Wine Analysis

Bryan Stapleton

2020-05-17

What factors influence the quality of wine? I will be looking at chemical properties of both red and white wines to predict the quality of the wine as rated as the median of at minimum three scores on a scale from one to ten. Our predictors include fixed acidity (g(tartaric acid)/dm^3), volatile acidity (g(acetic acid)/dm^3) , citric acid (g/dm^3), residual sugar (g/dm^3), chlorides (g(sodium chloride)/dm^3), free sulfur dioxide (mg/dm^3), total sulfur dioxide (mg/dm^3), density (g/cm^3), pH, sulphates (g(potassium sulphate)/dm^3), and alcohol (vol.%). I will first preform exploratory data analysis on both the red and white wine data sets, looking at the univariate distributions of our eleven predictors and response variable quality.
I will first provide a table of the Summary Statistics (min, quartiles, max, mean and standard deviations) for both red and white wines.

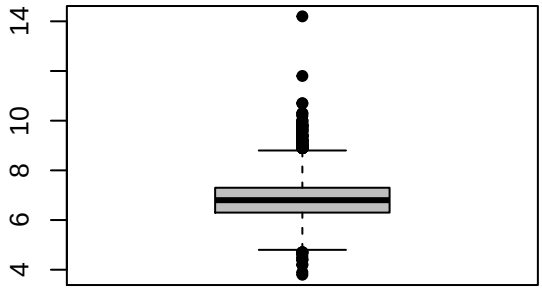
```
##
## Red Wine Summary Statistics
## ============================================================================
## Statistic             Min  Pctl(25) Median Pctl(75)  Max    Mean  St. Dev.
## ----------------------------------------------------------------------------
## alcohol              8.400  9.500   10.200  11.100  14.900 10.423  1.066
## chlorides            0.012  0.070   0.079   0.090   0.611  0.087   0.047
## citric.acid            0     0.1     0.3     0.4      1    0.271   0.195
## density              0.990  0.996   0.997   0.998   1.004  0.997   0.002
## fixed.acidity        4.600  7.100   7.900   9.200   15.900 8.320   1.741
## free.sulfur.dioxide    1      7       14      21      72   15.875  10.460
## pH                   2.740  3.210   3.310   3.400   4.010  3.311   0.154
## quality                3      5       6       6       8    5.636   0.808
## residual.sugar       0.900  1.900   2.200   2.600   15.500 2.539   1.410
## sulphates            0.330  0.550   0.620   0.730   2.000  0.658   0.170
## total.sulfur.dioxide   6      22      38      62     289   46.468  32.895
## volatile.acidity     0.120  0.390   0.520   0.640   1.580  0.528   0.179
## ----------------------------------------------------------------------------
```

```
##
## White Wine Summary Statistics
## ============================================================================
## Statistic             Min  Pctl(25) Median Pctl(75)  Max    Mean   St. Dev.
## ----------------------------------------------------------------------------
## alcohol              8.000  9.500   10.400  11.400  14.200 10.514   1.231
## chlorides            0.009  0.036   0.043   0.050   0.346   0.046   0.022
## citric.acid          0.000  0.270   0.320   0.390   1.660   0.334   0.121
## density              0.987  0.992   0.994   0.996   1.039   0.994   0.003
## fixed.acidity          4     6.3     6.8     7.3      14    6.855   0.844
## free.sulfur.dioxide    2      23      34      46     289   35.308   17.007
```

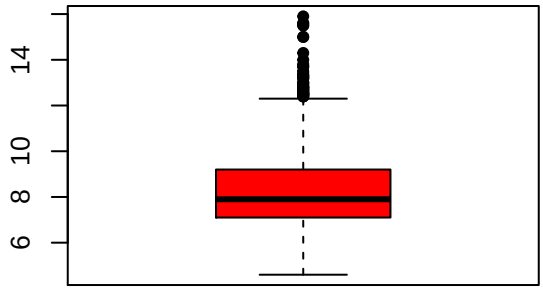
```
## pH                     3     3.1     3.2     3.3      4     3.188   0.151
## quality                3       5       6       6      9     5.878   0.886
## residual.sugar       0.600   1.700   5.200   9.900  65.800  6.391   5.072
## sulphates            0.220   0.410   0.470   0.550  1.080   0.490   0.114
## total.sulfur.dioxide   9     108     134     167    440   138.361  42.498
## volatile.acidity     0.080   0.210   0.260   0.320  1.100   0.278   0.101
## ------------------------------------------------------------------------
```

From the table you can see both red and white wine had a median quality score of six. The biggest distinction between red and white wines seems to be the free sulfur dioxide and total sulfur dioxide, for which white wine is significantly higher in both. A few of our predictors measure something related to the acidity of the wine including citric acid, fixed acidity, volatile acidity, and pH. Based on the summary output its difficult to see any major differences in these between red and white wine. Let's try looking at the box plots of these acidity variables to see if we can notice any distinctions.

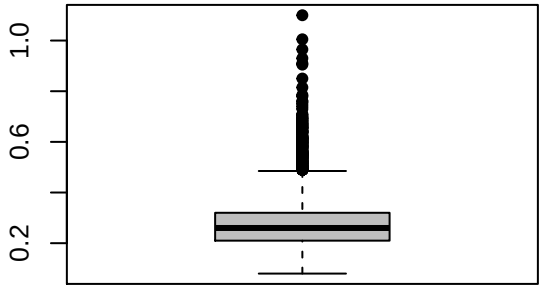
boxplots



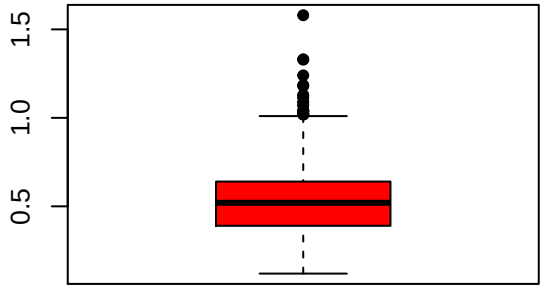
Fixed Acidity



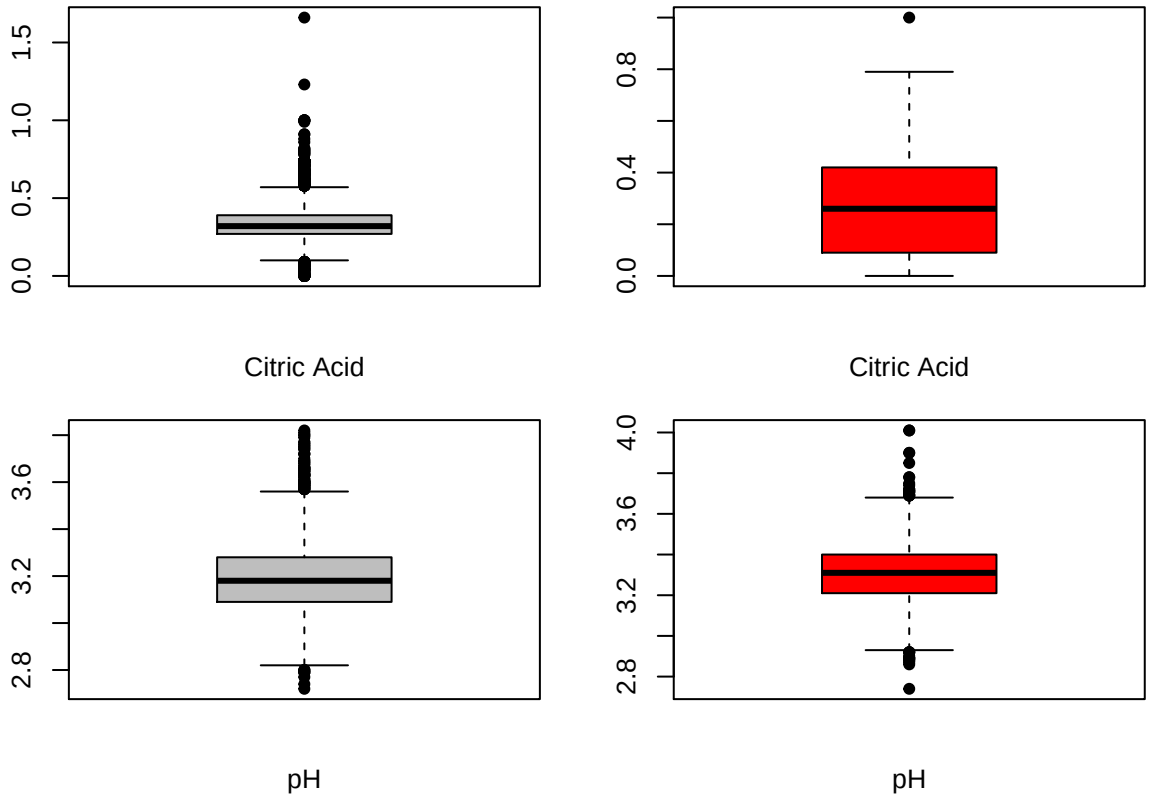
Fixed Acidity



Volatile Acidity

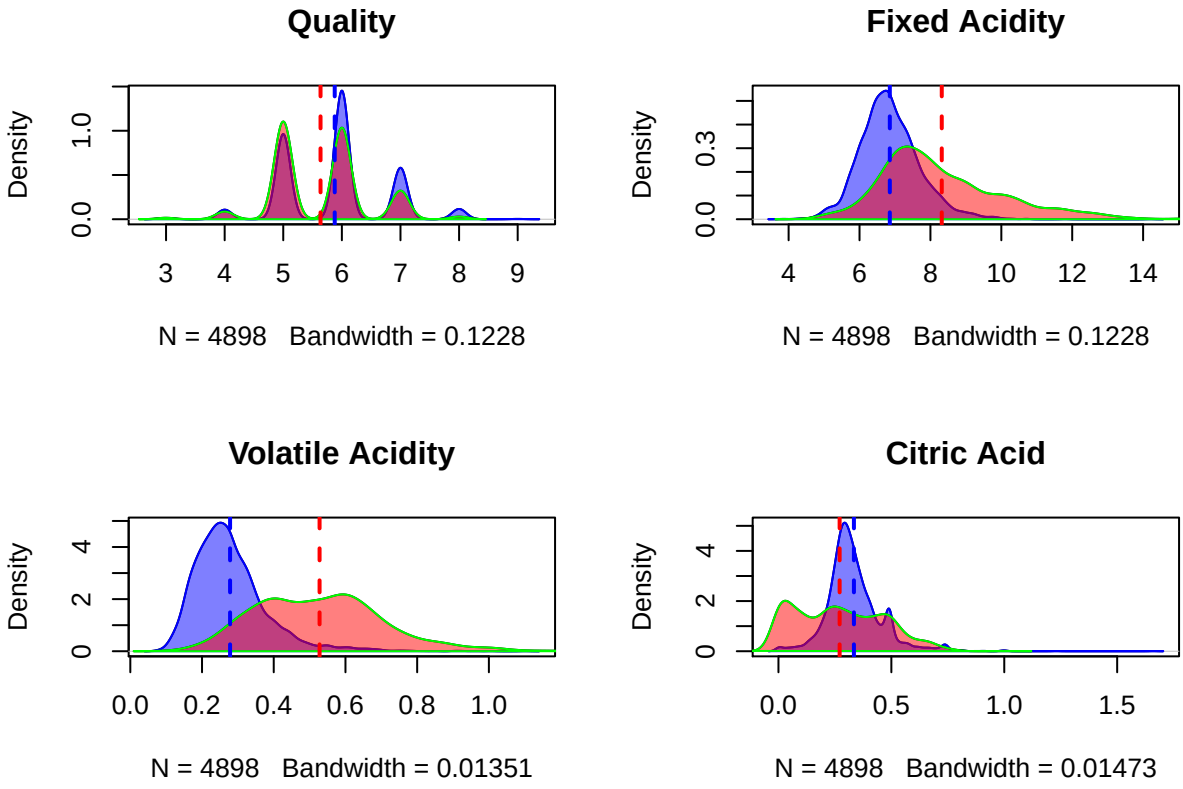


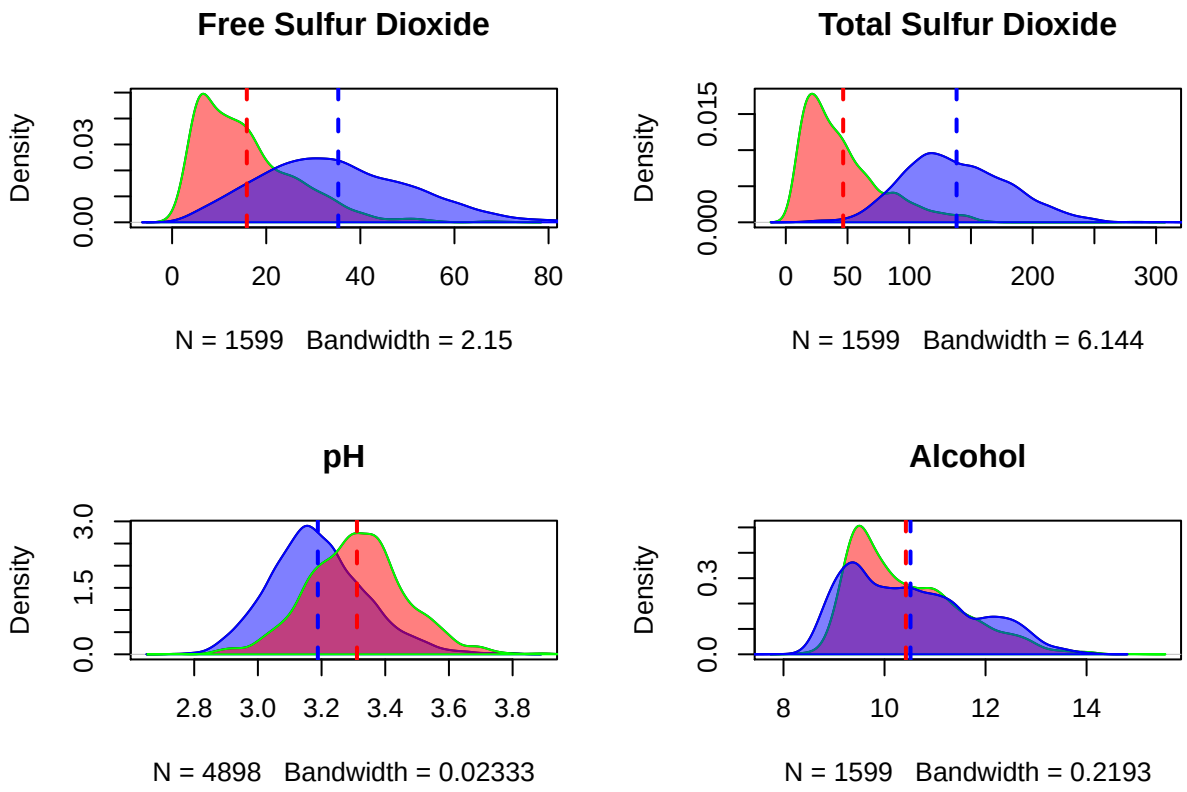
Volatile Acidity



Looks like we have a few potential outliers in the acidity variables in both red and white. Citric acid is dispersed more so in red wine. No obvious skewness in any of the variables. Let's take a look at the kernel density plots to see the distributions of these and some of the other relevant variables.

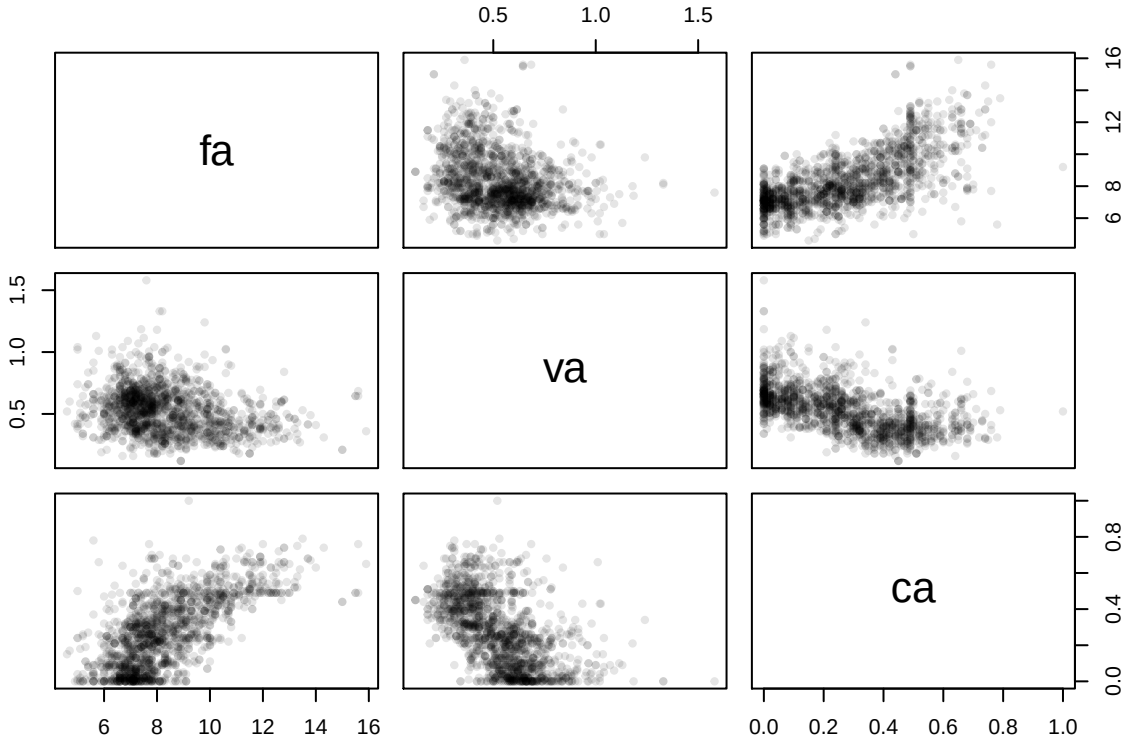
kernel density plots

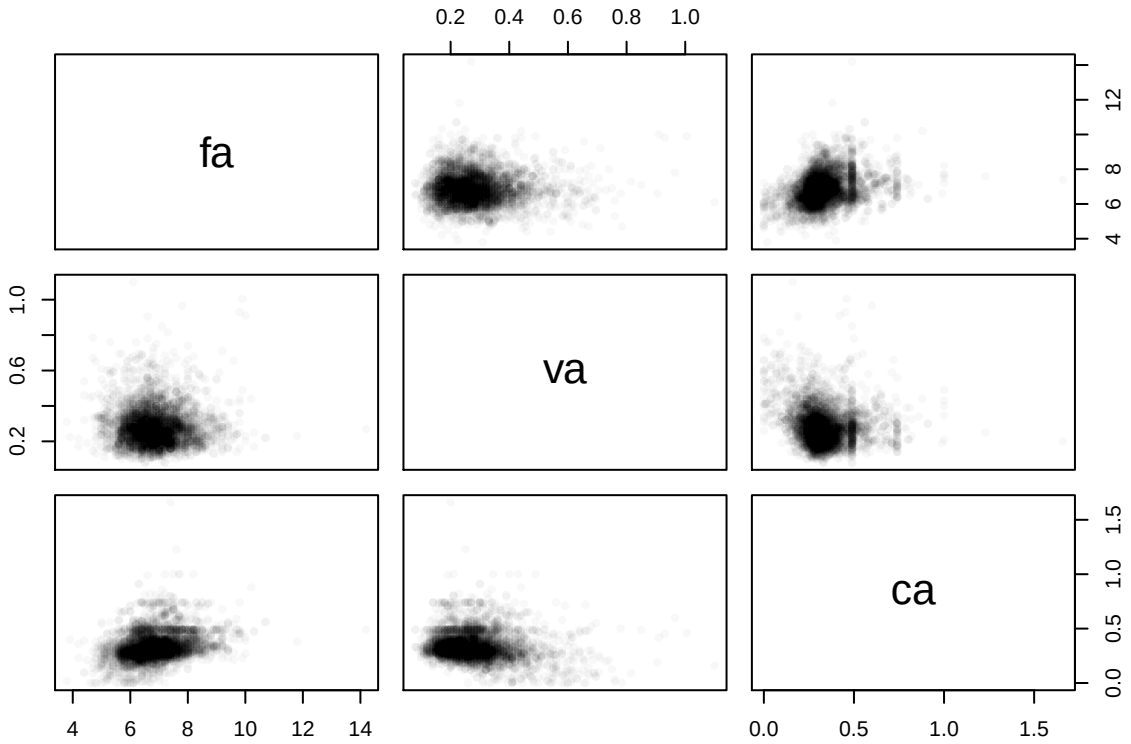




I've included the densities of predictors free sulfur dioxide, total sulfur dioxide, alcohol, as well as the variables that measure acidity. I have also included our response variable quality. I've added a line for the mean (blue line is mean of white wine; red line is mean of red wine). White wines quality tends to have more data fall into the 7+ region and have a slightly higher average score then reds. The distribution of total sulfur dioxide and free sulfur dioxide show significant differences between red and white wines, white containing around double the content. This difference can likely be explained by initial levels of antioxidants. Sulfur dioxide is added in as an antioxidant which is more important in white wines as they have less of other antioxidants than do red wines. (http://www.piwine.com/media/home-wine-making-basics/using_sulfur_dioxide.pdf) Lets go back to the acidity measurements and look at the pairwise relationships with some scatter plots.

Red(top) and White (bottom) Scatter plots





We start to see some vertical lines forming when plotting citric acid vs volatile acidity or citric acid vs fixed acidity, creating a grid like pattern. It could be the case that these are the same types of wine or come from the same grapes. There is no information in the data set on types of wine or wine brand to come to a conclusive result.
Let's try looking at some correlation tables.

correlation tables:

```
## [1] "Red Wine"
```

```
##         fa    va    ca    rs   chl   fsd   tsd   den    ph   sul   alc  qual
## fa    1.00 -0.26  0.67  0.11  0.09 -0.15 -0.11  0.67 -0.68  0.18 -0.06  0.12
## va   -0.26  1.00 -0.55  0.00  0.06 -0.01  0.08  0.02  0.23 -0.26 -0.20 -0.39
## ca    0.67 -0.55  1.00  0.14  0.20 -0.06  0.04  0.36 -0.54  0.31  0.11  0.23
## rs    0.11  0.00  0.14  1.00  0.06  0.19  0.20  0.36 -0.09  0.01  0.04  0.01
## chl   0.09  0.06  0.20  0.06  1.00  0.01  0.05  0.20 -0.27  0.37 -0.22 -0.13
## fsd  -0.15 -0.01 -0.06  0.19  0.01  1.00  0.67 -0.02  0.07  0.05 -0.07 -0.05
## tsd  -0.11  0.08  0.04  0.20  0.05  0.67  1.00  0.07 -0.07  0.04 -0.21 -0.19
## den   0.67  0.02  0.36  0.36  0.20 -0.02  0.07  1.00 -0.34  0.15 -0.50 -0.17
## ph   -0.68  0.23 -0.54 -0.09 -0.27  0.07 -0.07 -0.34  1.00 -0.20  0.21 -0.06
## sul   0.18 -0.26  0.31  0.01  0.37  0.05  0.04  0.15 -0.20  1.00  0.09  0.25
## alc  -0.06 -0.20  0.11  0.04 -0.22 -0.07 -0.21 -0.50  0.21  0.09  1.00  0.48
## qual  0.12 -0.39  0.23  0.01 -0.13 -0.05 -0.19 -0.17 -0.06  0.25  0.48  1.00
```

```
## [1] "White Wine"
```

```
##         fa    va    ca    rs   chl   fsd   tsd   den    ph   sul   alc  qual
## fa    1.00 -0.02  0.29  0.09  0.02 -0.05  0.09  0.27 -0.43 -0.02 -0.12 -0.11
## va   -0.02  1.00 -0.15  0.06  0.07 -0.10  0.09  0.03 -0.03 -0.04  0.07 -0.19
## ca    0.29 -0.15  1.00  0.09  0.11  0.09  0.12  0.15 -0.16  0.06 -0.08 -0.01
## rs    0.09  0.06  0.09  1.00  0.09  0.30  0.40  0.84 -0.19 -0.03 -0.45 -0.10
## chl   0.02  0.07  0.11  0.09  1.00  0.10  0.20  0.26 -0.09  0.02 -0.36 -0.21
## fsd  -0.05 -0.10  0.09  0.30  0.10  1.00  0.62  0.29  0.00  0.06 -0.25  0.01
## tsd   0.09  0.09  0.12  0.40  0.20  0.62  1.00  0.53  0.00  0.13 -0.45 -0.17
## den   0.27  0.03  0.15  0.84  0.26  0.29  0.53  1.00 -0.09  0.07 -0.78 -0.31
## ph   -0.43 -0.03 -0.16 -0.19 -0.09  0.00  0.00 -0.09  1.00  0.16  0.12  0.10
## sul  -0.02 -0.04  0.06 -0.03  0.02  0.06  0.13  0.07  0.16  1.00 -0.02  0.05
## alc  -0.12  0.07 -0.08 -0.45 -0.36 -0.25 -0.45 -0.78  0.12 -0.02  1.00  0.44
## qual -0.11 -0.19 -0.01 -0.10 -0.21  0.01 -0.17 -0.31  0.10  0.05  0.44  1.00
```

```
## [1] "differences between red and white"
```

```
##         fa    va    ca    rs   chl   fsd   tsd   den    ph   sul   alc  qual
## fa    0.00 -0.23  0.38  0.03  0.07 -0.10 -0.20  0.40 -0.26  0.20  0.06  0.24
## va   -0.23  0.00 -0.40 -0.06 -0.01  0.09 -0.01 -0.01  0.27 -0.23 -0.27 -0.20
## ca    0.38 -0.40  0.00  0.05  0.09 -0.16 -0.09  0.22 -0.38  0.25  0.19  0.24
## rs    0.03 -0.06  0.05  0.00 -0.03 -0.11 -0.20 -0.48  0.11  0.03  0.49  0.11
## chl   0.07 -0.01  0.09 -0.03  0.00 -0.10 -0.15 -0.06 -0.17  0.35  0.14  0.08
## fsd  -0.10  0.09 -0.16 -0.11 -0.10  0.00  0.05 -0.32  0.07 -0.01  0.18 -0.06
## tsd  -0.20 -0.01 -0.09 -0.20 -0.15  0.05  0.00 -0.46 -0.07 -0.09  0.24 -0.01
## den   0.40 -0.01  0.22 -0.48 -0.06 -0.32 -0.46  0.00 -0.25  0.07  0.28  0.13
## ph   -0.26  0.27 -0.38  0.11 -0.17  0.07 -0.07 -0.25  0.00 -0.35  0.08 -0.16
```

```
## sul   0.20 -0.23  0.25  0.03  0.35 -0.01 -0.09  0.07 -0.35  0.00  0.11  0.20
## alc   0.06 -0.27  0.19  0.49  0.14  0.18  0.24  0.28  0.08  0.11  0.00  0.04
## qual  0.24 -0.20  0.24  0.11  0.08 -0.06 -0.01  0.13 -0.16  0.20  0.04  0.00
```

We can see that quality is most strongly correlated with the alcohol content for both red and white wine, which is not hard to believe. I have also included a graphic showing the differences in correlation between red and white wines(correlation(red)- correlation(white)). The difference in correlation of alcohol content to residual sugar is fairly significant between red and white wines. There is a strong negative relationship between alcohol content and residual sugar in white wine but nearly no relationship in red. A similar situation is found between when looking at the correlation between density and residual sugar, with white wines having a much stronger correlation. We might want to look at the types of wine to learn more about the cause of these differences.

In order to predict wine quality based in its chemical properties, I will need to choose a regression model. We can first look at the red wine model with all eleven predictors.

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  21.965208449 2.119457e+01  1.0363599 3.001921e-01
## x1            0.024990553 2.594850e-02  0.9630827 3.356528e-01
## x2           -1.083590259 1.211013e-01 -8.9478019 9.872361e-19
## x3           -0.182563948 1.471762e-01 -1.2404449 2.149942e-01
## x4            0.016331270 1.500210e-02  1.0885992 2.764960e-01
## x5           -1.874225158 4.192832e-01 -4.4700697 8.373953e-06
## x6            0.004361333 2.171292e-03  2.0086353 4.474495e-02
## x7           -0.003264580 7.287285e-04 -4.4798298 8.004610e-06
## x8          -17.881163833 2.163310e+01 -0.8265650 4.086079e-01
## x9           -0.413653144 1.915974e-01 -2.1589710 3.100189e-02
## x10           0.916334413 1.143375e-01  8.0142971 2.127228e-15
## x11           0.276197699 2.648359e-02 10.4290143 1.123029e-24
```

We see from this model the coefficients of predictors fixed acidity, citric acid, residual sugar, and density are not statically significant (α= .05). To preform model selection, I used the LASSO method. LASSO is a model selection technique that shrinks certain coefficients of our model to zero, improving prediction accuracy and interpretability. Which variables get shrunk to zero is dependent on the tuning parameter λ, which I find using cross validation. I find the λ value that minimizes the mean error is .0077 for red wine. Using the glmnet() function I find the following model:

Red LASSO

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                s0
## V1   .
## V2  -1.0250289730
## V3   .
## V4   0.0007938082
## V5  -1.6808773647
## V6   0.0022193371
## V7  -0.0025990141
## V8   .
## V9  -0.3740801288
## V10  0.8137196641
## V11  0.2849033886
```

We can compare this to our MLR with all eleven predictors. Lasso has shrunk fixed acidity, citric acid, and density to zero in our model. Let's do the same thing with white wine.

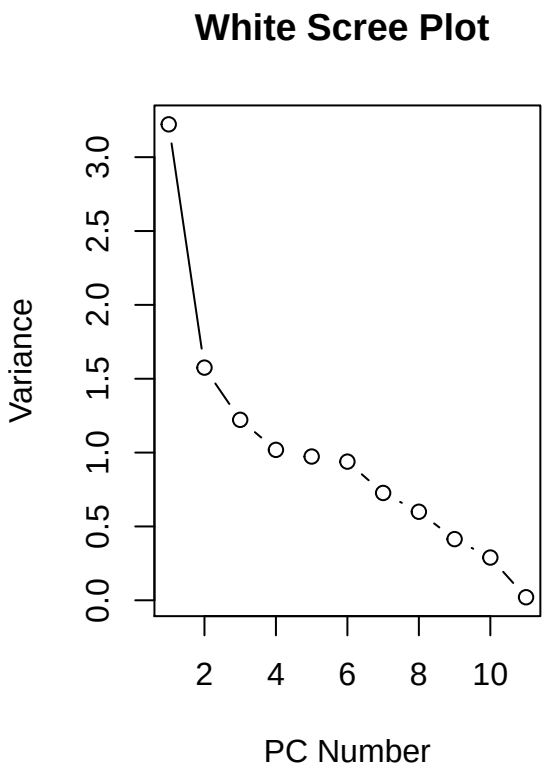
White LASSO

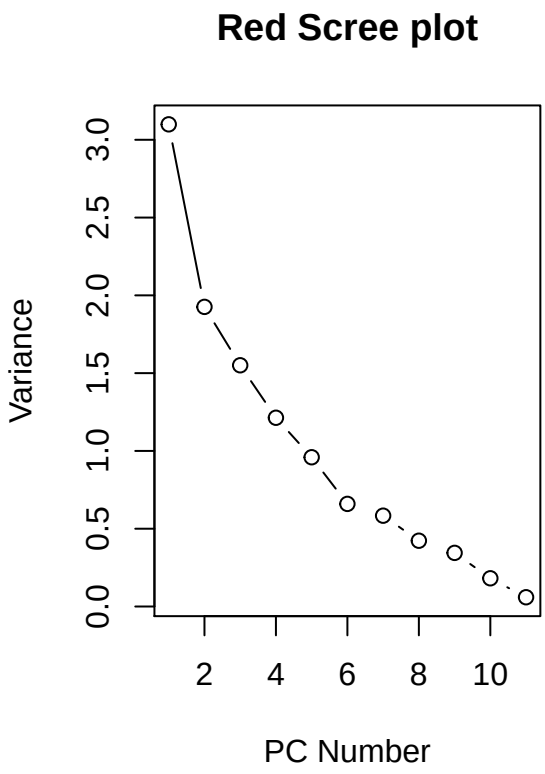
```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                s0
## V1   3.370687e-02
## V2  -1.874662e+00
## V3   .
## V4   6.785531e-02
## V5  -3.781317e-01
## V6   3.615447e-03
## V7  -2.401579e-04
## V8  -1.165558e+02
## V9   5.370167e-01
## V10  5.632926e-01
## V11  2.291073e-01
```

Comparing this to our White MLR, only citric acid variable is shrunk to zero. Coefficients citric acid, chlorides, and total sulfur dioxide are not statistically significant in the White MLR model.

Principal Componets

Principal component analysis (PCA) is used to reduce the dimensions of a data matrix. When there are a large number of predictors, PCA allows us to create a simpler way of summarize the X's without getting rid of much information. Using the prcomp() function I broke red and white wine down into eleven principal components. The scree plots are below.





As seen by the scree plot, the variance does not drop to zero. Looking at the summary table of the first eight components capture about 95% of the variance. Using the PCs as my X's in MLR I get the following model

PC Red

```
##                Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept)  5.63602251 0.016205345 347.7878723 0.000000e+00
## PC1          0.05062084 0.009208186   5.4973745 4.484918e-08
## PC2         -0.22508738 0.011680896 -19.2697013 1.619396e-74
## PC3         -0.25894571 0.013018236 -19.8909978 8.773966e-79
## PC4         -0.03237640 0.014717095  -2.1999176 2.795645e-02
## PC5         -0.08372128 0.016550789  -5.0584463 4.717994e-07
## PC6          0.02511361 0.019959566   1.2582242 2.084958e-01
## PC7          0.09525749 0.021216090   4.4898701 7.640954e-06
## PC8          0.08604865 0.024925632   3.4522154 5.705991e-04
## PC9          0.14200045 0.027612769   5.1425646 3.046844e-07
## PC10         0.09441745 0.038067601   2.4802575 1.323180e-02
## PC11        -0.06465991 0.066423681  -0.9734467 3.304797e-01
```

PC6(which may be measuring something related to acidity) and PC11 (unknown relationship) are not statistically significant (α= .05)

PC White

```
##                 Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept)  5.877909351 0.010735861 547.5023713  0.000000e+00
## PC1w        -0.146506296 0.005981379 -24.4937313 4.567948e-125
## PC2w         0.041008809 0.008554761   4.7936828  1.686062e-06
## PC3w        -0.174888433 0.009714133 -18.0035049  3.252620e-70
## PC4w         0.167386902 0.010638881  15.7335069  1.897220e-54
## PC5w        -0.042308540 0.010883038  -3.8875671  1.025968e-04
## PC6w         0.003660448 0.011081746   0.3303132  7.411774e-01
## PC7w        -0.022084861 0.012596042  -1.7533175  7.961023e-02
## PC8w        -0.183107625 0.013868768 -13.2028758  3.924281e-39
## PC9w        -0.357665581 0.016684212 -21.4373670  1.640922e-97
## PC10w        0.108756277 0.019955682   5.4498902  5.287630e-08
## PC11w       -0.480295045 0.074718958  -6.4280212  1.415966e-10
```

For the White Model PC6 and PC7 (unknown relationships) are not statistically significant (α= .05)

Generalized Additive Models

Generalized additive models (GAM) allow for nonlinear functions of each of our predictors while maintaining additivity. I apply the GAM model with all eleven predictors for red and white wine.

Red

```
## Anova for Nonparametric Effects
##                     Npar Df  Npar F     Pr(F)
## (Intercept)
## s(wine.red$fa, 4)         3  3.7217  0.011048 *
## s(wine.red$va, 4)         3  1.9932  0.113083
## s(wine.red$ca, 4)         3  2.8406  0.036714 *
## s(wine.red$rs, 4)         3  1.6058  0.186125
## s(wine.red$chl, 4)        3  2.3446  0.071258 .
## s(wine.red$fsd, 4)        3  3.0062  0.029351 *
## s(wine.red$tsd, 4)        3  3.2367  0.021460 *
## s(wine.red$den, 4)        3  1.9323  0.122401
## s(wine.red$ph, 4)         3  0.6279  0.597019
## s(wine.red$sul, 4)        3 22.9579 1.532e-14 ***
## s(wine.red$alc, 4)        3  4.8324  0.002369 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our ANOVA nonparametric effects output of the GAM() function we see that Sulphates and alcohol are significant at the α= .01 level, meaning we may have some nonlinear effects with these variables. Citric acid and Residual sugar are nonsignificant for linear and nonlinear effects.

White

We see nearly the opposite case where only sulphates is nonsignificant for nonlinear effects yet all other variables are. Citric acid is nonsignificant for linear effects again.

It appears that Citric acid does not play a significant role in determining quality in either red or white wine, while Alcohol content and Sulphates appear to have a significant influence over quality. pH plays more of an important role in determining quality in red wine than it does in white.

I combined the two data sets into one data set and created a binary variable qual6 which is zero when quality is less than 6 and 1 otherwise. I also created variable for type (red or white). We use this to run some logistic regressions.

I tried a few differnet logistic regression models, but the model with the highest AUC was still the full model with all eleven predictors. Here are some interpritations of the coefficents.

Interpretations of coefficients:

Fixed Acidity - For each additional g(tartaric acid)/dmˆ3, the odds of quality being less than 6 increases by 11 percent.

Volatile Acidity - For each additional g(acetic acid)/dmˆ3, the odds of quality being less than 6 are lower by .99 percent.

Citric Acid - For each additional g(citric acid)/dmˆ3 the odds of quality being less than 6 are lower by .39 percent.

My best achieved logistic model AUC was .804 when using all 12 predictors. I was able to achieve a slightly better AUC when using random forests and got an AUC of .9139.THe ROC curves are below.

ROC Curves