Final EDA

Tianshu Liu, Lincole Jiang, Jiong Ma

Contents

```
library(tidyverse)
library(summarytools)
library(corrplot)
library(caret)
library(vip)
library(rpart.plot)
library(ranger)
library(gridExtra)
```

1 Data Import

```
# import data
load("./recovery.RData")

set.seed(3196)
lts.dat <- dat[sample(1:10000, 2000),]
set.seed(2575)
lincole.dat <- dat[sample(1:10000, 2000),]
set.seed(5509)
amy.dat <- dat[sample(1:10000, 2000),]

dat1 <- lts.dat %>%
  merge(lincole.dat, all = TRUE) %>%
  na.omit() %>%
  select(-id) %>%
  mutate(
    gender = as.factor(gender),
    race = as.factor(race),
    smoking = as.factor(smoking),
    hypertension = as.factor(hypertension),
    diabetes = as.factor(diabetes),
    vaccine = as.factor(vaccine),
    severity = as.factor(severity),
    study = as.factor(study))

dat2 <- lts.dat %>%
  merge(amy.dat, all = TRUE) %>%
  na.omit() %>%
  select(-id) %>%
  mutate(
    gender = as.factor(gender),
    race = as.factor(race),
    smoking = as.factor(smoking),
    hypertension = as.factor(hypertension),
    diabetes = as.factor(diabetes),
    vaccine = as.factor(vaccine),
    severity = as.factor(severity),
    study = as.factor(study))

dat3 <- lincole.dat %>%
  merge(amy.dat, all = TRUE) %>%
  na.omit() %>%
```

```r
  select(-id) %>%
  mutate(
    gender = as.factor(gender),
    race = as.factor(race),
    smoking = as.factor(smoking),
    hypertension = as.factor(hypertension),
    diabetes = as.factor(diabetes),
    vaccine = as.factor(vaccine),
    severity = as.factor(severity),
    study = as.factor(study))

dat <- dat1
summary(dat)
```

```
##       age        gender   race     smoking      height          weight
##  Min.   :45.00   0:1842   1:2372   0:2223   Min.   :151.2   Min.   : 56.70
##  1st Qu.:57.00   1:1781   2: 172   1:1034   1st Qu.:166.2   1st Qu.: 75.40
##  Median :60.00            3: 716   2: 366   Median :170.2   Median : 80.20
##  Mean   :60.06            4: 363            Mean   :170.2   Mean   : 80.13
##  3rd Qu.:63.00                              3rd Qu.:174.2   3rd Qu.: 84.80
##  Max.   :77.00                              Max.   :188.6   Max.   :103.40
##       bmi        hypertension diabetes      SBP             LDL        vaccine
##  Min.   :19.70   0:1891       0:3065   Min.   :102.0   Min.   : 28.0   0:1469
##  1st Qu.:25.80   1:1732       1: 558   1st Qu.:125.0   1st Qu.: 97.0   1:2154
##  Median :27.60                         Median :130.0   Median :110.0
##  Mean   :27.73                         Mean   :130.2   Mean   :110.5
##  3rd Qu.:29.40                         3rd Qu.:136.0   3rd Qu.:124.0
##  Max.   :39.80                         Max.   :158.0   Max.   :174.0
##  severity study    recovery_time
##  0:3289   A: 728   Min.   :  3.00
##  1: 334   B:2171   1st Qu.: 28.00
##           C: 724   Median : 38.00
##                    Mean   : 42.87
##                    3rd Qu.: 49.00
##                    Max.   :365.00
```

```r
bin.dat <- dat %>%
  mutate(recovery_time = ifelse(recovery_time > 30, "gt30", "lt30")) %>%
  mutate(recovery_time = factor(recovery_time, levels = c("lt30", "gt30")))

summary(bin.dat)
```

```
##       age        gender   race     smoking      height          weight
##  Min.   :45.00   0:1842   1:2372   0:2223   Min.   :151.2   Min.   : 56.70
##  1st Qu.:57.00   1:1781   2: 172   1:1034   1st Qu.:166.2   1st Qu.: 75.40
##  Median :60.00            3: 716   2: 366   Median :170.2   Median : 80.20
##  Mean   :60.06            4: 363            Mean   :170.2   Mean   : 80.13
##  3rd Qu.:63.00                              3rd Qu.:174.2   3rd Qu.: 84.80
##  Max.   :77.00                              Max.   :188.6   Max.   :103.40
##       bmi        hypertension diabetes      SBP             LDL        vaccine
##  Min.   :19.70   0:1891       0:3065   Min.   :102.0   Min.   : 28.0   0:1469
##  1st Qu.:25.80   1:1732       1: 558   1st Qu.:125.0   1st Qu.: 97.0   1:2154
##  Median :27.60                         Median :130.0   Median :110.0
##  Mean   :27.73                         Mean   :130.2   Mean   :110.5
##  3rd Qu.:29.40                         3rd Qu.:136.0   3rd Qu.:124.0
##  Max.   :39.80                         Max.   :158.0   Max.   :174.0
##  severity study    recovery_time
##  0:3289   A: 728   lt30:1102
##  1: 334   B:2171   gt30:2521
##           C: 724
##
##
##
```

2 Data partition

```
# data partition
dat.matrix <- model.matrix(recovery_time ~ ., dat)[ ,-1]

set.seed(2023)
trainRows <- createDataPartition(y = dat$recovery_time, p = 0.8, list = FALSE)

train.dat <- dat[trainRows,]
train.bin.dat <- bin.dat[trainRows,]

train.dat.matrix <- model.matrix(~., train.dat)[, -1]
train.bin.dat.matrix <- train.dat.matrix %>%
  as.data.frame() %>%
 mutate(recovery_time = ifelse(recovery_time > 30, "gt30", "lt30")) %>%
  mutate(recovery_time = factor(recovery_time, levels = c("lt30", "gt30")))

train.x <- dat.matrix[trainRows,]
train.y <- dat$recovery_time[trainRows]
train.bin.y <- bin.dat$recovery_time[trainRows]

test.x <- dat.matrix[-trainRows,]
test.y <- dat$recovery_time[-trainRows]
test.bin.y <- bin.dat$recovery_time[-trainRows]
```

3 Exploratory analysis and data visualization

```
# data summary
st_options(plain.ascii = FALSE,
           style = "rmarkdown",
           dfSummary.silent = TRUE,
           footnote = NA,
           subtitle.emphasis = FALSE)
dfSummary(train.dat)
```

3.0.1 Data Frame Summary

train.dat
Dimensions: 2900 x 15
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	age [numeric]	Mean (sd) : 60.1 (4.5) min < med < max: 45 < 60 < 77 IQR (CV) : 6 (0.1)	33 distinct values	: : . : : : : : . . : : : : : .	2900 (100.0%)	0 (0.0%)
2	gender [factor]	1. 0 2. 1	1468 (50.6%) 1432 (49.4%)	IIIIIIIIII IIIIIIIII	2900 (100.0%)	0 (0.0%)
3	race [factor]	1. 1 2. 2 3. 3 4. 4	1909 (65.8%) 132 (4.6%) 568 (19.6%) 291 (10.0%)	IIIIIIIIIIIII III II	2900 (100.0%)	0 (0.0%)
4	smoking [factor]	1. 0 2. 1 3. 2	1763 (60.8%) 845 (29.1%) 292 (10.1%)	IIIIIIIIIIII IIIII II	2900 (100.0%)	0 (0.0%)
5	height [numeric]	Mean (sd) : 170.2 (6) min < med < max: 151.2 < 170.1 < 188.6 IQR (CV) : 8 (0)	312 distinct values	: : : : . : : . : : : : . : : : : .	2900 (100.0%)	0 (0.0%)
6	weight [numeric]	Mean (sd) : 80.2 (7) min < med < max: 57.1 < 80.3 < 103.4 IQR (CV) : 9.5 (0.1)	361 distinct values	. : . : : : : : : . : : : : . . : : : : : : .	2900 (100.0%)	0 (0.0%)
7	bmi [numeric]	Mean (sd) : 27.8 (2.7) min < med < max: 19.7 < 27.7 < 39.8 IQR (CV) : 3.6 (0.1)	160 distinct values	: . : : : : : . : : : : : : : : : : .	2900 (100.0%)	0 (0.0%)
8	hypertension [factor]	1. 0 2. 1	1514 (52.2%) 1386 (47.8%)	IIIIIIIIII IIIIIIIII	2900 (100.0%)	0 (0.0%)
9	diabetes [factor]	1. 0 2. 1	2446 (84.3%) 454 (15.7%)	IIIIIIIIIIIIIIII III	2900 (100.0%)	0 (0.0%)
10	SBP [numeric]	Mean (sd) : 130.2 (8.1) min < med < max: 104 < 130 < 158 IQR (CV) : 11 (0.1)	54 distinct values	: : . : : : . . : : : : . : : : : : :	2900 (100.0%)	0 (0.0%)
11	LDL [numeric]	Mean (sd) : 110.3 (19.9) min < med < max: 32 < 110 < 174 IQR (CV) : 27 (0.2)	116 distinct values	. : : : : : : : . : : : : : . : : : : : .	2900 (100.0%)	0 (0.0%)
12	vaccine [factor]	1. 0 2. 1	1192 (41.1%) 1708 (58.9%)	IIIIIIII IIIIIIIIIII	2900 (100.0%)	0 (0.0%)
13	severity [factor]	1. 0 2. 1	2619 (90.3%) 281 (9.7%)	IIIIIIIIIIIIIIIIII I	2900 (100.0%)	0 (0.0%)
14	study [factor]	1. A 2. B 3. C	580 (20.0%) 1750 (60.3%) 570 (19.7%)	IIII IIIIIIIIIIII III	2900 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
15	recovery_time [numeric]	Mean (sd) : 43 (30.5) min < med < max: 3 < 38 < 365 IQR (CV) : 21 (0.7)	144 distinct values	: : : : : : : : : .	2900 (100.0%)	0 (0.0%)

```
skimr::skim_without_charts(train.dat)
```

Table 2: Data summary

Name	train.dat
Number of rows	2900
Number of columns	15
Column type frequency:	
factor	8
numeric	7
Group variables	None

Variable type: factor

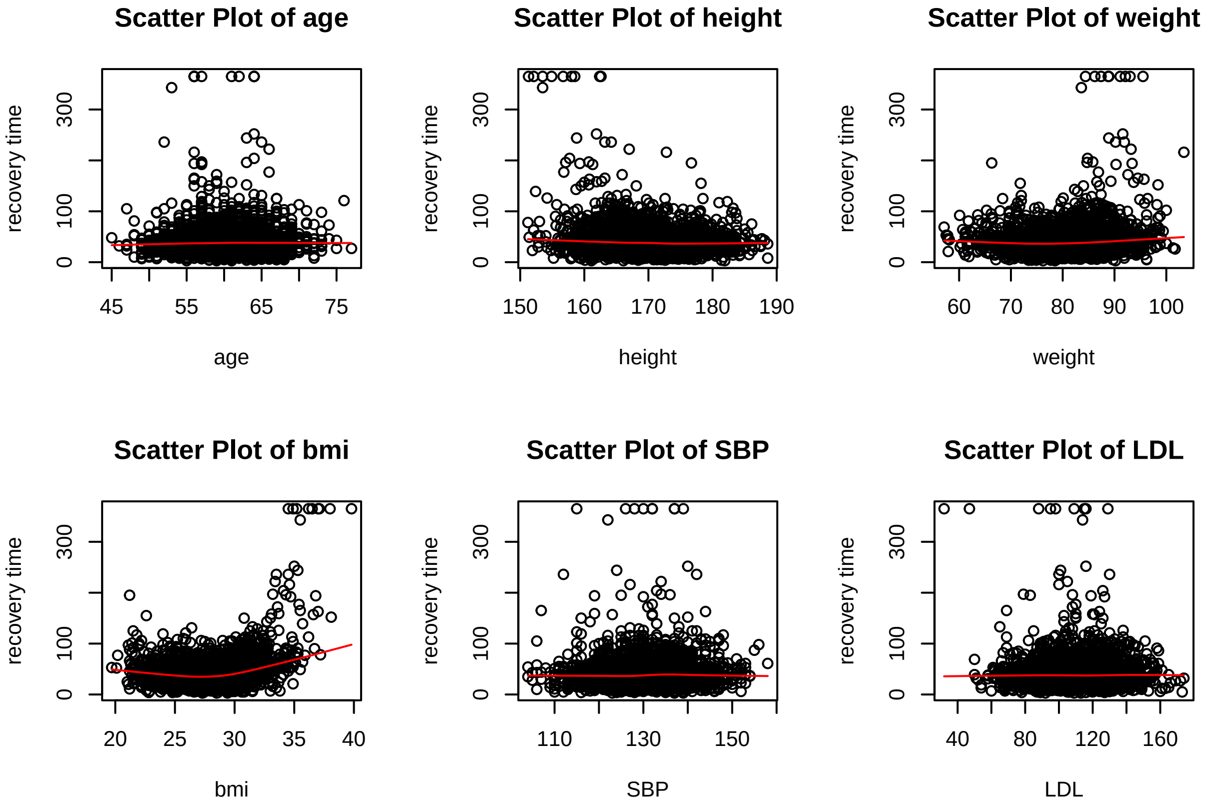
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1468, 1: 1432
race	0	1	FALSE	4	1: 1909, 3: 568, 4: 291, 2: 132
smoking	0	1	FALSE	3	0: 1763, 1: 845, 2: 292
hypertension	0	1	FALSE	2	0: 1514, 1: 1386
diabetes	0	1	FALSE	2	0: 2446, 1: 454
vaccine	0	1	FALSE	2	1: 1708, 0: 1192
severity	0	1	FALSE	2	0: 2619, 1: 281
study	0	1	FALSE	3	B: 1750, A: 580, C: 570

Variable type: numeric

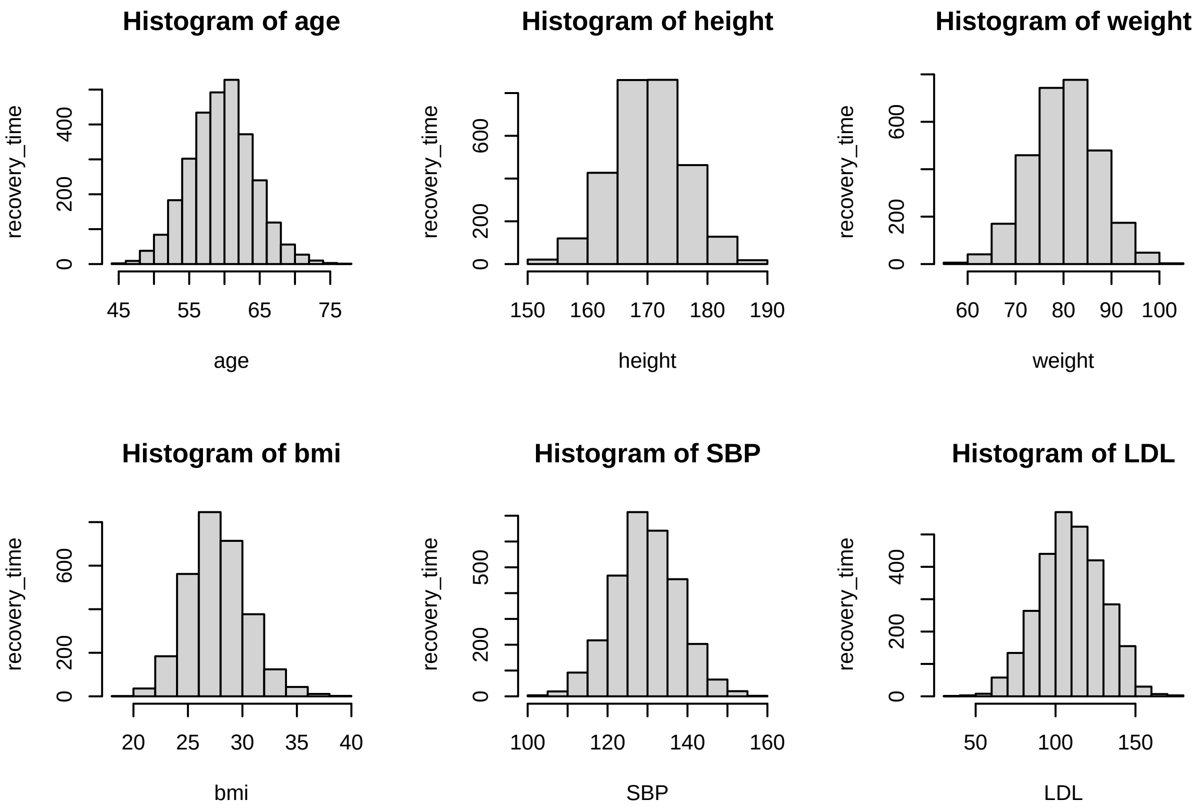
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	60.07	4.51	45.0	57.0	60.00	63.0	77.0
height	0	1	170.17	6.04	151.2	166.1	170.15	174.1	188.6
weight	0	1	80.20	7.00	57.1	75.4	80.30	84.9	103.4
bmi	0	1	27.76	2.73	19.7	25.9	27.70	29.5	39.8
SBP	0	1	130.19	8.08	104.0	125.0	130.00	136.0	158.0
LDL	0	1	110.27	19.87	32.0	97.0	110.00	124.0	174.0
recovery_time	0	1	43.02	30.51	3.0	28.0	38.00	49.0	365.0

```
##############################################################
## Remember to edit the next chunk if you do any modification here:)
##############################################################
```

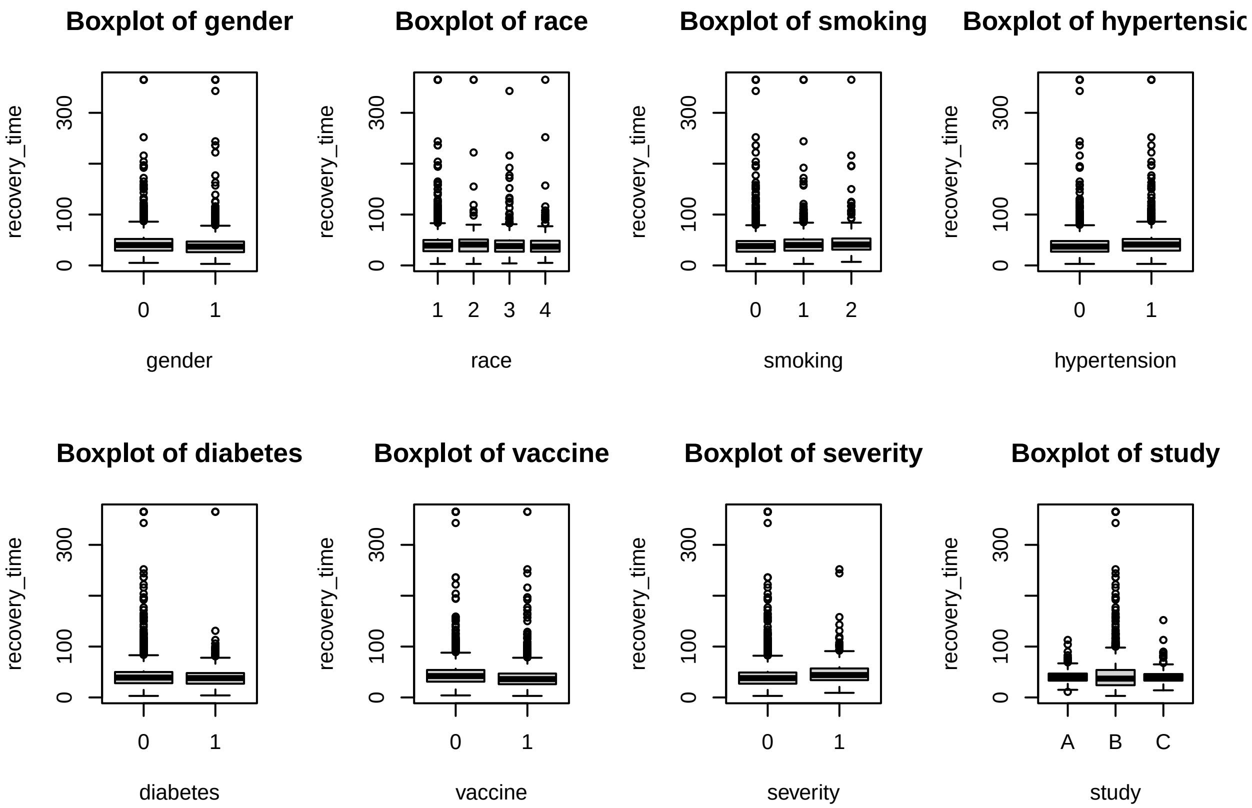
```r
# EDA
cts_var = c("age", "height", "weight", "bmi", "SBP", "LDL")
fct_var = c("gender", "race", "smoking", "hypertension", "diabetes", "vaccine", "severity", "study")

# scatter plot of continuous predictors
par(mfrow=c(2, 3))
for (i in 1:length(cts_var)){
  var = cts_var[i]
  plot(recovery_time~train.dat[,var],
       data = train.dat,
       ylab = "recovery time",
       xlab = var,
       main = str_c("Scatter Plot of ", var))
  lines(stats::lowess(train.dat[,var], train.dat$recovery_time), col = "red", type = "l")
}
```



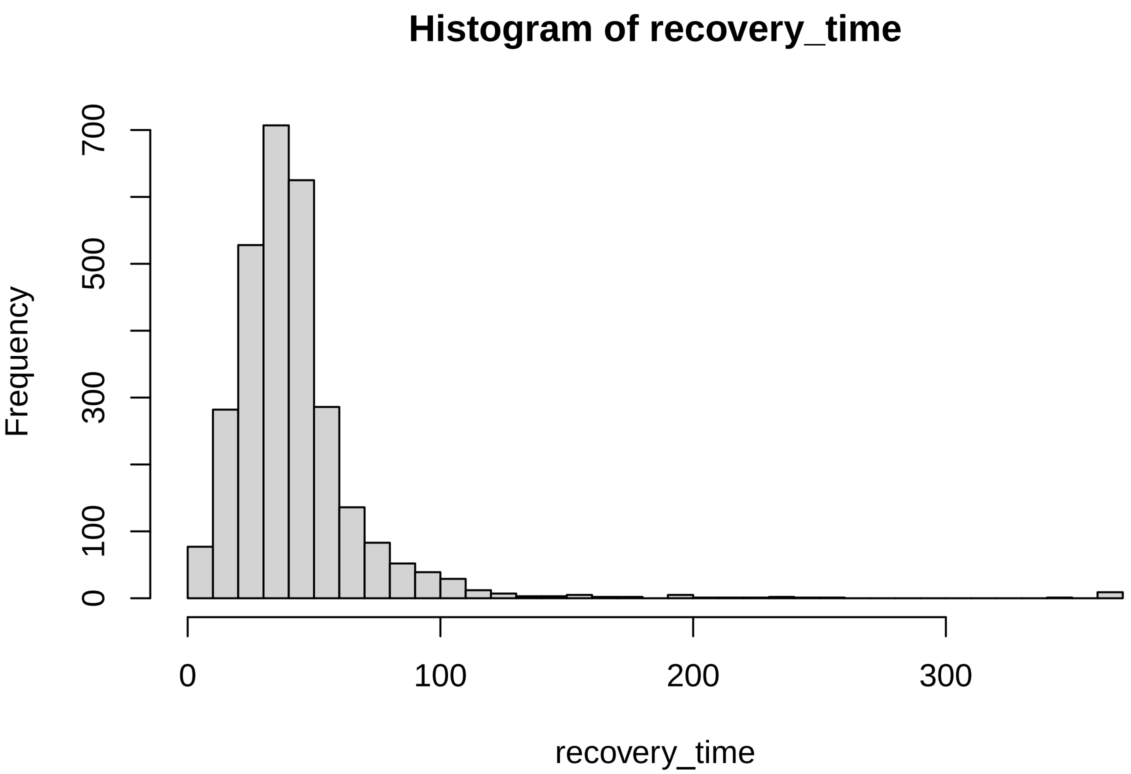
```r
for (i in 1:length(cts_var)){
  var = cts_var[i]
  hist(train.dat[,var],
       ylab = "recovery_time",
       xlab = var,
       main = str_c("Histogram of ", var))
}
```



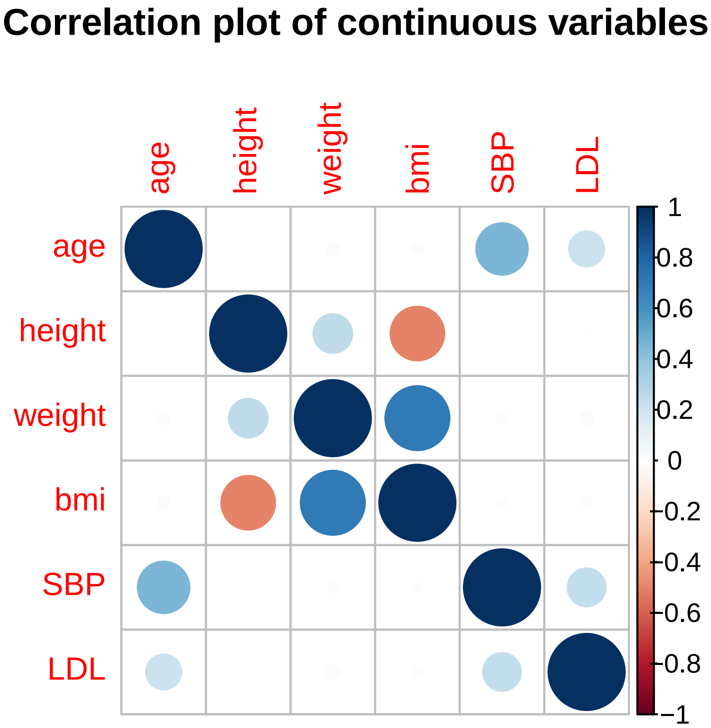
```r
# boxplot of categorical predictors
par(mfrow=c(2, 4))
for (i in 1:length(fct_var)){
  var = fct_var[i]
  plot(recovery_time~train.dat[,var],
       data = train.dat,
       ylab = "recovery_time",
       xlab = var,
       main = str_c("Boxplot of ", var))
}
```



```r
# histogram of response
par(mfrow=c(1, 1))
hist(train.dat$recovery_time,
     breaks = 50,
     main = "Histogram of recovery_time",
     xlab = "recovery_time")
```



```
# correlation
par(mfrow=c(1, 1))
corrplot(cor(train.dat[,cts_var]), method = "circle", type = "full",
         title = "Correlation plot of continuous variables",
         mar = c(2, 2, 4, 2))
```



```
# data summary
st_options(plain.ascii = FALSE,
           style = "rmarkdown",
           dfSummary.silent = TRUE,
           footnote = NA,
           subtitle.emphasis = FALSE)
dfSummary(train.bin.dat)
```

3.0.2 Data Frame Summary

train.bin.dat
Dimensions: 2900 x 15
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	age [numeric]	Mean (sd) : 60.1 (4.5) min < med < max: 45 < 60 < 77 IQR (CV) : 6 (0.1)	33 distinct values	: : . : : : : : . . : : : : : .	2900 (100.0%)	0 (0.0%)
2	gender [factor]	1. 0 2. 1	1468 (50.6%) 1432 (49.4%)	IIIIIIIIII IIIIIIIII	2900 (100.0%)	0 (0.0%)
3	race [factor]	1. 1 2. 2 3. 3 4. 4	1909 (65.8%) 132 (4.6%) 568 (19.6%) 291 (10.0%)	IIIIIIIIIIIII III II	2900 (100.0%)	0 (0.0%)
4	smoking [factor]	1. 0 2. 1 3. 2	1763 (60.8%) 845 (29.1%) 292 (10.1%)	IIIIIIIIIIII IIIII II	2900 (100.0%)	0 (0.0%)
5	height [numeric]	Mean (sd) : 170.2 (6) min < med < max: 151.2 < 170.1 < 188.6 IQR (CV) : 8 (0)	312 distinct values	: : : : . : : . : : : : . : : : : .	2900 (100.0%)	0 (0.0%)
6	weight [numeric]	Mean (sd) : 80.2 (7) min < med < max: 57.1 < 80.3 < 103.4 IQR (CV) : 9.5 (0.1)	361 distinct values	. : . : : : : : : . : : : : . . : : : : : : .	2900 (100.0%)	0 (0.0%)
7	bmi [numeric]	Mean (sd) : 27.8 (2.7) min < med < max: 19.7 < 27.7 < 39.8 IQR (CV) : 3.6 (0.1)	160 distinct values	: . : : : : : . : : : : : : : : : : .	2900 (100.0%)	0 (0.0%)
8	hypertension [factor]	1. 0 2. 1	1514 (52.2%) 1386 (47.8%)	IIIIIIIIII IIIIIIIII	2900 (100.0%)	0 (0.0%)
9	diabetes [factor]	1. 0 2. 1	2446 (84.3%) 454 (15.7%)	IIIIIIIIIIIIIIII III	2900 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
10	SBP [numeric]	Mean (sd) : 130.2 (8.1) min < med < max: 104 < 130 < 158 IQR (CV) : 11 (0.1)	54 distinct values	: : . : : : . . : : : : . : : : : : :	2900 (100.0%)	0 (0.0%)
11	LDL [numeric]	Mean (sd) : 110.3 (19.9) min < med < max: 32 < 110 < 174 IQR (CV) : 27 (0.2)	116 distinct values	. : : : : : : : . : : : : : . : : : : : .	2900 (100.0%)	0 (0.0%)
12	vaccine [factor]	1. 0 2. 1	1192 (41.1%) 1708 (58.9%)	IIIIIIII IIIIIIIIIII	2900 (100.0%)	0 (0.0%)
13	severity [factor]	1. 0 2. 1	2619 (90.3%) 281 (9.7%)	IIIIIIIIIIIIIIIIII I	2900 (100.0%)	0 (0.0%)
14	study [factor]	1. A 2. B 3. C	580 (20.0%) 1750 (60.3%) 570 (19.7%)	IIII IIIIIIIIIIII III	2900 (100.0%)	0 (0.0%)
15	recovery_time [factor]	1. lt30 2. gt30	887 (30.6%) 2013 (69.4%)	IIIIII IIIIIIIIIIIII	2900 (100.0%)	0 (0.0%)

```
skimr::skim_without_charts(train.bin.dat)
```

Table 6: Data summary

Name	train.bin.dat
Number of rows	2900
Number of columns	15
Column type frequency:	
factor	9
numeric	6
Group variables	None

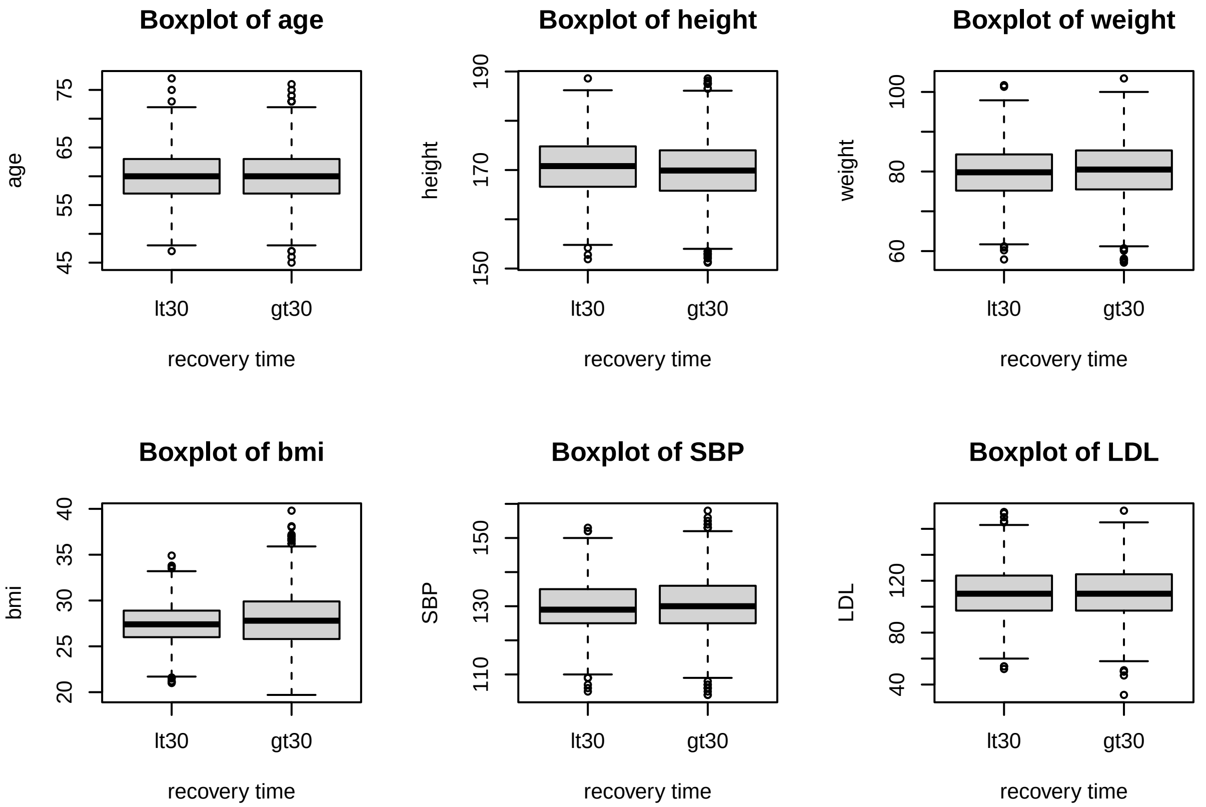
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1468, 1: 1432
race	0	1	FALSE	4	1: 1909, 3: 568, 4: 291, 2: 132
smoking	0	1	FALSE	3	0: 1763, 1: 845, 2: 292
hypertension	0	1	FALSE	2	0: 1514, 1: 1386
diabetes	0	1	FALSE	2	0: 2446, 1: 454
vaccine	0	1	FALSE	2	1: 1708, 0: 1192
severity	0	1	FALSE	2	0: 2619, 1: 281
study	0	1	FALSE	3	B: 1750, A: 580, C: 570
recovery_time	0	1	FALSE	2	gt3: 2013, lt3: 887

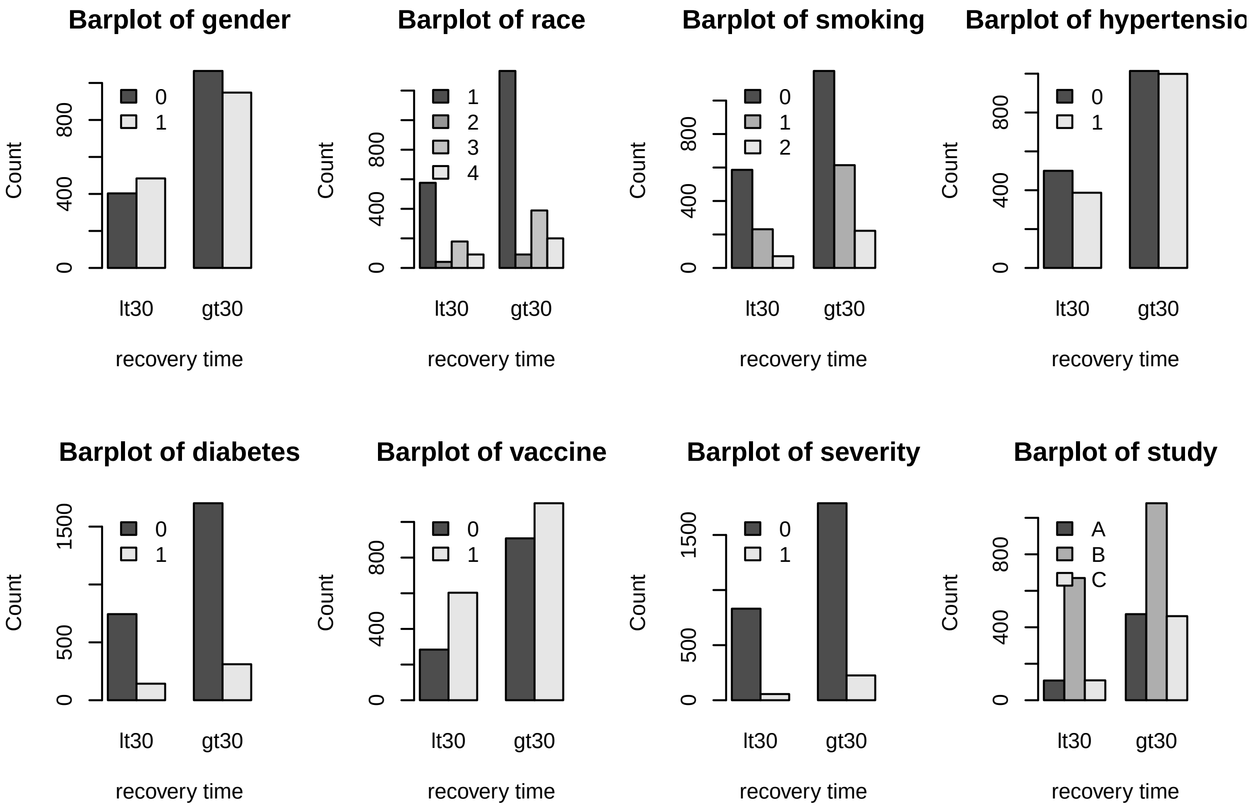
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	60.07	4.51	45.0	57.0	60.00	63.0	77.0
height	0	1	170.17	6.04	151.2	166.1	170.15	174.1	188.6
weight	0	1	80.20	7.00	57.1	75.4	80.30	84.9	103.4
bmi	0	1	27.76	2.73	19.7	25.9	27.70	29.5	39.8
SBP	0	1	130.19	8.08	104.0	125.0	130.00	136.0	158.0
LDL	0	1	110.27	19.87	32.0	97.0	110.00	124.0	174.0

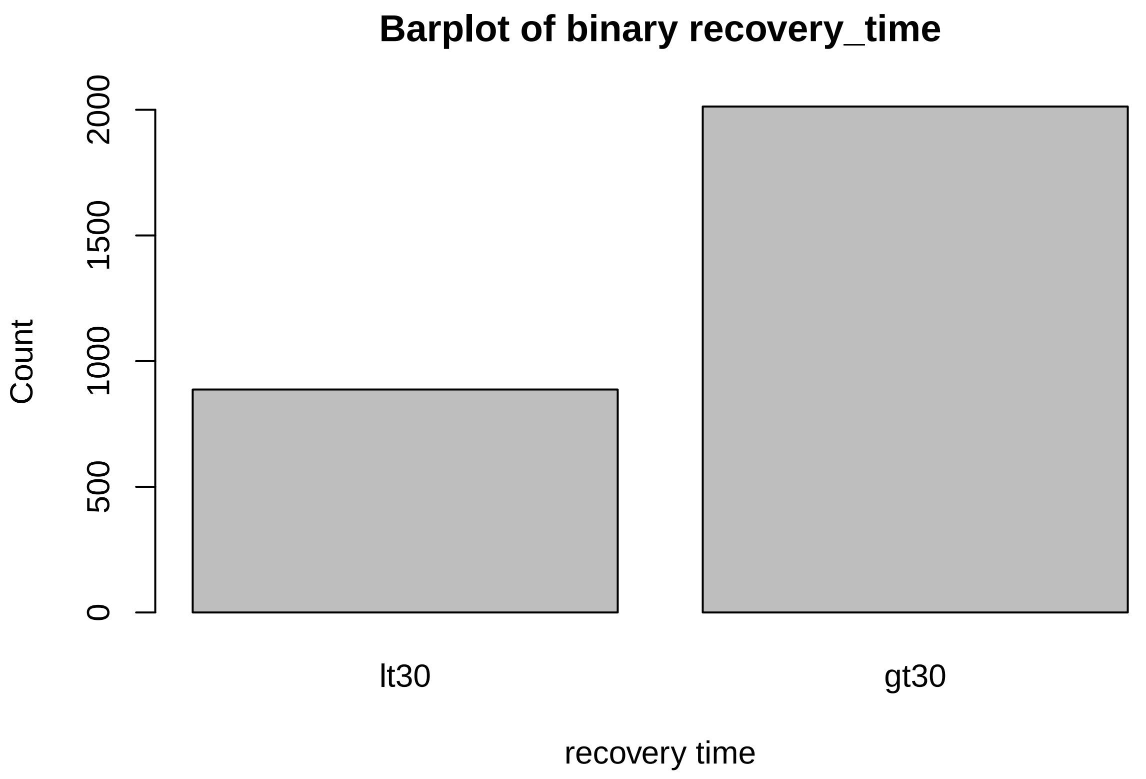
```r
###################################################################
## Remember to edit the next chunk if you do any modification here:)
###################################################################
# EDA

# boxplot of continuous predictors
par(mfrow=c(2, 3))
for (i in 1:length(cts_var)){
  var = cts_var[i]
  boxplot(train.bin.dat[,var]~recovery_time,
        data = train.bin.dat,
        xlab = "recovery time",
        ylab = var,
        main = str_c("Boxplot of ", var))
}
```



```r
# barplot of categorical predictors
par(mfrow=c(2, 4))
for (i in 1:length(fct_var)){
  var <- fct_var[i]
  counts <- table(train.bin.dat[,var], train.bin.y)
  barplot(counts, beside = TRUE, legend.text = TRUE,
        xlab = "recovery time",
        ylab = "Count",
        main = str_c("Barplot of ", var),
        args.legend = list(bty = 'n', x = 'topleft'))
}
```



```
# barplot of response
par(mfrow=c(1, 1))
counts <- table(train.bin.y)
barplot(counts,
        xlab = "recovery time",
        ylab = "Count",
        main = "Barplot of binary recovery_time")
```


Barplot of binary recovery_time
Count
0
500
1000
1500
2000
lt30
gt30
recovery time