Exercies

Natali Tckvitishvili

2024-05-19

Applied Statistics in R

Natali Tckvitishvili

Load libraries

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

Exercise 1

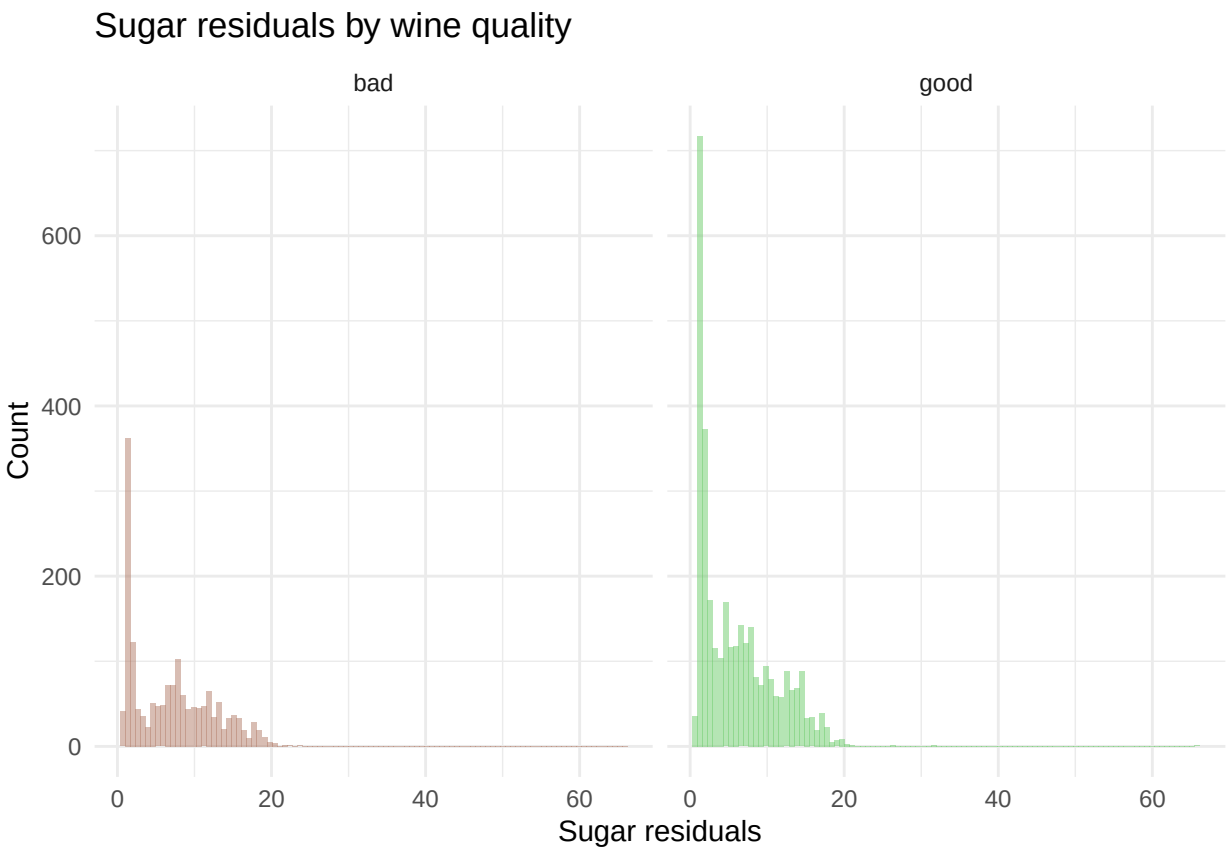
a. load data & add new variable - good

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

wine <- mutate(wine, good = ifelse(quality > 5, 1, 0))
```

b. residual.sugar analysis

- histograms for good and bad quality wines

```
good_labels <- c("0" = "bad", "1" = "good")

ggplot(wine, aes(x = residual.sugar, fill = as.factor(good))) +
  geom_histogram(position = "identity", alpha = .5, bins = 100) +
  scale_fill_manual(values=c("#b37d69", "#6dcc6b")) +
  facet_wrap(vars(good), labeller=labeller(good = good_labels)) +
  theme_minimal() +
  theme(legend.position="none") +
  labs (
    x = "Sugar residuals",
    y = "Count",
    title = "Sugar residuals by wine quality"
  )
```



According to the graphs, both bad and good quality wines have sugar residuals near zero, however, for good wines this number is higher than for bad ones. Moreover, sugar residuals of good quality wines have a smoother decrease in frequency, most of them have less sugar. Therefore, we can assume that sugar residuals may have negative correlation with wine quality.

- summary statistics

```
summary <- wine %>%
  group_by(good) %>%
  summarise(
    n = n(),
    mean = mean(residual.sugar),
    median = median(residual.sugar),
    sd = sd(residual.sugar),
    iqr = IQR(residual.sugar),
    max = max(residual.sugar),
    min = min(residual.sugar)
  )
data.frame(summary)
```

```
##   good    n     mean median       sd   iqr  max min
## 1    0 1640 7.054451  6.625 5.283594 9.325 23.5 0.6
## 2    1 3258 6.057658  4.750 4.929353 7.400 65.8 0.7
```