Build Prediction Model

Minghe Wang

2025-03-26

Exploratory Data Analysis

```
load("./data/dat1.RData")
load("./data/dat2.RData")

# no missing data
all(is.na(dat1))
```

```
## [1] FALSE
```

```
all(is.na(dat2))
```

```
## [1] FALSE
```

```
ifelse(all(names(dat1) == names(dat2)), "train and test data have same structure", "train and test data
```

```
## [1] "train and test data have same structure"
```

```
str(dat1)
```

```
## 'data.frame':    5000 obs. of  14 variables:
##  $ id          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age         : num  50 71 58 63 56 59 67 62 60 64 ...
##  $ gender      : int  0 1 1 0 1 1 0 1 0 1 ...
##  $ race        : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 3 4 1 4 1 ...
##  $ smoking     : Factor w/ 3 levels "0","1","2": 1 1 2 1 1 1 1 1 1 1 ...
##  $ height      : num  176 176 169 167 163 ...
##  $ weight      : num  68.3 69.6 76.9 90 83.9 86.8 91.4 87.7 85.7 76.6 ...
##  $ bmi         : num  22 22.6 27 32.1 31.7 30.8 29.7 28.1 29 31.5 ...
##  $ diabetes    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hypertension: num  0 1 0 1 0 1 1 0 0 1 ...
##  $ SBP         : num  130 149 127 138 123 132 133 130 129 134 ...
##  $ LDL         : num  82 129 101 93 97 108 89 96 120 135 ...
##  $ time        : num  76 82 168 105 193 143 63 78 61 88 ...
##  $ log_antibody: num  10.65 9.89 10.9 9.91 9.56 ...
```

Univariate analysis(continous & categorical)

```
continuous_var <- dat1 %>%
  select(age, height, weight, bmi, SBP, LDL, time, log_antibody)

categorical_var <- dat1 %>%
  select(gender, race, smoking, diabetes, hypertension) %>%
  mutate(
```

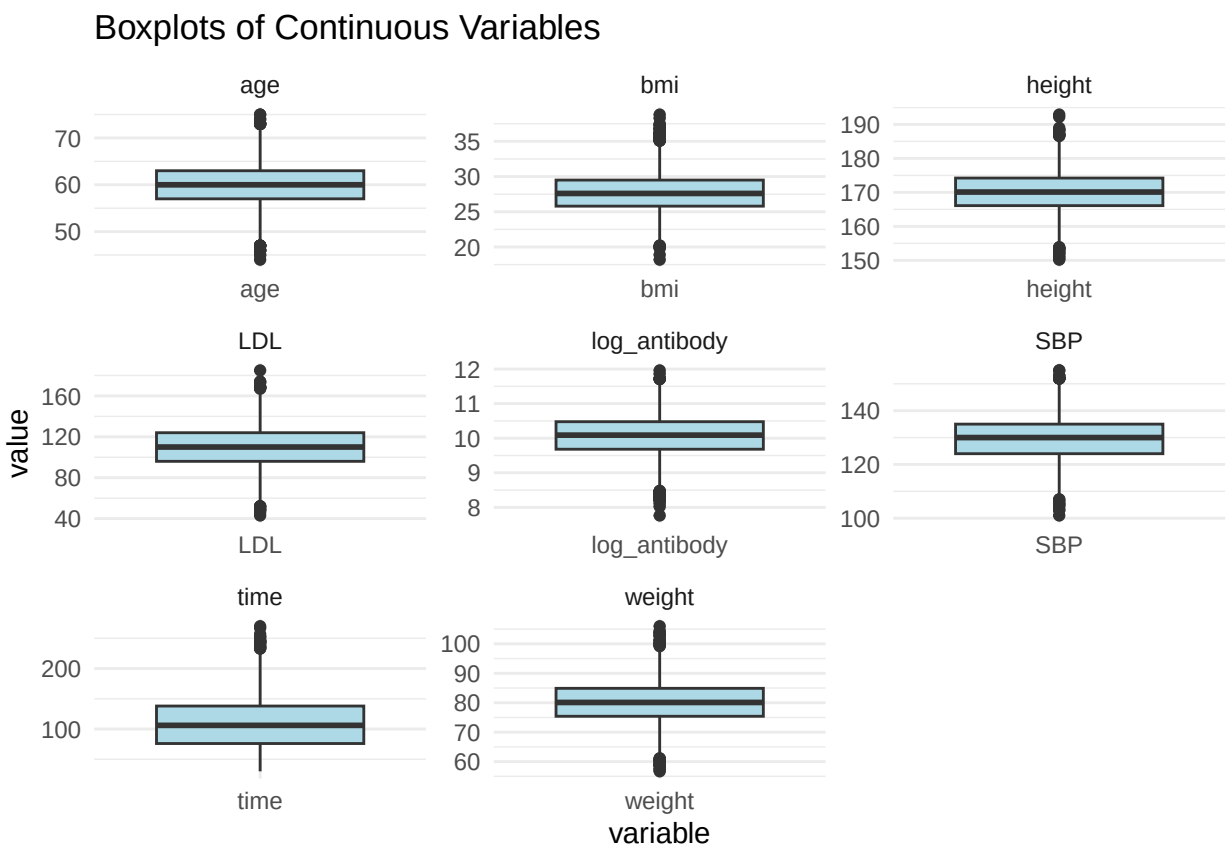
```
    # Convert binary variables to factors with labels
    gender = factor(gender, levels = c(0, 1), labels = c("Female", "Male")),
    diabetes = factor(diabetes, levels = c(0, 1), labels = c("No", "Yes")),
    hypertension = factor(hypertension, levels = c(0, 1), labels = c("No", "Yes"))
  )

# Continuous:
summary(continuous_var)
```

```
##       age             height          weight            bmi
##  Min.   :44.00   Min.   :150.2   Min.   : 56.70   Min.   :18.20
##  1st Qu.:57.00   1st Qu.:166.1   1st Qu.: 75.40   1st Qu.:25.80
##  Median :60.00   Median :170.1   Median : 80.10   Median :27.60
##  Mean   :59.97   Mean   :170.1   Mean   : 80.11   Mean   :27.74
##  3rd Qu.:63.00   3rd Qu.:174.2   3rd Qu.: 84.90   3rd Qu.:29.50
##  Max.   :75.00   Max.   :192.9   Max.   :106.00   Max.   :38.80
##       SBP              LDL              time        log_antibody
##  Min.   :101.0   Min.   : 43.0   Min.   : 30.0   Min.   : 7.765
##  1st Qu.:124.0   1st Qu.: 96.0   1st Qu.: 76.0   1st Qu.: 9.682
##  Median :130.0   Median :110.0   Median :106.0   Median :10.089
##  Mean   :129.9   Mean   :109.9   Mean   :108.9   Mean   :10.064
##  3rd Qu.:135.0   3rd Qu.:124.0   3rd Qu.:138.0   3rd Qu.:10.478
##  Max.   :155.0   Max.   :185.0   Max.   :270.0   Max.   :11.961
```

```
# Boxplots
continuous_var_long <- continuous_var %>%
  tidyr::pivot_longer(cols = everything(), names_to = "variable", values_to = "value")

ggplot(continuous_var_long, aes(x = variable, y = value)) +
  geom_boxplot(fill = "lightblue") +
  facet_wrap(~variable, scales = "free", ncol = 3) +
  theme_minimal() +
  labs(title = "Boxplots of Continuous Variables")
```

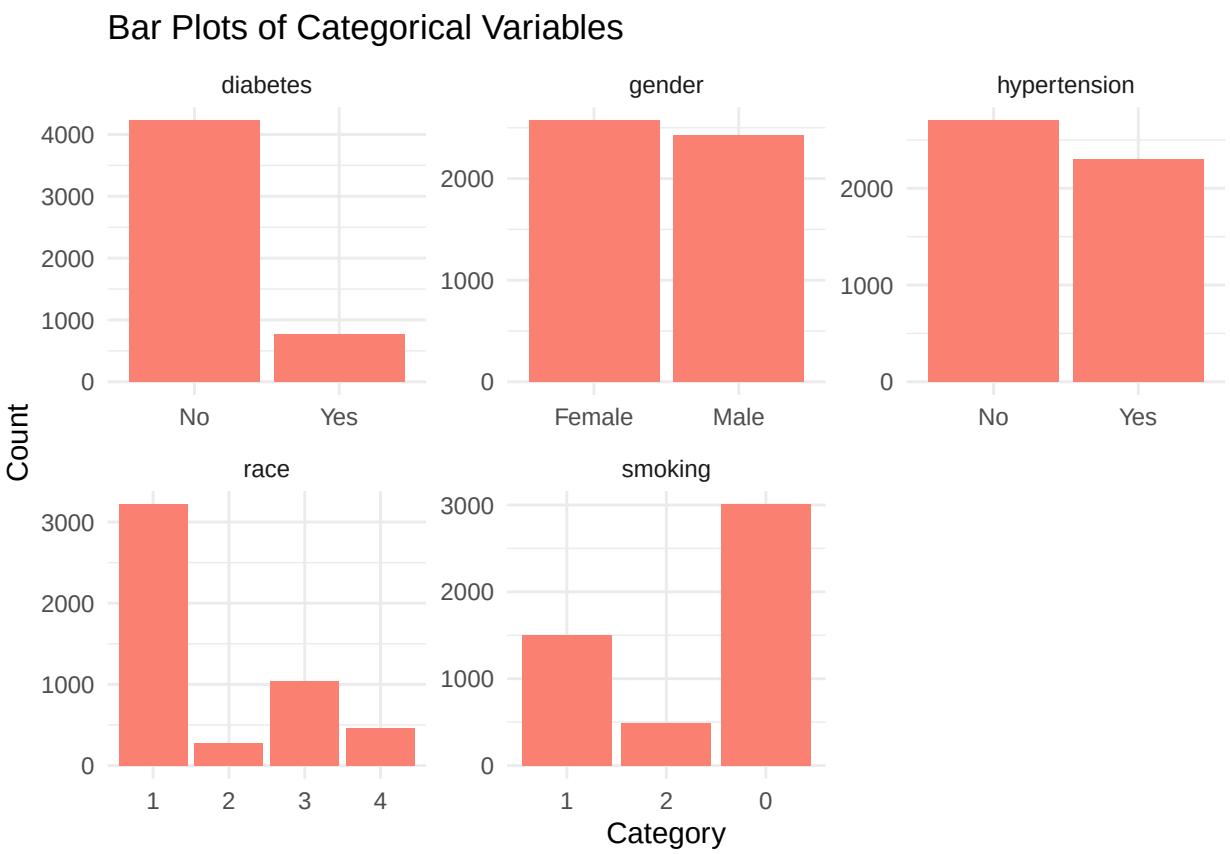



```
# Categorical:
summary(continuous_var)
```

```
##       age             height          weight           bmi
##  Min.   :44.00   Min.   :150.2   Min.   : 56.70   Min.   :18.20
##  1st Qu.:57.00   1st Qu.:166.1   1st Qu.: 75.40   1st Qu.:25.80
##  Median :60.00   Median :170.1   Median : 80.10   Median :27.60
##  Mean   :59.97   Mean   :170.1   Mean   : 80.11   Mean   :27.74
##  3rd Qu.:63.00   3rd Qu.:174.2   3rd Qu.: 84.90   3rd Qu.:29.50
##  Max.   :75.00   Max.   :192.9   Max.   :106.00   Max.   :38.80
##       SBP              LDL             time        log_antibody
##  Min.   :101.0   Min.   : 43.0   Min.   : 30.0   Min.   : 7.765
##  1st Qu.:124.0   1st Qu.: 96.0   1st Qu.: 76.0   1st Qu.: 9.682
##  Median :130.0   Median :110.0   Median :106.0   Median :10.089
##  Mean   :129.9   Mean   :109.9   Mean   :108.9   Mean   :10.064
##  3rd Qu.:135.0   3rd Qu.:124.0   3rd Qu.:138.0   3rd Qu.:10.478
##  Max.   :155.0   Max.   :185.0   Max.   :270.0   Max.   :11.961
```

```
# bar plots
categorical_var_long <- categorical_var %>%
  tidyr::pivot_longer(cols = everything(), names_to = "variable", values_to = "value")

ggplot(categorical_var_long, aes(x = value)) +
  geom_bar(fill = "salmon") +
  facet_wrap(~variable, scales = "free", ncol = 3) +
  theme_minimal() +
  labs(title = "Bar Plots of Categorical Variables", x = "Category", y = "Count")
```



According to the box plot for continuous variables:

- `Age`, `BMI`, and `SBP` appear reasonably normally distributed, with expected ranges for an adult population; `LDL` cholesterol and `time` since vaccination show a wider range and some outliers, which may impact linear models.
- `log_antibody` (response) appears fairly symmetrical, which supports its use as a continuous response in linear or GAM models.
- Correlations and non-linear trends should be assessed in the next step to guide model form.

According to the bar plot for categorical variables:

- `Gender` is fairly balanced between Female and Male;
- `Race` is skewed, with a majority of participants identifying as White (Category 1). Other racial/ethnic groups are underrepresented;
- `Smoking` status shows that the majority are never smokers (Category 0), with fewer current and former smokers;
- A large proportion of participants do not have `diabetes`;
- A moderate split exists for `hypertension`, which may contribute meaningfully to clinical outcome variation
- Demographically, the population is balanced by gender but skewed by race and smoking status.

Overall, we believe the response variable `log_antibody` is well-behaved, and further correlation analysis(eg. bivariate) is needed.

Model Training