Exercise : descriptive summary statistics

Dataset : EmployeeAttrition.csv

In [1]:
```
import pandas
import matplotlib.pyplot as plot
```

creating a dataframe from the csv file

In [2]:
```
df = pandas.read_csv("Y:\DA LAB\EmployeeAttrition.csv")
```

creating a summary statistics using describe()

In [3]:
```
df.describe(include='all')
```

Out[3]:

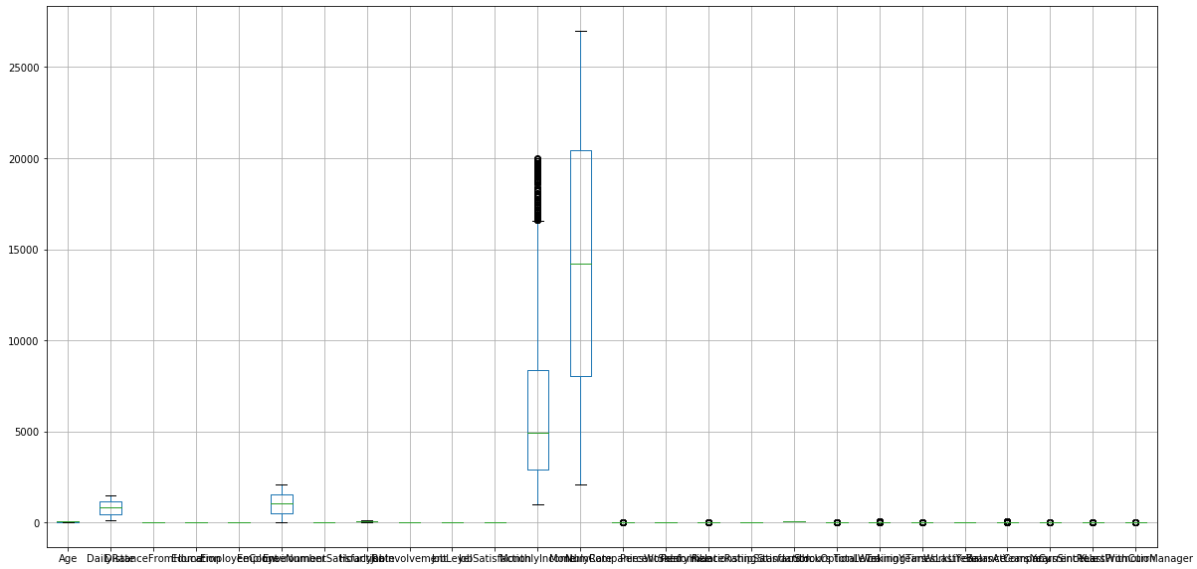
	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFrom
count	1470.000000	1470	1470	1470.000000	1470	1470.000000
unique	NaN	2	3	NaN	3	NaN
top	NaN	No	Travel_Rarely	NaN	Research & Development	NaN
freq	NaN	1233	1043	NaN	961	NaN
mean	36.923810	NaN	NaN	802.485714	NaN	9.192517
std	9.135373	NaN	NaN	403.509100	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

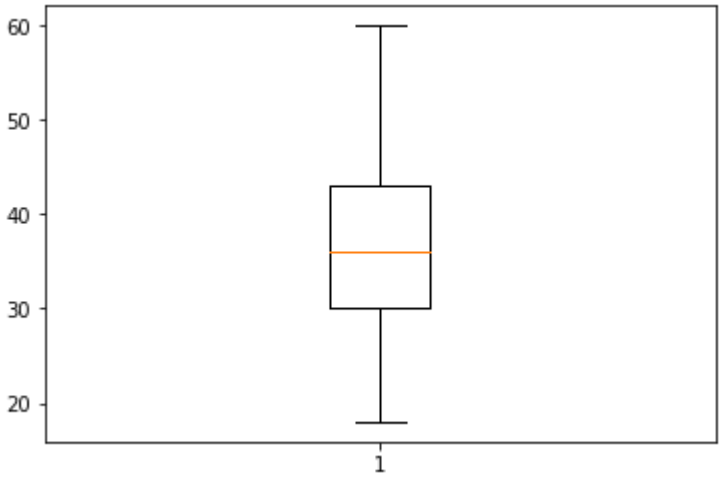
creating a boxplot of all numeric attributes using pandas boxplot()

In [4]:
```
df.boxplot(figsize='20,10')
```

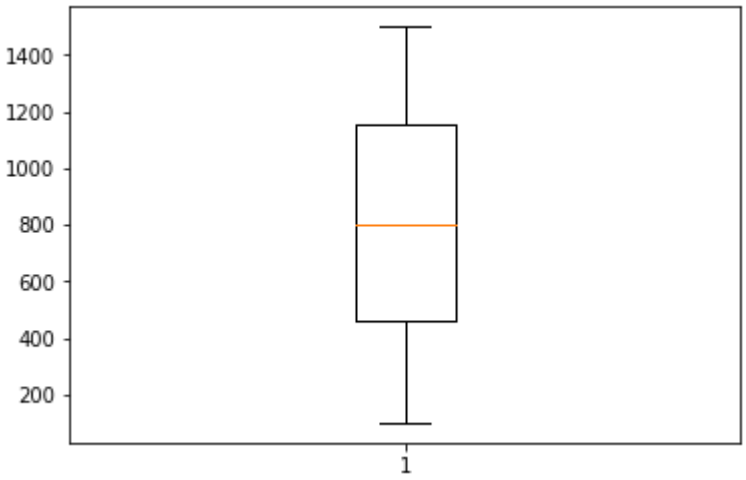
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x62b3910>



In [5]:
```
plot.boxplot(df['Age'])
plot.show()
```



In [6]:
```
plot.boxplot(df['DailyRate'])
plot.show()
```



Checking if there are any missing values

In [7]:
```
df.isnull().any()
```

Out[7]:
```
Age                         False
Attrition                   False
BusinessTravel              False
DailyRate                   False
Department                  False
DistanceFromHome            False
Education                   False
EducationField              False
EmployeeCount               False
EmployeeNumber              False
EnvironmentSatisfaction     False
Gender                      False
HourlyRate                  False
JobInvolvement              False
JobLevel                    False
JobRole                     False
JobSatisfaction             False
MaritalStatus               False
MonthlyIncome               False
MonthlyRate                 False
NumCompaniesWorked          False
Over18                      False
OverTime                    False
PercentSalaryHike           False
PerformanceRating           False
RelationshipSatisfaction    False
StandardHours               False
StockOptionLevel            False
TotalWorkingYears           False
TrainingTimesLastYear       False
WorkLifeBalance             False
YearsAtCompany              False
YearsInCurrentRole          False
YearsSinceLastPromotion     False
YearsWithCurrManager        False
dtype: bool
```

Creating a scatter plot of Age vs Attrition

In [9]:
```
plot.scatter(df['Age'],df['Attrition'])
plot.show()
```