Data Science II Midterm Project

Huanyu Chen

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

load("recovery.Rdata")
```

Exploratory Analysis and Data Visualization

Exploratory Analysis

In this dataset, `age`, `height`, `weight`, `bmi`, `SBP`, `LDL`, and `recovery_time` are continuous variables.

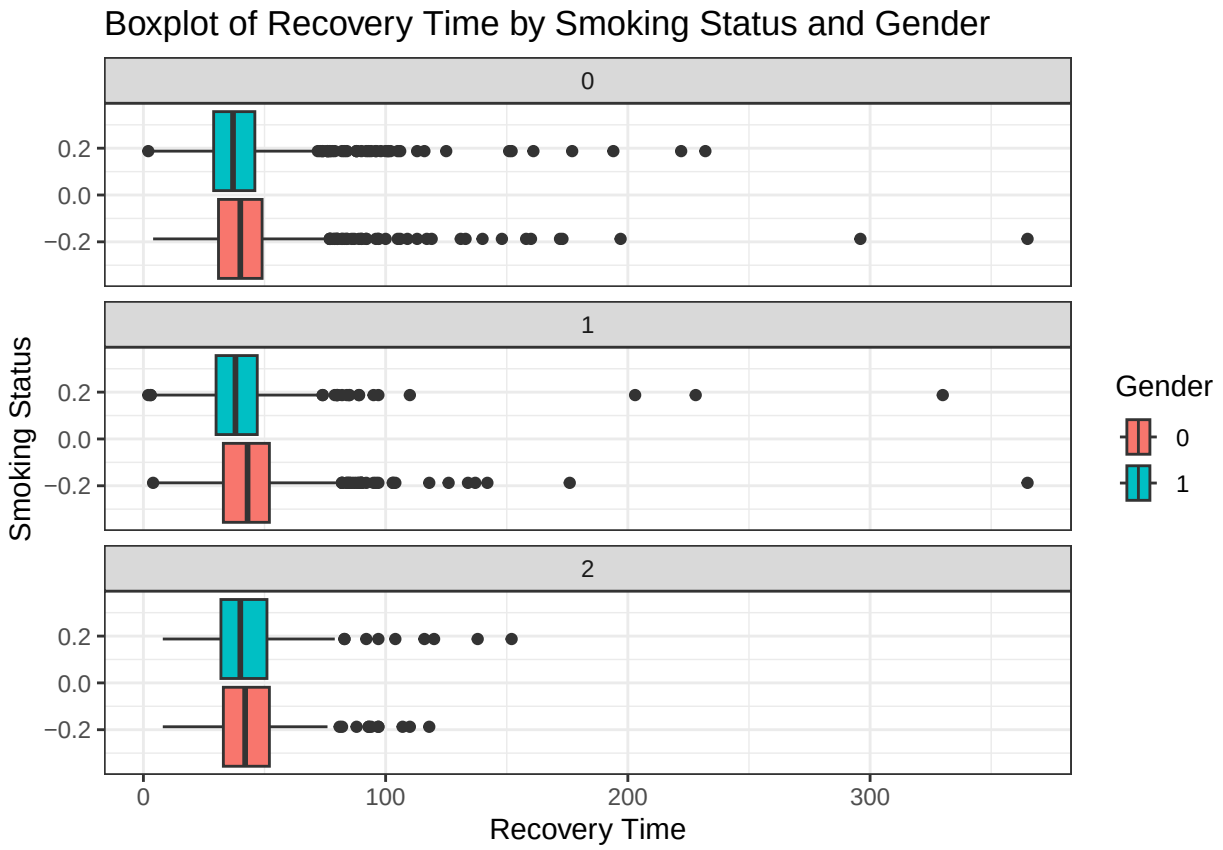
```
continuous_vars <- dat[, c("age", "height", "weight", "bmi",
                           "SBP", "LDL", "recovery_time")]
summary(continuous_vars)
```

```
##       age           height          weight           bmi
##  Min.   :42.0   Min.   :147.8   Min.   : 55.90   Min.   :18.80
##  1st Qu.:57.0   1st Qu.:166.0   1st Qu.: 75.20   1st Qu.:25.80
##  Median :60.0   Median :169.9   Median : 79.80   Median :27.65
##  Mean   :60.2   Mean   :169.9   Mean   : 79.96   Mean   :27.76
##  3rd Qu.:63.0   3rd Qu.:173.9   3rd Qu.: 84.80   3rd Qu.:29.50
##  Max.   :79.0   Max.   :188.6   Max.   :103.70   Max.   :38.90
##       SBP             LDL        recovery_time
##  Min.   :105.0   Min.   : 28.0   Min.   :  2.00
##  1st Qu.:125.0   1st Qu.: 97.0   1st Qu.: 31.00
##  Median :130.0   Median :110.0   Median : 39.00
##  Mean   :130.5   Mean   :110.5   Mean   : 42.17
##  3rd Qu.:136.0   3rd Qu.:124.0   3rd Qu.: 49.00
##  Max.   :156.0   Max.   :178.0   Max.   :365.00
```

Boxplot of Recovery Time by Smoking Status and Gender

Our analysis reveals a notable trend: across all smoking statuses, females (`gender` = 0) consistently exhibit longer recovery times compared to males. Interestingly, individuals who had never smoked had more outliers on the right side of the boxplot, suggesting a longer recovery time. This counter-intuitive finding suggests that individuals with healthier lifestyles, such as non-smokers, paradoxically require more time to recover from COVID-19.

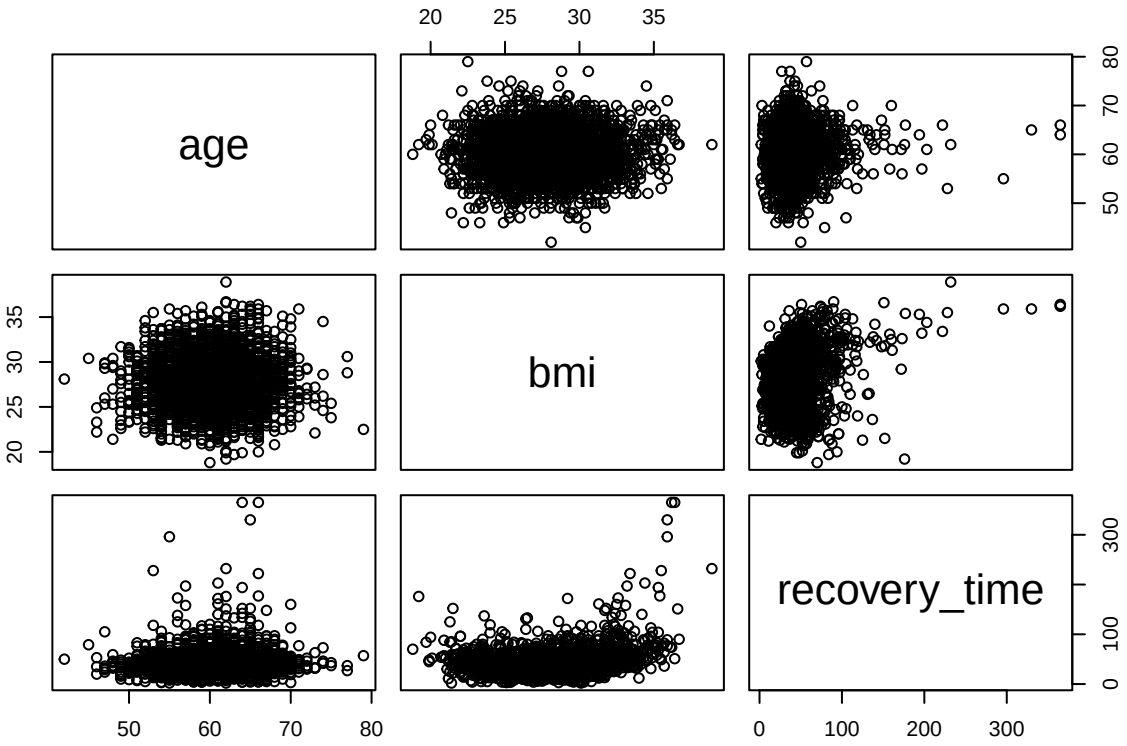
```
ggplot(dat, aes(x = recovery_time, fill = factor(gender))) +
  geom_boxplot() +
  labs(title = "Boxplot of Recovery Time by Smoking Status and Gender",
       x = "Recovery Time", y = "Smoking Status",
       fill = "Gender") +
  facet_wrap(~factor(smoking), ncol = 1) +
  theme_bw()
```



Pairs

Our exploration of the variables age, BMI, and recovery time reveals no clear linear relationships among them. It implies that other complex factors beyond these variables might be influencing the recovery time from COVID-19, highlighting the complexity of analysis about recovery time.

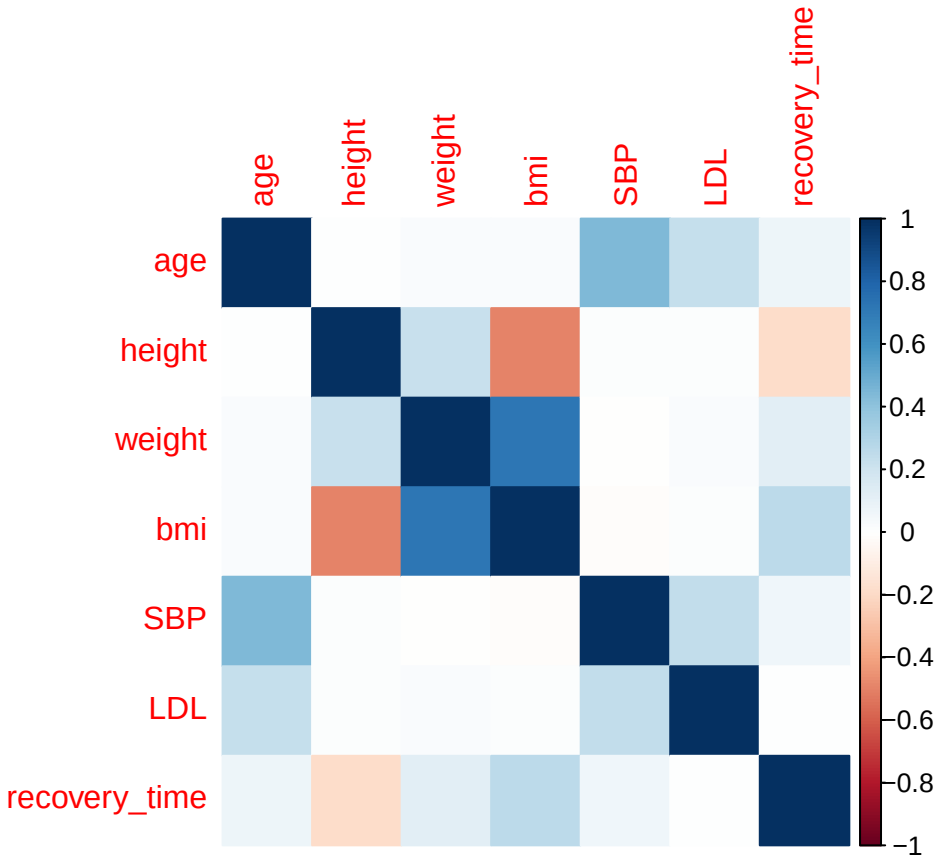
```
pairs(dat[, c("age", "bmi", "recovery_time")])
```



Correlation Table

The correlation analysis conducted on variables including "height," "weight," and "bmi" suggests a strong positive correlation among these attributes, which aligns with our common understanding. However, no significant correlations were observed between these attributes and other variables in the dataset.

```
correlation_matrix <- cor(dat[, c("age", "height", "weight", "bmi",
                                  "SBP", "LDL", "recovery_time")])
corrplot(correlation_matrix, method = "color")
```


age
height
weight
bmi
SBP
LDL
recovery_time
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Model Training

Results

Conclusions