Final Project

Devansh Taori (drt2131)

5/7/2022

Introduction

Since its creation thousands of years ago, wine has often been referred to as the eighth wonder of the world. Connoisseurs, dabblers, and novices alike seem to hold deep appreciation for the grape-based drink, with wine standing as the second most popular alcoholic beverage in the US today (according to a Gallup poll). Hailing from numerous different regions around the world, wine is extremely versatile: it comes in a dark, rich red or light white. It can be dry or sweet, and it can be paired with meals ranging from steak to dessert.

Despite its popularity, very little is known about what makes wine taste good. Some experts attribute it to the amount of years it has aged in an oak barrel, but there is a high degree of variability within aged bottles as well. Wine that sells for 30 dollars/bottle can often match the quality of wine that sells for even 100 dollars/bottle. Given that this is a relatively overlooked field, I wanted to explore what objective physiochemical properties influence the quality of wine. I was curious to see whether there is a relationship between any of these properties, and if there is a significant difference between red and white populations as well.

I collected datasets from UC Irvine that include samples from two different populations of wine (red and white) in the Vinho Verde region of northern Portugal. Portugal is a top ten wine exporting country, and this wine accounts for roughly 15% of total Portuguese production. I'll explain the source of this data and its properties later in this report, but at a high level, these datasets have numeric columns representing different physiochemical properties (i.e. alcohol content, amount of residual sugar, amount of citric acid, etc...) as well as a column named "quality" that represents a 0-10 score of the wine. This column is the median of 3 experts' ratings of the wine, and it ranks 0 as the worst and 10 as the best.

In this report, I mine this dataset toward a couple different objectives: to analyze (1) what physiochemical properties of wine improve its quality, (2) whether there is a significant non-obvious relationship between permutations of properties, and (3) to what degree the different populations of red and white influence the quality score. I utilize a number of algorithms to do so, ranging from a simple correlation matrix (which calculates the correlation between each variable) to more complex clustering (which groups rows based on implicit relationships) and principal components analysis (which groups columns based on implicit relationships) to linear regression methods like ordinary least squares and LASSO. The results of the report are quite fascinating and provide useful insight into the nature of wine production and consumption.

Target Audience

This report is targeted specifically towards wine producers, such as Quinta da Aveleda (the largest exporting company in Portugal's Vinho Verde region) and Quinta do Ameal (a small 18th-century property that produces white wines in the same region). Similar to the production of coffee bean, the quality of production from wineries fluctuates on a year-to-year basis based on the quality of grape. Unfortunately, the quality of grape is dependent on factors that wineries often can't control, such as yearly rainfall, weather conditions, and soil quality (according to Sonoma State University Wine Business Institute). As such, the quality of wine produced swings dramatically every year. Wineries that one year might be considered top producers

can fall heavily in the rankings next year. This is bad for their business since it prevents consistent, stable revenues from production and tarnishes their long-term reputation.

This report develops insight into what makes wine higher quality based on scientific properties of the wine. While wineries may not be able to control the quality of grape, they can certainly control – among other things – the residual sugar content, the citric acid content, the chloride content of wines, and/or the balance or concentration of all 3. The value provided to wine producers is thus the ability to objectively measure which properties are worth including in higher concentrations and which properties do not matter as much. Since the insights from this report can assist in both streamlining the production process while allowing wineries to iteratively improve their formulas, there is value from a business standpoint. Consistently generating quality wine ensures repeat customers each year and improves the brand name of the winery.

Source of Data

The datasets were sourced from UC Irvine's Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine+quality). Although the data can be downloaded from here, the original source is Professor Paulo Cortez at the Department of Information Systems in the University of Minho, Portugal (https://pcortez.dsi.uminho.pt/). His research interests center around data mining, data science and forecasting, business analytics, and the like. He donated this dataset to UC Irvine in 2009, and it has since been used in various statistical papers.

Professor Cortez approached constructing the datasets in a comprehensive manner. Living in Portugal, he found it easiest to collect data on wines from the Vinho Verde region in northern Portugal. At the same time, due to privacy and logistical issues, he could only collect data on physiochemical properties (the input) and sensory quality (the output). There was no available data on wine brand and wine selling price, so information on that front is lacking. That said, he was quite intentional regarding the attributes he decided to include. The data was collected from May 2004 to February 2007 using samples tested at the official certification entity, CVRVV. CVRVV is an inter-professional organization with the goal of improving the quality and marketing of Vinho Verde. The data was recorded by a computerized system, which automates the process of wine sample testing from producer requests to laboratory and sensory analysis. Each entry denotes a given test, and the final database was exported into a single sheet. To avoid discarding examples, only the most common physicochemical tests were selected.

The datasets include wine from two distinct populations: red populations and white populations in the Vinho Verde region. Since the red and white tastes are quite different, Professor Cortez constructed these datasets separately (albeit with the same column names in each for standardization purposes). In fact, the red dataset contains 1,599 samples while the white dataset contains 4,898 samples. Both red and white wine are produced in quite distinct manners. Red wine requires the grape to be soaked in the seed and skin of the fruit, giving it not only its richer flavor but also different chemical properties that dissolve over time. White wine, on the other hand, is produced from an unsoaked grape, contributing to its generally lighter flavor. As a result, the physiochemical tests done on red versus white often yield significantly different results. For the purpose of this report, having these two distinct samples is valuable because it introduces data from a population that is vastly different, starting from the production process to the testing process. Certain properties that balance well in red wines and contribute to a more favorable quality might detract from the flavor in white wines; as such, analyzing the quality question using different varieties is useful. Therefore, the additional dataset provides much needed clarity and complexity to the overall data.

There are 11 physiochemical properties included, and they are described here: (1) fixed acidity – the amount of nonvolatile acid that does not evaporate readily; (2) volatile acidity – the amount of acetic acid, which has a vinegar taste; (3) citric acid – the amount of citric acid, which can add freshness and flavor; (4) residual sugar – the amount of sugar remaining after fermentation stops; (5) chlorides – the amount of salt; (6) free sulfur dioxide – the concentration of the free form of SO2, which prevents microbial growth and oxidation; (7) total sulfur dioxide – the amount of free and bound forms of SO2, which in higher concentrations, becomes evident in the nose and taste of wine; (8) density – how dense the wine is; (9) pH – how acidic or basic a wine is; (10) sulphates – the amount of wine additive that acts as an antimicrobial and antioxidant; (11) alcohol – the alcohol content.

There is also a 12th column, quality, which represents a score from 0 (very bad) to 10 (excellent). Each sample is evaluated by a minimum of three judges or sensory assessors (using blind tastes), who graded the wine. The final sensory score is given by the median of these evaluations.

Exploratory Analysis

I first began by loading in the data and conducting some exploratory analysis.

```
#reading in the two datasets
red <- read.csv("/Volumes/GoogleDrive/My Drive/4th Year/Semester 2/Applied Data Mining/Project - Final/
white <- read.csv("/Volumes/GoogleDrive/My Drive/4th Year/Semester 2/Applied Data Mining/Project - Final

#calculating the dimensions for each dataset
dim(red)
```

```
## [1] 1599   12
```

```
dim(white)
```

```
## [1] 4898   12
```

As mentioned above, there are many more instances in the white wine set as compared to the red wine set. Knowing the meticulous process that the CVRVV went through to conduct accurate tests of the wines, I didn't expect many data quality issues. However, I wanted to ensure that any NAs in the dataset were removed ahead of time.

```
#evaluating if there are NAs that need to be removed
sum(is.na(red))
```

```
## [1] 0
```

```
sum(is.na(white))
```

```
## [1] 0
```

Seeing as there are no missing values, I decided to generate summary statistics for each column to evaluate other data quality issues. Potential red flags include numbers that are nonsensical (i.e. negative alcohol content).

```
#generating summary statistics to evaluate data quality issues
summary(red)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
```

```
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

```
summary(white)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

Overall, these numbers seem to make sense – there are no screaming inaccuracies. That said, most of these variables are in different units, which have so far been undefined. To interpret what these summary statistics mean, I searched and defined the units.

The units are: (1) fixed acidity – g/L of tartaric acid; (2) volatile acidity – g/L of acetic acid; (3) citric acid – g/L; (4) residual sugar – g/L; (5) chlorides – g/L of sodium chloride; (6) free sulfur dioxide – mg/L; (7) total sulfur dioxide – mg/L; (8) density – g/mL; (9) pH – scale from 0 to 14; (10) sulphates – g/L of potassium sulphate; (11) alcohol – % by volume.

Understanding the units, I interpreted the summary statistics. Mean and maximum residual sugar seem to differ widely between the red and white samples; mean is 2.5 g/L for red and 6.4 g/L for white, while max is 15.5 g/L for red and 65.8 g/L for white. As a whole, this implies that white wines are generally sweeter. The acidity measures seem relatively to be stable between both, as do density, pH, sulphates, and alcohol. The other major difference occurs with chlorides and free/total sulfur dioxide. Mean chlorides is nearly double for red than for white, implying that red is generally saltier. Meanwhile, mean and max total sulfur dioxide stand at 46.5 mg/L and 289 mg/L for red, while they are 138.4 mg/L and 440 mg/L for white. These

summary statistics reveal that certain properties of red and white varieties are similar, but many others are heavily divergent.

To synthesize the dataset into one, I combined the two sets and added a column called "type" that notes whether the sample is red or white.

```
#combining the datasets into one
red$type <- "red"
white$type <- "white"
wine <- rbind(red,white)
dim(wine)
```

```
## [1] 6497   13
```

Another useful exploratory task is analyzing the correlation between the features of the dataset. Constructing a correlation matrix is beneficial, which I did below.

```
#evaluating correlation between variables
cor(subset(wine,select=-c(type)))
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity            1.00000000       0.21900826  0.32443573    -0.11198128
## volatile.acidity         0.21900826       1.00000000 -0.37798132    -0.19601117
## citric.acid              0.32443573      -0.37798132  1.00000000     0.14245123
## residual.sugar          -0.11198128      -0.19601117  0.14245123     1.00000000
## chlorides                0.29819477       0.37712428  0.03899801    -0.12894050
## free.sulfur.dioxide     -0.28273543      -0.35255731  0.13312581     0.40287064
## total.sulfur.dioxide    -0.32905390      -0.41447619  0.19524198     0.49548159
## density                  0.45890998       0.27129565  0.09615393     0.55251695
## pH                      -0.25270047       0.26145440 -0.32980819    -0.26731984
## sulphates                0.29956774       0.22598368  0.05619730    -0.18592741
## alcohol                 -0.09545152      -0.03764039 -0.01049349    -0.35941477
## quality                 -0.07674321      -0.26569948  0.08553172    -0.03698048
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity          0.29819477         -0.28273543          -0.32905390
## volatile.acidity       0.37712428         -0.35255731          -0.41447619
## citric.acid            0.03899801          0.13312581           0.19524198
## residual.sugar        -0.12894050          0.40287064           0.49548159
## chlorides              1.00000000         -0.19504479          -0.27963045
## free.sulfur.dioxide   -0.19504479          1.00000000           0.72093408
## total.sulfur.dioxide  -0.27963045          0.72093408           1.00000000
## density                0.36261466          0.02571684           0.03239451
## pH                     0.04470798         -0.14585390          -0.23841310
## sulphates              0.39559331         -0.18845725          -0.27572682
## alcohol               -0.25691558         -0.17983843          -0.26573964
## quality               -0.20066550          0.05546306          -0.04138545
##                           density          pH    sulphates      alcohol
## fixed.acidity          0.45890998 -0.25270047  0.299567744 -0.095451523
## volatile.acidity       0.27129565  0.26145440  0.225983680 -0.037640386
## citric.acid            0.09615393 -0.32980819  0.056197300 -0.010493492
## residual.sugar         0.55251695 -0.26731984 -0.185927405 -0.359414771
## chlorides              0.36261466  0.04470798  0.395593307 -0.256915580
## free.sulfur.dioxide    0.02571684 -0.14585390 -0.188457249 -0.179838435
```

```
## total.sulfur.dioxide  0.03239451 -0.23841310 -0.275726820 -0.265739639
## density               1.00000000  0.01168608  0.259478495 -0.686745422
## pH                    0.01168608  1.00000000  0.192123407  0.121248467
## sulphates             0.25947850  0.19212341  1.000000000 -0.003029195
## alcohol              -0.68674542  0.12124847 -0.003029195  1.000000000
## quality              -0.30585791  0.01950570  0.038485446  0.444318520
##                          quality
## fixed.acidity        -0.07674321
## volatile.acidity     -0.26569948
## citric.acid           0.08553172
## residual.sugar       -0.03698048
## chlorides            -0.20066550
## free.sulfur.dioxide   0.05546306
## total.sulfur.dioxide -0.04138545
## density              -0.30585791
## pH                    0.01950570
## sulphates             0.03848545
## alcohol               0.44431852
## quality               1.00000000
```

From the above, I noted the interactions that have a >50% positive or negative correlation since that implies a significant linear relationship. Total sulfur dioxide and residual sugar have a 49.5% relationship. Density and residual sugar have a 55.3% relationship. Total sulfur dioxide and free sulfur dioxide have a 72.1% relationship. Alcohol and density have a -68.7% relationship. I took note of the relationships between these features and determined to explore them more down the road.

Of notable importance is that the variable that has the largest absolute value correlation with quality is alcohol. While this was unexpected because a higher alcohol content would seem to give wine a more repulsive flavor on face value, perhaps there is some correlation between the strength of the wine and its perceived flavor. This was another relationship that I noted to evaluate later on.

The two variables that have the highest correlation are total sulfur dioxide and free sulfur dioxide. This makes sense because total sulfur dioxide is computed as free plus bound forms of SO2, so the relationship should be highly correlated. That said, this relationship is potentially important. The percent of free SO2 with regard to total SO2 could give some indication into quality. Bound SO2 is bound to other chemicals in the wine such as aldehydes, pigments, or sugars, so a higher free SO2 percentage of the total could imply more poignant tastes and smells (especially if it has antimicrobial properties).

In order to explore this relationship further, I engineered a new feature, which calculates the percent of free SO2 versus total SO2 in wine. I also ran summary statistics for the feauture below.

```
#engineering a new feature
wine$free.sulfur.dioxide.percent.total <- wine$free.sulfur.dioxide / wine$total.sulfur.dioxide

#running summary statistics for the new feature
summary(wine$free.sulfur.dioxide.percent.total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02273 0.20207 0.26977 0.28677 0.34884 0.85714
```

As can be seen, on average, Vinho Verde wines in the dataset seem to have 28.7% of their SO2 in free form and the remainder in bound form. The 25th and 75th percentiles put that percentage at 20.2% and 34.9%. On the low end, the percent drops to 2.3%, while on the high end, it rises to 85.7%. This engineered feature was used for all the further analysis below and provides greater insight than what the initial data enables.

Before continuing, I needed to scale and center the variables. Since each column is in different units (i.e. alcohol is in ABV, density is in g/mL, and chlorides is in g/L), I decided to normalize with 0 as the minimum and 1 as the maximum. This concisely represents all the variables and their distributions without having variables of larger scales visually warp the variables of smaller scales. I felt that both red and white datasets should be normalized on the same scale since the units for each column for both are the same; thus, they are measured according to the same metrics. I excluded normalizing the quality column, however, since that is the output variable and is already on a 0-10 scale.

```
#normalizing the combined dataset
mins <- apply(subset(wine,select=-c(quality,type)),2,min)
maxs <- apply(subset(wine,select=-c(quality,type)),2,max)
wine_scaled <- data.frame(scale(subset(wine,select=-c(quality,type)),center=mins,scale=maxs-mins))

#adding back the quality and type variables
wine_scaled$quality <- wine$quality
wine_scaled$type <- wine$type

#running summary statistics again
summary(wine_scaled)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.2149   1st Qu.:0.1000   1st Qu.:0.1506   1st Qu.:0.01840
##  Median :0.2645   Median :0.1400   Median :0.1867   Median :0.03681
##  Mean   :0.2823   Mean   :0.1731   Mean   :0.1919   Mean   :0.07428
##  3rd Qu.:0.3223   3rd Qu.:0.2133   3rd Qu.:0.2349   3rd Qu.:0.11503
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.0000       Min.   :0.0000
##  1st Qu.:0.04817   1st Qu.:0.05556     1st Qu.:0.1636       1st Qu.:0.1008
##  Median :0.06312   Median :0.09722     Median :0.2581       Median :0.1500
##  Mean   :0.07813   Mean   :0.10252     Mean   :0.2529       Mean   :0.1463
##  3rd Qu.:0.09302   3rd Qu.:0.13889     3rd Qu.:0.3456       3rd Qu.:0.1905
##  Max.   :1.00000   Max.   :1.00000     Max.   :1.0000       Max.   :1.0000
##        pH           sulphates         alcohol
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3023   1st Qu.:0.1180   1st Qu.:0.2174
##  Median :0.3798   Median :0.1629   Median :0.3333
##  Mean   :0.3864   Mean   :0.1749   Mean   :0.3611
##  3rd Qu.:0.4651   3rd Qu.:0.2135   3rd Qu.:0.4783
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  free.sulfur.dioxide.percent.total    quality          type
##  Min.   :0.0000                    Min.   :3.000   Length:6497
##  1st Qu.:0.2149                    1st Qu.:5.000   Class :character
##  Median :0.2961                    Median :6.000   Mode  :character
##  Mean   :0.3164                    Mean   :5.818
##  3rd Qu.:0.3908                    3rd Qu.:6.000
##  Max.   :1.0000                    Max.   :9.000
```

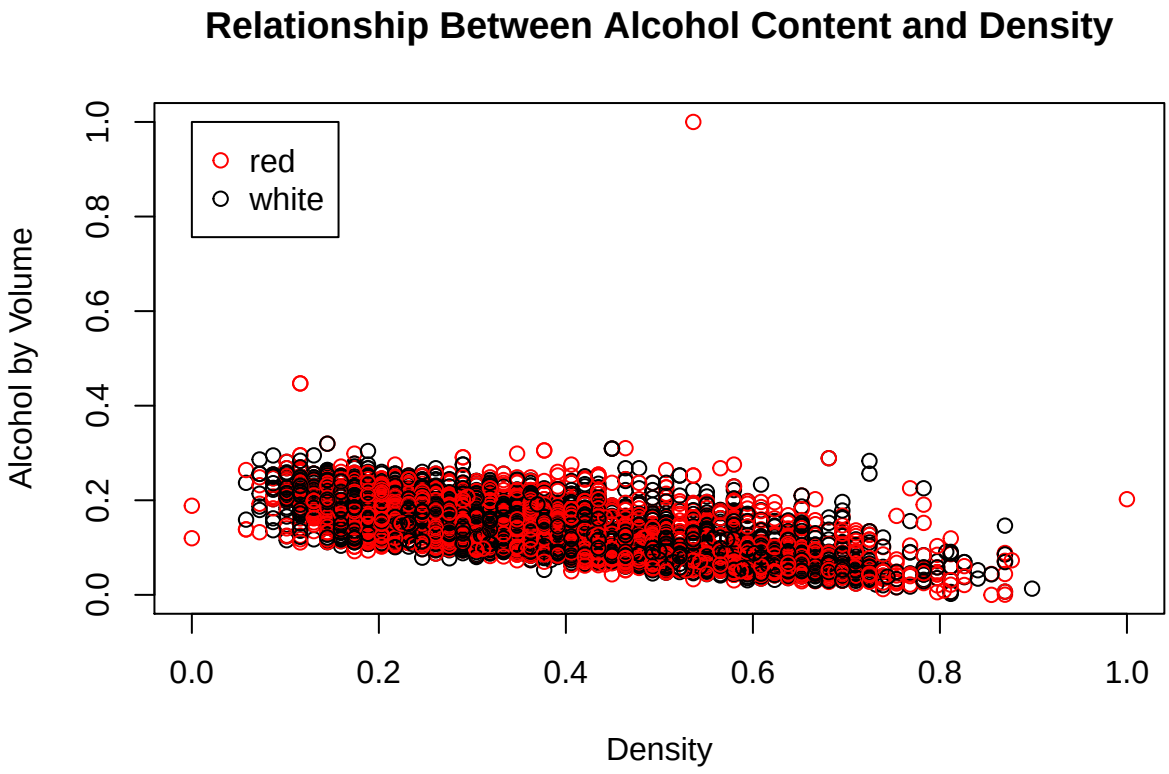
Revised summary statistics for each column are displayed above.

Relationships Between Features

Visualizations

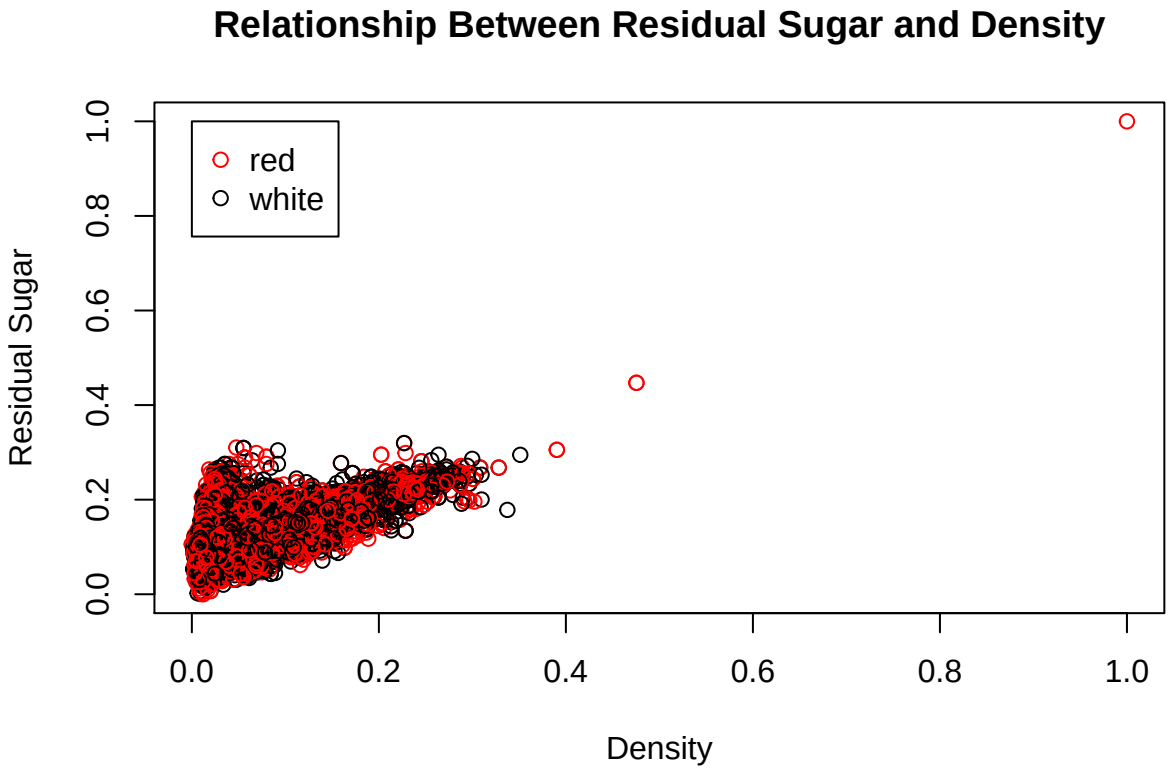
The next relevant aspect of this report is visualizing relationships between features in the dataset. I began by graphing the relationship between alcohol and density, since the correlation matrix above demonstrates a significant negative correlation. This makes sense because alcohol is less dense than water, so higher concentrations of alcohol mean that drinks are likely to be less dense in sum.

```
#some visualizations
library(ggplot2)
plot(wine_scaled$alcohol,wine_scaled$density,main="Relationship Between Alcohol Content and Density",yla
legend(0,1,unique(wine_scaled$type),col=rev(palette()[1:2]),pch=1)
```



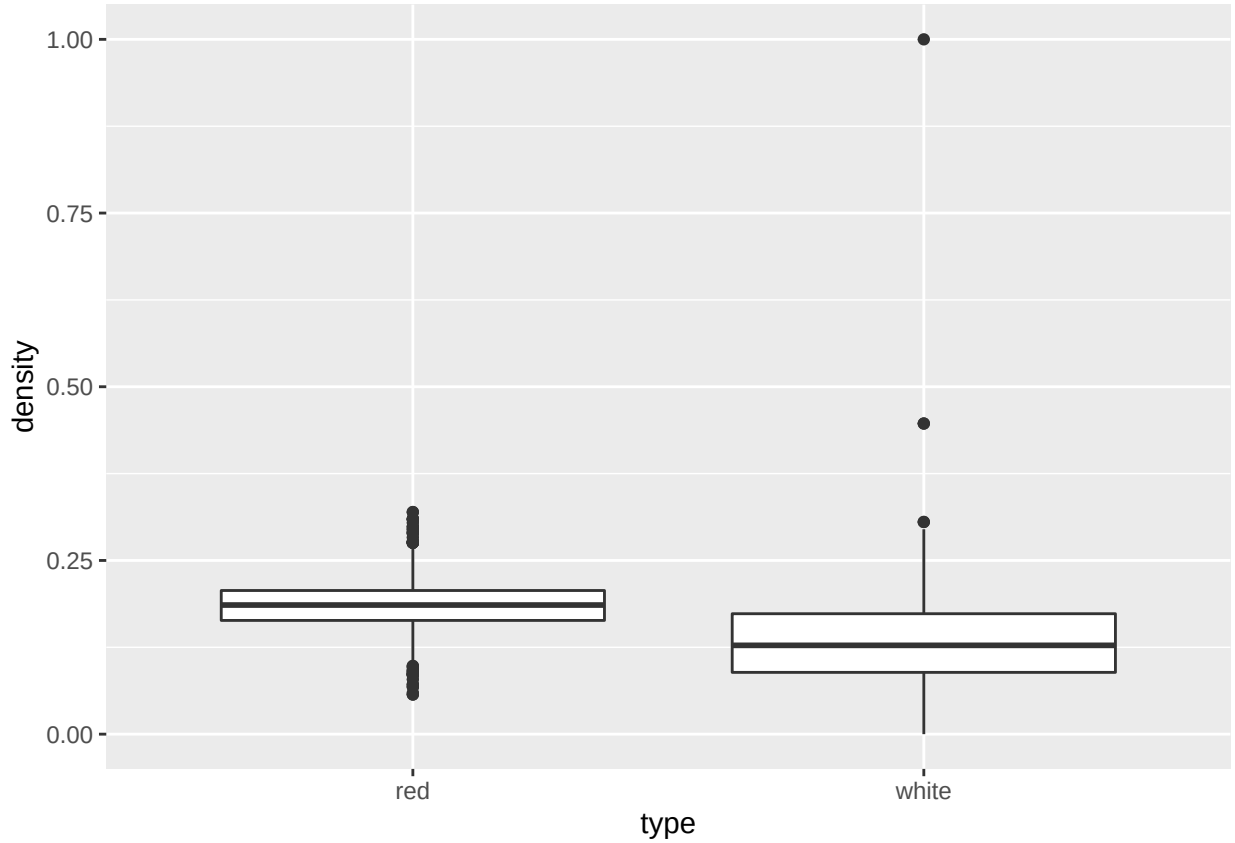
Next, I noticed that residual sugar and density have a significant positive relationship. By nature of the substance, sugars are more dense than water, so higher concentrations of residual sugar should increase the density of the drink as a whole.

```
plot(wine_scaled$residual.sugar,wine_scaled$density,main="Relationship Between Residual Sugar and Densi
legend(0,1,unique(wine_scaled$type),col=rev(palette()[1:2]),pch=1)
```

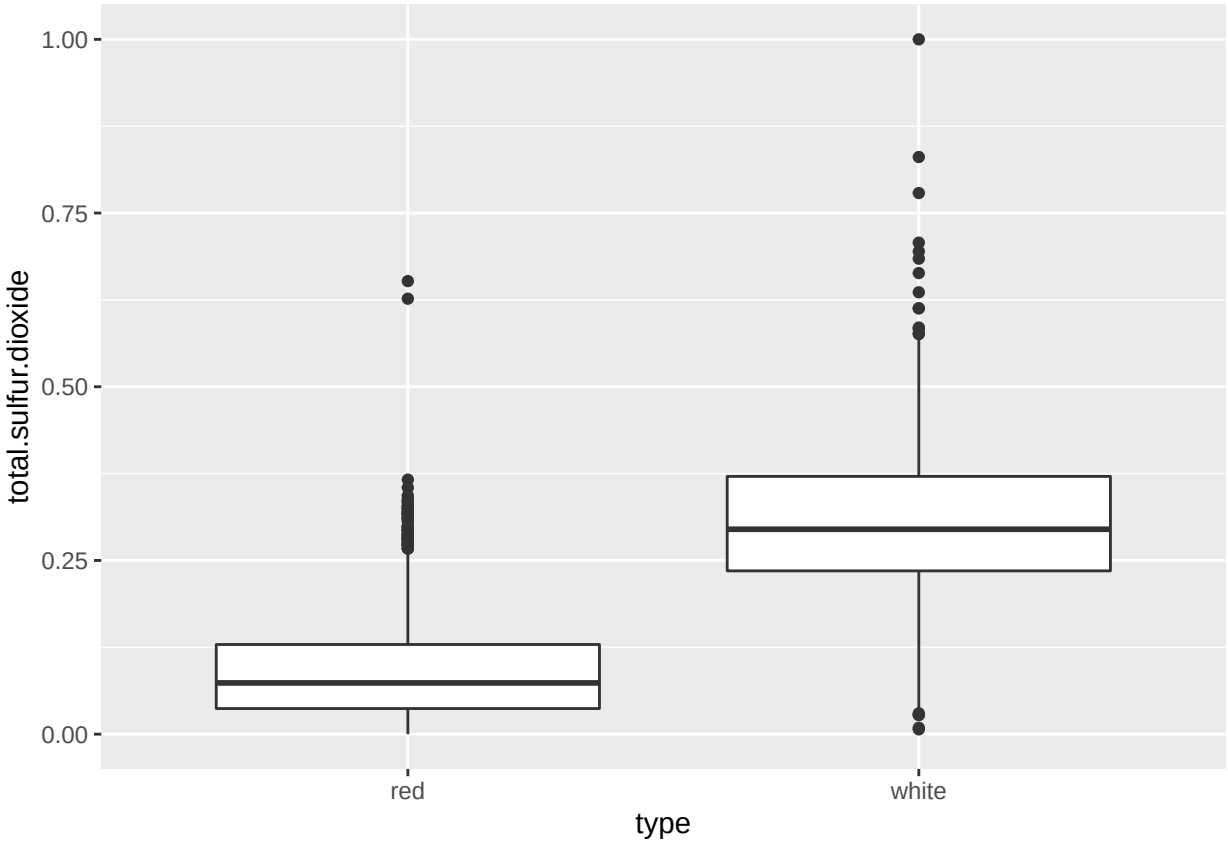


As noted earlier, there are significant differences in the red versus white populations. I wanted to visualize some of these differences, so I chose the properties that fluctuate the most in value between the two and plotted a boxplot. The properties graphed are density, total sulfur dioxide, free sulfur dioxide, sulphates, residual sugar, and quality.

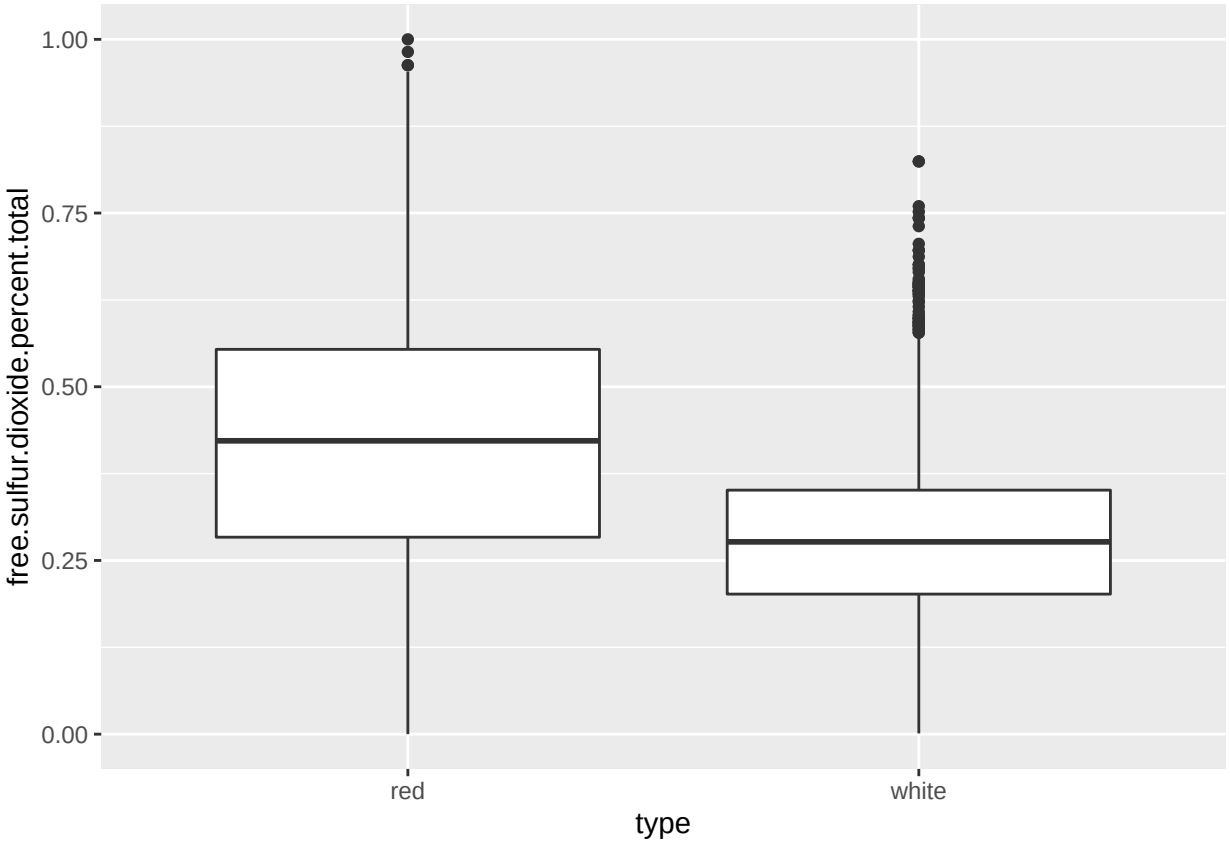
```
#visualizing the difference in red vs wine based on the summary statistics above
qplot(x = type, y = density, data = wine_scaled, geom = "boxplot")
```



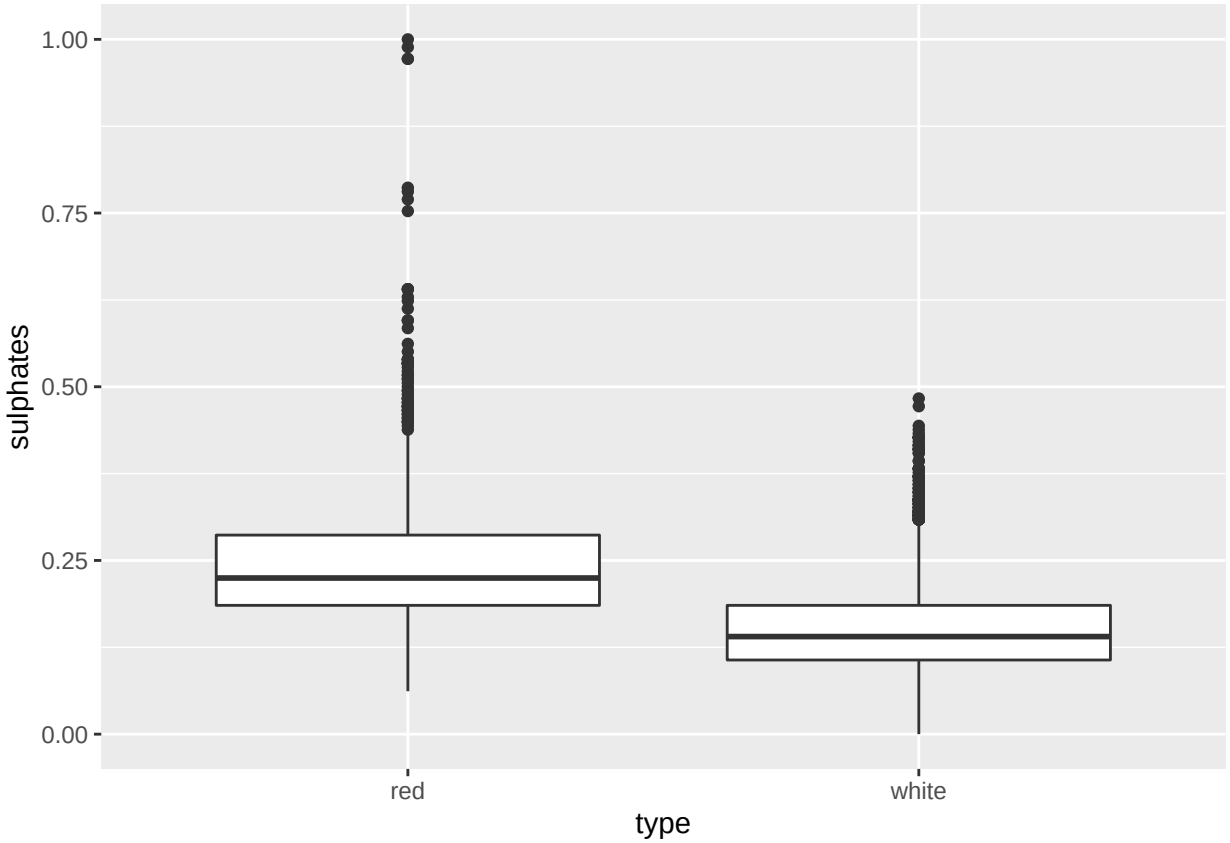

```
qplot(x = type, y = total.sulfur.dioxide, data = wine_scaled, geom = "boxplot")
```



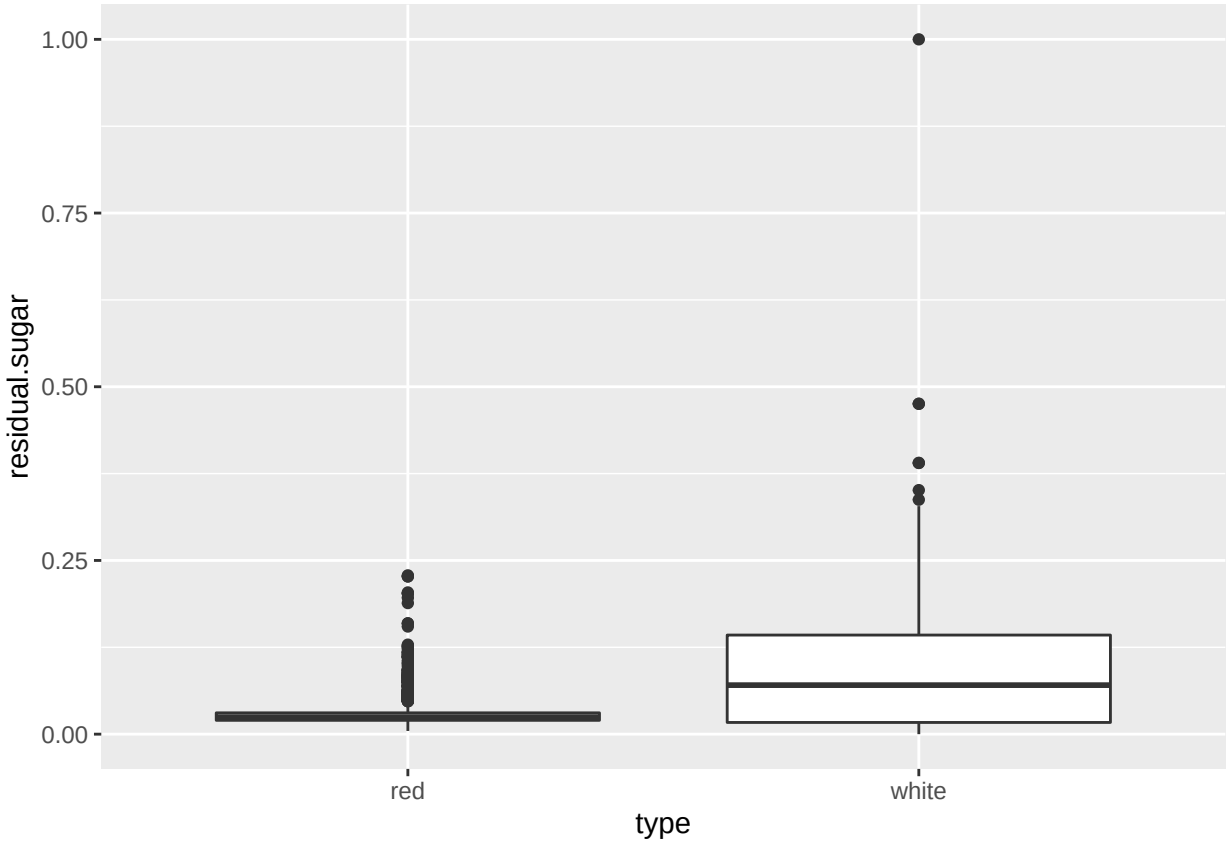

```
qplot(x = type, y = free.sulfur.dioxide.percent.total, data = wine_scaled, geom = "boxplot")
```



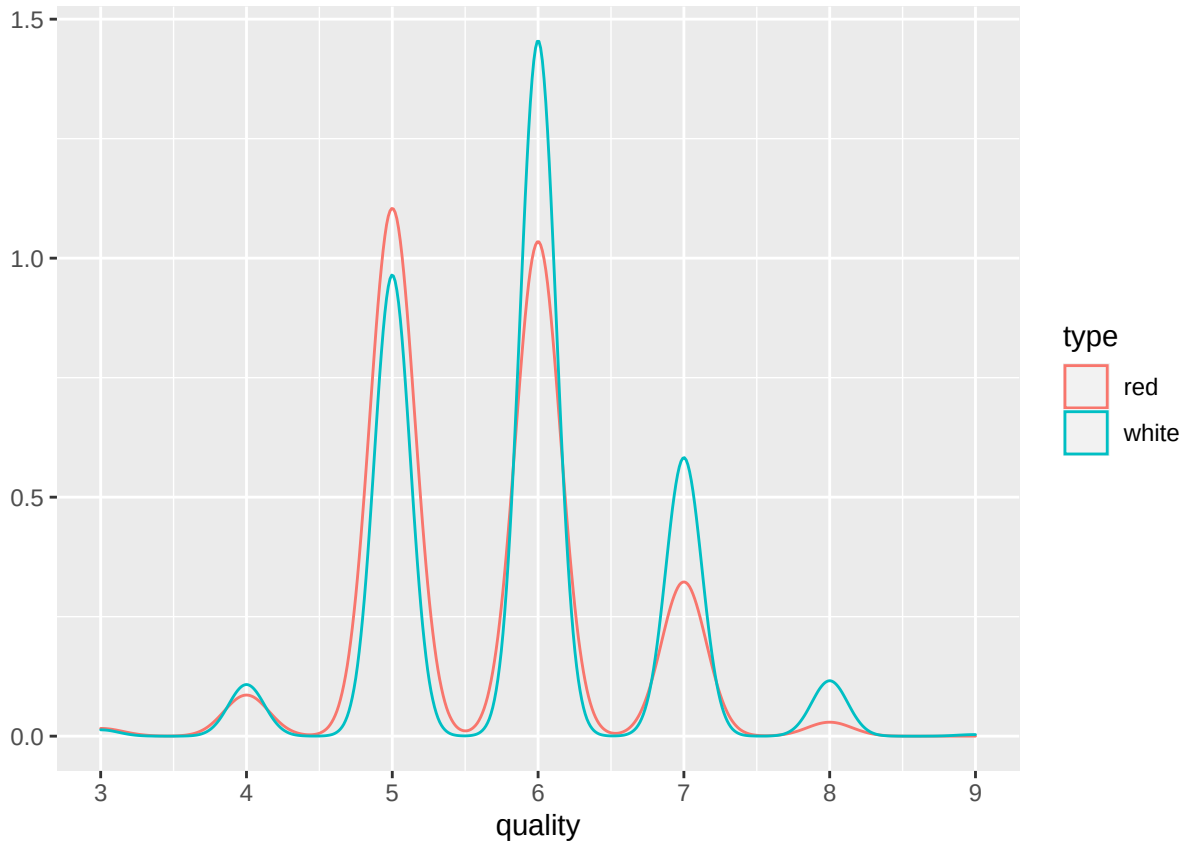

```
qplot(x = type, y = sulphates, data = wine_scaled, geom = "boxplot")
```




```
qplot(x = type, y = residual.sugar, data = wine_scaled, geom = "boxplot")
```




```
qplot(quality, data = wine_scaled, color = type, geom = "density") + scale_x_continuous(breaks = seq(3,
```



From the density graph, we can see that red wine is, on average, more dense than white wine. This is surprising considering white has higher residual sugar (which should make it more dense). However, this implies that the concentrations of the other properties in red wine increase its overall density more than in white.

White has a statistically significantly higher level of total sulfur dioxide. Interestingly enough, red wine has a higher proportion of free sulfur dioxide as a percent of total sulfur dioxide, which results in the SO2 being more present in the aroma and taste. This explains why red wine is often considered "richer" and "darker" than white wine.

Red also has a higher concentration of sulphates, explaining why it can age for much longer than white wine. As expected, the amount of residual sugar in red wine pales in comparison to white wine, a phenomenon that has been discussed above.

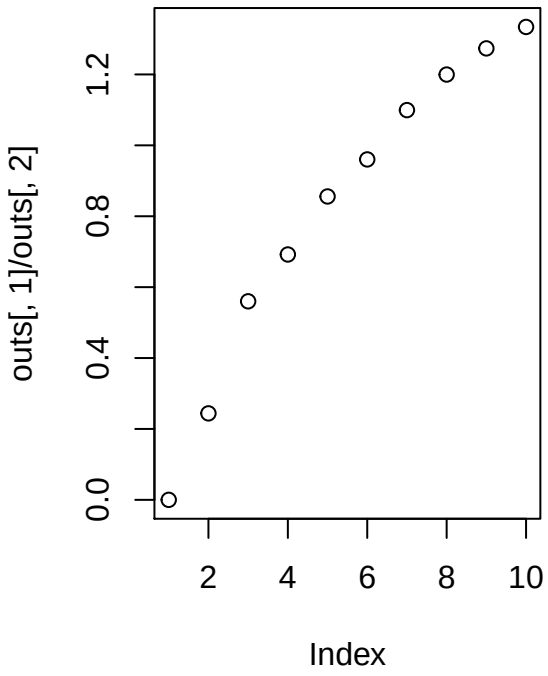
The final graph tells us the distribution of quality scores for both red and white. Although the averages of the samples are quite similar, the white sample has more of its quality scores centering around 6, with 6 being the peak average. On the other hand, the red sample has a higher percent of scores in 5. This indicates that although there is little difference in the distribution of quality scores, white wine receives (very) slightly higher rankings than red wine.

Clustering

Clustering is an algorithm that classifies entries into groups based on similarities. It involves automatically discovering natural groupings in data, generating insights that would not be obvious to someone simply looking at the dataset. I decided to run k-means clustering on the dataset in order to evaluate the relationships between rows. At the very least, we would expect the clustering algorithm to separate red from white

entries. On top of that, however, it would be valuable to see what other connections the algorithm is able to identify.

```
#running kmeans with k from 1 to 10
set.seed(420)
k_max <- 10
outs <- matrix(NA,ncol=2,nrow=k_max)
for (k_guess in seq_len(k_max)) {
  km_out <- kmeans(wine_scaled[,1:12],centers=k_guess)
  outs[k_guess,1] <- km_out$betweenss
  outs[k_guess,2] <- km_out$tot.withinss
}
par(mfrow=c(1,2))
plot(outs[,1] / outs[,2])
```



The plot above helps us identify the optimal number of clusters to choose. While this is more or less subjective, there seems to be a higher jump between 4 and 5 than 5 and anything following. As such, I chose the optimal number of clusters as 4.

```
#running kmeans with the optimal number of clusters
set.seed(420)
km_final <- kmeans(wine_scaled[,1:12],centers=4)

#comparing the mean values of each column in each cluster
cluster_summary <- c()
for (i in 1:4) {
```

```
   tempo <- c(apply(wine_scaled[km_final$cluster==i,-14],2,mean),table(wine_scaled[km_final$cluster==i,]
   cluster_summary <- rbind(cluster_summary,tempo)
}
rownames(cluster_summary) <- c("cluster_1","cluster_2","cluster_3","cluster_4")
print(cluster_summary)
```

```
##           fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## cluster_1     0.2425593        0.1380994   0.1975005     0.05017445 0.04741584
## cluster_2     0.2614890        0.1307472   0.2172131     0.15762526 0.07486104
## cluster_3     0.3862183        0.3038175   0.1616005     0.02949288 0.13303895
## cluster_4     0.2637061        0.1507906   0.1853116     0.05233547 0.07038040
##           free.sulfur.dioxide total.sulfur.dioxide    density        pH
## cluster_1          0.10610361           0.24434922 0.08134082 0.3705636
## cluster_2          0.16344765           0.38064640 0.18936405 0.3218977
## cluster_3          0.05054499           0.07395996 0.18870350 0.4581229
## cluster_4          0.08042322           0.27203174 0.13685248 0.4091766
##           sulphates   alcohol free.sulfur.dioxide.percent.total  quality  red
## cluster_1 0.1513753 0.5784577                          0.3126534 6.397534   85
## cluster_2 0.1531831 0.2038818                          0.3140424 5.586372    8
## cluster_3 0.2492140 0.3479166                          0.4741634 5.659786 1300
## cluster_4 0.1633149 0.3116286                          0.2084110 5.602868  206
##           white
## cluster_1  1618
## cluster_2  1665
## cluster_3     8
## cluster_4  1607
```

The 4 clusters seem to be doing a good job separating the red from white. While not perfect (because the tastes/properties of red and white can heavily overlap – even humans cannot distinguish the difference at times), clusters 1, 2, and 4 seem to focus on white wine, while cluster 3 focuses on red wine. Even within the 3 clusters of white wine, there is a separation along certain properties. For example, cluster 1 seems to contain the white wines with higher alcohol content, lower density, lower chlorides, and higher quality. On the other hand, clusters 2 and 4 seem more similar from a quality standpoint, but cluster 2 has a significantly lower alcohol content than cluster 4. In addition, cluster 2 has much higher free sulfur dioxide, total sulfur dioxide, and free sulfur dioxide percent of total sulfur dioxide. Residual sugar is also nearly 3x that of cluster 4 (meaning that these are likely much sweeter wines). Acidity levels are generally the same. When turning attention to cluster 3 (the red wines), sulfur dioxide concentrations pale in comparison to the whites. Sulphates, pH, chloride, and acidity levels are all, however, substantially higher. Despite these adjustments, quality seems relatively similar to clusters 2 and 4.

This clustering algorithm thus tells us that the red and white samples come from distinct populations. In addition, within the white sample, there are different groups that can be identified. One group (cluster 2) is generally sweeter (higher residual sugar) and has the lowest alcohol content. At the same time, this group has the lowest average quality ranking. One group (cluster 1) has very high alcohol content and is much less dense than the others. This group maintains the highest average quality ranking. The final group (cluster 4) sits somewhere in between the above two groups.

Since the results of this analysis seem useful, I will include the cluster assignment as a variable in the dataset.

```
wine_scaled$cluster <- as.factor(km_final$cluster)
```

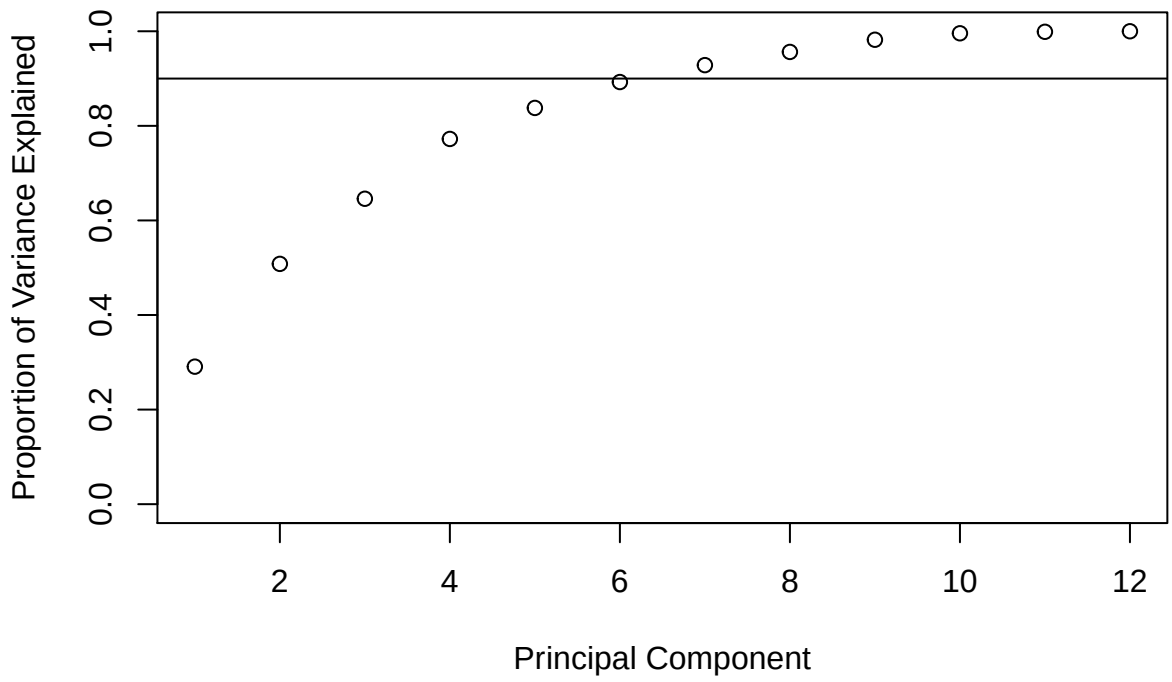
Principal Components Analysis

Now, while k-means clustering identifies relationships between rows in the dataset, principal components analysis identifies relationships between numeric columns. It groups properties (or features) together based on implicit relationships, correlations, and the like. It reduces dimensionality in the data and concisely represents the most important "principal components," which can then be used for analysis. As such, I ran PCA on the already scaled dataset.

```
#running PCA (centering and scaling is already done)
pca_out <- prcomp(wine_scaled[,1:12])

#plotting the ratio of cumulative sdev^2 over total sdev^2
plot(cumsum(pca_out$sdev^2)/sum(pca_out$sdev^2),ylim=c(0,1),xlab="Principal Component",ylab="Proportion

#let's preserve 90% of the variability
abline(h=0.9)
```



I chose the number of principal components to include (k) based on the number that would preserve 90% of the variability in the data. This is the preferable method over eyeballing for a k because there is no kink in the graph – the proportion of variability explained doesn't hit a certain "flat line" that makes it clear where the cutoff should be. Therefore, we need a more quantitative approach, such as choosing the smallest k that hits the 90% ratio. From the graph above, this happens when k is 6, so I will choose that as the number.

```
#selecting k and creating w
k <- 6
w <- pca_out$x[,1:k]
```

```r
#evaluating the top 3 factors that contribute the most to each principal component
for (i in 1:k) {
   print(paste("For principal component", i))
   print(head(pca_out$rotation[,i][order(pca_out$rotation[,i],decreasing=TRUE)],3))
}
```

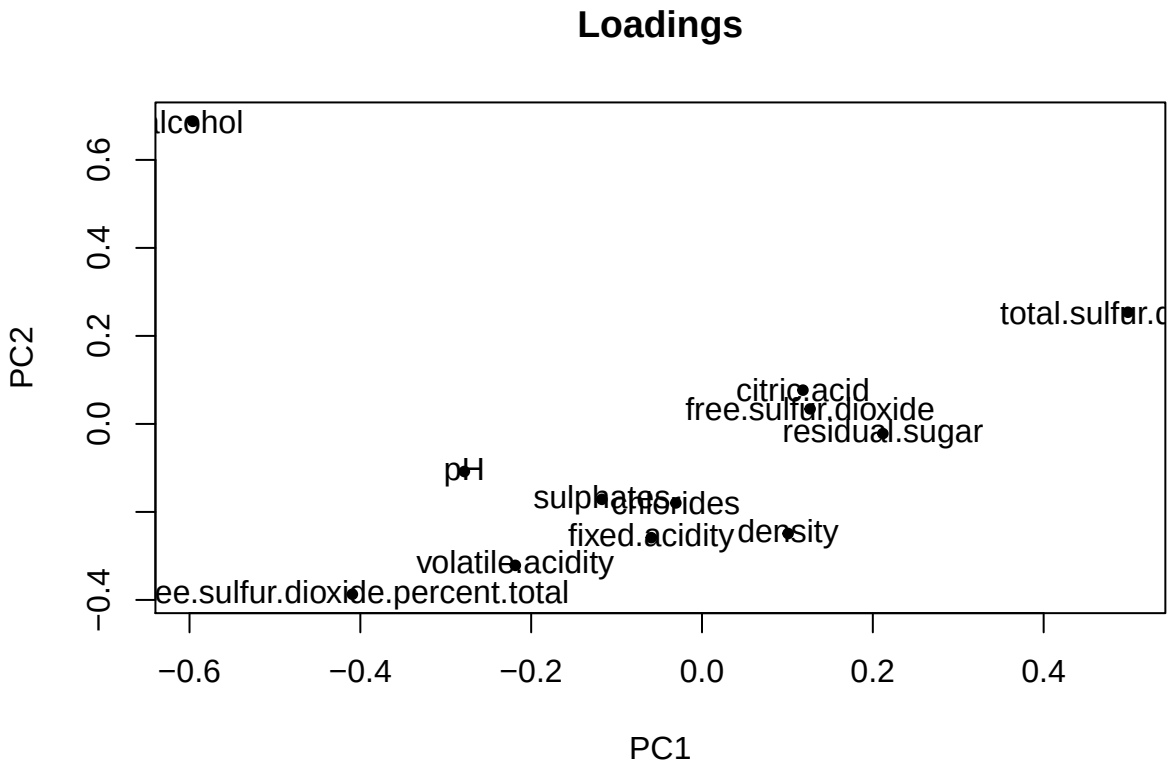
```
## [1] "For principal component 1"
## total.sulfur.dioxide       residual.sugar  free.sulfur.dioxide
##            0.4986990            0.2117710            0.1265624
## [1] "For principal component 2"
##              alcohol total.sulfur.dioxide          citric.acid
##           0.68758739           0.25375313           0.07673166
## [1] "For principal component 3"
## fixed.acidity   citric.acid       alcohol
##     0.5855261     0.2977093     0.1251793
## [1] "For principal component 4"
##               pH volatile.acidity        chlorides
##       0.39864130       0.33771440       0.04323249
## [1] "For principal component 5"
##                   volatile.acidity free.sulfur.dioxide.percent.total
##                         0.38583625                        0.06882218
##                     residual.sugar
##                         0.04381603
## [1] "For principal component 6"
##     volatile.acidity total.sulfur.dioxide              alcohol
##            0.6354656            0.5467119            0.2992904
```

Each of the principal components appears to emphasize a distinct theme. While there is some overlap between which features are significant, a few clearly identifiable patterns are observable. PC 1 seems to focus on SO2/residual sugar, while PC 2 focuses more on the alcohol content. PC 3 concerns itself with acidity (fixed acidity, citric acidity, etc...), PC 4 with pH levels (and some amounts of acidity), while PC 5 and 6 seem concerned with levels of acetic acid (represented as volatile acidity).

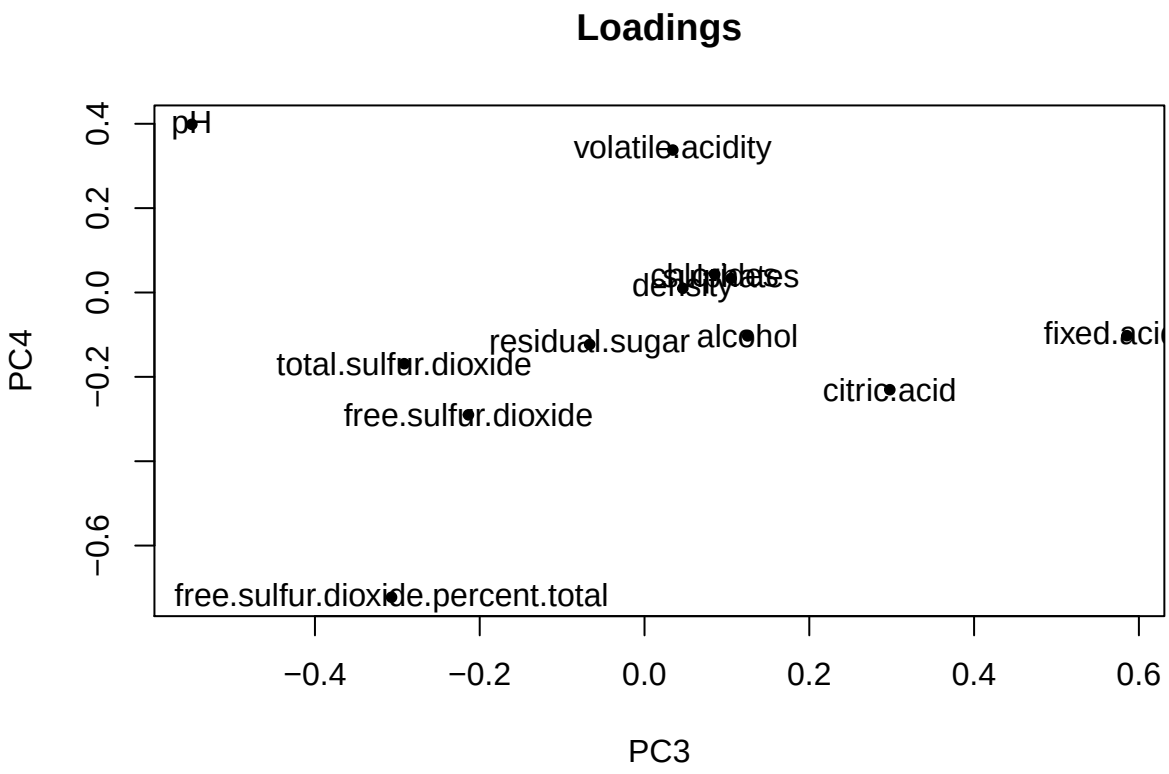
Despite the overlaps, the emphasis on which feature is most important is distinct for each PC. Ultimately, constructing this PC matrix reduces the dimensionality of the data and packages the columns that have an implicit relationship together.

Loadings plots are a useful graphical summary of how the algorithm works. These show the relationship between two PCs and how the variables factor into each PC – coded below.

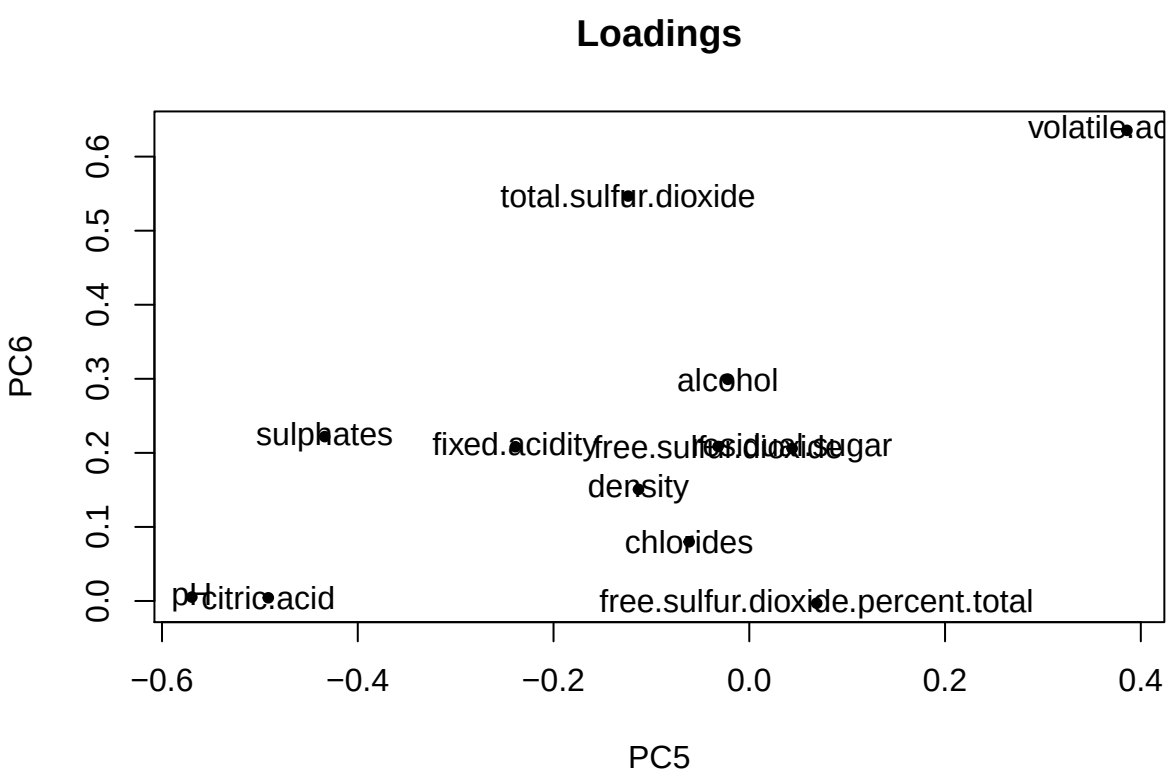
```r
#outputting loadings plot of PC 1 versus PC 2
plot(pca_out$rotation[,1:2],pch=20,bg="black",cex=1,main="Loadings")
text(pca_out$rotation[,1:2],labels=rownames(pca_out$rotation))
```




```
#outputting loadings plot of PC 3 versus PC 4
plot(pca_out$rotation[,3:4],pch=20,bg="black",cex=1,main="Loadings")
text(pca_out$rotation[,3:4],labels=rownames(pca_out$rotation))
```




```
#outputting loadings plot of PC 5 versus PC 6
plot(pca_out$rotation[,5:6],pch=20,bg="black",cex=1,main="Loadings")
text(pca_out$rotation[,5:6],labels=rownames(pca_out$rotation))
```



The loadings plots confirm the themes articulated above. As we can see, the first plot demonstrates that alcohol is a significant component of PC 2, while SO2/residual sugar concentrations influence PC 1. The other patterns described above are reflected in the remaining 2 graphs as well.

Regressions and Results

OLS

After analyzing the relationship between various features in the dataset, I recided to run a few regression algorithms to quantitatively answer the question: what physiochemical properties most significantly influence the quality of wine? I began with a simple OLS, then ran a LASSO model, then finally ran OLS on the PCA-adjusted data. I incorporated multiple models to confirm the accuracy of the results with multiple different algorithms. I used RMSE as the evaluation metric to ensure the robustness of each model. Those steps are articulated more below, but in order to do that, I first had to separate the dataset into a training set (70%) and testing set (30%).

```
#creating a training and testing set
set.seed(420)
test <- sample(seq_len(nrow(wine_scaled)),1950,replace=FALSE)
train <- seq_len(nrow(wine_scaled))[!seq_len(nrow(wine_scaled)) %in% test]
```

The first algorithm I ran is OLS regression. As the most standard algorithm, it is both the simplest to interpret and takes all the columns in the dataset (even non-numeric columns) into consideration.

```r
#training OLS on the training set
wine_scaled$type <- as.factor(wine_scaled$type)
ols_out <- lm(quality ~ ., data=wine_scaled[train,])
summary(ols_out)
```

```
##
## Call:
## lm(formula = quality ~ ., data = wine_scaled[train, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3117 -0.4784 -0.0314  0.4577  3.0645
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.71023    0.16354  34.916  < 2e-16 ***
## fixed.acidity                     0.95492    0.24640   3.875 0.000108 ***
## volatile.acidity                 -2.30541    0.14970 -15.400  < 2e-16 ***
## citric.acid                      -0.12416    0.15721  -0.790 0.429692
## residual.sugar                    4.14402    0.50421   8.219 2.65e-16 ***
## chlorides                        -0.32368    0.24365  -1.328 0.184086
## free.sulfur.dioxide              -0.73166    0.49440  -1.480 0.138967
## total.sulfur.dioxide              0.36000    0.23832   1.511 0.130968
## density                          -5.35503    1.00577  -5.324 1.06e-07 ***
## pH                                0.66723    0.14729   4.530 6.05e-06 ***
## sulphates                         1.22998    0.16125   7.628 2.90e-14 ***
## alcohol                           1.08730    0.17307   6.282 3.65e-10 ***
## free.sulfur.dioxide.percent.total  0.65883    0.16645   3.958 7.67e-05 ***
## typewhite                        -0.25493    0.07914  -3.221 0.001286 **
## cluster2                         -0.24811    0.05777  -4.295 1.79e-05 ***
## cluster3                         -0.05443    0.07482  -0.728 0.466935
## cluster4                         -0.19491    0.04398  -4.432 9.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7351 on 4530 degrees of freedom
## Multiple R-squared:  0.2949, Adjusted R-squared:  0.2924
## F-statistic: 118.4 on 16 and 4530 DF,  p-value: < 2.2e-16
```

The results of the regression are displayed above. From the coefficients that are statistically significant, the properties that have the biggest impact are volatile acidity (which, in higher concentrations, reduces the quality ranking), residual sugar (which raises the quality ranking significantly), and density (which lowers the quality ranking). Although these results appear to give some indication of the most important properties based on statistical significance and magnitude of coefficient, it is important to remember that many of these variables are collinear. The correlation matrix tells us that they are intimately related, which the OLS model does not account for. It assumes that these variables are all independent, which may not be the case. For example, while higher densities have a large negative impact on quality, this may be moreso because alcohol content is increasing, and alcohol and density have a negative linear relationship. As such, the results of this OLS model, while helpful, are incomplete.

Another important observation is that the coefficient on cluster 3 (which represents red wine) is statistically insignificant from cluster 1 (which represents one of the white wine groupings). With a p-value of 0.47, this result tells us that although clustering is a valuable algorithm for the insights gleaned in above sections,

the fact that a wine is "red" or "white" has little to do with its quality score. The other physiochemical properties are far more important than the color and the way in which the wine was produced.

To quantitatively evaluate the model's performance using an evaluation metric, I predicted the model on the testing set and used the predictions to calculate RMSE. I chose RMSE as the evaluation metric because it minimizes the difference between the predicted values and actual values. In this sense, it is the most appropriate metric for testing how accurate the model is. In addition, unlike R-squared, it can be calculated for multiple different models and evaluated universally between them.

```
#predicting OLS on the testing set
ols_pred <- predict(ols_out,wine_scaled[test,])

#calculating RMSE
sqrt(mean((wine_scaled[test,]$quality - ols_pred)^2))
```

```
## [1] 0.7168197
```

On an output range of 0-10, an RMSE of 0.72 is not bad. While a "good" RMSE cannot be determined in a vaccum, this gives some initial conviction that the OLS model is working decently well.

LASSO

The next model I decided to try is LASSO. It is a type of linear regression that utilizes shrinkage. In effect, data values are shrunk towards a central point, like the mean. When there are variables that are suspected to be statistically insignificant or correlated with one another, LASSO helps shrink the value of those variables to 0. This not only reduces the number of predictors, but it also helps choose only the most important variables. LASSO is perfect for the dataset that we have since it will penalize the less important features, making their respective coefficients zero. This provides us the benefit of feature selection and simple model creation.

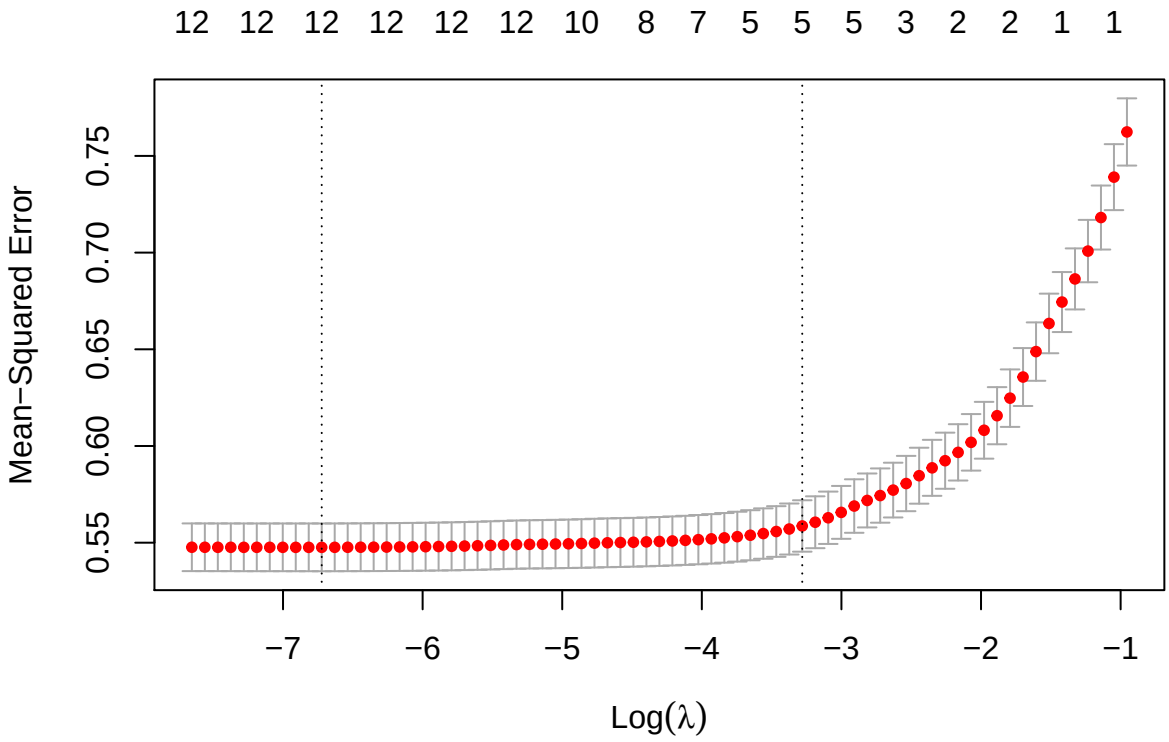
```
#running lasso on the train set with k-fold cross-validation to find optimal lambda
set.seed(420)
library(glmnet)
```

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

```
## Loaded glmnet 4.1-3
```

```
lasso.cv <- cv.glmnet(as.matrix(wine_scaled[train,1:12]),wine_scaled[train,13],alpha=1)

#plotting the change in cross validation error
plot(lasso.cv)
```




```
#determining the best lambda and betas
best_lambda <- which(lasso.cv$lambda==lasso.cv$lambda.1se)
best_betas <- lasso.cv$glmnet.fit$beta[, best_lambda]

#calculating the percent of coefficients shrunk to 0 for the lambda.1se model
length(best_betas[best_betas==0])/length(best_betas)
```

```
## [1] 0.5833333
```

The statistic above tells us that 58.3% of the parameters (or 7 of 12) of the parameters have been shrunk to zero by LASSO. As such, the model has chosen only the 5 most important predictors.

```
#demonstrating the coefficients that are important
coef(lasso.cv)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                    5.1781568
## fixed.acidity                  .
## volatile.acidity              -1.7186029
## citric.acid                    .
## residual.sugar                 0.4294669
## chlorides                      .
## free.sulfur.dioxide            .
## total.sulfur.dioxide           .
```

```
## density                           .
## pH                                .
## sulphates                         0.3745001
## alcohol                           1.9874991
## free.sulfur.dioxide.percent.total 0.4076915
```

The coefficient values above tell us that the 5 most significant properties that influence quality are alcohol, volatile acidity, free SO2 as a percent of total SO2, residual sugar, and sulphates. Alcohol has the largest positive effect, while volatile acidity (acetic acid) has the largest negative effect. This makes sense because acetic acid has a vinegar-esque taste, so higher contents of acetic acid can give the wine a repulsive flavor and aroma. Alcohol content also affects a wine's body (which determines quality) since alcohol is more viscous than water (a wine with more alcohol will have a fuller, richer body). Residual sugar (sweetness) and sulphates (antioxidation) are also important.

The feature I engineered, free SO2 as a percent of total SO2, is more important than the values of free SO2 and total SO2 independently. This tells us that the total amount of SO2 does not matter much for quality as long as the free SO2 percent of total is high. This result demonstrates that the feature I engineered is valuable and should be a metric that wine producers calculate to evaluate wine quality.

Similar to OLS, I predicted the LASSO model on the testing set and calculated RMSE.

```
#predicting lasso on the testing set
lasso_pred <- predict(lasso.cv,as.matrix(wine_scaled[test,1:12]))

#calculating RMSE
sqrt(mean((wine_scaled[test,]$quality - lasso_pred)^2))
```

```
## [1] 0.7329093
```

The RMSE of LASSO is 0.73, extremely close to the previous RMSE of 0.72. Although it is slightly higher, LASSO's ability to deal with multicollinearity makes its result just as valuable as OLS. As such, it is worth taking both models into consideration when drawing inferences.

OLS With PCA

The final model I attempted was OLS on the PCA-adjusted dataset. Since LASSO uses a biased estimator (because of regularization), I wanted to try an unbiased method that can deal with multicollinearity. PCA automatically reduces dimensionality and aggregates variables that are correlated into a singular principal component, so running OLS with PCA is useful. The results of the regression are below.

```
#running OLS with PCA on the training set
df_pca <- data.frame(w,wine_scaled$quality)
pca_ols <- lm(wine_scaled.quality ~ .,data=df_pca[train,])
summary(pca_ols)
```

```
##
## Call:
## lm(formula = wine_scaled.quality ~ ., data = df_pca[train, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6366 -0.4834 -0.0493  0.4855  3.1682
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.82261    0.01118 520.627  < 2e-16 ***
## PC1         -1.08587    0.05582 -19.453  < 2e-16 ***
## PC2          1.78874    0.06438  27.786  < 2e-16 ***
## PC3         -0.07129    0.08170  -0.873  0.38297
## PC4         -1.42582    0.08492 -16.789  < 2e-16 ***
## PC5         -1.19648    0.11711 -10.217  < 2e-16 ***
## PC6         -0.39830    0.12814  -3.108  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.754 on 4540 degrees of freedom
## Multiple R-squared:  0.2565, Adjusted R-squared:  0.2556
## F-statistic: 261.1 on 6 and 4540 DF,  p-value: < 2.2e-16
```

PC 3 (the component that deals with fixed and citric acidity) seems to be non-significant. This aligns with the results of LASSO, which shrunk the coefficients of those variables to 0. In addition, PC 2 (the component that deals with alcohol) seems to have the largest influence on quality, again confirming the results of LASSO. PC 1 (which relates to residual sugar/SO2 content) is also significant, but in the negative direction. This is the opposite result given by LASSO, which is interesting. PC 4, 5, and 6 all relate to overall acidity (pH) and acetic acid concentration (volatile acidity), and both LASSO and OLS with PCA model these coefficients as negative.

I then predicted OLS with PCA on the testing set and calculated RMSE.

```
#predicting OLS with PCA on the testing set
pca_pred <- predict(pca_ols,data.frame(w[test,]))

#calculating RMSE
sqrt(mean((wine_scaled[test,]$quality - pca_pred)^2))
```

```
## [1] 0.7353698
```

RMSE falls within the range described above, with 0.74 falling not far from LASSO's 0.73 and standard OLS's 0.72. Given the output range of 0-10, these values are quite decent. We can draw inferences from the models and be confident that the models have some explanatory power.

Summary of Results

To summarize, if we aggregate the results of the 3 mining algorithms, we end up with a few significant insights.

(1) Alcohol content has the most significant positive affect on wine quality. Since alcohol is less dense than water, wines that have higher concentrations of alcohol also end up being less dense. This results in a relationship where higher quality wines simultaneously have higher alcohol and lower density.

(2) Fixed acidity and citric acid are not determinants of quality, especially when compared to volatile acidity, which has a significant negative affect. Volatile acidity is the amount of acetic acid, and it adds a pungent, sharp, vinegar flavor to the wine. Acetic acid in doses too high creates a prickling sensation on the sides of the tongue and a salty aftertaste, and this can be overwhelming for drinkers of wine. As such, volatile acidity puts downward pressure on quality.

(3) The total concentration of SO2 is not important, especially because this value fluctuates heavily for red versus white wine. What is important is the percent of total SO2 that is free SO2, since free SO2 is unbound to chemicals. As such, free SO2 prevents microbial growth and oxidation, which adds freshness to the wine (preventing spoilage). If there is an imbalance between free and bound SO2, it can disrupt the taste and reduce quality.

(4) Finally, residual sugars (sweetness) and sulphates (antioxidants) are important, but the direction of their affect is unclear. Wines that are not sweet or salty enough likely lack flavor; at the same time, wines that are too sweet or too salty are overwhelming in taste. The quality impact of these two variables is more of a subjective determination.

The likelihood that the results are due to random chance is quite low. To evaluate this quantatively, I randomly predicted from the range 0 to 10 and calculated that algorithm's RMSE.

```
#randomly predicting
set.seed(420)
random_pred <- sample(0:10,dim(wine_scaled[test,][1]),replace=TRUE)

#calculating RMSE
sqrt(mean((wine_scaled[test,]$quality - random_pred)^2))
```

```
## [1] 3.295218
```

As we can see, the RMSE from random prediction is nearly 5x that of the 3 models. Thus, I am confident that the likelihood that these results are random chance is quite low. To evaluate this qualitatively, research on wine characteristics seems to confirm the above insights. For example, Wine Folly (https://winefolly.com/tips/the-spectrum-of-boldness-in-red-wines-chart/) writes that wines with higher alcohol tend to taste bolder (because higher alcohol wines collect more droplets on the side of the glass than low alcohol wines, the Marangoni effect). In addition, PubMed (https://pubmed.ncbi.nlm.nih.gov/20931186/) shows that higher levels of acetic acid (volatile acidity) reduce yeast fermentative performance, which is a significant driver of wine quality. Thus, volatile acidity, despite contributing a sometimes desirable vinegar taste, can often reduce overall quality. Finally, research from Iowa State University (https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/) confirms that percentage of free SO2 determines how much SO2 is available in active, molecular form to help protect the wine from oxidation and spoilage – and this is beneficial for quality.

Despite the results being robust, high uncertainty can affect conclusions by spreading the sample average widely around the population average, meaning that the sample may not necessarily represent the population. Not only does this introduce skepticism into the strength of the conclusions reached, but it also means that the results cannot be generalized accurately to other samples. It reduces the explanatory power of the results and weakens the claims of the inferences. One area in the current dataset where there might be high uncertainty is with regard to the "quality" column. Currently, quality represents the median of 3 scores assessed by judges. However, given that only 3 judges assessed the wine, there is a high degree of variability. In addition, the median is not necessarily representative of consensus opinion between the 3. Taking the average would be more accurate. This adds an element of skepticism into how the results truly apply to the subjective "quality" that you and I might taste in wine.

Conclusion

Most Important Insights

In conclusion, the most important insights from clustering, PCA, OLS, LASSO, and OLS With PCA helped me answer three critical questions.

(1) What physiochemical properties influence the quality of wine? OLS, LASSO, and OLS With PCA tell us that higher alcohol content and free SO2 as a percent of total SO2 improve quality, while higher volatile acidity (acetic acid) reduces quality. Although it is unlikely that the inclusion of alcohol itself improves quality, it is likely that higher quality brands want their bottles to be stronger (to compensate for the higher price) and are more skilled at mixing the alcohol into the drink. In addition, higher alcohol content changes the texture of the drink by bringing out tannins, which improves quality. These are properties that wine producers should be aware of and can calibrate carefully in their wines. These are also non-obvious insights because many of these relationships are not reflected on face value in the correlation matrix. For example, as an engineered feature, free SO2 as a percent of total SO2 is not in the matrix, and the original free SO2 and total SO2 variables have very little correlation with quality. Variables such as density appear to be relevant in the matrix, but we know that since density is just a function of alcohol (lower density), residual sugar (higher density), and other properties, it cannot be significant by itself. This is a result confirmed by the 3 models and gives conviction to the fact that hidden patterns have been revealed.

(2) What are significant relationships between properties? PCA tells us that total sulfur dioxide and residual sugar have interactions with one another, and fixed acidity and citric acid share a relationship as well. This aligns with the results above because both fixed acidity and citric acid appear insignificant (it wouldn't make sense if one was significant and the other wasn't). Volatile acidity and pH, both contributing to the vinegar taste of wine, also correlate with each another. The distinctness of each principal component allows us to see that alcohol, volatile acidity/pH, and SO2/residual sugar have divergent relationships from one other. Again, this is corroborated by the mining algorithms above, where each of those combinations of features has its own identifiable impact on the overall quality of wine.

(3) What are major differences between red and white wine, and how does that impact quality? K-means clustering allows us to draw a few important conclusions about red versus white wine. Generally, red wines have a much higher free SO2 percent of total SO2, meaning they have more antimicrobial properties. They also have a higher concentration of sulphates and acids (fixed acidity and volatile acidity). Chloride content is much higher as well, explaining the more salty taste in red wines. With regard to white wines, at least in this dataset, there is a huge degree of variability. The white wines with a higher average quality rating tend to have more alcohol and less chlorides. Other white wines tend to be much sweeter (higher residual sugar) but simultaneously lack alcohol content. There seems to be some tradeoff between alcohol and residual sugar, and this is corroborated by the fact that OLS With PCA yields opposite signs for those coefficients. Other differences included density and acidity.

Iterations Performed and Data Snooping

My project went through multiple iterations from beginning to end. When I started, I planned on using only one grouping algorithm (either clustering or PCA) and only one mining algorithm (OLS or LASSO). As I ran the initial analysis, I realized that adding different algorithms could expose unique angles to the data, and this would help me draw more valuable insights. I also ran the initial analysis without normalizing the data (which is incorrect since each column is in different units), and this threw vastly different results from what I got after normalization. Normalizing helped standardize the data and reveal patterns that I wouldn't have otherwise seen. Finally, after sending my rough draft to Krishna Menon, I incorporated some of his edits. In my graphs, I color-coded the points to show which are red versus white. I rewrote my introduction and conclusion to simplify the language and make it easier to understand. I also had initially chosen k for clustering and PCA by eyeballing, but he told me to rely on more quantitative methods, which is what I ended up doing.

Data snooping, also known as p-hacking, is the misuse of data analysis to find patterns in data that can be presented as statistically significant. This is done by performing many statistical tests on the data and only reporting those that come back with significant results. In my project, I attempted to reduce data snooping as much as possible by including all algorithms that I originally tried out. Even if many important coefficients (such as alcohol) had ended up being insignificant in standard OLS, I would have included the

results for the sake of completeness. That said, when utilizing clustering and PCA algorithms, I did try out multiple different k values to see which one would report the most complete results. Although choosing k is an inherently subjective task (so manually checking the result for relevance is not a significant issue), it could fall within the umbrella of data snooping and is something I tried to fix as quantitatively as possible.

Further Research

If I had more time, I would have addressed the following drawbacks in the report.

First, I would have liked a larger dataset with more rows (more observations) and more columns (more features). The current dataste contains nearly 6,500 observations, a decent amount of wine samples – but it is the only dataset on wine that is publicly available. All of these observations originate from the Vinho Verde region, which reduces the generalizability of the results. It also would have been nice to have other metrics such as the wine brand or bottle price, since those would have added more dimensions to the predictive accuracy of the models. While it was helpful that Professor Cortez automatically cleaned out incomplete and nonsensical entries, I would have appreciated having the full data from the CVRVV since cleaning is part of data mining!

Second, I would have attempted to run more algorithms, such as hierarchical clustering or ridge regression. Although I felt that k-means clustering was robust, hierarchical clustering uses a different algorithm to tackle the same grouping objective. At worst, this would have yielded different results. At best, it would have yielded more accurate results. In addition, while LASSO sets the value of insignificant coefficients to zero, ridge simply shrinks the coefficients close to zero without removing them. This might have yielded more complete results since it still factors in all the variables.

Third, I would have thought about asking a different question. I could have made classification the objective of the mining: can we determine what is a red or white wine based on the physiochemical properties? This is an interesting question because even experts have trouble distinguishing between the two in a blind taste test (https://www.realclearscience.com/blog/2014/08/the_most_infamous_study_on_wine_tasting.html). Perhaps some objective metrics (beyond sensory taste) could be used to effectively categorize the two.

Critique of Another Project

Krishna Menon's project is focused on analyzing a baseball dataset that includes statistics on 2,805 players since the founding of the MLB. The task is classification: to determine what factors influence players winning or losing the World Series.

The dataset includes 16 different variables (such as hits, runs scored, strikeouts, etc...) that aggregate statistics on a player. Although this dataset is quite robust – and gathers data over multiple decades – I would have tried to find and incorporate a supplementary dataset that adds two crucial features: (1) the coach/team under which the player played, and (2) the number of years the player has been playing. While this dataset may not exist readily, I feel that it wouldn't be too difficult to manually scrape. Having data on the coach/team would add much because coaches often direct strategy for the team and sculpt players over time. Perhaps there are a few coaches that helped multiple players attain wins. Additionally, since baseball is a team sport, the strength of the team would also determine how well a player performs in any given game. Finally, having information on the number of years a player played would be relevant because players with more experience (and more years) are more likely to have a win under their belt as compared to less experienced players. Altogether, this additional dataset would reveal non-obvious patterns in the data and help "mine" it for insights.

Another recommendation I have is utilizing different mining algorithms – specifically, logistic regression and LASSO. Currently, the analysis contains OLS With PCA, which is great for reducing dimensionality and multicollinearity. However, when the response variable is binary, OLS will produce linear probabilities that don't fit within 0 and 1. As a result, conducting inference with the model is difficult. Logistic regression solves

for this by constraining predictions to between 0 and 1, improving interpretability and accuracy. Similarly, while PCA is useful, principal components are often difficult to parse through, and it is sometimes unclear what each component is comprised of. LASSO solves for the issue of dimensionality because the coefficient on variables that are insignificant will be shrunk to zero. The resulting regression will only factor in highly relevant variables, and this will fine-tune the analysis. It also has the ability to deal with multicollinearity, which is an added benefit.