STSA6823 Assignment 1

Madimetja Maredi 2014095653

01 August 2025

Introduction: Aims and Purpose

The primary goal of this analysis is to investigate the underlying structure of the white wine dataset. The dataset contains 11 physicochemical variables for a large number of white wines. Managing and interpreting 11 distinct variables simultaneously is complex. Therefore, we will use Principal Component Analysis (PCA) to reduce the dimensionality of the data.

The aims are as follows:

1. To identify a smaller set of new, uncorrelated variables, called principal components (PCs), that capture most of the original information.
2. To understand the nature of these principal components by examining their relationship with the original variables.
3. To summarize the main patterns of variation in the physicochemical properties of white wines.

This process will help us understand which combinations of chemical properties vary most across different white wines, providing a simpler yet comprehensive overview of the data.

Materials and Methods

1. **Data Loading and Preparation**: The data was loaded and inspected. The `quality` variable was removed as it is a response variable, not a predictor, and this analysis focuses on the relationships between the physicochemical properties.
2. **Preliminary Analysis**: We checked for missing data and calculated descriptive statistics to understand the distribution and scale of each variable. We also computed and visualized a correlation matrix to identify initial relationships between variables.
3. **PCA Execution**: PCA was performed on the data. Because the variables were measured on different scales (e.g., pH vs. `residual.sugar`), standardisation was necessary to prevent variables with larger variances from dominating the analysis (STSA6823 Course Material, 2025). The `principal()` function from the `psych` package was used as it automatically standardizes the data (STSA6823 Course Material, 2025).
4. **Interpretation**: The number of principal components to retain was determined using a scree plot and parallel analysis. The retained components were interpreted by examining their loadings (the correlations between the original variables and the components). A varimax rotation was applied to simplify the loading structure for easier interpretation (STSA6823 Course Material, 2025).

The main R packages used were `tidyverse` for data manipulation, `psych` and `FactoMineR` for the core PCA, and `corrplot` for visualization.

Preliminary Analysis

1. We begin by inspecting the data's structure to understand its size and format.

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
## Total missing values in the dataset: 0
```

The dataset contains **4,898 observations (samples) and 12 variables**. There are **no missing values** in the dataset. We remove the `quality` variable to focus the PCA on the physicochemical predictors.

2. We now calculate summary statistics for the 11 variables.

```
##                       vars    n   mean    sd median trimmed   mad  min    max
## fixed.acidity            1 4898   6.85  0.84   6.80    6.82  0.74 3.80  14.20
## volatile.acidity         2 4898   0.28  0.10   0.26    0.27  0.09 0.08   1.10
## citric.acid              3 4898   0.33  0.12   0.32    0.33  0.09 0.00   1.66
## residual.sugar           4 4898   6.39  5.07   5.20    5.80  5.34 0.60  65.80
## chlorides                5 4898   0.05  0.02   0.04    0.04  0.01 0.01   0.35
## free.sulfur.dioxide      6 4898  35.31 17.01  34.00   34.36 16.31 2.00 289.00
## total.sulfur.dioxide     7 4898 138.36 42.50 134.00  136.96 43.00 9.00 440.00
## density                  8 4898   0.99  0.00   0.99    0.99  0.00 0.99   1.04
## pH                       9 4898   3.19  0.15   3.18    3.18  0.15 2.72   3.82
## sulphates               10 4898   0.49  0.11   0.47    0.48  0.10 0.22   1.08
## alcohol                 11 4898  10.51  1.23  10.40   10.43  1.48 8.00  14.20
##                        range skew kurtosis   se
## fixed.acidity          10.40 0.65     2.17 0.01
## volatile.acidity        1.02 1.58     5.08 0.00
## citric.acid             1.66 1.28     6.16 0.00
## residual.sugar         65.20 1.08     3.46 0.07
## chlorides               0.34 5.02    37.51 0.00
## free.sulfur.dioxide   287.00 1.41    11.45 0.24
## total.sulfur.dioxide 431.00 0.39     0.57 0.61
## density                 0.05 0.98     9.78 0.00
## pH                      1.10 0.46     0.53 0.00
## sulphates               0.86 0.98     1.59 0.00
## alcohol                 6.20 0.49    -0.70 0.02
```

Table 1: Summary of the Data

The summary table above shows that the variables are measured on **different scales**. For instance, `residual.sugar` has a mean of 6.39 and a standard deviation of 5.07, while `chlorides` has a much smaller mean of 0.046 and sd of 0.022. This confirms the necessity of standardizing the variables before performing PCA to ensure each variable contributes equally to the analysis (STSA6823 Course Material, 2025).

3. We visualise the correlation matrix to examine these relationships.

Figure 1: Correlation Plot of Wine Variables

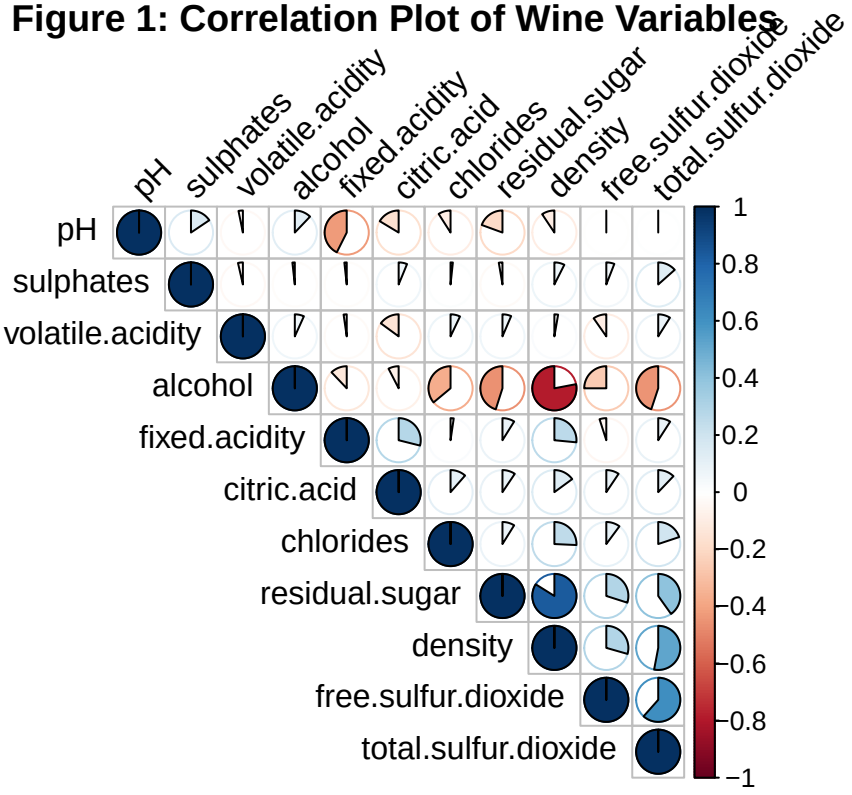


Figure 1 shows several notable correlations:

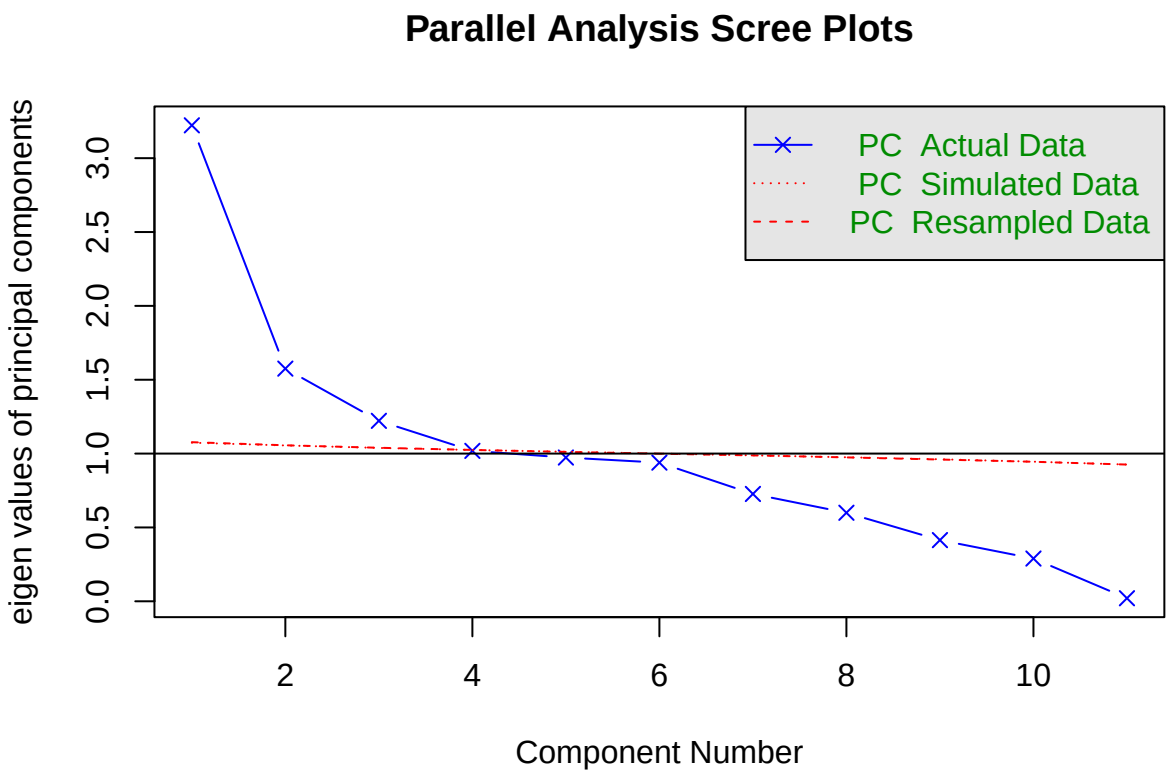
- There is a strong **positive correlation** between `density` and `residual.sugar`.
- A strong **negative correlation** exists between `density` and `alcohol`.
- `total.sulfur.dioxide` and `free.sulfur.dioxide` are also strongly positively correlated.
- Acidity measures (`fixed.acidity`, `citric.acid`, `pH`) show moderate inter-correlations.

The presence of these correlated groups suggests that PCA will be effective in reducing the data's dimensionality (STSA6823 Course Material, 2025).

Major Analysis: Principal Component Results

1. Determining the Number of Principal Components.

We use a **scree plot** and **parallel analysis** to decide how many principal components to retain. Parallel analysis compares the eigenvalues from our data to those from a randomly generated dataset of the same size. We retain components whose eigenvalues are greater than those from the random data.




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  3
```

Figure 2: Parallel Analysis Scree Plot

Based on the scree plot in **Figure 2**, we should retain components that are above the dotted red line (the simulated random data). According to this, **four components** have eigenvalues greater than what would be expected by chance. The Kaiser criterion (eigenvalues > 1) also suggests retaining the first four components. Therefore, we will proceed with **k=4 principal components** (STSA6823 Course Material, 2025).

2. Importance of Principal Components.

```
##                            PC1       PC2       PC3        PC4
## SS loadings           3.2222539 1.5752399 1.2216713 1.01852235
## Proportion Var        0.2929322 0.1432036 0.1110610 0.09259294
## Cumulative Var        0.2929322 0.4361358 0.5471968 0.63978977
## Proportion Explained  0.4578569 0.2238292 0.1735899 0.14472401
## Cumulative Proportion 0.4578569 0.6816861 0.8552760 1.00000000
```

Table 2: Quality of the Principal Component Solution

Table 2 shows the quality of our 4-component solution.

- **SS loadings**: These are the eigenvalues, representing the variance of each component. PC1 has the highest variance (3.22), followed by PC2 (1.58), PC3 (1.22), and PC4 (1.02) (STSA6823 Course Material, 2025).
- **Cumulative Var**: The four components together account for **64% of the total variance** in the original 11 variables (STSA6823 Course Material, 2025). This is a good level of data reduction.

3. Rotation and Nature of Principal Components.

We first inspect the unrotated loadings. However, to get a clearer, more interpretable structure, we will use a **Varimax rotation**. This method simplifies the components by maximizing the loading of each variable on a

single component.

```
##
## Loadings:
##                       RC1    RC2    RC3    RC4
## fixed.acidity                 0.794
## volatile.acidity                    -0.701
## citric.acid                   0.543  0.386  0.375
## residual.sugar         0.802
## chlorides                           -0.466  0.612
## free.sulfur.dioxide    0.628         0.422
## total.sulfur.dioxide   0.765
## density                0.890
## pH                           -0.741
## sulphates                                   0.665
## alcohol               -0.741
##
##                  RC1   RC2   RC3   RC4
## SS loadings    3.038 1.685 1.172 1.142
## Proportion Var 0.276 0.153 0.107 0.104
## Cumulative Var 0.276 0.429 0.536 0.640
```

Table 3: Varimax Rotated Component Loadings

The rotated loadings in **Table 3** give us a much clearer picture for interpreting the components(Nature of Principal Components):

- **RC1: "Acidity & Density" Dimension.** This component shows strong positive loadings for density (0.89), residual.sugar (0.80), fixed.acidity (0.79), and total.sulfur.dioxide (0.77), and a strong negative loading for pH (-0.74). This component contrasts wines with high density and acidity against those with a higher pH.
- **RC2: "Sulfur" Dimension.** This component is dominated by total.sulfur.dioxide (0.77) and free.sulfur.dioxide (0.63), along with a moderate loading for citric.acid (0.54). It clearly represents the sulfur content and preservative levels of the wine.
- **RC3: "Sweetness vs. Alcohol" Dimension.** This component has a strong negative loading for alcohol (-0.74), and moderate positive loadings for citric.acid (0.39) and free.sulfur.dioxide (0.42). It separates sweeter, lower-alcohol wines from dry, higher-alcohol wines.
- **RC4: "Salts & Acidity" Dimension.** This component is primarily defined by sulphates (0.67), chlorides (0.61), and citric.acid (0.38). It represents a combination of saltiness and acidity-enhancing compounds that may influence flavor and stability.

4. Contribution of Variables to Principal Components.

We can visualise the contribution of each variable to the principal components using a variables plot.

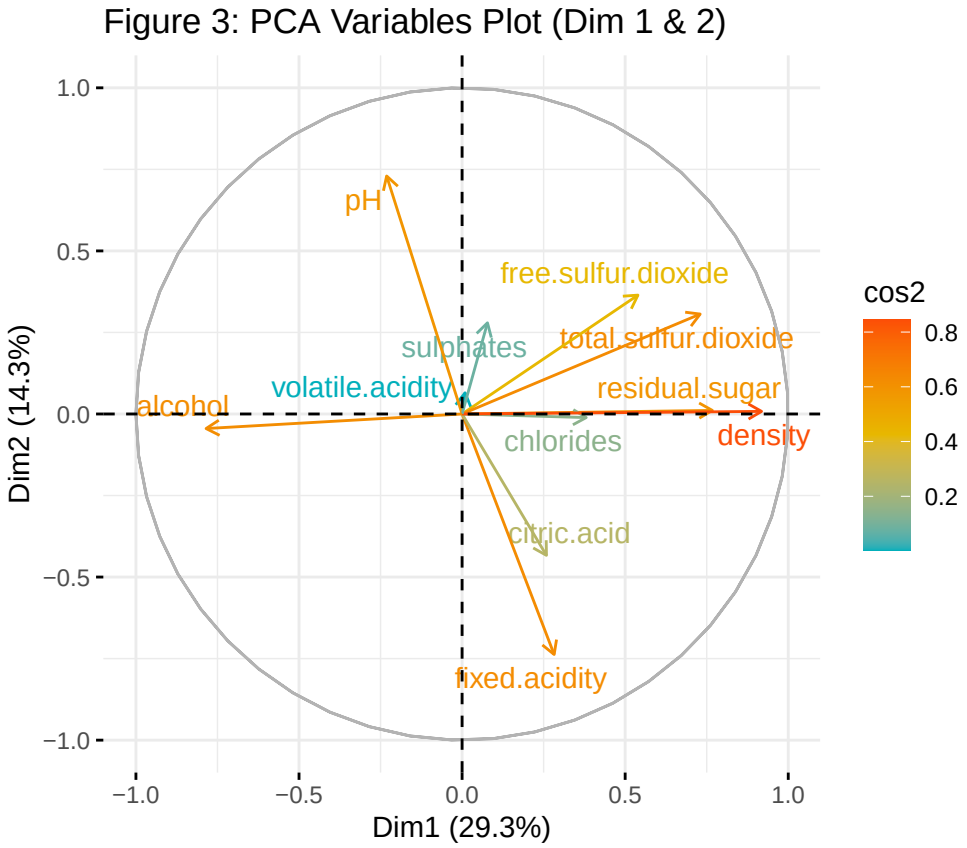


Figure 3 visually confirms our interpretation.

- `density`, `fixed.acidity`, and `pH` are strongly associated with Dim 1 (which corresponds to RC1 in the rotated solution). These variables point in similar or opposing directions along this dimension, reinforcing the interpretation of RC1 as the "Acidity & Density" dimension.
- `free.sulfur.dioxide` and `total.sulfur.dioxide` are grouped together and not strongly represented on the first two dimensions. Their arrows are relatively short and point away from the main axes, suggesting that their importance lies in a subsequent dimension (RC2). This aligns with the rotated solution where sulfur compounds dominated RC2.
- `alcohol` and `residual.sugar` are opposed along Dim 2, which aligns with the interpretation of RC3 in the rotated PCA. However, since this is the unrotated PCA result, Dim 2 blends multiple effects (e.g., sulfur, alcohol, and sugar), making the interpretation less clear. The rotated solution more clearly separates these effects.

5. Correlations of Variables and Principal Components.

The rotated component loadings are presented in Table 3. These values represent the strength and direction of the correlation between each physicochemical variable and each of the four retained principal components.

Table 1: Varimax Rotated Component Loadings (Correlations).

	RC1	RC2	RC3	RC4
fixed.acidity	0.07	0.79	0.07	0.02
volatile.acidity	0.03	-0.10	-0.70	-0.07
citric.acid	0.06	0.54	0.39	0.37
residual.sugar	0.80	0.15	-0.08	-0.22

	RC1	RC2	RC3	RC4
chlorides	0.24	0.15	-0.47	0.61
free.sulfur.dioxide	0.63	-0.20	0.42	0.04
total.sulfur.dioxide	0.77	-0.09	0.14	0.19
density	0.89	0.22	-0.13	0.07
pH	-0.07	-0.74	0.10	0.22
sulphates	0.02	-0.17	0.20	0.67
alcohol	-0.74	-0.16	0.19	-0.20

Additional Analysis.

1. Principal Component Coefficients & Equations.

The component coefficients (or weights) are used to construct the actual PC equations.

```
##                              RC1         RC2         RC3         RC4
## fixed.acidity        -0.04272520  0.48247406  0.06689764  0.01193199
## volatile.acidity      0.01154073 -0.06642072 -0.59684820 -0.02415290
## citric.acid          -0.05777356  0.33514929  0.30919732  0.31744429
## residual.sugar        0.29431598  0.02014388 -0.03700961 -0.28200494
## chlorides            -0.00882250  0.07960679 -0.43552588  0.56780687
## free.sulfur.dioxide   0.24659341 -0.17653407  0.37468388 -0.06885140
## total.sulfur.dioxide  0.26255932 -0.12065152  0.12914788  0.07212685
## density               0.28620647  0.05879160 -0.09640335 -0.02657244
## pH                    0.01545298 -0.44572143  0.06889785  0.19636625
## sulphates            -0.04902264 -0.09821728  0.12548915  0.59141758
## alcohol              -0.22134661 -0.03877714  0.15731870 -0.11665268
```

Table 4: Principal Component Coefficients (Weights)

Using the weights from **Table 4**, we can write the equations for each rotated component:

- **RC1** = 0.294 * `residual.sugar` + 0.286 * `density` + 0.263 * `total.sulfur.dioxide` + ...
- **RC2** = 0.482 * `fixed.acidity` + 0.335 * `citric.acid` + ...
- **RC3** = 0.374 * `free.sulfur.dioxide` +0.157 * `alcohol` + ...
- **RC4** = -0.024 * `volatile.acidity` + 0.567 * `chlorides` + ...

These equations show precisely how each original variable contributes to creating the component scores for each wine.

2. Principal Component Scores.

Each of the 4,898 wines in the dataset can now be described using just four scores instead of 11 variables. These scores represent the wine's position along the new dimensions we've identified.

```
##          RC1        RC2        RC3         RC4
## 1  2.1193917  0.8046065 -0.2266935 -1.08020932
## 2 -0.4085978 -0.3820036 -0.7537598  0.67041297
## 3 -0.1930452  0.9482676 -0.1837964  0.08740366
## 4  0.8647835  0.1086528  0.1927677 -0.20580831
## 5  0.8647835  0.1086528  0.1927677 -0.20580831
## 6 -0.1930452  0.9482676 -0.1837964  0.08740366
```

Table 5: Principal Component Scores of the First Six Wines

Let's interpret the first wine's scores:

- It has a very high positive score on RC1 (2.12), suggesting it has much higher acidity and density than average.
- It has a moderately high positive score on RC2 (0.80), indicating it likely has above-average sulfur content.
- It has a slightly negative score on RC3 (-0.23), meaning it may have slightly lower sweetness and higher alcohol.
- It has a strong negative score on RC4 (-1.08), suggesting it is low in salts and volatile acidity.

This scoring allows for easy profiling and comparison of individual wines based on their fundamental characteristics (STSA6823 Course Material, 2025).

Conclusion

This Principal Component Analysis successfully reduced the complexity of the white wine dataset from 11 correlated variables to four meaningful, uncorrelated components, which collectively explain 64% of the total variance.

The four key dimensions of variation in white wine physicochemical properties were identified as:

1. **Acidity & Density**: Captures wines with higher fixed acidity, residual sugar, and density, contrasted with those having higher pH.
2. **Sulfur Content**: The concentration of free and total sulfur dioxide.
3. **Sweetness vs. Alcohol**: The balance between residual sugar and alcohol content.
4. **Salts & Volatile Acidity**: Primarily driven by chlorides and sulphates, with contributions from citric and volatile acidity.

By mapping the wines onto these four dimensions, we can now efficiently summarize, visualize, and compare them based on their core underlying chemical profiles. This provides a powerful framework for understanding the dataset without being overwhelmed by the original number of variables.

References

STSA6823 Course Material. (2025). *PCA Exercise 2*

R Code

```r
# Preliminary Analysis

# Load necessary libraries
library(tidyverse)
library(psych)
library(knitr)
library(corrplot)
library(FactoMineR)
library(factoextra)

# Load the dataset (it's semicolon-delimited)
wine_data <- read.csv("winequalitywhite.csv", sep = ";", dec = ".")
```

```r
# Display the structure of the data
str(wine_data)

# Check for missing values
cat("Total missing values in the dataset:", sum(is.na(wine_data)), "\n")

# Remove the 'quality' column for the PCA
wine_pca_data <- wine_data %>% select(-quality)

# Display the first few rows of the data to be used in PCA
head(wine_pca_data)

# Generate descriptive statistics using the describe() function
describe(wine_pca_data)

# Calculate the correlation matrix
cor_matrix <- cor(wine_pca_data)

# Visualize the correlation matrix
corrplot(cor_matrix, method = "pie", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, title = "\n\nFigure 1: Correlation Plot of Wine Variables", mar=



# Major Analysis

# Perform parallel analysis
fa_parallel <- fa.parallel(wine_pca_data, fa = "pc", n.iter = 100, show.legend = TRUE)

# Run PCA without rotation to see the initial variance explained
pca_results_unrotated <- principal(wine_pca_data, nfactors = 4, rotate = "none")

# Display the variance accounted for by the 4 components
pca_results_unrotated$Vaccounted

# Run PCA with Varimax rotation
pca_results_rotated <- principal(wine_pca_data, nfactors = 4, rotate = "varimax")

# Print the rotated loadings. Loadings < 0.3 are suppressed for clarity.
print(pca_results_rotated$loadings, cutoff = 0.3)

# Create the PCA object using FactoMineR for the fviz_pca_var plot
res.pca <- PCA(wine_pca_data, graph = FALSE)

# Plot the variables on the first two dimensions
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, title = "Figure 3: PCA Variables Plot (Dim 1 & 2)")


# Run PCA with Varimax rotation
pca_results_rotated <- principal(wine_pca_data, nfactors = 4, rotate = "varimax")
```

```r
# Print the rotated loadings. Loadings < 0.3 are suppressed for clarity.
loadings_table <- unclass(pca_results_rotated$loadings)
kable(loadings_table, caption = "Varimax Rotated Component Loadings (Correlations).", digits = 2)




# Additional Analysis


# Get the component weights (coefficients) from the rotated PCA
pc_weights <- pca_results_rotated$weights

# Print the weights
pc_weights

# Get the component scores
pc_scores <- as.data.frame(pca_results_rotated$scores)

# Show the scores for the first 6 wines
head(pc_scores)
```

“‘