White wine quality

```
library(tidyverse) # to be able to use data visual.tools
library(foreign) #reading data in diff.formats
library(ggplot2) # plot
library(ggrepel) #ggrepel provides geoms for ggplot2
```

Descriptive Measures

The data set **winequality-white.csv** is data on the white wine quality. In our analysis we will only consider the following variables: **volatile.acidity:** Volatile acidity, **residual.sugar:** Residual sugar, **pH:** pH level and **quality:** Wine quality in a score between 0 and 10.

Read and modify data

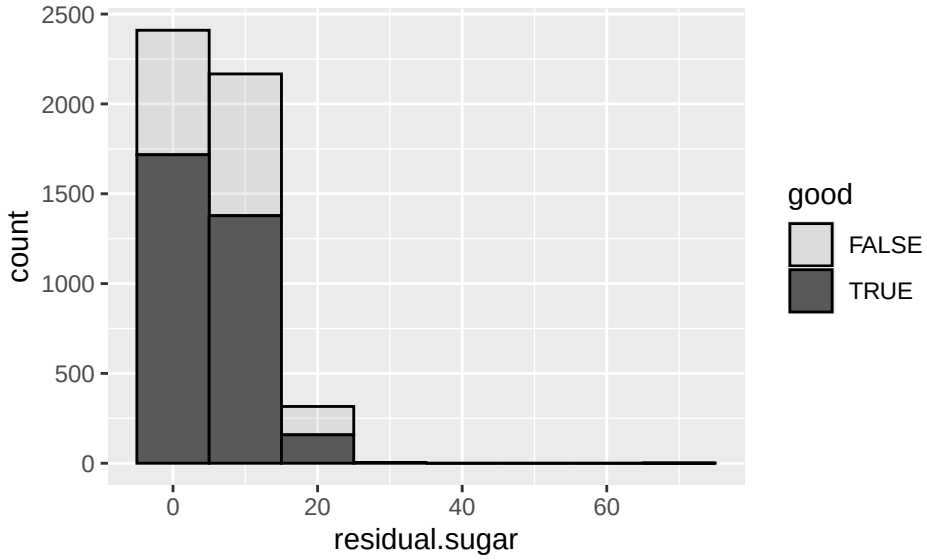
We will read the data and add to the df a new boolean variable **good** which is **TRUE** for values > 5:

```
winequality <- read.csv("winequality-white.csv", sep = ";")

#define new df with necessary variables
winequality.white <- winequality[,c("volatile.acidity", "residual.sugar", "pH", "quality")] %>%
  mutate(good = (quality > 5))#add new column with properties
```

Consider variable residual.sugar

We will plot frequency histograms of **residual.sugar** for good and bad wines in one plot, where we determine the binwidth with Freedman-Diaconis' rule.

```
#plot histograms
ggplot(winequality.white, aes(x = residual.sugar, alpha = good)) +
  geom_histogram(binwidth = 10, color = "black")
```



It looks like there are a couple of outliers here, but it's hard to tell which group they belong to.

Now, we calculate the summary statistics for both wine groups:

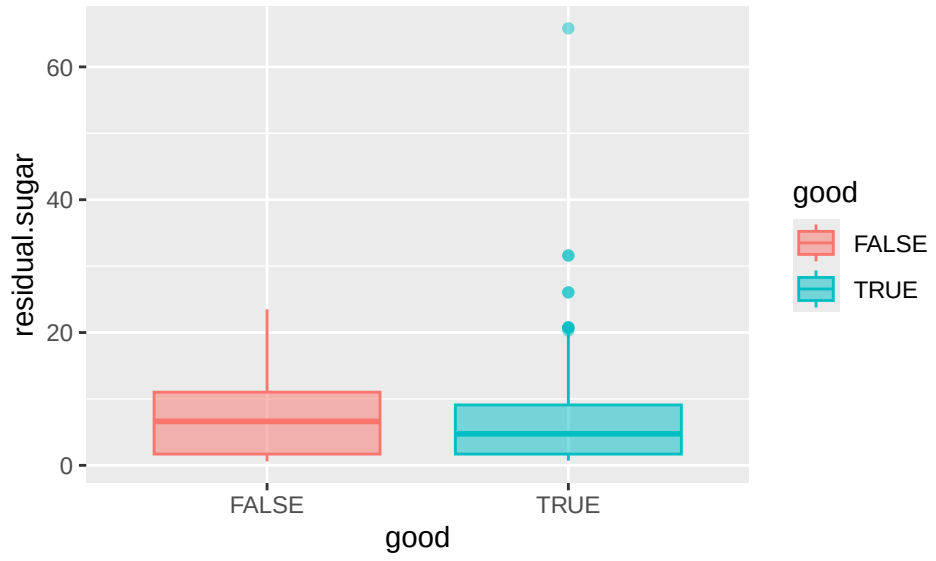
```
#define the matrix for summary statistics
t <- matrix(nrow = 6, ncol = 2)#number of sum.stat elem.

for (i in 1:2) {
  d <- filter(winequality.white, good == (i-1))$residual.sugar
  t[1, i] <- mean(d)
  t[2, i] <- median(d)
  t[3, i] <- sd(d)
  t[4, i] <- IQR(d)
  t[5, i] <- min(d)
  t[6, i] <- max(d)
}

rownames(t) <- c("mean", "median", "sd", "IQR", "min", "max")
colnames(t) <- c("bad", "good")

t
```

```
##              bad      good
## mean    7.054451  6.057658
## median  6.625000  4.750000
## sd      5.283594  4.929353
## IQR     9.325000  7.400000
## min     0.600000  0.700000
## max    23.500000 65.800000
```

Most summary statistics do not differ all that much between groups. The maximum value for residual sugar, however, is much larger for good wines.

We generate the boxplots:

```
ggplot(winequality.white, aes(x = good, y = residual.sugar, color = good, fill = good)) +
  geom_boxplot(alpha = 0.5)
```



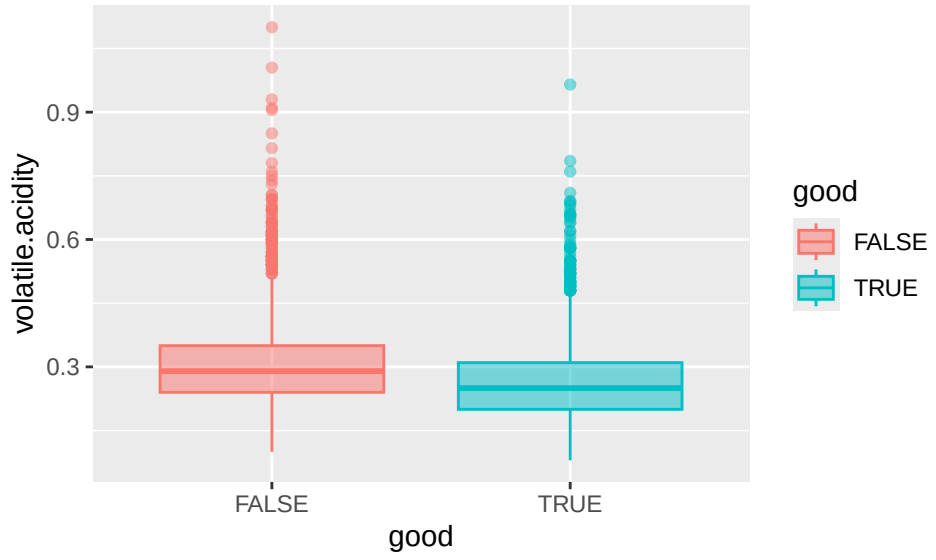
The boxplot confirms what we have seen earlier: There are outliers in the good wines, but none in the bad wines.

Consider volatile.acidity for good and bad wines.

We will use boxplots, histograms, QQ-plots, summary statistics and empirical distribution functions to compare this variable for good and bad wines, where we use the same code from above:

Boxplots:

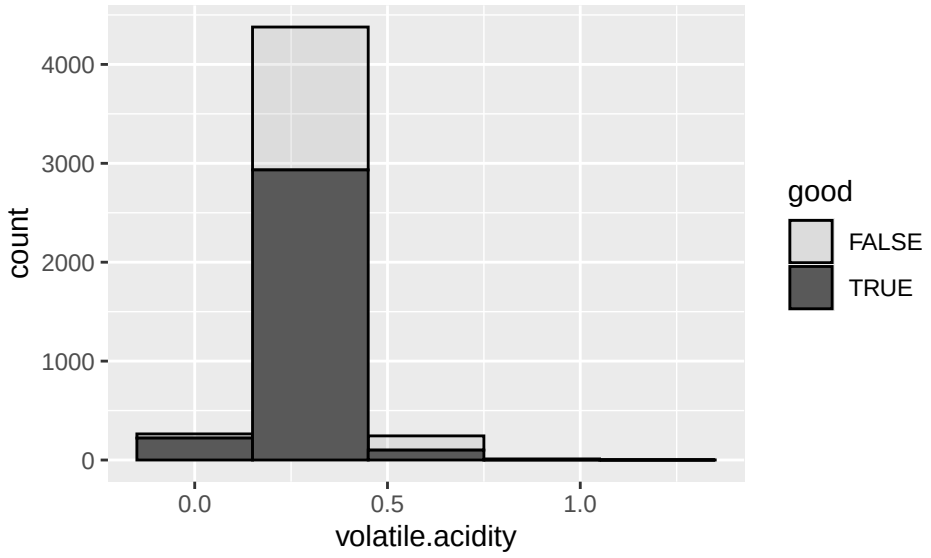
```
ggplot(winequality.white, aes(x = good, y = volatile.acidity, color = good, fill = good)) +
  geom_boxplot(alpha = 0.5)
```



The “boxes” are not so far apart here, but there are many outliers here in both groups.

Histogram:

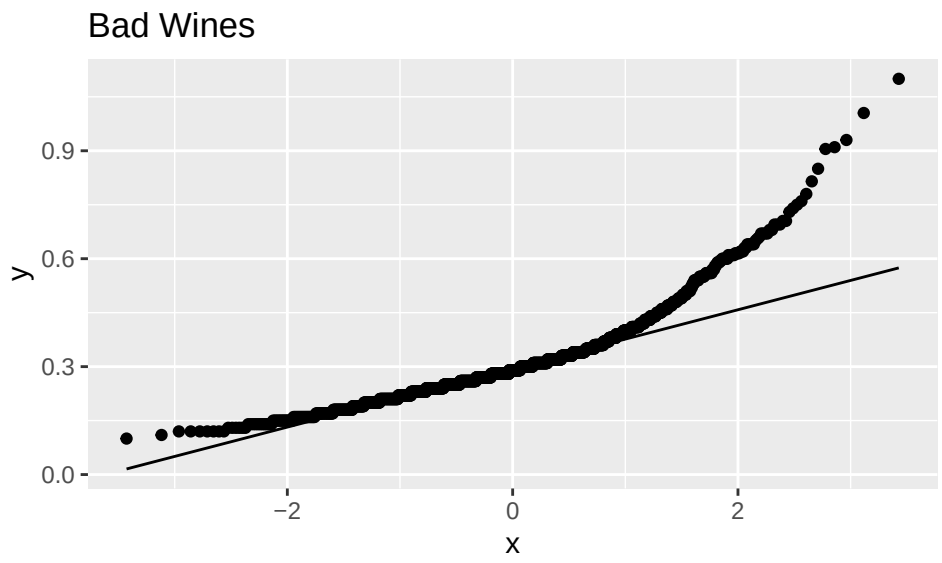
```
ggplot(winequality.white, aes(x = volatile.acidity, alpha = good)) +
  geom_histogram(binwidth = 0.3, color = "black")
```



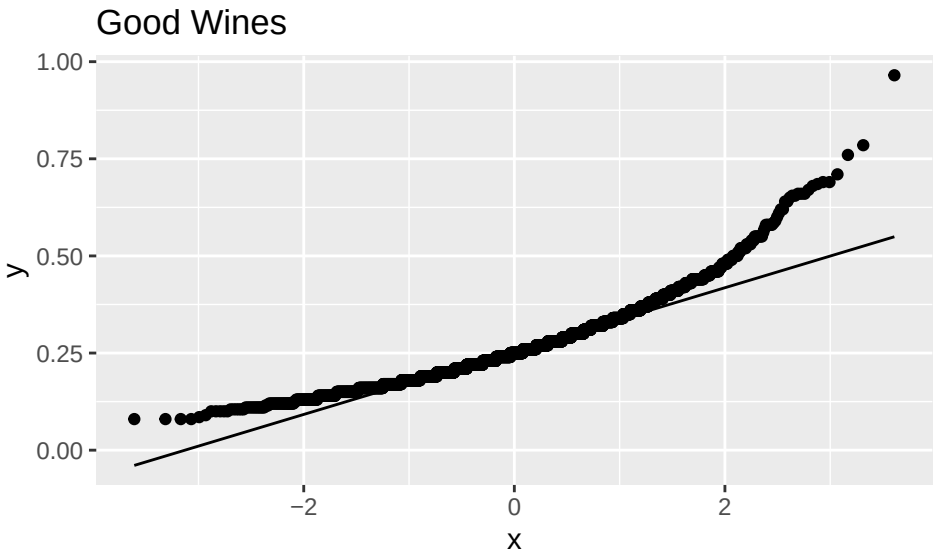
Most values are somewhere between 0 and 0.5, and as we have already seen, there are some outliers too.

QQ-Plots:

```
ggplot(filter(winequality.white, good == 0), aes(sample = volatile.acidity)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Bad Wines")
```



```
ggplot(filter(winequality.white, good == 1), aes(sample = volatile.acidity)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Good Wines")
```



We observe, that neither plot alludes to a normal distribution!

Summary statistics:

```
z <- matrix(nrow = 6, ncol = 2)

for (i in 1:2) {
  dz <- filter(winequality.white, good == (i-1))$volatile.acidity
  z[1, i] <- mean(dz)
  z[2, i] <- median(dz)
  z[3, i] <- sd(dz)
  z[4, i] <- IQR(dz)
  z[5, i] <- min(dz)
  z[6, i] <- max(dz)
}

rownames(z) <- c("mean", "median", "sd", "IQR", "min", "max")
colnames(z) <- c("bad", "good")

z
```

```
##              bad      good
## mean   0.3102652 0.2621209
## median 0.2900000 0.2500000
## sd     0.1125479 0.0901360
## IQR    0.1100000 0.1100000
## min    0.1000000 0.0800000
## max    1.1000000 0.9650000
```

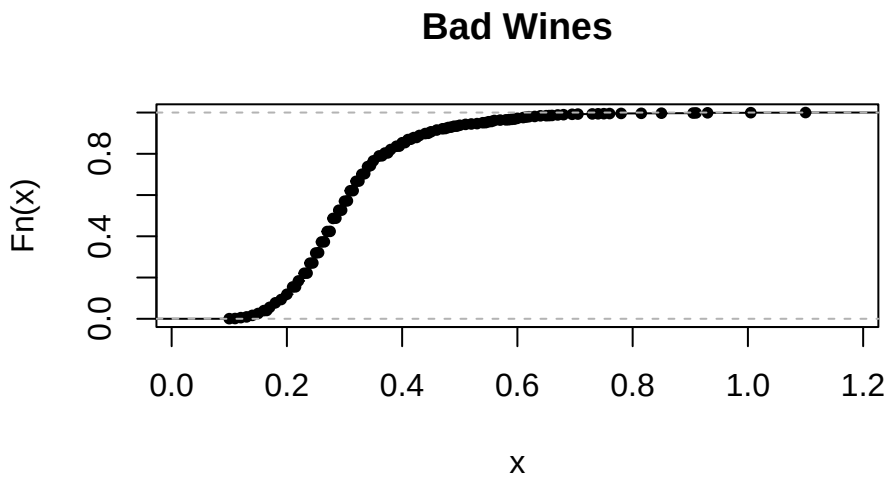
We observe that most summary statistics do not differ all that much between groups.

Empirical distributions functions:

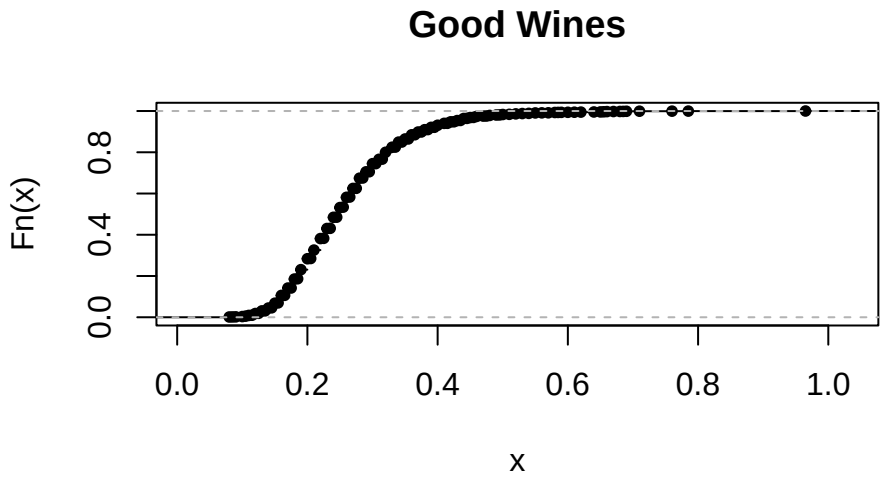
```
x1 <- filter(winequality.white, good == 0)
x2 <- filter(winequality.white, good == 1)

"Empirical Cumulative Distribution Function"
```

```
## [1] "Empirical Cumulative Distribution Function"
```

```
plot(ecdf(x1$volatile.acidity), main = "Bad Wines", pch = 20)
```




```
plot(ecdf(x2$volatile.acidity), main = "Good Wines", pch = 20)
```



All in all, it seems that good wines and bad wines are rather similar when it comes to volatile acidity.