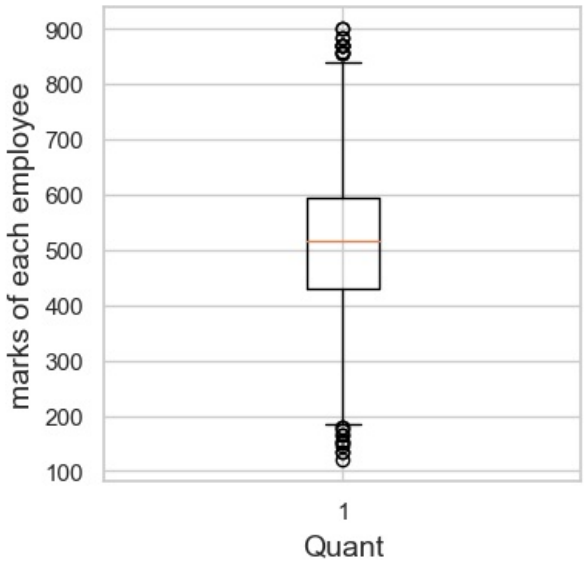
```
plot=plt.subplots(figsize=(4,4))
```

```
plt.boxplot(data['Logical'])
plt.xlabel('Logical', fontsize=14)
plt.ylabel('marks of each employee', fontsize=14)
plt.show()
```




This Box plot tells abouts the Quant coloumn it shows the marks of each employee and this coloumn have the many outliers like high extream outliers and low extream outliers

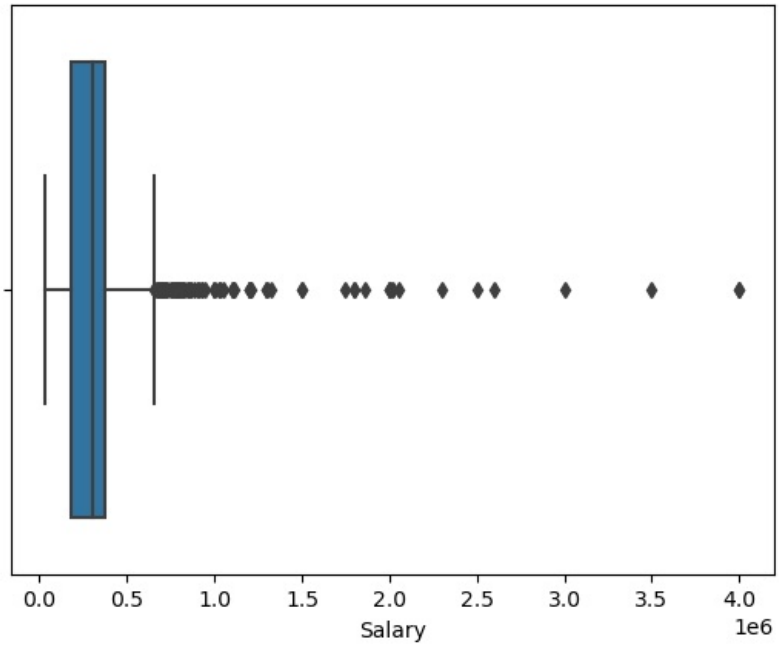
In [251…
```
plot=plt.subplots(figsize=(4,4))
plt.boxplot(data['Quant'])
plt.xlabel('Quant', fontsize=14)
plt.ylabel('marks of each employee', fontsize=14)
plt.show()
```




Boxplot for the Salary column indicates the presence of outliers, particularly towards the extremely higher end.

In [61]:
```
sns.boxplot(data=data,x=data['Salary'])
```

Out[61]:
```
<Axes: xlabel='Salary'>
```

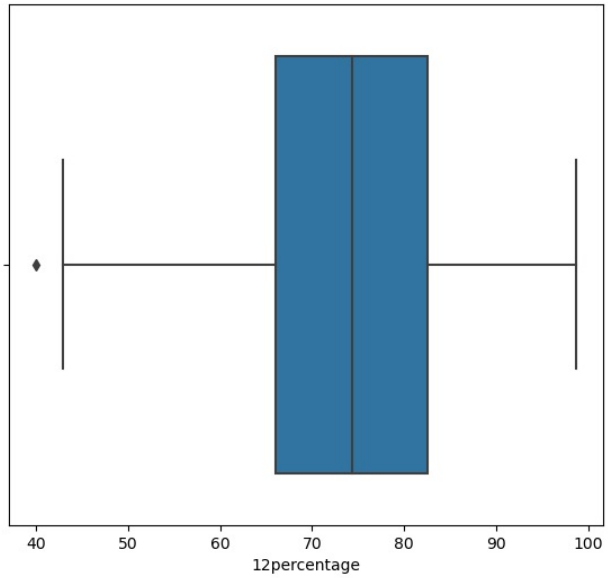


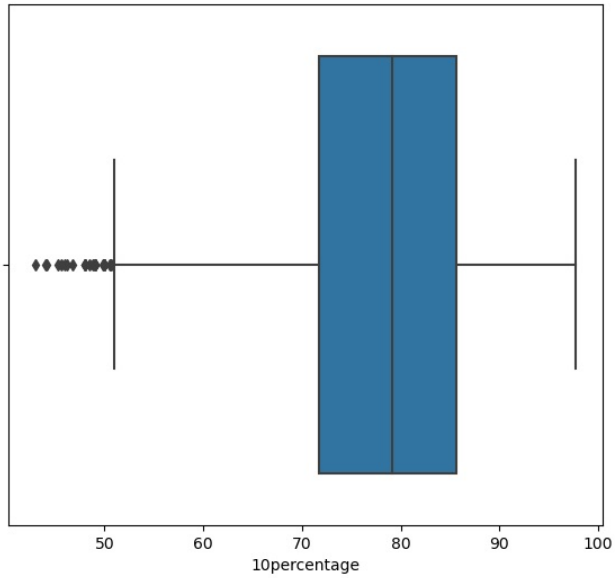

12th percentage has relatively fewer outliers towards lower percentages, with a median around 75%.

10th percentage also exhibits outliers at lower percentages, and its median is approximately 80%.

In [63]:
```
plt.figure(figsize=(15,6))
plt.subplot(121)
sns.boxplot(data=data, x='12percentage')

plt.subplot(122)
sns.boxplot(data=data, x='10percentage')
```

Out[63]:
```
<Axes: xlabel='10percentage'>
```



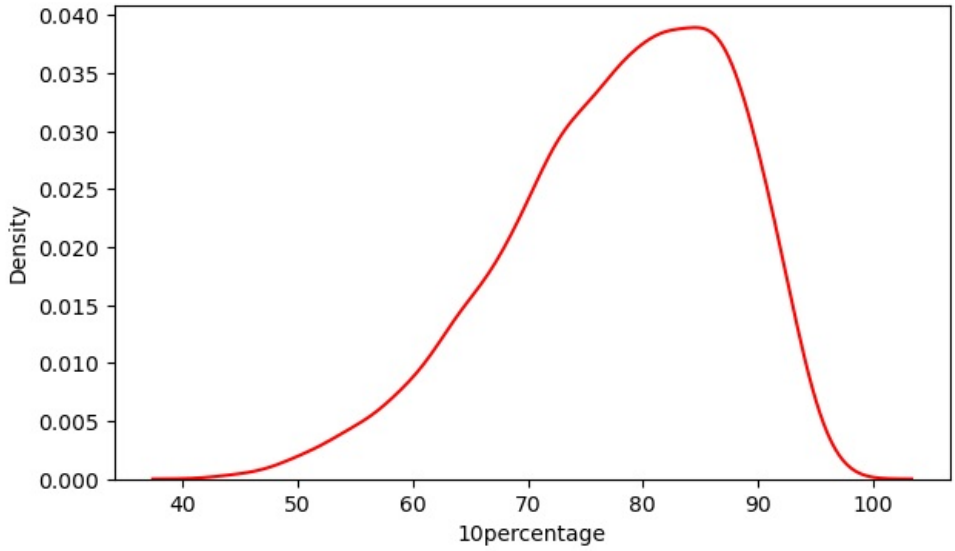




KDE plots for percentages

this kdeplot shows the dencity of the 10percentage of each employee its shows like a left skewed and the most of the employee's are 10percentage is approximatly 80 percentage¶

In [66]:
```
plot = plt.subplots(figsize=(7,4))
sns.kdeplot(data=data, x="10percentage",color="r")
```

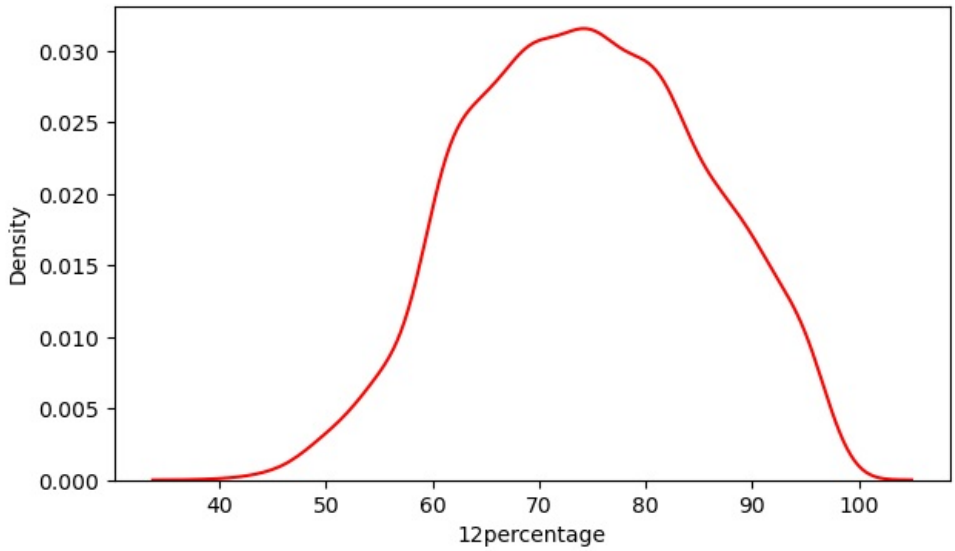
Out[66]:
```
<Axes: xlabel='10percentage', ylabel='Density'>
```




this kdeplot shows the dencity of the 12percentage of each employee its shows like a normal distribution and the most of the employee's are 12percentage is approximatly 70-80 percentage¶

In [67]:
```
plot = plt.subplots(figsize=(7,4))
sns.kdeplot(data=data, x="12percentage",color="r")
```

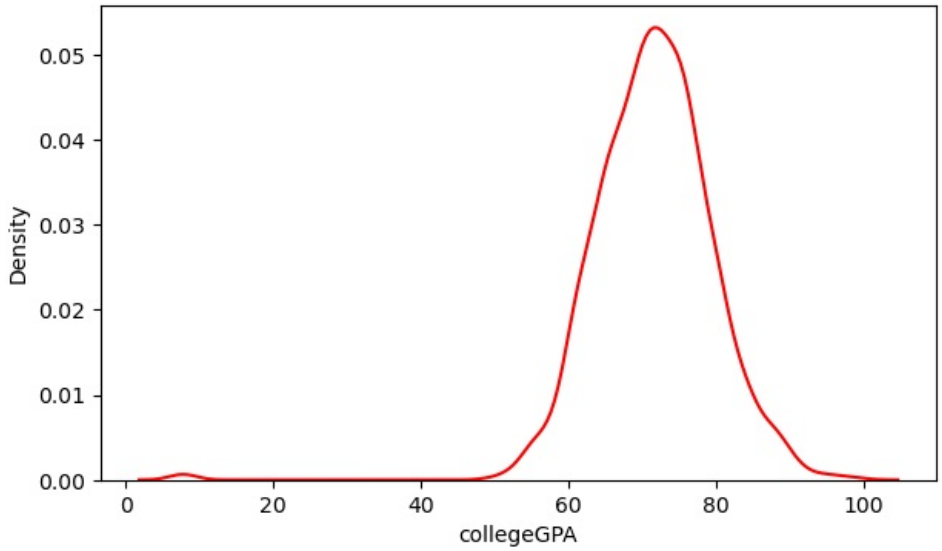
Out[67]:
```
<Axes: xlabel='12percentage', ylabel='Density'>
```




this kdeplot shows the dencity of the collegeGPA of each employee it is a normal distribution and the most of the employee's are collegeGPA is approximatly 65-7

In [68]:
```
plot = plt.subplots(figsize=(7,4))
sns.kdeplot(data=data, x="collegeGPA",color="r")
```

Out[68]:
```
<Axes: xlabel='collegeGPA', ylabel='Density'>
```

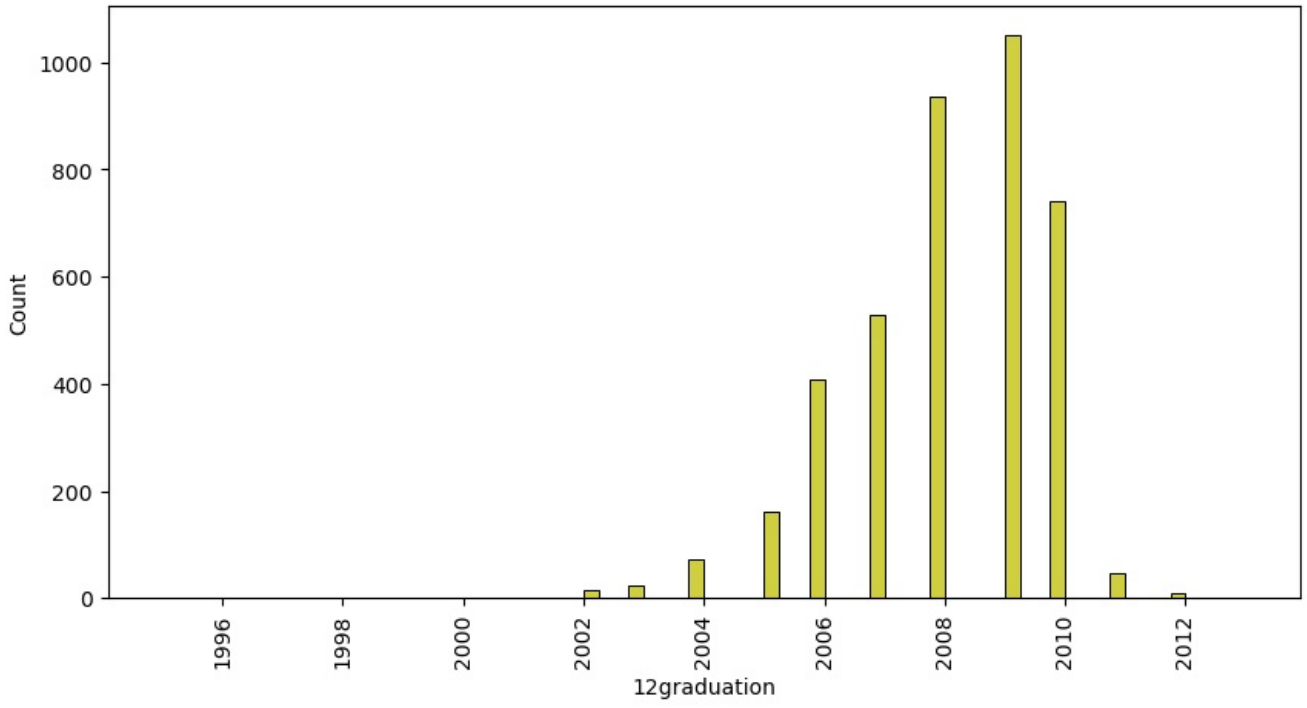



Hist and distribution plots for employees passed out and percentages

This hisogram tells about the most of the employee's are passed out in year 2009 in 12graduation¶

In [71]:
```
plot=plt.subplots(figsize=(10,5))
plt.xticks(rotation=90)
sns.histplot(data=data,x="12graduation",color="y")
```

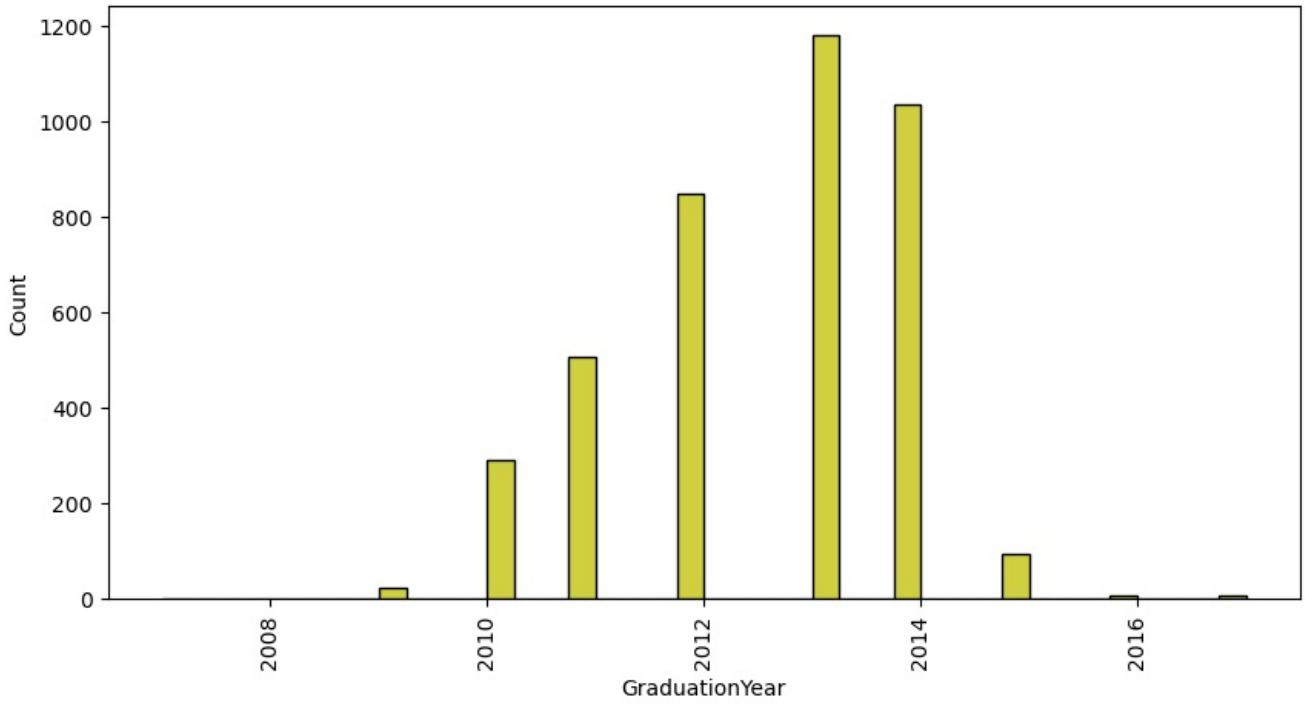
Out[71]:
```
<Axes: xlabel='12graduation', ylabel='Count'>
```




This hisogram tells about the most of the employee's are passed out in year 2013 in Graduation year¶

In [72]:
```
plot=plt.subplots(figsize=(10,5))
plt.xticks(rotation=90)
sns.histplot(data=data,x="GraduationYear",color="y")
```

Out[72]: <Axes: xlabel='GraduationYear', ylabel='Count'>




The data is unimodal, with one prominent peak around the “12percentage” value of approximately between 70 and 80. This suggests that the most common “12percentage” values in the dataset are in this range.

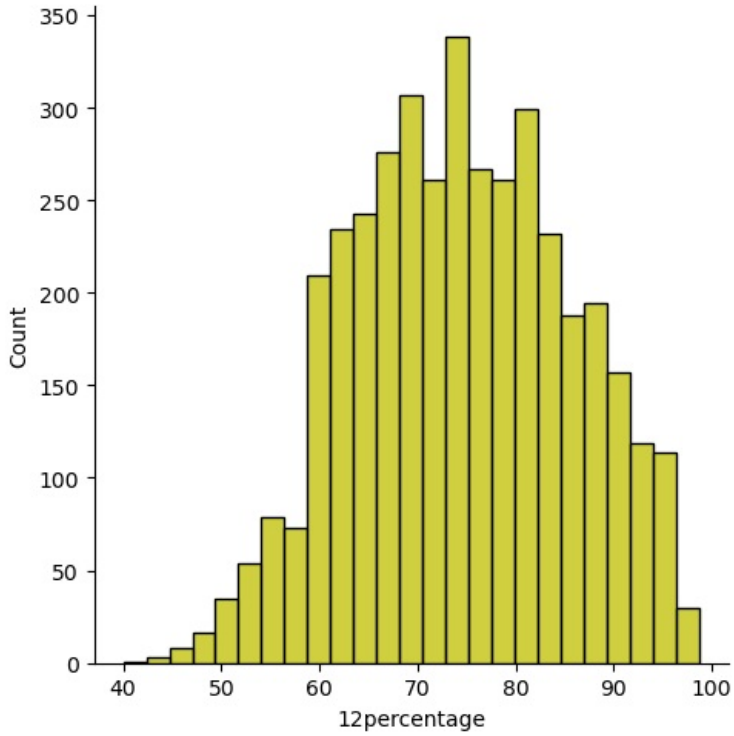
There are fewer occurrences of lower “12percentage” values (below 70) and higher “12percentage” values (above 80). This indicates that such values are less common in the dataset.

The distribution appears to be approximately symmetric around the peak, suggesting that the data is fairly evenly distributed on either side of the peak.

The range of the “12percentage” values is from about 40 to 100. This wide range indicates a significant variability in the data.

In [74]:
```
sns.displot(data=data, x='12percentage',kind='hist',bins=25,color="y")
```

Out[74]: <seaborn.axisgrid.FacetGrid at 0x229780c14d0>




The data is unimodal, with one prominent peak around the “10percentage” value of

approximately between 80 and 90. This suggests that the most common “10percentage” values in the dataset are in this range.

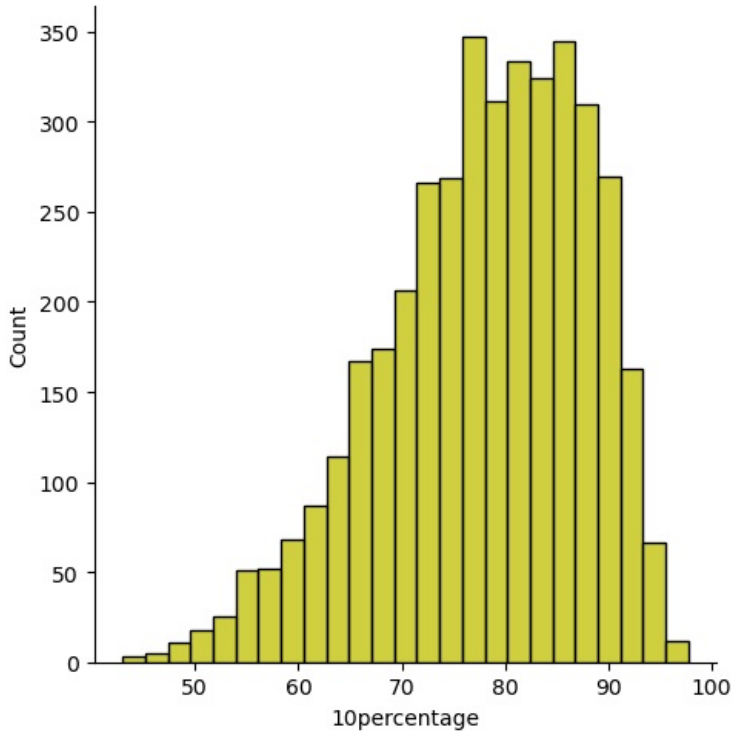
There are fewer occurrences of lower “10percentage” values (below 70) and higher “10percentage” values (above 90). This indicates that such values are less common in the dataset.

The distribution appears to be right-skewed, with a tail extending towards the higher “10percentage” values. This suggests that while most of the data points are clustered around 70-90, there are a few data points with significantly higher values.

The range of the “10percentage” values is from about 0 to 100. This wide range indicates a significant variability in the data.

In [75]:
```
sns.displot(data=data, x='10percentage',kind='hist',bins=25,color="y")
```

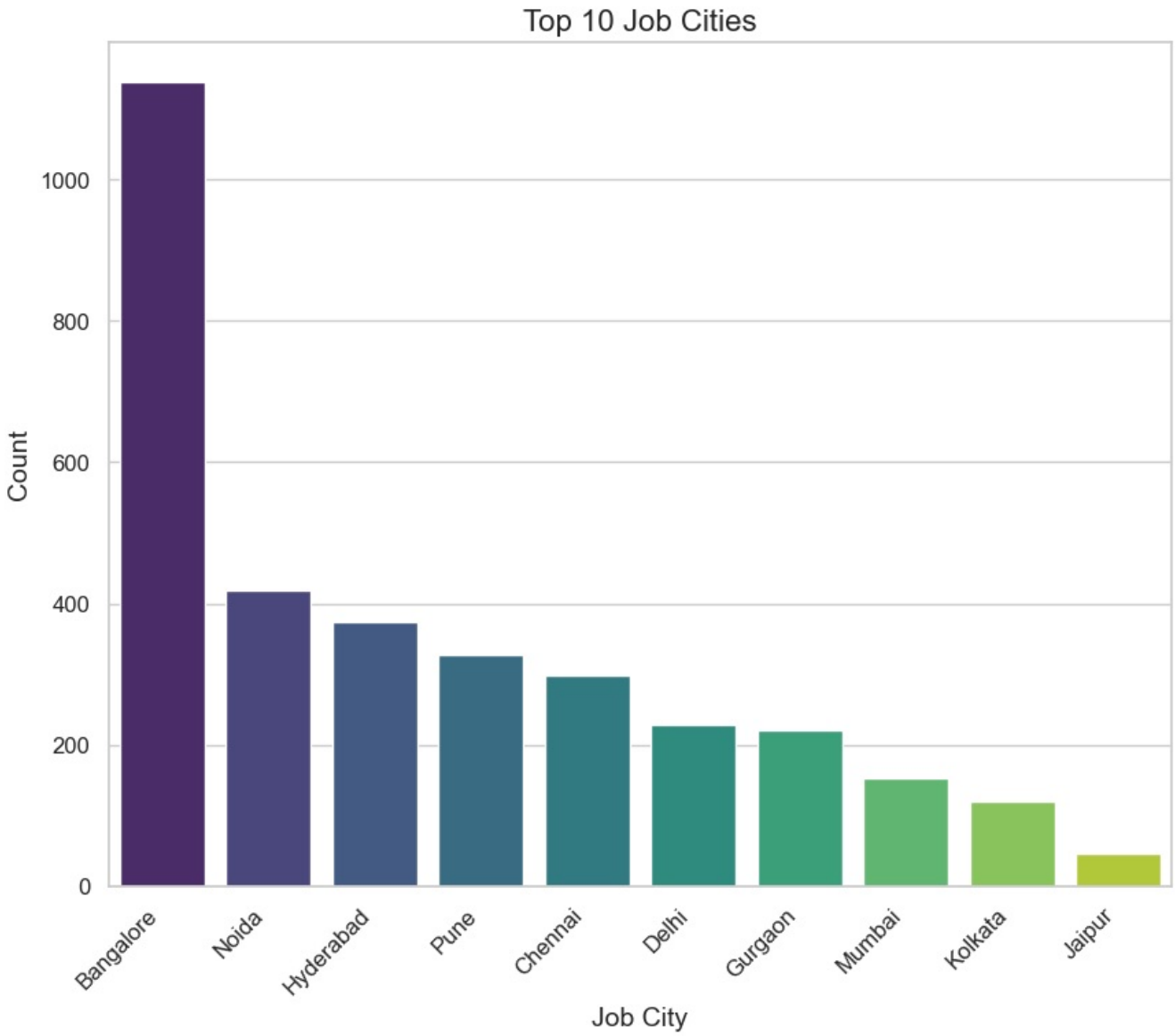
Out[75]:
```
<seaborn.axisgrid.FacetGrid at 0x229786dad10>
```




Count plot for categorical columns

This countplot tells the top 10 job cities where most of the employees are working in Bangalore

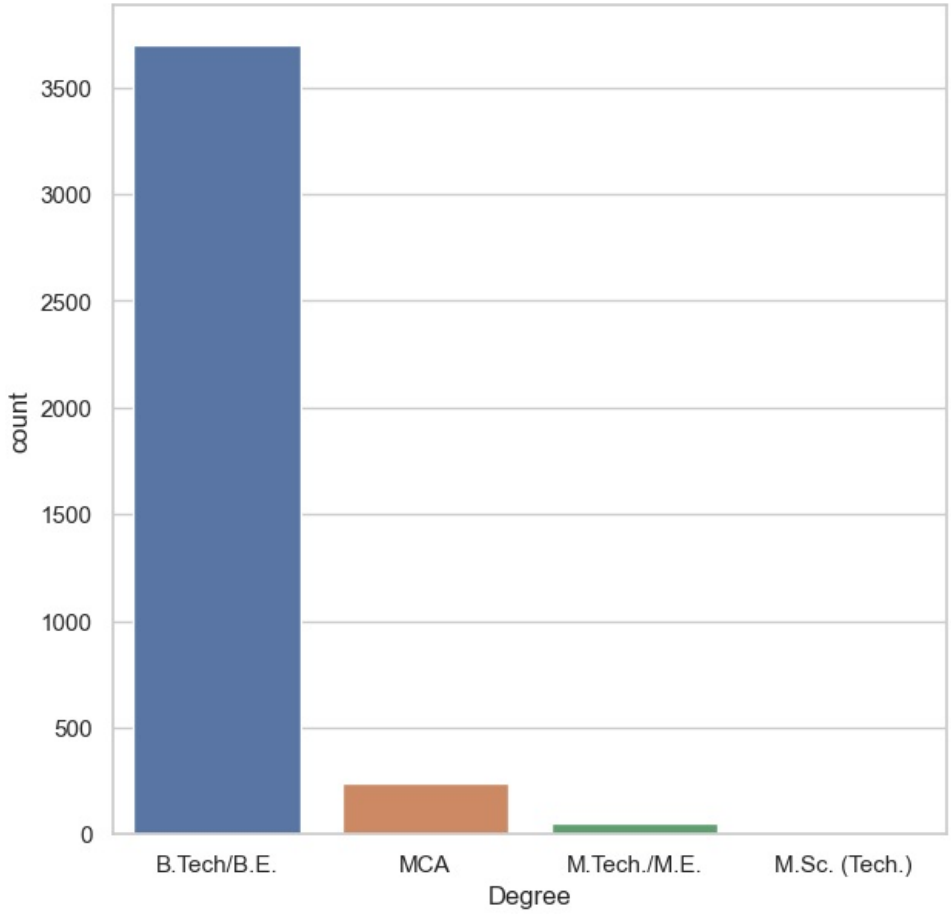
In [264…
```
top_cities = data['JobCity'].value_counts().nlargest(10).index
df_top_cities = data[data['JobCity'].isin(top_cities)]

fig, ax = plt.subplots(figsize=(10, 8))
sns.countplot(x='JobCity', data=df_top_cities, order=top_cities, palette='viridis')

plt.xticks(rotation=45, ha='right', fontsize=12)
plt.yticks(fontsize=12)
plt.xlabel('Job City', fontsize=14)
plt.ylabel('Count', fontsize=14)
plt.title('Top 10 Job Cities', fontsize=16)

plt.show()
```




This countplot tells about the all the employee's are which stream in the Degree coloumn¶

In [260…
```
plot=plt.subplots(figsize=(7,7))
sns.countplot(x='Degree', data=data)
```

Out[260]:
```
<Axes: xlabel='Degree', ylabel='count'>
```

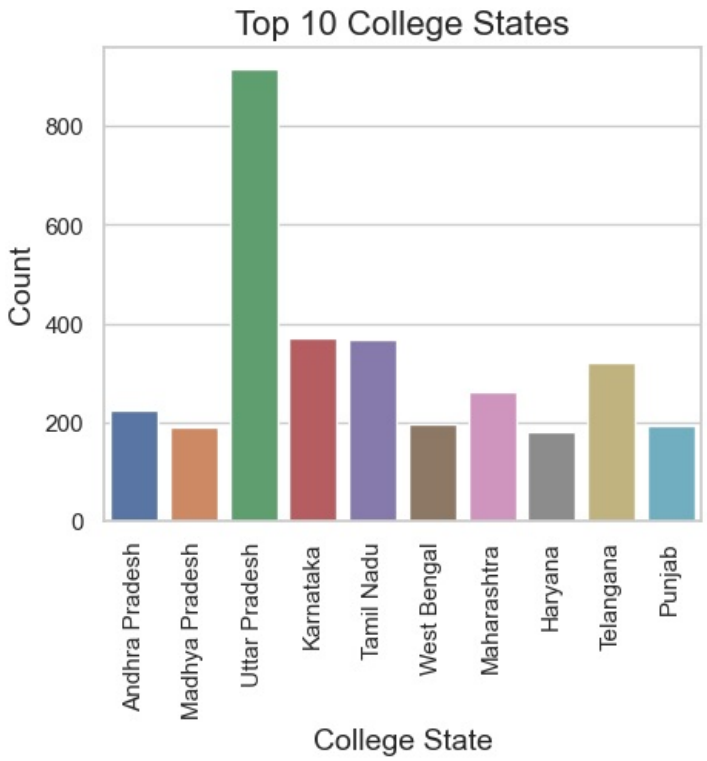



This countplot tells about thetop 10 college state employee's are from they are studied which state and the most of the employee's are from Uttar pradesh¶

In [262…

```
fig, ax = plt.subplots(figsize=(5, 4))
plt.xticks(rotation=90)

top_college_states = data['CollegeState'].value_counts().nlargest(10).index

filtered_data = data[data['CollegeState'].isin(top_college_states)]

sns.countplot(x=filtered_data['CollegeState'], ax=ax)

plt.xlabel('College State', fontsize=14)
plt.ylabel('Count', fontsize=14)
plt.title('Top 10 College States', fontsize=16)
```

```
plt.show()
```

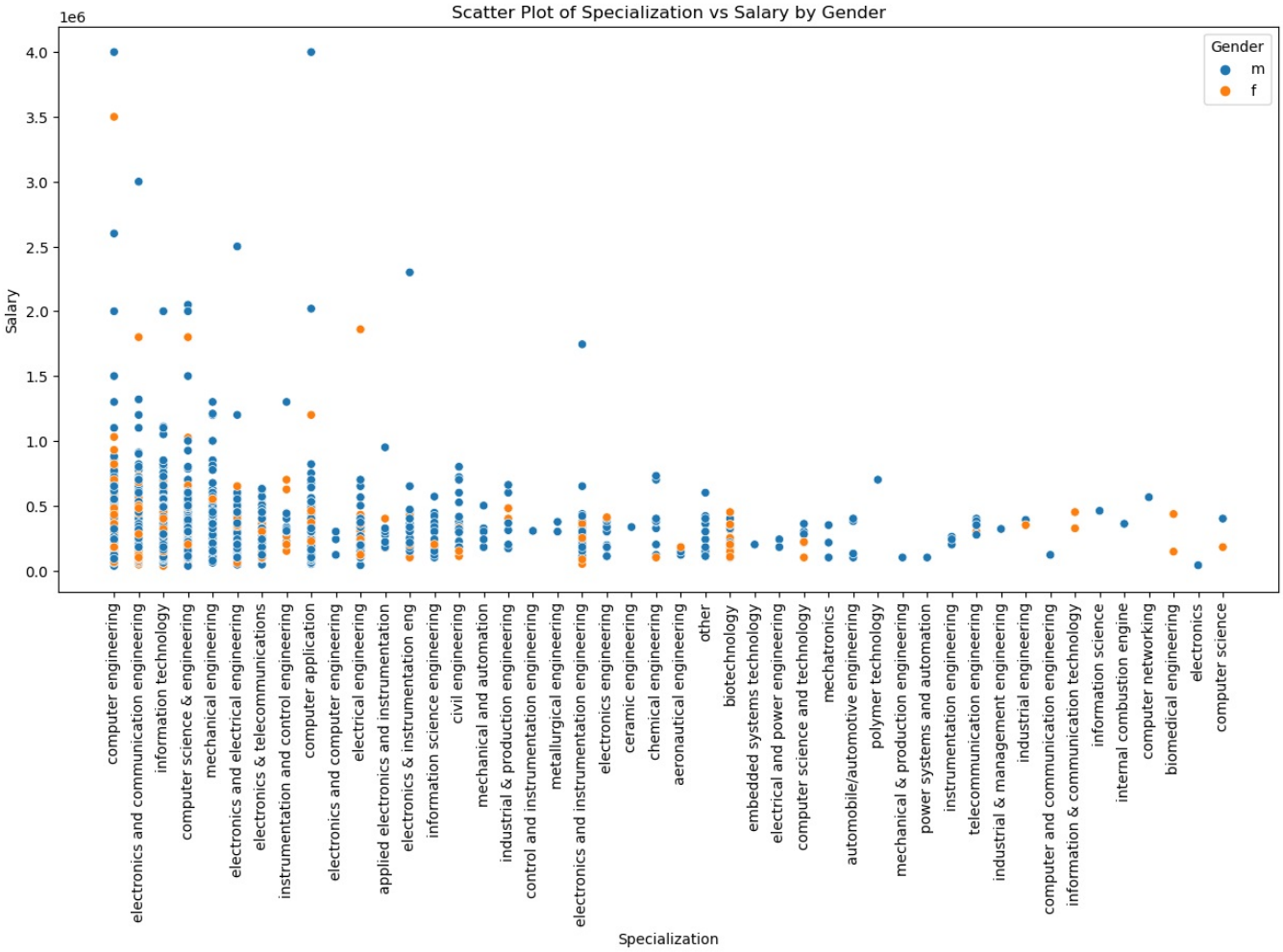



## Bivariate analysis

Scatterplot on Specialization,Salary comparing the Gender

This scatter plot tells about the specialization and salary of the employee's and Gender how many employee's are from male and the how many employee's are from female¶

In [137…
```
fig=plt.subplots(figsize=(15,7))
plt.xticks(rotation=90)
sns.scatterplot(x = "Specialization", y = "Salary", data = data, hue = "Gender",hue_order= ['m', 'f'])
plt.title('Scatter Plot of Specialization vs Salary by Gender')
plt.xlabel('Specialization')
plt.ylabel('Salary')
plt.legend(title='Gender')
plt.show()
```



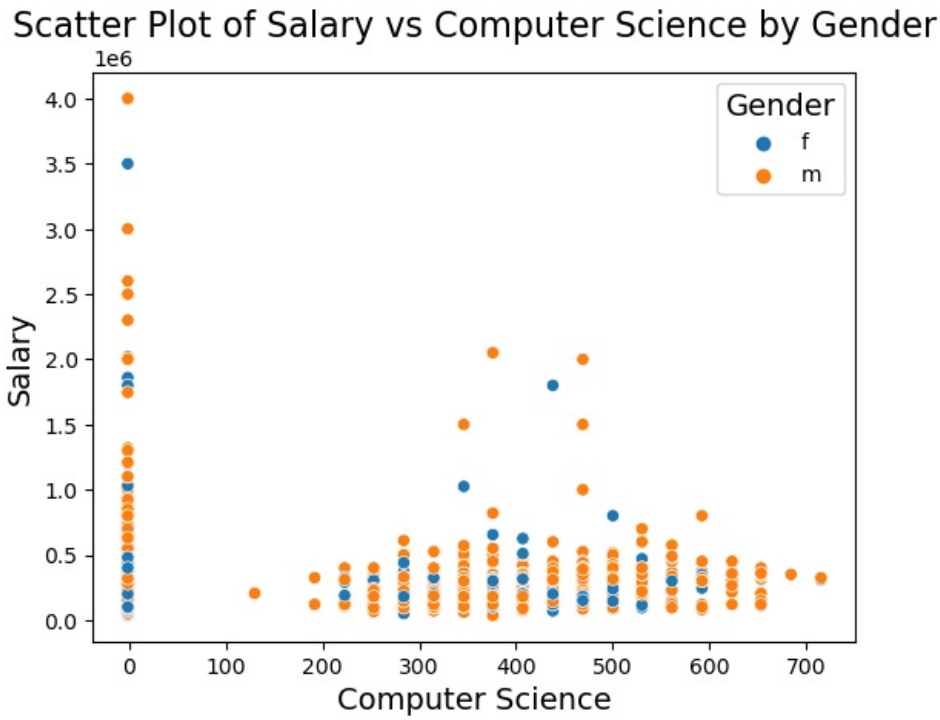

Relation between Salary and ComputerScience columns, in this the AMCAT score for computerScience graduate getting salary range is having high when compare to other high score AMCAT scored people. people who are getting high score.

Here we can observe that when compare to Gender, male is having more high paid Salary than Female and also more males percentage of employment is higher, and some female are also having salary little less based on AMCAT score is higher.

From the above scatterplot the relation between DOJ and Salary is more peoples joining start from 2008 to 2016, and very less people are joined below 2008.

In [151…
```
sns.scatterplot(data=data, y='Salary', x='ComputerScience',hue='Gender')
plt.title('Scatter Plot of Salary vs Computer Science by Gender', fontsize=16)
plt.xlabel('Computer Science', fontsize=14)
plt.ylabel('Salary', fontsize=14)
plt.legend(title='Gender', title_fontsize='14', loc='upper right')
```

Out[151]:
```
<matplotlib.legend.Legend at 0x2297969ce50>
```

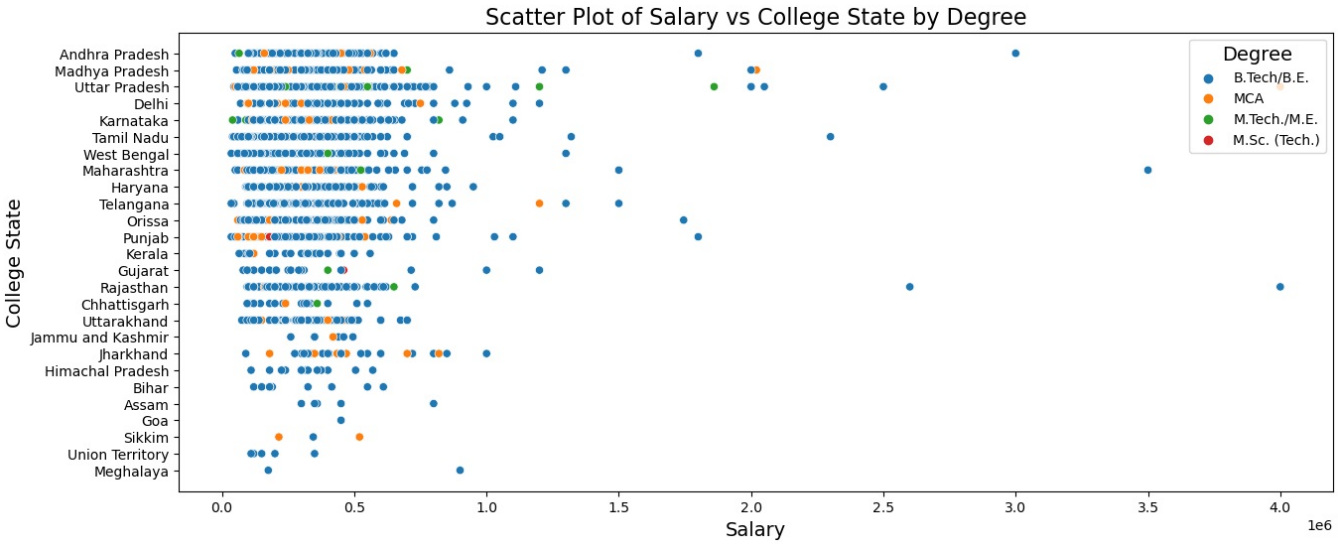



We observe that different states having different salary, from this majority of people are from BTech/B.E and the heighest paid salaried people from Gujarath and Madhya pradesh.

And from other people got high salary and from data majority of people are having salary in the range up to 1000k.

In [153…

```
plt.figure(figsize=(15,6))
sns.scatterplot(data=data, y='CollegeState', x='Salary', hue='Degree')
plt.title('Scatter Plot of Salary vs College State by Degree', fontsize=16)
plt.xlabel('Salary', fontsize=14)
plt.ylabel('College State', fontsize=14)
plt.legend(title='Degree', title_fontsize='14', loc='upper right')
plt.show()
```

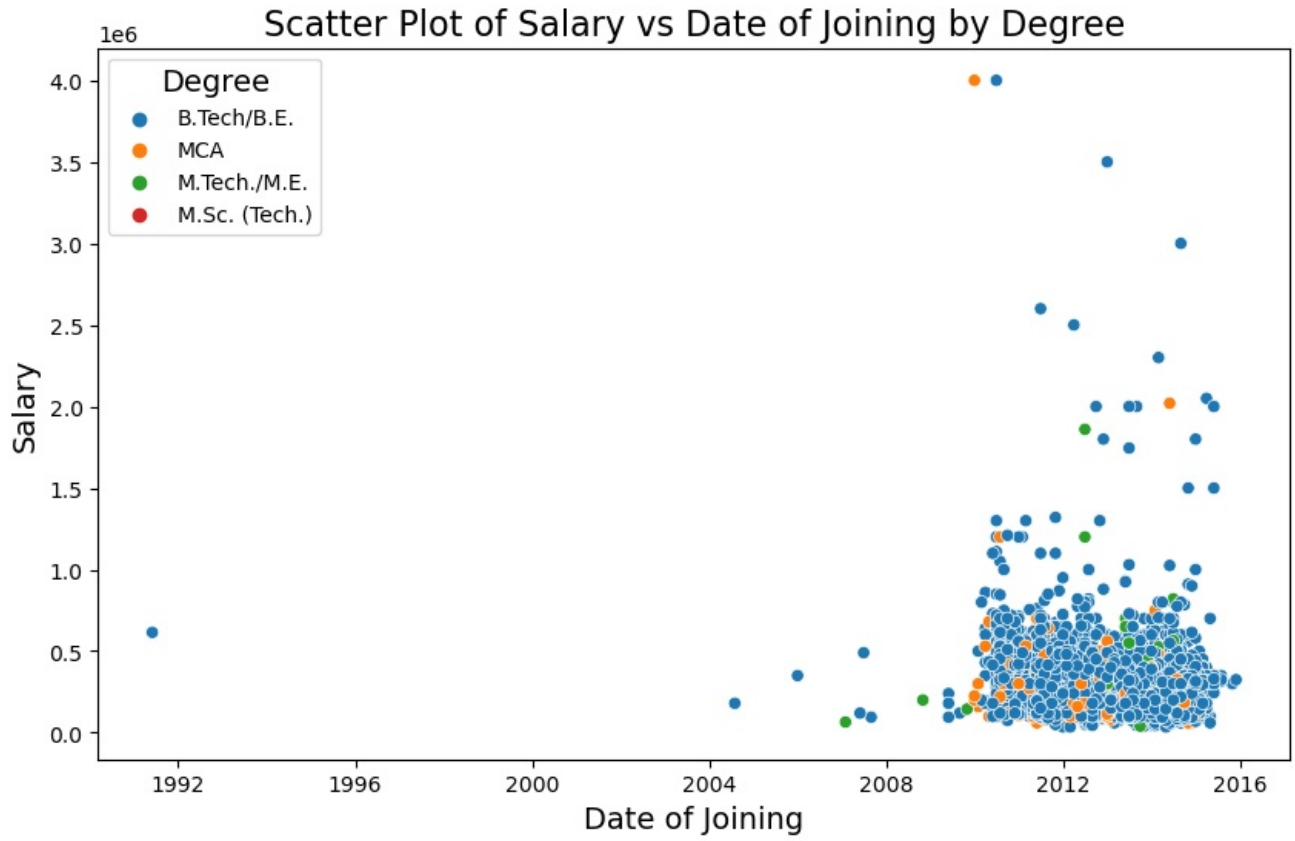



From the above scatterplot the relation between DOJ and Salary is more peoples joining start from 2009 to 2016, and very less people are joined below 2008.

The people joined in 2009 to 2016 having salary ranging below 1000k and very few people having above 1000k

In [156…

```
plt.figure(figsize=(10,6))
sns.scatterplot(data=data, x='DOJ', y='Salary',hue='Degree')
plt.title('Scatter Plot of Salary vs Date of Joining by Degree', fontsize=16)
plt.xlabel('Date of Joining', fontsize=14)
plt.ylabel('Salary', fontsize=14)
plt.legend(title='Degree', title_fontsize='14', loc='upper left')
plt.show()
```



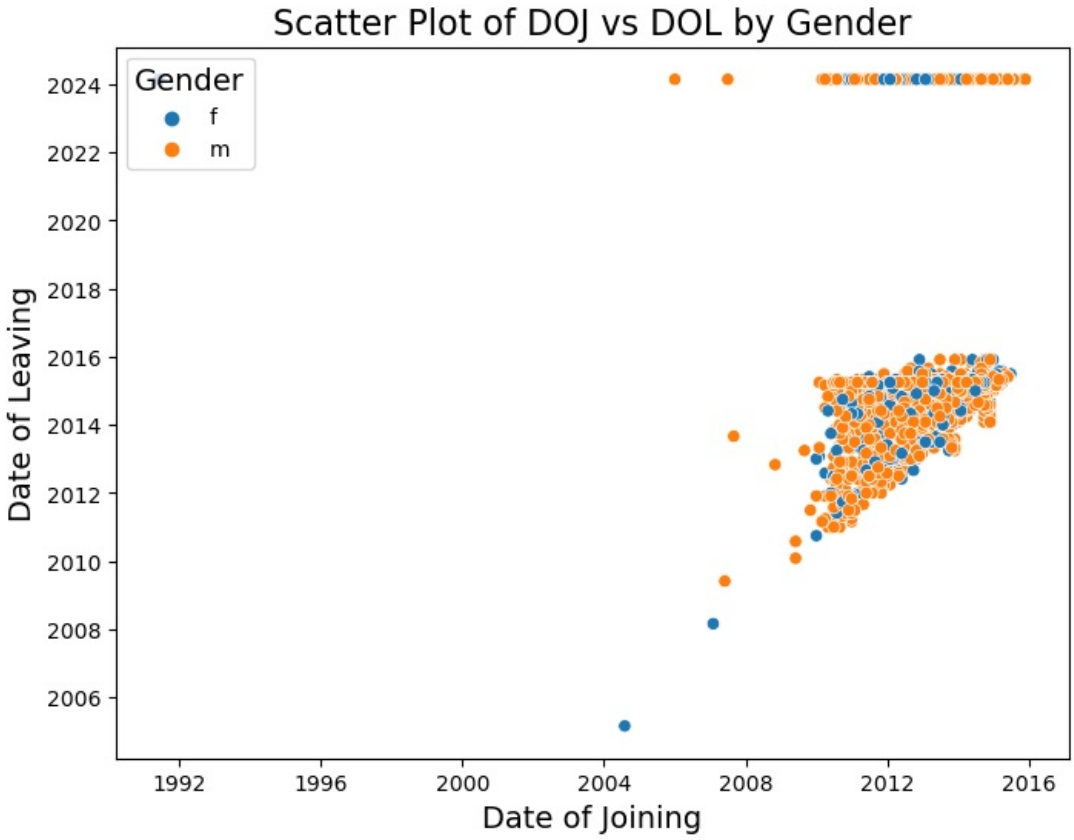

From the above plot that DOJ and DOL is based on Gender, people joined in 2008 to 2015 left the company in between 2012 to 2018

In those some people joined during 2010 to 2016 left the company in the year 2022.

In [190…

```
plt.figure(figsize=(8,6))
sns.scatterplot(data=data, x='DOJ', y='DOL', hue='Gender')
plt.title('Scatter Plot of DOJ vs DOL by Gender', fontsize=16)
plt.xlabel('Date of Joining', fontsize=14)
plt.ylabel('Date of Leaving', fontsize=14)
plt.legend(title='Gender', title_fontsize='14', loc='upper left')
```

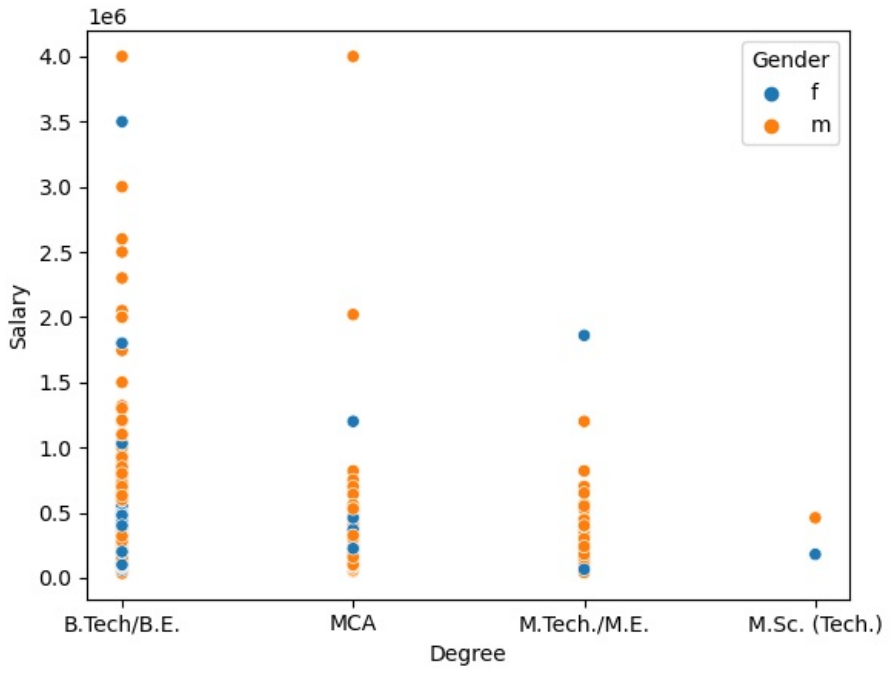
```
# Show the plot
plt.show()
```




- The corelation between Degree and salary based on Gender, we observe that the person having heighest salary from BTech/B.E and MCA graguate is male.
- From the above scatterplot we observe that males having employment when compare to female from the data.
- Scatterplot is drawn between Salary as x-axis and collegeGPA as y-axis and hue is separated them by the Gender. I observe more male and female are having their salary range in below 100k and thier collegeGPA in between 50% and 99%, in those majority of males are present in that data.

In [196…
```
sns.scatterplot(data=data,y='Salary', x="Degree",hue='Gender')
```

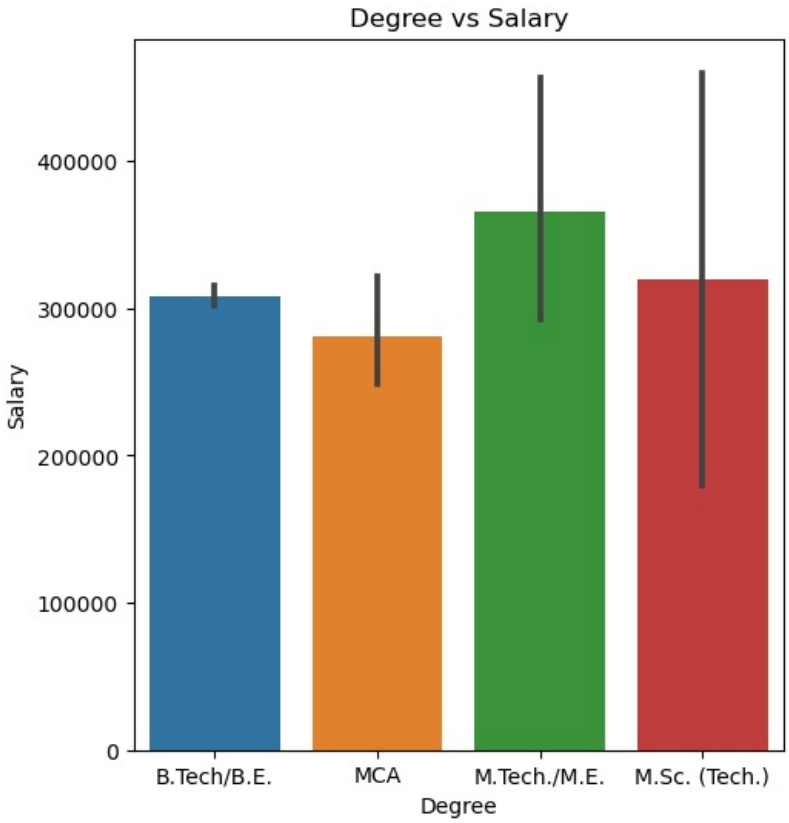
Out[196]:
```
<Axes: xlabel='Degree', ylabel='Salary'>
```




- Barplot for Degree and salary shows that M.Tech./M.E is the heighest salary and M.Sc (Tech) has second heighest salary but both have similar outliers in their salary
- BTech/B.E is having third heighest salary and very less outliers in the BTech/B.E Degree and last MCA has the last salary paid and high outliers than BTech/B.E.

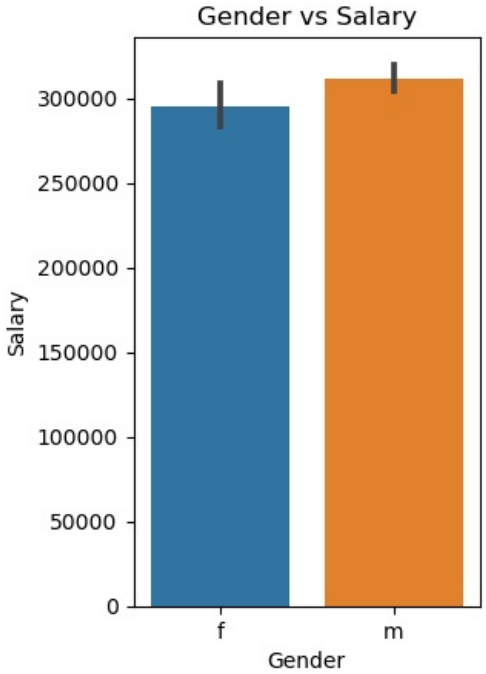
In [191…
```
plt.figure(figsize=(12, 6))

plt.subplot(121)
sns.barplot(x='Degree', y='Salary', data=data)
plt.title('Degree vs Salary')
```

Out[191]:
```
Text(0.5, 1.0, 'Degree vs Salary')
```




- The bar graph in the image, titled “Gender vs Salary”, provides insights into the salary distribution across two categories labeled ‘f’ and ‘m’. Here are some key observations:
- Both categories reach up near the 300,000 mark on the y-axis, indicating similar salary levels for both categories represented.
- Without additional context, it’s hard to provide more specific insights. If these categories refer to different groups within a company or industry, for example, this data might suggest that there is a similar salary range for both groups.

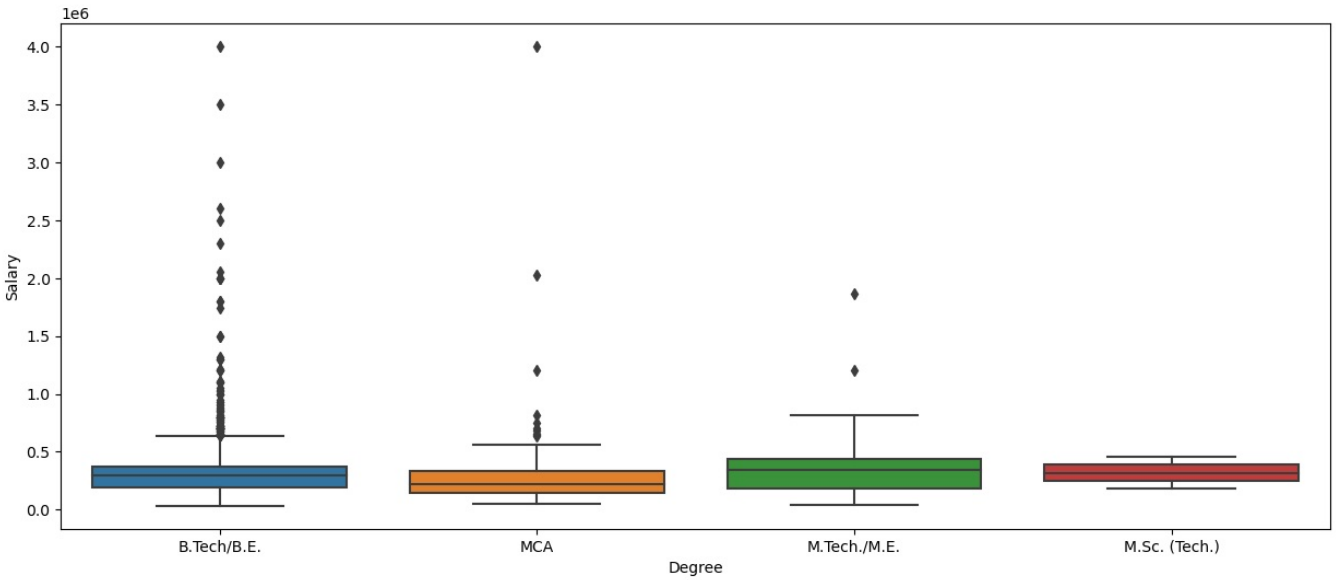
In [192…
```
plt.subplot(122)
sns.barplot(x='Gender', y='Salary', data=data)
plt.title('Gender vs Salary')

plt.show()
```



- Boxplot for Degree and salary shows that BTech/B.E graduates having outliers above 70k to very high and other degrees like MCA, M.Tech./M.E also having very less outliers when compare to BTech/B.E.
- There is no outliers for M.Sc (Tech) Degree and when compare to salary based on Degree M.Tech./M.E is having high median salary.
- M.Sc(Tech is having second high median salary and next BTech/B.E and last median salary goes to MCA graduates.

In [194…
```
plt.figure(figsize=(15,6))
sns.boxplot(data=data,y=data['Salary'], x=data['Degree'])
```

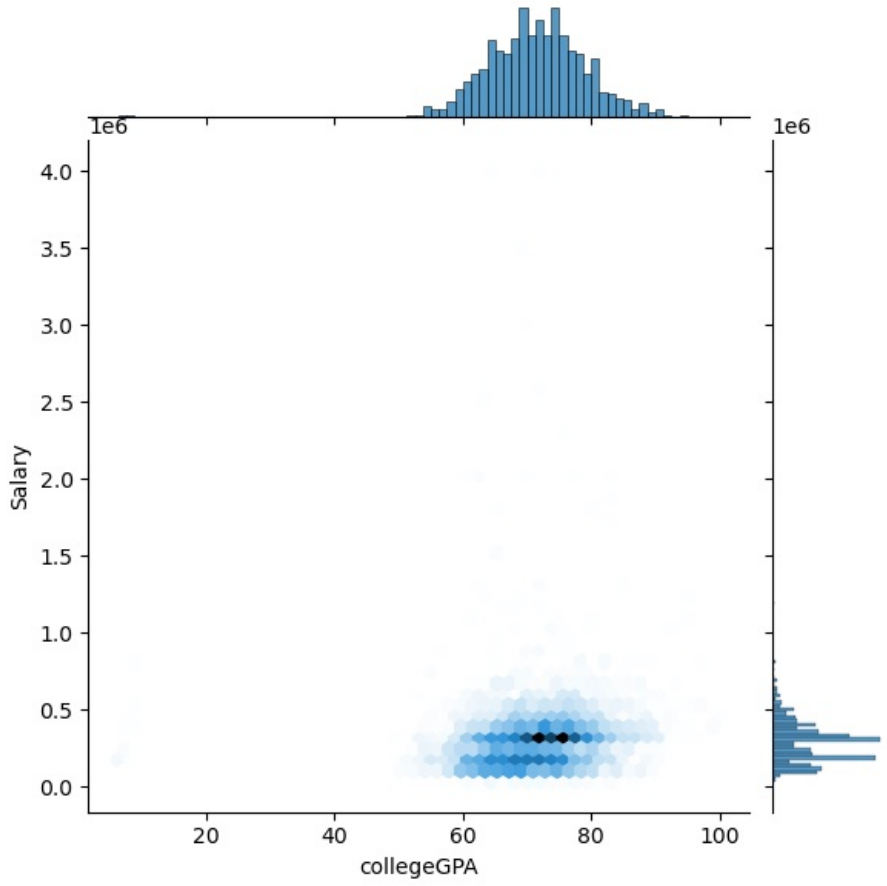
Out[194]:
```
<Axes: xlabel='Degree', ylabel='Salary'>
```




- From hexbinplot the xollegeGPA and salary shows that above 70 to 80 percentage of collegeGPA is having Salary in the range of 50k and most of them having percentage above 60.

In [198…
```
sns.jointplot(data=data, x='collegeGPA', y='Salary', kind='hex')
```

Out[198]:
```
<seaborn.axisgrid.JointGrid at 0x22978d4ec10>
```

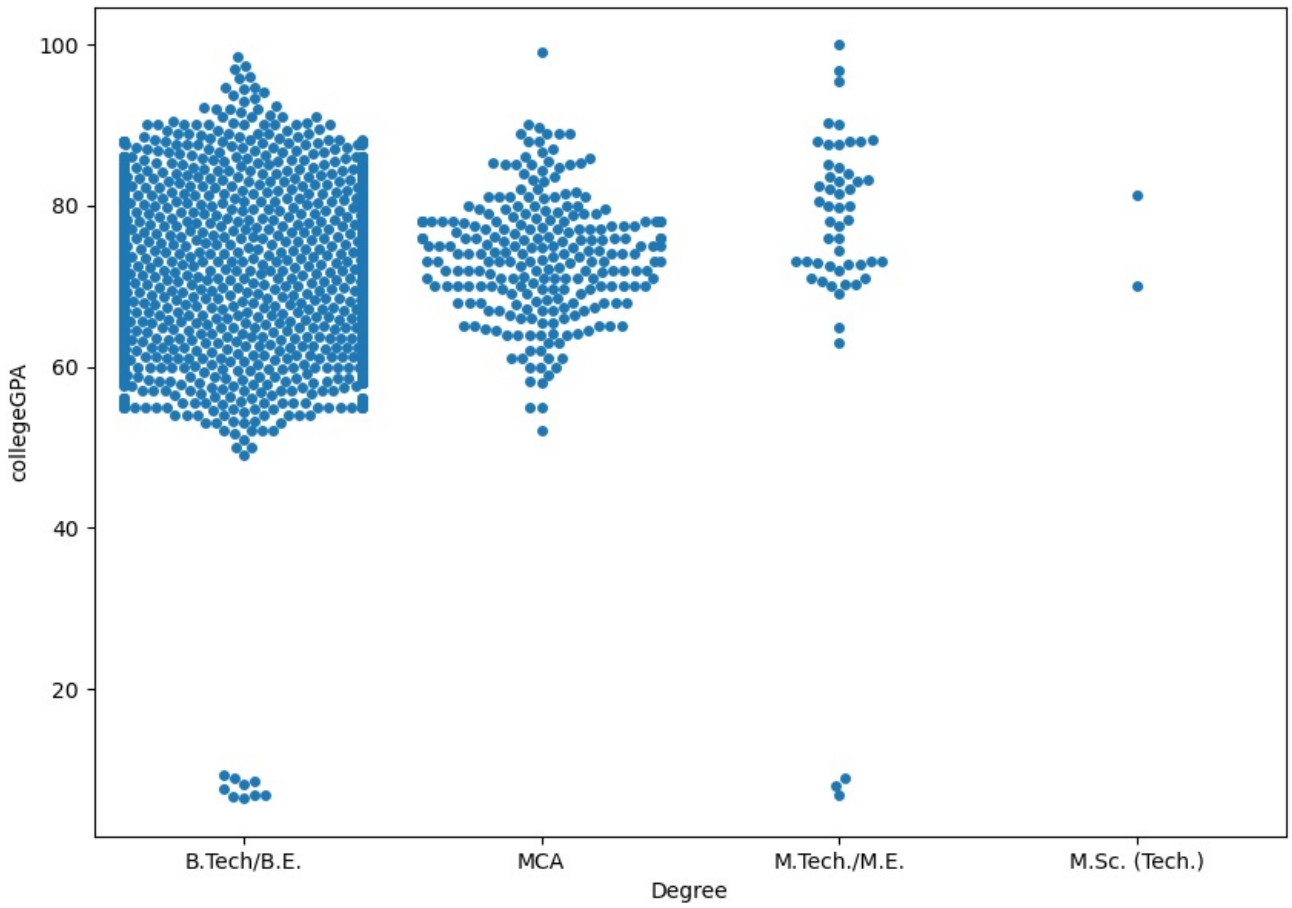



- This swarmplot tells about the Degree and collegeGPA and most of the employee's are from B.E/B.Tech

In [201…
```
import warnings
warnings.filterwarnings('ignore')
```

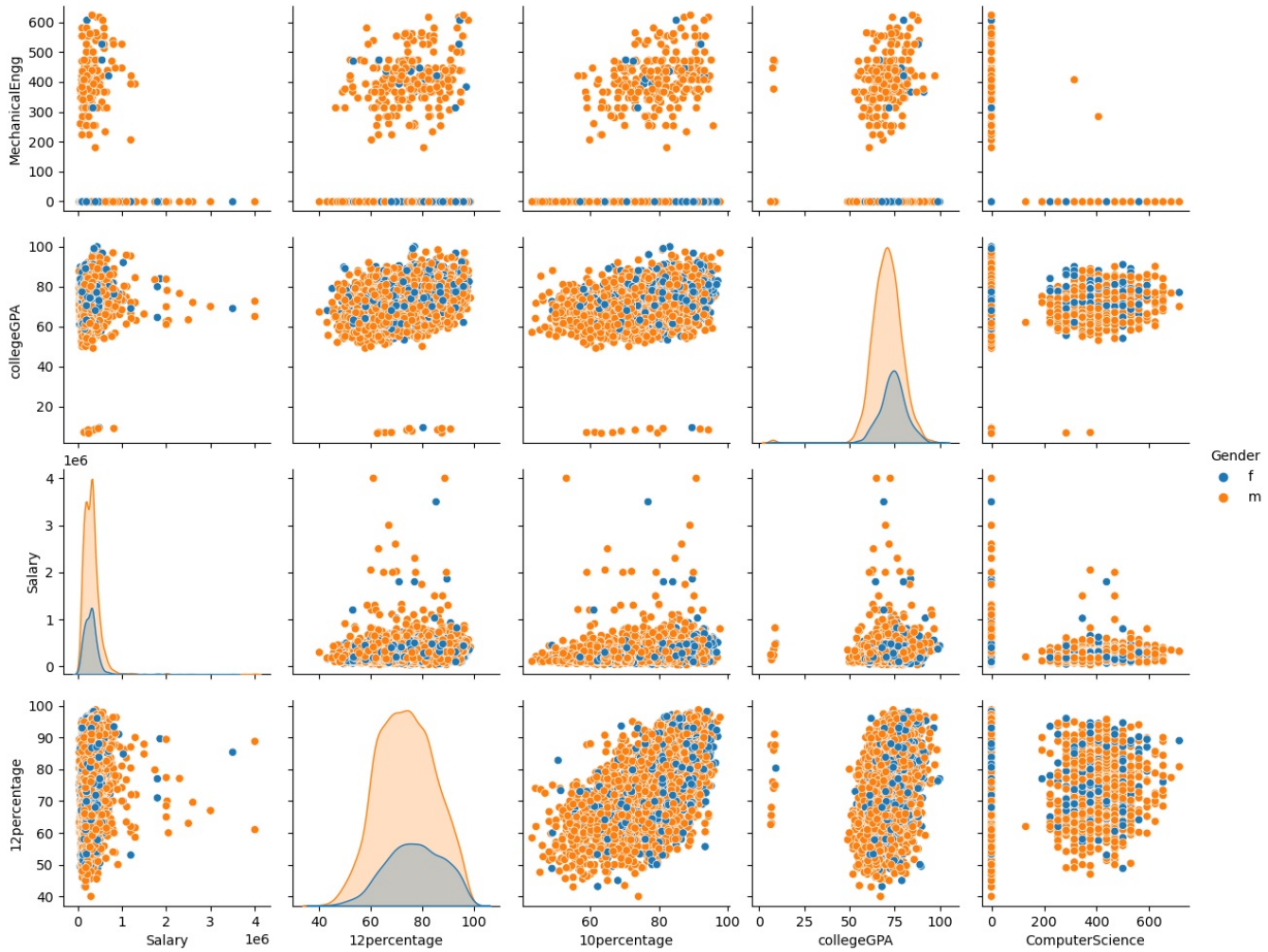
In [202…
```
plt.subplots(figsize=(10,7))
sns.swarmplot(data = data, x='Degree', y='collegeGPA')
```

Out[202]:
```
<Axes: xlabel='Degree', ylabel='collegeGPA'>
```



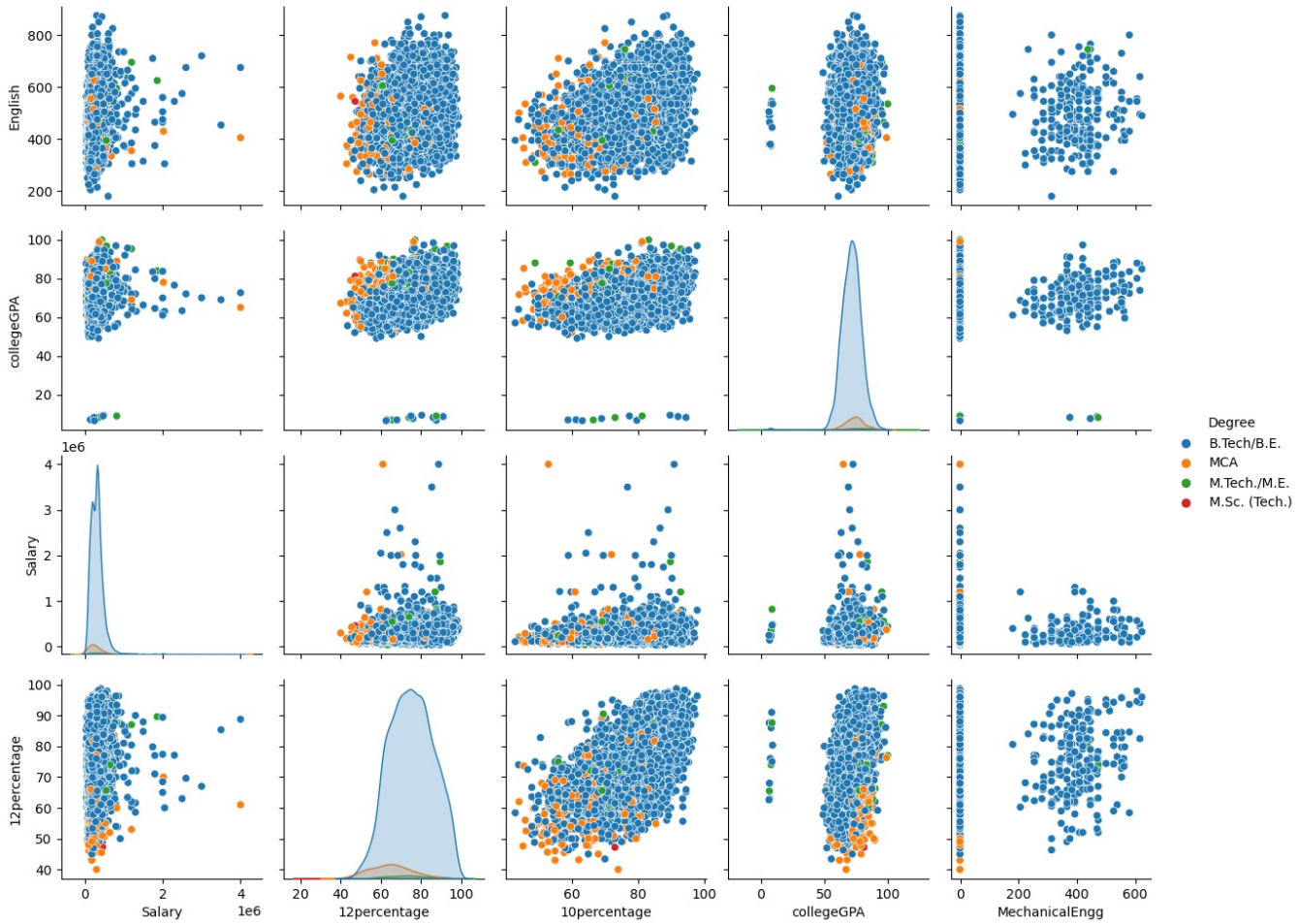

In [204…
```
sns.pairplot(data=data, x_vars=['Salary','12percentage','10percentage','collegeGPA','ComputerScience'], y_vars=
```

Out[204]:
```
<seaborn.axisgrid.PairGrid at 0x229028528d0>
```



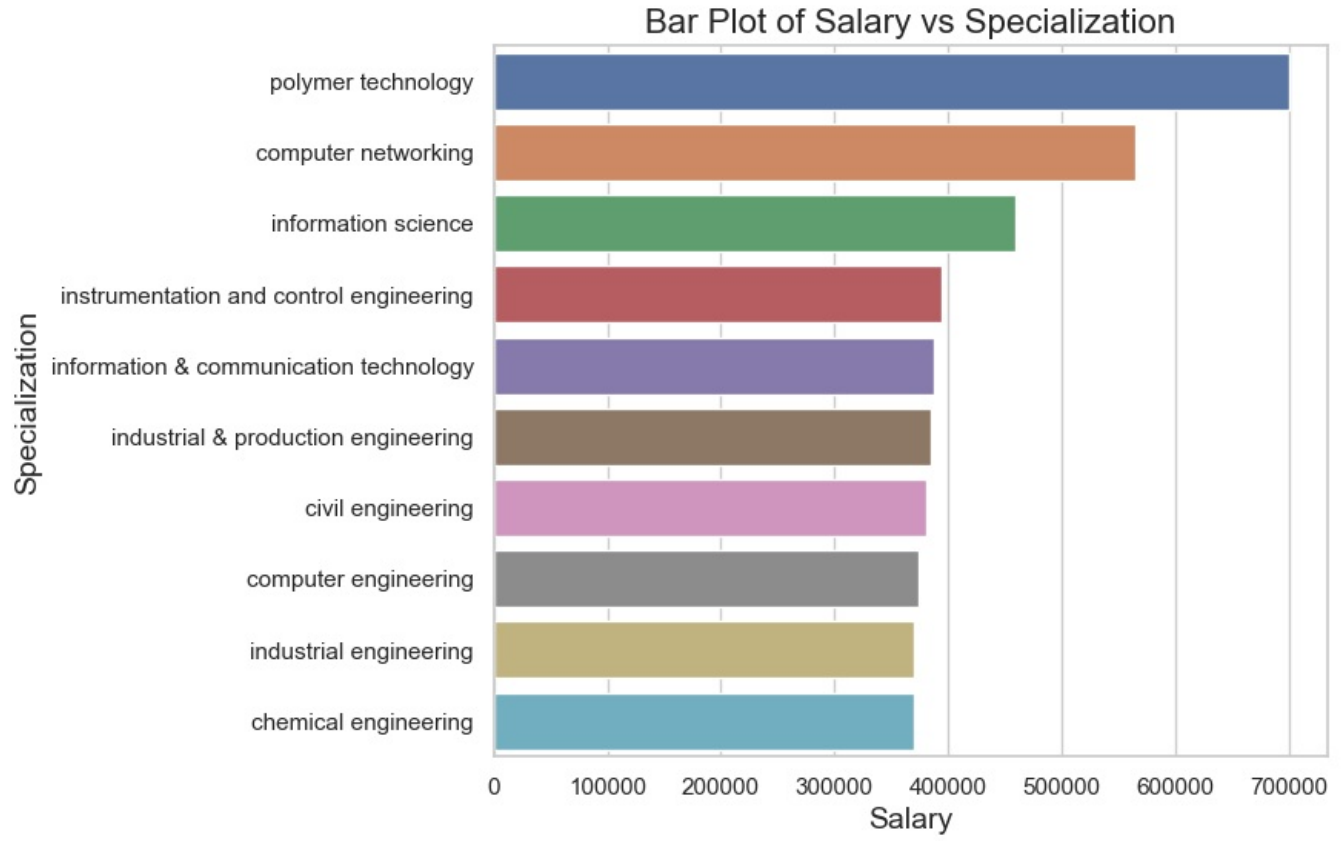

In [206…
```
sns.pairplot(data=data, x_vars=['Salary','12percentage','10percentage','collegeGPA','MechanicalEngg'],y_vars=['
```

Out[206]:
```
<seaborn.axisgrid.PairGrid at 0x2290091d390>
```




- From the above barplot the salary and specialization each and every speialization is having different salary and polymer technology is having the heighest salary and computer networking is next heighest salary.

In [272…
```
plt.figure(figsize=(7,6))
salary_by_sp = data.groupby('Specialization')['Salary'].mean().reset_index().sort_values(by='Salary',ascending=
sns.barplot(data=salary_by_sp, x='Salary', y='Specialization')
plt.title('Bar Plot of Salary vs Specialization', fontsize=16)
plt.xlabel('Salary', fontsize=14)
plt.ylabel('Specialization', fontsize=14)
plt.show()
```



# Research Questions

- Times of India article dated Jan 18, 2019 states that “After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate.” Test this claim with the data given to you.

In [211…
```
que_1=data[(data["Designation"]=="programmer analyst")|(data["Designation"]=="software engineer")|(data["Design
      |(data["Designation"]=="associate engineer")]
que_1
```

Out[211]:

| | ID | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | ... | ComputerScience | Me |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **19** | 466888 | 325000 | 2014-09-01 | 2024-02-22 22:01:18.975275 | software engineer | Pune | f | 1990-11-30 | 79.00 | cbse | ... | -1 | |
| **20** | 140069 | 320000 | 2010-11-01 | 2012-09-01 00:00:00.000000 | software engineer | Bangalore | f | 1988-07-25 | 91.20 | karnataka secondary school of examination | ... | -1 | |
| **21** | 339689 | 200000 | 2012-08-01 | 2013-12-01 00:00:00.000000 | software engineer | -1 | f | 1991-08-20 | 75.67 | up | ... | -1 | |
| **24** | 963123 | 335000 | 2014-06-01 | 2015-06-01 00:00:00.000000 | programmer analyst | Hyderabad | m | 1993-06-28 | 88.00 | state board | ... | 346 | |
| **31** | 1094324 | 340000 | 2014-08-01 | 2015-04-01 00:00:00.000000 | software engineer | Bangalore | m | 1992-10-23 | 77.20 | state board | ... | 407 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **3979** | 212055 | 550000 | 2013-07-01 | 2014-04-01 00:00:00.000000 | software engineer | Bangalore | m | 1989-07-22 | 69.16 | up board | ... | -1 | |
| **3981** | 1077872 | 220000 | 2014-09-01 | 2024-02-22 22:01:18.975275 | software engineer | Gurgaon | m | 1991-12-17 | 53.40 | cbse | ... | 530 | |
| **3984** | 305041 | 480000 | 2011-12-01 | 2024-02-22 22:01:18.975275 | software engineer | Gurgaon | f | 1990-01-18 | 89.80 | cbse | ... | -1 | |
| **3989** | 1204604 | 300000 | 2014-09-01 | 2024-02-22 22:01:18.975275 | software engineer | Bangalore | m | 1991-11-23 | 74.88 | state board | ... | 346 | |
| **3993** | 47916 | 280000 | 2011-10-01 | 2012-10-01 00:00:00.000000 | software engineer | New Delhi | m | 1987-04-15 | 52.09 | cbse | ... | -1 | |

692 rows × 38 columns

In [213…
```
que_1["Salary"]
```

Out[213]:
```
19      325000
20      320000
21      200000
24      335000
31      340000
         ...
3979    550000
3981    220000
3984    480000
3989    300000
3993    280000
Name: Salary, Length: 692, dtype: int64
```

In [218…
```
df2=que_1["Salary"]
df3=[]
for i in ab:
    df3.append(i)
print(df3)
```

```
[325000, 320000, 200000, 335000, 340000, 270000, 380000, 390000, 400000, 250000, 120000, 305000, 300000, 800000
, 325000, 450000, 310000, 340000, 640000, 330000, 305000, 500000, 315000, 300000, 325000, 340000, 305000, 33500
0, 375000, 235000, 450000, 300000, 410000, 240000, 440000, 305000, 325000, 360000, 300000, 300000, 300000, 5500
00, 310000, 300000, 265000, 275000, 170000, 245000, 425000, 300000, 395000, 305000, 250000, 560000, 390000, 305
000, 315000, 300000, 320000, 785000, 300000, 240000, 315000, 330000, 210000, 380000, 405000, 460000, 320000, 27
5000, 320000, 425000, 240000, 180000, 300000, 310000, 300000, 475000, 300000, 240000, 335000, 335000, 390000, 3
10000, 385000, 300000, 385000, 310000, 400000, 300000, 515000, 345000, 250000, 500000, 300000, 450000, 500000,
330000, 450000, 370000, 310000, 445000, 305000, 220000, 405000, 335000, 300000, 180000, 265000, 300000, 315000,
360000, 335000, 100000, 420000, 400000, 300000, 240000, 440000, 335000, 480000, 300000, 400000, 400000, 320000,
375000, 345000, 400000, 420000, 215000, 350000, 300000, 315000, 305000, 180000, 300000, 310000, 350000, 325000,
350000, 350000, 300000, 340000, 200000, 315000, 480000, 310000, 335000, 325000, 450000, 360000, 335000, 350000,
435000, 360000, 350000, 310000, 315000, 345000, 350000, 240000, 400000, 110000, 340000, 400000, 170000, 430000,
85000, 330000, 305000, 200000, 240000, 350000, 550000, 420000, 335000, 380000, 515000, 350000, 275000, 260000,
300000, 395000, 240000, 445000, 240000, 300000, 315000, 450000, 335000, 500000, 800000, 370000, 325000, 300000,
300000, 300000, 350000, 350000, 450000, 280000, 350000, 320000, 240000, 345000, 200000, 335000, 350000, 350000,
400000, 415000, 420000, 300000, 90000, 140000, 120000, 340000, 300000, 505000, 320000, 300000, 310000, 305000,
440000, 375000, 375000, 335000, 275000, 335000, 350000, 430000, 305000, 335000, 275000, 240000, 400000, 290000,
60000, 400000, 335000, 275000, 400000, 450000, 350000, 385000, 335000, 300000, 360000, 305000, 350000, 330000,
120000, 305000, 320000, 375000, 360000, 430000, 165000, 320000, 610000, 190000, 350000, 550000, 300000, 510000,
335000, 300000, 355000, 230000, 475000, 120000, 240000, 305000, 315000, 450000, 475000, 120000, 560000, 360000,
180000, 320000, 320000, 450000, 240000, 300000, 305000, 240000, 620000, 320000, 350000, 120000, 450000, 280000,
410000, 400000, 480000, 310000, 330000, 350000, 95000, 120000, 420000, 330000, 300000, 225000, 430000, 240000,
405000, 280000, 480000, 450000, 335000, 300000, 325000, 200000, 300000, 880000, 200000, 120000, 300000, 300000,
330000, 145000, 200000, 240000, 415000, 335000, 310000, 340000, 215000, 100000, 400000, 300000, 180000, 325000,
290000, 105000, 350000, 350000, 820000, 70000, 400000, 180000, 300000, 450000, 315000, 310000, 180000, 325000,
325000, 320000, 400000, 300000, 605000, 600000, 440000, 195000, 200000, 315000, 325000, 335000, 105000, 335000,
590000, 275000, 300000, 330000, 325000, 280000, 400000, 180000, 305000, 600000, 240000, 140000, 450000, 310000,
200000, 240000, 300000, 450000, 310000, 460000, 300000, 310000, 545000, 340000, 360000, 230000, 500000, 100000,
375000, 120000, 305000, 240000, 240000, 360000, 470000, 180000, 325000, 350000, 335000, 240000, 380000, 330000,
350000, 350000, 265000, 720000, 400000, 465000, 200000, 350000, 300000, 490000, 180000, 335000, 210000, 140000,
520000, 430000, 325000, 550000, 420000, 220000, 350000, 300000, 375000, 335000, 505000, 930000, 350000, 345000,
400000, 310000, 325000, 680000, 305000, 620000, 315000, 335000, 340000, 225000, 455000, 290000, 1000000, 335000
, 180000, 530000, 570000, 400000, 205000, 320000, 245000, 400000, 205000, 335000, 80000, 360000, 350000, 360000
, 250000, 120000, 150000, 390000, 300000, 430000, 300000, 240000, 355000, 460000, 240000, 300000, 300000, 18000
0, 485000, 85000, 350000, 400000, 320000, 820000, 315000, 325000, 120000, 500000, 345000, 305000, 110000, 22500
0, 310000, 220000, 440000, 305000, 330000, 135000, 370000, 310000, 300000, 330000, 420000, 400000, 200000, 3600
00, 335000, 505000, 300000, 220000, 350000, 365000, 325000, 350000, 335000, 300000, 250000, 340000, 450000, 355
000, 200000, 300000, 210000, 350000, 300000, 325000, 420000, 325000, 420000, 630000, 300000, 450000, 300000, 40
0000, 240000, 420000, 430000, 405000, 345000, 400000, 100000, 500000, 230000, 150000, 325000, 240000, 360000, 3
50000, 180000, 375000, 180000, 300000, 305000, 530000, 300000, 330000, 315000, 405000, 240000, 325000, 650000,
310000, 325000, 565000, 120000, 250000, 420000, 200000, 460000, 380000, 145000, 480000, 1500000, 350000, 320000
, 300000, 455000, 195000, 320000, 280000, 500000, 360000, 550000, 415000, 600000, 570000, 360000, 110000, 33500
0, 590000, 325000, 315000, 325000, 300000, 400000, 310000, 180000, 480000, 240000, 380000, 640000, 290000, 4000
00, 400000, 330000, 180000, 335000, 380000, 400000, 550000, 300000, 145000, 315000, 360000, 220000, 500000, 455
000, 475000, 160000, 90000, 310000, 550000, 310000, 310000, 315000, 925000, 320000, 400000, 325000, 300000, 280
000, 240000, 335000, 200000, 300000, 445000, 700000, 400000, 615000, 240000, 400000, 225000, 700000, 95000, 390
000, 325000, 420000, 95000, 295000, 180000, 180000, 500000, 300000, 180000, 400000, 300000, 450000, 350000, 405
000, 145000, 300000, 240000, 335000, 315000, 160000, 300000, 330000, 260000, 400000, 300000, 600000, 335000, 14
5000, 370000, 390000, 290000, 340000, 105000, 240000, 100000, 300000, 280000, 410000, 310000, 390000, 360000, 3
10000, 335000, 415000, 320000, 550000, 220000, 480000, 300000, 280000]
```

In [220…
```
import random
n=30
df4=random.sample(df3,n)
print(df4)
```

```
[320000, 340000, 300000, 100000, 200000, 110000, 310000, 300000, 400000, 400000, 325000, 315000, 335000, 360000
, 445000, 180000, 450000, 375000, 450000, 360000, 420000, 95000, 240000, 350000, 300000, 250000, 360000, 320000
, 350000, 350000]
```

In [221…
```
def t_score(sample_size, sample_mean, pop_mean, sample_std):
    numerator = sample_mean - pop_mean
    denomenator = sample_std / sample_size**0.5
    return numerator / denomenator
```

In [222…
```
import statistics
from scipy.stats import t,norm
```

In [223…
```
sample_size = 100
sample_mean =332250.0
pop_mean = 300000
sample_std=89621.3
```

In [224…
```
t_val = t_score(sample_size, sample_mean, pop_mean, sample_std)

print(t_val)
```

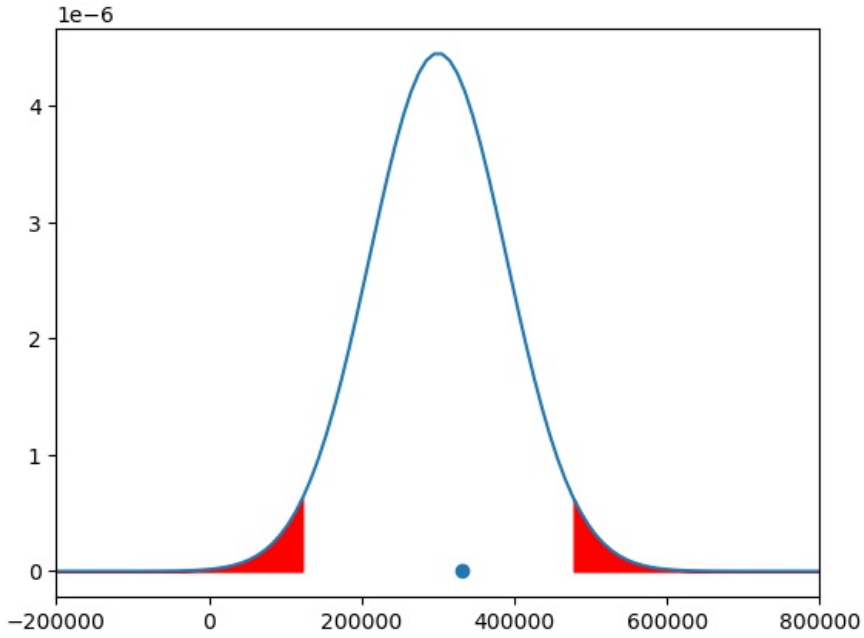
```
3.59847491611927
```

In [225…
```
confidence_level = 0.95

alpha = 1 - confidence_level

t_critical = t.ppf(1 - alpha/2,df=99)

print(t_critical)
```

```
1.9842169515086827
```

In [227…
```
h_min = -200000
h_max = 800000

mean = pop_mean
std = sample_std

x = np.linspace(h_min, h_max, 100)
y = norm.pdf(x, mean, std)
plt.xlim(h_min, h_max)
plt.plot(x, y)

t_critical_left = pop_mean + (-t_critical * std)
t_critical_right = pop_mean + (t_critical * std)

x1 = np.linspace(h_min, t_critical_left, 100)
y1 = norm.pdf(x1, mean, std)
plt.fill_between(x1, y1, color='red')

x2 = np.linspace(t_critical_right, h_max, 100)
y2 = norm.pdf(x2, mean, std)
plt.fill_between(x2, y2, color='red')

plt.scatter(sample_mean, 0)
plt.annotate("h_bar", (sample_mean, 0.7))
```

Out[227]:
```
Text(332250.0, 0.7, 'h_bar')
```




In [229…
```
if(t_val < t_critical):
    print("Reject Null Hypothesis")
else:
    print("Fail to reject Null Hypothesis")
```

```
Fail to reject Null Hypothesis
```

In [231…
```
p_value = 2 * (1.0 - norm.cdf(np.abs(t_val)))

print("p_value = ", p_value)

if(p_value > alpha):
    print("Reject Null Hypothesis")
else:
    print("Fail to reject Null Hypothesis")
```

```
p_value =  0.00032008871607791534
Fail to reject Null Hypothesis
```

- The claim that researchers is claiming is acceptable that salary range is up to 2.5 to 3 lakhs.

*Conclusion*:

- The dataset contains different salary range and different Designations and specializations, from that based on DOJ and DOL all the students are leaving the company in 2 years.
- All the columns data is cleaned and done Data analysis after cleaning, all the data is having different relation.
- From all the columns majority of Gender is having male and very less female is present in data set.
- After all the relation of univariate and bivariate analysis plots shows the relationship between all the columns and outliers in the dataset.
- After Exploratory data analysis the research based question on computerscience graduate salary is taken and done hypothesis testing the given mean salary is accepting or rejecting. The Hypothesis testing shows the claim that researchers are acceptable.

- The average salary for a graduates is having above 35k and below 50k, and with collegeGPA of 60% to 80%.

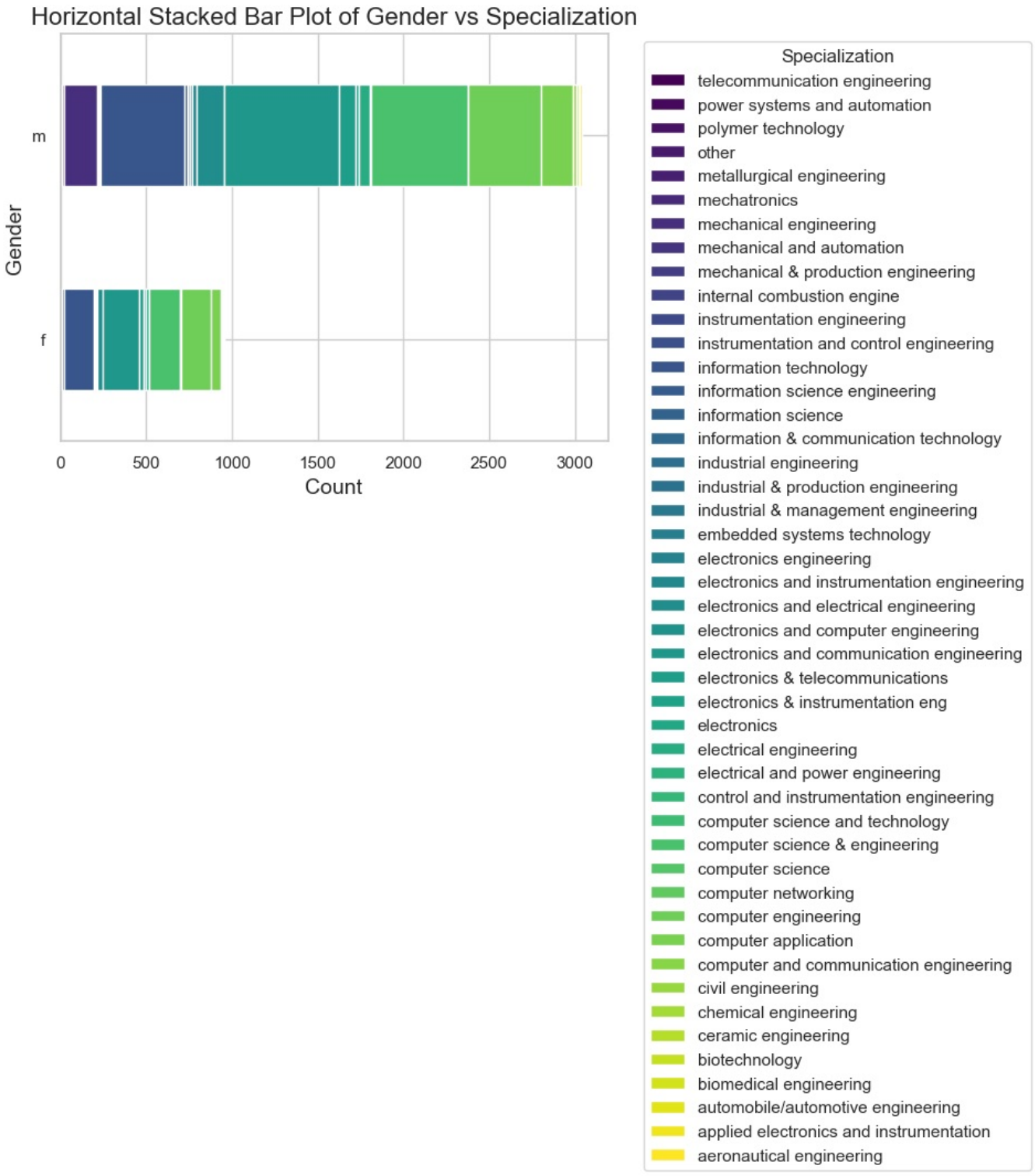
- Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)

In [234…
```
from scipy.stats import chi2_contingency

# Create a contingency table
con = pd.crosstab(data['Gender'], data['Specialization'])

# Perform chi-square test of independence
chi2, p_value, dof, expected = chi2_contingency(con)

# Determine if null hypothesis is rejected
alpha = 0.05
if p_value < alpha:
    print("Reject Null Hypothesis")
    print("There is a relationship between gender and specialization.")
else:
    print("Do Not Reject Null Hypothesis")
    print("There is no significant relationship between gender and specialization.")
    print("Reject Null Hypothesis")
    print("There is a relationship between gender and specialization.")
```

```
Reject Null Hypothesis
There is a relationship between gender and specialization.
```

In [240…
```
import matplotlib.pyplot as plt
import seaborn as sns

# Create a contingency table
con = pd.crosstab(data['Gender'], data['Specialization'])

# Plotting a horizontal stacked bar plot
plt.figure(figsize=(12, 8))
sns.set(style="whitegrid")

# Reverse the order of columns to have Specialization on one side
con = con[con.columns[::-1]]

con.plot(kind='barh', stacked=True, cmap='viridis')
plt.title('Horizontal Stacked Bar Plot of Gender vs Specialization', fontsize=16)
plt.xlabel('Count', fontsize=14)
plt.ylabel('Gender', fontsize=14)
plt.legend(title='Specialization', title_fontsize='12', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

```
<Figure size 1200x800 with 0 Axes>
```




*Conclusion*

- The dataset showcases a broad spectrum of salary ranges, encompassing a substantial number of outliers, indicative of a wide array of income levels.
- Educational performance displays a moderate variance, with some individuals attaining lower scores in 10th and 12th percentages.
- Personality traits exhibit diversity, encompassing varying degrees of conscientiousness, agreeableness, extraversion, neuroticism, and openness.
- A noticeable gender imbalance is observed, particularly in degree choices and college states.
- Positive correlations are evident between salary and factors such as college tier, GPA, and the domain of study.
- However, no distinct correlations emerge between salary and English scores, or personality traits like conscientiousness and agreeableness.
- The dataset underscores the diversity in educational backgrounds, career trajectories, and gender representation within the sampled population.

# Some Insights regarding data

*SalaryDistribution*:

- The dataset exhibits a broad salary range, spanning from 35,000 to 4,000,000, with an average (mean) salary of approximately 307,700 and a median of 300,000.
- A notable dispersion is observed in salary values around the mean, evidenced by a standard deviation of approximately 212,700.
- The presence of numerous outliers suggests substantial variability in salary levels within the dataset.

*EducationalPerformanceAnalysis*:

- There is a moderate range of variability in academic achievements.
- The mean percentages for 10th and 12th grades are around 77.9 and 74.5, respectively, with standard deviations of approximately 9.9 and 11.0.
- The average college GPA stands at 71.5, accompanied by a standard deviation of around 8.2.
- Notably, there are outliers, especially in 10th and 12th percentages, suggesting the presence of individuals with lower academic scores.

*AnalysisofPersonalityTraits*:

- There is noticeable variability in the distribution of personality traits around their respective means.
- The measured traits include conscientiousness, agreeableness, extraversion, neuroticism, and openness to experience.
- As an illustration, conscientiousness spans from roughly -4.13 to 1.99, with a mean in close proximity to 0 and a standard deviation of about 1.03.

*GenderComposition*:

- Within the dataset, there are two gender categories: 'm' (indicating male) and 'f' (indicating female).
- Around 76.1% of the individuals identify as male, whereas approximately 23.9% identify as female.

*Degree*

- 'B.Tech/B.E.' dominates the dataset, constituting approximately 92.5% of the degrees.
- 'MCA' follows as the second most prevalent, making up about 6.1%.
- 'M.Tech./M.E.' and 'M.Sc. (Tech.)' have lower representation, collectively accounting for about 1.4%.

*Specialization*:

- The dataset encompasses 46 distinct specializations.
- The most prevalent specialization is 'Electronics and Communication Engineering', followed by 'Computer Science & Engineering' and 'Information Technology'.

*GenderandDegree/Specialization/CollegeStateDistribution*:

- Males predominate in 'B.Tech/B.E.', 'MCA', 'Computer Engineering', 'Information Technology', 'Automobile/Automotive Engineering', and 'Electronics Engineering'.
- Females dominate in 'Biomedical Engineering' and 'Information & Communication Technology'.
- Some fields exhibit an equal gender split, such as 'Computer Science' and 'Telecommunication Engineering'.
- College states like Goa, Meghalaya, Andhra Pradesh, Gujarat, and Telangana demonstrate a higher male representation.

In [ ]: