Breast Cancer Data Set - Analysis I

Load Dataset

```
table = readtable("dataR2.csv");
```

Display details of the dataset such as attribute names, number of samples

```
summary(table);
```

```
Variables:

    Age: 116×1 double

        Properties:
            Description:  Age
        Values:

            Min          24
            Median       56
            Max          89

    BMI: 116×1 double

        Properties:
            Description:  BMI
        Values:

            Min         18.37
            Median     27.662
            Max        38.579

    Glucose: 116×1 double

        Properties:
            Description:  Glucose
        Values:

            Min          60
            Median       92
            Max         201

    Insulin: 116×1 double

        Properties:
            Description:  Insulin
        Values:

            Min         2.432
            Median     5.9245
            Max         58.46

    HOMA: 116×1 double

        Properties:
            Description:  HOMA
        Values:

            Min       0.46741
            Median     1.3809
            Max         25.05
    Leptin: 116×1 double

        Properties:
            Description:  Leptin
        Values:

            Min         4.311
            Median     20.271
            Max         90.28

    Adiponectin: 116×1 double

        Properties:
            Description:  Adiponectin
        Values:

            Min         1.656
            Median     8.3527
            Max         38.04

    Resistin: 116×1 double

        Properties:
            Description:  Resistin
        Values:

            Min          3.21
            Median     10.828
            Max          82.1

    MCP_1: 116×1 double

        Properties:
            Description:  MCP.1
        Values:

            Min        45.843
            Median     471.32
            Max        1698.4

    Classification: 116×1 double

        Properties:
            Description:  Classification
        Values:

            Min          1
            Median       2
            Max          2
```

Display first five records of the table

```
disp(table(1:5, :));
```

```
    Age     BMI      Glucose    Insulin     HOMA      Leptin    Adiponectin    Resistin    MCP_1     Classification
    ___    ______    _______    _______    _______    ______    ___________    ________    ______    ______________

    48       23.5      70        2.707     0.46741    8.8071      9.7024        7.9958     417.11          1
    83      20.69      92        3.115      0.7069    8.8438      5.4293         4.064     468.79          1
    82     23.125      91        4.498      1.0097    17.939      22.432        9.2772      554.7          1
    68      21.368      77         3.226     0.61272    9.8827      7.1696         12.766     928.22            1
    86      21.111      92         3.549     0.80539    6.6994      4.8192         10.576     773.92            1
```

Getting numeric values from the table

```
tablenum = table2array(table(:, 1:end-1)); %classification is excluded since it is
categorical value

%displaying first five records of numeric values
disp(tablenum(1:5, :));
```

```
   48.0000   23.5000   70.0000    2.7070    0.4674    8.8071    9.7024    7.9958  417.1140
   83.0000   20.6905   92.0000    3.1150    0.7069    8.8438    5.4293    4.0640  468.7860
   82.0000   23.1247   91.0000    4.4980    1.0097   17.9393   22.4320    9.2772  554.6970
   68.0000   21.3675   77.0000    3.2260    0.6127    9.8827    7.1696   12.7660  928.2200
   86.0000   21.1111   92.0000    3.5490    0.8054    6.6994    4.8192   10.5763  773.9200
```

Display mean and standard deviation vector

```
mean_vector = mean(tablenum);
disp(mean_vector)
```

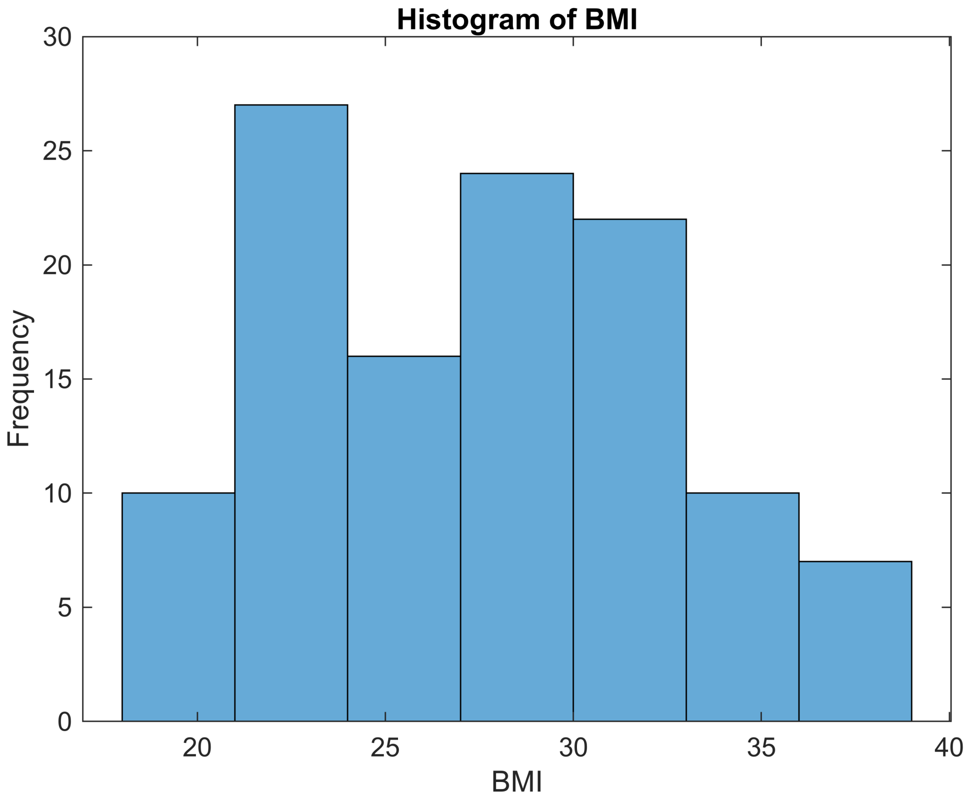
```
   57.3017   27.5821   97.7931   10.0121    2.6950   26.6151   10.1809   14.7260  534.6470
```

```
std_vector = std(tablenum);
disp(std_vector)
```

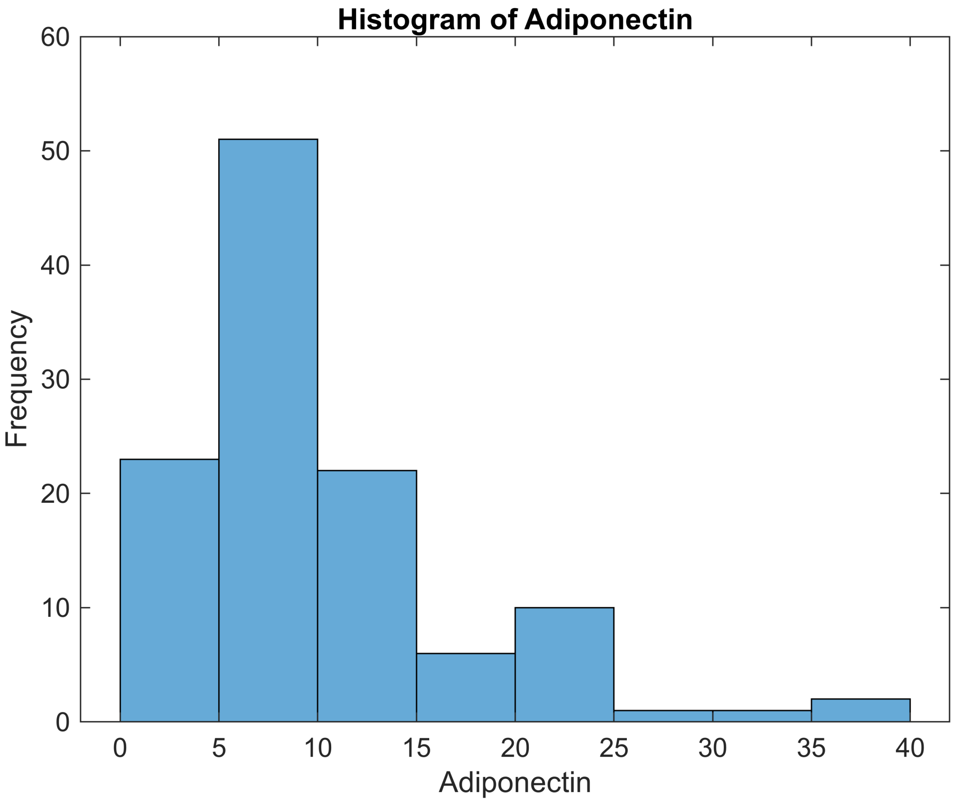
```
   16.1128    5.0201   22.5252   10.0678    3.6420   19.1833    6.8433   12.3906  345.9127
```

Display histogram of at least one attributes

```
% histogram of BMI

figure;
histogram(tablenum(:,2));

xlabel(table.Properties.VariableNames{2})
ylabel('Frequency')
title(['Histogram of ', table.Properties.VariableNames{2}]);
```

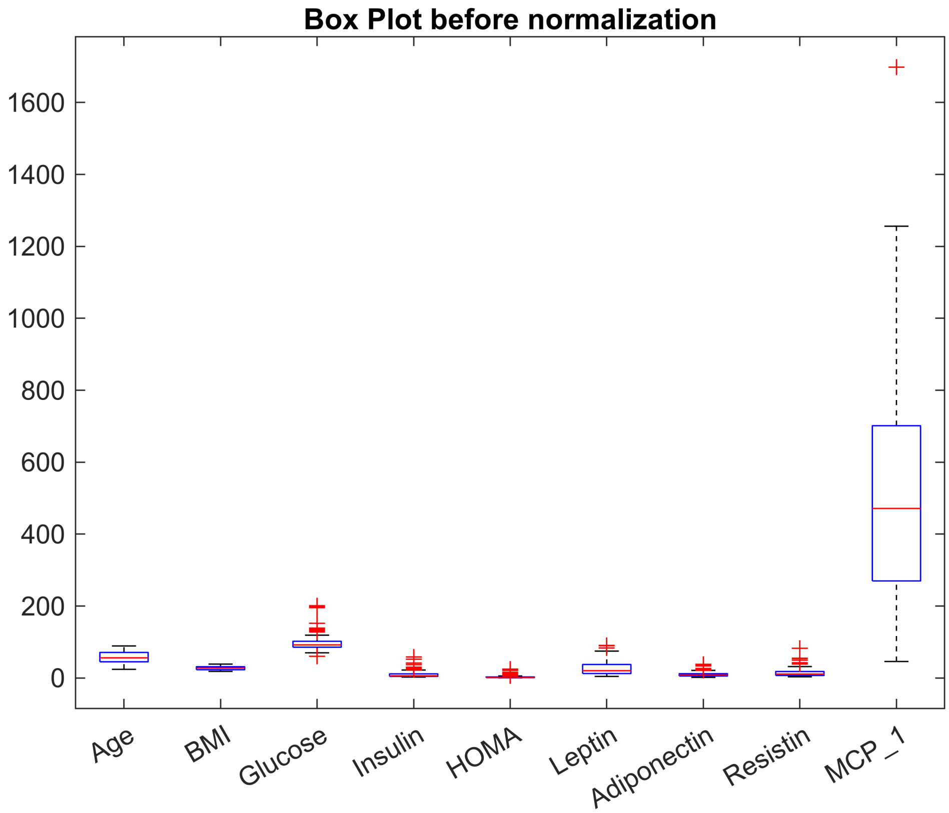


```
% histogram of Adiponectin

figure;
histogram(tablenum(:,7));

xlabel(table.Properties.VariableNames{7})
ylabel('Frequency')
title(['Histogram of ', table.Properties.VariableNames{7}]);
```



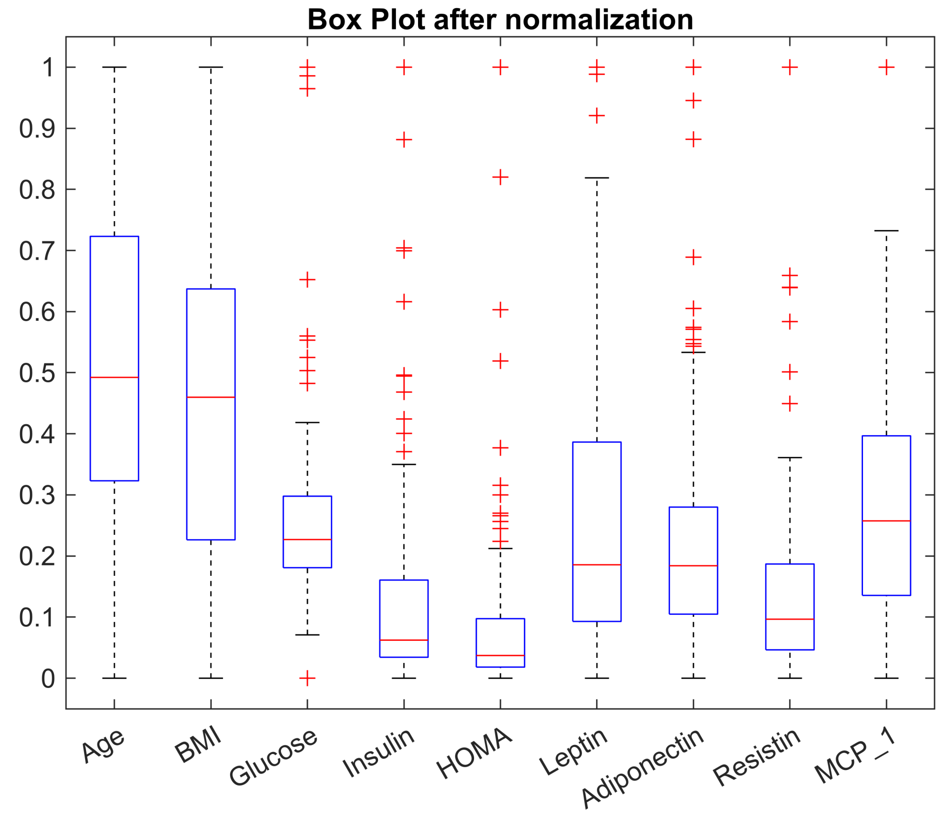
Display the box plot

```
figure;
boxplot(tablenum,'Labels', table.Properties.VariableNames(1:end-1));
title('Box Plot before normalization')
```



The boxplot shows that the variables have different scale. Let's normalize the data to have better comparison across attributes

```matlab
% Find minimum and maximum values for each attribute
tab_min = min(tablenum);
tab_max = max(tablenum);

% Compute the range for each attribute
temp = tab_max - tab_min;

% Initialize the normalized feature matrix
n_tab = zeros(size(tablenum));

% Perform min-max normalization for each attribute
for i = 1 : size(tablenum, 2)
    n_tab(:, i) = (tablenum(:, i) - tab_min(i)) ./ temp(i);
end

% tablenum = normalize(tablenum);
% % tablenum = (tablenum-mean_vector)./std_vector;
%
figure;
boxplot(n_tab, 'Labels', table.Properties.VariableNames(1:end-1));
title('Box Plot after normalization');
```


Box Plot after normalization

```
%displaying first five records of normalized numeric values
disp(n_tab(1:5, :));
```

```
    0.3692    0.2539    0.0709    0.0049         0    0.0523    0.2212    0.0607    0.2247
    0.9077    0.1148    0.2270    0.0122    0.0097    0.0527    0.1037    0.0108    0.2559
    0.8923    0.2353    0.2199    0.0369    0.0221    0.1585    0.5710    0.0769    0.3079
    0.6769    0.1483    0.1206    0.0142    0.0059    0.0648    0.1515    0.1211    0.5339
    0.9538    0.1356    0.2270    0.0199    0.0137    0.0278    0.0869    0.0934    0.4406
```