Project 1 - Wine Quality

Moustapha Dieng

February 18, 2018

Part I. Regression

Abstract: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009]

Description: In this project, I will be combining the two datasets and studying the effects of some attributes on the wine quality. The wines are ranked on a scale from 0 to 10.

Source: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 Download datasets here.

Initial Setup

```
# Load required librariess
library(ggplot2)
library(gridExtra)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(class)
library(DAAG)
```

I will be combining both datasets for the regression part of this project.

```
white <- read.csv("data/winequality-white.csv", header=TRUE, sep=";") # Read in white wine csv file.
red <- read.csv("data/winequality-red.csv", header=TRUE, sep=";") # Read in red wine csv file
wine1 <- rbind(white, red) # Combines the two data frames by rows.
```

Examining the Dataset

Structure of the dataset

```
str(wine1) # Structure of dataset
```

```
## 'data.frame':    6497 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
```

```
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

From the structure, we can see the dimension of the dataset, names and types of the attributes as well as a preview of their values.

Columns description

- Fixed acidity (g(tartaric acid)/dm3): The predominant fixed acids found in wines are tartaric, malic, citric, and succinic.
- Volatile acidity (g(acetic acid)/dm3): A measure of the wine's volatile (or gaseous) acids. The primary volatile acid in wine is acetic acid, which is also the primary acid associated with the smell and taste of vinegar.
- Citric acid (g/dm3): Citric acid is a weak organic acid, which is often used as a natural preservative or additive to food or drink to add a sour taste to food.
- Residual sugar (g/dm3): How much sugar is left in the wine after fermentation is complete. The amount of residual sugar tells you how sweet the wine is going to be.
- Chlorides (g(sodium chloride)/dm3): The amount of chlorides ions in the wine.
- Free sulfur dioxide (mg/dm3): Free sulfites are those available to react and thus exhibit both germicidal and antioxidant properties.
- Total sulfur dioxide (mg/dm3): Free and bound sulfites. The bound sulfites are those that have reacted (both reversibly and irreversibly) with other molecules within the wine medium.
- Density (g/cm3)
- pH
- Sulphates (g(potassium sulphate)/dm3): Preservatives that are widely used in winemaking (and most food industries) for its antioxidant and antibacterial properties.
- Alcohol (vol.%)
- Quality (score between 0 and 10)

Target column

The target column will be quality.

Summary of dataset

```r
summary(wine1) # Structure of dataset
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100
##  free.sulfur.dioxide total.sulfur.dioxide    density             pH           sulphates
##  Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300
##  Median : 29.00      Median :118.0        Median :0.9949   Median :3.210   Median :0.5100
```

```
##  Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313
##  3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000
##  Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000
##     alcohol          quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.30   Median :6.000
##  Mean   :10.49   Mean   :5.818
##  3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :14.90   Max.   :9.000
```

The summary provides useful statistics. For example, we can see that the minimum and maximum scores given to wines are 3 and 9 respectively. From the description of the dataset, we are provided with the information that wines are scored on a 0 to 10 scale.

Preview of dataset

```r
head(wine1) # Preview of dataset
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
## 1           7.0             0.27        0.36           20.7     0.045                  45
## 2           6.3             0.30        0.34            1.6     0.049                  14
## 3           8.1             0.28        0.40            6.9     0.050                  30
## 4           7.2             0.23        0.32            8.5     0.058                  47
## 5           7.2             0.23        0.32            8.5     0.058                  47
## 6           8.1             0.28        0.40            6.9     0.050                  30
##   total.sulfur.dioxide density   pH sulphates alcohol quality
## 1                  170  1.0010 3.00      0.45     8.8       6
## 2                  132  0.9940 3.30      0.49     9.5       6
## 3                   97  0.9951 3.26      0.44    10.1       6
## 4                  186  0.9956 3.19      0.40     9.9       6
## 5                  186  0.9956 3.19      0.40     9.9       6
## 6                   97  0.9951 3.26      0.44    10.1       6
```

Edge cases

```r
# Display number of wines with poor rating
paste("Number of poor quality wines: ", sum(wine1$quality == 3))
```

```
## [1] "Number of poor quality wines:  30"
```

```r
# Display number of wines with excellent rating
paste("Number of excellent quality wines: ", sum(wine1$quality == 9))
```

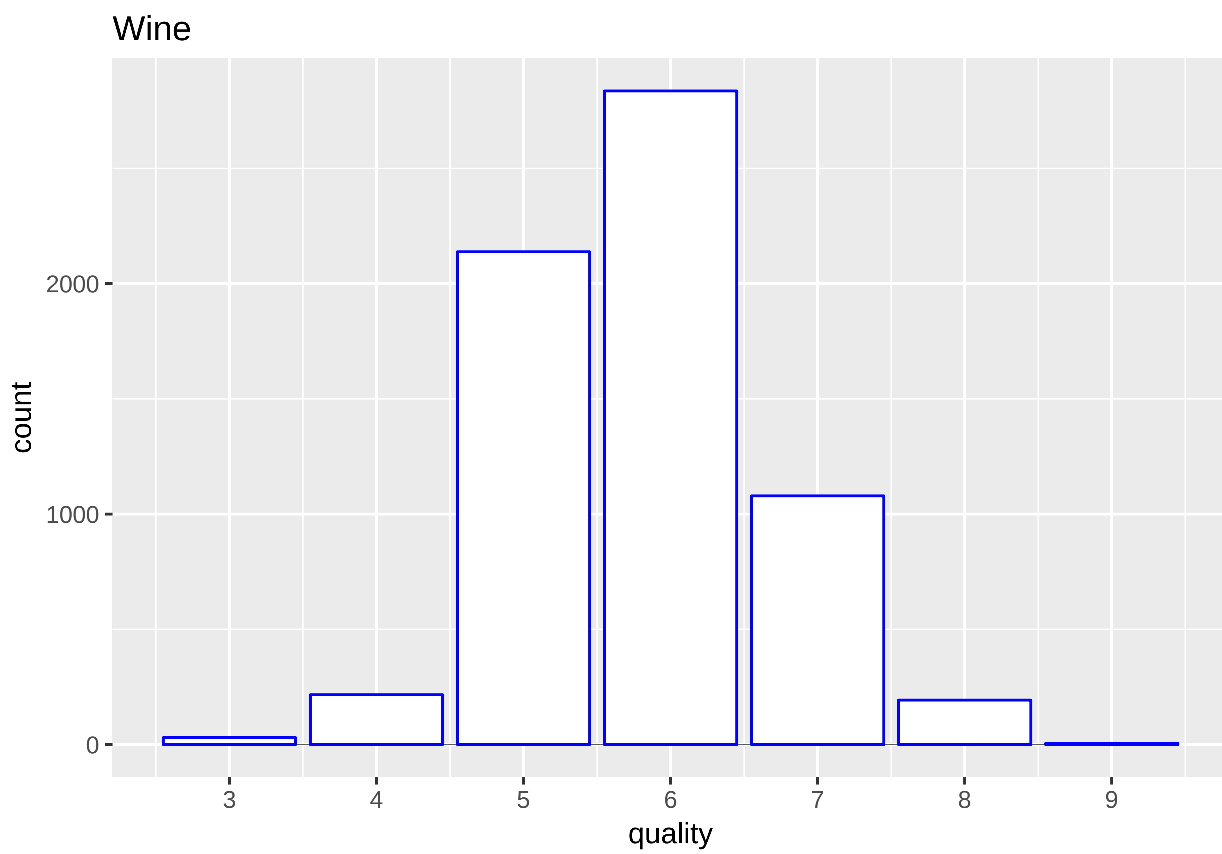
```
## [1] "Number of excellent quality wines:  5"
```

From these results, we can see that very few wines are rated as poor (score of 3) and even fewer as excellent (score of 9).

Graphs

Histogram

```r
ggplot(wine1, aes(x = quality)) +
  geom_histogram(binwidth = 0.5) +
  geom_bar(color = "blue", fill = "white") +
  scale_x_continuous(breaks = round(seq(min(wine1$quality), max(wine1$quality), by = 1),1)) +
  labs(title = "Wine") # Bar plot for wine scores
```



With the graph, we can visualize the summary of the quality scores of the wines. We can easily see that the mean is between 5 and 6 for the whole dataset.

Correlation plot

The correlation plot will be saved to then loaded from a png file for better readibility.

```
png(height = 600, width = 800, file = "corrPlot.png")
corr <- cor(wine1) # Compute matrix of correlation
corrplot(corr, method = "color", addCoef.col = "grey") # Save plot to png file for better readibility
dev.off() # Shutdown currrent graphics device
```

```
## pdf
##   2
```

From the correlation plot, we can observe a few strong correlations: - Residual.sugar and density @ 0.55 - free.sulfur.dioxide and total.sulfur.dioxide @ 0.72 - alcohol and density @ -0.69 It's also interesting to note that quality has a somewhat strong positive correlation with only one other variable, alcohol @ 0.44.

Algorithms

Linear Regression: Model 1 - quality~alcohol

```
set.seed(1234) # Set seed to 1234 to ensure reproducibility of results
i <- sample(nrow(wine1), nrow(wine1)*.75, replace=FALSE) # Sample from dataframe
train <-wine1[i,] # Initiliaze train set
test <- wine1[-i,] # Initialize test set
```

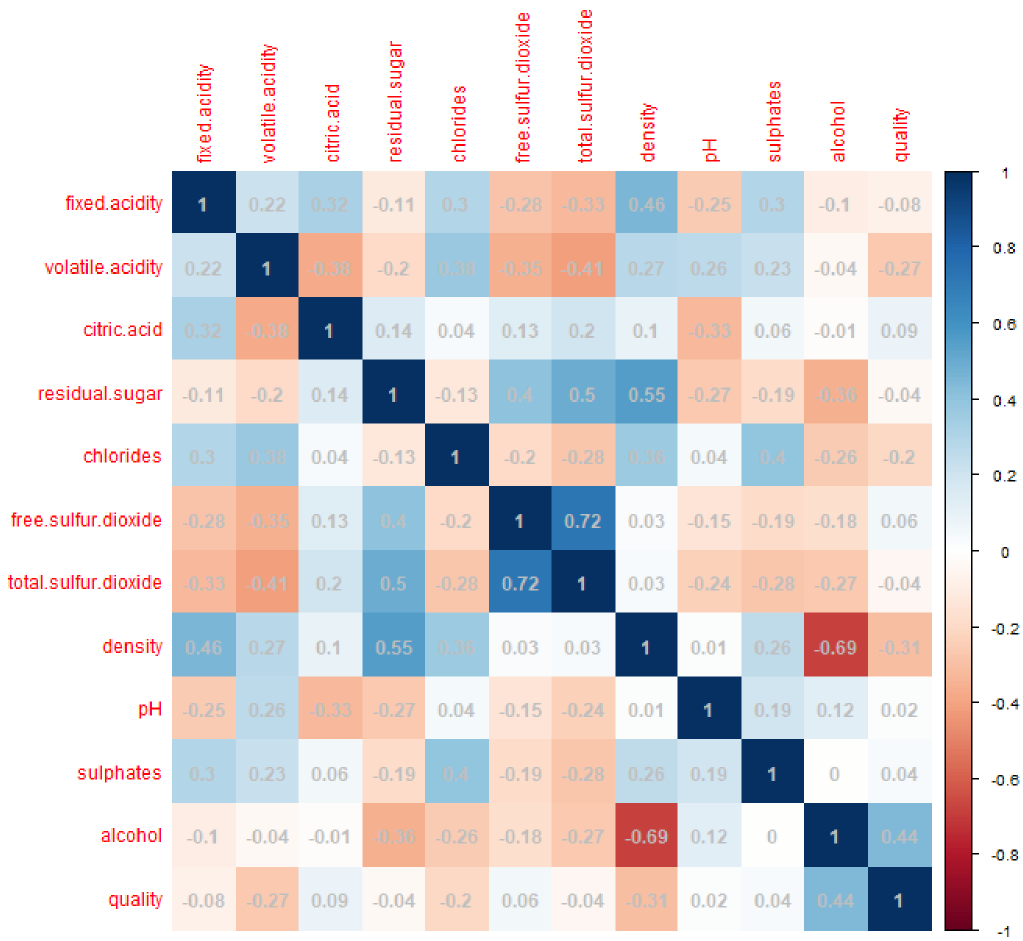

fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality

Figure 1:

```r
lm1 <- lm(quality~alcohol, data=train) # Create linear model lm1 from train set
summary(lm1) # Summary of linear model
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5037 -0.4904 -0.0461  0.5096  2.8365
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.38506    0.09846   24.22   <2e-16 ***
## alcohol      0.32688    0.00932   35.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7805 on 4870 degrees of freedom
## Multiple R-squared:  0.2016, Adjusted R-squared:  0.2015
## F-statistic:  1230 on 1 and 4870 DF,  p-value: < 2.2e-16
```

The p-value being much less than significance level of 0.05 indicates that there exists a strong relationship between quality and alcohol. However, looking at the R-squared which is only .2016, we know that only 20.16% of the total variation in quality can be explained by the linear relationship between quality and alcohol. We will therefore come up with a new linear model after calculating the MSE of our current model.

Mean Squared Error of lm1:

```r
mean(lm1$residuals^2) # Display MSE of lm1
```

```
## [1] 0.608906
```

Linear Regression: Model 2 - quality~.

```r
lm2 <- lm(quality~., data = train) # Create linear model lm2 from train set
summary(lm2) # Summary of linear model
```

```
##
## Call:
## lm(formula = quality ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7227 -0.4549 -0.0408  0.4635  2.7489
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.637e+01  1.353e+01   4.904 9.71e-07 ***
## fixed.acidity          8.411e-02  1.796e-02   4.684 2.89e-06 ***
## volatile.acidity      -1.314e+00  8.909e-02 -14.745  < 2e-16 ***
## citric.acid           -8.664e-02  9.131e-02  -0.949    0.343
## residual.sugar         4.901e-02  5.939e-03   8.252  < 2e-16 ***
## chlorides             -3.473e-01  3.746e-01  -0.927    0.354
```

```
## free.sulfur.dioxide   6.069e-03  8.559e-04   7.090 1.53e-12 ***
## total.sulfur.dioxide -2.627e-03  3.165e-04  -8.300  < 2e-16 ***
## density              -6.592e+01  1.382e+01  -4.771 1.88e-06 ***
## pH                    5.243e-01  1.037e-01   5.058 4.40e-07 ***
## sulphates             6.579e-01  8.779e-02   7.494 7.87e-14 ***
## alcohol               2.594e-01  1.901e-02  13.645  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7329 on 4860 degrees of freedom
## Multiple R-squared:  0.2975, Adjusted R-squared:  0.2959
## F-statistic: 187.1 on 11 and 4860 DF,  p-value: < 2.2e-16
```

By analyzing our summary, we can see that all variables but citric.acid and chlorides have strong significance. Moreover, our R-squared is now at 0.2975 vs .2016 for model 1 which shows that this model explain the variation in quality better.

Mean Squared Error of lm2:

```r
mean(lm2$residuals^2) # Display MSE of lm2
```

```
## [1] 0.5357894
```

The MSE decreased from 0.608906 for model 1 to .537894 which demonstrates that the second model is better than the first.

Linear Regression: Model 3 - quality~.-citric.acid-chlorides

```r
lm3 <- lm(quality~.-citric.acid-chlorides, data=train) # Create linear model lm3 from train set
summary(lm3) # Summary of linear model
```

```
##
## Call:
## lm(formula = quality ~ . - citric.acid - chlorides, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7069 -0.4575 -0.0424  0.4637  2.7597
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.927e+01  1.326e+01   5.224 1.82e-07 ***
## fixed.acidity         8.250e-02  1.725e-02   4.783 1.78e-06 ***
## volatile.acidity     -1.295e+00  8.155e-02 -15.875  < 2e-16 ***
## residual.sugar        5.022e-02  5.791e-03   8.673  < 2e-16 ***
## free.sulfur.dioxide   6.048e-03  8.539e-04   7.083 1.62e-12 ***
## total.sulfur.dioxide -2.648e-03  3.104e-04  -8.531  < 2e-16 ***
## density              -6.895e+01  1.352e+01  -5.098 3.56e-07 ***
## pH                    5.523e-01  1.015e-01   5.442 5.54e-08 ***
## sulphates             6.345e-01  8.588e-02   7.389 1.74e-13 ***
## alcohol               2.578e-01  1.894e-02  13.607  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7329 on 4862 degrees of freedom
```

```
## Multiple R-squared:  0.2972, Adjusted R-squared:  0.2959
## F-statistic: 228.5 on 9 and 4862 DF,  p-value: < 2.2e-16
```

Although all the variables in this model are significant, its R-squared is slightly lower than the second linear model .2972 vs .2975. So model 2 is still better in that aspect.

Mean Squared Error of lm3:

```r
mean(lm3$residuals^2) # Display MSE of lm3
```

```
## [1] 0.5360198
```

Model 3 MSE is slightly higher than model 2's .5360198 vs 0.5357894. In conclusion, model 2 is the clear winner as far as linear regression models are concerned. We can easily verify this with an analysis of variance table.

Anova

```r
anova(lm1, lm2, lm3)
```

```
## Analysis of Variance Table
##
## Model 1: quality ~ alcohol
## Model 2: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
## Model 3: quality ~ (fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol) - citric.acid - chlorides
##   Res.Df    RSS Df Sum of Sq       F Pr(>F)
## 1   4870 2966.6
## 2   4860 2610.4 10    356.22 66.3221 <2e-16 ***
## 3   4862 2611.5 -2     -1.12  1.0448 0.3518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, the anova results show model 2 having the lowest RSS.

Correlation of lm2

```r
pred <- predict(lm2, newdata=test) # Predict results based on test data
paste("Correlation for linear model 2: ", cor(pred, test$quality) * 100, "%") # Display correlation
```

```
## [1] "Correlation for linear model 2:  52.3551096300085 %"
```

Knn Regression

Choosing k: 10-fold cross-validation

```r
scaled_df <- data.frame(scale(wine1[, 1:12])) # Scale the data first
scaled_train1 <- scaled_df[i,] # Set train set
scaled_test1 <- scaled_df[-i,] # Set test set
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3) # Controls the computational nu
knn_fit <- train(quality~., data = scaled_train1, method = "knn",
 trControl=trctrl,
 preProcess = c("center", "scale"),
 tuneLength = 10) # Train the classifier
knn_fit # Checking results
```

```
## k-Nearest Neighbors
##
## 4872 samples
##   11 predictor
##
## Pre-processing: centered (11), scaled (11)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 4384, 4385, 4385, 4385, 4385, 4385, ...
## Resampling results across tuning parameters:
##
##   k   RMSE       Rsquared   MAE
##    5  0.8101316  0.3604485  0.6125434
##    7  0.7988927  0.3692147  0.6145831
##    9  0.7968764  0.3702581  0.6164263
##   11  0.7991591  0.3658126  0.6199600
##   13  0.7992014  0.3649383  0.6212856
##   15  0.8000629  0.3632530  0.6237091
##   17  0.8013424  0.3610918  0.6257696
##   19  0.8025709  0.3591034  0.6274675
##   21  0.8033286  0.3579150  0.6285386
##   23  0.8028368  0.3590846  0.6289210
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

Knn regression with k = 9

```
knnreg1 <- knnreg(scaled_train1[, 1:11],scaled_train1[, 12], k = 9) # 1:11 for training set predictors
```

Knn is not model-based therefore no summary is displayed

According to the cross-validation, our optimal k = 9.

Correlation of Knn Regression

```
pred1 <- predict(knnreg1, scaled_test1[, 1:11]) # Predict results
paste("Correlation for knn, k = 9: ", cor(pred1, test$quality) * 100, "%") # Displays correlation resul
```

```
## [1] "Correlation for knn, k = 9:  56.9386911547391 %"
```

Final comments

Knn with k = 9 has better correlation results than lm2 (formula(quality~.)). Knn with k = 9 correlation is around 56.94% while lm2 has correlation is around 52.36%. It doesn't seem possible to reliably predict the wine quality given the predictors we used.

Part II. Classification

I will be using the same dataset for the classification part but I will be adding a categorical variable to differenciate the wines by color.

Add color attribute

```
white$color <- 'white' # Add categorical attribute 'color' to white wine dataset
white$color <- as.factor(white$color) # Transfrom the attribute color into a factor
red$color <- 'red' # Add categorical attribute 'color' to red wine dataset
```

```r
red$color <- as.factor(red$color) # Transfrom the attribute color into a factor
wine2 <- rbind(white, red) # Combines the two data frames by rows.
```

Examining the Dataset

Structure of the dataset

```r
str(wine2) # Structure of dataset
```

```
## 'data.frame':    6497 obs. of  13 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ color               : Factor w/ 2 levels "white","red": 1 1 1 1 1 1 1 1 1 1 ...
```

Columns description

The columns are the same as in part I. However, I have added the 'color' column to categorize the wines by color.

Target column

The target column will be color. I will be attempting to predict the wine color using all other variables as predictors.

Summary of dataset

```r
summary(wine2) # Summary of dataset
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100
##  free.sulfur.dioxide total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300
##  Median : 29.00      Median :118.0        Median :0.9949   Median :3.210   Median :0.5100
##  Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313
##  3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000
##  Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000
##     alcohol         quality         color
##  Min.   : 8.00   Min.   :3.000   white:4898
##  1st Qu.: 9.50   1st Qu.:5.000   red  :1599
##  Median :10.30   Median :6.000
```

```
##  Mean   :10.49   Mean   :5.818
##  3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :14.90   Max.   :9.000
```

Note: The only difference is the added column 'color'. The red wines account for about 1/3 of the data.

Preview of dataset

```r
wine2[sample(nrow(wine2), 10), ] # Random sample of dataset
```

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
## 859            6.7             0.22        0.39           10.2     0.038                  60
## 1022           8.6             0.20        0.42            1.5     0.041                  35
## 3547           6.6             0.23        0.37            8.5     0.036                  46
## 1790           7.8             0.20        0.24            1.6     0.026                  26
## 906            8.4             0.19        0.42            1.6     0.047                   9
## 3271           6.4             0.15        0.29            1.8     0.044                  21
## 4614           5.8             0.27        0.20            7.3     0.040                  42
## 2851           6.7             0.24        0.29           14.9     0.053                  55
## 99             9.8             0.36        0.46           10.5     0.038                   4
## 2382           7.0             0.23        0.42            5.1     0.042                  37
##      total.sulfur.dioxide density   pH sulphates alcohol quality color
## 859                   149 0.99725 3.17      0.54    10.0       7 white
## 1022                  125 0.99250 3.11      0.49    11.4       7 white
## 3547                  153 0.99576 3.20      0.48     9.4       6 white
## 1790                  189 0.99100 3.08      0.74    12.1       7 white
## 906                   101 0.99400 3.06      0.65    11.1       4 white
## 3271                  115 0.99166 3.10      0.38    10.2       5 white
## 4614                  145 0.99442 3.15      0.48     9.8       5 white
## 2851                  136 0.99839 3.03      0.52     9.0       5 white
## 99                     83 0.99560 2.89      0.30    10.1       4 white
## 2382                  144 0.99518 3.50      0.59    10.2       6 white
```

Best/worst wines

```r
# Find the percentage of highest rated wine that are white
paste("White wines make up ",
      sum(wine2[which(wine2$quality == 9), 13] == 'white') /
        nrow(wine2[which(wine2$quality == 9), ]) * 100,
      "% of the excellent wines.")
```

```
## [1] "White wines make up  100 % of the excellent wines."
```

```r
# Find the percentage of worst rated wine that are white
paste("White wines make up ",
      sum(wine2[which(wine2$quality == 3), 13] == 'white') /
        nrow(wine2[which(wine2$quality == 3), ]) * 100,
      "% of the worst wines.")
```

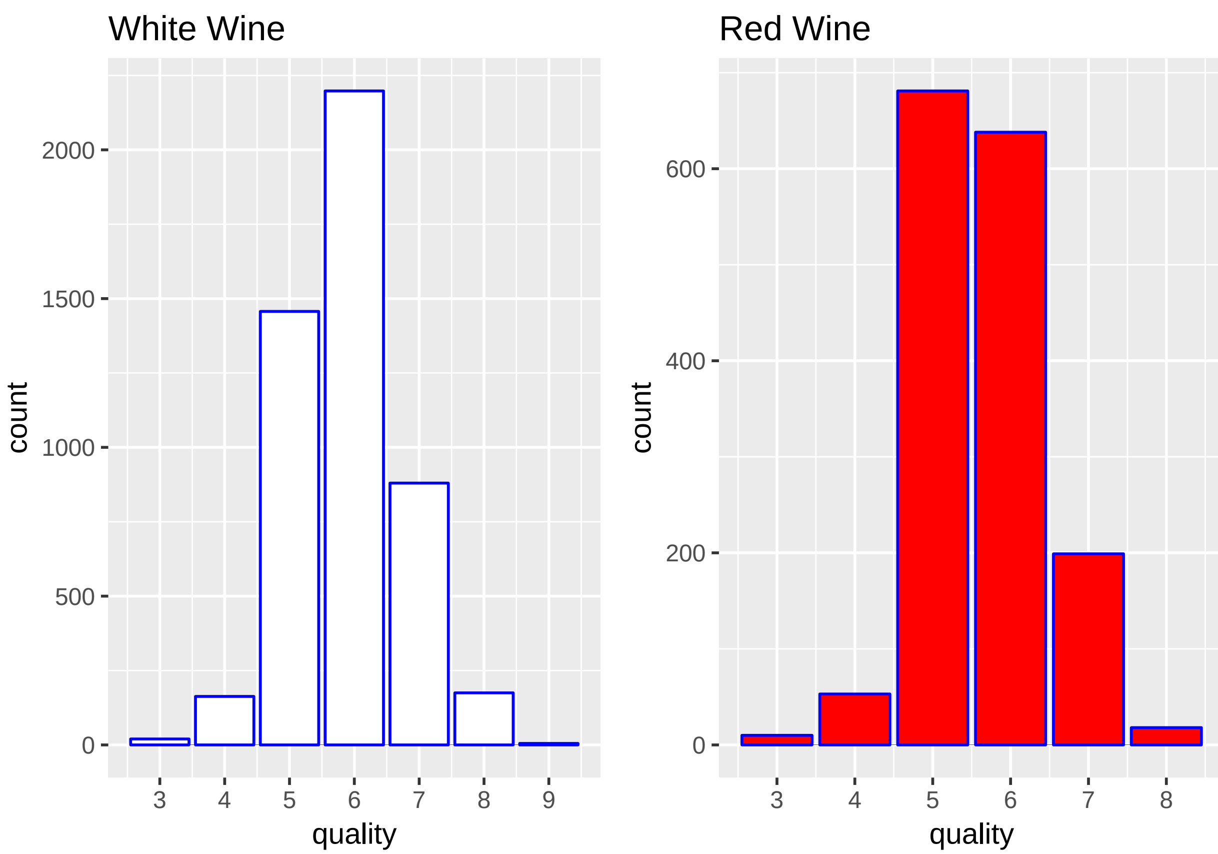
```
## [1] "White wines make up  66.6666666666667 % of the worst wines."
```

Graphs

Histogram

```r
plot_white <- ggplot(white, aes(x = quality)) +
 geom_histogram(binwidth = 0.5) +
 geom_bar(color = "blue", fill = "white") +
```

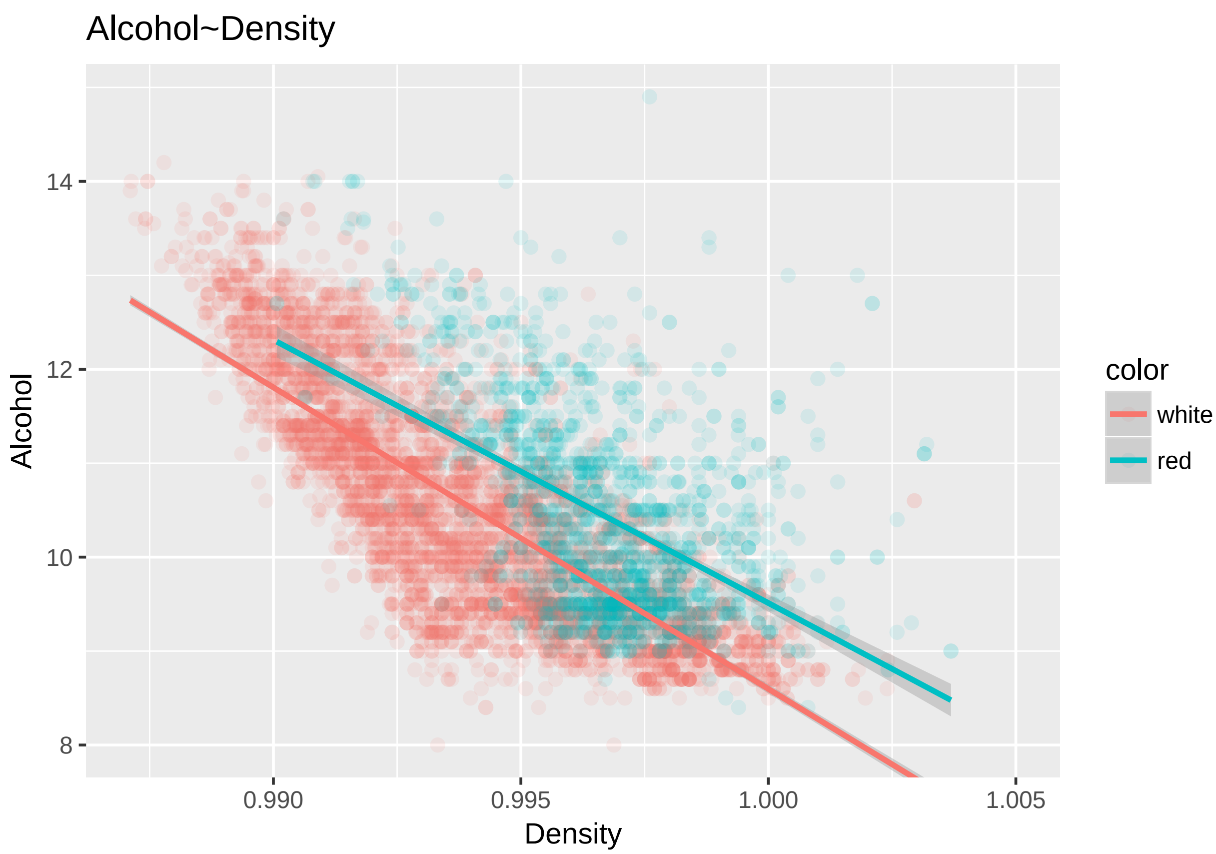
```
  scale_x_continuous(breaks = round(seq(min(wine2$quality), max(wine2$quality), by = 1), 1)) +
  labs(title = "White Wine") # Histogram for white wine scores
plot_red <- ggplot(red, aes(x = quality)) +
  geom_histogram(binwidth = 0.5) +
  geom_bar(color = "blue", fill = "red") +
  scale_x_continuous(breaks = round(seq(min(wine2$quality), max(wine2$quality), by = 1), 1)) +
  labs(title = "Red Wine")  # Histogram for red wine scores
grid.arrange(plot_white, plot_red, ncol = 2)  # Arrange plot side by side
```



From the graph, we can visualize the distribution of scores by wine color. There are a few distinctions to note which can help in identifying the type of wine.

Plot of Alcohol~Density

```
ggplot(wine2, aes(x = density, y = alcohol, color = color)) +
  geom_point(alpha = 0.1, position = position_jitter(h = 0), size = 2) +
  geom_smooth(method = 'glm') +
  coord_cartesian(xlim=c(min(wine2$density), 1.005), ylim=c(min(wine2$alcohol), max(wine2$alcohol))) +
  xlab('Density') +
  ylab('Alcohol') +
  ggtitle('Alcohol~Density')
```



The graph shows that red wines tend to be denser than white wines with the same alcohol volume.

Algorithms

Logistical Regression

```
set.seed(1234) # Set seed
i <- sample(nrow(wine2), nrow(wine2)*.67, replace=FALSE) # Sample from df
train <- wine2[i,] # Initialize train set
test <- wine2[-i,] # Initialize test set
glm1 <- glm(color~., data = train, family = binomial) # Create logistical regression
summary(glm1) # Summary of model
```

```
##
## Call:
## glm(formula = color ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.1012  -0.0288  -0.0032   0.0001   4.4069
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.252e+03  2.538e+02  -8.873  < 2e-16 ***
## fixed.acidity        -5.677e-01  2.812e-01  -2.019   0.0435 *
```

```
## volatile.acidity      7.344e+00  1.552e+00   4.733 2.22e-06 ***
## citric.acid          -4.171e+00  1.694e+00  -2.462   0.0138 *
## residual.sugar       -1.559e+00  1.749e-01  -8.917  < 2e-16 ***
## chlorides             2.400e+01  5.013e+00   4.788 1.68e-06 ***
## free.sulfur.dioxide   9.592e-02  1.719e-02   5.579 2.41e-08 ***
## total.sulfur.dioxide -6.573e-02  7.569e-03  -8.684  < 2e-16 ***
## density               2.260e+03  2.586e+02   8.740  < 2e-16 ***
## pH                   -4.050e+00  1.873e+00  -2.163   0.0306 *
## sulphates             3.590e+00  1.646e+00   2.181   0.0292 *
## alcohol               2.087e+00  3.600e-01   5.798 6.71e-09 ***
## quality               3.506e-01  2.632e-01   1.332   0.1829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4901.15  on 4351  degrees of freedom
## Residual deviance:  217.15  on 4339  degrees of freedom
## AIC: 243.15
##
## Number of Fisher Scoring iterations: 11
```

The model created is much better than the one created from the intercept alone as the residual deviance is much lower than the null deviance. Although, not all attributes are significant, I will be using this model to predict the type of wine.

Accuracy of prediction

```
probs1 <- predict(glm1, newdata=test, type='response') # Computes the probilities based on test data
pred2 <- ifelse(probs1 > 0.5, "red", "white" ) # Classify probabilities
table(Predicted = pred2, Actual = test$color) # Prediction table
```

```
##          Actual
## Predicted white  red
##     red       4  499
##     white  1633    9
```

```
paste("Accuracy of prediction: ", mean(pred2==test$color) * 100, "%") # Displays accuracy of prediction
```

```
## [1] "Accuracy of prediction:  99.3939393939394 %"
```

Knn Classification

Choosing k: 10-fold cross-validation

```
set.seed(4752)  # Set seed for reproducibility
i <- sample(nrow(wine2), nrow(wine2)*.67, replace=FALSE) # Sample from data frame
train <- wine2[i, ]
test <- wine2[-i, ]
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3) # Controls the computational nu
knn_fit <- train(color~., data = train, method = "knn",
                 trControl=trctrl,
                 preProcess = c("center", "scale"),
                 tuneLength = 10) # Train the classifier
knn_fit # Checking results
```

```
## k-Nearest Neighbors
##
## 4352 samples
##   12 predictor
##    2 classes: 'white', 'red'
##
## Pre-processing: centered (12), scaled (12)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 3917, 3918, 3916, 3917, 3916, 3917, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.9937970  0.9832486
##    7  0.9938738  0.9834430
##    9  0.9937202  0.9830307
##   11  0.9939500  0.9836592
##   13  0.9933374  0.9820171
##   15  0.9931848  0.9815979
##   17  0.9934910  0.9824305
##   19  0.9935681  0.9826407
##   21  0.9933384  0.9820218
##   23  0.9931852  0.9816229
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 11.
```

Knn classification with k = 11

```r
ind <- sample(2, nrow(scaled_df), replace = TRUE, prob=c(0.67, 0.33)) # Get random sample
scaled_train2 <- scaled_df[ind == 1, 1:12] # Initiliaze train set
scaled_test2 <- scaled_df[ind == 2, 1:12] # Initiliaze test set
trainLabels <- wine2[ind == 1, 13] # Set training label
testLabels <- wine2[ind == 2, 13] # Set test label
pred2 <- knn(train = scaled_train2, test = scaled_test2, cl = trainLabels, k = 11) # Predict results
```

Knn is not model-based therefore no summary is displayed

Accuracy of prediction

```r
table(Predicted = pred2, Actual = testLabels) # Prediction table
```

```
##          Actual
## Predicted white  red
##     white  1685    8
##     red       8  472
```

```r
paste("Accuracy of prediction: ", mean(pred2 == testLabels) * 100, "%") # Displays accuracy of predicti
```

```
## [1] "Accuracy of prediction:  99.263690750115 %"
```

Final comments

Both the logistical regression and knn algorithms are able to predict the type of wine extremely accurately: 99.40% for logistical regression vs 99.27% for knn with k = 11. In conclusion, logistical regression performs slightly better in this scenario.