Data 630 – Fall 2018

Assignment 5 – K-Means Clustering Analysis of Red Wine

Kenneth Lulie

Dr. Gates

University of Maryland University College

November 17, 2018

**Cluster Analysis of Red Wine using the Vinho Verde Dataset**

**Introduction**

The United States wine market was estimated to be worth over 60 billion USD in 2017 (Grand View Research, 2018). This market is comprised of 41 billion USD worth of domestic wine sales, and 20 billion USD of imported. There are four major types of wine: red, white, rose and champagne (Primermagazine, 2016). The type and color of wine is highly related to the taste, as during the fermenting process, the skins of the grapes give the color to the wine. The skins also determine the Tannin level which gives each type of wine its characteristic flavor. Tannins are a naturally occurring substance in grapes often described as bitter, causing a dry and pucker feeling in the mouth. Generally, the redder the wine the more tannins are in it as red wines have the grape skins kept in during the fermenting process, while white wines do not. Additionally, rose wines are pinkish in color because it is allowed to stay in contact with red grape skins for a limited amount of time. Champagne wine has significant levels of carbonation, either naturally during the fermenting process or through carbon dioxide injections. It is also sometimes called sparkling wine. Other factors that affect wine quality include the region where the grapes were harvested, the weather while those grapes were growing, the process used to ferment the wine, the age of the wine, and the type of wood in the barrels used to age the wine.

Despite the immense amount of money spent on wine every year, it has proven difficult to objectively determine which wines are the best. The famous Judgment of Paris of 1976 was a wine competition where top quality wines from France and California were judged in a blind taste test (Godoy, 2016). A California wine scored best in each category, shocking the wine industry as France was considered to be the best wine producing area in the world. This highlights the subjective nature of judging quality of wine, with region and price often used by buyers as a proxy for quality. A more objective analysis of wine based on chemical analysis rather than region, price, or general public opinion could prove to be a boon to developing wine vintners without an expensive established brand. An understanding based on chemical analysis could potentially allow them to improve their wine in a more scientific way by comparing the chemical testing results to the perceived quality of the wine. It could also assist them in developing a standardized method of quality control or even helping wine buyers pick their favorite wine the way that International Bitter Units have assisted beer drinkers in picking beers they will enjoy.

In this analysis, the impact of chemical measurements such as pH, sulphates, and alcohol will be used to create groupings of wines. The specific aim of this analysis is to use a K-Means clustering algorithm to discover clusters of wine for future analysis and investigation. These clusters will also be compared to the quality ratings provided for these wines to ensure that the clusters produced appear to be meaningful. The model will be tuned to different numbers of clusters to attempt to increase clustering relevance based on a holistic evaluation. Emphasis will be placed on reviewing centroids of clusters with high proportions of either low or high quality wine.

**Analysis and Model Demonstration**

**Subsection: Data Information, Cleaning, and Preprocessing**

The data used in this analysis comes from the UMUC provided list of approved datasets, originally provided by Cortez et al in 2009. The dataset is that of 1599 red variants of Portuguese Vinho Verde wine. The data includes a classification showing the quality each wine was ranked from 3 to 8.

The analysis of this study was implemented using Rstudio version 3.5.1. There were 1599 wines in the dataset provided. The data as provided contained 12 variables, specifically fixed\_acidity, volatile\_acidity, citric\_acid, residual\_sugar, chlorides, free\_sulfur\_dioxide, totalsulfur\_dioxide, density, pH, sulphates, alcohol and quality.

The data was reviewed for missing values and unique identifiers. No missing values or unique identifiers were found. The data was also reviewed outliers or other possible errors in the data. Some outliers were found, but they were judged to likely be accurate measurements and were kept in due to the number of observations in the dataset as well as the limited impact of outliers on the k-means clustering algorithm. The data was also reviewed for data types to ensure that the data was of the correct type. All variables other than quality were continuous variables and thus were scaled to prepare the data for the K-Means clustering algorithm. This ensures that each variable will be of equal importance to the algorithm. Additionally, as the Quality variable was a numeric variable with 6 values, it was transformed to a factorial variable. To assist in the evaluation of the results of the clustering algorithm, the data was also transformed into only three values, with wine rated at quality between 3-5 as low quality, 6 as medium quality, and 7 or 8 as high quality.

As the purpose of the analysis to create a model to assist in understanding which combination of variables leads to lower or higher perceived quality, all variables were retained in the dataset. Analysis of which variables impacted perceived quality the most was deemed out of scope for this analysis, but would be a natural study for a next step. The quality variable was removed from the dataset fed into the algorithm, but was retained for evaluation of the results.

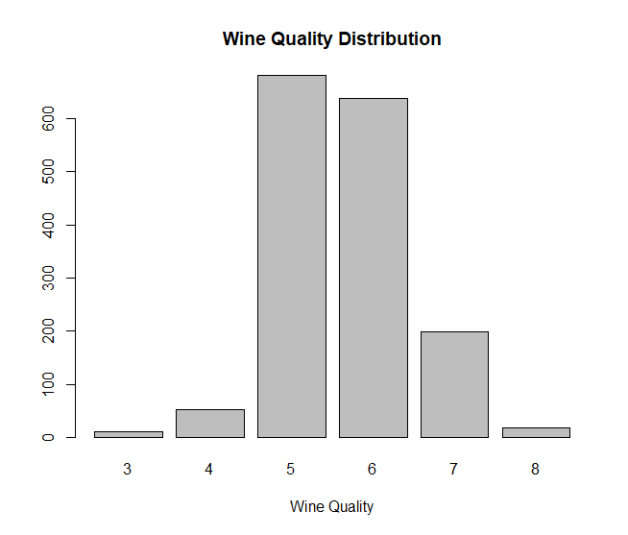
A descriptive analysis including mean, median, 1st and 3rd quartiles, min and max of the chemical measurement variables are provided in **Table 1**. Alcohol ranges from 8.4 to 14.9 with the mean at 10.42 and the median at 10.2 indicating the 14.9 score is an outlier. pH ranges from 2.74 to 4.01 showing that all the wines are fairly acidic. Density ranges from .9901 to 1.0037 indicating a very tight range and therefore unlikely to be an informative variable. The remaining chemical measurements appear to have

Figure 1. Distribution of Wine Quality before Data Preprocessing

Table 1. Distribution of Numeric Variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Fixed Acidity | | Volatile Acidity | | Citric Acid | | Residual Sugar | |
| Min | 4.6 | Min | 0.12 | Min | 0 | Min | 0.9 |
| 1st Qu. | 7.1 | 1st Qu. | 0.39 | 1st Qu. | 0.09 | 1st Qu. | 1.9 |
| Median | 7.9 | Median | 0.52 | Median | 0.26 | Median | 2.2 |
| Mean | 8.32 | Mean | 0.5278 | Mean | 0.27 | Mean | 2.539 |
| 3rd Qu. | 9.2 | 3rd Qu. | 0.64 | 3rd Qu. | 0.42 | 3rd Qu. | 2.6 |
| Max. | 15.9 | Max. | 1.58 | Max. | 1 | Max. | 15.5 |
| Chlorides | | Free Sulfur Dioxide | | Total Sulfur Dioxide | | Density | |
| Min | 0.012 | Min | 1 | Min | 6 | Min | 0.99 |
| 1st Qu. | 0.07 | 1st Qu. | 7 | 1st Qu. | 22 | 1st Qu. | 0.996 |
| Median | 0.079 | Median | 14 | Median | 38 | Median | 0.997 |
| Mean | 0.08747 | Mean | 15.87 | Mean | 46.5 | Mean | 0.997 |
| 3rd Qu. | 0.09 | 3rd Qu. | 21 | 3rd Qu. | 62 | 3rd Qu. | 0.998 |
| Max. | 0.611 | Max. | 72 | Max. | 289 | Max. | 1.004 |
| pH | | Sulphates | | Alcohol | | | |
| Min | 2.74 | Min | 0.33 | Min | | 8.4 | |
| 1st Qu. | 3.21 | 1st Qu. | 0.55 | 1st Qu. | | 9.5 | |
| Median | 3.31 | Median | 0.62 | Median | | 10.2 | |
| Mean | 3.311 | Mean | 0.6581 | Mean | | 10.42 | |
| 3rd Qu. | 3.4 | 3rd Qu. | 0.73 | 3rd Qu. | | 11.1 | |
| Max. | 4.01 | Max. | 2 | Max. | | 14.9 | |

sufficient range and distribution to be of use in clustering. A bar graph showing the distribution of quality shows that the majority of wine qualities are of either low or medium quality with only 217 wines scoring a high quality score of 7 or 8 **Figure 1.** More wines are scored as lower quality with 744 wines, than medium quality with 638 wines.

**Subsection: Analysis and Model Methods**

The preprocessed data was used in this analysis with the K-Means clustering algorithm partitioning the data into unsupervised subsets or clusters. There are many different types of clustering algorithms with different methods for building and evaluating the clusters created (Han, 2011). Different clustering methods may produce different results. However, almost all clustering algorithms are examples of learning by observation (unsupervised learning) instead of learning by examples (supervised learning). Clustering algorithms are useful when the focus is on discovering new knowledge of the relationships and groupings in the data, as opposed to supervised learning when the intent is typically to build a model to classify data into preexisting groups. The four major methods of clustering are Partitioning, Hierarchical, Density, and Grid based methods.

The K-Means Neural Net algorithm is a centroid-based partitioning clustering algorithm (Han, 2011). Partitioning clustering methods distribute the objects of a dataset in an arbitrary number of clusters denoted by the variable K. It uses the centroid, or center point of a cluster to represent each cluster, and the objective function of the algorithm is to assess the quality of partitioning so that objects within one cluster are similar to each other but dissimilar to objects in other clusters. It does this by comparing the mean of each cluster for each observation to the mean of that cluster. Specifically, for this analysis, for each object in each cluster, the Euclidian distance is measured from that object to its cluster center or centroid for each variable, that distance is squared, and the distances are then summed. The algorithm attempts to minimize the total sum of squares and thereby form clusters with objects similar to each other, yet different from objects in other clusters. As the method uses measures of distance, it is important to scale data before clustering so that data points with larger numbers will not have undue influence over the measurements.

The analysis reviewed the results of the K-Means clustering model using three values of K, with the ‘kmeans’ algorithm in R from the library “Cluster”. Each model used every observation in the dataset provided, with all variables used except for the ‘quality’ variable. Each model was measured on Within Cluster Sum of Squares, Between Cluster Sum of Squares divided by Total Sum of Squares, and the proportion and number of the wines within categorized as low, medium, or high quality. Within Cluster Sum of Squares is a measurement of the variability of the observations within each cluster, with a small number generally indicating a more compact cluster. Between Cluster Sum of Squares divided by Total Sum of Squares is a measure of the total variance in the dataset explained by the clustering. The centroids of the most interesting clusters, i.e. clusters with high proportions of low or high quality wine, were also examined to investigate differences in the centroid chemical measurements.

The first model was run on all observations using all remaining variables in the dataset, fixed\_acidity, volatile\_acidity, citric\_acid, residual\_sugar, chlorides, free\_sulfur\_dioxide, totalsulfur\_dioxide, density, pH, sulphates, alcohol. The clustering was created using the kmeans method from the Cluster library with K set to 3, number of random starts or ‘nstarts’ set to 10, and with the random seed set to 1234. The second model was trained with the same parameters, with the exception of k being set to 5. The third model was trained with the same parameters, with the exception of k being to 7.

**Results**

The first clustering model produced 3 clusters with a total of 506, 391, and 702 wines in each cluster **Table 2**. A visualization of the clustering was also created **(Figure 2)**. The centroid measurements are shown in **Appendix 1.** The model had 3 iterations while generating, and had a Between Sum of Squares/Total Sum of Squares measurement of 26.9%. The 3 clusters produced did not show high concentrations of any one specific quality rating in any of the groups, no clusters were mostly empty, and the Between Sum of Squares/Total Sum of Squares % was relatively low indicating that the clusters were not very dense and therefore more clusters may yield more insights.

Table 2. K-Means Model One – 3 Clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K-Means Model One - 3 Clusters | | | | | |
|  | Within Cluster Sum of Squares | Low | Medium | High | Total |
| Cluster 1 | 5,609.05 | 119 | 231 | 156 | 506 |
| Cluster 2 | 3,751.46 | 276 | 104 | 11 | 391 |
| Cluster 3 | 4,666.23 | 349 | 303 | 50 | 702 |
| Between Sum of Squares/Total Sum of Squares | | 26.90% | Model Iterations | | 3 |

The second clustering model produced 5 clusters with a total of 30, 322, 336, 361 and 550 wines in each cluster **Table 3**. A visualization of the clustering was also created **(Figure 3)**. The centroid measurements are shown in **Appendix 2.** The model had 4 iterations while generating, and had a Between Sum of Squares/Total Sum of Squares measurement of 40.8% showing that the clusters have become more dense. The 5 clusters produced showed more concentration in quality versus the 3 clusters previously produced indicating potentially more interesting clusters to review. Cluster 1 had only 30 observations, indicating that the number of clusters is starting to approach diminishing returns.

Table 3. K-Means Model Two - 5 Clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K-Means Model Two - 5 Clusters | | | | | |
|  | Within Cluster Sum of Squares | low | medium | high | Total |
| Cluster 1 | 494.99 | 20 | 9 | 1 | 30 |
| Cluster 2 | 2,899.01 | 237 | 78 | 7 | 322 |
| Cluster 3 | 2,482.43 | 30 | 191 | 115 | 336 |
| Cluster 4 | 2,711.53 | 95 | 180 | 86 | 361 |
| Cluster 5 | 2,761.95 | 362 | 180 | 8 | 550 |
| Between Sum of Squares/Total Sum of Squares | | 40.80% | Model Iterations | | 4 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K-Means Model Three- 7 Clusters | | | | | |
|  | Within Cluster Sum of Squares | low | medium | high | Total |
| Cluster 1 | 2,259.16 | 329 | 145 | 4 | 478 |
| Cluster 2 | 1,318.37 | 37 | 133 | 41 | 211 |
| Cluster 3 | 1,524.19 | 12 | 131 | 129 | 272 |
| Cluster 4 | 1,918.22 | 225 | 96 | 6 | 327 |
| Cluster 5 | 791.25 | 29 | 12 | 5 | 46 |
| Cluster 6 | 478.21 | 19 | 9 | 1 | 29 |
| Cluster 7 | 1,528.07 | 93 | 112 | 31 | 236 |
| Between Sum of Squares/Total Sum of Squares | | 48.80% | Model Iterations | | 5 |

The third clustering model produced 7 clusters with a total of 478, 211, 272, 327, 46, 29 and 236 wines in each cluster **Table 3**. A visualization of the clustering was also created **(Figure 4)**. The centroid measurements are shown in **Appendix 3.** The model had 5 iterations while generating, and had a Between Sum of Squares/Total Sum of Squares measurement of 48.8% showing minor improvement versus the previous model. The 7 clusters produced showed more concentration in quality versus the 5 clusters previously produced more interesting clusters to review, but two clusters had less than 50 observations each. The increase of relatively small clusters along with only the modest improvement in Sum of Squares from the previous model shows that diminishing returns are being reached at a level of 7 clusters and more clusters are unlikely to yield additional insights.

Of the three models, the third model with 7 clusters appeared to have the most interesting clusters to review based on proportions of low and high quality. Clusters 3 and 4 of this model were chosen for review of centroid measurements **Table 5**. Cluster 3 had the highest proportion high quality wine of any cluster produced with 47.4% of wine being high quality, 48.1% as medium quality, and the remaining 4.4% of low quality. Cluster 4 had the lowest quality proportion of wine with 68.8% low quality, 29.3% medium quality, and the remaining 1.8% as high quality. From the review of these clusters we can see the significant differences between the centroids are that higher quality wine has lower Volatile Acidity, higher Citric Acid, lower Free Sulfur Dioxide, lower Total Sulfur Dioxide and higher Alcohol than lower quality wines.

Table 5. Centroid Measurements of Low and High Quality Clusters

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| K Means Model 3 - Cluster 3 'Low Quality' and Cluster 4 'High Quality' Centroid Measurement Comparison | | | | | | | |
| Fixed Acidity | | Volatile Acidity | | Citric Acid | | Residual Sugar | |
| High Quality | 8.792 | High Quality | 0.349 | High Quality | 0.4182 | High Quality | 2.398 |
| Low Quality | 8.118 | Low Quality | 0.532 | Low Quality | 0.2798 | Low Quality | 2.388 |
| Chlorides | | Free Sulfur Dioxide | | Total Sulfur Dioxide | | Density | |
| High Quality | 0.07607 | High Quality | 12.35 | High Quality | 28.91 | High Quality | 0.9958 |
| Low Quality | 0.08555 | Low Quality | 26.21 | Low Quality | 85.87 | Low Quality | 0.9971 |
| pH | | Sulphates | | Alcohol | | | |
| High Quality | 3.271 | High Quality | 0.737 | High Quality | | 11.51 | |
| Low Quality | 3.296 | Low Quality | 0.6252 | Low Quality | | 9.798 | |

**Conclusion**

This analysis showed that the K-Means clustering algorithm combined with the variables provided could be used to potentially identify unknown clusters that could yield interesting insights. Additionally, it showed that there were clusters produced to investigate that could be of interest to vintners trying to improve the quality of their wine. The analysis also showed a potential relationship between lower Volatile Acidity, higher Citric Acid, lower Free Sulfur Dioxide, lower Total Sulfur Dioxide and higher Alcohol to higher quality wine. This approach could be used and refined as well to create clusters of wine for different purposes such as suggesting wines similar to other wines a consumer has enjoyed.

Future studies could improve on the limitations of this analysis. The dataset used was rather robust at 1,599 observations but only covered one type of wine grown in one region. Any attempt to generalize the information would need data from different regions. Second, the dataset did not have any information as to how the wine was created which could yield more insights. Third, no information was provided on how the quality of the wine was judged which could also provide more context for investigating the relationships discovered.

**References**

Han, Kamber, and Pei (2011). Data Mining: Concepts and Techniques, Third Edition Retrieved September 14, 2018 from http://hanj.cs.illinois.edu/cs412/bk3/01.pdf

U.S. Wine Market Size, Share & Trends Analysis Report, By Product (Table Wine, Dessert Wine, Sparkling Wine), By Distribution (On-trade, Off-trade), Competitive Landscape, And Segment Forecasts, 2018 - 2025. (n.d.). Retrieved from https://www.grandviewresearch.com/industry-analysis/us-wine-market

Learn About Wine: An Easy Explanation of Wine Types. (2017, March 30). Retrieved from https://www.primermagazine.com/2011/learn/an-easy-explanation-of-wine-types

Godoy, M. (2016, May 24). The Judgment Of Paris: The Blind Taste Test That Decanted The Wine World. Retrieved from https://www.npr.org/sections/thesalt/2016/05/24/479163882/the-judgment-of-paris-the-blind-taste-test-that-decanted-the-wine-world

Figure 2. Plot of K-Means Model One - 3 Clusters

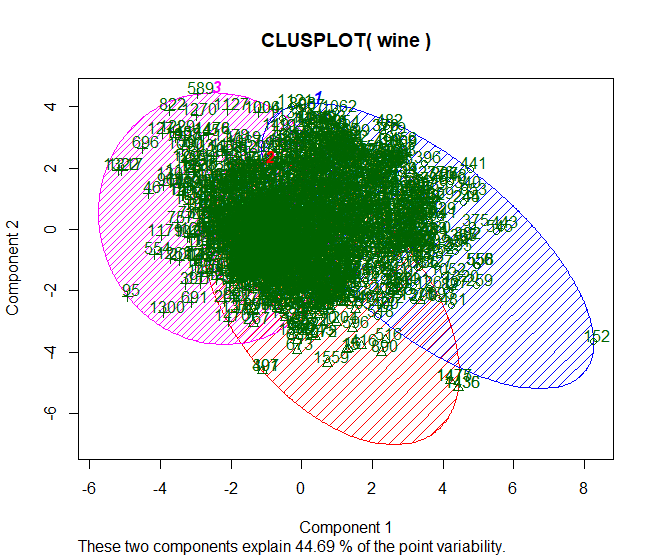


Figure 3. Plot of K-Means Model Two - 5 Clusters

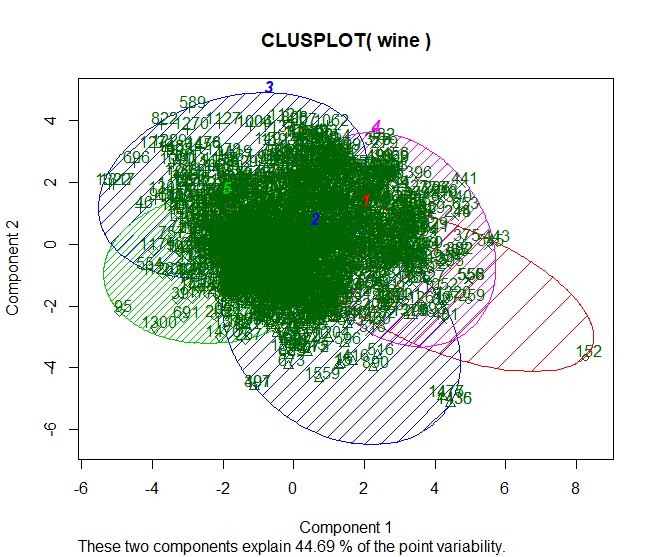
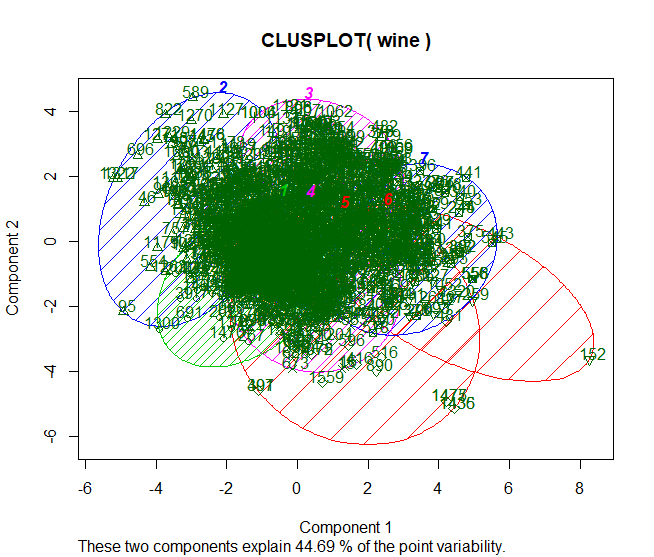


Figure 4. Plot of K-Means Model Three - 7 Clusters



**Appendix 1**

**K-Means Model 1 – Centroids Measurements**

1

> print(kc1)

K-means clustering with 3 clusters of sizes 506, 391, 702

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cluster means: | |  |  |  |  |  |  |  |
|  | fixed\_acidity volatile\_acidity | citric\_acid | residual\_sugar | chlorides | fre\_sulfur\_dioxide | totalsulfur\_dioxide | density | pH | sulphates |
| 1 | 0.9477509 -0.72929480 | 1.006462 | 0.037025 | 0.241189 | -0.42769 | -0.46438 | 0.363669 | -0.72752 | 0.587226 |
| 2 | -0.0659228 0.04973665 | 0.099367 | 0.362931 | -0.00296 | 1.022368 | 1.26901 | 0.296455 | -0.17171 | -0.18326 |
| 3 | -0.6464190 0.49797171 | -0.7808 | -0.22883 | -0.1722 | -0.26116 | -0.37209 | -0.42725 | 0.620031 | -0.3212 |
|  | alcohol quality | |  |  |  |  |  |  |  |
| 1 | 0.38154044 0.5632786 | | |  |  |  |  |  |  |
| 2 | -0.99365 |  |  |  |  |  |  |  |  |
| 3 | 0.03203569 -0.1596148 | | |  |  |  |  |  |  |

**Appendix 2**

**K-Means Model 2 – Centroids Measurements**

K-means clustering with 5 clusters of sizes 30, 322, 336, 361, 550



**Appendix 3**

**K-Means Model 3 – Centroids Measurements**

K-means clustering with 7 clusters of sizes 478, 211, 272, 327, 46, 29, 236



**Appendix 4**

> # Assignment 5

> # by Kenneth Lulie, Data 630 - Ami Gates

> # Created 11/16

> #Worked on 11/17

>

> #Set random seed so it is reproducible

> set.seed(1234)

>

> #Load cluster package

> library("cluster")

> #load dplyr to assist in column removal

> library(dplyr)

>

> #Read the CSV file. and set working directory

> setwd("D:/UMUC/630/Week 11/Assignment 5")

> origwine <- read.csv("winequality\_red.csv")

>

>

> #Set up a DF to work off of

> wine <- origwine

>

> #we can see that the vast majority of the observations are in 5 6 and 7. We will move all 3s and 4s to 5 and all 8s to 7s.

> #and then have low quality, medium quality, high quality.

> summary(origwine)

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide density

Min. : 4.60 Min. :0.1200 Min. :0.000 Min. : 0.900 Min. :0.01200 Min. : 1.00 Min. : 6.00 Min. :0.9901

1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.: 1.900 1st Qu.:0.07000 1st Qu.: 7.00 1st Qu.: 22.00 1st Qu.:0.9956

Median : 7.90 Median :0.5200 Median :0.260 Median : 2.200 Median :0.07900 Median :14.00 Median : 38.00 Median :0.9968

Mean : 8.32 Mean :0.5278 Mean :0.271 Mean : 2.539 Mean :0.08747 Mean :15.87 Mean : 46.47 Mean :0.9967

3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.: 2.600 3rd Qu.:0.09000 3rd Qu.:21.00 3rd Qu.: 62.00 3rd Qu.:0.9978

Max. :15.90 Max. :1.5800 Max. :1.000 Max. :15.500 Max. :0.61100 Max. :72.00 Max. :289.00 Max. :1.0037

pH sulphates alcohol quality

Min. :2.740 Min. :0.3300 Min. : 8.40 Min. :3.000

1st Qu.:3.210 1st Qu.:0.5500 1st Qu.: 9.50 1st Qu.:5.000

Median :3.310 Median :0.6200 Median :10.20 Median :6.000

Mean :3.311 Mean :0.6581 Mean :10.42 Mean :5.636

3rd Qu.:3.400 3rd Qu.:0.7300 3rd Qu.:11.10 3rd Qu.:6.000

Max. :4.010 Max. :2.0000 Max. :14.90 Max. :8.000

>

> # data looks fine.

> summary(wine)

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide density

Min. : 4.60 Min. :0.1200 Min. :0.000 Min. : 0.900 Min. :0.01200 Min. : 1.00 Min. : 6.00 Min. :0.9901

1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.: 1.900 1st Qu.:0.07000 1st Qu.: 7.00 1st Qu.: 22.00 1st Qu.:0.9956

Median : 7.90 Median :0.5200 Median :0.260 Median : 2.200 Median :0.07900 Median :14.00 Median : 38.00 Median :0.9968

Mean : 8.32 Mean :0.5278 Mean :0.271 Mean : 2.539 Mean :0.08747 Mean :15.87 Mean : 46.47 Mean :0.9967

3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.: 2.600 3rd Qu.:0.09000 3rd Qu.:21.00 3rd Qu.: 62.00 3rd Qu.:0.9978

Max. :15.90 Max. :1.5800 Max. :1.000 Max. :15.500 Max. :0.61100 Max. :72.00 Max. :289.00 Max. :1.0037

pH sulphates alcohol quality

Min. :2.740 Min. :0.3300 Min. : 8.40 Min. :3.000

1st Qu.:3.210 1st Qu.:0.5500 1st Qu.: 9.50 1st Qu.:5.000

Median :3.310 Median :0.6200 Median :10.20 Median :6.000

Mean :3.311 Mean :0.6581 Mean :10.42 Mean :5.636

3rd Qu.:3.400 3rd Qu.:0.7300 3rd Qu.:11.10 3rd Qu.:6.000

Max. :4.010 Max. :2.0000 Max. :14.90 Max. :8.000

>

>

> #Need to remove quality before clustering

> wine <- select(wine, -quality)

> #Scale variable before using in clustering

> wine <- scale(wine)

>

> # data looks fine.

> summary(wine)

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide

Min. :-2.1364 Min. :-2.27757 Min. :-1.39104 Min. :-1.1623 Min. :-1.60344 Min. :-1.4221 Min. :-1.2302

1st Qu.:-0.7005 1st Qu.:-0.76969 1st Qu.:-0.92903 1st Qu.:-0.4531 1st Qu.:-0.37111 1st Qu.:-0.8485 1st Qu.:-0.7438

Median :-0.2410 Median :-0.04367 Median :-0.05634 Median :-0.2403 Median :-0.17989 Median :-0.1792 Median :-0.2574

Mean : 0.0000 Mean : 0.00000 Mean : 0.00000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000

3rd Qu.: 0.5056 3rd Qu.: 0.62649 3rd Qu.: 0.76501 3rd Qu.: 0.0434 3rd Qu.: 0.05383 3rd Qu.: 0.4900 3rd Qu.: 0.4722

Max. : 4.3538 Max. : 5.87614 Max. : 3.74240 Max. : 9.1928 Max. :11.12355 Max. : 5.3656 Max. : 7.3728

density pH sulphates alcohol quality

Min. :-3.53762 Min. :-3.69924 Min. :-1.9359 Min. :-1.8983 Min. :-3.2641

1st Qu.:-0.60757 1st Qu.:-0.65494 1st Qu.:-0.6380 1st Qu.:-0.8661 1st Qu.:-0.7876

Median : 0.00176 Median :-0.00721 Median :-0.2251 Median :-0.2092 Median : 0.4507

Mean : 0.00000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000

3rd Qu.: 0.57664 3rd Qu.: 0.57574 3rd Qu.: 0.4239 3rd Qu.: 0.6353 3rd Qu.: 0.4507

Max. : 3.67890 Max. : 4.52687 Max. : 7.9162 Max. : 4.2011 Max. : 2.9273

>

>

>

>

> ### Data exploration

> counts <- table(origwine$quality)

> barplot(counts, main="Wine Quality Distribution",

+ xlab="Wine Quality")

>

>

>

>

> #set quality to factor in original

> #Transform all to 3 groups

> origwine$quality[origwine$quality == 3] <- 5

> origwine$quality[origwine$quality == 4] <- 5

> origwine$quality[origwine$quality == 8] <- 7

>

>

>

>

>

>

>

> ######## Model One - 3 clusters

>

>

>

> #Run the method and store the result in kc variable

> kc1<-kmeans(wine, 3, nstart = 10)

>

>

>

> #Cluster to class evaluation

> table(origwine$quality, kc1$cluster)

1 2 3

5 119 274 351

6 231 104 303

7 156 11 50

> #Do cluster plot

> clusplot(wine, kc1$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)

>

> print(kc1)

K-means clustering with 3 clusters of sizes 506, 389, 704

Cluster means:

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide density pH sulphates

1 0.94775088 -0.72929480 1.0064621 0.03702487 0.241189151 -0.4276930 -0.4643849 0.3636687 -0.7275167 0.5872260

2 -0.05970723 0.05595957 0.1004318 0.36566736 -0.001173487 1.0241225 1.2737317 0.2991146 -0.1762184 -0.1936646

3 -0.64820431 0.49325980 -0.7788889 -0.22866362 -0.172706284 -0.2584815 -0.3700325 -0.4266647 0.6202733 -0.3150580

alcohol quality

1 0.38154044 0.5632786

2 -0.55496332 -0.4406022

3 0.03241658 -0.1613987

Clustering vector:

[1] 3 2 3 1 3 3 3 3 3 2 3 2 3 1 2 2 2 1 3 1 2 2 1 2 3 3 3 1 3 3 3 3 2 2 3 3 3 1 3 2 2 2 1 3 3 3 2 1 3 2 3 3 3 2 2 2 1 2 3 3 2 2 3 3 3 3 3 3 1 3 3

[72] 2 2 3 2 1 1 3 3 2 3 1 2 1 3 3 1 3 2 3 2 1 1 3 3 3 3 3 3 3 3 3 3 2 3 2 1 3 2 2 3 2 2 1 3 1 3 3 3 2 3 3 3 3 2 2 3 3 3 3 2 3 3 3 3 2 3 3 2 2 2 3

[143] 3 3 3 2 3 2 3 3 1 1 2 2 2 2 2 2 3 3 3 3 3 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 2 2 3 2 2 2 3 2 3 3 2 3 1 3 3 1 2 3 3 3 1 1 2 2 1 1 3 1

[214] 2 3 2 3 3 3 2 2 2 3 3 2 2 1 3 2 3 3 3 2 3 3 3 3 3 3 3 1 1 2 1 1 3 3 2 3 3 1 3 1 2 3 2 1 3 1 1 2 3 3 2 1 1 2 1 3 1 2 1 1 3 2 2 3 1 1 1 1 1 3 1

[285] 2 2 1 2 1 2 1 1 1 3 1 1 2 3 3 3 3 1 3 3 2 1 3 1 1 3 1 2 2 2 2 3 2 2 1 2 1 2 3 2 2 2 1 1 1 1 1 1 2 3 3 1 1 2 1 1 1 1 1 1 1 2 3 1 1 3 1 3 3 1 2

[356] 3 1 1 1 1 2 1 1 1 1 1 1 1 1 1 3 1 1 2 1 1 1 1 1 2 1 1 1 1 2 3 2 3 2 1 3 1 1 2 1 1 2 1 1 3 2 3 1 1 3 1 1 1 1 2 2 2 2 1 2 2 1 2 1 3 1 3 3 1 3 3

[427] 3 3 3 1 1 2 1 1 1 1 2 1 1 3 1 1 1 1 3 3 1 1 3 1 1 1 3 1 3 1 3 2 1 2 1 3 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 1 3

[498] 2 1 2 3 1 1 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 1 2 1 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 3 1 1 2 1 3 1 1 2 3 1 1 1 3 1 1 3 1 1 1 1 1 1 1 2 2 2 1 1 3 3

[569] 1 3 1 3 1 2 1 1 3 2 2 1 1 1 1 1 2 3 1 2 3 1 2 2 2 1 3 2 1 1 3 1 3 1 3 1 2 3 1 2 1 3 2 1 3 1 1 2 2 1 1 1 2 2 1 3 2 2 3 3 3 2 3 1 3 2 2 3 2 2 3

[640] 1 2 2 2 2 2 3 3 3 1 2 1 2 1 1 1 2 1 1 3 3 3 2 3 1 1 2 1 1 1 1 2 3 2 3 1 1 1 3 2 1 1 3 2 3 2 3 3 2 3 1 3 2 1 2 2 3 3 3 2 1 2 3 3 2 3 3 3 3 3 1

[711] 2 2 3 3 2 3 3 3 3 3 3 2 3 2 3 3 3 3 3 3 1 3 3 2 3 3 3 3 2 3 3 2 3 2 2 3 2 2 3 3 3 3 2 3 1 3 3 3 3 2 2 3 3 3 3 3 2 2 2 3 2 2 2 1 1 3 3 3 3 2 3

[782] 3 2 3 3 2 2 2 2 2 2 2 3 3 1 1 2 1 1 1 2 3 3 3 3 1 1 1 3 3 3 1 1 3 1 1 1 1 3 2 3 3 3 3 3 3 1 3 3 3 3 3 2 2 3 3 2 2 1 3 1 3 2 2 1 3 3 3 3 3 1 1

[853] 2 2 2 3 2 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 1 1 3 3 2 3 3 3 1 3 2 3 3 1 3 2 2 3 1 3 3 3 1 3 1 3 1 3 3 3 3 2 3 3 3 3 1 1 1 1 3 1 3 2 2 3 1 2 3

[924] 2 1 2 2 2 1 1 3 3 2 3 3 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 3 3 1 3 3 1 1 1 1 2 1 3 1 1 1 3 1 2 2 2 3 1 1 3 2 1 1 3 1 2 3 1 3 2 3 2

[995] 2 3 3 3 3 3

[ reached getOption("max.print") -- omitted 599 entries ]

Within cluster sum of squares by cluster:

[1] 5609.049 3731.160 4686.527

(between\_SS / total\_SS = 26.9 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter" "ifault"

> kc1$betweenss

[1] 5149.264

> kc1$iter

[1] 3

>

>

> ######## Model Two - 5 clusters

>

>

>

> #Run the method and store the result in kc variable

> kc2<-kmeans(wine, 5, nstart = 10)

>

>

>

> #Cluster to class evaluation

> table(origwine$quality, kc2$cluster)

1 2 3 4 5

5 26 366 20 237 95

6 189 183 9 78 179

7 117 8 1 7 84

>

> #Do cluster plot

> clusplot(wine, kc2$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)

>

>

> print(kc2)

K-means clustering with 5 clusters of sizes 332, 557, 30, 322, 358

Cluster means:

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide density pH sulphates

1 -0.59791225 -0.47985627 -0.1106149 -0.2222480 -0.38084049 0.18026693 -0.2043088 -1.13931268 0.5261873 0.1980333

2 -0.48333933 0.67817840 -0.7859916 -0.2323414 -0.07620962 -0.39799332 -0.3684807 -0.09148229 0.4072406 -0.4198352

3 0.06530142 0.01031027 1.1038149 -0.4081571 5.52140920 -0.08682997 0.5005435 0.18049490 -1.6740232 3.6233581

4 -0.07317300 0.07088088 0.1037187 0.4486917 0.02664288 1.10928919 1.3832103 0.39002947 -0.1560665 -0.1827354

5 1.36684228 -0.67476547 1.1396916 0.1982304 -0.01489799 -0.53841497 -0.5232842 0.83296958 -0.8409305 0.3302828

alcohol quality

1 1.2149075 0.8311439

2 -0.4328307 -0.4652224

3 -0.8880034 -0.3748151

4 -0.6124736 -0.4645457

5 0.1720502 0.4022829

Clustering vector:

[1] 2 2 2 5 2 2 2 2 2 4 2 4 2 3 4 4 4 3 2 3 4 4 2 2 2 2 2 2 2 2 2 2 4 4 2 2 2 5 2 4 4 2 3 2 2 1 4 5 2 4 2 2 2 4 4 2 5 4 2 2 4 4 2 2 2 2 2 2 4 2 2

[72] 4 4 2 4 5 5 2 2 4 2 3 4 3 1 2 3 2 4 2 4 3 3 2 2 1 2 2 2 2 2 2 2 2 2 2 3 2 4 4 2 4 4 5 2 5 2 2 2 2 2 2 2 2 4 4 2 2 1 2 4 1 1 2 2 2 2 2 4 4 2 2

[143] 1 2 1 4 2 3 2 1 1 3 4 4 4 4 4 4 2 2 2 2 2 4 4 4 4 2 2 3 2 2 2 1 2 2 2 2 2 2 2 3 2 2 2 4 4 2 4 4 4 2 4 2 2 4 2 5 1 2 5 4 2 2 2 5 5 4 4 5 5 2 5

[214] 4 2 4 2 2 2 4 4 4 2 2 2 4 3 2 4 2 1 2 4 2 2 2 2 2 2 2 3 5 4 5 5 2 2 2 2 2 5 2 5 2 2 4 5 2 3 5 2 2 2 4 5 5 2 1 2 5 4 5 5 2 4 4 2 5 5 5 5 3 2 5

[285] 4 4 5 4 1 4 1 3 5 2 5 5 4 2 2 2 2 5 2 2 4 5 2 5 5 2 5 4 4 4 4 1 4 4 5 4 5 4 2 4 5 5 5 5 5 5 5 5 4 2 1 5 1 4 5 5 5 5 5 5 5 4 1 5 5 2 5 2 2 5 1

[356] 1 5 5 5 5 4 4 5 5 5 5 5 5 5 5 2 5 5 4 5 5 5 5 5 4 5 5 5 5 4 2 4 2 4 5 1 5 5 4 5 5 4 5 5 2 4 1 5 5 2 5 5 5 5 5 4 4 2 5 4 4 5 4 5 2 1 1 2 5 2 1

[427] 1 2 2 5 5 4 5 5 5 5 4 5 5 2 5 5 5 5 1 2 5 5 2 5 5 3 2 5 1 5 2 4 5 5 5 2 5 4 5 5 5 1 5 2 5 5 5 5 5 2 5 5 2 2 5 5 5 5 5 5 5 5 5 5 4 1 1 4 1 5 2

[498] 4 5 4 2 5 5 5 5 5 5 5 4 5 5 4 5 5 5 4 5 5 5 4 5 2 4 4 4 4 4 1 4 2 5 5 5 1 5 5 2 2 5 5 4 5 2 5 5 4 2 5 5 5 2 5 5 2 5 5 5 5 5 5 5 4 4 4 5 5 2 2

[569] 5 1 5 1 5 5 5 5 2 4 4 5 5 5 5 5 4 2 5 4 1 5 4 1 4 5 2 4 5 5 2 5 2 5 2 5 4 2 5 4 5 1 4 5 2 2 3 4 4 5 5 5 4 4 2 1 4 4 2 2 2 4 2 5 1 4 4 2 4 4 2

[640] 5 5 4 5 4 5 2 2 2 1 4 5 4 5 5 5 4 5 5 2 2 2 2 2 5 5 4 5 5 5 5 4 2 4 2 5 5 5 2 4 5 5 2 4 2 4 2 2 2 2 5 2 2 3 4 4 1 2 2 4 5 4 2 2 4 2 2 2 2 1 5

[711] 4 4 2 2 4 2 2 2 2 2 2 4 2 4 2 2 2 2 2 1 3 2 2 4 2 2 2 2 4 2 2 4 2 4 4 2 4 4 2 2 2 2 4 2 3 2 2 2 2 4 4 2 2 2 2 2 4 4 4 2 4 4 4 5 5 2 2 2 2 4 2

[782] 2 4 2 2 5 5 5 5 4 4 4 2 2 5 5 4 1 5 5 4 2 1 2 2 1 1 1 2 2 2 5 5 2 5 5 5 5 2 2 2 1 2 2 2 2 1 2 1 1 2 1 4 4 2 2 1 1 5 2 5 2 4 2 5 2 2 2 2 2 5 5

[853] 4 1 1 1 1 1 5 1 2 1 2 2 2 2 1 1 1 2 1 2 4 1 5 1 2 1 4 2 2 1 1 2 4 2 2 1 1 4 4 2 5 2 2 2 1 2 1 2 1 1 1 1 1 4 2 1 1 1 5 5 5 5 1 1 2 4 4 1 5 4 1

[924] 4 1 1 4 2 1 1 2 2 4 2 2 1 1 5 1 1 1 5 5 5 1 5 5 1 1 1 1 1 1 5 1 5 5 5 1 2 1 2 2 1 1 1 5 4 1 2 5 5 5 2 5 4 4 4 1 5 5 2 1 5 5 2 1 4 2 5 2 4 2 4

[995] 4 2 1 1 2 1

[ reached getOption("max.print") -- omitted 599 entries ]

Within cluster sum of squares by cluster:

[1] 2436.6014 2831.9529 494.9888 2899.0124 2687.3474

(between\_SS / total\_SS = 40.8 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter" "ifault"

> kc2$betweenss

[1] 7826.097

> kc2$iter

[1] 5

>

>

>

>

>

> ######## Model Three - 7 clusters

>

>

>

> #Run the method and store the result in kc variable

> kc3<-kmeans(wine, 7, nstart = 10)

>

>

> #Cluster to class evaluation

> table(origwine$qual, kc3$cluster)

1 2 3 4 5 6 7

5 94 237 17 332 33 12 19

6 121 93 12 151 125 127 9

7 36 6 5 6 39 124 1

>

> #Do cluster plot

> clusplot(wine, kc3$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)

>

>

>

> print(kc3)

K-means clustering with 7 clusters of sizes 251, 336, 34, 489, 197, 263, 29

Cluster means:

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide density pH sulphates

1 1.65342323 -0.57333475 1.24337899 0.11800161 0.06551024 -0.50032591 -0.4498654 1.146378841 -0.97957239 0.24525405

2 -0.12221727 0.02937481 0.05472912 -0.02140148 -0.04051855 0.99741466 1.2302571 0.205514449 -0.09858596 -0.17378852

3 -0.08560643 -0.03464133 0.41472600 4.96021580 0.29629520 1.74964380 1.6953018 1.224617401 -0.32535781 -0.02378189

4 -0.42894562 0.67485881 -0.78772665 -0.19743990 -0.04719279 -0.48452173 -0.4123976 0.008683312 0.34261413 -0.41615679

5 -1.08242828 0.34867306 -0.91678019 -0.27162233 -0.41565628 0.27983410 -0.1575764 -1.337410954 1.14751585 -0.13102562

6 0.18854297 -1.00380298 0.71503972 -0.11193260 -0.26778589 -0.32391768 -0.5290464 -0.549785979 -0.20768509 0.45281970

7 0.08180575 0.01794915 1.14382029 -0.39927103 5.60297844 -0.07045695 0.4742672 0.185744661 -1.68682883 3.71944476

alcohol quality

1 -0.05533854 0.13003630

2 -0.57165428 -0.42272483

3 -0.36379918 -0.02275403

4 -0.50844289 -0.51155803

5 1.18197158 0.50099298

6 0.98745807 1.03453689

7 -0.88228685 -0.36058191

Clustering vector:

[1] 4 4 4 1 4 4 4 5 4 2 4 2 5 7 2 2 2 7 4 7 2 2 4 4 4 4 4 4 4 4 4 4 2 3 4 4 4 6 4 2 2 4 7 4 4 5 2 1 4 2 4 4 4 2 2 4 1 2 4 4 2 2 4 4 4 4 4 4 2 4 4

[72] 2 2 4 2 1 1 4 4 2 4 7 2 7 6 4 7 4 2 4 2 7 7 4 5 5 4 4 4 4 4 4 4 4 4 4 7 4 2 2 4 2 2 1 4 6 4 4 4 2 4 4 4 4 2 2 4 4 6 4 2 5 5 4 4 4 4 4 2 2 4 4

[143] 5 4 5 2 5 2 4 6 6 7 2 2 2 2 2 2 4 2 4 4 4 3 3 2 2 4 4 7 4 4 4 5 4 4 4 4 4 4 4 7 4 4 4 2 2 4 2 2 2 4 2 4 4 2 4 1 5 4 6 2 4 4 4 1 1 2 2 1 6 4 1

[214] 2 4 2 4 4 4 2 2 2 4 4 4 2 7 4 2 4 5 4 2 4 4 4 4 4 4 4 7 1 2 1 1 4 4 4 4 4 1 4 1 2 4 2 1 4 7 1 2 4 4 2 1 1 4 6 4 1 2 1 1 4 3 2 4 1 6 6 1 7 4 6

[285] 2 2 1 2 6 1 6 7 1 4 1 1 2 4 4 4 4 1 4 4 2 1 4 1 1 6 1 2 2 2 2 6 2 2 6 2 6 2 4 2 3 3 1 1 1 1 1 1 2 4 5 1 6 2 1 1 1 1 1 1 1 2 5 1 1 4 1 4 4 1 2

[356] 5 1 6 1 1 2 2 1 1 1 6 1 1 1 6 2 6 6 2 1 1 1 6 1 2 6 1 6 6 2 4 2 4 2 1 5 1 1 2 1 1 3 1 1 4 3 6 1 1 4 6 1 1 1 1 2 2 4 1 2 3 1 2 1 4 6 5 4 6 4 5

[427] 5 4 4 1 6 2 1 1 1 1 2 1 1 4 1 1 1 6 5 4 1 1 4 1 1 7 4 1 2 6 4 2 1 1 6 4 1 2 1 1 1 6 1 2 1 6 1 1 1 4 1 1 4 4 3 6 1 1 1 1 1 1 1 1 2 6 6 2 3 6 4

[498] 2 6 2 4 3 3 6 6 6 6 1 2 1 1 2 1 1 1 3 1 1 6 2 1 2 2 2 2 2 2 2 2 4 6 1 1 6 1 6 4 4 1 1 2 1 4 1 1 2 4 1 1 1 4 1 1 5 1 1 1 1 1 1 1 2 2 2 1 1 4 4

[569] 1 5 1 5 1 1 1 1 4 2 2 1 1 1 1 1 2 4 1 2 5 1 2 2 2 1 4 3 1 1 4 1 4 1 4 1 2 4 6 2 1 5 2 1 4 4 7 2 2 1 1 1 2 2 4 6 4 4 4 4 4 2 4 1 4 4 2 4 2 2 4

[640] 6 1 2 1 2 1 4 4 4 6 3 1 2 1 6 1 2 1 1 4 4 4 4 4 1 1 2 1 1 1 1 2 4 2 4 1 1 1 4 2 1 1 4 2 4 2 4 4 4 4 1 4 4 7 2 2 5 4 4 2 1 2 4 4 2 4 4 4 4 5 1

[711] 2 2 4 4 2 4 4 4 4 4 4 2 4 2 4 4 4 4 4 5 7 4 4 2 4 4 4 4 2 4 4 2 4 1 1 4 2 2 4 4 4 4 2 4 7 5 5 4 4 2 2 4 4 4 4 4 2 2 2 4 2 2 2 6 6 4 4 4 4 2 4

[782] 4 2 4 4 1 1 2 2 2 2 2 4 4 6 1 2 6 1 1 2 4 5 4 4 6 6 6 4 4 4 1 1 4 1 1 1 6 4 4 4 5 4 4 4 4 6 4 5 5 4 5 2 2 4 4 5 5 6 4 6 4 2 2 6 4 4 4 4 4 1 1

[853] 2 2 2 5 2 6 1 5 4 5 4 4 4 4 5 5 5 5 5 5 2 6 6 6 4 5 2 4 4 5 6 4 2 4 4 6 5 3 2 4 1 4 4 5 6 5 6 4 6 5 5 5 5 2 2 5 5 6 6 3 6 6 6 6 5 3 2 5 1 2 5

[924] 3 6 2 2 4 6 6 4 4 2 4 4 6 6 1 6 5 6 6 1 1 6 6 6 6 6 6 6 6 6 6 6 6 6 6 5 4 6 4 4 6 6 6 6 2 6 4 6 6 6 6 6 2 2 2 6 1 6 4 5 6 1 5 6 2 4 6 4 2 2 2

[995] 2 4 5 5 4 5

[ reached getOption("max.print") -- omitted 599 entries ]

Within cluster sum of squares by cluster:

[1] 1636.4155 2025.5769 658.4632 2351.1973 1210.0044 1450.9580 478.2138

(between\_SS / total\_SS = 48.8 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter" "ifault"

> kc3$betweenss

[1] 9365.171

> kc3$iter

[1] 4

>

>

>

>

>

>

> ### Centroids 3 and 4 evaluation

> ## Create new DF, add kc3 cluster results

> C <- origwine

> C$index <- kc3$cluster

>

>

> ##Create DF with only objects where the cluster assigned was 3

> #careful on capitalization here

> C3 <- subset(C, index == 3)

> #use summary to get means.

> summary (C3)

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide density

Min. : 5.600 Min. :0.2800 Min. :0.0200 Min. : 6.100 Min. :0.0500 Min. : 3.00 Min. : 9.00 Min. :0.9932

1st Qu.: 6.800 1st Qu.:0.3950 1st Qu.:0.2075 1st Qu.: 7.425 1st Qu.:0.0700 1st Qu.:24.50 1st Qu.: 76.75 1st Qu.:0.9979

Median : 7.700 Median :0.5000 Median :0.3150 Median : 8.300 Median :0.0790 Median :30.50 Median : 93.00 Median :0.9990

Mean : 8.171 Mean :0.5216 Mean :0.3518 Mean : 9.532 Mean :0.1014 Mean :34.18 Mean :102.24 Mean :0.9991

3rd Qu.: 9.900 3rd Qu.:0.6450 3rd Qu.:0.4475 3rd Qu.:11.000 3rd Qu.:0.1163 3rd Qu.:39.50 3rd Qu.:121.00 3rd Qu.:1.0013

Max. :10.700 Max. :0.9000 Max. :0.7800 Max. :15.500 Max. :0.2350 Max. :72.00 Max. :289.00 Max. :1.0037

pH sulphates alcohol quality index

Min. :3.010 Min. :0.4800 Min. : 8.80 Min. :5.000 Min. :3

1st Qu.:3.170 1st Qu.:0.5400 1st Qu.: 9.20 1st Qu.:5.000 1st Qu.:3

Median :3.245 Median :0.5900 Median : 9.40 Median :5.500 Median :3

Mean :3.261 Mean :0.6541 Mean :10.04 Mean :5.647 Mean :3

3rd Qu.:3.370 3rd Qu.:0.7500 3rd Qu.:10.80 3rd Qu.:6.000 3rd Qu.:3

Max. :3.470 Max. :1.1400 Max. :12.30 Max. :7.000 Max. :3

> ##Repeat above for 4

>

> C4 <- subset(C, index == 4)

> summary (C4)

fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides fre\_sulfur\_dioxide totalsulfur\_dioxide density

Min. : 5.200 Min. :0.3200 Min. :0.0000 Min. :1.20 Min. :0.03900 Min. : 3.00 Min. : 7.0 Min. :0.9940

1st Qu.: 7.000 1st Qu.:0.5400 1st Qu.:0.0200 1st Qu.:1.80 1st Qu.:0.07400 1st Qu.: 6.00 1st Qu.:20.0 1st Qu.:0.9960

Median : 7.500 Median :0.6300 Median :0.0900 Median :2.10 Median :0.08000 Median :10.00 Median :31.0 Median :0.9967

Mean : 7.573 Mean :0.6487 Mean :0.1175 Mean :2.26 Mean :0.08525 Mean :10.81 Mean :32.9 Mean :0.9968

3rd Qu.: 8.100 3rd Qu.:0.7200 3rd Qu.:0.2100 3rd Qu.:2.40 3rd Qu.:0.08900 3rd Qu.:15.00 3rd Qu.:41.0 3rd Qu.:0.9974

Max. :10.100 Max. :1.5800 Max. :0.5400 Max. :7.80 Max. :0.26700 Max. :34.00 Max. :93.0 Max. :1.0010

pH sulphates alcohol quality index

Min. :2.880 Min. :0.3300 Min. : 9.000 Min. :5.000 Min. :4

1st Qu.:3.290 1st Qu.:0.5200 1st Qu.: 9.500 1st Qu.:5.000 1st Qu.:4

Median :3.360 Median :0.5700 Median : 9.800 Median :5.000 Median :4

Mean :3.364 Mean :0.5876 Mean : 9.881 Mean :5.333 Mean :4

3rd Qu.:3.440 3rd Qu.:0.6300 3rd Qu.:10.200 3rd Qu.:6.000 3rd Qu.:4

Max. :3.690 Max. :1.2000 Max. :12.200 Max. :7.000 Max. :4