P8106 Midterm Code

Group 2: Kate Colvin (KAC2301), Jeong Yun (Lizy) Choi (JC6452), and

Exploratory Analysis

Loading in data files

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

dat1 <- dat1 %>% janitor::clean_names()
dat2 <- dat2 %>%janitor::clean_names()
```

Producing summary table

Notes

Training and test data have the same distribution of demographic characteristics; there is a difference in time since vaccination between training and test data

```
# Combining data for summary table, data cleaning
dat1_com <- dat1 %>% mutate(set = "Training Data")
dat2_com <- dat2 %>% mutate(set = "Testing Data")

dat <- dat1_com %>%
  rbind(dat2_com) %>%
  rename(days_vaccinated = time) %>%
  mutate(race = as.character(race),
         smoking = as.character(smoking)) %>%
  mutate(race = case_match(
        race,
          "1" ~ "White",
          "2" ~ "Asian",
          "3" ~ "Black",
          "4" ~ "Hispanic"),
         gender = case_match(
          gender,
          1 ~ "Male",
          0 ~ "Female"),
         smoking = case_match(
          smoking,
          "0" ~ "Never",
          "1" ~ "Former",
          "2" ~ "Current"))

# Summary table
dat %>% select(!id) %>%
```

Table 1: Summary of Patient Testing and Training Data (N=6000)

Characteristic	**Overall** N = 6,000[1]	**Testing Data** N = 1,000[1]	**Training Data** N = 5,000[1]
Age	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)
Gender			
Female	3,082 (51%)	509 (51%)	2,573 (51%)
Male	2,918 (49%)	491 (49%)	2,427 (49%)
Race			
Asian	333 (5.6%)	55 (5.5%)	278 (5.6%)
Black	1,235 (21%)	199 (20%)	1,036 (21%)
Hispanic	548 (9.1%)	83 (8.3%)	465 (9.3%)
White	3,884 (65%)	663 (66%)	3,221 (64%)
Smoking			
Current	589 (9.8%)	103 (10%)	486 (9.7%)
Former	1,800 (30%)	296 (30%)	1,504 (30%)
Never	3,611 (60%)	601 (60%)	3,010 (60%)
Height (cm)	170.1 (166.1, 174.2)	170.2 (166.1, 174.2)	170.1 (166.1, 174.3)
Weight (kg)	80 (75, 85)	80 (75, 84)	80 (75, 85)
BMI	27.60 (25.80, 29.50)	27.60 (25.80, 29.60)	27.60 (25.80, 29.50)
Diabetes	929 (15%)	157 (16%)	772 (15%)
Hypertension	2,754 (46%)	456 (46%)	2,298 (46%)
Systolic Blood Pressure (mmHg)	130 (124, 135)	130 (124, 135)	130 (124, 135)
LDL Cholesterol (mg/dL)	110 (96, 124)	112 (96, 124)	110 (96, 124)
Time Since Vaccinated (days)	116 (82, 152)	171 (140, 205)	106 (76, 138)
Log-Transformed Antibody Level	10.06 (9.65, 10.45)	9.93 (9.50, 10.32)	10.09 (9.68, 10.48)

[1] Median (Q1, Q3); n (%)

```r
tbl_summary(
  by = set,
  label = list(age = "Age",
               gender = "Gender",
               race = "Race",
               smoking = "Smoking",
               height = "Height (cm)",
               weight = "Weight (kg)",
               bmi = "BMI",
               diabetes = "Diabetes",
               hypertension = "Hypertension",
               sbp = "Systolic Blood Pressure (mmHg)",
               ldl = "LDL Cholesterol (mg/dL)",
               days_vaccinated = "Time Since Vaccinated (days)",
               log_antibody = "Log-Transformed Antibody Level")) %>%
add_overall() %>%
modify_caption("Summary of Patient Testing and Training Data (N=6000)") %>%
as_gt() %>%
tab_options(table.font.size = 12)
```

Correlation matrix of numerical variables

Model Training

Results