# Homework

Anna Miletova, 89231151

2025-05-23

## Description of the problem.

The files winequality-red.csv and winequality-white.csv contain the following data on selected red and white wines: *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol and quality.*

### Libraries

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

### Data preparation

```
red.wine <- read.table("winequality-red.csv", header = TRUE, sep = ";")
white.wine <- read.table("winequality-white.csv", header = TRUE, sep = ";")

red.wine$wine.type <- "red"
white.wine$wine.type <- "white"

data <- rbind(red.wine, white.wine)
```

# Task 1

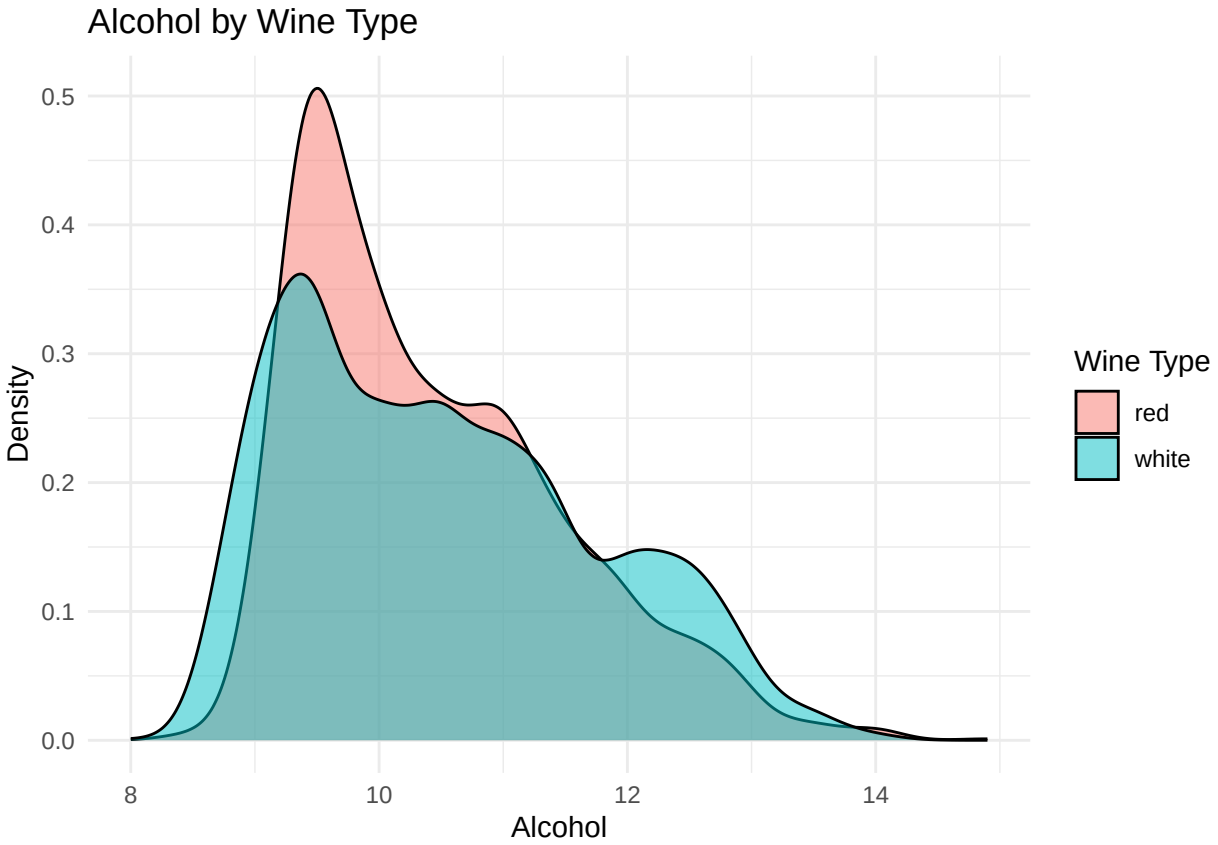
**Choose at least five variables from the list above and graphically depict the distribution of these variables for red and for white wines. Compare and interpret the results.**

```
plot_by_wine_type <- function (param, param_label) {
  ggplot(data, aes(x = {{ param }}, fill = wine.type)) +
    geom_density(alpha = 0.5) +
    labs(
      title = paste(param_label, "by Wine Type"),
      y = "Density",
      x = param_label,
      fill = "Wine Type"
    ) +
    theme_minimal()
}
```

The chosen variables are 1) alcohol, 2) pH, 3) quality, 4) fixed acidity, 5) volatile acidity and 6) residual sugar.

### 1) Alcohol

```
plot_by_wine_type(alcohol, "Alcohol")
```

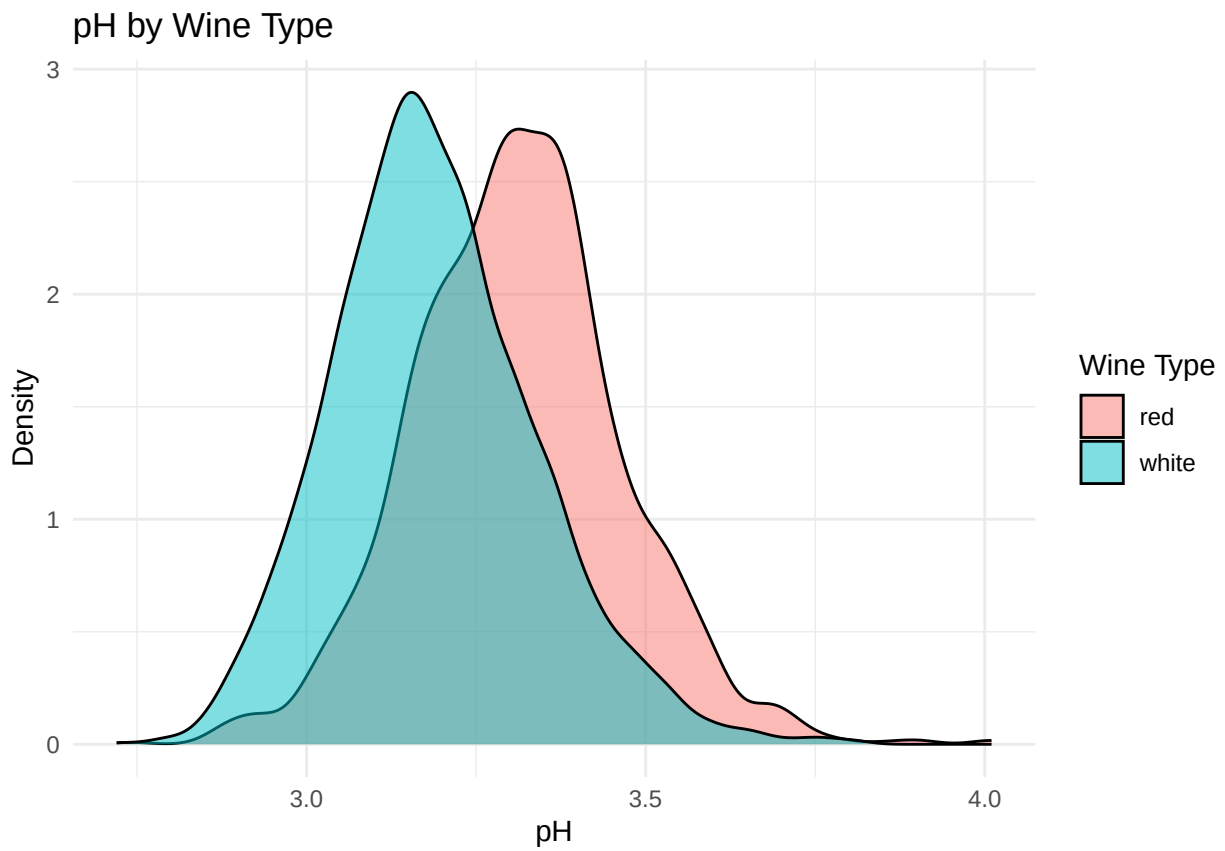



Density plot showing distribution of alcohol content for red and white wines. Both wine types show unimodal distribution with right-skewed density which implies that there are more wines of high alcohol content. The peaks of the both graphs are around the same point (~9.5%); the red wines graph is higher which indicates that a higher proportion of red wines have alcohol content around 9.5%, compared to white wines. Curve of the red wine is narrower than the curve of the white wine, which implies that white wine has more variability in alcohol content. The tail on the right suggests a few high-alcohol outliers but all of the wines in the dataset have alcohol content in an interval from 8 to 15%.

Red wines have **greater variablity** at alcohol content around **9-10%**, while white wines have slightly **bigger variability** at **higher alcohol contents** (above 12%).

**2) pH level**

```
plot_by_wine_type(pH, "pH")
```

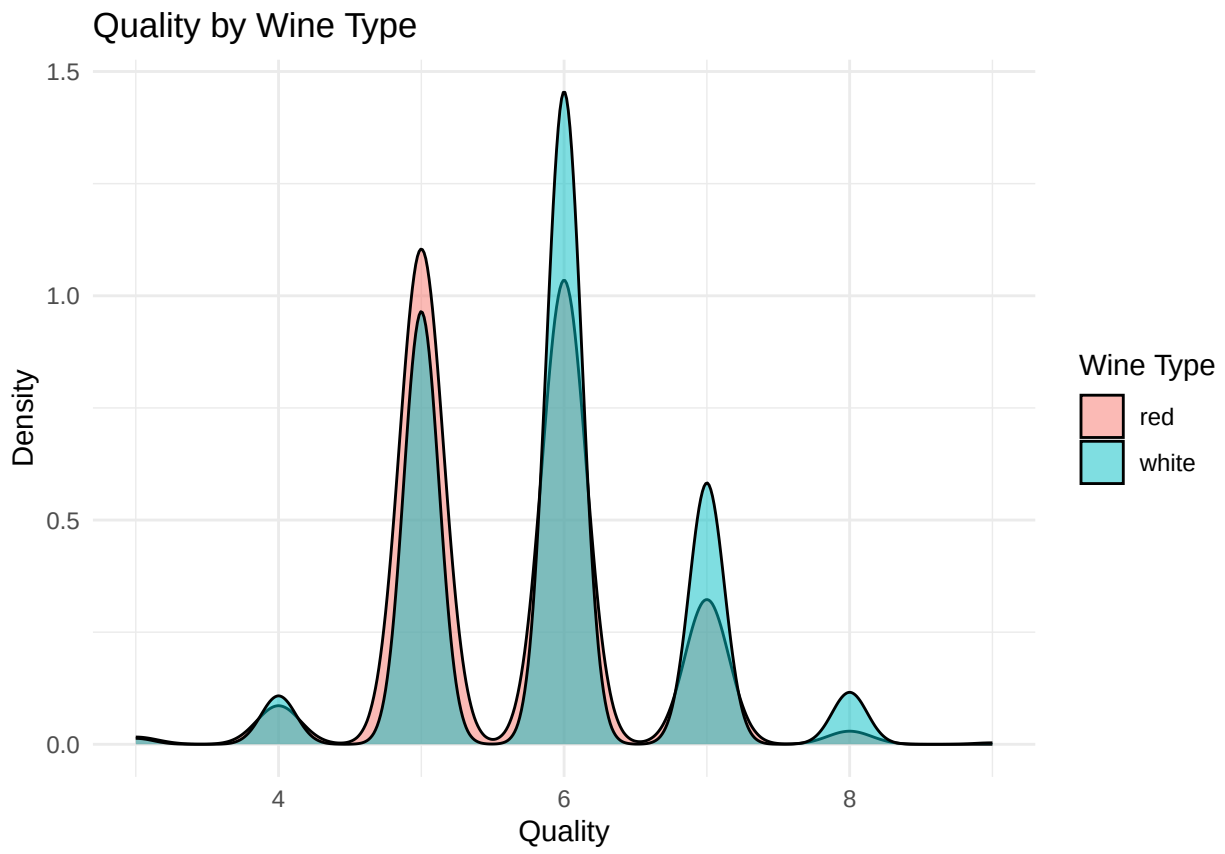



Density plot showing distribution of pH-level for red and white wines. Both wine types show unimodal distribution. White wine density is slightly right-skewed, while red wine density is nearly symmetrical. The curves overlap, nevertheless red wines tend to cluster at higher pH. Peak of the white wine distribution graph is around ~3.15 pH level, peak of the red wine distribution graph is around ~3.3 pH level. Narrow curve is shown for both wine types suggesting more consistent values. The long tail on the right may suggest a few outliers with pH level up to 4.0.

Generally, red wines have **slightly higher pH**, compared to white wines.

### 3) Quality

```
plot_by_wine_type(quality, "Quality")
```

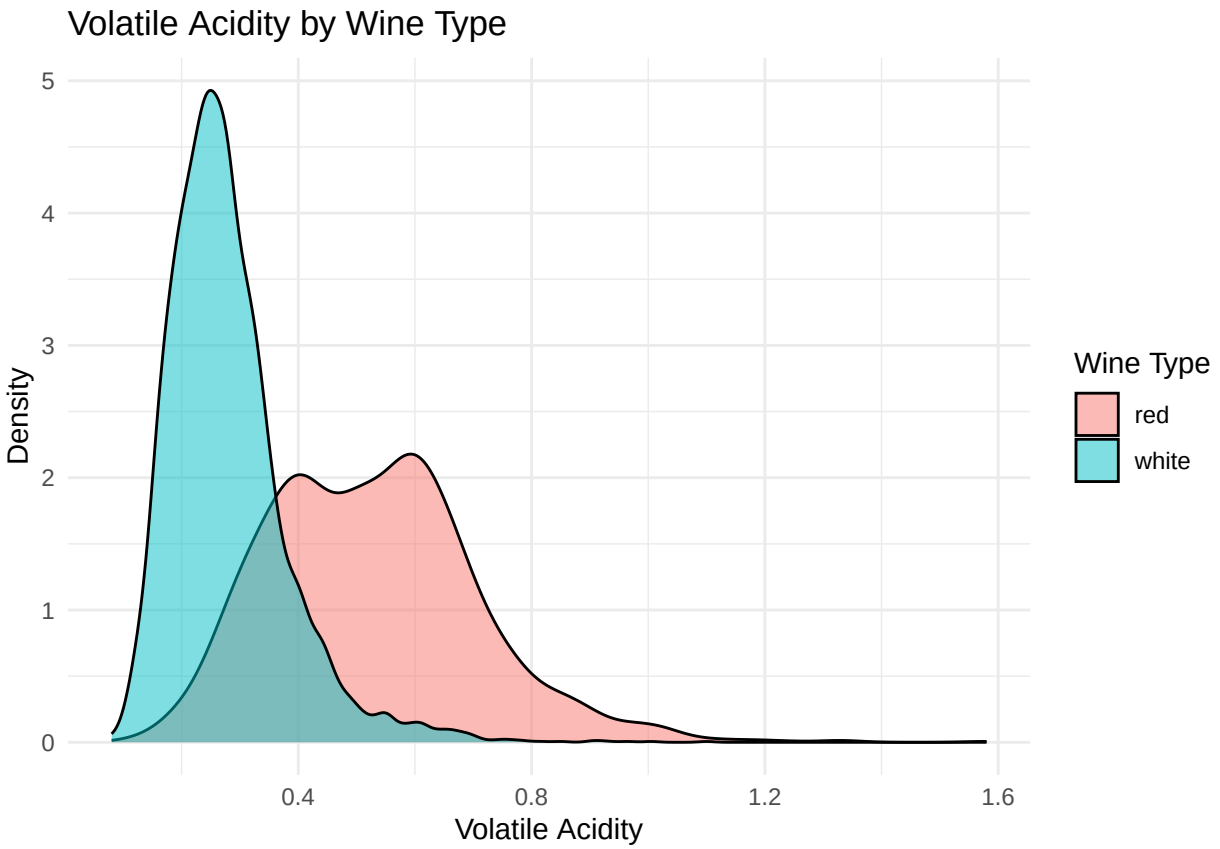



Density graphic is showing distribution of quality for red and white wines. The graphs of the both types are severely overlapped. Both wine types show multimodal distribution with peaks on the integer values (due to the integer type of the variable) with distinct peaks around quality level 5, 6 and 7. Both curves are not symmetric. The differences in the graphs is the highest peak – for red wine it is at quality 5, for white wine it is at quality 6 – moreover, the peaks at quality 7 and 8 are also higher for white wine, therefore, there are proportionally more white wines of higher quality in correspondence to the total number number of wine entries of a type.

The white wines achieve **higher quality ratings** and show **greater variability at higher quality levels** in comparison with red wines.

### 4) Volatile Acitidy

```
plot_by_wine_type(volatile.acidity, "Volatile Acidity")
```

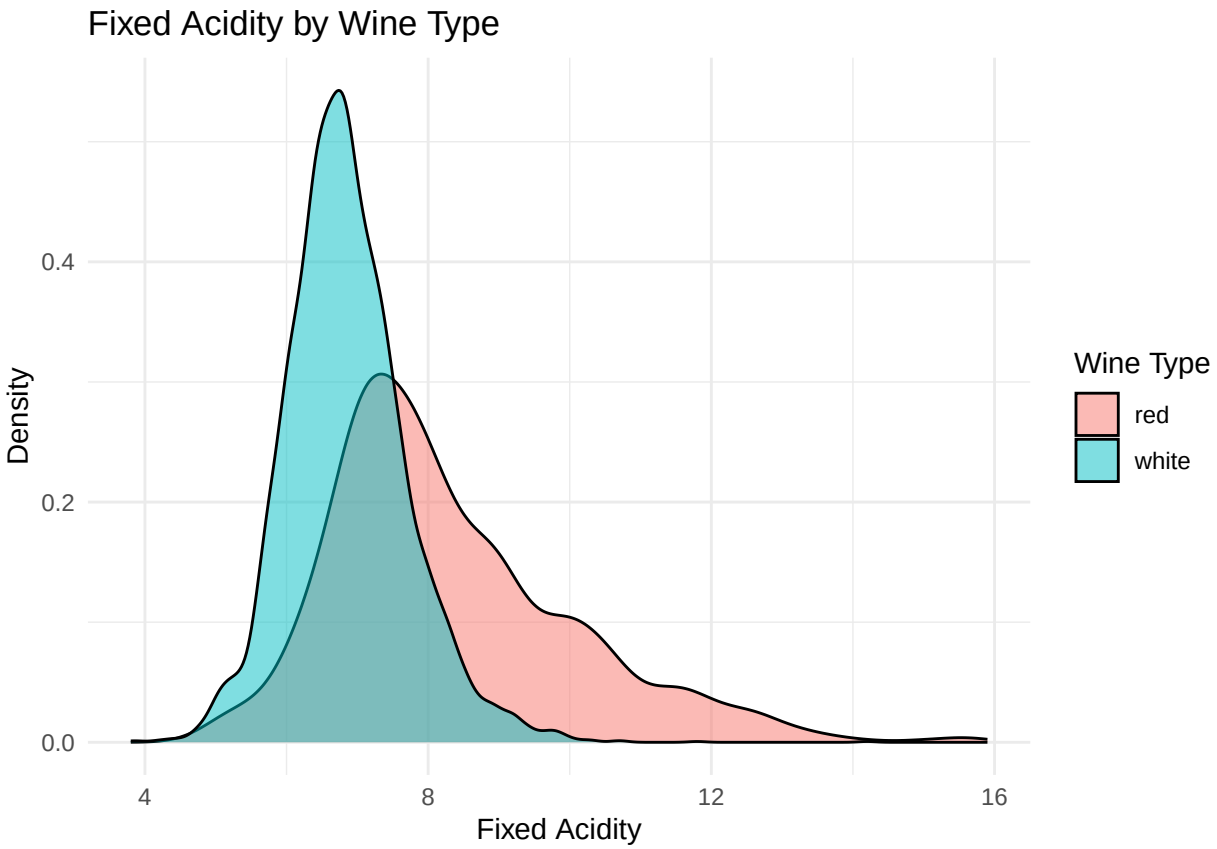



Density graphic is showing distribution of volatile acidity for red and white wines. White wines graph show unimodal, significantly right-skewed distribution with a peak around 0.25. White wines curve is very narrow, simultaneously has a long tail to the right, meaning their volatile acidity clusters around value 0.25 having outlying entries with volatile acidity up to almost 1.6. Red wines graph also has significantly right-skewed distribution. The distribution is implicitly bimodal, indicating two subgroups, with relatively close values around 0.6 (the highest pick of the graph) and around 0.4. Red wines curve is wider, implying more variability in volatile acidity among the type, compared to white wines. Long tail to the right indicates possible outliers at the values up to almost 1.6. Graphs of the wines overlap not significantly.

White wines usually have **volatile acidity around 0.25**, meanwhile red wines have **bigger variability on the whole interval from ~0.1 to ~1.6** with a significantly bigger variability on interval from ~0.4 to ~0.5.

### 5) Fixed Acidity

```
plot_by_wine_type(fixed.acidity, "Fixed Acidity")
```

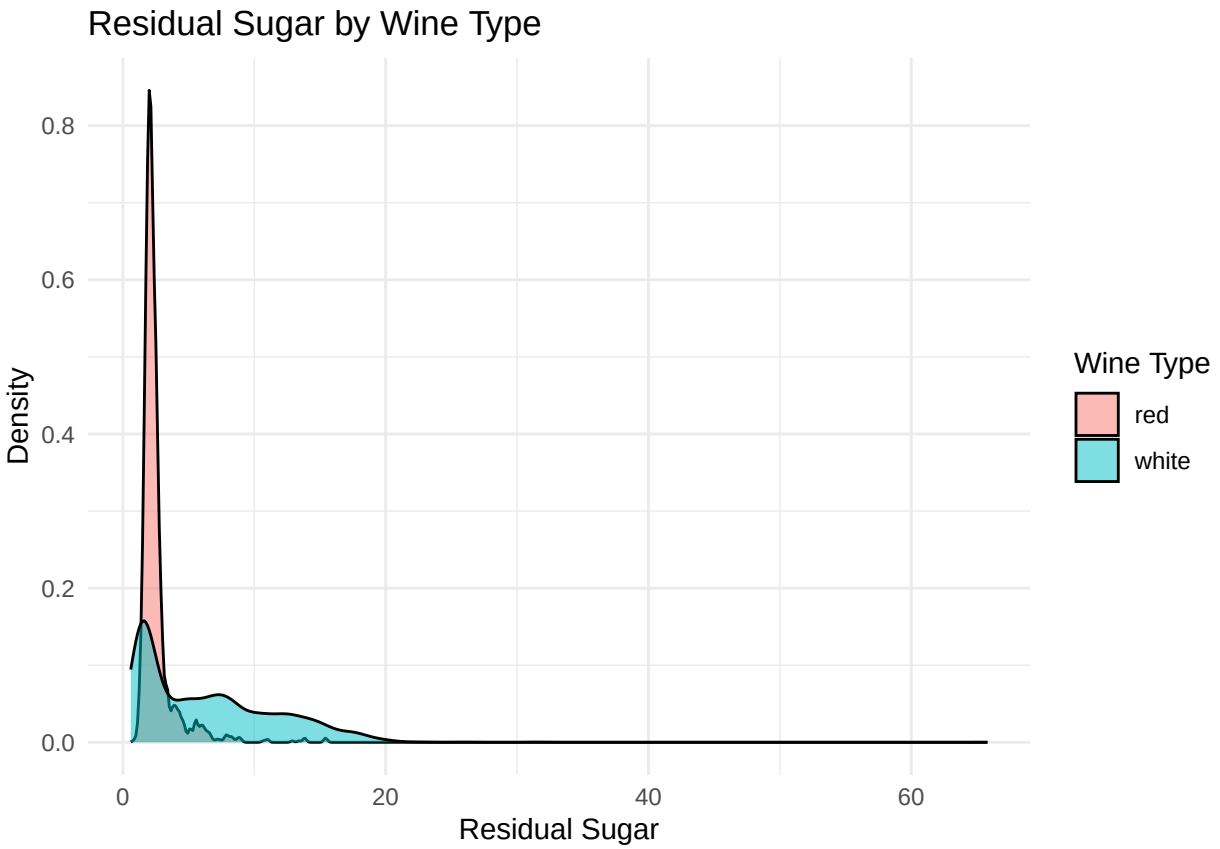



Density graphs of the fixed acidity show peak at a bit less than 7 for white wine and at around 7 for red wine. The white wine graph is narrow and it's pick is around 0.55; the red wine graph is wider and has a lower pick, meaning **bigger variability on the whole interval for red wines** and more **consistent values for white wine**. Right-skewed density and long tail on the right can be observed for both.

Roughly speaking, the graphs of the volatile acidity and fixed acidity are sufficiently similar, which shows **relationship between two criteria**.

### 6) Residual sugar

```
plot_by_wine_type(residual.sugar, "Residual Sugar")
```




Density graph of the residual sugar shows extremely narrow, right-skew density for red wine, and not-so-narrow, simultaneously extremely right-skew density for white wine. On a graph we see very long tail on the right, which indicates **outliers with extreme values, that go beyond 60,** meanwhile commonly the value is around 1-3 for red wine and is in range from 0 up to 20 for white wine.

Usual density graph wasn't informative enough because of the extreme skewness and extreme outlying values. It was decided to additionally use log-transformed plot to get more precise information on a given parameter for more typical values.

```
ggplot(data, aes(x = residual.sugar, fill = wine.type)) +
  geom_density(alpha = 0.5) +
  scale_x_log10() +

    labs(
      title = "Residual Sugar (log-transformed) by Wine Type",
      y = "Density",
      x = "Residual Sugar",
      fill = "Wine Type"
    )
```


After applying a log-transformation to the residual sugar variable, the density plot becomes more interpretable. White wines show a wider distribution with peaks around ~1.5, ~9 and ~15, indicating indicating several subgroups with typical residual sugar levels around 3, 30, and 80. Red wines are concentrated near lower values roughly around ~2.3, meaning most red wines have residual sugar around. Red wine graph still retains significant right-skewed density, but the log transformation helps reveal the primary distribution more clearly by reducing the impact of extreme outliers.

To sum up, red wine tends to **cluster at residual sugar value around 2.3**, while white wine has **greater variability on the interval from ~0.1 to almost 30**. The parameter has outliers up to the value of ~65.

## Task 2

### Calculate the quartiles and draw the box plot for (a) fixed acidity in red wines and (b) fixed acidity in white wines. Compare and interpret the results. Repeat the experiment for residual sugar and density.

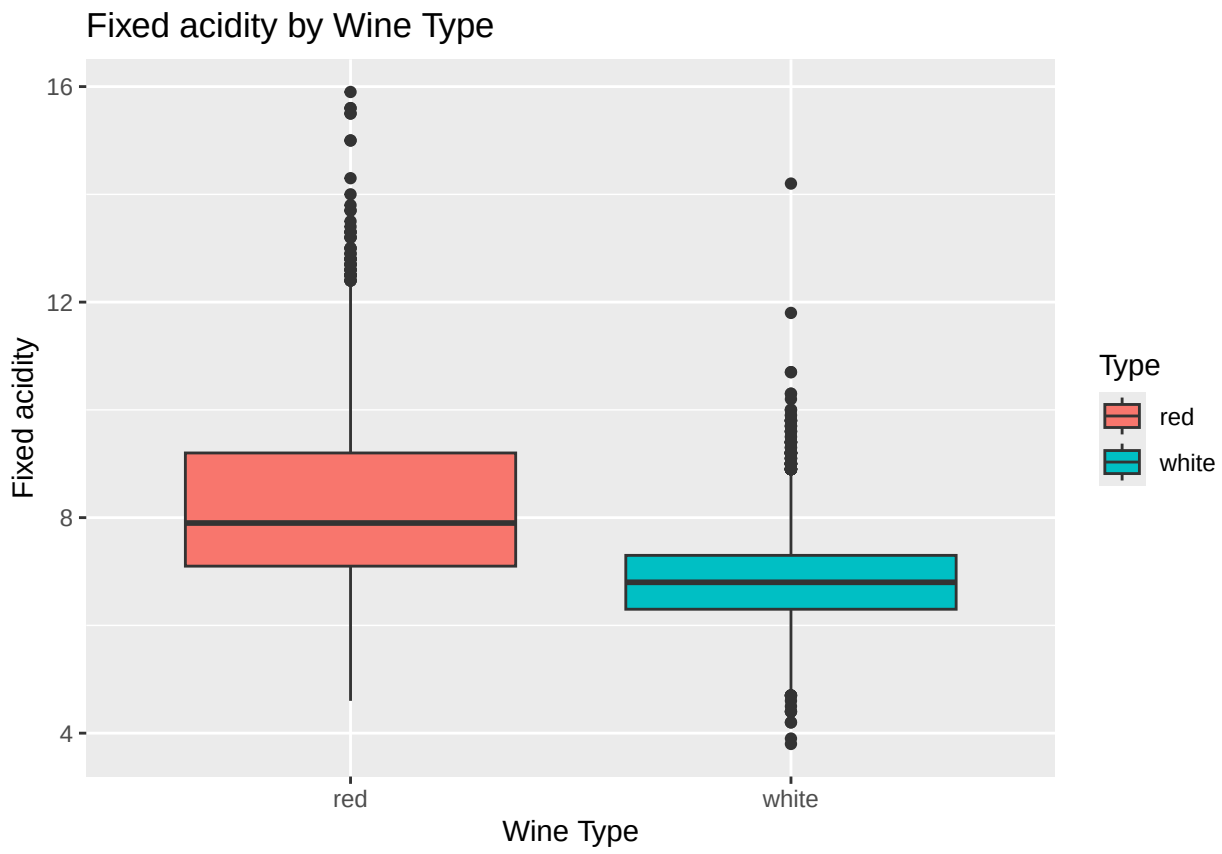
```
quartiles_and_box_plot <- function (param, param_label) {
  print(paste("(a) Quartiles for", param_label, "in red wines"))
  param_name <- deparse(substitute(param))

  red.wine[[param_name]] %>% summary %>% print

  print(paste("(b) Quartiles for", param_label, "in white wines"))
  white.wine[[param_name]] %>% summary %>% print
```

```
    ggplot(data, aes(x = wine.type, y = {{ param }}, fill = wine.type)) +
    geom_boxplot() +
    labs(
      title = paste(param_label, "by Wine Type"),
      x = "Wine Type",
      y = param_label,
      fill = "Type"
    )
}
```

**1) Fixed acidity**

```
quartiles_and_box_plot(fixed.acidity, "Fixed acidity")
```

```
## [1] "(a) Quartiles for Fixed acidity in red wines"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
## [1] "(b) Quartiles for Fixed acidity in white wines"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```



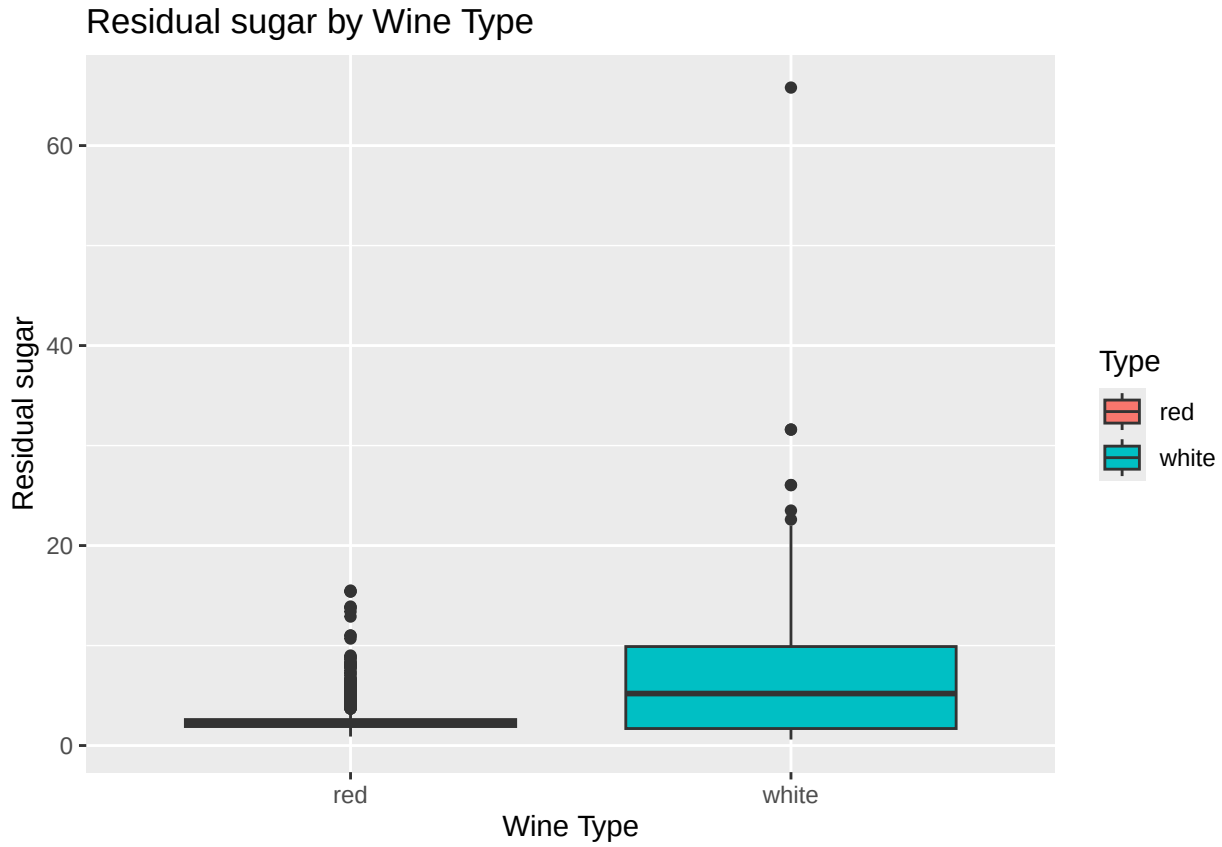

Box plot of fixed acidity for red and white wines shows that red wine has a higher median (7.9 vs 6.8) and a bigger IQR (2.1 vs 1) than white wine, indicating **more variability in fixed acidity level in a range**

**from ~7 to ~9 for red wine** and the fact that generally red wines have **higher fixed acidity level**, compared to white. Plot displays outliers with values more than 12 for red wine and outliers with values less than 5 and values more than 9 for white wine, indicating entries with **unusually high fixed acidity for both wines** and entries with **unusually low fixed acidity for white wine**.

**2) Residual sugar**

```
quartiles_and_box_plot(residual.sugar, "Residual sugar")
```

```
## [1] "(a) Quartiles for Residual sugar in red wines"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
## [1] "(b) Quartiles for Residual sugar in white wines"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```



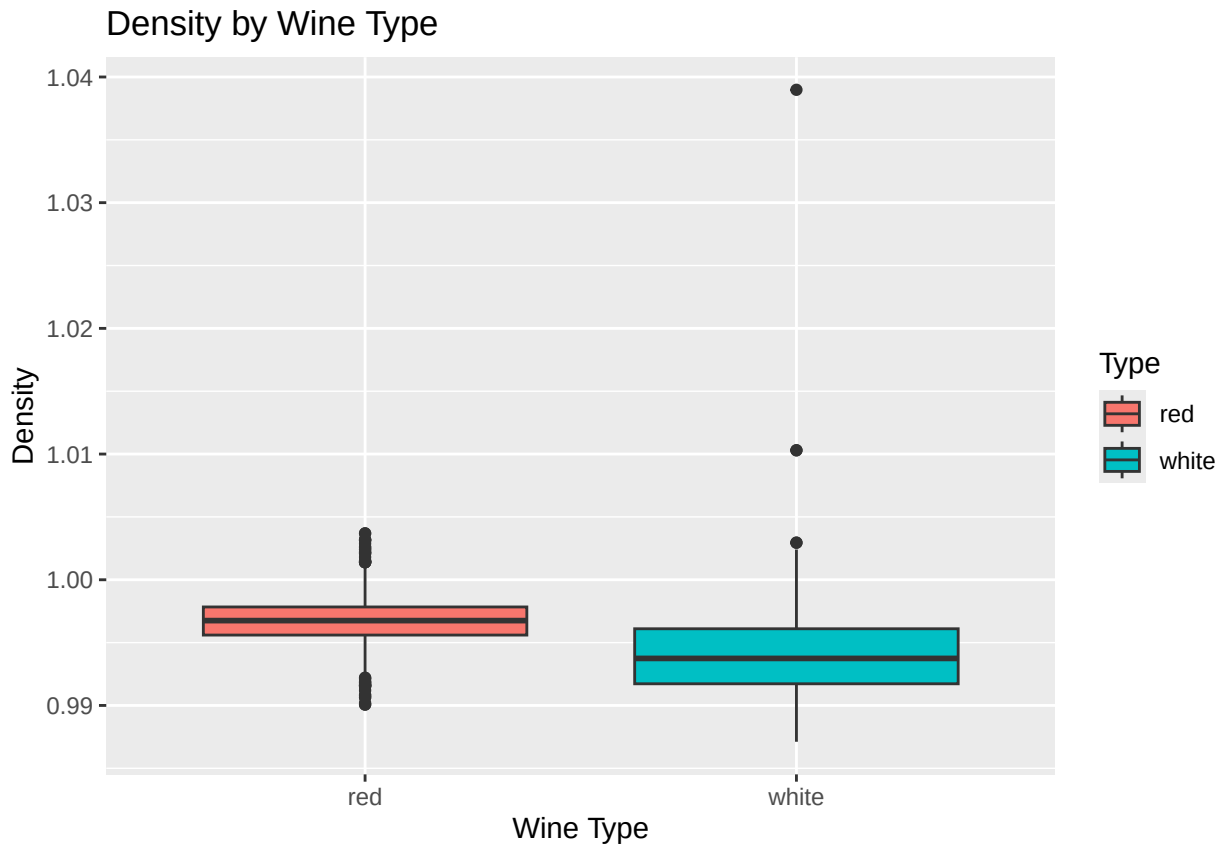

Box plot of residual sugar for red and white wines shows that white wine has a higher median (5.2 vs 2.2), suggesting that white wines usually have more residual sugar than red wines. White wines plot has much bigger IQR (8.2 vs 0.7), indicating the fact that **red wines are significantly clustered around median value** and white wines have bigger variability in terms of different residual sugar levels. Red wines plot has outliers which values are not that big (up to 15.5) and not that far from the whisker, meaning that there are **no entries with extreme residual sugar level for red wine**. White wine plot has a few outliers fairly close the whisker (on interval from ~20 to 35) and an extreme value - 65.8. Generally, the vast majority of **red wines have level of residual sugar around ~2**, while white wines have a **significantly larger**

**variability of entries** with the level from 1.7 to 9.9, and a few wines wines with **extreme values of more than 20 and up to 65.8**.

### 3) Density

```
quartiles_and_box_plot(density, "Density")
```

```
## [1] "(a) Quartiles for Density in red wines"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
## [1] "(b) Quartiles for Density in white wines"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```




Box plot of density for wines shows that red wines are more clustered around their median value of 0.9968 with a relatively small IQR (0.0022). White wines have lower median (0.9937) and wider IQR (0.9968) indicating greater spread of density values in white wines. Red wines show more compact distribution, whereas white wines vary more. In terms of outliers, white wines have a couple of distant from each other outliers with values of ~1.0025, ~1.01 and ~1.04, indicating a couple of entries with significantly extreme values of the parameter. Red wines' outliers remain not so distant from the whiskers, having the smallest value of 0.9901 and biggest value of 1.0037. **Red wines are more uniform** in terms of density level, compared to white wines.

## Task 3

**We would like to know whether mean fixed acidity of red wines of quality at most 6 is higher than mean fixed acidity of red wines of quality strictly larger than 6. Which statistical test would you use to answer this question and what are your conclusions? What would your answer be in the case of white wines? Repeat the experiment (for both red and white wines) for volatile acidity.**

```
red_low_q <- red.wine %>% filter(quality <= 6)
red_high_q <- red.wine %>% filter(quality > 6)

white_low_q <- white.wine %>% filter(quality <= 6)
white_high_q <- white.wine %>% filter(quality > 6)
```

To compare the mean fixed acidity between red wines of quality  6 and >6, a Welch two-sample t-test was used. This test is appropriate when comparing the means of two independent groups, especially when their variances may differ or the group sizes are unequal.

#### T-test for fixed acidity of red wines

```
t.test(red_low_q$fixed.acidity, red_high_q$fixed.acidity)
```

```
##
##  Welch Two Sample t-test
##
## data:  red_low_q$fixed.acidity and red_high_q$fixed.acidity
## t = -4.2635, df = 266.17, p-value = 2.797e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8919537 -0.3283941
## sample estimates:
## mean of x mean of y
##  8.236831  8.847005
```

T-test was conducted to compare the mean fixed acidity of red wines with quality scores  6 and >6. The results showed a statistically significant difference between the two groups (t = -4.26, $p < 0.001$), with the higher-quality red wines having a higher average fixed acidity (8.85) than the lower-quality ones (8.24).

Answer: **no**, mean fixed acidity of red wines of quality at most 6 is **NOT** higher than mean fixed acidity of red wines of quality strictly larger than 6.

#### T-test for fixed acidity of white wines

```
t.test(white_low_q$fixed.acidity, white_high_q$fixed.acidity)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  white_low_q$fixed.acidity and white_high_q$fixed.acidity
## t = 6.0401, df = 1856.5, p-value = 1.856e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1117293 0.2191758
## sample estimates:
## mean of x mean of y
##  6.890594  6.725142
```

The t-test on fixed acidity of white wine also showed a statistically significant difference between two groups ($t = 6.04$, $p < 0.001$), whereas the mean fixed acidity of white wine of higher-quality (6.89) is indeed larger than the mean fixed acidity of lower-quality white wine (6.73).

Answer: **yes**, mean fixed acidity of white wines of quality at most 6 **IS** higher than mean fixed acidity of white wines of quality strictly larger than 6.

### T-test for volatile acidity of red wines

```
t.test(red_low_q$volatile.acidity, red_high_q$volatile.acidity)
```

```
##
##  Welch Two Sample t-test
##
## data:  red_low_q$volatile.acidity and red_high_q$volatile.acidity
## t = 12.952, df = 325.29, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1200014 0.1629836
## sample estimates:
## mean of x mean of y
## 0.5470224 0.4055300
```

T-test was conducted to compare the mean volatile acidity of red wines with quality scores  6 and >6. The test shows a highly significant difference between the groups ($t = 12.95$, $p < 0.001$), with low-quality red wines having much higher volatile acidity (0.547) compared to high-quality ones (0.406).

Answer: **yes**, mean volatile acidity of red wines of quality at most 6 **IS** higher than mean volatile acidity of red wines of quality strictly larger than 6.

### T-test for volatile acidity of white wines

```
t.test(white_low_q$volatile.acidity, white_high_q$volatile.acidity)
```

```
##
##  Welch Two Sample t-test
##
## data:  white_low_q$volatile.acidity and white_high_q$volatile.acidity
## t = 4.9429, df = 1809.9, p-value = 8.409e-07
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  0.009924419 0.022980907
## sample estimates:
## mean of x mean of y
## 0.2818017 0.2653491
```

T-test on comparison of the mean volatile acidity of white wines with quality scores  6 and >6 also shows a significant difference between the groups ($t = 4.94$, $p < 0.001$), with low-quality white wines having higher volatile acidity (0.282) compared to high-quality ones (0.265).

Answer: **yes**, mean volatile acidity of white wines of quality at most 6 **IS** higher than mean volatile acidity of white wines of quality strictly larger than 6.

# Task 4

## We would like to know whether there is a correlation between the perceived quality of red wines and the fixed acidity level. Which statistical test would you use to answer this question and what are your conclusions? Furthermore, test the correlation between the perceived quality of red wines and residual sugar levels. Lastly, repeat both tests for white wines and compare the results.

Spearman correlation was chosen due to the significantly distinct outliers for fixed acidity and residual sugar for both types of wine.

### Fixed acidity

```
cor.test(red.wine$fixed.acidity, red.wine$quality, method="spearman")
```

```
## Warning in cor.test.default(red.wine$fixed.acidity, red.wine$quality, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  red.wine$fixed.acidity and red.wine$quality
## S = 603652045, p-value = 4.801e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1140837
```

The test shows a weak but statistically significant positive correlation ($rho = 0.114$, $p < 0.001$), suggesting that as **fixed acidity increases**, red wine **quality tends to increase** slightly.

```
cor.test(white.wine$fixed.acidity, white.wine$quality, method="spearman")
```

```
## Warning in cor.test.default(white.wine$fixed.acidity, white.wine$quality, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  white.wine$fixed.acidity and white.wine$quality
## S = 2.1239e+10, p-value = 3.183e-09
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.08448545
```

The correlation analysis shows a weak but statistically significant negative correlation (rho $-0.084$, p $<$ 0.001) between fixed acidity and quality in white wines. This suggests that **higher fixed acidity** is mildly associated with **lower quality** ratings in white wines.

**Negative correlation between fixed acidity and perceived quality in white wines contrasts with the positive association observed for red wines.**

### Residual sugar

```
cor.test(red.wine$residual.sugar, red.wine$quality, method="spearman")
```

```
## Warning in cor.test.default(red.wine$residual.sugar, red.wine$quality, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  red.wine$residual.sugar and red.wine$quality
## S = 659549989, p-value = 0.2002
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.03204817
```

A Spearman's rank correlation test showed no significant correlation between residual sugar and quality in red wines ( $= 0.032$, p $= 0.20 > 0.001$). This suggests that **residual sugar does not have** a consistent **impact on perceived quality** in **red wines**.

```
cor.test(white.wine$residual.sugar, white.wine$quality, method="spearman")
```

```
## Warning in cor.test.default(white.wine$residual.sugar, white.wine$quality, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  white.wine$residual.sugar and white.wine$quality
## S = 2.1191e+10, p-value = 8.822e-09
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.08206979
```

A Spearman's rank correlation analysis reveals a weak but statistically significant negative correlation between residual sugar and quality in white wines (rho = –0.082, $p < 0.001$). This indicates that **higher residual sugar levels** are slightly associated with **lower quality scores** in white wines. While the effect is small, the trend is consistent across the dataset.

**Therefore, despite there is no correlation between residual sugar and quality for red wine, there is a weak correlation between residual sugar and perceived quality for red wine.**

# Task 5

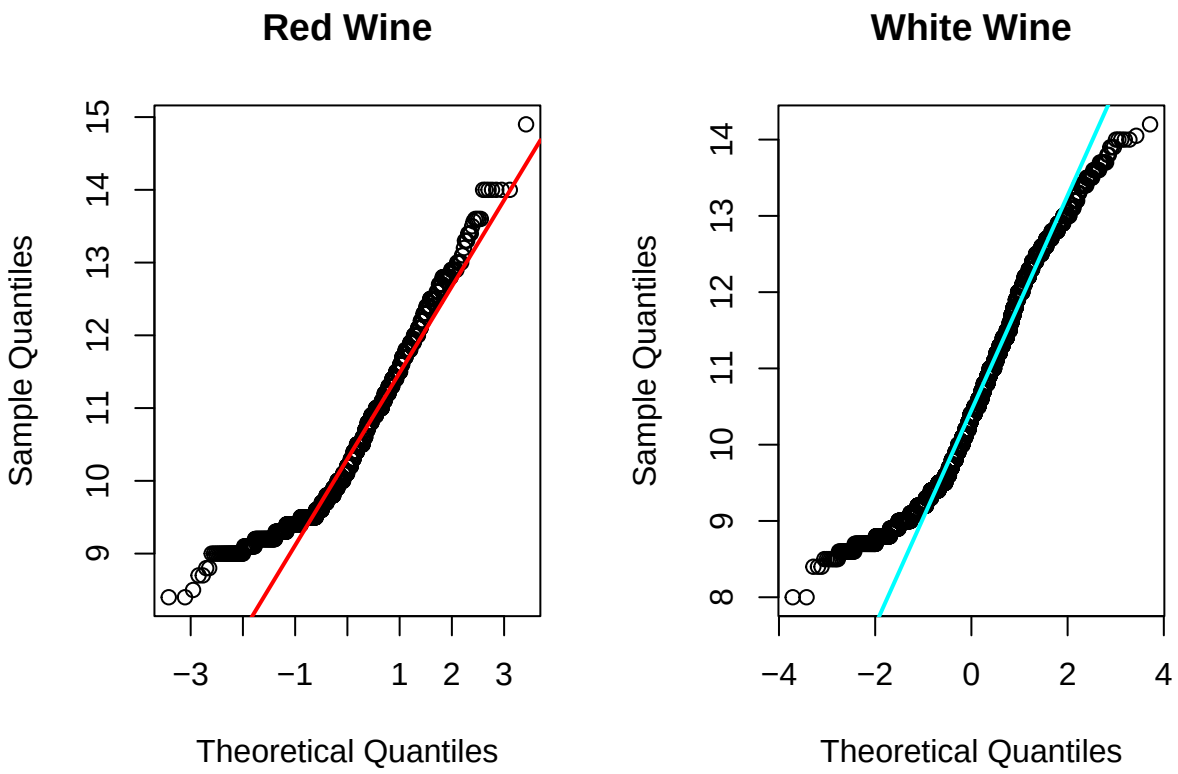
## For at least five of the above mentioned variables check whether the data for red wines is normally distributed. Repeat the analysis for white wines and interpret the results.

The chosen variables are 1) alcohol, 2) pH, 3) chlorides, 4) free sulfur dioxide and 5) total sulfur dioxide.

```
qq_by_param <- function (param) {
  param_name <- deparse(substitute(param))

  par(mfrow = c(1, 2))

  qqnorm(red.wine[[param_name]], main = "Red Wine")
  qqline(red.wine[[param_name]], col = "red", lwd = 2)

  qqnorm(white.wine[[param_name]], main = "White Wine")
  qqline(white.wine[[param_name]], col = "cyan", lwd = 2)

  par(mfrow = c(1, 1))
}
```

### 1) Alcohol

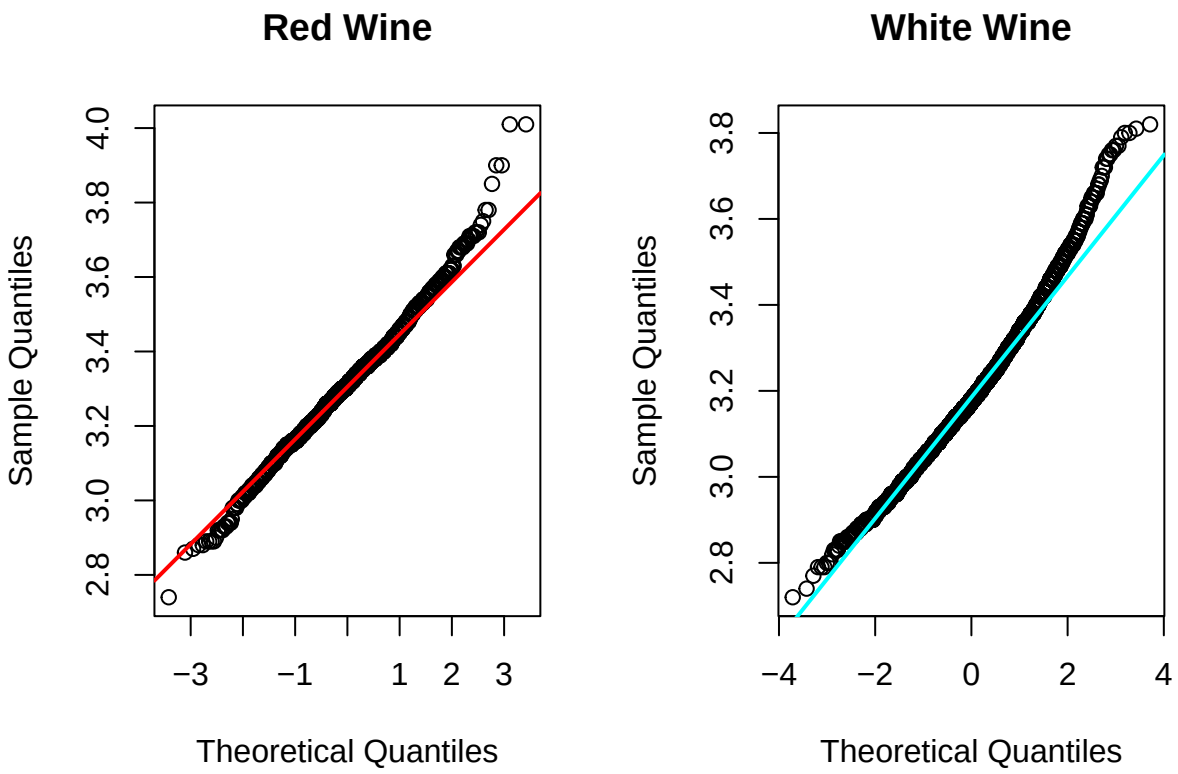
```
qq_by_param(alcohol)
```




Q-Q plot for alcohol content shows mild deviation from normal distribution for both types of wine. Deviation is right-skrewed but remains close to normal for red and white wine.

**2) pH level**

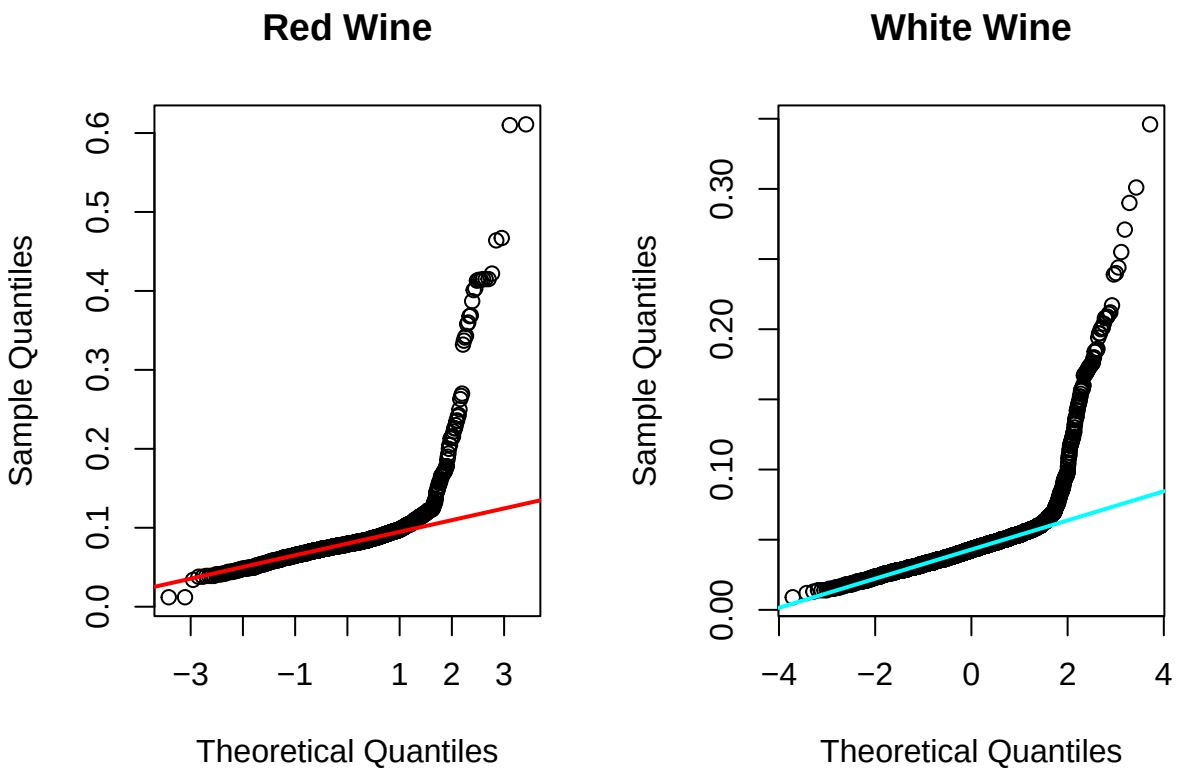
```
qq_by_param(pH)
```




Q-Q plot for pH level clearly shows the long tail on the right for distributions of both wine types. The rest of the graph is significantly close to normal distribution.

**3) Chlorides**

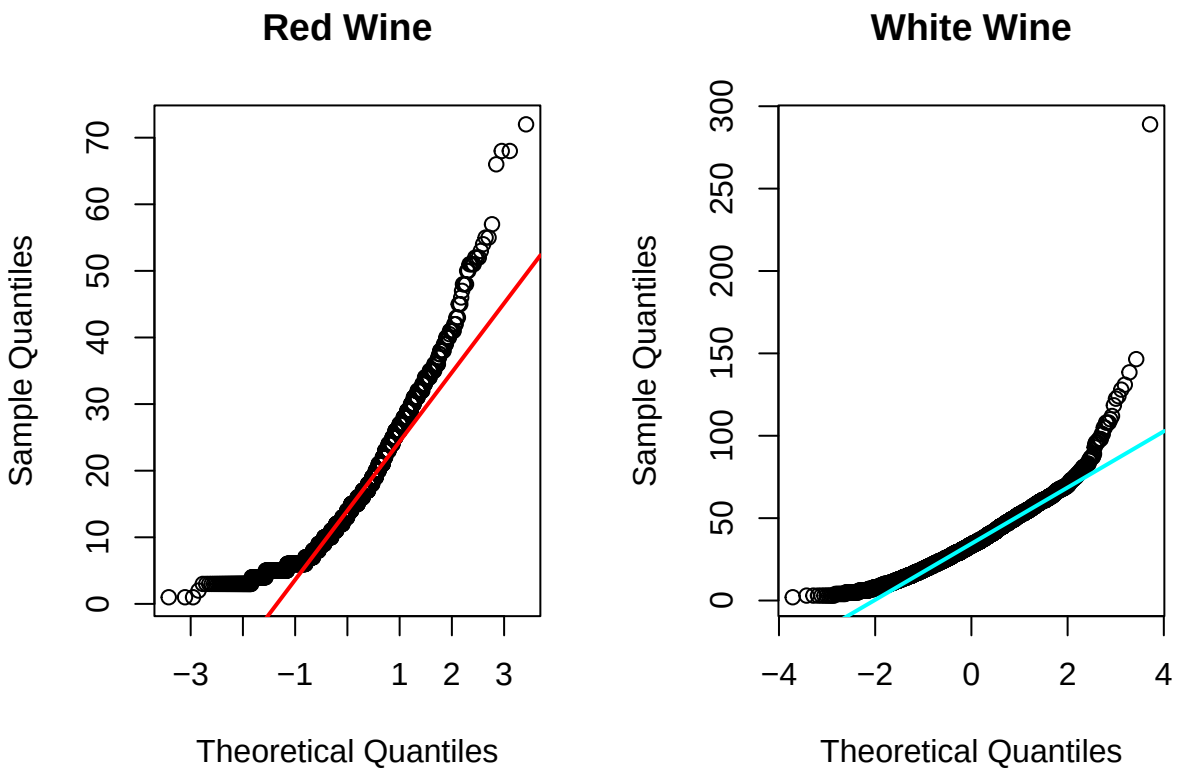
```
qq_by_param(chlorides)
```




Q-Q plot for chlorides for both wine types looks similar, having right tail dramatically above the line, meaning the strong right skew. The distribution has many small values and a few very large outliers.

**4) Free sulfur dioxide**

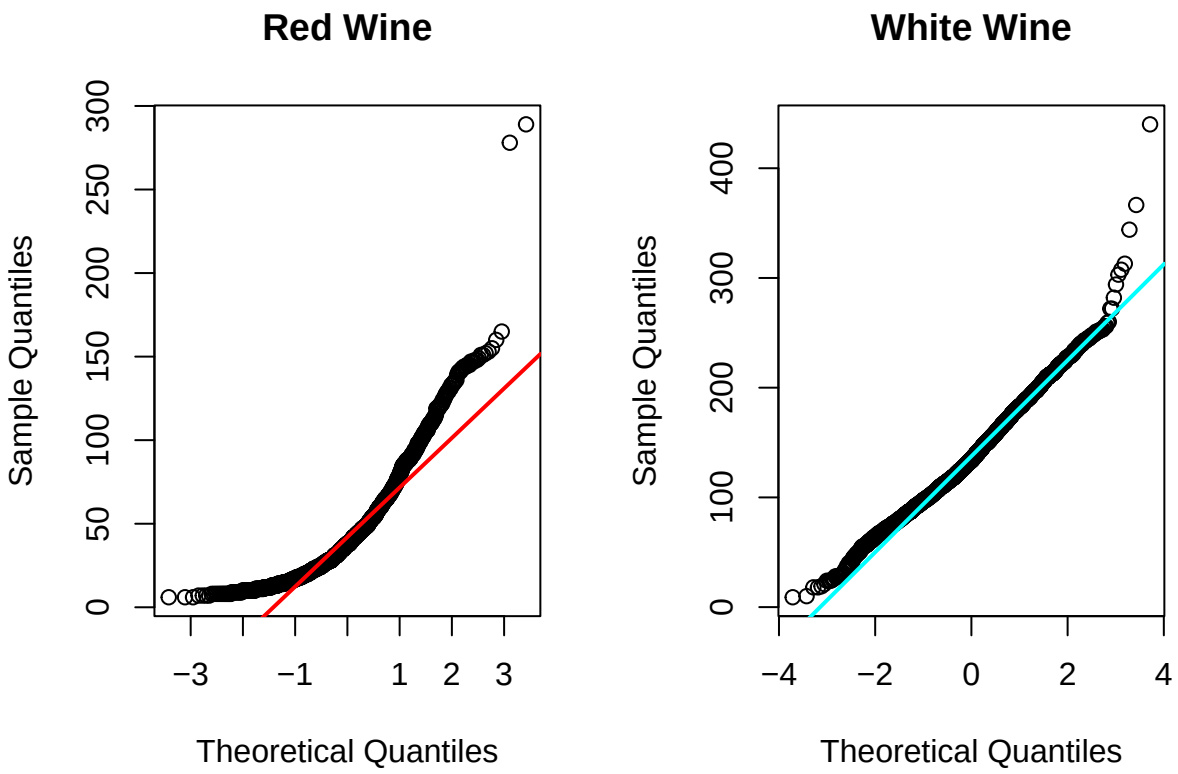
```
qq_by_param(free.sulfur.dioxide)
```




The Q-Q plot for free sulfur dioxide shows long right tail, strong outliers and pretty significant right skrew.

**5) Total sulfur dioxide**

```
qq_by_param(total.sulfur.dioxide)
```




The Q-Q plot for total sulfur dioxide also shows long right tail, strong outliers and pretty significant right skrew just as the plot for free sulfur dioxide for both wines.