Batch 6 - Data Mining 1 - Unsupervised Learning - Assignment 1

Subba Reddy Yeruva, ISB Id: 71610085

06 July 2016

**Note:** The analysis and insights are done in R markdown. The Rpubs link is as below.

The same word document can be found in Rpubs link.

<http://rpubs.com/yeruva_s_reddy/194972>

## 1. Marketing to Frequent Fliers

The file EastWestAirlinesCluster.xls (available on the textbook website <http://dataminingbook.com/>) contains information on 4000 passengers who belong to an airline's frequent flier program. For each passenger the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers.

1.1 Apply hierarchical clustering with Euclidean distance and Ward's method. Make sure to standardize the data first. How many clusters appear?

1.2 What would happen if the data were not standardized?

1.3 Compare the cluster centroids to characterize the different clusters and try to give each cluster a label.

1.4 To check the stability of the 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?

1.5 Use k-means clustering with the number of clusters that you found above in Part (a). Does the same picture emerge? If not, how does it contrast or validate the finding in Part c above?

1.6 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 cluster(s) and the corresponding offer(s).

## 1.1 Number of Clusters

As per the nature of the columns in EastWestAirlinesCluster.xls, The data is mixed data. Hence the categorical variables are converted into binary and then clustering is applied.

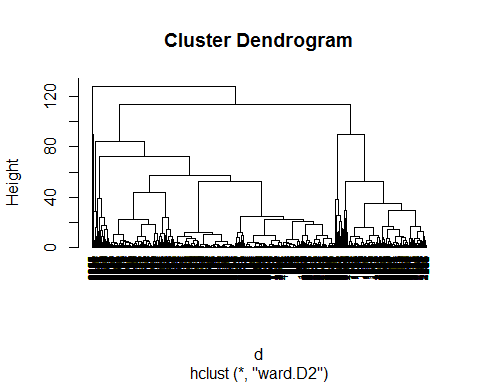
As per the generated dendogram below, if slicing is done at height 100 then there are 3 clusters formed.

As per the dendogram branching is done into 2 parts and left branch alone formed one cluster and 2nd branch(Right Branch) again splitted into two more branches. The two branches of right branch are formed into an cluster. Hence in total 3 clusters with maximum number of customers.

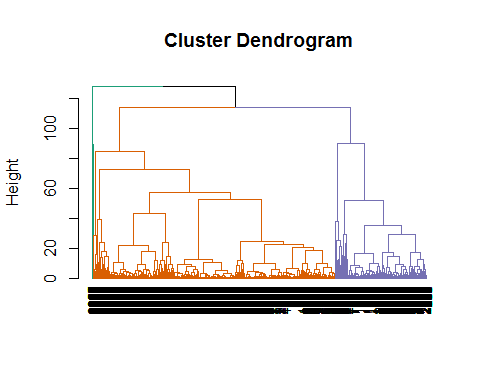
## R code illustrating the procees to create Hierarchical Clustering

## Libraries  
library("dummies")  
library("dendextend")  
library("dendextendRcpp")  
library("gridExtra")  
library("cluster")  
library("factoextra")  
library("MASS")  
library("fpc")  
  
## Set the working directory to where the Excel file is  
setwd('J:\\ISB Business Analytics\\Data Mining\\Data Mining Assignment 1')  
  
## Input file read  
input <- read.csv("EastWestAirlinesClusterCSV.csv",header=TRUE)  
  
## 1. Loading and preparing data  
mydatawd <- input[,2:11]  
  
#Creating Dummy variables for categorical data  
mydata <- dummy.data.frame(mydatawd, names = "cc1\_miles", omit.constants=FALSE )  
mydata <- dummy.data.frame(mydatawd, names = "cc2\_miles", omit.constants=FALSE )  
mydata <- dummy.data.frame(mydatawd, names = "cc3\_miles", omit.constants=FALSE )  
  
# Standardize Data  
my\_data <- scale(mydata)  
  
# 2. Compute dissimilarity matrix  
d <- dist(my\_data, method = "euclidean")  
  
# Hierarchical clustering using Ward's method  
res.hc <- hclust(d, method = "ward.D2" )

## Cluster Dendogram in Ward's Method



## Different clusters with differentation of color



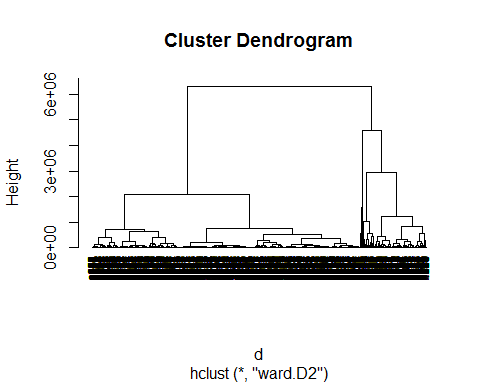
## 1.2 Standardization of Data

Stanardization/Normalization of data is not done then below issues will influence the model.

-> Distance measure will be wrongly clacualted, if all the variables are not with equal weight.

-> Laregest scale dominating the measure

The below dendogarm generated with the given data set without standardization is as below. As per the insights large value units are influencing the small values.



## 1.3 Cluster Centroids

The below are the metrics of cluster centriods.

### Cluster 2 - Label - High Networth frequent fliers

Cluster 2 metrics are leading in progressive way except non-flight bonus transactions.  
These are the segment of customers, who are associated with EastWest Airlines since long time.

### Cluster 3 - Label - Non Frequent travellers

Cluster 3 metrics depicts that the flight miles and flight transactions in last 12 months is zero and their  
non-flight bonus transactions are leading than the other clusters.

### Cluster 1 - Label - Middle class travellers

Cluser 1 metrics depicts that the level of spending is average to cluster 2 and 3.

### Centriod Metrics - Column versus Clusters

## [,1] [,2] [,3]  
## Balance 73547.288828 115682.00 89401.333  
## Qual\_miles 144.071142 295.50 0.000  
## cc1\_miles 2.057114 3.75 3.000  
## cc2\_miles 1.014529 1.00 1.000  
## cc3\_miles 1.009519 3.00 2.000  
## Bonus\_miles 17064.150802 74597.50 47920.000  
## Bonus\_trans 11.576653 32.75 17.000  
## Flight\_miles\_12mo 460.035822 825.00 0.000  
## Flight\_trans\_12 1.373246 2.75 0.000  
## Days\_since\_enroll 4117.809118 5058.75 3863.333

## 1.4 Stability of Clusters Validation

The Process to check for the stability of clusters check. Create different Dendogarms with the sample of 95% data. Here Dend1 and Dend2 are with sample of 95% data.

# Hierarchical clustering using Ward's method  
  
Dend1 <- as.dendrogram(res.hc)  
  
#Random Sample1 with 95% of data  
input2=input[sample(nrow(my\_data),replace=F,size=0.95\*nrow(input)),]  
  
d <- dist(input2, method = "euclidean")  
res2.hc <- hclust(d, method = "ward.D2" )  
Dend2 <- as.dendrogram(res2.hc)  
  
#Random Sample2 with 95% of data  
input3=input[sample(nrow(my\_data),replace=F,size=0.95\*nrow(input)),]  
  
d <- dist(input3, method = "euclidean")  
res3.hc <- hclust(d, method = "ward.D2" )  
Dend3 <- as.dendrogram(res3.hc)

### Global Comparison of Dendograms:

# Global Comparison of Dendograms  
# Total Population Metrics  
Dend1

## 'dendrogram' with 2 branches and 3999 members total, at height 128.0329

# Random Sample 1 Metrics  
Dend2

## 'dendrogram' with 2 branches and 3799 members total, at height 6094226

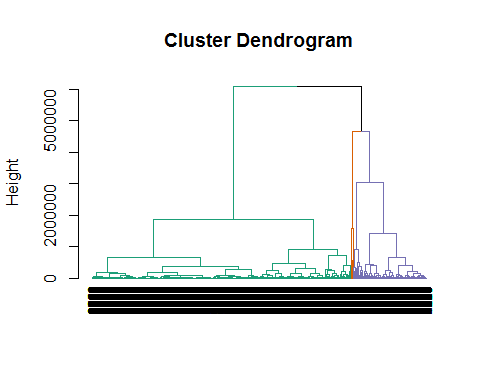
# Random Sample 2 Metrics  
Dend3

## 'dendrogram' with 2 branches and 3799 members total, at height 6548637

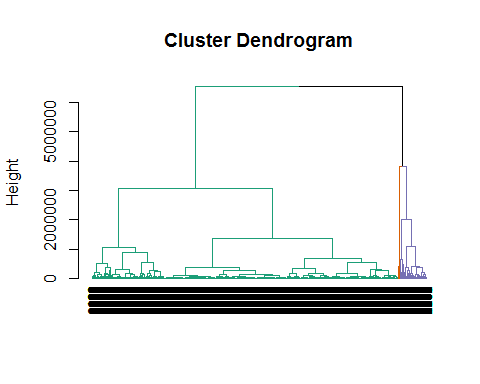
# Comparison of Population, Random sample 1 and Random sample 2 dendogarms.  
all.equal(Dend3, Dend2, Dend1, use.edge.length = TRUE)

## [1] "Difference in branch heights - Mean relative difference: 0.05321841"

#### Sample 1 Dendogarm:



#### Sample 2 Dendogarm:



### Conclusion

From above stastics of dendogram, dendogram comparison and dendogram visualizations, the dendogarams are not same(The Pictures are not same).For every instance of sample, new insights are emerged.

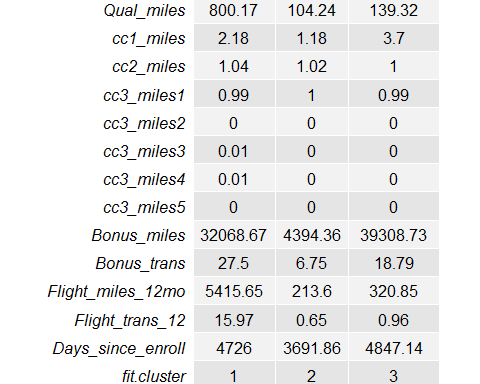
## 1.5 K-Means Clustering

The below R code depicts the process to get the aggregate metrics for Hierarchial clustering and K-Means.

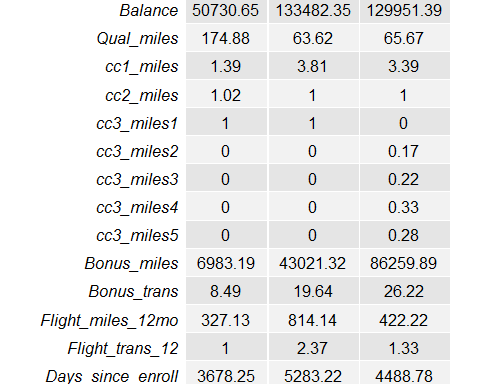
## K-means clustering  
set.seed(123)  
fit <- kmeans(my\_data, 3) # 3 cluster solution  
  
#Aggregation of k-means  
  
mydatak <- data.frame(mydata, fit$cluster) # append cluster membership  
temp <- aggregate(mydatak, by=list(fit$cluster), FUN=mean)  
  
#to find the size of clusters   
ClusterCo <- aggregate(mydatak, by=list(fit$cluster), FUN=sum)   
#to find the cluster size  
d <- transform(ClusterCo, clusterSize = fit.cluster / Group.1)  
d <- transform(d, fit.cluster= fit.cluster/ clusterSize)  
temp$clusterSize <- d$clusterSize  
temp$clusterPCT <- (d$clusterSize\*100)/3999  
# transpose to change from horizontal to vertical  
temp2 <- t(temp)  
  
round\_df <- function(x, digits) {  
 # round all numeric variables  
 # x: data frame   
 # digits: number of digits to round  
 numeric\_columns <- sapply(x, class) == 'numeric'  
 x[numeric\_columns] <- round(x[numeric\_columns], digits)  
 x  
}  
  
temp4 <- round\_df(temp2, 2)  
  
#Hierarchical Aggregate calculations  
  
# Hierarchical clustering using Ward's method  
  
#set.seed(123)  
groups <- cutree(res.hc, k=3) # cut tree into 3 clusters  
membership <-as.matrix(groups)  
membership <- data.frame(membership)  
names(membership) <- c("cluster")  
mydatao <- data.frame(mydata, membership$cluster) # append cluster membership  
  
temp <- aggregate(mydatao, by=list(membership$cluster), FUN=mean)  
  
temp2 <- t(temp)  
  
round\_df <- function(x, digits) {  
 # round all numeric variables  
 # x: data frame   
 # digits: number of digits to round  
 numeric\_columns <- sapply(x, class) == 'numeric'  
 x[numeric\_columns] <- round(x[numeric\_columns], digits)  
 x  
}  
  
temp5 <- round\_df(temp2, 2)

### K-Means Aggregate Metrics with Mean

Along with the primary dataset metrics, each cluster size( clusterSize ) and percentage of cluster size( clusterPCT ) in given population are calculated and tabulated.



### Hierarchical Cluster Aggregate Metrics with Mean



### Conclusion

From above metrics, none of the clusters are matching, Hence the outcomes of K-means cluster and Hierarchial clustering are not same.

## 1.6 Cluster Target and Offers

From the metrics of K-Means Aggregate Metrics with Mean, The below are the target offers to the customers.

### Cluster 2

Business Proportion : 62.64% of customers fall into this cluster. There is potential business growth   
 opportunity, if the passengers are turned into regular travellers.   
Potential Offers :   
 1. If frequent flyer credit card is used then the reward points per travel   
 percentage can increased.   
 2. If the number of travel checkins are more than 5 in an year, Then extra  
 bonus points can be awarded to the customer.

### Cluster 3

Business Proportion : This cluster is second largest with 33.31% of customers.   
Potential Offers :   
 1. If Bonus miles are used with the proportion of 80:20 then the customers can  
 be awarded with 50% off with frequent flyer credit card.

## 2. Wine data from the UCI machine learning repository

Step 1: Download the Wine data from the UCI machine learning repository (<http://archive.ics.uci.edu/ml/datasets/Wine>)

Step 2: Do a Principal Components Analysis (PCA) on the data. Please include (copy-paste) the relevant software outputs in your submission while answering the following questions.

2.1 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)

2.2 What are the social and business values of those insights, and how the value of those insights can be harnessed?

Step 3: Do a cluster analysis using (i) all chemical measurements (ii) using two most significant PC scores. Please include (copy-paste) the relevant software outputs in your submission while answering the following questions.

2.3 Any more insights you come across during the clustering exercise?

2.4 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), qualitatively?

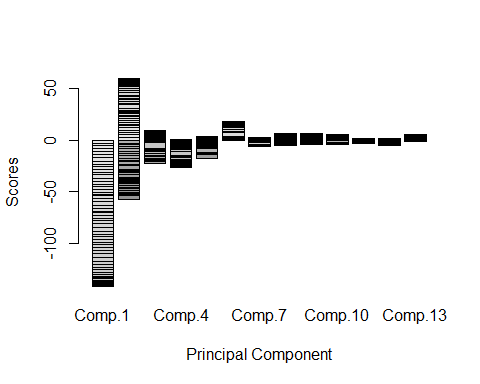
2.5 Could you suggest a subset of the chemical measurements that can separate wines more distinctly? How did you go about choosing that subset? How do the rest of the measurements that were not included while clustering, vary across those clusters?

## 2.1 PCA Insights

### R Code Depicting PCA Process

setwd('J:\\ISB Business Analytics\\Data Mining\\Data Mining Assignment 1\\WIne DataSet')  
## Principal Component Analysis  
  
input <- read.csv("WineData.csv",header=TRUE)  
  
mydata <- input[1:178,2:14]  
  
# Pricipal Components Analysis  
# entering raw data and extracting PCs   
# from the correlation matrix   
wine.pca <- princomp(mydata, cor=TRUE)

### Barplot Illustrating different pricipal components scores



### Variance Accounted for Pricipal Components

summary(wine.pca) # print variance accounted for

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 2.1692972 1.5801816 1.2025273 0.9586313 0.92370351  
## Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294  
## Cumulative Proportion 0.3619885 0.5540634 0.6652997 0.7359900 0.80162293  
## Comp.6 Comp.7 Comp.8 Comp.9  
## Standard deviation 0.80103498 0.74231281 0.59033665 0.53747553  
## Proportion of Variance 0.04935823 0.04238679 0.02680749 0.02222153  
## Cumulative Proportion 0.85098116 0.89336795 0.92017544 0.94239698  
## Comp.10 Comp.11 Comp.12 Comp.13  
## Standard deviation 0.50090167 0.47517222 0.41081655 0.321524394  
## Proportion of Variance 0.01930019 0.01736836 0.01298233 0.007952149  
## Cumulative Proportion 0.96169717 0.97906553 0.99204785 1.000000000

### Insights

As per the barplot above the principal components 1,2,3,4 and 5 are highly negatively correlated.

As per the summary statistics, the cumulative proportion of variance till principal componet is 80%.So 5 principal components can be considered for the analysis.

As per the standrad deviation metrics above the principal componets 1 through 5 are having above and equal to 1. It indicate the better predictors.

## 2.2 Social and Business Value Insights

The social and business insights of each principal component can drwan with metrics of loading with help of correlation effect of each component.

loadings(wine.pca) # loadings

##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## Alcohol -0.144 -0.484 -0.207 -0.266 0.214   
## Malic\_Acid 0.245 -0.225 0.537 0.537 -0.421  
## Ash -0.316 0.626 -0.214 -0.143 0.154 0.149  
## Ash\_Alcalinity 0.239 0.612 -0.101 0.287  
## Magnesium -0.142 -0.300 0.131 -0.352 0.727 -0.323  
## Total\_Phenols -0.395 0.146 0.198 -0.149   
## Flavanoids -0.423 0.151 0.152 -0.109   
## Nonflavanoid\_Phenols 0.299 0.170 -0.203 -0.501 -0.259 -0.595  
## Proanthocyanins -0.313 0.149 0.399 0.137 -0.534 -0.372  
## Color\_Intensity -0.530 -0.137 -0.419 0.228  
## Hue -0.297 0.279 -0.428 -0.174 0.106 -0.232  
## OD280\_OD315 -0.376 0.164 0.166 0.184 -0.101 0.266   
## Proline -0.287 -0.365 -0.127 -0.232 -0.158 0.120   
## Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13  
## Alcohol 0.396 -0.509 0.212 -0.226 -0.266   
## Malic\_Acid -0.309 0.122   
## Ash -0.170 0.308 -0.499 -0.141   
## Ash\_Alcalinity 0.428 -0.200 0.479   
## Magnesium -0.156 -0.271   
## Total\_Phenols -0.406 -0.286 -0.320 0.304 -0.304 -0.464   
## Flavanoids -0.187 -0.163 0.832   
## Nonflavanoid\_Phenols -0.233 -0.196 0.216 0.117 0.114   
## Proanthocyanins 0.368 0.209 0.134 -0.237 -0.117   
## Color\_Intensity -0.291 0.604   
## Hue 0.437 -0.522 0.259   
## OD280\_OD315 -0.137 0.524 0.601 -0.157   
## Proline 0.120 0.576 0.162 0.539   
##   
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8  
## SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000  
## Proportion Var 0.077 0.077 0.077 0.077 0.077 0.077 0.077 0.077  
## Cumulative Var 0.077 0.154 0.231 0.308 0.385 0.462 0.538 0.615  
## Comp.9 Comp.10 Comp.11 Comp.12 Comp.13  
## SS loadings 1.000 1.000 1.000 1.000 1.000  
## Proportion Var 0.077 0.077 0.077 0.077 0.077  
## Cumulative Var 0.692 0.769 0.846 0.923 1.000

### Food additives rich with advantages

-> The first principal component is highly negative correlated with Flavanoids, Total\_Phenols.  
  
-> The third principal component is positively correlated with Ash, Ash Alcalinity which are good for health.

### Health Harzardous composition of chemicals

-> The principal component 2 is highly negative correlated with alcohol, color intensity and etc.  
-> The principal component 4 is highly positively correlated with alchocal whis is harzardous to health.

## 2.3 Clustering Insights on Wine Dataset

### R Code for all measurements using k-means clustering

setwd('J:\\ISB Business Analytics\\Data Mining\\Data Mining Assignment 1\\WIne DataSet')  
inputk1 <- read.csv("WineData.csv",header=TRUE)  
winek <- inputk1[1:178,2:14]  
wineks <- scale(winek)  
wine.km <- kmeans(wineks, 3)

### Size of clusters

#Size of clusters  
wine.km$size

## [1] 65 51 62

### Metrics of clusters

#Size of clusters  
wine.km$centers

## Alcohol Malic\_Acid Ash Ash\_Alcalinity Magnesium  
## 1 -0.9234669 -0.3929331 -0.4931257 0.1701220 -0.49032869  
## 2 0.1644436 0.8690954 0.1863726 0.5228924 -0.07526047  
## 3 0.8328826 -0.3029551 0.3636801 -0.6084749 0.57596208  
## Total\_Phenols Flavanoids Nonflavanoid\_Phenols Proanthocyanins  
## 1 -0.07576891 0.02075402 -0.03343924 0.05810161  
## 2 -0.97657548 -1.21182921 0.72402116 -0.77751312  
## 3 0.88274724 0.97506900 -0.56050853 0.57865427  
## Color\_Intensity Hue OD280\_OD315 Proline  
## 1 -0.8993770 0.4605046 0.2700025 -0.7517257  
## 2 0.9388902 -1.1615122 -1.2887761 -0.4059428  
## 3 0.1705823 0.4726504 0.7770551 1.1220202

#aggregate(inputk1[-1], by=list(cluster=wine.km$cluster), mean)

### Insights

From Below plot the insights are as below.

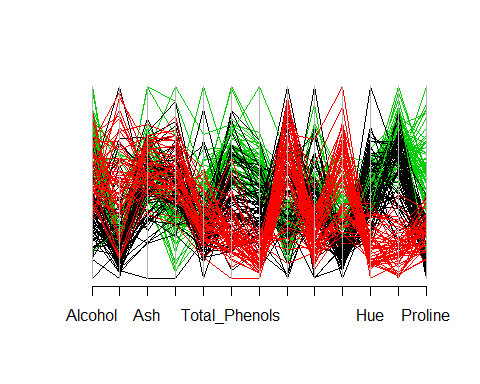
--> The Cluster representing Green colour, the composition of chemicals Alcohol,OD280\_OD315 , Proline and Total\_Phenols and Flavanoids is more compared to other chemical proportions.

These metrics indicate that the composition of checmicals is harzardous to health.

--> The Cluster representing black colour, the composition of chemicals Flavanoids, Hue, OD280\_OD315 and Proanthocyanins is less compared to other chemical proportions.

--> The Cluster representing red colour, the composition of chemicals Alcohol, Malic\_Acid, Magnesium and Color\_Intensity is less compared to other chemical proportions.

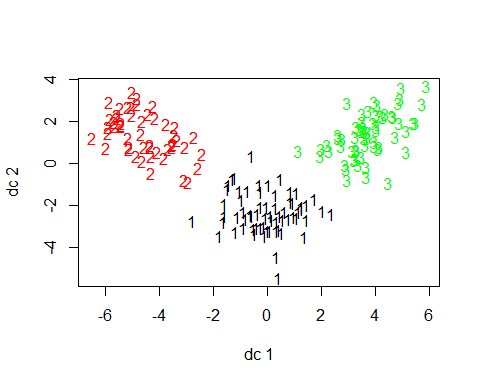
The metrics indicate that the compisition of checmicals is good for health.



## 2.4 Seperabale Clusters or not and other metrics

### Seperabale Clusters

As per the below plot the clusters are clearly seperable as there is no interminging of the clusters.



### Number of Clusters

The optimum number of clusters are 3 from above metrics. The optimum number of clusters are decided based on   
NBClust function results.

## 2.5 Chemical Measurements and Conclusion

### Subset of chemicals

From above statistics the Seven chemical compositions Alchohol, Magnesium, Flavanoids, Color\_Intensity, Hue,  
OD280\_OD315 and Proline will seperate the wines more distinctly.

### Decision Methodology

Based on the metrics of PCA and K-means, Identified the chemicals which are affecting the clusters to the   
extreme either low or high.

### Otherthan Suset of chemicals

The other chemicals which are not part of subset of chemicals are consistent across the clusters in terms of  
chemical composition.

## 3 References

1. Journal of Statistical Software

<https://www.jstatsoft.org/index>

1. Dendogarms <https://cran.r-project.org/web/packages/dendextend/vignettes/introduction.html>

<https://rpubs.com/gaston/dendrograms>

<http://www.sthda.com/english/wiki/static-and-interactive-heatmap-in-r-unsupervised-machine-learning#at_pco=smlwn-1.0&at_si=577bd1ceccc50574&at_ab=per-2&at_pos=0&at_tot=1>

<http://www.sthda.com/english/wiki/cluster-analysis-in-r-unsupervised-machine-learning#at_pco=smlwn-1.0&at_si=577bd0960cfd2397&at_ab=per-2&at_pos=0&at_tot=1>

<http://www.sthda.com/english/rpkgs/factoextra/fviz_dend.html>

1. R Mark Down Cheat sheet <http://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf>

<http://www.jacolienvanrij.com/Tutorials/tutorialMarkdown.html>

1. K-Means <https://en.wikipedia.org/wiki/K-means_clustering>

Wine DataSet: <https://datayo.wordpress.com/2015/05/06/using-k-means-to-cluster-wine-dataset/>

R-in Action: <http://www.r-bloggers.com/k-means-clustering-from-r-in-action/>

<https://cran.r-project.org/web/views/Cluster.html>

R- Useful <https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means>

1. Data Mining Book <https://mineracaodedados.files.wordpress.com/2012/07/data-mining-in-excel.pdf>
2. Science direct <http://www.sciencedirect.com/science/article/pii/S0020025507002319>
3. PCA <https://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_principal_components_analysis.pdf>

<http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/>

<https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html>

<http://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/>

PCA R:- <https://stat.ethz.ch/~mmarloes/teaching/fall08/Rscript-PCA.R>

imp: <https://www.ime.usp.br/~pavan/pdf/MAE0330-PCA-R-2013>

1. Nice Plots <http://www.r-bloggers.com/using-r-to-replicate-common-spss-multiple-regression-output/>
2. NBClust Paper

<https://www.google.co.in/url?sa=t&rct=j&q>=&esrc=s&source=web&cd=10&cad=rja&uact=8&ved=0ahUKEwj-7N6HleLNAhWDt48KHcuZBCkQFghYMAk&url=https%3A%2F%2Fwww.jstatsoft.org%2Farticle%2Fview%2Fv061i06%2Fv61i06.pdf&usg=AFQjCNHAkhpWbWNxKheZ0Z\_itE-KdjHfvw&bvm=bv.126130881,d.c2I

1. For PDF <http://miktex.org/download>