Final Project

Jihyun Lee

11/7/2020

```
library(dplyr)
library(kableExtra)
```

1 Introduction

1.1 Data

This wine quality dataset is from [1]. The dataset includes *vinho verde*, a unique product from the Minho (northwest) region of Portugal. The data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of vinho verde. The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis.

The outcome variable is `wine quality`, which was measured by a minimum score of three sensory assessors using blind tastes in a scale that ranges from 0 (very bad) to 10 (excellent). There are 11 attributes of the wine based on physicochemical tests: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Originally, two datasets were created separately, one for red wine ($n = 1599$) and another for white wine ($n = 4988$). In this report, we use a merged dataset and create a dummy variable to indicate the wine type, `red`. Thus, in total, the dataset includes 11 numerical attributes (covariates), one dummy variable, and one numerical outcome. There is no missing value in this dataset.

1. Descriptive statistics

   Table 1 shows the descriptive statistics of 11 attributes by wine type.

2. Histogram

Table 1: Descriptive statistics of 11 attributes

	White			Red		
	Min	Max	Mean	Min	Max	Mean
fixed.acidity	3.8	14.2	6.855	4.6	15.9	8.320
volatile.acidity	0.08	1.10	0.278	0.12	1.58	0.528
citric.acid	0	1.66	0.334	0	1.00	0.271
residual.sugar	0.6	65.8	6.391	0.9	15.5	2.539
chlorides	0.009	0.346	0.046	0.012	0.611	0.087
free.sulfur.dioxide	2	289	35.308	1	72	15.875
total.sulfur.dioxide	9	440	138.361	6	289	46.468
density	0.987	1.039	0.994	0.990	1.004	0.997
pH	2.72	3.82	3.188	2.74	4.01	3.311
sulphates	0.22	1.08	0.490	0.33	2.00	0.658
alcohol	8.0	14.2	10.514	8.4	14.9	10.423

1.2 Research Questions

References

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