Proposal

STA 210 - Project

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

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

```
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)
```

```
Joining, by = c("fixed.acidity", "volatile.acidity", "citric.acid",
"residual.sugar", "chlorides", "free.sulfur.dioxide", "total.sulfur.dioxide",
"density", "pH", "sulphates", "alcohol", "quality", "color")
```

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

Introduction

About 234 million hectoliters of wine were consumed in 2020, worldwide, with the US making up approximetly 14% of that consumption. Since Wine composition and wine quality varies widely, it raises the question: what makes a good wine?

To answer that question, we will analyze the wine quality dataset from Vinho Verde vinyard in Portogal, and more importantly try to narrow down our question to make it possible for it to be supported by evidence. Below is the introduction to our research:

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" 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 wine quality. We believe this is important because by understanding what chemical compositions affect wine qualities, it may shed some light in future direction of improving/preserving wine quality.

Out goal is to produce a classification model that best explains how different chemical compositions of the Portuguese "Vinho Verde" wine affects the variation of the wine quality. …

Data description

The Wine Quality dataset was collected from Vinho Verde wine Samples, from the North of Portugal. The data was originally donated in 2009 by Professor Cortez. The specific mechanism of the collection of the data was lab work done on different wines to measure their chemical attributites (like acidity etc.). The quality of the wine however was obtained through the average rating of three wine experts. The dataset is divided into two: Red wine and White wine. Red Wine has 1599 observations, and white wine has 4898 observations (each observation being a specific wine). Information about the wine include but are not limited to:PH,Density,Acidity, and alcohol content.

Each observation is a specific wine from Vinho Verde vinyard. Thus, there might be a little uncertainity in collecting the exact numbers for each numbers. This uncertainity shouldn't be significant in our analysis or project. Thus, we will assume that the values from the

```
glimpse(wine)
```

```
Rows: 6,497
Columns: 13
$ fixed.acidity        <dbl> 6.1, 6.6, 6.9, 8.5, 7.4, 6.0, 6.2, 7.8, 8.3, 6.4,~
$ volatile.acidity     <dbl> 0.340, 0.340, 0.400, 0.240, 0.250, 0.395, 0.220, ~
$ citric.acid          <dbl> 0.29, 0.27, 0.43, 0.39, 0.36, 0.00, 0.20, 0.30, 0~
$ residual.sugar       <dbl> 2.20, 6.20, 6.20, 10.40, 13.20, 1.40, 20.80, 1.80~
$ chlorides            <dbl> 0.036, 0.059, 0.065, 0.044, 0.067, 0.042, 0.035, ~
$ free.sulfur.dioxide  <dbl> 25, 23, 42, 20, 53, 7, 58, 43, 11, 44, 53, 33, 36~
$ total.sulfur.dioxide <dbl> 100, 136, 178, 142, 178, 55, 184, 179, 24, 140, 1~
$ density              <dbl> 0.98938, 0.99570, 0.99552, 0.99740, 0.99760, 0.99~
$ pH                   <dbl> 3.06, 3.30, 3.11, 3.20, 3.01, 3.37, 3.11, 3.43, 3~
$ sulphates            <dbl> 0.44, 0.49, 0.53, 0.53, 0.48, 0.38, 0.53, 0.41, 0~
$ alcohol              <dbl> 11.8, 10.1, 9.4, 10.0, 9.0, 11.2, 9.0, 9.0, 12.1,~
$ quality              <int> 6, 6, 5, 6, 6, 4, 6, 5, 7, 7, 6, 6, 6, 5, 6, 5, 5~
```

```
$ color                <chr> "white", "white", "white", "white", "white", "whi~
```

There are 6497 observations and 13 variables (14 if you include the new response variable added later).

```
any(is.na(wine))
```

```
[1] FALSE
```

There are no NAs in our data, so we shouldn't be concerned about missing data.

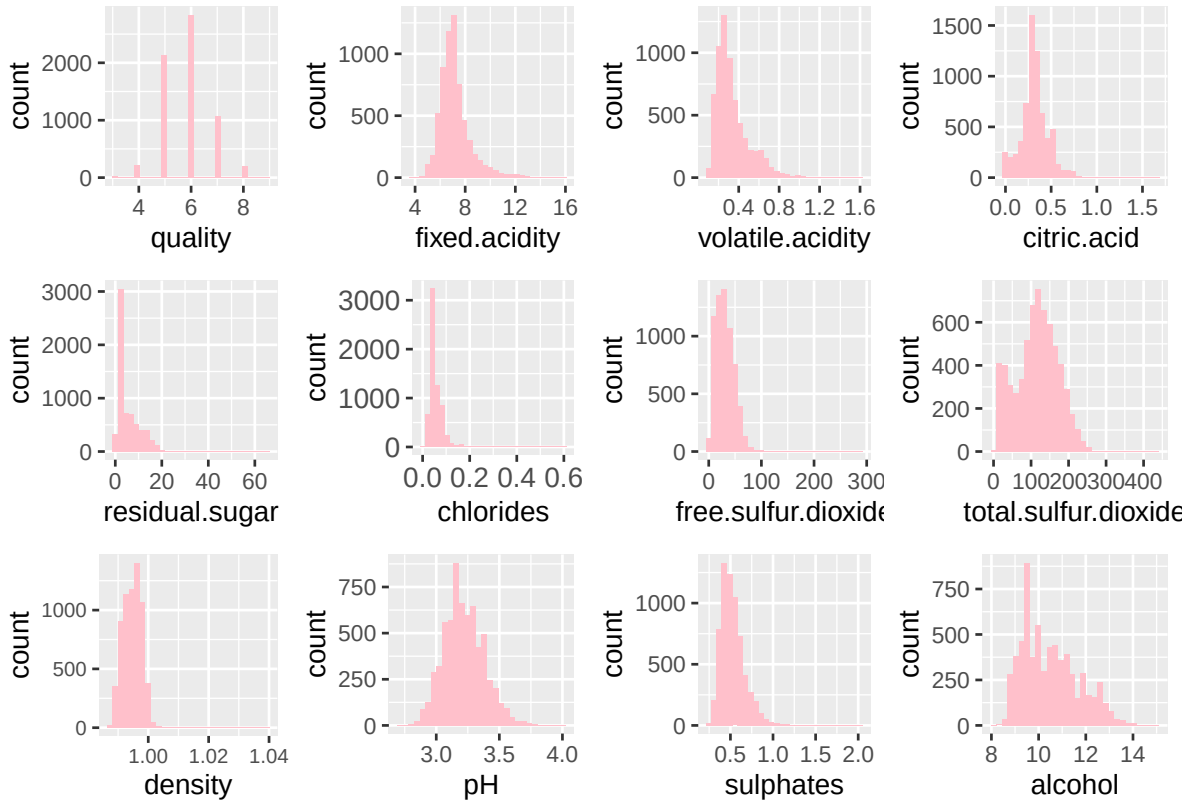
```
summary(wine)
```

```
 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
```

Here are some important summary statistics that might be useful in our project

```
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)
```



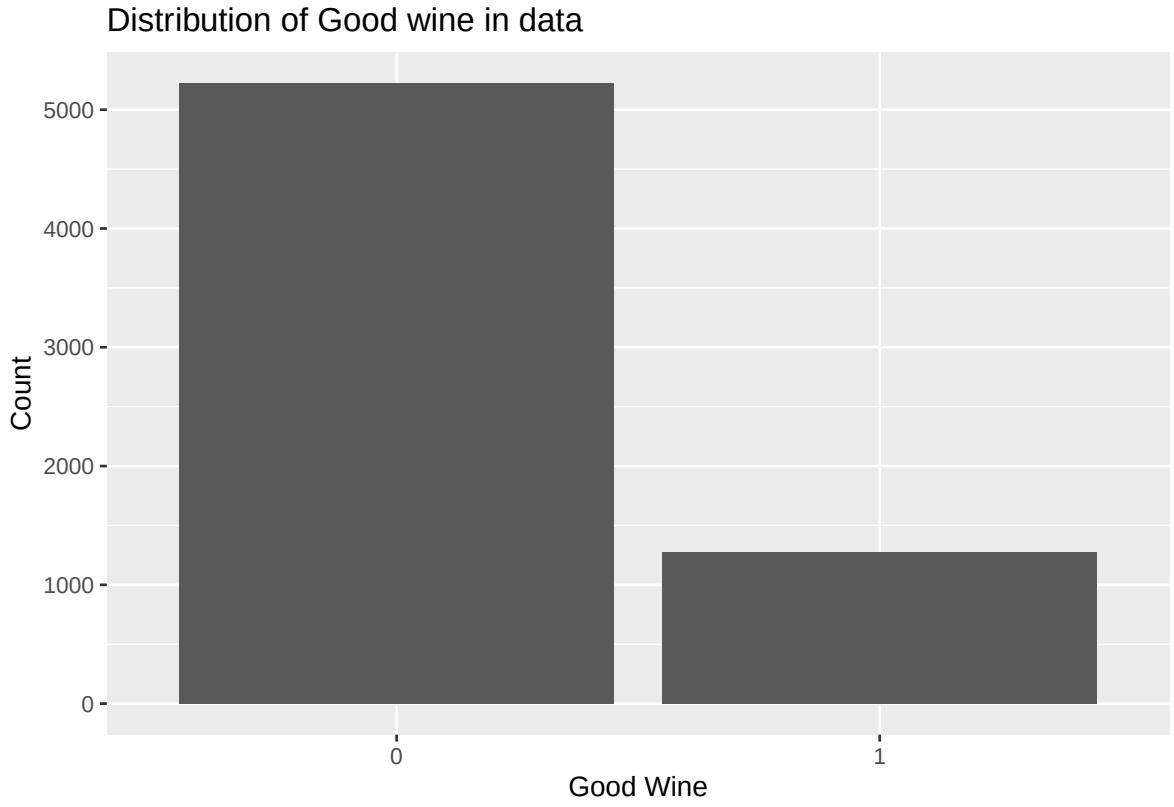
...

Analysis approach

Our response variable will be based on the quality rating of each wine. We will divide wines into two categeries: "bad or subpar wine" and "Good wine". "Bad or subpar wine" will be wine with a quality rating lower then 7 and "Good wine" is any wine with a quality rating equal or greater than 7. Thus, our response variable will be a categrical variable based on the "quality" variable with the responses:"bad or subpar wine" and "Good wine".

```
# Creating the categrical factor and visualizing it
wine<-wine%>%
```

```
  mutate(good_wine=if_else(quality>=7,"1","0"))
wine<-wine%>%
  mutate(good_wine_names=if_else(good_wine=="1","Good wine","Bad or subpar wine"))
ggplot(wine,aes(x=good_wine))+
  geom_bar()+
  labs(title="Distribution of Good wine in data",
       y="Count",
       x="Good Wine"
  )
```



As we can observe the sample is unbalanced with respect to good wine.

```
# Visualizing the response variable with respect to white and red wine.
wine %>%
  count(color, good_wine_names) %>%
  pivot_wider(names_from = good_wine_names, values_from = n) %>%
  kable()
```

color	Bad or subpar wine	Good wine
red	1382	217
white	3838	1060

...

All variables other than "quality" and "good_wine" will be used in in our model as predicators:11 numerical predicators and 1 categrical predicator.

The Following is our Project Plan:

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 good_wine (the outcome variable) and the other predictor variables, alcohol and density seem to be the strongest predictors for quality of wine (good_wine). We will also explore the relationship between the color of wine and it's quality .

Since good_wine is a categorical variable that can take the values "1" and "0", we will conduct logistic regression and fit two LR models for predicting quality: the first is a full model and the second is a reduced model that accounts for collinearity. 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.