SDS 323: Exercises 3 Report

Nikhil Ajjarapu   Nevyn Duarte   Rithvik Saravanan

April 20, 2020

Predictive model building

What causes what?

1) Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)

This is because of the fallacy "correlation implies causation". As mentioned in the podcast, this fallacy can cause us to have irrational beliefs. In this specific example, even if there is some correlation between the variables of "Crime" and "Police", that doesn't necessarily mean that the police is the reason crime is changing. There could (and most likely are) other stronger explanations for changes in crime such as poverty, etc. Thus, all other variables must be controlled for in order to run this regression and draw any meaningful conclusions from it.

2) How were the researchers from UPenn able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below, from the researchers' paper.

EFFECT OF POLICE ON CRIME

TABLE 2

TOTAL DAILY CRIME DECREASES ON HIGH-ALERT DAYS

	(1)	(2)
High Alert	−7.316*	−6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

Figure 1: The dependent variable is the daily total number of crimes in D.C. This table present the estimated coefficients and their standard errors in parenthesis. The first column refers to a model where the only variable used in the High Alert dummy whereas the model in column (2) controls form the METRO ridership. * refers to a significant coeficient at the 5% level, ** at the 1% level.

The UPenn researchers were able to isolate this effect by measuring the effect of police on crime when there was a high number of police in an area for a reason unrelated to crime. In the example mentioned in the podcast, they said that in Washington D.C. there are often a lot of cops for events that may attract terroristic threats, which allowed them to isolate the event. When the amount of crime was measured during those times, it had significantly dropped. In addition, they also measured the number of tourists measured by metro ridership (as shown in the chart), to check if the number of police on high-alert days had any influence on the number of tourists (potential victims) out and about. The table shows that the ridership was unchanged by the number of police on high terror days, which shows that there is in fact an inverse relationship between the number of police present and the amount of crime that occurs.

3) Why did they have to control for Metro ridership? What was that trying to capture?

They controlled for Metro ridership to answer the question of whether the drop in crime was actually because of an increased police presence, or because there were just less potential victims (tourists and others who use the metro) around because they were scared by the high-alert police. As mentioned above, it was shown that ridership was not affected, which is further evidence that police themselves do have an effect on crime.

4) Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

TABLE 4

REDUCTION IN CRIME ON HIGH-ALERT DAYS: CONCENTRATION ON THE NATIONAL MALL

	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert × District 1	−2.621**	−2.621*	−2.621*
	(.044)	(1.19)	(1.225)
High Alert × Other Districts	−.571	−.571	−.571
	(.455)	(.366)	(.364)
Log(midday ridership)	2.477*	2.477**	2.477**
	(.364)	(.522)	(.527)
Constant	−11.058**	−11.058	−11.058+
	(4.211)	(5.87)	(5.923)

Figure 2: The dependent variable is the daily total number of crimes in D.C. District 1 refers to a dummy variable associated with crime incidents in the first police district area. This table present the estimated coefficients and their standard errors in parenthesis.* refers to a significant coeficient at the 5% level, ** at the 1% level.

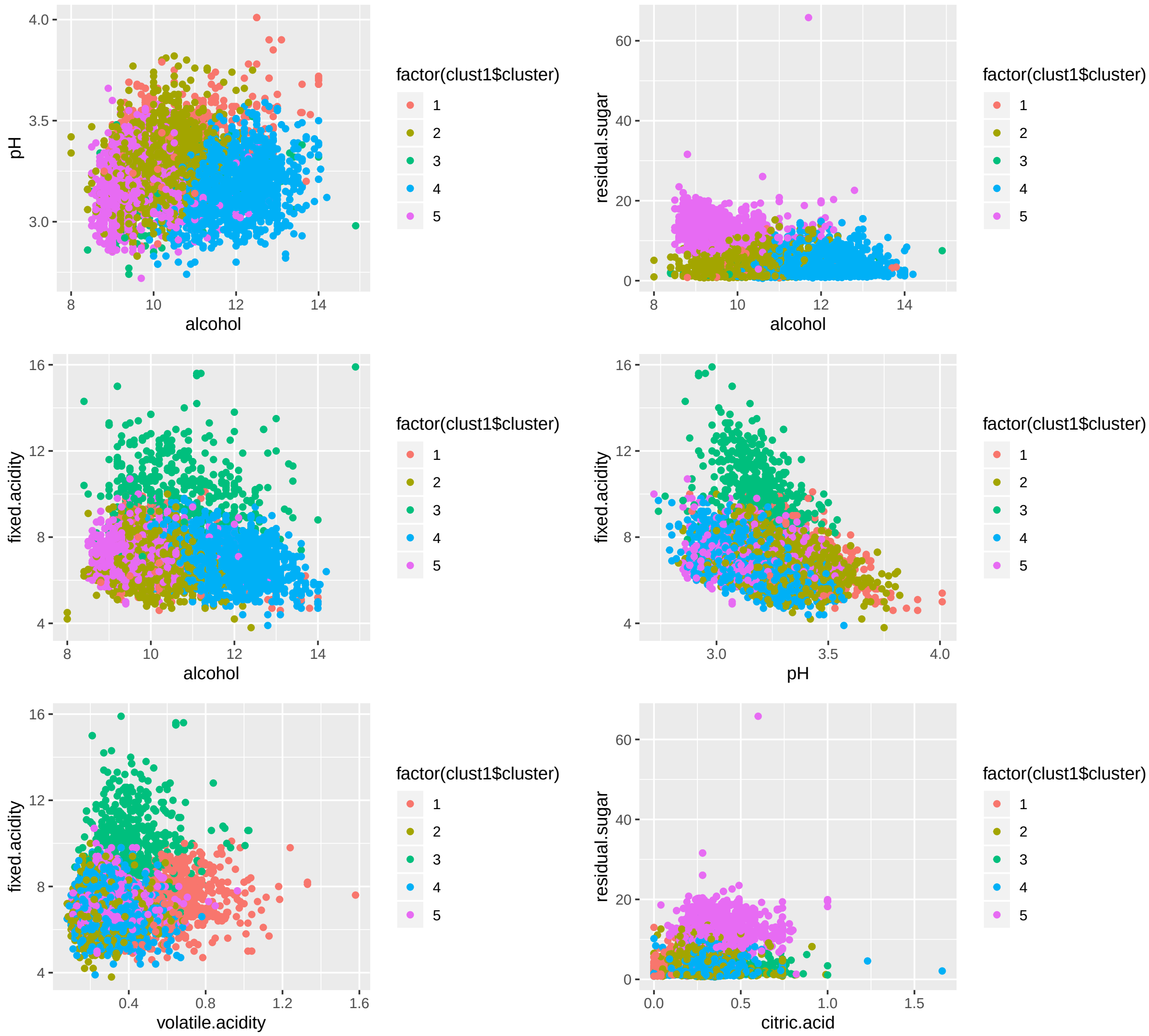
The model being estimated here is a linear model with a few variables as well as a constant to fit the data, where the dependent variable is crime. From the table, it seems to be that the theory that police influence crime holds especially strongly in District 1, but it still does hold some (albeit weak) weight in other districts as well. It seems the tourist theory mentioned earlier also holds true, as metro ridership has a positive coefficient as well. All in all, it seems that the police have a relatively strong effect on crime in District 1, and a much more moderate effect on crime in other districts after controlling for various other factors.

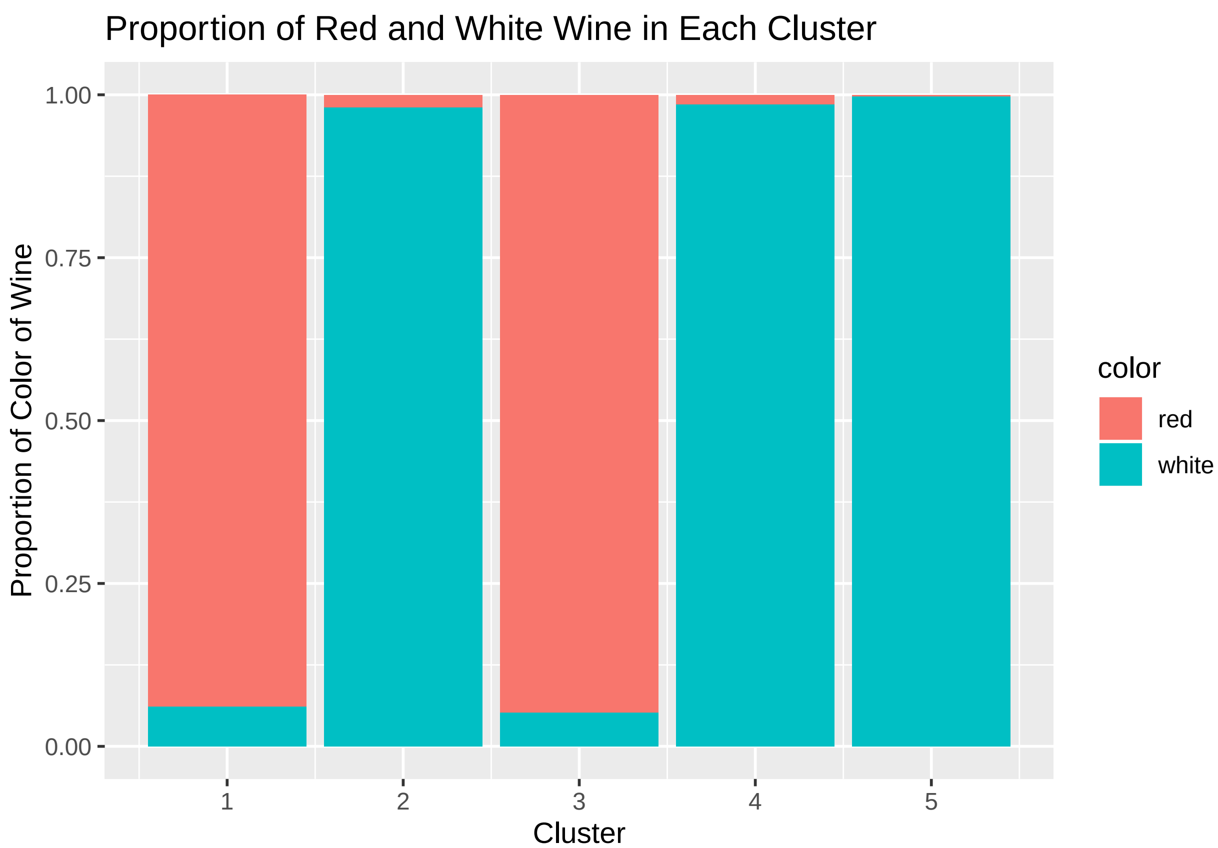
Clustering and PCA

```
##       fixed.acidity    volatile.acidity         citric.acid
##          7.21530706          0.33966600          0.31863322
```

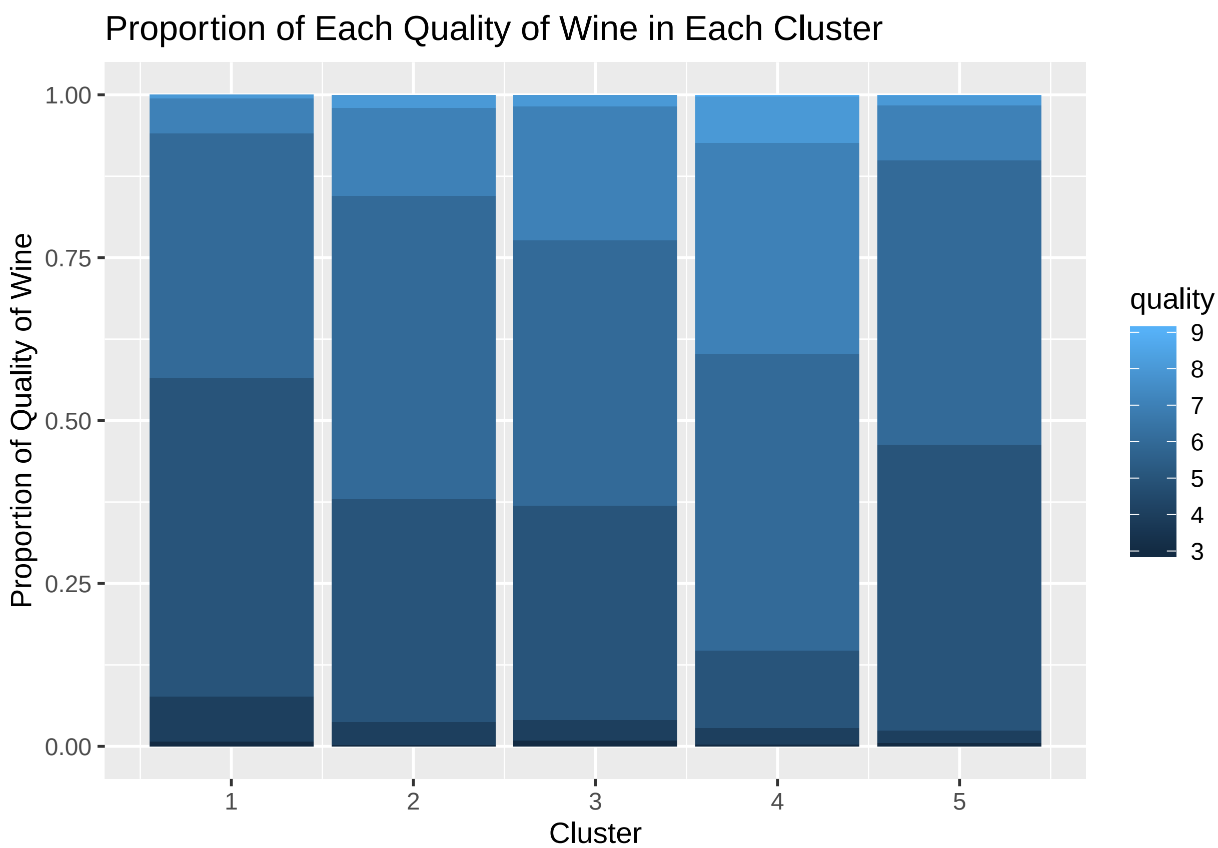
```
##       residual.sugar            chlorides  free.sulfur.dioxide
##           5.44323534           0.05603386          30.52531938
## total.sulfur.dioxide              density                   pH
##         115.74457442           0.99469663           3.21850085
##            sulphates              alcohol
##           0.53126828          10.49180083

##        fixed.acidity     volatile.acidity          citric.acid
##          1.296433758          0.164636474          0.145317865
##       residual.sugar            chlorides  free.sulfur.dioxide
##          4.757803743          0.035033601         17.749399772
## total.sulfur.dioxide              density                   pH
##         56.521854523          0.002998673          0.160787202
##            sulphates              alcohol
##          0.148805874          1.192711749

##   fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## 1    0.05585635        1.6798020 -1.27766503     -0.6244299  0.6595856
## 2   -0.37091549       -0.4911407 -0.02340931     -0.3354886 -0.1687018
## 3    1.97093351        0.4710655  0.96050413     -0.5645007  1.2317014
## 4   -0.31307067       -0.3421027  0.08188139     -0.4182985 -0.5685509
## 5   -0.15889880       -0.3556940  0.30997074      1.4159306 -0.1548387
##   free.sulfur.dioxide total.sulfur.dioxide    density          pH  sulphates
## 1         -0.79552056           -1.1578503  0.4665105  0.97316911  0.4166445
## 2          0.08025269            0.3576911 -0.3000622  0.25692567 -0.1425171
## 3         -0.89115243           -1.2367349  0.9435414 -0.09254086  1.3841836
## 4         -0.13106965           -0.1163211 -1.1907912 -0.32018406 -0.4024693
## 5          0.92254482            0.9863668  0.8800752 -0.48783181 -0.2737318
##      alcohol
## 1 -0.1997366
## 2 -0.2875634
## 3  0.0454428
## 4  1.1746971
## 5 -0.8344902

##
##
## Average Data of Cluster 1 :
##
##        fixed.acidity     volatile.acidity          citric.acid
##           7.28772112           0.61622268           0.13296566
##       residual.sugar            chlorides  free.sulfur.dioxide
##           2.47232050           0.07914152          16.40530697
## total.sulfur.dioxide              density                   pH
##          50.30072841           0.99609555           3.37497399
##            sulphates              alcohol
##           0.59326743          10.25357267
##
##
## Average Data of Cluster 2 :
##
##        fixed.acidity     volatile.acidity          citric.acid
##           6.73443971           0.25880633           0.31523143
```

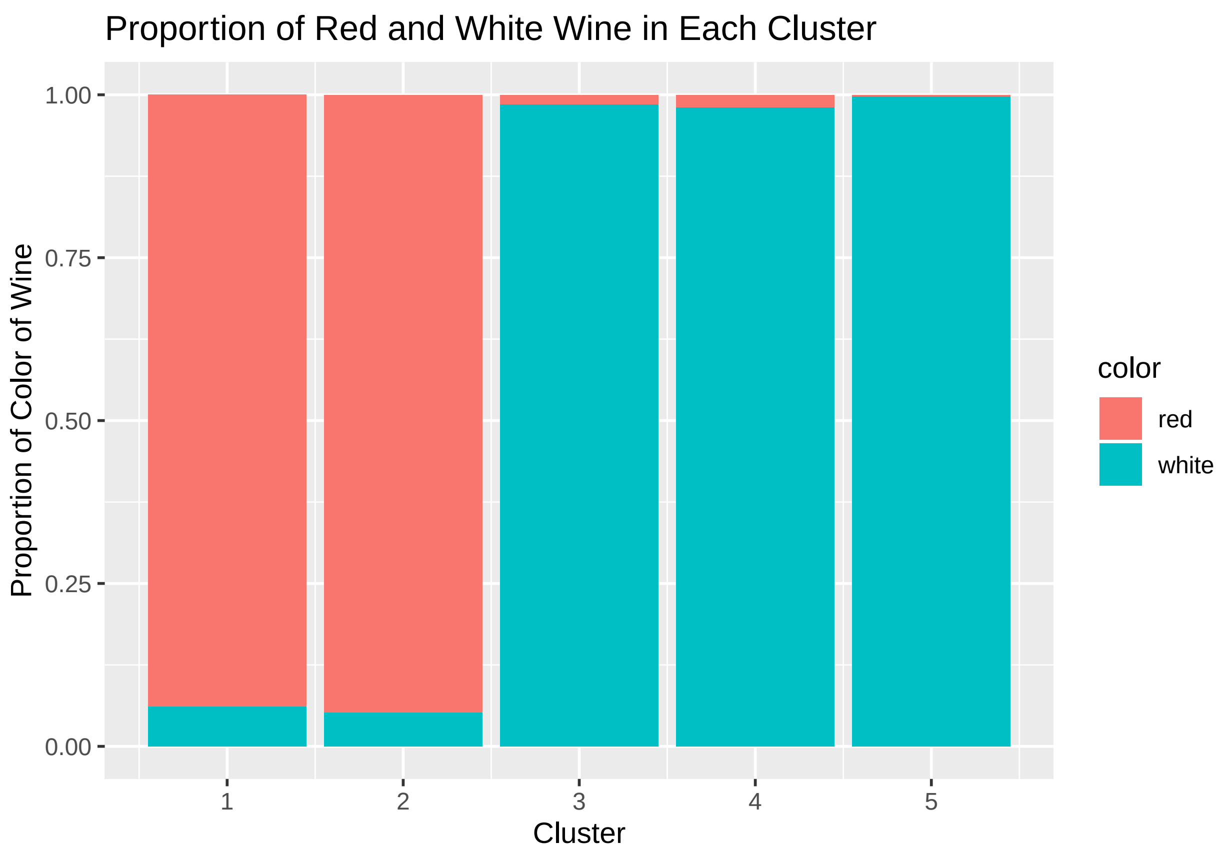
```
##       residual.sugar            chlorides  free.sulfur.dioxide
##           3.84704629           0.05012363          31.94975639
## total.sulfur.dioxide              density                   pH
##         135.96193666           0.99379685           3.25981121
##            sulphates              alcohol
##           0.51006090          10.14882054
##
##
## Average Data of Cluster 3 :
##
##        fixed.acidity     volatile.acidity          citric.acid
##            9.7704918            0.4172206            0.4582116
##       residual.sugar            chlorides  free.sulfur.dioxide
##            2.7574516            0.0991848           14.7078987
## total.sulfur.dioxide              density                   pH
##           45.8420268            0.9975260            3.2036215
##            sulphates              alcohol
##            0.7372429           10.5460010
##
##
## Average Data of Cluster 4 :
##
##        fixed.acidity     volatile.acidity          citric.acid
##           6.80943168           0.28334341           0.33053204
##       residual.sugar            chlorides  free.sulfur.dioxide
##           3.45305320           0.03611548          28.19891173
## total.sulfur.dioxide              density                   pH
##         109.16989117           0.99112584           3.16701935
##            sulphates              alcohol
##           0.47137848          11.89287586
##
##
## Average Data of Cluster 5 :
##
##        fixed.acidity     volatile.acidity          citric.acid
##           7.00930529           0.28110580           0.36367750
##       residual.sugar            chlorides  free.sulfur.dioxide
##          12.17995539           0.05060931          46.89993627
## total.sulfur.dioxide              density                   pH
##         171.49585723           0.99733569           3.14006373
##            sulphates              alcohol
##           0.49053537           9.49649458
```

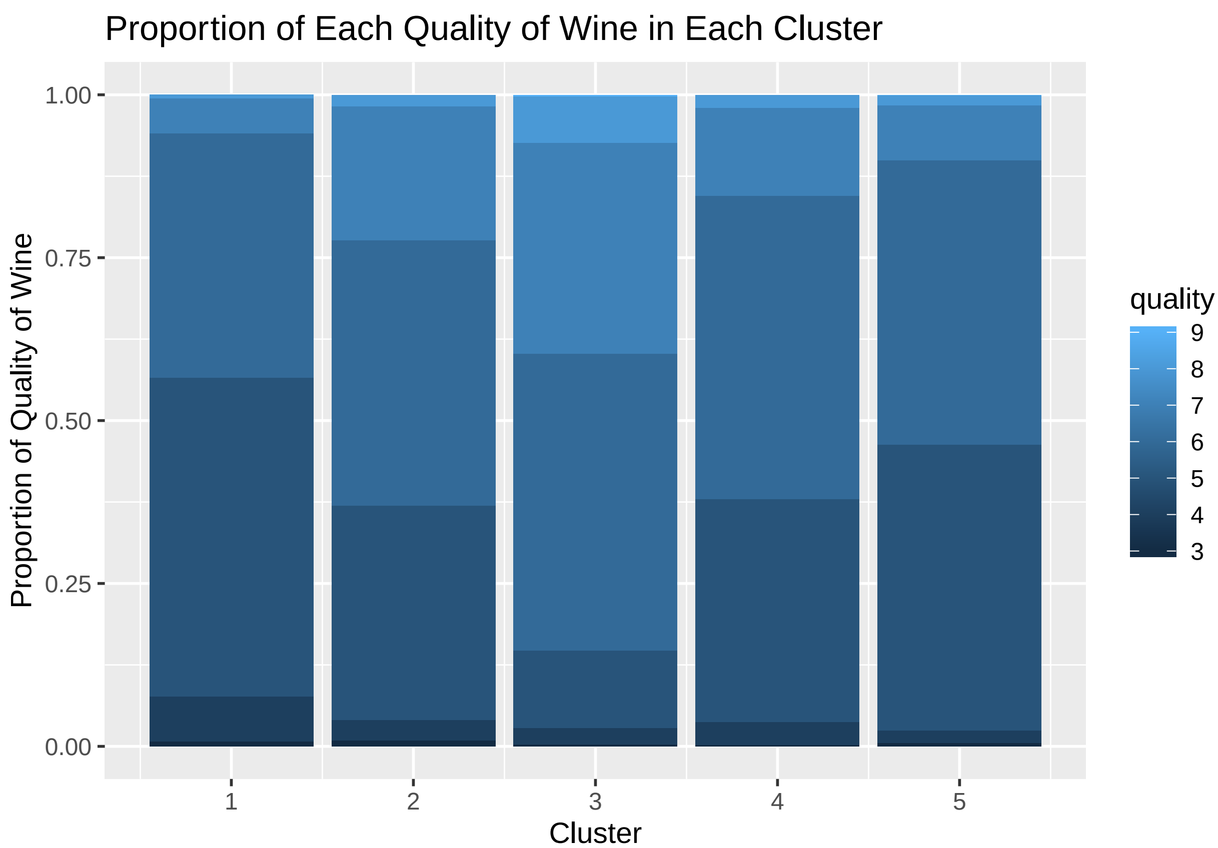

pH
alcohol
residual.sugar
alcohol
fixed.acidity
alcohol
fixed.acidity
pH
fixed.acidity
volatile.acidity
residual.sugar
citric.acid
factor(clust1$cluster)
1
2
3
4
5

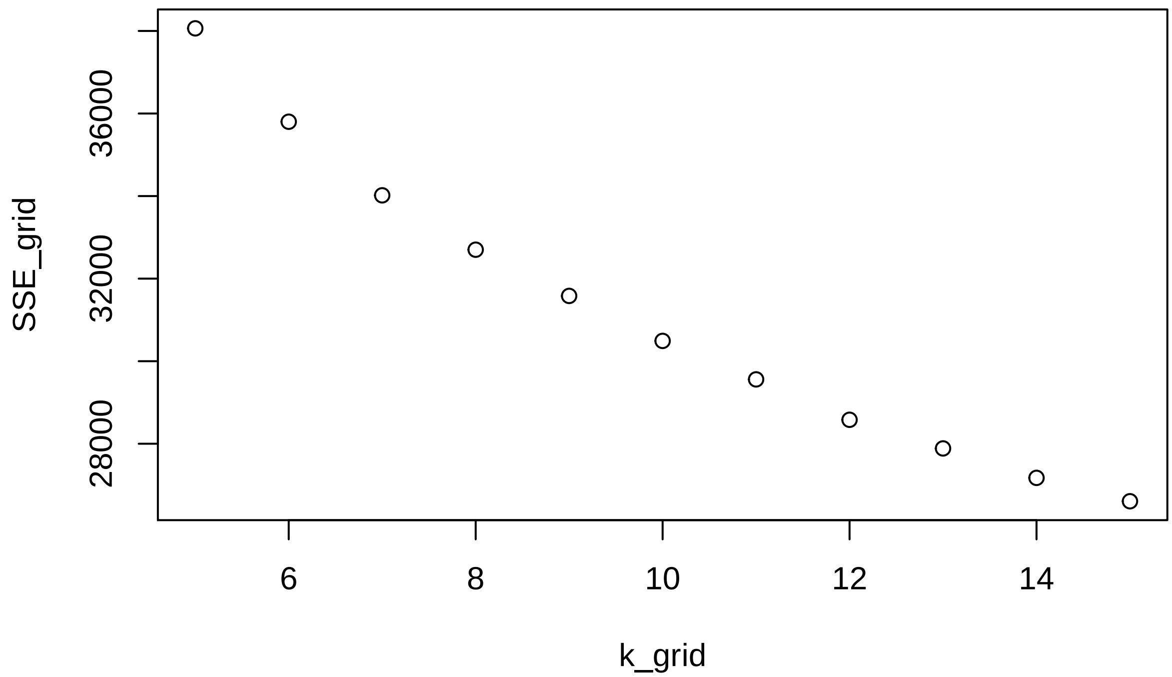


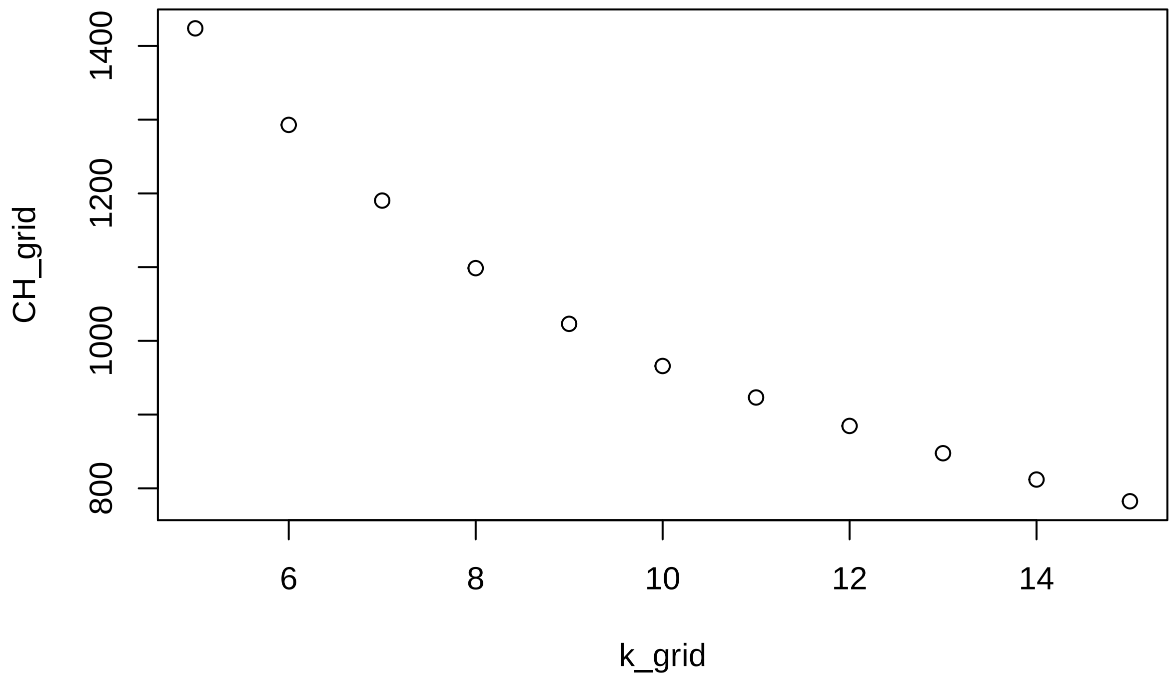
Proportion of Red and White Wine in Each Cluster
1.00
0.75
0.50
0.25
0.00
Proportion of Color of Wine
1
2
3
4
5
Cluster
color
red
white

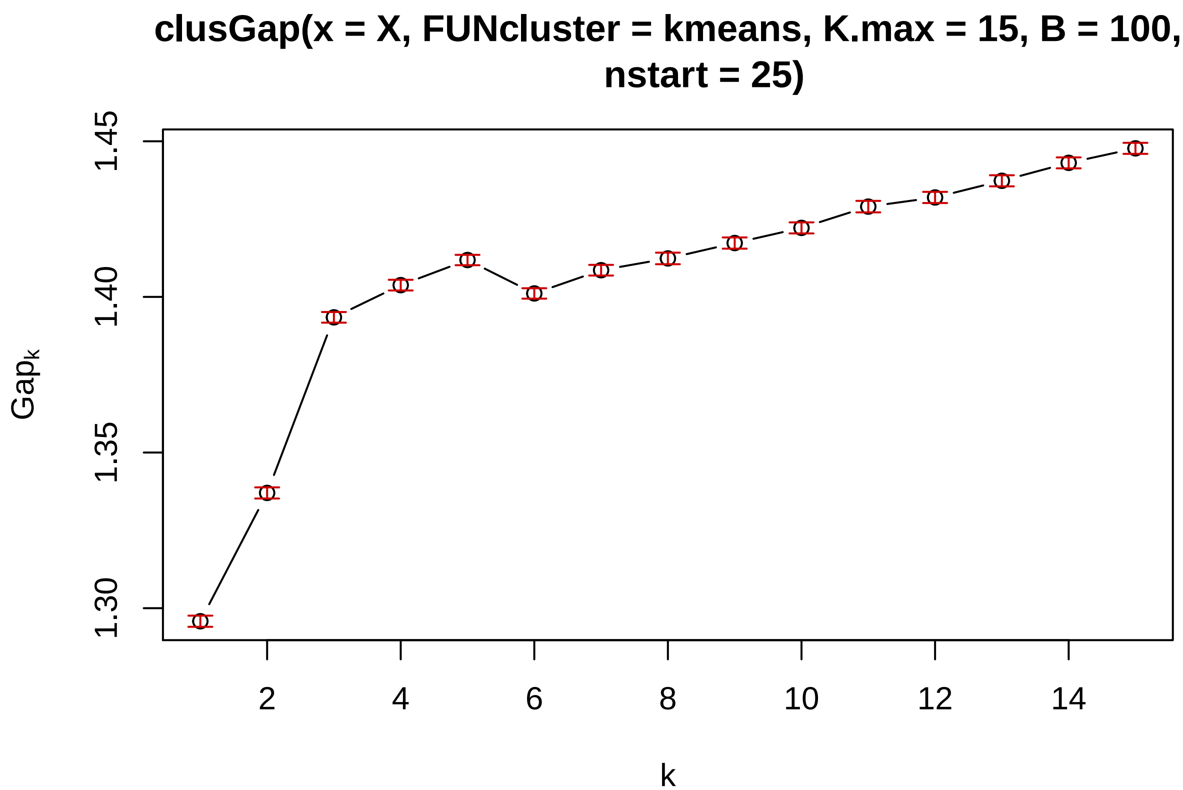

Proportion of Each Quality of Wine in Each Cluster
Proportion of Quality of Wine
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster
quality
9
8
7
6
5
4
3


Proportion of Red and White Wine in Each Cluster
1.00
0.75
0.50
0.25
0.00
Proportion of Color of Wine
1
2
3
4
5
Cluster
color
red
white


Proportion of Each Quality of Wine in Each Cluster
Proportion of Quality of Wine
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster
quality
9
8
7
6
5
4
3


SSE_grid
36000
32000
28000
6
8
10
12
14
k_grid


CH_grid
800
1000
1200
1400
6
8
10
12
14
k_grid



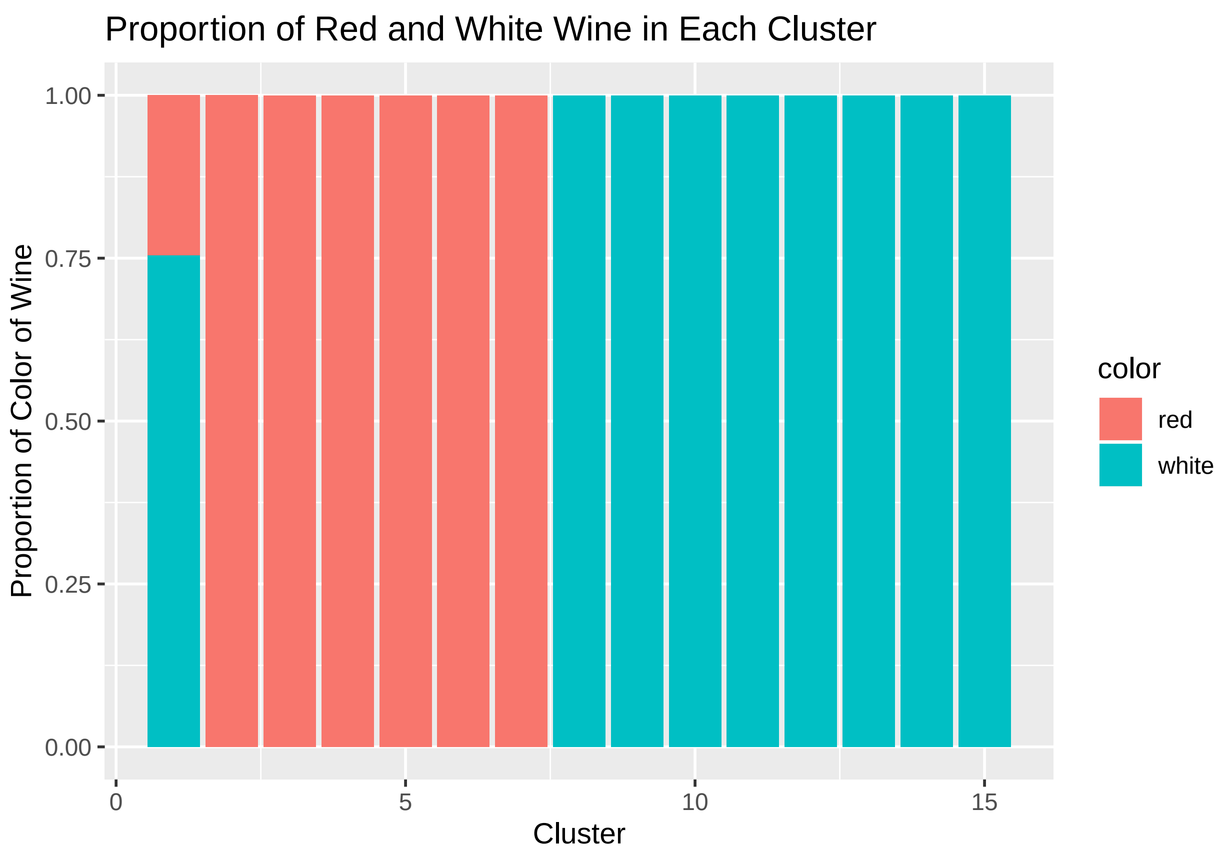

```
## Clustering Gap statistic ["clusGap"] from call:
## clusGap(x = X, FUNcluster = kmeans, K.max = 15, B = 100, nstart = 25)
## B=100 simulated reference sets, k = 1..15; spaceH0="scaledPCA"
##  --> Number of clusters (method 'firstSEmax', SE.factor=1): 5
##           logW    E.logW      gap      SE.sim
##  [1,] 8.873237 10.169038 1.295801 0.001795682
##  [2,] 8.748376 10.085400 1.337025 0.001788170
##  [3,] 8.635368 10.028801 1.393433 0.001707526
##  [4,] 8.584878  9.988660 1.403783 0.001719962
##  [5,] 8.546191  9.958041 1.411851 0.001662067
##  [6,] 8.529867  9.930981 1.401113 0.001670242
##  [7,] 8.503832  9.912409 1.408577 0.001715145
##  [8,] 8.482593  9.894954 1.412361 0.001857292
##  [9,] 8.465828  9.883132 1.417304 0.001806153
## [10,] 8.449872  9.872044 1.422172 0.001775965
## [11,] 8.432989  9.861997 1.429008 0.001845866
## [12,] 8.420307  9.852259 1.431952 0.001793817
## [13,] 8.406257  9.843558 1.437301 0.001783471
## [14,] 8.392131  9.835196 1.443065 0.001764047
## [15,] 8.379247  9.826973 1.447727 0.001774021
```

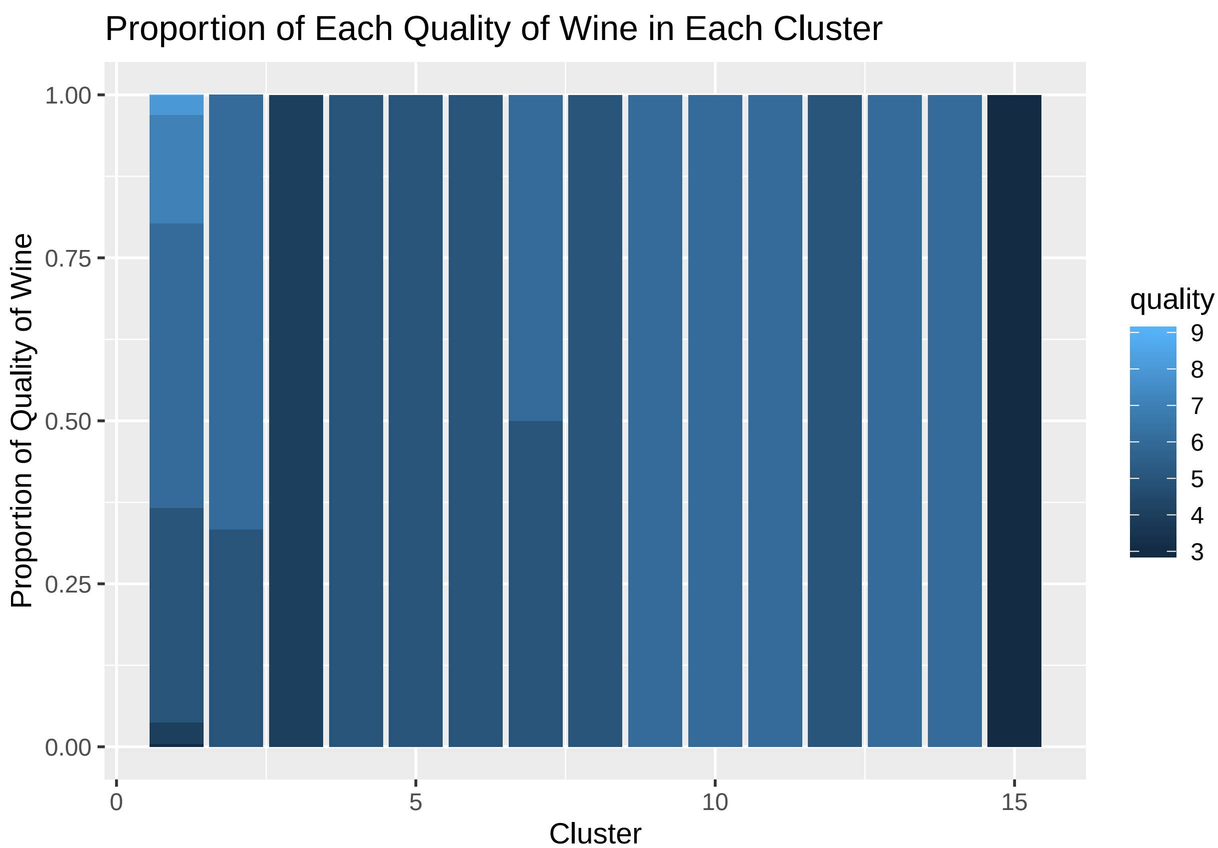
```
## K-means total within-cluster distances: 38063.17
```

```
## K-means++ total within-cluster distances: 38063.17
```

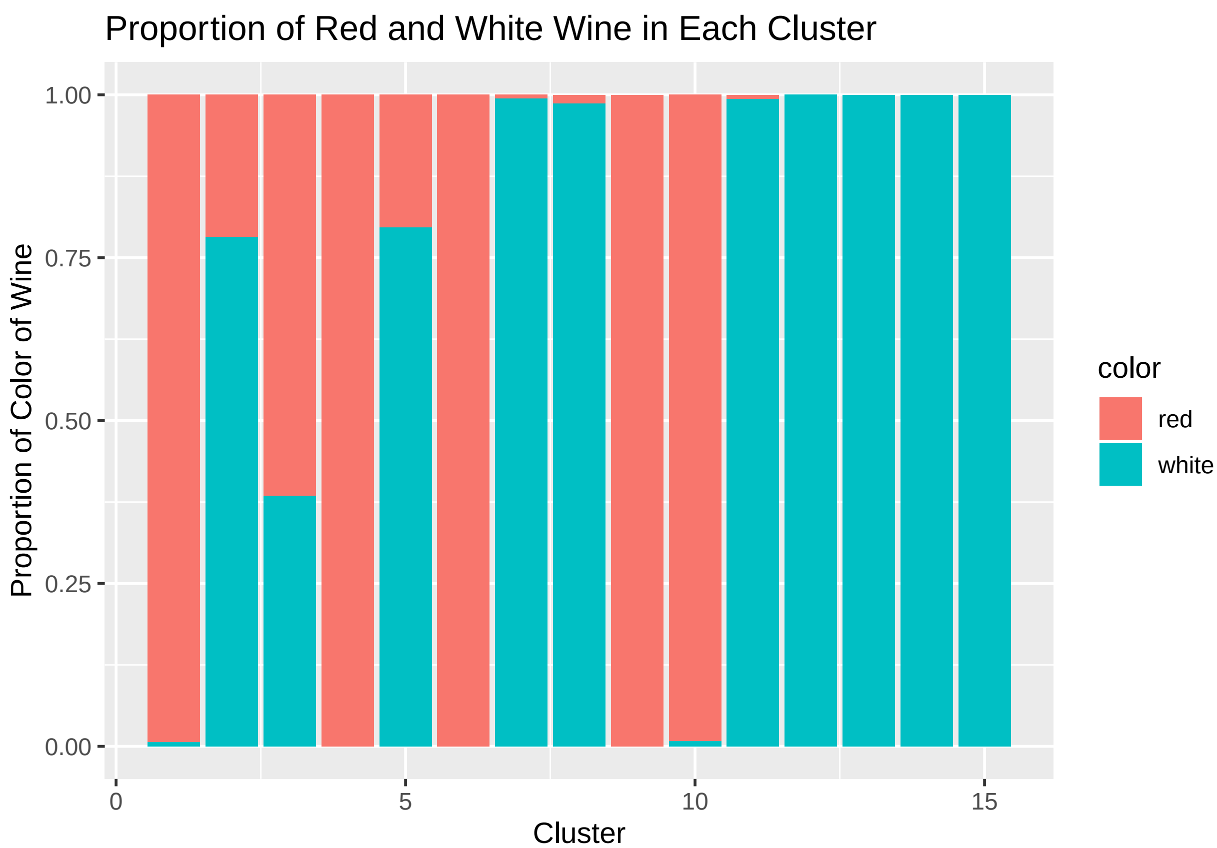
```
## K-means between-cluster distances: 33392.83
```

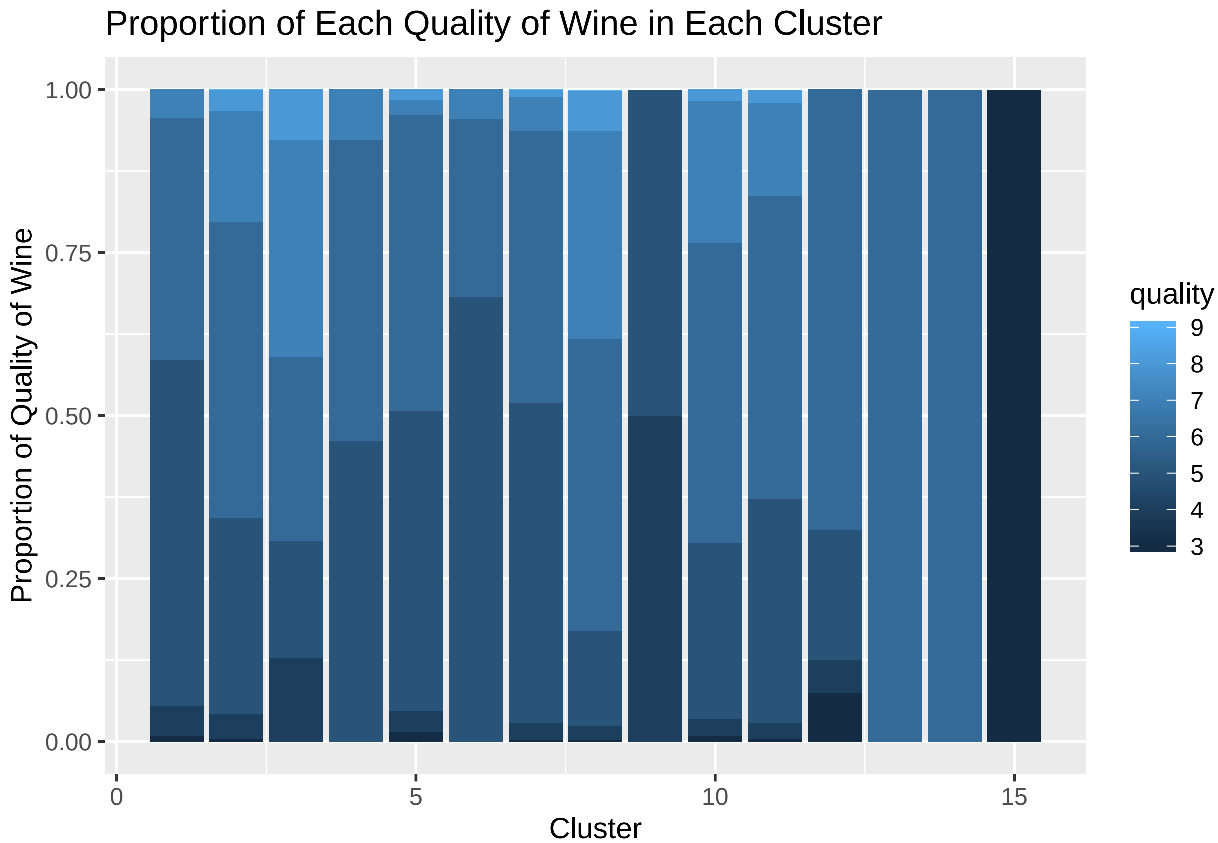
```
## K-means++ between-cluster distances: 33392.83

##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 6477    3    1    1    1    1    4    1    1    1    2    1    1    1    1
```



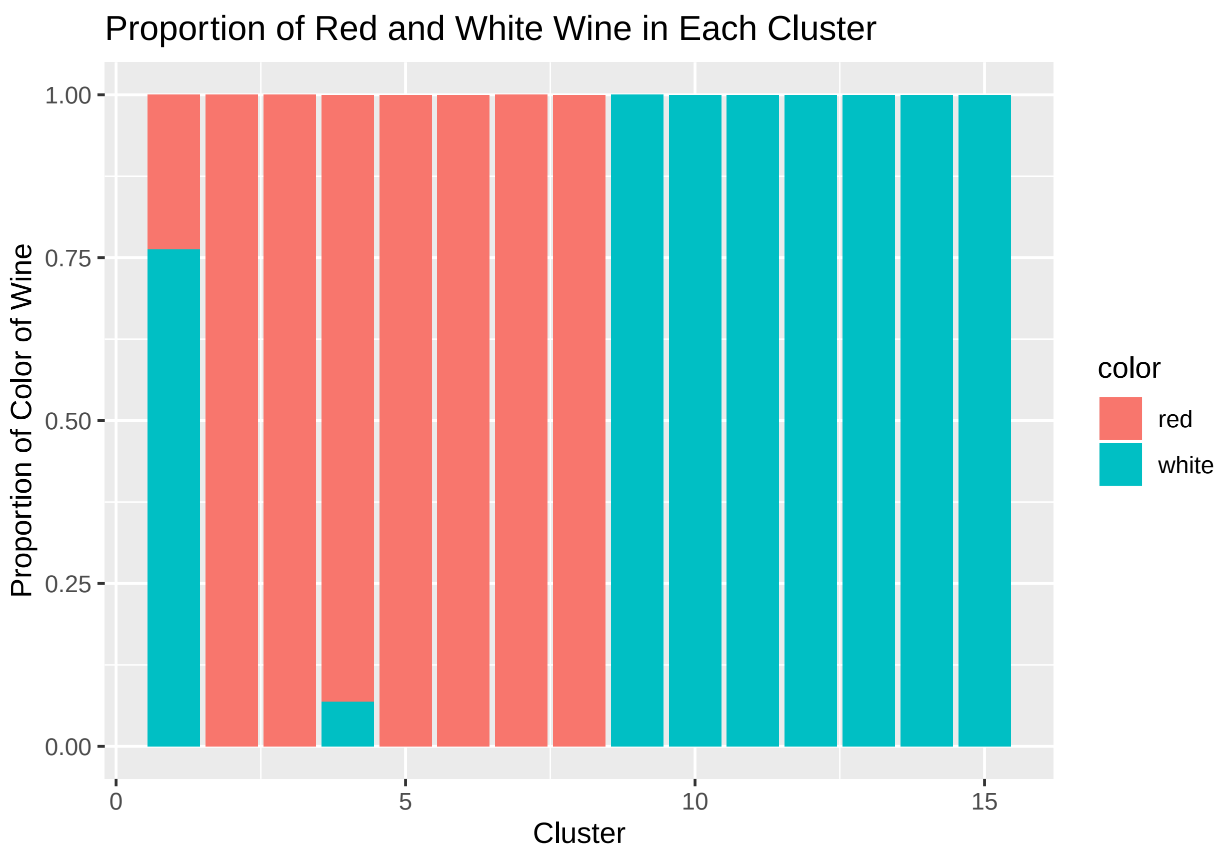


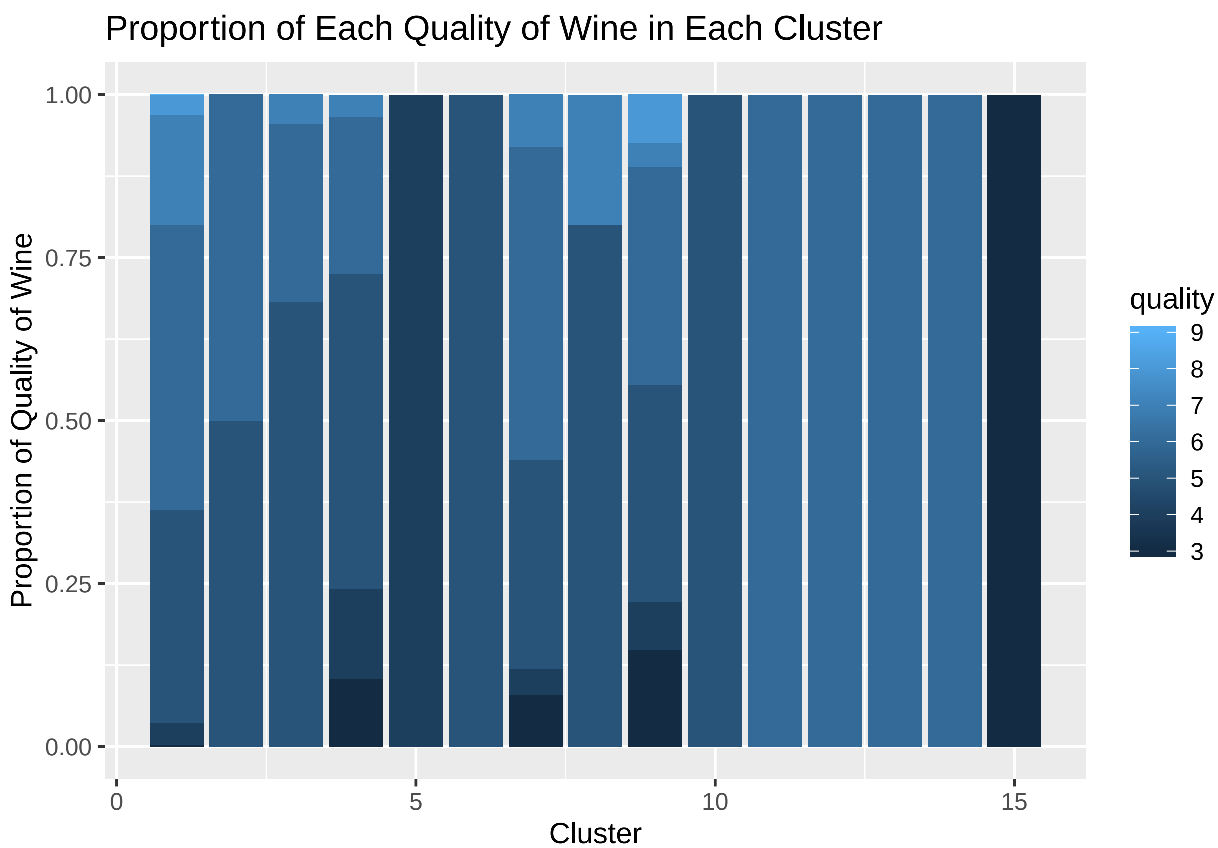

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##  866 2344   39   13  128   22  962 1516    2  115  446   40    2    1    1
```


Proportion of Red and White Wine in Each Cluster
Proportion of Color of Wine
1.00
0.75
0.50
0.25
0.00
0
5
10
15
Cluster
color
red
white

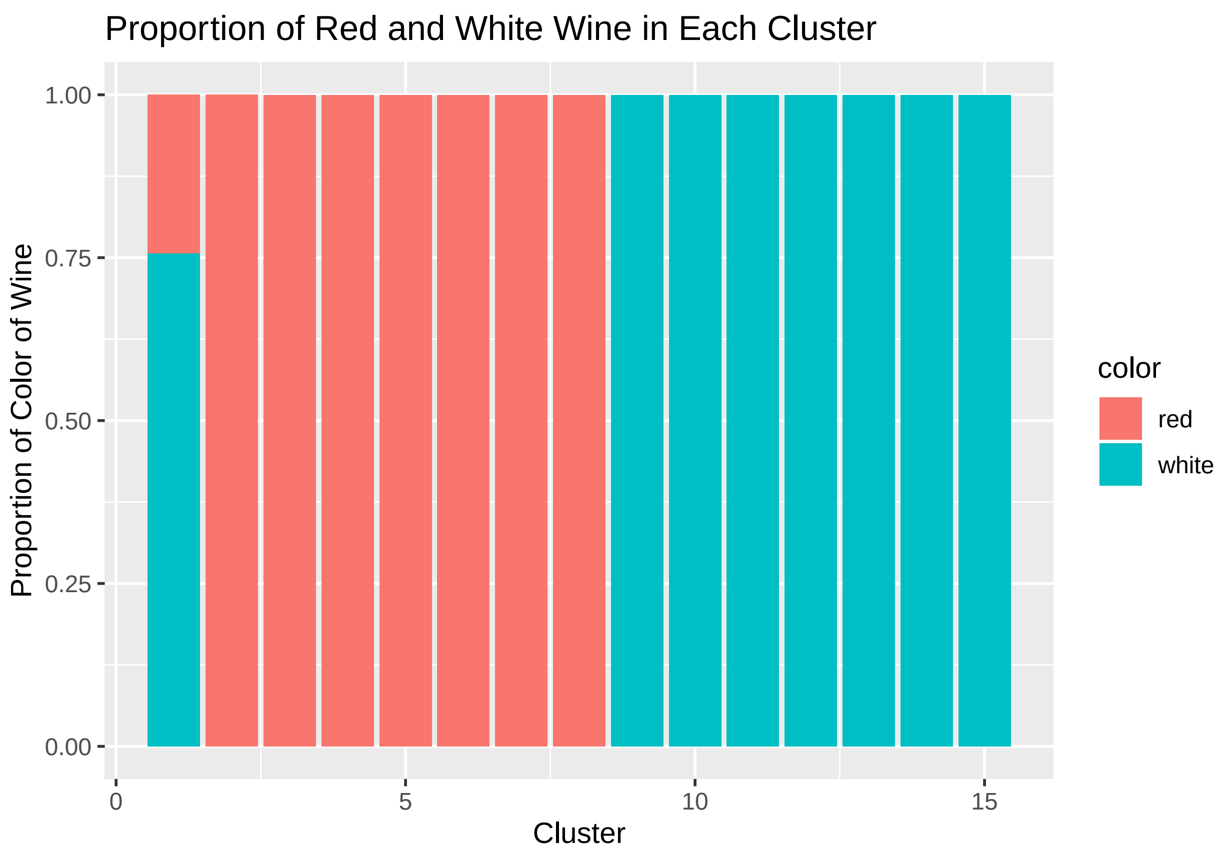


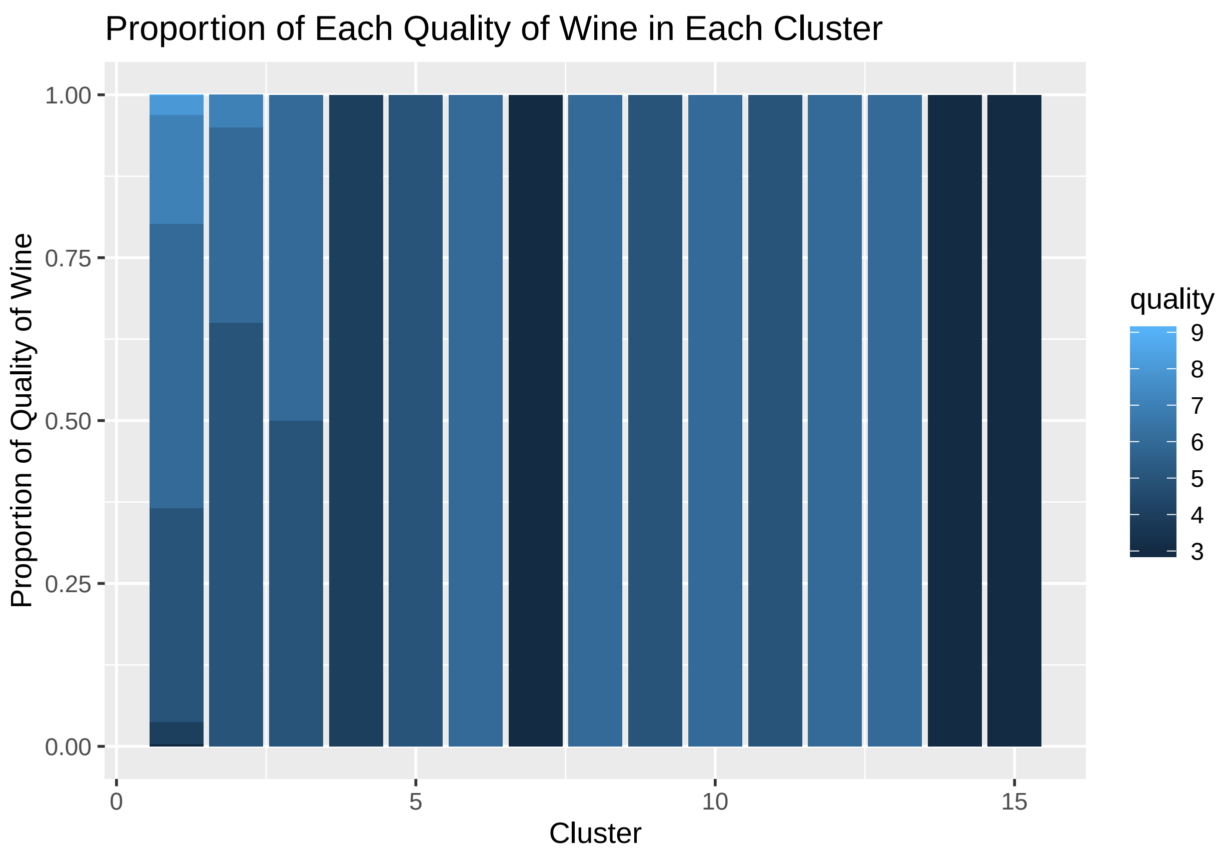

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 6373    6   22   29    1    1   25    5   27    1    2    1    2    1    1
```


Proportion of Red and White Wine in Each Cluster
Proportion of Color of Wine
1.00
0.75
0.50
0.25
0.00
0
5
10
15
Cluster
color
red
white




```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 6461   20    4    1    1    1    1    1    1    1    1    1    1    1    1
```


Proportion of Red and White Wine in Each Cluster
1.00
0.75
0.50
0.25
0.00
Proportion of Color of Wine
0
5
10
15
Cluster
color
red
white



Market segmentation