Semester Project Report

Of

EDA PROJECTS

INT 353

on

Wine Quality Analysis Report

SUBMITTED TO:

Anjana R Ma'am

Date: 1st July, 2023

To

20th Nov, 2023

SUBMITTED BY:

ABHAY SAHANI

Registration No: 12115826

Introduction

Wine, a centuries-old elixir cherished for its complexity, nuances, and cultural significance, has always captivated the senses and intrigued connoisseurs. In the world of viticulture and oenology, the quest for crafting the perfect bottle is a relentless pursuit, where countless factors converge to define a wine's quality. The Wine Quality Analysis Report embarks on a journey to explore the multifaceted landscape of wine quality, employing the lens of data-driven inquiry.

The profound intersection of nature and artistry in winemaking manifests in the sensory symphony captured within each bottle. Factors such as grape variety, terroir, weather conditions, and winemaking techniques intermingle to create wines with distinctive profiles. At the heart of this study lies the Wine Quality Dataset, a treasure trove of information encompassing chemical attributes and quality ratings of diverse wines.

Domain Knowledge

Wine quality assessment is a critical aspect of the wine industry. Wine quality is influenced by a variety of factors, including grape variety, climate, soil, winemaking techniques, and aging. Wine experts and enthusiasts often evaluate wines based on attributes such as acidity, sweetness, alcohol content, and aroma. The wine quality rating is a subjective assessment, typically ranging from 1 (low) to 10 (high), with higher ratings indicating better quality.

Data Understanding

Structure of the Wine Quality Dataset

The Wine Quality Dataset is a structured dataset that encompasses a rich array of attributes related to various wines, both red and white. Understanding the dataset's structure is essential before diving into the analysis. The dataset consists of the following columns:

1. **Fixed Acidity**: Represents the non-volatile acids in the wine, typically measured in grams per liter (g/L).
2. **Volatile Acidity**: Reflects the volatile acidity, which contributes to a wine's vinegar-like taste, measured in g/L.
3. **Citric Acid**: Denotes the citric acid content in the wine, also measured in g/L.
4. **Residual Sugar**: Indicates the amount of residual sugar left after fermentation, measured in g/L.
5. **Chlorides**: Represents the salt content in the wine, measured in g/L.
6. **Free Sulfur Dioxide**: Measures the free form of sulfur dioxide, an antioxidant and antimicrobial, in mg/L.
7. **Total Sulfur Dioxide**: Represents the total sulfur dioxide content, which includes both free and bound forms, in mg/L.
8. **Density**: Denotes the density of the wine, typically measured in g/cm^3.
9. **pH**: Reflects the acidity or basicity of the wine on a scale from 0 (very acidic) to 14 (very basic).
10. **Sulphates**: Indicates the concentration of sulphates in the wine, measured in g/L.
11. **Alcohol**: Represents the alcohol content of the wine, typically measured as a percentage of alcohol by volume (ABV).
12. **Quality (Wine Quality Rating)**: Provides the quality rating of the wine on a scale from 1 (lowest) to 10 (highest), where higher ratings signify better quality.

Data Overview

The dataset comprises 6497 rows, each representing a distinct wine sample, and 12 columns denoting the wine's attributes and quality rating. The inclusion of both red and white wines adds diversity to the dataset, making it suitable for comprehensive analysis.

Data Types and Statistics

- The majority of attributes in the dataset are of numeric data types, with some represented as decimal values (e.g., alcohol content) and others as integer values (e.g., quality rating).
- The 'Quality' column, being ordinal, is categorical in nature and represents the dependent variable for potential predictive modelling.
- Summary statistics, such as mean, standard deviation, minimum, maximum, and quartiles, offer initial insights into the distribution and central tendencies of the numeric attributes.

Data Quality and Missing Values

- Preliminary assessment indicates that the dataset exhibits overall good data quality. There are no glaring data quality issues, such as missing values or extreme outliers.
- The absence of missing values simplifies data preparation, allowing for more straightforward analysis and modelling.

Dataset Diversity

- The dataset represents a diverse selection of wines, with varying quality ratings, chemical compositions, and characteristics.
- The inclusion of both red and white wines introduces an interesting dimension for comparative analysis.

Reasons for Choosing the Dataset

1. Wine is a popular and widely consumed beverage, making this dataset relevant and interesting for analysis.
2. The dataset offers a diverse set of attributes, allowing for a comprehensive exploration of factors affecting wine quality.
3. Wine quality analysis can provide insights that are valuable for winemakers, sommeliers, and wine enthusiasts.
4. The subjective nature of wine quality ratings provides an opportunity for exploratory data analysis and modelling.

Questions for Analysis

1. What are the names and data types of the columns?
2. . What are the basic summary statistics?
3. Are there any categorical variables and missing values? If so print it
4. Are there any outliers in the data? If so use box plots, histograms and visualize .
5. . Is the data balanced or imbalanced? Visualize
6. What is the target variable (if any) .
7. What are the units of measurement for numerical(example : time , currency ,date, distance)
8. Do you have domain clarification? Brief it
9. Are there any time-based trends or patterns?
10. Are there any correlations between variables? Calculate correlations.
11. Create a histogram or bar plot to visualize the distribution of wine quality ratings. Do certain types or varieties of wine tend to have higher quality ratings?
12. Group the data by wine type or variety and calculate summary statistics or visualize differences in quality ratings. Is there a relationship between alcohol content and wine quality?
13. Analyze the correlation or visually plot the relationship between alcohol content and quality. Are there any seasonal patterns in wine quality or sales?
14. If you have time-related data, create time series plots and analyze whether wine quality or sales follow any seasonal trends. Do different acidity levels (e.g., citric acid, volatile acidity) impact wine quality differently?
15. Analyze the correlations or create scatter plots between acidity levels and quality. What is the average price of wines with different quality ratings?
16. If you have price data, analyze how price varies with wine quality. Are there regional differences in wine quality?
17. If your dataset includes information about the region where the wine was produced, analyze whether wines from certain regions

tend to have higher quality ratings. What is the relationship between pH levels and wine quality?

18. Analyze the correlations or create scatter plots between pH levels and quality. Are there any interactions between variables that affect wine quality?
19. Explore interactions between multiple variables (e.g., alcohol content, acidity) to see if they jointly influence wine quality. What is the average age of the wines in the dataset?
20. Is there a relationship between residual sugar and sweetness perception (quality) of wines?

Git Hub Link:

https://github.com/its-AbhaySahani/Wine-Quality-Prediction-Analysis.git