MATH1324 Assignment 3

pH value of Red and White Wines

Sarthak Sirari (S3766477)

Last updated: 02 June, 2019

Introduction

Over the years, we have developed a taste for wines and this has lead to invention of tastes of Red and White Wines.

This variation in the taste of Wine is done by adding or removing different ingredients when making the wine, changing the order of adding different ingredients and also by manipulating the fermentation process by adding some chemical.

We will be observing the variance of pH value of in different Red and WHite Wine samples of the Portuguese “Vinho Verde” wine.

Problem Statement

The purpose is to determine if there is a statistically significant mean difference between the pH values of the Red and White Wines.

To determine this, we will be performing several statistical tasks and visualisation of the data.

We will be performing a Two Sample t-Test on our data set to conclude our investigation.

Data

Our data set is an open data set available at UCI Machine Learning repository present under the link (http://archive.ics.uci.edu/ml/datasets/Wine+Quality).

This Wine Quality Data Set contains two data sets of the Portuguese "Vinho Verde" wine. One data set is of Red Wine samples and another one is of the White Wine samples.

This data set is of Classification or Regression type.

There are total 12 attributes available in the data set like fixed acidity, citric acid, total sulfur dioxide, density, etc. However, we will be picking only the pH value of both the Red and White wines.

The data is available since 2009 and there are no missing values.

Data Cont.

Since we will be performing our test only on the pH value observation, we will first load both the data set CSVs of Whiet and Red Wines and then extract the observations of pH values and then start perming our investigation.

The data set which will be further investigated will have only two columns as follong: - 1. pH: Contains the pH value observation of different Wine samples. 2. Type: Contains the type of Wine the pH observation belongs to.

```
# Set working directory
setwd("C:\\Users\\Abhisar")
```

```
# Load the data set of Red Wine
red_wine <- read_csv2("winequality-red.csv")
```

```
# View loaded data set of Red Wine
View(red_wine)
```

```
# Load the data set of White Wine
white_wine <- read_csv2("winequality-white.csv")
```

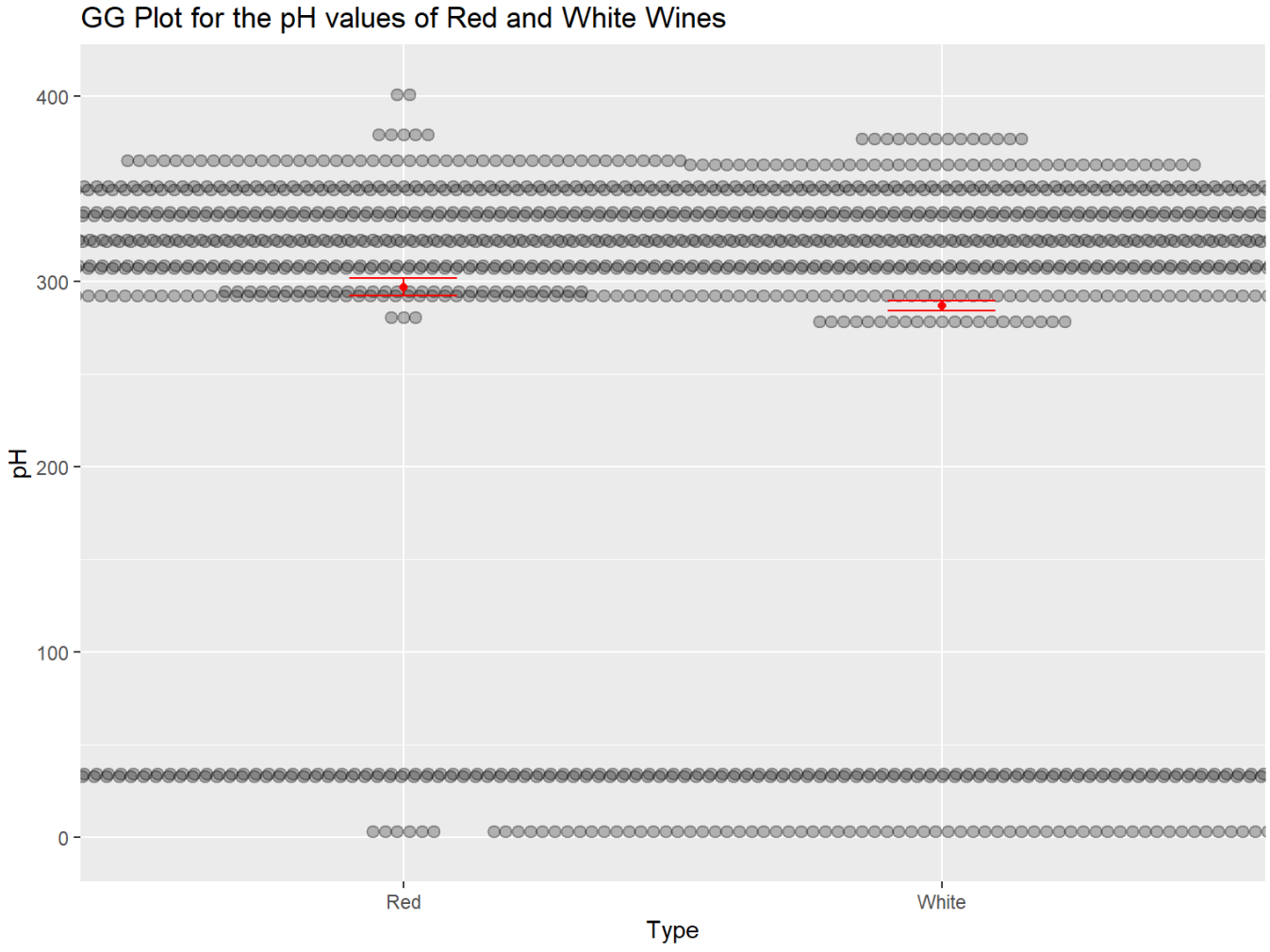
```
# View loaded data set of White Wine
View(white_wine)
```

```
# Selecting the pH value column and adding a Type column
red_wine_pH <- red_wine %>% select(pH) %>% mutate(Type="Red")
white_wine_pH <- white_wine %>% select(`pH`) %>% mutate(Type="White")

# Combining the filtered pH value data of both Red and White Wines.
wine_pH <- rbind(white_wine_pH,red_wine_pH)
view(wine_pH)
```

Descriptive Statistics and Visualisation

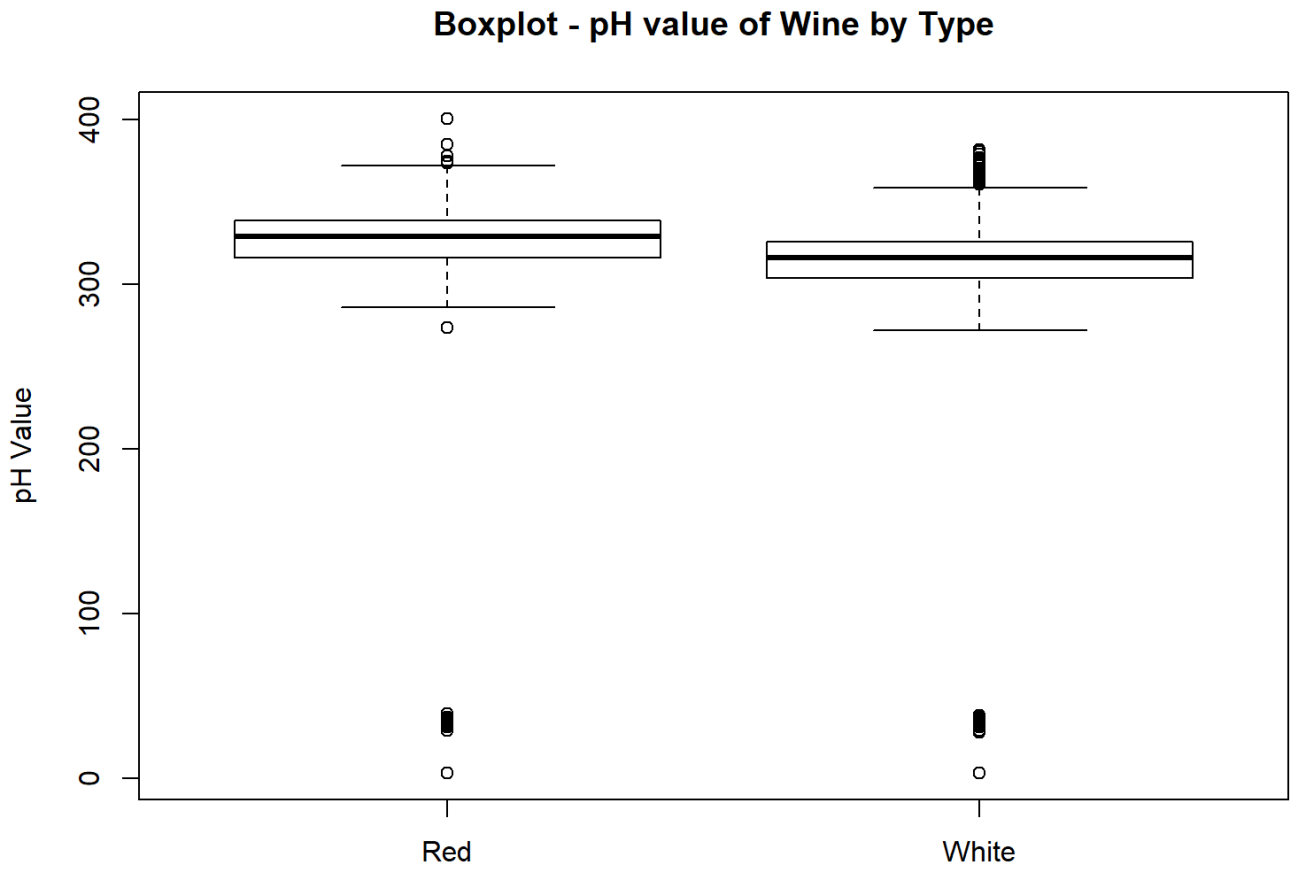
```
# GG-Plot of the pH value observations of Red and White Wines
p1 <- ggplot(data = wine_pH, aes(x = Type, y = pH))
p1 + geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 1/2, alpha = .25) +
    stat_summary(fun.y = "mean", geom = "point", colour = "red") +
    stat_summary(fun.data = "mean_cl_normal", colour = "red",
                 geom = "errorbar", width = .2) +
    ggtitle("GG Plot for the pH values of Red and White Wines")
```



From the above plot, it can be clearly seen that the mean pH value of Red Wine is greater than that of White Wine.

Descriptive Statistics and Visualisation Cont.

```
# Box Plot of the pH value observations of Red and White Wines
wine_pH %>% boxplot(pH ~ Type, data = ., ylab="pH Value", main = "Boxplot - pH value of Wine by Type")
```



From the above plot, it can be observed that the median pH value for Red Wine is greater than that of White Wine.

Decsriptive Statistics Cont.

Following is the summary of statistics data of our chosen data set categorised by the type of Wine, i.e. Red and White.

```
wine_pH %>% group_by(Type) %>% summarise(Min = min(pH,na.rm = TRUE),
                                         Q1 = quantile(pH,probs = .25,na.rm = TRUE),
                                         Median = median(pH, na.rm = TRUE),
                                         Q3 = quantile(pH,probs = .75,na.rm = TRUE),
                                         Max = max(pH,na.rm = TRUE),
                                         Mean = mean(pH, na.rm = TRUE),
                                         SD = sd(pH, na.rm = TRUE),
                                         n = n(),
                                         Missing = sum(is.na(pH))) -> wine_pH_summary
knitr::kable(wine_pH_summary)
```

Type	Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
Red	3	316	329	339	401	297.0025	96.37154	1599	0
White	3	304	316	326	382	287.0768	92.06556	4898	0

Hypothesis Testing

We will perform the Two Samples t-test on our data set to check for a statistically significant mean difference in the pH Values of White Wine and Red Wine by performing the following steps: -

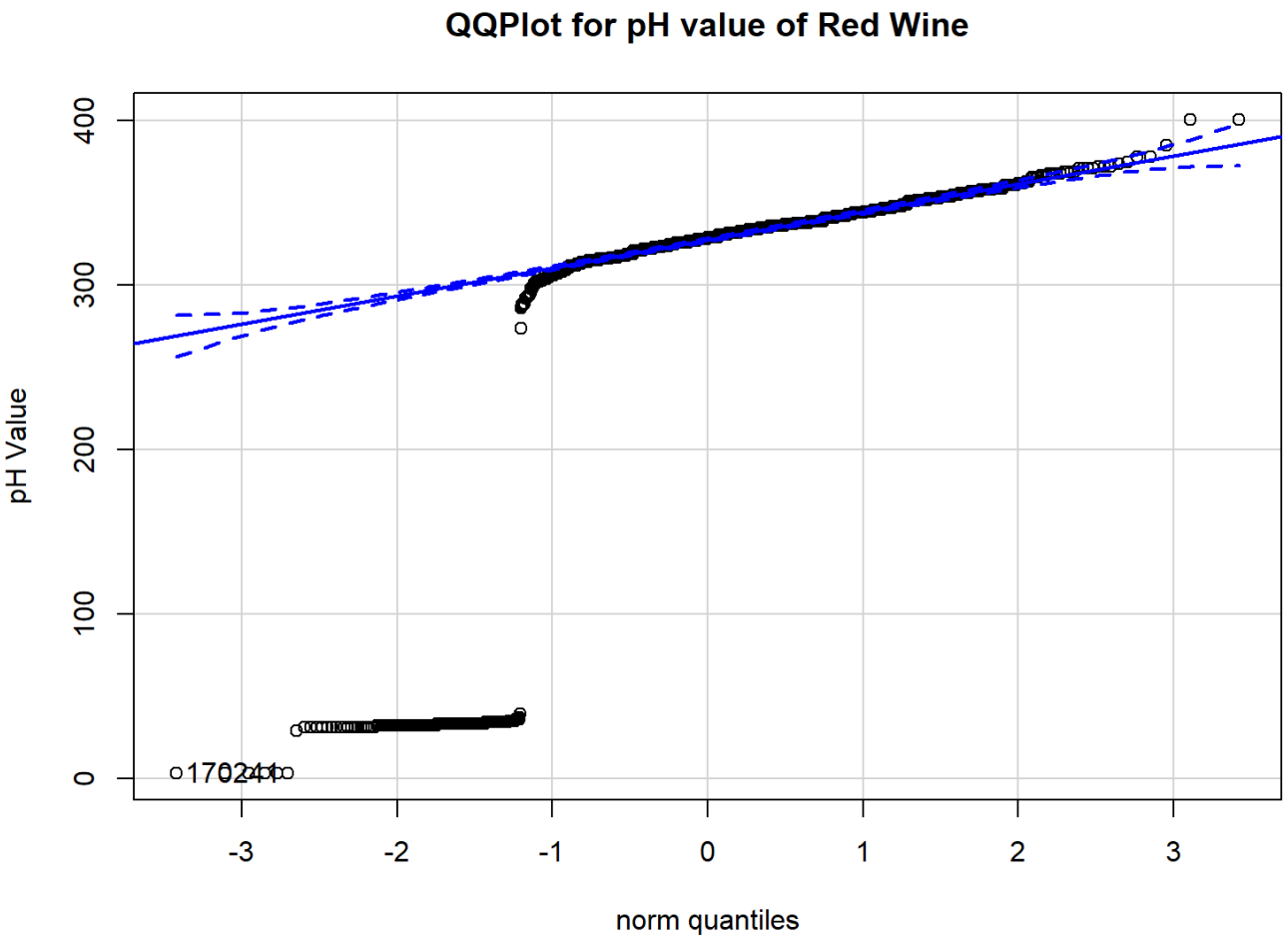
1. We will first check for the normality of data by visualising a QQ-Plot.
2. Then we perform the test of Homogeneity of Variance.
3. Further we will perform the Two Sample t-test on our data set.

After performing the above steps, we will observe the results and look for the appropriate conclusion.

Hypothesis Testing : QQ-Plot

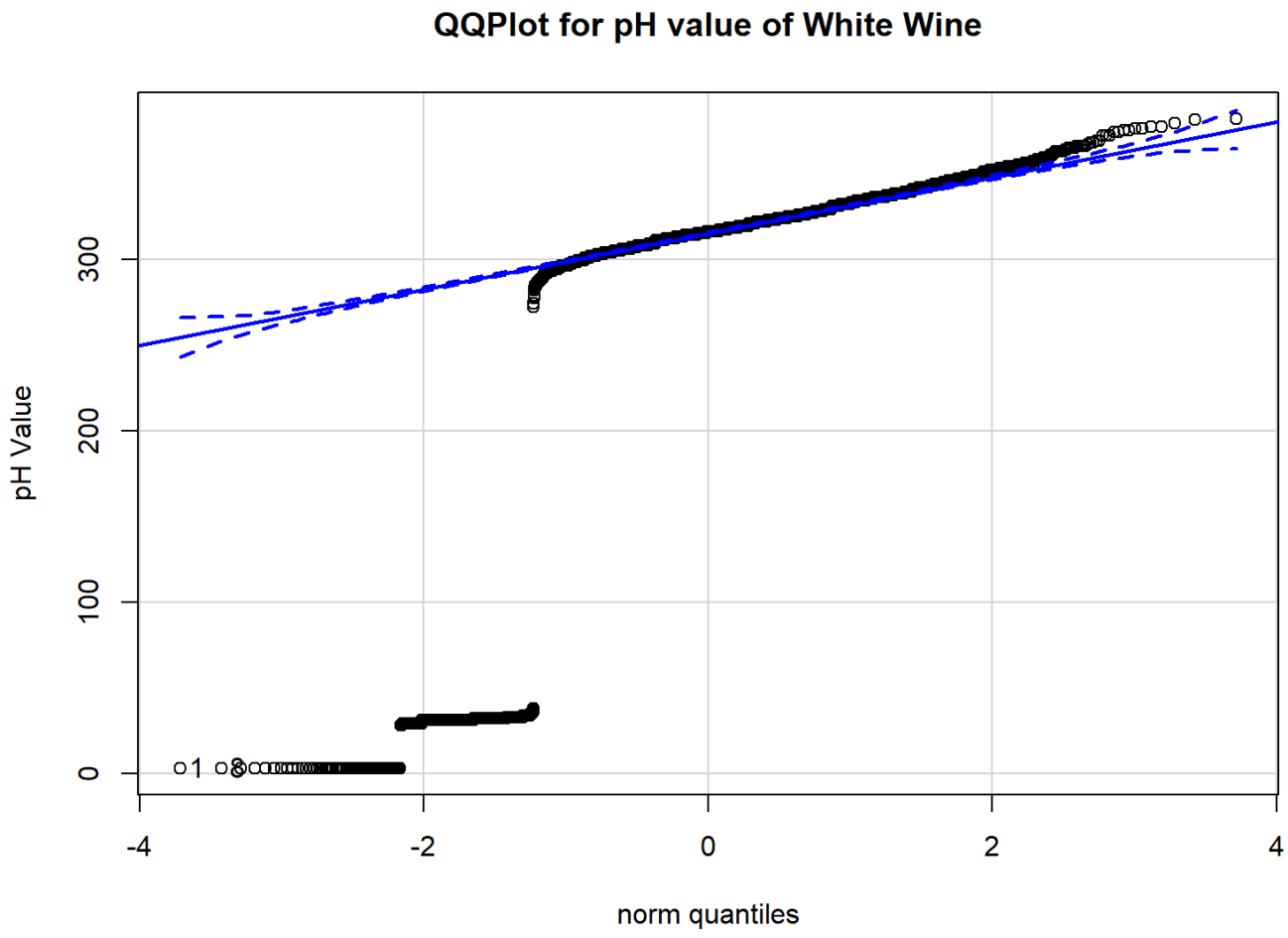
We visualised a QQ Plot to compare our observations in the data set with what the normal observations should be. We check if our values falls ourside the dashed line or not to further go for the test of Homogeneity of Variance.

```
red_wine_pH$pH %>% qqPlot(dist="norm", ylab="pH Value", main = "QQPlot for pH value of Red Wine")
```




```
## [1] 170 241
```

```
white_wine_pH$pH %>% qqPlot(dist="norm", ylab="pH Value", main = "QQPlot for pH value of White Wine")
```


QQPlot for pH value of White Wine
pH Value
0
100
200
300
-4
-2
0
2
4
norm quantiles
1
8

```
## [1] 1 8
```

Hypothesis Testing: Levene Test

In order to test Homogeneity of Variance, we use Levene's Test. The Levene's test has following statistical hypotheses: -

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_A : \sigma_1^2 \neq \sigma_2^2$$

```
leveneTest(wine_pH$pH ~ wine_pH$Type)
```

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	1	1.146913	0.2842362
	6495	*NA*	*NA*

2 rows

Hypthesis Testing: Two sample t-Test

For the two-sample t-test, following are the statistical hypotheses: -

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

Since the p-value for the Levene's Test was found to be 0.284, is greater than 0.05, so we perform the Two Sample t-Test assuming unequal variance.

```
t.test(
    pH ~ Type,
    data = wine_pH,
    var.equal = FALSE,
    alternative = "two.sided"
)
```

```
##
##  Welch Two Sample t-test
##
## data:  pH by Type
## t = 3.615, df = 2616.3, p-value = 0.000306
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.54179 15.30968
## sample estimates:
##   mean in group Red mean in group White
##            297.0025            287.0768
```

Discussion

We performed the Two Samples t-test on our data set to check for a statistically significant mean difference in the pH Values of White Wine and Red Wine.

The strength of our investigation is that our data set consisted of a large number of observations of both Red and White Wine samples.

The weakness of our investigation is that our data set contained a lot of outliers in observations of both Red and White Wine samples.

Our results could be used in the future by a new liqour company to determine the pH value of the Red Wines and White Wines. Also, to determine the difference in the pH values of both Red and White Wines should be kept to get the best selling product.

We plot a QQ Plot for both Red and White wine samples. When we check the plots, there were a set of values falling inside as well as outside of the dashed lines.

Since there were some values which fall outside the dashed lines, so we performed the test of Homogeneity of Variance, we use Levene's Test, which showed the p-value of 0.284, is greater than 0.05, so we perform the Two Sample t-Test assuming unequal variance.

The result of the performed Two Sample t-test are as following:-

1. The p-value was found to be 0.0003, which is less than 0.05.

2. The 95 percent confidence interval is [4.541, 15.309]

3. The t-value is 3.615, which does not fall within the 95% CI.

To conclude, the results of our investigation suggest that pH value of Red Wine is significantly higher than the pH value of White Wine.

References

1. Intro to Stats MATH1324 Module Notes, https://astral-theory-157510.appspot.com/secured/MATH1324_Module_01.html

2. UCI Machine Learning Respository, Wine Quality Data Set, http://archive.ics.uci.edu/ml/datasets/Wine+Quality