# Predictive Analysis of Wine Data Providing Solution to a Business Problem

## An analytics workflow on predicting wine quality using their physicochemical features

**Yao Luo**[1]

[1] Farmer School of Business, Miami University, Oxford, OH, USA

E-mail: luoy25@miamioh.edu

### Abstract

A business problem was brought up by the sales team of a wine company, Miami Wine. Miami Wine is looking to cut costs by only bringing in experts for high-quality wine but are lacking the power of predicting the quality of the wine before bringing in experts. In this article, a complete analytics workflow will be presented, focusing on Understanding and processing the data, Analytical methods and tools, Results, and limitations for the proposed business problem. A multiple linear regression model will be presented as the solution and several statistical metrics will be used to evaluate the significance of the model and its prediction power. How the proposed solution will create gain and reduce pain to the sales team as well as how the model can be implemented and improved will be discussed. The inevitable limitations of the model will also be discussed, and recommendations for improvements to the current model will also be given.

Keywords: UCI Wine data [1], Business Applications, Multiple Linear Regression, Stepwise, Evaluation, Limitations, Predictive Modelling, Analytics Workflow, Methods, Reproducibility

## 1. Background

Miami Wine is a well-known wine company that produces and sells Vinho Verdo wine globally. To advertise the wine's quality, the company brings in wine experts to examine the wine and its expert rating will be a market point of the wine. The company's sales team realized that it is very costly to bring in experts for every single wine they produced. The current analyst on the company's sales team has been trying to find a way to lower the expert rating costs. One of the suggestions is that they would like to reduce the amount of wine that needed to be rated by an expert. They realized that some of the wine ended up with low expert ratings, meaning that these ratings are useless for advertising, but they still have to spend money on bringing the experts in. This may not seem to be a problem for just one wine, but it gets very costly for a large amount of wine that the company produced. The sales team suggested not bring in experts for the wine that is likely to receive a low rating.

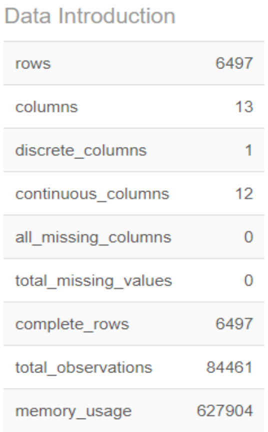
The issue for the sales team is that they lack the power in knowing and predicting how each wine will be rated by the expert. They have no idea whether each wine will be rated high or low before getting an expert. However, they have historical data on the expert ratings on both red and white wines they produced as well as the chemical features of this wine. They wondered if it is possible to make predictions of the quality ratings only using the chemical features of the wine.

In this article, the wine data will be analyzed and a solution will be provided to the wine company’s sales team targeting their needs. The analytics workflow will be introduced from preprocessing the data to evaluating the solution. Some discussions and the drawback of the solution will also be mentioned in the later section of the article.

## 2. Data & Pre-processing

The original data [2] given includes red and white wine data each in a separate CSV file. Both datasets have 12 variables including 11 input variables on chemical features of the wine, and one output variable representing wine quality with scores between 0 to 10. Missing data was checked by summing all blank rows in each column for both datasets using tidyverse library and no missing value was found in both datasets. Therefore no additional imputation was needed.

Since both datasets have the same data structure without any missing values, they were combined into one dataset called *wine* using *rbind* function from the *dplyr* package [3]. A new dummy variable *wine.type* was created representing whether the wine was a red or white wine with white wine coded as the base. Variable *wine.type* was then recoded as a factor based on the categorical nature of the variable. The final dataset has 6497 observations of 13 variables with *wine.type* being a factor while all other variables being numeric (Figure 1).



Data Introduction

| | |
|---|---|
| rows | 6497 |
| columns | 13 |
| discrete_columns | 1 |
| continuous_columns | 12 |
| all_missing_columns | 0 |
| total_missing_values | 0 |
| complete_rows | 6497 |
| total_observations | 84461 |
| memory_usage | 627904 |

**Figure 1.** *The metadata for the finalized wine dataset*

To explore the data more, *DataExplorer* library [4] was used to generate scatter plots, histograms, and bar plots using *plot_scatterplot*, *plot_histogram*, and *plot_bar* functions respectively. No obvious trend or relationships were observed through the scatterplots generated (Figure 2 & Figure 3). From the histogram (Figure 4), most numeric variables appeared to be right-skewed and unimodal which makes sense since these chemical variables can not have negative values. Both PH and quality variables appeared to be normally distributed. Total sulfur dioxide seemed to have a bimodal distribution.

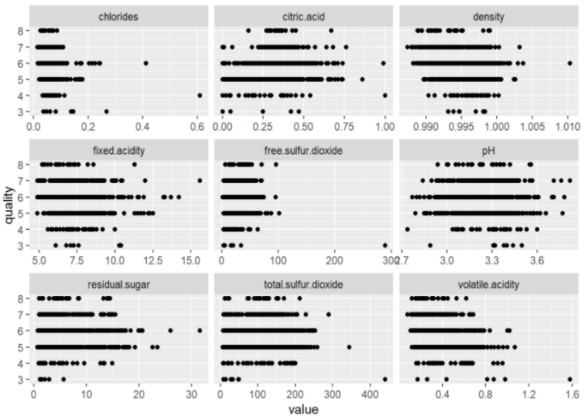


**Figure 2.** *Scatterplots for input variables against quality (1)*

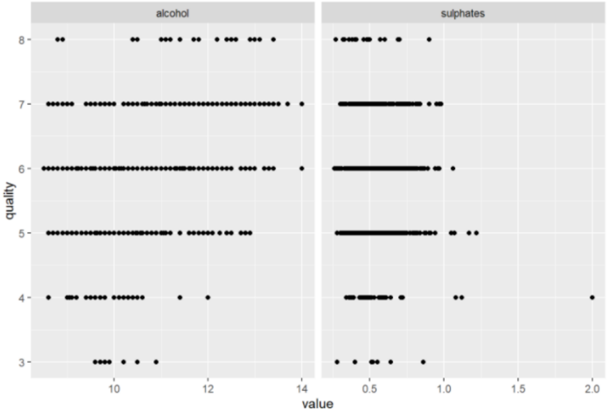


**Figure 3.** *Scatterplots for input variables against quality (2)*

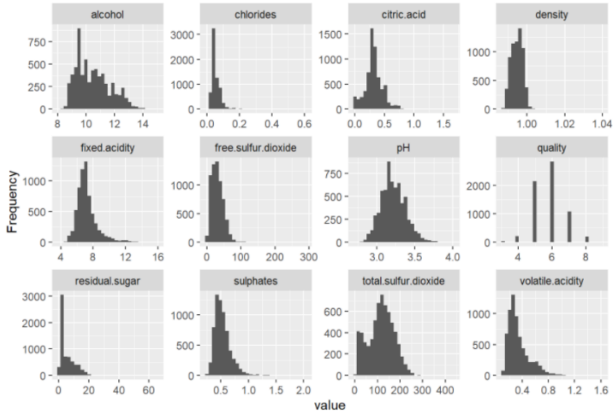


**Figure 4.** *Histogram for distribution of all numerical variables*

## 3. Methods

The method used modeling is a multiple linear regression as such models generally have low variance. Before modeling the finalized wine data, a training and validation split was performed so that the prediction capacity can be evaluated later on. The wine dataset was split into 75% training data and 25% testing data setting the seed to be 13. A training index was generated using the random generation and was used for the splitting process. The final training data was stored in a dataframe named *train* with 4873 rows. The final validation data was stored in a dataframe named *valid* with 1624 rows.

To model the finalized data, a regression model was created. Only the *train* data was used for the creation of the model. A full model using all variables was first generated using the multiple linear regression *lm* functions. From there, a stepwise model was then derived from the full model using the *stepAIC* function from the *MASS* library [5] with the direction set to both directions and trace set to FALSE since we do not care about the elimination process.

### *3.1 Reproducibility*

For both the data preprocessing steps and model evaluation process, some functions were created and used to enhance analytics workflow performance and for reproducibility. A function called *Metadata* was created with the aim to produce an overview of the data combining the *introduce* function from the *DataExplorer* library as well as *knitr::kable* function from *KableExtra* package [6] to generate a table with desired summary statistics.

Another function named *Evaluation* was created for evaluating model performance where *MLmetrics* library was used. The function inputs the created model as well as predicted y-value and actual y-value. Several evaluation metrics were calculated and reported when the function is called including Adjusted R-Squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) using *adjr2* function, *mse* function, *rmse* function, and *mae* function respectively from the *MLmetrics* library [7].

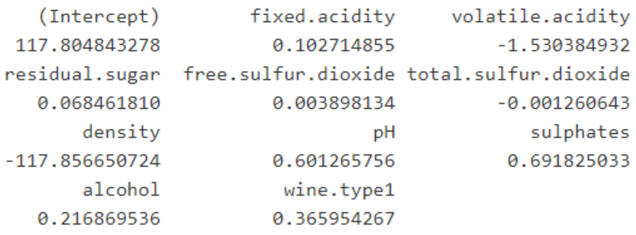
## 4. Analytics Result

A statistical model is provided to the sales team with some levels of prediction power on expert ratings using the chemical features of the wine. In this way, the sales team will only need to collect the chemical features of each new wine they produced and run it through the model to get an estimation of the predicted expert rating to help the sales team on deciding whether they are going to actually bring in the expert for that particular wine.

The final stepwise model was a multiple linear regression with 10 inputs and can be expressed by the equation below:

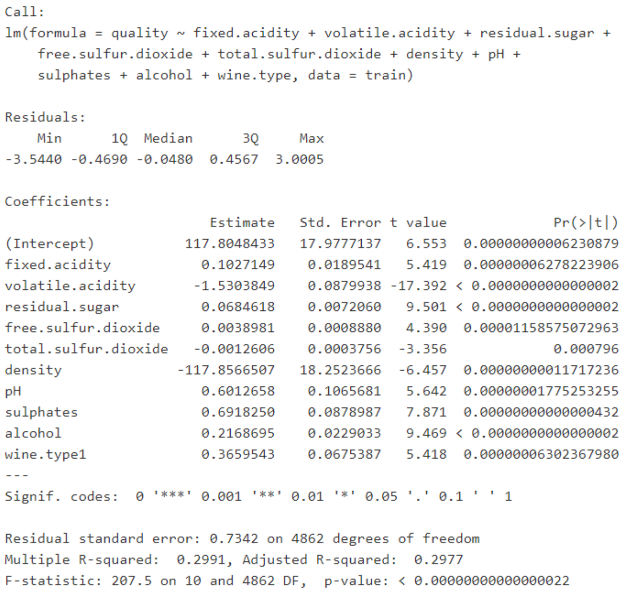
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10}$$

where the following are values from $\beta_0$ to $\beta_{10}$ and from $x_1$ to $x_{10}$:



```
  (Intercept)         fixed.acidity     volatile.acidity
 117.804843278          0.102714855         -1.530384932
residual.sugar  free.sulfur.dioxide total.sulfur.dioxide
   0.068461810          0.003898134         -0.001260643
       density                   pH            sulphates
-117.856650724          0.601265756          0.691825033
       alcohol           wine.type1
   0.216869536          0.365954267
```

Based on the summary below (Figure 5), the overall model appeared to be statistically significant due to the extremely small p-value (less than 0.001) from the F-statistics. All input variables used in the model were statistically significant based on the small p-values from the individual t-tests for each separate variable (All p-values less than 0.001). The model has a Root Mean Squared Error of 0.7342 on 4862 degrees of freedom. The Adjusted R-squared is 0.2977.



```
Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar + 
    free.sulfur.dioxide + total.sulfur.dioxide + density + pH + 
    sulphates + alcohol + wine.type, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5440 -0.4690 -0.0480  0.4567  3.0005 

Coefficients:
                        Estimate   Std. Error t value              Pr(>|t|)
(Intercept)          117.8048433   17.9777137   6.553   0.00000000006230879
fixed.acidity          0.1027149    0.0189541   5.419   0.00000006278223906
volatile.acidity      -1.5303849    0.0879938 -17.392 < 0.0000000000000002
residual.sugar         0.0684618    0.0072060   9.501 < 0.0000000000000002
free.sulfur.dioxide    0.0038981    0.0008880   4.390   0.00001158575072963
total.sulfur.dioxide  -0.0012606    0.0003756  -3.356              0.000796
density             -117.8566507   18.2523666  -6.457   0.00000000011717236
pH                     0.6012658    0.1065681   5.642   0.00000001775253255
sulphates              0.6918250    0.0878987   7.871   0.00000000000000432
alcohol                0.2168695    0.0229033   9.469 < 0.0000000000000002
wine.type1             0.3659543    0.0675387   5.418   0.00000006302367980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7342 on 4862 degrees of freedom
Multiple R-squared:  0.2991, Adjusted R-squared:  0.2977 
F-statistic: 207.5 on 10 and 4862 DF,  p-value: < 0.00000000000000022
```

**Figure 5.** *Stepwise regression summary output*

This model will be mainly used as a predictive model inputting wine chemical features to predict its quality score. To ensure the model predictive ability, some evaluation methods were applied. The metrics used are: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and adjusted R-squared. All metrics were calculated from the function created called Evaluation

(See Section 3.1 for details). The evaluation was performed on both training and validation datasets to ensure that there is no potential overfitting issue.

When evaluated from the training dataset, an Adjusted R-Squared of 0.2977 was produced, with MSE being 0.5379, RMSE being 0.7334, MAE being 0.5686. When evaluated from the validation dataset, an Adjusted R-Squared of 0.2977 was produced, with MSE being 0.5357, RMSE being 0.7319, MAE being 0.57. When comparing the two sets of metrics, no overfitting issues were found since the metrics results appear to be similar between the training and validation data. The model has the same Adjusted R-Squared for both datasets being 0.2977.

## 5. Discussions & Business Value

The model has an adjusted R-squared of 0.2977 meaning that on average 29.77% of the variation in wine quality can be interpreted by the model. Although this improves the chances of predicting the wine quality from none to around 30 percent, the model is still very inaccurate, meaning that the sales team won't be able to trust a lot of the predictions made by the model. To some extent, the model can be used as a descriptive model to help the sales team understand what features of the wine are more impactful when it comes to getting a better or lower expert rating.

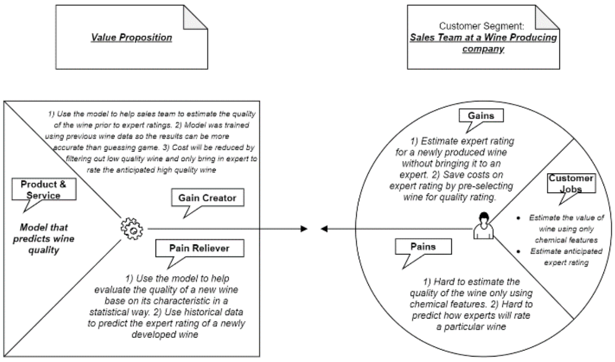

Value Proposition
Customer Segment:
Sales Team at a Wine Producing company
1) Use the model to help sales team to estimate the quality of the wine prior to expert ratings. 2) Model was trained using previous wine data so the results can be more accurate than guessing game. 3) Cost will be reduced by filtering out low quality wine and only bring in expert to rate the anticipated high quality wine
Product & Service
Model that predicts wine quality
Gain Creator
Pain Reliever
1) Use the model to help evaluate the quality of a new wine base on its characteristic in a statistical way. 2) Use historical data to predict the expert rating of a newly developed wine
Gains
1) Estimate expert rating for a newly produced wine without bringing it to an expert. 2) Save costs on expert rating by pre-selecting wine for quality rating.
Customer Jobs
• Estimate the value of wine using only chemical features
• Estimate anticipated expert rating
Pains
1) Hard to estimate the quality of the wine only using chemical features. 2) Hard to predict how experts will rate a particular wine

**Figure 6.** *Business Value Proposition*

The proposed business values were added in the above graph (Figure 6), we need to better understand the sales team's needs before we give further recommendations. Our main objective is to reduce expert rating cost by reducing the proportion of the wine that are likely not able to receive a high rating from an expert. Looking at the model, it cannot accurately filter out low-quality wine as we expected.

### *5.1 Current Model Limitations & Improvements*

Though it is clear that business value was brought to the sales team with the creation of the model, there are still some limitations and rooms for improvements for the current model.

During the data exploration process, plot_correlation functions from the DataExplorer package was used to generate a correlation matrix (Figure 7). From the plot, some relatively strong correlations were noticed:

- Between variable density and variable alcohol (-0.69)
- Between variable free.sulfur.dioxide and variable total.sulfur.dioxide (0.72)

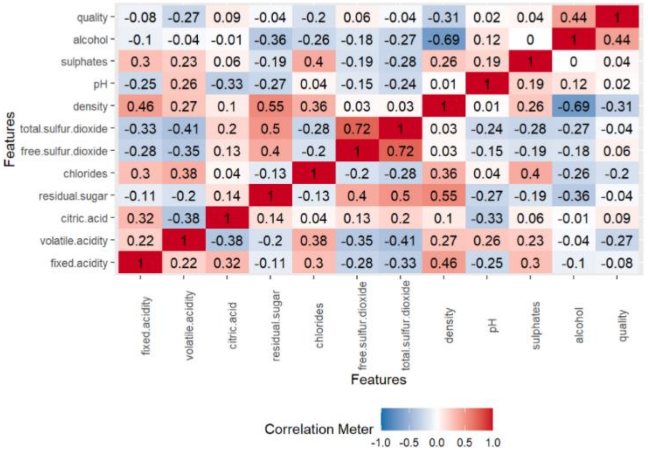


**Figure 7.** *Correlation Matrix*

These strong correlations may indicate possible multicollinearity which is something to take into consideration. The final model has used the variable that is highly correlated meaning that this could be a possible limitation of the model.

In addition, as it showed from the model evaluation, the Adjusted R-Squared being 29.77% can be relatively low. There may be other models such as random forests that can possibly improve the accuracy and these models can be attempted in the future. However, there is always the trade-off between bias and variance that needs to be taken into consideration. Our current model has relatively low variance meaning that the model would not change much when having a different training dataset. The drawback is that the current stepwise model seems to have a relatively high bias meaning that a larger error exists as we performed a simpler model to approximate a complex problem.

To maintain and improve the model, the sales team can continue to add newly collected wine data to the database each time when an expert was brought in for a rating. The model can be modified and retrained with the updated dataset. Some other future improvements of the model include but are not limited to adding interaction terms and higher-level terms to the existing linear model, conducting a PCA for correlated variables to reduce dimensions, or try other types of machine learning models.

## 6. Business Recommendations

After learning about the current model with its limitations and possible improvements, we are able to provide Miami Wine's sales team with some recommendations targetting their business need. As we have discussed in the discussion section, though we have a statistically significant stepwise regression model, it is not very accurate when used to make predictions. We would like to look into why that happened and make some strategic business recommendations based on that.

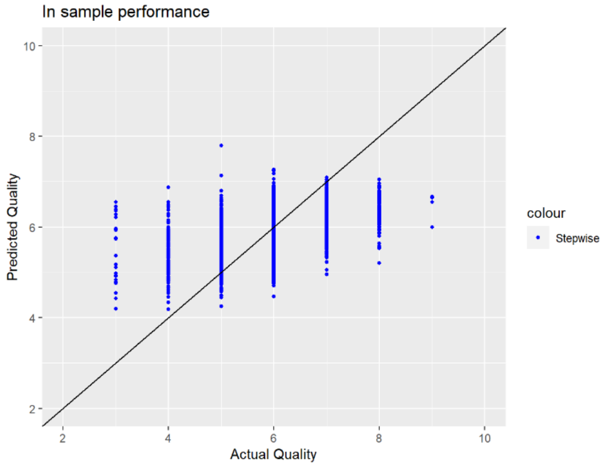


**Figure 8.** *In sample Predicted vs. Actual Scatter plot*

Based on the in-sample performance graphic above (Figure 8), we can see that the model prediction is not well-matched with the actual quality. To be more specific, we can see that the prediction on lower grade wine tends to be relatively high while the predictions for higher grade wine tends to be relatively low. Especially with the actual wine quality of less than 4 or higher than 8, the predictions are completely off. And that is the main reason our overall prediction is lower than what we wanted.

### *5.1 Data & Business Process Limitations*

So what caused the lower and higher end of the prediction to be significantly off from its actual value? There are two possible reasons, leading to two recommendations for Miami Wine's Sales team. The first issue is the limitation of the data we used to build the model. From the histogram below (Figure 9), we can see that the majority of the data we used falls into the middle range where most of the quality was between 5 to 7. This is not very helpful in the model building process as we are more interested in the wine from the lower quality range. The model needs to have more data with lower quality wine to better understand what features caused them to be rated poorly. To improve that, the company can collect more data especially with the ones that have low ratings, and fit the model again using more low rating data. This may be helpful to improve model performance.

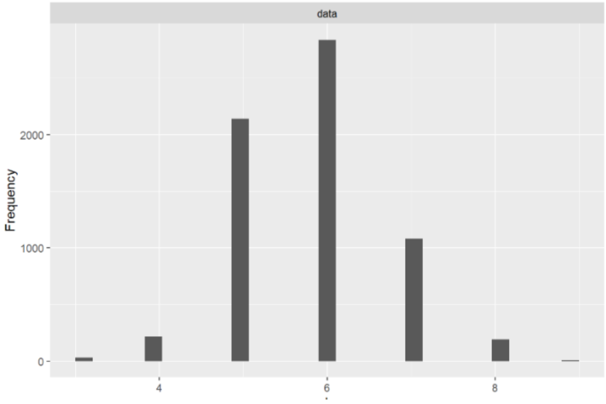


**Figure 9.** *Histogram for Quality Variable*

The other limitation is from the wine rating process itself. Imagine we collected more data like mentioned above and realized that the model was not improved much, and is still not powerful in predicting the wine quality, we need to start thinking about the wine rating process itself. This means that the data we currently have on hand can only provide us with this much information and we can't do any better by adjusting the model. The wine rating process needs to be taken into consideration and find out what other factors impact the ratings of wine. A possible solution for the sales team is to observe the rating process of a wine rating expert. Some levels of interviews may also be conducted to find out why they rate the wine in a certain way. Perhaps it is factors like the smell of the wine, or the color of the liquid, or even whether there is bitterness in the taste. Once the other important factors have been found, these data need to be collected along with wine ratings and new models can be built based on these new features.

### *5.1 Effectively Using the Current Model*

Although it is ideal to find a wine expert and collect more data to make better models, one may ask what we can do at the current stage with the current model. It is possible to make this model relatively helpful depending on our end goal and criteria regardless of the model's poor performance.

Our overall goal is to reduce the amount of wine rated so that the cost of bringing in experts can be reduced. However, we need to be relatively conservative and not eliminate any wine that is actually high-quality wine. The model can help us select the wine to be eliminated for all the predicted values being 5 or lower. Because based on the in-sample performance result (Figure 8), we can see that all the values predicted to be lower than 5 have an actual quality lower than 7. Depending on the criteria of the business, we can change the threshold of elimination as long as we made sure it is lower than 5 because once the predicted value hits 6, some of the actual high-quality wine will be eliminated. The only downside is that a lot of low-quality wine will still be rated higher than 5 by the model and we will end up bring in experts for them making it a bit more costly. However, we are still reducing the cost from rating all wine to cutting out a small portion of fairly low-quality wine to not be rated. And this small reduction meets the objective and solves the sales teams' pain.

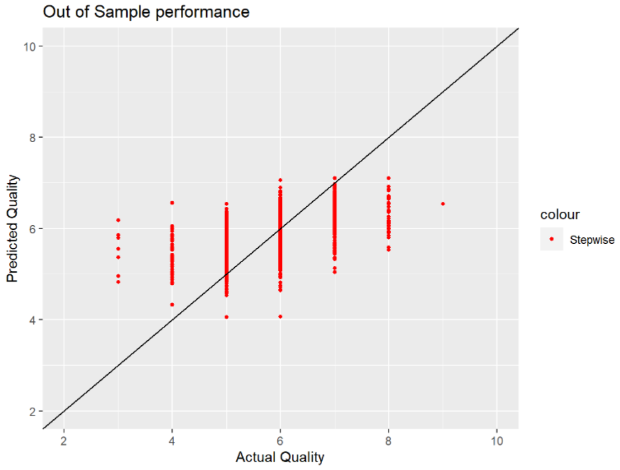


**Figure 10.** *Out of Sample Predicted vs. Actual Scatter plot*

Just to make sure this pattern is reliable, we can double-check on the out of sample data to ensure its performance (Figure 10). We can see that all wine that has a predicted score of 5 or below has, in reality, a quality less than 6 on the validation data. Although we weren't able to eliminate the very low rated wine, the model is helpful to ensure that we are not eliminating any high-quality wine to be rated.

## 6. Conclusion

In this article, a business problem was brought up by the sales team of a wine company in need of a way to estimate wine quality before bringing in an expert. A complete analytics workflow was provided on how the data was pre-processed, analyzed modeled to provide the sales team a complete solution. A multiple linear regression model was presented as the solution and its performance was evaluated using different metrics. The model will improve the business processes and can be implemented by the company to help the sales team with their problem. There are some limitations of the model and some ways for the model to be improved were discussed in the article. Lastly, why the model performs poorly was discussed and some business recommendations were made to help to resolve the issue. One possible way of effectively using the current proposed model disregarding its limitations was also discussed.

## Acknowledgment

The datasets used in this article are obtained from the UCI Machine Learning Repository. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine [8].

## References

[1] Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez

[2] A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009

[3] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

[4] Boxuan Cui (2020). DataExplorer: Automate Data Exploration and Treatment. R package version 0.8.1. https://CRAN.R-project.org/package=DataExplorer

[5] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[6] Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra

[7] Yachen Yan (2016). MLmetrics: Machine Learning Evaluation Metrics. R package version 1.1.1. https://CRAN.R-project.org/package=MLmetrics

[8] Cvrvv. Vinho Verde, www.vinhoverde.pt/en/.