Introduction to Probability and Statistics

Assignment

In this assignment, we will use the dataset of diabetes patients taken from here.

```
import pandas as pd
import numpy as np

df =
pd.read_csv("https://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.
txt",sep='\t')
df.head()
```

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
0	59	2	32.1	101.0	157	93.2	38.0	4.0	4.8598	87	151
1	48	1	21.6	87.0	183	103.2	70.0	3.0	3.8918	69	75
2	72	2	30.5	93.0	156	93.6	41.0	4.0	4.6728	85	141
3	24	1	25.3	84.0	198	131.4	40.0	5.0	4.8903	89	206
4	50	1	23.0	101.0	192	125.4	52.0	4.0	4.2905	80	135

In this dataset, columns as the following:

- Age and sex are self-explanatory
- BMI is body mass index
- BP is average blood pressure
- S1 through S6 are different blood measurements
- Y is the qualitative measure of disease progression over one year

Let's study this dataset using methods of probability and statistics.

Task 1: Compute mean values and variance for all values

```
print("Mean : ")
print(np.mean(df))
print("")
print("Variance : ")
print(np.var(df))
```

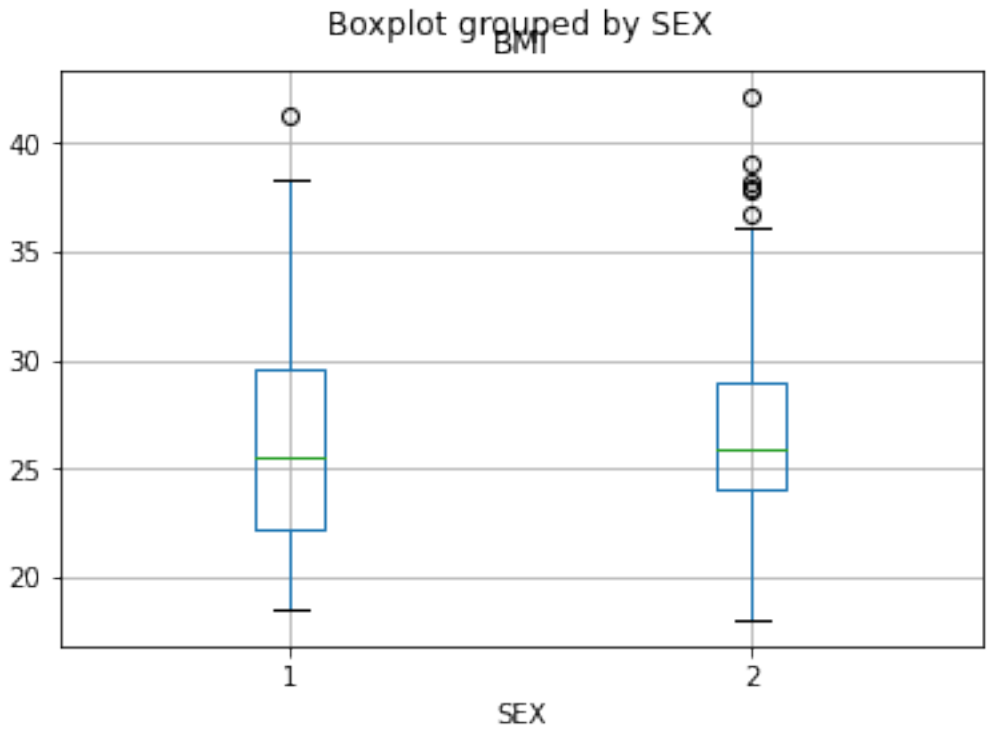
```
Mean :
AGE     48.518100
SEX      1.468326
BMI     26.375792
BP      94.647014
S1     189.140271
S2     115.439140
S3      49.788462
S4       4.070249
S5       4.641411
S6      91.260181
```

```
Y      152.133484
dtype: float64

Variance :
AGE     171.457817
SEX       0.248997
BMI      19.475636
BP      190.871586
S1     1195.007473
S2      922.862835
S3      166.915093
S4        1.661493
S5        0.272274
S6      131.866695
Y      5929.884897
dtype: float64
```

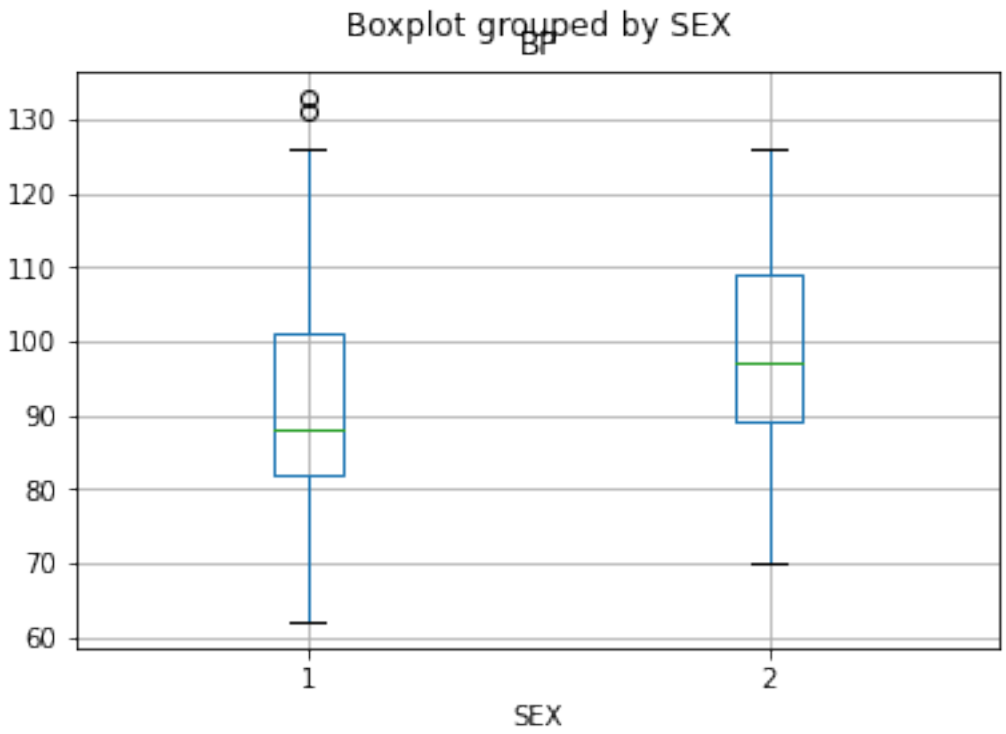
Task 2: Plot boxplots for BMI, BP and Y depending on gender

```
import matplotlib.pyplot as plt
df.boxplot(column='BMI',by='SEX')
plt.show()
```



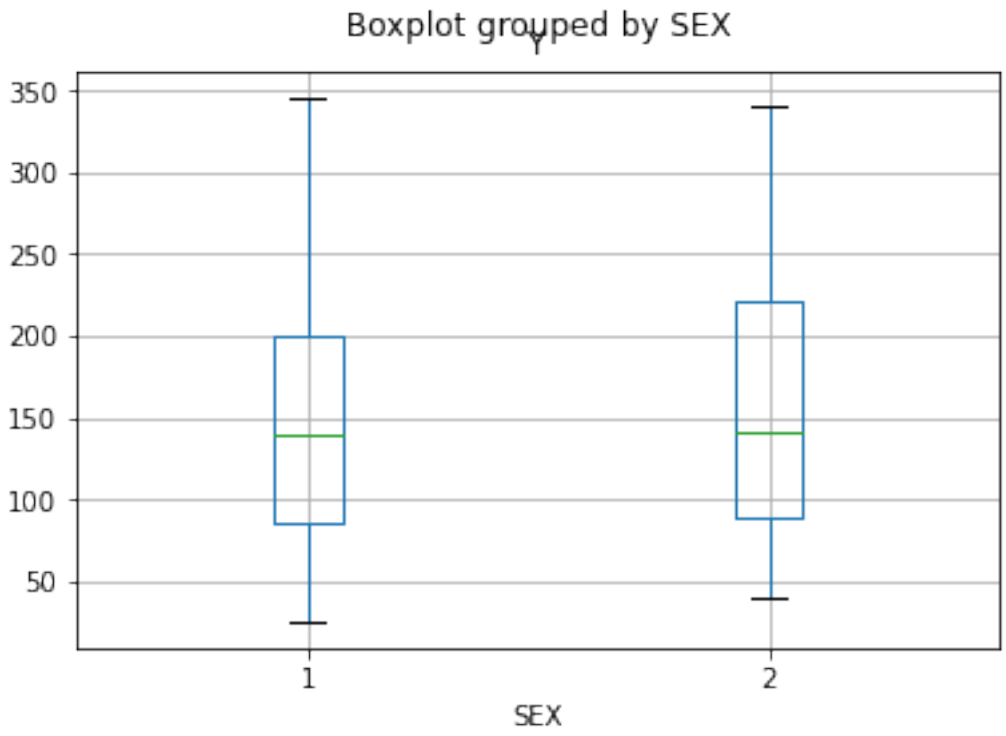
We could see that there's so many outlier for Sex 2 and an outlier for Sex 1 on this BMI's boxplot.

```
df.boxplot(column='BP',by='SEX')
plt.show()
```



We could see that there's only two outlier for Sex 1 and no outlier for Sex 2 on this BP's boxplot.

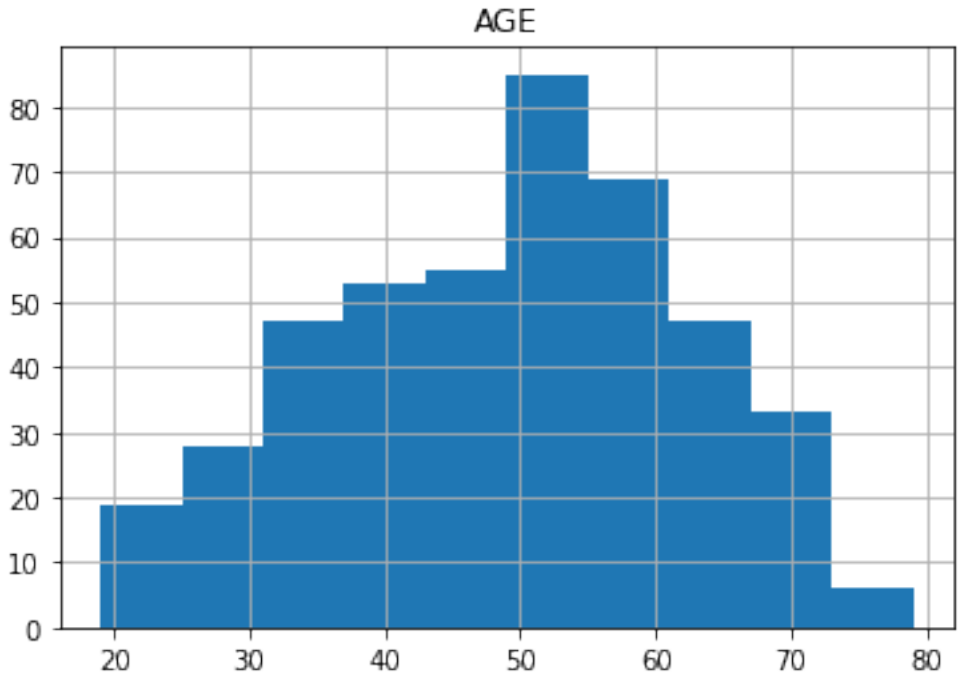
```
df.boxplot(column='Y',by='SEX')
plt.show()
```



We could see that there's no outlier for both Sex 1 and Sex 2 on this Y's boxplot.

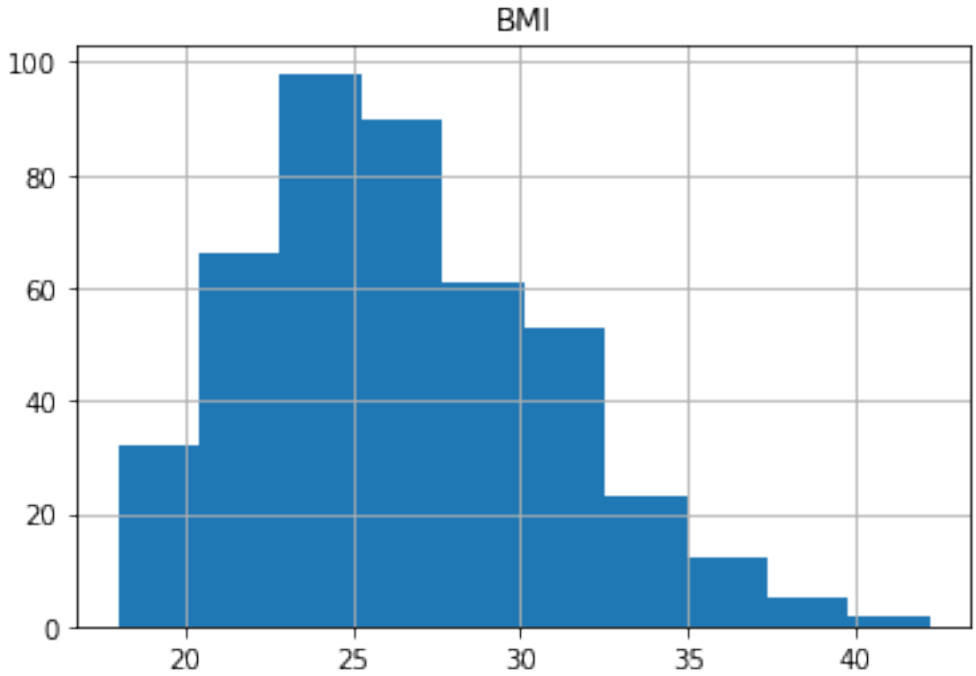
Task 3: What is the the distribution of Age, Sex, BMI and Y variables?

```
df.hist(column='AGE')
plt.show()
```



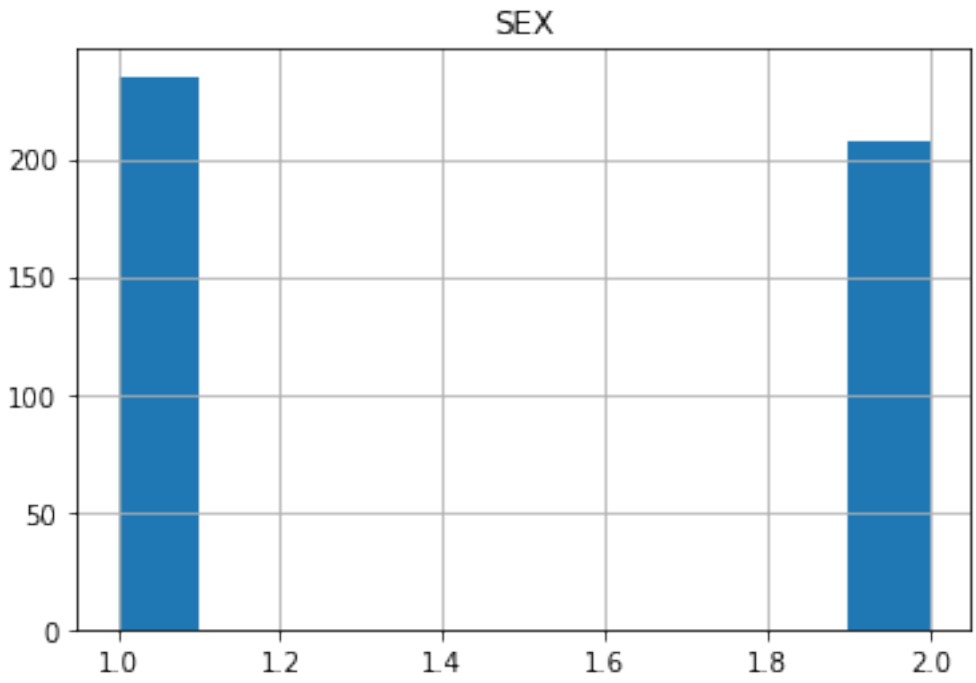
The histogram of Age above have a left-skewed distribution. As we can see, the age that most often appears in the data is 50 years old, but the mean is 48.52 , the mean of the data is to the left of the peak so the data is left-skewed distribution.

```
df.hist(column='BMI')
plt.show()
```



The histogram of BMI above have a right-skewed distribution. As we can see, the BMI that most often appears in the data is 24, but the mean is 26.37 , the mean of the data is to the right of the peak so the data is right-skewed distribution.

```
df. hist(column='SEX')
plt.show()
```



The histogram of SEX above only gathered in numbers 1 and 2. This is because numbers 1 and 2 represent 2 types of gender and are not numerical data.