Green_Buildings

2) Predictive Model for Green Buildings

Approach

We set out to build a predictive model for price of buildings given characteristics of the building and its surrounding area. Once attaining this model, we hope to determine the average change in rental income per square foot associated with green certification. In order to determine the best predictive model, we combined two approaches to model optimization: stepwise selection and lasso fits.

First we used stepwise selection to determine which variables, including interactions contributed the most to variation in the data. We began with a baseline model of five variables that we estimated would have the largest impact on property rent: Gas_Costs, Electricity_Costs, net utility cost, class_a, and green_rating. From the resulting model, we were able to determine one prediction for the impact of green rating on property price.

In order to confirm our results, we incorporated the important interactions discovered by the stepwise selected model to run a lasso fit to create a regularized model from all the available variables from the dataset. The resulting model incorporated only the variables that contributed substantially to the variance in the data. From this resulting model, we were able to hold other relevant variables constant to determine the isolated impact of green certification on property price.

Results

The initial stepwise selection on the baseline medium model yielded the following optimal model for predicting property price.

```
## Warning: package 'gamlr' was built under R version 3.4.4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Start:  AIC=40635.58
## Rent ~ Gas_Costs + Electricity_Costs + net + class_a + green_rating
##
##                                  Df Sum of Sq     RSS   AIC
## + Gas_Costs:Electricity_Costs     1    129532 1280710 39884
## + Gas_Costs:class_a               1      5984 1404257 40604
## + Gas_Costs:net                   1      5026 1405216 40610
## + Electricity_Costs:class_a       1      2678 1407564 40623
## + Electricity_Costs:net           1      1724 1408517 40628
## <none>                                       1410241 40636
## + Electricity_Costs:green_rating  1       118 1410124 40637
## + net:green_rating                1       114 1410127 40637
## + net:class_a                     1        83 1410158 40637
## + Gas_Costs:green_rating          1        25 1410216 40637
## + class_a:green_rating            1         0 1410241 40638
## - green_rating                    1      1513 1411754 40642
## - net                             1     10880 1421121 40694
## - Gas_Costs                       1     17978 1428220 40733
## - class_a                         1     71371 1481612 41020
## - Electricity_Costs               1    283436 1693677 42066
##
## Step:  AIC=39884.15
## Rent ~ Gas_Costs + Electricity_Costs + net + class_a + green_rating +
##     Gas_Costs:Electricity_Costs
##
##                                  Df Sum of Sq     RSS   AIC
## + Electricity_Costs:class_a       1     12027 1268682 39812
## + Electricity_Costs:net           1       797 1279913 39881
## + Gas_Costs:class_a               1       558 1280152 39883
## <none>                                       1280710 39884
## + Gas_Costs:net                   1       217 1280493 39885
## + Gas_Costs:green_rating          1       212 1280498 39885
## + Electricity_Costs:green_rating  1       146 1280563 39885
## + net:class_a                     1        58 1280652 39886
## + net:green_rating                1        51 1280659 39886
## + class_a:green_rating            1        26 1280684 39886
## - green_rating                    1      1238 1281947 39890
## - net                             1      4348 1285058 39909
## - class_a                         1     54304 1335014 40207
## - Gas_Costs:Electricity_Costs     1    129532 1410241 40636
##
## Step:  AIC=39812.37
## Rent ~ Gas_Costs + Electricity_Costs + net + class_a + green_rating +
##     Gas_Costs:Electricity_Costs + Electricity_Costs:class_a
##
##                                  Df Sum of Sq     RSS   AIC
## + Gas_Costs:class_a               1      2230 1266452 39801
## + Electricity_Costs:green_rating  1      1204 1267478 39807
## <none>                                       1268682 39812
## + Electricity_Costs:net           1       285 1268397 39813
## + net:class_a                     1       138 1268544 39814
## + net:green_rating                1        70 1268613 39814
## + class_a:green_rating            1        61 1268621 39814
## + Gas_Costs:green_rating          1        60 1268622 39814
```

```
## + Gas_Costs:net                   1        48 1268634 39814
## - green_rating                    1      1051 1269734 39817
## - net                             1      3834 1272516 39834
## - Electricity_Costs:class_a       1     12027 1280710 39884
## - Gas_Costs:Electricity_Costs     1    138881 1407564 40623
##
## Step:  AIC=39800.61
## Rent ~ Gas_Costs + Electricity_Costs + net + class_a + green_rating +
##     Gas_Costs:Electricity_Costs + Electricity_Costs:class_a +
##     Gas_Costs:class_a
##
##                                     Df Sum of Sq     RSS   AIC
## + Gas_Costs:Electricity_Costs:class_a  1      2951 1263502 39784
## + Electricity_Costs:green_rating       1      1550 1264903 39793
## + Electricity_Costs:net                1       439 1266013 39800
## <none>                                             1266452 39801
## + Gas_Costs:net                        1       175 1266278 39802
## + class_a:green_rating                 1        97 1266355 39802
## + net:green_rating                     1        62 1266391 39802
## + net:class_a                          1        60 1266392 39802
## + Gas_Costs:green_rating               1        13 1266439 39803
## - green_rating                         1      1162 1267614 39806
## - Gas_Costs:class_a                    1      2230 1268682 39812
## - net                                  1      3581 1270033 39821
## - Electricity_Costs:class_a            1     13700 1280152 39883
## - Gas_Costs:Electricity_Costs          1    132772 1399224 40578
##
## Step:  AIC=39784.37
## Rent ~ Gas_Costs + Electricity_Costs + net + class_a + green_rating +
##     Gas_Costs:Electricity_Costs + Electricity_Costs:class_a +
##     Gas_Costs:class_a + Gas_Costs:Electricity_Costs:class_a
##
##                                     Df Sum of Sq     RSS   AIC
## + Electricity_Costs:green_rating       1    1752.5 1261749 39776
## + Electricity_Costs:net                1     523.5 1262978 39783
## <none>                                             1263502 39784
## + Gas_Costs:net                        1     234.4 1263267 39785
## + Gas_Costs:green_rating               1     124.2 1263377 39786
## + net:green_rating                     1      85.8 1263416 39786
## + class_a:green_rating                 1      75.8 1263426 39786
## + net:class_a                          1       5.3 1263496 39786
## - green_rating                         1    1149.9 1264651 39789
## - Gas_Costs:Electricity_Costs:class_a  1    2950.5 1266452 39801
## - net                                  1    3347.0 1266849 39803
##
## Step:  AIC=39775.51
## Rent ~ Gas_Costs + Electricity_Costs + net + class_a + green_rating +
##     Gas_Costs:Electricity_Costs + Electricity_Costs:class_a +
##     Gas_Costs:class_a + Electricity_Costs:green_rating + Gas_Costs:Electricity_Costs:class_a
##
##                                     Df Sum of Sq     RSS   AIC
## + Electricity_Costs:net                1     528.1 1261221 39774
## <none>                                             1261749 39776
## + Gas_Costs:net                        1     246.0 1261503 39776
```

```
## + class_a:green_rating                 1     146.9 1261602 39777
## + Gas_Costs:green_rating               1      95.1 1261654 39777
## + net:green_rating                     1      42.1 1261707 39777
## + net:class_a                          1       5.4 1261744 39777
## - Electricity_Costs:green_rating       1    1752.5 1263502 39784
## - Gas_Costs:Electricity_Costs:class_a  1    3153.5 1264903 39793
## - net                                  1    3440.6 1265190 39795
##
## Step:  AIC=39774.24
## Rent ~ Gas_Costs + Electricity_Costs + net + class_a + green_rating +
##     Gas_Costs:Electricity_Costs + Electricity_Costs:class_a +
##     Gas_Costs:class_a + Electricity_Costs:green_rating + Electricity_Costs:net +
##     Gas_Costs:Electricity_Costs:class_a
##
##                                       Df Sum of Sq     RSS   AIC
## <none>                                              1261221 39774
## + class_a:green_rating                 1     149.3 1261072 39775
## - Electricity_Costs:net                1     528.1 1261749 39776
## + Gas_Costs:green_rating               1      96.7 1261124 39776
## + net:green_rating                     1      85.4 1261136 39776
## + Gas_Costs:net                        1       4.3 1261217 39776
## + net:class_a                          1       0.0 1261221 39776
## - Electricity_Costs:green_rating       1    1757.1 1262978 39783
## - Gas_Costs:Electricity_Costs:class_a  1    3241.1 1264462 39792

## lm(formula = Rent ~ Gas_Costs + Electricity_Costs + net + class_a +
##     green_rating + Gas_Costs:Electricity_Costs + Electricity_Costs:class_a +
##     Gas_Costs:class_a + Electricity_Costs:green_rating + Electricity_Costs:net +
##     Gas_Costs:Electricity_Costs:class_a, data = gb)

## Electricity_Costs:green_rating
##                      -1.256927
```

Based on this initial stepwise selected model, we combined the impact of the main effect of a green rating and the interaction effect between green rating and electricity cost to determine the anticipated effect of a green rating on property price. Green rating by itself had an effect of approximately a $5.58 increase in rent price. However, when incorporating the interaction effect of electricity cost and green rating, the overall rent in green buildings is on average $1.26 cheaper per square foot than in non-green buildings.

Incorporating the important interactions identified by the stepwise model selection, we created a new regularized model using lasso fit. From the coefficients of each variable identified, we were able to identify the effect of green rating on property price in this model.

```
## 25 x 1 sparse Matrix of class "dgCMatrix"
##                                        seg100
## intercept                        1.043502e+00
## cluster                          8.895361e-05
## size                             1.282580e-07
## empl_gr                          1.305377e-02
## leasing_rate                     2.479714e-03
## stories                         -3.459884e-04
## age                             -7.839404e-04
## renovated                       -4.731742e-02
## class_a                          1.032433e-02
```

```
## class_b                               5.028472e-02
## green_rating                          .
## net                                  -1.504949e-01
## amenities                             1.762456e-02
## cd_total_07                          -8.798106e-05
## hd_total07                            1.150198e-05
## total_dd_07                           .
## Precipitation                         1.353562e-02
## Gas_Costs                             4.946163e+01
## Electricity_Costs                     5.304793e+01
## Gas_Costs:Electricity_Costs          -1.799936e+03
## class_a:Electricity_Costs             4.778497e+00
## class_a:Gas_Costs                     .
## green_rating:Electricity_Costs       -6.774646e-01
## net:Electricity_Costs                 .
## class_a:Gas_Costs:Electricity_Costs -6.865535e+01
```

```
## [1] -0.02089218
```

This model did not find green ratings to have a substantial main effect on property price on its own. However, from the interaction effect between green rating and electricity costs, property prices seem to be approximately $0.21 less on average in green buildings versus non-green buildings.

Conclusion

From our analysis, we determined that green rated buildings on average tend to have cheaper rent per square foot than non-green buildings, driven mainly by the resulting lower electricity costs of the property. The range for this decrease in rent prices may be between $0.21 and $1.26 in price per square foot.

What Causes What?

1) We cannot simply run a broad comparison of number of cops on the street and crime levels in order to draw causality between cops and crime because there could be alternative reasons for causality. For example, there could be more cops on the street in cities that have more crime in an attempt to combat the higher crime levels.
2) The researchers were able to find a natural experiment in DC where terrorism threats meant that there would be high levels of cops on the street independent of the amount of other crimes on high alert days. This means researchers did not have to worry about the confounding effect of more crime on cop numbers. They could simply compare daily crime amounts on high alert days to normal days. On high alert days when there were more cops in the city, researchers found that daily crimes decreased by about 7.
3) Researchers controlled for Metro ridership because they hypothesized that there may be a confounding variable in a decrease in crime on high alert days simply because more people would stay at home in response to the threat or seeing a large number of cops in the street. Metro ridership provides a proxy for the number of people out in the streets. By controlling for Metro ridership, they are able to determine whether the effect between crimes and cops holds true even holding the number of people out and about to be constant.
4) Table 4 is the results of running the same model estimating crime amounts and high alert days, holding metro ridership constant, on just the national mall then on all other districts. It found that there was a much more substantial decrease in crime on high alert days in the National Mall compared to all other districts. Perhaps this is because the largest concentration of the policeman increase happens in the National Mall on high alert terror days.

Clustering and PCA of Wine

With this dataset, we explored both PCA and a clustering alogrithm on the 11 chemical properties in the “wine.csv” dataset. These 11 different chemical properties include:

1. fixed.acidity
2. volatile.acidity
3. citric.acid
4. residual.sugar
5. chlorides
6. free.sulfur.dioxide
7. total.sulfur.dioxide
8. density
9. pH
10. sulphates
11. alcohol

Objectives

Using these properties, we aim to create clusters which distinguishes red and white wine, and if possible, sort the higher and lower quality wines into different clusters.

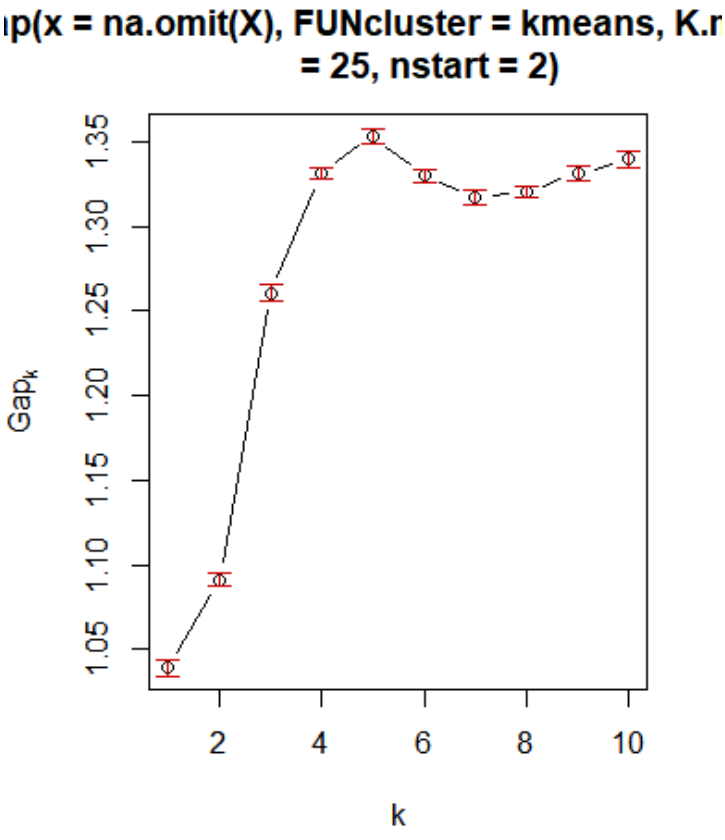
Clustering Algorithm (K-means)

Our choice was to use k-means clustering instead of hierarchial clustering because it allowed for better interpretability and made more sense to cluster red and white wines together.

We first ran clusGap in the following code snippet below:

```
library(cluster)
wine_gap = clusGap(na.omit(X) ,FUN=kmeans, nstart=2, K.max=10,B=25)
plot(wine_gap)
```

This code gave us the following elbow plot which we used to choose the optimal amount of clusters for k-means.



From this plot, we were able to see the dip around 5, which we confirmed by looking at the output of wine_gap. This would be the amount of k-means clusters we would use for the wine.csv data.

```
> wine_gap
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = na.omit(X), FUNcluster = kmeans, K.max = 10, B = 25,     nstart = 2)
B=25 simulated reference sets, k = 1..10; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstSEmax', SE.factor=1): 5
          logW   E.logW      gap      SE.sim
 [1,] 11.63833 12.67760 1.039271 0.004926141
 [2,] 11.16051 12.25156 1.091050 0.003856928
 [3,] 10.82628 12.08691 1.260627 0.005145008
 [4,] 10.65465 11.98593 1.331285 0.003704118
 [5,] 10.53122 11.88465 1.353434 0.004376664
 [6,] 10.45671 11.78644 1.329725 0.003719523
 [7,] 10.39759 11.71446 1.316875 0.004480327
 [8,] 10.33239 11.65299 1.320597 0.003128621
 [9,] 10.27823 11.60931 1.331075 0.004167755
[10,] 10.23165 11.57129 1.339641 0.004851971
```

With 5 clusters, we ran the k-means algorithm and obtained the following table.

```
##                       Cen1        Cen2        Cen3        Cen4        Cen5
## RedCount        636.000000   32.000000   25.000000    4.000000  902.000000
## WhiteCount       35.000000 1610.000000 1629.000000 1565.000000   59.000000
## Avg Red Qual      5.872642    5.343750    6.360000    6.500000    5.455654
## Avg White Qual    4.914286    5.764596    6.294659    5.623642    4.779661
```

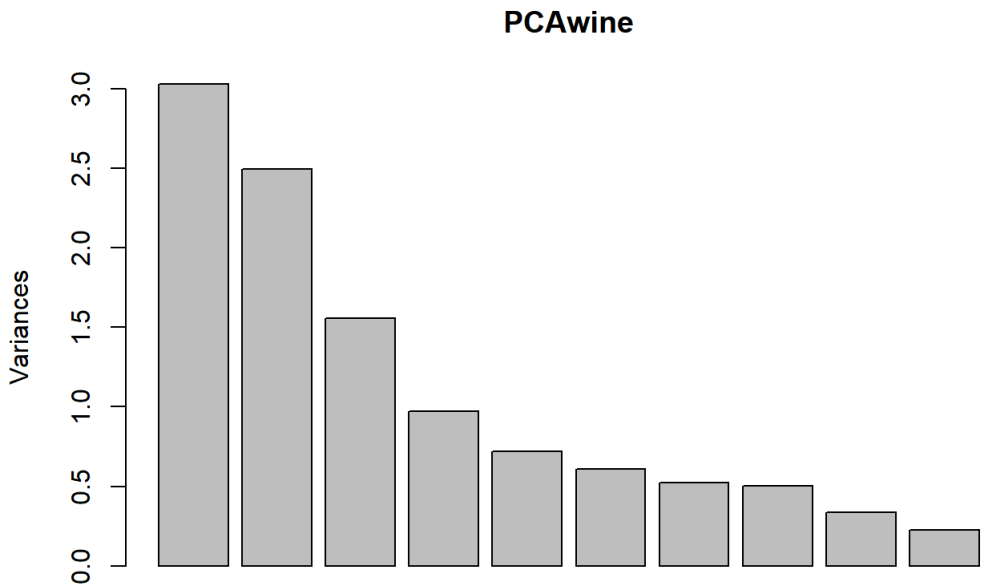
Analysis of the Clustering Algorithm

From this table, we can clearly see that the five different centers were able to differentiate red and white wine pretty accurately. Each of the centers was able to cluster the type of wine based only based on the chemical properties. In each of the centers, there is a clear majority of what type of wine is classified.

This technique was able to differentiate red and white wine accurately, but the quality of wine across the clusters was not easily differentiable. The range of quality between the clusters was from around 5.3 to 6.5 for red wine and 4.7 to 6.3 for white wine. Since the original range is between 1-10, this clustering algorithm depicts that in each cluster, the quality of wine averages out to 5 or 6. This technique does not seem capable of sorting the higher from the lower quality of wines.

PCA

With PCA, we first had to clean the data by subtracting quality and color and then we ran prcomp which gave us the following plot.

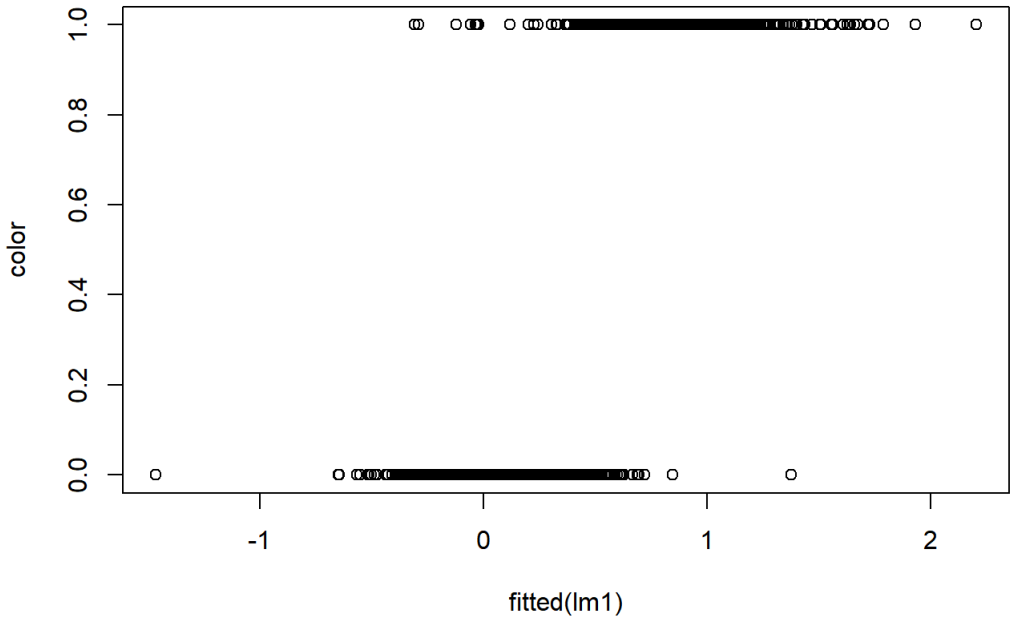


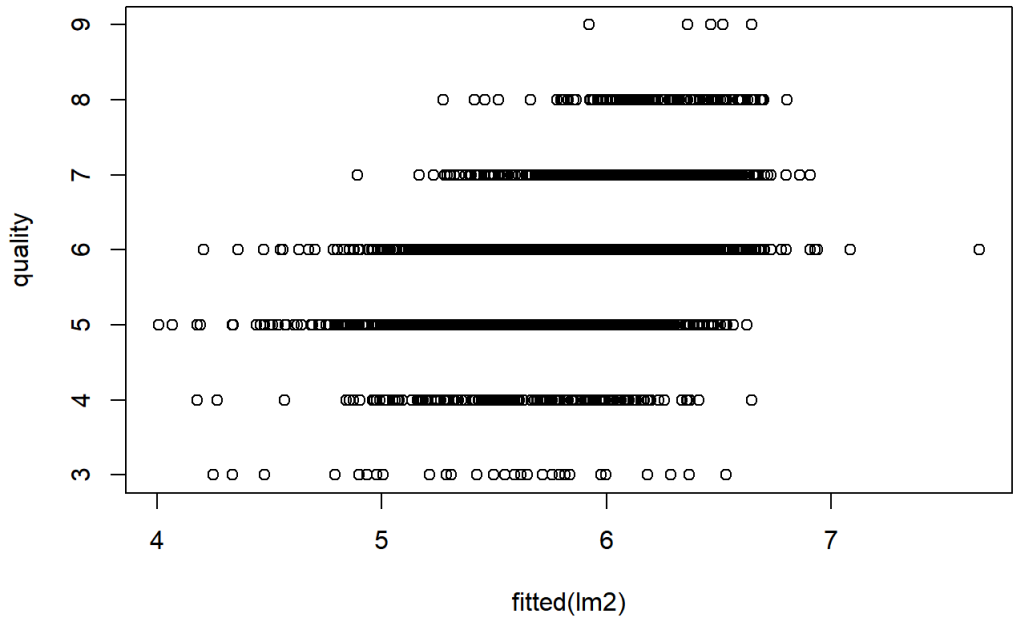
We took the first five PCA's since these are the highest values and would impact the linear models the most out of the total 11. The values below depict how each of the variables would affect the linear model in distinguishing red vs white wine.

```
##                         PC1   PC2   PC3   PC4   PC5
## fixed.acidity         -0.24  0.34 -0.43  0.16 -0.15
## volatile.acidity      -0.38  0.12  0.31  0.21  0.15
## citric.acid            0.15  0.18 -0.59 -0.26 -0.16
## residual.sugar         0.35  0.33  0.16  0.17 -0.35
## chlorides             -0.29  0.32  0.02 -0.24  0.61
## free.sulfur.dioxide    0.43  0.07  0.13 -0.36  0.22
## total.sulfur.dioxide   0.49  0.09  0.11 -0.21  0.16
## density               -0.04  0.58  0.18  0.07 -0.31
## pH                    -0.22 -0.16  0.46 -0.41 -0.45
## sulphates             -0.29  0.19 -0.07 -0.64 -0.14
## alcohol               -0.11 -0.47 -0.26 -0.11 -0.19
```

Next, we plotted PC1 vs. PC2 and there was a clear separation between red and white wine.

From this data, we fitted a linear model to predict color and quality from the first five PCA's.





Analysis of PCA

With the first fitted plot, there is a clear distinction between 0 and 1, which denotes white and red wine respectively. Again, this proves that using the first 5 PCA's, we are able to separate red and white wines from eachother.

On the other hand, fitting the same PCA's against the quality, we can see that the model has trouble separating the wine based on qualtiy. An example of this would be at a guess of 6, the actual quality has more of a density around 7 or 8, which is on the higher end of the quality of wines. Further, it is not as interpretable to distinguish the quality of wines easily as it is with the color.

Conclusion

After exploring both PCA and k-means clustering, we decided that k-means clustering made more sense for us for this data. This method was easy to understand because it had the goal of clustering red and white wines together, which it was able to do easily. This interpretability was the reason we believe k-means clustering was the right choice. For both PCA and k-means, they were able distinguish red and white wine, although both had difficulty sorting the higher from the lower quality wines. In each of the algorithms, they quality of wine seemed to average out toward the middle instead of creating clear distinctions between "higher" and "lower" quality.

Market Segmentation

Stephenson Gokingco, Akash Thakkar, Caroline Hao, James Cornejo

4/19/2020

Exercises 3: Market Segmentation

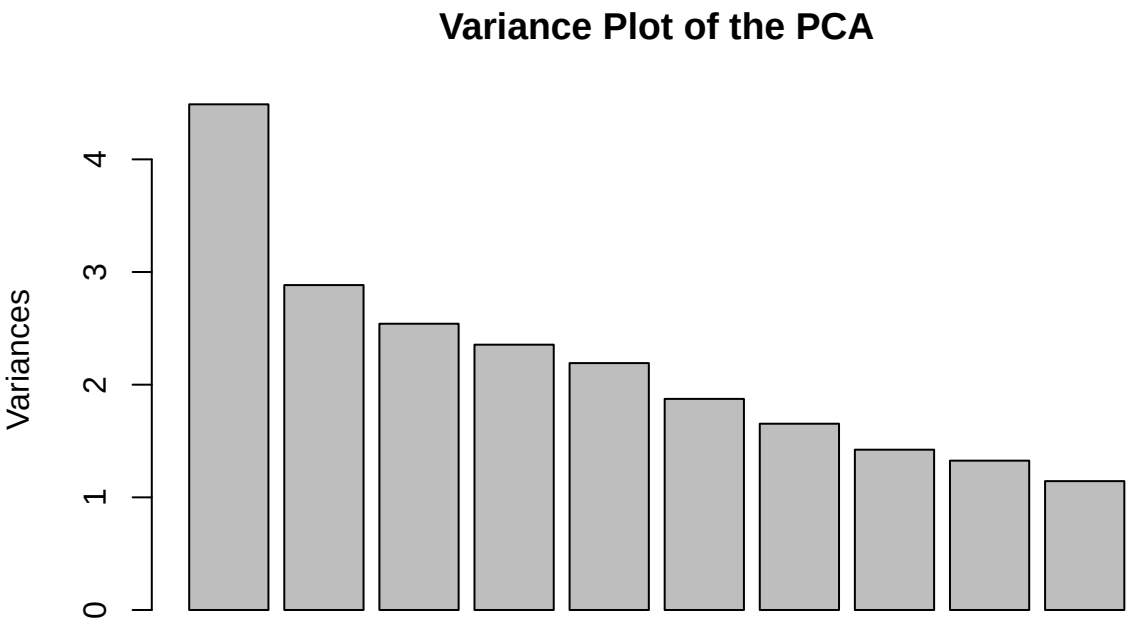
The data in social_marketing.csv was collected in the course of a market-research study using followers of the Twitter account of a large consumer brand, NutrientH20. The goal here was for NutrientH20 to understand its social-media audience a little bit better, so that it could hone its messaging a little more sharply.

After initally trying out a principal component analysis that was rendered uninterpretable, we determined that a market segment is best defined as a cluster. We used k-means clustering (specifically k-means plus plus) and were able to to come up with some interesting, well-supported insights about the audience that will give NutrientH20 some insight as to how they might position their brand to maximally appeal to each market segment.

Let's run the PCA for this dataset.

```
PCAtwit = prcomp(smdata, scale=TRUE)
```

We get the following variance plot:



Then, we get get the cumulative variances of each of the components.

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1186 1.69825 1.59398 1.53458 1.48030 1.36887 1.28581
## Proportion of Variance 0.1213 0.07795 0.06867 0.06365 0.05922 0.05064 0.04468
```

```
## Cumulative Proportion  0.1213 0.19926 0.26793 0.33157 0.39080 0.44144 0.48613
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.19279 1.15128 1.06941 1.01265 0.99753 0.96701 0.96097
## Proportion of Variance 0.03845 0.03582 0.03091 0.02772 0.02689 0.02527 0.02496
## Cumulative Proportion  0.52458 0.56040 0.59131 0.61902 0.64592 0.67119 0.69615
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.94183 0.93296 0.9164 0.90195 0.85845 0.83461 0.80496
## Proportion of Variance 0.02397 0.02352 0.0227 0.02199 0.01992 0.01883 0.01751
## Cumulative Proportion  0.72012 0.74365 0.7663 0.78833 0.80825 0.82707 0.84459
##                           PC22   PC23   PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.75300 0.6963 0.6856 0.65291 0.64875 0.63743 0.63625
## Proportion of Variance 0.01532 0.0131 0.0127 0.01152 0.01137 0.01098 0.01094
## Cumulative Proportion  0.85991 0.8730 0.8857 0.89724 0.90861 0.91960 0.93054
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.61513 0.60162 0.59424 0.58682 0.54965 0.48442 0.47576
## Proportion of Variance 0.01023 0.00978 0.00954 0.00931 0.00817 0.00634 0.00612
## Cumulative Proportion  0.94076 0.95055 0.96009 0.96940 0.97756 0.98390 0.99002
##                           PC36   PC37
## Standard deviation     0.43757 0.4216
## Proportion of Variance 0.00517 0.0048
## Cumulative Proportion  0.99520 1.0000
```

The first component accounts for over 12% of variation in the data, which is still significant considering that there are over 37 components to account for. We decided to take a closer look at the first 8 components because together they account for over 50% of the variation.

```
##                     PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
## numid              0.00  0.00  0.01  0.00  0.01 -0.01  0.01 -0.01
## chatter           -0.13  0.20  0.07  0.11 -0.19  0.46 -0.11  0.07
## current_events    -0.10  0.06  0.05  0.03 -0.06  0.14  0.04 -0.05
## travel            -0.12  0.04  0.42 -0.15 -0.01 -0.16  0.08  0.31
## photo_sharing     -0.18  0.30 -0.01  0.15 -0.23  0.21 -0.13  0.02
## uncategorized     -0.09  0.15 -0.03  0.02  0.06 -0.04  0.19 -0.05
## tv_film           -0.10  0.08  0.09  0.09  0.21  0.06  0.50 -0.22
## sports_fandom     -0.29 -0.32 -0.05  0.06 -0.03  0.01 -0.07 -0.11
## politics          -0.13  0.01  0.49 -0.20 -0.06 -0.13 -0.07  0.01
## food              -0.30 -0.24 -0.11 -0.07  0.07  0.02  0.04  0.09
## family            -0.24 -0.20 -0.05  0.07 -0.01  0.05 -0.10 -0.02
## home_and_garden   -0.12  0.05  0.02 -0.01  0.04  0.04  0.09 -0.10
## music             -0.12  0.14 -0.01  0.08  0.07 -0.01  0.15 -0.09
## news              -0.13 -0.04  0.34 -0.18 -0.03 -0.09 -0.14 -0.46
## online_gaming     -0.07  0.08  0.06  0.22  0.48 -0.01 -0.29  0.06
## shopping          -0.13  0.21  0.05  0.10 -0.20  0.43 -0.09  0.03
## health_nutrition  -0.12  0.15 -0.23 -0.46  0.17  0.08 -0.04  0.04
## college_uni       -0.09  0.12  0.09  0.26  0.49  0.00 -0.19  0.04
## sports_playing    -0.13  0.11  0.04  0.18  0.37 -0.03 -0.22  0.06
## cooking           -0.19  0.31 -0.19  0.01 -0.12 -0.36 -0.06 -0.05
## eco               -0.15  0.09 -0.03 -0.12  0.02  0.18  0.00  0.04
## computers         -0.14  0.04  0.37 -0.14 -0.06 -0.14 -0.01  0.36
## business          -0.14  0.10  0.11  0.01 -0.05  0.07  0.09  0.14
## outdoors          -0.14  0.11 -0.14 -0.41  0.15  0.04 -0.06 -0.08
## crafts            -0.19 -0.02  0.00  0.02  0.04  0.08  0.24  0.03
## automotive        -0.13 -0.03  0.19 -0.04 -0.06  0.06 -0.24 -0.59
## art               -0.10  0.06  0.05  0.06  0.16  0.03  0.49 -0.16
## religion          -0.30 -0.32 -0.09  0.07 -0.02 -0.03  0.02  0.10
```

```
## beauty           -0.20  0.21 -0.15  0.15 -0.19 -0.37 -0.02 -0.06
## parenting        -0.29 -0.30 -0.09  0.05 -0.04 -0.01 -0.04  0.06
## dating           -0.11  0.07  0.03 -0.03 -0.01  0.00  0.03  0.17
## school           -0.28 -0.20 -0.08  0.09 -0.09  0.01  0.02  0.08
## personal_fitness -0.14  0.14 -0.22 -0.44  0.16  0.09 -0.04  0.04
## fashion          -0.18  0.28 -0.14  0.14 -0.17 -0.36 -0.03 -0.02
## small_business   -0.12  0.09  0.10  0.08  0.03  0.05  0.21  0.01
## spam             -0.01  0.00  0.01 -0.02  0.02  0.01  0.07  0.01
## adult            -0.03 -0.01  0.00 -0.02  0.01  0.02  0.07  0.00
```

We have now reached a point where our PCA no longer becomes interpretable. Take a look at PC4, for example. The coefficients with largest magnitude are health_nutrition, outdoors, and personal_fitness. They all have negative signs associated with them. Does it make sense that there's a significant population of the NutrientH20 followers that do not tweet about the aforementioned topics? Maybe we'll have better luck running a k-means clustering analysis.

We will use the gap statistic method and the associated elbow plot to determine our K.

```
# Using the gap statistic, we will identify how many K clusters to make
twit_gap = clusGap(na.omit(smz), FUN=kmeans, nstart=2, K.max=20, B=20)
plot(twit_gap, main="Gap Statistic Elbow Plot")
```

K = 11 based on this analysis, which means we will have 11 clusters. Let's run k-means (specifically k-means++) and then summarize the characteristics of each cluster. These will be our market segments. For the sake of space, we did not include the elbow plot in the output file because 10+ pages of "Warning: did not converge in 10 iterations" would be spammed onto the report.

```
##    clustall chatter     current_events travel      photo_sharing uncategorized
## 1   1        0.07205880  0.27684711     0.28872703 -0.09071828    0.11259854
## 2   2       -0.17184985 -0.02227120    -0.15924764 -0.12785633    0.14860478
## 3   3       -0.15681715  0.09697331    -0.10453108 -0.09474573   -0.12090010
## 4   4       -0.10074138 -0.08996728    -0.03151808 -0.01665036   -0.04036475
## 5   5       -0.06213139  0.16622169    -0.05308555  1.22964152    0.51034347
## 6   6       -0.07757068  0.07409674    -0.18475717 -0.21709746   -0.10007124
## 7   7        1.00582902  0.08086089    -0.05072419 -0.01920094    0.75613012
## 8   8       -0.13109160  0.33268481     0.22135146 -0.08508261    0.67765190
## 9   9       -0.37479524 -0.20567172    -0.21922126 -0.42260052   -0.18116914
## 10 10        1.50914339  0.36348654    -0.21124821  1.20204950   -0.01633342
## 11 11       -0.09398965  0.11141712     3.29662325 -0.11846913   -0.09301041
##    tv_film     sports_fandom politics    food        family
## 1  -0.11619101  0.1406567     0.15052740  0.04049422 -0.05999555
## 2  -0.14993850 -0.1935502    -0.20658489  0.46157676 -0.08790054
## 3  -0.09868784  2.1225673    -0.22430813  1.87801214  1.54671749
## 4   0.10182545 -0.1348585    -0.17556194 -0.09402141  0.20416883
## 5  -0.14155956 -0.2135820    -0.13328994 -0.20581519  0.02775310
## 6  -0.01109243  0.6735642     1.22753623 -0.15479446  0.23056840
## 7  -0.07966224 -0.1436920    -0.15304279 -0.14548260 -0.11646894
## 8   2.74377902 -0.1157061    -0.08671196  0.14837353 -0.10325113
## 9  -0.22210914 -0.3205036    -0.30012581 -0.35837060 -0.30234015
## 10 -0.13792348 -0.1952677    -0.13323702 -0.30432422 -0.03632034
## 11 -0.05994830 -0.2060808     3.13800828  0.16692765 -0.08573765
##    home_and_garden music        news          online_gaming shopping
## 1   0.23510191      0.014182641 -0.0006901999  0.08935906   -0.23782264
## 2   0.13189086     -0.007926127 -0.0693130514 -0.11337913   -0.07318340
## 3   0.16005496      0.034845690 -0.1073123737 -0.07599581   -0.01969005
## 4   0.07559319     -0.046742341 -0.1883744546  3.62504401   -0.14515721
```

```
## 5   0.12275302      0.552025123 -0.0896988537 -0.02522995    0.19818994
## 6   0.15441195     -0.084744503  2.6684320541 -0.12158476   -0.18526890
## 7   0.56667470     -0.033921188 -0.1321873380 -0.06811514   -0.09842075
## 8   0.31378507      0.988839687  0.0062555939 -0.17735805    0.01701595
## 9  -0.20254983     -0.230312042 -0.3094135434 -0.23268005   -0.39555688
## 10  0.04623985      0.156492505 -0.2690816185 -0.16229063    1.51647049
## 11  0.05701464     -0.036766273  1.1553429038 -0.16991614   -0.07664518
##    health_nutrition college_uni sports_playing cooking     eco
## 1   0.05086059       0.12732753 -0.11129036    -0.05898219  0.447999475
## 2   2.23271700      -0.21723844 -0.03552022     0.41241504  0.560938155
## 3  -0.14796034      -0.12728883  0.09986750    -0.09592143  0.179980862
## 4  -0.18192529       3.31074191  2.16306002    -0.12383210 -0.062788118
## 5  -0.06485575      -0.01707877  0.18369563     2.83814877 -0.007660224
## 6  -0.24261383      -0.19454633 -0.09431724    -0.23258480 -0.104846226
## 7  -0.09291457      -0.05088371  0.29590176    -0.13774308  0.131339840
## 8  -0.15949644       0.35310570  0.13312316    -0.13773569  0.090945563
## 9  -0.31629874      -0.25371771 -0.26414199    -0.32597937 -0.275070348
## 10 -0.21452397      -0.10795936 -0.09082811    -0.22232502  0.312528619
## 11 -0.16895848      -0.04273982  0.04112939    -0.18478229  0.148774765
##    computers    business    outdoors    crafts      automotive   art
## 1   0.297532905 -0.34600901  0.29780310  0.21793373  0.124535640  0.331675372
## 2  -0.084056248  0.03874142  1.74411261  0.07268689 -0.173715971 -0.081170068
## 3   0.092386853  0.10090702 -0.08240533  0.69018210  0.118930866 -0.027539252
## 4  -0.078746547 -0.09933284 -0.13482855  0.03174133  0.071799074  0.276744741
## 5   0.052960658  0.21270142  0.02775562  0.07737173  0.007230587  0.007586915
## 6  -0.190381036 -0.12077228  0.30838881 -0.16666750  2.601885558 -0.159669425
## 7   0.004718573  0.42354263  0.06035697  0.39076157 -0.188632571 -0.020922250
## 8  -0.148265614  0.32555266 -0.09658814  0.74609787 -0.220773544  2.649127214
## 9  -0.253943904 -0.24351309 -0.32791124 -0.29222895 -0.307706536 -0.238866916
## 10 -0.036656325  0.31191686 -0.25914428  0.01952022  0.102400580 -0.213147217
## 11  2.917712151  0.56772911 -0.03966612  0.20365713 -0.131551892 -0.156745281
##    religion    beauty     parenting   dating       school
## 1   0.12070179 -0.1007020  0.18658414 -0.009528244  0.092448236
## 2  -0.16178364 -0.2121507 -0.09307818  0.050909102 -0.198952336
## 3   2.32464576  0.3223732  2.20141590 -0.100691182  1.676294758
## 4  -0.19280175 -0.2337056 -0.13256181 -0.035035380 -0.233774761
## 5  -0.12459055  2.6421544 -0.06230650 -0.045175440  0.128963124
## 6  -0.18038539 -0.1758057  0.03790115 -0.083900294 -0.001000724
## 7   0.01478152  0.2608590  0.08932623  4.795290740  1.258301264
## 8   0.01176547  0.0111083 -0.19778695 -0.145185951 -0.042291982
## 9  -0.29879097 -0.2735124 -0.32271615 -0.215342053 -0.327569905
## 10 -0.28017728 -0.2279146 -0.20820178 -0.153700007 -0.039685216
## 11  0.12374692 -0.1885232  0.01727783  0.235832395 -0.118556959
##    personal_fitness fashion      small_business spam        adult         count
## 1   0.12183236      -0.020449872  0.31428826    12.41886450  3.750222155   49
## 2   2.15850413      -0.129213135 -0.13913048    -0.07768727  0.022104412  751
## 3  -0.09660409       0.001873498  0.08763194    -0.07768727 -0.004535327  653
## 4  -0.18012600      -0.080086079  0.11324834    -0.07768727 -0.018996040  347
## 5  -0.04641446       2.729474485  0.16452116    -0.07768727  0.012070866  468
## 6  -0.22642717      -0.226440874 -0.15874612    -0.07768727 -0.107959548  424
## 7  -0.04637783       0.819438154  0.36483196    -0.07768727 -0.049018237  194
## 8  -0.15533283      -0.046502945  0.81129898    -0.07768727 -0.038144166  407
## 9  -0.33528922      -0.300456377 -0.20992523    -0.07768727 -0.012033089 3287
## 10 -0.16392786      -0.153819787  0.14583909    -0.07768727 -0.037254089  959
```

```
## 11 -0.15210745      -0.179902514  0.38503485    -0.07768727 -0.146847738  343
##    pct
## 1   0.6216696
## 2   9.5280386
## 3   8.2846993
## 4   4.4024359
## 5   5.9375793
## 6   5.3793453
## 7   2.4613042
## 8   5.1636640
## 9  41.7026135
## 10 12.1669627
## 11  4.3516874
```

Let's discuss what we have here. Our findings show that we have the following market segments: 1) "Lurkers" (cluster 9): 41.7%, fall slightly below average in the number of tweets in all categories make up 40%+ of the Twitter users, they are the silent and quiet majority; 2) "Influencers" (cluster 8): 12.2%, above average chatter (by 1.5 stdevs), shopping (by 1.5 stdevs), and photo-sharing (by 1.2 stdevs), they most likely partner up with brands and get paid to promote through social media outlets; 3) "Gym Rats" (cluster 7): 9.5%, significantly above average health/nutrition (2.2 stdevs) and personal fitness (2.2 stdevs) related content, and above average outdoors (1.8 stdevs) related content, these people like to frequent the gym; 4) "Media/Artists" (cluster 6): 5.2%, significantly above average tv/film (2.7 stdevs) and art (2.6 stdevs) related content, these are what you would call "creatives."; 5) "College Students" (cluster 10): 4.4%, significantly above average online-gaming (3.6 stdevs), college/uni (3.3 stdevs), and sports_playing (2.2 stdevs) related content, these are the hardcore gamers that likely spend more time playing Smash at the local video game store tournament than writing up a detailed statistical report for their stats class; 6) "Spam" (cluster 11): 0.6%, significantly above average spam (12.4 stdevs) and adult (3.8 stdevs) related content, as mentioned in the problem description, these are the spammers/trolls.

One of the major tenets of marketing is to define a target and serve their needs specifically without regard to anyone outside of that target demographic. Our research shows that NutrientH20 could tailor its messaging to fit the needs of the influencers. This could include posting aesthetic photos or making "challenge" posts where they call upon their general follower base to post and/or hashtag their products in creative ways. Then, NutrientH20 can reach out to the users who did a fine job and offer them an influencer partnership. This would increase engagement and expand marketing in the long run.