P8106 Data ScienceII Midterm EDA

Yueran Zhang (yz4188)

2023-03-26

Loading in the Data

```
# Data Import
# Set as my last four digits of your UNI
set.seed(4188)

# Load objects into my workspace
setwd("/Users/yueranzhang/Desktop/DSII/Midterm/Dataset")
load(file = "recovery.RData")

# Generate a random sample of 2000 participants
dat <- dat[sample(1:10000, 2000),]
```

```
head(dat)
```

```
##        id age gender race smoking height weight  bmi hypertension diabetes SBP
## 2748 2748  59      0    1       1  167.4   75.1 26.8            1        0 137
## 7241 7241  64      0    3       1  172.5   73.5 24.7            1        0 135
## 5578 5578  52      1    4       0  179.7   82.3 25.5            1        0 137
## 8792 8792  62      0    1       2  168.2   81.6 28.8            0        0 129
## 2706 2706  65      0    1       1  173.3   81.4 27.1            1        0 134
## 8293 8293  63      0    1       0  167.1   72.6 26.0            0        0 128
##      LDL vaccine severity study recovery_time
## 2748  86       1        0     B            33
## 7241 132       0        0     B            54
## 5578 135       1        0     B            46
## 8792 104       0        0     C            49
## 2706 133       0        0     B            82
## 8293 147       1        0     C            27
```

Cleaning the Data

Though there seems to be many numeric variables, not all of them are true numerical variables. Some are displayed as numbers but are really factors. These variables will be converted from int to factor.

```
set.seed(4188)

to_be_factors <- c("gender", "hypertension", "diabetes", "vaccine", "severity", "study")
dat[to_be_factors] <- lapply(dat[to_be_factors], factor)
```

```
dat$gender <- recode(dat$gender, '1' = 'Male',
                                 '0' = "Female")

dat$race <- recode(dat$race, '1' = 'White',
                             '2' = "Asian",
                             '3' = 'Black',
                             '4' = 'Hispanic')

dat$smoking <- recode(dat$smoking, '0' = 'Never smoked',
                                   '1' = 'Former smoke',
                                   '2' = 'Current smoker')

dat$hypertension <- recode(dat$hypertension, '0' = 'No',
                                             '1' = 'Yes')

dat$diabetes <- recode(dat$diabetes, '0' = 'No',
                                     '1' = 'Yes')

dat$vaccine <- recode(dat$vaccine, '0' = 'Not vaccinated',
                                   '1' = 'Vaccinated')

dat$severity <- recode(dat$severity, '0' = "Not severe",
                                     '1' = 'Severe')


# count the missing values by column wise
print("Count of missing values by column wise")
```

```
## [1] "Count of missing values by column wise"
```

```
sapply(dat, function(x) sum(is.na(x)))
```

```
##            id           age        gender          race       smoking
##             0             0             0             0             0
##        height        weight           bmi  hypertension      diabetes
##             0             0             0             0             0
##           SBP           LDL       vaccine      severity         study
##             0             0             0             0             0
## recovery_time
##             0
```

```
str(dat)
```

```
## 'data.frame':    2000 obs. of  16 variables:
##  $ id           : int  2748 7241 5578 8792 2706 8293 1975 9748 6169 3139 ...
##  $ age          : num  59 64 52 62 65 63 58 58 62 59 ...
##  $ gender       : Factor w/ 2 levels "Female","Male": 1 1 2 1 1 1 2 2 2 1 ...
##  $ race         : Factor w/ 4 levels "White","Asian",..: 1 3 4 1 1 1 3 2 3 4 ...
##  $ smoking      : Factor w/ 3 levels "Never smoked",..: 2 2 1 3 2 1 1 1 1 3 ...
##  $ height       : num  167 172 180 168 173 ...
##  $ weight       : num  75.1 73.5 82.3 81.6 81.4 72.6 82.7 80.9 84.3 77.1 ...
```

```
##  $ bmi          : num  26.8 24.7 25.5 28.8 27.1 26 26.4 26.9 26 27.7 ...
##  $ hypertension : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 1 2 1 2 1 ...
##  $ diabetes     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SBP          : num  137 135 137 129 134 128 134 119 136 124 ...
##  $ LDL          : num  86 132 135 104 133 147 117 80 126 78 ...
##  $ vaccine      : Factor w/ 2 levels "Not vaccinated",..: 2 1 2 1 1 2 1 2 2 2 ...
##  $ severity     : Factor w/ 2 levels "Not severe","Severe": 1 1 1 1 1 1 1 1 1 1 ...
##  $ study        : Factor w/ 3 levels "A","B","C": 2 2 2 3 2 3 1 3 2 2 ...
##  $ recovery_time: num  33 54 46 49 82 27 37 35 31 40 ...
```

```
set.seed(4188)
dat <- subset(dat, select = c(3:16))
dim(dat)
```

```
## [1] 2000   14
```

```
cat(paste("Number of Numeric Variables: ", sum(sapply(dat, is.numeric))))
```

```
## Number of Numeric Variables:  6
```

```
cat(paste("\nNumber of Categorical Variables: ", sum(sapply(dat, is.factor))))
```

```
##
## Number of Categorical Variables:  8
```

Order and ID variables do not convey any useful information and is dropped. There are **2000 observations** and **14 variables**. 1 of the 14 variables include **recovery time**, which is the target variable.

Looking at the Target Feature

```
cat(paste("Mean recovery time: ", round(mean(dat$recovery_time))))
```

```
## Mean recovery time:  43
```

```
cat(paste("\nMedian recovery time: ", median(dat$recovery_time)))
```

```
##
## Median recovery time:  39
```

```
cat(paste("\nMax recovery time: ", max(dat$recovery_time)))
```

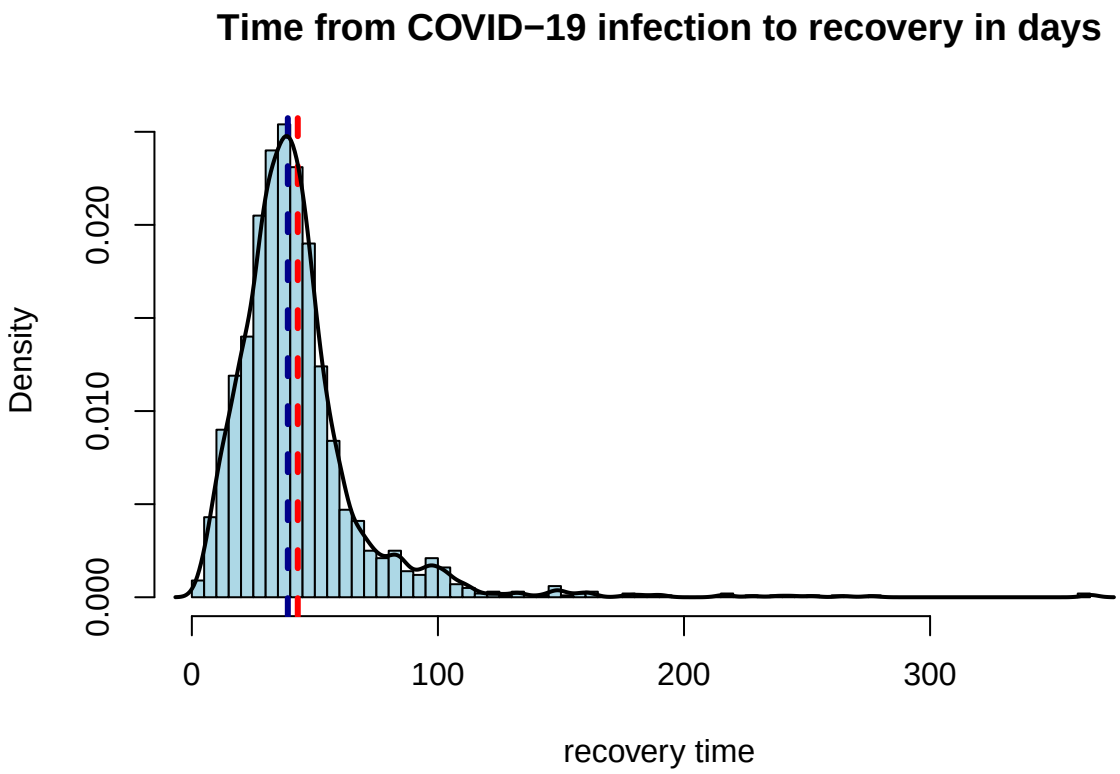
```
##
## Max recovery time:  365
```

```
cat(paste("\nMin recovery time: ", min(dat$recovery_time)))
```

```
##
## Min recovery time:  3
```

```
set.seed(4188)
options(repr.plot.height = 4.5, repr.plot.width = 8)
options(scipen=10000)

hist(dat$recovery_time, col = 'light blue', main = 'Time from COVID-19 infection to recovery in days', x
abline(v = mean(dat$recovery_time), col = 'red', lty = 2, lwd = 3)
abline(v = median(dat$recovery_time), col = 'dark blue', lty = 2, lwd = 3)
lines(density(dat$recovery_time), col = 'black', lwd = 2)
```



The histogram of the recovery time is a little bit of right skewed. The mean is higher than the median. There are also a good number of outliers.

Looking at Other Features

In order to predict the recovery time, let's look at how other variables influence recovery time.

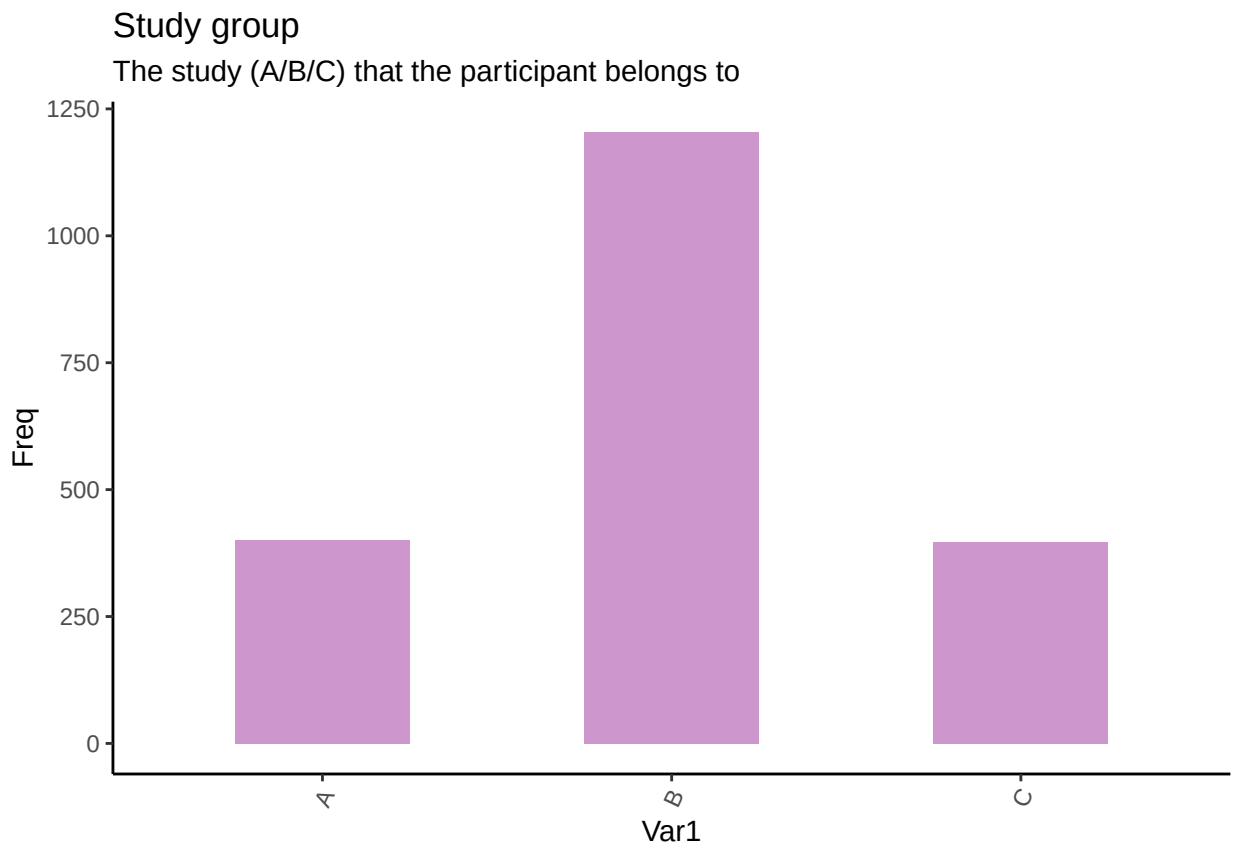
What study group were the participants from?

```
set.seed(4188)

table(dat$study)
```

```
##
##    A    B    C
##  400 1204  396
```

```
theme_set(theme_classic())
options(repr.plot.height = 4, repr.plot.width = 8)
ggplot(as.data.frame.table(table(dat$study)), aes(Var1, Freq)) +
    geom_bar(stat = "identity", width = 0.5, fill = "plum3") +
    labs(title = "Study group", subtitle = "The study (A/B/C) that the participant belongs to") +
    theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



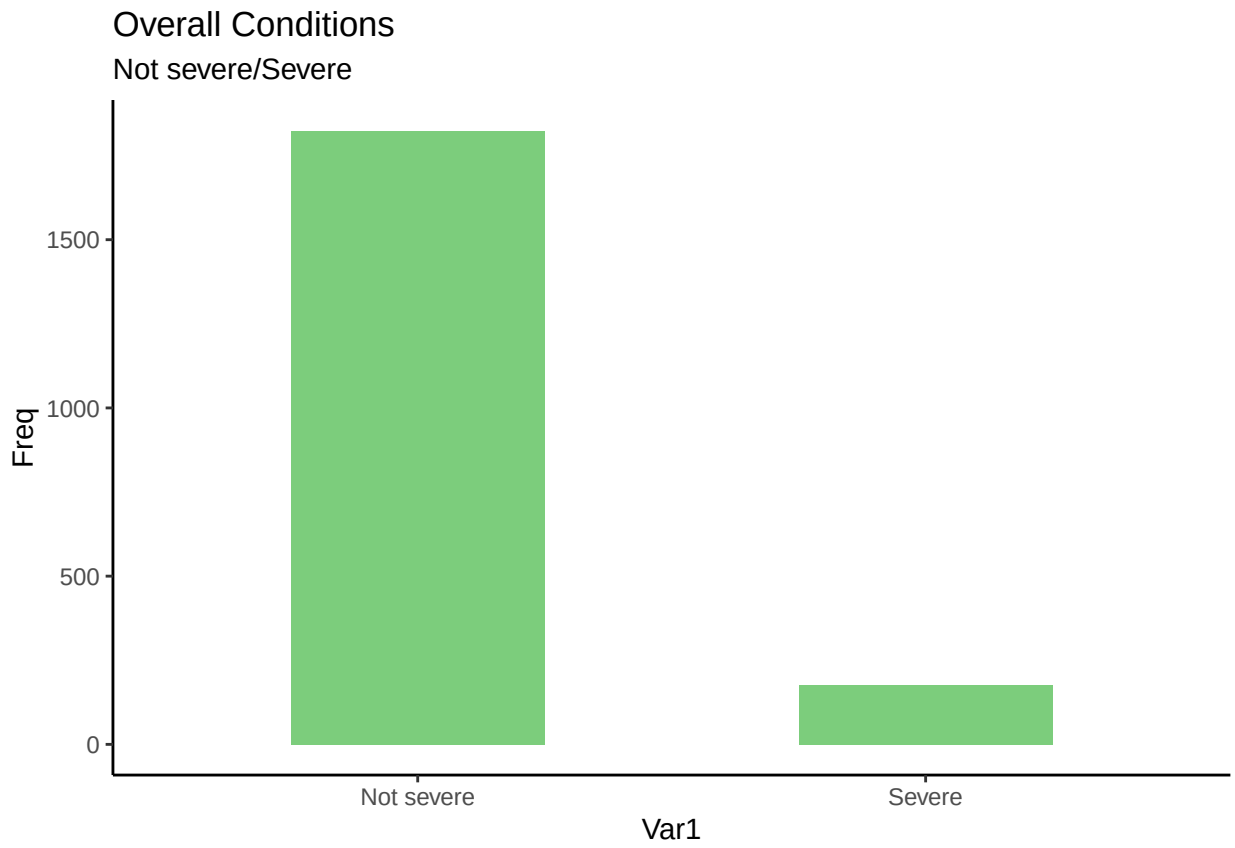
From the above plot, we can know that most participants are from the Study B, around 1250 out of 2000(62.5%). People from study A is around 350 and from study C is around 400.

What was the condition of the participants' severity of COVID-19 infection?

```
table(dat$severity)
```

```
##
## Not severe     Severe
##       1824        176
```

```
options(repr.plot.height = 3, repr.plot.width = 7)
ggplot(as.data.frame.table(table(dat$severity)), aes(Var1, Freq)) +
    geom_bar(stat = "identity", width = 0.5, fill = "palegreen3") +
    labs(title = "Overall Conditions", subtitle = "Not severe/Severe ")
```



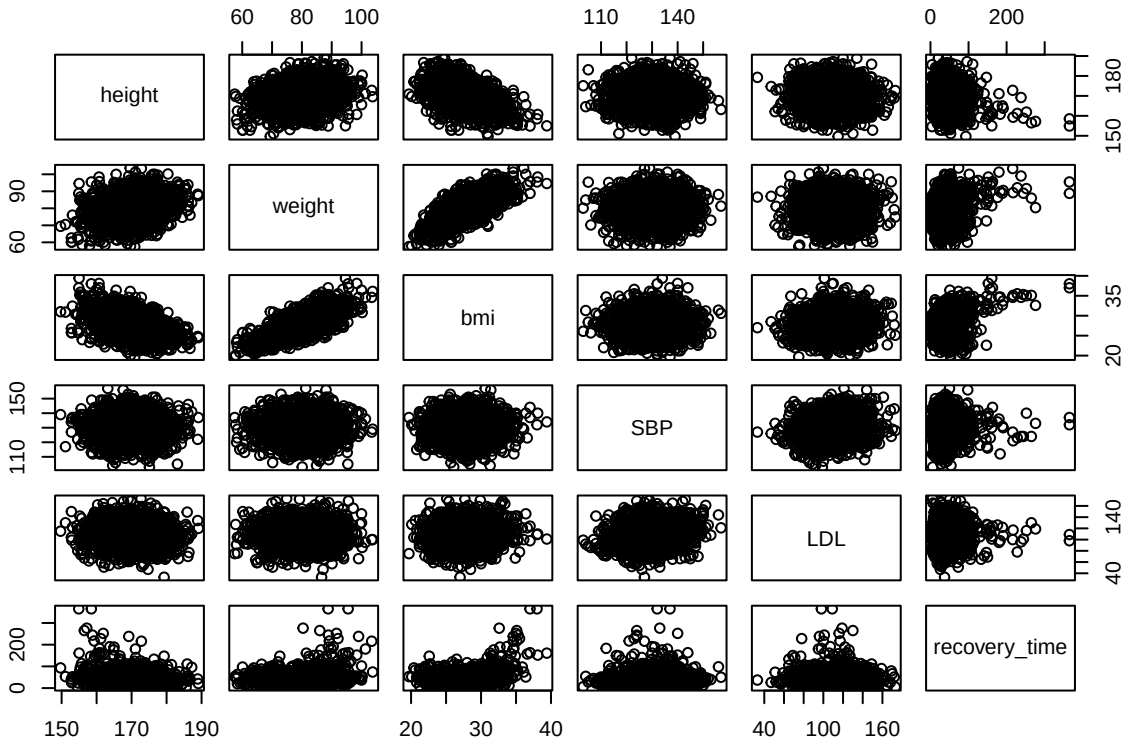
From the above information, to our relief, we know that most participants are not severe condition.

How Other Features Compare with Recovery time?

Numerical Features

Let's look at a few key numerical features at once and see how they correlate with Recovery time.

```
options(repr.plot.height = 7, repr.plot.width = 7)
num_vars <- c("height", "weight", "bmi", "SBP", "LDL", "recovery_time")
plot(dat[, num_vars])
```

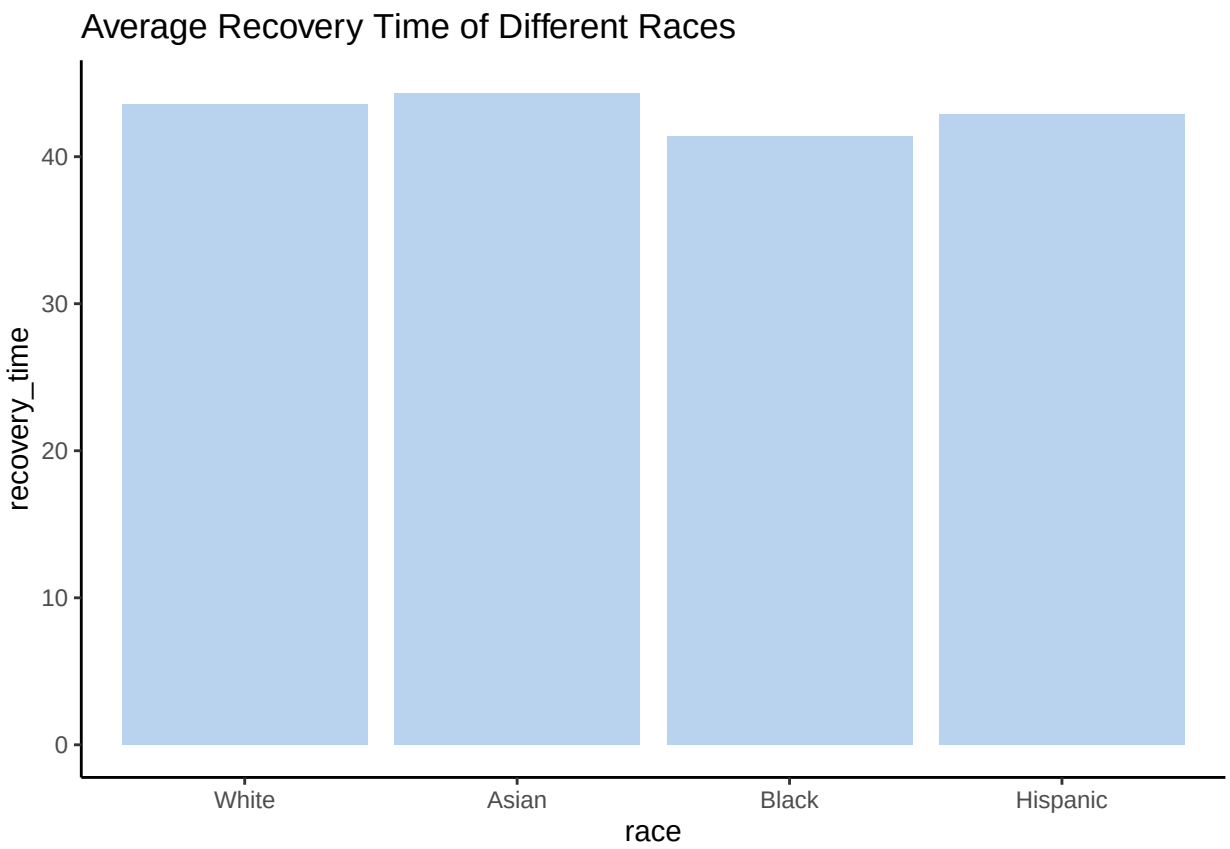


Most variables are little correlated with one another. Notably, **weight and bmi** seem positively correlated with recovery time; **height** seems negatively correlated with recovery time, and **SBP and LDL** seem no significant correlated with recovery time.

Categorical Features

Let's compare race with the time from COVID-19 infection to recovery in days.

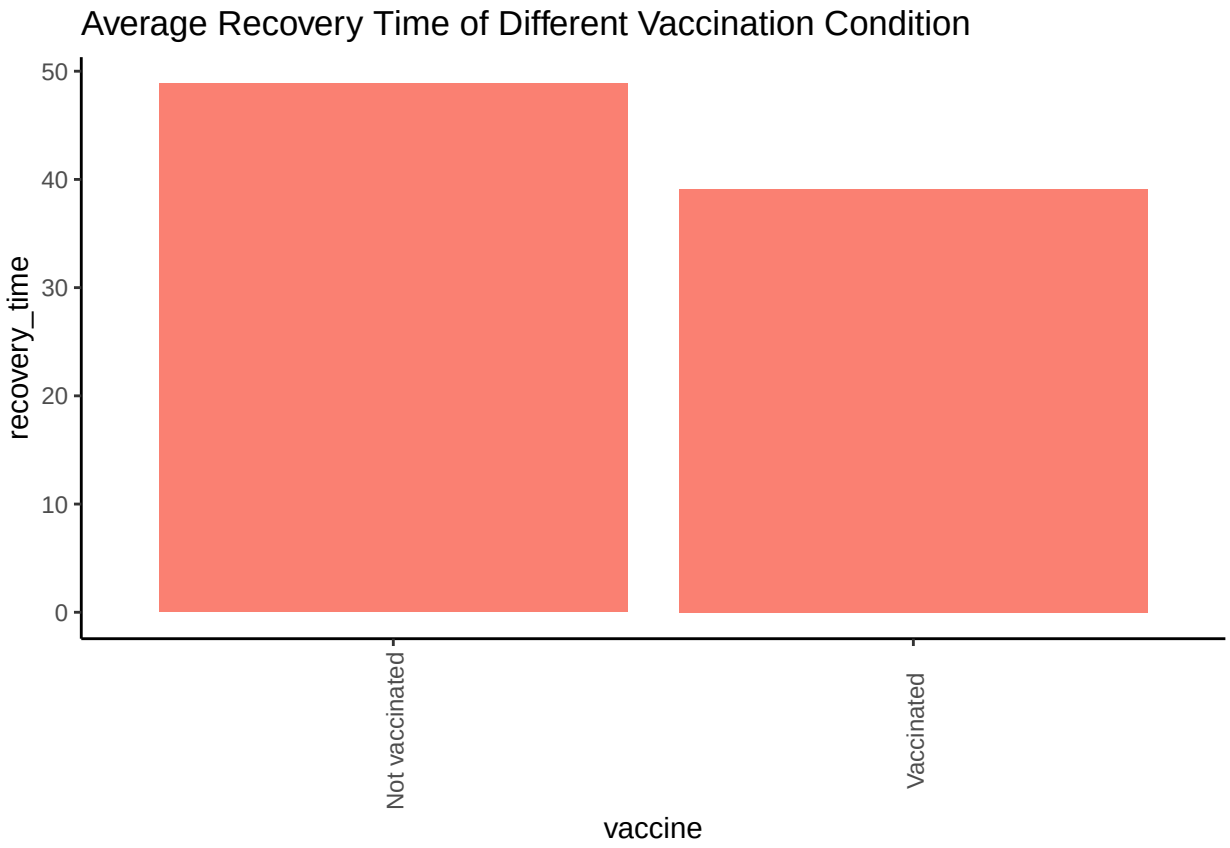
```
options(repr.plot.height = 4, repr.plot.width = 5)
ggplot(dat, aes(x=race, y=recovery_time)) +
    stat_summary(fun.y="mean", geom="bar", fill = "slategray2") +
    labs(title = "Average Recovery Time of Different Races")
```



From the graph, we can tell that there is no significant difference of recovery time among various race.The average of recovery days is more than 40 days. People self-identified as Asian tend to have the longest average recovery time and self-identified as American African tend to have the average shortest recovery time.

Let's also look at how vaccination condition relate to recovery time.

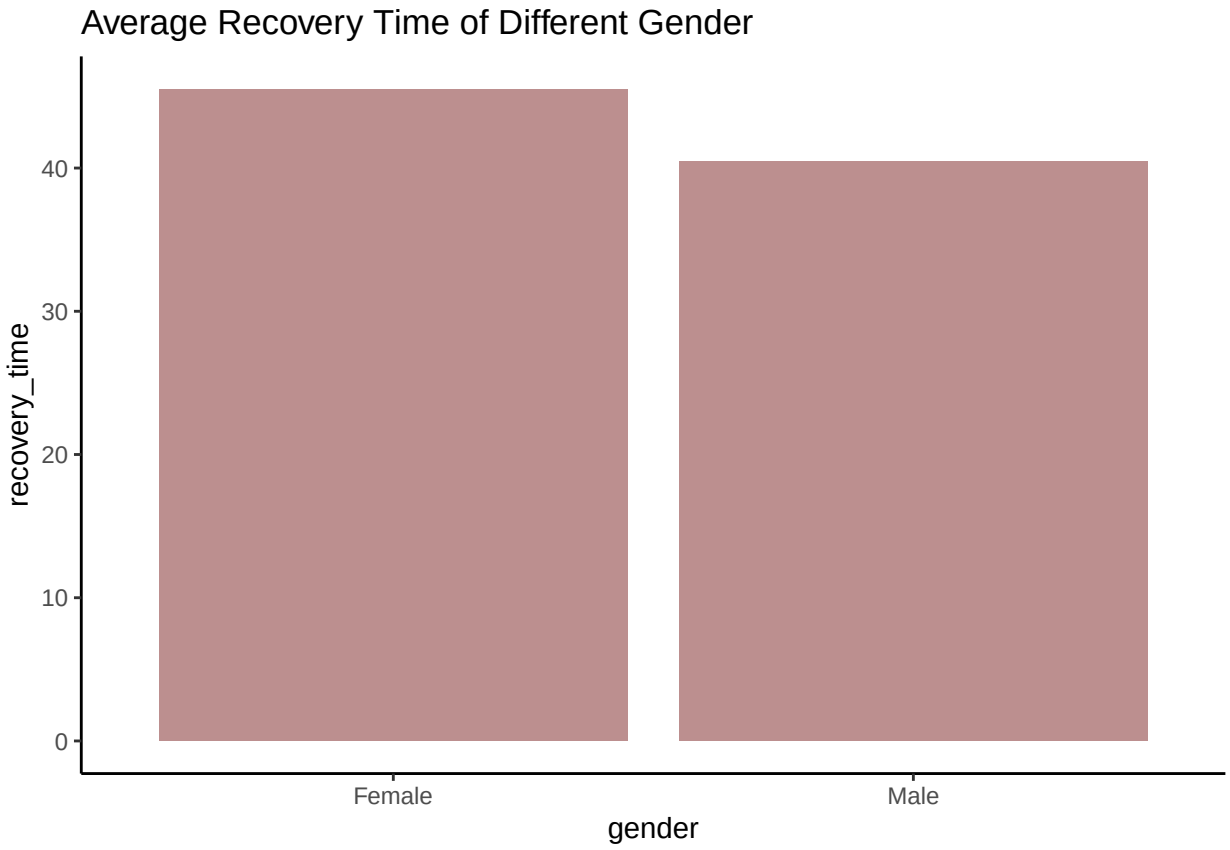
```
options(repr.plot.height = 4, repr.plot.width = 7)
ggplot(dat, aes(x=vaccine, y=recovery_time)) +
    stat_summary(fun.y="mean", geom="bar", fill = "salmon") +
    labs(title = "Average Recovery Time of Different Vaccination Condition") +
    theme(axis.text.x = element_text(angle=90, vjust=0.6))
```



From the graph, we can assume that vaccine is a related factors with recovery time. Since people in the not vaccined group(around 50 days of recovery time) have longer recovery time than the people in vaccined group (around 40 days of recovery time).

Which gender cost more days to recover?

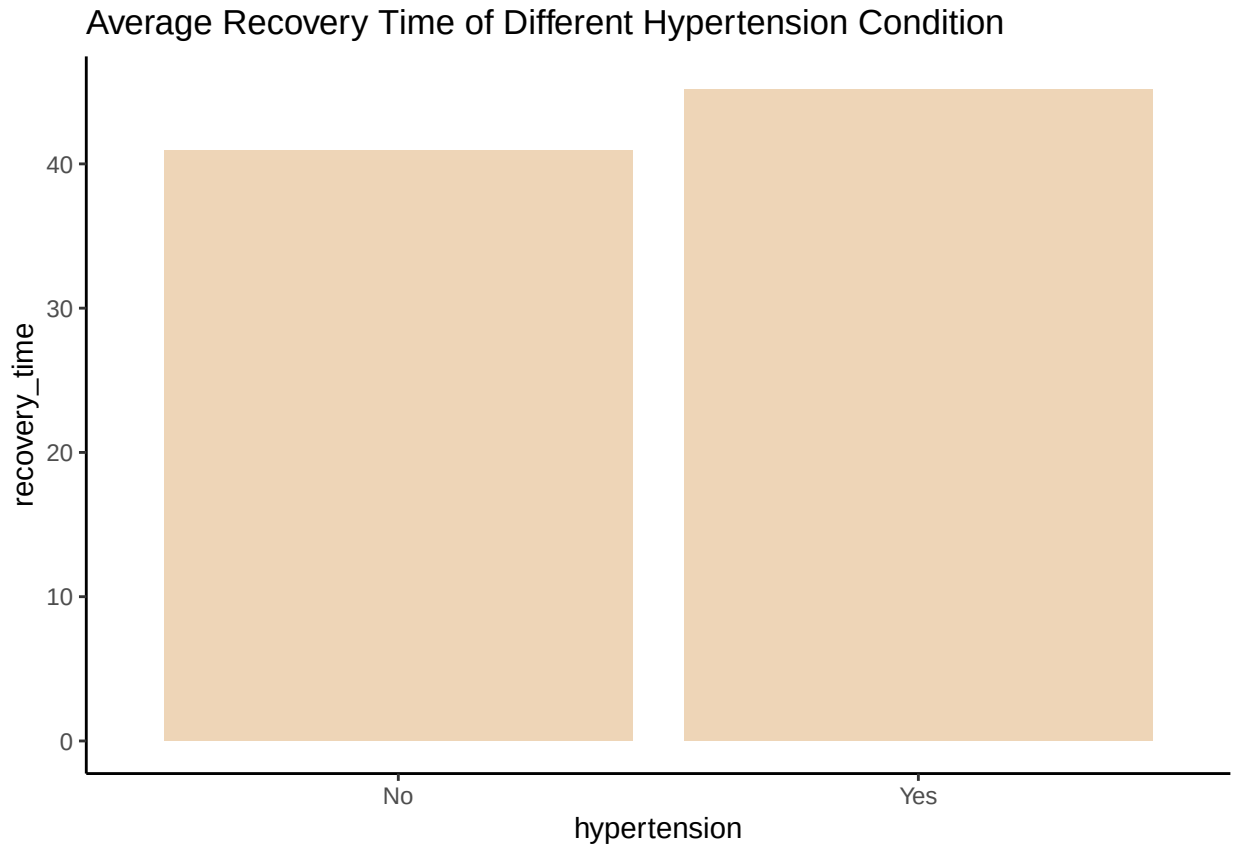
```
ggplot(dat, aes(x=gender, y=recovery_time)) +
    stat_summary(fun.y="mean", geom="bar", fill = "rosybrown") +
    labs(title = "Average Recovery Time of Different Gender")
```



From the above graph, it seems that female may need longer time to recover after COVID-19 infection. We can assume that gender tends to influence recovery time.

What about the hypertension? Would this variable affect the target?

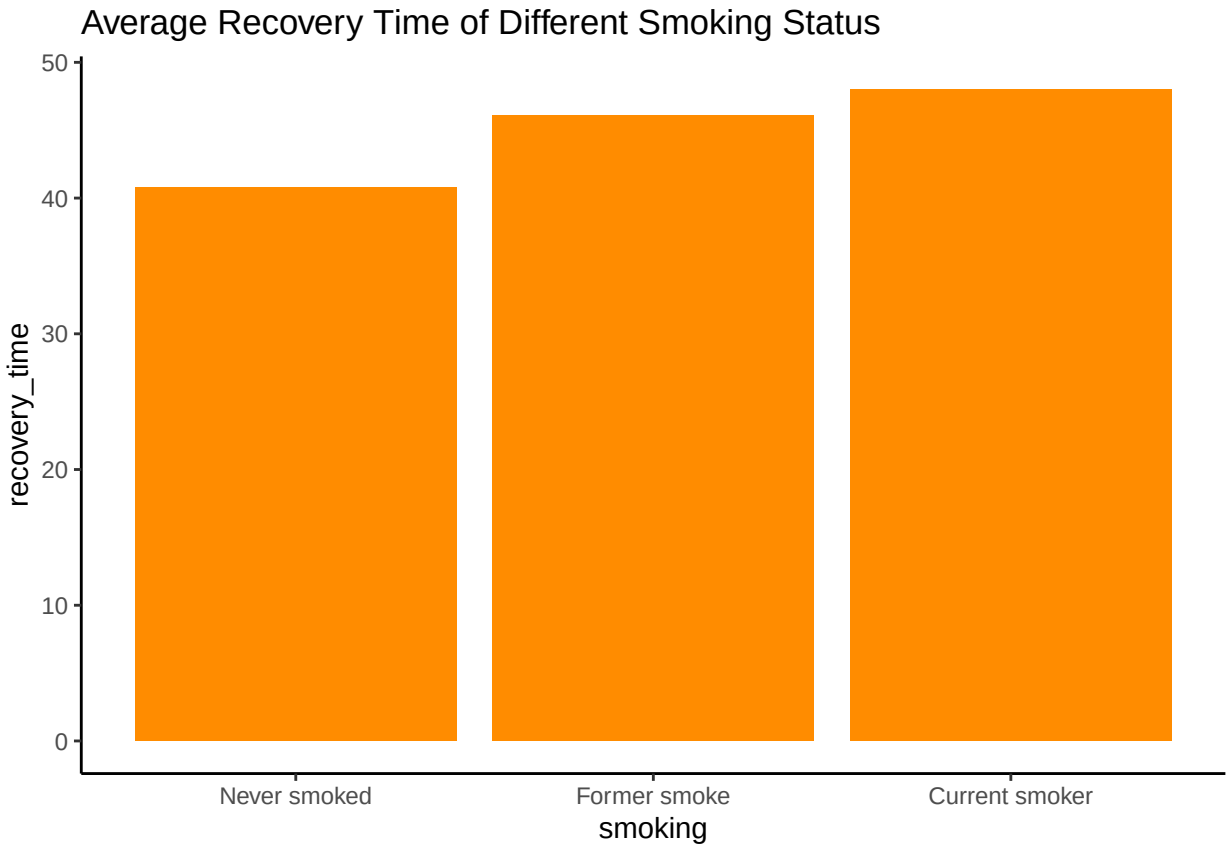
```
ggplot(dat, aes(x=hypertension, y=recovery_time)) +
    stat_summary(fun.y="mean", geom="bar", fill = "bisque2") +
    labs(title = "Average Recovery Time of Different Hypertension Condition")
```



From the above graph, people with hypertension need more time to recover after COVID-19 infection.

Let's focus on smoking status this time!

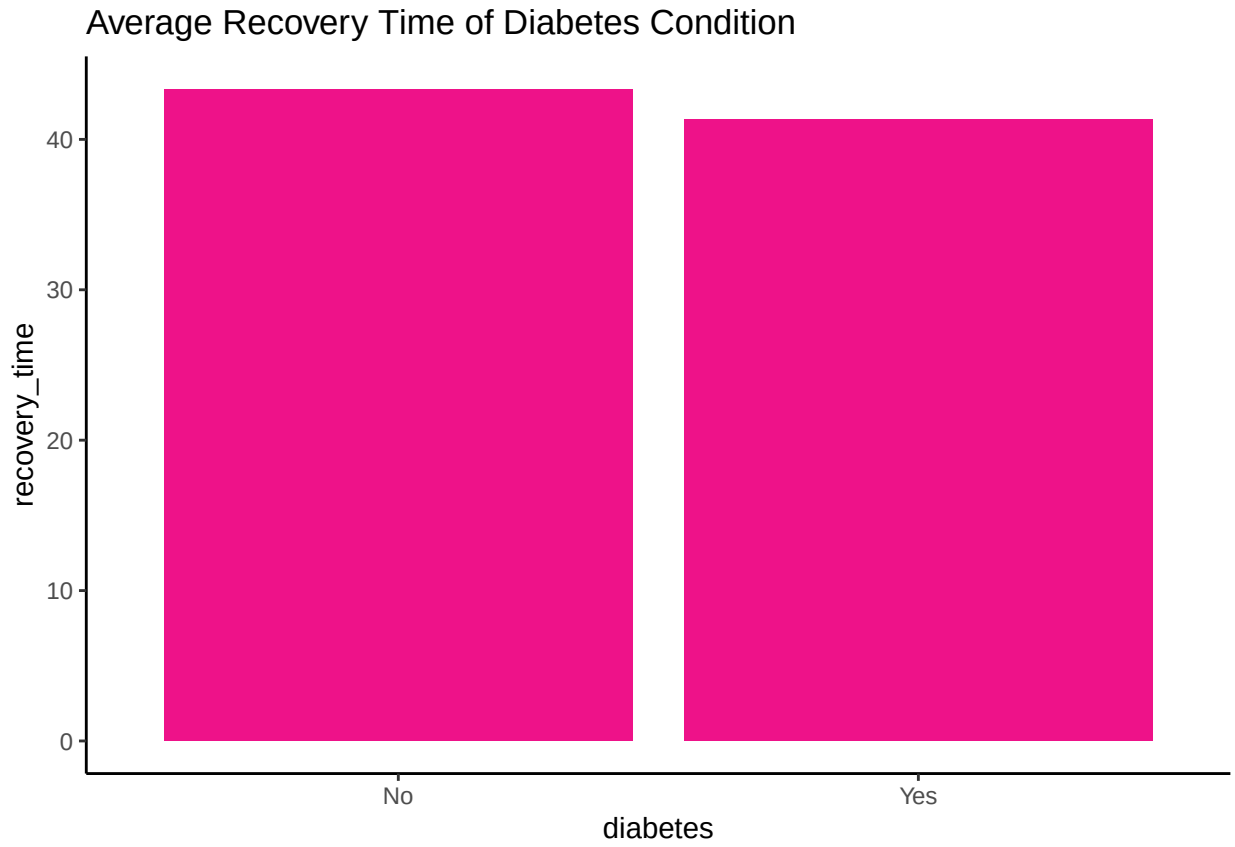
```
ggplot(dat, aes(x=smoking, y=recovery_time)) +
    stat_summary(fun.y="mean", geom="bar", fill = "darkorange") +
    labs(title = "Average Recovery Time of Different Smoking Status")
```



If people never smoke, they tend to less recovery time(around 40 days). People with smoking history or still are smokers are more likely to need more time to recovery from COVID-19.

The last one.. I promise!! (For variable 'diabetes')

```
ggplot(dat, aes(x=diabetes, y=recovery_time)) +
    stat_summary(fun.y="mean", geom="bar", fill = "deeppink2") +
    labs(title = "Average Recovery Time of Diabetes Condition")
```



There seems no significant difference of recovery time for people with or without diabetes.

- In conclusion, from these different graphs, we can determine which features tend to influence recovery time.