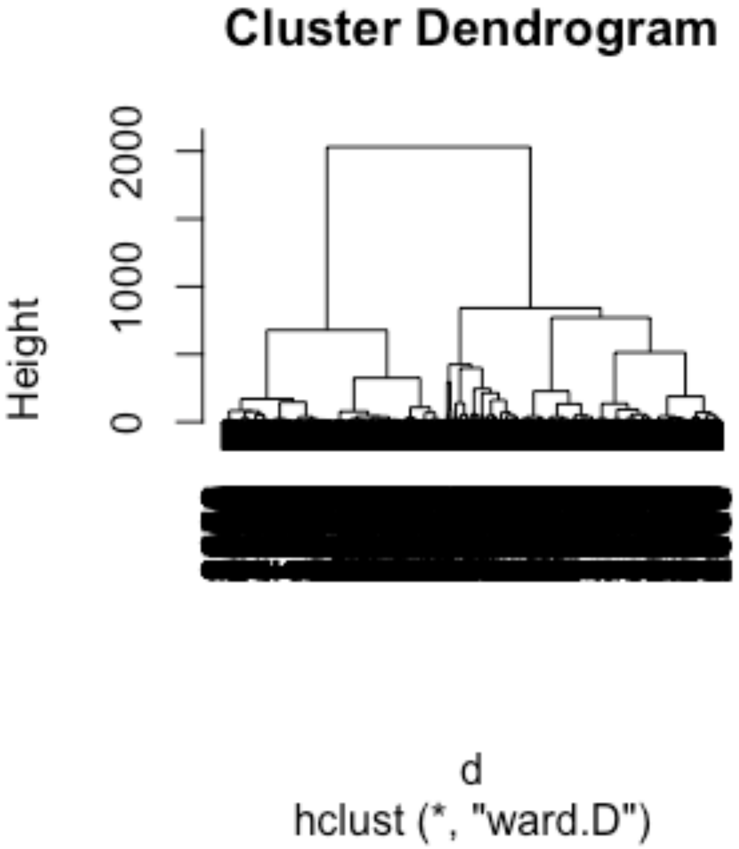
1a.Do you need to normalize the data before applying any clustering technique? Why or why not?

A. Yes we need to normalise the data as few columns are categorical like cc1,cc2,cc3. Few columns are summation of number of trips and few are sum of miles. So as the units of all the columns are not same we need to normalise them.

B.) Apply hierarchical clustering with Euclidean distance and Ward's method. How many clusters do appear?

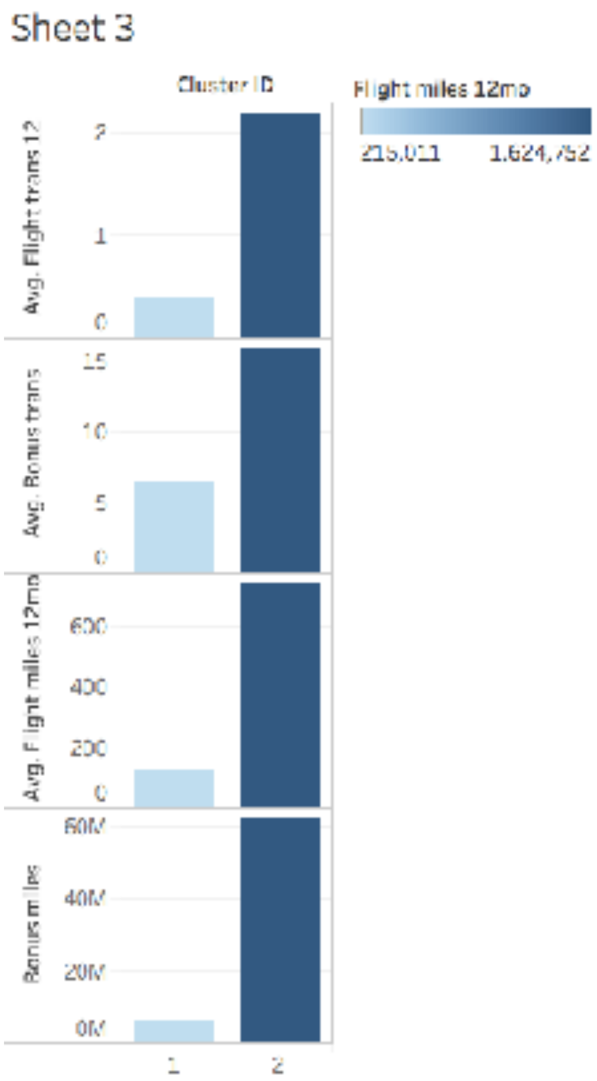
A. The dendrogram cluster figure is drawn below.



In the above figure having two clusters is a good idea. When I looked at the cluster data.

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

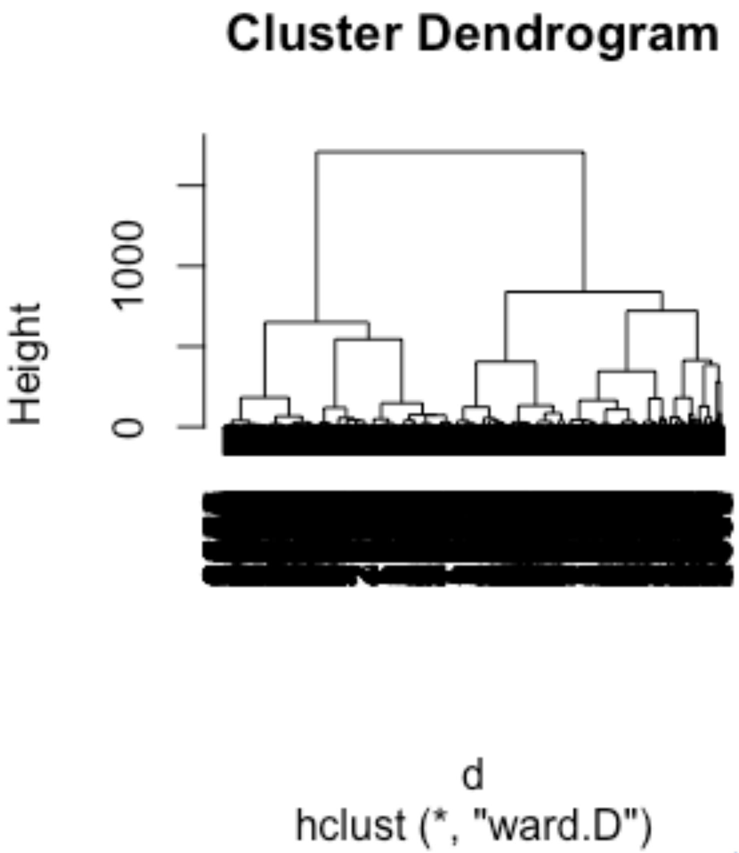
A. When I look at the two clusters using tableau, in the below figure. It clearly gives me two clusters one with frequent fliers and one with non frequent fliers. SO using this custers the airline industry can figure out their selling plan.



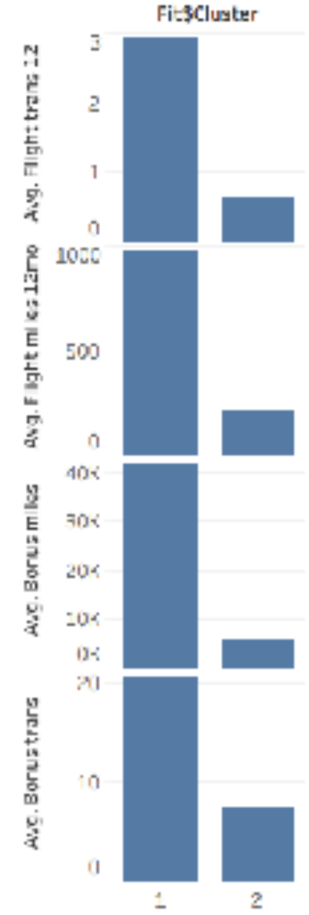
Average of Flight trans 12, average of Bonus trans, average of Flight miles 12mo and sum of Bonus miles for each Cluster ID. Color shows sum of Flight miles 12mo.

D.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 emerge?

A> using the random function I have taken 95% data and created the dendrogram using Euclidian distance and wards algorithm. Again I see there is a clear possibility of two clusters for our analysis. And even the tableau report shows that the output what we got above is similar to that of 95%data

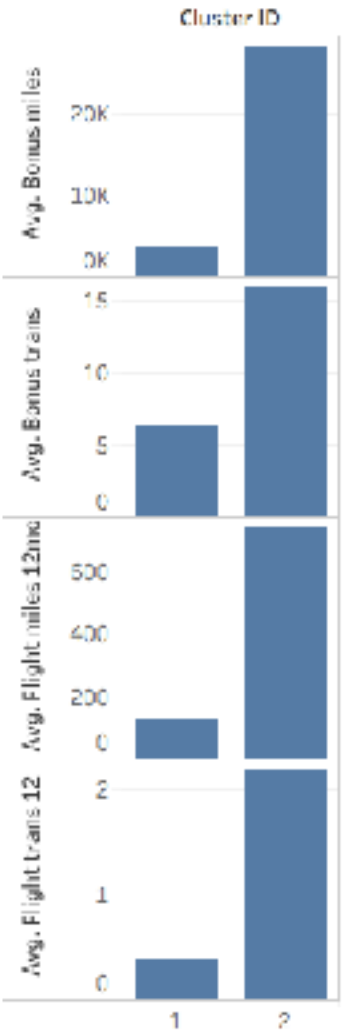


Sheet 5



Average of Flight trans 12, average of Flight miles 12mo, average of Bonus miles and average of Bonus trans for each Fit$Cluster.

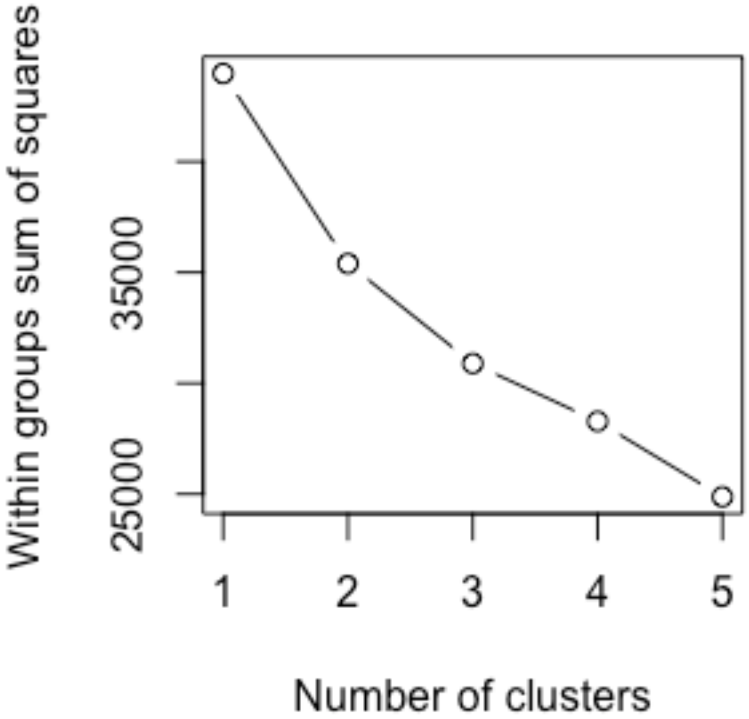
Sheet 4



Average of Bonus miles, average of Bonus trans, average of Flight miles 12mo and average of Flight trans 12 for each Cluster ID.

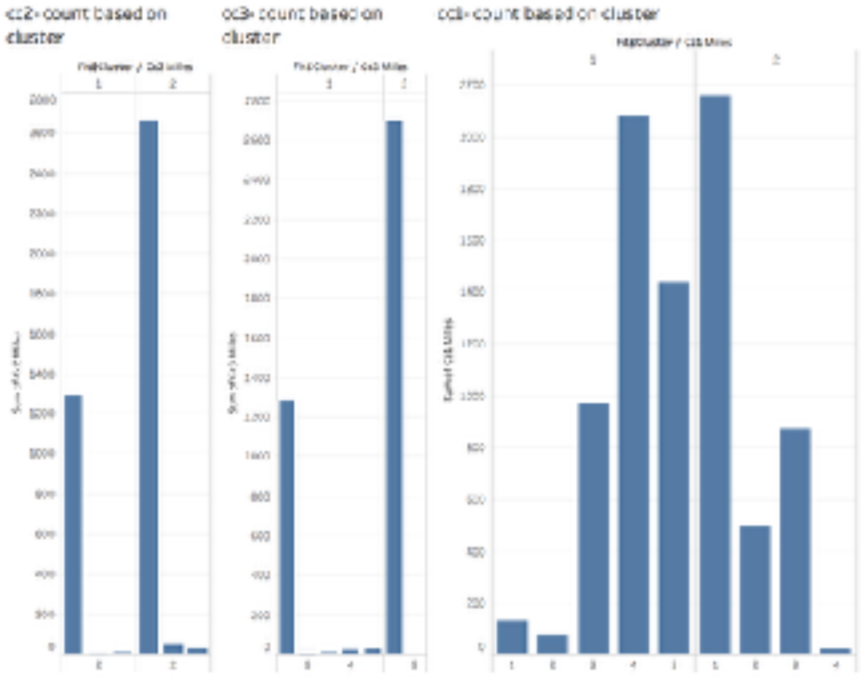
E.e) Cluster all passengers again using k-means clustering. How many clusters do you want to go with? How did you decide on the number of clusters? Explain your choice on the number of clusters.

A. I will choose two clusters to go with. I have decided it using the scree plot and the elbow rule. Elbow rule states that where there is a maximum decrease in legth that should be chosen as number of cluster . From below graph it is evident that it is at 2.

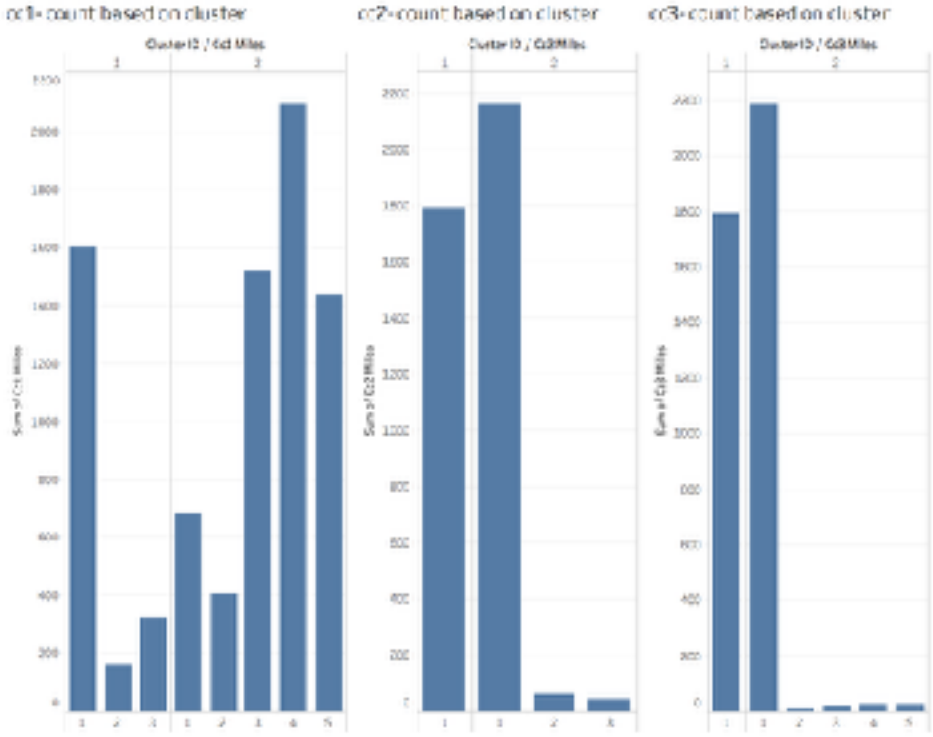


-
-
-
- f) How do the characteristics of the clusters, obtained in Part (e), contrast or validate the finding in Part c above?

- A. Both the clustering technique divides data in two clusters. When I looked on miles and sum of transactions it looks stronger on cluster with frequent flier list but the count is different as we look at the below graph . So I check the cc miles list. KMean cluster looks more organised it has divided the frequent flier credit card , rewards credit card and small business credit card. Means will show a better way to divide the customer for publicising.
KMeans cluster cc miles graph.



**Hierarchial
clustering for credit card miles:**



- **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).**

A. The data clearly shows that cluster-1 are our frequent flier list and second one are people who travel occasionally.

A. Cluster 1 offer: we can provide them with hassle free travel and free food. As they are regular customer they wont go for other websites.

B. Cluster 2 offer: as they travel less frequently we can assume they travel only for occasion so we advertise them with cheaper rates during festive season to attract them.

****

2 A) Enumerate the insights you gathered during your PCA exercise. Please do not clutter your report with too many insignificant insights as it will dilute the value of your other significant findings.

A. After doing PCA and creating cluster with 2 PC components. When looked at the below graph I could see

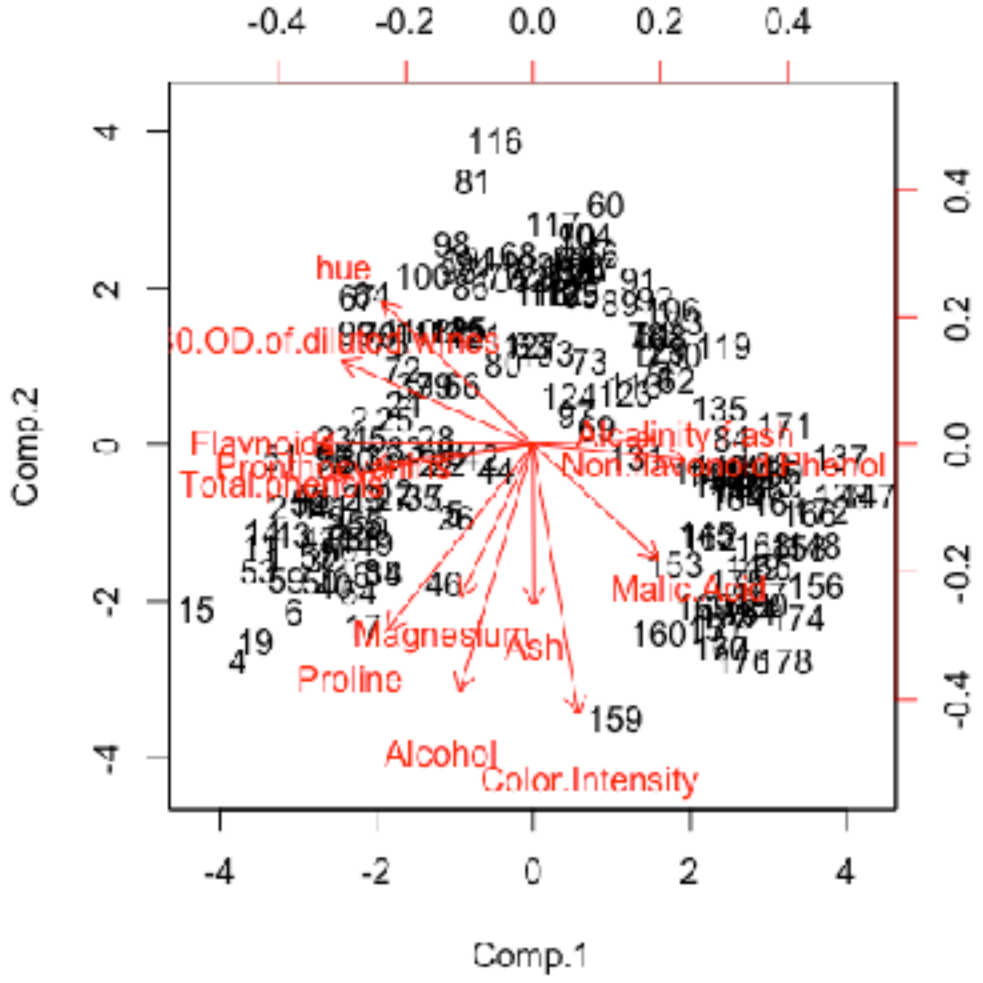
1. Malic Acic, Non Flavournoids phenol and alkalinity of ash is directly related in wine.

2. Flavournoid phenol, total Phenol and OD280/OD328 Hue and proline are directly related.

3.Magnesium, Ash,Alcohol, Color intensity are not much helpful in creating cluster they are more or less same.

B. What are the social and/or business values of those insights, and how the value of those insights can be harnessed—enumerate actionable recommendations for the identified stakeholder in this analysis?

A.
Below is the pilot which shows that biplot of pea analysis of two components of PCA. Clearly if we divide the below graph based on above research right side of the graph is cluster 2 and the left side of the graph is cluster 1.



Below is the graph which represent the 3 stake holders when we look the graph it. Is very clear that

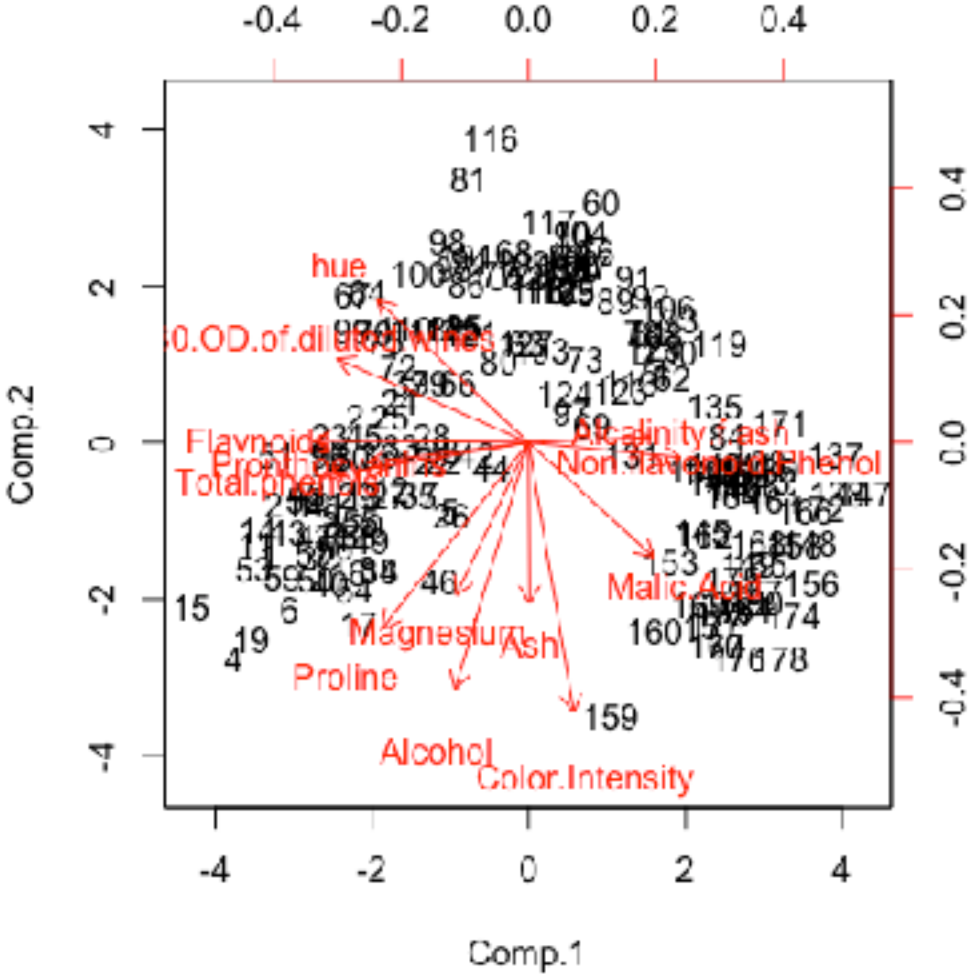
1. Stakeholder 1 belongs to cluster 1 and makes RED WINE
2. Stake holder 2 is spread out in graph so makes both RED and WHITE WINE.
3. Stake holder 3 makes only WHITE WINE.

c. Any more insights you come across during the clustering exercise?

A. When I did cluster analysis using two components and K-Means .

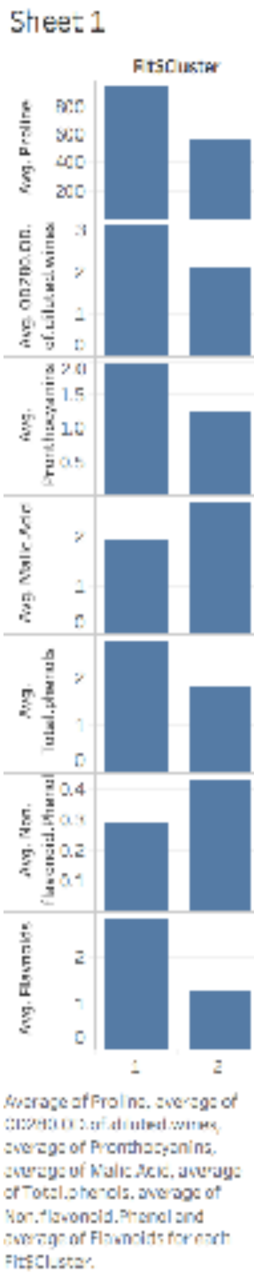
So first I constructed a screeplot graph using the below scaled version.

Screeplot



Looking at the above screeplot I decided to go with two cluster analysis. And checked the average of each column using the two clusters. First let me display the values which are largely different in both the clusters.

The below graph has been created in tableau.



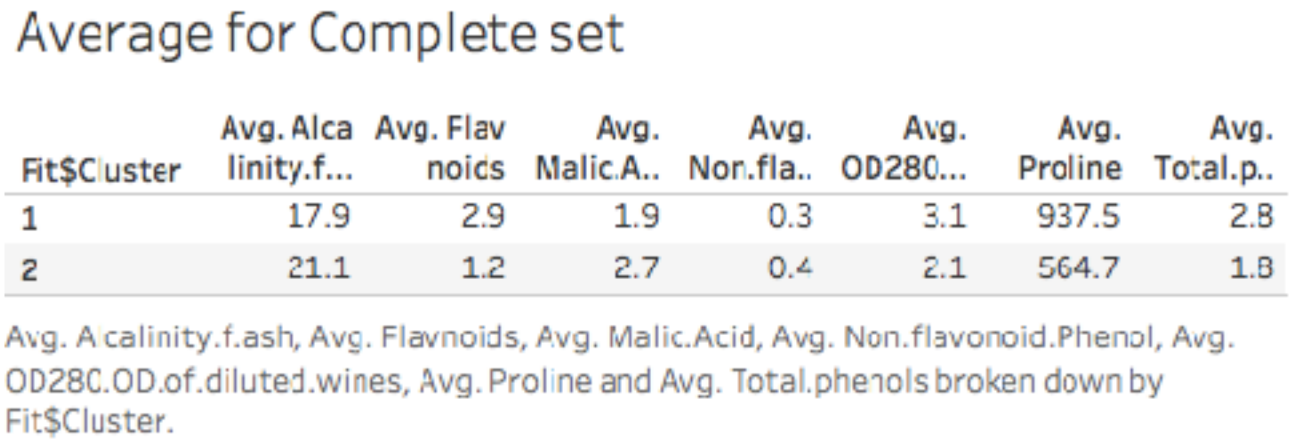
Sheet 1

Average of Proline, average of OD280.OD315.of.diluted.wines, average of Proanthocyanins, average of Malic.Acid, average of Total.phenols, average of Non.Flavonoid.Phenol and average of Flavnoids for each FitSCluster.

The below table represents the difference of content which helped us creating this cluster.

Alcohol content, ash ,Magnesium,color intensity, hue these features were more or less are same in the cluster.

When we look at Both the PCA components average cluster and the complete data set cluster both have same averages and similar cluster creation.



Average for Complete set

Fit$Cluster	Avg. Alcalinity.f...	Avg. Flavnoids	Avg. Malic.A..	Avg. Non.fla..	Avg. OD280...	Avg. Proline	Avg. Total.p..
1	17.9	2.9	1.9	0.3	3.1	937.5	2.8
2	21.1	1.2	2.7	0.4	2.1	564.7	1.8

Avg. Alcalinity.f.ash, Avg. Flavnoids, Avg. Malic.Acid, Avg. Non.flavonoid.Phenol, Avg. OD280.OD.of.diluted.wines, Avg. Proline and Avg. Total.phenols broken down by Fit$Cluster.

D.Are there clearly separable clusters of wines? How many clusters did you go with? How the clusters obtained in part (i) are different from or similar to clusters obtained in part (ii),

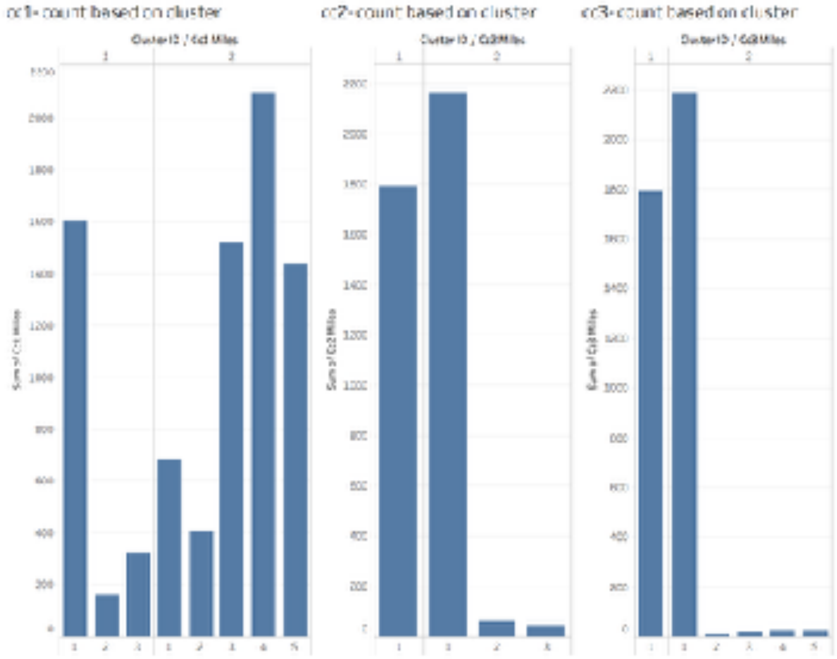
In the end after research on internet I found that the wine which have higher context on Flavonoid phenol are red win and non Flavonoid are White wine.

Names for cluster
Cluster 1 - Red Wine
Cluster 2- White wine

Average using the PCA

Column Name	Cluster 1	Cluster 2	Significance of column
Malic acid	1.9	2.7	Sour flavour
Alcalinity of wine	17.9	21.1	How basic is the wine
phenol	2.8	1.8	It affects taste,color and texture
Flavoids	2.9	1.2	Type of phenol
Non flavour flavonoid	0.3	0.4	Type of phenol
Pronthocyannis	1.95	1.24	Type of phenol
OD280/OD315	3.1191	2.12	Protein measurement
Proline	937.5	564.7	Amino acids

Hierachial cluster on cc miles


cc1= count based on cluster
Cluster ID / Cc1 Miles
Sum of Cc1 Miles
cc2= count based on cluster
Cluster ID / Cc2 Miles
Sum of Cc2 Miles
cc3= count based on cluster
Cluster ID / Cc3 Miles
Sum of Cc3 Miles