I. Objective:

This project aims to divide the given dataset into clusters based on the similarities between the observations. In this project we measure the dissimilarities between the observations by four methods; Euclidean, Canberra, Manhattan, and Maximum. We are also going to use four methods to measure distances between clusters; Single, Ward.d2, Complete, and Average. We will use two methods for obtaining the clusters. First method is Hierarchal clustering, we use the hierarchal algorithm which produces a cluster Dendrogram based on which we can decide on the number of clusters we are going to divide the data into. The Second method is K-means method which requires the number of clusters k to be known from the beginning. To determine the number of clusters in the data, we compute the within sum of squares for various values of k = 2, …, n – 1 (at most), then plot WSS(k) versus k. Finally, for the goodness of fit we compute R^2 which is equal to the trace of the between sum of squares divided by the trace of the total sum of squares.

II. Source and description of variables:

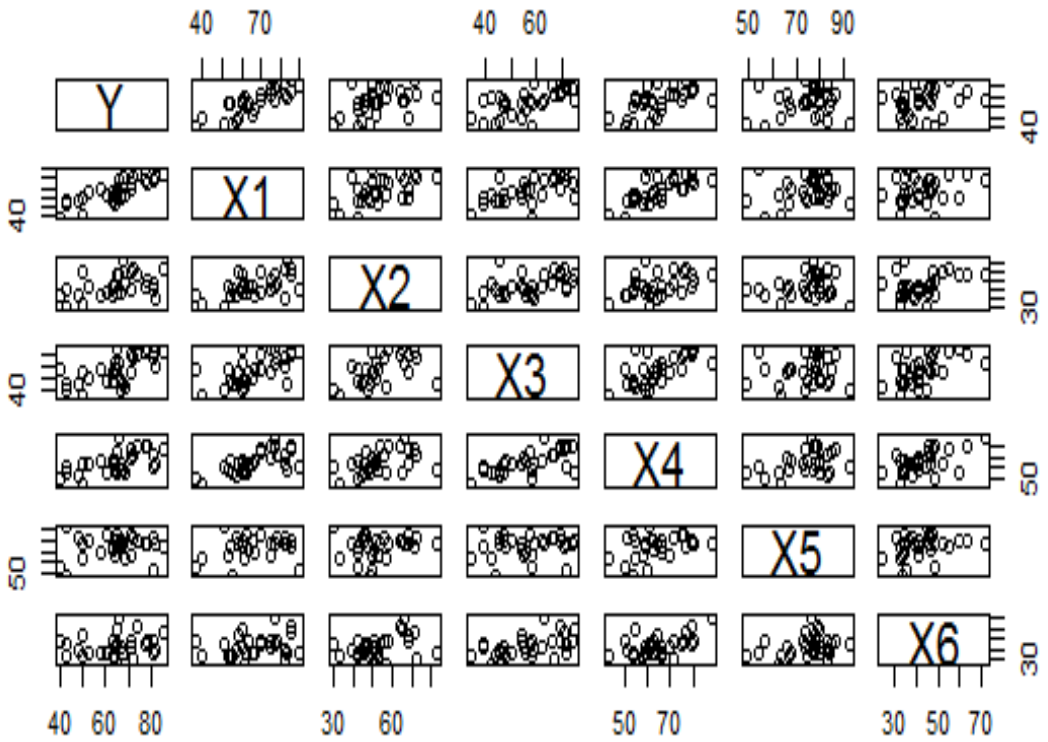
A recent survey of the clerical employees of a large financial organization included questions related to employee satisfaction with their supervisors. There was a question designed to measure the overall performance of a supervisor, as well as questions that were related to specific activities involving interaction between supervisor and employee. An exploratory study was undertaken to try to explain the relationship between specific supervisor characteristics and overall satisfaction with supervisors as perceived by the employees. Initially, six questionnaire items were chosen as possible explanatory variables. Table 3.2 gives the description of the variables in the study. As can be seen from the list, there are two broad types of variables included in the study. Variables XI, X2, and X5 relate to direct interpersonal relationships between employee and supervisor, whereas variables X3 and X4 are of a less personal nature and relate to the job as a whole. Variable x6 is not a direct evaluation of the supervisor but serves more as a general measure of how the employee perceives his or her own progress in the company. The data for the analysis were generated from the individual employee response to the items on the survey questionnaire. The response on any item ranged from 1 through 5, indicating very satisfactory to very unsatisfactory, respectively. A dichotomous index was created to each item by collapsing the response scale to two categories: {1,2}, to be interpreted as a favorable response, and

{3,4,5}, representing an unfavorable response. The data were collected in 30 departments selected at random from the organization. Each department had approximately 35 employees and one supervisor. The data to be used in the analysis, given in Table 3.3, were obtained by aggregating responses for departments to get the proportion of favorable responses for each item for each department. The resulting data therefore consist of 30 observations on seven variables, one observation for each department. We refer to this data set as the Supervisor Performance data.

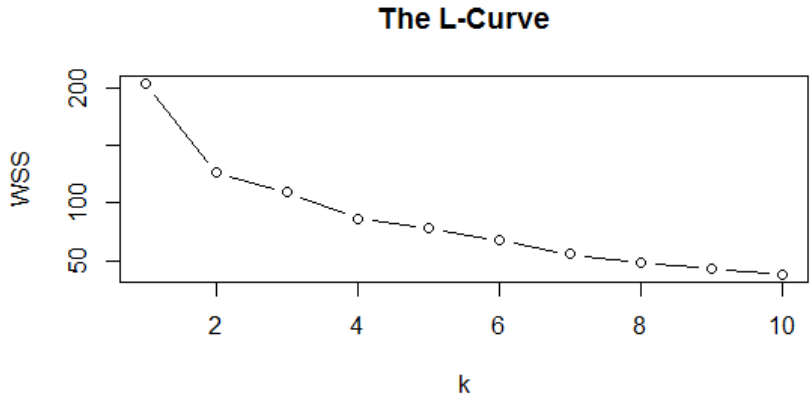
Table 3.2 Description of Variables in Supervisor Performance Data

Variable	Description
Y	Overall rating of job being done by supervisor
X_1	Handles employee complaints
X_2	Does not allow special privileges
X_3	Opportunity to learn new things
X_4	Raises based on performance
X_5	Too critical of poor performance
X_6	Rate of advancing to better jobs

III. Plot of the pairs:


Y
X1
X2
X3
X4
X5
X6
40 70
40 60
50 70 90
40
30
40
50
50
30
40 60 80
30 60
50 70
30 50 70

IV. Determining the number of clusters

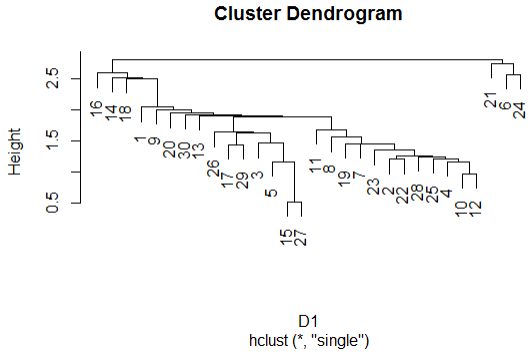


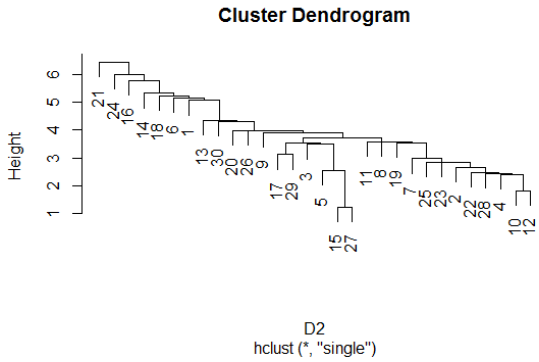
By looking at the L-curve and according to the elbow curve we can pick the number 2.

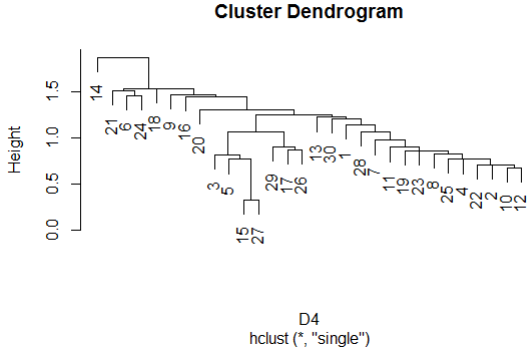
V. Hierarchal Algorithm:

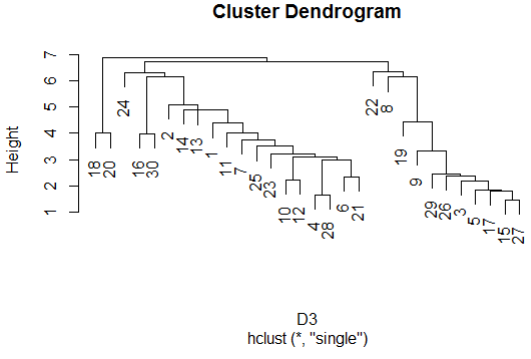
Plots of the Dendrograms

Using Single method and the four distance methods:

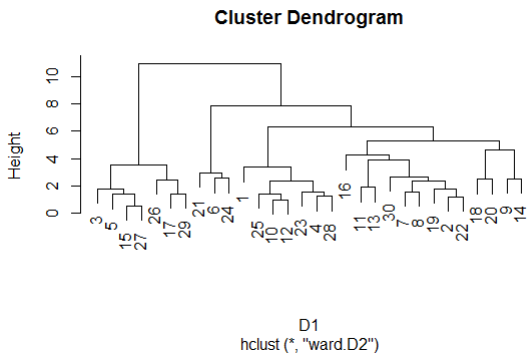


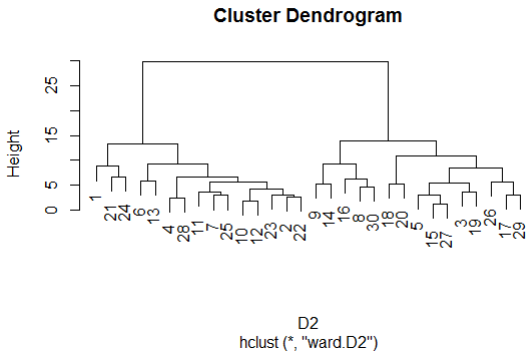


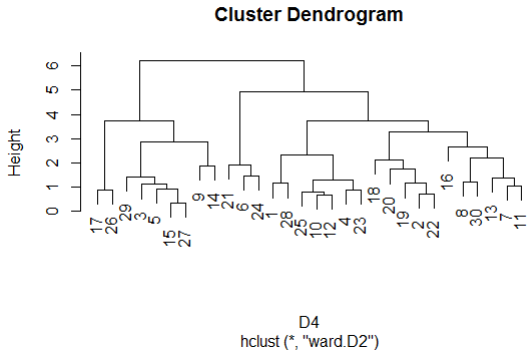


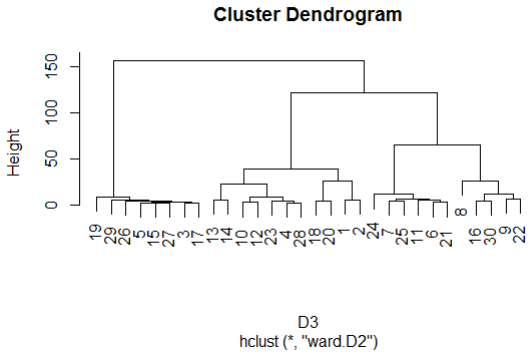


Using Ward.D2 method and the four between distance methods:

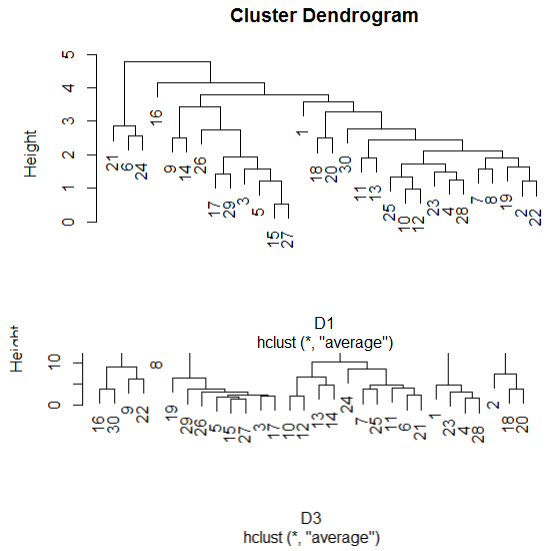


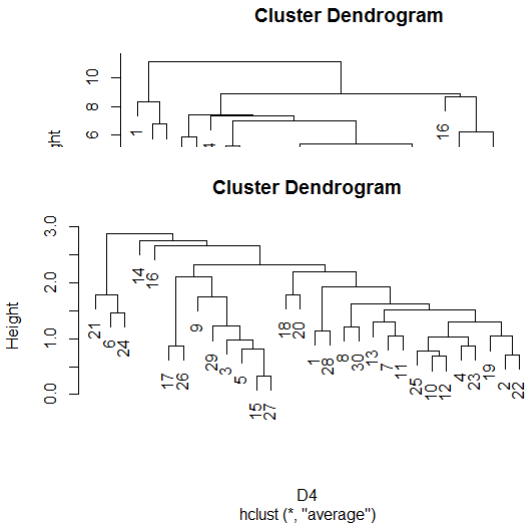




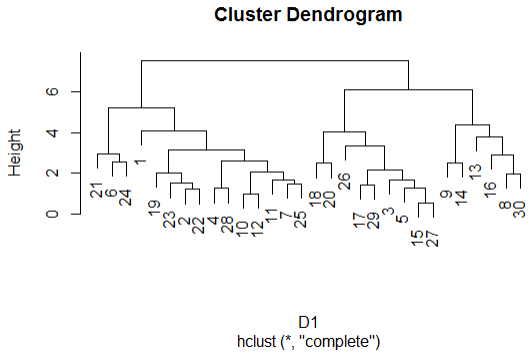


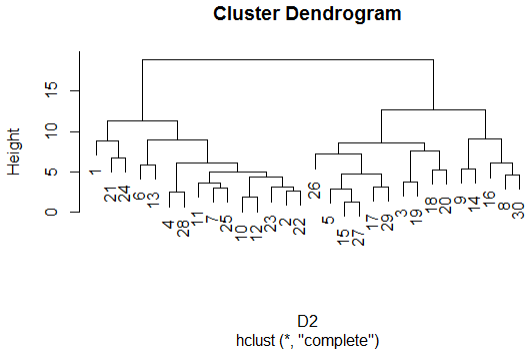
Using Average method and the four between distance methods:

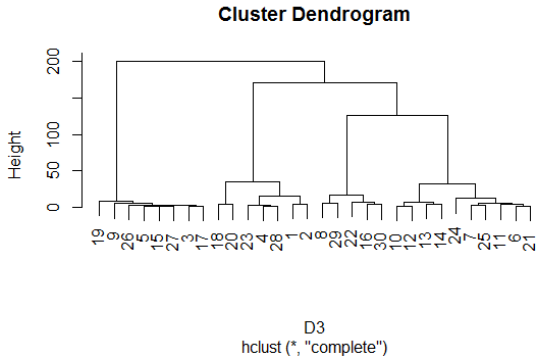


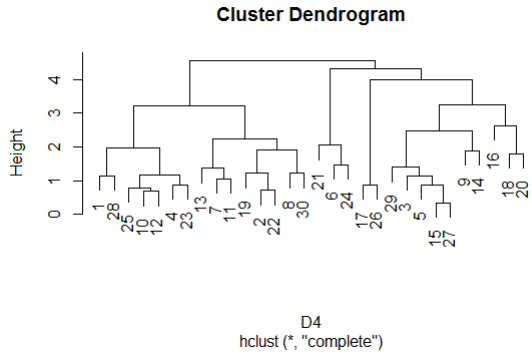


Using Complete method and the four distance methods:





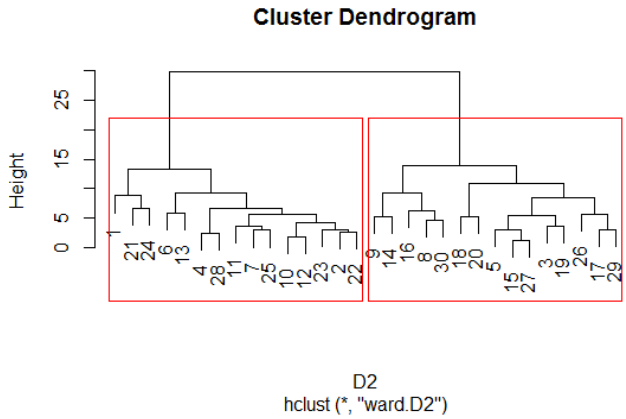




Analysis:

Using the hierarchal method, as seen in the above graphs all methods leaded to the conclusion that we cannot split the data into more than one cluster.

The only one that would allow for a split into two clusters is using the Ward.d2 distances between clusters and the Manhattan distance between observations.



```
+     }
> centers=aggregate(x, list(clusters), mean)
> centers
  Group.1         Y         X1         X2         X3         X4         X5         X6
1       1 -0.638047 -0.7410324 -0.6156983 -0.6958045 -0.6315787 -0.2189679 -0.4924494
2       2  0.638047  0.7410324  0.6156983  0.6958045  0.6315787  0.2189679  0.4924494
>
>
```

I obtained these two centers for the two groups.

Using the goodness of fit test, I obtain an R^2 of 0.5614695, this is a very low R2 which shows that the clustering is not very successful.

VI. K-Means method:

I tried this method by setting the number of clusters to 2 and obtained:

```
> kmc$centers
           Y         X1         X2         X3         X4         X5         X6
1 -0.6733039 -0.7350866 -0.5012765 -0.6649193 -0.5959122 -0.1848594 -0.4065947
2  0.7694902  0.8400990  0.5728874  0.7599078  0.6810425  0.2112679  0.4646797
> clusters=kmc$cluster
>
>
```

Using the goodness of fit test, we obtain an R^2 of 0.3786105 which is even lower than the one obtaine d from the hierarchal clustering method.

VII. Conclusion:

One can conclude that this dataset doesn't contain any clusters.