White Wine Analysis

First is regression models for white wine analysis.

1. Linear Regression

```r
# Linear Regression
lm_model <- lm(quality ~ alcohol, data = wine_data)

# Summary of the model
summary(lm_model)
```

```
Call:
lm(formula = quality ~ alcohol, data = wine_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5317 -0.5286  0.0012  0.4996  3.1579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.582009   0.098008   26.34   <2e-16 ***
alcohol     0.313469   0.009258   33.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7973 on 4896 degrees of freedom
Multiple R-squared:  0.1897,    Adjusted R-squared:  0.1896
F-statistic:  1146 on 1 and 4896 DF,  p-value: < 2.2e-16
```

`We took `quality` as a function of `alcohol` and did the analysis. As per the Random forest analysis, we got to know that alcohol has the greatest effect on the quality of alcohol

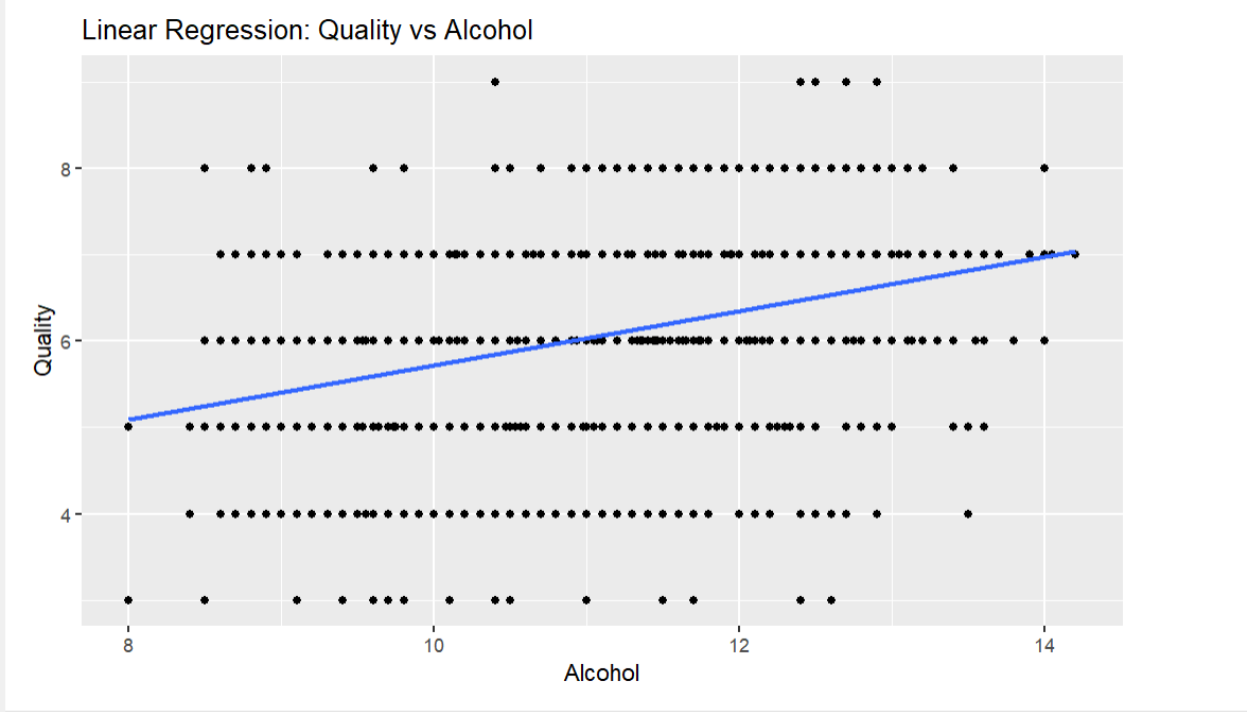
The summary of residuals shows that the median is close to zero and suggests that the model is not systematically over or under-predicting.

- The **Intercept** is 2.582009, meaning when the alcohol level is zero, the predicted quality is approximately 2.58. The standard error of the intercept is 0.098008, and its t-value is 26.34, which is highly significant (p-value < 2e-16).

- The **slope** for alcohol is 0.31469, meaning for each unit increase in alcohol, the quality score is expected to increase by approximately 0.31. The standard error of this estimate is very small (0.009258), indicating a precise estimate.

- In **Multiple R-squared:** The value of 0.1897 indicates that approximately 18.97% of the variability in the quality score can be explained by the alcohol level. It's a measure of the model's goodness of fit.

In summary, the output suggests that there is a statistically significant positive relationship between alcohol content and wine quality.



This is a scatter plot with a fitted regression line, depicting the relationship between alcohol content (x-axis) and quality (y-axis) of wine from a linear regression analysis.

Interpretation of the plot:

- **Data Points (Black Dots):** Each dot represents an individual observation in the dataset, showing the wine's quality score associated with its alcohol content.

- **Regression Line (Blue Line):** The blue line is the linear regression line, which represents the best linear fit to the data points. It shows the average relationship between alcohol content and quality score.

- **Trend:** The regression line has a positive slope, indicating that there is a general positive correlation between alcohol content and quality score. In other words, higher alcohol content is associated with a higher quality rating of wine.

- **Outliers:** There do not appear to be extreme outliers, but there is some spread in the quality scores at higher alcohol levels. This might indicate variability in quality that is not explained by alcohol alone.

Overall, the visualization suggests that there is a trend where wines with higher alcohol content tend to have higher quality ratings, but the variation in the data indicates that alcohol content is not the only factor that determines the quality of wine.

2. Multiple Regression

```
   ```{r}
# Multiple Regression
multiple_lm <- lm(quality ~ alcohol + pH , data = wine_data)

# Summary of the model
summary(multiple_lm)

```
```

```
Call:
lm(formula = quality ~ alcohol + pH, data = wine_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5568 -0.5186 -0.0102  0.4965  3.1542

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.742207   0.250108   6.966  3.7e-12 ***
alcohol     0.309342   0.009316  33.207  < 2e-16 ***
pH          0.277016   0.075920   3.649 0.000266 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7963 on 4895 degrees of freedom
Multiple R-squared:  0.1919,    Adjusted R-squared:  0.1916
F-statistic: 581.3 on 2 and 4895 DF,  p-value: < 2.2e-16
```

For multiple regression quality of wine is predicted based on both alcohol and pH levels.

- **Residuals:** The summary of residuals indicates the spread of the errors (differences between observed and predicted values) of the model. They range from a minimum of -3.5568 to a maximum of 3.1542. The median is very close to zero (-0.0102), which suggests that there is no systematic bias in the predictions.

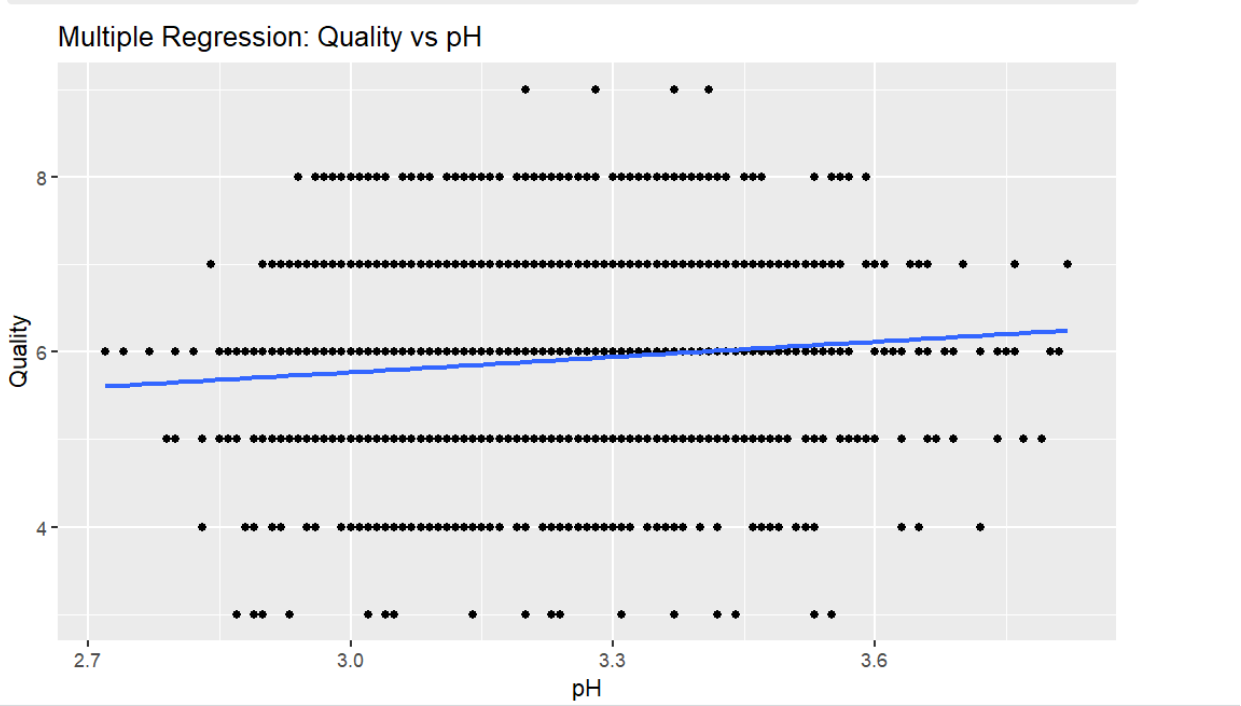
- **Coefficients:**

  - The **Intercept** is 1.742207, implying that when both alcohol and pH are at zero, the expected wine quality score is 1.742.

  - The coefficient for **alcohol** is 0.309342, meaning that with each unit increase in alcohol, holding pH constant, the quality score is expected to increase by about 0.309.

  - The coefficient for **pH** is 0.277016, meaning that with each unit increase in pH, holding alcohol constant, the quality score is expected to increase by about 0.277

In summary, the model suggests that both alcohol and pH levels are significant predictors of wine quality. The model is statistically significant and likely a better fit than a simple linear regression with only one predictor.



This scatter plot with a fitted regression line depicts the relationship between pH level (x-axis) and quality (y-axis) of wine as part of a multiple regression analysis.

- **Regression Line (Blue Line):** The blue line is the fitted regression line, which represents the average effect of pH on wine quality when holding other variables in the multiple regression model constant. This line shows a slight positive slope, suggesting a weak positive relationship between pH level and wine quality.

- **Trend:** The slight upward trend of the regression line implies that there is a small positive association between the pH of the wine and its quality rating. However, the relationship is not strong.

3. Logistic Regression

```r
# Create a binary outcome variable
wine_data$high_quality <- ifelse(wine_data$quality > 5, 1, 0)
# Logistic Regression
logistic_model <- glm(high_quality ~ alcohol + pH ,
                      family = binomial(), data = wine_data)
# Summary of the model
summary(logistic_model)
```

```
Call:
glm(formula = high_quality ~ alcohol + pH, family = binomial(),
    data = wine_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.21366    0.74110 -12.432   <2e-16 ***
alcohol      0.81620    0.03331  24.505   <2e-16 ***
pH           0.46817    0.22121   2.116   0.0343 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6245.4  on 4897  degrees of freedom
Residual deviance: 5429.7  on 4895  degrees of freedom
AIC: 5435.7

Number of Fisher Scoring iterations: 4
```

Here, `high_quality ~ alcohol + pH` indicates that the binary outcome `high_quality` is predicted based on alcohol content and pH levels.

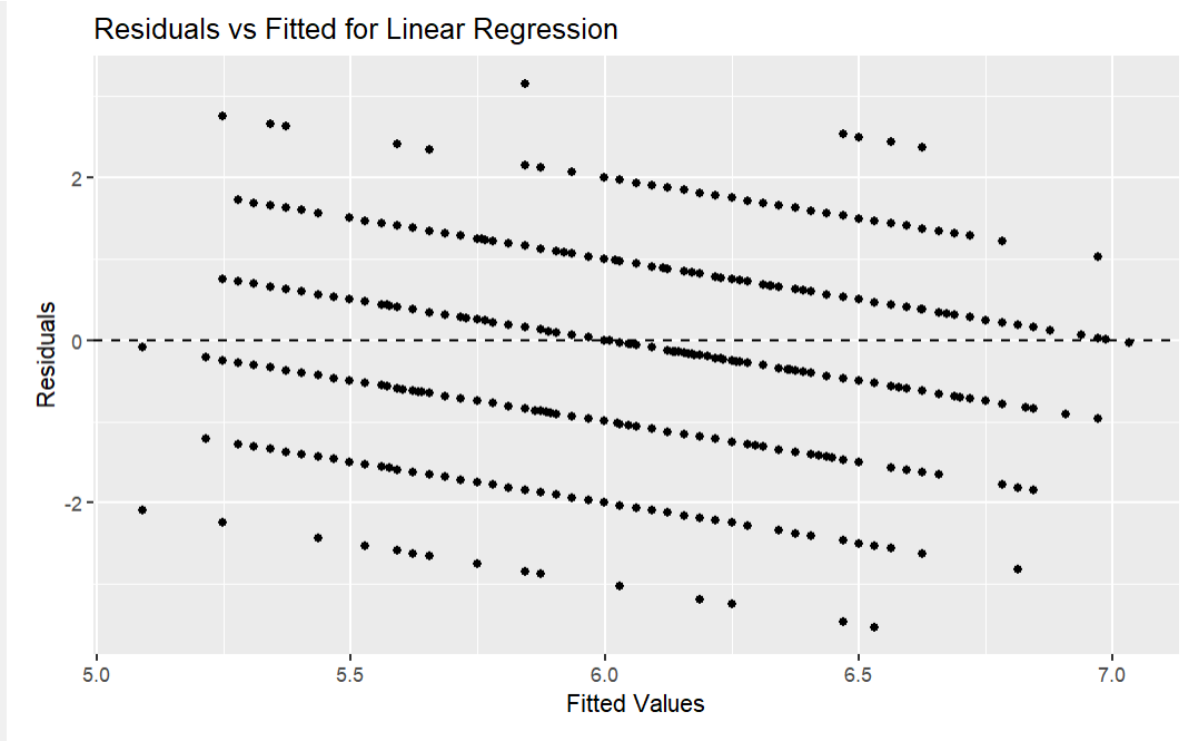
- **Coefficients:**

- The **Intercept** is -9.21366, which is the log-odds of a wine being high quality when the alcohol and pH are at zero. The intercept's z-value is -12.432 with a highly significant p-value (<2e-16), indicating that it is statistically significant.

- The coefficient for **alcohol** is 0.81620, suggesting that for each one-unit increase in alcohol, the log-odds of the wine being high quality increase by 0.81620, holding pH constant. The z-value is 24.505, with a highly significant p-value (<2e-16).

- The coefficient for **pH** is 0.46817, suggesting that for each one-unit increase in pH, the log-odds of the wine being high quality increase by 0.46817, holding alcohol constant. The z-value is 2.116 with a p-value of 0.0343, which is statistically significant at the 5% level.

- **Number of Fisher Scoring iterations:** This indicates the number of iterations it took for the algorithm to converge on a solution, which is 4 in this case.



This is a residuals versus fitted values plot for a linear regression model. It's a diagnostic tool used to detect non-linearity and outliers. Here's how to interpret it:

- **Horizontal Axis (Fitted Values):** These are the predicted values by the model. For linear regression, we expect the relationship between the independent variables and the dependent variable to be linear.

- **Vertical Axis (Residuals):** These are the differences between the observed values and the fitted values. If the model is well-fitted, the residuals should be randomly scattered around zero without any systematic patterns.

- **Pattern in the Plot:**

- The residuals should be centered around the horizontal line at zero, which would indicate that the model's predictions are unbiased.

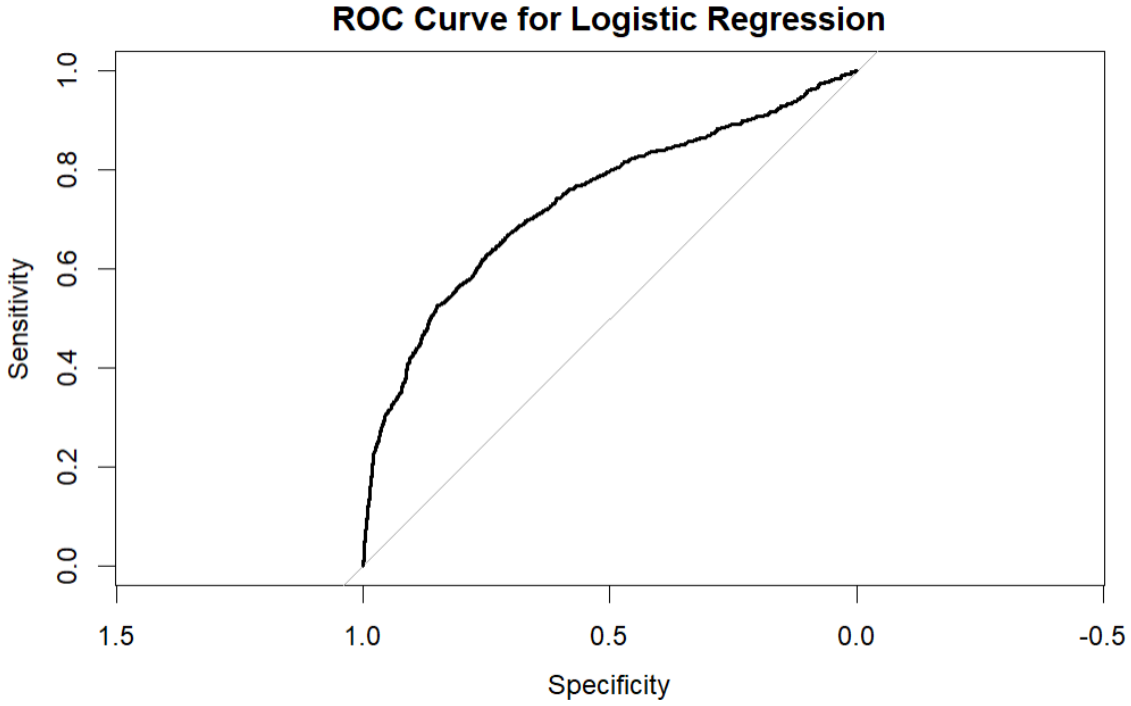
- The spread of residuals should be roughly constant across the range of fitted values to satisfy the assumption of homoscedasticity (constant variance of residuals).

- **Interpretation:**

- The dashed line represents zero, where ideally most of the residuals would cluster if the model's predictions were perfect.

- The plot shows that the residuals appear to be randomly scattered around the zero line, which is good as it suggests no obvious patterns of non-linearity.

- The absence of clear patterns or curves suggests that the assumption of linearity is not violated.



This is a Receiver Operating Characteristic (ROC) curve for a logistic regression model. The ROC curve is a graphical representation of a classifier's performance across all classification thresholds. Here's how to interpret it:

- **ROC Curve:** The curve shows the trade-off between sensitivity and specificity (1 - FPR). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the

test. In other words, the closer the curve is to the top-left corner, the better the model is at distinguishing between the positive and negative classes.

- **Diagonal Line:** The gray diagonal line represents a completely random classifier (AUC = 0.5). A good model has an ROC curve that bows up toward the top left, away from the diagonal line.

- **Interpretation of the Curve:**

 - This ROC curve is above the diagonal line, indicating that the logistic regression model has a good classification ability, better than random chance.

 - The curve approaches the top-left corner, suggesting high sensitivity and specificity, but it's not perfect. A perfect test would have a point in the top-left corner of the plot.

 - There's no clear 'elbow' in this ROC curve, which sometimes can be used to choose a threshold for classification.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  701  551
         1  939 2707

               Accuracy : 0.6958
                 95% CI : (0.6827, 0.7087)
    No Information Rate : 0.6652
    P-Value [Acc > NIR] : 2.544e-06

                  Kappa : 0.2744

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4274
            Specificity : 0.8309
         Pos Pred Value : 0.5599
         Neg Pred Value : 0.7425
             Prevalence : 0.3348
         Detection Rate : 0.1431
   Detection Prevalence : 0.2556
      Balanced Accuracy : 0.6292

       'Positive' Class : 0
```

The confusion matrix and associated statistics is used for a classification model, likely from a logistic regression given the context of previous discussions. Here's a breakdown of each part:

- **Confusion Matrix:**

  - True negatives (TN): 701 instances were correctly predicted as class 0.

  - False positives (FP): 551 instances were incorrectly predicted as class 1 when they were actually class 0.

  - False negatives (FN): 939 instances were incorrectly predicted as class 0 when they were actually class 1.

  - True positives (TP): 2707 instances were correctly predicted as class 1.

- **Accuracy:** The model correctly predicted 69.58% of the instances. The 95% Confidence Interval (CI) for this estimate is between 68.27% and 70.87%, which gives a range where the true accuracy is likely to fall.

- **Balanced Accuracy:** The average of sensitivity and specificity is 62.92%, which accounts for any class imbalance..