P8106 Midterm - Code

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

Exploratory Analysis

Loading in Data

```
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 and log-transformed antibody levels 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) %>%
  tbl_summary(
```

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]	**p-value**[2]
Age	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)	0.9
Gender				0.7
Female	3,082 (51%)	509 (51%)	2,573 (51%)	
Male	2,918 (49%)	491 (49%)	2,427 (49%)	
Race				0.6
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				0.8
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)	0.7
Weight (kg)	80 (75, 85)	80 (75, 84)	80 (75, 85)	0.8
BMI	27.60 (25.80, 29.50)	27.60 (25.80, 29.60)	27.60 (25.80, 29.50)	0.9
Diabetes	929 (15%)	157 (16%)	772 (15%)	0.8
Hypertension	2,754 (46%)	456 (46%)	2,298 (46%)	0.8
Systolic Blood Pressure (mmHg)	130 (124, 135)	130 (124, 135)	130 (124, 135)	0.3
LDL Cholesterol (mg/dL)	110 (96, 124)	112 (96, 124)	110 (96, 124)	0.4
Time Since Vaccinated (days)	116 (82, 152)	171 (140, 205)	106 (76, 138)	<0.001
Log-Transformed Antibody Level	10.06 (9.65, 10.45)	9.93 (9.50, 10.32)	10.09 (9.68, 10.48)	<0.001

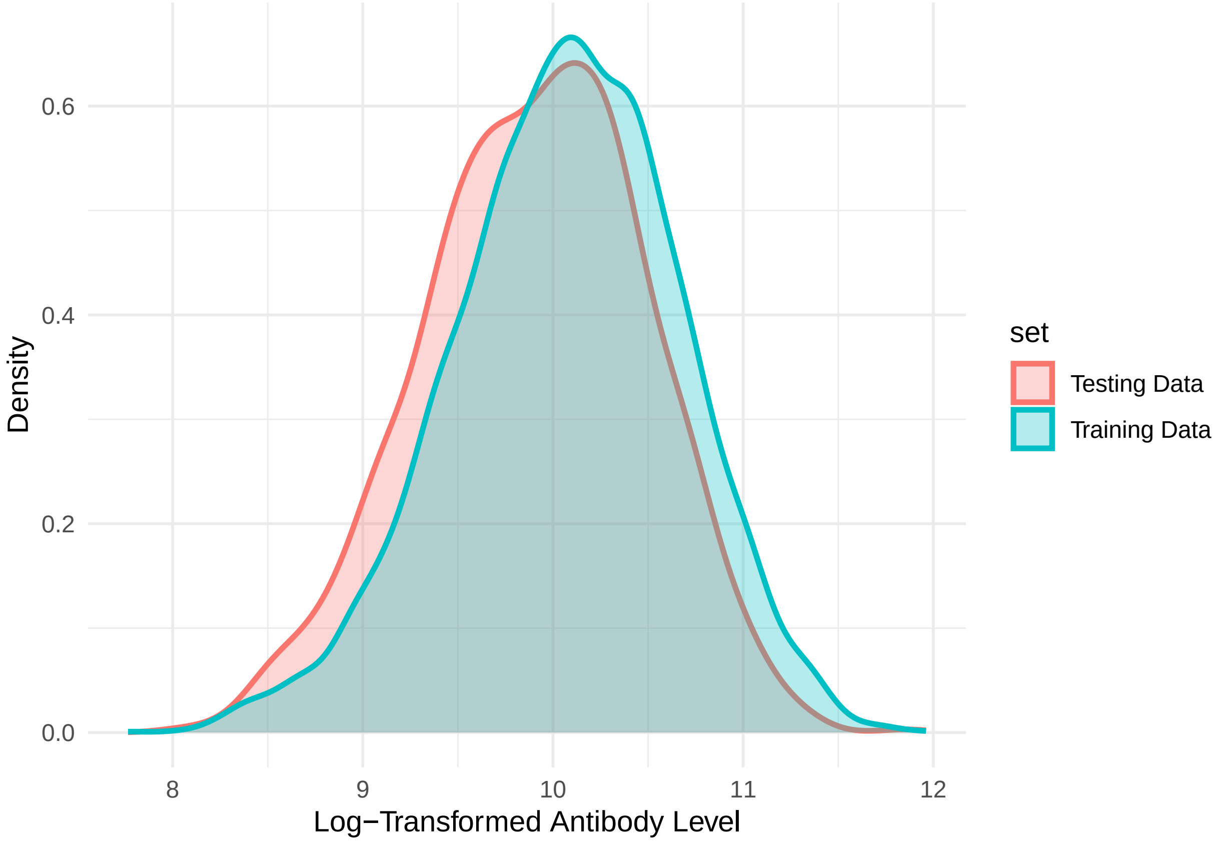
[1] Median (Q1, Q3); n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

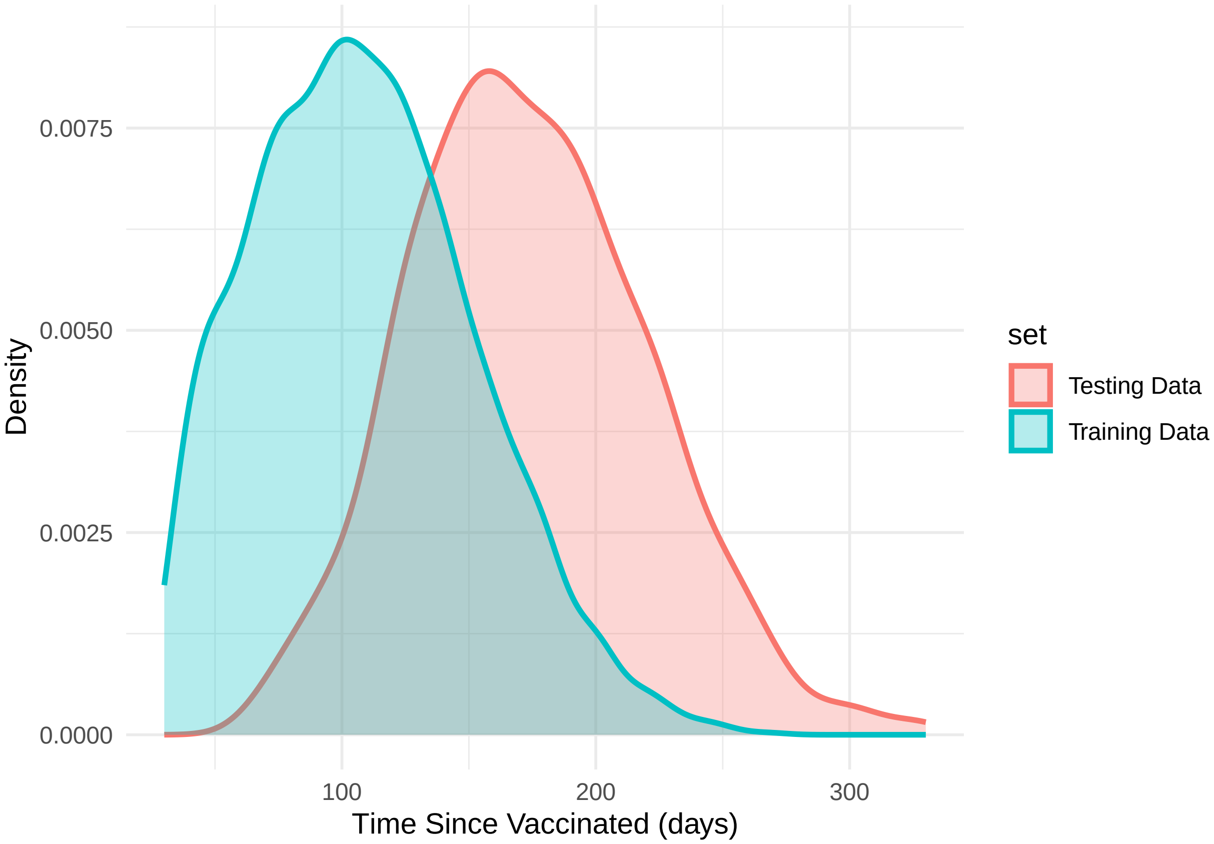
```
    by = set,
    label = list(age = "Age",
                 gender = "Gender",
                 race = "Race",
                 smoking = "Smoking",
                 height = "Height (cm)",
                 weight = "Weight (kg)",
                 bmi = "BMI",
                 diabetes = "Diabetes",
                 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() %>%
  add_p() %>%
  modify_caption("Summary of Patient Testing and Training Data (N=6000)") %>%
  as_gt() %>%
  tab_options(table.font.size = 10)
```

Histograms of Differing Variables by Training and Test Set

```
# Antibody level
plot_sets <- dat %>%
  ggplot(aes(x = log_antibody, fill = set, color = set)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  labs(x = "Log-Transformed Antibody Level", y = "Density") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
plot_sets
```



```
# Time since vaccination (days)
plot_days <- dat %>%
  ggplot(aes(x = days_vaccinated, fill = set, color = set)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  labs(x = "Time Since Vaccinated (days)", y = "Density") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))

plot_days
```



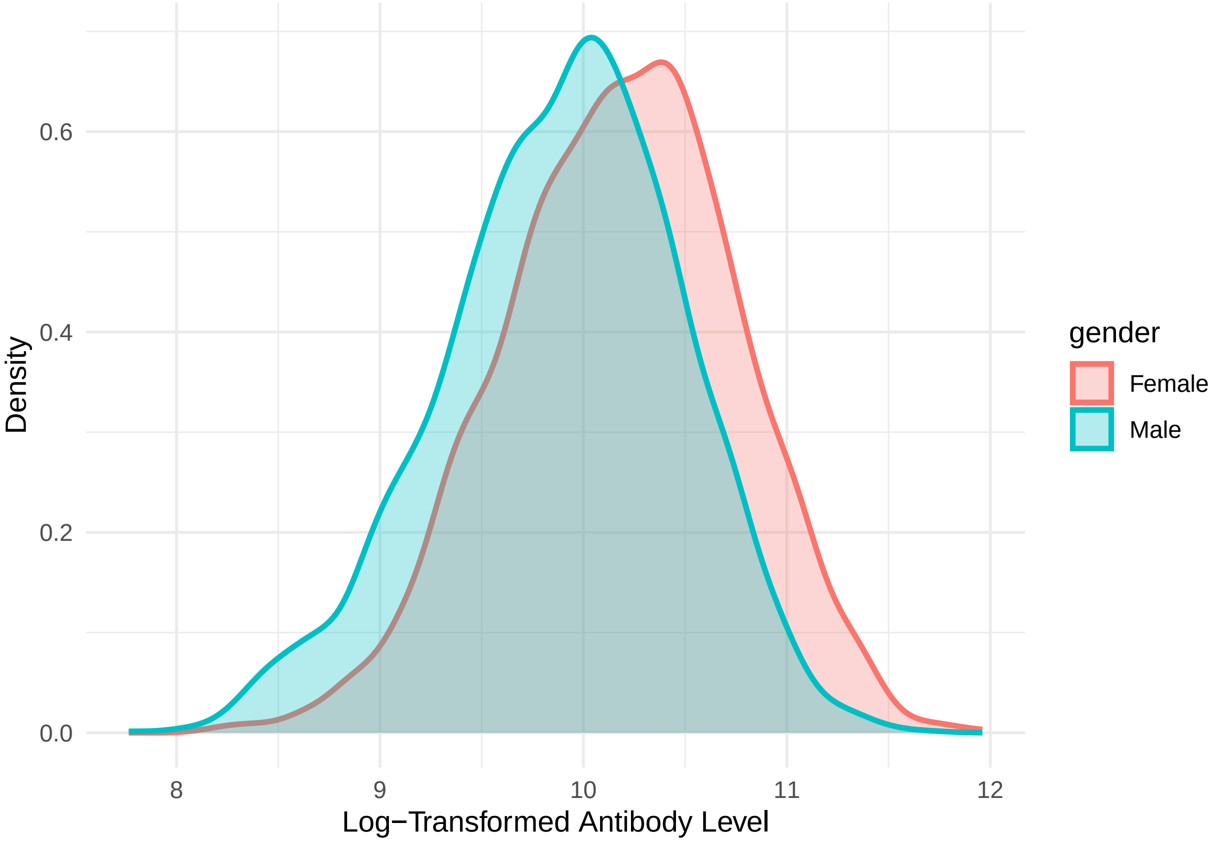
Plots of Log(Antibody), by Categorical Variables

```
# Antibody level, by gender
plot_gender <- dat %>%
  ggplot(aes(x = log_antibody, fill = gender, color = gender)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  labs(x = "Log-Transformed Antibody Level", y = "Density") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
plot_gender
```

Table 2: Log-Transformed Antibody Level, by Gender

Characteristic	**Overall** N = 6,000[1]	**Female** N = 3,082[1]	**Male** N = 2,918[1]	**p-value**[2]
log_antibody	10.06 (9.65, 10.45)	10.20 (9.79, 10.58)	9.93 (9.51, 10.30)	<0.001

[1] Median (Q1, Q3)
[2] Wilcoxon rank sum test



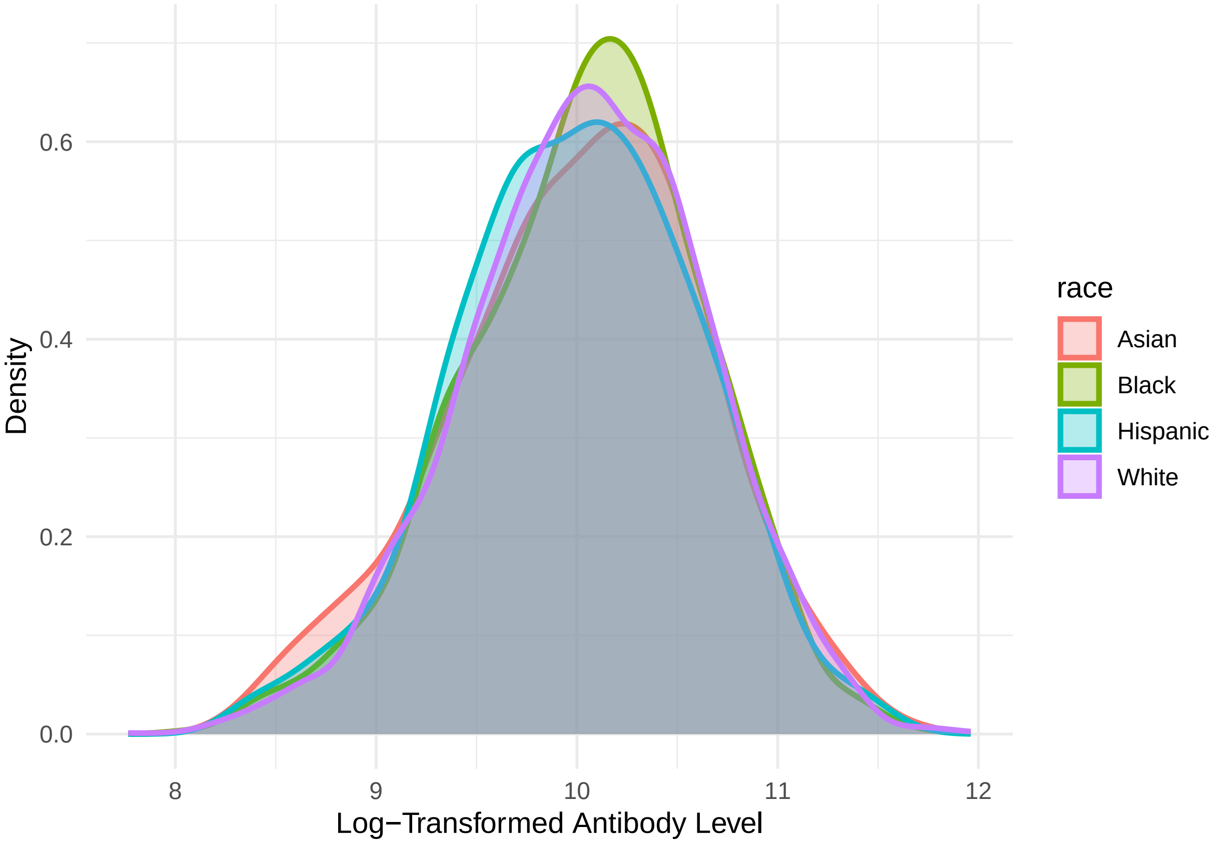
```
dat %>% select(gender, log_antibody) %>%
  tbl_summary(by = gender) %>%
    add_p() %>%
    add_overall() %>%
  modify_caption("Log-Transformed Antibody Level, by Gender") %>%
  as_gt() %>%
  tab_options(table.font.size = 10)
```

```
# Antibody level, by race
plot_race <- dat %>%
  ggplot(aes(x = log_antibody, fill = race, color = race)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  labs(x = "Log-Transformed Antibody Level", y = "Density") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
plot_race
```

Table 3: Log-Transformed Antibody Level, by Race

Characteristic	**Overall** N = 6,000[1]	**Asian** N = 333[1]	**Black** N = 1,235[1]	**Hispanic** N = 548[1]	**White** N = 3,884[1]	**p-value**[2]
log_antibody	10.06 (9.65, 10.45)	10.06 (9.62, 10.44)	10.08 (9.65, 10.44)	10.03 (9.61, 10.42)	10.06 (9.65, 10.46)	0.4

[1] Median (Q1, Q3)
[2] Kruskal-Wallis rank sum test



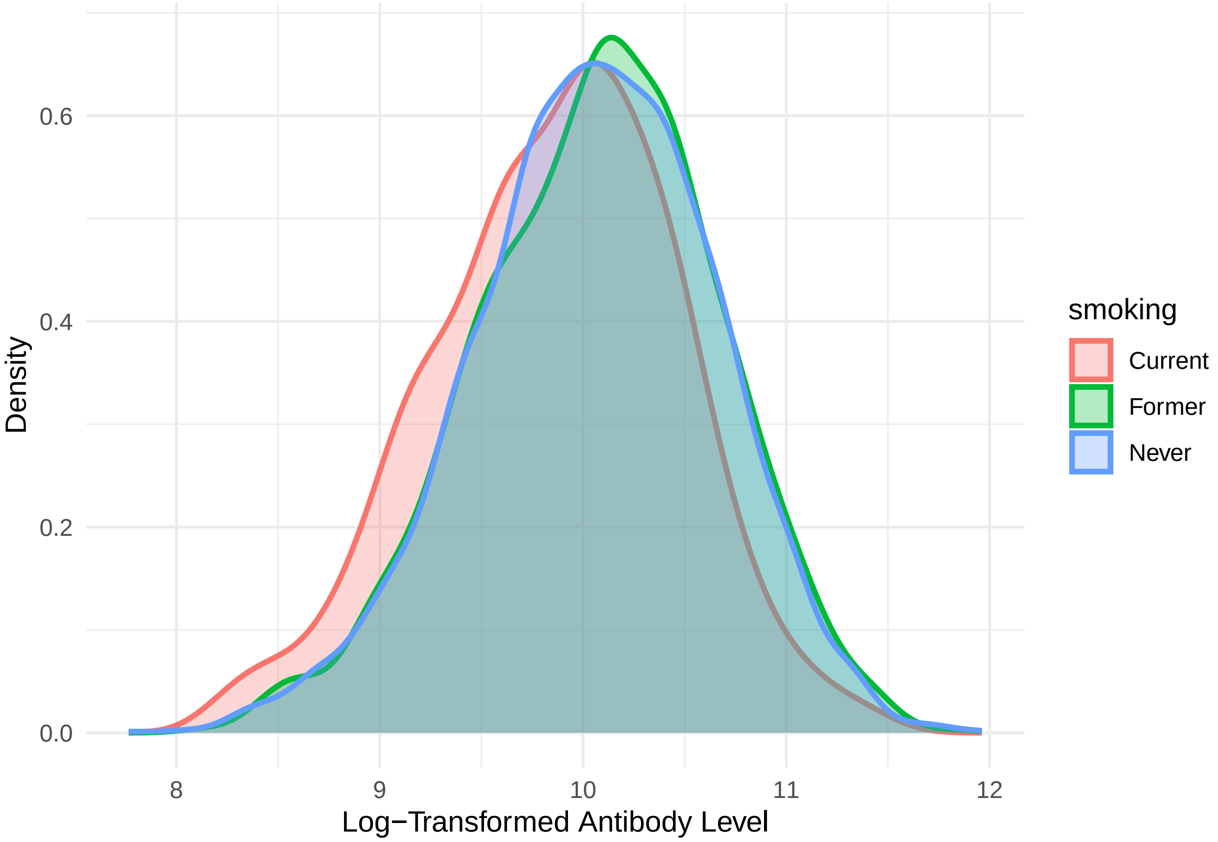
```r
dat %>% select(race, log_antibody) %>%
  tbl_summary(by = race) %>%
    add_p() %>%
    add_overall() %>%
  modify_caption("Log-Transformed Antibody Level, by Race") %>%
  as_gt() %>%
  tab_options(table.font.size = 8)
```

```r
# Antibody level, by smoking status
plot_smoking <- dat %>%
  ggplot(aes(x = log_antibody, fill = smoking, color = smoking)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  labs(x = "Log-Transformed Antibody Level", y = "Density") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
plot_smoking
```

Table 4: Log-Transformed Antibody Level, by Smoking Status

Characteristic	**Overall** N = 6,000[1]	**Current** N = 589[1]	**Former** N = 1,800[1]	**Never** N = 3,611[1]	**p-value**[2]
log_antibody	10.06 (9.65, 10.45)	9.91 (9.46, 10.28)	10.10 (9.66, 10.48)	10.07 (9.68, 10.46)	<0.001

[1] Median (Q1, Q3)
[2] Kruskal-Wallis rank sum test

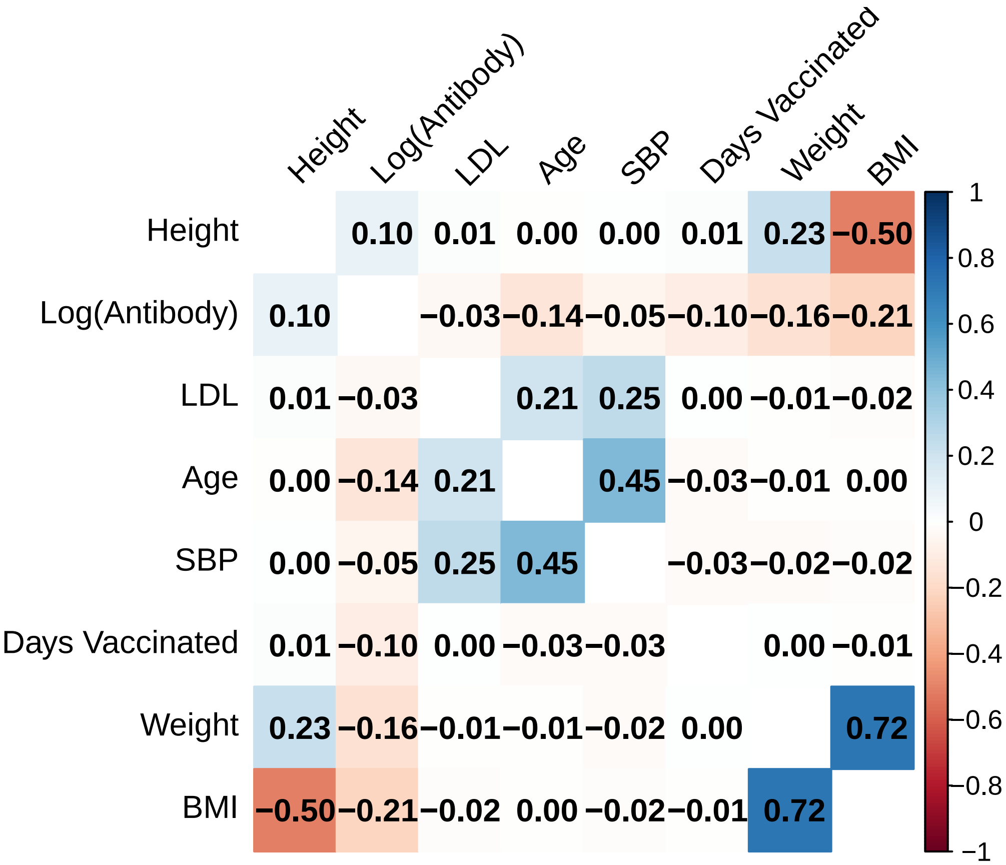


```r
dat %>% select(smoking, log_antibody) %>%
  tbl_summary(by = smoking) %>%
    add_p() %>%
    add_overall() %>%
  modify_caption("Log-Transformed Antibody Level, by Smoking Status") %>%
  as_gt() %>%
  tab_options(table.font.size = 10)
```

Correlation Matrix of Numerical Variables

Potentially will try to reorder axes but idk why its not going lol !

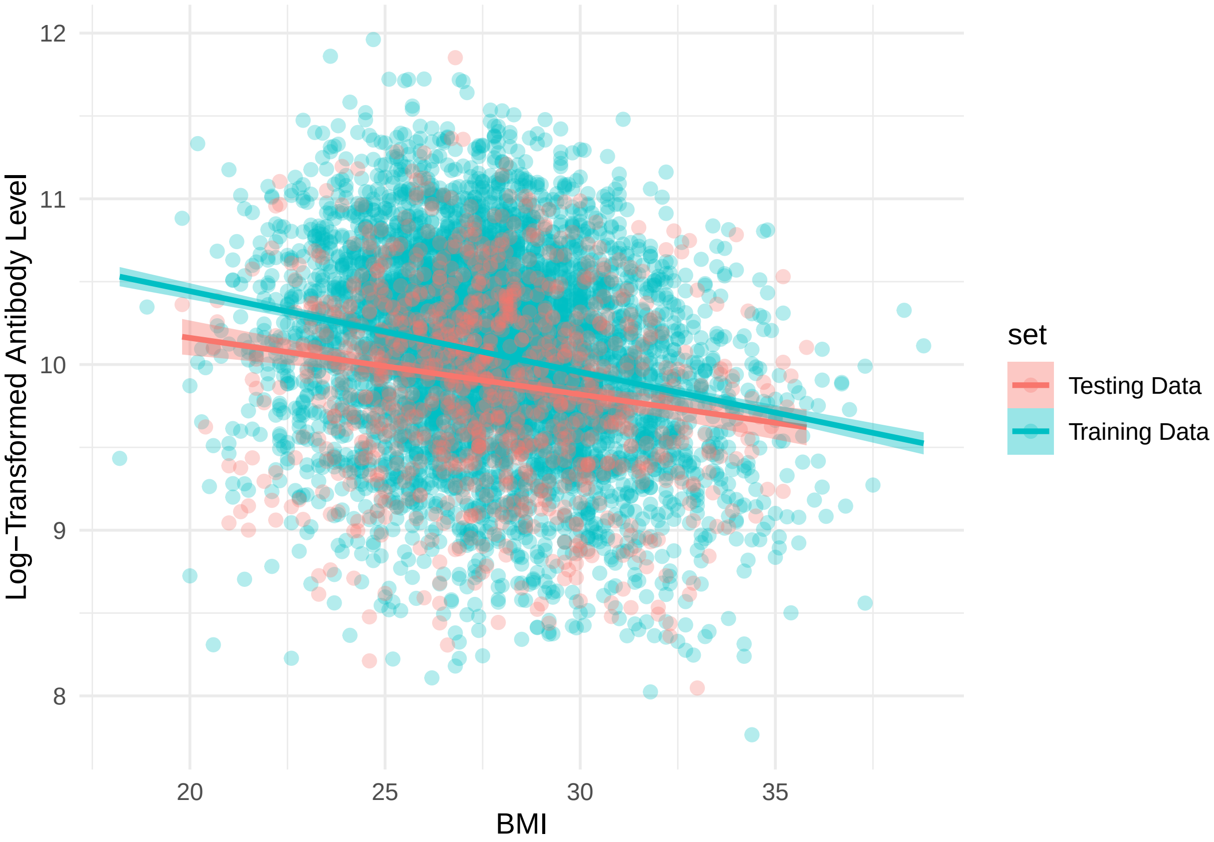
```r
cor_matrix <- dat %>%
  select(age, height, weight, bmi, sbp, ldl, days_vaccinated, log_antibody) %>%
  rename("Age" = age,
         "Height" = height,
         "Weight" = weight,
         "BMI" = bmi,
         "SBP" = sbp,
         "LDL" = ldl,
         "Days Vaccinated" = days_vaccinated,
         "Log(Antibody)" = log_antibody) %>%
  cor()

cor_plot <- corrplot(cor_matrix,  method = "color",
         addCoef.col = "black",
         tl.col = "black",
         tl.srt = 45,
         order = 'hclust',
         diag = F)
```



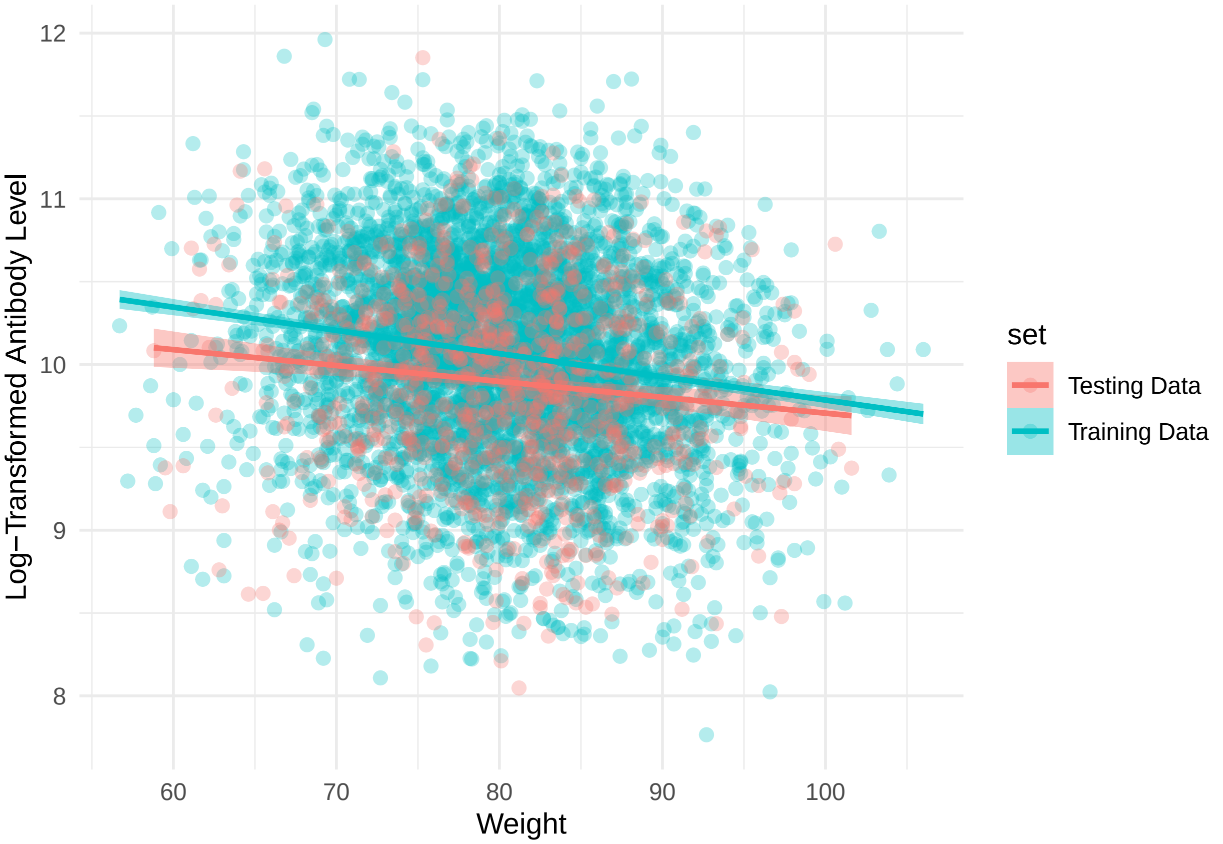
Plots of Log(Antibody) vs. Selected Numerical Variables

Not sure how I feel about these lol

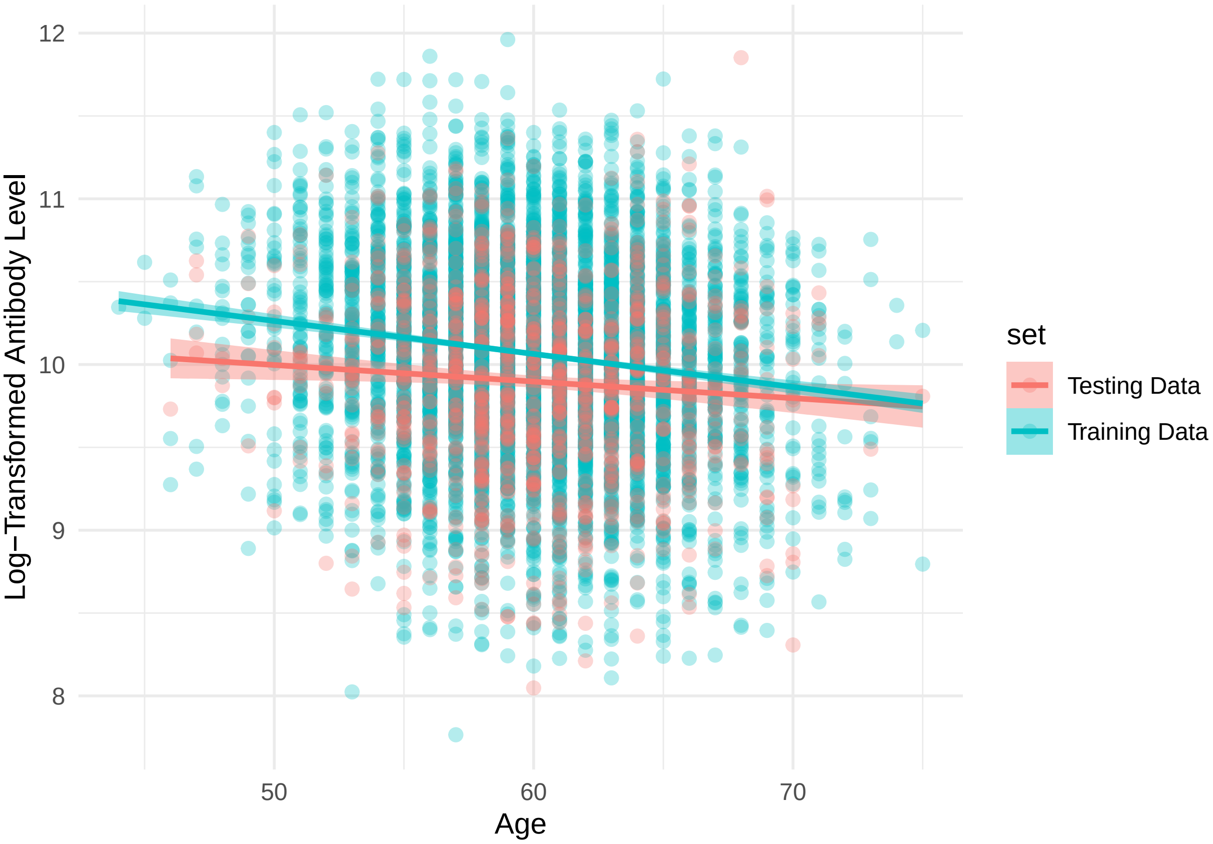
```
# Antibody level vs. BMI
plot_bmi <- dat %>%
  ggplot(aes(x = bmi, y = log_antibody, fill = set, color = set)) +
  geom_point(alpha = 0.3, size = 2) +
  geom_smooth(method = "lm") +
  labs(y = "Log-Transformed Antibody Level", x = "BMI") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
plot_bmi
```



```
# Antibody level vs. Weight
plot_weight <- dat %>%
  ggplot(aes(x = weight, y = log_antibody, fill = set, color = set)) +
  geom_point(alpha = 0.3, size = 2) +
  geom_smooth(method = "lm") +
  labs(y = "Log-Transformed Antibody Level", x = "Weight") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
plot_weight
```



```
# Antibody level vs. Age
plot_age <- dat %>%
  ggplot(aes(x = age, y = log_antibody, fill = set, color = set)) +
  geom_point(alpha = 0.3, size = 2) +
  geom_smooth(method = "lm") +
  labs(y = "Log-Transformed Antibody Level", x = "Age") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 14),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
plot_age
```



Model Training

Results