White Wine Quality Kmeans Investigation

Kevin McBeth

March 6, 2018

Introduction

The white wine data set, obtainable at https://archive.ics.uci.edu/ml/datasets/Wine+Quality, examines the quality of white wine in northern Portugal based on characterstics such as residual sugar, pH, density, sulfur dioxide, chlorides, and alcohol content. The purpose of this assignment is to explore kmeans on the features to help explore the data.

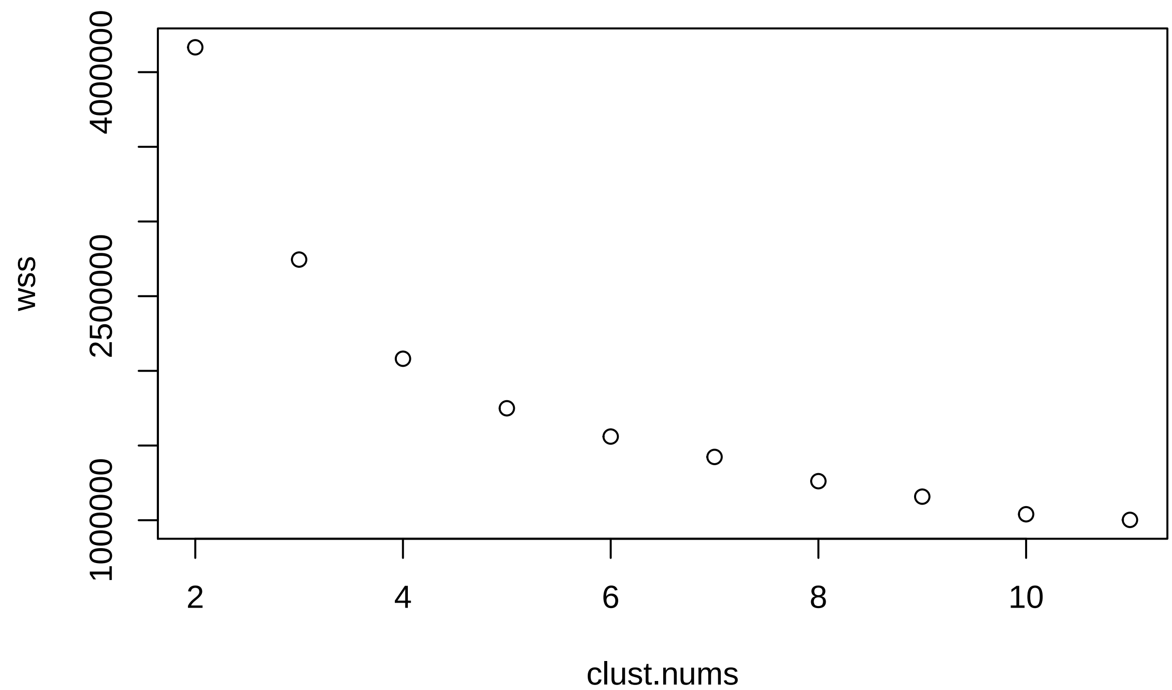
Through kmeans fitting, a cluster number of FIVE gave the best performance when optimizing the maximum of the silhoutte score and the minimizing of the within sum of squares (WSS). The clustering primarily occurs around total sulfur dioxide and free sulfur dioxide, with some clustering based upon sugar and alcohol content.

Cluster Selection

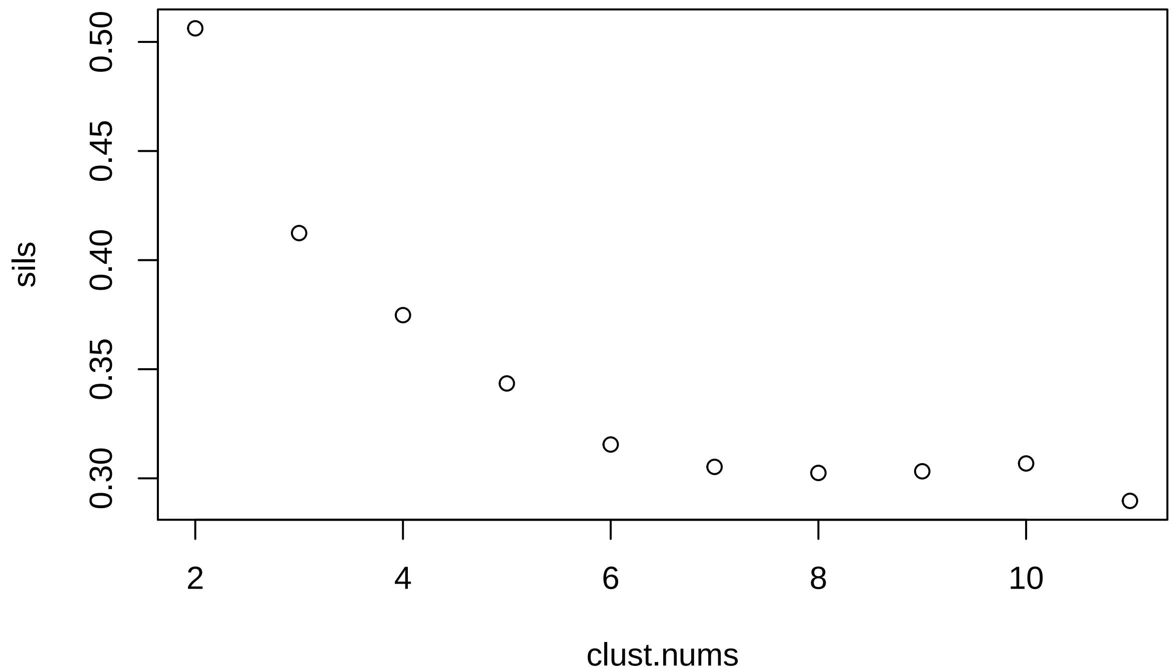
Selecting the number of clusters is the most important hyperparameter function to select for unsupervised learning with kmeans. To do this, we ran a loop to obtain the within sum of squares and the silhoutte score for each number of clusters. During this investigation, we determined seven clusters to be ideal, as it is at the inflection point of the within sum of squares vs clusters plot, and the inflection point of the silhoutte score vs clusters plot.

This is important due to the silhoutte score representing how well seperated clusters are, which our goal was to maximize, while minimizing the within sum of squares metric which measures variance between cluster points and the centers. One issue with only focusing on the within sum of squares is that it tends to overfit when attempting to use new data. Given that our data set is only 4898 observations large, it is improbable that enough significant data was collected to represent the entire range of possiblities for white wine in northern Portugal.

```r
wss <- c()
sils <- c()
clust.nums <- 2:11
for (i in clust.nums) {
  km <- kmeans(wine.dt.notargets, centers = i)
  sil<- silhouette(km$cluster, dist(wine.dt.notargets))
  sils <- c(sils, summary(sil)$si.summary[4]) #gets the mean silhoutte score
  wss <- c(wss, km$tot.withinss)
}

plot(clust.nums, wss)
```




```r
plot(clust.nums, sils)
```




```r
#5 clusters works well due to high silhoutte score and low wss.
km <- kmeans(wine.dt.notargets, centers=5)
```

Cluster Averages

Investingating the cluster averages, we determined that the largest variance is in total and free Sulfur Dioxide for the cluster centers, having a max/min = 2.86. I consider these two variables to be highly similar based on a correlation plot later in this document. Looking at the plots section of this paper shows that the algorithm assigns clusters primarily by amount of sulfur dioxide. This feature is not well correlated to a potential target variable, quality. The highest correlated component of our features compared to quality is alcohol, which has a more moderate spread at 1.18 when looking at the output of our summary statistics on the centers of our clusters.

```r
km$centers
```

```
##   fixed acidity volatile acidity citric acid residual sugar  chlorides
## 1      6.813370        0.2799025   0.3158357       3.450557 0.04015042
## 2      6.777090        0.2700066   0.3230678       4.734200 0.04193153
## 3      6.960524        0.2855388   0.3537160       8.824018 0.05114804
## 4      6.840321        0.2724726   0.3359326       7.006975 0.04732367
## 5      7.010969        0.3073980   0.3557908      10.033801 0.05228571
##   free sulfur dioxide total sulfur dioxide   density       pH sulphates
## 1            18.86212             77.53482 0.9918326 3.175864 0.4691086
## 2            28.24753            113.13989 0.9927783 3.191257 0.4843779
## 3            47.02971            179.05690 0.9959135 3.183112 0.5075831
## 4            37.70415            145.32367 0.9944275 3.198770 0.4855643
```

```
## 5            55.29847           221.74617 0.9968072 3.178265 0.5180357
##     alcohol
## 1 11.255687
## 2 10.959480
## 3  9.831101
## 4 10.397542
## 5  9.541582
```

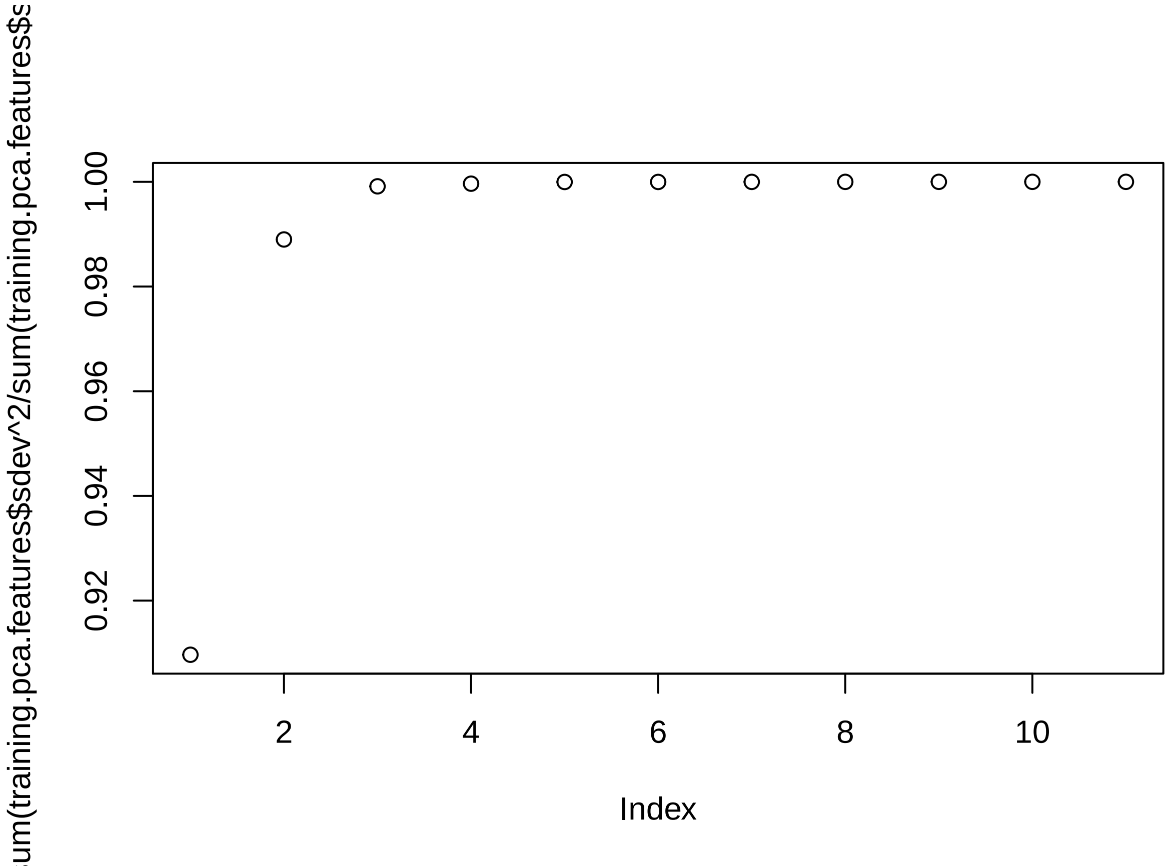
```r
summary(km$centers)
```

```
##  fixed acidity   volatile acidity  citric acid     residual sugar
##  Min.   :6.777   Min.   :0.2700   Min.   :0.3158   Min.   : 3.451
##  1st Qu.:6.813   1st Qu.:0.2725   1st Qu.:0.3231   1st Qu.: 4.734
##  Median :6.840   Median :0.2799   Median :0.3359   Median : 7.007
##  Mean   :6.880   Mean   :0.2831   Mean   :0.3369   Mean   : 6.810
##  3rd Qu.:6.961   3rd Qu.:0.2855   3rd Qu.:0.3537   3rd Qu.: 8.824
##  Max.   :7.011   Max.   :0.3074   Max.   :0.3558   Max.   :10.034
##    chlorides       free sulfur dioxide total sulfur dioxide
##  Min.   :0.04015   Min.   :18.86       Min.   : 77.53
##  1st Qu.:0.04193   1st Qu.:28.25       1st Qu.:113.14
##  Median :0.04732   Median :37.70       Median :145.32
##  Mean   :0.04657   Mean   :37.43       Mean   :147.36
##  3rd Qu.:0.05115   3rd Qu.:47.03       3rd Qu.:179.06
##  Max.   :0.05229   Max.   :55.30       Max.   :221.75
##     density            pH          sulphates         alcohol
##  Min.   :0.9918   Min.   :3.176   Min.   :0.4691   Min.   : 9.542
##  1st Qu.:0.9928   1st Qu.:3.178   1st Qu.:0.4844   1st Qu.: 9.831
##  Median :0.9944   Median :3.183   Median :0.4856   Median :10.398
##  Mean   :0.9944   Mean   :3.185   Mean   :0.4929   Mean   :10.397
##  3rd Qu.:0.9959   3rd Qu.:3.191   3rd Qu.:0.5076   3rd Qu.:10.959
##  Max.   :0.9968   Max.   :3.199   Max.   :0.5180   Max.   :11.256
```

PCA Analysis

Conducting a basic PCA test on our data shows that the most important features in describing the variance of our data set are sulfur dioxide, alcohol, and sugar. We discussed the first two features as having the majority of impact on our clustering algorithm. This checks with the cumulative sum, as the principal components of sulfur dioxide and alcohol describe 99% of the variance of our data set. That is, principal components one and two make up 99% of the variance, and PC1 and PC2 are dominated by free and total sulfur dioxide and alcohol.

```r
training.pca.features <- prcomp(wine.dt.notargets)
plot(cumsum(training.pca.features$sdev^2 / sum(training.pca.features$sdev^2)))
```




```
training.pca.features
```

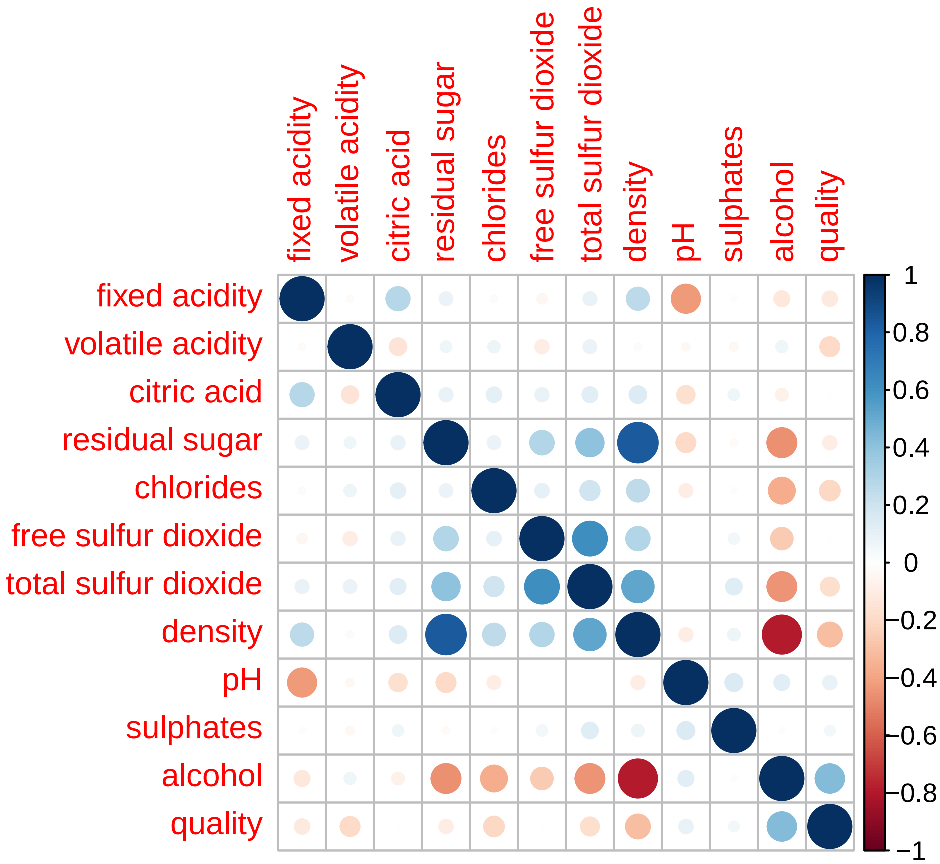
```
## Standard deviations (1, .., p=11):
##  [1] 4.394899e+01 1.297894e+01 4.643382e+00 1.036542e+00 8.286788e-01
##  [6] 1.361319e-01 1.195400e-01 1.069865e-01 9.296260e-02 1.990118e-02
## [11] 5.628848e-04
##
## Rotation (n x k) = (11 x 11):
##                               PC1           PC2           PC3
## fixed acidity        -1.544402e-03 -9.163498e-03 -1.290026e-02
## volatile acidity     -1.690037e-04 -1.545470e-03 -9.288874e-04
## citric acid          -3.386506e-04  1.403069e-04 -1.258444e-03
## residual sugar       -4.732753e-02  1.494318e-02 -9.951917e-01
## chlorides            -9.757405e-05 -7.182998e-05 -7.849881e-05
## free sulfur dioxide  -2.618770e-01  9.646854e-01  2.639318e-02
## total sulfur dioxide -9.638576e-01 -2.627369e-01  4.278881e-02
## density              -3.596983e-05 -1.836319e-05 -4.468979e-04
## pH                   -3.384655e-06 -4.169856e-05  7.017342e-03
## sulphates            -3.409028e-04 -3.611112e-04  2.142053e-03
## alcohol               1.250375e-02  6.455196e-03  8.272268e-02
##                              PC4           PC5           PC6
## fixed acidity         0.147657857 -0.9849646813 -0.0734101708
## volatile acidity     -0.015451710  0.0039780757  0.1066747709
## citric acid           0.005004529 -0.0416921666  0.0166103959
## residual sugar       -0.084200484  0.0008080231 -0.0060314933
## chlorides             0.006573232  0.0014977852  0.0142782518
```

```
## free sulfur dioxide   0.006381109 -0.0078746905 -0.0004473015
## total sulfur dioxide -0.010613506  0.0017527656  0.0007002095
## density               0.001151657 -0.0003284420 -0.0036344486
## pH                   -0.017027136  0.0755059384 -0.9282430660
## sulphates            -0.002600913  0.0035382620 -0.3479102306
## alcohol              -0.985062967 -0.1493611788  0.0049668641
##                               PC7           PC8           PC9
## fixed acidity        -4.866972e-02 -0.0049631343 -0.0010124310
## volatile acidity     -3.247150e-01  0.1622433608  0.9251575355
## citric acid           8.616026e-01 -0.3523204110  0.3619176845
## residual sugar       -1.540992e-04 -0.0001407293 -0.0017096094
## chlorides             1.293099e-02  0.0014986845  0.0309588623
## free sulfur dioxide  -9.947390e-04  0.0004937578  0.0013462404
## total sulfur dioxide  3.447211e-05 -0.0003549051 -0.0007640255
## density               8.362378e-05  0.0001295313  0.0012631325
## pH                   -1.512742e-01 -0.3122773702  0.1084427293
## sulphates             3.560404e-01  0.8671788798  0.0129083897
## alcohol               3.942837e-03 -0.0019414812 -0.0143625037
##                              PC10          PC11
## fixed acidity         2.217085e-03  7.708019e-04
## volatile acidity     -2.613508e-02  6.331732e-04
## citric acid          -2.204267e-02  3.395396e-04
## residual sugar        6.105257e-04  3.715513e-04
## chlorides             9.993001e-01  4.696472e-03
## free sulfur dioxide  -7.478473e-06 -6.883221e-06
## total sulfur dioxide -2.871786e-05  3.885406e-06
## density               4.704068e-03 -9.999807e-01
## pH                    1.231226e-02  3.467970e-03
## sulphates            -1.331315e-03  1.411536e-03
## alcohol               7.042806e-03 -1.125942e-03
```

Correlation Plot

When looking at correlated variables, we confirmed the suspicion that free and total sulfur dioxide are strongly correlated. Of interesting note, the strongest variable correlated to the quality of white wine is alcohol's positive correlation. Other interesting things to note from the plot are densities correlation with residual sugar and alcohol, something which makes physical and chemical sense.

```
corrplot(cor(wine.dt))
```

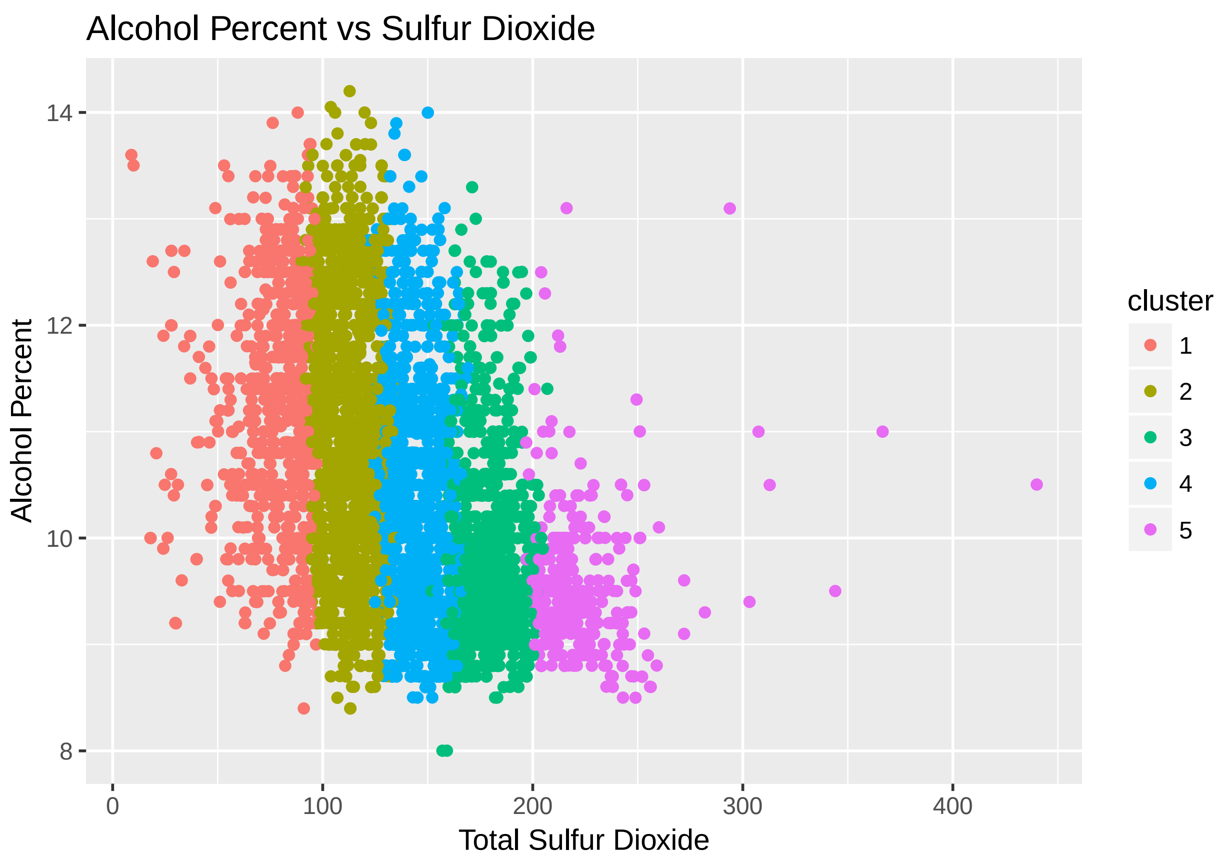


Final Plots

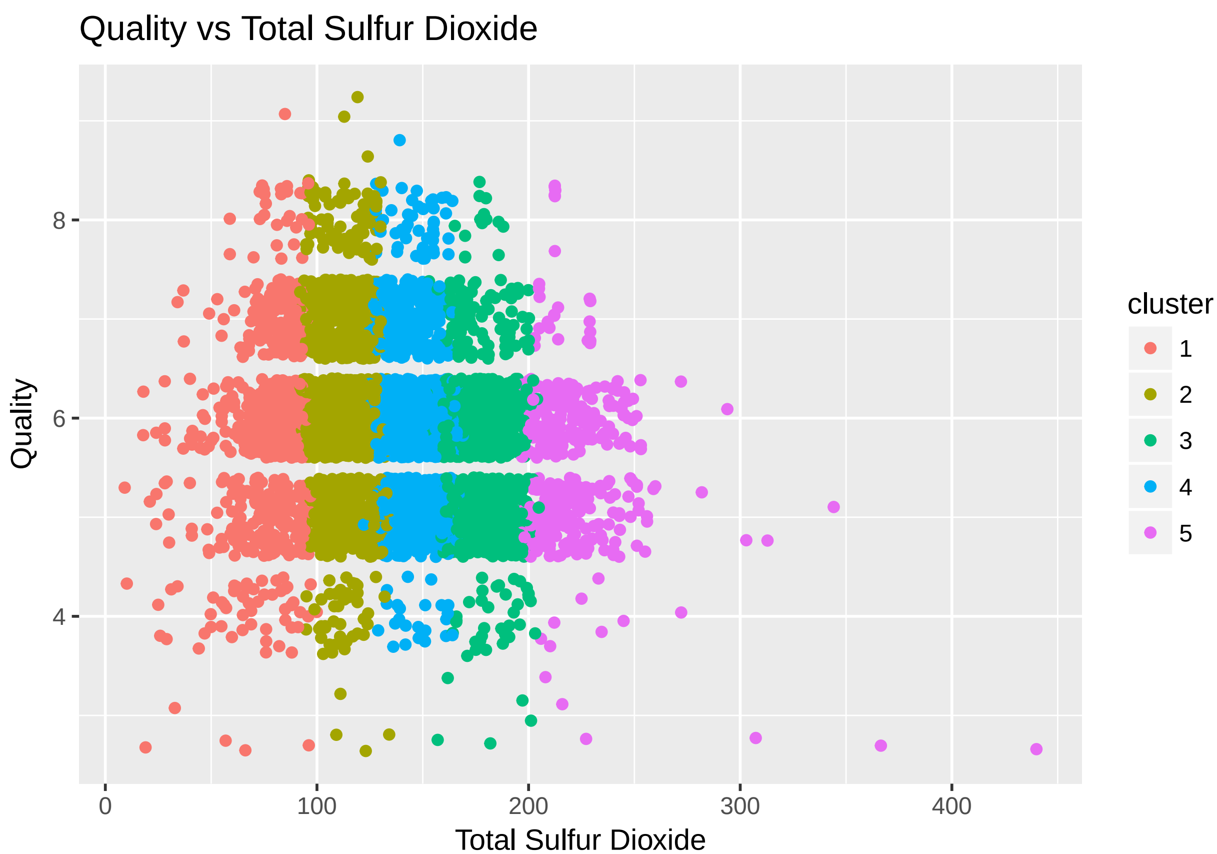
Investigating the two plots below, one sees the strong tendency for the kmeans algorithm to cluster around different amounts of total sulfur dioxide. This was expected from both our principal component analysis, and the investigation into the cluster centers in the Cluster Averages section. Figure two below shows the lack of correlation between wine quality and total sulfur dioxide despite the emphasis on clustering around the high variance and leveraged variable of total sulfur dioxide.

Looking at the most strongly correlated variable, alcohol, one can see little grouping of the clusters due to the domination by total sulfur dioxide, but a slight postive overall correlation between alcohol and quality.

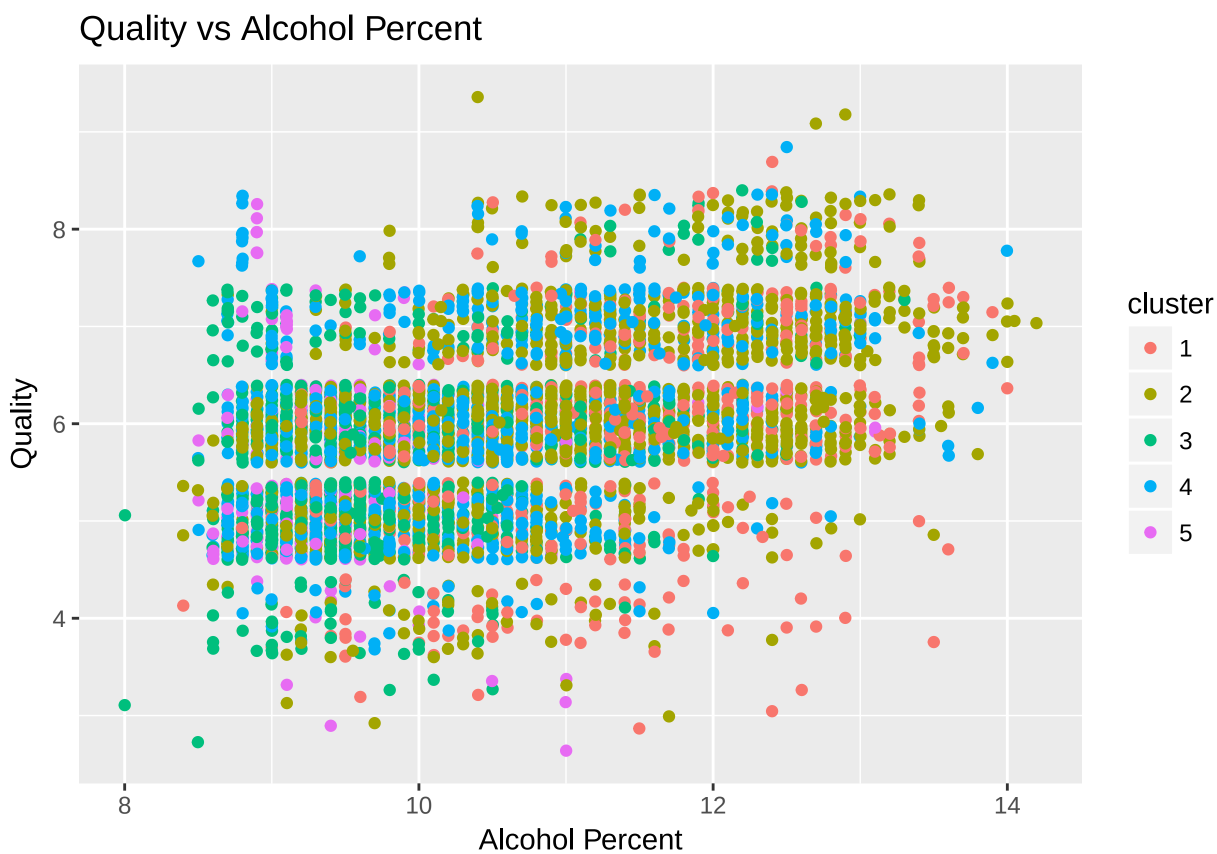
```
cluster = as.factor(km$cluster)

qplot(jitter(wine.dt$`total sulfur dioxide`, 2),
      jitter(wine.dt$alcohol, 2),
      xlab = "Total Sulfur Dioxide",
      ylab = "Alcohol Percent",
      main  = "Alcohol Percent vs Sulfur Dioxide",
      colour = cluster)
```




```
qplot(jitter(wine.dt$`total sulfur dioxide`, 2),
      jitter(wine.dt$quality, 2),
      xlab = "Total Sulfur Dioxide",
      ylab = "Quality",
      main  = "Quality vs Total Sulfur Dioxide",
      colour = cluster)
```




```r
qplot(jitter(wine.dt$alcohol, 2),
      jitter(wine.dt$quality, 2),
      xlab = "Alcohol Percent",
      ylab = "Quality",
      main  = "Quality vs Alcohol Percent",
      colour = cluster)
```



Conclusion

This investigation shows that the kmeans algorithm has a tendency to show bias for high variance variables, which might not be correlated with a desired output. If supervised learning was performed on the dataset, the defining variable for the unsupervised learning algorithm, total sulfur dioxide, would not alone perform well. This indicates that for meaningful results, we should take efforts to adjust the data to allow for more indicative variables on quality to become cluster centers.

Some things that can be tried are scaling the data before hand, even though this is unlikely to have much of an impact due to the biases of the kmeans algorithm focusing on high variance clusters. Clipping might have better performance for future predictions with a supervised learning such as knn.