ASSIGNMENT SUBMISSION FORM

This will be the first page of your assignment

Course Name : **Unsupervised Machine Learning - 1**
Assignment Title : **Individual Assignment 1**
Submitted by : **Unnati Khinvasara**

Student Name	PG ID
Unnati Khinvasara	12120097

ISB Honour Code

- I will represent myself in a truthful manner.
- I will not fabricate or plagiarise any information with regard to the curriculum.
- I will not seek, receive or obtain an unfair advantage over other students.
- I will not be a party to any violation of the ISB Honour Code.
- I will personally uphold and abide, in theory and practice, the values, purpose and rules of the ISB Honour Code.
- I will report all violations of the ISB Honour Code by members of the ISB community.
- I will respect the rights and property of all in the ISB community.
- I will abide by all the rules and regulations that are prescribed by ISB.

Note: Lack of awareness of the ISB Honour Code is never an excuse for a violation. Please go through the Honour Code in the student handbook, understand it completely. Please also pay attention to the following points:

- Please do not share your assignment with your fellow students under any circumstances if the Honour Code scheme prohibits it. The HCC considers both parties to be guilty of an Honour Code violation in such circumstances.
- If the assignment allows you to refer to external sources, please make sure that you cite all your sources. Any material that is taken verbatim from an external source (website, news article etc.) must be in quotations. A much better practice is to paraphrase the source material (it still must be cited).

(Please start writing your assignment below)

Data Background

We have been given dataset of 400 passengers of an Airlines frequent flier program for performing clustering with a goal to target segments of passengers with mileage offers.

The Data Dictionary used for the same is as follows –

East-West Airlines is trying to learn more about its customers. Key issues are their flying patterns, earning and use of frequent flyer rewards, and use of the airline credit card. The task is to identify customer segments via clustering.

Source: Based upon real business data; company names have been changed.
(c) 2016 Galit Shmueli and Peter Bruce

Field Name	Data Type	Max Data Length	Raw Data or Telcom Created Field?	Description
ID#	NUMBER		Telcom	Unique ID
Balance	NUMBER	8	Raw	Number of miles eligible for award travel
Qual_miles	NUMBER	8	Raw	Number of miles counted as qualifying for Topflight status
cc1_miles	CHAR	1	Raw	Number of miles earned with freq. flyer credit card in the past 12 months:
cc2_miles	CHAR	1	Raw	Number of miles earned with Rewards credit card in the past 12 months:
cc3_miles	CHAR	1	Raw	Number of miles earned with Small Business credit card in the past 12 months:
note: miles bins:				1 = under 5,000
				2 = 5,000 - 10,000
				3 = 10,001 - 25,000
				4 = 25,001 - 50,000
				5 = over 50,000
Bonus_miles	NUMBER		Raw	Number of miles earned from non-flight bonus transactions in the past 12 months
Bonus_trans	NUMBER		Raw	Number of non-flight bonus transactions in the past 12 months
Flight_miles_12mo	NUMBER		Raw	Number of flight miles in the past 12 months
Flight_trans_12	NUMBER		Raw	Number of flight transactions in the past 12 months
Days_since_enroll	NUMBER		Telcom	Number of days since Enroll_date
Award?	NUMBER		Telcom	Dummy variable for Last_award (1=not null, 0=null)

Further, we have combined certain variables to form a new meaningful variable. For eg – Bonus Miles per Transaction Ratio is created out of “bonus_miles & “bonus_trans” and same with Flight Miles per Transaction Ratio. This gives a better representation of the data at hand telling us miles accumulated per transaction. Higher Ratio means that fewer transactions were made to accrue higher miles (pointing at spending capacity / flight distance category inputs respectively.

Also, while analysis we have noted that all 3 cc_miles variables are ordinal and majority belong to only 1 class. Hence it does not add value in segmentation of data. And this, we have removed these values.

Further, variables such as ID (unique) and Awards (categorical) have also been deleted in final analysis.

We have finally worked with a set of only 5 variables x 3999 observations. The variables used are as follows –

- Balance
- Qualification Miles
- Bonus Miles per Transaction Ratio
- Flight Miles per Transaction Ratio
- Days Since Enrollment

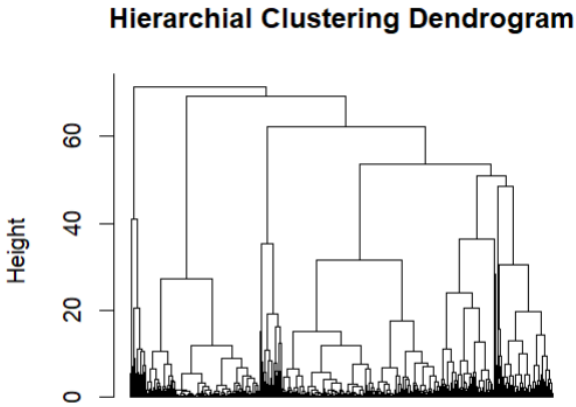
Q(a) - Do you need to standardize the data before applying any clustering technique? Why or why not?

Standardization (or Normalization) Techniques are required on the data before applying any clustering techniques to get the variables / data points on the same scale since every variable has a diverse range of values which can skew the end clustering result in the favour of the heavy weighted variables.

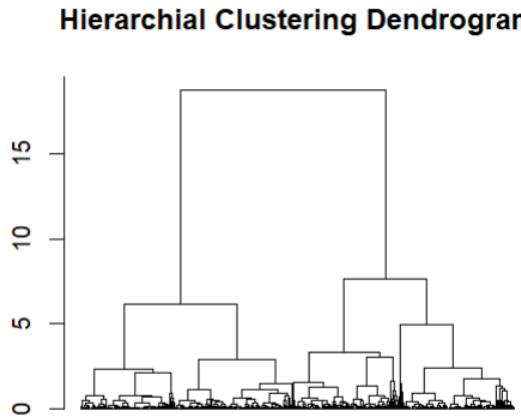
Thus, in the current assignment data has been scaled using standardization as well as normalization technique. The scalar which ends up giving better cluster segmentation is finally adopted.

Q(b) - Apply hierarchical clustering with Euclidean distance and Ward's method. How many clusters do appear?

Using Euclidean Distance and Ward's Method, we get the following dendrogram. The number of clusters in hierarchical clustering are dependent on the vertical heights of the branches in the dendrograms.



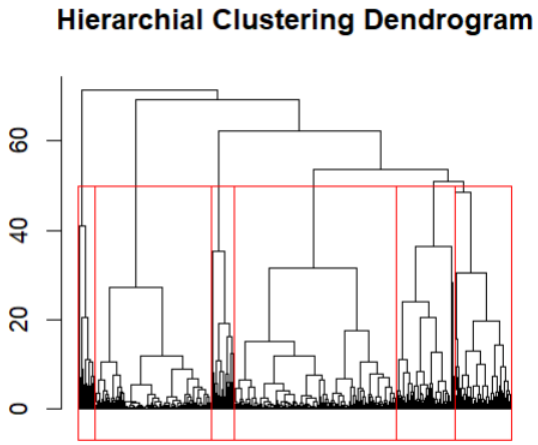
(Using Standardization scaling method)

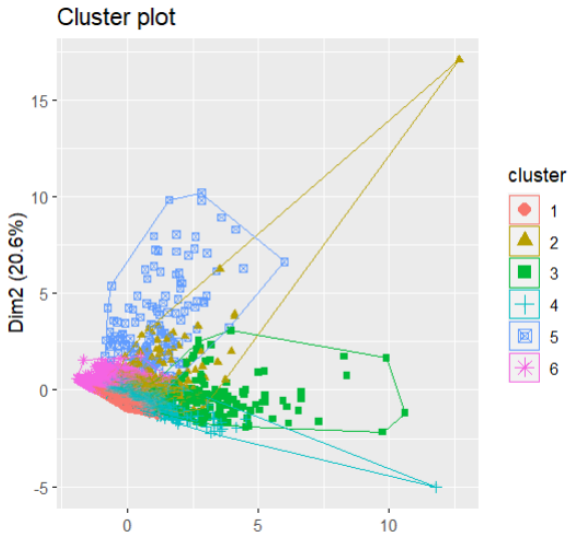


(Using Normalization scaling method)

Looking at the above, we note that the dendrograms of the same dataset using different scaling methods are varied. However, the vertical height difference using the standardization method is more pronounced (and further in the analysis will help in better segment analysis), hence we go ahead with the standardization method dendrogram.

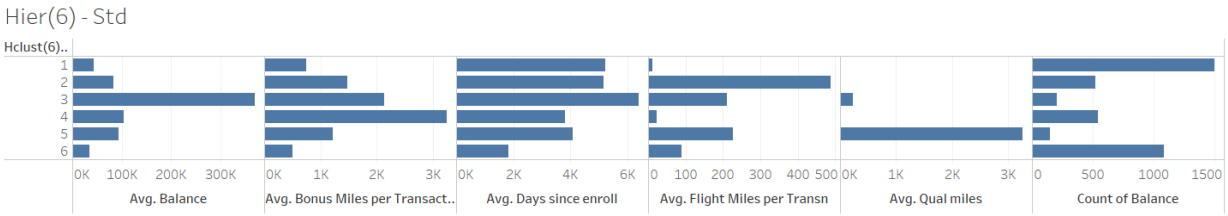
No. of clusters using the dendrogram depends on the business problem as well as the dendrogram height differentiation. Another popular way of finding optimal number of clusters is by visual inspection of the dataset. Currently, we can go ahead with 6 clusters as those are distinct (pictorial representation given as below).





Q(c) - Compare cluster centroids to characterize different clusters and try to give each cluster a label—a meaningful name that characterizes the cluster.

To compare cluster centroids, we use the average of variable within each cluster. A representation of the same is as follows –



Hclust(6) - Std	Avg. Balance	Avg. Bonus Miles per Transaction	Avg. Days since enroll	Avg. Flight Miles per Transn	Avg. Qual miles	Count of Balance
1	44,253	739	5,214	11	9	1,501
2	83,694	1,472	5,173	488	18	522
3	370,967	2,126	6,403	210	233	205
4	104,256	3,249	3,801	22	17	540
5	93,246	1,225	4,081	227	3,261	149
6	35,101	510	1,822	90	9	1,082

According to the cluster centroids found above, we can characterize the different clusters as below –

Cluster 1 – General Low Mile Flyers **1501 passengers**
This cluster has the highest number of passengers with an above average days since enrollment. However, they have a low average balance as well as low bonus miles / flight miles. Thus, we can characterize these passengers as

Cluster 2 - International Flyers **522 passengers**
This cluster has the highest average flight miles per transaction ratio which indicates that these passengers are international / long flight flyers which gives them high flight mile rewards per transaction.

Cluster 3 – Irregular yet High Miles Category Flyers **205 passengers**
This cluster is characterized by a low number of flyers enrolled since a long time in the frequent flyers program with the highest accumulated average balance as well as a decent bonus & flight ratio. This indicates that this is a small group yet high potential group since it has an unused accumulated balance lying in their miles program.

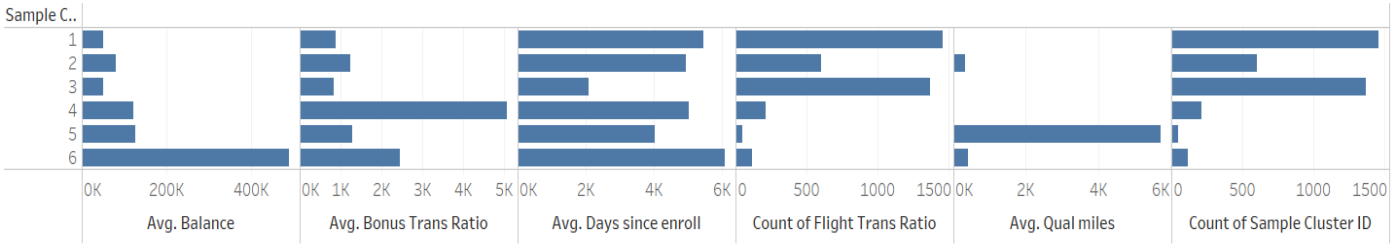
Cluster 4 – Non-Flight Bonus Mile Flyers **540 passengers**
This cluster has the highest average of non-flight bonus miles per transaction coupled with a low flight miles transaction ratio. This indicates that these are small route flyers who have accumulated miles through non-flight related bonus benefits like Credit Card Reward Program.

Cluster 5 – Elite Flyers **149 passengers**
This is a small cluster which has almost the majority of the flyers with high qualification miles, i.e., topflight mile status. This indicates that these are regular flying elite passengers.

Cluster 6 – New Small Budget Flyers **1082 passengers**
This cluster has the lowest days since enroll making them recent joiners to the frequent flyers program which have a low average mile balance and low transaction ratios.

Q(d) - 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 emerges?

If we take a random sample of 95% and perform the same clustering and analysis, the following picture emerges.



To simplify further, we have plotted down the cluster wise differentiation below –

Clusters	Cluster No. in Full Data		Cluster No. in Sample Data	
General Low Mile Flyers	1	1,501	2	603
International Flyers	2	522	1	1,459
Irregular yet High Mile Flyers	3	205	6	113
Non-Flight Bonus Mile Flyers	4	540	4	212
Elite Flyers	5	149	5	46
New Small-Budget Flyers	6	1,082	3	1,367
		3,999		**3,800**

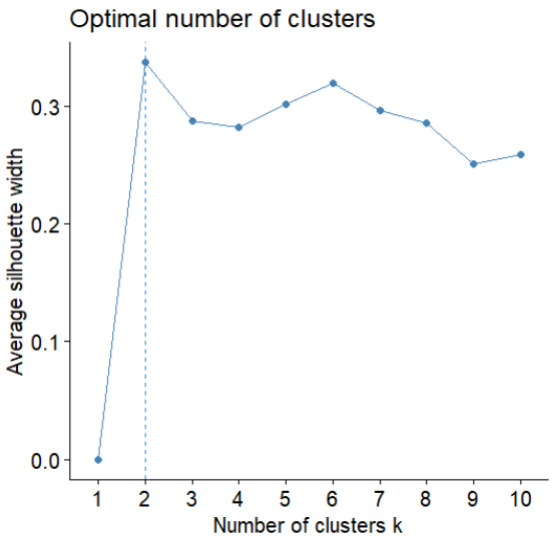
From the above, we can see that Clusters 3 to 6 have more or less remained same, however Cluster 1 and Cluster 2 have gotten overlapped. A different linkage or scaler might help reduce this overlap.

Q(e) - 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.

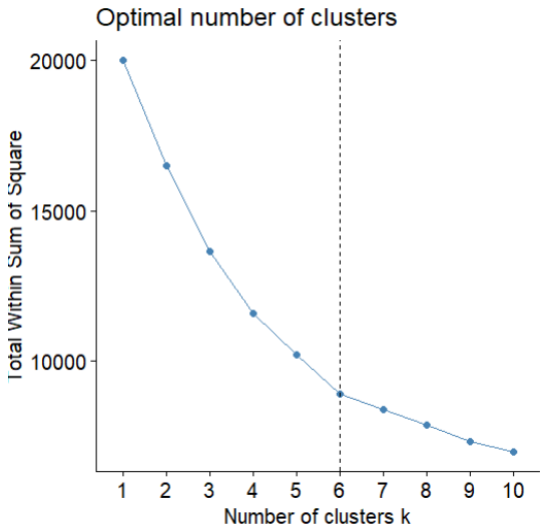
Number of Clusters in K-Means is decided based on within-cluster sum of squares which can be visualized with either an elbow plot or the silhouette score or other techniques. Here we have used a mix of silhouette score and elbow plot.

First, when we plot down the silhouette score, the optimal cluster number comes to be 2. However, it is worth noting that the second peak of the silhouette score is at 6.

Further, when we plot down the elbow plot, we see a slight dip at 6 and hence we can conclude that the optimal number of clusters could be 6.



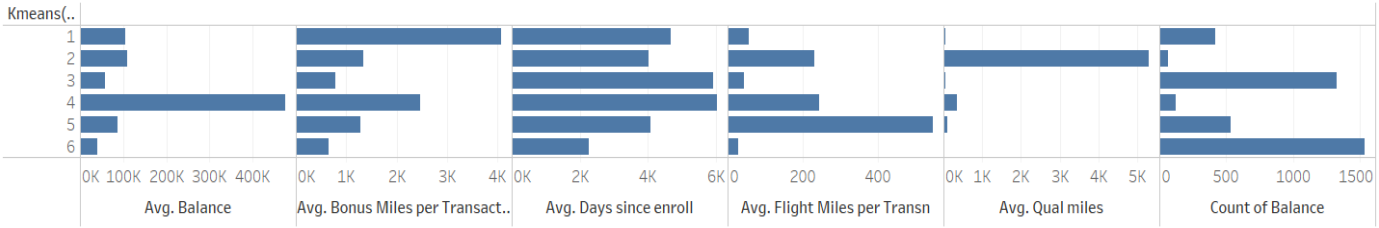
Silhouette Score Visualization



Elbow Plot Visualization

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

The Segmentation Analysis on the K Means Clusters is as follows -



Further, if we compare the Hierarchical Clustering with the K Means, the following picture emerges. We can see that the clusters are pretty similar quantum wise and property wise with the earlier analysis. The only small overlap is for Cluster 1 & 6 which are New Small Budget Flyers and General Low Miles Flyers.

Clusters	Hierarchical		K Means Clustering	
General Low Mile Flyers	1	1,501	3	1,325
International Flyers	2	522	5	530
Irregular yet High Mile Flyers	3	205	4	124
Non-Flight Bonus Mile Flyers	4	540	1	419
Elite Flyers	5	149	2	63
New Small-Budget Flyers	6	1,082	6	1,538
		3,999		**3,999**

Q(g) 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).

The clusters which can be targeted by the Airlines for offers are as follows –

Cluster 1 – General Low Mile Flyers could be targeted for higher miles per flight. This would help increase the frequency of flyers.

Cluster 2 – International Flyers could be targeted for Qualifying Bonuses (i.e. Top-Flight Status rewards) on in-flight purchases / shopping)

Cluster 3 – Irregular yet High Mile Flyers could be given free upgrades in the flights based on their miles accumulation. This would encourage this group to fly more often.

Cluster 4 – Non-flight Bonus Mile Flyers are people who spend and earn miles through non-flight bonuses. The Airlines could offer them higher rewards for flight transactions.

Cluster 6 – This cluster is the newest joiner group and could be encouraged for loyalty benefit rewards on flying continuously through the airlines.