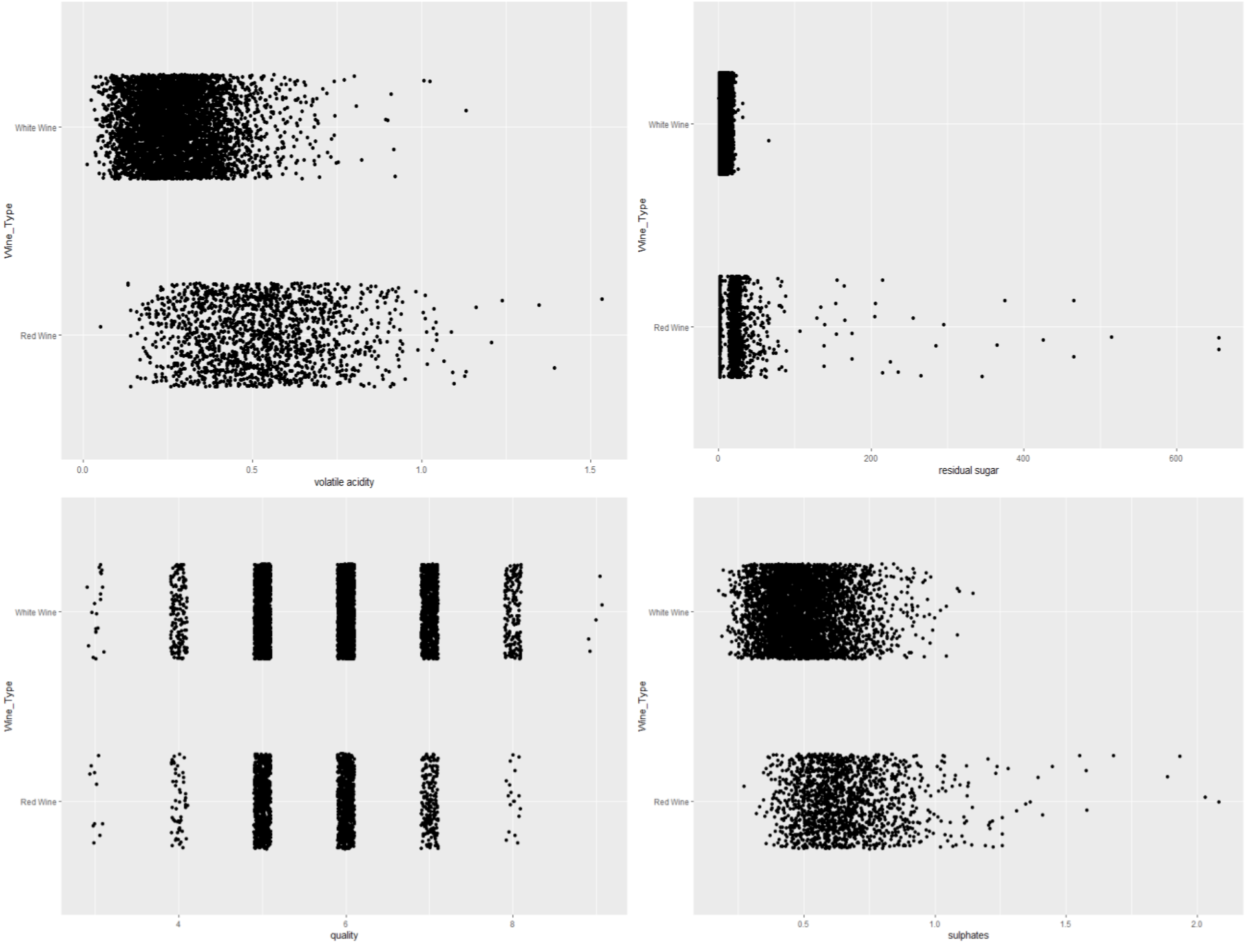
INDIVIDUAL ASSIGNMENT-5

Problem Statement and Data:

Red Wine and White Wine are the two most consumed wines globally. Each wine exhibits a distinct smell, taste and color which is influenced by the raw material and fermentation process used to make these wines. A particular wine exhibits unique characteristics based on the concentration of different factors like ‘Acidity’, ‘Sugar’, ‘Sulfur Dioxides’, ‘Density’, ‘pH’, ‘Alcohol’, ‘Quality’ etc. The variation in proportion of these factors distinguish a red wine from a white wine.



A jitter plot of four random features is plotted. The first plot indicates that a wine is highly likely to be white if volatile acidity is low. Second plot shows that higher residual sugar could make a wine red. Third plot shows that both type of wines shows some discrete qualities and are equally likely to be red ow white. The fourth plot indicates that lower sulphates in a wine would likely make it white.

The current study would try to predict if a particular wine sample is a red wine, or a white wine based on different wine features using logistic regression. The prediction is modelled using the datasets related to red and white vinho verde wine samples, from the north of Portugal.

Planning:

The two datasets contain samples for Red and White wines. A column containing the factor ‘Red Wine’, or ‘White Wine’ was added in both datasets and the data types of different columns was corrected. The datasets were then merged, null values were removed, and outliers were removed for analysis.

INDIVIDUAL ASSIGNMENT-5

The analysis would try to predict if the outcome will be a ‘Red Wine’ (Factor = 0) or ‘White Wine’ (Factor =1). The predictor variables are ‘fixed acidity’, ‘volatile acidity’, ‘citric acid’, ‘residual sugar’, ‘chlorides’, ‘free sulfur dioxide’, ‘total sulfur dioxide’, ‘density’, ‘pH’, ‘sulphates’, ‘alcohol’ and ‘quality’. Multiple logit regression models will be created using these outcome and predictor variables and a model with all significant features will be finalized and verified for assumptions.

Analysis:

The first model was built by including all the features and few variables like ‘Citric Acid’, ‘Density’, ’pH’ and ‘Alcohol’ were insignificant. These variables were removed, and a new model was constructed using ‘Volatile Acidity’, ‘Residual Sugar’, ‘Total sulfur dioxide, ‘Sulphates’, Quality and ‘Free sulfur dioxide’ which were all significant. The 95% confidence intervals were calculated to give β estimates. The β estimates were then transformed to odds ratio by taking exponent. The odds ratio of intercept is between 246.8 and 35443.5. This means there is a significant difference between the odds of a White or Red wine in general, at the 5% level of significance. The coefficient for ‘volatile acidity’ is 5.16E-08 to 7.75E-06, which is way less than 1. This means that, with the presence of volatile acidity, the wine is less likely to be White (specifically, 5.16E-08 to 7.75E-06 times less likely), at the 5% level of significance. Similarly, ‘residual sugar’ is between 0.7 to 0.78, ‘sulphates’ between 9.42E-07 to 1.05E-04 and ‘free sulfur dioxide’ between 0.91 to 0.98. On the other hand, ‘total sulfur dioxide’ is between 1.067 to 1.093, hence a wine is more likely to be white if there is presence of ‘total sulfur dioxide’ (specifically, 1.067 to 1.093 times more likely). Similarly, ‘quality’ is between 1.15 to 2.21, and a wine is more likely to be white if a wine exhibits higher quality.

Nagelkerke’s R_N^2 (pseudo R-sq) yielded a value of 0.94 or 94%, which indicates an excellent goodness of fit of the predictive model on accurately predicting the wine type.

All significant variables in the final model approximately satisfy the linearity assumption as indicated by the scatter plot between predictor variables and the logit of the outcome (Appendix-1).

Most of the significant variables have a VIF value less than 1.3, except the sulfur dioxides which have a VIF of 2.3 and 2.7. These values indicate that there is moderate correlation, but does not significantly affect our results, since the values do not exceed 5. The model is free from significant multicollinearity.

The errors or residuals are correlated as indicated by the Durbin-Watson test; hence the errors are not independent of each other and could affect the model’s accuracy.

Conclusion:

A white wine significantly differs from the red wine and a particular wine sample could belong to a type of wine depending on the presence of several features. The study predicted the significant features which distinguishes a white wine from a red wine and found that the presence of higher ‘Total Sulfur Dioxide’ and ‘Quality’ would increase the chances of a wine being white, whereas the presence of higher ‘Volatile Acidity’, ‘Residual Sugar’, ‘Sulphates’ and ‘Free Sulfur Dioxide’ would increase a wine’s chances of being red. The prediction model serves as a good guide for predicting the likeliness of an unknown wine sample belonging to the categories of a white or red wine. Although, the predictive model performs well in making predictions for the available dataset of different wines, its accuracy could be further improved by adding more data samples from wines. The next versions of this study could use more data points, splitting the data into training and test samples, and better accuracy indicators of the predictive model by comparing the predicted results with the test data observations.