SDS 323 Exercises 3

Rylan Keniston

4/20/2020

Predictive Model Building

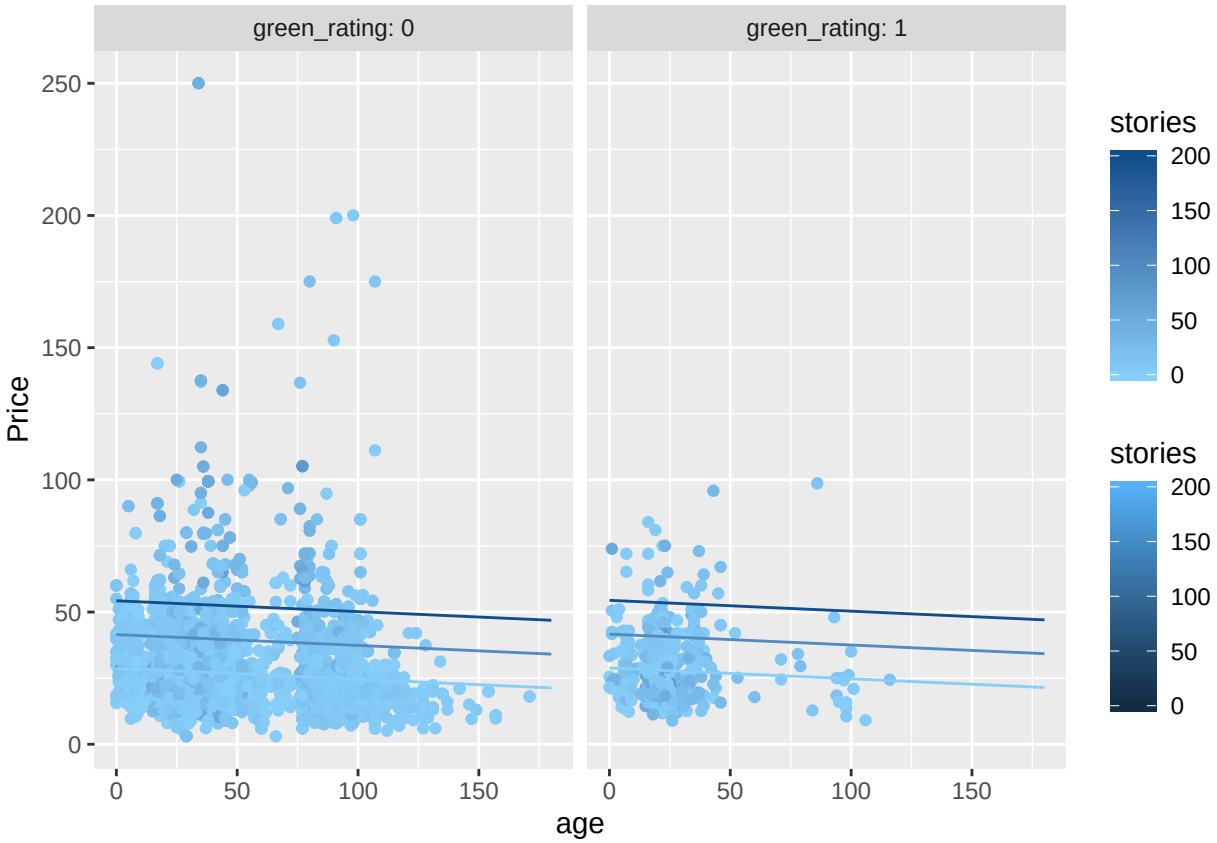
The data being analyzed regarding commercail rental properties from across the Unites States contains twenty-one variables, ranging from the rent amount charged to tenants to the annual precipitation of the buildings' geographoic region to whether or not the building is economically-green certified.

What is the best model for predicting price?

There are many different models that can be used to predict the prices of the buildings in our dataset. Once the different models have been established, the Roor Mean Square Errors (RMSE) of each model are compared to identify which predictive model is the most accurate. Although the prices of the buildings is not a variable in the data set, I have combined rent, gas costs, and electricity costs in order to create a price variable for each building. Whether or not a building is LEED certified and whether or not a building is Energystar certified were decided to keep seperately, as each of the two indicates it's own specific kinds of green certifications. the building ID column has also been removed.

Linear regression

The first model I have used to predict building prices was a linear model. The data has been split into two different sets, the training set that contains a sample of 80% of the buildings observations, and the testing set that contains the remaining 20% of observations. The training set is used to create a multiple linear regression model using variables that are presumed to have a significant affect on price. Once the regression model had been determined, the prices of the buildings in the testing set had then been predicted using the model and then a RMSE value was calculated between the predicted prices and the actual prices of the testing set. This process has been repeated 100 times to get an accurate RMSE, which was calculated by finding the average RMSE of all 100 tests ran.



Here is a visual of the predicted values of the linear model shown against the actual values for green versus non green buildings (nongreen=0, green=1), as well as a summary of the result. Looking at the coefficients, the predictive formula for price was able to be determined.

```
## [1] "Price = 28.69 + -0.04*(age) +  0.13*(stories) +  0.16*(green_rating)"
```

The adjusted R-squared value is pretty low, which reflects a model that fits the data poorly. We can also conclude from the plots that the model might not be a great fit since the residuals for both green and non-green buildings seem to be pretty large. The calculated RMSE value will be compared to our other models later on.

```
## RMSE = 14.92661
```

```
## R-squared = 0.02018105
```

Stepwise linear regression

Now, instead of coming up with a linear model using variables that I think significantly impact building prices, I used stepwise selections to take the base model I created and add variables that will make the multiple regression model more accurate in it's prediction of price. After the linear model using the stepwise selection method was created, it's statistic values were found.

```
## R-squared = 0.6371991
```

```
## RMSE = 15.36776
```

The r-squared value reveals that this model fits the data pretty well. This value is higher than the previous linear regression, hinting that this stepwise selected model is more accurate at predicting building prices. We udes the same 100 training/testing procedure used earlier to once again calculate a reliable RMSE value. The RMSE for our stepwise linear model is very similar to that of our first linear model, indicating the two models may not be too different in their predicting accuracies.

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? This is because the assumption you're making about the correlation may not be accurate. For example, in the podcast they discuss how the terrorist threat level impacts the amount of police out in the public, which in turn affects crime amount in that area. So yes, "Crime" and "Police" may have a correlation, but it is actually do to a confounding variable. Picking a few different cities would not give rable data since these confounding variables differ depending on the city.

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.

The researchers were able to isolate this affect by looking at confounding variables during times when police on the streets was higher during for non-crime related reasons. They noticed that the times when the terrorist threat level was raised to orange seem to have aligned with their other data, which led them to question if raised threat levels, which leads more cops to be on the streats, was actually the reason crime was lower. Before they could correctly make this conclusion, they checked to see if the there is any correlation between orange level terrorist threat days and the amount of tourists that visited D.C., which is shown in the table that there in fact is not a correlation.

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

The researchers controlled Metro ridership to determine if the reduction in crime was actually becuase the criminals were staying home due to the raised terrorist threat level, which would be seen because other people of the general public would also stay home and would not be riding the metro.

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?

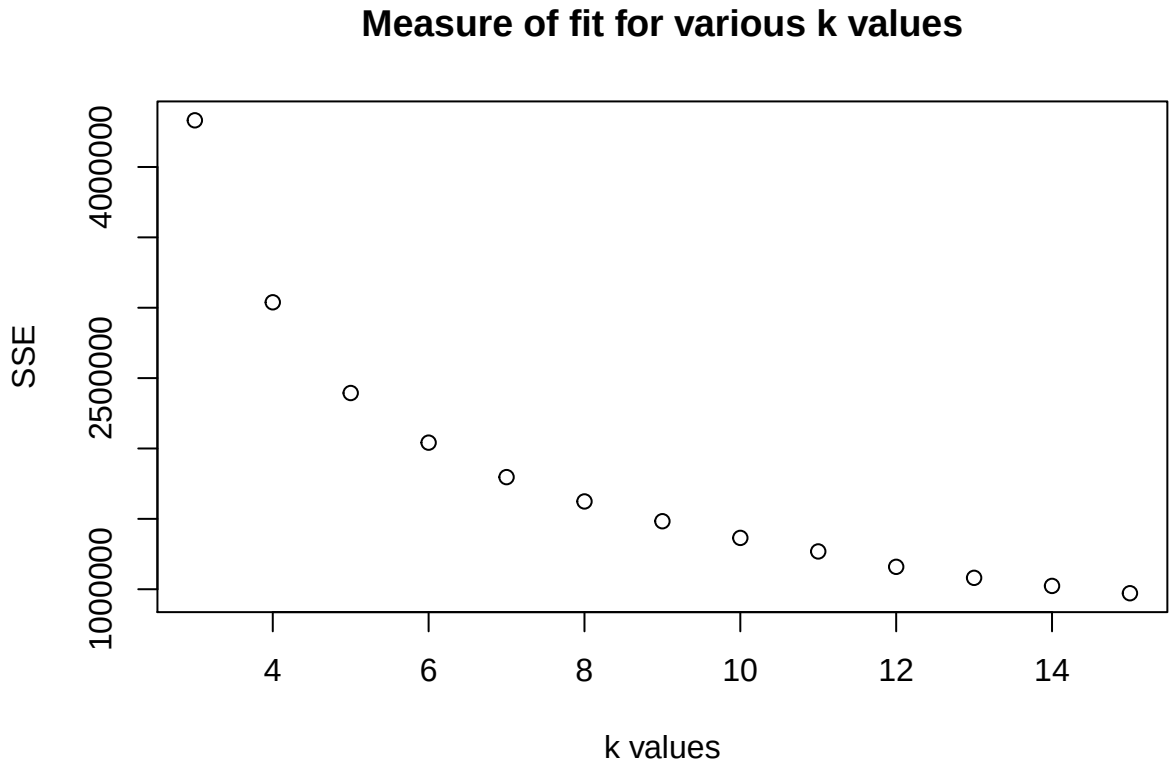
The model seems to be a predictive model describing the relationship between each variable used and crime. The two variables "High Alert x District 1" and "High Alert x Other Districts" are used to described how orange level threats in District 1 and how orabge level threats in other districs impact crime. Raised alerts in Distric 1 has a significant impact on crime, while raised alerts in other districs has a much smaller impact on crime. The "Log(midday ridership)" variable reaffirms the earlier belief that people are still out an about on high alert days. Overall, I think the model shows that reductions in crime due to more police being out is actually caused by raised terrorist threat levels.

Clustering and PCA

Our data set of various bottles of wine has descriptive properties that include 11 chemical properties and two other variables, color and quality (on a 1-10 scale). The goal was to find a model that could distinguish red wines from white wines, and potentially even sort the higher from the lower quality wines. Below is the summary of the dataset.

K-means++

First, the clusterning method of K-means++ was used to assign each wine bottle to a common centroid. This was used instead of the basic K-means clustering because K-means++ uses the bias of distance when choosing the starting centroid points, which helps reduce final cluster errors. To find out how many k number of clusters will best fit the data, we looked at a plot of trial k's and their assocaited SSE's, which describes how well using that number of k-means fits the data.

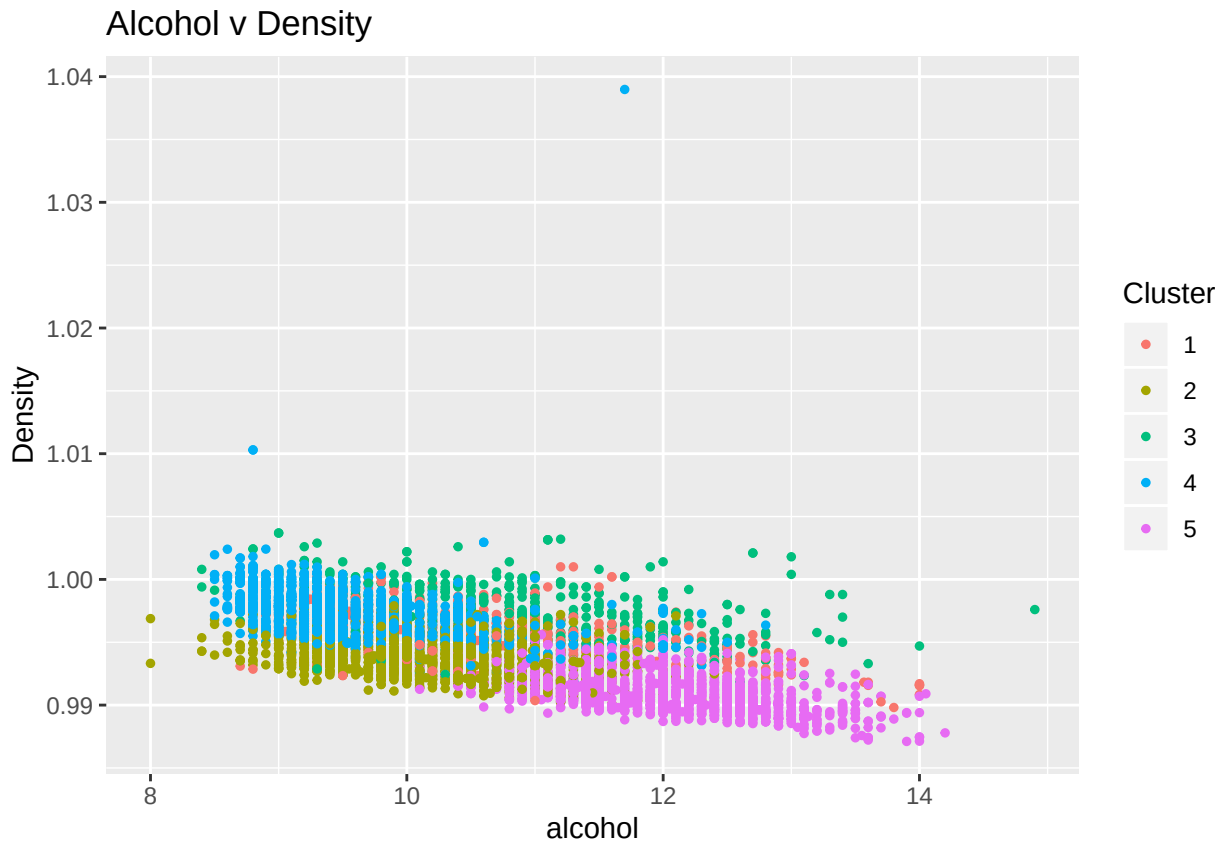


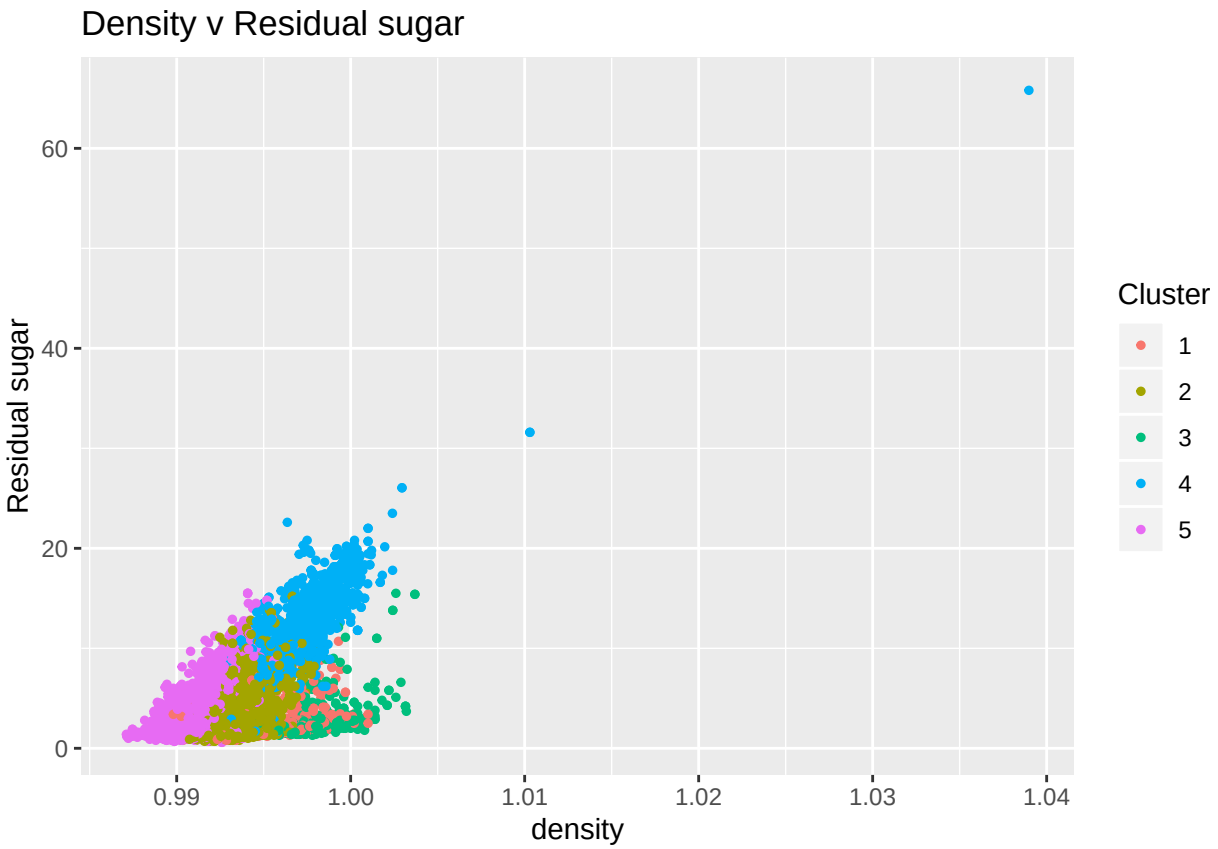
Looking at the plot above, the "elbow" point is estimated to be about 5, which is the reason that k=5 clusters was chosen. After running the wine bottle data through the K-means++ algorithm using 5 clusters, each cluster's average wine bottle was found and the chemical properties of those 5 bottles were determined:

```
##   fixed acity volatile acidity citric acid residual sugar  chlorides
## 1  7.28772112       16.4053070  10.2535727     0.05012363  0.5100609
## 2  0.61622268       50.3007284   6.7344397    31.94975639 10.1488205
## 3  0.13296566        0.9960955   0.2588063   135.96193666  6.7344397
## 4  2.47232050        3.3749740   0.3152314     0.99379685  0.2588063
## 5  0.07914152        0.5932674   3.8470463     3.25981121  0.3152314
##   free sulfur dioxide total sulfur oxide      density         pH   sulphates
## 1          3.84704629          3.2598112   0.31523143  0.9937968  0.25880633
## 2          0.05012363          0.5100609   3.84704629  3.2598112  0.31523143
## 3         31.94975639         10.1488205   0.05012363  0.5100609  3.84704629
## 4        135.96193666          6.7344397  31.94975639 10.1488205  0.05012363
## 5          0.99379685          0.2588063 135.96193666  6.7344397 31.94975639
##       alcohol
## 1 135.9619367
```

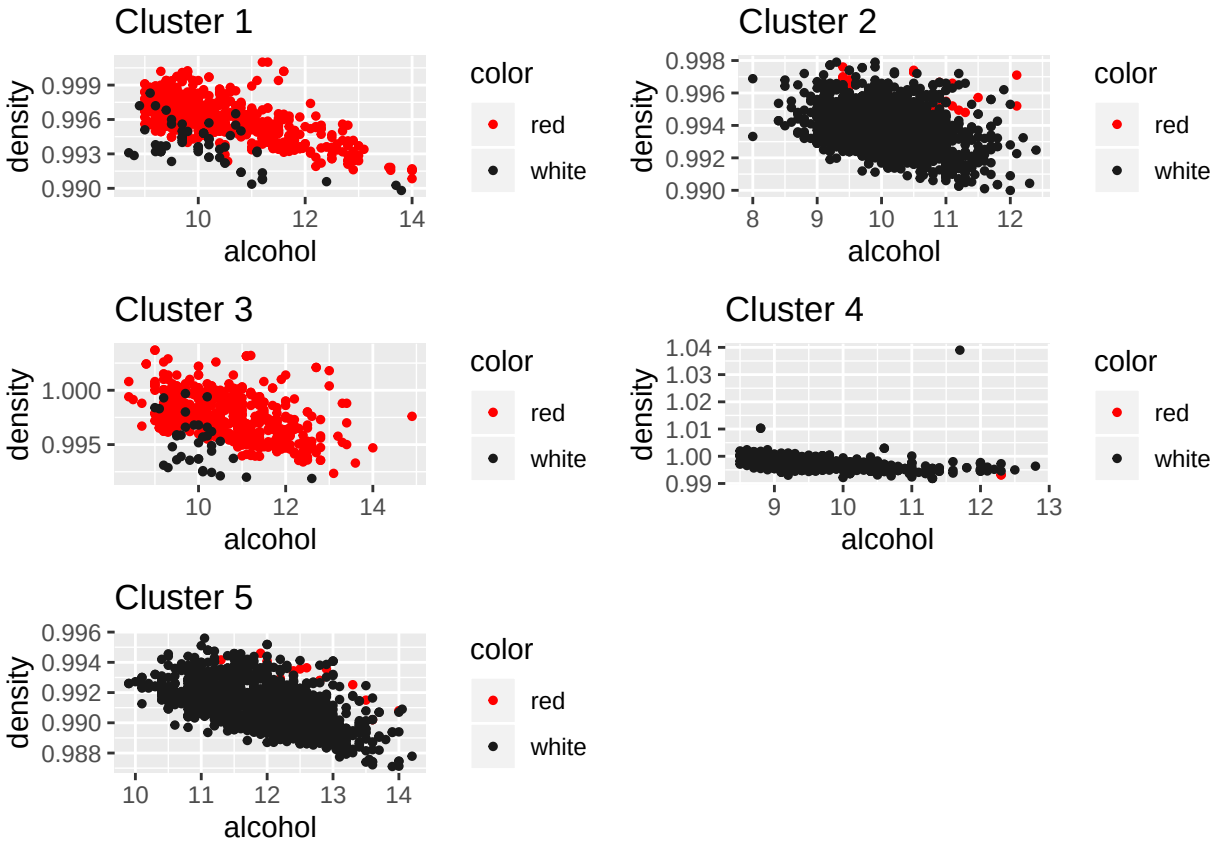
```
## 2   0.9937968
## 3   3.2598112
## 4   0.5100609
## 5  10.1488205
```

To determine whether K-means successfully formed reasonable clusters of wine, we looked at single plots of each cluster. We were able to roughly visualize the results of our clustering approach by looking at the clusters on a few of the chemical properties.





Looking at the two plots above, each cluster seemed to be relatively centrally located in a specific region of each plot, which lead us to believe that K-means++ created reasonable clusters of wines from the data. To see if the clustering method could distinguish red wines from white wines, single clusters were looked at, with color determining whether the observations in that cluster were either a red or a white wine.



From the graphs above, clusters 1, 2, and 4 are all predominantly saturated with white wines, while clusters 3 and 5 contain nearly all red wines. This was a good indicator that K-means++ performs well when sorting wine by color.

PCA

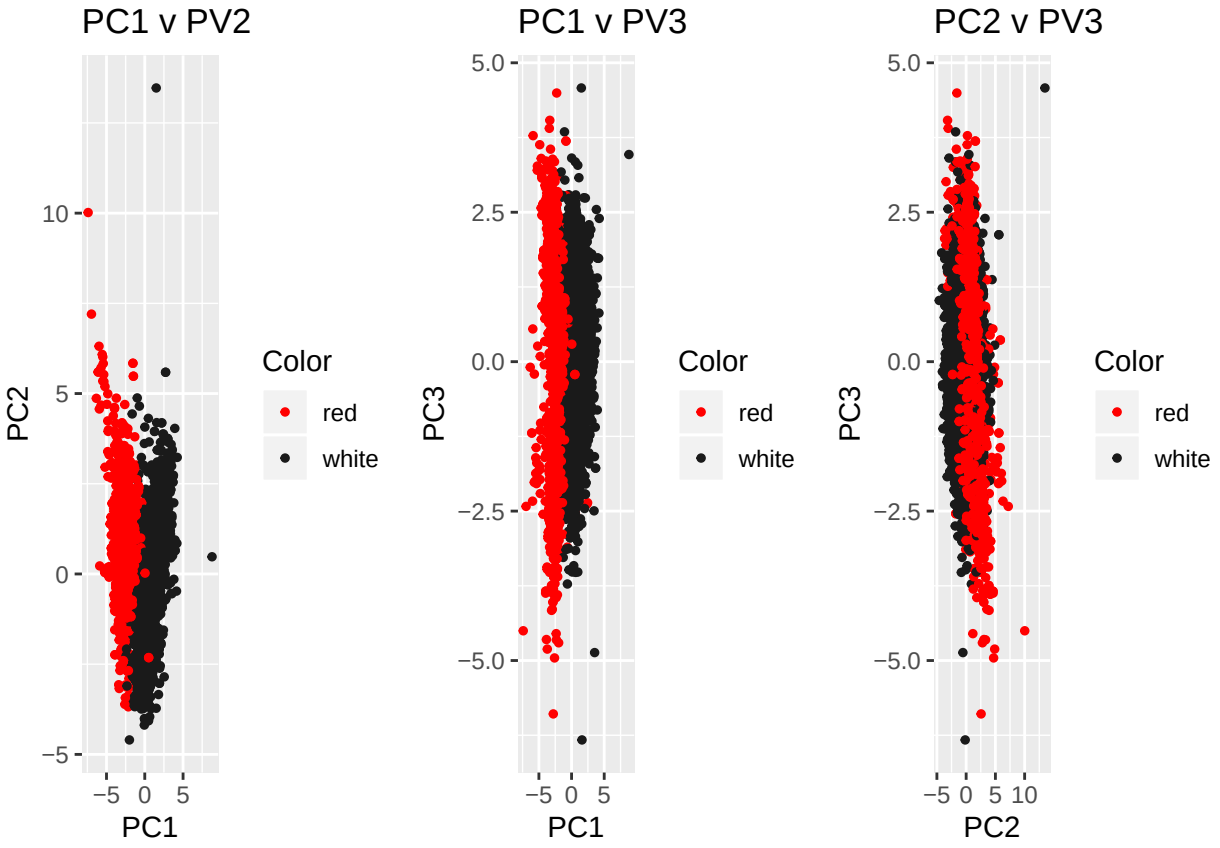
Secondly, instead of clustering to sort the wine bottles in our data, we have created a principal component analysis (PCA) to reduce the number of variables used when describing the data and distinguishing the observations. From the original chemical elements, the PCA algorithm created 11 new summary variables variables, named PC1 to PC11.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.90009
##                           PC8     PC9   PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

Each summary variable is a linear combination that maximizes the amount of variability retained from the original data. Combined, the first four of our summary variables explain about 73% of the variation in our data. The linear combinations of these components are:

```
##                          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
```

To understand how well PCA performs at identifying similar wines, we looked at plots of our top three summary variables.



In two out of our three graphs, PV1 v PV2 and PV1 v PV3, the observations seem to cluster by wine color pretty well in their own regions. The third graph of PV2 v PV3 still shows significant clustering among wine colors, except the overlapping of the clusters makes it harder to 100% tell how significant the clusters are. Overall, PCA has shown to perform well in sorting the data. Regarding the actual values of our summary variables, PC1 looks to be positive for white wines and negative for red wines. For PC2, it seems that half of each color cluster is positive and half is megative. The same goes for PC3.

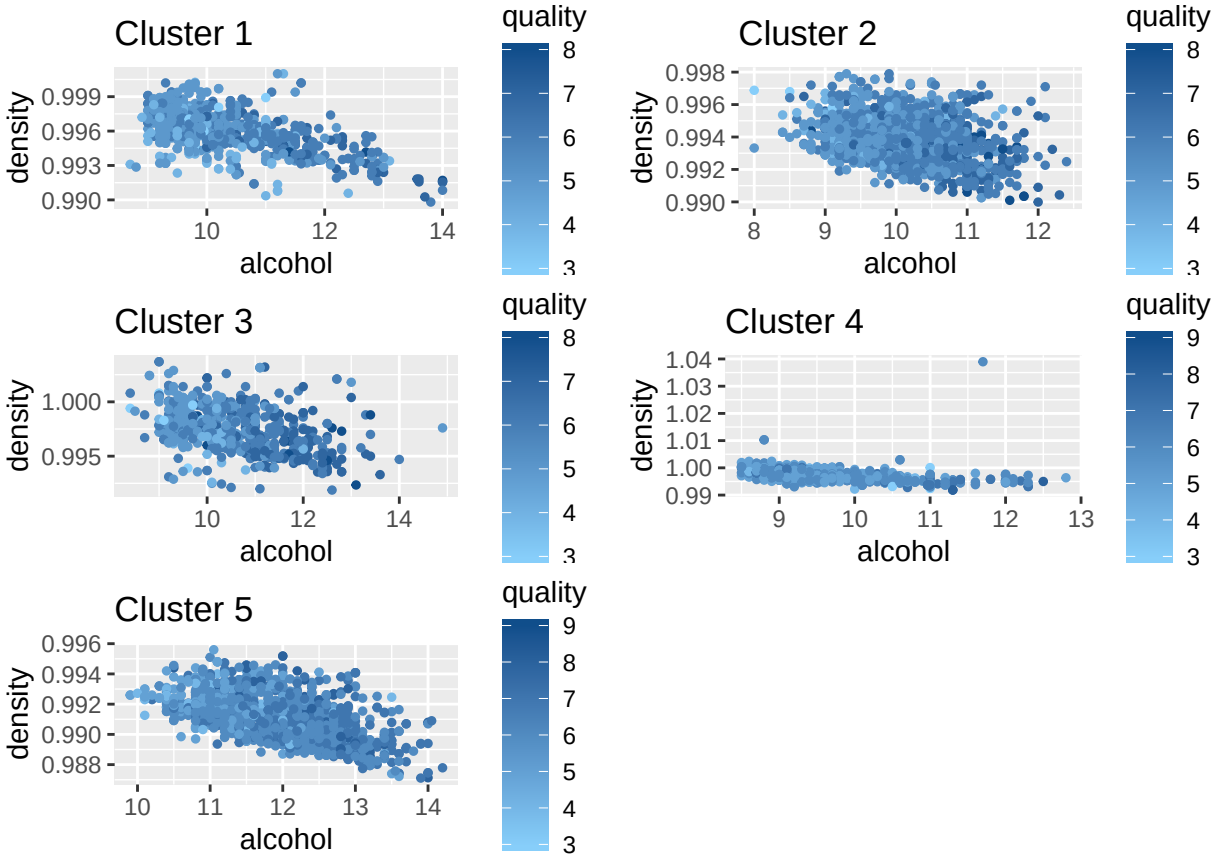
Which dimensionality technique makes more sense to you for this data?

The dimensionality reduction technique that makes the most sense to me for this data is K-means++ clustering. The single cluster plots showing which points in each cluster were white annd red showed the

high accuracy that the technique had on distinguishing out data. Although PCA is helpful since it reduces the amount of noise created by variables, I had a hard time determining whether its performance was adequate.

Convince yourself that your chosen method is easily capable of distinguishing the reds from the whites. Does this technique also seem capable of sorting the higher from the lower quality wine?

The visuals of red and white wines in each cluster frome earlier has convinced me that K-means++ is easily capable of distinguishing the reds from the whites. To see whether this method is also capable of sorting the higher from the lower quality wines, I will recreate the plot from before, except labeling each point by quality instead of color.



This clustering method does not seem to distinguish higher quality and lower quality wines very well. Each cluster has a mix of wines of all quality types. However, thepattern that higher quality wines typically have a higher alcohol concentration and a lower density than lower quality wines can be observed. It is highly possible that there is too much noise in the data set dor clustering to be succeddult. In that case, using PCA might have been a better option for sorting by quality. Another explanation could be that the qualities of the wines may be influenced more by personal preference of those who rated them than the other variables, which may explain why K-means++ was not able to to use the 11 chemical variables to distinguish quality.

Market segmentation

The social marketing data collected by NutrientH20's advertizing firm contains 36 categories that single tweets were categorized into. Our goal was to analyze the data and give feedback to the firm if any interesting

details regarding market segmentations were found. Market segments can be described as a subgroups of a population that consists of peoople with similar cahractereistics related to that market. Since it is a nutrition company, we are focused on discovering groupings of categories that may be associated with interest in nutrition. Before we could decide in which direction to go to analyze the data, we altered the original dataset, changing each entry from a count to a frequency so we could see the proportions of categories that each user tweeted about. Since the "spam" and "adult" categories are not of high interest to the company, those categories were removed when calculating each frequency.

```
##      chatter current_events     travel photo_sharing uncategorized    tv_film
## 1 0.03278689     0.00000000 0.03278689    0.03278689    0.03278689 0.01639344
## 2 0.10000000     0.10000000 0.06666667    0.03333333    0.03333333 0.03333333
## 3 0.12765957     0.06382979 0.08510638    0.06382979    0.02127660 0.10638298
## 4 0.04761905     0.23809524 0.09523810    0.09523810    0.00000000 0.04761905
## 5 0.16666667     0.06666667 0.00000000    0.20000000    0.03333333 0.00000000
## 6 0.17647059     0.11764706 0.05882353    0.20588235    0.00000000 0.02941176
##   sports_fandom   politics       food     family home_and_garden      music
## 1    0.01639344 0.00000000 0.06557377 0.01639344      0.03278689 0.00000000
## 2    0.13333333 0.03333333 0.06666667 0.06666667      0.03333333 0.00000000
## 3    0.00000000 0.04255319 0.02127660 0.02127660      0.02127660 0.02127660
## 4    0.00000000 0.04761905 0.00000000 0.04761905      0.00000000 0.00000000
## 5    0.00000000 0.06666667 0.00000000 0.03333333      0.00000000 0.00000000
## 6    0.02941176 0.00000000 0.05882353 0.02941176      0.02941176 0.02941176
##        news online_gaming   shopping college_uni sports_playing    cooking
## 1 0.0000000           0.0 0.01639344  0.00000000     0.03278689 0.08196721
## 2 0.0000000           0.0 0.00000000  0.00000000     0.03333333 0.00000000
## 3 0.0212766           0.0 0.04255319  0.00000000     0.00000000 0.04255319
## 4 0.0000000           0.0 0.00000000  0.04761905     0.00000000 0.00000000
## 5 0.0000000           0.1 0.06666667  0.13333333     0.00000000 0.03333333
## 6 0.0000000           0.0 0.14705882  0.00000000     0.00000000 0.00000000
##          eco  computers   business   outdoors     crafts automotive       art
## 1 0.01639344 0.01639344 0.00000000 0.03278689 0.01639344 0.00000000 0.0000000
## 2 0.00000000 0.00000000 0.03333333 0.00000000 0.06666667 0.00000000 0.0000000
## 3 0.02127660 0.00000000 0.00000000 0.00000000 0.04255319 0.00000000 0.1702128
## 4 0.00000000 0.00000000 0.04761905 0.00000000 0.14285714 0.00000000 0.0952381
## 5 0.00000000 0.03333333 0.00000000 0.03333333 0.00000000 0.00000000 0.0000000
## 6 0.00000000 0.02941176 0.02941176 0.00000000 0.00000000 0.02941176 0.0000000
##     religion     beauty  parenting     dating    school personal_fitness
## 1 0.01639344 0.00000000 0.01639344 0.01639344 0.0000000        0.1803279
## 2 0.00000000 0.00000000 0.00000000 0.03333333 0.1333333        0.0000000
## 3 0.00000000 0.02127660 0.00000000 0.02127660 0.0000000        0.0000000
## 4 0.00000000 0.04761905 0.00000000 0.00000000 0.0000000        0.0000000
## 5 0.00000000 0.00000000 0.00000000 0.00000000 0.0000000        0.0000000
## 6 0.00000000 0.00000000 0.00000000 0.00000000 0.0000000        0.0000000
##     fashion small_business
## 1 0.0000000     0.00000000
## 2 0.0000000     0.00000000
## 3 0.0212766     0.00000000
## 4 0.0000000     0.00000000
## 5 0.0000000     0.03333333
## 6 0.0000000     0.00000000
```

The summary of the data above can be used to get a rough idea of how often each interest/category was mentioned. The category that had the highest frequency of being categoriezd into it was "chatter". This

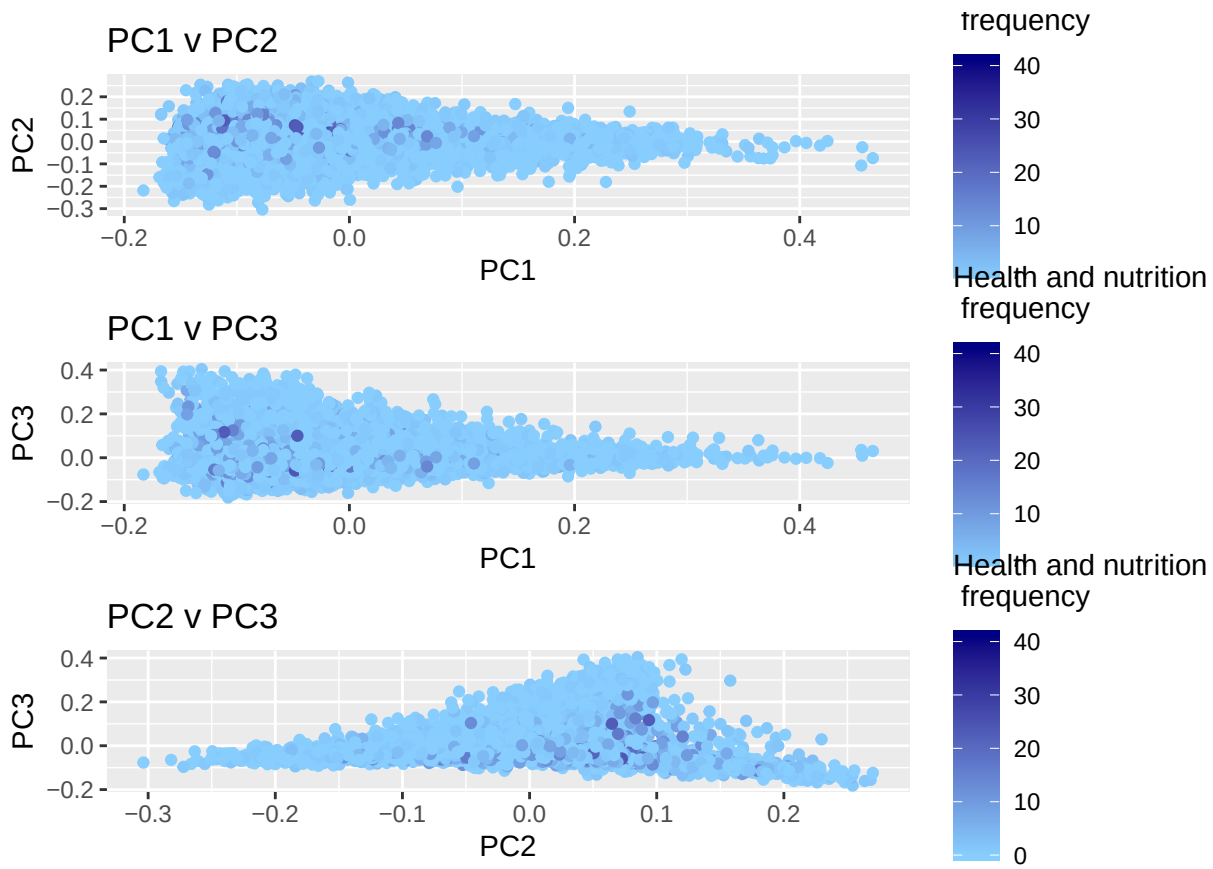
does not come as a surprise since social media is heavily used to keep in contact with or to "chatter" with others.

We determined that principal component analysis could be effective in reducing the 34 categories we were considering into just a few summary variables. These PC variables retain the maximum variance from the data set they are pulled from, and allowed us to categorize the main types of audiences that follow NutrientH2O's twitter account.

After all 34 summary variable had been calculated, we were abpe tp see that the first six component variables describe about 56%% of the variance in our data. More specifically, we observed and evaluated the first three, PC1, PC2, and PC3, which combine to explain almost 40% of the data, to evaulate any market segmentations.

```
##                       PC1     PC2     PC3
## chatter            0.8514 -0.1159  0.0660
## current_events     0.0770 -0.0579  0.0050
## travel            -0.0513 -0.2924 -0.0480
## photo_sharing      0.3310  0.2462 -0.1170
## uncategorized      0.0167  0.0077  0.0069
## tv_film           -0.0329 -0.0379  0.1003
## sports_fandom     -0.0968 -0.1276 -0.0444
## politics          -0.1095 -0.5367 -0.1516
## food              -0.0985 -0.0324 -0.0222
## family            -0.0165 -0.0282 -0.0002
## home_and_garden    0.0059 -0.0053 -0.0016
## music              0.0023  0.0242  0.0193
## news              -0.1075 -0.3050 -0.0996
## online_gaming     -0.0887  0.1100  0.5621
## shopping           0.2162  0.0350 -0.0213
## college_uni       -0.0979  0.1146  0.6668
## sports_playing    -0.0150  0.0257  0.1090
## cooking           -0.1332  0.5432 -0.3382
## eco                0.0116  0.0032 -0.0080
## computers         -0.0192 -0.0942 -0.0321
## business           0.0122 -0.0063 -0.0042
## outdoors          -0.0573  0.0234 -0.0467
## crafts            -0.0026 -0.0071 -0.0024
## automotive        -0.0150 -0.1164 -0.0283
## art               -0.0452  0.0011  0.0375
## religion          -0.0953 -0.0384 -0.0309
## beauty            -0.0315  0.1381 -0.0908
## parenting         -0.0607 -0.0353 -0.0340
## dating             0.0060 -0.0047 -0.0041
## school            -0.0195 -0.0053 -0.0298
## personal_fitness  -0.0949  0.1294 -0.1130
## fashion           -0.0333  0.2277 -0.1294
## small_business     0.0027 -0.0040  0.0109
```

We then plotted each pair of components against each other and have color-labeled each point based on its associated "health and nutrition" frequency. By doing this, we are able to identify what other categories are mentioned by twitter users who are interested in health and nutrition.



Each pairs-model shows distinguishable clustering of frequency rates for heath and nutrition related tweets. Out first summary variable, PC1, has the highest health and nutrition frequency observations as typically being negative. The second component identifies most of these high frequency observations as being slightly above zero. For PC3 the high frequencies tend to be slightly below zero. These models infer that the PCA method has performed moderately well at distinguishing twitter users by health and nutrition frequency. The clusters of the high frequencies are didentifiable, but not enough to make strong claims. From here, we can look at the top categories that are the most significant of each PC.

```
## PC1: #1 chatter  #2 photo_sharing  #3 shopping  #4  current_events

## PC2: #1 cooking  #2 photo_sharing  #3 fashion  #4  beauty

## PC3: #1 college_uni  #2 online_gaming  #3 sports_playing  #4  tv_film
```

The most significant market segment that may be used to predict health and nutrition interest includes chatter, photo sharing, shopping, and current events, and it accounts for amost 18% of the data collected. The second most influential market segment consists of those who tweet about cooking, photo sharing, fashion, and beauty. Another market segment is identified by college status, online gaming, playing sports, and TV/film. These can be used to generalize specific target audiences and use social media advertizing strategies to reach them.