Data Science Project:  
Red Wine Quality

by: Nicholas Katzenberger, Adam Fiorito, and Laszlo Feledy

ENGR 241

Professor Ronghuan Xu

&

Professor Sarath Chandra Kumar Jagupilla

1, December 2023

I pledge my honor that I have abided by the Stevens Honor System.

Signed: Nicholas Katzenberger, Adam Fiorito, and Laszlo Feledy

# Table of Contents

[Table of Contents 2](#_Toc152694984)

[Introduction 3](#_Toc152694985)

[Data Analysis 6](#_Toc152694986)

[Linear Regression Modeling 9](#_Toc152694987)

[Interpretation and Conclusions 12](#_Toc152694988)

[References 13](#_Toc152694989)

# Introduction

The data analyzed in this report originates from data published by Professor Paulo Cortez of the University of Minho in Guimarães, Portugal. Professor Paulo Cortez works in the Department of Information Systems in the engineering department of the university. This data set is a part of a two series set where 12 attributes of red and white "Vinho Verde" wine were analyzed. This report focuses on the red wine section and its regression data.

The variables used in this report are essential to understanding data science as well as making any conclusions about the data.

The population mean—represented by μ —is the average of all the values in a particular data set. The equation for the mean is computed as:

Where ∑Xi represents the sum of all the elements within a set of data. And n represents the size of the population.

The standard deviation is “a measure of how dispersed the data is in relation to the mean” as defined by the [National Library of Medicine](https://www.nlm.nih.gov/oet/ed/stats/02-900.html#:~:text=A%20standard%20deviation%20(or%20%CF%83,data%20are%20more%20spread%20out.).

The equation for standard deviation is

The square of standard deviation is variance (σ2), or the measure of variability in a data set.

F distribution studies the behavior of two variances from samples taken from two different normal populations. The F statistic is any statistical test in which the test statistic has an F-distribution under the null hypothesis.

It is calculated by: or

MSR or the Mean Sum Regression sum of squares measures the squared difference between the predicted values and the mean of the dependent variable in a data set. It is given by:

MSE or the Mean Squared Error: measures the average of squares of the errors in a data set. It is given by:

The variable p equals the number of parameters in the regression model. For example, if a regression model has two data sets that it compares, p would equal 2.

SSR or the Sum of Squares Regression describes how well a line fits the data. It is the is the sum of the differences between the predicted value and the mean of the dependent variable . It is calculated by

SSE or the Sum of Squares Error is the is the difference between the observed and predicted values.

Also, SST is the Total Sum of Squares, or SSR + SSE. We can keep track of all these variables in a regression model using an ANOVA (Analysis of Variance) table: shown below: A table with numbers and symbols

Description automatically generated

R2 is the coefficient of determination. It represents the proportion of the variance for a dependent variable that’s explained by an independent variable. In other words, it measures the proportion of variance in the dependent variable that is explained in the independent variable.

It is given by:

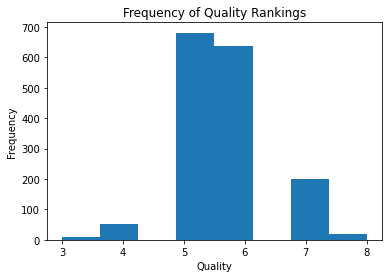
Where the values of the equation are indicated in the ANOVA table.

The t statistic is used to determine if there is a significant difference between the means of two groups. It also determines how the two means are related. tcalc is the calculated value of the t statistic. It is given by but only for Simple Linear Regression. Otherwise, it is calculated by .

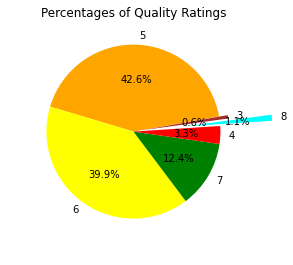
is the estimated regression coefficient. This represents the change in the dependent variable for each one-unit change in the independent variable, assuming all other variables are held constant. is the standard error of the estimated regression coefficient. The standard error is the measure of variability of the sampling distribution of a statistic. In other words, it is an estimation of the standard deviation of the regression coefficient.

tcrit is calculated by using a Statistic table to find where α is a significance level typically ranging from 0.05 to 0.01. tcrit and tcalc are compared to determine if a model is significant: when the model is significant, otherwise it is not.

# Data Analysis



The above plot displays a histogram of the Quality column named in the data. From the data, it is observed that most of the wines were ranked as a 5 or a 6 out of 10. Considering that no wines were ranked above an 8, we began to wonder what affected the quality of each wine. As seen on the graph, the maximum quality ranking is 8 and the lowest quality ranking is 3.



This pie chart visualizes the histogram of the quality ratings in a more colorful way. The rankings are differentiated through color and size instead of just size and quantity. The average quality is 5.636, the standard deviation of quality is 0.808, and the variance is 0.652. while the minimum is 2.74 There were no significant outliers in this data that did not oppose the format of the rest of the data.

A graph of blue bars

Description automatically generatedA graph of blue bars

Description automatically generated with medium confidence

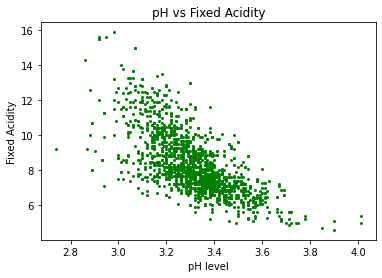
One consistent correlation I found within the data is that Quality increased on average when the percentage of volatile acid decreased. It can be seen on the graph that higher quality wines have a lower percentage of volatile acid. However, this was not the case for Citric Acid, as a similar chart demonstrates that lower quality wines had significantly less percentage of citric acid. The data for both acids was adjusted to exclude outliers from the data that consisted of wines with less than or equal to 0 g/mol or 0 mg/L. These were likely errors in the data and were not considered for this model. It is likely that a wine of quality 5 commonly has the most acidic percentage overall because most wines were rated as this quality. Therefore, it is possible that quality has a joint correlation between acidity and another factor. However, there is a slight decrease in acidity for both graphs as quality is rated in the higher end.

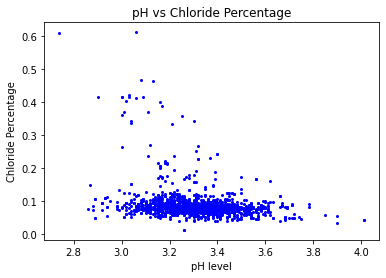
A graph of blue rectangular bars

Description automatically generated with medium confidenceA blue graph with white text

Description automatically generated

The diagram on the left is a histogram of the resulting data after dividing the column titled free sulfur dioxide by the total sulfur dioxide column for all rows. Most of the wines in the data have a ratio of free sulfur and total sulfur between 20% and 50%. The diagram on the right shows the data from the left diagram compared to the quality ratings of each of the wines. From this graph, it can be determined that the ratio of sulfur dioxide does not impact the quality of the wine. This is because wines with qualities 3-8 do not deviate the percentage values significantly from a free sulfur to total sulfur ratio of 70%.



This diagram is a scatter plot between pH and Fixed Acidity. There is a clear downward trend that visualizes pH levels increasing as Fixed Acidity content decreases. This makes sense in a logical sense since the pH level is a measurement of the acidity of a chemical, with lower levels indicating acidity and higher levels indicating basicness.

A chart of a ph and sugar content

Description automatically generated

To test the relationship between the chemicals in the wine data and their acidity, we constructed scatter plots in a similar manner to measuring the fixed acidity. The Chloride Percentage graph had more of a downward trend than the Residual Sugar Content graph. The mean pH level for the entire data set is: 3.31. It has a standard deviation of 0.154 and a variance of 0.024. Some of the points on this graph stray far from the rest of the data, but this information is not statistically impossible, therefore we decided to keep these points within the data sets for further analysis. The max pH value is 4.01 while the minimum pH value is 2.74.

# Linear Regression Modeling

Simple Linear Regression Modeling

R-squared: 0.466 α = 0.05

F-statistic: 1396

Prob (F-statistic): 4.06e-220

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source of Variation | Deg. of Freedom | Sum of Squares | Mean Sum of Squares | F |
| pH | 1 | 2259.62 | 2259.62 | 1396.21 |
| Residual | 1597 | 2584.58 | 1.618397 |  |
| Total | 1598 | 4,844.20 |  |  |

H0 = | tcalc| < | tcrit|, H1 = | tcalc| > | tcrit|

tcalc = = = 37.366

tcrit = t0.025,∞ = 1.96

| tcalc| > | tcrit| meaning we can reject H0, thus this model is significant

Line equation: y = -7.7023x + 33.8228

A chart of a ph versus fixed acidity

Description automatically generated

Line generated from Simple Linear Regression calculations and formula from above. The slope and y intercept were calculated by the python program. There were a small number of outliers from the rest of the data that did not significantly impact the results of the regression.

Multi Linear Regression Model

R2 0.074

F-statistic: 64.97

Prob (F-statistic): 7.50e-28

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source of Variation | Deg. of Freedom | Sum of Squares | Mean Sum of Squares | F |
| Chlorides | 1 | 2.675 | 2.867 | 129.934 |
| Residual Sugar | 1 | 0.192 | 0.192136 |  |
| Residual | 1596 | 35.221 | 0.022 |  |
| Total | 1598 | 38.088 |  |  |

H0 = | tcalc| < | tcrit|, H1 = | tcalc| > | tcrit|

=0.003

Tcalc = = = -10.84 (chlorides)

Tcalc = = -2.6 (sugar)

tcrit = t0.025,∞ = 1.96

| tcalc| > | tcrit| is true for both cases, thus we can reject H0 and say that the entire model is significant.

(axis inverted)

Y = -1.168x +3.977 (chlorides graph)

Y = -128.205x +436.641 (sugar graph)

Actual: (Shown below)

Y = -0.8564x + 3.4058 (chlorides graph)

Y = -0.0078x +3.4058 (sugar graph)

Y = -0.8564x1 -0.0078x2 + 3.4058 (multi-linear model)

A graph of residual sugar

Description automatically generatedA graph of a ph versus chloride percentage

Description automatically generatedThese best fit lines were generated from the equations above. The slope and y intercept were calculated by the python program. The axes are inverted because the multi linear regression model used pH as the dependent variable, and the ANOVA chart represents this change. The negative regression coefficients represent that the pH column compared to the Residual Sugar Content and Chlorides columns have an indirect relationship.

# Interpretation and Conclusions

From analyzing the data in multiple different ways, the group learned how the concepts discussed in class applied to data sets. For example, modeling a linear regression model demonstrated how the ANOVA tables represented calculations obtained with the data. This data also aided in proving the significance of the model since the F statistic translated to the t statistic in simple linear regression. Alternatively, significance was also proved by determining the regression coefficient and the standard error of the estimated regression coefficient to calculate the t statistic. Comparing this to the critical value of t proved significance.

The team also learned about the importance of filtering out data that did not match with the rest of the dataset, also known as outliers. Outliers in the data can significantly change regression models and compress data visualizations (such as graphs) to a point where most of the data points are difficult to distinguish between. Removing outliers from the data set enables a more accurate representation of the data. In addition, looking at data in different ways helps others to see different trends in a population that is otherwise not seen. Exploring data in this way enabled us to create unique data visualizations that account for unseen correlations in the data. Along with learning some new tricks in Python, the group ultimately gained a better understanding of data science.

# References

1. Fernando, J. (2023, May 12). *The Correlation Coefficient: What It Is, What It Tells Investors* (M. James & V. Velasquez, Eds.). Investopedia; Dotdash Meredith. <https://www.investopedia.com/terms/c/correlationcoefficient.asp>
2. National Library of Medicine. (2012). *Finding and Using Health Statistics*. Www.nlm.nih.gov; National Library of Medicine. [https://www.nlm.nih.gov/oet/ed/stats/02-900.html#:~:text=A%20standard%20deviation%20(or%20%CF%83](https://www.nlm.nih.gov/oet/ed/stats/02-900.html%23:~:text=A%20standard%20deviation%20(or%20%CF%83)
3. Standard Normal Cumulative Probability Table <https://stevens0-my.sharepoint.com/personal/nkatzenb_stevens_edu/Documents/Statistics_Tables.pdf>
4. Python Code (Attached with submission)