# Hierarchical Clustering:

**1.Business Objective:Clustering for the crime data**

Step 1: Normalize or standardize the dataset as the variables are in different scale except the X variable.

Step 2: Calculate the distance between the each data point that is one record to another calculate the distance using the Euclidean distance.

Step 3: Calculate the distance between cluster to cluster using “complete” under linkage.

Step 4: Dendrogram is used to represent the clusters formed.Dendrogram is cut into 5 clusters in beginning to determine the distance between each clusters.

Step 5: New variable is created to store the cluster number for each rows.

Step 6: Mean of each variable is calculated using the aggregate function. Below is the mean values of all the variables.

Group.1 Murder Assault UrbanPop Rape

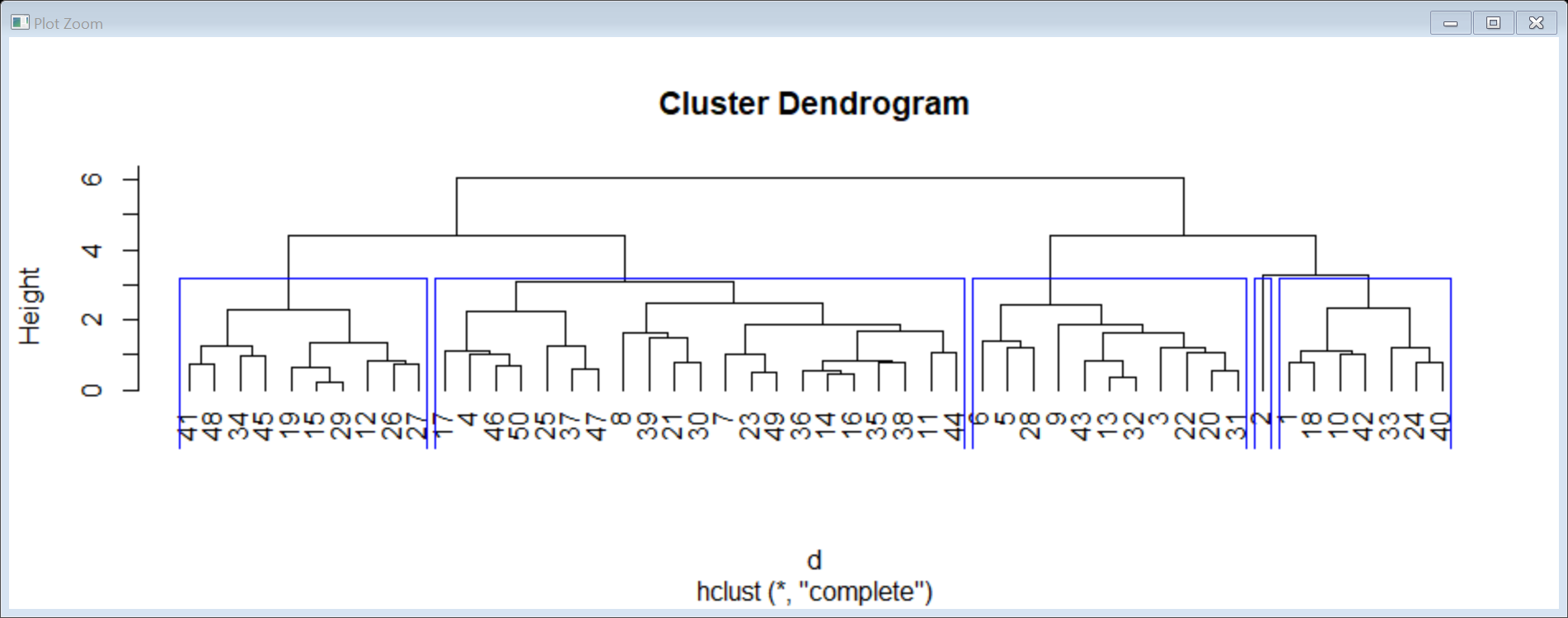
1 1 14.671429 251.2857 54.28571 21.68571

2 2 10.000000 263.0000 48.00000 44.50000

3 3 11.054545 264.0909 79.09091 32.61818

4 4 5.871429 134.4762 70.76190 18.58095

5 5 3.180000 78.7000 49.30000 11.63000



Step 7: Stability of the cluster is determined by using sample split method using train and test. In this stability test a 95% probability was calculated. After the sample split the values don’t seem to be appropriate so unable to determine the stability of the cluster.

Group.1 Murder Assault UrbanPop Rape

1 1 14.71667 246.6667 55.33333 21.55

2 2 10.00000 263.0000 48.00000 44.50

3 3 10.95000 265.0000 79.60000 32.37

4 4 5.80500 135.5500 71.05000 18.46

5 5 3.18000 78.7000 49.30000 11.63

Step 8: Stability of the cluster is determined by using sample split method using train and test. In this stability test a 70% probability was calculated. There seems to be lot of improvement in the data set.

Group.1 Murder Assault UrbanPop Rape

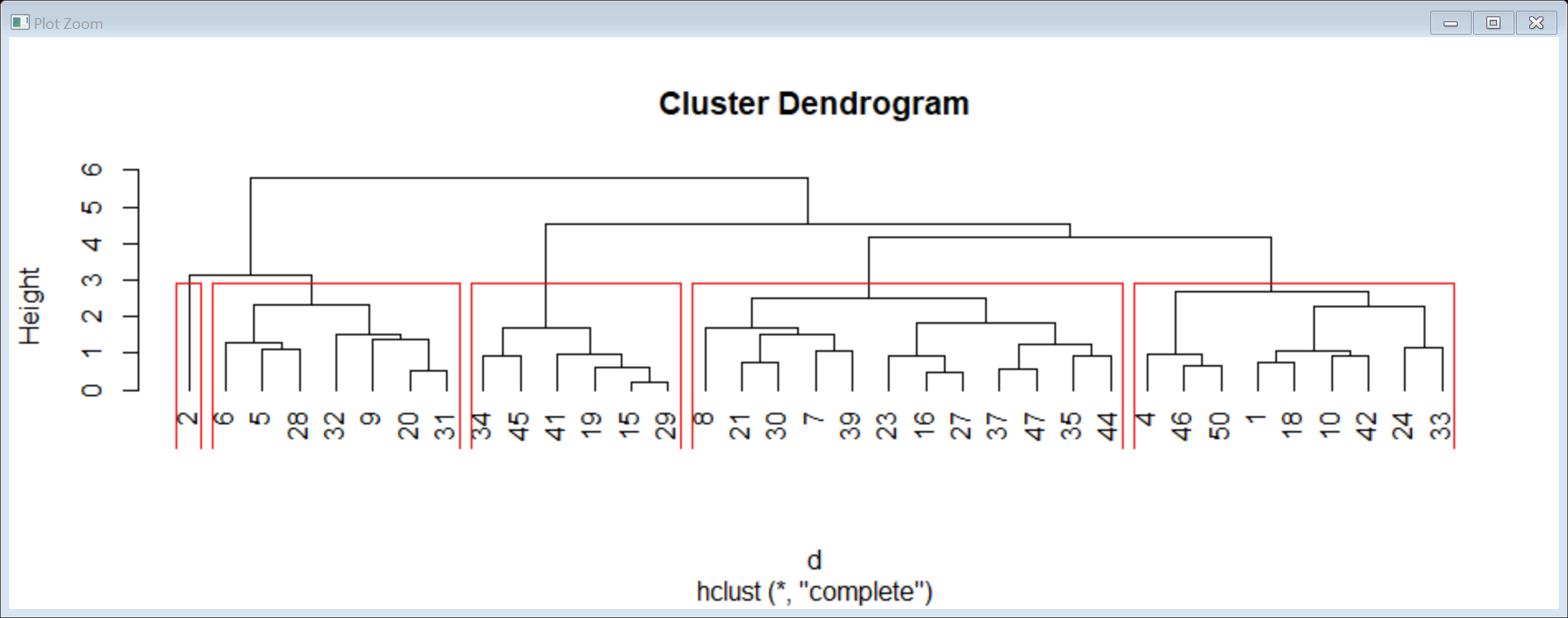
1 1 14.680000 262.40000 55.40000 21.56000

2 2 10.000000 263.00000 48.00000 44.50000

3 3 11.287500 268.25000 81.37500 32.15000

4 4 5.058333 98.58333 53.91667 14.23333

5 5 4.744444 146.77778 76.44444 18.46667



Step 9: Naming the clusters.

Cluster 3 – Most Frequent worry

Cluster 1 – More Persistent worry

Cluster 2 – More Frequent worry

Cluster 5 – Persistent worry

Cluster 4 – Frequent worry

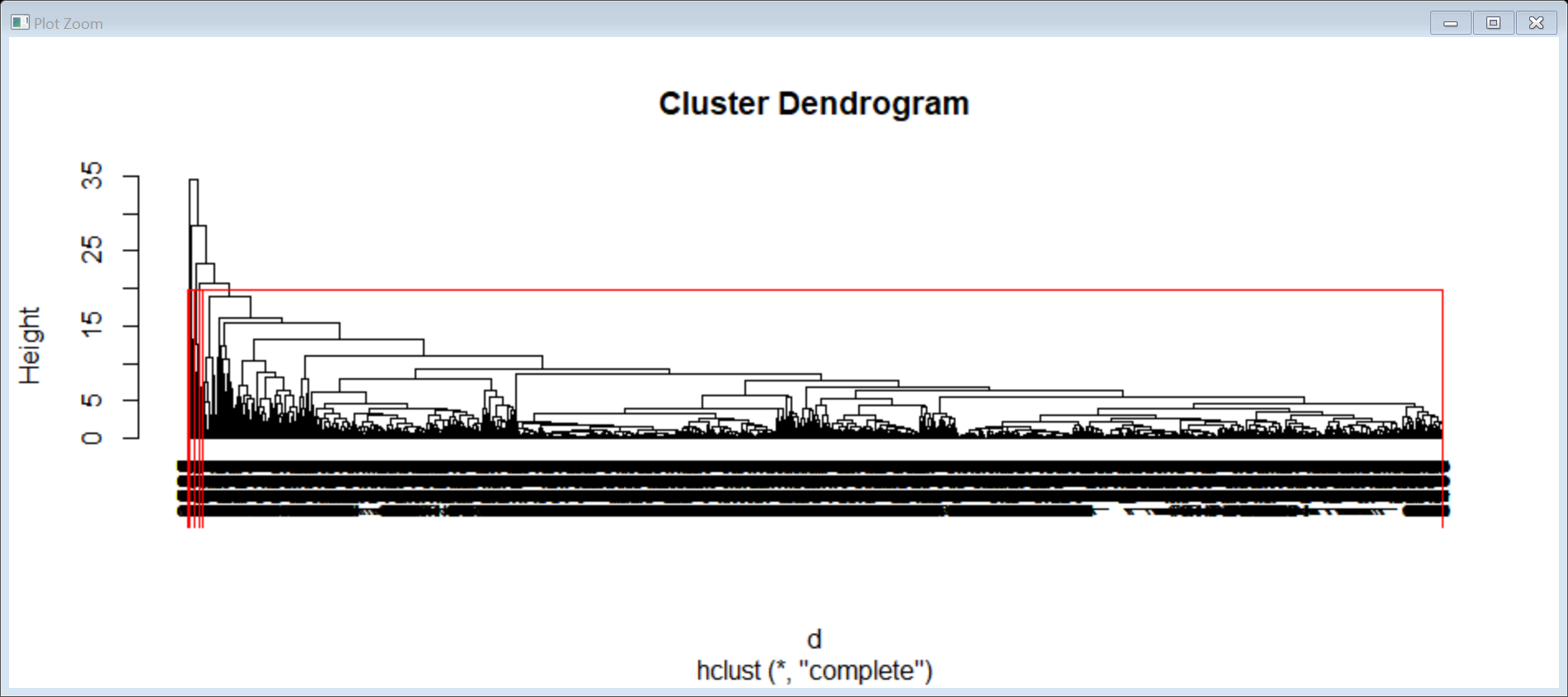
**2.Business Objective:Clustering for the different types of mileage offers**

Step 1: Normalize or standardize the dataset as the variables are in different scale except the ID variable.

Step 2: Calculate the distance between the each data point that is one record to another calculate the distance using the Euclidean distance.

Step 3: Calculate the distance between cluster to cluster using “complete” under linkage.

Step 4: Dendrogram is used to represent the clusters formed.Dendrogram is cut into 5 clusters in beginning to determine the distance between each clusters.Due to huge amount of data the dendrogram is not clear.



Step 5: New variable is created to store the cluster number for each rows.

Step 6: Mean of each variable is calculated using the aggregate function. Below is the mean values of all the variables.

Group.1 Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles

1 1 70225.28 111.3107 2.051366 1.014676 1.000759

2 2 138061.40 78.8000 3.466667 1.000000 4.066667

3 3 973710.31 746.6154 2.538462 1.000000 1.000000

4 4 102951.00 8275.8667 2.266667 1.000000 1.000000

5 5 131999.50 347.0000 2.500000 1.000000 1.000000

Bonus\_miles Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12

1 16762.41 11.46053 435.8350 1.303391

2 93927.87 28.06667 506.6667 1.600000

3 29401.08 16.61538 1576.3846 6.615385

4 17568.40 12.66667 627.3333 2.333333

5 65634.25 69.25000 19960.0000 49.250000

Days\_since\_enroll

1 4106.883

2 4613.867

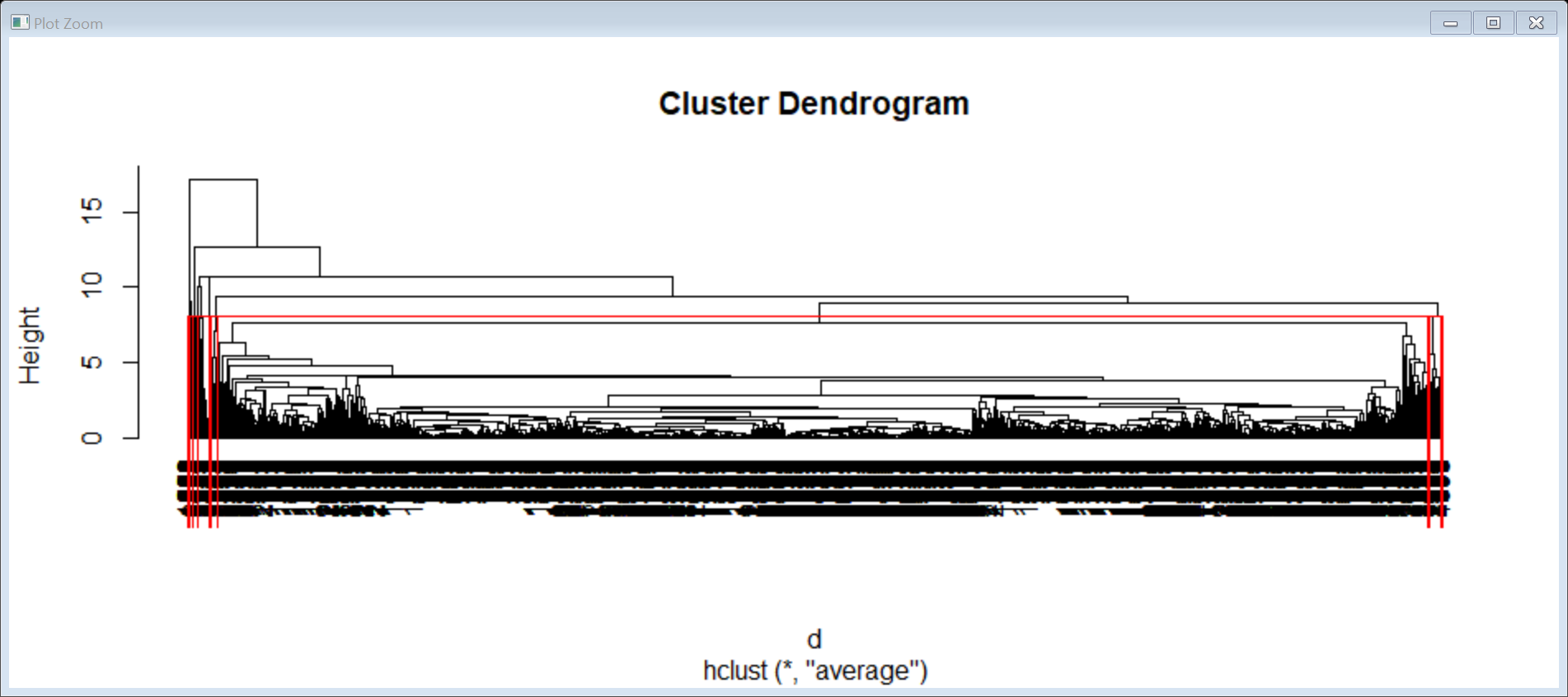
3 6972.846

4 4737.467

5 2200.250

Step 7: Stability of the cluster is determined by using sample split method using train and test. In this stability test a 95% probability was calculated. After the sample split the values don’t seem to be appropriate so unable to determine the stability of the cluster.

Step 8: Stability of the cluster is determined by using sample split method using train and test. In this stability test a 70% probability was calculated. There seems to be lot of improvement in the data set.



Group.1 Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles

1 1 73838.69 140.7010 2.063241 1.015092 1.000719

2 2 149080.00 107.4545 3.000000 1.000000 4.000000

3 3 1704838.00 0.0000 1.000000 1.000000 1.000000

4 4 114468.33 0.0000 1.000000 1.000000 1.000000

5 5 53232.00 888.0000 4.000000 1.000000 1.000000

Bonus\_miles Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12

1 17108.46 11.43730 431.9015 1.34567

2 77952.36 27.54545 654.5455 2.00000

3 17108.00 32.00000 4823.0000 23.00000

4 16461.67 14.66667 13928.3333 12.00000

5 80696.00 65.00000 22100.0000 45.00000

Days\_since\_enroll

1 4133.417

2 4457.000

3 7283.000

4 2210.667

5 3831.000

Step 9: Naming the clusters.

Cluster 5 - Highest Offer

Cluster 3 –Higher Offer

Cluster 2 – High Offer

Cluster 4 – Medium Offer

Cluster 1 – Low Offer