Wine recognition data analysis

Sharmin Akhter

December 9, 2022

Contents

1. Introduction **2**

2. Exploring the dataset **3**
Columns . . . 3

3. Data Analysis **4**
Missing values . . . 4
Outlier detection and count . . . 5
Clean outliers . . . 6
Correlation Matrix . . . 7
Physical Interpretation . . . 8

4. Exploring the Relationships Between Variables **8**
Box Plot . . . 8
Physical Interpretation . . . 11
Violin Plot . . . 11
Histogram . . . 15
Density Plot . . . 18

5. Summary **21**

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#library(caret)
#library(data.table)
#library(zoo)
#library(leaps)
#library(imputeTS)
library(dplyr)
library(MASS)
```

```
##
```

```
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
#library(mosaic)
```

1. Introduction

The title of this database is Wine recognition Data Analysis. The purpose of this project to check the quality of wine from given attributes. The data is obtained from

```
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
```

From the above information we can see that each wine was grown in the same city in Italy by three different cultivers. Below are the columns of the dataset:

- **Class**: predictor
- **Alcohol**: Numeric
- **Malic Acid**: Numeric
- **Ash**: Numeric
- **Alcalinity** of Ash: Numeric
- **Magnesium**: Integer
- **Total Phenols**: Numeric
- **Flavanoids**: Numeric
- **Nonflavanoids Phenols**: Numeric
- **Proanthocyanins**: Numeric
- **Color Intensity**: Numeric
- **Hue**: Numeric
- **OD280/OD315 of diluted wines**: Numeric
- **Proline**: Numeric

All the cultivers has 3 classes as follows: class 1 59 class 2 71 class 3 48

and number of attributes 13, all are continuous.

2. Exploring the dataset

```
winedata<- read.csv("wine.txt", header = F)
head(winedata)
```

```
##   V1    V2   V3   V4   V5  V6   V7   V8   V9  V10  V11  V12  V13  V14
## 1  1 14.23 1.71   NA 15.6 127 2.80 3.06 0.28 2.29 5.64 1.04 3.92 1065
## 2  1 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.05 3.40 1050
## 3  1 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.03 3.17 1185
## 4  1 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.86 3.45 1480
## 5  1 13.24 2.59 2.87   21 118 2.80 2.69 0.39 1.82 4.32 1.04 2.93  735
## 6  1 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.05 2.85 1450
```

Columns

```
##  [1] "Class"                         "Alcohol"
##  [3] "Malic_acid"                    "Ash"
##  [5] "Alcalinity_of_ash"             "Magnesium"
##  [7] "Total_phenols"                 "Flavanoids"
##  [9] "Nonflavanoid_phenols"          "Proanthocyanins"
## [11] "Color_intensity"               "Hue"
## [13] "OD280_OD315_of_diluted_wines" "Proline"
```

#There are total 14 variables, specified to 2 types of data: number and integer. The 1st variable should be factor. Now I will change integer to factor. Then look at the structur eof the data.

```
winedata$Class <- as.factor(winedata$Class)
str(winedata)
```

```
## 'data.frame':    178 obs. of  14 variables:
##  $ Class                       : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Alcohol                     : num  14.2 13.2 13.2 14.4 13.2 ...
##  $ Malic_acid                  : chr  "1.71" "1.78" "2.36" "1.95" ...
##  $ Ash                         : num  NA 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
##  $ Alcalinity_of_ash           : chr  "15.6" "11.2" "18.6" "16.8" ...
##  $ Magnesium                   : chr  "127" "100" "101" "113" ...
##  $ Total_phenols               : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
##  $ Flavanoids                  : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
##  $ Nonflavanoid_phenols        : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
##  $ Proanthocyanins             : chr  "2.29" "1.28" "2.81" "2.18" ...
##  $ Color_intensity             : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
##  $ Hue                         : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
##  $ OD280_OD315_of_diluted_wines: num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
##  $ Proline                     : int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
```

#Here I am trying to get an idea of 3 classes. Below are the mean of class 1(59), class 2(71), class 3(48)

```
## [1] 0.3314607
```

```
## [1] 0.3988764
```

```
## [1] 0.2696629
```

#From the structure we can see that the Malic acid, Alcalinity of ash, Magnesium, Proanthocyanins are non numeric. Now change to non numeric value to numeric

```
## 'data.frame':    178 obs. of  14 variables:
##  $ Class                       : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ Alcohol                     : num  14.2 13.2 13.2 14.4 13.2 ...
##  $ Malic_acid                  : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
##  $ Ash                         : num  NA 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
##  $ Alcalinity_of_ash           : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
##  $ Magnesium                   : num  127 100 101 113 118 112 96 121 97 98 ...
##  $ Total_phenols               : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
##  $ Flavanoids                  : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
##  $ Nonflavanoid_phenols        : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
##  $ Proanthocyanins             : num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
##  $ Color_intensity             : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
##  $ Hue                         : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
##  $ OD280_OD315_of_diluted_wines: num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
##  $ Proline                     : num  1065 1050 1185 1480 735 ...
```

After changed the non numeric values to numeric its introduced some NAs. In my next section I will analyze the dataset

3. Data Analysis

Missing values

```
## [1] 18
```

#From the summary of wine data we can see that there are some missing values on the data which are not symmetric. So we will replace missing values with the median.

```
newwinedata<- winedata %>% mutate(across(where(is.numeric), ~replace_na(., median(., na.rm=TRUE))))
str(newwinedata)
```

```
## 'data.frame':    178 obs. of  14 variables:
##  $ Class                     : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Alcohol                   : num  14.2 13.2 13.2 14.4 13.2 ...
##  $ Malic_acid                : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
##  $ Ash                       : num  2.36 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
##  $ Alcalinity_of_ash         : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
##  $ Magnesium                 : num  127 100 101 113 118 112 96 121 97 98 ...
##  $ Total_phenols             : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
##  $ Flavanoids                : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
##  $ Nonflavanoid_phenols      : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
##  $ Proanthocyanins           : num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
##  $ Color_intensity           : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
##  $ Hue                       : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
##  $ OD280_OD315_of_diluted_wines: num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
##  $ Proline                   : num  1065 1050 1185 1480 735 ...
```

```
sum(is.na(newwinedata))
```

```
## [1] 0
```

#Summary of new data without NAs

```
##  Class     Alcohol        Malic_acid         Ash        Alcalinity_of_ash
##  1:59   Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :11.20
##  2:71   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210   1st Qu.:17.50
##  3:48   Median :13.05   Median :1.870   Median :2.360   Median :19.50
##         Mean   :13.00   Mean   :2.337   Mean   :2.365   Mean   :19.59
##         3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.547   3rd Qu.:21.50
```

```
##         Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00
##    Magnesium       Total_phenols     Flavanoids    Nonflavanoid_phenols
##  Min.   :   70.0   Min.   :0.980   Min.   :0.340   Min.   :0.1300
##  1st Qu.:   88.0   1st Qu.:1.742   1st Qu.:1.205   1st Qu.:0.2700
##  Median :   98.0   Median :2.350   Median :2.130   Median :0.3400
##  Mean   :  660.7   Mean   :2.294   Mean   :2.023   Mean   :0.3625
##  3rd Qu.:  107.0   3rd Qu.:2.800   3rd Qu.:2.842   3rd Qu.:0.4375
##  Max.   :99999.0   Max.   :3.880   Max.   :5.080   Max.   :0.6600
##  Proanthocyanins Color_intensity        Hue
##  Min.   :0.410   Min.   :      1   Min.   :0.4800
##  1st Qu.:1.250   1st Qu.:      3   1st Qu.:0.7825
##  Median :1.550   Median :      5   Median :0.9650
##  Mean   :1.586   Mean   :  55623   Mean   :0.9574
##  3rd Qu.:1.950   3rd Qu.:      6   3rd Qu.:1.1200
##  Max.   :3.580   Max.   :9899999   Max.   :1.7100
##  OD280_OD315_of_diluted_wines    Proline
##  Min.   :1.270                Min.   : 278.0
##  1st Qu.:1.938                1st Qu.: 500.5
##  Median :2.780                Median : 673.5
##  Mean   :2.609                Mean   : 746.9
##  3rd Qu.:3.170                3rd Qu.: 985.0
##  Max.   :4.000                Max.   :1680.0
```

#Lets check the dimension of the data

```
## [1] 178  14
```

###The wine dataset has 178 observations, 13 predictors and 1 outcome (Class). All of the predictors are numeric values, outcomes are integer.

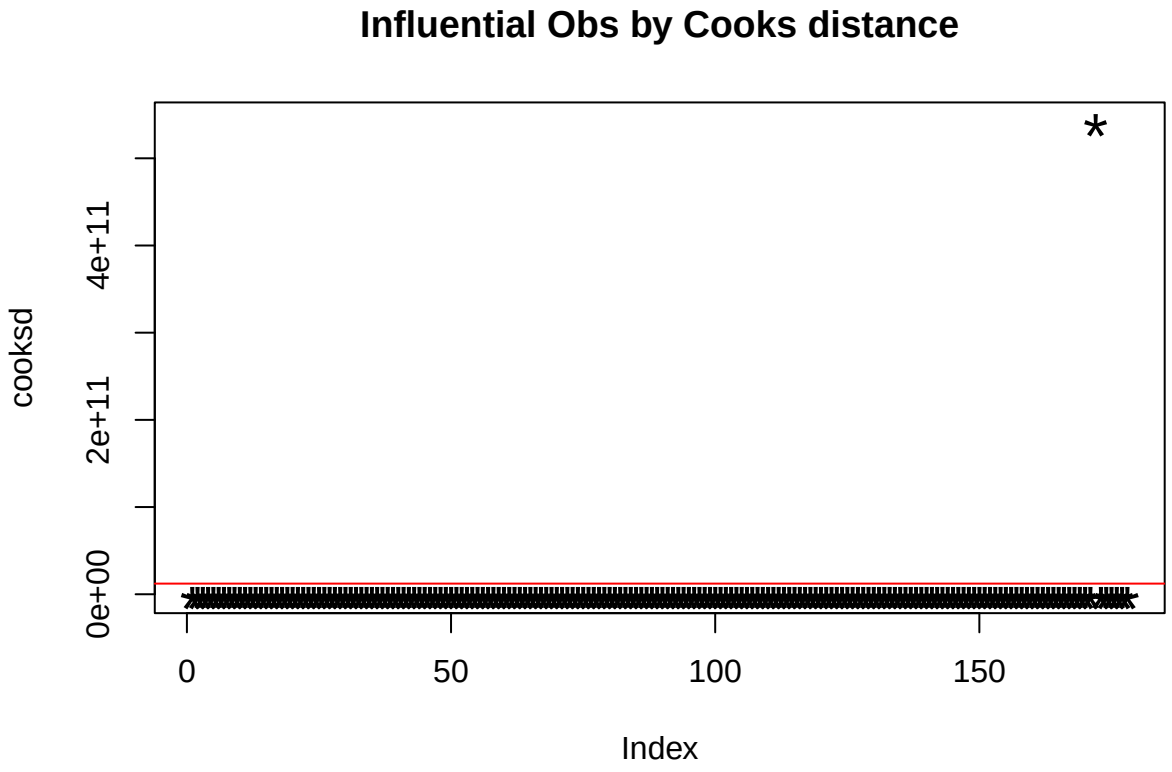
The summary shows that some of the variables has wide range compared to the IQR, which may indicate spread in the data and the presence of outliers. We investigate further by producing boxplots for each of the variables:

Outlier detection and count

```
##  [1] 124 138 174  26 122  60  67 101  74 128   2  14  70  79  96 177 111 152 159
## [20] 160 172 116
```

###Use cooks distance to detect influential observations

```
mod<- lm(as.integer(Class) ~., data = newwinedata)
cooksd<- cooks.distance(mod)
plot(cooksd, pch = "*", cex = 2, main = "Influential Obs by Cooks distance")
abline(h = 4*mean(cooksd, na.rm = T), col = "red")
```



Clean outliers

```
clean_outliers = as.numeric(rownames(newwinedata[cooksd > 4 * mean(cooksd, na.rm=T),]))
outliers = c(outliers, clean_outliers[!clean_outliers %in% outliers ])

clean_Data = newwinedata[-outliers,]
summary(clean_Data)
```

```
##  Class     Alcohol        Malic_acid         Ash        Alcalinity_of_ash
##  1:56   Min.   :11.41   Min.   :0.740   Min.   :1.710   Min.   :12.00
##  2:59   1st Qu.:12.37   1st Qu.:1.607   1st Qu.:2.237   1st Qu.:17.48
##  3:41   Median :13.06   Median :1.870   Median :2.360   Median :19.50
##         Mean   :13.04   Mean   :2.331   Mean   :2.373   Mean   :19.44
##         3rd Qu.:13.71   3rd Qu.:3.132   3rd Qu.:2.540   3rd Qu.:21.12
##         Max.   :14.83   Max.   :5.190   Max.   :2.920   Max.   :27.00
##    Magnesium      Total_phenols     Flavanoids    Nonflavanoid_phenols
##  Min.   : 70.00   Min.   :0.980   Min.   :0.340   Min.   :0.1300
##  1st Qu.: 88.00   1st Qu.:1.715   1st Qu.:1.215   1st Qu.:0.2700
##  Median : 98.00   Median :2.335   Median :2.120   Median :0.3400
##  Mean   : 98.55   Mean   :2.284   Mean   :2.024   Mean   :0.3589
##  3rd Qu.:106.00   3rd Qu.:2.800   3rd Qu.:2.885   3rd Qu.:0.4300
##  Max.   :134.00   Max.   :3.880   Max.   :3.930   Max.   :0.6600
##  Proanthocyanins Color_intensity       Hue          OD280_OD315_of_diluted_wines
##  Min.   :0.410   Min.   : 1.280   Min.   :0.5400   Min.   :1.270
##  1st Qu.:1.235   1st Qu.: 3.250   1st Qu.:0.7975   1st Qu.:2.007
##  Median :1.535   Median : 4.750   Median :0.9600   Median :2.780
##  Mean   :1.538   Mean   : 5.002   Mean   :0.9577   Mean   :2.621
##  3rd Qu.:1.870   3rd Qu.: 6.200   3rd Qu.:1.1125   3rd Qu.:3.170
##  Max.   :2.960   Max.   :10.680   Max.   :1.4500   Max.   :4.000
```

```
##     Proline
##  Min.   : 278.0
##  1st Qu.: 507.5
##  Median : 675.0
##  Mean   : 757.6
##  3rd Qu.:1023.8
##  Max.   :1680.0
```

```
str(clean_Data)
```

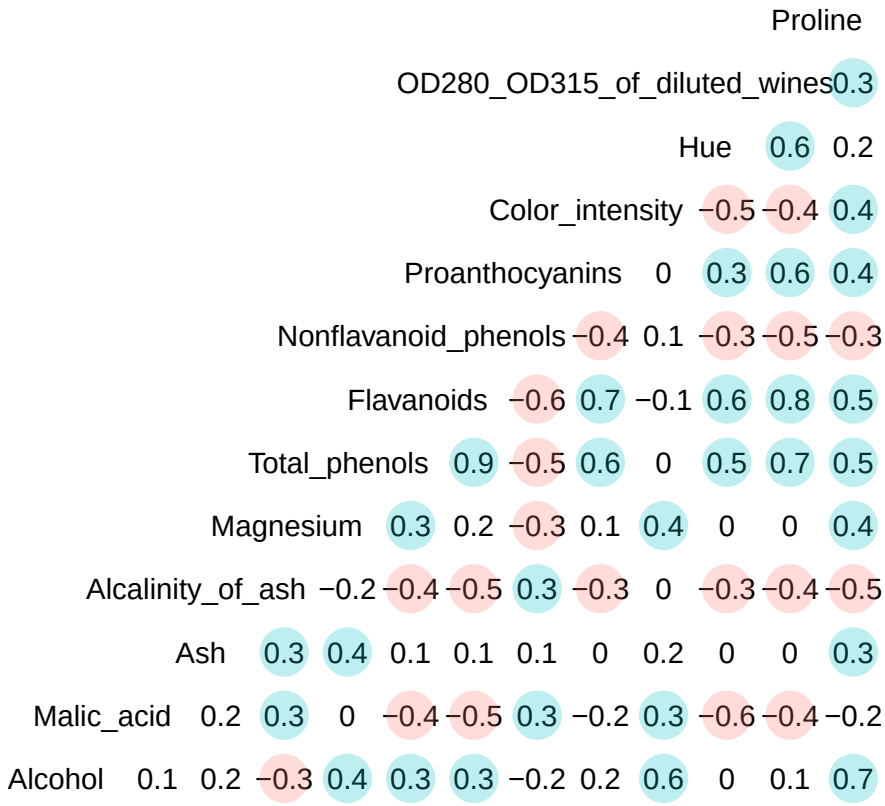
```
## 'data.frame':    156 obs. of  14 variables:
##  $ Class                        : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Alcohol                      : num  14.2 13.2 14.4 13.2 14.2 ...
##  $ Malic_acid                   : num  1.71 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 2.16 ...
##  $ Ash                          : num  2.36 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 2.3 ...
##  $ Alcalinity_of_ash            : num  15.6 18.6 16.8 21 15.2 14.6 17.6 14 16 18 ...
##  $ Magnesium                    : num  127 101 113 118 112 96 121 97 98 105 ...
##  $ Total_phenols                : num  2.8 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 2.95 ...
##  $ Flavanoids                   : num  3.06 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 3.32 ...
##  $ Nonflavanoid_phenols         : num  0.28 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 0.22 ...
##  $ Proanthocyanins              : num  2.29 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 2.38 ...
##  $ Color_intensity              : num  5.64 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 5.75 ...
##  $ Hue                          : num  1.04 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 1.25 ...
##  $ OD280_OD315_of_diluted_wines: num  3.92 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 3.17 ...
##  $ Proline                      : num  1065 1185 1480 735 1450 ...
```

Correlation Matrix

```
options(repr.plot.width=6, repr.plot.height=4)

ggcorr(clean_Data[,2:14], geom = "blank", label = TRUE,
       hjust = 0.9, layout.exp = 2) +
  geom_point(size = 8, aes(color = coefficient > 0,
                           alpha = abs(coefficient) > 0.25)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = FALSE, alpha = FALSE)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
```



Physical Interpretation

In this correlation matrix we used 'ggcor' function from 'ggally' to show the correlation coefficients for each of the variables in the data sets. The diagonal elements of the matrix are labeled with the names of the variables. Here we choose absolute value of coefficients greater than 25% to show the transparency of the points on the matrix. The point with value greater than 25% has lower transparency and less or equal 25% has higher transparency. Also from this matrix we can tells that the points with red color are negatively correlated with each pair of the variables and point with blue color are negatively correlated. Overall we can say correlation matrix provides a useful overview of the relationships between the different variables in the data sets, and can help us gain insights into the data.

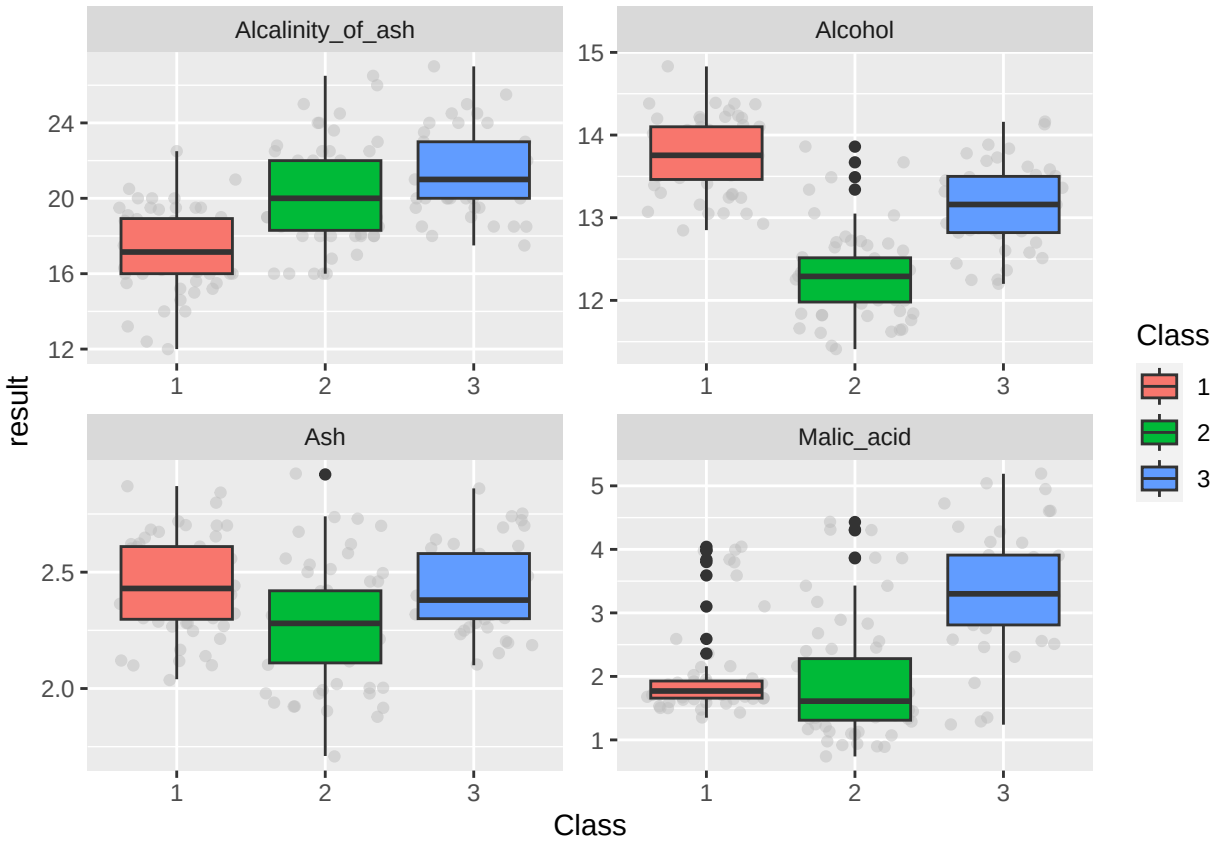
4. Exploring the Relationships Between Variables

In order to understand our data sets we are interested in visualizing the each variables by Class. In this section we will show our work by creating Box plot, Violin plot, Histogram and at the last Density distribution plot to analyze the distribution of each variable using the ggplot package.

Box Plot

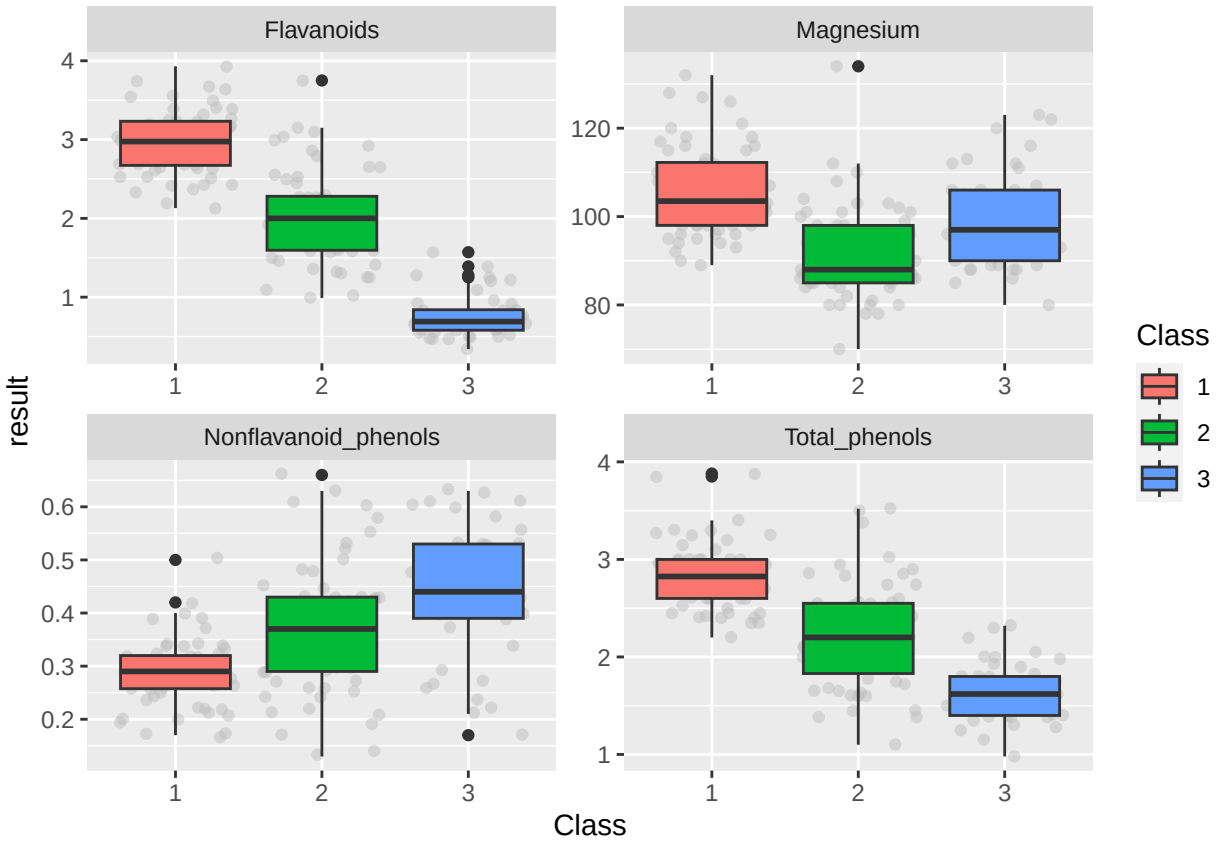
Distribution of Alcohol, Malic_acid, Ash, Alcalinity_of_ash in the dataset using Boxplot.

```
clean_Data%>% gather(2:5, key = "variables", value = "result") %>%
  ggplot(aes(Class, result, fill = Class)) +
  geom_jitter(color = "grey", alpha = 0.5)+
  geom_boxplot()+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```



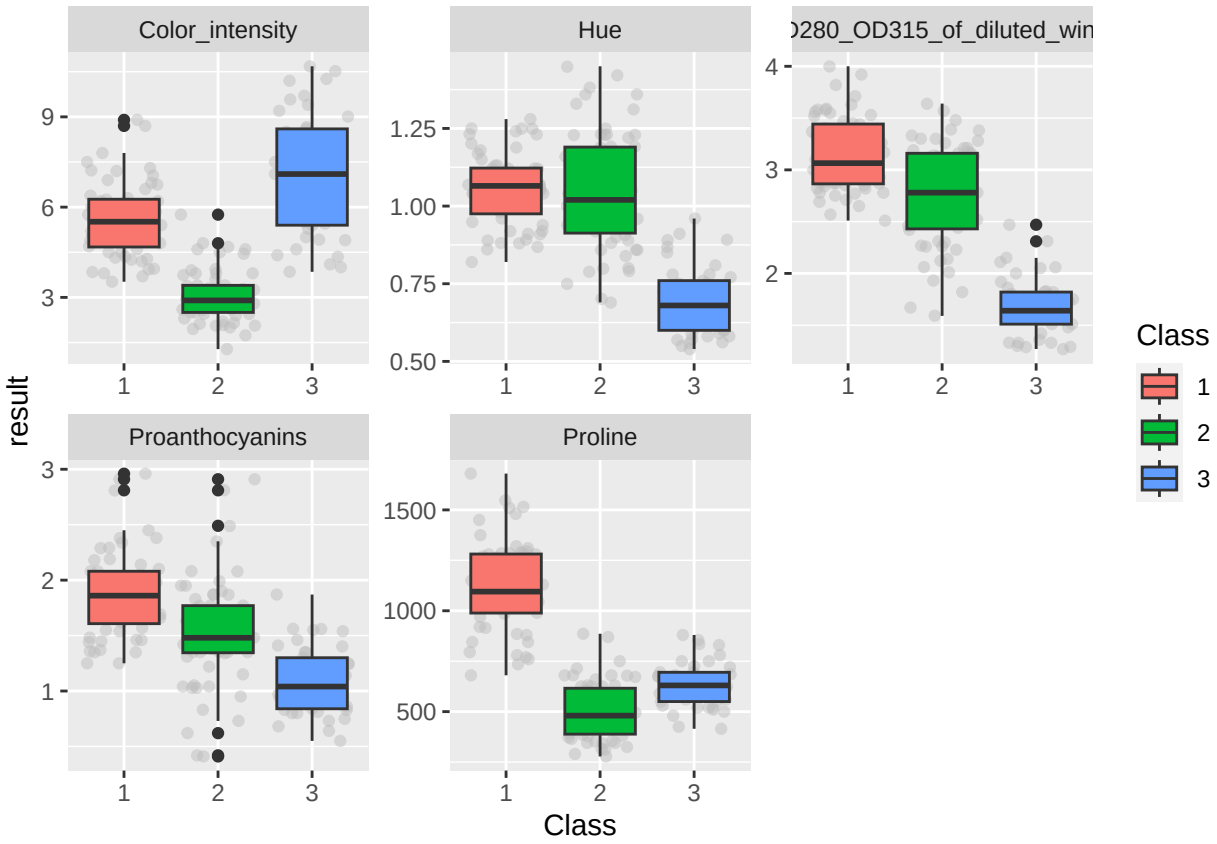
Distribution of Magnesium, Total_phenols, Nonflavanoid_phenols, Flavanoids in the dataset using Boxplot.

```
clean_Data%>% gather(6:9, key = "variables", value = "result") %>%
  ggplot(aes(Class, result, fill = Class)) +
  geom_jitter(color = "grey", alpha = 0.5)+
  geom_boxplot()+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```



Distribution of Proanthocyanins, Color_intensity, Hue, OD280_OD315_of_diluted_wines, Proline in the dataset using Boxplot.

```
clean_Data%>% gather(10:14, key = "variables", value = "result") %>%
  ggplot(aes(Class, result, fill = Class)) +
  geom_jitter(color = "grey", alpha = 0.5)+
  geom_boxplot()+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```



Physical Interpretation

- The Box plot visiualized the distribution of each of the variables by their class.
- In these plot we use facet with free scaling to see the range, median and quartiles for each variable. It also allows us to see the spread and central tendency of the data.
- By using jittered scatter plot we want to see the overall shape of the data distribution, while box plot highlighting any outliers or unusual values.
- By comparing the distributions of the different variables, we can see how they relate to each other and identify any potential patterns or trends in the data.

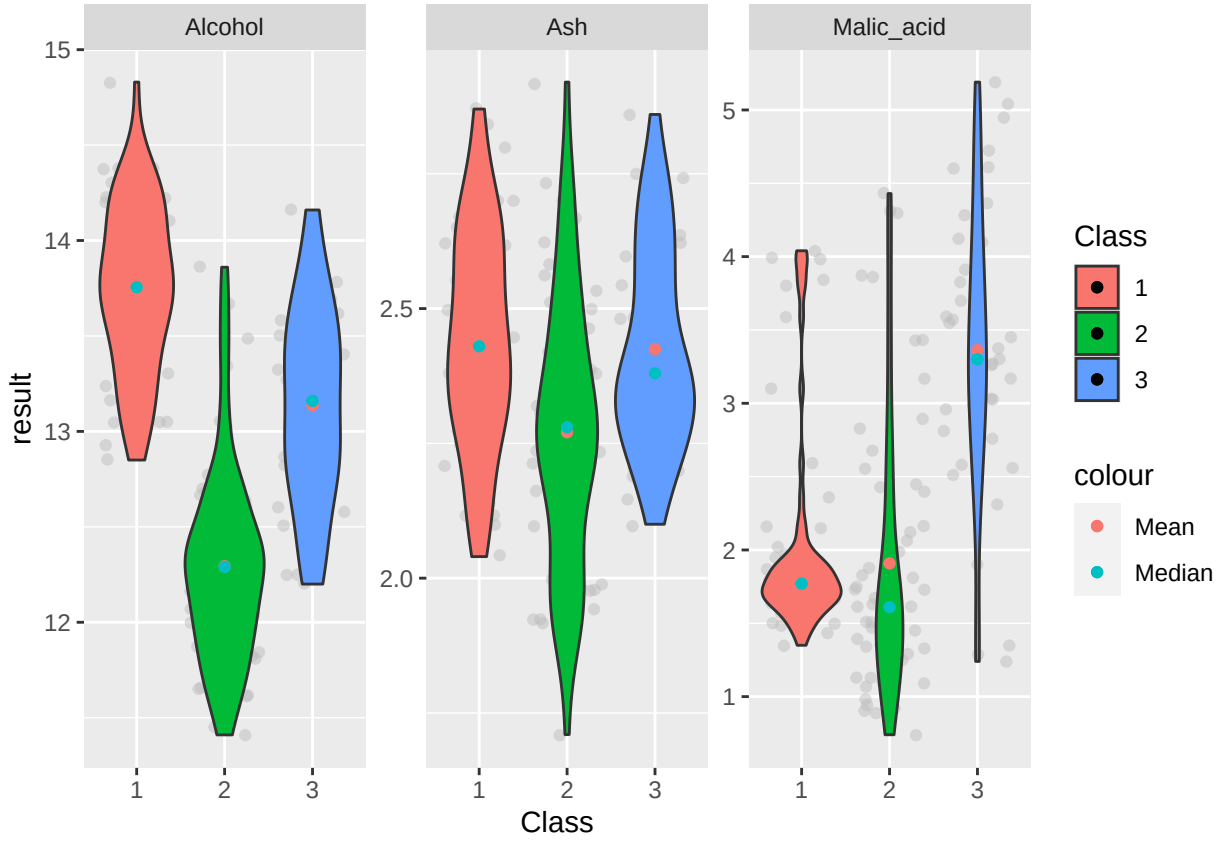
From overall figure we can see the except Alcalinity_of_ash, hue, Proline all other variables has outliers.

Violin Plot

Distribution of Alcohol, Ash, Malic_acid in the dataset using Boxplot.

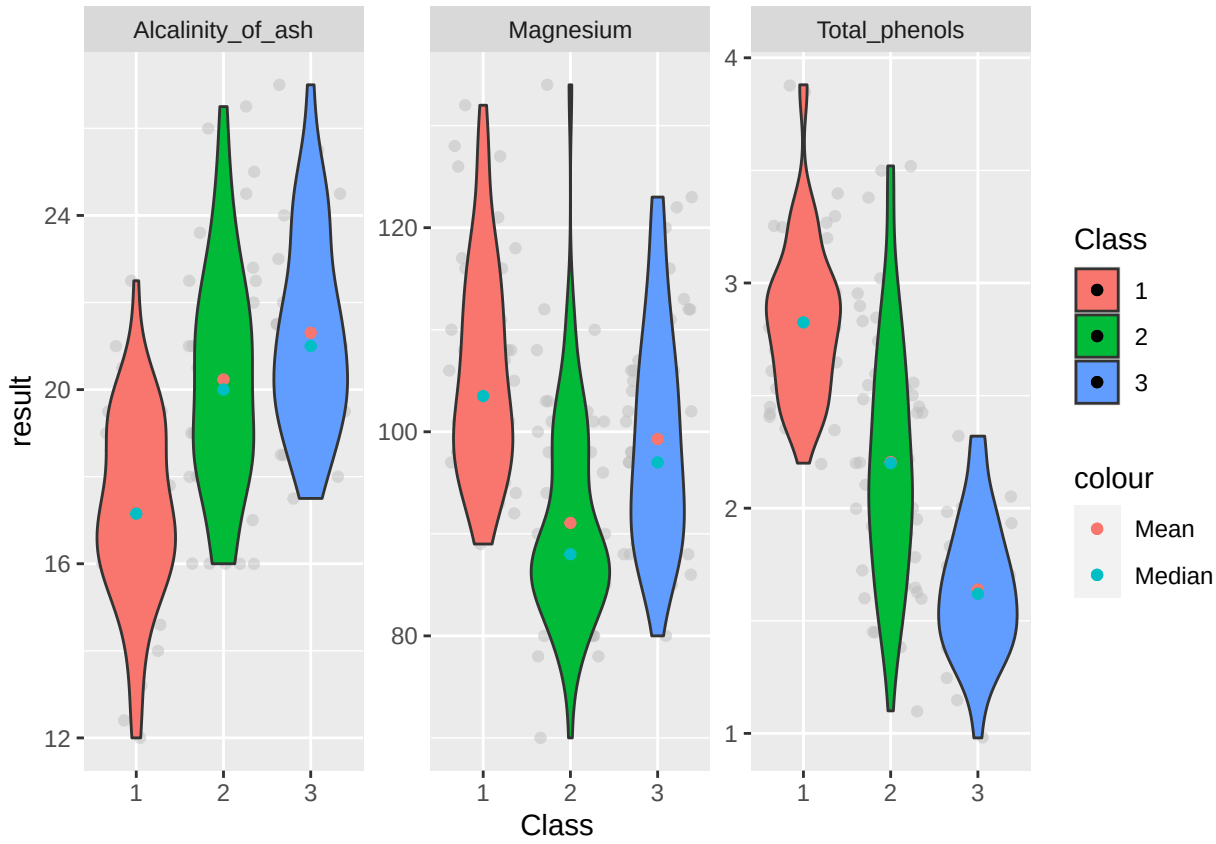
```
clean_Data%>% gather(2:4, key = "variables", value = "result") %>%
  ggplot(aes(Class, result, fill = Class)) +
  geom_jitter(color = "grey", alpha = 0.5)+
  geom_violin()+
  stat_summary(fun = "mean",
               geom = "point",
               aes(color = "Mean")) +
  stat_summary(fun = "median",
               geom = "point",
```

```
              aes(color = "Median"))+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```

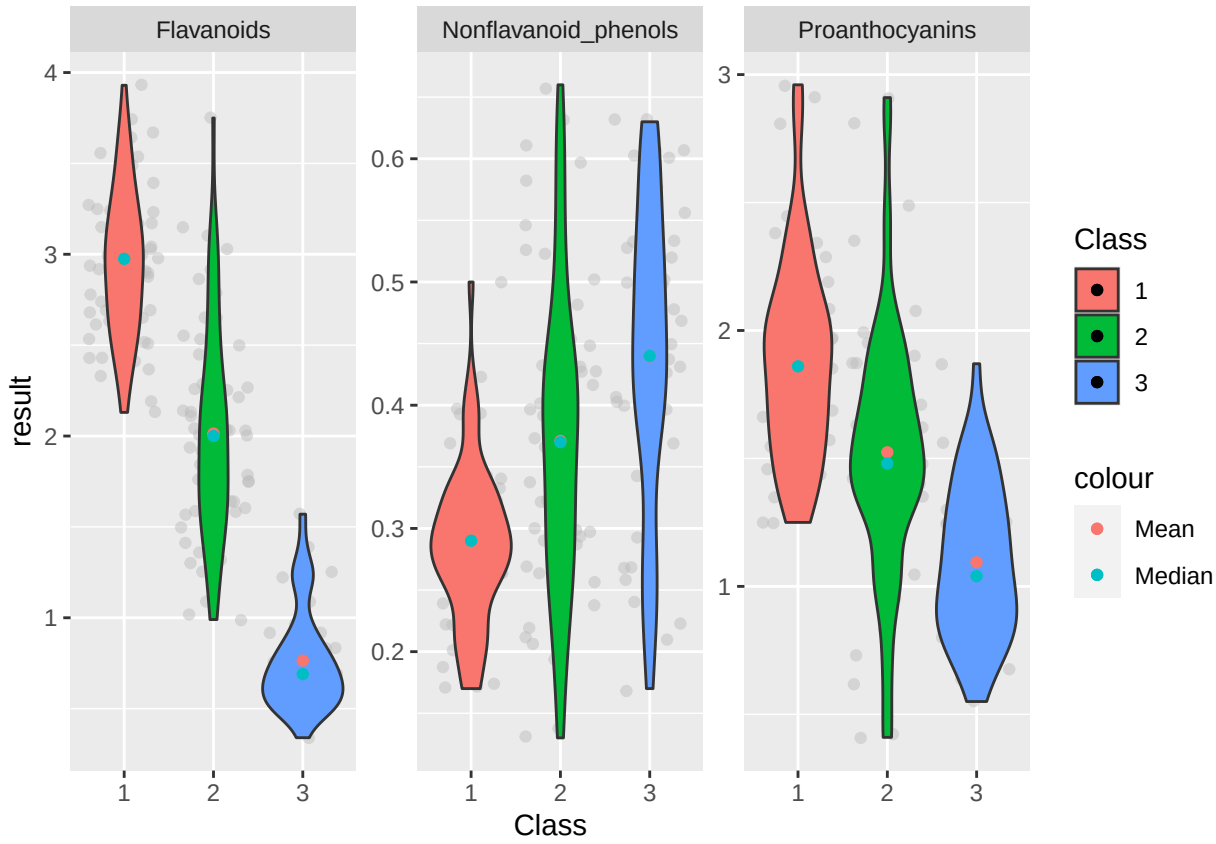


Distribution of Magnesium, Total_phenols, Nonflavanoid_phenols, Flavanoids in the dataset using Boxplot.

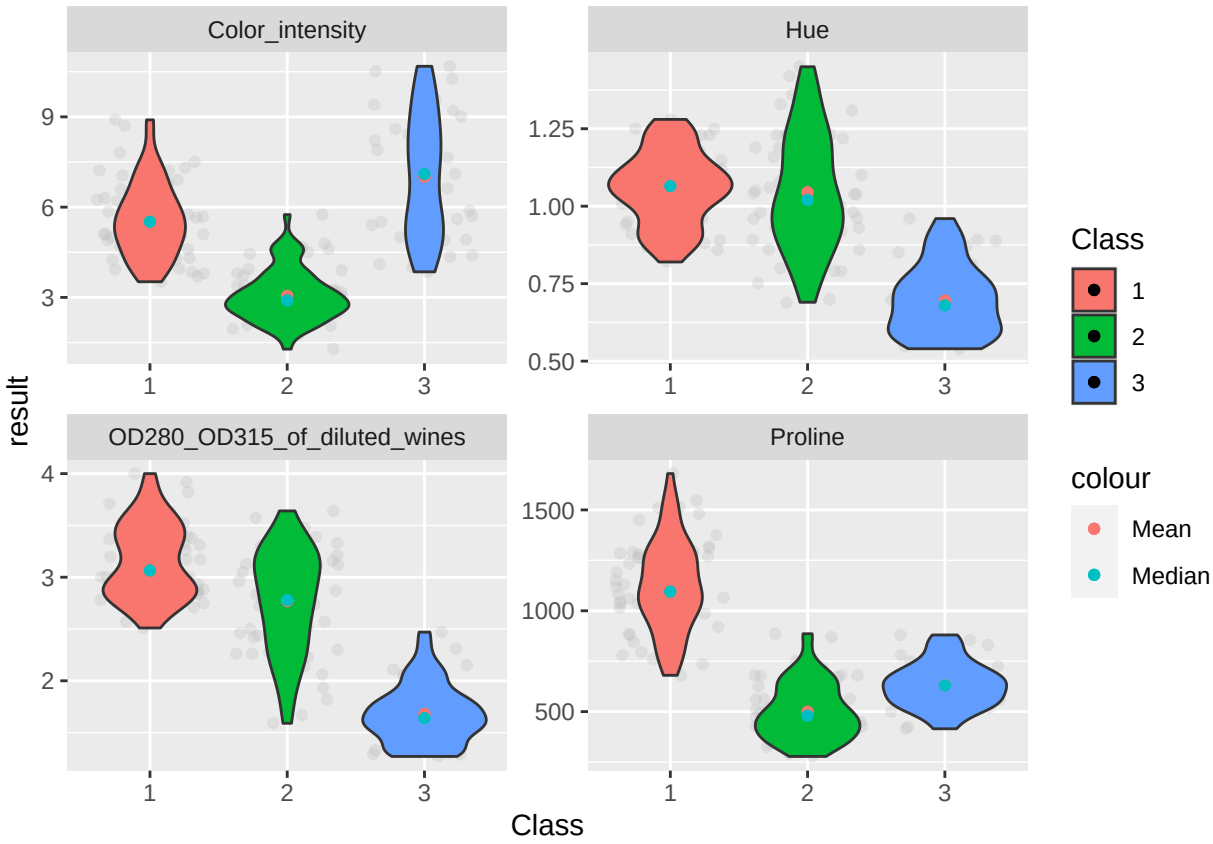
```
clean_Data%>% gather(5:7, key = "variables", value = "result") %>%
  ggplot(aes(Class, result, fill = Class)) +
  geom_jitter(color = "grey", alpha = 0.5)+
  geom_violin()+
  stat_summary(fun = "mean",
               geom = "point",
               aes(color = "Mean")) +
  stat_summary(fun = "median",
               geom = "point",
               aes(color = "Median"))+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```




```
clean_Data%>% gather(8:10, key = "variables", value = "result") %>%
  ggplot(aes(Class, result, fill = Class)) +
  geom_jitter(color = "grey", alpha = 0.5)+
  geom_violin()+
  stat_summary(fun = "mean",
               geom = "point",
               aes(color = "Mean")) +
  stat_summary(fun = "median",
               geom = "point",
               aes(color = "Median"))+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```




```
clean_Data%>% gather(11:14, key = "variables", value = "result") %>%
  ggplot(aes(Class, result, fill = Class)) +
  geom_jitter(color = "grey", alpha = 0.3)+
  geom_violin()+
  stat_summary(fun = "mean",
               geom = "point",
               aes(color = "Mean")) +
  stat_summary(fun = "median",
               geom = "point",
               aes(color = "Median"))+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```



Physical Interpretation

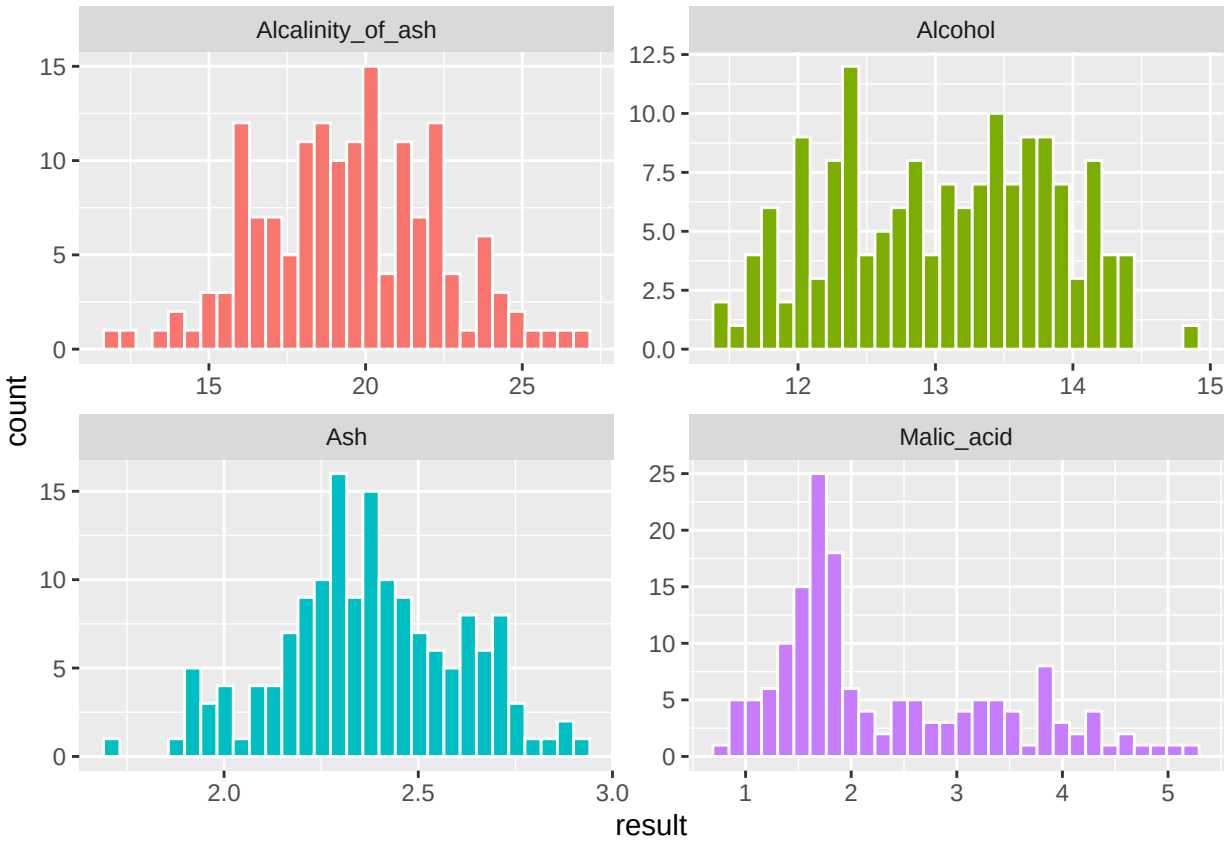
This figure shows the distribution of multiple variables, separated by class, using a combination of a jitter plot, violin plot, and summary statistics. The jitter plot allows us to visualize the density of the data, while the violin plot shows the distribution of the data. The summary statistics, such as the mean and median, provide additional information about the distribution of each variable.

Histogram

distribution of values for each of the columns

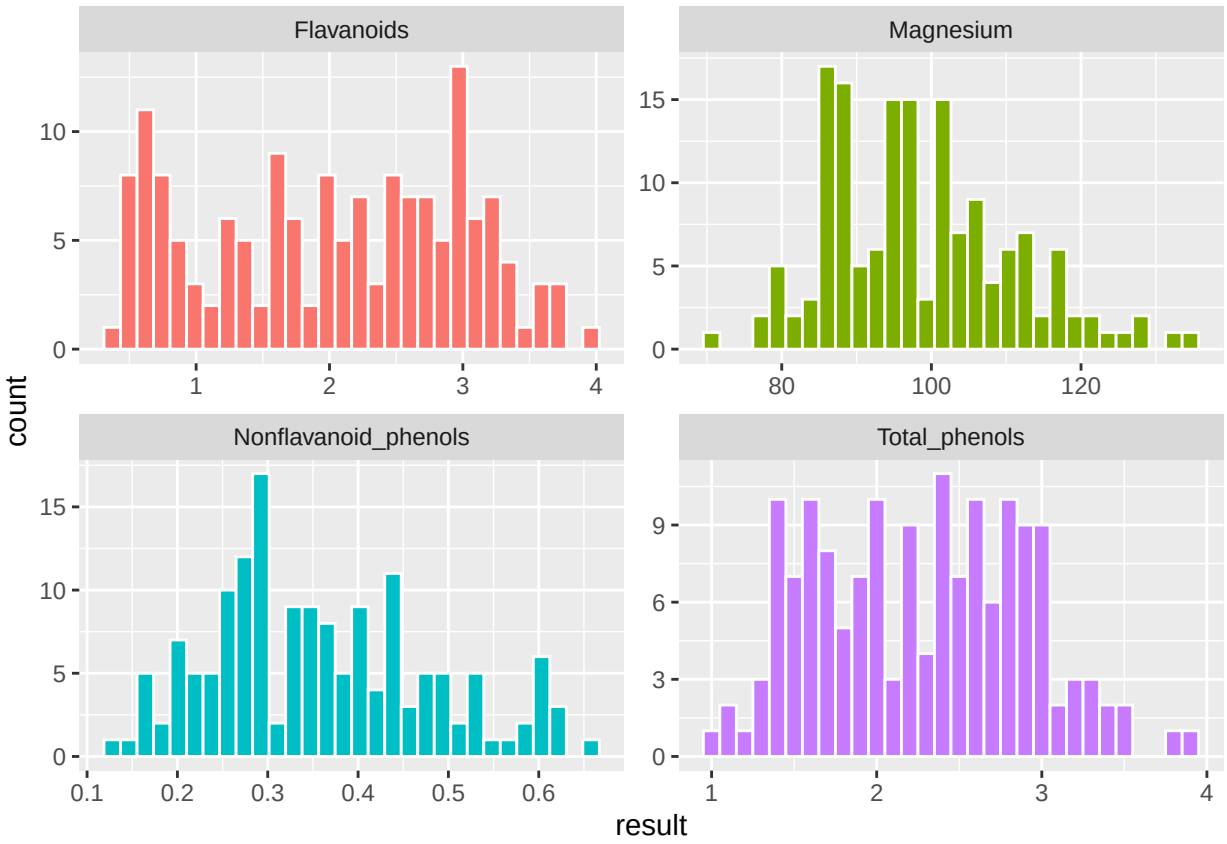
```
clean_Data %>% gather(2:5, key = "variables", value = "result") %>%
  ggplot(aes(result)) +
  geom_histogram(aes(fill = variables), color = "white")+
  theme_get()+
  facet_wrap(.~variables, scale = "free") +
  theme(legend.position = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



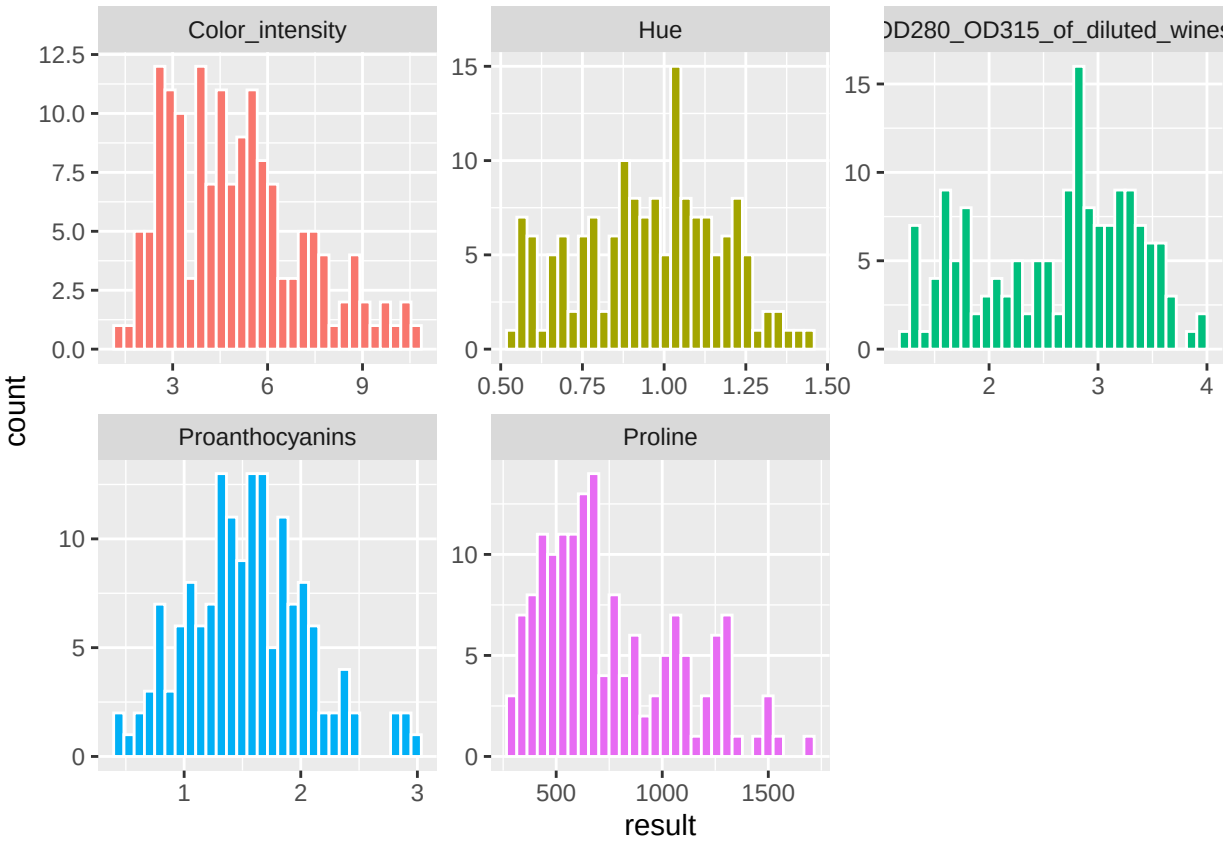

```
clean_Data %>% gather(6:9, key = "variables", value = "result") %>%
  ggplot(aes(result)) +
  geom_histogram(aes(fill = variables), color = "white")+
  theme_get()+
  facet_wrap(.~variables, scale = "free") +
  theme(legend.position = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
clean_Data %>% gather(10:14, key = "variables", value = "result") %>%
  ggplot(aes(result)) +
  geom_histogram(aes(fill = variables), color = "white")+
  theme_get()+
  facet_wrap(.~variables, scale = "free") +
  theme(legend.position = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

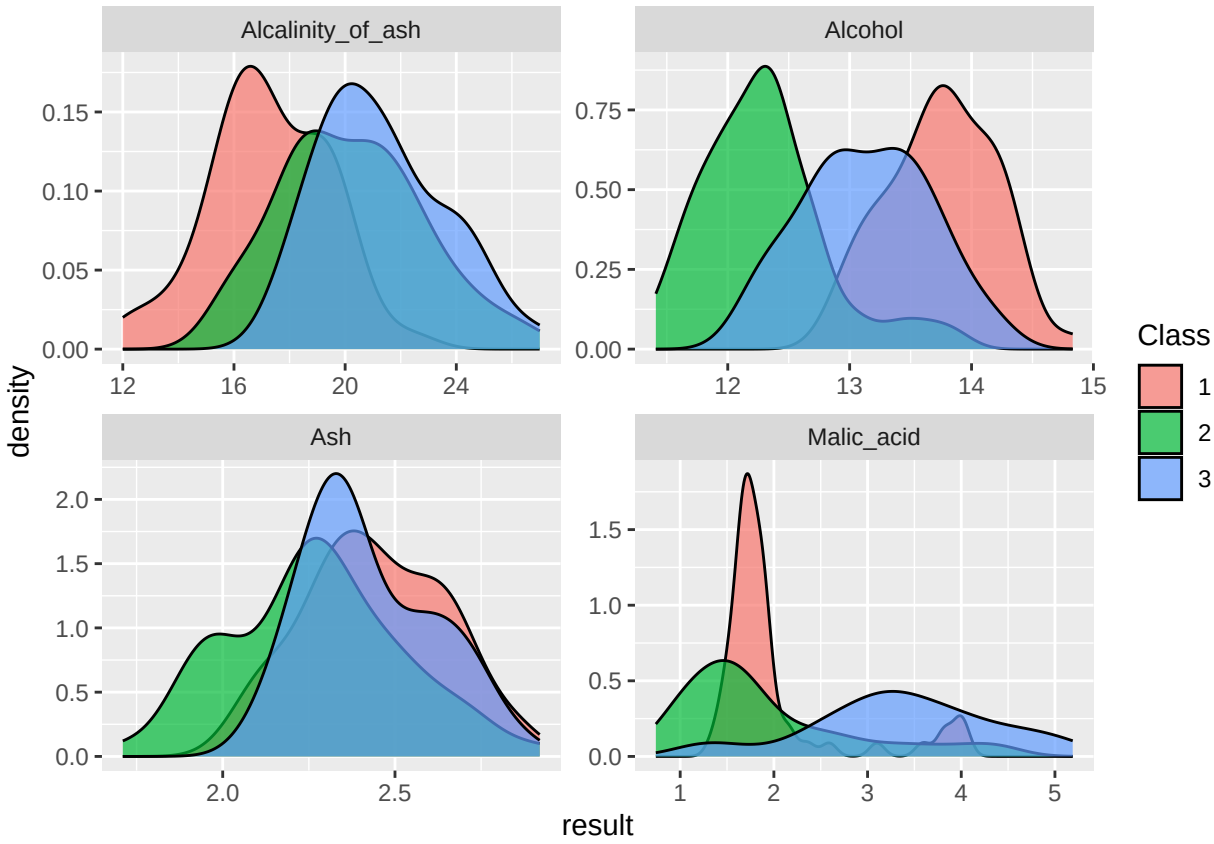


Next, I want to visualize the variables by class. To do this, I will make distributions of the variables and overlap them by class.

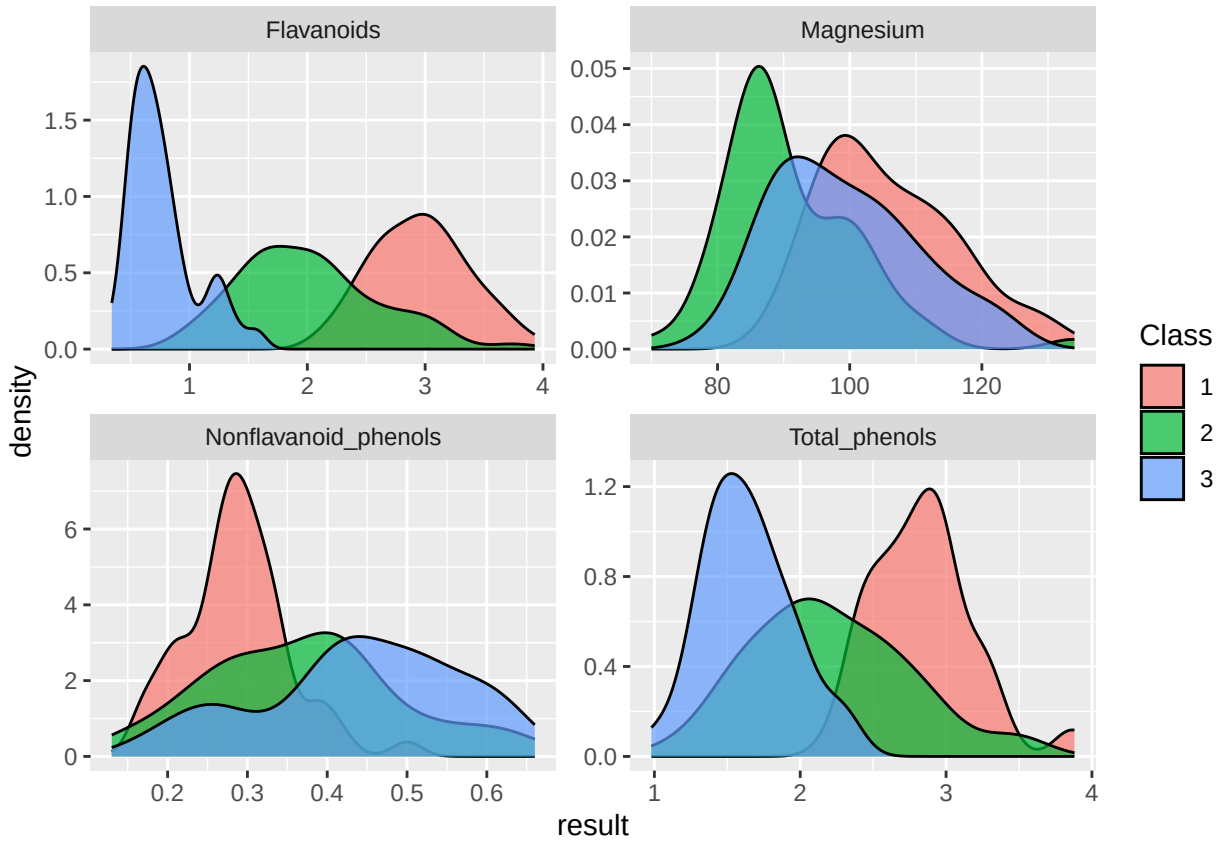
Density Plot

A density plot is a graphical representation of the distribution of a numeric variable. It shows the frequency of the values on the x-axis, and the density of the values on the y-axis. Its useful for visualizing the overall shape of a distribution and identifying any potential outliers or unusual values.

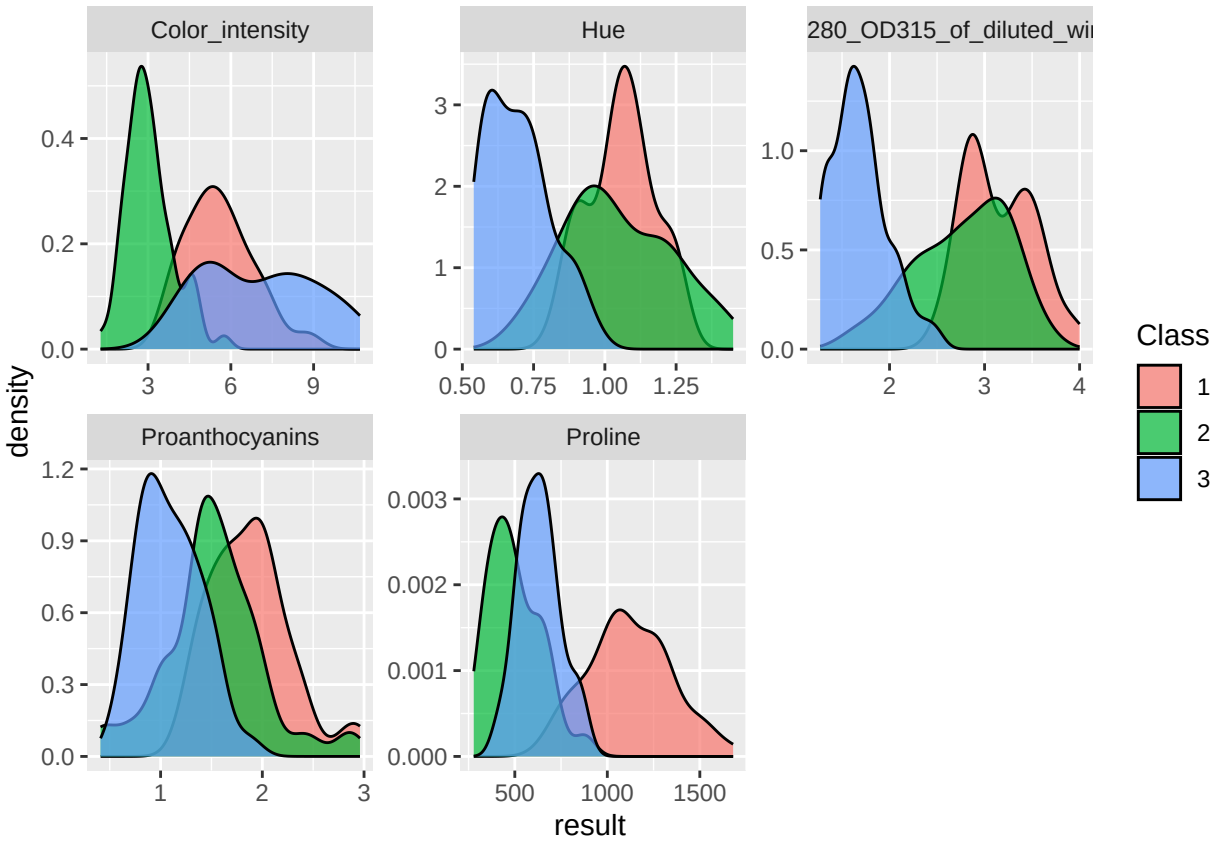
```
clean_Data %>% gather(2:5, key = "variables", value = "result") %>%
  ggplot(aes(result, fill = Class)) +
  geom_density(alpha = 0.7)+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```




```
clean_Data %>% gather(6:9, key = "variables", value = "result") %>%
  ggplot(aes(result, fill = Class)) +
  geom_density(alpha = 0.7)+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```




```
clean_Data %>% gather(10:14, key = "variables", value = "result") %>%
  ggplot(aes(result, fill = Class)) +
  geom_density(alpha = 0.7)+
  theme_get()+
  facet_wrap(.~variables, scale = "free")
```



5. Summary

Overall in this wine recognition data sets analysis we start with exploratory data analysis. We looked for outliers, missing values/ unusual values and clean those values by different method in this section. In the next section to understand the relationship between different variables by class we introduced 4 different plots. Each plots shows the relation of each variables by class. First we make box plot to see if there is any outliers or not, We choose violin plot with summary statistics for closer view of variables. Next, by applying histogram we can see the distribution of each variable in each class, their maximum minimum values and other statistics summary. Atlast we saw the distribution of each variable in density plot and all the variables are well distributed by class.