R Notebook

Code

Clustering and dimensionality reduction

Question:

The data in wine.csv (https://github.com/jgscott/STA380/blob/master/data/wine.csv) contains information on 11 chemical properties of 6500 different bottles of *vinho verde* wine from northern Portugal. In addition, two other variables about each wine are recorded:

- whether the wine is red or white
- the quality of the wine, as judged on a 1-10 scale by a panel of certified wine snobs.

Run PCA, tSNE, and any clustering algorithm of your choice on the 11 chemical properties (or suitable transformations thereof) and summarize your results. Which dimensionality reduction technique makes the most sense to you for this data? Convince yourself (and me) that your chosen approach is easily capable of distinguishing the reds from the whites, using only the "unsupervised" information contained in the data on chemical properties. Does your unsupervised technique also seem capable of distinguishing the higher from the lower quality wines? Present appropriate numerical and/or visual evidence to support your conclusions.

To clarify: I'm not asking you to run a supervised learning algorithms. Rather, I'm asking you to see whether the differences in the labels (red/white and quality score) emerge naturally from applying an unsupervised technique to the chemical properties. This should be straightforward to assess using plots.

Order of Question Execution:

I have put the order of question execution below. I decided to start with the very basic models, before implementing the PCA and tSNE models. This was simply because I followed the systematic way we learned the clustering models in class.

1. Basic Clustering Models:
    - k-means clustering
    - k-means++ clustering
    - Hierarchical clustering with a cluster Dendogram
2. PCA Model
3. tSNE Model

K-Means Clustering Model:

Step 1: import packages and read wine data. Note, the wine data is in my downloads.

Hide

```
library(ggplot2)
library(ClusterR)  #for kmeans++
library(foreach)
library(mosaic)

#Download the file from my desktop and make sure I include the header:
library(readr)
wine <- read.csv("~/Desktop/wine.csv", header = TRUE)
```

Run a simple summary first: this allows us to see what type of features we are dealing with.

Hide

```
summary(wine)
```

```
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       f
ree.sulfur.dioxide
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   M
in.   :  1.00
 1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800   1
st Qu.: 17.00
 Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700   M
edian : 29.00
 Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603   M
ean   : 30.53
 3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500   3
rd Qu.: 41.00
 Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100   M
ax.   :289.00
 total.sulfur.dioxide    density             pH          sulphates         alcohol
quality
 Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
Min.   :3.000
 1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
1st Qu.:5.000
 Median :118.0        Median :0.9949   Median :3.210   Median :0.5100   Median :10.30
Median :6.000
 Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49
Mean   :5.818
 3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
3rd Qu.:6.000
 Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90
Max.   :9.000
    color
 Length:6497
 Class :character
 Mode  :character
```

As the summary shows there are many variables, bulleted are key takeaways:

- Color is a string (red/white).
- The length of the data is 6497: we have 6497 wines to analyze.
- Quality is measured in integers, unlike the 11 chemical properties, measured as floats.
- While quality is rated on a 0-10 scale there are no wines with a quality rating lower than 3 and higher than 9.
- The ranges of many chemicals are very large, e.g. total sulfur dioxide ranges from 6.0 to 440.0. *This makes me wonder if the huge range comes from very different wine qualities, or differing chemical make ups for red and white wine.*

Step 2: Center and scale data.

X will **not** include quality or color.

Hide

```
#0 for white, 1 for red
wine$color_numeric <- as.numeric(factor(wine$color)) - 1

#Now I am going to set X as all of my variables
X = wine[, c(1:11)]
X = scale(X, center=TRUE, scale=TRUE)
X = scale(X, center=TRUE, scale=TRUE)
```

Checking the column headings:

Hide

```
colnames(X)
```

```
 [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"          "residual.suga
r"
 [5] "chlorides"            "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
 [9] "pH"                   "sulphates"            "alcohol"
```

Hide

```
#Extract the centers and scales from the rescaled data (which are named attributes)
mu = attr(X,"scaled:center")
sigma = attr(X,"scaled:scale")
```

Step 3: Run k-means cluster:

To start, I am going to run a k-means with 10 clusters and 25 starts.

Hide

```
clust1 = kmeans(X, 10, nstart=25)
```

Step 4: Looking at the clusters:

Looking at the z-scores of these centroids is not super helpful:

Hide

```
#What are the clusters?
clust1$center
```

```
   fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides free.sulfur.diox
ide
1    -0.67882313      -0.57877938 -0.15159923     -0.4314757 -0.3735109          0.14153
925
2    -0.47911881      -0.25674908  0.02522201     -0.4522692 -0.6031688         -0.06969
754
3    -0.25130305       1.84426625 -1.67765994     -0.6502460  0.5868665         -0.85167
330
4     3.18883876       0.42767333  1.45662268     -0.5297886  0.9360649         -1.05515
594
5     0.86772438       1.42787171 -0.38073867     -0.5538212  0.9391720         -0.65590
076
6     0.82124746       1.11863021  1.27788912     -0.4840963  9.1464421         -0.69065
168
7     0.73451686       0.07234959  0.59494454     -0.5811424  0.5698290         -0.83361
426
8     0.11764317      -0.46742334  0.17679425     -0.2890164 -0.2623345         -0.34797
522
9    -0.34071279      -0.36947415  0.09816867      0.5931168 -0.1081189          1.33208
909
10   -0.07325184      -0.35568087  0.43746397      1.8714238 -0.1595805          0.59484
847
   total.sulfur.dioxide    density         pH   sulphates     alcohol
1            0.29004495 -0.5169654  0.9217350  0.04654917  0.03922083
2           -0.11704767 -1.3847361 -0.1755866 -0.41216890  1.48121606
3           -1.32138348  0.3218454  1.3309727  0.44835721  0.01994967
4           -1.45865214  1.3285041 -0.4481953  1.27654757  0.08450639
5           -0.78349922  0.8732173  0.2426831  0.32677333 -0.69962939
6           -0.71319742  0.8112587 -0.9132206  3.70100796 -0.87347243
7           -1.20634843  0.3460569  0.4345013  1.60091644  0.53912490
8            0.02976291 -0.4378959 -0.7589997 -0.50875294 -0.09617225
9            1.18147258  0.2916537 -0.1964055 -0.30443681 -0.57247138
10           0.84377609  1.2190533 -0.6268219 -0.21005161 -0.99972852
```

By converting these clusters back (through multiplying by our sigma + mu), results are far more relevant and interpretable.

Looking at Cluster 1:

Hide

```
clust1$center[1,]*sigma + mu
```

```
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar
chlorides
         -0.67882313          -0.57877938          -0.15159923          -0.43147568
-0.37351088
 free.sulfur.dioxide total.sulfur.dioxide              density                   pH
sulphates
          0.14153925           0.29004495          -0.51696536           0.92173499
0.04654917
             alcohol
          0.03922083
```

The cluster center value for volatile acidity is positive. This indicates that, on average, the wines within cluster 1 tend to have a slightly higher volatile acidity value versus the overall mean value of the entire data set. Conversely, the cluster center value for alcohol is negative, indicating that, on average, the wines within cluster 1 tend to have a slightly lower alcohol concentration, in comparison to the mean value of the entire data set.

Cluster 6:

Hide

```
clust1$center[6,]*sigma + mu
```

```
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar
chlorides
           0.8212475            1.1186302            1.2778891          -0.4840963
9.1464421
 free.sulfur.dioxide total.sulfur.dioxide              density                   pH
sulphates
          -0.6906517           -0.7131974            0.8112587          -0.9132206
3.7010080
             alcohol
          -0.8734724
```

Looking at the same two variables for ease of comparison:

The cluster center value for alcohol is positive. This indicates that, on average, the wines within cluster 6 tend to have a slightly higher alcohol concentration versus the overall mean value of the entire data set. Conversely, the cluster center value for volatile acidity is negative, indicating that, on average, the wines within cluster 6 tend to have a slightly lower volatile acidity values, in comparison to the mean value of the entire data set.

Looking at which wines belong to which clusters:

Unlike other data sets, this was not useful as the names of the specific wines were taken out of the data set for privacy reasons. As a result, when we look at cluster one, we simply get the indexes which do not reveal anything:

Hide

```
#Which wines are in which clusters?
which(clust1$cluster == 1)
```

```
  [1]   50  495  891 1029 1132 1145 1236 1584 1601 1608 1613 1616 1622 1624 1626 1627 16
31 1633 1643 1644 1649
 [22] 1650 1651 1652 1653 1659 1665 1668 1672 1674 1680 1709 1716 1720 1725 1728 1729 17
30 1740 1746 1793 1803
 [43] 1804 1809 1810 1811 1813 1814 1822 1823 1824 1826 1828 1838 1845 1849 1850 1853 18
54 1855 1860 1867 1868
 [64] 1869 1891 1893 1895 1900 1902 1903 1909 1910 1911 1912 1918 1919 1920 1927 1938 19
39 1943 1944 1945 1946
 [85] 1950 1957 1959 1962 1964 1965 1966 1976 1978 1979 1986 1988 1992 1997 2000 2002 20
08 2019 2036 2037 2039
[106] 2041 2046 2049 2051 2052 2053 2054 2058 2071 2073 2074 2079 2081 2085 2086 2088 20
94 2098 2102 2103 2107
[127] 2109 2110 2115 2118 2129 2130 2138 2139 2140 2142 2145 2146 2148 2157 2161 2181 21
84 2187 2188 2191 2195
[148] 2198 2222 2224 2229 2232 2234 2235 2236 2238 2239 2240 2248 2254 2258 2266 2268 22
73 2286 2289 2290 2314
[169] 2315 2318 2319 2321 2326 2328 2341 2342 2358 2359 2364 2368 2369 2370 2388 2391 23
93 2398 2399 2411 2412
[190] 2422 2423 2426 2430 2434 2443 2444 2445 2450 2452 2454 2455 2462 2464 2466 2469 24
79 2487 2489 2490 2491
[211] 2492 2494 2495 2500 2510 2511 2513 2515 2516 2520 2524 2525 2542 2545 2552 2557 25
58 2563 2568 2573 2576
[232] 2578 2581 2584 2585 2588 2593 2596 2603 2604 2605 2606 2608 2611 2624 2625 2626 26
28 2630 2638 2646 2656
[253] 2658 2659 2664 2669 2680 2684 2686 2692 2695 2697 2702 2705 2706 2716 2718 2728 27
36 2753 2754 2755 2760
[274] 2765 2769 2770 2773 2777 2782 2783 2791 2796 2800 2801 2803 2804 2808 2813 2820 28
21 2833 2834 2840 2841
[295] 2844 2848 2850 2855 2856 2861 2862 2864 2865 2879 2889 2892 2896 2911 2917 2918 29
22 2923 2924 2931 2935
[316] 2940 2952 2954 2959 2960 2961 2965 2978 2981 2984 2985 2990 2993 3002 3003 3014 30
16 3022 3029 3031 3061
[337] 3065 3075 3080 3081 3082 3086 3090 3122 3153 3155 3172 3175 3176 3189 3204 3207 32
15 3219 3225 3230 3248
[358] 3249 3265 3271 3276 3278 3279 3281 3289 3292 3294 3298 3299 3300 3305 3306 3308 33
11 3313 3315 3316 3319
[379] 3320 3321 3323 3324 3329 3349 3351 3353 3354 3362 3363 3371 3377 3378 3379 3381 33
84 3391 3394 3397 3399
[400] 3403 3410 3416 3425 3434 3437 3440 3441 3449 3450 3452 3460 3464 3470 3471 3484 35
06 3522 3534 3538 3548
[421] 3553 3556 3568 3569 3571 3576 3592 3600 3601 3602 3604 3607 3608 3611 3612 3620 36
21 3623 3632 3636 3638
[442] 3641 3642 3644 3649 3657 3661 3663 3668 3670 3671 3672 3675 3676 3677 3678 3680 36
84 3685 3689 3699 3704
[463] 3713 3728 3730 3734 3743 3746 3758 3760 3777 3784 3788 3794 3801 3807 3811 3813 38
18 3833 3838 3847 3851
[484] 3855 3868 3870 3872 3873 3874 3877 3880 3881 3890 3892 3906 3922 3923 3940 3943 39
53 3954 3956 3962 3964
[505] 3969 3970 3979 3980 3981 3983 3986 3987 3990 3992 3993 3995 3999 4007 4009 4010 40
13 4014 4016 4022 4023
[526] 4024 4025 4026 4028 4039 4047 4048 4049 4052 4053 4067 4068 4071 4080 4095 4099 41
12 4116 4121 4126 4127
```

```
[547] 4128 4130 4137 4138 4139 4144 4145 4147 4150 4153 4154 4155 4157 4158 4167 4169 41
70 4172 4187 4191 4196
[568] 4198 4200 4203 4204 4206 4212 4226 4227 4229 4230 4245 4246 4247 4257 4261 4284 42
93 4295 4311 4318 4323
[589] 4326 4348 4350 4357 4358 4371 4377 4402 4403 4405 4409 4412 4413 4419 4425 4427 44
28 4434 4435 4441 4447
[610] 4448 4453 4455 4456 4458 4459 4462 4464 4465 4466 4472 4474 4493 4495 4500 4504 45
21 4530 4534 4536 4546
[631] 4547 4551 4558 4564 4591 4602 4621 4625 4632 4636 4674 4684 4690 4698 4709 4728 47
36 4743 4759 4777 4782
[652] 4785 4793 4808 4826 4829 4834 4867 4874 4894 4911 4917 4919 4927 4933 4949 4954 49
61 4972 4974 4983 4999
[673] 5006 5009 5053 5057 5072 5081 5094 5096 5098 5101 5104 5105 5106 5112 5121 5124 51
25 5143 5144 5149 5156
[694] 5186 5191 5192 5193 5198 5201 5202 5207 5211 5212 5252 5254 5258 5261 5274 5275 52
95 5309 5315 5320 5323
[715] 5324 5326 5327 5338 5349 5351 5352 5353 5356 5357 5358 5378 5394 5418 5419 5426 54
30 5453 5454 5457 5475
[736] 5532 5534 5536 5539 5541 5544 5545 5548 5550 5556 5558 5559 5566 5578 5579 5583 55
88 5593 5594 5596 5609
[757] 5624 5629 5677 5678 5682 5683 5706 5709 5712 5733 5735 5742 5772 5777 5788 5803 58
07 5827 5834 5839 5849
[778] 5851 5859 5898 5904 5905 5907 5924 5928 5939 5942 5943 5956 5957 5975 5977 5980 59
83 5986 6029 6034 6035
[799] 6037 6038 6044 6045 6059 6061 6066 6070 6083 6087 6106 6119 6125 6132 6156 6158 61
59 6162 6163 6165 6167
[820] 6178 6181 6185 6188 6231 6235 6236 6238 6246 6247 6263 6264 6267 6268 6269 6271 62
72 6274 6275 6279 6284
[841] 6292 6303 6305 6307 6308 6311 6312 6317 6318 6321 6324 6326 6331 6332 6333 6341 63
42 6343 6344 6350 6353
[862] 6367 6375 6380 6381 6395 6396 6401 6405 6409 6410 6411 6425 6426 6431 6432 6433 64
35 6438 6439 6441 6452
[883] 6453 6464 6466 6468 6474 6476 6482 6486 6491 6492
```

Next I looked at the mean value of each wine quality within the clusters. I was hoping to find some large variations within this summary, so that the quality of wines were very distinguishable. Notwithstanding, all of the wine qualities are within 1 point of each other:

Hide

```
average_quality_cluster1 <- mean(wine$quality[clust1$cluster == 1])
print(paste("Average quality of wines in cluster 1:", average_quality_cluster1))
```

```
[1] "Average quality of wines in cluster 1: 6.03923766816143"
```

Hide

```
average_quality_cluster2 <- mean(wine$quality[clust1$cluster == 2])
print(paste("Average quality of wines in cluster 2:", average_quality_cluster2))
```

```
[1] "Average quality of wines in cluster 2: 6.46709916589435"
```

Hide

```
average_quality_cluster3 <- mean(wine$quality[clust1$cluster == 3])
print(paste("Average quality of wines in cluster 3:", average_quality_cluster3))
```

```
[1] "Average quality of wines in cluster 3: 5.45008460236887"
```

Hide

```
average_quality_cluster4 <- mean(wine$quality[clust1$cluster == 4])
print(paste("Average quality of wines in cluster 4:", average_quality_cluster4))
```

```
[1] "Average quality of wines in cluster 4: 5.90804597701149"
```

Hide

```
average_quality_cluster5 <- mean(wine$quality[clust1$cluster == 5])
print(paste("Average quality of wines in cluster 5:", average_quality_cluster5))
```

```
[1] "Average quality of wines in cluster 5: 5.26059322033898"
```

Hide

```
average_quality_cluster6 <- mean(wine$quality[clust1$cluster == 6])
print(paste("Average quality of wines in cluster 6:", average_quality_cluster6))
```

```
[1] "Average quality of wines in cluster 6: 5.16666666666667"
```

Hide

```
average_quality_cluster7 <- mean(wine$quality[clust1$cluster == 7])
print(paste("Average quality of wines in cluster 7:", average_quality_cluster7))
```

```
[1] "Average quality of wines in cluster 7: 6.06547619047619"
```

Hide

```
average_quality_cluster8 <- mean(wine$quality[clust1$cluster == 8])
print(paste("Average quality of wines in cluster 2:", average_quality_cluster8))
```

```
[1] "Average quality of wines in cluster 2: 5.65196998123827"
```

Hide

```
average_quality_cluster9 <- mean(wine$quality[clust1$cluster == 9])
print(paste("Average quality of wines in cluster 9:", average_quality_cluster9))
```

```
[1] "Average quality of wines in cluster 9: 5.57752808988764"
```

Hide

```
average_quality_cluster10 <- mean(wine$quality[clust1$cluster == 10])
print(paste("Average quality of wines in cluster 10:", average_quality_cluster10))
```

```
[1] "Average quality of wines in cluster 10: 5.69204545454545"
```

Since the question addressed both the quality and the color, I have repeated the mean cluster values for the color numeric column I created, previously. These results are far more promising and distinguishable: clusters 1-5 are all below 0.2, indicating that the majority of the wines within these clusters are white wine, and the remaining 5 clusters are all above 0.99, indicating that they are mainly red wines.

Hide

```
average_color_cluster1 <- mean(wine$color_numeric[clust1$cluster == 1])
print(paste("Average color (numeric) of wines in cluster 1:", average_color_cluster1))
```

```
[1] "Average color (numeric) of wines in cluster 1: 0.991031390134529"
```

Hide

```
average_color_cluster2 <- mean(wine$color_numeric[clust1$cluster == 2])
print(paste("Average color (numeric) of wines in cluster 2:", average_color_cluster2))
```

```
[1] "Average color (numeric) of wines in cluster 2: 0.990732159406858"
```

Hide

```
average_color_cluster3 <- mean(wine$color_numeric[clust1$cluster == 3])
print(paste("Average color (numeric) of wines in cluster 3:", average_color_cluster3))
```

```
[1] "Average color (numeric) of wines in cluster 3: 0.0592216582064298"
```

Hide

```
average_color_cluster4 <- mean(wine$color_numeric[clust1$cluster == 4])
print(paste("Average color (numeric) of wines in cluster 4:", average_color_cluster4))
```

```
[1] "Average color (numeric) of wines in cluster 4: 0.0114942528735632"
```

Hide

```
average_color_cluster5 <- mean(wine$color_numeric[clust1$cluster == 5])
print(paste("Average color (numeric) of wines in cluster 5:", average_color_cluster5))
```

```
[1] "Average color (numeric) of wines in cluster 5: 0.0741525423728814"
```

Hide

```
average_color_cluster6 <- mean(wine$color_numeric[clust1$cluster == 6])
print(paste("Average color (numeric) of wines in cluster 6:", average_color_cluster6))
```

```
[1] "Average color (numeric) of wines in cluster 6: 0.166666666666667"
```

Hide

```
average_color_cluster7 <- mean(wine$color_numeric[clust1$cluster == 7])
print(paste("Average color (numeric) of wines in cluster 7:", average_color_cluster7))
```

```
[1] "Average color (numeric) of wines in cluster 7: 0.113095238095238"
```

Hide

```
average_color_cluster8 <- mean(wine$color_numeric[clust1$cluster == 8])
print(paste("Average color (numeric) of wines in cluster 8:", average_color_cluster8))
```

```
[1] "Average color (numeric) of wines in cluster 8: 0.99718574108818"
```

Hide

```
average_color_cluster9 <- mean(wine$color_numeric[clust1$cluster == 9])
print(paste("Average color (numeric) of wines in cluster 9:", average_color_cluster9))
```

```
[1] "Average color (numeric) of wines in cluster 9: 0.996629213483146"
```

Hide

```
average_color_cluster10 <- mean(wine$color_numeric[clust1$cluster == 10])
print(paste("Average color (numeric) of wines in cluster 10:", average_color_cluster10))
```

```
[1] "Average color (numeric) of wines in cluster 10: 0.998863636363636"
```

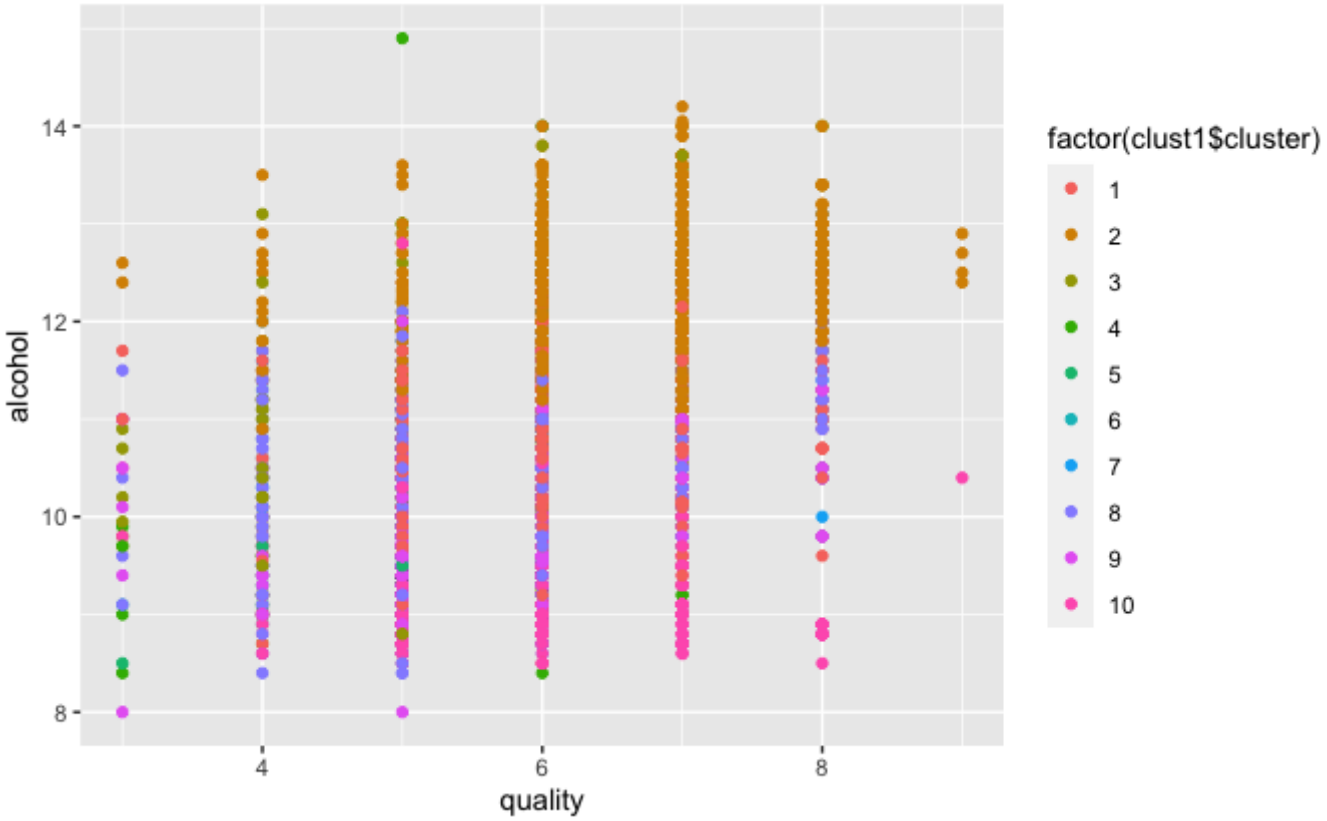
Visual Depictions (Graphs):

To go with the mean values, I have also produced some plots to that color is easily distinguishable, and quality is not *as distinguishable.*

Plot 1: Scatterplot showing different wine quality's alcohol levels:

Hide

```
# A few plots with cluster membership shown
# qplot is in the ggplot2 library
qplot(quality, alcohol, data=wine, color=factor(clust1$cluster))
```

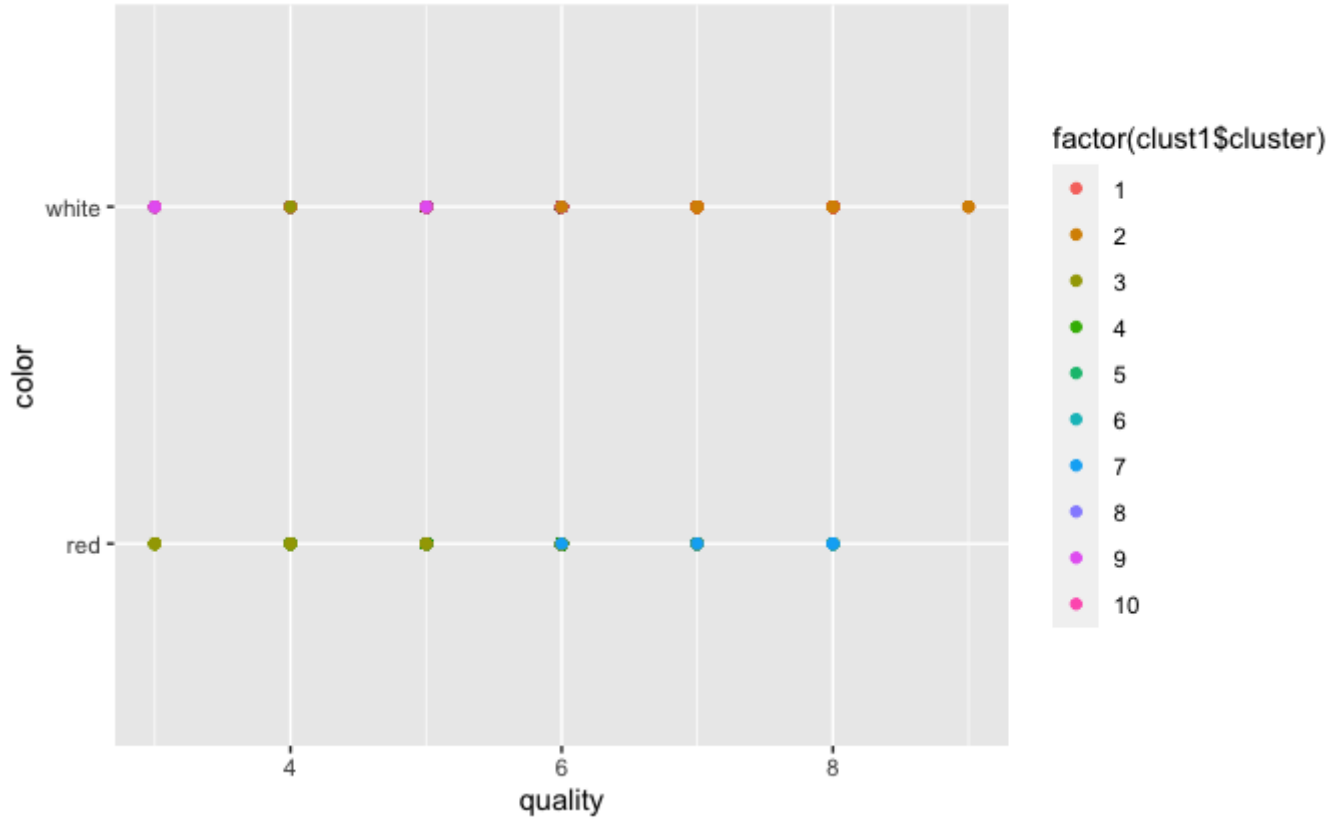


For the first plot I actually played around with plotting each of the chemical elements against quality, and they all had similar patterns, so I have only kept alcohol. Looking at cluster 6 (turquoise color), it is clear that all types of qualities are represented in this cluster. There are turquoise points at every quality level. However, cluster six seems to have higher alcohol levels in comparison to the pink and purple points (clusters 9-10). This cements the findings from the mean values for quality, in that quality of wines is not easily distinguishable through k-means clustering.

Plot 2: Scatterplot showing just color and quality (cluster centers)

Hide

```
qplot(quality, color, data=wine, color=factor(clust1$cluster))
```



The second scatterplot shows just color and quality, using the cluster centers. From this plot it seems like we can distinguish between quality a little better, when we separate the red and white wines. For example, cluster 2 seems to have higher quality red wines, while cluster 4 has lower quality red wines. Notwithstanding, we have to be careful here as there are some clusters that we cannot easily see (such as clusters 9 and 10) as they have been plotted underneath cluster 7, for example, due to overlaps.

K-Means Summary:

From the k-means clustering method, red and white wines are easily distinguished, using only the “unsupervised” information contained in the data on chemical properties. However, this technique does not also seem capable of distinguishing between higher and lower quality wines.

K-Means++ Clustering Model:

Using the same values in X (i.e. all of the chemical elements), I have implemented a kmeans++ model.

Hide

```
# Using kmeans++ initialization
clust2 = KMeans_rcpp(X, clusters=10, num_init=25, initializer = 'kmeans++')
cat("Total SSE for cluster kmeans++:", clust2$total_SSE)
```

```
Total SSE for cluster kmeans++: 71456
```

Hide

```
cat("Total SSE for cluster kmeans:", clust1$totss)
```

```
Total SSE for cluster kmeans: 71456
```

I have now repeated the same analysis for kmeans++ as I did for kmeans:

Hide

```
print("K-means++ Quality and Color mean values per cluster")
```

```
[1] "K-means++ Quality and Color mean values per cluster"
```

Hide

```
average_quality_cluster1 <- mean(wine$quality[clust2$cluster == 1])
print(paste("Average quality of wines in cluster 1:", average_quality_cluster1))
```

```
[1] "Average quality of wines in cluster 1: 5.58041958041958"
```

Hide

```
average_quality_cluster2 <- mean(wine$quality[clust2$cluster == 2])
print(paste("Average quality of wines in cluster 2:", average_quality_cluster2))
```

```
[1] "Average quality of wines in cluster 2: 5.41071428571429"
```

Hide

```
average_quality_cluster3 <- mean(wine$quality[clust2$cluster == 3])
print(paste("Average quality of wines in cluster 3:", average_quality_cluster3))
```

```
[1] "Average quality of wines in cluster 3: 5.76493256262042"
```

Hide

```
average_quality_cluster4 <- mean(wine$quality[clust2$cluster == 4])
print(paste("Average quality of wines in cluster 4:", average_quality_cluster4))
```

```
[1] "Average quality of wines in cluster 4: 5.82007952286282"
```

Hide

```
average_quality_cluster5 <- mean(wine$quality[clust2$cluster == 5])
print(paste("Average quality of wines in cluster 5:", average_quality_cluster5))
```

```
[1] "Average quality of wines in cluster 5: 5.37037037037037"
```

Hide

```
average_quality_cluster6 <- mean(wine$quality[clust2$cluster == 6])
print(paste("Average quality of wines in cluster 6:", average_quality_cluster6))
```

```
[1] "Average quality of wines in cluster 6: 5.70339976553341"
```

Hide

```
average_quality_cluster7 <- mean(wine$quality[clust2$cluster == 7])
print(paste("Average quality of wines in cluster 7:", average_quality_cluster7))
```

```
[1] "Average quality of wines in cluster 7: 5.29166666666667"
```

Hide

```
average_quality_cluster8 <- mean(wine$quality[clust2$cluster == 8])
print(paste("Average quality of wines in cluster 2:", average_quality_cluster8))
```

```
[1] "Average quality of wines in cluster 2: 5.83034872761546"
```

Hide

```
average_quality_cluster9 <- mean(wine$quality[clust2$cluster == 9])
print(paste("Average quality of wines in cluster 9:", average_quality_cluster9))
```

```
[1] "Average quality of wines in cluster 9: 5.92700729927007"
```

Hide

```
average_quality_cluster10 <- mean(wine$quality[clust2$cluster == 10])
print(paste("Average quality of wines in cluster 10:", average_quality_cluster10))
```

```
[1] "Average quality of wines in cluster 10: 6.50844091360477"
```

Hide

```
average_color_cluster1 <- mean(wine$color_numeric[clust2$cluster == 1])
print(paste("Average color (numeric) of wines in cluster 1:", average_color_cluster1))
```

```
[1] "Average color (numeric) of wines in cluster 1: 0.996503496503496"
```

Hide

```
average_color_cluster2 <- mean(wine$color_numeric[clust2$cluster == 2])
print(paste("Average color (numeric) of wines in cluster 2:", average_color_cluster2))
```

```
[1] "Average color (numeric) of wines in cluster 2: 0.8125"
```

Hide

```
average_color_cluster3 <- mean(wine$color_numeric[clust2$cluster == 3])
print(paste("Average color (numeric) of wines in cluster 3:", average_color_cluster3))
```

```
[1] "Average color (numeric) of wines in cluster 3: 0.0635838150289017"
```

Hide

```
average_color_cluster4 <- mean(wine$color_numeric[clust2$cluster == 4])
print(paste("Average color (numeric) of wines in cluster 4:", average_color_cluster4))
```

```
[1] "Average color (numeric) of wines in cluster 4: 0.984095427435388"
```

Hide

```
average_color_cluster5 <- mean(wine$color_numeric[clust2$cluster == 5])
print(paste("Average color (numeric) of wines in cluster 5:", average_color_cluster5))
```

```
[1] "Average color (numeric) of wines in cluster 5: 0.0459770114942529"
```

Hide

```
average_color_cluster6 <- mean(wine$color_numeric[clust2$cluster == 6])
print(paste("Average color (numeric) of wines in cluster 6:", average_color_cluster6))
```

```
[1] "Average color (numeric) of wines in cluster 6: 0.998827667057444"
```

Hide

```
average_color_cluster7 <- mean(wine$color_numeric[clust2$cluster == 7])
print(paste("Average color (numeric) of wines in cluster 7:", average_color_cluster7))
```

```
[1] "Average color (numeric) of wines in cluster 7: 0"
```

Hide

```
average_color_cluster8 <- mean(wine$color_numeric[clust2$cluster == 8])
print(paste("Average color (numeric) of wines in cluster 8:", average_color_cluster8))
```

```
[1] "Average color (numeric) of wines in cluster 8: 0.997172478793591"
```

Hide

```
average_color_cluster9 <- mean(wine$color_numeric[clust2$cluster == 9])
print(paste("Average color (numeric) of wines in cluster 9:", average_color_cluster9))
```

```
[1] "Average color (numeric) of wines in cluster 9: 0.0109489051094891"
```

Hide

```
average_color_cluster10 <- mean(wine$color_numeric[clust2$cluster == 10])
print(paste("Average color (numeric) of wines in cluster 10:", average_color_cluster10))
```

```
[1] "Average color (numeric) of wines in cluster 10: 0.973187686196624"
```

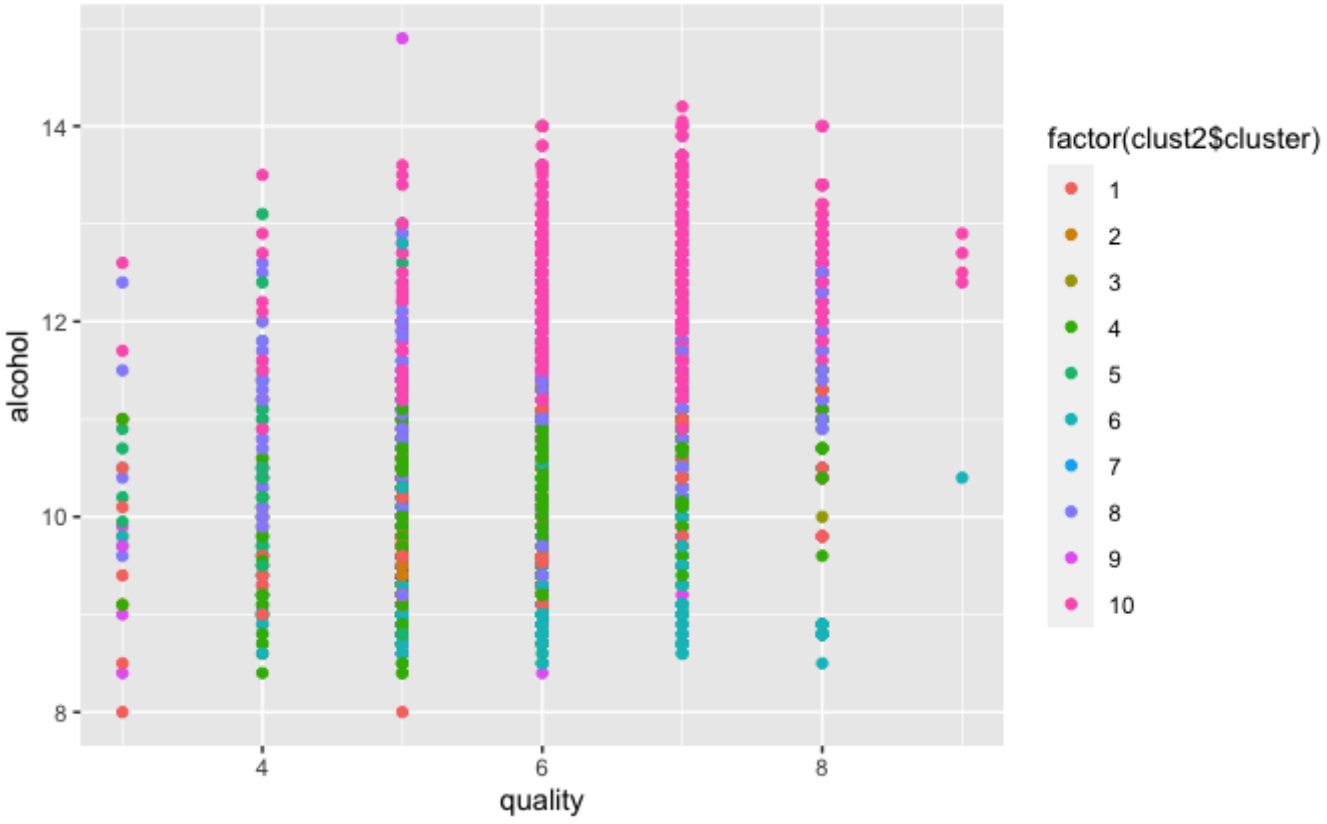
The kmeans++ clustering method seems to perform worse in distinguishing quality, if we simply look at the mean values. Aside from cluster 10, which is much higher, the other clusters seem to average around 5.3-5.9. This is a smaller range than before. Looking at the average color, it seems to separate some clusters better, and others worse. For example, cluster 7 is entirely white wines, and cluster 8 is very nearly all red wines. This is a huge success! Yet, cluster 2 seems to be more of a mix, with a value of 0.8125, which is the closest value to 0.5 out of kmeans and kmeans++ clusters.

Plotting:

A similar issue of quality is shown in the same plot as before: clusters have representation in each quality score, as most clearly shown by cluster 10 in the scatterplot below:

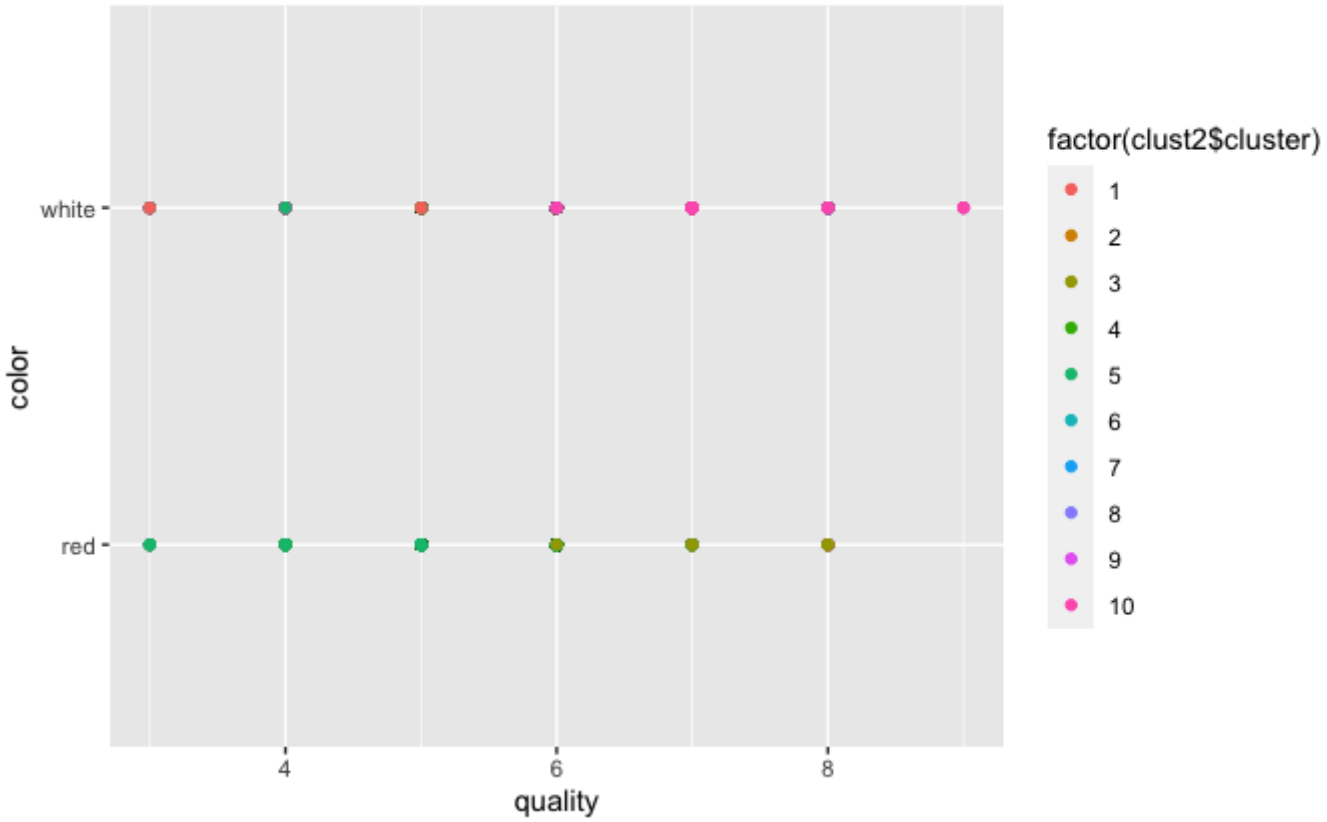
Hide

```
# A few plots with cluster membership shown
# qplot is in the ggplot2 library
qplot(quality, alcohol, data=wine, color=factor(clust2$cluster))
```



Hide

```
qplot(quality, color, data=wine, color=factor(clust2$cluster))
```



Scatterplot 2 shows very similar results to kmeans.

Summary for kmeans++:

In summary, kmeans++ is preferred to kmeans, as we have some more robust results when identifying color wines within the clusters. This is unsurprising given that K-Means++ is merely an improved version of the K-Means algorithm that enhances the initialization step, leading to faster convergence and better clustering quality. Notwithstanding, I am still unsatisfied with the quality side of things, so hopefully this can be improved in the subsequent models.

Hierarchical Clustering Model:

Next, hierarchical clustering was executed. As the summary below shows, these clusters are not evenly split, and there is just 1 wine in cluster 9 and cluster 10.

Hide

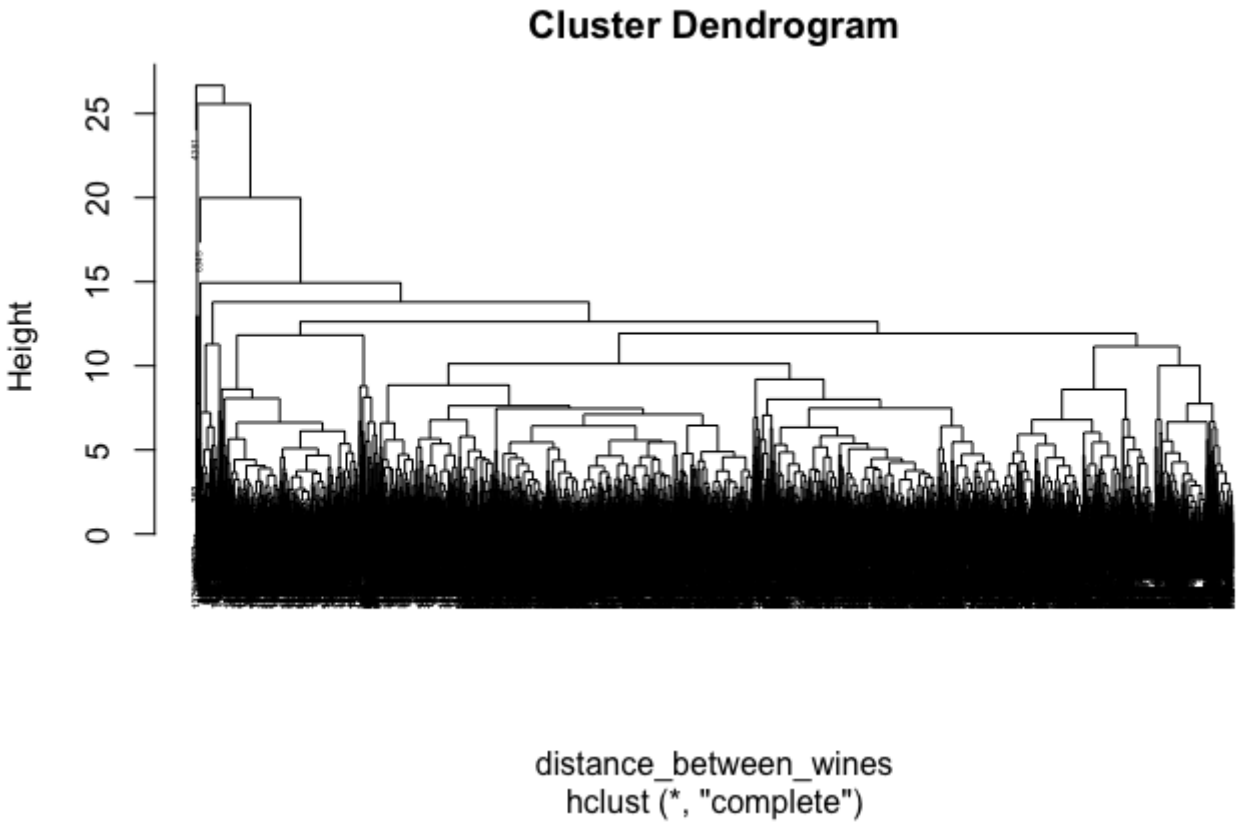
```
# First form a pairwise distance matrix
distance_between_wines = dist(X)

# Now run hierarchical clustering
h1 = hclust(distance_between_wines, method='complete')

# Cut the tree into 10 clusters
cluster3 = cutree(h1, k=10)
summary(factor(cluster3))
```

```
   1    2    3    4    5    6    7    8    9   10
 866 3899  128  128   22 1448    2    2    1    1
```

Next, I have plotted a dendogram to visually depict this form of clustering. However, as shown in the plot, it is really uninterpretable. Thus, I will look at the cluster center means again, and see if there are any obvious patterns for color and quality.

Hide

```
# Plot the dendrogram
plot(h1, cex=0.3)
```



While the dendogram is not very helpful, the cluster mean values for color and quality are. Where most of the wines are contained (in cluster 2), we see that there is a wider mix of red and white wines (0.8 numeric values), with a quality value of six. Since this class is so large, I also printed the range and medium values to see if the mean was skewed (for quality). The median value is 6, however, the range is 6, with the lowest value 3, and highest 9. This is the same range as the entire data set, which insinuates that the highest and lowest quality wines from wine.csv are contained within cluster 2. Obviously, this is not a good sign by way of distinguishing quality between clusters.

Hide

```
print("Hierarchical Quality and Color mean values per cluster")
```

```
[1] "Hierarchical Quality and Color mean values per cluster"
```

Hide

```
average_quality_cluster1 <- mean(wine$quality[cluster3 == 1])
print(paste("Average quality of wines in cluster 1:", average_quality_cluster1))
```

```
[1] "Average quality of wines in cluster 1: 5.39376443418014"
```

Hide

```
average_quality_cluster2 <- mean(wine$quality[cluster3 == 2])
print(paste("Average quality of wines in cluster 2:", average_quality_cluster2))
```

```
[1] "Average quality of wines in cluster 2: 6.00615542446781"
```

Hide

```
median_quality_cluster2 <- median(wine$quality[cluster3 == 2])
print(paste("Median quality of wines in cluster 2:", median_quality_cluster2))
```

```
[1] "Median quality of wines in cluster 2: 6"
```

Hide

```
min_quality_cluster2 <- min(wine$quality[cluster3 == 2])
print(paste("Minimum quality of wines in cluster 2:", min_quality_cluster2))
```

```
[1] "Minimum quality of wines in cluster 2: 3"
```

Hide

```
max_quality_cluster2 <- max(wine$quality[cluster3 == 2])
print(paste("Maximum quality of wines in cluster 2:", max_quality_cluster2))
```

```
[1] "Maximum quality of wines in cluster 2: 9"
```

Hide

```
average_quality_cluster3 <- mean(wine$quality[cluster3 == 3])
print(paste("Average quality of wines in cluster 3:", average_quality_cluster3))
```

```
[1] "Average quality of wines in cluster 3: 5.875"
```

Hide

```
average_quality_cluster4 <- mean(wine$quality[cluster3 == 4])
print(paste("Average quality of wines in cluster 4:", average_quality_cluster4))
```

```
[1] "Average quality of wines in cluster 4: 5.484375"
```

Hide

```
average_quality_cluster5 <- mean(wine$quality[cluster3 == 5])
print(paste("Average quality of wines in cluster 5:", average_quality_cluster5))
```

```
[1] "Average quality of wines in cluster 5: 5.36363636363636"
```

Hide

```
average_quality_cluster6 <- mean(wine$quality[cluster3 == 6])
print(paste("Average quality of wines in cluster 6:", average_quality_cluster6))
```

```
[1] "Average quality of wines in cluster 6: 5.60151933701657"
```

Hide

```
average_quality_cluster7 <- mean(wine$quality[cluster3 == 7])
print(paste("Average quality of wines in cluster 7:", average_quality_cluster7))
```

```
[1] "Average quality of wines in cluster 7: 4.5"
```

Hide

```
average_quality_cluster8 <- mean(wine$quality[cluster3 == 8])
print(paste("Average quality of wines in cluster 2:", average_quality_cluster8))
```

```
[1] "Average quality of wines in cluster 2: 6"
```

Hide

```
average_quality_cluster9 <- mean(wine$quality[cluster3 == 9])
print(paste("Average quality of wines in cluster 9:", average_quality_cluster9))
```

```
[1] "Average quality of wines in cluster 9: 6"
```

Hide

```
average_quality_cluster10 <- mean(wine$quality[cluster3 == 10])
print(paste("Average quality of wines in cluster 10:", average_quality_cluster10))
```

```
[1] "Average quality of wines in cluster 10: 3"
```

Similar to kmeans and kmeans++, hierarchical clustering does a good job in distinguishing between colors. Aside from clusters 2 and 4, the color is more distinguishable than before. In fact, clusters 5,7,8,9, and 10 all include just one color of wine. This is somewhat expected for these clusters since they contain very few wines, particularly clusters 7-10, however, still a good sign.

Hide

```
average_color_cluster1 <- mean(wine$color_numeric[cluster3 == 1])
print(paste("Average color (numeric) of wines in cluster 1:", average_color_cluster1))
```

```
[1] "Average color (numeric) of wines in cluster 1: 0.0069284064665127"
```

Hide

```
average_color_cluster2 <- mean(wine$color_numeric[cluster3 == 2])
print(paste("Average color (numeric) of wines in cluster 2:", average_color_cluster2))
```

```
[1] "Average color (numeric) of wines in cluster 2: 0.857912285201334"
```

Hide

```
average_color_cluster3 <- mean(wine$color_numeric[cluster3 == 3])
print(paste("Average color (numeric) of wines in cluster 3:", average_color_cluster3))
```

```
[1] "Average color (numeric) of wines in cluster 3: 0.0078125"
```

Hide

```
average_color_cluster4 <- mean(wine$color_numeric[cluster3 == 4])
print(paste("Average color (numeric) of wines in cluster 4:", average_color_cluster4))
```

```
[1] "Average color (numeric) of wines in cluster 4: 0.796875"
```

Hide

```
average_color_cluster5 <- mean(wine$color_numeric[cluster3 == 5])
print(paste("Average color (numeric) of wines in cluster 5:", average_color_cluster5))
```

```
[1] "Average color (numeric) of wines in cluster 5: 0"
```

Hide

```
average_color_cluster6 <- mean(wine$color_numeric[cluster3 == 6])
print(paste("Average color (numeric) of wines in cluster 6:", average_color_cluster6))
```

```
[1] "Average color (numeric) of wines in cluster 6: 0.994475138121547"
```

Hide

```
average_color_cluster7 <- mean(wine$color_numeric[cluster3 == 7])
print(paste("Average color (numeric) of wines in cluster 7:", average_color_cluster7))
```

```
[1] "Average color (numeric) of wines in cluster 7: 0"
```

Hide

```
average_color_cluster8 <- mean(wine$color_numeric[cluster3 == 8])
print(paste("Average color (numeric) of wines in cluster 8:", average_color_cluster8))
```

```
[1] "Average color (numeric) of wines in cluster 8: 1"
```

Hide

```
average_color_cluster9 <- mean(wine$color_numeric[cluster3 == 9])
print(paste("Average color (numeric) of wines in cluster 9:", average_color_cluster9))
```

```
[1] "Average color (numeric) of wines in cluster 9: 1"
```

Hide

```
average_color_cluster10 <- mean(wine$color_numeric[cluster3 == 10])
print(paste("Average color (numeric) of wines in cluster 10:", average_color_cluster10))
```

```
[1] "Average color (numeric) of wines in cluster 10: 1"
```

PCA Model:

First, I created a correlation heatmap next which looks at all of our numerical data (i.e. the data minus the wine color column).

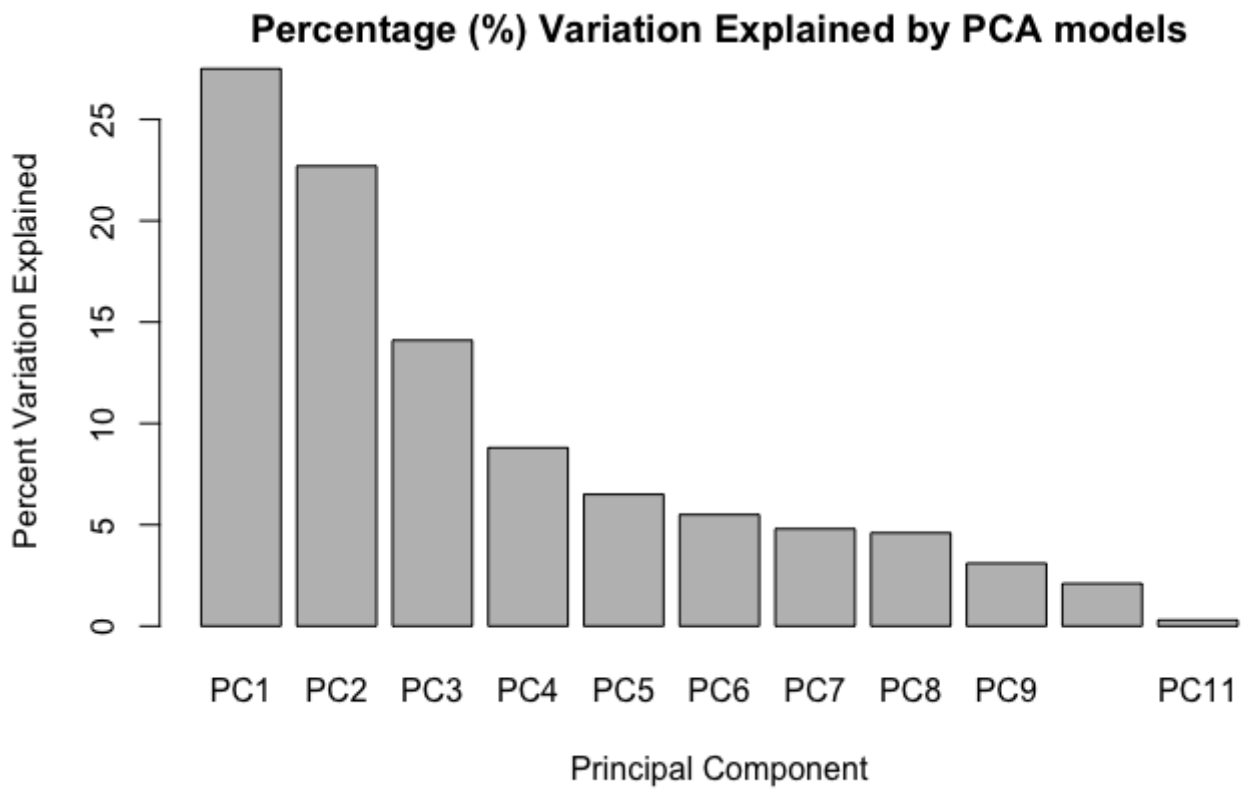
Hide

```
library(ggplot2)
library(dplyr)

pca_result <- prcomp(wine[1:11], scale. = TRUE)

#Create a df with PCA results
pca_data <- data.frame(PC1 = pca_result$x[, 1],
                       PC2 = pca_result$x[, 2],
                       color = wine$color)
pca_data
```

Variance Plot:

Hide

```
pca_var <- pca_result$sdev^2
pca_var_percent <- round(pca_var/sum(pca_var)*100, 1)
barplot(pca_var_percent, xlab="Principal Component",
        ylab="Percent Variation Explained",
        names.arg = paste("PC", 1:length(pca_var_percent), sep = ""),
        main = "Percentage (%) Variation Explained by PCA models")
```



Hide

```
cumulative_var_percent <- cumsum(pca_var_percent)

for (i in 1:length(pca_var_percent)) {
  cat(paste("PC", i, ": Eigenvalue =", pca_var[i], ", Explained Variance =", pca_var_per
cent[i], "%, Cumulative Variance =", cumulative_var_percent[i], "%\n"))
}
```

```
PC 1 : Eigenvalue = 3.02986864855658 , Explained Variance = 27.5 %, Cumulative Variance
= 27.5 %
PC 2 : Eigenvalue = 2.49382602722136 , Explained Variance = 22.7 %, Cumulative Variance
= 50.2 %
PC 3 : Eigenvalue = 1.55634695306159 , Explained Variance = 14.1 %, Cumulative Variance
= 64.3 %
PC 4 : Eigenvalue = 0.970552078671011 , Explained Variance = 8.8 %, Cumulative Variance
= 73.1 %
PC 5 : Eigenvalue = 0.719874915952048 , Explained Variance = 6.5 %, Cumulative Variance
= 79.6 %
PC 6 : Eigenvalue = 0.607311710889117 , Explained Variance = 5.5 %, Cumulative Variance
= 85.1 %
PC 7 : Eigenvalue = 0.523158763207031 , Explained Variance = 4.8 %, Cumulative Variance
= 89.9 %
PC 8 : Eigenvalue = 0.501510290337152 , Explained Variance = 4.6 %, Cumulative Variance
= 94.5 %
PC 9 : Eigenvalue = 0.337024045392155 , Explained Variance = 3.1 %, Cumulative Variance
= 97.6 %
PC 10 : Eigenvalue = 0.227695764459189 , Explained Variance = 2.1 %, Cumulative Variance
= 99.7 %
PC 11 : Eigenvalue = 0.032830802252775 , Explained Variance = 0.3 %, Cumulative Variance
= 100 %
```

Remember, the “best summary” is the one that preserves as much of the variance in the original data. As a result, I have produced a variance plot above.

From this variance plot (variance scores of the number summaries) we can see that there are 3 distinct bins that have a variance above 15%, and then a slightly smaller fourth bin just below 10%. The other bins are relatively small.

When we look at the summary, we account for just under 90% of the varaince with 7 PCs.

Next, I am going to look at the first few PCs to answer the question: “Which variables does this load heavily on positively and negatively?” Note, I have only printed the first four PCs.

Hide

```
summary(pca_var)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03283 0.41927 0.60731 1.00000 1.26345 3.02987
```

The table above shows a feature-centric view. This looks at the vectors and the direction they point to. Below, I have individually looked at the columns of these tables, with the PC numbers ordered:

Hide

```
wine2 <- wine[, -c(12:14)]
PCAwine = prcomp(wine2, scale=TRUE)
round(PCAwine$rotation[,1:4],2)
```

```
                       PC1   PC2   PC3   PC4
fixed.acidity        -0.24  0.34 -0.43  0.16
volatile.acidity     -0.38  0.12  0.31  0.21
citric.acid           0.15  0.18 -0.59 -0.26
residual.sugar        0.35  0.33  0.16  0.17
chlorides            -0.29  0.32  0.02 -0.24
free.sulfur.dioxide   0.43  0.07  0.13 -0.36
total.sulfur.dioxide  0.49  0.09  0.11 -0.21
density              -0.04  0.58  0.18  0.07
pH                   -0.22 -0.16  0.46 -0.41
sulphates            -0.29  0.19 -0.07 -0.64
alcohol              -0.11 -0.47 -0.26 -0.11
```

Hide

```
loadings_summary = PCAwine$rotation %>%
  as.data.frame() %>%
  rownames_to_column('Features')

# This seems to pick out characteristics of
# well-received dramas with positive loadings?
loadings_summary %>%
  select(Features, PC1) %>%
  arrange(desc(PC1))
```

PC1 is highly loaded on total sulfur dioxide and free sulfur dioxide. It is negatively loaded on pH, fixed acidity, chlorides, and sulphates.

Hide

```
loadings_summary %>%
  select(Features, PC2) %>%
  arrange(desc(PC2))
```

PC2 is highly loaded on density and fixed acidity. It is negatively loaded on pH.

Hide

```
loadings_summary %>%
  select(Features, PC3) %>%
  arrange(desc(PC3))
```

PC3 is highly loaded on pH and volatile acidity. It is negatively loaded on sulphates, alcohol, and fixed acidity.

Hide

```
loadings_summary %>%
  select(Features, PC4) %>%
  arrange(desc(PC4))
```

PC4 is highly loaded on volatile acidity and residual sugar. It is negatively loaded on most other features!

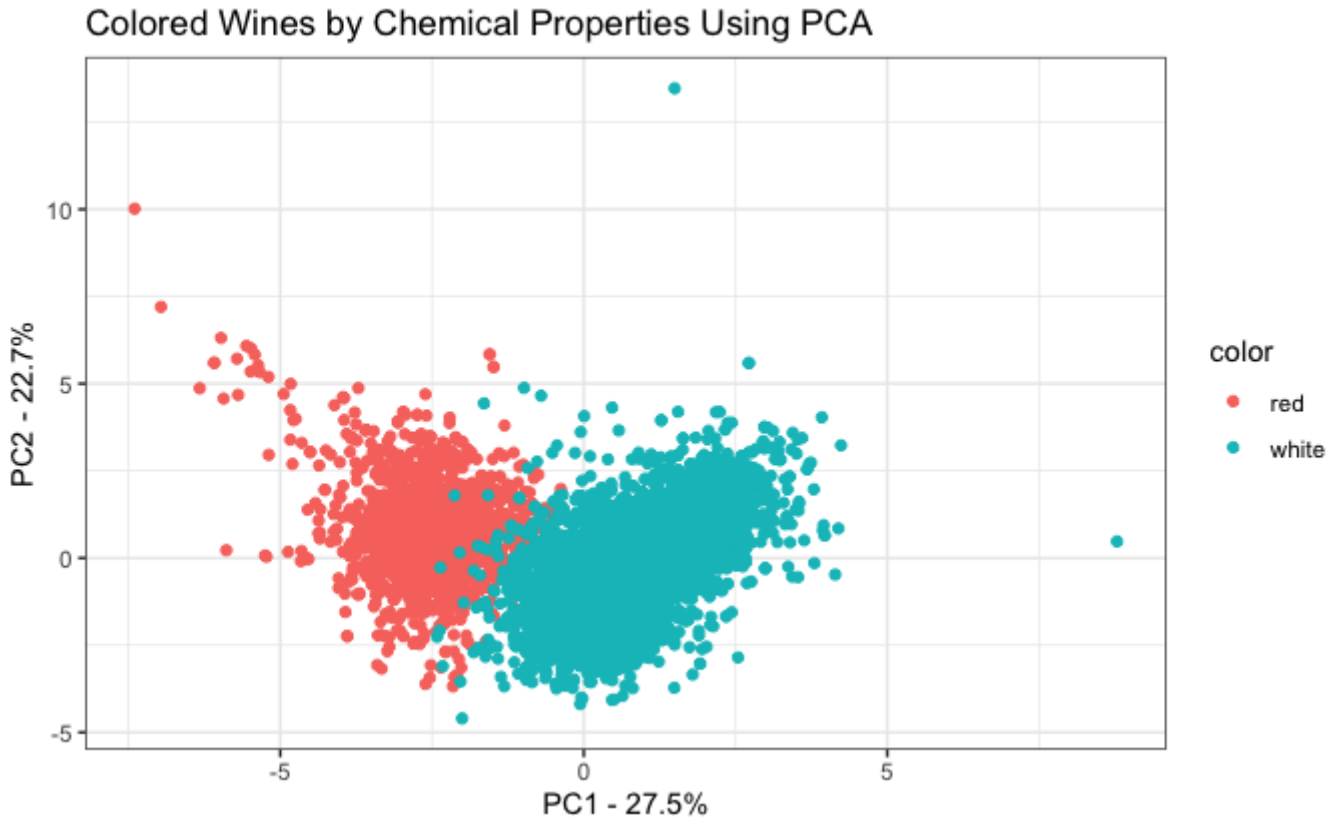
Scatterplot:

For PCA I have created a scatterplot to show that this model distinguishes between red and white wines, without using the mean values. PCA does not return lots of little clusters, so it is far better to graphically show that it recognizes color.

I wanted to keep this plot 2-dimensional, so I have only taken the top 2 PCAs, which account for 50% of the variance. In a perfect world, I would use the first four.

Hide

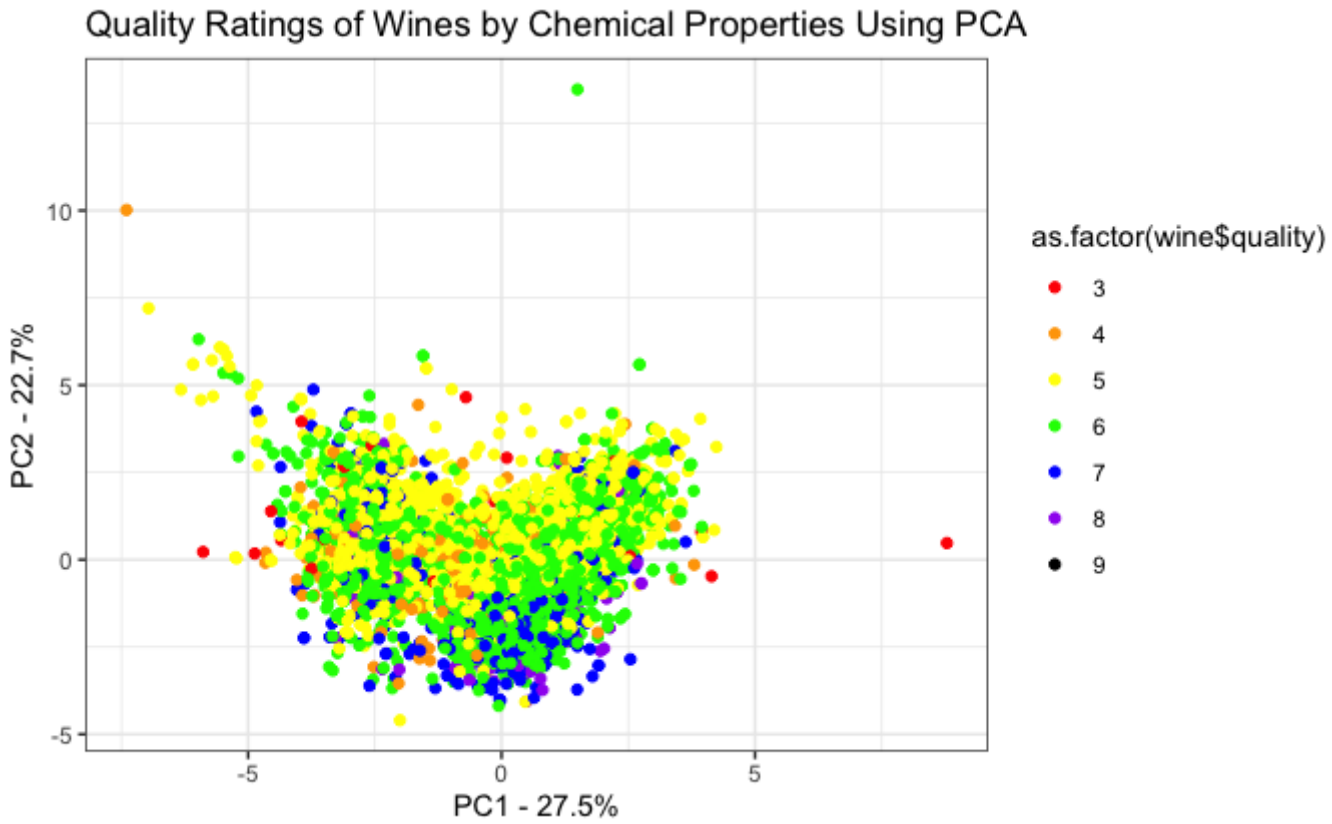
```
# scatter plot
ggplot(data=pca_data, aes(x=PC1, y=PC2, color=color)) +
  geom_point() + theme_bw() +
  xlab(paste("PC1 - ", pca_var_percent[1], "%", sep="")) +
  ylab(paste("PC2 - ", pca_var_percent[2], "%", sep="")) +
  ggtitle("Colored Wines by Chemical Properties Using PCA")
```



I have tried to do the same for quality, but again, this is hard because quality appears to be indistinguishable.

Hide

```
ggplot(data=pca_data, aes(x=PC1, y=PC2, color=as.factor(wine$quality))) +
  geom_point() +
  scale_color_manual(values = c("3" = "red", "4" = "orange", "5" = "yellow",
                                "6" = "green", "7" = "blue", "8" = "purple", "9" = "blac
k")) +
  theme_bw() +
  xlab(paste("PC1 - ", pca_var_percent[1], "%", sep="")) +
  ylab(paste("PC2 - ", pca_var_percent[2], "%", sep="")) +
  ggtitle("Quality Ratings of Wines by Chemical Properties Using PCA")
```



tSNE Model:

The tSNE model is created on python, and at the bottom of the code is the summary for all models.