# **MIDTERM PROJECT MSDA STUDENTS (GROUP 3)**

**Business Problem 1**: Which attributes of wine impact the quality score?

(Regression – continuous output)

**Business Problem 2**: How many attributes of wine are needed to predict whether their

quality score will be low or high?

(Classification - changing quality score to a binary variable [1-5 is low and 6-10 is high])

Aganbi, Eferhire

Basiri, Hedieh

Koneru, Anita

Mobley, Sofia

Narayanabhatla, Hari Teja

**Business Question 1:** Which attributes of wine impact the quality score?

**Dataset Overview**

The dataset contains information about 1143 entries representing 1143 wine samples; with 13 variables per sample. The independent variables in the dataset included: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and Id. The dependent variable we selected was “quality”; which is the quality of a given wine measured on a score of 0 – 10. Using the df[‘quality’].value\_count() function, the ’quality’ scorings were: 5 = 483, 6 = 462, 7 = 143, 4 = 33, 8 = 16, and 3 = 6. Our hypothesis is that, the higher the score, the better the quality of wine; nevertheless, finding solution (s) to the business problem of focus would allow confirmation of the grading system. Within the dataset, no variables had missing values, the mean scoring for quality is 5.65 and the maximum is 8. Detailed descriptive statistics; including mean, std, and maximum values for all variables in the dataset have been presented in Cell 4 of the attached Jupyter Notebook using the df.describe() function.

**Exploratory Data Analysis and Pre-processing**

For exploratory data analysis (EDA), a number of pre-processing and processing methods were employed. These methods allowed the identification of patterns and summarizing of the dataset’s main characteristics; the methods are described in the following section, along with their interpretation with respect to the business problem.

**Boxplots**

Boxplots allow graphical depictions of the dataset through their quartiles at five data points -lowest, first, median, third and maximum values; boxplots also display any outliers among the features of a dataset. In Figure 1, we present two boxplots of our dataset; [A] is a plot of the variable ‘density’ against the dependent variable ‘quality’, and [B] displays the frequency in the occurrence of outliers in four features: quality, density, pH and citric acid. In general, the mean density was 0.996, and wine samples with ‘quality’ scoring of ≥6 had mean density values close to or less than the mean density value. From Figure 1, the density variable appeared to have more outliers; with occurrence frequency increasing at density values below 0.992 and higher than 1.000 (Figure 1B). Another variable observed with a high number of outliers, from the boxplots is pH; outliers were more at pH values higher than 3.6 and lower than 3.0.



Figure 1: [A] Wine quality versus density, [B] frequency in the occurrence of outliers in four variables – quality, density, pH and citric acid.

**Scatterplots**

Like most data visualization methods used in EDA, Scatter plots allow determination of relationships (correlation) between variables; upon which some conclusions could be drawn. Figure 2 shows three scatterplots of selected variables (‘quality’, ‘density’, ‘alcohol’, ‘pH’, and ‘citric acid’) from our dataset.



Figure 2:Scatter plots of selected wine variables: [A] quality vs density/pH [B] alcohol vs density/quality , and [C] citric acid vs quality.

In Figure 2[A], there appeared to be a correlation (negative? Or Positive?) between wine quality, density and alcohol with most of the wine samples clustering around quality score ‘5’, ‘6’, and ‘7’; colour graduation became lighter with high alcohol and this was sometimes accompanied by lower pH values. Also, there seemed to be an increase in density with lower alcohol, and hence darker colouration. There was no noticeable relationship between pH versus quality, density and alcohol. Figure 2[B] considered a plot of ‘alcohol’ vs ‘density’, using colour = ‘quality’; the higher the quality scoring, the lighter the colour and vice versa. Though most of the wine samples clustered around density range 0.996 – 0.998, there appeared to be no discernible relationship between alcohol and density. In Figure 2[C], the plot shows wine with lower quality scoring (< 5) had had darker colour shades, with no apparent correlation between citric acid values and wine quality; though majority of the wine samples had parallel clustering at quality score 5, 6, and 7. The conclusion drawn from the scatter plots was that wine quality had a relationship with the variables ‘density’ and ‘alcohol’.

**Heatmap**

Next, we looked at heatmaps so that we could better visualize, and determine the correlations between the predictors and the outcome variable (‘quality’). At first glance, we see that volatile acidity, total sulfur dioxide and density had weak negative correlation with wine quality. The attribute with the most positive correlation to quality was alcohol; while citric acid and sulphates had weaker positive correlations. The heatmap provided us with more insight on attributes that were more likely to determine wine quality scoring; it was clear that volatile acidity was not a desirable characteristics when deciding wine quality, whereas alcohol is. Nevertheless, the map also enabled us to decide which features were unrelated (e.g. residual sugar), those that are similar and likely to have same effect (e.g. free sulfur dioxide and total sulfur dioxide); these attributes would be dropped prior to modelling.

We also used a Pair plot to further confirm whether our variables were correlated, using ‘quality’ as color hue; the goal was to allow for selection of features that were most related to wine quality scoring. Some insights gained from the Pair plots include: (i) fixed acidity is positively correlated with citric acid and density, but negatively correlated with pH; (ii) there is a negative correlation between density and alcohol, as well as pH and citric acid; (iii) in terms of single variable distribution, quality scores 5, 6, and 7 were predominant for all features in the plots. Figure 3 shows the heatmap for the features in our wine dataset; refer to Cell 17 of the attached Jupyter Notebook for Pair plots.



Figure 3: Heat map for features in wine dataset; wine quality determination most affected by alcohol.

**Histogram plots**

We used a groupby('quality').mean() function to reclassify the predictors and plotted a histogram of obtained mean values against quality. Total sulfur dioxide was the predominant feature across all quality scoring groups; with quality score ‘5’ having the highest value of 55.29 and score ‘3’, with the least value of 24.5. Another feature of interest was alcohol; the mean values for the scoring groups were in close range, though wine with the predominant quality score ‘5’ had mean alcohol value of 9.90, see Figure 4 for the histogram chart for wine variables. In a different histogram plot (see attached Notebook), the trend for the probability distribution for wine quality scoring was displayed as: score 3 < score 8 < score 7 < score 6 < score 5.



Figure 4: Histogram plot of mean values of wine variables versus wine quality.

Given the above insights from the EDA conducted, the following attributes ……. were selected for the modelling part of the project.

**Modelling - Linear Regression Analyses**

Different linear regression methods were used to predict the wine quality based on the wine physio-chemical properties. The ridge regression calculated R-Squared as low as 0.326 and RMSE = 0.65. Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.

On the other, Lasoo regression calculated R-Squared equal to 0.304 and RMSE= 0.665. Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

**Business Question 2:** How many attributes of wine are needed to predict whether their

quality score will be low or high?

**Dataset overview**

**Conclusion**

**References**

*Wine Quality Dataset*. Kaggle. https://www.kaggle.com/yasserh/wine-quality-dataset