EDA

Chen Liang (cl4469), Xinyi Shang (xs2529), Yuki Joyama (yj2803)

2024-03-20

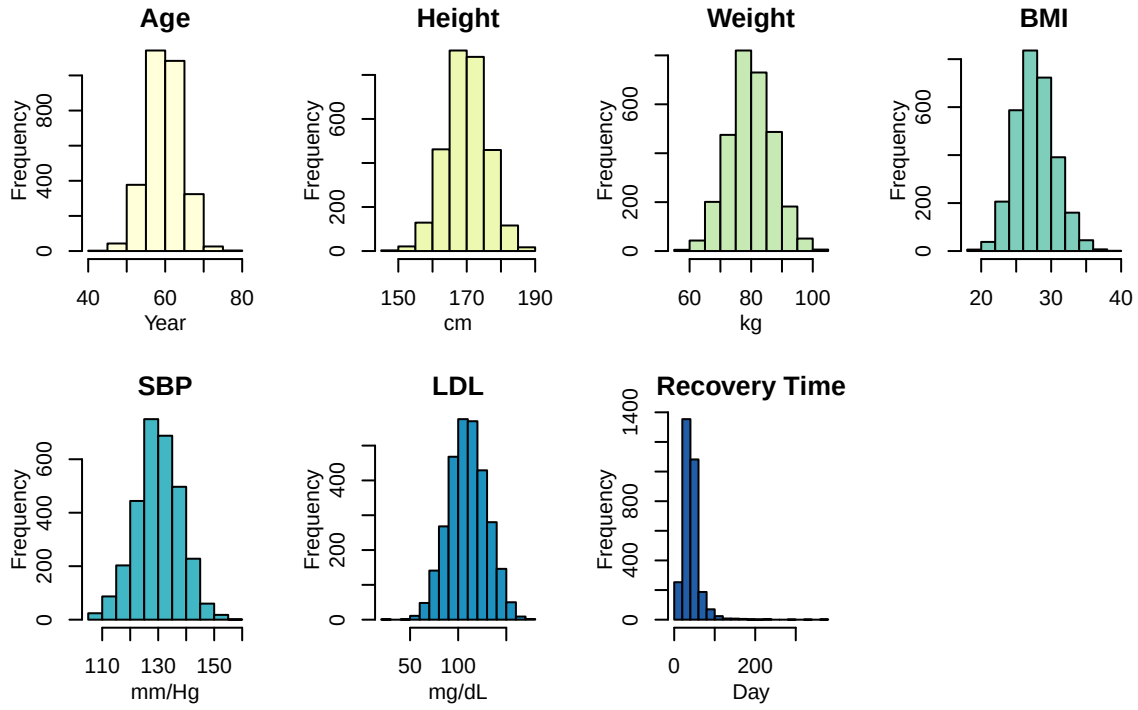
```
# read RData file
df_recov <- get(load("./data/recovery.RData")) |>
  janitor::clean_names()

summary(df_recov)
```

```
##        id              age           gender       race     smoking
##  Min.   :   1.0   Min.   :42.0   Min.   :0.0000   1:1967   0:1822
##  1st Qu.: 750.8   1st Qu.:57.0   1st Qu.:0.0000   2: 158   1: 859
##  Median :1500.5   Median :60.0   Median :0.0000   3: 604   2: 319
##  Mean   :1500.5   Mean   :60.2   Mean   :0.4853   4: 271
##  3rd Qu.:2250.2   3rd Qu.:63.0   3rd Qu.:1.0000
##  Max.   :3000.0   Max.   :79.0   Max.   :1.0000
##      height          weight            bmi         hypertension
##  Min.   :147.8   Min.   : 55.90   Min.   :18.80   Min.   :0.0000
##  1st Qu.:166.0   1st Qu.: 75.20   1st Qu.:25.80   1st Qu.:0.0000
##  Median :169.9   Median : 79.80   Median :27.65   Median :0.0000
##  Mean   :169.9   Mean   : 79.96   Mean   :27.76   Mean   :0.4973
##  3rd Qu.:173.9   3rd Qu.: 84.80   3rd Qu.:29.50   3rd Qu.:1.0000
##  Max.   :188.6   Max.   :103.70   Max.   :38.90   Max.   :1.0000
##     diabetes           sbp             ldl           vaccine
##  Min.   :0.0000   Min.   :105.0   Min.   : 28.0   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:125.0   1st Qu.: 97.0   1st Qu.:0.000
##  Median :0.0000   Median :130.0   Median :110.0   Median :1.000
##  Mean   :0.1543   Mean   :130.5   Mean   :110.5   Mean   :0.596
##  3rd Qu.:0.0000   3rd Qu.:136.0   3rd Qu.:124.0   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :156.0   Max.   :178.0   Max.   :1.000
##     severity        study           recovery_time
##  Min.   :0.000   Length:3000        Min.   :  2.00
##  1st Qu.:0.000   Class :character   1st Qu.: 31.00
##  Median :0.000   Mode  :character   Median : 39.00
##  Mean   :0.107                      Mean   : 42.17
##  3rd Qu.:0.000                      3rd Qu.: 49.00
##  Max.   :1.000                      Max.   :365.00
```

Histogram

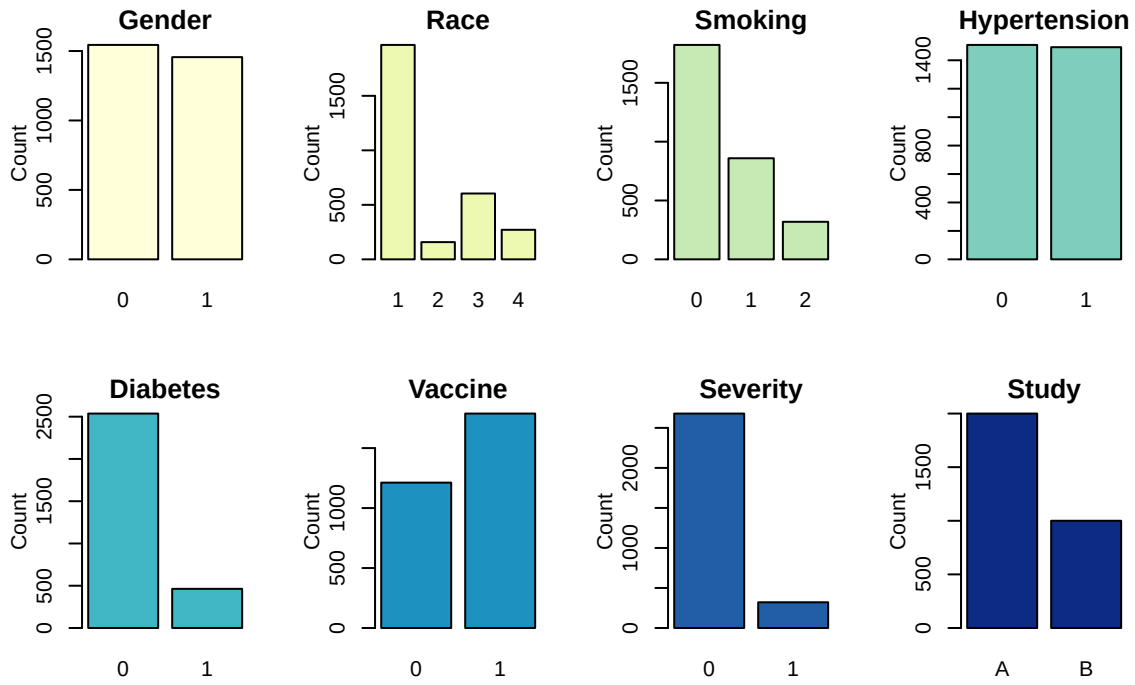
```
cate_recov = df_recov |>
  select(gender, race, smoking, hypertension, diabetes, vaccine, severity, study)
```

```r
conti_recov = df_recov |>
  select(age, height, weight, bmi, sbp, ldl, recovery_time)

#ggplot(gather(conti_recov, cols, value), aes(x = value)) +
#       geom_histogram(binwidth = 20) + facet_grid(.~cols)

#library(Hmisc)
#hist.data.frame(conti_recov)

par(mfrow = c(2, 4),  # Layout: 2 rows, 4 columns
    oma = c(2, 2, 3, 1),  # Outer margins
    mar = c(4, 4, 2, 1),  # Inner margins
    mgp = c(2, 1, 0))     # Margins for labels and title

colors <- brewer.pal(8, "YlGnBu")

# Plot each histogram using a color from the Set3 palette
hist(conti_recov$age, main = "Age", xlab = "Year", ylab = "Frequency", col = colors[1])
hist(conti_recov$height, main = "Height", xlab = "cm", ylab = "Frequency", col = colors[2])
hist(conti_recov$weight, main = "Weight", xlab = "kg", ylab = "Frequency", col = colors[3])
hist(conti_recov$bmi, main = "BMI", xlab = " ", ylab = "Frequency", col = colors[4])
hist(conti_recov$sbp, main = "SBP", xlab = "mm/Hg", ylab = "Frequency", col = colors[5])
hist(conti_recov$ldl, main = "LDL", xlab = "mg/dL", ylab = "Frequency", col = colors[6])
hist(conti_recov$recovery_time, main = "Recovery Time", xlab = "Day", ylab = "Frequency", col = colors[
```



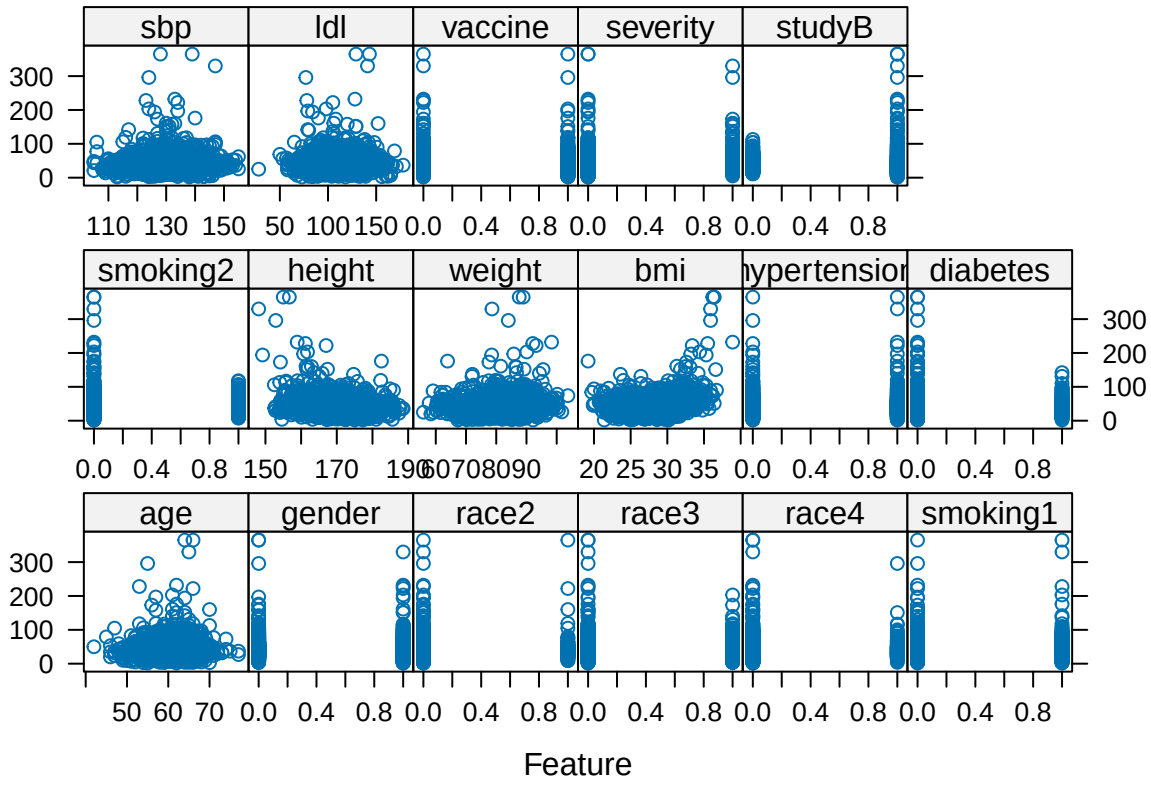
Bar plot

```
cate_recov = df_recov |>
  select(gender, race, smoking, hypertension, diabetes, vaccine, severity, study)
#ggplot(gather(cate_recov, cols, value), aes(x = value)) +
#       geom_bar(binwidth = 20) + facet_grid(.~cols)

# Setting up the plotting area
par(mfrow = c(2, 4),  # Layout: 2 rows, 4 columns
    oma = c(2, 2, 3, 1),  # Outer margins
    mar = c(4, 4, 2, 1),  # Inner margins
    mgp = c(2, 1, 0))     # Margins for labels and title

barplot(table(cate_recov$gender), main = "Gender", ylab = "Count", , col = colors[1])
barplot(table(cate_recov$race), main = "Race", ylab = "Count", , col = colors[2])
barplot(table(cate_recov$smoking), main = "Smoking", ylab = "Count", col = colors[3])
barplot(table(cate_recov$hypertension), main = "Hypertension", ylab = "Count", col = colors[4])
barplot(table(cate_recov$diabetes), main = "Diabetes", ylab = "Count", col = colors[5])
barplot(table(cate_recov$vaccine), main = "Vaccine", ylab = "Count", col = colors[6])
barplot(table(cate_recov$severity), main = "Severity", ylab = "Count", col = colors[7])
barplot(table(cate_recov$study), main = "Study", ylab = "Count", col = colors[8])
```

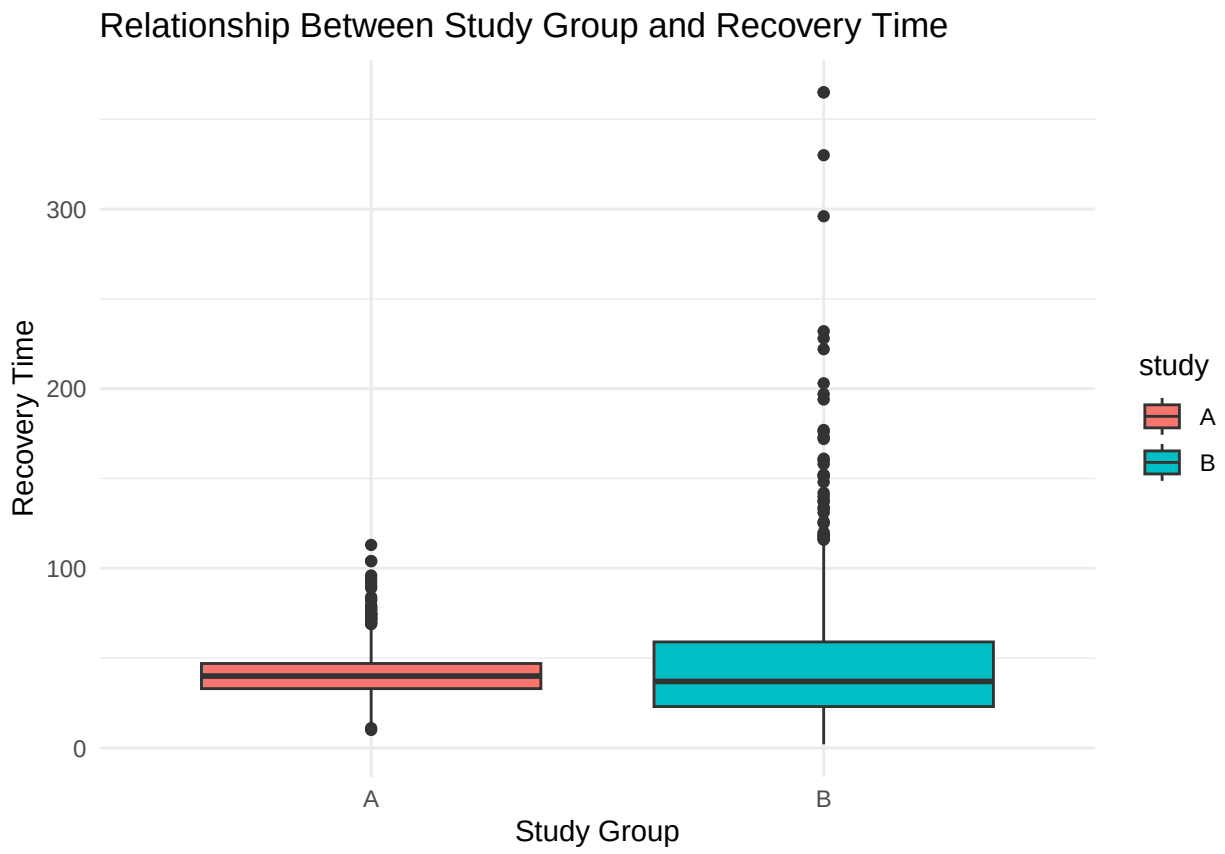



```
data_split = initial_split(df_recov, prop = .80)
train = training(data_split) |>
```

```
select( -id)
test = testing(data_split) |>
select( -id)

# Training data
x = model.matrix(recovery_time~.,train)[, -1]
y = train$recovery_time
# Testing data
x2 <- model.matrix(recovery_time~.,test)[, -1]
y2 <- test$recovery_time

featurePlot(x = x,
            y = y,
            plot = "scatter",
            auto.key = list(columns = 3))
```




```
ggplot(data = df_recov, aes(x = study, y = recovery_time, fill = study)) +
  geom_boxplot() +
  labs(x = "Study Group", y = "Recovery Time",
       title = "Relationship Between Study Group and Recovery Time") +
  theme_minimal()
```



Correlation

```
numeric_df_recov <- df_recov |>
  mutate(race = as.numeric(race)) |>
  mutate(smoking = as.numeric(smoking)) |>
  mutate(study = as.numeric(as.factor(study))) |>
  select_if(is.numeric)


# Compute the correlation matrix
correlation_matrix <- cor(numeric_df_recov)

# Plot the correlation matrix
corrplot(correlation_matrix, method = "circle", type = "upper", order = "hclust")
```

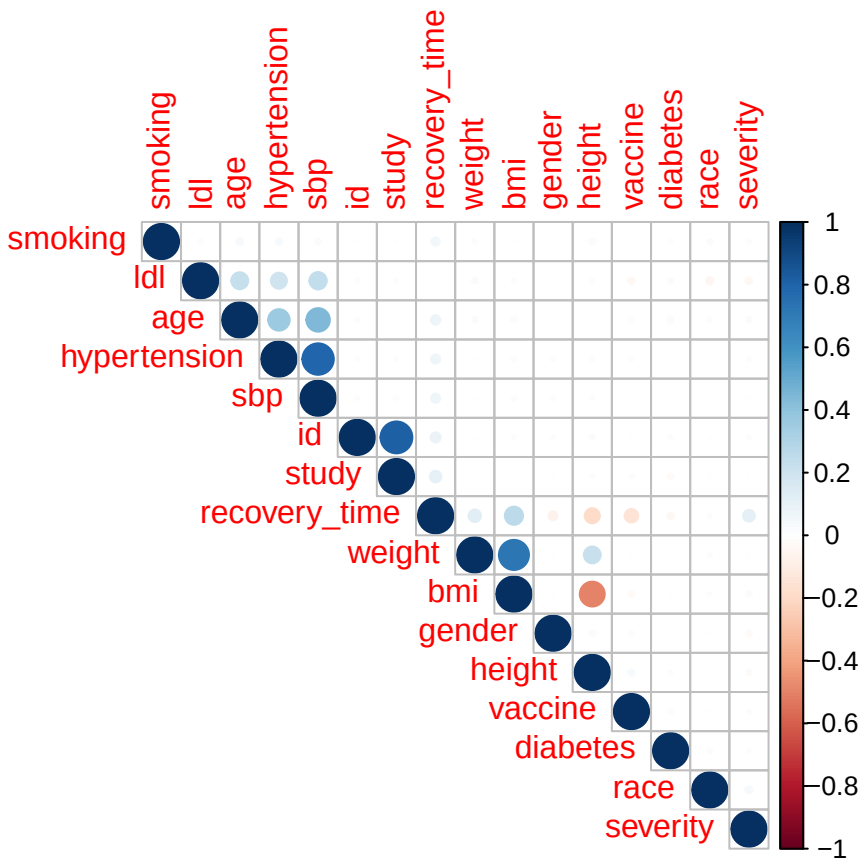


Table 1

```
theme_gtsummary_journal(journal = "nejm")

table_1 = df_recov |>
  select(!id) |>
  mutate(
    gender = case_when(
      gender == 1 ~ "Male",
      gender == 0 ~ "Female"
    ),
    race = case_when(
      race == 1 ~ "White",
      race == 2 ~ "Asian",
      race == 3 ~ "Black",
      race == 4 ~ "Hispanic"
    ),
    smoking = case_when(
      smoking == 0 ~ "Never smoked",
      smoking == 1 ~ "Former smoker",
      smoking == 2 ~ "Current smoker"
    ),
    hypertension = case_when(
```

```
      hypertension == 0 ~ "No hypertension",
      hypertension == 1 ~ "Hypertension"
    ),
    diabetes = case_when(
      diabetes == 0 ~ "No diabetes",
      diabetes == 1 ~ "Diabetes"
    ),
    vaccine = case_when(
      vaccine == 0 ~ "Not vaccinated",
      vaccine == 1 ~ "Vaccinated"
    ),
    severity = case_when(
      severity == 0 ~ "Not severe",
      severity == 1 ~ "Severe"
    )
  ) |>
  tbl_summary(
    by = study,
    statistic = list(
      all_continuous() ~ "{mean} / {median} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    digits = all_continuous() ~ 1,
    label = list(
      age ~ "Age",
      gender ~ "Gender",
      race ~ "Race",
      smoking ~ "Smoking",
      height ~ "Height",
      weight ~ "Weight",
      bmi ~ "BMI",
      hypertension ~ "Hypertension",
      diabetes ~ "Diabetes",
      sbp ~ "SBP",
      ldl ~ "LDL",
      vaccine ~ "Vaccine",
      severity ~ "Severity",
      recovery_time ~ "Recovery time"
    )
  ) |>
  # modify_caption("Table 1: Baseline Characteristics") |>
  as_flex_table() |>
  line_spacing(space = 0, part = "body")

table_1
```

Characteristic	**A**, N = 2,000[1]	**B**, N = 1,000[1]
Age	60.2 / 60.0 (4.5)	60.2 / 60.0 (4.4)
Gender		
Female	1,036 (52%)	508 (51%)

[1]Mean / Median (SD); n (%)

Characteristic	**A**, N = 2,000[1]	**B**, N = 1,000[1]
Male	964 (48%)	492 (49%)
Race		
Asian	108 (5.4%)	50 (5.0%)
Black	408 (20%)	196 (20%)
Hispanic	172 (8.6%)	99 (9.9%)
White	1,312 (66%)	655 (66%)
Smoking		
Current smoker	218 (11%)	101 (10%)
Former smoker	557 (28%)	302 (30%)
Never smoked	1,225 (61%)	597 (60%)
Height	169.9 / 169.9 (5.9)	170.0 / 170.0 (6.0)
Weight	79.9 / 79.6 (7.1)	80.0 / 80.0 (7.2)
BMI	27.8 / 27.7 (2.8)	27.8 / 27.6 (2.8)
Hypertension		
Hypertension	1,002 (50%)	490 (49%)
No hypertension	998 (50%)	510 (51%)
Diabetes		
Diabetes	322 (16%)	141 (14%)
No diabetes	1,678 (84%)	859 (86%)
SBP	130.6 / 131.0 (8.0)	130.3 / 130.0 (7.9)
LDL	110.3 / 110.0 (19.8)	110.7 / 110.0 (19.8)
Vaccine		
Not vaccinated	797 (40%)	415 (42%)
Vaccinated	1,203 (60%)	585 (59%)
Severity		
Not severe	1,785 (89%)	894 (89%)
Severe	215 (11%)	106 (11%)
Recovery time	40.4 / 40.0 (11.2)	45.7 / 37.0 (36.6)

[1]Mean / Median (SD); n (%)