Regression Analysis: Estimating Alcohol Content Level in Wine

Gabriel Vasquez

2024-12-13

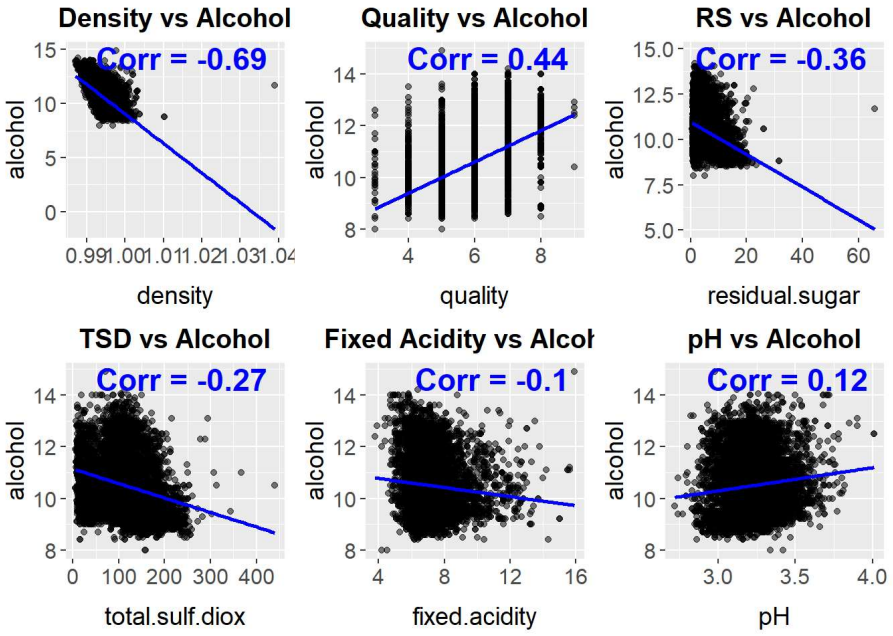
Introduction

We know the consumers' perception of their preferred quality of wine is often linked to flavor and alcohol content. Red wine typically has higher alcohol content than white wine, while white wine generally provides more of a sweeter taste because of its lower alcohol content. The balance between sweetness and alcohol content involves a complex process called fermentation - the process in which yeast converts simple sugars to ethanol(alcohol) through various chemical reaction instances. The sweetness and alcohol content in a typical wine beverage is determined by the amount of residual sugar left after fermentation.

For our project, we were interested in determining if we can derive a regression model that can be used to estimate the amount of alcohol content in a wine beverage. We questioned, "which combination of fermentation features can provide the best alcohol content estimate in wine?" Our goal was to investigate which combination of these features in have the most influence on estimating alcohol content and build a regression on top of them. We observed wine samples from the Portuguese wine company "Vihno Verde" to examine which chemicals and fermentation features most influence the amount of alcohol content in wine.

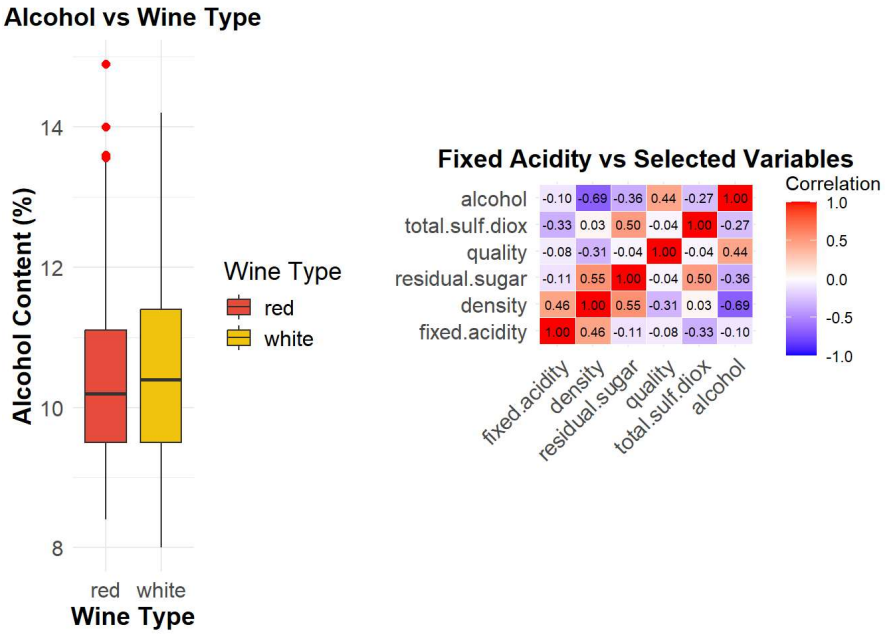
The dataset is comprised of 6,497 red and white wine samples. There are 13 total variables. They are described as follows:
1. **Fixed Acidity** (g/dm3) -Wines with balanced acidity support better yeast activity, which can indirectly affect alcohol production by promoting efficient fermentation.
2 **Volatile Acidity** (g/dm3) -A measure of the wine's gaseous acids that contribute to the smell and taste of vinegar in wine.
3. **Citric Acid** (g/dm³) -Citric acid is less influential on alcohol content but can enhance nutrients to yeast.
4. **Residual Sugar** (g/dm³) -Lower residual sugar usually means more sugar was converted to alcohol during fermentation, leading to higher alcohol content. Residual sugar is unfermented sugar.
5. **Chlorides** (Sodium Chloride, g/dm³) -This effects the salinity of wine and are generally unrelated to alcohol content.
6. **Free Sulfur Dioxide** (mg/dm³) -Excessive amount can inhibit fermentation, reducing alcohol production. Preserves wine.
7. **Total Sulfur Dioxide** (mg/dm³) -Similar to free sulfur dioxide, total sulfur dioxide preserves wine but may inhibit yeast activity if levels are too high, impacting alcohol levels.
8. **Density** (g/cm³) -Wine density is an indicator of alcohol content. As sugar ferments into alcohol, wine density decreases.
9. **pH** - An optimal pH range fosters healthy yeast, ensuring efficient sugar conversion to alcohol.
10. **Sulphates** (g/dm³) -Sulfates stabilize by influencing fermentation conditions. Doesn't directly support alcohol production.
11. **Quality** (scaled: 0 to 10) -While quality is subjective, higher-quality wines might result from well-managed fermentation, potentially correlating with balanced alcohol levels.
12. **Wine Type** -Red and white. Different color profiles have different variations of fermentation practices and sugar content.
13. **Alcohol** (Volume) – the amount of alcohol in a wine sample.

EDA & Model Selection



We checked all correlations with our response variable (alcohol) from a scatter plot matrix. We selected Density, Residual Sugar, and Total Sulfur Dioxide as our predictors because they had the highest correlations compared to all other independent variables. These 3 features are noted by our research sources to have very important influences on alcohol production during fermentation. We also selected Quality because of its high correlation coefficient with alcohol, compared to the other independent variables.

We removed all noticeable outliers. We didn't believe this would affect the integrity of the model because we only removed so little number of observations. From the summary statistics, Total Sulfur Dioxide is very skewed. Also, Wine Type failed to follow the given fact that red wine has higher alcohol content than white wine. This is because the data set is comprised more of white wine samples than red wine. Also, since it appears that there is a minimal difference between the two levels in Wine Type, we considered Wine Type too insignificant to include in our regression model. We excluded it as a potential option.



We found Fixed Acidity to be a promising variable to include into the model later. Fixed Acidity has some correlation with the density of wine and the total sulfur dioxide in wine. Higher Fixed Acidity levels lead to less density in wine. Fixed Acidity consists of tartaric acid, which is a relatively light molecule compared to other components in wine. When tartaric acid increases, it displaces heavier molecules - lowering the overall density in wine.

Also,higher Fixed Acidity allows for greater proportion of Sulfur Dioxide to be present in wine. The variable pH plays the middle man between these two molecules because higher Fixed Acidity generally means lower pH. And, lower pH levels indicates more sulfur dioxide molecules to be present - higher Total Sulfur Dioxide - in wine. This developed evidence to possibly include Fixed Acidity and pH into future models. We began further looking into interactions between the predictors themselves.

To determine which variables to include in our second-order regression model, we first identified potential interaction terms that might significantly impact the model's performance. We accomplished this by creating simple models and testing various combinations of the selected predictors. For example, we explored whether density interacted with other predictors such as quality, residual sugar, and total sulfur dioxide.

Through this testing, we observed that the plots for density:residual_sugar and residual_sugar:total_sulf_diox showed clear intersections. These intersections indicate interaction because adjusting the values of residual sugar alters the slope of the line for density. This change in slope reflects that density and residual sugar interact. Similarly, we observed the same trend for residual_sugar:total_sulf_diox, where changes in the slope of total_sulf_diox caused corresponding changes in residual sugar. The intersecting slopes for these predictor combinations suggest strong interactions. In terms of what these intersections mean, after a mean density value of 0.5 we can see that as the residual sugar increases it leads to a higher level of alcohol concentration. As the slope of the residual sugar predictor increases we see the slopes adapt to different behavior after a density value of 0.5. For residual_sugar and total_sulf_diox, we see that after a predictor value of -0.5 the lower amounts of residual sugar lead to higher concentration counts, which indicates strong interaction between total_sulf_diox.

In contrast, other predictor combinations did not show intersecting slopes, indicating no significant interaction. When the slope of one predictor changes, the effect on the other predictor remains relatively consistent (no drastic changes or intersections). These non-interacting combinations are excluded from our appendix, as most predictor combinations lack significant interaction effects. After the addition of these variables, we also check multi-colinearity by checking the VIF scores and all are under 5 (no multi-colinearity).

After capturing these relationships amongst our selected predictors, we ran an ANOVA test to see if our findings were beneficial to your regression model. When we ran the ANOVA, we saw that the inclusion of our interactive terms didn't contribute a ton to the Residual R^2 score (pushed to .70), but all of the assumptions of a linear regression model still pass. Our RSS for the interactive term model seems to be lower than our first order model given an F-score of 24.44 with a p-value of < 2e-16 which indicates that our interactive model's interactive variables contribute to the coverage of residuals in our data.

Regression Analysis

Our initial **first order model**:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

x_1 = Density, x_2 = Quality, x_3 = Residual Sugar, x_4 = Total Sulfur Dioxide

After we had built our first-order linear model, the constant variance and linearity assumptions for all predictors and the residuals appeared too questionable. Also, the distribution of the residuals was not passing normality assumption both graphically and statically. We didn't want to continue further with the next model build until we exhausted all our options to get our assumptions to pass. Transforming the predictors and the response variable satisfied the assumptions enough. Although, linearity was still a bit questionable.

Our diagnostic tools, such as scatter plots and variance inflation factor scores between the predictors, showed no clear existing multicollinearity. Interestingly though, we noticed that the Density vs Residual scatter plot showed a quadratic pattern. This pattern suggested that the relationship between Density and Alcohol is nonlinear. A higher-order term of Density may need to be included in the regression model for better capture the underlying relationship. We planned considering a polynomial regression in one of our next model selections.

We aimed to identify the best model for estimating alcohol content in wines by starting with including key predictors such as **density**, **quality**, **residual sugar**, **total sulfur dioxide**, and **fixed acidity**, which are known to impact alcohol production. Through exploratory analysis, we identified the need for interaction terms to capture complex relationships and quadratic terms to address non-linear effects. Density showed a non-linear relationship with alcohol content. Including a quadratic term improved the model significantly, as seen by an increase in Adjusted R-squared.

Interactions between predictors like density × residual sugar and fixed acidity × total sulfur dioxide were theoretically meaningful, as they reflect fermentation dynamics and chemical interactions. Adding these terms captured the interplay between predictors and further improved the model fit. We compared models using several metrics:
1. Adjusted R-squared: The final model had the highest Adjusted R-squared, explaining 79.7% of the variability in alcohol content.
2. Residual Sum of Squares (RSS): The final model significantly reduced RSS (from 2295.4 in the initial model to 1873.1 in the final model).
3. AIC/BIC: The final model showed lower AIC and BIC values, indicating a better trade-off between model complexity and performance.
4. ANOVA: A statistically significant F-statistic confirmed that adding interaction and quadratic terms improved the model.

Finally, we validated the model's assumptions. Linearity, normality, and homoscedasticity appeared satisfied. Based on these results, we selected this second-order model because it seemed best to balance accuracy, complexity, and interpretability amongst our predictors for estimating alcohol content.

Limitations

We were satisfied with our final model's Adjusted R-Square, Residual Standard Error, the β-coefficients, and the significance of each variable we used. The major issue we ran into was from our residual analysis. The linear assumptions were not satisfied enough. After performing our transformations to the variables in the beginning, we noticed an improvement to normality and constant variance. Linearity also improved, but just not as much as normality and constant variance. We continued with the model anyways. We were expecting that all assumptions would at least be satisfied enough after the transformations. We determined later that we should have considered that linearity failed.

We didn't realize it right away, but we came to learn that the transformations of the variables made it too difficult to interpret the model. The transformations seemed plausible because of two reasons. One, each variable's measurements were scaled differently. And, two, they improved the assumptions check during the residual analysis. However, once we derived our models, we were too unsure as to how to interpret the alcohol content based on the conditions of the predictors.

Conclusion

Since the assumptions were not fully satisfied and the interpretation of our model was too difficult, we agreed that the model was not ideal for estimating alcohol in wine. Our goal was to derive a regression model based on the fermentation features that best estimated alcohol content in a wine sample. We wanted to have at least a 0.70 Adjusted R-square score. We wanted the model to be as simple as possible because we suspected there would more variable interactions that would make the model too complicate if included. We discovered that Fixed Acidity had more to contribute to the model than we expected. We also discovered that pH is not as contributing as we thought it would be.

Lessons Learned & Future Work

We learned that our transformations did not work the way we wanted. We transformed the predictors to help satisfy the assumptions. We standardized the predictors because of their difference in scale. We didn't know this would impact their coefficient estimates. We, also, noticed after building our simple model, the large differences in magnitude of the β-coefficients and large standard errors for some of the coefficients made it difficult to interpret the model. So, we considered that standardizing would help, especially once we included our interaction terms. However, we were not mindful enough about the log-transformations and the response transformation that we did before hand.

Also, we thought that selecting the highest, significant independent variables would be a good starting point for building our model1. We had initially included all the significant variables from research into the model, 9 variables, but this made the model more complicated than we needed. It turns out, we should have been more thoughtful of selecting our predictors. For example, Fixed Acidity had one of the lowest correlations with Alcohol, yet when we included it in the model, it turned out to be more useful than we imagined. We'll be more thorough with our EDA and include a balance of predictor variables that have enough correlation with Alcohol and significance to the fermentation process.

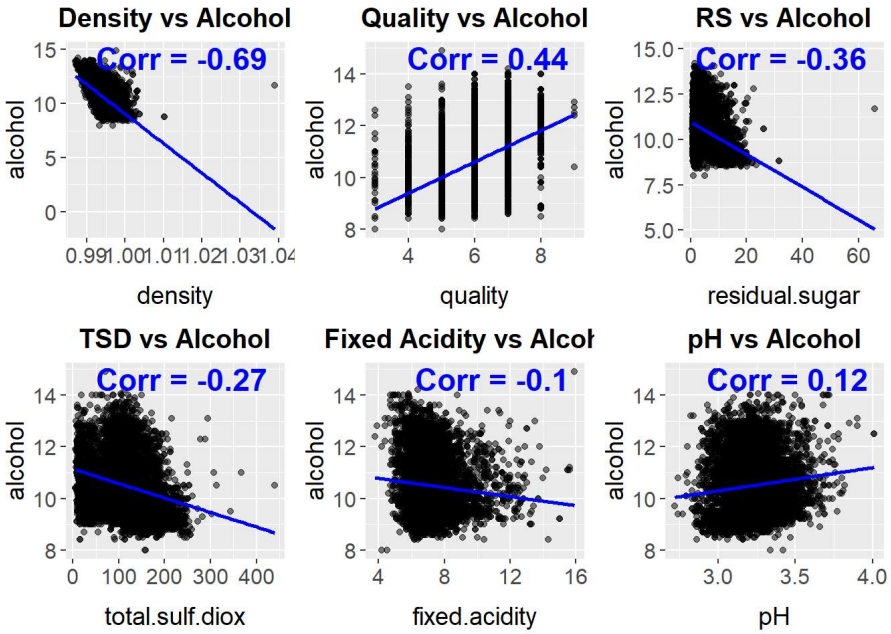
Appendix

EDA and Model Selection

Scatter Plot Matrix (Independent Variables vs Alcohol)

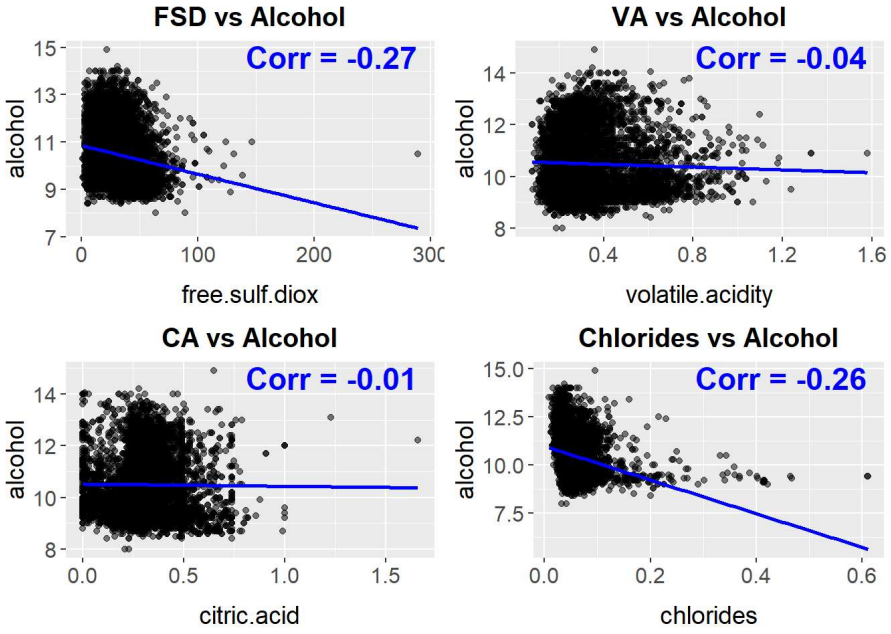
```r
# Calculate correlations to include in the ggplots
cor_density <- cor(wines$density, wines$alcohol, use = "complete.obs")
cor_quality <- cor(wines$quality, wines$alcohol, use = "complete.obs")
cor_residual_sugar <- cor(wines$residual.sugar, wines$alcohol, use = "complete.obs")
cor_total_sulfur_dioxide <- cor(wines$total.sulf.diox, wines$alcohol, use = "complete.obs")
cor_fixedAcidity <- cor(wines$fixed.acidity, wines$alcohol, use = "complete.obs")
cor_pH <- cor(wines$pH, wines$alcohol, use = "complete.obs")
cor_volAcidity <- cor(wines$volatile.acidity, wines$alcohol, use = "complete.obs")
cor_citricAcid <- cor(wines$citric.acid, wines$alcohol, use = "complete.obs")
cor_chlorides <- cor(wines$chlorides, wines$alcohol, use = "complete.obs")
cor_freeSD <- cor(wines$free.sulf.diox, wines$alcohol, use = "complete.obs")
cor_sulphates <- cor(wines$sulphates, wines$alcohol, use = "complete.obs")

## Function to Generate Plots
generate_plot <- function(data, xvar, yvar, cor_value, title) {
  ggplot(data, aes_string(x = xvar, y = yvar)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    annotate("text", x = Inf, y = Inf, label = paste("Corr =", round(cor_value, 2)),
             hjust = 1.1, vjust = 1.2, color = "blue", size = 6, fontface = "bold") +
    ggtitle(title) +
    theme(
      axis.text.x = element_text(size = 12),
      axis.text.y = element_text(size = 12),
      axis.title.x = element_text(size = 14,margin = margin(t = 10)),
      axis.title.y = element_text(size = 14),
      plot.title = element_text(size = 15, face = "bold", hjust = 0.5)
    )
}

# Generate individual plots
density_plot <- generate_plot(wines, "density", "alcohol", cor_density, "Density vs Alcohol")

quality_plot <- generate_plot(wines, "quality", "alcohol", cor_quality, "Quality vs Alcohol")

residual_sugar_plot <- generate_plot(wines, "residual.sugar", "alcohol", cor_residual_sugar, "RS vs Alcohol")

total_sulfDiox_plot <- generate_plot(wines, "total.sulf.diox", "alcohol", cor_total_sulfur_dioxide, "TSD vs Alcohol")

fixed_acidity_plot <- generate_plot(wines, "fixed.acidity", "alcohol", cor_fixedAcidity, "Fixed Acidity vs Alcohol")

pH_plot <- generate_plot(wines, "pH", "alcohol", cor_pH, "pH vs Alcohol")

volatile_Acidity_plot <- generate_plot(wines, "volatile.acidity", "alcohol", cor_volAcidity, "VA vs Alcohol")

citric_Acid_plot <- generate_plot(wines, "citric.acid", "alcohol", cor_citricAcid, "CA vs Alcohol")

chlorides_plot <- generate_plot(wines, "chlorides", "alcohol", cor_chlorides, "Chlorides vs Alcohol")

free_sulf_diox_plot <- generate_plot(wines, "free.sulf.diox", "alcohol", cor_total_sulfur_dioxide, "FSD vs Alcohol")

grid.arrange(density_plot, quality_plot, residual_sugar_plot, total_sulfDiox_plot,
             fixed_acidity_plot, pH_plot,
             nrow = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



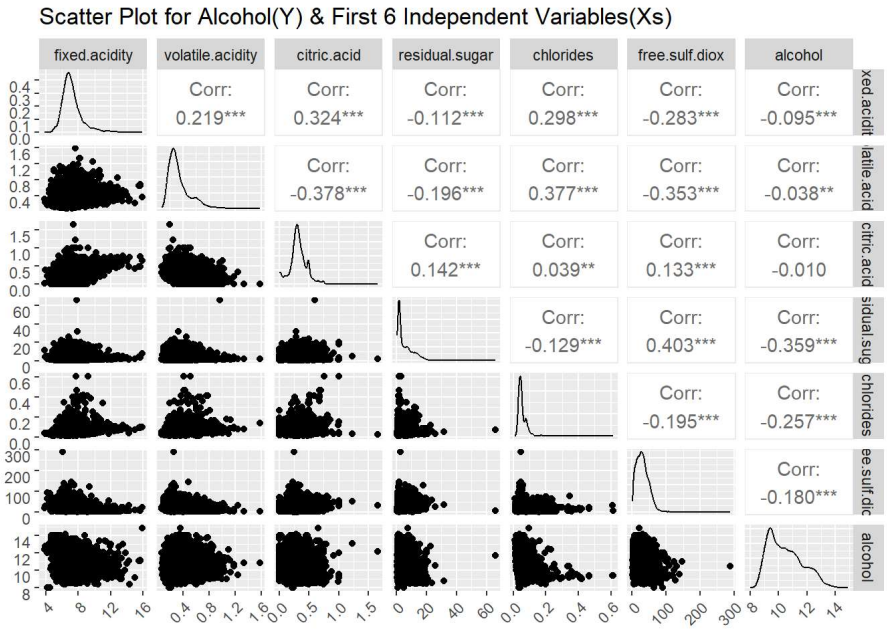

```
grid.arrange( free_sulf_diox_plot, volatile_Acidity_plot, citric_Acid_plot,
              chlorides_plot,
              nrow = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



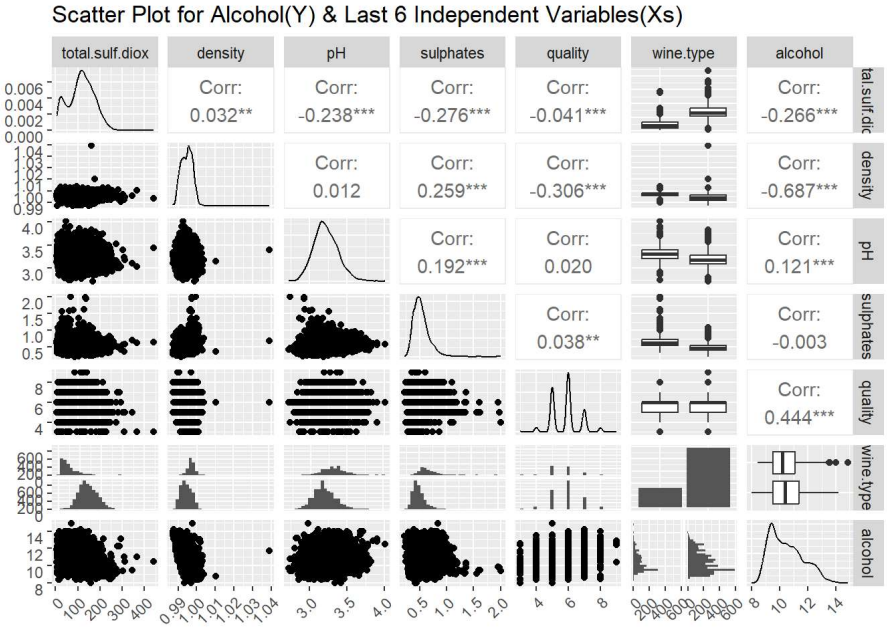
Observing Correlations between predictors and response variable

```
wines_1stHalf <- wines %>% select(fixed.acidity,
                                  volatile.acidity,
                                  citric.acid,
                                  residual.sugar,
                                  chlorides,
                                  free.sulf.diox,
                                  alcohol)
ggpairs(wines_1stHalf, title = "Scatter Plot for Alcohol(Y) & First 6 Independent Variables(Xs)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(axis.text.y = element_text(vjust = 1))
```




```
wines_2ndHalf <- wines %>% select(total.sulf.diox,
                                  density,
                                  pH,
                                  sulphates,
                                  quality,
                                  wine.type,
                                  alcohol)
ggpairs(wines_2ndHalf,
        title = "Scatter Plot for Alcohol(Y) & Last 6 Independent Variables(Xs)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(axis.text.y = element_text(vjust = 1))
```

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



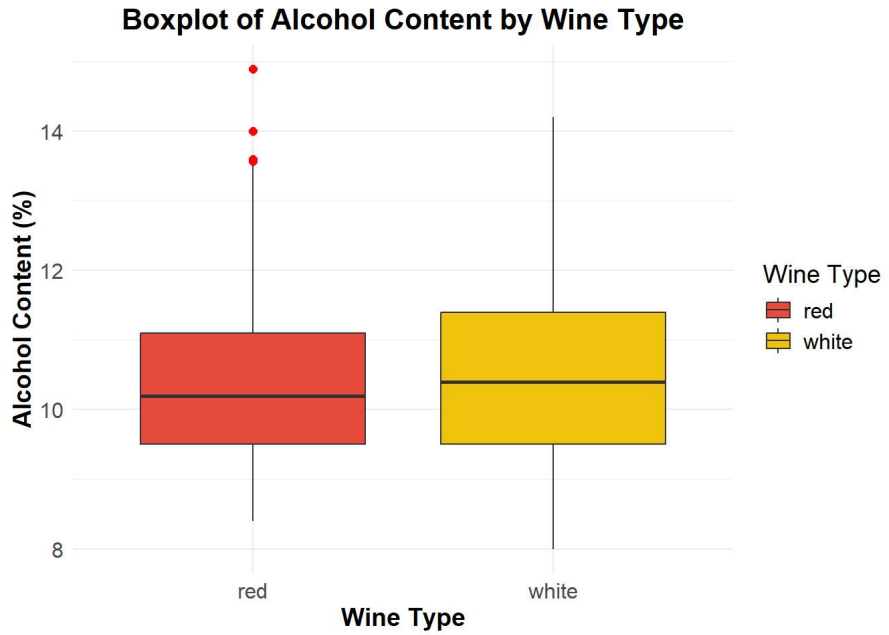
Summary Statistics

```
summary(wines[, c("density", "quality", "residual.sugar", "total.sulf.diox")])
```

```
##     density          quality      residual.sugar   total.sulf.diox
##  Min.   :0.9871   Min.   :3.000   Min.   : 0.600   Min.   :  6.0
##  1st Qu.:0.9923   1st Qu.:5.000   1st Qu.: 1.800   1st Qu.: 77.0
##  Median :0.9949   Median :6.000   Median : 3.000   Median :118.0
##  Mean   :0.9947   Mean   :5.818   Mean   : 5.443   Mean   :115.7
##  3rd Qu.:0.9970   3rd Qu.:6.000   3rd Qu.: 8.100   3rd Qu.:156.0
##  Max.   :1.0390   Max.   :9.000   Max.   :65.800   Max.   :440.0
```

Wine Type Boxplot

```
ggplot(wines, aes(x = wine.type, y = alcohol, fill = wine.type)) +
  geom_boxplot(outlier.color = "red", outlier.shape = 19, outlier.size = 2) +
  scale_fill_manual(values = c("red" = "#E74C3C", "white" = "#F1C40F")) +
  labs(title = "Boxplot of Alcohol Content by Wine Type",
       x = "Wine Type",
       y = "Alcohol Content (%)",
       fill = "Wine Type") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    axis.title.x = element_text(size = 14, face = "bold"),
    axis.title.y = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12)
  )
```

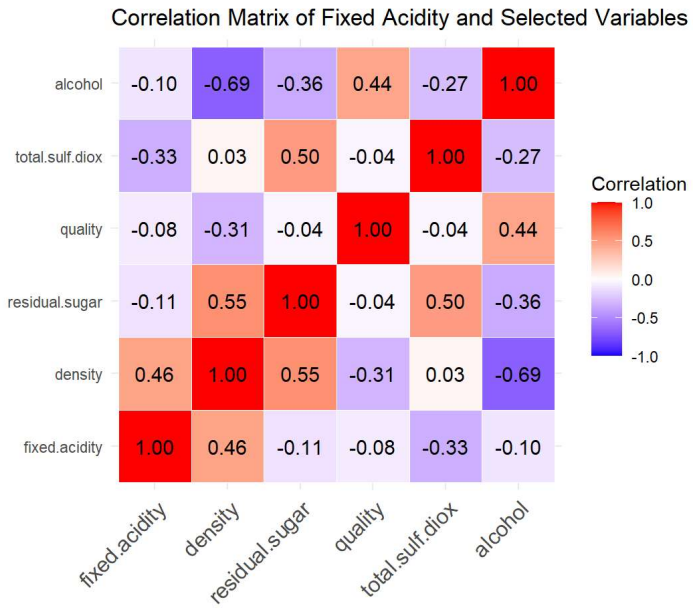


Observing Fixed Acidity's Correlation With The Predictors

```
selected_vars <- wines %>% dplyr::select(fixed.acidity, density, residual.sugar, quality, total.sulf.diox, alcohol)
corr_matrix <- cor(selected_vars, use = "complete.obs")

fixed_acidity_corr <- corr_matrix["fixed.acidity", ]
print(fixed_acidity_corr)
```

```
##   fixed.acidity         density  residual.sugar         quality total.sulf.diox
##      1.00000000      0.45890998     -0.11198128     -0.07674321     -0.32905390
##         alcohol
##     -0.09545152
```

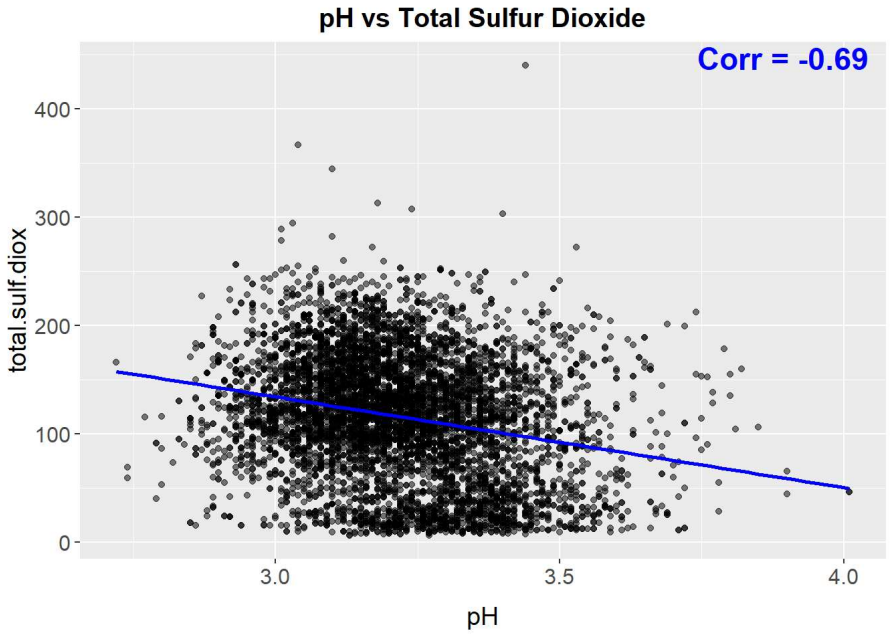
```
corr_melted <- melt(corr_matrix)

ggplot(data = corr_melted, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  geom_text(aes(label = sprintf("%.2f",value)), color = "black", size = 4) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1)) +
  coord_fixed() +
  labs(title = "Correlation Matrix of Fixed Acidity and Selected Variables",
       x = "", y = "")
```



Observing pH

```
# Check correlation with pH and Total Sulfur Dioxide as Total Sulfur Dioxide is reportably inlfuenced by different pH levels
cor_ph_TSD <- cor(wines$pH, wines$total.sulf.diox, use = "complete.obs")
pH_plot <- generate_plot(wines, "pH", "total.sulf.diox", cor_density, "pH vs Total Sulfur Dioxide")
pH_plot
```

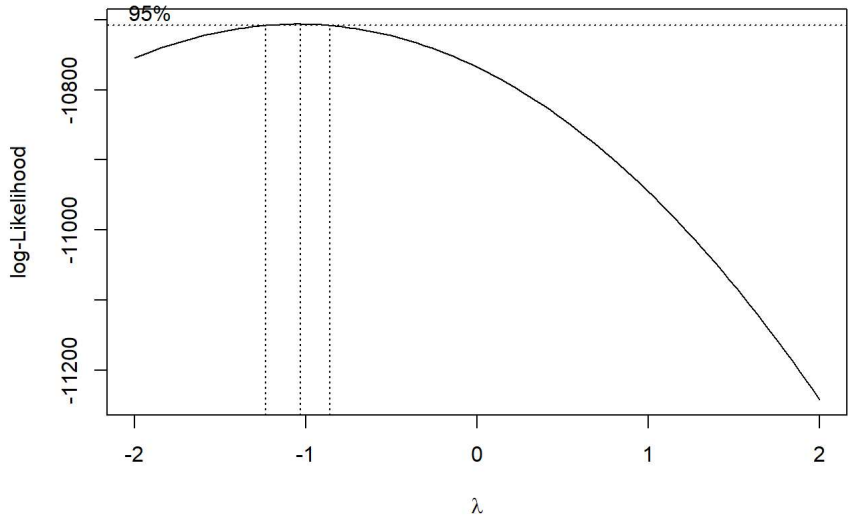
```
## `geom_smooth()` using formula = 'y ~ x'
```



Remove outliers, Transform Predictors, and Transform Response Variable

```
# Transform 2 variables
wines$residual.sugar <- log(wines$residual.sugar)
wines$total.sulf.diox <- log(wines$total.sulf.diox)

#Removing Residual outliers and 2 Density outliers
wines <- wines[-c(354, 653, 3017, 4381, 378, 395, 559, 564, 633, 651, 3244, 3254, 556),]
wines <- wines[-c(440, 478, 552,  554,  603,  881, 1426, 1427, 1466, 1468, 3198, 3242, 3251, 3922, 5206, 5210)]
```

```
model3 <- lm(alcohol ~ density + quality + residual.sugar + total.sulf.diox , data = wines)
boxcox_result <- boxcox(model3)
```




```
optimal_lambda <- boxcox_result$x[which.max(boxcox_result$y)]
optimal_lambda
```

```
## [1] -1.030303
```

```
wines$alcohol_trans <- (wines$alcohol^(optimal_lambda) - 1) / optimal_lambda
```

Standardize all predictors because of instable coefficients from Model1

```
mean_x1 <- mean(wines$density)
mean_x2 <- mean(wines$quality)
mean_x3 <- mean(wines$residual.sugar)
mean_x4 <- mean(wines$total.sulf.diox)
mean_FA <- mean(wines$fixed.acidity)
mean_pH <- mean(wines$pH)

sd1 <- sd(wines$density)
sd2 <- sd(wines$quality)
sd3 <- sd(wines$residual.sugar)
sd4 <- sd(wines$total.sulf.diox)
sdFA <- sd(wines$fixed.acidity)
sdpH <- sd(wines$pH)

wines$density <- (wines$density - mean_x1)/sd1
wines$quality <- (wines$quality - mean_x2)/sd2
wines$residual.sugar <- (wines$residual.sugar  - mean_x3)/sd3
wines$total.sulf.diox <- (wines$total.sulf.diox - mean_x4)/sd4
wines$fixed.acidity <- (wines$fixed.acidity - mean_FA)/sdFA
wines$pH <- (wines$pH - mean_pH)/sdpH
```

Summary statistics after standardizing

```
summary(wines[, c("density", "quality", "residual.sugar", "total.sulf.diox")])
```

```
##     density            quality        residual.sugar    total.sulf.diox
##  Min.   :-2.57221   Min.   :-3.2299   Min.   :-2.1267   Min.   :-3.8860
##  1st Qu.:-0.80240   1st Qu.:-0.9384   1st Qu.:-0.8539   1st Qu.:-0.2916
##  Median : 0.06722   Median : 0.2073   Median :-0.2621   Median : 0.2885
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.77464   3rd Qu.: 0.2073   3rd Qu.: 0.8886   3rd Qu.: 0.6797
##  Max.   : 5.30532   Max.   : 3.6444   Max.   : 2.4657   Max.   : 2.1328
```

Model 1

```
# Null Model
null_model <- lm(alcohol_trans ~ 1, data = wines) #y = B0

#Full model1
model1 <- lm(alcohol_trans ~ density + quality + residual.sugar + total.sulf.diox , data = wines)

# Stepwise
model1 <- step(null_model, scope = list(lower = null_model, upper = model1),
                      direction = "both",
                      test = "F")
```

```
## Start:  AIC=-60000.18
## alcohol_trans ~ 1
##
##                   Df Sum of Sq     RSS    AIC F value    Pr(>F)
## + density          1  0.309706 0.31107 -64478 6453.66 < 2.2e-16 ***
## + quality          1  0.118357 0.50242 -61370 1527.01 < 2.2e-16 ***
## + residual.sugar   1  0.070089 0.55068 -60775  825.01 < 2.2e-16 ***
## + total.sulf.diox  1  0.023505 0.59727 -60248  255.09 < 2.2e-16 ***
## <none>                         0.62077 -60000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-64478.35
## alcohol_trans ~ density
##
##                   Df Sum of Sq     RSS    AIC  F value    Pr(>F)
## + total.sulf.diox  1  0.046058 0.26501 -65515 1126.389 < 2.2e-16 ***
## + quality          1  0.032338 0.27873 -65188  751.936 < 2.2e-16 ***
## + residual.sugar   1  0.000585 0.31048 -64489   12.219 0.0004763 ***
## <none>                         0.31107 -64478
## - density          1  0.309706 0.62077 -60000 6453.659 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-65515.38
## alcohol_trans ~ density + total.sulf.diox
##
##                   Df Sum of Sq     RSS    AIC F value    Pr(>F)
## + quality          1   0.03047 0.23454 -66305  841.76 < 2.2e-16 ***
## + residual.sugar   1   0.02875 0.23626 -66258  788.57 < 2.2e-16 ***
## <none>                         0.26501 -65515
## - total.sulf.diox  1   0.04606 0.31107 -64478 1126.39 < 2.2e-16 ***
## - density          1   0.33226 0.59727 -60248 8125.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-66305.27
## alcohol_trans ~ density + total.sulf.diox + quality
##
##                   Df Sum of Sq     RSS    AIC F value    Pr(>F)
## + residual.sugar   1  0.017603 0.21694 -66809  525.72 < 2.2e-16 ***
## <none>                         0.23454 -66305
## - quality          1  0.030467 0.26501 -65515  841.76 < 2.2e-16 ***
## - total.sulf.diox  1  0.044187 0.27873 -65188 1220.81 < 2.2e-16 ***
## - density          1  0.243292 0.47783 -61693 6721.76 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-66809.14
## alcohol_trans ~ density + total.sulf.diox + quality + residual.sugar
##
##                   Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>                         0.21694 -66809
## - residual.sugar   1  0.017603 0.23454 -66305  525.72 < 2.2e-16 ***
## - quality          1  0.019319 0.23626 -66258  576.98 < 2.2e-16 ***
## - total.sulf.diox  1  0.061719 0.27866 -65188 1843.29 < 2.2e-16 ***
## - density          1  0.214690 0.43163 -62350 6411.84 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = alcohol_trans ~ density + total.sulf.diox + quality +
##     residual.sugar, data = wines)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.022146 -0.003798 -0.000040  0.003654  0.032031
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      8.833e-01  7.186e-05 12291.93   <2e-16 ***
## density         -7.946e-03  9.924e-05   -80.07   <2e-16 ***
## total.sulf.diox -3.753e-03  8.742e-05   -42.93   <2e-16 ***
## quality          1.868e-03  7.776e-05    24.02   <2e-16 ***
## residual.sugar   2.360e-03  1.029e-04    22.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005786 on 6479 degrees of freedom
## Multiple R-squared:  0.6505, Adjusted R-squared:  0.6503
## F-statistic:  3015 on 4 and 6479 DF,  p-value: < 2.2e-16
```

Multicollinearity Check (model1)

```
cor_dens_RS <- cor(model1$model$density, model1$model$residual.sugar, use = "complete.obs")
cor_dens_RS
```

```
## [1] 0.5130359
```

```
cor_TSD_RS <- cor(model1$model$total.sulf.diox, model1$model$residual.sugar, use = "complete.obs")
cor_TSD_RS
```

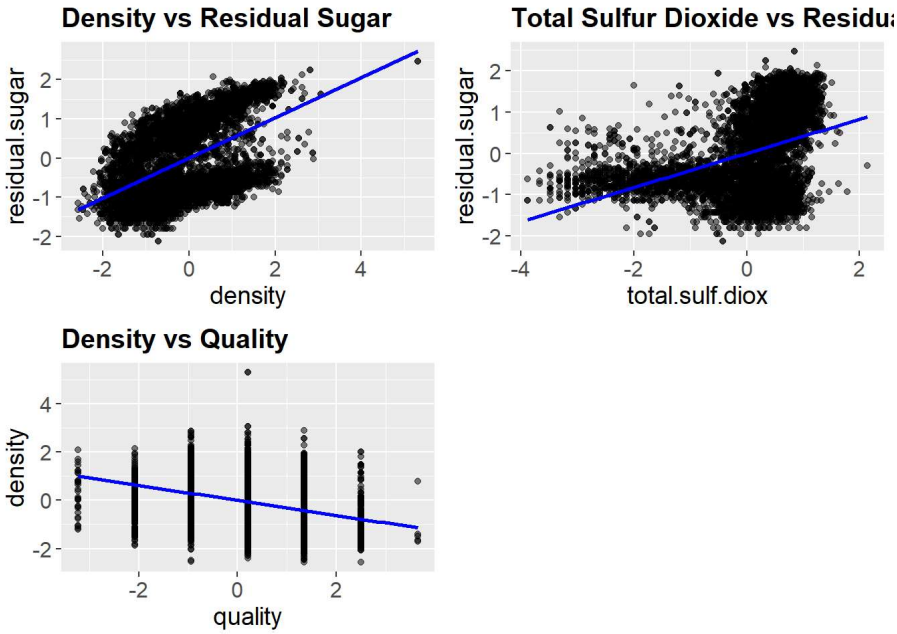
```
## [1] 0.412371
```

```
cor_dens_Quality <- cor(model1$model$density, model1$model$quality, use = "complete.obs")
cor_dens_Quality
```

```
## [1] -0.3110886
```

```
generate_plot <- function(data, xvar, yvar, cor_value, title) {
  ggplot(data, aes_string(x = xvar, y = yvar)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    ggtitle(title) +
    theme(
      axis.text.x = element_text(size = 12),
      axis.text.y = element_text(size = 12),
      axis.title.x = element_text(size = 14),
      axis.title.y = element_text(size = 14),
      plot.title = element_text(size = 15, face = "bold")
    )
    #theme_minimal()
}

# Generate individual plots
dens_RS_plot <- generate_plot(wines, "density", "residual.sugar", cor_dens_RS, "Density vs Residual Sugar")

TSD_RS_plot <- generate_plot(wines, "total.sulf.diox", "residual.sugar", cor_TSD_RS, "Total Sulfur Dioxide vs Residual Suga
r")

dens_Quality_plot <- generate_plot(wines,  "quality", "density", cor_dens_Quality, "Density vs Quality")

grid.arrange(dens_RS_plot, TSD_RS_plot, dens_Quality_plot, nrow = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```




```
vif(model1)
```

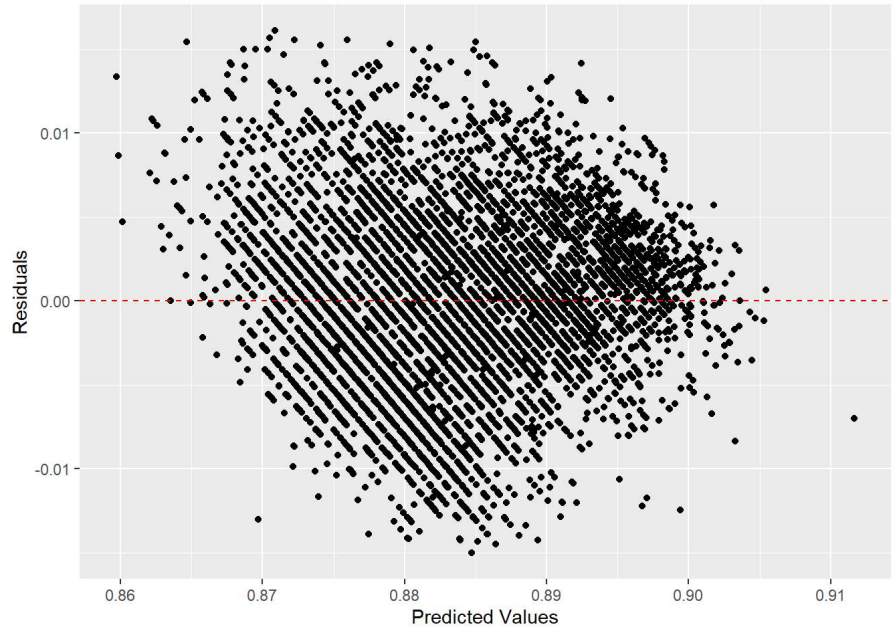
```
##         density total.sulf.diox         quality  residual.sugar
##        1.906680        1.479732        1.170888        2.050479
```

Residual Analysis (model1)

```
## REMOVING ANY REMAINING OUTLIERS SEE IN PREVIOUS (FITTED VS RESIDUALS) SCATTER PLOT
residuals <- residuals(model1)

wines$residuals <- residuals

filtered_wines <- subset(wines, abs(residuals) <= 0.015)

model1 <- lm(alcohol_trans ~ density + quality + residual.sugar + total.sulf.diox, data = filtered_wines)
summary(model1)
```

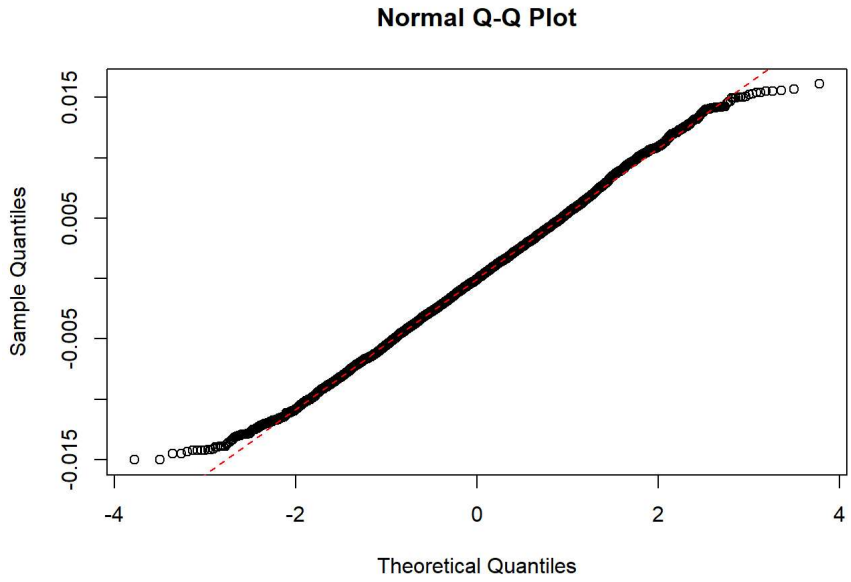
```
##
## Call:
## lm(formula = alcohol_trans ~ density + quality + residual.sugar +
##     total.sulf.diox, data = filtered_wines)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0150154 -0.0036830 -0.0000667  0.0036058  0.0160989
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      8.832e-01  6.750e-05 13083.82   <2e-16 ***
## density         -8.381e-03  9.470e-05   -88.49   <2e-16 ***
## quality          1.708e-03  7.329e-05    23.31   <2e-16 ***
## residual.sugar   2.500e-03  9.702e-05    25.77   <2e-16 ***
## total.sulf.diox -3.810e-03  8.251e-05   -46.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0054 on 6397 degrees of freedom
## Multiple R-squared:  0.6919, Adjusted R-squared:  0.6917
## F-statistic:  3591 on 4 and 6397 DF,  p-value: < 2.2e-16
```

```
#Update dataset
wines_updated <- model1$model
ggplot(data = model1, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```



Normality of Residuals (Q-Q Plot)

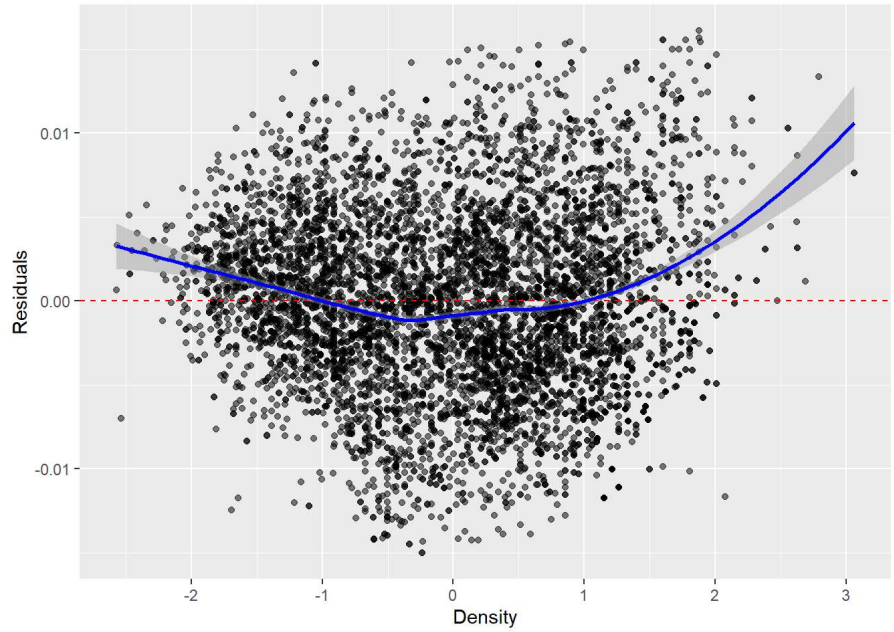
```
#install.packages("nortest")  # Install if not already installed
qqnorm(residuals(model1), main = "Normal Q-Q Plot")
qqline(residuals(model1), col = "red", lty = 2)
```



Density vs Residual

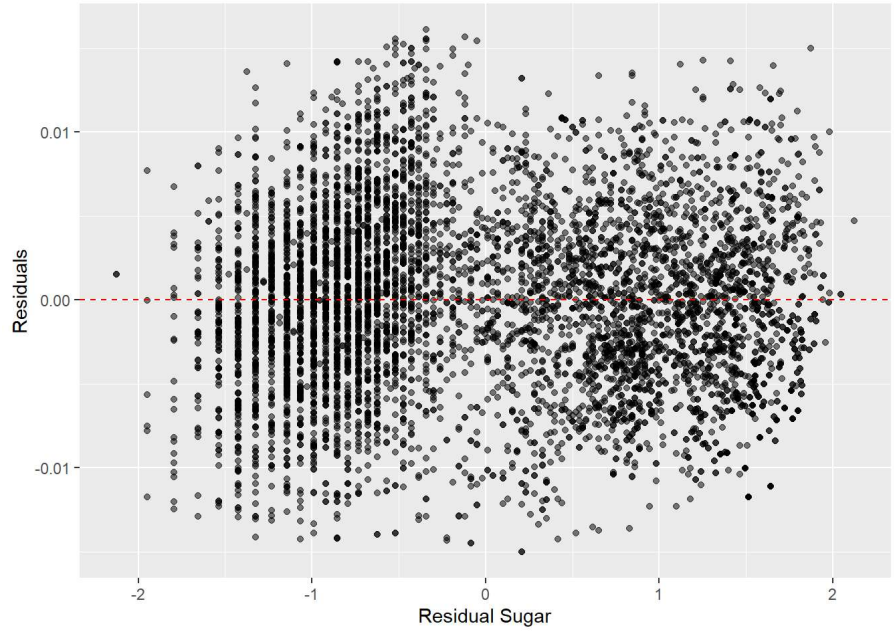
```
ggplot(data = model1, aes(x = density, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", color = "blue", linetype = "solid", se = TRUE) +
  labs(x = "Density", y = "Residuals")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



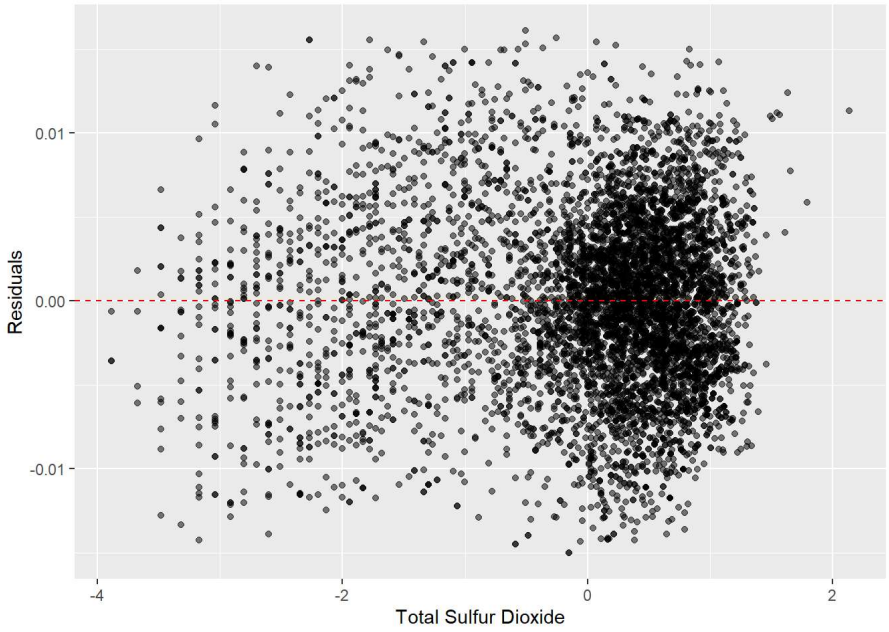
Residual Sugar vs Residual

```
ggplot(data = model1, aes(x = residual.sugar, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Residual Sugar", y = "Residuals")
```



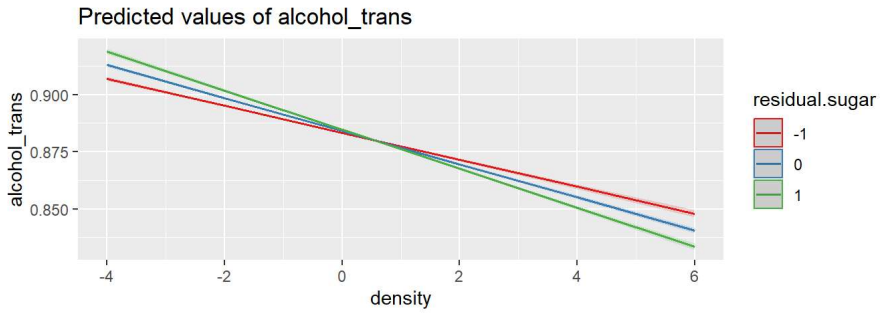
Total Sulfur Dioxide vs Residual

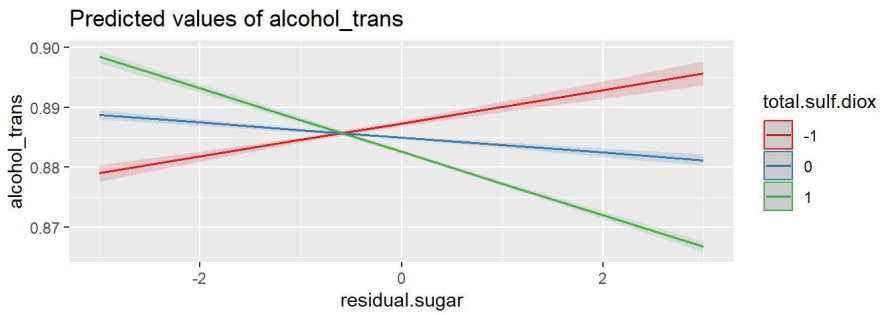
```
ggplot(data = model1, aes(x = total.sulf.diox, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Total Sulfur Dioxide", y = "Residuals")
```



Checking for Interactions

```
fit1 <- lm(alcohol_trans ~ density*residual.sugar, wines)
interaction1 <- plot_model(fit1,
                           type = "int",
                           mdrt.values = "meansd")

fit2 <- lm(alcohol_trans ~ residual.sugar*total.sulf.diox, wines)
interaction2 <- plot_model(fit2,
                           type = "int",
                           mdrt.values = "meansd")
grid.arrange(interaction1, interaction2)
```





Checking Fixed.Acidity and pH's improvement to Model1

```
model1_FA <- lm(alcohol_trans ~ density + quality + residual.sugar + total.sulf.diox + fixed.acidity, data = wines)
summary(model1_FA)
```

```
##
## Call:
## lm(formula = alcohol_trans ~ density + quality + residual.sugar +
##     total.sulf.diox + fixed.acidity, data = wines)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0221746 -0.0034249  0.0000125  0.0032699  0.0289057
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      8.833e-01  6.649e-05 13285.83   <2e-16 ***
## density         -9.871e-03  1.087e-04   -90.77   <2e-16 ***
## quality          1.489e-03  7.286e-05    20.43   <2e-16 ***
## residual.sugar   3.410e-03  1.004e-04    33.97   <2e-16 ***
## total.sulf.diox -3.261e-03  8.224e-05   -39.66   <2e-16 ***
## fixed.acidity    2.851e-03  8.630e-05    33.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005354 on 6478 degrees of freedom
## Multiple R-squared:  0.7009, Adjusted R-squared:  0.7007
## F-statistic:  3036 on 5 and 6478 DF,  p-value: < 2.2e-16
```

```
model1_pH <- lm(alcohol_trans ~ density + quality + residual.sugar + total.sulf.diox + pH, data = wines)
summary(model1_pH)
```

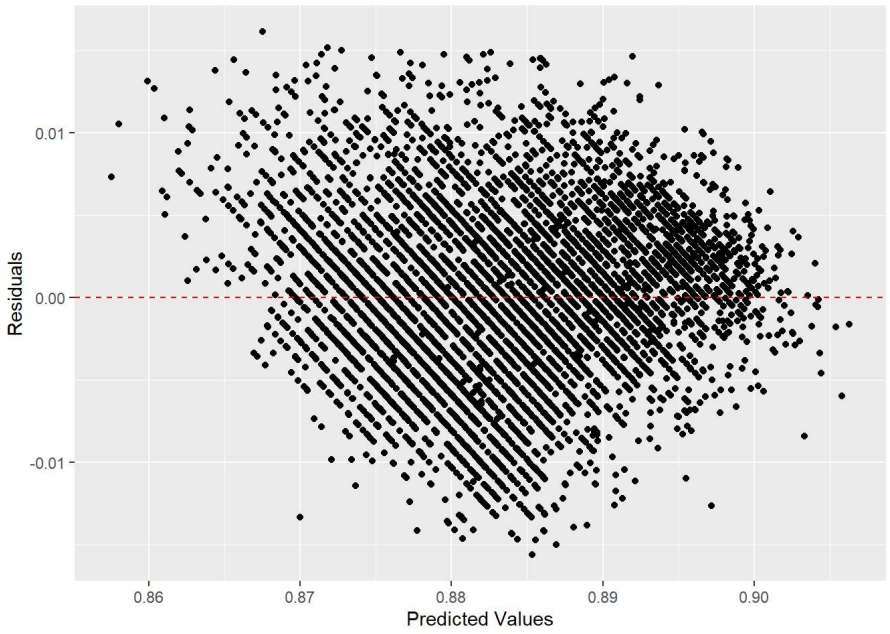
```
##
## Call:
## lm(formula = alcohol_trans ~ density + quality + residual.sugar +
##     total.sulf.diox + pH, data = wines)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.022762 -0.003806 -0.000097  0.003513  0.032025
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      8.833e-01  7.057e-05 12517.29   <2e-16 ***
## density         -8.133e-03  9.819e-05   -82.83   <2e-16 ***
## quality          1.792e-03  7.652e-05    23.42   <2e-16 ***
## residual.sugar   2.682e-03  1.032e-04    26.00   <2e-16 ***
## total.sulf.diox -3.612e-03  8.633e-05   -41.84   <2e-16 ***
## pH               1.156e-03  7.451e-05    15.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005682 on 6478 degrees of freedom
## Multiple R-squared:  0.6631, Adjusted R-squared:  0.6628
## F-statistic:  2550 on 5 and 6478 DF,  p-value: < 2.2e-16
```

Checking Interactions between predictor variables

```
## Building model with interactions for testing
model_interactions = lm(alcohol_trans ~ density + quality + residual.sugar + total.sulf.diox + density*residual.sugar + tota
l.sulf.diox*residual.sugar, data = filtered_wines)

summary(model_interactions)
```

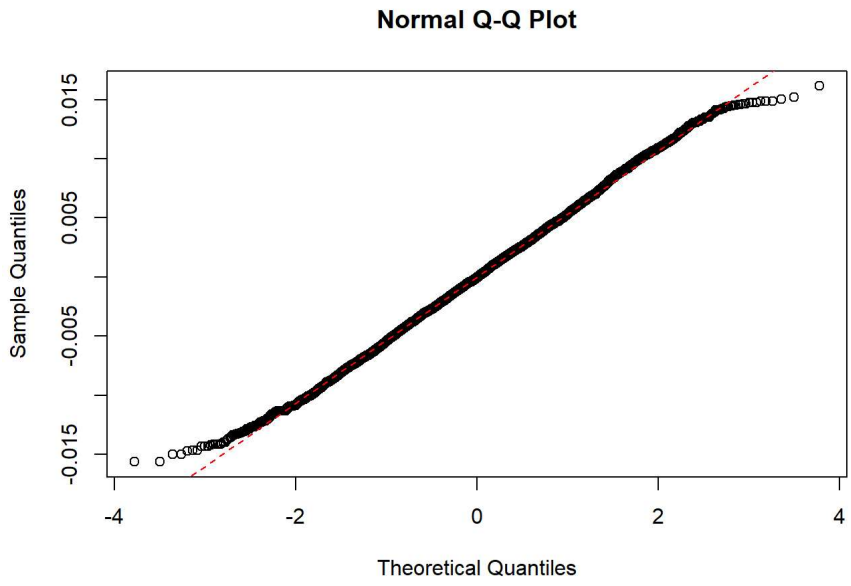
```
##
## Call:
## lm(formula = alcohol_trans ~ density + quality + residual.sugar +
##     total.sulf.diox + density * residual.sugar + total.sulf.diox *
##     residual.sugar, data = filtered_wines)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0156228 -0.0036194 -0.0000505  0.0035683  0.0161363
##
## Coefficients:
##                                  Estimate Std. Error   t value Pr(>|t|)
## (Intercept)                     8.837e-01  8.708e-05 10148.336  < 2e-16 ***
## density                        -8.191e-03  9.661e-05   -84.779  < 2e-16 ***
## quality                         1.708e-03  7.285e-05    23.444  < 2e-16 ***
## residual.sugar                  2.777e-03  1.031e-04    26.931  < 2e-16 ***
## total.sulf.diox                -4.017e-03  9.955e-05   -40.355  < 2e-16 ***
## density:residual.sugar         -2.896e-04  7.661e-05    -3.781 0.000158 ***
## residual.sugar:total.sulf.diox -8.929e-04  1.171e-04    -7.623 2.84e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005364 on 6395 degrees of freedom
## Multiple R-squared:  0.6961, Adjusted R-squared:  0.6958
## F-statistic:  2441 on 6 and 6395 DF,  p-value: < 2.2e-16
```

```
ggplot(data = model_interactions, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted Values", y = "Residuals")
```



Normality of Residuals (Q-Q Plot)

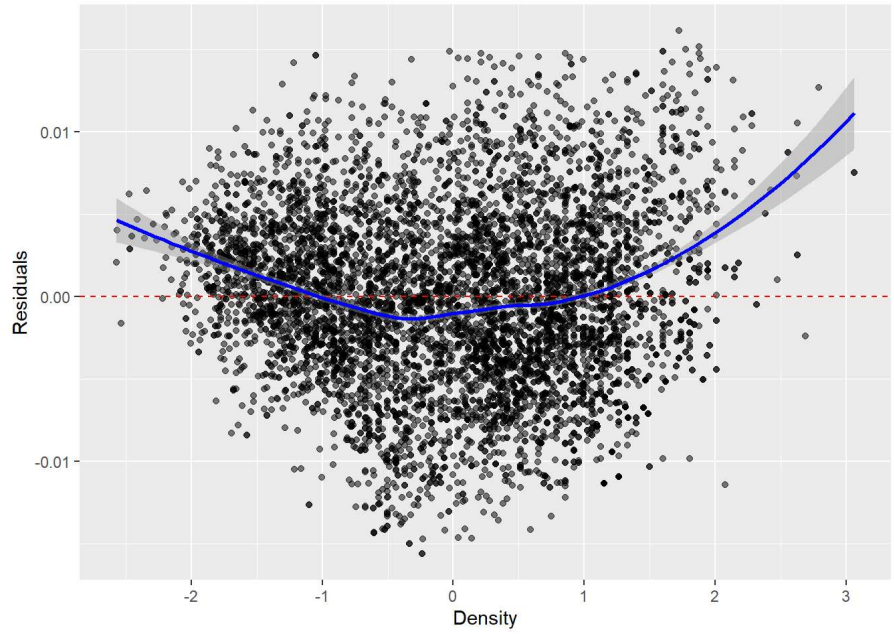
```
#install.packages("nortest")  # Install if not already installed
qqnorm(residuals(model_interactions), main = "Normal Q-Q Plot")
qqline(residuals(model_interactions), col = "red", lty = 2)
```



Density vs Residual

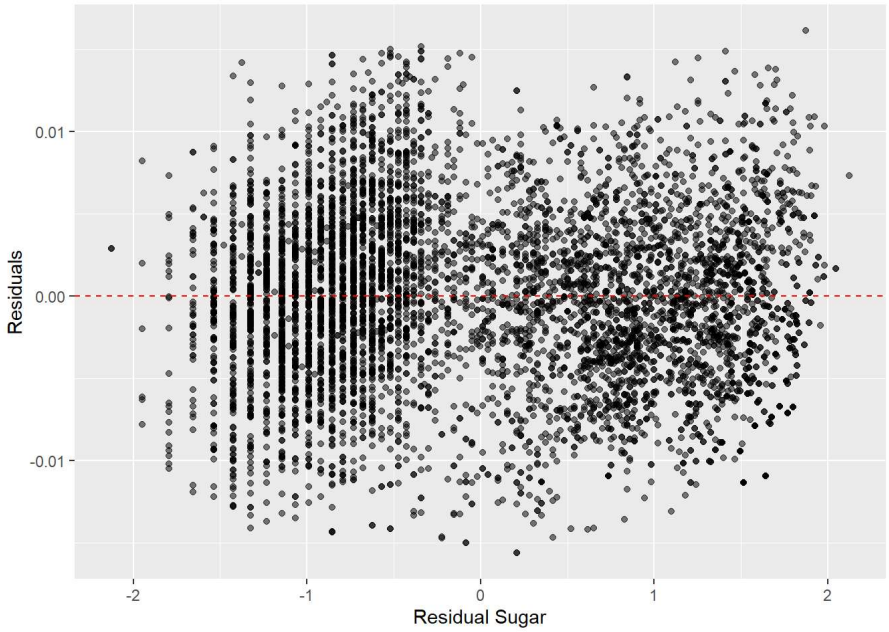
```
ggplot(data = model_interactions, aes(x = density, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", color = "blue", linetype = "solid", se = TRUE) +
  labs(x = "Density", y = "Residuals")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



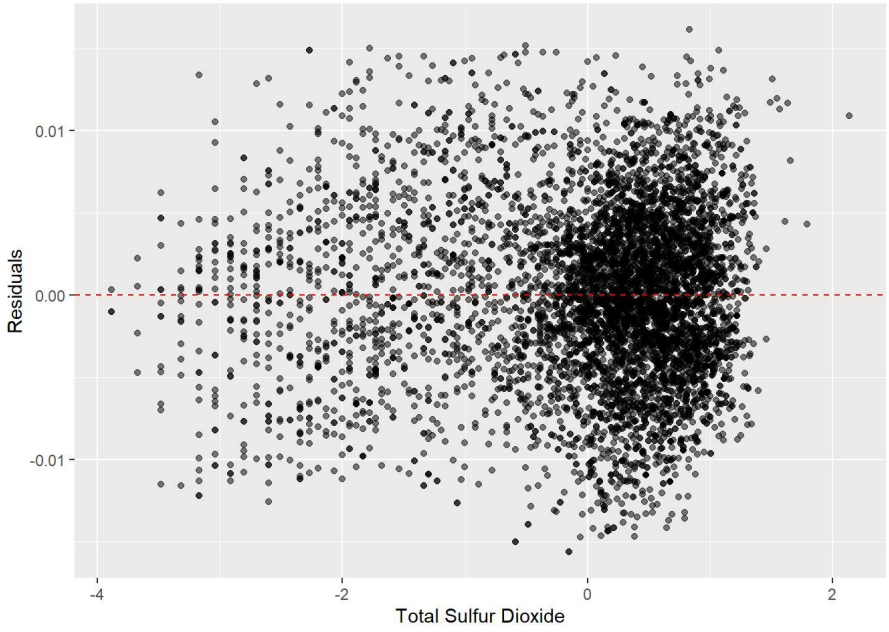
Residual Sugar vs Residual

```
ggplot(data = model_interactions, aes(x = residual.sugar, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Residual Sugar", y = "Residuals")
```



Total Sulfur Dioxide vs Residual

```
ggplot(data = model_interactions, aes(x = total.sulf.diox, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Total Sulfur Dioxide", y = "Residuals")
```




```
vif(model_interactions, type='predictor')
```

```
## GVIFs computed for predictors
```

```
##                    GVIF Df GVIF^(1/(2*Df))           Interacts With
## density         3.068973  3        1.205495           residual.sugar
## quality         1.183928  1        1.088085                       --
## residual.sugar  1.183928  5        1.017027 density, total.sulf.diox
## total.sulf.diox 2.291628  3        1.148217           residual.sugar
##                                      Other Predictors
## density                           quality, total.sulf.diox
## quality         density, residual.sugar, total.sulf.diox
## residual.sugar                                   quality
## total.sulf.diox                          density, quality
```

```
anova(model1, model_interactions)
```

```
## Analysis of Variance Table
##
## Model 1: alcohol_trans ~ density + quality + residual.sugar + total.sulf.diox
## Model 2: alcohol_trans ~ density + quality + residual.sugar + total.sulf.diox +
##     density * residual.sugar + total.sulf.diox * residual.sugar
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   6397 0.18656
## 2   6395 0.18400  2 0.0025573 44.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions between predictors allow the model to capture more complex relationships. Based on theoretical insights and prior analysis, we included the following interactions: - **Density × Residual Sugar**: Both are strongly tied to fermentation dynamics.
- **Quality × pH**: Quality may depend on the acidity balance.
- **Fixed Acidity × Total Sulfur Dioxide**: Both influence yeast activity during fermentation.

Code for Interaction Terms:

```
wines <- wines %>%
  mutate(density_residual_sugar = density * residual.sugar,
         quality_pH = quality * pH,
         fixed_acidity_sulfur = fixed.acidity * total.sulf.diox)
```

Quadratic Terms

To capture non-linear relationships, we included quadratic terms for: - **Density**: Alcohol levels decrease non-linearly as density increases.
- **Residual Sugar**: Small changes in sugar levels can have non-linear effects on fermentation.

Code for Quadratic Terms:

```
wines <- wines %>%
  mutate(density_squared = density^2,
         residual_sugar_squared = residual.sugar^2)
```

Final Model

We fit an updated regression model with both interaction and quadratic terms to improve predictive performance. The final model formula is as follows:

$$\text{alcohol} \sim \text{fixed.acidity} + \text{residual.sugar} + \text{density} + \text{total.sulf.diox} + \text{pH} + \text{quality} + \text{density\_residual\_sugar} + \text{quality\_pH} + \text{fixed\_aci}$$

Code:

```
wines.mlr_interactions <- lm(alcohol ~ fixed.acidity + residual.sugar + density +
                             total.sulf.diox + pH + quality +
                             density_residual_sugar + quality_pH + fixed_acidity_sulfur +
                             density_squared + residual_sugar_squared,
                             data = wines)
summary(wines.mlr_interactions)
```

```
##
## Call:
## lm(formula = alcohol ~ fixed.acidity + residual.sugar + density +
##     total.sulf.diox + pH + quality + density_residual_sugar +
##     quality_pH + fixed_acidity_sulfur + density_squared + residual_sugar_squared,
##     data = wines)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08955 -0.32391 -0.03115  0.28268  2.69686
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)            10.232486   0.013628  750.821   <2e-16 ***
## fixed.acidity           0.624426   0.011776   53.027   <2e-16 ***
## residual.sugar          0.674124   0.010923   61.719   <2e-16 ***
## density                -1.486100   0.012814 -115.978   <2e-16 ***
## total.sulf.diox        -0.336189   0.008861  -37.942   <2e-16 ***
## pH                      0.422083   0.008624   48.942   <2e-16 ***
## quality                 0.094962   0.007303   13.003   <2e-16 ***
## density_residual_sugar -0.126755   0.014113   -8.982   <2e-16 ***
## quality_pH              0.003891   0.006374    0.610    0.542
## fixed_acidity_sulfur    0.071561   0.005995   11.937   <2e-16 ***
## density_squared         0.219865   0.007725   28.461   <2e-16 ***
## residual_sugar_squared  0.131677   0.013491    9.760   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5202 on 6472 degrees of freedom
## Multiple R-squared:  0.8097, Adjusted R-squared:  0.8094
## F-statistic:  2504 on 11 and 6472 DF,  p-value: < 2.2e-16
```

Model Assumptions

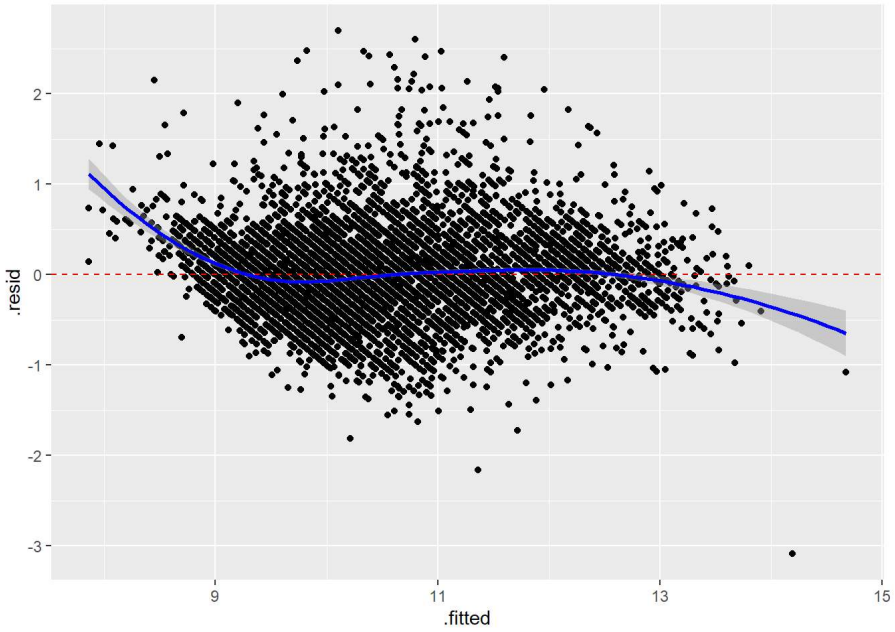
We validated that the updated model meets key regression assumptions:

1. **Linearity**: The residuals vs. fitted values plot showed no patterns or curvature, confirming linear relationships between predictors and alcohol.
2. **Normality of Residuals**: The Q-Q plot indicated that residuals were approximately normally distributed. Minor deviations in the tails were noted but not significant.
3. **Homoscedasticity**: The Scale-Location plot showed constant variance of residuals, with no major patterns of increasing or decreasing variance.

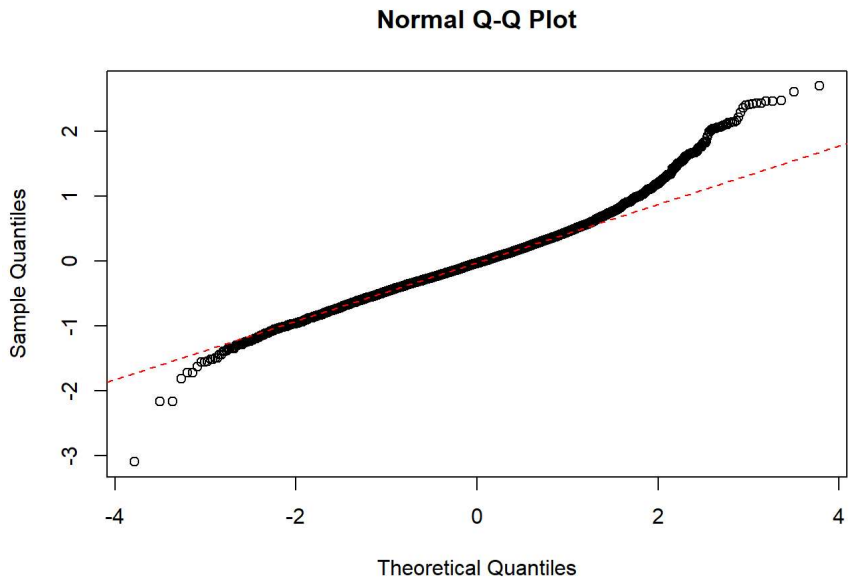
Code for Assumptions Checks:

```
# Linearity
ggplot(wines.mlr_interactions, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", color = "blue", linetype = "solid", se = TRUE)
```

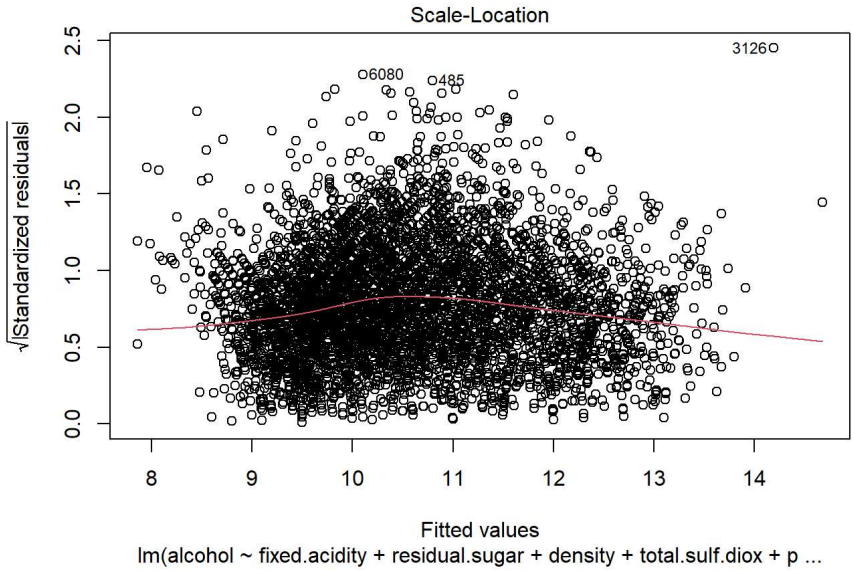
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Normality
qqnorm(residuals(wines.mlr_interactions))
qqline(residuals(wines.mlr_interactions), col = "red", lty = 2)
```




```
# Homoscedasticity
plot(wines.mlr_interactions, which = 3)
```



Model Comparison

We compared the updated model with the initial model using ANOVA and adjusted R-squared.

Findings: 1. **Initial Model**: Adjusted R-squared = **0.7514**
- Includes only main effects: fixed acidity, residual sugar, density, total sulfur dioxide, pH, quality.

2. **Updated Model**: Adjusted R-squared = **0.797**
    - Adds interaction and quadratic terms: density_residual_sugar, quality_pH, fixed_acidity_sulfur, density_squared, residual_sugar_squared.

Code for Model Comparison:

```
wines.mlr <- lm(alcohol ~ fixed.acidity + residual.sugar + density + total.sulf.diox + pH + quality, data = wines)

summary(wines.mlr)
```

```
##
## Call:
## lm(formula = alcohol ~ fixed.acidity + residual.sugar + density +
##     total.sulf.diox + pH + quality, data = wines)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7066 -0.3719 -0.0291  0.3286  4.4809
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.490689   0.007063 1485.30   <2e-16 ***
## fixed.acidity    0.601632   0.010906   55.17   <2e-16 ***
## residual.sugar   0.658223   0.011634   56.58   <2e-16 ***
## density         -1.454570   0.012763 -113.97   <2e-16 ***
## total.sulf.diox -0.305184   0.008986  -33.96   <2e-16 ***
## pH               0.396277   0.008872   44.67   <2e-16 ***
## quality          0.128455   0.007860   16.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5687 on 6477 degrees of freedom
## Multiple R-squared:  0.7724, Adjusted R-squared:  0.7722
## F-statistic:  3663 on 6 and 6477 DF,  p-value: < 2.2e-16
```

```
anova(wines.mlr, wines.mlr_interactions)
```

```
## Analysis of Variance Table
##
## Model 1: alcohol ~ fixed.acidity + residual.sugar + density + total.sulf.diox +
##     pH + quality
## Model 2: alcohol ~ fixed.acidity + residual.sugar + density + total.sulf.diox +
##     pH + quality + density_residual_sugar + quality_pH + fixed_acidity_sulfur +
##     density_squared + residual_sugar_squared
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   6477 2095.1
## 2   6472 1751.3  5    343.75 254.07 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The updated model significantly reduces residual error and explains more variance in alcohol content.