CS105 Project Report

Data Modelling of Wine Quality based on Chemical Properties

G1T6 - Wine Connoisseurs

Edwin Tok Wei Liang (edwin.tok.2019@sis.smu.edu.sg)

Harshit Jain (harshitj.2019@economics.smu.edu.sg)

Hartono Tjakrawinata @ Jonathan Chow (hartonot.2019@sis.smu.edu.sg)

Tay Wei Jie (weijie.tay.2019@sis.smu.edu.sg)

ABSTRACT

In this report, we investigated the relationship between the chemical properties and quality rating of the Portuguese "Vinho Verde" wine. Initial exploratory data analysis and correlation analysis was performed to filter out the chemical attributes. A Multiple Linear Regression model was then applied to investigate the relationship between the chemical attributes of wine and its rating. We performed clustering to further investigate the effect of the features on the ratings. Despite the relatively low r^2 value, we feel that the linear model is still a reasonable estimate of the wine ratings for practical applications. Wines with higher alcohol content seems to score a higher rating, while wines with high volatile acidity and total sulfur dioxide do poorer.

INTRODUCTION

Taste is fundamentally based on the interaction between the chemical compounds of our food and the receptors on our taste buds. There are several chemical factors that may influence a wine's taste - acidity, pH level, density, residual sugar, alcohol concentration produced during fermentation etc. Taste preferences is one of the least understood concepts of the human body. As such, our team intends to investigate how different chemical properties of the wine affects its taste rating to help develop a more robust method to predict the reception of the wine.

DATASET

The Wine Quality dataset is separated into two sections corresponding to the red and white variants of the Portuguese "Vinho Verde" wine. The dataset contains a total of 6,497 instances, with 1,599 samples for red wine and 4,898 samples for white wine[1].

2.1 Overall Description

There are 11 independent numerical variables - fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and 1 dependent numerical variable - wine quality.

2.2 Data Preprocessing

All attributes were ensured to contain no missing data and only contain values where are numerical in nature.

APPROACH

3.1 Exploratory Data Analysis

The primary outcome of the EDA is to filter out the chemical attributes to obtain the minimal set of attributes on which to perform multiple linear regression. To begin with, we performed Independent 2 sample t-test for all the features between the red and white wine. We noted that all 12 features were significantly different using a confidence interval of 95%. Hence, we decided to perform the analysis for red and white wine separately to exclude possible inherent chemical differences due to the type of the wine affecting the regression model.

3.1.1 Visual Analysis – Boxplot & Histogram

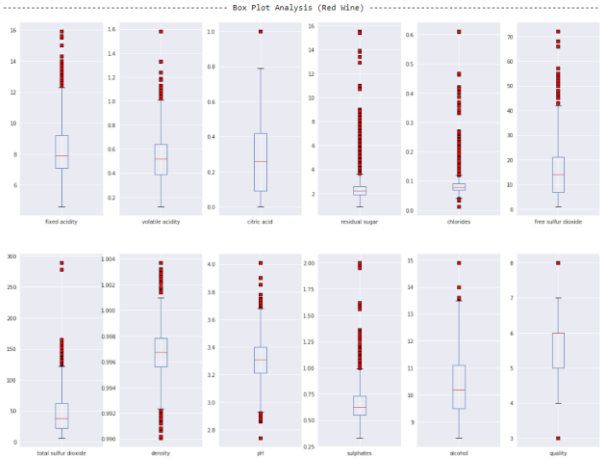


Figure 1: Red Wine - Box Plot Analysis

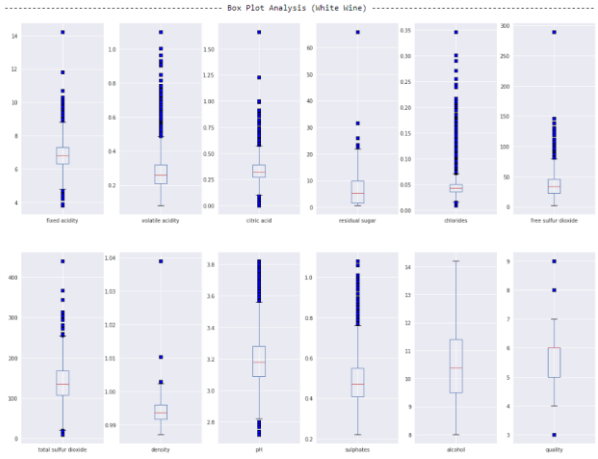


Figure 2: White Wine - Box Plot Analysis

*Histogram can be found in the Jupyter notebook

From visual observation, we noted that the rating of the wine is largely normally distributed, with most readings in the region of 5 – 7, which is consistent with the expected behavior of a ratings dataset[2]. Residual sugar, chlorides and

sulphates have many outliers in the upper quartile. This indicates that there are many readings on the higher end which may skew the regression model. Comparing red wine to white wine, residual sugar and total sulfur dioxide is on average lower for red wine while sulphates is higher for red wine.

3.1.2 Correlation Analysis

Next, we performed a correlation analysis to evaluate the relationship between the 11 independent quantitative variables to the wine ratings. The analysis was done in 2 stages. Firstly, the independent attributes that have a low correlation (<15%) with wine ratings were removed. This step selects for attributes which are likely to have a higher impact on wine ratings. Subsequently, the remaining independent attributes, after the first stage, with high correlation (> 50%) with each another were removed to select for mutually independent attributes. For each pair of attributes with high correlation (>50%), the attribute with lower correlation with wine quality was dropped. Finally, the attributes selected for red wine and white wine are shown below.

Red Wine	Volatile Acidity, Total Sulfur Dioxide, Density, Sulphates, Alcohol
White Wine	Volatile Acidity, Chlorides, Total Sulfur Dioxide, Alcohol

Table 1: Final Selected Attributes for Both Wines

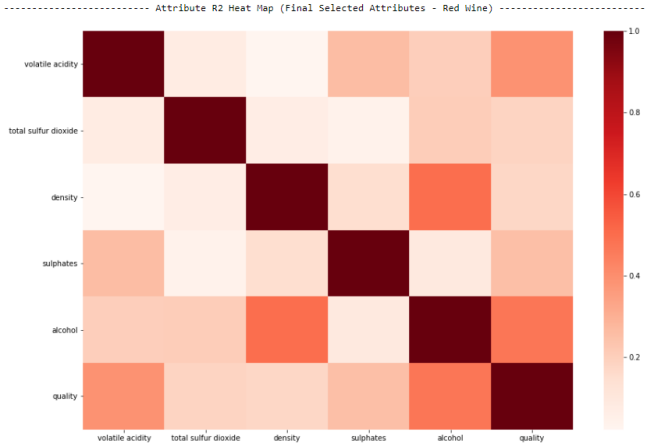


Figure 3: Red Wine - Final Selected Attributes

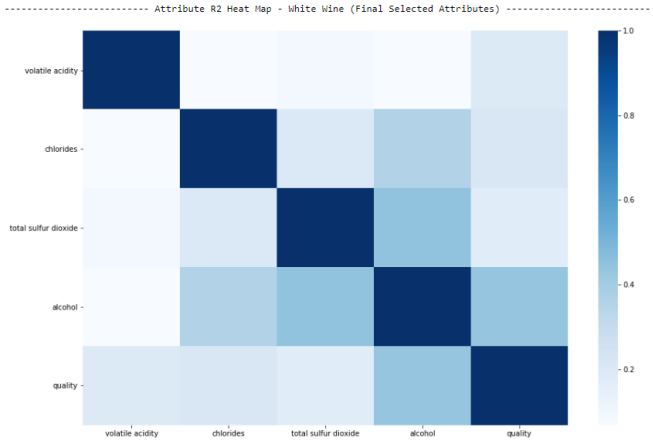


Figure 4: White Wine - Final Selected Attributes

Attribute	***Impact on quality of wine***
Volatile Acidity	Small amounts may contribute to complexity and aroma while excessive amounts may destroy wine's fruitiness and makes it taste like vinegar[5]
Alcohol	Helps balance sweetness and acidity while excessive amounts may overpower the fruitiness[6]
Sulphates	Prevents oxidation and preserves taste[7]
Total Sulfur Dioxide	Prevents bacteria and oxidation, but makes wine impure[8]
Chlorides	Gives wine an undesirable salty taste, reducing quality[9]
Density	Higher density reduces the wine's sweetness and fluidity, reducing quality[10]

Table 2: Wine Attribute's Impact on Quality

3.2 Multiple Linear Regression

3.2.1 Model Building Process

We used the final selected attributes as our predictors and the wine rating as our response variable. As we have multiple predictor attributes, we performed multiple linear regression. Before fitting, we split the entire dataset into 50 / 50 train and test set, i.e. 50% is used to train the model, while the remaining 50% is the "unseen" data used to test the model. We used the `model_selction` package from `sklearn` to perform a random split on our dataset. The model was built using the `LinearRegression` class from `sklearn`. A r^2 score was then calculated for both the training and test data set to measure the "goodness of fit" for the model on both datasets. The dataset splitting and model fitting process was then repeated with 100 different random seeds (1-100) to generate 100 different sample sets and their corresponding models.

3.2.2 Model Selection Process

In order to filter for models not overfitted to the data, the 3 models with the lowest difference in the r^2 score between the train and test datasets was selected. This ensures that the performance of the models generated can more likely be replicated over a new dataset and is not overly fitted to the sample dataset.

3.3 Evaluation and Sensitivity Analysis

The models will be tested against multiple tolerance ranges of 0, 0.25, 0.5, 0.75, 1, ie. given a tolerance range of 0.5, a predicted rating of 5.75 and an actual rating of 5 will be evaluated as having 0.25 error. To allow for a fairer comparison across the 3 models who have different `y-test` datasets, we performed our sensitivity analysis over the entire dataset so that the analysis of all 3 models will be based on the same data. Their corresponding mean absolute error (MAE) and MAE as a percentage of the average quality scores was then used to evaluate the models. This is to better reflect real-life applications where quality scores are given in integers compared to the exact values generated from regression and ratings being highly subjective.

3.4 K Clustering

The red and wine dataset will be split into 3 clusters using the `KMeans` function from the `sklearn.cluster`. The differentiators of each cluster were then determined by running independent 2 sample t-test between the clusters for each feature with a confidence interval of 95%. The actual score distribution of each cluster was then analyzed.

RESULTS AND DISCUSSION

After performing the initial model building process. The distribution of the 100 random samples and their corresponding models is shown in Figure 7 -10. There is an inverse relationship between the training r^2 score and the test r^2 score where a high training r^2 score often results in a low r^2 test score.

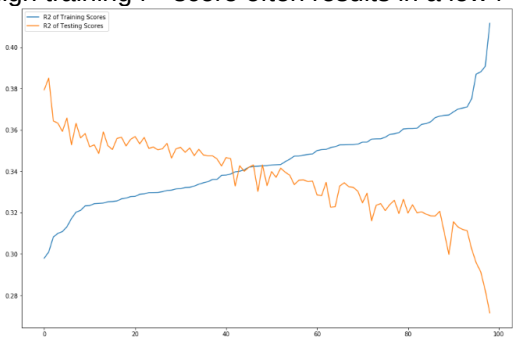


Figure 5: Red Wine – Plot of R2 scores

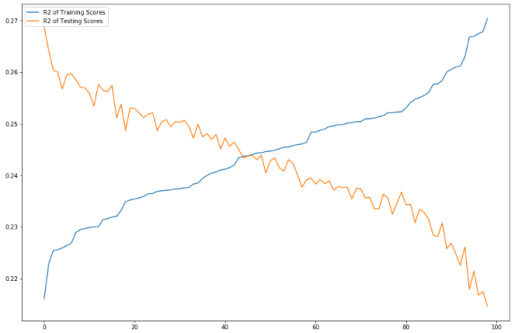


Figure 6: White Wine - Plot of R2 scores

4.1 100 Repeated Cross Validation Results

Model	Random Seed	Intercept	Volatile Acidity	Total Sulfur Dioxide	Density	Sulphates	Alcohol	R^2 Difference
1st	72	15.2	-1.16	-0.00211	-12.1	0.741	0.270	0.000056
2nd	63	4.47	-1.25	-0.00241	-1.63	0.814	0.290	0.000434
3rd	31	-10.5	-1.34	-0.00244	13.3	0.501	0.314	0.000665

Table 3: Red Wine – 3 Best Models after Model Building

Model	Random Seed	Intercept	Volatile Acidity	Chlorides	Total Sulfur Dioxide	Alcohol	R^2 Difference
1st	92	2.81	-2.01	-1.86	0.00144	0.335	0.000048
2nd	90	2.74	-1.90	-2.07	0.00157	0.337	0.000156
3rd	5	2.73	-1.92	-2.13	0.00139	0.343	0.000343

Table 4: White Wine – 3 Best Models after Model Building

4.2 Model Evaluation and Sensitivity Analysis Results

Model	Range (0)	Range (0.25)	Range (0.5)	Range (0.75)	Range (1)
1st	0.511489 (0.09075)	0.297862 (0.05285)	0.161634 (0.02868)	0.082378 (0.01462)	0.041838 (0.00742)
2nd	0.507449 (0.09004)	0.297375 (0.05276)	0.162749 (0.02888)	0.083221 (0.01477)	0.041640 (0.00739)
3rd	0.507656 (0.09007)	0.298780 (0.05301)	0.163844 (0.02907)	0.083923 (0.01489)	0.041850 (0.00743)

*Range(x) - x indicates the tolerance range applied

*The MAE values are indicated above, with the MAE as a % of mean indicated in brackets below

Table 5: Sensitivity Analysis (MAE) for Red Wine

Model	Range (0)	Range (0.25)	Range (0.5)	Range (0.75)	Range (1)
1st	0.602178 (0.10245)	0.381810 (0.06496)	0.224236 (0.03815)	0.129594 (0.02205)	0.075248 (0.01280)
2nd	0.601658 (0.10236)	0.381309 (0.06487)	0.223993 (0.03811)	0.129855 (0.02209)	0.075573 (0.01286)
3rd	0.602174 (0.10245)	0.381714 (0.06494)	0.223942 (0.03810)	0.129669 (0.02206)	0.075555 (0.01285)

*Range(x) - x indicates the tolerance range applied
*The MAE values are indicated above, with the MAE as a % of mean indicated in brackets below

Table 6: Sensitivity Analysis (MAE) for White Wine

For red wine, the 3 models vary significantly in their coefficients and y-intercept. We feel that Model 2 is more appropriate for practical application. It has a y-intercept of 4.47 as compared to 15.2 and -10.5 for the other 2 models. Taking into consideration that the expected value of the wine quality should be in the region of 5 – 7 as mentioned previously, the 2nd model represents this the most. Furthermore, the 2nd model has the best performance during the sensitivity analysis, with the lowest MAE for the ranges of 0, 0.25 and 1, and the second lowest MAE for 0.5 and 1.

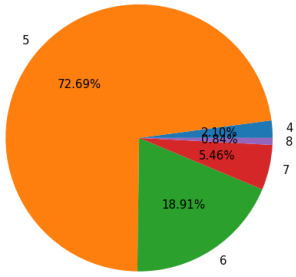
For white wine, the 3 models are actually very similar with only negligible differences between their coefficients and intercept values. Looking further into the sensitivity analysis, their performance is also similar in nature. As such, we would recommend the 1st model since it has the most consistent performance.

$$\boldsymbol{Red\ Wine\ Rating} = 4.47 - (1.25 \times VolatileAcidity) - (0.00241 \times TotalSulfurDioxide) - (1.63 \times Density) + (0.814 \times Sulphates) + (0.290 \times Alcohol) \qquad \boldsymbol{R^2} = 0.343$$

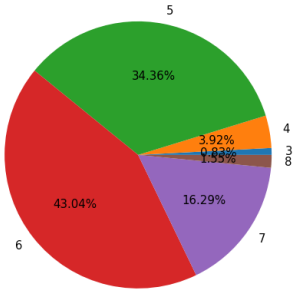
$$\boldsymbol{White\ Wine\ Rating} = 2.81 - (2.01 \times VolatileAcidity) - (1.86 \times Chlorides) + (0.00144 \times TotalSulfurDioxide) + (0.335 \times Alcohol) \qquad \boldsymbol{R^2} = 0.244$$

Figure 7: Selected Best Models for Red & White Wine

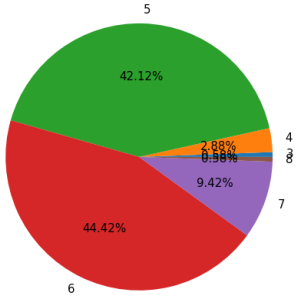
4.3 Clustering Analysis Results



Cluster 1 – Red Wine significantly higher in *Volatile Acidity* and *Total Sulfur Dioxide*

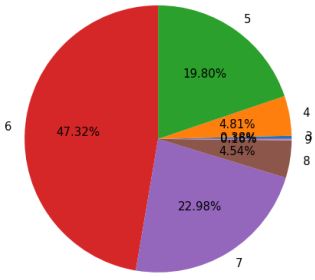


Cluster 2 – Red Wine Significantly higher in *Alcohol*

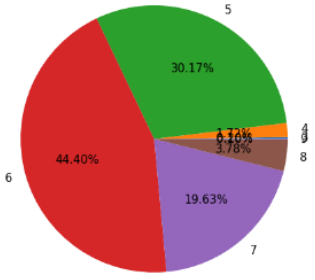


Cluster 3 – Red wine relatively balanced in all their chemical features

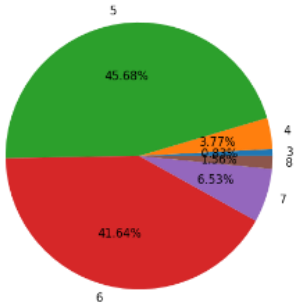
Figure 8: Red Wine - Actual Wine Quality Distribution by Cluster



Cluster 1 – White Wine significantly higher in *Alcohol*



Cluster 2 – White Wine relatively balanced in all their chemical features.



Cluster 3 – White wine significantly higher in *Volatile Acidity*, *Total Sulfur Dioxide* and *Chlorides*

Figure 9: White Wine - Actual Wine Quality Distribution by Cluster

Volatile acidity seems to be the one of the most important attributes in the determination of wine quality, being present in both red and wine rating models. This could be because volatile acidity not only affects the taste but also the aroma of the wine[3] and given that up to 80% of taste can be affected[4] by our sense of smell, changes in volatile acidity may significantly change the wine quality profile. Furthermore, the clusters with significantly higher volatile acidity have a lower score on average as compared to the other clusters. This could be because volatile acidity in larger quantities can result in an unpleasant pungent vinegary flavor[3]. Hence, the model is accounting for the cases where there is excessive volatile acidity. Interestingly, residual sugar, the attribute seemingly most related to taste, has extremely low correlation with quality (<0.15). This could be because of the highly varied nature of sweetness preference resulting in it being ineffective to be used as a predictor for quality. For both wines, the cluster with high alcohol has the highest proportion of the higher rated scores(≥ 7) which is also similarly reflected in our regression where alcohol is positively correlated with quality. However, the high alcohol cluster also has the highest proportion of low scores (3-4). This corresponds with secondary research where alcohol may have an opposing effect on quality based on its quantity.

The multiple linear regression model performs better for red wine as compared to white wine with an r^2 value of 0.343 to 0.244. However, both models still have a relatively low r^2, which may be due to the unpredictability and significant variance of the human taste. The MAE weighted against the average still indicates a relatively low error margin of 0.7% (tolerance range of 1) - 10% (tolerance range of 0). In practical application, a tolerance range of 1 may still be acceptable given that most taste preferences are subjective. Furthermore, most ratings are given in integers as compared to the linear regression model which predicts an exact value.

However, we do note that the current dataset is heavily concentrated in the region of 5 - 7 quality scores, hence the model may be biased to these ratings. Further analysis can be done on datasets where the quality scores are more evenly distributed to improve on the predictive ability of the model.

CONCLUSION

In this project, we have successfully carried out the EDA on the wine quality data and performed multiple linear regression and clustering on the selected attributes. Although the r^2 values of our eventual models may be relatively low, given the practical application of this model as a quality rating, the margins of error after sensitivity analysis suggests that the multiple linear regression model is still a reasonable estimate of the wine quality. Wines with higher alcohol content seems to score a higher rating, while wines with high volatile acidity and total sulfur dioxide do poorer.

CONTRIBUTIONS OF EACH MEMBER

Edwin Tok Wei Liang	Planned overall schedule, conducted research, wrote the introduction, dataset description and EDA, assisted in the coding, formatted the report and Jupyter Notebook
Tay Wei Jie	Contributed most of the coding, conceptualized and developed the linear regression model and sensitivity analysis, wrote the model development approach, results and discussion, edited and formatted the entire report and Jupyter Notebook.
Harshit Jain	Wrote an initial draft on introduction and dataset, contributed some ideas for the results and EDA, contributed some coding on EDA.
Jonathan Chow	Contributed some ideas for the EDA

REFERENCES

1. P.Cortez, A.Cerdeira, F.Almeida, T.Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
2. Dalvi, N.N., Kumar, R., & Pang, B. (2013). Para 'Normal' Activity: On the Distribution of Average Ratings. *ICWSM*.
3. Goode, J. (2018). Flawless : understanding faults in wine . Oakland, California: University of California Press.
4. Styger, Gustav & Prior, Bernard & Bauer, Florian. (2011). Wine flavor and aroma. Journal of industrial microbiology & biotechnology. 38. 1145-59. 10.1007/s10295-011-1018-4.
5. GORMAN-MCADAMS, M. (2013, March 25). Wine Words: Volatile Acidity. Retrieved from thekitch: https://www.thekitchn.com/wine-words-volatile-acidity-186893
6. How does alcohol content affect wine? (2015, April 16). Retrieved from tennessean: https://www.tennessean.com/story/life/food/2015/04/17/alcohol-content-affect-wine/25779589/
7. Bodin, G. (n.d.). SULFITES IN WINE - DYNAMIS. Retrieved from dahu.bio: https://www.dahu.bio/en/knowledge/oenology/sulfites-in-wine-dynamis
8. Woolf, S. (2016, March 18). Sulfites in wine: friend or foe? Retrieved from Decanter: https://www.decanter.com/learn/wine-terminology/sulfites-in-wine-friend-or-foe-295931/
9. Chloride concentration in red wines: influence of terroir and grape type. (2015, February 15). Retrieved from scielo: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-20612015000100095
10. Why is Wine Density Important? . (n.d.). Retrieved from ChemWine: https://www.chemwine.com/home/why-is-wine-density-important-1-sl6yl