W203 Final Project: Vinho Verde Wine Quality Analysis

Luc Robitaille, Andrew Bailey, Vish Sivakumar, Arun Sundrenath

2022-04-12

```
df_red <- read.csv(file = 'data/winequality-red.csv', sep=';')
df_white <- read.csv(file = 'data/winequality-white.csv', sep=';')
```

Background: The North Western portion of Portugal has a designated region from which vinho verde originates. This region is a proud producer of wine and has taken care to aggregate a large portion of data to understand what attributes are associated with the best wines. To do this a team of researchers observed over 6,000 samples of wines taking objective sensory data about the samples chemical attributes as well as having a sommelier grade the wine on a 10-point categorical scale. The test data recorded were operationalized with sensory data as well as subjective data regarding wine quality, of which the ratings were normalized by taking the median of at least 3 evaluations made by wine experts; scaling the quality between 0 (very bad) and 10 (very excellent). As a research team we hope to explore these chemical qualities and how they can be modeling indicators for high-quality wines, to help improve the product of the Vinho Verde vineyards. Through contemporary research surrounding the quality of wine, our team pre-posits the theory that quality of wine is the outcome of a system which relies on the balancing effects of different chemical properties. Therefore, we hypothesize that any one single variable will not be a strong indicator of a quality wine; rather that a system of indicators interacting together will support a better predictor of quality. Thus, our null hypothesis is that there is no one equation of physicochemical properties that will predict the quality of a bottle of wine.

Research Question: What are the most important attributes that create a great wine? Do red and white wines have different markers of great quality? Can we build a model which takes the chemical profile of wine and explains the overall quality of the product?

Explanatory Data Analysis: To perform our Exploratory Data Analysis (EDA) the team decided to run our analysis on the red and white wine datasets separately from each other in order to better observe the qualities of both types of wines; according to our research belief that red and white wine will have different marking indicators of quality. Furthermore, to preserve the quality of the dataset for linear model building and theory testing we separated the data to 30% EDA and 70% Linear Regression. Figures 1-12 show that there is a marked difference in the quantity of each of the wine making variables recorded in the data collection phase of this study. However, we further question, is there a parallel relationship between the quality indicators for red or white wine, or are the indicator variables different?

```
df_red$wine_type <- "red"
df_white$wine_type <-"white"
print("Table 1: Summary Red Wine")
```

```
## [1] "Table 1: Summary Red Wine"
```

```
summary(df_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
##   wine_type
##  Length:1599
##  Class :character
##  Mode  :character
##
##
##
```

```
print("Table 2: Summary White Wine")
```

```
## [1] "Table 2: Summary White Wine"
```

```
summary(df_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
##   wine_type
##  Length:4898
```

```
##  Class :character
##  Mode  :character
##
##
##
```

We were able to confirm that varieties of vinho verde wine should be treated differently and distinctly by running paired t-test across several physicochemical properties of both types.

```
df_wine_red_sample <- df_red[sample(nrow(df_red), 750), ]
df_wine_white_sample <- df_white[sample(nrow(df_white), 750), ]

t.test(df_wine_red_sample$fixed.acidity, df_wine_white_sample$fixed.acidity, paired=T)
```

```
##
##  Paired t-test
##
## data:  df_wine_red_sample$fixed.acidity and df_wine_white_sample$fixed.acidity
## t = 20.339, df = 749, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.246438 1.512762
## sample estimates:
## mean of the differences
##                  1.3796
```

```
t.test(df_wine_red_sample$total.sulfur.dioxide, df_wine_white_sample$total.sulfur.dioxide, paired=T)
```

```
##
##  Paired t-test
##
## data:  df_wine_red_sample$total.sulfur.dioxide and df_wine_white_sample$total.sulfur.dioxide
## t = -49.546, df = 749, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -97.49372 -90.06228
## sample estimates:
## mean of the differences
##                 -93.778
```

```
t.test(df_wine_red_sample$density, df_wine_white_sample$density, paired=T)
```

```
##
##  Paired t-test
##
## data:  df_wine_red_sample$density and df_wine_white_sample$density
## t = 19.697, df = 749, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.002419088 0.002954659
## sample estimates:
## mean of the differences
##             0.002686873
```

```
t.test(df_wine_red_sample$sulphates, df_wine_white_sample$sulphates, paired=T)
```

```
##
```

```
##  Paired t-test
##
## data:  df_wine_red_sample$sulphates and df_wine_white_sample$sulphates
## t = 24.903, df = 749, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1509735 0.1768132
## sample estimates:
## mean of the differences
##               0.1638933
```

```
summary(all_wine)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
##  Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
##  Max.   :0.61100   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.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
##   wine_type          Wine_Type
##  Length:6497        Length:6497
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$fixed.acidity, red_eda_df$fixed.acidity))

fig1 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 1: Red And White \n Distribution of Fixed Acidity \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$volatile.acidity, red_eda_df$volatile.acidity))
```

```
fig2 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 2: Red And White \n Distribution of Volatile Acidity \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$citric.acid, red_eda_df$citric.acid))

fig3 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 3: Red And White \n Distribution of Citric Acid \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$residual.sugar, red_eda_df$residual.sugar))

fig4 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 4: Red And White \n Distribution of Residual Sugar \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$chlorides, red_eda_df$chlorides))

fig5 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 5: Red And White \n Distribution of Chlorides \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c(white_eda_df$free.sulfur.dioxide, red_eda_df$free.sulfur.dioxide))

fig6 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 6: Red And White \n Distribution of Free Sulfur Dioxide \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$total.sulfur.dioxide, red_eda_df$total.sulfur.dioxide))

fig7 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 7: Red And White \n Distribution of Total Sulfur Dioxide \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$pH, red_eda_df$pH))
```

```r
fig8 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 8: Red And White \n Distribution of pH \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$density, red_eda_df$density))

fig9 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 9: Red And White \n Distribution of Density \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$sulphates, red_eda_df$sulphates))

fig10 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 10: Red And White \n Distribution of Sulphates \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$alcohol, red_eda_df$alcohol))

fig11 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig11: Red And White \n Distribution of Alcohol \n 30% Random Sample EDA")

d <- data.frame(
  type = c( white_eda_df$Wine_Type, red_eda_df$Wine_Type),
  value = c( white_eda_df$quality, red_eda_df$quality))

fig12 <- ggplot(data = d, aes(x=value, fill = type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(values=c("#404080", "#69b3a2")) +
  labs(title = "Fig 12: Red And White \n Distribution of Quality \n 30% Random Sample EDA")


grid.arrange(fig1, fig2,fig3, fig4, ncol = 2, nrow = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Fig 1: Red And White
Distribution of Fixed Acidity
30% Random Sample EDA

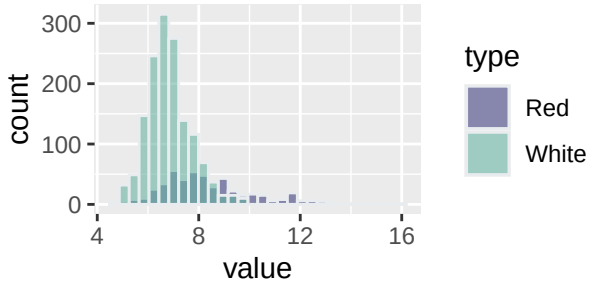


Fig 2: Red And White
Distribution of Volatile Acidity
30% Random Sample EDA

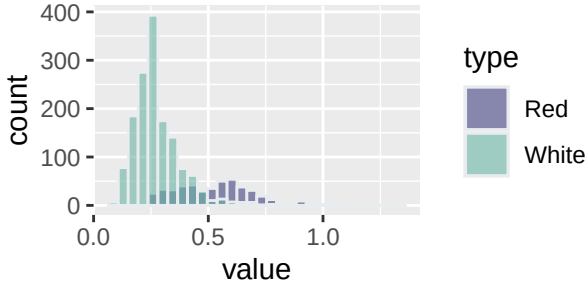


Fig 3: Red And White
Distribution of Citric Acid
30% Random Sample EDA

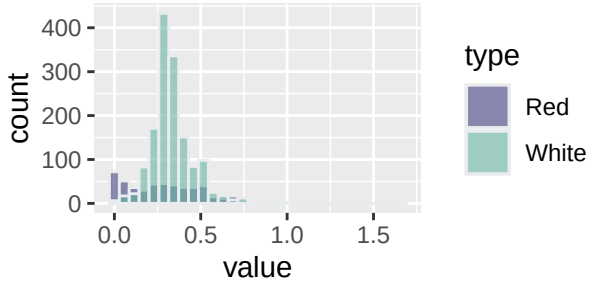
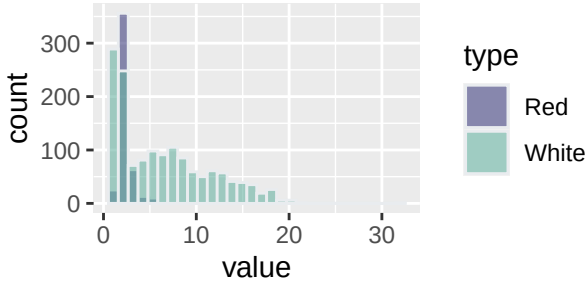


Fig 4: Red And White
Distribution of Residual Sugar
30% Random Sample EDA




```
grid.arrange(fig5, fig6, fig7, fig8, ncol = 2, nrow = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Fig 5: Red And White
Distribution of Chlorides
30% Random Sample EDA

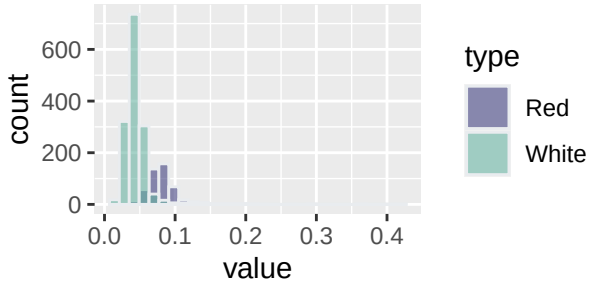


Fig 6: Red And White
Distribution of Free Sulfur Dioxide
30% Random Sample EDA

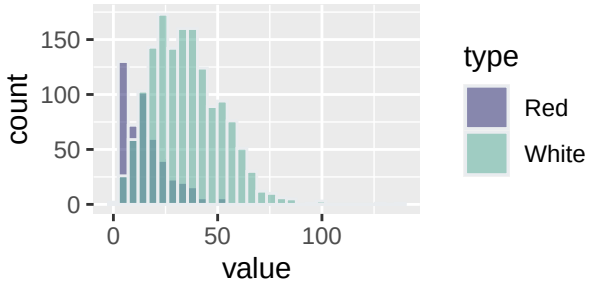


Fig 7: Red And White
Distribution of Total Sulfur Dioxide
30% Random Sample EDA

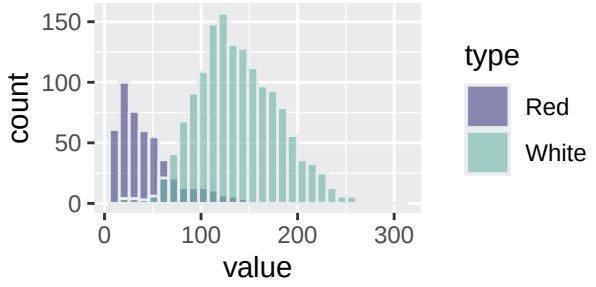
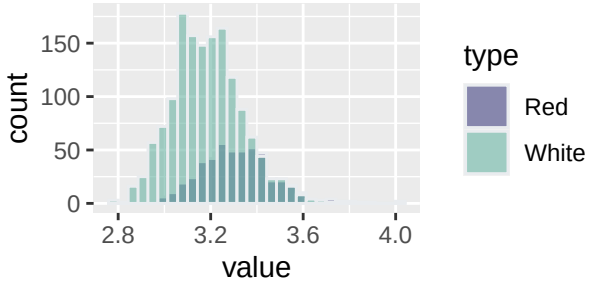
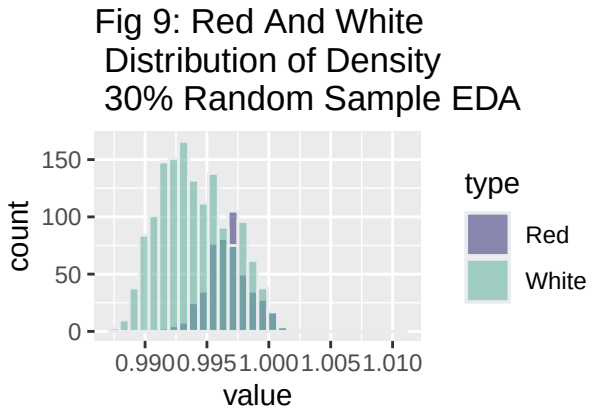


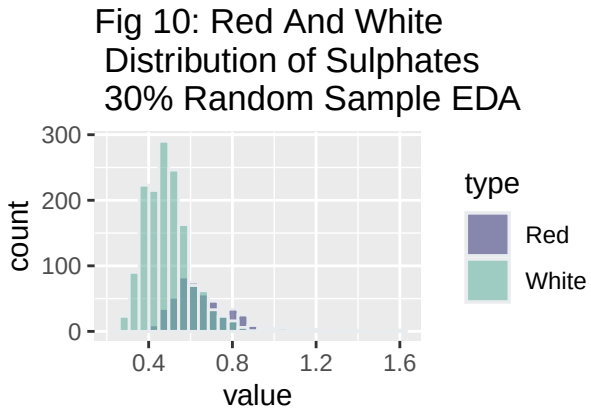
Fig 8: Red And White
Distribution of pH
30% Random Sample EDA

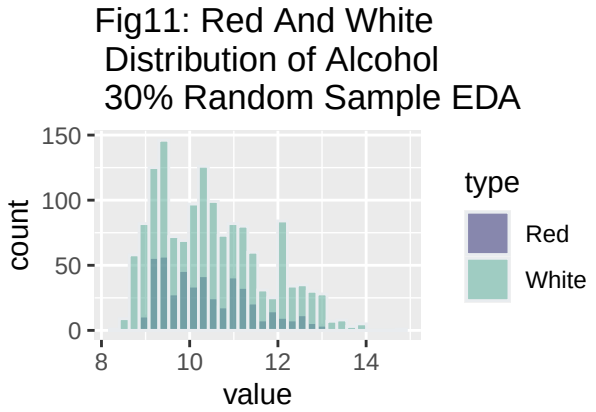


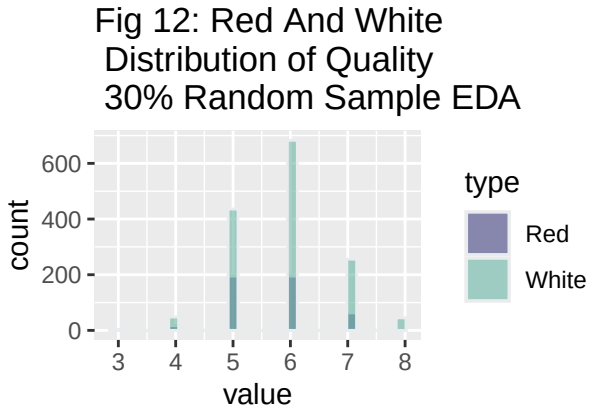

```
grid.arrange(fig9, fig10, fig11, fig12, ncol = 2, nrow = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```









> By looking at the distribution of individual physicochemical properties of the different varieties of vinho verde wine, we then confirmed with a t-test that there were significant differences between the varieties and they should be modeled differently.

Given that the datasets we are studying are very large we will pursue a linear model under the Large Sample framework. In accordance with this framework we will model around variables which approach a normal distribution, have low correlations with other variables, and have a linear relationship with our outcome variable, quality. From our initial EDA, our dataset appears to have i.i.d underpinnings, with variables generally approaching a normal distribution.

Independent, identically distributed (i.i.d) data is an assumption of our dataset that will support the Central Limit Theorem (CLM), which is a statistical theorem of large datasets which will produce powerful properties for our linear regression models. An important piece of proving i.i.d is examining experimental theory and operationalization. In the case of this data set our chemical variables are all independent from another wine's same variable measurement; e.g. the measurement of one wine's acidity will not influence the acidity of another wine. For our outcome variable, Quality, the score is operationalized by taking the median of three independent sommelier's ratings (ordinal, 1-10 scale). The effect of taking a median between three independent ratings helps to decrease the effect of bias in any of the ratings which supports the i.i.d assumptions. Additionally, the i.i.d assumption requires that each data-point comes from one underlying distribution. In the case of our dataset, the data is pulled from a large sample of wines all from the Vinho Verde region of portugal. The central location of these entries provides a common distribution for all of the properties of the various wines. However, this reduces the generalizability of our results to mostly just wines in the Vinho Verde region of Portugal.

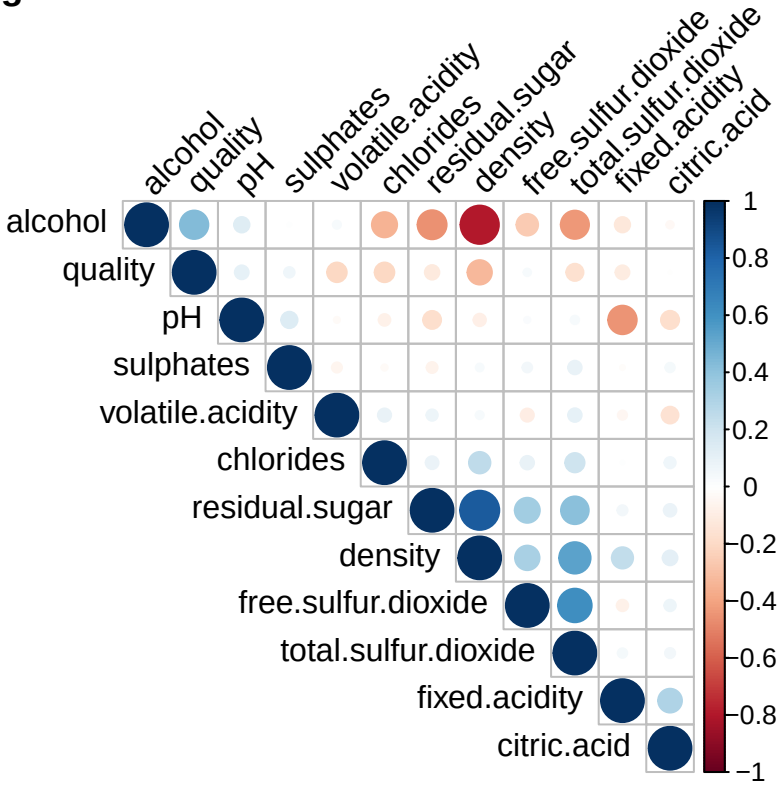
A final EDA check which is important to perform is the independence of individual variables relative to each other across the dataset. Figures 13 and 14 demonstrate that we have a reasonable amount of differentiation from each other to perform our linear regression analysis.

```
all_numeric_vals_white = subset(white_eda_df,select=-c(Wine_Type,qual_bin,wine_type))
cor_tab_white <- cor(all_numeric_vals_white)
round(cor_tab_white, 2)
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity                  1.00            -0.05        0.31           0.06
## volatile.acidity              -0.05             1.00       -0.15           0.07
## citric.acid                    0.31            -0.15        1.00           0.09
## residual.sugar                 0.06             0.07        0.09           1.00
## chlorides                      0.01             0.10        0.07           0.09
## free.sulfur.dioxide           -0.07            -0.09        0.07           0.34
## total.sulfur.dioxide           0.05             0.10        0.06           0.42
## density                        0.25             0.04        0.12           0.83
## pH                            -0.44            -0.02       -0.18          -0.18
## sulphates                     -0.01            -0.06        0.05          -0.07
## alcohol                       -0.13             0.04       -0.03          -0.46
## quality                       -0.11            -0.21        0.00          -0.11
##                       chlorides free.sulfur.dioxide total.sulfur.dioxide density
## fixed.acidity              0.01               -0.07                 0.05    0.25
## volatile.acidity           0.10               -0.09                 0.10    0.04
## citric.acid                0.07                0.07                 0.06    0.12
## residual.sugar             0.09                0.34                 0.42    0.83
## chlorides                  1.00                0.10                 0.20    0.25
## free.sulfur.dioxide        0.10                1.00                 0.61    0.33
## total.sulfur.dioxide       0.20                0.61                 1.00    0.54
## density                    0.25                0.33                 0.54    1.00
## pH                        -0.07                0.02                 0.04   -0.08
## sulphates                 -0.02                0.05                 0.10    0.03
## alcohol                   -0.35               -0.25                -0.43   -0.80
## quality                   -0.21                0.03                -0.16   -0.32
##                           pH sulphates alcohol quality
## fixed.acidity          -0.44     -0.01   -0.13   -0.11
## volatile.acidity       -0.02     -0.06    0.04   -0.21
## citric.acid            -0.18      0.05   -0.03    0.00
## residual.sugar         -0.18     -0.07   -0.46   -0.11
## chlorides              -0.07     -0.02   -0.35   -0.21
## free.sulfur.dioxide     0.02      0.05   -0.25    0.03
## total.sulfur.dioxide    0.04      0.10   -0.43   -0.16
## density                -0.08      0.03   -0.80   -0.32
## pH                      1.00      0.15    0.13    0.10
## sulphates               0.15      1.00    0.01    0.06
## alcohol                 0.13      0.01    1.00    0.44
## quality                 0.10      0.06    0.44    1.00
```

```
fig13 <- corrplot(cor_tab_white, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, main = "Fig 13: Correlation between Variables White Wine", mar=
```

Fig 13: Correlation between Variables White Wine



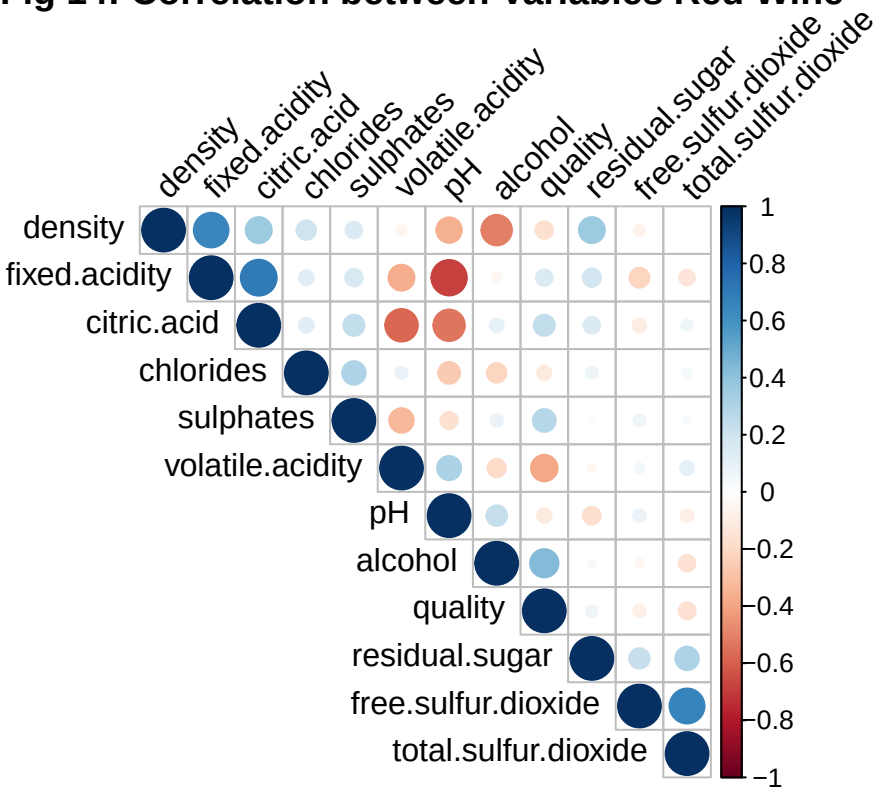

```
all_numeric_vals_red = subset(red_eda_df,select=-c(Wine_Type,qual_bin,wine_type))
cor_tab_red <- cor(all_numeric_vals_red)
round(cor_tab_red, 2)
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity                  1.00            -0.36        0.70           0.18
## volatile.acidity              -0.36             1.00       -0.58          -0.04
## citric.acid                    0.70            -0.58        1.00           0.15
## residual.sugar                 0.18            -0.04        0.15           1.00
## chlorides                      0.12             0.08        0.12           0.08
## free.sulfur.dioxide           -0.22             0.05       -0.10           0.23
## total.sulfur.dioxide          -0.13             0.11        0.07           0.30
## density                        0.65            -0.05        0.37           0.37
## pH                            -0.68             0.32       -0.54          -0.17
## sulphates                      0.17            -0.33        0.24           0.02
## alcohol                       -0.05            -0.18        0.10           0.03
## quality                        0.15            -0.38        0.24           0.07
##                       chlorides free.sulfur.dioxide total.sulfur.dioxide density
## fixed.acidity              0.12              -0.22                -0.13    0.65
## volatile.acidity           0.08               0.05                 0.11   -0.05
## citric.acid                0.12              -0.10                 0.07    0.37
## residual.sugar             0.08               0.23                 0.30    0.37
## chlorides                  1.00              -0.01                 0.05    0.21
## free.sulfur.dioxide       -0.01               1.00                 0.67   -0.07
## total.sulfur.dioxide       0.05               0.67                 1.00    0.01
## density                    0.21              -0.07                 0.01    1.00
```

```
## pH                     -0.26              0.09             -0.09   -0.35
## sulphates               0.30              0.08              0.03    0.15
## alcohol                -0.21             -0.04             -0.15   -0.51
## quality                -0.11             -0.09             -0.16   -0.17
##                         pH sulphates alcohol quality
## fixed.acidity        -0.68      0.17   -0.05    0.15
## volatile.acidity      0.32     -0.33   -0.18   -0.38
## citric.acid          -0.54      0.24    0.10    0.24
## residual.sugar       -0.17      0.02    0.03    0.07
## chlorides            -0.26      0.30   -0.21   -0.11
## free.sulfur.dioxide   0.09      0.08   -0.04   -0.09
## total.sulfur.dioxide -0.09      0.03   -0.15   -0.16
## density              -0.35      0.15   -0.51   -0.17
## pH                    1.00     -0.16    0.23   -0.11
## sulphates            -0.16      1.00    0.09    0.28
## alcohol               0.23      0.09    1.00    0.44
## quality              -0.11      0.28    0.44    1.00
```

```
fig14 <- corrplot(cor_tab_red, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, main = "Fig 14: Correlation between Variables Red Wine", mar=c(
```

Fig 14: Correlation between Variables Red Wine




```
fig13
```

```
## $corr
##                         alcohol      quality          pH    sulphates
## alcohol             1.000000000  0.438505618  0.13008872  0.007154234
## quality             0.438505618  1.000000000  0.10431981  0.060524660
```

```
## pH                    0.130088721  0.104319813  1.00000000  0.145468256
## sulphates             0.007154234  0.060524660  0.14546826  1.000000000
## volatile.acidity      0.036288629 -0.208260656 -0.02047712 -0.055936232
## chlorides            -0.345308865 -0.209852896 -0.07462093 -0.021608975
## residual.sugar       -0.458040434 -0.114473396 -0.17579408 -0.065420028
## density              -0.798902162 -0.323507225 -0.08377036  0.033774062
## free.sulfur.dioxide  -0.254056070  0.030924020  0.02098188  0.052050311
## total.sulfur.dioxide -0.433444760 -0.160074005  0.03899260  0.099766585
## fixed.acidity        -0.127709012 -0.105669804 -0.44247074 -0.013545909
## citric.acid          -0.030173723 -0.004117709 -0.17930874  0.048335475
##                      volatile.acidity   chlorides residual.sugar     density
## alcohol                    0.03628863 -0.34530887    -0.45804043 -0.79890216
## quality                   -0.20826066 -0.20985290    -0.11447340 -0.32350722
## pH                        -0.02047712 -0.07462093    -0.17579408 -0.08377036
## sulphates                 -0.05593623 -0.02160898    -0.06542003  0.03377406
## volatile.acidity           1.00000000  0.09560415     0.07017739  0.03570130
## chlorides                  0.09560415  1.00000000     0.08964335  0.25037326
## residual.sugar             0.07017739  0.08964335     1.00000000  0.83436669
## density                    0.03570130  0.25037326     0.83436669  1.00000000
## free.sulfur.dioxide       -0.09363044  0.09922001     0.34139876  0.32731141
## total.sulfur.dioxide       0.10056558  0.20242305     0.41817142  0.53766452
## fixed.acidity             -0.04636288  0.00708508     0.05569686  0.24686402
## citric.acid               -0.15363065  0.06896667     0.08630008  0.11963022
##                      free.sulfur.dioxide total.sulfur.dioxide fixed.acidity
## alcohol                      -0.25405607          -0.43344476   -0.12770901
## quality                       0.03092402          -0.16007401   -0.10566980
## pH                            0.02098188           0.03899260   -0.44247074
## sulphates                     0.05205031           0.09976659   -0.01354591
## volatile.acidity             -0.09363044           0.10056558   -0.04636288
## chlorides                     0.09922001           0.20242305    0.00708508
## residual.sugar                0.34139876           0.41817142    0.05569686
## density                       0.32731141           0.53766452    0.24686402
## free.sulfur.dioxide           1.00000000           0.61399090   -0.07187455
## total.sulfur.dioxide          0.61399090           1.00000000    0.04925928
## fixed.acidity                -0.07187455           0.04925928    1.00000000
## citric.acid                   0.07217376           0.05601818    0.30895129
##                       citric.acid
## alcohol              -0.030173723
## quality              -0.004117709
## pH                   -0.179308740
## sulphates             0.048335475
## volatile.acidity     -0.153630649
## chlorides             0.068966667
## residual.sugar        0.086300080
## density               0.119630215
## free.sulfur.dioxide   0.072173757
## total.sulfur.dioxide  0.056018181
## fixed.acidity         0.308951287
## citric.acid           1.000000000
##
## $corrPos
##                   xName               yName  x  y         corr
## 1               alcohol             alcohol  1 12  1.000000000
## 2               quality             alcohol  2 12  0.438505618
```

```
## 3               quality              quality  2 11  1.000000000
## 4                    pH              alcohol  3 12  0.130088721
## 5                    pH              quality  3 11  0.104319813
## 6                    pH                   pH  3 10  1.000000000
## 7             sulphates              alcohol  4 12  0.007154234
## 8             sulphates              quality  4 11  0.060524660
## 9             sulphates                   pH  4 10  0.145468256
## 10            sulphates            sulphates  4  9  1.000000000
## 11    volatile.acidity              alcohol  5 12  0.036288629
## 12    volatile.acidity              quality  5 11 -0.208260656
## 13    volatile.acidity                   pH  5 10 -0.020477123
## 14    volatile.acidity            sulphates  5  9 -0.055936232
## 15    volatile.acidity     volatile.acidity  5  8  1.000000000
## 16            chlorides              alcohol  6 12 -0.345308865
## 17            chlorides              quality  6 11 -0.209852896
## 18            chlorides                   pH  6 10 -0.074620929
## 19            chlorides            sulphates  6  9 -0.021608975
## 20            chlorides     volatile.acidity  6  8  0.095604150
## 21            chlorides            chlorides  6  7  1.000000000
## 22      residual.sugar              alcohol  7 12 -0.458040434
## 23      residual.sugar              quality  7 11 -0.114473396
## 24      residual.sugar                   pH  7 10 -0.175794084
## 25      residual.sugar            sulphates  7  9 -0.065420028
## 26      residual.sugar     volatile.acidity  7  8  0.070177390
## 27      residual.sugar            chlorides  7  7  0.089643351
## 28      residual.sugar       residual.sugar  7  6  1.000000000
## 29             density              alcohol  8 12 -0.798902162
## 30             density              quality  8 11 -0.323507225
## 31             density                   pH  8 10 -0.083770365
## 32             density            sulphates  8  9  0.033774062
## 33             density     volatile.acidity  8  8  0.035701295
## 34             density            chlorides  8  7  0.250373262
## 35             density       residual.sugar  8  6  0.834366693
## 36             density              density  8  5  1.000000000
## 37  free.sulfur.dioxide              alcohol  9 12 -0.254056070
## 38  free.sulfur.dioxide              quality  9 11  0.030924020
## 39  free.sulfur.dioxide                   pH  9 10  0.020981878
## 40  free.sulfur.dioxide            sulphates  9  9  0.052050311
## 41  free.sulfur.dioxide     volatile.acidity  9  8 -0.093630438
## 42  free.sulfur.dioxide            chlorides  9  7  0.099220011
## 43  free.sulfur.dioxide       residual.sugar  9  6  0.341398758
## 44  free.sulfur.dioxide              density  9  5  0.327311409
## 45  free.sulfur.dioxide  free.sulfur.dioxide  9  4  1.000000000
## 46 total.sulfur.dioxide              alcohol 10 12 -0.433444760
## 47 total.sulfur.dioxide              quality 10 11 -0.160074005
## 48 total.sulfur.dioxide                   pH 10 10  0.038992596
## 49 total.sulfur.dioxide            sulphates 10  9  0.099766585
## 50 total.sulfur.dioxide     volatile.acidity 10  8  0.100565580
## 51 total.sulfur.dioxide            chlorides 10  7  0.202423050
## 52 total.sulfur.dioxide       residual.sugar 10  6  0.418171418
## 53 total.sulfur.dioxide              density 10  5  0.537664521
## 54 total.sulfur.dioxide  free.sulfur.dioxide 10  4  0.613990902
## 55 total.sulfur.dioxide total.sulfur.dioxide 10  3  1.000000000
## 56        fixed.acidity              alcohol 11 12 -0.127709012
```

```
## 57        fixed.acidity              quality 11 11 -0.105669804
## 58        fixed.acidity                   pH 11 10 -0.442470741
## 59        fixed.acidity            sulphates 11  9 -0.013545909
## 60        fixed.acidity     volatile.acidity 11  8 -0.046362881
## 61        fixed.acidity            chlorides 11  7  0.007085080
## 62        fixed.acidity       residual.sugar 11  6  0.055696858
## 63        fixed.acidity              density 11  5  0.246864022
## 64        fixed.acidity  free.sulfur.dioxide 11  4 -0.071874554
## 65        fixed.acidity total.sulfur.dioxide 11  3  0.049259281
## 66        fixed.acidity        fixed.acidity 11  2  1.000000000
## 67          citric.acid              alcohol 12 12 -0.030173723
## 68          citric.acid              quality 12 11 -0.004117709
## 69          citric.acid                   pH 12 10 -0.179308740
## 70          citric.acid            sulphates 12  9  0.048335475
## 71          citric.acid     volatile.acidity 12  8 -0.153630649
## 72          citric.acid            chlorides 12  7  0.068966667
## 73          citric.acid       residual.sugar 12  6  0.086300080
## 74          citric.acid              density 12  5  0.119630215
## 75          citric.acid  free.sulfur.dioxide 12  4  0.072173757
## 76          citric.acid total.sulfur.dioxide 12  3  0.056018181
## 77          citric.acid        fixed.acidity 12  2  0.308951287
## 78          citric.acid          citric.acid 12  1  1.000000000
##
## $arg
## $arg$type
## [1] "upper"
```

```
fig14
```

```
## $corr
##                         density fixed.acidity citric.acid    chlorides
## density               1.000000000    0.65493027  0.36938698  0.207004224
## fixed.acidity         0.654930273    1.00000000  0.70158484  0.120373882
## citric.acid           0.369386983    0.70158484  1.00000000  0.123368653
## chlorides             0.207004224    0.12037388  0.12336865  1.000000000
## sulphates             0.150356424    0.16747691  0.24306603  0.300913632
## volatile.acidity     -0.051980381   -0.36446782 -0.57805794  0.081487156
## pH                   -0.352065394   -0.68294578 -0.53878639 -0.258119053
## alcohol              -0.506940882   -0.04943936  0.10255921 -0.210060404
## quality              -0.169763803    0.15156557  0.24174962 -0.110122573
## residual.sugar        0.367640537    0.18286838  0.15254390  0.079435875
## free.sulfur.dioxide  -0.069508059   -0.21657181 -0.10151979 -0.009587733
## total.sulfur.dioxide  0.005718127   -0.13488848  0.06876455  0.047403046
##                        sulphates volatile.acidity          pH     alcohol
## density               0.15035642      -0.05198038 -0.35206539 -0.50694088
## fixed.acidity         0.16747691      -0.36446782 -0.68294578 -0.04943936
## citric.acid           0.24306603      -0.57805794 -0.53878639  0.10255921
## chlorides             0.30091363       0.08148716 -0.25811905 -0.21006040
## sulphates             1.00000000      -0.32640610 -0.16214069  0.08746852
## volatile.acidity     -0.32640610       1.00000000  0.31699836 -0.18490331
## pH                   -0.16214069       0.31699836  1.00000000  0.23130021
## alcohol               0.08746852      -0.18490331  0.23130021  1.00000000
## quality               0.28300462      -0.38233594 -0.11456463  0.43680537
## residual.sugar        0.01959596      -0.03860524 -0.17224624  0.03355432
## free.sulfur.dioxide   0.07843620       0.05005701  0.08780527 -0.04260088
```

```
## total.sulfur.dioxide  0.03141743       0.10602232 -0.08841754 -0.15109212
##                          quality residual.sugar free.sulfur.dioxide
## density              -0.16976380     0.36764054        -0.069508059
## fixed.acidity         0.15156557     0.18286838        -0.216571807
## citric.acid           0.24174962     0.15254390        -0.101519794
## chlorides            -0.11012257     0.07943587        -0.009587733
## sulphates             0.28300462     0.01959596         0.078436198
## volatile.acidity     -0.38233594    -0.03860524         0.050057015
## pH                   -0.11456463    -0.17224624         0.087805269
## alcohol               0.43680537     0.03355432        -0.042600883
## quality               1.00000000     0.07024869        -0.086501024
## residual.sugar        0.07024869     1.00000000         0.228295427
## free.sulfur.dioxide  -0.08650102     0.22829543         1.000000000
## total.sulfur.dioxide -0.15963601     0.30102872         0.665321230
##                      total.sulfur.dioxide
## density                       0.005718127
## fixed.acidity                -0.134888482
## citric.acid                   0.068764554
## chlorides                     0.047403046
## sulphates                     0.031417434
## volatile.acidity              0.106022324
## pH                           -0.088417541
## alcohol                      -0.151092120
## quality                      -0.159636006
## residual.sugar                0.301028715
## free.sulfur.dioxide           0.665321230
## total.sulfur.dioxide          1.000000000
##
## $corrPos
##                    xName                yName  x  y         corr
## 1                density              density  1 12  1.000000000
## 2          fixed.acidity              density  2 12  0.654930273
## 3          fixed.acidity        fixed.acidity  2 11  1.000000000
## 4            citric.acid              density  3 12  0.369386983
## 5            citric.acid        fixed.acidity  3 11  0.701584835
## 6            citric.acid          citric.acid  3 10  1.000000000
## 7              chlorides              density  4 12  0.207004224
## 8              chlorides        fixed.acidity  4 11  0.120373882
## 9              chlorides          citric.acid  4 10  0.123368653
## 10             chlorides            chlorides  4  9  1.000000000
## 11             sulphates              density  5 12  0.150356424
## 12             sulphates        fixed.acidity  5 11  0.167476914
## 13             sulphates          citric.acid  5 10  0.243066034
## 14             sulphates            chlorides  5  9  0.300913632
## 15             sulphates            sulphates  5  8  1.000000000
## 16     volatile.acidity              density  6 12 -0.051980381
## 17     volatile.acidity        fixed.acidity  6 11 -0.364467825
## 18     volatile.acidity          citric.acid  6 10 -0.578057944
## 19     volatile.acidity            chlorides  6  9  0.081487156
## 20     volatile.acidity            sulphates  6  8 -0.326406098
## 21     volatile.acidity     volatile.acidity  6  7  1.000000000
## 22                   pH              density  7 12 -0.352065394
## 23                   pH        fixed.acidity  7 11 -0.682945783
## 24                   pH          citric.acid  7 10 -0.538786390
```

```
## 25                  pH            chlorides  7  9 -0.258119053
## 26                  pH            sulphates  7  8 -0.162140692
## 27                  pH     volatile.acidity  7  7  0.316998359
## 28                  pH                   pH  7  6  1.000000000
## 29             alcohol              density  8 12 -0.506940882
## 30             alcohol        fixed.acidity  8 11 -0.049439357
## 31             alcohol          citric.acid  8 10  0.102559205
## 32             alcohol            chlorides  8  9 -0.210060404
## 33             alcohol            sulphates  8  8  0.087468519
## 34             alcohol     volatile.acidity  8  7 -0.184903307
## 35             alcohol                   pH  8  6  0.231300208
## 36             alcohol              alcohol  8  5  1.000000000
## 37             quality              density  9 12 -0.169763803
## 38             quality        fixed.acidity  9 11  0.151565570
## 39             quality          citric.acid  9 10  0.241749621
## 40             quality            chlorides  9  9 -0.110122573
## 41             quality            sulphates  9  8  0.283004625
## 42             quality     volatile.acidity  9  7 -0.382335944
## 43             quality                   pH  9  6 -0.114564630
## 44             quality              alcohol  9  5  0.436805369
## 45             quality              quality  9  4  1.000000000
## 46       residual.sugar              density 10 12  0.367640537
## 47       residual.sugar        fixed.acidity 10 11  0.182868376
## 48       residual.sugar          citric.acid 10 10  0.152543896
## 49       residual.sugar            chlorides 10  9  0.079435875
## 50       residual.sugar            sulphates 10  8  0.019595958
## 51       residual.sugar     volatile.acidity 10  7 -0.038605243
## 52       residual.sugar                   pH 10  6 -0.172246241
## 53       residual.sugar              alcohol 10  5  0.033554317
## 54       residual.sugar              quality 10  4  0.070248693
## 55       residual.sugar       residual.sugar 10  3  1.000000000
## 56  free.sulfur.dioxide              density 11 12 -0.069508059
## 57  free.sulfur.dioxide        fixed.acidity 11 11 -0.216571807
## 58  free.sulfur.dioxide          citric.acid 11 10 -0.101519794
## 59  free.sulfur.dioxide            chlorides 11  9 -0.009587733
## 60  free.sulfur.dioxide            sulphates 11  8  0.078436198
## 61  free.sulfur.dioxide     volatile.acidity 11  7  0.050057015
## 62  free.sulfur.dioxide                   pH 11  6  0.087805269
## 63  free.sulfur.dioxide              alcohol 11  5 -0.042600883
## 64  free.sulfur.dioxide              quality 11  4 -0.086501024
## 65  free.sulfur.dioxide       residual.sugar 11  3  0.228295427
## 66  free.sulfur.dioxide  free.sulfur.dioxide 11  2  1.000000000
## 67 total.sulfur.dioxide              density 12 12  0.005718127
## 68 total.sulfur.dioxide        fixed.acidity 12 11 -0.134888482
## 69 total.sulfur.dioxide          citric.acid 12 10  0.068764554
## 70 total.sulfur.dioxide            chlorides 12  9  0.047403046
## 71 total.sulfur.dioxide            sulphates 12  8  0.031417434
## 72 total.sulfur.dioxide     volatile.acidity 12  7  0.106022324
## 73 total.sulfur.dioxide                   pH 12  6 -0.088417541
## 74 total.sulfur.dioxide              alcohol 12  5 -0.151092120
## 75 total.sulfur.dioxide              quality 12  4 -0.159636006
## 76 total.sulfur.dioxide       residual.sugar 12  3  0.301028715
## 77 total.sulfur.dioxide  free.sulfur.dioxide 12  2  0.665321230
## 78 total.sulfur.dioxide total.sulfur.dioxide 12  1  1.000000000
```

```
##
## $arg
## $arg$type
## [1] "upper"
```

Regression Model Assumptions for Large Sample Are the assumptions of the large-sample model met so that you can use an OLS regression to produce consistent estimates? Unique BLP Exists

From the analysis of the data for train data set on red wine and white wine shows unique estimates exits for the features for the individual parameters

```
model_red_all<- lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.su
model_white_all <- lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free
summary(model_red_all)
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + volatile.acidity, data = train_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70641 -0.36264 -0.04507  0.46526  1.99482
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.265255  25.629789  -0.049   0.9606
## fixed.acidity         0.012547   0.031271   0.401   0.6883
## volatile.acidity     -1.105935   0.149416  -7.402 2.65e-13 ***
## citric.acid          -0.138168   0.181885  -0.760   0.4476
## residual.sugar        0.005385   0.017543   0.307   0.7589
## chlorides            -1.580355   0.525633  -3.007   0.0027 **
## free.sulfur.dioxide   0.004907   0.002686   1.827   0.0680 .
## total.sulfur.dioxide -0.003638   0.000911  -3.993 6.95e-05 ***
## density               5.200041  26.139468   0.199   0.8424
## pH                   -0.355486   0.232861  -1.527   0.1271
## sulphates             0.802013   0.141404   5.672 1.80e-08 ***
## alcohol               0.296482   0.032141   9.224  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6656 on 1107 degrees of freedom
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3431
## F-statistic: 54.09 on 11 and 1107 DF,  p-value: < 2.2e-16
```

```
summary(model_white_all)
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + volatile.acidity, data = train_white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.7535 -0.4957 -0.0318  0.4569  3.1360
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.153e+02  2.621e+01   8.213 3.02e-16 ***
## fixed.acidity         1.039e-01  2.654e-02   3.916 9.17e-05 ***
## volatile.acidity     -1.797e+00  1.340e-01 -13.412  < 2e-16 ***
## citric.acid          -6.519e-02  1.146e-01  -0.569 0.569621
## residual.sugar        1.017e-01  9.961e-03  10.210  < 2e-16 ***
## chlorides            -3.934e-01  6.700e-01  -0.587 0.557186
## free.sulfur.dioxide   2.907e-03  9.837e-04   2.955 0.003150 **
## total.sulfur.dioxide  8.860e-05  4.546e-04   0.195 0.845492
## density              -2.158e+02  2.656e+01  -8.126 6.12e-16 ***
## pH                    8.257e-01  1.306e-01   6.322 2.92e-10 ***
## sulphates             7.170e-01  1.189e-01   6.031 1.80e-09 ***
## alcohol               1.127e-01  3.348e-02   3.365 0.000773 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7396 on 3416 degrees of freedom
## Multiple R-squared:  0.2909, Adjusted R-squared:  0.2886
## F-statistic: 127.4 on 11 and 3416 DF,  p-value: < 2.2e-16
```

```
library(stargazer)
stargazer(
  model_red_all,
  model_white_all,
  type='text',title = c("Table 3: Summary of Red Wine","Table 4: Summary of White Wine"),
  header=FALSE,
  star.cutoffs=c(0.05,0.01,0.001)
)
```

```
##
## Table 3: Summary of Red Wine
## ================================================================================
##                                        Dependent variable:
##                          ---------------------------------------------------
##                                                quality
##                                   (1)                          (2)
## --------------------------------------------------------------------------------
## fixed.acidity                      0.013                      0.104***
##                                   (0.031)                    (0.027)
##
## volatile.acidity                -1.106***                  -1.797***
##                                   (0.149)                    (0.134)
##
## citric.acid                       -0.138                      -0.065
##                                   (0.182)                    (0.115)
##
## residual.sugar                     0.005                      0.102***
##                                   (0.018)                    (0.010)
##
## chlorides                       -1.580**                     -0.393
##                                   (0.526)                    (0.670)
##
```

```
## free.sulfur.dioxide            0.005                    0.003**
##                               (0.003)                  (0.001)
##
## total.sulfur.dioxide         -0.004***                  0.0001
##                               (0.001)                  (0.0005)
##
## density                       5.200                   -215.803***
##                              (26.139)                  (26.556)
##
## pH                            -0.355                    0.826***
##                               (0.233)                  (0.131)
##
## sulphates                    0.802***                   0.717***
##                               (0.141)                  (0.119)
##
## alcohol                      0.296***                   0.113***
##                               (0.032)                  (0.033)
##
## Constant                      -1.265                   215.288***
##                              (25.630)                  (26.212)
##
## ------------------------------------------------------------------------
## Observations                   1,119                     3,428
## R2                             0.350                     0.291
## Adjusted R2                    0.343                     0.289
## Residual Std. Error      0.666 (df = 1107)         0.740 (df = 3416)
## F Statistic          54.094*** (df = 11; 1107) 127.382*** (df = 11; 3416)
## ========================================================================
## Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

Variation in X's Table below shows the variance analysis of the data set for both the Red and White wine datasets. From the variance statistics we can conclude that the Var in most of the feature dataset for the Red and White wine dataset is > 0.2 which should be the case for satiating the assumption for the variation in the feature dataset for the large sample assumption. For both Red and White wine variance dataset, we do see a trend with the Total Sulfur Dioxide and Free Sulfur Dioxide tends to be < 0.2 for the variance statistics, this could be due to non normality of the both the datasets and possible tight distribution of both the features. More investigation into these features is necessary to uncover more insights.

```
var_1 <- var.test(df_red$quality,df_red$fixed.acidity)
var_2 <- var.test(df_red$quality,df_red$volatile.acidity)
var_3 <- var.test(df_red$quality,df_red$citric.acid)
var_4 <- var.test(df_red$quality,df_red$residual.sugar)
var_5 <- var.test(df_red$quality,df_red$chlorides)
var_6 <- var.test(df_red$quality,df_red$free.sulfur.dioxide)
var_7 <- var.test(df_red$quality,df_red$total.sulfur.dioxide)
var_8 <- var.test(df_red$quality,df_red$density)
var_9 <- var.test(df_red$quality,df_red$pH)
var_10 <- var.test(df_red$quality,df_red$sulphates)
var_11 <- var.test(df_red$quality,df_red$alcohol)


df_var_red <- data.frame( name=  c('Fixed Acidity', 'Volatile Acidity','Citric Acid', 'Residual Sugars'
                          var_stat =c(var_1$statistic,var_2$statistic,var_3$statistic,var_4$statistic,va
```

```
var_12 <- var.test(df_white$quality,df_white$fixed.acidity)
var_13 <- var.test(df_white$quality,df_white$volatile.acidity)
var_14 <- var.test(df_white$quality,df_white$citric.acid)
var_15 <- var.test(df_white$quality,df_white$residual.sugar)
var_16 <- var.test(df_white$quality,df_white$chlorides)
var_17 <- var.test(df_white$quality,df_white$free.sulfur.dioxide)
var_18 <- var.test(df_white$quality,df_white$total.sulfur.dioxide)
var_19 <- var.test(df_white$quality,df_white$density)
var_20 <- var.test(df_white$quality,df_white$pH)
var_21 <- var.test(df_white$quality,df_white$sulphates)
var_22 <- var.test(df_white$quality,df_white$alcohol)

df_var_white <- data.frame(name=  c('Fixed Acidity', 'Volatile Acidity','Citric Acid', 'Residual Sugars
                           var_stat =c(var_12$statistic,var_13$statistic,var_14$statistic,var_15$statis

print("Table 5: Variance table for Red Wine")
```

```
## [1] "Table 5: Variance table for Red Wine"
```

```
gt(df_var_red,rowname_col = df_var_red$name)
```

name	var_stat
Fixed Acidity	2.151365e-01
Volatile Acidity	2.034061e+01
Citric Acid	1.718608e+01
Residual Sugars	3.280695e-01
chlorides	2.944137e+02
Free Sulfur Dioxide	5.960509e-03
Total Sulfur Dioxides	6.026864e-04
Density	1.830890e+05
pH	2.736159e+01
Sulphates	2.269784e+01
Alcohol	5.742702e-01

```
print("Table 6: Variance table for white Wine")
```

```
## [1] "Table 6: Variance table for white Wine"
```

```
gt(df_var_white,rowname_col = df_var_white$name)
```

name	var_stat
Fixed Acidity	1.101447e+00
Volatile Acidity	7.720385e+01
Citric Acid	5.355502e+01
Residual Sugars	3.048910e-02
chlorides	1.643202e+03
Free Sulfur Dioxide	2.711756e-03
Total Sulfur Dioxides	4.342849e-04
Density	8.768136e+04
pH	3.439978e+01
Sulphates	6.022061e+01
Alcohol	5.179224e-01

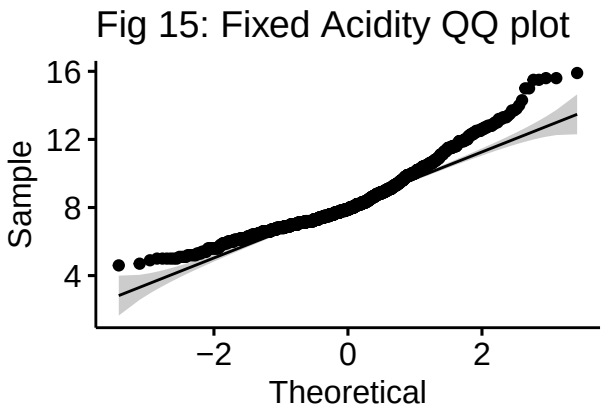
Normality

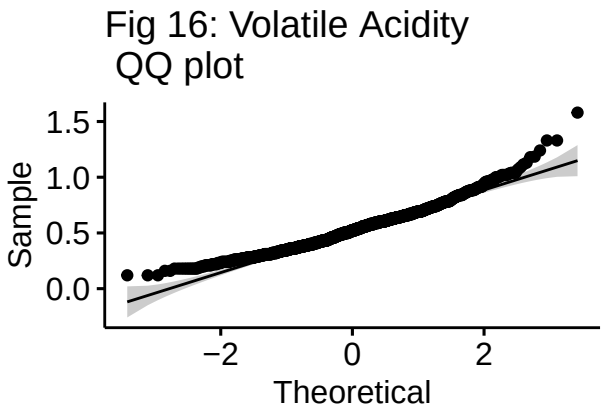
A good way to analyze the normality of the dataset is to look at the qq plot, this is a visual check for the normality of the dataset, the plot shown below shows most of the features except the pH dataset does not appear to be normally distributed, there is significant effect of the outliers on all of the features which is probably pushing the dataset to deviate from normal distribution.

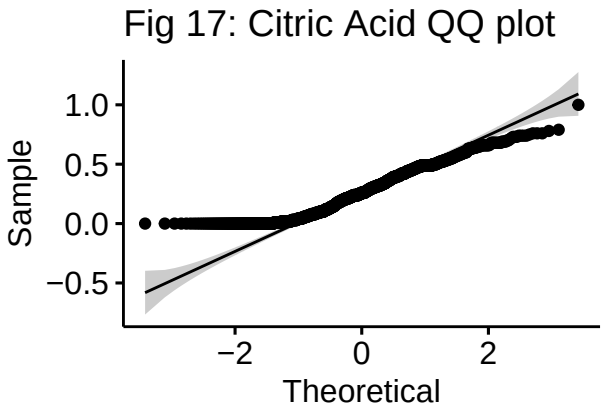
```
install.packages('ggpubr')
```

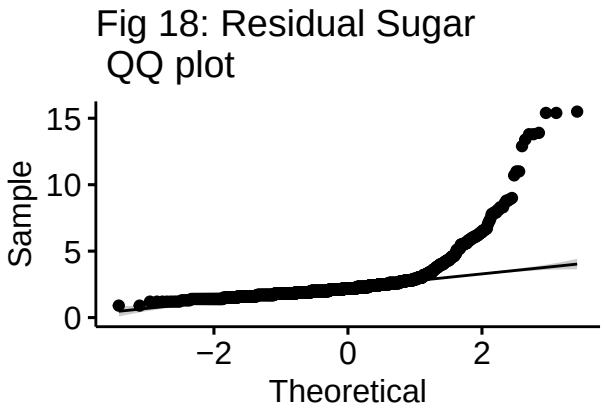
```
## Installing package into '/opt/r'
## (as 'lib' is unspecified)
```

```
library(gridExtra)
library(lattice)
library(ggpubr)
fig1 <- ggqqplot(df_red$fixed.acidity,title='Fig 15: Fixed Acidity QQ plot')
fig2 <- ggqqplot(df_red$volatile.acidity,title='Fig 16: Volatile Acidity \n QQ plot')
fig3 <- ggqqplot(df_red$citric.acid,title='Fig 17: Citric Acid QQ plot')
fig4 <- ggqqplot(df_red$residual.sugar,title='Fig 18: Residual Sugar \n QQ plot')
fig5 <- ggqqplot(df_red$chlorides,title='Fig 19: Chlorides QQ plot')
fig6 <- ggqqplot(df_red$free.sulfur.dioxide,title='Fig 20: Free Sulfur Dioxide \n QQ plot')
fig7 <- ggqqplot(df_red$total.sulfur.dioxide,title='Fig 21: Total Sulfur Dioxide \n QQ plot')
fig8 <- ggqqplot(df_red$density,title='Fig 22: Density QQ plot')
fig9 <- ggqqplot(df_red$pH,title='Fig 23: pH QQ plot')
fig10 <- ggqqplot(df_red$sulphates,title='Fig 24: Sulphates QQ plot')
fig11 <- ggqqplot(df_red$alcohol,title='Fig 25: Alcohol QQ plot')
fig12 <- ggqqplot(df_red$quality,title='Fig 26: Quality QQ plot')

grid.arrange(fig1, fig2,fig3, fig4, ncol = 2, nrow = 2)
```

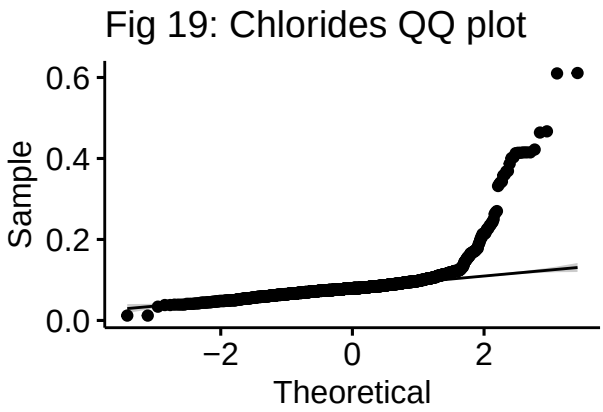


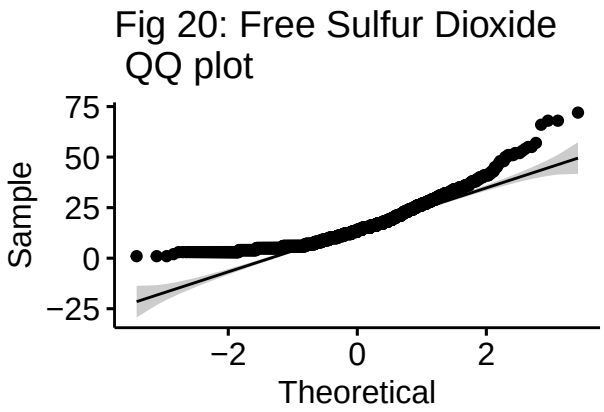


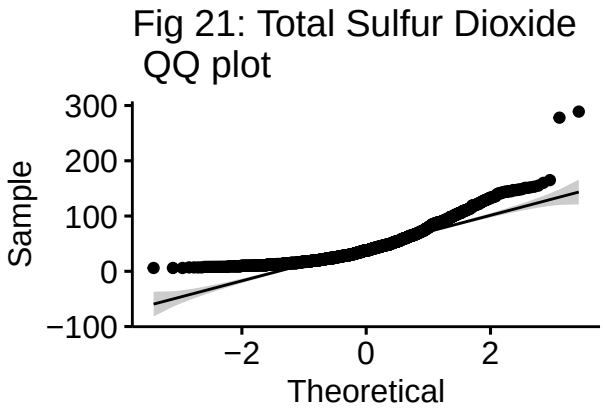


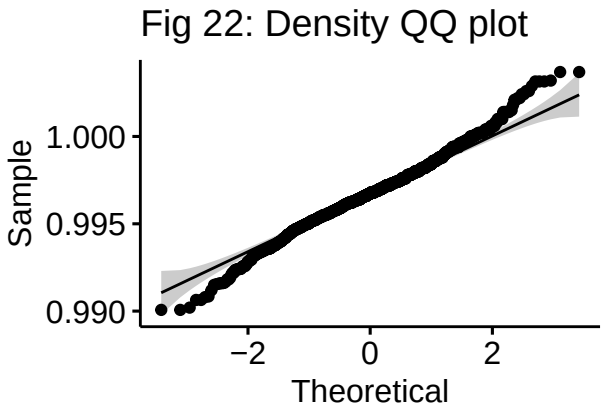



```
grid.arrange( fig5, fig6,fig7, fig8, ncol = 2, nrow = 2)
```










```
grid.arrange( fig9, fig10, fig11,fig12, ncol = 2, nrow = 2)
```

Fig 23: pH QQ plot

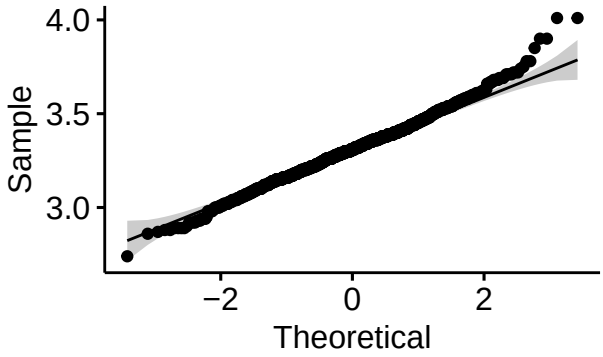


Fig 24: Sulphates QQ plot

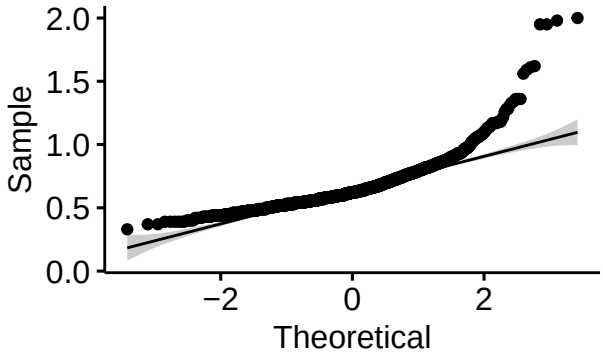


Fig 25: Alcohol QQ plot

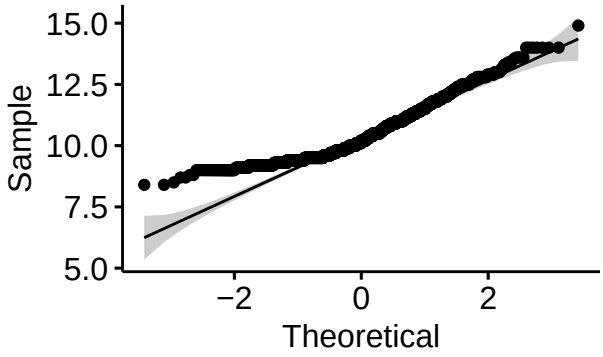
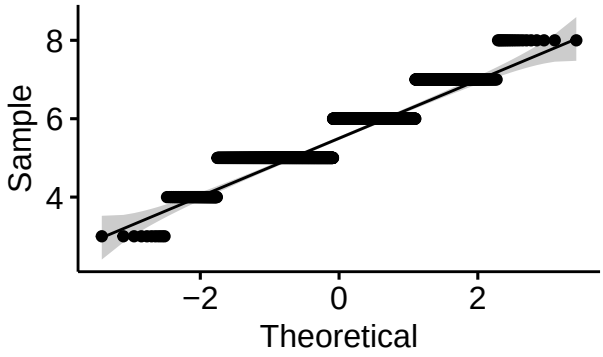


Fig 26: Quality QQ plot



IID Best way to check for IID is to look at correlation analysis between the features for the dataset, in this case we will refer to the EDA correlation of analysis of the dataset from Red wine and White Wine. Based on the analysis, the majority of the data features seem to be IID, except for pH and fixed acidity which demonstrate negative correlation.

Linear Regression Analysis Linear regression methodology for this dataset will be made based on the findings from the large sample assumptions. In general we will divide the analysis into the following parts to make the model creation part to be more systematic. Model fitting will be carried out on a randomly separated regression data set, which has been separated out from the exploratory dataset. Overall, we are using 70% of the total dataset for building the model and 30% for utilizing in the test (theory) phase. 1.)General model with all the features to understand significant effects 2.)Stepwise removing the features which are not able explain the total variance in the quality variable for both wine dataset 3.)Understanding the effect of transformation on the X's to improve the model parameters and explain the variance 4.)Any possible effect of interaction between the significant variables for either dataset 5.)Utilizing the most appropriate model to predict the dataset for both Red and White wine datasets

#Linear model for Red and White wine

```
model_red_all<- lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.su
model_white_all <- lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free
summary(model_red_all)
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + volatile.acidity, data = df_red)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
## volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
## chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
## total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density              -1.788e+01  2.163e+01  -0.827   0.4086
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
summary(model_white_all)
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + volatile.acidity, data = df_white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
## fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
## volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## citric.acid           2.209e-02  9.577e-02   0.231  0.81759
## residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
## chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
## free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
## total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
## density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
## sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
## alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

```
library(stargazer)
stargazer(
  model_red_all,
  model_white_all,
  type='text',title = c("Summary of Red and Whites Wine"),
  header=FALSE,font.size = "huge",
  star.cutoffs=c(0.05,0.01,0.001)
)
```

```
##
## Summary of Red and Whites Wine
## ================================================================================
##                                        Dependent variable:
##                          ---------------------------------------------------
##                                               quality
##                                  (1)                          (2)
## --------------------------------------------------------------------------------
## fixed.acidity                    0.025                       0.066**
##                                 (0.026)                     (0.021)
##
## volatile.acidity               -1.084***                   -1.863***
##                                 (0.121)                     (0.114)
##
## citric.acid                     -0.183                       0.022
##                                 (0.147)                     (0.096)
##
## residual.sugar                   0.016                      0.081***
##                                 (0.015)                     (0.008)
##
## chlorides                      -1.874***                     -0.247
##                                 (0.419)                     (0.547)
##
## free.sulfur.dioxide             0.004*                      0.004***
##                                 (0.002)                     (0.001)
##
## total.sulfur.dioxide           -0.003***                    -0.0003
##                                 (0.001)                     (0.0004)
##
## density                         -17.881                    -150.284***
##                                (21.633)                     (19.075)
##
## pH                              -0.414*                     0.686***
##                                 (0.192)                     (0.105)
##
## sulphates                      0.916***                     0.631***
##                                 (0.114)                     (0.100)
##
## alcohol                        0.276***                     0.193***
##                                 (0.026)                     (0.024)
##
## Constant                        21.965                     150.193***
```

```
##                                (21.195)                  (18.804)
##
## ------------------------------------------------------------------------
## Observations                     1,599                     4,898
## R2                               0.361                     0.282
## Adjusted R2                      0.356                     0.280
## Residual Std. Error      0.648 (df = 1587)         0.751 (df = 4886)
## F Statistic          81.348*** (df = 11; 1587) 174.344*** (df = 11; 4886)
## ========================================================================
## Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

Model Summary: For the initial regression model build we regress on all of the indicator variables in an attempt to understand the significance of each of the features. From the linear regression model below the main conclusions are
1.) overall the linear model is poor at explaining the variance in quality, with R2 adj. Only at 34.3%. 2.) variables such as fixed acidity, citric acid, residual sugar, free Sulfur Dioxide, density, pH show insignificant influence on explaining the variance in red wine quality. We will pursue variable transformation in later models to understand if these variables display higher significance. 3.) for significant variables: chlorides, sulfates, alcohol, total sulfur dioxide and volatile acidity are statistically significant at explaining the variance in red wine quality. 4.) Based on the linear estimate, Volatile acidity seems to have an overall negative effect on the quality of the wine, for Ceteris paribus every one unit change in volatile acidity, it decreases the overall Red wine quality by -1.1 units. The same feature seems to have a slightly higher effect in the case of White wine, as the estimate is at -1.8, suggesting higher impact of the white wine quality over the amounts of volatile acidity of the wine. 5.) Residual sugar is a significant variable for white wine with a much lower estimate of 0.102 compared to Red wine; the variable is not significant at all. 6.) Effect of chlorides is significant with Red wine compared to white wine, as Red wine has a liner estimate at -1.6 suggesting the overall quality of the Red wine is significantly influenced by the presence of chlorides, and for every 1 unit of chloride present it reduces the overall wine score by -1.6 units with all being equal. 7.) Free sulfur dioxide is a significant variable with White wine, but its overall effect is much lower at estimate of -0.003, for Red wine this variable is shown to be not a significant effect on measuring the quality of the wine. Total sulfur dioxide seems to have the exact opposite effect of free sulfur dioxide on Red wine as this variable is significant but its effect on the overall quality is negligibly small. 8.) Density has a significant higher estimate for Red wine compared to White wine, the overall std. Error is also quite high. This could be an effect of the distribution or an effect of outlier on the overall model, we may need to remove some of the outliers for white wine to completely understand its effect. 9.) pH is measure of the amount of H+ ions in the solution, so far we have seen significance from variables such as fixed acidity (it's the measure of the acidity of the wine), Volatile acidity(it's the volatile organic acid components, probably the by product of fermenting process), in case of white wine, pH is a significant variable and has a higher estimate for its estimator at 0.83, suggesting higher pH white wine tends to score higher in overall quality of the White wine.

#Linear model for Red and White Wine Model1 Residual analysis

```
print("Formal Test of Residual Normality- Red Wine")
```

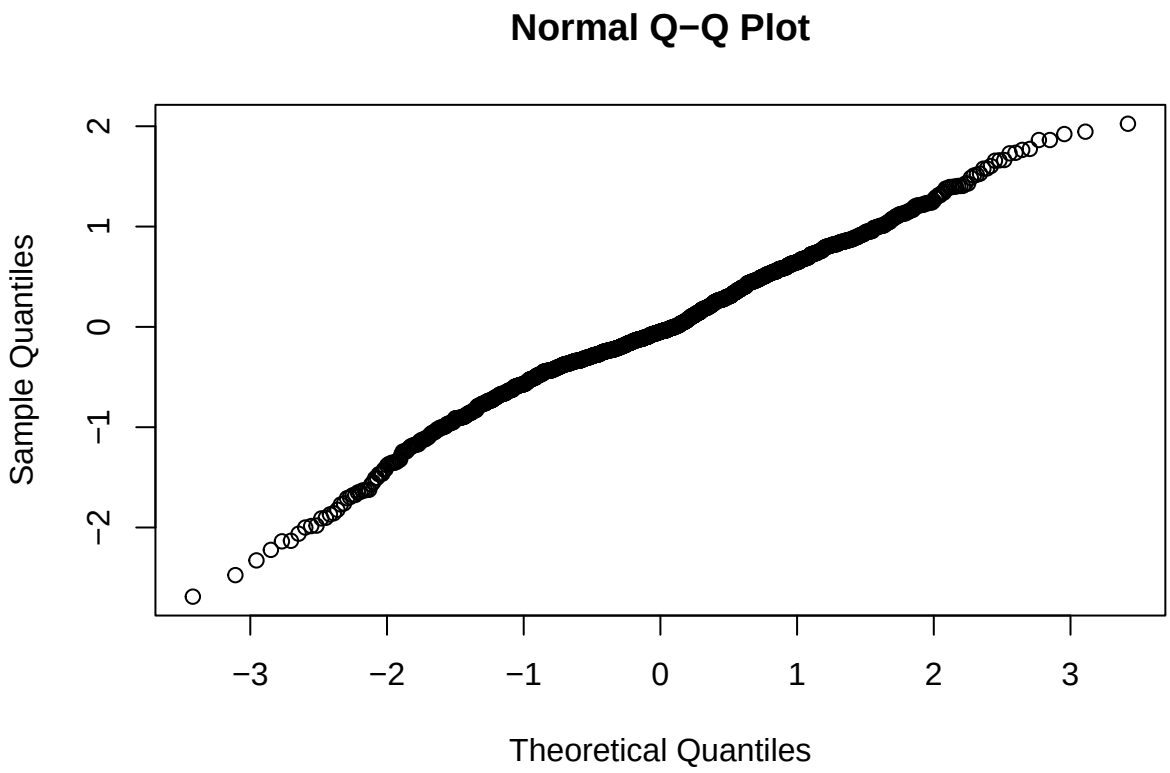
```
## [1] "Formal Test of Residual Normality- Red Wine"
```

```
set.seed(56423)
shapiro.test(sample(model_red_all$residuals,size=5000,replace = TRUE))
```

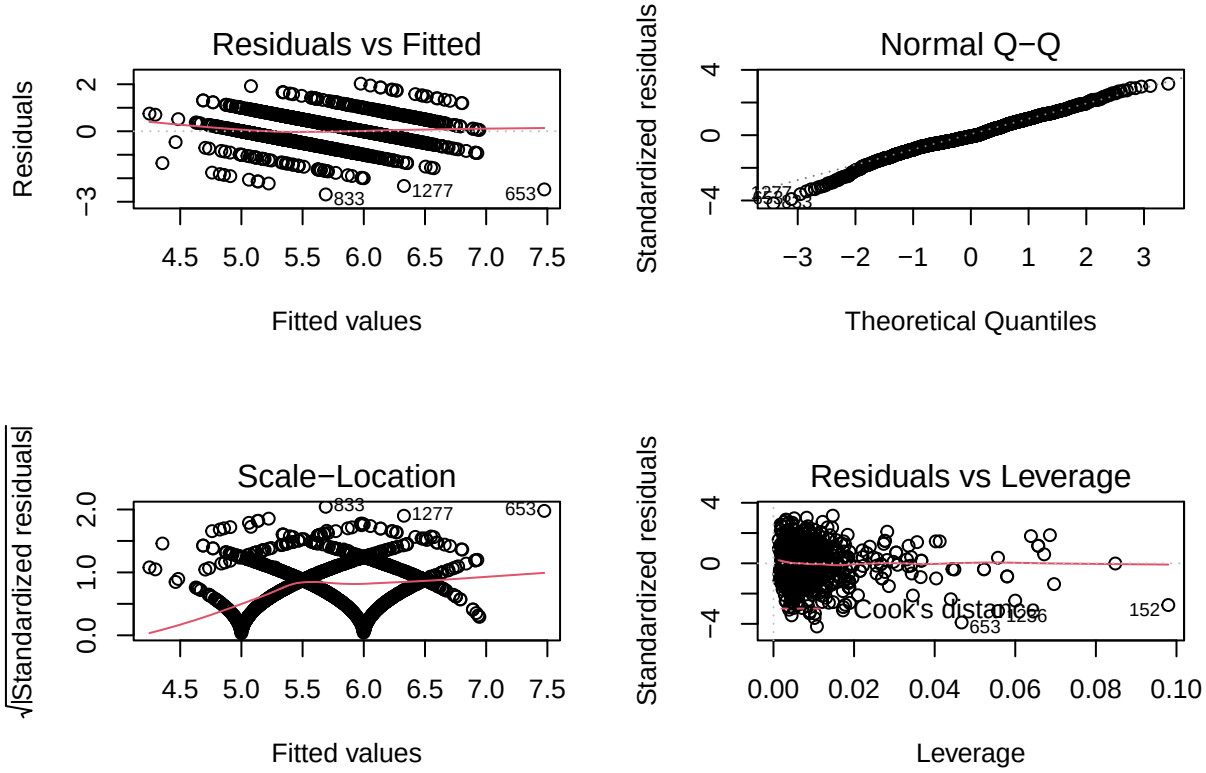
```
##
##  Shapiro-Wilk normality test
##
## data:  sample(model_red_all$residuals, size = 5000, replace = TRUE)
```

```
## W = 0.99074, p-value < 2.2e-16
```

```
qqnorm(model_red_all$residuals)
```




```
par(mfrow=c(2,2))
plot(model_red_all)
```



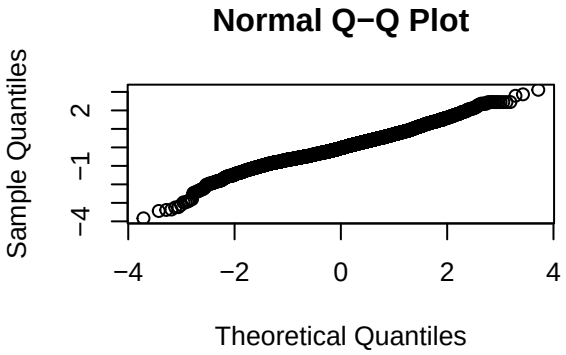

```
print("Formal Test of Residual Normality- White Wine")
```

```
## [1] "Formal Test of Residual Normality- White Wine"
```

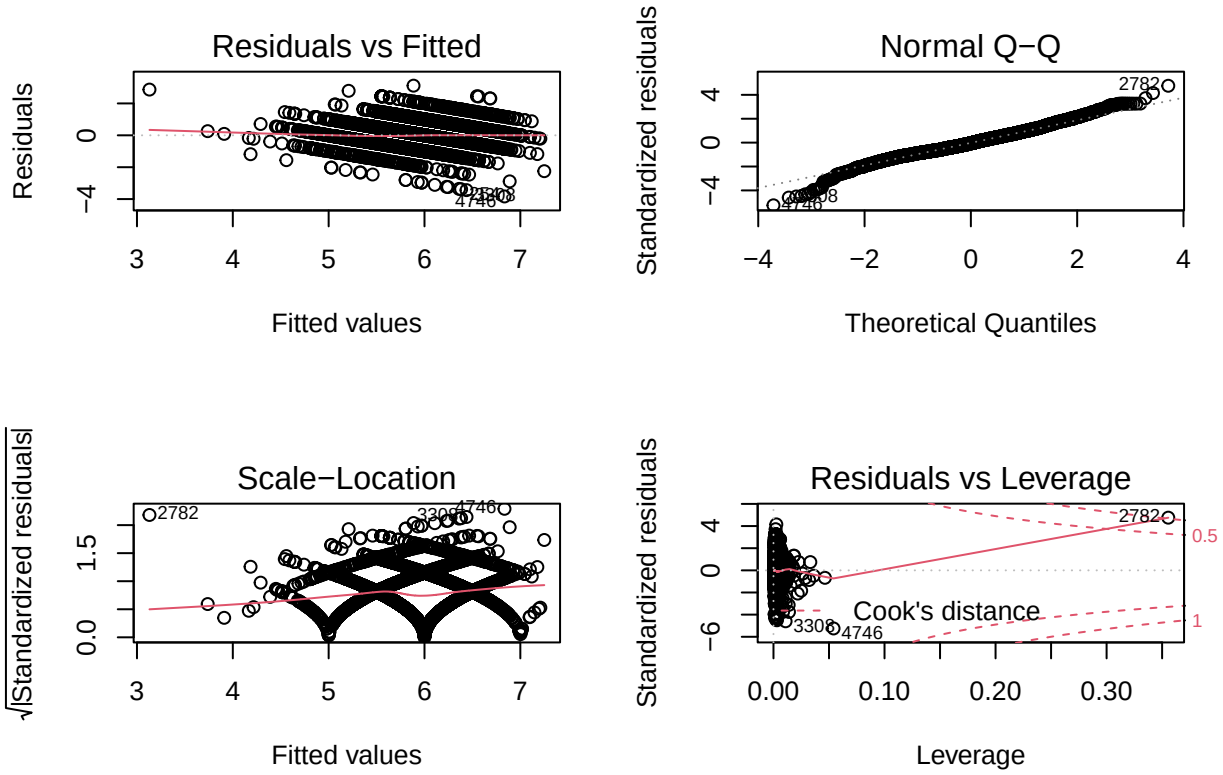
```
set.seed(56423)
shapiro.test(sample(model_white_all$residuals,size=5000,replace = TRUE))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(model_white_all$residuals, size = 5000, replace = TRUE)
## W = 0.9898, p-value < 2.2e-16
```

```
qqnorm(model_white_all$residuals)
par(mfrow=c(2,2))
```




```
plot(model_white_all)
```



>Residual Analysis for Model: Residual analysis for the both the Red and White wines train only data suggests the model fitted isn't good at explaining the large variance seen in the quality feature, and the residual Vs fitted does have a significant dispersion away from 0, this could be due to several factors such as 1.)Overall data suggests there is a linear relationship between quality and the X's, it seems to suggest the model might be more effective at fu 2.)Likert scale effect of the quality variable, we do see the same effect on the QQ plot for quality which suggest our Y variable is not continuous but more like a Likert scale 3.)The QQ plot of the residuals suggests they are not normally distributed 4.)Residual Vs leverage data shows some points having significant effect on the overall quality variable, this could be due to outliers in the dataset. 5.)Overall the residual and model summary analysis is very similar for both Red and White wine dataset Next steps in the model would be to understand the inflation factor or VIF for both the models to better understand if adding all 12 X variables has an effect on the inflationary effect of certain variables. Next steps in the model would be to understand the inflation factor or VIF for both the models to better understand if adding all 12 X variables has an effect on the inflationary effect of certain variables.

#Linear model for Red and White Wine-Model1 VIF analysis

```
#create vector of VIF values
library(car)
```

```
## Loading required package: carData
```

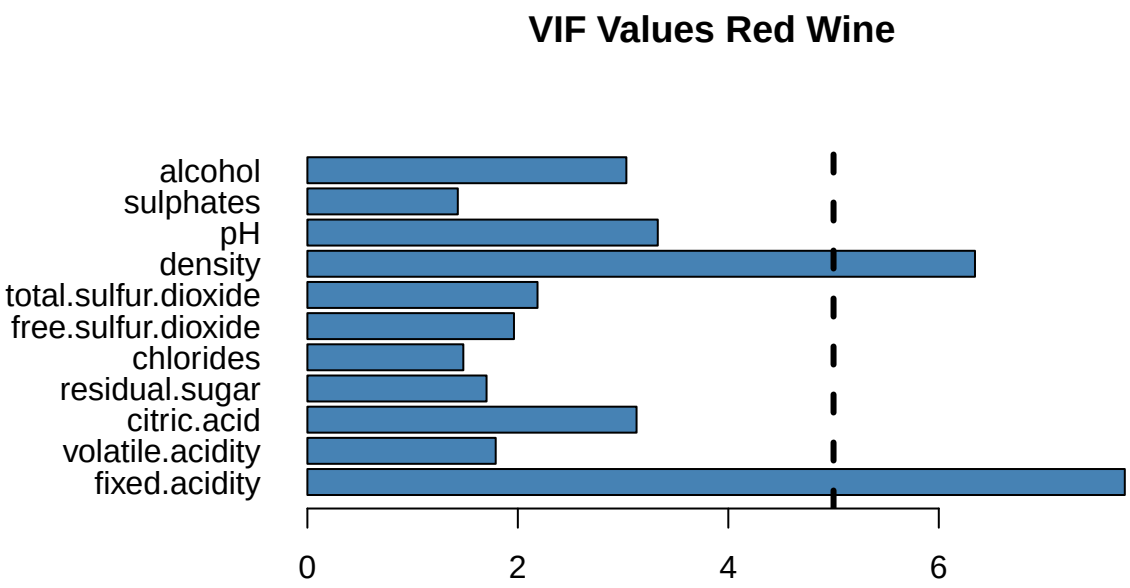
```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

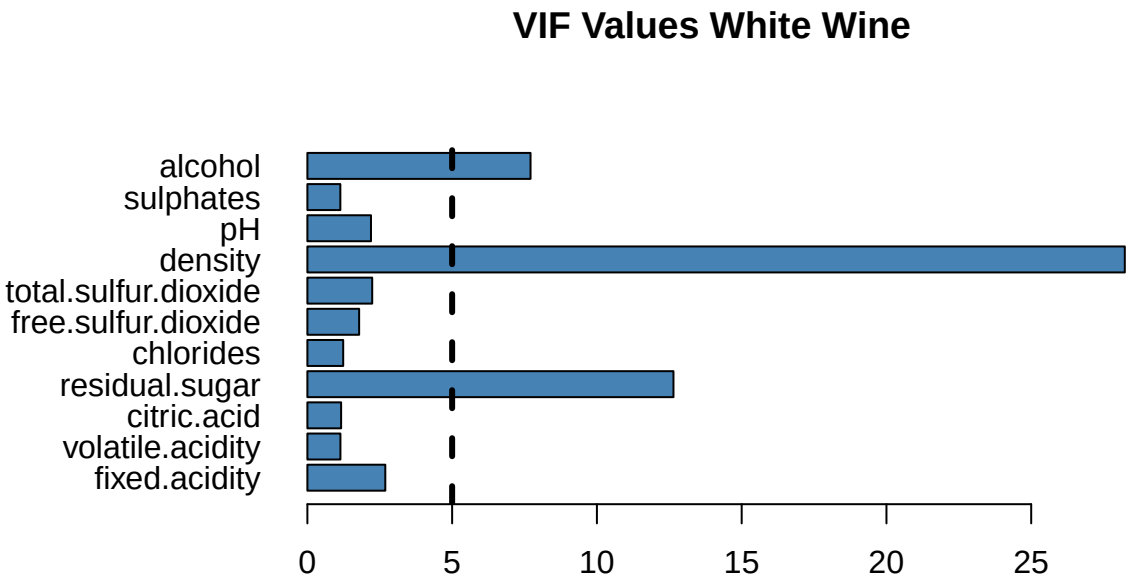
```
## The following object is masked from 'package:purrr':
```

```
##
##     some
```

```
vif_values <- vif(model_red_all)

#create horizontal bar chart to display each VIF value
par(mar=c(5,6,4,1)+2)
barplot(vif_values, main = "VIF Values Red Wine", horiz = TRUE, col = "steelblue",
        axes=TRUE, cex.names=1, las=1)

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```




```
vif_values_w <- vif(model_white_all)

#create horizontal bar chart to display each VIF value
par(mar=c(5,6,4,1)+2)
barplot(vif_values_w, main = "VIF Values White Wine", horiz = TRUE, col = "steelblue",
        axes=TRUE, cex.names=1, las=1)

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```



>Variance Inflation Factor(VIF) analysis: Variance Inflation Factor(VIF) analysis for both Red and Wine models shows the density to have a significant inflatory effect for both types of wine, in general a VIF value of <10 is considered acceptable and <5 is nominal. VIF is also an indicator of the collinearity of the variables, the above graph suggests significant collinearity with variables such as density, residual sugar. The first explanatory mode needs more improvements for both Red and White dataset and we will be exploring the effect of removing some of the insignificant variables and effect of transformation on the new models.

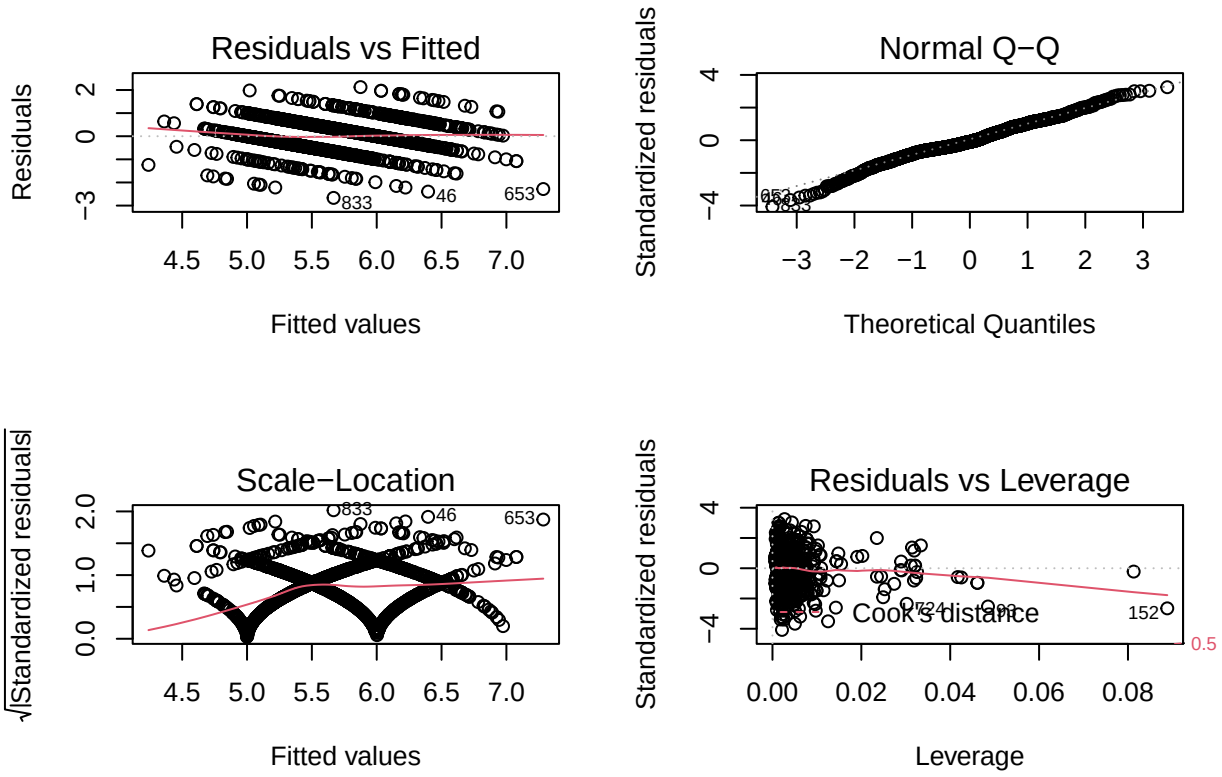
Improved model for Red Wine

```
model_red_i <- lm(quality ~ volatile.acidity+chlorides+sulphates+alcohol, data=df_red)
summary(model_red_i)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + sulphates +
##     alcohol, data = df_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66747 -0.38061 -0.06736  0.46303  2.12482
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.77676    0.19870  13.975  < 2e-16 ***
## volatile.acidity -1.16746    0.09737 -11.990  < 2e-16 ***
## chlorides        -1.64511    0.39387  -4.177 3.12e-05 ***
```

```
## sulphates          0.87356    0.11057   7.900 5.14e-15 ***
## alcohol            0.29209    0.01625  17.975  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6554 on 1594 degrees of freedom
## Multiple R-squared:  0.3431, Adjusted R-squared:  0.3414
## F-statistic: 208.1 on 4 and 1594 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model_red_i)
```



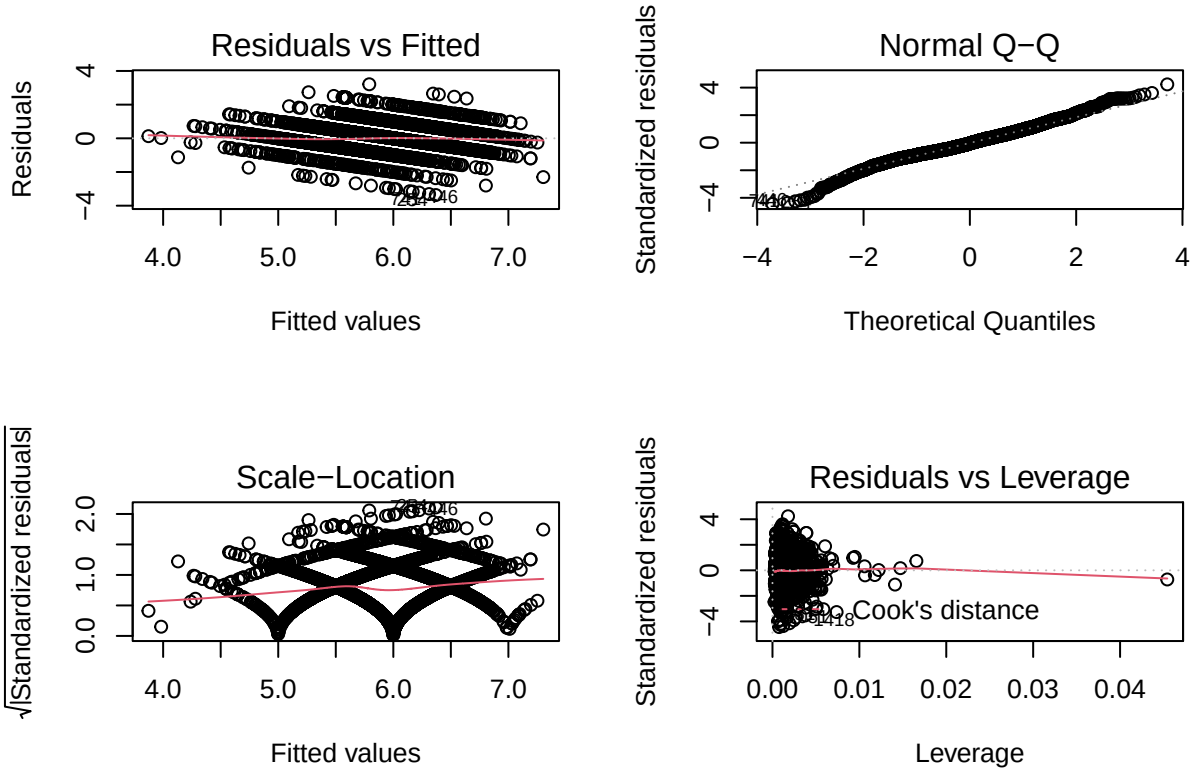
Improved model for White Wine

```
model_white_i <- lm(quality ~ volatile.acidity+sulphates+alcohol+residual.sugar+fixed.acidity
                    , data=df_white)
summary(model_white_i)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + sulphates + alcohol +
##     residual.sugar + fixed.acidity, data = df_white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3693 -0.4939 -0.0341  0.4634  3.2090
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.671921   0.158935  16.811  < 2e-16 ***
## volatile.acidity -2.102798   0.108510 -19.379  < 2e-16 ***
## sulphates         0.443294   0.095155   4.659 3.27e-06 ***
## alcohol           0.370957   0.009972  37.199  < 2e-16 ***
## residual.sugar    0.027559   0.002412  11.427  < 2e-16 ***
## fixed.acidity    -0.073315   0.012958  -5.658 1.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7588 on 4892 degrees of freedom
## Multiple R-squared:  0.2667, Adjusted R-squared:  0.266
## F-statistic: 355.9 on 5 and 4892 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model_white_i)
```



#Comparison beween the models for Red and White Wine

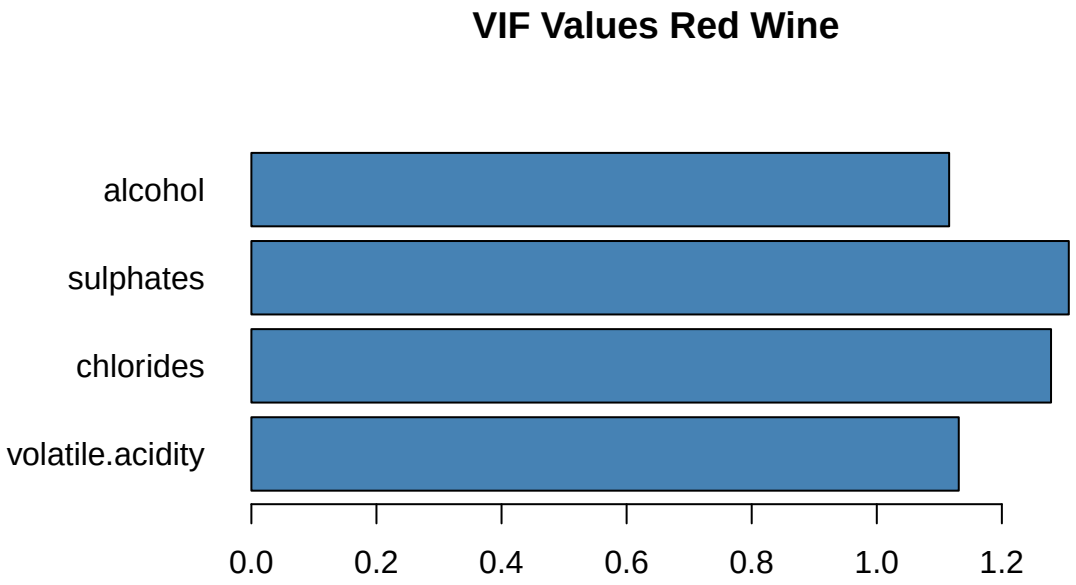
```
library(stargazer)
stargazer(
  model_red_i,
  model_white_i,
  type='text',title = c("Summary of Red and White Wine Model"),
  header=TRUE,font.size = "huge",
```

```
  star.cutoffs=c(0.05,0.01,0.001)
)
```

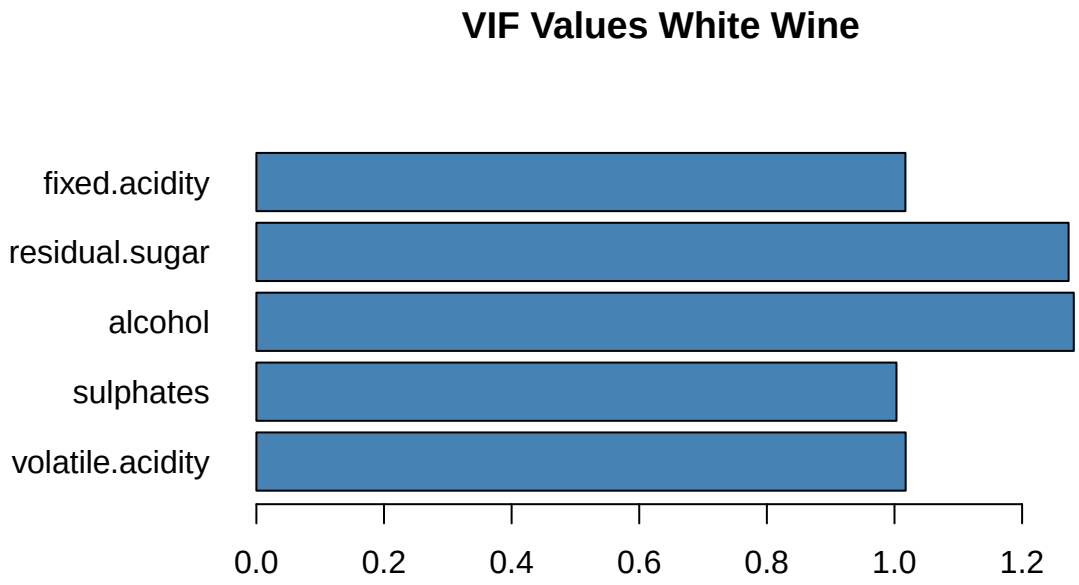
```
##
## Summary of Red and White Wine Model
## ================================================================================
##                                        Dependent variable:
##                          ---------------------------------------------------
##                                              quality
##                                  (1)                          (2)
## --------------------------------------------------------------------------------
## volatile.acidity               -1.167***                    -2.103***
##                                 (0.097)                      (0.109)
##
## chlorides                      -1.645***
##                                 (0.394)
##
## sulphates                      0.874***                     0.443***
##                                 (0.111)                      (0.095)
##
## alcohol                        0.292***                     0.371***
##                                 (0.016)                      (0.010)
##
## residual.sugar                                              0.028***
##                                                              (0.002)
##
## fixed.acidity                                               -0.073***
##                                                              (0.013)
##
## Constant                       2.777***                     2.672***
##                                 (0.199)                      (0.159)
##
## --------------------------------------------------------------------------------
## Observations                     1,599                        4,898
## R2                               0.343                        0.267
## Adjusted R2                      0.341                        0.266
## Residual Std. Error       0.655 (df = 1594)            0.759 (df = 4892)
## F Statistic           208.125*** (df = 4; 1594) 355.890*** (df = 5; 4892)
## ================================================================================
## Note:                                          *p<0.05; **p<0.01; ***p<0.001
```

VIF analysis for the refined model

```
vif_values_r <- vif(model_red_i)

#create horizontal bar chart to display each VIF value
par(mar=c(5,6,4,1)+2)
barplot(vif_values_r, main = "VIF Values Red Wine", horiz = TRUE, col = "steelblue",
        axes=TRUE, cex.names=1, las=1)

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```




```
vif_values_w_i <- vif(model_white_i)

#create horizontal bar chart to display each VIF value
par(mar=c(5,6,4,1)+2)
barplot(vif_values_w_i, main = "VIF Values White Wine", horiz = TRUE, col = "steelblue",
        axes=TRUE, cex.names=1, las=1)

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```



>Linear modeling - Improved Modeling Parameters 1.)Overall R2 adj for both Red and White wine is still quite low suggesting the linear model suggested is not fully capable of explaining the variance seen with the quality of the wine. 2.)The values of significant estimators haven't changed significantly, this is a good indication that we have a unique BLP for each of the parameters, which satisfies the large sample assumptions for linear regression. 3.)Residual analysis for the new model shows similar trend between the fitted value and the residuals, the dispersion observed in the residual is an indication of the likert like scale effect of the quality parameter for both Red and White wine 4.)Removing variable such as density and residual sugar, shows improvement in the VIF, which is a good indicator of collinearity of the dataset, it seems with the new model we don't observe as much collinearity as the previous model

#Anova analysis of the models

```
anova(model_red_all,model_red_i ,test='F')
```

```
## Analysis of Variance Table
##
## Model 1: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + volatile.acidity
## Model 2: quality ~ volatile.acidity + chlorides + sulphates + alcohol
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1587 666.41
## 2   1594 684.61 -7   -18.201 6.1922 3.596e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 anova(model_white_all,model_white_i ,test='F')
```

```
## Analysis of Variance Table
```

```
##
## Model 1: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + volatile.acidity
## Model 2: quality ~ volatile.acidity + sulphates + alcohol + residual.sugar +
##     fixed.acidity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4886 2758.3
## 2   4892 2816.5 -6   -58.168 17.173 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova Test for selecting the best model From the anova table for the red wine dataset, it shows the second model which is the more refined model with removing non significant terms does show to be a more significant model compared to the model with all the terms. This holds true for white wine data sets too. Based on this analysis we can conclude that for good quality Red wine we want less of volatile acidity, chlorides and more of sulfates and alcohol. Whereas for white wine for a high quality wine we would want less of volatile acidity and fixed acidity components and more of sulfates and alcohol concentration. Though this is an oversimplification of how good a wine is, one could generalize in terms of these chemical components from the given dataset.

```
#Alternative methods - Binomial Logistic Regression model

#Red wine Logistic Regression
df_red$quality_ind<-0
df_red$quality_ind[df_red$quality>6]<-1

model_red_logit <- glm(formula = quality_ind ~
                  `fixed.acidity`
                + `volatile.acidity`
                #+ `citric.acid`
                + `residual.sugar`
                + chlorides
                #+ `free.sulfur.dioxide`
                + `total.sulfur.dioxide`
                + density
                #+ pH
                + sulphates
                + alcohol
                , data = df_red, family = binomial)

summary(model_red_logit)
```

```
##
## Call:
## glm(formula = quality_ind ~ fixed.acidity + volatile.acidity +
##     residual.sugar + chlorides + total.sulfur.dioxide + density +
##     sulphates + alcohol, family = binomial, data = df_red)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0158  -0.4314  -0.2220  -0.1255   2.9883
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)           2.268e+02  9.163e+01   2.475 0.013336 *
## fixed.acidity         2.812e-01  8.029e-02   3.502 0.000462 ***
## volatile.acidity     -2.913e+00  6.467e-01  -4.504 6.66e-06 ***
## residual.sugar        2.328e-01  7.009e-02   3.322 0.000893 ***
## chlorides            -8.441e+00  3.259e+00  -2.590 0.009593 **
## total.sulfur.dioxide -1.360e-02  3.447e-03  -3.946 7.95e-05 ***
## density              -2.409e+02  9.202e+01  -2.618 0.008835 **
## sulphates             3.699e+00  5.287e-01   6.997 2.62e-12 ***
## alcohol               7.823e-01  1.120e-01   6.983 2.88e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  872.08  on 1590  degrees of freedom
## AIC: 890.08
##
## Number of Fisher Scoring iterations: 6
```

Logistic regression models are generally benchmarked using AIC (Akaike Information Criterian) or BIC (Bayesian Information Criterian) The logit model for red wine yields a pretty high AIC of 890

```
#White wine Logistic Regression
df_white$quality_ind<-0
df_white$quality_ind[df_white$quality>6]<-1

model_white_logit <- glm(formula = quality_ind ~
                  `fixed.acidity`
                + `volatile.acidity`
                #+ `citric.acid`
                + `residual.sugar`
                + chlorides
                + `free.sulfur.dioxide`
                #+ `total.sulfur.dioxide`
                + density
                + pH
                + sulphates
                #+ alcohol
                , data = df_white, family = binomial)

summary(model_white_logit)
```

```
##
## Call:
## glm(formula = quality_ind ~ fixed.acidity + volatile.acidity +
##     residual.sugar + chlorides + free.sulfur.dioxide + density +
##     pH + sulphates, family = binomial, data = df_white)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3155  -0.6744  -0.4115  -0.1802   2.7756
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          7.405e+02  3.448e+01  21.473  < 2e-16 ***
## fixed.acidity        6.006e-01  6.349e-02   9.459  < 2e-16 ***
## volatile.acidity    -3.550e+00  4.601e-01  -7.714 1.21e-14 ***
## residual.sugar       3.298e-01  1.925e-02  17.134  < 2e-16 ***
## chlorides           -1.249e+01  3.772e+00  -3.310 0.000933 ***
## free.sulfur.dioxide  8.048e-03  2.470e-03   3.259 0.001118 **
## density             -7.644e+02  3.558e+01 -21.485  < 2e-16 ***
## pH                   3.657e+00  3.301e-01  11.079  < 2e-16 ***
## sulphates            2.266e+00  3.284e-01   6.901 5.17e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5116.8  on 4897  degrees of freedom
## Residual deviance: 4147.9  on 4889  degrees of freedom
## AIC: 4165.9
##
## Number of Fisher Scoring iterations: 5
```

The logit model for white wine yields an evenworse AIC of 4165 Given that this attempt at an alternative model is out of the covered content in the course and because model performance was sub-par based on our existing knowledge, we decided not to delve further into this approach.

Omitted Variables:

Age: The age of wine can be a strong indicator of quality, as it gives more time for chemicals to interact and flavor to develop. Year: Wine quality can often be demarcated by the year the grapes were grown. This variable was not collected in the experiment which could be a key indicator in objective quality of the wine. Wine Type: Further separation of data to the type of wine could hold information about what chemical attributes are better indicators for different types of wine.

Conclusion

Creating an excellent wine is not an exact science provided just the physicochemical characteristics of the drink. It's difficult to predict what makes an excellent wine from the major chemical variables as wine has hundreds to thousands of microbial characteristics that affect the taste, smell, and appearance of the product. Our team fails to reject the null hypothesis that: "there is no one equation of physicochemical properties that will predict the quality of a bottle of wine". However, we do believe that it is possible to predict what makes a bad bottle of wine. Through binning our data into good and bad wines it is easy to see a higher dispersion of chemical measurements on our bad bottles of wine. Thus in a future study we propose an outlier regression model that would measure for quantities of chemicals above the level accepted for good wines. Practical Uses for our causal model include quality assurance for the Vinho Verde vineyards as it would be highly beneficial to understand if a wine bottle is going to be judged poorly prior to it becoming bottled.

Sources

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009