# BIOST 546 HW 1

My-Anh Doan

2023-01-09

Q1. In this problem, we will make use of the dataset `Medical_Cost_2.RData`.

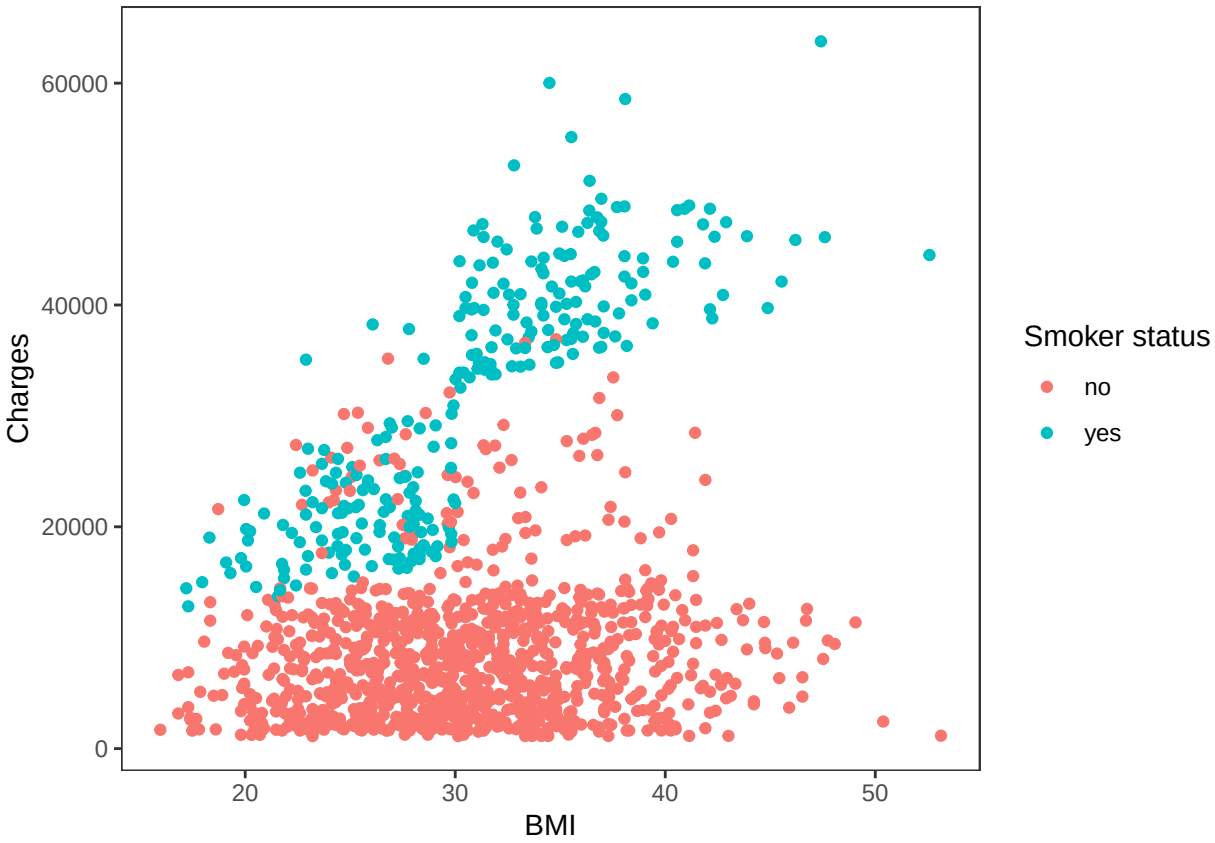
(a) Load the dataset with the command `load` and check if there are missing data.

(b) If any, remove the missing data using the command `na.omit`.

```
# load data to environment
load("Medical_Cost_2.RData") # dataframe with 1338 rows and 7 variables

# check for missing data; if returned value is non-zero, there are missing data
sum(is.na(df) == TRUE) # 60
## [1] 60

# remove missing data
df <- na.omit(df) # dataframe with 1278 rows and 7 variables
```

(c) We decide to focus on the outcome variable `charges` (individual medical costs billed by health insurance) and the predictors `bmi` (body mass index) and `smoker` (whether the subjects is a smoker or not). Make a scatterplot with `bmi` on the x-axis, `charges` on the y-axis, and with the color of each dot representing whether the subject is a smoker or not.

```
library(ggplot2)

ggplot(df, aes(x = bmi, y = charges)) +
  geom_point(aes(color = smoker)) +
  labs(x = "BMI",
       y = "Charges",
       color = "Smoker status") +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank())
```


60000
40000
20000
0
Charges
20
30
40
50
BMI
Smoker status
no
yes