R_Project

NWUDO CHIKAEZE FIDELIS JUNIOR

2022-11-08

INTRODUCTION

1. Title of Database: Wine recognition data

2. Sources:

    (a) Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

    (b) Stefan Aeberhard, email: stefan@coral.cs.jcu.edu.au

    (c) July 1991

Description: The wine dataset contains the results of a chemical analysis of wines grown in a specific area of Italy. Three types of wine are represented in the 178 samples, with the results of 13 chemical analyses recorded for each sample. The Cultivar variable has been transformed into a categoric variable.

• Number of samples: 178 • Class labels: [1, 2, 3] • Distribution: [59, 71, 48]

Dataset Attributes:

Features - they are all continuous variables and they include;

-
    1) Alcohol (%)
-
    2) Malic acid (g/L)
-
    3) Ash(g/L)
-
    4) Alcalinity of ash(g/L)
-
    5) Magnesium (mg/L)
-
    6) Total phenols (mg/L)
-
    7) Flavanoids
-

8) Nonflavanoid phenols

-

9) Proanthocyanins

-

10) Color intensity

-

11) Hue

-

12) OD280/OD315 of diluted wines

-

13) Proline

Target variable(categorical): - Cultivars

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   0.3.5
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
##
## Attaching package: 'gridExtra'
##
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
##
## corrplot 0.92 loaded
##
## Loading required package: xts
##
## Loading required package: zoo
##
##
## Attaching package: 'zoo'
##
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
```

```
##
##
## Attaching package: 'xts'
##
##
## The following objects are masked from 'package:dplyr':
##
##     first, last
##
##
##
## Attaching package: 'PerformanceAnalytics'
##
##
## The following object is masked from 'package:graphics':
##
##     legend
```

CONVERTING NON NUMERIC COLUMNS TO NUMERIC.

To convert non-numeric columns to numeric in a dataframe in R, you can use the as.numeric() function. This function will attempt to convert values in a non-numeric column to numeric, and will return a numeric vector with the converted values.

DATA SUMMARY

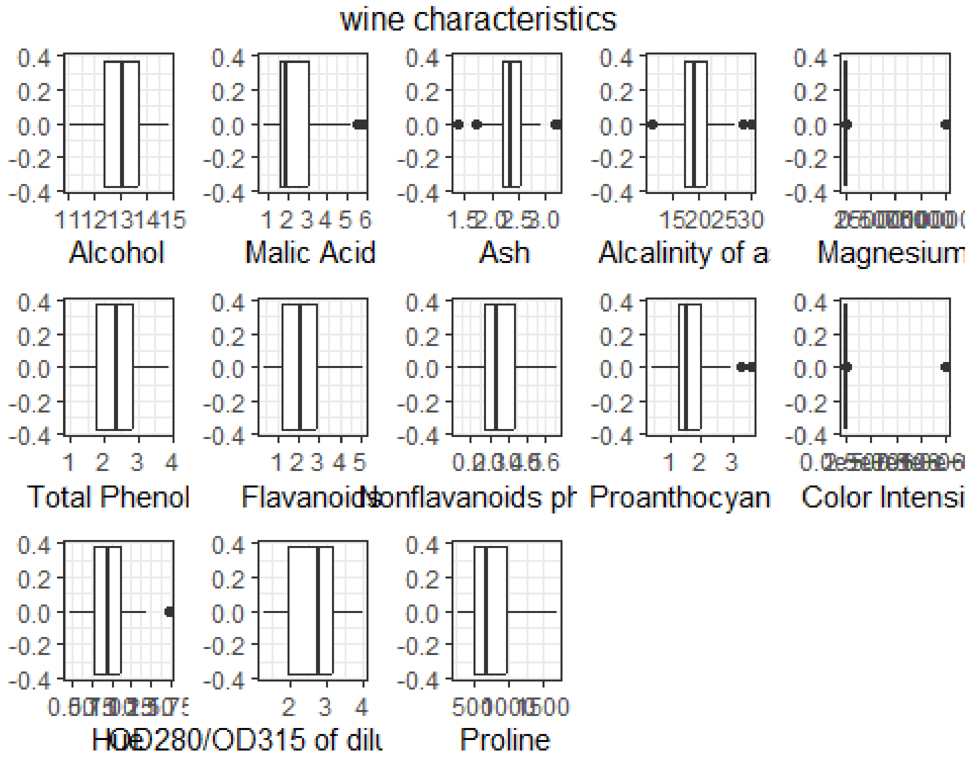
The purpose of the summary() function is to provide a quick and easy way to obtain a summary of the data in an object. This can be useful for exploring and understanding the characteristics of the data, identifying potential outliers or anomalies, and checking the validity and quality of the data. The summary() function can also be useful for comparing the data in different objects or groups, or for identifying trends or patterns in the data.

```
##    Cultivars        Alcohol        Malic Acid         Ash
##  Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360
##  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.597   1st Qu.:2.210
##  Median :2.000   Median :13.05   Median :1.870   Median :2.360
##  Mean   :1.938   Mean   :13.00   Mean   :2.343   Mean   :2.365
##  3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.105   3rd Qu.:2.555
##  Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230
##                                  NA's   :2       NA's   :3
##  Alcalinity of ash   Magnesium       Total Phenols     Flavanoids
##  Min.   :11.20     Min.   :   70.0   Min.   :0.980   Min.   :0.340
##  1st Qu.:17.40     1st Qu.:   88.0   1st Qu.:1.740   1st Qu.:1.200
##  Median :19.50     Median :   98.0   Median :2.350   Median :2.130
##  Mean   :19.59     Mean   :  670.3   Mean   :2.294   Mean   :2.022
##  3rd Qu.:21.50     3rd Qu.:  107.0   3rd Qu.:2.800   3rd Qu.:2.860
##  Max.   :30.00     Max.   :99999.0   Max.   :3.880   Max.   :5.080
##  NA's   :5         NA's   :3         NA's   :1       NA's   :1
##  Nonflavanoids phenols Proanthocyanins Color Intensity        Hue
##  Min.   :0.1300        Min.   :0.410   Min.   :      1   Min.   :0.4800
##  1st Qu.:0.2700        1st Qu.:1.250   1st Qu.:      3   1st Qu.:0.7825
##  Median :0.3400        Median :1.550   Median :      5   Median :0.9650
##  Mean   :0.3627        Mean   :1.586   Mean   :  55623   Mean   :0.9574
##  3rd Qu.:0.4400        3rd Qu.:1.950   3rd Qu.:      6   3rd Qu.:1.1200
##  Max.   :0.6600        Max.   :3.580   Max.   :9899999   Max.   :1.7100
##  NA's   :1             NA's   :1
##  OD280/OD315 of diluted wines    Proline
##  Min.   :1.270                Min.   : 278.0
##  1st Qu.:1.930                1st Qu.: 500.5
##  Median :2.780                Median : 673.5
##  Mean   :2.608                Mean   : 746.9
##  3rd Qu.:3.170                3rd Qu.: 985.0
##  Max.   :4.000                Max.   :1680.0
##  NA's   :1
```

From the analysis, I realize there are missing values in my data for the, total phenols, flavanoids, non flavanoids, and OD280/OD315 of diluted wines columns.So i went further to write a program that replaces all missing values with the median observations and a program that detects all high leverage points and deletes all rows that have those points.

```
##   Cultivars Alcohol Malic Acid  Ash Alcalinity of ash Magnesium Total
Phenols
## 1         1   14.23       1.71 2.36              15.6       127
2.80
## 2         1   13.20       1.78 2.14              11.2       100
2.65
## 3         1   13.16       2.36 2.67              18.6       101
2.80
## 4         1   14.37       1.95 2.50              16.8       113
3.85
## 5         1   13.24       2.59 2.87              21.0       118
2.80
## 6         1   14.20       1.76 2.45              15.2       112
3.27
##   Flavanoids Nonflavanoids phenols Proanthocyanins Color Intensity  Hue
## 1       3.06                  0.28            2.29            5.64 1.04
## 2       2.76                  0.26            1.28            4.38 1.05
## 3       3.24                  0.30            2.81            5.68 1.03
## 4       3.49                  0.24            2.18            7.80 0.86
## 5       2.69                  0.39            1.82            4.32 1.04
## 6       3.39                  0.34            1.97            6.75 1.05
##   OD280/OD315 of diluted wines Proline
## 1                         3.92    1065
## 2                         3.40    1050
## 3                         3.17    1185
## 4                         3.45    1480
## 5                         2.93     735
## 6                         2.85    1450
```

- Plotting boxplots for all numeric columns to see if we have outliers. The blank white portion of the boxplots without grids is our Interquartile range(IQR). The black thick line that passes through the blank portion of our boxplots is our median. Any data points that lie 1.5 times of IQR above Q3 and below Q1 are outliers. The tail from the left hand side is our lower bound(q1 - (1.5 * IQR)). Moving towards the blank region from the left we have the 25th percentile(q1) before crossing the median and then the 75th percentile(q3 which is the end of the blank region. The tail at the right end is our upper bound(q3 + (1.5 * IQR)).



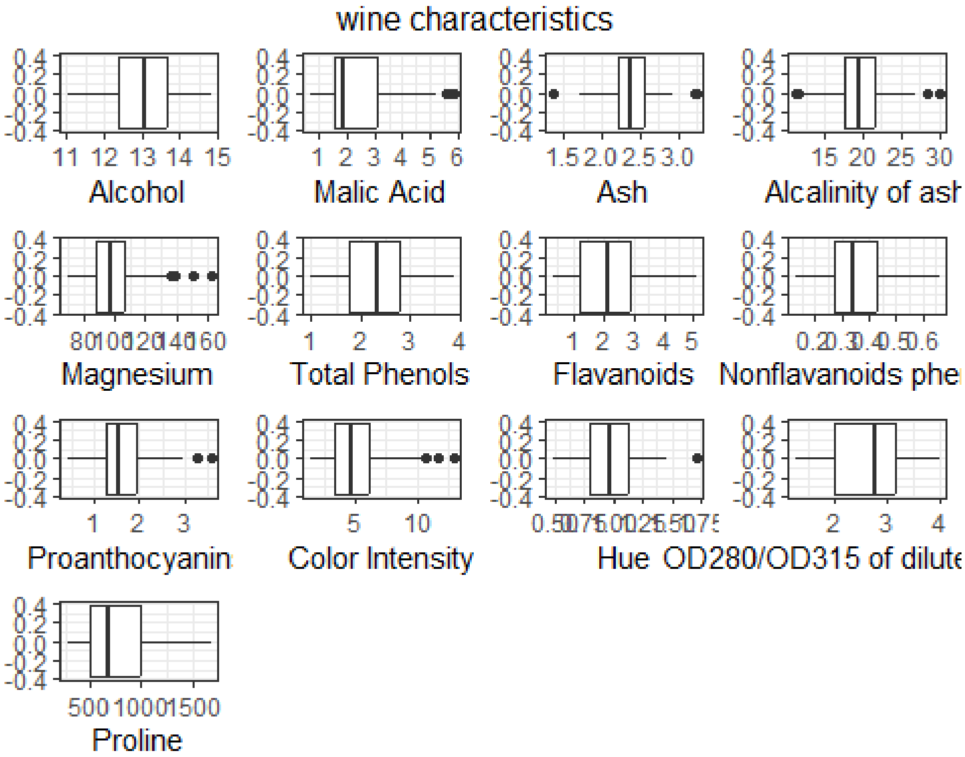
- Using matrix approach to detect outliers. I computed the hat matrix and printed the values of the diagonal of the hat matrix which are greater than the condition 2p/n (p = predictors, n = total observations) and considered them to be high leverage points.

```
## [1]  60  74 111 122 159 172 177
```

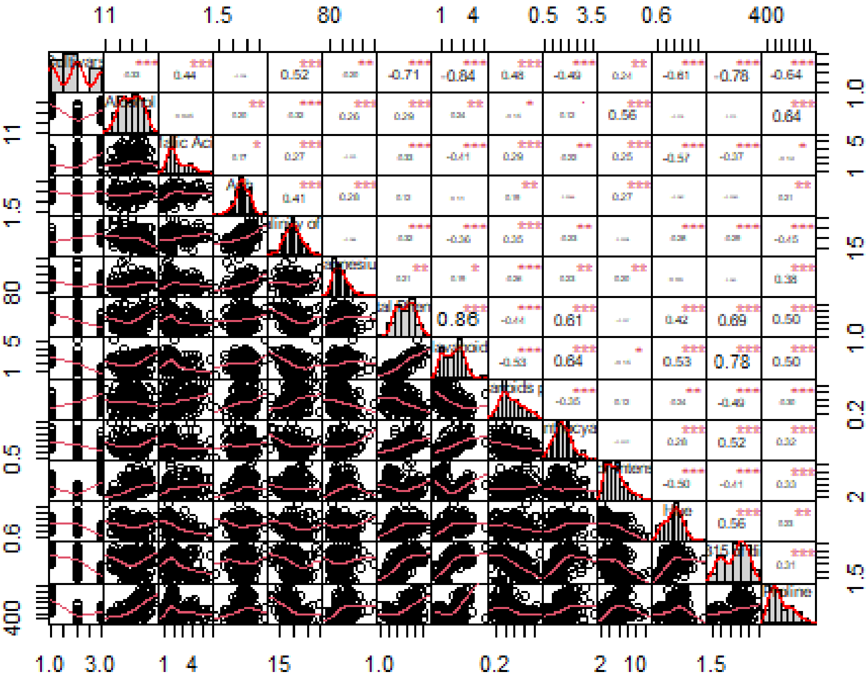
- I went further to compute the cook's distance which measures the influence of the ith observation (yi) on all fitted values. The ith observation is influential if the cook's distance is greater than F(0.5, p, n-p). In my work, I went ahead to remove all observations that were inflential.

```
## [1] 0.9567062

## [1] 542073156263
## [1] 3690.944

## 172 177
## 172 177
##   Cultivars Alcohol Malic Acid  Ash Alcalinity of ash Magnesium Total
Phenols
## 1         1   14.23       1.71 2.36              15.6       127
2.80
## 2         1   13.20       1.78 2.14              11.2       100
2.65
## 3         1   13.16       2.36 2.67              18.6       101
2.80
## 4         1   14.37       1.95 2.50              16.8       113
3.85
## 5         1   13.24       2.59 2.87              21.0       118
2.80
## 6         1   14.20       1.76 2.45              15.2       112
3.27
##   Flavanoids Nonflavanoids phenols Proanthocyanins Color Intensity  Hue
## 1       3.06                  0.28            2.29            5.64 1.04
## 2       2.76                  0.26            1.28            4.38 1.05
## 3       3.24                  0.30            2.81            5.68 1.03
## 4       3.49                  0.24            2.18            7.80 0.86
## 5       2.69                  0.39            1.82            4.32 1.04
## 6       3.39                  0.34            1.97            6.75 1.05
##   OD280/OD315 of diluted wines Proline
## 1                         3.92    1065
## 2                         3.40    1050
## 3                         3.17    1185
## 4                         3.45    1480
## 5                         2.93     735
## 6                         2.85    1450
```

Boxplots after outliers have been removed



- From the above plots, we can see that the high leverage points that were previously in Magnesium and Color Intensity have been detected and removed.



- Seeing the Correlation (statistical relationship) among the various features of the dataset.
- A positive correlation means a variable increases/decreases as other other variable increases/decreases respectively.
- Negative Correlation means a variable increases/decreases as the other variable decreases/increases respectively.
- The values more close to one indicate strong correlation, the values below 0.5 indicate a low correlation and the values above 0.5 indicate a high correlation.

I compared all the variables in the correlation matrix above. I was interested in the variables that influenced the wine quality as well as relationships between the features themselves. These were my main findings:

- Our wine types have high positive correlation with Alcalinity of ash.
- They wines also have a high negative correlation with Total phenols, flavanoids, Hue, Dilution and Proline.
- Our classes have low negative correlation with Alcohol, Magnesium and Proanthocyanins.
- They is a low positive correlation between malic acid, nonflavanoids phenols and color intensity.
- Also, the alcohol content has a high positive correlation with color intensity and proline.
- Total phenols has a high positive correlation with flavanoids, proanthocyanins, dilution and proline.
- Flavanoids have a high positive correlation with proanthocyanins, hue, dilution and proline.

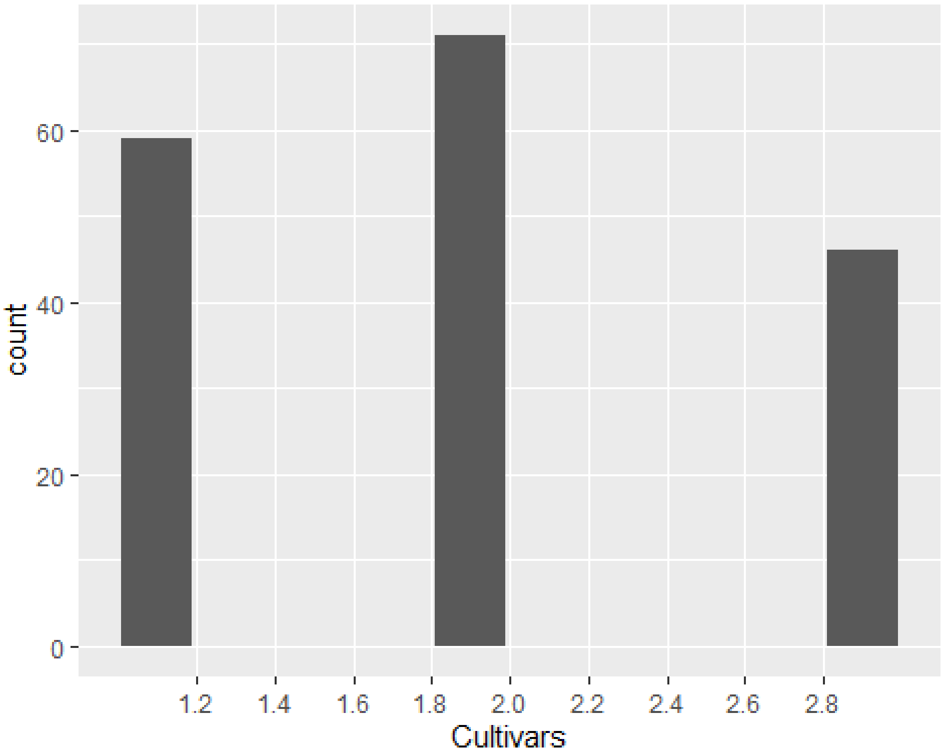
TEST FOR MULTICOLLINEARITY

- Multicollinearity occurs when independent variables in a regression model are correlated. Multicollinearity causes a lot of things such as the estimated standard deviations of the regression coefficients becoming large when predictor variables in the model are highly correlated. Also the extra sum of squares associated with a predictor variables may vary. There are informal ways to check for multicollinearity but in this project I used the variance inflation factor.

- I used variance inflation factor (VIF) method which is a formal method of detecting the presence of multicollinearity to measure how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. Any VIF value greater than 4 or 5 is considered a severe case of multicollinearity.

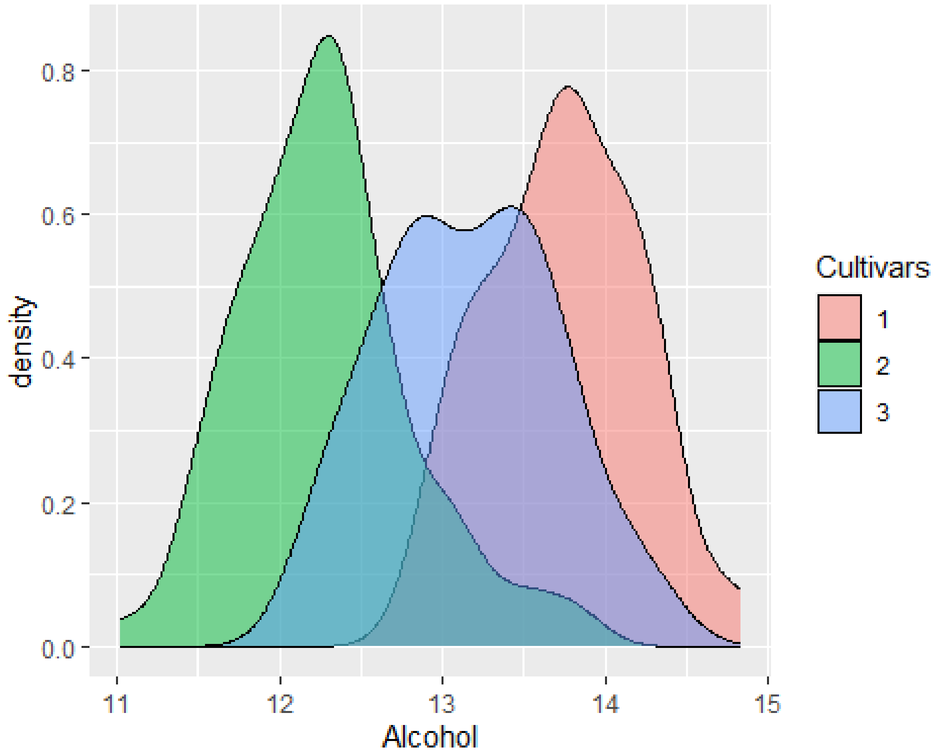
```
## [1] 2.769792
```

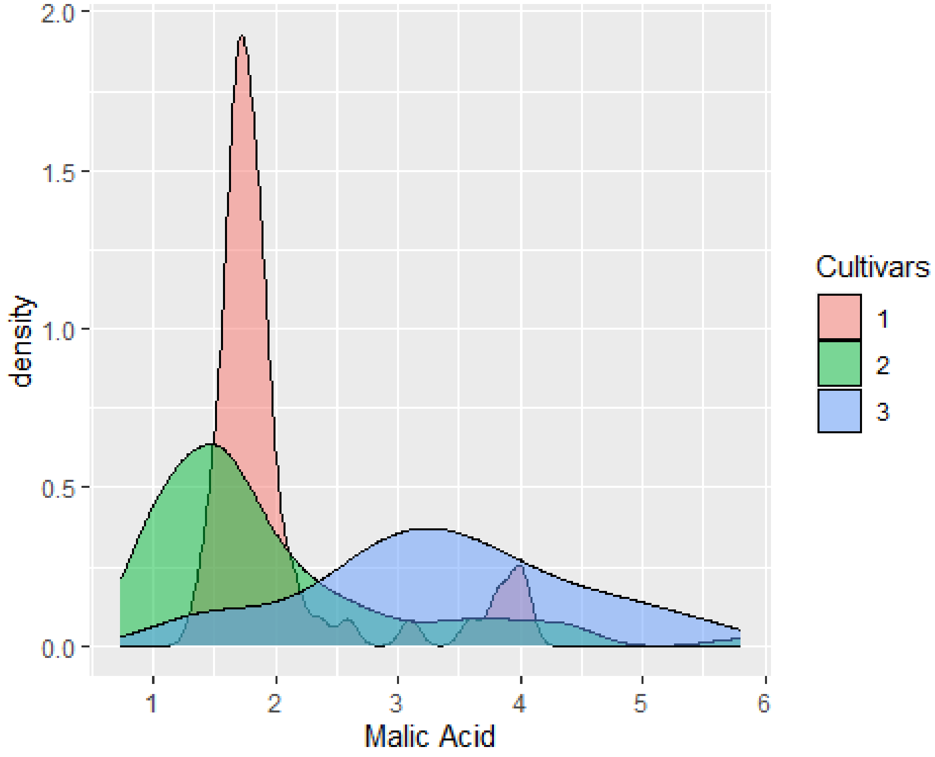
- From the analysis, the mean of the variance inflation factor is 2.769792 and it indicates that it is not a severe case of of multicollinearity.

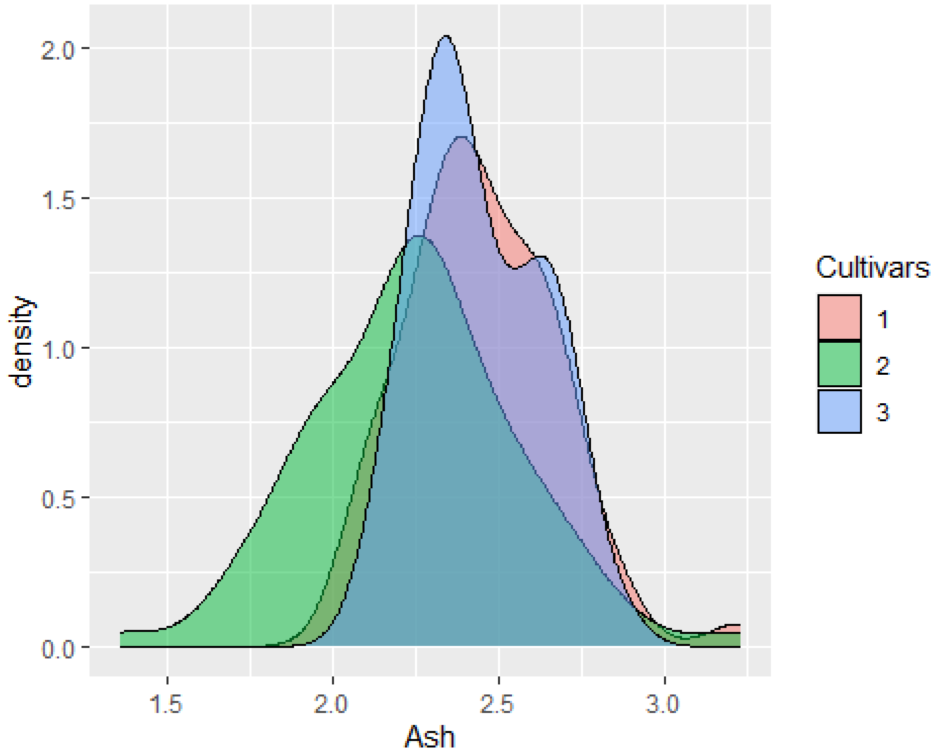


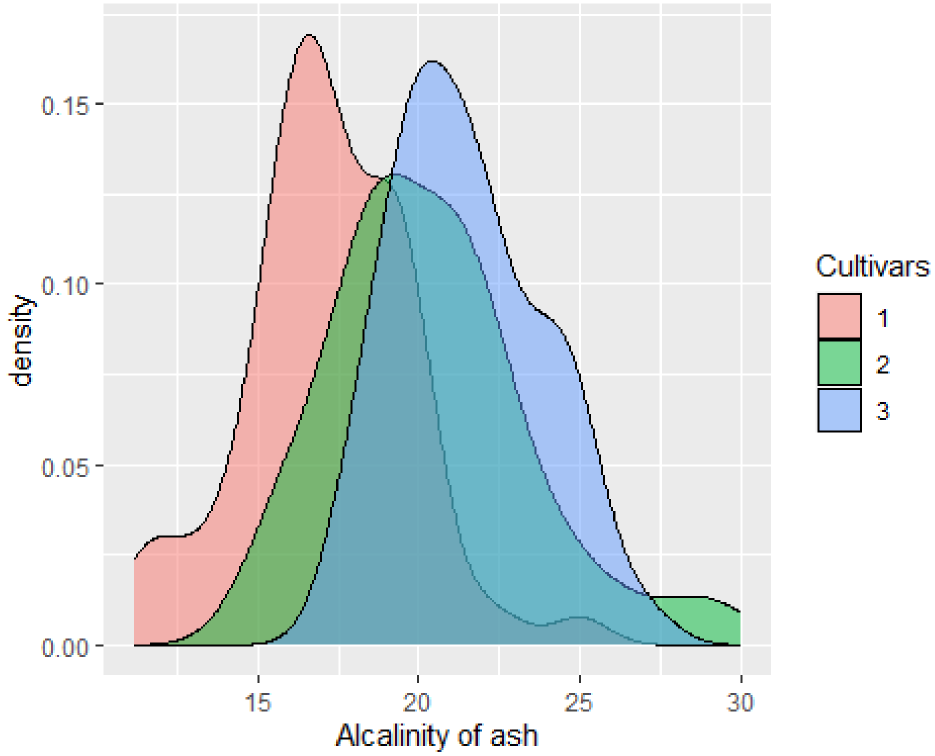
- From the bar chart above, cultivar 2 had the highest number of occurrence in the data while cultivar 3 has the lowest count in our data.

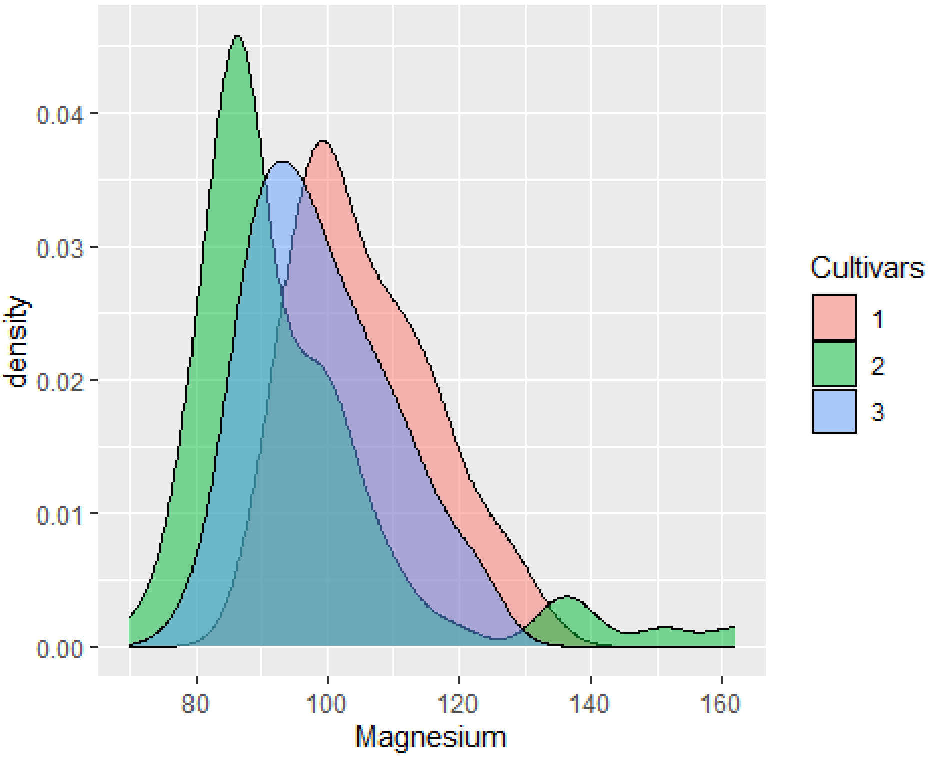
- Next, I want to visualize the variables by class. To do this, I will create density plots of the variables and overlap them by class. The density plots can provide useful insights into the distributions of the different variables within each cultivar of wine and can help you to understand how the different variables are distributed within each cultivar.

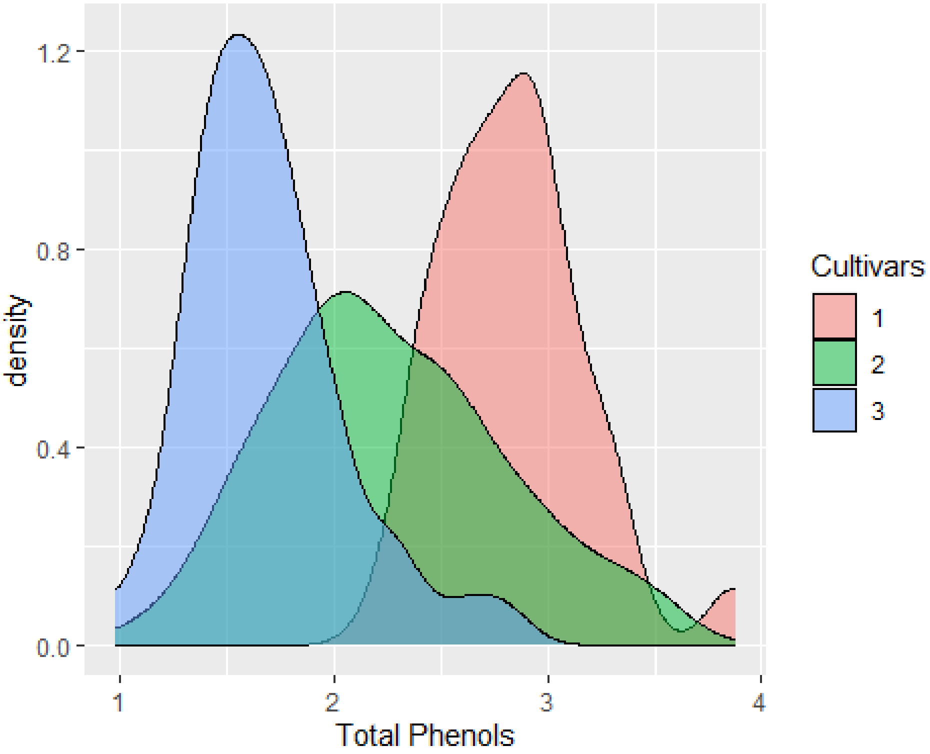

0.8
0.6
0.4
0.2
0.0
density
11
12
13
14
15
Alcohol
Cultivars
1
2
3

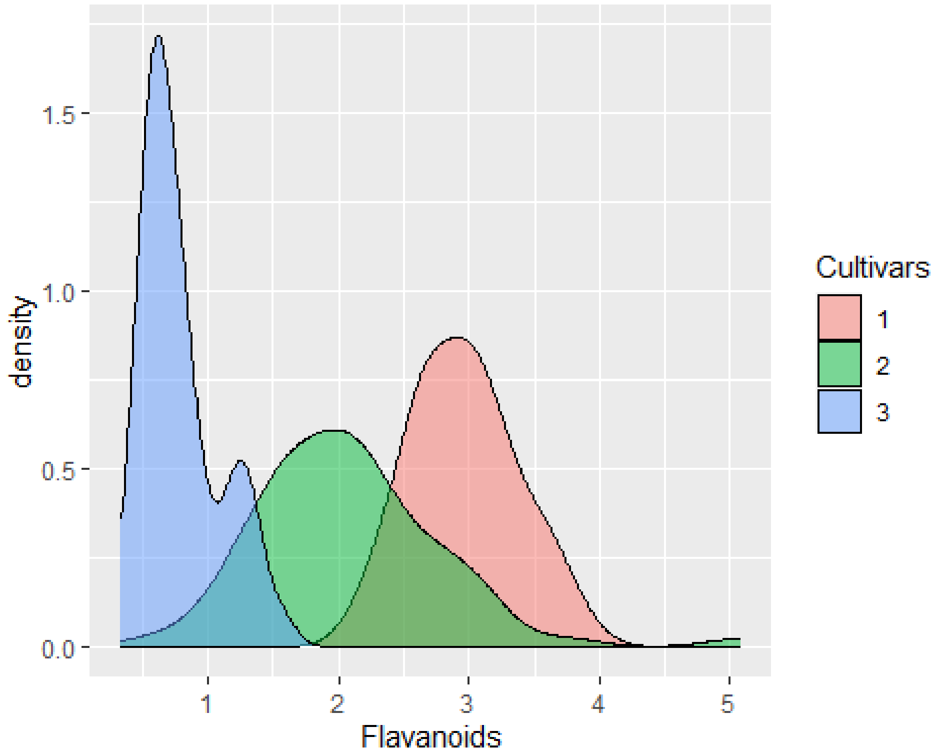

2.0
1.5
1.0
0.5
0.0
density
1
2
3
4
5
6
Malic Acid
Cultivars
1
2
3

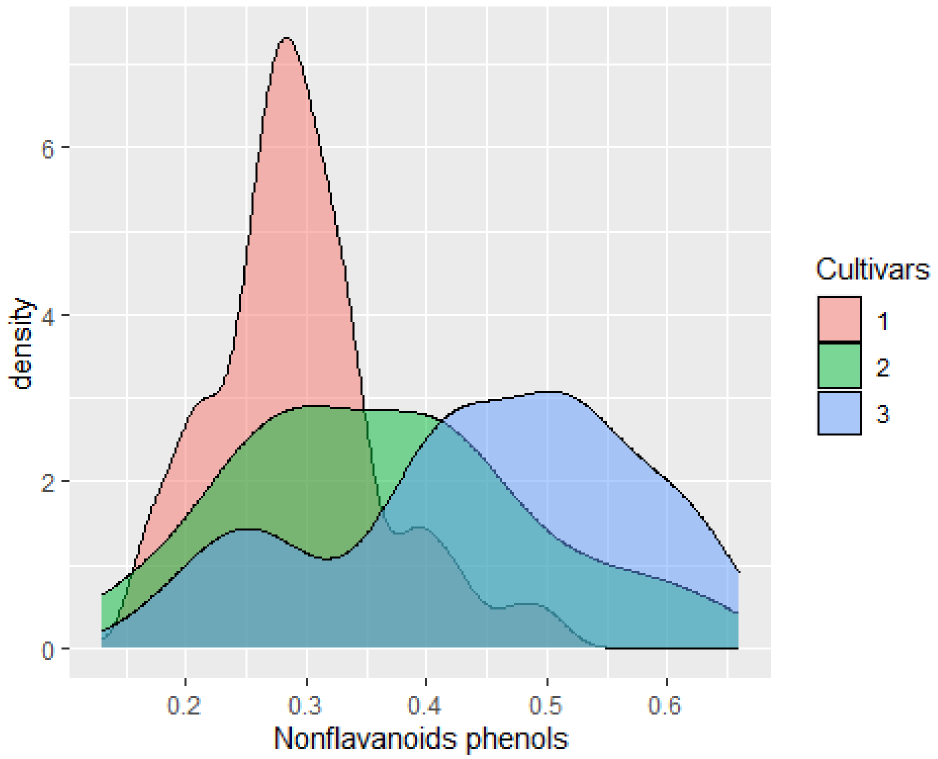

2.0
1.5
1.0
0.5
0.0
density
1.5
2.0
2.5
3.0
Ash
Cultivars
1
2
3

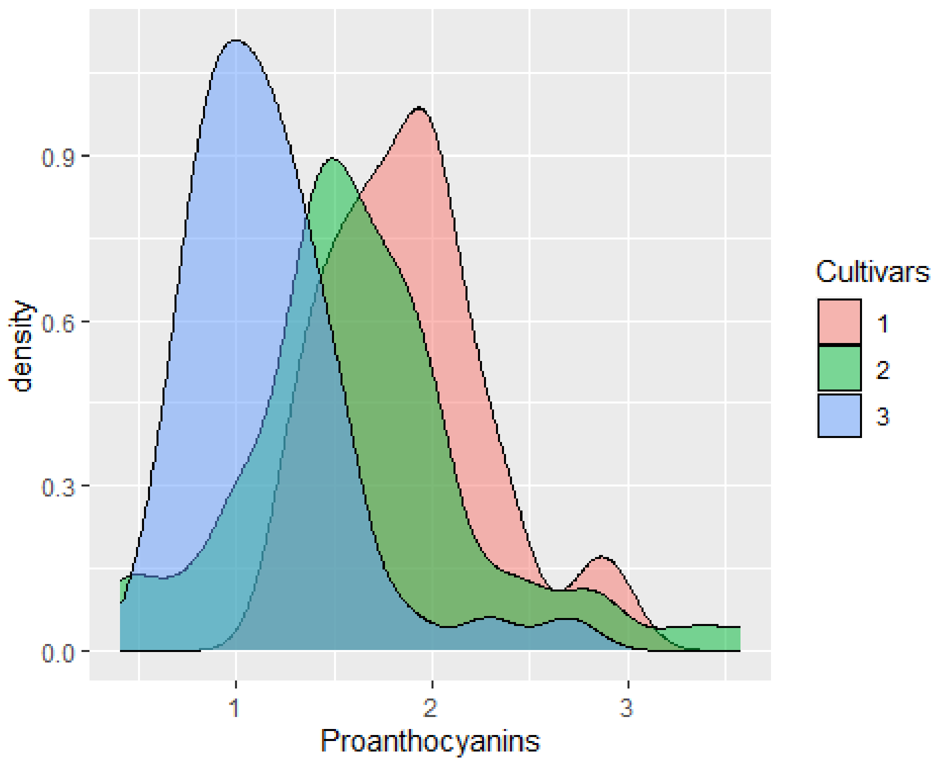

0.15
0.10
0.05
0.00
density
15
20
25
30
Alcalinity of ash
Cultivars
1
2
3

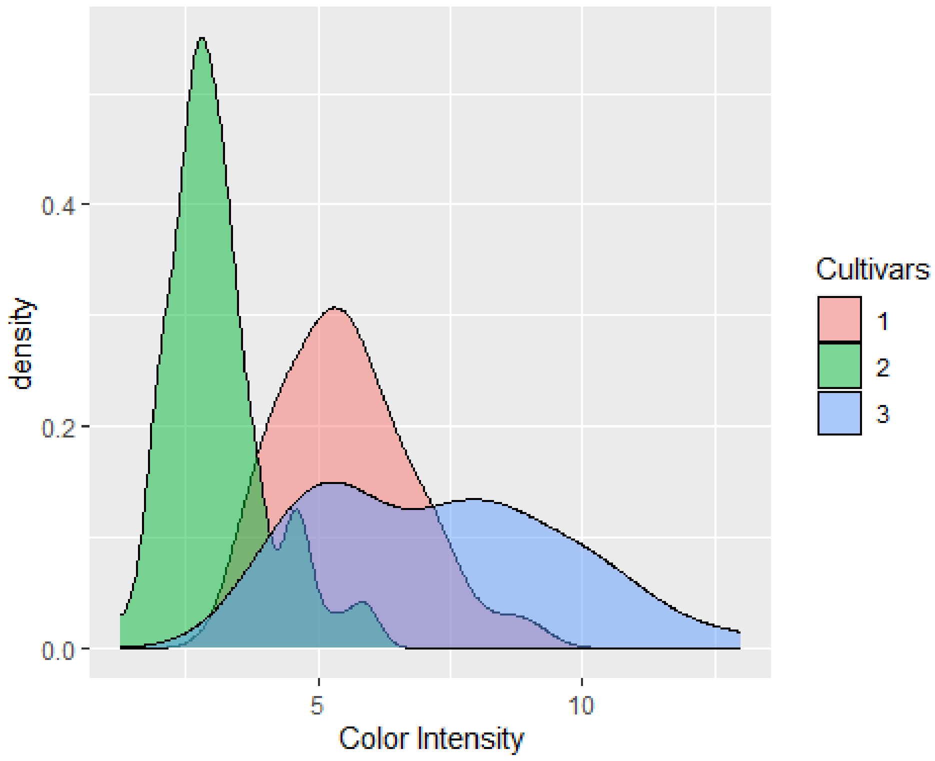

density
0.04
0.03
0.02
0.01
0.00
80
100
120
140
160
Magnesium
Cultivars
1
2
3

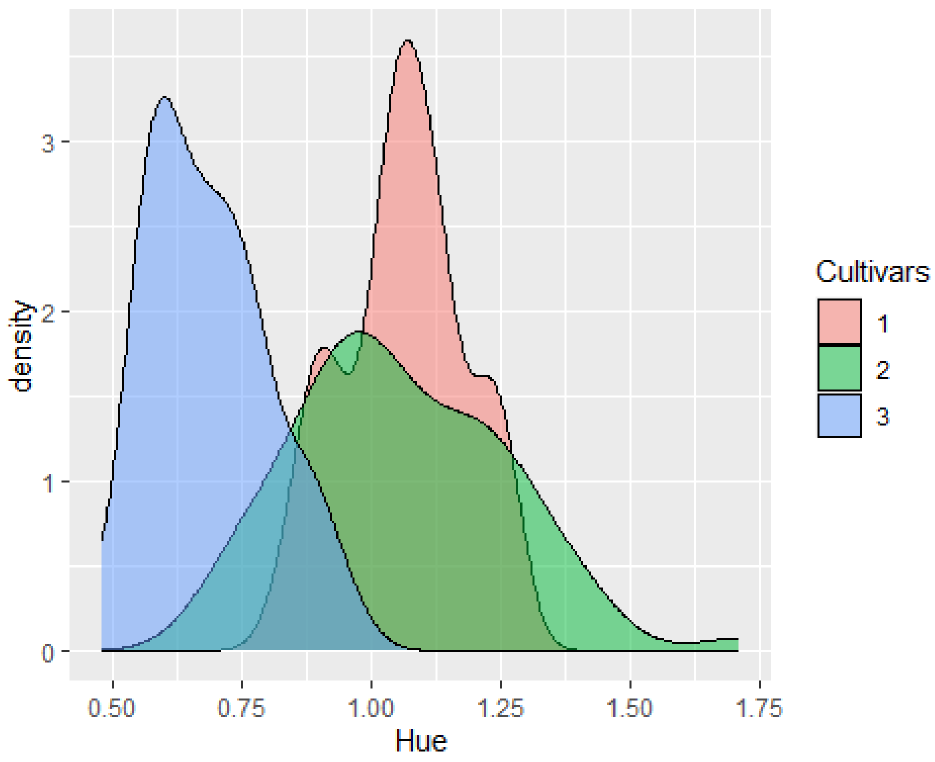

density
1.2
0.8
0.4
0.0
1
2
3
4
Total Phenols
Cultivars
1
2
3


density
0.0
0.5
1.0
1.5
1
2
3
4
5
Flavanoids
Cultivars
1
2
3

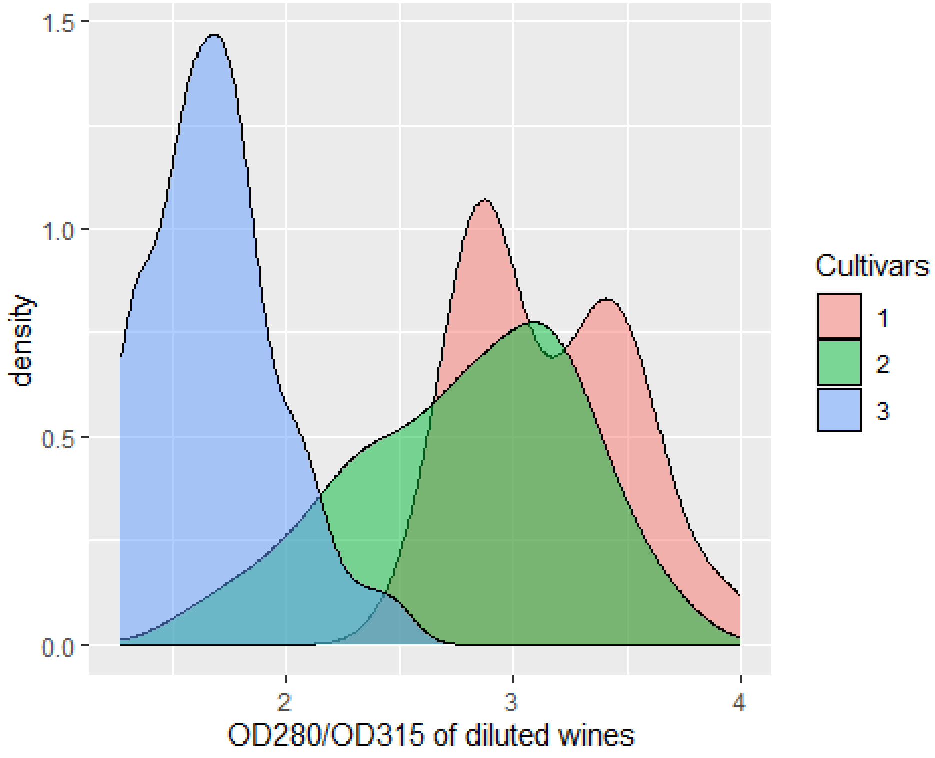

density
0
2
4
6
0.2
0.3
0.4
0.5
0.6
Nonflavanoids phenols
Cultivars
1
2
3


density
0.0
0.3
0.6
0.9
1
2
3
Proanthocyanins
Cultivars
1
2
3

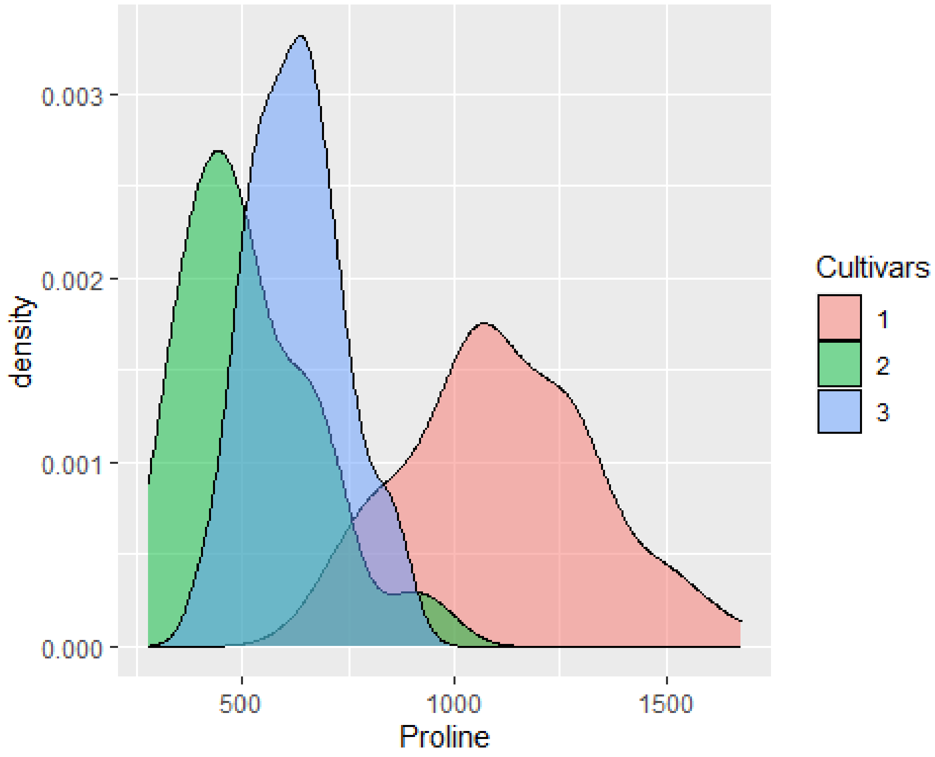

density
0.0
0.2
0.4
5
10
Color Intensity
Cultivars
1
2
3



density
3
2
1
0
0.50
0.75
1.00
1.25
1.50
1.75
Hue
Cultivars
1
2
3



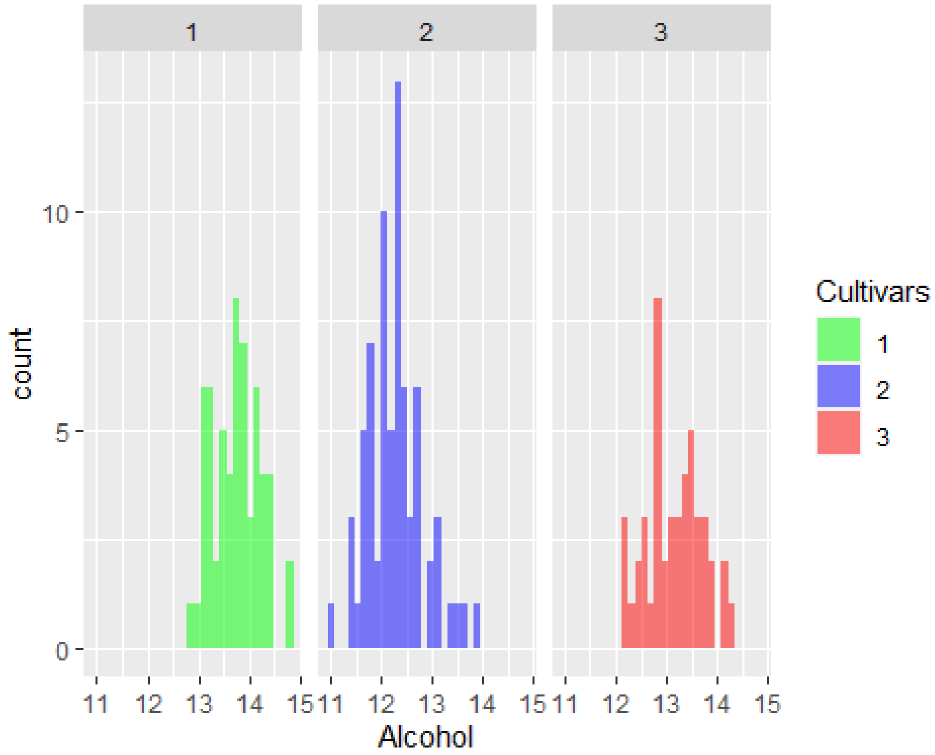
density
1.5
1.0
0.5
0.0
2
3
4
OD280/OD315 of diluted wines
Cultivars
1
2
3

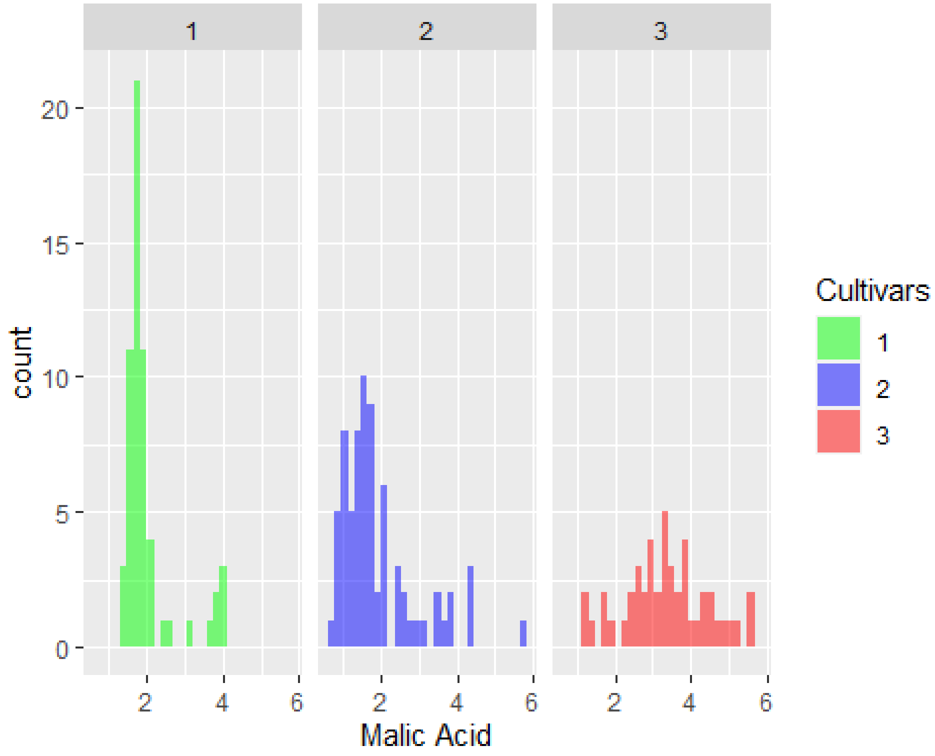


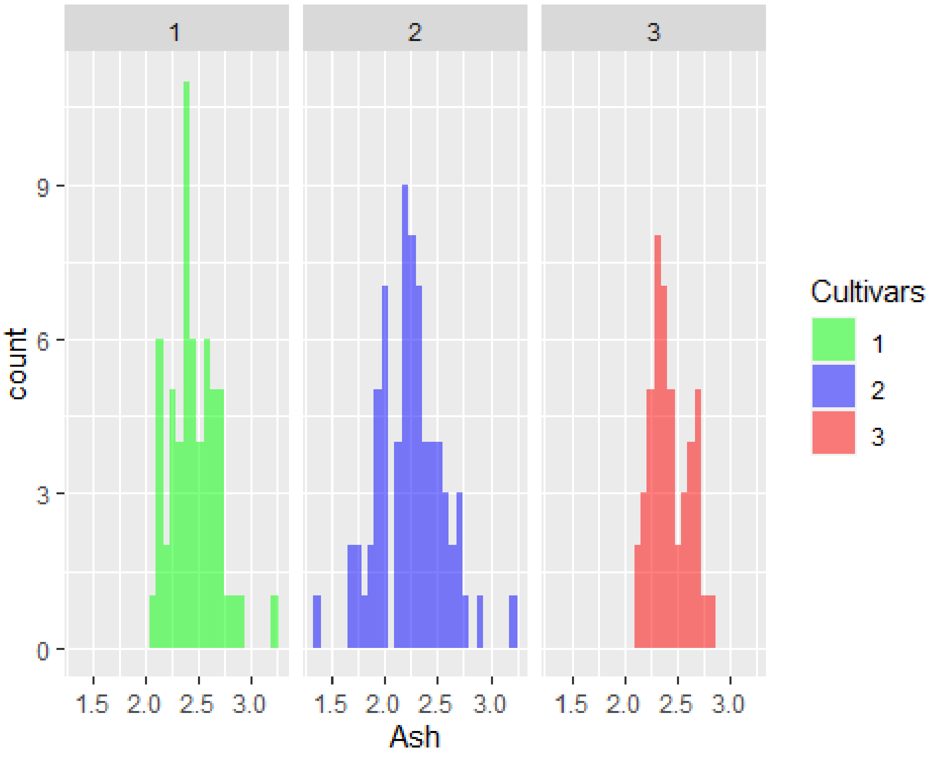
From the plots above, their density distribution vary in density. I can say that:

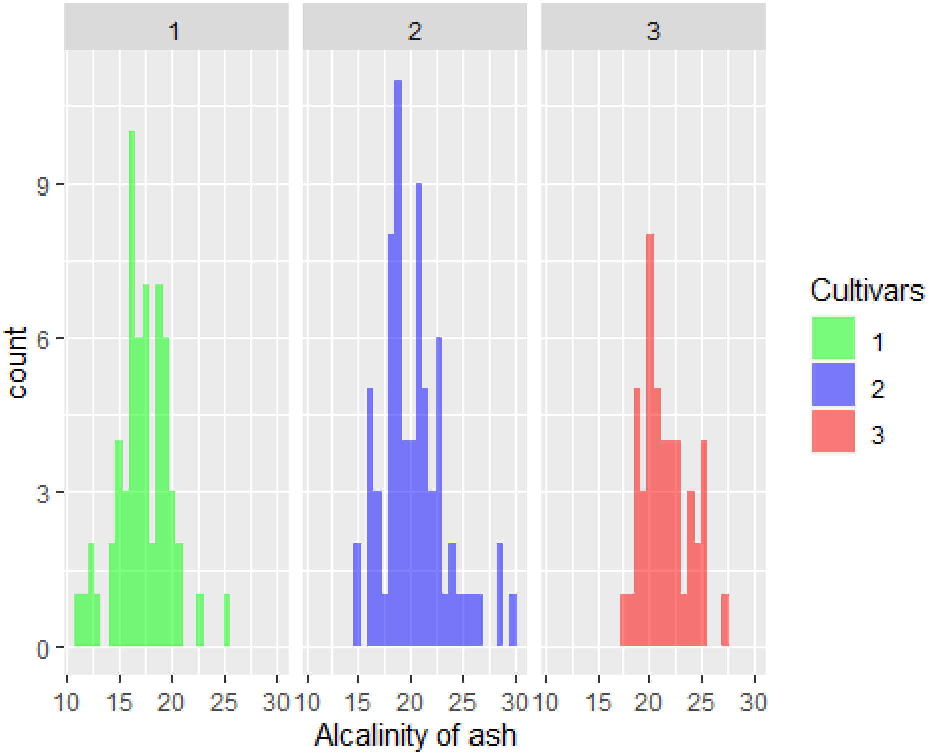
- The distribution for cultivar 3 is symmetric with a range from about 11.6 to 15. Meanwhile the distribution of cultivar 1 and 2 are asymmetric (right skewed)
- From the density plot of malic acid, the distribution of malic acid is wider for cultivar 3 compared to the other cultivars. Cultivar 1 is seen to have a high malic content with right skewed distribution to the right indicating outlying points.
- The distribution of ash for the different wine types are sort of normal with cultivar 2 having a wider range.
- In the case of nonflavanoids, the distribution of cultivars 3 is skewed to the left (asymmetric) with cultivar 1 having a higher Nonflavanoid phenol content.
- In the density plots for proanthocyanins, flavanoids, total phenols and magnesium, all the cultivar types are skewed to the right.

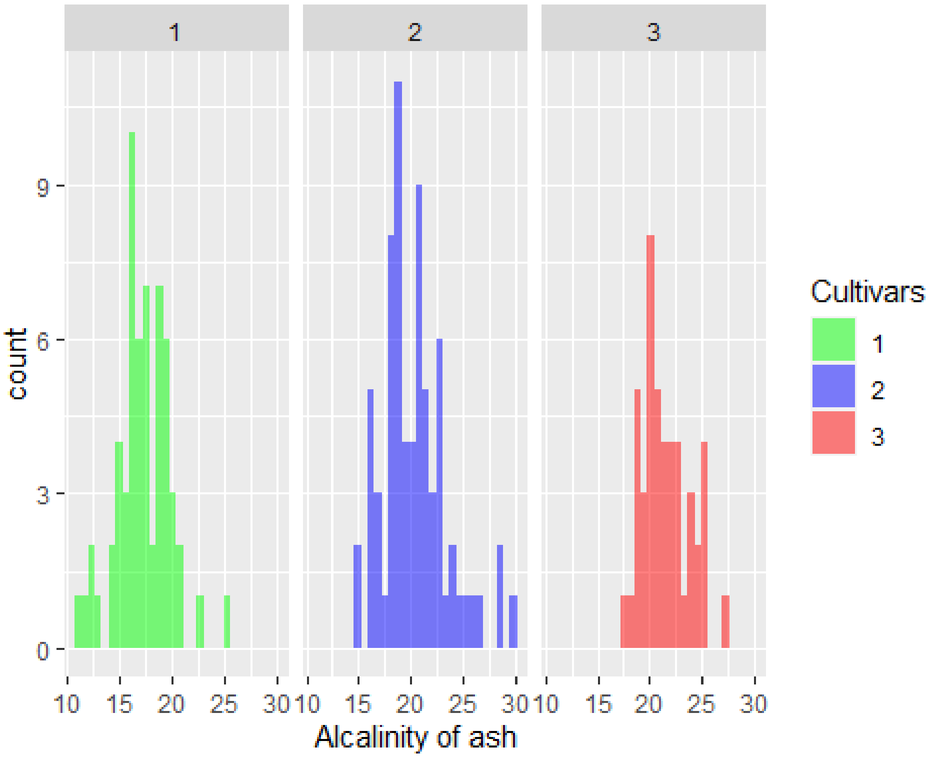
In the plots below, I created individual histograms for each class for each variable. Each histogram shows the distribution of a different wine chemical property, with the bars colored according to the type of wine (Cultivars) and arranged in a grid according to the wine type.

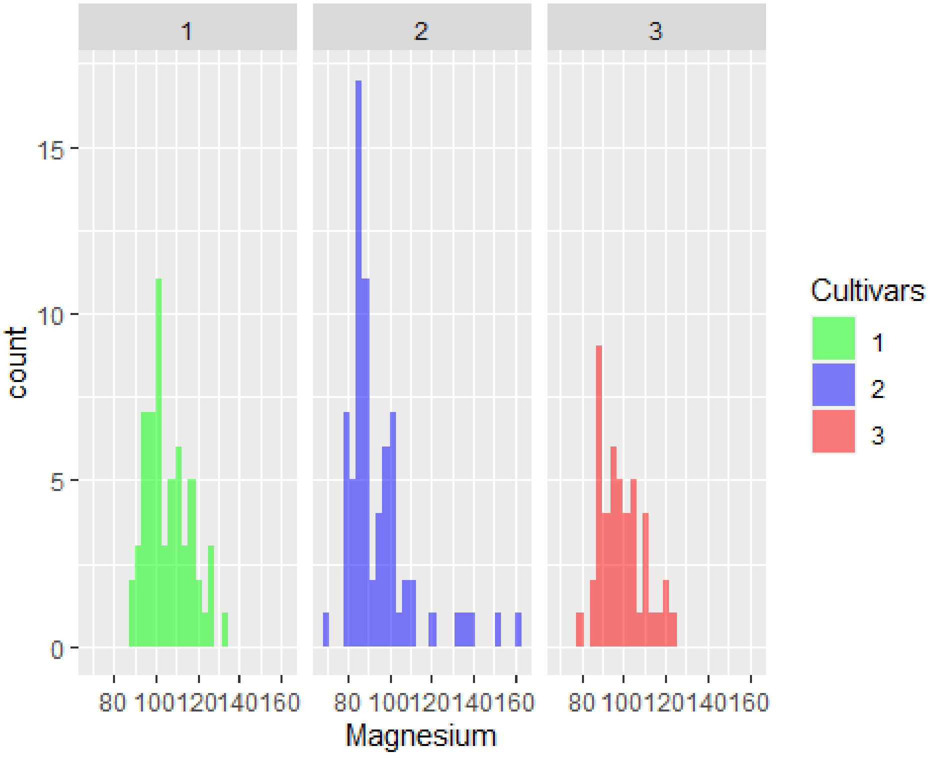

1
2
3
count
0
5
10
11 12 13 14 15
Alcohol
Cultivars
1
2
3

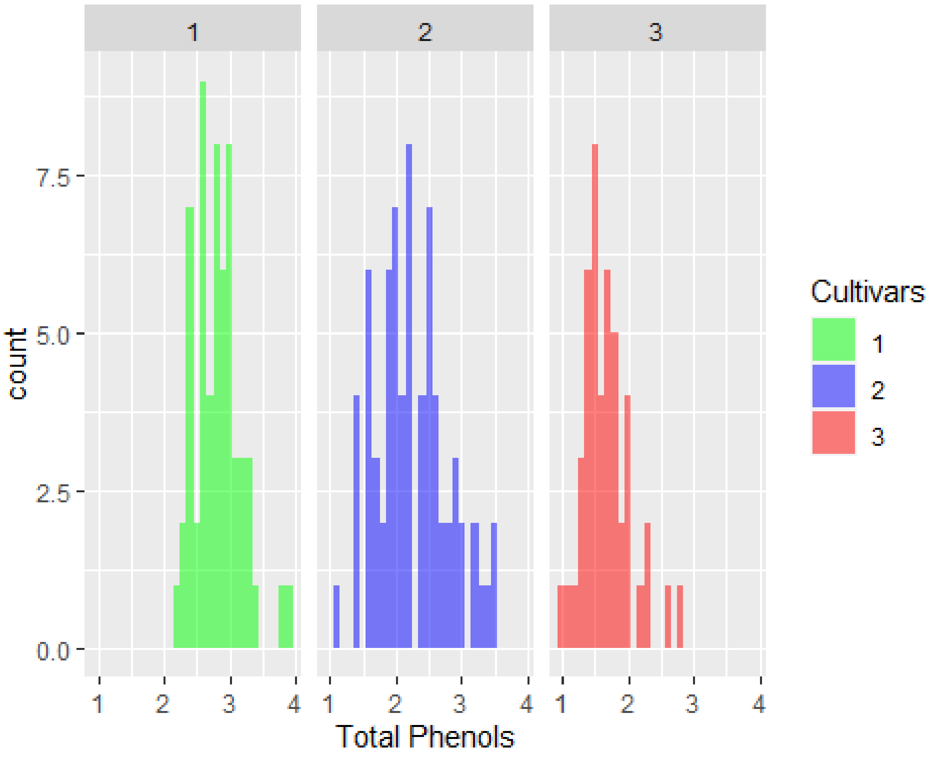

1
2
3
count
0
5
10
15
20
2 4 6
Malic Acid
Cultivars
1
2
3

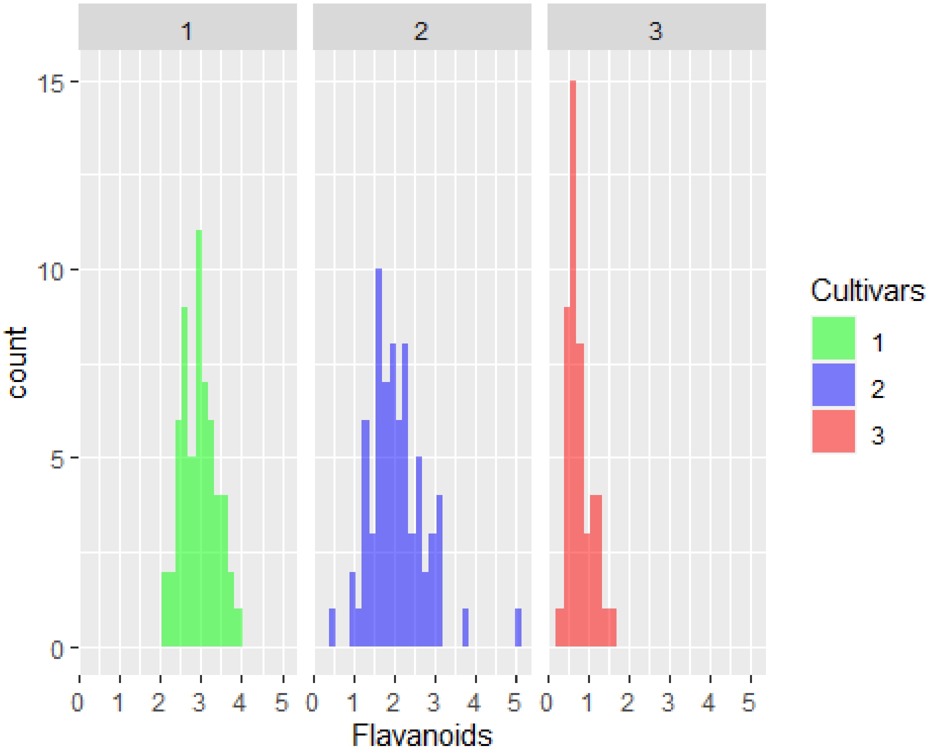

1
2
3
count
9
6
3
0
1.5
2.0
2.5
3.0
Ash
Cultivars
1
2
3

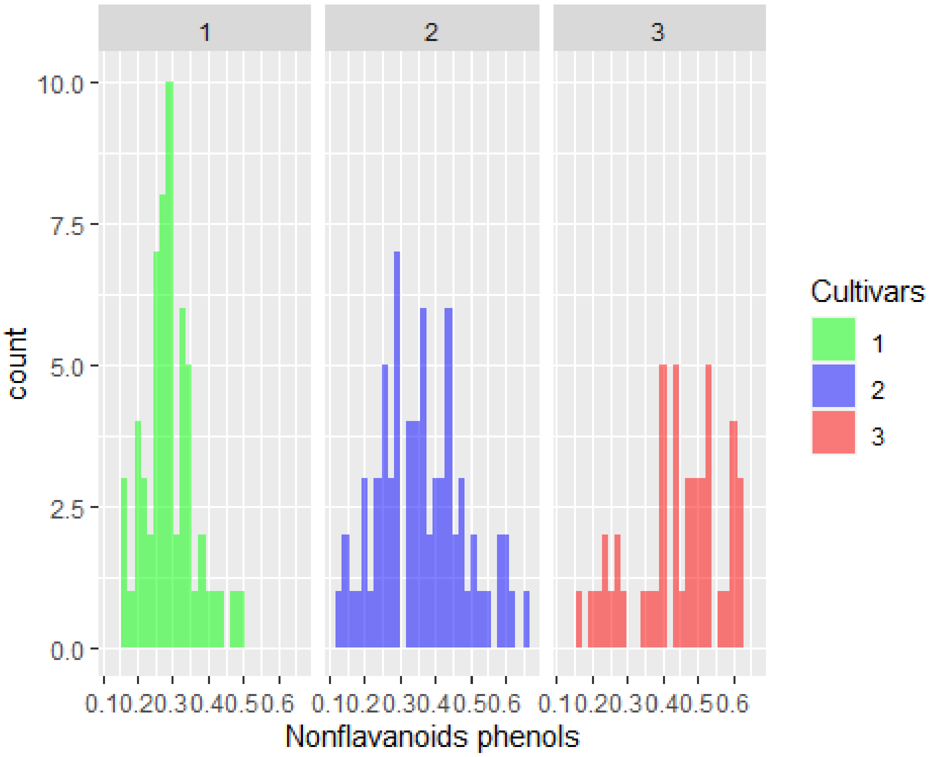

1
2
3
count
9
6
3
0
10
15
20
25
30
Alcalinity of ash
Cultivars
1
2
3

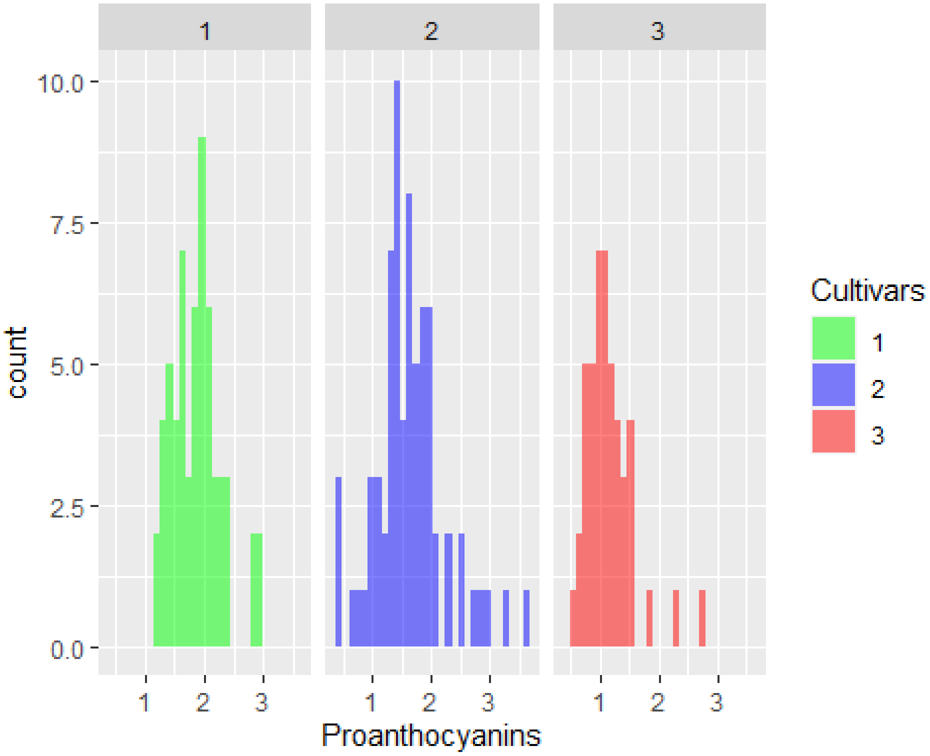

1
2
3
count
0
3
6
9
10 15 20 25 30
10 15 20 25 30
10 15 20 25 30
Alcalinity of ash
Cultivars
🟩 1
🟦 2
🟥 3

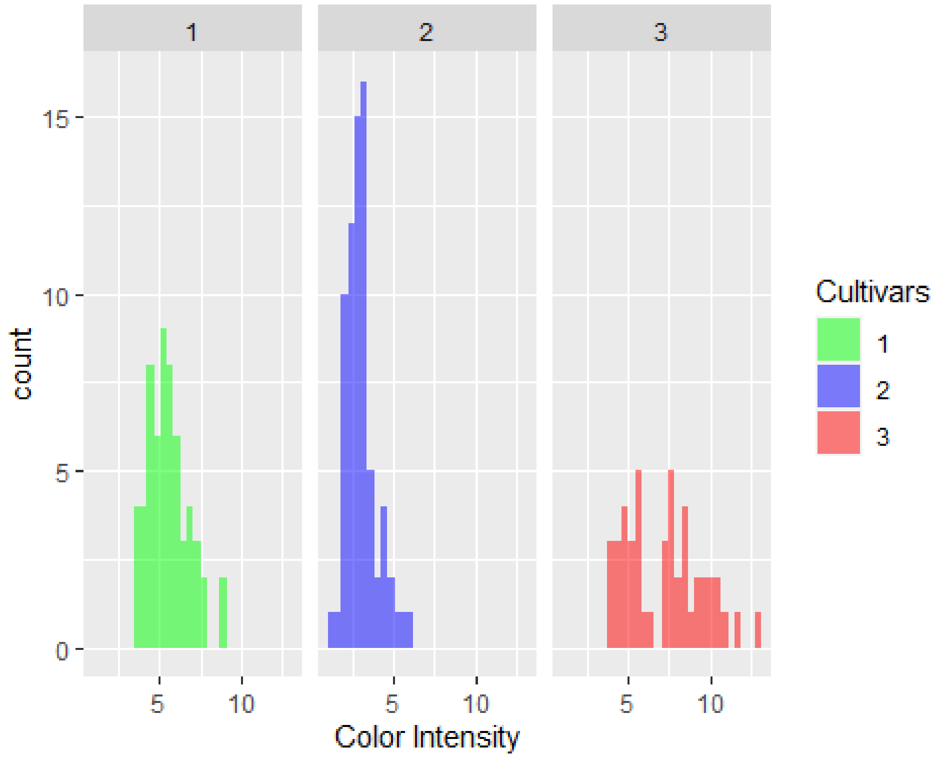

1
2
3
count
0
5
10
15
80 100120140160
80 100120140160
80 100120140160
Magnesium
Cultivars
🟩 1
🟦 2
🟥 3

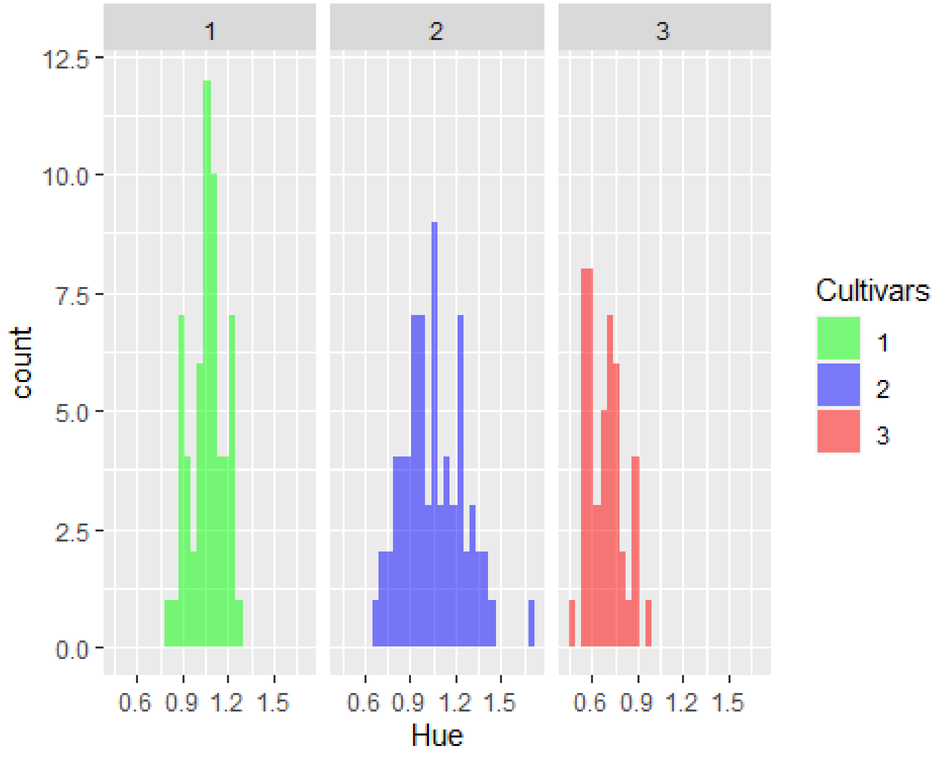

1
2
3
count
7.5
5.0
2.5
0.0
1 2 3 4
1 2 3 4
1 2 3 4
Total Phenols
Cultivars
🟩 1
🟦 2
🟥 3

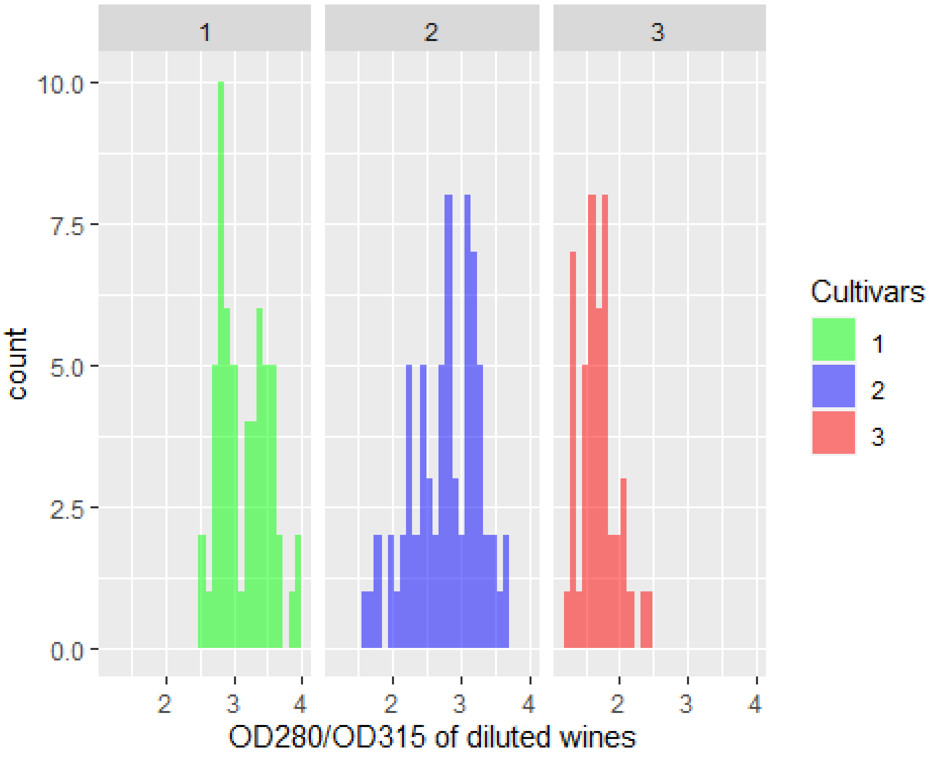

1
2
3
count
15
10
5
0
0 1 2 3 4 5
0 1 2 3 4 5
0 1 2 3 4 5
Flavanoids
Cultivars
🟩 1
🟦 2
🟥 3

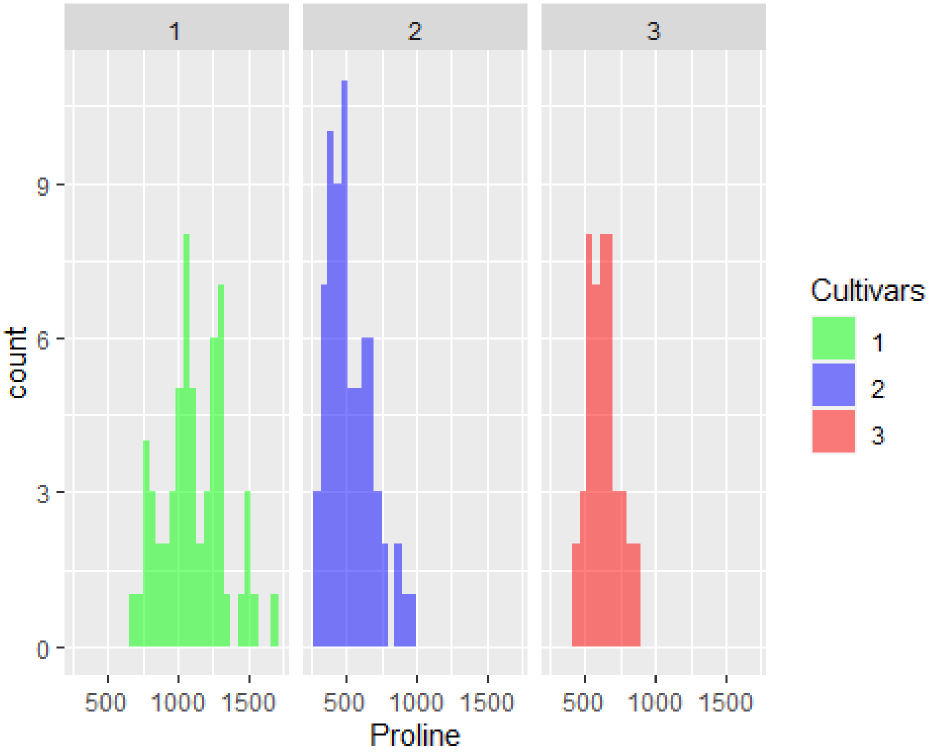

1
2
3
10.0
7.5
5.0
2.5
0.0
count
0.10.20.30.40.50.6
0.10.20.30.40.50.6
0.10.20.30.40.50.6
Nonflavanoids phenols
Cultivars
1
2
3


1
2
3
10.0
7.5
5.0
2.5
0.0
count
1
2
3
1
2
3
1
2
3
Proanthocyanins
Cultivars
1
2
3


1
2
3
count
15
10
5
0
5
10
5
10
5
10
Color Intensity
Cultivars
1
2
3


1
2
3
count
12.5
10.0
7.5
5.0
2.5
0.0
0.6
0.9
1.2
1.5
0.6
0.9
1.2
1.5
0.6
0.9
1.2
1.5
Hue
Cultivars
1
2
3



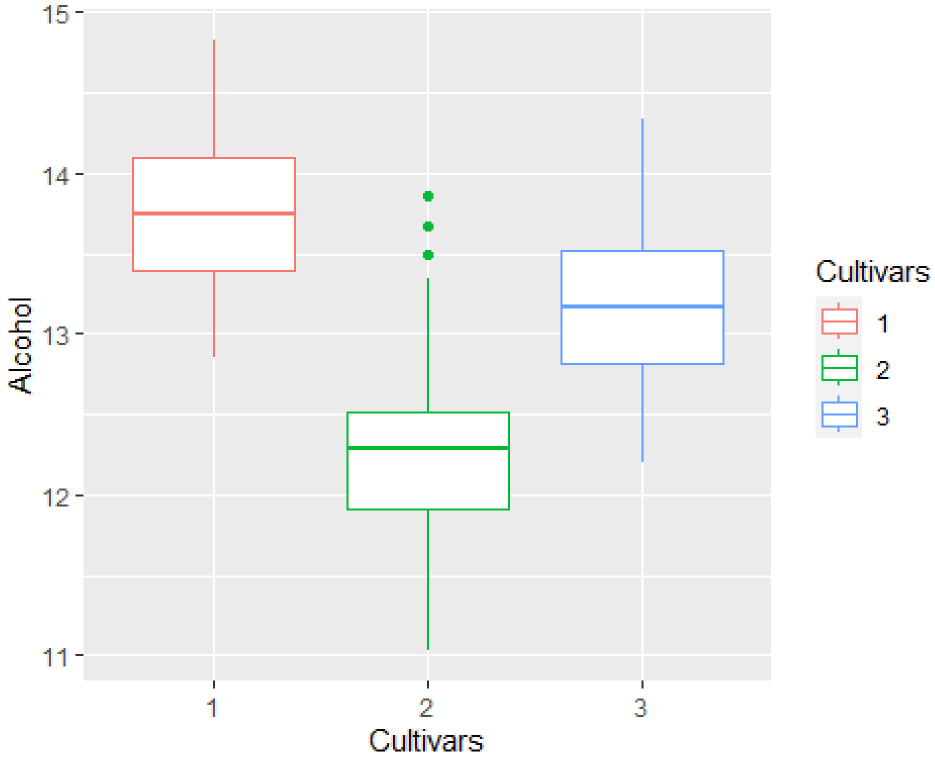


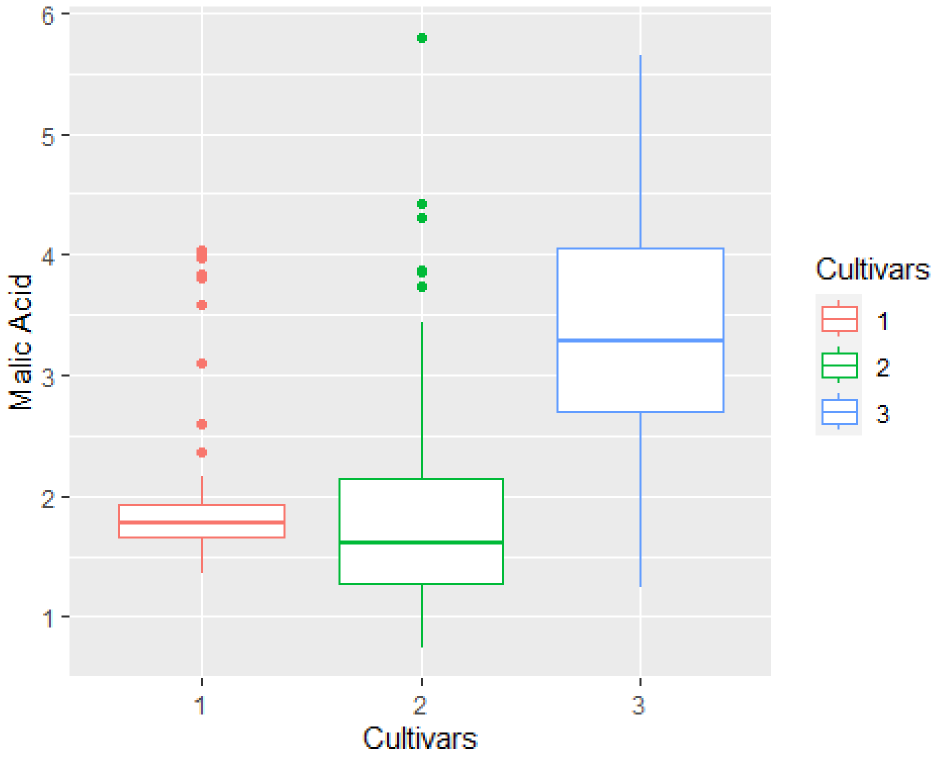
- The majority of the plots above have a non-symmetric multi-modal distribution having several peaks.

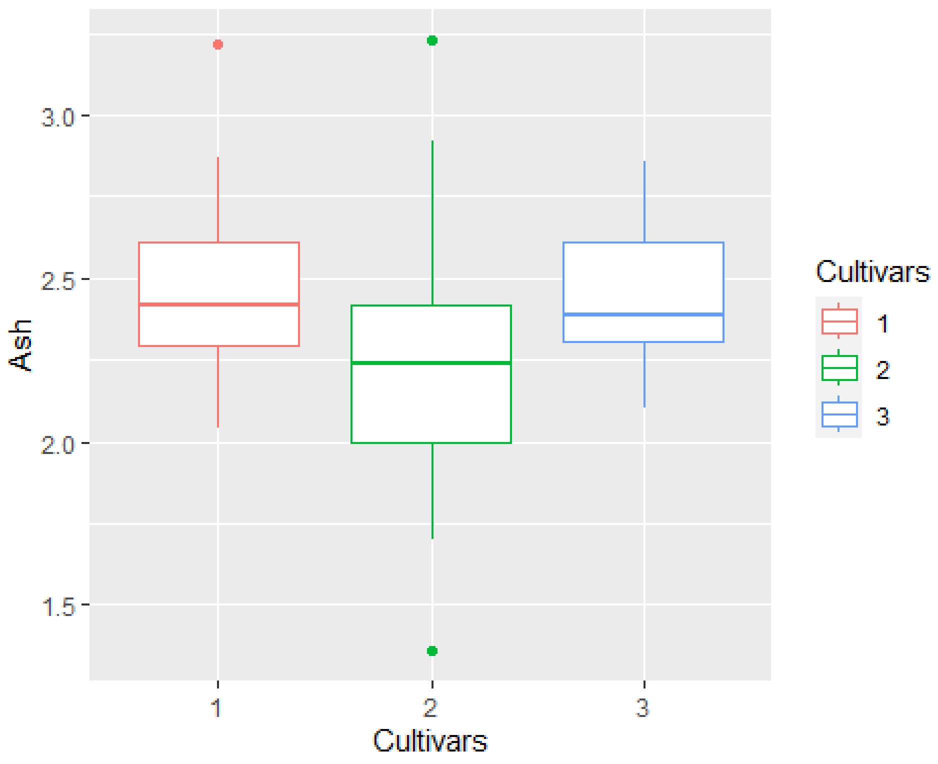
- The histogram of cultivar 3 and proline, shows a symmetric bimodal distribution with just two peaks.

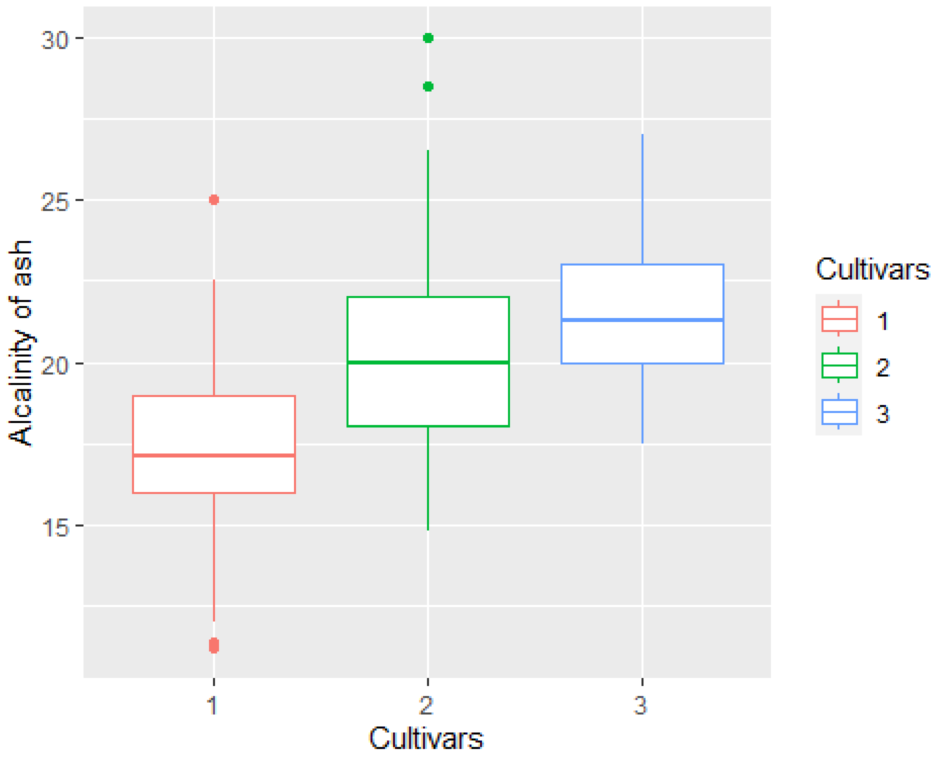
- If the histogram is skewed to the left or right, this indicates that the predictor levels are not evenly distributed and there is a tail of values in one direction. We can see this in the histogram of malic acid and cultivar 2. There, we can see an outlying point close to 6 which seems far away from the rest of the data points.

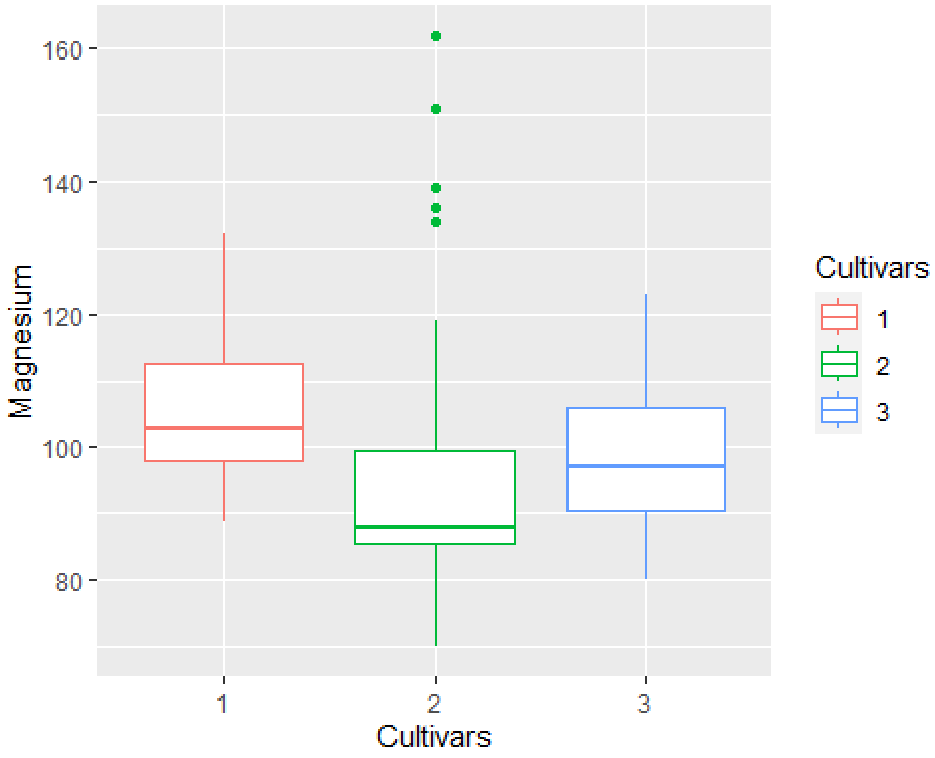
In the code that I provided, boxplots are created for each of the predictor variables in the wine data, with the color indicating the different cultivars (classes) of wine. The x-axis of each plot represents the Cultivars and the y-axis represents the variable being plotted. This means that for each variable, the boxplot will show the distribution of that variable for each cultivar, allowing you to compare the distributions of the different cultivars and see how they differ.

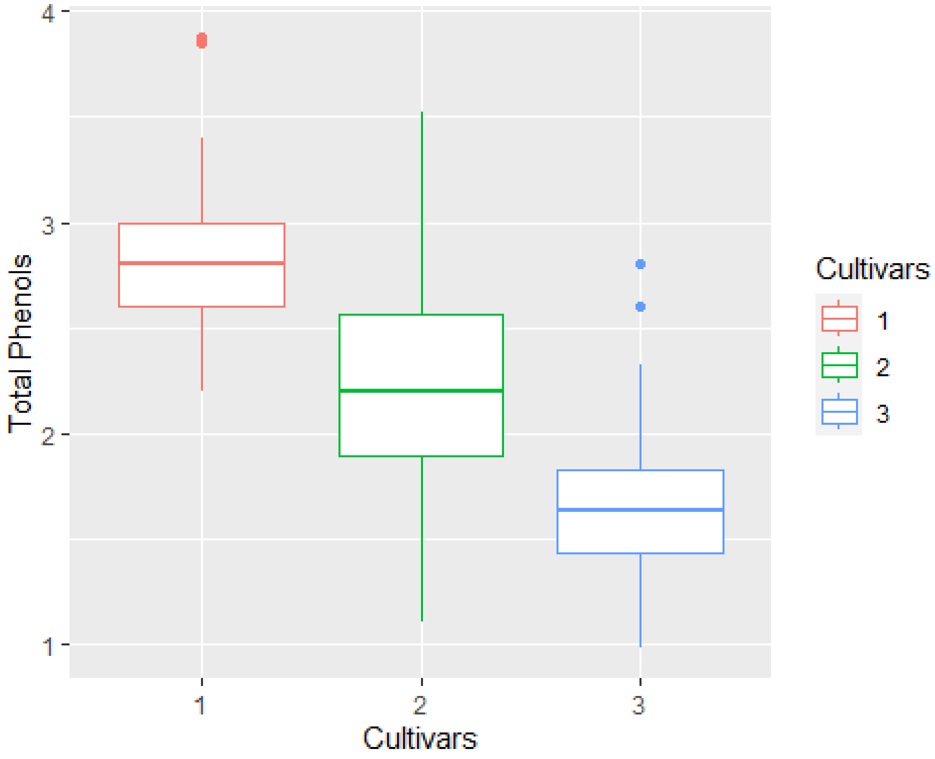

Alcohol
15
14
13
12
11
1
2
3
Cultivars
Cultivars
1
2
3

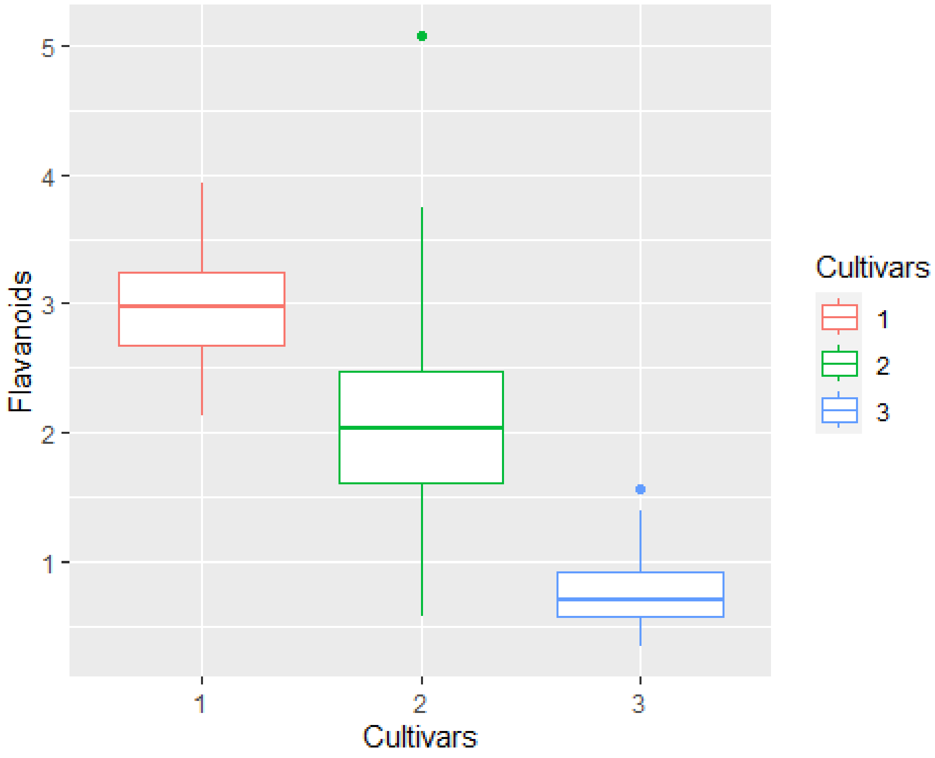

Malic Acid
6
5
4
3
2
1
1
2
3
Cultivars
Cultivars
1
2
3


Ash
3.0
2.5
2.0
1.5
1
2
3
Cultivars
Cultivars
1
2
3

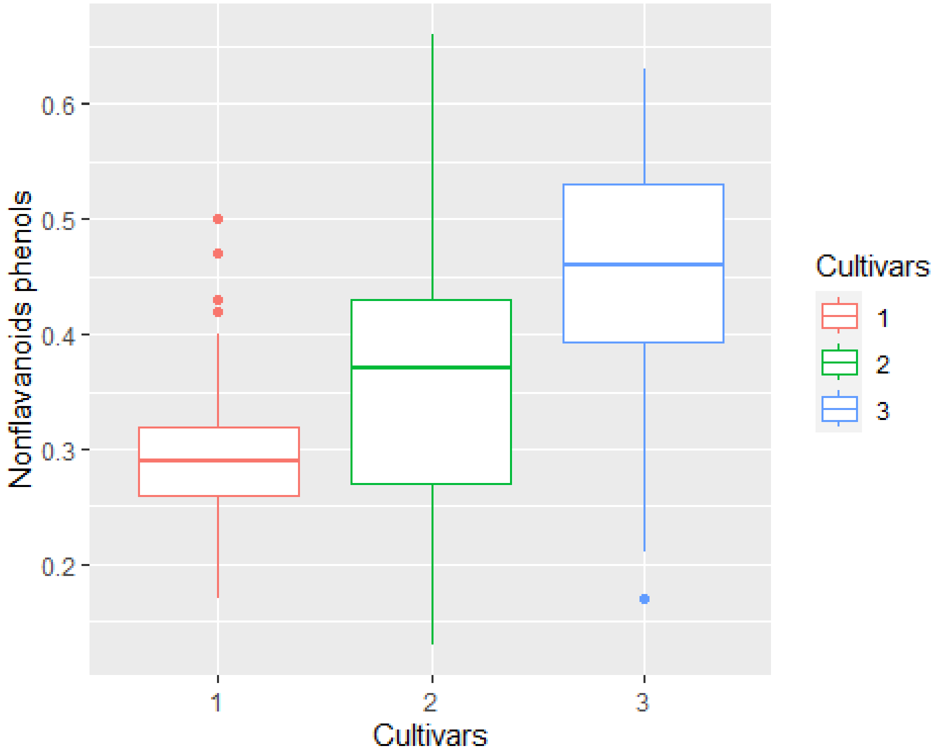

Alcalinity of ash
30
25
20
15
1
2
3
Cultivars
Cultivars
1
2
3


Magnesium
160
140
120
100
80
1
2
3
Cultivars
Cultivars
1
2
3

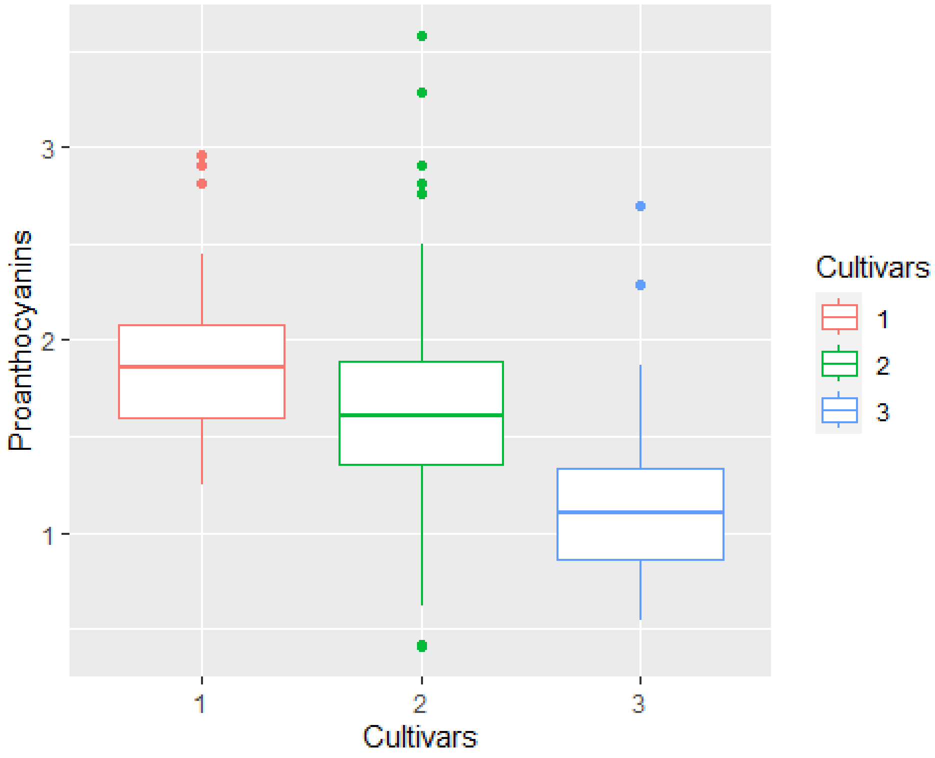

Total Phenols
4
3
2
1
1
2
3
Cultivars
Cultivars
1
2
3

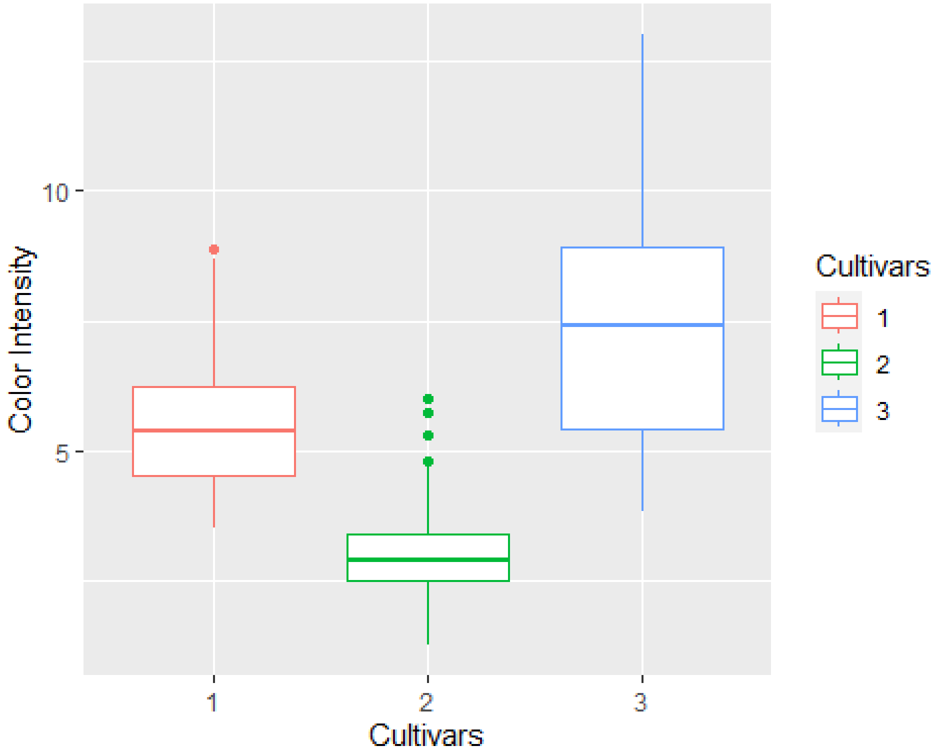


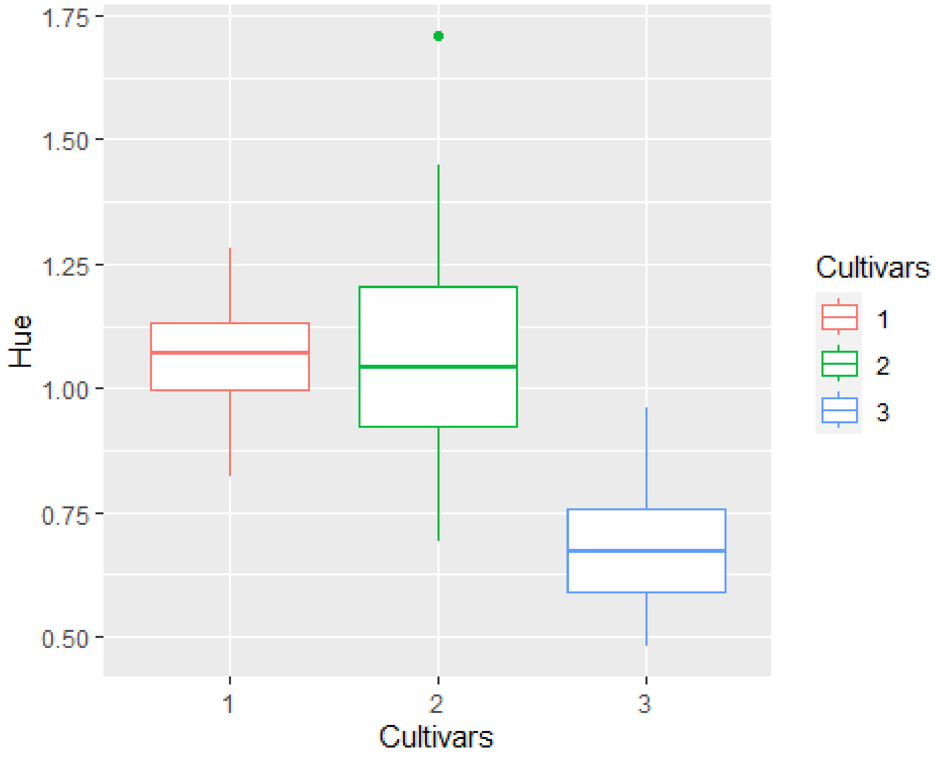
Flavanoids
5
4
3
2
1
Cultivars
1
2
3
Cultivars
1
2
3

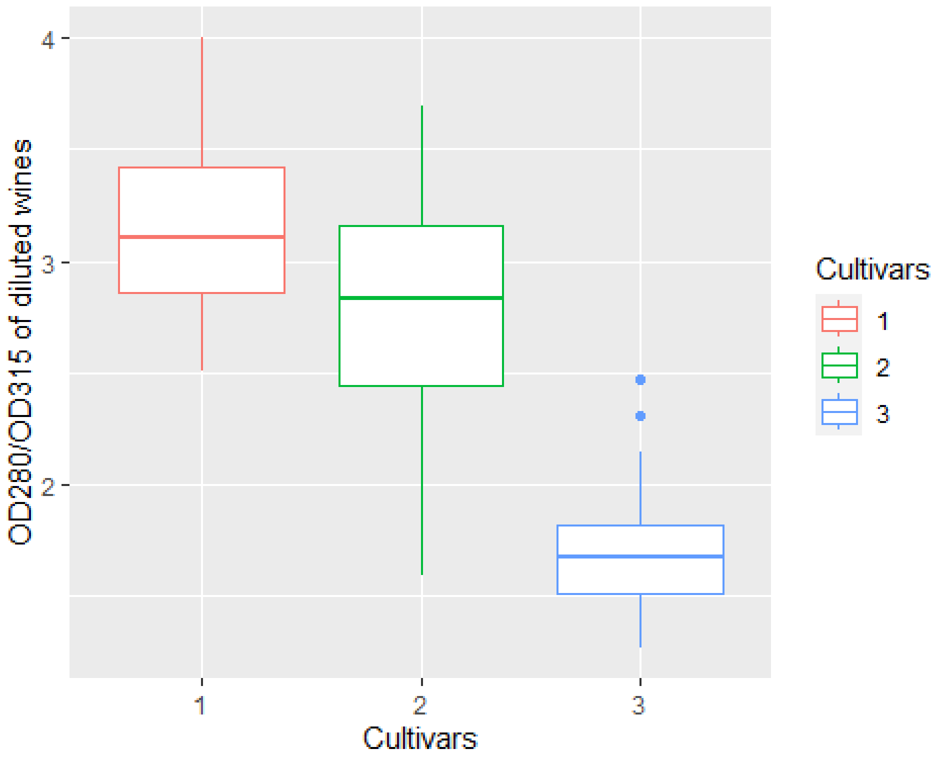


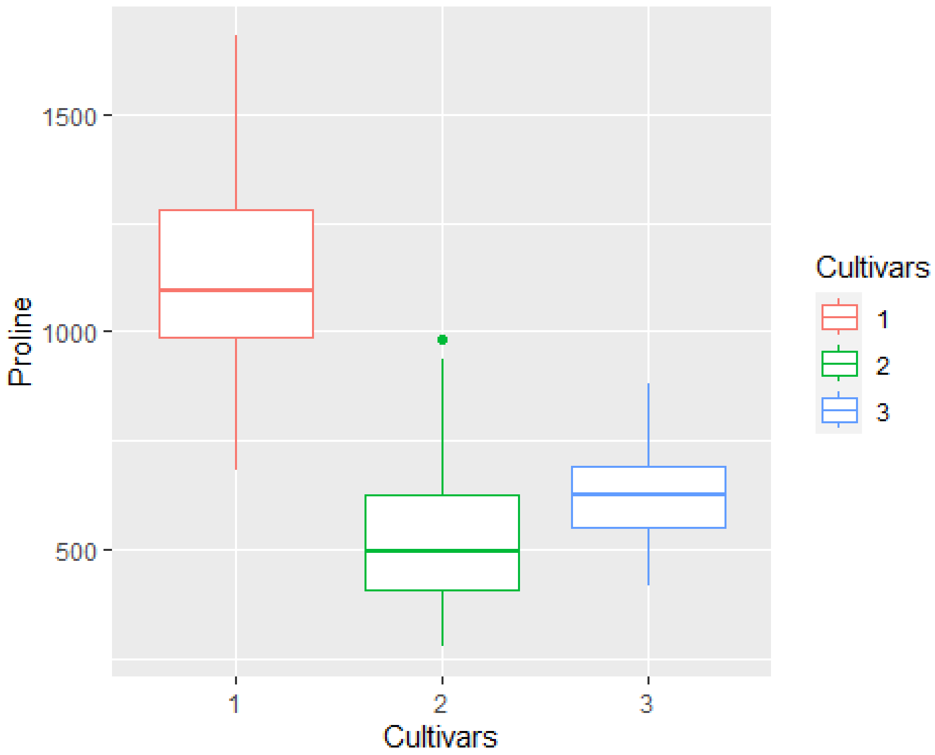
Nonflavanoids phenols
0.6
0.5
0.4
0.3
0.2
Cultivars
1
2
3
Cultivars
1
2
3


Proanthocyanins
3
2
1
1
2
3
Cultivars
Cultivars
1
2
3


Color Intensity
10
5
1
2
3
Cultivars
Cultivars
1
2
3


1.75
1.50
1.25
1.00
0.75
0.50
Hue
1
2
3
Cultivars
Cultivars
1
2
3


4
3
2
OD280/OD315 of diluted wines
1
2
3
Cultivars
Cultivars
1
2
3



- The box itself represents the middle 50% of the predictor values in the wine 2 dataset with a line in the middle indicating the median.
- The boxplots show that cultivar 1 wine have higher alcohol content while cultivar 2 wine has the lowest alcohol content.
- The upper and lower whiskers represent the minimum and maximum predictor values in the dataset. Any points outside of the whiskers are considered outliers. For example, examining the boxplots of the wine classes and alcohol, we can see that the minimum value of alcohol for class 1 cultivar is around 12.7 and the maximum value is about 14.8
- The height of the box and the position of the median gives us an idea of the central tendency of the predictor. For example, from the boxplots of cultivars and magnesium we can see that the outlier points in the boxplot of cultivar 2 and magnesium does not affect the central tendency (median).
- A short box with a median that is closer to the bottom of the box indicates that the data is skewed towards lower values.
- A wide box indicates that there is a lot of variation in the predictor while a narrow box indicates that there is less variation.

CONCLUSION

In this work, I pre-processed the data by cleaning it taking note of missing values and outliers. I replaced all my missing values with median observation and I used the matrix method to identify high leverage points. The final part of this work was concerned with my visualization of all variables and how they affect each other. Data pre-processing and visualization is a critical task in data analysis as it can help analyst make proper predictions.