Quality of Wine

Parth Rana

Table of Contents

INTRODUCTION

When it comes to determining the quality of the greatest bottle of red wine, a few scientific factors and a few essential signals can provide an instant response. When it comes to winemaking in general, the obvious element is that the grape comes first. The grape is one-of-a-kind and a vital ingredient in deciding how wine is graded. Unlike white wine, which isn't fermented with grape skins and seeds, red wine gets a lot of its flavor, texture, taste, fragrance, and overall quality from the grape skins, which provide the wine with acidity as it ages. Volatile acidity, residual sugar fixed acidity, citric acid, and total sulfur

When these traits are combined and, in certain circumstances, changed, they may be used to determine the wine's quality. Furthermore, the "outcome" of a very difficult scoring system in which all the mentioned qualities are "inputs" may be used to judge the quality of red wine. In this article, we'll look at how to use these inputs by looking at how they interact with one another. On a scale of one to ten to 10, with 10 being the best, these qualities will be used to predict the quality of red wine in the future.

Methodology

Data exploration, data analysis, and model validation will be the main topics of this study. A five-number summary, one or many histograms, scatterplots, and correlation coefficients will be included in the data exploration section of this report.

A comprehensive model, model selection, a final model, predictions, and outcomes will be included in the analysis section.

The testing and training values will be separated into a format that will forecast the wine's final quality during the model validation step.

In general, the report will begin with a thorough introduction to the data set(s) and a demonstration of certain data visualization techniques. After that, an example of cross-validation will be shown, in which the model's and data's correctness will be assessed. The cross-validation score will be used to decide which wine is of the best quality. Following that, a conclusion will be made. The scoring model will resemble the graph (Figure 1).

Data Exploration:

I have utilized a variety of outputs in the data exploration stage to show an ideal model that fits our data (Figure 1).

Descriptive Statistics:

An example of descriptive statistics is a five-number summary (Figure 2-3) was one method I used to summarize the distribution of red wine quality. I discovered that the mean value for wine quality is 5.649 and the median value for wine quality is 6. Red wine has a minimum rating of 3 and a maximum quality of 8. Also, the interquartile range of wine quality runs from 5 - 6 after analyzing the median 50% of the data.

Histogram

A histogram was the second tool our team utilized to examine the distribution of wine quality. The outcome of the univariate approach we used to produce the histogram is shown in Figures 3 and 4. Because the numbers for the mean, median, and mode of wine quality were all identical, our team concluded that the data set was normally distributed. Figure 4-7 provided substantial evidence that the data set for red wine quality was regularly distributed. Looking at Figures 4-7, we can see that the histogram has an asymmetric bell shape, suggesting that the data is distributed normally. Furthermore, because the data was found to be normally distributed, our team established that no alteration was required for the distribution of wine quality.

Scatterplots

The sgscatter approach was used to visualize and evaluate the relationship between wine quality and the other factors. I discovered that the relationship between wine quality and fixed acidity has a slight positive linear relationship after looking at Figure 8. The relationship between wine quality and volatile acidity is inversely proportional. A positive linear relationship exists between wine quality and citric acid. There is a positive linear relationship between wine quality and residual sugar. The relationship between wine quality and chlorides is inversely proportional. A positive linear relationship exists between wine quality and free sulfur dioxide. A positive linear relationship exists between red wine quality and total sulfur dioxide. The relationship between wine quality and density is inversely proportional. The relationship between wine quality and pH is inversely proportional. A positive linear relationship exists between red wine quality and sulfates. Lastly, there is a positive linear relationship between wine quality and alcohol. Because the regression lines for these correlations are linear, in this case, we should fit our data to a linear regression model.

Pearson Correlation Coefficients:

Following that, I looked at the correlations between red wine quality and the other eleven factors. The Pearson correlation coefficient values were produced using the corr method. These numbers allowed us to evaluate the relationship between red wine quality and the other eleven factors, as well as validate our scatterplot findings. The outcome of the corr process is shown in Figures 9 and 10. Our experts determined that the correlation value between wine quality and fixed acidity is 0.13040 by looking at Figure 9. This correlation coefficient reveals that wine quality and fixed acidity have a weak positive relationship.

Wine quality and volatile acidity have a correlation coefficient of -0.39334. This correlation value suggests that wine quality and volatile acidity have a moderate negative relationship. The association between the quality of wine and citric acid is 0.23246. This correlation value reveals that wine quality and citric acid have a weak positive relationship. Wine quality and residual sugar have a correlation coefficient of 0.03582. This correlation value reveals that wine quality and residual sugar have a very weak positive relationship. The correlation between the quality of wine and chlorides is -0.13234.

This correlation result reveals that wine quality and chlorides have a mild negative relationship. -0.06352 is the correlation value between wine quality and free sulfur dioxide. A very slight negative link exists between red wine quality and free sulfur dioxide, as seen by this correlation value. -0.20153 is the correlation value between wine quality and total sulfur dioxide. A slight negative link exists between wine quality and total sulfur dioxide, as seen by this correlation value. The density of red wine has a correlation value of -0.17307. A slight negative link exists between wine quality and density, as seen by this correlation value.

-0.04782 is the correlation value between red wine quality and pH. This correlation value reveals that wine quality and pH have a very weak negative relationship. Finally, the correlation coefficient between the quality of red wine and its alcohol content is 0.49761. This correlation value suggests a weak positive relationship between the quality of wine and the amount of alcohol consumed. The pearson correlation coefficients were also examined by our team to look for any collinearity between the independent variables. The absence of coefficient values larger than 0.899 indicates that the independent variables are not collinear.

Data Analysis

Full Model:

The initial step in the data analysis stage was to fit a comprehensive model to our data. Our whole model's analysis of parameter and variance estimations. I utilized the /vif option in the reg process to ensure that the control (independent) variables were not multicollinear. Figure 11 indicates that all VIF values were < 10, and the corr technique revealed that all pearson correlation coefficients were < 0.89, thus we concluded that multicollinearity among many of the

independent variables was not an issue. Also, discovered many factors with a p-value > than 0.05. I realized that these characteristics might not be beneficial in forecasting wine quality, therefore I made sure this was taken into account throughout the model selection process. The chart in Figure 11 was also utilized to assess the goodness-of-fit. Since all x-variables equal 0, the null hypothesis is true, but now at least one x-variable is not equal 0, the alternative hypothesis is true. The F-value is 45.42, and the p-value is < than 0.001, as shown in Figure 11. Because the p-value is so low, we can rule out the null hypothesis and conclude there is at most one predictor that is strongly linked to wine quality. Eventually, I calculated that the corrected R squared value is 0.3595 using the data in Figure 11. Our whole model can explain 35.95 percent of the variation in wine quality, according to this corrected R squared value.

Selection Method:

Given in Figure 12 for the next step. debated whether the method of selection would be appropriate for the data set I was going to utilize. The stepwise procedure was chosen as the method of selection. This strategy proved to be the most effective and is also one of the most popular. I discovered that just seven variables should have been included in the final model by using this option. I also found the order of relevance runs from alcohol to volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, and lastly free sulfur_dioxide using the stepwise technique. During this process, the stepwise option determines that all these variables have had no multicollinearity issues and seem to be statistically significant. For our situation, these seven variables are the most interesting and helpful.

Final Model:

Afterward, I decided the final model will include the seven variables stated in the selection technique above. We repeated the regression technique, this time omitting the stepwise option and using only the seven variables. I noticed our adj-R^2 and R^2 were close to identical, both at 38% percent. I can tell the selection approach worked since after running the vif option, I could also see that all of the variables in the model are important but don't have any difficulties with multicollinearity. I also opted to look at the studentized versus projected residuals and the probability value during the regression phase. Also, noticed the normal probability plot appeared to be in good shape and had a slight curve indicating that there were no normality issues (Figure 14) Figure 14 indicates a linear relationship between wine quality and the control variables, indicating that the normality test was not broken. And noticed there were several data over the 3 and -3 limit in the studentized vs projected residuals, indicating there were outliers. The following item will go through this in further detail. Finally, because Figure 13 reveals a scratch a pattern, the requirements of independent and constant variance were broken. Since the quality variable was confined to integers within 1 and 10, I decided there was nothing that can be done about it and chose to leave it alone.

Outlier and Influential Points Removal Process:

In this process, I saw there were a significant number of outliers within data after looking at the residuals as well as the probability plot from the final model. As a result, we employed the impact R option for our regression technique to identify outliers and influential points in our data. Afterward, the data set had an extraordinarily high number of outliers and influential points. The whole elimination method began with just eliminating the observations in the *Cook's D* and Studentized Residual portions of the influence or option that had a pointed head. This was the initial removal step and I was able to delete roughly 9 observations in the first round. The first elimination wave is depicted in detail in Figure 15. Then decided to use the impact R option on the freshly created dataset, which was free of the outliers from the very first round. There were still many outliers and influential spots, but there were far fewer. Then I had deleted a couple more observations from the data during the second round of elimination. This is a positive indicator because it is less than the previous quantity.

Conclusion and Results

I repeated the reg method as a final attempt to construct our final model after correcting for outliers and influential spots. Figure 15 depicts our final model's analysis of variance and also parameter estimations after removing outliers and important points.
Quality = 4.21544 - 1.07937volatile_acidity - 1.60077chlorides - 0.00240total_sulfur_dioxide + 0.93053sulphates + 0.25449alcohol was determined. From this final model, increasing volatile acidity by 1 g/L reduces wine quality by -1.07937. The quality of wine will drop by - 1.60077 if chlorides rise by 1 g/L. The quality of red wine will drop by - 0.00240 if total_sulfur_dioxide levels rise by 1 mg/L. If sulphates rise by 1 g/L, the quality of red wine rises by 1.23919. The quality of wine will rise by 0.25449 if the alcohol concentration is increased by 1%. Starting with alcohol, sulphates, volatile_acidity, total_sulfur_dioxide, chlorides, and free_sulphur_dioxide is by far the most important to least important predictors, as shown in Figure 15. Lastly, 0.3806 is the modified R squared value. Volatile_acidity, chlorides, free_sulfur dioxide, total_sulfur dioxide, sulphates, and alcohol concentration account for 38.38 percent of the variance in wine quality.

Predictions:

I first predicted wine quality using 0.17g/L volatile acidity, 0.15 percent citric acid, 1.14g/L residual sugar, 0.037g/L chlorides, 20mg/L free sulfur dioxide, 65mg/L total sulfur dioxide, 0.88 density, 4.3 pH, 0.77g/L sulphates, and 7.6 percent alcohol, yielding an output statistic of 5.1811 with a 95 percent confidence interval of 4.9524 to 5.4098 and a 95 percent prediction (Figure 16). The anticipated value for this time was 5.0721, with a 95 percent confidence interval of 5.0027 to 5.1415 and a 95 percent prediction interval of 3.8509 to 6.2934. (Figure 16)

Model Validation

Splitting the Dataset:

To do model validation on our data, we opted to divide it into two sets: one is for training and another for testing. We picked this split since we have a large sample size which will enable us to perform model validation on both our training and test sets. We decided to split the data into two sets, with 75 percent of the observations in the training set and 25 percent in the test set. This corresponds to a sampling rate of .75 in our code. Another predictor variable in this dataset is "new_y," which duplicates the "quality" variable's value again for observations that have been chosen for our training dataset. We'll fill the "new_y" variable with data to determine how successful our model predictions are.

Model Selection:

In our training dataset, we selected to run two methods of model selection: Stepwise and CP. Utilizing a variety of analyses, we can evaluate which model will perform best on our data set. Until we can compare the two models, we must first pick that one of the many models developed for each model choice is the best. We can determine that the stepwise model (Figure 17) that utilizes 7 of our 11 cumulative variables provides the best stepwise final model for our training data by simply evaluating R Squared for every model constructed. Alcohol, sulphates, total sulfur dioxide, volatile acidity chlorides, are a few of the eleven variables considered. We established that the model comprising 11 of our 11 total variables provides the best CP final model for the training data set for our CP model selection (Figure 17-18). The R^2 values for stepwise and CP are 0.33 and 0.3236.

Performance Statistics:

We can execute model validation using our sets given we've final models for stepwise and CP model selection. The final models are being used to generate two data sets, “outm1” and “outm2”, which include the predicted dependent variables for observations with absent "new_y" variables. Outm1 will calculate these projected y variables using the stepwise final model, whereas outm2 will use the CP final model. We may output both data sets if we want to examine exactly what projected y variable was generated for each observation in our testing dataset. Therefore, I am primarily concerned with generating performance statistics that will allow us to evaluate the results of both models' overall sets of data, both training, and test. The discrepancy between both the observed y for our training dataset and the predicted y, which is our "new_y" variable, is the initial step. The performance statistics with each model are then calculated. This will provide us the Root mean squared, mean absolute error and r^2 of every model for every test set, that we can use to measure model performance as well as evaluate the training and test sets of each model in each model. Testing Model 1 is the most fitting model in this case. (Figure 19-20)

APPENDIX

Figure 1 - Dataset

dataset of winequality

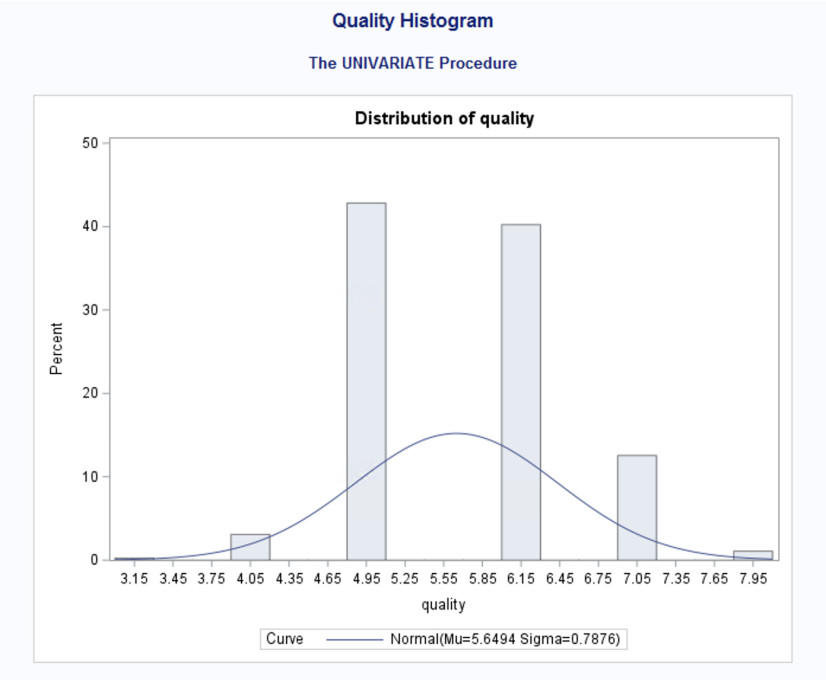
Obs	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
8	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
9	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
10	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
11	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
12	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
13	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
14	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
15	8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
16	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
17	8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
18	8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
19	7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
20	7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
21	8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
22	7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
23	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
24	8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5
25	6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
26	6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
27	7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5
28	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
29	7.1	0.71	0	1.9	0.08	14	35	0.9972	3.47	0.55	9.4	5
30	7.8	0.645	0	2	0.082	8	16	0.9964	3.38	0.59	9.8	6

Figure 2-3 - Descriptive Statistics

Basic Statistical Measures			
Location		Variability	
Mean	5.649433	Std Deviation	0.78755
Median	6.000000	Variance	0.62024
Mode	5.000000	Range	5.00000
		Interquartile Range	1.00000

Quantiles (Definition 5)	
Level	Quantile
100% Max	8
99%	8
95%	7
90%	7
75% Q3	6
50% Median	6
25% Q1	5
10%	5
5%	5
1%	4
0% Min	3

Figure 4-7 - Histogram



Quality Histogram

The UNIVARIATE Procedure
Fitted Normal Distribution for quality

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	5.649433
Std Dev	Sigma	0.787555

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.256744	Pr > D	<0.010
Cramer-von Mises	W-Sq	20.713546	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	115.324974	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	4.00000	3.81731
5.0	5.00000	4.35402
10.0	5.00000	4.64014
25.0	5.00000	5.11823
50.0	6.00000	5.64943
75.0	6.00000	6.18063
90.0	7.00000	6.65872
95.0	7.00000	6.94484
99.0	8.00000	7.48156

Figure 8 - Scatterplot

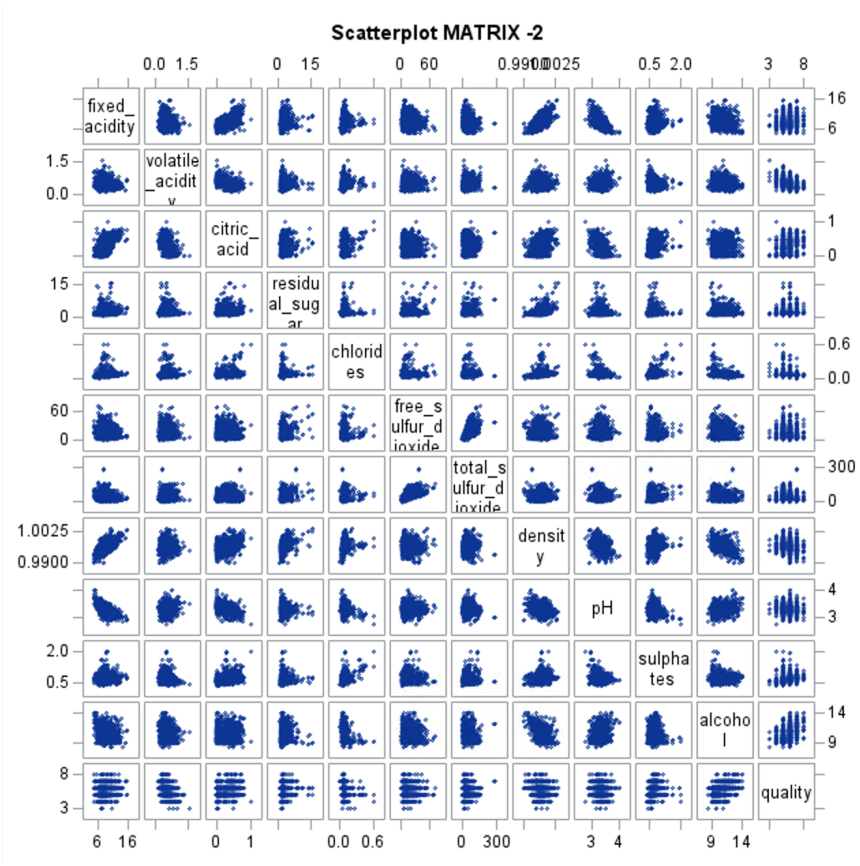


Figure 9-10 - Correlation Tables

Pearson Correlation Coefficients, N = 1586
Prob > |r| under H0: Rho=0

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1.00000	-0.25556 <.0001	0.67080 <.0001	0.11565 <.0001	0.09121 0.0003	-0.16302 <.0001	-0.11523 <.0001	0.66905 <.0001	-0.68355 <.0001	0.18086 <.0001	-0.06660 0.0080	0.13040 <.0001
volatile_acidity	-0.25556 <.0001	1.00000	-0.55061 <.0001	0.00507 0.8402	0.06303 0.0120	-0.00437 0.8618	0.08320 0.0009	0.02183 0.3849	0.22911 <.0001	-0.26066 <.0001	-0.20564 <.0001	-0.39334 <.0001
citric_acid	0.67080 <.0001	-0.55061 <.0001	1.00000	0.14401 <.0001	0.20404 <.0001	-0.06601 0.0085	0.03279 0.1918	0.36408 <.0001	-0.54088 <.0001	0.31246 <.0001	0.11038 <.0001	0.23246 <.0001
residual_sugar	0.11565 <.0001	0.00507 0.8402	0.14401 <.0001	1.00000	0.06122 0.0147	0.20061 <.0001	0.20289 <.0001	0.36654 <.0001	-0.08449 0.0008	0.00967 0.7003	0.02736 0.2762	0.03582 0.1539
chlorides	0.09121 0.0003	0.06303 0.0120	0.20404 <.0001	0.06122 0.0147	1.00000	0.00231 0.9268	0.04901 0.0510	0.19914 <.0001	-0.26652 <.0001	0.37118 <.0001	-0.22137 <.0001	-0.13234 <.0001
free_sulfur_dioxide	-0.16302 <.0001	-0.00437 0.8618	-0.06601 0.0085	0.20061 <.0001	0.00231 0.9268	1.00000	0.67038 <.0001	-0.02497 0.3202	0.07680 0.0022	0.04697 0.0615	-0.06832 0.0065	-0.06352 0.0114
total_sulfur_dioxide	-0.11523 <.0001	0.08320 0.0009	0.03279 0.1918	0.20289 <.0001	0.04901 0.0510	0.67038 <.0001	1.00000	0.07461 0.0029	-0.06501 0.0096	0.04042 0.1076	-0.21307 <.0001	-0.20153 <.0001
density	0.66905 <.0001	0.02183 0.3849	0.36408 <.0001	0.36654 <.0001	0.19914 <.0001	-0.02497 0.3202	0.07461 0.0029	1.00000	-0.34065 <.0001	0.15031 <.0001	-0.49731 <.0001	-0.17307 <.0001
pH	-0.68355 <.0001	0.22911 <.0001	-0.54088 <.0001	-0.08449 0.0008	-0.26652 <.0001	0.07680 0.0022	-0.06501 0.0096	-0.34065 <.0001	1.00000	-0.19588 <.0001	0.20581 <.0001	-0.04782 0.0569
sulphates	0.18086 <.0001	-0.26066 <.0001	0.31246 <.0001	0.00967 0.7003	0.37118 <.0001	0.04697 0.0615	0.04042 0.1076	0.15031 <.0001	-0.19588 <.0001	1.00000	0.09277 0.0002	0.25088 <.0001
alcohol	-0.06660 0.0080	-0.20564 <.0001	0.11038 <.0001	0.02736 0.2762	-0.22137 <.0001	-0.06832 0.0065	-0.21307 <.0001	-0.49731 <.0001	0.20581 <.0001	0.09277 0.0002	1.00000	0.49761 <.0001
quality	0.13040 <.0001	-0.39334 <.0001	0.23246 <.0001	0.03582 0.1539	-0.13234 <.0001	-0.06352 0.0114	-0.20153 <.0001	-0.17307 <.0001	-0.04782 0.0569	0.25088 <.0001	0.49761 <.0001	1.00000

Correlation

The CORR Procedure

12 Variables:	fixed_acidity volatile_acidity citric_acid residual_sugar chlorides free_sulfur_dioxide total_sulfur_dioxide density pH sulphates alcohol quality

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
fixed_acidity	1586	8.32100	1.72864	13197	4.70000	15.60000
volatile_acidity	1586	0.52689	0.17790	835.65500	0.12000	1.58000
citric_acid	1586	0.27120	0.19454	430.13000	0	1.00000
residual_sugar	1586	2.52642	1.37855	4007	0.90000	15.50000
chlorides	1586	0.08754	0.04719	138.83600	0.01200	0.61100
free_sulfur_dioxide	1586	15.92560	10.45951	25258	1.00000	72.00000
total_sulfur_dioxide	1586	46.50189	32.89064	73752	6.00000	289.00000
density	1586	0.99675	0.00188	1581	0.99007	1.00369
pH	1586	3.31037	0.15339	5250	2.74000	4.01000
sulphates	1586	0.65842	0.16981	1044	0.33000	2.00000
alcohol	1586	10.41646	1.05750	16521	8.40000	14.00000
quality	1586	5.64943	0.78755	8960	3.00000	8.00000

Figure 11 - Selection

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	alcohol		1	0.2476	0.2476	333.489	521.30	<.0001
2	volatile_acidity		2	0.0884	0.3360	110.359	210.83	<.0001
3	sulphates		3	0.0185	0.3546	65.1807	45.42	<.0001
4	total_sulfur_dioxide		4	0.0092	0.3638	43.7325	22.89	<.0001
5	chlorides		5	0.0073	0.3711	27.1067	18.38	<.0001
6	pH		6	0.0051	0.3762	16.0039	13.03	0.0003
7	residual_sugar		7	0.0022	0.3784	12.4040	5.58	0.0182
8	free_sulfur_dioxide		8	0.0014	0.3798	10.8907	3.51	0.0612

Figure 12 - Final outliers and points

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	374.12963	41.56996	107.58	<.0001
Error	1576	608.95486	0.38639		
Corrected Total	1585	983.08449			

Root MSE	0.62160	R-Square	0.3806
Dependent Mean	5.64943	Adj R-Sq	0.3770
Coeff Var	11.00296		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	51.63622	17.13240	3.01	0.0026	0	0
fixed_acidity	1	0.07149	0.01755	4.07	<.0001	0.15691	3.77752
volatile_acidity	1	-1.07937	0.11563	-9.33	<.0001	-0.24381	1.73573
citric_acid	1	-0.21826	0.13981	-1.56	0.1187	-0.05391	3.03447
residual_sugar	1	0.04845	0.01394	3.48	0.0005	0.08481	1.51532
chlorides	1	-1.60077	0.39213	-4.08	<.0001	-0.09592	1.40467
total_sulfur_dioxide	1	-0.00240	0.00052154	-4.61	<.0001	-0.10040	1.20702
density	1	-49.24789	17.17579	-2.87	0.0042	-0.11781	4.29528
sulphates	1	0.93053	0.10892	8.54	<.0001	0.20064	1.40344
alcohol	1	0.25449	0.02185	11.64	<.0001	0.34172	2.19101

Figure 13 - Residual plot (Studentzied residual vs predicted value)

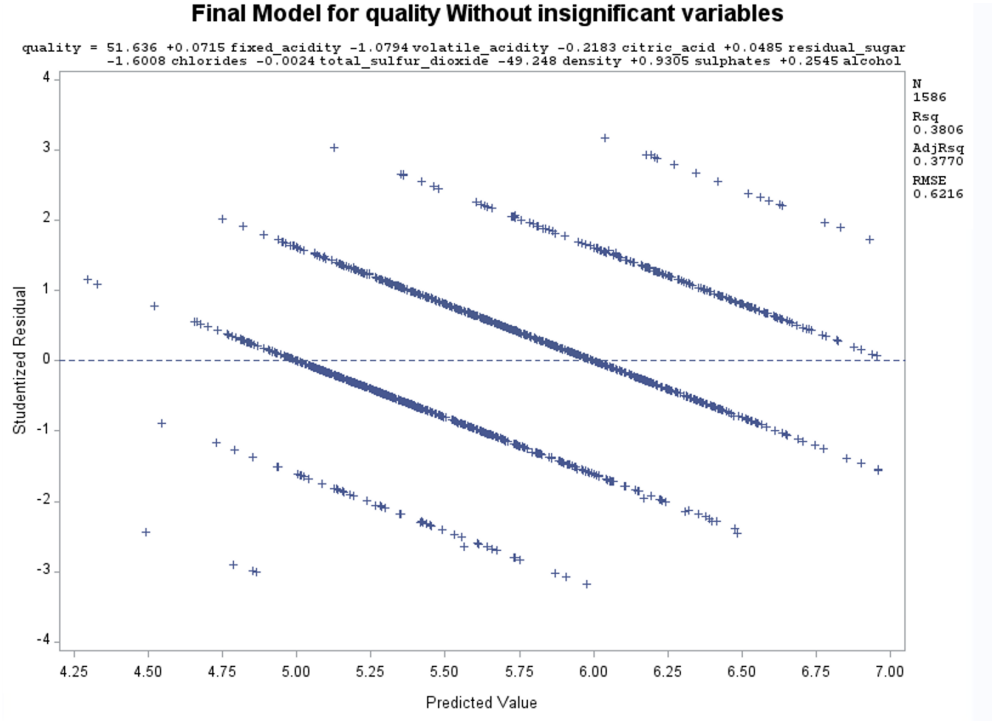
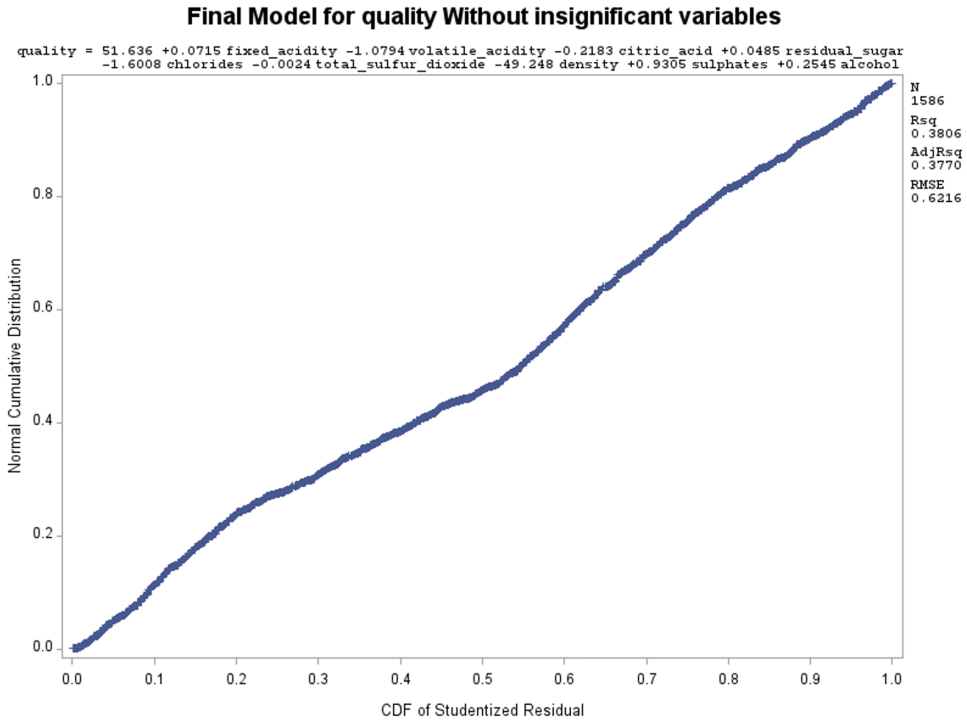
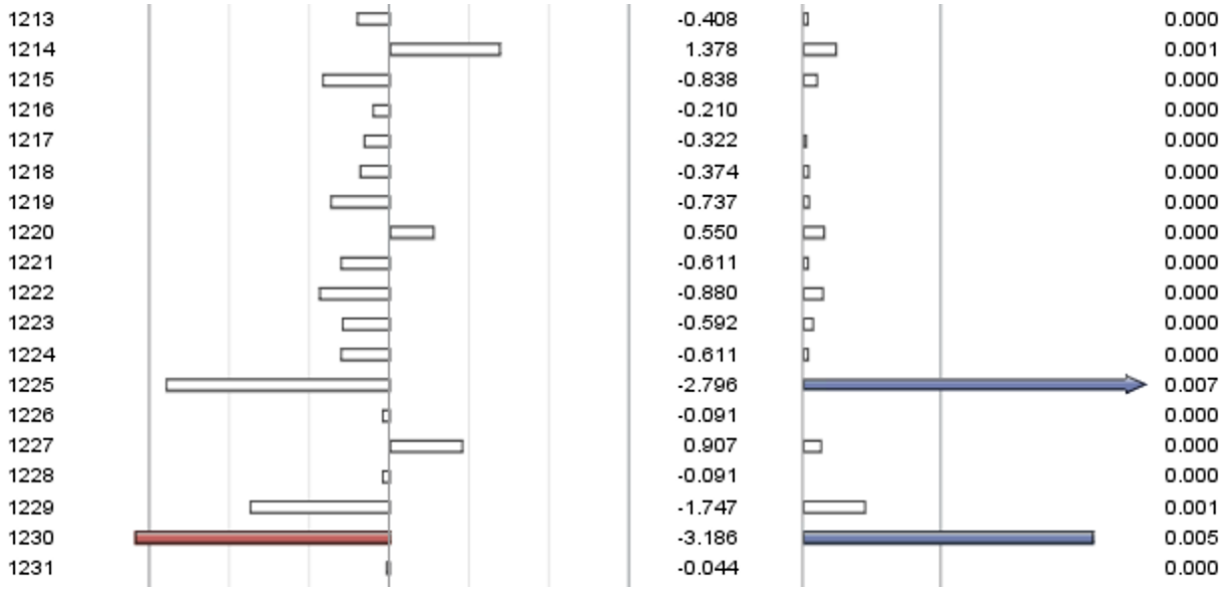


Figure 14 - Normality Plot



Outlier 15 - Removal Process




```
*Removing outliers and Influential points;
data wine_quality;
set wine_quality;
if _n_ in (1506, 1479, 1277, 1236, 900, 833, 814, 691, 653, 460, 440, 391, 46) then delete;
run;
```

Figure 16 Final Model (post outlier / influential points removal)

Output Statistics							
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	5	5.0721	0.0354	5.0027	5.1415	3.8509	6.2934

Figure 17 - Stepwise Model Selection (all steps)

Stepwise Selection: Step 1

Variable alcohol Entered: R-Square = 0.2301 and C(p) = 279.4546

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	167.71191	167.71191	355.11	<.0001
Error	1188	561.07465	0.47229		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.95648	0.19606	47.03109	99.58	<.0001
alcohol	0.35274	0.01872	167.71191	355.11	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable volatile_acidity Entered: R-Square = 0.3303 and C(p) = 90.6945

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	240.74770	120.37385	292.77	<.0001
Error	1187	488.03885	0.41115		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.21517	0.20587	100.28331	243.91	<.0001
volatile_acidity	-1.39502	0.10467	73.03580	177.64	<.0001
alcohol	0.30247	0.01787	117.82172	286.56	<.0001

Bounds on condition number: 1.0466, 4.1865

Stepwise Selection: Step 3

Variable sulphates Entered: R-Square = 0.3510 and C(p) = 53.3275

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	255.82002	85.27334	213.83	<.0001
Error	1186	472.96653	0.39879		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.72764	0.21771	62.59986	156.97	<.0001
volatile_acidity	-1.23669	0.10625	54.02551	135.47	<.0001
sulphates	0.65680	0.10684	15.07232	37.79	<.0001
alcohol	0.29946	0.01760	115.39141	289.35	<.0001

Bounds on condition number: 1.112, 9.6859

Stepwise Selection: Step 4

Variable total_sulfur_dioxide Entered: R-Square = 0.3603 and C(p) = 37.6029

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	262.60620	65.65155	166.88	<.0001
Error	1185	466.18035	0.39340		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.95217	0.22289	69.01606	175.43	<.0001
volatile_acidity	-1.22284	0.10558	52.76959	134.14	<.0001
total_sulfur_dioxide	-0.00230	0.00055439	6.78618	17.25	<.0001
sulphates	0.68293	0.10630	16.23840	41.28	<.0001
alcohol	0.28588	0.01779	101.61795	258.31	<.0001

Bounds on condition number: 1.1131, 17.245

Stepwise Selection: Step 5

Variable chlorides Entered: R-Square = 0.3686 and C(p) = 23.8629

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	268.63253	53.72651	138.24	<.0001
Error	1184	460.15402	0.38864		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.11909	0.22555	74.32020	191.23	<.0001
volatile_acidity	-1.16690	0.10590	47.18744	121.42	<.0001
chlorides	-1.65686	0.42076	6.02633	15.51	<.0001
total_sulfur_dioxide	-0.00234	0.00055113	7.03256	18.10	<.0001
sulphates	0.87969	0.11687	22.01837	56.65	<.0001
alcohol	0.26869	0.01821	84.60382	217.69	<.0001

Bounds on condition number: 1.313, 29.611

Stepwise Selection: Step 6

Variable residual_sugar Entered: R-Square = 0.3722 and C(p) = 18.9804

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	271.26761	45.21127	116.90	<.0001
Error	1183	457.51895	0.38674		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.08126	0.22547	72.22860	186.76	<.0001
volatile_acidity	-1.16250	0.10565	46.81981	121.06	<.0001
residual_sugar	0.03416	0.01309	2.63508	6.81	0.0092
chlorides	-1.73936	0.42092	6.60399	17.08	<.0001
total_sulfur_dioxide	-0.00264	0.00056098	8.53661	22.07	<.0001
sulphates	0.89166	0.11668	22.58707	58.40	<.0001
alcohol	0.26499	0.01822	81.79511	211.50	<.0001

Bounds on condition number: 1.315, 42.172

Stepwise Selection: Step 7

Variable fixed_acidity Entered: R-Square = 0.3746 and C(p) = 16.4219

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	273.01294	39.00185	101.15	<.0001
Error	1182	455.77361	0.38560		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.81577	0.25740	46.14246	119.67	<.0001
fixed_acidity	0.02360	0.01109	1.74533	4.53	0.0336
volatile_acidity	-1.10749	0.10862	40.08604	103.96	<.0001
residual_sugar	0.02935	0.01326	1.88858	4.90	0.0271
chlorides	-1.75720	0.42038	6.73748	17.47	<.0001
total_sulfur_dioxide	-0.00239	0.00057184	6.74071	17.48	<.0001
sulphates	0.86185	0.11734	20.80088	53.94	<.0001
alcohol	0.27102	0.01841	83.53310	216.63	<.0001

Bounds on condition number: 1.334, 58.724

Stepwise Selection: Step 8

Variable density Entered: R-Square = 0.3795 and C(p) = 9.0619

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	276.59656	34.57457	90.30	<.0001
Error	1181	452.18999	0.38289		
Corrected Total	1189	728.78655			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	60.75627	18.94072	3.93968	10.29	0.0014
fixed_acidity	0.06213	0.01676	5.26359	13.75	0.0002
volatile_acidity	-1.04109	0.11039	34.05443	88.94	<.0001
residual_sugar	0.05413	0.01550	4.66975	12.20	0.0005
chlorides	-1.83979	0.41977	7.35512	19.21	<.0001
total_sulfur_dioxide	-0.00250	0.00057104	7.36639	19.24	<.0001
density	-58.07798	18.98392	3.58362	9.36	0.0023
sulphates	0.94113	0.11977	23.64255	61.75	<.0001
alcohol	0.22220	0.02432	31.96608	83.49	<.0001

Bounds on condition number: 4.1514, 123.86

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection							
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	alcohol		1	0.2301	0.2301	279.455	355.11	<.0001
2	volatile_acidity		2	0.1002	0.3303	90.6945	177.64	<.0001
3	sulphates		3	0.0207	0.3510	53.3275	37.79	<.0001
4	total_sulfur_dioxide		4	0.0093	0.3603	37.6029	17.25	<.0001
5	chlorides		5	0.0083	0.3686	23.8629	15.51	<.0001
6	residual_sugar		6	0.0036	0.3722	18.9804	6.81	0.0092
7	fixed_acidity		7	0.0024	0.3746	16.4219	4.53	0.0336
8	density		8	0.0049	0.3795	9.0619	9.36	0.0023

Figure Model - 17-18 Validation (Training and Test Split)

Training and Testing for Quality

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	WINE_QUALITY
Random Number Seed	137287
Sampling Rate	0.75
Sample Size	1190
Selection Probability	0.750315
Sampling Weight	0
Output Data Set	XV_ALL

Obs	Selected	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	new_y
1	1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	5
2	1	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5	5
3	1	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5	5
4	1	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	6
5	1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	5
6	1	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	5
7	0	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5	.
8	1	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7	7
9	0	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7	.
10	0	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5	.
11	1	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5	5
12	1	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5	5
13	1	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5	5

Figure 19 Validation Model 1

Model Selection for Quality Model

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	1586
Number of Observations Used	1190
Number of Observations with Missing Values	396

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	267.00542	38.14363	97.63	<.0001
Error	1182	461.78114	0.39068		
Corrected Total	1189	728.78655			

Root MSE	0.62504	R-Square	0.3664
Dependent Mean	5.63193	Adj R-Sq	0.3626
Coeff Var	11.09818		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	24.36671	16.30715	1.49	0.1354
fixed_acidity	1	0.06107	0.01885	3.24	0.0012
volatile_acidity	1	-1.25351	0.12631	-9.92	<.0001
citric_acid	1	-0.29299	0.15300	-1.91	0.0557
total_sulfur_dioxide	1	-0.00175	0.00057922	-3.03	0.0025
density	1	-21.88369	16.34751	-1.34	0.1809
sulphates	1	0.70307	0.10964	6.41	<.0001
alcohol	1	0.28091	0.02217	12.67	<.0001

Figure 20 Validation Statistics Model 1

Validation stats for Quality model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	396	0.80006	5.70202

Correlation stats for Quality model

The CORR Procedure

1 Variables:	quality

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
quality	396	5.70202	0.80006	2258	3.00000	8.00000

Pearson Correlation Coefficients, N = 396
Prob > |r| under H0: Rho=0

	quality
quality	1.00000