amcat-task

October 3, 2024

[13]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

[14]:
```
df=pd.read_csv('data.xlsx - Sheet1.csv')
df.head()
```

[14]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	…	ComputerScience	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	conscientiousness	agreeableness	extraversion	nueroticism
0	train	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	…	-1	-1	-1	-1	-1	0.9737	0.8128	0.5269	1.35490
1	train	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	…	-1	-1	-1	-1	-1	-0.7335	0.3789	1.2396	-0.10760
2	train	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	…	-1	-1	-1	-1	-1	0.2718	1.7109	0.1637	-0.86820
3	train	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	…	-1	-1	-1	-1	-1	0.0464	0.3448	-0.3440	-0.40780
4	train	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.0	…	-1	-1	-1	-1	-1	-0.8810	-0.2793	-1.0697	0.09163

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

[5 rows x 39 columns]
```

[ ]:
```
df.tail()
```

[ ]:
```
     Unnamed: 0      ID    Salary           DOJ           DOL  \
3993      train   47916  280000.0  10/1/11 0:00  10/1/12 0:00
3994      train  752781  100000.0   7/1/13 0:00   7/1/13 0:00
3995      train  355888  320000.0   7/1/13 0:00       present
3996      train  947111  200000.0   7/1/14 0:00   1/1/15 0:00
3997      train  324966  400000.0   2/1/13 0:00       present

                     Designation           JobCity Gender           DOB  \
3993            software engineer        New Delhi       m  4/15/87 0:00
3994             technical writer         Hyderabad      f  8/27/92 0:00
3995  associate software engineer         Bangalore      m   7/3/91 0:00
3996          software developer  Asifabadbanglore      f  3/20/92 0:00
3997     senior systems engineer           Chennai      f  2/26/91 0:00

      10percentage  … ComputerScience  MechanicalEngg  ElectricalEngg  \
3993         52.09  …              -1              -1              -1
3994         90.00  …              -1              -1              -1
3995         81.86  …              -1              -1              -1
3996         78.72  …             438              -1              -1
3997         70.60  …              -1              -1              -1

     TelecomEngg  CivilEngg  conscientiousness agreeableness extraversion  \
3993          -1         -1            -0.1082        0.3448       0.2366
3994          -1         -1            -0.3027        0.8784       0.9322
3995          -1         -1            -1.5765       -1.5273      -1.5051
3996          -1         -1            -0.1590        0.0459      -0.4511
3997          -1         -1            -1.1128       -0.2793      -0.6343

      nueroticism  openess_to_experience
3993      0.64980                -0.9194
3994      0.77980                -0.0943
3995     -1.31840                -0.7615
3996     -0.36120                -0.0943
3997      1.32553                -0.6035

[5 rows x 39 columns]
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 39 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Unnamed: 0             3998 non-null   object 
 1   ID                     3998 non-null   int64  
 2   Salary                 3998 non-null   float64
 3   DOJ                    3998 non-null   object 
 4   DOL                    3998 non-null   object 
 5   Designation            3998 non-null   object 
 6   JobCity                3998 non-null   object 
 7   Gender                 3998 non-null   object 
 8   DOB                    3998 non-null   object 
 9   10percentage           3998 non-null   float64
 10  10board                3998 non-null   object 
 11  12graduation           3998 non-null   int64  
 12  12percentage           3998 non-null   float64
 13  12board                3998 non-null   object 
 14  CollegeID              3998 non-null   int64  
 15  CollegeTier            3998 non-null   int64  
 16  Degree                 3998 non-null   object 
 17  Specialization         3998 non-null   object 
 18  collegeGPA             3998 non-null   float64
 19  CollegeCityID          3998 non-null   int64  
 20  CollegeCityTier        3998 non-null   int64  
 21  CollegeState           3998 non-null   object 
 22  GraduationYear         3998 non-null   int64  
 23  English                3998 non-null   int64  
 24  Logical                3998 non-null   int64  
 25  Quant                  3998 non-null   int64  
 26  Domain                 3998 non-null   float64
 27  ComputerProgramming    3998 non-null   int64  
 28  ElectronicsAndSemicon  3998 non-null   int64  
 29  ComputerScience        3998 non-null   int64  
 30  MechanicalEngg         3998 non-null   int64  
 31  ElectricalEngg         3998 non-null   int64  
 32  TelecomEngg            3998 non-null   int64  
 33  CivilEngg              3998 non-null   int64  
 34  conscientiousness      3998 non-null   float64
 35  agreeableness          3998 non-null   float64
 36  extraversion           3998 non-null   float64
 37  nueroticism            3998 non-null   float64
 38  openess_to_experience  3998 non-null   float64
dtypes: float64(10), int64(17), object(12)
```

```
memory usage: 1.2+ MB
```

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

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

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

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

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

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

```
          CivilEngg  conscientiousness  agreeableness  extraversion  \
count  3998.000000        3998.000000    3998.000000   3998.000000
mean      2.683842          -0.037831       0.146496      0.002763
std      36.658505           1.028666       0.941782      0.951471
min      -1.000000          -4.126700      -5.781600     -4.600900
25%      -1.000000          -0.713525      -0.287100     -0.604800
50%      -1.000000           0.046400       0.212400      0.091400
75%      -1.000000           0.702700       0.812800      0.672000
max     516.000000           1.995300       1.904800      2.535400

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

[8 rows x 27 columns]
```

[ ]:
```
df.shape
```

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

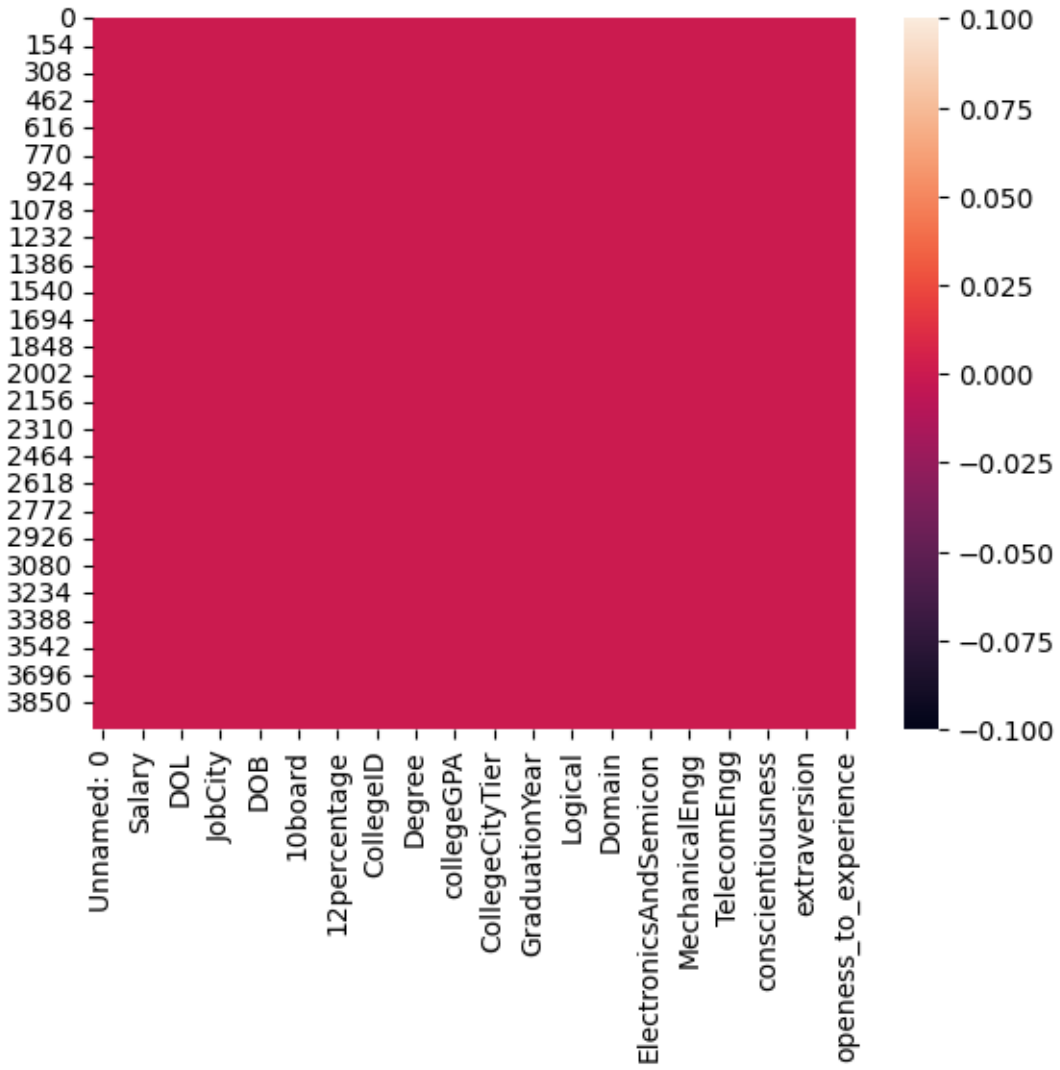
[ ]:
```
#checking the null values
df.isnull().sum()
```

[ ]:
```
Unnamed: 0             0
ID                     0
Salary                 0
DOJ                    0
DOL                    0
Designation            0
JobCity                0
Gender                 0
DOB                    0
10percentage           0
10board                0
12graduation           0
12percentage           0
12board                0
CollegeID              0
CollegeTier            0
Degree                 0
Specialization         0
```

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

[ ]:
```
sns.heatmap(df.isnull())
```

[ ]:
```
<Axes: >
```



Observation: we can see that in above heatmap there is no null value present in our data if there an another color part occur in our heatmap then we can say that there are many null value occurs in our data

[ ]:
```
#Univariate - Non Visual Statistical Analysis
def numerical_univariate_analysis(numerical_data):
    for col_name in numerical_data:
        print("*"*10,col_name,"*"*10)
        print(numerical_data[col_name].
 ↪agg(['min','max','mean','median','std','skew','kurt']))
        print()
```

[ ]:
```
numerical_univariate_analysis(df[['10percentage','12percentage','collegeGPA','Salary']])
```

```
********** 10percentage **********
```

```
min       43.000000
max       97.760000
mean      77.925443
median    79.150000
std        9.850162
skew      -0.591019
kurt      -0.110284
Name: 10percentage, dtype: float64

********** 12percentage **********
min       40.000000
max       98.700000
mean      74.466366
median    74.400000
std       10.999933
skew      -0.032607
kurt      -0.630737
Name: 12percentage, dtype: float64

********** collegeGPA **********
min        6.450000
max       99.930000
mean      71.486171
median    71.720000
std        8.167338
skew      -1.249209
kurt      10.234244
Name: collegeGPA, dtype: float64

********** Salary **********
min       3.500000e+04
max       4.000000e+06
mean      3.076998e+05
median    3.000000e+05
std       2.127375e+05
skew      6.451081e+00
kurt      8.093000e+01
Name: Salary, dtype: float64
```

[ ]:
```python
#Univariate - Non Visual Statistical Analysis
def discrete_univariate_analysis(discrete_data):
    for col_name in discrete_data:
        print("*"*10, col_name, "*"*10)
        print(discrete_data[col_name].agg(['count', 'nunique', 'unique']))
        print('Value Counts: \n', discrete_data[col_name].value_counts())
        print()
```

```
[ ]: discrete_univariate_analysis(df[['Designation','JobCity','Gender']])
```

```
********** Designation **********
count                                                    3998
nunique                                                   419
unique     [senior quality engineer, assistant manager, s…
Name: Designation, dtype: object
Value Counts:
 Designation
software engineer                    539
software developer                   265
system engineer                      205
programmer analyst                   139
systems engineer                     118
                                    …
cad drafter                            1
noc engineer                           1
human resources intern                 1
senior quality assurance engineer      1
jr. software developer                 1
Name: count, Length: 419, dtype: int64

********** JobCity **********
count                                                    3998
nunique                                                   339
unique     [Bangalore, Indore, Chennai, Gurgaon, Manesar,…
Name: JobCity, dtype: object
Value Counts:
 JobCity
Bangalore           627
-1                  461
Noida               368
Hyderabad           335
Pune                290
                   …
Tirunelvelli          1
Ernakulam             1
Nanded                1
Dharmapuri            1
Asifabadbanglore      1
Name: count, Length: 339, dtype: int64

********** Gender **********
count        3998
nunique         2
unique     [f, m]
Name: Gender, dtype: object
```

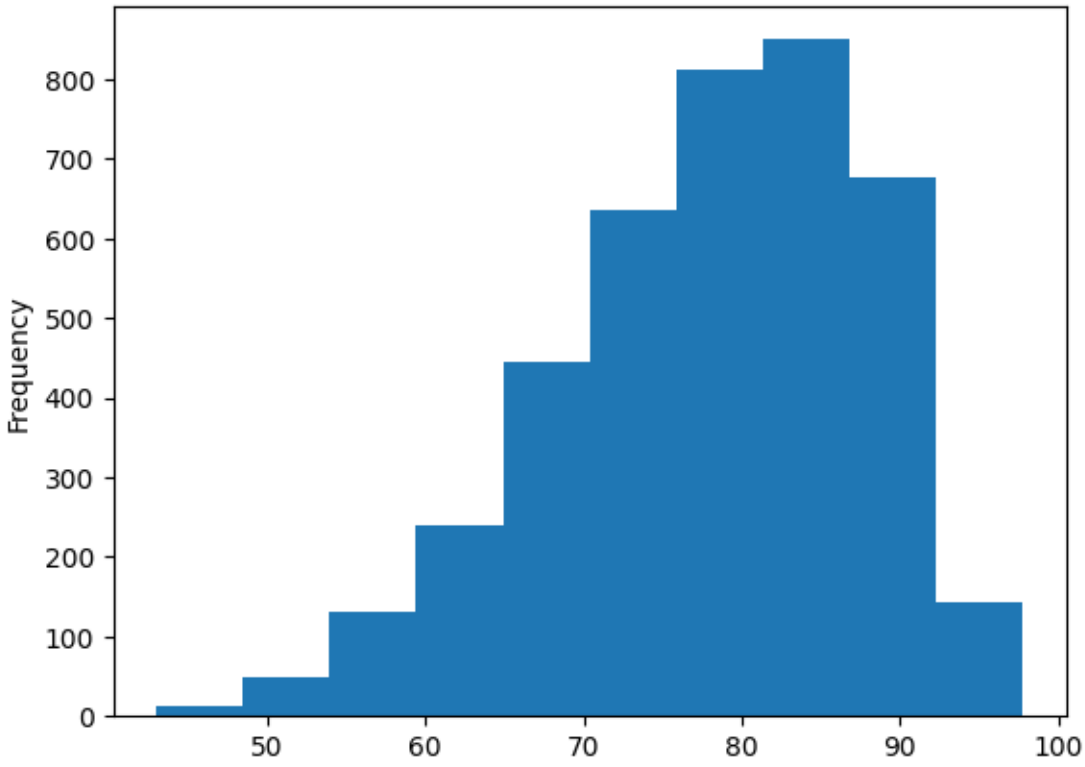
```
Value Counts:
 Gender
m    3041
f     957
Name: count, dtype: int64
```

[ ]:

##Univariate Plotting For Numerical Columns - Histogram, pdf and Box Plot

[ ]:
```
df['10percentage'].plot(kind='hist')
```

[ ]:
```
<Axes: ylabel='Frequency'>
```

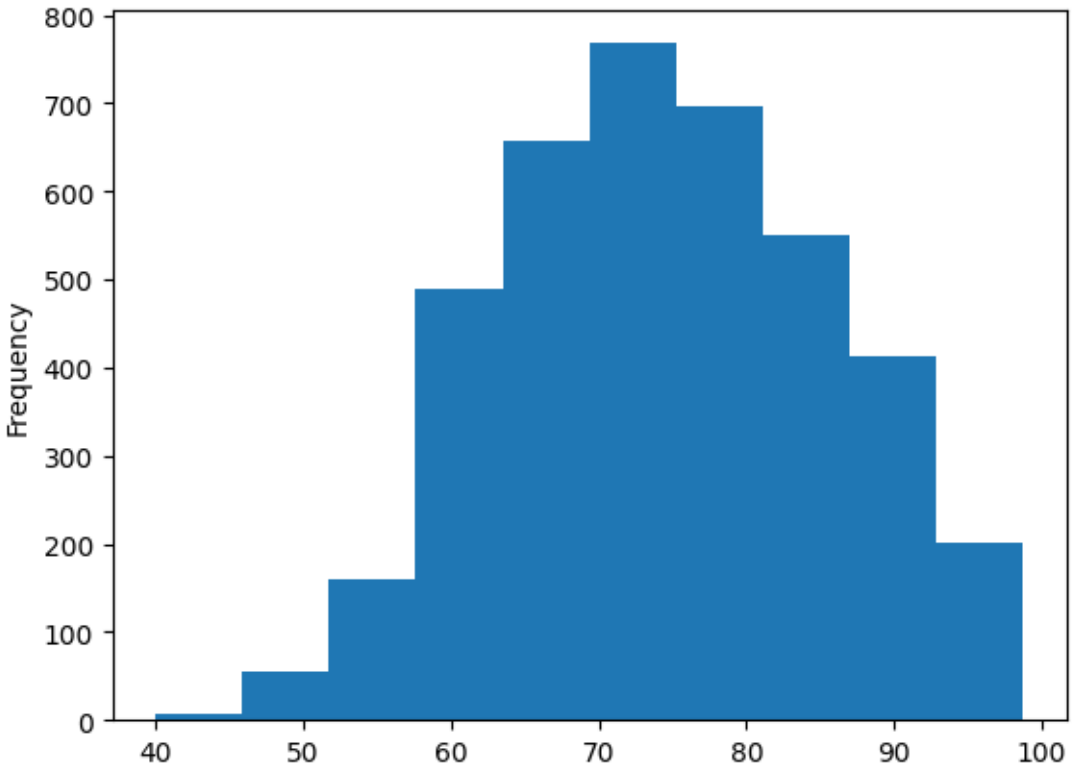


[ ]:
```
#Observation: 1.Maximum 10 percentage occurs between 70 to 90
```

[ ]:

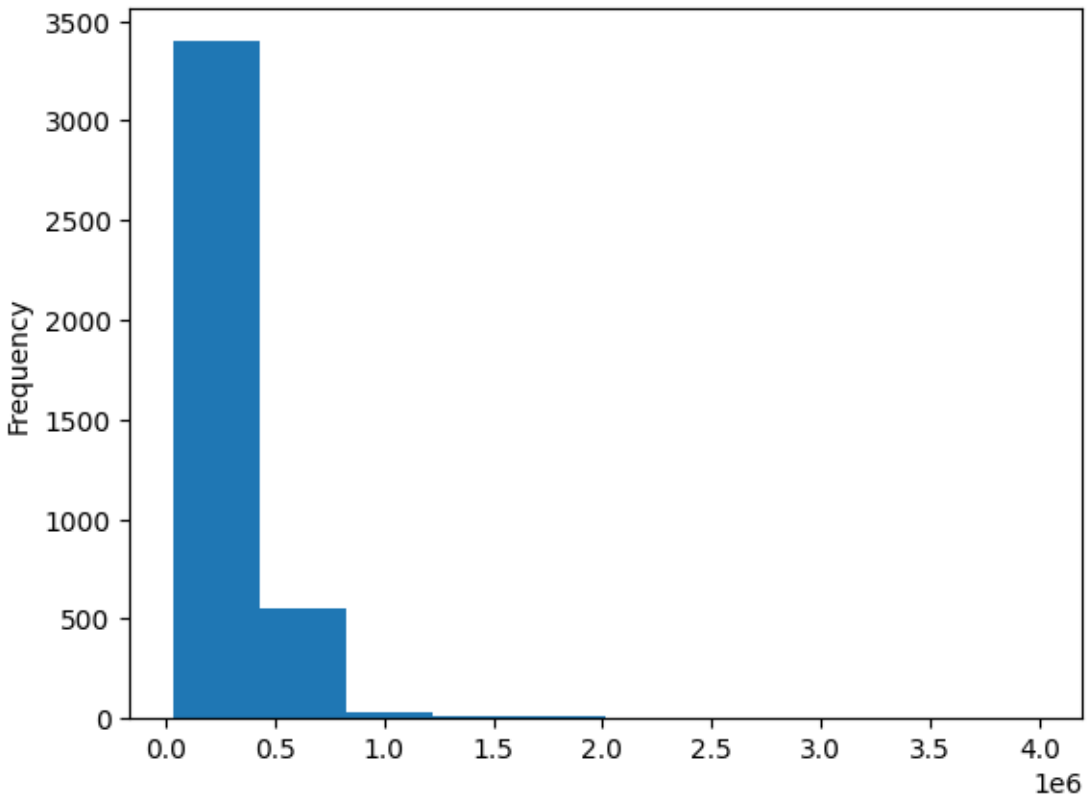
[ ]:
```
df['12percentage'].plot(kind='hist')
```

[ ]:
```
<Axes: ylabel='Frequency'>
```



[ ]:
```
df['Salary'].plot(kind='hist')
```

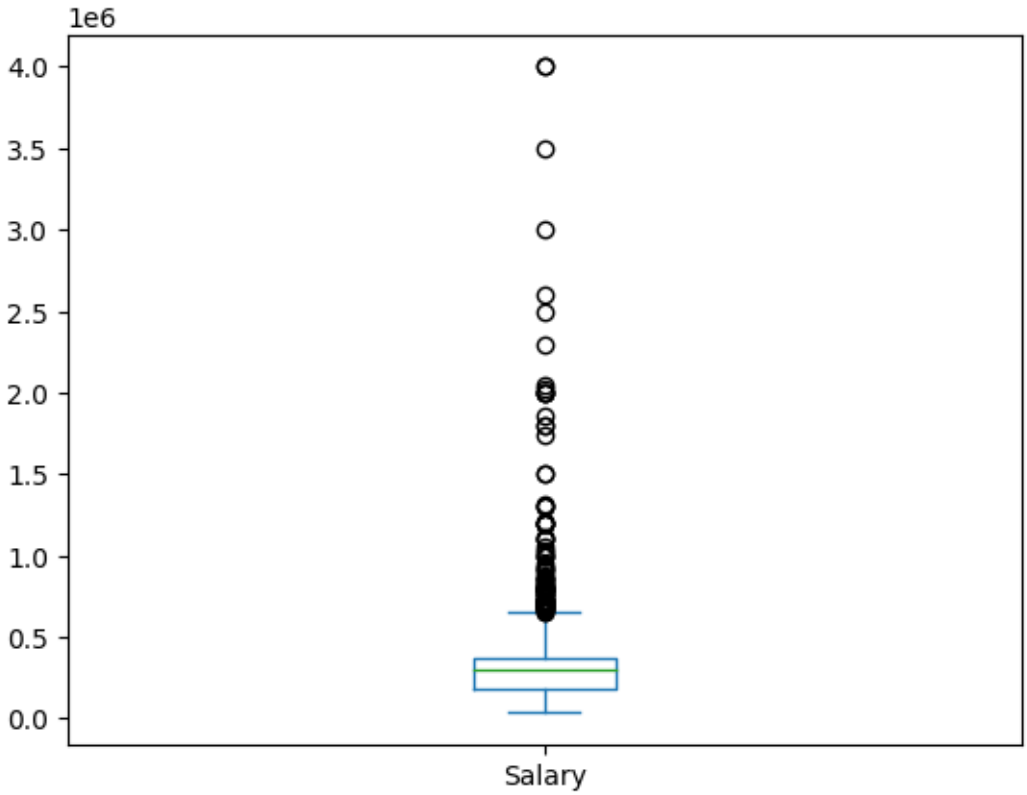
[ ]:
```
<Axes: ylabel='Frequency'>
```



Observation:Maximum salary occur between 0.1 to 0.4 and from 1.0 to 4.0 there is no salary value occurs

[ ]:
```
#here we can use the box plot on the univarite numerical data

df['Salary'].plot(kind='box')
```

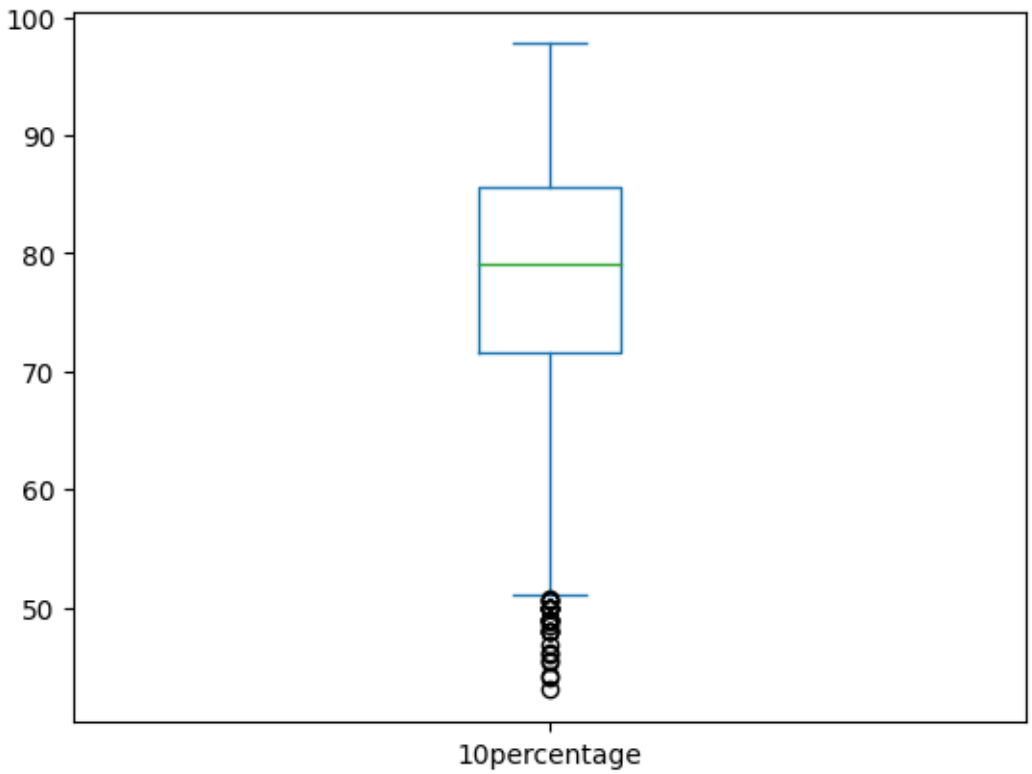
[ ]: <Axes: >



Observation: The boxplot indicates that most salaries are between 0.5 and 1 million INR, with a median around 0.75 million INR, and some candidates have much higher salaries as outliers above 4 million INR

[ ]:
```
df['10percentage'].plot(kind='box')
```

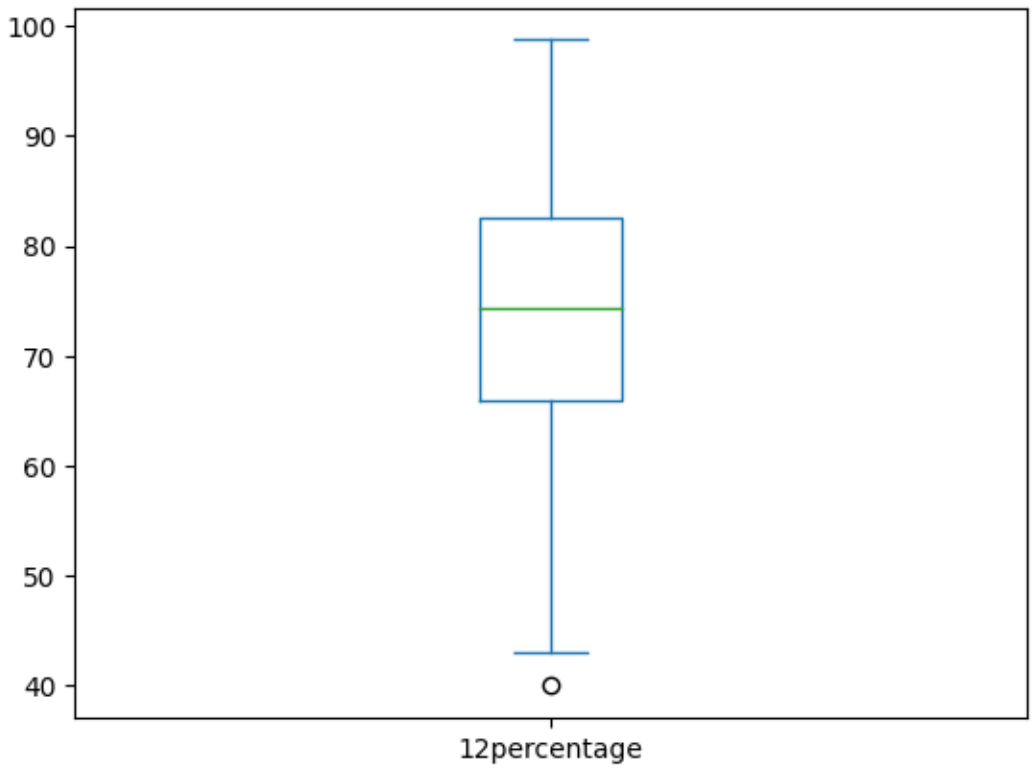
[ ]:
```
<Axes: >
```



Observation:This boxplot indicates that “10percentage” data falls between approximately 70% and 88% . There are a significant number of outliers on the lower end, suggesting some unusual occurrences in that range.

[ ]:
```
df['12percentage'].plot(kind='box')
```

[ ]: <Axes: >



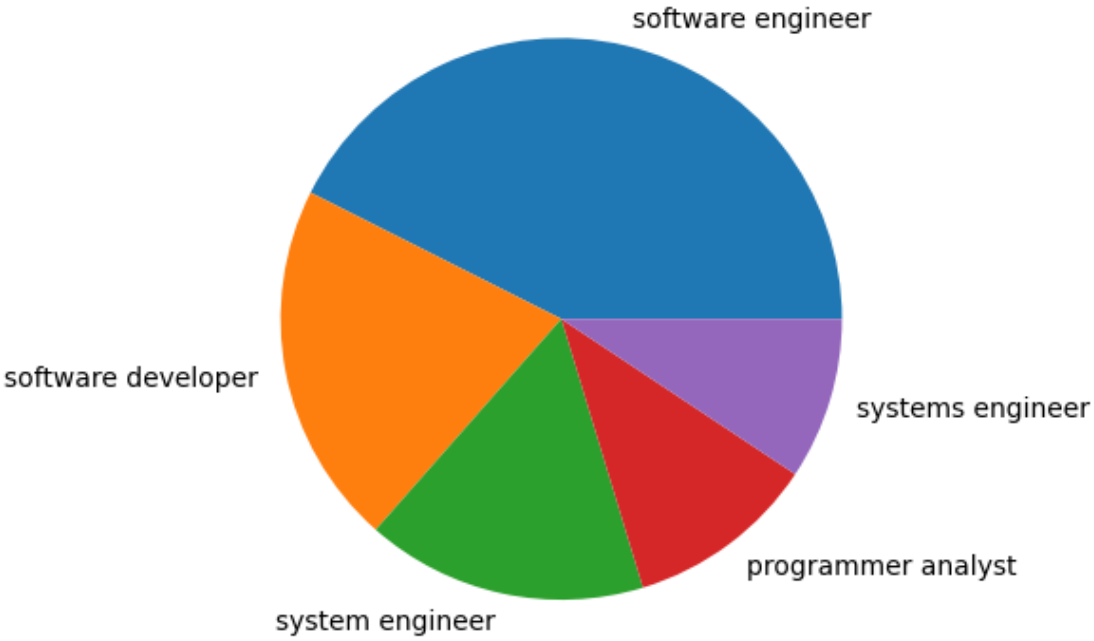
Observation: Maximum percentage occurs between 65 % to 82% At the lower end there is outlier value occurs

##Univariate Plotting For categorical Columns - piechart

[ ]:
```
Designation_name=df.Designation.value_counts().index
Designation_value=df.Designation.value_counts().values
```

[ ]:
```
plt.pie(Designation_value[:5],labels=Designation_name[:5])
```

[ ]:
```
([<matplotlib.patches.Wedge at 0x7be221238f40>,
  <matplotlib.patches.Wedge at 0x7be221238e20>,
  <matplotlib.patches.Wedge at 0x7be221239810>,
  <matplotlib.patches.Wedge at 0x7be221239c90>,
  <matplotlib.patches.Wedge at 0x7be22123a110>],
 [Text(0.25426762753772525, 1.070209312978698, 'software engineer'),
  Text(-1.079980418855396, -0.20890738351940463, 'software developer'),
  Text(-0.2329730153737593, -1.0750458474444977, 'system engineer'),
  Text(0.6571177481545666, -0.8821543317698279, 'programmer analyst'),
  Text(1.0531776140862001, -0.317516791977525, 'systems engineer')])
```

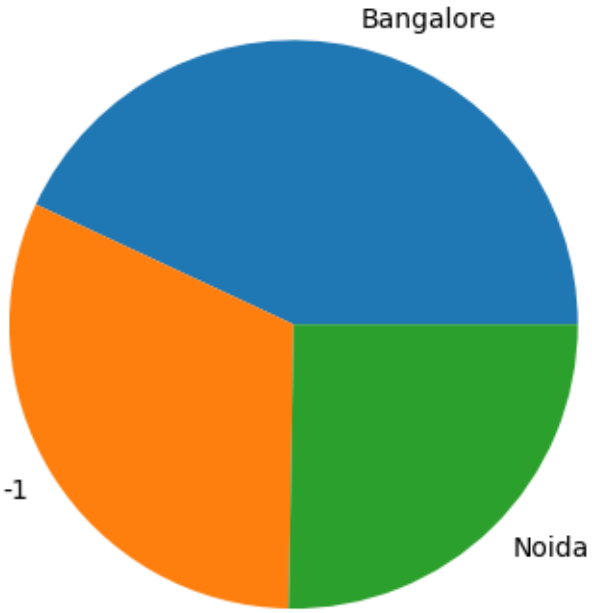


Observation: In above pie chart we can see the top 5 Designation in given data

[ ]:
```
jobcity_name=df.JobCity.value_counts().index
jobcity_values=df.JobCity.value_counts().values
```

[ ]:
```
plt.pie(jobcity_values[:3],labels=jobcity_name[:3])
```

[ ]:
```
([<matplotlib.patches.Wedge at 0x7be2210a44f0>,
  <matplotlib.patches.Wedge at 0x7be2210a43d0>,
  <matplotlib.patches.Wedge at 0x7be2210a4dc0>],
 [Text(0.23782607579436885, 1.0739826617186383, 'Bangalore'),
  Text(-0.9326571627742322, -0.5832243279613082, '-1'),
  Text(0.7710755363658346, -0.7845014450070443, 'Noida')])
```



Observation: Here we can see top 3 job cities are present in given data

[ ]:
```
df['JobCity'].value_counts()
```

[ ]:
```
JobCity
Bangalore           627
-1                  461
Noida               368
Hyderabad           335
Pune                290
                   ...
Tirunelvelli          1
Ernakulam             1
Nanded                1
Dharmapuri            1
Asifabadbanglore      1
Name: count, Length: 339, dtype: int64
```

[ ]:
```
df['JobCity'].unique()
```

[ ]:
```
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)
```

[ ]:
```
df['JobCity'].nunique()
```

[ ]:
```
339
```

##Bivariate Analysis

[23]:
```
discrete_data = df.select_dtypes(include=['object'])

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

#Bivariate - Non Visual Statistical Analysis

[ ]:
```
#numerical vs numerical
numerical_data.corr()
```

[ ]:

	ID	Salary	10percentage	12graduation	\
ID	1.000000	-0.247294	0.044547	0.673102	
Salary	-0.247294	1.000000	0.177373	-0.161383	
10percentage	0.044547	0.177373	1.000000	0.269957	
12graduation	0.673102	-0.161383	0.269957	1.000000	
12percentage	0.007069	0.170254	0.643378	0.259166	
CollegeID	0.284540	-0.118690	0.021082	0.254021	
CollegeTier	0.035160	-0.179332	-0.126042	0.027691	
collegeGPA	0.047144	0.130103	0.312538	0.086001	
CollegeCityID	0.284540	-0.118690	0.021082	0.254021	
CollegeCityTier	-0.035977	0.015384	0.116707	-0.003016	
GraduationYear	0.027539	-0.010053	-0.013799	0.014457	
English	0.135505	0.178219	0.350780	0.147925	
Logical	0.102215	0.179275	0.316014	0.105887	

```
Quant                  -0.055134  0.230627      0.317640      0.001379
Domain                 -0.125639  0.104656      0.078563     -0.034163
ComputerProgramming     0.018859  0.115665      0.053600     -0.047995
ElectronicsAndSemicon -0.115601  0.000665      0.085179     -0.005891
ComputerScience         0.482626 -0.100720     -0.018933      0.293439
MechanicalEngg         -0.026147  0.018475      0.050364      0.035459
ElectricalEngg          0.104454 -0.047598      0.074419      0.123751
TelecomEngg            -0.049272 -0.022691      0.049378      0.023470
CivilEngg              -0.017871  0.037639      0.030002     -0.004727
conscientiousness       0.175557 -0.064148      0.067657      0.103329
agreeableness           0.024837  0.057423      0.136645      0.041182
extraversion            0.120979 -0.010213     -0.004679      0.061956
nueroticism            -0.146289 -0.054685     -0.132496     -0.074369
openess_to_experience   0.031359 -0.011312      0.036692     -0.015069

                       12percentage  CollegeID  CollegeTier  collegeGPA  \
ID                         0.007069   0.284540     0.035160    0.047144
Salary                     0.170254  -0.118690    -0.179332    0.130103
10percentage               0.643378   0.021082    -0.126042    0.312538
12graduation               0.259166   0.254021     0.027691    0.086001
12percentage               1.000000   0.022336    -0.100771    0.346137
CollegeID                  0.022336   1.000000     0.067054    0.017240
CollegeTier               -0.100771   0.067054     1.000000   -0.086781
collegeGPA                 0.346137   0.017240    -0.086781    1.000000
CollegeCityID              0.022336   1.000000     0.067054    0.017240
CollegeCityTier            0.130462   0.007757    -0.101494    0.017471
GraduationYear            -0.012933  -0.000172    -0.005557    0.008706
English                    0.212888  -0.022792    -0.183843    0.106478
Logical                    0.243571  -0.047094    -0.182811    0.196610
Quant                      0.312413  -0.114672    -0.251103    0.217380
Domain                     0.074099  -0.073857    -0.061436    0.107252
ComputerProgramming        0.080818  -0.033760    -0.073644    0.136596
ElectronicsAndSemicon      0.117112  -0.020438    -0.031573    0.029855
ComputerScience           -0.043534   0.102303     0.001053    0.007601
MechanicalEngg             0.037635  -0.009291    -0.021548   -0.031765
ElectricalEngg             0.064001   0.022933     0.002594    0.052258
TelecomEngg                0.044201   0.025620     0.000007   -0.005226
CivilEngg                  0.005910   0.005749    -0.033722   -0.018950
conscientiousness          0.058299   0.076432     0.055174    0.069582
agreeableness              0.103998  -0.005264    -0.038055    0.068282
extraversion              -0.007486   0.005917     0.009970   -0.032684
nueroticism               -0.094369  -0.008973     0.023778   -0.074859
openess_to_experience      0.006332  -0.010678    -0.019179    0.028071

                       CollegeCityID  CollegeCityTier  …  ComputerScience  \
ID                          0.284540        -0.035977  …         0.482626
Salary                     -0.118690         0.015384  …        -0.100720
```

10percentage	0.021082	0.116707	…	-0.018933
12graduation	0.254021	-0.003016	…	0.293439
12percentage	0.022336	0.130462	…	-0.043534
CollegeID	1.000000	0.007757	…	0.102303
CollegeTier	0.067054	-0.101494	…	0.001053
collegeGPA	0.017240	0.017471	…	0.007601
CollegeCityID	1.000000	0.007757	…	0.102303
CollegeCityTier	0.007757	1.000000	…	-0.010643
GraduationYear	-0.000172	0.008152	…	0.024089
English	-0.022792	0.050462	…	0.059500
Logical	-0.047094	0.020353	…	0.044481
Quant	-0.114672	0.007896	…	-0.043379
Domain	-0.073857	0.009250	…	0.058762
ComputerProgramming	-0.033760	0.064272	…	0.253039
ElectronicsAndSemicon	-0.020438	0.041083	…	-0.273707
ComputerScience	0.102303	-0.010643	…	1.000000
MechanicalEngg	-0.009291	-0.052395	…	-0.124355
ElectricalEngg	0.022933	0.010311	…	-0.083798
TelecomEngg	0.025620	0.049876	…	-0.148095
CivilEngg	0.005749	-0.033392	…	-0.052613
conscientiousness	0.076432	0.014763	…	0.090155
agreeableness	-0.005264	0.005565	…	0.039866
extraversion	0.005917	-0.008203	…	0.102153
nueroticism	-0.008973	0.004442	…	-0.112652
openess_to_experience	-0.010678	-0.016790	…	0.058039

	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	\
ID	-0.026147	0.104454	-0.049272	-0.017871	
Salary	0.018475	-0.047598	-0.022691	0.037639	
10percentage	0.050364	0.074419	0.049378	0.030002	
12graduation	0.035459	0.123751	0.023470	-0.004727	
12percentage	0.037635	0.064001	0.044201	0.005910	
CollegeID	-0.009291	0.022933	0.025620	0.005749	
CollegeTier	-0.021548	0.002594	0.000007	-0.033722	
collegeGPA	-0.031765	0.052258	-0.005226	-0.018950	
CollegeCityID	-0.009291	0.022933	0.025620	0.005749	
CollegeCityTier	-0.052395	0.010311	0.049876	-0.033392	
GraduationYear	-0.066844	0.008525	0.004226	0.001696	
English	-0.002477	0.032438	-0.005822	-0.007724	
Logical	-0.009861	0.012003	-0.012947	-0.011286	
Quant	0.019933	0.020975	0.021387	0.000528	
Domain	0.048472	0.042875	0.024442	0.017569	
ComputerProgramming	-0.284891	-0.138224	-0.248269	-0.088249	
ElectronicsAndSemicon	-0.109434	0.036968	0.387140	0.002863	
ComputerScience	-0.124355	-0.083798	-0.148095	-0.052613	
MechanicalEngg	1.000000	-0.040522	-0.070947	0.076201	
ElectricalEngg	-0.040522	1.000000	-0.051469	-0.020059	

TelecomEngg	-0.070947	-0.051469	1.000000	-0.031492
CivilEngg	0.076201	-0.020059	-0.031492	1.000000
conscientiousness	-0.010858	0.029806	-0.004946	-0.017526
agreeableness	-0.028586	-0.015454	-0.014627	-0.034254
extraversion	-0.017748	0.004467	-0.039050	-0.031822
nueroticism	0.036148	-0.030870	0.020638	0.010555
openess_to_experience	-0.027988	-0.012585	-0.000141	-0.031201

	conscientiousness	agreeableness	extraversion	\
ID	0.175557	0.024837	0.120979	
Salary	-0.064148	0.057423	-0.010213	
10percentage	0.067657	0.136645	-0.004679	
12graduation	0.103329	0.041182	0.061956	
12percentage	0.058299	0.103998	-0.007486	
CollegeID	0.076432	-0.005264	0.005917	
CollegeTier	0.055174	-0.038055	0.009970	
collegeGPA	0.069582	0.068282	-0.032684	
CollegeCityID	0.076432	-0.005264	0.005917	
CollegeCityTier	0.014763	0.005565	-0.008203	
GraduationYear	-0.013235	-0.002877	0.008397	
English	0.034943	0.194990	0.018755	
Logical	0.025876	0.167207	-0.006949	
Quant	-0.005639	0.103443	-0.028616	
Domain	-0.039478	0.051944	-0.024647	
ComputerProgramming	0.012862	0.076934	0.043504	
ElectronicsAndSemicon	-0.026483	-0.024286	-0.044458	
ComputerScience	0.090155	0.039866	0.102153	
MechanicalEngg	-0.010858	-0.028586	-0.017748	
ElectricalEngg	0.029806	-0.015454	0.004467	
TelecomEngg	-0.004946	-0.014627	-0.039050	
CivilEngg	-0.017526	-0.034254	-0.031822	
conscientiousness	1.000000	0.481820	0.355537	
agreeableness	0.481820	1.000000	0.454369	
extraversion	0.355537	0.454369	1.000000	
nueroticism	-0.330312	-0.207480	-0.096491	
openess_to_experience	0.395649	0.591541	0.435074	

	nueroticism	openess_to_experience
ID	-0.146289	0.031359
Salary	-0.054685	-0.011312
10percentage	-0.132496	0.036692
12graduation	-0.074369	-0.015069
12percentage	-0.094369	0.006332
CollegeID	-0.008973	-0.010678
CollegeTier	0.023778	-0.019179
collegeGPA	-0.074859	0.028071
CollegeCityID	-0.008973	-0.010678

```
CollegeCityTier           0.004442              -0.016790
GraduationYear           -0.000417               0.016855
English                  -0.155528               0.067979
Logical                  -0.178781               0.048420
Quant                    -0.131895               0.020377
Domain                   -0.017928               0.010412
ComputerProgramming      -0.084344               0.043133
ElectronicsAndSemicon     0.021026              -0.013460
ComputerScience          -0.112652               0.058039
MechanicalEngg            0.036148              -0.027988
ElectricalEngg           -0.030870              -0.012585
TelecomEngg               0.020638              -0.000141
CivilEngg                 0.010555              -0.031201
conscientiousness        -0.330312               0.395649
agreeableness            -0.207480               0.591541
extraversion             -0.096491               0.435074
nueroticism               1.000000              -0.065795
openess_to_experience    -0.065795               1.000000

[27 rows x 27 columns]
```

[ ]:
```
#numerical vs categorical
data=df.groupby('Gender')
data['collegeGPA'].agg(['min','max','mean'])
```

[ ]:
```
          min    max       mean
Gender
f        9.30  99.93  74.048056
m        6.45  98.40  70.679947
```

[ ]:
```
demo=df.groupby('Gender')
demo['Salary'].agg(['min','max','mean'])
```

[ ]:
```
             min        max           mean
Gender
f       35000.0  3500000.0  294937.304075
m       35000.0  4000000.0  311716.211772
```

[ ]:
```
#categorical vs categorical
pd.crosstab(df['Gender'],df['Designation'],normalize='index')
```

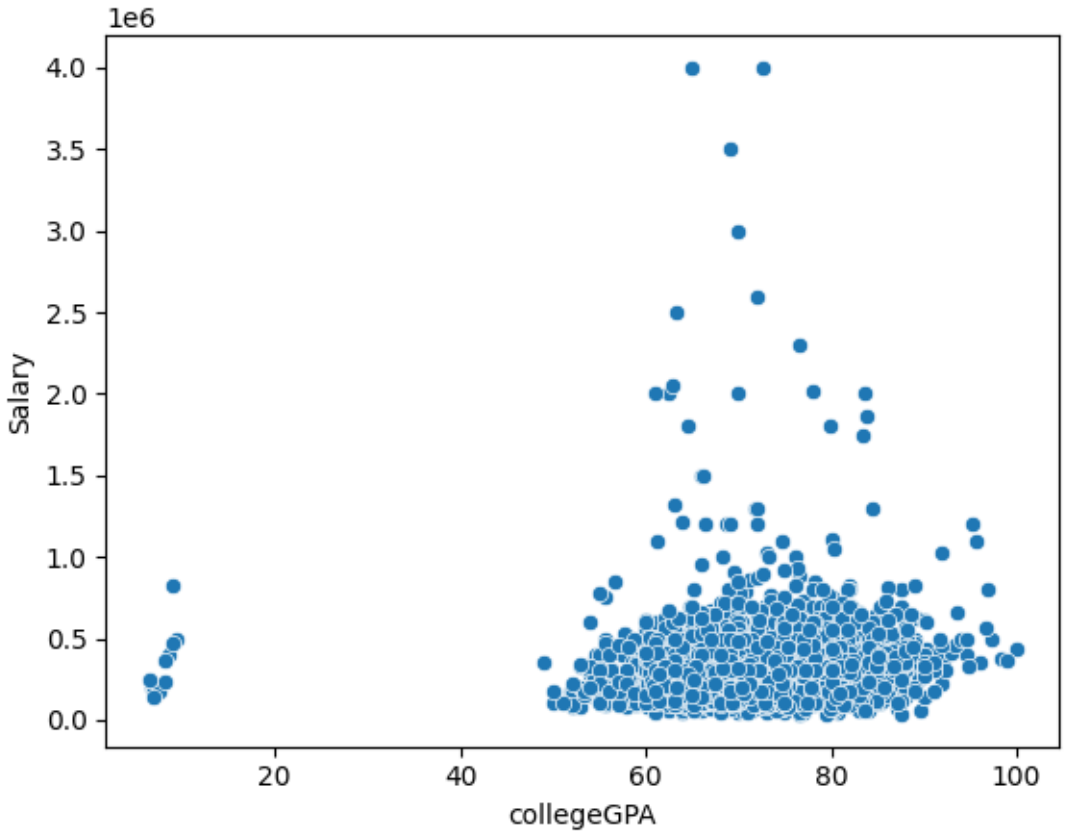
[ ]:
```
Designation  .net developer  .net web developer  account executive  \
Gender
f                  0.008359            0.001045           0.002090
m                  0.008550            0.000987           0.000658

Designation  account manager  admin assistant  administrative coordinator  \
```

```
Gender
f                   0.000000         0.000000                    0.000000
m                   0.000329         0.000658                    0.000329

Designation  administrative support  aircraft technician  android developer  \
Gender
f                          0.001045             0.001045           0.014629
m                          0.000000             0.000000           0.010523

Designation  application developer  …  ux designer  visiting faculty  \
Gender                             …
f                         0.016719  …     0.001045          0.000000
m                         0.011838  …     0.000329          0.000329

Designation  web application developer  web designer  \
Gender
f                             0.001045      0.005225
m                             0.001644      0.001315

Designation  web designer and joomla administrator  web designer and seo  \
Gender
f                                        0.000000              0.001045
m                                        0.000329              0.000000

Designation  web developer  web intern  website developer/tester  \
Gender
f                 0.017764    0.001045                  0.000000
m                 0.012167    0.000000                  0.000329

Designation  windows systems administrator
Gender
f                                0.001045
m                                0.000000

[2 rows x 419 columns]
```

#Bivariate Plotting For Num vs Num Columns - Line Plot, Scatter Plot, Hexbin Plot, Heat Map and Pair Plot

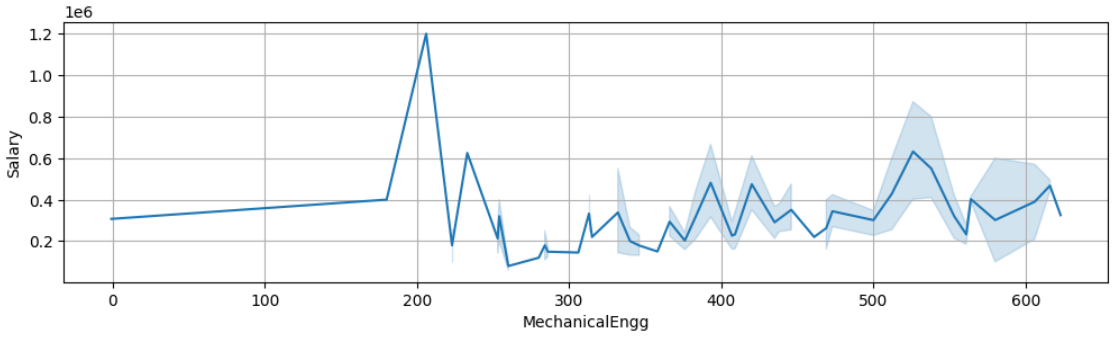
[ ]:
```
sns.scatterplot(data=df,y='Salary',x='collegeGPA')
```

[ ]:
```
<Axes: xlabel='collegeGPA', ylabel='Salary'>
```



Observation: There seems to be a positive correlation between college GPA and salary, especially in the range of 60 to 90 GPA. There are some outliers with high salaries despite lower GPAs.

[ ]:
```
fig, ax = plt.subplots(figsize=(12,3))
sns.lineplot(data=df,x='MechanicalEngg',y='Salary',ax=ax)
plt.grid(True)
plt.show()
```

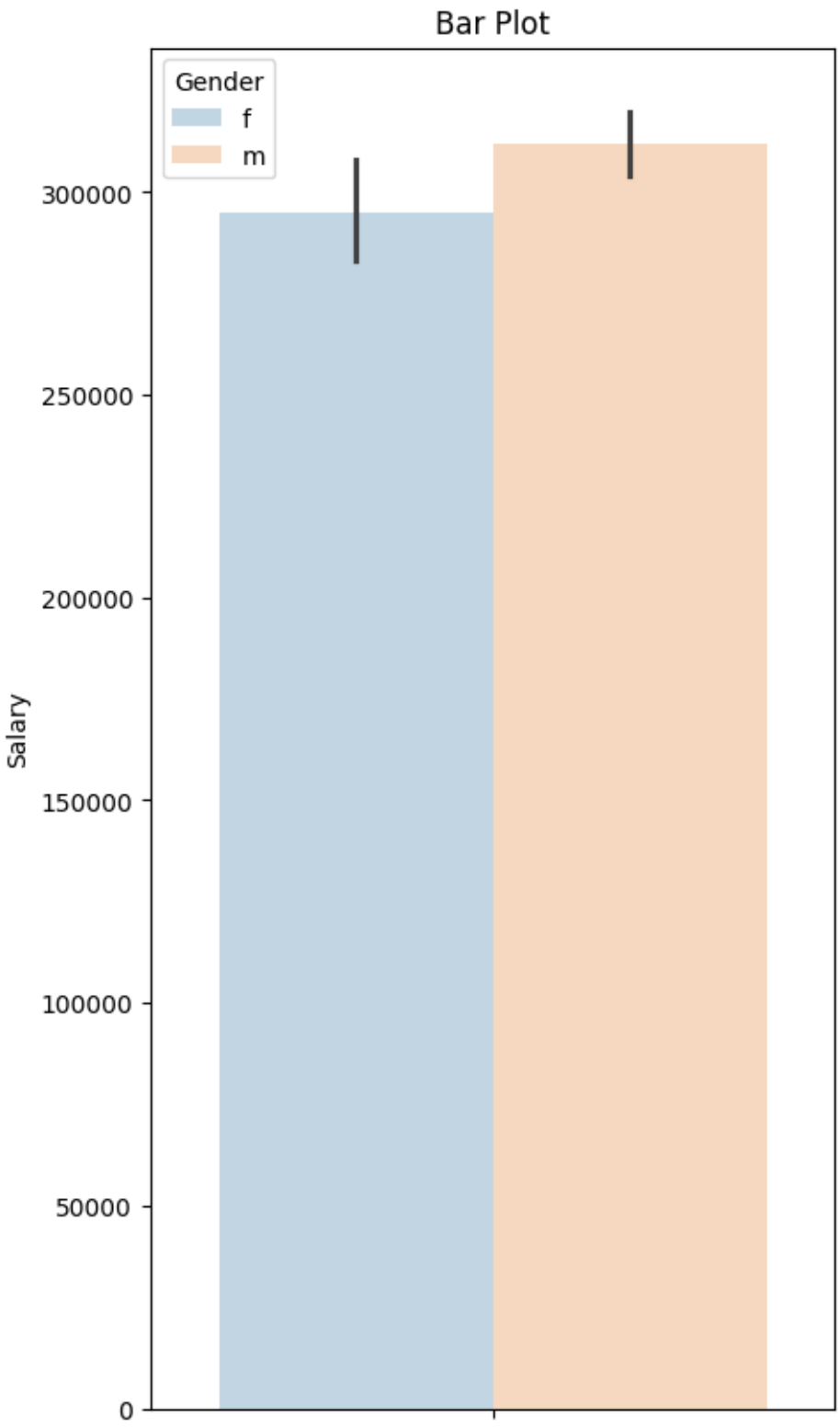


Observation:maximum salary for mechanical engg id approx 1.2 and most of mechanical engg have good salary

##Bivariate Plotting For Num vs Categorical Columns

[20]:
```
fig, ax = plt.subplots(figsize=(5,10))

ax.set_title("Bar Plot")
sns.barplot(data=df, y='Salary', hue="Gender", alpha=0.3, ax=ax)

plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

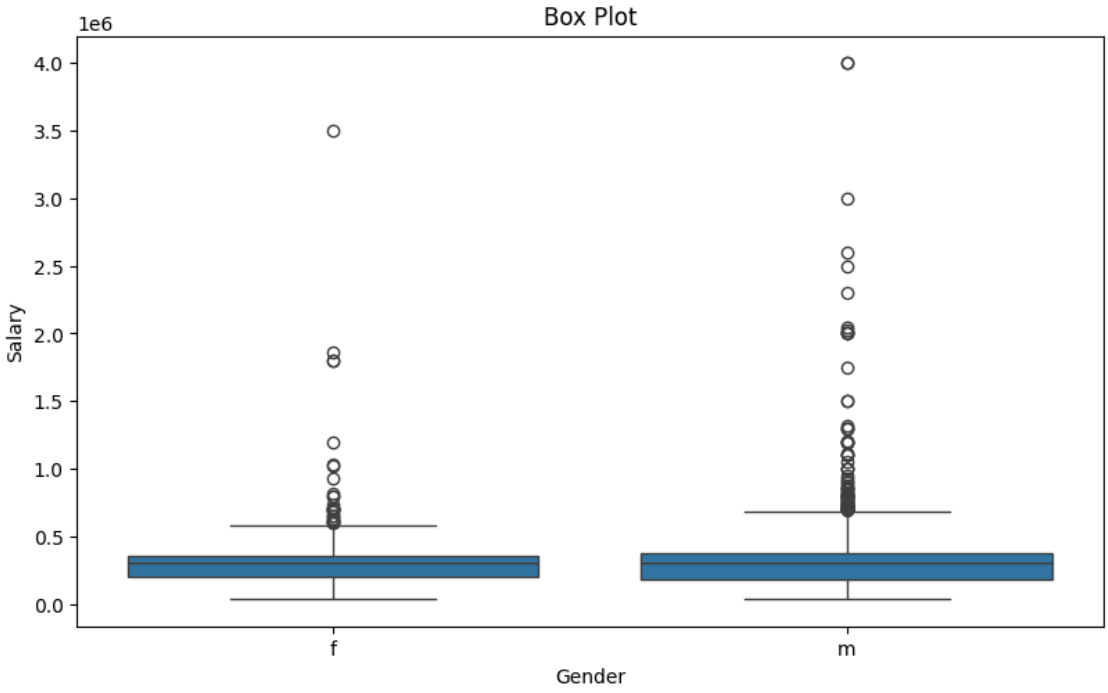


Obsevation: We observe that Male gender earn more salary than Female gender.

[21]:
```
fig, axs = plt.subplots(figsize=(8, 5), constrained_layout=True)
axs.set_title("Box Plot")
sns.boxplot(data=df, x='Gender', y='Salary', ax=axs)
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:640:
FutureWarning: SeriesGroupBy.grouper is deprecated and will be removed in a
future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```

[21]: `<Axes: title={'center': 'Box Plot'}, xlabel='Gender', ylabel='Salary'>`

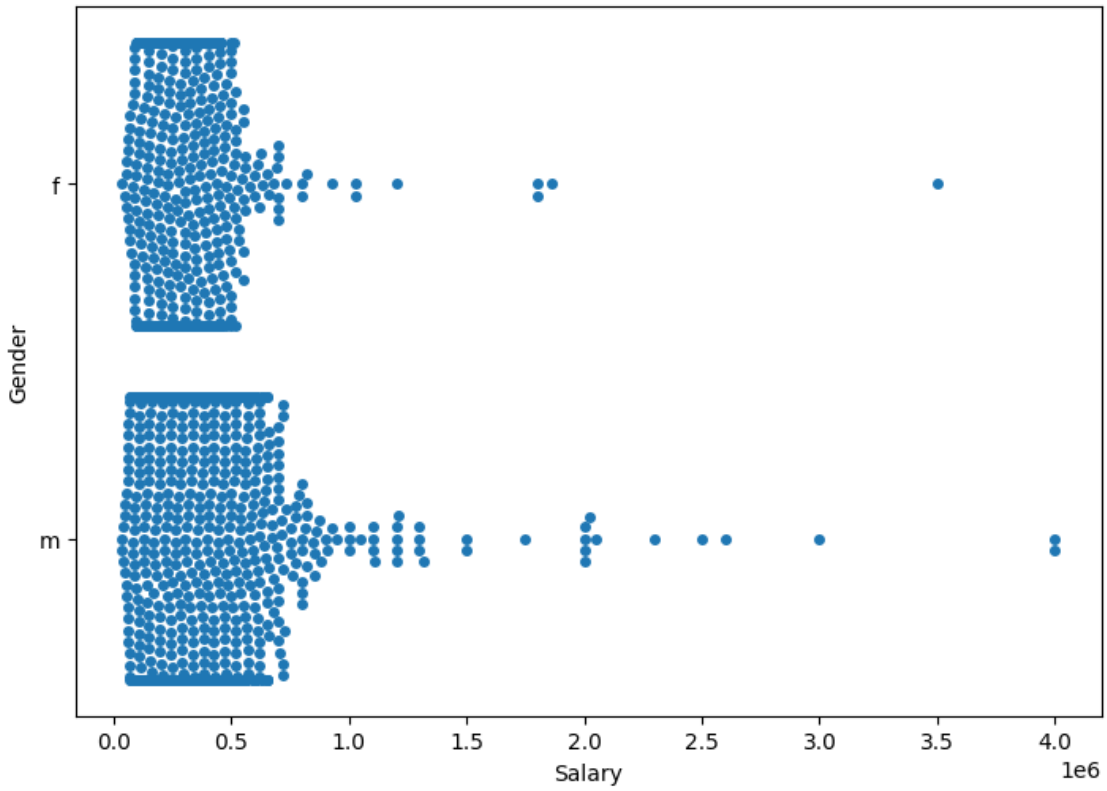


Observation :The box plot seems to show that there is no significant difference in median salary between females (f) and males (m). However, males exhibit greater variability in salary, including some significantly higher outliers

[22]:
```
fig,ax=plt.subplots(figsize=(7,5),constrained_layout=True)
sns.swarmplot(data=df,x='Salary',y='Gender',ax=ax)
plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:3398:
```

```
UserWarning: 51.3% of the points cannot be placed; you may want to decrease the
size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:3398:
UserWarning: 76.0% of the points cannot be placed; you may want to decrease the
size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:3398:
UserWarning: 69.5% of the points cannot be placed; you may want to decrease the
size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:3398:
UserWarning: 85.9% of the points cannot be placed; you may want to decrease the
size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

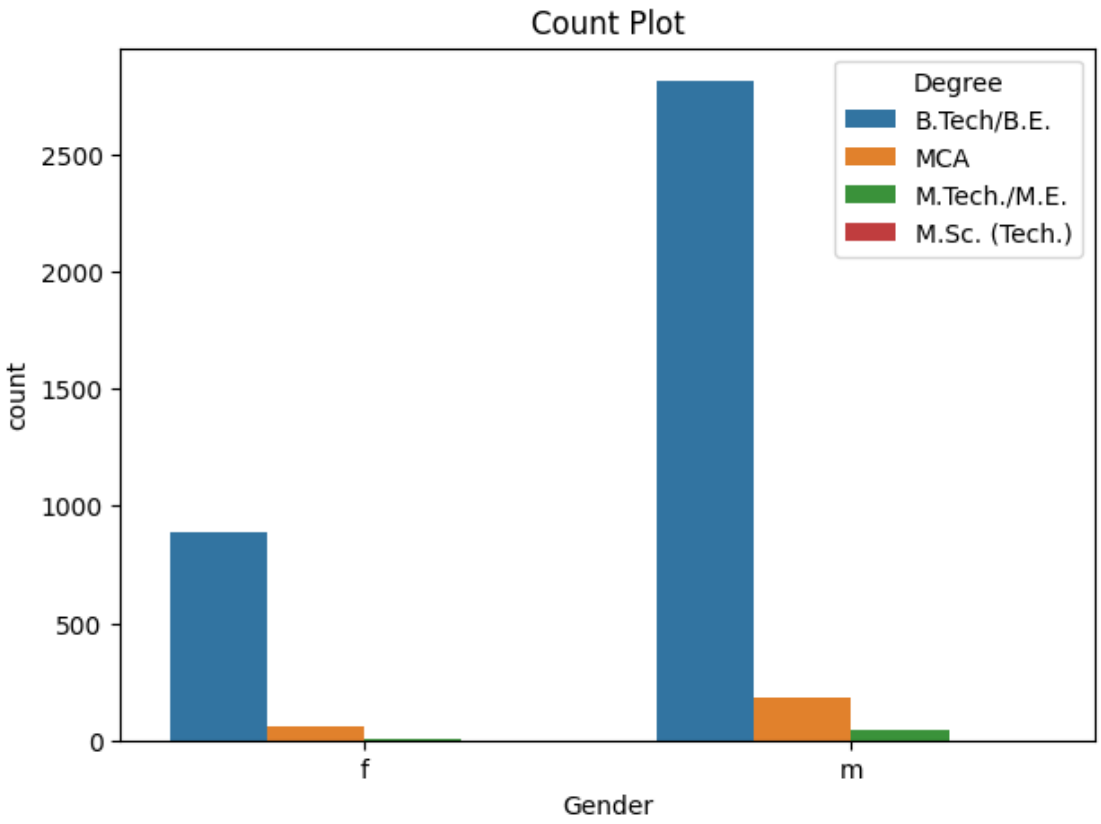


Observation :The box plot seems to show that there is no significant difference in median salary between females (f) and males (m). However, males exhibit greater variability in salary, including some significantly higher outliers

##Bivariate Plotting For Cat vs Categorical Columns - Grouped Count Plot

[18]:
```python
fig, ax = plt.subplots(figsize=(7,5))

ax.set_title("Count Plot")
sns.countplot(data=df, x='Gender', hue='Degree', ax=ax)

plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```



Obseevation:In above plot we can see that both gender having maximum count in B.Tech Degree and less count in M.Tech Degree