AMEO EDA

In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
df = pd.read_csv('E:/file2/Downloads/aspiring_minds_employability_outcome_2015.csv')
df.head()
```

Out[2]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10pe
0	train	203097	420000.0	01-06-2012 00:00	present	senior quality engineer	Bangalore	f	19-02-1990 00:00	
1	train	579905	500000.0	01-09-2013 00:00	present	assistant manager	Indore	m	04-10-1989 00:00	
2	train	810601	325000.0	01-06-2014 00:00	present	systems engineer	Chennai	f	03-08-1992 00:00	
3	train	267447	1100000.0	01-07-2011 00:00	present	senior software engineer	Gurgaon	m	05-12-1989 00:00	
4	train	343523	200000.0	01-03-2014 00:00	01-03-2015 00:00	get	Manesar	m	27-02-1991 00:00	

5 rows × 39 columns

In [3]:

```
data=df.iloc[:,1:]
data.head()
```

Out[3]:

	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	
0	203097	420000.0	01-06-2012 00:00	present	senior quality engineer	Bangalore	f	19-02-1990 00:00	84.3	o ec
1	579905	500000.0	01-09-2013 00:00	present	assistant manager	Indore	m	04-10-1989 00:00	85.4	
2	810601	325000.0	01-06-2014 00:00	present	systems engineer	Chennai	f	03-08-1992 00:00	85.0	
3	267447	1100000.0	01-07-2011 00:00	present	senior software engineer	Gurgaon	m	05-12-1989 00:00	85.6	
4	343523	200000.0	01-03-2014 00:00	01-03-2015 00:00	get	Manesar	m	27-02-1991 00:00	78.0	

5 rows × 38 columns

In [4]:

```
data.columns.to_list()
```

Out[4]:

```
['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']
```

In [5]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 38 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   ID                     3998 non-null   int64  
 1   Salary                 3998 non-null   float64
 2   DOJ                    3998 non-null   object 
 3   DOL                    3998 non-null   object 
 4   Designation            3998 non-null   object 
 5   JobCity                3998 non-null   object 
 6   Gender                 3998 non-null   object 
 7   DOB                    3998 non-null   object 
 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  CollegeTier            3998 non-null   int64  
 15  Degree                 3998 non-null   object 
 16  Specialization         3998 non-null   object 
 17  collegeGPA             3998 non-null   float64
 18  CollegeCityID          3998 non-null   int64  
 19  CollegeCityTier        3998 non-null   int64  
 20  CollegeState           3998 non-null   object 
 21  GraduationYear         3998 non-null   int64  
 22  English                3998 non-null   int64  
 23  Logical                3998 non-null   int64  
 24  Quant                  3998 non-null   int64  
 25  Domain                 3998 non-null   float64
 26  ComputerProgramming    3998 non-null   int64  
 27  ElectronicsAndSemicon  3998 non-null   int64  
 28  ComputerScience        3998 non-null   int64  
 29  MechanicalEngg         3998 non-null   int64  
 30  ElectricalEngg         3998 non-null   int64  
 31  TelecomEngg            3998 non-null   int64  
 32  CivilEngg              3998 non-null   int64  
 33  conscientiousness      3998 non-null   float64
 34  agreeableness          3998 non-null   float64
 35  extraversion           3998 non-null   float64
 36  nueroticism            3998 non-null   float64
 37  openess_to_experience  3998 non-null   float64
dtypes: float64(10), int64(17), object(11)
memory usage: 1.2+ MB
```

In [6]:

```
data.isnull().sum()
```

Out[6]:

```
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
```

In [7]:

```
data.describe()
```

Out[7]:

	ID	Salary	10percentage	12graduation	12percentage	CollegeID
count	3.998000e+03	3.998000e+03	3998.000000	3998.000000	3998.000000	3998.000000
mean	6.637945e+05	3.076998e+05	77.925443	2008.087544	74.466366	5156.851426
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933	4802.261482
min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000	2.000000
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000	494.000000
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000
max	1.298275e+06	4.000000e+06	97.760000	2013.000000	98.700000	18409.000000

8 rows × 27 columns

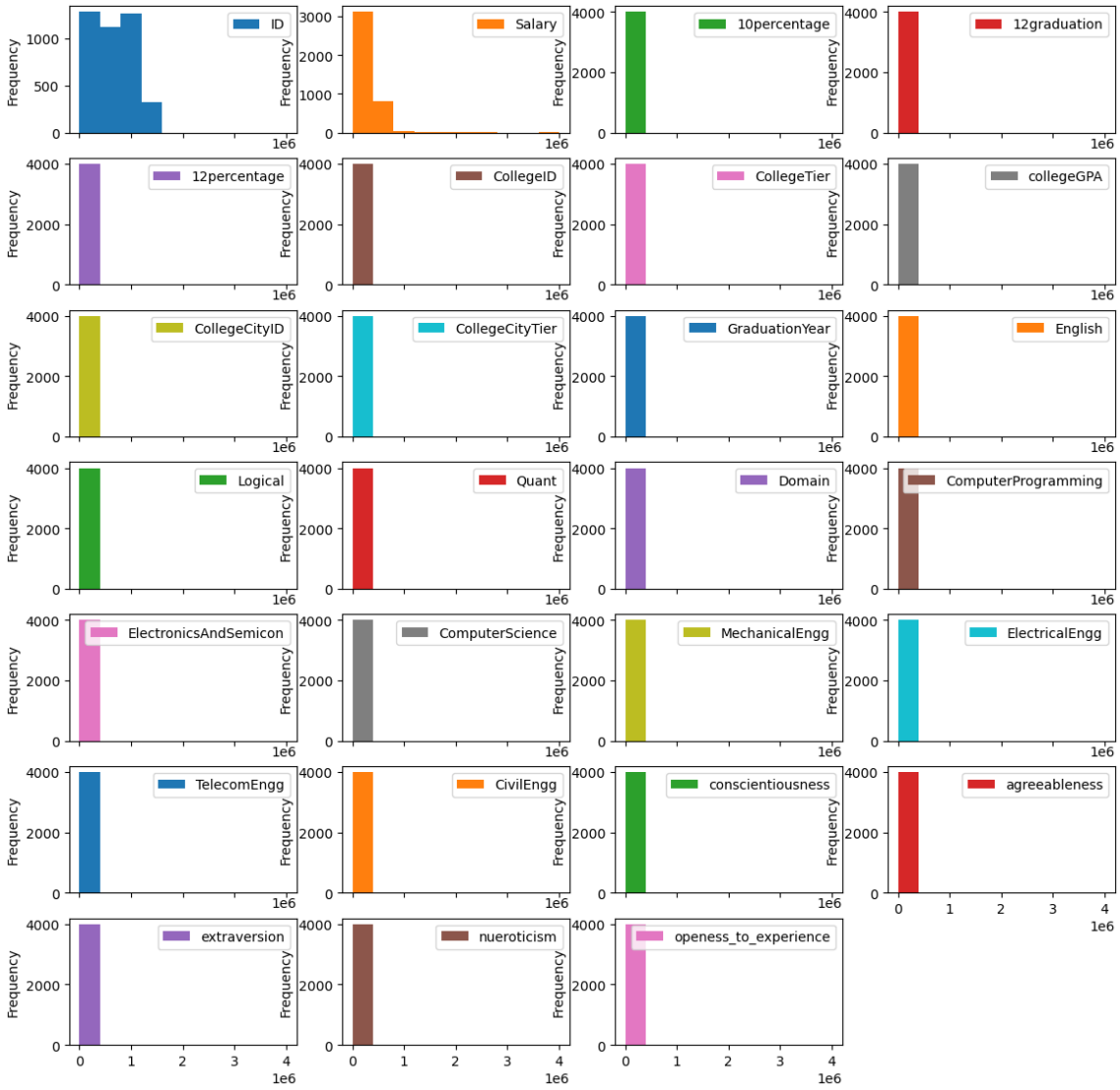
In [8]:

```
numerical_data = data.select_dtypes(['int64','float64'])
numerical_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 27 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   ID                     3998 non-null   int64  
 1   Salary                 3998 non-null   float64
 2   10percentage           3998 non-null   float64
 3   12graduation           3998 non-null   int64  
 4   12percentage           3998 non-null   float64
 5   CollegeID              3998 non-null   int64  
 6   CollegeTier            3998 non-null   int64  
 7   collegeGPA             3998 non-null   float64
 8   CollegeCityID          3998 non-null   int64  
 9   CollegeCityTier        3998 non-null   int64  
 10  GraduationYear         3998 non-null   int64  
 11  English                3998 non-null   int64  
 12  Logical                3998 non-null   int64  
 13  Quant                  3998 non-null   int64  
 14  Domain                 3998 non-null   float64
 15  ComputerProgramming    3998 non-null   int64  
 16  ElectronicsAndSemicon  3998 non-null   int64  
 17  ComputerScience        3998 non-null   int64  
 18  MechanicalEngg         3998 non-null   int64  
 19  ElectricalEngg         3998 non-null   int64  
 20  TelecomEngg            3998 non-null   int64  
 21  CivilEngg              3998 non-null   int64  
 22  conscientiousness      3998 non-null   float64
 23  agreeableness          3998 non-null   float64
 24  extraversion           3998 non-null   float64
 25  nueroticism            3998 non-null   float64
 26  openess_to_experience  3998 non-null   float64
dtypes: float64(10), int64(17)
memory usage: 843.5 KB
```

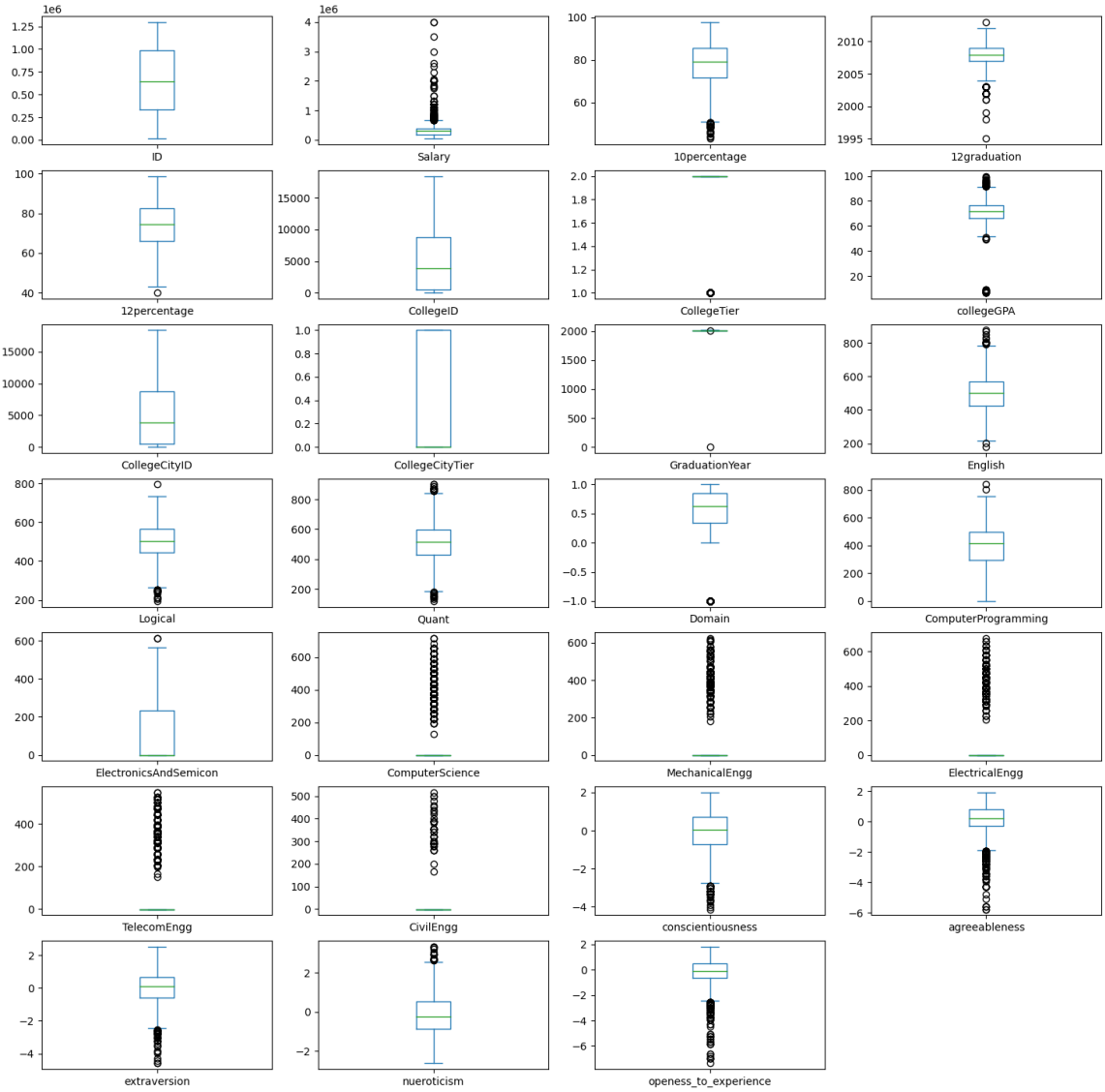
In [9]:

```
numerical_data.plot(kind='hist',subplots=True,figsize=(14,14),layout=(7,4));
```



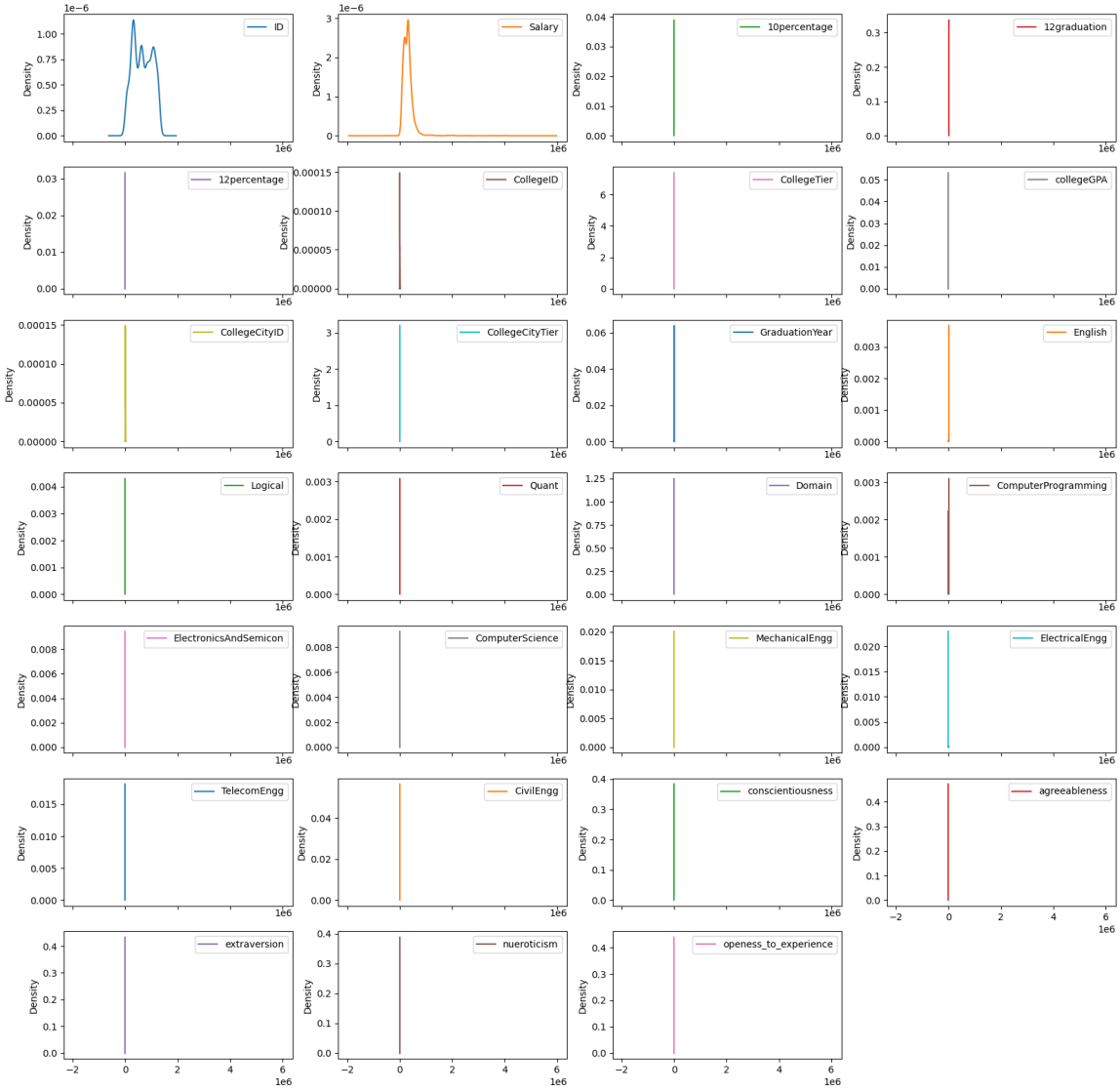
In [10]:

```
numerical_data.plot(kind='box',subplots=True,figsize=(18,18),layout=(7,4));
```



In [11]:

```
numerical_data.plot(kind='kde',subplots=True,figsize=(20,20),layout=(7,4));
```



In [12]:

```
categorical_data = data.select_dtypes(['object'])
categorical_data.head()
```

Out[12]:

	DOJ	DOL	Designation	JobCity	Gender	DOB	10board	12board	Degre
0	01-06-2012 00:00	present	senior quality engineer	Bangalore	f	19-02-1990 00:00	board ofsecondary education,ap	board of intermediate education,ap	B.Tech/B.E
1	01-09-2013 00:00	present	assistant manager	Indore	m	04-10-1989 00:00	cbse	cbse	B.Tech/B.E
2	01-06-2014 00:00	present	systems engineer	Chennai	f	03-08-1992 00:00	cbse	cbse	B.Tech/B.E
3	01-07-2011 00:00	present	senior software engineer	Gurgaon	m	05-12-1989 00:00	cbse	cbse	B.Tech/B.E
4	01-03-2014 00:00	01-03-2015 00:00	get	Manesar	m	27-02-1991 00:00	cbse	cbse	B.Tech/B.E

In [13]:

```
categorical_data = categorical_data.drop(columns=['DOJ','DOL','DOB'])
categorical_data.head()
```

Out[13]:

	Designation	JobCity	Gender	10board	12board	Degree	Specialization	Colle
0	senior quality engineer	Bangalore	f	board ofsecondary education,ap	board of intermediate education,ap	B.Tech/B.E.	computer engineering	
1	assistant manager	Indore	m	cbse	cbse	B.Tech/B.E.	electronics and communication engineering	
2	systems engineer	Chennai	f	cbse	cbse	B.Tech/B.E.	information technology	
3	senior software engineer	Gurgaon	m	cbse	cbse	B.Tech/B.E.	computer engineering	
4	get	Manesar	m	cbse	cbse	B.Tech/B.E.	electronics and communication engineering	

In [ ]:

```
```

In [ ]:

```
```

In [14]:

```
jobcity=data['JobCity'].value_counts()
```

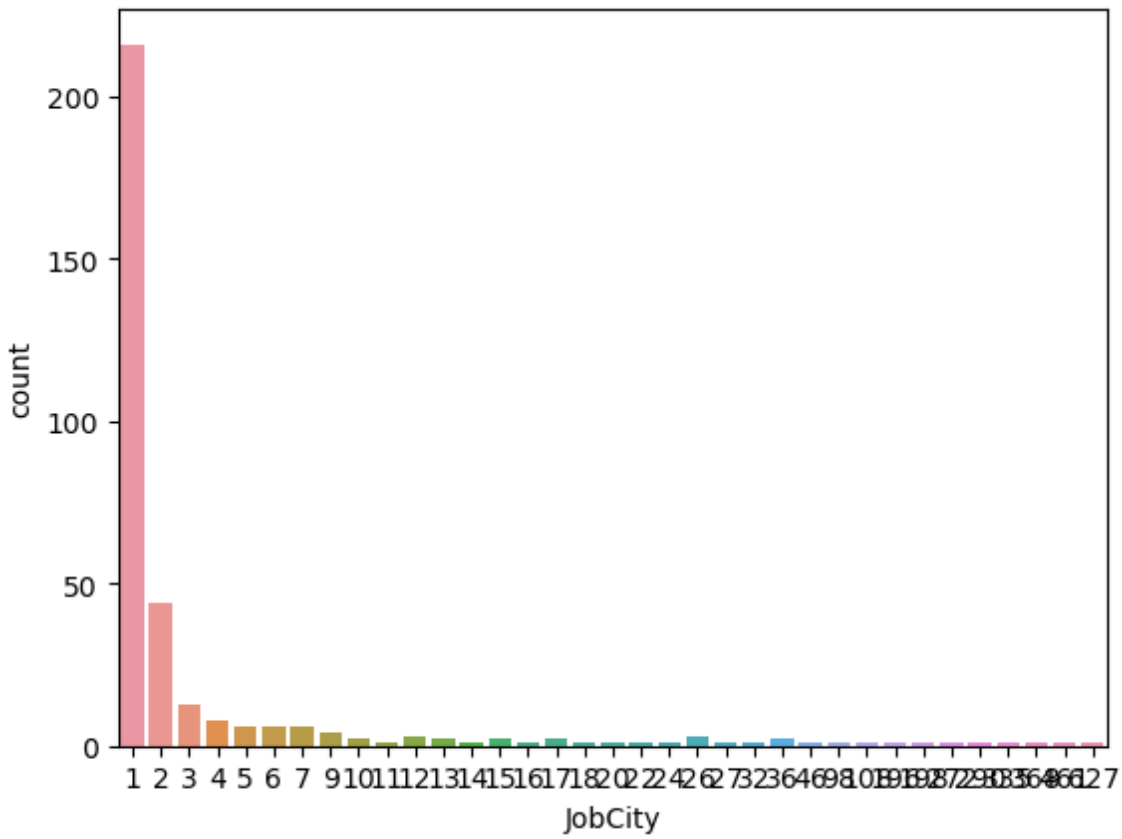
In [15]:

```
import seaborn as sns
sns.countplot(jobcity)
```

```
D:\Users\Safuvan\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or misinterp
retation.
  warnings.warn(
```

Out[15]:

```
<AxesSubplot:xlabel='JobCity', ylabel='count'>
```



In [16]:

```
data['JobCity'].isnull().sum()
```

Out[16]:

```
0
```

In [17]:

```
data['Designation'].value_counts()
```

Out[17]:

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

In [18]:

```
data['Degree'].value_counts()
```

Out[18]:

```
B.Tech/B.E.      3700
MCA               243
M.Tech./M.E.       53
M.Sc. (Tech.)       2
Name: Degree, dtype: int64
```

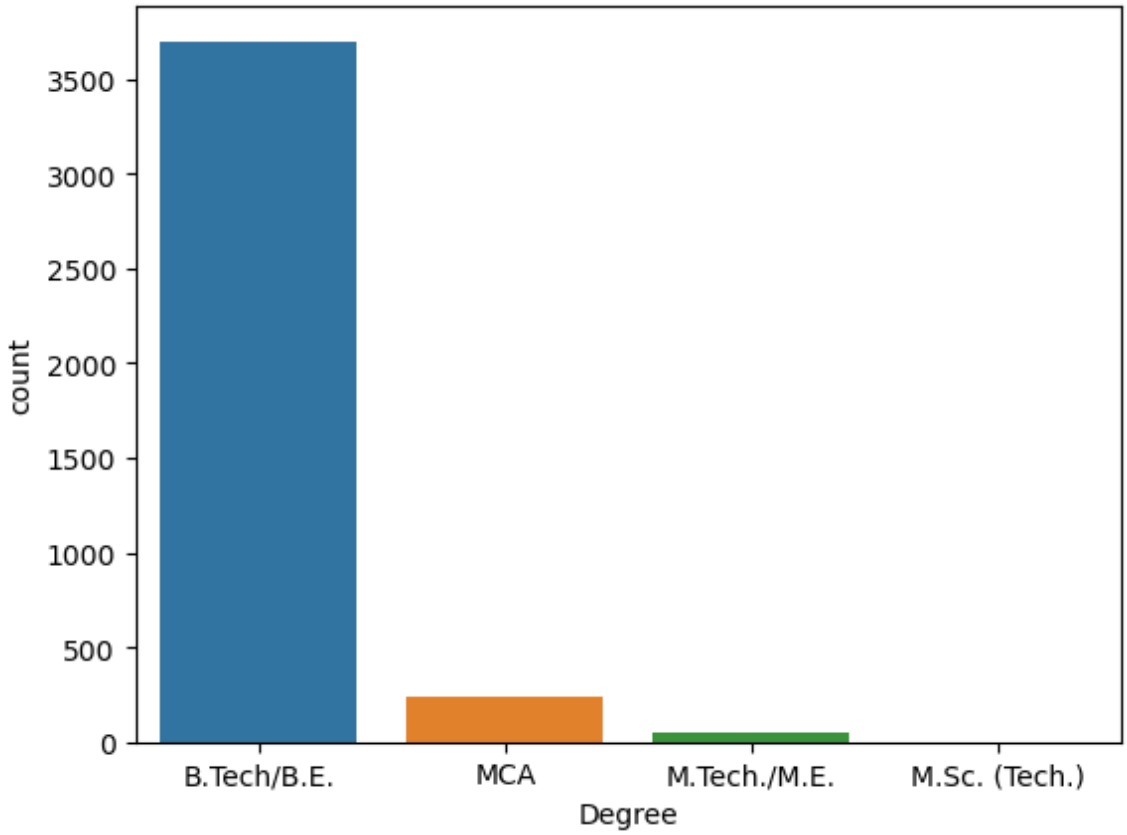
In [19]:

```
sns.countplot(data['Degree'])
```

```
D:\Users\Safuvan\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or misinterp
retation.
  warnings.warn(
```

Out[19]:

```
<AxesSubplot:xlabel='Degree', ylabel='count'>
```



In [20]:

```
data['DOB']=pd.to_datetime(data['DOB'])
```

In [21]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 38 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     3998 non-null   int64
 1   Salary                 3998 non-null   float64
 2   DOJ                    3998 non-null   object
 3   DOL                    3998 non-null   object
 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  CollegeTier            3998 non-null   int64
 15  Degree                 3998 non-null   object
 16  Specialization         3998 non-null   object
 17  collegeGPA             3998 non-null   float64
 18  CollegeCityID          3998 non-null   int64
 19  CollegeCityTier        3998 non-null   int64
 20  CollegeState           3998 non-null   object
 21  GraduationYear         3998 non-null   int64
 22  English                3998 non-null   int64
 23  Logical                3998 non-null   int64
 24  Quant                  3998 non-null   int64
 25  Domain                 3998 non-null   float64
 26  ComputerProgramming    3998 non-null   int64
 27  ElectronicsAndSemicon  3998 non-null   int64
 28  ComputerScience        3998 non-null   int64
 29  MechanicalEngg         3998 non-null   int64
 30  ElectricalEngg         3998 non-null   int64
 31  TelecomEngg            3998 non-null   int64
 32  CivilEngg              3998 non-null   int64
 33  conscientiousness      3998 non-null   float64
 34  agreeableness          3998 non-null   float64
 35  extraversion           3998 non-null   float64
 36  nueroticism            3998 non-null   float64
 37  openess_to_experience  3998 non-null   float64
dtypes: datetime64[ns](1), float64(10), int64(17), object(10)
memory usage: 1.2+ MB
```

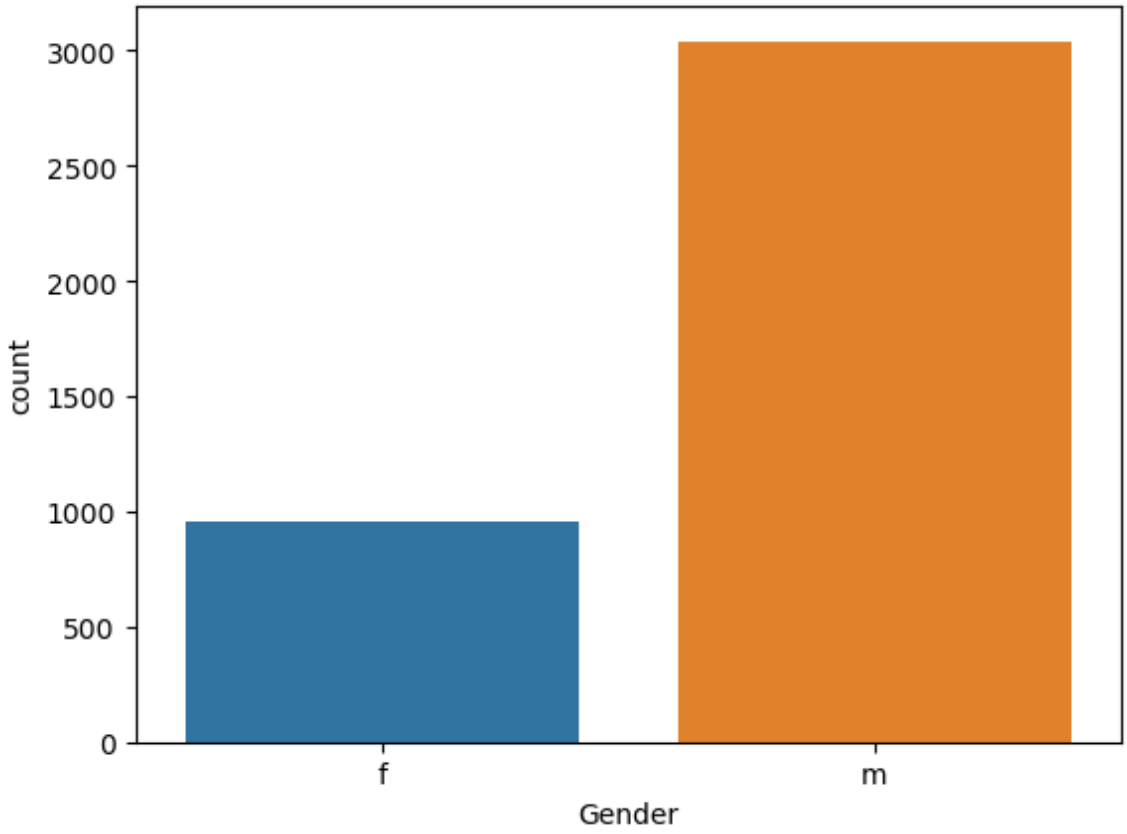
In [22]:

```
sns.countplot(data['Gender'])
```

```
D:\Users\Safuvan\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or misinterp
retation.
  warnings.warn(
```

Out[22]:

```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```

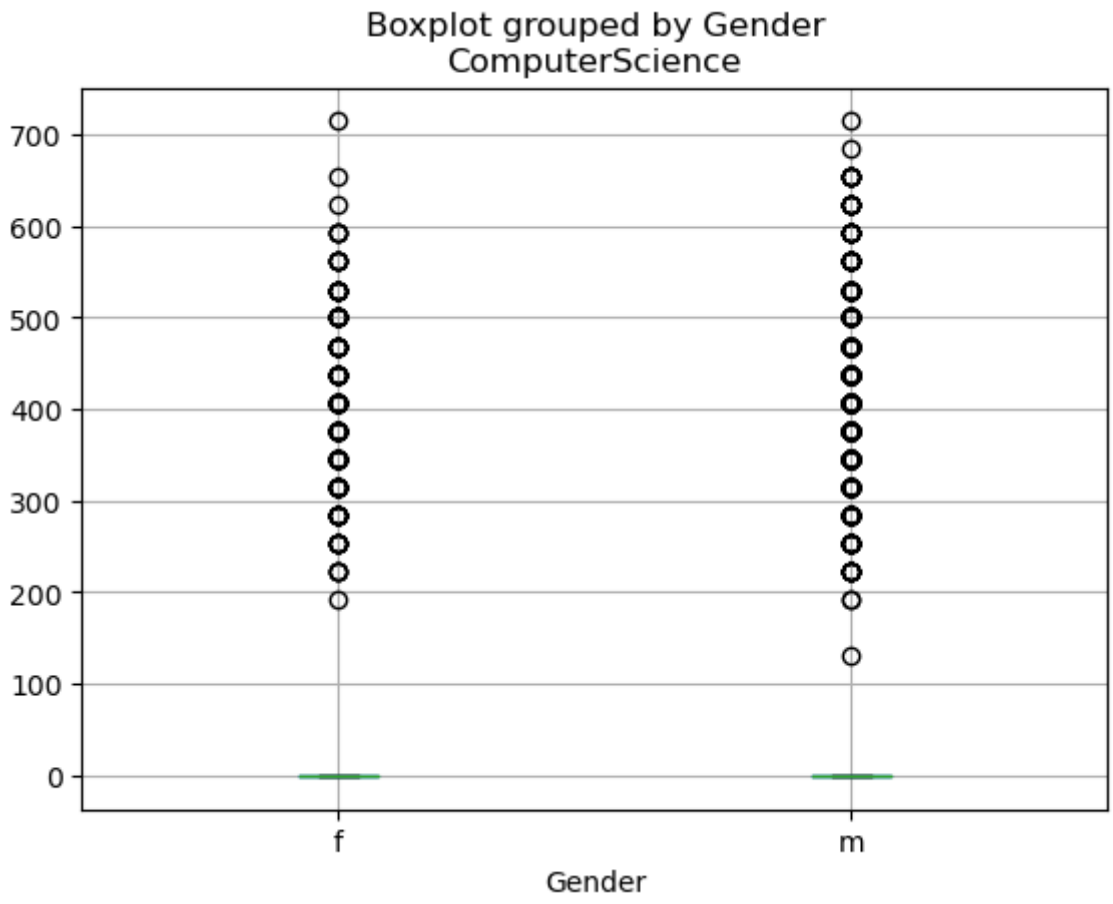


In [23]:

```
data.boxplot(by='Gender',column='ComputerScience')
```

Out[23]:

```
<AxesSubplot:title={'center':'ComputerScience'}, xlabel='Gender'>
```



In [24]:

```
data['ComputerScience'].unique()
```

Out[24]:

```
array([ -1, 407, 346, 376, 500, 438, 315, 253, 469, 192, 530, 284, 223,
       561, 684, 592, 623, 653, 130, 715], dtype=int64)
```

In [25]:

```
print('mean :',data['ComputerScience'].mean())
print('median :',data['ComputerScience'].median())
```

```
mean : 90.7423711855928
median : -1.0
```

In [26]:

```
print('min :',data['ComputerScience'].min())
print('max :',data['ComputerScience'].max())
```

```
min : -1
max : 715
```

In [27]:

```
print('mean :',data['ComputerScience'].std())
```

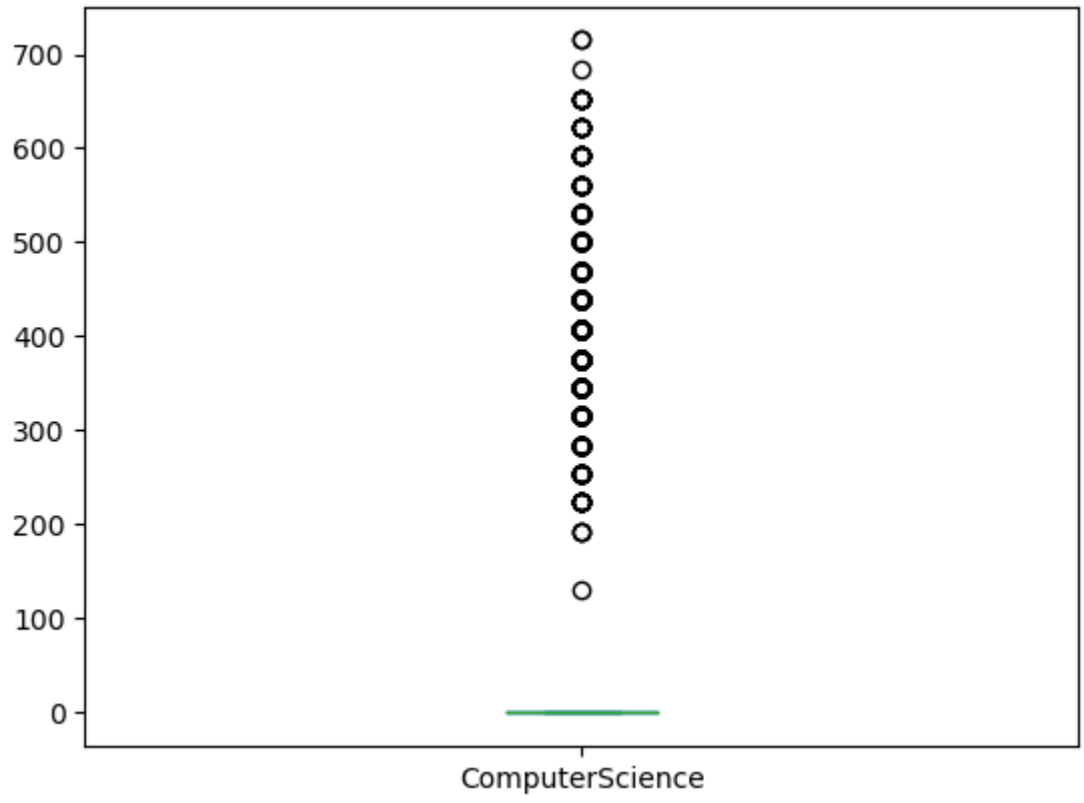
```
mean : 175.2730830755835
```

In [28]:

```
data['ComputerScience'].plot(kind='box')
```

Out[28]:

```
<AxesSubplot:>
```



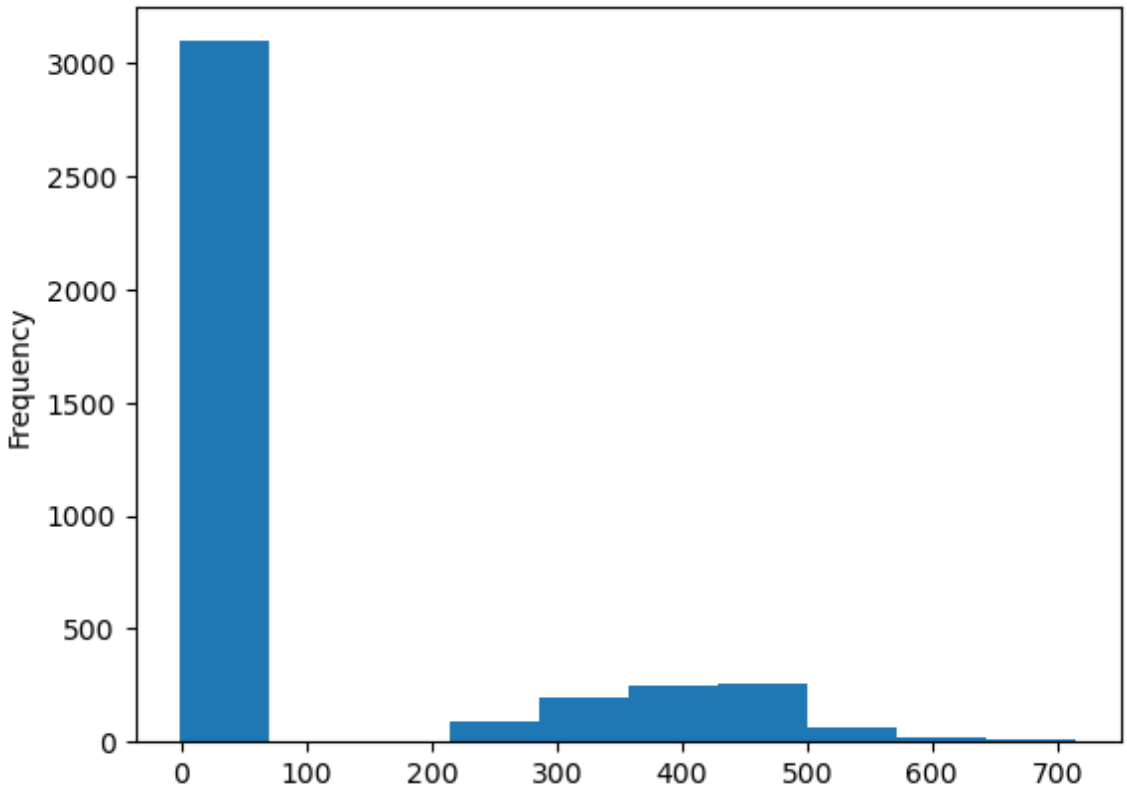
In [ ]:

In [29]:

```
data['ComputerScience'].plot(kind='hist')
```

Out[29]:

```
<AxesSubplot:ylabel='Frequency'>
```



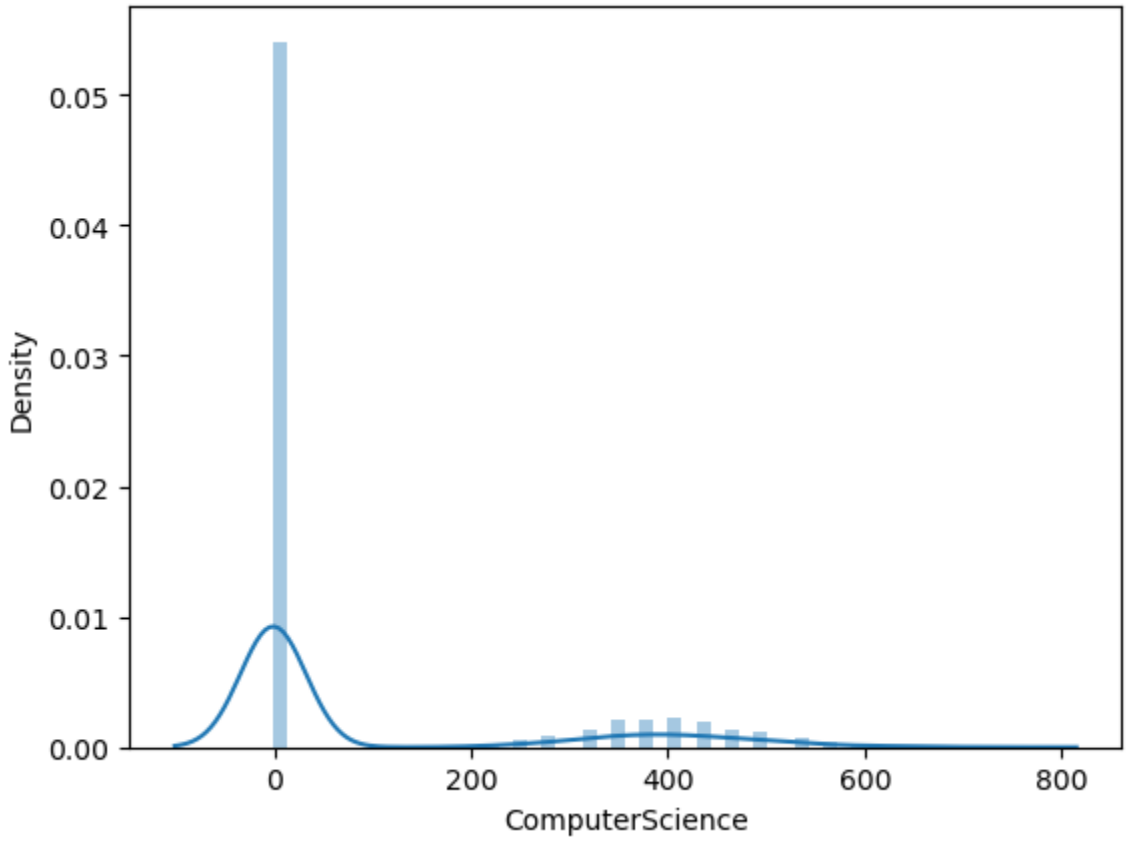
In [30]:

```
sns.distplot(data['ComputerScience'])
```

```
D:\Users\Safuvan\anaconda3\lib\site-packages\seaborn\distributions.py:261
9: FutureWarning: `distplot` is a deprecated function and will be removed
in a future version. Please adapt your code to use either `displot` (a fig
ure-level function with similar flexibility) or `histplot` (an axes-level
function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[30]:

```
<AxesSubplot:xlabel='ComputerScience', ylabel='Density'>
```

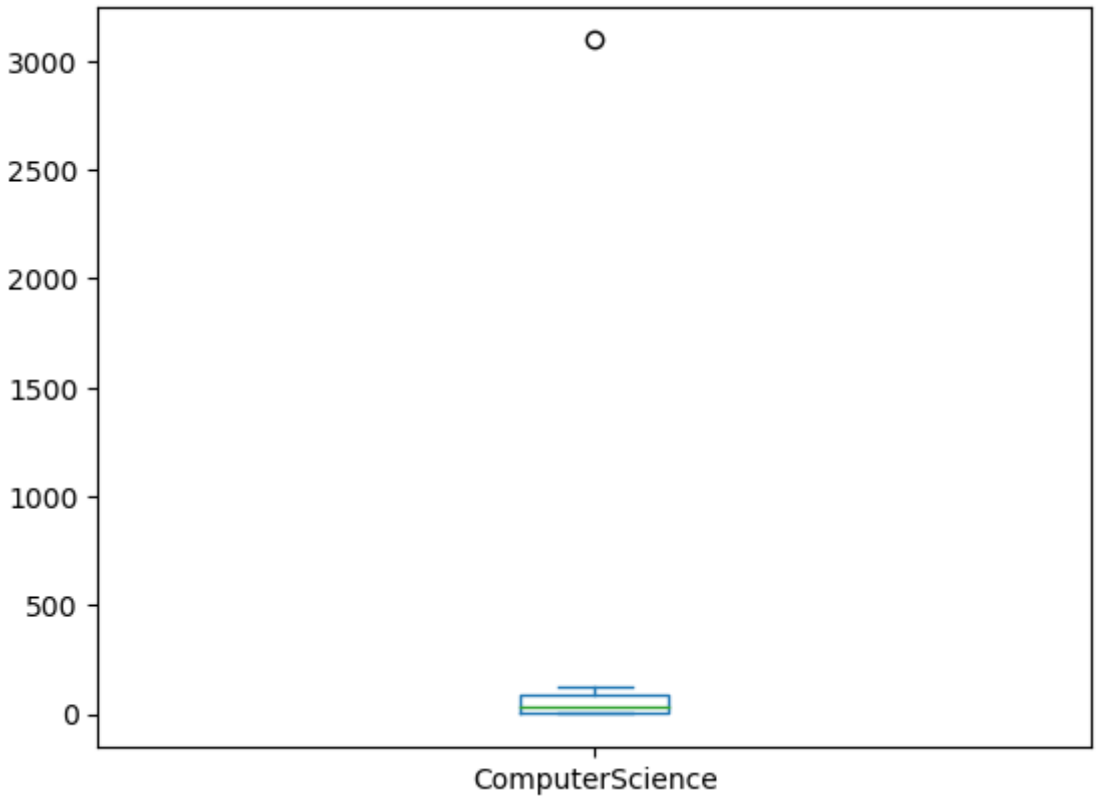


In [31]:

```
data['ComputerScience'].value_counts().plot(kind='box')
```

Out[31]:

```
<AxesSubplot:>
```



In [32]:

```
data=data[data.ComputerScience <= 450]
```

In [33]:

```
data['ComputerScience'].unique()
```

Out[33]:

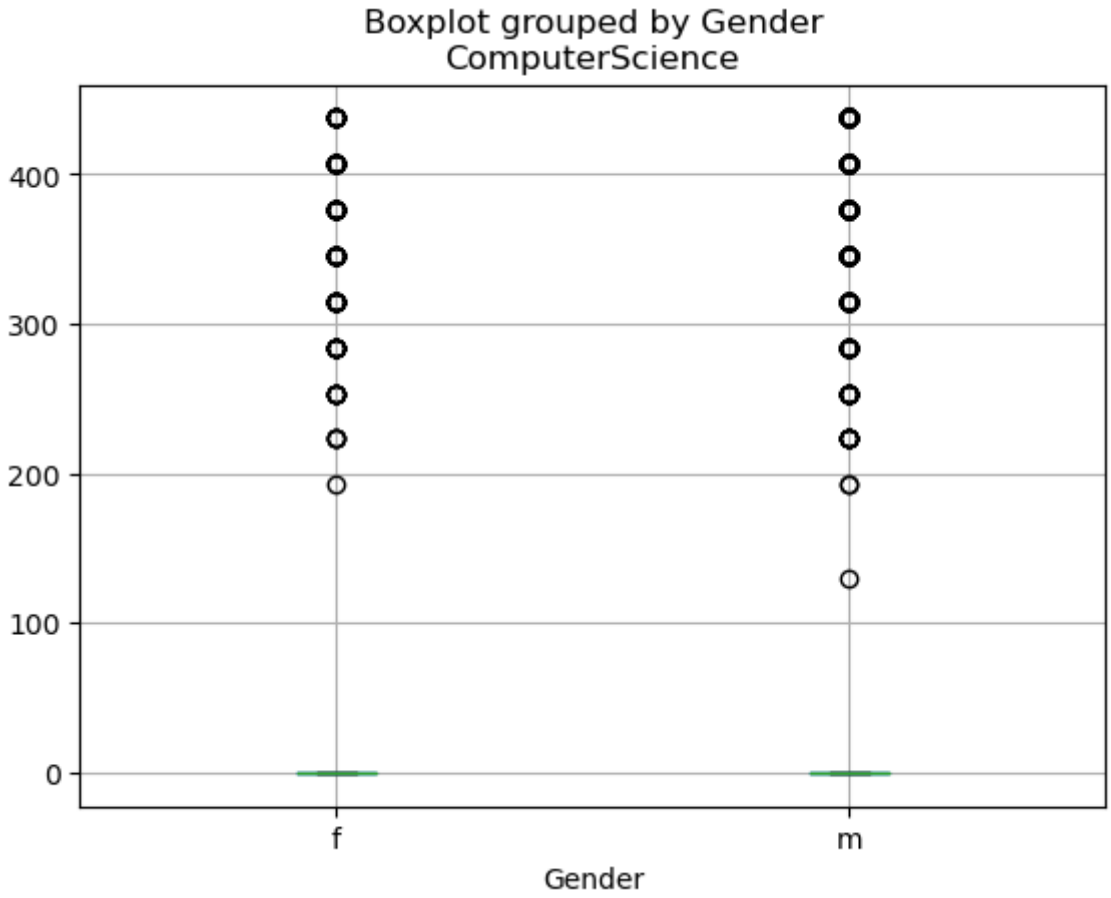
```
array([ -1, 407, 346, 376, 438, 315, 253, 192, 284, 223, 130], dtype=int6
4)
```

In [34]:

```
data.boxplot(by='Gender',column='ComputerScience')
```

Out[34]:

```
<AxesSubplot:title={'center':'ComputerScience'}, xlabel='Gender'>
```

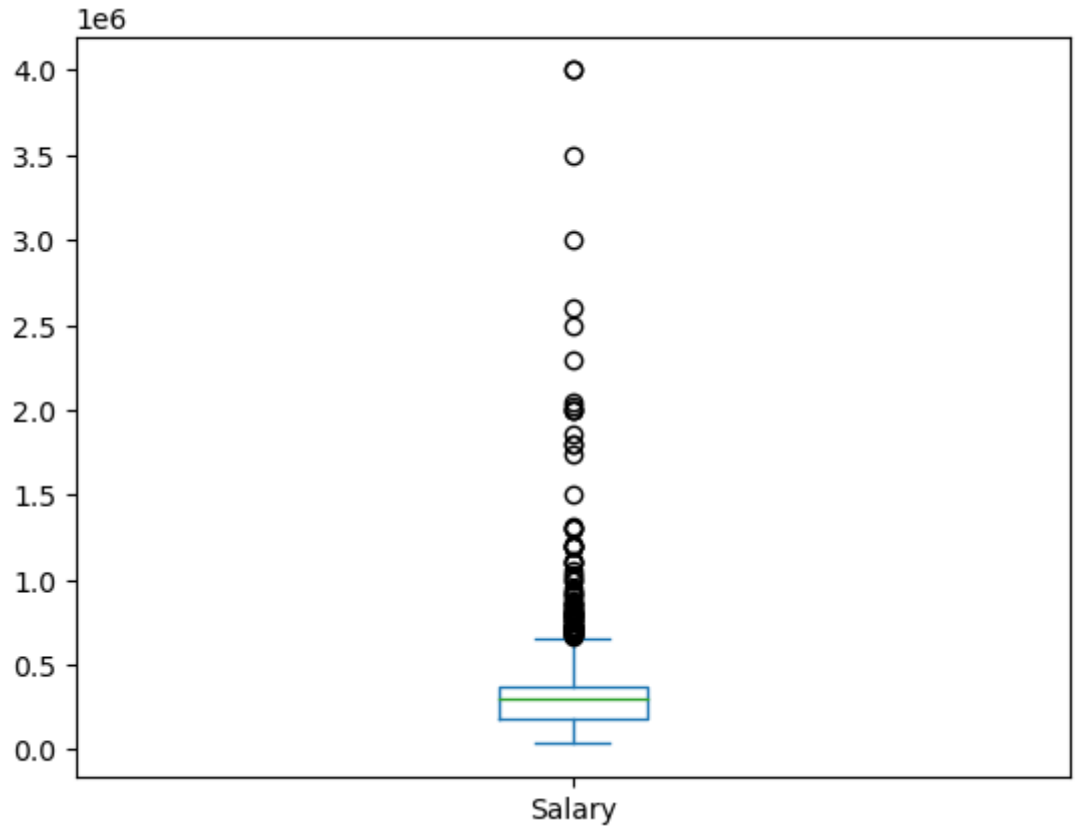


In [35]:

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

Out[35]:

```
<AxesSubplot:>
```



In [36]:

```
print('median:',data['Salary'].median())
print('mean:',data['Salary'].mean())
print('min:',data['Salary'].min())
print('max:',data['Salary'].max())
```

```
median: 300000.0
mean: 308392.1620901093
min: 35000.0
max: 4000000.0
```

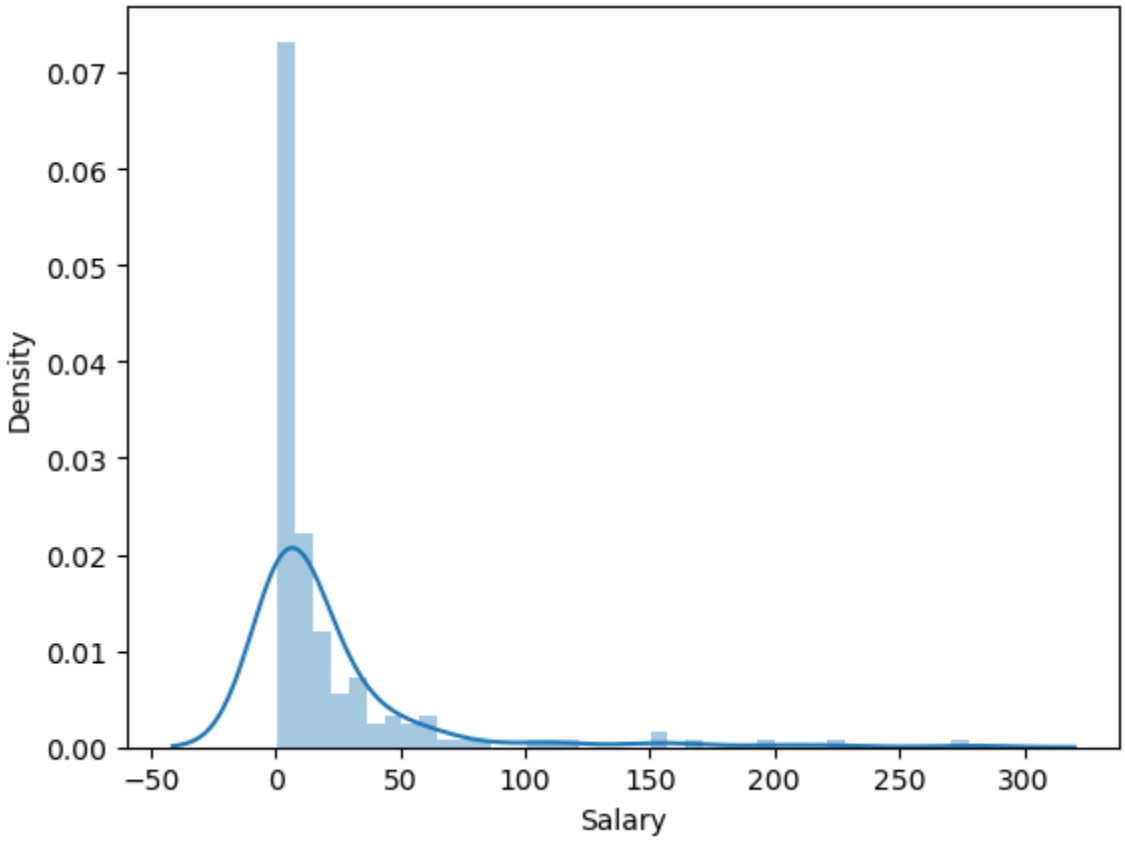
In [37]:

```
salary=data['Salary'].value_counts()
sns.distplot(salary)
```

```
D:\Users\Safuvan\anaconda3\lib\site-packages\seaborn\distributions.py:261
9: FutureWarning: `distplot` is a deprecated function and will be removed
in a future version. Please adapt your code to use either `displot` (a fig
ure-level function with similar flexibility) or `histplot` (an axes-level
function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[37]:

```
<AxesSubplot:xlabel='Salary', ylabel='Density'>
```



In [38]:

```
q1 = data['Salary'].quantile(0.25)
q3 = data['Salary'].quantile(0.75)
IQR = q3-q1
print(IQR)
print(q1)
```

```
195000.0
180000.0
```

In [39]:

```
lower = q1 - 1.5*IQR
upper = q3 + 1.5*IQR
print(lower)
print(upper)
```

```
-112500.0
667500.0
```

In [47]:

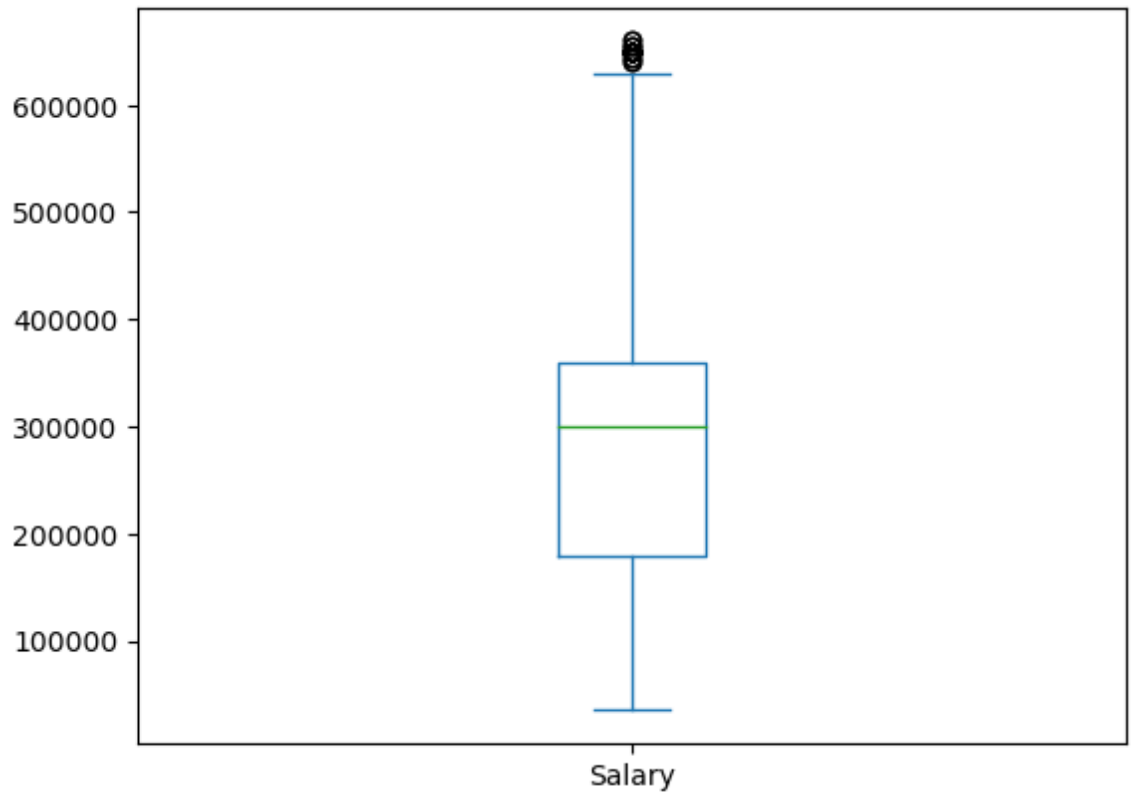
```
data2=data[(data['Salary'] >= lower)&(data['Salary'] <= upper)]
```

In [48]:

```
data2['Salary'].plot(kind='box')
```

Out[48]:

```
<AxesSubplot:>
```



In [54]:

```
print(data2['Salary'].shape)
print(data['Salary'].shape)
```

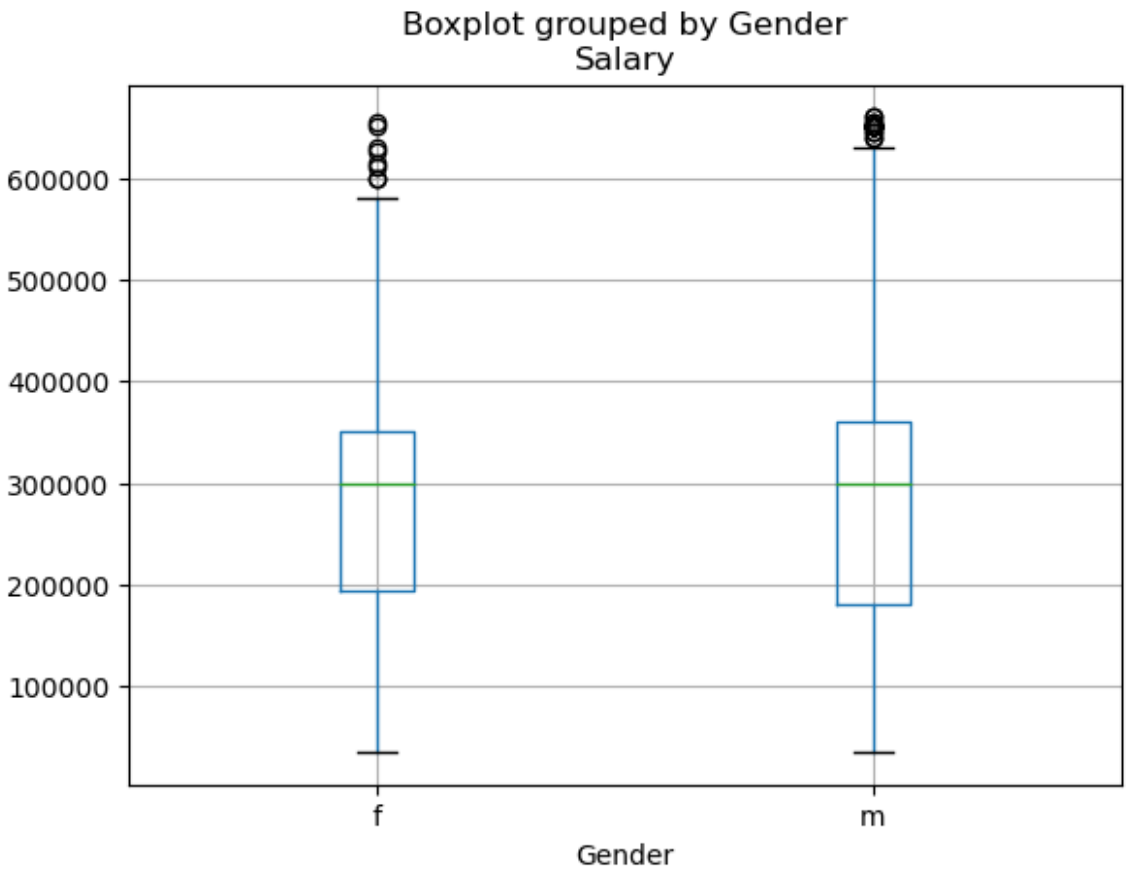
```
(3650,)
(3751,)
```

In [55]:

```
data2.boxplot(by='Gender', column='Salary')
```

Out[55]:

```
<AxesSubplot:title={'center':'Salary'}, xlabel='Gender'>
```



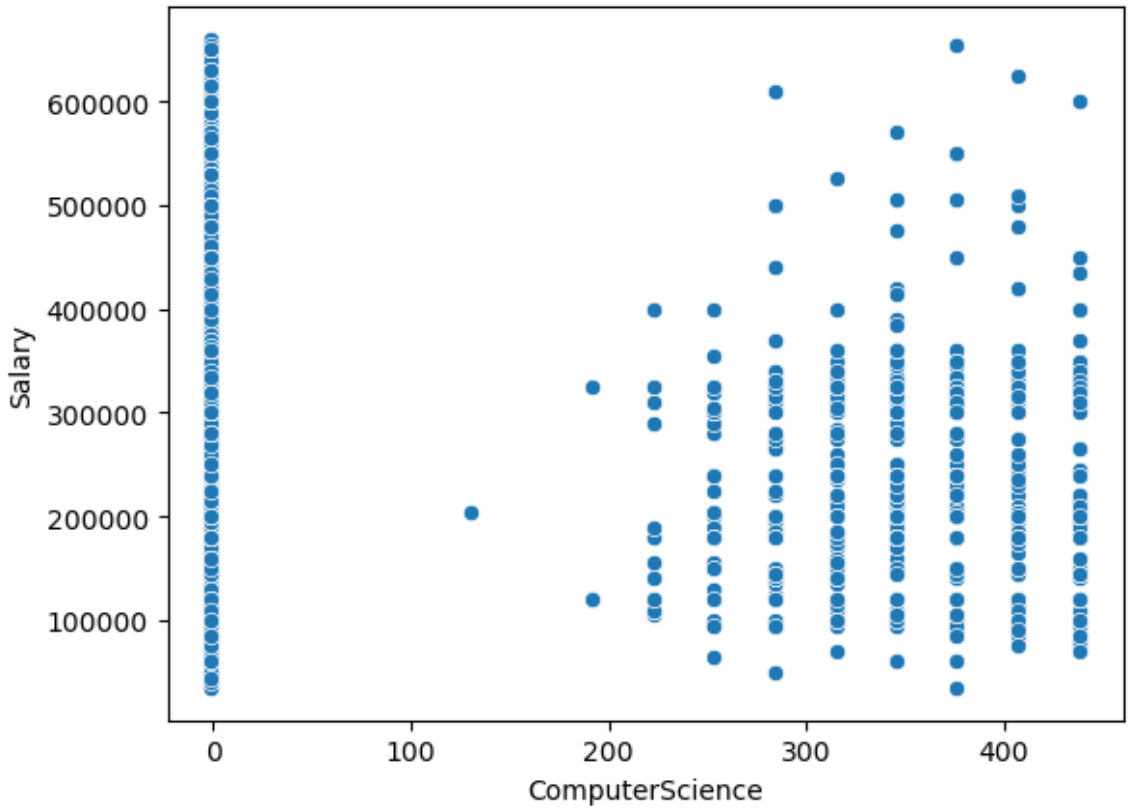
In [56]:

```
x=data2['ComputerScience']
y=data2['Salary']
sns.scatterplot(x,y)
```

```
D:\Users\Safuvan\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variables as keyword args: x, y. From vers
ion 0.12, the only valid positional argument will be `data`, and passing o
ther arguments without an explicit keyword will result in an error or misi
nterpretation.
  warnings.warn(
```

Out[56]:

```
<AxesSubplot:xlabel='ComputerScience', ylabel='Salary'>
```



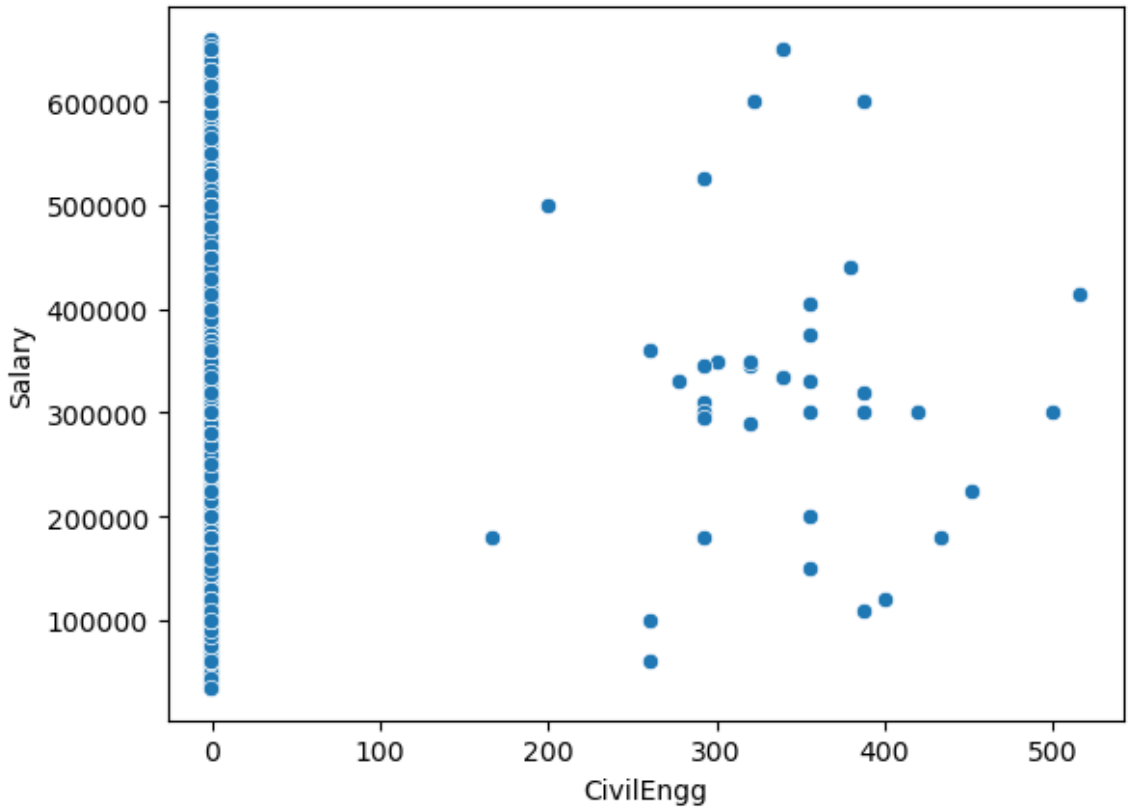
In [57]:

```
x=data2['CivilEngg']
y=data2['Salary']
sns.scatterplot(x,y)
```

```
D:\Users\Safuvan\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variables as keyword args: x, y. From vers
ion 0.12, the only valid positional argument will be `data`, and passing o
ther arguments without an explicit keyword will result in an error or misi
nterpretation.
  warnings.warn(
```

Out[57]:

```
<AxesSubplot:xlabel='CivilEngg', ylabel='Salary'>
```



In [ ]: