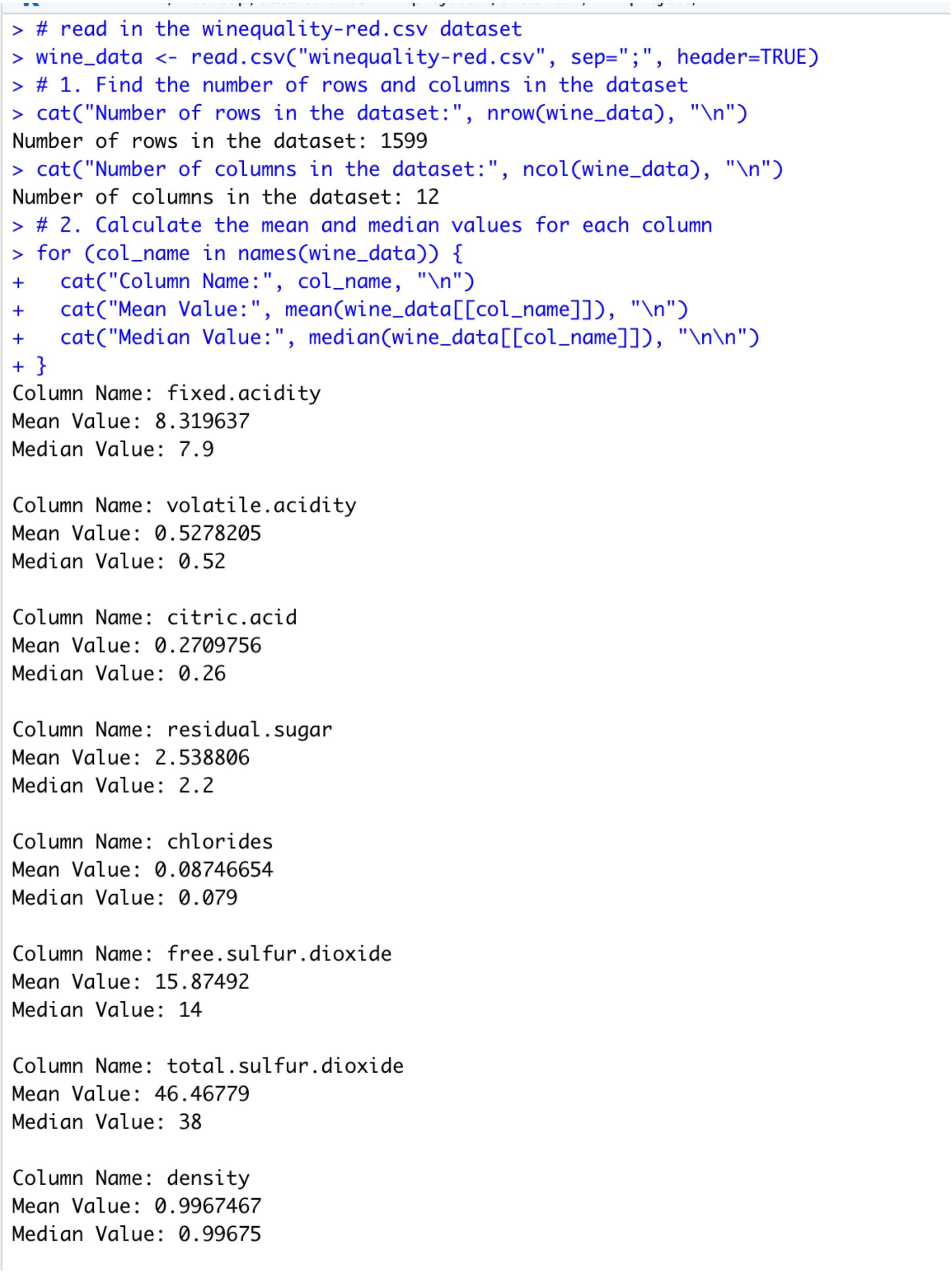
Data Science

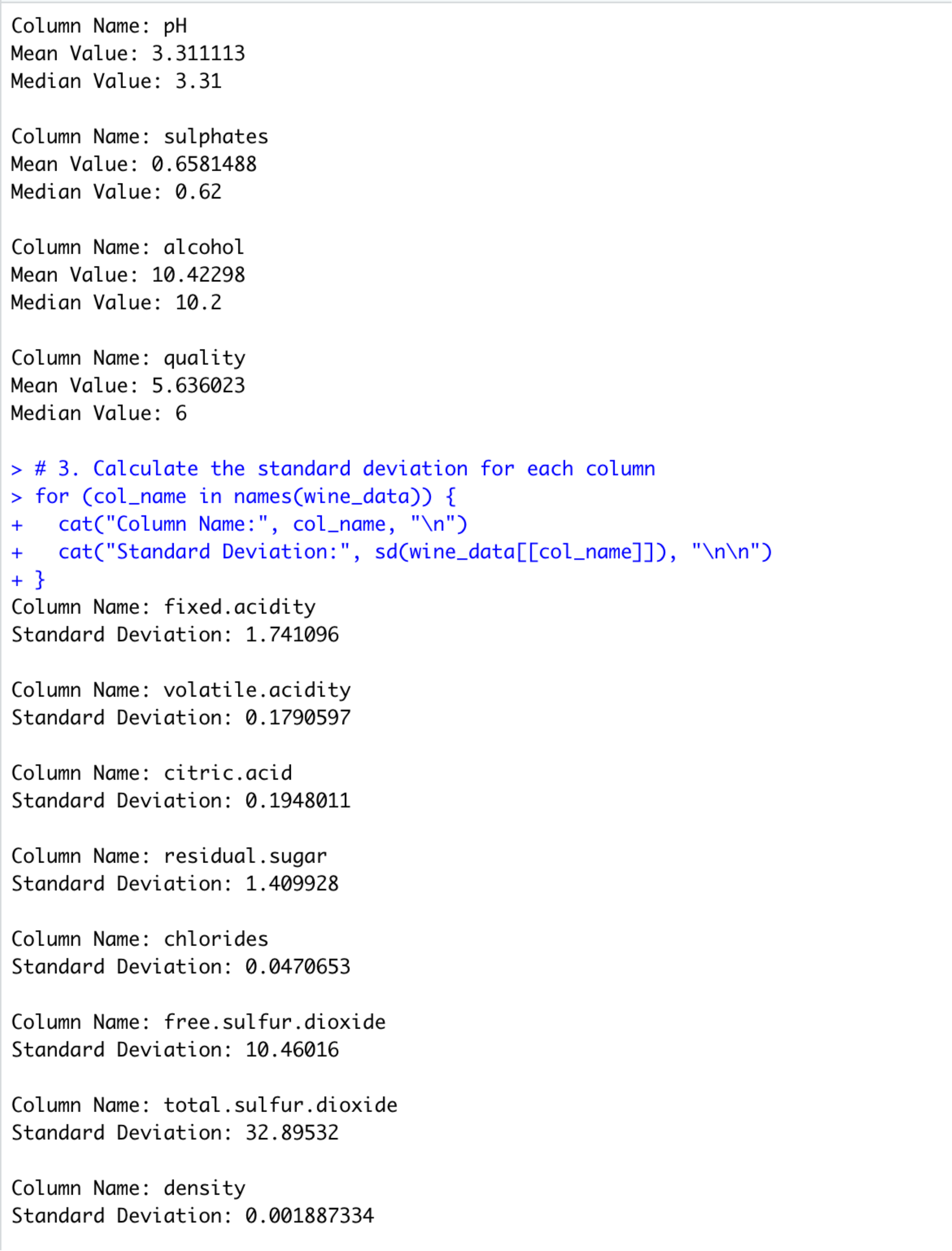
Miniproject 1: R

By: Shubham Vats
02/15/2023

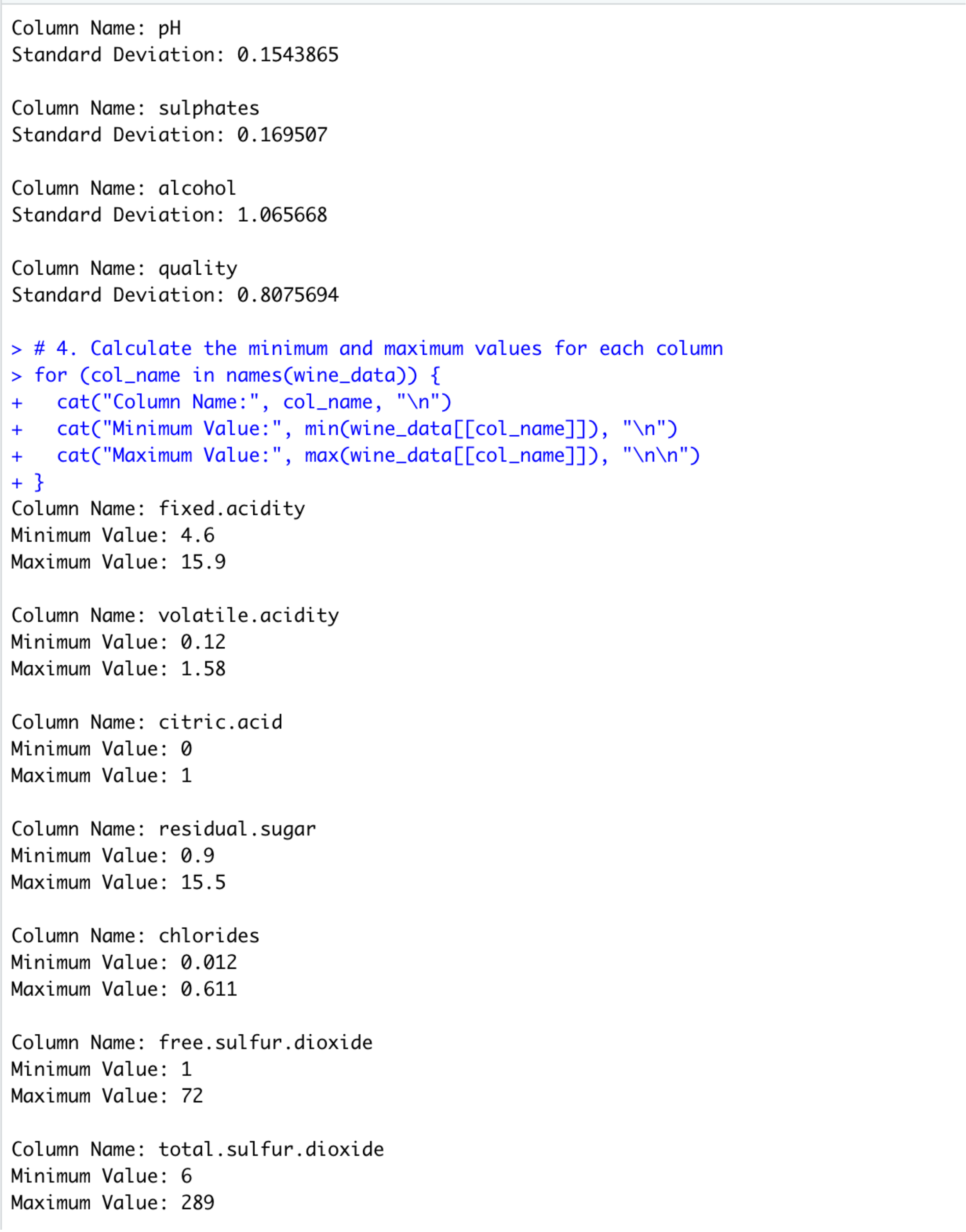
Output File-



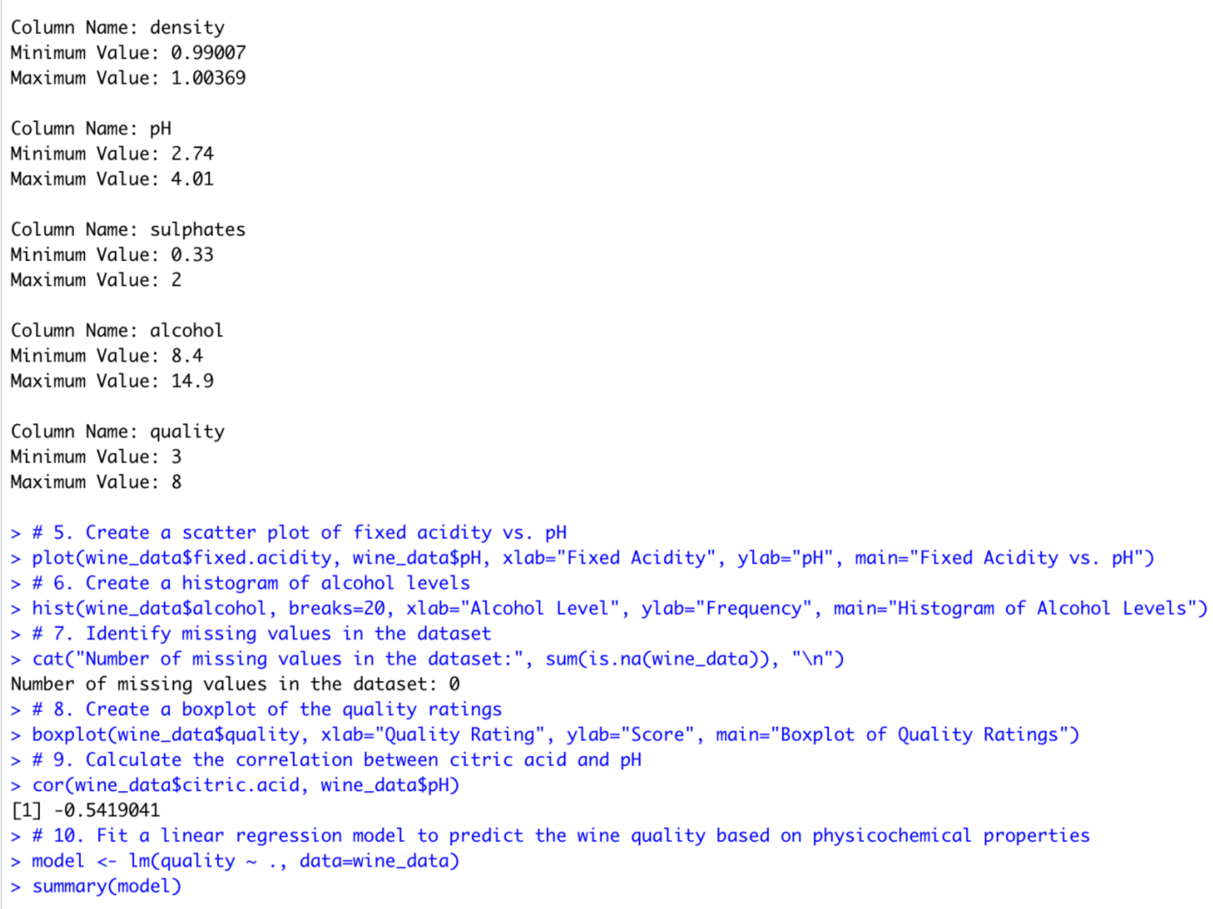
```
> # read in the winequality-red.csv dataset
> wine_data <- read.csv("winequality-red.csv", sep=";", header=TRUE)
> # 1. Find the number of rows and columns in the dataset
> cat("Number of rows in the dataset:", nrow(wine_data), "\n")
Number of rows in the dataset: 1599
> cat("Number of columns in the dataset:", ncol(wine_data), "\n")
Number of columns in the dataset: 12
> # 2. Calculate the mean and median values for each column
> for (col_name in names(wine_data)) {
+   cat("Column Name:", col_name, "\n")
+   cat("Mean Value:", mean(wine_data[[col_name]]), "\n")
+   cat("Median Value:", median(wine_data[[col_name]]), "\n\n")
+ }
Column Name: fixed.acidity
Mean Value: 8.319637
Median Value: 7.9

Column Name: volatile.acidity
Mean Value: 0.5278205
Median Value: 0.52

Column Name: citric.acid
Mean Value: 0.2709756
Median Value: 0.26

Column Name: residual.sugar
Mean Value: 2.538806
Median Value: 2.2

Column Name: chlorides
Mean Value: 0.08746654
Median Value: 0.079

Column Name: free.sulfur.dioxide
Mean Value: 15.87492
Median Value: 14

Column Name: total.sulfur.dioxide
Mean Value: 46.46779
Median Value: 38

Column Name: density
Mean Value: 0.9967467
Median Value: 0.99675
```



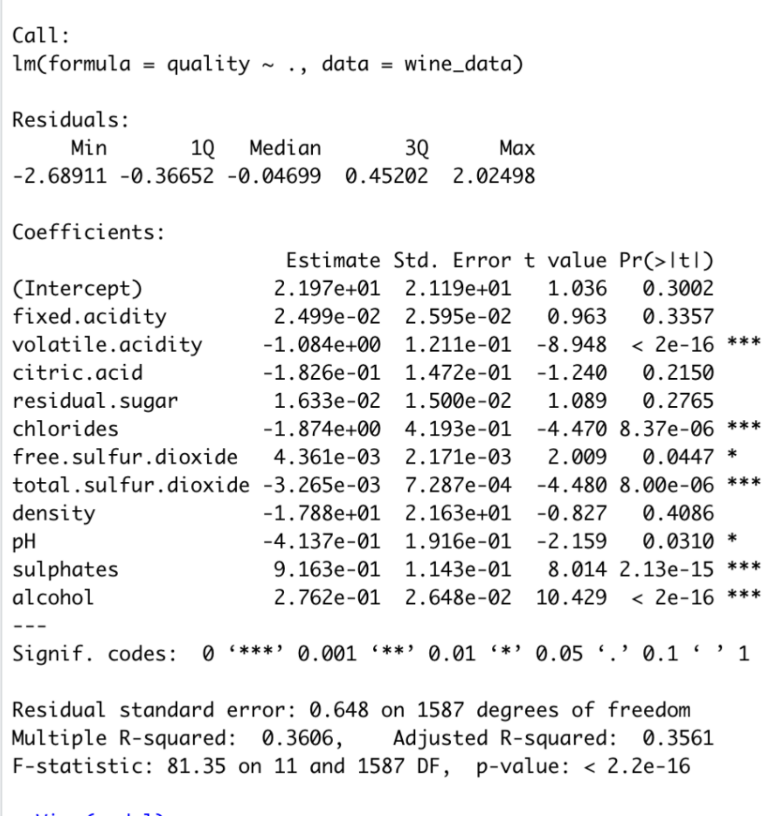
```
Column Name: pH
Mean Value: 3.311113
Median Value: 3.31

Column Name: sulphates
Mean Value: 0.6581488
Median Value: 0.62

Column Name: alcohol
Mean Value: 10.42298
Median Value: 10.2

Column Name: quality
Mean Value: 5.636023
Median Value: 6

> # 3. Calculate the standard deviation for each column
> for (col_name in names(wine_data)) {
+   cat("Column Name:", col_name, "\n")
+   cat("Standard Deviation:", sd(wine_data[[col_name]]), "\n\n")
+ }
Column Name: fixed.acidity
Standard Deviation: 1.741096

Column Name: volatile.acidity
Standard Deviation: 0.1790597

Column Name: citric.acid
Standard Deviation: 0.1948011

Column Name: residual.sugar
Standard Deviation: 1.409928

Column Name: chlorides
Standard Deviation: 0.0470653

Column Name: free.sulfur.dioxide
Standard Deviation: 10.46016

Column Name: total.sulfur.dioxide
Standard Deviation: 32.89532

Column Name: density
Standard Deviation: 0.001887334
```



```
Column Name: pH
Standard Deviation: 0.1543865

Column Name: sulphates
Standard Deviation: 0.169507

Column Name: alcohol
Standard Deviation: 1.065668

Column Name: quality
Standard Deviation: 0.8075694

> # 4. Calculate the minimum and maximum values for each column
> for (col_name in names(wine_data)) {
+   cat("Column Name:", col_name, "\n")
+   cat("Minimum Value:", min(wine_data[[col_name]]), "\n")
+   cat("Maximum Value:", max(wine_data[[col_name]]), "\n\n")
+ }
Column Name: fixed.acidity
Minimum Value: 4.6
Maximum Value: 15.9

Column Name: volatile.acidity
Minimum Value: 0.12
Maximum Value: 1.58

Column Name: citric.acid
Minimum Value: 0
Maximum Value: 1

Column Name: residual.sugar
Minimum Value: 0.9
Maximum Value: 15.5

Column Name: chlorides
Minimum Value: 0.012
Maximum Value: 0.611

Column Name: free.sulfur.dioxide
Minimum Value: 1
Maximum Value: 72

Column Name: total.sulfur.dioxide
Minimum Value: 6
Maximum Value: 289
```

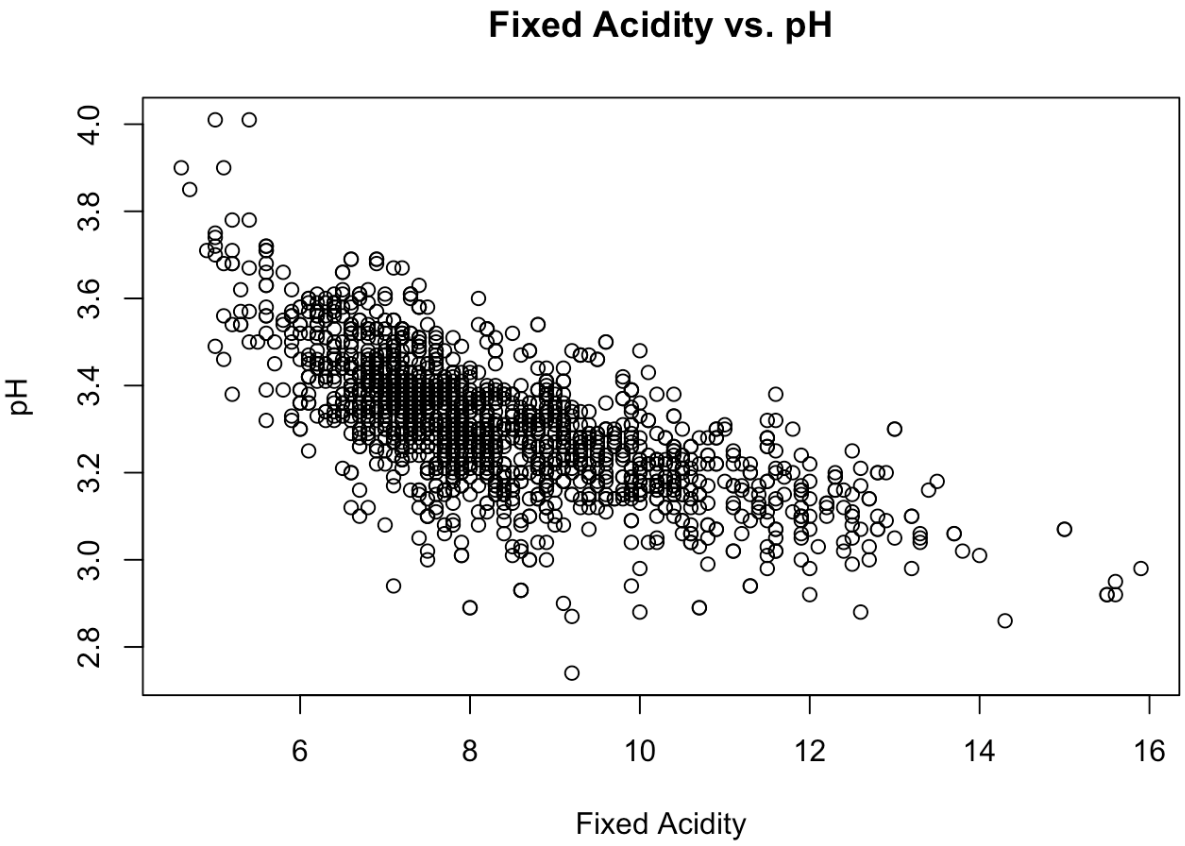


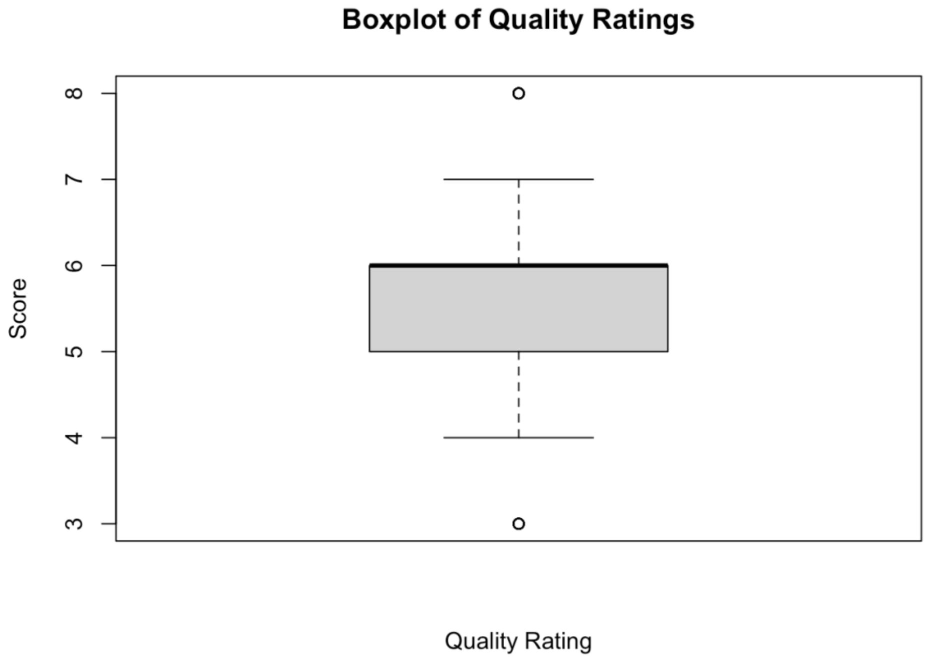
```
Column Name: density
Minimum Value: 0.99007
Maximum Value: 1.00369

Column Name: pH
Minimum Value: 2.74
Maximum Value: 4.01

Column Name: sulphates
Minimum Value: 0.33
Maximum Value: 2

Column Name: alcohol
Minimum Value: 8.4
Maximum Value: 14.9

Column Name: quality
Minimum Value: 3
Maximum Value: 8

> # 5. Create a scatter plot of fixed acidity vs. pH
> plot(wine_data$fixed.acidity, wine_data$pH, xlab="Fixed Acidity", ylab="pH", main="Fixed Acidity vs. pH")
> # 6. Create a histogram of alcohol levels
> hist(wine_data$alcohol, breaks=20, xlab="Alcohol Level", ylab="Frequency", main="Histogram of Alcohol Levels")
> # 7. Identify missing values in the dataset
> cat("Number of missing values in the dataset:", sum(is.na(wine_data)), "\n")
Number of missing values in the dataset: 0
> # 8. Create a boxplot of the quality ratings
> boxplot(wine_data$quality, xlab="Quality Rating", ylab="Score", main="Boxplot of Quality Ratings")
> # 9. Calculate the correlation between citric acid and pH
> cor(wine_data$citric.acid, wine_data$pH)
[1] -0.5419041
> # 10. Fit a linear regression model to predict the wine quality based on physicochemical properties
> model <- lm(quality ~ ., data=wine_data)
> summary(model)
```

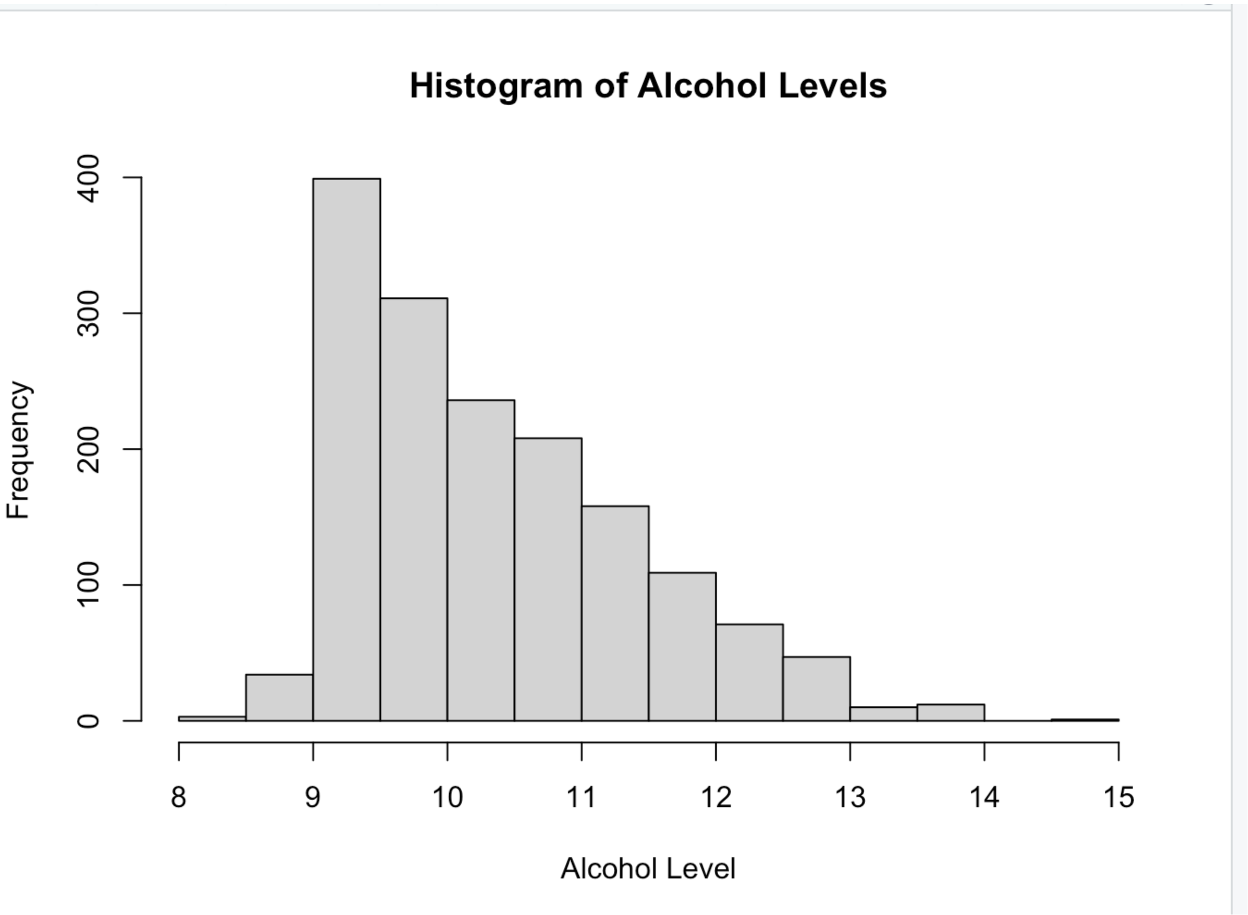


```
Call:
lm(formula = quality ~ ., data = wine_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.68911 -0.36652 -0.04699  0.45202  2.02498 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.197e+01  2.119e+01   1.036   0.3002    
fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357    
volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150    
residual.sugar        1.633e-02  1.500e-02   1.089   0.2765    
chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *  
total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density              -1.788e+01  2.163e+01  -0.827   0.4086    
pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *  
sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.648 on 1587 degrees of freedom
Multiple R-squared:  0.3606,	Adjusted R-squared:  0.3561 
F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Related Graphs-






Histogram of Alcohol Levels
Frequency
0
100
200
300
400
8
9
10
11
12
13
14
15
Alcohol Level