# Data Science 2 Final

Group 6

## Contents

# Set up

## Load libraries and data

```
library(caret)
library(mgcv)
library(earth)
library(tidyverse)
library(summarytools)
library(corrplot)
library(ggpubr)

setwd("D:/CUMC/Y2S2/DS2/Final/ds2_final")

load("./recovery.RData")
```

## Subset 2 df and keep unique observations

```
set.seed(2543)

dat1 <- dat[sample(1:10000, 2000),]

set.seed(4017)

dat2 <- dat[sample(1:10000, 2000),]

dat_bind <- unique(rbind(dat1, dat2))
```

# Exploratory analysis and data visualization

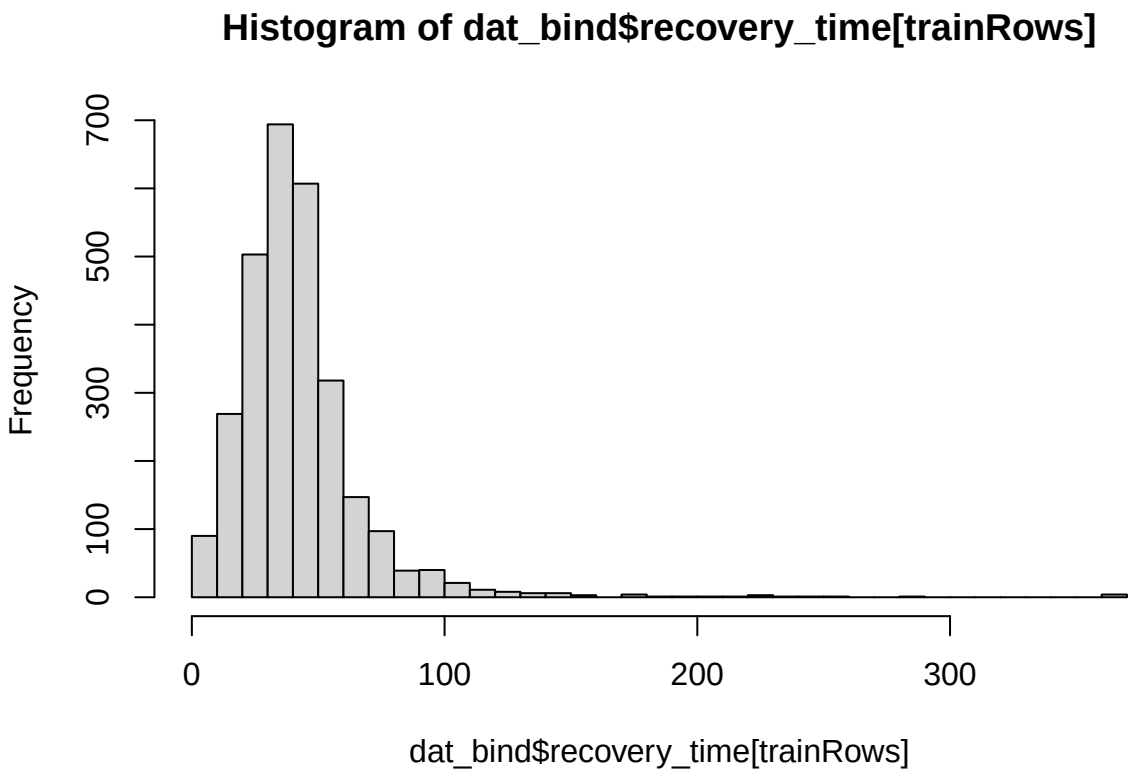
## Data Partition

Here, we mainly want to investigate the EDA of the training dataset. Therefore, we will start with the data partition.

```
set.seed(2460)

trainRows <- createDataPartition(y = dat_bind$recovery_time,
                                 p = 0.8, list = FALSE)
```

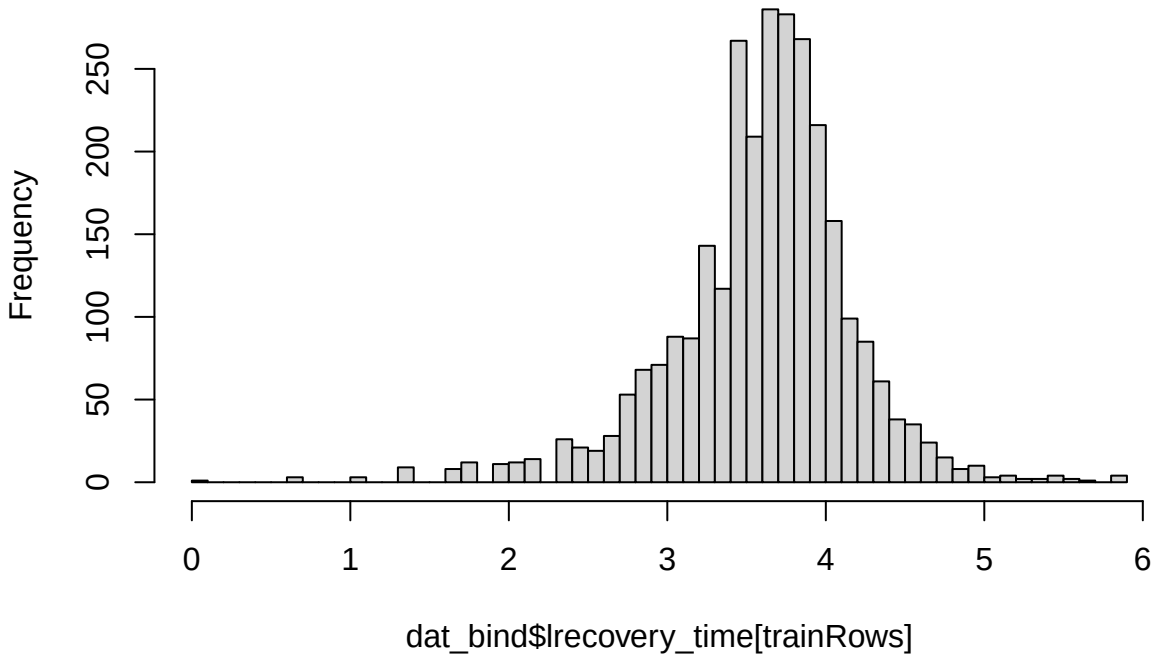
## Understanding the outcome variable `recovery_time`

```
# check the outcome variable
hist(dat_bind$recovery_time[trainRows], breaks = 50)
```




The distribution of the outcome variable `recovery_time` is heavily right-skewed. To account for this, I will take the log-transformation of the outcome and use that variable for following analyses.

```
dat_bind = dat_bind %>%
  na.omit(dat_bind) %>%
  mutate(lrecovery_time = log(recovery_time)) %>%
  select(-recovery_time, -id)

# log-transformation helped with making it more normal
hist(dat_bind$lrecovery_time[trainRows], breaks = 50)
```




### Summary of the dataset

```
st_options(plain.ascii = F,
           style = "rmarkdown",
           dfSummary.silent = T,
           footnote = NA,
           subtitle.emphasis = F)

dfSummary(dat_bind[trainRows, -1])
```

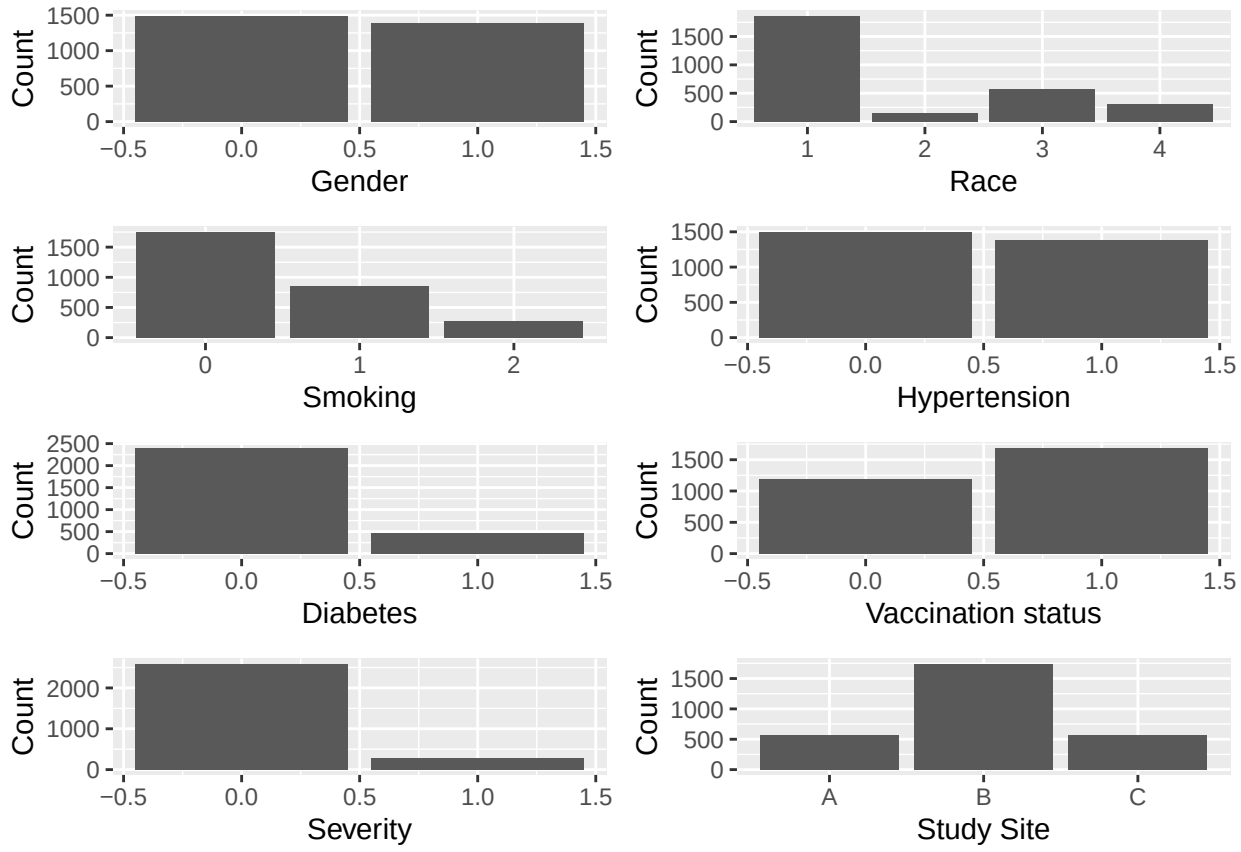
```
## ### Data Frame Summary
## **dat_bind**
## **Dimensions:** 2878 x 14
## **Duplicates:** 0
##
## ------------------------------------------------------------------------------------------------------
## No   Variable          Stats / Values              Freqs (% of Valid)    Graph
## ---- ----------------- --------------------------- --------------------- ----------------------------
## 1    gender\           Min  : 0\                   0 : 1490 (51.8%)\     IIIIIIIIII \
##      [integer]         Mean : 0.5\                 1 : 1388 (48.2%)      IIIIIIIII
##                        Max  : 1
##
## 2    race\             1\. 1\                      1863 (64.7%)\         IIIIIIIIIII \
```

```
##      [factor]          2\. 2\                    145 ( 5.0%)\          I \
##                        3\. 3\                    569 (19.8%)\          III \
##                        4\. 4                     301 (10.5%)           II
##
## 3    smoking\          1\. 0\                    1753 (60.9%)\         IIIIIIIIIII \
##      [factor]          2\. 1\                    846 (29.4%)\          IIIII \
##                        3\. 2                     279 ( 9.7%)           I
##
## 4    height\           Mean (sd) : 170.1 (5.9)\  311 distinct values   \ \ \ \ \ \ . :\
##      [numeric]         min < med < max:\                               \ \ \ \ \ \ : :\
##                        150.7 < 170.4 < 190.6\                          \ \ \ \ . : : .\
##                        IQR (CV) : 7.9 (0)                              \ \ \ \ : : : :\
##                                                                        \ \ . : : : : : .
##
## 5    weight\           Mean (sd) : 79.9 (7)\     358 distinct values   \ \ \ \ \ \ \ \ :\
##      [numeric]         min < med < max:\                               \ \ \ \ \ \ : : .\
##                        55.9 < 80.1 < 111.6\                            \ \ \ \ \ \ : : :\
##                        IQR (CV) : 9.5 (0.1)                            \ \ \ \ : : : :\
##                                                                        \ \ . : : : : : :
##
## 6    bmi\              Mean (sd) : 27.7 (2.7)\   162 distinct values   \ \ \ \ \ \ . :\
##      [numeric]         min < med < max:\                               \ \ \ \ \ \ : :\
##                        19.7 < 27.5 < 38.1\                             \ \ \ \ . : : :\
##                        IQR (CV) : 3.6 (0.1)                            \ \ \ \ : : : : .\
##                                                                        \ \ : : : : : : : .
##
## 7    hypertension\     Min  : 0\                 0 : 1499 (52.1%)\     IIIIIIIIII \
##      [numeric]         Mean : 0.5\               1 : 1379 (47.9%)      IIIIIIIII
##                        Max  : 1
##
## 8    diabetes\         Min  : 0\                 0 : 2403 (83.5%)\     IIIIIIIIIIIIIII \
##      [integer]         Mean : 0.2\               1 :  475 (16.5%)      III
##                        Max  : 1
##
## 9    SBP\              Mean (sd) : 130.2 (8)\    51 distinct values    \ \ \ \ \ \ \ \ : .\
##      [numeric]         min < med < max:\                               \ \ \ \ \ \ . : :\
##                        106 < 130 < 157\                                \ \ \ \ \ \ : : : :\
##                        IQR (CV) : 11 (0.1)                             \ \ \ \ : : : : :\
##                                                                        \ \ : : : : : : : .
##
## 10   LDL\              Mean (sd) : 110.6 (19.9)\ 121 distinct values   \ \ \ \ \ \ \ \ \ :\
##      [numeric]         min < med < max:\                               \ \ \ \ \ \ \ \ : : .\
##                        28 < 111 < 178\                                 \ \ \ \ \ \ \ \ : : :\
##                        IQR (CV) : 27 (0.2)                             \ \ \ \ \ \ . : : : :\
##                                                                        \ \ \ \ . : : : : : .
##
## 11   vaccine\          Min  : 0\                 0 : 1189 (41.3%)\     IIIIIIII \
##      [integer]         Mean : 0.6\               1 : 1689 (58.7%)      IIIIIIIIII
##                        Max  : 1
##
## 12   severity\         Min  : 0\                 0 : 2589 (90.0%)\     IIIIIIIIIIIIIIII \
##      [integer]         Mean : 0.1\               1 :  289 (10.0%)      II
##                        Max  : 1
##
```

```
## 13   study\            1\. A\                    571 (19.8%)\          III \
##      [character]       2\. B\                    1741 (60.5%)\         IIIIIIIIIII \
##                        3\. C                     566 (19.7%)           III
##
## 14   lrecovery_time\   Mean (sd) : 3.6 (0.6)\    147 distinct values   \ \ \ \ \ \ \ \ \ \ \ \ :\
##      [numeric]         min < med < max:\                               \ \ \ \ \ \ \ \ \ \ \ \ :\
##                        0 < 3.7 < 5.9\                                  \ \ \ \ \ \ \ \ \ \ : :\
##                        IQR (CV) : 0.6 (0.2)                            \ \ \ \ \ \ \ \ \ \ : :\
##                                                                        \ \ \ \ \ \ \ \ : : : :
## ---------------------------------------------------------------------------------------------------
```

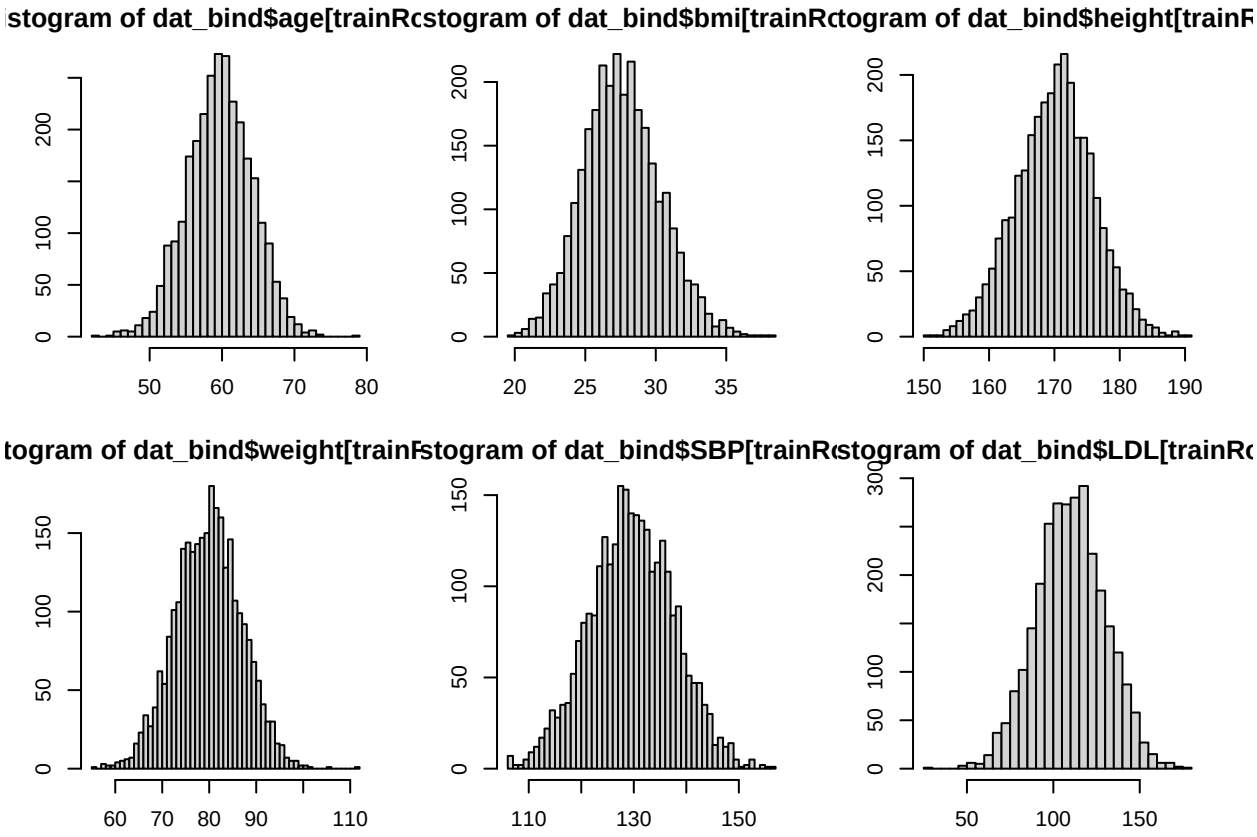
### Understand categorical variables

```
gender = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = gender)) +  geom_bar() +  labs(x = "Gender", y = "Count")

race = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = race)) + geom_bar() + labs(x = "Race",y = "Count")

smoking = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = smoking)) + geom_bar() + labs(x = "Smoking", y = "Count")

hypertension = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = hypertension)) + geom_bar() + labs(x = "Hypertension",
                                                    y = "Count")
diabetes = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = diabetes)) + geom_bar() + labs(x = "Diabetes",y = "Count")

vaccine = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = vaccine)) + geom_bar() + labs(x = "Vaccination status",
                                               y = "Count")
severity = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = severity)) + geom_bar() + labs(x = "Severity", y = "Count")

study = (dat_bind[trainRows, -1]) %>%
  ggplot(aes(x = study)) + geom_bar() + labs(x = "Study Site", y = "Count")

cat_combined_plot = ggarrange(gender, race, smoking, hypertension,
                              diabetes, vaccine,severity, study,
                          ncol = 2, nrow = 4)

cat_combined_plot
```



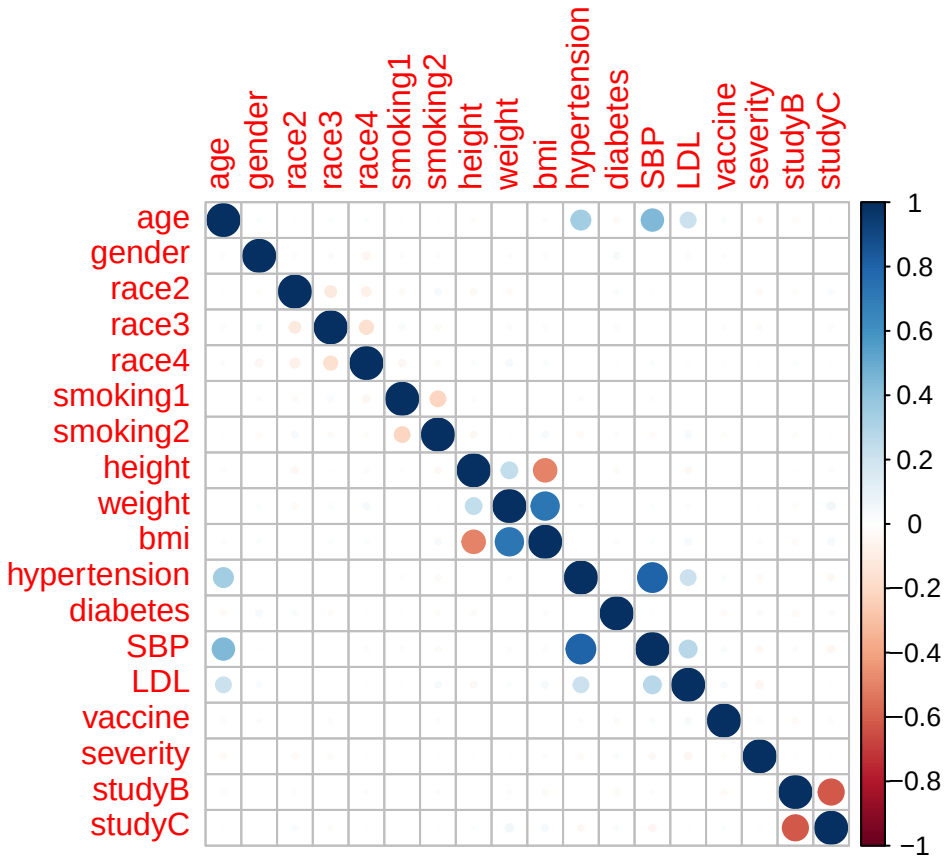

### Understand continuous variables

```
par(mar = c(3, 3, 2, 2), mfrow = c(2, 3))

age = hist(dat_bind$age[trainRows], breaks = 50)

bmi = hist(dat_bind$bmi[trainRows], breaks = 50)

height = hist(dat_bind$height[trainRows], breaks = 50)

weight = hist(dat_bind$weight[trainRows], breaks = 50)

SBP = hist(dat_bind$SBP[trainRows], breaks = 50)

LDL = hist(dat_bind$LDL[trainRows], breaks = 50)
```



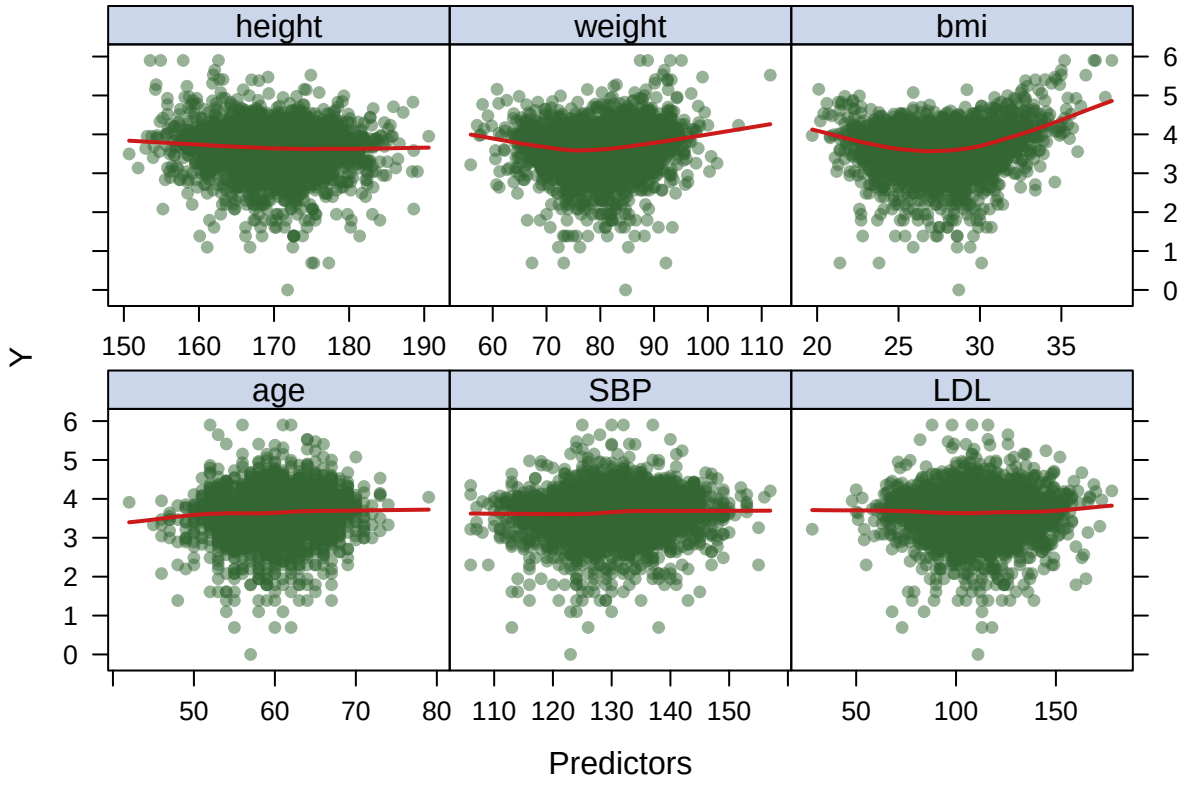

## Understand the correlation between continuous predictors ****

```
correlation <- model.matrix(lrecovery_time ~ ., dat_bind)[trainRows,-1]

corrplot(cor(correlation), method = "circle", type = "full")
```




**Understand the relationship with continuous predictors and the outcome**

```
theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.2, .4, .2, .5)
theme1$plot.symbol$pch <- 16
theme1$plot.line$col <- rgb(.8, .1, .1, 1)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <- rgb(.0, .2, .6, .2)
trellis.par.set(theme1)

# plotting continuous predictors
featurePlot(x = model.matrix(lrecovery_time ~ ., dat_bind)[trainRows,c("age", "SBP", "LDL", "height", "
            y = dat_bind$lrecovery_time[trainRows],
            plot = "scatter",
            span = .5,
            labels = c("Predictors","Y"),
            type = c("p", "smooth"),
            layout = c(3,2))
```




## Considering variables based on the EDA

From the correlation plot, we can observe that `bmi` is highly correlated with `weight` and `height`, which makes sense because BMI is calculated by weight divided by the square of height. This demonstrates collinearity between the variables, and to account for this, I will remove the `bmi` variable for the predictions.

Also, I believe that the `study` variable is more of a geographical indicator to distinguish different study sites, and it will not be critical in predicting recovery time. Therefore, I will also remove the `study` variable.

Lastly, I will remove variables `race` and `smoking` since I have created dummy variables for them and I will use the dummy variables in further analyses.

```
final = dat_bind %>%
  mutate(
  # create dummy variables for categorical variables
    # set up 3 dummy variables for `race`, reference = White:
         race_2 = ifelse(race == 2, 1, 0),
         race_3 = ifelse(race == 3, 1, 0),
         race_4 = ifelse(race == 4, 1, 0),
    # set up 2 dummy variables for `smoking`, reference = Never smoked:
         smoking_1 = ifelse(smoking == 1, 1, 0),
         smoking_2 = ifelse(smoking == 2, 1, 0))

# remove variables that will not be used
final = final %>%
  select(-bmi, -study, -race, -smoking)
```

```r
# partition again based on the new outcome variable
set.seed(2460)

trainRows_new <- createDataPartition(y = final$lrecovery_time, p = 0.8, list = FALSE)

x <- model.matrix(lrecovery_time ~ ., final)[trainRows_new,-1]

y <- final$lrecovery_time[trainRows_new]

x2 <- model.matrix(lrecovery_time ~ ., final)[-trainRows_new,-1]

y2 <- final$lrecovery_time[-trainRows_new]

ctrl1 <- trainControl(method = "repeatedcv", number = 10, repeats = 5)
```