Red Wine Quality Analysis based on Machine Learning Techniques

Jianhong Dong *

Foster School of Business, University of Washington PACCAR Hall, 4273 E Stevens Way NE, Seattle, WA 98195

* Corresponding Author Email: jdong2@uw.edu

Abstract. The red wine industry is growing at a tremendous speed as more and more people start to drink wine. Therefore, the industry is becoming competitive and wine companies need to make better quality wines to stand out. This paper used machine learning techniques to analyze 1599 wine samples each with 11 input variables in order to find the variables that have the most impact on wine's general quality. The linear regression model used in the paper shows the most influential variables on quality are alcohol and acid. In addition, a heat map was adopted to show all the correlation between the variables. To go deeper, box plot and 3D scatter plot were used to support the finding through linear regression model and have a more detailed conclusion on the variables that have the most impact on quality. These results shed light on what are the most influential variables on wine's quality.

Keywords: wine industry, wine quality, quality improvement.

1. Introduction

Wines have a history of 6000 years, winemaking was first found in the southern Caucasus area, then the methods of growing grapes and making wines traveled to a lot of other countries, e.g., Syria, Egypt, and Mesopotamia. Wines were used for religious purposes instead of for people to consume for a long time until the 17th century. It took another 2 centuries for North America to finally have its first vineyards in California about 200 years ago [1].

However, nowadays in America, people spend more than 70 billion dollars on wine every year. Approximately every 1 out of 3 adults in America drinks wine regularly, which means drinking more than 14 times a week for men and more than 7 times a week for women [2]. That is a lot more than people would assume, which makes the wine industry very huge and profitable. Between different wines, about 60% of Americans prefer red wine over white wine. It is now believed that moderate consumption of wine (5 oz per day) or any alcohol would be beneficial to people, it can increase survival rate and decrease risk of heart diseases [3]. That makes the quality of wines important for those who drink wines regularly. Some common indicators of red wines' quality are the origin of the grape that makes the wine, like its species, ground type, climate [4], as well as the polyphenolic compounds like flavonoids, anthocyanins and tannins in wines [5]. Some past related study show that different indicators of wines have different impact on the wine's quality, some indicators can be identified as more influential than others by using machine learning techniques [6]. Different indicators will have different accuracy to predict wines quality if they are used in a quality prediction model, indicators like chemical compound found in wines normally can predict the quality better [7, 8].

The research focus of this paper is using machine learning techniques to analyze what variables of wines have stronger impact on the wines quality and can be used to better predict the quality so that the wine companies can use this as a reference to make better wines with the resources they have. This study will use different models to show the correlations between the variables to identify the most influential indicators of wines.

2. Data & Method

The dataset that was used is the "Red Wine Quality" dataset from Kaggle, the dataset has 1599 samples with 12 variables, of which 11 of them are input variables and 1 is the output variable — quality, which is what this research is about. The 11 input variables and the definitions are given as follows,

Fixed acidity: low volatility organic acids
Volatile acidity: short chain organic acids that can be extracted (amount of acetic acid)
Citric acid: Added to wine to finish in order to add a fresh flavor to wine
Residual sugar: Sugar added remaining after fermentation stops
Chloride: Amount of salt in wine(limited)
Free SO2: so2 directly in acidic media, prevents microbial growth
Total SO2 = free so2 + combined so2
Density: Mass per 1 unit of wine or must at 20-degree temperature (avg between 1.08-1.09)
pH: how acidic wines are (normally 2.5-4.5)
Sulphate: Produced by yeast through fermentation, Protects wine from oxidation, bacteria
Alcohol: Percent of alcohol

3. Results & Discussion

After understanding the data, the two main objectives are to find out are what variables have the biggest impact on the wine quality and what are the possible correlations between those variables. Before started, some hypotheses were made based on the facts of those variables. First one is that residual sugar would be an important variable that impacts the quality of the wine because it was found during the research that cheaper wines normally have more sugar in them. Another one is that one variable that is thought would be influential on quality is the acid, In this paper, "acid" is defined as the combination of the three variables: both fixed and volatile acidity and citric acid. Because from the research on the internet, too much volatile acidity would not be desired and citric acid keeps wines fresh. After finishing hypotheses, python was used as the programing tool to find out if they are correct and to achieve the objectives. The first thing was to illustrate a general view of the data summary of the dataset (shown as Table 1) by simply using the describe function in python to help better understand the data like each data's mean, range, maximum and minimum value, and distribution.

Table 1. Summary of the statistical descriptions of the datasets.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00
mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	1.00	3.31	0.66	10.42	5.64
std	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00
25%	7.10	0.39	0.09	1.90	0.07	7.00	22.00	1.00	3.21	0.55	9.50	5.00
50%	7.90	0.52	0.26	2.20	0.08	14.00	38.00	1.00	3.31	0.62	10.20	6.00
75%	9.20	0.64	0.42	2.60	0.09	21.00	62.00	1.00	3.40	0.73	11.10	6.00
max	15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00

Then to find out what variables have the biggest correlation with the quality of wine. Linear regression model would be the best option to find relationships between variables, so Linear Regression Model was used on some of the input variables to find out which ones of them are the most influential to the quality [9]. The only reason it was not used for every input variable is that some of the variables are basically the same for all kinds of wine after the research on the variables, like the pH value and the density would be something that are mostly the same and not able to change. After having some Linear Regression Model laid out together (as shown in Figure 1), It was found that volatile acidity, citric acid, total sulfur dioxide, sulphates and alcohol seem to have relatively stronger correlations (both positive and negative) than the rest. Like predicted, volatile acidity and citric acid do have big impacts on the wine's quality, but surprisingly, residual sugar, which was thought to be the most influential, turns out to not seem to have any correlation at all.

Subsequently, Pearson correlation heat map (seen from Figure 2) was created in order to show the correlation more clearly with specific number value. Besides, the correlation between different input is presented [10]. This heat map can directly show the correlation coefficient between every two variables. In this heat map, the light color (light orange to white) means it has a strong positive correlation between the two variables and a dark color (dark purple and black) means it has a strong negative relationship between the two variables, any color in the middle that's showing there is no significant relationship between the two variables.

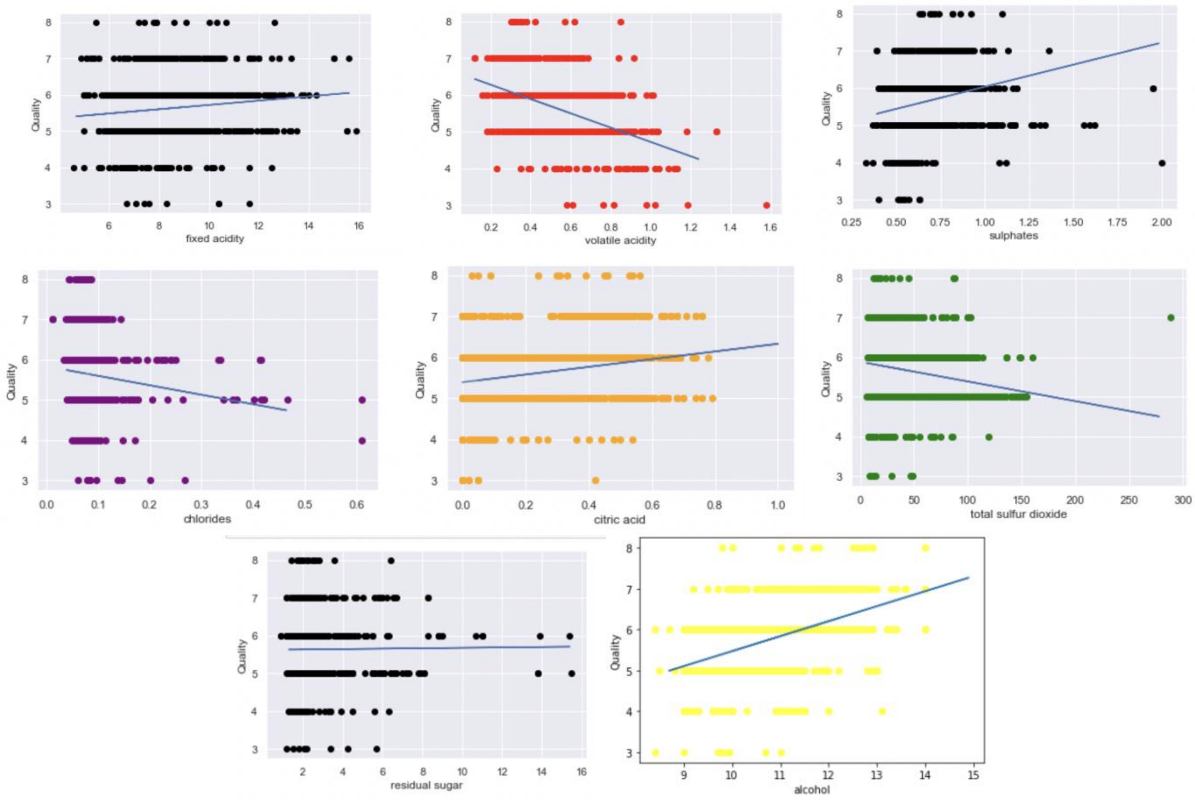

Figure 1. Linear Regression Model

Seen from Fig. 2, the first focus was on the quality again to find out exactly how much impact those influential variables found in the linear regression model have. According to the results, volatile acidity has a correlation coefficient of -0.39 with quality, citric acid has a correlation coefficient of 0.23 with quality, sulphates have a correlation coefficient of 0.25 with quality and alcohol has a correlation coefficient of 0.48 with quality. The variable that has the strongest correlation with the quality turns out to be the alcohol percent of the wines, which is different from the hypothesis of the study before the analysis. Now that the it is found that the input variable that has the biggest impact on the wine quality is its alcohol percent. The study then wanted to find out if there are any relationships between the input variables using the same heat map. In this study, an absolute value of

0.2 or higher would be the minimum to be considered as having a correlation between two variables and pH value and density are not considered. However, after going through the whole map, It was found that there are very few correlations between any two variables, chloride and citric acid (0.2), total sulfur dioxide and residual sugar (0.2), sulphates and chloride (0.37), that's about all and those relationships would not be considered valuable to the research because they are not that related to those influential variables.

With this in mind, this study will specifically study the three variables that have the most impact on the quality, which are alcohol, volatile acidity and citric acid. Study will do more detailed research on alcohol first then do volatile acidity and citric acid together as a group since they are both acids. A box plot was made to show the alcohol percent distributions for every quality category as illustrated in Figure 3, because seeing the distribution of every quality wines can also better show the whole sample distribution and how alcohol affects the quality of wines. From the bot plot, it's easy to see that as quality of wines go up from (5-8), the alcohol percent goes up as well with a higher mean, a narrower range and fewer and fewer outliers. This means higher quality wines are even more strict with having a higher alcohol percent. Higher the quality, the more likely its alcohol percentage is high with few exceptions. This supports how alcohol has the biggest impact on the quality even more.

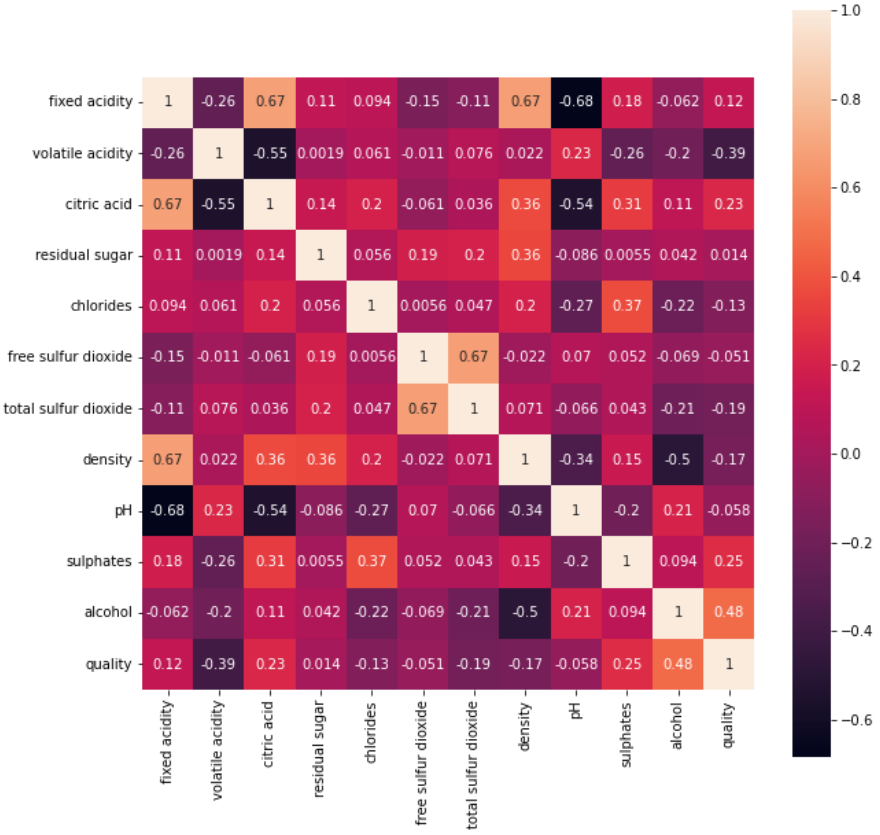

Figure 2. Heat map.

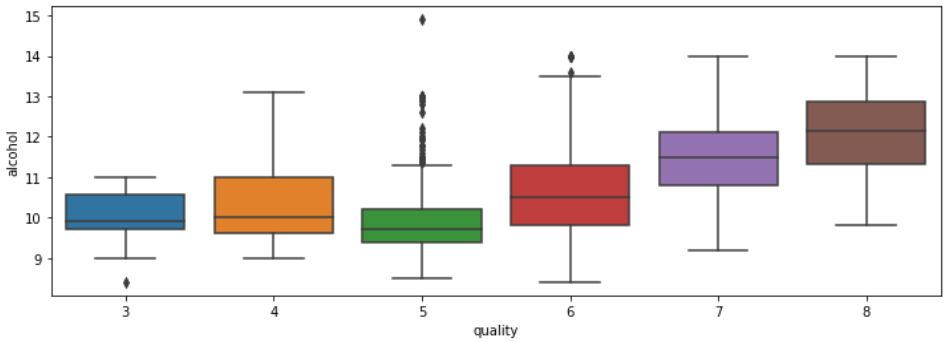

Figure 3. Alcohol Box plot.

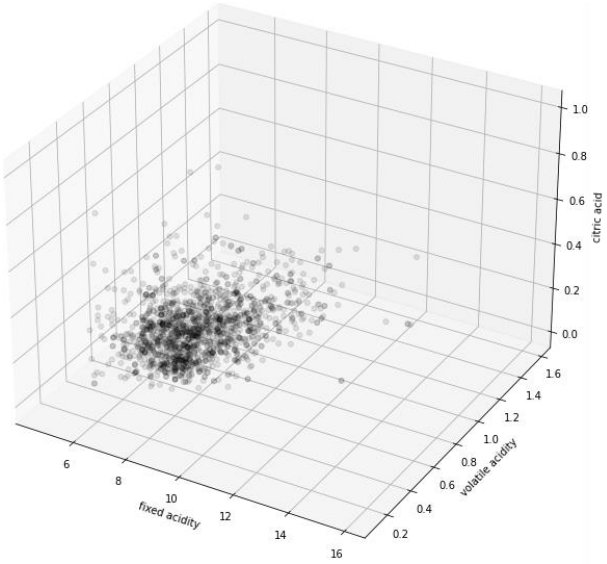

Figure 4. Scatter 3D chart.

After alcohol, study will focus and go deeper with the second biggest impact variable – the acid. From the previous finding, it was found that the quality has a strong negative correlation with volatile acidity, a positive correlation with the citric acid and a less strong relationship with fixed acidity. As mentioned before, those three together are what was identified as the "acid" variable. However, instead of showing the relationship between acid and quality, a 3D scatter plot (depicted in Fig. 4) was decided to be made to show the relationship between the three kinds of acid also with the acid distribution [11]. This plot shows the higher the fixed acidity is, the higher the citric acid is, and the correlation is strong, also as fixed acidity increases, the volatile acidity decreases, which makes the fixed acidity an influential variable on the quality. Fig. 5 is a flow chart showing the steps that were took.

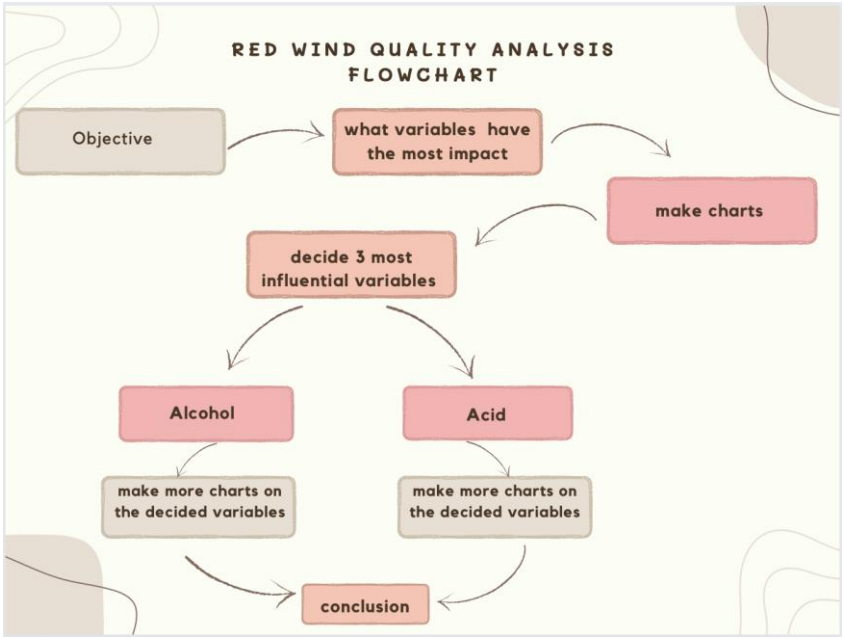

Figure 5. Flow chart.

4. Limitations & PROSPECTS

The biggest limitation of the study is that it does not give an insight of how quality was impact by those variables other than having the relationships quality has with those them. One thing that can be improved is to not only have qualities as numbers, but also group them in categories based on their

quality value (good, bad, middle). Subsequently, there will be some more insights on what makes a wine good or bad. For a possible future topic, now that it's been found how alcohol percentage can have a huge impact on the wine's general quality, it will be very interesting and also beneficial for wine companies to do some research on how much it would cost to make wines with different alcohol percentage, and what are the average sales value for each alcohol percentage wines, once they have that, they can decide what is the best alcohol percentage they should make more for a company in terms of maximizing profit. It's necessary process they want to go through that is worth the time and money if they want to keep growing as a business.

5. Conclusion

In conclusion, this paper investigates red wine quality based on the input variables and the graph made with those variables. With those graphs created, it is found that the key to higher quality is higher alcohol percent, and with a high alcohol percent, what also decides wine quality would be fixed acidity and citric acid, the more wines have those, the better the quality of wines. those findings can be used by wine industry companies if they want to make better quality wine with limited resources or budget so they can know what the priorities in red wine making are. In the future, they can use the findings to make better quality wines alone with more profit as businesses. Overall, these results offer a guideline for all the red wine companies in terms of improving their products.

References

[1] Soleas G. J., Eleftherios P. D., and David M. G. "Wine as a biological fluid: history, production, and role in disease prevention." Journal of clinical laboratory analysis 11.5: 287-313 (1997).

[2] Abernathy C. "Press Release: Frequent Wine Drinking Population in the US in Decline, Led by Younger Consumers, Though Overall Participation in Wine Category Up." Wine Intelligence, Courtney Abernathy Retrieved from: https://Www.wineintelligence.com/Wp-Content/Uploads/2018/07/logo5.Png, 13 Jan. 2020,

[3] Szmitko P. E., and Subodh V. "Red wine and your heart." Circulation 111.2: e10-e11 (2005).

[4] Coarfă E., and Mona Ea Popa. "Some relevant quality indicators of red wine from three grapes cultivars–a minireview." Scientific Bulletin. Series F. Biotechnologies 22 (2018): 70-80.

[5] Schamel G. "Individual and Collective Repuatation Indicators of Wine Quality." Datasets, (2000).

[6] Gupta Y. "Selection of important features and predicting wine quality using machine learning techniques." Procedia Computer Science 125: 305-312 (2018).

[7] Dahal K. R., et al. "Prediction of wine quality using machine learning algorithms." Open Journal of Statistics 11.2: 278-289 (2021).

[8] Bhardwaj P., et al. "A machine learning application in wine quality prediction." Machine Learning with Applications 8: 100261 (2022).

[9] Sklearn.linear_model. Linearregression. Scikit, Retrieved from: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[10] Piyushagni5. "Multi-Dimensional Data Visualization: EDA." Kaggle, Kaggle, 29 Sept. 2020, Retrieved from: https://www.kaggle.com/code/piyushagni5/multi-dimensional-data-visualization-eda.

[11] "3D Scatterplot#." 3D Scatterplot - Matplotlib 3.5.3 Documentation, Retrieved from: https://matplotlib.org/stable/gallery/mplot3d/scatter3d.html.