Data Mining Exercise 4

Kylie Taylor, Larisa Barreto, David Fraire

4/25/2019

Clustering and PCA of Wines

The 'wine' dataset contains 11 different chemical properties of wines produced from the vine "vino verde", grown only in Northern Portugal. There are 6,497 total observations and 13 variables detailing observed chemical elements of the wines, color and quality of wine. The chemical properties measured are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH level, sulfates, and alcohol percentage. We were also given the have color of the wine, red or white, and quality, rated 1 through 10.

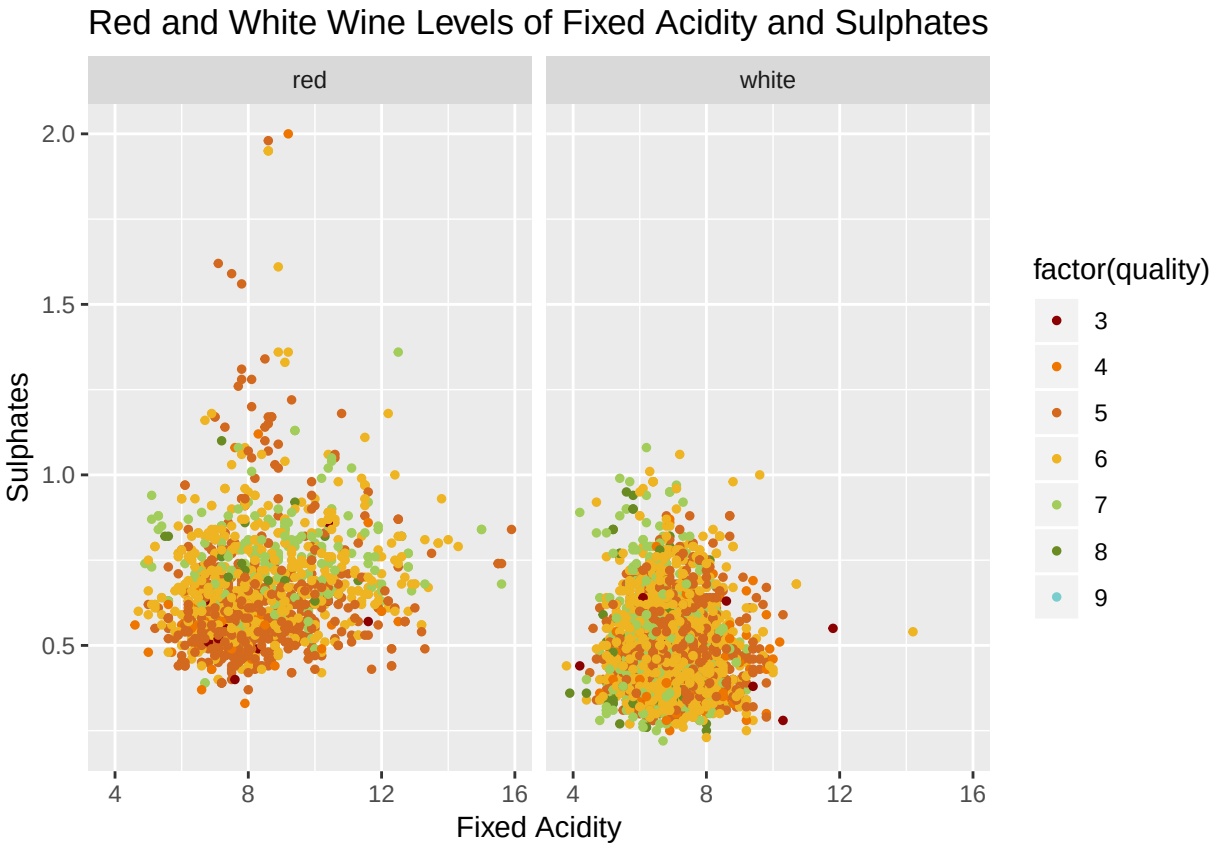
To start things off, we looked at the summary statistics of all the variables in the Wine dataset, as one should do with any statistical analysis. The summary stats below reveal that all variables appear to behaving well and contain no outlandish values. Some interesting findings were that pH only ranged from 2.7 to 4.0, meaning wine is more acidic than neutral. The quality of wine only ranged from 3 to 9, this means there are only 7 ratings of wines in this dataset, instead of all 10. The final finding was that there were disproportionate more white wines, 4,898, than red wines, 1,599, in the dataset.

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0
##  Median :0.04700   Median : 29.00      Median :118.0
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
##  Median :0.9949   Median :3.210   Median :0.5100   Median :10.30
##  Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49
##  3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
##  Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality        color
##  Min.   :3.000   red  :1599
##  1st Qu.:5.000   white:4898
##  Median :6.000
##  Mean   :5.818
##  3rd Qu.:6.000
##  Max.   :9.000
```

In order to see relationships across each variable we created a Correlogram that demonstrates several interesting plots. First, the distributions shown along the diagonal of the Correlogram show us the difference between red and white wines for each variable.

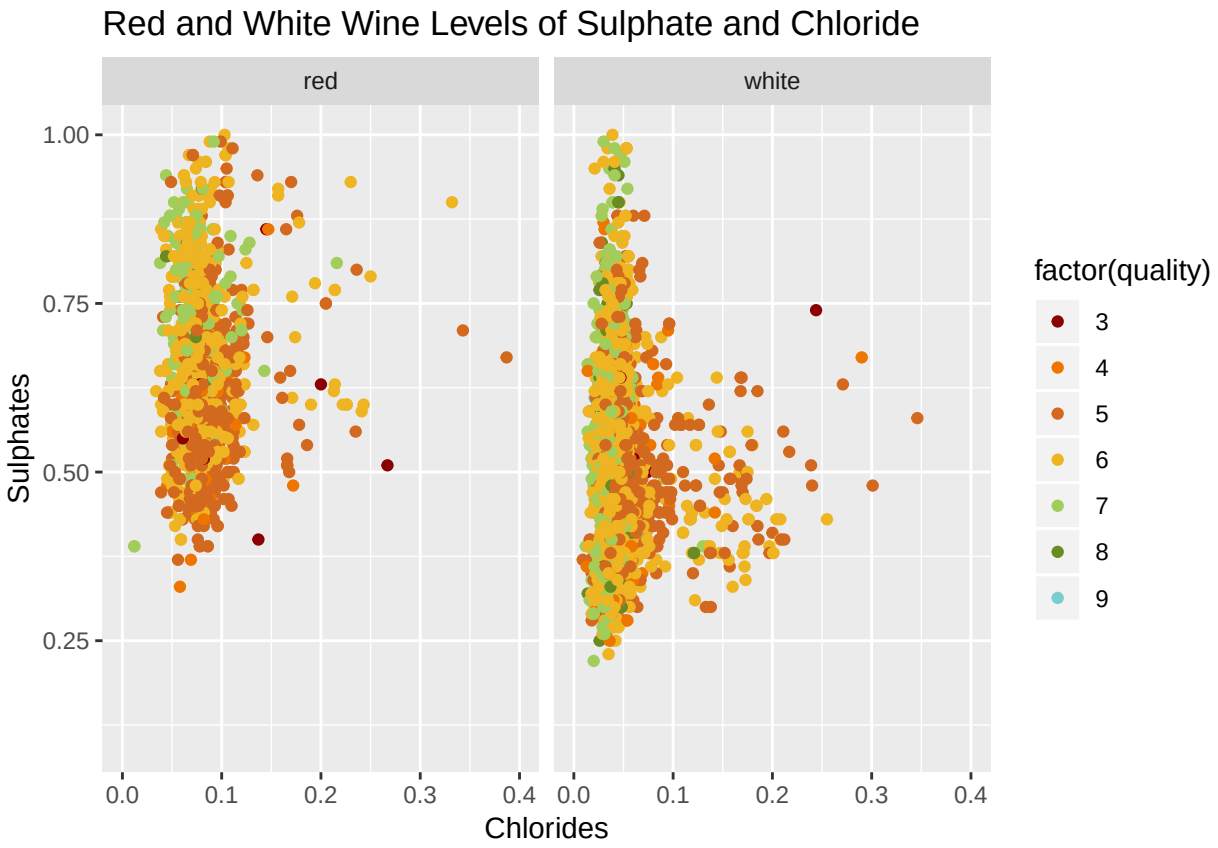
According to the diagram, red wines are shown to have higher levels of density, pH, chlorides, and sulphates than white wines.

On the other hand, white wines within our dataset show higher levels of residual sugars and sulfur dioxide.

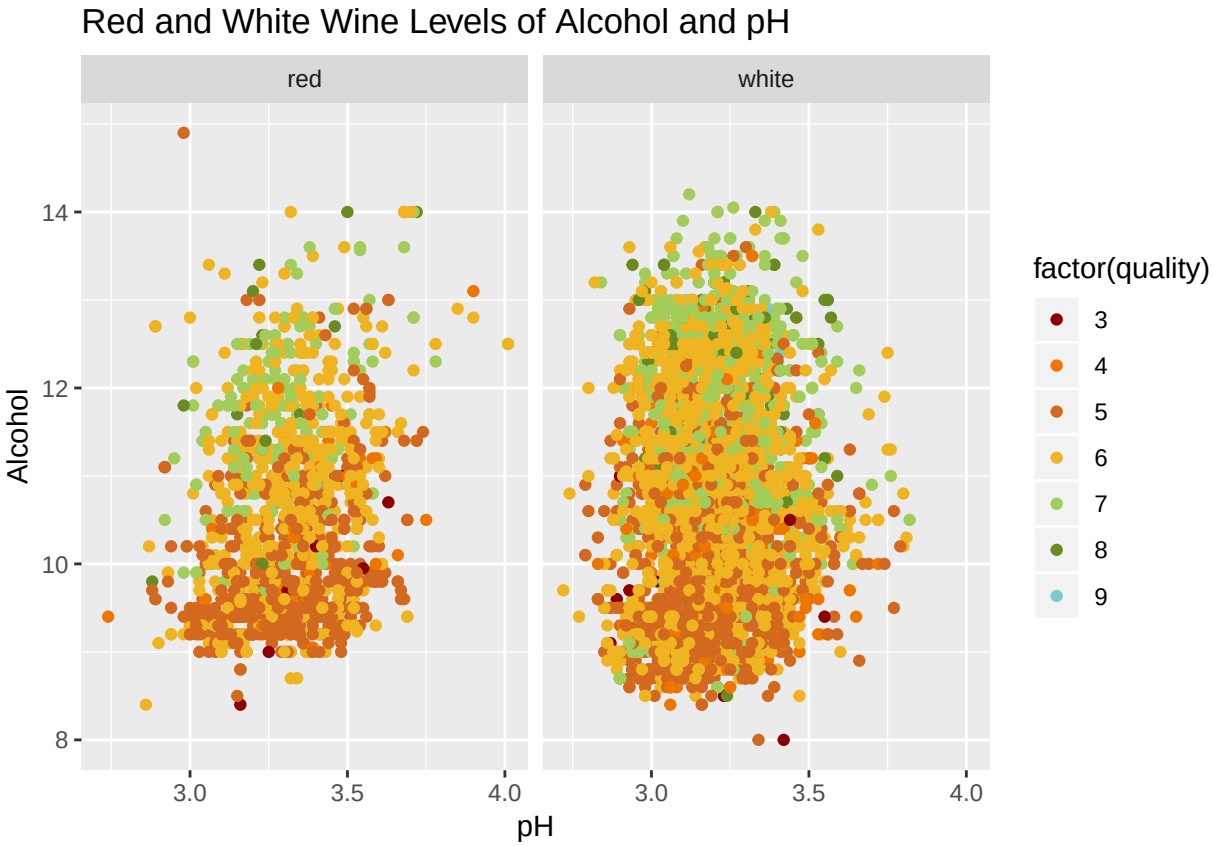


The graph shows us that red wines in our dataset cluster around a fixed acidity level of 8, and sulphate levels of .5 to .8 . White wines cluster between fixed acidity levels of 5 and 8, and sulfate levels of .25 to .75.

```
## Warning: Removed 61 rows containing missing values (geom_point).
```

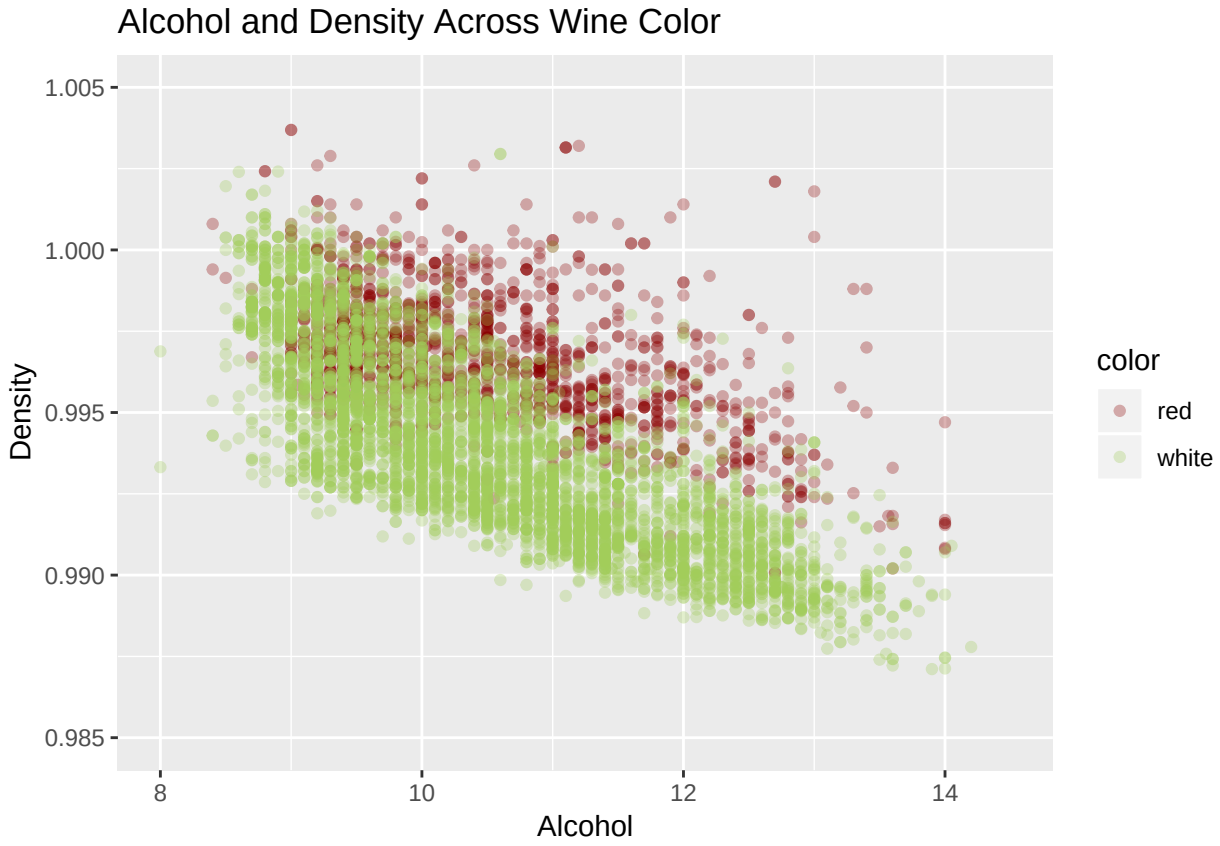


This demonstrates to us that red wines cluster at a chloride level of .005 and a sulfate level between .3 and .8. White wines cluster at a lower level of chloride level between 0.01 and 0.08.

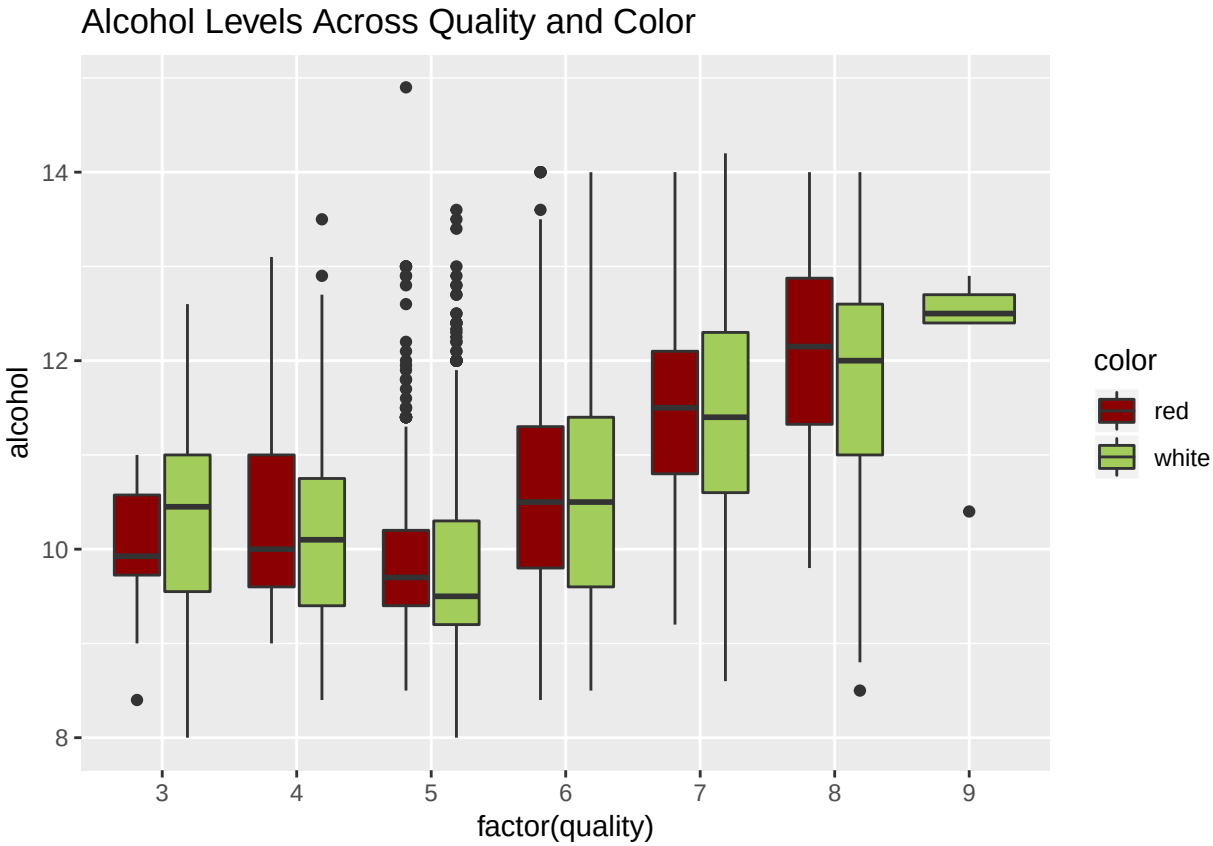


In this graph we see interesting clustering as alcohol levels increase. Across both colors of wines, the higher quality wines tend to have higher levels of alcohol. Lower quality wines tend to have lower levels of alcohol.

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



This plot of Alcohol and density levels shows us a clear negative correlation across both variables. As alcohol levels increase, the density of the wine will decrease. Furthermore, we see that red wines tend to have higher density than in comparison to white wines.



Plotting box plots across each level of quality and alcohol confirms that regardless of color of wine, higher alcohol content is associated with a higher quality or vice versa.

Once again, we see that this dataset is dominated by several mid-quality wines and that the higher quality wines tend to have higher alcohol content.

Since our working knowledge of chemical properties of wine is minimal to say the least, we scaled and standardized all the numerical variables. Now, instead of doing an analysis on the absolute values of the variables, all observations will be expressed as standard deviations away form a respective variable's mean. This implies all standardized variables will have a mean of zero.

To begin, we ran two K-means ++ for both color and quality, respectively. We ran a K-means ++ clustering with two clusters, in hopes that K-means would be able to identify clusters specific to red and white wines. We used 25 starts and found that the algorithm was able to identify two clusters, with the following coordinates for the first cluster:

```
##        fixed.acidity     volatile.acidity          citric.acid
##           6.84794989           0.27377097           0.33553945
##       residual.sugar            chlorides  free.sulfur.dioxide
##           6.41230068           0.04505757          35.62859805
## total.sulfur.dioxide              density                   pH
##         138.66980741           0.99400200           3.18764962
##            sulphates              alcohol              quality
##           0.48890246          10.52740319           5.90308552
```

And the following coordinates for the second cluster:

```
##        fixed.acidity     volatile.acidity          citric.acid
##           8.27883693           0.53043765           0.26968825
##       residual.sugar            chlorides  free.sulfur.dioxide
```

```
##           2.63770983           0.08781115          15.75089928
## total.sulfur.dioxide              density                 pH
##          49.37410072           0.99670764           3.30781775
##            sulphates              alcohol              quality
##           0.65392086          10.38872902           5.57314149
```

The coordinates above reveal that the two clusters do have different values for the 11 chemicals, although some are similar in location. We do know that the first cluster contains 1,601 members, which is awfully close to 1,599 (the number of red wines), and the second cluster contains 4,896 members (close to the number of whites). Since we are personally unable to visualize a 13 dimensional space, we decided to make some plots of the cluster membership across several properties.

Below we see three boxplots of color of wine against 'residual sugar', 'total sulfur dioxide', and 'volatile acidity'. The color denotes which cluster that observation belongs to. It is fairly obvious that the K-means ++ did a good job of determining which wines are red and white, since there aren't many observations shared between the two groups. This can be seen on the plots below where there are only but a few green dots mixed with the red, and vice versa.

Next, we wanted to see if K-means ++ could do equally as well of a job in determining the quality rating of the wine. Since no wines in this dataset were rated 1, 2, or 10, we only included 7 clusters, with 25 starts. Reading tables of the coordinated for all seven clusters was too much to decipher, so we made similar plots as above.

The three boxplots below are the quality rating against 'residual sugar', 'total sulfur dioxide', and 'volatile acidity'. The color denotes membership of each cluster. Simply by examining the plots, it is obvious that K-means ++ did not accomplish clustering quality very well at all. There are clearly shared members across each rating of quality, seen by the scattering of colors across the plots.

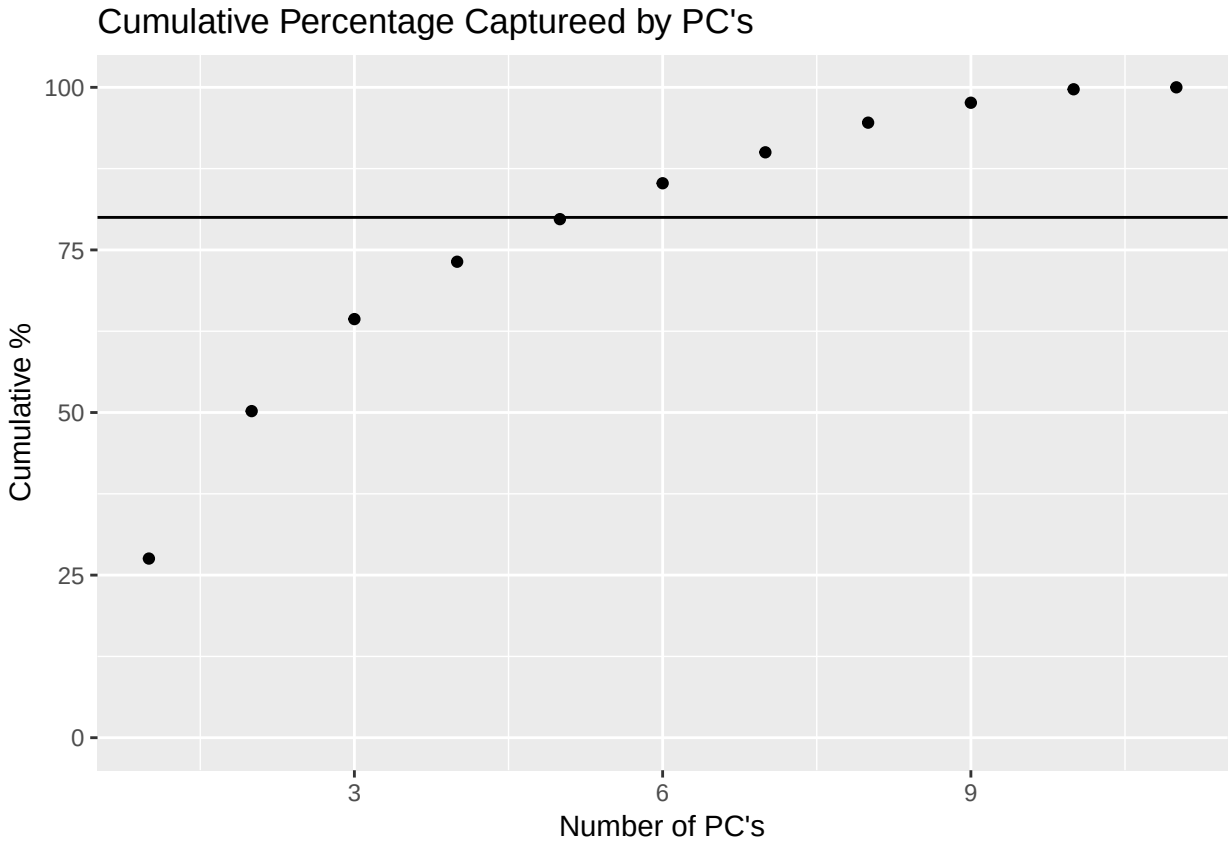
To verify the conclusions above, we find that total within cluster average distance for the 2 cluster model is 62,783.79, and for the 7 cluster model is 39,465.86. The between cluster average distance for the 2 cluster model is 21,664.21, and for the 7 cluster model is 44,982.14. These values reveal that the 2 cluster model has two large clusters that are closer together than the 7 cluster model, which has smaller clusters.

PCA

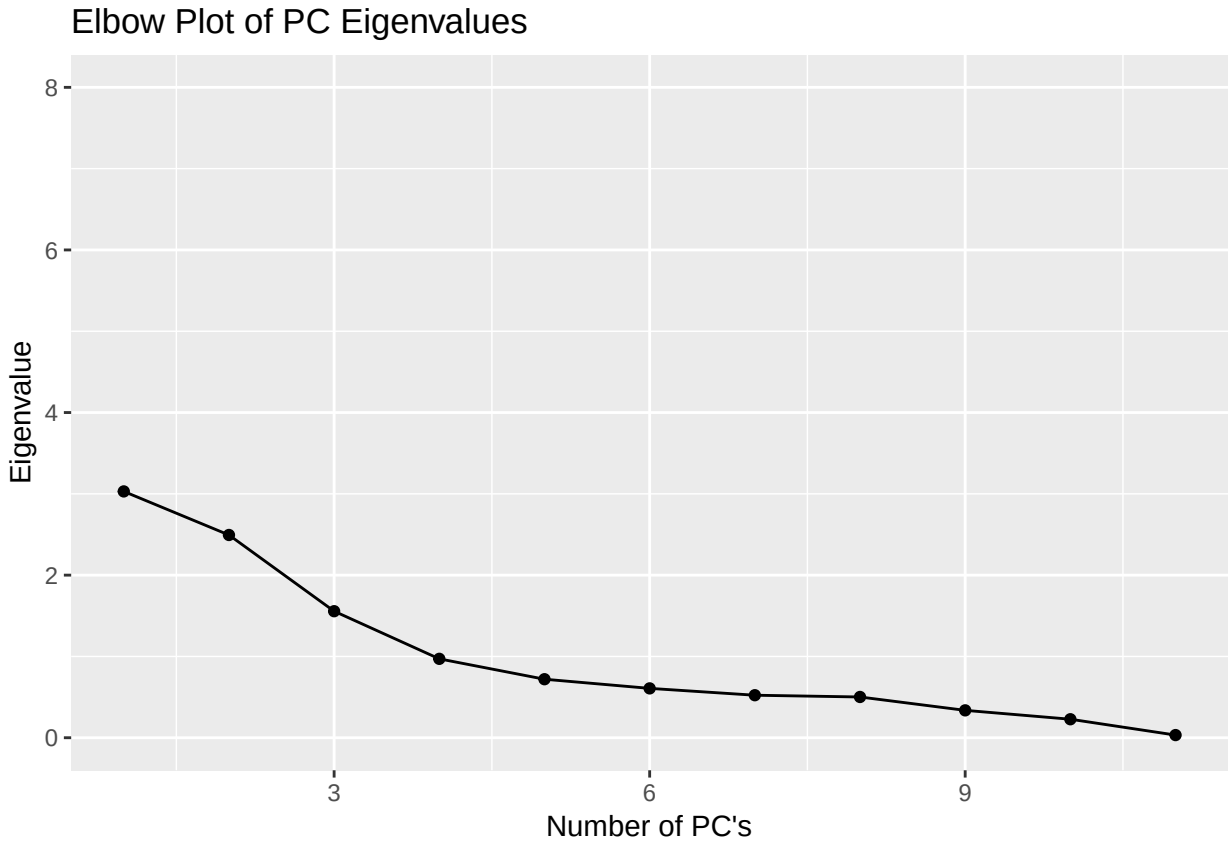
The next portion of our analysis is to see how well Principal Component Analysis does in distinguishing between color of win and quality. We ran the principal components and found that the first two components capture nearly 50% of the variation in the dataset, and the fifth component captures 77% of the cumulative variation in the data.

```
## Importance of components:
##                           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Standard deviation       1.741  1.579  1.248 0.9852 0.8485 0.7793 0.7233
## Proportion of Variance   0.275  0.227  0.141 0.0882 0.0654 0.0552 0.0476
## Cumulative Proportion    0.275  0.502  0.644 0.7319 0.7973 0.8525 0.9001
##                           Comp.8 Comp.9 Comp.10 Comp.11
## Standard deviation       0.7082 0.5805  0.4772 0.18119
## Proportion of Variance 0.0456 0.0306  0.0207 0.00298
## Cumulative Proportion  0.9457 0.9763  0.9970 1.00000
```

The visual below helps to see the cumulative percent of variation captured in all components. We embedded a horizontal line at 80%, making it easier to see that five components captures quite a bit of information.



Another plot that suggests how many principal components to include is an elbow plot of the eigenvalues. The general rule is to not pick any PC's that have an eigenvalue less than one, or where the "elbow" or kink in the plot occurs. Naturally, our elbow plot is not that transparent. By examining this plot, we found that either four or five components is suggested.



We settled on analyzing five principal components, but the most challenging part is left; determining what characteristic of wine each component is capturing. Below we have made a heat map of the loadings of each variable in our data set on the five components. This plot is quite a bit to take in, which is why we are going to explain what it is telling us right now. Dark orange reveals that a particular variable has a very negative loading/correlation with that component. A medium orange reveals that the variable has a more neutral, or close to zero, correlation with that component. A very light yellow or white shade reveals a very positive loading/correlation on that component.

Component 1 : Very negative loading of 'free sulfur dioxide', 'total sulfur dioxide'; moderately negative loading of 'residual sugar'; and high positive loadings of 'volatile acidity', 'chlorides' and 'sulfates'

Component 2 : Very negative loading of 'alcohol' and 'quality'; positive loading of 'residual sugar' and 'total sulfur dioxide'; and very positive loading of 'density'

Component 3 : Very negative loading of 'pH' and 'volatile acid', and very positive loading of 'quality', 'fixed acidity' and 'citric acid'.

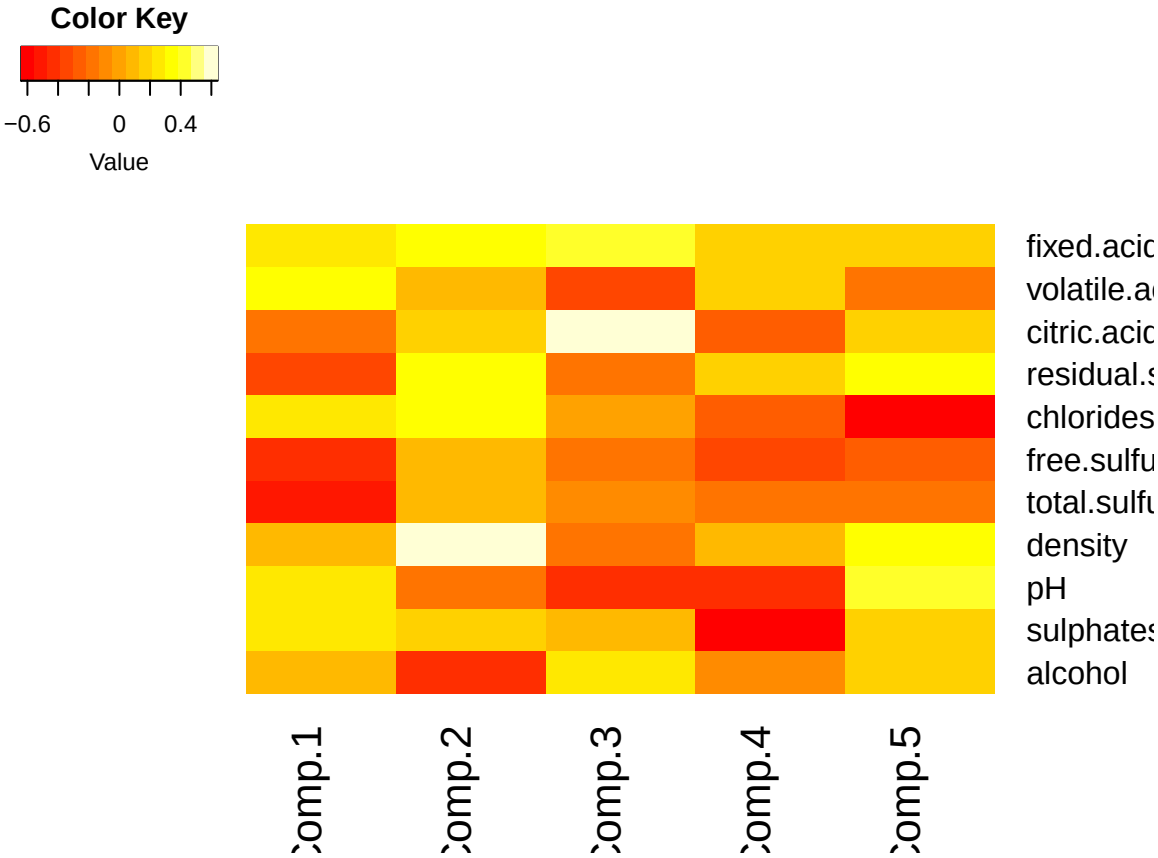
Component 4 : Negative loading of 'quality', 'sulfates', and 'pH'

Component 5 : Positive loading of 'residual sugar', 'density', 'alcohol' and 'quality'

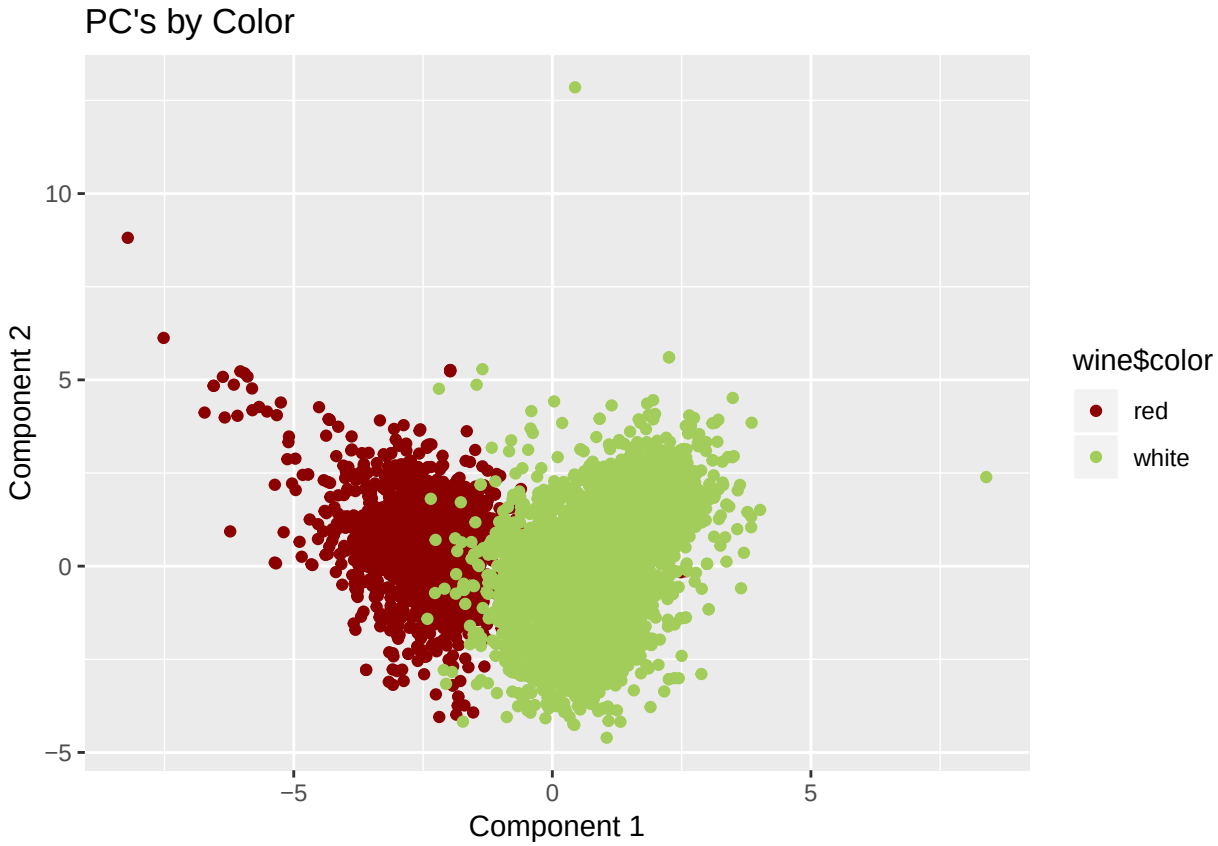
```
## Warning: package 'gplots' was built under R version 3.5.2
```

```
##
## Attaching package: 'gplots'
```

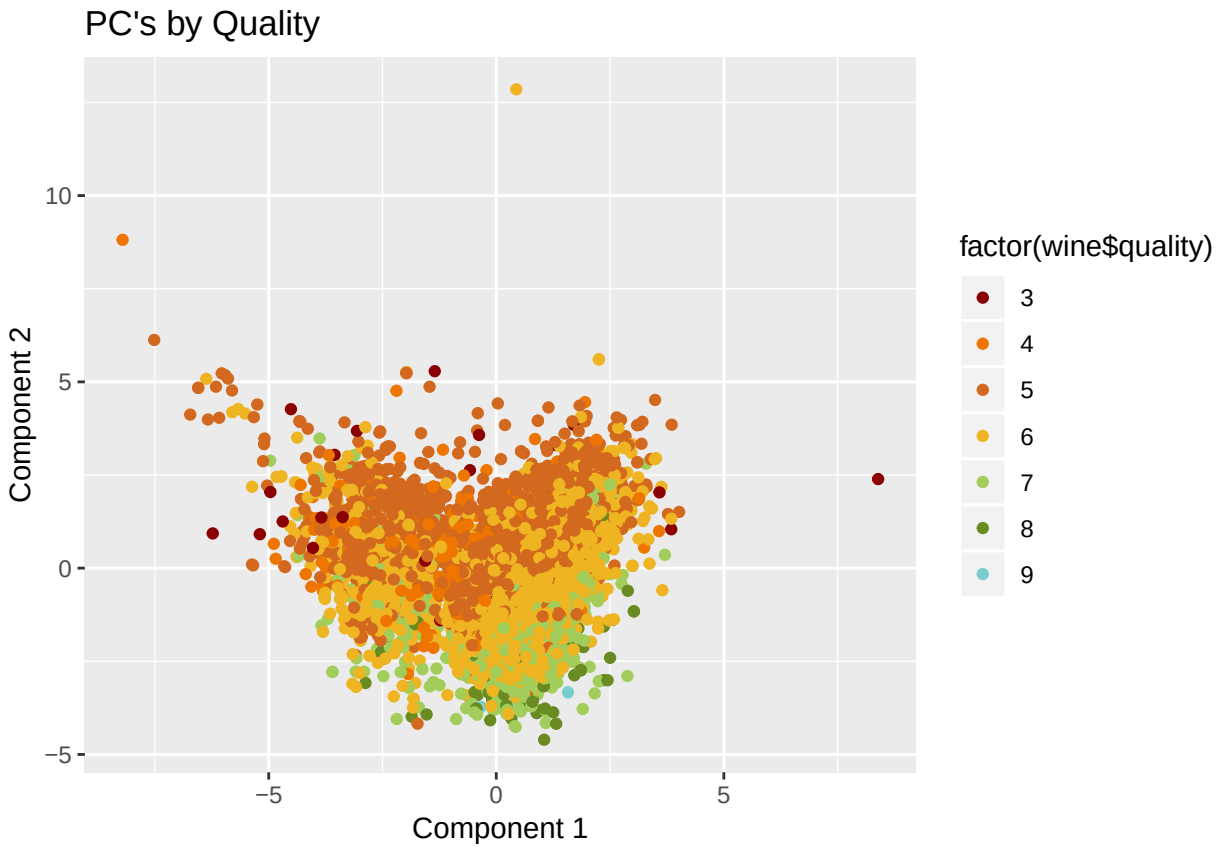
```
## The following object is masked from 'package:stats':
##
##     lowess
```



These findings lead us to think about what these components are explaining, which leads us back to the plots we created earlier in this analysis. We found that white wines have higher levels of 'residual sugar' and 'sulfur dioxide' and lower levels of 'chlorides' and 'sulfates' in comparison to red wines. This leads us to believe that the first component distinguishes between the color wines. We verified this by plotting our first and second components, and coloring each observation by which color of wine it is. We see clear clustering across red and white wines for the first component.



We then want to determine what the second component is suggesting. We believe that it is distinguishes alcohol levels and quality. Keep in mind alcohol levels and quality are about 50% correlated. High quality wines generally have higher 'alcohol', and lower 'density' and 'chloride' levels. This matches up with what the heat map reveals about component 2. To verify our inclinations about component 2, we plot the same plot as above, except the colors now reveal the quality of the wine. We see that higher quality wines tend to have more negative values of component 2, whereas lower and medium quality wines have higher values.



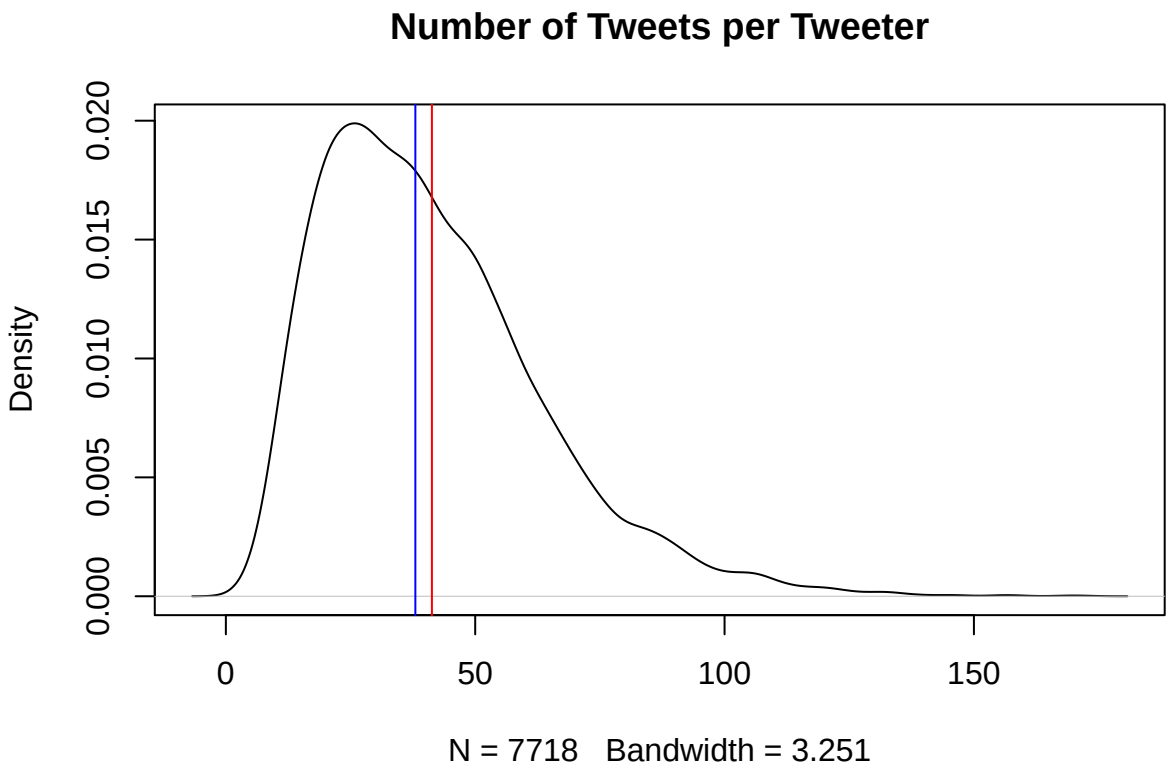
To finalize our thoughts on dimensionality reduction of this wine dataset, we feel that Principal Component Analysis does a much better job in finding a distinction between color of wine and the quality ratings of wine. Just like with any statistical analysis, there are many dynamics and layers to a data set and a question that may not always render textbook results. In the case of this particular data set, we felt like the reason quality rating was so hard to distinguish between was because it was the only variable that was calculated subjectively, possibly causing some bias in the values. Nonetheless, we made out best shot at trying to determine which characteristics of wine are associated with higher quality, and found that PCA is a much better method than K-means ++.

Market Segmentation

This data set contains data from a market-research study for "NutrientH2O" to understand their social-media consumer base better. The data was collected over seven days in June 2014, where each tweet was categorized by content into 32 different categories. We decided to drop individuals who we thought had a high chance of being Twitter bots. Those accounts with over a 20% share of tweets categorized as "spam" or "adult" were designated as likely bots and removed from the dataset.

Summary statistics reveal that the three most tweeted categories are 'health/nutrition', 'politics', and 'cooking'. The three least tweeted categories are 'small business', 'business' and 'eco'.

We see that the number of posts about "NurtientH2O" has high variation across all categories for individual tweeters. The counts appear to have right skew with a mean of 41.3 tweets and median of 38 tweets per tweeter. The most amount of tweets a single twitter user wrote about "NutrientH2O" is 342, whereas the least is 6. There were also 637,835 total tweets made about "NutrientH2O" in the seven week period in June 2014.



In order to determine which market segment "NutrientH2O" should target with their advertising, we used a Principal Component Analysis and Hierarchical Clustering. We feel that our computers were likely to determine which categories cluster together better than assumptions we make about twitter users. Below is a summary of the PCA output, we see that 75% of the variation in tweets is captured by the first 15 components. This is good because 15 is smaller than 35, meaning PCA is an improvement.

To make sense of these components, we decided to look at the loadings of the first four components.

Component 1: Top ten most positive loadings for fashion, cooking, crafts, beauty, family, school, sports_fandom, parenting, religion, and food. The most negative loadings were for adult, online_gaming, college_uni, uncategorized, current_events, tv_film, art, dating, travel, and home_and_garden.

Component 2: Top ten most positive loadings for music, personal_fitness, uncategorized, health_nutrition, chatter, beauty, shopping, fashion, photo_sharing, and cooking. Ten most negative loadings for sports_fandom, religion, parenting, food, family, school, news, automotive, crafts, and politics.

Component 3: Top ten most positive loadings for chatter, college_uni, tv_film, small_business, business, automotive, news, computers, travel, and politics. Ten most negative loadings for health_nutrition, personal_fitness, cooking, beauty, outdoors, fashion, food, religion, parenting, and school.

Component 4: Top ten most positive loadings for music, tv_film, shopping, chatter, fashion, beauty, photo_sharing, sports_playing, online_gaming, and college_uni. Ten most negative loadings for health_nutrition, personal_fitness, outdoors, politics, news, travel, computers, eco, food, and automotive.

From these loadings, we conclude that the first component defines a market segment for families, or in particular a mother. The second component defines a market segment for millennials, or young individuals. The third component defines a market segment for businesses or professional individuals. The fourth and last component defines a market segment for individuals younger than millennials, or gen-Z.

The final portion of our analysis using the algorithm of the HCPC (hierarchical clustering principal component) method. This method includes the following steps:

```
1) Compute principal components
```

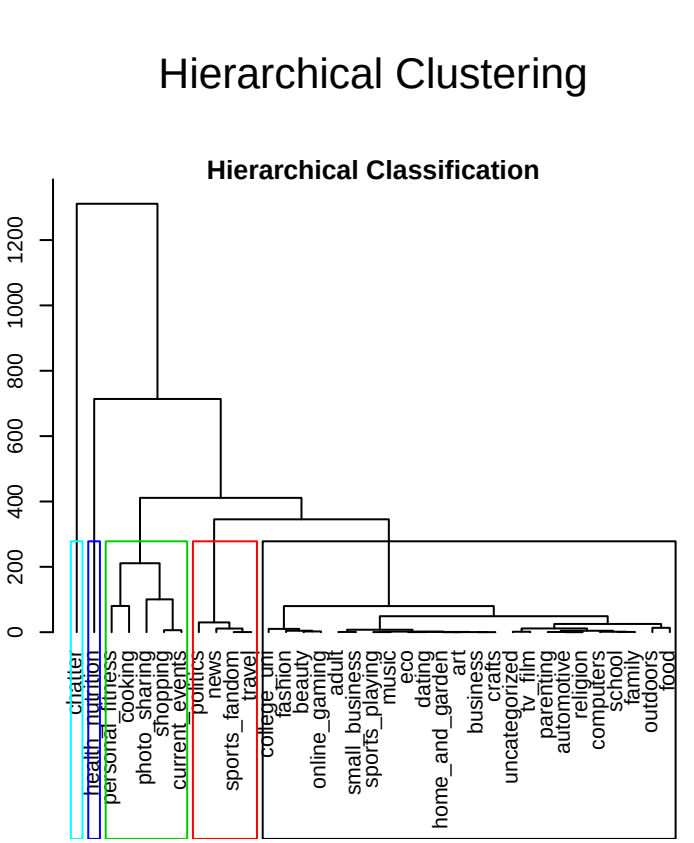
2) Compute hierarchical clustering: Hierarchical clustering is performed using the Ward's criterion on the selected principal components.

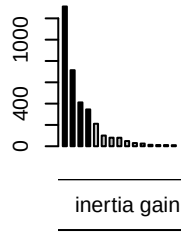
3) Choose the number of clusters based on the hierarchical tree: An initial partitioning is performed by cutting the hierarchical tree.

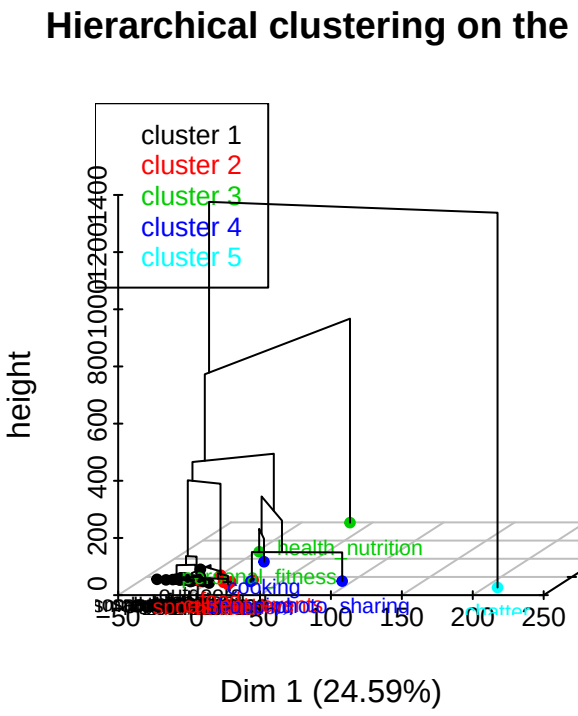
4) Perform K-means clustering to improve the initial partition obtained from hierarchical clustering.

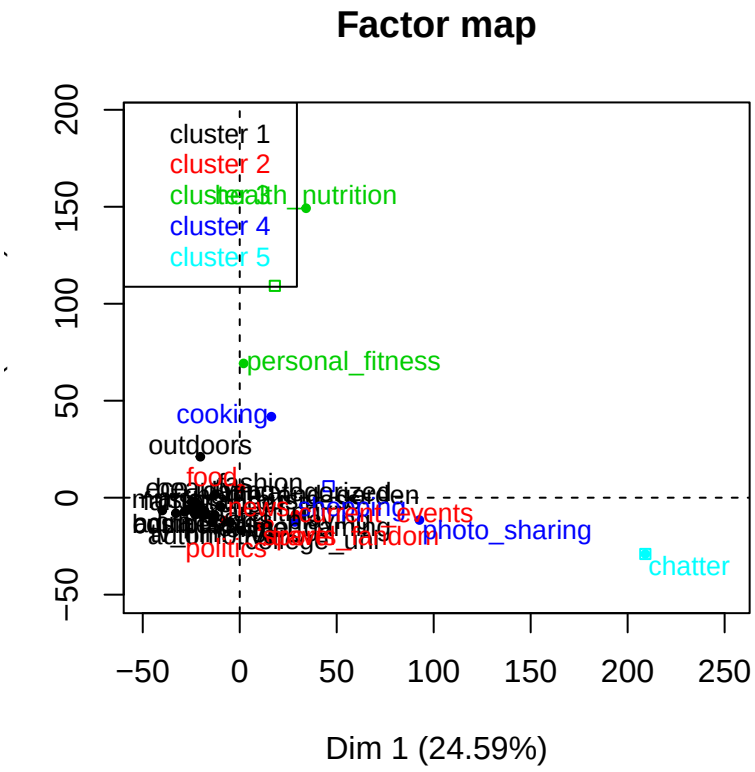
The dendrogram below shows how this method split up the PC's and the hierarchical clustering of all the categories. The second and third plot depict the hierarchical clusters in the factor space and the 2D factor space respectively.

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```









To maximally appeal to each market segment "NutrientH2O" should advertise in categories that are specific to each principal component, or market segment. They should not pay too much attention to tweet that at categorized as "chatter", but should not completely disregard those tweets, as they show up in many early components. The market segment for families/mothers should be "NutrientH2O"'s main focus, as their actions on twitter explain most of the traffic "NutrientH2O" has their account compared to all other market groups. If "NutrientH2O" were to focus on one category, 'nutrition and health' would gain the most attention as far as tweets from followers. Each market segment will take to advertisements differently, therefore "NutrientH2O" should make the focus of advertisements specific to the key interests of each group. For example, advertisements for schooling and food for mothers, beauty and nutrition for millenials, business and politics for business men, and music and photos for gen-Z.

Association Rules for Grocery Purchases

We downloaded the list of grocery items, cobbled together a few utilities, to ultimatley have a data frame that can work with our Apriori algorithm. After testing out several sensible thresholds we let the Apriori algorithm find association rules for a support and confidence of 0.05% and 20% respectively. This led to us identifying 435 association rules. They are listed below by decreasing lift.

Interestingness of a pattern is expressed in terms of how it affects the belief system. We found with our support and confidence levels, there were many rules that seem intuitive- a shopper purchasing both bottled beer and red/blush wine is 57 times more likely to purchase liquor. Similarly, a shopper purchasing both bottled beer and liquor wine is 30 times more likely to purchase red/blush wine. Most of our discovered item sets are logical and unsurprising. Some notable rules include rule 6, which implies a shopper who purchases berries and pork is 10 times more likely to purchase beef. Rule 12 implies that a shopper who purchases root vegetables and turkey will be 7 times more likely to also purchase tropical fruit. Rule 34 implies that a shopper who purachases potato products is 5 times more likely to also purchase a pastry. We see that some rules, such as rule 5, are unrelated to all others. We also see an exclusive relationship between alcoholic drinks and find their only to food items are through soda and bottled water rules.

We have included a net that shows some of the most prominent rules and the edges between goods. We chose

to include 49 rules with the highest lift that have a confidence over 10% and support over 0.05%. While the graph is sometimes hard to follow, it can reveal quite a bit of information about the inner workings of grocery stores.