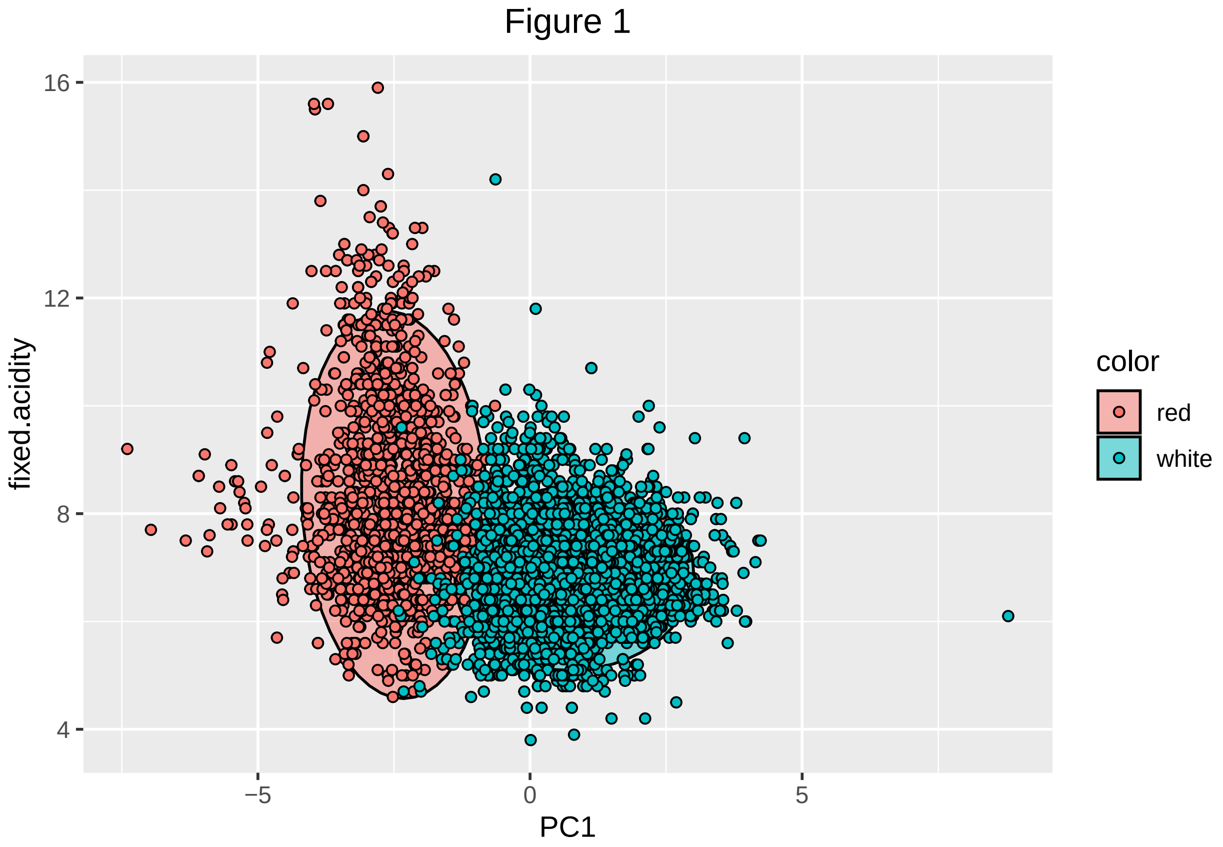
Exercise 4

Question 1

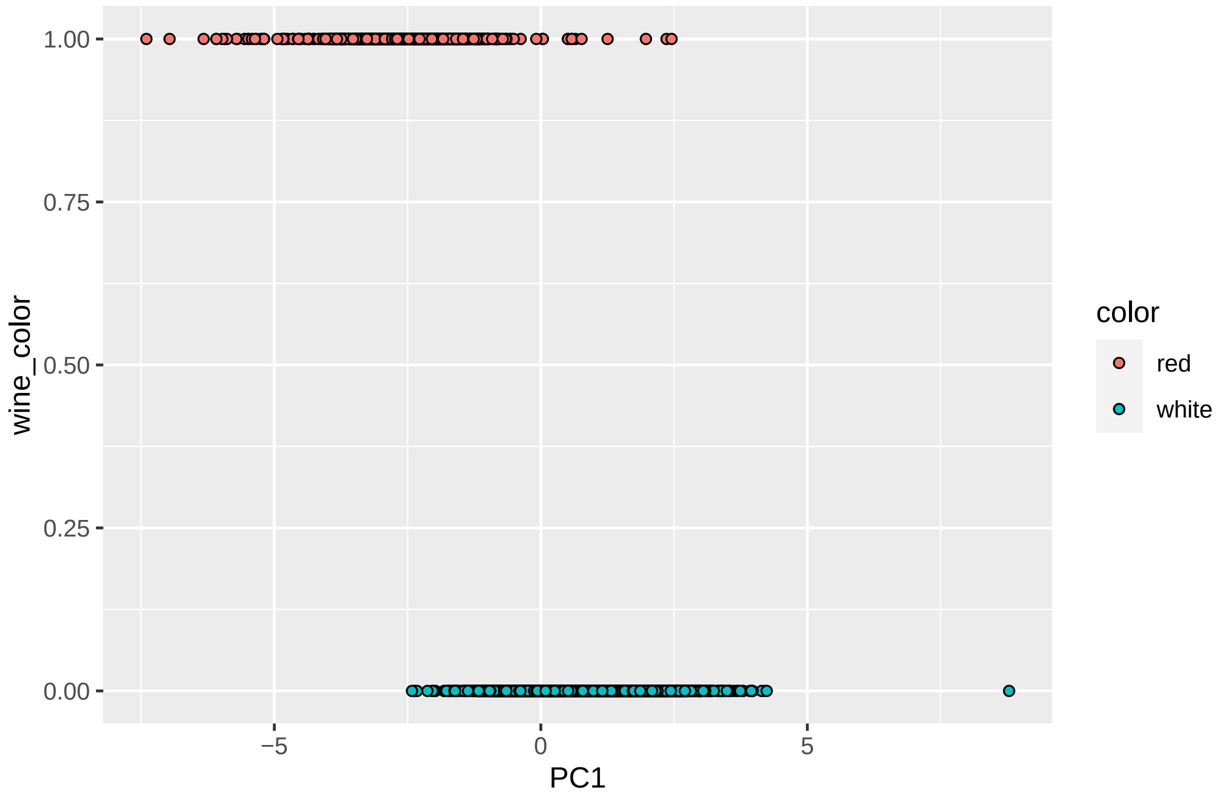
Principal Component Analysis (PCA)

Using principal component analysis (PCA), I was able to use a data set consisting of 11 chemical properties of wine to determine whether the observed wine was categorized as red or white. Figure 1 displays the data collected for the chemical property, fixed acidity, and PC1, and determines which of the points are predicted to be white or red. Additionally, fixed acidity and PC1 have a moderate correlation of approximately -0.415. It should also be noted that PC1 is more or less correlated, depending on the chemical property. Total sulfur dioxide has the highest correlation of 0.848, while density has the lowest correlation of -0.078.



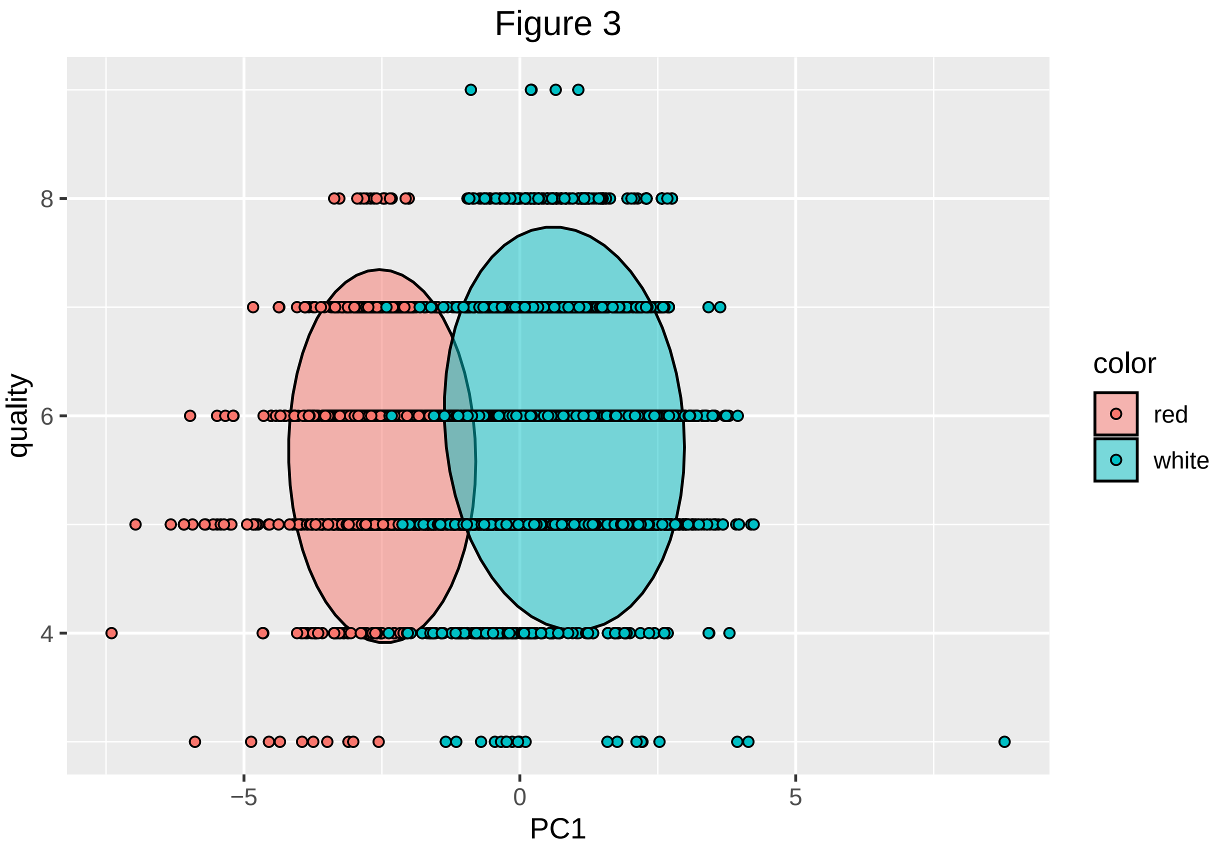
```
##                               [,1]
## fixed.acidity        -0.41566573
## volatile.acidity     -0.66276622
## citric.acid           0.26525521
## residual.sugar        0.60212615
## chlorides            -0.50498499
## free.sulfur.dioxide   0.75007124
## total.sulfur.dioxide  0.84842512
## density              -0.07821904
## pH                   -0.38065695
## sulphates            -0.51198691
## alcohol              -0.18526996
```

Figure 2



```
## [1] -0.8254209
```

Figure 2 displays PC1 and the determined wine_color, red or white. Overall, PCA did a good job at determining the color of wine, with a correlation of -0.825.

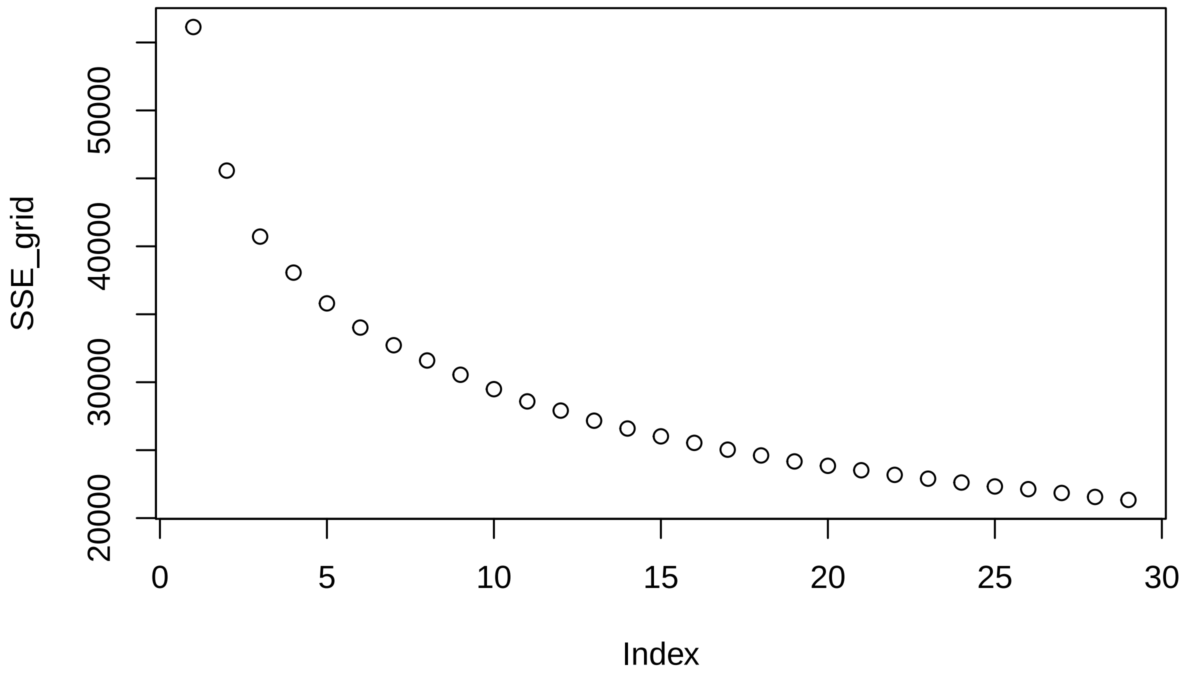


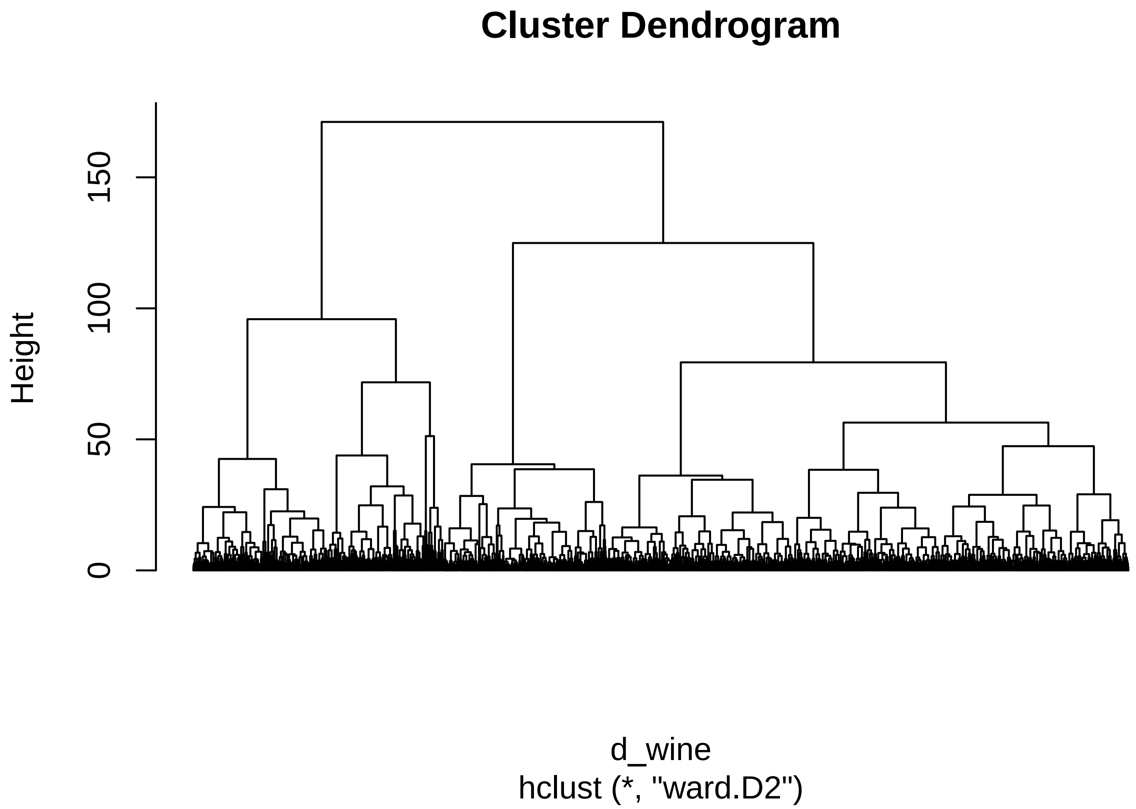
```
## [1] 0.07614681
```

Figure 3 displays the rated quality of wine and PC1 of either red or white wines. PCA did a poor job at determining the rated quality of wine, with a correlation of -0.076.

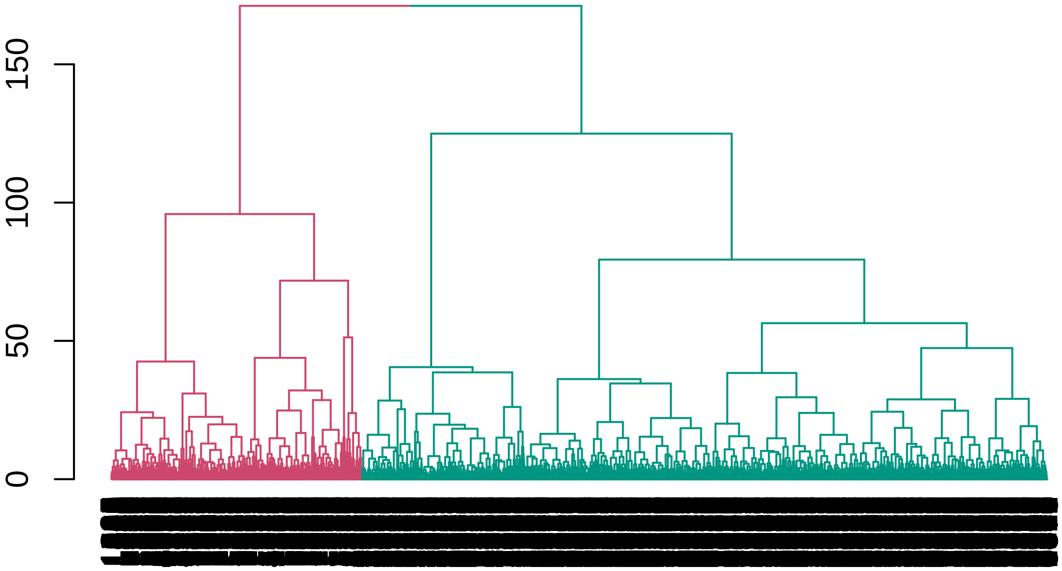
Hierarchical Clustering

As a separate approach, I decided to use hierarchical clustering. Here, I applied the elbow method and determined that 2 was the optimal number of clusters for the data set.





Once centering and scaling the data for the 11 chemical properties, I used euclidean distances to measure the distances between the points, and created a resulting variable named, d_wine. Thereafter, the Ward method proved to be the best approach in creating a well-balanced denodrogram, which resulted in the following cluster dendrogram.



```
## cluster
##    1    2
## 1741 4756
```

I created a second dendrogram to give an appropriate visualization of the optimal tree cut at k = 2. On the left-hand side, the section in red is cluster 1 and the resulting clustering of the red wines. On the right-hand side, the section in green is cluster 2 and is the resulting clustering of the white wines. Cluster 1 and cluster 2 contain 1,741 and 4,756 wines, respectively, which is very close to the proportionate amounts of red and white wines in the data set.

```
## [1] 0.9274149
```

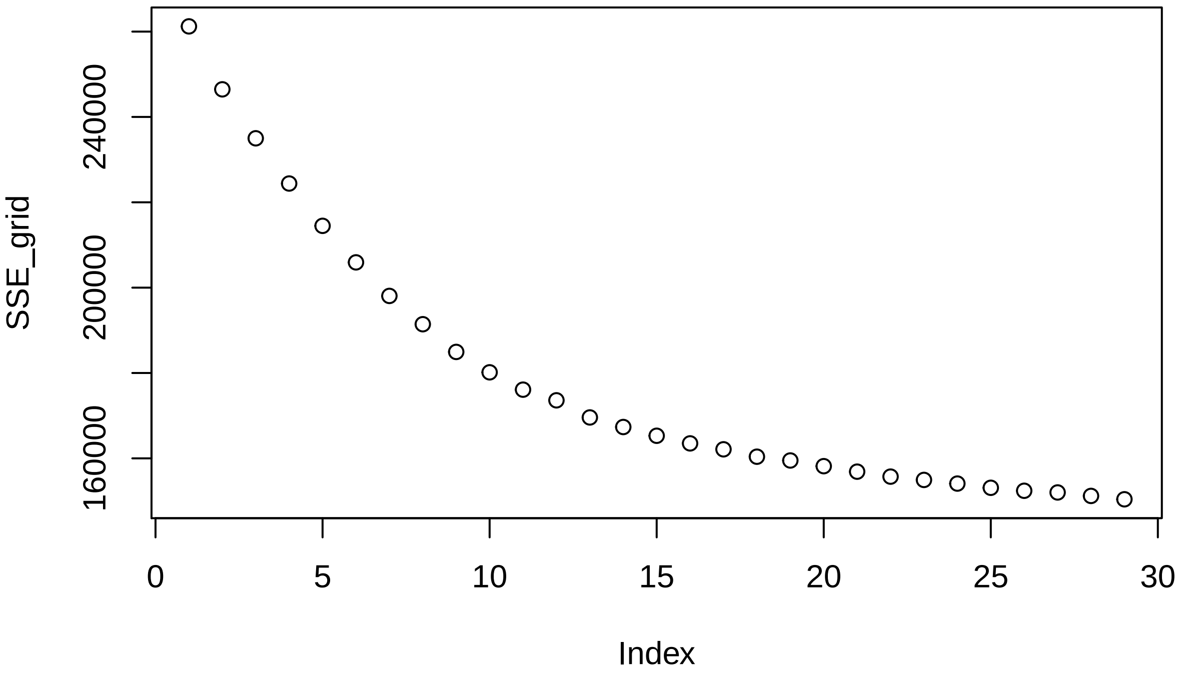
```
## [1] -0.1193233
```

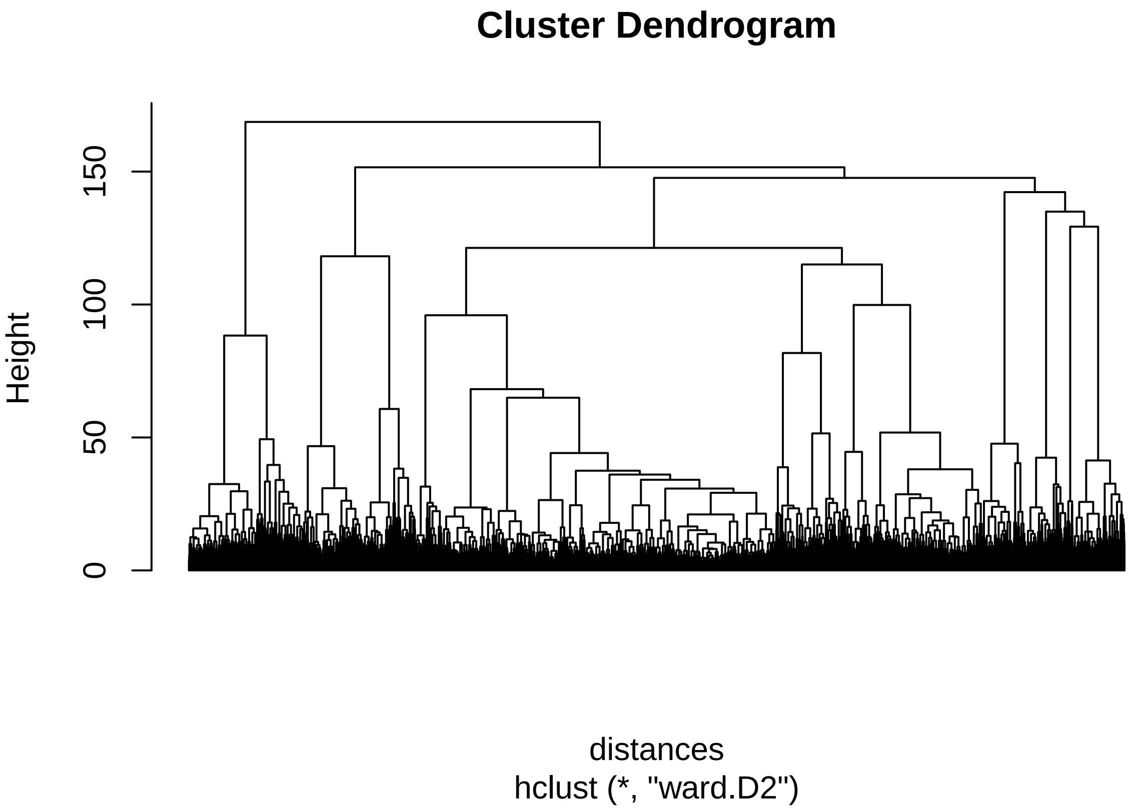
Using the resulting clustering to determine the rate of success in verifying whether the wine is red or white based on the 11 chemical properties. The correlation of determined wine colors between the clusters and the data set is 0.927, which is a better outcome than the resulting outcome found using PCA, which had a correlation of -0.825. With a correlation of -0.119, hierarchical clustering also better predicted the rated quality of wines, although only slightly; PCA had a correlation of -0.076. To summarize, hierarchical clustering was the better approach with this particular data set.

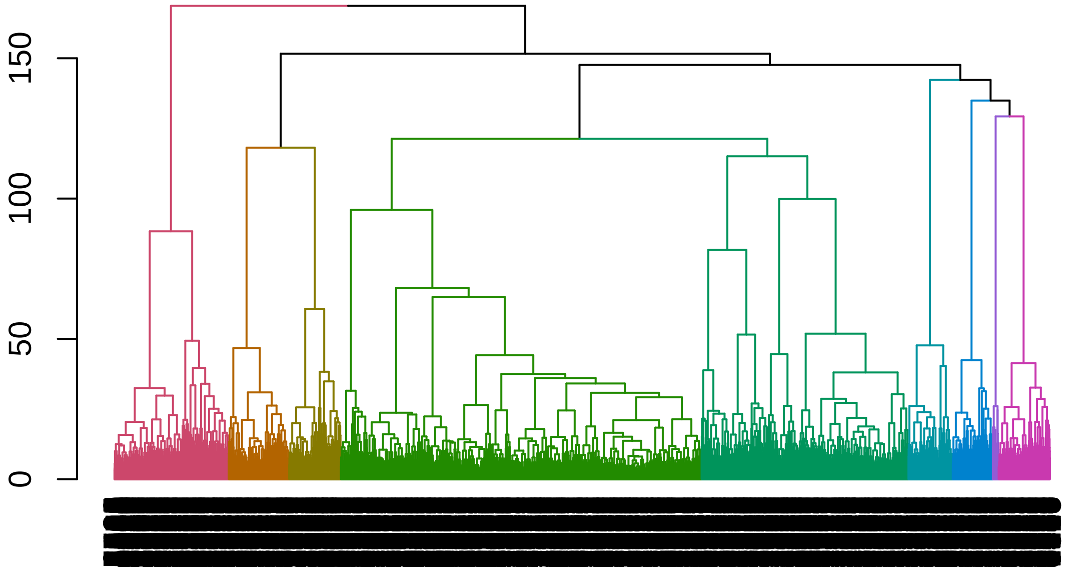
Question 2

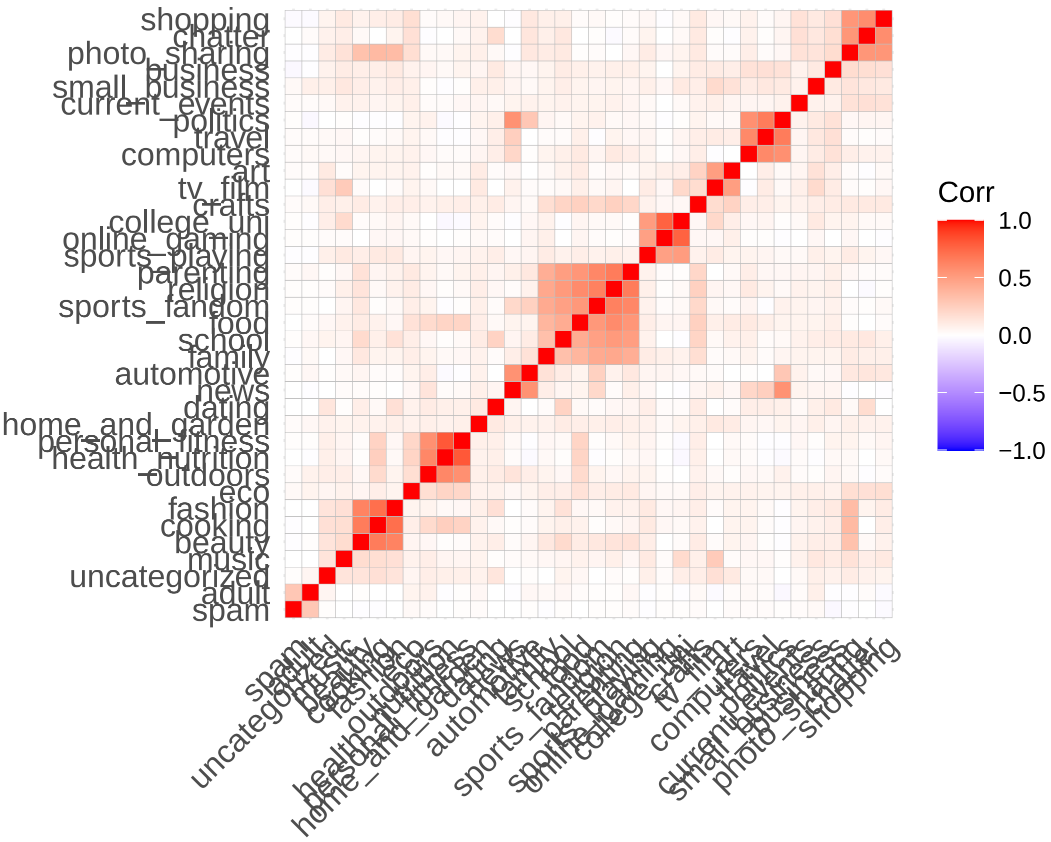
To best approach this objective, I decided to use hierarchical clustering. I removed the first column of the data set, which contained non-numerical values, and named the resulting X_social. I user the euclidean method to find the distances between the points of ther data set. Thereafter, the Ward method proved to be the most successful in creating a cluster dendrogram.

which is represented by the following SSE plot.




Cluster Dendrogram
150
100
50
0
Height
distances
hclust (*, "ward.D2")


150
100
50
0


shopping
chatter
photo_sharing
business
small_business
current_events
politics
travel
computers
art
tv_film
crafts
college_uni
online_gaming
sports_playing
parenting
religion
sports_fandom
food
school
family
automotive
news
dating
home_and_garden
personal_fitness
health_nutrition
outdoors
eco
fashion
cooking
beauty
music
uncategorized
adult
spam
spam
adult
uncategorized
music
beauty
cooking
fashion
eco
outdoors
health_nutrition
personal_fitness
home_and_garden
dating
news
automotive
family
school
food
sports_fandom
religion
parenting
sports_playing
online_gaming
college_uni
crafts
tv_film
art
computers
travel
politics
current_events
small_business
business
photo_sharing
chatter
shopping
Corr
1.0
0.5
0.0
−0.5
−1.0