Exercise 3

Alt Nayani and Conor McKinley

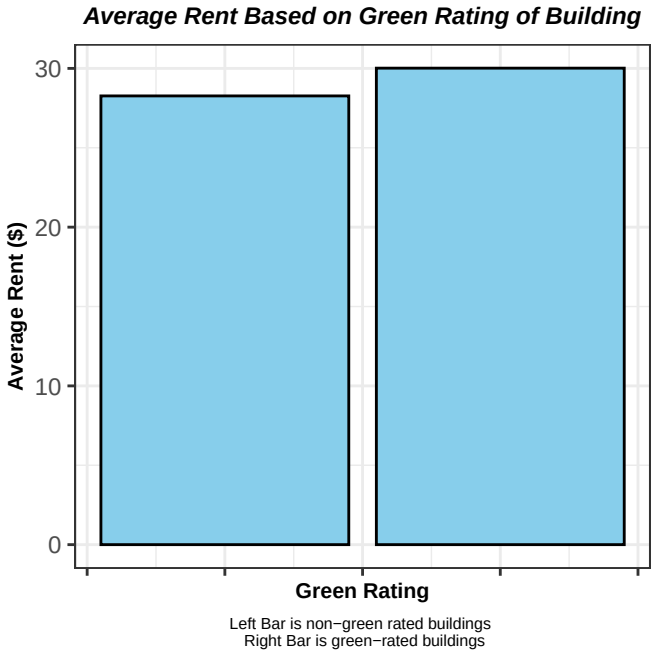
4/23/2020

(1) Predictive Model Building - Green Rating

Question: What is the effect of a green rating (i.e LEED or EnergyStar Certified) on the rent that can be charged of the occupants?

This question has many implications such as if a building owner should incur additional costs to recieve a green rating in hopes for an increased rent rate.

Let us first begin by looking at the raw green v. non-green rated buildings.



While the initial bar graph shows an increased rent price for green rated buildings, there are many potential cofounding variables such as green rated buildings are usually newer and located in better areas. Therefore, we must isolate the effect of the green rated buildings.

To do this we have tried two different methods: linear regression using forward selection and Principle Components Analysis (PCA).

First is the linear regression. Below can be found the fitted model:

```
## lm(formula = Rent ~ class_a + cluster + renovated + size + age +
##     class_b + empl_gr + green_rating + cluster:size + class_a:age +
##     class_a:renovated + cluster:age + class_a:size + size:age +
##     renovated:age + age:class_b + cluster:renovated + age:empl_gr +
##     renovated:green_rating + size:green_rating + renovated:size +
##     renovated:class_b, data = greenbuilding_train)
```

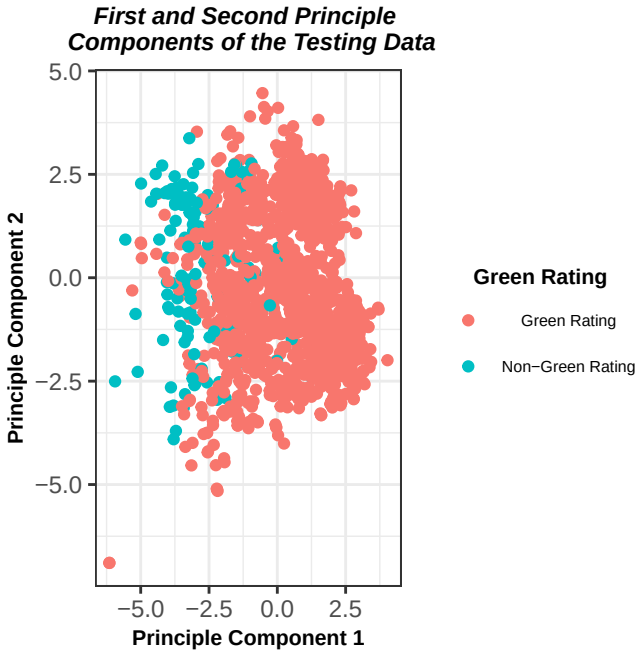
Next, we tested the fitted model on multiple testing datasets and compared the error (Root Mean Sqaured Error) against a null model.

Table 1: Average RMSE of Each Model

	x
Handmade	13.99
Null	15.04

Next, Principle Components Analysis was used. As seen in the forward selection model, we split the data into a training and testing set. Below is the training set graphed by the first and second principle components and colored based on the green rating of the building.

The components then were projected onto the testing data. Below is the testing data plotted using the components from the training data.



Conclusion:

Both Forward Selection Linear Regression and PCA was used to determine the impact of a green rating on the rent price for a building. Using the forward selection linear regression we saw the model performed quite well out of sample with a RMSE = 1400 compared to the null of 1500. The model portrayed the isolated impact of the green building certificate as $5 per square foot. Meaning there are obvious economic benefits for attempting to attain a green building certificate, holding all other building features constant.

Next, to confirm the benefit of a green rated building, principle components analysis was used. The training and testing data looked very similar in reference to the relative location of green buildings (both on the left-side of the first component). The first principle component accounts for 15% of the variance and by adding the second principle component the cumulative variance accounted for increases to 30%.

(2) Effect of Police on Crime

1. There are a couple different reasons why the relationship between crime and police cannot be easily interpretable from a simple regression.

    a) There may be confounding variables that affect the results such as the population density of the city, amount of guns per capita, etc. All of the confounding variables would make the underlying affect of police on crime difficult to assess.
    b) There is an inherent problem with the regression. In a city with high crime, there are more cops than cities without crime. Therefore, the marginal affect of determining the effect of adding a cop is again is difficult to determine.

2. Due to problem described above, the underlying data had certain problems. Therefore, the researchers had to determine a method to see the effect of police holding the crime rate stable. To be able to achieve this they focused on days of 'high alert' (orange level for terrorist activity), which created more cop presence without the cop presence being determined by the level of crime. What the researchers found was that police decreases the level of crime on average by using the high alert data. They took the study a level further by controlling for METRO ridership, and again found conclusive evidence that police presence does in fact lower crime rates.

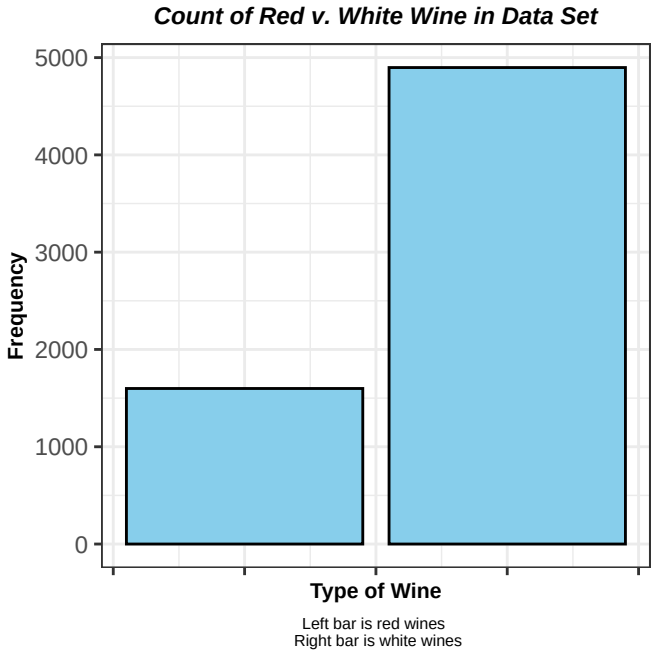
3. The researchers controlled for METRO ridership because there was a fear of another cofounding variable skewing the results of the study. The researchers were worried that due to the elevated level of the alert, less tourist and crowds were likely to be found, and thus a lower probability of a crime occurring. The researchers found that the tourists levels were not affected by the level of alert and the result was still statistically significant after accounting for the METRO ridership.

4. The model has three main variables: alert level (encoded as a dummy variable with 0 being low and 1 being high), location (encoded as a categorical variable between a constant, District 1 and Other Districts) and Log(midday ridership). The model can be written as: crime = constant + HighAlert*District1* + *HighAlert*Other Districts + Log(midday ridership) It can be seen that the high alert is a significant indicator of crime in District 1 but not in other districts. This makes sense because other districts are not Washington D.C. with alert levels and reported data. Since High Alert in District 1 was a negative value, it shows how an added cop does decrease crime by 2.621 crimes per day. Furthermore, midday ridership is also significant with a positive value. This shows as midday ridership increases by 1% point (since it is on a log scale), crime increases by 2.477 per day at a 95% confidence level.

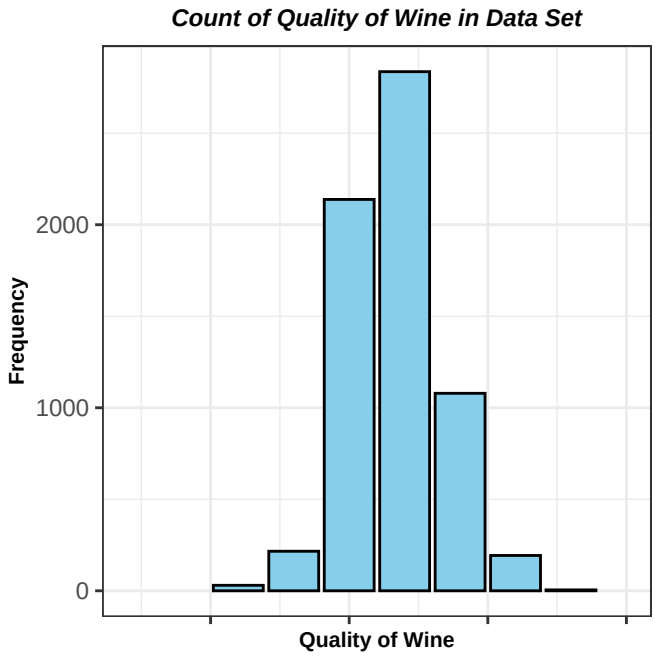
(3) Clustering and PCA

Question: What is the best method of unsupervised learning to determine the color and quality of wine? To first determine how to best to distinguish wines by different chemical properties, let's take an initial look at the raw data.

First, let us look at the split between red and white wines.

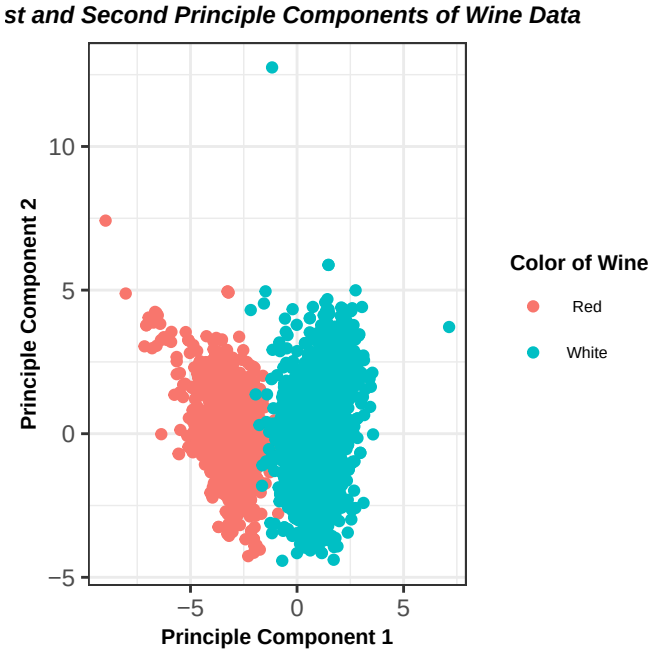


We can see the split between red and white wines is overwhelmingly white wines (~75%). Next, we will look at the frequency of quality ratings by certified wine experts.



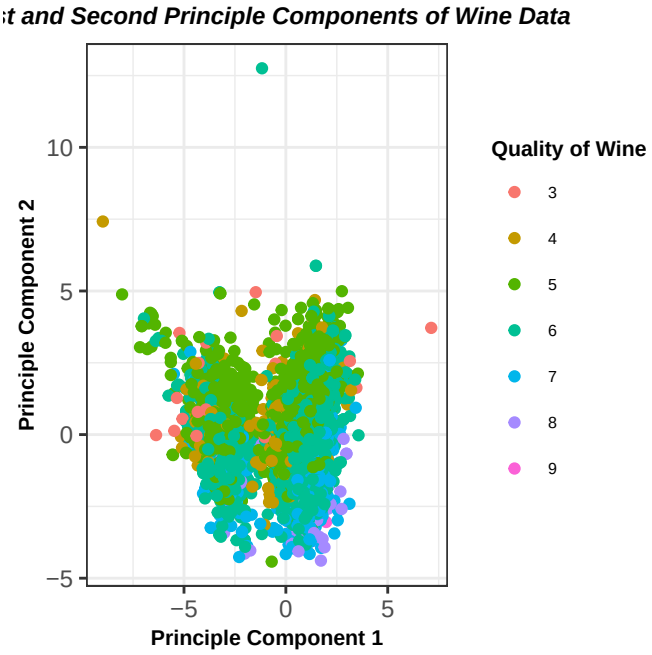
The majority of experts rated the wines as a '5' or '6'. It is important to note that the lowest rated wine was a '3' with the highest rated wine being a '8'.

To be able to distinguish the red wines from the white wines, we will try two different dimension reduction techniques. We will begin with Principle Components Analysis.

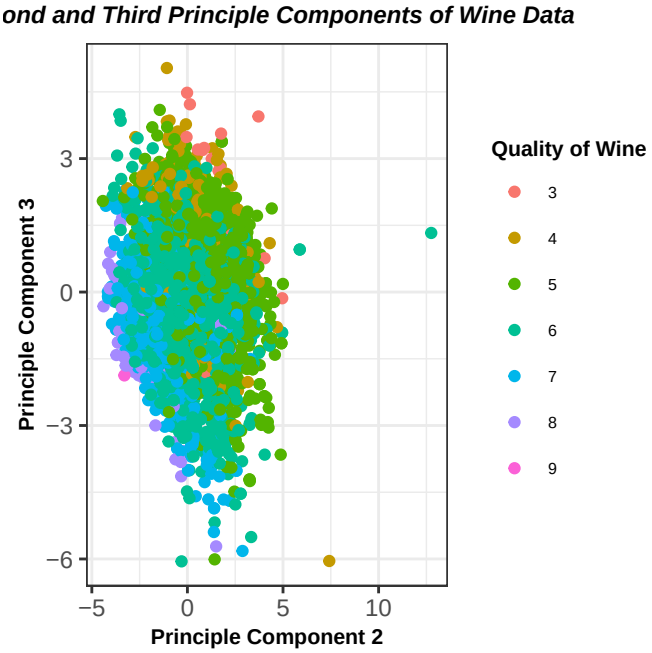


After running the PCA, the first two components explain approximately 50% of the dataset, while the first three components explain approximately 63% of the dataset. Furthermore, the graph provides critical insight on PCA's ability to distinguish between the colors of the wines by clearly segmenting the into two different clusters.

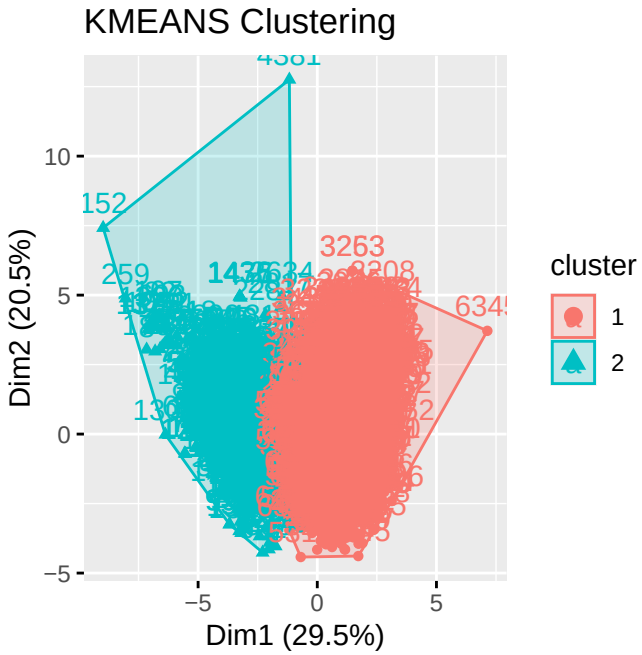
To take this analysis a step further, we will now use PCA to distinguish the quality rating given by the expert wine tasters.

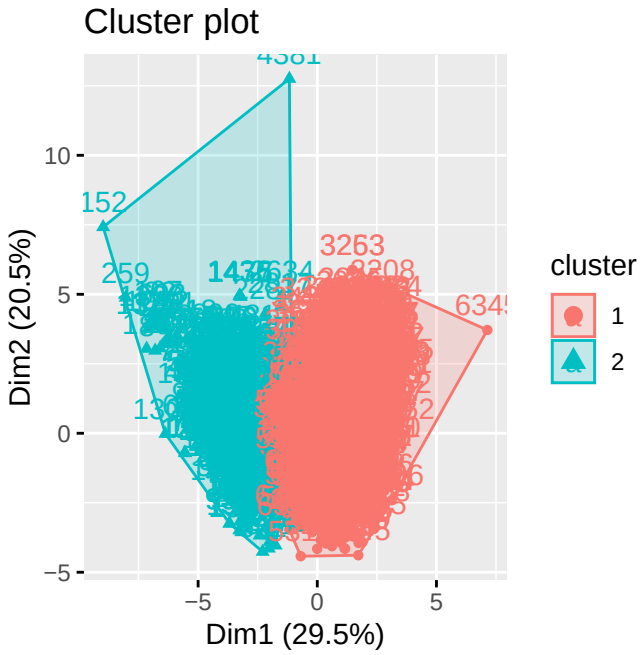


Unfortunately after plotting the first two components, it is hard to distinguish once cluster from another. Therefore, we attempted to plot the second and third component as this quality of the wines were a less significant feature of the dataset compared to the actual type of wine.



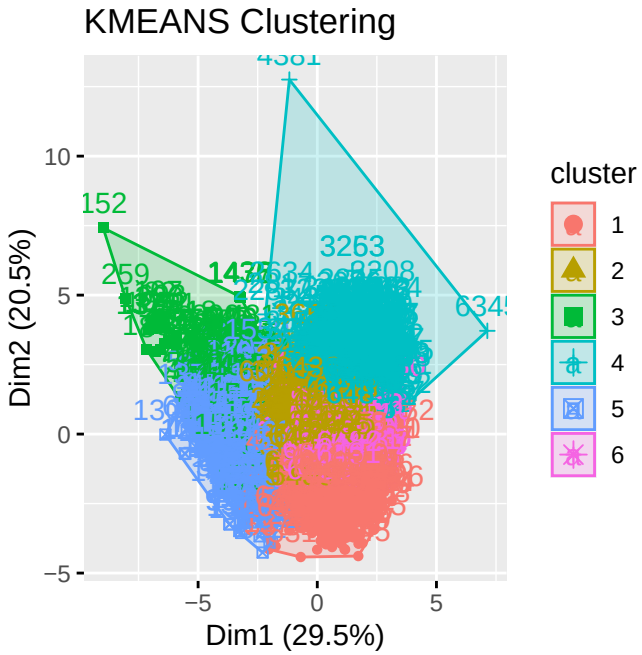
Next, we attempted the same analysis using a K-means clustering algorithm. We again tried the clustering on the color of the wine witha k = 2 (since we are expecting two clusters).

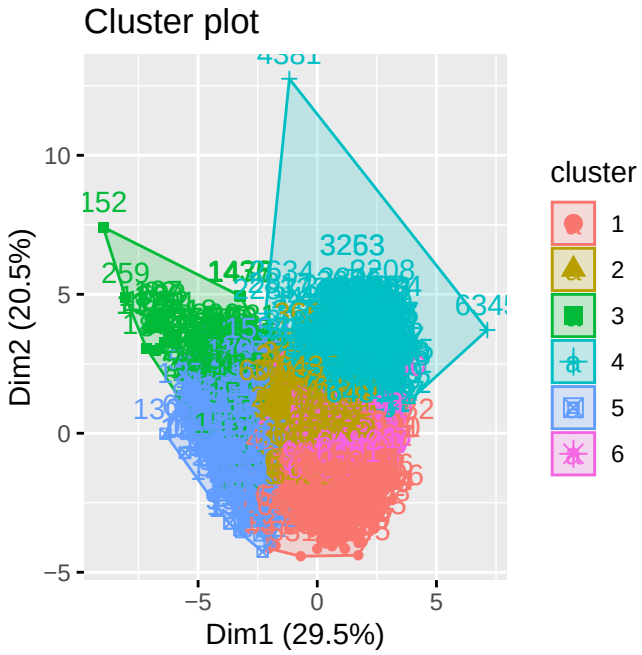




We can see the K-Means Clustering Algorithm clustered the two wines correctly for the most part, just as seen in the Principle Components Analysis.

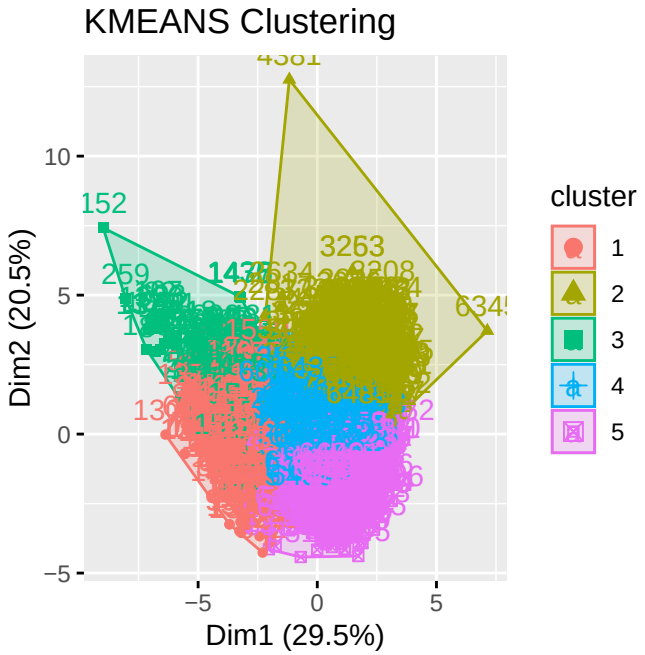
Again, we tried the same analysis on the quality of wine. However, rather than using k = 10, as the number of clusters we used k = 6. This was due to the ratings of the wine experts only included ratings from 3 to 8 (6 unique values).

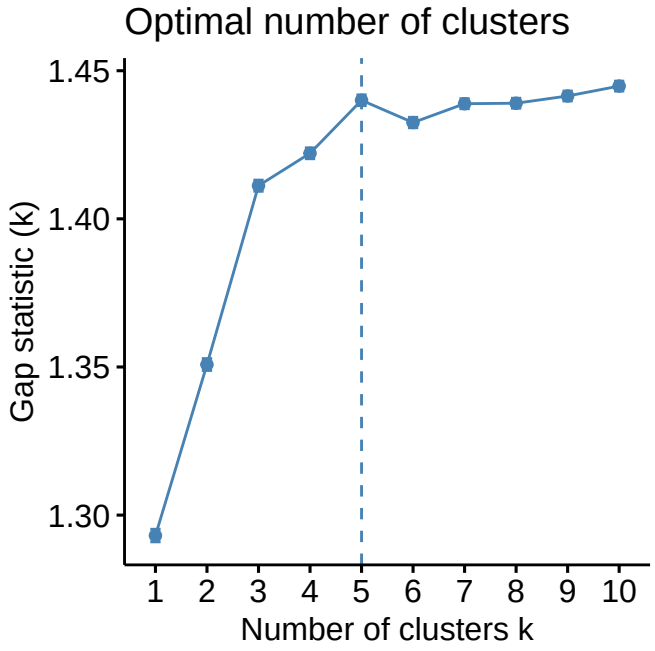




The clustering for the quality of the wine was more difficult to ascertain. The clusters that are seen are more forced rather than natural breaks in the dataset. Therefore, we can conclude that the k-means clustering algorithm does a good job at clustering the different classes of wine but not necessarily the quality.

Finally, the last component of the analysis is to find an optimal number of clusters. To determine the optimal number of clusters, a gap statistic was used. Based off the analysis done using the gap statistic, the optimal number of clusters in the data set is k = 5.



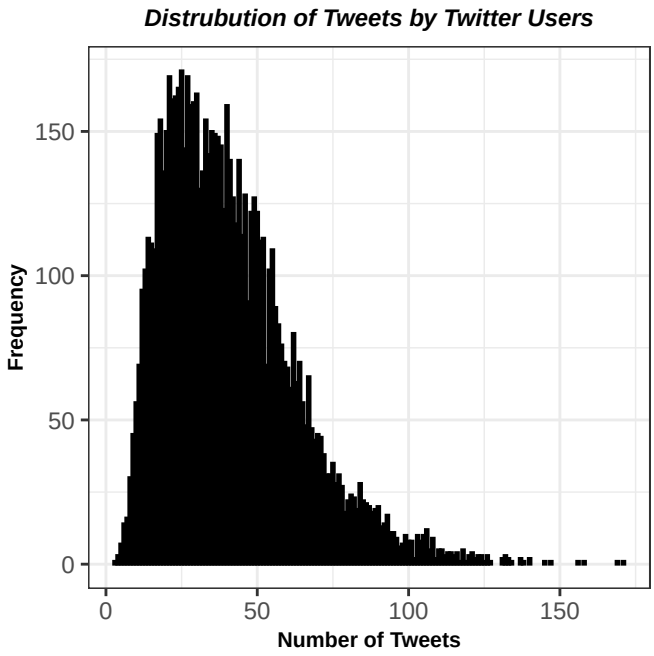


In conclusion, both methods PCA and k-means clustering did a good job at distinguishing the red wine from the white wine based on the chemical properties in the dataset. Both clearly placed the majority of the wines in the correct category. However, the k-means clustering did not clearly separate the data into six categories of the quality of wine. On the other hand, the PCA analysis was able to distinguish the quality of wine using the second and third components. It is clear on the PC2 v. PC3 graph that those with a lower value of PC2 and PC3 had a higher quality while those with a higher PC2 and PC3 had lower qualities.

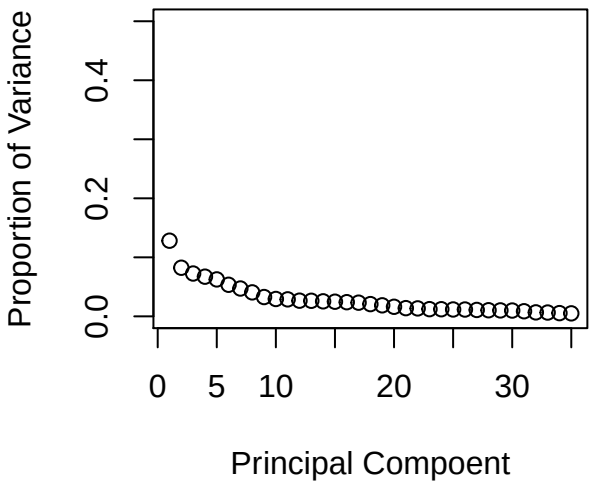
(4) Market Segmentation

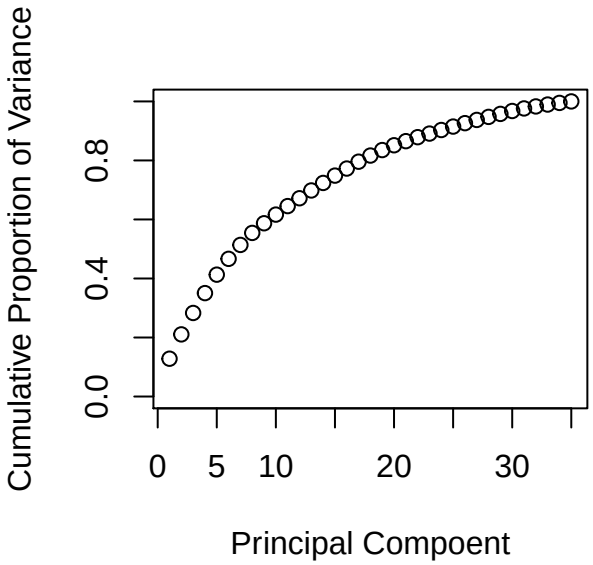
Our question is to determine different market segments that may be found through the data collected in the market-research study. The market-research study collected user tweet data that was grouped into categories based on the 'topic' of the tweet.

Before diving into methods of segmenting the data to find different market segments, lets take a look at the distribution of users tweet counts.



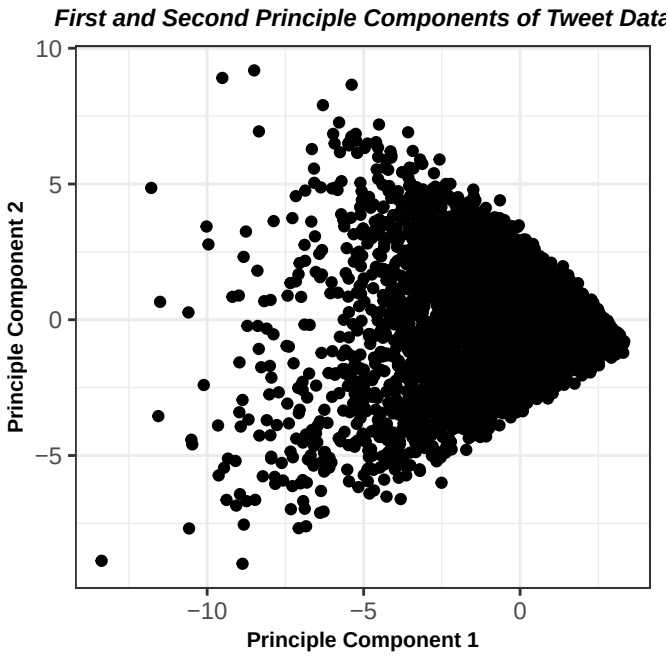
The distribution seems to be right skewed with a few large outliers on the right tail. Now to focus on segmenting the data, we will begin by using Principal Component's Analysis.





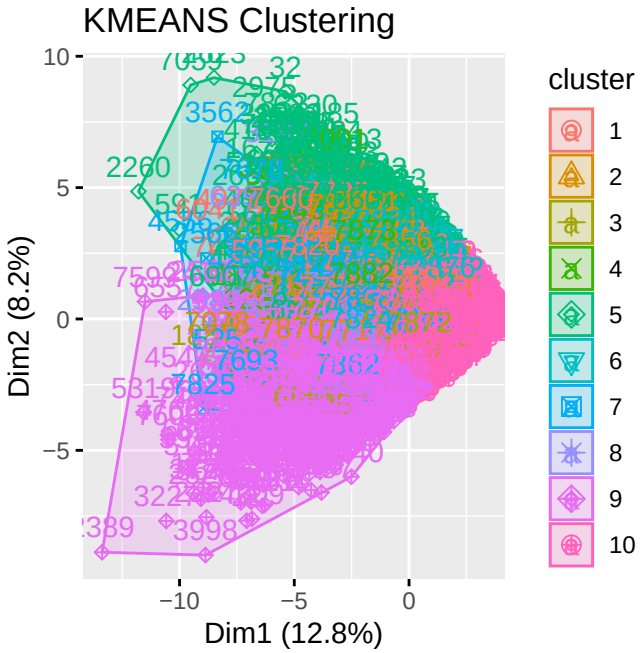
It can be seen that the first two components represent appoximately a fifth of the total variance of the data. We will be focusing on the first two components for our analysis to find the major underlying segments in the data.

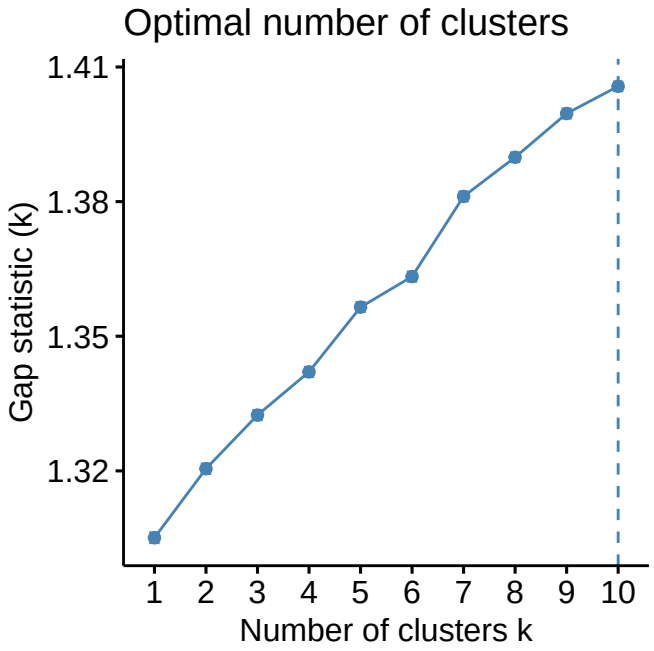
Let us begin by plotting the first and second components.

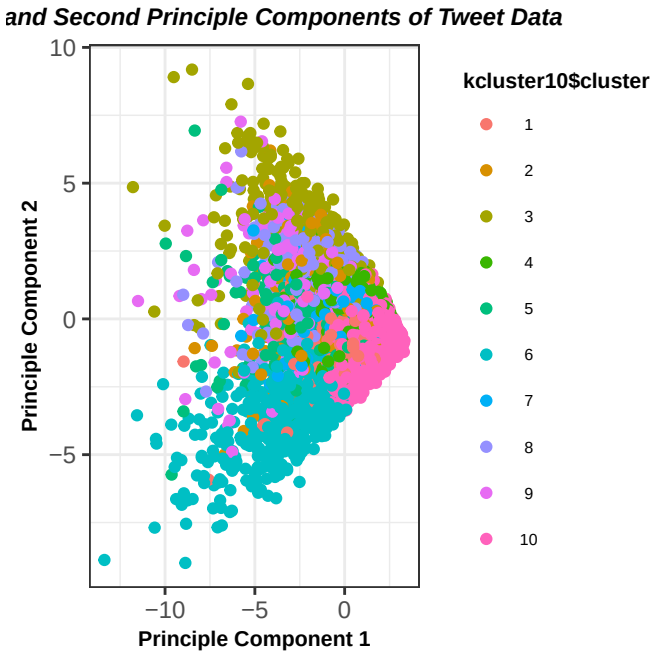


We can see a general funneling of pattern that can represent some strong correlations in users and their tweets. However, without being able to segment the data based on such clusters, the PCA does not provide us much quantifiable market segments.

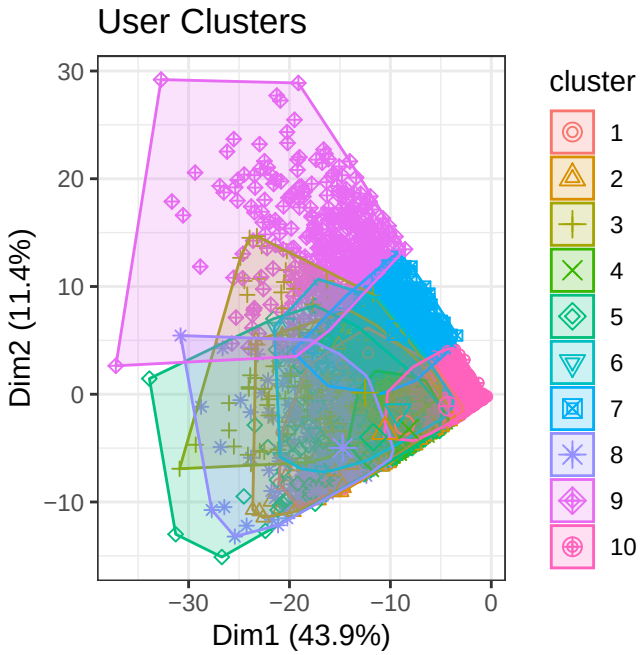
Let us replot the same graph, but color the points using a k-means clustering algorithm with a k value of 10. A k value of 10 was chosen due to the gap statistic graph below provided insight on its potential as the optimal number of clusters in the data set.







The coloring of the clustering allows us to see more clearly the different user groups. A cleaner representation of these user groups could be found as follows:



Now knowing the different user groups that are similar based off tweet data, we can now define each of these user clusters as a different market segment. Knowing each market segment has many potential upsides in a business perspective. An example of such is in marketing where you want to market similar items across a user segment that has a higher propensity to purchase such an item.