P8106 Midterm - Code

Kate Colvin (KAC2301), Jeong Yun (Lizy) Choi (JC6452), and Flora Pang (FP2513)

Exploratory Analysis

Loading in Data

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

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

Producing Summary Table

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(
    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)
```

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

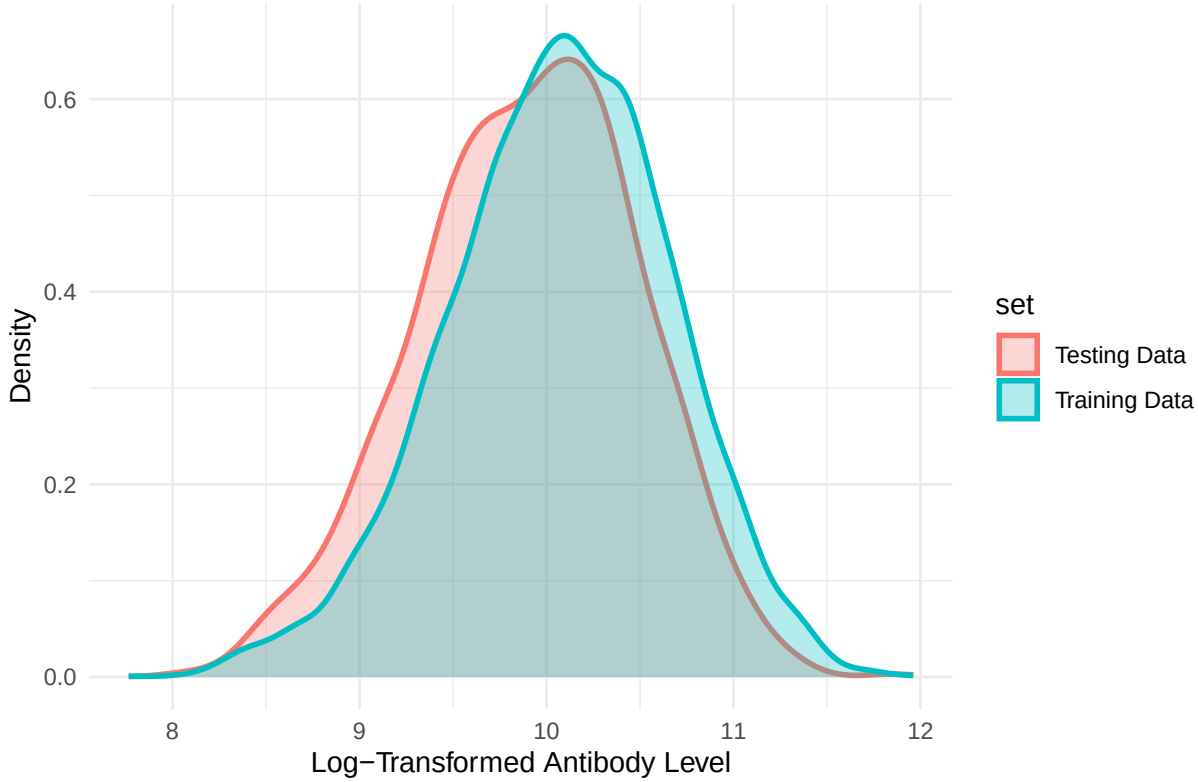
Characteristic	**Overall** N = 6,000[1]	**Testing Data** N = 1,000[1]	**Training Data** N = 5,000[1]	**p-value**[2]
Age	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)	0.9
Gender				0.7
Female	3,082 (51%)	509 (51%)	2,573 (51%)	
Male	2,918 (49%)	491 (49%)	2,427 (49%)	
Race				0.6
Asian	333 (5.6%)	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

Histograms of Differing Variables by Training and Test Set

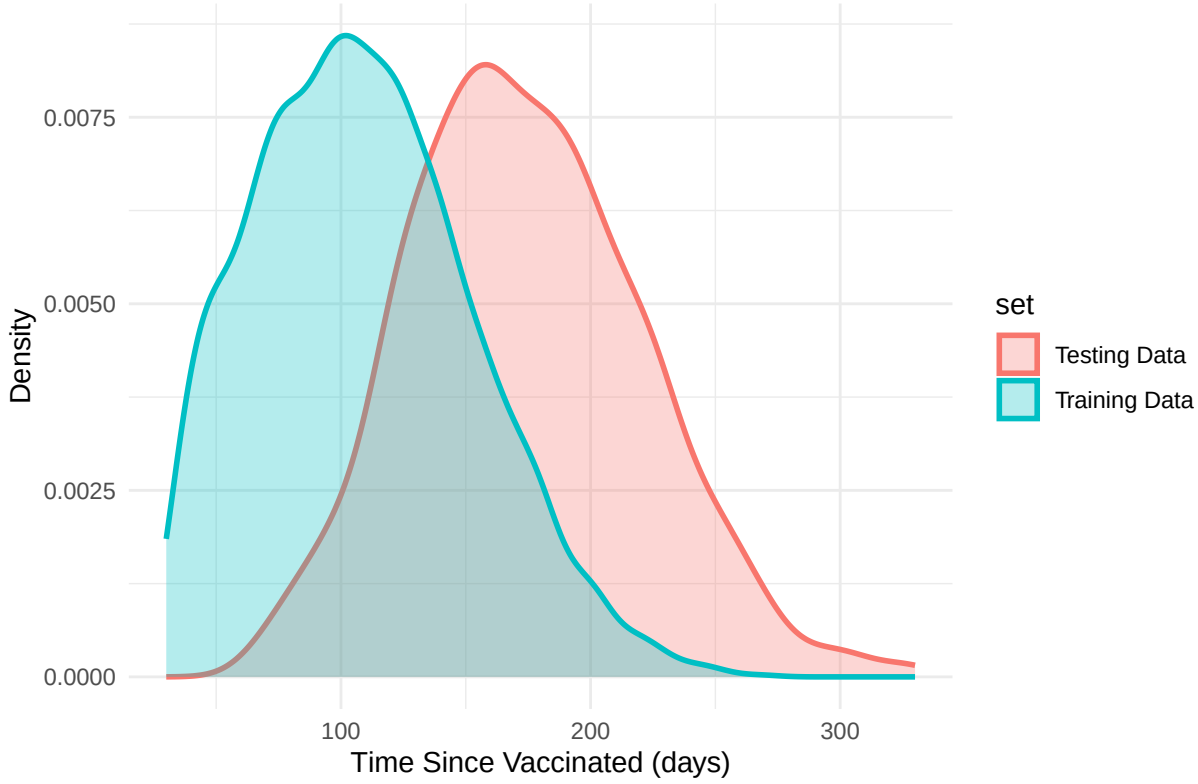
```
# Antibody level
plot_sets <- dat %>%
  ggplot(aes(x = log_antibody,
             fill = set,
             color = set)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  labs(x = "Log-Transformed Antibody Level",
       y = "Density",
       title = "Figure 1: Distribution of Log-Transformed Antibody Level, by Data Set") +
  theme_minimal()

# 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",
       title = "Figure 2: Distribution of Days Since Vaccination, by Data Set") +
  theme_minimal()

plot_sets
```

Figure 1: Distribution of Log−Transformed Antibody Level, by Data Set



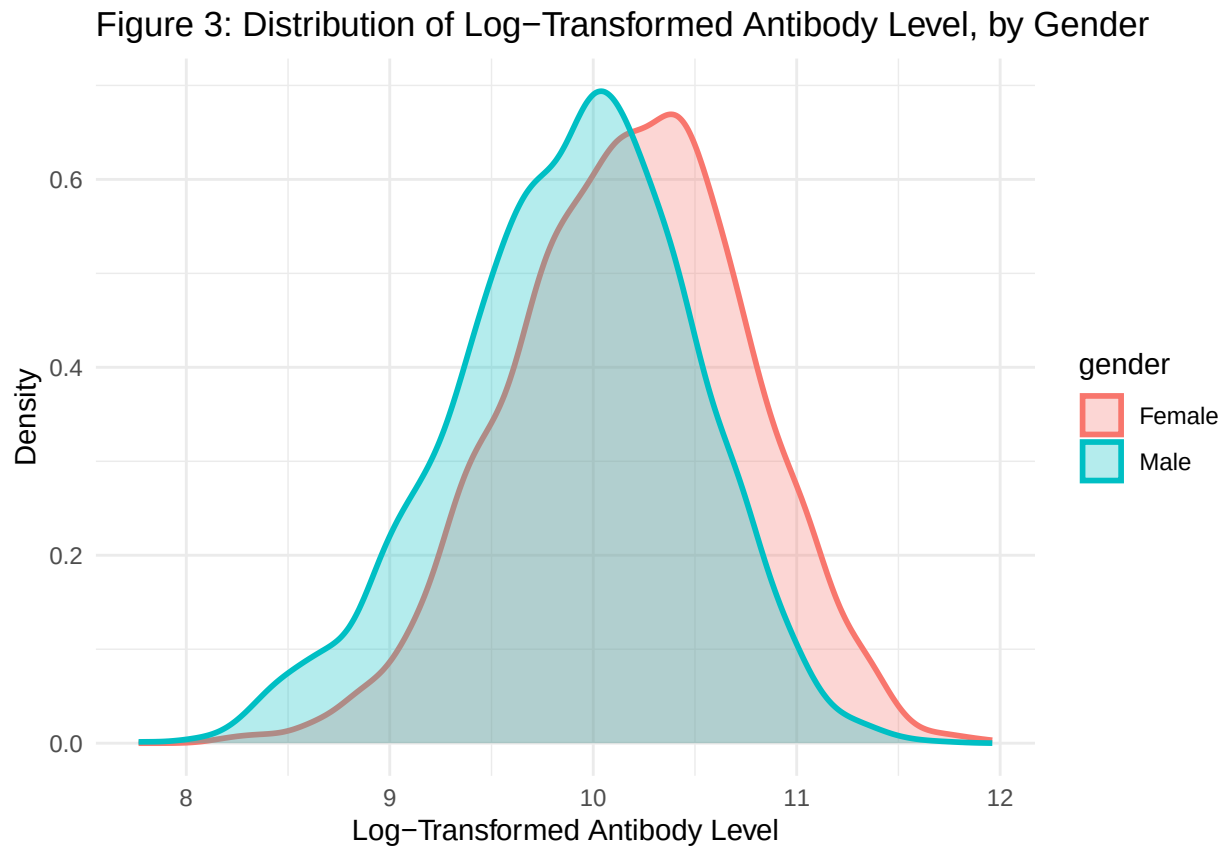

```
plot_days
```

Figure 2: Distribution of Days Since Vaccination, by Data Set



Plots of Log-Transformed Antibody Level, 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",
       title = "Figure 3: Distribution of Log-Transformed Antibody Level, by Gender") +
  theme_minimal()
plot_gender
```



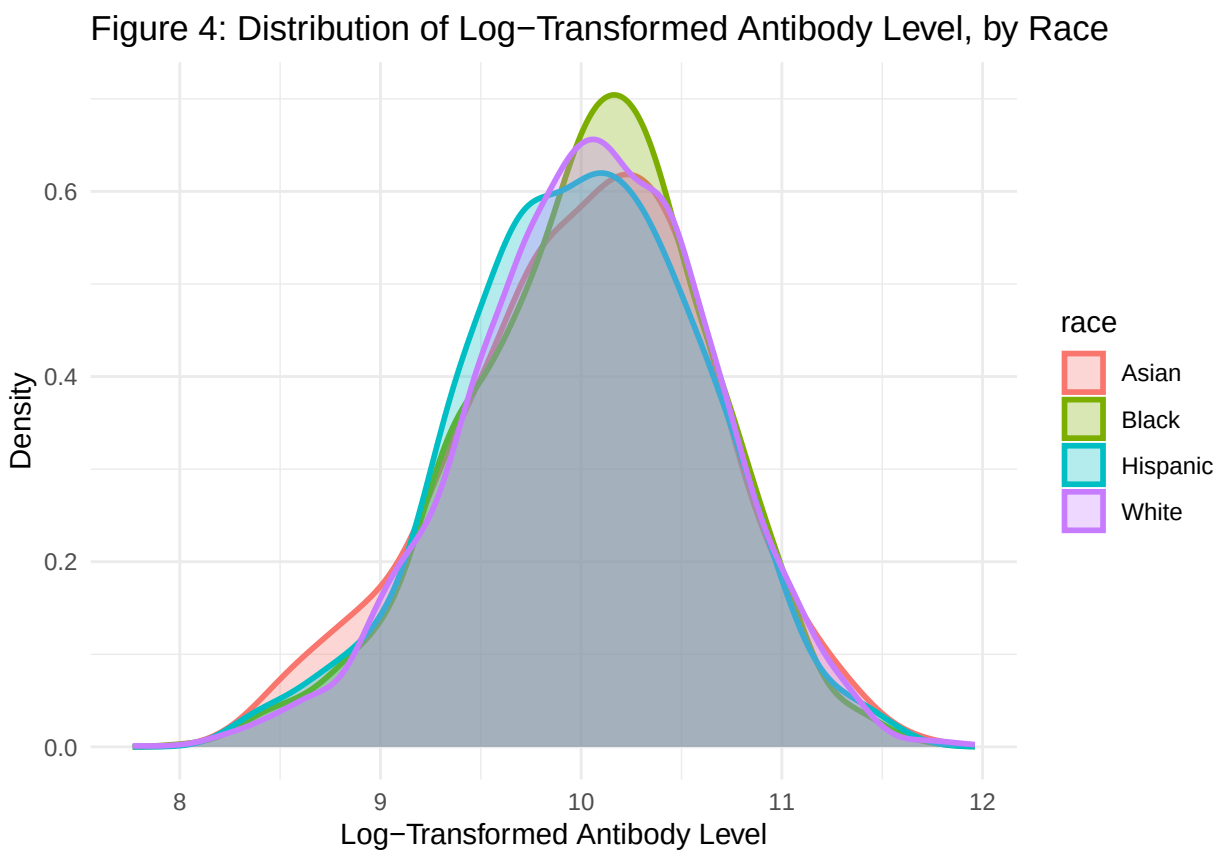

```
strip_markdown <- function(x) {gsub("\\*\\*", "", x)}

dat %>% select(gender, log_antibody) %>%
  tbl_summary(by = gender) %>% add_p() %>%
  modify_caption("Log-Transformed Antibody Level, by Gender") %>%
  as_kable() %>%
  footnote(general_title = "", general = "Median (Q1, Q3), Wilcoxon Rank Sum Test") %>%
  strip_markdown()
```

Table 2: Log-Transformed Antibody Level, by Gender

Characteristic	Female N = 3,082	Male N = 2,918	p-value
log_antibody	10.20 (9.79, 10.58)	9.93 (9.51, 10.30)	<0.001

Median (Q1, Q3), Wilcoxon Rank Sum Test

```r
# 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",
       title = "Figure 4: Distribution of Log-Transformed Antibody Level, by Race") +
  theme_minimal()

plot_race
```



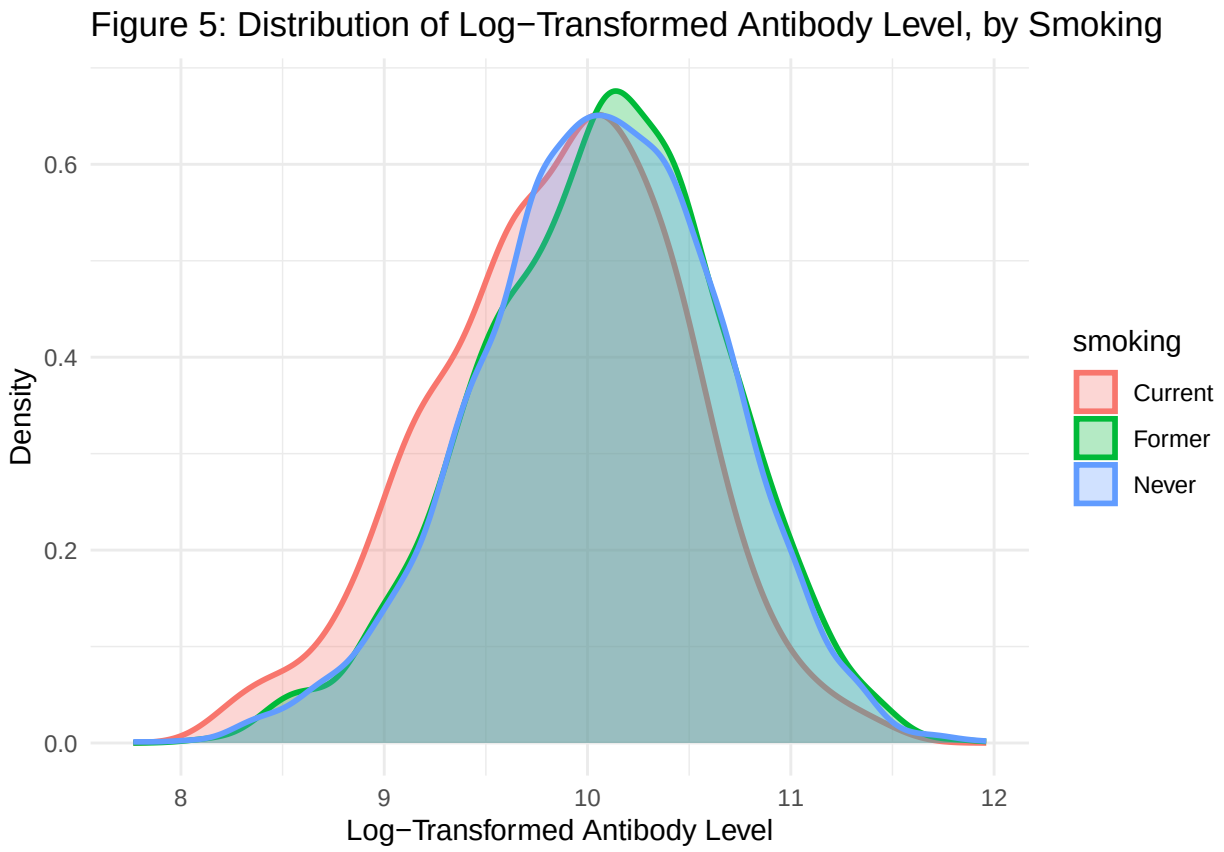

```r
dat %>%
  select(race, log_antibody) %>%
  tbl_summary(by = race) %>%
    add_p() %>%
  modify_caption("Log-Transformed Antibody Level, by Race") %>%
  as_kable() %>%
  footnote(general_title = "",
           general = "Median (Q1, Q3), Kruskal-Wallis Rank Sum Test") %>%
  strip_markdown()
```

Table 3: Log-Transformed Antibody Level, by Race

Characteristic	Asian N = 333	Black N = 1,235	Hispanic N = 548	White N = 3,884	p-value
log_antibody	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

Median (Q1, Q3), Kruskal-Wallis Rank Sum Test

```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",
       title = "Figure 5: Distribution of Log-Transformed Antibody Level, by Smoking") +
  theme_minimal()
plot_smoking
```




```r
dat %>% select(smoking, log_antibody) %>%
  tbl_summary(by = smoking) %>%
    add_p() %>%
  modify_caption("Log-Transformed Antibody Level, by Smoking Status") %>%
  as_kable() %>%
  footnote(general_title = "",
           general = "Median (Q1, Q3), Kruskal-Wallis Rank Sum Test") %>%
  strip_markdown()
```

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

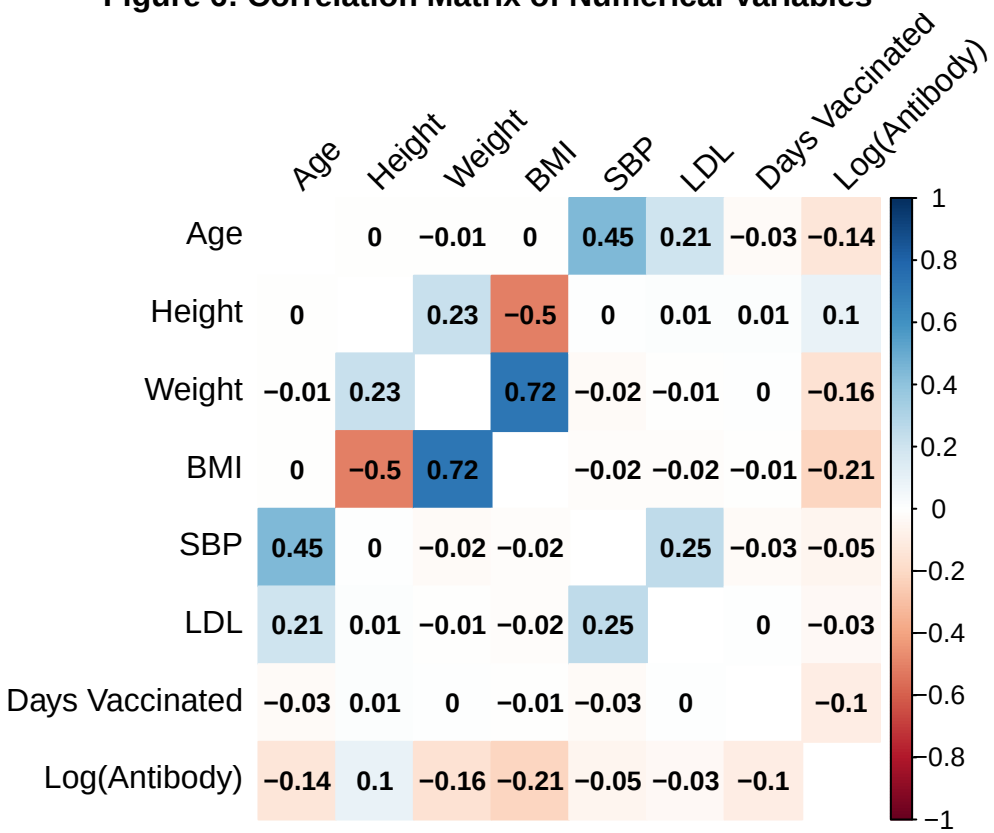
Characteristic	Current N = 589	Former N = 1,800	Never N = 3,611	p-value
log_antibody	9.91 (9.46, 10.28)	10.10 (9.66, 10.48)	10.07 (9.68, 10.46)	<0.001

Median (Q1, Q3), Kruskal-Wallis Rank Sum Test

Correlation Matrix of Numerical Variables

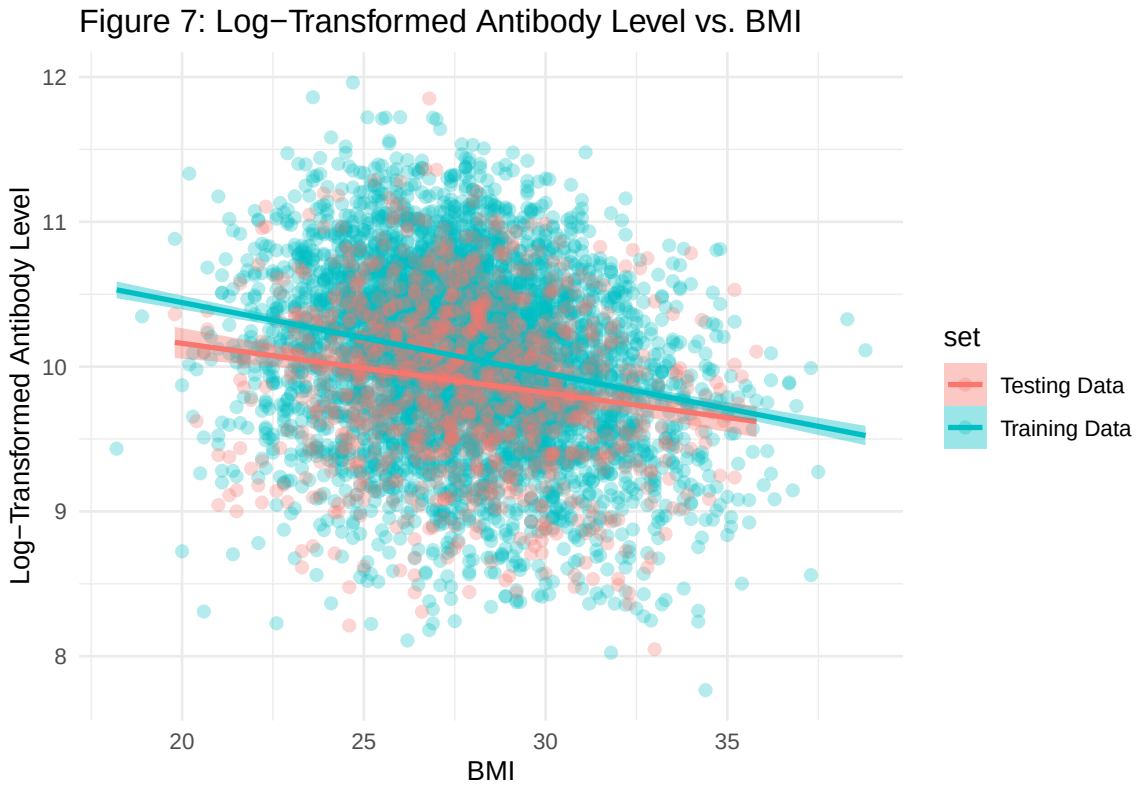
```
cor_matrix <- dat %>%
  select(age, height, weight, bmi, sbp, ldl, days_vaccinated, log_antibody) %>%
  rename("Age" = age,
         "Height" = height,
         "Weight" = weight,
         "BMI" = bmi,
         "SBP" = sbp,
         "LDL" = ldl,
         "Days Vaccinated" = days_vaccinated,
    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

cor_plot <- corrplot(cor_matrix,
                     main = "Figure 6: Correlation Matrix of Numerical Variables",
                     mar=c(0,0,1,0), cex.main = 1,
                     method = "color",
                     addCoef.col = "black",
                     tl.col = "black",
                     number.cex = 0.8,
                     tl.srt = 45,
                     order = 'original',
                     diag = F)
```

Figure 6: Correlation Matrix of Numerical Variables



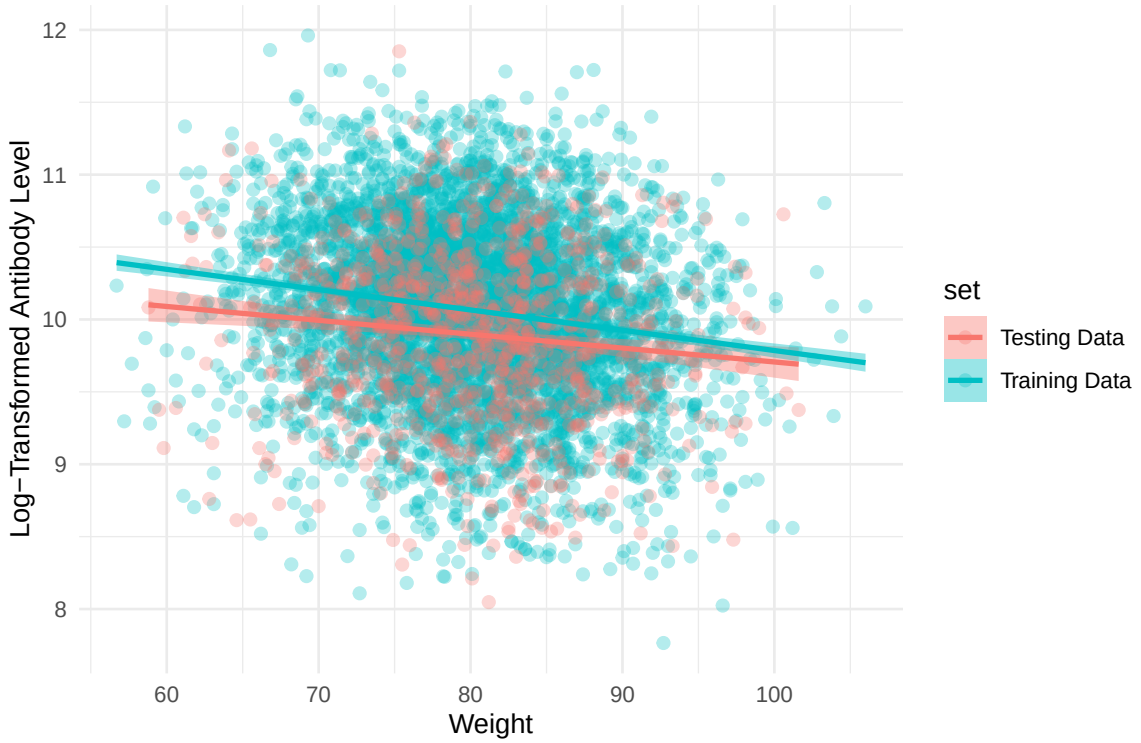
Plots of Log-Transformed Antibody Level vs. Selected Numerical Variables

```
# 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",
       title = "Figure 7: Log-Transformed Antibody Level vs. BMI") +
  theme_minimal()

# 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",
       title = "Figure 8: Log-Transformed Antibody Level vs. Weight") +
  theme_minimal()

# 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",
       title = "Figure 9: Log-Transformed Antibody Level vs. Age") +
  theme_minimal()

plot_bmi
```



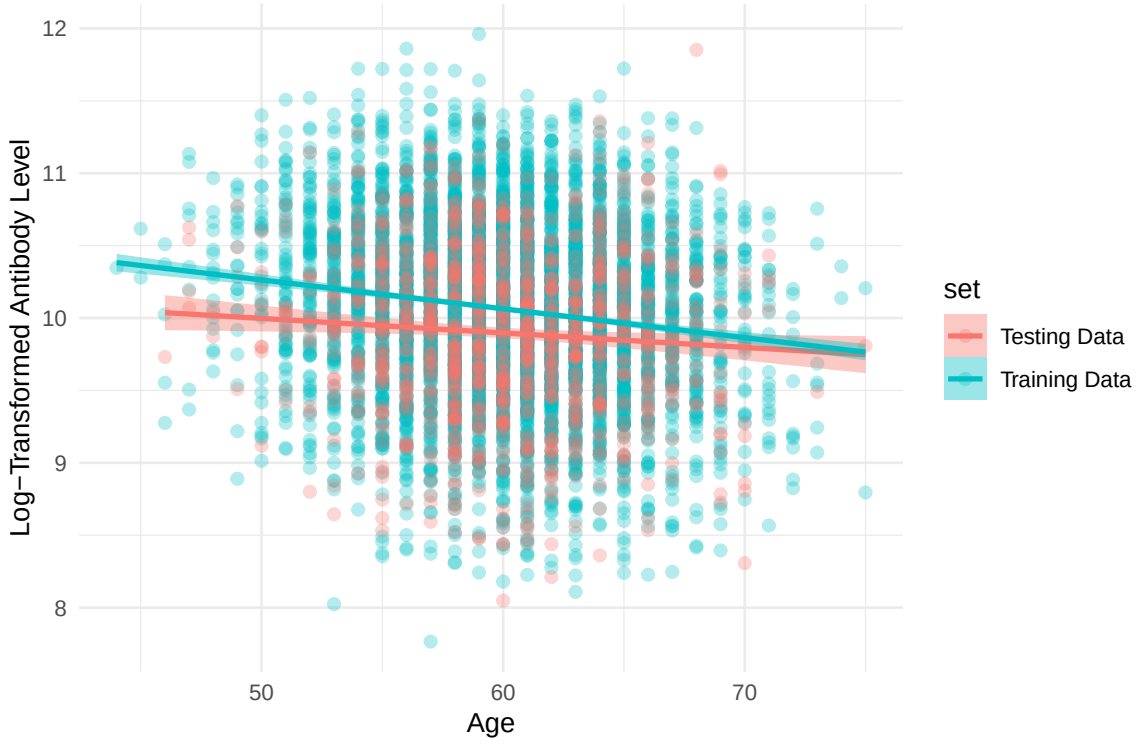
```
plot_weight
```

Figure 8: Log−Transformed Antibody Level vs. Weight




```
plot_age
```

Figure 9: Log−Transformed Antibody Level vs. Age



Model Selection and Training

Since the response variable (log_antibody) is continuous, this project will consider the following models:

- Multiple Linear Regression (MLR) - as a baseline.
- LASSO Regression – to improve predictive performance by selecting important predictors.
- MARS model – allow remain in regression but also capture nonlinear effects

After comparing model performance, the best model will be based on cross-validation results.

Data Pre-processing

```
# Converting categorical variables into factors
dat1 <- dat1 %>%
  mutate(
    gender = factor(gender, levels = c(0, 1), labels = c("Female", "Male")),
    race = factor(race, levels = c(1, 2, 3, 4), labels = c("White", "Asian", "Black", "Hispanic")),
    smoking = factor(smoking, levels = c(0, 1, 2), labels = c("Never", "Former", "Current")),
    diabetes = factor(diabetes),
    hypertension = factor(hypertension)
  )

dat2 <- dat2 %>%
  mutate(
    gender = factor(gender, levels = c(0, 1), labels = c("Female", "Male")),
    race = factor(race, levels = c(1, 2, 3, 4), labels = c("White", "Asian", "Black", "Hispanic")),
    smoking = factor(smoking, levels = c(0, 1, 2), labels = c("Never", "Former", "Current")),
    diabetes = factor(diabetes),
    hypertension = factor(hypertension)
  )
```

```
sum(is.na(dat1))
```

```
## [1] 0
```

```
sum(is.na(dat2))
```

```
## [1] 0
```

```
dat1 <- dat1 %>% select(-id)
dat2 <- dat2 %>% select(-id)
```

Training multiple linear regression model

```
mlr_model <- lm(log_antibody ~ ., data = dat1)
summary(mlr_model)
```

```
##
## Call:
## lm(formula = log_antibody ~ ., data = dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14396 -0.35840  0.02944  0.37802  1.65090
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26.6751961  2.3149812  11.523  < 2e-16 ***
## age            -0.0205979  0.0019385 -10.626  < 2e-16 ***
## genderMale     -0.2974929  0.0155977 -19.073  < 2e-16 ***
## raceAsian      -0.0060422  0.0344613  -0.175   0.8608
## raceBlack      -0.0075295  0.0196815  -0.383   0.7021
## raceHispanic   -0.0417571  0.0273309  -1.528   0.1266
## smokingFormer   0.0219907  0.0173992   1.264   0.2063
## smokingCurrent -0.1934834  0.0269576  -7.177 8.15e-13 ***
## height         -0.0821381  0.0135622  -6.056 1.49e-09 ***
## weight          0.0859034  0.0143481   5.987 2.29e-09 ***
## bmi            -0.2977935  0.0412612  -7.217 6.10e-13 ***
## diabetes1       0.0112795  0.0215643   0.523   0.6010
## hypertension1  -0.0179106  0.0260931  -0.686   0.4925
## sbp             0.0015181  0.0017049   0.890   0.3733
## ldl            -0.0001645  0.0004028  -0.409   0.6829
## time           -0.0003011  0.0001795  -1.677   0.0936 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5503 on 4984 degrees of freedom
## Multiple R-squared:  0.1513, Adjusted R-squared:  0.1488
## F-statistic: 59.25 on 15 and 4984 DF,  p-value: < 2.2e-16
```

```
# Evaluating model performance on validation data
mlr_pred <- predict(mlr_model, newdata = dat2)
mlr_rmse <- sqrt(mean((mlr_pred - dat2$log_antibody)^2))
mlr_adj_r2 <- summary(mlr_model)$adj.r.squared
mlr_rmse
```

```
## [1] 0.568318
```

```
mlr_adj_r2
```

```
## [1] 0.1487817
```

Training LASSO regression model

Standardizing numerical variables for LASSO

```
num_vars <- c("age", "height", "weight", "bmi", "sbp", "ldl", "time") #only continuous variable

preprocess_params <- preProcess(dat1[, num_vars], method = c("center", "scale"))
dat1[, num_vars] <- predict(preprocess_params, dat1[, num_vars])
dat2[, num_vars] <- predict(preprocess_params, dat2[, num_vars])
```

```
# Preparing the data matrices for glmnet
x_train <- model.matrix(log_antibody ~ ., dat1)[, -1]  # Removing the intercept
y_train <- dat1$log_antibody

x_valid <- model.matrix(log_antibody ~ ., dat2)[, -1]
y_valid <- dat2$log_antibody
```

```
set.seed(123)
lasso_model <- cv.glmnet(x_train, y_train, alpha = 1) # LASSO with cross validation
best_lambda <- lasso_model$lambda.min
```

```r
lasso_final <- glmnet(x_train, y_train, alpha = 1, lambda = best_lambda) # final model is based on opti
```

```r
# predicting with LASSO on validation data

lasso_pred <- predict(lasso_final, newx = x_valid)
lasso_rmse <- sqrt(mean((lasso_pred - y_valid)^2))
lasso_rmse
```

```
## [1] 0.5683693
```

```r
ss_total_lasso <- sum((y_valid - mean(y_valid))^2)
ss_res_lasso <- sum((lasso_pred - y_valid)^2)
r_squared_lasso <- 1 - ss_res_lasso / ss_total_lasso
n <- length(y_valid)
p <- length(coef(lasso_final)) - 1
lasso_adj_r2 <- 1 - ((1 - r_squared_lasso) * (n - 1) / (n - p - 1))

system.time(glmnet(x_train, y_train, alpha = 1, lambda = best_lambda))
```

```
##    user  system elapsed
##   0.002   0.000   0.001
```

Training MARS model

```r
mars_model <- earth(log_antibody ~ ., data = dat1)
summary(mars_model)
```

```
## Call: earth(formula=log_antibody~., data=dat1)
##
##                   coefficients
## (Intercept)         10.8474469
## genderMale          -0.2962905
## smokingCurrent      -0.2051269
## h(-0.215005-age)     0.0726888
## h(age- -0.215005)   -0.1034029
## h(bmi- -1.46481)    -0.2327459
## h(0.0216075-bmi)    -0.1710074
## h(-1.19454-time)    -1.4557278
## h(time- -1.19454)   -0.0978688
##
## Selected 9 of 10 terms, and 5 of 15 predictors
## Termination condition: RSq changed by less than 0.001 at 10 terms
## Importance: bmi, genderMale, time, age, smokingCurrent, raceAsian-unused, ...
## Number of terms at each degree of interaction: 1 8 (additive model)
## GCV 0.2787787    RSS 1384.431    GRSq 0.2166152    RSq 0.2216218
```

```r
# predicting with MARS on validation data
mars_pred <- predict(mars_model, newdata = dat2)
mars_rmse <- sqrt(mean((mars_pred - dat2$log_antibody)^2))


mars_rmse
```

```
## [1] 0.5327718
```

- The MARS model achieves the lowest RMSE. Therefore MARS will be used as the preferred model for

predicting log_antibody. Although further fine-tuning and additional feature exploration could further enhance the model's predictive power.

MARS model tuning

```
tune_grid <- expand.grid(degree = 1:3, nprune = seq(5, 50, by = 5))

train_control <- trainControl(method = "cv", number = 10)
mars_tune <- train(log_antibody ~ ., data = dat1, method = "earth",
                   trControl = train_control, tuneGrid = tune_grid)

mars_tune$bestTune
```

```
##   nprune degree
## 2     10      1
```

```
train_control <- trainControl(method = "cv", number = 10)

# Train the MARS model with best tuning parameters
mars_model_tune <- train(log_antibody ~ .,
                         data = dat1,
                         method = "earth",
                         trControl = train_control,
                         tuneGrid = data.frame(nprune = 10, degree = 1))

summary(mars_model_tune)
```

```
## Call: earth(x=matrix[5000,15], y=c(10.65,9.889,1...), keepxy=TRUE, degree=1,
##             nprune=10)
##
##                     coefficients
## (Intercept)           10.8474469
## genderMale            -0.2962905
## smokingCurrent        -0.2051269
## h(-0.215005-age)       0.0726888
## h(age- -0.215005)     -0.1034029
## h(bmi- -1.46481)      -0.2327459
## h(0.0216075-bmi)      -0.1710074
## h(-1.19454-time)      -1.4557278
## h(time- -1.19454)     -0.0978688
##
## Selected 9 of 10 terms, and 5 of 15 predictors (nprune=10)
## Termination condition: RSq changed by less than 0.001 at 10 terms
## Importance: bmi, genderMale, time, age, smokingCurrent, raceAsian-unused, ...
## Number of terms at each degree of interaction: 1 8 (additive model)
## GCV 0.2787787    RSS 1384.431    GRSq 0.2166152    RSq 0.2216218
```

```
mars_tune_pred <- predict(mars_model_tune, newdata = dat2)
mars_tune_rmse <- sqrt(mean((mars_tune_pred - dat2$log_antibody)^2))

# Compute R-squared
ss_total_mars <- sum((dat2$log_antibody - mean(dat2$log_antibody))^2)
ss_res_mars <- sum((mars_tune_pred - dat2$log_antibody)^2)
r_squared_mars <- 1 - ss_res_mars / ss_total_mars
```

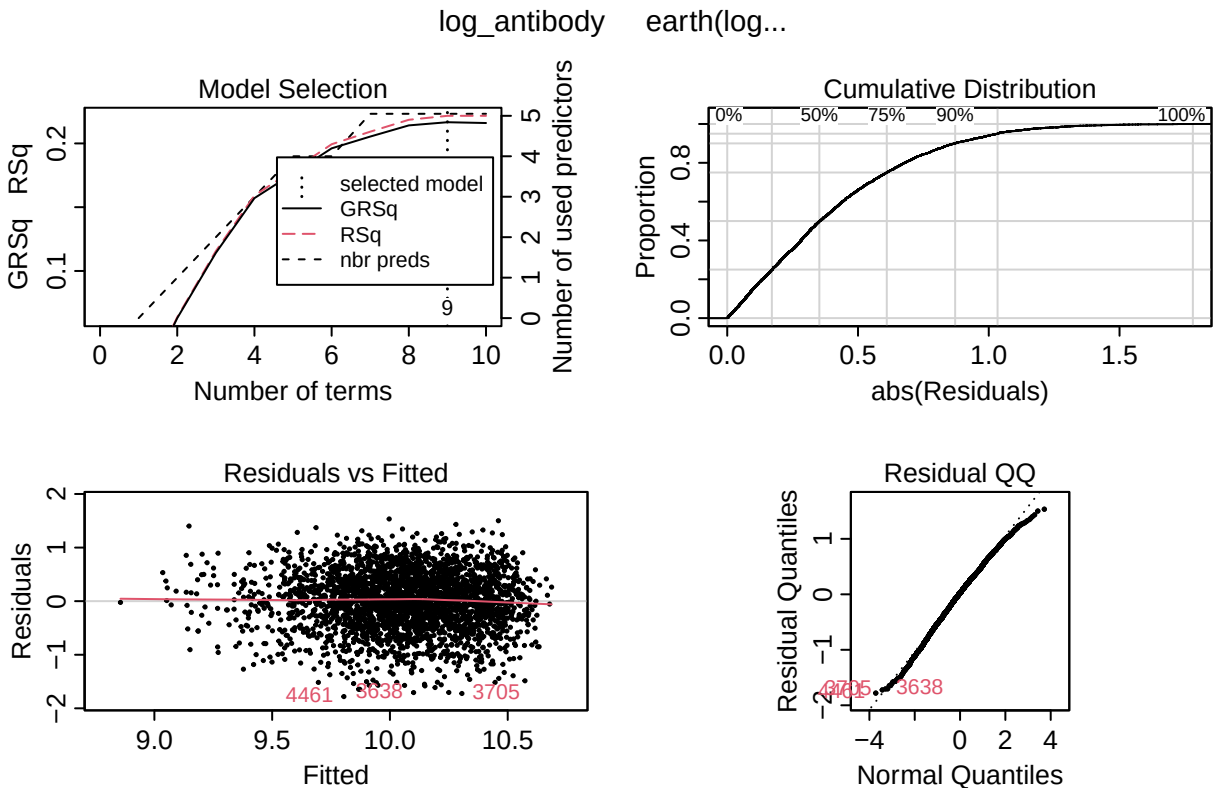
```r
# Get n and number of selected terms (p)
n <- nrow(dat2)
p <- length(mars_model_tune$finalModel$selected.terms)

# Adjusted R-squared for MARS
mars_adj_r2 <- 1 - ((1 - r_squared_mars) * (n - 1) / (n - p - 1))

# Print result
mars_adj_r2
```

```
## [1] 0.1689031
```

```r
mars_tune_rmse
```

```
## [1] 0.5327718
```

```r
plot(mars_model)
```




```r
plot(mars_model, which = 1)
```

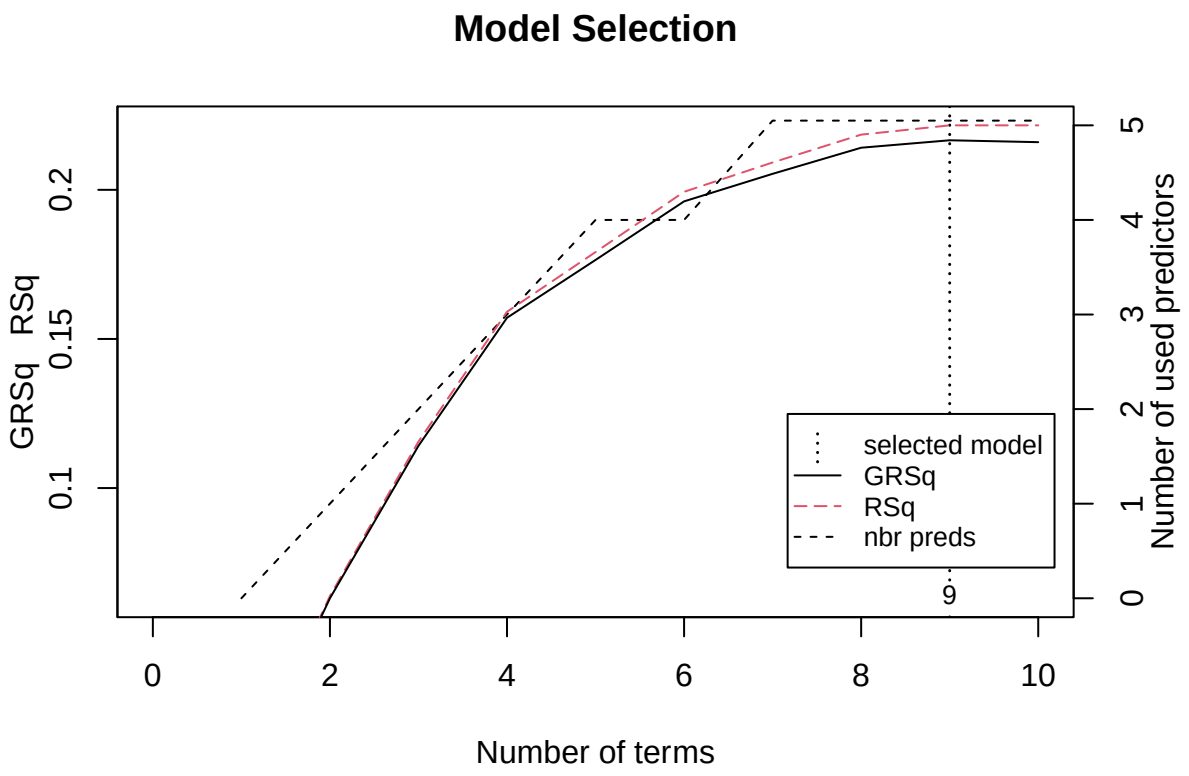

Model Selection
GRSq RSq
0.1
0.15
0.2
0
2
4
6
8
10
Number of terms
Number of used predictors
0
1
2
3
4
5
selected model
GRSq
RSq
nbr preds
9

Table: Model Performance on Independent Test Set (dat2)

Model	RMSE ↓	Adjusted R-squared↑	Notes
Multiple Linear Regression (MLR)	0.568	0.149	Baseline model; assumes linear relationships
LASSO Regression	0.568	0.048	Performs variable selection via L1 regularization
MARS (Final, Tuned)	0.533	0.169	Best performance; captures non-linear effects

Results

```
# Summary table
model_perf <- tibble(
  Model = c("Multiple Linear Regression (MLR)",
            "LASSO Regression",
            "MARS (Final, Tuned)"),
  RMSE = c(mlr_rmse, lasso_rmse, mars_tune_rmse),
  `Adjusted R-squared` = c(mlr_adj_r2, lasso_adj_r2, mars_adj_r2),
  Notes = c("Baseline model; assumes linear relationships",
            "Performs variable selection via L1 regularization",
            "Best performance; captures non-linear effects")
)

model_perf %>%
  gt() %>%
  tab_header(
    title = "Table: Model Performance on Independent Test Set (dat2)"
  ) %>%
  cols_label(
    Model = "Model",
    RMSE = "RMSE ↓",
    `Adjusted R-squared` = "Adjusted R-squared↑",
    Notes = "Notes"
  ) %>%
  fmt_number(columns = c(RMSE, `Adjusted R-squared`), decimals = 3) %>%
  tab_options(table.font.size = "small")
```