DETERMINATION OF PROPER GROUPING APPROACHES FOR LARGE DATA SETS BY APPLYING CLUSTER ANALYSIS METHODS IN MATLAB

for the Bachelor of Science (General) Degree

By
O.G.V.V Bandara
SC/2019/10723
Supervisor:
Prof. L.A.L.W. Jayasekara

Department of Mathematics
University of Ruhuna
Matara.

1)INTRODUCTION

What is cluster analysis?

- Cluster analysis intends to group data sets
- The grouping is done such that similarities between clusters in high and dissimilarities is less
- Within the certain cluster the similarities are greater between data
- Most of the time the criteria used for grouping is the distance

Some of the common clustering methods

- Connectivity models
- Centroid Models
- Distribution models
- Density models
- Subspace models
- Group models
- Graph based models
- Signed graph models
- Neural models

2)PROBLEM STATEMENT

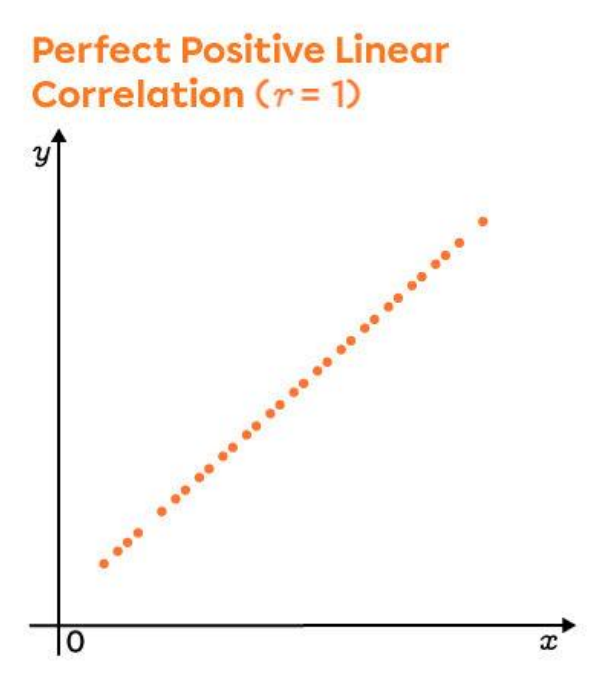
1)grouping large data sets which have no visible

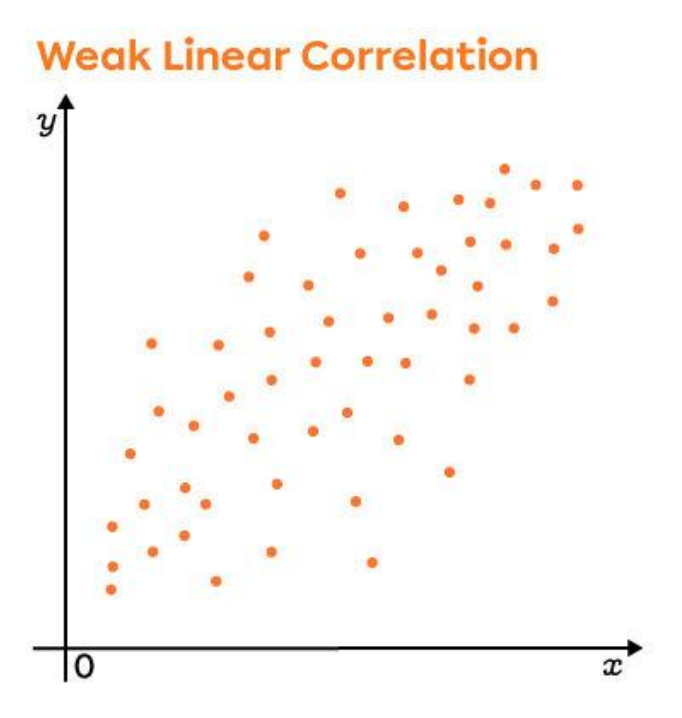
Relationship

When it comes to raw data, there may be correlations, but some data sets may have no visible relationship

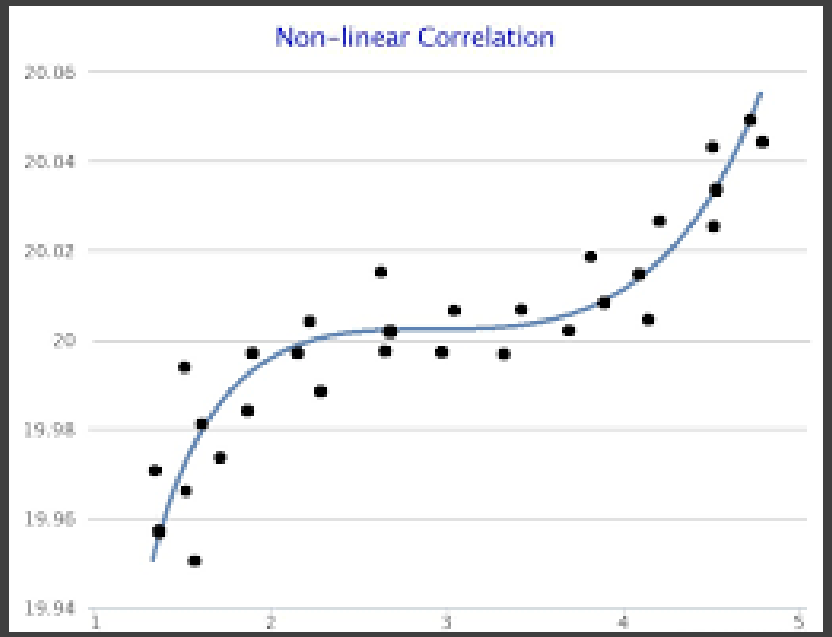
For such data to be usefull they should be converted to informations

- Data with linear correlations

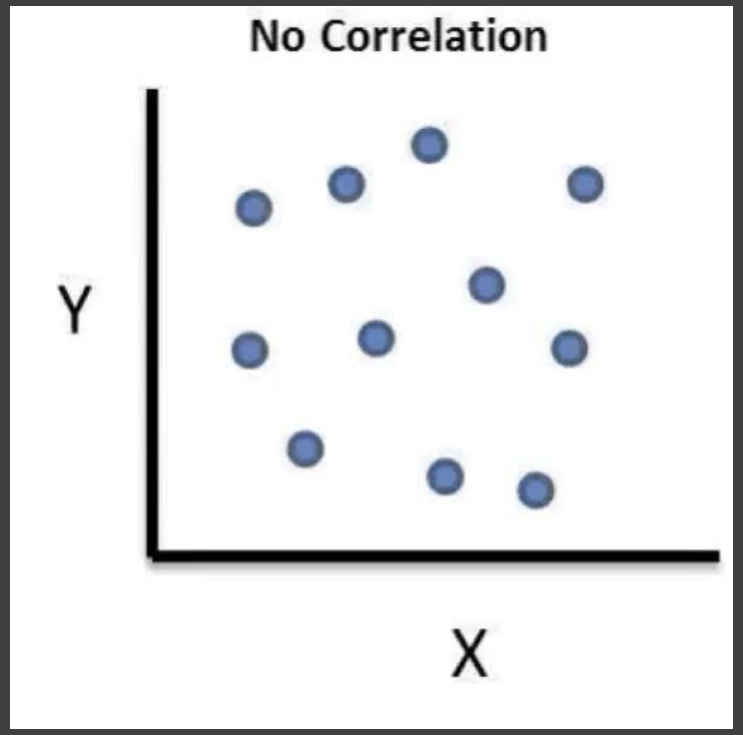




- Data with non linear correlations



- Data with no correlation



2)The tedious process of repetitive calculations

- Cluster analysis methods involve large and tedious calculations which consumes time to be performed repetitively.
- The human error involved makes the efficiency of the calculation process reduce.
- Obtaining a dendrogram without the use of computer is impractical for large data or p dimensional matrices

3)OBJECTIVES

1) The objective of cluster analysis is to identify similarities between observation and create a criteria in order to group them such that the data is useful to the Researcher.

2) A code is developed through MATLAB such that the grouping process can be done easily by the user.

3) The code enables the user to choose their preferred grouping method and obtain a visual representation of the grouped data.

4) By using a code the user is able to input data one time without the tedious process of multiple calculations.

4)Background of Study

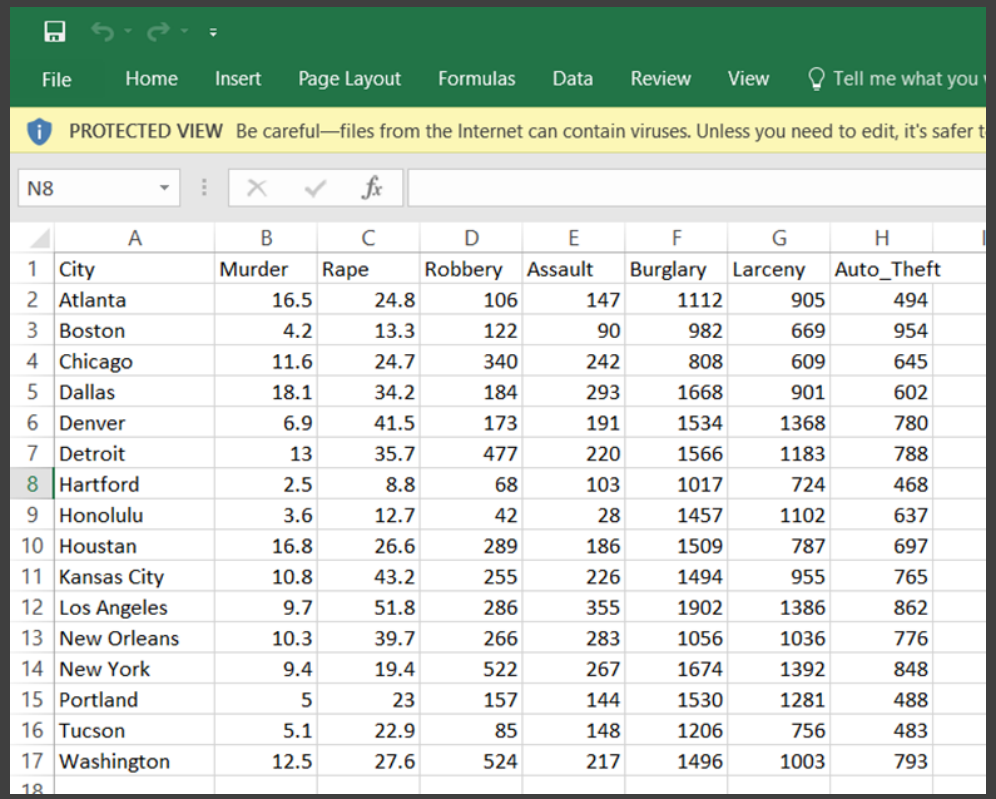
- The type of clustering depends on the input data
- Two common clustering approaches can be identified
  - Hierarchical clustering
  - Partitioning

Hierarchical clustering	Partitioning
typically starts with 'n' clusters such that it comprises one for each observation. And ends with a single cluster containing all 'n' observations.	The observations are divided into 'g' clusters.This can be done by starting with an initial partitioning or with cluster centers and then reallocating the observations according to some optimality criterion.

5)METHEDOLOGY

- A code was developed using MATLAB
- The user can use the code to input the data in the form of a spreadsheet
- The user is able to then select the preferred clustering method
- The code outputs a dendrogram which is a visual representation of the grouping

The sample input data obtained from a reference



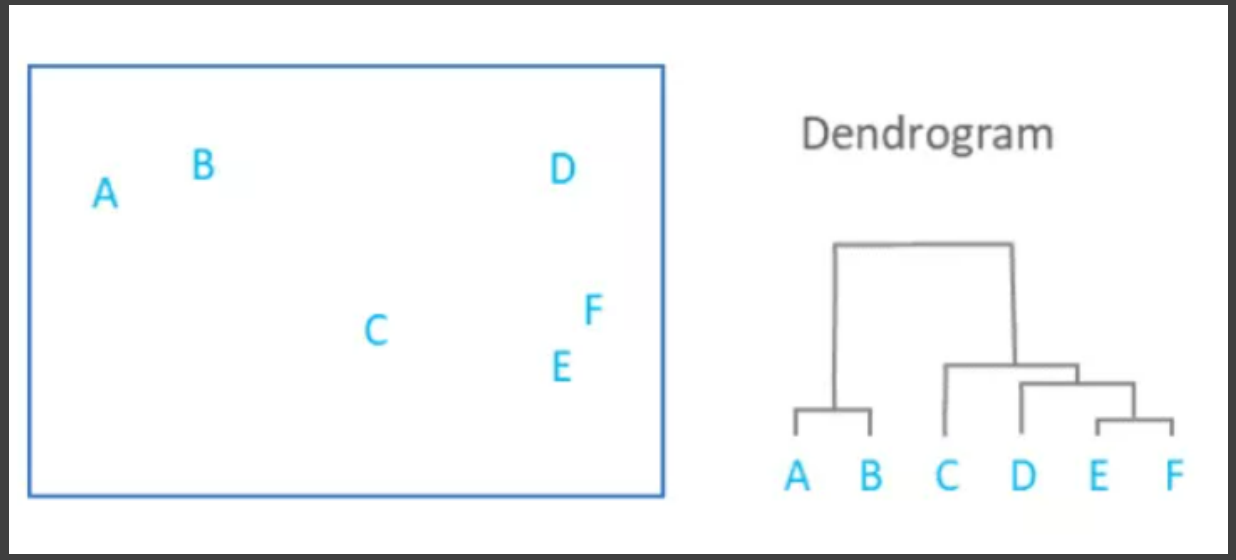
	A	B	C	D	E	F	G	H
1	City	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
2	Atlanta	16.5	24.8	106	147	1112	905	494
3	Boston	4.2	13.3	122	90	982	669	954
4	Chicago	11.6	24.7	340	242	808	609	645
5	Dallas	18.1	34.2	184	293	1668	901	602
6	Denver	6.9	41.5	173	191	1534	1368	780
7	Detroit	13	35.7	477	220	1566	1183	788
8	Hartford	2.5	8.8	68	103	1017	724	468
9	Honolulu	3.6	12.7	42	28	1457	1102	637
10	Houstan	16.8	26.6	289	186	1509	787	697
11	Kansas City	10.8	43.2	255	226	1494	955	765
12	Los Angeles	9.7	51.8	286	355	1902	1386	862
13	New Orleans	10.3	39.7	266	283	1056	1036	776
14	New York	9.4	19.4	522	267	1674	1392	848
15	Portland	5	23	157	144	1530	1281	488
16	Tucson	5.1	22.9	85	148	1206	756	483
17	Washington	12.5	27.6	524	217	1496	1003	793

What is a dendrogram?

- A dendrogram is a diagram that shows the hierarchical relationship between objects.

- The main use of a dendrogram is to work out the best way to allocate objects to clusters.

- The data can be represented in the form of a scatter plot, but identifying grouping in a scatter plot is inefficient.
- The dendrogram represents the data in manner such that the grouping can be visually identified



Euclidean distance

- A common measure of distance is the Euclidean Distance between two vectors. The two vectors are p dimensional vectors.
  - $(x_1, x_2, \dots\dots, x_p)^T, (y_1, y_2, \dots\dots, y_p)^T$

- $d(x, y) = \sqrt{(x - y)^T (x - y)}$

The methods of clustering

The clustering methods used in this project were Hierarchical clustering methods.

1)Single linkage

- In the single linkage method, the distance between two clusters A and B is defined as the minimum distance between a point in A and a point in B.

- $D(A,B) = min\{d(y_i, y_j)\}$

2)Complete linkage

- the distance between two clusters A and B is defined as the maximum distance between a point in A and a point in B.

- $D(A, B) = max\{d(y_i, y_j)\}$

3)Centroid method

- In the centroid method, the distance between two clusters A and B is defined as the Euclidean distance between the mean vectors (often called centroids) of the two clusters

$$D(A, B) = d(\bar{y}_A, \bar{y}_B)$$

Mean distance

$$\bar{y}_A = \sum_{i=1}^{n_A} \frac{y_i}{n_A}$$

4)Median method

- If two clusters A and B are combined using the centroid method, and if A contains a larger number of items than B, then the new centroid.

- $\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}$

- To avoid weighting the mean vectors according to cluster size, we can use the median(midpoint) of the line joining A and B as the point for computing new distances to other clusters.

- $m_{AB} = \frac{1}{2}(y_A + y_B)$

5)Average method

- In average linkage approach, The distance between two clusters A and B is defined as the average of n_A, n_B distances between n_A points in A and n_B points in B.

- $D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j)$

6)Wards method

- Ward's method, also called the incremental sum of squares method, uses the within cluster (squared) distances and the between-cluster (squared) distances.

- If AB is the cluster obtained by combining clusters A and B, then the sum of within-cluster distances are

- The equations for each quantity
- $SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A)$

- $SSE_B = \sum_{i=1}^{n_A} (y_i - \bar{y}_B)'(y_i - \bar{y}_B)$

- $SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB})$

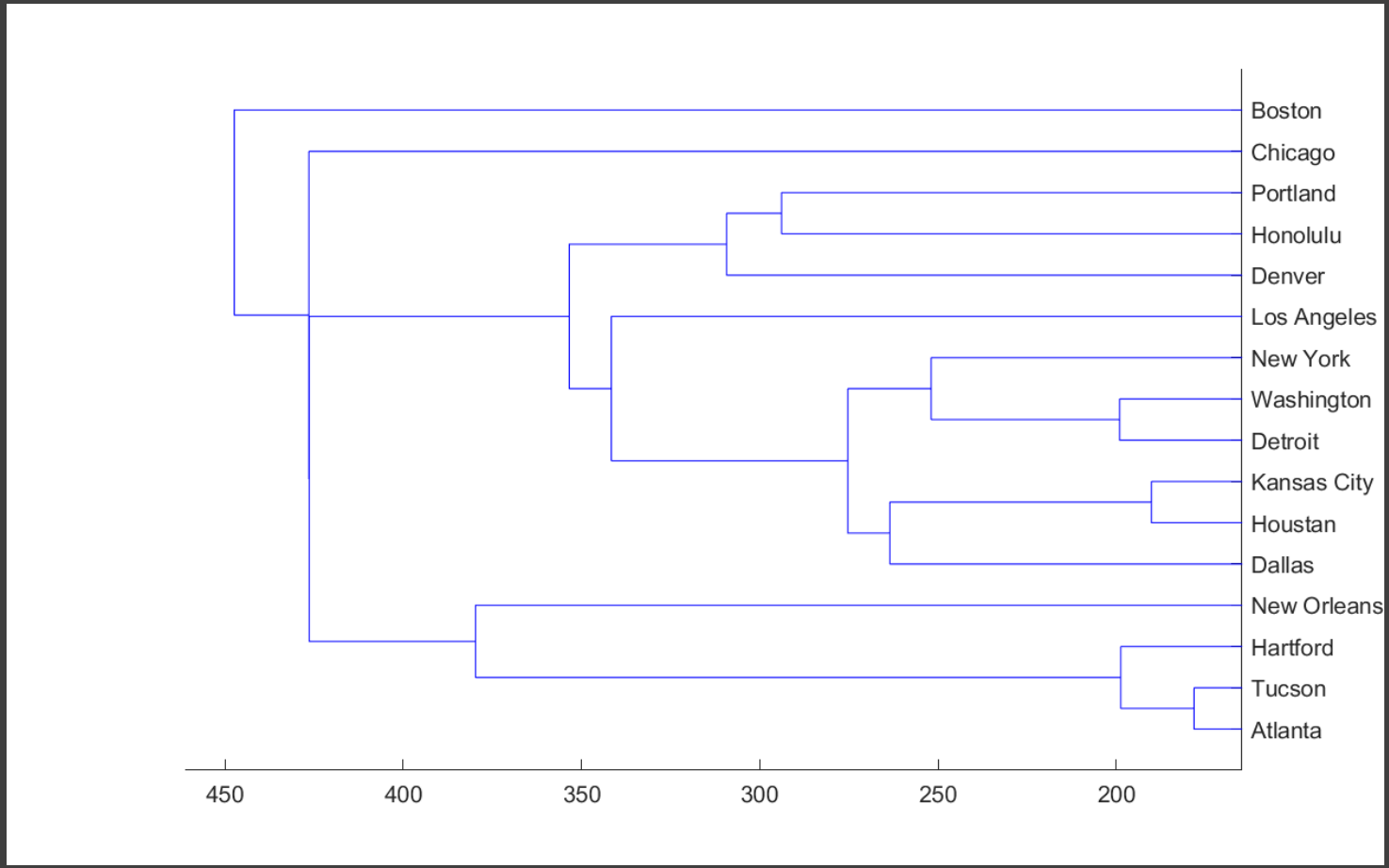
- Where
- $\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}$

- The increase in SSE is defined as

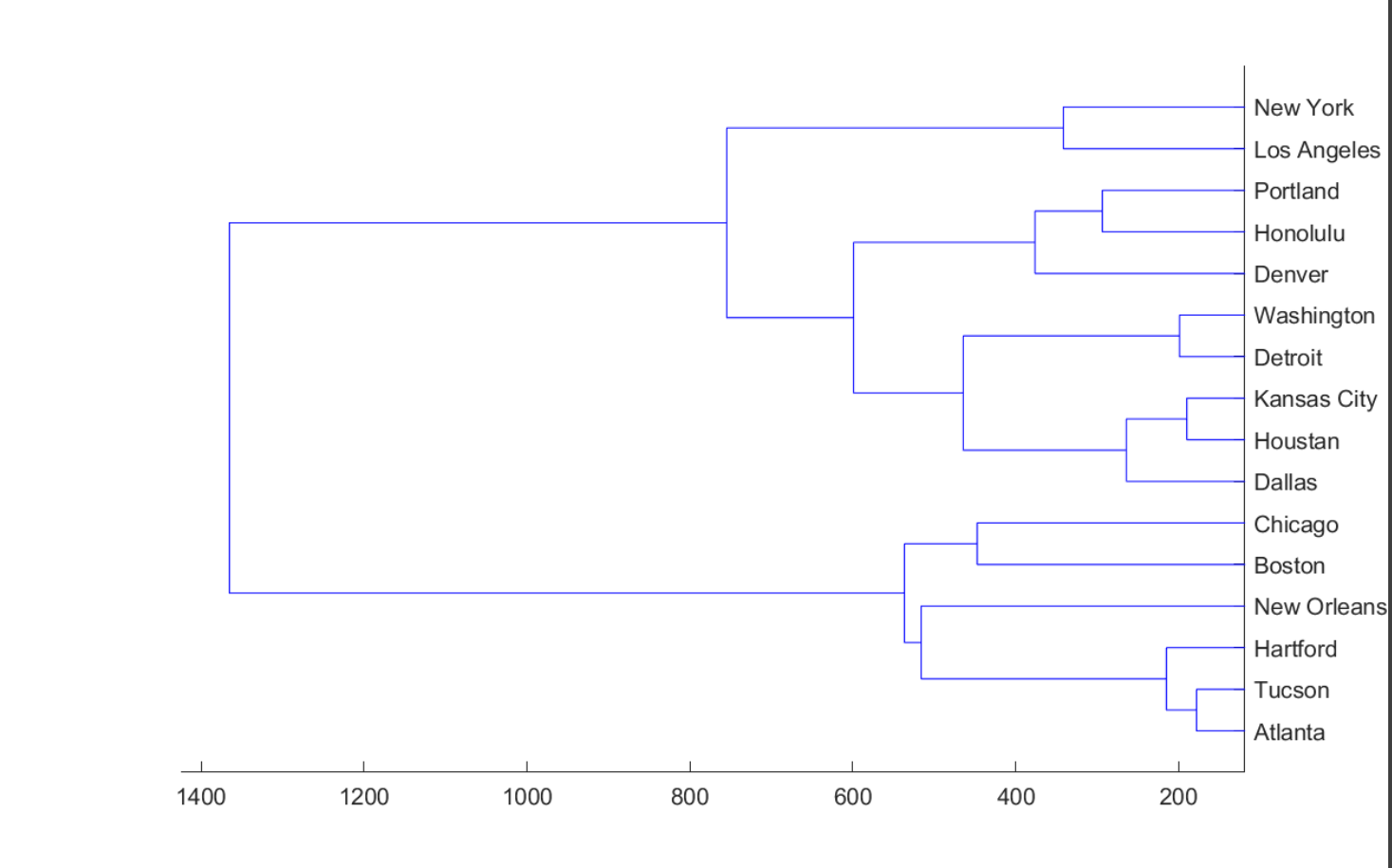
- $I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$

6)DISCUSSION

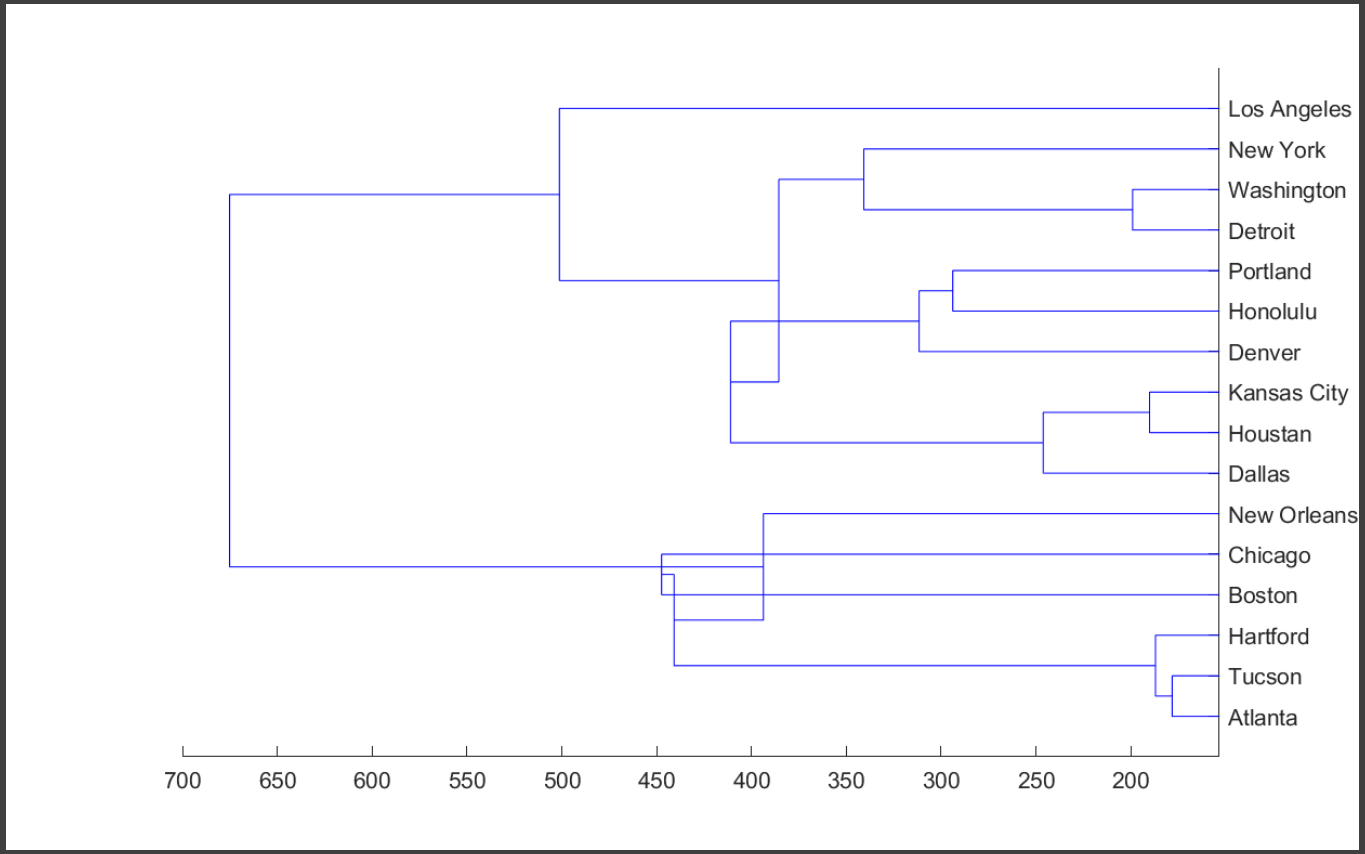
1)Dendrogram obtained by MATLAB through single linkage clustering



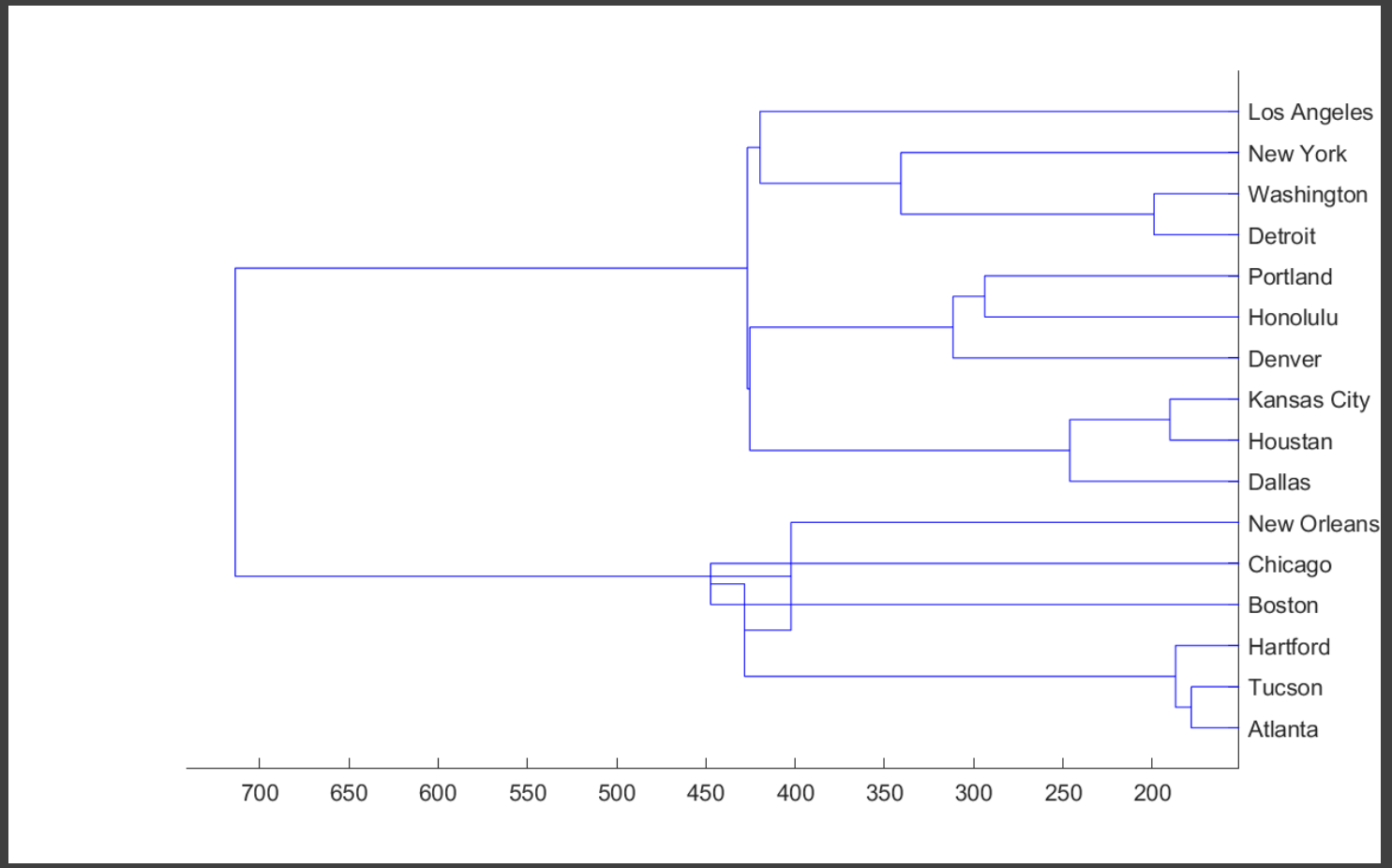
2)Dendrogram obtained by matlab through complete linkage clustering



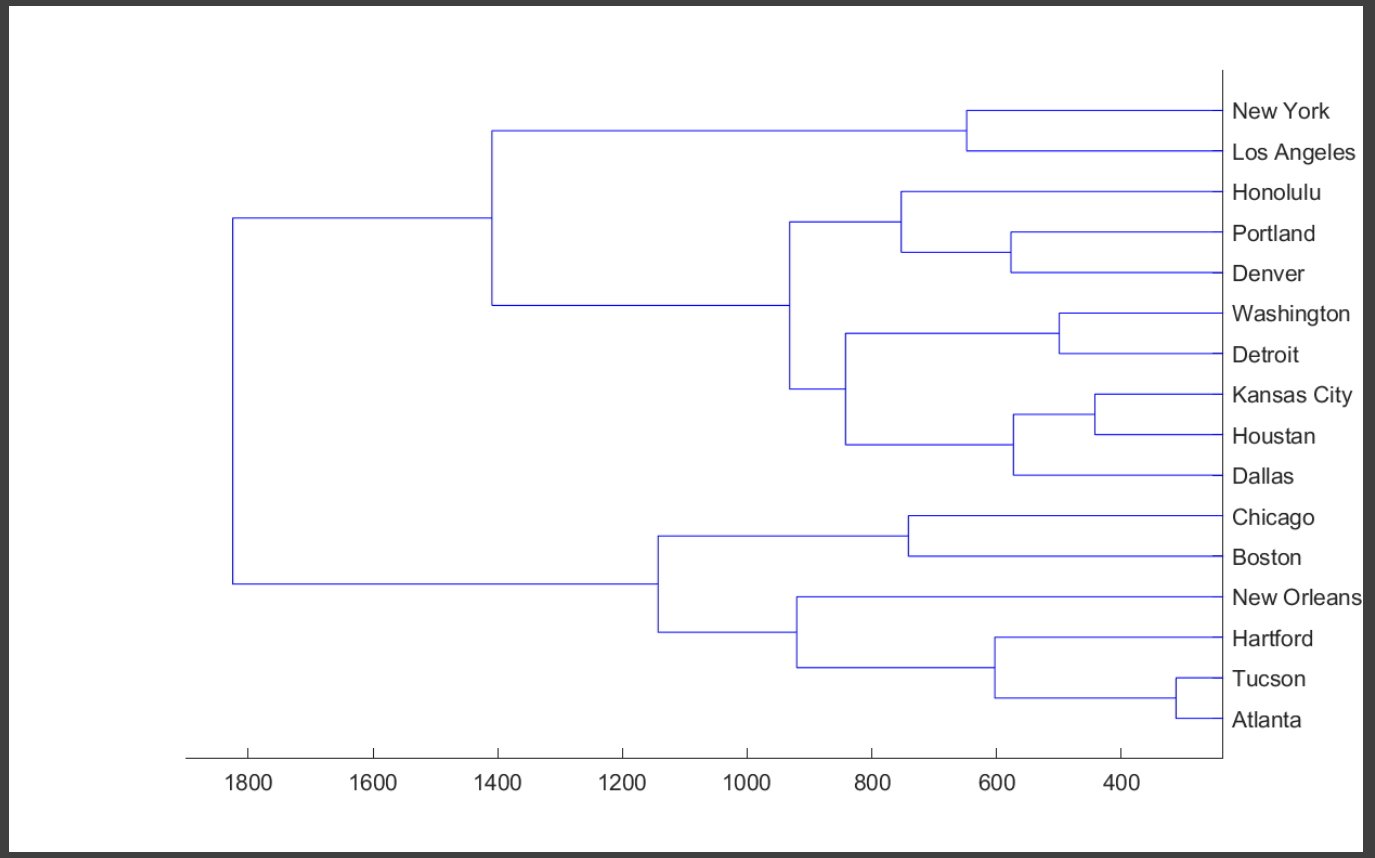
3)Dendrogram obtained in MATLAB through centroid method clustering



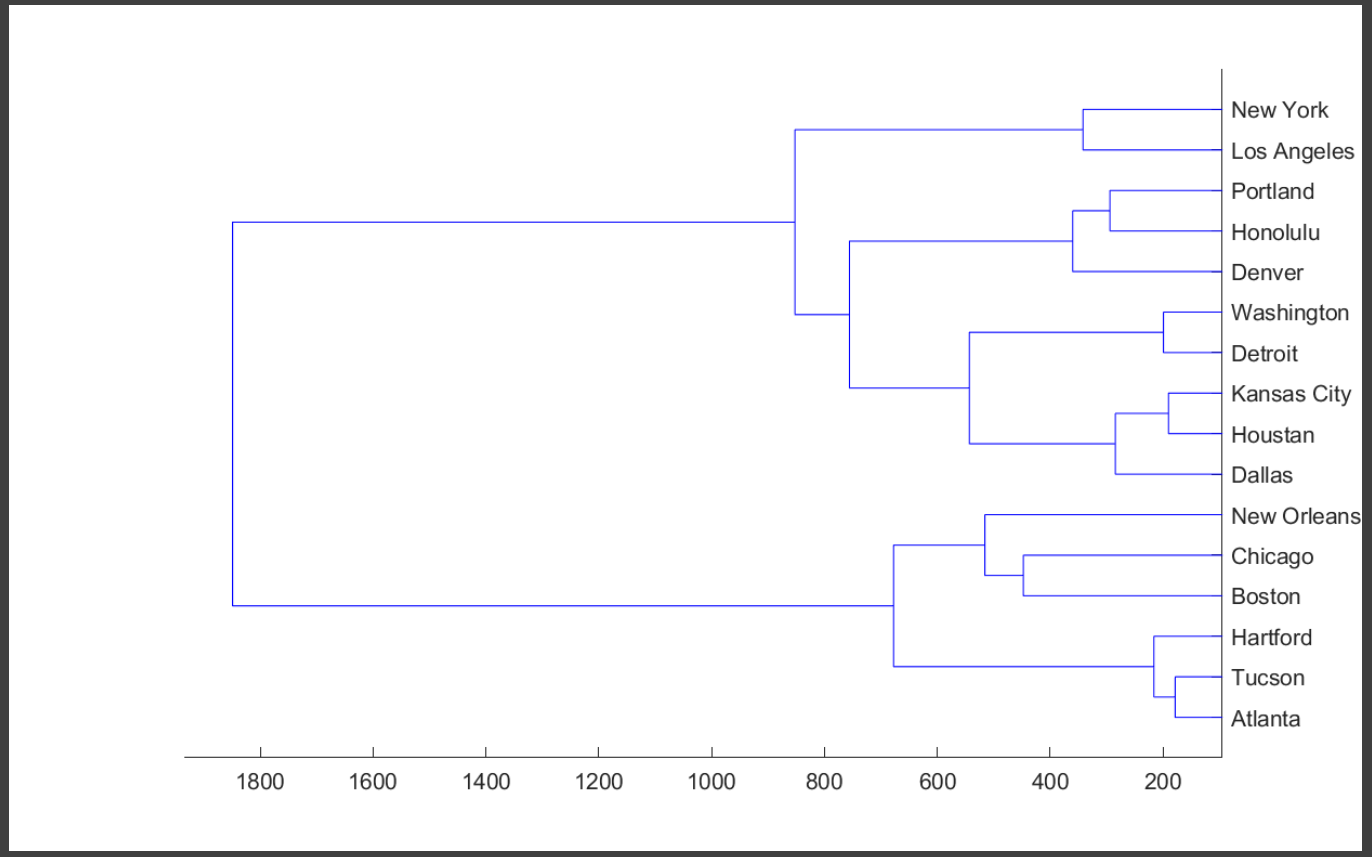
4)Dendrogram obtained through median method clustering



5)Dendrogram obtained through average linkage clustering



6)Dendrogram obtained in MATLAB through wards method clustering



7)CONCLUSION

- The code developed has the ability to obtain data from a user and allow the user to select a suitable clustering method
- The code outputs a dendrogram which is a visual representation of the grouping
- This code is useful in converting raw data to useful information through grouping

THANK YOU

Q & A