Wine Quality Prediction

We are interested in finding out what features can be used to discern a wine quality. In order to analyse the relation between wine quality and these features, we would be using a wine quality dataset provided by UCI Machine Learning Repository [1].

Wine balance is usually determined by the acidity, alcohol content, sweetness and its density [2] and has an effect on wine quality. In the analysis presented, we would be building a **linear regression model**, which can be used as a tool to explain the relationship among wine quality and various features, inculcating the prior knowledge regarding factors mentioned above.

Exploring the dataset

Wine dataset provided at the UCI Machine Learning Repository [1] features white [3] and red wine data [3]. A new dataset is formed by merging these datasets and removing duplicate data (1197 rows). The final data features 11 columns (refer Appendix for summary of central tendency and dispersion of each variable) shared among numerical features; fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol and 5320 rows. There is no missing data in these 5320 observations. The summary of relevant numerical features used in this report is explained in the table 1 below:

Feature Name	Min	Max	Mean	Median	Std dev
volatile acidity	0.08	0.41	0.3411	0.3	0.1682
alcohol	8	14.9	10.55	10.4	1.1859
residual sugar	0.6	65.8	5.048	2.7	4.5002
sulphates	0.22	2	0.5334	0.511	0.1497
density	0.9871	1.039	0.9945	0.9947	0.003

Table 1: Central tendency and measure of dispersion

There are some outliers observed in these variables however these values are not removed as we do not know about the possible ranges of these features. Furthermore, the influence of outliers on regression model fit can be accessed post fitting the linear regression by analysing standardized residual errors.

Quality: it is an ordinal variable, which tells about the median wine quality rating provided by the experts and can take any value from 0 to 10. In this dataset the range for this outcome variable is 3 to 9 [wine quality distribution can be referred from Appendix]. For the purpose of linear regression analysis, it is assumed as a continuous and unbounded numerical variable.

Methods and assumptions :

Hierarchical method is used to select the features, the priori based features are ‘volatile acidity’, ‘residual sugar’, ‘Alcohol’ and ‘density’ (as related by above mentioned wine balance information). As can be seen from the dataset-summary, the features have non-zero variance for each model of linear regression following other assumptions are made :

1. The quality is considered as a predictor variable. Since it is a linear regression model, we assume the quality varies linearly. Furthermore, any predicted value can be interpreted as described by the function f(x) defined below:

f(x) = GIF[x+0.5] for x C [3,9], f(x) = 3 for x <3, f(x) = 9 for x >9 ; GIF is greatest integer function(refer appendix for interpreting the function values)

2. All predictors are uncorrelated with external variables and are not highly correlated with each other.

Taking the above mentioned model assumptions as basis, we can build a regression model using these priori based feature variables.

Regression models: Various regression models based on the selected features are explained below:
volatile acidity, residual sugar, Alcohol and density : The model summary for this model gives the following result :

	Adjusted ΔR^2	B	SE B	t-value	p-value
Model coefficient	0.28				< 2.2e-16
(Intercept)		-25.77	6.03	-4.27	0.0000199
volatile acidity		-1.35	0.07	-18.81	< 2e-16
alcohol		0.38	0.01	32.09	< 2e-16
density		28.11	6.00	4.68	0.00000292
residual sugar		0.002	0.002	0.65	0.517

Table 2 : Base feature based linear regression model summary

Looking at the adjusted R^2 from the above mentioned summary (see Table 2), it is found that the model can be generalized to explain the 28.01% of the total variance on given wine dataset. However, the coefficient of the 'residual sugar' parameter is not significant as p-value is 0.517 which is greater than the 0.05 (for 95% confidence interval).

Going further, the correctness of this model is checked by carrying out the test for assumptions regarding the model residuals and multicollinearity shared among the variables. In order to verify the assumptions of the multicollinearity, VIF test is carried out on the fitted predictive model. It is found that the value of mean VIF is 2 (greater than 1), raising the concern of multicollinearity among variables. To explore it further, the correlation between all the selected variables are calculated as shown in the correlation matrix below (refer appendix for complete correlation matrix).

volatile acidity	residual sugar	density	alcohol	
1	-0.164	0.308	-0.065	volatile acidity
-0.164	1	0.521	-0.305	residual sugar
0.308	0.521	1	-0.668	density
-0.065	-0.305	-0.668	1	alcohol

The figure indicates that the density has a high linear correlation with all the other variables, while other features are not highly correlated with each other. This means that the density can be derived from the other variables. This conclusion is also supported by the literature [4] as well.

This leaves us with the three features: **volatile acidity, residual sugar and alcohol.** In order to test these features and their contribution towards the proportion of the total variance explained, a linear regression model is fitted. Value of the adjusted R^2 for this model is 0.277 with significant linear coefficients.

To dive deep and to improve the variance explained by the base model, in addition to above three selected variables (i.e. volatile acidity, residual sugar and alcohol), we tried all the features available except density (as it showed multicollinearity with other variables) to form a four feature model. The four features for which the best model (based on high value of adjusted R^2 and low value of AIC) is formulated are **volatile acidity, residual sugar, alcohol and sulphates**. The summary and the confidence interval of the regression coefficient of the trained linear regression model is shown below:

	Adjusted ΔR^2	B	SE B	t-value	p-value
Model coefficient	0.29				< 2.2e-16
(Intercept)		2.08	0.11	18.38	< 2e-16
volatile acidity		-1.30	0.06	-20.77	< 2e-16
alcohol		0.35	0.01	38.89	< 2e-16
residual sugar		0.01	0.002	5.46	4.92e-08
sulphates		0.70	0.07	9.91	< 2e-16

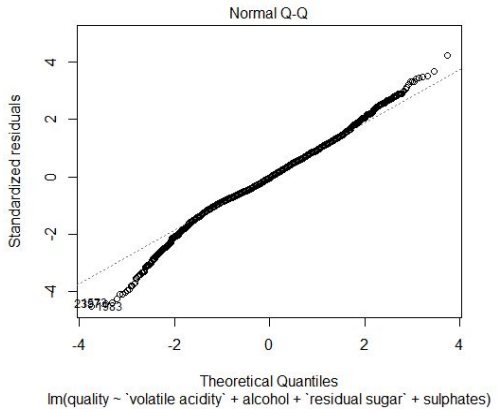
Table 3: Four feature based linear regression model summary

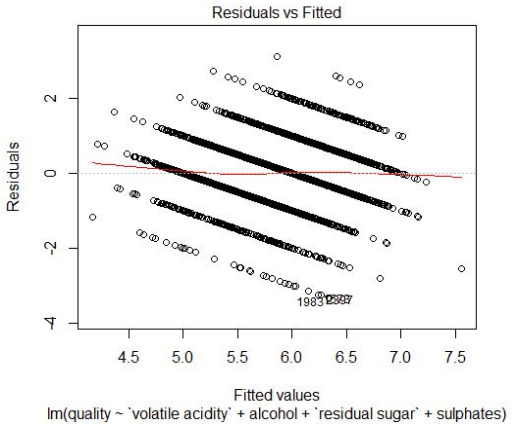
Confidence Interval for Coefficient (B)	2.50%	97.50%
(Intercept)	1.86	2.3
volatile acidity	-1.43	-1.18
alcohol	0.33	0.37
residual sugar	0.008	0.02
sulphates	0.56	0.84

Table 4: CI for model features(4 feature model)

Value of the adjusted R^2 shown in the above model summary (refer Table 3) is 0.29 (and AIC value of 11971.30) which is greater than the above fitted regression model with the three features. Also, this model's feature coefficient values do not cross zero in 95% of the cases as described by table 4 ensuring the significant effect of each variable on wine quality .To ascertain the significance of change in the adjusted R^2 values for these models, both the models are compared using the ANOVA test (F-statistic = 12.53 and p-value = 0.003). As F-statistics turns out to be greater than 1 and is significant (p-value <0.05), it can be said that this model's R^2 value is improved compared to the three features model explained above.

The remaining assumption of linear regression model regarding linearity and homoscedasticity of residuals and multicollinearity among variables are investigated. The value of variance inflation factor (VIF) for each variable lies between 1 and 2 (with mean VIF = 1.11) which indicates that the features are not collinear and do not share high interrelation as mean VIF is close to 1. We can comment on the nature of residuals looking at the plot shown below:

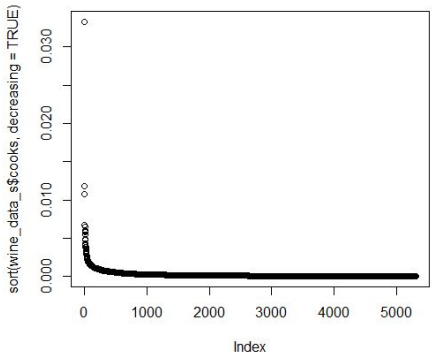




Above Q-Q plot of the standardised residuals reflects the normal distribution of the residuals of the model. Moreover, the plot for Residuals vs. Fitted represents (as shown above) the homoscedasticity and linearity of the residuals as points are approximately uniformly scattered around the linear red line centred at 0. The independence of the residual is checked using the durbin watson test (the results are shown below).Value of test statistics (i.e. D-W statistics) is 1.98 which is very close to 2 with the p-value greater than 0.05. Value of D-W statistics represents no autocorrelation between the residuals following the assumption of independence. The large p-value (0.458) fails to reject the null hypothesis of no autocorrelation and therefore, the model can be generalised. Moreover, the effect of outliers and the influential points are required to be taken into account.

lag	Autocorrelation	D-W Statistic	p-value
1	0.00958151	1.980682	0.458

Alternative hypothesis: rho != 0



The plot on the left shows the cook distance of the residual points with respect to their index. The value of the maximum cook distance is 0.0331 which is very less than 1, explaining the absence of the influential points. Moreover, a total of 304 data samples from 5320 data points (approx. 5.7 % of total samples) lie outside -1.96 to +1.96 Z-score range. These numbers show that residual follows the normal distribution and we need not to remove the outliers from our samples. As this model follows all the assumptions required, it can be concluded that a linear regression model based on **volatile acidity, residual sugar, alcohol and sulphates** can explain 29% of variance in predicting the wine quality. Moreover, If no other parameters are changed, varying volatile acidity in wine by one unit will decrease wine quality by 1.43 to 1.18 (in 95% of the cases), changing the alcohol content in wine by 1 unit will increase the wine quality by 0.33 to 0.37 (in 95% of the cases), varying the residual sugar by a unit seems to have small

effect as it increases the wine quality only by 0.008 to 0.02 (in 95% of the cases) whereas sulphates have the maximum positive effect on the wine quality as it varies outcome variable by 0.56 to 0.84 (in 95% of the cases) against its 1 unit variation in it.

Linear regression models featuring more numbers of features might be able to explain more of the model variance (refer appendix for table featuring variation highest adjusted R^2 values and lower AIC values with respect to number of features used in linear regression models). This is observed as we built the model with five features: **volatile acidity, alcohol, residual sugar, sulphates and total sulfur dioxide** (these five features provide the best adjusted R^2 and AIC values among all combinations of five features). Total variance explained by the 5-feature model explained above is 0.291 (and an AIC value of 11906.51) However, the percentage increase in the adjusted R^2 is 0.24% compared to the variance explained by the improved linear regression model with four features.

On carrying out further analysis, It is found that when the model is fitted with all the **11 features (all variables),** the model is able to explain 0.306 of the variance (and an AIC value of 11804.56) which is 0.016 higher than the variance explained by the improved linear regression model with four features. However, this increase in variance is obtained at the cost of using approximately 3 times more number of variables compared to the four feature models explained above. Therefore, it can be said that the sufficient amount of variance per number of features is explained with a linear regression model when built with four features.

The best achieved explained variance by linear regression (adjusted R^2) is .306 (with all features). This low value of explained variance indicates that there might be some non-linear relationship among wine quality and dataset features. A non-linear method such as Support vector regression (SVM) or neural network might perform better on wine quality prediction (as is described within the dataset description). Moreover, If the problem can be treated as wine quality classification logistic regression, K-nearest neighbours, SVM or neural networks based models might result in better performance.

Conclusion:

We analysed the different features for predicting the wine quality using the linear regression models and have following concluding points:

1. Linear model fitted with volatile acidity, residual sugar, alcohol and sulphates features is able to explain 29.03% of the total variance.
2. With increase in the number of features, the variance explained by the model increases but after a certain number of variables increase in the adjusted R^2 values are not relatively large.
3. Wine data- linear regression model (with four features : volatile acidity, residual sugar, alcohol and sulphates) follows all the assumptions required for the linear regression model which leads to the generalised model. However, different nonlinear models can be used for improving prediction performance on the wine data used.

References:

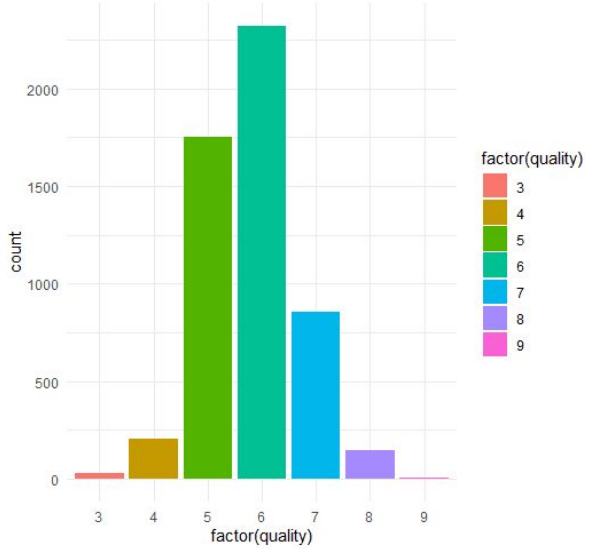
1. *Index of /ml/machine-learning-databases*. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases. [Accessed: 28-Mar-2020].
2. "How to Discern Wine Quality," *dummies*. [Online]. Available: https://www.dummies.com/food-drink/drinks/wine/how-to-discern-wine-quality/. [Accessed: 28-Mar-2020].
3. *Index of /ml/machine-learning-databases/wine-quality*. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/. [Accessed: 28-Mar-2020].
4. B. Coffey, "Why is Wine Density Important?," *ChemWine*, 03-Dec-2019. [Online]. Available: https://www.chemwine.com/home/why-is-wine-density-important-1-sl6yl. [Accessed: 28-Mar-2020].

Appendix

- The measures of central tendency and dispersion for each variable are described below:

Feature Name	Min	Max	Mean	Median	Std dev
volatile acidity	0.08	0.41	0.3411	0.3	0.1682
alcohol	8	14.9	10.55	10.4	1.1859
residual sugar	0.6	65.8	5.048	2.7	4.5002
sulphates	0.22	2	0.5334	0.511	0.1497
density	0.9871	1.039	0.9945	0.9947	0.003
Fixed acidity	3.8	15.9	7.215	7	1.31967
Citric acid	0	1.66	0.3185	0.31	0.14716
Chlorides	0.009	0.611	0.05669	0.047	0.03686
free sulpher dioxide	1	289	30.04	28	17.8054
total sulpher dioxide	6	440	114.1	116	56.77422
pH	2.72	4.01	3.225	3.21	0.16038

- Wine Quality distribution among 5320 samples can be seen from the bar plot plotted below:
- A total of 30 wines are categorized as “3” rated, 206 wines are categorized as “4” rated,1752 wines are categorized as “5” rated,2323 wines are categorized as “6” rated, 856 wines are categorized as “7” rated,148 wines are categorized as “8” rated and 5 wines are categorized as “9” rated.



- The table for interpreting the output of regression model as described by the function definition f(x) is listed below

(f(x) = GIF[x+0.5] for x C [3,9], f(x) = 3 for x <3, f(x) = 9 for x >9)

Quality Predicted Value	Interpretation of Quality
Less than 3.5	3
between 3.5 and 4.5	4
between 4.5 and 5.5	5
between 5.5 and 6.5	6
between 6.5 and 7.5	7
between 7.5 and 8.5	8
Above 8.5	9

- The complete correlation matrix for wine quality dataset is depicted below :

Fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulphur dioxide	total sulphur dioxide	density	pH	sulphates	alcohol	quality	
1	0.215	0.33	-0.104	0.289	-0.282	-0.327	0.478	-0.271	0.305	-0.103	-0.0801	Fixed acidity
0.215	1	-0.384	-0.164	0.368	-0.349	-0.401	0.308	0.247	0.228	-0.065	-0.265	volatile acidity
0.33	-0.384	1	0.146	0.0552	0.131	0.195	0.094	-0.345	0.0592	-0.005	0.098	citric acid
-0.104	-0.164	0.146	1	-0.123	0.399	0.488	0.521	-0.235	-0.175	-0.305	-0.056	residual sugar
0.289	0.368	0.0552	-0.123	1	-0.187	-0.27	0.372	0.025	0.405	-0.27	-0.202	chlorides
-0.282	-0.349	0.131	0.399	-0.187	1	0.72	0.006	-0.142	-0.198	-0.17	0.054	free sulphur dioxide
-0.327	-0.401	0.195	0.488	-0.27	0.72	1	0.006	-0.223	-0.276	-0.249	-0.05	total sulphur dioxide
0.478	0.308	0.094	0.521	0.372	0.006	0.006	1	0.034	0.283	-0.668	-0.326	density
-0.271	0.247	-0.345	-0.235	0.025	-0.142	-0.223	0.034	1	0.168	0.097	0.0397	pH
0.305	0.228	0.0592	-0.175	0.405	-0.198	-0.276	0.283	0.168	1	-0.0172	0.041	sulphates
-0.103	-0.065	-0.005	-0.305	-0.27	-0.17	-0.249	-0.668	0.097	-0.0172	1	0.469	alcohol
-0.0801	-0.265	0.098	-0.056	-0.202	0.054	-0.05	-0.326	0.0397	0.041	0.469	1	quality

- The table below features variation in best adjusted R^2 values and AIC values with respect to number of features used.

Number of features	Adjusted R2	AIC
1	0.22	12415.35
2	0.2753	12026
3	0.28644	11945
4	0.2909	11912
5	0.2966	11870
6	0.3	11839.75
7	0.3029	11824.38
8	0.3038	11818.43
9	0.3057	11804.99
10	0.306	11803.53
11	0.306	11804.58