

Proposal

STA 210 - Project

Team 2 - Alicia Gong, Ashley Chen, Abdel Shehata, Claire Tan

```
library(tidyverse)
library(tidymodels)
library(dplyr)
library(ggplot2)
library(cowplot)

redwine <- read.csv("winequality-red.csv", sep = ";")
whitewine <- read.csv("winequality-white.csv", sep = ";")
redwine<-redwine%>%mutate(color="red")
whitewine<-whitewine%>%mutate(color="white")
wine<-redwine%>%full_join(whitewine)
```

Introduction

Project Goal: To identify variables that are important in explaining variation in the response.

We are interested in what factors contribute to the quality of Portuguese “Vinho Verde” red wine. The goal of this dataset is to model wine quality based on physicochemical tests. We believe that this dataset can also be used to analyze the relationship between different chemical compositions and the ratings of red wine quality. We believe this is important because by understanding what chemical compositions affect red wine qualities, it may shed some light in future direction of improving/preserving red wine quality.

Our goal is to produce a regression model that best explains how different chemical compositions of the Portuguese “Vinho Verde” red wine affects the variation of the red wine quality.

Data description

```
wine<- slice(wine, sample(1:n()))  
glimpse(wine)
```

Rows: 6,497

Columns: 13

```
$ fixed.acidity      <dbl> 6.8, 7.1, 7.2, 9.5, 7.2, 6.0, 6.5, 6.6, 10.3, 6.1~  
$ volatile.acidity  <dbl> 0.240, 0.210, 0.200, 0.860, 0.240, 0.360, 0.180, ~  
$ citric.acid       <dbl> 0.37, 0.28, 0.28, 0.26, 0.29, 0.06, 0.41, 0.04, 0~  
$ residual.sugar    <dbl> 7.45, 2.70, 1.60, 1.90, 3.00, 1.40, 14.20, 2.50, ~  
$ chlorides         <dbl> 0.043, 0.034, 0.028, 0.079, 0.036, 0.066, 0.039, ~  
$ free.sulfur.dioxide <dbl> 59, 23, 13, 13, 17, 27, 47, 48, 21, 25, 12, 44, 2~  
$ total.sulfur.dioxide <dbl> 188, 111, 168, 28, 117, 128, 129, 111, 51, 133, 7~  
$ density           <dbl> 0.99579, 0.99405, 0.99203, 0.99712, 0.99411, 0.99~  
$ pH                <dbl> 3.20, 3.35, 3.17, 3.25, 3.36, 3.26, 3.28, 3.14, 3~  
$ sulphates         <dbl> 0.50, 0.64, 1.06, 0.62, 0.68, 0.55, 0.72, 0.32, 0~  
$ alcohol           <dbl> 9.4, 10.2, 11.5, 10.0, 10.1, 9.3, 10.3, 11.9, 11.~  
$ quality           <int> 6, 4, 6, 5, 6, 5, 7, 5, 6, 7, 5, 5, 5, 5, 6, 5, 5~  
$ color             <chr> "white", "white", "white", "red", "white", "white~
```

There are 6497 observations and 13 variables.

```
summary(wine)
```

fixed.acidity	volatile.acidity	citric.acid	residual.sugar
Min. : 3.800	Min. : 0.0800	Min. : 0.0000	Min. : 0.600
1st Qu.: 6.400	1st Qu.: 0.2300	1st Qu.: 0.2500	1st Qu.: 1.800
Median : 7.000	Median : 0.2900	Median : 0.3100	Median : 3.000
Mean : 7.215	Mean : 0.3397	Mean : 0.3186	Mean : 5.443
3rd Qu.: 7.700	3rd Qu.: 0.4000	3rd Qu.: 0.3900	3rd Qu.: 8.100
Max. : 15.900	Max. : 1.5800	Max. : 1.6600	Max. : 65.800
chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
Min. : 0.00900	Min. : 1.00	Min. : 6.0	Min. : 0.9871
1st Qu.: 0.03800	1st Qu.: 17.00	1st Qu.: 77.0	1st Qu.: 0.9923
Median : 0.04700	Median : 29.00	Median : 118.0	Median : 0.9949
Mean : 0.05603	Mean : 30.53	Mean : 115.7	Mean : 0.9947
3rd Qu.: 0.06500	3rd Qu.: 41.00	3rd Qu.: 156.0	3rd Qu.: 0.9970
Max. : 0.61100	Max. : 289.00	Max. : 440.0	Max. : 1.0390
pH	sulphates	alcohol	quality

Min.	:2.720	Min.	:0.2200	Min.	: 8.00	Min.	:3.000
1st Qu.:	3.110	1st Qu.:	0.4300	1st Qu.:	9.50	1st Qu.:	5.000
Median	:3.210	Median	:0.5100	Median	:10.30	Median	:6.000
Mean	:3.219	Mean	:0.5313	Mean	:10.49	Mean	:5.818
3rd Qu.:	3.320	3rd Qu.:	0.6000	3rd Qu.:	11.30	3rd Qu.:	6.000
Max.	:4.010	Max.	:2.0000	Max.	:14.90	Max.	:9.000

color
Length:6497
Class :character
Mode :character

```
p1 <- ggplot(data = wine, aes(x = quality) ) +
  geom_histogram(fill = "pink")

p2 <- ggplot(data = wine, aes(x = fixed.acidity) ) +
  geom_histogram(fill = "pink")

p3 <- ggplot(data = wine, aes(x = volatile.acidity) ) +
  theme(axis.text=element_text(size=9)) +
  geom_histogram(fill = "pink")

p4 <- ggplot(data = wine, aes(x = citric.acid) ) +
  theme(axis.text = element_text(size=9)) +
  geom_histogram(fill = "pink")

p5 <- ggplot(data = wine, aes(x = residual.sugar) ) +
  geom_histogram(fill = "pink")

p6 <- ggplot(data = wine, aes(x = chlorides) ) +
  theme(axis.text = element_text(size = 11)) +
  geom_histogram(fill = "pink")

p7 <- ggplot(data = wine, aes(x = free.sulfur.dioxide) ) +
  theme(axis.text = element_text(size=9)) +
  geom_histogram(fill = "pink")

p8 <- ggplot(data = wine, aes(x = total.sulfur.dioxide) ) +
  theme(axis.text = element_text(size=9)) +
  geom_histogram(fill = "pink")
```

```

p9 <- ggplot(data = wine, aes(x = density) ) +
  theme(axis.text = element_text(size = 7.5)) +
  geom_histogram(fill= "pink")

p10 <- ggplot(data = wine, aes(x = pH) ) +
  geom_histogram(fill = "pink")

p11 <- ggplot(data = wine, aes(x = sulphates) ) +
  theme(axis.text = element_text(size=9)) +
  geom_histogram(fill= "pink")

p12 <- ggplot(data = wine, aes(x = alcohol) ) +
  geom_histogram(fill= "pink")

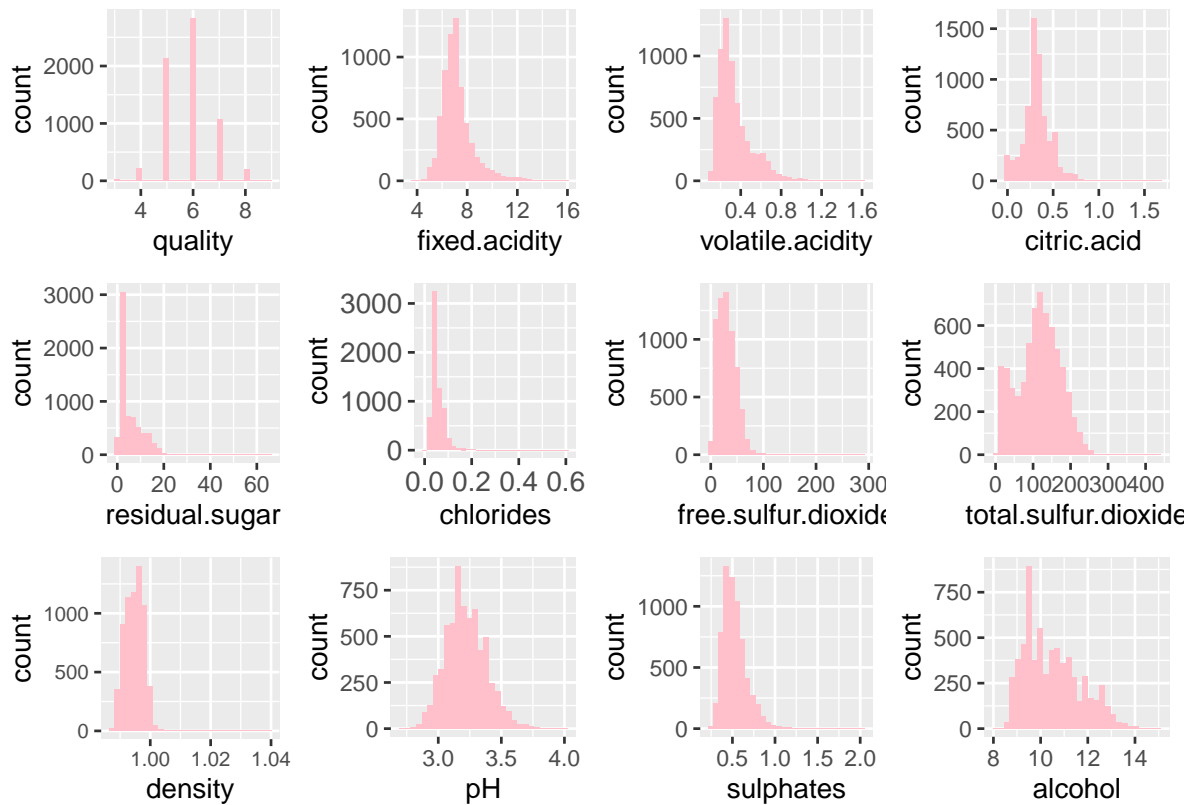
plot_grid(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, p12, ncol = 4, nrow = 3)

```

```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```



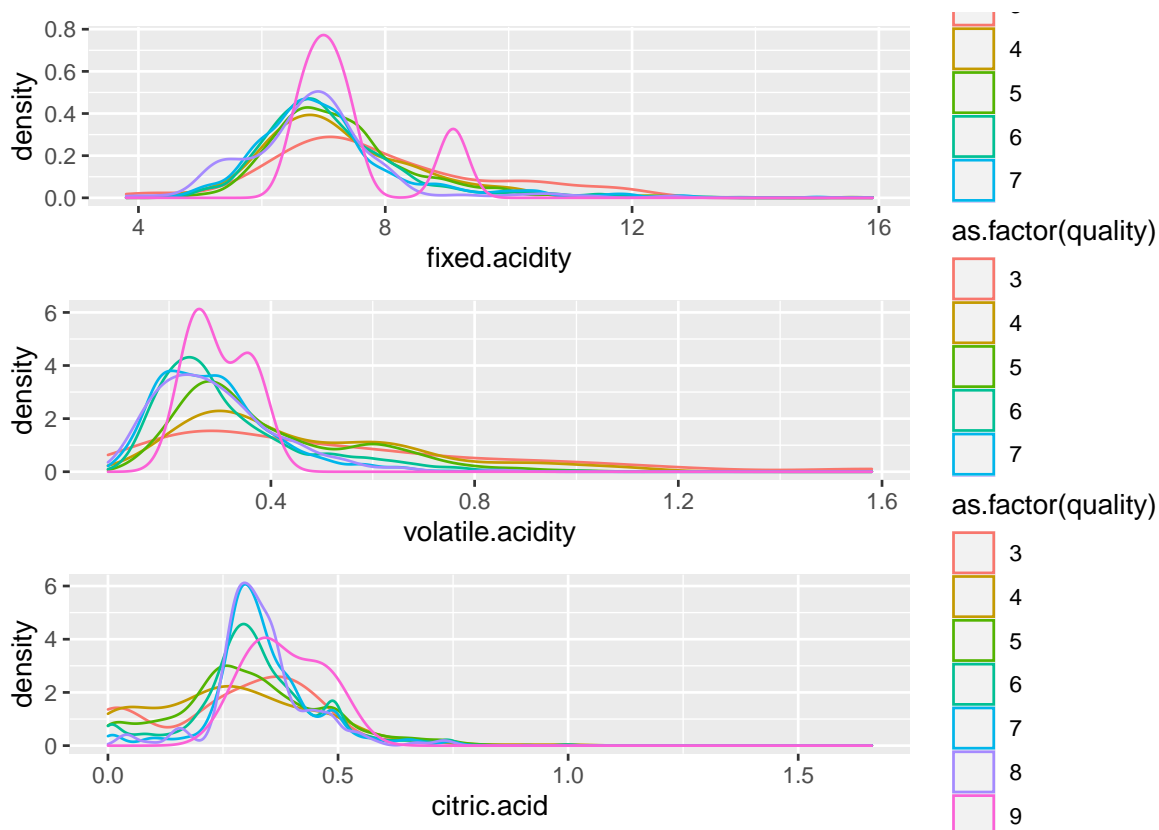
Most of the variables are normally distributed. Variables like fixed.acidity, volatile.acidity, citric.acid, residual.sugar, free.sulfur.dioxide, total.sulfur.dioxide, sulphates, and alcohol are right-skewed.

```
d1 <- ggplot(wine, aes(x = fixed.acidity, color = as.factor(quality))) +
  geom_density()

d2 <- ggplot(wine, aes(x = volatile.acidity, color = as.factor(quality))) +
  geom_density()

d3 <- ggplot(wine, aes(x = citric.acid, color = as.factor(quality))) +
  geom_density()

plot_grid(d1, d2, d3, ncol = 1, nrow = 3)
```



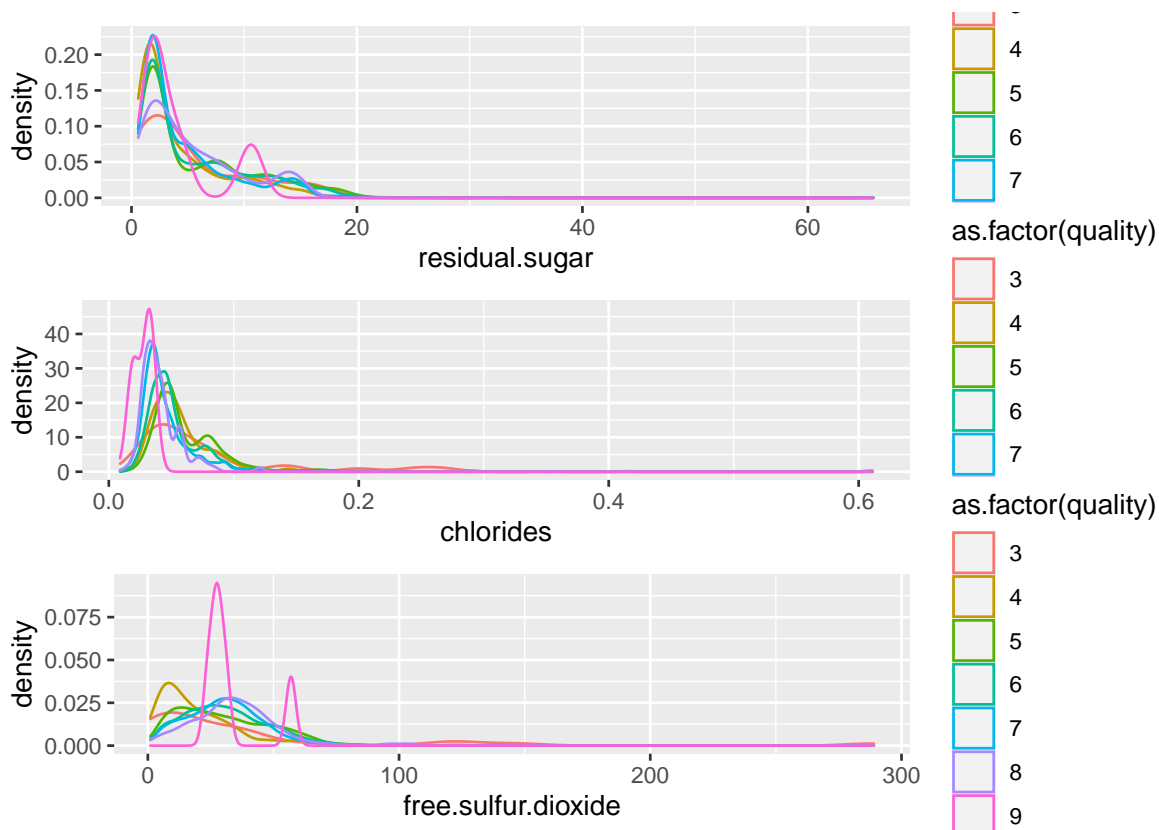
Wine with quality points of 9 has the highest peak of density of fixed.acidity at approximate 7 g/dm³; wine with quality points of 9 has the highest peak of density of volatile.acidity at approximate 0.3 g/dm³; red wine with quality points of 9 has the highest peak of density of citric.acid at approximate 0.03 g/dm³.

```
d4 <- ggplot(wine, aes(x = residual.sugar, color = as.factor(quality))) +
  geom_density()

d5 <- ggplot(wine, aes(x = chlorides, color = as.factor(quality))) +
  geom_density()

d6 <- ggplot(wine, aes(x = free.sulfur.dioxide, color = as.factor(quality))) +
  geom_density()

plot_grid(d4, d5, d6, ncol = 1, nrow = 3)
```



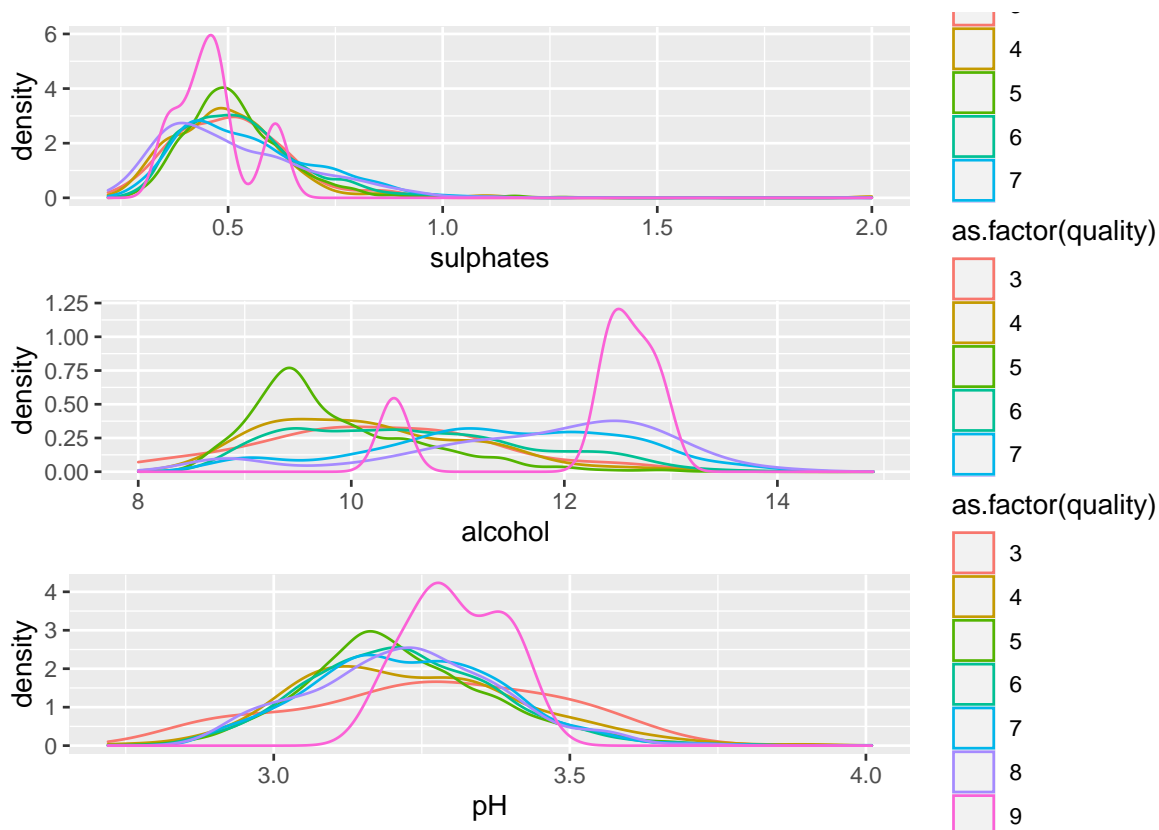
wine with quality points of 9 has the highest peak of density of chlorides at approximate 0.07 g/dm³.

```
d7 <- ggplot(wine, aes(x = sulphates, color = as.factor(quality))) +
  geom_density()

d8 <- ggplot(wine, aes(x = alcohol, color = as.factor(quality))) +
  geom_density()

d9 <- ggplot(wine, aes(x = pH, color = as.factor(quality))) +
  geom_density()

plot_grid(d7, d8, d9, ncol = 1, nrow = 3)
```

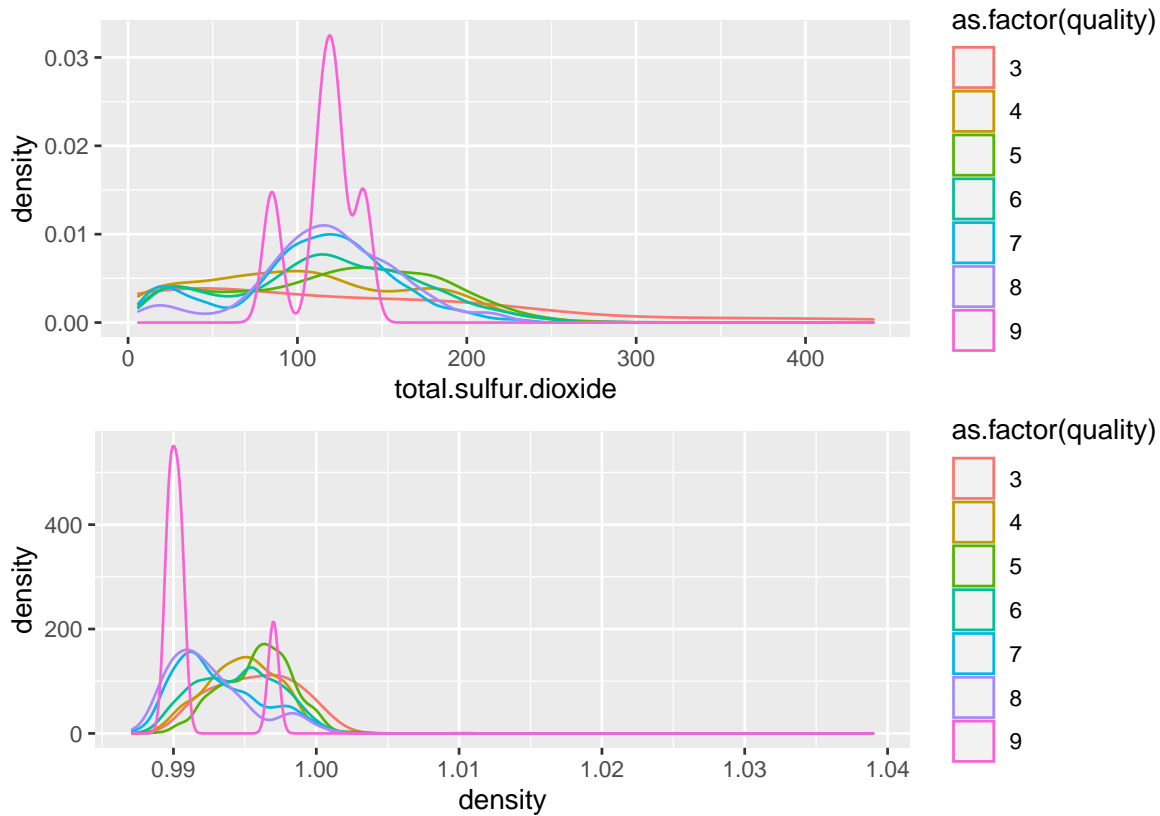


Red wine with quality points of 9 has the highest peak of density of alcohol at approximate 13 vol.

```
d10 <- ggplot(wine, aes(x = total.sulfur.dioxide, color = as.factor(quality))) +
  geom_density()

d11 <- ggplot(wine, aes(x = density, color = as.factor(quality))) +
  geom_density()

plot_grid(d10, d11, ncol = 1, nrow = 2)
```

Red wine with quality points of 9 has the highest peak of density of the density of the liquid at approximate 0.99 g/cm^3 .

Analysis approach

First, we will make visualizations and calculate summary statistics as part of exploratory data analysis. This will give us a better idea of which predictor variables we should focus on. After visualizing the relationships between our quality (the outcome variable) and the other predictor variables, alcohol and density seem to be the strongest predictors for quality of wine. We will also explore the relationship between chemical composition of wine and color of wine.

Since quality is a categorical variable that can take the values 3-9, we will conduct multinomial logistic regression and fit two multiLR models for predicting quality: the first is a reduced model with only alcohol and density as predictors, and the second is a full model. Similarly, since color is a binary variable, we will fit two LR models for predicting color, one reduced and one full. These models will be compared using adjusted R-squared, AIC, and BIC. Then, we will check the conditions for inference. For linearity, we will examine empirical logit plots between each level of the response and the quantitative predictor variables. We will check

randomness and independence based on the context of the data and how the observations were collected.

For prediction, we will build two models for each outcome variable based on our previous evaluations of the relationship between the predictor and response variables, then conduct CV and evaluate which model is preferred. We will then fit the models to the testing data and again evaluate the performance of these models using a confusion matrix and ROC curves. Lastly, we will make predictions for some example observations.

Data dictionary

The data dictionary can be found [here](#).