Decision 520Q

Dexter Nguyen

Red Wine Quality Prediction
Using Regression Modeling and Machine Learning

November 15, 2020

Contents

Business Understanding

The red wine industry shows a recent exponential growth as social drinking is on the rise. Nowadays, industry players are using product quality certifications to promote their products. This is a time-consuming process and requires the assessment given by human experts, which makes this process very expensive. Also, the price of red wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Another vital factor in red wine certification and quality assessment is physicochemical tests, which are laboratory-based and consider factors like acidity, pH level, sugar, and other chemical properties. The red wine market would be of interest if the human quality of tasting can be related to wine's chemical properties so that certification and quality assessment and assurance processes are more controlled. This project aims to determine which features are the best quality red wine indicators and generate insights into each of these factors to our model's red wine quality.

Data Understanding

My analysis will use Red Wine Quality Data Set, available on the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/wine+quality). I obtained the red wine samples from the north of Portugal to model red wine quality based on physicochemical tests. The dataset contains a total of 12 variables, which were recorded for 1,599 observations. This data will allow us to create different regression models to determine how different independent variables help predict our dependent variable, quality. Knowing how each variable will impact the red wine quality will help producers, distributors, and businesses in the red wine industry better assess their production, distribution, and pricing strategy.

Data Preparation

Data Cleaning

My first step was to clean and prepare the data for analysis. I went through different steps of data cleaning. First, I checked the data types focusing on numerical and categorical to simplify the correlation's computation and visualization. Second, I tried to identify any missing values existing in our data set. Last, I researched each column/feature's statistical summary to detect any problem like outliers and abnormal distributions.

Data Exploration and Transformation

To see which variables are likely to affect the quality of red wine the most, I ran a correlation analysis of our independent variables against our dependent variable, quality. This analysis ended up with a list of variables of interest that had the highest correlations with quality (Figure 1). In order of highest correlation, these variables are:

1. Alcohol: the amount of alcohol in wine
2. Volatile acidity: are high acetic acid in wine which leads to an unpleasant vinegar taste
3. Sulphates: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant
4. Citric Acid: acts as a preservative to increase acidity (small quantities add freshness and flavor to wines)

5. Total Sulfur Dioxide: is the amount of free + bound forms of SO2
6. Density: sweeter wines have a higher density
7. Chlorides: the amount of salt in the wine
8. Fixed acidity: are non-volatile acids that do not evaporate readily
9. pH: the level of acidity
10. Free Sulfur Dioxide: it prevents microbial growth and the oxidation of wine
11. Residual sugar: is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between - sweetness and sourness (wines > 45g/ltrs are sweet)

Starting with our dependent variable, quality, I found the popularity of the medium/average values of quality: 5 and 6. Considering the dependent variable's transformation, I found out that our data is normally distributed (Figure 2). This conclusion can be verified by running a QQ plot, which shows no need to transform our data (Figure 3).

Next, for independent numerical variables, the first step to further analyze the relationship with our dependent variable was to create density plots visualizing the spread of the data (Figure 4, 5, 6). It can be seen that most red wines' pH levels are always between 3-4 and chlorides – the amount of salt is most prevalent at level 0.1. After analyzing the density plots, I plotted the interaction between our numeric variables of interest and our dependent variable of quality (Figure 7, 8, 9). Three different patterns can be observed. First, there are positive relationships between quality and critic.acid, alcohol, and sulphates. Even though wines with a higher level of alcohol may make them less popular, they should be highly rated in quality. Second, there are negative relationships between quality and volatile.acidity, density, and pH. It is reasonable that less sweet wines and a lower level of acidity are favored in quality testings.

Last, these independent variables show no significant relationship with quality: residual.sugar, chlorides, and total.sulfur.dioxide.

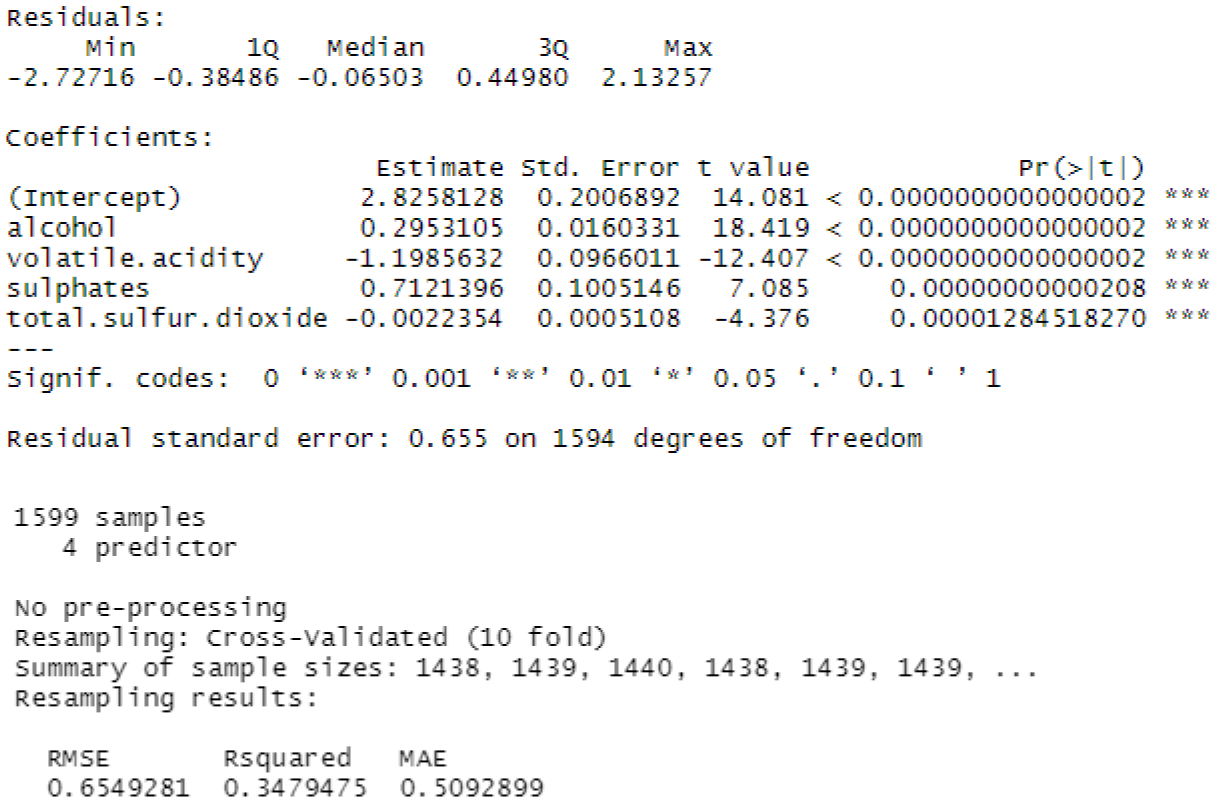
To dive deep into relationships within independent variables and with quality, I built different three-dimensional plots. When inspecting the two variables, alcohol and volatile.acidity with quality (Figures 10), we can see that with red wines' alcohol level between 9% to 12%, the level of volatile acidity decreases as the wines' alcohol level increases. For higher alcohol content (>12%), the pattern reverses, implying high-quality wines' popularity. Keep researching the alcohol variable, I selected the citric.acid and visualized their interactions with quality (Figure 11). Interestingly, for wines with an alcohol percentage level below 14, as the level of citric acid increases, there is a rise in red wines' quality. The only exception was at alcohol 14%, where the citric acid level drops as the wine's quality increases. Finally, an interaction analysis using chlorides in relationships with alcohol and quality shows that the wines' quality decreases when chloride level decreases at the alcohol before 12%. However, the quality of red wine increases as the chloride level increases at the alcohol level from 12% (Figure 12).

Last, we considered if the collinearity problem existed in our data. As a result of correlation analysis and VIF verification, we discovered some variables with slightly high correlations. To deal with such a potential problem, we will take advantage of the LASSO regularization technique in the next modeling part.

Modeling

Based on the EDA and correlation analysis, three potential models were used in the modeling part.

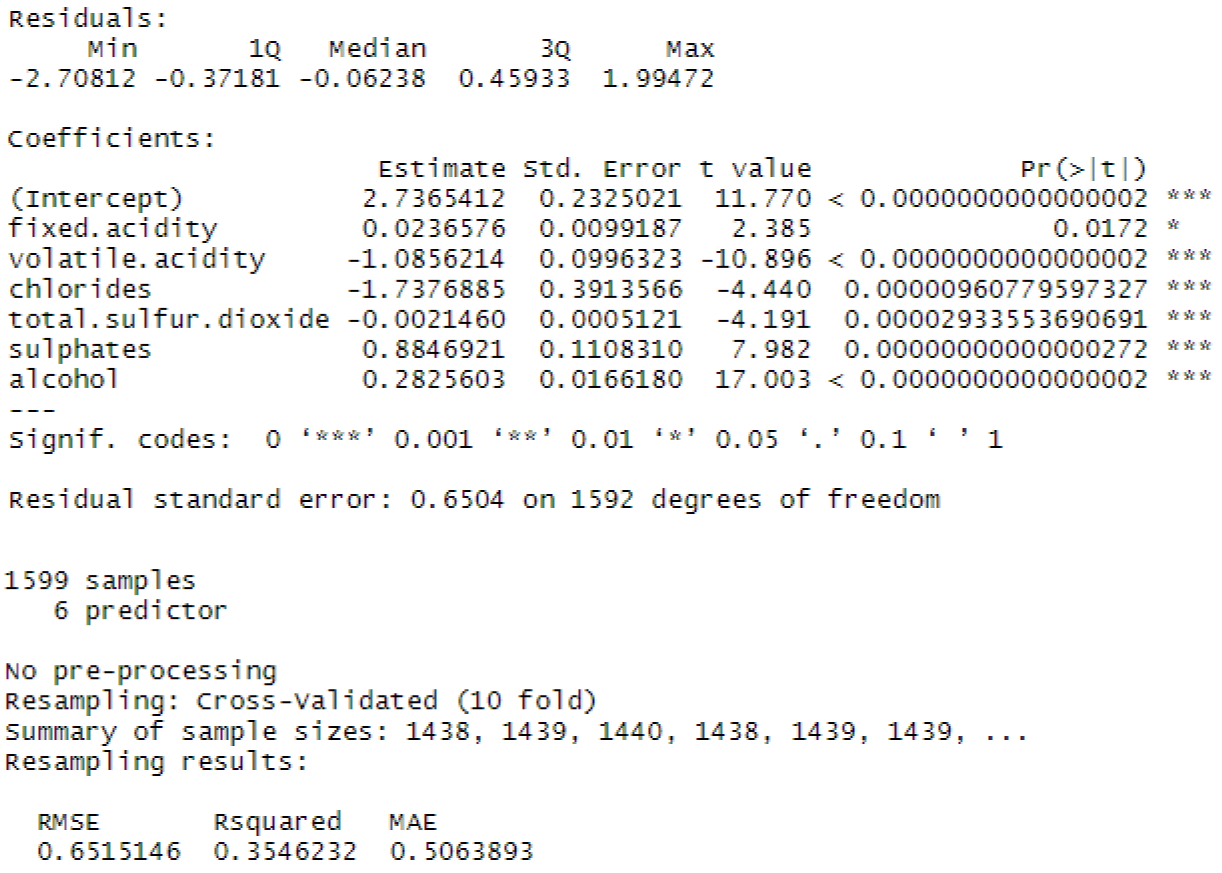
***Model 1*:** Since the correlation analysis shows that quality is highly correlated with a subset of variables (our "Top 5"), I employed multi-linear regression to build an optimal prediction model for the red wine quality. Removing a non-significant independent variable from the initial model, we got "Model 1", which included our "Top 4" explanatory variables. Using K-Fold Cross Validation, we have Model 1 summary as below:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.72716 -0.38486 -0.06503  0.44980  2.13257 

Coefficients:
                       Estimate Std. Error t value             Pr(>|t|)    
(Intercept)           2.8258128  0.2006892  14.081 < 0.0000000000000002 ***
alcohol               0.2953105  0.0160331  18.419 < 0.0000000000000002 ***
volatile.acidity     -1.1985632  0.0966011 -12.407 < 0.0000000000000002 ***
sulphates             0.7121396  0.1005146   7.085     0.00000000000208 ***
total.sulfur.dioxide -0.0022354  0.0005108  -4.376     0.00001284518270 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.655 on 1594 degrees of freedom


 1599 samples
    4 predictor

 No pre-processing
 Resampling: Cross-Validated (10 fold) 
 Summary of sample sizes: 1438, 1439, 1440, 1438, 1439, 1439, ... 
 Resampling results:

   RMSE       Rsquared   MAE      
   0.6549281  0.3479475  0.5092899
```

In Model 1, all identified variables are highly correlated with our target variable (quality) and show statistical significance. Alcohol and sulphates have positive relationships with quality, implying that the more level of alcohol and sulphates will translate into a higher quality of red wine. Reversely, there are negative relationships between both volatile.acidity and total.sulfur.dioxide and quality, showing that people expect a low level of acetic acid and SO2 in high-quality wine. A large amount of acetic acid may lead to an unpleasant vinegar taste, for example.

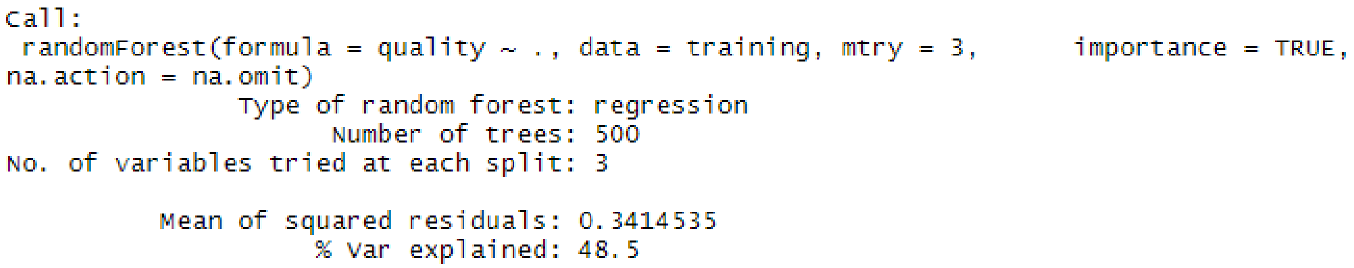
Model 2: Next, using the LASSO method, I came up with the second model ("Model 2") that performs both variable selection and regularization. This resulted in a subset of predictors (our "Top 6") that minimizes prediction error for a quantitative response variable - quality. This subset includes six variables: fixed.acidity, volatile.acidity, chlorides, total.sulfur.dioxide, sulphates, and alcohol. Applying K-Fold Cross Validation again, we got Model 2 summary as below:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.70812 -0.37181 -0.06238  0.45933  1.99472 

Coefficients:
                      Estimate Std. Error t value             Pr(>|t|)    
(Intercept)          2.7365412  0.2325021  11.770 < 0.0000000000000002 ***
fixed.acidity        0.0236576  0.0099187   2.385               0.0172 *  
volatile.acidity    -1.0856214  0.0996323 -10.896 < 0.0000000000000002 ***
chlorides           -1.7376885  0.3913566  -4.440   0.00000960779597327 ***
total.sulfur.dioxide -0.0021460  0.0005121  -4.191   0.00002933553690691 ***
sulphates            0.8846921  0.1108310   7.982   0.0000000000000272 ***
alcohol              0.2825603  0.0166180  17.003 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6504 on 1592 degrees of freedom


1599 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold) 
Summary of sample sizes: 1438, 1439, 1440, 1438, 1439, 1439, ... 
Resampling results:

  RMSE       Rsquared   MAE      
  0.6515146  0.3546232  0.5063893
```

All these six variables are highly correlated with our target variable (quality) and show highly statistical significance. Compared with Model 1, the new model has additional two variables: fixed.acidity and chlories, whose marginal impacts on quality are in different directions. A negative estimate coefficient of chlorides means that higher quality wine should have a smaller amount of salt. Meanwhile, there is a slight positive relationship between fixed acidity and quality, implying that non-volatile acids that do not evaporate readily should be an indicator of high-quality wine.

Model 3: Last, I ran Random Forest as a machine learning regression tree algorithm used in the modeling process. This helps to create a random sample of multiple regression decision trees and merges them to obtain a more stable and accurate prediction through cross-validation. We call this "Model 3", with its summary as below:



```
Call:
 randomForest(formula = quality ~ ., data = training, mtry = 3,      importance = TRUE,
na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.3414535
                    % Var explained: 48.5
```

Diving deep into variable selection, we have the top 10 predictors most important to the model (Figure 13). It is done by using MDI (Gini Importance or Mean Decrease in Impurity) that calculates each feature's importance as the sum over the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits. In comparison with Model 1 and Model 2, we have additional insights into such variables as density and pH.

Evaluation

After running our three models, I used three metrics: R-squared, RMSE, and MAE, to evaluate our model prediction performance. As we expected from Figure 14, Model 3 is the best in terms of all three metrics, with R-Squared: 48.50%, RMSE: 0.5843, and MAE: 0.4222. Model 1 and Model 2, whose predictors selected from our correlation analysis and regularization techniques, meanwhile, don't record much difference in terms of these performance metrics.

It is reasonable that Random Forest in Model 3 gives us superior "predictions". However, from a perspective of "marginal impact" interpretation, Model 1 and Model 2 may be the winners even though their performance measurements are behind. In the context of our business question focusing on the prediction of red wine quality, Model 3 will be the best choice.

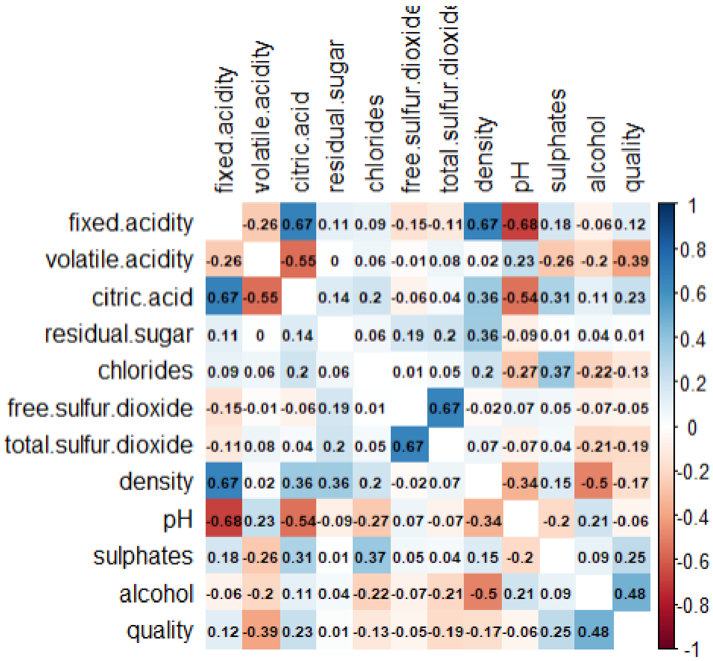
Deployment

By analyzing the physicochemical tests samples data of red wines from the north of Portugal, I was able to create a model that can help industry producers, distributors, and sellers predict the quality of red wine products and have a better understanding of each critical and up-to-date features. I have found that the Model 3 - Random Forest-based feature sets performed better than others. In general, using Model 3 as our best model for prediction, I determined four of the features as the most influential: volatile acidity, citric acid, sulphates, and alcohol. To be more specific, high-quality wines seem to have lower volatile acidity, higher alcohol, and medium-high sulphate values. Meanwhile, lower-quality wines tend to have low values for citric acid.

However, this analysis has some limitations. First, the main problem came from the fact that our data set was unbalanced. A majority of the quality values were “regular” (5 and 6), which made no significant contribution to finding an optimal model. These values made it harder to identify each factor's different influence on a "high" or "low" quality of the wine, which was the main focus of this analysis. In order to improve our predictive model, we need more balanced data. Another limitation worth mentioned from the data set was it only had 12 attributes, which can narrow down the accuracy of our predicting quality of red wine. The solution for this is to include more relevant data features, like the year of harvest, brew time, location, or wine type. In the future, we also can try other performance measures and other machine learning techniques for better performance and comparison of results. This analysis will help wine businesses predict the red wines' quality based on certain attributes and make and sell good associated products.

Appendix: Referenced Figures

Figure 1. Correlation Matrix



	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
fixed.acidity		-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile.acidity	-0.26		-0.55	0	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.2	-0.39
citric.acid	0.67	-0.55		0.14	0.2	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
residual.sugar	0.11	0	0.14		0.06	0.19	0.2	0.36	-0.09	0.01	0.04	0.01
chlorides	0.09	0.06	0.2	0.06		0.01	0.05	0.2	-0.27	0.37	-0.22	-0.13
free.sulfur.dioxide	-0.15	-0.01	-0.06	0.19	0.01		0.67	-0.02	0.07	0.05	-0.07	-0.05
total.sulfur.dioxide	-0.11	0.08	0.04	0.2	0.05	0.67		0.07	-0.07	0.04	-0.21	-0.19
density	0.67	0.02	0.36	0.36	0.2	-0.02	0.07		-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34		-0.2	0.21	-0.06
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.2		0.09	0.25
alcohol	-0.06	-0.2	0.11	0.04	-0.22	-0.07	-0.21	-0.5	0.21	0.09		0.48
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.17	-0.06	0.25	0.48	

Figure 2.

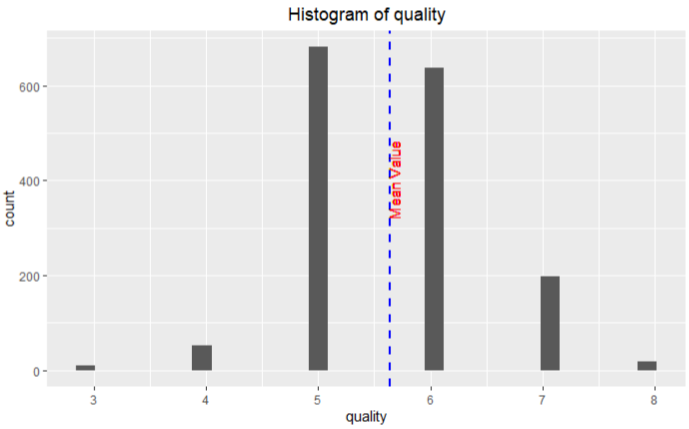


Figure 3.

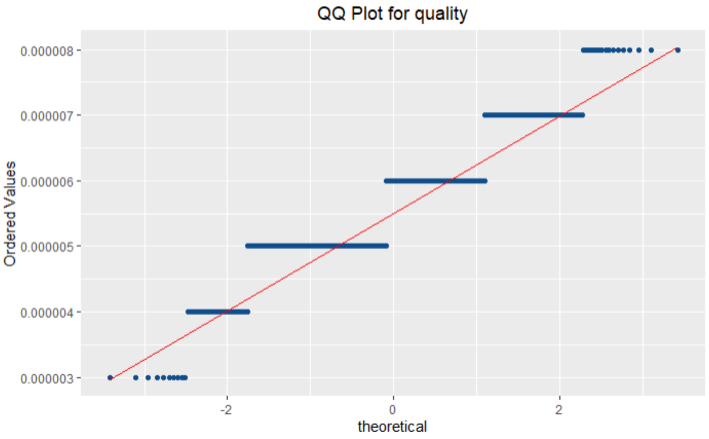


Figure 4.

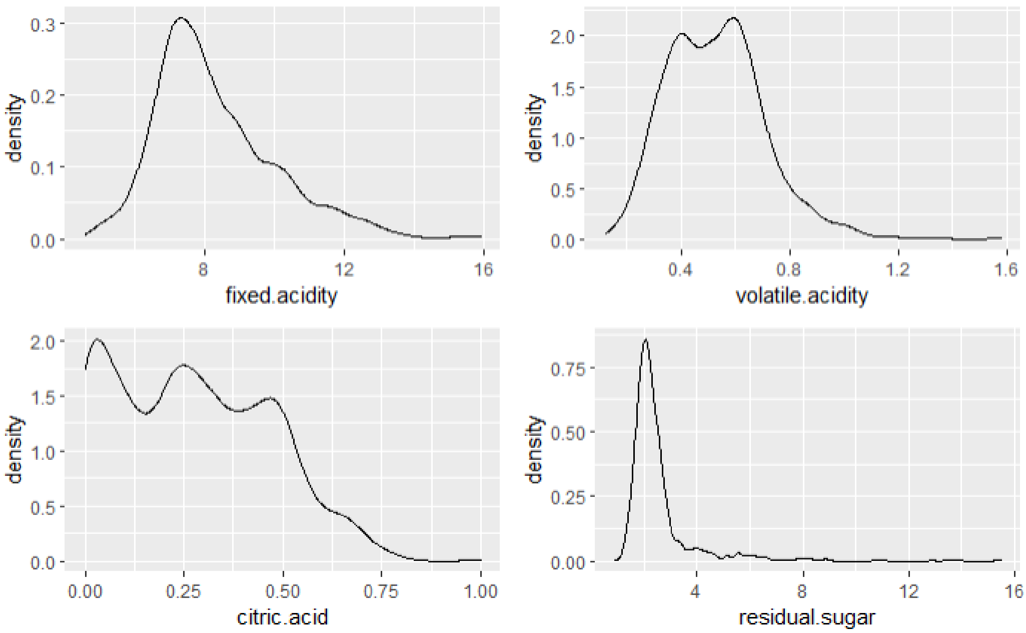


Figure 5.

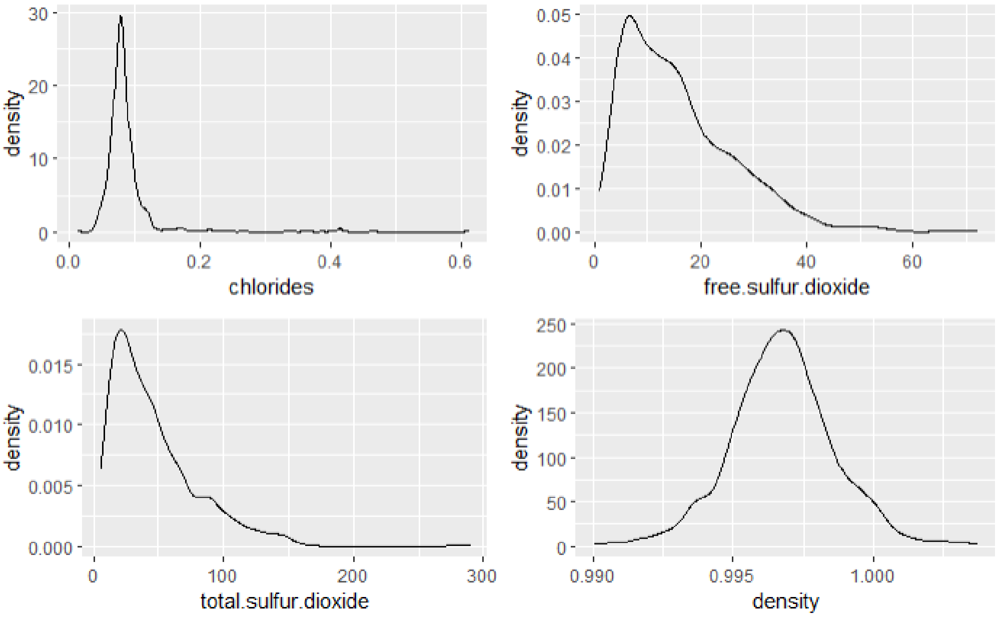


Figure 6.

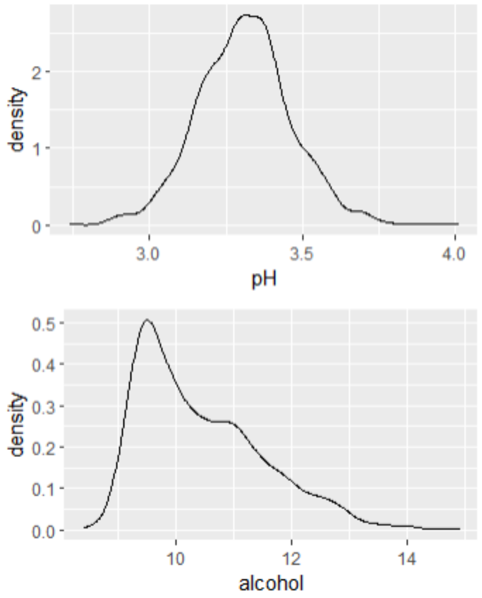


Figure 7.

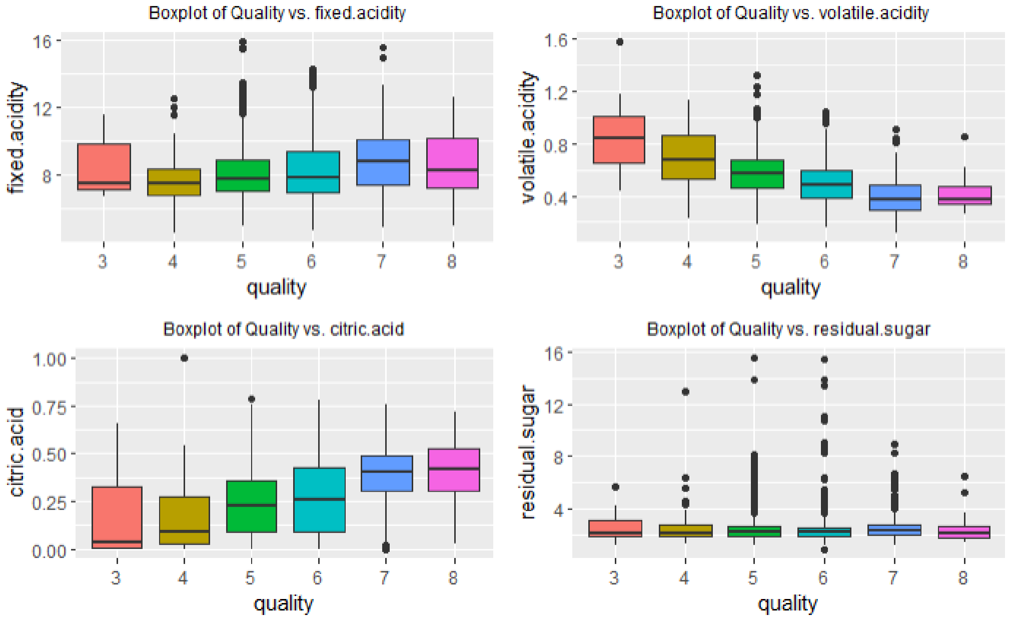


Figure 8.

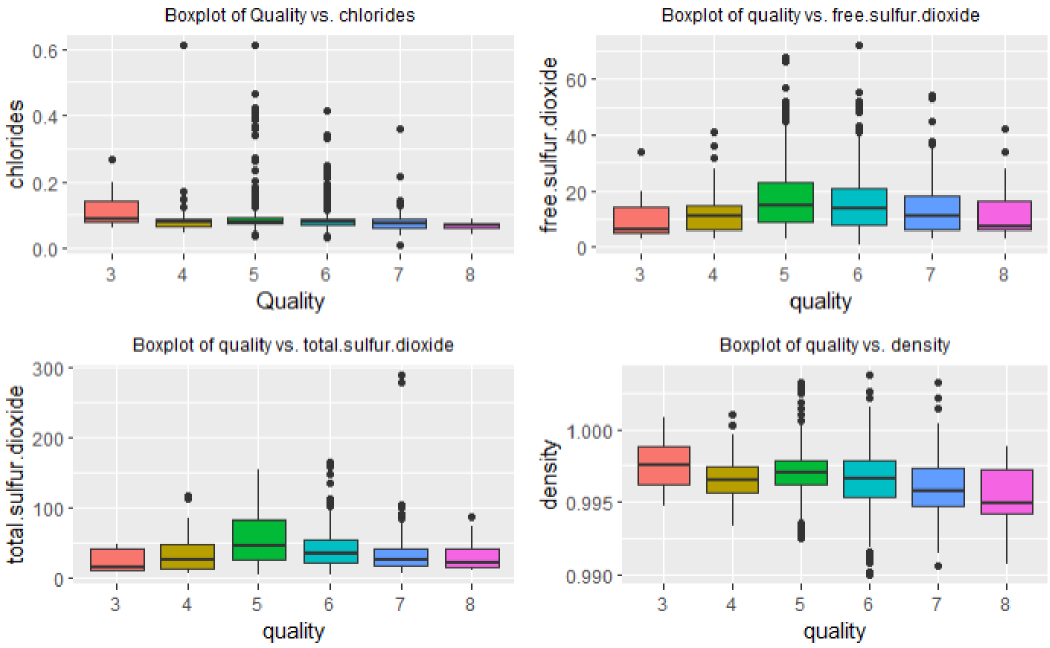


Figure 9.

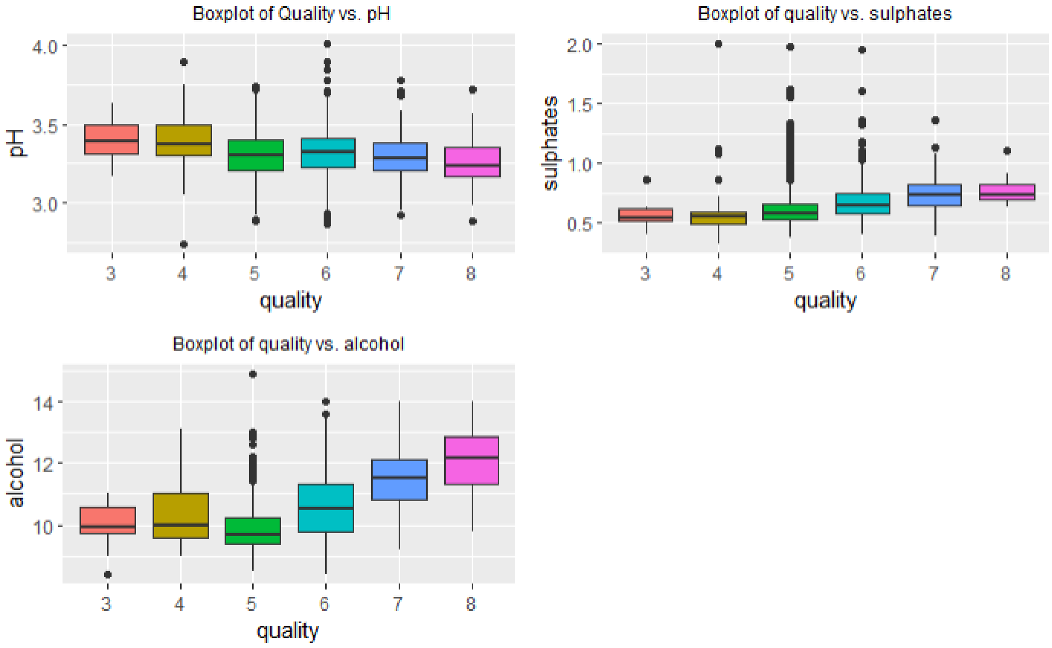


Figure 10.

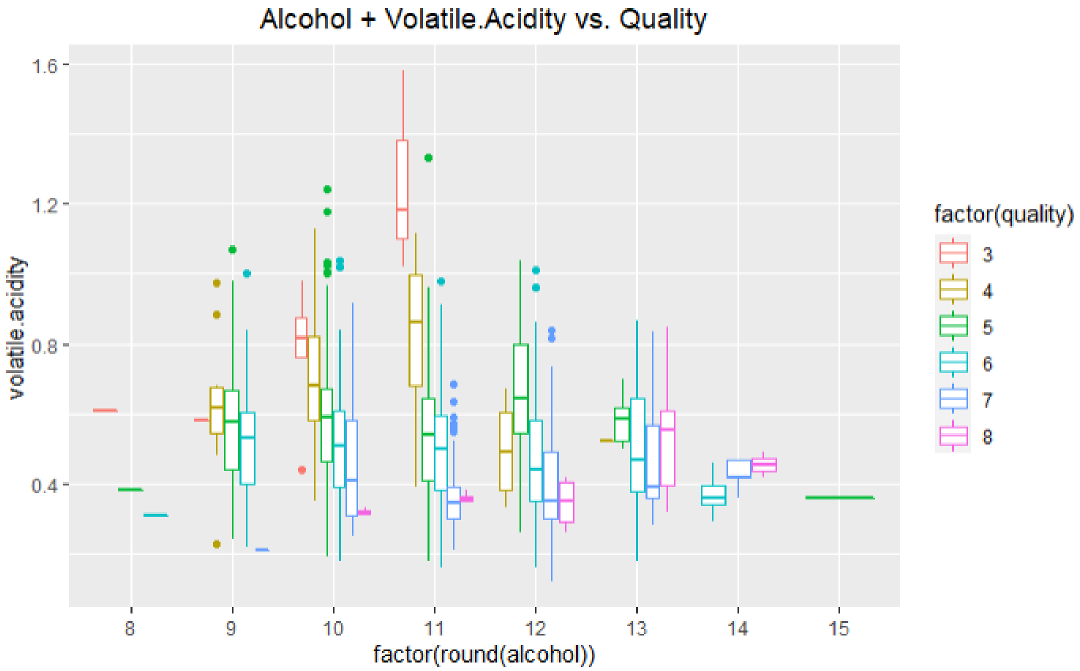


Figure 11.

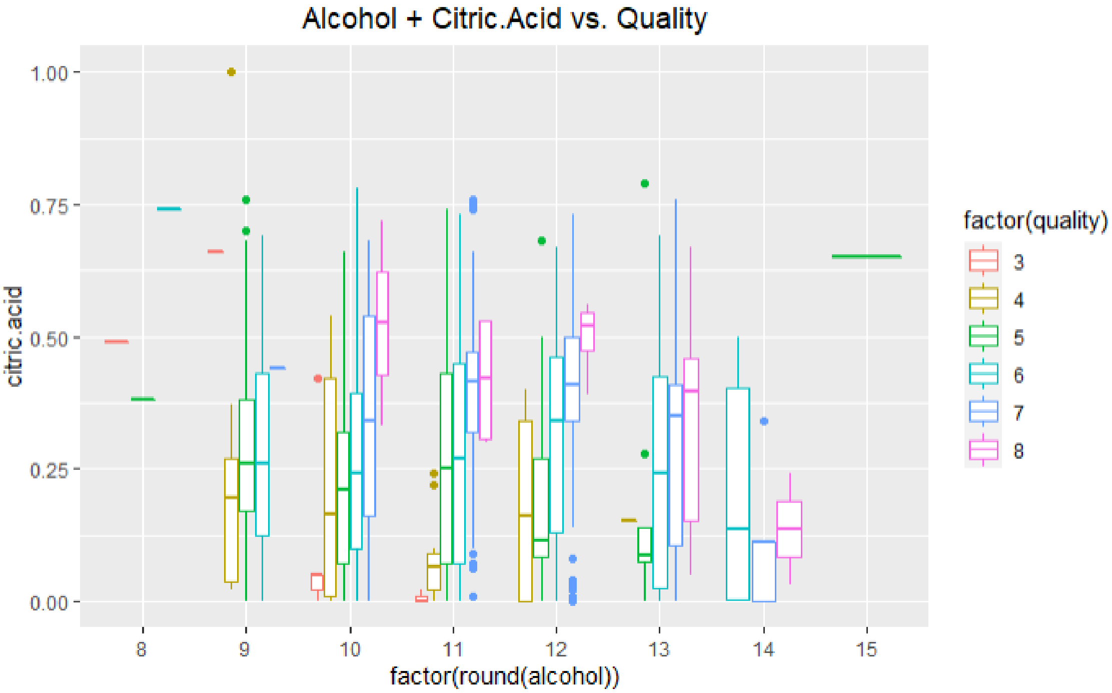


Figure 12.

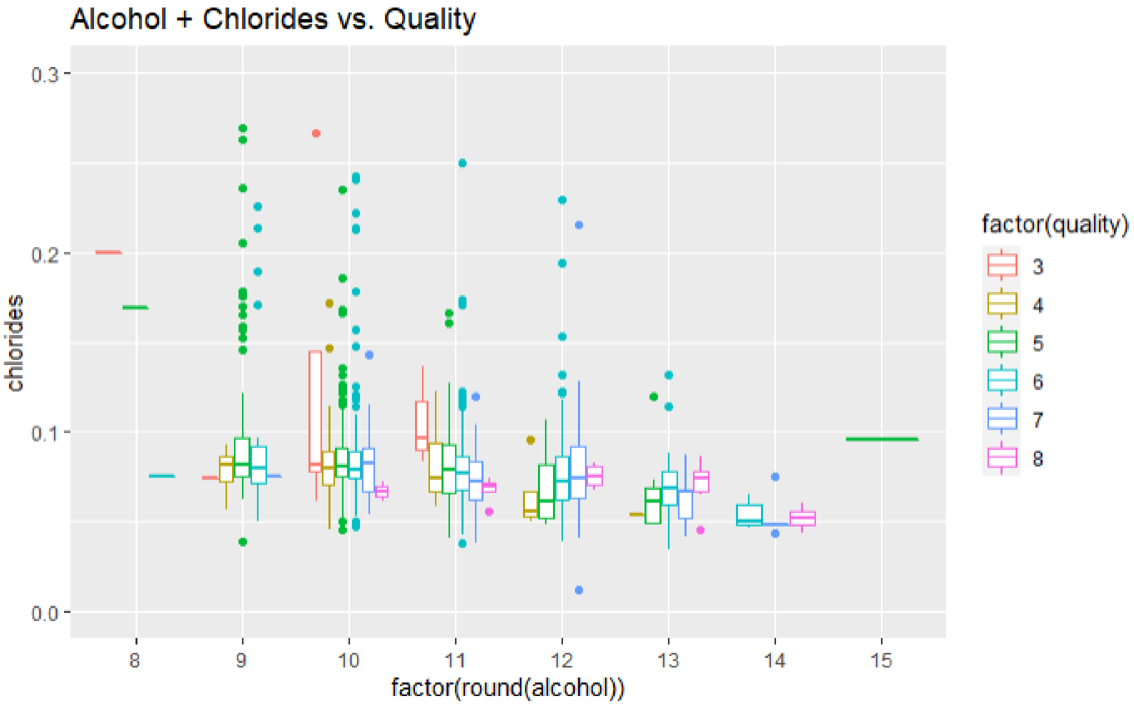


Figure 13.

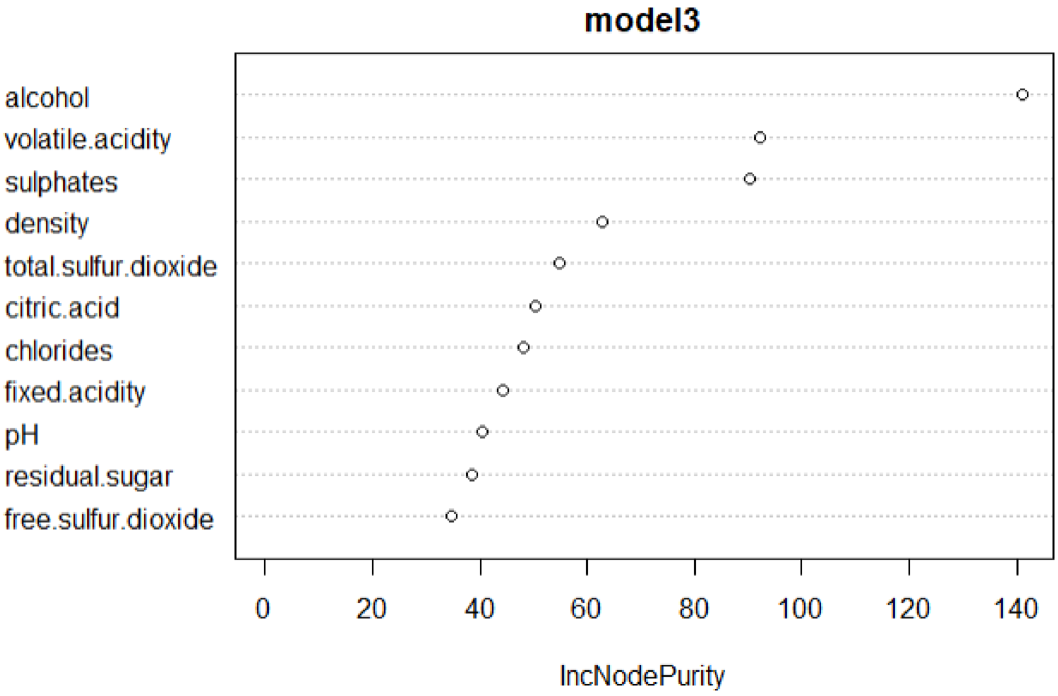
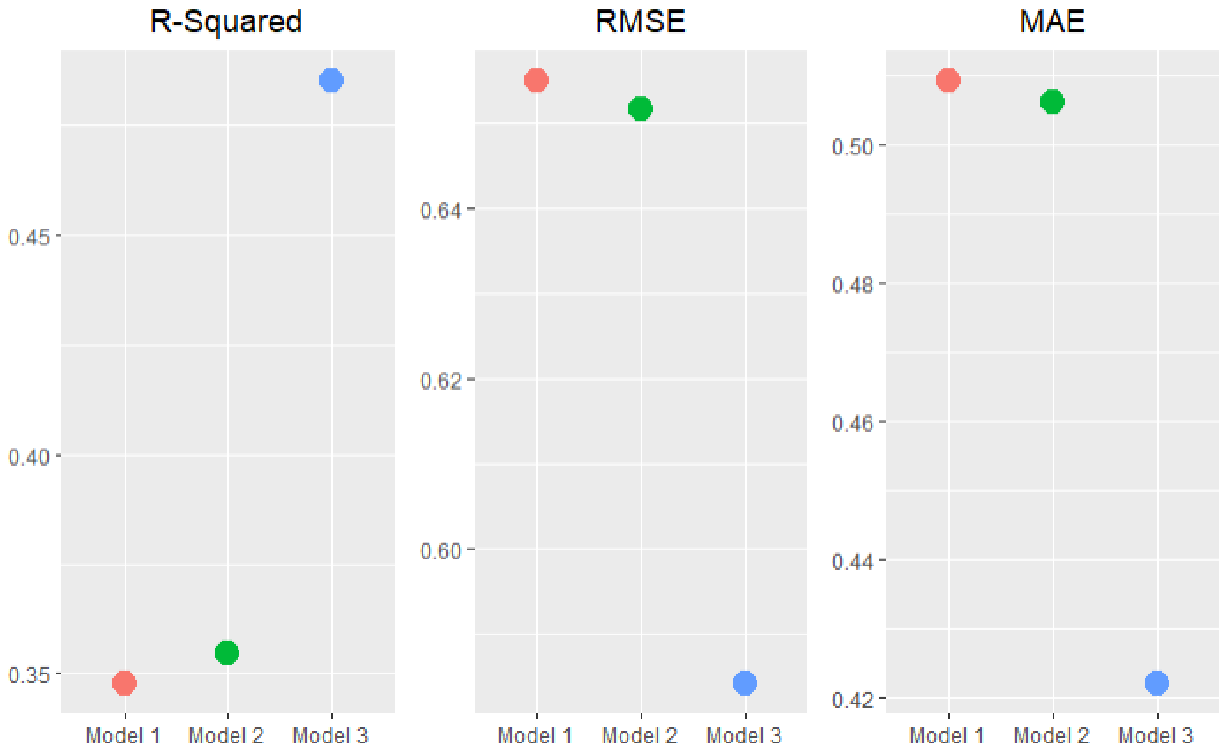


Figure 14.


R-Squared
0.45
0.40
0.35
Model 1
Model 2
Model 3
RMSE
0.64
0.62
0.60
Model 1
Model 2
Model 3
MAE
0.50
0.48
0.46
0.44
0.42
Model 1
Model 2
Model 3