Problem Set 4

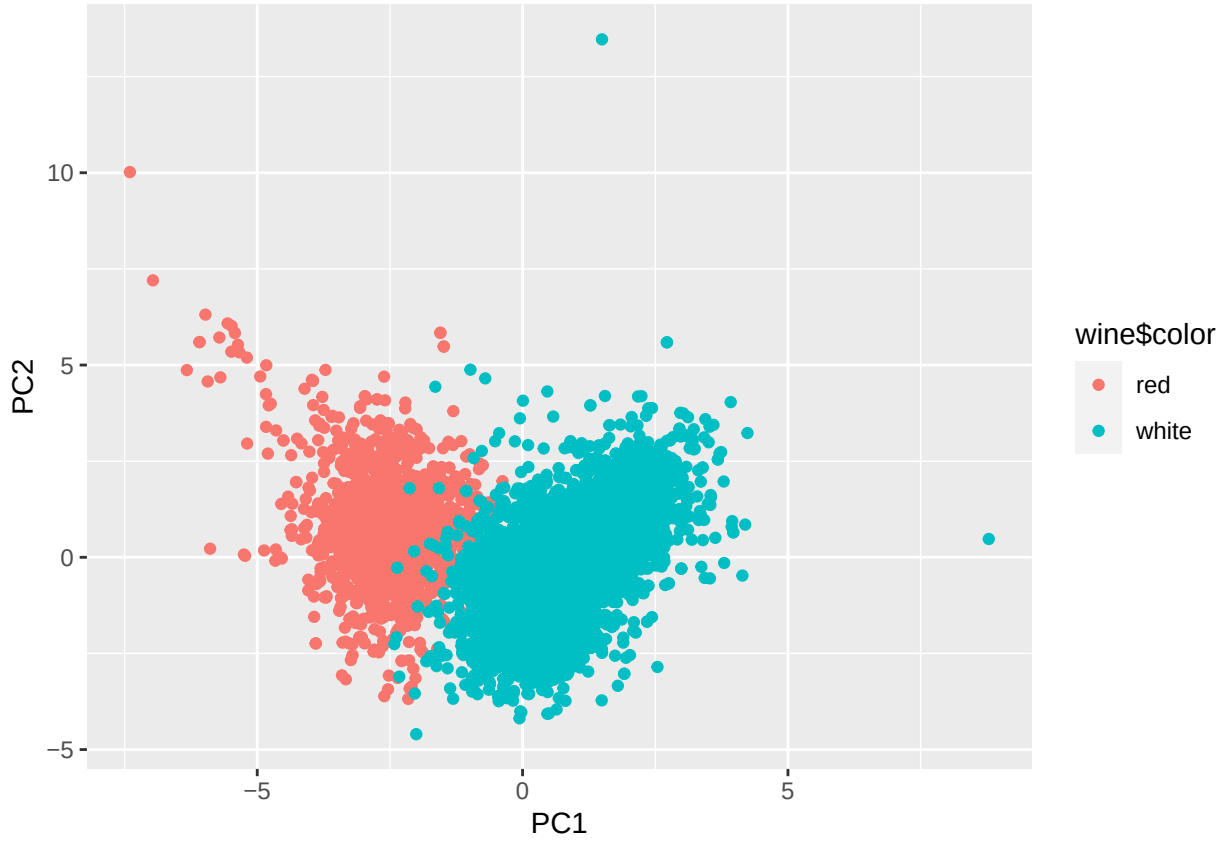
By Jae You (jhy479)

1)Clustering and PCA

Below are the Principle Component Analysis on the 11 chemical properties as well as the standard deviation and variance of the first five principle components.

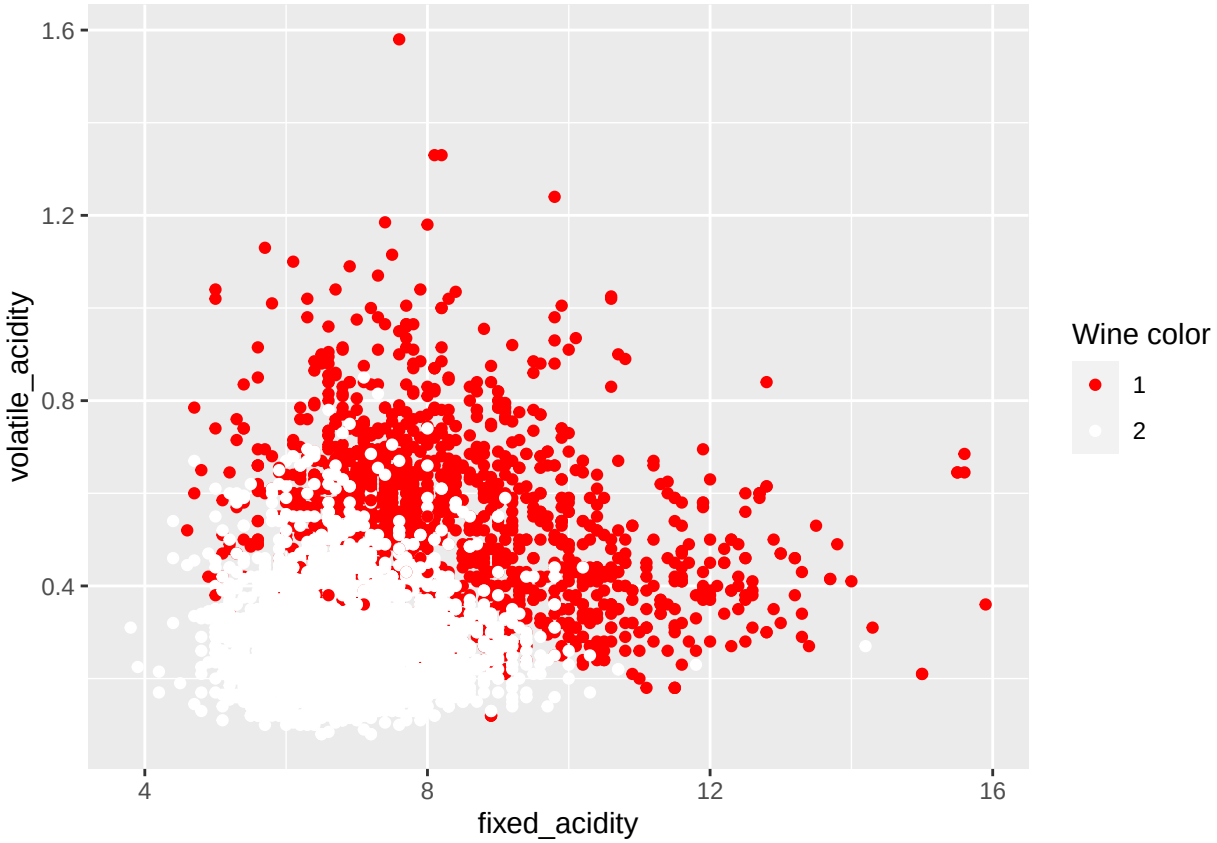
```
##                              PC1       PC2       PC3       PC4       PC5
## fixed.acidity        -0.238799  0.336355 -0.434301  0.164346 -0.147480
## volatile.acidity     -0.380757  0.117550  0.307259  0.212785  0.151456
## citric.acid           0.152388  0.183299 -0.590570 -0.264300 -0.155349
## residual.sugar        0.345920  0.329914  0.164688  0.167443 -0.353362
## chlorides            -0.290113  0.315258  0.016679 -0.244744  0.614391
## free.sulfur.dioxide   0.430914  0.071933  0.134224 -0.357279  0.223532
## total.sulfur.dioxide  0.487418  0.087266  0.107462 -0.208420  0.158134
## density              -0.044937  0.584037  0.175606  0.072725 -0.306561
## pH                   -0.218686 -0.155869  0.455324 -0.414551 -0.453376
## sulphates            -0.294135  0.191716 -0.070042 -0.640536 -0.136577
## alcohol              -0.106437 -0.465058 -0.261101 -0.106803 -0.188892
```

```
## Importance of first k=5 (out of 11) components:
##                            PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732
```



The plot of PC1 and PC2 for the color of the wine is pretty accurate at distinguishing red and white wines, with an even mix in both groups.

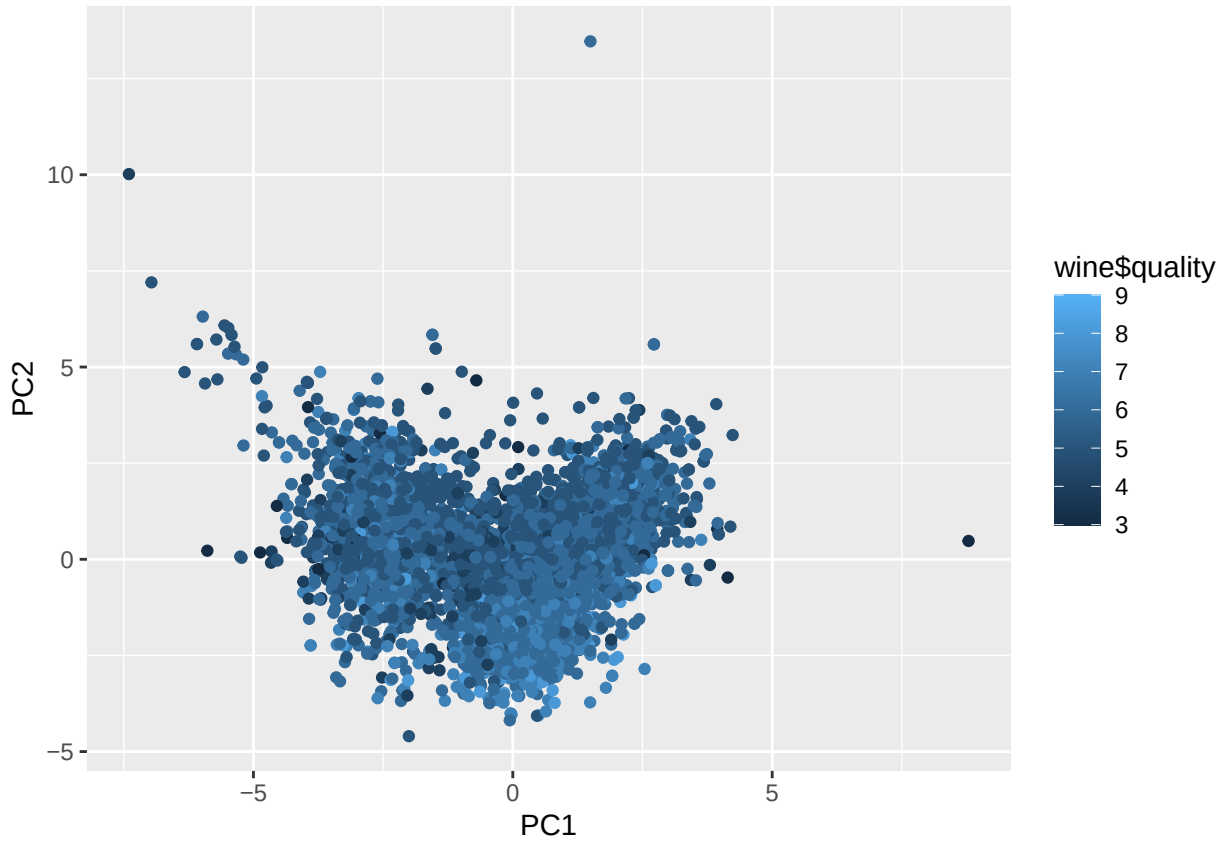
Below is the plot done using kmeans as the clustering method. I plotted the variables fixed.acidity and volatile.acidity.



I noticed that the colors encroached on each other and thus did not perform a solid 2-cluster group. There are some points far inside the white "territory" that are identified as red wine when they fall far outside of the red cluster.

In conclusion, I would prefer the PCA method for distinguishing between the two wine colors. Not only was it simple to use, but also requires far less fine-tuning to get closer to the PCA results using kmeans.

However, the PCA method is not great when it comes to characterizing wine quality. Below is the plot of PC1 and PC2 for the wine quality.

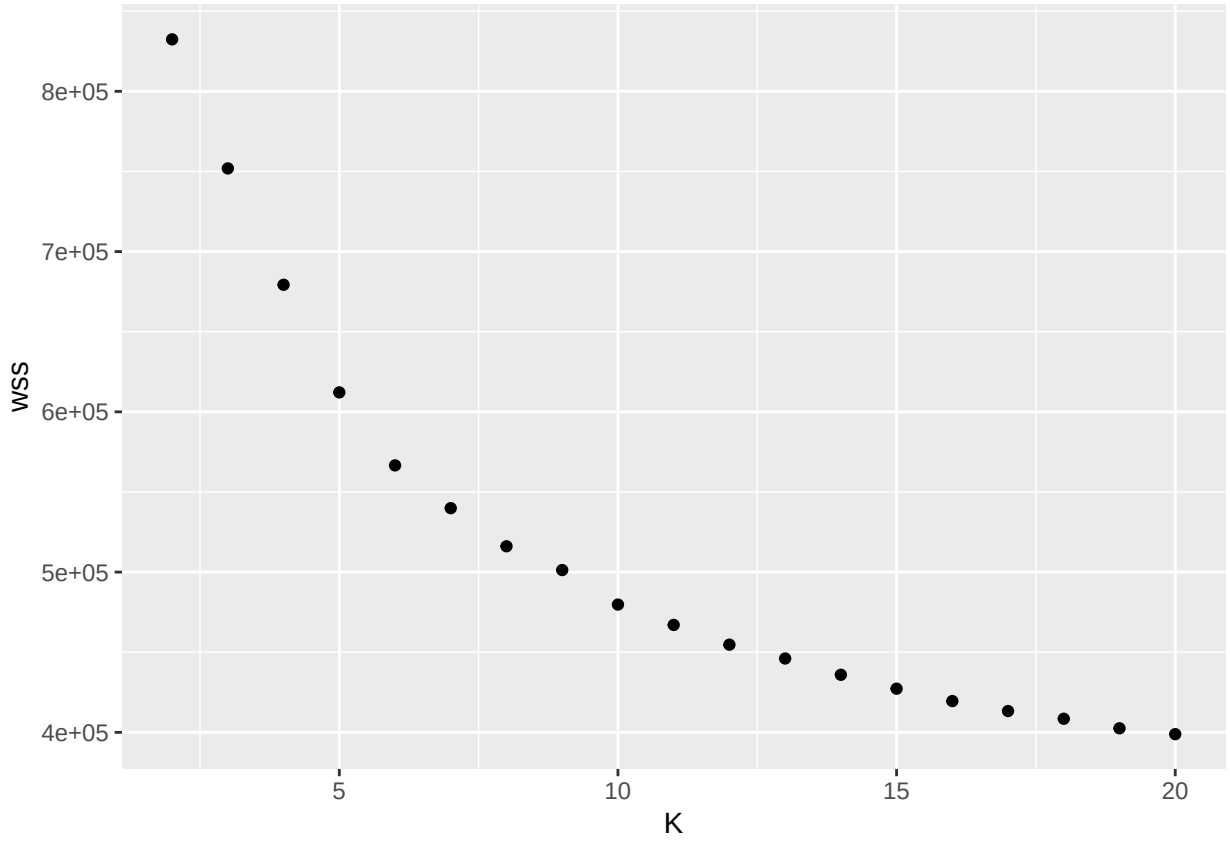


The plot for PCA is not good at distinguishing the wine qualities and it seems like there is too much noise.

2) Market segmentation

I dropped the X column because it was a non-numeric column that listed out the people who were tweeting. Additionally, I removed the chatter variable as annotators may use this label along with the uncategorized label, so I kept the "official" variable that was recommended to the annotators to use. Lastly, I removed the spam and adult(pornographic) variables since there may be some "bot tweets" that may slip through into the data set. As such, it may not define the market segment accurately.

I decided to use kmeans to define the market segments. I started with the elbow method in order to determine the number of clusters.



Here, I chose to consider K=7 for 7 clusters.

```
##                       one
## health_nutrition 58.138654
## personal_fitness 16.312589
## cooking          13.550406
## photo_sharing     9.508657
## politics          5.804076

##                       two
## college_uni      32.807116
## online_gaming    29.323847
## photo_sharing     9.955952
## health_nutrition  9.517856
## cooking           7.283567

##                     three
## health_nutrition 6.767523
## photo_sharing    6.497171
## politics         4.694676
## cooking          4.668731
## college_uni      4.131760
```

```
##                       four
## sports_fandom    15.018589
## religion         11.387253
## food              9.644803
## health_nutrition  9.492259
## photo_sharing     8.893450

##                       five
## photo_sharing    20.438176
## shopping          8.155660
## health_nutrition  8.026214
## cooking           6.068330
## politics          6.061371

##                        six
## politics         31.373096
## travel           15.759790
## news             11.985397
## photo_sharing     9.011993
## health_nutrition  8.879562

##                      seven
## cooking          43.130873
## photo_sharing    19.096445
## health_nutrition 11.760328
## fashion          11.359945
## politics          5.949792
```

To NutrientH20, I present my definition of the market segment as 7 clusters.

The first cluster seems to consist of older adults who wish to stay fit and healthy, as seen by how they are involved in politics while also dominating the health and exercise aspects.

The second cluster consists of students in college who wish to stay healthy.

The third has adults who post frequently on the media about what they cook and wear, presumably the female gender.

The fourth cluster is interested in the current state of the world and keeps track of politics, the traveling status, and the news.

The fifth cluster seems to be those interested in shopping and cooking healthy foods as well as uploading photos online.

The sixth are religious adults who have an interest in sports and food. As it is specifically food and not cooking, this cluster may be providing for many people, perhaps buying bread in bulk for church or even just providing nutritious food to friends and family on a daily basis.

The seventh cluster consists of young college students who are highly interested in games and their phones, as well as those who are concerned about how healthy their diet is. We may be able to assume that both students who stay at their computer for long periods of time and busy college students may have unhealthy diets or consume fast foods or delivery services frequently, and are looking to improve their nutrition standards.

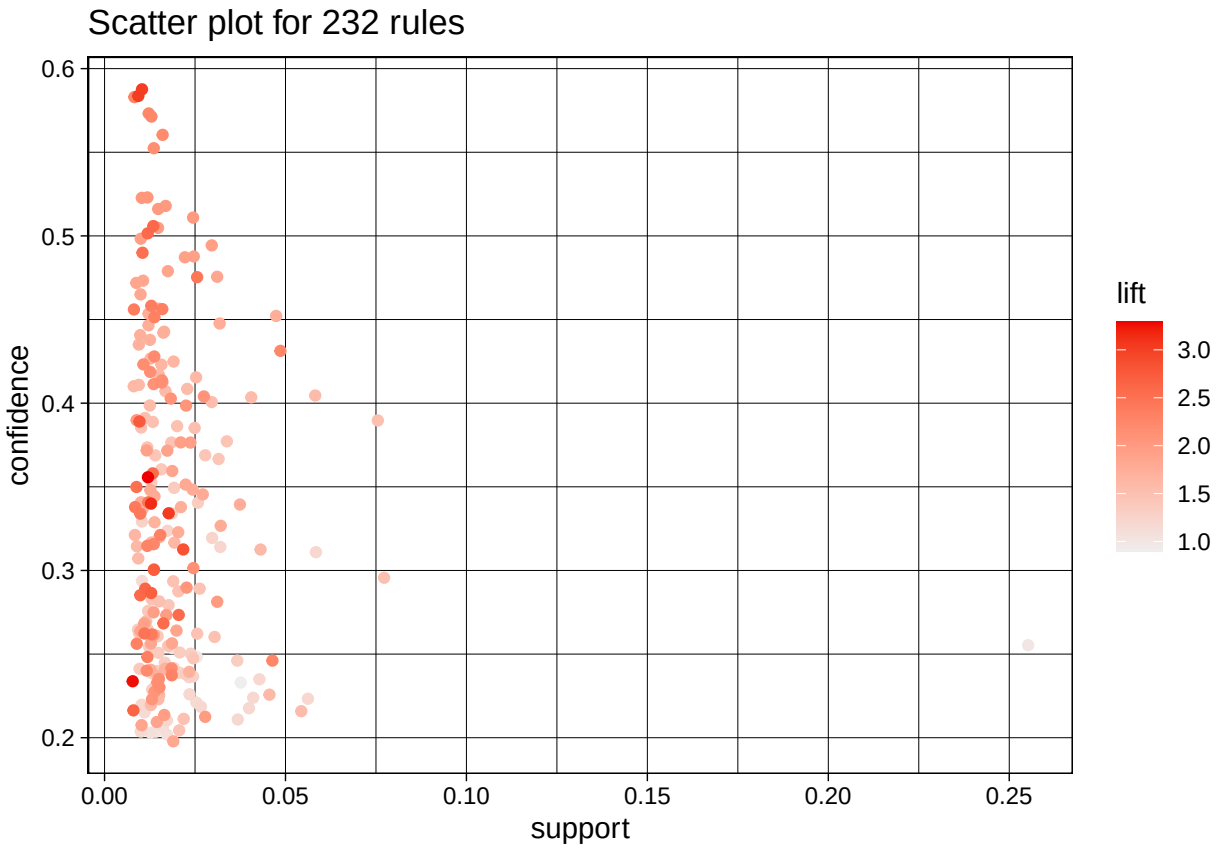
In conclusion, there are seven distinct market segments, which can better help NutrientH20 appeal to each segment.

3) Association rules for grocery purchases

I used association rule mining to find out the rules for grocery baskets.

By lowering the support level below 0.1, I was able to have more than 1 rule occur.

Below is the scatter plot of 232 rules.



I filtered more rules out by targeting those with a lift greater than 2 and looked at the first 10 rules.

```
##                      lhs                    rhs    support confidence   coverage
## [1]              {onions} => {other vegetables} 0.01423488  0.4590164 0.03101169
## [2]             {berries} =>           {yogurt} 0.01057448  0.3180428 0.03324860
## [3]      {hamburger meat} => {other vegetables} 0.01382816  0.4159021 0.03324860
## [4]       {cream cheese } =>           {yogurt} 0.01240468  0.3128205 0.03965430
## [5]             {chicken} =>  {root vegetables} 0.01087951  0.2535545 0.04290798
## [6]             {chicken} => {other vegetables} 0.01789527  0.4170616 0.04290798
## [7]  {frozen vegetables} =>  {root vegetables} 0.01159126  0.2410148 0.04809354
## [8]                {beef} =>  {root vegetables} 0.01738688  0.3313953 0.05246568
## [9]                {curd} =>           {yogurt} 0.01728521  0.3244275 0.05327911
## [10]               {pork} =>  {root vegetables} 0.01362481  0.2363316 0.05765125
##          lift count
## [1]  2.372268   140
## [2]  2.279848   104
## [3]  2.149447   136
## [4]  2.242412   122
## [5]  2.326221   107
## [6]  2.155439   176
## [7]  2.211176   114
## [8]  3.040367   171
```

```
## [9]  2.325615   170
## [10] 2.168210   134
```

From the subset, I noticed that meats such as beef, pork, and chicken are bought with root vegetables, dairy such as curd and cheese are paired with yogurt as well as berry-flavored yogurt, and vegetables and onions are paired with other vegetables.

Conceptually, these rules make sense, as meats are paired with vegetables (a hamburger isn't complete without lettuce, onions, pickles and so on), and vegetables go well with other vegetables, such as for stew, curry, and so on. As for berries and yogurt, blueberry greek yogurt is very delicious.