Exercises 3

Isaac Hulsey (idh285) & Tristan Robins (tjr2695)

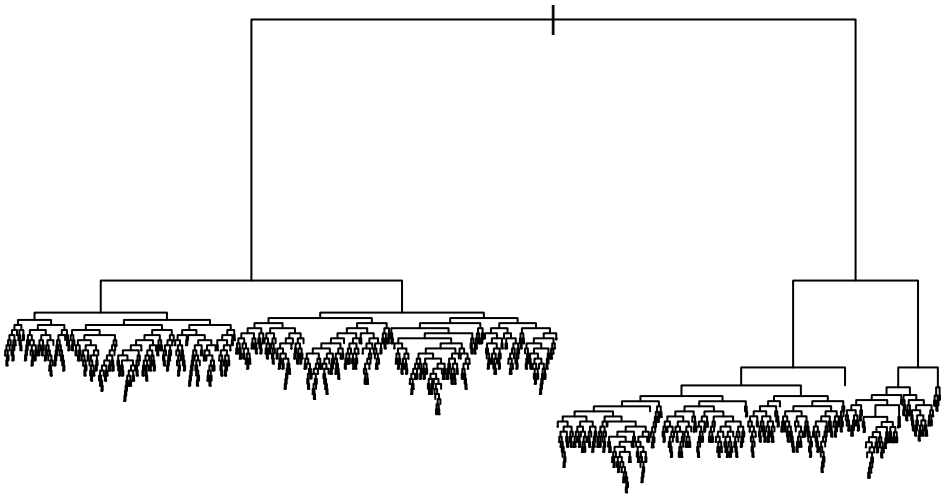
4/20/2020

Exercises 3

Question 1: Predictive Model Building

```
## [1] 2344
```

```
## [1] 1272
```



```
mean((yhat_test_lms - green_test$Rent)^2) %>% sqrt
```

```
## [1] 15.73678
```

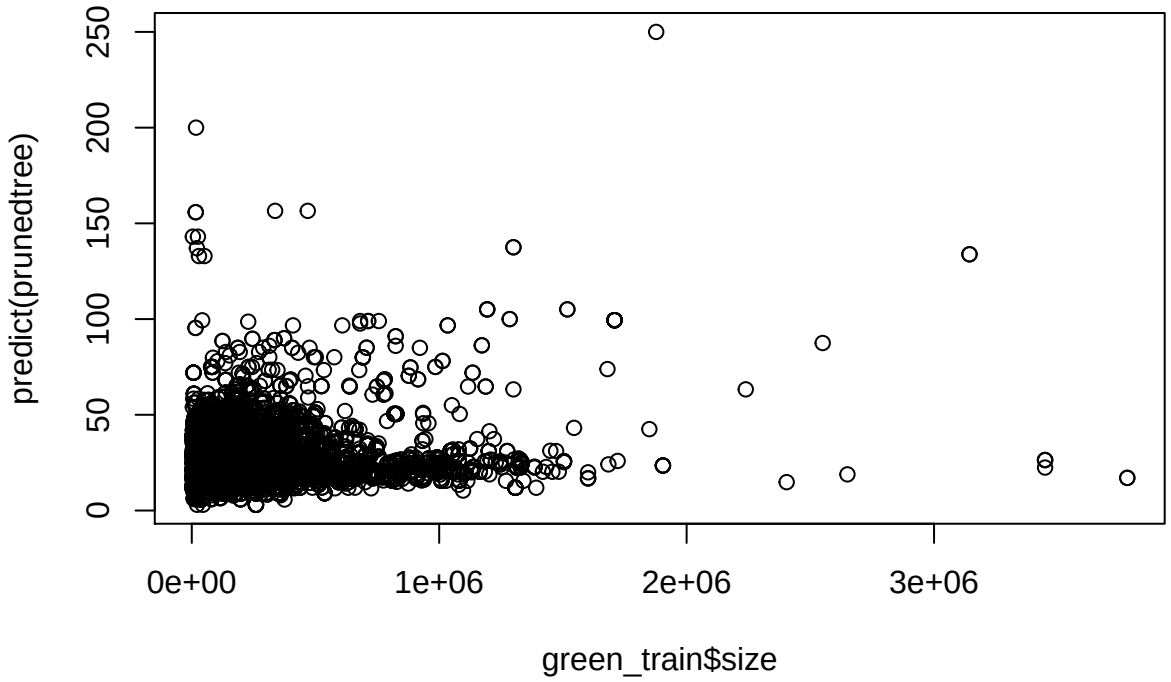
```
mean((yhat_test_tree - green_test$Rent)^2) %>% sqrt
```

```
## [1] 8.445067
```

In this problem, we are addressing the best outcome of predicting rental price charged to tenants of a buildings based off various factors defined in the dataset.

To accomplish this task, a tree model was utilized (rather than a simple linear model). The outcome of RMSE suggests the tree model was the better model compared to the linear model, given the reduction in RMSE. This tree model incorporated the best 1000 branches of the tree (which cut the total number of branches by over 50%). The linear model predicted rent price using key distinguishing qualities of a building rather than cluster qualities.

Clearly, the tree model was better able to address the fit of rental price charged to the tenant. But how can we find the average change in rental price per square foot, holding all else fixed?



To find the average change in rental price, the graph above presents a decent answer. We see that when square footage increases in these buildings by approximately 1 million square feet, we only see an uptick in rental price charged per tenant of between $0 and $50 on average, according to the best fit tree model.

Question 2: What Causes What?

Why you can't just regress "crime" on "police" to get a causal effect of police presence on crime:

If you simply regress crime on police you miss out on a larger story that might be going on. For instance, maybe a city has deployed more police simply because there is a crime wave, and as a result, you would see a positive "casual" effect if this scenario was seen time-and-time again in the data. It wouldn't be a causal effect of police presence on crime, but, rather, it would be a measurement of how crime effects police presence.

How did researchers from UPenn isolate a causal effect between police presence and crime, and why did they include metro data?

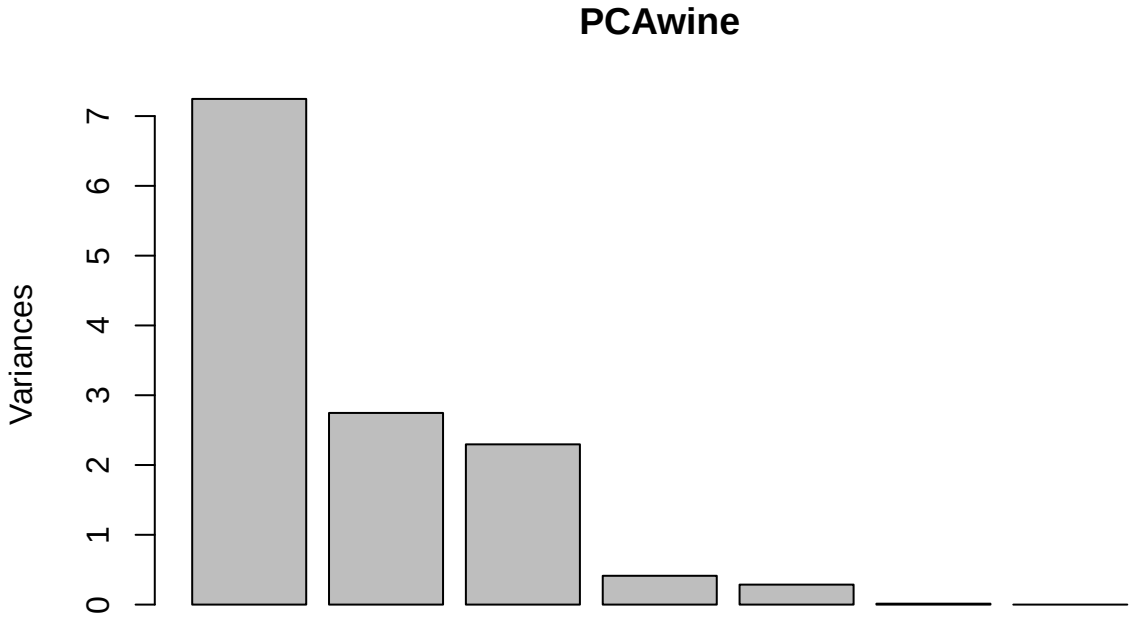
The researchers at UPenn noticed that on "high alert" days for terroism in Washintgon DC there is a spike in police officers present."High alert" days are random in the context that they have no correlation with any underlying crime waves. By including a "high alert" dummy

variable, the researchers were able to get closer to a causal effect than just regressing crime on police. The researchers noticed that there in fact was a decrease in crime on "high alert" days which indicates that police officers have a negative effect (they decrease) crime levels. However, the researchers at UPenn thought of an issue with the model. Perhaps on "high alert" days in DC, there's less travel going on and people are staying at home. To account for this, the researcher's factored in metro data, and with the inclusion of the data, they noticed that the causal effect of police officers was still negative but a little weaker than before. By combining these two features into their regression, the researchers at UPenn were able to establish a good case for the causal effect of police officers on crime levels.

Capturing geographical effects:

I do want to start off with a concern. The table doesn't make it very clear what district is being left out of the data in order to allow for proper interpretation of the causal effect. With that concern being clear and understanding that my interpretation is flawed due to the lack of clarification, it seems that district one has a much higher reduction in crime than the surrounding districts. In fact, the other distrcits don't have a statistically significant negative effect.

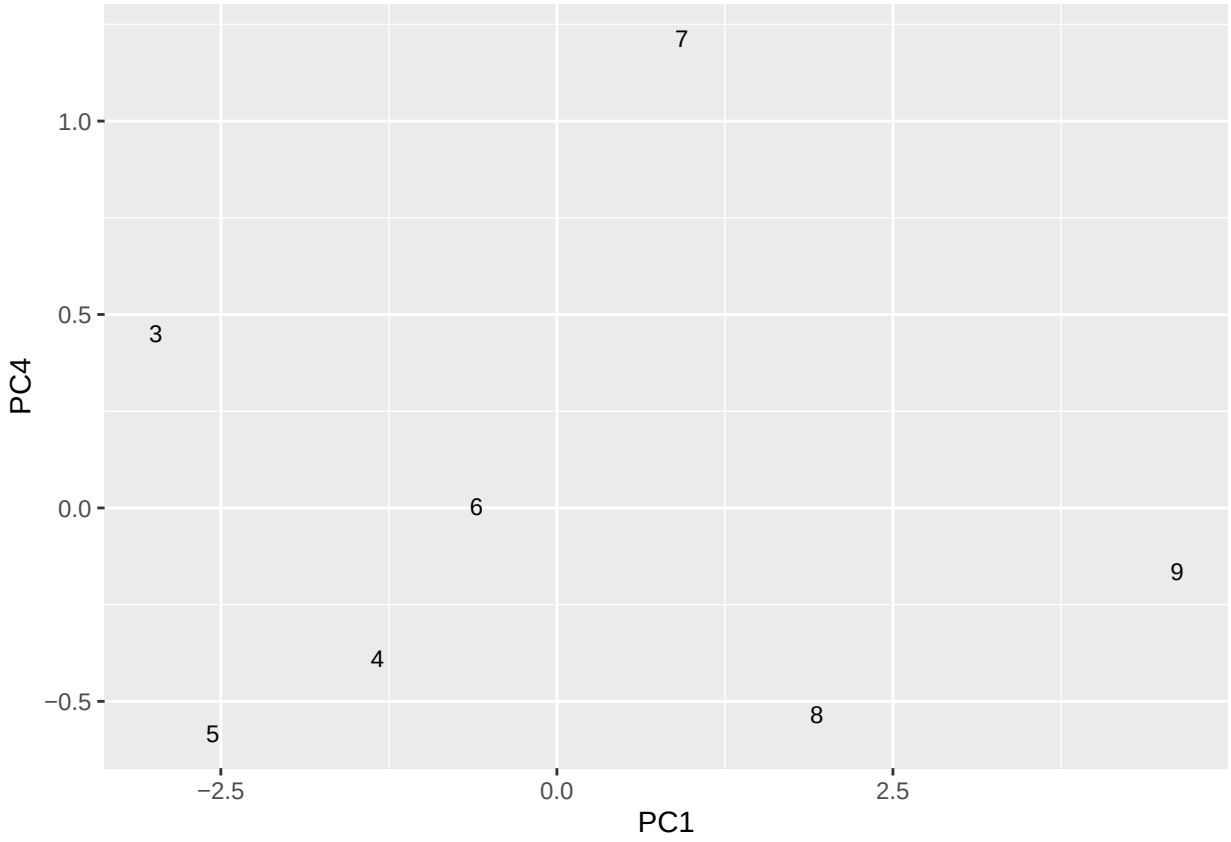
Question 3: PCA and Clustering

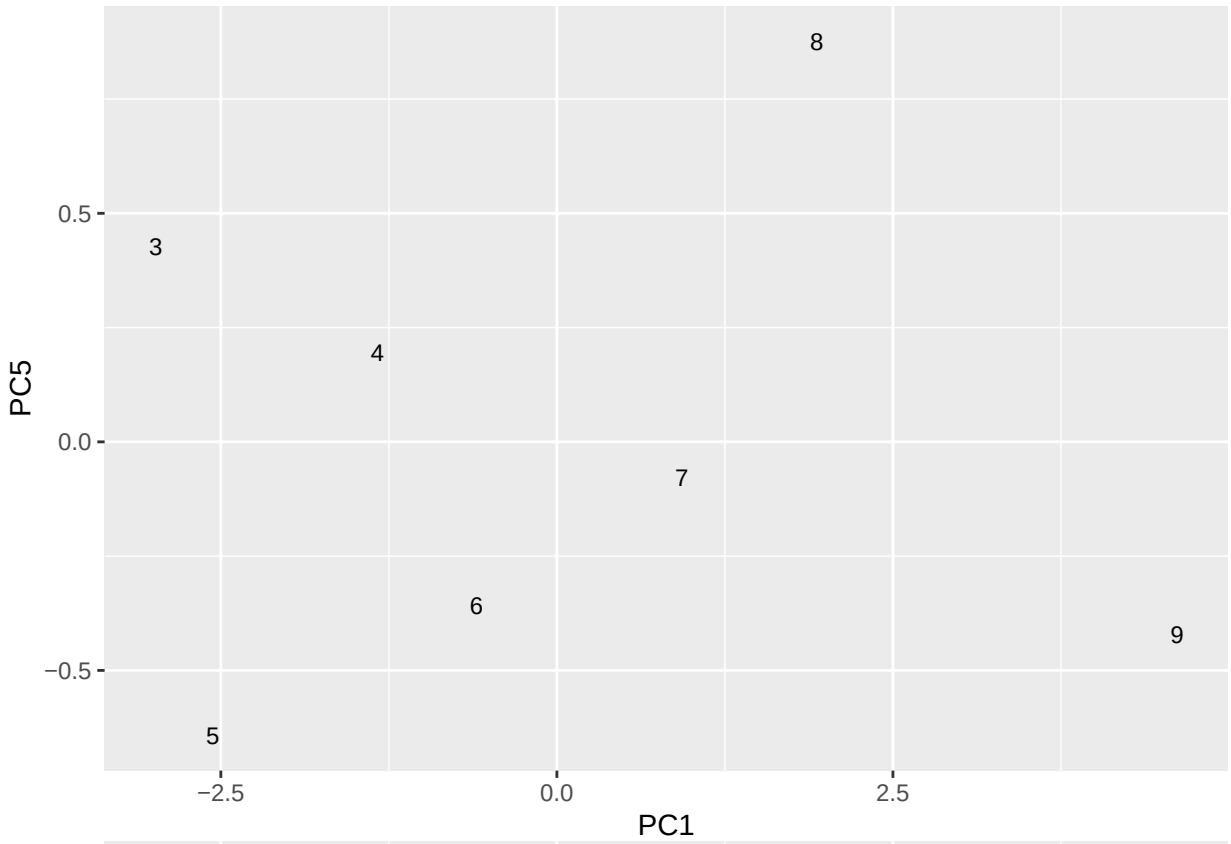


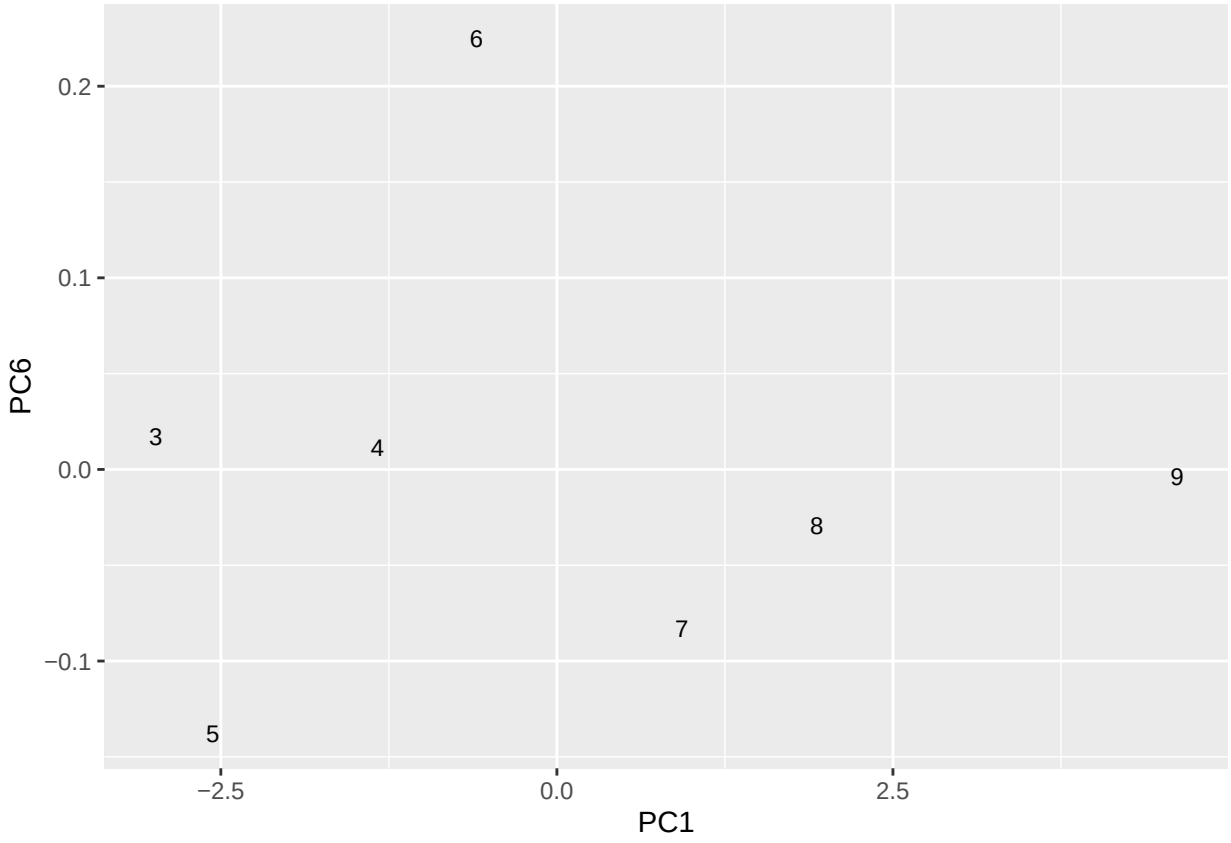

```
## Importance of components:
##                             PC1    PC2    PC3     PC4     PC5    PC6       PC7
## Standard deviation     2.6918 1.6572 1.5153 0.64239 0.53495 0.1138 1.018e-14
## Proportion of Variance 0.5574 0.2112 0.1766 0.03174 0.02201 0.0010 0.000e+00
## Cumulative Proportion  0.5574 0.7686 0.9453 0.97699 0.99900 1.0000 1.000e+00
```

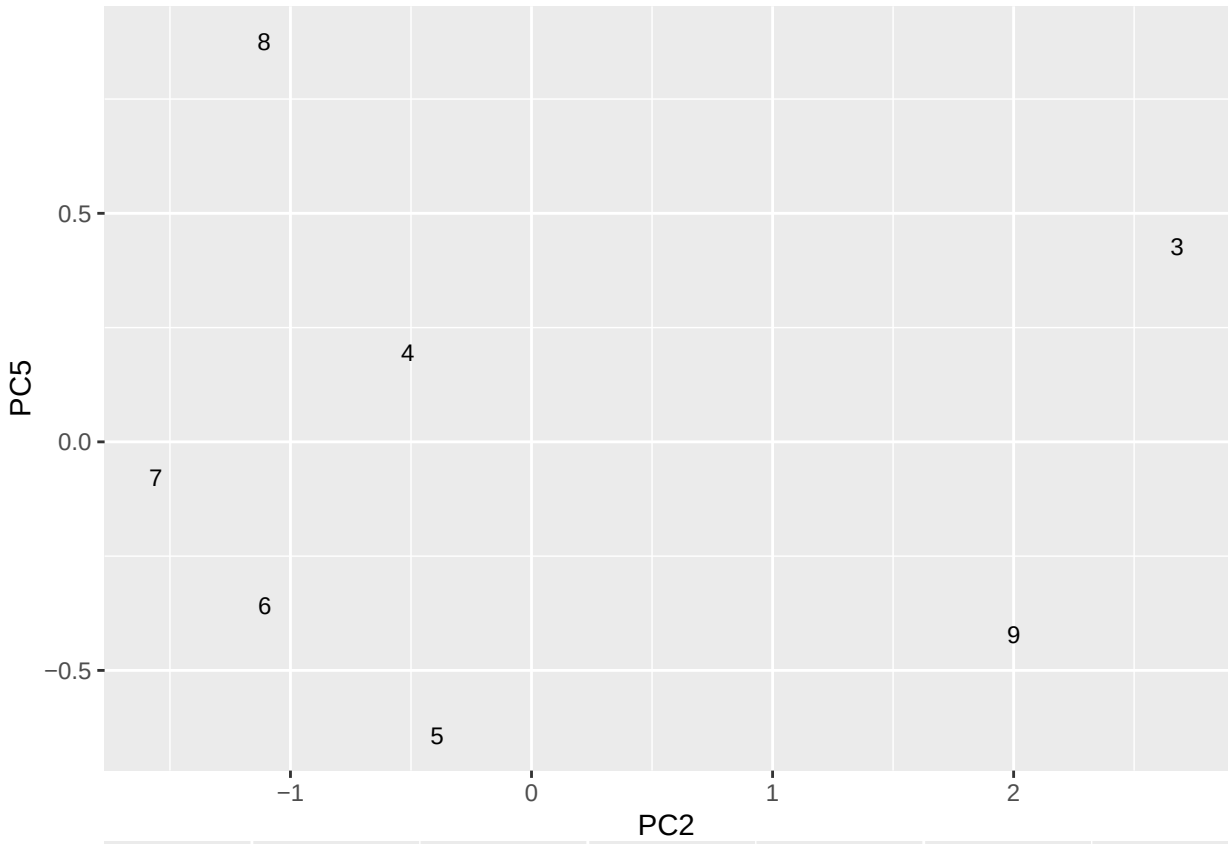
```
##                         PC1   PC2   PC3   PC4   PC5   PC6   PC7
## fixed.acidity         -0.16  0.51 -0.08  0.33 -0.34  0.23  0.07
## volatile.acidity      -0.28  0.31 -0.22 -0.11  0.35 -0.34  0.03
## citric.acid            0.33  0.03  0.20  0.09 -0.57 -0.19  0.19
## residual.sugar        -0.18 -0.17  0.53 -0.28 -0.08  0.38 -0.32
## chlorides             -0.36  0.11  0.00  0.06  0.17  0.18 -0.16
## free.sulfur.dioxide    0.03  0.31  0.52  0.37  0.42  0.23  0.16
## total.sulfur.dioxide -0.06  0.29  0.55 -0.31 -0.07 -0.51  0.01
```

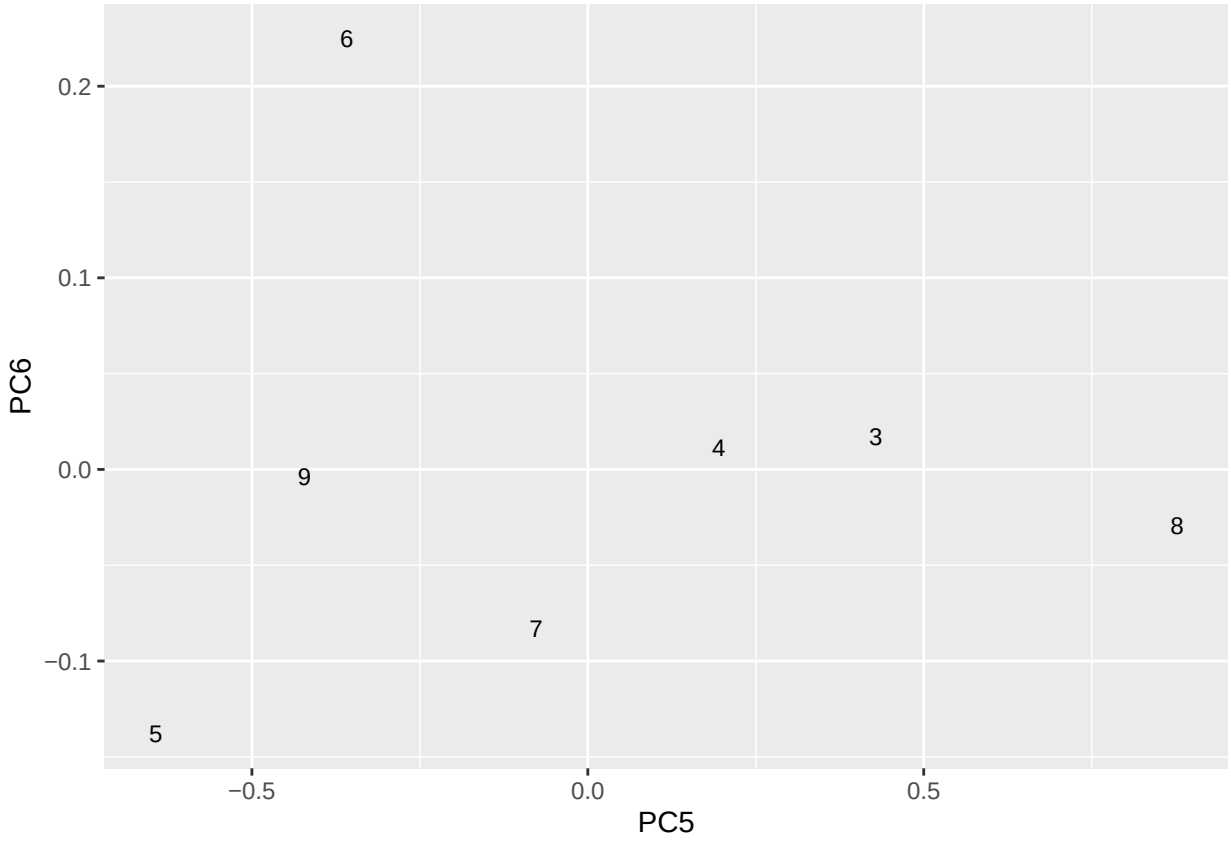
```
## density              -0.37  0.03  0.00 -0.11 -0.22  0.28  0.68
## pH                    0.22  0.47 -0.12  0.13 -0.19  0.10 -0.47
## sulphates            -0.18 -0.44  0.16  0.65 -0.07 -0.14 -0.14
## alcohol               0.36  0.02  0.08  0.25  0.31 -0.17  0.31
## red                  -0.37  0.00  0.01  0.15 -0.13 -0.29 -0.06
## white                 0.37  0.00 -0.01 -0.15  0.13  0.29  0.09
```

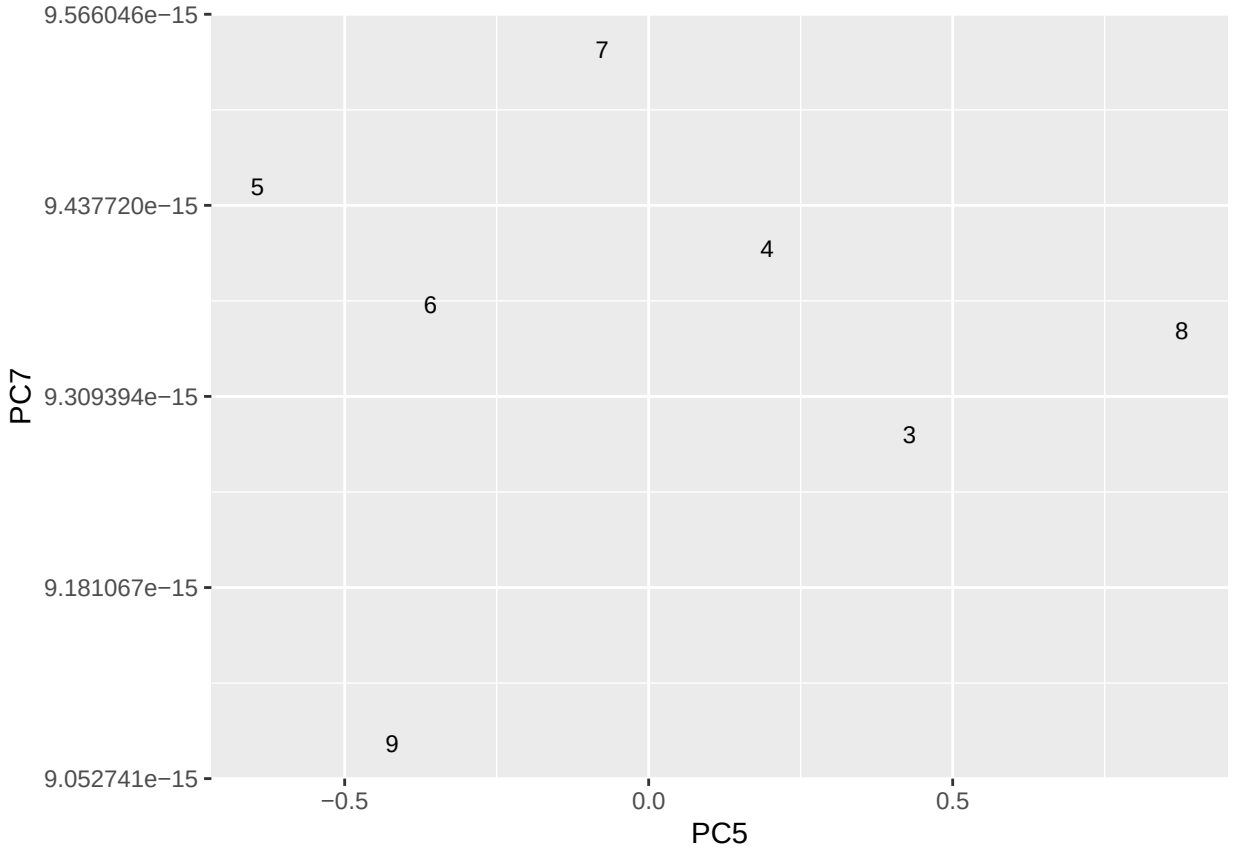



8
0.5
3
4
PC5
0.0
7
6
9
−0.5
5
−2.5
0.0
2.5
PC1


6
0.2
0.1
3
4
PC6
0.0
9
8
7
−0.1
5
−2.5
0.0
2.5
PC1


8
3
4
7
6
9
5
0.5
0.0
−0.5
PC5
−1
0
1
2
PC2


6
4
3
9
8
7
5
0.2
0.1
0.0
−0.1
PC6
−0.5
0.0
0.5
PC5



In this problem, we are trying to address how to distinguish between different qualities of wines and whether it is possible to do with 11 chemcial properties.

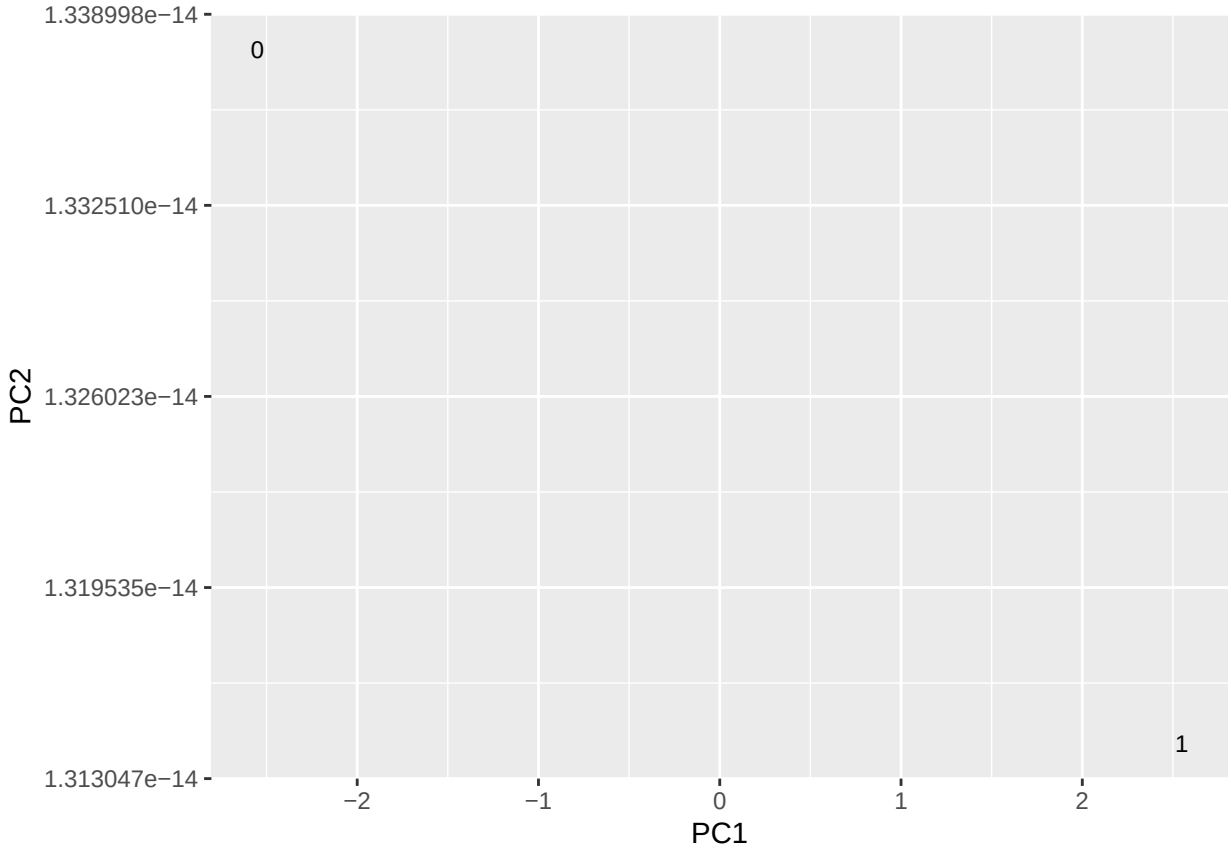
To do this, PCA was conducted by grouping on quality and examining the relationships between each of the PC's presented in the data.

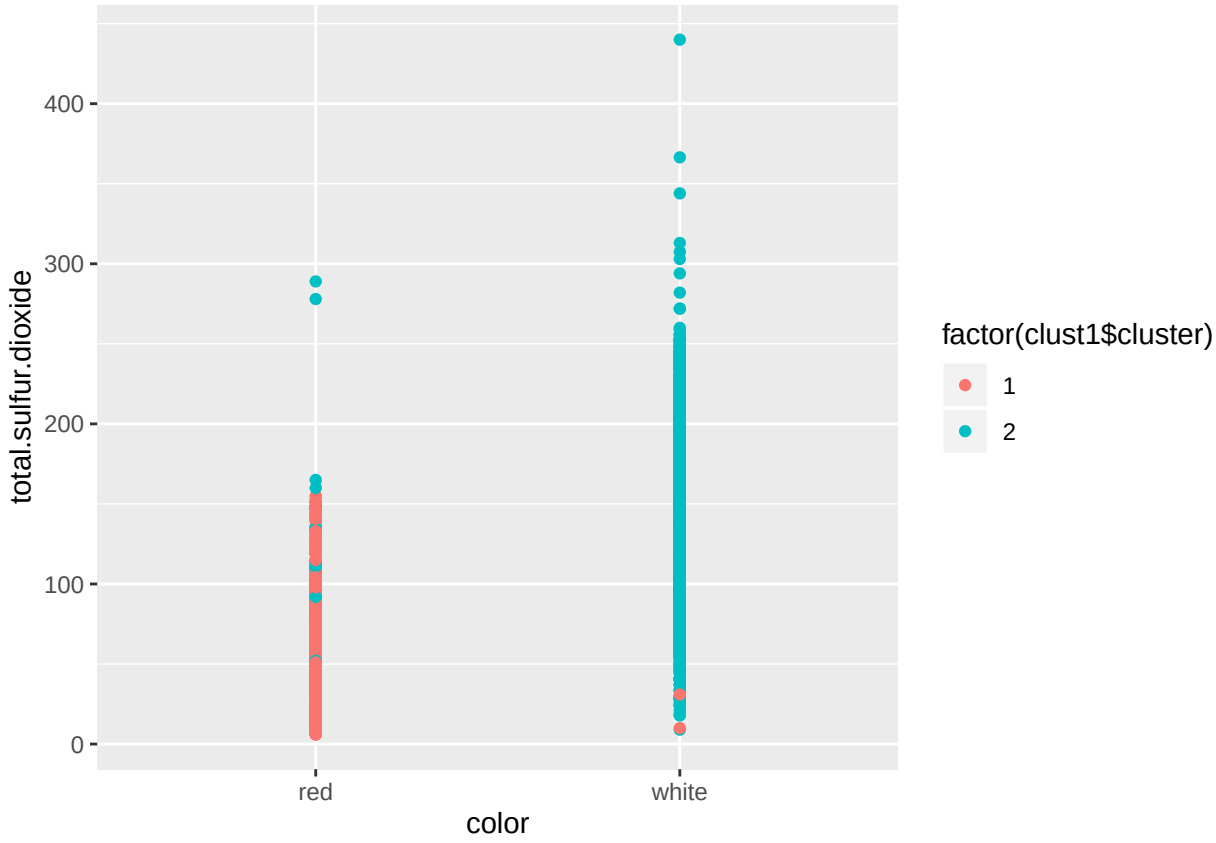
Only the best PC graphs were included but, when combined, suggest an ability to strongly distinguish between the quality of wine presented in the data. Further, clusters could be ran on this, but the PCA provides so much distinction that clutering is not especially necessary.

Each relationship between PC's shows that there each wine quality ranking is distinct. There is very little overlap between qualities which tells us that the PCA does a great job at distinguishing between the different qualities of wine.


```
## Importance of components:
##                             PC1       PC2
## Standard deviation     3.606 1.873e-14
## Proportion of Variance 1.000 0.000e+00
## Cumulative Proportion  1.000 1.000e+00
```

```
##                         PC1   PC2
## fixed.acidity          0.28 -0.12
## volatile.acidity       0.28 -0.10
## citric.acid           -0.28  0.05
## residual.sugar        -0.28  0.09
## chlorides              0.28 -0.09
## free.sulfur.dioxide   -0.28  0.10
## total.sulfur.dioxide -0.28  0.09
## density                0.28 -0.09
## pH                     0.28 -0.09
## sulphates              0.28 -0.12
## alcohol               -0.28 -0.94
## quality               -0.28 -0.10
## white                 -0.28  0.09
```





Further within this problem, we were tasked to sort on color if possible.

To do this, both a PCA graph of the distinctions between red and white wine, and a clustering algorithm of 50 starts were conducted to group the wines.

The PC graph suggests that we can group based on PC's wherein 1 signifies red wine, while 0 signifies white wine. The PC graph suggests that it is possible to sort; however, it is too perfect to take at face value. As such, the clustering algorithm, although it undoubtedly contains errors due to some of the traits of white and red wines overlapping at times, shows a high degree of ability to cluster on color, given the 11 chemical properites of the wine that were assessed in the dataset.

The solution could be changed to incorporate 4 clusters, but anything more than 4 clusters starts to provide too little distinction between colors; as such, 2 clusters seemed to provide the most intuitive approach.

Question 4: Market Segmentation

This data set has a good bit of categories. To reduce the feature size and to get an idea of where the market is concentrated, I am going to run a principal component analysis and see what the correlation with the data is with five pricncipal components.

```
## Importance of components:
##                            PC1    PC2    PC3    PC4     PC5
## Standard deviation     1.2794 1.0103 0.9880 0.9505 0.68024
## Proportion of Variance 0.3274 0.2041 0.1952 0.1807 0.09254
## Cumulative Proportion  0.3274 0.5315 0.7268 0.9075 1.00000
```

```
##                       PC1        PC2         PC3        PC4         PC5
## chatter        0.65230740 -0.1790585 -0.03576448 -0.2083243 -0.70551755
## current_events 0.32472042  0.2499770 -0.40617255  0.8165953  0.01625326
## travel         0.07477987  0.8639270 -0.26119252 -0.4238859 -0.01171740
## photo_sharing  0.65474125 -0.1495622  0.01776467 -0.2198171  0.70732589
## uncategorized  0.18647671  0.3697443  0.87476148  0.2485445 -0.03916125
```

From the first principal component, it looks like NutrientH20's most important categories are chatter, current events, travel, photo sharing, and uncategorized. From the correlation table, chatter and photo sharing are incredibly close to each other when it comes to correlation with the first principal component. The first principal component accounts for about a third of the variance. From this information, it appeares the bulk of NutrientH20's followers are split amongst people who like taking photos and just chatting.

The second principal component which accounts for an additional fifth of the variance in NutrientH20's followers shows a high correlation with travel. This shares a negative correlation with chatter and photo sharing; I feel like travel detracts from photo sharing and chatter because people who normally just would chatter or share photos might be talking a lot about a vacation they just had or sharing pictures of that vacation. Maybe these people are the more affluent of NutrientH20's followers.

The third principal component accounts for roughly a fifth of the variance as well, and it is highly correlated with uncategorized. This is negatively correlated with everything except for photo sharing (which is near zero) and uncategorized. A theory I have is these are "meme" accounts because a "meme" category isn't anywhere in the data, and most "meme pages" don't typically chat with their followers or talk about current events or mention their travel plans, but they do post funny photos which might explain the slight correlation with photo sharing.

The fourth principal component accounts for a little less than a fifth of the variance, and it is most strongly correlated with current events. my hypothesis is that this is picking up on something to do with more political users, and maybe even a certain political party. This component is negatively correlated with travel, photo sharing, and chatter which makes me think that these followers don't post much about their personal lives and are more of a sort of activist type of poster.

The fifth principal component acounts for the roughly remaining tenth of the variance in the data. It is roughly correlated to the same extent but in different magnitudes to both photo sharing and chatter. Photo sharing is the positive correlation and chatter being the negative. This would be a sort of instagramer who would be posting on twitter.

Taking all of this into consideration here would be my recommendation of what NutreientH20's market looks like in five categoires. People who don't travel a lot but are somewhat politcally minded would be the first category. There may be a certain political party associated with this group, but it cannot be told with what is in the data. The second category being made up of more affluent or oversharing people about their vacations who also care about current events. The third category which probably is made up of younger people would be potential meme pages. The fourth group being people very concerned with current events and not posting much personally. This might be indicative of another certain political party. The last category would be instagrammers who also post on twitter for extra exposure.