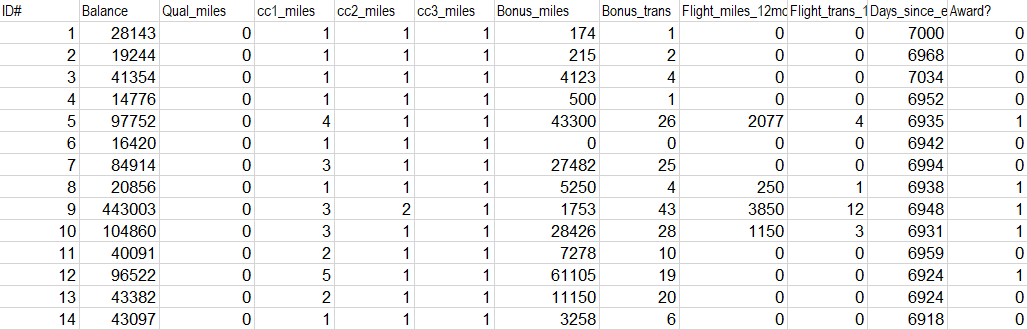
**East West Airlines Cluster Analysis**

1. Do you need to normalize the data before applying any clustering technique? Why or why not?

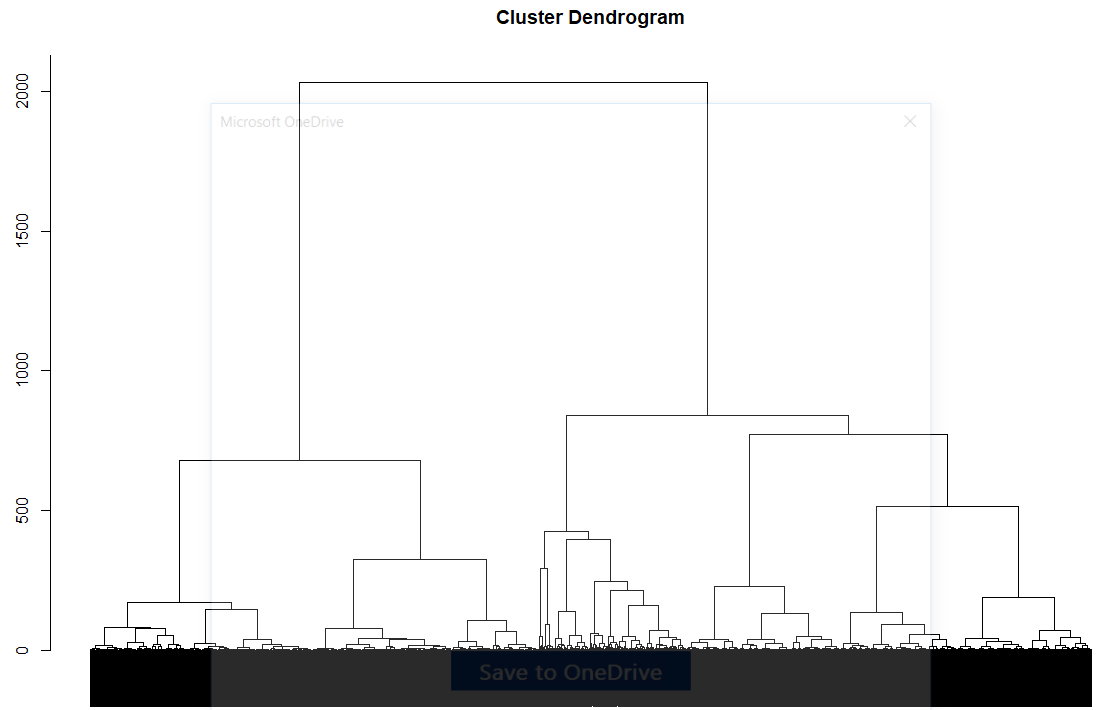
Yes, we need to normalize the data before applying any data. The reason is scale will be biased while calculating distance between clusters and also within clusters. Also, if we do not normalize the data, the large values will impact the variables having small values while calculating distance.

In East West Airlines data, by having physical look, it is observed that columns like Balance, Bonus\_miles, Days\_since\_enroll having high values when compared with other variables. Hence these column will highly skew the analysis if not normalized.

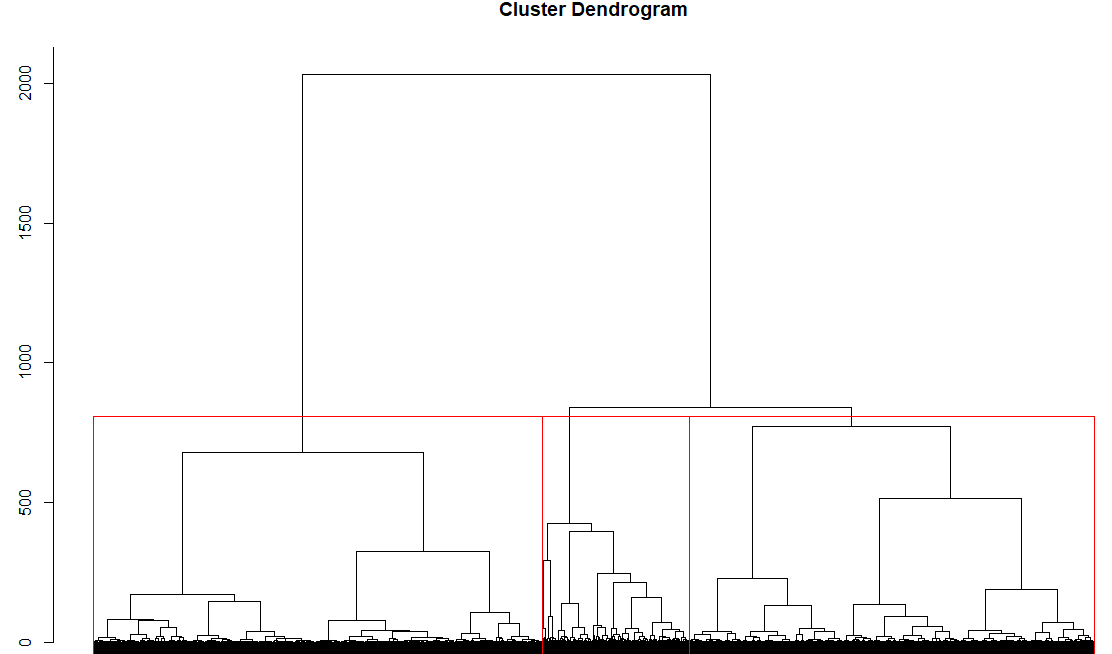


1. Apply hierarchical clustering with Euclidean distance and Ward’s method. How many clusters do appear?

Ward's minimum variance method is a special case of the [objective function](https://en.wikipedia.org/wiki/Objective_function) approach originally presented by Joe H. Ward. Ward suggested a general [agglomerative hierarchical clustering](https://en.wikipedia.org/wiki/Agglomerative_hierarchical_clustering) procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. This objective function could be "any function that reflects the investigator's purpose."



Ideally for hierarchical clustering, we can generate n clusters with each with single item. If we cut the cluster at 600, then we will get 3 primary clusters

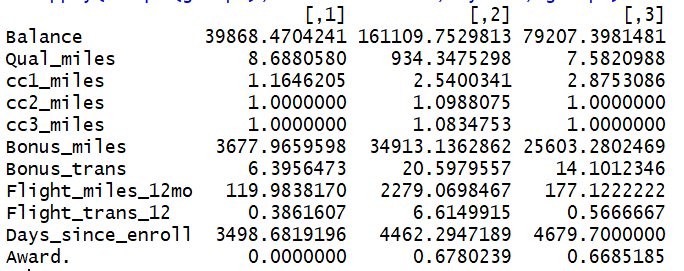


Here I used Ward.D method.

The difference between ward.D and ward.D2 is the difference between the two clustering criteria that in the manuscript are called Ward1 and Ward2.

It basically boils down to the fact that the Ward algorithm is directly correctly implemented in just Ward2 (ward.D2), but Ward1 (ward.D) can also be used, if the Euclidean distances (from dist()) are squared before inputing them to the hclust() using the ward.D as the method.

1. Compare cluster centroids to characterize different clusters and try to give each cluster a label—a meaningful name that characterizes the cluster.

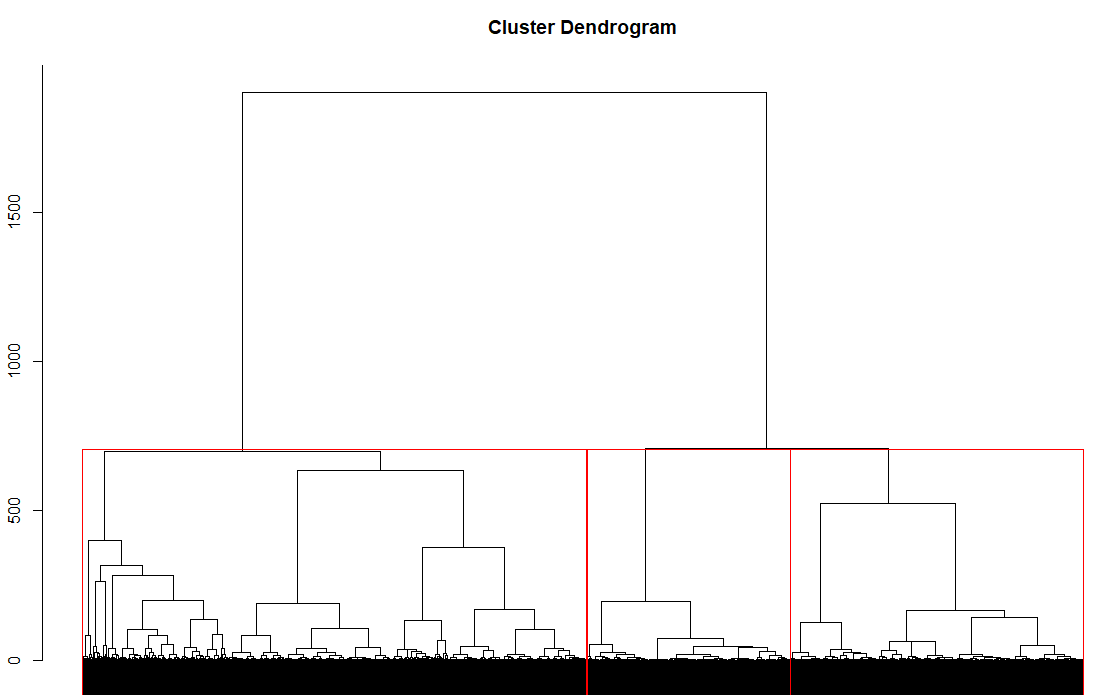


Cluster2 🡪 Flight\_miles in last 12 months is very much higher than other 2 clusters, also Qual\_miles in top flight is also high, hence this cluster can be tagged as “**FREQUENT BUSINESS CLASS TRAVELERS**”

Cluster3 🡪 Flight\_miles in less than cluster2 but very much higher than Cluster1. Also, Qual\_miles is very less, hence this cluster can be tagges as “**FREQUENT ECONOMY CLASS TRAVELERS**”

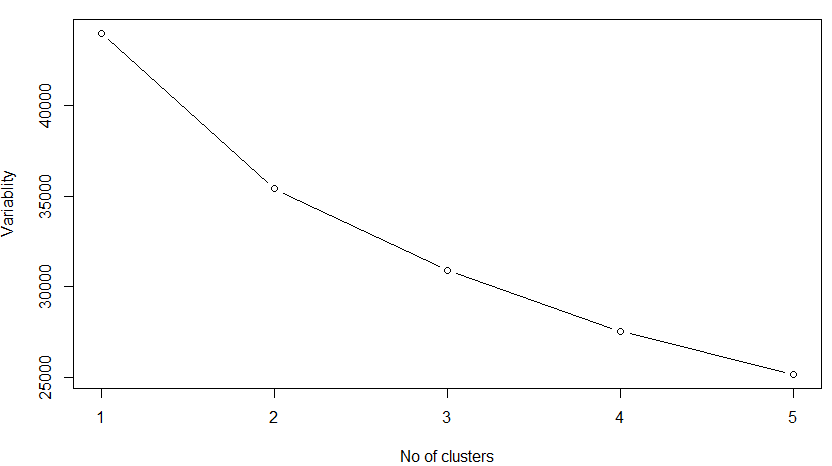
Cluster1 🡪 These are fliers apart from other 2 category, which can be tagged as “**OCCASIONAL TRAVELERS**”

1. To check the stability of clusters, remove a random 5% of the data (by taking a random sample of 95% of the records), and repeat the analysis. Does the same picture emerge?



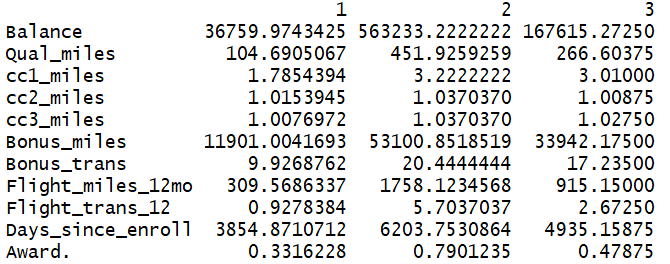
If we compare the new dendrogram with old one, we can see the changes in scale when clustering groups though the picture looks same. Hence even 5% change in samples, it will impact the clustering groups

1. Cluster all passengers again using k-means clustering. How many clusters do you want to go with? How did you decide on the number of clusters? Explain your choice on the number of clusters.



From the Elbow diagram, we can see at K=3, cluster variability is less when compared with previous clusters, hence we can go with 3 clusters for K-Means algorithm for anaysis

1. How do the characteristics of the clusters, obtained in Part (e), contrast or validate the finding in Part c above?



Per comparing with previous dendrogram hierarchical clustering, we can see similar pattern exists in K-means clustering technique

Cluster2 🡪 high flight\_miles in 12 months & also high Qual\_miles, hence this is **FREQUENT BUSINESS CLASS TRAVELERS**

Cluster3 🡪 Have flight miles less than cluster2 but comparatively much higher than cluster1. Hence this is “**FREQUENT ECONOMY CLASS TRAVELERS**”

Cluster1 🡪 Have much low values in all categories. Hence this is “**OCCASIONAL TRAVELERS**”

1. Which cluster(s) would you target for offers, and what type of offers would you target to customers in that cluster? Include proper reasoning in support of your choice of the cluster(s) and the corresponding offer(s).



We classified Cluster1 as “Occasional Travels” which are very high in no’s (3118), contributing more than 75% of total travelers. Hence, we need to target more offers to convert them in to “Frequent Travelers” first. **Target is discount in flight price, more bonus for more no of travels can be given.**