Descriptive statistics for employee attrition data set

creating a data frame from the .csv file

In [2]:
```
import pandas as pd
df=pd.read_csv('Z:\da lab\EmployeeAttrition.csv')
df
```

Out[2]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Educ
0	41	Yes	Travel_Rarely	1102	Sales	1	2
1	49	No	Travel_Frequently	279	Research & Development	8	1
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2
3	33	No	Travel_Frequently	1392	Research & Development	3	4
4	27	No	Travel_Rarely	591	Research & Development	2	1
5	32	No	Travel_Frequently	1005	Research & Development	2	2
6	59	No	Travel_Rarely	1324	Research & Development	3	3
7	30	No	Travel_Rarely	1358	Research & Development	24	1
8	38	No	Travel_Frequently	216	Research & Development	23	3
9	36	No	Travel_Rarely	1299	Research & Development	27	3
10	35	No	Travel_Rarely	809	Research & Development	16	3
11	29	No	Travel_Rarely	153	Research & Development	15	2
12	31	No	Travel_Rarely	670	Research & Development	26	1
13	34	No	Travel_Rarely	1346	Research & Development	19	2
14	28	Yes	Travel_Rarely	103	Research & Development	24	3
15	29	No	Travel_Rarely	1389	Research & Development	21	4
16	32	No	Travel_Rarely	334	Research & Development	5	2
17	22	No	Non-Travel	1123	Research & Development	16	2
18	53	No	Travel_Rarely	1219	Sales	2	4

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Edu
19	38	No	Travel_Rarely	371	Research & Development	2	3
20	24	No	Non-Travel	673	Research & Development	11	2
21	36	Yes	Travel_Rarely	1218	Sales	9	4
22	34	No	Travel_Rarely	419	Research & Development	7	4
23	21	No	Travel_Rarely	391	Research & Development	15	2
24	34	Yes	Travel_Rarely	699	Research & Development	6	1
25	53	No	Travel_Rarely	1282	Research & Development	5	3
26	32	Yes	Travel_Frequently	1125	Research & Development	16	1
27	42	No	Travel_Rarely	691	Sales	8	4
28	44	No	Travel_Rarely	477	Research & Development	7	4
29	46	No	Travel_Rarely	705	Sales	2	4
...	...	...	...	...	...	...	...
1440	36	No	Travel_Frequently	688	Research & Development	4	2
1441	56	No	Non-Travel	667	Research & Development	1	4
1442	29	Yes	Travel_Rarely	1092	Research & Development	1	4
1443	42	No	Travel_Rarely	300	Research & Development	2	3
1444	56	Yes	Travel_Rarely	310	Research & Development	7	2
1445	41	No	Travel_Rarely	582	Research & Development	28	4
1446	34	No	Travel_Rarely	704	Sales	28	3
1447	36	No	Non-Travel	301	Sales	15	4
1448	41	No	Travel_Rarely	930	Sales	3	3
1449	32	No	Travel_Rarely	529	Research & Development	2	3

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Edu
1450	35	No	Travel_Rarely	1146	Human Resources	26	4
1451	38	No	Travel_Rarely	345	Sales	10	2
1452	50	Yes	Travel_Frequently	878	Sales	1	4
1453	36	No	Travel_Rarely	1120	Sales	11	4
1454	45	No	Travel_Rarely	374	Sales	20	3
1455	40	No	Travel_Rarely	1322	Research & Development	2	4
1456	35	No	Travel_Frequently	1199	Research & Development	18	4
1457	40	No	Travel_Rarely	1194	Research & Development	2	4
1458	35	No	Travel_Rarely	287	Research & Development	1	4
1459	29	No	Travel_Rarely	1378	Research & Development	13	2
1460	29	No	Travel_Rarely	468	Research & Development	28	4
1461	50	Yes	Travel_Rarely	410	Sales	28	3
1462	39	No	Travel_Rarely	722	Sales	24	1
1463	31	No	Non-Travel	325	Research & Development	5	3
1464	26	No	Travel_Rarely	1167	Sales	5	3
1465	36	No	Travel_Frequently	884	Research & Development	23	2
1466	39	No	Travel_Rarely	613	Research & Development	6	1
1467	27	No	Travel_Rarely	155	Research & Development	4	3
1468	49	No	Travel_Frequently	1023	Sales	2	3
1469	34	No	Travel_Rarely	628	Research & Development	8	3

1470 rows × 35 columns

In [3]:
```
df.describe()
```

Out[3]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	Em
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	147
mean	36.923810	802.485714	9.192517	2.912925	1.0	102
std	9.135373	403.509100	8.106864	1.024165	0.0	602
min	18.000000	102.000000	1.000000	1.000000	1.0	1.0
25%	30.000000	465.000000	2.000000	2.000000	1.0	491
50%	36.000000	802.000000	7.000000	3.000000	1.0	102
75%	43.000000	1157.000000	14.000000	4.000000	1.0	155
max	60.000000	1499.000000	29.000000	5.000000	1.0	206

8 rows × 26 columns

calculating mean

In [4]:
```
df.mean()
```

Out[4]:
```
Age                          36.923810
DailyRate                   802.485714
DistanceFromHome              9.192517
Education                     2.912925
EmployeeCount                 1.000000
EmployeeNumber             1024.865306
EnvironmentSatisfaction       2.721769
HourlyRate                   65.891156
JobInvolvement                2.729932
JobLevel                      2.063946
JobSatisfaction               2.728571
MonthlyIncome              6502.931293
MonthlyRate               14313.103401
NumCompaniesWorked            2.693197
PercentSalaryHike            15.209524
PerformanceRating             3.153741
RelationshipSatisfaction      2.712245
StandardHours                80.000000
StockOptionLevel              0.793878
TotalWorkingYears            11.279592
TrainingTimesLastYear         2.799320
WorkLifeBalance               2.761224
YearsAtCompany                7.008163
YearsInCurrentRole            4.229252
YearsSinceLastPromotion       2.187755
YearsWithCurrManager          4.123129
dtype: float64
```

In [5]:
```
df.median()
```

Out[5]:
```
Age                          36.0
DailyRate                   802.0
DistanceFromHome              7.0
Education                     3.0
EmployeeCount                 1.0
EmployeeNumber             1020.5
EnvironmentSatisfaction       3.0
HourlyRate                   66.0
JobInvolvement                3.0
JobLevel                      2.0
JobSatisfaction               3.0
MonthlyIncome              4919.0
MonthlyRate               14235.5
NumCompaniesWorked            2.0
PercentSalaryHike            14.0
PerformanceRating             3.0
RelationshipSatisfaction      3.0
StandardHours                80.0
StockOptionLevel              1.0
TotalWorkingYears            10.0
TrainingTimesLastYear         3.0
WorkLifeBalance               3.0
YearsAtCompany                5.0
YearsInCurrentRole            3.0
YearsSinceLastPromotion       1.0
YearsWithCurrManager          3.0
dtype: float64
```

In [20]:
```
df.var()
```

Out[20]:
```
Age                         8.345505e+01
DailyRate                   1.628196e+05
DistanceFromHome            6.572125e+01
Education                   1.048914e+00
EmployeeCount               0.000000e+00
EmployeeNumber              3.624333e+05
EnvironmentSatisfaction     1.194829e+00
HourlyRate                  4.132856e+02
JobInvolvement              5.063193e-01
JobLevel                    1.225316e+00
JobSatisfaction             1.216270e+00
MonthlyIncome               2.216486e+07
MonthlyRate                 5.066288e+07
NumCompaniesWorked          6.240049e+00
PercentSalaryHike           1.339514e+01
PerformanceRating           1.301936e-01
RelationshipSatisfaction    1.169013e+00
StandardHours               0.000000e+00
StockOptionLevel            7.260346e-01
TotalWorkingYears           6.054056e+01
TrainingTimesLastYear       1.662219e+00
WorkLifeBalance             4.991081e-01
YearsAtCompany              3.753431e+01
YearsInCurrentRole          1.312712e+01
YearsSinceLastPromotion     1.038406e+01
YearsWithCurrManager        1.273160e+01
dtype: float64
```

In [7]:
```
df.std()
```

Out[7]:
```
Age                          9.135373
DailyRate                  403.509100
DistanceFromHome             8.106864
Education                    1.024165
EmployeeCount                0.000000
EmployeeNumber             602.024335
EnvironmentSatisfaction      1.093082
HourlyRate                  20.329428
JobInvolvement               0.711561
JobLevel                     1.106940
JobSatisfaction              1.102846
MonthlyIncome             4707.956783
MonthlyRate               7117.786044
NumCompaniesWorked           2.498009
PercentSalaryHike            3.659938
PerformanceRating            0.360824
RelationshipSatisfaction     1.081209
StandardHours                0.000000
StockOptionLevel             0.852077
TotalWorkingYears            7.780782
TrainingTimesLastYear        1.289271
WorkLifeBalance              0.706476
YearsAtCompany               6.126525
YearsInCurrentRole           3.623137
YearsSinceLastPromotion      3.222430
YearsWithCurrManager         3.568136
dtype: float64
```

In [8]:
```
df.sum()
```

Out[8]:
```
Age                                                        54278
Attrition                 YesNoYesNoNoNoNoNoNoNoNoNoNoYesNoNoNoNoNoNoY...
BusinessTravel            Travel_RarelyTravel_FrequentlyTravel_RarelyTra...
DailyRate                                                1179654
Department                SalesResearch & DevelopmentResearch & Developm...
DistanceFromHome                                           13513
Education                                                   4282
EducationField            Life SciencesLife SciencesOtherLife SciencesMe...
EmployeeCount                                               1470
EmployeeNumber                                           1506552
EnvironmentSatisfaction                                     4001
Gender                    FemaleMaleMaleFemaleMaleMaleFemaleMaleMaleMale...
HourlyRate                                                 96860
JobInvolvement                                              4013
JobLevel                                                    3034
JobRole                   Sales ExecutiveResearch ScientistLaboratory Te...
JobSatisfaction                                             4011
MaritalStatus             SingleMarriedSingleMarriedMarriedSingleMarried...
MonthlyIncome                                            9559309
MonthlyRate                                             21040262
NumCompaniesWorked                                          3959
Over18                    YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY...
OverTime                  YesNoYesYesNoNoYesNoNoNoNoYesNoNoYesNoYesYesNo...
PercentSalaryHike                                          22358
PerformanceRating                                           4636
RelationshipSatisfaction                                    3987
StandardHours                                             117600
StockOptionLevel                                            1167
TotalWorkingYears                                          16581
TrainingTimesLastYear                                       4115
WorkLifeBalance                                             4059
YearsAtCompany                                             10302
YearsInCurrentRole                                          6217
YearsSinceLastPromotion                                     3216
YearsWithCurrManager                                        6061
dtype: object
```

In [9]:
```
df.min()
```

Out[9]:
```
Age                                      18
Attrition                                No
BusinessTravel                    Non-Travel
DailyRate                                102
Department                   Human Resources
DistanceFromHome                           1
Education                                  1
EducationField               Human Resources
EmployeeCount                              1
EmployeeNumber                             1
EnvironmentSatisfaction                    1
Gender                                Female
HourlyRate                                30
JobInvolvement                             1
JobLevel                                   1
JobRole          Healthcare Representative
JobSatisfaction                            1
MaritalStatus                       Divorced
MonthlyIncome                           1009
MonthlyRate                             2094
NumCompaniesWorked                         0
Over18                                     Y
OverTime                                  No
PercentSalaryHike                         11
PerformanceRating                          3
RelationshipSatisfaction                   1
StandardHours                             80
StockOptionLevel                           0
TotalWorkingYears                          0
TrainingTimesLastYear                      0
WorkLifeBalance                            1
YearsAtCompany                             0
YearsInCurrentRole                         0
YearsSinceLastPromotion                    0
YearsWithCurrManager                       0
dtype: object
```

In [10]:
```
df.count()
```

Out[10]:
```
Age                         1470
Attrition                   1470
BusinessTravel              1470
DailyRate                   1470
Department                  1470
DistanceFromHome            1470
Education                   1470
EducationField              1470
EmployeeCount               1470
EmployeeNumber              1470
EnvironmentSatisfaction     1470
Gender                      1470
HourlyRate                  1470
JobInvolvement              1470
JobLevel                    1470
JobRole                     1470
JobSatisfaction             1470
MaritalStatus               1470
MonthlyIncome               1470
MonthlyRate                 1470
NumCompaniesWorked          1470
Over18                      1470
OverTime                    1470
PercentSalaryHike           1470
PerformanceRating           1470
RelationshipSatisfaction    1470
StandardHours               1470
StockOptionLevel            1470
TotalWorkingYears           1470
TrainingTimesLastYear       1470
WorkLifeBalance             1470
YearsAtCompany              1470
YearsInCurrentRole          1470
YearsSinceLastPromotion     1470
YearsWithCurrManager        1470
dtype: int64
```

In [11]:
```
df.max()
```

Out[11]:
```
Age                                60
Attrition                         Yes
BusinessTravel          Travel_Rarely
DailyRate                       1499
Department                     Sales
DistanceFromHome                  29
Education                          5
EducationField        Technical Degree
EmployeeCount                      1
EmployeeNumber                  2068
EnvironmentSatisfaction            4
Gender                          Male
HourlyRate                       100
JobInvolvement                     4
JobLevel                           5
JobRole          Sales Representative
JobSatisfaction                    4
MaritalStatus                 Single
MonthlyIncome                 19999
MonthlyRate                   26999
NumCompaniesWorked                 9
Over18                             Y
OverTime                         Yes
PercentSalaryHike                 25
PerformanceRating                  4
RelationshipSatisfaction           4
StandardHours                     80
StockOptionLevel                   3
TotalWorkingYears                 40
TrainingTimesLastYear              6
WorkLifeBalance                    4
YearsAtCompany                    40
YearsInCurrentRole                18
YearsSinceLastPromotion           15
YearsWithCurrManager              17
dtype: object
```

In [12]:
```
df.describe(include='all')
```

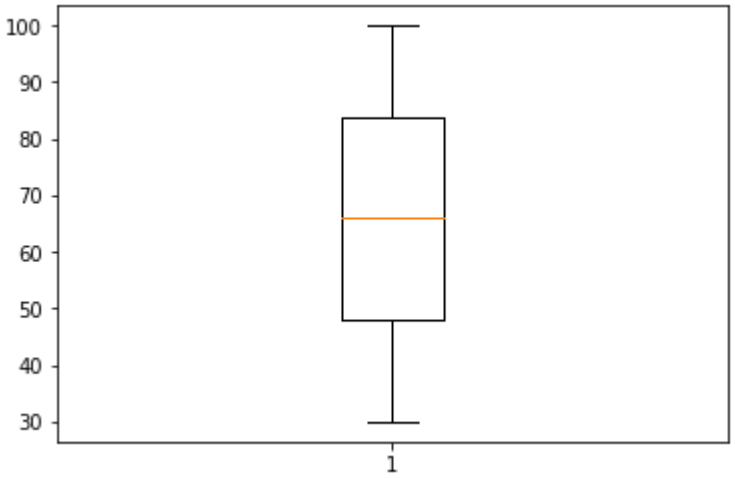
Out[12]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFrom
count	1470.000000	1470	1470	1470.000000	1470	1470.000000
unique	NaN	2	3	NaN	3	NaN
top	NaN	No	Travel_Rarely	NaN	Research & Development	NaN
freq	NaN	1233	1043	NaN	961	NaN
mean	36.923810	NaN	NaN	802.485714	NaN	9.192517
std	9.135373	NaN	NaN	403.509100	NaN	8.106864
min	18.000000	NaN	NaN	102.000000	NaN	1.000000
25%	30.000000	NaN	NaN	465.000000	NaN	2.000000
50%	36.000000	NaN	NaN	802.000000	NaN	7.000000
75%	43.000000	NaN	NaN	1157.000000	NaN	14.000000
max	60.000000	NaN	NaN	1499.000000	NaN	29.000000

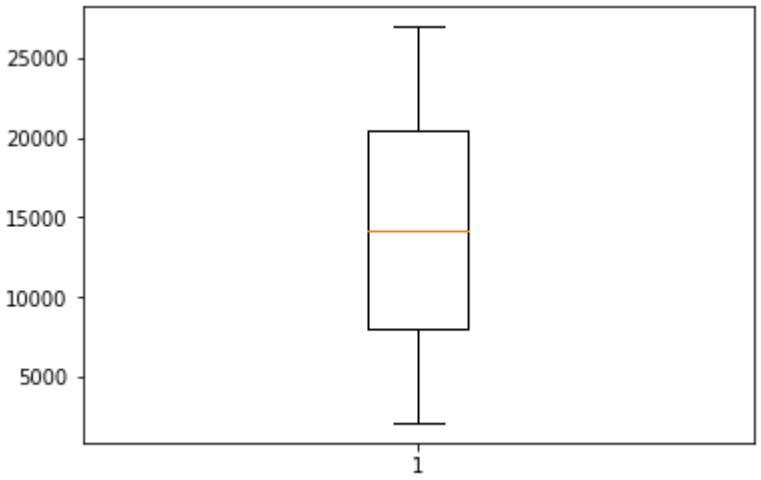
11 rows × 35 columns

creating a boxplot of all the numerical attributes using pandas boxplot()

In [16]:
```
import matplotlib.pyplot as pl
pl.boxplot(df['HourlyRate'])
pl.show()
```



In [18]:
```
pl.boxplot(df['MonthlyRate'])
pl.show()
```



In [19]:
```
df.boxplot(figsize='30,10')
```

Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x72ed990>