Problem (Seed# 28)

For this assignment you will use the churn dataset from the last assignment to explore the two clustering algorithms we discussed in class - hierarchical clustering and k-means.

Suppose we are interested in understanding the characteristics of the customers who leave. Focus on these cases by first removing the negative cases (result = STAY). Partition the resulting dataset using your own random number seed. Then, using the training set, cluster the data using both clustering models. Do not use the class labels in your clustering models.

Choose an appropriate value for the parameter k (number of clusters). You will need to justify this choice quantitatively, and separately for the two methods. Note that your two choices may not be the same. Document the approach you took for this choice and justify your conclusion, graphically if possible.

For both methods, you should next interpret your clusters. Give a short English description of what makes each cluster "special," that is, what feature values make it different from the other clusters, or different from the general trends in the data.

For each method, show a scatter plot of as many pairs of dimensions as you can (using the "Data" button). Which dimension pairs show a good separation of the clusters? Describe what you can learn from these plots.

Which of the two models do you think makes more sense? Justify your answer based on what you understand about the data.

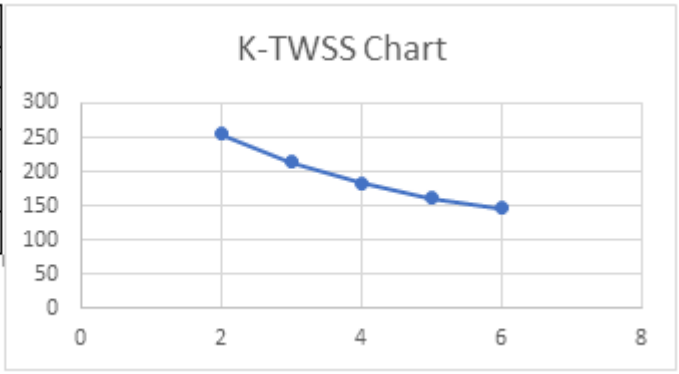
Data Preparation

Based on the churn data, the idea is to find the similarities between the group of customers who are likely to leave. All the customers who are likely to stay were removed from the data set to zoom in on the customers who are likely to leave. Since the different features are with different scales, all the numeric features were transformed to 0-1 scaling before clustering was done.

K-Means Clustering

Clustering was done with different values of K and TWSS was calculated. The table below shows the different TWSS for their corresponding K values. Even though the drop in TWSS is very gradual, it seems like the inflection point on the curve is at k=4.

K	TWSS
2	255
3	214
4	183
5	161
6	146



Cluster centers were also analyzed for each of the features when k=3, 4 and 5. It was found that the clustering made more sense when k was 4. Given below is the cluster centers when k=4

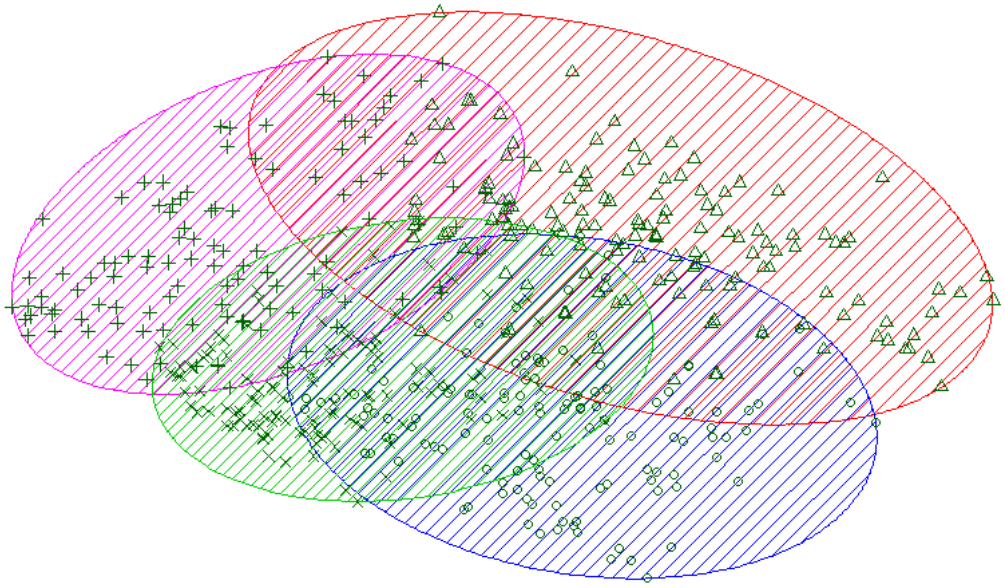
```
Cluster centers:

  R01_INCOME R01_OVERAGE R01_LEFTOVER R01_HOUSE R01_HANDSET_PRICE
1  0.3110232   0.6967624   0.18926923 0.2723344         0.1966422
2  0.8140426   0.4828086   0.19306852 0.3666250         0.7127478
3  0.3858046   0.2049149   0.71111768 0.3624906         0.2869543
4  0.2675132   0.1194541   0.07569987 0.3429928         0.1950701
  R01_OVER_15MINS_CALLS_PER_MONTH R01_AVERAGE_CALL_DURATION
1                      0.71478214                0.40808416
2                      0.37032540                0.39637827
3                      0.12764670                0.05639098
4                      0.07890123                0.56537530
```

Grouping in each of the 4 clusters was found to be very similar and can be explained as below.

Cluster	Cluster Description
1	Customers who went over their allotted minutes (High overage, long calls and low leftover mins)
2	High End Customers (High income and expensive handsets)
3	Customers who don't use all their allotted minutes (Low overage, high leftover mins, short calls)
4	Customers who utilized most of their allotted minutes (Low overage, low leftover mins)

Different groups also have low similarity. The below discriminant plot gives an idea about this.



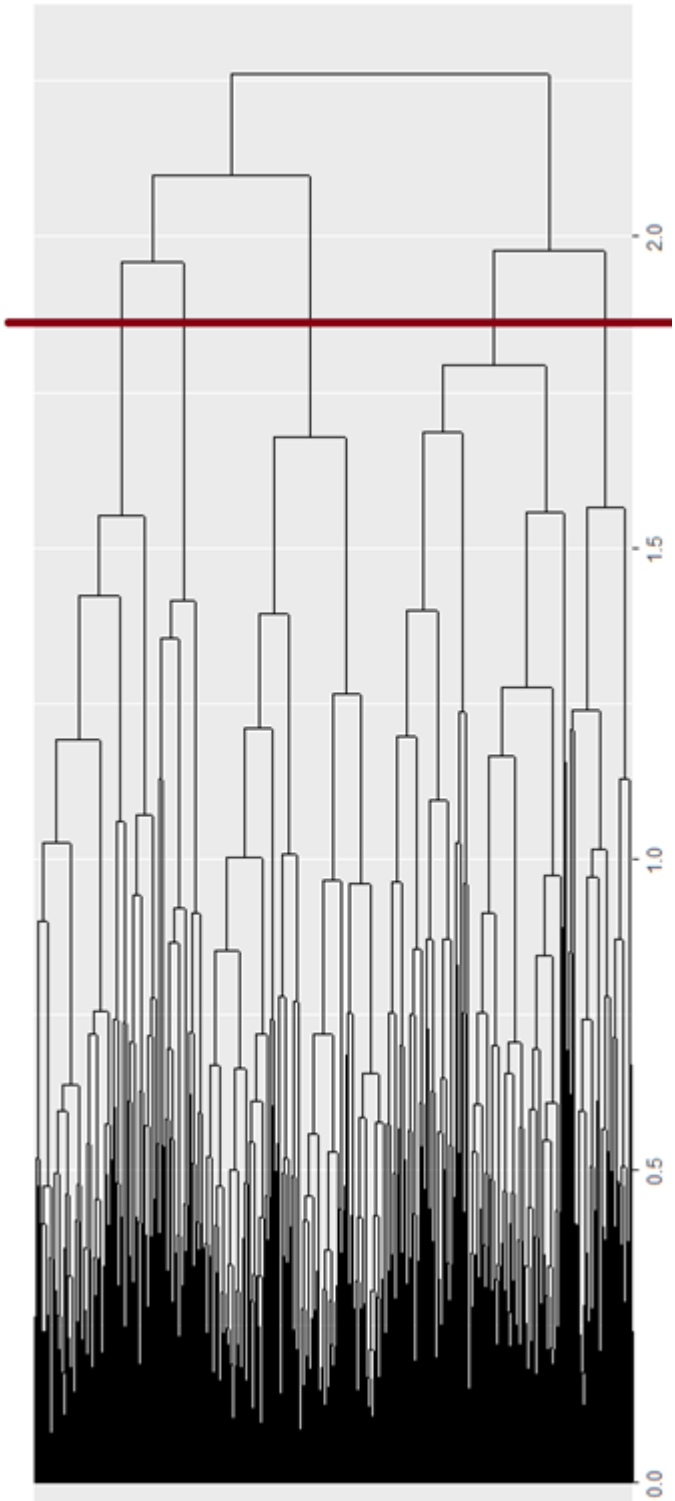
The 4 clusters were also found to be almost balanced.

Cluster Info	1	2	3	4
Cluster Sizes	129	142	114	118
Within cluster sum of squares	44.65677	66.57778	42.46026	29.63526

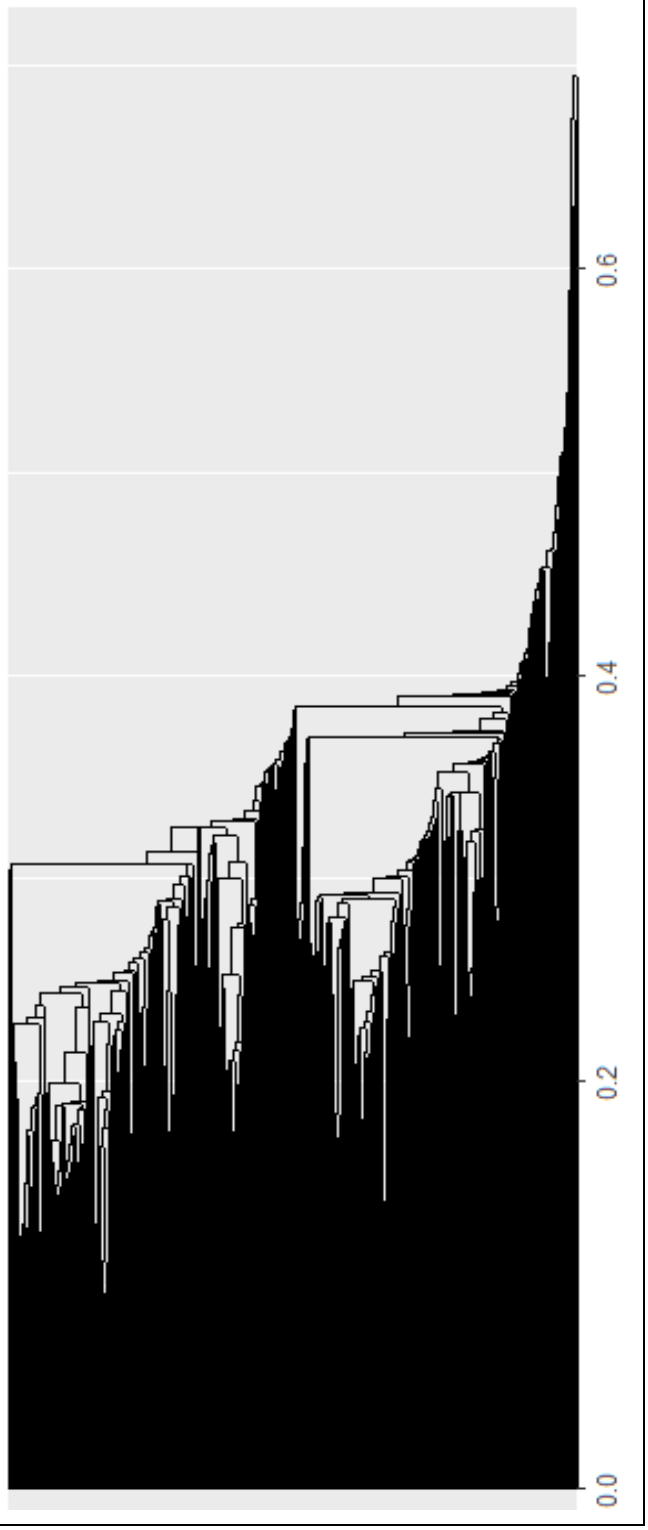
Hierarchical Clustering

Hierarchical clustering was done using single link and total(complete) link methods.

Complete/ Total Link Dendrogram



Simple Link Dendrogram



Complete/ Total link clustering was found to be very much balanced than the simple link method.

The total link cluster was found to be very much balanced than the single link method. A comparison of cluster sizes with both the methods can be found below.

Complete/ Total Link Dendrogram	***Simple Link Dendrogram***
`$cluster.size` `[1] 153 153  40 102  55`	`$cluster.size` `[1] 497   2   1   2   1`

Based on the above findings, complete link method was used to create the hierarchical clustering dendrogram.

Dendrogram was scanned from top to bottom and divided it once big spaces were gone. The line divided the tree at 5 points, making 5 clusters. K was found to be 5

Cluster centers were also analyzed for each of the features when k=5. Given below is the cluster centers when k=4

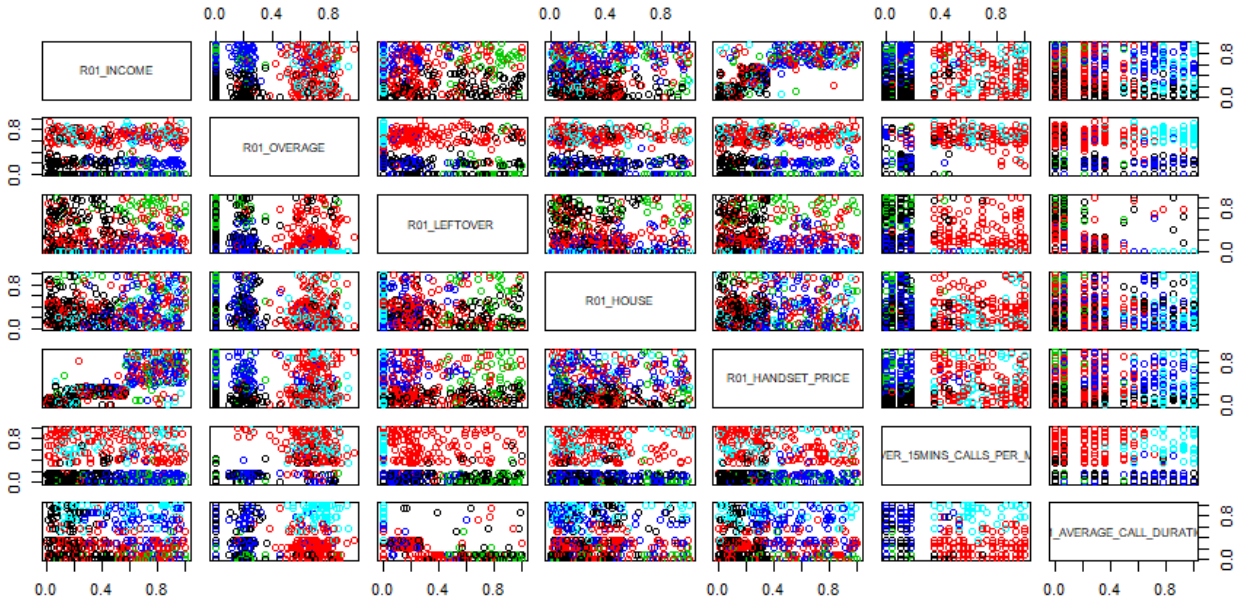
```
Cluster means:

     R01_INCOME R01_OVERAGE R01_LEFTOVER R01_HOUSE
[1,]  0.1950434   0.1651727   0.36439744 0.3063488
[2,]  0.4923346   0.6644491   0.30263641 0.3320413
[3,]  0.6648503   0.1055357   0.75505618 0.4833397
[4,]  0.7118460   0.1963585   0.09385327 0.3427013
[5,]  0.4889670   0.7112338   0.00000000 0.3095664
     R01_HANDSET_PRICE R01_OVER_15MINS_CALLS_PER_MONTH
[1,]         0.1456400                       0.09894073
[2,]         0.3922847                       0.66125761
[3,]         0.5146023                       0.06896552
[4,]         0.5655469                       0.10446247
[5,]         0.3908735                       0.70783699
     R01_AVERAGE_CALL_DURATION
[1,]                0.30578898
[2,]                0.22549020
[3,]                0.06071429
[4,]                0.51400560
[5,]                0.83506494
```

Grouping in each of the 5 clusters was found to be very similar and can be explained as below.

Cluster	Cluster Description
1	Low end customers (Low income and Low handset price)
2	Customers who are the highest callers with some leftover minutes
3	Customers who are the lowest callers with high leftover minutes
4	High end customers (High income and High handset price)
5	Customers who are the highest callers with no leftover minutes

Different plots generated with those 5 clusters are given below.



The best plot that describe our cluster is

Cluster Evaluation Summary

K-means clustering with 4 clusters gave a better result because the different groups had very low similarity. K-means clustering also gave a better-balanced cluster.

Two clusters (2 and 5) in the hierarchical clustering were very similar in nature which wasn't ideal.

Thank you for your time.