# P8106 Midterm - Report

Group 2: Kate Colvin (KAC2301), Jeong Yun (Lizy) Choi (JC6452), and Flora Pang (FP2513)

## Introduction

In this project, our team explored the dataset collected from a study on evaluating antibody responses to a newly authorized vaccine. The primary outcome of interest is the log-transformed antibody level measured via dried blood spots. The dataset includes a range of demographic and clinical predictors such as age, gender, race/ethnicity, smoking status, BMI, chronic conditions, and time since vaccination.

Our goal is to develop a predictive model that characterizes how these factors influence antibody responses and asses how well this model generalizes to a new independent dataset collected at a later time point. By doing so, we hope to identify key predictors of antibody levels and evaluate the robustness/generalizability of our model across different dataset.

## Exploratory Analysis

notes

for hist - more recent vaccines = higher antibody level ?

for scatter plots - testing slopes more flat?

## Model Training

## Results

```
## The following errors were returned during `as_gt()`:
## x For variable `age` (`set`) and "p.value" statistic: The package "cardx" (>=
##   0.2.3) is required.
## x For variable `bmi` (`set`) and "p.value" statistic: The package "cardx" (>=
##   0.2.3) is required.
## x For variable `days_vaccinated` (`set`) and "p.value" statistic: The package
##   "cardx" (>= 0.2.3) is required.
## x For variable `diabetes` (`set`) and "p.value" statistic: The package "cardx"
##   (>= 0.2.3) is required.
## x For variable `gender` (`set`) and "p.value" statistic: The package "cardx"
##   (>= 0.2.3) is required.
## x For variable `height` (`set`) and "p.value" statistic: The package "cardx"
##   (>= 0.2.3) is required.
## x For variable `hypertension` (`set`) and "p.value" statistic: The package
##   "cardx" (>= 0.2.3) is required.
## x For variable `ldl` (`set`) and "p.value" statistic: The package "cardx" (>=
##   0.2.3) is required.
## x For variable `log_antibody` (`set`) and "p.value" statistic: The package
##   "cardx" (>= 0.2.3) is required.
## x For variable `race` (`set`) and "p.value" statistic: The package "cardx" (>=
##   0.2.3) is required.
## x For variable `sbp` (`set`) and "p.value" statistic: The package "cardx" (>=
##   0.2.3) is required.
## x For variable `smoking` (`set`) and "p.value" statistic: The package "cardx"
##   (>= 0.2.3) is required.
## x For variable `weight` (`set`) and "p.value" statistic: The package "cardx"
##   (>= 0.2.3) is required.
```

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** |
|---|---|---|---|---|
| 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 (%)

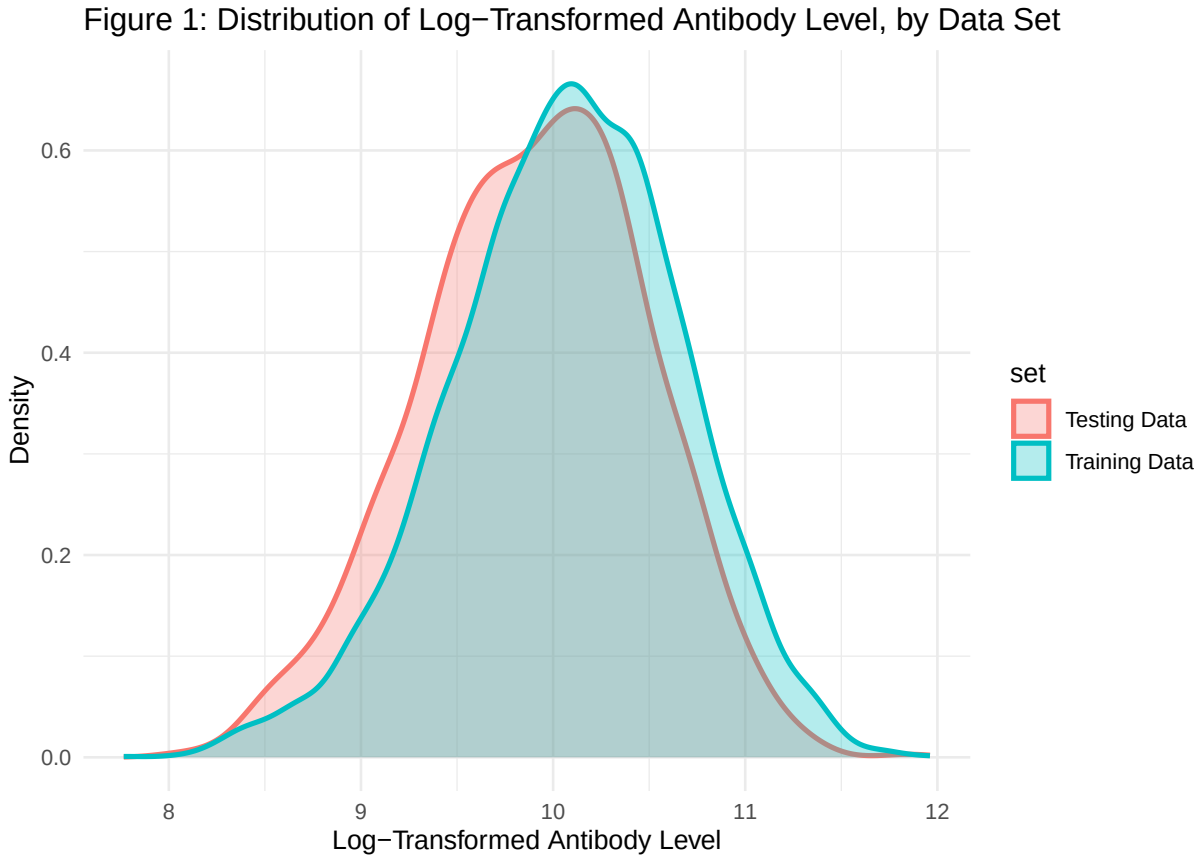

Figure 2: Distribution of Days Since Vaccination, by Data Set

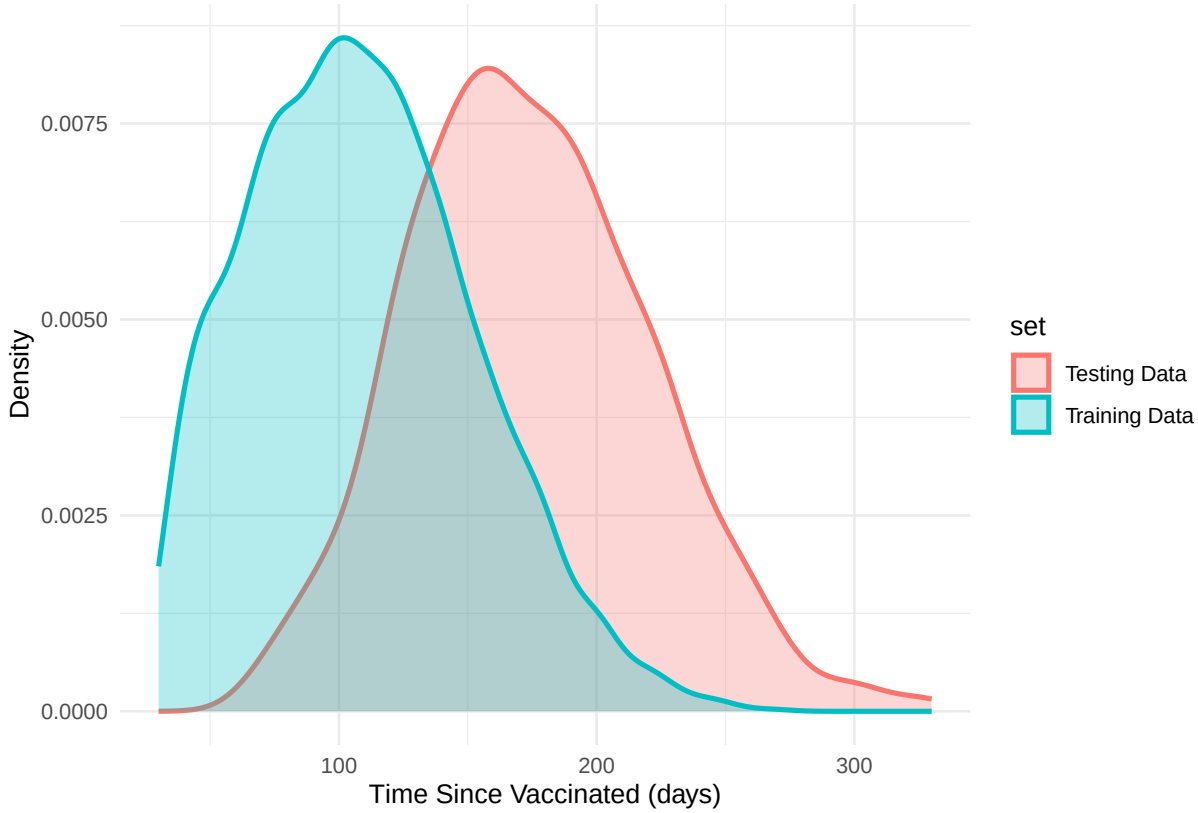


Figure 3: Distribution of Log−Transformed Antibody Level, by Gender

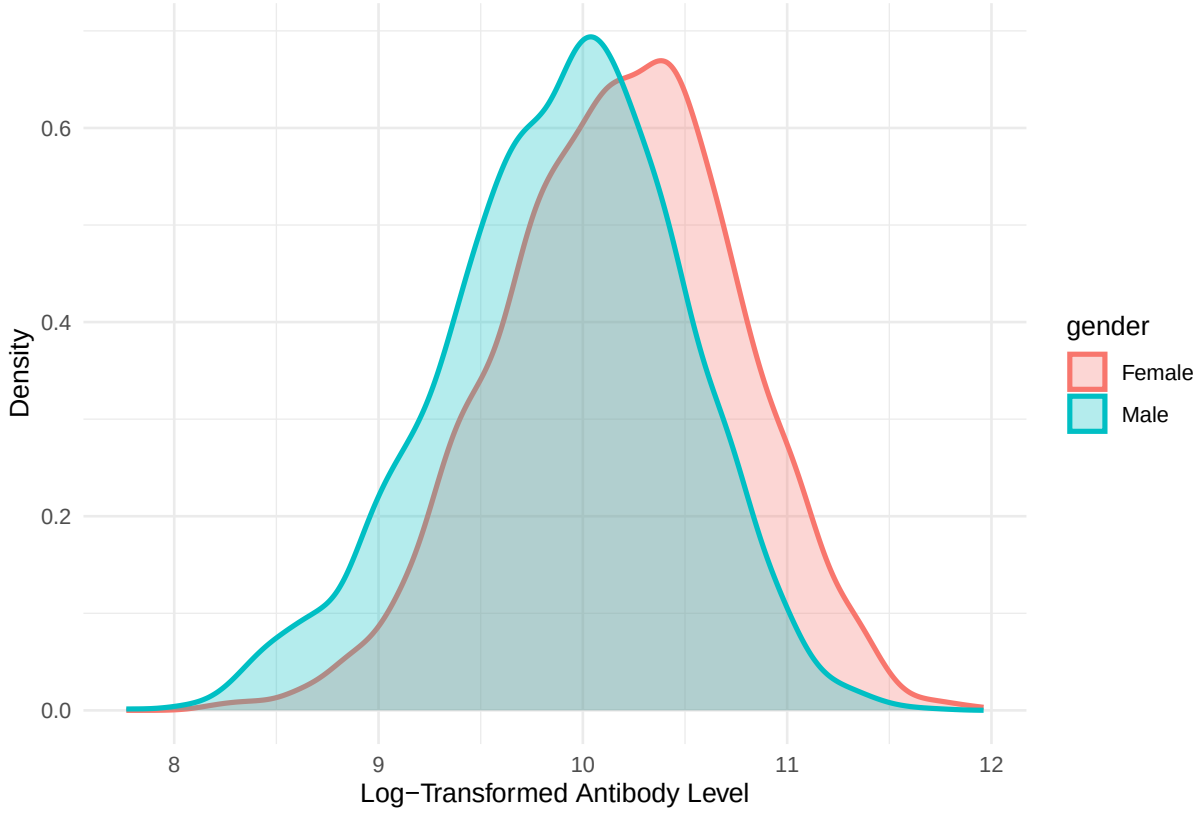

Figure 4: Distribution of Log−Transformed Antibody Level, by Race

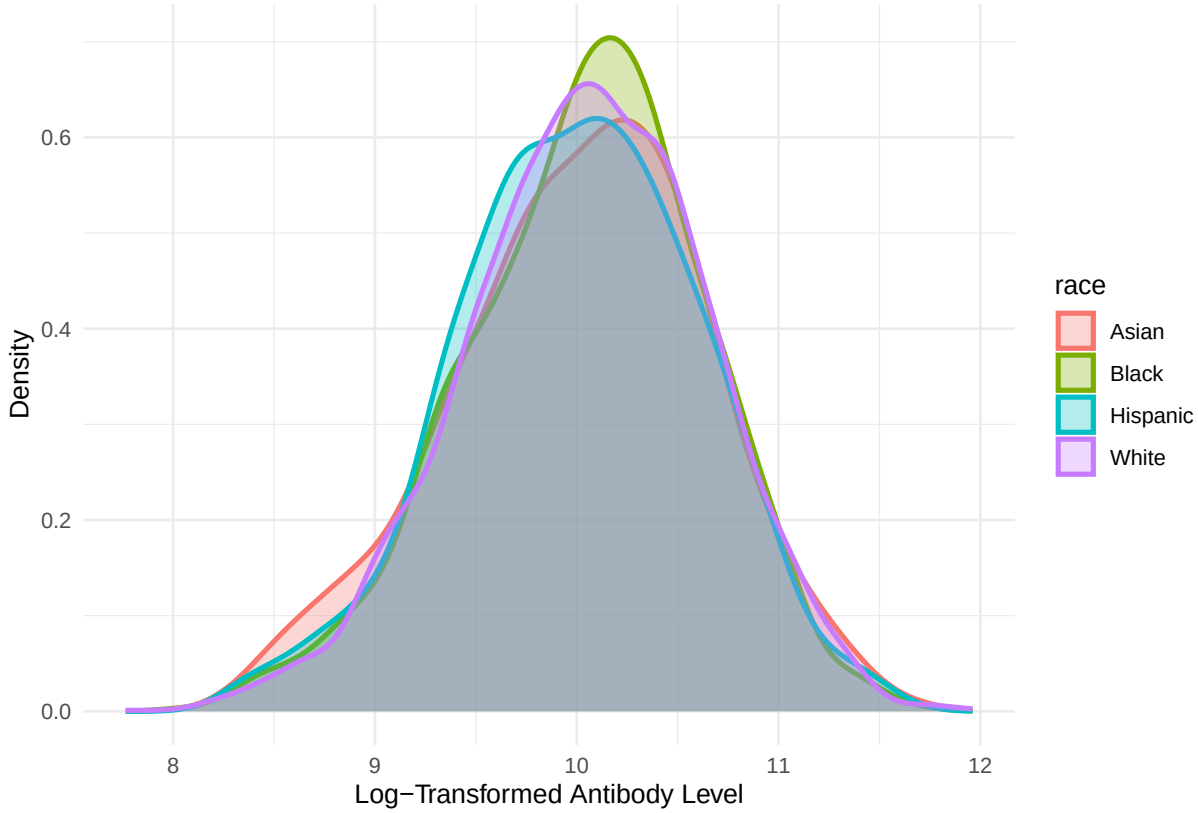


Figure 5: Distribution of Log−Transformed Antibody Level, by Smoking

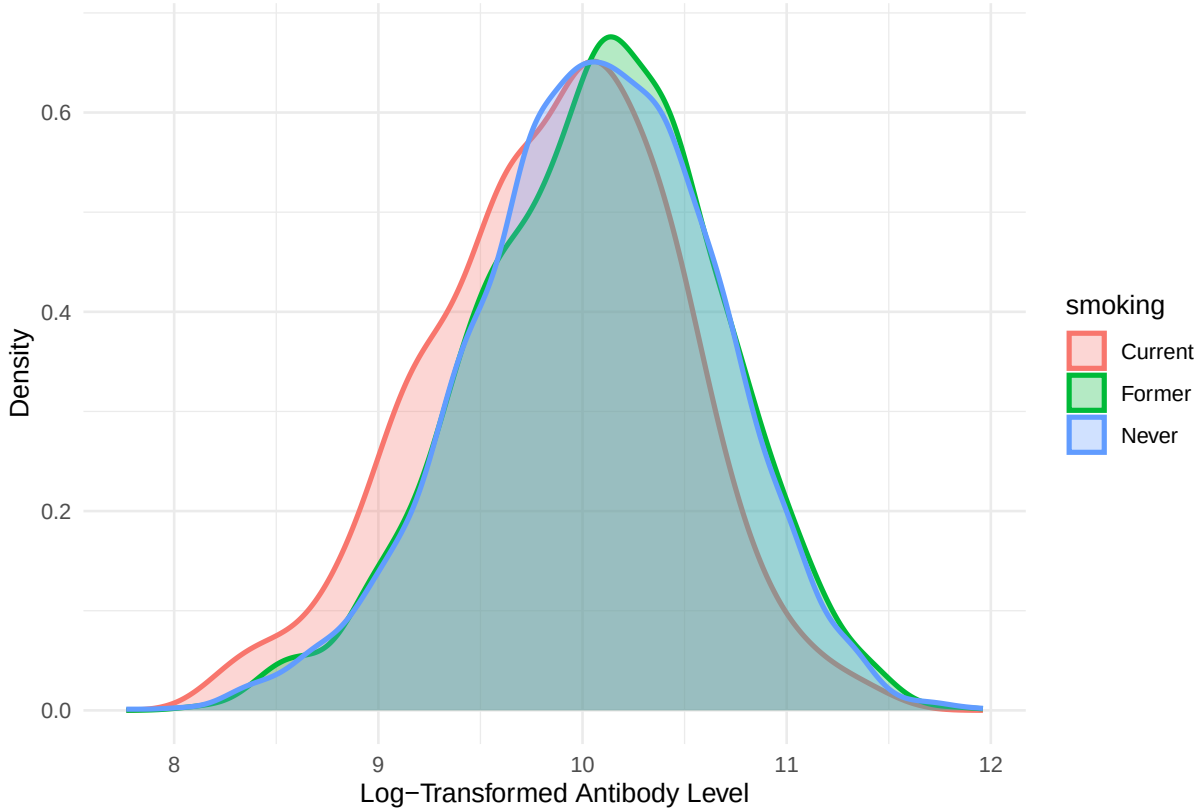

Table 2: Log-Transformed Antibody Level, by Gender

| **Characteristic** | **Female** N = 3,082[1] | **Male** N = 2,918[1] | **p-value** |
|---|---|---|---|
| log_antibody | 10.20 (9.79, 10.58) | 9.93 (9.51, 10.30) | |

[1] Median (Q1, Q3)

Table 3: Log-Transformed Antibody Level, by Race

| **Characteristic** | **Asian** N = 333[1] | **Black** N = 1,235[1] | **Hispanic** N = 548[1] | **White** N = 3,884[1] | **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) | |

[1] Median (Q1, Q3)

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

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

[1] Median (Q1, Q3)

```
## The following errors were returned during `as_gt()`:
## x For variable `log_antibody` (`gender`) and "p.value" statistic: The package
##   "cardx" (>= 0.2.3) is required.

## The following errors were returned during `as_gt()`:
## x For variable `log_antibody` (`race`) and "p.value" statistic: The package
##   "cardx" (>= 0.2.3) is required.

## The following errors were returned during `as_gt()`:
## x For variable `log_antibody` (`smoking`) and "p.value" statistic: The package
##   "cardx" (>= 0.2.3) is required.
```

**Figure 6: Correlation Matrix of Numerical Variables**

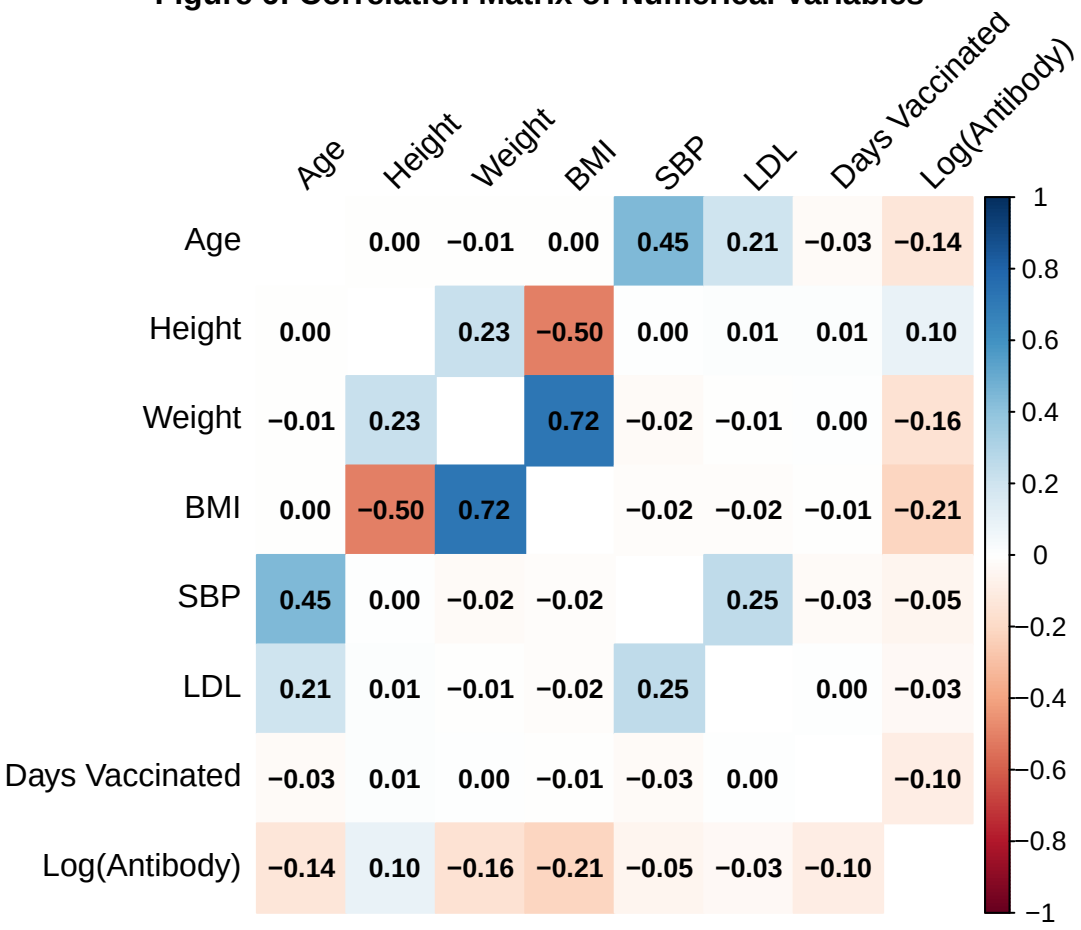

Figure 7: Log−Transformed Antibody Level vs. BMI

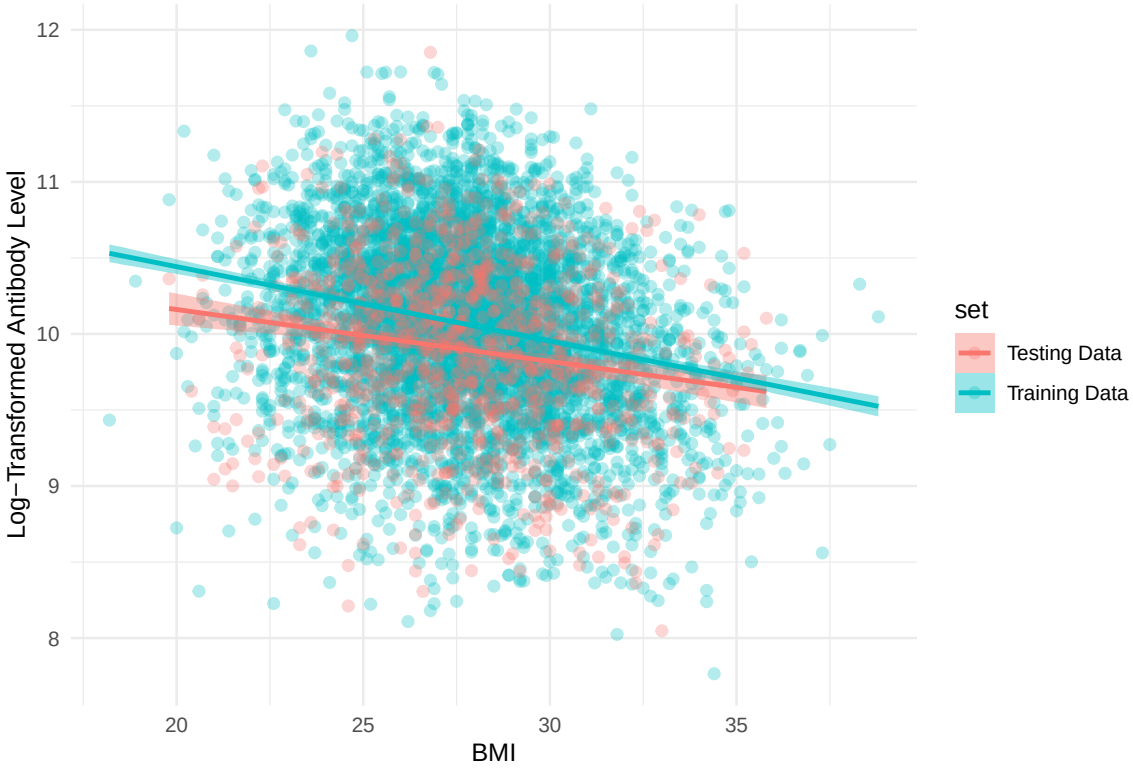


Figure 8: Log−Transformed Antibody Level vs. Weight

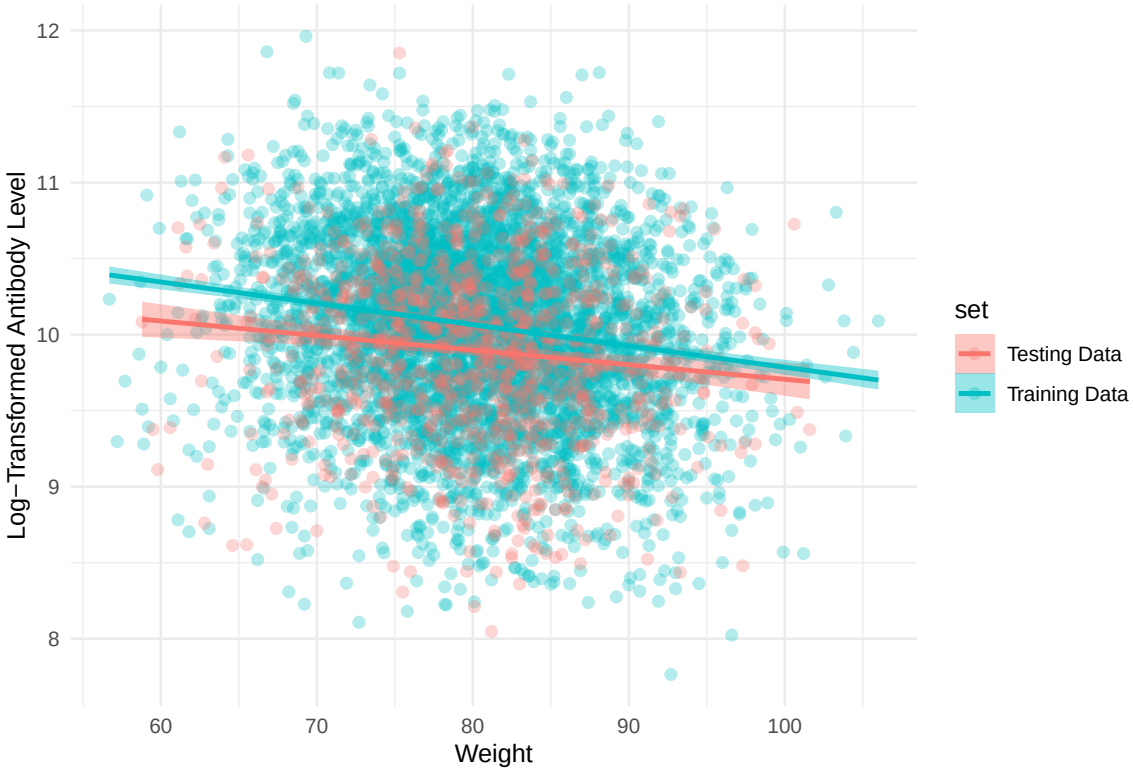

Figure 9: Log−Transformed Antibody Level vs. Age