Midterm Result

Yiying Wu (yw3996)

Input dataset

Exploratory analysis and data visualization

In this section, use appropriate visualization techniques to explore the dataset and identify any patterns or relationships in the data.

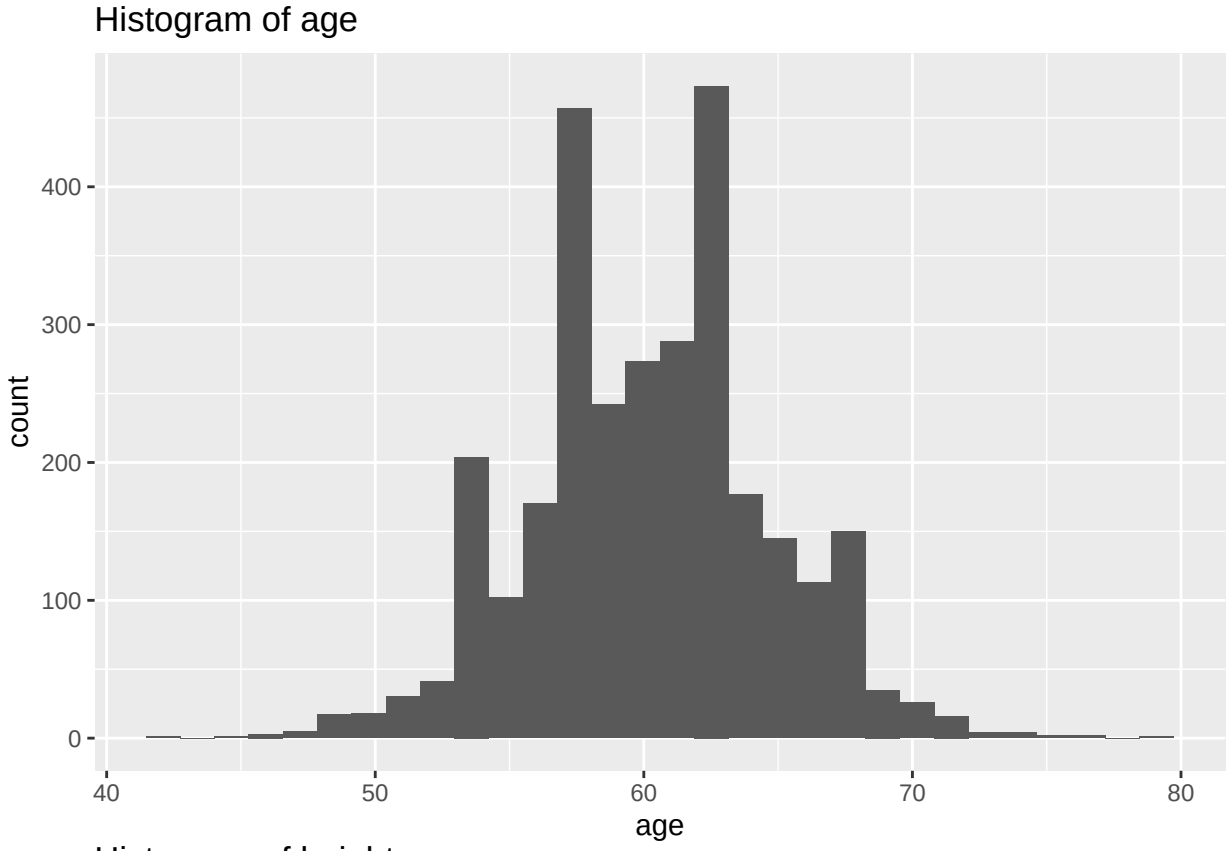
Summary statistics

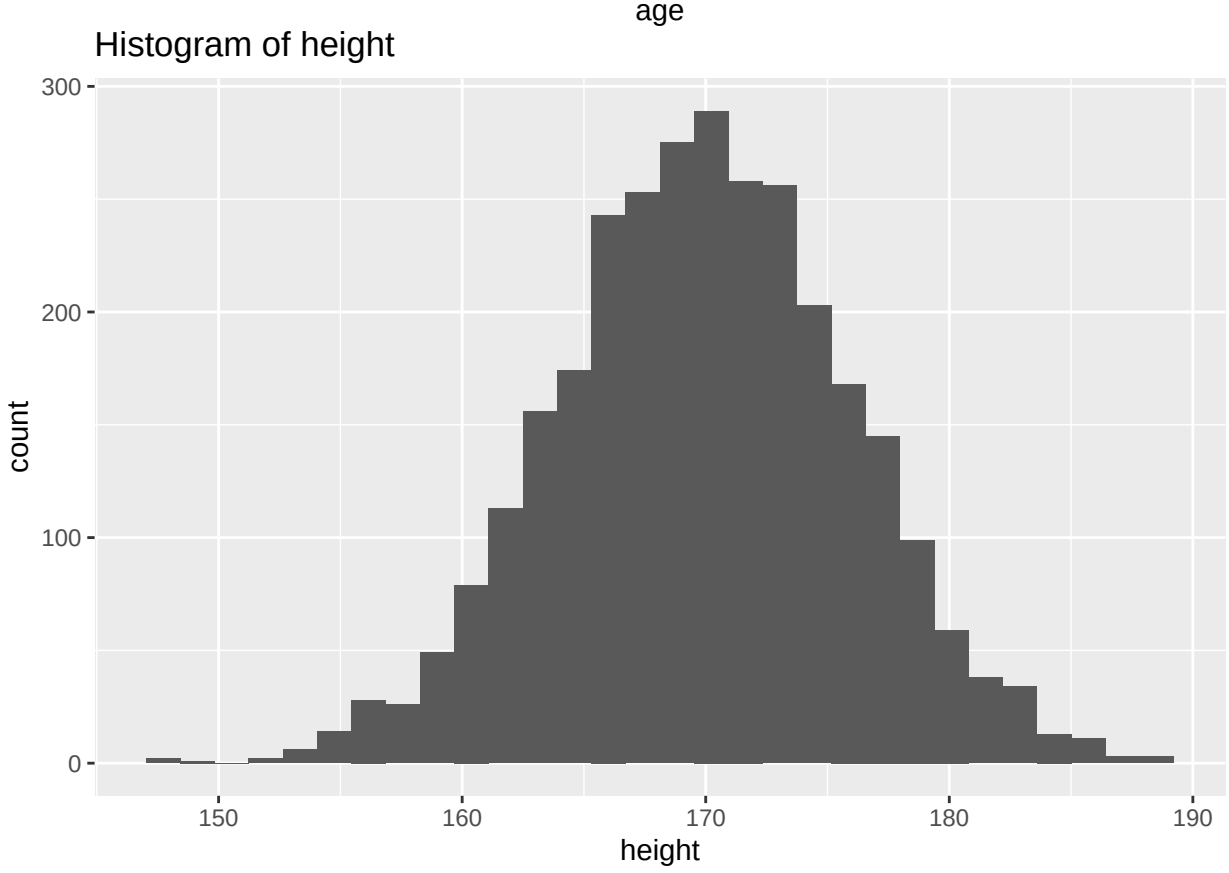
Table 1: Summary of Dataset

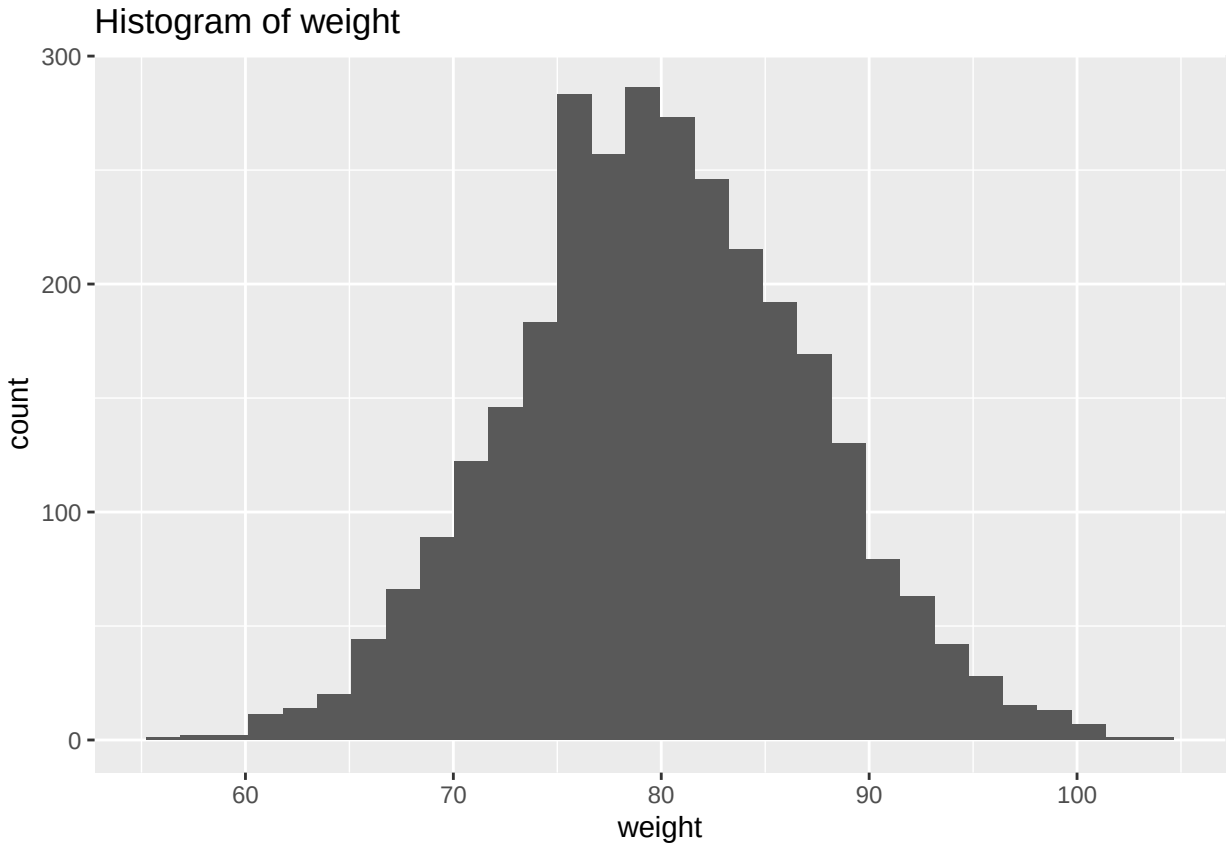
Characteristic	**N = 3,000**[1]
age	60.0 (57.0, 63.0)
gender	
male	1,544 (51%)
female	1,456 (49%)
race	
White	1,967 (66%)
Asian	158 (5.3%)
Black	604 (20%)
Hispanic	271 (9.0%)
smoking	
Never smoked	1,822 (61%)
Former smoker	859 (29%)
Current smoker	319 (11%)
height	169.9 (166.0, 173.9)
weight	80 (75, 85)
bmi	27.65 (25.80, 29.50)
hypertension	1,492 (50%)
diabetes	463 (15%)
SBP	130 (125, 136)
LDL	110 (97, 124)
vaccine	
Not vaccinated	1,212 (40%)
Vaccinated	1,788 (60%)
severity	
Not severe	2,679 (89%)
Severe	321 (11%)
study	
A	2,000 (67%)
B	1,000 (33%)
recovery_time	39 (31, 49)

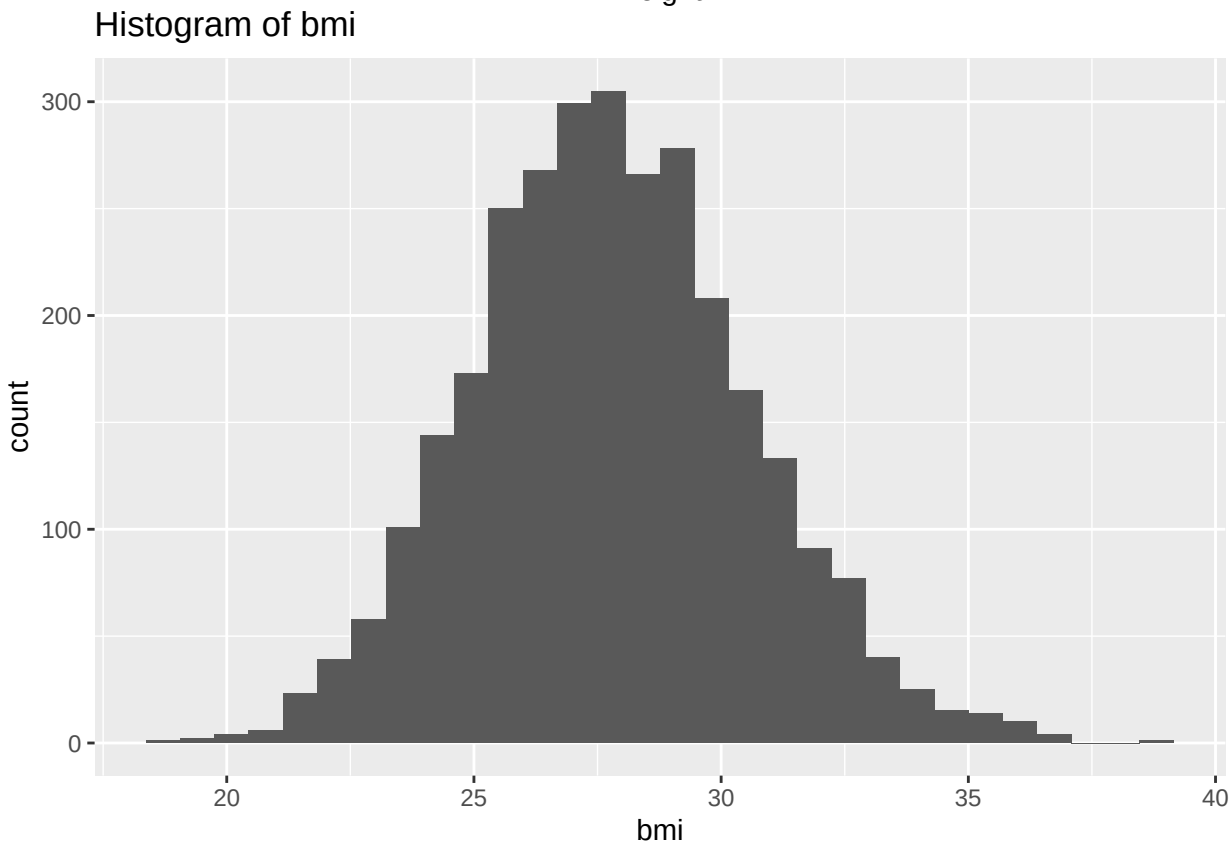
[1] Median (IQR); n (%)

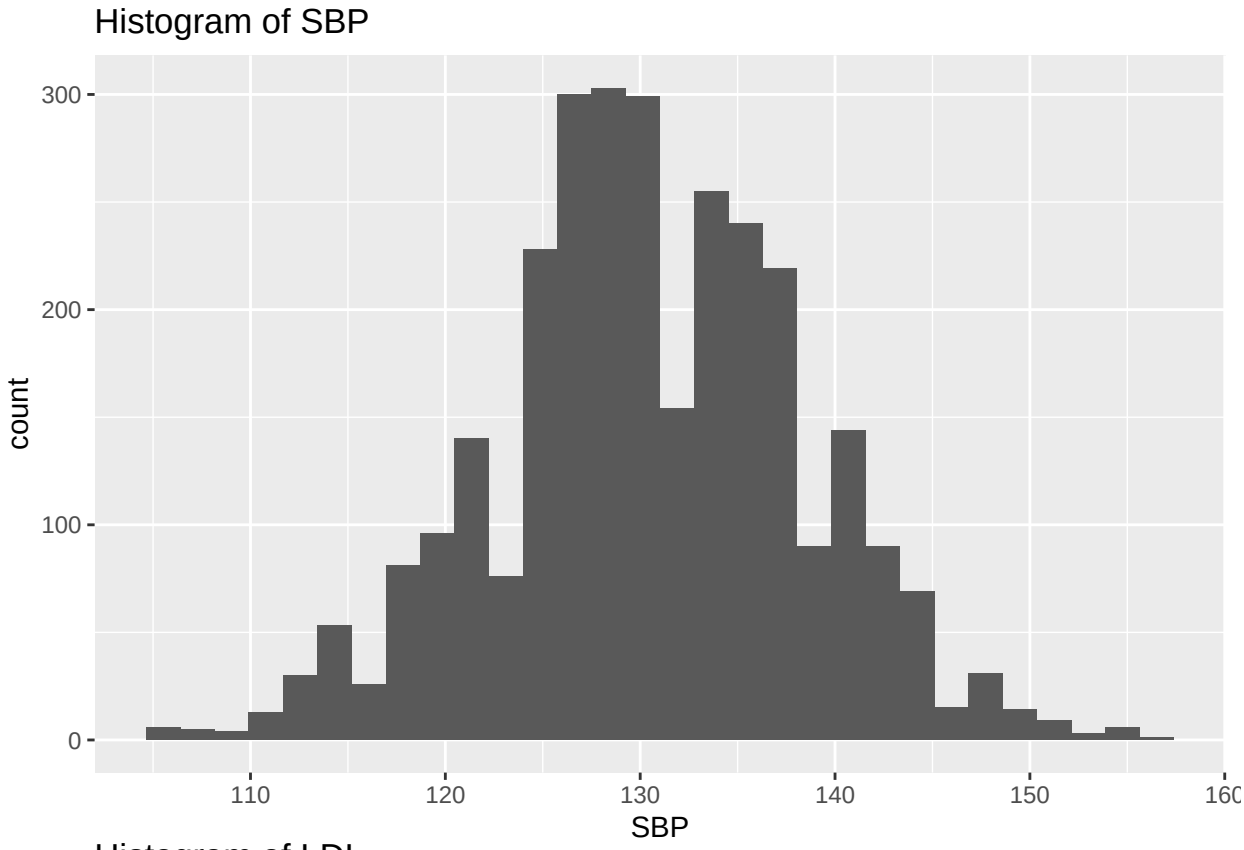
Visualizations for the numerical variables

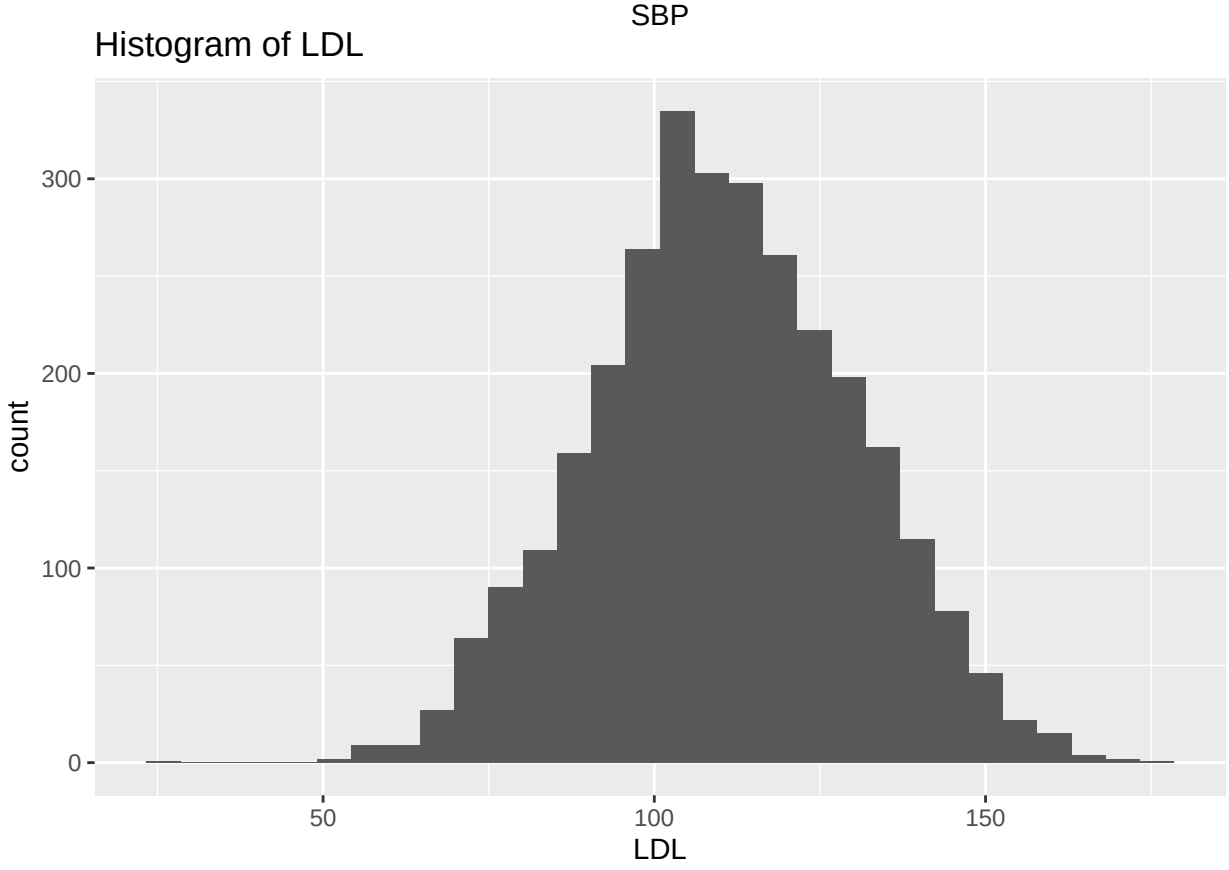


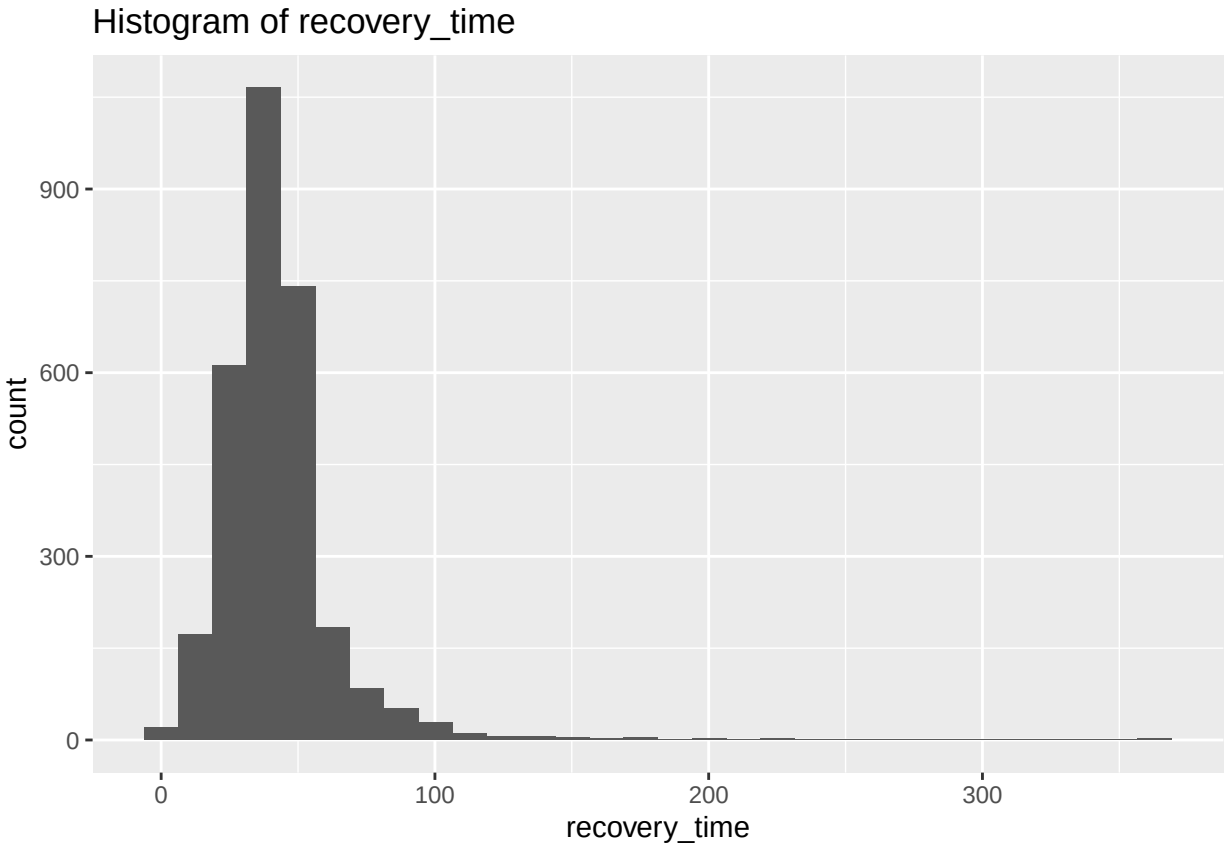



Histogram of weight
count
weight
300
200
100
0
60
70
80
90
100

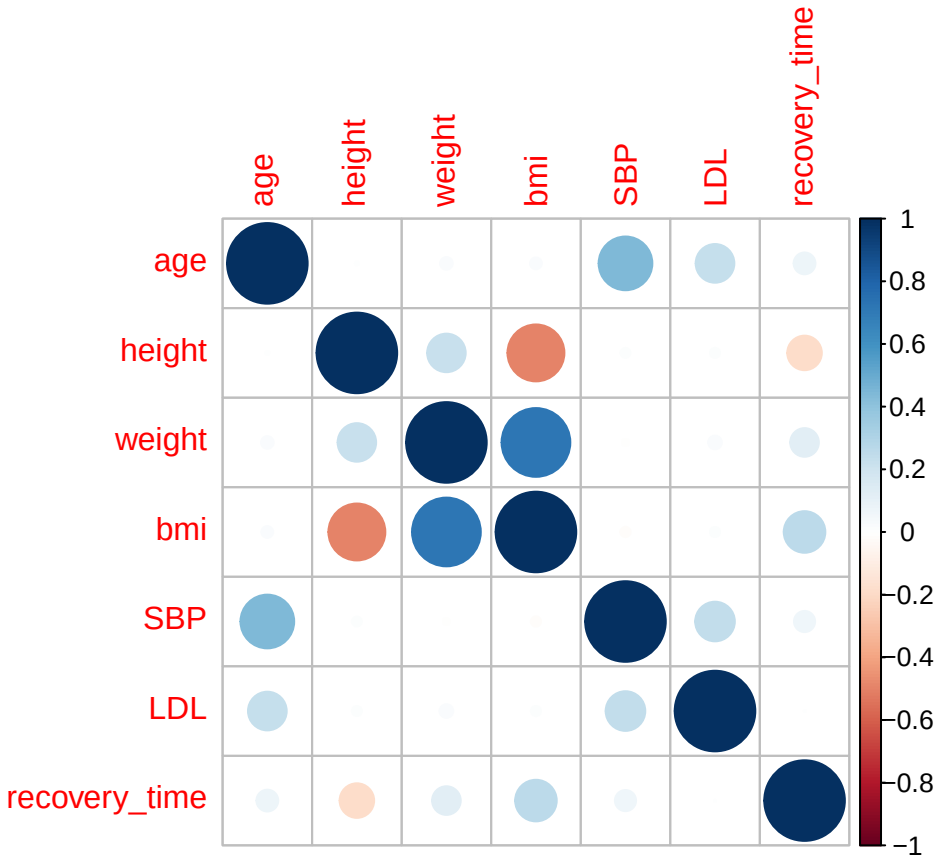

Histogram of bmi
count
bmi
300
200
100
0
20
25
30
35
40


Histogram of SBP
count
300
200
100
0
110
120
130
140
150
160
SBP


Histogram of LDL
count
300
200
100
0
50
100
150
LDL



correlation plot



Model training

In this section, describe the models you used to predict the time to recovery from COVID-19. Briefly state the assumptions made by using the models. Provide a detailed description of the model training procedure and how you obtained the final model.

MARS

GAM

lasso

Elastic net

ctrl1 <- trainControl(method = "cv", number = 10) tuneGrid = expand.grid(alpha = seq(0, 1, length = 21), lambda = exp(seq(-25, 5, length = 100)))