1 Sample statistics

Immanuel Klein

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

The task is about analyzing the white wine quality dataset from the UCI Machine Learning Repository, focusing on two variables: `volatile acidity` and `residual sugar`. The problem requires adding a binary variable indicating good or bad wine called `quality`. We then compare `volatile acidity` and `residual sugar` across wines that are good (`quality` > 5) or bad (`quality` 5). For both variables, the analysis includes generating and interpreting histograms, summary statistics, boxplots, QQ-plots, and empirical distribution functions to compare the distributions between good and bad wines.

Exercise (a)

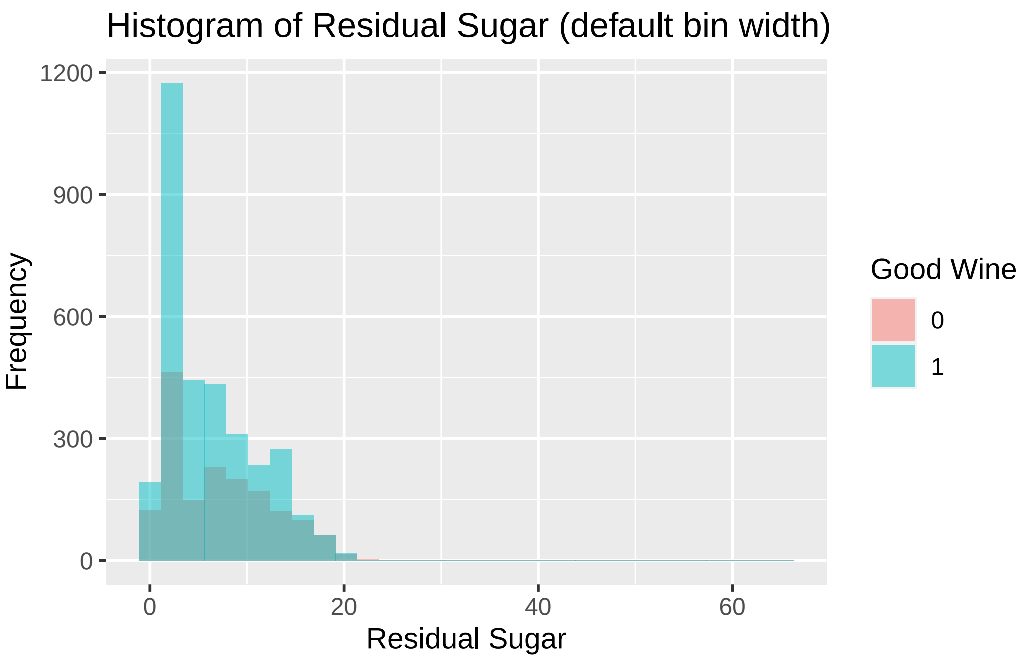
```
# Reading the whole data set
winequality.white <- read.csv("wine+quality/winequality-white.csv", sep = ";")

# Using only volatile.acidity and residual.sugar
# and adding binary variable good which is 1 if quality > 5 and 0 otherwise.
working.df <- winequality.white %>%
  mutate(good = ifelse(quality > 5, 1, 0)) %>%
  select(volatile.acidity, residual.sugar, good)

head(working.df)
```

```
  volatile.acidity residual.sugar good
1             0.27           20.7    1
2             0.30            1.6    1
3             0.28            6.9    1
4             0.23            8.5    1
5             0.23            8.5    1
6             0.28            6.9    1
```

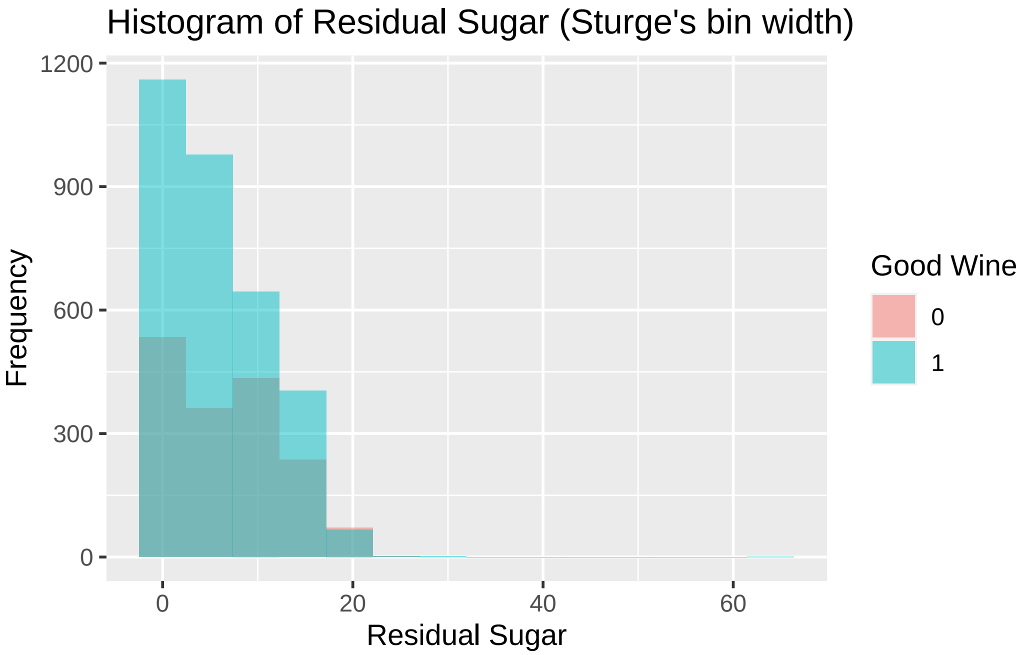
Exercise (b)

Histograms

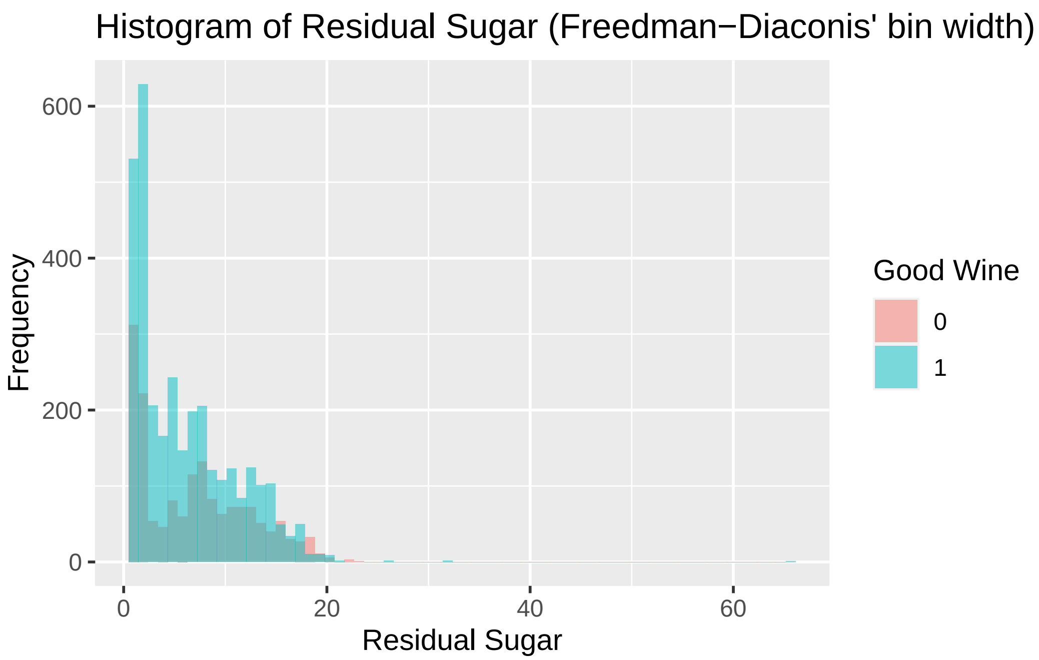
```
# Plotting the histogram with the default bin width
ggplot(working.df) +
  aes(x = residual.sugar, fill = factor(good)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 30) +
  labs(title = "Histogram of Residual Sugar (default bin width)",
       x = "Residual Sugar",
       y = "Frequency",
       fill = "Good Wine")
```



```
# Plotting the histogram and calculating Sturge's bin width
ggplot(working.df) +
  aes(x = residual.sugar, fill = factor(good)) +
  geom_histogram(position = "identity", alpha = 0.5,
                 binwidth = diff(range(working.df$residual.sugar)) /
                            (1 + log2(length(working.df$residual.sugar)))) +
  labs(title = "Histogram of Residual Sugar (Sturge's bin width)",
       x = "Residual Sugar",
       y = "Frequency",
       fill = "Good Wine")
```



```
# Plotting the histogram and calculating Freedman-Diaconis' bin width
ggplot(working.df) +
  aes(x = residual.sugar, fill = factor(good)) +
  geom_histogram(position = "identity", alpha = 0.5,
                 binwidth = 2 * IQR(working.df$residual.sugar) /
                               (length(working.df$residual.sugar)^(1/3))) +
  labs(title = "Histogram of Residual Sugar (Freedman-Diaconis' bin width)",
       x = "Residual Sugar",
       y = "Frequency",
       fill = "Good Wine")
```



As `residual sugar` increases, the number of both good and bad wines decreases. This trend might be due to underrepresentation of wines with high `residual sugar` in the dataset, or it could be a general thing that such wines are less represented in the population. Additionally, there is a noticeable increase in the number of bad wines with moderate levels of `residual sugar`. If we disregard the low `residual sugar` range - where both good and bad wines are most common - we can see that many bad wines have moderate `residual sugar` content.

Summary Statistics

```
summary.stats <- working.df %>%
  group_by(good) %>%
  summarise(
    "Mean" = mean(residual.sugar, na.rm = TRUE),
    "Median" = median(residual.sugar, na.rm = TRUE),
    "Standard Deviation" = sd(residual.sugar, na.rm = TRUE),
    "IQR" = IQR(residual.sugar, na.rm = TRUE),
    "min" = min(residual.sugar, na.rm = TRUE),
    "max" = max(residual.sugar, na.rm = TRUE)
  )

summary.stats
```

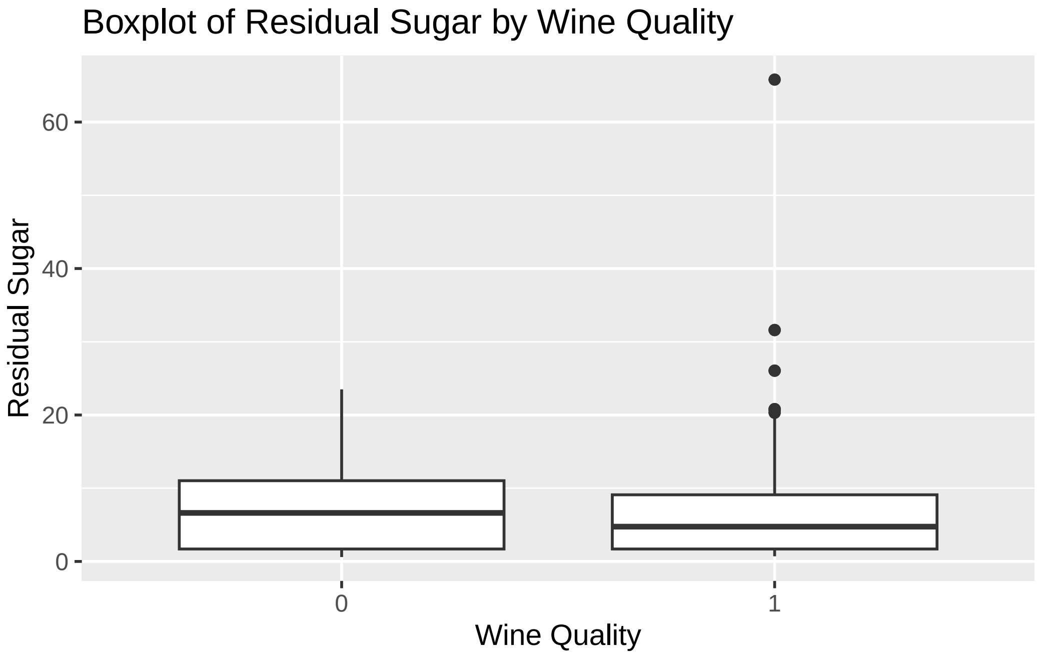
```
# A tibble: 2 x 7
```

```
   good  Mean Median `Standard Deviation`   IQR   min   max
  <dbl> <dbl>  <dbl>                <dbl> <dbl> <dbl> <dbl>
1     0  7.05   6.62                 5.28  9.33   0.6  23.5
2     1  6.06   4.75                 4.93  7.4    0.7  65.8
```

- The mean `residual sugar` is higher for bad wines (7.05) compared to good wines (6.06).
- The median `residual sugar` is also higher for bad wines (6.625) compared to good wines (4.75). On average, bad wines seem to have slightly higher `residual sugar` content than good wines.
- The standard deviation is slightly higher for bad wines (5.28) compared to good wines (4.93), so there is more variability in the `residual sugar` content among bad wines.
- The IQR is also higher for bad wines (9.325) compared to good wines (7.4). This suggests that bad wines have a wider spread in their `residual sugar` in the central portion of the data.
- The minimum value of residual sugar is very similar for both categories (0.6 for bad wines and 0.7 for good wines).
- The maximum value of `residual sugar` is much higher for good wines (65.8) compared to bad wines (23.5). This difference indicates that some good wines have extremely high levels of `residual sugar` (outliers).

Box Plot

```
boxplot <- ggplot(working.df) +
  aes(x = factor(good), y = residual.sugar) +
  geom_boxplot() +
  labs(title = "Boxplot of Residual Sugar by Wine Quality",
       x = "Wine Quality",
       y = "Residual Sugar")

boxplot
```

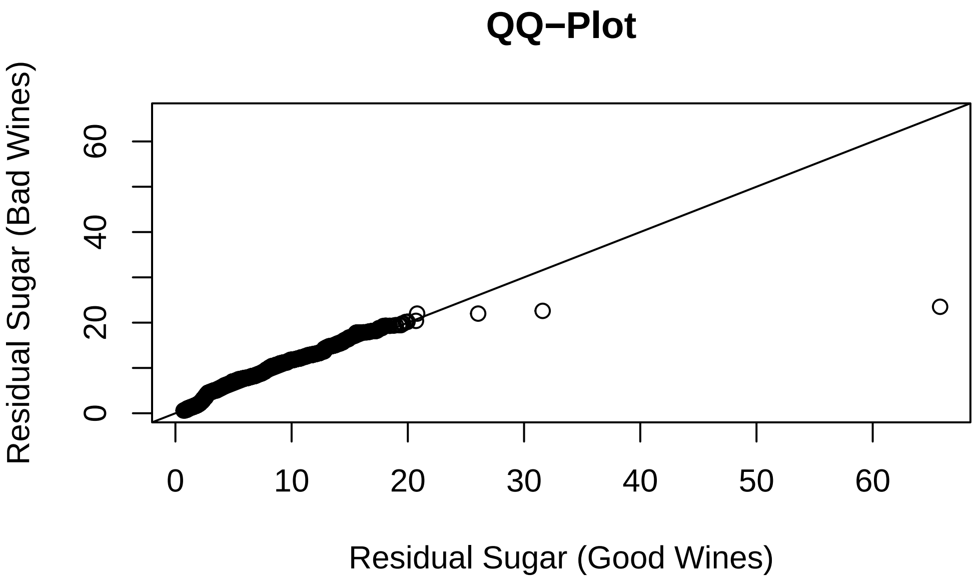


The box plot visualization reveals several key points consistent with the summary statistics:

- Good wines show a wider overall range and more outliers compared to bad wines.
- Bad wines have a higher IQR, indicating greater variability in the middle 50% of their `residual sugar` content.
- The range of bad wines is entirely contained within the whiskers, showing fewer extreme values beyond the central data points.

QQ-Plot

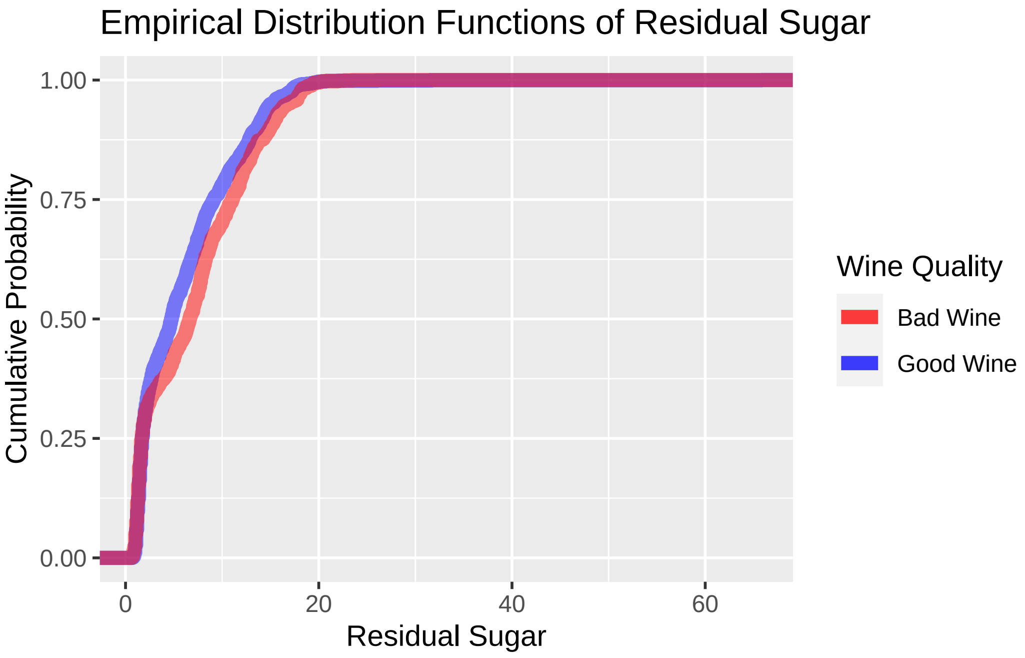
```r
residual.sugar.good <- working.df %>%
  filter(good == 1) %>%
  pull(residual.sugar)
residual.sugar.bad <- working.df %>%
  filter(good == 0) %>%
  pull(residual.sugar)

xlim.range <- range(residual.sugar.good, residual.sugar.bad)
ylim.range <- range(residual.sugar.good, residual.sugar.bad)

# Create QQ-plot
qqplot(residual.sugar.good, residual.sugar.bad,
       xlab = "Residual Sugar (Good Wines)",
       ylab = "Residual Sugar (Bad Wines)",
       main = "QQ-Plot",
       xlim = xlim.range,
       ylim = ylim.range)
abline(0, 1)
```



- If the distributions were identical, the points would fall on the 45-degree line. Here, we can see a smaller deviation from this line in the middle part and a strong deviation in the tails.
- Almost exclusively, points fall slightly above the line (indicating higher `residual sugar` for bad wines).
- The outliers of good wines are apparent.

Empirical Distribution

```
working.good <- working.df %>% filter(good == 1)
working.bad <- working.df %>% filter(good == 0)

ggplot() +
  stat_ecdf(data = working.good,
            aes(x = residual.sugar, color = "Good Wine"),
            geom = "step",
            linewidth = 2.5,
            alpha = 0.5) +
  stat_ecdf(data = working.bad,
            aes(x = residual.sugar, color = "Bad Wine"),
            geom = "step", linewidth = 2.5,
            alpha = 0.5) +
  scale_color_manual(values = c("Good Wine" = "blue", "Bad Wine" = "red")) +
  labs(title = "Empirical Distribution Functions of Residual Sugar",
       x = "Residual Sugar",
       y = "Cumulative Probability",
       color = "Wine Quality")
```



- Since the median residual sugar content of good wines is lower, the curve rises more steeply initially, indicating that a higher proportion of good wines have lower `residual sugar` content.

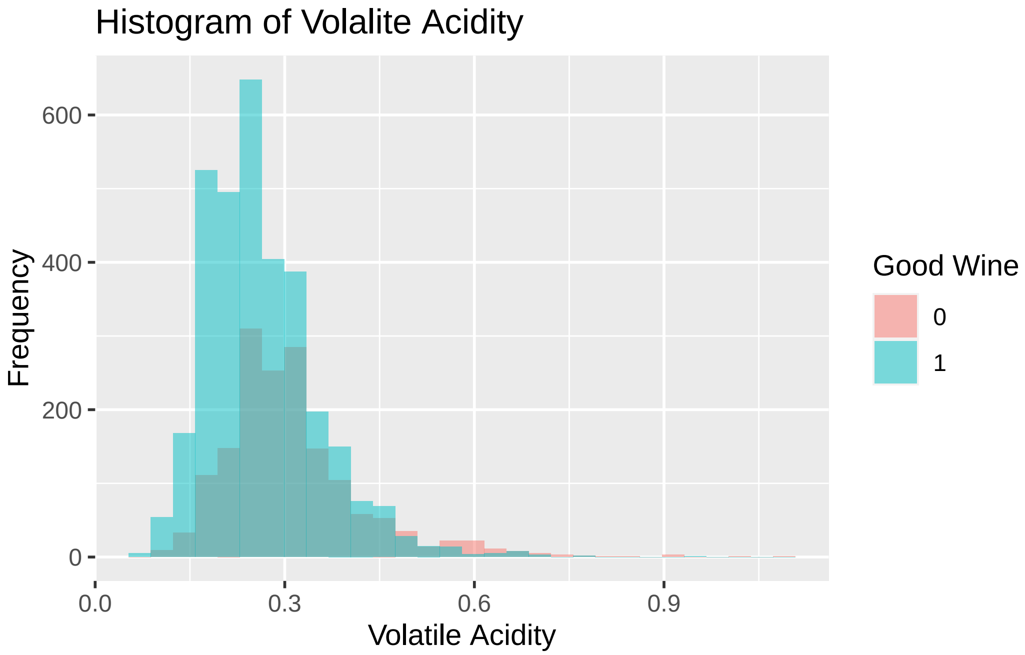
- Given the higher median residual sugar content of bad wines, the slope is less steep in the middle part compared to good wines, indicating that a lower proportion of bad wines have lower `residual sugar` content.

Exercise (c)

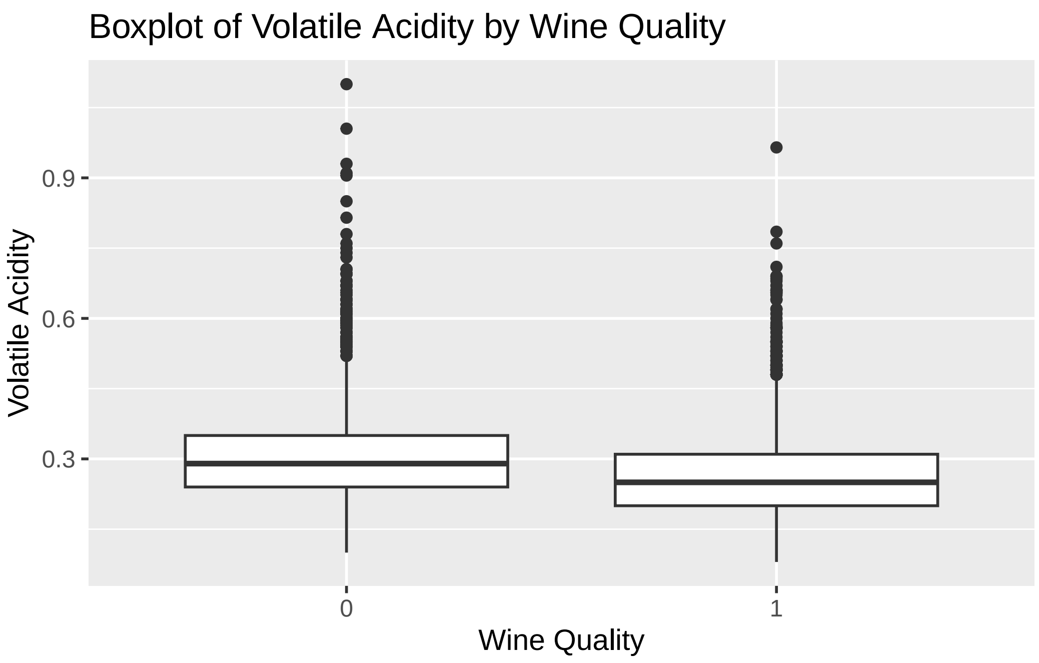
```
# Summary statistics
summary.stats <- working.df %>%
  group_by(good) %>%
  summarise(
    "Mean" = mean(volatile.acidity, na.rm = TRUE),
    "Median" = median(volatile.acidity, na.rm = TRUE),
    "Standard Deviation" = sd(volatile.acidity, na.rm = TRUE),
    "IQR" = IQR(volatile.acidity, na.rm = TRUE),
    "min" = min(volatile.acidity, na.rm = TRUE),
    "max" = max(volatile.acidity, na.rm = TRUE)
  )

summary.stats
```

```
# A tibble: 2 x 7
   good  Mean Median `Standard Deviation`   IQR   min   max
  <dbl> <dbl>  <dbl>                <dbl> <dbl> <dbl> <dbl>
1     0 0.310   0.29               0.113   0.11  0.1  1.1
2     1 0.262   0.25               0.0901  0.11  0.08 0.965
```

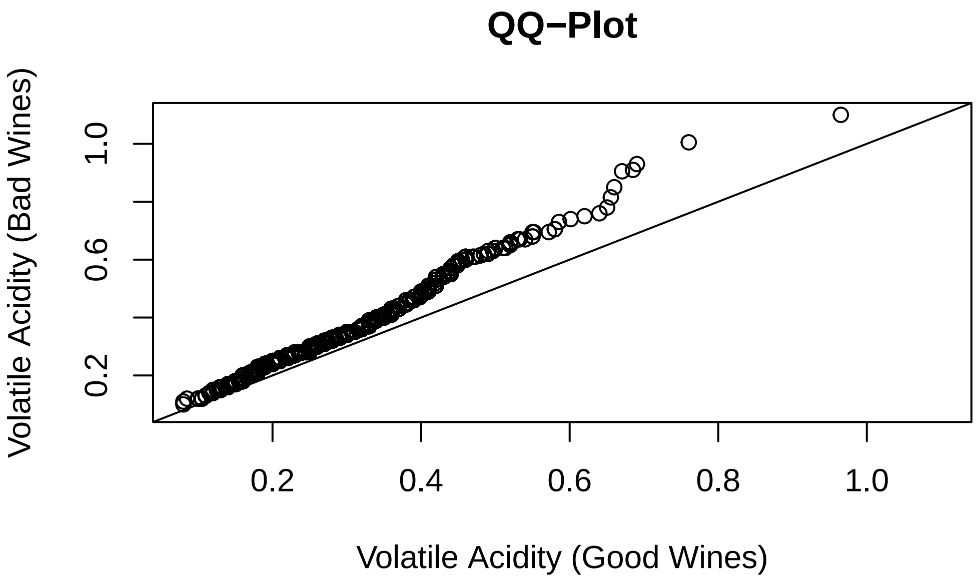
```
# Creating a histogram
ggplot(working.df) +
  aes(x = volatile.acidity, fill = factor(good)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 30) +
  labs(title = "Histogram of Volalite Acidity",
       x = "Volatile Acidity",
       y = "Frequency",
       fill = "Good Wine")
```



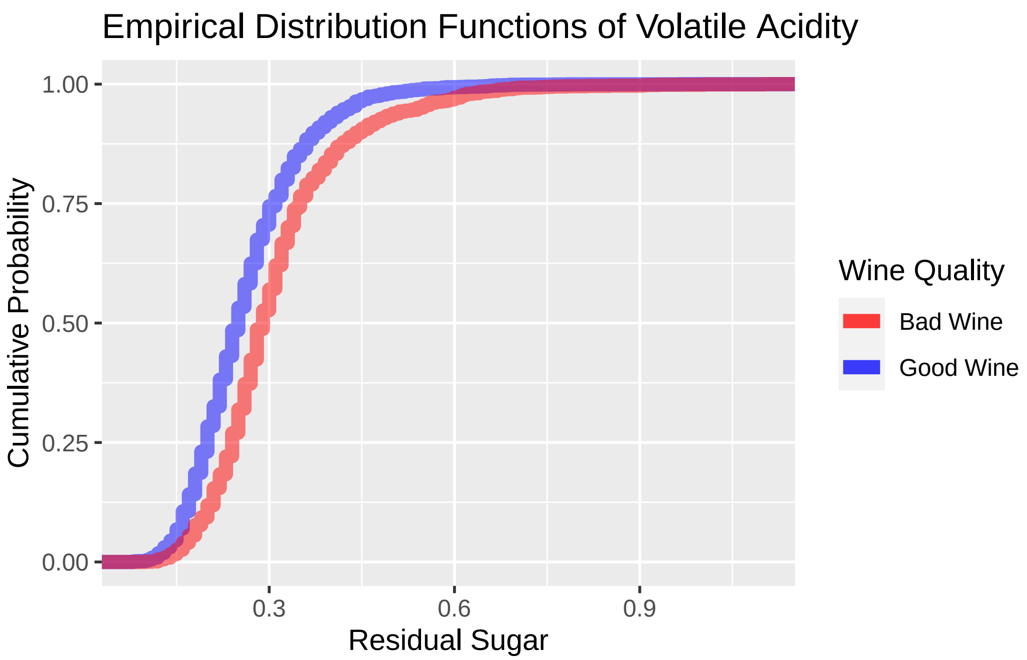
```r
# Creating box plots
boxplot <- ggplot(working.df) +
  aes(x = factor(good), y = volatile.acidity) +
  geom_boxplot() +
  labs(title = "Boxplot of Volatile Acidity by Wine Quality",
       x = "Wine Quality",
       y = "Volatile Acidity")

boxplot
```



```r
# Creating a QQ plot
volatile.acidity.good <- working.df %>%
  filter(good == 1) %>%
  pull(volatile.acidity)
volatile.acidity.bad <- working.df %>%
  filter(good == 0) %>%
  pull(volatile.acidity)

xlim.range <- range(volatile.acidity.good, volatile.acidity.bad)
ylim.range <- range(volatile.acidity.good, volatile.acidity.bad)

qqplot(volatile.acidity.good, volatile.acidity.bad,
       xlab = "Volatile Acidity (Good Wines)",
       ylab = "Volatile Acidity (Bad Wines)",
       main = "QQ-Plot",
       xlim = xlim.range,
       ylim = ylim.range)
abline(0, 1)
```



```r
# Creating an ECDF plot
working.good <- working.df %>% filter(good == 1)
working.bad <- working.df %>% filter(good == 0)

ggplot() +
  stat_ecdf(data = working.good,
            aes(x = volatile.acidity, color = "Good Wine"),
            geom = "step",
            linewidth = 2.5,
            alpha = 0.5) +
  stat_ecdf(data = working.bad,
            aes(x = volatile.acidity, color = "Bad Wine"),
            geom = "step",
            linewidth = 2.5, alpha = 0.5) +
  scale_color_manual(values = c("Good Wine" = "blue", "Bad Wine" = "red")) +
  labs(title = "Empirical Distribution Functions of Volatile Acidity",
       x = "Residual Sugar",
       y = "Cumulative Probability",
       color = "Wine Quality")
```



The summary statistics suggest that good wines have lower `volatile acidity` on average, with less variability and fewer extreme values compared to bad wines. This reflects in the histograms, boxplots, QQ plots, and ECDFs as follows:

- Histograms: Good wines have a sharper peak and a little bit of a narrower spread.
- Boxplots: Good wines have lower medians and a somewhat lesser spread with fewer "extreme" outliers compared to bad wines.
- QQ Plots: Good wines shows consistently lower values compared to bad wines for the same quantiles resulting in all points being above the diagonal.
- ECDFs: Good wines' ECDF start to rise at a lower value.

Lower `volatile acidity` seems to be associated with higher quality wines, as could have been expected.