7COM1079-0901-2024 - Team Research and Development Project

Final report title: (*the topic of your research.)*

Group ID:

Dataset number:

Prepared by: *[Name and ID of submitting student first],*

*[Name and ID of other group members]*

***Please make sure*** *the document spelled correctly (including image labels, section headings, and table of contents). Please use correct punctuation.*

*Make sure your report is grammatically correct.*

University of Hertfordshire

Hatfield, 2024

Contents

[1. Introduction 2](#_Toc186725640)

[1:1 Problem statement and research motivation 2](#_Toc186725641)

[1:2. The data set 2](#_Toc186725642)

[1:3. Research question 2](#_Toc186725643)

[1:4. Null hypothesis and alternative hypothesis (H0/H1) 2](#_Toc186725644)

[2. Background research 3](#_Toc186725645)

[2.1. Research papers 3](#_Toc186725646)

[2.2. Why RQ is of interest 3](#_Toc186725647)

[3. Visualisation 4](#_Toc186725648)

[3.1. Appropriate plot for the RQ 4](#_Toc186725649)

[3.2. Additional information relating to understanding the data 4](#_Toc186725650)

[3.3. Useful information for the data understanding 5](#_Toc186725651)

[4. Analysis 5](#_Toc186725652)

[4.1. Statistical test used to test the hypotheses and output 5](#_Toc186725653)

[4.2. The null hypothesis is rejected /not rejected based on the p-value 5](#_Toc186725654)

[Evaluation – group’s experience at 7COM1079 5](#_Toc186725655)

[5.1. What went well 5](#_Toc186725656)

[5.2. Points for improvement 6](#_Toc186725657)

[5.3. Group’s time management 6](#_Toc186725658)

[5.4. Project’s overall judgement 6](#_Toc186725659)

[5.5. Comment on the GitHub log output 6](#_Toc186725660)

[6. Conclusion 6](#_Toc186725661)

[6.1. Results Explained 6](#_Toc186725662)

[6.2. Interpretation of the results 6](#_Toc186725663)

[6.3. Reasons and/or implications for future work, limitations of your study 6](#_Toc186725664)

[7. Reference List 7](#_Toc186725665)

[8. Appendices 7](#_Toc186725666)

# 1. Introduction

## 1:1 Problem statement and research motivation

The ever-growing wine industry has seen a decline in both consumption and production recently. The growth of the industry relies on the satisfaction of consumers. Thus, the relationship between alcohol content and wine quality is a critical topic in the wine industry, as prior research suggests that alcohol content can significantly impact the sensory characteristics of a wine (Arancha de-la-Fuente-Blanco, et al., 2024). Ensuring the wine industry relishes its glory days again with the help of our research provided us the motivation to figure out whether wine with higher alcohol content has a significantly higher quality rating than then the wines with lower alcohol content.

## 1:2. The data set

Our dataset “wine quality-white” was derived from Kaggle containing a total of 4898 samples of wine along with 12 features next to them. Although our research primarily prioritises on the feature of Alcohol (percentage of Alcohol) and wine quality (rating from 1-10), other features such as density, PH, residual sugar, citric acid etc are also key components of a wine sample which are present in the dataset that can lead many more compelling and insightful research. The dataset is ideal for studying the impact of alcohol content on wine quality.

## 1:3. Research question

To ascertain whether the proportion of alcohol quantity connects to alcohol quality we needed to craft an excellent research question. Therefore, it lead to the research question of –

Is there a difference in the mean quality rating between wines with low alcohol content and wines with high alcohol content?

We plan to answer our research questions through statistical testing revolving around our hypothesis. We want use visualizations such as histogram to test the normality of the data, boxplots to illustrate differences and statistical test such as T-test, Wilcoxon test to evaluate significance. We intend to use the language R through R Studio, utilizing the dataset to explore the effects of alcohol content on wine quality.

## 1:4. Null hypothesis and alternative hypothesis (H0/H1)

Over the course of this assessment, we want to investigate the effect of alcohol content on the quality ratings of white wine. To establish this, we approach it scientifically, establishing two competing hypotheses-

**Null Hypothesis**: There is no difference in the mean quality ratings between the wines with low alcohol content and wines with high alcohol content.

**Alternative Hypothesis**: There is a difference in the mean quality ratings between wines with low alcohol content and wines with high alcohol content.

Though our statistical testing we will be able to evaluate the hypothesis under defined significance level leading us to either reject or accept the null hypothesis.

# 2. Background research

## 2.1. Research papers

Research centring around wine is not a new phenomenon as the earliest production of wine date up to 6000 BC. However, the modern technology provides greater opportunity to research on wine and its elements. So, it is not an unfamiliar event that we found multiple journals that worked with the same dataset as ours.

The research paper titled “Prediction of wine quality using machine learning algorithms (Dahal, et al., 2021)” harnessed our white wine dataset to establish relation between different wine components and the quality using machine learning algorithms. Noteworthy corelation between alcohol content and wine quality was observed. Therefore, it provides us with the basis to study the relation of alcohol with wine quality for different levels.

An article entitled “The influence of Alcohol on the sensory perception of red wines (King, et al., 2013)” established the idea of alcohol content influencing sensory attributes like aroma and taste for the red wines. The article manages to proof the wines containing higher level of alcohols tend to be more intense especially associating with sensory perceptions. Senses are highly corelated with the quality of wine providing us with evidence that alcohol content might influence overall wine quality further solidifying our reason to research on our chosen topic.

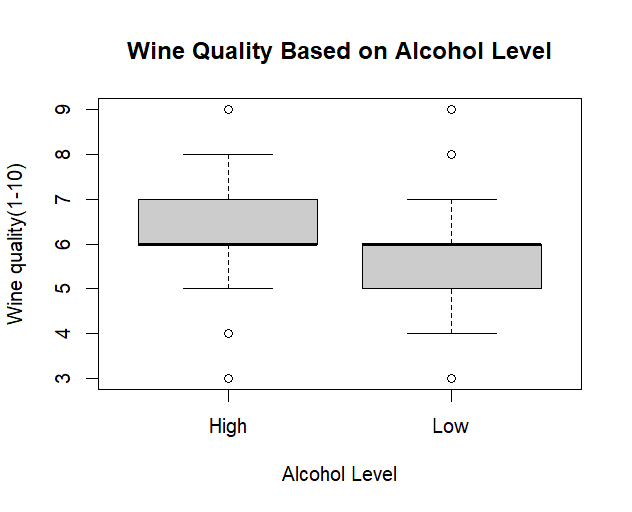
The article titled “Impact of dealcoholisation on quality properties in white wine at various alcohol content levels (Loredana Liguori , et al., 2019)” emphasizes on the effect of alcohol on white wine’s taste, aroma and volatile compounds through the process of dealcoholisation. The research paper suggests that alcohol exerts a significant influence on the overall texture of a wine.

## 2.2. Why RQ is of interest

Numerous studies have suggested that Alcohol is a key determinant of wine quality as it plays a pivotal role in the sensory attributes such as taste, aroma and texture. Only a handful of research compares the wine quality across different alcohol levels in white wine, establishing a sustainable gap in the research. Our study addresses the gap by comparing means of wine with different alcohol levels statistically and visually demonstrating how wine quality is influenced by alcohol level. Future studies might explore how different wine components other than alcohol and wine quality corelate applying advanced machine learning to improve predictions, contributing further to wine analysis.

# 3. Visualisation

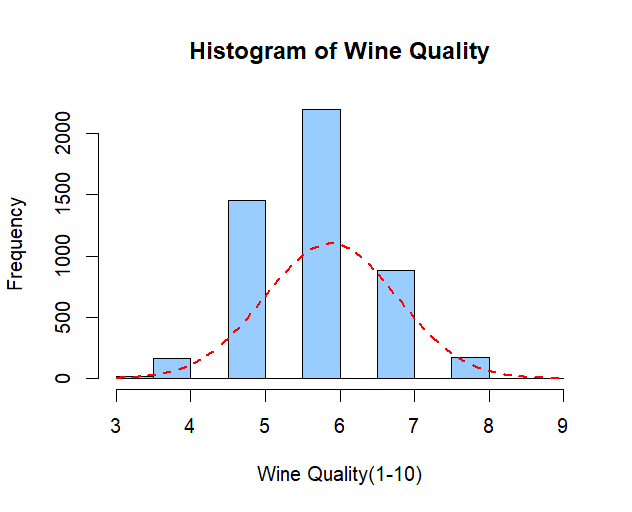
## 3.1. Appropriate plot for the RQ



*Figure 1 (Boxplot of Alcohol level with respect to wine quality)*

The visualisation of Boxplot is the most appropriate for our research as it effectively demonstrates how the two-alcohol level high and low differs from each other in variability and potential outliers. The upward shift of wine with higher alcohol content hints the possibility of difference in the means of the two alcohol levels. The visualization aligns with the premise of our research.

## 3.2. Additional information relating to understanding the data



*Figure 2 (histogram with bell curve on top of dependent variable)*

The Histogram of our dependent variable (wine quality) is essential for our research as this will allow us to form the right statistical test. The bell curve and the histogram structure suggest normality of the data explaining T-test is more suitable for testing our hypothesises. Wilcox test would have been a better alternative if the data wasn’t distributed normally. Thus, the histogram emerges as an essential tool to for selecting the appropriate statistical test.

## 3.3. Useful information for the data understanding

The boxplot indicates that wine with high alcohol generally tend to have higher quality ratings along with outliers reflecting data’s range. The Histogram combined with the bell curve hints that the distribution of dependent variable wine data is mostly normal. This implies that T-test is more appropriate for testing our hypothesis.

# 4. Analysis

## 4.1. Statistical test used to test the hypotheses and output

To examine the hypothesis the independent T-test was chosen as the dependent variable Wine quality is mostly normally distributed because it follows the bell curve shape. Normality is the primary condition of the T-test which makes it superior to the Wilcoxon test in our specific scenario. The Wilcoxon Test is non-parametric and doesn’t require normality. T-test will effectively compare the mean of the two different alcohol levels aligning with our research question.

## 4.2. The null hypothesis is rejected /not rejected based on the p-value

According to the result of the T-test the null hypothesis stating there is no difference in the mean quality between the wines with low alcohol content and wines with high alcohol content can be rejected. The p-value (<2.2e-16) is significantly smaller in contrast to the common significance level of 0.05 provides compelling evidence to decisively refuse the null hypothesis. The 95 percent confidence interval further of 0.63966262 and 0.7332534 further give more evidence that the difference in mean lies within this range confirming a difference exists between means of two different alcohol levels supporting to reject the null hypothesis.

# Evaluation – group’s experience at 7COM1079

## 5.1. What went well

The intention of making sure our team achieves all the results and completes the task consistently and perfectly within due time, we maintained constant contact and cooperation through routine-wise video meetings. The meetings ensured that we understood what we were assigned to do and sort out our lacking. We also divided the works using Trello and divided the work within ourselves to ensure timely completion. Clear and concise results were achievable because we carried out the whole research methodically.

## 5.2. Points for improvement

Although our research as a team was mostly successful there were certainly areas where we could have improved. At the beginning our task division was not very concise which led us to initial delays. Solving the issue would have made our work smoother. Scheduled video meetings were at times difficult as individual members were available at different times leading to cancellation occasionally. Updates through chat messages and Trello comment box in the beginning would have improved our overall task quality.

## 5.3. Group’s time management

Our group used Trello to efficiently assign and manage tasks with due dates. We created boards for each project with To-do, doing, and done columns to track progress. Each task was outlined in a card, including hint, deadline, and assigned team member. Our group’s approach to time management using Trello features like due dates and reminders ensured a swift workflow, on-time completion, necessary corrections, and lastly, a well-structured report.

## 5.4. Project’s overall judgement

Our project effectively examined how alcohol content impacts the quality of white wine. By applying statistical tests and visualizations, we were able to assess differences in quality ratings. The findings met our initial expectations, though there is room for further analysis of additional features. Overall, the team was pleased with the outcome.

## 5.5. Comment on the GitHub log output

# 6. Conclusion

## 6.1. Results Explained

The results demonstrate that there is a difference in the mean quality between the wines with low alcohol content and wines with high alcohol content. The p value being lower than 2.2e-16 enables us to reject the null hypothesis. High Alcohols had a mean quality rating of 6.26 in comparison to 5.57 for low alcohol wines with a confidence interval of 0.6396 to 0.7333. The results prove that wines have a positive influence on the wine quality ratings.

## 6.2. Interpretation of the results

The result aligns with our research question establishing that the content of alcohol influences the overall wine quality. Higher Alcohol being rated on the higher side suggests that alcohol does indeed improve sensory attributes like aroma and taste. This information is helpful for both the winemakers and wine consumers as this will enable them to optimize alcohol for better quality. The influence of alcohol on wine quality suggests the importance of alcohol’s implications on wine production, marketing and consumer preferences.

## 6.3. Reasons and/or implications for future work, limitations of your study

# 7. Reference List

A. d.-l.-F.-B., I. A.-P. A. E., . M.-P. S.-N. & . V. F., 2024. The relevant and complex role of ethanol in the sensory properties of model wines.

Dahal, K. R., J. N. D., H. R. B. & S. G., 2021. Prediction of Wine Quality Using Machine Learning Algorithms.

King, E. S., R. L. D. & H. H., 2013. The influence of alcohol on the sensory perception of red wines.

L. L. et al., 2019. Impact of dealcoholization on quality properties in white wine at various alcohol content levels.

# 8. Appendices

# loading necessary library for CSV file

library(readr)

# reading the winequality-white data set and value

wine <- read.csv("winequality-white.csv", header = TRUE, sep = ";")

View(wine) # viewing data set in another window

# calculating mean to divide the alcohol into subclasess

mean\_alcohol <- mean(wine$alcohol)

# division of two sub classes based on mean // "low" and "High" categories

wine$alcohol\_class <- ifelse(wine$alcohol <= mean\_alcohol, "Low", "High")

#showing first two rows of the dataset

head(wine,2)

# plotting the histogram

hist(wine$quality, main = "Histogram of Wine Quality", xlab = "Wine Quality(1-10)", ylab="Frequency", col = "lightblue", border = "black", freq = TRUE)

# Plotting the normal bell curve

mean\_quality <- mean(wine$quality)

sd\_quality <- sd(wine$quality)

x\_values <- seq(min(wine$quality), max(wine$quality), length.out = 100)

normal\_curve <- dnorm(x\_values, mean = mean\_quality, sd = sd\_quality)

lines(x\_values, normal\_curve \* length(wine$quality) \* diff(hist(wine$quality, plot = FALSE)$mids)[1], col = "red", lwd = 2, lty = 2)

#box plot of the two sub classes in relation to wine quality

boxplot(wine$quality ~ wine$alcohol\_class, data = wine, xlab = "Alcohol Level", ylab = "Wine quality(1-10)", main = "Wine Quality Based on Alcohol Level")

# Performing a t-test to compare the means of the two classes

t\_test\_result <- t.test(wine$quality ~ wine$alcohol\_class)

print(t\_test\_result)