IE 6318 Data Mining and Analytics

Homework 1

Data Exploration

Submitted by

Sarika Kshatriya

[Sarika.kshatriya@mavs.uta.edu](mailto:Sarika.kshatriya@mavs.uta.edu)

1.Iris Dataset :

Used Iris dataset provided in Blackboard. The attributes of the dataset are as follows:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:   
   -- Iris Setosa   
   -- Iris Versicolour   
   -- Iris Virginica

Each attribute consists of 50 instances per class as there are three different classes, therefore each feature consists of 150 readings. Replaced the names of the three classes with numbers 1,2 and 3 and saved the file in csv format.

After loading the csv file into Matlab separated all the columns.

**Code:**

% Converted the given Iris dataset excel to csv file

load Iris.csv;

%replaced the names of the three classes by numerical numbers as:

% 1- Iris Setosa ,2- Iris Versicolour,3- Iris Virginica

Class= [1,2,3];

Y =[Class];

X = Iris (:,1:4);

Y = Iris (:,5);

Sepal\_length = Iris(:,1);

Sepal\_width = Iris (:,2);

Petal\_length = Iris (:,3);

Petal\_width = Iris (:,4);

class = Y;

names = char('Sepal length','Sepal width','Petal length','Petal width');

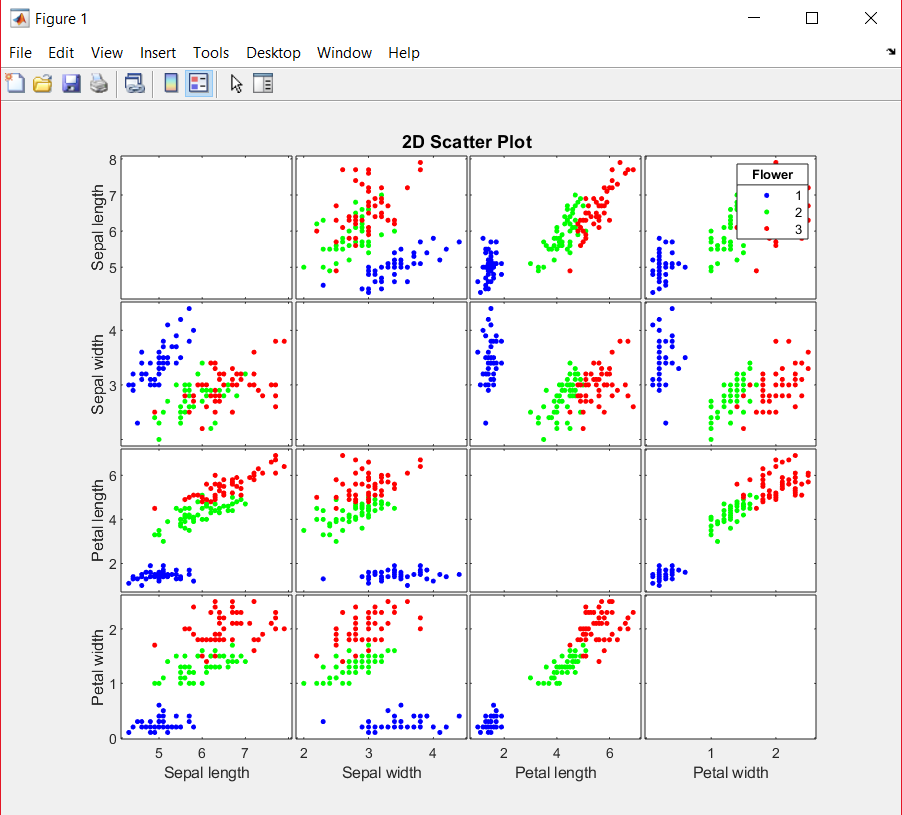
2.Explore the Iris dataset:

1) 2D scatter plots of the four attributes.

**Code:**

%2d scatter plot for the four attributes

gplotmatrix(X,[],Y,'','',[],'on','none',names,[]);



The above 2D scatter plot of the four attributes shows how they are grouped with respect to the classes. In gplotmatrix function X is the data set used by the four attributes and Y is the grouping variable for the 3 classes. The three classes are shown in different color.

From plot we can conclude class setosa is independent, while class virginica and versicolor are correlated.

2) 3D scatter plot of three attributes (sepal length, sepal width, petal width).

**Code**:

%3d scatter plot

scatter3(Sepal\_length(1:50),Sepal\_width(1:50),Petal\_width(1:50),'r');

hold on;

scatter3(Sepal\_length(50:100),Sepal\_width(50:100),Petal\_width(50:100),'g');

hold on;

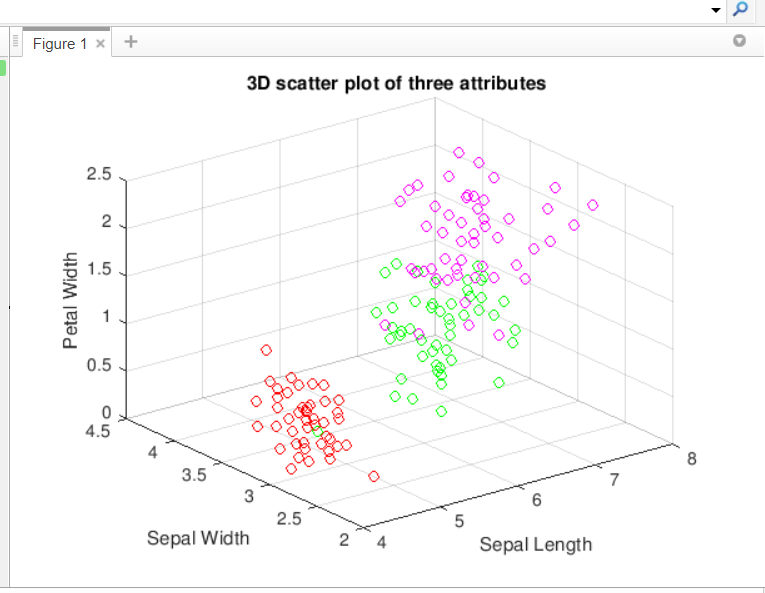
scatter3(Sepal\_length(100:150),Sepal\_width(100:150),Petal\_width(100:150),'magenta');

title('3D scatter plot of three attributes ');

xlabel('Sepal Length');

ylabel('Sepal Width');

zlabel('Petal Width');



The above 3d plot shows relationship between 3 attributes. setosa is having upward trend but not spread widely. In contrast to setosa class the versicolor and veginica spread wider.. In conclusion, we can say that there is a good correlation between the attribute’s sepal length, sepal width and petal width.

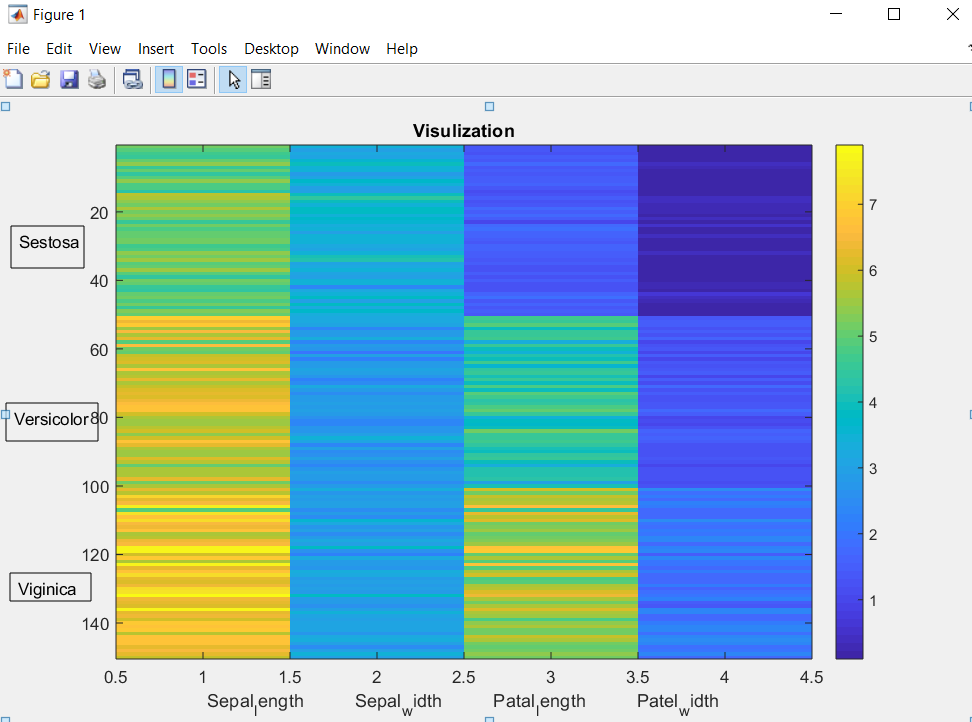
3) Visualization of the feature matrix (column 1-4).

**Code**:

V=[sepal\_length,sepal\_width,petal\_length,petal\_width];

imagesc(V);

colorbar;



The above figure illustrates the standardized data matrix for Iris data set. As the data matrix is rectangular array of values we can visualize it as an image by associating each entry of the data matrix with a pixel in the image. The color of pixel is determined by the value of the corresponding entry of the matrix.To sum up, this visualization of the data matrix is a technique to explore the associations between variables and their interactions where reduction of data dimensions is not necessary.

4) Histogram of the four attributes for the three classes.

a. Histogram for Sepal\_length for the 3 classes.

**Code**:

%Histogram for Sepal\_length for the 3 classes

h1= Iris(1:50,1);

h2= Iris(51:100,1);

h3= Iris(101:150,1);

histogram(h1);

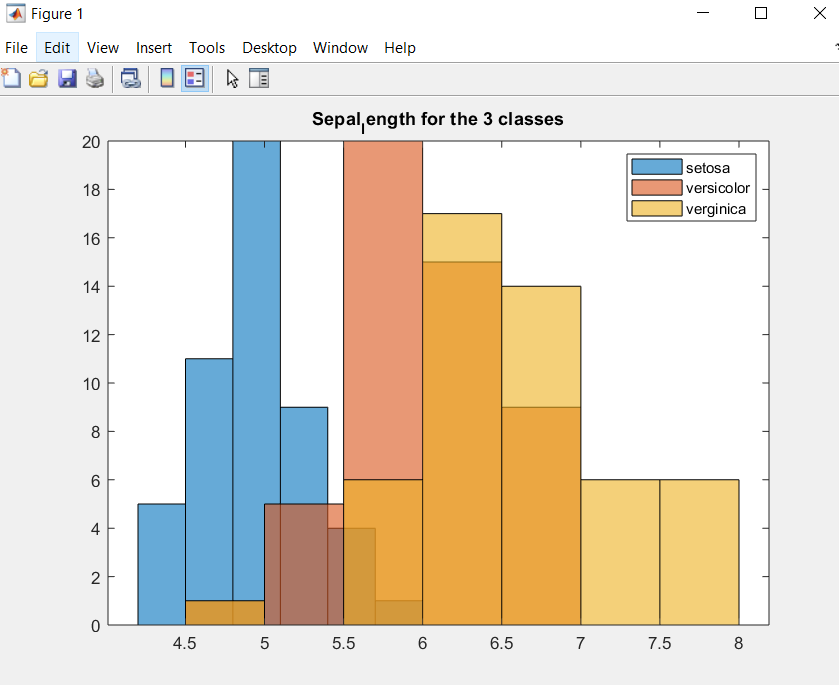
hold on;

histogram(h2);

hold on;

histogram(h3);

legend('setosa','versicolor','verginica')



b. Histogram for Sepal\_width for the 3 classes.

**Code**:

%Histogram for Sepal\_width for the 3 classes

h1= Iris(1:50,2);

h2= Iris(51:100,2);

h3= Iris(101:150,2);

histogram(h1);

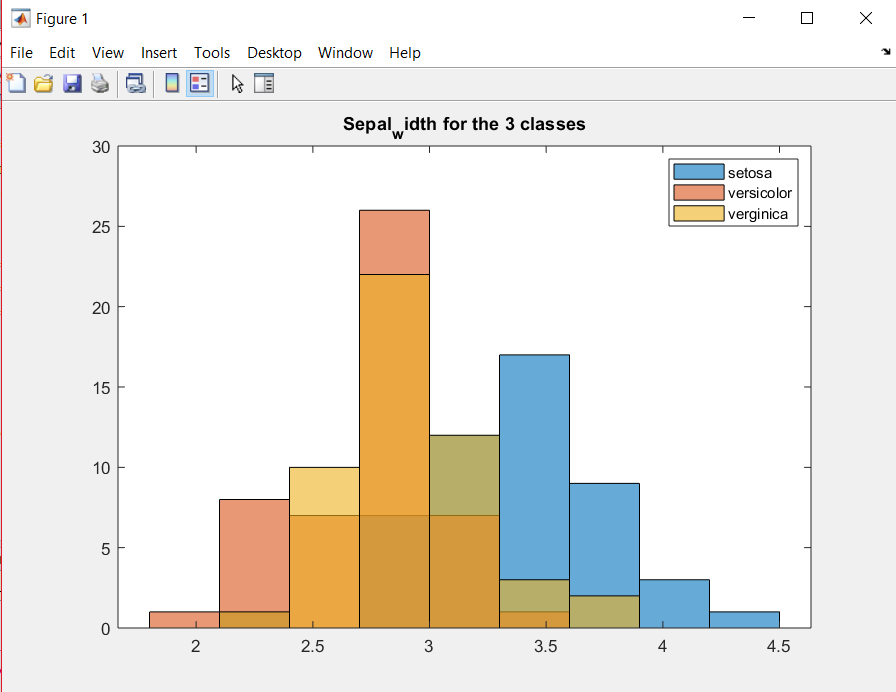
hold on;

histogram(h2);

hold on;

histogram(h3);

legend('setosa','versicolor','verginica')



c. Histogram for Petal\_length for the 3 classes.

**Code**:

%Histogram for Petal\_length for the 3 classes

h1= Iris(1:50,3);

h2= Iris(51:100,3);

h3= Iris(101:150,3);

histogram(h1);

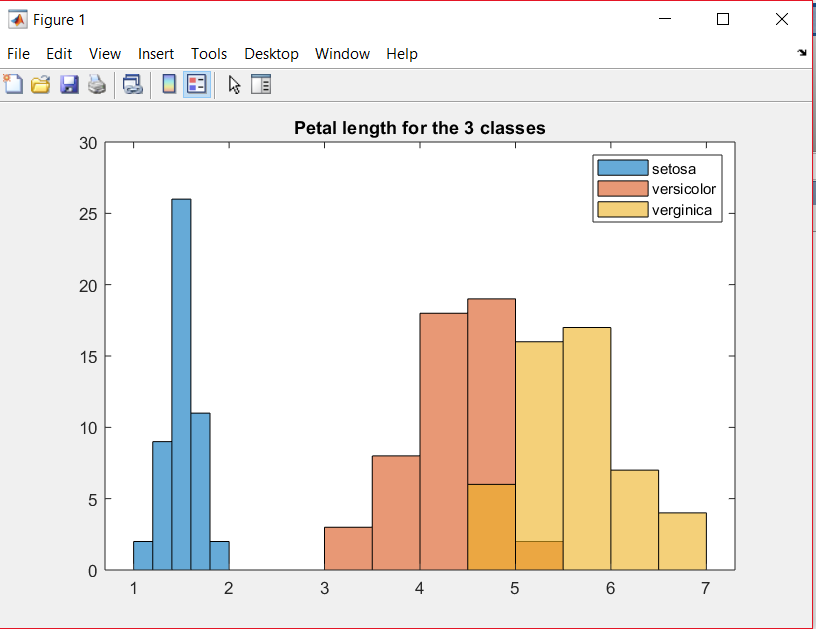
hold on;

histogram(h2);

hold on;

histogram(h3);

legend('setosa','versicolor','verginica')



d. Histogram for Petal\_width for the 3 classes.

**Code**:

%Histogram for Petal\_width for the 3 classes

h1= Iris(1:50,4);

h2= Iris(51:100,4);

h3= Iris(101:150,4);

histogram(h1);

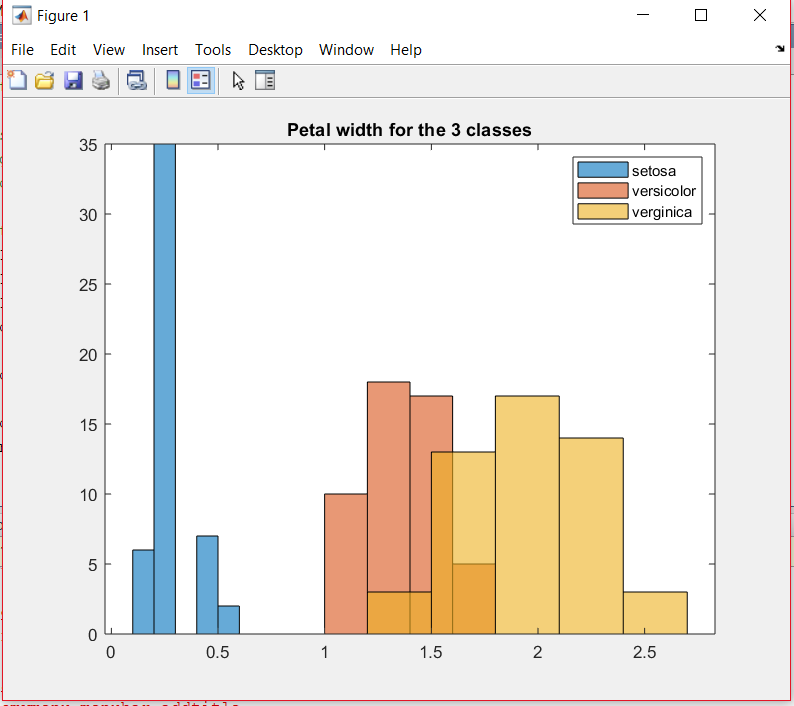
hold on;

histogram(h2);

hold on;

histogram(h3);

legend('setosa','versicolor','verginica')



A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of [continuous](https://statistics.laerd.com/statistical-guides/types-of-variable.php) data. This allows the inspection of the data for its underlying distribution. The above 4 Histograms splits the data into intervals, called bins. In this case, class has been split into bins.

We can notice that, there are no "gaps" between the bars (although some bars might be "absent"). This is because a histogram represents a continuous data set, and as such, there are no gaps in the histograms.

From all above histograms we can say that it is possible to classify iris data as class setosa, versicolor and Virginica by sepal length, sepal width, petal length and petal width.

5) Boxplots of the four attributes for the three classes.

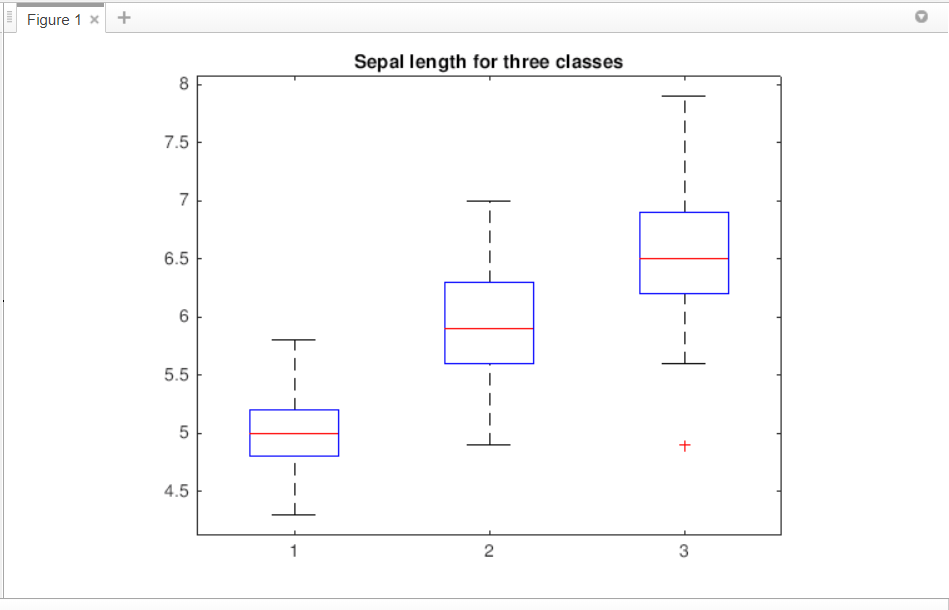
Boxplots give us mean, median, first and third quartiles. Also shows us the outlies if any them are present. Box plots can be used to compare attributes.

**a.** Boxplot for the Sepal\_length

**Code:**

% Boxplot for the Sepal\_length

boxplot(sepal\_length,Y);



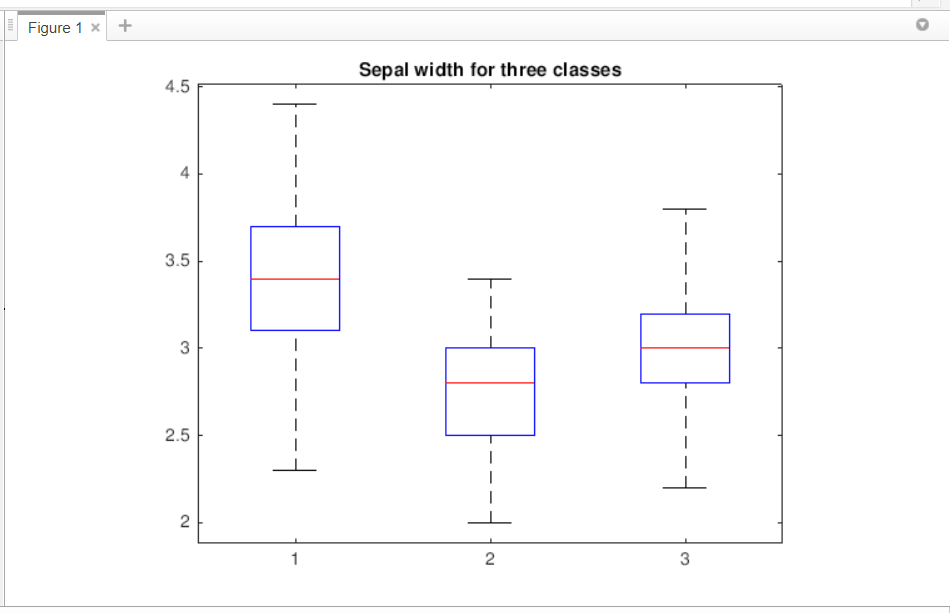
In the above boxplot, the whisker for virginica is farthest to the top than others. Also, virginica contains an outliner while setosa and versicolor don’t show any. The median line in the virginica plot is not centered in the box, which means the sample is slightly skewed.

**b.** Boxplot for the Sepal\_width

**Code:**

% Boxplot for the Sepal\_width

boxplot(Sepal\_width,Y);



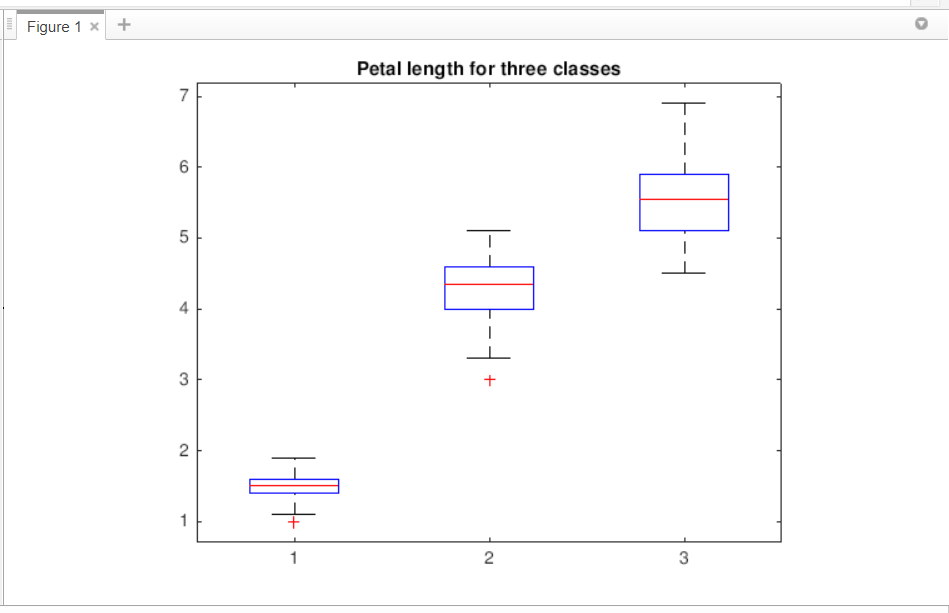
In the above boxplot, the median line in the versicolor is not centered in the box, which indicated that the sample is slightly skewed. Also, there is no any outliers.

**c.** Boxplot for the Petal\_length

**Code:**

% Boxplot for the Petal\_length

boxplot(Petal\_length,Y);



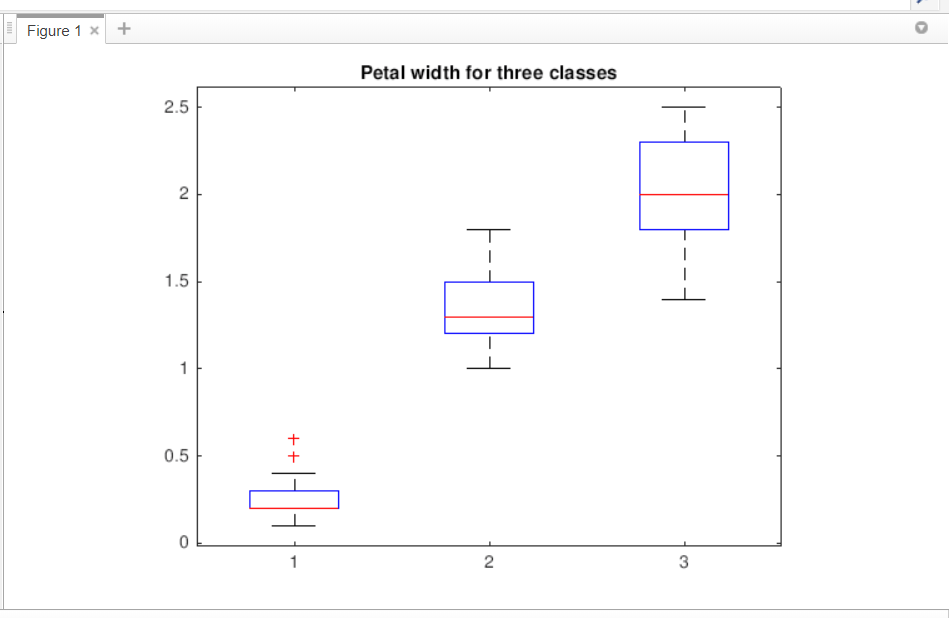
In the above boxplot, setosa and versicolor show outliers and whiskers for setosa has a low range. We can conclude that samples not skewed.

**d.** Boxplot for the Petal\_width

**Code:**

% Boxplot for the Petal\_width

boxplot(Petal\_width,Y);

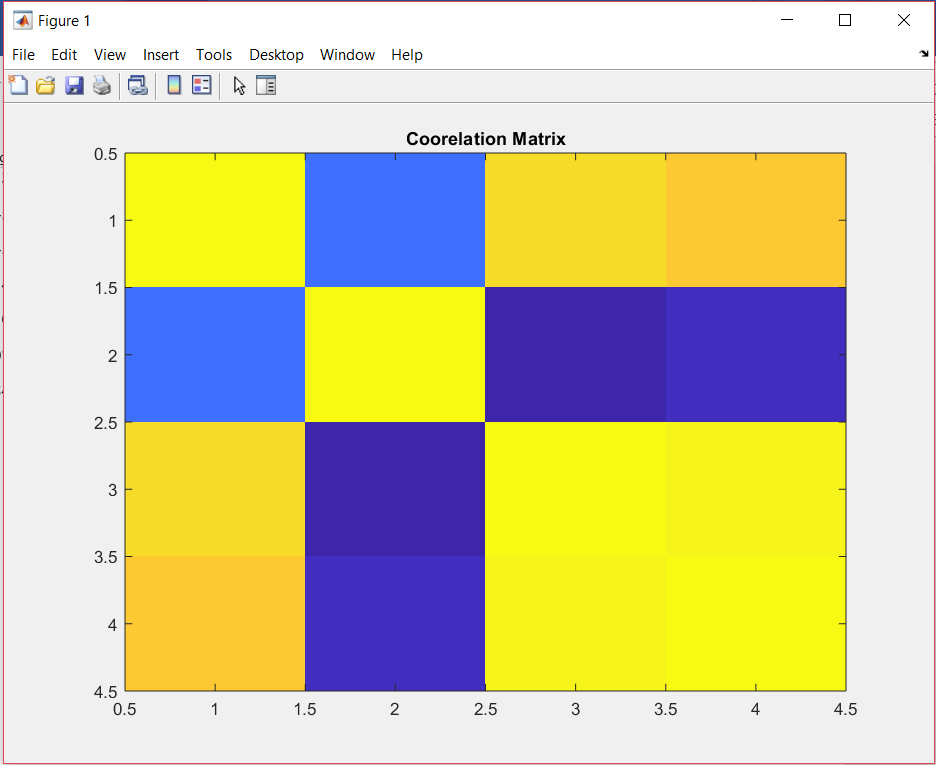


In the above boxplot, none of the classes have their median lines centered to the box and hence the samples are skewed.

6) Calculate the correlation matrix of the four attributes and visualize the correlation matrix.

**Code:**

imagesc(corrcoef(X));



A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two or more variables.

The direction of a correlation is either positive or negative. In a negative correlation, the variables move in inverse, or opposite, directions. In other words, as one variable increases, the other variable decreases.

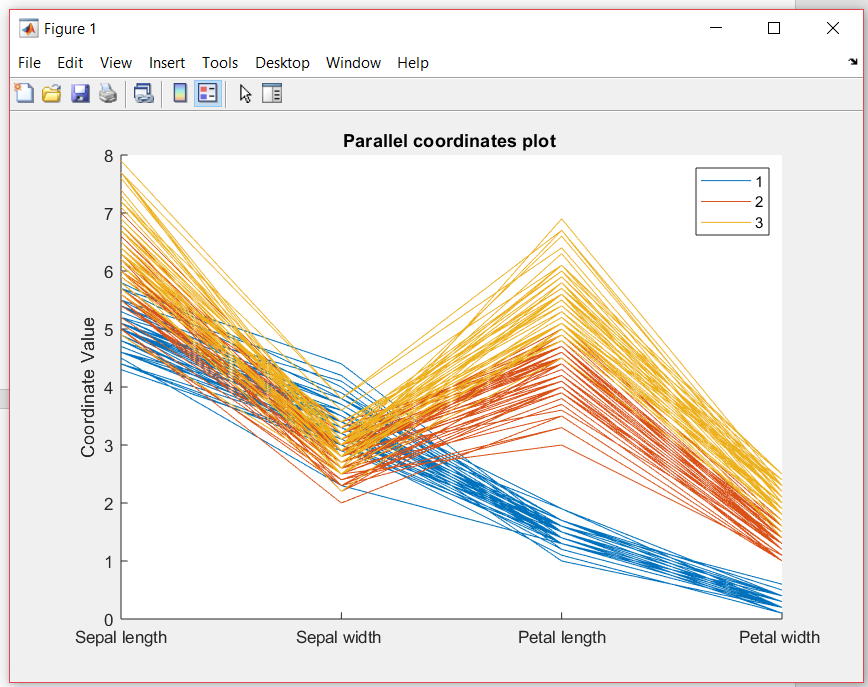
The above figure shows statistical relationship between four variables. The positive correlations are displayed in blue and negative correlations are displayed in yellow color. Hence, from correlation matrix we can predict the positive or negative correlations.

7) Parallel coordinates plot of the four attributes.

**Code**:

%Parallel coordinates plot of the four attributes

parallelcoords(Iris(:,1:4),'group',Y,'labels',names);



Parallel coordinate plot is used to plot the attribute values of high dimensional data. Instead of using perpendicular axes, we are using parallel axes.

In above plot is over-cluttered for sepal length and sepal width, which is illegible since they're very data-dense. While is it very sorted and less dense for petal length and petal width.

3. Practice Data Distance Measures:

1) Make a function for Minkowski Distance. (3 function inputs: vector A, vector B, and order r)

**Code:**

function [MD] = minkowskiDistance(A,B,r)

MD=(sum((abs(A-B)).^r).^(1/r));

end

2) Make a function for T-statistics Distance. (3 function inputs: time series A, time series B)

**Code:**

function [tstatd] = tStasticsDistance(A,B)

tstatd = ((abs(mean(A)-mean(B)))/std(A-B));

end

3) Make a function for Mahalanobis Distance. (3 function inputs: vector A, vector B, and covariance matrix M)

**Code:**

function [ MD ] = mahalanobis(A,B,M)

P = inv(M);

MD = (A-B)\*P\*(transpose(A-B));

end

4. Assume a new iris sample S has a feature vector of [5.0000, 3.5000, 1.4600, 0.2540]. Calculate the distances between the new sample and the 150 samples in the iris dataset using the distance functions you made.

1) Calculate Minkowski distances with r = 1, 2, 100, respectively, and plot the obtained distances.

a. Minkowski distances with r = 1

**Code:**

%Minkowski Distance with r=1

load Iris.csv;

A = Iris (:,1:4);

B = [5.0000,3.5000,1.4600,0.2540];

r=1;

MD=minkowskiDist(A,B,r);

plot(MD);

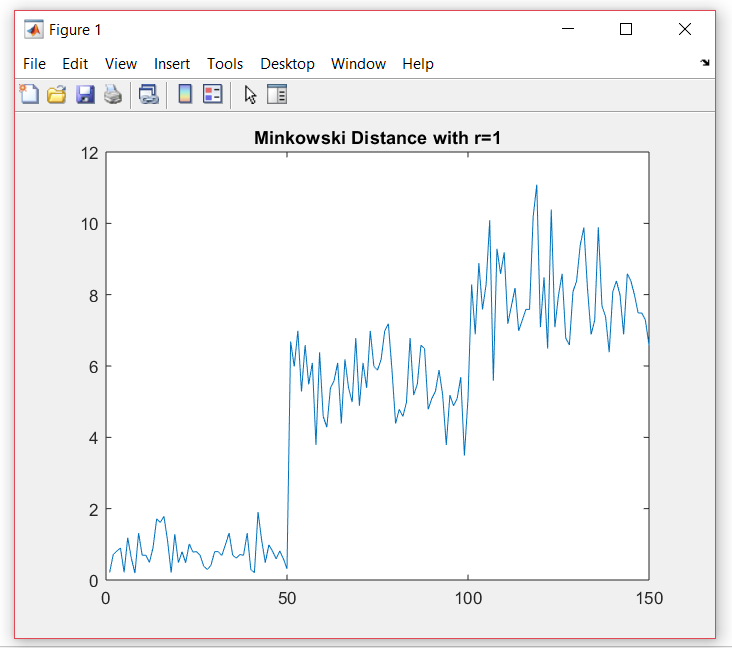
function [MD] = minkowskiDist(A,B,r)

for i=1:150

MD(i)=(sum((abs(A(i,:)-B)).^r).^(1/r));

end

end



b.Minkowski distances with r = 2

**Code:**

%Minkowski Distance with r=2

load Iris.csv;

A = Iris (:,1:4);

B = [5.0000,3.5000,1.4600,0.2540];

r=2;

MD=minkowskiDist(A,B,r);

plot(MD);

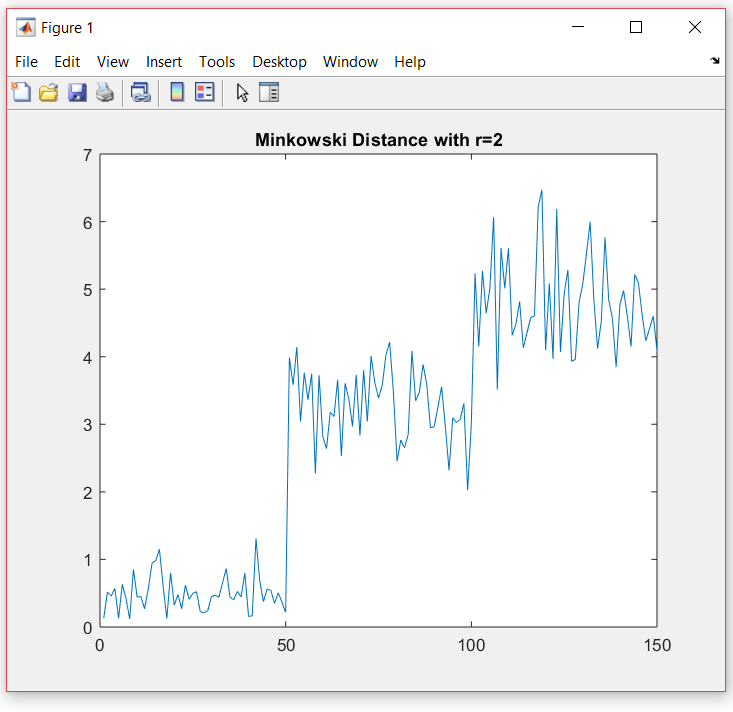
function [MD] = minkowskiDist(A,B,r)

for i=1:150

MD(i)=(sum((abs(A(i,:)-B)).^r).^(1/r));

end

end



c. Minkowski distances with r = 100

**Code:**

%Minkowski Distance with r=100

load Iris.csv;

A = Iris (:,1:4);

B = [5.0000,3.5000,1.4600,0.2540];

r=100;

MD=minkowskiDist(A,B,r);

plot(MD);

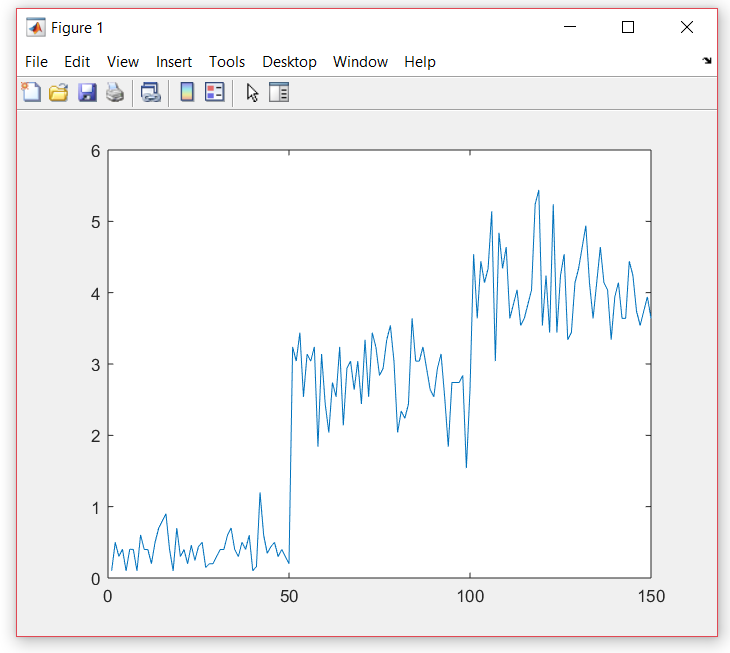
function [MD] = minkowskiDist(A,B,r)

for i=1:150

MD(i)=(sum((abs(A(i,:)-B)).^r).^(1/r));

end

end



2) Calculate Mahalanobis distances and plot the obtained distances.

**Code:**

%Mahalanobis distances

load Iris.csv;

A = Iris (:,1:4);

B = [5.0000,3.5000,1.4600,0.2540];

M = cov(A);

MD=mahalanobisDist(A,B,M);

plot(MD);

function [ MD ] = mahalanobisDist(A,B,M)

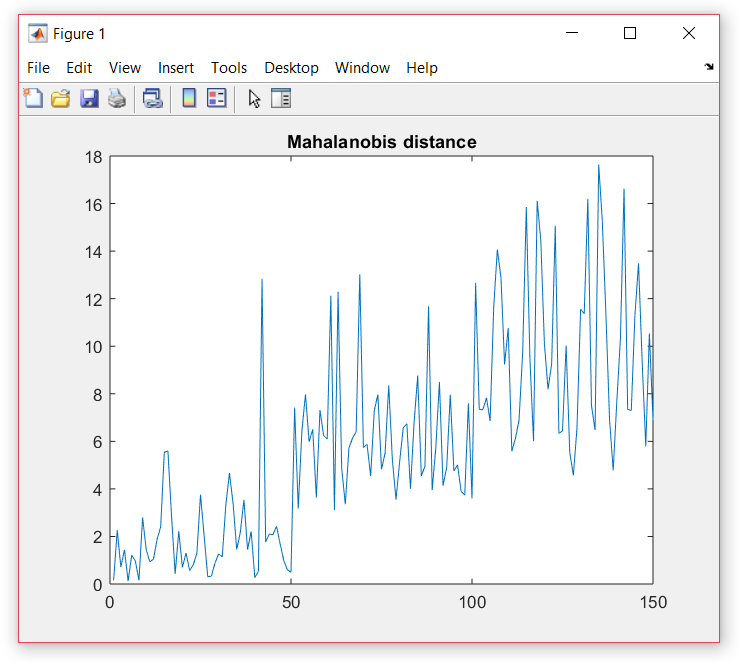
j= inv(M);

for i= 1:150

MD(i)= (A(i,:)-B)\*j\*(transpose(A(i,:)-B));

end

end



5. We provide a dataset with two time series in HW1\_DataMining.txt file. Perform the following analysis:

1) Plot the generated two time series in one plot.

Generate two time series data by the code: **X = mvnrnd([0;0],[1 .3;.3 1],100);**

Else Use HW1\_DataMining.txt file,import it and plot it.

**Code:**

%Plot the generated two time series in one plot

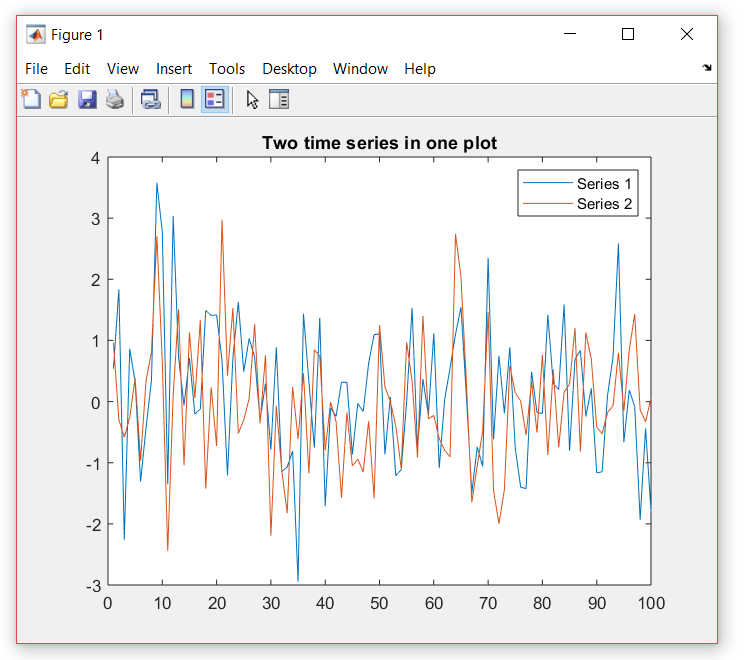
filename = 'HW1\_DataMining.txt ';

A= importdata(filename);

plot(A);

title('Two time series in one plot');

legend('Series 1','Series 2');



2) Calculate the T-statistics distance between the two time series.

**Code:**

% Converted the given text file to csv file

load HW1\_DataMining.csv;

Series1 = HW1\_DataMining(:,1);

Series2 = HW1\_DataMining (:,2);

tsd=tStasticsDist(Series1,Series2);

disp(tsd);

function [tstatd] = tStasticsDist(A,B)

for i=1:100

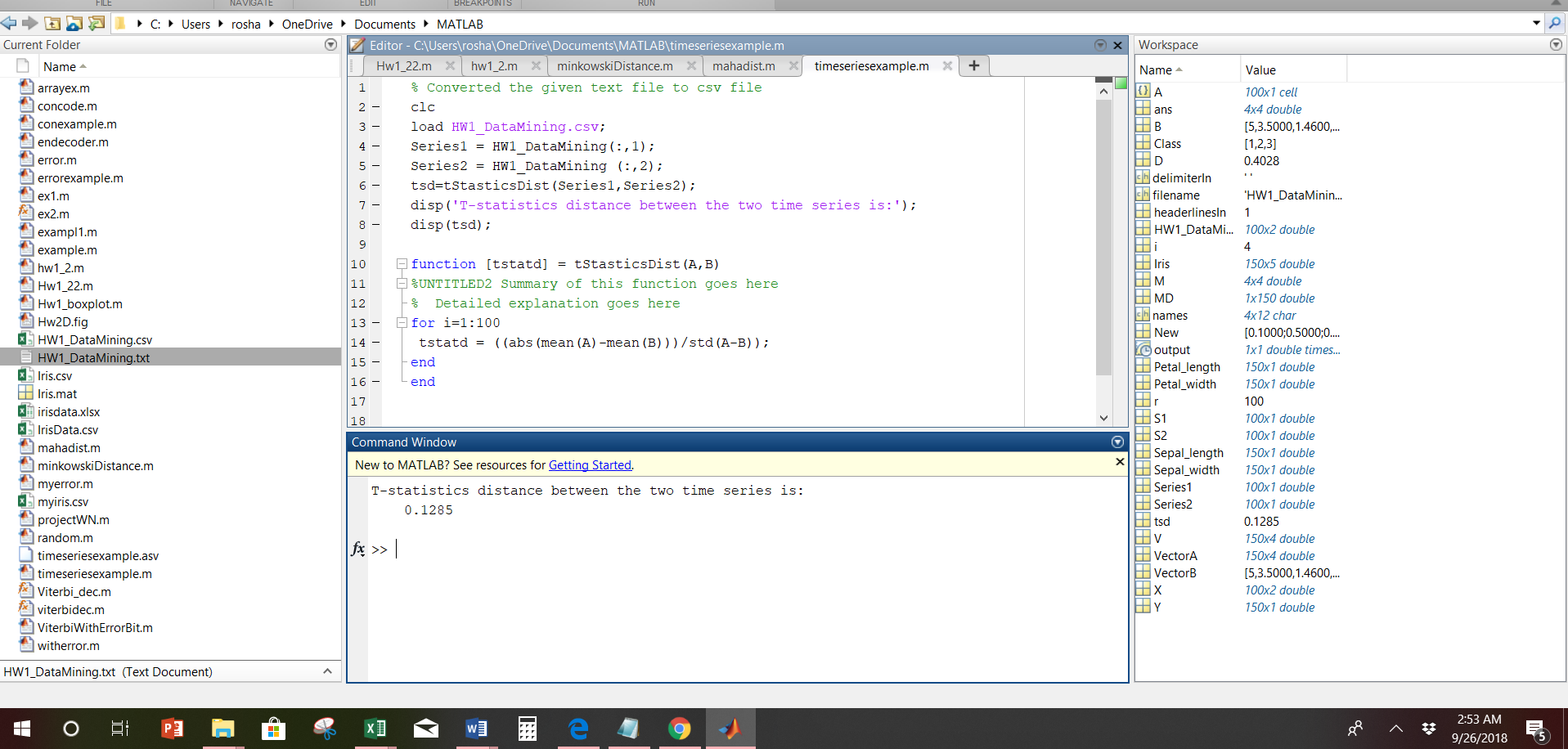
tstatd = ((abs(mean(A)-mean(B)))/std(A-B));

end

end

**T-statistics distance between the two time series is:**

**0.1285**



3)Calculate the correlation of the two time series.

**Code:**

% Converted the given text file to csv file

load HW1\_DataMining.csv;

Series1 = HW1\_DataMining(:,1);

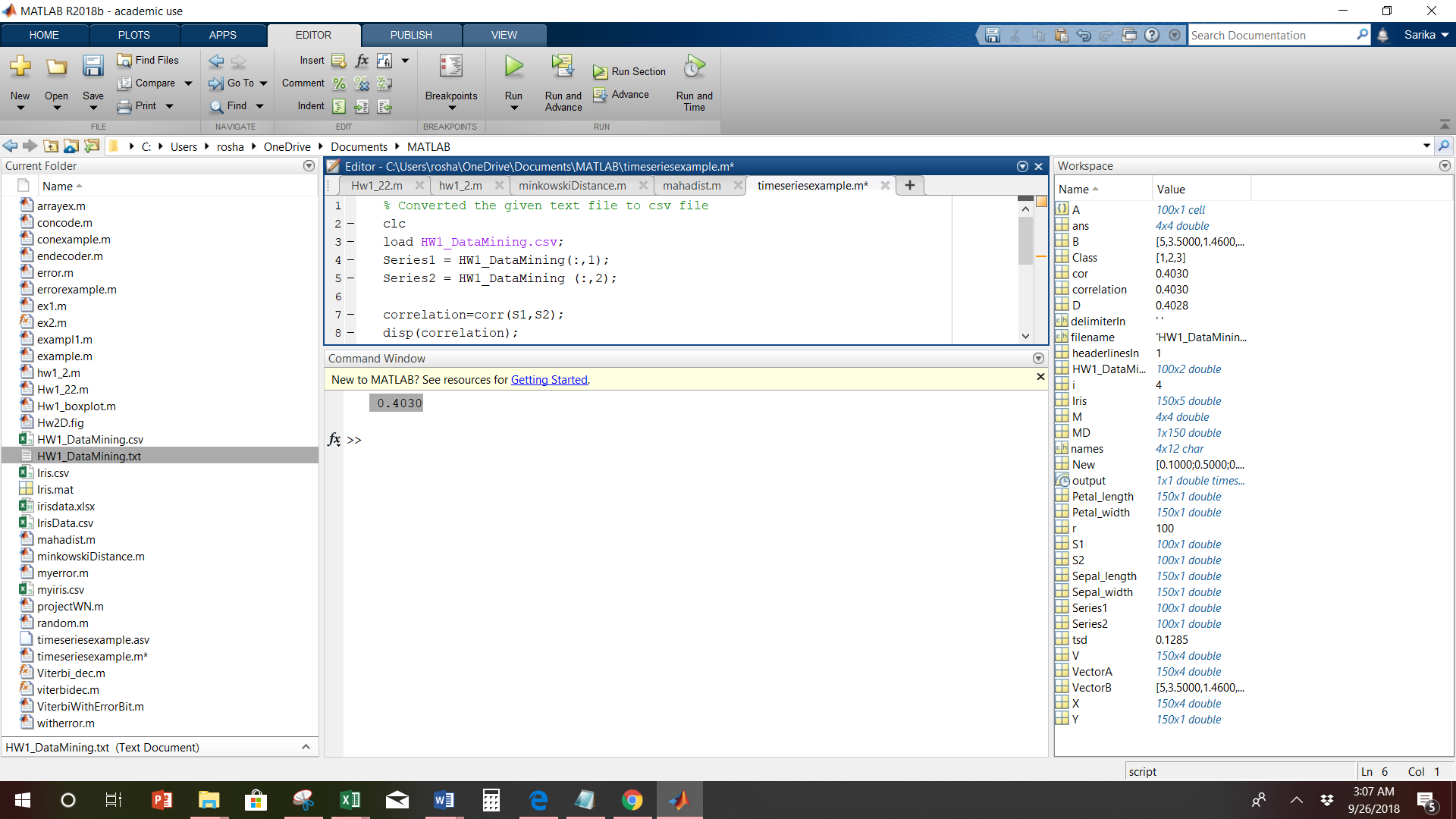
Series2 = HW1\_DataMining (:,2);

correlation=corr(S1,S2);

disp(correlation);

**Correlation of the two time series is:**

**0.4030**



4) Normalize the feature matrix of the IRIS dataset such that after normalization each feature has a mean of 0 and a standard deviation of 1.

**Code:**

%Normalize the feature matrix of the IRIS dataset

load Iris.csv;

X = Iris (:,1:4);

%Z and N gives same results

%Z=zscore(X);

%disp(Z);

%or

N= normalize(X);

disp(N);

plot(N);

**Output after Normalization:**

-0.8977 1.0286 -1.3368 -1.3086

-1.1392 -0.1245 -1.3368 -1.3086

-1.3807 0.3367 -1.3935 -1.3086

-1.5015 0.1061 -1.2801 -1.3086

-1.0184 1.2592 -1.3368 -1.3086

-0.5354 1.9511 -1.1668 -1.0465

-1.5015 0.7980 -1.3368 -1.1776

-1.0184 0.7980 -1.2801 -1.3086

-1.7430 -0.3552 -1.3368 -1.3086

-1.1392 0.1061 -1.2801 -1.4396

-0.5354 1.4899 -1.2801 -1.3086

-1.2600 0.7980 -1.2234 -1.3086

-1.2600 -0.1245 -1.3368 -1.4396

-1.8638 -0.1245 -1.5068 -1.4396

-0.0523 2.1818 -1.4501 -1.3086

-0.1731 3.1043 -1.2801 -1.0465

-0.5354 1.9511 -1.3935 -1.0465

-0.8977 1.0286 -1.3368 -1.1776

-0.1731 1.7205 -1.1668 -1.1776

-0.8977 1.7205 -1.2801 -1.1776

-0.5354 0.7980 -1.1668 -1.3086

-0.8977 1.4899 -1.2801 -1.0465

-1.5015 1.2592 -1.5635 -1.3086

-0.8977 0.5674 -1.1668 -0.9155

-1.2600 0.7980 -1.0534 -1.3086

-1.0184 -0.1245 -1.2234 -1.3086

-1.0184 0.7980 -1.2234 -1.0465

-0.7769 1.0286 -1.2801 -1.3086

-0.7769 0.7980 -1.3368 -1.3086

-1.3807 0.3367 -1.2234 -1.3086

-1.2600 0.1061 -1.2234 -1.3086

-0.5354 0.7980 -1.2801 -1.0465

-0.7769 2.4124 -1.2801 -1.4396

-0.4146 2.6430 -1.3368 -1.3086

-1.1392 0.1061 -1.2801 -1.4396

-1.0184 0.3367 -1.4501 -1.3086

-0.4146 1.0286 -1.3935 -1.3086

-1.1392 0.1061 -1.2801 -1.4396

-1.7430 -0.1245 -1.3935 -1.3086

-0.8977 0.7980 -1.2801 -1.3086

-1.0184 1.0286 -1.3935 -1.1776

-1.6223 -1.7390 -1.3935 -1.1776

-1.7430 0.3367 -1.3935 -1.3086

-1.0184 1.0286 -1.2234 -0.7845

-0.8977 1.7205 -1.0534 -1.0465

-1.2600 -0.1245 -1.3368 -1.1776

-0.8977 1.7205 -1.2234 -1.3086

-1.5015 0.3367 -1.3368 -1.3086

-0.6561 1.4899 -1.2801 -1.3086

-1.0184 0.5674 -1.3368 -1.3086

1.3968 0.3367 0.5335 0.2638

0.6722 0.3367 0.4202 0.3948

1.2761 0.1061 0.6469 0.3948

-0.4146 -1.7390 0.1368 0.1328

0.7930 -0.5858 0.4768 0.3948

-0.1731 -0.5858 0.4202 0.1328

0.5515 0.5674 0.5335 0.5259

-1.1392 -1.5083 -0.2600 -0.2603

0.9138 -0.3552 0.4768 0.1328

-0.7769 -0.8164 0.0801 0.2638

-1.0184 -2.4308 -0.1466 -0.2603

0.0684 -0.1245 0.2501 0.3948

0.1892 -1.9696 0.1368 -0.2603

0.3100 -0.3552 0.5335 0.2638

-0.2939 -0.3552 -0.0899 0.1328

1.0345 0.1061 0.3635 0.2638

-0.2939 -0.1245 0.4202 0.3948

-0.0523 -0.8164 0.1935 -0.2603

0.4307 -1.9696 0.4202 0.3948

-0.2939 -1.2777 0.0801 -0.1293

0.0684 0.3367 0.5902 0.7880

0.3100 -0.5858 0.1368 0.1328

0.5515 -1.2777 0.6469 0.3948

0.3100 -0.5858 0.5335 0.0017

0.6722 -0.3552 0.3068 0.1328

0.9138 -0.1245 0.3635 0.2638

1.1553 -0.5858 0.5902 0.2638

1.0345 -0.1245 0.7035 0.6569

0.1892 -0.3552 0.4202 0.3948

-0.1731 -1.0471 -0.1466 -0.2603

-0.4146 -1.5083 0.0234 -0.1293

-0.4146 -1.5083 -0.0332 -0.2603

-0.0523 -0.8164 0.0801 0.0017

0.1892 -0.8164 0.7602 0.5259

-0.5354 -0.1245 0.4202 0.3948

0.1892 0.7980 0.4202 0.5259

1.0345 0.1061 0.5335 0.3948

0.5515 -1.7390 0.3635 0.1328

-0.2939 -0.1245 0.1935 0.1328

-0.4146 -1.2777 0.1368 0.1328

-0.4146 -1.0471 0.3635 0.0017

0.3100 -0.1245 0.4768 0.2638

-0.0523 -1.0471 0.1368 0.0017

-1.0184 -1.7390 -0.2600 -0.2603

-0.2939 -0.8164 0.2501 0.1328

-0.1731 -0.1245 0.2501 0.0017

-0.1731 -0.3552 0.2501 0.1328

0.4307 -0.3552 0.3068 0.1328

-0.8977 -1.2777 -0.4300 -0.1293

-0.1731 -0.5858 0.1935 0.1328

0.5515 0.5674 1.2703 1.7052

-0.0523 -0.8164 0.7602 0.9190

1.5176 -0.1245 1.2136 1.1811

0.5515 -0.3552 1.0436 0.7880

0.7930 -0.1245 1.1569 1.3121

2.1214 -0.1245 1.6103 1.1811

-1.1392 -1.2777 0.4202 0.6569

1.7591 -0.3552 1.4403 0.7880

1.0345 -1.2777 1.1569 0.7880

1.6384 1.2592 1.3270 1.7052

0.7930 0.3367 0.7602 1.0500

0.6722 -0.8164 0.8736 0.9190

1.1553 -0.1245 0.9869 1.1811

-0.1731 -1.2777 0.7035 1.0500

-0.0523 -0.5858 0.7602 1.5742

0.6722 0.3367 0.8736 1.4431

0.7930 -0.1245 0.9869 0.7880

2.2422 1.7205 1.6670 1.3121

2.2422 -1.0471 1.7804 1.4431

0.1892 -1.9696 0.7035 0.3948

1.2761 0.3367 1.1003 1.4431

-0.2939 -0.5858 0.6469 1.0500

2.2422 -0.5858 1.6670 1.0500

0.5515 -0.8164 0.6469 0.7880

1.0345 0.5674 1.1003 1.1811

1.6384 0.3367 1.2703 0.7880

0.4307 -0.5858 0.5902 0.7880

0.3100 -0.1245 0.6469 0.7880

0.6722 -0.5858 1.0436 1.1811

1.6384 -0.1245 1.1569 0.5259

1.8799 -0.5858 1.3270 0.9190

2.4837 1.7205 1.4970 1.0500

0.6722 -0.5858 1.0436 1.3121

0.5515 -0.5858 0.7602 0.3948

0.3100 -1.0471 1.0436 0.2638

2.2422 -0.1245 1.3270 1.4431

0.5515 0.7980 1.0436 1.5742

0.6722 0.1061 0.9869 0.7880

0.1892 -0.1245 0.5902 0.7880

1.2761 0.1061 0.9302 1.1811

1.0345 0.1061 1.0436 1.5742

1.2761 0.1061 0.7602 1.4431

-0.0523 -0.8164 0.7602 0.9190

1.1553 0.3367 1.2136 1.4431

1.0345 0.5674 1.1003 1.7052

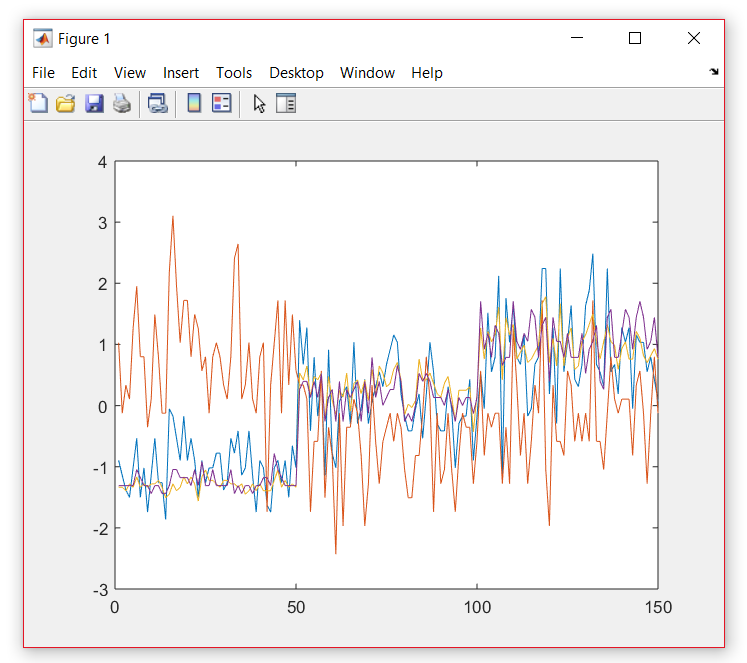
1.0345 -0.1245 0.8169 1.4431

0.5515 -1.2777 0.7035 0.9190

0.7930 -0.1245 0.8169 1.0500

0.4307 0.7980 0.9302 1.4431

0.0684 -0.1245 0.7602 0.7880



The above plot shows normalization of the feature matrix of the IRIS dataset such that each feature has a mean of 0 and a standard deviation of 1.

References:

<https://in.mathworks.com/help/>

<https://stackoverflow.com/questions/20972221/minkowski-distance-and-pdist>

Lecture2\_Explore Data Slides