Draft-1

STA 210 - Summer 2022

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

6-8-2022

Setup

```
library(tidyverse)
library(tidymodels)
library(dplyr)
library(ggplot2)
library(cowplot)
library(knitr)
library(recipes)
library(caret)
library(InformationValue)
library(ISLR)
library(MASS)
library(nnet)
```

```
redwine <- read.csv("data/winequality-red.csv", sep = ";")
whitewine <- read.csv("data/winequality-white.csv", sep = ";")
redwine <- redwine %>% mutate(color="red")
whitewine <- whitewine %>% mutate(color="white")
wine <- redwine %>% full_join(whitewine)
wine <- slice(wine, sample(1:n()))
```

Load packages and data:

Introduction and Data

Introduction About 234 million hectoliters of wine were consumed in 2020, worldwide, with the US making up approximetly 14% of that consumption (Karlsson 2020). 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. "Vinho Verde" is the kind of wine exclusively produced in the demarcated region of Vinho Verde in northwestern Portugal. The Vinho Verde wine has its own unique system of production and is only produced from the indigenous grape varieties of the region. Vinho Verde region is one of the largest and oldest wine regions in the world, and is home to thousands of producers, generating a wealth of economic activity and jobs, and strongly contributes to the development of Minho province and the country. The Vinho Verde wine also enjoys high reputation worldwide. It is recurrently awarded in national and international competitions. 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 that this dataset can also be used to analyze what chemical factors are attributable to the final rating of Vinho Verde wine. Our research may shed light on future research and development directions for improving the quality of Vinho Verde wine, which may also contribute to the competitiveness of Portuguese wine industry.

Our 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 Introduction 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 the Vinho Verde region. Thus, there might be a little uncertainity in collecting the exact numbers for each numbers. However, since the Vinho Verde region is a vast region spreading 15500 hectareas of vineyards in far-north Portugal, this uncertainity shouldn't be significant in our analysis or project. Thus, we will assume that the datas are independent and random

```
glimpse(wine)
```

```
Rows: 6,497
Columns: 13
$ fixed.acidity        <dbl> 9.8, 7.0, 6.4, 5.2, 7.1, 6.2, 7.3, 7.3, 8.0, 7.7,~
$ volatile.acidity     <dbl> 0.39, 0.23, 0.26, 0.37, 0.28, 0.37, 0.23, 0.21, 0~
$ citric.acid          <dbl> 0.43, 0.32, 0.21, 0.33, 0.35, 0.24, 0.37, 0.21, 0~
$ residual.sugar       <dbl> 1.65, 1.80, 8.20, 1.20, 3.50, 6.10, 1.80, 1.60, 1~
$ chlorides            <dbl> 0.068, 0.048, 0.050, 0.028, 0.028, 0.032, 0.032, ~
$ free.sulfur.dioxide  <dbl> 5.0, 25.0, 51.0, 13.0, 35.0, 19.0, 60.0, 35.0, 40~
$ total.sulfur.dioxide <dbl> 11, 113, 182, 81, 91, 86, 156, 133, 131, 27, 127,~
$ density              <dbl> 0.99478, 0.99150, 0.99542, 0.99020, 0.99022, 0.98~
$ pH                   <dbl> 3.19, 3.11, 3.23, 3.37, 2.96, 3.04, 3.11, 3.38, 3~
$ sulphates            <dbl> 0.46, 0.47, 0.48, 0.38, 0.33, 0.26, 0.35, 0.46, 0~
$ alcohol              <dbl> 11.4, 11.1, 9.5, 11.7, 12.1, 13.4, 11.1, 10.0, 10~
$ quality              <int> 5, 6, 5, 6, 5, 8, 6, 6, 5, 5, 6, 5, 6, 8, 7, 6, 7~
$ color                <chr> "red", "white", "white", "white", "white", "white~
```

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

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

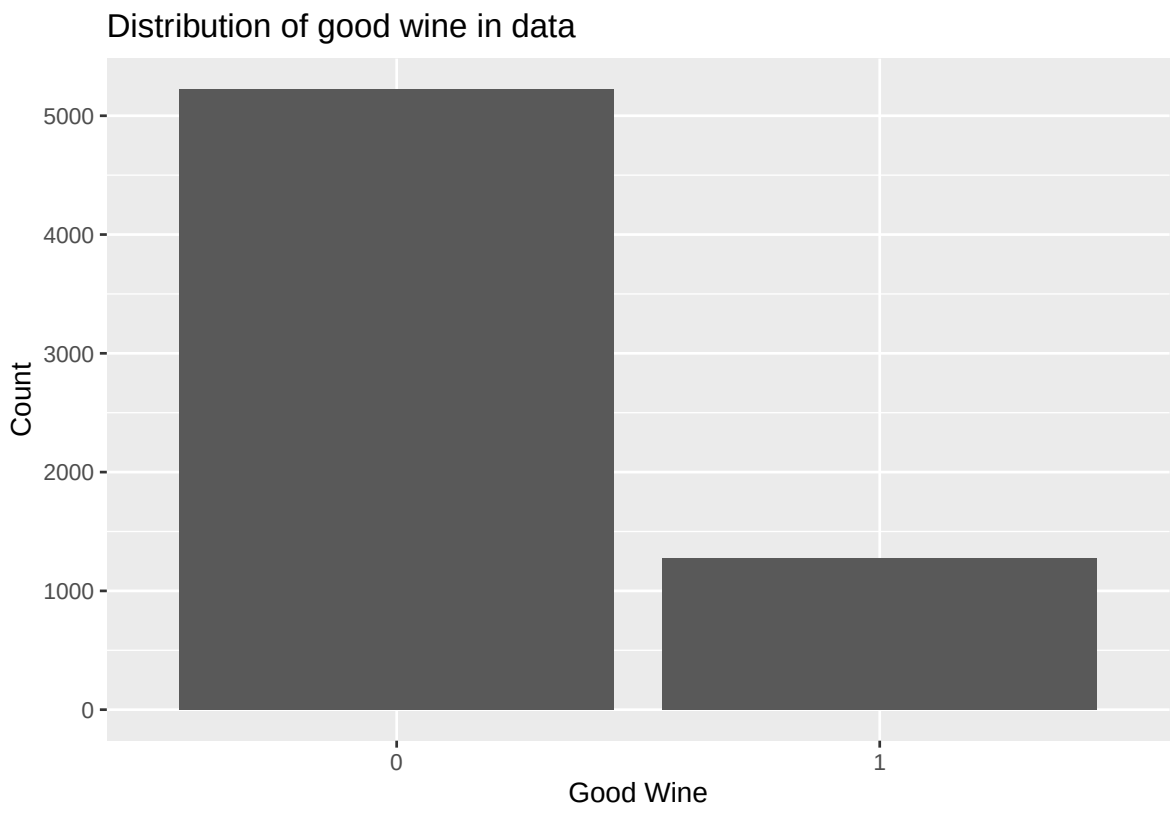
```
[1] FALSE
```

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

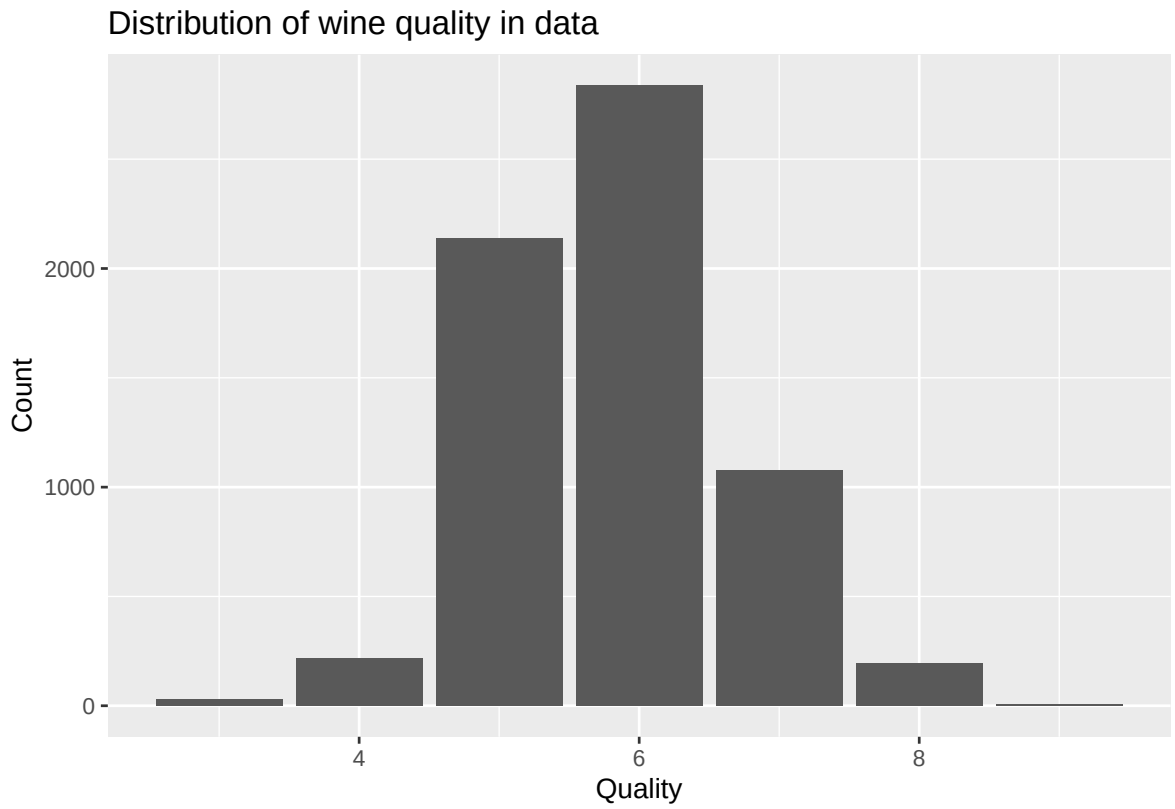
```
wine <- wine %>%
  mutate(good_wine = if_else(quality >= 7,"1","0"))
wine <- wine %>%
  mutate(good_wine = as.factor(good_wine))
wine <- wine %>%
  mutate(good_wine_names = if_else(good_wine=="1","Good wine","Bad or subpar wine"))

no1 <- colnames(wine)[1:11]
colnames(wine)[1:11] = paste("c_", no1, sep = "")
```

Data Wrangling

```
ggplot(wine, aes(x = good_wine)) +
  geom_bar() +
  labs(title = "Distribution of good wine in data",
       y = "Count",
       x = "Good Wine"
  )
```




```
ggplot(wine, aes(x = quality)) +
  geom_bar() +
  labs(title = "Distribution of wine quality in data",
       y = "Count",
       x = "Quality"
  )
```



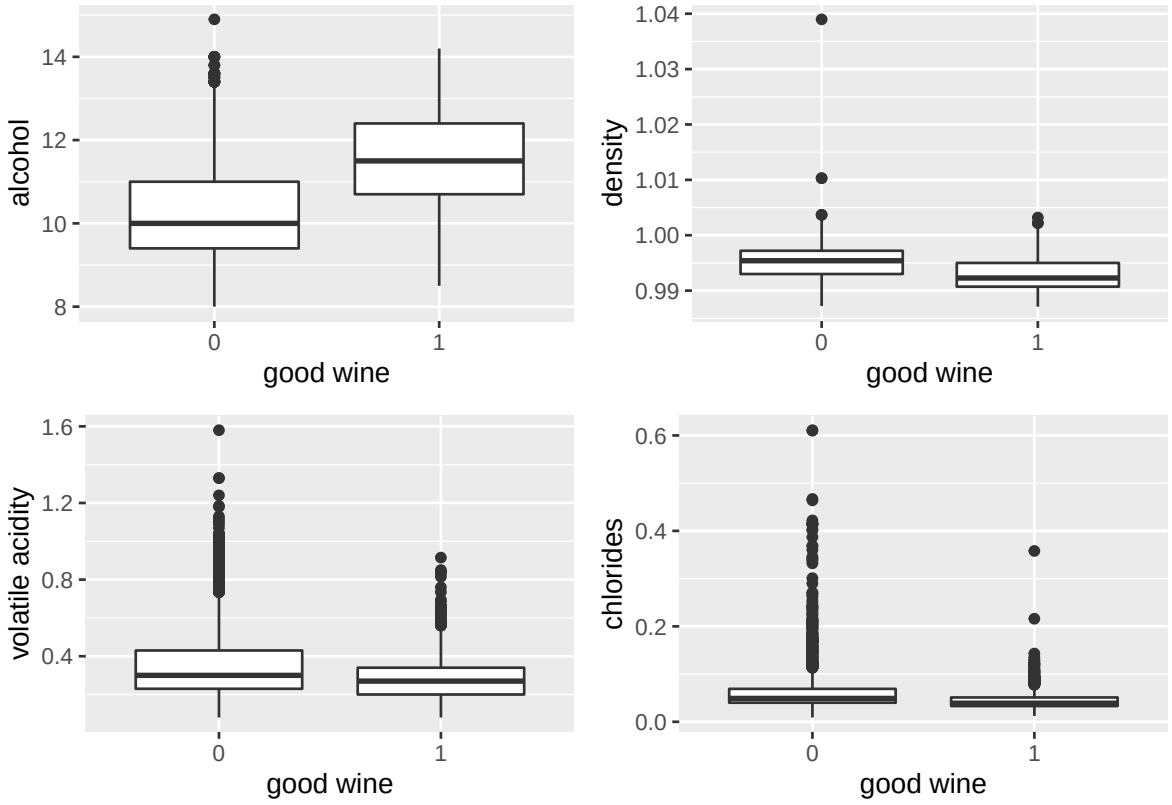

```
p1  <- ggplot(data = wine, aes(x = quality) ) +
  geom_bar(fill = "pink") +
  labs(x = "quality")

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

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

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

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

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

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

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

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

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

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

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

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


```
#cor(wine$c_alcohol, wine$quality)
#cor(wine$c_density, wine$quality)
#cor(wine$c_volatile.acidity, wine$quality)
#cor(wine$c_chlorides, wine$quality)
#cor(wine$c_residual.sugar, wine$quality)
#cor(wine$c_fixed.acidity, wine$quality)
#cor(wine$c_free.sulfur.dioxide, wine$quality)
#cor(wine$c_total.sulfur.dioxide, wine$quality)
#cor(wine$c_pH, wine$quality)
#cor(wine$c_sulphates, wine$quality)
# cor(wine$c_citric.acid, wine$quality)
```

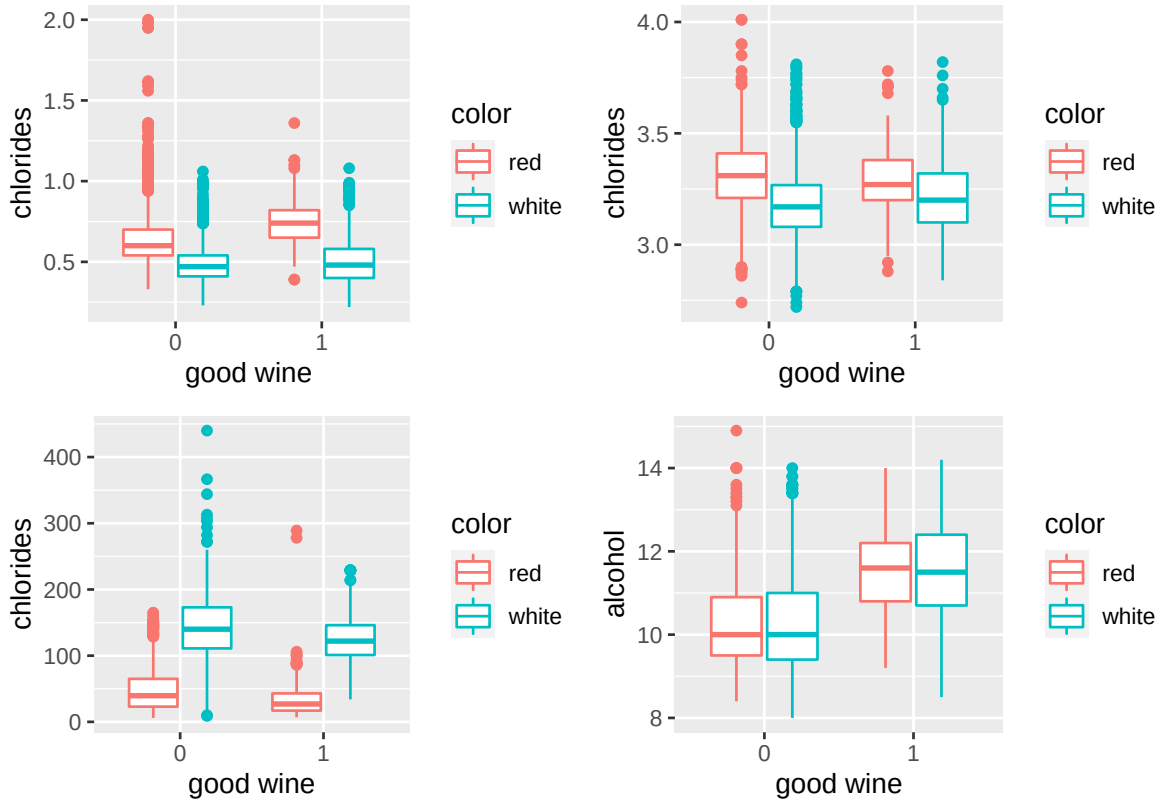
```
a1 <- ggplot(wine, aes(y = c_alcohol, x = as.factor(good_wine))) +
  geom_boxplot() +
  labs(x = "good wine", y = "alcohol")
```

```
a2 <- ggplot(wine, aes(y = c_density, x = as.factor(good_wine))) +
  geom_boxplot() +
  labs(x = "good wine", y = "density")

a3 <- ggplot(wine, aes(y = c_volatile.acidity, x = as.factor(good_wine))) +
  geom_boxplot() +
  labs(x = "good wine", y = "volatile acidity")

a4 <- ggplot(wine, aes(y = c_chlorides, x = as.factor(good_wine))) +
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

plot_grid(a1, a2, a3, a4, ncol = 2, nrow = 2)
```

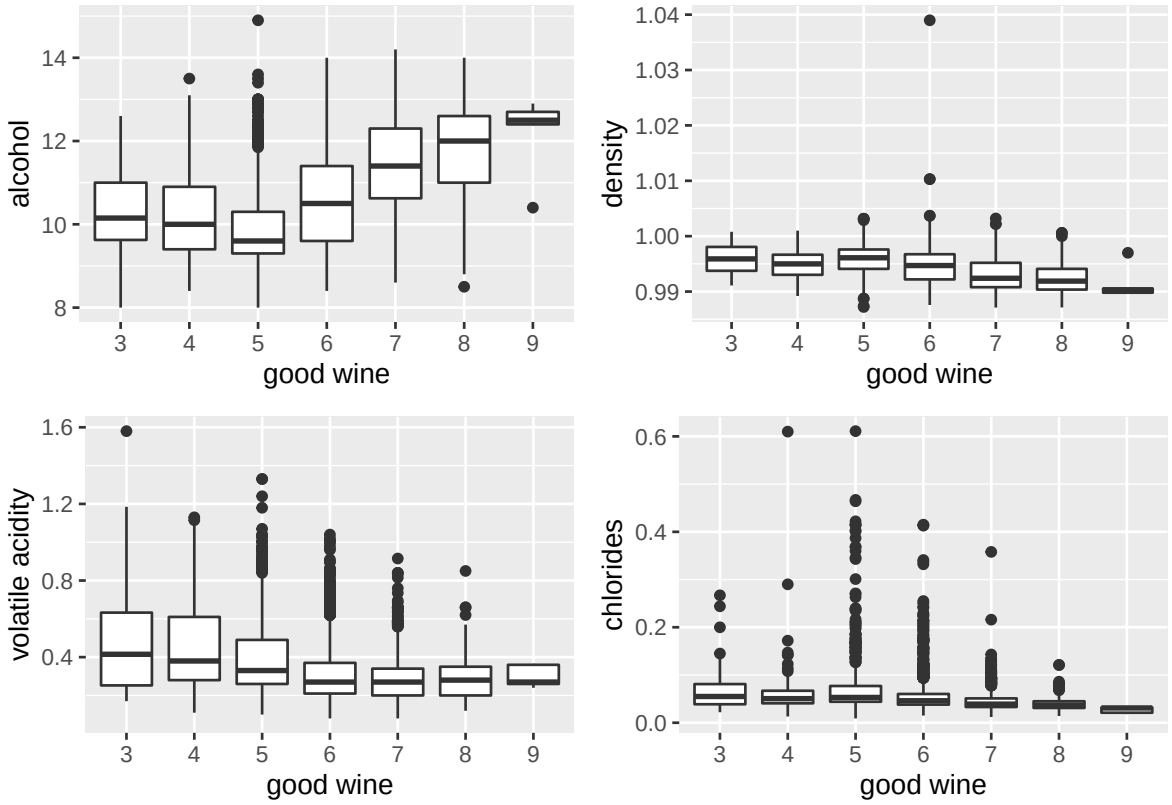



```
# exploring interaction effects
a5 <- ggplot(wine, aes(y = c_sulphates, x = as.factor(good_wine), color = color)) +
```

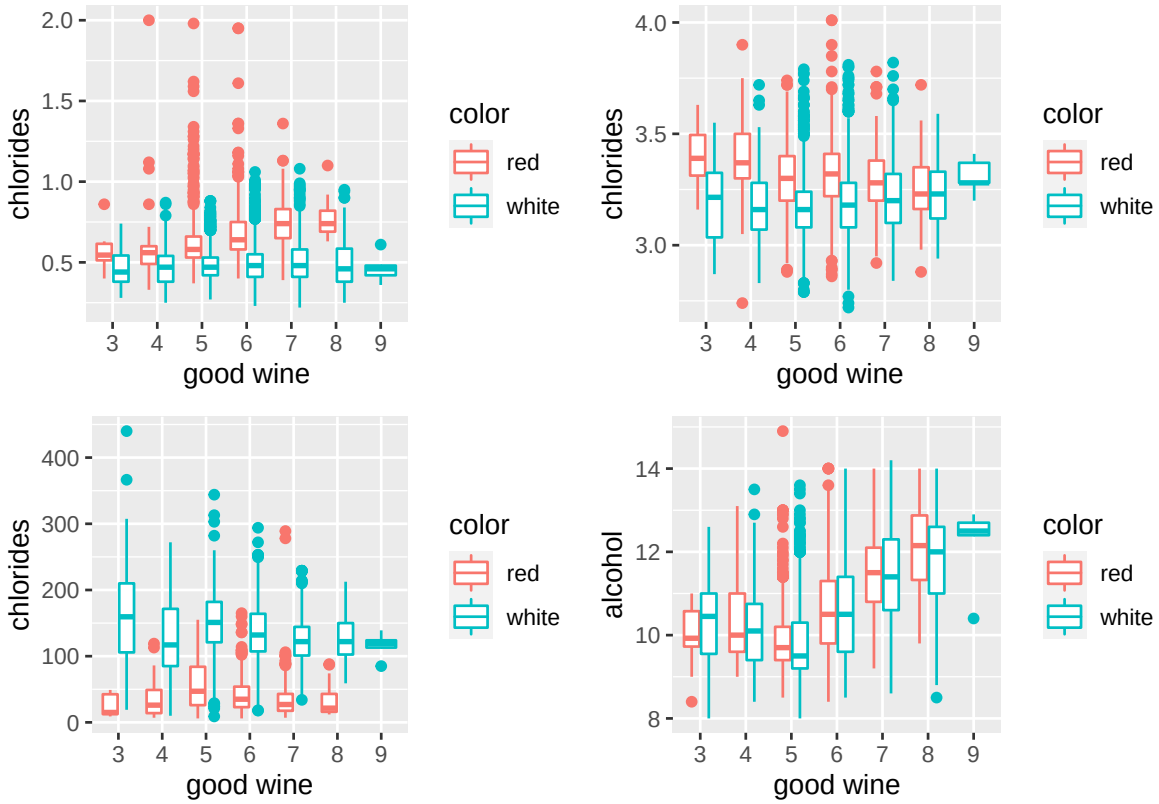
```
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

a6 <- ggplot(wine, aes(y = c_pH, x = as.factor(good_wine), color = color)) +
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

a7 <- ggplot(wine, aes(y = c_total.sulfur.dioxide, x = as.factor(good_wine), color = color))
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

a8 <- ggplot(wine, aes(y = c_alcohol, x = as.factor(good_wine), color = color)) +
  geom_boxplot() +
  labs(x = "good wine", y = "alcohol")

plot_grid(a5, a6, a7, a8, ncol = 2, nrow = 2)
```



```r
a1 <- ggplot(wine, aes(y = c_alcohol, x = as.factor(quality))) +
  geom_boxplot() +
  labs(x = "good wine", y = "alcohol")

a2 <- ggplot(wine, aes(y = c_density, x = as.factor(quality))) +
  geom_boxplot() +
  labs(x = "good wine", y = "density")

a3 <- ggplot(wine, aes(y = c_volatile.acidity, x = as.factor(quality))) +
  geom_boxplot() +
  labs(x = "good wine", y = "volatile acidity")

a4 <- ggplot(wine, aes(y = c_chlorides, x = as.factor(quality))) +
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

plot_grid(a1, a2, a3, a4, ncol = 2, nrow = 2)
```



```r
# exploring interaction effects
a5 <- ggplot(wine, aes(y = c_sulphates, x = as.factor(quality), color = color)) +
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

a6 <- ggplot(wine, aes(y = c_pH, x = as.factor(quality), color = color)) +
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

a7 <- ggplot(wine, aes(y = c_total.sulfur.dioxide, x = as.factor(quality), color = color)) +
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

a8 <- ggplot(wine, aes(y = c_alcohol, x = as.factor(quality), color = color)) +
  geom_boxplot() +
  labs(x = "good wine", y = "alcohol")

plot_grid(a5, a6, a7, a8, ncol = 2, nrow = 2)
```



Exploratory Data Analysis

Methodology

```
set.seed(222)

wine_split <- initial_split(wine, prop = 3/4)
wine_train <- training(wine_split)
wine_test <- training(wine_split)
```

```
wine_spec <- logistic_reg() %>% set_engine("glm")
```

```
wine_rec1 <- recipe(
  good_wine ~., data = wine_train) %>%
```

```
  step_rm(good_wine_names, quality) %>%
  step_center(all_numeric_predictors())%>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors())
```

```
wine_wflow1 <- workflow() %>%
  add_model(wine_spec) %>%
  add_recipe(wine_rec1)
```

```
wine_fit1 <- wine_wflow1 %>%
  fit(data = wine_train)
kable(tidy(wine_fit1), digits = 3)
```

Logistic Model: Reduced

term	estimate	std.error	statistic	p.value
(Intercept)	-1.454	0.219	-6.636	0.000
c_fixed.acidity	0.482	0.078	6.207	0.000
c_volatile.acidity	-3.878	0.459	-8.441	0.000
c_citric.acid	-0.586	0.404	-1.452	0.146
c_residual.sugar	0.210	0.031	6.879	0.000
c_chlorides	-8.934	3.118	-2.866	0.004
c_free.sulfur.dioxide	0.013	0.004	3.524	0.000
c_total.sulfur.dioxide	-0.004	0.002	-2.711	0.007
c_density	-388.101	77.362	-5.017	0.000
c_pH	2.442	0.419	5.823	0.000
c_sulphates	2.389	0.328	7.291	0.000
c_alcohol	0.507	0.094	5.407	0.000
color_white	-0.682	0.286	-2.380	0.017

Since the p-value of the citric acid coefficient is well above our significance level of 0.05, we perform an Anova test:

```
wine_rec2 <- recipe(
  good_wine ~., data = wine_train) %>%
  step_rm(good_wine_names,quality,c_citric.acid) %>%
  step_center(all_numeric_predictors()) %>%
  step_dummy(all_nominal_predictors()) %>%
```

```
  step_zv(all_predictors())

wine_wflow2 <- workflow() %>%
  add_model(wine_spec) %>%
  add_recipe(wine_rec2)

wine_fit2 <- wine_wflow2 %>%
  fit(data = wine_train)
```

```
fit_engine1 <- extract_fit_engine(wine_fit1)
fit_engine2 <- extract_fit_engine(wine_fit2)

anova(fit_engine2, fit_engine1, test = "Chisq") %>%
    kable(digits = 3)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
4860	3742.045	NA	NA	NA
4859	3739.911	1	2.134	0.144

Based on these results, we should remove citric acid.

```
wine_rec_full <- recipe(good_wine ~., data = wine_train) %>%
  step_rm(good_wine_names,quality,c_citric.acid) %>% # i don't think we should remove citric
  step_dummy(color) %>%
  step_interact(terms = ~starts_with("c_"):starts_with("color")) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors())
```

```
wine_flow_model <- workflow() %>%
  add_model(wine_spec) %>%
  add_recipe(wine_rec_full)
```

```
wine_fit_test <- wine_flow_model %>%
  fit(data = wine_train)
```

```
tidy(wine_fit_test,conf.int = T) %>%
  kable(digits = 3)
```

Full Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	201.576	128.917	1.564	0.118	-51.501	454.769
c_fixed.acidity	0.274	0.146	1.883	0.060	-0.013	0.560
c_volatile.acidity	-3.150	0.788	-3.999	0.000	-4.728	-1.637
c_residual.sugar	0.215	0.099	2.182	0.029	0.010	0.400
c_chlorides	-8.150	4.372	-1.864	0.062	-17.766	-0.805
c_free.sulfur.dioxide	0.017	0.015	1.091	0.275	-0.013	0.046
c_total.sulfur.dioxide	-0.019	0.006	-2.991	0.003	-0.032	-0.007
c_density	-217.245	131.599	-1.651	0.099	-475.716	41.064
c_pH	0.215	1.174	0.183	0.855	-2.109	2.500
c_sulphates	3.492	0.639	5.462	0.000	2.224	4.746
c_alcohol	0.874	0.152	5.759	0.000	0.581	1.177
color_white	447.773	168.987	2.650	0.008	117.678	780.483
c_fixed.acidity_x_color_white	0.265	0.179	1.484	0.138	-0.084	0.618
c_volatile.acidity_x_color_white	-0.456	0.967	-0.471	0.637	-2.336	1.458
c_residual.sugar_x_color_white	0.084	0.107	0.783	0.434	-0.118	0.304
c_chlorides_x_color_white	-6.861	6.287	-1.091	0.275	-18.848	5.891
c_free.sulfur.dioxide_x_color_white	-0.006	0.016	-0.376	0.707	-0.037	0.025
c_total.sulfur.dioxide_x_color_white	0.018	0.007	2.713	0.007	0.006	0.031
c_density_x_color_white	-454.672	171.983	-2.644	0.008	-793.257	-118.703
c_pH_x_color_white	3.076	1.273	2.416	0.016	0.595	5.593
c_sulphates_x_color_white	-1.252	0.753	-1.662	0.096	-2.729	0.235
c_alcohol_x_color_white	-0.770	0.201	-3.834	0.000	-1.168	-0.380

As we can see from some variables p values and confidence interval, we can drop some of those valuables if we were to conduct to a hypothesis test since their p value would exceed 0.05, meaning that we would not have enough to reject the null hypothesis. (better wording later)

```
wine_full_reduced <- recipe(good_wine ~., data = wine_train) %>%
  step_rm(good_wine_names, quality, c_citric.acid) %>%
  step_dummy(color)%>%
  step_interact(terms = ~starts_with("c_"):starts_with("color")) %>%
  step_rm(c_sulphates_x_color_white, c_free.sulfur.dioxide_x_color_white, c_chlorides_x_colo
```

```
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors())
```

```
wine_full_reduced_workflow<- workflow() %>%
  add_model(wine_spec) %>%
  add_recipe(wine_full_reduced)

wine_fit_test <- wine_full_reduced_workflow %>%
  fit(data = wine_train)

tidy(wine_fit_test,conf.int = T) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	286.536	86.865	3.299	0.001	117.163	458.206
c_fixed.acidity	0.394	0.069	5.732	0.000	0.260	0.530
c_volatile.acidity	-3.361	0.451	-7.461	0.000	-4.254	-2.488
c_residual.sugar	0.259	0.032	8.108	0.000	0.197	0.322
c_chlorides	-12.021	3.278	-3.667	0.000	-18.704	-5.896
c_free.sulfur.dioxide	0.011	0.004	3.196	0.001	0.004	0.018
c_total.sulfur.dioxide	-0.015	0.004	-3.737	0.000	-0.024	-0.008
c_density	-301.199	87.066	-3.459	0.001	-473.285	-131.465
c_sulphates	2.458	0.336	7.306	0.000	1.796	3.115
c_alcohol	0.811	0.124	6.532	0.000	0.571	1.059
color_white	257.245	78.819	3.264	0.001	102.590	411.830
c_total.sulfur.dioxide_x_color_white	0.013	0.004	3.172	0.002	0.005	0.022
c_density_x_color_white	-262.974	78.393	-3.355	0.001	-416.712	-109.152
c_pH_x_color_white	2.736	0.402	6.806	0.000	1.951	3.528
c_alcohol_x_color_white	-0.585	0.148	-3.955	0.000	-0.880	-0.299

```
AIC_fit <- logistic_reg() %>%
  set_engine("glm") %>%
  fit(good_wine~.-c_citric.acid-quality-good_wine_names, data = wine_train)

AIC_fit <- repair_call(AIC_fit, data = wine_train)
AIC_fit_engine <- AIC_fit %>% extract_fit_engine()
```

```
best_AIC_model <- stepAIC(AIC_fit_engine, direction="forward", trace=FALSE)
```

```
best_AIC_model %>% tidy()
```

Stepwise

```
# A tibble: 12 x 5
   term                    estimate std.error statistic  p.value
   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)             376.       76.1          4.94 7.83e- 7
 2 c_fixed.acidity           0.460     0.0759       6.05 1.43e- 9
 3 c_volatile.acidity       -3.66      0.433       -8.45 3.03e-17
 4 c_residual.sugar          0.212     0.0305       6.93 4.16e-12
 5 c_chlorides              -9.17      3.10        -2.95 3.13e- 3
 6 c_free.sulfur.dioxide     0.0127    0.00357      3.57 3.54e- 4
 7 c_total.sulfur.dioxide   -0.00446   0.00156     -2.87 4.14e- 3
 8 c_density              -396.       77.1         -5.14 2.76e- 7
 9 c_pH                      2.47      0.418        5.91 3.32e- 9
10 c_sulphates               2.37      0.327        7.24 4.59e-13
11 c_alcohol                 0.485     0.0925       5.24 1.57e- 7
12 colorwhite               -0.706     0.286       -2.47 1.34e- 2
```

Multnomial Regression

```
full_fit1 <- multinom_reg() %>%
  set_engine("nnet") %>%
  fit(as.factor(quality)~.-good_wine_names-good_wine, data = wine_train)

full_fit1 <- repair_call(full_fit1, data = wine_train)
tidy(full_fit1)
```

Data Editing for Regression

```
# A tibble: 78 x 6
   y.level term                   estimate std.error statistic   p.value
   <chr>   <chr>                     <dbl>     <dbl>     <dbl>     <dbl>
 1 4       (Intercept)              -2.80      0.0872    -32.1   4.77e-226
```

```
 2 4       c_fixed.acidity         -0.463     0.142      -3.27  1.09e-  3
 3 4       c_volatile.acidity      -1.52      0.549      -2.77  5.54e-  3
 4 4       c_citric.acid            1.47      0.597       2.46  1.39e-  2
 5 4       c_residual.sugar        -0.0849    0.0507     -1.68  9.38e-  2
 6 4       c_chlorides            -19.2       0.0384   -499.    0
 7 4       c_free.sulfur.dioxide   -0.0576    0.0179     -3.22  1.26e-  3
 8 4       c_total.sulfur.dioxide   0.00107   0.00594     0.180 8.57e-  1
 9 4       c_density               20.3       0.0873    233.    0
10 4       c_pH                    -2.02      0.420      -4.81  1.55e-  6
# ... with 68 more rows
```

```
full_fit1_engine <- full_fit1 %>% extract_fit_engine()
newmodel <- stepAIC(full_fit1_engine, direction="both", trace=FALSE)
```

```
tidy(newmodel)
```

```
# A tibble: 72 x 6
   y.level term                   estimate std.error statistic       p.value
   <chr>   <chr>                     <dbl>     <dbl>     <dbl>         <dbl>
 1 4       (Intercept)             -5.63      0.0903    -62.4   0
 2 4       c_fixed.acidity         -0.358     0.137      -2.60  0.00927
 3 4       c_volatile.acidity      -1.77      0.525      -3.38  0.000732
 4 4       c_residual.sugar        -0.0931    0.0454     -2.05  0.0402
 5 4       c_chlorides            -17.0       0.0351   -485.    0
 6 4       c_free.sulfur.dioxide   -0.0543    0.0186     -2.91  0.00357
 7 4       c_total.sulfur.dioxide   0.00268   0.00591     0.454 0.650
 8 4       c_density               21.9       0.0905    242.    0
 9 4       c_pH                    -1.96      0.446      -4.39  0.0000112
10 4       c_sulphates              3.94      0.662       5.95  0.00000000271
# ... with 62 more rows
```

Results

Model selection- Logistic

```
wine_fit1_eg <- wine_fit1 %>% extract_fit_engine()
wine_fit2_eg <- wine_fit_test %>% extract_fit_engine()
```

```r
# 1 = reduced, 2 = full
wine_test_pred1 <- predict(wine_fit1, wine_test, type = "prob") %>%
  bind_cols(wine_test)
wine_test_pred1
```

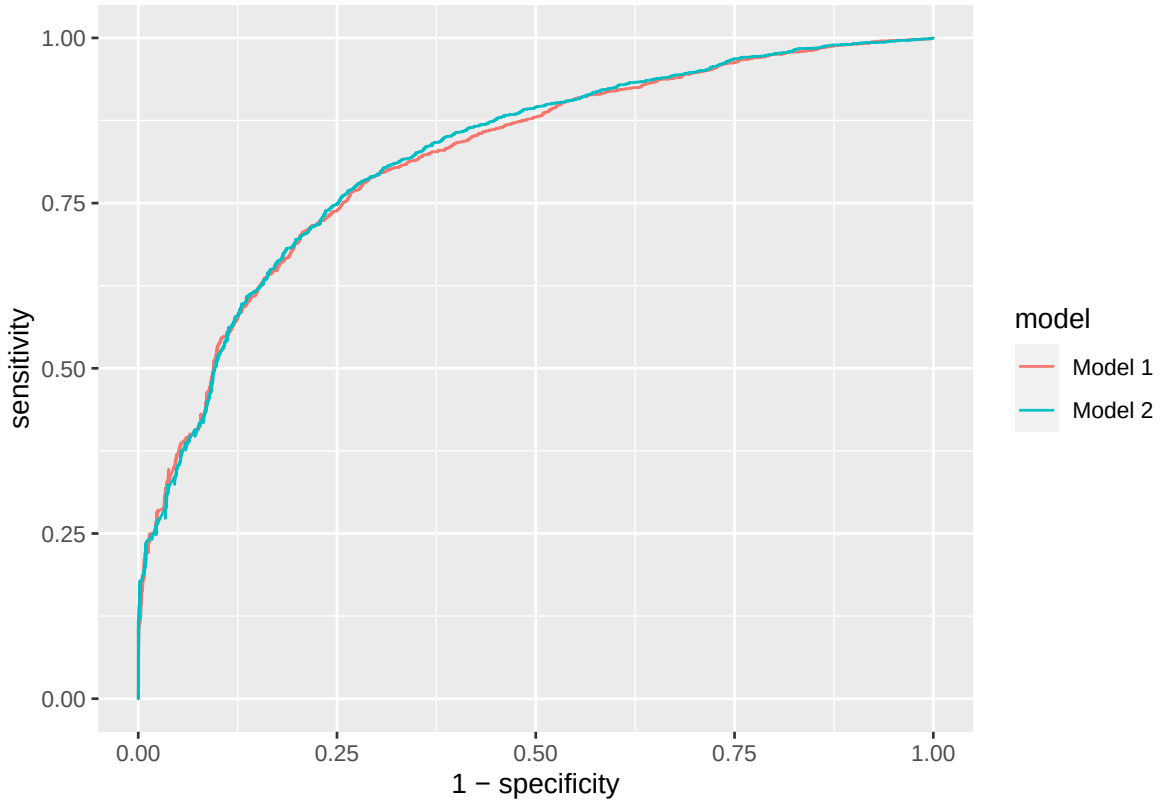
```
# A tibble: 4,872 x 17
   .pred_0 .pred_1 c_fixed.acidity c_volatile.acidity c_citric.acid
     <dbl>   <dbl>           <dbl>              <dbl>         <dbl>
 1   0.989  0.0112             6.6               0.2          0.14
 2   0.846  0.154              6                 0.34         0.29
 3   0.932  0.0684             6.8               0.44         0.37
 4   0.732  0.268              7                 0.29         0.33
 5   0.964  0.0359             6.8               0.25         0.27
 6   0.876  0.124              6.5               0.17         0.33
 7   0.955  0.0451             7.3               0.4          0.3
 8   0.816  0.184              6.2               0.3          0.32
 9   0.971  0.0289             7.8               0.55         0
10   0.580  0.420              6.7               0.24         0.3
# ... with 4,862 more rows, and 12 more variables: c_residual.sugar <dbl>,
#   c_chlorides <dbl>, c_free.sulfur.dioxide <dbl>,
#   c_total.sulfur.dioxide <dbl>, c_density <dbl>, c_pH <dbl>,
#   c_sulphates <dbl>, c_alcohol <dbl>, quality <int>, color <chr>,
#   good_wine <fct>, good_wine_names <chr>
```

```r
wine_test_pred2 <- predict(wine_fit_test, wine_test, type = "prob") %>%
  bind_cols(wine_test)
wine_test_pred2
```

```
# A tibble: 4,872 x 17
   .pred_0 .pred_1 c_fixed.acidity c_volatile.acidity c_citric.acid
     <dbl>   <dbl>           <dbl>              <dbl>         <dbl>
 1   0.992 0.00827             6.6               0.2          0.14
 2   0.843 0.157               6                 0.34         0.29
 3   0.915 0.0847              6.8               0.44         0.37
 4   0.710 0.290               7                 0.29         0.33
 5   0.963 0.0368              6.8               0.25         0.27
 6   0.866 0.134               6.5               0.17         0.33
 7   0.986 0.0142              7.3               0.4          0.3
 8   0.779 0.221               6.2               0.3          0.32
 9   0.978 0.0220              7.8               0.55         0
10   0.573 0.427               6.7               0.24         0.3
```

```
# ... with 4,862 more rows, and 12 more variables: c_residual.sugar <dbl>,
#   c_chlorides <dbl>, c_free.sulfur.dioxide <dbl>,
#   c_total.sulfur.dioxide <dbl>, c_density <dbl>, c_pH <dbl>,
#   c_sulphates <dbl>, c_alcohol <dbl>, quality <int>, color <chr>,
#   good_wine <fct>, good_wine_names <chr>
```

```
wine_test_pred1 %>%
  roc_curve(truth = as.factor(good_wine), .pred_0) %>%
  mutate(model = "Model 1") %>%
  bind_rows(wine_test_pred2 %>%
              roc_curve(truth = as.factor(good_wine), .pred_0) %>%
              mutate(model = "Model 2")) %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity, color = model)) +
  geom_line()
```




```
wine_test_pred1 %>%
  roc_auc(truth = as.factor(good_wine), .pred_0)
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.816
```

```
wine_test_pred2 %>%
  roc_auc(truth = as.factor(good_wine), .pred_0)
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.820
```

Based on the roc_auc, the full model performs slightly better than the reduced model.

```
glance(wine_fit1_eg)$AIC
```

```
[1] 3765.911
```

```
glance(wine_fit2_eg)$AIC
```

```
[1] 3736.542
```

```
glance(wine_fit1_eg)$BIC
```

```
[1] 3850.298
```

```
glance(wine_fit2_eg)$BIC
```

```
[1] 3833.91
```

The full model has lower AIC and BIC.

```
anova(wine_fit1_eg, wine_fit2_eg)%>%tidy()
```

```
# A tibble: 2 x 4
  Resid..Df Resid..Dev    df Deviance
      <dbl>      <dbl> <dbl>    <dbl>
1      4859      3740.    NA     NA
2      4857      3707.     2     33.4
```

Drop-in-Deviance Test:

H_0: the β_j of all the additional interactive terms equal to 0. H_1: at least one additional interactive term have β_j that does not equal 0.

```
pchisq(38.508, 4, lower.tail = FALSE)
```

```
[1] 8.802476e-08
```

The p-value is very small, smaller than the critical value of 0.05 under 95% CI. So we can reject the null hypothesis and conclude that there is enough evidence showing that there is at least 1 beta_j does not equal to 0. The interactive terms have significant effects so we'll choose the full model. ————————————-

```
wine2<-wine%>%mutate(quality=factor(quality,levels=0:10))
```

```
set.seed(22)

wine_split2 <- initial_split(wine2, prop = 3/4)
wine_train <- training(wine_split2)
wine_test <- testing(wine_split2)
```

```
full_fit1<- multinom_reg() %>%
  set_engine("nnet") %>%
  fit(quality~.,
  data = wine_train)
```

```
full_fit1 <- repair_call(full_fit1, data = wine_train)
full_fit1_eg <- full_fit1 %>% extract_fit_engine()
```

```
newmodel <- stepAIC(full_fit1_eg, direction="both",trace=FALSE)
```

```
full_fit1%>%tidy()
```

Multinominal Regression

```
# A tibble: 90 x 6
   y.level term                   estimate std.error statistic   p.value
   <chr>   <chr>                     <dbl>     <dbl>     <dbl>     <dbl>
 1 4       (Intercept)               3.93      0.172      22.9  9.23e-116
 2 4       c_fixed.acidity          -0.612     0.133      -4.61 3.94e-  6
 3 4       c_volatile.acidity       -1.21      0.543      -2.22 2.63e-  2
 4 4       c_citric.acid             0.793     0.488       1.62 1.04e-  1
 5 4       c_residual.sugar         -0.0598    0.0565     -1.06 2.90e-  1
 6 4       c_chlorides             -10.7       0.0870   -123.   0
 7 4       c_free.sulfur.dioxide    -0.0858    0.0147     -5.83 5.57e-  9
 8 4       c_total.sulfur.dioxide    0.00934   0.00745     1.25 2.10e-  1
 9 4       c_density                17.9       0.172     104.   0
10 4       c_pH                     -4.03      0.658      -6.13 8.91e- 10
# ... with 80 more rows
```

```
newmodel%>%tidy()
```

```
# A tibble: 78 x 6
   y.level term                   estimate std.error statistic  p.value
   <chr>   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
 1 4       (Intercept)              22.5       0.171     131.    0
 2 4       c_fixed.acidity          -0.576     0.133      -4.34  1.40e- 5
 3 4       c_volatile.acidity       -1.40      0.543      -2.59  9.69e- 3
 4 4       c_residual.sugar         -0.0523    0.0577     -0.906 3.65e- 1
 5 4       c_chlorides             -11.0       0.0834   -132.    0
 6 4       c_free.sulfur.dioxide    -0.0843    0.0147     -5.73  1.02e- 8
 7 4       c_total.sulfur.dioxide    0.00825   0.00746     1.11  2.69e- 1
 8 4       c_pH                     -4.18      0.701      -5.97  2.44e- 9
 9 4       c_sulphates               5.20      0.445      11.7   1.47e-31
10 4       c_alcohol                -0.217     0.239      -0.907 3.65e- 1
# ... with 68 more rows
```

```
newmodel$AIC
```

```
[1] 6581.262
```

```
glance(full_fit1)$AIC
```

```
[1] 6602.803
```

```
training_pred <- predict(full_fit1,wine_test)
accuracy <- mean(training_pred$.pred_class == wine_test$quality)
```

```
training_pred$.pred_class<-newmodel%>%predict(wine_test)
training_pred2<-training_pred%>%mutate(training_pred2=factor(.pred_class,levels=0:10))

accuracy2 <- mean(training_pred2$training_pred2 == wine_test$quality)
accuracy
```

```
[1] 0.6947692
```

```
accuracy2
```

```
[1] 0.6953846
```

```
full_fit1_aug <- augment(full_fit1, new_data = wine_test)
full_fit1_aug
```

```
# A tibble: 1,625 x 23
   c_fixed.acidity c_volatile.acidity c_citric.acid c_residual.sugar c_chlorides
             <dbl>              <dbl>         <dbl>            <dbl>       <dbl>
 1             7.1               0.28          0.35              3.5       0.028
 2             7.3               0.21          0.21              1.6       0.046
 3             7.1               0.23          0.3               2.6       0.034
 4             7.4               0.66          0                 1.8       0.075
 5             7.1               0.21          0.35              2.5       0.04
 6             6.4               0.21          0.34             16.0       0.04
 7             5.6               0.18          0.31              1.5       0.038
 8             8.2               0.34          0.37              1.9       0.057
 9             7.1               0.68          0                 2.2       0.073
```

```
10             7.2               0.39          0.62              11          0.047
# ... with 1,615 more rows, and 18 more variables: c_free.sulfur.dioxide <dbl>,
#   c_total.sulfur.dioxide <dbl>, c_density <dbl>, c_pH <dbl>,
#   c_sulphates <dbl>, c_alcohol <dbl>, quality <fct>, color <chr>,
#   good_wine <fct>, good_wine_names <chr>, .pred_class <fct>, .pred_3 <dbl>,
#   .pred_4 <dbl>, .pred_5 <dbl>, .pred_6 <dbl>, .pred_7 <dbl>, .pred_8 <dbl>,
#   .pred_9 <dbl>
```

```
full_fit1_conf<-full_fit1_aug %>%
  count(quality, .pred_class, .drop=FALSE) %>%
  pivot_wider(names_from = .pred_class, values_from = n)

full_fit1_conf
```

```
# A tibble: 11 x 12
   quality   `0`   `1`   `2`   `3`   `4`   `5`   `6`   `7`   `8`   `9`  `10`
   <fct>   <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
 1 0           0     0     0     0     0     0     0     0     0     0     0
 2 1           0     0     0     0     0     0     0     0     0     0     0
 3 2           0     0     0     0     0     0     0     0     0     0     0
 4 3           0     0     0     0     0     2     2     0     0     0     0
 5 4           0     0     0     0     2    45    23     0     0     0     0
 6 5           0     0     0     0     3   326   212     0     0     0     0
 7 6           0     0     0     0     0   161   532     0     0     1     0
 8 7           0     0     0     0     0     0     0   269     0     0     0
 9 8           0     0     0     0     0     0     0    45     0     0     0
10 9           0     0     0     0     0     0     0     2     0     0     0
11 10          0     0     0     0     0     0     0     0     0     0     0
```