GR5291 Project Wine Quality Analysis

Saier Gong (sg3772)

5/2/2020

Part 1 Data Introduction

```
## [1] 6463   13
```

1.1 Data Introduction

This wine quality data set is downloaded from Kaggle, it has 6497 observations and 13 columns, including one categorical explanatory variable "type", 11 quantitative explanatory variables, and one response variable. The reponse variable is the quality of wine, from 1 to 10, which represents from low quality to high quality.

After cleaning the data set, removing some rows with missing values, the data set has 6463 observations. I also modify the response variable:

- if the quality score is higher than 5, set it as 1, representing high quality;
- if the quality score is lower than 5, set it as 0, representing low quality.

So now, the data set has 6463 observation rows, 1 categorical variable, 11 quantitative variables and a $0, 1$ response variable.

```
## # A tibble: 6 x 2
##   variable         feature
##   <chr>            <chr>
## 1 type             categorical
## 2 fixed.acidity    quantitative
## 3 volatile.acidity quantitative
## 4 citric.acid      quantitative
## 5 residual.sugar   quantitative
## 6 chlorides        quantitative
```

```
## # A tibble: 7 x 2
##   variable             feature
##   <chr>                <chr>
## 1 free.sulfur.dioxide  quantitative
## 2 total.sulfur.dioxide quantitative
## 3 density              quantitative
## 4 pH                   quantitative
## 5 sulphates            quantitative
## 6 alcohol              quantitative
## 7 quality              dependent
```

1.2 Data Collection

Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

1.3 Interesting Question

Through analyzing this wine quality data, I want to show the following interesting topics:

- which feature or combination of several features may contribute most to the taste of wine;
- how the type of wine may influence the taste of wine;
- only according to the provided variables, how to make make and choose good wine.

Part 2 Conclusions

According to my analysis, I have the following conclusions about the interesting topics:

- The volatile acidity variable has the greatest impact on the quanlity of wines;
- Given other quantitative variables the same, the estimated odds of high quality in white wine is about 6 of the odds of high quality in red wine. Therefore, the type of wine has an influence on quality, and it is more likely to buy fake white wine.
- With more alcohol and sulphates and with less volatile acidity, the quality of wine is increasing.