**Text

Description automatically generated**

**Red Wine Dataset Analysis**

**PROJECT  REPORT**

CSE 240 Data Science With R

**TERM II(Year 2)**

***Submitted by***

**AADITHYA PRABHA R       E0321008**

**SRI RAKSHA S                       E0321062**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**(ARTFICIAL INTELLIGENCE & DATA ANALYTICS)**

**Sri Ramachandra Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116**

**JANUARY, 2022**

|  |  |  |
| --- | --- | --- |
| S.NO | TABLE OF CONTENTS | PAGE NO. |
|  | **ABSTRACT** | 3 |
| **1** | INTRODUCTION | 4 |
| **2.** | OBJECTIVE | 5 |
| **3** | TECHNOLOGY USED and PACKAGES | 6-7 |
| **4.** | DATASET ANALYSIS | 8-9 |
| **5** | MACHINE LEARNING ALGORITHMS | 10-17 |
| **6** | CONCLUSION | 18 |

ABSTRACT

* Dataset analysis refers to the process of examining, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.
* It typically involves a combination of statistical and machine learning techniques and can be applied to a wide range of fields, including finance, healthcare, marketing, and more.
* The specific steps involved in a dataset analysis will vary depending on the nature of the data and the goals of the analysis, but common tasks include data visualization, feature engineering, model selection, and performance evaluation.

INTRODUCTION

* The pH of red wine typically ranges from 3.3 to 3.6, making it slightly acidic.
* The acidity in red wine is primarily due to the presence of tartaric acid, which is naturally occurring in grapes.
* The pH level of a wine can affect its taste and aging potential, with lower pH levels resulting in a more acidic and tannic wine.
* Alcohol content in red wine ranges from 12% to 15% by volume.
* High alcohol content in wine can contribute to a feeling of warmth and dryness in the mouth and can also affect the balance of flavours in the wine.
* The alcohol content in wine can also impact the aging potential of a wine and higher alcohol content wine will age faster.

OBJECTIVE

The objective of data analysis for red wine quality could be to understand the factors that contribute to the quality of red wine and how they are related. This could include:

* Identifying key chemical and physical properties of red wine that are associated with quality, such as pH, alcohol content, and acidity.
* Investigating the relationship between these properties and the sensory characteristics of red wine, such as taste, aroma, and color.
* Examining the impact of wine production methods, such as grape variety, viticulture, and winemaking techniques, on wine quality.
* Analyzing data from wine tasting evaluations to understand how experts and consumers perceive the quality of different red wines.
* Developing predictive models to identify the characteristics of high-quality red wines and identify wines that have the potential to improve with aging.
* Using statistical methods to identify the significant factors that contribute to the quality of red wine and understand how they are related.
* Identifying any patterns or trends in the data that could be used to improve the quality of red wine in the future.
* Evaluating the performance of different models and techniques used in the analysis to choose the best approach for predicting wine quality.

TECHNOLOGY USED and PACKAGES

R-STUDIO



RStudio is an open-source integrated development environment (IDE) for the R programming language. It is designed to make working with R more efficient and user-friendly.

RStudio is widely used by data scientists, statisticians, and researchers to analyze data, create statistical models, and generate reports and visualizations. It's also often used in data science education and training.

Packages:

1.plotly

Plotly is an R package that allows you to create interactive, web-based plots and visualizations using the Plotly.js library.

2.factoextra

Factoextra is an R package that provides additional functionalities for exploratory factor analysis (EFA) and principal component analysis (PCA).

3.e1071

The e1071 R package is a package for classification and other machine learning techniques. It provides a collection of functions for training and evaluating models, as well as for pre-processing and post-processing data.

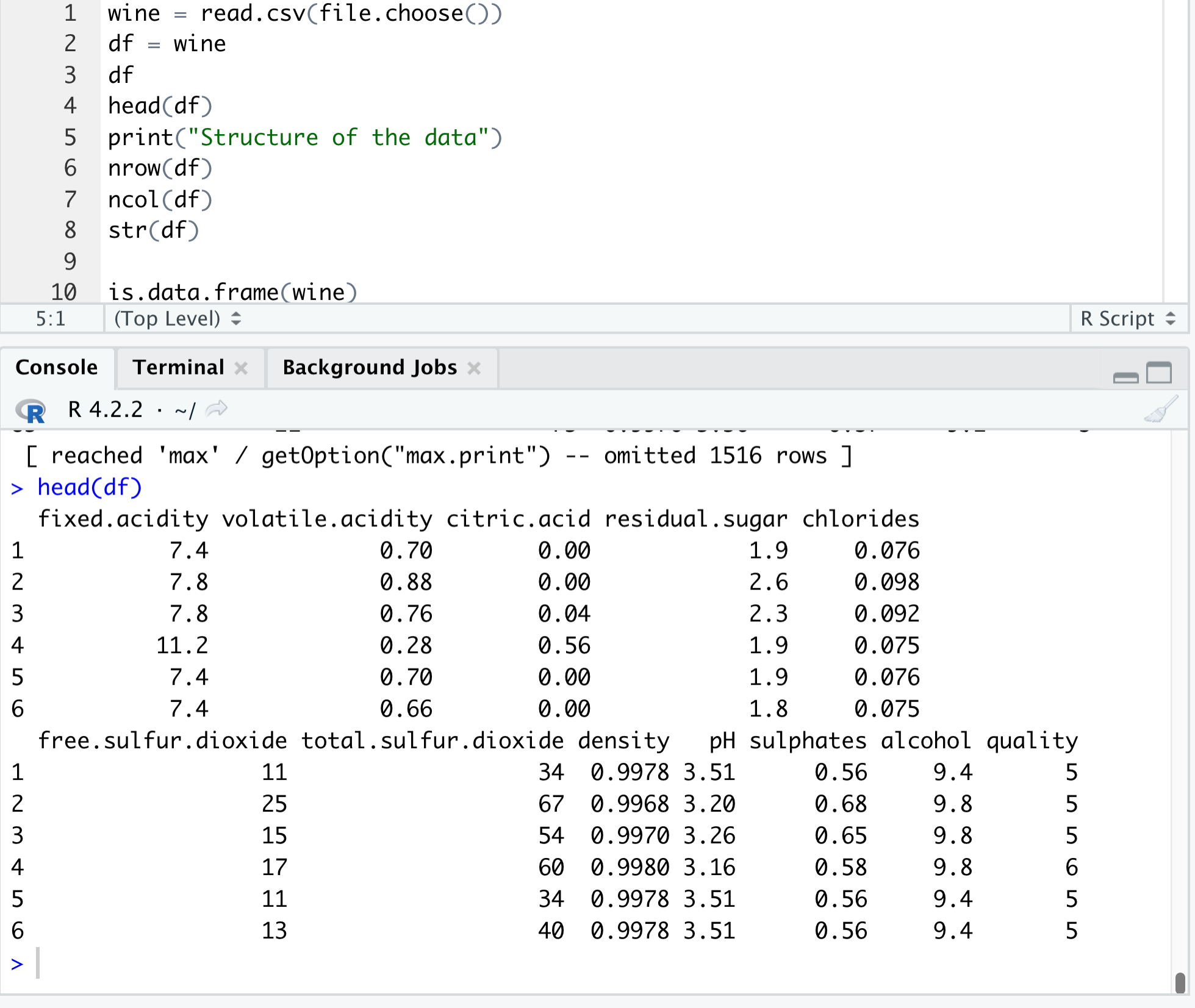
4.caTools

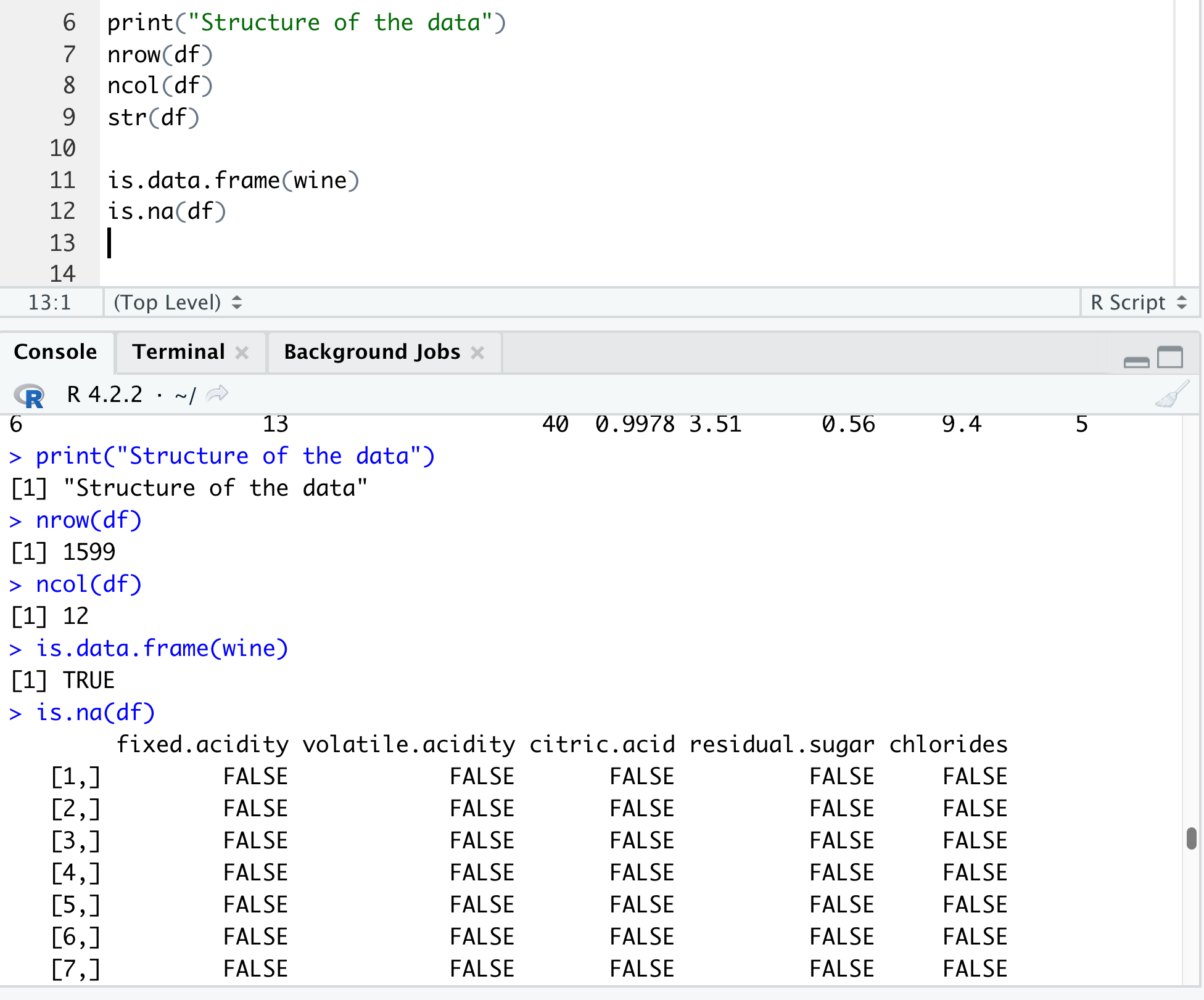
caTools is an R package that provides tools for reading, writing, and manipulating data in various file formats. It provides a set of functions for working with compressed files, as well as for working with character encodings.

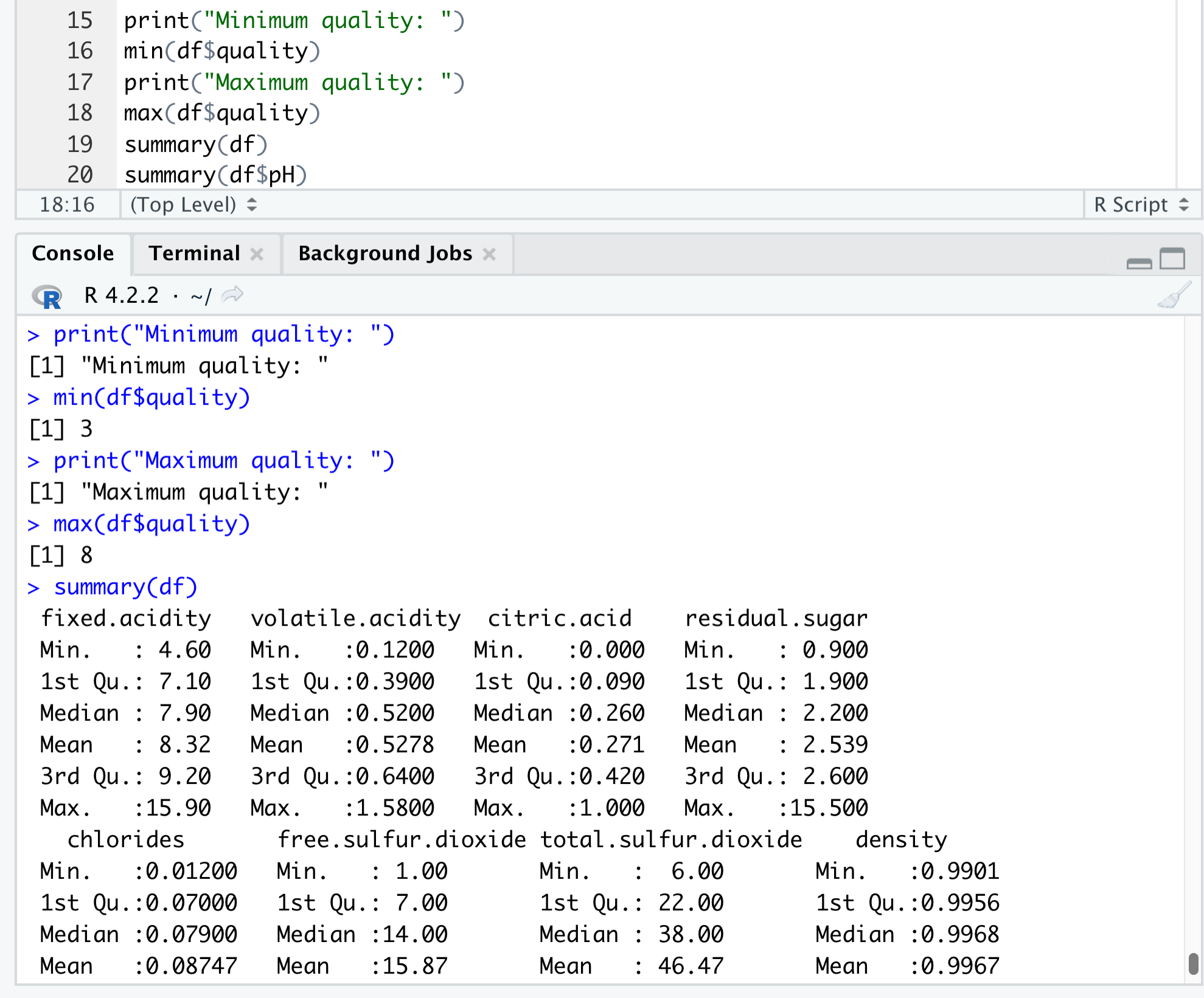
5.class

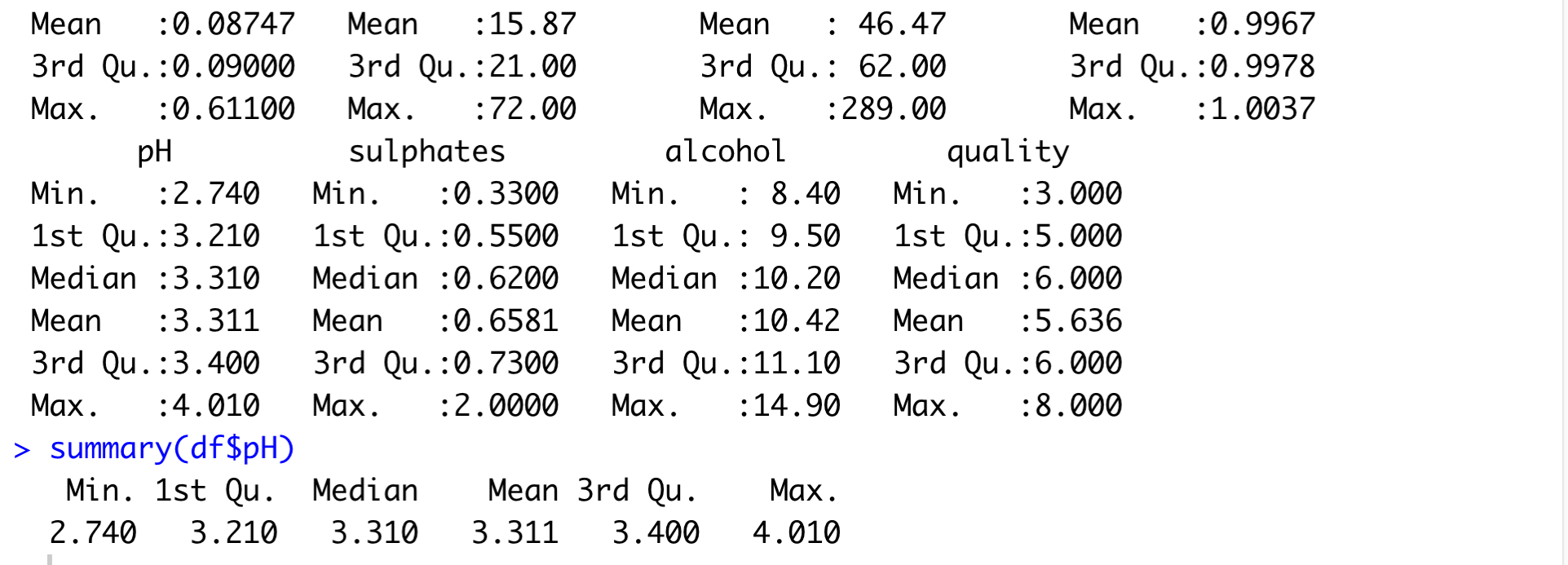
The class R package is a package for classification and other machine learning techniques. It provides a set of functions for training and evaluating models, as well as for pre-processing and post-processing data.

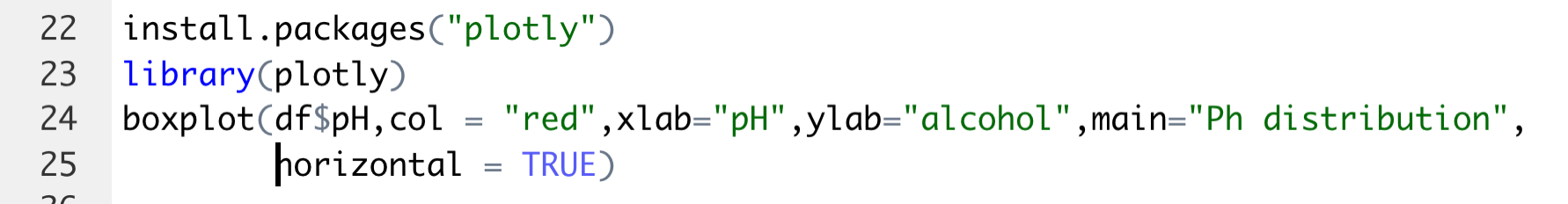
DATASET ANALYSIS

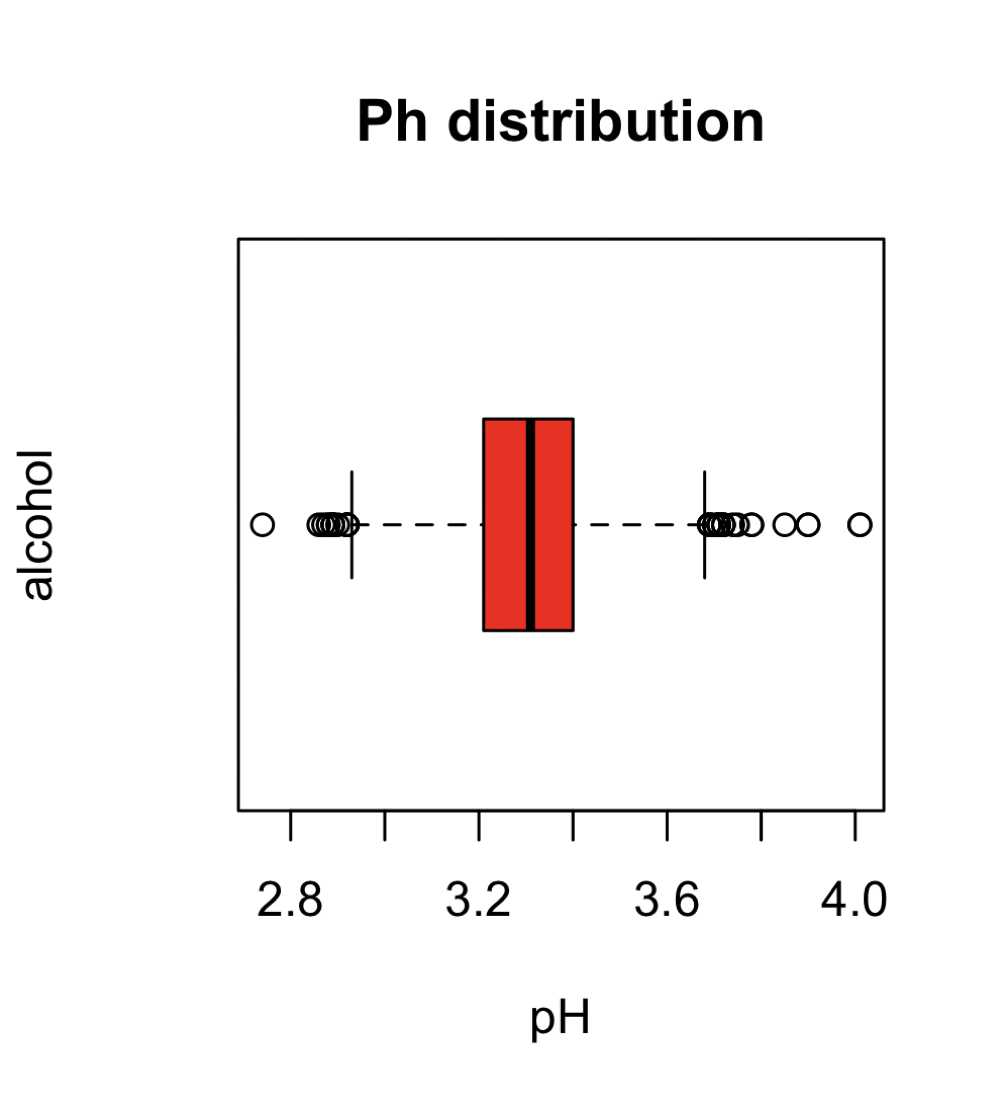












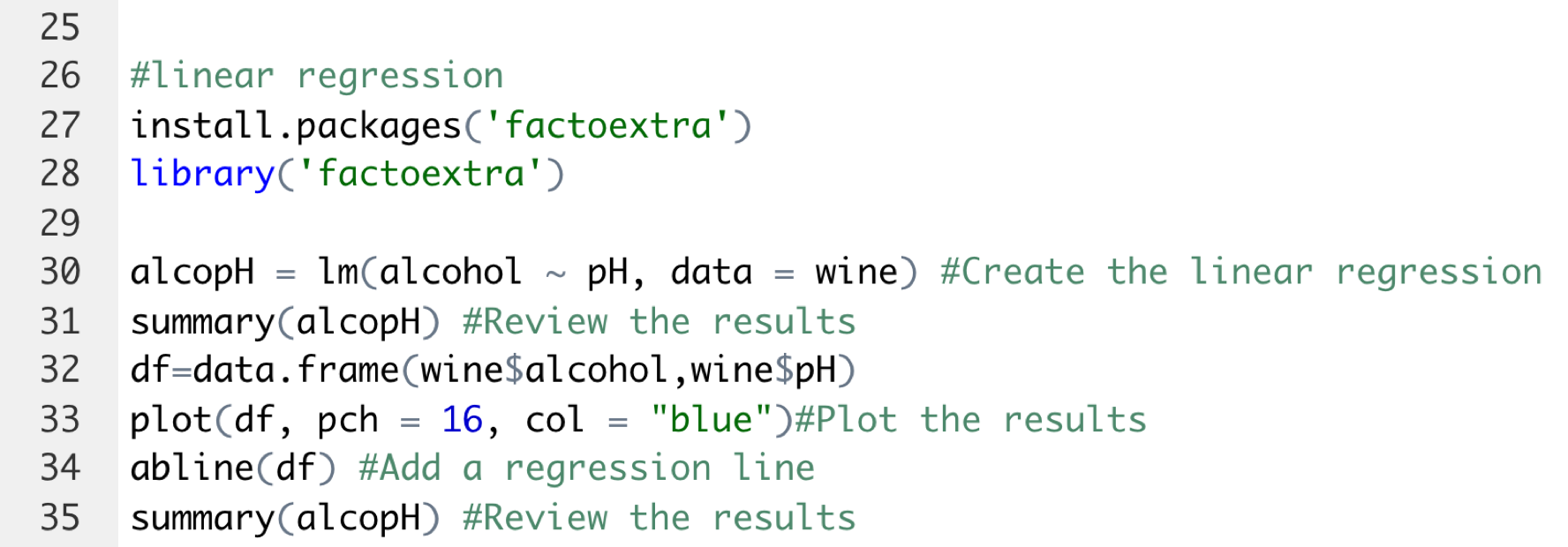
MACHINE LEARNING ALGORITHMS

LINEAR REGRESSION

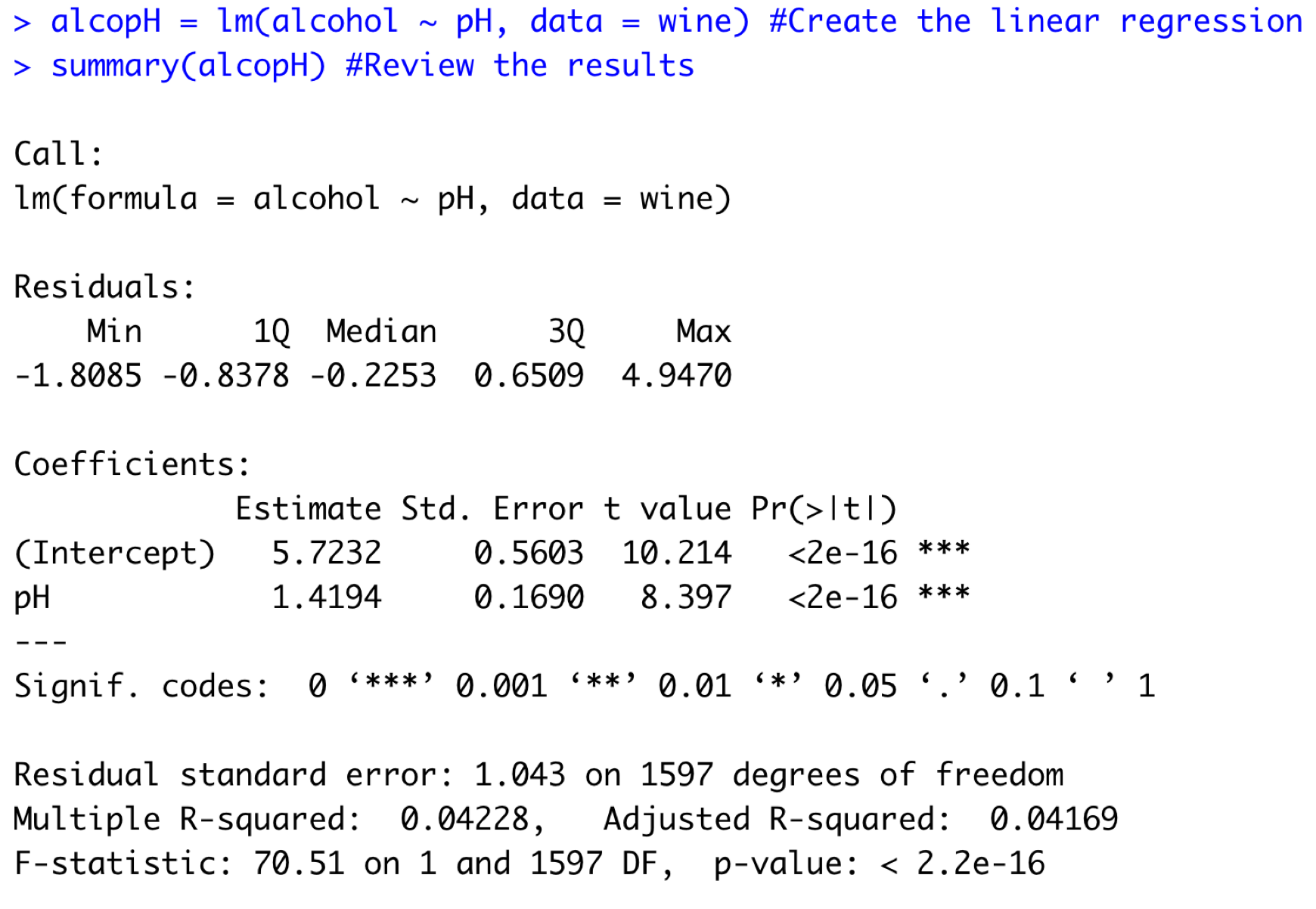
Linear regression is a statistical method used to model the relationship between a dependent variable (y) and one or more independent variables (x).

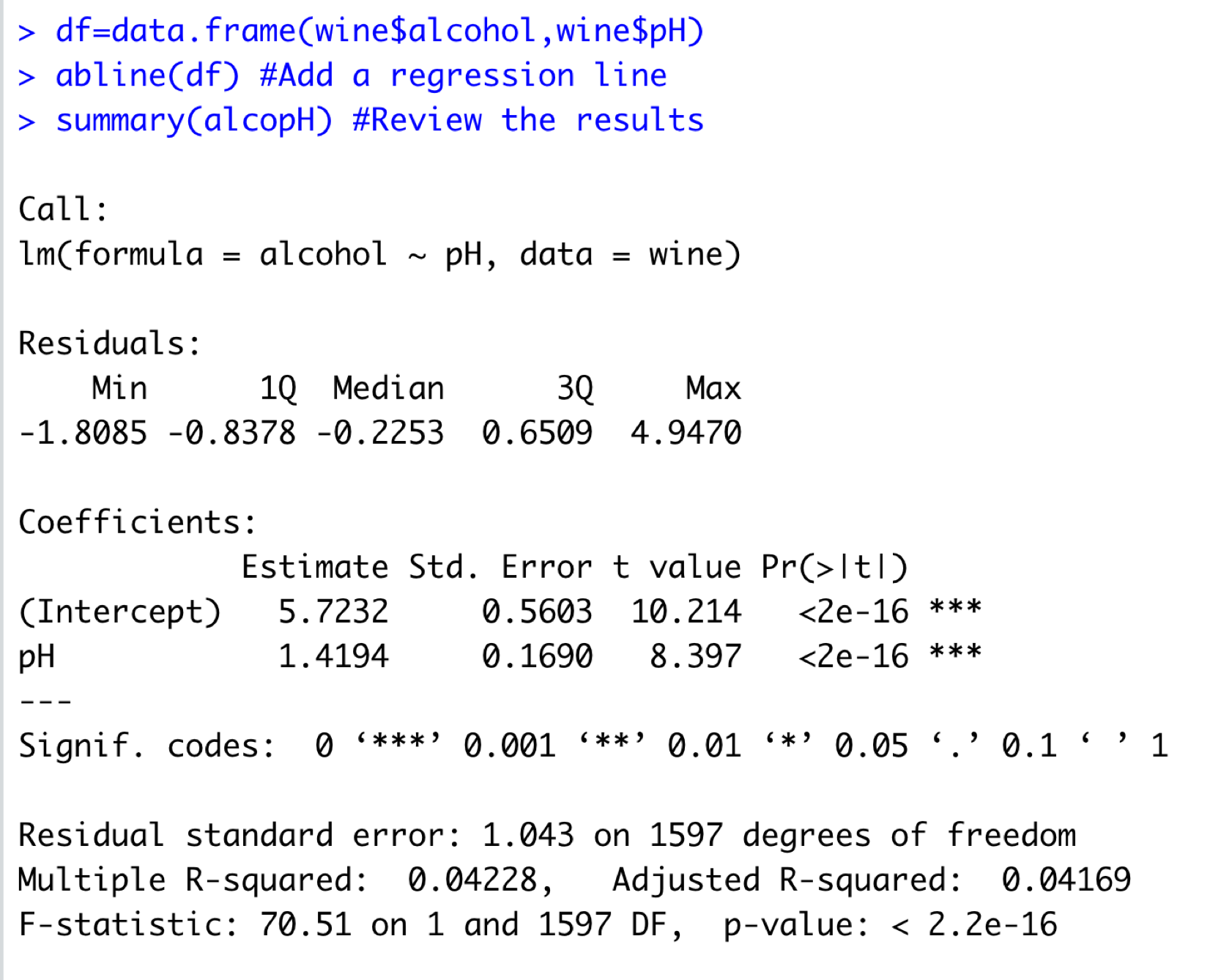
Steps for linear regression

* 1. Collect and organize data in a spreadsheet with columns for the independent variable(s) and dependent variable
  2. Plot the data on a scatter plot to check for linearity and homoscedasticity
  3. Use a linear regression software or calculator to find the equation of the line of best fit ( y = mx + b)
  4. Use the equation to make predictions for new data points
  5. Evaluate the model by calculating the R-squared value and checking residuals plot
  6. Communicate the results and interpret the coefficients of the equation in the context of the problem.



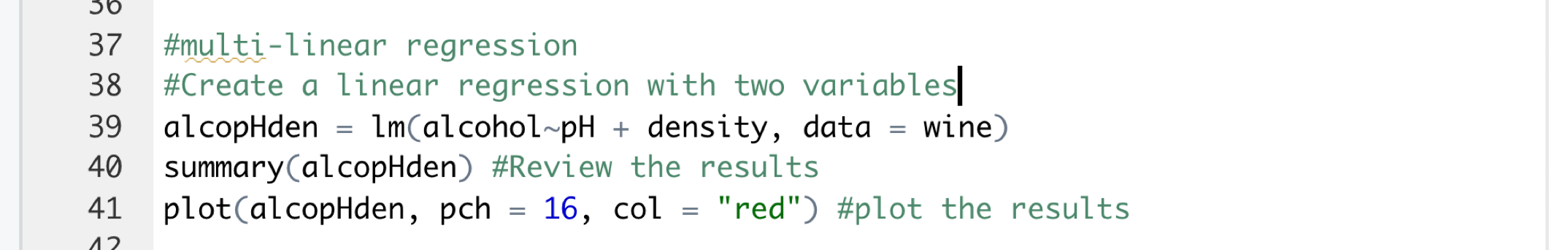


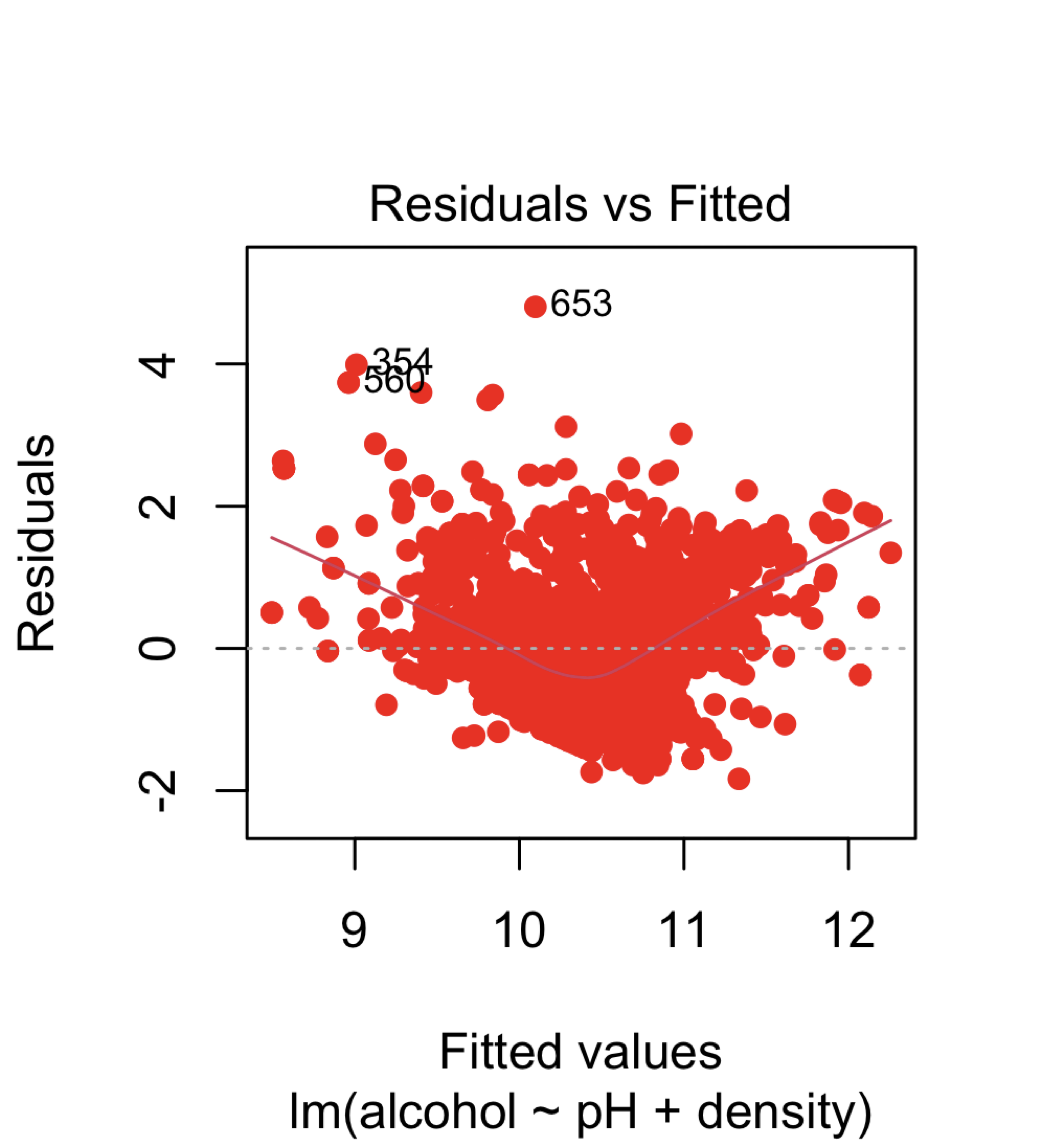


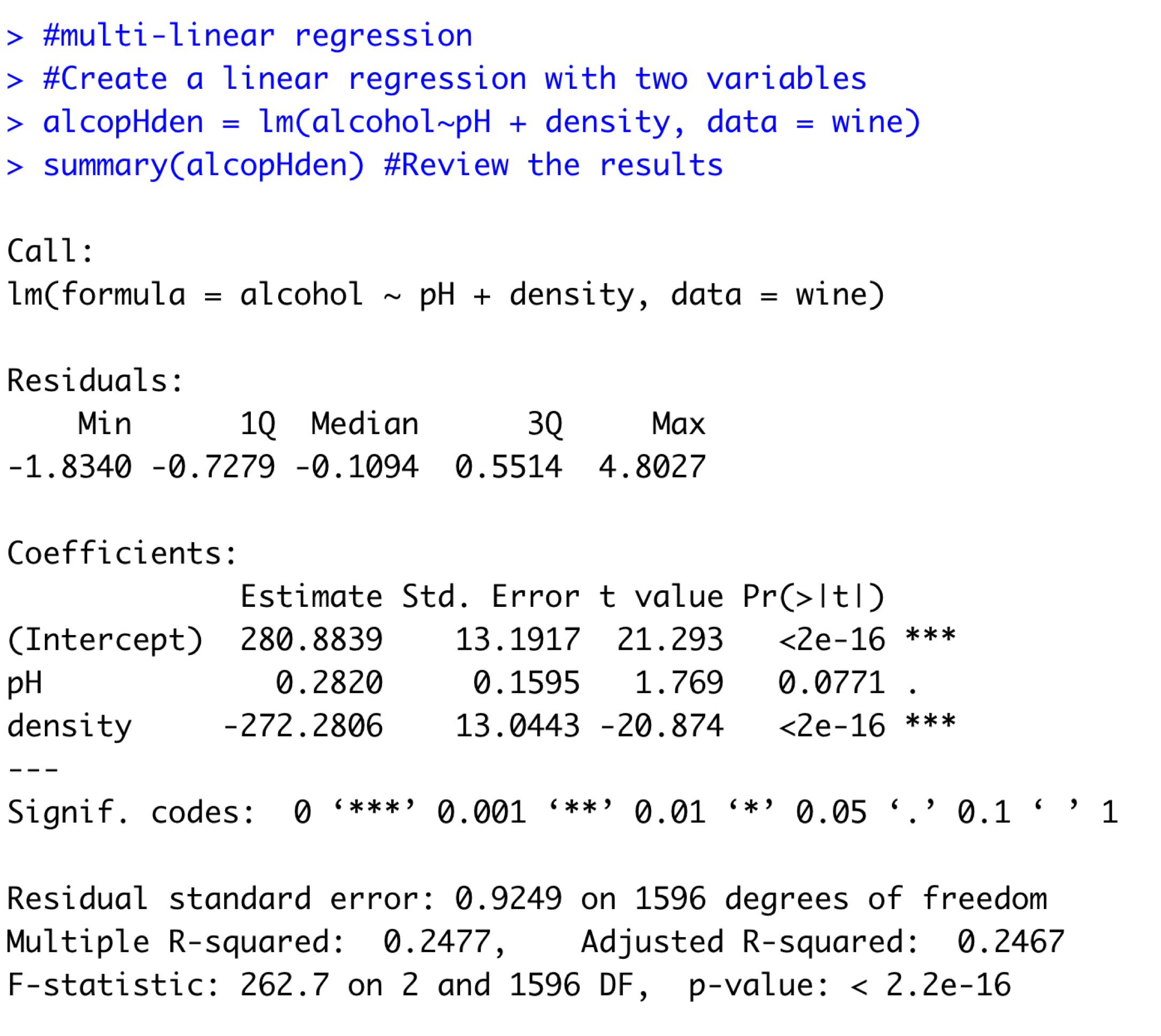


MUTLI LINEAR REGRESSION

Multi-linear regression is a type of linear regression that is used to model the relationship between a dependent variable and multiple independent variables.







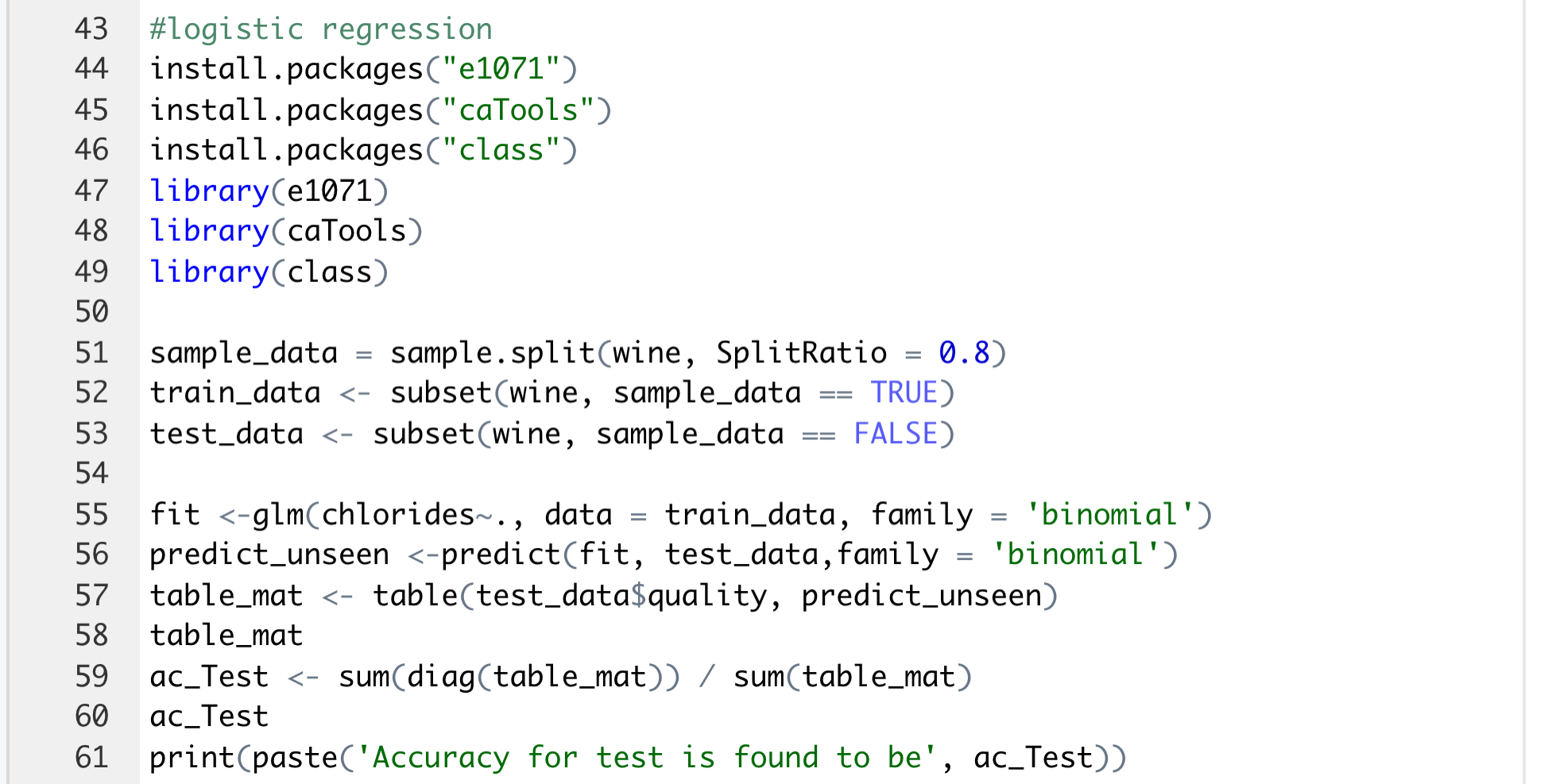
LOGISITIC REGRESSION

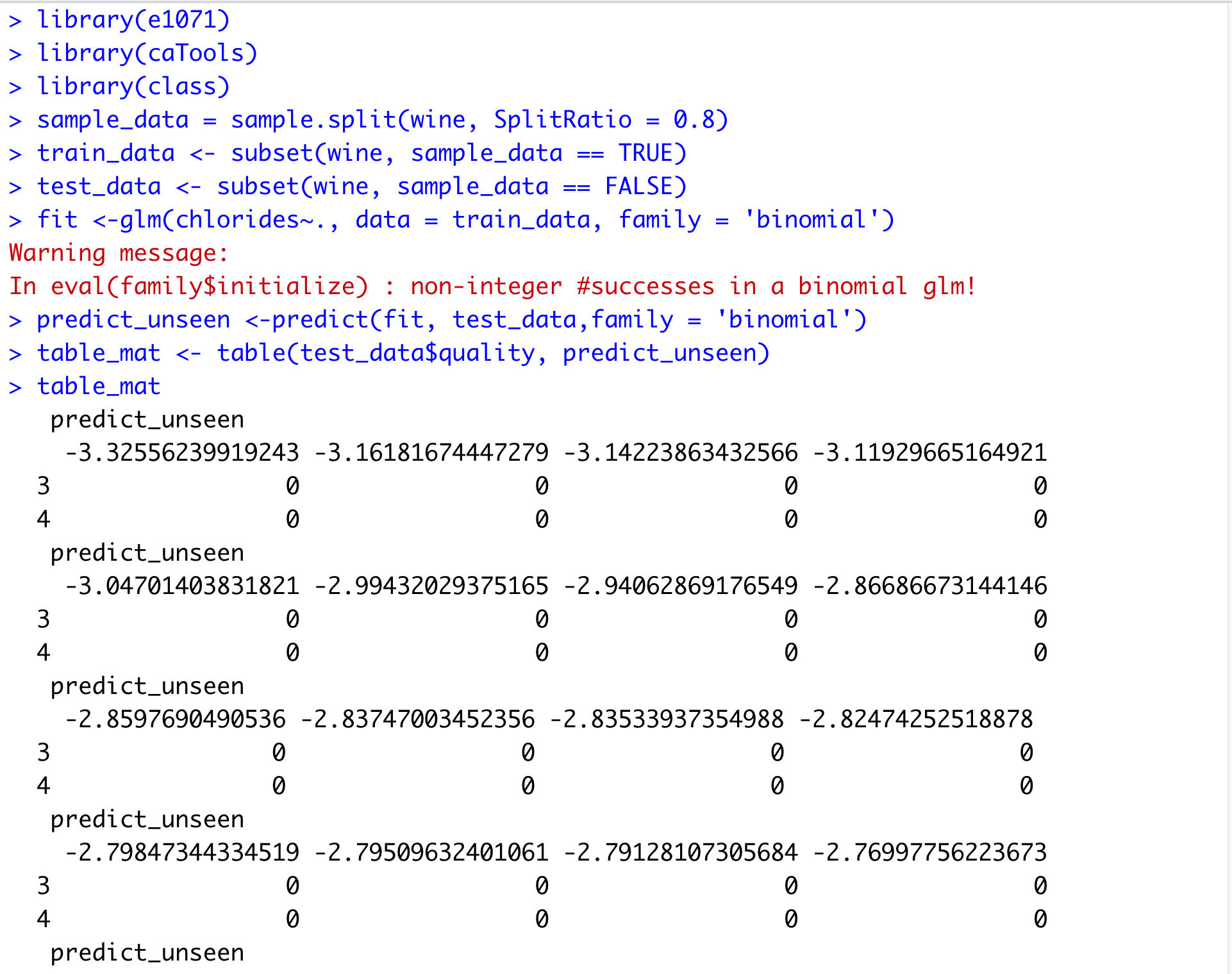
Logistic regression is a type of statistical analysis used for predicting a binary outcome (1 / 0, Yes / No, True / False) based on one or more predictor variables.

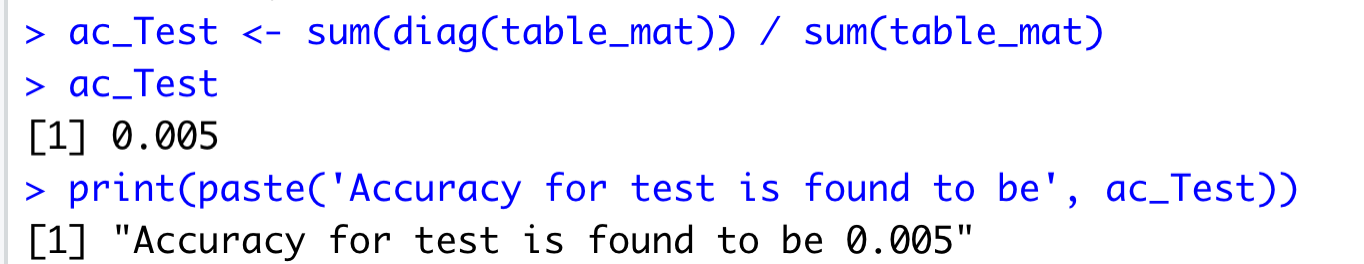
Steps for logistic regression:

1. Collect and organize data in a spreadsheet with columns for the independent variable(s) and binary dependent variable
2. Use a logistic regression software or calculator to find the equation of the logistic curve
3. Use the equation to make predictions for new data points
4. Evaluate the model by calculating the Deviance and checking residuals plot
5. Communicate the results and interpret the coefficients of the equation in the context of the problem.
6. Examine the goodness of fit by calculating the confusion matrix and evaluating the accuracy, precision, recall and F1 score.
7. Evaluate the model by calculating the area under the ROC curve.
8. Identify and handle any outliers and influential points, if any

If necessary, use model selection techniques such as forward and backward selection, stepwise regression and regularization technique to select the best model.







|  |  |
| --- | --- |
|  |  |
|  |  |

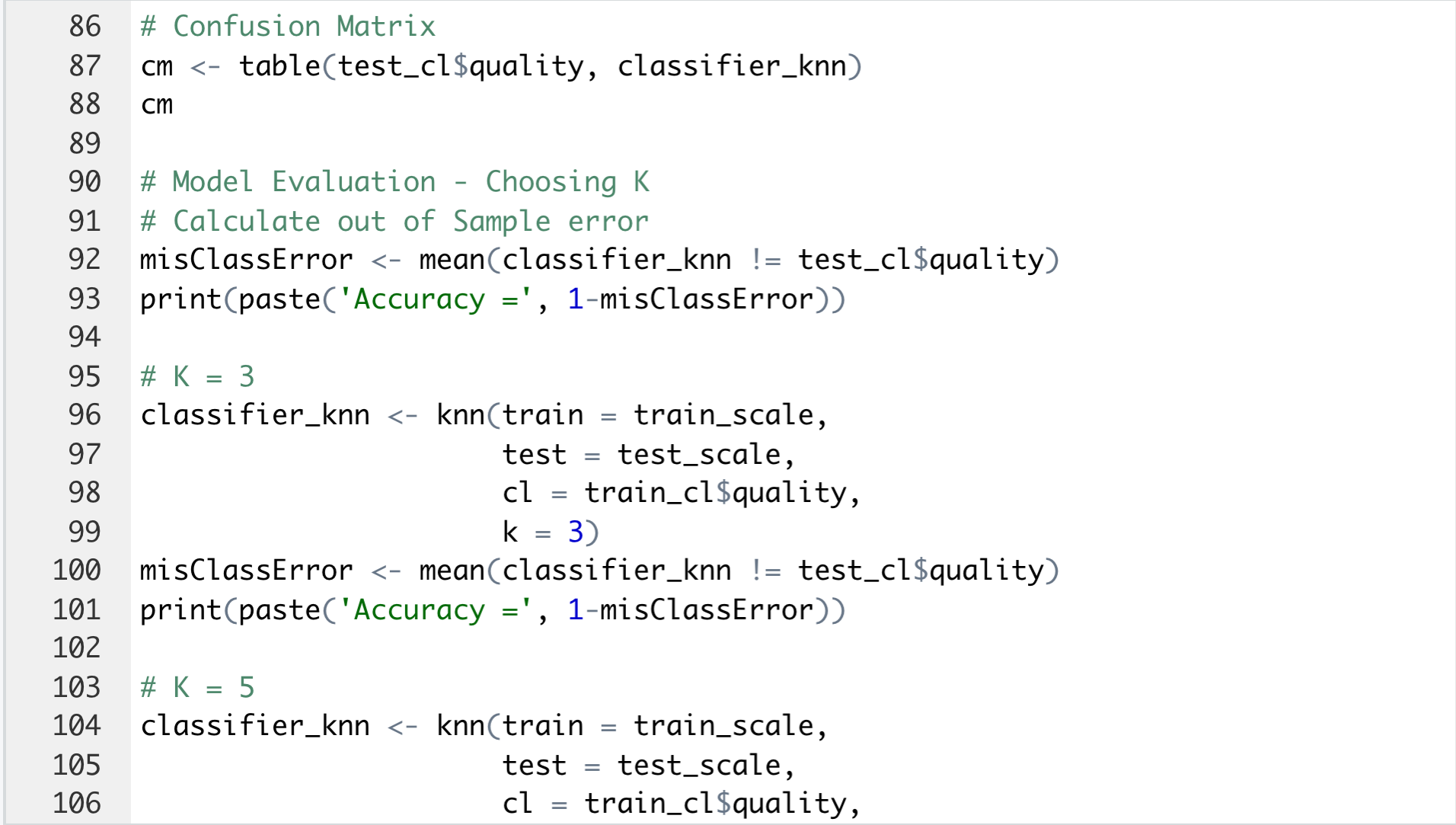
k-Nearest Neighbors (k-NN)

k-Nearest Neighbors (k-NN) is a type of supervised machine learning algorithm used for classification and regression. The basic idea behind k-NN is to predict the class or value of a new data point by looking at the k number of closest "neighbors" from the training data.

Steps for k-NN

* Collect and organize data in a spreadsheet with columns for the independent variable(s) and dependent variable
* Split the data into a training set and a test set
* Choose the value of k (number of nearest neighbors to consider)
* For each point in the test set, calculate the distance to all points in the training set
* Select the k closest points in the training set
* For a classification problem, assign the majority class among the k nearest points to the test point. For a regression problem, assign the mean value among the k nearest points to the test point
* Compare the predicted class or value to the actual class or value and calculate the accuracy
* Communicate the results, including any limitations and assumptions, to the appropriate audience

Repeat the process with different values of k to determine the best value of k for the problem.





|  |  |
| --- | --- |
|  |  |

CONCLUSION

In conclusion, data analysis for red wine quality is a complex task that involves understanding the chemical and physical properties of red wine and how they relate to sensory characteristics and wine production methods.

A thorough analysis of wine quality data can provide insights into the factors that contribute to the quality of red wine and how they are related.

Additionally, by using statistical methods, patterns and trends can be identified that can be used to improve the quality of red wine in the future.

Additionally, R packages such as Plotly and Factoextra as well as e1071 and caTools can help in the data visualization and data manipulation, which is an important step in data analysis.

To achieve the best results, it is important to use appropriate statistical techniques and tools, while also taking into account the limitations and assumptions of the methods used.