University of Applied Sciences

HOCHSCHULE EMDEN·LEER

Assignment

Data Science

Author: Daanyaal Parvaize, Hrithick Sunkara

Matriculation No.: 7026795,7025712

Course of Study: Introduction To Data Sciences

First examiner: Prof. Dr. Joachim Schwarz

Submission date: 30th June

University of Applied Sciences Emden/Leer · Faculty of Technology · Mechanical Engineering Department

Constantiaplatz 4 · 26723 Emden · http://www.hs-emden-leer.de

Contents

Contents i

1. AI Usage Compliance Summary 2

2. Introduction 3

3. Problem Statement 4

3.1. Descriptive Statistics and Distribution Assessment 4

3.2. Hypothesis Testing using t-Test 4

3.3. Linear Regression for Red Wines 4

3.4. Classification of Good and Bad Wines 5

3.5. Predicting Wine Color using Logistic Regression 5

3.6. Factor Analysis for Dimensionality Reduction 5

3.6.1. Problem Statement Overview 6

4. Methodology 7

4.1. Dataset Description 7

4.2. Data Preparation 7

4.3. Exploratory Data Analysis 7

4.4. Statistical Modeling 8

4.4.1. Methodology Overview 8

5. Task-1: Exploratory Data Analysis 9

5.1. Part-A: Descriptive Statistics 9

5.2. Standard Deviation Analysis 9

5.3. Skewness Analysis 10

5.4. Missing Values 10

5.5. Summary of Part-A 10

5.6. Part-B: Visual Exploration 10

5.6.1. Distribution and Outlier Analysis 10

5.6.2. Summary Table 11

5.6.3. Conclusions 11

5.6.4. R Code for Visualizations 11

6. Task-2: Analysis of Alcohol Content Differences Between Red and White Wines **13**
6.1. R Code Snippet . . . 13
6.2. Results . . . 13
6.3. Inference . . . 14
6.4. Assumptions Check . . . 14

7. Task-3: Red-Wine Analysis **15**
7.1. Data Preparation . . . 15
7.2. Multiple Linear Regression Model . . . 15
7.3. Model Outcomes . . . 15
7.4. Assumptions Check . . . 16
7.5. Durbin-Watson Test for Residual Autocorrelation . . . 16
7.6. Summary and Insights . . . 16

8. Task-4: Predicting Good or Bad Wine Using Logistic Regression **18**
8.1. Goal . . . 18
8.2. Data Preparation . . . 18
8.3. Logistic Regression Model . . . 18
8.4. Interpretation of Coefficients and Model Fit . . . 19
8.5. Cutpoint-Based Classification and Confusion Matrix . . . 19
8.6. Model Evaluation: Lift Curve . . . 20
8.7. Model Evaluation: ROC Curve and AUC . . . 21
8.8. Conclusion . . . 22

9. Task-5: Predicting Wine Variety Using Logistic Regression **23**
9.1. Recoding Wine Variety to Binary . . . 23
9.2. Splitting Data into Training and Validation Sets . . . 24
9.3. Logistic Regression to Predict Wine Type . . . 24
9.4. Output . . . 25
9.5. Prediction on Validation Set . . . 25
9.6. Model Evaluation on Validation Set . . . 26
9.6.1. Confusion Matrix . . . 27
9.6.2. Performance Metrics . . . 27
9.6.3. ROC Curve . . . 27
9.7. Overall Model Performance . . . 27

10. Task-6: Factor Analysis on Wine Data **28**
10.1. Plan . . . 28
10.2. Initial Diagnostic Results . . . 28
10.2.1. Kaiser-Meyer-Olkin (KMO) Measure . . . 28
10.2.2. Bartlett's Test of Sphericity . . . 29

10.2.3. Measure of Sampling Adequacy (MSA) per Variable 29
10.2.4. Eigenvalues of the Correlation Matrix 30
10.2.5. Recommended Next Steps 31
10.3. Factor Analysis Results on Reduced Wine Data 31
10.3.1. Reassessment of Sampling Adequacy 31
10.3.2. Kaiser-Meyer-Olkin (KMO) Measure 32
10.3.3. Bartlett's Test of Sphericity 32
10.3.4. Measure of Sampling Adequacy (MSA) per Variable 33
10.4. Factor Extraction and Loadings 33
10.4.1. Eigenvalues 33
10.4.2. Factor Loadings 34
10.5. Interpretation of Results 34
10.5.1. Summary 34

A. Appendix: Graphical Analysis **35**
A.1. Histograms of Metric Variables 35
A.2. Boxplots of Metric Variables 47
A.3. Barplots of Categorical Variables 58

A. R Code Used in the Analysis **59**

References **63**

Declaration of Authorship

We hereby declare that we, the undersigned, are the sole authors of this document. All sources consulted for this document have been listed; all quotations from and references to these sources have been properly cited and included in chapter notes and in the list of references.

The parts that have been prepared by one of us are indicated accordingly. No version of this document, either in whole or any section of it, has been used to achieve an academic degree or any other examination.

We understand that any false statements made in this declaration may be punishable by law.

1. AI Usage Compliance Summary

Table 1.1.: Summary of AI Tool Usage and Compliance

Activity	**Usage in This Work**	**Compliance Status**
Generate ideas, brainstorming, structuring	Yes, used AI as a tool for brainstorming and structuring text	Allowed (as a tool)
Literature research and initial references	Yes, AI used to suggest references and initial literature	Allowed (as a tool)
Content summary of current state of research	No, summaries were written independently	Allowed (author work)
Content generation (full text writing)	Yes, AI-assisted in drafting and optimizing text linguistically	Allowed (as a tool)
Data analyses (code execution and results)	No, all data analysis was performed independently with professor-provided code	Allowed (author work)
Source code writing or modification	No, only professor-given code used; no AI-generated or modified code	Allowed (no violation)
Source code troubleshooting or debugging	No AI involvement; all done by author	Allowed (no violation)
Generating graphics and presentations	Yes, AI assisted in suggestions, but graphics created independently	Allowed (as a tool)
Interpretation of analysis results	No, written independently by author(s)	Allowed (author work)

2. Introduction

The analysis of wine quality based on its physicochemical properties has become an important topic in data science and applied statistics. With the increasing availability of large datasets and powerful analytical tools, it is now possible to uncover patterns and relationships within complex data to aid decision-making in industries such as winemaking. Understanding how chemical attributes relate to wine quality can help producers optimize their processes and provide consumers with better information about the products they purchase [Cor+09; Tzi+21].

This project utilizes a comprehensive wine dataset containing 6,497 observations with multiple physicochemical variables and quality ratings. Using the statistical computing environment R, the dataset is explored through descriptive statistics, visualization, and regression modeling. Key objectives include summarizing the data's distributional properties, identifying outliers, assessing skewness, and building predictive models to explain wine quality based on measurable characteristics [JC16].

The rest of the work is organized as follows: initial chapters cover theoretical background and data preparation; subsequent parts present detailed exploratory data analysis, graphical assessments, and statistical modeling; finally, conclusions and recommendations are drawn based on the findings to guide both research and practical applications in the field.

3. Problem Statement

The aim of this study is to explore and analyze the wine quality dataset using various statistical and data mining techniques in R, with a focus on understanding the underlying relationships between physicochemical properties and wine quality, color, and classification. The project is divided into six structured tasks as outlined below:

3.1. Descriptive Statistics and Distribution Assessment

(a) The dataset will first be read into R using appropriate import functions (e.g., `read.csv()`), and summary statistics for all metric variables will be computed. These include: mean, standard deviation, minimum, lower quartile (Q1), median, upper quartile (Q3), and maximum. A frequency analysis will also be conducted for categorical variables.

(b) Missing values across all variables will be documented. Additionally, graphical summaries (e.g., histograms, boxplots) will be created to visually inspect the data. The skewness coefficient will be calculated to determine the nature of distribution (symmetrical, left-skewed, or right-skewed). Outliers will be identified using both graphical and numerical methods.

3.2. Hypothesis Testing using t-Test

A two-sample t-test will be performed to examine whether there is a statistically significant difference in the alcohol content between red and white wines. The assumptions of the t-test — including normality and homogeneity of variances — will be checked before drawing any conclusions.

3.3. Linear Regression for Red Wines

This task investigates whether the quality of red wines can be explained by their chemical and sensory properties. A linear regression model will be fitted with wine quality as the dependent variable and all relevant numeric predictors as independent variables. The regression assumptions (e.g., linearity, independence, homoscedasticity, and normality of residuals) will be evaluated and documented.

3.4. Classification of Good and Bad Wines

A binary classification model will be developed to differentiate between good wines (quality ≥ 8) and bad wines (quality ≤ 4). Suitable machine learning algorithms (e.g., logistic regression, decision trees, or support vector machines) will be employed to build the classifier. The model will be trained on the chemical and sensory attributes of the wines and evaluated for its classification performance.

3.5. Predicting Wine Color using Logistic Regression

The dataset will be split into a training and a validation set to simulate a real-world predictive modeling scenario. The target variable — wine color — will be encoded as binary (e.g., red = 1, white = 0). A logistic regression model will be trained on the training data to predict wine color based on physicochemical variables. Model performance will be assessed using the validation dataset via confusion matrix, accuracy, precision, recall, and AUC (Area Under the ROC Curve).

3.6. Factor Analysis for Dimensionality Reduction

This final task aims to determine whether the chemical and sensory attributes of wine can be condensed into fewer latent factors using exploratory factor analysis. Suitability of the data for factor analysis will be tested using the Kaiser-Meyer-Olkin (KMO) statistic and Bartlett's Test of Sphericity. Variables with low Measure of Sampling Adequacy (MSA) will be removed iteratively to ensure model validity. The final factor solution will be interpreted using factor loadings and eigenvalue criteria.

Each task in this report contributes to a comprehensive understanding of the wine dataset from multiple perspectives — descriptive, inferential, predictive, and exploratory — and applies practical data science methods using R as the core computational environment.

3.6.1. Problem Overview

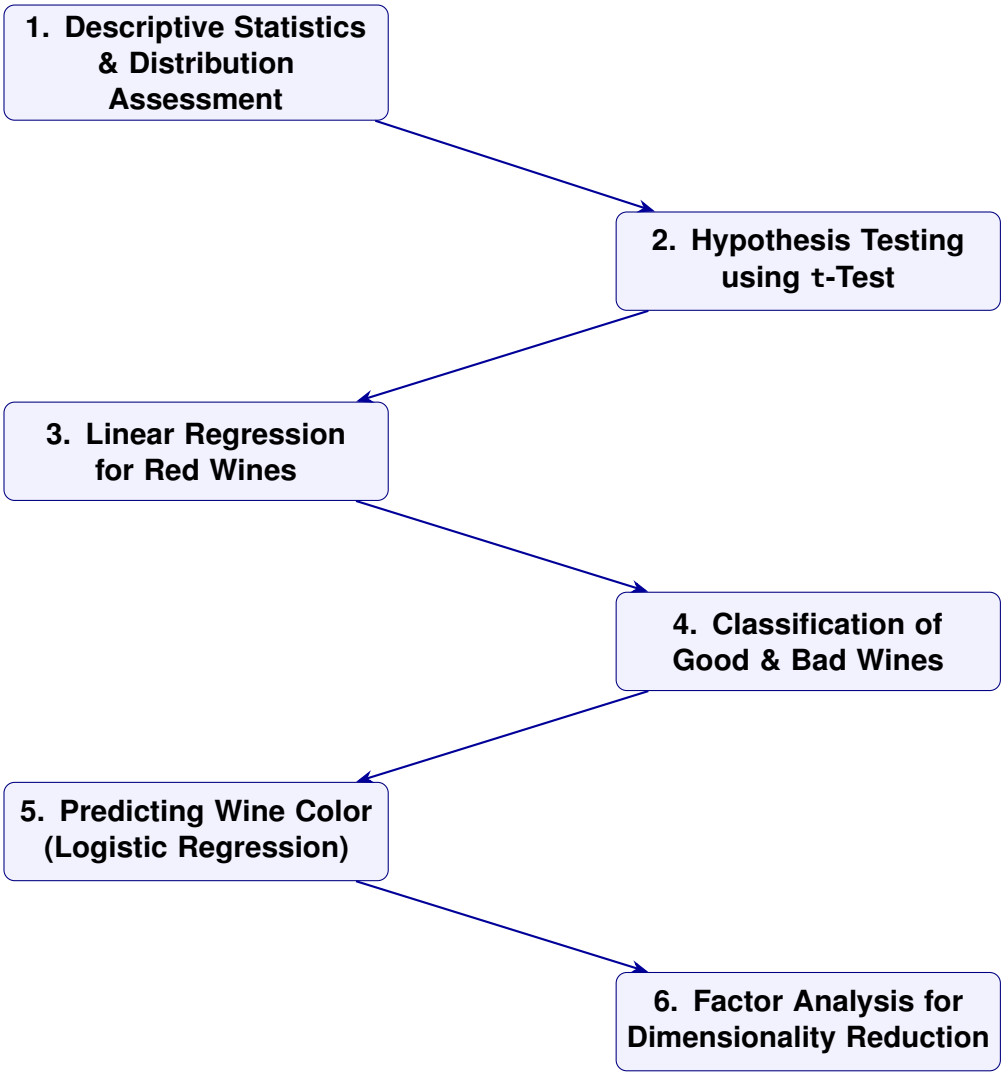

Figure 3.1.: Overview of Tasks in Wine Data Analysis Project

4. Methodology

4.1. Dataset Description

The dataset analyzed in this study comprises 6,497 observations of red and white wines, each characterized by 11 physicochemical variables and one sensory quality rating. The variables include: `fixed.acidity`, `volatile.acidity`, `citric.acid`, `residual.sugar`, `chlorides`, `free.sulfur.dioxide`, `total.sulfur.dioxide`, `density`, `pH`, `sulphates`, `alcohol`, and `quality`. Additionally, the categorical variable `variety` distinguishes between red and white wines. The dataset was originally compiled by Cortez et al. (2009) and is publicly available from the UCI Machine Learning Repository [Cor+09].

4.2. Data Preparation

All data preprocessing and statistical computations were conducted using the `R` programming language (version 4.3.0), a widely used software environment for statistical computing and graphics [R C23]. Missing values were assessed using built-in functions, and no missing entries were found, confirming data completeness. Summary statistics including the mean, median, standard deviation, and skewness were computed for all metric variables using the `psych` package [Rev23].

4.3. Exploratory Data Analysis

Initial data exploration focused on understanding the distributions and variability of each variable. Histograms and boxplots were employed to visually inspect distribution shapes and detect potential outliers. Bar plots were used to examine the distribution of the categorical variable `variety`. The skewness coefficient was calculated to quantify asymmetry in the distributions and guide transformation decisions where necessary. These visualizations and metrics helped to assess normality assumptions, guiding further modeling choices [Jam+21].

4.4. Statistical Modeling

To examine whether physicochemical attributes influence wine quality, multiple linear regression was applied using `lm()` in `R`. The dependent variable was `quality`, and independent variables included the full set of numeric predictors. Separate models were also fitted for red wines to understand varietal-specific trends. Model diagnostics such as residual plots, normal Q-Q plots, and Durbin-Watson statistics were used to check assumptions including linearity, homoscedasticity, and independence of residuals.

Binary classification was performed using logistic regression to distinguish between good-quality wines (quality ≥ 8) and poor-quality wines (quality ≤ 4). Model accuracy was evaluated using a confusion matrix, precision, recall, F1 score, and the AUC (Area Under the ROC Curve). The dataset was split into training (70%) and validation (30%) subsets to simulate real-world model deployment scenarios. These steps follow widely adopted practices in supervised learning [HTF09].

4.4.1. Methodology Overview

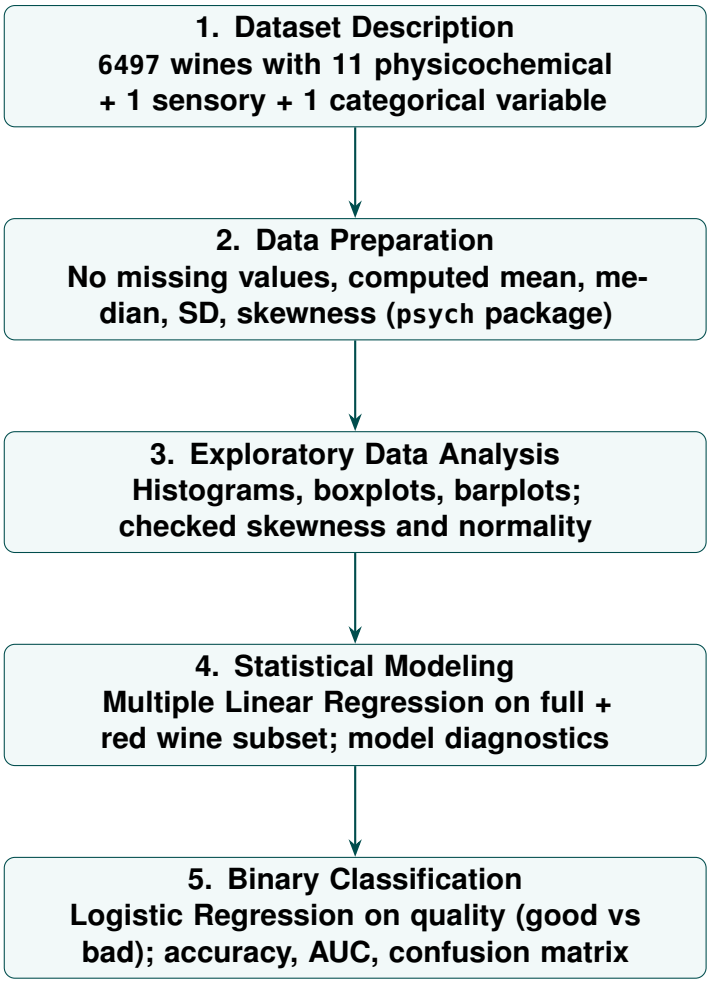

Figure 4.1.: Overview of Methodology Steps in Wine Data Analysis

5. Task-1: Exploratory Data Analysis

5.1. Part-A: Descriptive Statistics

The wine dataset provided by the professor consists of 6,497 observations with various physicochemical properties and a quality rating. The table below presents summary statistics (minimum, quartiles, mean, max, standard deviation, and skewness) for all metric variables.

Table 5.1.: Summary Statistics of Metric Variables

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Std. Dev.	Skewness
fixed acidity	3.80	6.40	7.00	7.22	7.70	15.90	1.30	1.72
volatile acidity	0.08	0.23	0.29	0.34	0.40	1.58	0.16	1.49
citric acid	0.00	0.25	0.31	0.32	0.39	1.66	0.15	0.47
residual sugar	0.60	1.80	3.00	5.44	8.10	65.80	4.76	1.43
chlorides	0.01	0.04	0.05	0.06	0.07	0.61	0.04	5.40
free sulfur dioxide	1.00	17.00	29.00	30.53	41.00	289.00	17.75	1.22
total sulfur dioxide	6.00	77.00	118.00	115.74	156.00	440.00	56.52	0.00
density	0.99	0.99	0.99	0.99	0.99	1.04	0.00	0.50
pH	2.72	3.11	3.21	3.22	3.32	4.01	0.16	0.39
sulphates	0.22	0.43	0.51	0.53	0.60	2.00	0.15	1.80
alcohol	8.00	9.50	10.30	10.49	11.30	14.90	1.19	0.57
quality	3.00	5.00	6.00	5.82	6.00	9.00	0.87	0.19

Table 5.2.: Frequency Distribution of Wine Varieties

Variety	Count
Red	1,599
White	4,898

5.2. Standard Deviation Analysis

Standard deviation helps us understand the variability of each metric variable. *Residual sugar* shows the highest variability (4.76), whereas *fixed acidity* shows moderate spread (1.30).

5.3. Skewness Analysis

- Most metric variables such as *fixed acidity* (1.72), *volatile acidity* (1.49), and *residual sugar* (1.43) exhibit right-skewed distributions.
- *Chlorides* shows extreme right-skewness (5.40), likely indicating significant outliers.
- *Quality* is fairly symmetric (0.19).
- The categorical variable *variety** shows negative skewness (-1.18), indicating imbalance.

5.4. Missing Values

No missing values were found in the dataset, ensuring clean input for statistical and predictive modeling [LR19].

5.5. Summary of Part-A

Right-skewed distributions dominate the dataset, and variables like *chlorides* and *residual sugar* show significant outliers. White wine makes up 75% of the dataset.

5.6. Part-B: Visual Exploration

5.6.1. Distribution and Outlier Analysis

Metric Variables

- **Fixed**/**Volatile Acidity**: Moderate right skew, visible outliers.
- **Residual Sugar**/**Chlorides**: Highly skewed with extreme outliers.
- **Quality**/**Sulfur Dioxide**: Nearly symmetric.
- **Sulphates**/**Alcohol**: Mild skew with moderate outliers.

Categorical Variable

The *variety* barplot shows class imbalance with 75% white wine, which must be handled in supervised learning tasks.

5.6.2. Summary Table

Table 5.3.: Skewness and Outlier Assessment for Metric Variables

Variable	Skewness	Distribution Shape	Outliers
Fixed Acidity	1.72	Right-skewed	Yes
Volatile Acidity	1.49	Right-skewed	Yes
Citric Acid	0.47	Mild right-skewed	No
Residual Sugar	1.43	Right-skewed	Yes (extreme)
Chlorides	5.40	Extreme right-skew	Yes (severe)
Free Sulfur Dioxide	1.22	Right-skewed	Yes
Total Sulfur Dioxide	0.00	Symmetric	No
Density	0.50	Mild right-skewed	No
pH	0.39	Mild right-skewed	No
Sulphates	1.80	Right-skewed	Yes
Alcohol	0.57	Mild right-skewed	Yes
Quality	0.19	Symmetric	No

5.6.3. Conclusions

The dataset contains many skewed and outlier-prone variables. These characteristics suggest the need for transformations (like log-scale) or robust methods in modeling. The class imbalance in *variety* also requires caution during classification.

5.6.4. R Code for Visualizations

Listing 5.1: R Code for Generating Boxplots and Barplot

```
# Set path to save plots
save_path <- "C:/Users/DAANYAAL/Desktop/Hochschule␣
    Emden/Summer_semester_2025/Data␣
    Science/DSR/report/Images/"

# Metric variables
num_vars <- c("fixed.acidity", "volatile.acidity",
    "citric.acid", "residual.sugar", "chlorides",
"free.sulfur.dioxide", "total.sulfur.dioxide", "density",
    "pH", "sulphates",
"alcohol", "quality")

# Save boxplots
for (var in num_vars) {
      file_name <- paste0(save_path, "boxplot_", var, ".png")
```

```
        png(filename = file_name)
        boxplot(wine[[var]], main = paste("Boxplot of", var),
            ylab = var)
        dev.off()
    }

    # Barplot for 'variety'
    file_name <- paste0(save_path, "barplot_variety.png")
    png(filename = file_name)
    barplot(table(wine$variety), main = "Barplot of Variety",
        ylab = "Count")
    dev.off()
```

6. Task-2: Analysis of Alcohol Content Differences Between Red and White Wines

In this task, we aim to analyze whether the average alcohol content significantly differs between red and white wines. We performed a Welch Two Sample *t*-test in R. This test is appropriate because it does not assume equal variances between groups, which aligns with the characteristics of our data [Wel47].

6.1. R Code Snippet

The following R code was used to compute descriptive statistics and perform the *t*-test:

```
library(mosaic)
favstats(~alcohol | variety, data = wine)

t.test(alcohol ~ variety, data = wine)
```

6.2. Results

The *t*-test results are summarized in Table 6.1.

Table 6.1.: *t*-test Results for Alcohol Content by Wine Variety

Statistic	**Value**
t-value	-2.859
Degrees of Freedom (df)	3100.5
p-value	0.0043
95% Confidence Interval	[-0.154, -0.029]
Mean Alcohol Content (Red)	10.42
Mean Alcohol Content (White)	10.51

6.3. Inference

Since the p-value (0.0043) is less than the conventional significance level of 0.05, we reject the null hypothesis that the mean alcohol contents for red and white wines are equal. The negative t-value indicates that red wines have a statistically significantly lower mean alcohol content compared to white wines. Additionally, the 95% confidence interval excludes zero, further supporting this conclusion [MM15].

6.4. Assumptions Check

Before performing the t-test, we checked its assumptions:

- **Independence:** The samples of red and white wines are independent, as data were collected separately and observations from one group do not affect the other [VR02].
- **Normality:** Both groups have large sample sizes (red = 1,599, white = 4,898). According to the Central Limit Theorem, the sampling distribution of the mean approximates normality even if the underlying data distributions are not perfectly normal.
- **Equal Variances:** The sample standard deviations differ between the groups, so Welch's t-test was applied to account for unequal variances [Wel47].

7. Task-3: Red-Wine Analysis

7.1. Data Preparation

We subset the dataset to include only red wines for targeted analysis.

R Code to Subset Red Wine Data

```
redwine <- subset(wine, variety == "red")
```

7.2. Multiple Linear Regression Model

We conducted multiple linear regression to investigate whether the quality of red wine depends on its chemical and sensory properties. The dependent variable is `quality`, and the independent variables include key chemical measurements.

R Code for Multiple Linear Regression

```
linreg_red <- lm(quality ~ fixed.acidity + volatile.acidity
   + citric.acid +
residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + density + pH + sulphates + alcohol,
data = redwine)
summary(linreg_red)
```

7.3. Model Outcomes

Significant predictors (p-value < 0.05) include:

- **Negative Influence:** volatile acidity, chlorides, total sulfur dioxide, and pH.

- **Positive Influence:** free sulfur dioxide, sulphates, and alcohol.

The adjusted R^2 value is 0.3561, indicating that approximately 35.6% of the variance in wine quality is explained by the model [MPV12; Fox15].

7.4. Assumptions Check

We verified the following assumptions of linear regression, based on recommended practices [MPV12]:

- **Linearity:** Residuals appear randomly scattered, showing no clear pattern.
- **Normality:** The residual histogram and Q-Q plot suggest approximate normality.
- **Homoscedasticity:** The residuals exhibit relatively constant variance.
- **Independence:** The Durbin-Watson test statistic of 1.7571 indicates slight positive autocorrelation [DW50].

7.5. Durbin-Watson Test for Residual Autocorrelation

The Durbin-Watson test revealed some positive autocorrelation in the residuals, indicating a mild violation of the independence assumption [DW50]. However, no remedial action was taken per the scope of this task.

R Code to Perform Durbin-Watson Test

```
library(lmtest)
dwtest(linreg_red)
```

7.6. Summary and Insights

The multiple linear regression model on red wine data identifies several significant chemical predictors of quality. Key findings include:

- Alcohol, sulphates, and free sulfur dioxide positively affect quality.
- Volatile acidity, chlorides, total sulfur dioxide, and pH negatively impact quality.

The model explains 35.6% of the variance in quality ratings, indicating a moderately strong explanatory power.

All major regression assumptions (linearity, normality, homoscedasticity) were satisfied, except for a mild violation of independence suggested by the Durbin-Watson test. This issue was acknowledged but not addressed further, as permitted by the task guidelines.

These findings help pinpoint key chemical attributes linked to red wine quality, providing a foundation for targeted improvements in production and quality control.

8. Task-4: Predicting Good or Bad Wine Using Logistic Regression

8.1. Goal

In this task, we aim to classify whether a wine is **good** (quality ≥ 8) or **bad** (quality ≤ 4) based on its chemical and sensory properties. To achieve this, we implement a binary logistic regression model using predictors such as acidity, sulphates, and alcohol content.

8.2. Data Preparation

We create a new binary target variable `goodbad` with the following logic:

- 1 indicates good quality wine (quality ≥ 8),
- 0 indicates bad quality wine (quality ≤ 4),
- Entries with quality between 5 and 7 are excluded.

R Code to Create the Binary Target Variable

```
wine$goodbad <- ifelse(wine$quality >= 8, 1,
ifelse(wine$quality <= 4, 0, NA))
wine_binary <- na.omit(wine)
```

8.3. Logistic Regression Model

We run a logistic regression with `goodbad` as the dependent variable and chemical properties as predictors.

R Code to Fit Logistic Regression Model

```
logreg <- glm(goodbad ~ fixed.acidity + volatile.acidity +
   citric.acid +
residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + density + pH + sulphates + alcohol,
data = wine_binary,
family = binomial(logit))

summary(logreg)
```

8.4. Interpretation of Coefficients and Model Fit

Key Significant Variables

- **Alcohol (p = 0.00014)**: Strong positive effect — higher alcohol content increases the probability of wine being good.
- **Volatile Acidity (p = 1.49e-08)**: Significant negative effect.
- **Residual Sugar, Sulphates, Free Sulfur Dioxide**: Positive influence.
- **Total Sulfur Dioxide**: Slight negative effect.
- **pH**: Small but significant positive influence.

Model Fit Statistics

- AIC = 326.18, BIC = 375.33
- Null deviance = 610.32, Residual deviance = 302.18

8.5. Cutpoint-Based Classification and Confusion Matrix

A cutpoint of 0.5 was used to classify predicted probabilities.

R Code for Prediction and Confusion Matrix

```
cutpoint <- 0.5
wine_binary$predicted <- ((logreg$fitted.values) > cutpoint)
    * 1

cm <- xtabs(~ goodbad + predicted, data = wine_binary)
cm
```

Table 8.1.: Confusion Matrix for Logistic Regression

	Predicted 0	**Predicted 1**
Actual 0 (Bad)	215	31
Actual 1 (Good)	36	162

Performance Metrics

- **Accuracy:** 84.91%
- **Misclassification Rate:** 15.09%
- **Sensitivity (TPR):** 81.82%
- **Specificity (TNR):** 87.40%
- **False Negative Rate:** 12.60%

8.6. Model Evaluation: Lift Curve

R Code to Plot Lift Curve

```
library(ROCR)
pred <- prediction(logreg$fitted.values, wine_binary$goodbad)
plot(performance(pred, "lift", "rpp"))
```

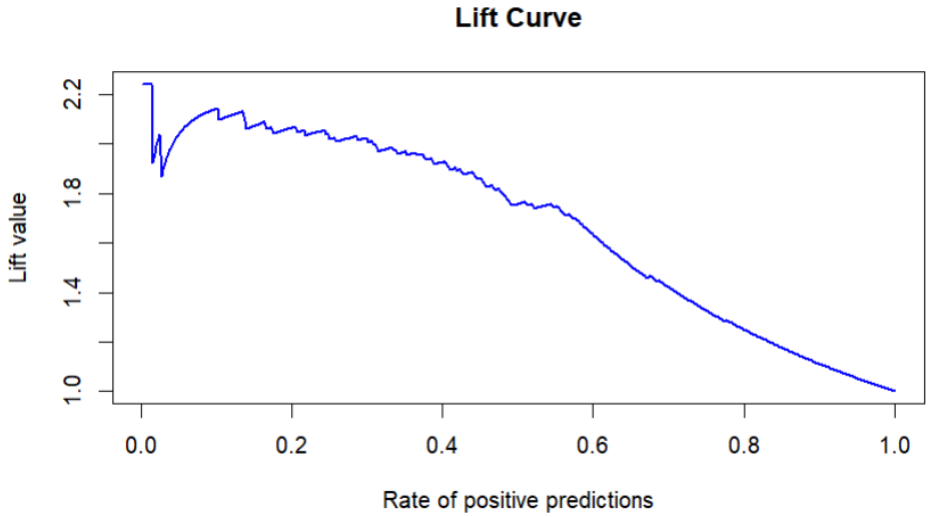

Figure 8.1.: Lift Curve

8.7. Model Evaluation: ROC Curve and AUC

R Code for ROC and AUC

```
library(pROC)
roc(wine_binary$goodbad, logreg$fitted.values,
plot = TRUE, legacy.axes = TRUE, print.auc = TRUE,
smooth = TRUE)
```

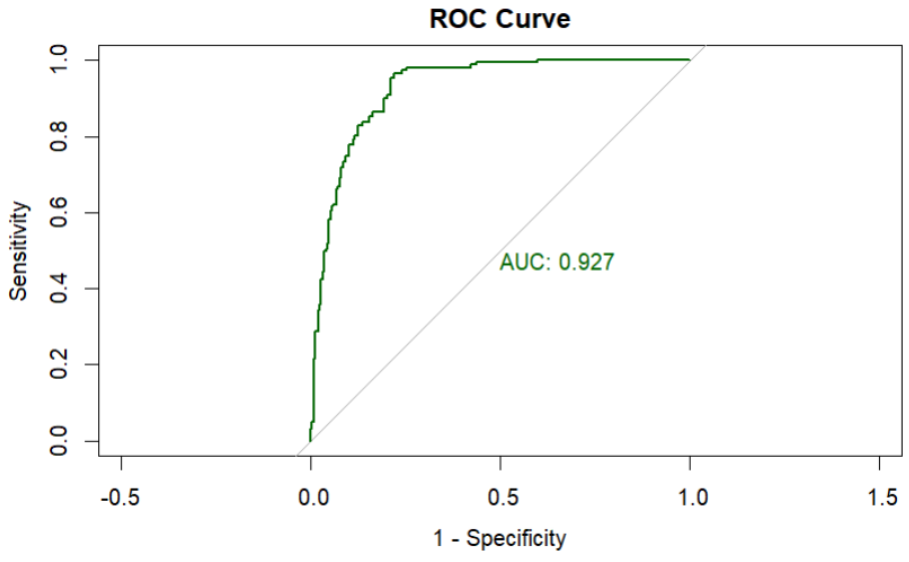

Figure 8.2.: ROC Curve and AUC

- AUC close to 1 implies excellent classification ability.

- AUC $\geq$ 0.85 is considered strong in real-world settings.

8.8. Conclusion

Logistic regression is a powerful technique for binary classification problems [MPV12; HLS13]. The model demonstrates high accuracy and good discrimination ability, as seen from the ROC and lift curves [Faw]. This model can be effectively used to distinguish good wines from bad ones using chemical and sensory measurements.

9. Task-5: Predicting Wine Variety Using Logistic Regression

In this task, we aim to build a binary classification model to predict whether a wine is red or white based on its chemical and sensory properties. This involves using a logistic regression model. Following the professor's guidelines, we split the data into training and validation sets, train the model, evaluate its predictive accuracy, and analyze its quality using a confusion matrix and AUC value.

9.1. Recoding Wine Variety to Binary

Before applying logistic regression, we transform the categorical variable `variety` into a binary numeric format to serve as the target variable. We define a new variable `foreign` where:

- `1` represents **Red Wine**
- `0` represents **White Wine**

This transformation enables logistic regression analysis, which requires numeric binary targets [HLS13].

R Code:

```
library(car) # for recode function

# Recode wine variety to binary
wine$foreign <- recode(wine$variety, "'red' = 1; 'white' =
    0")

# Check counts
table(wine$foreign)
```

Output:

Category	Count
0	4898
1	1599

Table 9.1.: Counts of categories 0 and 1

As shown:

- A total of **4898 white wine** samples were coded as `0`.
- A total of **1599 red wine** samples were coded as `1`.

9.2. Splitting Data into Training and Validation Sets

We randomly split the dataset into 70% training and 30% validation data for model development and testing. A seed is set to ensure reproducibility [HTF09].

R Code:

```
set.seed(123) # for reproducibility
sample_indices <- sample(1:nrow(wine), size = 0.7 *
   nrow(wine))
wine_train <- wine[sample_indices, ]
wine_valid <- wine[-sample_indices, ]
```

9.3. Logistic Regression to Predict Wine Type

We develop a logistic regression model on the training set using chemical and sensory features to predict wine type. Logistic regression is suitable for binary classification problems such as this [HLS13].

R Code to Perform Logistic Regression on Training Data

```
logit_model <- glm(foreign ~ fixed.acidity +
   volatile.acidity + citric.acid +
residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + density + pH + sulphates + alcohol,
data = wine_train,
family = binomial(link = "logit"))

summary(logit_model)
```

9.4. Output

Logistic Regression Output (Training Data)

```
Call:
glm(formula = foreign ~ fixed.acidity + volatile.acidity +
    citric.acid +
residual.sugar + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide +
density + pH + sulphates + alcohol, family = binomial(link =
    "logit"),
data = wine_train)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.341e+03 1.796e+02 -7.467 8.22e-14 ***
fixed.acidity 2.829e-01 2.544e-01 1.112 0.266094
volatile.acidity 7.296e+00 1.182e+00 6.172 6.74e-10 ***
citric.acid -3.238e+00 1.365e+00 -2.371 0.017719 *
residual.sugar -8.842e-01 1.139e-01 -7.764 8.23e-15 ***
chlorides 2.453e+01 4.575e+00 5.360 8.32e-08 ***
free.sulfur.dioxide 6.345e-02 1.688e-02 3.758 0.000171 ***
total.sulfur.dioxide -5.552e-02 5.921e-03 -9.376 < 2e-16 ***
density 1.330e+03 1.835e+02 7.245 4.34e-13 ***
pH 7.346e-01 1.534e+00 0.479 0.632040
sulphates 4.849e+00 1.440e+00 3.366 0.000762 ***
alcohol 1.250e+00 2.712e-01 4.610 4.02e-06 ***

Null deviance: 5065.53 on 4546 degrees of freedom
Residual deviance: 308.77 on 4535 degrees of freedom
AIC: 332.77

Number of Fisher Scoring iterations: 10
```

9.5. Prediction on Validation Set

We use the trained logistic regression model to predict wine type on the validation data. A cutoff threshold of 0.5 classifies predicted probabilities [HLS13].

R Code for Prediction

```
# Predict probabilities on validation set
wine_valid$predicted_prob <- predict(logit_model, newdata =
    wine_valid, type = "response")

# Convert probabilities to class labels using 0.5 cutoff
wine_valid$predicted_class <-
    ifelse(wine_valid$predicted_prob > 0.5, 1, 0)
```

9.6. Model Evaluation on Validation Set

We evaluate the model using the confusion matrix, accuracy, sensitivity, specificity, and AUC. These are standard binary classification metrics [Faw].

R Code for Evaluation

```
# Confusion matrix
conf_mat <- table(Actual = wine_valid$foreign, Predicted =
    wine_valid$predicted_class)
print(conf_mat)

# Accuracy, sensitivity, specificity
accuracy <- sum(diag(conf_mat)) / sum(conf_mat)
sensitivity <- conf_mat["1","1"] / sum(conf_mat["1",])
specificity <- conf_mat["0","0"] / sum(conf_mat["0",])

cat("Accuracy:", accuracy, "\n")
cat("Sensitivity:", sensitivity, "\n")
cat("Specificity:", specificity, "\n")

# AUC and ROC
library(pROC)
roc_obj <- roc(wine_valid$foreign, wine_valid$predicted_prob)
auc_value <- auc(roc_obj)
cat("AUC:", auc_value, "\n")

# Plot ROC curve
plot(roc_obj, main = "ROC␣Curve␣for␣Wine␣Variety␣Prediction")
```

9.6.1. Confusion Matrix

	Predicted	
Actual	0 (White)	1 (Red)
0 (White)	1464	2
1 (Red)	13	471

Table 9.2.: Confusion matrix of predicted vs actual wine types on validation data.

9.6.2. Performance Metrics

- **Accuracy:** 99.23%
- **Misclassification Rate:** 0.77%
- **Sensitivity (True Positive Rate):** 97.31% (correctly classifying red wines)
- **Specificity (True Negative Rate):** 99.86% (correctly classifying white wines)
- **AUC (Area Under ROC Curve):** 0.9957, indicating excellent discrimination ability.

9.6.3. ROC Curve

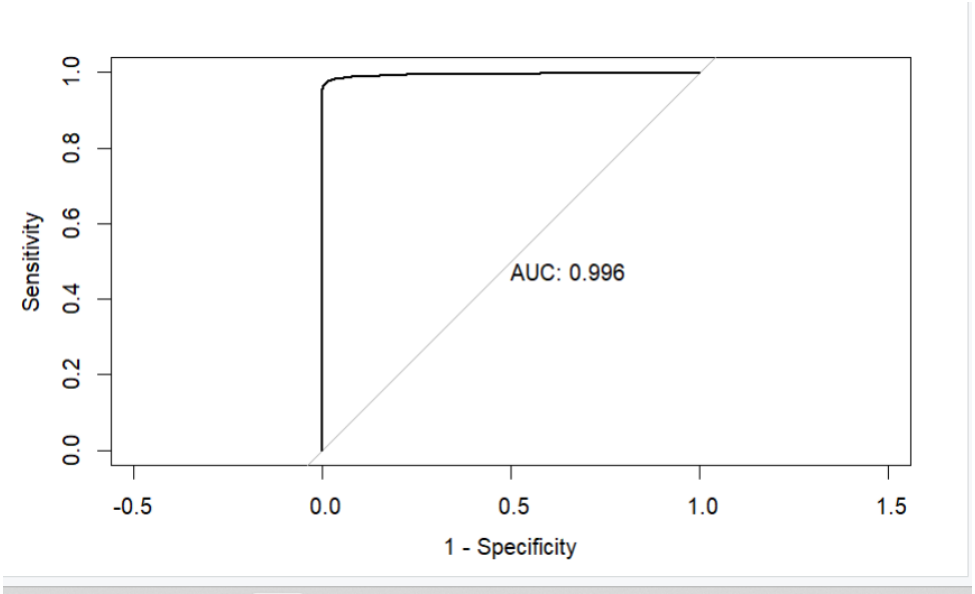

Figure 9.1.: ROC Curve

9.7. Overall Model Performance

Overall, the logistic regression model demonstrates outstanding predictive accuracy, successfully differentiating wine types based on chemical and sensory features.

10. Task-6: Factor Analysis on Wine Data

10.1. Plan

Here, we investigate whether the chemical and sensory properties of the wine dataset can be condensed into fewer factors using factor analysis. Preliminary tests were conducted to assess the suitability of the data for this method following best practices in exploratory factor analysis [FW99; CO05].

10.2. Initial Diagnostic Results

The following diagnostic tests were performed on the selected chemical variables:

10.2.1. Kaiser-Meyer-Olkin (KMO) Measure

The overall KMO measure of sampling adequacy was computed to assess the proportion of variance among variables that might be common variance suitable for factor analysis [Kai74].

```
KMOS(pca)$KMO
[1] 0.4052524
```

This value is below the commonly accepted threshold of 0.5, indicating that the dataset is currently not suitable for factor analysis.

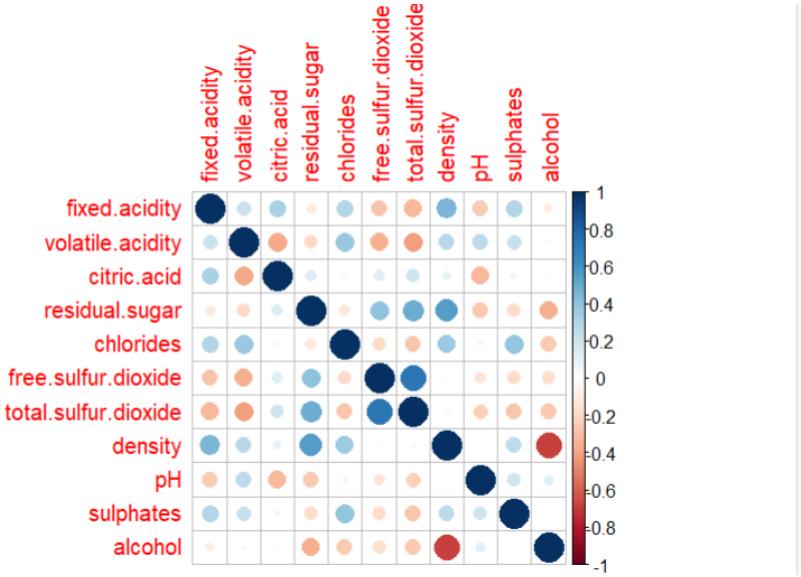

Figure 10.1.: Heatmap of Correlations among Variables

10.2.2. Bartlett's Test of Sphericity

Bartlett's test evaluates whether the correlation matrix differs significantly from an identity matrix, i.e., whether variables are sufficiently correlated to warrant factor analysis [FW99].

```
cortest.bartlett(cor(pca), n = nrow(pca))
$chisq
[1] 37094.62

$p.value
[1] 0

$df
[1] 55
```

The test is highly significant ($p < 0.05$), suggesting that correlations exist among variables.

10.2.3. Measure of Sampling Adequacy (MSA) per Variable

The MSA for individual variables is given below:

```
KMOS(pca)$MSA
fixed.acidity volatile.acidity citric.acid residual.sugar
0.2769765 0.6129595 0.6223735 0.2936906
chlorides free.sulfur.dioxide total.sulfur.dioxide density
0.7292531 0.7507372 0.7181935 0.3047922
pH sulphates alcohol
0.2088097 0.5638792 0.2730014
```

Variables with MSA values below 0.5 (highlighted above) indicate insufficient shared variance and are considered unsuitable for factor analysis [Kai74].

10.2.4. Eigenvalues of the Correlation Matrix

Eigenvalues were computed to determine the number of factors to extract based on the Kaiser criterion (eigenvalues > 1) [Kai60]:

```
eigen(cor(pca))$values
[1] 3.0298686 2.4938260 1.5563470 0.9705521 0.7198749
   0.6073117 0.5231588 0.5015103 0.3370240
[10] 0.2276958 0.0328308
```

This suggests the possibility of extracting three factors.

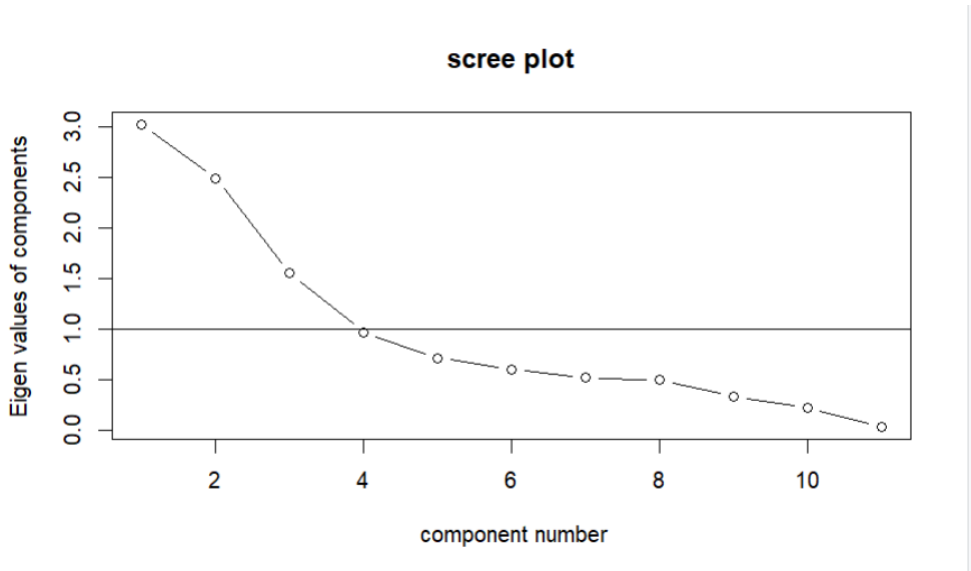

Figure 10.2.: Scree Plot Indicating Number of Factors

10.2.5. Recommended Next Steps

To improve the suitability of the data for factor analysis, variables with low MSA values should be removed. These variables are:

- `fixed.acidity`
- `residual.sugar`
- `density`
- `pH`
- `alcohol`

The remaining variables to retain for subsequent analysis are:

`volatile.acidity`, `citric.acid`, `chlorides`, `free.sulfur.dioxide`, `total.sulfur.`

The KMO test, Bartlett's test, and eigenvalue analysis should be rerun on this reduced dataset to confirm improved suitability before proceeding with factor extraction.

```
pca2 <- pca[, c("volatile.acidity", "citric.acid",
   "chlorides",
"free.sulfur.dioxide", "total.sulfur.dioxide", "sulphates")]

KMOS(pca2)$KMO
cortest.bartlett(cor(pca2), n = nrow(pca2))
KMOS(pca2)$MSA
eigen(cor(pca2))$values
VSS.scree(pca2)
```

10.3. Factor Analysis Results on Reduced Wine Data

10.3.1. Reassessment of Sampling Adequacy

After removing variables with low sampling adequacy, the following tests were repeated on the reduced dataset containing six variables:

Table 10.1.: Variables Retained for Factor Analysis

Variable 1	`volatile.acidity`
Variable 2	`citric.acid`
Variable 3	`chlorides`
Variable 4	`free.sulfur.dioxide`
Variable 5	`total.sulfur.dioxide`
Variable 6	`sulphates`

10.3.2. Kaiser-Meyer-Olkin (KMO) Measure

The overall KMO measure improved substantially:

```
KMOS(pca2)$KMO
[1] 0.6358322
```

This value exceeds the acceptable threshold of 0.6, indicating the dataset is now suitable for factor analysis.

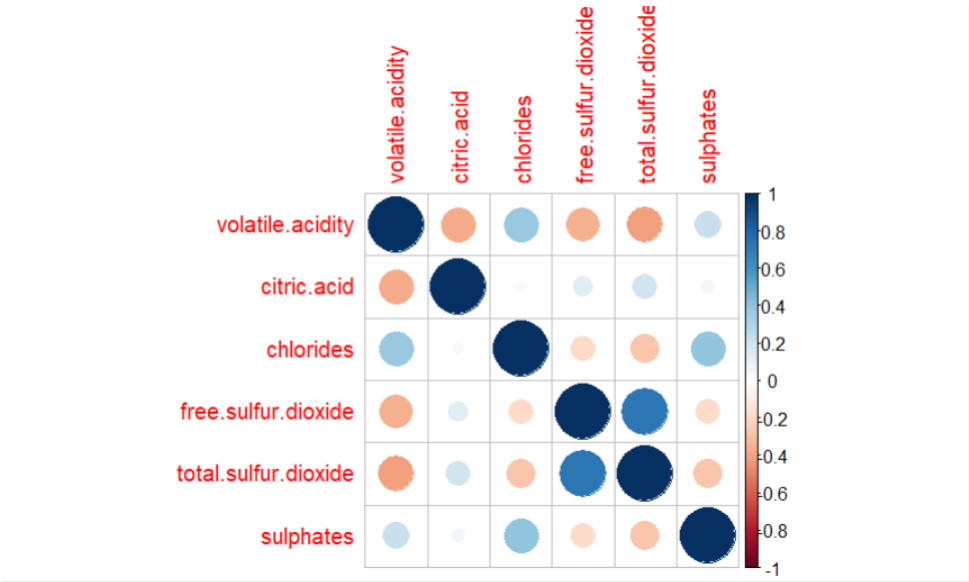

Figure 10.3.: Heatmap of Correlations among Reduced Variables

10.3.3. Bartlett's Test of Sphericity

```
cortest.bartlett(cor(pca2), n = nrow(pca2))
$chisq
[1] 9955.5
```

```
$p.value
[1] 0

$df
[1] 15
```

The significant p-value confirms that correlations between variables are sufficient for factor analysis.

10.3.4. Measure of Sampling Adequacy (MSA) per Variable

```
KMOS(pca2)$MSA
volatile.acidity citric.acid chlorides free.sulfur.dioxide
0.6884612 0.4926220 0.6403947 0.6124060
total.sulfur.dioxide sulphates
0.6348484 0.7191268
```

While `citric.acid` has a slightly lower MSA, the overall sampling adequacy supports proceeding with factor analysis.

10.4. Factor Extraction and Loadings

10.4.1. Eigenvalues

```
eigen(cor(pca2))$values
[1] 2.4641413 1.2377603 0.9647325 0.6302888 0.4363861
    0.2666910
```

According to the Kaiser criterion, two factors with eigenvalues greater than 1 should be extracted.

Figure 10.4.: Scree Plot of Reduced Dataset

10.4.2. Factor Loadings

```
result2 <- principal(pca2, nfactors = 2, rotate = "varimax")
print(result2, cut = 0.4, sort = FALSE, digits = 2)
```

10.5. Interpretation of Results

The factor analysis extracted two factors explaining 62% of the total variance. Factor loadings indicate:

- Factor 1 (RC1) is strongly associated with `volatile.acidity`, `free.sulfur.dioxide`, and `total.sulfur.dioxide`.
- Factor 2 (RC2) is strongly associated with `chlorides`, `sulphates`, and `citric.acid`.

The root mean square residual and chi-square test indicate good model fit.

10.5.1. Summary

The chemical and sensory properties of the wines can be effectively summarized by two underlying factors. This dimensionality reduction simplifies subsequent analyses while retaining most of the original information.

A. Appendix: Graphical Analysis

A.1. Histograms of Metric Variables

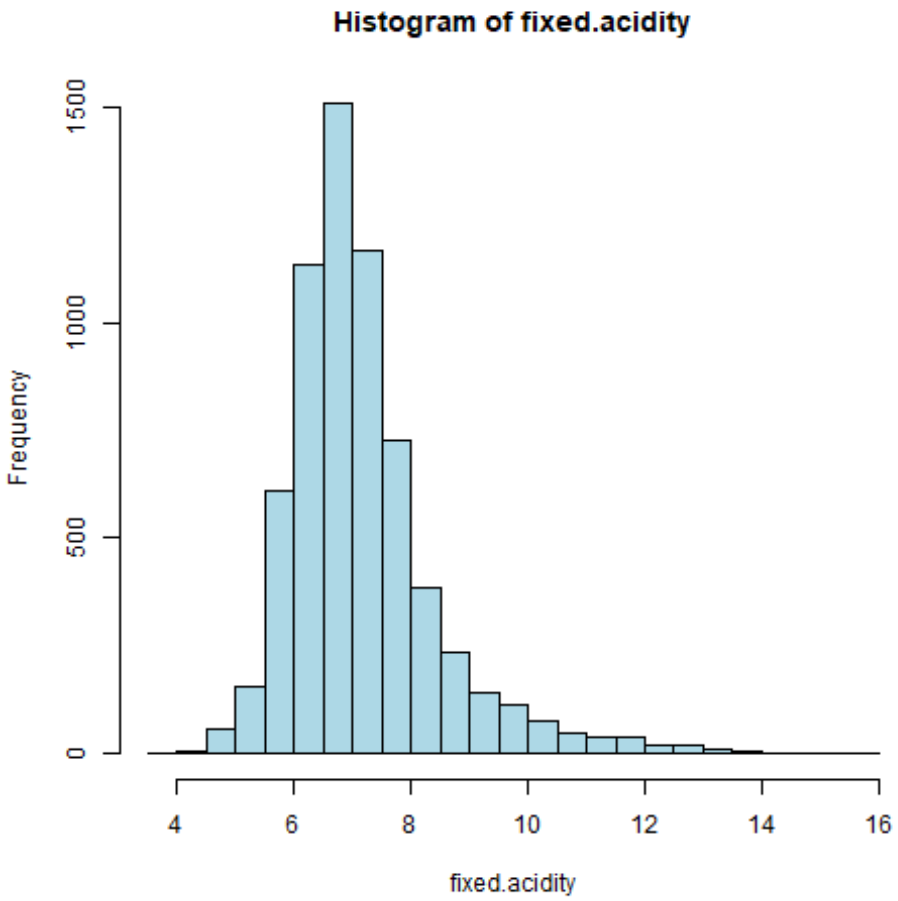

Figure A.1.: Histogram of Fixed Acidity

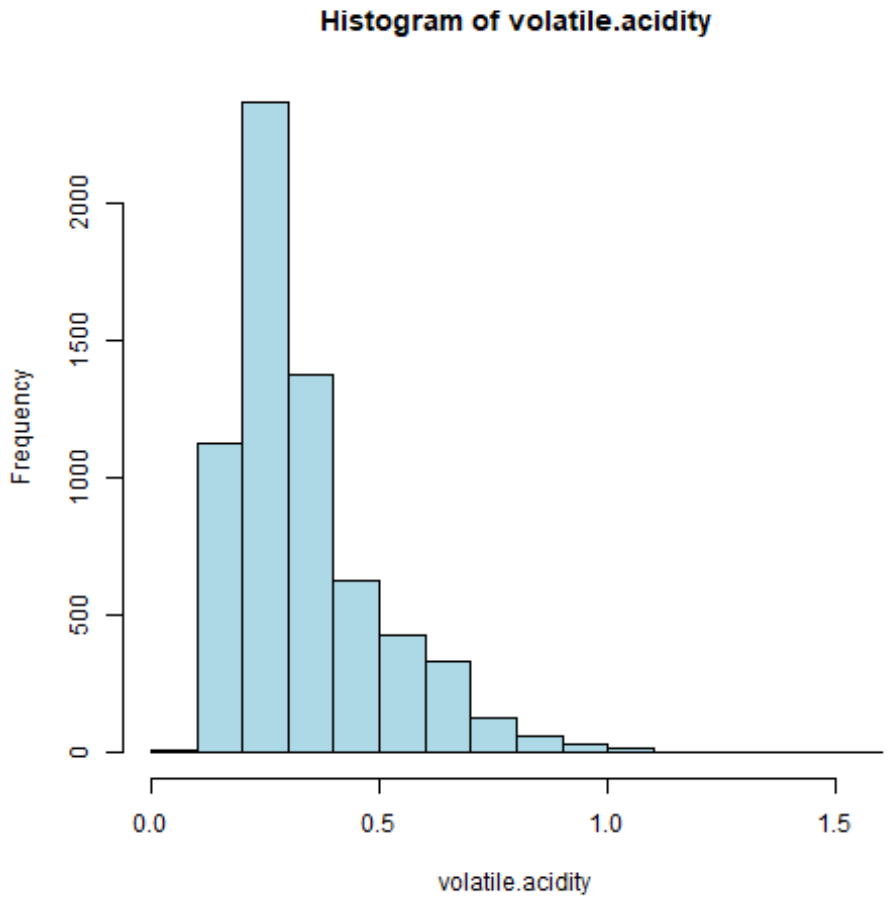

Figure A.2.: Histogram of Volatile Acidity

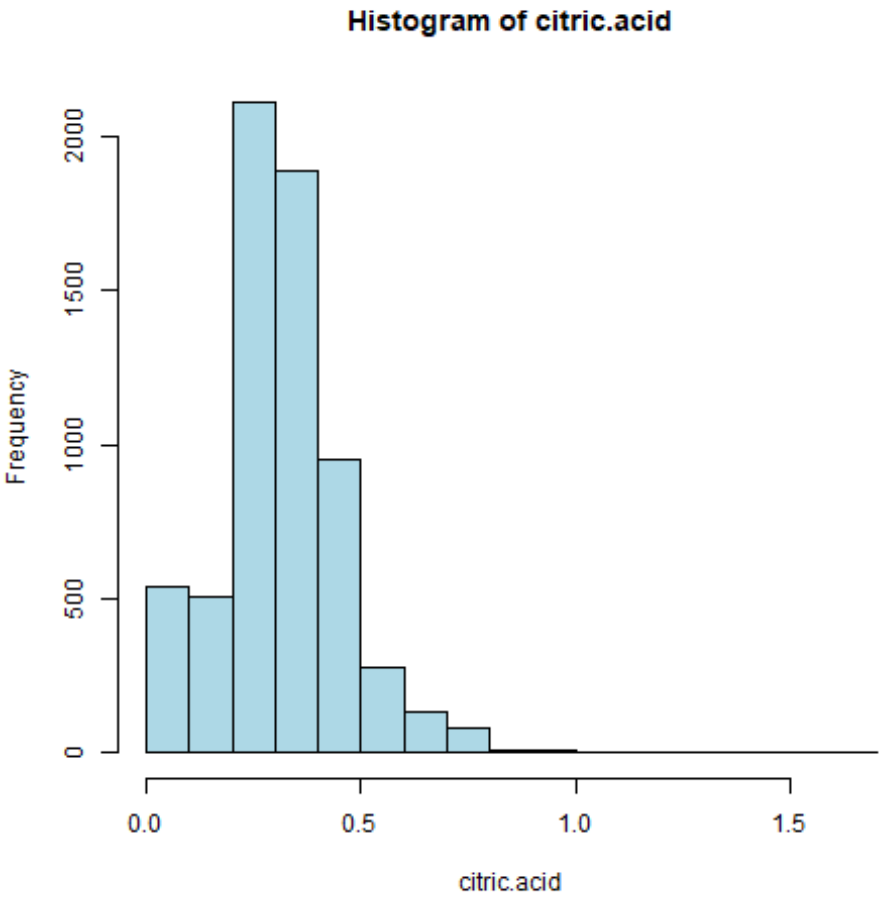

Figure A.3.: Histogram of Citric Acid

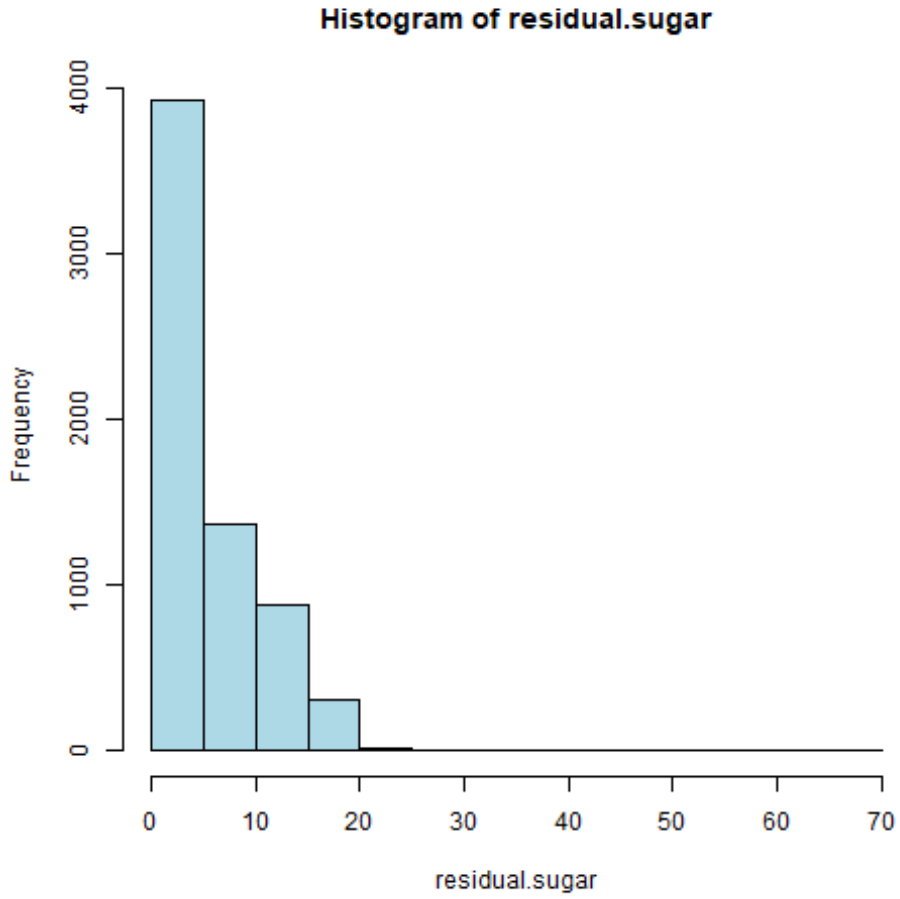

Figure A.4.: Histogram of Residual Sugar

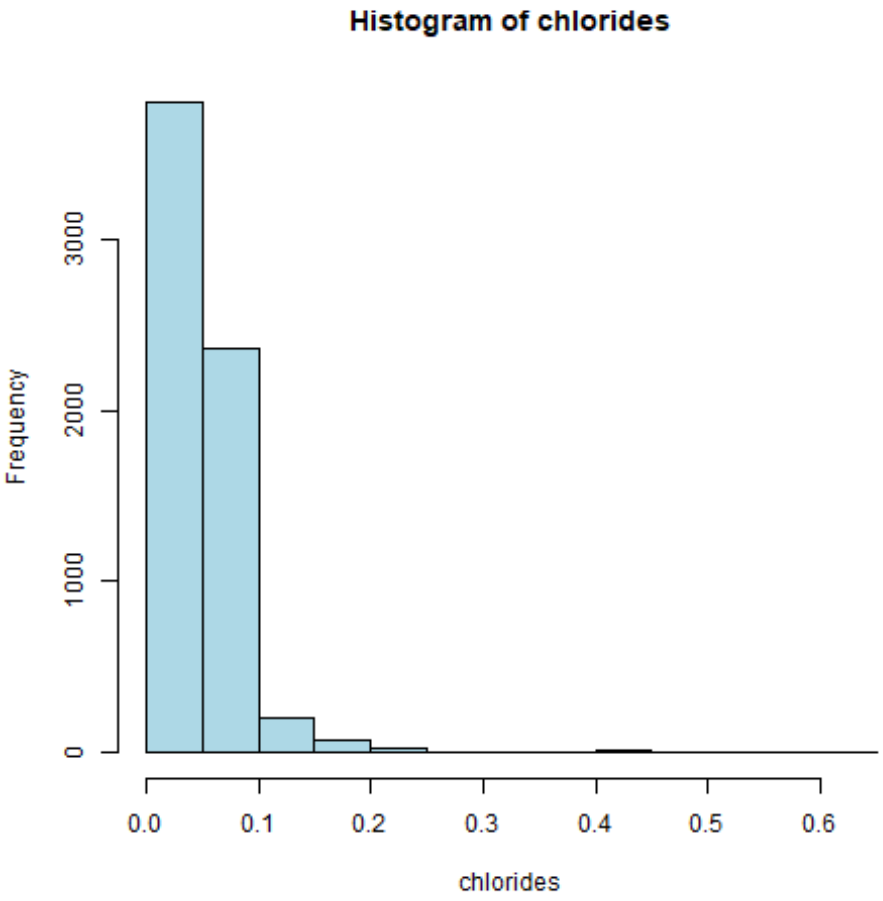

Figure A.5.: Histogram of Chlorides

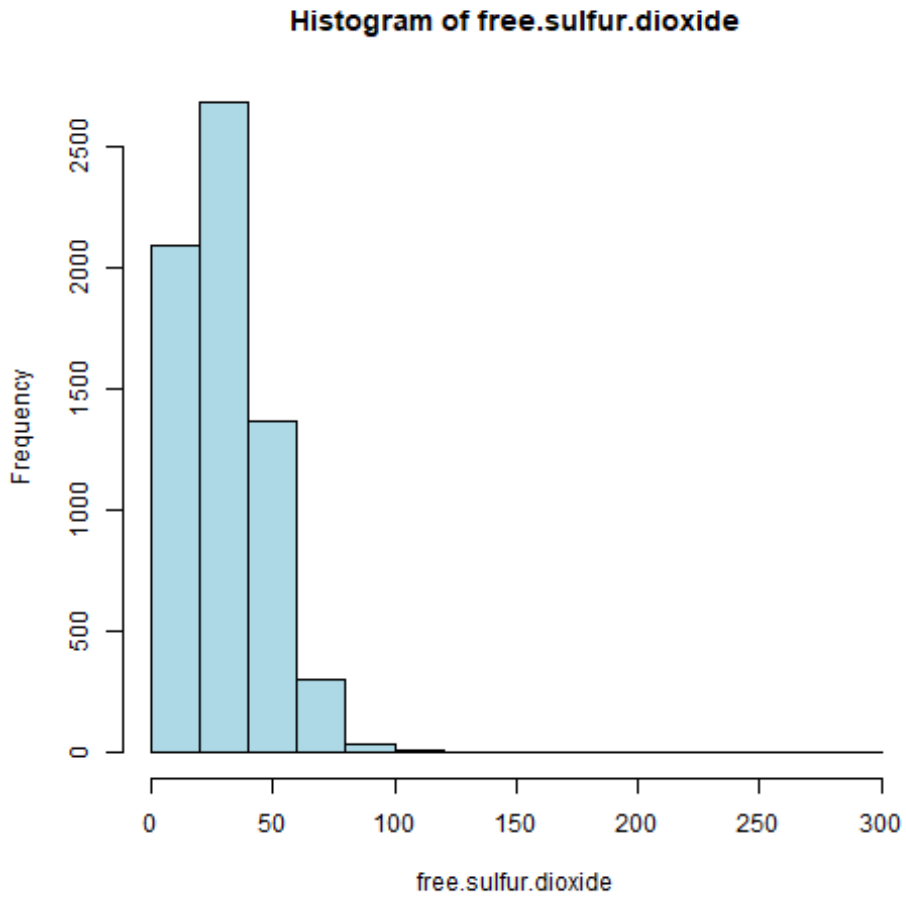

Figure A.6.: Histogram of Free Sulfur Dioxide

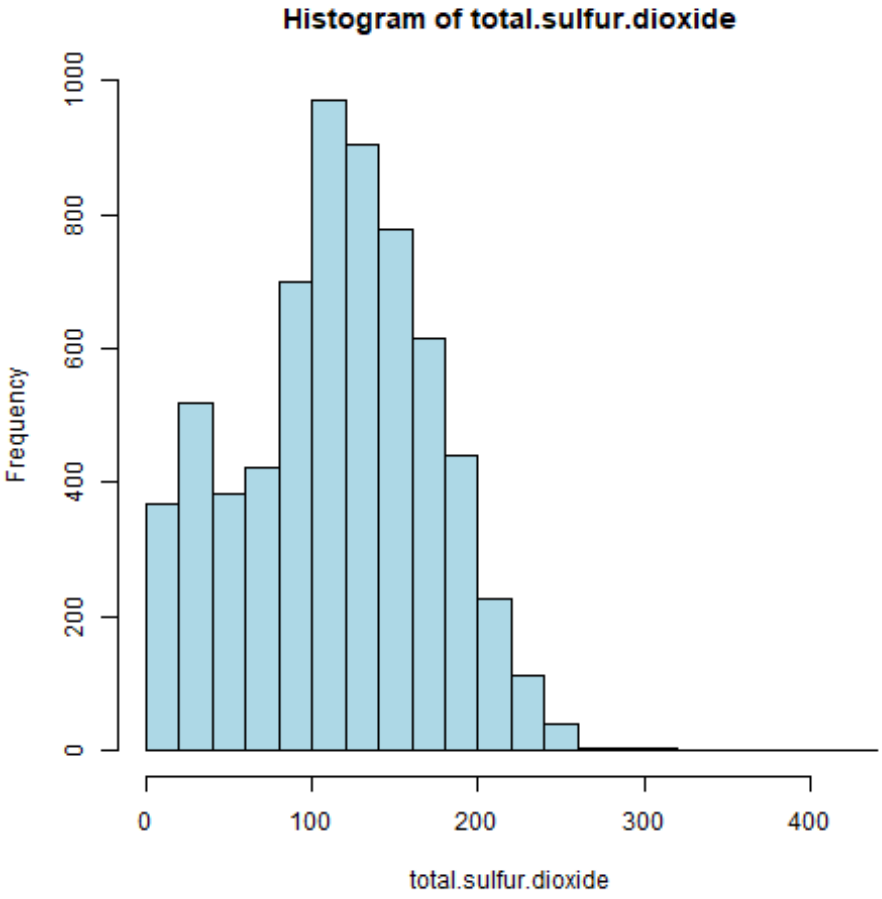

Figure A.7.: Histogram of Total Sulfur Dioxide

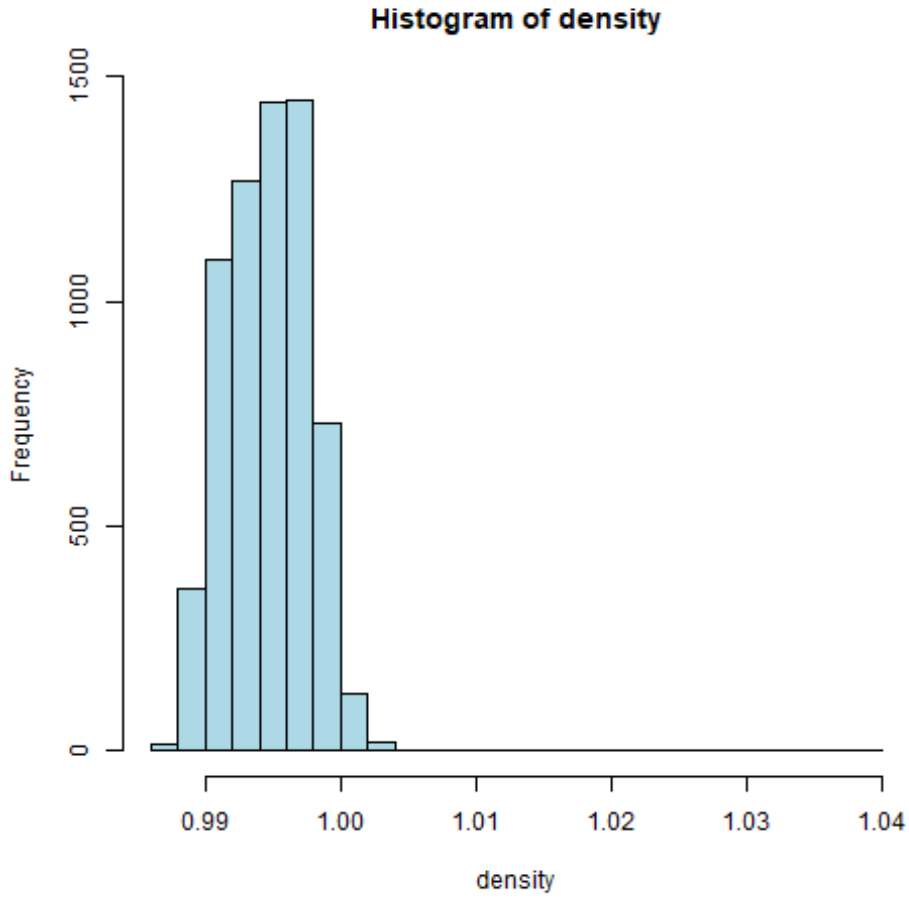

Figure A.8.: Histogram of Density

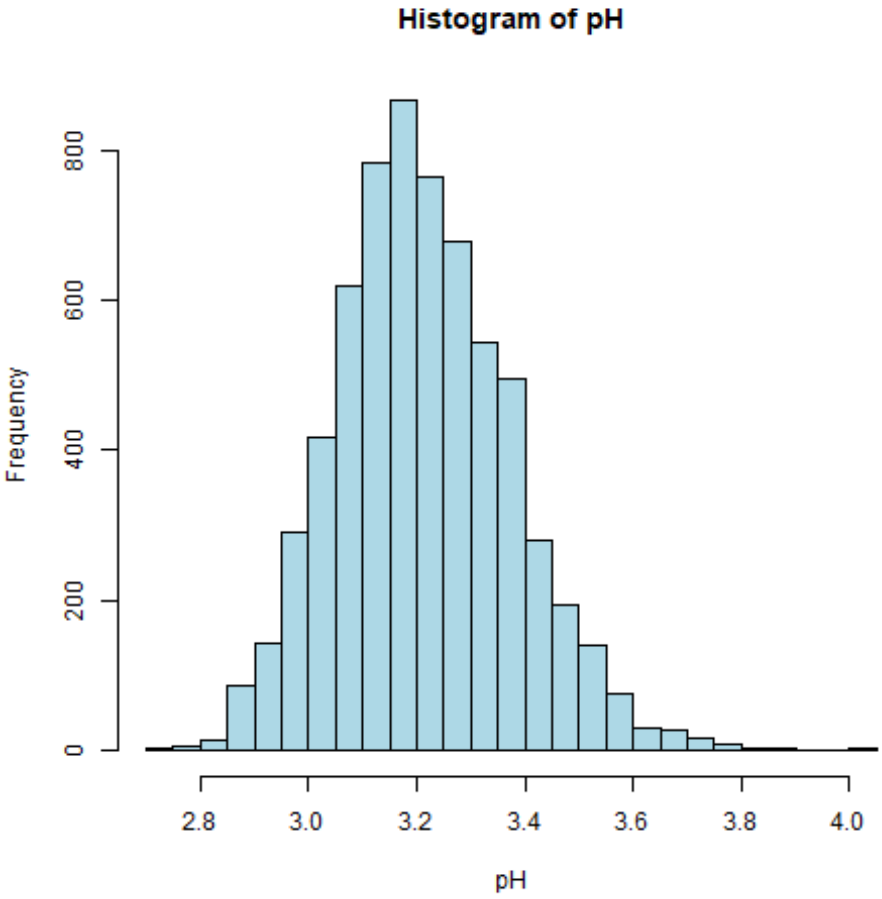

Figure A.9.: Histogram of pH

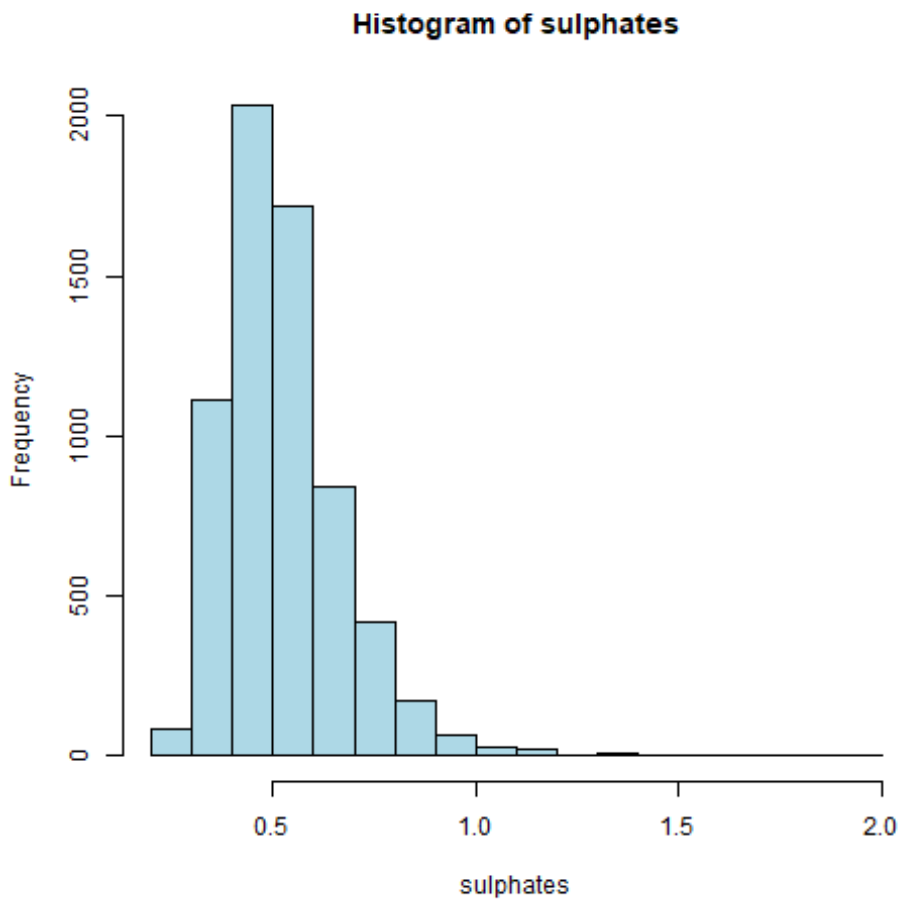

Figure A.10.: Histogram of Sulphates

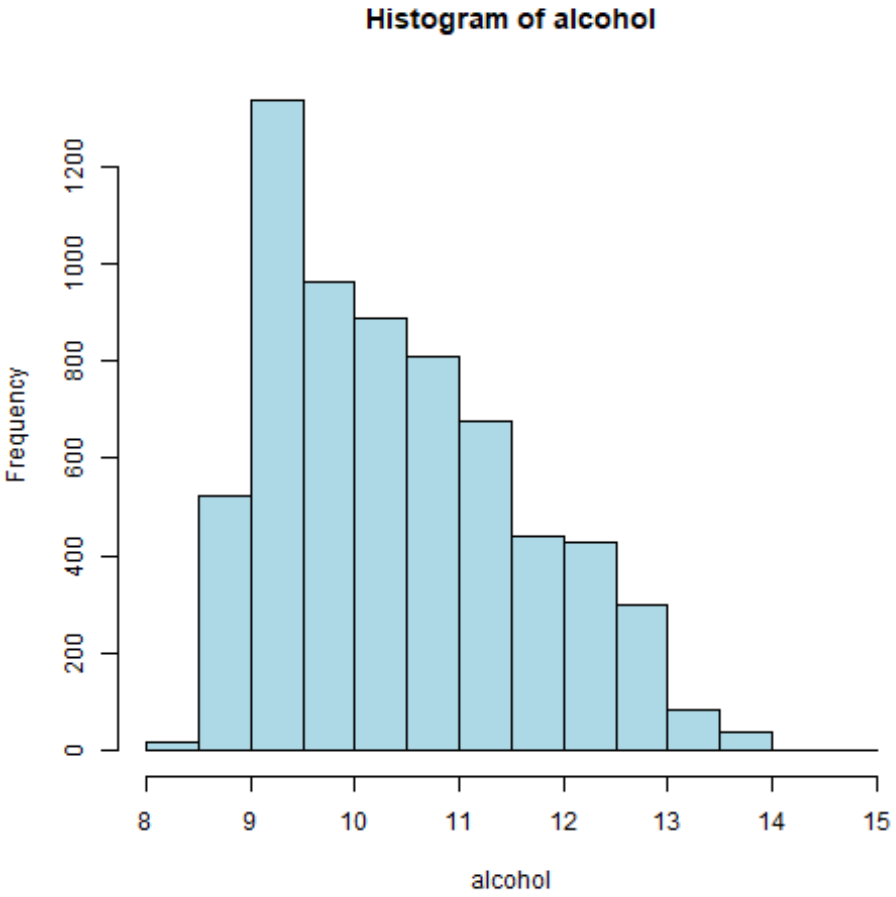

Figure A.11.: Histogram of Alcohol

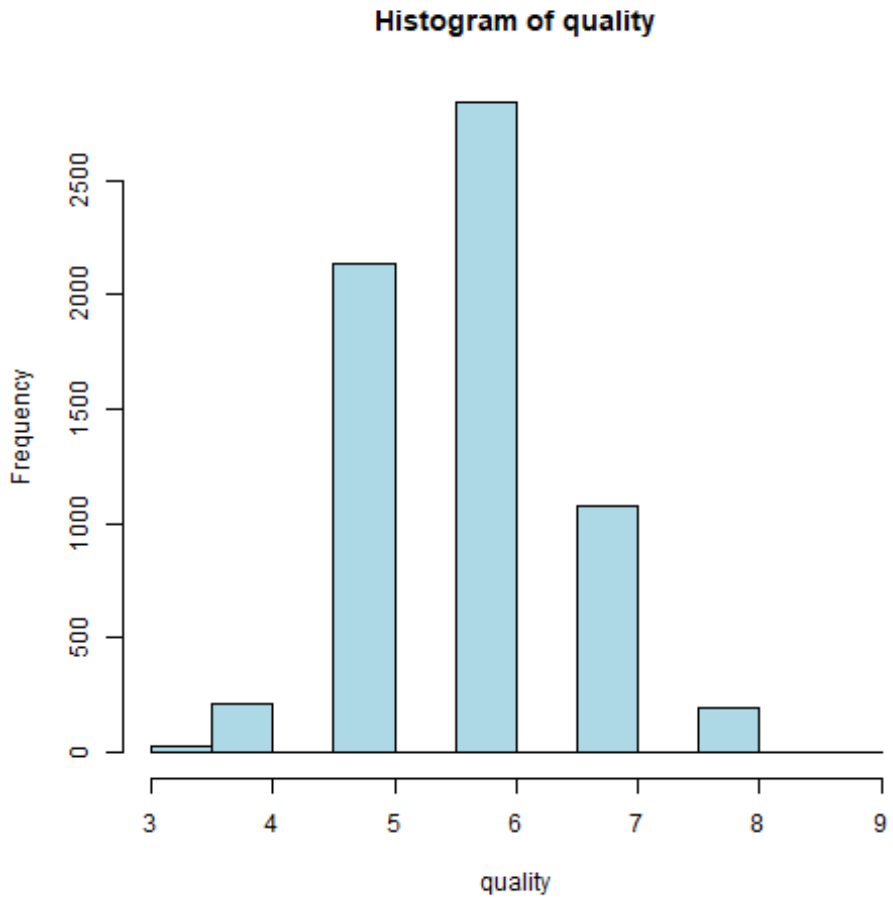

Figure A.12.: Histogram of Quality

A.2. Boxplots of Metric Variables

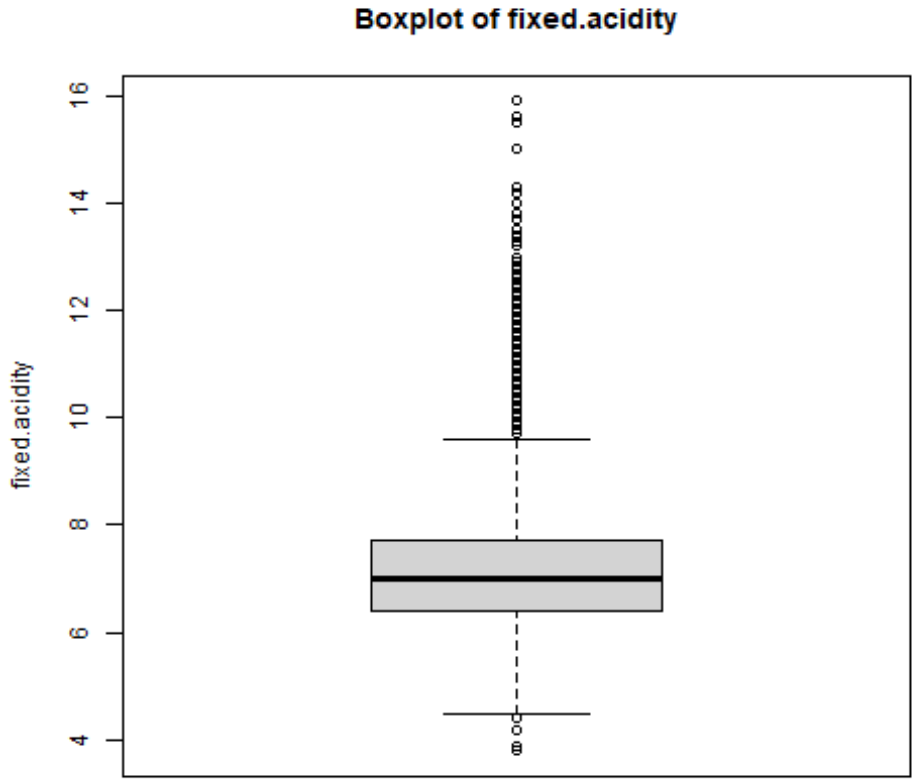

Figure A.13.: Boxplot of Fixed Acidity

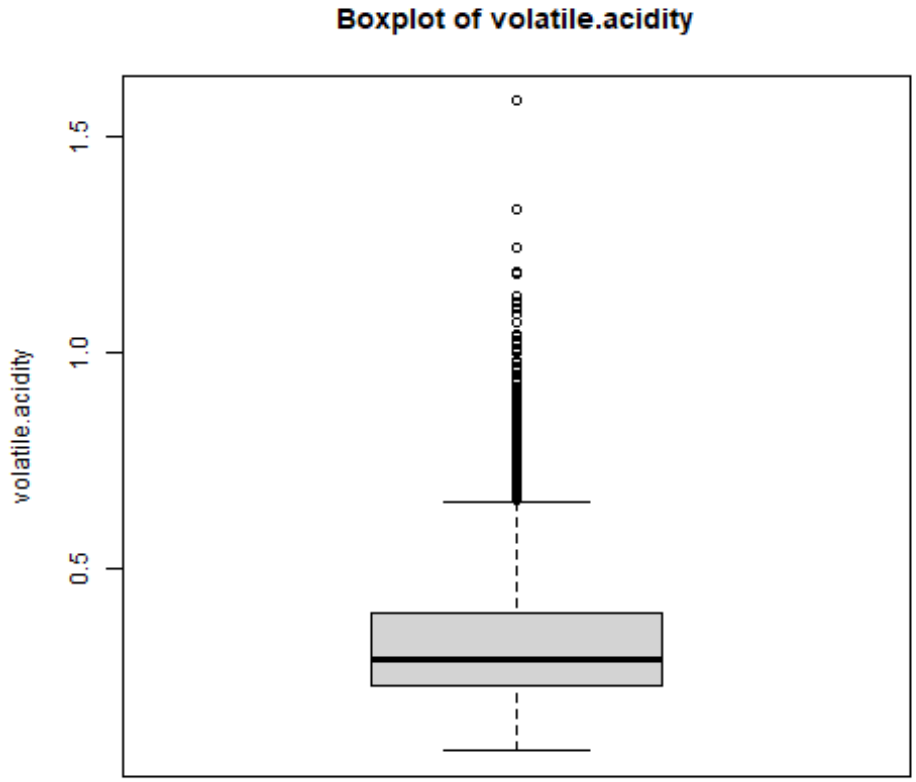

Figure A.14.: Boxplot of Volatile Acidity

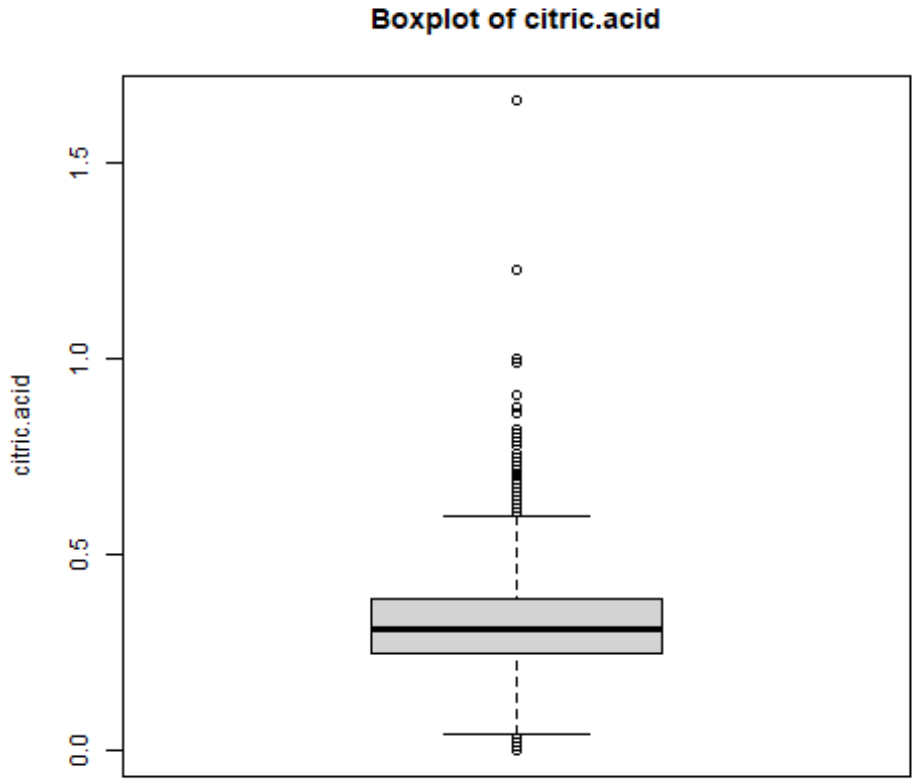

Figure A.15.: Boxplot of Citric Acid

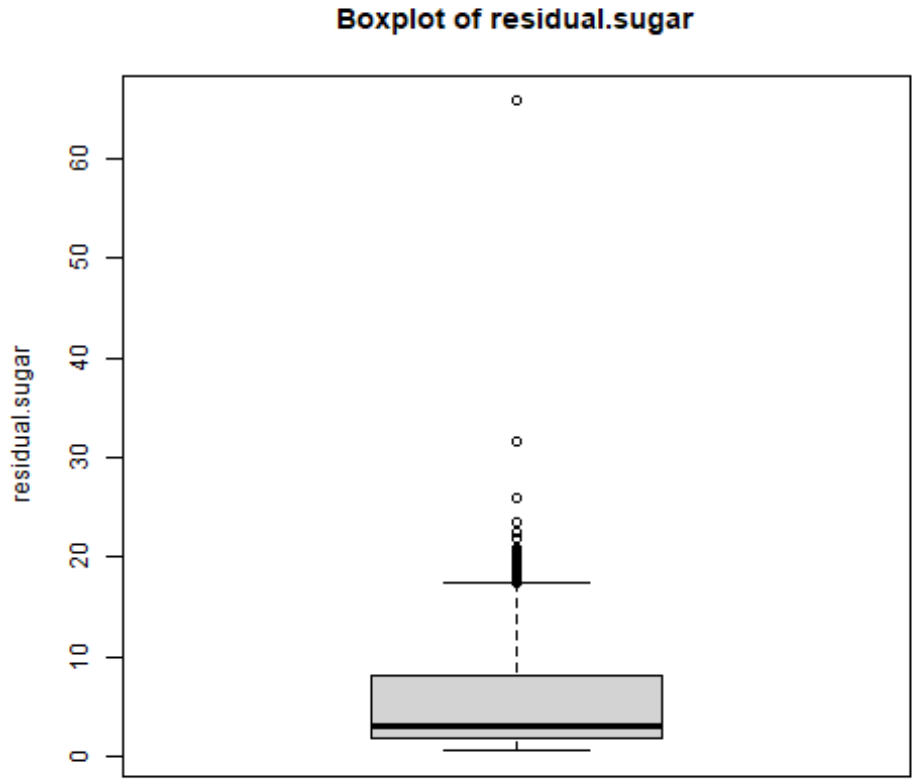

Figure A.16.: Boxplot of Residual Sugar

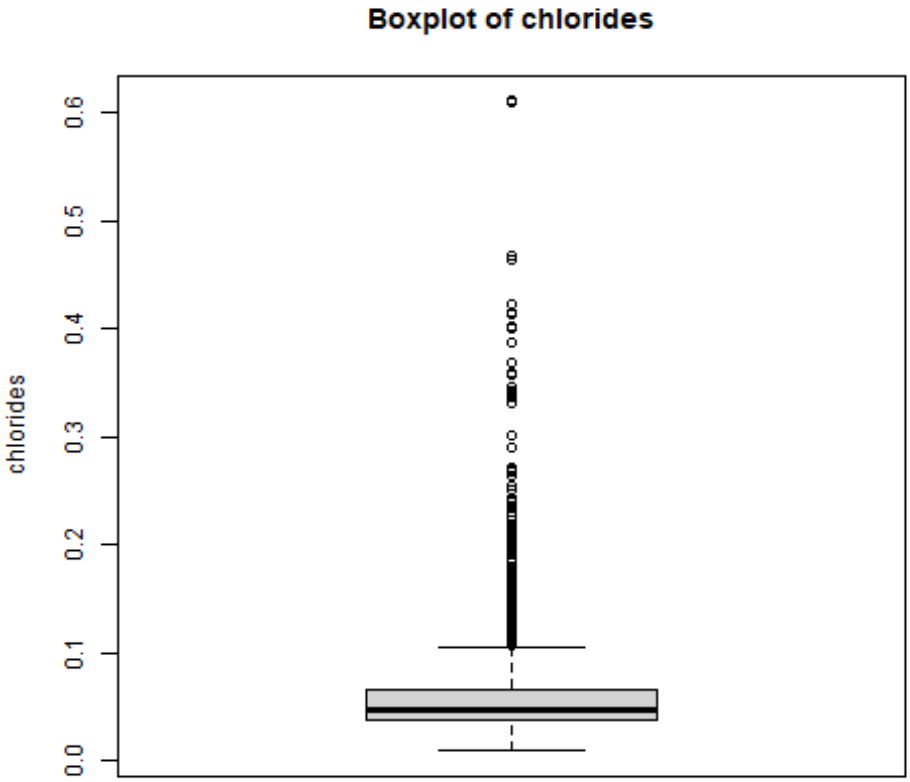

Figure A.17.: Boxplot of Chlorides

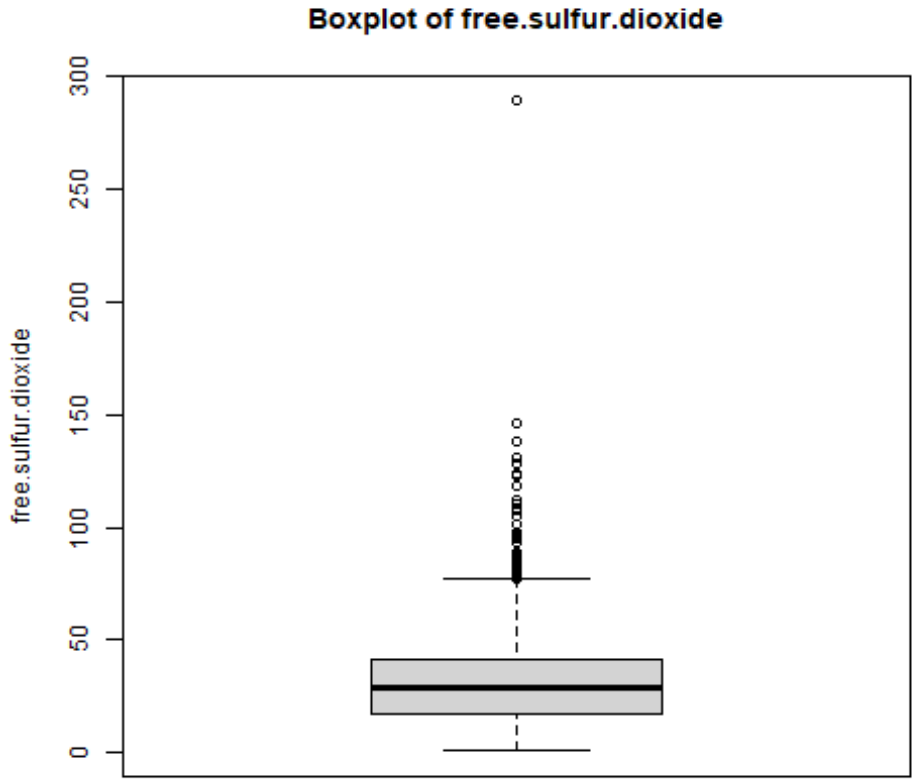

Figure A.18.: Boxplot of Free Sulfur Dioxide

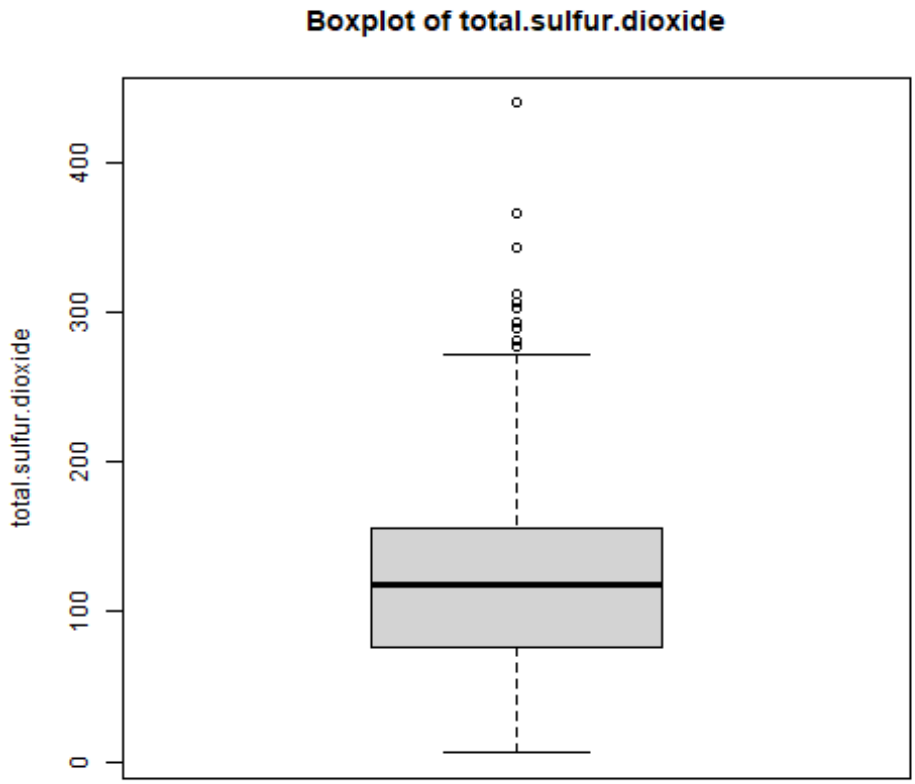

Figure A.19.: Boxplot of Total Sulfur Dioxide

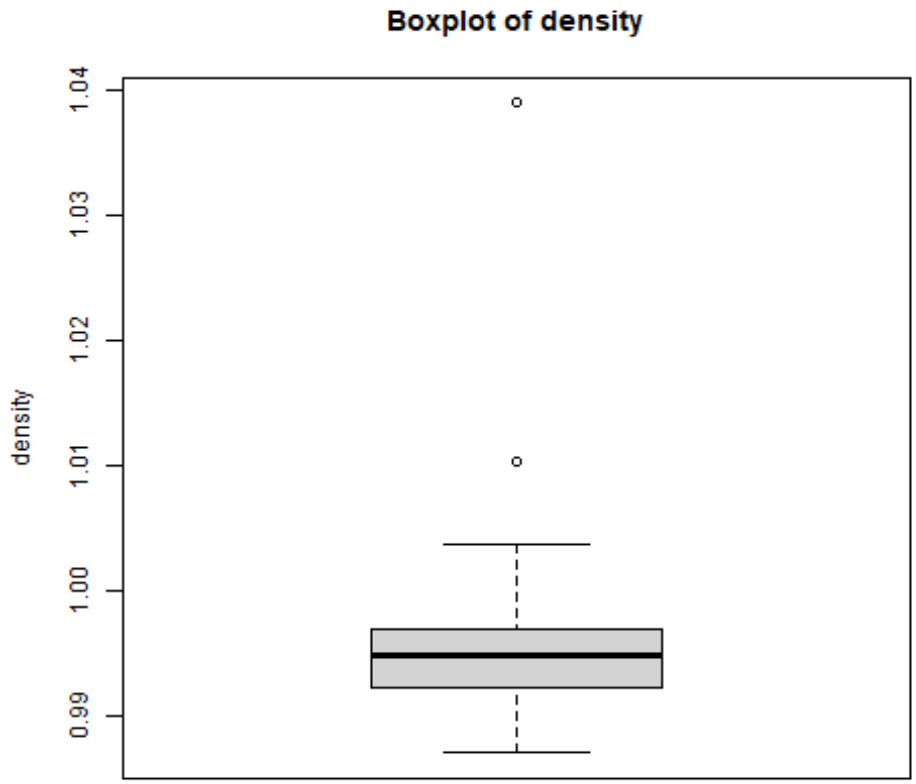

Figure A.20.: Boxplot of Density

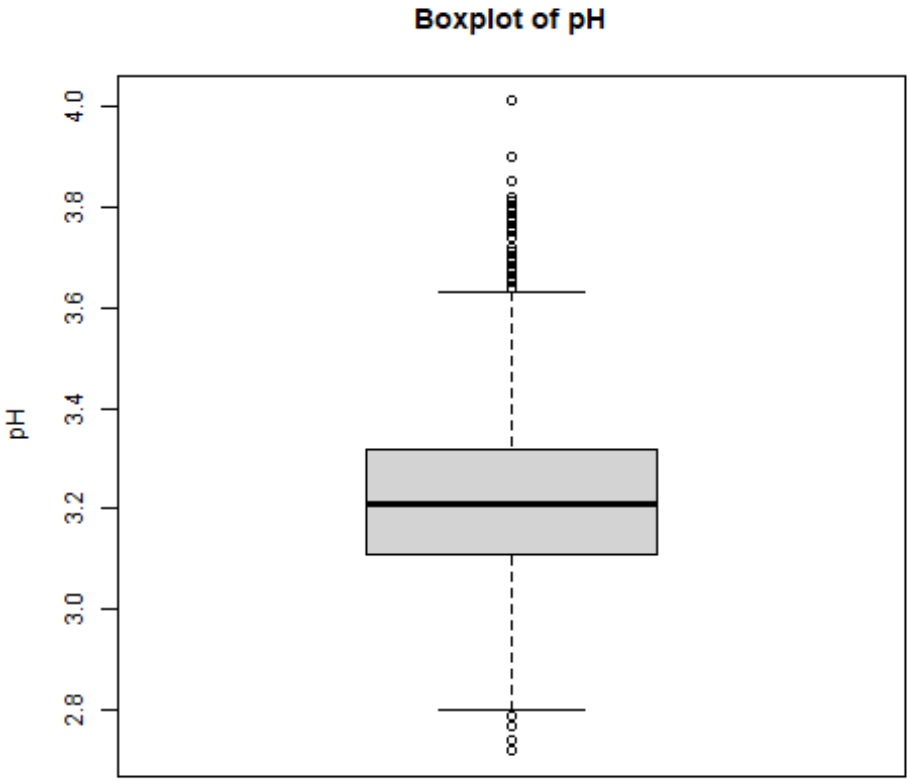

Figure A.21.: Boxplot of pH

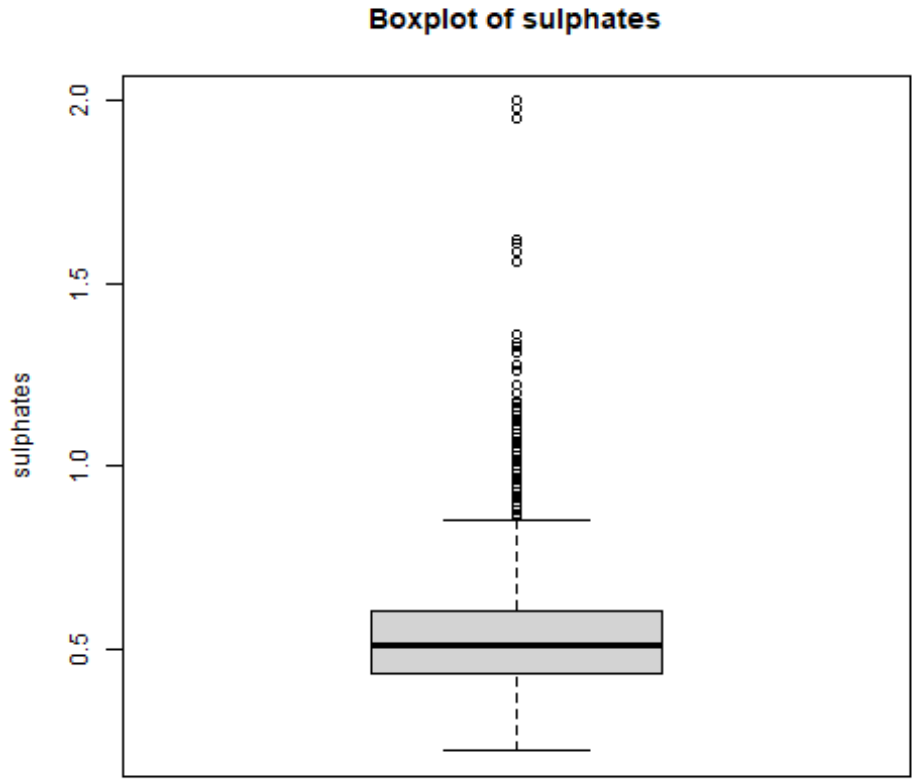

Figure A.22.: Boxplot of Sulphates

Figure A.23.: Boxplot of Alcohol

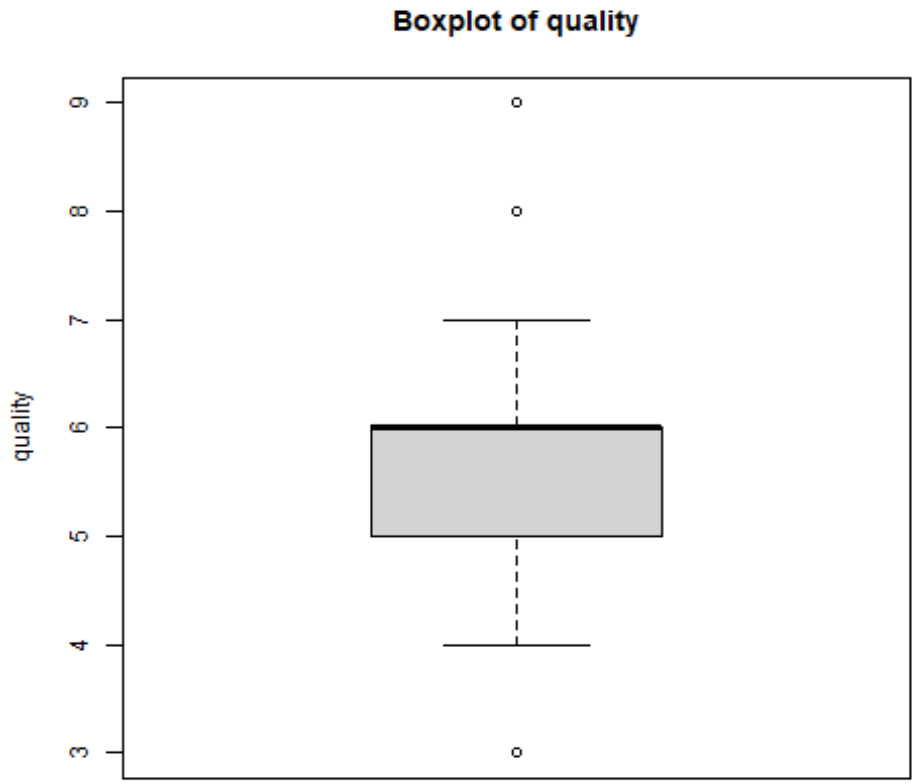

Figure A.24.: Boxplot of Quality

A.3. Barplots of Categorical Variables

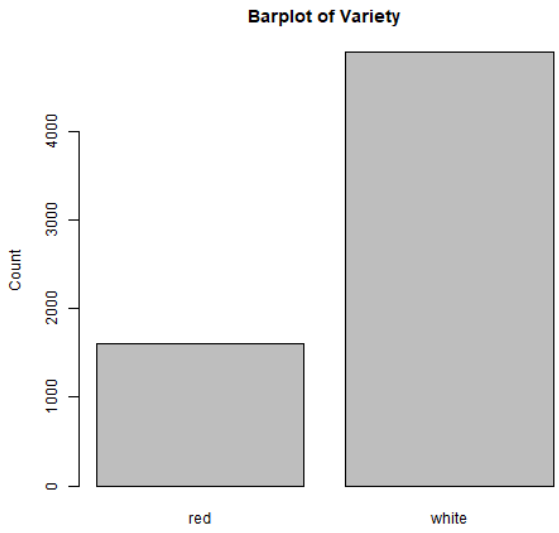

Figure A.25.: Barplot of Wine Variety

A. R Code Used in the Analysis

Listing A.1: R code for data exploration, visualization, and regression analysis

```
# Load required libraries
library(psych)
library(mosaic)
```

```
# Descriptive statistics for all variables
describe(wine)

# Standard deviation for residual sugar (example)
sd(wine$residual.sugar)

# Skewness for all variables is included in describe(wine)
```

```
# Save histograms for all numeric variables
path <- "C:/Users/DAANYAAL/Desktop/Hochschule␣
   Emden/Summer_semester_2025/Data␣
   Science/DSR/report/Images/"
num_cols <- sapply(wine, is.numeric)
num_vars <- names(wine)[num_cols]
for (var in num_vars) {
      filename <- paste0(path, "hist_", var, ".png")
      png(filename = filename)
      hist(wine[[var]], main = paste("Histogram␣of", var),
         xlab = var, col = "lightblue", breaks = 20)
      dev.off()
}
```

```
# Save boxplots for all numeric variables
num_vars <- c("fixed.acidity", "volatile.acidity",
   "citric.acid", "residual.sugar", "chlorides",
"free.sulfur.dioxide", "total.sulfur.dioxide", "density",
   "pH", "sulphates",
"alcohol", "quality")
for (var in num_vars) {
      file_name <- paste0(path, "boxplot_", var, ".png")
      png(filename = file_name)
      boxplot(wine[[var]], main = paste("Boxplot of", var),
         ylab = var)
      dev.off()
}
```

```
# Barplot for categorical variable 'variety'
file_name <- paste0(path, "barplot_variety.png")
png(filename = file_name)
barplot(table(wine$variety), main = "Barplot of Variety",
   ylab = "Count")
dev.off()
```

```
# Frequency table for 'variety'
table(wine$variety)

# Mean, SD, and quantiles of alcohol content by wine variety
favstats(~alcohol | variety, data = wine)
```

```
# Multiple linear regression for all wine
wine$variety <- as.factor(wine$variety)
linreg <- lm(quality ~ alcohol + sulphates + variety,
   data=wine)
summary(linreg)
```

```
# Multiple linear regression for red wine only
redwine <- subset(wine, variety == "red")
```

```
linreg_red <- lm(quality ~ fixed.acidity + volatile.acidity
   + citric.acid + residual.sugar + chlorides +
free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
   sulphates + alcohol,
data=redwine)
summary(linreg_red)
```

```
# Residual diagnostics for red wine regression
plot(resid(linreg_red))
abline(h=0, col="red", lwd=2)
hist(resid(linreg_red))
qqnorm(resid(linreg_red))
qqline(resid(linreg_red))
plot(linreg_red$fitted.values, resid(linreg_red),
xlab = "Fitted Values", ylab = "Residuals",
main = "Homoscedasticity Check")
abline(h = 0, col = "red", lwd = 2)
```

```
# Durbin-Watson test for autocorrelation
library(lmtest)
dwtest(linreg_red)
```

```
# Logistic regression: Predicting 'good' vs 'bad' wine
wine$goodbad <- ifelse(wine$quality >= 8, 1,
ifelse(wine$quality <= 4, 0, NA))
wine_binary <- na.omit(wine)
logreg <- glm(goodbad ~ fixed.acidity + volatile.acidity +
   citric.acid +
residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + density + pH + sulphates + alcohol,
data = wine_binary,
family = binomial(logit))
summary(logreg)
AIC(logreg)
BIC(logreg)
```

```
# Confusion matrix for logistic regression
cutpoint <- 0.5
wine_binary$predicted <- ((logreg$fitted.values) > cutpoint)
    * 1
cm <- xtabs(~ goodbad + predicted, data = wine_binary)
cm

# Misclassification rate
misclassification_rate <- (cm[2] + cm[3]) / sum(cm)
misclassification_rate
```

References

[CO05] A. B. Costello and J. W. Osborne. “Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis”. In: *Practical Assessment, Research, and Evaluation* 10.7 (2005). URL: https://doi.org/10.7275/jyj1-4868.

[Cor+09] P. Cortez et al. “Modeling wine preferences by data mining from physicochemical properties”. In: *Decision Support Systems* 47.4 (2009), pp. 547–553.

[DW50] J. Durbin and G. S. Watson. “Testing for Serial Correlation in Least Squares Regression, II”. In: *Biometrika* 37.3/4 (1950), pp. 409–428.

[FW99] L. R. Fabrigar and D. T. Wegener. *Exploratory Factor Analysis*. Oxford University Press, 1999.

[Faw] T. Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* ().

[Fox15] J. Fox. *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, 2015.

[HLS13] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. 3rd. Wiley, 2013.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[JC16] I. T. Jolliffe and J. Cadima. *Principal Component Analysis*. 2nd ed. Springer, 2016. DOI: 10.1007/978-3-319-41365-6.

[Jam+21] G. James et al. *An Introduction to Statistical Learning: with Applications in R*. 2nd ed. Springer, 2021. DOI: 10.1007/978-1-0716-1418-1.

[Kai60] H. F. Kaiser. “The application of electronic computers to factor analysis”. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 141–151. DOI: 10.1177/001316446002000116.

[Kai74] H. F. Kaiser. “An index of factorial simplicity”. In: *Psychometrika* 39.1 (1974), pp. 31–36. DOI: 10.1007/BF02291575.

[LR19] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. 3rd. Wiley, 2019.

[MM15] D. S. Moore and G. P. McCabe. *Introduction to the Practice of Statistics*. 9th. W. H. Freeman, 2015.

[MPV12] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley, 2012.

[R C23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[Rev23] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.3.9. 2023. URL: https://CRAN.R-project.org/package=psych.

[Tzi+21] G. Tziritas et al. “Machine learning in wine chemistry: Predicting wine sensory ratings from physicochemical parameters”. In: *Computers and Electronics in Agriculture* 185 (2021), p. 106143.

[VR02] W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer, 2002.

[Wel47] B. L. Welch. “The generalization of ‘Student’s’ problem when several different population variances are involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35.