

# Wine

Hồ Nghĩa Phương

Hồ Nghĩa Phương - 21110154

**Đề tài:** Phân loại chất lượng và loại rượu vang. Phân tích dữ liệu để tìm thông tin hữu ích giúp doanh nghiệp cải thiện chất lượng rượu vang.

## Nguồn gốc dữ liệu

Bộ dữ liệu chất lượng rượu vang bao gồm các đặc tính hóa học của cả hai biến thể màu đỏ và trắng của rượu vang ‘Vinho Verde’ xuất xứ từ phía bắc Bồ Đào Nha. Nó bao gồm các biến như nồng độ cồn và mức độ axit, cùng với xếp hạng chất lượng. Bộ dữ liệu này thường được sử dụng để điều tra mối quan hệ giữa các tính chất hóa học này và chất lượng cảm nhận của rượu vang.

*Nguồn:* <https://archive.ics.uci.edu/dataset/186/wine+quality>

## Giới thiệu các biến

- **fixed acidity:** lượng gram axit tartaric trên mỗi đè-xi-mét khối.
- **volatile acidity:** lượng gram axit axetic trên mỗi đè-xi-mét khối.
- **citric acid:** lượng gram axit citric trên mỗi đè-xi-mét khối.
- **residual sugar:** lượng gram đường còn lại trên mỗi đè-xi-mét khối.
- **chlorides:** lượng gram natri clorua trên mỗi đè-xi-mét khối.
- **free sulfur dioxide:** lượng gram lưu huỳnh dioxit tự do trên mỗi đè-xi-mét khối.
- **total sulfur dioxide:** lượng gram tổng lượng lưu huỳnh dioxit trên mỗi đè-xi-mét khối.
- **density:** mật độ của rượu tính bằng gram trên mỗi đè-xi-mét khối.
- **pH:** giá trị pH của rượu.
- **sulphates:** lượng gram kali sulphate trên mỗi đè-xi-mét khối.
- **alcohol:** phần trăm thể tích cồn trong rượu.
- **quality:** điểm chất lượng từ 0 (đại diện cho chất lượng thấp) đến 10 (đại diện cho chất lượng cao).
- **type:** loại rượu vang (đỏ hoặc trắng)

## Mục tiêu

Hai biến mục tiêu là quality và type. Điểm chất lượng được liệt kê dưới dạng các giá trị thứ tự từ 3 đến 9. Trong đó, với loại rượu vang đỏ, phạm vi điểm là từ 3 tới 8, với loại rượu vang trắng,

## Khai báo thư viện

```

library(tidyverse)
library(corrplot)
library(ggplot2)
library(factoextra)
library(dplyr)
library(ggfortify)
library(corrgram)
library(psych)

```

## Nhập dữ liệu

```

# Read and clean white wine data
wine_white <- read.csv("wine+quality/winequality-white.csv", sep = ";") |>
  janitor::clean_names() |>
  mutate(type = "white") # Add a column to identify white wine

# Read and clean red wine data
wine_red <- read.csv("wine+quality/winequality-red.csv", sep = ";") |>
  janitor::clean_names() |>
  mutate(type = "red") # Add a column to identify red wine

```

Bộ dữ liệu rượu vang trắng có 4898 quan trắc, trong khi bộ dữ liệu rượu vang đỏ có 1599 quan trắc. Để đơn giản hóa việc phân tích và tận dụng các đặc điểm chung, các bộ dữ liệu này đã được hợp nhất thành một bộ dữ liệu duy nhất.

```

# Combine the two datasets into one
# Assuming wine_white and wine_red are your datasets
wine <- bind_rows(wine_white, wine_red) |>
  mutate(quality = factor(quality),
        type = factor(type, levels = c("red", "white"))) |>
  tibble::as.tibble()

## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## i Please use `as_tibble()` instead.
## i The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

# Verify the structure of 'wine' dataframe
str(wine)

```

```

## tibble [6,497 x 13] (S3:tbl_df/tbl/data.frame)
## $ fixed_acidity      : num [1:6497] 7 6.3 8.1 7.2 7.2 ...
## $ volatile_acidity   : num [1:6497] 0.27 0.3 0.28 0.23 0.23 ...
## $ citric_acid        : num [1:6497] 0.36 0.34 0.4 0.32 0.32 ...
## $ residual_sugar     : num [1:6497] 20.7 1.6 6.9 8.5 8.5 ...
## $ chlorides          : num [1:6497] 0.045 0.049 0.05 0.058 0.058 ...
## $ free_sulfur_dioxide: num [1:6497] 45 14 30 47 47 ...
## $ total_sulfur_dioxide: num [1:6497] 170 132 97 186 186 ...
## $ density             : num [1:6497] 1.001 0.994 0.995 0.996 0.996 ...
## $ p_h                 : num [1:6497] 3 3.3 3.26 3.19 3.19 ...
## $ sulphates           : num [1:6497] 0.45 0.49 0.44 0.4 0.44 ...
## $ alcohol              : num [1:6497] 8.8 9.5 10.1 9.9 9.9 ...
## $ quality              : Factor w/ 7 levels "3","4","5","6",...: 4 4 4 4 4 4 4 4 ...

```

```
## $ type : Factor w/ 2 levels "red","white": 2 2 2 2 2 2 2 2 2 2 ...
```

Dữ liệu wine chứa 11 biến đặc trưng đại diện cho các đặc tính hóa học của rượu vang, chẳng hạn như độ axit cố định, lượng đường còn lại, clo, mật độ, v.v. Biến type được thêm vào để phân biệt hai loại rượu vang.

## Trực quan hóa

### Khuôn mẫu

```
# Chuyển đổi dữ liệu thành dạng dài
wine_long <- reshape2::melt(wine, id.vars = c("type", "quality"))

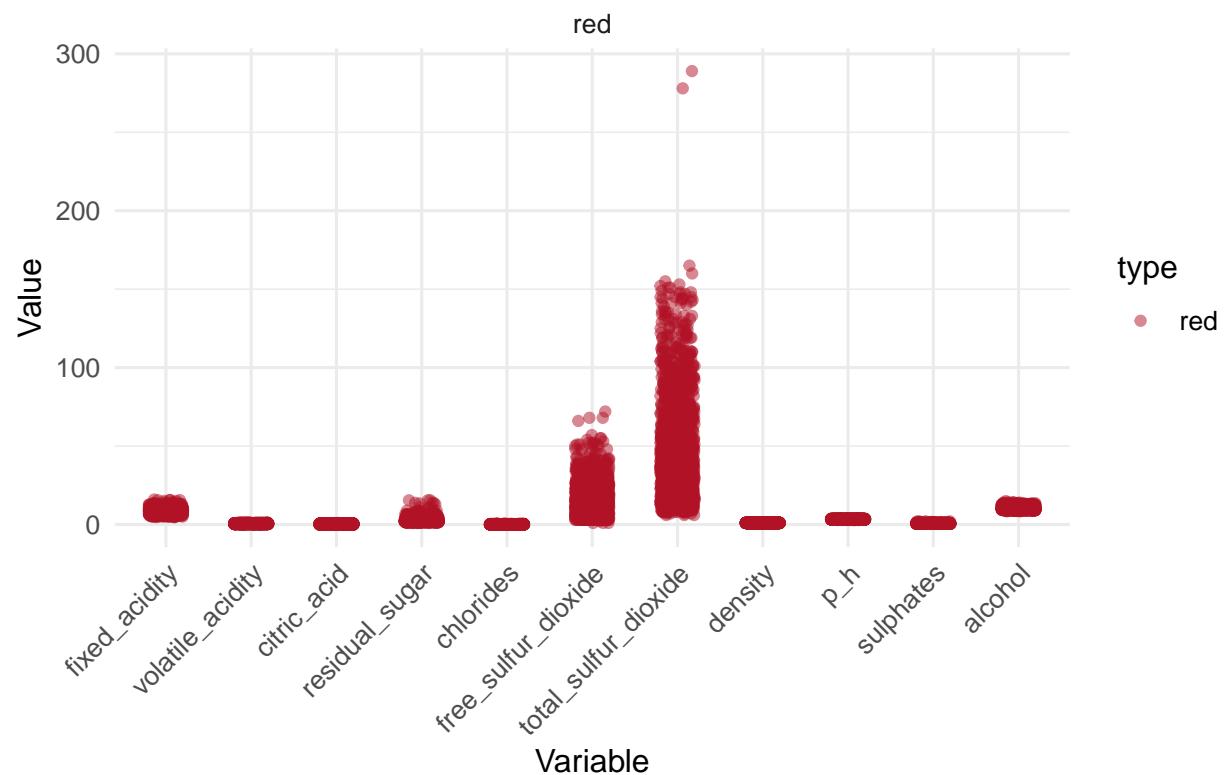
# Tạo một danh sách các biến
variables <- c(
  "fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar",
  "chlorides", "free_sulfur_dioxide", "total_sulfur_dioxide",
  "density", "pH", "sulphates", "alcohol"
)

# Vẽ biểu đồ cho từng loại rượu
for (t in c("red", "white")) {
  wine_long_subset <- wine_long[wine_long$type == t, ]

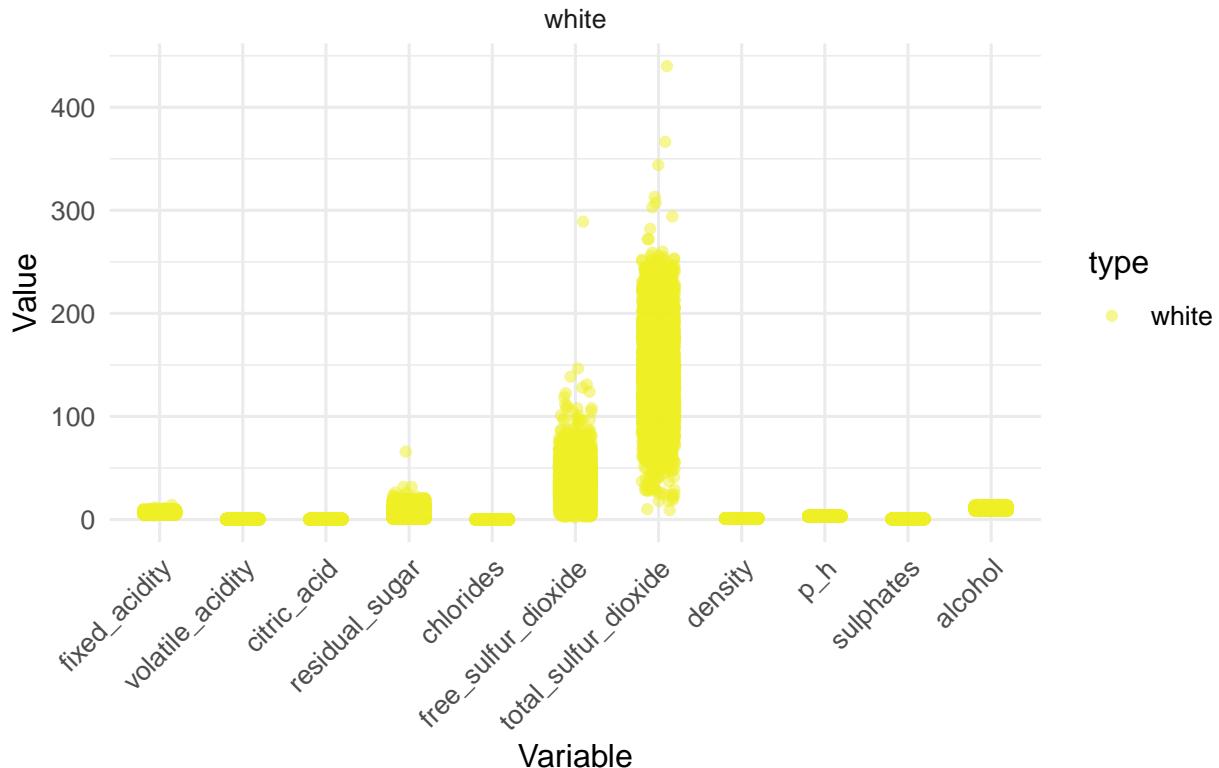
  p <- ggplot(wine_long_subset, aes(x = variable, y = value)) +
    geom_jitter(aes(colour = type), width = 0.2, height = 0, alpha = 0.5) +
    labs(title = paste("Dot Plot for", t, "Wine"),
        x = "Variable",
        y = "Value") +
    theme_minimal(base_size = 12) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_colour_manual(values = c("red" = "#b11226", "white" = "#EFEF26")) +
    facet_wrap(~type, scales = "free_x", ncol = 1) # Vẽ hai biểu đồ trong hai khung hình riêng biệt

  print(p)
}
```

## Dot Plot for red Wine



## Dot Plot for white Wine



**Nhận xét:** Thông qua hai biểu đồ của các biến định lượng theo từng loại rượu vang, các đặc tính hóa học theo từng loại rượu đều thể hiện theo cùng một khuôn mẫu, chỉ khác về mức độ (cụ thể là ở một vài biến thì rượu trắng có giá trị cao hơn) nên việc kết hợp hai bộ dữ liệu để phân tích như trên là hợp lý.

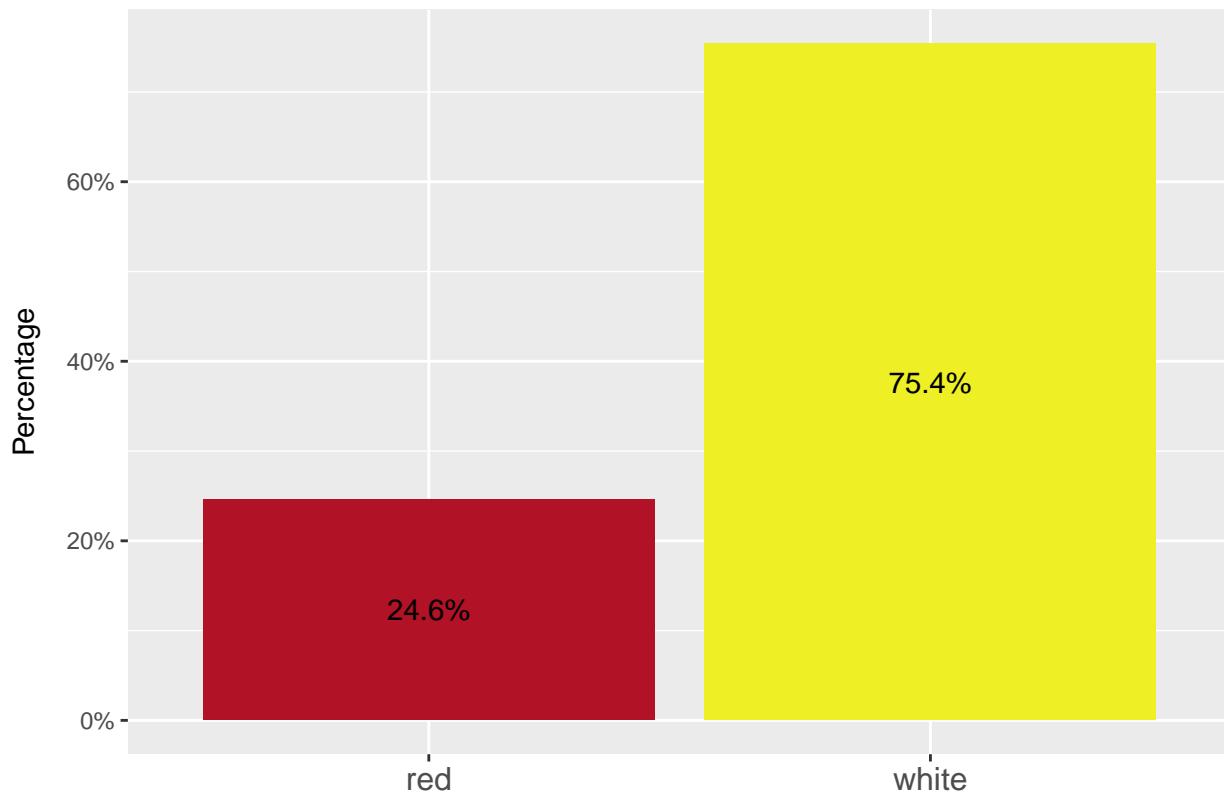
### Phân phối các biến đặc trưng

**Type:** biến định tính phân loại rượu vang với hai giá trị “red” và “white”.

```
# Calculate the percentage of each wine type
df_percentage <- wine |>
  group_by(type) |>
  summarise(percentage = n() / nrow(wine) * 100)

# Plotting the proportion of Red and White Wine
ggplot(df_percentage, aes(x = type, y = percentage, fill = type)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = sprintf("%.1f%%", percentage)),
            position = position_stack(vjust = 0.5)) +
  labs(title = "Proportion of Red and White Wine",
       x = NULL,
       y = "Percentage") +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  scale_fill_manual(values = c("red" = "#b11226", "white" = "#EFEF26")) +
  theme(axis.title.y = element_text(margin = margin(r = 10)),
        axis.text.x = element_text(size = 12),
        legend.position = "none",
        plot.title = element_text(hjust = 0.5))
```

## Proportion of Red and White Wine



**Nhận xét:** Cỡ mẫu của dữ liệu rượu vang đỏ thấp hơn nhiều so với rượu vang trắng, điều này có thể gây khó khăn trong quá trình phân tích.

**Quality:** chất lượng rượu vang được đánh giá theo thang điểm từ 0 đến 10 gán cho từng mẫu rượu dựa trên các đánh giá của các chuyên gia rượu vang.

```
# Calculate percentage of each quality level by wine type
quality_distribution <- wine |>
  group_by(type, quality) |>
  summarise(count = n(), .groups = 'drop') |>
  group_by(type) |>
  mutate(percentage = round(count / sum(count) * 100, 1))

# Calculate total quality distribution
total_quality_distribution <- wine |>
  group_by(quality) |>
  summarise(total_count = n()) |>
  mutate(percentage = round(total_count / sum(total_count) * 100, 1),
        type = "Total", .groups = 'drop')

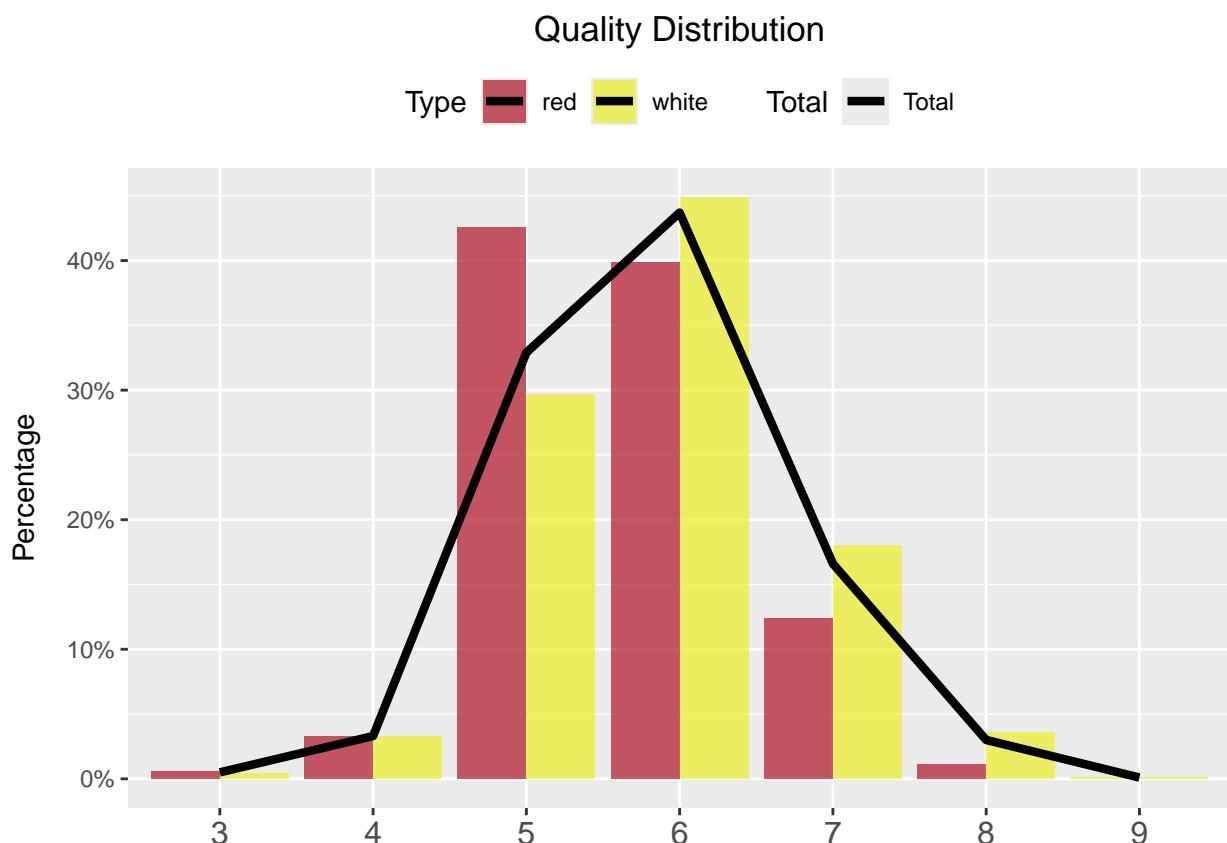
# Convert 'type' to factor with levels
total_quality_distribution$type <- factor(total_quality_distribution$type,
                                         levels = c("Total"))

# Plotting
ggplot(quality_distribution, aes(x = as.factor(quality), y = percentage,
                                   fill = type)) +
  geom_bar(stat = "identity", position = "dodge", alpha = 0.7) +
  geom_line(data = total_quality_distribution,
```

```

aes(x = as.factor(quality), y = percentage, group = type, color = type),
  linewidth = 1.5, show.legend = TRUE) + # Sửa size thành linewidth
  labs(title = "Quality Distribution",
    x = NULL,
    y = "Percentage") +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  scale_fill_manual(values = c("red" = "#b11226", "white" = "#EFEF26"),
    name = "Type") +
  scale_color_manual(values = c("Total" = "black"),
    name = "Total") +
  theme(
    axis.title.y = element_text(margin = margin(r = 10)),
    axis.text.x = element_text(size = 12),
    legend.position = "top", # Move legend to top for better visibility
    plot.title = element_text(hjust = 0.5)
  )

```



**Nhận xét:** Hầu hết mẫu rượu có số điểm từ 4 đến 8, các mẫu rượu có điểm 3 và 9 là cực kì ít. Do mẫu rượu vang trắng có nhiều quan trắc hơn rượu vang đỏ nên việc chỉ có các mẫu rượu trắng có điểm 9 là điều dễ hiểu. Điểm đặc biệt ở đây là mặc dù cỡ mẫu rượu vang trắng lớn hơn rất nhiều nhưng tỷ lệ rượu vang trắng có điểm 3 lại ít hơn so với tỷ lệ rượu có điểm 3.

Biểu đồ histogram và jitter của các biến đặc trưng hóa học

```

variables <- c("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar",
  "chlorides", "free_sulfur_dioxide", "total_sulfur_dioxide", "density",
  "p_h", "sulphates", "alcohol")

```

```

plots <- list()

for (var in variables) {
  freq_data <- wine |>
    group_by(type, .data[[var]]) |>
    count()

  p_bar <- ggplot(freq_data, aes(x = .data[[var]], y = n, fill = type)) +
    geom_bar(stat = "identity", position = "dodge", alpha = 0.7) +
    labs(title = paste("Frequency of", var, "by Type"),
         x = var,
         y = "Frequency") +
    scale_fill_manual(values = c("red" = "#b11226", "white" = "#EFEF26")) +
    theme_minimal() +
    theme(
      axis.title.y = element_text(margin = margin(r = 10)),
      axis.text.x = element_text(size = 12),
      legend.position = "none",
      plot.title = element_text(hjust = 0.5)
    )

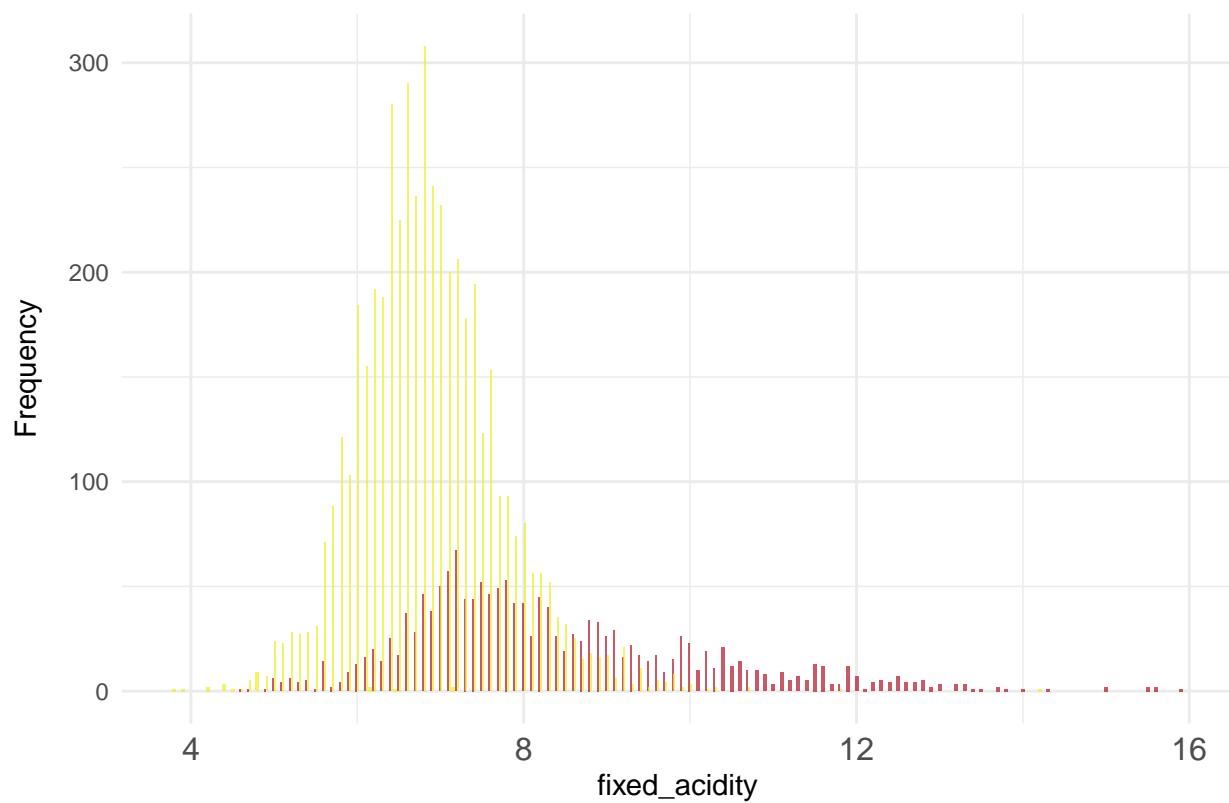
  p_jitter <- ggplot(wine, aes(x = quality, y = .data[[var]], color = type)) +
    geom_jitter(alpha = 0.7, width = 0.3) +
    labs(title = paste("Distribution of", var, "by Quality"),
         x = "Quality",
         y = var) +
    scale_color_manual(values = c("red" = "#b11226", "white" = "#EFEF26")) +
    theme_minimal() +
    theme(
      axis.title.y = element_text(margin = margin(r = 10)),
      axis.text.x = element_text(size = 12),
      legend.position = "none",
      plot.title = element_text(hjust = 0.5)
    )

  plots[[paste(var, "bar")]] <- p_bar
  plots[[paste(var, "jitter")]] <- p_jitter
}

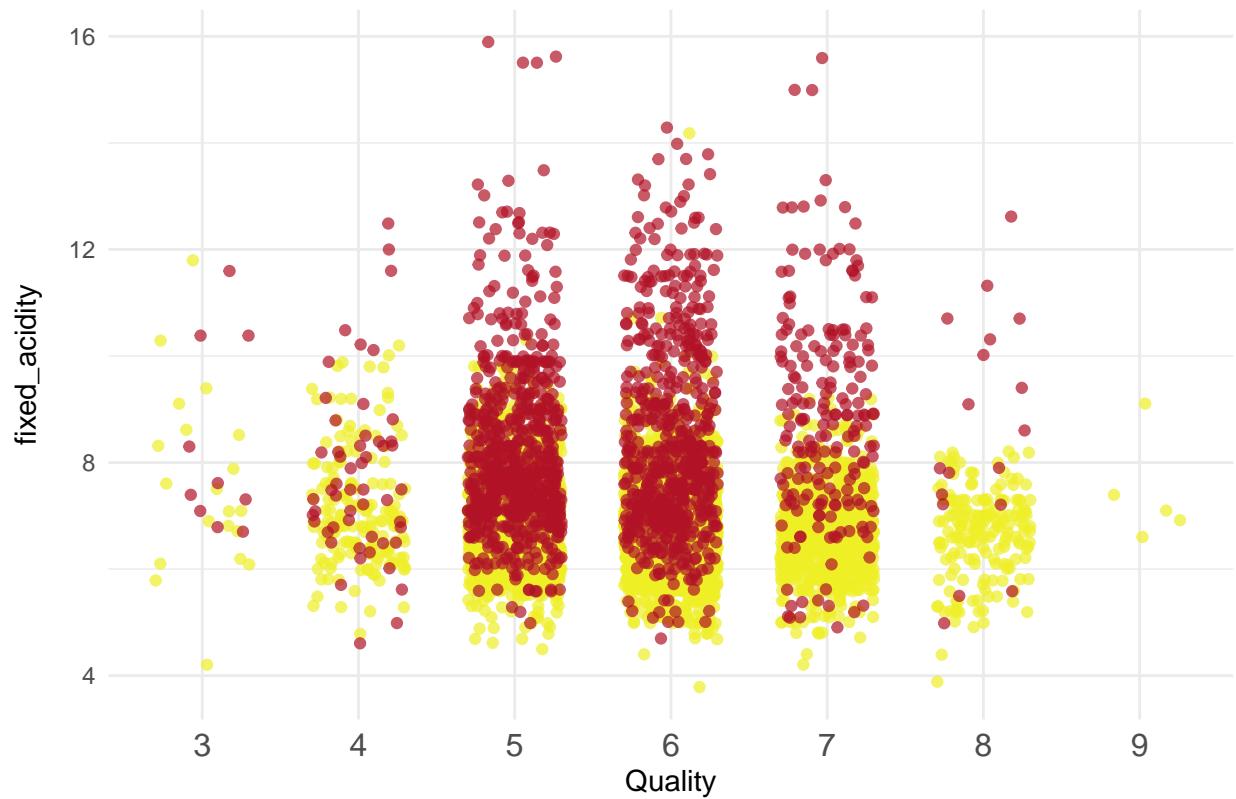
for (plot in plots) {
  print(plot)
}

```

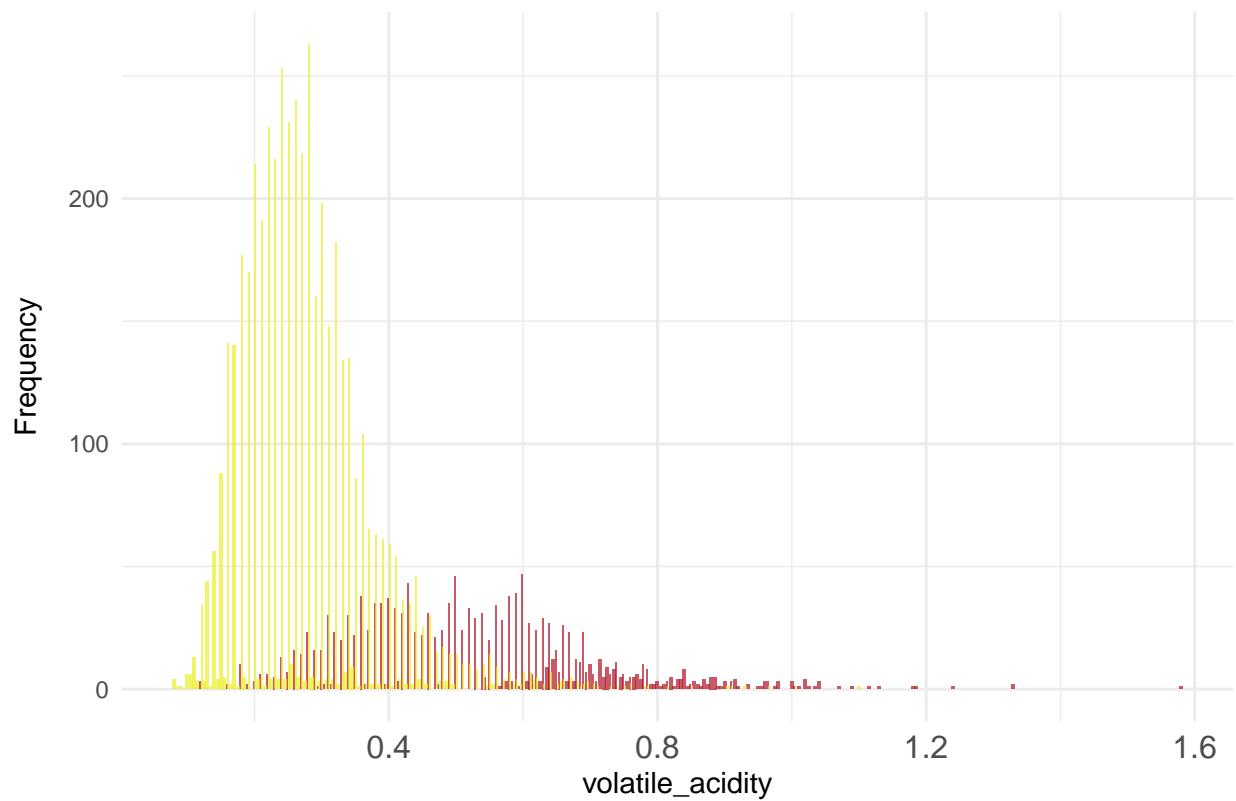
Frequency of fixed\_acidity by Type



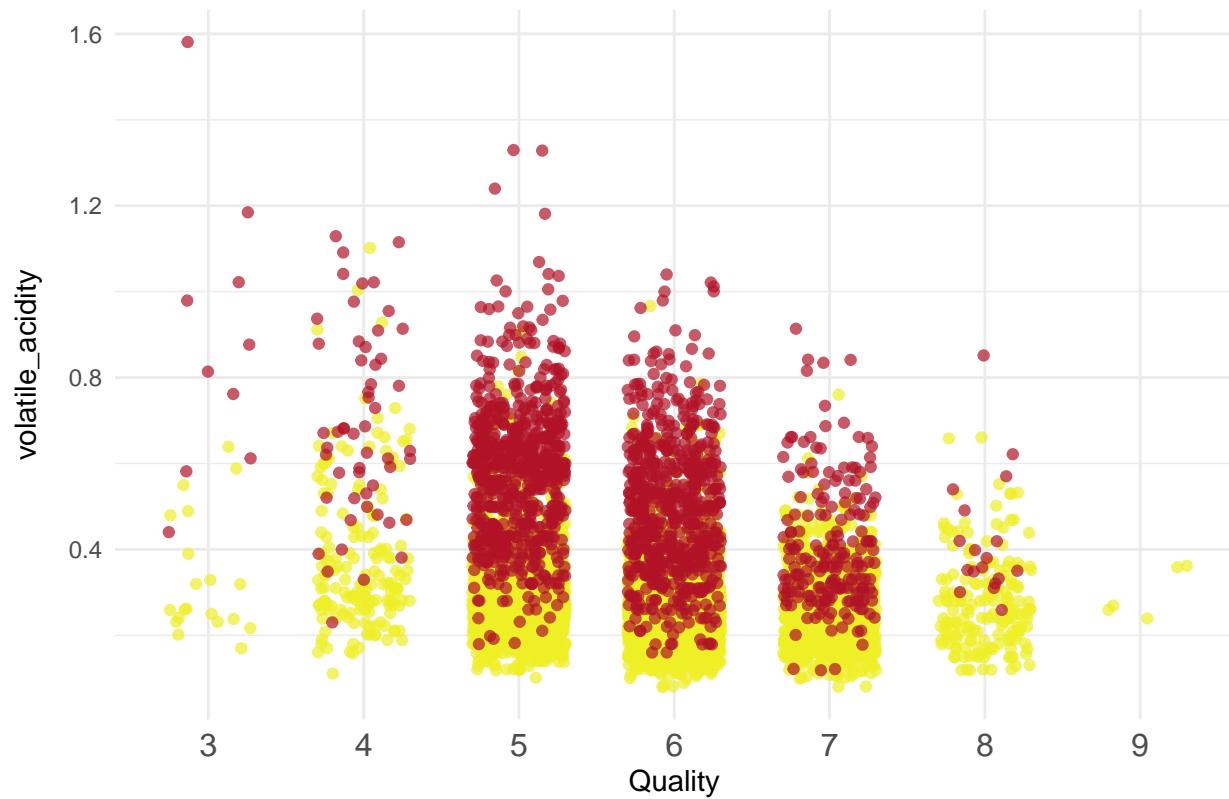
### Distribution of fixed\_acidity by Quality



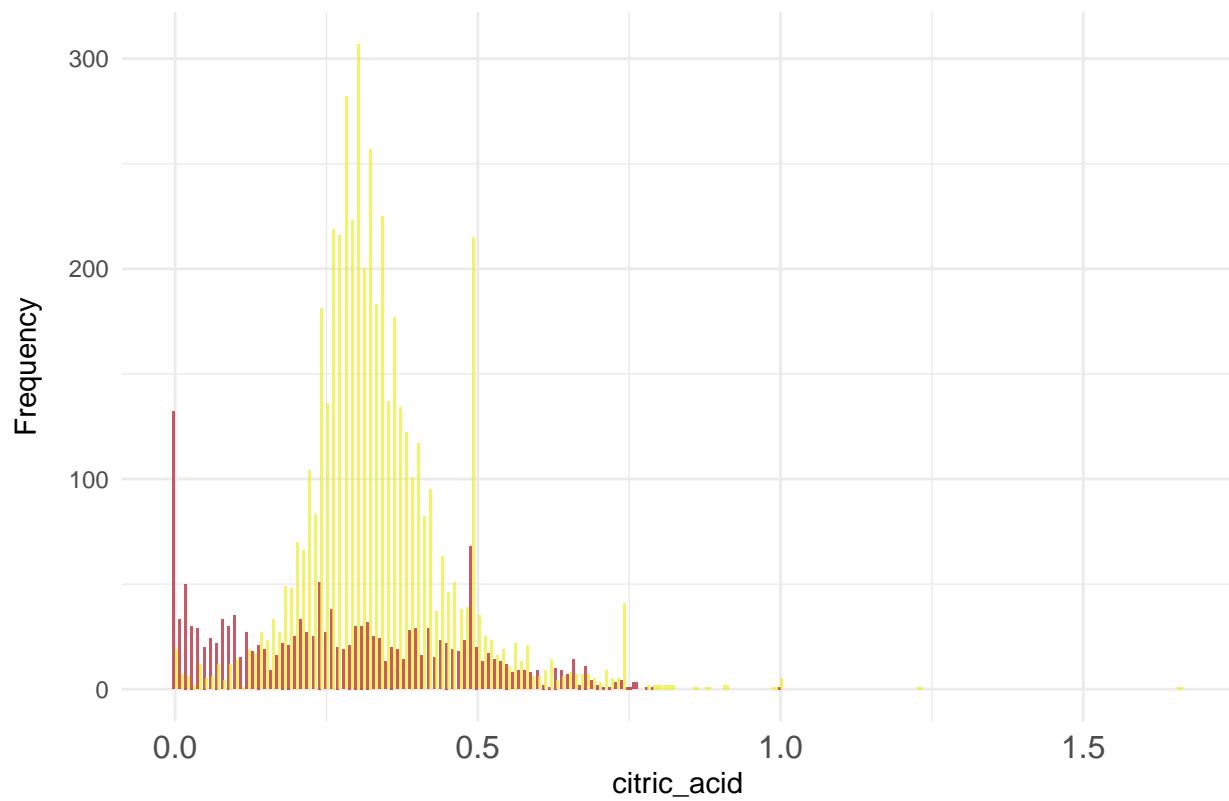
Frequency of volatile\_acidity by Type



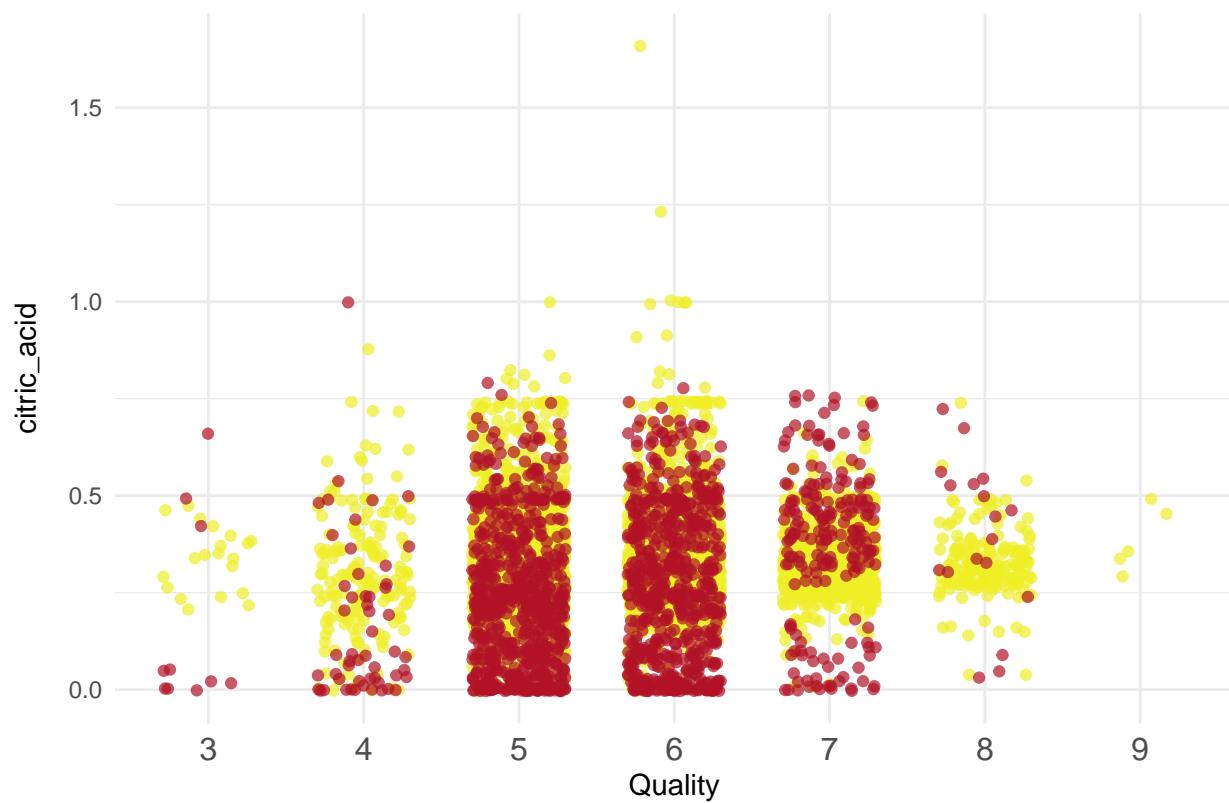
Distribution of volatile\_acidity by Quality



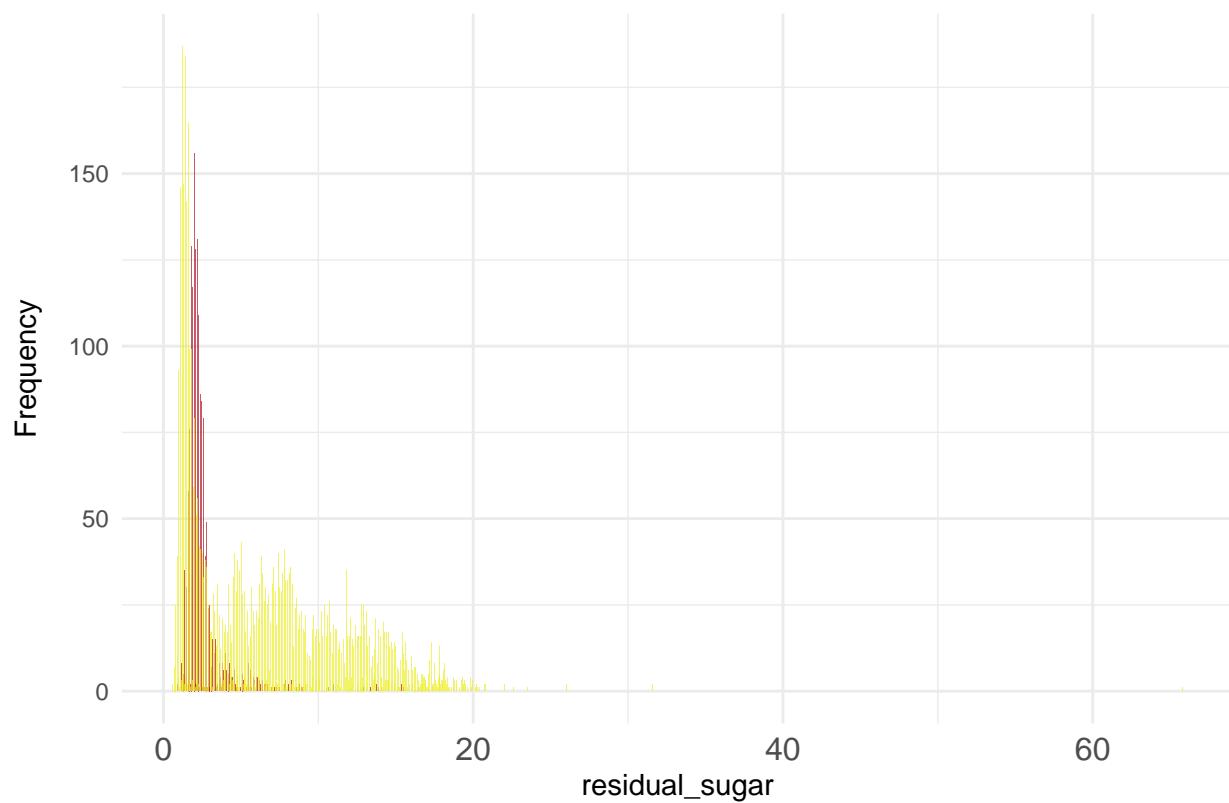
### Frequency of citric\_acid by Type



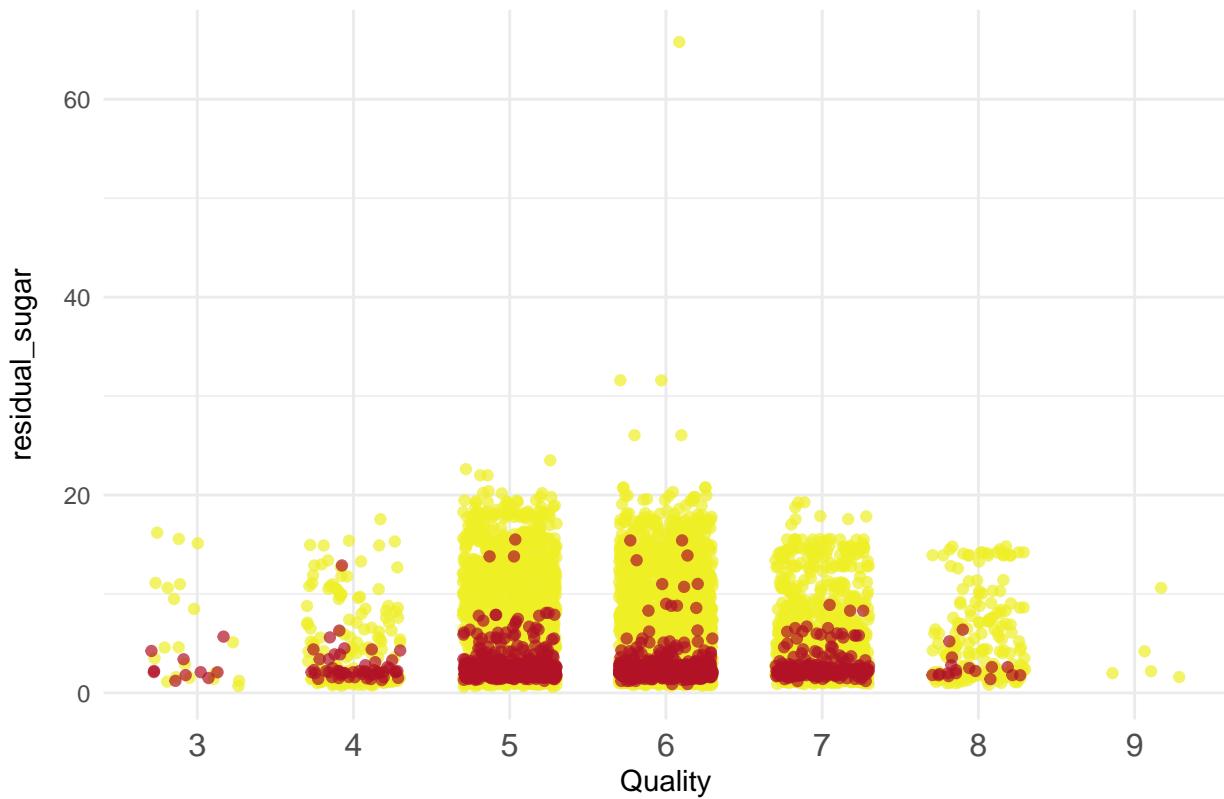
Distribution of citric\_acid by Quality



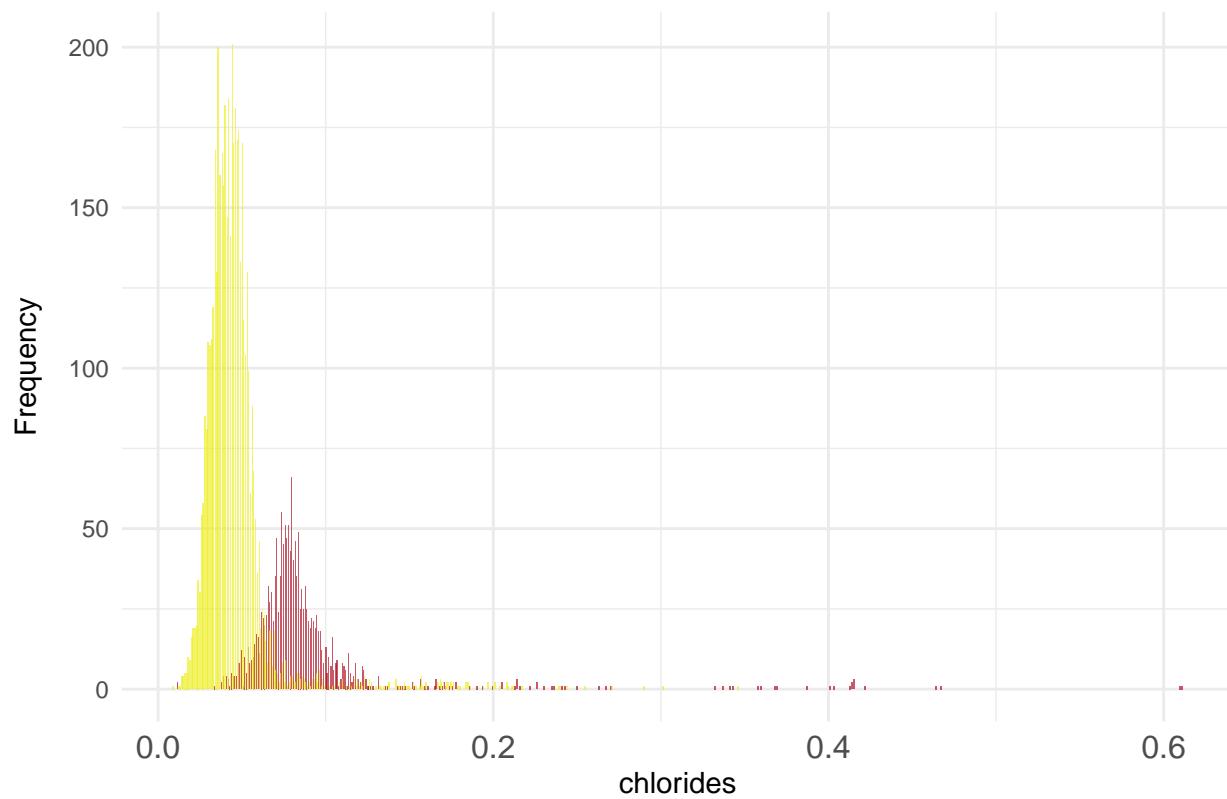
Frequency of residual\_sugar by Type



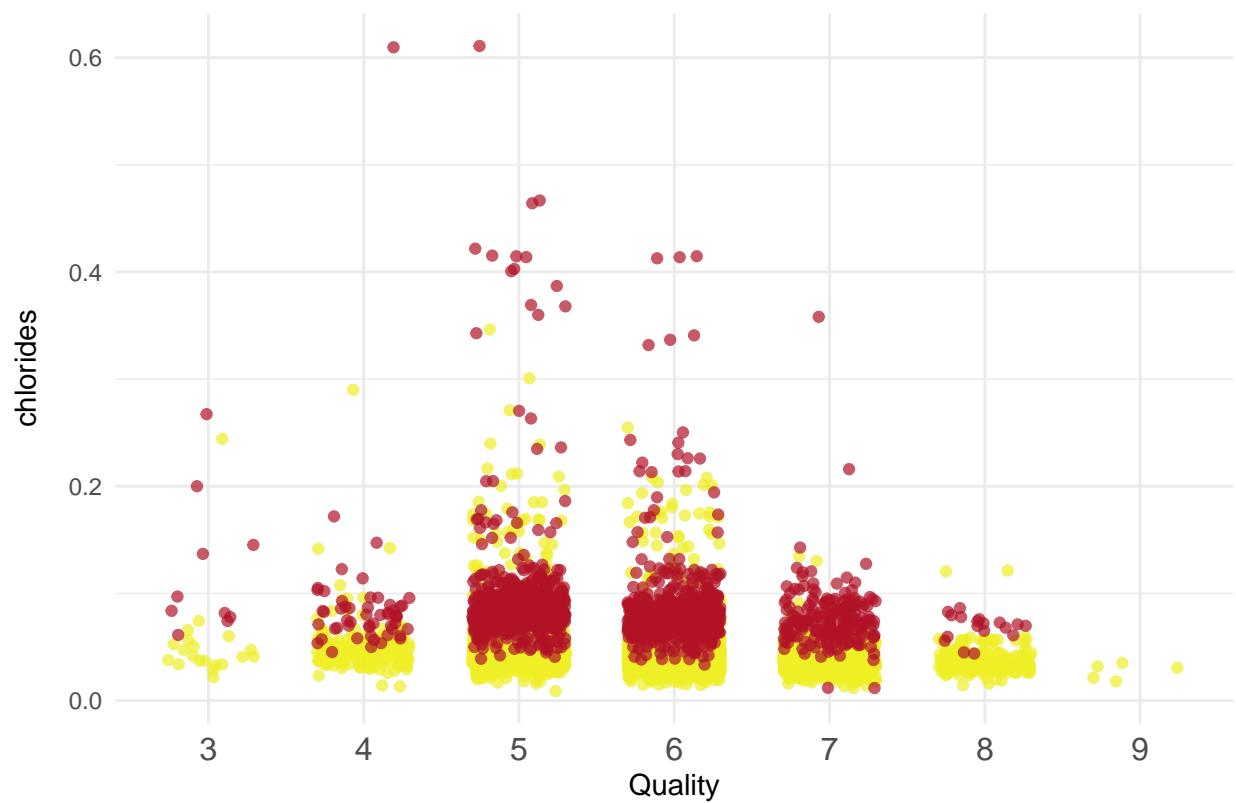
Distribution of residual\_sugar by Quality



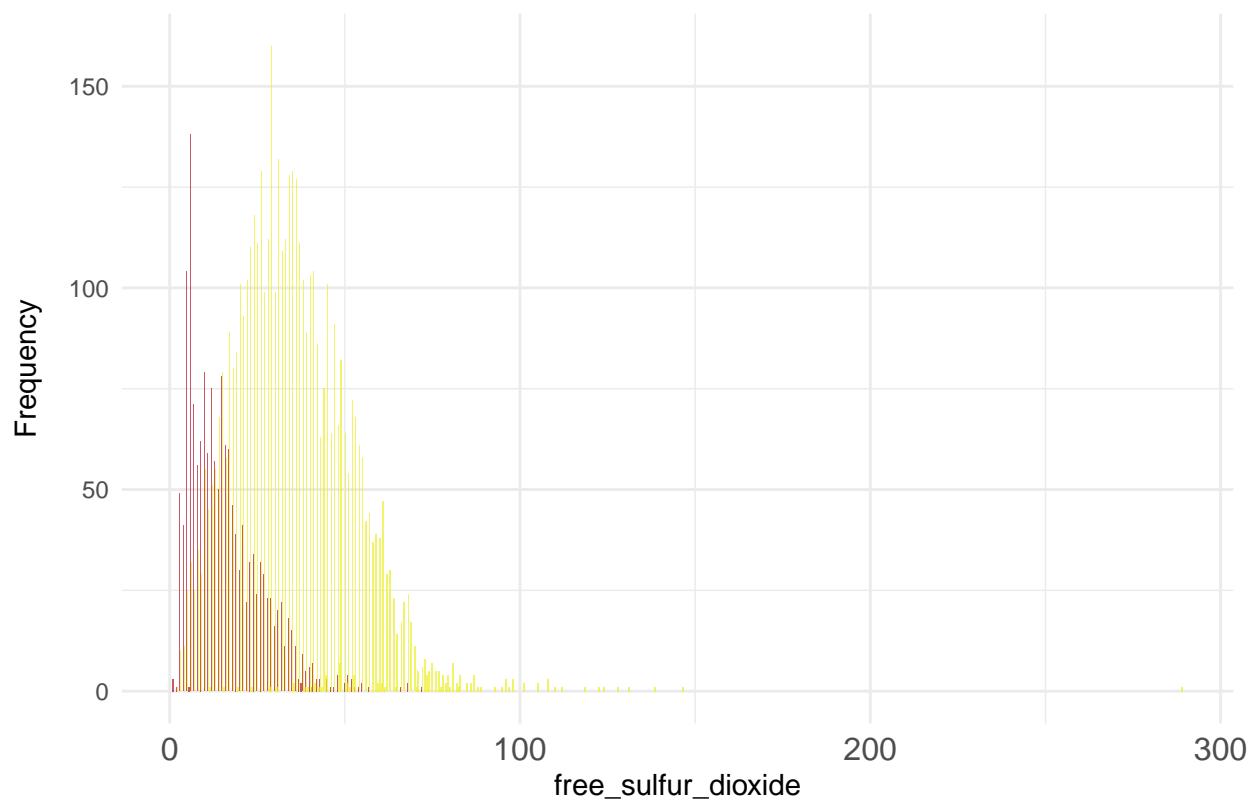
### Frequency of chlorides by Type



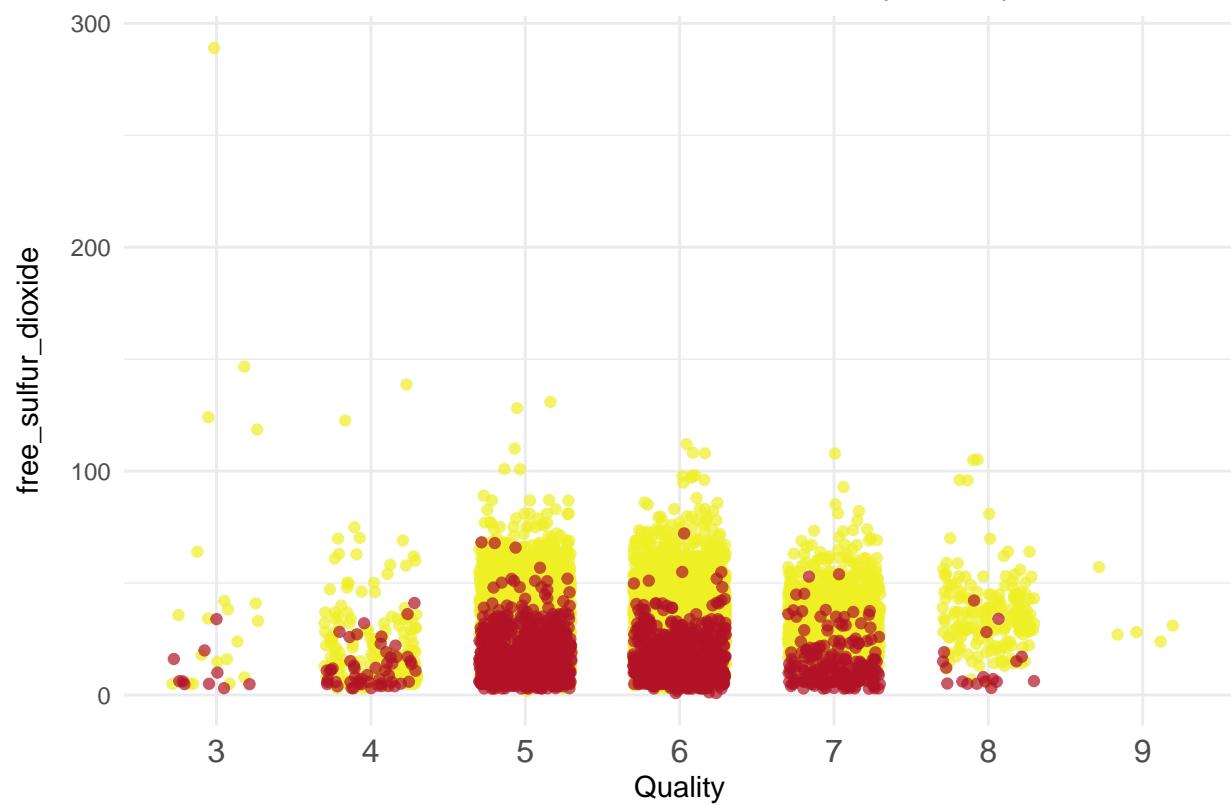
### Distribution of chlorides by Quality



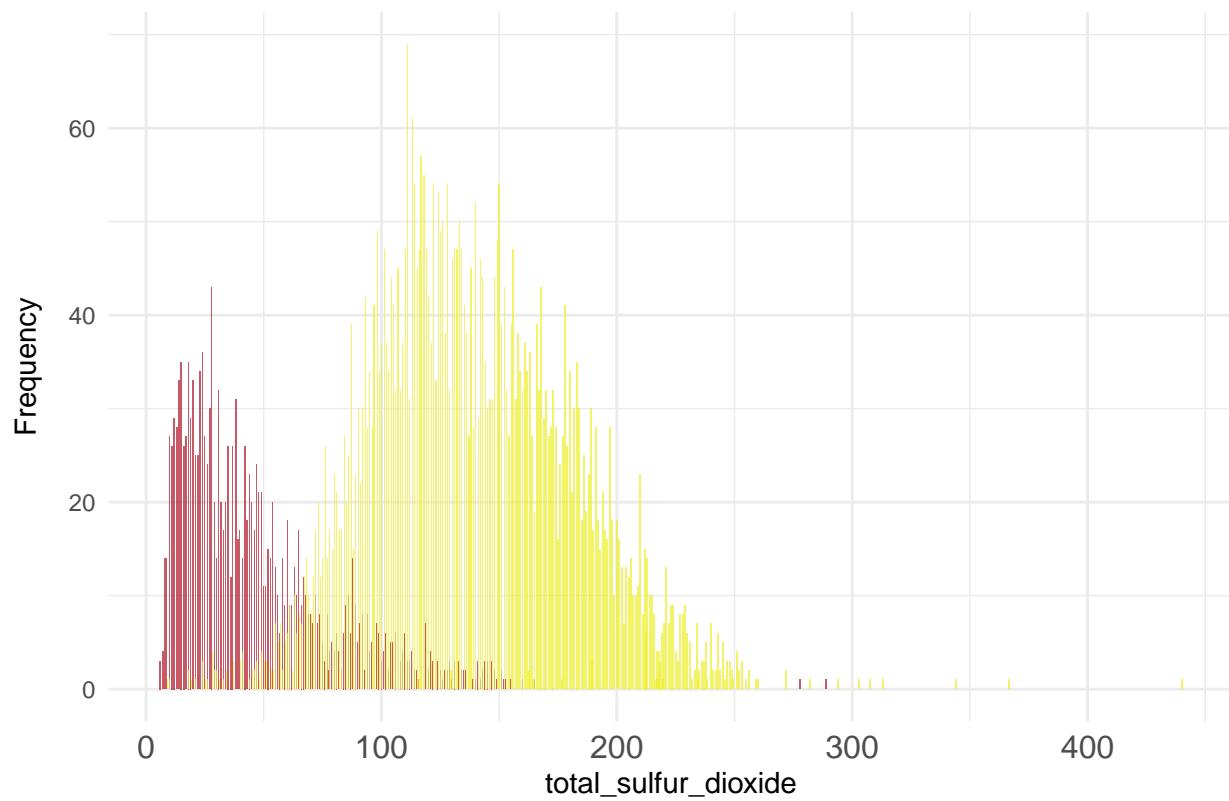
Frequency of free\_sulfur\_dioxide by Type



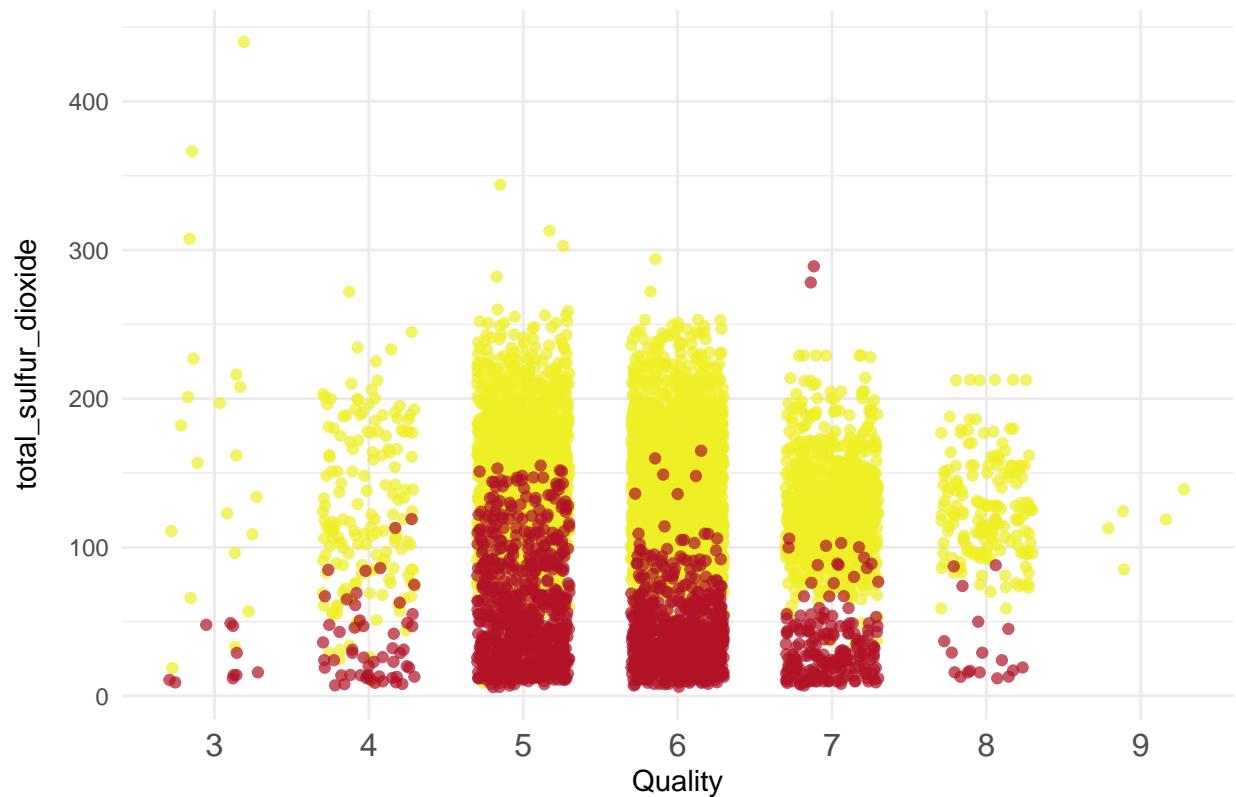
Distribution of free\_sulfur\_dioxide by Quality



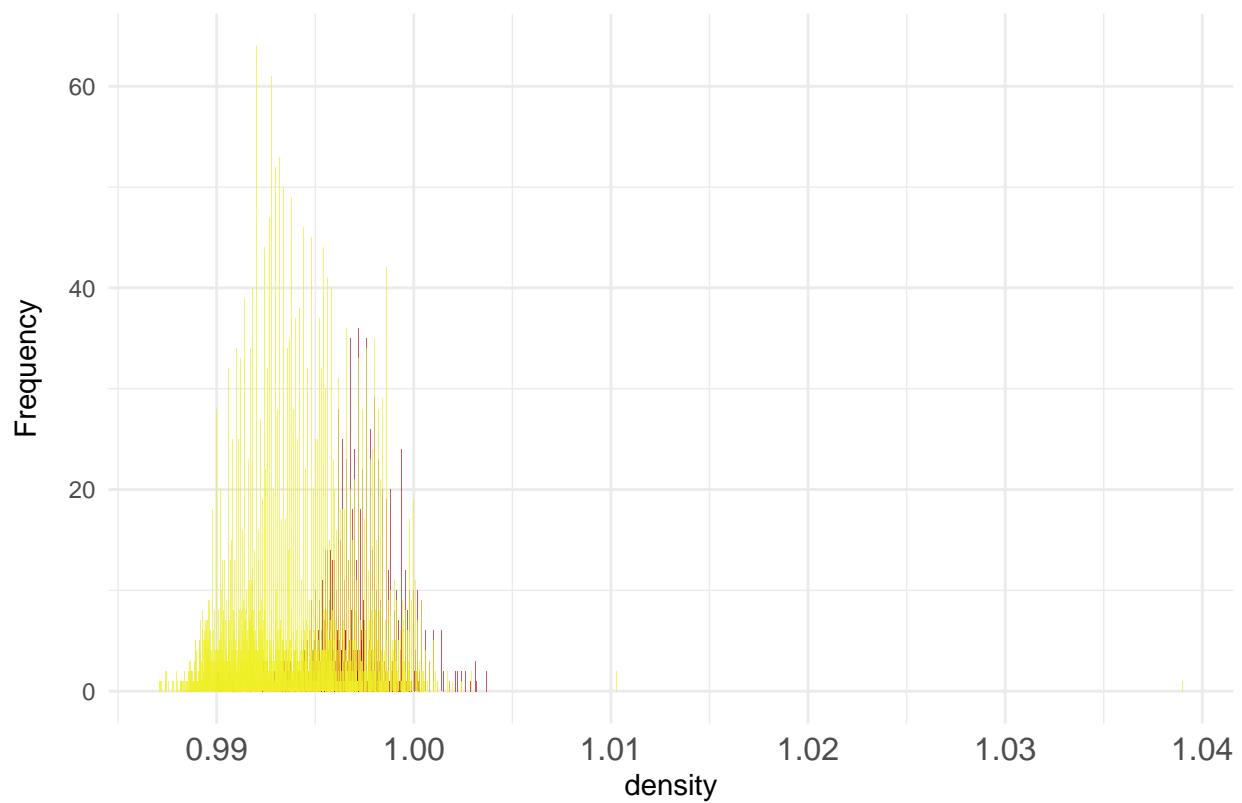
Frequency of total\_sulfur\_dioxide by Type



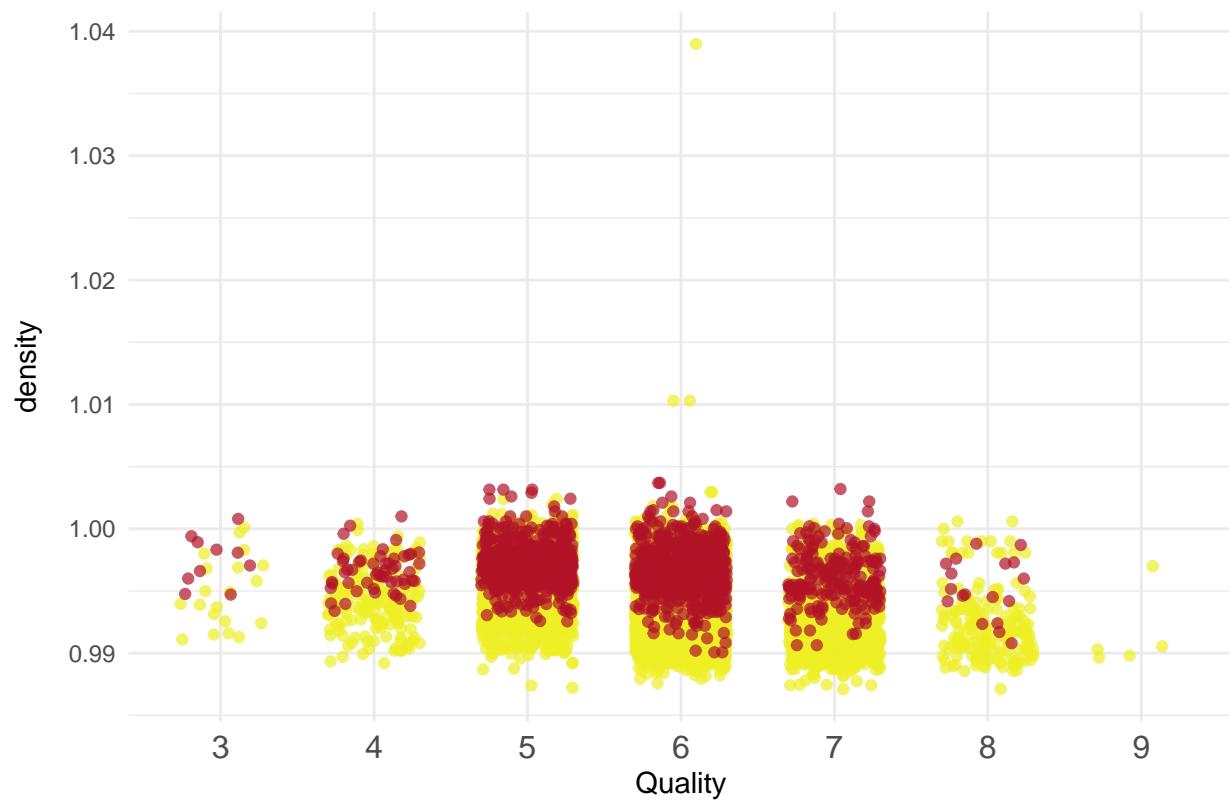
Distribution of total\_sulfur\_dioxide by Quality



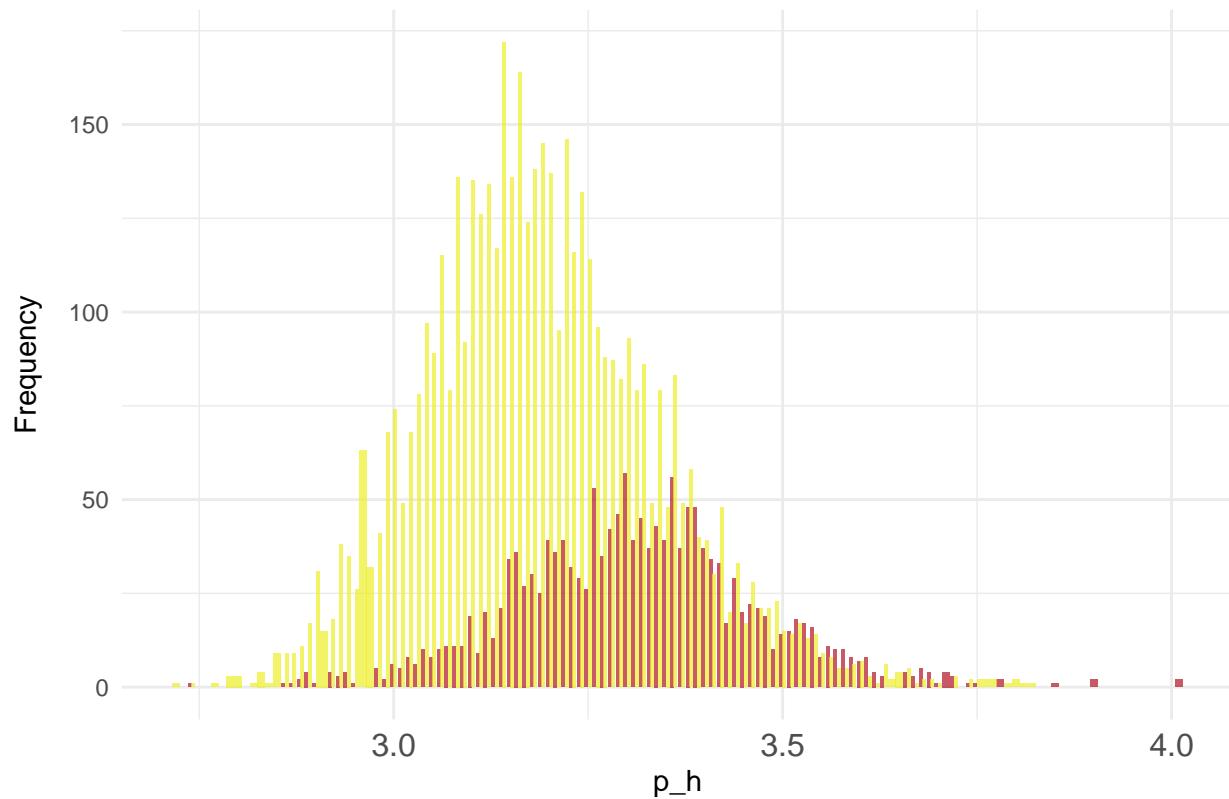
Frequency of density by Type



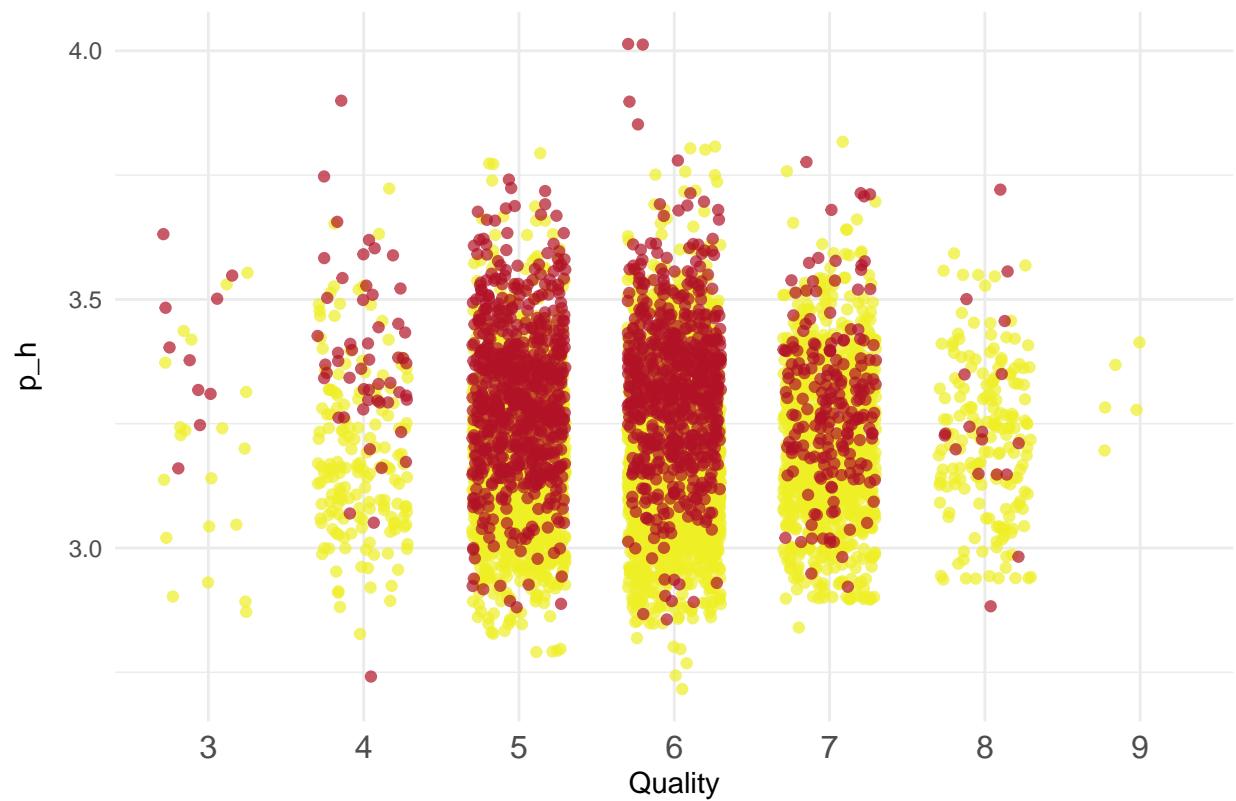
Distribution of density by Quality



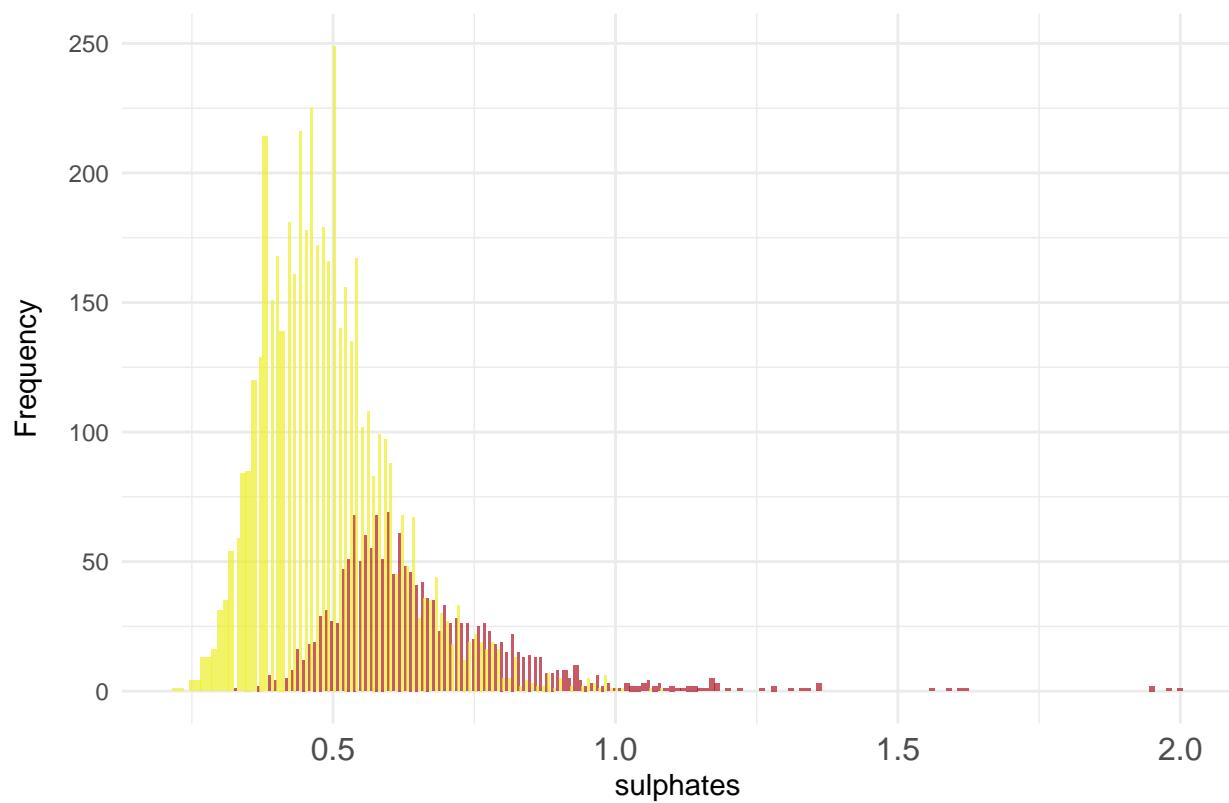
Frequency of p\_h by Type



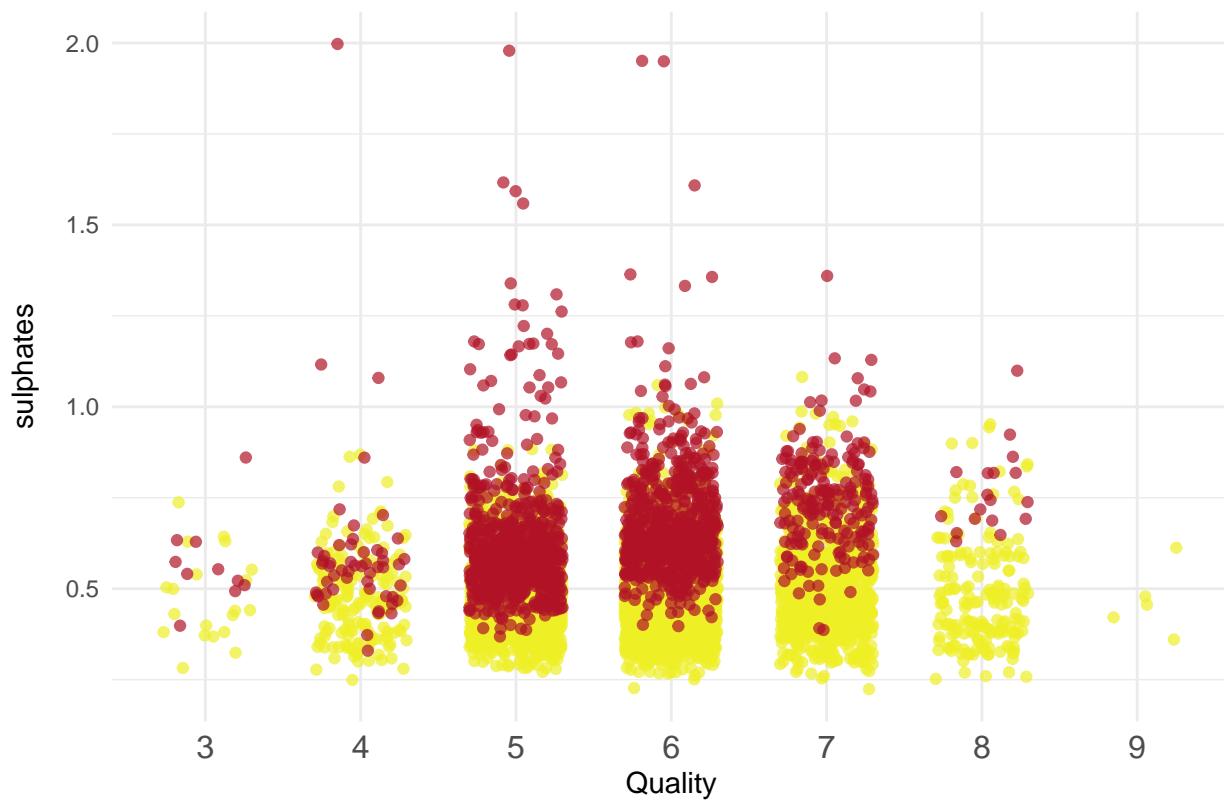
Distribution of p\_h by Quality



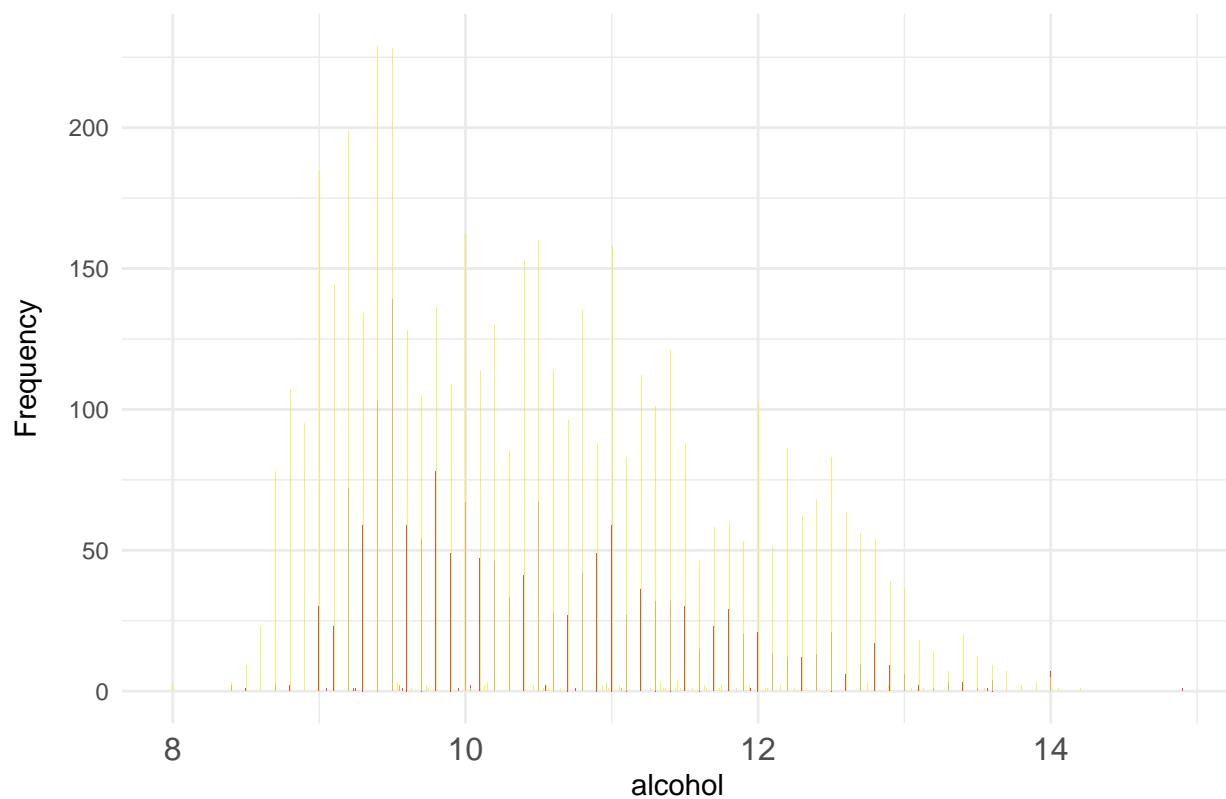
Frequency of sulphates by Type



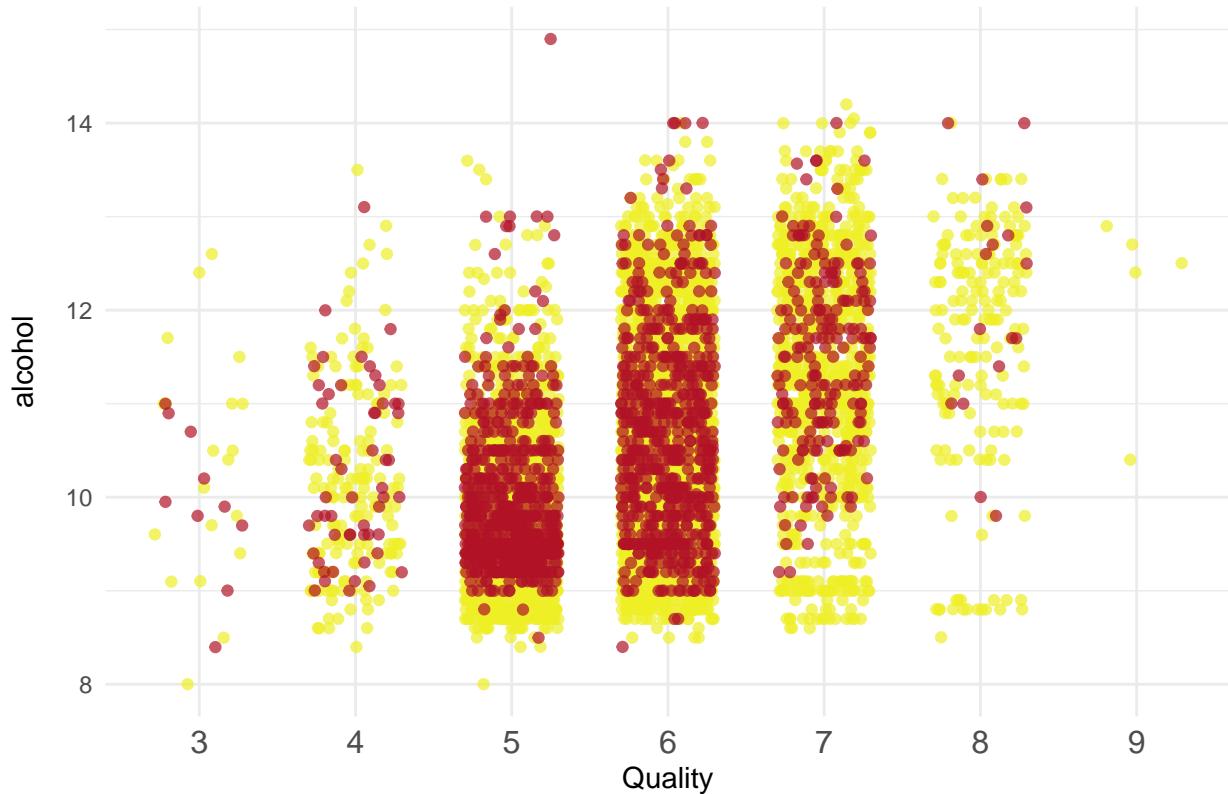
### Distribution of sulphates by Quality



Frequency of alcohol by Type



Distribution of alcohol by Quality



### Giá trị khuyết

Kiểm tra giá trị khuyết

```
missing_summary <- wine |>
  summarise_all(~ sum(is.na(.)))

print(missing_summary, width = Inf)

## # A tibble: 1 x 13
##   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
##       <int>           <int>      <int>          <int>      <int>
## 1            0              0          0            0            0
##   free_sulfur_dioxide total_sulfur_dioxide density    p_h sulphates alcohol
##       <int>           <int>      <int>     <int>      <int>      <int>
## 1            0              0            0            0            0            0
##   quality type
##   <int> <int>
## 1      0    0
```

Nhận xét: Kết quả cho thấy bộ dữ liệu không có giá trị khuyết.

### Giá trị ngoại lai

Phương pháp dò tìm và xử lý giá trị ngoại lai ở mục này dựa trên mục **4.7. Detecting Outliers and Cleaning Data** trong sách Applied Multivariate Statistical Analysis của Richard Johnson và Dean Wichern và **Fig. 1.6.** trong sách An Introduction to Applied Multivariate Analysis with R của Brian Everitt và

Torsten Hothorn.

Kiểm tra dấu hiệu univariate outlier

```
wine_long <- reshape2::melt(wine, id.vars = c("type", "quality"))

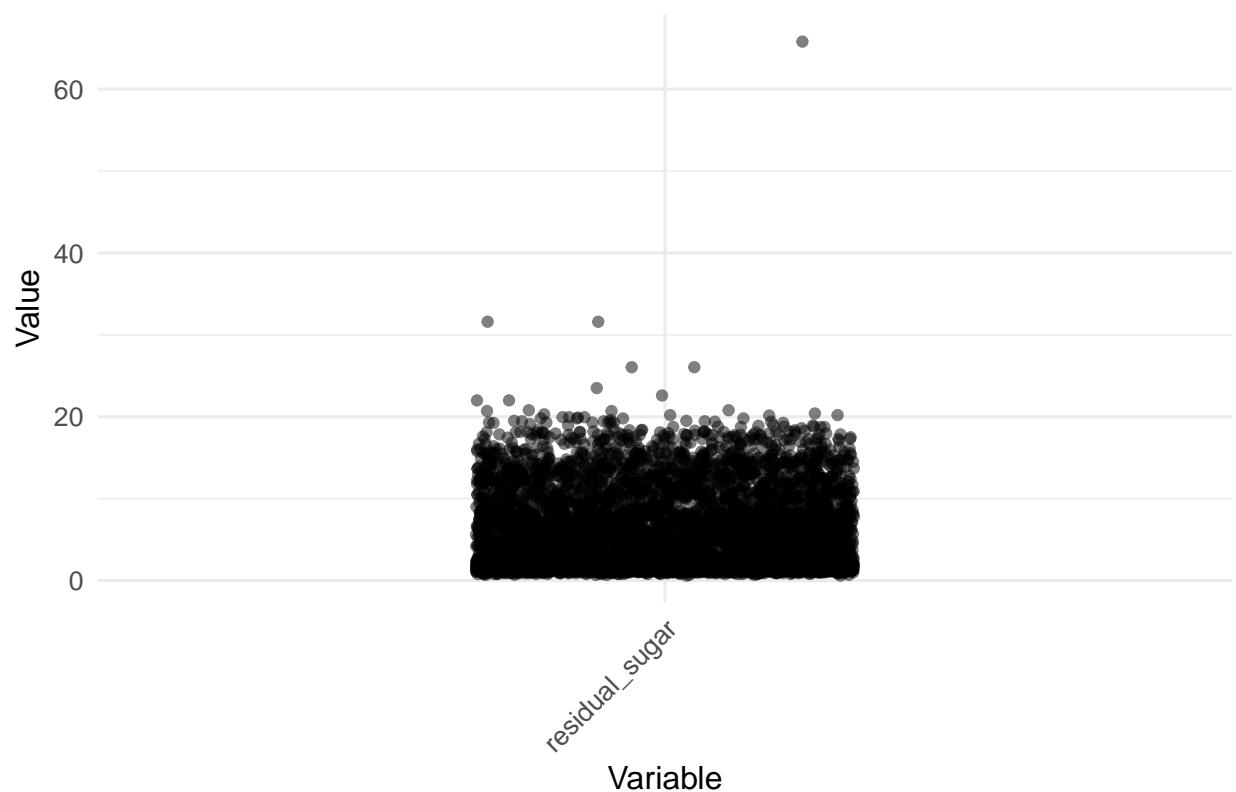
var_groups <- list(
  "residual_sugar" = "residual_sugar",
  "free_sulfur_dioxide, and total_sulfur_dioxide" = c("free_sulfur_dioxide", "total_sulfur_dioxide"),
  "fixed_acidity and alcohol" = c("fixed_acidity", "alcohol"),
  "volatile_acidity, citric_acid, chlorides, and sulphates" = c("volatile_acidity",
    "citric_acid", "chlorides", "sulphates"),
  "density" = "density",
  "p_h" = "p_h"
)

for (group in names(var_groups)) {
  vars <- var_groups[[group]]
  wine_long_subset <- if (length(vars) > 1) {
    wine_long[wine_long$variable %in% vars, ]
  } else {
    wine_long[wine_long$variable == vars, ]
  }

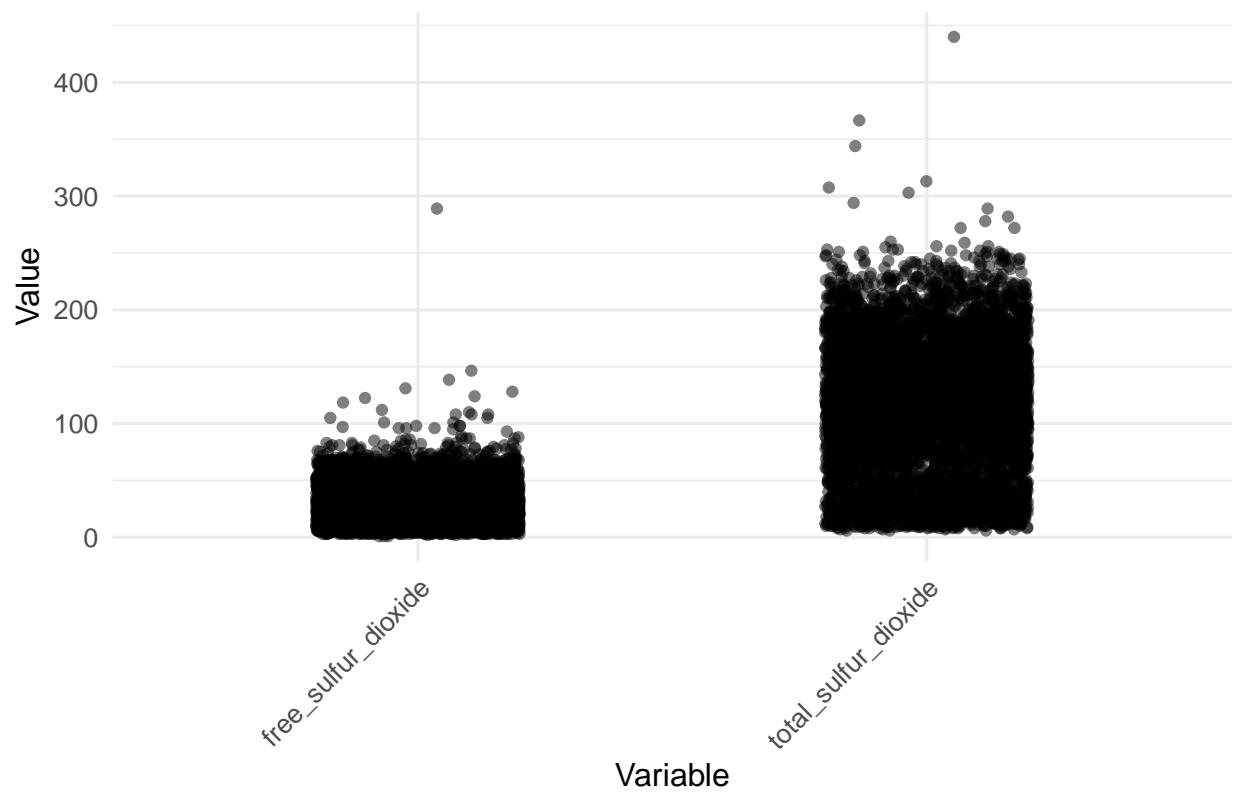
  p <- ggplot(wine_long_subset, aes(x = variable, y = value)) +
    geom_jitter(width = 0.2, height = 0, alpha = 0.5) +
    labs(title = paste("Dot Plot for", group),
        x = "Variable",
        y = "Value") +
    theme_minimal(base_size = 12) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(p)
}
```

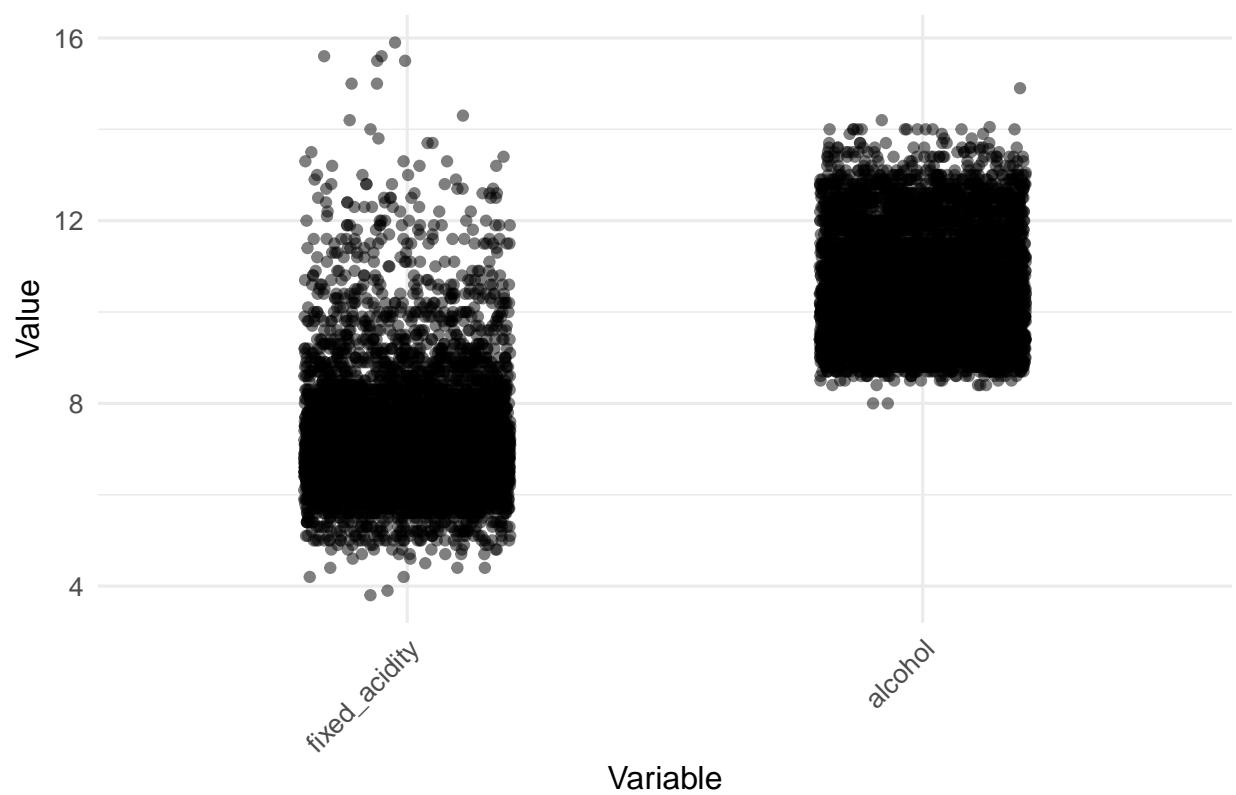
Dot Plot for residual\_sugar



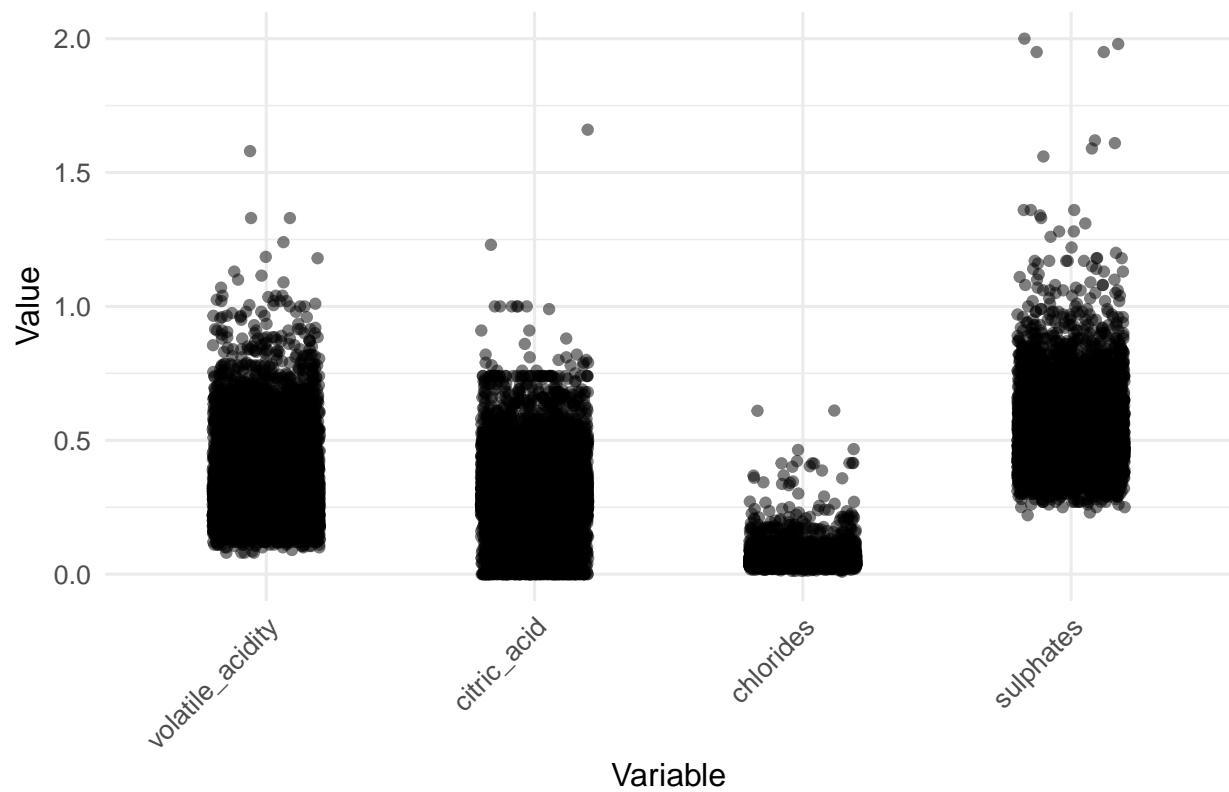
Dot Plot for free\_sulfur\_dioxide, and total\_sulfur\_dioxide



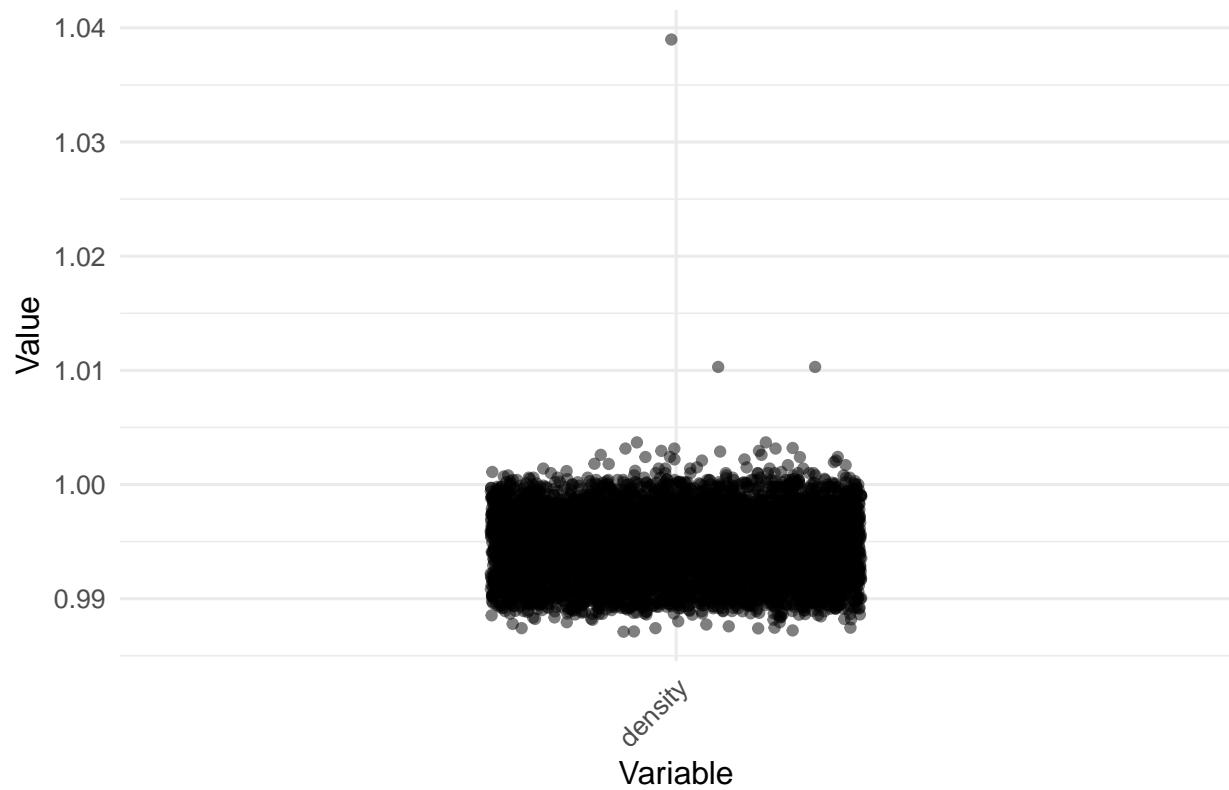
Dot Plot for fixed\_acidity and alcohol



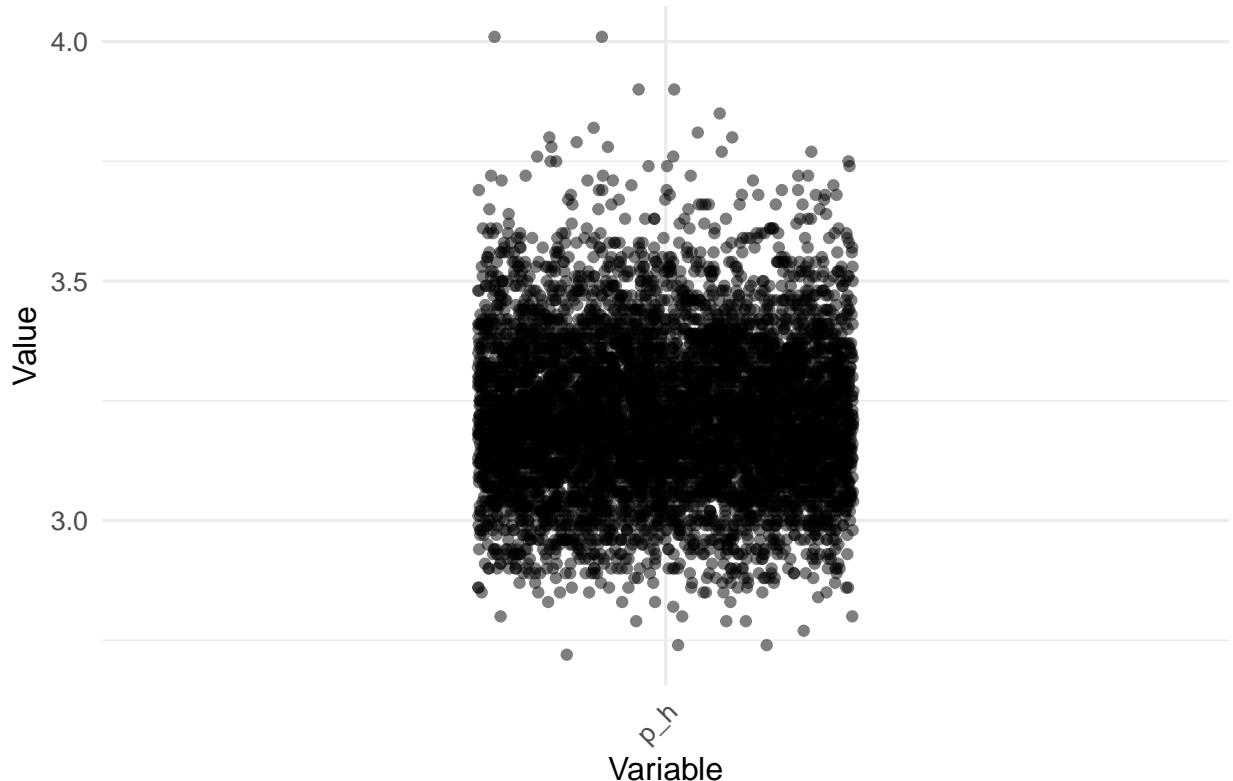
Dot Plot for volatile\_acidity, citric\_acid, chlorides, and sulphates



Dot Plot for density



## Dot Plot for p\_h



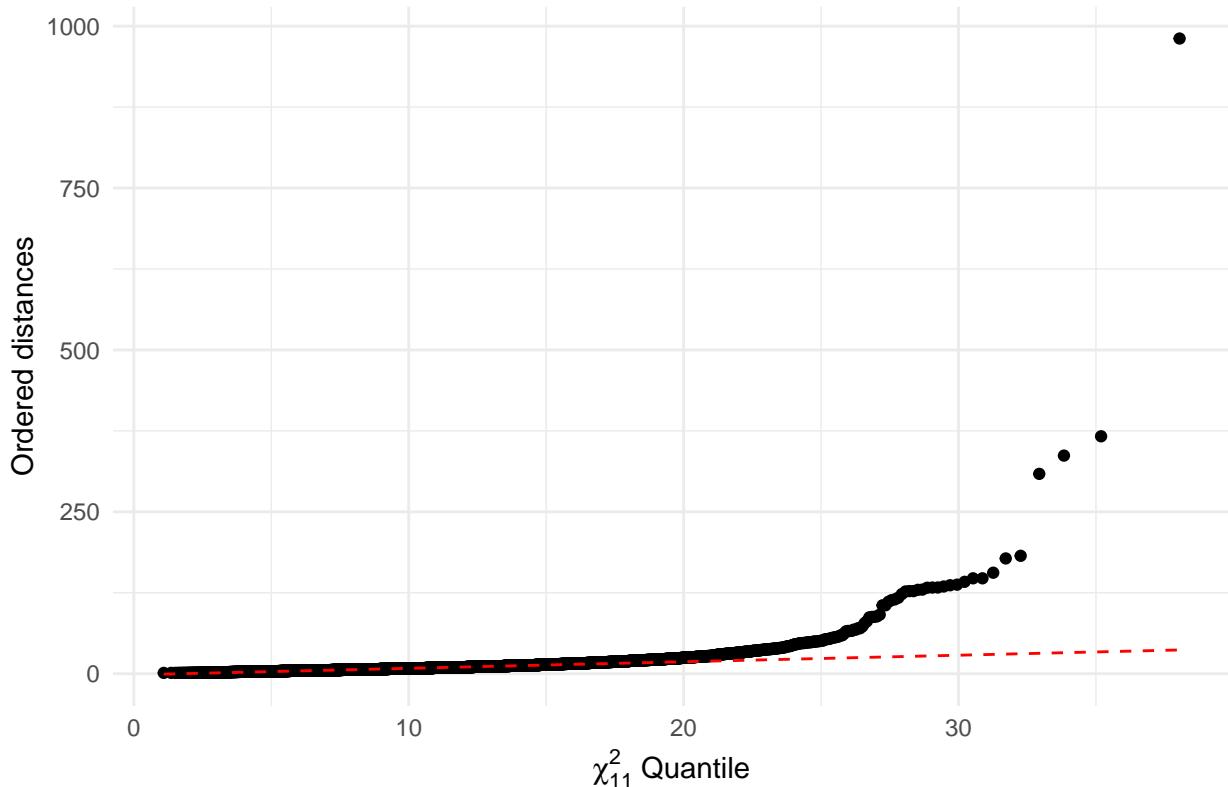
Nhận xét: Có thể quan sát được vài biến có giá trị ngoại lai đơn biến.

Kiểm tra dấu hiệu multivariate outlier

```
# Compute Mahalanobis distances
x <- wine |>
  select_if(is.numeric)
cm <- colMeans(x)
S <- cov(x)
d <- apply(x, 1, function(x) t(x - cm) %*% solve(S) %*% (x - cm))
dist_data <- data.frame(distances = d)

# Create the Q-Q plot
ggplot(dist_data, aes(sample = distances)) +
  stat_qq(distribution = qchisq, dparams = list(df = 11)) +
  stat_qq_line(distribution = qchisq, dparams = list(df = 11),
               linetype = "dashed", color = "red") +
  ggtitle("Chi-Square Q-Q Plot of Mahalanobis Distances") +
  xlab(expression(paste(chi[11]^2, " Quantile"))) +
  ylab("Ordered distances") +
  theme_minimal()
```

### Chi-Square Q-Q Plot of Mahalanobis Distances



Nhận xét: Bộ dữ liệu rõ ràng tồn tại giá trị ngoại lai nhiều chiều.

Xử lý giá trị ngoại lai bằng cách loại bỏ theo giá trị chuẩn hóa và khoảng cách Mahalanobis

```
# Calculate standardized values for numeric variables
standardized_wine <- wine |>
  mutate(across(where(is.numeric), ~ (. - mean(.)) / sd(.), .names = "z_{col}"))

# Calculate Mahalanobis distances
numeric_data <- wine |>
  select_if(is.numeric)
means <- colMeans(numeric_data)
cov_matrix <- cov(numeric_data)
mahalanobis_distances <- mahalanobis(numeric_data, means, cov_matrix)

# Add Mahalanobis distances to the data frame
wine_with_distances <- standardized_wine |>
  mutate(mahalanobis_distances = mahalanobis_distances)

# Identify outliers based on standardized values
z_threshold <- 4
chi_sq_threshold <- qchisq(0.95, df = ncol(numeric_data))

outliers_standardized <- wine_with_distances |>
  rowwise() |>
  mutate(is_outlier_standardized = any(c_across(starts_with("z_")) > z_threshold | c_across(starts_with
ungroup()
```

Phần trăm số quan trắc bị loại bỏ

```
(1 - nrow(cleaned_data)/nrow(wine))*100
```

```
## [1] 8.126828
```

**Nhân xét:** Số quan trắc bị loại bỏ xấp xỉ 8% là chấp nhận được (dưới 10%).

Kiểm tra tổng quát dữ liệu sau khi làm sạch

Kiểm tra lại dấu hiệu univariate outlier:

```
cleaned_long <- reshape2::melt(cleaned_data, id.vars = c("type", "quality"))

var_groups <- list(
  "residual_sugar" = "residual_sugar",
  "free_sulfur_dioxide, and total_sulfur_dioxide" = c( "free_sulfur_dioxide",
    "fixed acidity and alcohol" = c("fixed acidity", "alcohol").
```

```

"volatile_acidity, citric_acid, chlorides, and sulphates" = c("volatile_acidity", "citric_acid", "chl
"density" = "density",
"p_h" = "p_h"
)

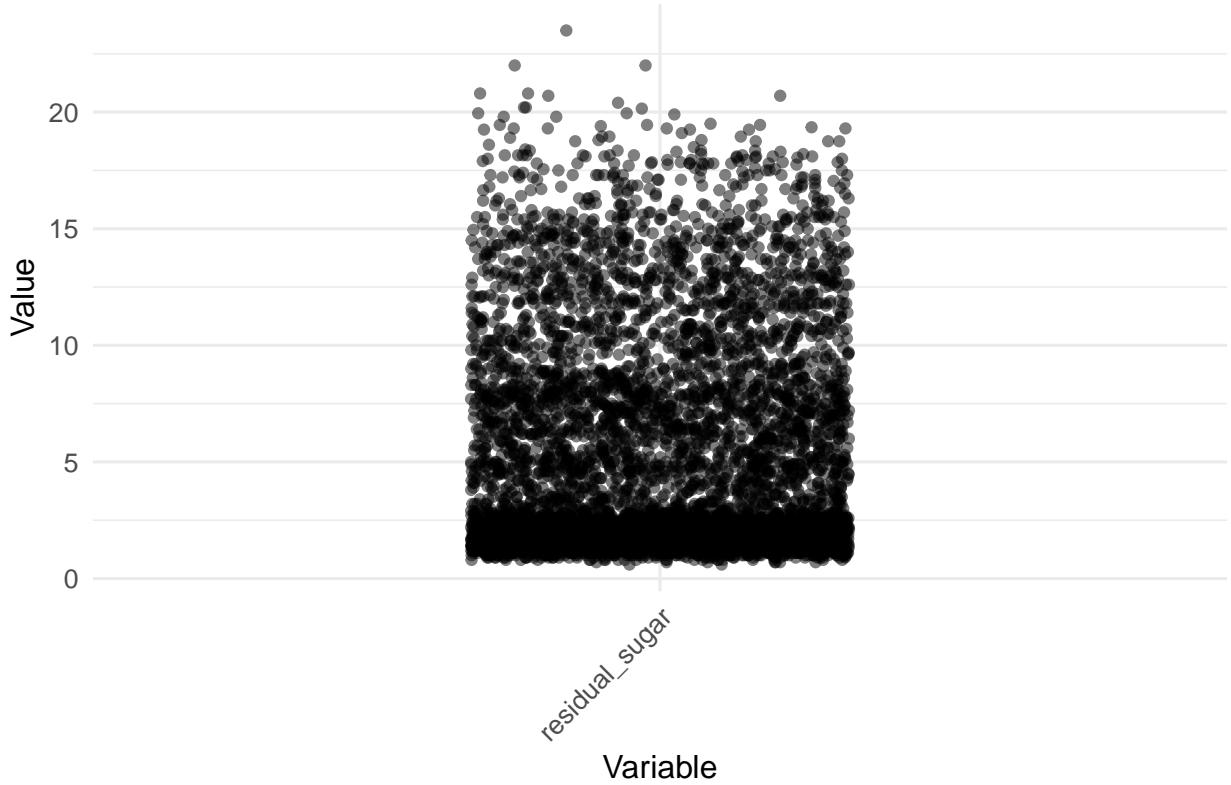
for (group in names(var_groups)) {
  vars <- var_groups[[group]]
  cleaned_long_subset <- if (length(vars) > 1) {
    cleaned_long[cleaned_long$variable %in% vars, ]
  } else {
    cleaned_long[cleaned_long$variable == vars, ]
  }

  p <- ggplot(cleaned_long_subset, aes(x = variable, y = value)) +
    geom_jitter(width = 0.2, height = 0, alpha = 0.5) +
    labs(title = paste("Dot Plot for", group),
        x = "Variable",
        y = "Value") +
    theme_minimal(base_size = 12) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

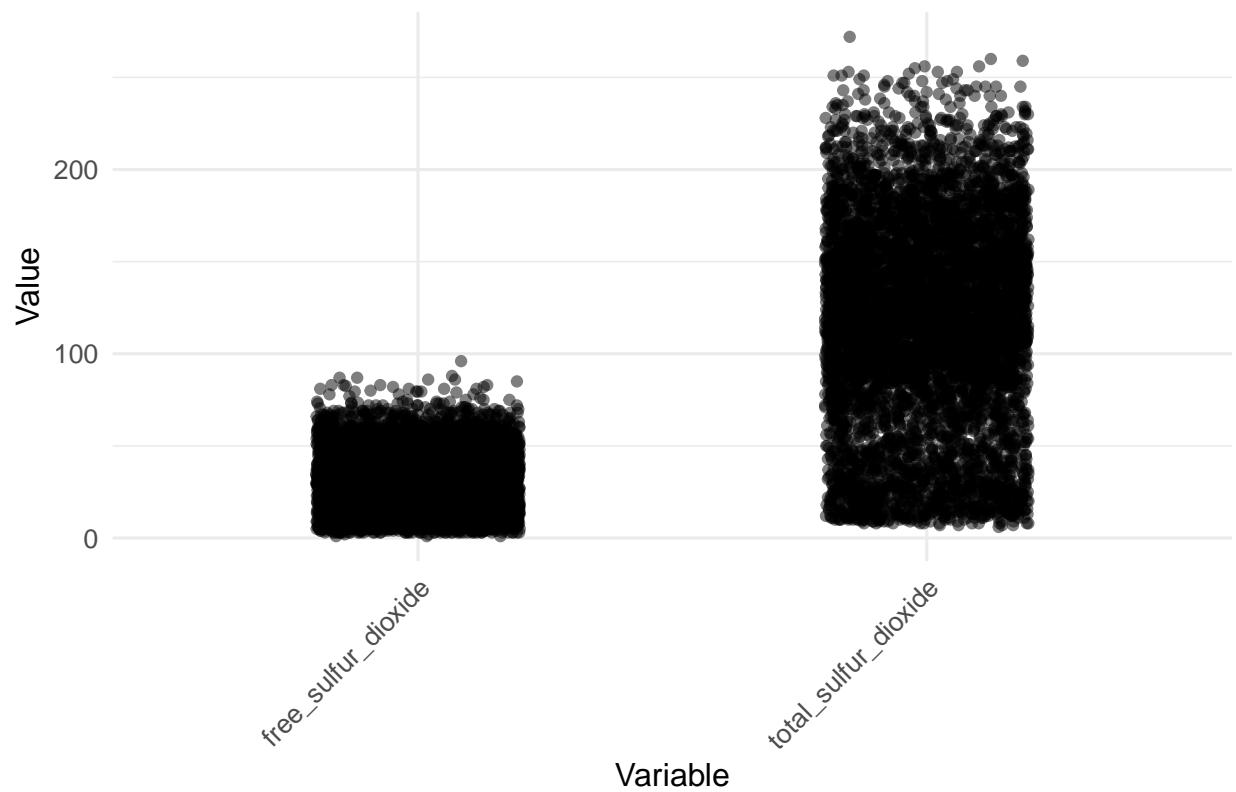
  print(p)
}

```

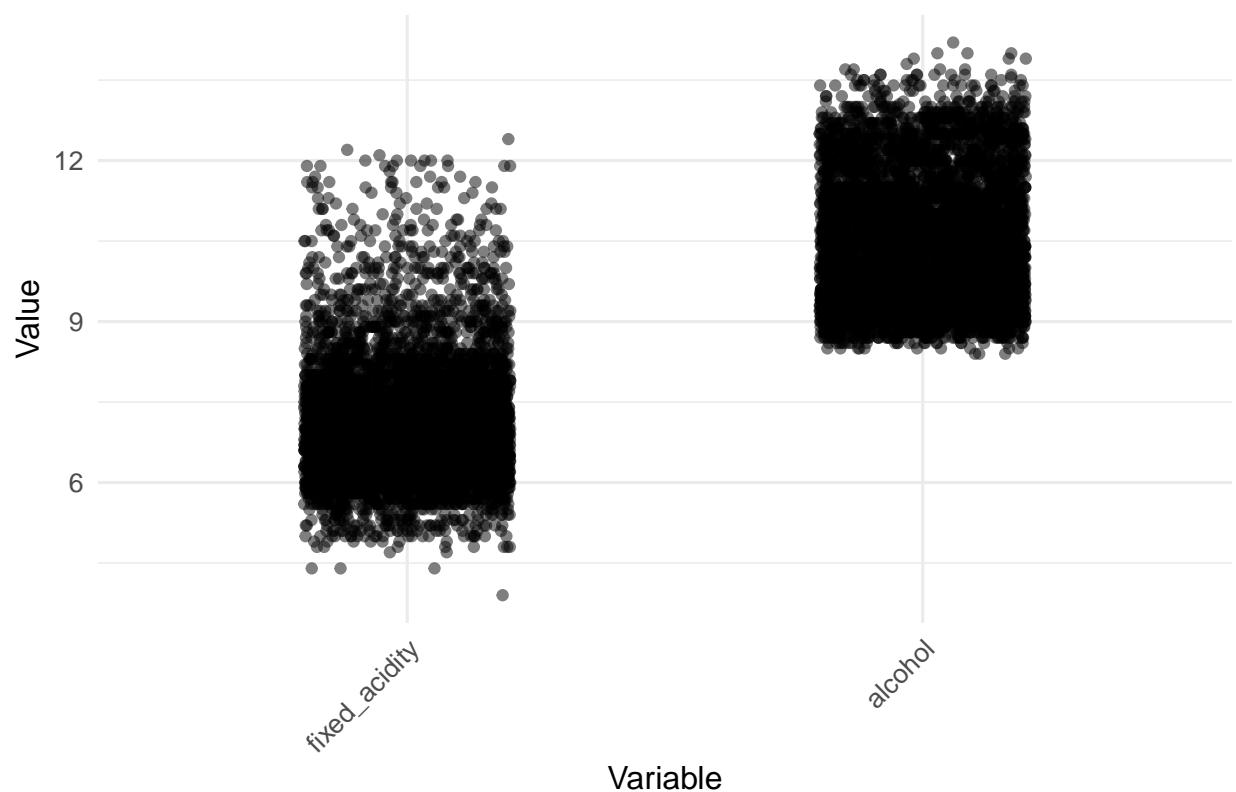
Dot Plot for residual\_sugar



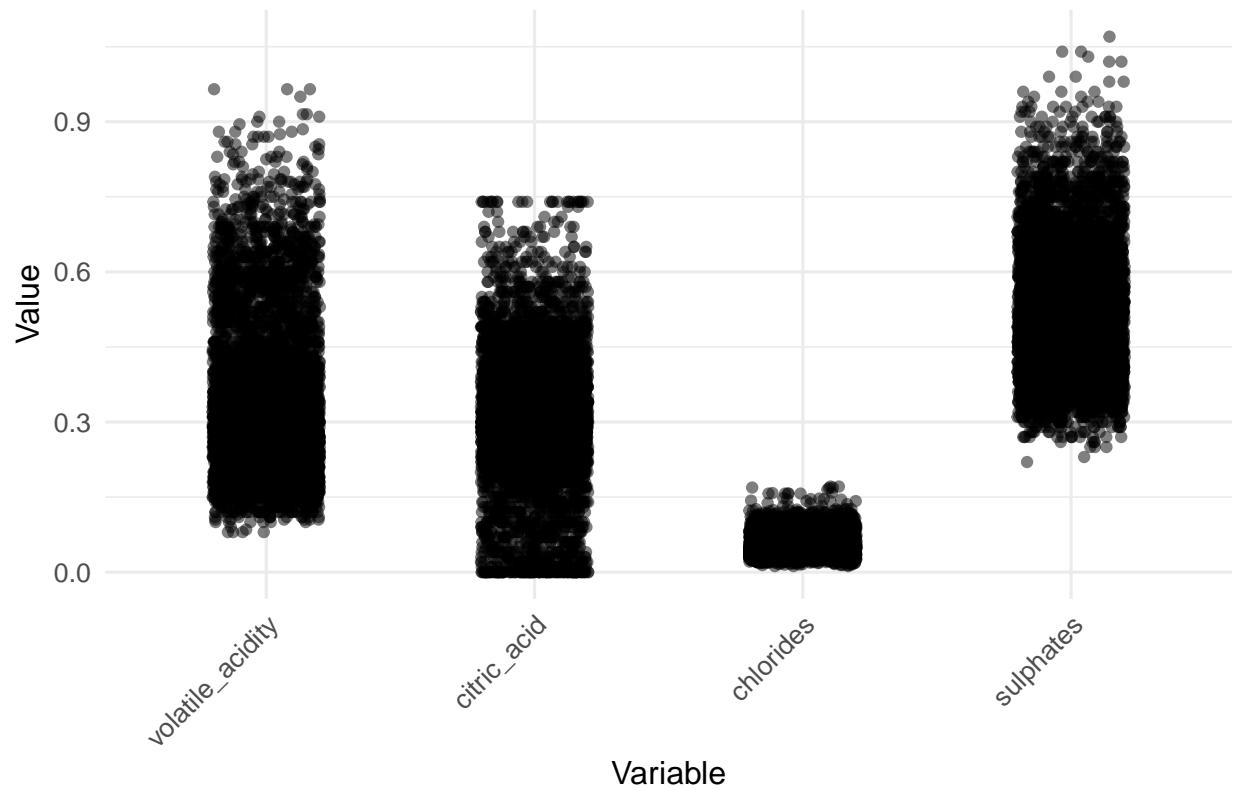
Dot Plot for free\_sulfur\_dioxide, and total\_sulfur\_dioxide



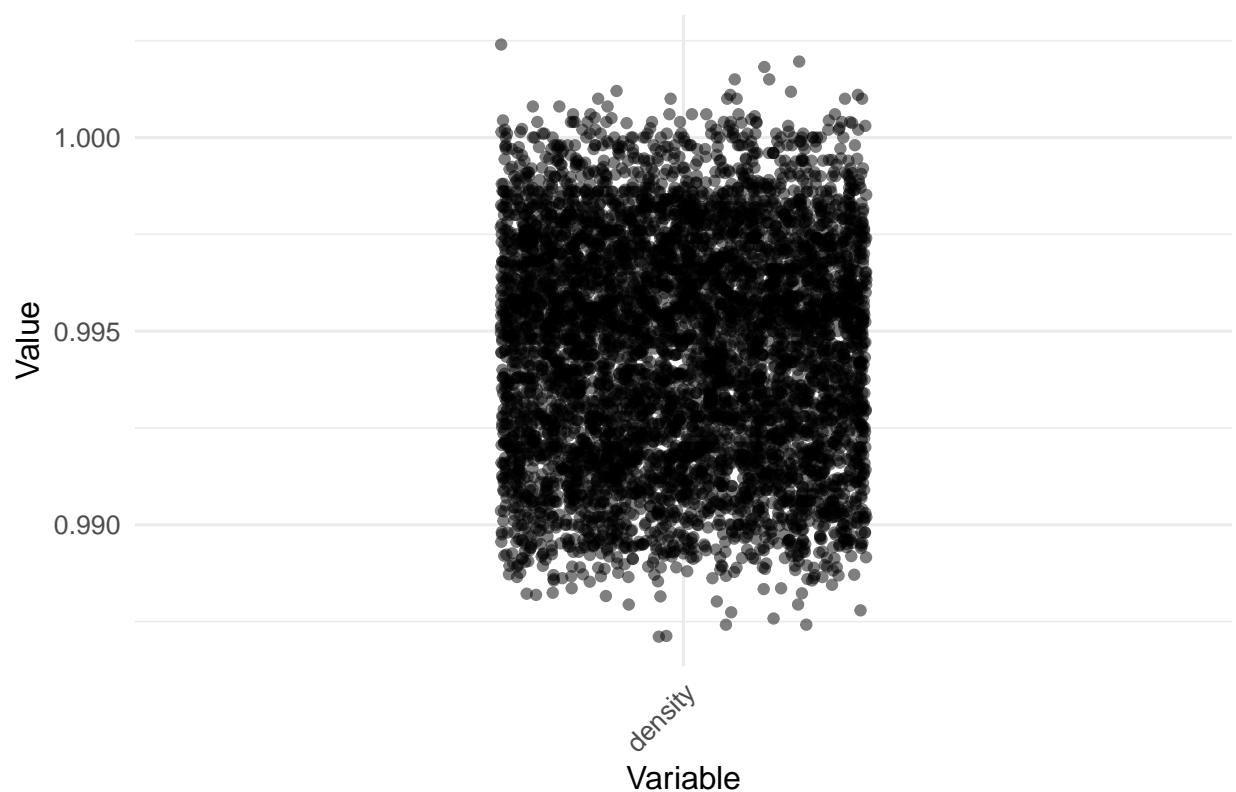
Dot Plot for fixed\_acidity and alcohol



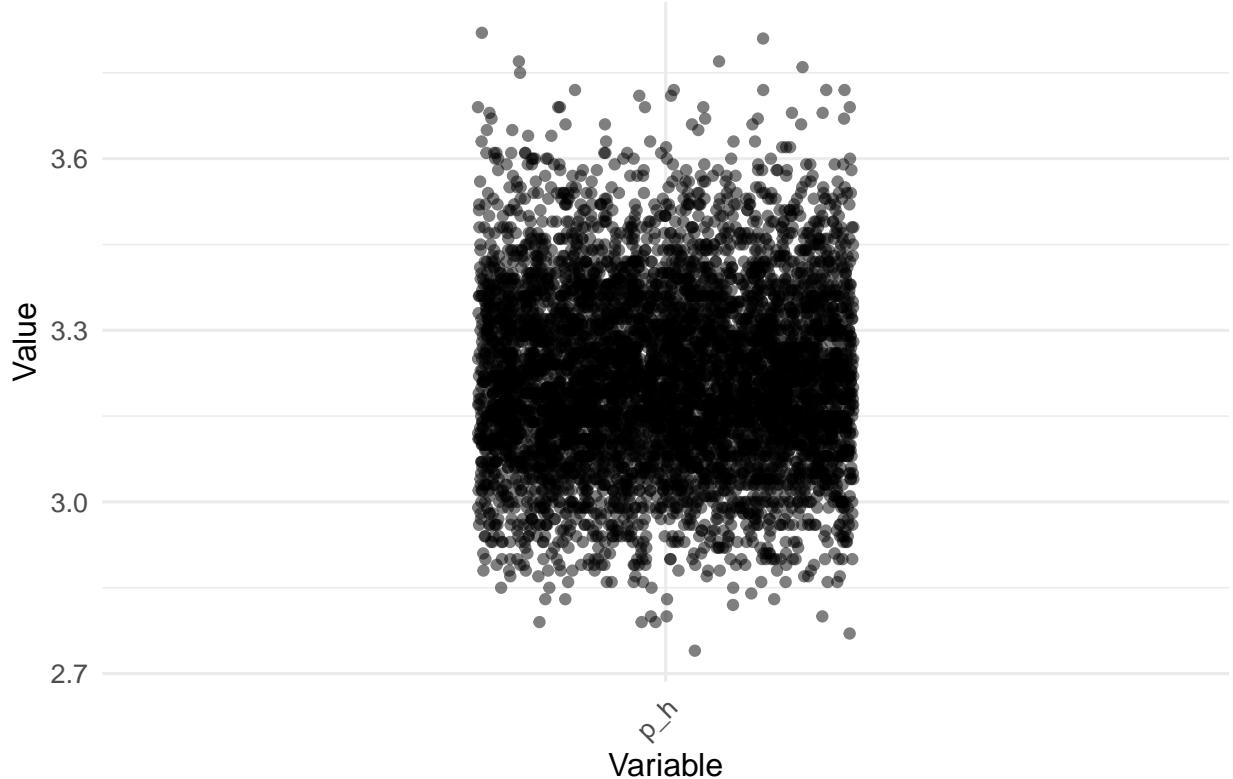
Dot Plot for volatile\_acidity, citric\_acid, chlorides, and sulphates



Dot Plot for density



## Dot Plot for p\_h



Nhận xét: Số lượng ngoại lai đơn biến giảm đi đáng kể sau khi làm sạch.

Kiểm tra lại dấu hiệu multivariate outlier

```
# Select numeric columns
x <- cleaned_data |>
  select(where(is.numeric))

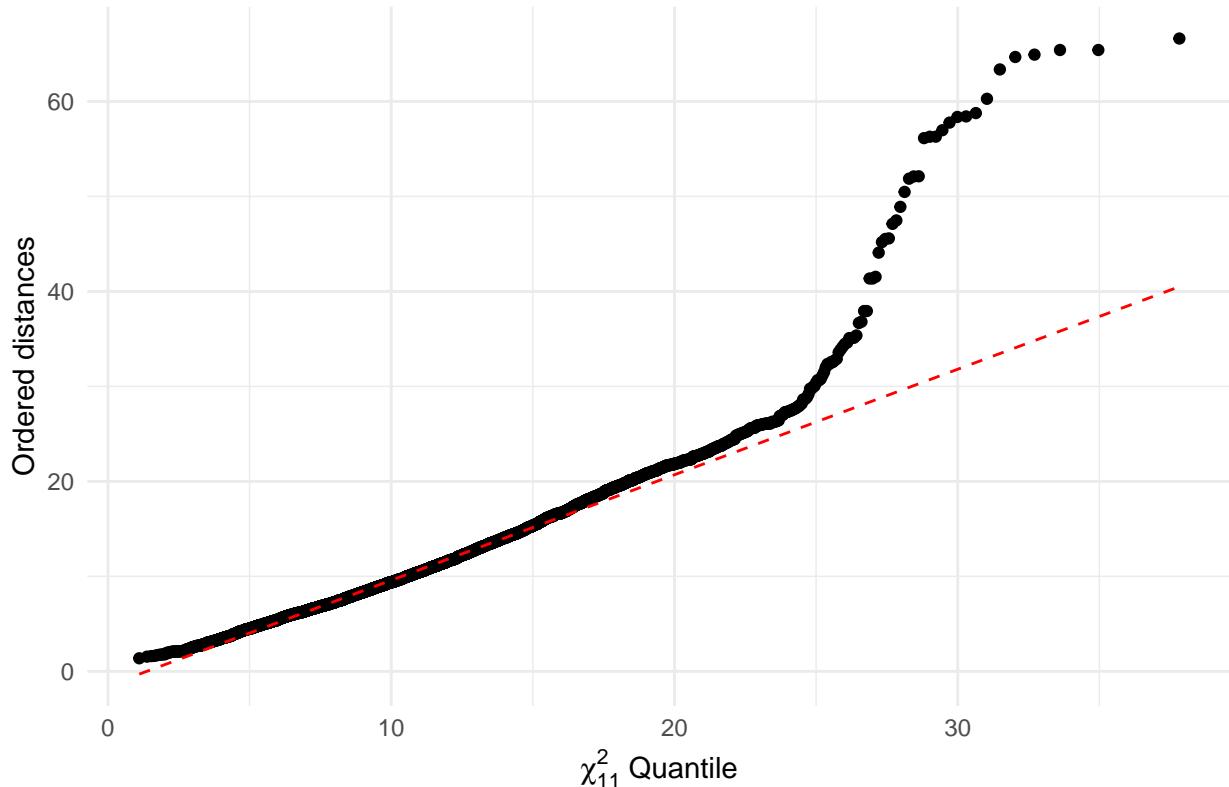
# Compute column means and covariance matrix
cm <- colMeans(x)
S <- cov(x)

# Compute Mahalanobis distances
d <- apply(x, 1, function(x) t(x - cm) %*% solve(S) %*% (x - cm))

# Convert distances to a data frame
dist_data <- data.frame(distances = d)

# Create the Q-Q plot
ggplot(dist_data, aes(sample = distances)) +
  stat_qq(distribution = qchisq, dparams = list(df = 11)) +
  stat_qq_line(distribution = qchisq, dparams = list(df = 11), linetype = "dashed", color = "red") +
  ggtitle("Chi-Square Q-Q Plot of Mahalanobis Distances") +
  xlab(expression(paste(chi[11]^2, " Quantile"))) +
  ylab("Ordered distances") +
  theme_minimal()
```

## Chi–Square Q–Q Plot of Mahalanobis Distances



**Nhận xét:** Không còn những điểm khoảng cách Malahanobis nằm lệch lề loi phia trên bên phải như lúc ban đầu. Dấu hiệu giá trị ngoại lai nhiều chiều đã biến mất.

So sánh phân phối trước và sau khi làm sạch

```
# Assuming cleaned_data and wine_quality are your datasets

numeric_vars <- colnames(cleaned_data)[sapply(cleaned_data, is.numeric)]

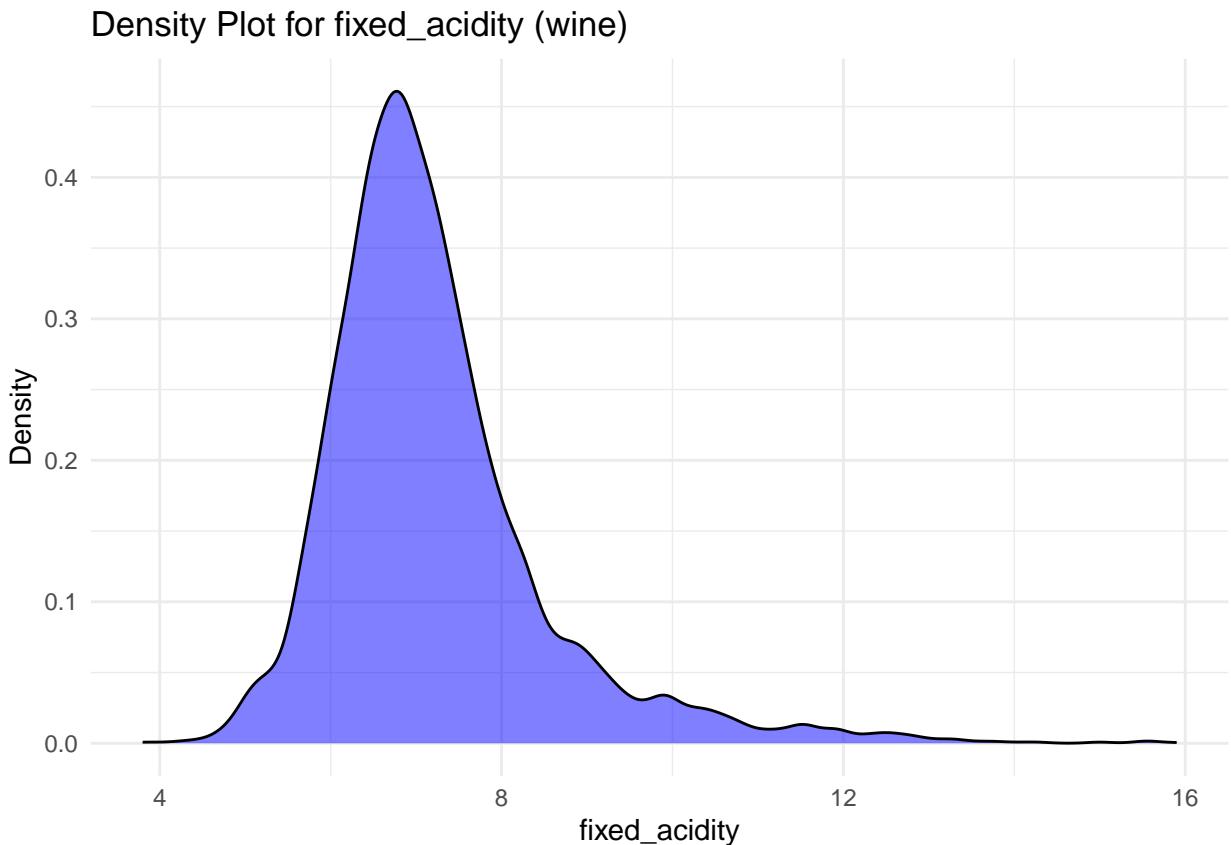
# Create a list to store ggplot objects
density_plots <- lapply(numeric_vars, function(var) {

  # Create ggplot for wine_quality
  p1 <- ggplot(wine, aes(x = !!sym(var))) +
    geom_density(fill = "blue", alpha = 0.5) +
    labs(x = var, y = "Density", title = paste("Density Plot for", var, "(wine)")) +
    theme_minimal()

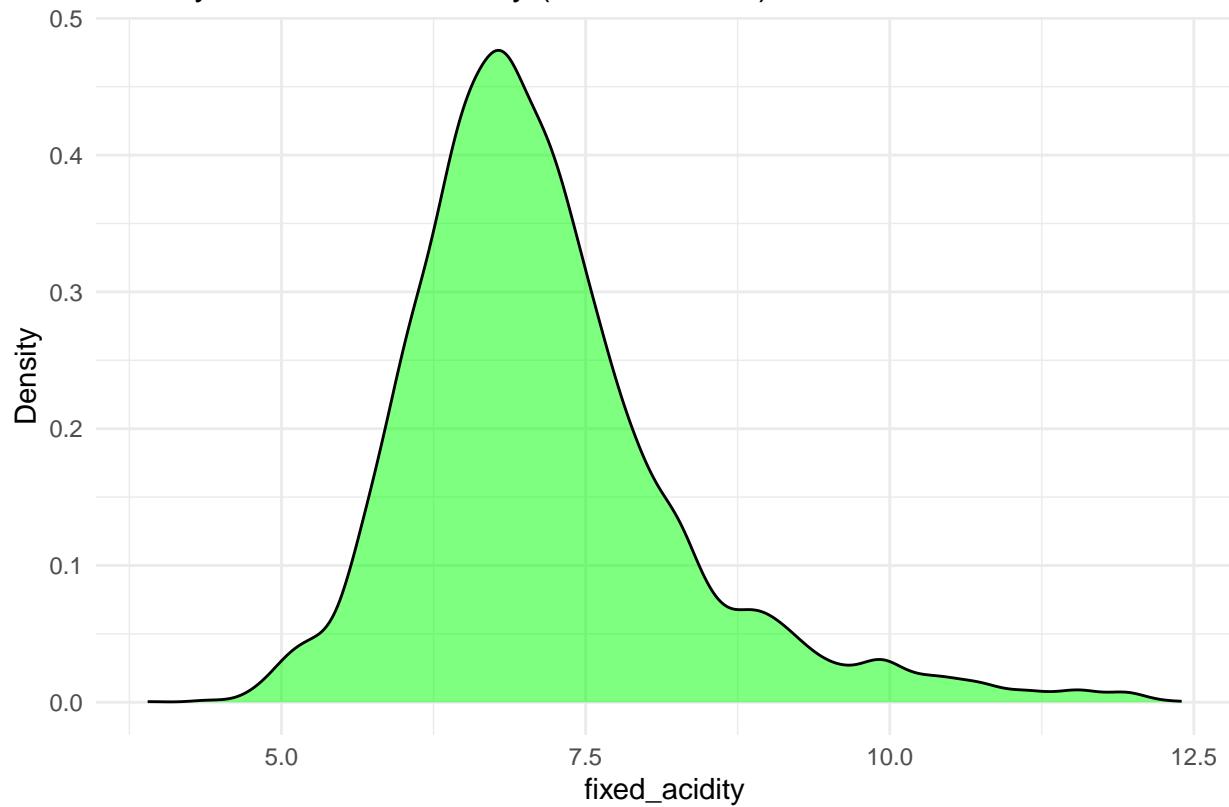
  # Create ggplot for cleaned_data
  p2 <- ggplot(cleaned_data, aes(x = !!sym(var))) +
    geom_density(fill = "green", alpha = 0.5) +
    labs(x = var, y = "Density", title = paste("Density Plot for", var, "(cleaned_data)")) +
    theme_minimal()

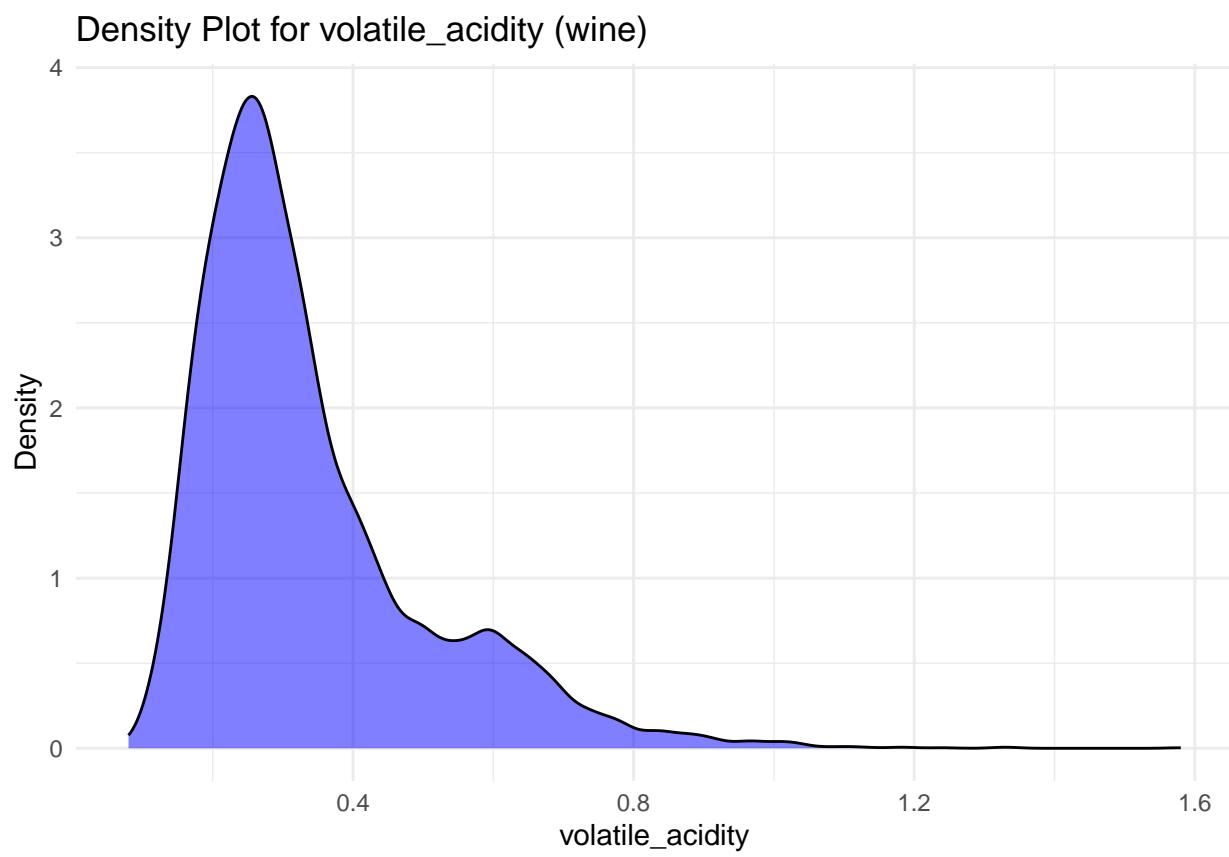
  # Return a list with both plots
  list(p1, p2)
})
```

```
# Plotting side by side
for (i in seq_along(density_plots)) {
  print(density_plots[[i]][[1]]) # Print wine_quality plot
  print(density_plots[[i]][[2]]) # Print cleaned_data plot
}
```

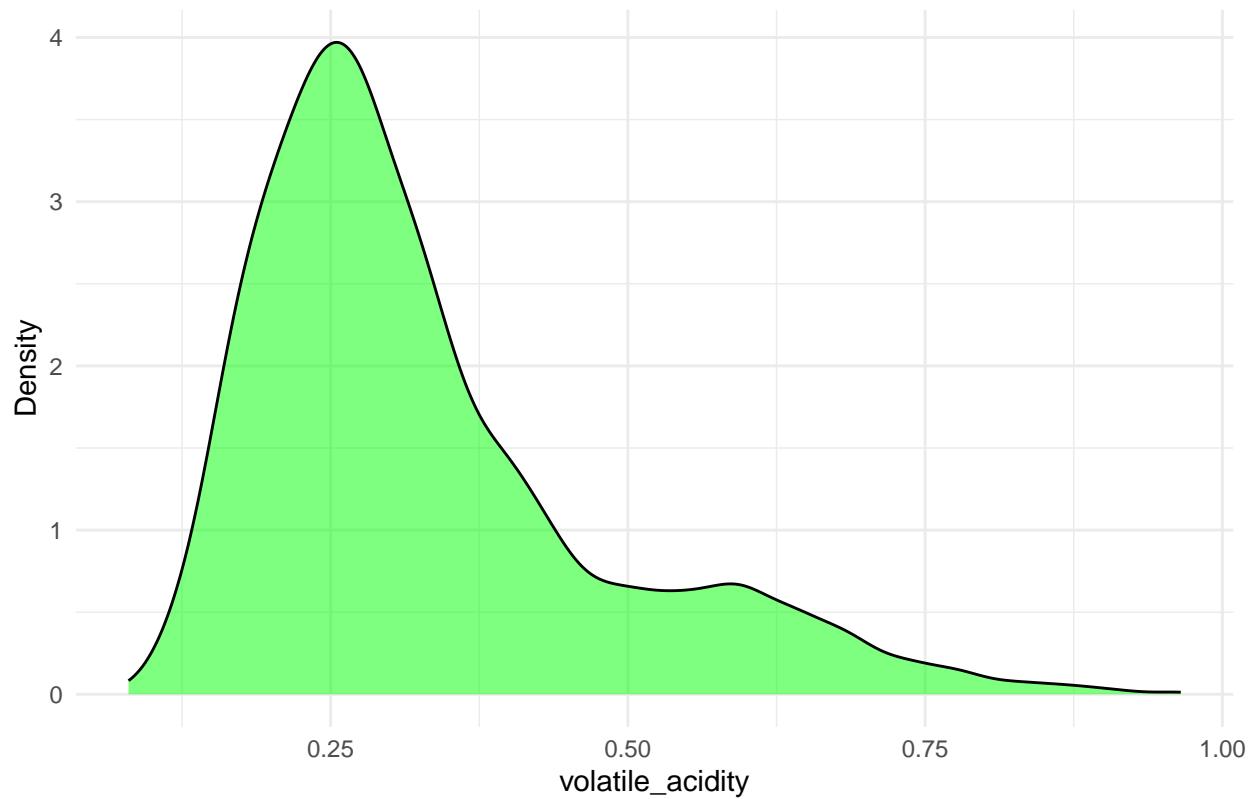


Density Plot for fixed\_acidity (cleaned\_data)

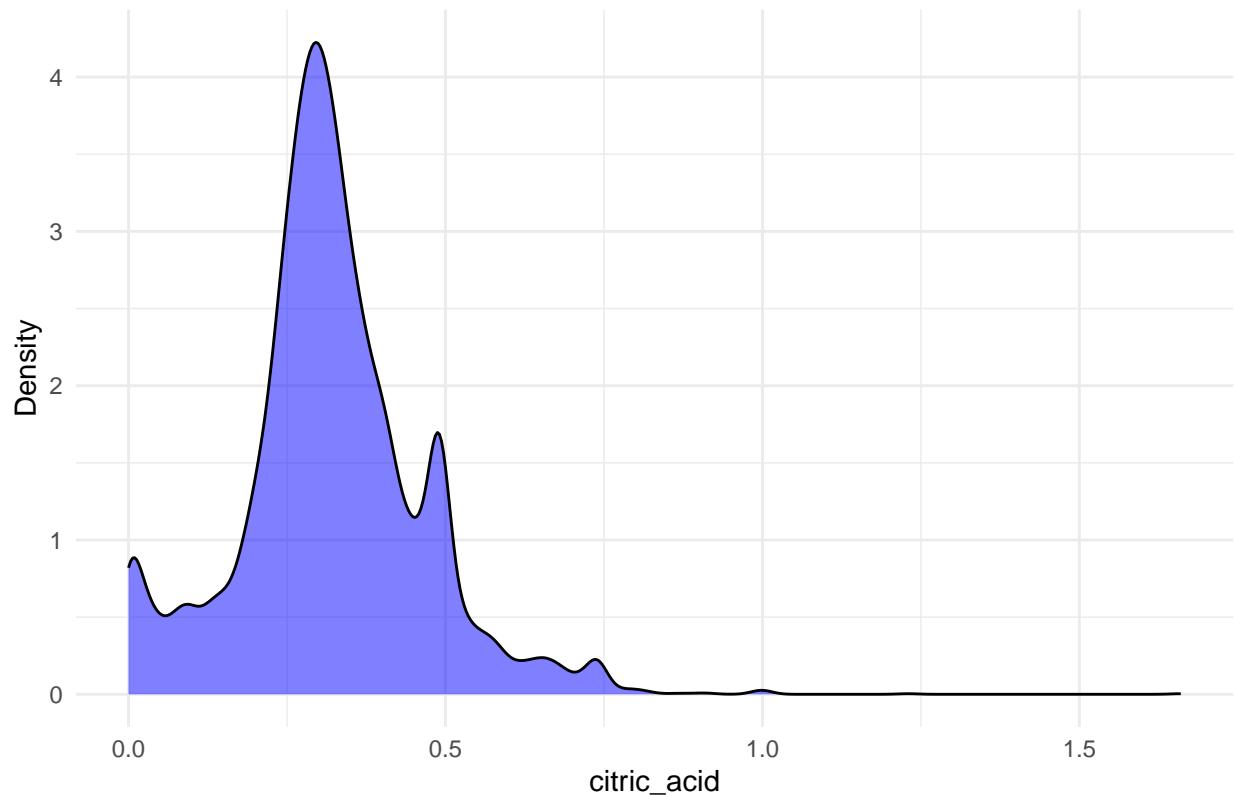




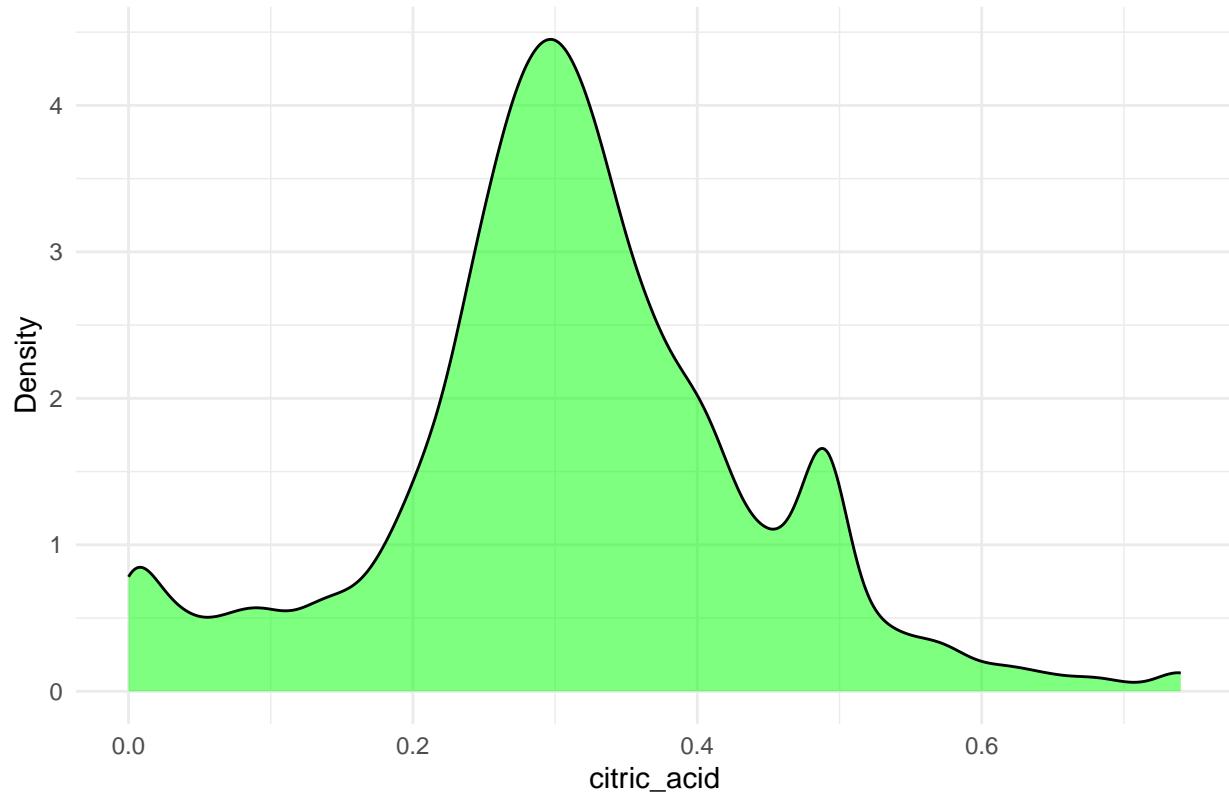
Density Plot for volatile\_acidity (cleaned\_data)



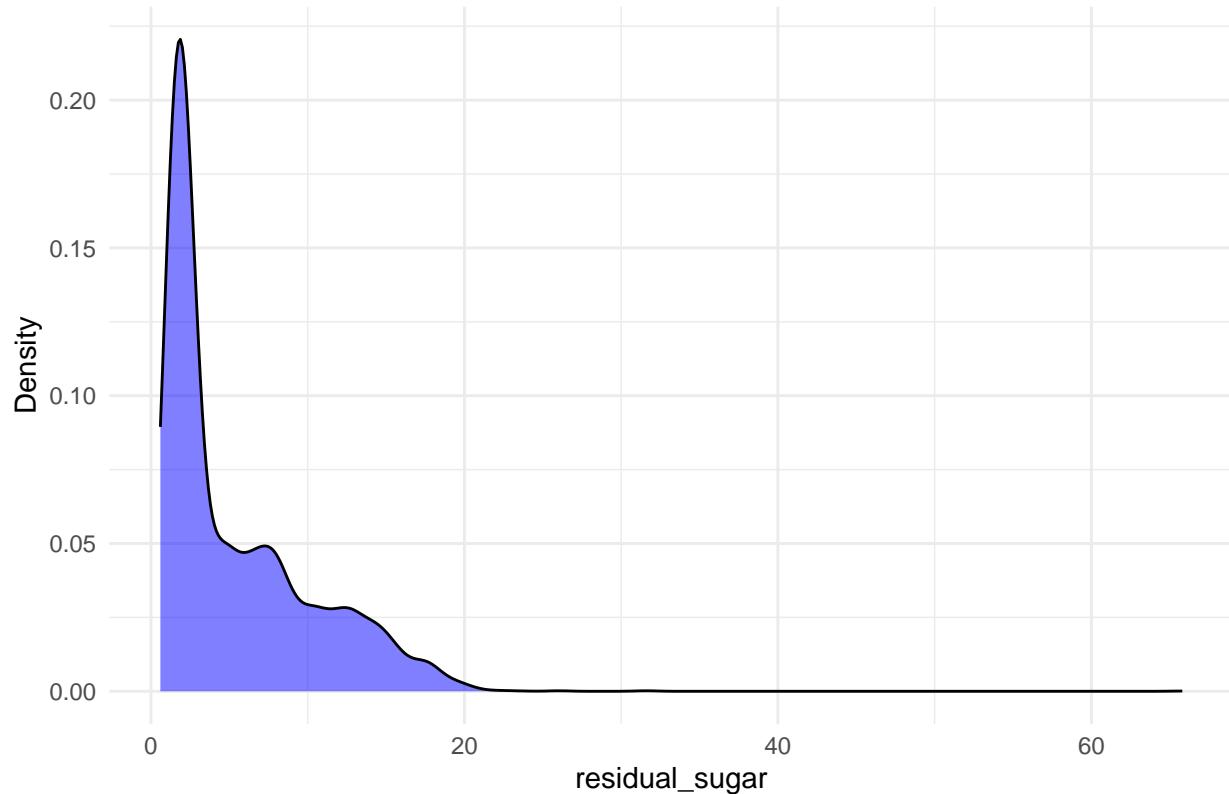
Density Plot for citric\_acid (wine)



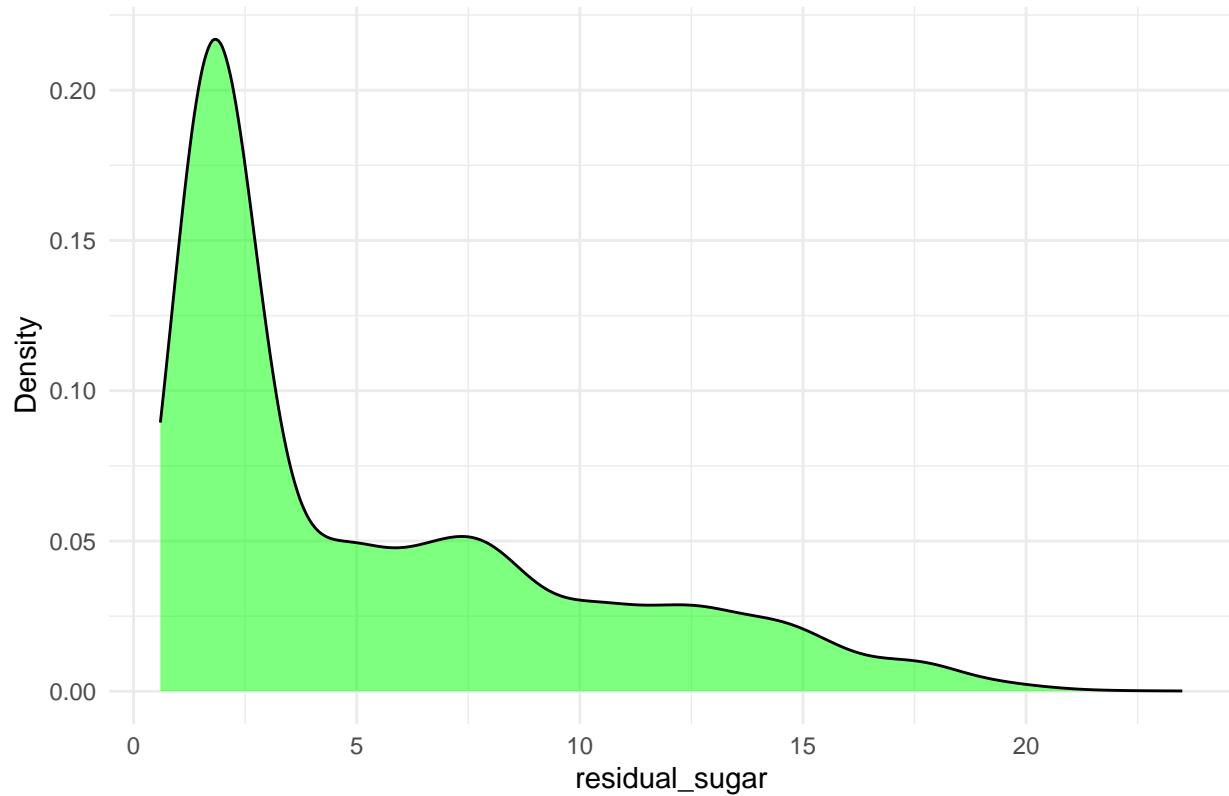
Density Plot for citric\_acid (cleaned\_data)



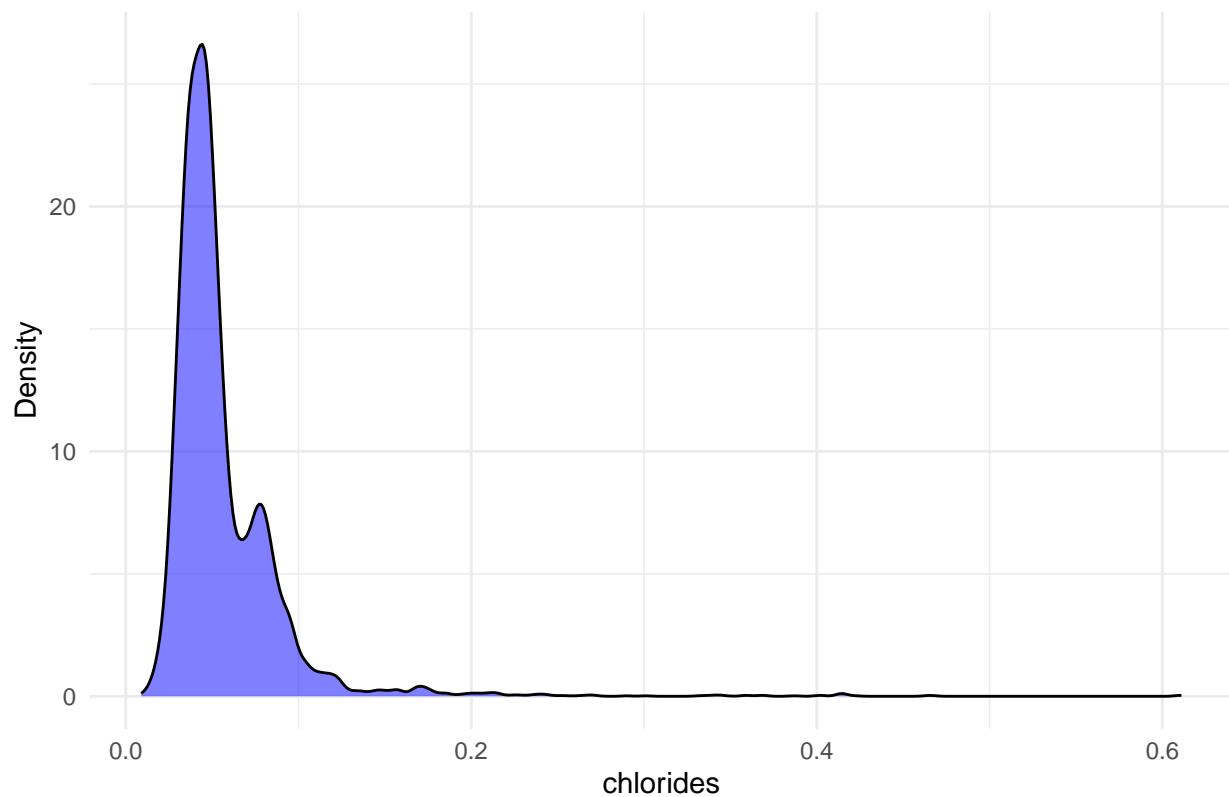
Density Plot for residual\_sugar (wine)



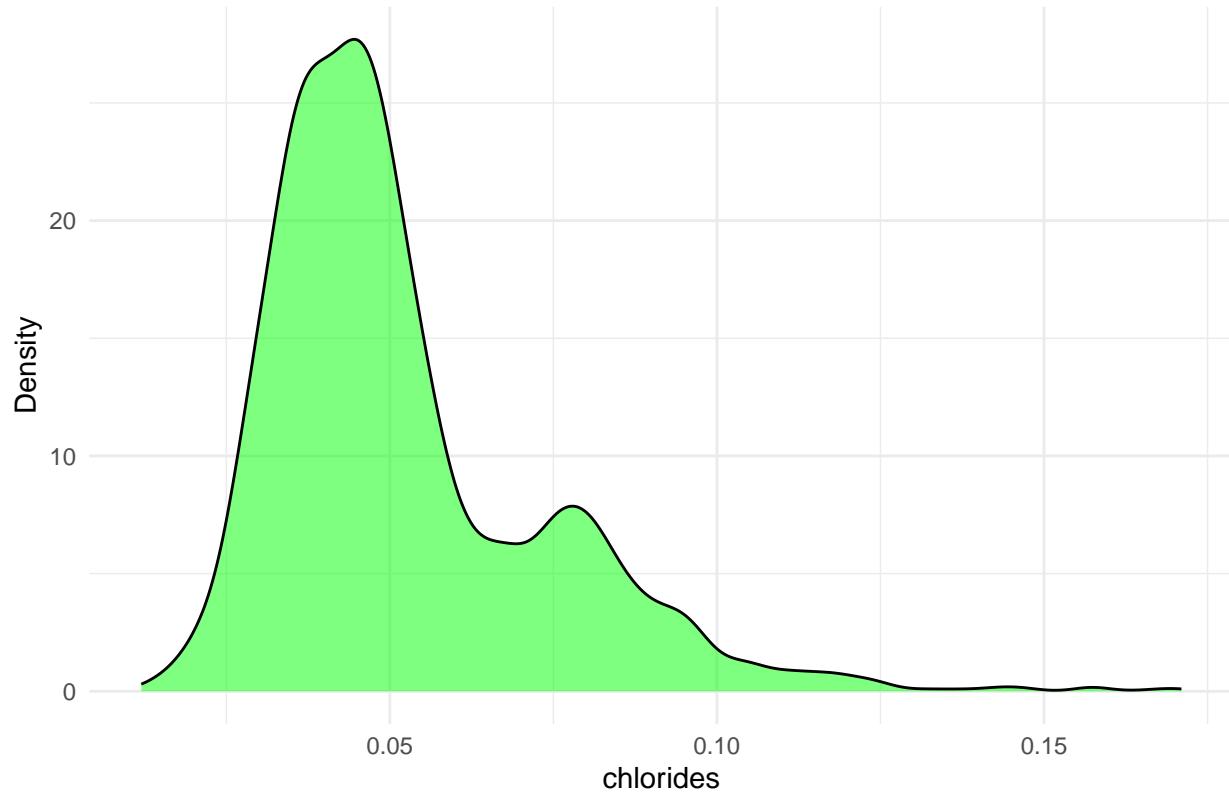
Density Plot for residual\_sugar (cleaned\_data)



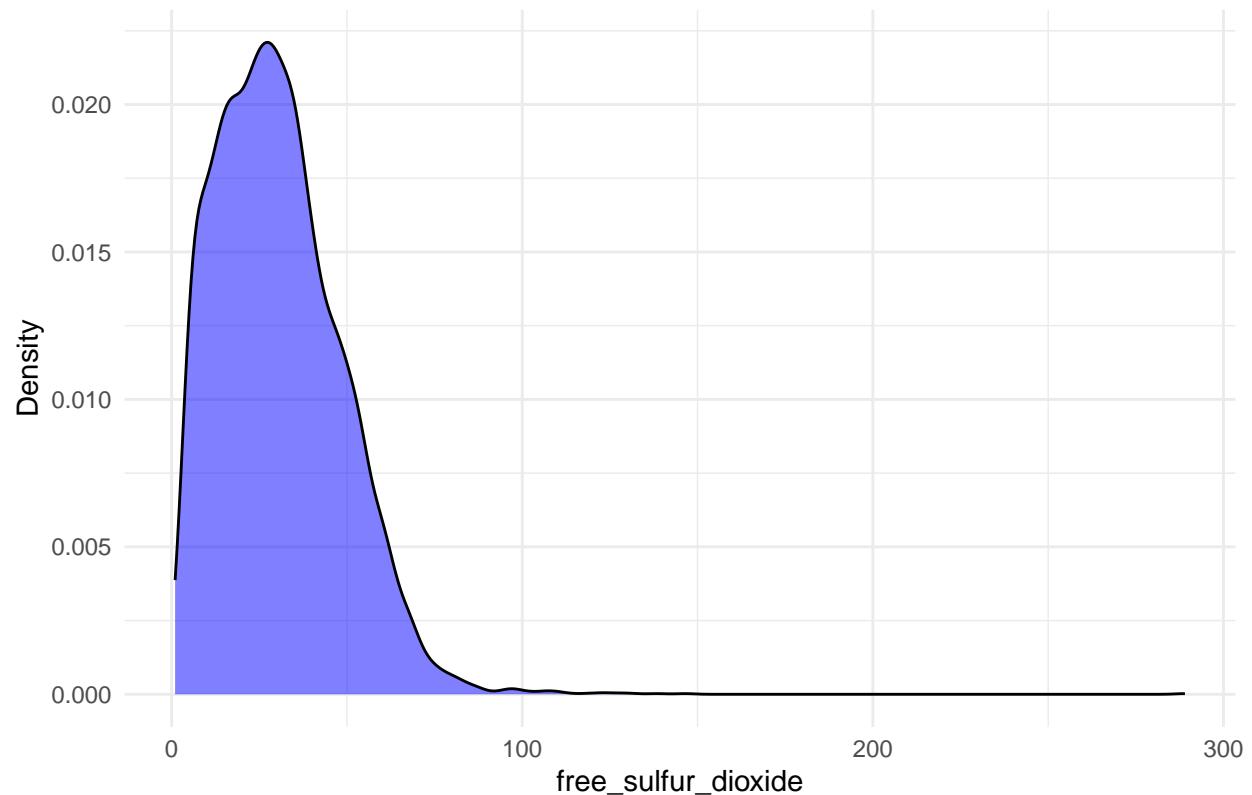
Density Plot for chlorides (wine)



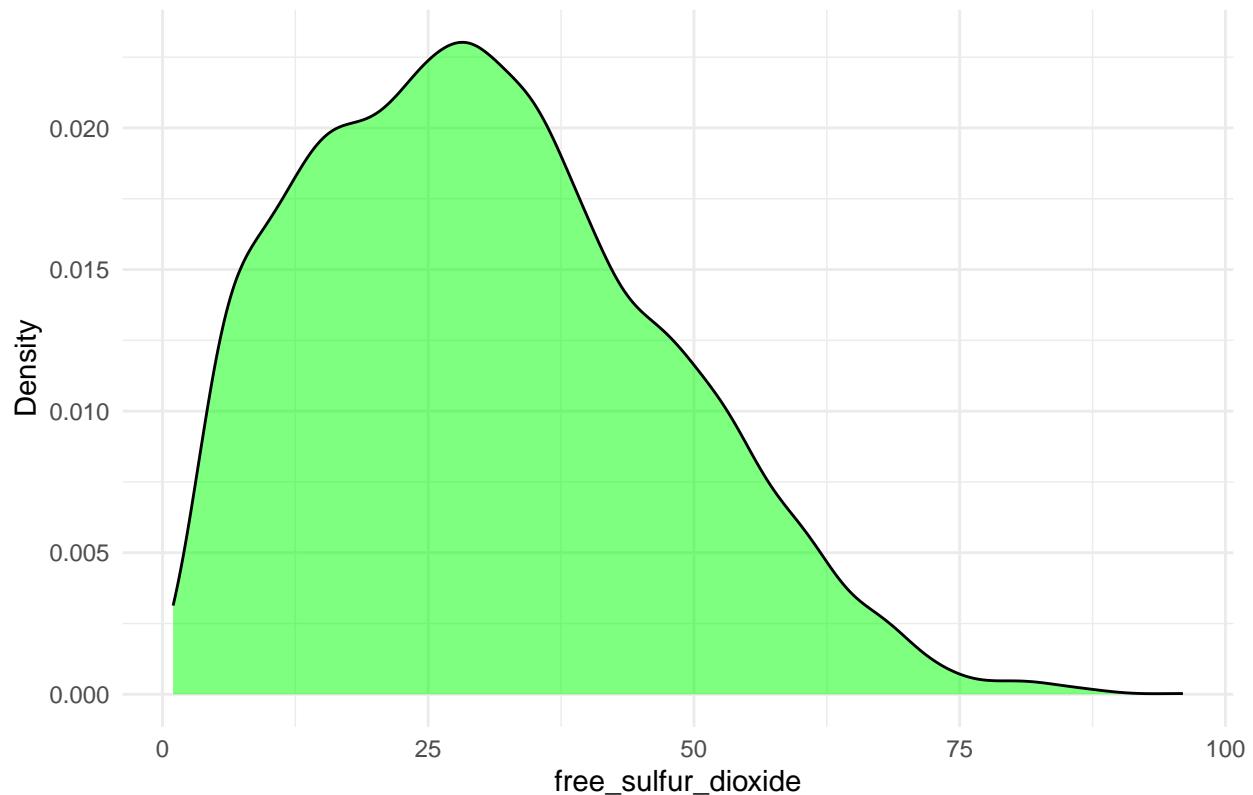
Density Plot for chlorides (cleaned\_data)



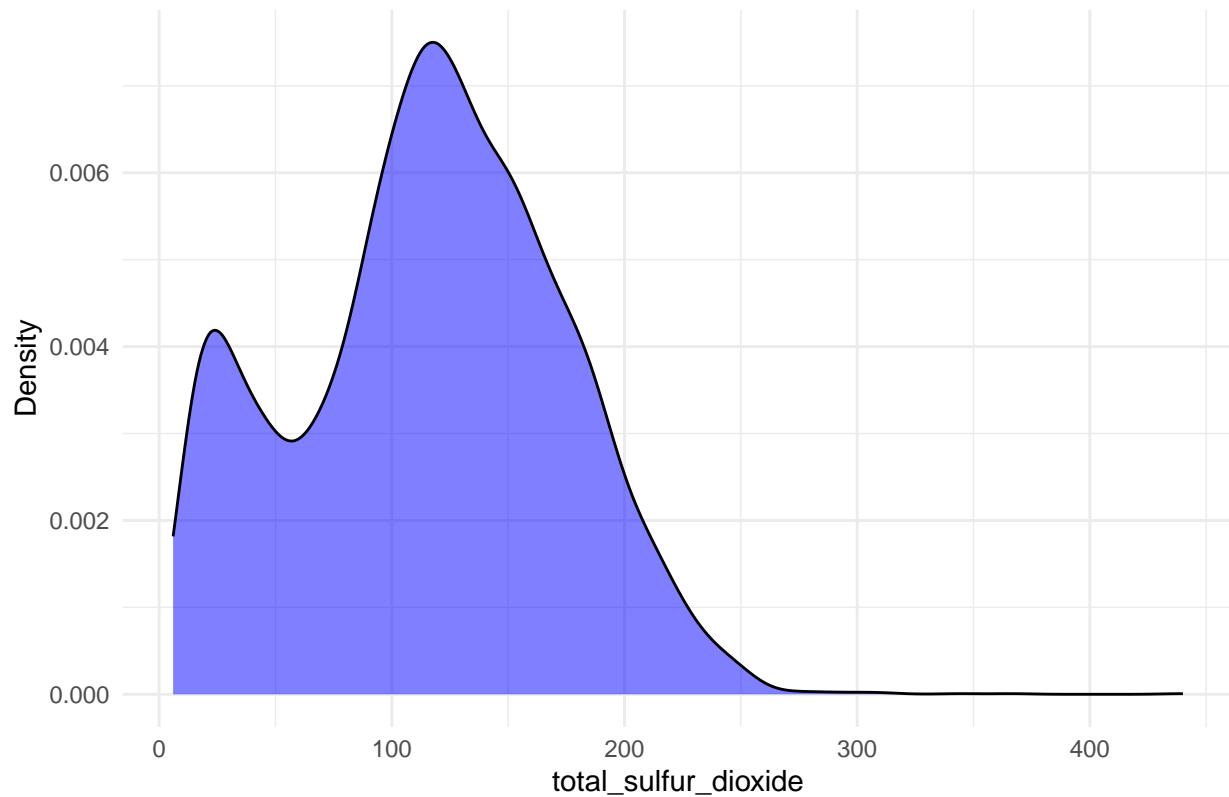
Density Plot for free\_sulfur\_dioxide (wine)



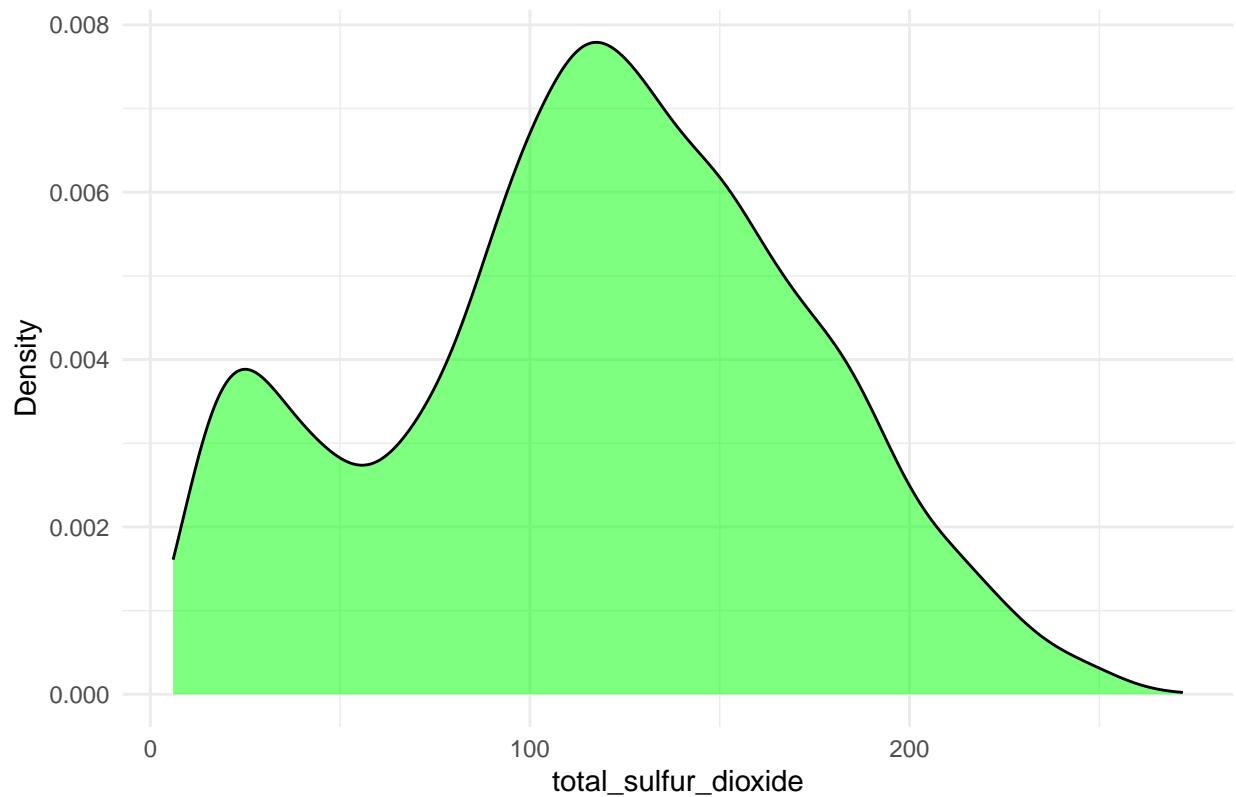
Density Plot for free\_sulfur\_dioxide (cleaned\_data)



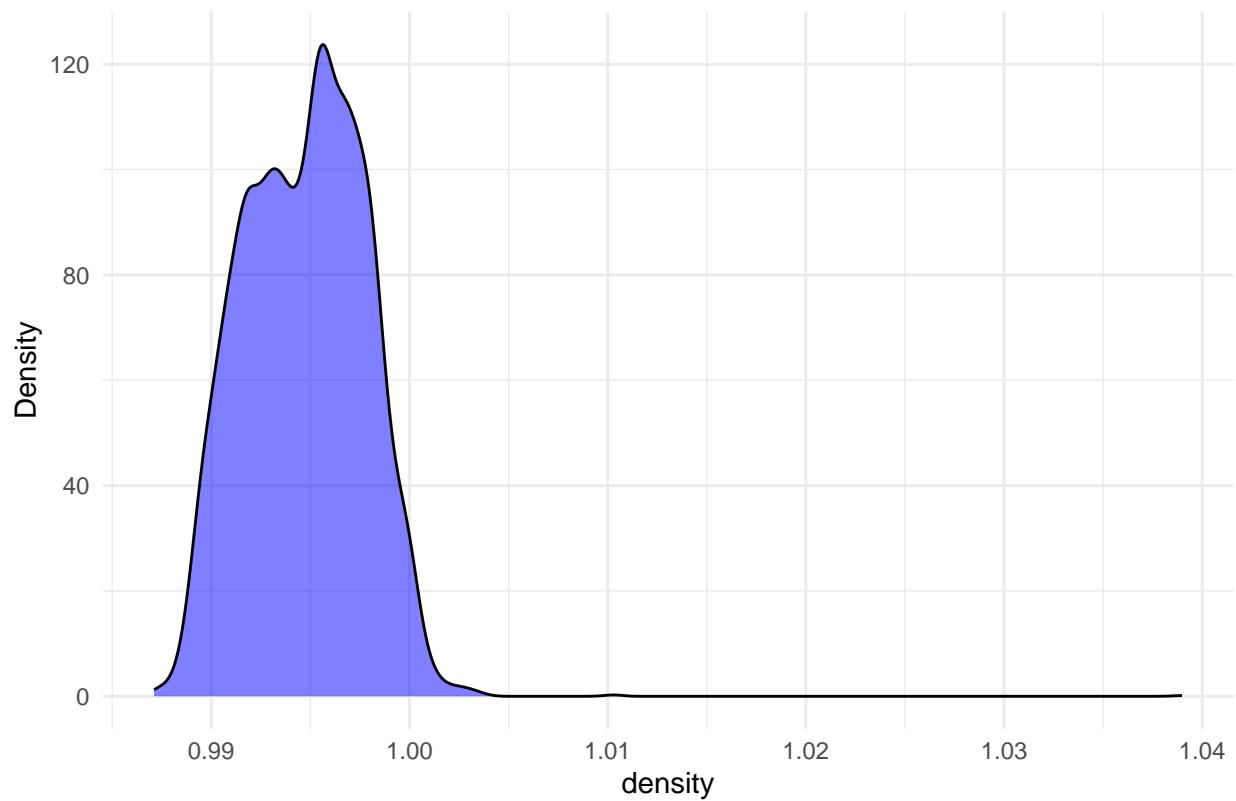
Density Plot for total\_sulfur\_dioxide (wine)



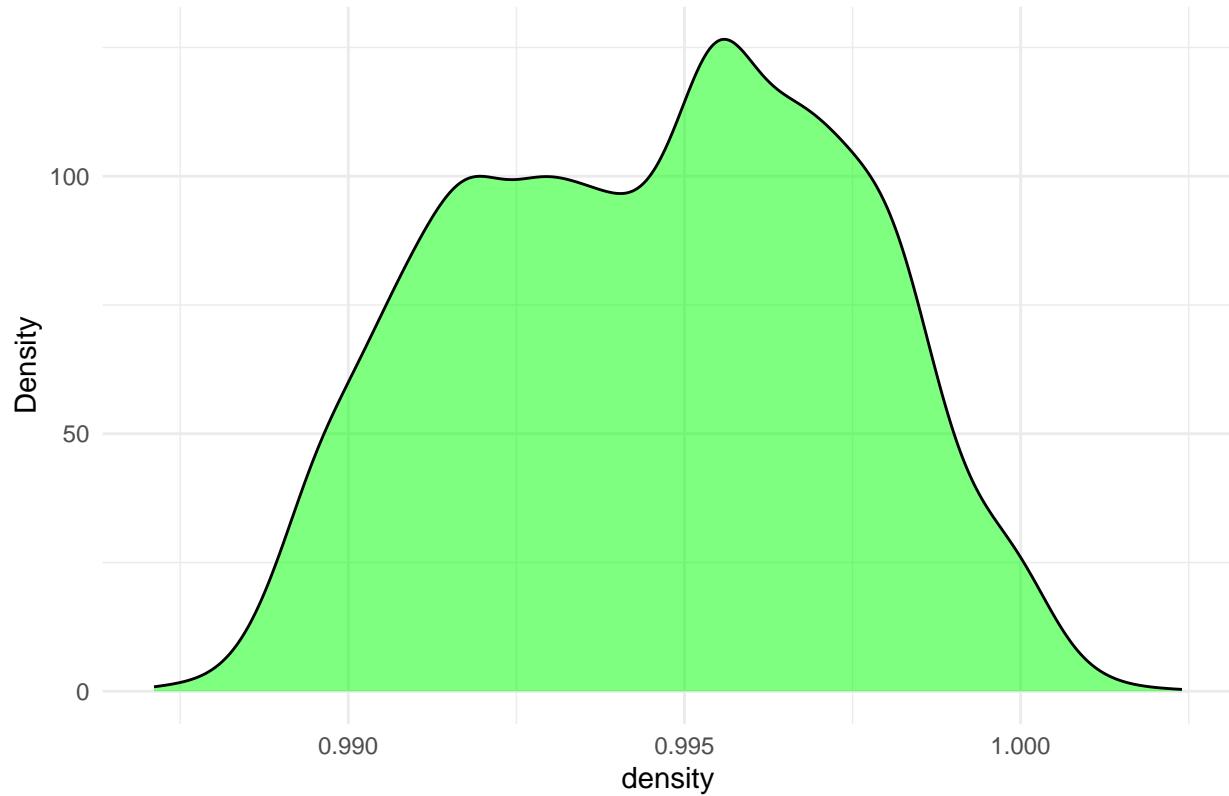
Density Plot for total\_sulfur\_dioxide (cleaned\_data)



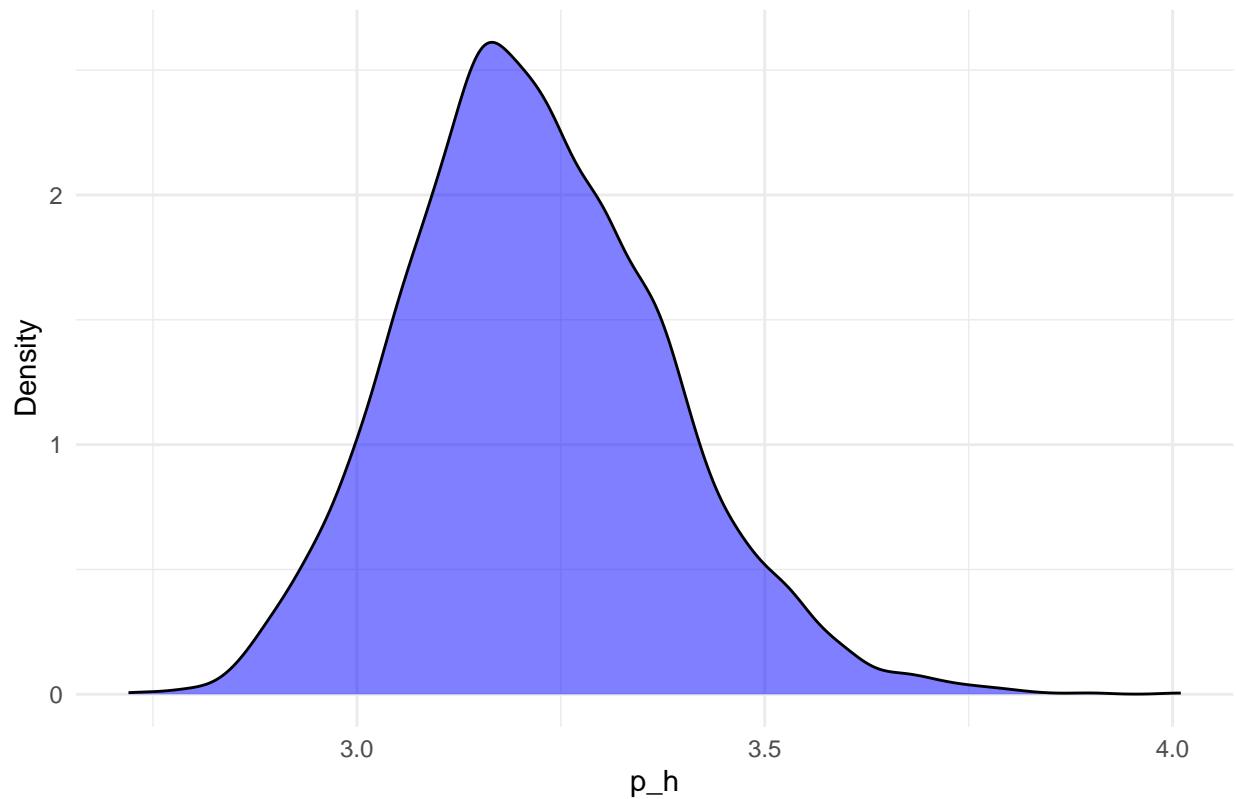
Density Plot for density (wine)



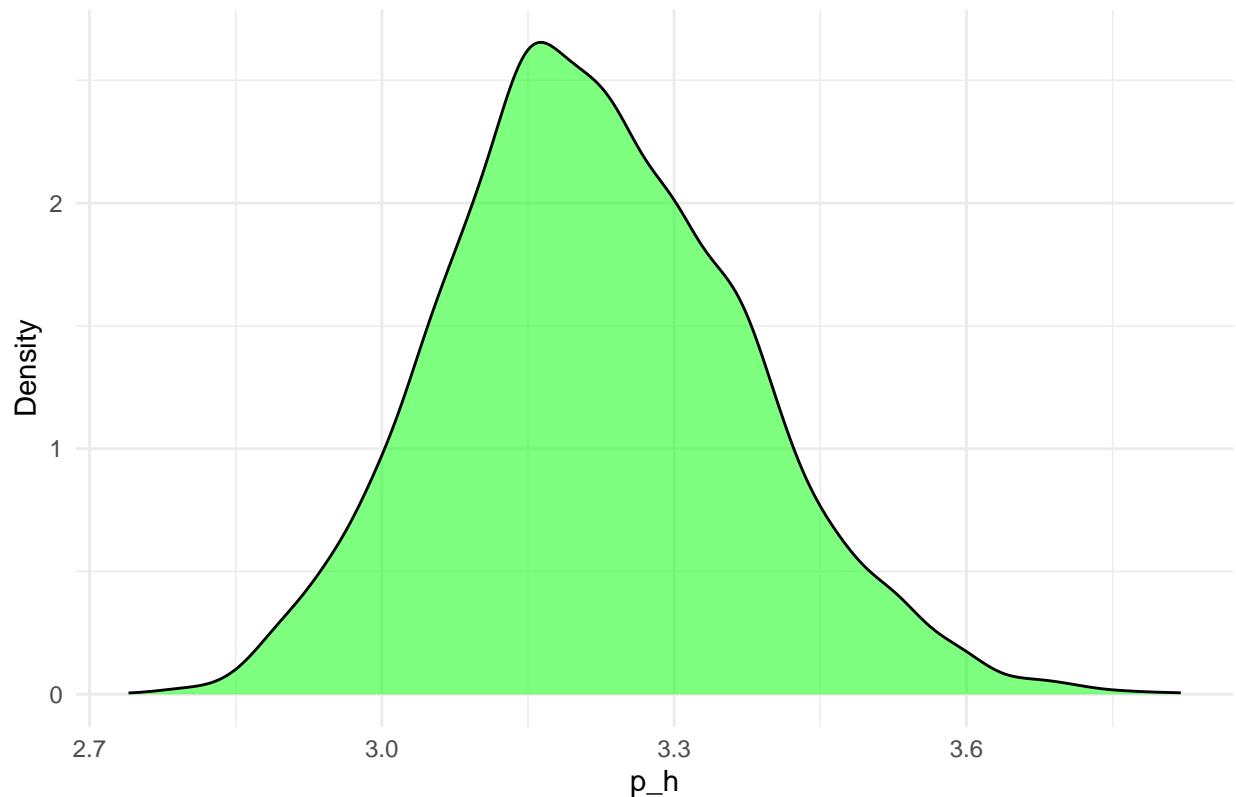
Density Plot for density (cleaned\_data)



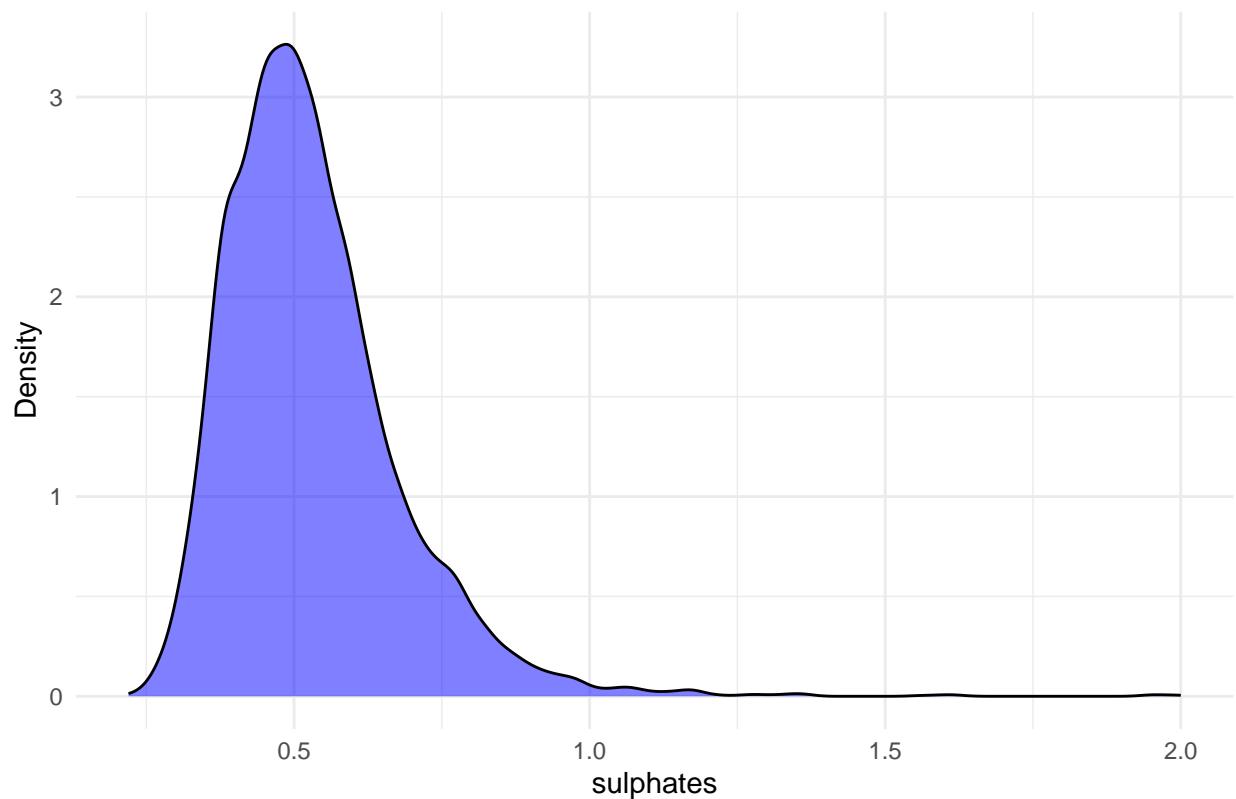
Density Plot for  $p_h$  (wine)



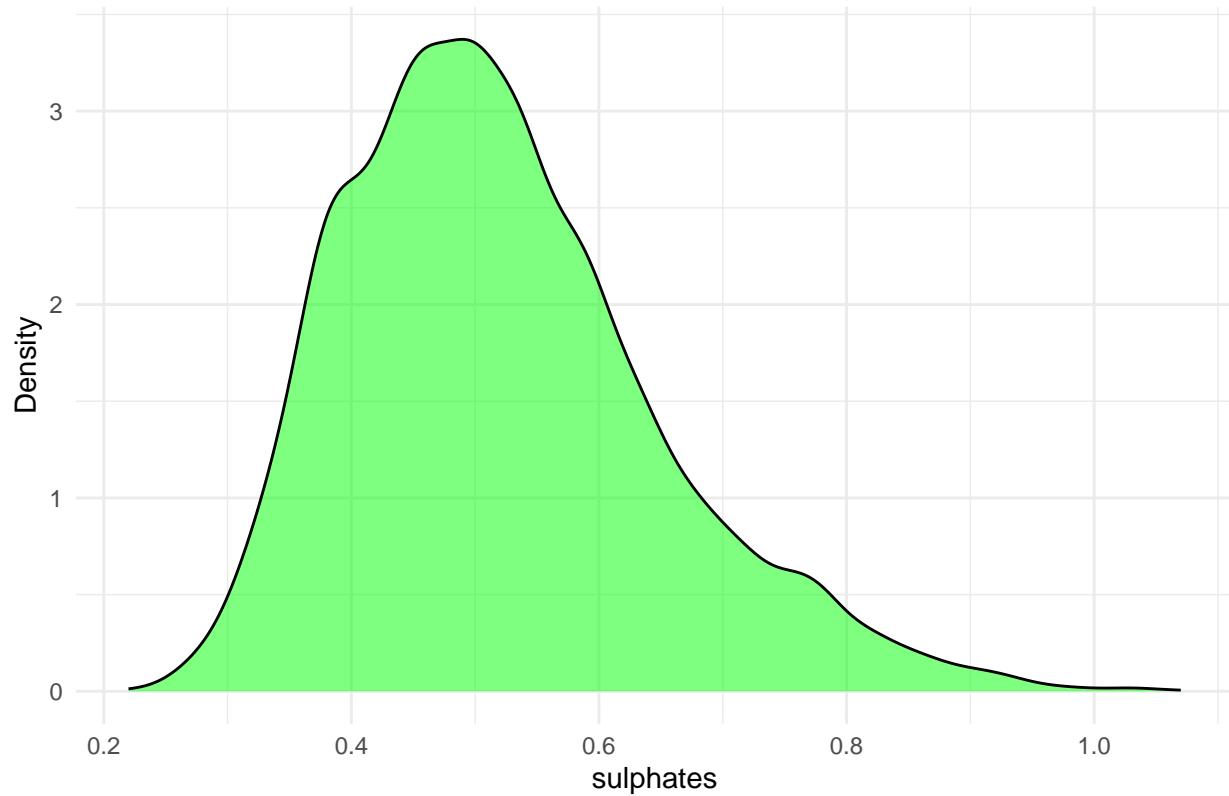
Density Plot for p\_h (cleaned\_data)



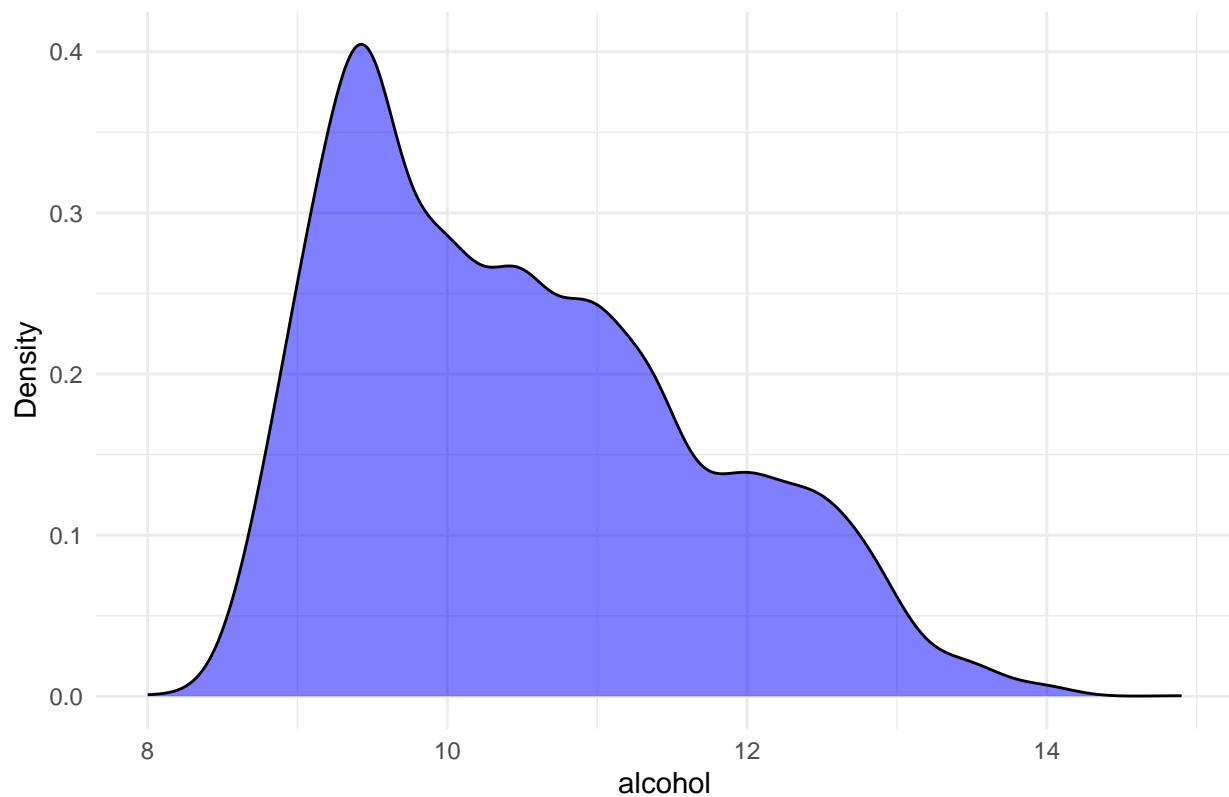
Density Plot for sulphates (wine)



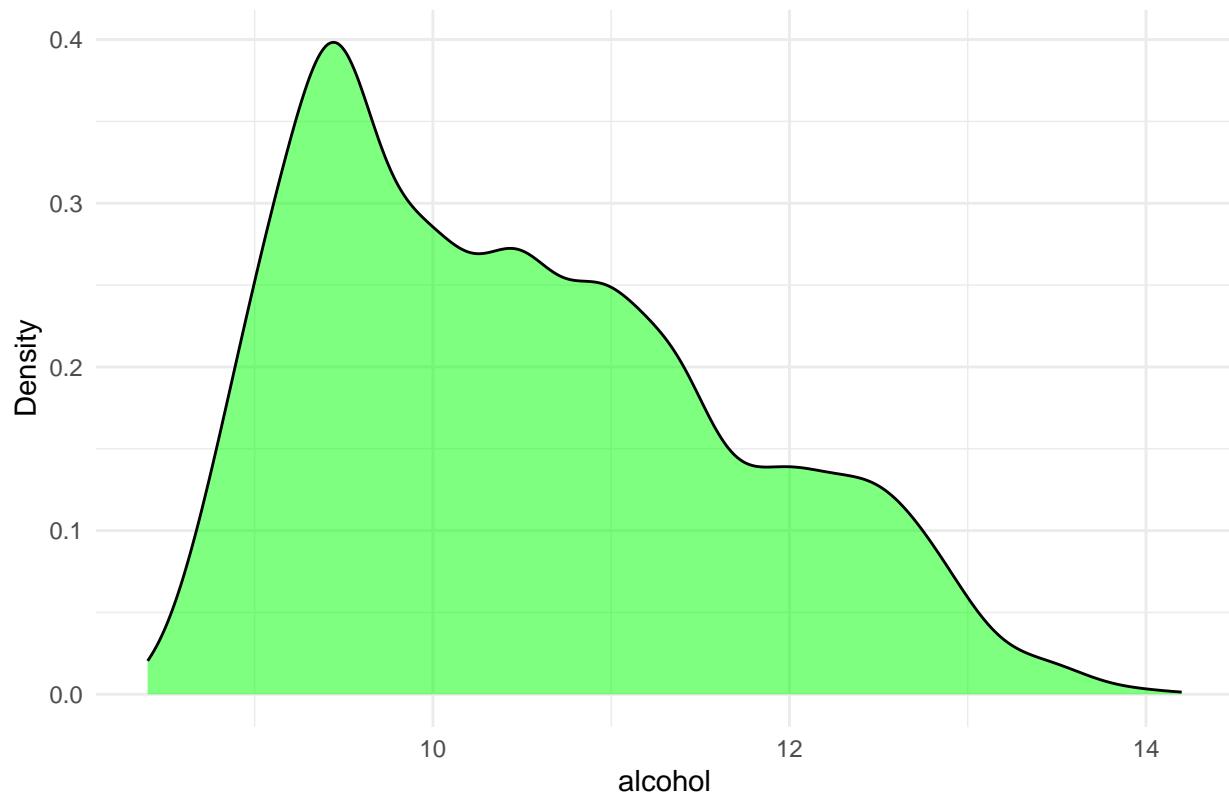
Density Plot for sulphates (cleaned\_data)



Density Plot for alcohol (wine)



Density Plot for alcohol (cleaned\_data)



**Nhận xét:** Hình dạng phân phối của các biến trước và sau khi làm sạch đều tương đồng, sự khác biệt ở đây có lẽ là đuôi của phân phối sau khi làm sạch ngắn lại so với trước do loại bỏ giá trị ngoại lai.

## Phân tích thành phần chính

Ma trận tương quan của dữ liệu

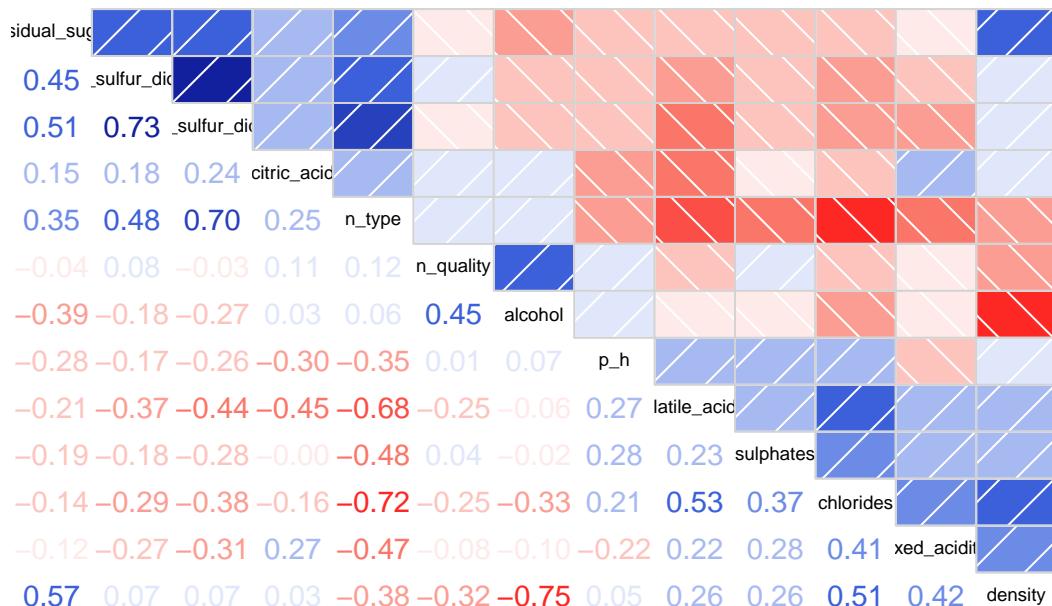
```
# Select numeric columns

corr_data <- cleaned_data |>
  mutate(n_type = as.numeric(type),
        n_quality = as.numeric(quality))

# Select only numeric columns
numeric_corr <- corr_data |>
  select_if(is.numeric)

#corrplot
corrgram(numeric_corr, type = "data",
          lower.panel = panel.cor,
          upper.panel = panel.shade,
          main = "Corrgram for wine quality dataset",
          order = TRUE, cex.labels = 0.9)
```

## Corrgram for wine quality dataset



**Nhận xét:** Theo ma trận tương quan mẫu, biến loại rượu vang gần như không có sự tương quan với biến alcohol. Ngược lại, biến chất lượng rượu vang chỉ tương quan thuận với alcohol và tương quan nghịch với density. Điều này sẽ được dùng làm dấu hiệu nhận biết thành phần chính để dự phân loại rượu vang và chất lượng của mẫu rượu.

Kiểm tra trung bình và độ lệch chuẩn

```
# Perform PCA on the selected numeric columns
pca_data <- cleaned_data[,numeric_vars]
colMeans(pca_data)

##          fixed_acidity      volatile_acidity       citric_acid
## 7.13472106          0.32846122          0.31007371
##      residual_sugar      chlorides free_sulfur_dioxide
## 5.47796113          0.05188524         30.47914223
##      total_sulfur_dioxide      density            p_h
## 116.74945552          0.99455560          3.21852069
##      sulphates      alcohol
## 0.51945887          10.49166192

apply(pca_data,2,sd)

##          fixed_acidity      volatile_acidity       citric_acid
## 1.109341169          0.149494087          0.129937064
##      residual_sugar      chlorides free_sulfur_dioxide
## 4.659685499          0.021050608          16.342732176
##      total_sulfur_dioxide      density            p_h
## 54.707032702          0.002872084          0.154324306
##      sulphates      alcohol
```

```
##          0.126126793          1.173282228
```

**Nhận xét:** Nhận thấy có sự chênh lệch lớn giữa các trung bình và độ lệch chuẩn, cần phải chuẩn hóa dữ liệu. Thực hiện phân tích thành phần chính với dữ liệu chuẩn hóa (tham số scale. = TRUE)

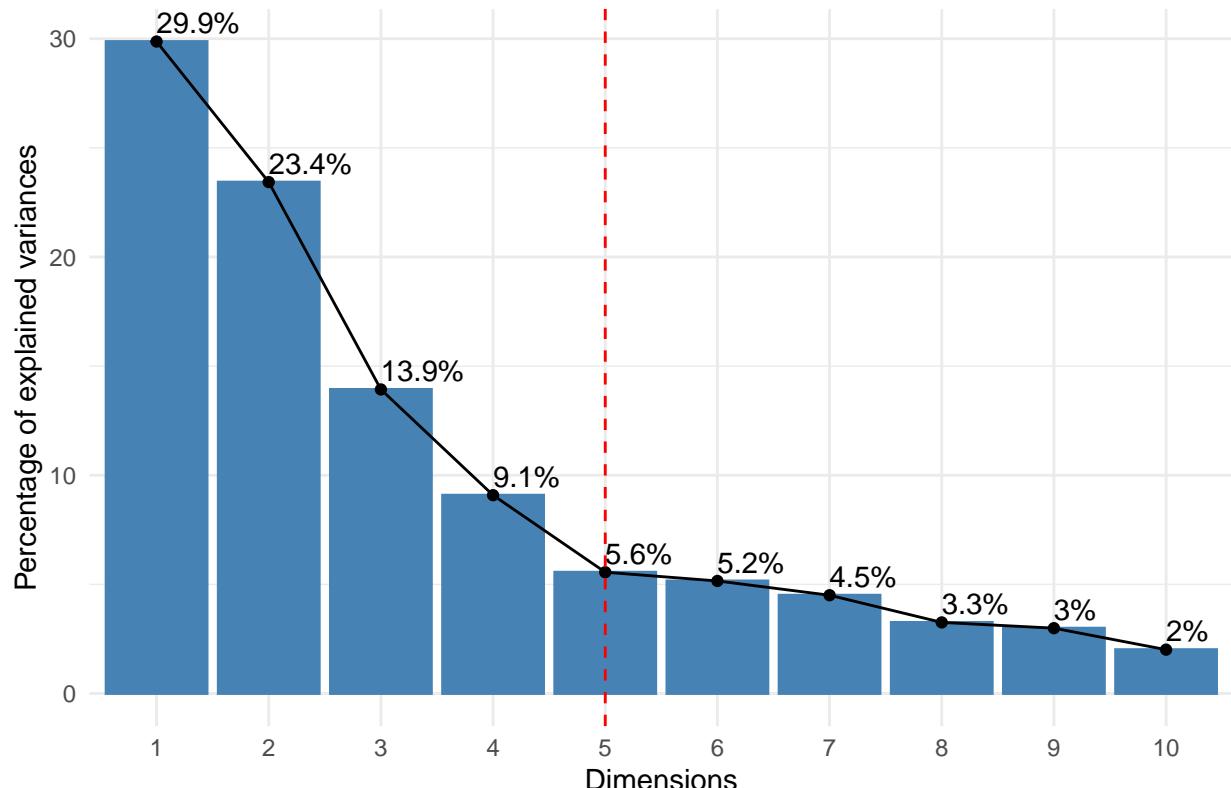
```
pca <- prcomp(pca_data, scale. = TRUE)  
summary(pca)
```

```
## Importance of components:  
##           PC1    PC2    PC3    PC4    PC5    PC6    PC7  
## Standard deviation   1.8126 1.6054 1.2379 0.99974 0.78182 0.75313 0.70379  
## Proportion of Variance 0.2987 0.2343 0.1393 0.09086 0.05557 0.05156 0.04503  
## Cumulative Proportion 0.2987 0.5330 0.6723 0.76316 0.81873 0.87029 0.91532  
##           PC8    PC9    PC10   PC11  
## Standard deviation   0.59857 0.57386 0.47013 0.15120  
## Proportion of Variance 0.03257 0.02994 0.02009 0.00208  
## Cumulative Proportion 0.94789 0.97783 0.99792 1.00000
```

Trực quan hóa bằng biểu đồ

```
fviz_eig(pca, addlabels = TRUE) +  
  geom_vline(xintercept = 5, linetype = "dashed", color = "red") +  
  labs(title = "Scree Plot of PCA Eigenvalues") +  
  theme_minimal()
```

Scree Plot of PCA Eigenvalues



**Nhận xét:** Dựa theo quy tắc khuỷu tay, độ dốc thay đổi lớn nhất tại thành phần chính thứ năm nên ta có thể chọn bốn thành phần chính đầu tiên giải thích được khoảng 76% phương sai của dữ liệu.

Quan sát bốn thành phần chính đầu tiên

```
pca$rotation[, 1:4]
```

```
##          PC1        PC2        PC3        PC4
## fixed_acidity  0.23091323  0.24410394 -0.55847923  0.0812407520
## volatile_acidity  0.39764733  0.10284429  0.22525514  0.2534339268
## citric_acid    -0.20824878  0.07993712 -0.58709639 -0.3140175885
## residual_sugar -0.30212641  0.38128481  0.11833546  0.1082436246
## chlorides       0.37682450  0.30614368  0.01001342  0.0005078118
## free_sulfur_dioxide -0.40063259  0.15246048  0.16581006 -0.2606340367
## total_sulfur_dioxide -0.45110089  0.16293274  0.13507437 -0.1405220535
## density         0.09371276  0.59335980  0.04679913 -0.0190413789
## p_h              0.24115334 -0.08547208  0.43802645 -0.5228874951
## sulphates       0.28241062  0.12945150 -0.09765420 -0.6680721801
## alcohol          0.03492564 -0.50834440 -0.17100766 -0.1052440299
```

#### Nhận xét:

- Thành phần chính thứ nhất có hệ số của biến alcohol gần như là 0 nên thành phần chính này sẽ là thành phần phân loại rượu vang.
- Thành phần chính thứ hai có hệ số của biến alcohol và density lớn nhất nên sẽ là thành phần phân loại chất lượng rượu vang, có một điểm khác so với ma trận tương quan mẫu là thành phần này tương quan nghịch với alcohol và tương quan thuận với density nên phương trình này càng bé thì khả năng mẫu rượu chất lượng càng cao.
- Trọng số cao của citric\_acid và mối quan hệ với residual\_sugar, total\_sulfur\_dioxide và free\_sulfur\_dioxide cho thấy rằng thành phần chính thứ ba phản ánh các yếu tố hóa học liên quan đến hương vị và sự bảo quản của rượu nên có thể gọi đây là thành phần hương vị chua của rượu.
- Trọng số cao của density và mối quan hệ với sulphates, alcohol cho thấy rằng thành phần chính thứ tư phản ánh các yếu tố liên quan đến sự ổn định của rượu và các yếu tố ảnh hưởng đến cảm nhận của rượu nên có thể coi thành phần cấu trúc và đặc trưng của rượu.

Tạo thêm biến phân loại chất lượng rượu vang để trực quan hóa

```
cleaned_data <- cleaned_data |>
  mutate(quality_class = case_when(
    type == "red" & quality == 3 ~ "poor_red",
    type == "white" & quality == 3 ~ "poor_white",
    quality %in% 4:8 ~ "normal",
    type == "white" & quality == 9 ~ "excellent_white",
  ))

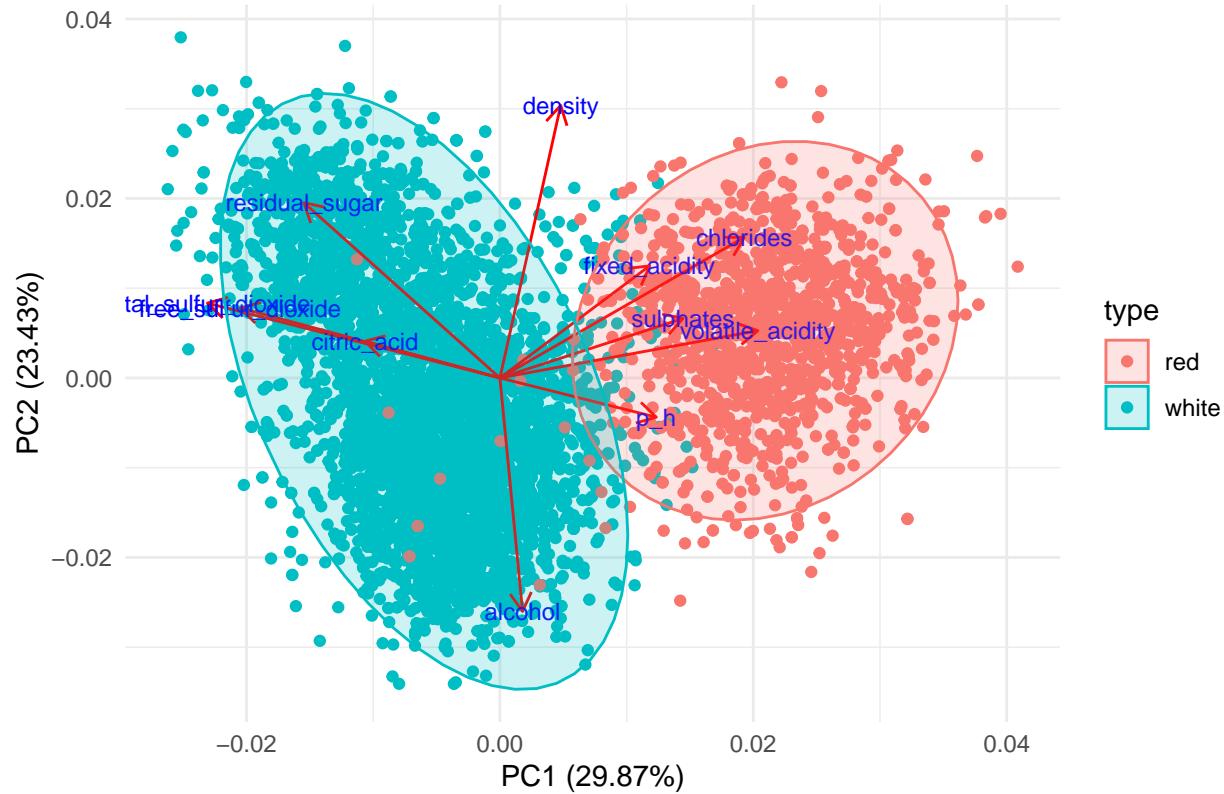
# Chuyển quality_class thành factor
cleaned_data$quality_class <- factor(cleaned_data$quality_class,
                                         levels = c("poor_red", "poor_white", "normal", "excellent_white"))
```

Nhận xét: Cách đặt này dựa theo biểu đồ phần trăm hai loại rượu vang theo điểm chất lượng.

Quan sát đồ thị autoplot của hai thành phần chính đầu tiên

```
autoplot(pca, data = cleaned_data, colour = "type",
         loadings = TRUE, loadings.label = TRUE,
         x = 1, y = 2,
         loadings.label.size = 3, loadings.label.color = "blue",
         frame = TRUE, frame.type = "norm") +
  ggtitle("PCA Plot Colored by Type") +
  theme_minimal()
```

### PCA Plot Colored by Type



```
pca$rotation[, "PC1"]
```

```
##          fixed_acidity      volatile_acidity      citric_acid
## 0.23091323        0.39764733       -0.20824878
## residual_sugar      chlorides    free_sulfur_dioxide
## -0.30212641        0.37682450       -0.40063259
## total_sulfur_dioxide density           p_h
## -0.45110089        0.09371276        0.24115334
## sulphates          alcohol
## 0.28241062        0.03492564
```

**Nhận xét:** Thành phần chính thứ nhất phân hóa được hai loại rượu vang. Trong đó, biến alcohol và density có ảnh hưởng gần như là không nên sẽ loại bỏ 2 biến này khi phân loại rượu, khi các biến residual\_sugar, free\_sulfur\_dioxide, total\_sulfur\_dioxide, citric\_acid càng tăng tức là phương trình càng giảm thì khả năng cao sẽ là rượu trắng và ngược lại mẫu rượu sẽ là rượu đỏ khi các biến fixed\_acidity, chlorides, sulphates, volatile\_acidity, p\_h càng tăng.

Quan sát biều đồ thành phần chính thứ hai theo thành phần chính thứ nhất

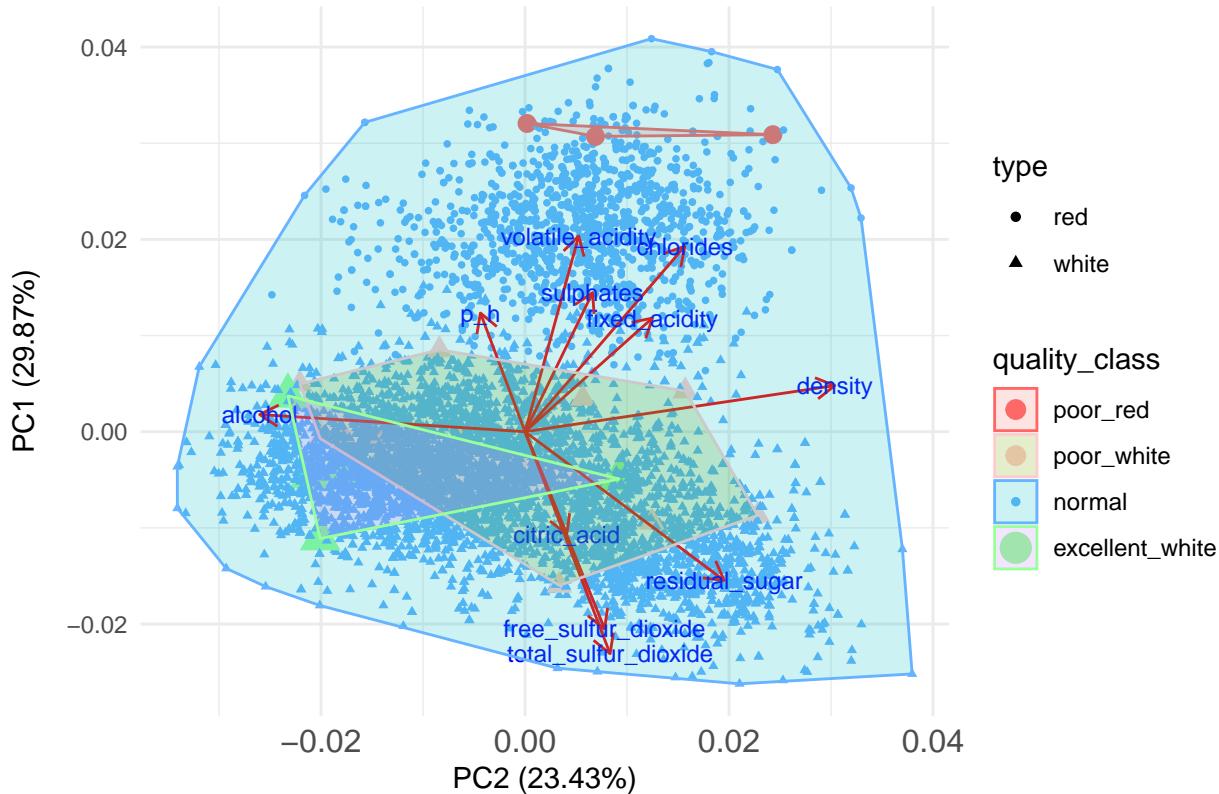
```
# Áp dụng autoplot cho PCA với các thông số đã chỉnh sửa
autoplot(pca, data = cleaned_data, colour = 'quality_class', shape = 'type', size = 'quality_class',
         x = 2, y = 1,
         loadings = TRUE, loadings.label = TRUE,
         loadings.label.size = 3, loadings.label.colour = "blue",
         frame = TRUE, frame.type = 'convex') +
  ggtitle("Combined PCA Plot Colored by Quality Class, Shaped by Type, and Sized by Quality Class") +
  theme_minimal() +
  scale_color_manual(values = c("poor_red" = "#ff6666", "poor_white" = "#ffcccc", "normal" = "#66b3ff"),
```

```

scale_shape_manual(values = c("red" = 16, "white" = 17)) +
scale_size_manual(values = c("poor_red" = 3, "poor_white" = 3, "normal" = 1, "excellent_white" = 5)) +
theme(
  axis.title.y = element_text(margin = margin(r = 10)),
  axis.text.x = element_text(size = 12),
  legend.position = "right",
  plot.title = element_text(hjust = 0.5)
)

```

I PCA Plot Colored by Quality Class, Shaped by Type, and Sized by Quality Class



```
pca$rotation[, "PC2"]
```

##	fixed_acidity	volatile_acidity	citric_acid
##	0.24410394	0.10284429	0.07993712
##	residual_sugar	chlorides	free_sulfur_dioxide
##	0.38128481	0.30614368	0.15246048
##	total_sulfur_dioxide	density	p_h
##	0.16293274	0.59335980	-0.08547208
##	sulphates	alcohol	
##	0.12945150	-0.50834440	

**Nhận xét:** Do bộ dữ liệu chứa quá nhiều quan trắc có điểm chất lượng từ 4 đến 8 nên ta sẽ bỏ qua và tập trung vào phân loại chất lượng rượu vang tè và xuất sắc. Ta không thể phân loại rượu vang đỏ tè và xuất sắc do không có rượu vang đỏ loại xuất sắc được ghi nhận. Vì thế do đặc trưng của bộ dữ liệu này, ta sẽ tập trung phân loại rượu vang trắng tè và xuất sắc. Trong đó, các biến chlorides, fixed\_acidity, density, residual\_sugar càng tăng tức là kết quả của phương trình thì khả năng rượu vang trắng dobr tè càng cao và ngược lại khi biến alcohol càng tăng tức là kết quả phương trình càng thấp thì khả năng rượu vang trắng thuộc loại xuất sắc. Các biến còn lại gần như ảnh hưởng rất thấp nên dường như không có nghĩa đối với sự phân loại.

## Phân tích nhân tố

```
factor_data <- cleaned_data[, -c(12:17)]  
  
factor_res <- fa(r = factor_data, covar = TRUE, nfactors = 1)  
summary(factor_res)  
  
##  
## Factor analysis with Call: fa(r = factor_data, nfactors = 1, covar = TRUE)  
##  
## Test of the hypothesis that 1 factor is sufficient.  
## The degrees of freedom for the model is 44 and the objective function was 4.74  
## The number of observations was 5969 with Chi Square = 28239.65 with prob < 0  
##  
## The root mean square of the residuals (RMSA) is 0.23  
## The df corrected root mean square of the residuals is 0.26  
##  
## Tucker Lewis Index of factoring reliability = 0.998  
## RMSEA index = 0.328 and the 10 % confidence intervals are 0.324 0.331  
## BIC = 27857.1
```

**Nhận xét:** Kết quả phân tích nhân tố cho thấy rằng một yếu tố MR1 có thể giải thích 82% phương sai tổng thể của dữ liệu. Kiểm định giả thuyết cho thấy mô hình một yếu tố là hợp lý, với RMSR thấp, cho thấy mô hình phù hợp tốt với dữ liệu.

Quan sát hệ số của nhân tố

```
factor_res$loadings  
  
##  
## Loadings:  
##          MR1  
## fixed_acidity     -0.370  
## volatile_acidity  
## citric_acid  
## residual_sugar      2.600  
## chlorides  
## free_sulfur_dioxide 13.050  
## total_sulfur_dioxide 50.197  
## density  
## p_h  
## sulphates  
## alcohol           -0.345  
##  
##          MR1  
## SS loadings    2697.121  
## Proportion Var 245.193
```

**Nhận xét:**

- Tổng hợp các biến: Nhân tố này có tải trọng cao cho các biến như total\_sulfur\_dioxide, free\_sulfur\_dioxide, và residual\_sugar. Đây là những yếu tố quan trọng trong bảo quản rượu, giúp bảo vệ rượu khỏi sự oxi hóa và sự phát triển của vi khuẩn hoặc nấm mốc.
- Chất lượng rượu: Các thành phần này cũng liên quan đến chất lượng của rượu trong quá trình bảo quản. Ví dụ, total\_sulfur\_dioxide và free\_sulfur\_dioxide là những yếu tố chính trong việc duy trì chất lượng của rượu qua thời gian.

- Chỉ số phân tách: total\_sulfur\_dioxide và free\_sulfur\_dioxide có tải trọng rất cao trong nhân tố này, cho thấy chúng là các yếu tố chính trong việc bảo quản rượu. Các biến như residual\_sugar và alcohol cũng ảnh hưởng đến sự ổn định của rượu.

## **Đề xuất khác:**

Do phân phối của các biến đặc trưng hóa học của rượu vang bị lệch trái rất nhiều nên có thể áp dụng mô hình hồi quy Poisson để phân tích.

## **Kết luận:**

Về cơ bản, mục tiêu đề ra ban đầu đã được hoàn thành, mặc dù có chút khó khăn khi phân loại rượu do đặc trưng của bộ dữ liệu nhưng việc phân tích đã mang lại kết quả hữu ích.