Dheeraj Goud Borlla (dxb160130)

ASSIGNMENT – #4: (CATEGORICAL AND MIXED DATA CLUSTERING)

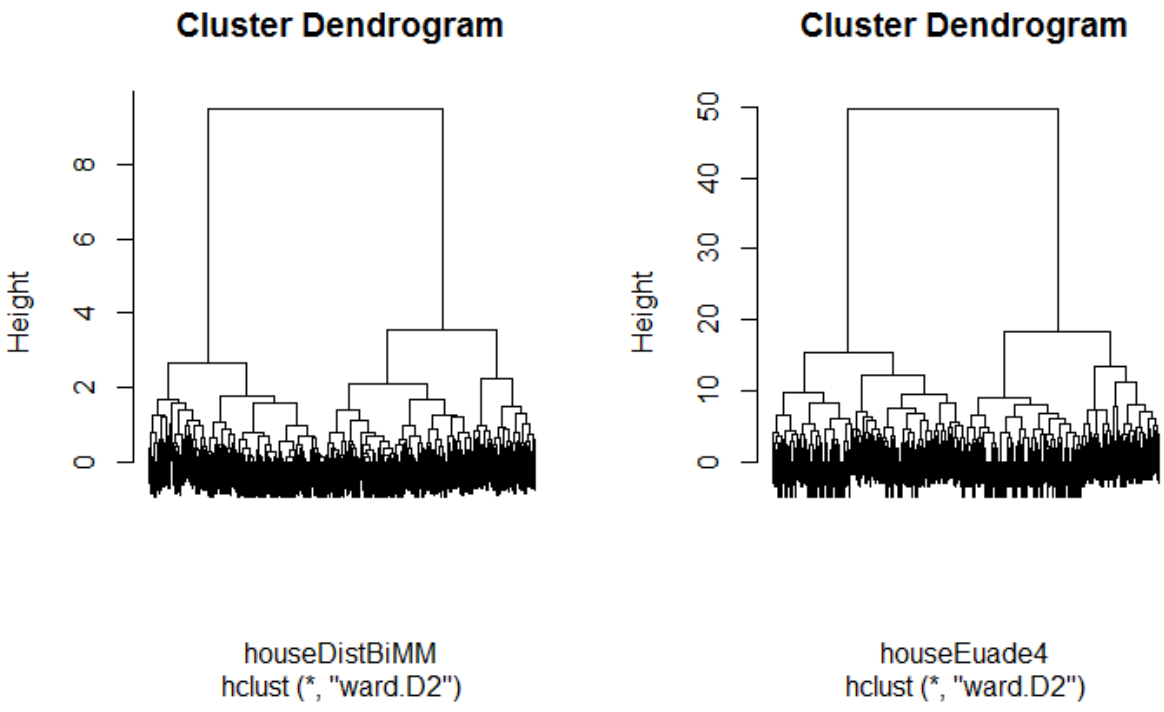
***PURPOSE*:** In this assignment, we apply clustering methods and try to cluster the categorical dataset (non-numerical data) and the mixed dataset (numerical and categorical data).

DATASETS: Congressional voting records (Link) and Wholesale Customers data (Link).

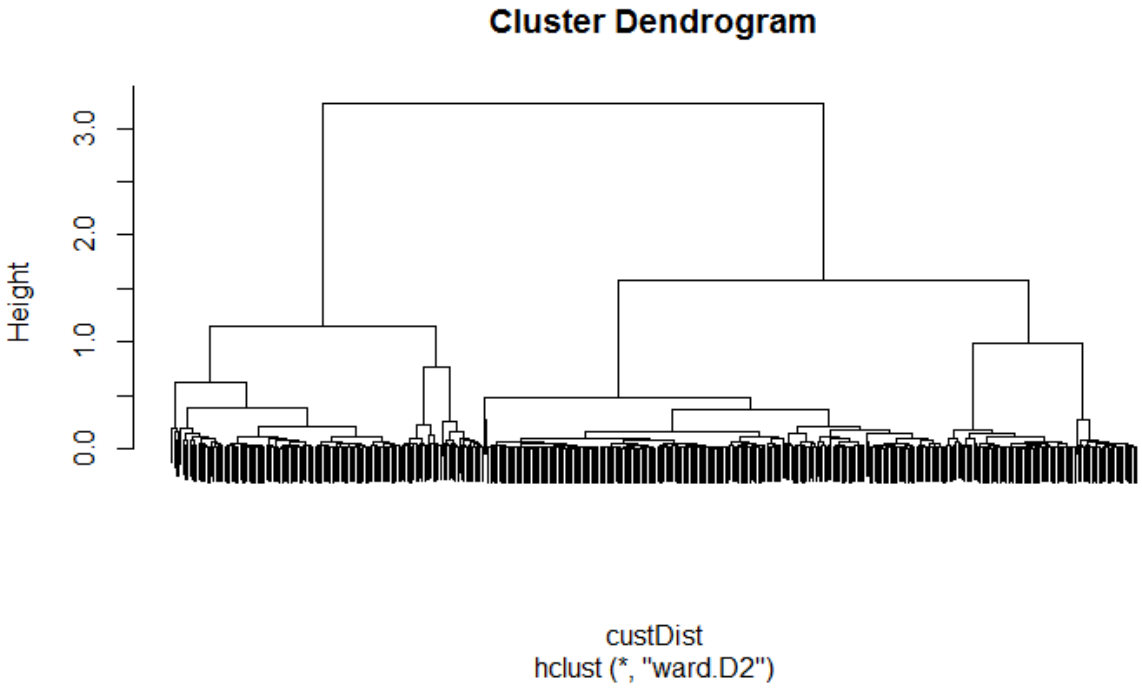
APPROACH:

- ***Congressional Voting Records* –** This dataset contains the information about votes for each of the congressmen on the 16 key votes.
- Since the data set contains data other than numerical data we can't use apply the K-means and hierarchal clustering methods.
- So, we replace the variables with dummy variables, that is categorical values are changed into binary values (either 0 or 1).
- We see that there are two groups namely, Republic and Democrat. We used the voting data of 16 different voting's and tried to observe whether the clustering results are matched.
- We applied the model matrix to the data, this transforms into dummy variables. Then we use the dist() metrics with which we can apply hierarchal clustering method.
- When we plot the dendrogram plot (first plot of graph1), it seems that two clusters are reasonable and 90.34% of the data are clustered exactly.
- Now, we applied the acm.disjonctif() from 'ade4' package. With this we give binary values for all the factors and this dataset is best suited because all are categorical values.
- Then again the dist() metrics is used and the hierarchal clustering is used to cluster the data. From the dendrogram plot (second plot of graph1), we can say that two clusters are reasonable and 93.22% of the data are clustered exactly.
- ***Wholesale Customer Data* –** This is a mixed dataset, has both numerical data and categorical values.
- For this dataset, we used daisy () method from the cluster package. This is a dist() metric "Gower", mainly used to handle the mixed data.
- Now, the clustering methods can be used on the transformed data.
- The dendrogram plot shown in graph 2. We see that the four clusters seem reasonable.
- So, we can cluster the customer's data into four groups.

GRAPHS:



(Graph 1)



(Graph 2)

SUMMARY:

- K-means clustering assumes the data, we wish to cluster, is numerical data because it works on the distance matrix, that is if the input data is numerical it takes the distance of the variables and try to cluster them but K-means fails for the categorical data and same goes with the hierarchal clustering too.
- So, we transform the categorical data (factors) into dummy variables, dummy variables (0 or 1). That is, for variable with two factors we add two dummy variables and give value either 0 or 1, the factor value for that particular row (0 1 or 1 0).
- Now, the transformed dataset has values 0 or 1 which are numerical. So, the distance metrics can be used on the dataset and Clustering methods can be used to cluster the dataset that contains non-numerical data.