Wine Constituents Analysis

Anshuman Sharma, Abhishek Motlani, Lisa Fukutoku

Data – Wine Constituents

The data consists of the quantities of 13 constituents of wines based on a chemical analysis.
The wines are grown in the same region in Italy but derived from three different cultivars, which we call type 1,2 and 3.

The attributes are:

1. **Alcohol**
2. **Malic acid** – tartness and acidity in wine
3. **Ash** – inorganic residue after burning
4. **Alcalinity of ash** – acid neutralizing capacity
5. **Magnesium** – influences aroma, flavor, and texture
6. **Total phenols** – collective measurement of all phenolic compounds
7. **Flavanoids** – color, flavor, and health benefits
8. **Nonflavanoid phenols** – astringency and bitterness
9. **Proanthocyanins** – color, mouthfeel, and aging potential
10. **Color intensity** – depth and concentration of color
11. **Hue** – dominant shade or tint of color
12. **OD280/OD315 of diluted wines** – protein content and wine stability
13. **Proline** – taste, texture, and aging potential

Data Set Snippet

	Wine <int>	Alcohol <dbl>	Malic.acid <dbl>	Ash <dbl>	Acl <dbl>	Mg <int>	Phenols <dbl>	Flavanoids <dbl>
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39

Nonflavanoid.phenols <dbl>	Proanth <dbl>	Color.int <dbl>	Hue <dbl>	OD <dbl>	Proline <int>
0.28	2.29	5.64	1.04	3.92	1065
0.26	1.28	4.38	1.05	3.40	1050
0.30	2.81	5.68	1.03	3.17	1185
0.24	2.18	7.80	0.86	3.45	1480
0.39	1.82	4.32	1.04	2.93	735
0.34	1.97	6.75	1.05	2.85	1450

- The first “Wine” column, indicating the type of wine 1, 2, and 3, is categorical.
- 13 columns, expressing 13 different wine constituents, are continuous.
- Total of 178 instances – there was no missing data.

Exploratory Data Analysis (EDA)

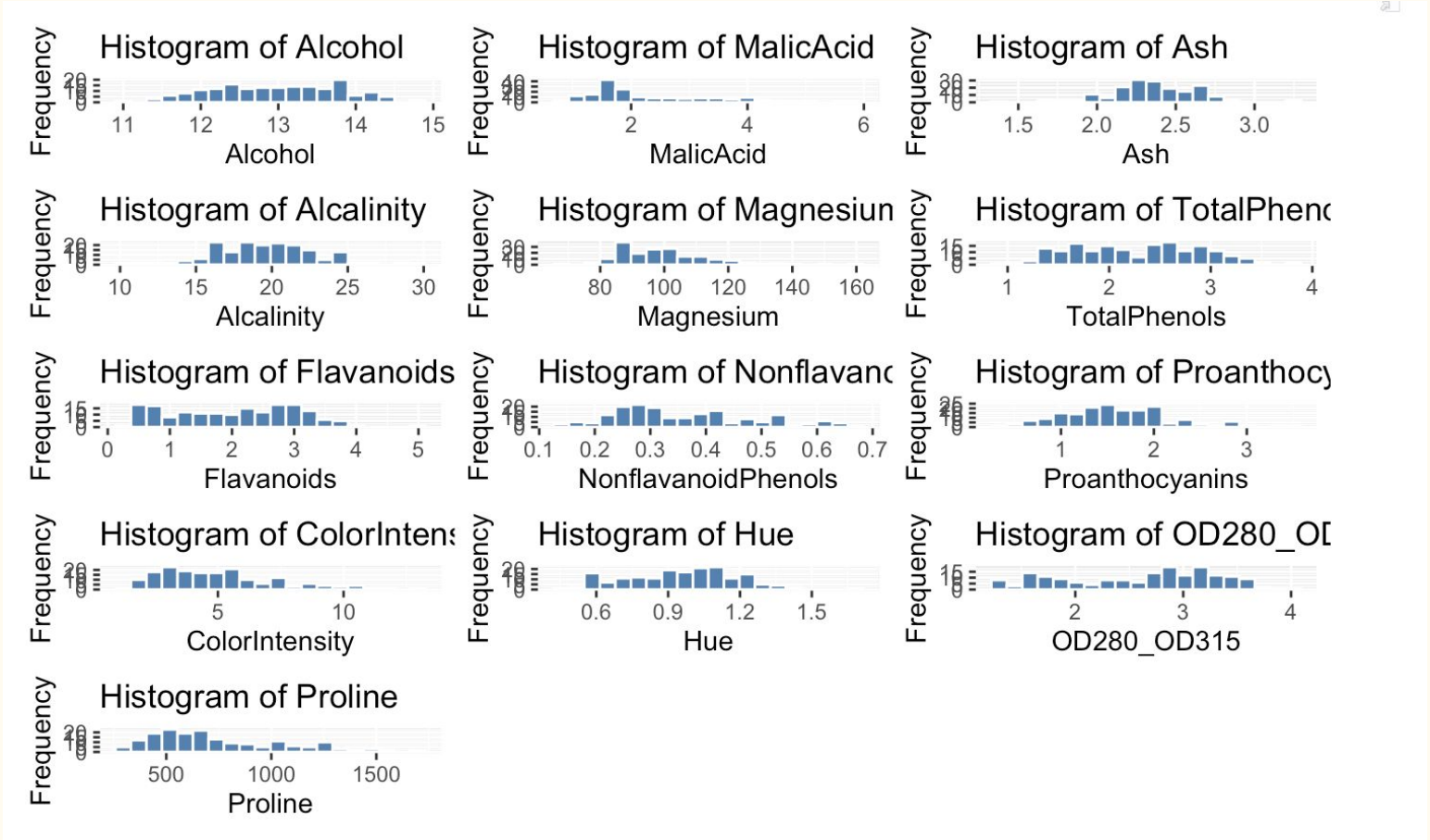
Based on the output of summary function on the dataset, these are some major findings:

- Alcohol Content Range:
  - The alcohol content in the dataset ranges from 11.03% to 14.83%, with a median value of 13.05% and a mean value of 12.99%.
  - This indicates that the majority of wines have an alcohol content around the median value.

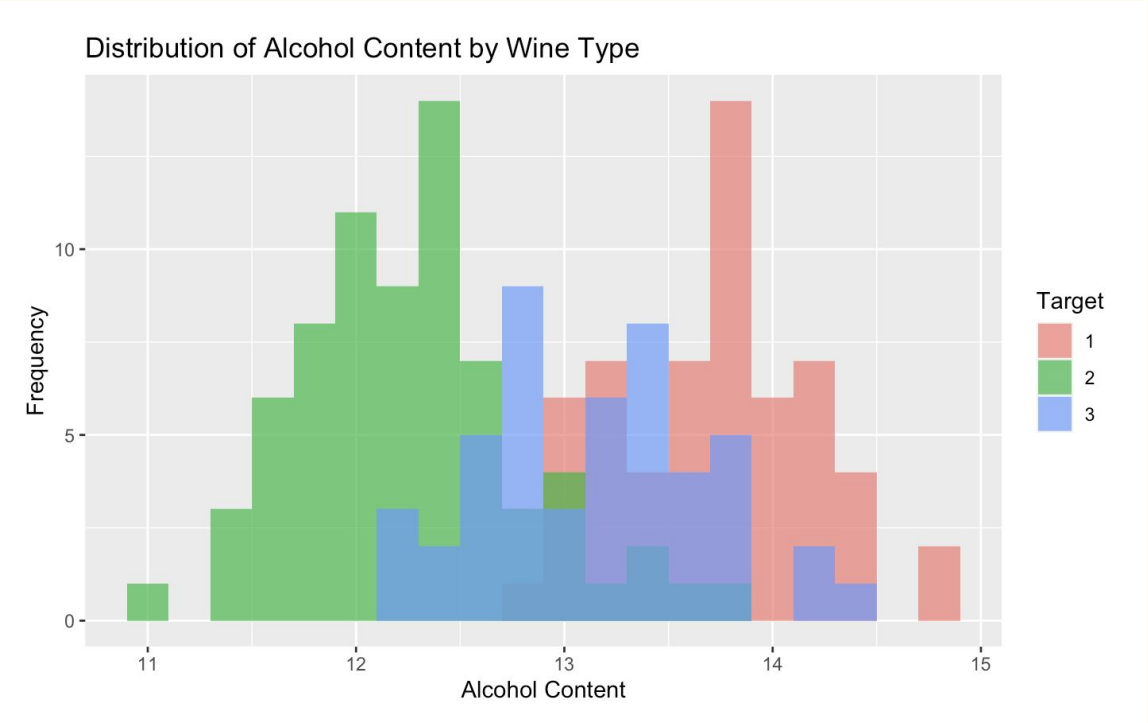
- Variability in Attributes:
  - The dataset shows variations in other attributes such as MalicAcid, Ash, Alcalinity, Magnesium, TotalPhenols, Flavanoids, NonflavanoidPhenols, Proanthocyanins, ColorIntensity, Hue, OD280_OD315, and Proline.
  - These attributes exhibit different ranges, means, and distributions, suggesting diversity in the dataset.

- Distribution Shape:
  - Attributes like MalicAcid, Ash, Alcalinity, ColorIntensity, and OD280_OD315 have a roughly symmetrical distribution, as their median and mean values are quite close.
  - On the other hand, attributes such as Magnesium, Flavanoids, Proanthocyanins, Hue, and Proline show some level of skewness, as their median and mean values differ.

Distribution of Each Constituent



Distribution of Alcohol Level by Types of Wine



Based on this distribution, it seems that there are significant differences in alcohol levels depending on the wine type.

Therefore, we further examine the associations of Alcohol content among different types of wine to prove the point in our second question.

Questions

1. Is there any correlation between Alcohol content and other chemical attributes?
2. Is there a significant difference in the mean Alcohol content between three different types of wine?
3. Can we identify distinct types of wines based on their constituents?

First Question

- Is there any correlation between Alcohol content and other chemical attributions?

We will examine whether there is an association between the Alcohol level and other constituents.

Methods of Analysis for Question 1

The correlation analysis was performed to investigate the relationship between Alcohol content and various constituents of the dataset.

- The correlation coefficient between Alcohol and each constituent was calculated using the Pearson correlation method.
- Heat Map
  - The correlation coefficient represents the strength and direction of the linear relationship between two variables. It ranges from -1 to +1.
- Bar Plot
  - It was taken the absolute values of correlation coefficients to identify the prominent constituents.

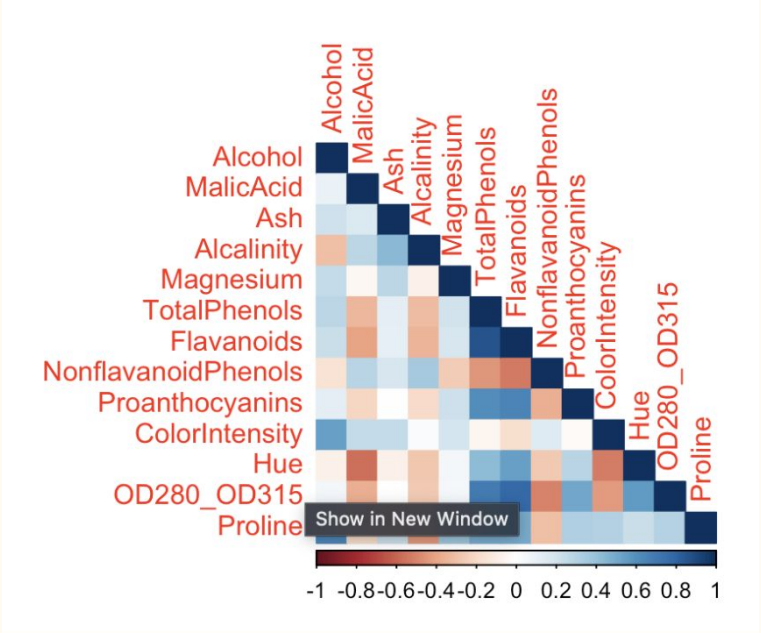
Findings

Alcohol	MalicAcid	Ash	Alcalinity	Magnesium
1.00000000	0.09996298	0.21096440	-0.30334986	0.25874233
TotalPhenols	Flavanoids	NonflavanoidPhenols	Proanthocyanins	ColorIntensity
0.28454303	0.23013326	-0.15144545	0.12756072	0.54788293
Hue	OD280_OD315	Proline		
-0.07537498	0.05741673	0.64106760		

From the correlation coefficients table and the heat map,

- Alcohol content has a positive correlation with Malic Acid (0.099), Ash (0.211), Magnesium (0.259), Total Phenols (0.285), Flavanoids (0.23), Proanthocyanins (0.128), Color Intensity (0.548), OD280_OD315 (0.057), and Proline(0.641).
    - This suggests that as the Alcohol content increases, there is a tendency for these constituents to also increase.

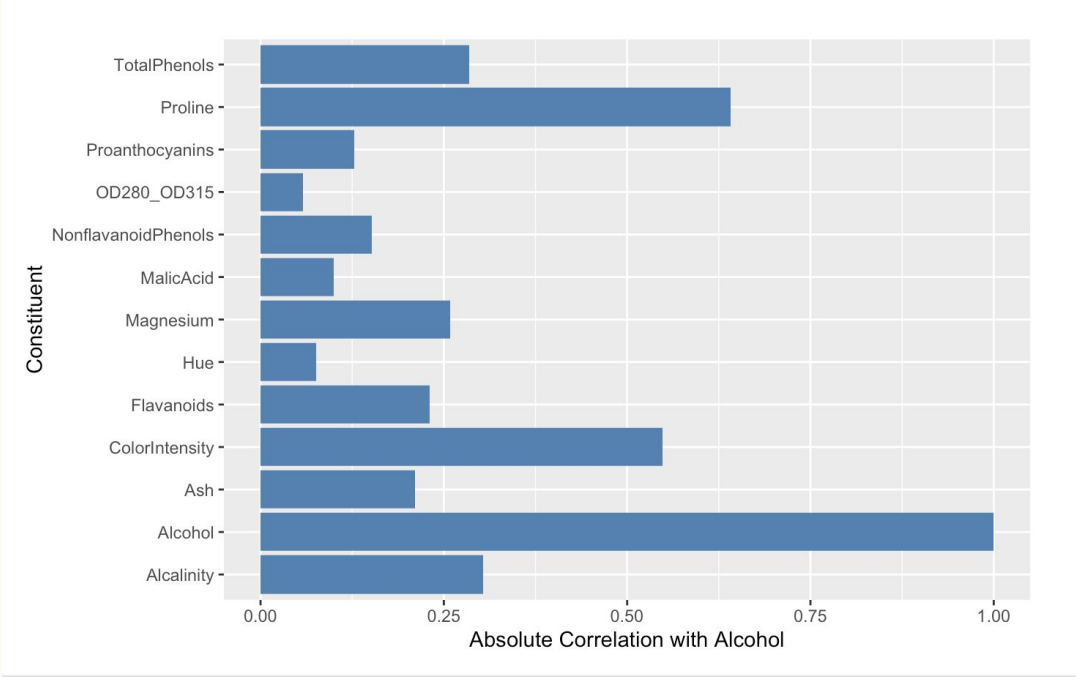
- Alcohol content has a negative correlation with Alcalinity (-0.303), Nonflavanoid Phenols (-0.151), and Hue (0.075).
    - This indicates that as the Alcohol content increases, the Alcalinity tends to decrease.



Bar Plot:

- The absolute values of correlation coefficients were obtained to identify the major impacts of each constituent on the alcohol level.

Proline, Color Intensity, and Alcalinity are identified as the three most prominent constituents based on their correlation coefficients.



Second Question

- Is there a significant difference in the mean Alcohol content between three different types of wine?

We will investigate whether there is a notable difference in the average Alcohol level among the different wine types.

Method of Analysis for Question 2

Multivariate analysis was performed to compare the alcohol content across wine types.

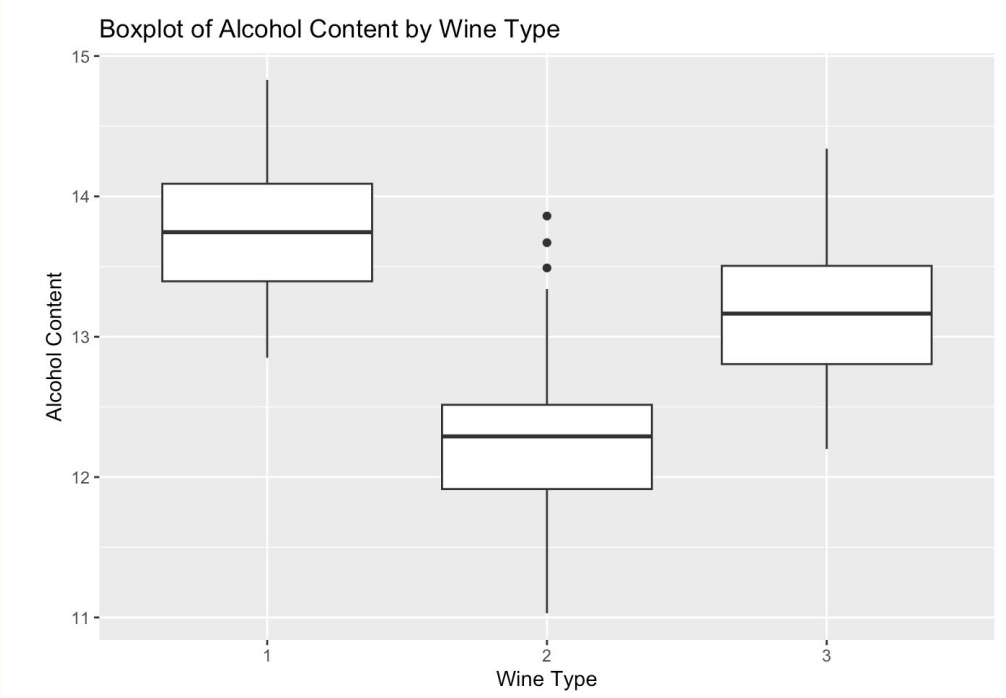
- MANOVA test (Multivariate Analysis of Variance)
- Box Plot
  - A box plot visualization was created to compare alcohol content by wine type.
  - It shows the distribution of alcohol content for each wine type.
- Hotelling T^2 Test
  - Pairwise Hotelling's T^2 tests were conducted to confirm the significant differences between each pair of wine types.

Findings

MANOVA Results:

- $H_0 : \mu_1 = \mu_2 = \mu_3$ vs H_a : the null is not true
- F-statistic: 118.48
- p-value: $<$2.2e-16
- We reject the null hypothesis

It indicates a significant difference in the mean Alcohol content among wine types.



Hotelling T^2 Test:

- $H_0 : \mu_1 = \mu_2$ / $H_0 : \mu_1 = \mu_3$ / $H_0 : \mu_2 = \mu_3$
- H_a : the null is not true

- The p-values for all pairwise comparisons are 0, we reject the null hypothesis.
- This indicates significant differences in the mean Alcohol content between each pair of wine types.

This confirms our conclusion from the MANOVA test.

Alcohol content plays a role in determining the wine type, highlighting the importance of considering alcohol content in assessing wine types.

```
Comparison between wine type 1 and wine type 2:
Test stat:  334.85
Numerator df:  3
Denominator df:  125
P-value:  0

Comparison between wine type 1 and wine type 3:
Test stat:  173.8
Numerator df:  3
Denominator df:  102
P-value:  0

Comparison between wine type 2 and wine type 3:
Test stat:  226.17
Numerator df:  3
Denominator df:  115
P-value:  0
```

Third Question

- Can we identify distinct groups of wines based on their chemical constituents?

We will examine whether we are able to classify the types of wines based on the data of constituents.

Method of Analysis for Question 3

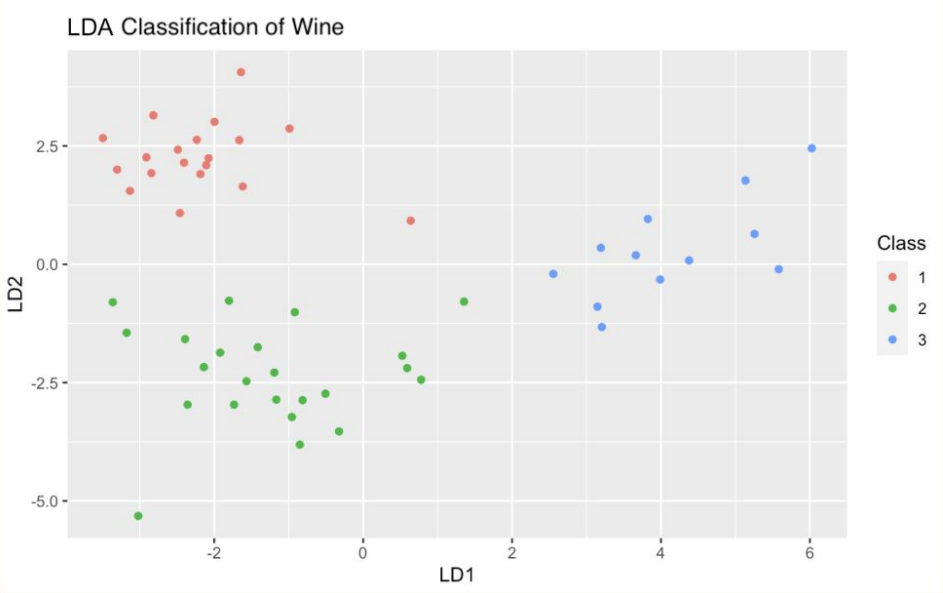
- Linear Discriminant Analysis (LDA)
  - It is a dimensionality reduction technique commonly used for classification and clustering tasks.
  - It aims to find a linear combination of features that maximizes the separation between different classes or clusters.
  - It is particularly useful when the goal is to find a low-dimensional representation of the data that preserves the class or cluster information.

Findings

By applying LDA to the wine dataset, we were able to identify distinct groups or clusters of wines based on their chemical attributes.

- The LDA model achieved an impressive accuracy of **98.2%** on the test data, indicating its effectiveness in distinguishing different wine types.
- The scatter plot visualization of the LDA classification clearly shows the separation of wines based on their chemical attributes.
- This suggests that the chemical composition of wines plays a significant role in determining their type or category.

The identified clusters can provide valuable insights for wine producers, distributors, and connoisseurs, enabling them to better understand the characteristics and quality of different wine types.



Conclusion

- Through our R implementation and statistical analysis on wine dataset, we drew the conclusions as following:
  - Proline, Color Intensity, and Alcalinity are identified as the three most prominent constituents based on their correlation coefficients against alcohol level, meaning that these components have an association with Alcohol.
  - There is a significant difference in the average Alcohol level among three wine types, suggesting that the wine types are distinct in terms of their alcohol levels.
  - Distinct types of wines are able to be classified correctly based on their chemical constituents with 98% of the accuracy.

- Our findings highlight the importance of considering constituents of wines in distinguishing wine types.
- Overall, this study provided valuable insights into the differences and associations of various constituents of wines, grown in the same region but derived from three different cultivars.

Reference

- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository, Wine Data Set [https://archive.ics.uci.edu/ml/datasets/Wine]. Irvine, CA: University of California, School of Information and Computer Science.

Appendix – Major R Code

```r
# Create a data frame with the attributes of interest
attributes <- dataset[, c("Alcohol", "MalicAcid", "Ash", "Alcalinity", "Magnesium", "TotalPhenols", "Flavanoids",
"NonflavanoidPhenols", "Proanthocyanins", "ColorIntensity", "Hue", "OD280_OD315", "Proline")]

# Create a list to store the histogram plots
histogram_plots <- list()

# Generate histograms for each attribute
for (col in colnames(attributes)) {
  histogram_plot <- ggplot(attributes, aes(x = .data[[col]])) +
    geom_histogram(fill = "steelblue", color = "white", bins = 20) +
    labs(x = col, y = "Frequency") +
    ggtitle(paste("Histogram of", col))

  histogram_plots[[col]] <- histogram_plot
}

# Combine the histogram plots into a single grid
histogram_grid <- cowplot::plot_grid(plotlist = histogram_plots, ncol = 3)

# Display the histogram grid
histogram_grid
```

```
variables <- c("Alcohol", "MalicAcid", "Ash", "Alcalinity", "Magnesium", "TotalPhenols", "Flavanoids", "Nonflavan
oidPhenols", "Proanthocyanins", "ColorIntensity", "Hue", "OD280_OD315", "Proline")
cor_matrix <- cor(dataset[, variables])

cor_alcohol <- cor_matrix["Alcohol", ]
print(cor_alcohol)
```

```
##            Alcohol           MalicAcid                 Ash          Alcalinity
##         1.00000000          0.09996298          0.21096440         -0.30334986
##          Magnesium        TotalPhenols          Flavanoids NonflavanoidPhenols
##         0.25874233          0.28454303          0.23013326         -0.15144545
##     Proanthocyanins      ColorIntensity                 Hue         OD280_OD315
##         0.12756072          0.54788293         -0.07537498          0.05741673
##            Proline
##         0.64106760
```

Plotting, to observe things more clearly

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor_matrix, method = "color", type = "lower")
```

```r
# Load the necessary libraries
library(stats)

# Create a data frame with the variables of interest
manova_data <- dataset[, c("Target", "Proline", "ColorIntensity", "Alcalinity")]

# Perform MANOVA
manova_result <- manova(cbind(Proline, ColorIntensity, Alcalinity) ~ Target, data = manova_data)

# Summarize the MANOVA results
summary(manova_result)
```

```r
# Create matrices for each wine type
wine_type1 <- subset(dataset, Target == 1, select = c("Alcohol", "ColorIntensity", "Alcalinity"))
wine_type2 <- subset(dataset, Target == 2, select = c("Alcohol", "ColorIntensity", "Alcalinity"))
wine_type3 <- subset(dataset, Target == 3, select = c("Alcohol", "ColorIntensity", "Alcalinity"))

# Perform pairwise Two-Sample Hotelling's T2 tests
result_1_2 <- hotelling.test(as.matrix(wine_type1), as.matrix(wine_type2))
result_1_3 <- hotelling.test(as.matrix(wine_type1), as.matrix(wine_type3))
result_2_3 <- hotelling.test(as.matrix(wine_type2), as.matrix(wine_type3))

# Print the results
cat("Comparison between wine type 1 and wine type 2:\n")
```

Can we identify distinct groups or clusters of wines based on their chemical attributes?

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
# Separate the features and the target variable
X <- dataset[, 2:13]  # Features
y <- dataset$Target   # Target variable

# Split the data into training and test sets
set.seed(123)
train_indices <- sample(1:length(y), 0.7 * length(y))  # 70% for training
train_data <- X[train_indices, ]
train_labels <- y[train_indices]
test_data <- X[-train_indices, ]
test_labels <- y[-train_indices]

# Train the LDA model
lda_model <- lda(train_data, train_labels)

# Predict the labels for the test data
lda_pred <- predict(lda_model, test_data)

# Calculate the accuracy of the model
accuracy <- sum(lda_pred$class == test_labels) / length(test_labels)
accuracy
```

```
## [1] 0.9814815
```