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

To understand how useful PCA and clustering can be, we can turn to the data we have on wine. The dataset that we used for this exercise contains information on 11 chemical properties of 6500 different bottles of

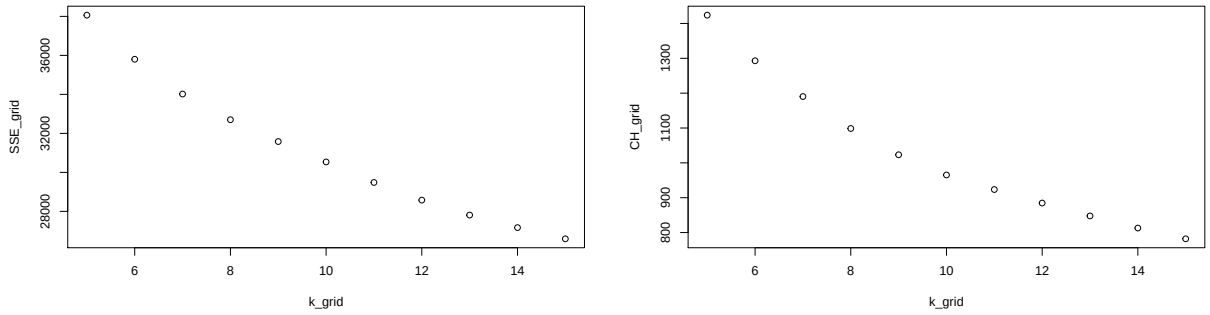
vinho verde wine from northern Portugal. This dataset also records two important features of each bottle of wine: the color (red or white) and quality (on a scale of 1-10).

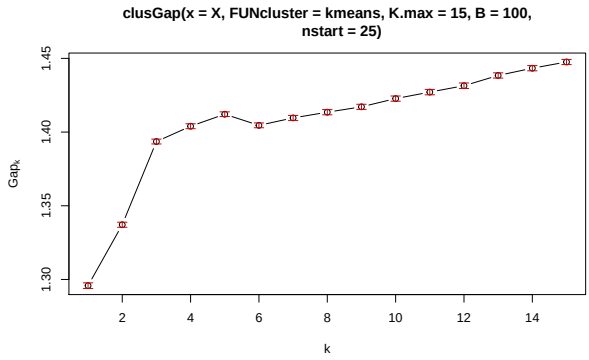
To get a better idea of what this dataset is composed of, we can look at the scaled values of the overall dataset. Below is the data for the average (μ) bottle of wine in the dataset and the standard deviation (σ) of the dataset.

```
Average data point in the dataset: 7.215307 0.339666 0.3186332 5.443235 0.05603386 30.52532 115.7446 0.9
```

```
Standard deviation of the data points in the dataset: 1.296434 0.1646365 0.1453179 4.757804 0.0350336 1
```

To identify the number of clusters to use in our *K-means* clustering algorithms, we utilized the elbow and CH plots shown below. Since there is no significant k value identified by either plot, we used a gap statistic plot to help us identify the value of k.






```
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 10.169043 1.295805 0.001907427
 [2,] 8.748376 10.085504 1.337128 0.001715956
 [3,] 8.635368 10.028942 1.393574 0.001645230
 [4,] 8.584878  9.988837 1.403959 0.001676683
 [5,] 8.546191  9.958311 1.412121 0.001587591
 [6,] 8.526578  9.931138 1.404561 0.001636778
 [7,] 8.502931  9.912527 1.409596 0.001681346
```

```
 [8,] 8.481650  9.895135 1.413485 0.001767710
 [9,] 8.466194  9.883329 1.417135 0.001736533
[10,] 8.449507  9.872221 1.422714 0.001760244
[11,] 8.434986  9.862102 1.427115 0.001834765
[12,] 8.420963  9.852409 1.431445 0.001872988
[13,] 8.405374  9.843790 1.438416 0.001777163
[14,] 8.392050  9.835376 1.443325 0.001767372
[15,] 8.379616  9.827163 1.447547 0.001742667
```

From this gap statistic plot, we can see that the function is non-increasing from k=5 to k=6. Thus, we initialized both of our algorithms with generating 5 clusters.

K-means Clustering

Using this information, we first ran a *K-means* clustering algorithm on this dataset with k=5. For reference, below are the average values of the bottles in each cluster.

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

```
Average Data of Cluster 1 :

       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 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
          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
```

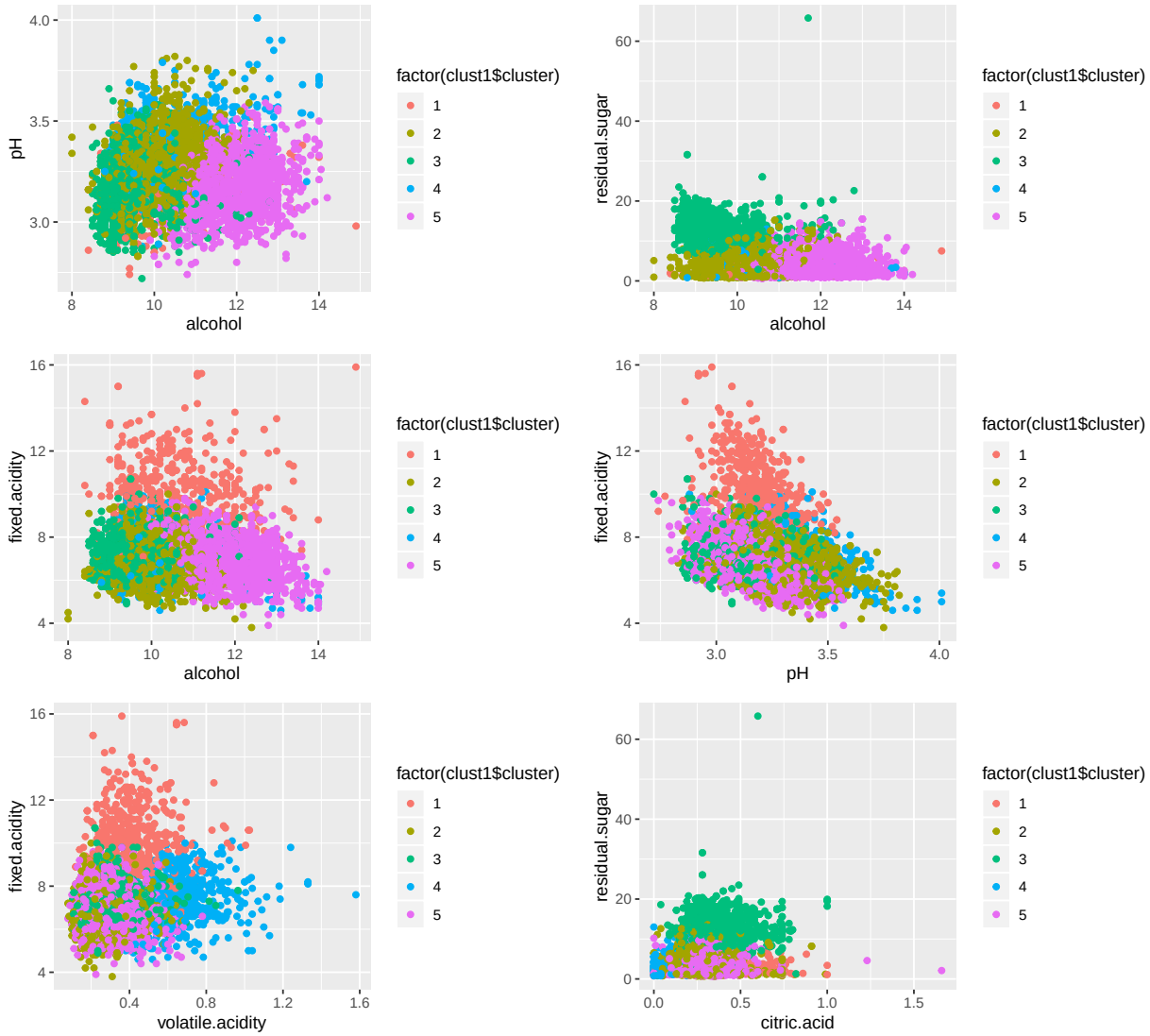
```
Average Data of Cluster 4 :
```

```
       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 5 :
```

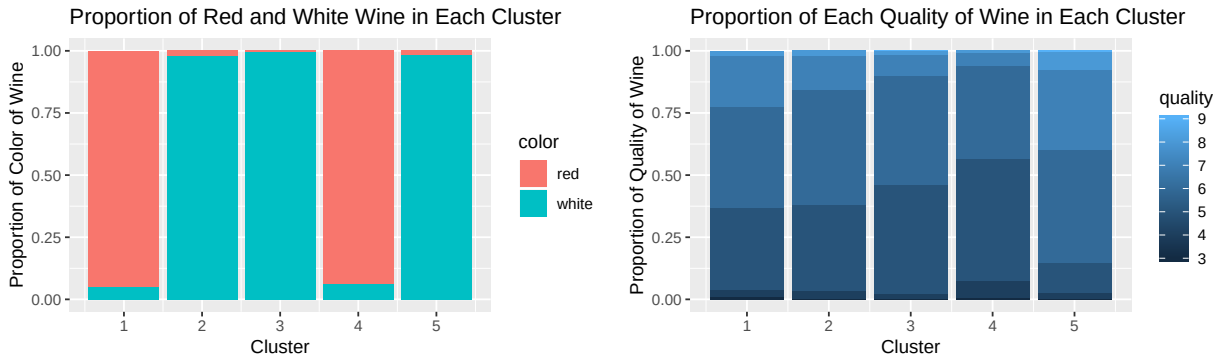
```
       fixed.acidity     volatile.acidity          citric.acid
          6.80943168           0.28334341           0.33053204
      residual.sugar            chlorides  free.sulfur.dioxide
          3.45305320           0.03611548          28.19891173
total.sulfur.dioxide              density                   pH
        109.16989117           0.99112584           3.16701935
           sulphates              alcohol
           0.47137848         11.89287586
```

To get a rough idea of how this basic clustering algorithm performs, we can observe the accuracy of the clustering on some of the features.



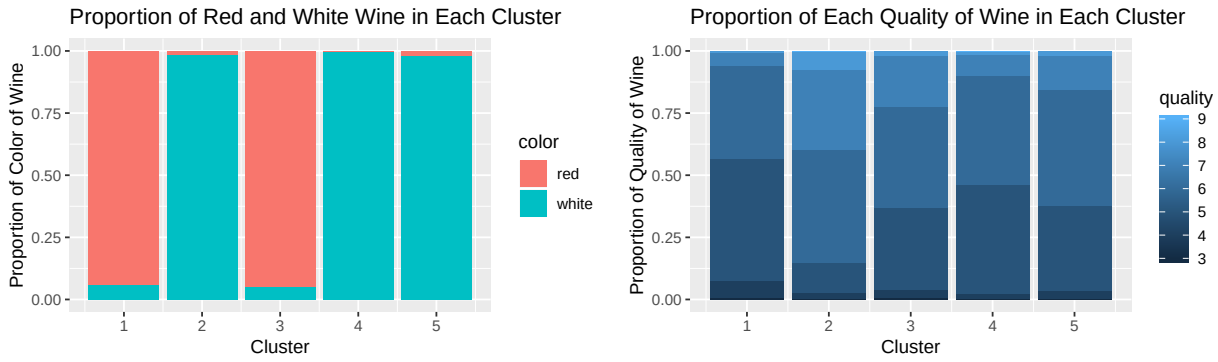
It is important to note that *K-means* generates reasonable clusterings because each cluster has a relatively well-defined region in each plot, but the issue persists that there is a significant amount of overlap between the individual clusters in each of the plots.

To get a better visualization of the clusters, we can look at how well they group the two additional features: color and quality.



These plots show a more descriptive explanation of the clusters. From the plot of red vs. white by cluster above, we can see that clusters 1 and 3 are predominantly bottles of red wine while clusters 2, 4, and 5 are predominantly bottles of white wine. Additionally, the plot of quality by cluster shows that clusters 2, 3, and 4 are roughly mid- to high-quality bottles of wine while clusters 1 and 5 contain comparatively lower-quality wine. Since the distinction between the clusters in terms of quality is somewhat ambiguous, we can look at other methods of clustering.

To improve our clustering, we ran a *K-means++* clustering algorithm with k=5. Below are the same plots that were generated to observe a meaningful description of each cluster.



From the plot of red vs. white by cluster above, we can see that clusters 2 and 4 are predominantly bottles of red wine while clusters 1, 3, and 5 are predominantly bottles of white wine. Additionally, the plot of quality by cluster shows that clusters 2, 4, and 5 have low- to mid-quality wine while clusters 1 and 3 are roughly mid- to high-quality bottles of wine. This clusteirng algorithm shows noticeably better groupings from the basic *K-means* apprach because each cluster is more distinct. To verify this, we can look at the within-cluster and between-cluster average distances for the two clustering algorithms.

```
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
```

However, since our value for k is relatively small, the distance within and between clusters between the *K-means* and *K-means++* clustering algorithms is not very distinguishable. So, these measures are not the best way to convey that *K-means++* is preferred over *K-means* for the wine dataset. Given this, we can see that the clusters from *K-means++* show more understandable groupings than the clusters from *K-means*. Since the main goal of clustering is to cluster the data in a manner that makes it easy to interpret, we can acknowledge that *K-means++* accomplishes this goal better for the wine dataset.

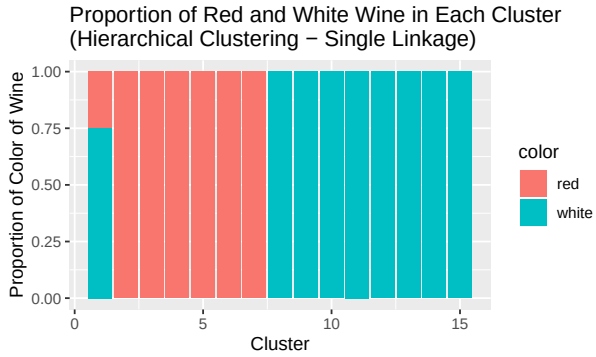
Hierarchical Clustering

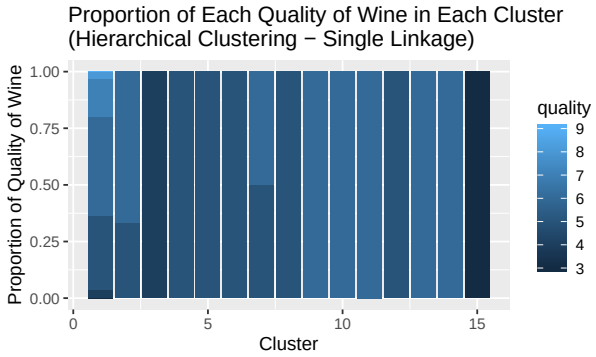
To see if we could further build better clusterings, we used hierarchical clustering of the wine data with 15 clusters.

Below is the number of wine bottles in each cluster for hierarchical clustering with single linkage. Since the clusters are very unbalanced (cluster 1 has significantly more data points than clusters 2-15), we determined

that single linkage was not a viable method. For reference, we have also included the plots of the color of wine by cluster and the quality of wine by cluster. Since single linkage is not a feasible method, these plots are not very useful for identifying clusters.

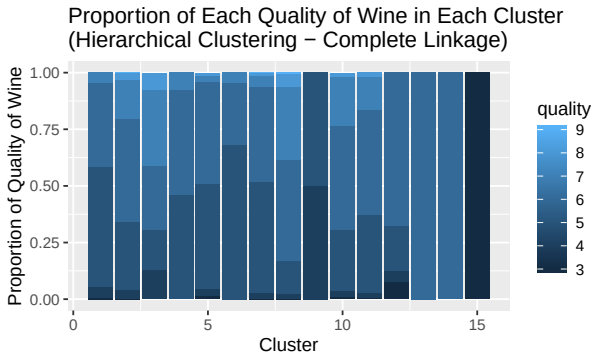
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6477	3	1	1	1	1	4	1	1	1	2	1	1	1	1





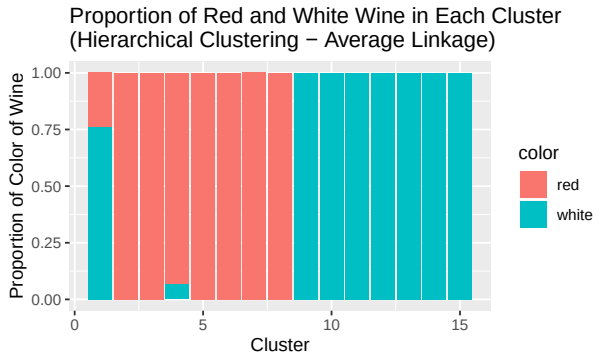
Below is the number of wine bottles in each cluster for hierarchical clustering with complete linkage. Since the clusters are relatively well balanced, we determined that complete linkage was a reasonable method. Looking at the plot of the color of wine by cluster, we can observe that clusters 1, 3, 4, 6, 9, and 10 are predominantly bottles of red wine while clusters 2, 5, 7, 8, 11, 12, 13, 14, and 15 are predominantly bottles of white wine. However, clusters 13, 14, and 15 are not that useful because they have comparatively few data points. From the plot of quality of wine by cluster, we can see that WRITE MORE HERE. This shows that complete linkage is a useful and insightful method for clustering the wine data.

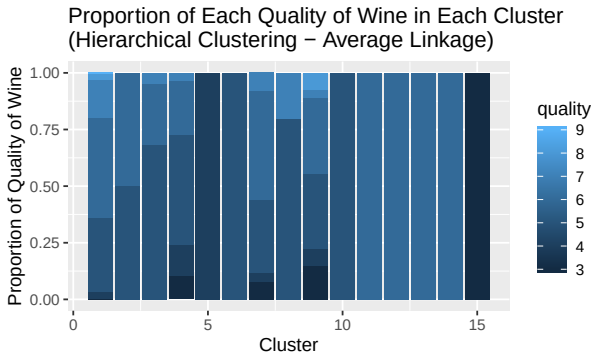
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
866	2344	39	13	128	22	962	1516	2	115	446	40	2	1	1



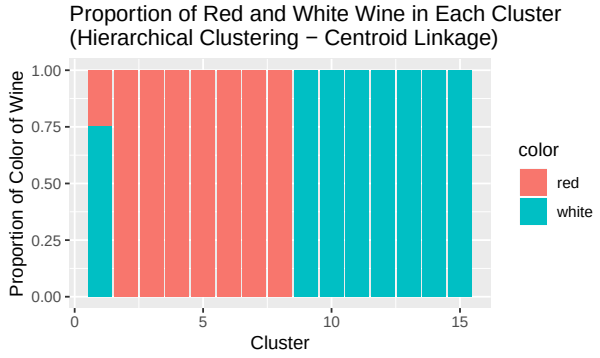
For reference, we have also included the plots for running hierarchical clustering on the wine dataset with both average linkage and centroid linkage. Since the number of bottles per cluster is very unbalanced for both of these types of clustering (similar to single linkage), we determined that these types of linkage are also not viable to cluster the wine data.

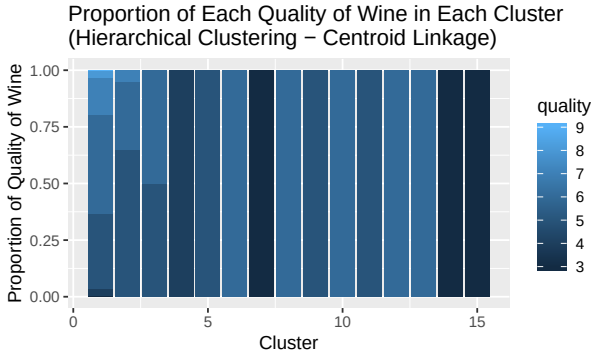
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6373	6	22	29	1	1	25	5	27	1	2	1	2	1	1





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





Market segmentation