Draft-1

STA 210 - Summer 2022

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

6-8-2022

Setup

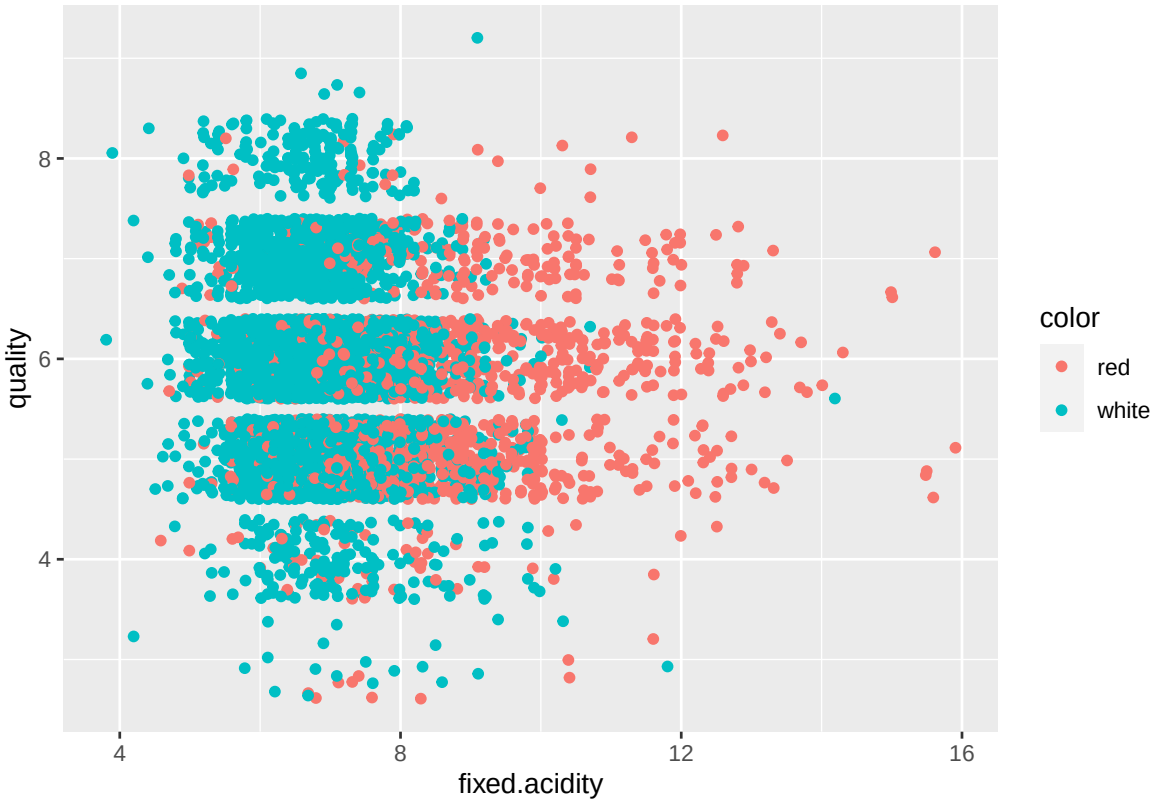
```
library(tidyverse)
library(tidymodels)
library(dplyr)
library(ggplot2)
library(cowplot)
library(knitr)
library(recipes)
library(caret)
library(InformationValue)
library(ISLR)
library(MASS)
library(nnet)
```

```
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)
```

Load packages and data:

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

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

ggplot(data = wine, aes(x = fixed.acidity, y = quality, color = color)) +
  geom_jitter()
```



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> 6.4, 8.0, 8.3, 6.1, 6.3, 8.2, 6.9, 7.2, 7.1, 7.5,~
$ volatile.acidity     <dbl> 0.35, 0.43, 0.26, 0.32, 0.32, 0.34, 0.54, 0.58, 0~
$ citric.acid          <dbl> 0.21, 0.40, 0.37, 0.37, 0.32, 0.38, 0.26, 0.03, 0~
$ residual.sugar       <dbl> 2.10, 12.40, 1.40, 1.80, 1.50, 2.50, 12.70, 2.30,~
$ chlorides            <dbl> 0.051, 0.168, 0.076, 0.051, 0.037, 0.080, 0.049, ~
$ free.sulfur.dioxide  <dbl> 46, 29, 8, 13, 12, 12, 59, 7, 28, 26, 70, 66, 44,~
$ total.sulfur.dioxide <dbl> 171, 190, 23, 200, 76, 57, 195, 28, 128, 180, 189~
$ density              <dbl> 0.99320, 0.99910, 0.99740, 0.99450, 0.98993, 0.99~
```

```
$ pH                  <dbl> 3.16, 3.07, 3.26, 3.49, 3.30, 3.30, 3.26, 3.35, 3~
$ sulphates           <dbl> 0.50, 0.64, 0.70, 0.44, 0.46, 0.47, 0.54, 0.52, 0~
$ alcohol             <dbl> 9.5, 9.2, 9.6, 10.5, 12.3, 9.0, 10.5, 10.0, 10.5,~
$ quality             <int> 5, 5, 6, 4, 6, 6, 6, 5, 5, 6, 4, 6, 6, 8, 6, 6, 6~
$ color               <chr> "white", "white", "red", "white", "white", "red",~
```

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

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

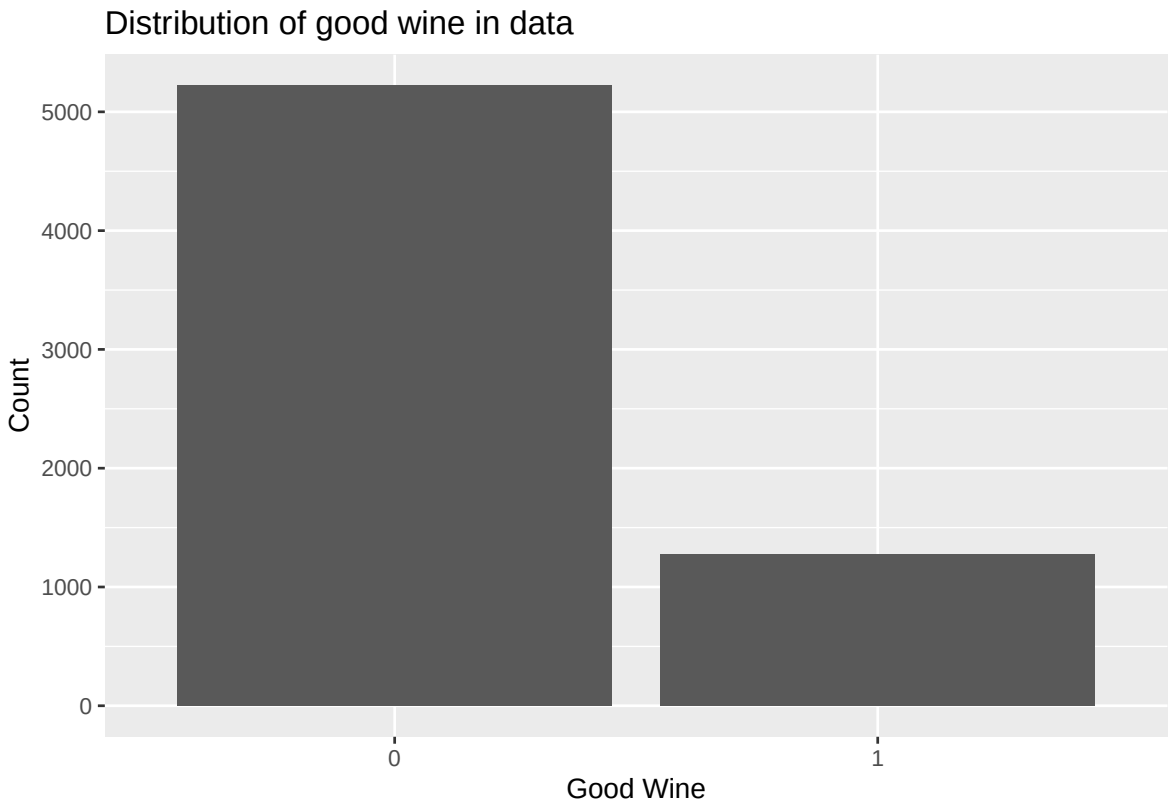
```
[1] FALSE
```

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

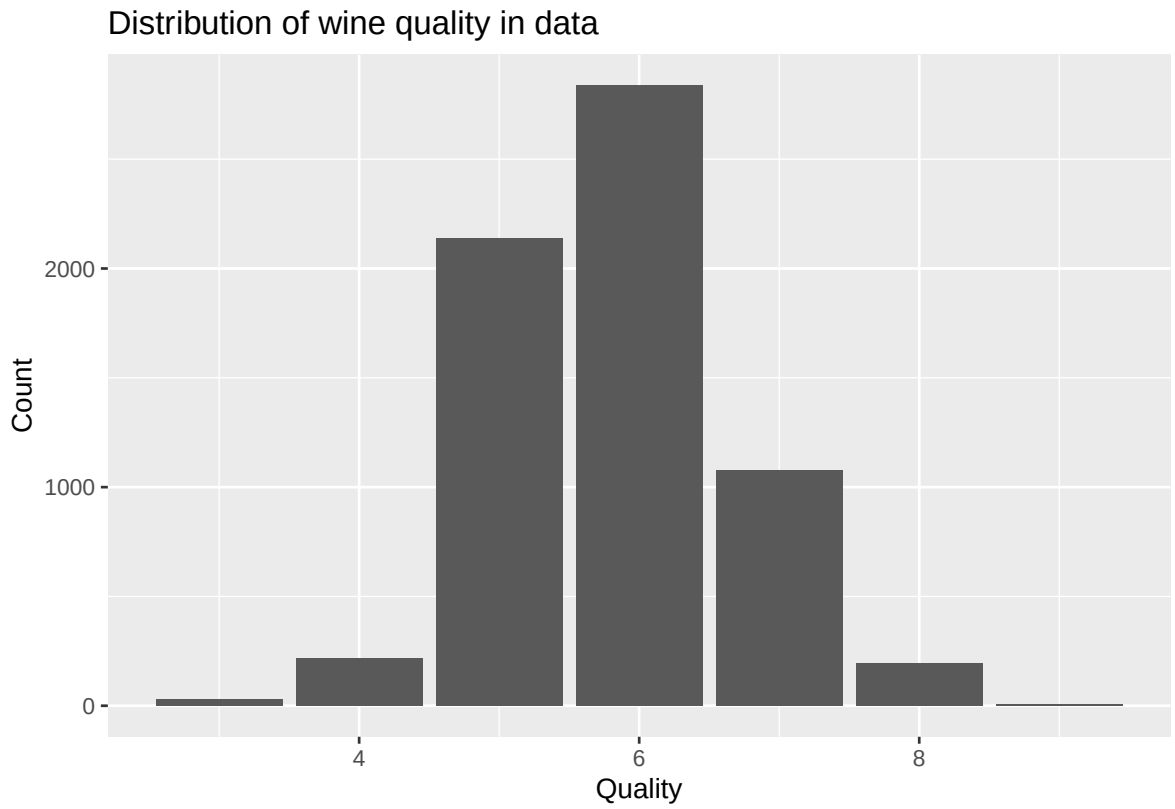
```
wine <- wine %>%
  mutate(good_wine = if_else(quality >= 7,"1","0"))
wine <- wine %>%
  mutate(good_wine = as.factor(good_wine))
wine <- wine %>%
  mutate(good_wine_names = if_else(good_wine=="1","Good wine","Bad or subpar wine"))

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"
  )
```



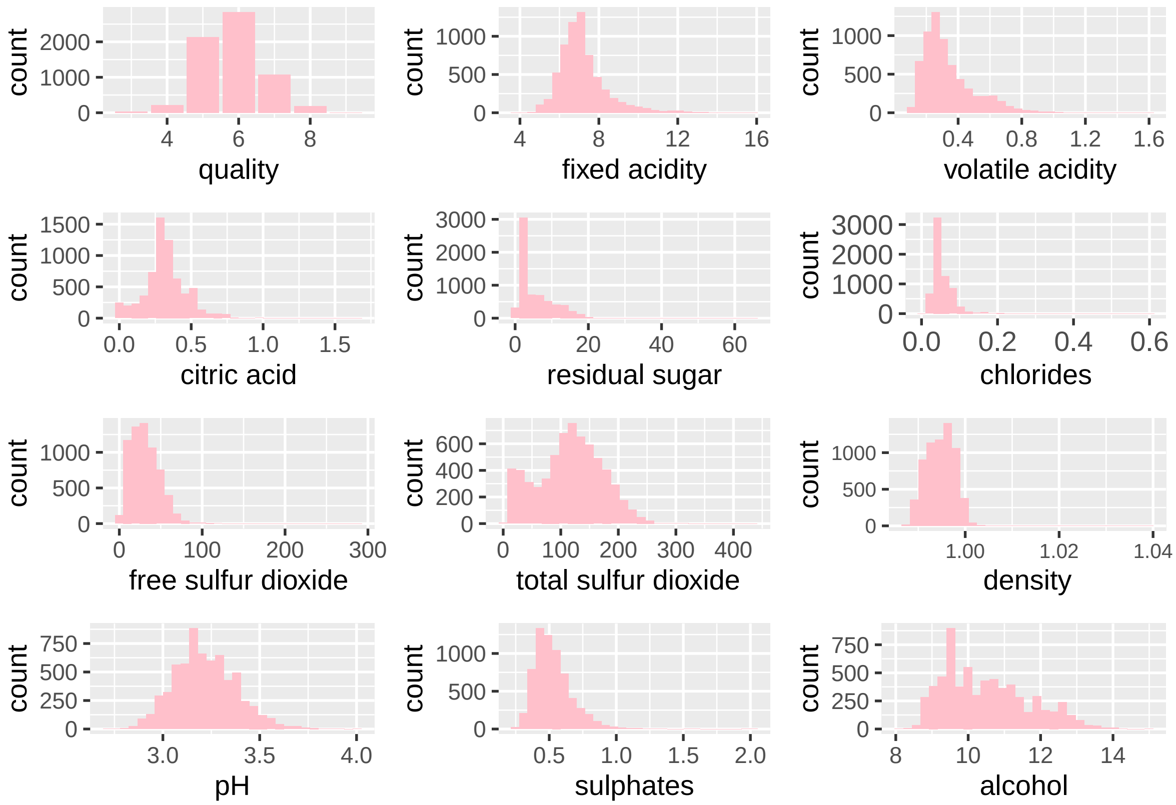

```
p1  <- ggplot(data = wine, aes(x = quality) ) +
  geom_bar(fill = "pink") +
  labs(x = "quality")

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")
```

```
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)
```

Exploratory Data Analysis

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

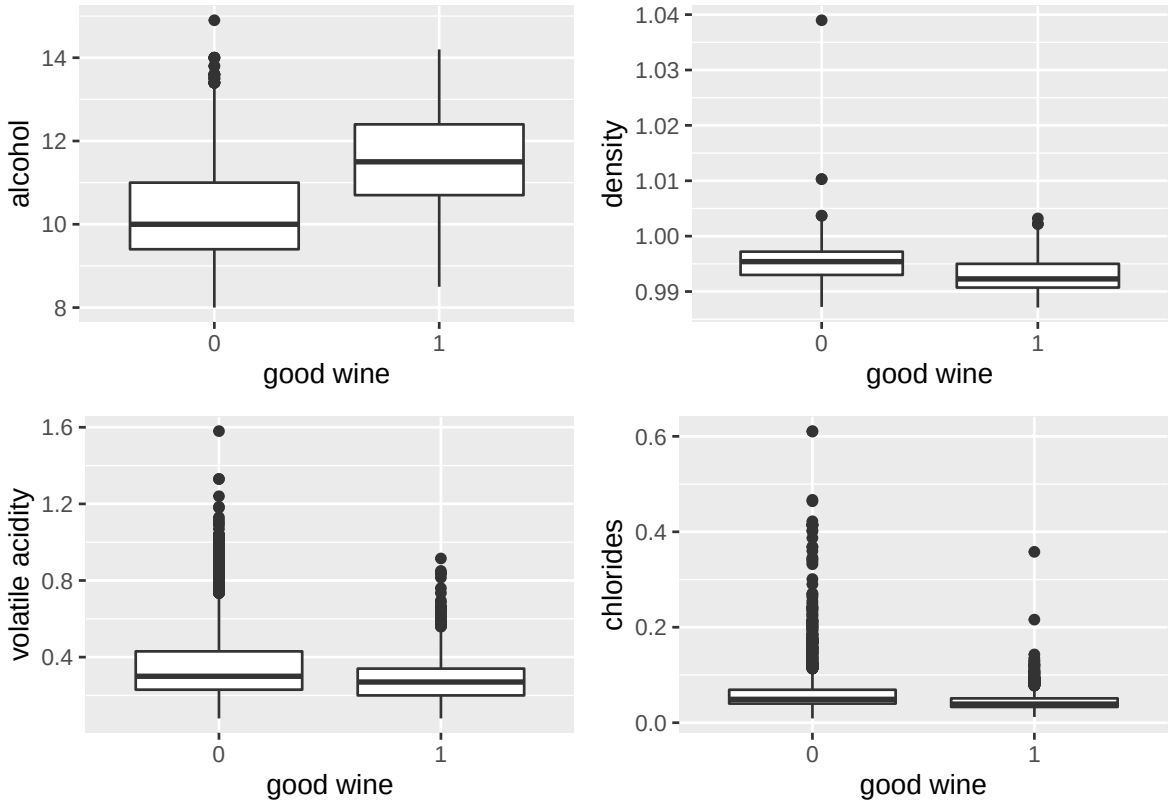
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`.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



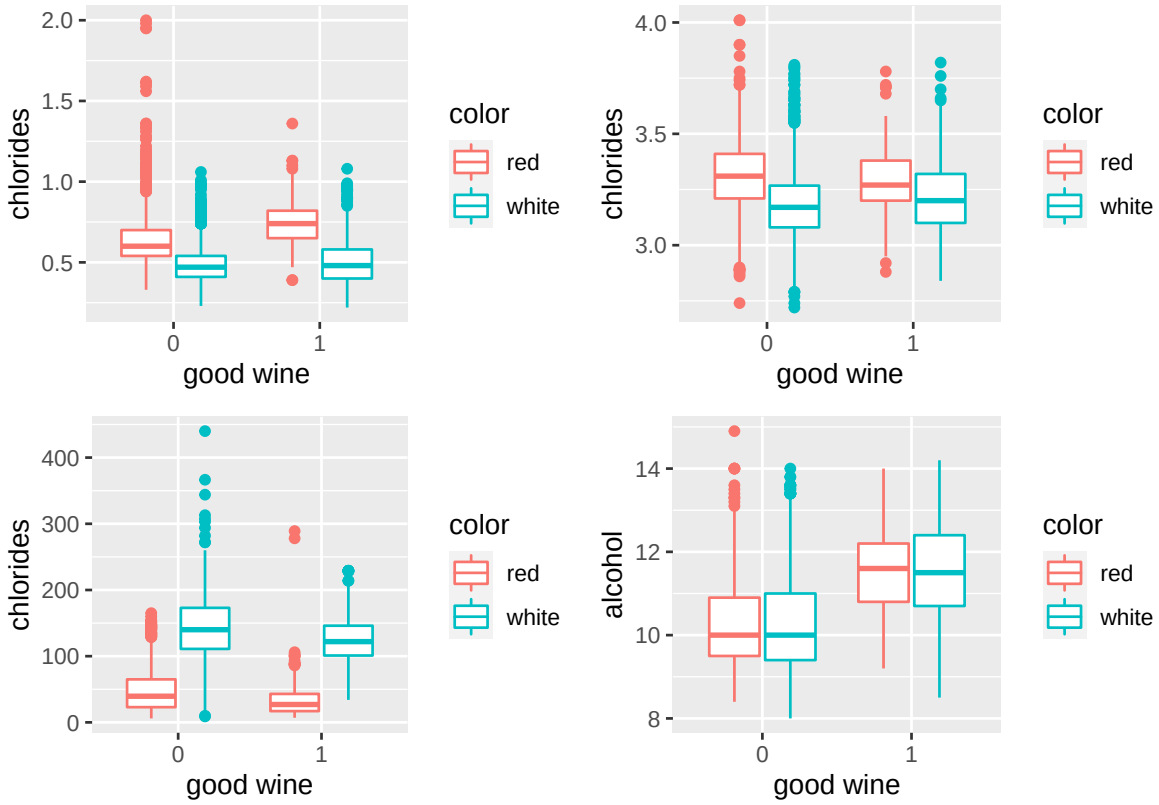

```
#cor(wine$c_alcohol, wine$quality)
#cor(wine$c_density, wine$quality)
#cor(wine$c_volatile.acidity, wine$quality)
#cor(wine$c_chlorides, wine$quality)
#cor(wine$c_residual.sugar, wine$quality)
#cor(wine$c_fixed.acidity, wine$quality)
#cor(wine$c_free.sulfur.dioxide, wine$quality)
```

```
#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)
```

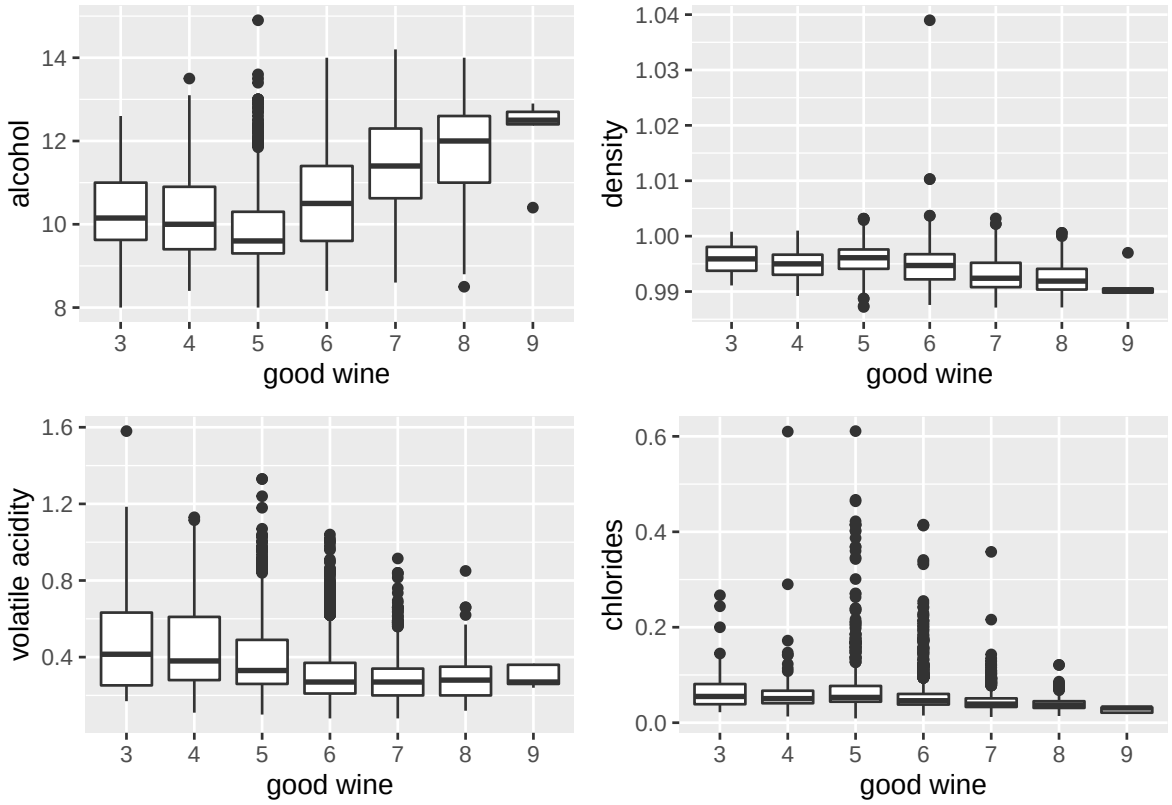
```
a1 <- ggplot(wine, aes(y = c_alcohol, x = as.factor(good_wine))) +
  geom_boxplot() +
  labs(x = "good wine", y = "alcohol")

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)
```



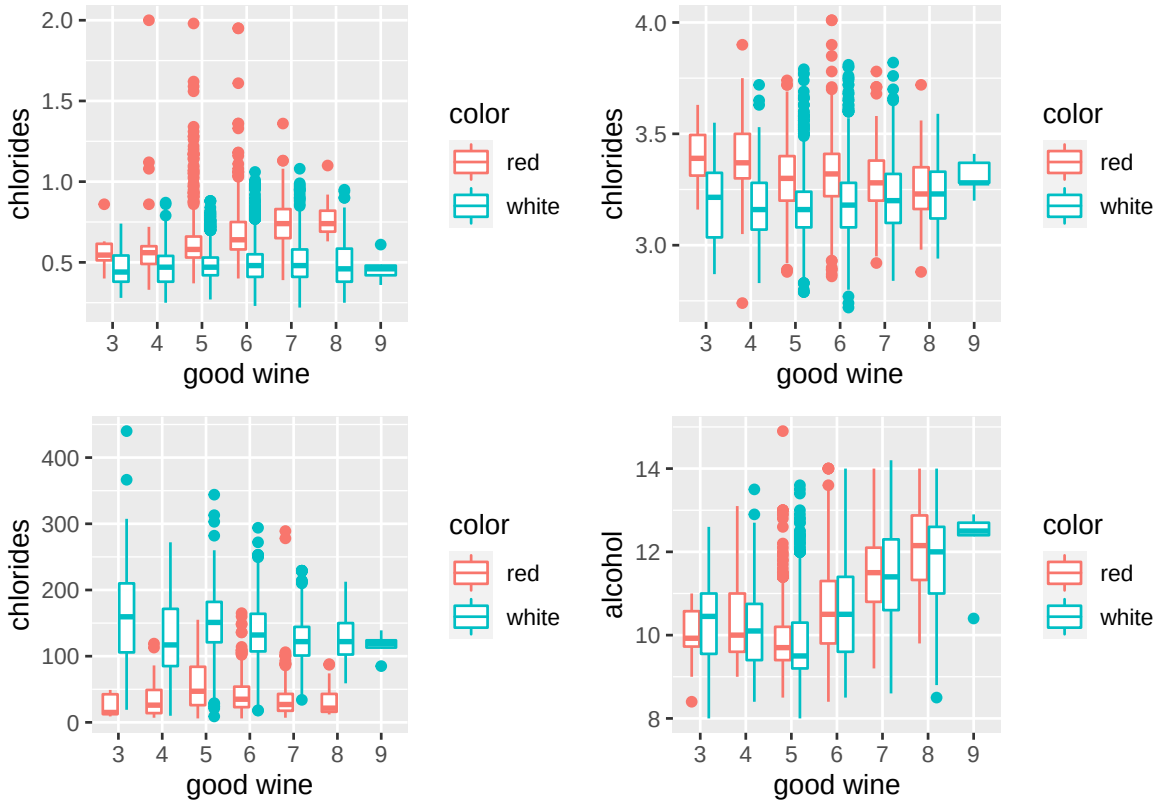

```
# exploring interaction effects
a5 <- ggplot(wine, aes(y = c_sulphates, x = as.factor(good_wine), color = color)) +
  geom_boxplot() +
  labs(x = "good wine", y = "chlorides")

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)
```




```
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)
```




```
# 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)
```



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)	-1.274	0.221	-5.771	0.000
c_fixed.acidity	0.486	0.076	6.359	0.000
c_volatile.acidity	-3.752	0.444	-8.444	0.000
c_citric.acid	-0.231	0.395	-0.585	0.558
c_residual.sugar	0.214	0.030	7.148	0.000
c_chlorides	-5.629	2.641	-2.132	0.033
c_free.sulfur.dioxide	0.010	0.003	2.972	0.003
c_total.sulfur.dioxide	-0.004	0.002	-2.270	0.023
c_density	-409.360	76.091	-5.380	0.000
c_pH	2.415	0.414	5.838	0.000
c_sulphates	2.448	0.332	7.370	0.000
c_alcohol	0.487	0.092	5.282	0.000
color_white	-0.805	0.289	-2.787	0.005

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	3839.753	NA	NA	NA
4859	3839.409	1	0.344	0.558

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)	231.822	120.558	1.923	0.054	-3.788	469.562
c_fixed.acidity	0.348	0.137	2.535	0.011	0.079	0.618
c_volatile.acidity	-3.206	0.734	-4.365	0.000	-4.678	-1.796
c_residual.sugar	0.251	0.087	2.877	0.004	0.072	0.418
c_chlorides	-3.831	3.005	-1.275	0.202	-10.561	1.351
c_free.sulfur.dioxide	0.003	0.014	0.234	0.815	-0.023	0.030
c_total.sulfur.dioxide	-0.013	0.006	-2.288	0.022	-0.024	-0.002
c_density	-248.822	123.071	-2.022	0.043	-491.564	-8.359
c_pH	0.352	1.101	0.320	0.749	-1.821	2.502
c_sulphates	3.732	0.642	5.814	0.000	2.474	5.001
c_alcohol	0.848	0.146	5.826	0.000	0.566	1.138
color_white	390.882	161.325	2.423	0.015	75.330	708.050
c_fixed.acidity_x_color_white	0.127	0.171	0.738	0.460	-0.209	0.463
c_volatile.acidity_x_color_white	-0.296	0.917	-0.323	0.747	-2.080	1.518
c_residual.sugar_x_color_white	0.034	0.096	0.357	0.721	-0.151	0.230
c_chlorides_x_color_white	-10.022	5.290	-1.895	0.058	-20.416	0.440
c_free.sulfur.dioxide_x_color_white	0.005	0.014	0.364	0.716	-0.022	0.033
c_total.sulfur.dioxide_x_color_white	0.012	0.006	2.116	0.034	0.001	0.024
c_density_x_color_white	-395.547	164.179	-2.409	0.016	-718.282	-74.371
c_pH_x_color_white	2.765	1.205	2.296	0.022	0.413	5.140
c_sulphates_x_color_white	-1.640	0.758	-2.162	0.031	-3.134	-0.153
c_alcohol_x_color_white	-0.711	0.195	-3.654	0.000	-1.096	-0.333

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)	254.800	82.824	3.076	0.002	93.048	418.061
c_fixed.acidity	0.397	0.066	5.991	0.000	0.268	0.529
c_volatile.acidity	-3.395	0.435	-7.807	0.000	-4.258	-2.552
c_residual.sugar	0.273	0.031	8.775	0.000	0.212	0.334
c_chlorides	-7.754	2.805	-2.764	0.006	-13.555	-2.608
c_free.sulfur.dioxide	0.008	0.003	2.415	0.016	0.002	0.015
c_total.sulfur.dioxide	-0.012	0.004	-3.201	0.001	-0.020	-0.005
c_density	-269.715	82.953	-3.251	0.001	-433.266	-107.756
c_sulphates	2.501	0.338	7.403	0.000	1.838	3.162
c_alcohol	0.820	0.121	6.769	0.000	0.587	1.062
color_white	333.101	75.861	4.391	0.000	184.644	482.229
c_total.sulfur.dioxide_x_color_white	0.011	0.004	2.884	0.004	0.004	0.019
c_density_x_color_white	-339.121	75.398	-4.498	0.000	-487.321	-191.558
c_pH_x_color_white	2.819	0.394	7.161	0.000	2.050	3.593
c_alcohol_x_color_white	-0.624	0.144	-4.329	0.000	-0.911	-0.345

```
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)             392.       74.9          5.24 1.62e- 7
 2 c_fixed.acidity           0.477     0.0747       6.38 1.76e-10
 3 c_volatile.acidity       -3.67      0.420       -8.73 2.49e-18
 4 c_residual.sugar          0.215     0.0300       7.17 7.45e-13
 5 c_chlorides              -5.77      2.63        -2.20 2.81e- 2
 6 c_free.sulfur.dioxide     0.0102    0.00340      2.99 2.75e- 3
 7 c_total.sulfur.dioxide   -0.00365   0.00158     -2.31 2.07e- 2
 8 c_density              -413.       75.9         -5.44 5.34e- 8
 9 c_pH                      2.43      0.413        5.87 4.31e- 9
10 c_sulphates               2.44      0.331        7.35 1.97e-13
11 c_alcohol                 0.478     0.0910       5.26 1.47e- 7
12 colorwhite               -0.821     0.288       -2.86 4.29e- 3
```

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)             -4.73      0.126    -37.5    0
 2 4       c_fixed.acidity         -0.444     0.148     -3.00   2.71e- 3
 3 4       c_volatile.acidity      -2.52      0.607     -4.14   3.43e- 5
 4 4       c_citric.acid           -1.06      0.606     -1.76   7.86e- 2
```

```
 5 4       c_residual.sugar          0.0286    0.0640      0.447 6.55e- 1
 6 4       c_chlorides             -13.2       0.0684   -193.    0
 7 4       c_free.sulfur.dioxide    -0.0898    0.0142     -6.31  2.78e-10
 8 4       c_total.sulfur.dioxide   -0.00125   0.00676    -0.185 8.53e- 1
 9 4       c_density                20.8       0.125     166.    0
10 4       c_pH                     -2.37      0.488      -4.87  1.14e- 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: 66 x 6
   y.level term                   estimate std.error statistic  p.value
   <chr>   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
 1 4       (Intercept)             17.6       0.118     149.    0
 2 4       c_fixed.acidity         -0.527     0.145      -3.64  2.77e- 4
 3 4       c_volatile.acidity      -2.80      0.560      -5.00  5.87e- 7
 4 4       c_residual.sugar         0.0539    0.0668      0.807 4.20e- 1
 5 4       c_chlorides             -9.43      0.0915   -103.    0
 6 4       c_free.sulfur.dioxide   -0.0976    0.0143     -6.85  7.58e-12
 7 4       c_total.sulfur.dioxide   0.00428   0.00636     0.673 5.01e- 1
 8 4       c_pH                    -2.74      0.461      -5.94  2.77e- 9
 9 4       c_sulphates              3.38      0.648       5.21  1.84e- 7
10 4       c_alcohol               -0.0226    0.168      -0.135 8.93e- 1
# ... with 56 more rows
```

Results

Model selection- Logistic

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

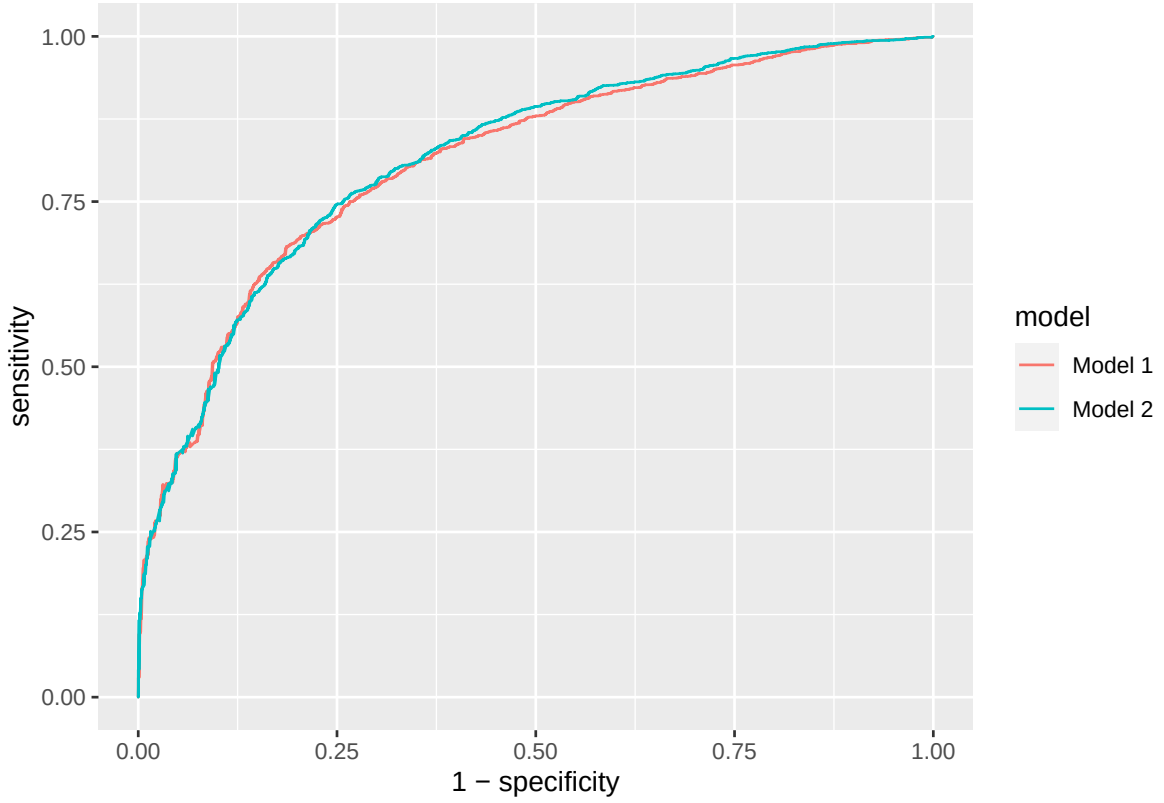
```
# 1 = reduced, 2 = full
wine_test_pred1 <- predict(wine_fit1, wine_test, type = "prob") %>%
  bind_cols(wine_test)
wine_test_pred1
```

```
# 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   0.976  0.0244             7.5              0.61          0.26
 2   0.345  0.655              5.6              0.21          0.4
 3   0.922  0.0776             6.4              0.67          0.08
 4   0.971  0.0290             8.8              0.7           0
 5   0.879  0.121              7.9              0.255         0.26
 6   0.475  0.525              6.5              0.24          0.36
 7   0.621  0.379              6.8              0.14          0.18
 8   0.500  0.500              8.3              0.3           0.49
 9   0.952  0.0476             6.5              0.25          0.5
10   0.722  0.278              6                0.16          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_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   0.989  0.0108             7.5              0.61          0.26
 2   0.285  0.715              5.6              0.21          0.4
 3   0.948  0.0521             6.4              0.67          0.08
 4   0.976  0.0241             8.8              0.7           0
 5   0.906  0.0943             7.9              0.255         0.26
 6   0.446  0.554              6.5              0.24          0.36
 7   0.660  0.340              6.8              0.14          0.18
 8   0.406  0.594              8.3              0.3           0.49
 9   0.938  0.0616             6.5              0.25          0.5
10   0.716  0.284              6                0.16          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         0.812
```

```
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         0.816
```

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

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

```
[1] 3865.409
```

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

```
[1] 3823.909
```

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

```
[1] 3949.796
```

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

```
[1] 3921.278
```

The full model has lower AIC and BIC.

```
anova(wine_fit1_eg, wine_fit2_eg)
```

```
Analysis of Deviance Table

Model 1: ..y ~ c_fixed.acidity + c_volatile.acidity + c_citric.acid +
    c_residual.sugar + c_chlorides + c_free.sulfur.dioxide +
    c_total.sulfur.dioxide + c_density + c_pH + c_sulphates +
    c_alcohol + color_white
Model 2: ..y ~ c_fixed.acidity + c_volatile.acidity + c_residual.sugar +
```

```
    c_chlorides + c_free.sulfur.dioxide + c_total.sulfur.dioxide +
    c_density + c_sulphates + c_alcohol + color_white + c_total.sulfur.dioxide_x_color_white
    c_density_x_color_white + c_pH_x_color_white + c_alcohol_x_color_white
  Resid. Df Resid. Dev Df Deviance
1      4859     3839.4
2      4857     3793.9  2   45.501
```

Drop-in-Deviance Test

```
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.

Conclusion: we'll use 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)
```

```
Warning in nnet::multinom(formula = quality ~ ., data = data, trace = FALSE):
groups '0' '1' '2' '10' are empty
```

```
full_fit1<- repair_call(full_fit1, data = wine_train)
full_fit1_fixed<-full_fit1 %>% extract_fit_engine()
```

```r
newmodel<-stepAIC(full_fit1_fixed,direction="both")
```

```
Start:  AIC=6712.88
quality ~ c_fixed.acidity + c_volatile.acidity + c_citric.acid +
    c_residual.sugar + c_chlorides + c_free.sulfur.dioxide +
    c_total.sulfur.dioxide + c_density + c_pH + c_sulphates +
    c_alcohol + color + good_wine + good_wine_names

Warning in nnet::multinom(formula = quality ~ c_volatile.acidity + c_citric.acid
+ : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity + c_citric.acid
+ : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in stepAIC(full_fit1_fixed, direction = "both"): 0 df terms are changing
AIC

                        Df    AIC
- c_density              6 6702.0
- c_citric.acid          6 6707.3
- c_chlorides            6 6711.2
<none>                     6712.9
- c_residual.sugar       6 6718.9
- c_sulphates            6 6721.6
- c_pH                   6 6721.7
- c_fixed.acidity        6 6727.8
- c_total.sulfur.dioxide 6 6729.1
- c_free.sulfur.dioxide  6 6760.9
- color                  6 6762.5
- c_alcohol              6 6795.1
- c_volatile.acidity     6 6901.1

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Step:  AIC=6701.97
quality ~ c_fixed.acidity + c_volatile.acidity + c_citric.acid +
    c_residual.sugar + c_chlorides + c_free.sulfur.dioxide +
    c_total.sulfur.dioxide + c_pH + c_sulphates + c_alcohol +
    color + good_wine + good_wine_names

Warning in nnet::multinom(formula = quality ~ c_volatile.acidity + c_citric.acid
+ : groups '0' '1' '2' '10' are empty
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity + c_citric.acid
+ : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in stepAIC(full_fit1_fixed, direction = "both"): 0 df terms are changing
AIC
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

                         Df    AIC
- c_citric.acid           6 6696.5
- c_chlorides             6 6700.6
<none>                      6702.0
- c_sulphates             6 6710.0
- c_pH                    6 6711.1
+ c_density               6 6712.9
- c_total.sulfur.dioxide  6 6719.6
- c_fixed.acidity         6 6722.1
- c_residual.sugar        6 6725.8
- c_free.sulfur.dioxide   6 6750.6
- color                   6 6763.5
- c_volatile.acidity      6 6892.7
- c_alcohol               6 7080.8

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Step:  AIC=6696.5
quality ~ c_fixed.acidity + c_volatile.acidity + c_residual.sugar +
    c_chlorides + c_free.sulfur.dioxide + c_total.sulfur.dioxide +
    c_pH + c_sulphates + c_alcohol + color + good_wine + good_wine_names

Warning in nnet::multinom(formula = quality ~ c_volatile.acidity +
c_residual.sugar + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity + c_residual.sugar
+ : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in stepAIC(full_fit1_fixed, direction = "both"): 0 df terms are changing
AIC

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

                         Df    AIC
- c_chlorides             6 6694.7
<none>                      6696.5
+ c_citric.acid           6 6702.0
- c_sulphates             6 6704.2
- c_pH                    6 6706.1
+ c_density               6 6707.3
- c_total.sulfur.dioxide  6 6716.8
- c_fixed.acidity         6 6718.3
```

```
- c_residual.sugar        6 6720.6
- c_free.sulfur.dioxide   6 6745.2
- color                   6 6755.5
- c_volatile.acidity      6 6904.5
- c_alcohol               6 7075.4

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Step:  AIC=6694.68
quality ~ c_fixed.acidity + c_volatile.acidity + c_residual.sugar +
    c_free.sulfur.dioxide + c_total.sulfur.dioxide + c_pH + c_sulphates +
    c_alcohol + color + good_wine + good_wine_names

Warning in nnet::multinom(formula = quality ~ c_volatile.acidity +
c_residual.sugar + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity + c_residual.sugar
+ : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty

Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
Warning in stepAIC(full_fit1_fixed, direction = "both"): 0 df terms are changing
AIC
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
Warning in nnet::multinom(formula = quality ~ c_fixed.acidity +
c_volatile.acidity + : groups '0' '1' '2' '10' are empty
```

```
                        Df    AIC
<none>                     6694.7
+ c_chlorides            6 6696.5
+ c_citric.acid          6 6700.6
- c_sulphates            6 6700.9
- c_pH                   6 6703.4
+ c_density              6 6705.2
- c_fixed.acidity        6 6714.0
- c_total.sulfur.dioxide 6 6716.4
- c_residual.sugar       6 6720.6
- c_free.sulfur.dioxide  6 6744.1
- color                  6 6758.8
- c_volatile.acidity     6 6904.6
- c_alcohol              6 7107.3
```

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

```
# A tibble: 90 x 6
   y.level term                estimate std.error statistic   p.value
   <chr>   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
 1 4       (Intercept)             6.93     0.212      32.7  1.17e-234
 2 4       c_fixed.acidity        -0.592    0.144      -4.12 3.74e-  5
```

```
 3 4       c_volatile.acidity      -0.827    0.539       -1.53 1.25e-  1
 4 4       c_citric.acid            1.54     0.475        3.24 1.22e-  3
 5 4       c_residual.sugar        -0.0604   0.0604      -1.00 3.17e-  1
 6 4       c_chlorides            -13.5      0.0583    -231.   0
 7 4       c_free.sulfur.dioxide   -0.0901   0.0170      -5.30 1.14e-  7
 8 4       c_total.sulfur.dioxide   0.0189   0.00803      2.35 1.86e-  2
 9 4       c_density                9.37     0.211       44.5  0
10 4       c_pH                    -2.64     0.716       -3.69 2.26e-  4
# ... with 80 more rows
```

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

```
# A tibble: 72 x 6
   y.level term                  estimate std.error statistic      p.value
   <chr>   <chr>                    <dbl>     <dbl>     <dbl>        <dbl>
 1 4       (Intercept)              12.1      0.173      70.2   0
 2 4       c_fixed.acidity          -0.441    0.142      -3.11  0.00187
 3 4       c_volatile.acidity       -1.19     0.515      -2.31  0.0211
 4 4       c_residual.sugar         -0.0419   0.0613     -0.683 0.495
 5 4       c_free.sulfur.dioxide    -0.0948   0.0172     -5.53  0.0000000323
 6 4       c_total.sulfur.dioxide    0.0213   0.00789     2.69  0.00705
 7 4       c_pH                     -2.07     0.705      -2.94  0.00330
 8 4       c_sulphates               0.647    0.525       1.23  0.218
 9 4       c_alcohol                 0.140    0.258       0.545 0.586
10 4       colorwhite               -1.09     0.656      -1.66  0.0976
# ... with 62 more rows
```

```
newmodel$AIC
```

```
[1] 6694.68
```

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

```
[1] 6712.881
```

```
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)
```

```
# I'm getting errors for everything i've commented below and for the rest of the doc
# ?
#fit2_aug <- augment(wine_fit2, new_data = wine_test)

#fit2_conf<-fit2_aug %>%
  #count(good_wine, .pred_class, .drop=FALSE) %>%
  #pivot_wider(names_from = .pred_class, values_from = n)

#fit2_conf
```

```
# predicted <- predict(wine_fit2, wine_test)
# predicted <- predicted %>%
  # mutate(.pred_class = as.numeric(.pred_class))

# optimal <- optimalCutoff(as.numeric(wine_test$good_wine), predicted)[1]

# mis1 <- misClassError(as.numeric(wine_test$good_wine), predicted, threshold = optimal)
# accuracy <- mean(as.numeric(wine_test$good_wine) == as.numeric(predicted$.pred_class))
```

```
# newmodel$AIC
# glance(full_fit1)$AIC

# training_pred <- predict(newmodel, wine_test)
# training_pred <-data_frame(training_pred)
# accuracy <- mean(wine_test$quality == training_pred$training_pred)
```