RED WINE QUALITY ANALYSIS

Matt Zhang
Sophie Lyu
Michael Wu
Michelle Tan

AGENDA

01 BUSINESS OBJECTIVE

02 PROJECT ROADMAP

03 MODEL EVALUATION & SELECTION

04 BUSINESS INSIGHTS

05 LIMITATIONS & IMPROVEMENT

BUSINESS PROBLEM

The market is interested if human quality of tasting can be correlated with wine's chemical properties so that certification and quality assessment and assurance processes are less time consuming and more controlled. This project aims to determine which features are the best quality red wine indicators and generate insights into each of these 12 factors to our model's red wine quality.

ROADMAP

The data collects features of red wine from Vinho Verde, Portugue. Due to privacy and logistic issues, grape type, wine brand, and selling prices are not included. [1]

DATA COLLECTION

- The data includes 1599 samples with 11 independent variables and 1 responsive variable.
- Although the responsive variable follows a normal pattern, most of the independent variables are slightly skewed to the right and contain correlations. [2]

EXPLORATORY DATA ANALYSIS

MODELLING

Both supervised and unsupervised methods are employed in the stage of modelling, which includes K-means, linear regression, logistic regression, decision tree classification, random forests and SVM.

PERFORMANCE EVALUATION

With a focus on accuracy score, other measurements like MSE, RMSE are also considered for model evaluation.

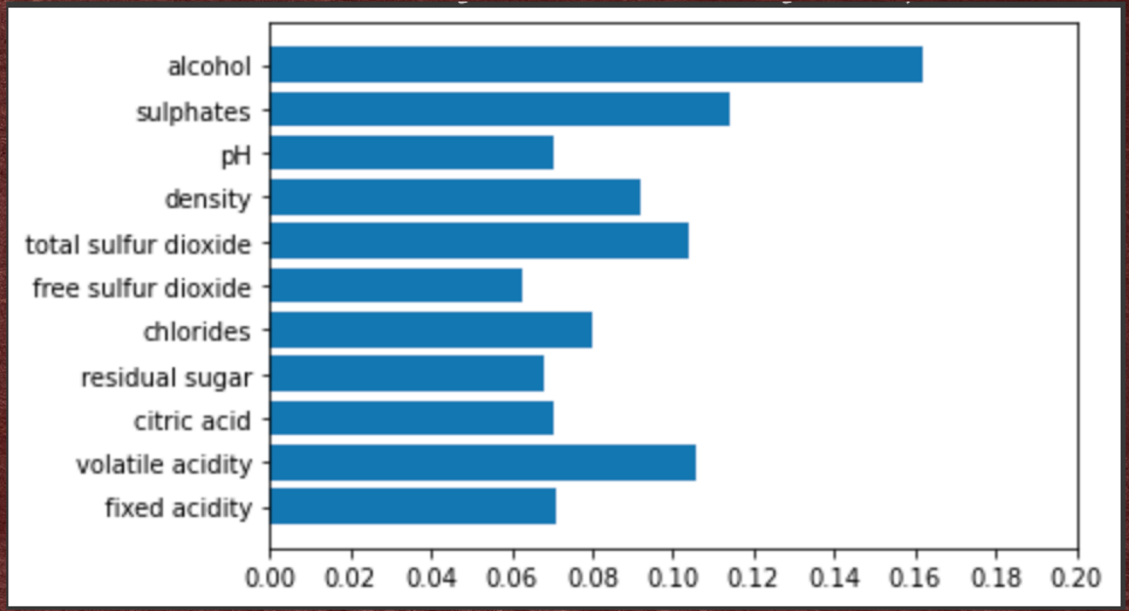
1. https://archive.ics.uci.edu/ml/datasets/wine+quality
2. See more details in appendix.

MODEL EVALUATION & SELECTION

	Accuracy	CVAccuracy	MAE	MSE	RMSE
Decision Tree	0.58	0.57	0.456	0.527	0.726
SVM	0.54	0.54	0.690	1.269	1.126
Random Forest	0.67	0.68	0.356	0.402	0.634
Neural Network	0.57	0.582	0.742	1.458	1.208
Linear Regression	NA	NA	0.520	0.430	0.656

- The Random Forest method implemented gives us the best accuracy score.
- The MAE, MSE, RMSE of the Random Forest are smaller than the other methods implemented.
- Thus, the Random Forest Model is selected as the model to predict the wine quality.

FEATURE IMPORTANCE



With the Random Forest Model, the importance of each variable is calculated. Alcohol, sulphates, volatile acidity are the most relevant features out of 10 features we have.

SUMMARY & BUSINESS INSIGHTS

AUTOMATION

Brings Revenue

MANUFACTURERS

- Better control and increase wine quality
- Better classify wine categories and identify customer segments

CERTIFICATE AGENCIES

- Reduce manpower expense
- Facilitate certification assessment and assurance processes

LIMITATIONS & IMPROVEMENT

DATA INSUFFICIENCY

Solution:
Data Augmentation

Takes the pre-existing samples and generate more training samples; includes more relevant data features, like the year of harvest, brew time, etc

DATA IMBALANCE

Solution:
Data Recollection

Collects more data from other wine datasets but with similar attributes distribution and complex qualities distribution.

DATA INACCURACY

Solution:
Data Trace-Back

Trace back to several factors, including human errors, data drift, and data decay of the original data


THANK YOU

APPENDIX1 : CHECK THE SOURCE DATA

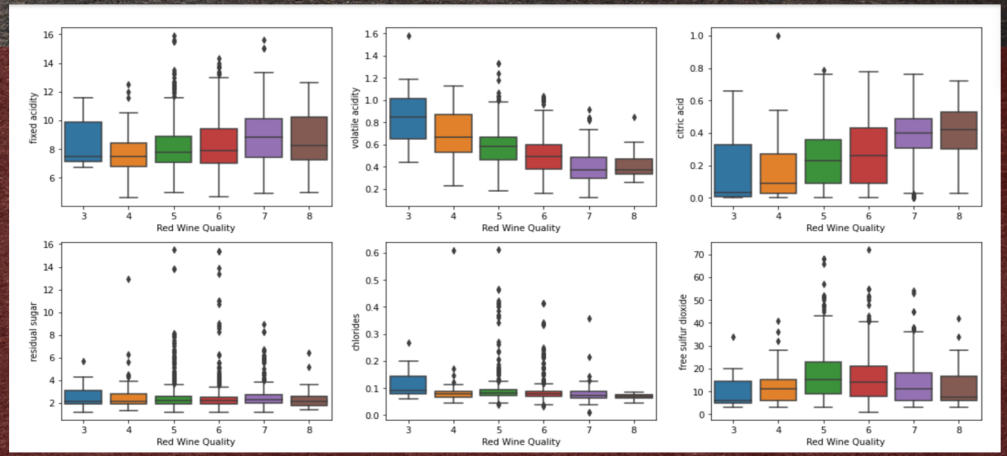
RED WINE RAW DATASET (https://archive.ics.uci.edu/ml/datasets/wine+quality)

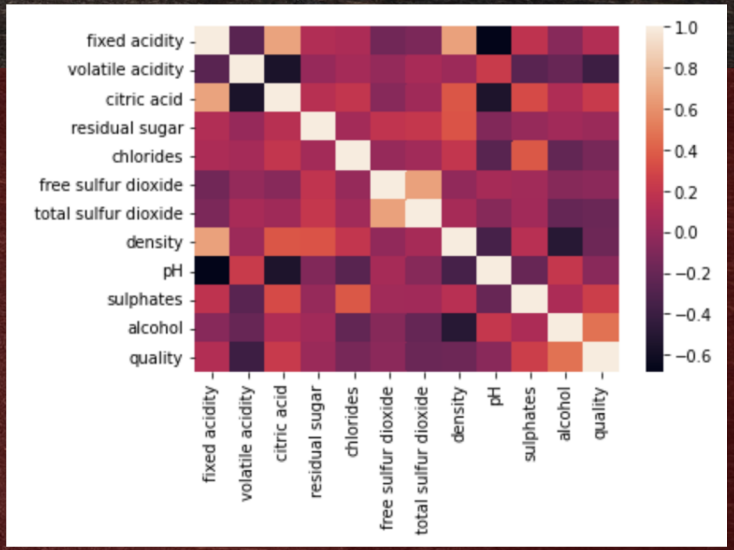
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

DATA DISTRIBUTION

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

APPENDIX 2: EXPLORATORY DATA ANALYSIS

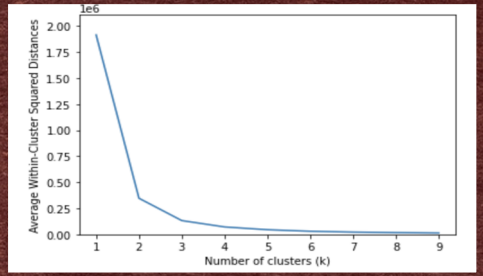




OUR DISCOVERY

1. Most of the data have right skewness.
2. There are relatively strong correlation between quality and volatile acidity, citric acid, chlorides, free sulfur dioxide, sulphates, alcohol.

APPENDIX 3: UNSUPERVISED MODEL



K-MEANS

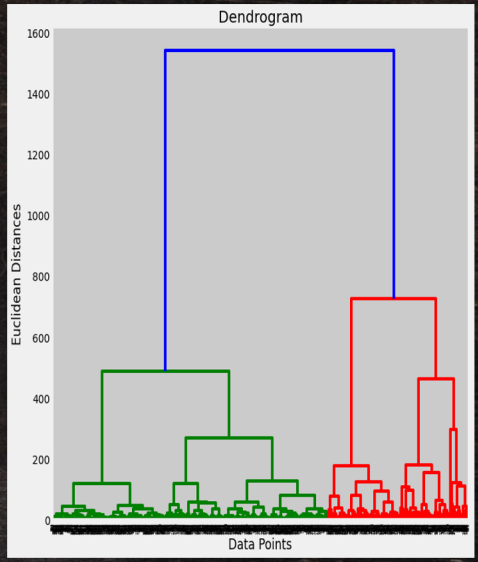
- Elbow method: Sum of Squared Error or Within-cluster sum of square (WSS)

- Computes Silhouette a point is similar to its own cluster ot to other clusters.

Both of the methods for K-Means indicate having two clusters as the optimal choice

DENDROGRAM

- shows the hierarchical relationship between objects work out the best way to allocate objects to clusters



Dendrogram also indicates two clusters.