Peter Pappas, ppappas4@depaul.edu

Nikhil Bhalala,nbhalal1@depaul.edu

Diya Malaney, dmalaney@depaul.edu

Maria Alexandra Theodorescu, mtheodo4@depaul.edu

Alec Favia, [afavia2@depaul.edu](mailto:afavia2@depaul.edu)

**On Cloud Wine**

**Wine Density Prediction**

**Introduction:**

Our goal for this analysis is to determine which other physicochemical properties of red wine are most important/significant in predicting wine density (which is going to be our dependent variable). The density of a wine is mainly determined by certain factors, and we want to know which of those are most relevant/significant. The dataset originally included wine quality as well (based on sensory data), but given that is a subjective variable, we decided to remove it as part of our data cleaning/preprocessing and only work with measurable variables.

In order to make sense of our results, it is important to perform some research about our data - the qualities of wine - in particular the type of wine we are analyzing. Generally, density is measured by taking the units of mass and dividing them by volume (g/cm^3). Background research on the density of red wine suggests that its density is generally about 0.99 g/cm³. A variety of acids make up about 0.4% of red wine and, on average, it contains about 12% ethyl alcohol. The dataset description states that the dataset is based on the red variant of a Portugese “Vinho Verde” wine. This type of wine traditionally gained the property of effervescence by being bottled prior to the dissipation of the carbon dioxide produced through the initial fermentation. At present, this is done by injecting CO2 before bottling. The residual sugar in wine making, therefore, is the amount of sugar remaining in the wine after the fermentation process. Amount of residual sugar in dry wine may be as high as 9g/L and the alcohol percentage of red wine typically ranges from 12%-15%. This is validated when we examine our dataset, as we see the variables for those specific features have values that are within these ranges.

**Methodology:**

We obtained our dataset through kaggle: (**https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)**

1. Clean Dataset
   1. Our dataset was overall clean and did not require many adjustments.
   2. The dataset originally included wine quality (based on sensory data), but given that is a subjective variable, we decided to remove it and only work with measurable variables (as mentioned in the introduction above).
2. Import data
   1. We imported our dataset into SAS, and printed it out to ensure it was created properly.
3. Pre-processing steps
   1. Check for categorical vs numerical variables
      1. Our variables were all numerical, so we did not need to create any dummy variables for our analysis.
4. Exploration steps
   1. Histograms, descriptives, etc.
      1. We created a histogram and generated the descriptive statistics in order to tell if the variable of density has a normal distribution and whether it requires any transformations. We also generated scatterplots and coefficient matrices to get a sense of the correlations between each variable. This also helped us determine if multicollinearity may be an issue (it did not end up being an issue).
5. Analysis steps
   1. Fit the full model and check for all assumptions and diagnostics
      1. We fit a model including all of the variables.
         * Multicollinearity
           + To check for multicollinearity, we needed to verify the VIF statistic for each variable. Given that all of the values were below 10, this was not an issue (which validates the previous conclusion we arrived at from the correlation coefficient values).
         * Residual Plots: checking for constant variance, independence, linearity, normality
           + The residual plots indicated that some assumptions are violated. This meant that we needed to transform some of the x variables. We decided to use a square root transformation.
         * P-values
           + P-values for individual predictors show that there are insignificant predictors (with p-value above 0.05) present in the full model.
         * Adj R-sq
           + This indicates how well the full model explains the variation in density. The higher the Adj R-sq, the better the model.
6. Split data into train/test
   1. 75% of our data was selected for training, leaving the remaining 25% for testing
   2. Model selection using training set
      1. We use variable selection methods to select the predictors to remain in our final model using the selected training set
7. Come up with the final model and re-check diagnostics (residuals, outliers/influential points etc.)
   1. We must check the diagnostics for the original and final models (same as explained above)
   2. We remove any outliers/influential points, to improve the model further
8. Test the performance using the test set
   1. We check different metrics in order to test performance
9. Compare performance
   1. We use the appropriate metrics to further compare test and training set performance
10. Compute predictions

**Analysis, Results, and Findings:**

1. **Exploratory Steps**
   1. **Scatterplots & Pearson Correlation Coefficients**

The scatterplots and pearson correlation coefficients of density vs all of the independent variables (**1a & 1b** in appendix) indicate weak correlation values overall. Density has positive correlations with fixed\_acidity, volatile\_acidity, citric\_acid, residual\_sugar, chlorides, total\_sulfur\_dioxide, and sulphates. The remaining correlations are negative. The strongest correlations overall are between density and fixed\_acidity (0.668, positive), and density and alcohol (-0.496, negative). We know that multicollinearity occurs when two or more of the independent variables used in regression are moderately or highly correlated, so checking for this is important. We could potentially see high correlations among the independent variables (i.e., extreme multicollinearity) which increases the likelihood of further errors later on - with the β estimates, standard errors, and so forth. Second, the regression results may be confusing and misleading, we might find that the p-values for both βˆ 1 and βˆ 2 (the least squares estimates) are nonsignificant.

* 1. **Descriptive Statistics/Histogram for Density**

The descriptive statistics for Density (**1c** in appendix) indicate that the average density is ~0.9967 and the 50th percentile, or the median, is also ~0.9967 (meaning the majority of observations have a density of ~0.9967). The minimum density is 0.990 and the maximum is 1.003. For the middle 50, the density ranges from 0.9956 - 0.9978. 75% of our observations have a density of 0.9956 or more.

The histogram (**1d** in appendix) suggests that our data follows a normal distribution, as it looks symmetric, so the dependent variable will most likely not need any transformation. We can tell that the most common density is around 0.9964 and there are several visible potential outliers that will need to be removed later in the analysis.

1. **Analysis Steps**
   1. **Full model assumptions & diagnostics**

Our hypotheses are (as will be referenced):

**Ho**: There is no relationship between the independent variables and the dependent variable.

**Ha:** There is a relationship between density and some x variables.

When fitting the full model with all 10 predictors and density as our dependent variable **(2a.i)**, we get an Adj. R-Square value of 0.8414, meaning that 84.14% of the variation in the density can be explained using the 10 variables in the full model. T-tests on individual parameter values show that all variables except citric acid have a significant effect on density (the p-values are less than 0.05). The p-value for citric\_acid is 0.7319, which is evidently above the threshold, therefore failing the t-test, meaning it is not a significant predictor (so we can’t reject the hypothesis that citric\_acid has no effect on density).

Moreover, to ensure that none of our predictors are collinear, we can use the VIF statistic to check for collinearity. If the value is greater than or equal to 10, the predictor is collinear. Given that all of the VIF values are below 10, it is safe to say that multicollinearity is not an issue.

Another relevant set of values that we can look at are the standardized coefficients. The standardized estimates will indicate the influence of each independent variable on the dependent variable (density). When examining the absolute values of these estimates for our full model, we see that fixed\_acidity has the strongest influence (~0.86), then alcohol (~ -0.522), and third pH (0.414). These will also be re-checked for the final model, but we can already get a sense of which predictors are strongest in terms of influencing our model the most at this point in the analysis.

We generate our residual plots **(2a.ii)** and check for the model assumptions, which will help us determine if there are any issues in the regression analysis and whether there is an important variation in Y that is not explained by the model. If the model assumptions hold, we expect a random scattering of the points across the horizontal axis. If the points have a specific shape, this is a problem we need to correct. In the case that the assumptions are violated, we must correct either outliers, the variables themselves through transformations, or adjust the model technique.

The first assumption to verify is constant variance, or homoscedasticity. For this, we can check the residuals vs predicted values and the residuals vs each x-variable (what we hope to see is a constant spread).

**Constant Variance**: The assumption of homoscedasticity is satisfied for residual plots 1-3, 5, 9. Residual plots 4, 6, 7, 8, 11-12 seem to have more of a funnel shape, therefore are appearing to violate it.

Next, we check for independence. For this, we will need to examine the same plots as we did for constant variance. A random scattering of the observations across the zero line means the assumption is not violated - any visible pattern suggests an issue.

**Independence**: The observations for plots 1-3, 5, 9 seem relatively randomly scattered along the ‘0 line’, but that does not seem to be the case for plots 4, 6, 7, 8, 11-12 , where they are clearly more concentrated in certain areas.

We then check the scatterplots for each x variable and the plot residuals vs each x-variable, to determine if linearity is a problem. Here, we inspect for a linear association.

**Linearity**: Linearity seems to be more satisfied for some of the x variables (ex: fixed\_acidity and alcohol), but not for others (ex: chlorides and total\_sulfur dioxide).

Lastly, the fourth model assumption that needs to be verified is normality. Here, we check the normal probability plot of the residuals. A good model should have a diagonal line (not necessarily perfect), or a 45 degree angle. A curvy shape indicates violation of normality.

**Normality**: Plot 10 shows a nearly straight diagonal line (at a 45 degree angle), therefore does not violate the assumption of normality.

The last step is to determine if there are any visible outliers by looking at all the residual plots except for the normal probability plot. The residual plots indicate that there are several outliers that will need to be removed in the subsequent steps (points that fall outside of the 3 and -3 range). Given that some of our model assumptions are violated, we will need to use transformations as we work our way towards determining a reliable and effective final model.

* 1. **Important Criteria and Methods for Assessing and Choosing our Model**

As we know, the most popular methods out there for defining the best model include optimizing coefficient of determination r2 or Adj-R2, optimizing Mallows’ cp statistics, and using different selection methods (stepwise, backward and forward).

R-square and AdjR-square is variability of Y explained by the model. AdjR-square is a better measure, so selecting the model that has the highest AdjR-square is the best fit model. Optimize Mallows’ cp statistics is to compare the predictive ability of the subset model to that of the full model. The method selects the model that has the cp=p where p=k+1. The most important benefit of using cp is it allows you to select model size- getting a good model that contains as few variables as possible.

On the other hand of selection methods, stepwise is a modification of the forward-selection method. This particular method selects the best model by navigating through the dynamic data, and selecting the models that have the highest R-square or AdjR-square, low RMSE or MSE and all predictors are significant. Then, going to the other side of the spectrum, backward is the opposite of forward selection which begins with all x-variables in the model. The method selects a best model by having highest R-square or AdjR-square, low RMSE or MSE and all predictors are significant. Forward begins with no x-variables in the model it adds x- variable to the model if the statistic for a x-variable to be added is significant(e.g. at a=0.05). The method selects a best model by having the highest R-square or AdjR-square, low RMSE or MSE and all predictors are significant.

1. **Determining the Final Model**
   1. **Transformations**

After seeing that the residual plots show points that are not randomly scattered or the normal probability plot shows “ S” shape- there are various solutions that the team can implement into our study. What the team found to be the easiest solution is to apply transformation on the response variable Y to stabilize the variance. Most common transformations are log(Y)(only if Y>0) ,Sqrt(Y)(only if Y>=0) ,Square Y=Y^2 , cubic Y=Y^3 and Inverse Y=1/Y(only for y is not = 0). Transforming response and/or predictor variables can potentially remedy a number of model problems - we can transform the response (y) value only and/or we can transform the predictor (X) values.

In part three section a, when analyzing the histogram that is displayed, this determines if the independent variable density needs to be transformed. As computed earlier in the histogram, this visual representation of the distribution of density over the data shows a symmetrical relationship. This concludes that the Y variable does not need to be transformed.

Getting more particular and analyzing the associations between density and the predictors, the residual plots present that there are a few x variables that violate the linearity, which may improve with a transformation (square root). These variables are residual sugar, chlorides, total sulfur dioxide, and sulphates. We created dummy variables through applying the square root transformation so as to improve the x variables and stabilize the variance. We created “sqrtSulphates” for sulphates, “sqrtVolatile” for volatile\_acidity, “sqrtRsugar” for residual\_sugar, “sqrtTotalsulfur” for total\_sulfur\_dioxide, “sqrtFreesulfur” for free\_sulfur\_dioxide” and “sqrtChl” for chlorides. We used these variables for the rest of the analysis. Our next step is to run a proc print to see the new transformed dataset. Following the proc print of the new dataset, we created scatterplots (Figure 3a) in order to determine if transforming the variables was useful in improving the x variables. From the scatterplots we can see that by transforming the variables, there was a positive effect as the scatterplots have improved for most of the variables.

**Figure 3ai.**

As we see, after transformation of the model the R-Square increases by 2% ( 0.86 which is 86%)and the Adj-R-square value increases by 1% (0.8591 which is 85.91%) . RMSE value slightly reduced( which is 0.00070838). The F-value has increased (which is 975.54) indicating a stronger support to this model. The F-test p-value is equal to 0.0001. and all predictors are significant except citric\_acid ( which is 0.7888). Overall there is some improvement in the model.

As we can see, after transformation of the model, the absolute value of the standardized coefficient most of fixed\_acidity (the most influential predictor) is 0.84436. Ph follows with a coefficient value of 0.40389, sqrtRsugar with 0.32575, then sqrtSulphates which has a value of 0.11962.

* 1. **Training/Testing and Model Selection**

For model validation we used training and testing sets. We used this test because it shows how well the model predicts new data as well as making sure our model is usable. For part three section b, when training and testing the Wine Quality data, 75% of the data was selected for training in order to get an accurate representation and the other 25% was used for the testing set. Using this we were able to have a total of 1200 observations. We used proc surveyselect to split the date into the test and training sets. For the model selection, there were two selection methods that were used, the stepwise and cp methods. Both model selections resulted in having nine predictors which were determining the total density in wine. The p-values for both models were well below ɑ= 0.05, which showed that all predictors are significant through these selection methods. The f-value for the stepwise method was 708.25 while the f-value for the cp selection method was at 722.15. Then, the cp selection method had a slightly bigger Adjusted R-Squared at 0.8452, while the stepwise Adjusted R-Square was at 0.8427. The computed Adjusted R-Squared identifies the predictors fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, ph levels, sulphates, and the amount of alcohol have a 84.52% influence on the density in wine.

1. **Outliers and Influential Points** 
   1. **Outliers**

Looking at the residual plots, we knew we had many outliers we had to remove in order to perfect our model. Outliers are any observation that is different from the majority of the other cases and isn't explained well by the model. An influential point is an observation that has excessive influence on the fit of the regression model. They can appear for different reasons such as, valid natural variation, measurement error which is the difference between a measured value of quantity and its true value, incorrect selection of a sample, and data entry errors that may reflect coding errors in the dataset that would cause a result of model misspecification. They can distort estimates of regression coefficients and model accuracy. Influential points will also always change your regression line whereas outliers may or may not.

* 1. **Influential Points**

After the influential points and outliers are removed the model will significantly change the statistical measure (e.g. r ,parameter estimates). To remove outliers during the data exploration stage – box plots and histograms, percentile table, extreme value, and multivariate scatter plots will help to remove outliers from the model. In the data analysis stage a residual that is larger than 3 (in absolute value)—or, equivalently, a standardized residual that is larger than 3 (in absolute value)—is considered to be an outlier and normal probability plot. To remove influential points from stats- at the data analysis stage- leverage statistics, COV Ratio, dffits, dfbetas and cook’s distance ,D.

* 1. **Model After Diagnostics**

At this stage, the model is:   
new\_y = 0.9758 + 0.0023sqrtSulphates + 0.001 sqrtVolatile + 0.0018 sqrtRsugar - 0.0001 sqrtFreesulfur + 507E-7 sqrtTotalsulfur + 0.0024 sqrtChl + 0.0009 fixed\_acidity + 0.0051 pH - 0.0009 alcohol

Next, we proceed to check on outliers and influential points (figure 4). We do a final model analysis with options r, influence, vif and stb. We remove 13 outliers and influential points from the dataset. There are a total of 1586 observations left after removing the outliers and influential points. After removing the outliers and influential points we rerun the model analysis. Our final model has improved significantly. The adjusted Rsquare has increased from 0.8427 to 0.8712 which means that at this stage, 87.12% of the variance in our dependent variable has been captured and explained. If the regression explains the variation in Y, the F statistic should be large. The F value has also increased, it has gone from 848.59 to 894.50. The final model after removing the outliers and influential points is:

new\_y = 0.9763 + 0.0023sqrtSulphates + 0.0009 sqrtVolatile + 0.0018 sqrtRsugar - 0.0001 sqrtFreesulfur + 401E-7 sqrtTotalsulfur + 0.0022 sqrtChl + 0.0009 fixed\_acidity + 0.005 pH - 0.0009 alcohol

The independent variables “sqrtSulphates”, “sqrtVolatile”, “sqrtRsugar”, “sqrtTotalsulfur”, “sqrtChl”, “fixed\_acidity” and “pH” are positively associated while the variables “sqrtFreesulfur” and “alcohol” are negatively associated.

After removing influential and outlier points we found a very solid model. As we can see, after removing influential and outlier points, the absolute value of the standardized coefficient most of fixed\_acidity (the most influential predictor) is 0.84516. Ph follows with a coefficient value of 0.40222, sqrtRsugar with 0.32834, then sqrtSulphates which has a value of 0.12232.

1. **Testing the performance of the final model**

To verify if this model is good, we needed to check the linear association and model assumptions satisfied first as we have done above. Second, we need to compare so that we check significance on the y variable. We have done that and removed one independent variable. Finally, we have dropped 13 outliers and influential points so that we can improve F-value, decrease the RMSE, and increase adj-R2 by 1%.

We used several techniques for checking model adequacy, (eg. tests of overall model adequacy, partial f-test and R-squared) we do this because checking model adequacy involves determining whether the regression model adequately fits the sample data. Model validation, however, involves assessing how the fitted regression model will perform in practice — how successful it will be when applied to new or future data. One of the most effective ways of validating a regression model is to use the model to predict the y for a new sample. By directly comparing the predicted values to the observed values of the new data, we can determine the accuracy of the predictions and use this information to assess how well the model performs in practice.

For our data, we had come up with a final model computed for the training set and predicted the values for test sets. We got predicted values for the missing new\_y in the test set for the final model. After we set in out=outm2 which defines datasets containing final model predicted values for the test set. So for the missing values, we get predicted y for the test set for the final model with p=yhat. We compute predicted y values for the test set, where (new\_y= .). The model performance on the test set is then computed based on y\_hat, which is our predicted value.

We computed performance stats RMSE, MAE and R-squared for the test set for the final model, to summarize the results of the validations for the final model. In our case, we use ‘d’ as the difference between observed and predicted in the test set difference for the final model. Then, we compute predictive statistics: root mean square error (RMSE) and mean absolute error (MAE). We can see in the above table RMSE is 0 .00066 and MAE is 0.00050 (as pictured in 4b). After we compute the correlation of observed and predicted values in the test set for the final model, we look at the R for the test set (use this to compute R2) and then CV-R2.

Now, we compare training and testing performance of the final model. For training, we have RMSE is 0.00067 and r-square is 0.8723 (which is 87.23% variance explanation) and adj-R-square is 0.8716, which is 87.16% (figure 4a). Testing RMSE is 0.00066 and MAE is 0.00050 and r-square is 0.93365^2, which is 0.8717 or 87.17%.

Overall, we have noticed testing is better than training with RMSE(.00067 >.00066 ), and r-square is slightly better in the training by using cross validated r-square (.8723 - .8717 = .0006 in favor of training). We tried to determine an optimized model by fixing some issues, therefore this is the best model we came up with if we want to use linear regression on the data.

1. **Conclusions**

Initially the dataset had a variable called “wine quality” and we thought that this would be our dependent variable, but because it was a subjective variable we did not feel comfortable using it. We decided to use density as our dependent variable. First, we created scatterplots with density against all other variables. Density ended up having a positive correlation with fixed\_acidity, volatile\_acidity, citric\_acid, residual\_sugar, chlorides, total\_sulfur\_dioxide, and sulphates. From there we worked to check for multicollinearity and found that there was none. After that we wanted to check on the descriptive statistics and histogram of density and found that we did not need to do any transformations because of its normal distribution. After this we came up with our hypothesis: Ho: There is no relationship between the independent variables and the dependent variable. Ha: There is a relationship between density and some x variables. When fitting the full model with all ten variables we initially got a Adj. R-Square of .8414. We then ran a t-test and found that citric acid had a significant effect on density and p-value was .7319 so we decided to remove it from our model because it failed the t-test. Then we checked for collinearity using the VIF statistic and found that multicollinearity would not be an issue. After computing those statistics, we generated our residual plots to check for shape and outliers to see if we needed to do any transformations. After these steps were completed, we verified constant variance, checked for independence, linearity, and verified normality. Next we determined r-squared, Adj-r-squared, low RMSE, MSE, and Mallows’ cp statistics using the stepwise and cp selection methods. Then we used residual plots and checked to see if there was an “S” shape to it if so we needed to use transformation on Y to stabilize the variance. In doing so we determined that we should use squrt(Y) to transform our data. Next we need to validate our model and used a training and testing set in order to do so.

We knew from an analysis of our residual plots that we had many outliers and influential points that needed to be removed as they distort estimates of regression coefficients and model accuracy. In the data exploration state we used boxplots and histograms to remove outliers, points above +3 or -3 are considered outliers and have to be removed to improve the quality of the model. In the data analysis stage we did a final model analysis with options r, influence, vif and stb. In doing so, we removed 13 outliers and influential points from the dataset. Our adjusted R Square improved and at this stage, and 87.12% of the variance in our dependent variable has been captured and explained. We found that the standardized coefficient fixed\_acidity is the most influential predictor of our model. Next, we needed to test the performance of the final model. We checked the linear association and model assumptions were satisfied before the comparing so that we checked their significance on the y variable.

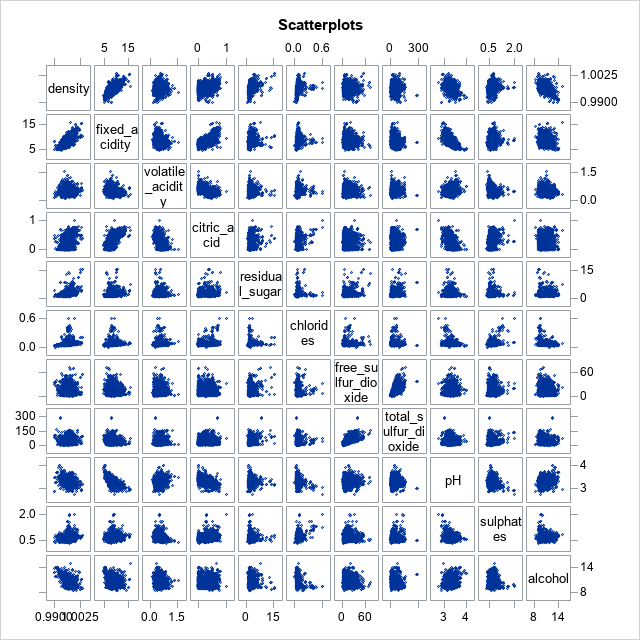
We used several techniques to test our model accuracy, we next also had to check our model validation which focuses on how well the fitted regression model will perform in practice and how it will be when applied to new or future data. We did this by using our model to predict the y for a new sample. This is effective because it directly compares the observed values to the predicted values and we are able to assess the accuracy of the model in practice. We computed the predicted y values for the test set where (new\_y= .). We computed our model performance on the test set based on y\_hat which is our predicted value. We ran performance stats RMSE, MAE and R-squared for the test set and then we compared the training and testing performance for the final model.

Through our results we determined that testing is better than training as it has a higher RMSE and the R-square is slightly better in the train when using cross validated R-square.

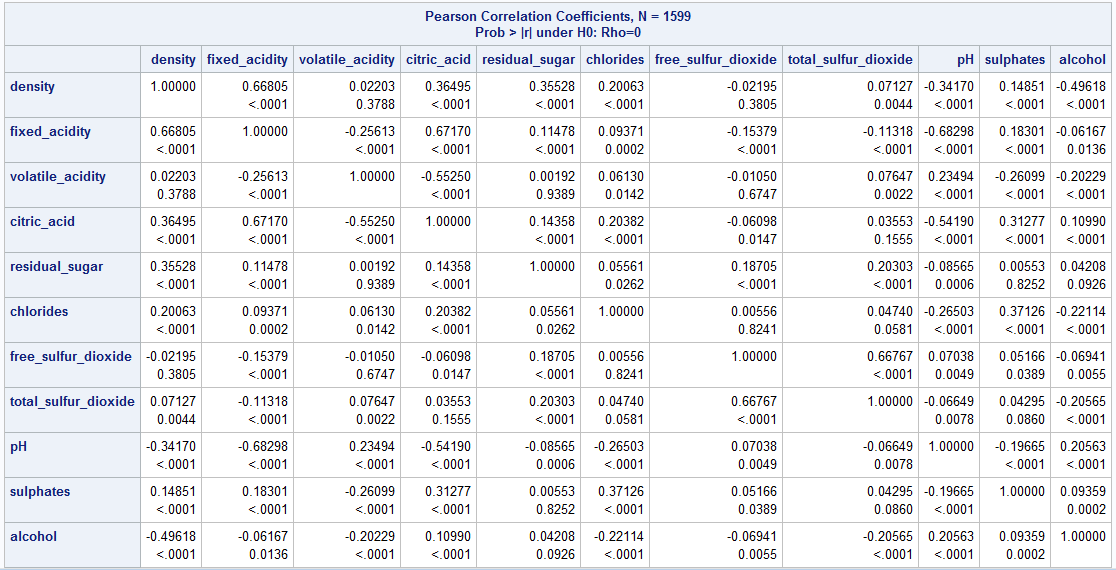
Through our various tests and methods to optimize the model, we were able to determine that this is the best model if we want to use linear regression on the wine data.

**Appendix:**

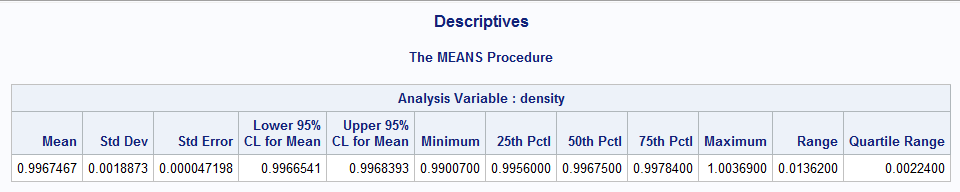
1. **Exploratory Steps**
   1. **Scatterplots**

****

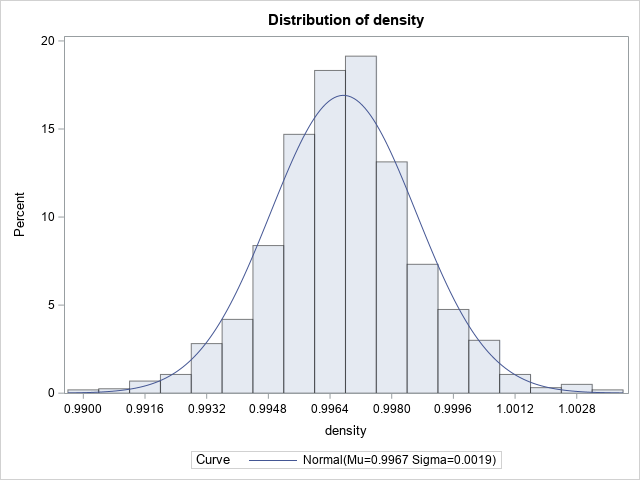
* 1. **Pearson Correlation Values**



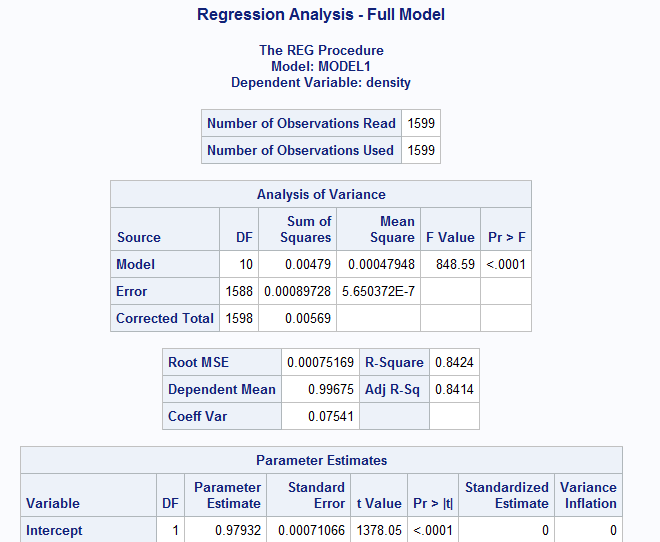
* 1. **Descriptive Statistics for Density**

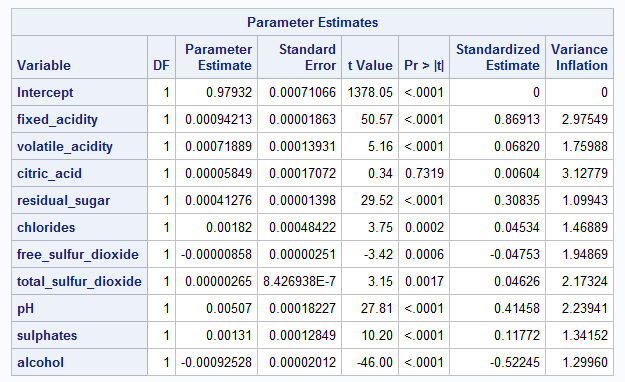
****

* 1. **Histogram for Density**

****

1. **Analysis Steps**
   1. **Fit Full Model/Checking Assumptions & Diagnostics**
      1. **Model**

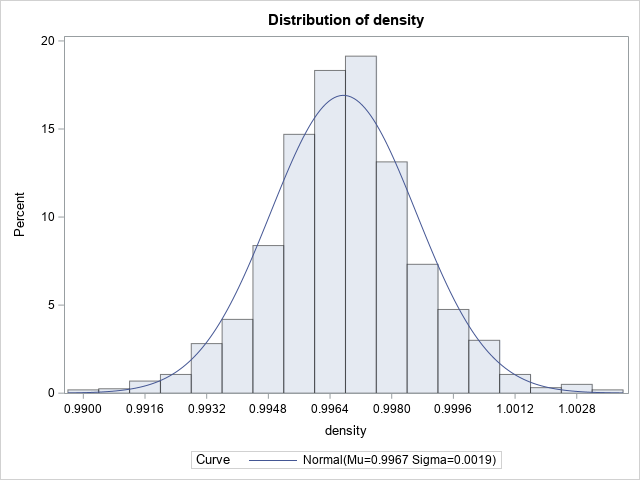
****

****

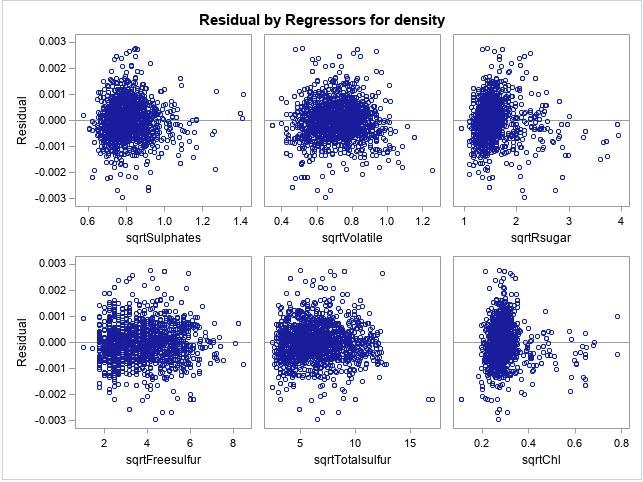
* + 1. **Residual Plots**

| **1** | **2** | **3** |
| --- | --- | --- |
| **4** | **5** | **6** |
| **7** | **8** | **9** |
| **10** | **11** | **12** |

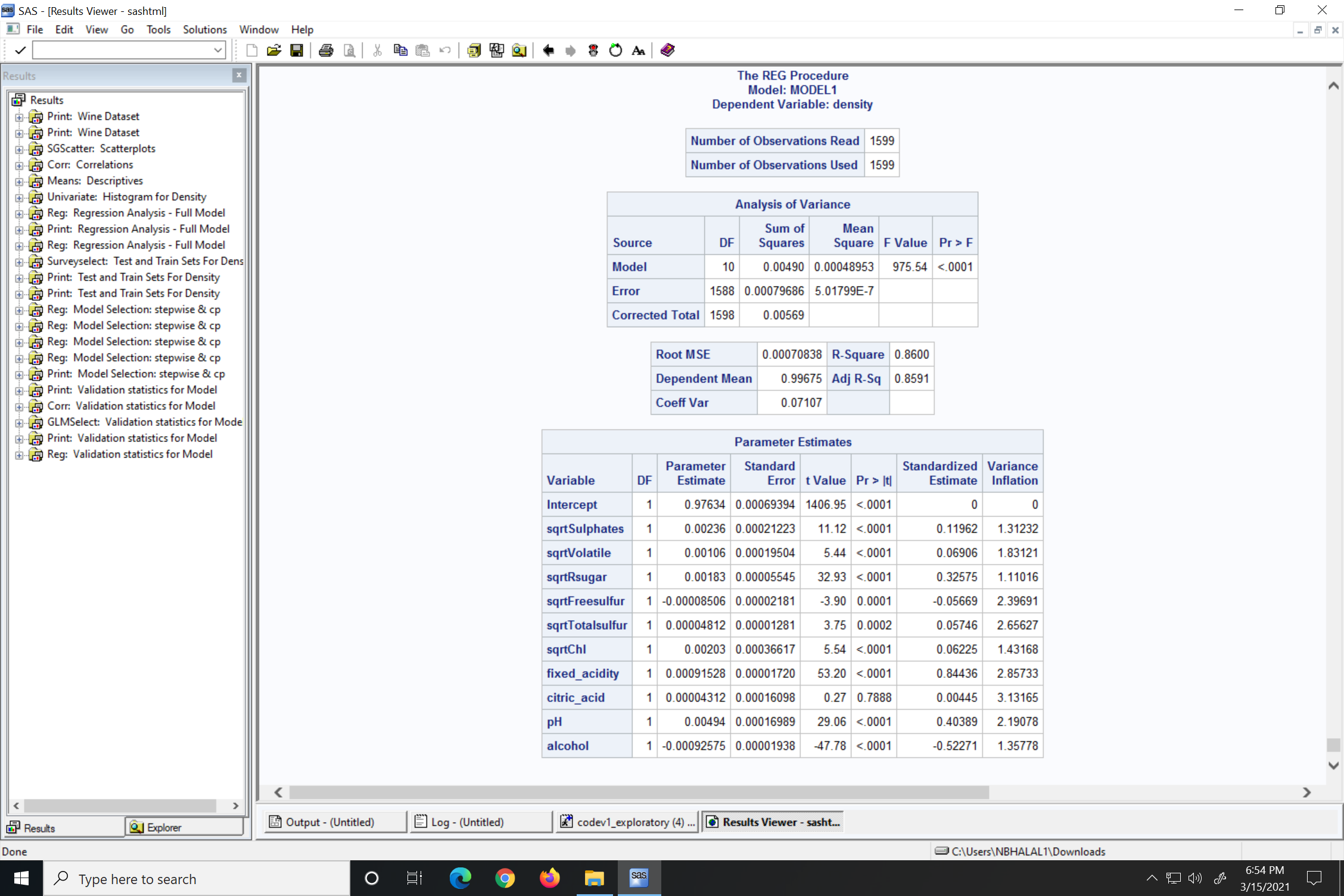
1. **Determining the Final Model** 
   1. **Transformations**

****

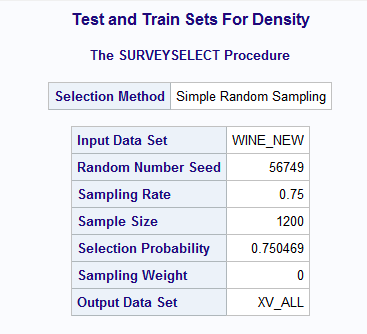
**Figure 3a   
(4,6,7,8,11,12)**

****

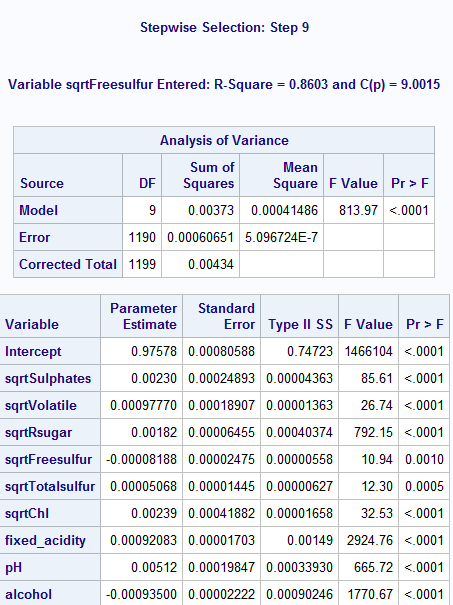
**Figure 3ai.**



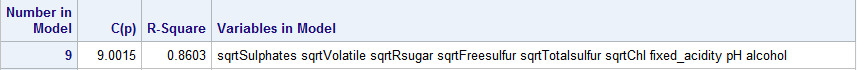
* 1. Training and Testing

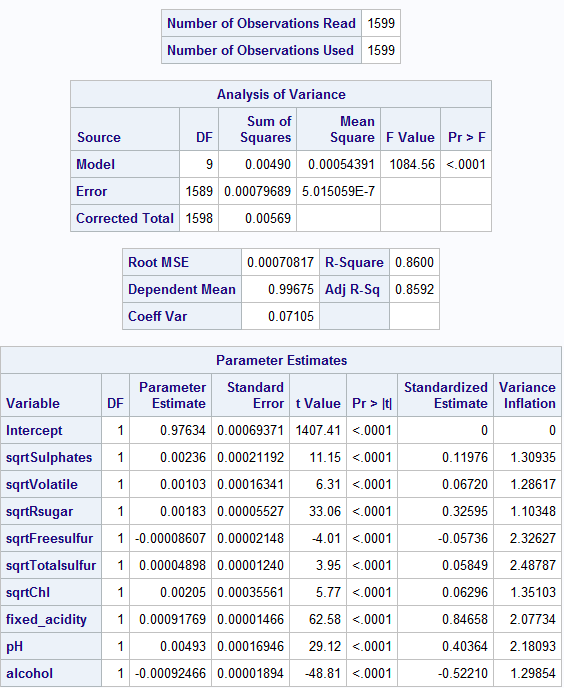


**Model 1:**

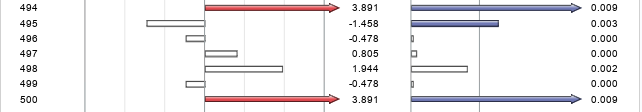


**Model 2:**

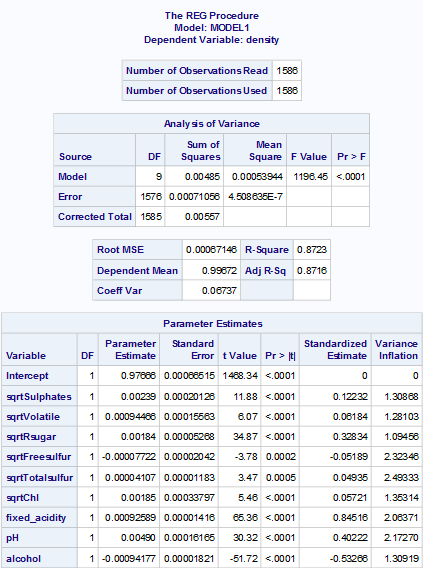


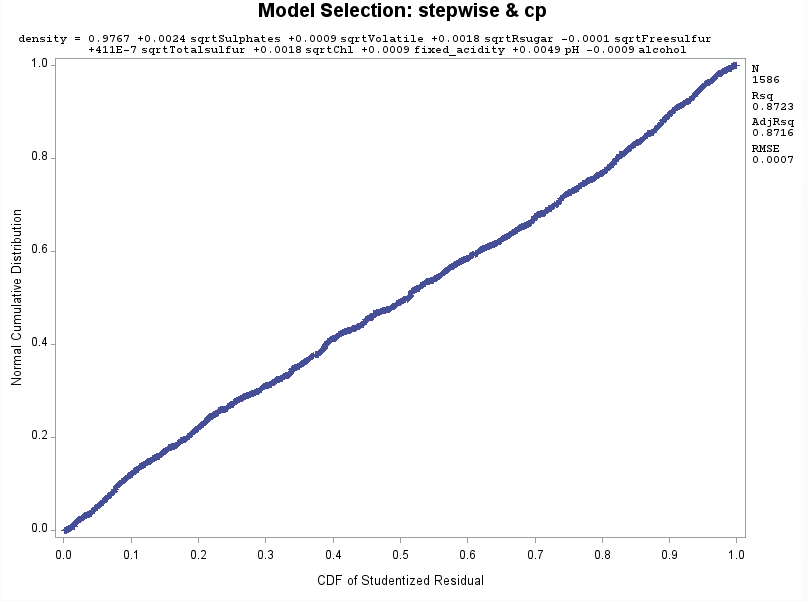


1. **Outliers and Influential Points**



**4a. Refitted model after removing problematic observations**





**4b. Validating the model**



