Determining Wine Quality based on Different Factors

Winemaking also referred to as Vinification is the manufacture of Wine. Fruits are fermented into alcohol to produce wine. Quality is a major factor that determines the price of wine. There are many factors that determine the wine Quality like Balance of alcohol, residual sugar, pH and chemical composition. In addition, factors such as plant Environment, Climate, soil composition and Sunlight play a major role.

The dataset is available for public use at UCI machine learning repository. The dataset is related to Portuguese wine both white and red type. The Dataset describes the physicochemical inputs to different wine samples. The Quality of Wine is determined by Experts and is ranked from 0 to 10. The corresponding physicochemical data have been recorded. The Goal is to find which variable has a major impact on the quality of wine. External factors which have been proved to have impact on quality and therefore not considered. The Data list factors such as fixed Acidity, citric acid, residual sugar, chlorides, Sulphur dioxide content, density, pH, sulphates and Alcohol.

Relationship between variables can be found by regression analysis. A scatter plot of the dependent and the independent variable gives a trendline. The analysis generates a regression equation for the independent variable and dependent variable. The fit of the equation can be verified by co-efficient of determination. A positive slope denotes positive linear relationship and negative slope determines a negative relationship and a slope of zero denotes no correlation. From the analysis it can be observed that alcohol content shows slightly positive meaning an increase in density will lead to decrease in quality of wine with a co-efficient of determination of 0.0258 which can be interpreted as 2.5% of variance can be predicted from Alcohol. pH value shows negative correlation meaning with increase in acidic nature of wine there is a slight increase in quality and co-efficient of determination is 0.0011 which means the pH value do not actually make much difference in quality of wine. A multiple regression which can be used to predict the dependent variable based on multiple variable can be utilized.

February 1, 2019

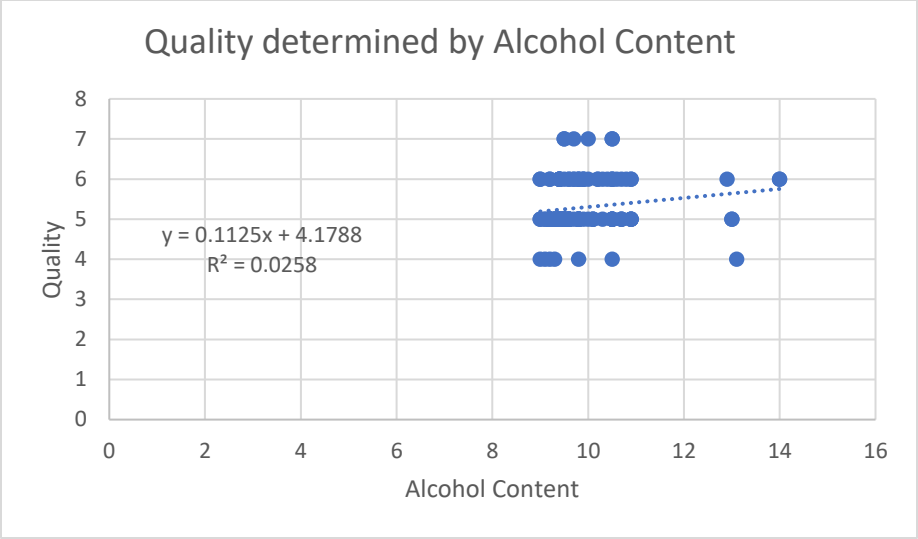

Fig.1: Linear Regression of Quality based on Alcohol Content.

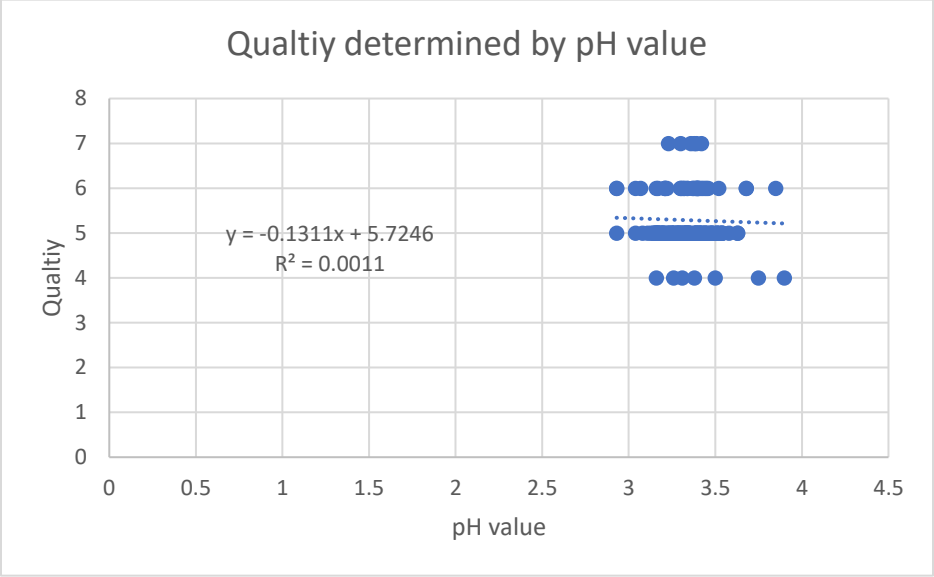

Fig.2: Linear Regression of Quality based on pH value.

A multiple Regression analysis on quality of product as dependent variable and pH and Alcohol as an independent variable can be plotted. Using the analysis, a regression equation can be generated.

$$y = 5.705 - 0.63224(pH\ value) + 0.1713(Alcohol\ Content)$$

The Equation can be interpreted as: The Quality increases by a point with every 0.6 decrease in pH value or Quality increases by a point with every 0.6 points move in the pH scale towards Acidity with Alcohol content being constant. It can also be interpreted as Quality increases by a point with 0.17 increase in concentration with pH value being constant.

The Equation can be tested for significance by using t test. T test is used to check if variable is significant. We can postulate two hypothesis, null hypothesis and alternate hypothesis. T test is conducted by comparing p value with 0.05. The p value of alcohol content is 0.011 and p value of pH content is 0.09. Therefore, alcohol content variable is significant in determining the quality of wine and pH value is not significant in determining the quality of wine. F test can also be conducted with the same hypothesis and comparing Significance F value with 0.05 significance. It can be observed that Significance F is 0.037 which is less than 0.05 which proves the overall significance of the regression equation.

Multiple Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple F	0.21153
R Square	0.044745
Adjusted I	0.031385
Standard E	0.611177
Observati	146

ANOVA

	df	SS	MS	F	Significance F
Regressio	2	2.50203	1.251015	3.349107	0.037891
Residual	143	53.41578	0.373537		
Total	145	55.91781			

	Coefficier	Standard E	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	5.705771	1.069655	5.334216	3.67E-07	3.591392	7.82015	3.591392	7.82015
pH	-0.63224	0.375782	-1.68246	0.094662	-1.37504	0.110567	-1.37504	0.110567
alcohol	0.171381	0.067062	2.555571	0.011646	0.038821	0.303942	0.038821	0.303942

Fig.3: Summary output of multiple Regression

Residual plots can be used to identify outliers and to check the relationship between dependent and independent variable.

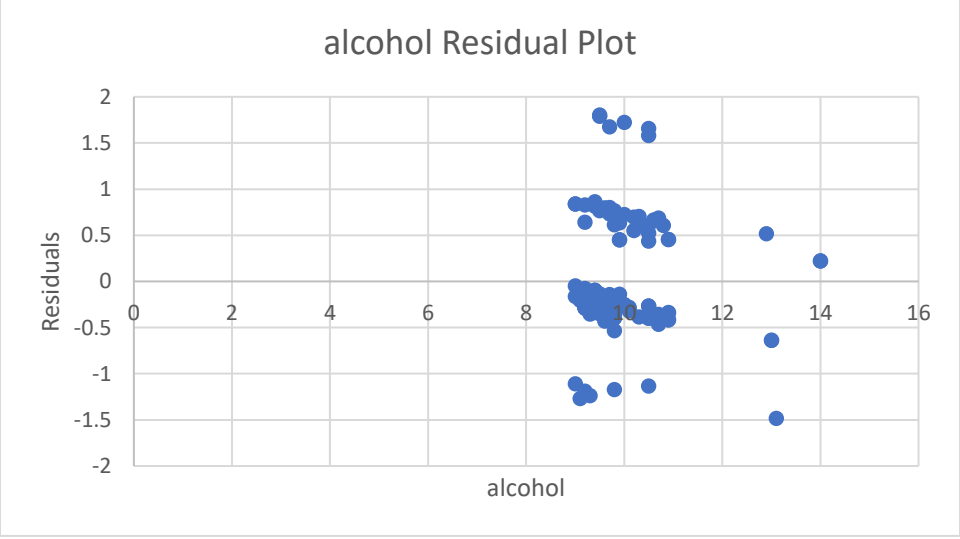

Fig.4: Residual plot of Alcohol against residuals

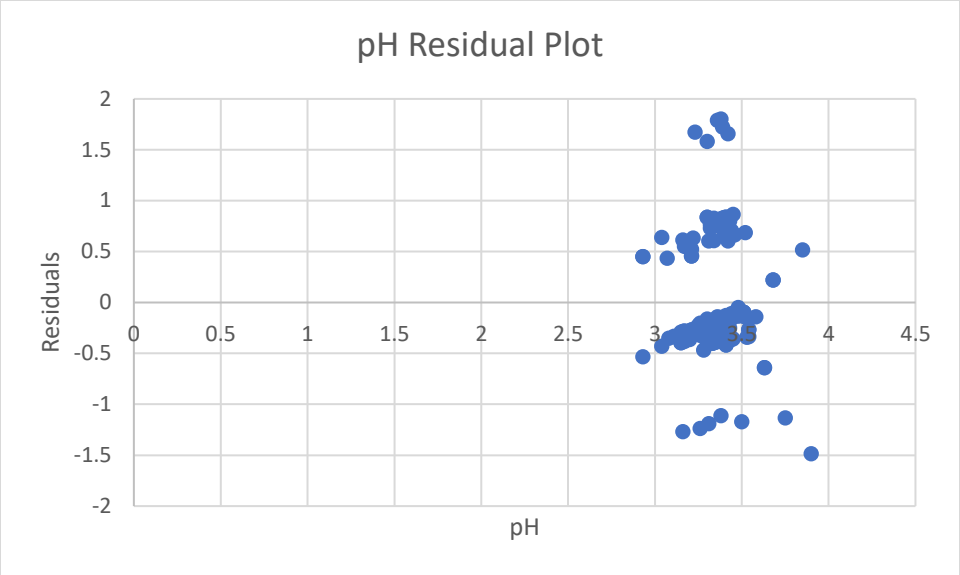

Fig.5: Residual plot of residuals vs pH value

Independent residual plots show how fit the regression model fits the dataset. Outliers can be identified by using standardized Residual plot. Any data point with value greater than 2 or less than -2 can be considered as outliers. In this scenario, 3 points are outliers.

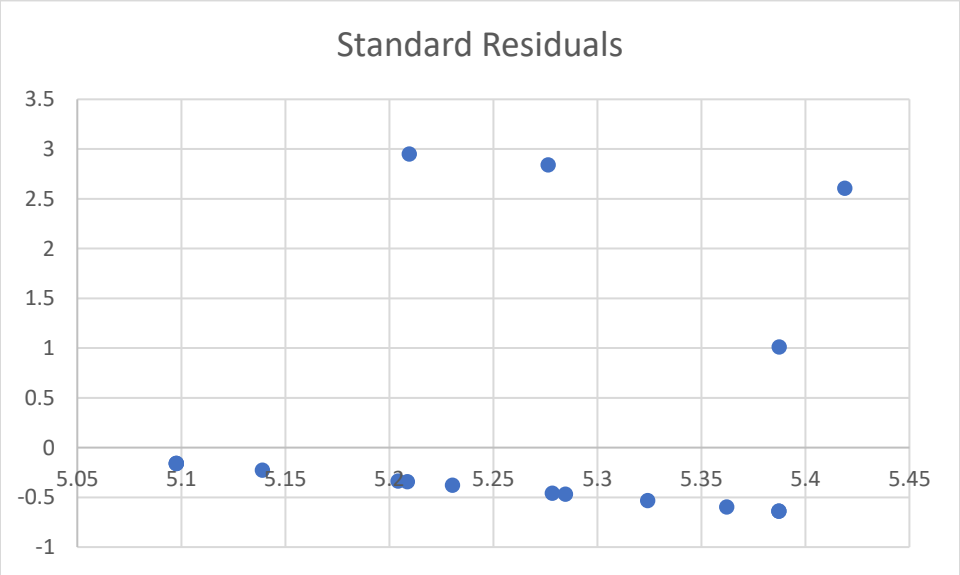

Fig. 6: Standard Residual plot for the dataset

It can be concluded that Alcohol content has a significant relationship with the quality of wine. pH value has little impact on the quality, since the co-efficient of determination is very low additional factors also contribute to the quality. The analysis is useful in optimizing the quality of the wine by modifying the alcohol content, residual sugar and other parameters. The analysis can

be extended to other physicochemical factors to determine the optimum range of quality and can also be applied to physical characters like soil moisture, sunlight and environment. The analysis can be applied to other datasets to get an insight on best operating conditions, therefore maximizing the yield and profits for operation.

Works cited

UCI Machine Learning, "Red Wine Quality", 2009,https://archive.ics.uci.edu/ml/datasets/wine+quality

Stat Trek, "Statistics Dictionary", 2018, https://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination

Dummies.com, "How to Discern Wine Quality", 2018, https://www.dummies.com/food-drink/drinks/wine/how-to-discern-wine-quality/