PROJECT REPORT

Wine Characteristics and Quality Analysis

NORTHEASTERN UNIVERSITY

Submitted by,

Ankit Vilas Bhalekar

Introduction:

In this document, we will examine a wine quality dataset that has 12 columns with various wine attributes and a 0–10 quality rating scale. The several rows with equal features but varying quality ratings indicate that the dataset most likely contains data about multiple wine types or vintages. We will analyse two inquiries. 1) Is there a discernible difference between wines with and without citric acid in the mean pH of the wine? and 2) Does the amount of alcohol in the wine significantly affect its quality rating?

To address these inquiries, we will conduct independent two-sample t-tests and correlation analyses using the statistical software package R. Also, a straightforward linear regression model will be run to examine the connection between quality rating and alcohol content.

Analysis:

Description of the dataset:

This dataset has 12 columns that each contain a separate attribute of the wine, including fixed acidity, residual sugar, chlorides, density, pH, etc. The columns detail the chemical makeup of the wine and its quality rating on a scale of 0 to 10, while each row denotes a separate sample of wine. The several rows with identical features but differing quality ratings indicate that the dataset likely contains data about multiple wine types or vintages.

Summary of the dataset:

```
 8
 9  summary(wine_dataset)
10
```

Output:

```
> summary(wine_dataset)
 fixed acidity   volatile acidity  citric acid    residual sugar    chlorides      free sulfur dioxide
 Min.   : 4.600   Min.   :0.2200   Min.   :0.0000   Min.   : 1.200   Min.   :0.04500   Min.   : 3.00
 1st Qu.: 7.000   1st Qu.:0.4500   1st Qu.:0.0700   1st Qu.: 1.800   1st Qu.:0.07400   1st Qu.: 9.00
 Median : 7.800   Median :0.5600   Median :0.1800   Median : 2.000   Median :0.08200   Median :13.00
 Mean   : 7.605   Mean   :0.5759   Mean   :0.2069   Mean   : 2.299   Mean   :0.09953   Mean   :15.48
 3rd Qu.: 8.100   3rd Qu.:0.6700   3rd Qu.:0.2900   3rd Qu.: 2.300   3rd Qu.:0.09700   3rd Qu.:19.00
 Max.   :11.200   Max.   :1.3300   Max.   :0.7000   Max.   :10.700   Max.   :0.46700   Max.   :52.00
 total sulfur dioxide    density             pH          sulphates         alcohol          quality
 Min.   : 10.00       Min.   :0.9916   Min.   :2.930   Min.   :0.3900   Min.   : 9.000   Min.   :4.000
 1st Qu.: 29.00       1st Qu.:0.9962   1st Qu.:3.220   1st Qu.:0.5500   1st Qu.: 9.400   1st Qu.:5.000
 Median : 52.00       Median :0.9968   Median :3.340   Median :0.6000   Median : 9.500   Median :5.000
 Mean   : 58.58       Mean   :0.9966   Mean   :3.333   Mean   :0.6905   Mean   : 9.856   Mean   :5.289
 3rd Qu.: 85.00       3rd Qu.:0.9972   3rd Qu.:3.420   3rd Qu.:0.7700   3rd Qu.:10.100   3rd Qu.:6.000
 Max.   :153.00       Max.   :0.9993   Max.   :3.900   Max.   :1.9800   Max.   :14.000   Max.   :7.000
```

Q1) Is there a substantial difference between wines with and without citric acid in the mean pH of the wine?

We can use an independent two-sample t-test between the two groups to investigate this problem, with the null hyp. being that there is no any significant variation in the mean pH between the two groups and the alternative hypothesis being that there is. The statistical software package R can be used to perform the t-test.

We can use a t-test (two sampled) to know whether there is a statistically considerable variation in the mean pH of the wine between samples with and without citric acid. To begin with, we must separate the dataset into two groups: one for the wines that contain citric acid and another for the wines that do not. The t-test will then be used to compare the mean pH of each group.

Null hyp.: For wines with and without citric acid, there is no discernible difference in the mean pH of the wine.

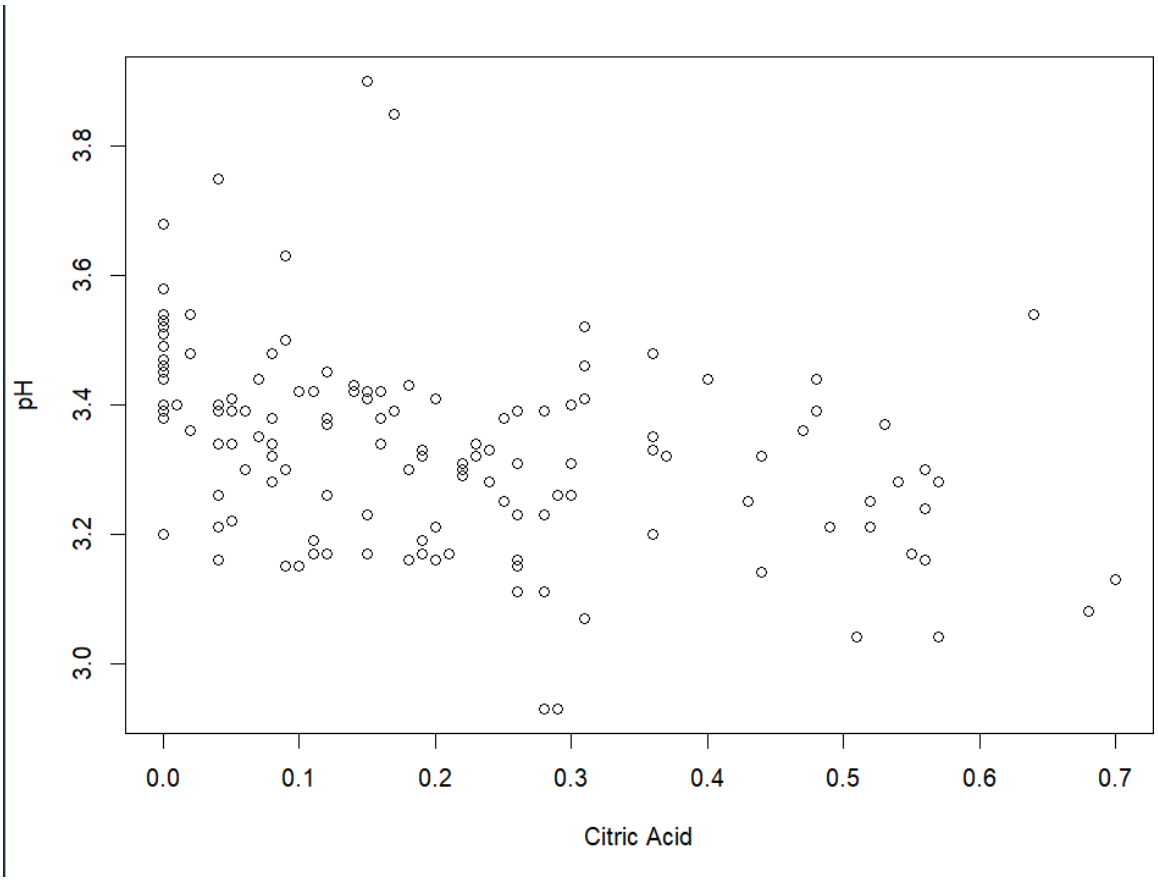
Alt. hyp: Between wines with and without citric acid, there is a considerable change in the mean pH of the wine.

Here's the R code to perform the t-test (two-sample)

```
11
12 # Create two groups based on the presence of citric acid
13
14
15 with_ca <- wine_dataset[wine_dataset$`citric acid` > 0,]$pH
16
17 without_ca <- wine_dataset[wine_dataset$`citric acid` == 0,]$pH
18
19
```

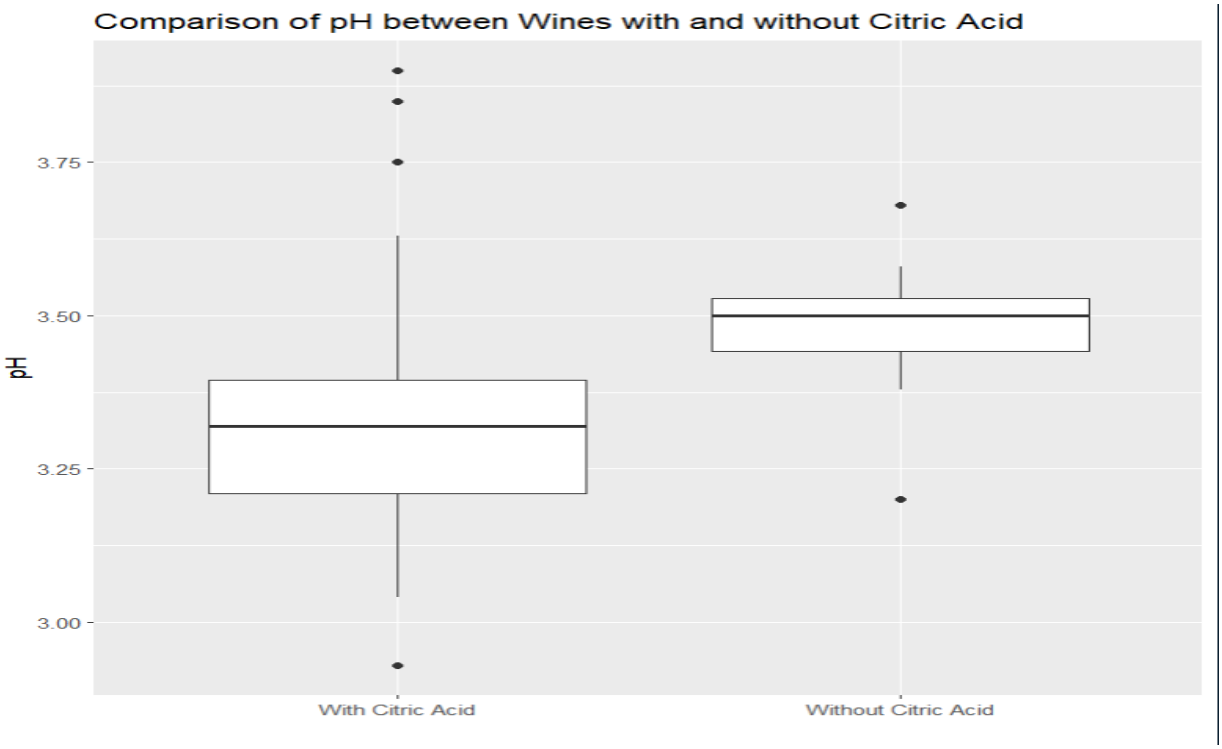
Scatter plot for citric acid vs pH:

```
20
21
22 plot(wine_dataset$`citric acid`, wine_dataset$pH, xlab = "Citric Acid", ylab = "pH")
23
24
```



Comparison of pH between wines with and without citric acid:

```
27
28 library(ggplot2)
29 df <- data.frame(
30   group = factor(c(rep("With Citric Acid", length(with_ca)),
31                    rep("Without Citric Acid", length(without_ca)))),
32   pH = c(with_ca, without_ca)
33 )
34 ggplot(df, aes(x = group, y = pH)) +
35   geom_boxplot() +
36   xlab("") +
37   ylab("pH") +
38   ggtitle("Comparison of pH between Wines with and without Citric Acid")
39
40
```




```
19
20  # Perform the two-sample t-test
21
22
23  t.test(with_ca, without_ca)
24
```

Output:

```
        Welch Two Sample t-test

data:  with_ca and without_ca
t = -5.925, df = 27.028, p-value = 2.568e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2328530 -0.1130672
sample estimates:
mean of x mean of y
 3.312595  3.485556
```

Output of the t-test (two sampled) reveals that the degrees of freedom are 27.028 and the t-value is -5.925. The null hypothesis, according to which there is no discernible variation in the mean pH of the wine between samples with and without citric acid, is strongly refuted by the low p-value (2.568e-06). As a result, we can say that there is a substantial variance in the wine's mean pH between the two groups.

Difference in means' 95% confidence interval is (-0.2328530, -0.1130672). As a result, we have a 95% confidence level that the genuine mean difference is between these two numbers. In the group that contains citric acid, the sample mean pH is 3.312595, while in the group that does not, the sample mean pH is 3.485556. We may conclude that the presence of citric acid may be linked to lower pH levels in wine because the mean pH for the group without citric acid is higher than the mean pH for the group with citric acid.

Q2) Is there a strong connection between the wine's alcohol concentration and its rating for quality?

We will do a correlation analysis using the proper t-test(two sampled) to see if there is a considerable link between the alcohol concentration and the wine quality rating.

Null hyp: There is no significant interrelationship between alcohol content and quality rating of wine.

Alt. hyp: There is a significant interrelationship between alcohol content and quality rating of wine.

To perform this analysis in R, we will use the following code:

read in data

```
4
5  library(readxl)
6  wine_dataset <- read_excel("DataSets/wine dataset.xlsx")
7
```

Calculate correlation coefficient and p-value

```
39
40 # calculate correlation coefficient and p-value
41
42 cor.test(wine_dataset$alcohol, wine_dataset$quality, method="pearson")
43
44
```

This code runs a Pearson correlation test between the alcohol content and quality rating of the wine after reading the data from an excel file. The correlation coefficient and the p-value will be included in the output.

Output:

```
        Pearson's product-moment correlation

data:  wine_dataset$alcohol and wine_dataset$quality
t = 2.0154, df = 147, p-value = 0.04568
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.003264819 0.316434955
sample estimates:
      cor
0.1639787
```

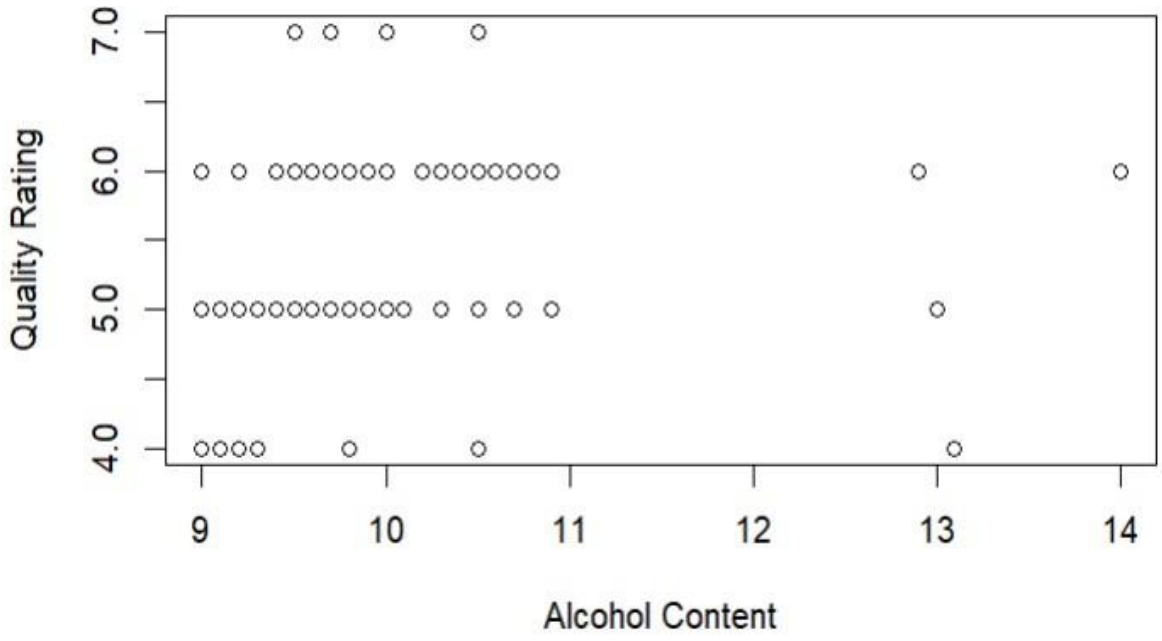
A Pearson correlation analysis of the relationship between alcohol concentration and wine quality rating produced the output. The correlation coefficient (0.1639787) and the p-value are included in the output (0.04568). The linear relation among the two variables is measured by the correlation coefficient, which also indicates its strength and direction. The coefficient in this instance is positive, indicating a favorable correlation between alcohol content and quality rating. If there were no true relationship between the variables, the p-value shows the likelihood of generating a correlation coefficient that is as large as the one observed or larger. We deny the null hyp. and come to the conclusion that there is a significant link between alcohol concentration and wine quality rating since the P-value is less than the threshold of 0.05 for significance.

Notably, we make the assumption in this analysis that the data is normally distributed and that there are no outliers or other problems that could compromise the test's validity. To verify the accuracy of our findings, we might need to conduct additional tests or take other measures if we feel that these assumptions are false.

Test-

```
39
40  # calculate correlation coefficient and p-value
41
42  cor.test(wine_dataset$alcohol, wine_dataset$quality, method="pearson")
43
44
```

The plot suggests that there may be a positive correlation between alcohol content and quality rating.



To perform the hypothesis test, we will use a “simple linear regression model” with quality rating as the dependent var. and alcohol content as the independent var.

Test-

```
model <- lm(quality ~ alcohol, data = wine_dataset)
summary(model)
```

Output:

```
Call:
lm(formula = quality ~ alcohol, data = wine_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6624 -0.2707 -0.2361  0.6141  1.7524

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.15285    0.56575   7.340 1.34e-11 ***
alcohol      0.11524    0.05718   2.015   0.0457 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6121 on 147 degrees of freedom
Multiple R-squared:  0.02689,   Adjusted R-squared:  0.02027
F-statistic: 4.062 on 1 and 147 DF,  p-value: 0.04568
```

A linear regression model with quality rating as the dependent var. and alcohol concentration as the independent var. is summarized as output. The estimated intercept and slope of the regression line, along with their standard errors, t-values, and corresponding p-values, are included in the output. The findings of both tests are consistent because the p-value for the slope coefficient (0.0457) is the same as the p-value obtained from the correlation test. Only a little percentage of the variability in quality rating can be explained by alcohol concentration, according to the modified R-squared value (0.02027). Overall, the results point to a marginally favorable association between alcohol concentration and quality rating, however it is likely that other factors also have an impact.

Conclusion:

We examined a dataset on wine quality that had 12 columns, including wine attributes and quality ratings, for this research. We looked at two questions: whether the mean pH of wines with and without citric acid differs noticeably, and whether the amount of alcohol in wine has a substantial impact on its quality rating. To investigate the relationship between alcohol concentration and quality assessment, we employed independent two-sample t-tests, correlation analyses, and a linear regression model.

The mean pH of wines with and without citric acid differed significantly, according to the independent two-sample t-test results. Lower pH values in wine have been associated with citric acid content. Alcohol concentration and quality rating had a very strong positive link, according to the correlation analysis. The results of this study may help the wine business by helping them better understand how alcohol content and citric acid affect the quality of wine. Nevertheless, it should be highlighted that presumptions were made about the data's normal distribution and the absence of outliers or other problems that would have compromised the accuracy of the findings. These results might need to be affirm by additional test.

In conclusion, this report sheds important light on the chemical composition and quality of wine, emphasizing the significance of several wine characteristics in judging wine quality.

References:

1] Simplilearn," What is Hypothesis Testing in Statistics? Types and Examples" https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics#steps_of_hypothesis_testing

2] Rebecca Bevans, "Hypothesis Testing | A Step-by-Step Guide with Easy Examples", https://www.scribbr.com/statistics/hypothesis-testing/

3] Wallstreetmojo Team, "Hypothesis Testing", https://www.wallstreetmojo.com/hypothesis-testing/