Exploratory Data Analysis

Dipika Subramaniam

Contents

Paper Requirements: **1**

Getting Started **1**

Distributions of Chlorides in red and white **2**

Quality comparison between Red and White **3**

Correlation between Residual Sugars and Density in Red **4**

Paper Requirements:

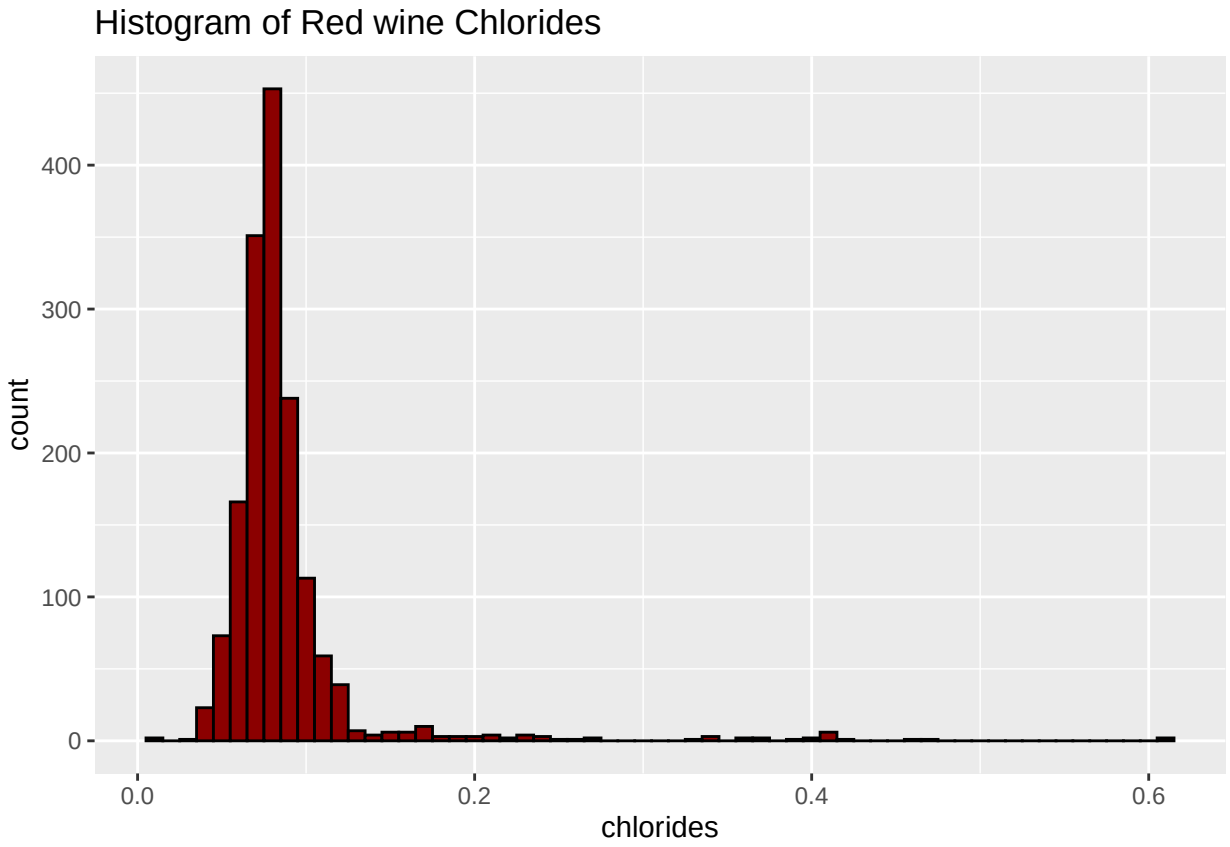
Submit at least 4 graphical displays according to the standards used in class. Include both one- and two-variable plots. For each, write a sentence or two describing the conclusions suggested by the plots. In your paper, you'll need to interrogate these conclusions more carefully. The purpose here is just preliminary analysis.

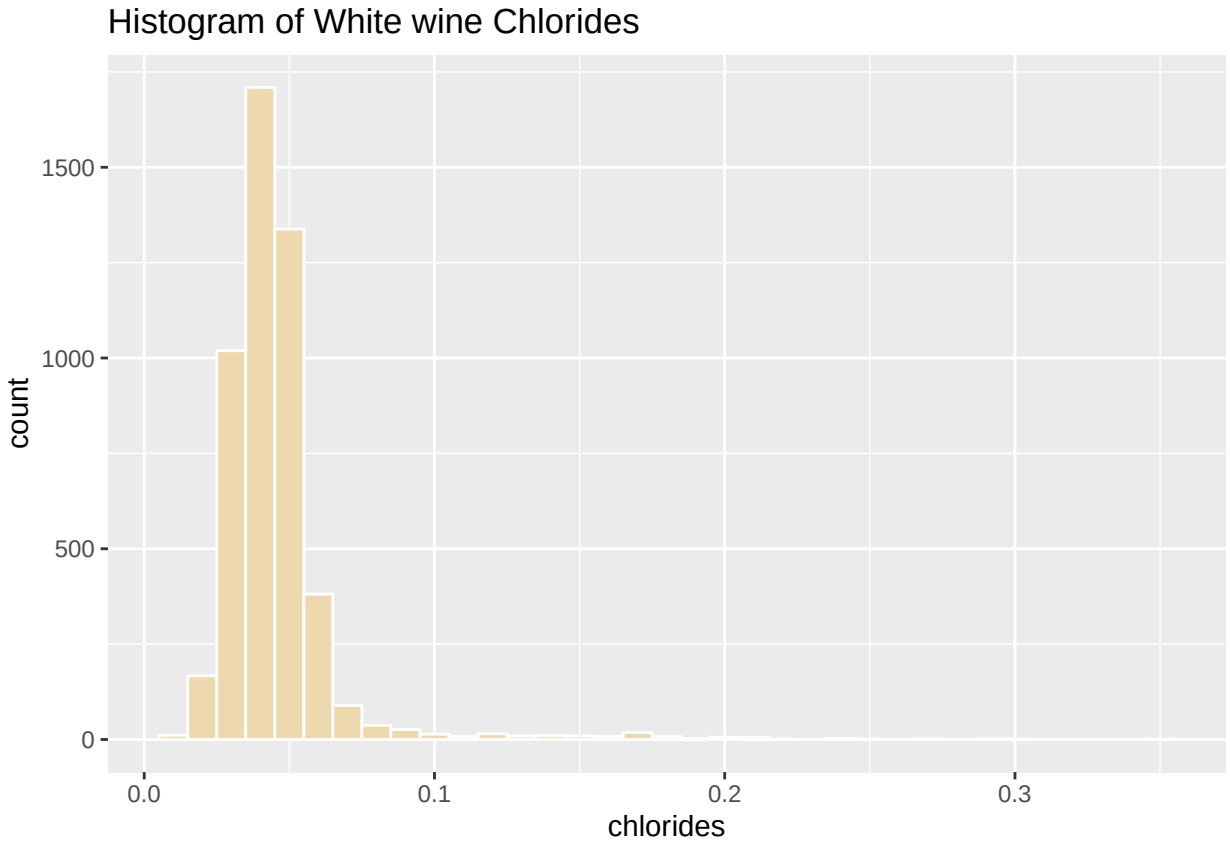
Getting Started

The `white_data` contains sample data for 4898 white wines from some unknown population. The `red_data` contains sample data for 1599 red wines from some unknown population. The variables are summarized below, as given by the winequality.names file and [Cortez et al., 2009]. This paper was made in R Markdown.

```
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)
```

Distributions of Chlorides in red and white

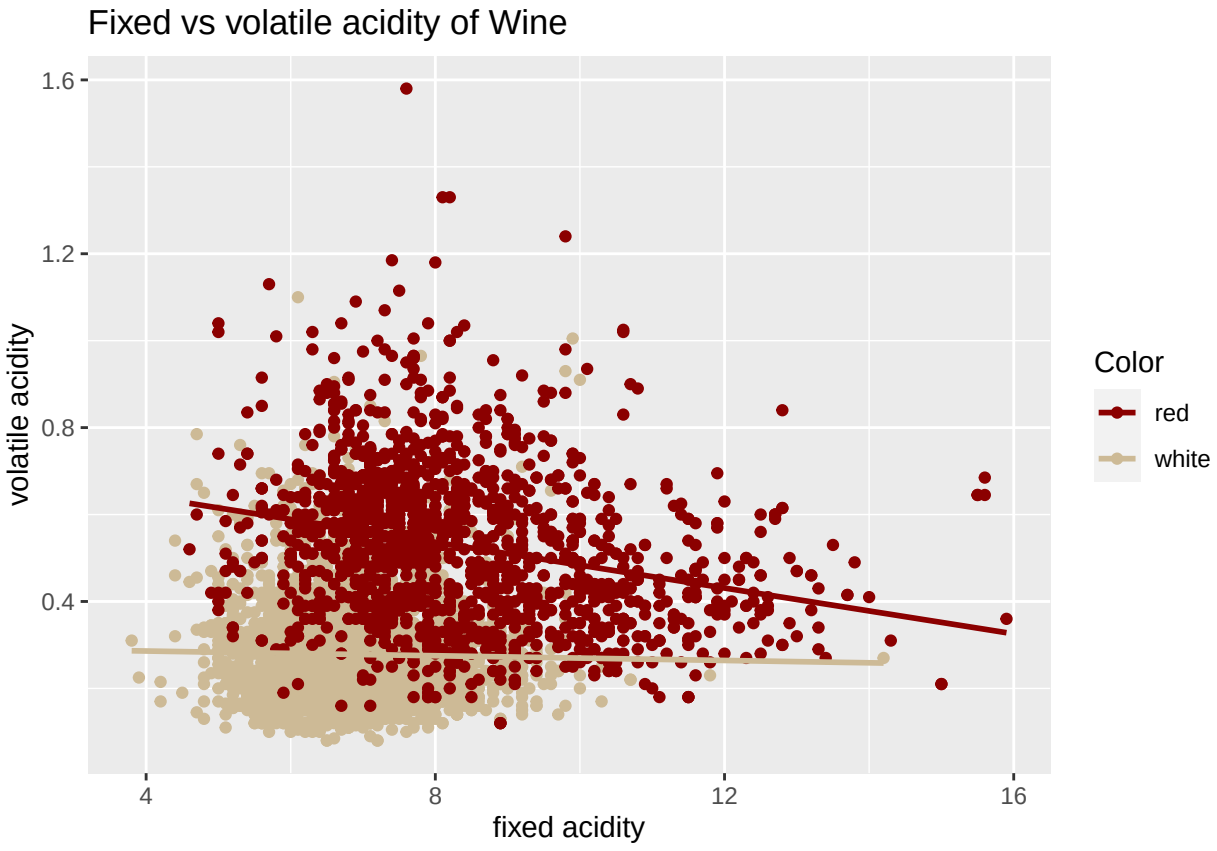




From the previous histograms, we can visually approximate the chlorides to have a normal distribution in both red and white.
We can use this observation in the paper, we can calculate confidence intervals and do some hypothesis testing to see if the means are equal. There are many different tests and exercises that we could try out here.
Off the top of my head, it seems as though they both have right skews, and I wonder if I need to remove outliers of chlorides >0.2 for red and >0.1 for white. Visually we can see that white has a mean of 0.04, and red has a mean of 0.08.

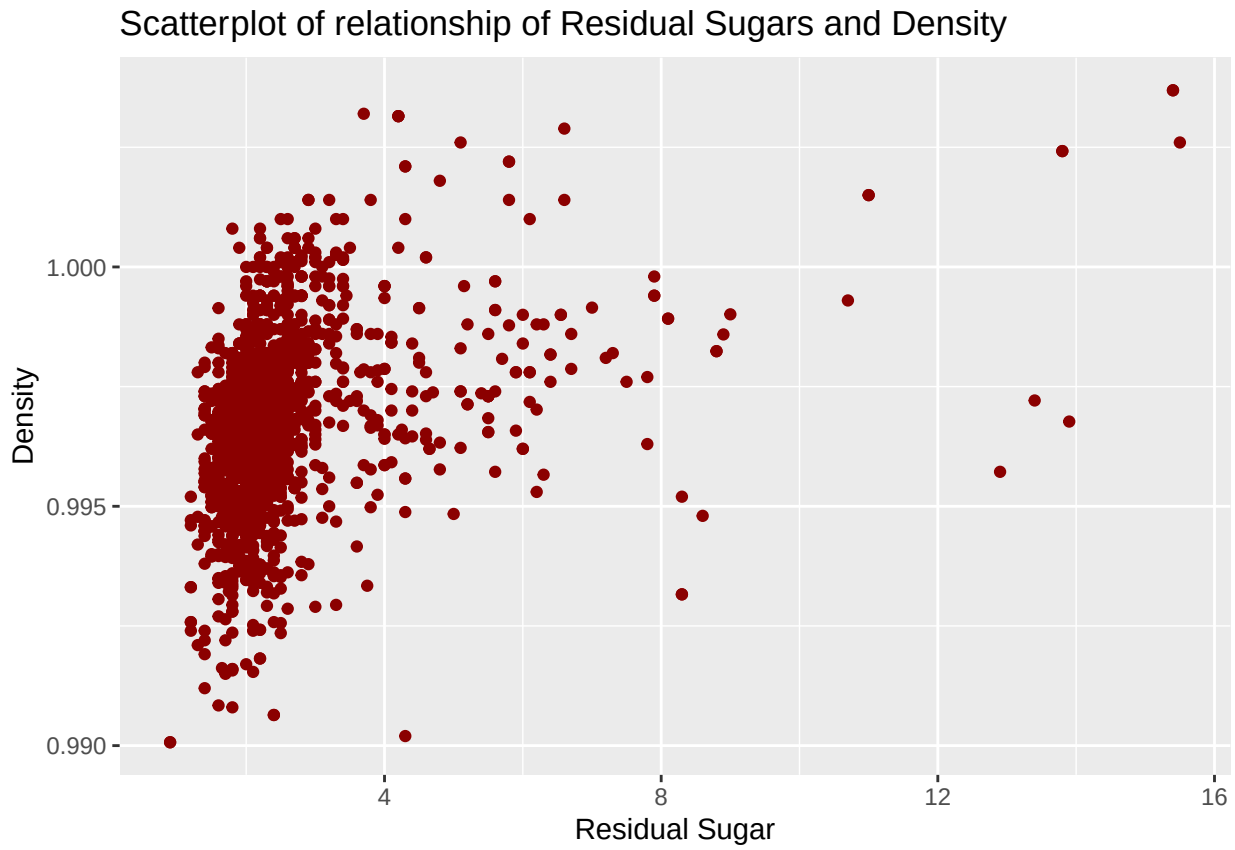
Quality comparison between Red and White

Here, we have merged two datasets into one, called total_data, where the red and white are distinguished by a variable named "color".



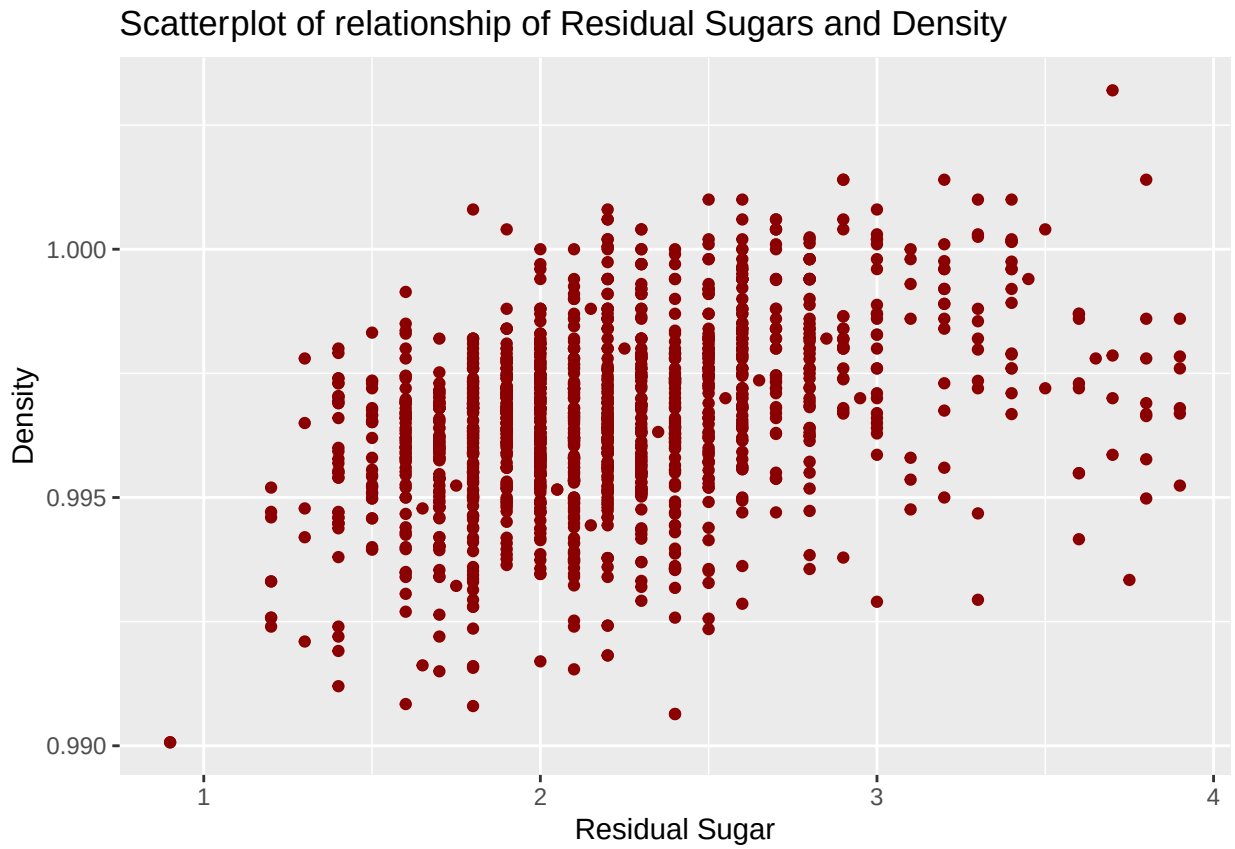
I'm not yet sure of the relationship between fixed and volatile acidity related to color of wine, but I'm including it in this data analysis because I want to investigate why the red is more spread than the white.

Correlation between Residual Sugars and Density in Red




```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.355
```

Here we have made a graph of residual sugars vs density. The correlation is 0.355. As you can see, there are a few values with residual sugars > 4, but the majority is a blob that lies under that value.
If we remove a few outliers, we could get a clear view and possibly a stronger correlation between residual sugars and density.




```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.395
```

We found a slightly stronger correlation of 0.395, which is still a weak correlation. Even though its weak, this data could be useful for the paper.