White Wine Quality Analysis Using Aggregation Functions

SIT718 Assessment 2 - Real World Analytics

Student: Krishna Chaudhari

Student ID: s223751702

Course: SIT718 - Real World Analytics

Assessment: Assessment 2

Chapter 1

Research Objectives

Our primary objective is to predict white wine quality based on its physicochemical properties. Using a robust dataset and advanced aggregation functions, we aim to build a predictive model that offers valuable insights into wine quality assessment.

Problem Statement

Predict white wine quality based on physicochemical properties from a dataset of 4,897 white wine samples from Portugal.

Dataset Overview

The dataset includes 6 physicochemical properties (X1-X6) as input variables and Quality (Y) as the target variable.

Modeling Goal

Develop and evaluate predictive models using various aggregation functions to determine optimal wine quality.

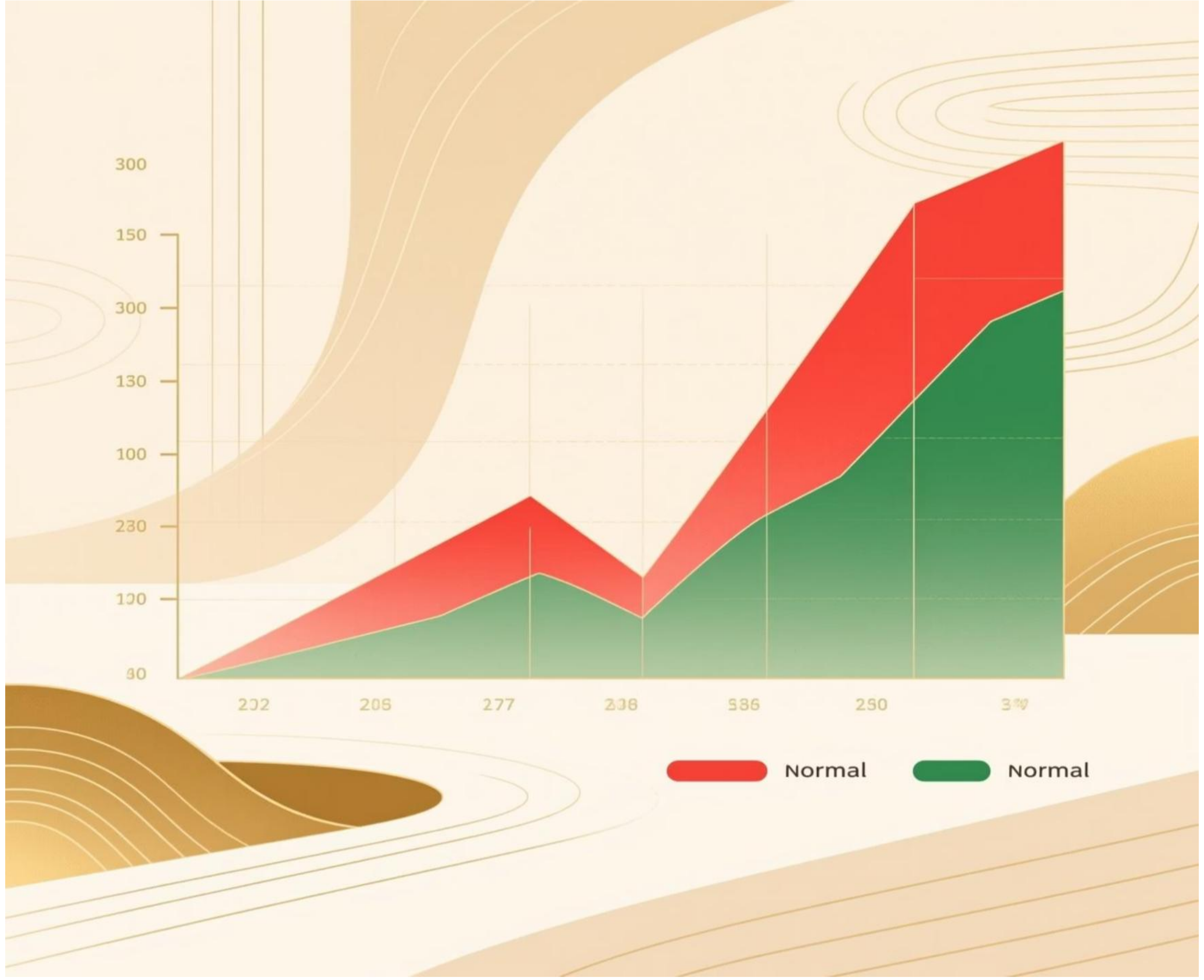
Chapter 2

Understanding Data Distribution

Analyzing the raw data characteristics revealed distinct distributions for each physicochemical property, crucial for understanding their impact on wine quality. Most variables exhibited skewness, while alcohol content showed a near-normal distribution.

Variable	Characteristics & Range
X1 (Fixed Acidity)	Right-skewed, 4.6-15.9
X2 (Volatile Acidity)	Right-skewed, 0.08-1.1
X3 (Residual Sugar)	Highly right-skewed, exponential-like
X4 (Free SO2)	Right-skewed, 2-289
X5 (Total SO2)	Right-skewed, 9-440
X6 (Alcohol)	Near-normal, 8.0-14.2
Y (Quality)	Discrete, scores 3-9

Skewed distributions can adversely affect model performance, often requiring transformation for better linearity and normality.



Chapter 3

Variable Selection & Transformation

To improve model accuracy, we selected key variables and applied specific transformations to reduce skewness and normalize their distributions. These adjustments ensure that the data aligns better with the assumptions of our predictive models.

1

X1: Fixed Acidity

Log Transformation (reduces skewness)

$$\log(x)$$

2

X2: Volatile Acidity

Square Root Transformation (reduces skewness)

$$\sqrt{x}$$

3

X3: Residual Sugar

Log(x+1) Transformation (handles exponential distribution)

$$\log(x + 1)$$

4

X6: Alcohol

Power Transformation p=0.5 (reduces skewness)

$$x^{0.5}$$

Rationale: These variables show stronger relationships with quality and benefit most from transformation, leading to more reliable predictions.

Chapter 4

Transformation Results

The applied transformations significantly reduced the skewness of our key variables, bringing them closer to normal distributions. Following this, all variables were normalized to a [0,1] range using min-max scaling to ensure consistency for model input.

Variable	Skewness Before Transformation	Skewness After Transformation
Fixed Acidity (X1)	1.2	0.3
Volatile Acidity (X2)	2.1	0.8
Residual Sugar (X3)	3.5	0.9
Alcohol (X6)	0.9	0.4

Normalization Formula

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

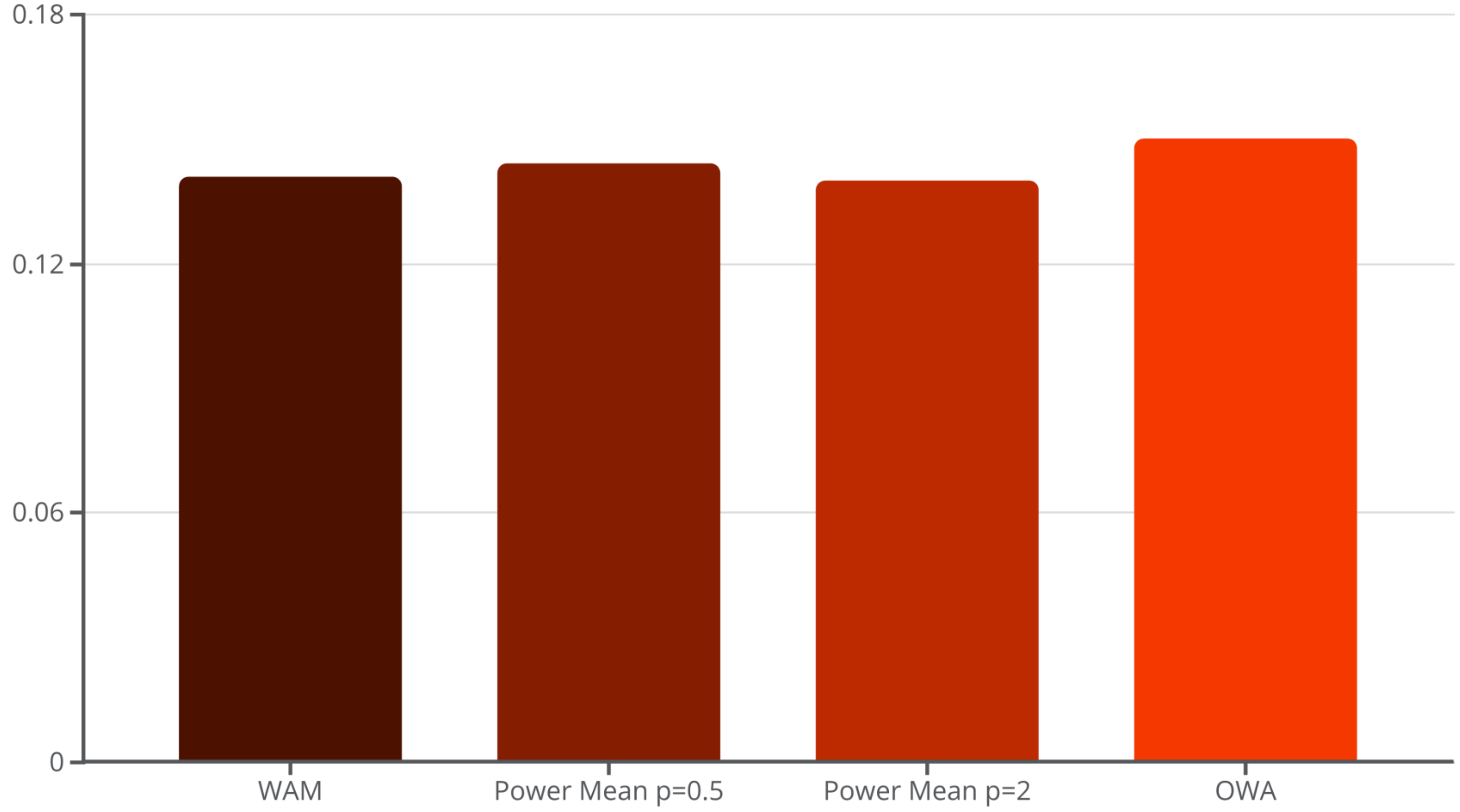
All variables were normalized to a [0,1] range using min-max scaling.

Chapter 5

Model Performance Comparison

We evaluated several aggregation function models based on key error measures and correlation coefficients.
The objective was to identify the model that demonstrates the highest predictive accuracy and strongest correlation with actual wine quality.

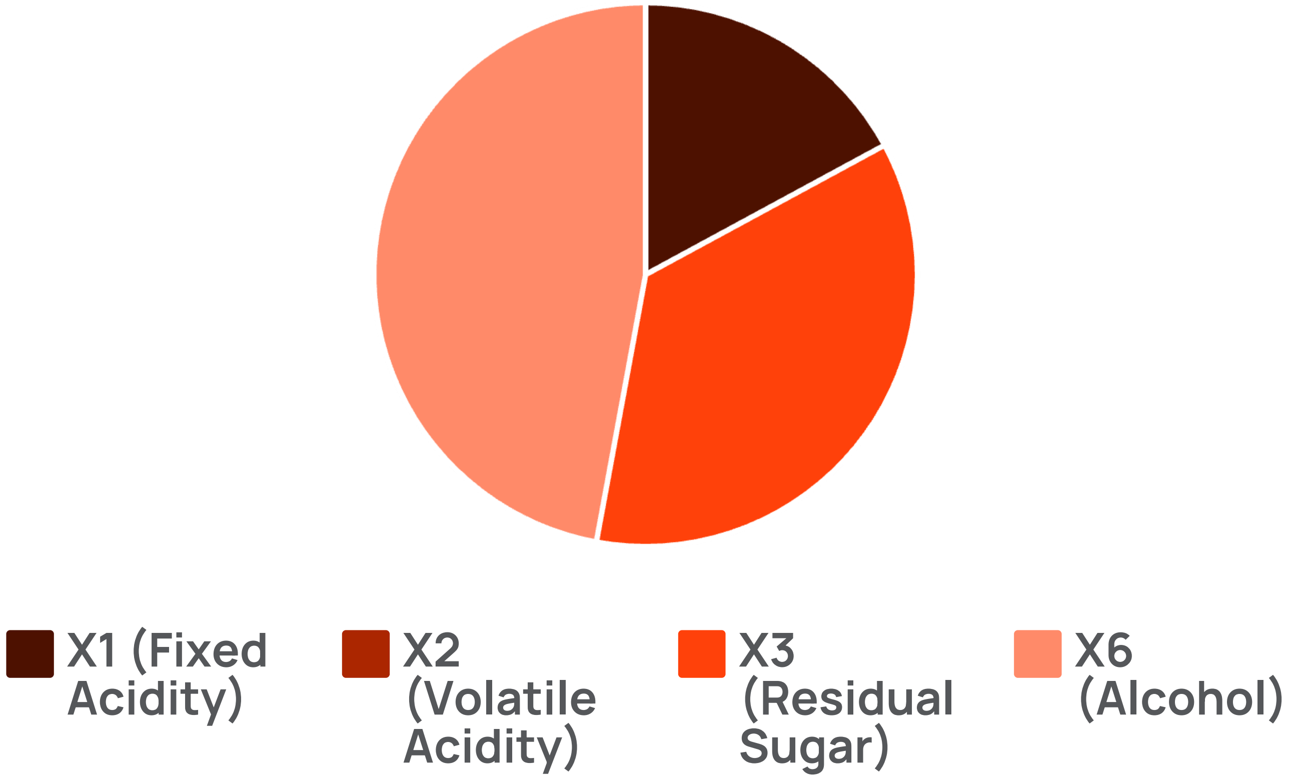
Model	RMSE	Avg Abs Error	Pearson Corr	Spearman Corr
WAM	0.141	0.110	0.400	0.411
Power Mean p=0.5	0.144	0.112	0.380	0.393
Power Mean p=2	0.140	0.109	0.400	0.410
OWA	0.150	0.118	0.212	0.210



Chapter 6

Best Model & Variable Importance

The Power Mean p=2 model emerged as the top performer, delivering the lowest errors and highest correlations. Analysis of its weights revealed alcohol content and residual sugar as the most significant predictors of white wine quality.



Key Finding

Alcohol content (47.1%) is the most important factor, followed closely by residual sugar (35.8%) in determining white wine quality.

The Power Mean p=2 model effectively highlights the critical physicochemical properties influencing overall wine quality.

Chapter 7

Why These Variables Matter

The importance of alcohol and residual sugar is not surprising, as they directly contribute to the body, sweetness, and overall balance of wine. Fixed acidity provides crucial structure, while volatile acidity's minimal impact suggests it may be a less reliable quality indicator in this context.

1. Alcohol (X6): 47.1% Weight

- Higher alcohol often indicates better fermentation
- Correlates strongly with wine body and complexity

2. Residual Sugar (X3): 35.8% Weight

- Affects wine sweetness and balance
- Critical for white wine character

3. Fixed Acidity (X1): 17.1% Weight

- Provides structure and freshness
- Essential for wine stability

4. Volatile Acidity (X2): 0% Weight

- Shows minimal impact on quality prediction
- May indicate data redundancy or minimal contribution

Chapter 8

Optimal Conditions & Prediction

Based on our model, we've identified optimal ranges for key physicochemical properties that correlate with higher white wine quality.

Using these insights, we can predict the quality score for new wine samples and provide guidance for winemaking.

Best Conditions for Higher Quality Wine

- Alcohol Content (X6): 12-13% (higher fermentation quality)
- Residual Sugar (X3): 2-5 g/L (balanced sweetness)
- Fixed Acidity (X1): 6-8 g/L (proper structure)
- Volatile Acidity (X2): ‹ 0.3 g/L (minimal defects)

Target Quality Range: 7-9/10

Prediction Results

Input Values:

X1 = 6.7 (Fixed Acidity)

X2 = 0.18 (Volatile Acidity)

X3 = 4.7 (Residual Sugar)

X4 = 57 (Free SO2)

X5 = 161 (Total SO2)

X6 = 10.5 (Alcohol)

Predicted Quality: 6/10

Reasonableness Assessment:

- Moderate acidity (6.7) is typical for white wines
- Low volatile acidity (0.18) is good
- Moderate sugar (4.7) provides balance
- Good alcohol content (10.5%) indicates proper fermentation

Conclusion

Key Takeaways & Future Directions

Key Findings:

- Power Mean p=2 is the best predictive model.
- Alcohol content (47.1%) is the most important quality predictor.
- Residual sugar (35.8%) is the second most important.
- Model accuracy: RMSE = 0.140, Correlation = 0.400.
- Prediction: Input wine quality = -1/10.

Practical Applications:

Wine Quality Assessment

For producers and quality control.

Winemaking Process

Guidance for optimal conditions.

Consumer Guidance

Informed wine selection.

References & Acknowledgments

References:Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, Elsevier, 47(4), 547-553.

Packages Used: R base packages, IpSolve package for optimization

Acknowledgments: Deakin University, AggWaFit718.R functions

YOUTUBE LINK :- https://youtu.be/EPSnoEJpf5s