Exercise 3

Max Kutschinski

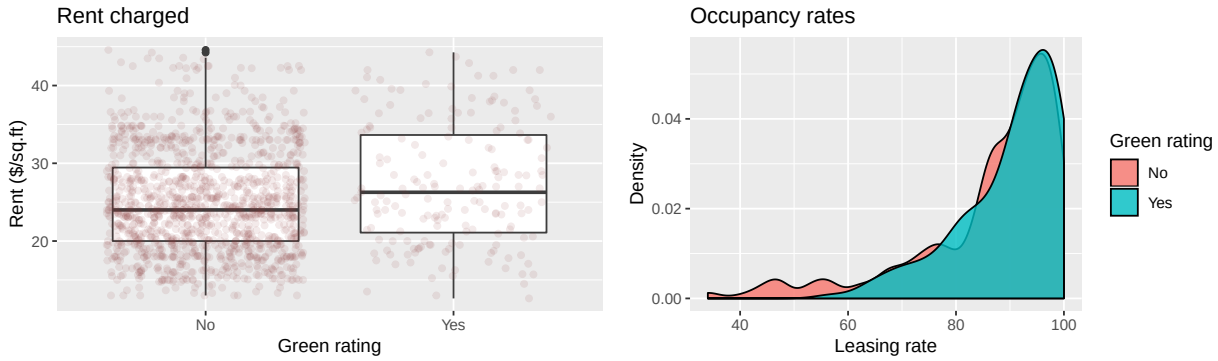
4/21/2020

Predictive model building

Overview

The dataset contains data on 7,894 commercial rental properties from across the United States. Of these, 685 properties have been awarded either LEED or EnergyStar certification as a green building. Each of these 685 buildings was matched to a cluster of nearby commercial buildings in the CoStar database, where each small cluster contains one green-certified building, and all non-rated buildings within a quarter-mile radius of the certified building. On average, each of the 685 clusters contains roughly 12 buildings, for a total of 7,894 data points. Some examples of features in the dataset are the building's age, electricity costs, number of stories, and average rent within the geographic region.

Below are two plots comparing the rent charged, as well as the occupancy rates between green and non-green buildings.



These two plots suggest that green rated bulidings not only benefit from charging more rent, but are also less likely to be vacant. Thus, there clearly seems to be an incentive for "going green". The goal of this exercise is to find a model that does a good job of predicting price and to use this model to quantify the average change in rental income associated with green certification.

Methods

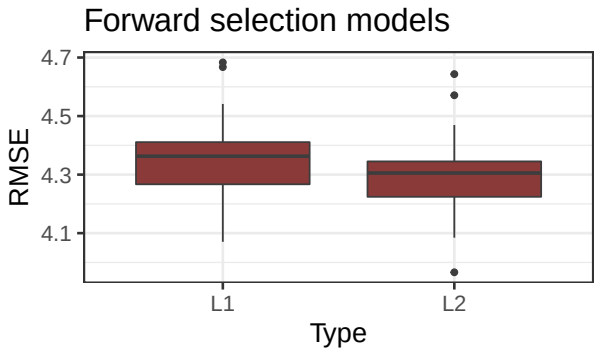
This problem will be approached using four different models. The baseline (null) model will be a linear model that uses only main effects, which are chosen through a forward selection alogorithm. This model will then be compared to a similar linear model that runs on the same algorithm, with the exception of accounting for interactions.

Since a forward selection algorithm that includes interactions is likely to produce a model with a lot of features, there exists a risk of overfitting. Thus, lasso regression techniques will be used with the goal of reducing model complexity as well as simplifying feature selection. This will produce a simpler third model for comparison. Lastly, the problem will be approached from a standpoint of KNN regression. This model will use the same features as the null model for simplicity.

Since there exists random variation due to the particular choice of data points that end up in the train/test splits, models will be compared using their average-out-of-sample RMSE over multiple different train/test splits.

Results

The figure below captures the RMSE results for each linear model using forward selection. L1 represents the model that only includes main effects, and L2 represents the model that includes main effects, as well as interactions.

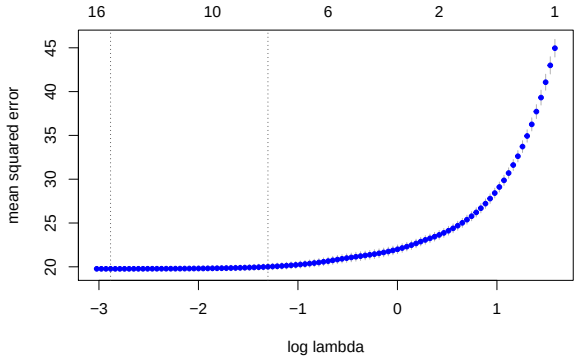


As expected, L2 performs better than the null model. Note that the RMSE is slightly lower.

```
## L1 RMSE: 4.359
```

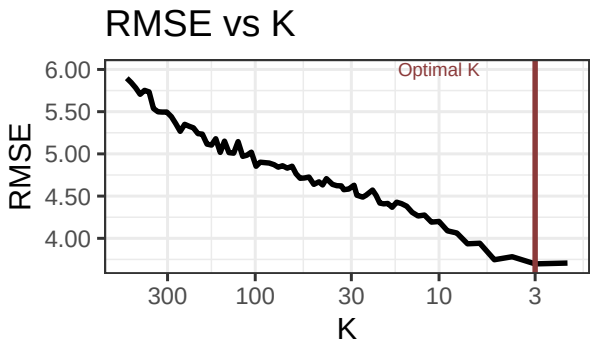
```
## L2 RMSE: 4.301
```

As mentioned above, L2 is a fairly complex model that has around 77 coefficients. Thus, the third model features lasso regularization where out-of-sample deviance is diplayed below as a function of log lamda. Note that this model uses 10 fold cross validation.



Overall, using lasso regularization results in a RMSE value of 4.159.

The last model that was used to predict price is built on KNN regression. Below is a plot of RMSE vs K, which resulted from running KNN regression on the features included in the null model.



Using the optimal K value over multiple train/test splits yields a RMSE of 4.354.

.....

To quantify the average change in rental income per square foot associated with green certification, it is sufficient to look at the "green_rating" coefficient of the best performing model. In this case, the lasso model performed the best by having the lowest RMSE. The estimated coefficients of this model are displayed below.

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                1
## (Intercept)        1.039777e+01
## size               .
## empl_gr            .
## leasing_rate       .
## stories            5.652947e-03
## age               -4.900504e-03
## renovated         -9.549770e-01
## class_a            1.641173e-02
## class_b           -9.051800e-14
## green_rating       1.566784e+00
## net               -1.194650e-01
## amenities          1.520095e-01
## cd_total_07       -1.470150e-03
## hd_total07         .
## total_dd_07       -2.116838e-04
## Precipitation      .
## Gas_Costs         -9.163972e+01
## Electricity_Costs  .
## cluster_rent       7.977577e-01
```

Thus, the average change in rental income per square foot associated with green certification, holding other features of the building constant, is around 1.57

Conclusion

Overall, it seems like the lasso regularization model serves as a good model to predict a building's rent, since it outperformed all the other models. Furthermore it appears that green certified buildings charge higher

rent on average and are thus potentially a good investment opportunity.

What causes what?

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? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)

This is a typical case of correlation versus causation. We can't simply assume that correlation implies causation, and therefore need to consider that there might be other reasons for crime levels to differ besides the number of cops in the streets. In addition, high crime areas naturally have a higher number of cops in the streets. As mentioned in the podcast, a solution would be to compare cities where the number of cops is low to cities that have a lot of cops for reasons unrelated to crime (such as terror threats). Furthermore, it is important to control for confounding variables such as poverty and income levels in order to obtain meaningful results.

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.

EFFECT OF POLICE ON CRIME

TABLE 2

TOTAL DAILY CRIME DECREASES ON HIGH-ALERT DAYS

	(1)	(2)
High Alert	−7.316*	−6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

Figure 1: The dependent variable is the daily total number of crimes in D.C. This table present the estimated coefficients and their standard errors in parenthesis. The first column refers to a model where the only variable used in the High Alert dummy whereas the model in column (2) controls form the METRO ridership. * refers to a significant coeficient at the 5% level, ** at the 1% level.

In order to isolate this effect, UPenn researchers had to find an area that gets a lot of police for reasons unrelated to crime. They determined that Washington D.C. would serve as a great example due to their terrorism alert system. Since Washington D.C. is likely to be a terrorism target, additional police units are dispatched when the threat level rises. To ensure that high alert days did not yield lower tourist traffic, which could mean fewer potential victims, the researchers kept track of these numbers by measuring METRO ridership. The results of this study indicate that on high terror days, crime levels dropped with additional police in the area, while METRO ridership was unchanged, suggesting that there seems to be an inverse relationship between the number of cops and crime activity.

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

The researchers controlled for METRO ridership because they were considering whether tourisms were less likely to visit Washington or go out and about due to a high terror alert, which could have a negative effect on the number of victims. When the number of victims go down, there are less opportunities for crime to happen, resulting in lower expected crime levels. The results indicate that METRO ridership did not diminish on high alert days, suggesting that the number of victims was largely unchanged and that this did not prove itself as a confounding variable.

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?

TABLE 4

REDUCTION IN CRIME ON HIGH-ALERT DAYS: CONCENTRATION ON THE NATIONAL MALL

	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert × District 1	−2.621**	−2.621*	−2.621*
	(.044)	(1.19)	(1.225)
High Alert × Other Districts	−.571	−.571	−.571
	(.455)	(.366)	(.364)
Log(midday ridership)	2.477*	2.477**	2.477**
	(.364)	(.522)	(.527)
Constant	−11.058**	−11.058	−11.058˙
	(4.211)	(5.87)	(5.923)

Figure 2: The dependent variable is the daily total number of crimes in D.C. District 1 refers to a dummy variable associated with crime incidents in the first police district area. This table present the estimated coefficients and their standard errors in parenthesis.* refers to a significant coeficient at the 5% level, ** at the 1% level.

The first column of the table above summarizes a linear regression of the daily total number of crimes on crime incidents in district 1, crime incidents in other districts, and METRO ridership. Overall, this model suggests a reduction in crim on high alert days. The coefficient on the district 1 feature is significant at the 1% level and larger in magnitude than the feature capturing crime incidents in other districts. Thus, it seems like crime incidents have dropped substantially the first police district area. Furthermore, the model reaffirms previous results that suggested METRO ridership did not diminish on high alert days.

Clustering and PCA

Introduction

This is an exercise about PCA and Clustering. The goal is two use both techniques on a set of wine data and to see which dimensionality reduction technique makes more sense.

Data and Methods

The dataset contains information on 11 chemical properties of 6500 different bottles of vinho verde wine from northern Portugal. In addition, two other variables about each wine are recorded:

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

The 11 chemical properties are:

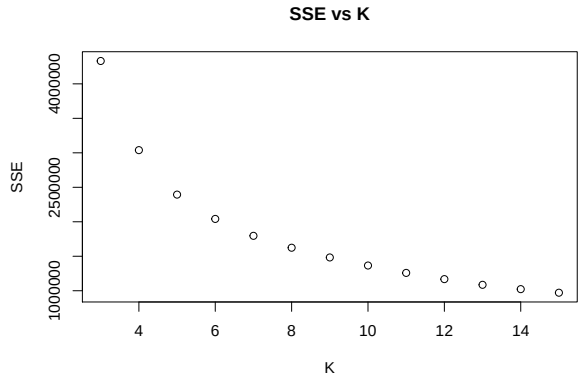
1) fixed acidity
2) volatile acidity
3) citric acid
4) residual sugar
5) chlorides
6) free sulfur dioxide
7) total sulfur dioxide
8) density
9) pH
10) sulphates
11) alcohol

The clustering algorithm that will be used is K-means++, since it ensures smarter initialization of the centroids and improves the quality of the clustering compared to regular K-means.

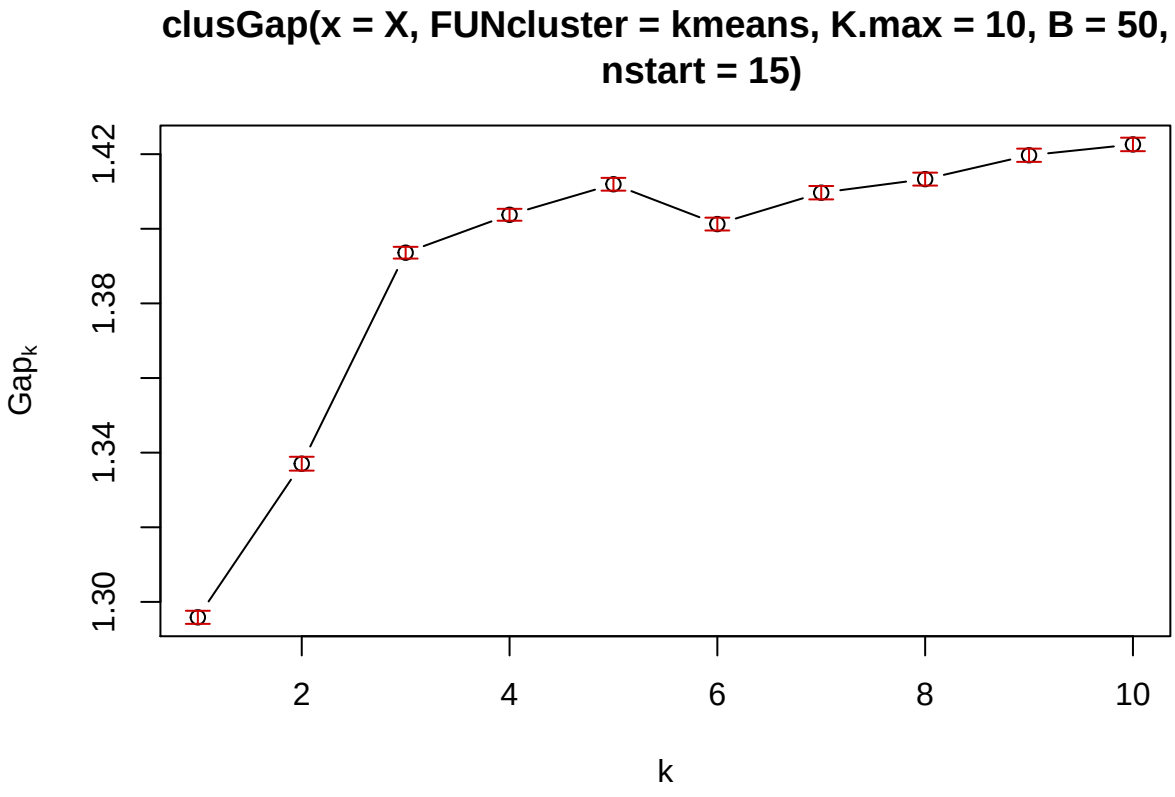
Results

K-means++

The following figure represents a plot of SEE versus K. This plot is used to see if there is an "elbow" visibile, which indicates the optimal value of K.



It looks like the optimal K could be anywhere between 4 to 6. To verify this result, the gap statistic is used.

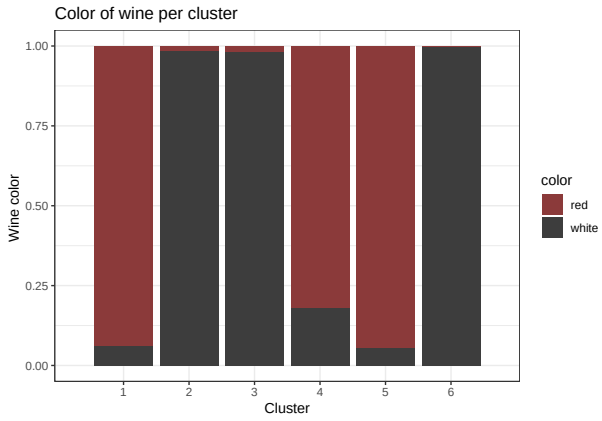


Here it becomes clear that k=6 is the "optimal" value for the number of clusters.

Below is a summary of each chemical property per cluster.

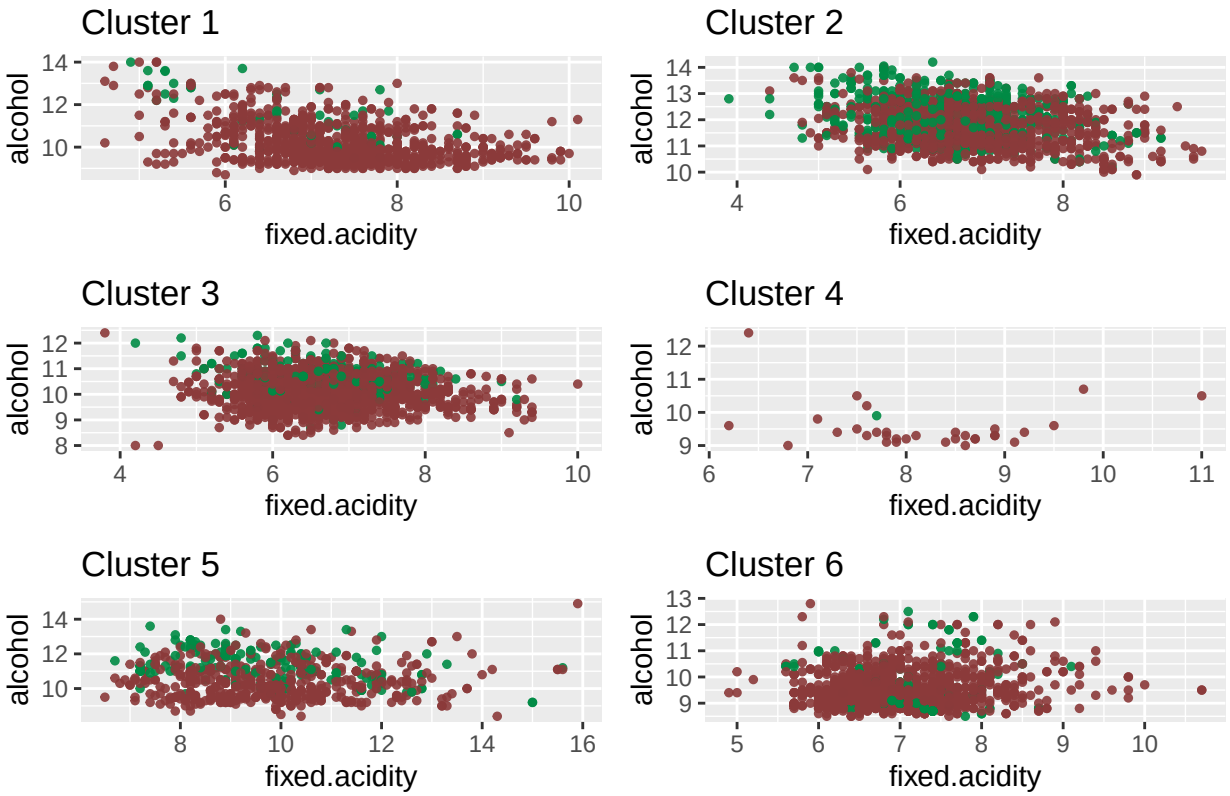
```
##           fixed acity volatile acidity citric acid residual sugar chlorides
## Cluster 1        6.73           136.19        0.41           1.00      0.36
## Cluster 2        0.26             0.99        0.45           3.21     12.20
## Cluster 3        0.32             3.26        2.78           0.72      0.05
## Cluster 4        3.84             0.51        0.08          10.61     46.84
## Cluster 5        0.05            10.15       14.68           7.01    171.48
## Cluster 6       32.08             9.79       45.30           0.28      1.00
##           free sulfur dioxide total sulfur oxide density    pH sulphates
## Cluster 1                3.14               3.44    0.47  0.08     10.23
## Cluster 2                0.49               0.04   11.89 16.40      8.18
## Cluster 3                9.50              28.21    7.29 50.41      0.51
## Cluster 4                6.81             109.15    0.62  1.00      0.50
## Cluster 5                0.28               0.99    0.13  3.38      3.22
## Cluster 6                0.33               3.17    2.47  0.59      0.36
##           alcohol
## Cluster 1   18.33
## Cluster 2   76.61
## Cluster 3    1.00
## Cluster 4    3.08
## Cluster 5    1.05
## Cluster 6    9.56
```

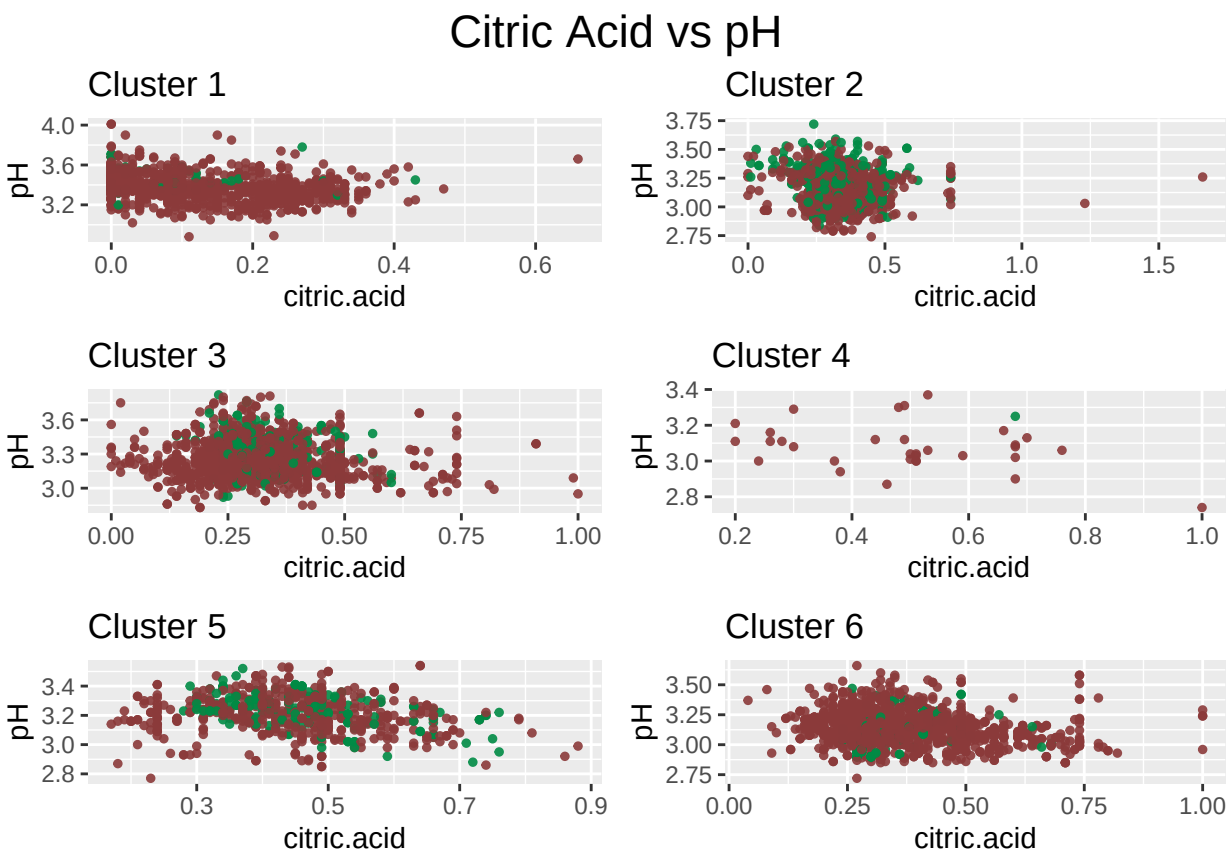
The plot below shows the percentage of red and white wine in each cluster. This figure indicates that clusters 1,4, and 5 contain mostly red wine, while clusters 2,3, and 6 contain mostly white wine.



The next figure can be used to assess whether the clustering algorithm was successful in separating wine quality between clusters. Displayed are plots of different chemical relationships in each cluster. To make things more interpretable, high quality wine will be defined as having a quality level of 7 or above, and low quality wine will be defined as everything below 7. The plots below highlight high quality wine with the color green and low quality wine with the color red.

Fixed acidity vs Alcohol





From these two different chemical relationships, it appears as if there is no strong separation between qualities among clusters. However, there seems to be a weak separation that indicates that clusters 2 and 5 have a higher proportion of high quality wines, while clusters 1,3, 4 and 6 contain more low quality wines.

Principal Component Analysis (PCA)

PCA was used on the scaled dataset. A summary of the results is shown below.

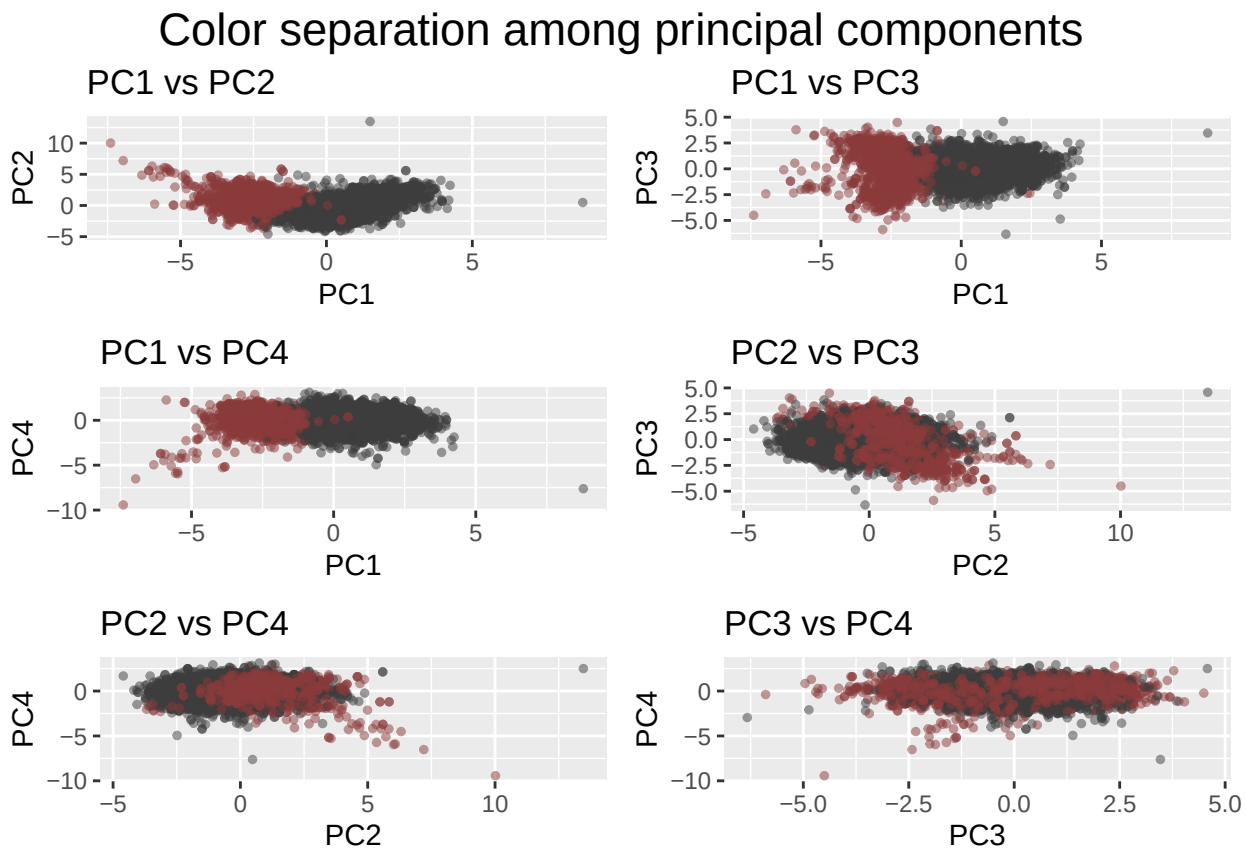
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 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
```

These summary statistics indicate that the first four principle components account for about 75% percent of the explainability of total features. Furthermore, it is apparent that the proportion of variance that is added per additional component decreases. Thus, the first four principal components are the ones that will be used to reduce dimensionality while retaining variation in the dataset. The first four principal components are described below.

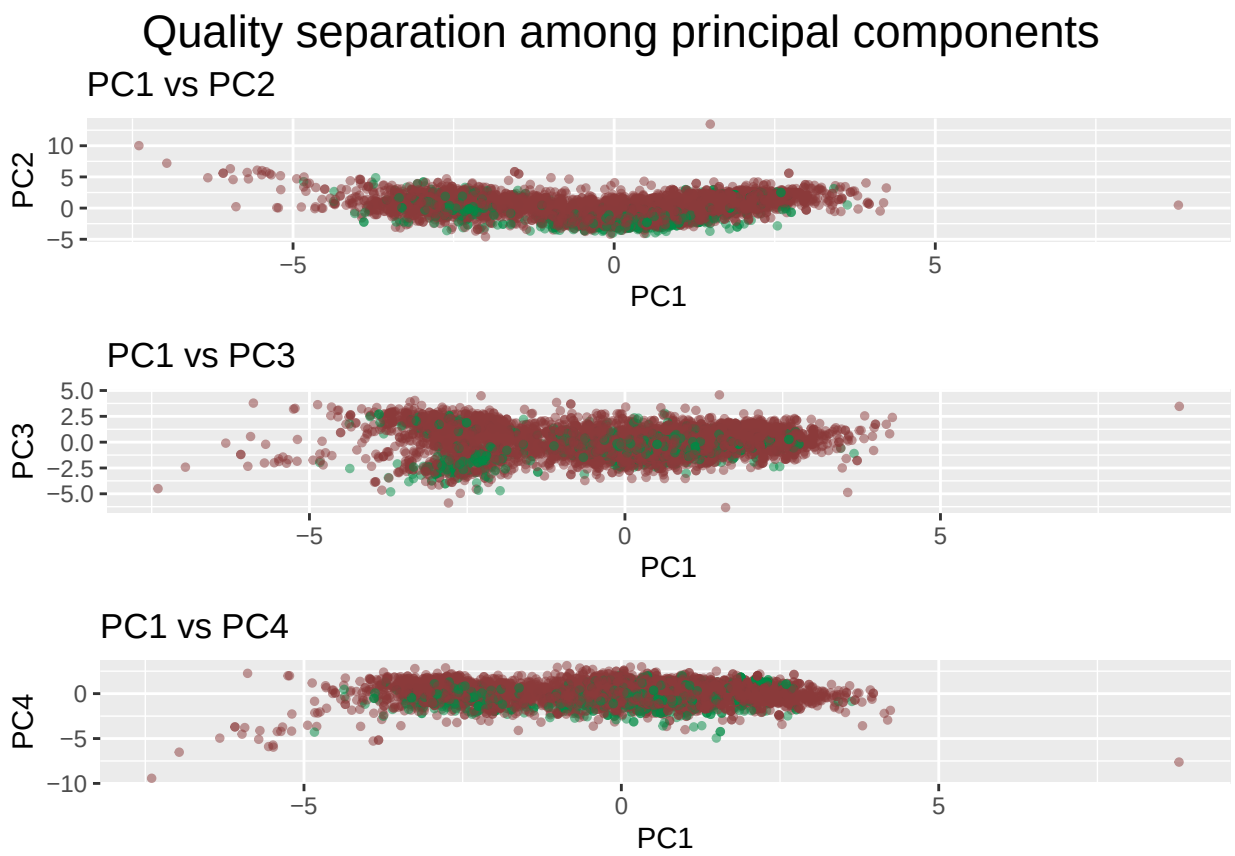
```
##                         PC1    PC2    PC3    PC4
```

```
## 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
```

The following plots demonstrate how well these principal components can cluster the dataset by wine color. Here the color red indicates red wine and gray represents white wine.



Based on these plots, any combination of PC1 does a good job at separating red wines from white wines. How about quality? Again, green is used to indicate a wine quality of 7 or above, and red is representative of anything below that level.



Based on the strongest principal components, it doesn't seem like there is a lot of separation based on quality.

Conclusion

Overall, the K-means++ clustering algorithm did a good job at serparating wines based on color, but not that great in separating them by quality. PCA did a similarly good, if not better job, when it came to clustering wine based on color, but also failed to separate wine based on quality. I personally prefered PCA in this instance due to its understandable and effective way of dimensionality reduction. PCA enable me to reduce 11 features into 4 principal components while retaining almost 75% of the variability of the dataset.

Market Segmentation

Introduction

The goal for a large consumer brand is to understand its social-media audience a little bit better, so that it could hone its messaging a little more sharply.

Data and Methods

The data from this dataset was collected in the course of a market-research study using followers of the Twitter account of a large consumer brand. It contains every Twitter post by a random sample of followers over a seven-day period in June 2014. Each feature of the dataset represents one of 36 pre-specified interest categories that a follower's post might fall in (e.g politics, sports, family, etc.). Two interests of note here are

"spam" (i.e. unsolicited advertising) and "adult" (posts that are pornographic or otherwise explicit). There are a lot of spam and pornography "bots" on Twitter; while these have been filtered out of the data set to some extent, there will certainly be some that slip through. There's also an "uncategorized" label, which is there to capture posts that don't fit at all into any of the listed interest categories.

Since the goal is to gain insights about certain market segments, PCA will be used to reduce this dataset to a more manageable scale and to gain potential insights.

Results

Below is a summary of the principal components that are obtained from the datset after excluding irrelevant variables such as spam, adult, and personal ID.

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027 1.36885 1.28577
## Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205 0.04592
## Cumulative Proportion  0.1247 0.20479 0.27536 0.34077 0.40164 0.45369 0.49961
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.19277 1.15127 1.06930 1.00566 0.96785 0.96131 0.94405
## Proportion of Variance 0.03952 0.03682 0.03176 0.02809 0.02602 0.02567 0.02476
## Cumulative Proportion  0.53913 0.57595 0.60771 0.63580 0.66182 0.68749 0.71225
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.93297 0.91698 0.9020 0.85869 0.83466 0.80544 0.75311
## Proportion of Variance 0.02418 0.02336 0.0226 0.02048 0.01935 0.01802 0.01575
## Cumulative Proportion  0.73643 0.75979 0.7824 0.80287 0.82222 0.84024 0.85599
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.69632 0.68558 0.65317 0.64881 0.63756 0.63626 0.61513
## Proportion of Variance 0.01347 0.01306 0.01185 0.01169 0.01129 0.01125 0.01051
## Cumulative Proportion  0.86946 0.88252 0.89437 0.90606 0.91735 0.92860 0.93911
##                           PC29    PC30    PC31   PC32    PC33    PC34    PC35
## Standard deviation     0.60167 0.59424 0.58683 0.5498 0.48442 0.47576 0.43757
## Proportion of Variance 0.01006 0.00981 0.00957 0.0084 0.00652 0.00629 0.00532
## Cumulative Proportion  0.94917 0.95898 0.96854 0.9769 0.98346 0.98974 0.99506
##                           PC36
## Standard deviation     0.42165
## Proportion of Variance 0.00494
## Cumulative Proportion  1.00000
```

Here, the first 7 principal components explain about 50% of the variability in the dataset, which is a good amount considering how many features there are. Below are the summarized features for each of these 7 principal components.

```
##                     PC1   PC2   PC3   PC4   PC5   PC6   PC7
## chatter           -0.13  0.20 -0.07  0.11 -0.19  0.46 -0.11
## current_events    -0.10  0.06 -0.05  0.03 -0.06  0.14  0.04
## travel            -0.12  0.04 -0.42 -0.15 -0.01 -0.16  0.09
## photo_sharing     -0.18  0.30  0.01  0.15 -0.23  0.21 -0.13
## uncategorized     -0.09  0.15  0.03  0.02  0.06 -0.04  0.19
## tv_film           -0.10  0.08 -0.09  0.09  0.21  0.06  0.50
## sports_fandom     -0.29 -0.32  0.05  0.06 -0.03  0.01 -0.07
## politics          -0.13  0.01 -0.49 -0.20 -0.06 -0.13 -0.07
## food              -0.30 -0.24  0.11 -0.07  0.07  0.02  0.04
```

```
## family           -0.24 -0.20  0.05  0.07 -0.01  0.05 -0.10
## home_and_garden  -0.12  0.05 -0.02 -0.01  0.04  0.04  0.09
## music            -0.12  0.14  0.01  0.08  0.07 -0.01  0.15
## news             -0.13 -0.04 -0.34 -0.18 -0.03 -0.09 -0.14
## online_gaming    -0.07  0.08 -0.06  0.22  0.48 -0.01 -0.29
## shopping         -0.13  0.21 -0.05  0.10 -0.20  0.43 -0.09
## health_nutrition -0.12  0.15  0.23 -0.46  0.17  0.08 -0.04
## college_uni      -0.09  0.12 -0.09  0.26  0.49  0.00 -0.19
## sports_playing   -0.13  0.11 -0.04  0.18  0.37 -0.03 -0.22
## cooking          -0.19  0.31  0.19  0.01 -0.12 -0.36 -0.06
## eco              -0.15  0.09  0.03 -0.12  0.02  0.18  0.00
## computers        -0.14  0.04 -0.37 -0.14 -0.06 -0.14 -0.01
## business         -0.14  0.10 -0.11  0.01 -0.05  0.07  0.09
## outdoors         -0.14  0.11  0.14 -0.41  0.15  0.04 -0.06
## crafts           -0.19 -0.02  0.00  0.02  0.04  0.08  0.24
## automotive       -0.13 -0.03 -0.19 -0.04 -0.06  0.06 -0.24
## art              -0.10  0.06 -0.05  0.06  0.16  0.03  0.49
## religion         -0.30 -0.32  0.09  0.07 -0.02 -0.03  0.02
## beauty           -0.20  0.21  0.15  0.15 -0.19 -0.37 -0.02
## parenting        -0.29 -0.30  0.09  0.05 -0.04 -0.01 -0.04
## dating           -0.11  0.07 -0.03 -0.03 -0.01  0.00  0.03
## school           -0.28 -0.20  0.08  0.09 -0.09  0.01  0.02
## personal_fitness -0.14  0.14  0.22 -0.44  0.16  0.09 -0.04
## fashion          -0.18  0.28  0.14  0.14 -0.17 -0.36 -0.03
## small_business   -0.12  0.09 -0.10  0.08  0.03  0.05  0.21
## spam             -0.01  0.00 -0.01 -0.02  0.02  0.01  0.07
## adult            -0.03 -0.01  0.00 -0.02  0.01  0.02  0.07
```

From the first principal component, the coefficients with the largest absolute magnitude are food, religion, sports fandom, and parenting. These are things that young adults would most likely talk about.

The most largest positive coefficients in PC2 are Photo sharing, cooking, and fashion, which definitely share the same target group namely young adults. Nowadays, the majority of the youth is obsessed with looking good for instagram and impressing their followers.

The largest negative coefficients in PC3 tell another interesting story. These coefficients are travel, politics, news, and computers. They all have to do with work or education and therefore suggest that a large part of the company's following is in the workforce and likes to know what is going on in the world.

The largest coefficients in absolute value in PC4 are personal fitness and outdoors. Clearly, this suggests that another large segment of the company's following is into fitness and health related activities.

In PC5, the coefficients that stand out are online gaming, college, and sports playing. In PC6, the coefficients that stand out are chatter and shopping, which is likely coming from a female audience. IN PC7, the coefficient that stands out more than any other is TV and film.

Conclusion

Overall, the company seems to reach a broad range of audiences. That being said, two marget segments that stand out from my analysis are health and beauty(food, sports, fashion, photo sharing), as well as intellectuality (politics, news, computers, education).