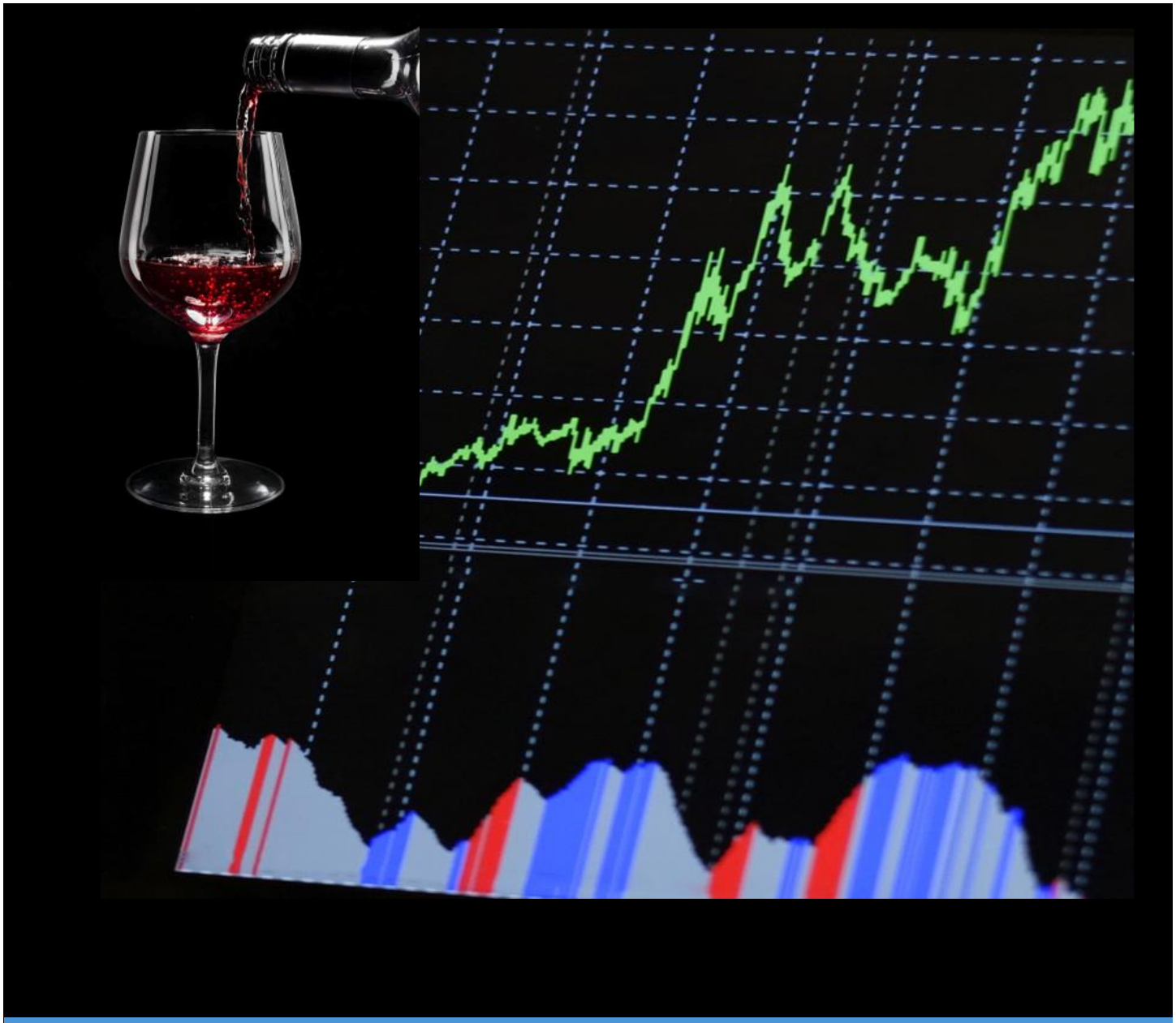

Factors Shaping Red Wine Quality; Insights from Data Exploration

NOROFF

Florence Braut

12/11/24

Data Analysis

Table of Contents

INTRODUCTION.

This report analyses the **Red Wine Quality dataset** (Cortez et al., 2009), which captures detailed information about the physicochemical properties of red wine samples and their corresponding quality ratings. The objective is to uncover meaningful insights from the dataset, such as identifying trends and patterns that influence wine quality.

By understanding these relationships, the report aims to develop a data-driven method for predicting wine quality. This prediction is crucial for informing purchasing decisions, especially considering the significant financial investment required for each pallet of wine.

A systematic approach ensures that only high-quality wines are selected, ultimately supporting the restaurant chain's reputation and customer satisfaction.

EXPLORATORY DATA ANALYSIS.

Data Collection

The dataset used in this analysis was sourced from the **UCI Machine Learning Repository** (Cortez et al., 2009) and made available on Kaggle (https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009). It was subsequently downloaded and imported into Excel for further analysis.

Dataset Information

This report examines the Red Wine Quality dataset, which contains physicochemical measurements of **1,599 red wine samples** along with their sensory quality scores. The dataset provides a unique opportunity to explore the relationships between chemical attributes, such as acidity, PH, alcohol, and sugar levels, and their corresponding quality scores, which are rated on a scale from **3 to 8.**

Data Cleaning

To ensure the dataset is accurate, consistent, and ready for exploration, the following cleaning tasks were performed:

- **Missing Data**: The dataset has no missing values, ensuring full completeness.
- **Correcting Data Types**: Volatile Acidity, Chlorides, and Density were adjusted the data types to ensure they align with the appropriate format were formatted to display 2 or 3 decimal points for precision respectively.
- **Handling Duplicates**: After reviewing the data, 240 duplicate rows, identified and removed, reducing the dataset to **1,359** unique records. This ensures only work with distinct data points.
- **Outliers**: Some variables contained values that could be considered outliers. However, after further inspection, it was determined that these outliers fall within plausible ranges for wine characteristics and do not pose a risk to the analysis. Only 2 rows with extreme values in

Total Sulfur Dioxide were removed, bringing the final count to **1,357** rows. (figure 1)

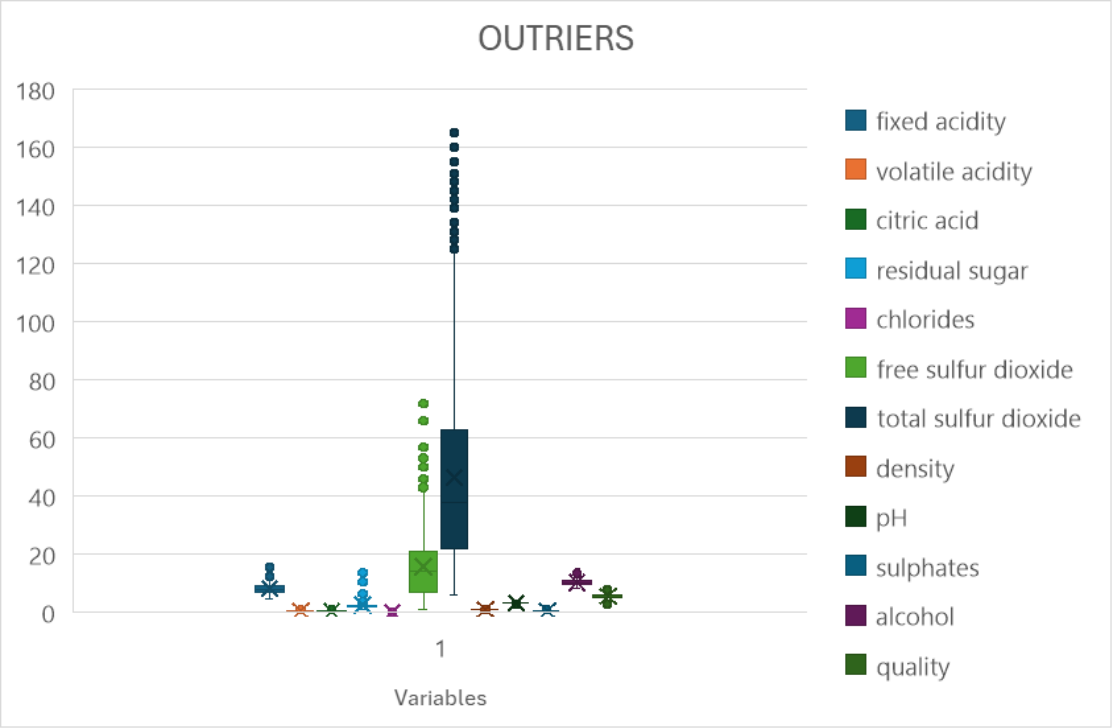

Figure 1: Boxplot highlighting outliers across wine quality-related variables

Initial Assumptions and Hypotheses

The red wine quality dataset includes several physicochemical properties that may influence wine quality. Based on industry knowledge and previous studies in winemaking, I hypothesize that certain factors, such as alcohol content and volatile acidity, have a significant impact on the final wine quality. (figure 2)

Assumptions:

- Higher alcohol content might be associated with higher wine quality.
- Lower volatile acidity might be led to higher quality wine.
- Lower chlorides and fixed acids might be led to higher quality wine.

- High residual sugar levels decrease the quality wines.
- The balance between sulfur dioxide levels and quality is critical.
- There might be optimal ranges for properties like pH and density that correspond to higher quality wines

Figure 2: These graphs illustrates the initial assumptions and hypothesis for wine quality

Hypotheses:

- Higher alcohol content might be correlates with better wine quality.
- Lower volatile acidity might be positively correlated with better wine quality.
- High residual sugar might be negatively impacts the quality of wines.
- High total sulfur dioxide levels are detrimental to wine quality.
- Higher pH values are inversely related to quality.

Data Analysis

Summary Statistics

To understand the characteristics of the red wine dataset, key summary statistics were calculated for each variable, including mean, median, minimum, maximum, and standard deviation. These metrics provide insights into the central tendencies, variability, and range of physicochemical properties, as well as the quality scores of the wines.

Variables	Mean	Medium	Min	Max	STD Dev
fixed acidity	8.31	7.9	4.6	15.9	1.7375581
volatile acidity	0.53	0.52	0.12	1.58	0.1828914
citric acid	0.27	0.26	0	1	0.19498055
residual sugar	2.51	2.2	0.9	15.5	1.33448378
chlorides	0.09	0.08	0.01	0.61	0.04943402
free sulfur dioxide	15.86	14	1	72	10.41688
total sulfur dioxide	46.48	38	6	165	32.1602308
sulphates	0.66	0.62	0.33	2	0.17063415
alcohol	10.43	10.2	8.4	14.9	1.08009349
density	1.00	0.997	0.99	1.004	0.00186646
pH	3.31	3.31	2.74	4.01	0.15466518
quality	5.62	6	3	8	0.82218178

Table 1 The screen shot showing statistics summary.

The summary statistics (table 1) reveals moderate variability in physicochemical properties. For example, fixed acidity ranges from 4.6 to 15.9 g/L, with an average of 8.31 g/L, indicating that most wines are moderately acidic. Volatile acidity, which impacts sourness, has a mean of 0.53 g/L, but some samples reach 1.58 g/L, potentially contributing to lower quality.

Alcohol content ranges from 8.4% to 14.9%, with higher alcohol levels generally associated with better wine quality. Total sulfur dioxide, a preservative, has a wide range (6–165 ppm), but most wines remain within

acceptable limits for red wines (<100 ppm). The average quality score is 5.62, with most wines rated between 5 and 6, indicating a predominantly average-quality selection.

Feature Distributions (in Figure 3)

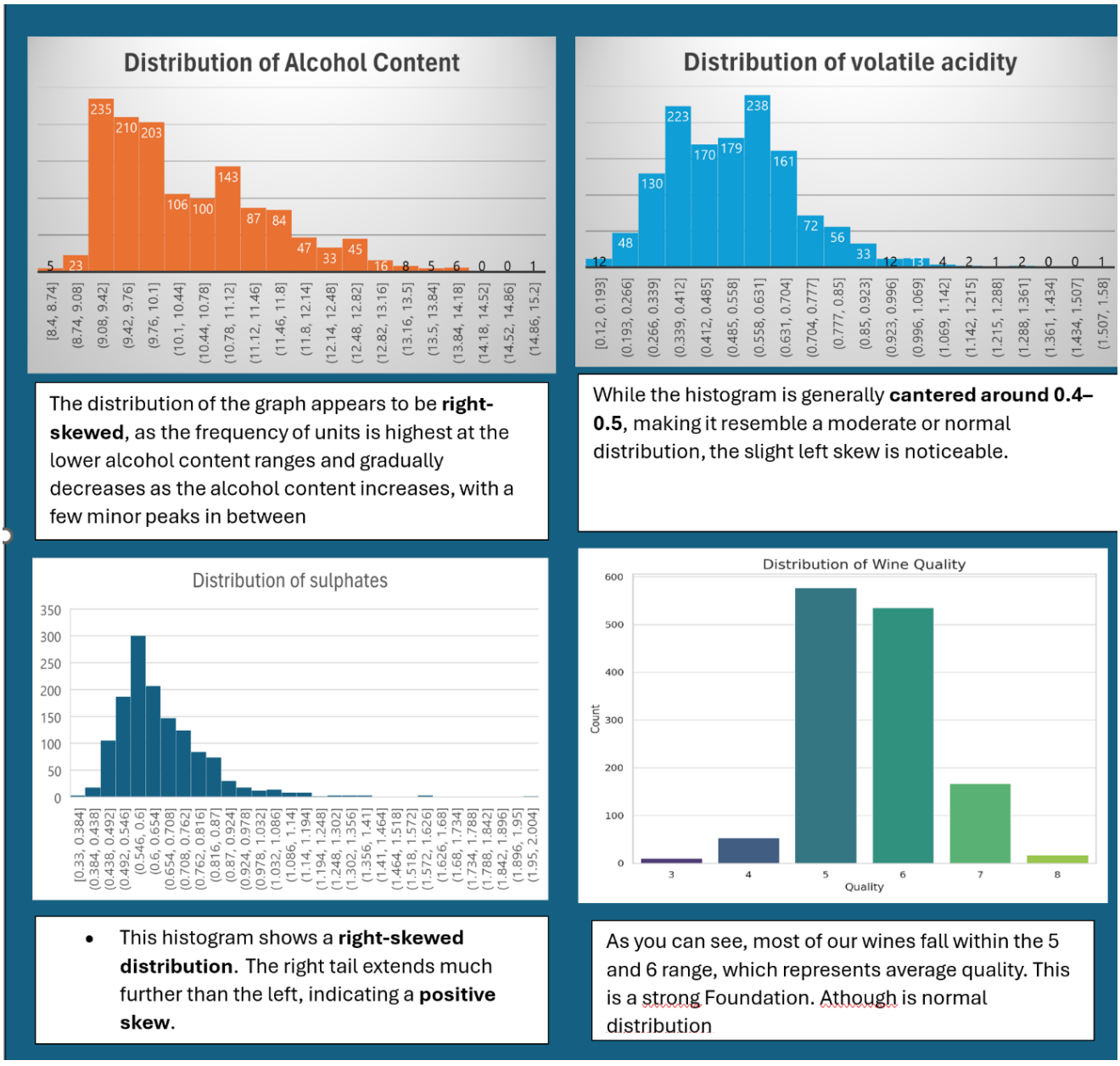

Figure 3 Screenshot showing the distribution of key attributes for wine quality.

Correlation Analysis

	quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	ee sulfur dioxid	tal sulfur dioxid	sulphates	alcohol	density	pH
quality	1.00											
fixed acidity	0.12	1.00										
volatile acidity	-0.39	-0.26	1.00									
citric acid	0.22	0.67	-0.55	1.00								
residual sugar	0.00	0.11	0.01	0.13	1.00							
chlorides	-0.13	0.08	0.06	0.21	0.03	1.00						
free sulfur dioxid	-0.06	-0.14	-0.02	-0.05	0.15	0.00	1.00					
total sulfur dioxic	-0.20	-0.11	0.09	0.03	0.16	0.05	0.67	1.00				
sulphates	0.25	0.19	-0.26	0.33	-0.01	0.39	0.06	0.05	1.00			
alcohol	0.48	-0.06	-0.20	0.10	0.05	-0.22	-0.09	-0.25	0.09	1.00		
density	-0.18	0.67	0.02	0.37	0.34	0.19	-0.02	0.09	0.14	-0.50	1.00	
pH	-0.05	-0.69	0.24	-0.55	-0.07	-0.27	0.06	-0.06	-0.22	0.22	-0.36	1.00

Figure 4 Relationships between key wine quality attributes

Variables	Correlation
alcohol	0.476166
sulphates	0.251397
citric acid	0.226373
fixed acidity	0.124052
residual sugar	0.013732
free sulfur dioxide	-0.050656
pH	-0.057731
chlorides	-0.128907
density	-0.174919
total sulfur dioxid	-0.1851
volatile acidity	-0.390558

Figure 6

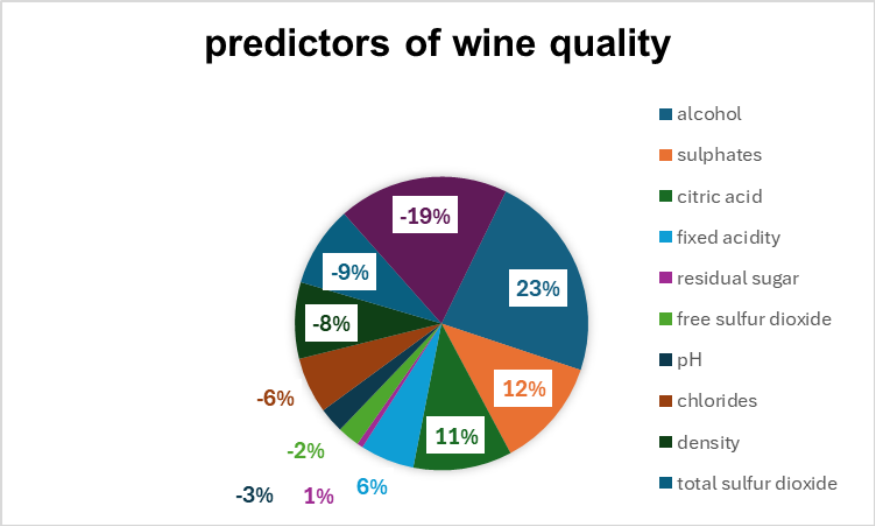

Figure 5 attributes impacting wine quality, arranged by their correlation strength."

A correlation matrix (Figure 4, clearer in figure 5 & 6) reveals:

- **Key Predictors**: Alcohol, sulphates, and citric acid are the most promising predictors of higher-quality wine.
- **Negative Influencers**: Volatile acidity should be carefully monitored as it has the most significant negative impact on quality.
- Other variables, while weakly correlated, may still contribute to quality when combined in a multivariate model.

Relationship Between Features and Quality

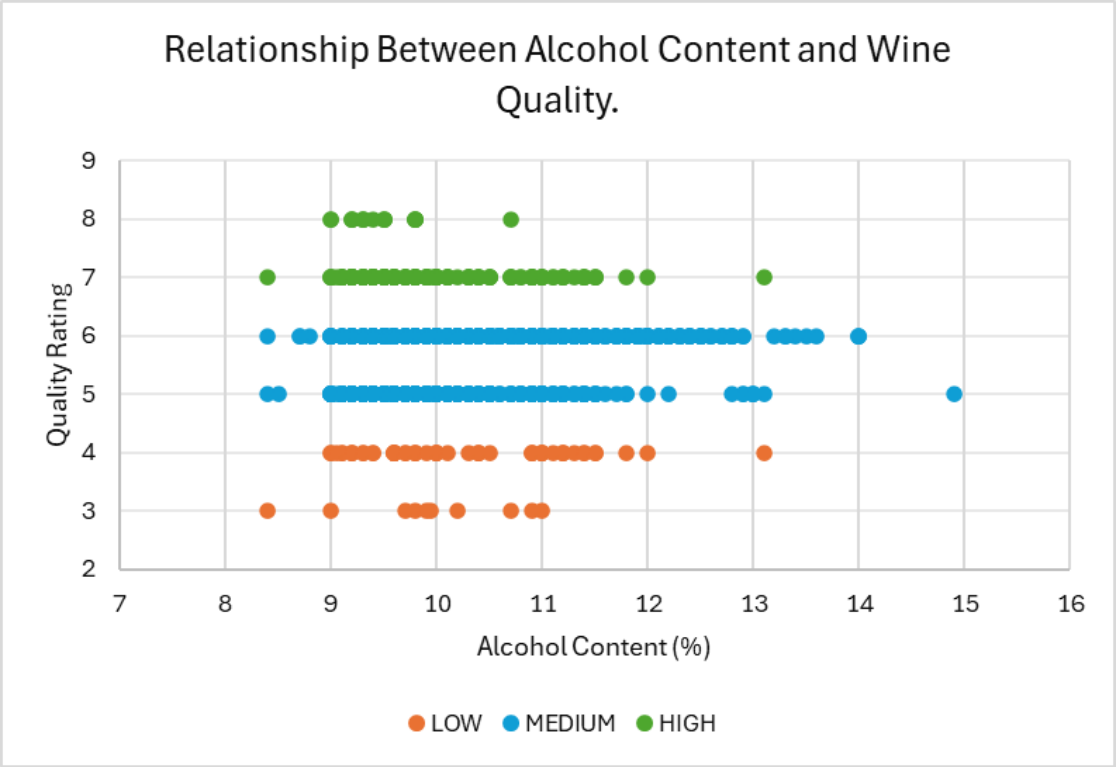

Figure 7 This chart shows relationship volatile acidity to wine quality

Scatter plots (Figure 7) reveals:

- Wines with **higher alcohol content** tend to score better in quality ratings. Notice how the green dots cluster in the 6–8 range.
- **Medium-quality wines** (in blue) are spread across a wide alcohol range, with no distinct clustering.
- Wines with **low alcohol content (below 10%)** tend to have lower quality ratings (in orange).

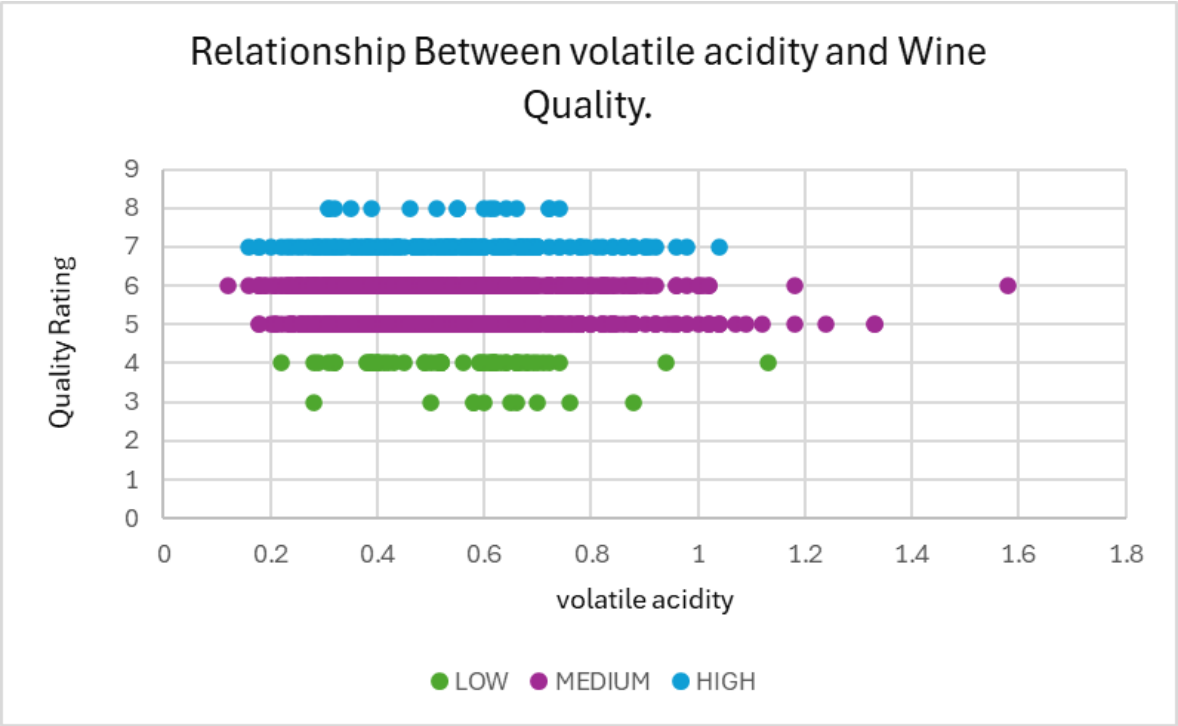

Figure 8 This chart shows volatile acidity to wine quality.

Scatter plots (Figure 8) reveals:

- Wines with high volatile acidity levels (above ~1.2) are predominantly of lower quality (ratings of 3 or 4).
- High-quality wines (ratings of 7 or 8) are more concentrated at lower volatile acidity levels, primarily below 0.6.
- Medium-quality wines (ratings of 5 and 6) are distributed across a broader range of volatile acidity values but tend to cluster below 0.8.
- Low-quality wines (ratings of 3 and 4) are more frequent at higher volatile acidity levels (above 0.8).

Trends, Patterns and Anomalies

These findings help identify factors that influence wine quality ratings.

1. Trends

- Higher alcohol content and moderate sulphate levels consistently correlate with higher wine quality. Conversely, higher volatile acidity and chloride levels are linked to lower quality.

2. Patterns

- Wine quality ratings cluster around medium ratings (5-6), with fewer wines rated as exceptional (7-8) or poor (3-4).

3. Anomalies

- Outliers include wines with very high volatile acidity (>1.0) and chlorides (>0.5), which consistently score lower in quality.

Discussion

ANOVA Test

A one-way ANOVA test was performed to determine if there are significant differences in wine quality based on categorical variables.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
fixed acidity	1357	11278.3	8.31	3.02
volatile acidity	1357	718.96	0.53	0.03
citric acid	1357	368.74	0.27	0.04
residual sugar	1357	3412.7	2.51	1.78
chlorides	1357	119.64	0.09	0.00
free sulfur dioxide	1357	21524	15.86	108.59
total sulfur dioxide	1357	63070	46.48	1,035.04
sulphates	1357	894.16	0.66	0.03
alcohol	1357	14152.92	10.43	1.17
density	1357	1352.545	1.00	0.00
pH	1357	4491.98	3.31	0.02
quality	1357	7628	5.62	0.68

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2560731.6	11	232793.7776	2428.28866	0	1.789235643
Within Groups	1559954.7	16272	95.86742369			
Total	4120686.3	16283				

Table 2 ANOVA results

ANOVA Results (table 2) The results show a highly significant F-statistic of 2428.29 (p-value = 0), which exceeds the critical F-value of 1.789. This indicates that the means of the groups are significantly different.

The results provide strong evidence to reject the null hypothesis, indicating that certain variables significantly influence variations in wine quality. These results validate the need for further exploration of group-level differences through post hoc analyses, to gain a deeper understanding of the relationships between key variables and wine quality.

Variables	Coefficient (2 dec)	p-values	Significance
quality-total sulfur dioxide	45.95	0.00	Significance
quality-volatile acidity	-1.77	0.00	Significance
quality-alcohol	9.90	0.00	Significance
quality-fixed acidity	7.78	0.00	Significance
quality-free sulfur dioxide	15.33	0.00	Significance
quality-pH	2.78	0.00	Significance
quality-residual sugar	1.99	0.00	Significance
quality-density	0.47	0.23	Not Significance
quality-chlorides	-0.44	0.26	Not Significance
quality-citric acid	-0.26	0.51	Not Significance
quality-sulphates	0.13	0.74	Not Significance

Table 3

The summary results (table 3) from the Post Hoc Analysis conducted using XLSTAT, highlight the key attributes that significantly influence wine quality. These findings further validate the importance of these attributes in determining variations in wine quality.

Regression Analysis

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.603987832							
R Square	0.364801302							
Adjusted R Square	0.359606368							
Standard Error	0.658190228							
Observations	1357							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	11	334.635433	30.421403	70.22251488	3.1971E-124			
Residual	1345	582.6733363	0.433214376					
Total	1356	917.3087693						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.978449513	22.73453587	0.218981797	0.826697452	-39.620556	49.577455	-39.620556	49.577455
fixed acidity	0.0058581	0.028019146	0.209074881	0.834421443	-0.04910788	0.06082408	-0.04910788	0.06082408
volatile acidity	-1.104864297	0.12999931	-8.499001222	5.00644E-17	-1.35988775	-0.84984084	-1.35988775	-0.84984084
citric acid	-0.159975156	0.161297064	-0.991804517	0.321471284	-0.47639634	0.15644602	-0.47639634	0.15644602
residual sugar	-8.41724E-05	0.016934741	-0.0049704	0.996034948	-0.03330555	0.03313721	-0.03330555	0.03313721
chlorides	-1.997587139	0.445586679	-4.483049499	7.98418E-06	-2.87170759	-1.12346669	-2.87170759	-1.12346669
free sulfur dioxide	0.004338999	0.002418281	1.794249276	0.072997904	-0.00040501	0.00908301	-0.00040501	0.00908301
total sulfur dioxide	-0.003336091	0.000830293	-4.017968298	6.19532E-05	-0.0049649	-0.00170728	-0.0049649	-0.00170728
sulphates	0.92986044	0.12585579	7.388300874	2.60105E-13	0.682965448	1.17675543	0.682965448	1.17675543
alcohol	0.290329643	0.02849913	10.18731592	1.57459E-23	0.234422064	0.34623722	0.234422064	0.34623722
density	-0.588553341	23.18236533	-0.025387976	0.979749268	-46.066079	44.8889723	-46.066079	44.8889723
pH	-0.473866988	0.208239248	-2.275589217	0.023027319	-0.88237603	-0.06535795	-0.88237603	-0.06535795

Table 4 : this table shows Regression results.

Regression Analysis Summary

The regression model explained 36.48% of the variance in wine quality, as indicated by the R^2 value of 0.3648. The adjusted R^2 of 0.3596, which accounts for the number of predictors in the model, further reflects its generalizability. The overall model was statistically significant ($F = 70.22$, $p < 0.001$), suggesting that the selected predictors collectively have a meaningful impact on wine quality.

Based on the **Model Insights** (table 4), the following summarizes the significant predictors of wine quality, highlighting both positive (yellow) and negative (blue) influences, with p-values less than 0.05 (in red).

Volatile acidity, chlorides, pH, and total sulfur dioxide exhibited significant negative effects, indicating that lower levels of these attributes contribute positively to wine quality. This reinforces the importance of controlling these components during production to maintain or improve wine quality.

On other hand, sulphates and alcohol showed strong positive relationships with wine quality. Higher concentrations of these components were associated with improved wine quality, underlining the benefit of higher alcohol content and sulphate levels.

Additionally, pH displayed a negative relationship with quality, where lower pH (higher acidity) led to a slight reduction in the overall wine rating.

These findings provide valuable insights, emphasizing the need for careful regulation and balance of chemical properties in the winemaking process to optimize wine quality.

Non-Significant Predictors

Some variables, including fixed acidity, citric acid, residual sugar, free sulfur dioxide, and density, did not show a statistically significant impact on wine quality (p-value > 0.05). While these factors may still influence quality in certain contexts, their effects were not robust in this model.

Validation of Initial Assumptions and Hypotheses on Wine Quality

Based on the regression results, i can evaluate whether my initial assumptions and hypotheses hold:

- **Volatile Acidity**: Initially, I might hypothesize that volatile acidity negatively affects wine quality. The regression results support this assumption strongly.
- **Alcohol Content**: If I hypothesized that higher alcohol content would lead to better wine quality, the results validate this assumption.
- **Chlorides and Sulfur Dioxide**: I might have assumed that the presence of certain chemicals like chlorides and sulfur dioxide could negatively impact wine quality. The regression results support this hypothesis.
- Other Components (e.g., pH, fixed acidity): If I hypothesized that other components, such as pH or fixed acidity, would have significant effects on wine quality, the results suggest otherwise. With high p-values for these variables, our assumption about their strong influence is not supported.

In conclusion, while the hypotheses regarding volatile acidity, alcohol content, and chemical components like chlorides and sulfur dioxide hold true, the role of other factors like fixed acidity, citric acid, and pH is not as significant as initially assumed. (See fugure 9)

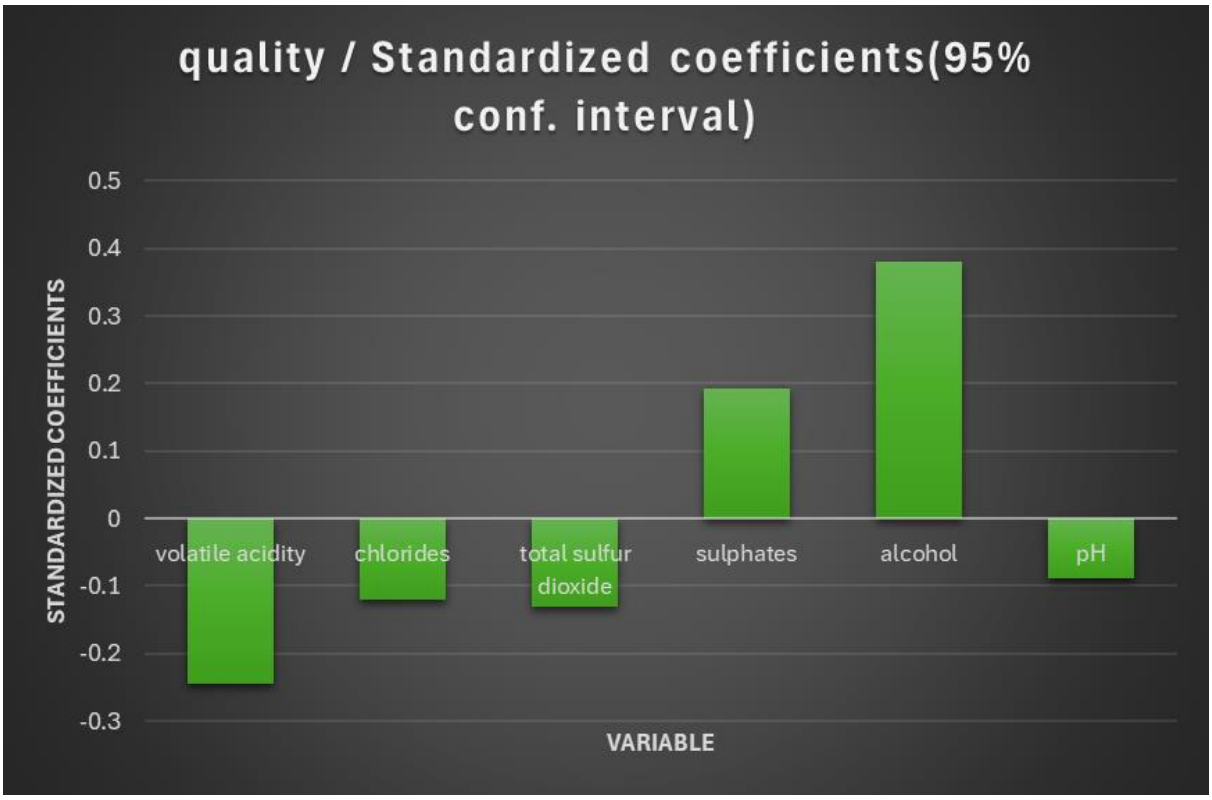

Figure 9. This graph illustrates the final assumptions and hypotheses derived from regression analysis.

Predictive Model for Wine Quality Prediction.

Objective:

To provide insights into the factors affecting wine quality and guide decisions on large-scale wine orders.

This analysis aims to predict wine quality based on physicochemical properties. The goal is to balance accuracy, ensuring managers make informed ordering decisions. Two predictive models were developed:

1. **Significant Attributes Model** (using only the most impactful features).

	WITH 2 DECIMAL POINTS					
Source	***:tual Coefficien***	**prediction valu**	**ERROR**	**MAE**	**RMSE**	**R2**
Atributes	Actual	prediction		0.68	0.91	0.46
chlorides	-2.00	-0.12	-1.88			
volatile acidity	-1.10	-0.25	-0.86			
total sulfur dioxide	0.00	-0.13	0.13			
sulphates	0.93	0.19	0.74			
alcohol	0.29	0.38	-0.09			
pH	-0.47	-0.09	-0.38			

Table 5 : presents the impactful features in the Significant Attributes Model for predicting wine quality.

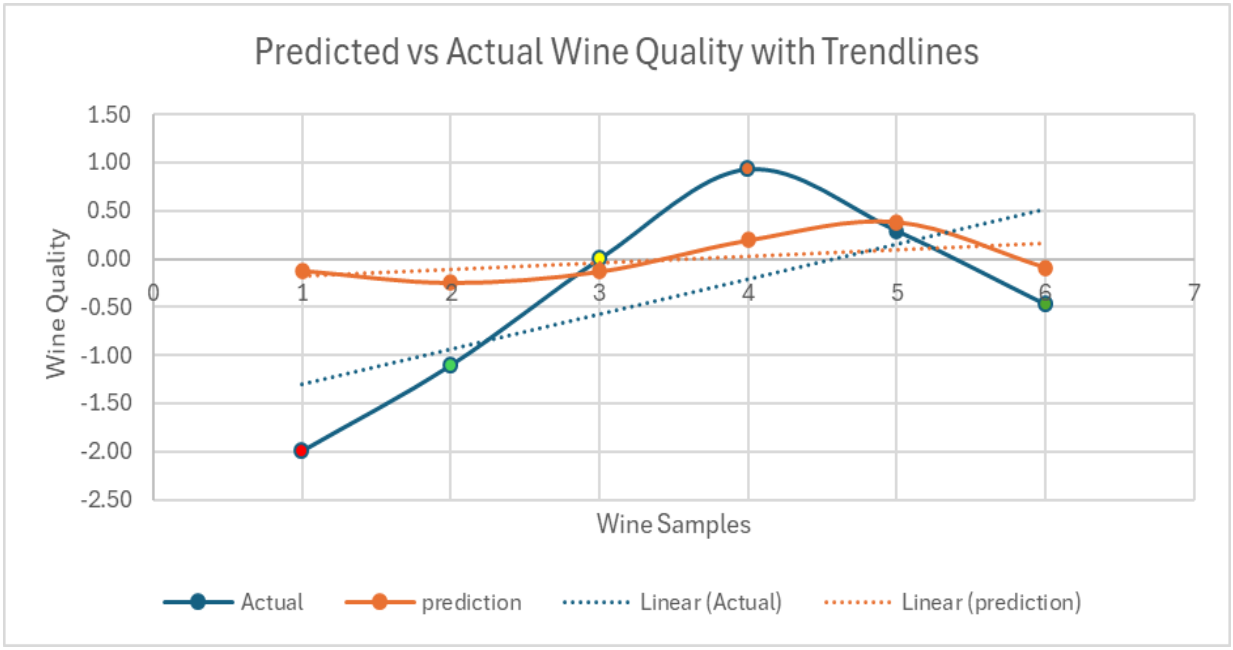

Figure 10 : This graph presents the features in the Significant Attributes Model for predicting wine quality.

Insights from Coefficients.

Significant Attributes Model (table 5, figure 10)**:**

- **Chlorides**: Strong negative impact on quality (−2.00).
- **Volatile Acidity**: Moderate negative impact (−1.10), reflecting poor quality at higher acidity levels.
- **Sulphates**: Positive impact (+0.93), indicating better quality with increased sulphates. This aligning with existing research (Mendes et al., 2019).
- **Alcohol**: Slightly positive effect (+0.29), showing higher alcohol improves quality.
- **pH**: Minor negative effect (−0.47).

All Attributes Model (using all available features).

Source	*ctual Coeffici*	prediction v	ERROR	MAE	RMSE	R2
Atributes	Actual	prediction		0.44	0.70	0.43
fixed acidity	0.01	0.01	-0.01			
volatile acidity	-1.10	-0.25	-0.86			
citric acid	-0.16	-0.04	-0.12			
residual sugar	0.00	0.00	0.00			
chlorides	-2.00	-0.12	-1.88			
free sulfur dioxide	0.00	0.05	-0.05			
total sulfur dioxide	0.00	-0.13	0.13			
sulphates	0.93	0.19	0.74			
alcohol	0.29	0.38	-0.09			
density	-0.59	0.00	-0.59			
pH	-0.47	-0.09	-0.38			

Table 6: This table displays all attributes used in the model to predict wine quality, highlighting the significance of each feature

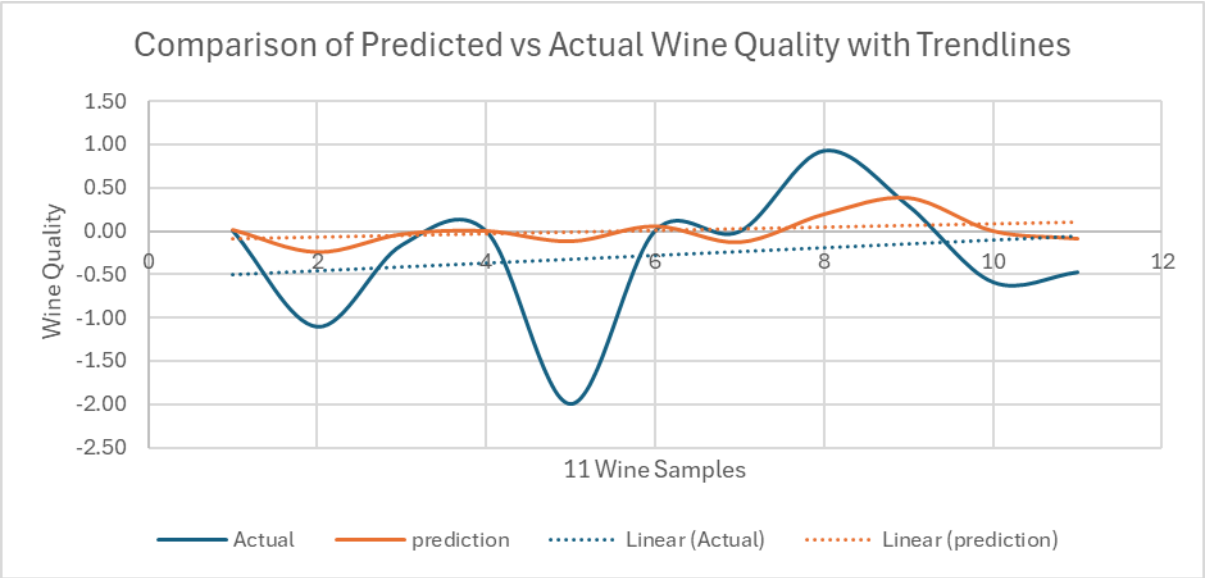

Figure 11

Insights from Coefficients.

All Attributes Model (table 6, figure 11)**:**

- Similar trends as the significant model, with some additional insights:
  - **Density** negatively impacts quality (−0.59).
  - **Citric Acid** and **Residual Sugar** have minimal influence on quality.

Model Performance

Model Type	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	R2R^2R2 (Explained
Significant Attributes	0.68	0.91	0.46
All Attributes	0.44	0.7	0.43

Table 7: presents the MAE, RMSE, and R2R^2R2 values, which evaluate the accuracy and performance.

- **MAE & RMSE** (table 7): All **attributes** model offers better error reduction, meaning its predictions are closer to actual values. This suggests that both can be useful for decision-making (Hastie et al., 2009).
- **R2R^2R2**: Both models perform similarly in explaining wine quality variance, with slightly better results for the significant attributes model.

Recommendations.

- Use the **Significant Attributes Model** for regular orders due to its simplicity.
- For high-value purchases, rely on the **All-Attributes Model** to optimize wine quality and minimize risk.

while the All-Attributes Model is better suited for high-value purchases (Shmueli et al., 2017).

CONCLUSION.

This analysis demonstrates the value of predictive modelling in estimating wine quality based on physicochemical attributes. Key predictors, including alcohol, volatile acidity, and sulphates, play a significant role in quality determination. These insights enable informed purchasing decisions, reducing financial risks. The findings also highlight the potential for data-driven strategies to optimize inventory selection in the wine industry.

Personal Reflection

Reflecting on the journey of compiling this report, I can confidently say that performing Exploratory Data Analysis (EDA) has become much easier for me over time. During my summer holiday, I dedicated a lot of time to studying and familiarizing myself with various datasets and analytical techniques. This commitment has made me feel more comfortable and confident in my ability to handle complex data tasks.

I really enjoy working with datasets now, especially when it comes to uncovering insights that can drive real-world decisions. This experience has not only honed my technical skills but also sparked my interest in pursuing freelance jobs once I receive my results. I look forward to applying the knowledge and skills I've gained to future projects and opportunities.