Homework 4

Brandon Chung, Jiaxin Zheng, Andreina Arias, and Stephanie Chiang

Fall 2025

OBJECTIVE: To build multiple linear regression and binary logistic regression models on the training data to predict the probability that: - a person will crash their car - the amount of money it will cost if the person does crash their car

1. DATA EXPLORATION

Explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

VARIABLE NAME DEFINITION THEORETICAL EFFECT

- INDEX Identification Variable (do not use) None
- TARGET_FLAG Was Car in a crash? 1=YES 0=NO None
- TARGET_AMT If car was in a crash, what was the cost None
- AGE Age of Driver Very young people tend to be risky. Maybe very old people also.
- BLUEBOOK Value of Vehicle Unknown effect on probability of collision, but probably effect the payout if there is a crash
- CAR_AGE Vehicle Age Unknown effect on probability of collision, but probably effect the payout if there is a crash
- CAR_TYPE Type of Car Unknown effect on probability of collision, but probably effect the payout if there is a crash
- CAR_USE Vehicle Use Commercial vehicles are driven more, so might increase probability of collision
- CLM_FREQ # Claims (Past 5 Years) The more claims you filed in the past, the more you are likely to file in the future
- EDUCATION Max Education Level Unknown effect, but in theory more educated people tend to drive more safely
- HOMEKIDS # Children at Home Unknown effect
- HOME_VAL Home Value In theory, home owners tend to drive more responsibly
- INCOME Income In theory, rich people tend to get into fewer crashes
- JOB Job Category In theory, white collar jobs tend to be safer
- KIDSDRIV # Driving Children When teenagers drive your car, you are more likely to get into crashes
- MSTATUS Marital Status In theory, married people drive more safely
- MVR_PTS Motor Vehicle Record Points If you get lots of traffic tickets, you tend to get into more crashes
- OLDCLAIM Total Claims (Past 5 Years) If your total payout over the past five years was high, this suggests future payouts will be high
- PARENT1 Single Parent Unknown effect
- RED_CAR A Red Car Urban legend says that red cars (especially red sports cars) are more risky. Is that true?

- REVOKED License Revoked (Past 7 Years) If your license was revoked in the past 7 years, you probably are a more risky driver.
- SEX Gender Urban legend says that women have less crashes then men. Is that true?
- TIF Time in Force People who have been customers for a long time are usually more safe.
- TRAVTIME Distance to Work Long drives to work usually suggest greater risk
- URBANICITY Home/Work Area Unknown
- YOJ Years on Job People who stay at a job for a long time are usually more safe

2. DATA PREPARATION

```
train_df <- read.csv("https://raw.githubusercontent.com/Chung-Brandon/621-Data-Mining/refs/heads/main/in

test_df <- read.csv("https://raw.githubusercontent.com/Chung-Brandon/621-Data-Mining/refs/heads/main/in
```

```
# Clean column values and convert to appropriate data types
train_df <- train_df |>
  mutate(across(
    c(TARGET_FLAG, PARENT1, MSTATUS, SEX, EDUCATION, JOB, CAR_USE, CAR_TYPE, RED_CAR, REVOKED,  URBANIC
    ~ as.factor(gsub("z_|<", "", .x))
  )) |>
  mutate(across(
    c(INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM),
    ~ as.numeric(gsub("[\\$,]", "", .x))
  ))

test_df <- test_df |>
  mutate(across(
    c(TARGET_FLAG, PARENT1, MSTATUS, SEX, EDUCATION, JOB, CAR_USE, CAR_TYPE, RED_CAR, REVOKED,  URBANIC
    ~ as.factor(gsub("z_|<", "", .x))
  )) |>
  mutate(across(
    c(INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM),
    ~ as.numeric(gsub("[\\$,]", "", .x))
  ))

glimpse(train_df)
```

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <fct> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <dbl> 67349, 91449, 16039, NA, 114986, 125301, 18755, 107961, 62~
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, No, No, N~
## $ HOME_VAL    <dbl> 0, 257252, 124191, 306251, 243925, 0, NA, 333680, 0, 0, 0,~
## $ MSTATUS     <fct> No, No, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Yes,~
## $ SEX         <fct> M, M, F, M, F, F, F, M, F, M, F, F, M, M, F, F, M, F, F, F~
```

```
## $ EDUCATION   <fct> PhD, High School, High School, High School, PhD, Bachelors~
## $ JOB         <fct> Professional, Blue Collar, Clerical, Blue Collar, Doctor, ~
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Commercial~
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 16970, 1120~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <fct> Minivan, Minivan, SUV, Minivan, SUV, Sports Car, SUV, Van,~
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no, yes, y~
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, 5028, 0,~
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, No,~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Urban/ Ur~
```

```r
glimpse(test_df)
```

```
## Rows: 2,141
## Columns: 26
## $ INDEX       <int> 3, 9, 10, 18, 21, 30, 31, 37, 39, 47, 60, 62, 63, 64, 68, ~
## $ TARGET_FLAG <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ TARGET_AMT  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ KIDSDRIV    <int> 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 1, 0, 0, 0, 2, 0, 0, 0~
## $ AGE         <int> 48, 40, 44, 35, 59, 46, 60, 54, 36, 50, 42, 41, 37, 36, 34~
## $ HOMEKIDS    <int> 0, 1, 2, 2, 0, 0, 0, 0, 2, 0, 0, 2, 2, 3, 3, 2, 2, 0, 0, 0~
## $ YOJ         <int> 11, 11, 12, NA, 12, 14, 12, 12, 12, 8, NA, 7, 13, 12, 12, ~
## $ INCOME      <dbl> 52881, 50815, 43486, 21204, 87460, NA, 37940, 33212, 13054~
## $ PARENT1     <fct> No, Yes, Yes, Yes, No, No, No, No, Yes, No, No, Yes, Yes, ~
## $ HOME_VAL    <dbl> 0, 0, 0, 0, 0, 207519, 182739, 158432, 344195, 0, 176275, ~
## $ MSTATUS     <fct> No, No, No, No, No, Yes, Yes, Yes, No, No, Yes, No, No, Ye~
## $ SEX         <fct> M, M, F, M, M, M, F, M, F, F, M, F, F, M, M, F, F, M, F, F~
## $ EDUCATION   <fct> Bachelors, High School, High School, High School, High Sch~
## $ JOB         <fct> Manager, Manager, Blue Collar, Clerical, Manager, Professi~
## $ TRAVTIME    <int> 26, 21, 30, 74, 45, 7, 16, 27, 5, 22, 24, 29, 62, 15, 26, ~
## $ CAR_USE     <fct> Private, Private, Commercial, Private, Private, Commercial~
## $ BLUEBOOK    <dbl> 21970, 18930, 5900, 9230, 15420, 25660, 11290, 24000, 2720~
## $ TIF         <int> 1, 6, 10, 6, 1, 1, 1, 4, 4, 4, 1, 1, 4, 6, 4, 5, 4, 6, 4, ~
## $ CAR_TYPE    <fct> Van, Minivan, SUV, Pickup, Minivan, Panel Truck, Sports Ca~
## $ RED_CAR     <fct> yes, no, no, no, yes, no, no, no, no, no, no, yes, no, no,~
## $ OLDCLAIM    <dbl> 0, 3295, 0, 0, 44857, 2119, 0, 0, 0, 0, 0, 0, 0, 2045, 0, ~
## $ CLM_FREQ    <int> 0, 1, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 1, 0, 2, 1~
## $ REVOKED     <fct> No, No, No, Yes, No, No, No, No, No, No, No, Yes, No, No, ~
## $ MVR_PTS     <int> 2, 2, 0, 0, 4, 2, 0, 5, 0, 3, 2, 2, 0, 2, 0, 4, 8, 1, 2, 2~
## $ CAR_AGE     <int> 10, 1, 10, 4, 1, 12, 1, NA, 9, 1, 11, 1, NA, 16, 1, 4, NA,~
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Rural/ Ru~
```

Next, we view at a summary of each features. We see majority values are 0.

```r
train_df |>
  summary() |>
  kable() |>
  kable_styling()
```

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
Min. : 1	0:6008	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000	Min. : 0.0
1st Qu.: 2559	1:2153	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.: 9
Median : 5133	NA	Median : 0	Median :0.0000	Median :45.00	Median :0.0000	Median :1
Mean : 5152	NA	Mean : 1504	Mean :0.1711	Mean :44.79	Mean :0.7212	Mean :10.
3rd Qu.: 7745	NA	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:1
Max. :10302	NA	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000	Max. :23.
NA	NA	NA	NA	NA's :6	NA	NA's :454

Both datasets feature 26 variables; the training set contains 8161 observations and the testing set has 2141.

There are 5 columns contain missing data. Since AGE is only missing in 6 rows, those observations can be dropped. For the rest of the variables, we can impute missing data using the MICE package.

```
colSums(is.na(train_df))
```

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           6           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##         454         445           0         464           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##         510           0
```

```
train_df_imp <- train_df |> drop_na(AGE)
train_df_imp <- complete(mice(train_df_imp, m = 1, method = "pmm", seed = 123))
```

```
##
##  iter imp variable
##   1   1  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   1  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   1  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   1  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   1  YOJ  INCOME  HOME_VAL  CAR_AGE
```

```
colSums(is.na(train_df_imp))
```

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           0           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##           0           0           0           0           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##           0           0
```

The testing data also had missing data used mice a well to impute missing data without imputing data to the target variables. The target variables were turned to empty strings to avoid mice from filling in those variables.

```
# specify methods for each column
methVar <- make.method(test_df)

# set target columns to empty strings "" so mice will not touch them
methVar[methVar != ""] <- "pmm"
methVar["TARGET_FLAG"] <- ""
methVar["TARGET_AMT"]  <- ""

imp <- (mice(test_df, m = 1, method = methVar, seed = 123))
```

```
##
##  iter imp variable
##   1   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE

## Warning: Number of logged events: 2
```

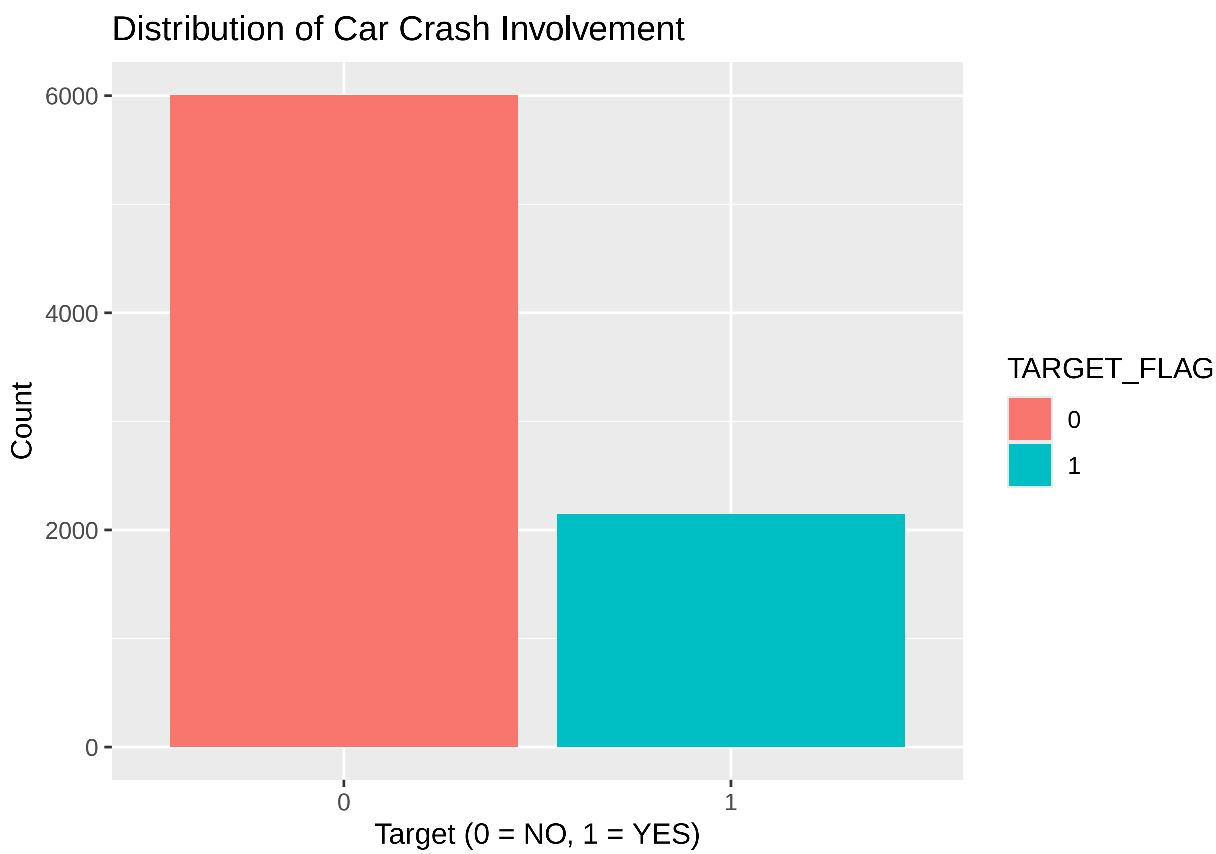
```
test_df_imp <- complete(imp)

test_df_imp$TARGET_AMT <- as.numeric(test_df$TARGET_AMT) #ensure the target variable is numeric

colSums(is.na(train_df_imp))
```

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           0           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##           0           0           0           0           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##           0           0
```

Distributions

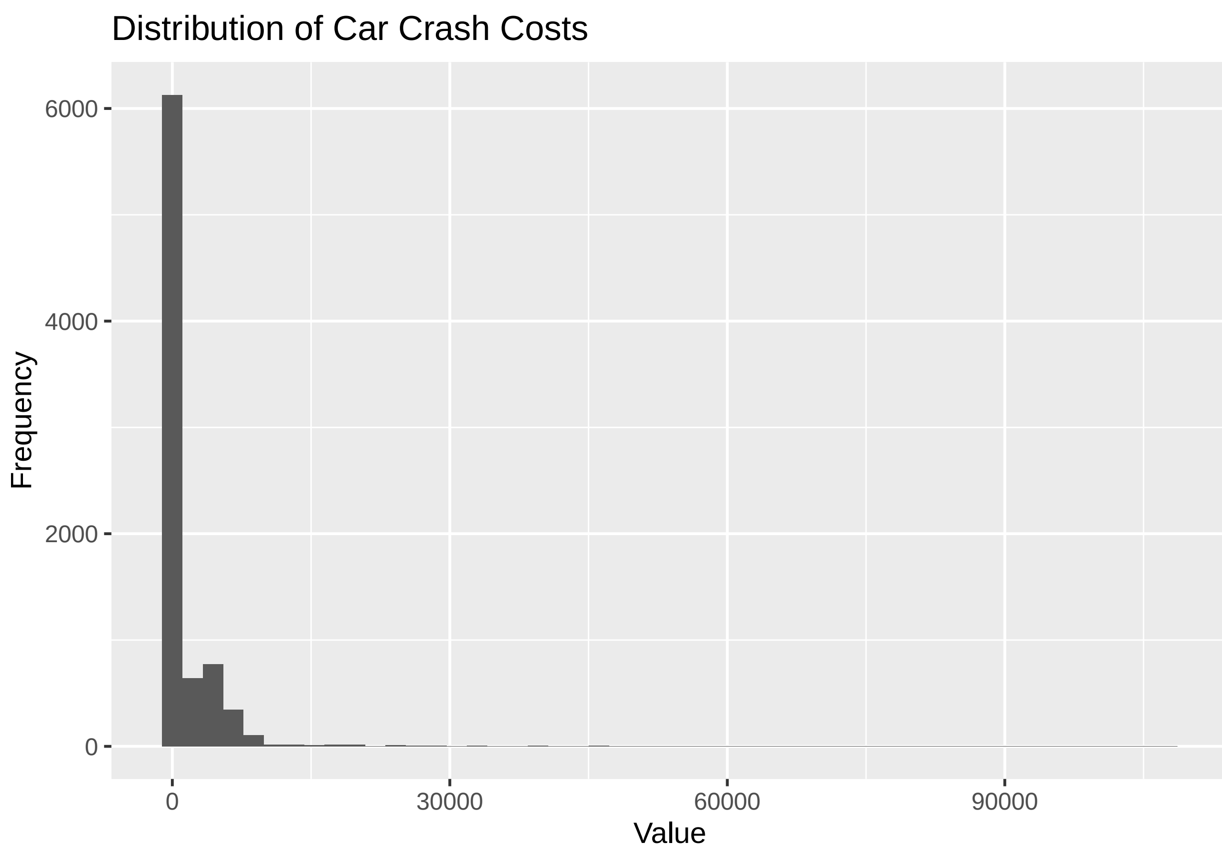
Now we can visualize the distributions for our target variables in the training set.

```
ggplot(train_df_imp, aes(x = TARGET_FLAG, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(
    title = "Distribution of Car Crash Involvement",
    x = "Target (0 = NO, 1 = YES)",
    y = "Count"
  )
```

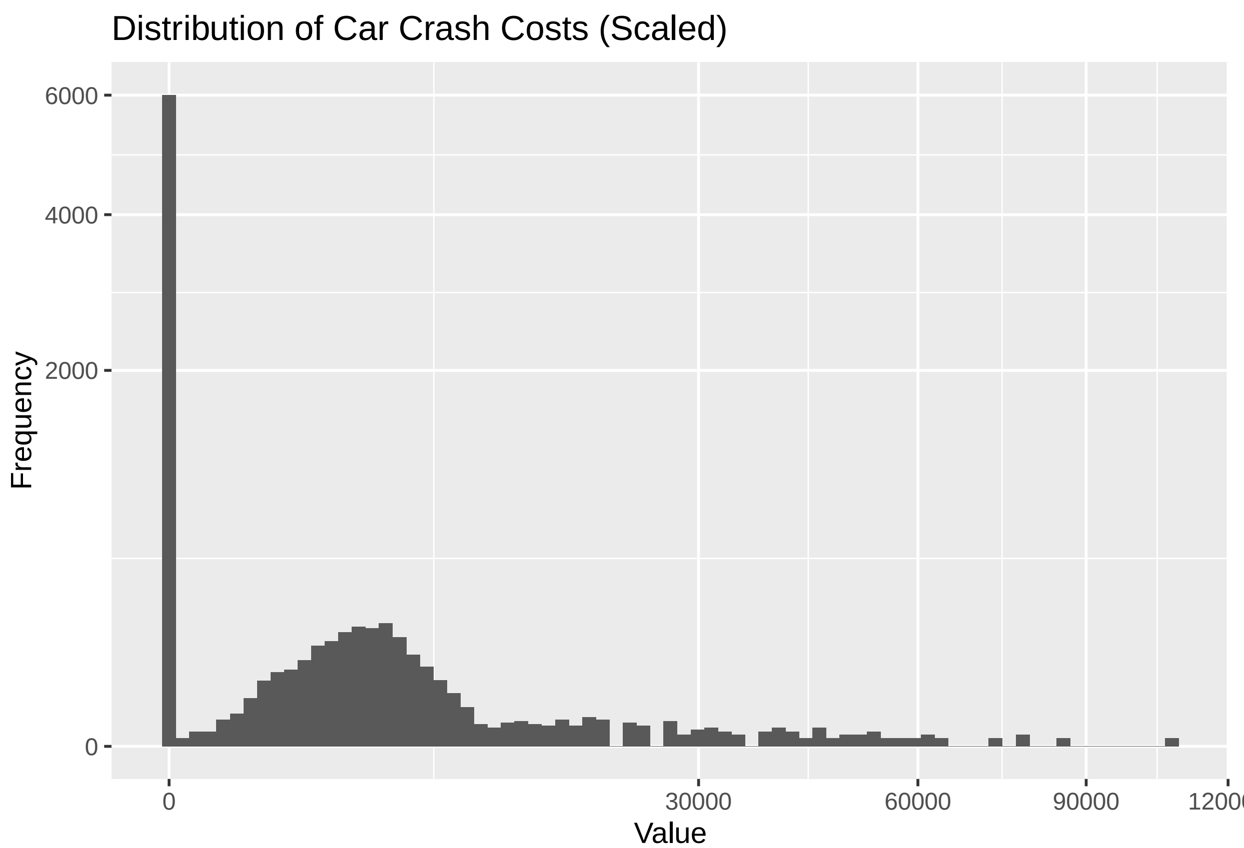


- The distribution of TARGET_AMT is extremely right-skewd. - There are approximately 80% on the rows have a target amount of 0.

```
ggplot(train_df_imp, aes(x = TARGET_AMT)) +
  geom_histogram(bins = 50) +
  labs(title = "Distribution of Car Crash Costs", x = "Value", y = "Frequency")
```




```r
ggplot(train_df_imp, aes(x = TARGET_AMT)) +
  geom_histogram(bins = 75) +
  scale_x_sqrt() +
  scale_y_sqrt() +
  labs(title = "Distribution of Car Crash Costs (Scaled)", x = "Value", y = "Frequency")
```

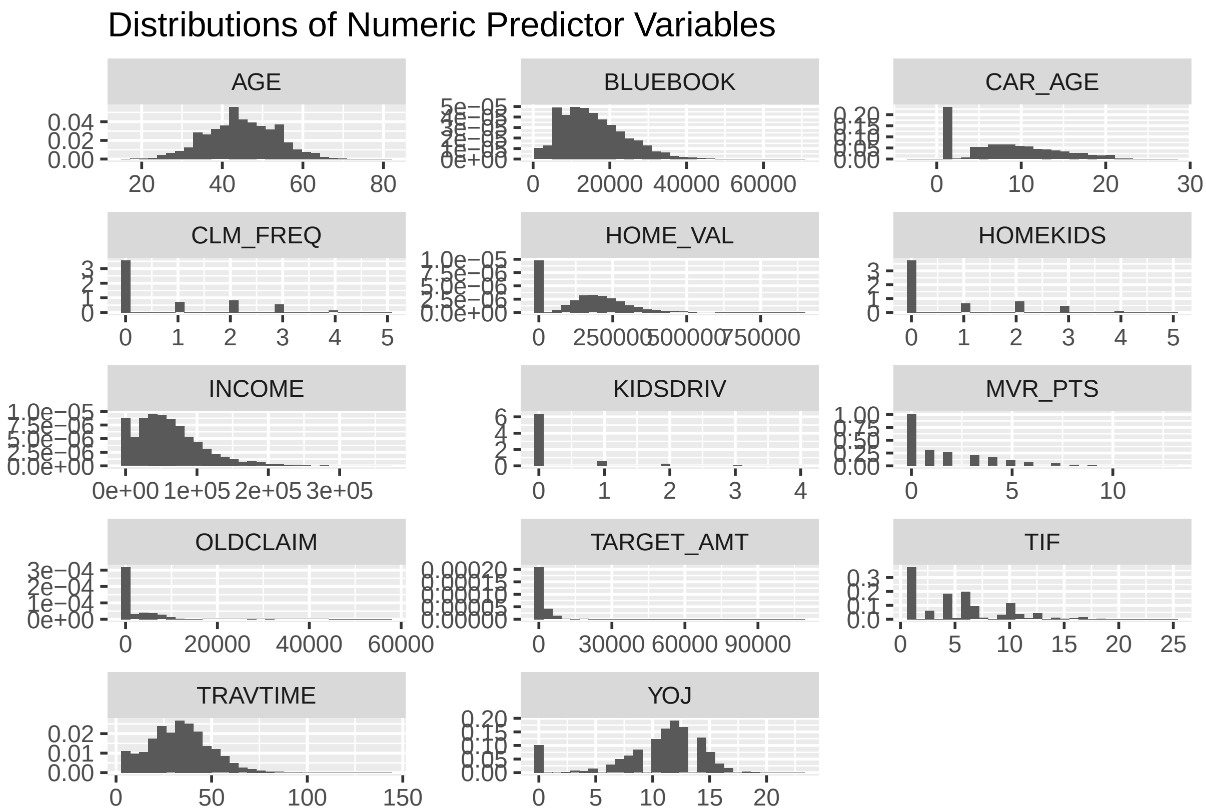


Distributions for some of the predictor variables can be seen below as well. - The distribution profile show specifically right skew with long tails in variables INCOME, OLDCLAIM, BLUEBOOK, HOME_VAL,TARGET_AMT.These variable contain many small values and a small number of extremely large values, this indicating the presence of outliers. - AGE, TRAVTIME, and YOJ are exhibit a mix of distribution shapes.

```
# Select numeric columns
num_vars <- train_df_imp |>
  as_tibble() |>
  dplyr::select(where(is.numeric)) |>
  dplyr::select(-any_of(c("target", "INDEX")))

# Reshape to long format
long_df <- num_vars |>
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value")

# Histograms
ggplot(long_df, aes(x = value)) +
  geom_histogram(aes(y = ..density..), bins = 30) +
  facet_wrap(~ variable, scales = "free", ncol = 3) +
  labs(title = "Distributions of Numeric Predictor Variables", y = "", x = "")
```

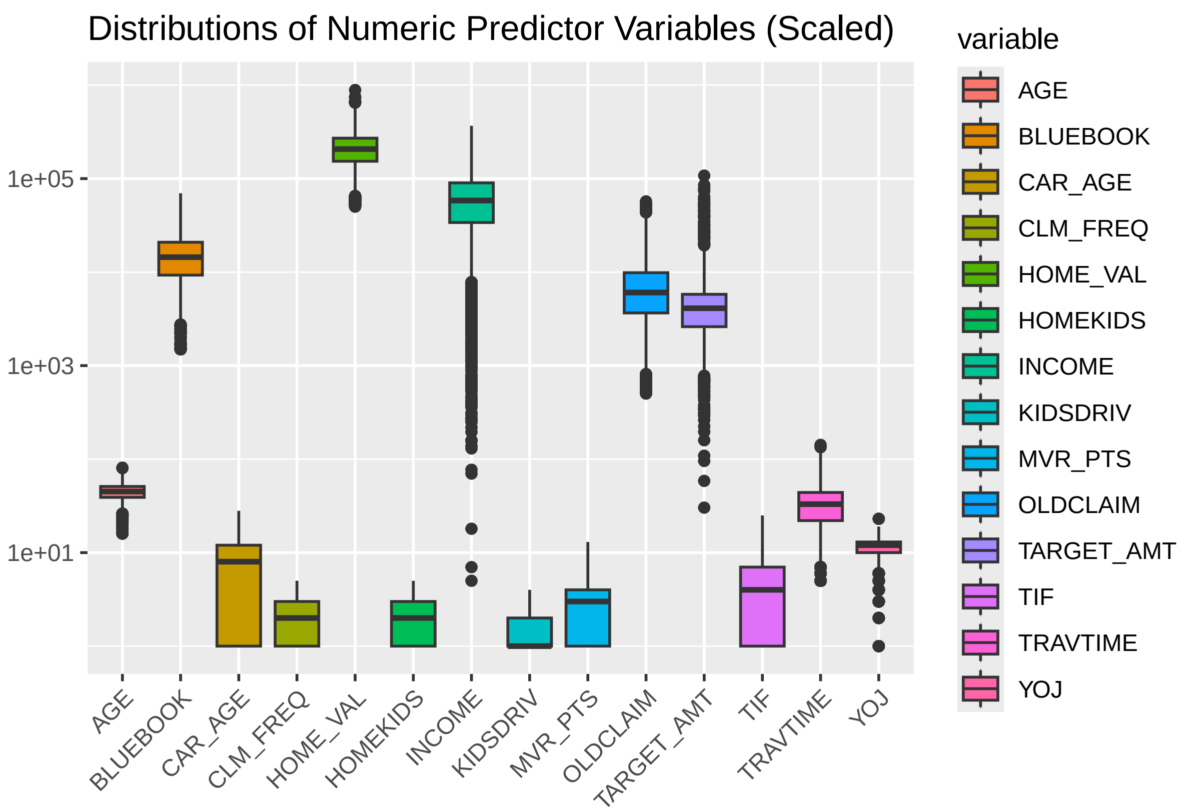
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
```

```
## generated.
```



- Several variables, such as INCOME, HOME_VAL, BLUEBOOK, and TARGET_AMT show a large spread and long right tails with many outliers, this is consistent with what was seen in the histograms above.
- In contrast, variables such as AGE, YOJ, and TIF exhibit more compact and symmetric distributions, with fewer extreme values.

```r
# Box plots
ggplot(long_df, aes(x = variable, y = value, fill = variable)) +
  geom_boxplot() +
  scale_y_log10() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distributions of Numeric Predictor Variables (Scaled)", y = "", x = "")
```



And here is a correlation matrix for the numeric variables.

- The correlation heatmap shows that most numeric predictor variables in the dataset have very weak linear relationships with one another. The majority of correlation values fall between –0.10 and 0.30. This is helpful for regression modeling because it reduces instability in coefficient estimates.

A few moderate correlations appear:

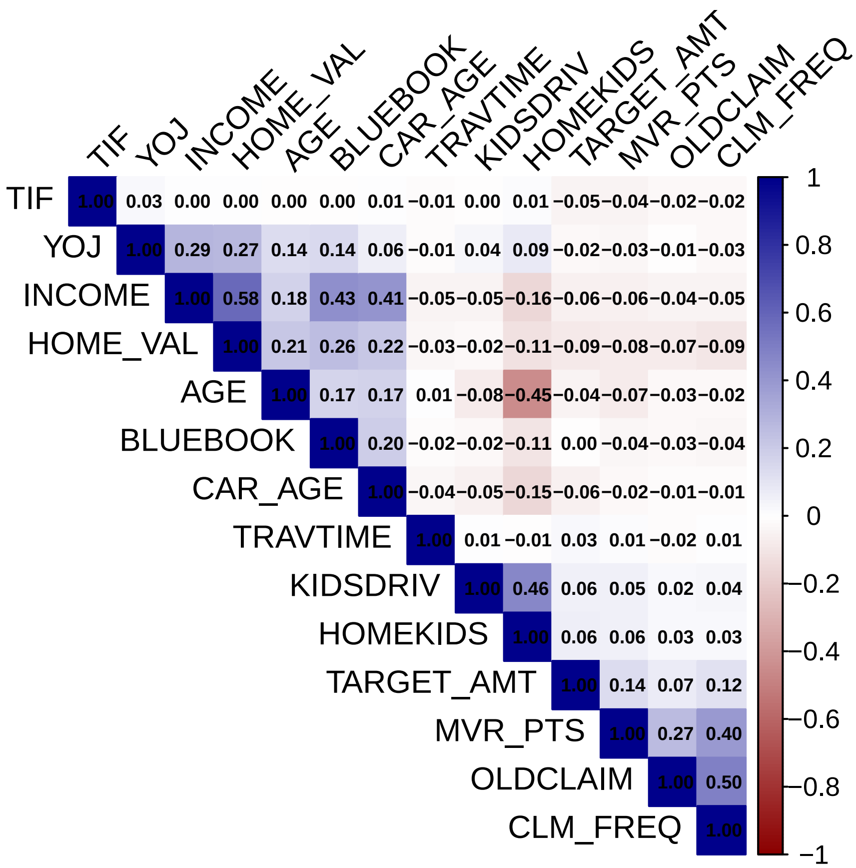
- INCOME and HOME_VAL (0.58): Higher-income customers tend to live in higher-valued homes, which is expected.
- CLM_FREQ and OLDCLAIM (0.50): Customers with more past claims tend to have higher total past claim costs.
- HOME_VAL and BLUEBOOK (0.22): House value is weakly related to vehicle value.
- CAR_AGE and BLUEBOOK show a small negative correlation (–0.15): Newer cars tend to have higher value.

```
# Compute the correlation matrix
corr_matrix <- cor(num_vars)
corr_matrix[!is.finite(corr_matrix)] <- 0
hclust(as.dist(1 - corr_matrix))
```

```
##
## Call:
```

```
## hclust(d = as.dist(1 - corr_matrix))
##
## Cluster method   : complete
## Number of objects: 14
```

```r
# Plot correlation heatmap
corrplot(
  corr_matrix,
  method = "color",
  type = "upper",
  order = "hclust",
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45,
  number.cex = 0.6,
  col = colorRampPalette(c("darkred", "white", "darkblue"))(200)
)
```




```r
# Get upper triangle of the correlation matrix
upper_tri <- upper.tri(corr_matrix)

# Find pairs with correlation > 0.8 (absolute)
high_corr_pairs <- which(abs(corr_matrix) > 0.8 & upper_tri, arr.ind = TRUE)

# Display variable pairs and their correlation
```

```r
data.frame(
  Var1 = rownames(corr_matrix)[high_corr_pairs[, 1]],
  Var2 = colnames(corr_matrix)[high_corr_pairs[, 2]],
  Correlation = corr_matrix[high_corr_pairs]
)
```

```
## [1] Var1        Var2        Correlation
## <0 rows> (or 0-length row.names)
```

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. b. c. d. e. Fix missing values (maybe with a Mean or Median value) Create flags to suggest if a variable was missing Transform data by putting it into buckets Mathematical transforms such as log or square root (or use Box-Cox) Combine variables (such as ratios or adding or multiplying) to create new variables

Some numerical variables seemed to be skewed such as: "INCOME", "TRAVTIME", "OLDCLAIM", "HOME_VAL" and "BLUEBOOK" were skewed to the right. Since right-skewed variables violate regression assumptions, we did a log tranformation to make the variables normally distributed. We transformed skewed numeric variables that were not count variables. For the transformation we just added new variables just in case we wanted to use the non-transformed variable for modeling. The transformation can help reduces sensitivity to outliers and help to stabilizes variance.

```r
Skewed_vars<- c("INCOME", "HOME_VAL", "BLUEBOOK", "OLDCLAIM", "TRAVTIME")
```

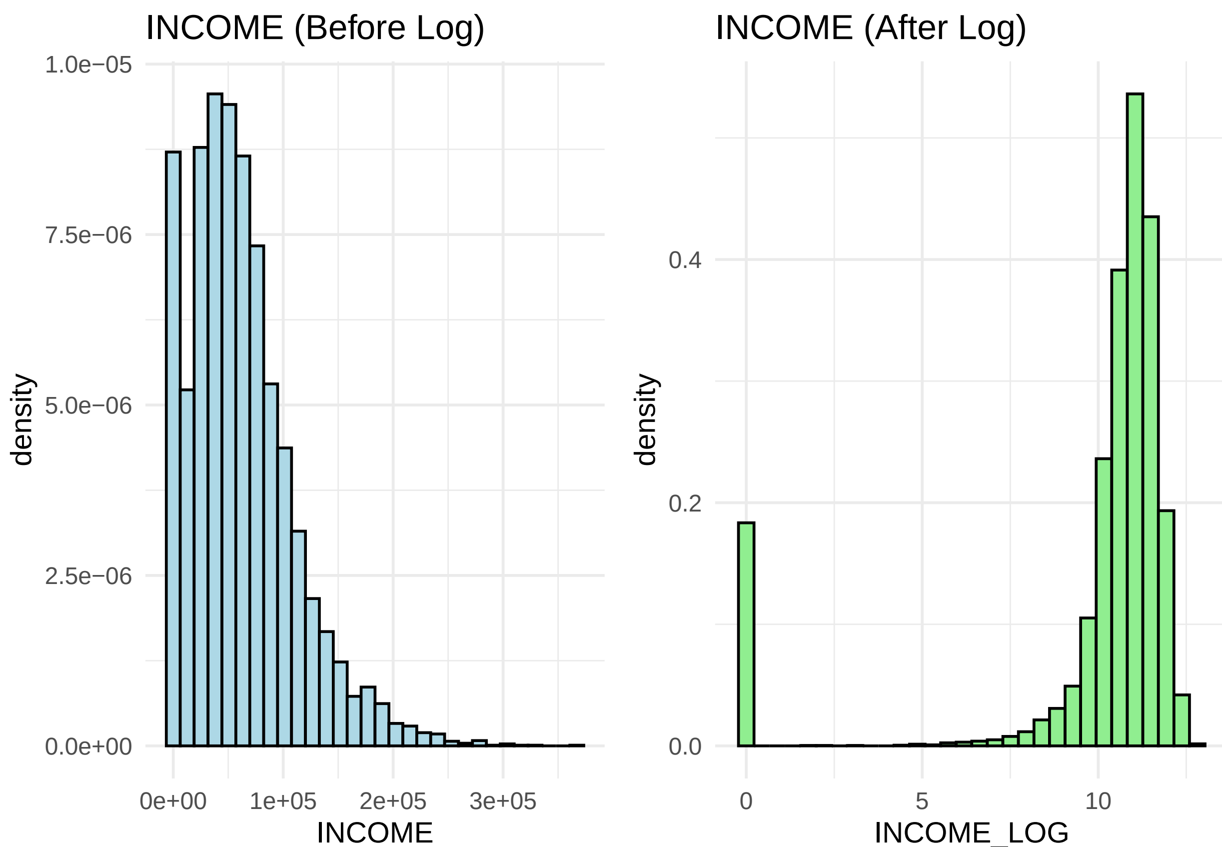
```r
#Applied log transformation using log(x+1) to avoid log(0)
for(v in Skewed_vars){
  new_name<-paste0(v,"_LOG")
  train_df_imp[[new_name]]<-log(train_df_imp[[v]]+1)
}
```

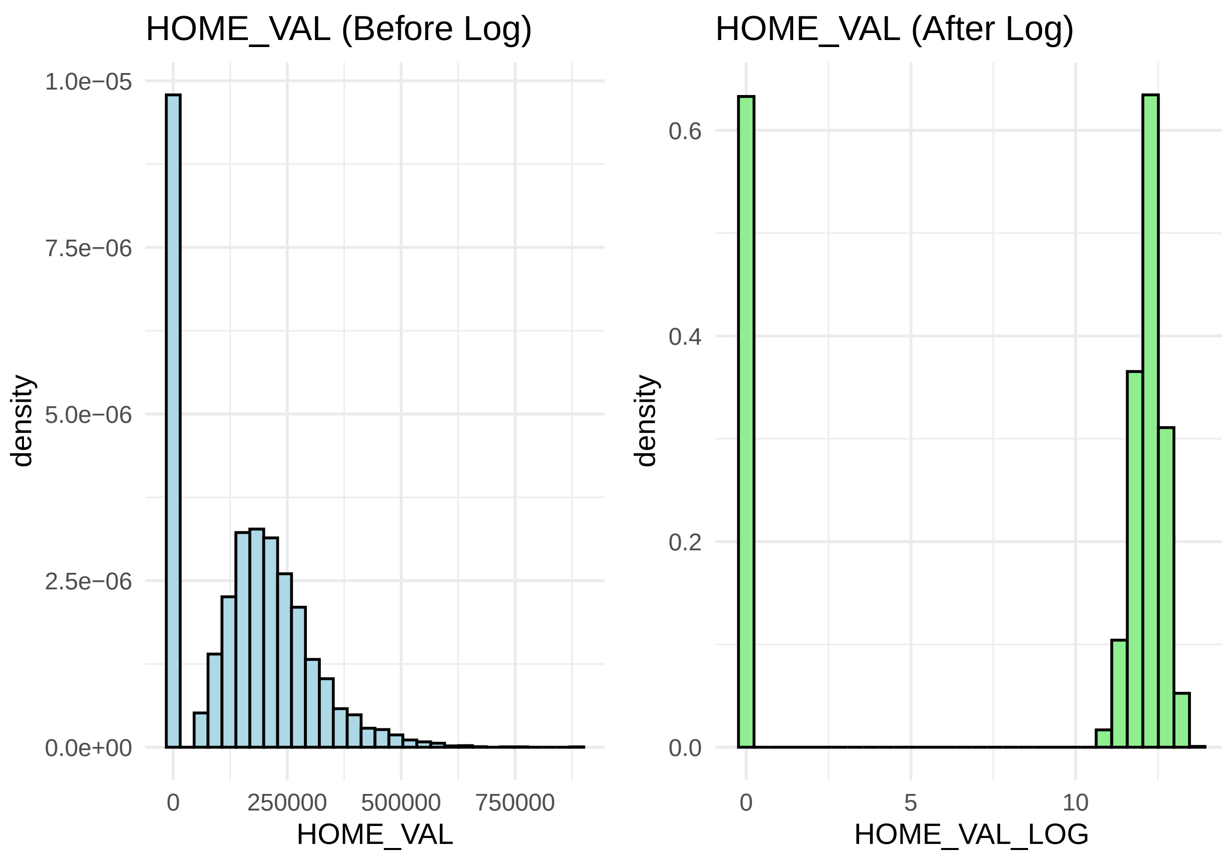
```r
#Also Applied transformation to test dataframe
for(v in Skewed_vars){
  new_name <- paste0(v, "_LOG")
  test_df_imp[[new_name]] <- log(test_df_imp[[v]] + 1)
}
```

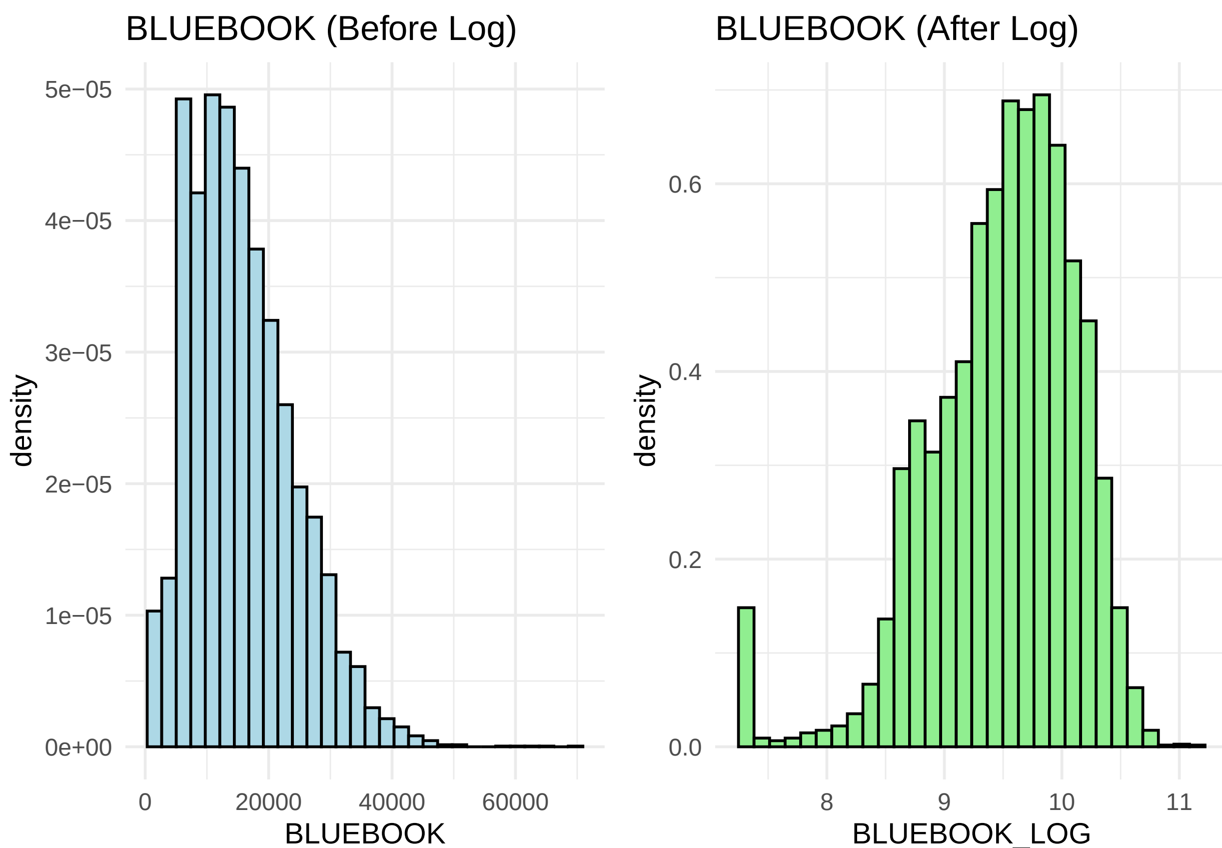
Comparsion graph

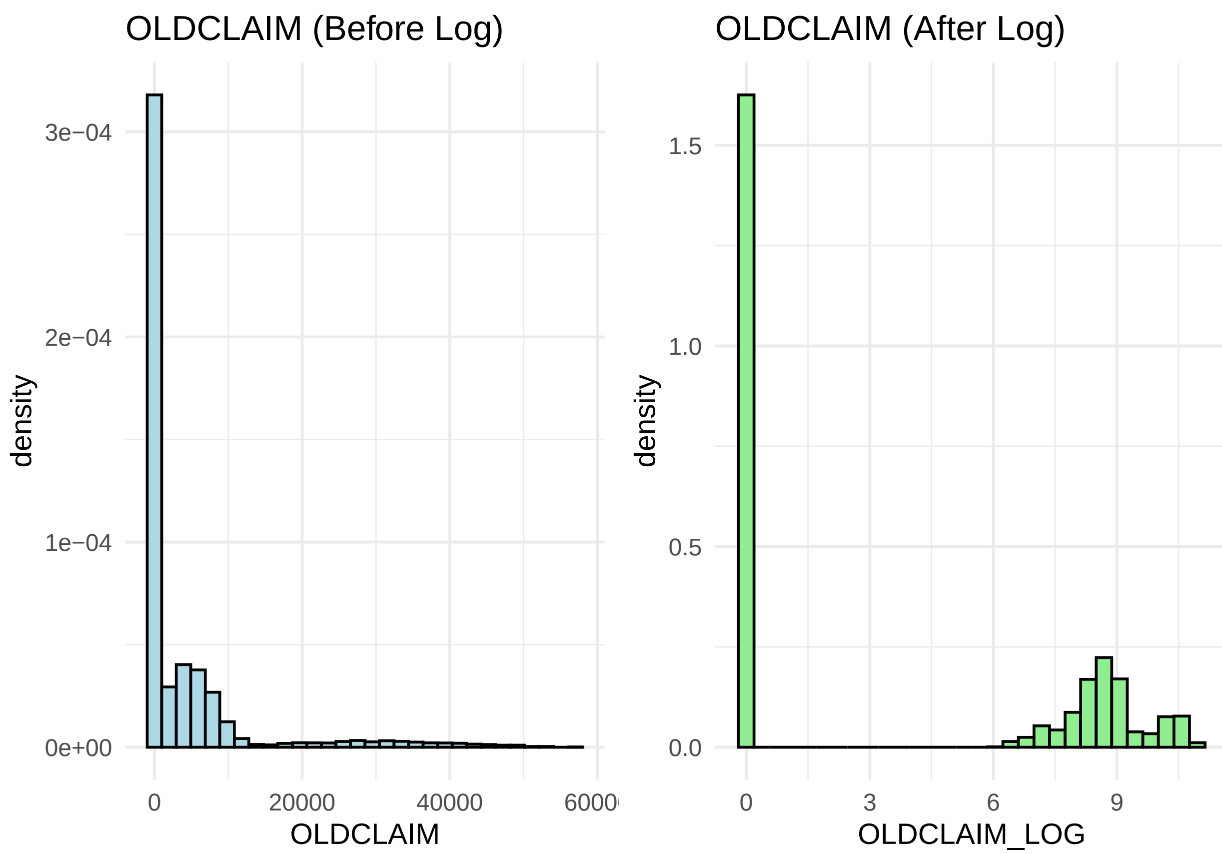
```r
plot_list <- list()

for (v in Skewed_vars) {
  new_v <- paste0(v, "_LOG")

  p1 <- ggplot(train_df_imp, aes(x = .data[[v]])) +
    geom_histogram(aes(y = after_stat(density)), fill = "lightblue", color = "black", bins = 30) +
    ggtitle(paste0(v, " (Before Log)")) +
    theme_minimal()

  p2 <- ggplot(train_df_imp, aes(x = .data[[new_v]])) +
    geom_histogram(aes(y = after_stat(density)), fill = "lightgreen", color = "black", bins = 30) +
    ggtitle(paste0(v, " (After Log)")) +
```

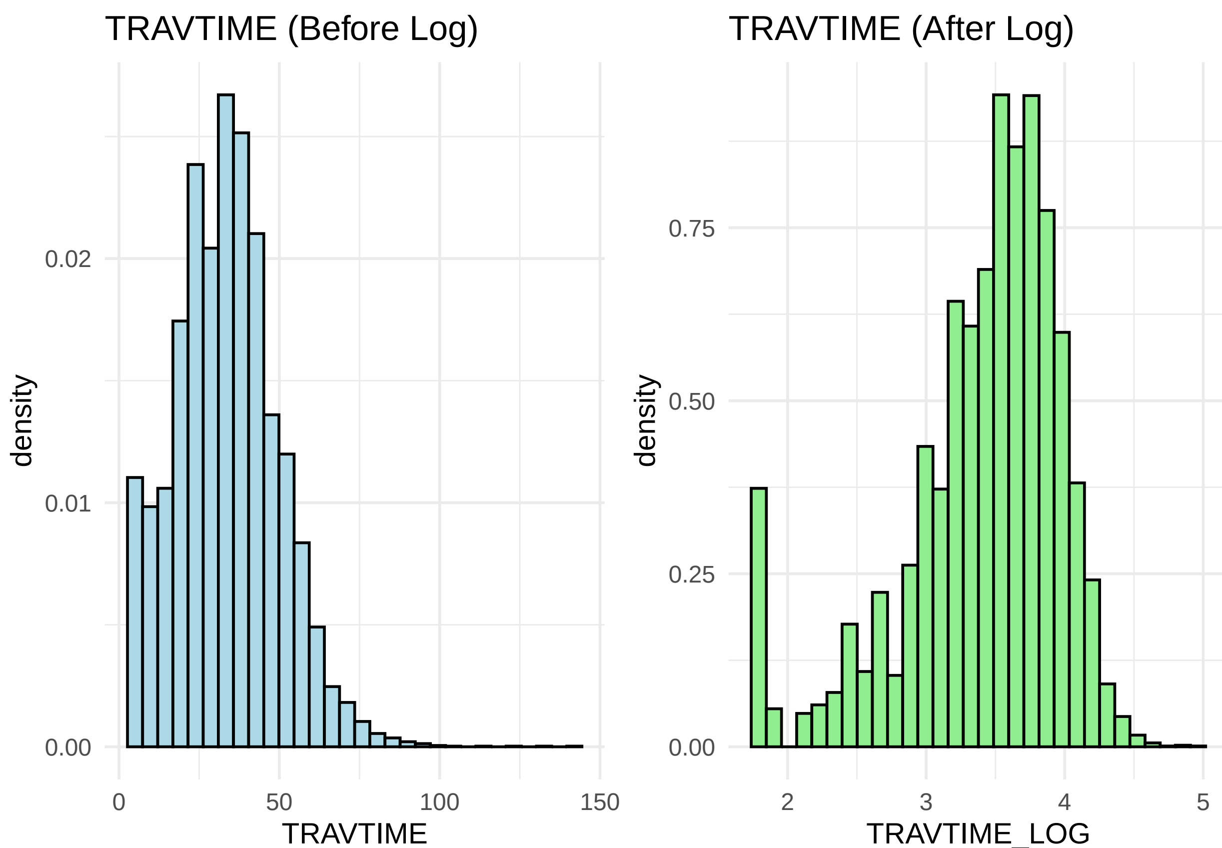
```
    theme_minimal()

# Display graphs together
  grid.arrange(p1, p2, ncol = 2)
}
```




HOME_VAL (Before Log)
1.0e−05
7.5e−06
5.0e−06
2.5e−06
0.0e+00
density
0
250000
500000
750000
HOME_VAL
HOME_VAL (After Log)
0.6
0.4
0.2
0.0
density
0
5
10
HOME_VAL_LOG


BLUEBOOK (Before Log)
density
5e−05
4e−05
3e−05
2e−05
1e−05
0e+00
0
20000
40000
60000
BLUEBOOK
BLUEBOOK (After Log)
density
0.6
0.4
0.2
0.0
8
9
10
11
BLUEBOOK_LOG


OLDCLAIM (Before Log)
density
3e−04
2e−04
1e−04
0e+00
0
20000
40000
OLDCLAIM
OLDCLAIM (After Log)
density
1.5
1.0
0.5
0.0
0
3
6
9
OLDCLAIM_LOG



3. BUILD MODELS

Using the training data set, build at least two different multiple linear regression models and three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

The three logistic regression models built to predict Target flag (whether a person will crash their car), full logistic model with all predictors, reduced logistic model, and transformed logistic model; these models are different in complexity and interpretability which will allow us to compare performance and coefficients.

Logistic model for Target flag using all variables, as target flag is a binary variable. This model could be used for predicting crash probability. The coefficient show change in crash probability with each predictor, a positive coefficients means a higher crash probability and a negative coefficient would mean a lower crash probability. OLDClaim LOG and CLM FREQ had positive variables as expected but MVR_PTS should have been positive as well, but it was skewed to the right and not transformed as it is a count variable.

```
# Removing Index from training and test data

train_df <- subset(train_df, select = -INDEX)
train_df_imp <- subset(train_df_imp, select = -INDEX)
```

```r
test_df <- subset(test_df, select = -INDEX)

# Removing TARGET_AMT from data for logistic regression for TARGET_FLAG

train_df_log <- subset(train_df, select = -TARGET_AMT)
train_df_imp_log <- subset(train_df_imp, select = -TARGET_AMT)
test_df_log <- subset(test_df, select = -TARGET_AMT)
```

Logistic Model 1: All Predictors Without Transformed

To compare whether transformations materially improve interpretability or predictive accuracy. Log transformations were applied to reduce skewness (like OLDCLAIM, HOME VAL, BLUEBOOK), improving coefficient stability. CAR USE and CAR TYPE were included because exploratory analysis indicated differences in crash likelihood across usage categories and vehicle types. Based on these coeficient CLM_FREQ positively associated with crash probability and AGE, HOMEVAL, and BLUEBOOK negatively associated, suggesting older, wealthier drivers are safer. Also, the car type show the bigger the vechile the higher the probability the person will crash.

```r
logist_model1 <- glm(
  TARGET_FLAG ~ . - INCOME_LOG - HOME_VAL_LOG - BLUEBOOK_LOG - OLDCLAIM_LOG - TRAVTIME_LOG,
  data = train_df_imp_log,
  family = binomial
)

summary(logist_model1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - INCOME_LOG - HOME_VAL_LOG - BLUEBOOK_LOG -
##     OLDCLAIM_LOG - TRAVTIME_LOG, family = binomial, data = train_df_imp_log)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.268e+00  3.381e-01  -9.664  < 2e-16 ***
## KIDSDRIV                      3.911e-01  6.129e-02   6.380 1.77e-10 ***
## AGE                          -1.742e-03  4.033e-03  -0.432 0.665757
## HOMEKIDS                      4.357e-02  3.727e-02   1.169 0.242411
## YOJ                          -6.791e-03  8.485e-03  -0.800 0.423527
## INCOME                       -3.913e-06  1.127e-06  -3.473 0.000514 ***
## PARENT1Yes                    3.749e-01  1.098e-01   3.415 0.000639 ***
## HOME_VAL                     -1.201e-06  3.471e-07  -3.460 0.000540 ***
## MSTATUSYes                   -5.036e-01  8.527e-02  -5.905 3.52e-09 ***
## SEXM                          9.099e-02  1.121e-01   0.812 0.417004
## EDUCATIONHigh School          3.876e-01  8.841e-02   4.384 1.16e-05 ***
## EDUCATIONMasters              9.824e-02  1.399e-01   0.702 0.482568
## EDUCATIONPhD                  2.506e-01  1.806e-01   1.388 0.165272
## JOBBlue Collar                2.951e-01  1.853e-01   1.592 0.111284
## JOBClerical                   3.810e-01  1.972e-01   1.932 0.053379 .
## JOBDoctor                    -4.571e-01  2.671e-01  -1.711 0.087068 .
## JOBHome Maker                 2.030e-01  2.113e-01   0.961 0.336651
## JOBLawyer                     9.662e-02  1.690e-01   0.572 0.567478
## JOBManager                   -5.613e-01  1.714e-01  -3.274 0.001059 **
```

```
## JOBProfessional                  1.567e-01  1.783e-01   0.879 0.379617
## JOBStudent                       1.849e-01  2.158e-01   0.857 0.391441
## TRAVTIME                         1.463e-02  1.883e-03   7.770 7.85e-15 ***
## CAR_USEPrivate                  -7.650e-01  8.743e-02  -8.750  < 2e-16 ***
## BLUEBOOK                        -2.018e-05  5.277e-06  -3.824 0.000131 ***
## TIF                             -5.559e-02  7.352e-03  -7.561 4.00e-14 ***
## CAR_TYPEPanel Truck              5.535e-01  1.605e-01   3.448 0.000565 ***
## CAR_TYPEPickup                   5.546e-01  1.001e-01   5.542 2.99e-08 ***
## CAR_TYPESports Car               1.026e+00  1.299e-01   7.894 2.93e-15 ***
## CAR_TYPESUV                      7.692e-01  1.113e-01   6.911 4.82e-12 ***
## CAR_TYPEVan                      6.124e-01  1.261e-01   4.857 1.19e-06 ***
## RED_CARyes                      -2.376e-02  8.665e-02  -0.274 0.783958
## OLDCLAIM                        -1.412e-05  3.911e-06  -3.611 0.000306 ***
## CLM_FREQ                         1.968e-01  2.856e-02   6.891 5.56e-12 ***
## REVOKEDYes                       8.903e-01  9.133e-02   9.748  < 2e-16 ***
## MVR_PTS                          1.132e-01  1.362e-02   8.307  < 2e-16 ***
## CAR_AGE                         -5.609e-04  7.528e-03  -0.075 0.940598
## URBANICITYHighly Urban/ Urban  2.384e+00  1.128e-01  21.137  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9404.0  on 8154  degrees of freedom
## Residual deviance: 7289.3  on 8118  degrees of freedom
## AIC: 7363.3
##
## Number of Fisher Scoring iterations: 5
```

```
pred_probs <- predict(logist_model1, train_df_imp, type = "response")
pred_class <- ifelse(pred_probs > 0.5, 1, 0)

# Confusion matrix on training data
confusionMatrix(
  factor(pred_class),
  factor(train_df_imp$TARGET_FLAG),
  positive = "1"
)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5545 1236
##          1  462  912
##
##                Accuracy : 0.7918
##                  95% CI : (0.7828, 0.8006)
##     No Information Rate : 0.7366
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3932
##
##  Mcnemar's Test P-Value : < 2.2e-16
```

```
##
##             Sensitivity : 0.4246
##             Specificity : 0.9231
##          Pos Pred Value : 0.6638
##          Neg Pred Value : 0.8177
##              Prevalence : 0.2634
##          Detection Rate : 0.1118
##    Detection Prevalence : 0.1685
##       Balanced Accuracy : 0.6738
##
##        'Positive' Class : 1
##
```

Logistic Model 2: Reduced Predictors without Transformed

To compare the outcomes with the reduced transformed.

```
logist_model2 <- step(logist_model1, direction = "backward")
```

```
## Start:  AIC=7363.31
## TARGET_FLAG ~ (KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG) - INCOME_LOG -
##     HOME_VAL_LOG - BLUEBOOK_LOG - OLDCLAIM_LOG - TRAVTIME_LOG
##
##                Df Deviance    AIC
## - CAR_AGE      1   7289.3 7361.3
## - RED_CAR      1   7289.4 7361.4
## - AGE          1   7289.5 7361.5
## - YOJ          1   7290.0 7362.0
## - SEX          1   7290.0 7362.0
## - HOMEKIDS     1   7290.7 7362.7
## <none>              7289.3 7363.3
## - PARENT1      1   7301.0 7373.0
## - HOME_VAL     1   7301.3 7373.3
## - INCOME       1   7301.6 7373.6
## - OLDCLAIM     1   7302.6 7374.6
## - BLUEBOOK     1   7304.2 7376.2
## - EDUCATION    3   7310.4 7378.4
## - MSTATUS      1   7323.8 7395.8
## - KIDSDRIV     1   7330.1 7402.1
## - CLM_FREQ     1   7336.2 7408.2
## - JOB          8   7350.5 7408.5
## - TIF          1   7348.6 7420.6
## - TRAVTIME     1   7349.9 7421.9
## - MVR_PTS      1   7358.9 7430.9
## - CAR_USE      1   7367.2 7439.2
## - CAR_TYPE     5   7381.2 7445.2
## - REVOKED      1   7382.8 7454.8
## - URBANICITY   1   7934.5 8006.5
##
```

```
## Step:  AIC=7361.32
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## - RED_CAR     1   7289.4 7359.4
## - AGE         1   7289.5 7359.5
## - YOJ         1   7290.0 7360.0
## - SEX         1   7290.0 7360.0
## - HOMEKIDS    1   7290.7 7360.7
## <none>            7289.3 7361.3
## - PARENT1     1   7301.0 7371.0
## - HOME_VAL    1   7301.3 7371.3
## - INCOME      1   7301.6 7371.6
## - OLDCLAIM    1   7302.6 7372.6
## - BLUEBOOK    1   7304.2 7374.2
## - EDUCATION   3   7313.9 7379.9
## - MSTATUS     1   7323.8 7393.8
## - KIDSDRIV    1   7330.1 7400.1
## - CLM_FREQ    1   7336.2 7406.2
## - JOB         8   7350.5 7406.5
## - TIF         1   7348.6 7418.6
## - TRAVTIME    1   7349.9 7419.9
## - MVR_PTS     1   7358.9 7428.9
## - CAR_USE     1   7367.3 7437.3
## - CAR_TYPE    5   7381.3 7443.3
## - REVOKED     1   7382.8 7452.8
## - URBANICITY  1   7934.6 8004.6
##
## Step:  AIC=7359.39
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## - AGE         1   7289.6 7357.6
## - SEX         1   7290.0 7358.0
## - YOJ         1   7290.0 7358.0
## - HOMEKIDS    1   7290.7 7358.7
## <none>            7289.4 7359.4
## - PARENT1     1   7301.1 7369.1
## - HOME_VAL    1   7301.3 7369.3
## - INCOME      1   7301.7 7369.7
## - OLDCLAIM    1   7302.7 7370.7
## - BLUEBOOK    1   7304.2 7372.2
## - EDUCATION   3   7314.1 7378.1
## - MSTATUS     1   7323.9 7391.9
## - KIDSDRIV    1   7330.2 7398.2
## - CLM_FREQ    1   7336.2 7404.2
## - JOB         8   7350.7 7404.7
## - TIF         1   7348.7 7416.7
```

```
## - TRAVTIME    1   7350.0 7418.0
## - MVR_PTS     1   7358.9 7426.9
## - CAR_USE     1   7367.4 7435.4
## - CAR_TYPE    5   7381.5 7441.5
## - REVOKED     1   7382.9 7450.9
## - URBANICITY  1   7934.7 8002.7
##
## Step:  AIC=7357.57
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## - SEX         1   7290.1 7356.1
## - YOJ         1   7290.4 7356.4
## <none>            7289.6 7357.6
## - HOMEKIDS    1   7291.6 7357.6
## - PARENT1     1   7301.7 7367.7
## - INCOME      1   7301.8 7367.8
## - HOME_VAL    1   7301.8 7367.8
## - OLDCLAIM    1   7302.9 7368.9
## - BLUEBOOK    1   7305.1 7371.1
## - EDUCATION   3   7314.2 7376.2
## - MSTATUS     1   7324.1 7390.1
## - KIDSDRIV    1   7330.8 7396.8
## - CLM_FREQ    1   7336.3 7402.3
## - JOB         8   7351.5 7403.5
## - TIF         1   7348.8 7414.8
## - TRAVTIME    1   7350.0 7416.0
## - MVR_PTS     1   7359.4 7425.4
## - CAR_USE     1   7367.4 7433.4
## - CAR_TYPE    5   7381.5 7439.5
## - REVOKED     1   7383.1 7449.1
## - URBANICITY  1   7935.9 8001.9
##
## Step:  AIC=7356.1
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## - YOJ         1   7290.9 7354.9
## - HOMEKIDS    1   7292.0 7356.0
## <none>            7290.1 7356.1
## - PARENT1     1   7302.1 7366.1
## - HOME_VAL    1   7302.2 7366.2
## - INCOME      1   7302.4 7366.4
## - OLDCLAIM    1   7303.4 7367.4
## - EDUCATION   3   7314.7 7374.7
## - BLUEBOOK    1   7312.2 7376.2
## - MSTATUS     1   7324.6 7388.6
## - KIDSDRIV    1   7331.3 7395.3
```

```
## - CLM_FREQ    1   7336.9 7400.9
## - JOB         8   7351.8 7401.8
## - TIF         1   7349.4 7413.4
## - TRAVTIME    1   7350.7 7414.7
## - MVR_PTS     1   7359.8 7423.8
## - CAR_USE     1   7367.8 7431.8
## - REVOKED     1   7383.8 7447.8
## - CAR_TYPE    5   7398.4 7454.4
## - URBANICITY  1   7936.6 8000.6
##
## Step:  AIC=7354.87
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY
##
##              Df Deviance    AIC
## - HOMEKIDS    1   7292.5 7354.5
## <none>            7290.9 7354.9
## - HOME_VAL    1   7303.1 7365.1
## - PARENT1     1   7303.1 7365.1
## - INCOME      1   7303.8 7365.8
## - OLDCLAIM    1   7304.3 7366.3
## - EDUCATION   3   7315.3 7373.3
## - BLUEBOOK    1   7313.3 7375.3
## - MSTATUS     1   7326.6 7388.6
## - KIDSDRIV    1   7332.5 7394.5
## - CLM_FREQ    1   7337.8 7399.8
## - JOB         8   7352.2 7400.2
## - TIF         1   7350.3 7412.3
## - TRAVTIME    1   7351.2 7413.2
## - MVR_PTS     1   7361.2 7423.2
## - CAR_USE     1   7369.2 7431.2
## - REVOKED     1   7384.7 7446.7
## - CAR_TYPE    5   7399.5 7453.5
## - URBANICITY  1   7936.9 7998.9
##
## Step:  AIC=7354.46
## TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL + MSTATUS +
##     EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## <none>            7292.5 7354.5
## - INCOME      1   7305.0 7365.0
## - HOME_VAL    1   7305.2 7365.2
## - OLDCLAIM    1   7306.0 7366.0
## - EDUCATION   3   7317.2 7373.2
## - BLUEBOOK    1   7315.2 7375.2
## - PARENT1     1   7315.3 7375.3
## - MSTATUS     1   7326.8 7386.8
## - CLM_FREQ    1   7339.4 7399.4
## - JOB         8   7355.0 7401.0
## - KIDSDRIV    1   7349.8 7409.8
```

```
## - TIF         1   7351.5 7411.5
## - TRAVTIME    1   7352.4 7412.4
## - MVR_PTS     1   7363.2 7423.2
## - CAR_USE     1   7370.5 7430.5
## - REVOKED     1   7387.1 7447.1
## - CAR_TYPE    5   7401.5 7453.5
## - URBANICITY  1   7938.6 7998.6
```

```
formula(logist_model2)
```

```
## TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL + MSTATUS +
##     EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY
```

```
pred_probs <- predict(logist_model2, train_df_imp, type = "response")
pred_class <- ifelse(pred_probs > 0.5, 1, 0)

# Confusion matrix on training data
confusionMatrix(
  factor(pred_class),
  factor(train_df_imp$TARGET_FLAG),
  positive = "1"
)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5552 1246
##          1  455  902
##
##                Accuracy : 0.7914
##                  95% CI : (0.7824, 0.8002)
##     No Information Rate : 0.7366
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3904
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4199
##             Specificity : 0.9243
##          Pos Pred Value : 0.6647
##          Neg Pred Value : 0.8167
##              Prevalence : 0.2634
##          Detection Rate : 0.1106
##    Detection Prevalence : 0.1664
##       Balanced Accuracy : 0.6721
##
##        'Positive' Class : 1
##
```

Logistic Model 3: All Predictors with Transformations

The baseline- logistic model uses all variables except Target amt, as it is defined for claimana and should be used to predict the likelihood someone would crash. This model could be used for predicting crash probability. The coefficient show change in crash probability with each predictor, a positive coefficients means a higher crash probability and a negative coefficent would mean a lower crash probability. MVR_PTS and CLM_FREQ had positive variables as expected.

```
logist_model3 <- glm(
  TARGET_FLAG ~ . - INCOME - HOME_VAL - BLUEBOOK - OLDCLAIM - TRAVTIME,
  data = train_df_imp_log,
  family = binomial
)

summary(logist_model3)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - INCOME - HOME_VAL - BLUEBOOK -
##     OLDCLAIM - TRAVTIME, family = binomial, data = train_df_imp_log)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.1558576  0.6723053  -1.719 0.085570 .
## KIDSDRIV                      0.4126953  0.0614931   6.711 1.93e-11 ***
## AGE                          -0.0047765  0.0040649  -1.175 0.239970
## HOMEKIDS                      0.0148857  0.0376524   0.395 0.692587
## YOJ                           0.0292065  0.0107289   2.722 0.006484 **
## PARENT1Yes                    0.3724065  0.1098345   3.391 0.000697 ***
## MSTATUSYes                   -0.5013048  0.0878438  -5.707 1.15e-08 ***
## SEXM                          0.1243239  0.1083885   1.147 0.251373
## EDUCATIONHigh School          0.4325627  0.0876530   4.935 8.02e-07 ***
## EDUCATIONMasters              0.0532185  0.1403764   0.379 0.704604
## EDUCATIONPhD                  0.0496531  0.1761907   0.282 0.778086
## JOBBlue Collar                0.3866485  0.1842718   2.098 0.035883 *
## JOBClerical                   0.5101121  0.1937472   2.633 0.008467 **
## JOBDoctor                    -0.3924582  0.2642226  -1.485 0.137456
## JOBHome Maker                 0.0514228  0.2196766   0.234 0.814920
## JOBLawyer                     0.1788500  0.1679545   1.065 0.286934
## JOBManager                   -0.5048479  0.1699760  -2.970 0.002977 **
## JOBProfessional               0.2268496  0.1773153   1.279 0.200771
## JOBStudent                   -0.0836751  0.2254450  -0.371 0.710522
## CAR_USEPrivate               -0.7692292  0.0877971  -8.761  < 2e-16 ***
## TIF                          -0.0533418  0.0073443  -7.263 3.79e-13 ***
## CAR_TYPEPanel Truck           0.4632667  0.1489124   3.111 0.001865 **
## CAR_TYPEPickup                0.5895862  0.0999902   5.896 3.71e-09 ***
## CAR_TYPESports Car            1.0387666  0.1282857   8.097 5.62e-16 ***
## CAR_TYPESUV                   0.8136462  0.1078678   7.543 4.59e-14 ***
## CAR_TYPEVan                   0.6062681  0.1247455   4.860 1.17e-06 ***
## RED_CARyes                   -0.0314509  0.0866317  -0.363 0.716574
## CLM_FREQ                      0.0907825  0.0435077   2.087 0.036926 *
## REVOKEDYes                    0.7069977  0.0815059   8.674  < 2e-16 ***
## MVR_PTS                       0.1027535  0.0140730   7.301 2.85e-13 ***
## CAR_AGE                      -0.0005622  0.0075064  -0.075 0.940295
```

```
## URBANICITYHighly Urban/ Urban  2.3764133  0.1131900  20.995  < 2e-16 ***
## INCOME_LOG                    -0.1097580  0.0178294  -6.156 7.46e-10 ***
## HOME_VAL_LOG                  -0.0272343  0.0069209  -3.935 8.32e-05 ***
## BLUEBOOK_LOG                  -0.2936087  0.0592957  -4.952 7.36e-07 ***
## OLDCLAIM_LOG                   0.0222095  0.0124852   1.779 0.075263 .
## TRAVTIME_LOG                   0.4294576  0.0544868   7.882 3.23e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9404.0  on 8154  degrees of freedom
## Residual deviance: 7271.5  on 8118  degrees of freedom
## AIC: 7345.5
##
## Number of Fisher Scoring iterations: 5
```

```
pred_probs <- predict(logist_model3, train_df_imp, type = "response")
pred_class <- ifelse(pred_probs > 0.5, 1, 0)

# Confusion matrix on training data
confusionMatrix(
  factor(pred_class),
  factor(train_df_imp$TARGET_FLAG),
  positive = "1"
)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5547 1237
##          1  460  911
##
##                Accuracy : 0.7919
##                  95% CI : (0.7829, 0.8007)
##     No Information Rate : 0.7366
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3932
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4241
##             Specificity : 0.9234
##          Pos Pred Value : 0.6645
##          Neg Pred Value : 0.8177
##              Prevalence : 0.2634
##          Detection Rate : 0.1117
##    Detection Prevalence : 0.1681
##       Balanced Accuracy : 0.6738
##
##        'Positive' Class : 1
##
```

Logistic Model 4: Reduced Predictors with Transformations

Reduced logistic regression model with transformed variables. This model includes only a subset of variables chosen based on statistical significance and predictive relevance from Model 1. Many continuous predictors are log-transformed to reduce skewness and linearize relationships.

In the model, OLDCLAIM LOG is positively and highly significantly associated with the probability of a crash, indicating that larger prior claims increase risk. CLM FREQ shows a small positive association, but it is not statistically significant, so its effect is uncertain. Demographic and financial factors such as AGE, HOME VAL LOG, and BLUEBOOK LOG are negatively associated with crash probability, suggesting that older and wealthier drivers tend to be safer. Regarding vehicle type, most larger vehicles—such as SUVs, pickups, vans, and panel trucks—are associated with higher crash probability compared to the reference category. Sports cars also have the highest positive coefficient, likely reflecting riskier driving behavior rather than vehicle size alone. Overall, the model indicates that prior claims, driver age, wealth, vehicle use, and type are important predictors of crash likelihood.

```r
logist_model4 <- step(logist_model3, direction = "backward")
```

```
## Start:  AIC=7345.54
## TARGET_FLAG ~ (KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG) - INCOME - HOME_VAL -
##     BLUEBOOK - OLDCLAIM - TRAVTIME
##
##                Df Deviance    AIC
## - CAR_AGE       1   7271.5 7343.5
## - RED_CAR       1   7271.7 7343.7
## - HOMEKIDS      1   7271.7 7343.7
## - SEX           1   7272.9 7344.9
## - AGE           1   7272.9 7344.9
## <none>              7271.5 7345.5
## - OLDCLAIM_LOG  1   7274.7 7346.7
## - CLM_FREQ      1   7275.9 7347.9
## - YOJ           1   7279.0 7351.0
## - PARENT1       1   7283.1 7355.1
## - HOME_VAL_LOG  1   7287.1 7359.1
## - EDUCATION     3   7296.1 7364.1
## - BLUEBOOK_LOG  1   7295.9 7367.9
## - MSTATUS       1   7303.7 7375.7
## - INCOME_LOG    1   7309.9 7381.9
## - KIDSDRIV      1   7316.6 7388.6
## - MVR_PTS       1   7325.3 7397.3
## - TIF           1   7326.2 7398.2
## - JOB           8   7349.8 7407.8
## - TRAVTIME_LOG  1   7336.5 7408.5
## - REVOKED       1   7345.5 7417.5
## - CAR_USE       1   7349.7 7421.7
## - CAR_TYPE      5   7365.6 7429.6
## - URBANICITY    1   7903.3 7975.3
##
## Step:  AIC=7343.55
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + MSTATUS +
```

```
##     SEX + EDUCATION + JOB + CAR_USE + TIF + CAR_TYPE + RED_CAR +
##     CLM_FREQ + REVOKED + MVR_PTS + URBANICITY + INCOME_LOG +
##     HOME_VAL_LOG + BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG
##
##                Df Deviance    AIC
## - RED_CAR       1   7271.7 7341.7
## - HOMEKIDS      1   7271.7 7341.7
## - SEX           1   7272.9 7342.9
## - AGE           1   7272.9 7342.9
## <none>              7271.5 7343.5
## - OLDCLAIM_LOG  1   7274.7 7344.7
## - CLM_FREQ      1   7275.9 7345.9
## - YOJ           1   7279.0 7349.0
## - PARENT1       1   7283.1 7353.1
## - HOME_VAL_LOG  1   7287.1 7357.1
## - BLUEBOOK_LOG  1   7295.9 7365.9
## - EDUCATION     3   7301.8 7367.8
## - MSTATUS       1   7303.7 7373.7
## - INCOME_LOG    1   7310.0 7380.0
## - KIDSDRIV      1   7316.6 7386.6
## - MVR_PTS       1   7325.3 7395.3
## - TIF           1   7326.2 7396.2
## - JOB           8   7349.8 7405.8
## - TRAVTIME_LOG  1   7336.6 7406.6
## - REVOKED       1   7345.5 7415.5
## - CAR_USE       1   7349.8 7419.8
## - CAR_TYPE      5   7365.7 7427.7
## - URBANICITY    1   7903.4 7973.4
##
## Step:  AIC=7341.68
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + MSTATUS +
##     SEX + EDUCATION + JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ +
##     REVOKED + MVR_PTS + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG
##
##                Df Deviance    AIC
## - HOMEKIDS      1   7271.8 7339.8
## - SEX           1   7272.9 7340.9
## - AGE           1   7273.0 7341.0
## <none>              7271.7 7341.7
## - OLDCLAIM_LOG  1   7274.9 7342.9
## - CLM_FREQ      1   7276.0 7344.0
## - YOJ           1   7279.1 7347.1
## - PARENT1       1   7283.3 7351.3
## - HOME_VAL_LOG  1   7287.2 7355.2
## - BLUEBOOK_LOG  1   7295.9 7363.9
## - EDUCATION     3   7302.0 7366.0
## - MSTATUS       1   7303.8 7371.8
## - INCOME_LOG    1   7310.0 7378.0
## - KIDSDRIV      1   7316.9 7384.9
## - MVR_PTS       1   7325.4 7393.4
## - TIF           1   7326.3 7394.3
## - JOB           8   7350.1 7404.1
## - TRAVTIME_LOG  1   7336.7 7404.7
```

```
## - REVOKED       1   7345.7 7413.7
## - CAR_USE       1   7350.0 7418.0
## - CAR_TYPE      5   7366.0 7426.0
## - URBANICITY    1   7903.4 7971.4
##
## Step:  AIC=7339.83
## TARGET_FLAG ~ KIDSDRIV + AGE + YOJ + PARENT1 + MSTATUS + SEX +
##     EDUCATION + JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY + INCOME_LOG + HOME_VAL_LOG + BLUEBOOK_LOG +
##     OLDCLAIM_LOG + TRAVTIME_LOG
##
##                Df Deviance    AIC
## - SEX           1   7273.0 7339.0
## <none>              7271.8 7339.8
## - AGE           1   7273.9 7339.9
## - OLDCLAIM_LOG  1   7275.0 7341.0
## - CLM_FREQ      1   7276.1 7342.1
## - YOJ           1   7280.4 7346.4
## - PARENT1       1   7287.2 7353.2
## - HOME_VAL_LOG  1   7287.4 7353.4
## - BLUEBOOK_LOG  1   7296.1 7362.1
## - EDUCATION     3   7302.2 7364.2
## - MSTATUS       1   7304.5 7370.5
## - INCOME_LOG    1   7311.6 7377.6
## - MVR_PTS       1   7325.5 7391.5
## - TIF           1   7326.4 7392.4
## - KIDSDRIV      1   7330.0 7396.0
## - JOB           8   7350.3 7402.3
## - TRAVTIME_LOG  1   7336.7 7402.7
## - REVOKED       1   7346.0 7412.0
## - CAR_USE       1   7350.1 7416.1
## - CAR_TYPE      5   7366.4 7424.4
## - URBANICITY    1   7903.5 7969.5
##
## Step:  AIC=7339.04
## TARGET_FLAG ~ KIDSDRIV + AGE + YOJ + PARENT1 + MSTATUS + EDUCATION +
##     JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY + INCOME_LOG + HOME_VAL_LOG + BLUEBOOK_LOG + OLDCLAIM_LOG +
##     TRAVTIME_LOG
##
##                Df Deviance    AIC
## - AGE           1   7274.8 7338.8
## <none>              7273.0 7339.0
## - OLDCLAIM_LOG  1   7276.2 7340.2
## - CLM_FREQ      1   7277.4 7341.4
## - YOJ           1   7281.6 7345.6
## - PARENT1       1   7288.2 7352.2
## - HOME_VAL_LOG  1   7288.4 7352.4
## - EDUCATION     3   7303.5 7363.5
## - BLUEBOOK_LOG  1   7305.3 7369.3
## - MSTATUS       1   7305.9 7369.9
## - INCOME_LOG    1   7312.9 7376.9
## - MVR_PTS       1   7326.6 7390.6
## - TIF           1   7327.5 7391.5
```

```
## - KIDSDRIV      1   7330.8 7394.8
## - JOB           8   7351.4 7401.4
## - TRAVTIME_LOG  1   7338.1 7402.1
## - REVOKED       1   7347.5 7411.5
## - CAR_USE       1   7351.0 7415.0
## - CAR_TYPE      5   7383.1 7439.1
## - URBANICITY    1   7905.0 7969.0
##
## Step:  AIC=7338.76
## TARGET_FLAG ~ KIDSDRIV + YOJ + PARENT1 + MSTATUS + EDUCATION +
##     JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY + INCOME_LOG + HOME_VAL_LOG + BLUEBOOK_LOG + OLDCLAIM_LOG +
##     TRAVTIME_LOG
##
##                Df Deviance    AIC
## <none>              7274.8 7338.8
## - OLDCLAIM_LOG  1   7278.0 7340.0
## - CLM_FREQ      1   7279.0 7341.0
## - YOJ           1   7282.4 7344.4
## - HOME_VAL_LOG  1   7290.6 7352.6
## - PARENT1       1   7295.3 7357.3
## - EDUCATION     3   7305.6 7363.6
## - MSTATUS       1   7306.6 7368.6
## - BLUEBOOK_LOG  1   7309.0 7371.0
## - INCOME_LOG    1   7313.5 7375.5
## - TIF           1   7329.0 7391.0
## - MVR_PTS       1   7329.2 7391.2
## - KIDSDRIV      1   7332.2 7394.2
## - TRAVTIME_LOG  1   7339.3 7401.3
## - JOB           8   7354.8 7402.8
## - REVOKED       1   7349.6 7411.6
## - CAR_USE       1   7352.3 7414.3
## - CAR_TYPE      5   7383.8 7437.8
## - URBANICITY    1   7908.6 7970.6
```

```
formula(logist_model4)
```

```
## TARGET_FLAG ~ KIDSDRIV + YOJ + PARENT1 + MSTATUS + EDUCATION +
##     JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY + INCOME_LOG + HOME_VAL_LOG + BLUEBOOK_LOG + OLDCLAIM_LOG +
##     TRAVTIME_LOG
```

```
summary(logist_model4)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + YOJ + PARENT1 + MSTATUS +
##     EDUCATION + JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY + INCOME_LOG + HOME_VAL_LOG + BLUEBOOK_LOG +
##     OLDCLAIM_LOG + TRAVTIME_LOG, family = binomial, data = train_df_imp_log)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                      -1.024284   0.607345  -1.686 0.091701 .
## KIDSDRIV                          0.420756   0.055264   7.614 2.67e-14 ***
## YOJ                               0.028332   0.010262   2.761 0.005768 **
## PARENT1Yes                        0.430005   0.094935   4.529 5.91e-06 ***
## MSTATUSYes                       -0.484854   0.085329  -5.682 1.33e-08 ***
## EDUCATIONHigh School              0.437975   0.080283   5.455 4.89e-08 ***
## EDUCATIONMasters                  0.043986   0.134407   0.327 0.743470
## EDUCATIONPhD                      0.025534   0.171478   0.149 0.881630
## JOBBlue Collar                    0.389370   0.184185   2.114 0.034514 *
## JOBClerical                       0.518526   0.193583   2.679 0.007394 **
## JOBDoctor                        -0.398345   0.263824  -1.510 0.131073
## JOBHome Maker                     0.031617   0.218758   0.145 0.885081
## JOBLawyer                         0.165509   0.167695   0.987 0.323660
## JOBManager                       -0.514913   0.169772  -3.033 0.002422 **
## JOBProfessional                   0.218896   0.177198   1.235 0.216711
## JOBStudent                       -0.072148   0.225065  -0.321 0.748540
## CAR_USEPrivate                   -0.764788   0.087658  -8.725  < 2e-16 ***
## TIF                              -0.053147   0.007341  -7.240 4.50e-13 ***
## CAR_TYPEPanel Truck               0.511876   0.142608   3.589 0.000331 ***
## CAR_TYPEPickup                    0.588641   0.099871   5.894 3.77e-09 ***
## CAR_TYPESports Car                0.957686   0.108213   8.850  < 2e-16 ***
## CAR_TYPESUV                       0.741755   0.085908   8.634  < 2e-16 ***
## CAR_TYPEVan                       0.639808   0.121348   5.273 1.35e-07 ***
## CLM_FREQ                          0.089786   0.043487   2.065 0.038955 *
## REVOKEDYes                        0.710936   0.081472   8.726  < 2e-16 ***
## MVR_PTS                           0.103327   0.014058   7.350 1.98e-13 ***
## URBANICITYHighly Urban/ Urban     2.380316   0.113242  21.020  < 2e-16 ***
## INCOME_LOG                       -0.108636   0.017575  -6.181 6.36e-10 ***
## HOME_VAL_LOG                     -0.027417   0.006915  -3.965 7.34e-05 ***
## BLUEBOOK_LOG                     -0.323606   0.055087  -5.874 4.24e-09 ***
## OLDCLAIM_LOG                      0.022491   0.012480   1.802 0.071519 .
## TRAVTIME_LOG                      0.427533   0.054439   7.853 4.05e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9404.0  on 8154  degrees of freedom
## Residual deviance: 7274.8  on 8123  degrees of freedom
## AIC: 7338.8
##
## Number of Fisher Scoring iterations: 5
```

```r
pred_probs <- predict(logist_model4, train_df_imp, type = "response")
pred_class <- ifelse(pred_probs > 0.5, 1, 0)

# Confusion matrix on training data
confusionMatrix(
  factor(pred_class),
  factor(train_df_imp$TARGET_FLAG),
  positive = "1"
)
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction    0    1
##          0 5550 1243
##          1  457  905
##
##                Accuracy : 0.7915
##                  95% CI : (0.7826, 0.8003)
##     No Information Rate : 0.7366
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3912
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4213
##             Specificity : 0.9239
##          Pos Pred Value : 0.6645
##          Neg Pred Value : 0.8170
##              Prevalence : 0.2634
##          Detection Rate : 0.1110
##    Detection Prevalence : 0.1670
##       Balanced Accuracy : 0.6726
##
##        'Positive' Class : 1
##
```

Linear Model 1: All Predictors Without Transformed

To make predictions on claim amount, this model would be the baseline.

```r
# Removing TARGET_FLAG from training and testing data for linear modeling of TARGET_AMT

train_df_imp_lin <- subset(train_df_imp, select = -TARGET_FLAG)
test_df_lin <- subset(test_df, select = -TARGET_FLAG)
```

```r
linear_model1 <- lm(TARGET_AMT ~ .- INCOME_LOG - HOME_VAL_LOG - BLUEBOOK_LOG - OLDCLAIM_LOG - TRAVTIME_L

summary(linear_model1)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - INCOME_LOG - HOME_VAL_LOG - BLUEBOOK_LOG -
##     OLDCLAIM_LOG - TRAVTIME_LOG, data = train_df_imp_lin)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5817  -1701   -767    351 103812
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -6.082e+02  5.640e+02  -1.078 0.280881
## KIDSDRIV                         3.160e+02  1.133e+02   2.790 0.005282 **
```

```
## AGE                              4.966e+00  7.089e+00   0.700 0.483637    
## HOMEKIDS                         7.772e+01  6.553e+01   1.186 0.235681    
## YOJ                             -3.455e+00  1.489e+01  -0.232 0.816450    
## INCOME                          -5.407e-03  1.897e-03  -2.851 0.004369 ** 
## PARENT1Yes                       5.719e+02  2.023e+02   2.827 0.004706 ** 
## HOME_VAL                        -5.628e-04  6.066e-04  -0.928 0.353516    
## MSTATUSYes                      -5.701e+02  1.481e+02  -3.849 0.000120 ***
## SEXM                             3.719e+02  1.840e+02   2.021 0.043281 *  
## EDUCATIONHigh School             1.779e+02  1.549e+02   1.148 0.250877    
## EDUCATIONMasters                 2.901e+02  2.242e+02   1.294 0.195783    
## EDUCATIONPhD                     6.117e+02  2.915e+02   2.099 0.035878 *  
## JOBBlue Collar                   4.971e+02  3.212e+02   1.547 0.121782    
## JOBClerical                      4.859e+02  3.421e+02   1.420 0.155591    
## JOBDoctor                       -5.329e+02  4.083e+02  -1.305 0.191808    
## JOBHome Maker                    2.523e+02  3.669e+02   0.688 0.491632    
## JOBLawyer                        2.049e+02  2.949e+02   0.695 0.487205    
## JOBManager                      -5.001e+02  2.880e+02  -1.736 0.082559 .  
## JOBProfessional                  4.309e+02  3.084e+02   1.397 0.162436    
## JOBStudent                       2.106e+02  3.761e+02   0.560 0.575560    
## TRAVTIME                         1.197e+01  3.223e+00   3.713 0.000207 ***
## CAR_USEPrivate                  -7.603e+02  1.578e+02  -4.819 1.47e-06 ***
## BLUEBOOK                         1.561e-02  8.647e-03   1.806 0.071014 .  
## TIF                             -4.825e+01  1.219e+01  -3.959 7.59e-05 ***
## CAR_TYPEPanel Truck              2.829e+02  2.761e+02   1.025 0.305495    
## CAR_TYPEPickup                   3.864e+02  1.697e+02   2.278 0.022763 *  
## CAR_TYPESports Car               1.020e+03  2.179e+02   4.681 2.90e-06 ***
## CAR_TYPESUV                      7.502e+02  1.794e+02   4.182 2.92e-05 ***
## CAR_TYPEVan                      5.233e+02  2.125e+02   2.462 0.013818 *  
## RED_CARyes                      -5.254e+01  1.495e+02  -0.351 0.725226    
## OLDCLAIM                        -1.060e-02  7.438e-03  -1.424 0.154355    
## CLM_FREQ                         1.394e+02  5.510e+01   2.530 0.011426 *  
## REVOKEDYes                       5.493e+02  1.736e+02   3.165 0.001557 ** 
## MVR_PTS                          1.748e+02  2.594e+01   6.739 1.70e-11 ***
## CAR_AGE                         -2.389e+01  1.273e+01  -1.877 0.060549 .  
## URBANICITYHighly Urban/ Urban    1.662e+03  1.395e+02  11.919  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4545 on 8118 degrees of freedom
## Multiple R-squared:  0.07117,    Adjusted R-squared:  0.06706 
## F-statistic: 17.28 on 36 and 8118 DF,  p-value: < 2.2e-16
```

Linear Model 2: Reduced Predictors Without Transformed

```r
linear_model2<- step(linear_model1, direction = "backward")
```

```
## Start:  AIC=137394.9
## TARGET_AMT ~ (KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 + 
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + 
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ + 
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG + 
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG) - INCOME_LOG - 
```

```
##     HOME_VAL_LOG - BLUEBOOK_LOG - OLDCLAIM_LOG - TRAVTIME_LOG
##
##              Df  Sum of Sq        RSS    AIC
## - YOJ         1    1112829 1.6767e+11 137393
## - RED_CAR     1    2551742 1.6767e+11 137393
## - AGE         1   10134704 1.6768e+11 137393
## - HOME_VAL    1   17780701 1.6768e+11 137394
## - HOMEKIDS    1   29048453 1.6770e+11 137394
## - EDUCATION   3  116858515 1.6778e+11 137395
## <none>                     1.6767e+11 137395
## - OLDCLAIM    1   41907585 1.6771e+11 137395
## - BLUEBOOK    1   67337122 1.6773e+11 137396
## - CAR_AGE     1   72768597 1.6774e+11 137396
## - SEX         1   84384775 1.6775e+11 137397
## - CLM_FREQ    1  132199902 1.6780e+11 137399
## - KIDSDRIV    1  160778001 1.6783e+11 137401
## - PARENT1     1  165098600 1.6783e+11 137401
## - INCOME      1  167882584 1.6783e+11 137401
## - REVOKED     1  206871090 1.6787e+11 137403
## - TRAVTIME    1  284669259 1.6795e+11 137407
## - MSTATUS     1  305920949 1.6797e+11 137408
## - TIF         1  323724741 1.6799e+11 137409
## - JOB         8  635092840 1.6830e+11 137410
## - CAR_TYPE    5  607765851 1.6827e+11 137414
## - CAR_USE     1  479609362 1.6815e+11 137416
## - MVR_PTS     1  937963469 1.6861e+11 137438
## - URBANICITY  1 2934277503 1.7060e+11 137534
##
## Step:  AIC=137393
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - RED_CAR     1    2583649 1.6767e+11 137391
## - AGE         1    9246605 1.6768e+11 137391
## - HOME_VAL    1   17732411 1.6769e+11 137392
## - HOMEKIDS    1   27939582 1.6770e+11 137392
## - EDUCATION   3  117385626 1.6779e+11 137393
## <none>                     1.6767e+11 137393
## - OLDCLAIM    1   42175316 1.6771e+11 137393
## - BLUEBOOK    1   67197994 1.6774e+11 137394
## - CAR_AGE     1   72668804 1.6774e+11 137394
## - SEX         1   84491243 1.6775e+11 137395
## - CLM_FREQ    1  132504760 1.6780e+11 137397
## - KIDSDRIV    1  162755347 1.6783e+11 137399
## - PARENT1     1  165441550 1.6783e+11 137399
## - INCOME      1  171214430 1.6784e+11 137399
## - REVOKED     1  207222948 1.6788e+11 137401
## - TRAVTIME    1  284272996 1.6795e+11 137405
## - MSTATUS     1  312435644 1.6798e+11 137406
## - TIF         1  324212833 1.6799e+11 137407
## - JOB         8  635159796 1.6830e+11 137408
```

```
## - CAR_TYPE    5  609771449 1.6828e+11 137413
## - CAR_USE     1  481219501 1.6815e+11 137414
## - MVR_PTS     1  939631112 1.6861e+11 137437
## - URBANICITY  1 2933248642 1.7060e+11 137532
##
## Step:  AIC=137391.1
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - AGE         1    9432621 1.6768e+11 137390
## - HOME_VAL    1   17243154 1.6769e+11 137390
## - HOMEKIDS    1   27817519 1.6770e+11 137390
## - EDUCATION   3  117858154 1.6779e+11 137391
## <none>                     1.6767e+11 137391
## - OLDCLAIM    1   42307629 1.6771e+11 137391
## - BLUEBOOK    1   67967365 1.6774e+11 137392
## - CAR_AGE     1   73040376 1.6774e+11 137393
## - SEX         1   91511213 1.6776e+11 137394
## - CLM_FREQ    1  132172012 1.6780e+11 137396
## - KIDSDRIV    1  163544918 1.6783e+11 137397
## - PARENT1     1  165875106 1.6784e+11 137397
## - INCOME      1  171800351 1.6784e+11 137397
## - REVOKED     1  207303164 1.6788e+11 137399
## - TRAVTIME    1  283655873 1.6795e+11 137403
## - MSTATUS     1  312660009 1.6798e+11 137404
## - TIF         1  323843383 1.6799e+11 137405
## - JOB         8  636677616 1.6831e+11 137406
## - CAR_TYPE    5  611689334 1.6828e+11 137411
## - CAR_USE     1  480965539 1.6815e+11 137412
## - MVR_PTS     1  939514216 1.6861e+11 137435
## - URBANICITY  1 2931769825 1.7060e+11 137530
##
## Step:  AIC=137389.5
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - HOME_VAL    1   15430964 1.6770e+11 137388
## - HOMEKIDS    1   20091357 1.6770e+11 137389
## - EDUCATION   3  121203600 1.6780e+11 137389
## <none>                     1.6768e+11 137390
## - OLDCLAIM    1   42187435 1.6772e+11 137390
## - CAR_AGE     1   72799224 1.6775e+11 137391
## - BLUEBOOK    1   76871033 1.6776e+11 137391
## - SEX         1   99265586 1.6778e+11 137392
## - CLM_FREQ    1  133388848 1.6781e+11 137394
## - PARENT1     1  159916441 1.6784e+11 137395
## - INCOME      1  176719999 1.6786e+11 137396
## - KIDSDRIV    1  181442153 1.6786e+11 137396
```

```
## - REVOKED     1  206197630 1.6789e+11 137398
## - TRAVTIME    1  284883156 1.6797e+11 137401
## - MSTATUS     1  311954623 1.6799e+11 137403
## - TIF         1  323108458 1.6800e+11 137403
## - JOB         8  631398802 1.6831e+11 137404
## - CAR_TYPE    5  629558850 1.6831e+11 137410
## - CAR_USE     1  482469176 1.6816e+11 137411
## - MVR_PTS     1  933474247 1.6861e+11 137433
## - URBANICITY  1 2926209016 1.7061e+11 137529
##
## Step:  AIC=137388.3
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + MSTATUS +
##     SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF +
##     CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - HOMEKIDS    1   22289847 1.6772e+11 137387
## <none>                     1.6770e+11 137388
## - EDUCATION   3  123896326 1.6782e+11 137388
## - OLDCLAIM    1   42025960 1.6774e+11 137388
## - CAR_AGE     1   70509464 1.6777e+11 137390
## - BLUEBOOK    1   76665722 1.6777e+11 137390
## - SEX         1   97804438 1.6779e+11 137391
## - CLM_FREQ    1  136571590 1.6783e+11 137393
## - PARENT1     1  154935616 1.6785e+11 137394
## - KIDSDRIV    1  181200540 1.6788e+11 137395
## - REVOKED     1  208207491 1.6790e+11 137396
## - TRAVTIME    1  286690985 1.6798e+11 137400
## - INCOME      1  296728586 1.6799e+11 137401
## - TIF         1  323387309 1.6802e+11 137402
## - JOB         8  623642849 1.6832e+11 137403
## - CAR_TYPE    5  626421056 1.6832e+11 137409
## - CAR_USE     1  482234372 1.6818e+11 137410
## - MSTATUS     1  548908441 1.6824e+11 137413
## - MVR_PTS     1  936347461 1.6863e+11 137432
## - URBANICITY  1 2925421152 1.7062e+11 137527
##
## Step:  AIC=137387.4
## TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS + SEX + EDUCATION +
##     JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - EDUCATION   3  122723668 1.6784e+11 137387
## <none>                     1.6772e+11 137387
## - OLDCLAIM    1   41734993 1.6776e+11 137387
## - CAR_AGE     1   71509300 1.6779e+11 137389
## - BLUEBOOK    1   73274972 1.6779e+11 137389
## - SEX         1   92981999 1.6781e+11 137390
## - CLM_FREQ    1  136657162 1.6785e+11 137392
## - REVOKED     1  210218593 1.6793e+11 137396
## - PARENT1     1  278222687 1.6800e+11 137399
## - KIDSDRIV    1  281565317 1.6800e+11 137399
```

```
## - TRAVTIME    1  284231927 1.6800e+11 137399
## - INCOME      1  293171568 1.6801e+11 137400
## - TIF         1  320101852 1.6804e+11 137401
## - JOB         8  631327071 1.6835e+11 137402
## - CAR_TYPE    5  625999868 1.6834e+11 137408
## - CAR_USE     1  479341033 1.6820e+11 137409
## - MSTATUS     1  530362886 1.6825e+11 137411
## - MVR_PTS     1  940827635 1.6866e+11 137431
## - URBANICITY  1 2926016969 1.7064e+11 137526
##
## Step:  AIC=137387.3
## TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS + SEX + JOB +
##     TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## <none>                     1.6784e+11 137387
## - OLDCLAIM    1   41654117 1.6788e+11 137387
## - BLUEBOOK    1   71628045 1.6791e+11 137389
## - SEX         1   87070026 1.6793e+11 137390
## - CAR_AGE     1   90325552 1.6793e+11 137390
## - CLM_FREQ    1  137391230 1.6798e+11 137392
## - REVOKED     1  212478648 1.6805e+11 137396
## - INCOME      1  270824651 1.6811e+11 137398
## - PARENT1     1  271187980 1.6811e+11 137399
## - TRAVTIME    1  276704526 1.6812e+11 137399
## - KIDSDRIV    1  282924792 1.6812e+11 137399
## - TIF         1  314920711 1.6816e+11 137401
## - JOB         8  621015876 1.6846e+11 137401
## - CAR_USE     1  454348561 1.6829e+11 137407
## - CAR_TYPE    5  629496006 1.6847e+11 137408
## - MSTATUS     1  528346071 1.6837e+11 137411
## - MVR_PTS     1  944263934 1.6878e+11 137431
## - URBANICITY  1 2920994033 1.7076e+11 137526
```

```r
summary(linear_model2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS +
##     SEX + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY,
##     data = train_df_imp_lin)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5804  -1693   -763    333 103735
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -8.933e+01  4.459e+02  -0.200 0.841209
## KIDSDRIV                       3.780e+02  1.021e+02   3.701 0.000216 ***
## INCOME                        -5.698e-03  1.573e-03  -3.621 0.000295 ***
## PARENT1Yes                     6.411e+02  1.769e+02   3.623 0.000292 ***
```

```
## MSTATUSYes                      -6.038e+02  1.194e+02  -5.058 4.34e-07 ***
## SEXM                             3.303e+02  1.609e+02   2.053 0.040088 *
## JOBBlue Collar                   2.505e+02  2.710e+02   0.924 0.355299
## JOBClerical                      2.751e+02  2.982e+02   0.923 0.356253
## JOBDoctor                       -3.415e+02  3.768e+02  -0.906 0.364787
## JOBHome Maker                    1.016e+02  3.392e+02   0.299 0.764650
## JOBLawyer                        1.147e+02  2.876e+02   0.399 0.690025
## JOBManager                      -7.109e+02  2.683e+02  -2.649 0.008081 **
## JOBProfessional                  1.166e+02  2.688e+02   0.434 0.664480
## JOBStudent                       7.334e+01  3.312e+02   0.221 0.824733
## TRAVTIME                         1.179e+01  3.221e+00   3.660 0.000254 ***
## CAR_USEPrivate                  -7.362e+02  1.570e+02  -4.690 2.77e-06 ***
## BLUEBOOK                         1.592e-02  8.547e-03   1.862 0.062608 .
## TIF                             -4.755e+01  1.218e+01  -3.905 9.51e-05 ***
## CAR_TYPEPanel Truck              3.049e+02  2.750e+02   1.108 0.267703
## CAR_TYPEPickup                   4.033e+02  1.694e+02   2.381 0.017305 *
## CAR_TYPESports Car               1.034e+03  2.164e+02   4.778 1.80e-06 ***
## CAR_TYPESUV                      7.565e+02  1.785e+02   4.239 2.27e-05 ***
## CAR_TYPEVan                      5.318e+02  2.122e+02   2.506 0.012239 *
## OLDCLAIM                        -1.056e-02  7.435e-03  -1.420 0.155617
## CLM_FREQ                         1.420e+02  5.506e+01   2.579 0.009923 **
## REVOKEDYes                       5.564e+02  1.735e+02   3.207 0.001345 **
## MVR_PTS                          1.752e+02  2.590e+01   6.761 1.46e-11 ***
## CAR_AGE                         -2.311e+01  1.105e+01  -2.091 0.036541 *
## URBANICITYHighly Urban/ Urban  1.658e+03  1.394e+02  11.892  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4545 on 8126 degrees of freedom
## Multiple R-squared:  0.07021,    Adjusted R-squared:  0.06701
## F-statistic: 21.92 on 28 and 8126 DF,  p-value: < 2.2e-16
```

```r
formula(linear_model2)
```

```
## TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS + SEX + JOB +
##     TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY
```

Linear Model 3: All Predictors With Transformed

For comparsion with models that have non tranformed variables

```r
linear_model3 <- lm(TARGET_AMT ~ .- INCOME - HOME_VAL - BLUEBOOK - OLDCLAIM - TRAVTIME, data = train_df_

summary(linear_model3)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - INCOME - HOME_VAL - BLUEBOOK -
##     OLDCLAIM - TRAVTIME, data = train_df_imp_lin)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##  -5489  -1687   -769    354 104037
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2295.993   1165.409  -1.970 0.048859 *
## KIDSDRIV                        327.894    113.370   2.892 0.003835 **
## AGE                               3.019      7.112   0.425 0.671193
## HOMEKIDS                         53.903     66.165   0.815 0.415275
## YOJ                              22.774     18.293   1.245 0.213184
## PARENT1Yes                      578.177    201.936   2.863 0.004205 **
## MSTATUSYes                     -553.242    151.997  -3.640 0.000275 ***
## SEXM                            376.046    179.431   2.096 0.036134 *
## EDUCATIONHigh School            237.207    153.304   1.547 0.121831
## EDUCATIONMasters                238.780    224.016   1.066 0.286498
## EDUCATIONPhD                    375.478    281.682   1.333 0.182574
## JOBBlue Collar                  595.661    319.693   1.863 0.062466 .
## JOBClerical                     660.500    336.428   1.963 0.049649 *
## JOBDoctor                      -508.220    408.302  -1.245 0.213272
## JOBHome Maker                   291.881    372.939   0.783 0.433855
## JOBLawyer                       267.343    294.391   0.908 0.363841
## JOBManager                     -449.145    287.699  -1.561 0.118524
## JOBProfessional                 502.263    307.661   1.633 0.102608
## JOBStudent                      132.249    393.723   0.336 0.736959
## CAR_USEPrivate                 -756.953    157.833  -4.796 1.65e-06 ***
## TIF                             -47.111     12.190  -3.865 0.000112 ***
## CAR_TYPEPanel Truck             297.467    260.603   1.141 0.253713
## CAR_TYPEPickup                  402.999    169.365   2.379 0.017361 *
## CAR_TYPESports Car             1020.824    216.043   4.725 2.34e-06 ***
## CAR_TYPESUV                     751.632    174.245   4.314 1.62e-05 ***
## CAR_TYPEVan                     503.726    211.540   2.381 0.017278 *
## RED_CARyes                      -58.047    149.476  -0.388 0.697775
## CLM_FREQ                         67.217     85.752   0.784 0.433144
## REVOKEDYes                      424.791    156.756   2.710 0.006745 **
## MVR_PTS                         169.384     26.678   6.349 2.28e-10 ***
## CAR_AGE                         -23.956     12.726  -1.882 0.059808 .
## URBANICITYHighly Urban/ Urban  1636.301    139.949  11.692  < 2e-16 ***
## INCOME_LOG                      -89.452     30.947  -2.890 0.003857 **
## HOME_VAL_LOG                    -16.205     12.609  -1.285 0.198768
## BLUEBOOK_LOG                    163.858    103.035   1.590 0.111806
## OLDCLAIM_LOG                     13.723     24.147   0.568 0.569837
## TRAVTIME_LOG                    335.287     89.168   3.760 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4546 on 8118 degrees of freedom
## Multiple R-squared:  0.07061,    Adjusted R-squared:  0.06649
## F-statistic: 17.13 on 36 and 8118 DF,  p-value: < 2.2e-16
```

Linear Model 4: Reduced Predictors With Transformed

```r
linear_model4 <- step(linear_model3, direction = "backward")
```

```
## Start:  AIC=137399.8
## TARGET_AMT ~ (KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG) - INCOME - HOME_VAL -
##     BLUEBOOK - OLDCLAIM - TRAVTIME
##
##                Df  Sum of Sq        RSS    AIC
## - EDUCATION     3   83764537 1.6785e+11 137398
## - RED_CAR       1    3116619 1.6777e+11 137398
## - AGE           1    3724528 1.6777e+11 137398
## - OLDCLAIM_LOG  1    6674831 1.6778e+11 137398
## - CLM_FREQ      1   12698169 1.6778e+11 137398
## - HOMEKIDS      1   13716509 1.6778e+11 137399
## - YOJ           1   32030904 1.6780e+11 137399
## - HOME_VAL_LOG  1   34133834 1.6780e+11 137399
## <none>                       1.6777e+11 137400
## - BLUEBOOK_LOG  1   52266252 1.6782e+11 137400
## - CAR_AGE       1   73234748 1.6784e+11 137401
## - SEX           1   90771157 1.6786e+11 137402
## - REVOKED       1  151762251 1.6792e+11 137405
## - PARENT1       1  169416146 1.6794e+11 137406
## - INCOME_LOG    1  172663430 1.6794e+11 137406
## - KIDSDRIV      1  172875568 1.6794e+11 137406
## - MSTATUS       1  273793538 1.6804e+11 137411
## - TRAVTIME_LOG  1  292195343 1.6806e+11 137412
## - TIF           1  308685747 1.6808e+11 137413
## - CAR_TYPE      5  609247307 1.6838e+11 137419
## - JOB           8  748968743 1.6852e+11 137420
## - CAR_USE       1  475337908 1.6824e+11 137421
## - MVR_PTS       1  833078192 1.6860e+11 137438
## - URBANICITY    1 2825204941 1.7059e+11 137534
##
## Step:  AIC=137397.9
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + MSTATUS +
##     SEX + JOB + CAR_USE + TIF + CAR_TYPE + RED_CAR + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG
##
##                Df  Sum of Sq        RSS    AIC
## - RED_CAR       1    3598045 1.6786e+11 137396
## - AGE           1    4654470 1.6786e+11 137396
## - OLDCLAIM_LOG  1    6718740 1.6786e+11 137396
## - CLM_FREQ      1   12585069 1.6786e+11 137397
## - HOMEKIDS      1   13877963 1.6787e+11 137397
## - YOJ           1   33923562 1.6789e+11 137398
## - HOME_VAL_LOG  1   38252754 1.6789e+11 137398
## <none>                       1.6785e+11 137398
## - BLUEBOOK_LOG  1   49228358 1.6790e+11 137398
## - SEX           1   86572738 1.6794e+11 137400
```

```
## - CAR_AGE       1  121470587 1.6797e+11 137402
## - REVOKED       1  153844373 1.6801e+11 137403
## - PARENT1       1  166003727 1.6802e+11 137404
## - KIDSDRIV      1  173689080 1.6803e+11 137404
## - INCOME_LOG    1  183979035 1.6804e+11 137405
## - MSTATUS       1  266455930 1.6812e+11 137409
## - TRAVTIME_LOG  1  284014359 1.6814e+11 137410
## - TIF           1  303762559 1.6816e+11 137411
## - CAR_TYPE      5  609946016 1.6846e+11 137417
## - CAR_USE       1  447327201 1.6830e+11 137418
## - JOB           8  847504748 1.6870e+11 137423
## - MVR_PTS       1  833642710 1.6869e+11 137436
## - URBANICITY    1 2819644634 1.7067e+11 137532
##
## Step:  AIC=137396.1
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + MSTATUS +
##     SEX + JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG
##
##                Df  Sum of Sq        RSS    AIC
## - AGE           1    4867735 1.6786e+11 137394
## - OLDCLAIM_LOG  1    6813624 1.6786e+11 137394
## - CLM_FREQ      1   12371730 1.6787e+11 137395
## - HOMEKIDS      1   13826346 1.6787e+11 137395
## - YOJ           1   33554295 1.6789e+11 137396
## - HOME_VAL_LOG  1   37636262 1.6789e+11 137396
## <none>                       1.6786e+11 137396
## - BLUEBOOK_LOG  1   49866193 1.6791e+11 137397
## - SEX           1   92176080 1.6795e+11 137399
## - CAR_AGE       1  122188318 1.6798e+11 137400
## - REVOKED       1  153732425 1.6801e+11 137402
## - PARENT1       1  166621510 1.6802e+11 137402
## - KIDSDRIV      1  174572217 1.6803e+11 137403
## - INCOME_LOG    1  183240986 1.6804e+11 137403
## - MSTATUS       1  266262583 1.6812e+11 137407
## - TRAVTIME_LOG  1  283166674 1.6814e+11 137408
## - TIF           1  303349198 1.6816e+11 137409
## - CAR_TYPE      5  611947046 1.6847e+11 137416
## - CAR_USE       1  446914288 1.6830e+11 137416
## - JOB           8  849536593 1.6871e+11 137421
## - MVR_PTS       1  833232415 1.6869e+11 137434
## - URBANICITY    1 2817725753 1.7067e+11 137530
##
## Step:  AIC=137394.3
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 + MSTATUS +
##     SEX + JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + OLDCLAIM_LOG + TRAVTIME_LOG
##
##                Df  Sum of Sq        RSS    AIC
## - OLDCLAIM_LOG  1    6820600 1.6787e+11 137393
## - HOMEKIDS      1    9639346 1.6787e+11 137393
## - CLM_FREQ      1   12499280 1.6787e+11 137393
```

```
## - HOME_VAL_LOG  1   36939207 1.6790e+11 137394
## <none>                       1.6786e+11 137394
## - YOJ           1   41416724 1.6790e+11 137394
## - BLUEBOOK_LOG  1   54966017 1.6792e+11 137395
## - SEX           1   96908236 1.6796e+11 137397
## - CAR_AGE       1  121058293 1.6798e+11 137398
## - REVOKED       1  152980168 1.6801e+11 137400
## - PARENT1       1  162936018 1.6802e+11 137400
## - KIDSDRIV      1  190539013 1.6805e+11 137402
## - INCOME_LOG    1  193861072 1.6805e+11 137402
## - MSTATUS       1  265962239 1.6813e+11 137405
## - TRAVTIME_LOG  1  283322006 1.6814e+11 137406
## - TIF           1  303072148 1.6816e+11 137407
## - CAR_USE       1  448751128 1.6831e+11 137414
## - CAR_TYPE      5  625784271 1.6849e+11 137415
## - JOB           8  844835231 1.6871e+11 137419
## - MVR_PTS       1  829618807 1.6869e+11 137433
## - URBANICITY    1 2813937018 1.7067e+11 137528
##
## Step:  AIC=137392.6
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 + MSTATUS +
##     SEX + JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG +
##     BLUEBOOK_LOG + TRAVTIME_LOG
##
##                Df  Sum of Sq        RSS    AIC
## - HOMEKIDS      1    9915206 1.6788e+11 137391
## - HOME_VAL_LOG  1   37375762 1.6791e+11 137392
## <none>                       1.6787e+11 137393
## - YOJ           1   41903117 1.6791e+11 137393
## - BLUEBOOK_LOG  1   55273914 1.6792e+11 137393
## - SEX           1   97192211 1.6796e+11 137395
## - CLM_FREQ      1   99266117 1.6797e+11 137395
## - CAR_AGE       1  121691454 1.6799e+11 137397
## - PARENT1       1  162483609 1.6803e+11 137399
## - REVOKED       1  165851175 1.6803e+11 137399
## - KIDSDRIV      1  190523133 1.6806e+11 137400
## - INCOME_LOG    1  193767962 1.6806e+11 137400
## - MSTATUS       1  266847149 1.6813e+11 137404
## - TRAVTIME_LOG  1  282849713 1.6815e+11 137404
## - TIF           1  304191130 1.6817e+11 137405
## - CAR_USE       1  450314331 1.6832e+11 137412
## - CAR_TYPE      5  630032122 1.6850e+11 137413
## - JOB           8  848903491 1.6872e+11 137418
## - MVR_PTS       1  923088612 1.6879e+11 137435
## - URBANICITY    1 2885568697 1.7075e+11 137530
##
## Step:  AIC=137391.1
## TARGET_AMT ~ KIDSDRIV + YOJ + PARENT1 + MSTATUS + SEX + JOB +
##     CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY + INCOME_LOG + HOME_VAL_LOG + BLUEBOOK_LOG +
##     TRAVTIME_LOG
##
##                Df  Sum of Sq        RSS    AIC
```

```
## - HOME_VAL_LOG  1   38160904 1.6792e+11 137391
## <none>                       1.6788e+11 137391
## - YOJ           1   51182838 1.6793e+11 137392
## - BLUEBOOK_LOG  1   53591657 1.6793e+11 137392
## - SEX           1   94029600 1.6797e+11 137394
## - CLM_FREQ      1   99289987 1.6798e+11 137394
## - CAR_AGE       1  123231789 1.6800e+11 137395
## - REVOKED       1  167278023 1.6804e+11 137397
## - INCOME_LOG    1  204064662 1.6808e+11 137399
## - MSTATUS       1  257057257 1.6813e+11 137402
## - PARENT1       1  259900760 1.6814e+11 137402
## - KIDSDRIV      1  270503381 1.6815e+11 137402
## - TRAVTIME_LOG  1  280611051 1.6816e+11 137403
## - TIF           1  302329138 1.6818e+11 137404
## - CAR_USE       1  449130706 1.6833e+11 137411
## - CAR_TYPE      5  630020938 1.6851e+11 137412
## - JOB           8  860804920 1.6874e+11 137417
## - MVR_PTS       1  926462842 1.6880e+11 137434
## - URBANICITY    1 2885886068 1.7076e+11 137528
##
## Step:  AIC=137391
## TARGET_AMT ~ KIDSDRIV + YOJ + PARENT1 + MSTATUS + SEX + JOB +
##     CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY + INCOME_LOG + BLUEBOOK_LOG + TRAVTIME_LOG
##
##                Df  Sum of Sq        RSS    AIC
## <none>                       1.6792e+11 137391
## - YOJ           1   51807582 1.6797e+11 137392
## - BLUEBOOK_LOG  1   52809358 1.6797e+11 137392
## - SEX           1   91596369 1.6801e+11 137393
## - CLM_FREQ      1  103056765 1.6802e+11 137394
## - CAR_AGE       1  122757775 1.6804e+11 137395
## - REVOKED       1  170913939 1.6809e+11 137397
## - INCOME_LOG    1  213579483 1.6813e+11 137399
## - PARENT1       1  259169748 1.6817e+11 137402
## - KIDSDRIV      1  273899405 1.6819e+11 137402
## - TRAVTIME_LOG  1  284181817 1.6820e+11 137403
## - TIF           1  303740963 1.6822e+11 137404
## - CAR_USE       1  446353873 1.6836e+11 137411
## - CAR_TYPE      5  628972873 1.6854e+11 137411
## - JOB           8  841952714 1.6876e+11 137416
## - MSTATUS       1  564089417 1.6848e+11 137416
## - MVR_PTS       1  935911078 1.6885e+11 137434
## - URBANICITY    1 2880984646 1.7080e+11 137528
```

```r
summary(linear_model4)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + YOJ + PARENT1 + MSTATUS +
##     SEX + JOB + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY + INCOME_LOG + BLUEBOOK_LOG +
##     TRAVTIME_LOG, data = train_df_imp_lin)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -5482  -1681   -766    343 103878 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     -1865.26    1128.94  -1.652 0.098527 .  
## KIDSDRIV                          372.21     102.23   3.641 0.000274 ***
## YOJ                                27.76      17.53   1.583 0.113370    
## PARENT1Yes                        628.53     177.48   3.541 0.000400 ***
## MSTATUSYes                       -637.96     122.10  -5.225 1.79e-07 ***
## SEXM                              327.41     155.51   2.105 0.035288 *  
## JOBBlue Collar                    452.19     262.54   1.722 0.085037 .  
## JOBClerical                       543.91     282.07   1.928 0.053853 .  
## JOBDoctor                        -436.83     375.63  -1.163 0.244894    
## JOBHome Maker                     149.56     347.45   0.430 0.666872    
## JOBLawyer                         213.74     285.85   0.748 0.454641    
## JOBManager                       -608.01     266.42  -2.282 0.022506 *  
## JOBProfessional                   268.52     264.74   1.014 0.310476    
## JOBStudent                        115.69     339.91   0.340 0.733599    
## CAR_USEPrivate                   -729.72     157.01  -4.648 3.41e-06 ***
## TIF                               -46.70      12.18  -3.834 0.000127 ***
## CAR_TYPEPanel Truck               328.88     259.89   1.265 0.205743    
## CAR_TYPEPickup                    417.35     169.09   2.468 0.013599 *  
## CAR_TYPESports Car               1028.57     214.64   4.792 1.68e-06 ***
## CAR_TYPESUV                       755.38     173.46   4.355 1.35e-05 ***
## CAR_TYPEVan                       519.52     211.26   2.459 0.013950 *  
## CLM_FREQ                          109.11      48.86   2.233 0.025562 *  
## REVOKEDYes                        445.86     155.03   2.876 0.004039 ** 
## MVR_PTS                           173.79      25.82   6.730 1.81e-11 ***
## CAR_AGE                           -26.74      10.97  -2.437 0.014817 *  
## URBANICITYHighly Urban/ Urban    1639.98     138.89  11.808  < 2e-16 ***
## INCOME_LOG                        -97.73      30.40  -3.215 0.001310 ** 
## BLUEBOOK_LOG                      162.97     101.94   1.599 0.109942    
## TRAVTIME_LOG                      330.33      89.07   3.708 0.000210 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4546 on 8126 degrees of freedom
## Multiple R-squared:  0.0698, Adjusted R-squared:  0.06659 
## F-statistic: 21.78 on 28 and 8126 DF,  p-value: < 2.2e-16
```

```r
formula(linear_model4)
```

```
## TARGET_AMT ~ KIDSDRIV + YOJ + PARENT1 + MSTATUS + SEX + JOB + 
##     CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + 
##     CAR_AGE + URBANICITY + INCOME_LOG + BLUEBOOK_LOG + TRAVTIME_LOG
```

4. SELECT MODELS

Decide on the criteria for selecting the best multiple linear regression model and the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models. For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. For the binary logistic regression model, will you use a metric such as log likelihood, AIC, ROC curve, etc.? Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set.

Assessing Logistic Models

Our logistic regression model 1 the model with all variables (except target amt) with log transformation. This model was the best choice, as the accuracy difference amog the other models were minimal, it had the highest accuracy, the highest sensitivity, meaning it was slightly better at correctly identifying positive cases. This is important in situations where detecting positive is more critical than minimizing false positives. Although its AIC (7345.540) is slightly higher than some reduced models, the trade-off in favor of sensitivity justifies choosing Full Log in this context. Therefore, Full Log provides a good balance between performance and the practical goal of maximizing true positive detection.

```r
compare_logistic_models <- function(models, data, response_var, threshold = 0.5) {

  results <- data.frame(
    Model = character(),
    AIC = numeric(),
    Accuracy = numeric(),
    Sensitivity = numeric(),
    Specificity = numeric(),
    stringsAsFactors = FALSE
  )

  for (i in seq_along(models)) {
    model <- models[[i]]
    model_name <- names(models)[i]

    # Predicted probabilities
    probs <- predict(model, newdata = data, type = "response")
    preds <- factor(ifelse(probs > threshold, 1, 0), levels = c(0, 1))


    # Actual values
    actual <- factor(data[[response_var]], levels = c(0, 1))


    # Confusion matrix
    cm <- confusionMatrix(as.factor(preds), as.factor(actual), positive = "1")

    # Append results
    results <- rbind(results, data.frame(
      Model = model_name,
      AIC = AIC(model),
      Accuracy = cm$overall["Accuracy"],
      Sensitivity = cm$byClass["Sensitivity"],
      Specificity = cm$byClass["Specificity"]
    ))
```

```r
  }

  return(results)
}


models <- list(
  Full_Log = logist_model3,
  Full_NoLog = logist_model1,
  Reduced_Log = logist_model4,
  Reduced_NoLog = logist_model2
)


compare_logistic_models(models, train_df_imp_log, response_var = "TARGET_FLAG")
```

```
##                     Model      AIC  Accuracy Sensitivity Specificity
## Accuracy       Full_Log 7345.540 0.7919068   0.4241155   0.9234227
## Accuracy1    Full_NoLog 7363.311 0.7917842   0.4245810   0.9230897
## Accuracy2   Reduced_Log 7338.760 0.7915389   0.4213222   0.9239221
## Accuracy3 Reduced_NoLog 7354.455 0.7914163   0.4199255   0.9242550
```

Assessing Linear Models

Our linear model 2, the reduced model without transformed data performed the best with AIC and only very slightly worse on Rsquared.

```r
model_comparison <- data.frame(
  Model = c("linear_model1", "linear_model2", "linear_model3", "linear_model4"),
  AIC = c(
    AIC(linear_model1),
    AIC(linear_model2),
    AIC(linear_model3),
    AIC(linear_model4)
  ),
  Adjusted_R2 = c(
    summary(linear_model1)$adj.r.squared,
    summary(linear_model2)$adj.r.squared,
    summary(linear_model3)$adj.r.squared,
    summary(linear_model4)$adj.r.squared
  )
)

print(model_comparison)
```

```
##           Model      AIC Adjusted_R2
## 1 linear_model1 160539.8  0.06705563
## 2 linear_model2 160532.2  0.06700925
## 3 linear_model3 160544.7  0.06649053
## 4 linear_model4 160535.9  0.06659172
```

Predictions

Predictions were made and saved to csv "Homework 4 Predictions". Logistic regression model 3 was used to predict car crash in variable TARGET_FLAG, and multiple linear regression model 2 was used to predict amount of claim for the car crash in variable TARGET_AMT. Multiple linear regression was only used on observations in the test data where our logistic regression predicts a car crash. Observations where there is not a predicted car crash defaults to a 0$ claim in TARGET_AMT.

```r
# Generate predicted probabilities
probs <- predict(logist_model3, newdata = test_df_imp, type = "response")

# Convert to binary predictions using a threshold (e.g., 0.5)
preds <- ifelse(probs > 0.5, 1, 0)

# Add predictions to your test data frame
test_df_imp$TARGET_FLAG <- preds
```

```r
pred_probs <- predict(logist_model3, train_df_imp, type = "response")
pred_class <- ifelse(pred_probs > 0.5, 1, 0)

# Confusion matrix on training data
confusionMatrix(
  factor(pred_class),
  factor(train_df_imp$TARGET_FLAG),
  positive = "1"
)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5547 1237
##          1  460  911
##
##                Accuracy : 0.7919
##                  95% CI : (0.7829, 0.8007)
##     No Information Rate : 0.7366
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3932
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4241
##             Specificity : 0.9234
##          Pos Pred Value : 0.6645
##          Neg Pred Value : 0.8177
##              Prevalence : 0.2634
##          Detection Rate : 0.1117
##    Detection Prevalence : 0.1681
##       Balanced Accuracy : 0.6738
##
##        'Positive' Class : 1
##
```

```r
# Predict claim amount only for predicted claimants
test_df_imp$TARGET_AMT <- ifelse(
  test_df_imp$TARGET_FLAG == 1,
  predict(linear_model2, newdata = test_df_imp),
  0
)
```

```r
# Creating predictions df
testdf_predictions <- test_df_imp
```

```r
# Predicted Values for TARGET_FLAG and TARGET_AMT
testdf_predictions$TARGET_FLAG
```

```
##    [1] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
##   [38] 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 1 1 0 0 0 0 1 0
##   [75] 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1
##  [112] 0 0 0 1 0 0 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0
##  [149] 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 1 1 1 1 1 0 0 0 0
##  [186] 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
##  [223] 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 1
##  [260] 0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0
##  [297] 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 1
##  [334] 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0
##  [371] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [408] 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
##  [445] 0 0 0 0 1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 1 0 0 0
##  [482] 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1
##  [519] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
##  [556] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0
##  [593] 0 0 1 1 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0
##  [630] 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [667] 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [704] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
##  [741] 1 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 1 0 0 0
##  [778] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [815] 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
##  [852] 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0
##  [889] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0
##  [926] 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
##  [963] 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1000] 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0
## [1037] 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 1 0 0 0
## [1074] 1 0 0 0 0 1 0 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1
## [1111] 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0
## [1148] 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 1 0 0
## [1185] 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0
## [1222] 0 1 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0
## [1259] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0
## [1296] 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0
## [1333] 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0
## [1370] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0
## [1407] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0
## [1444] 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
```

```
## [1481] 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0
## [1518] 1 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1
## [1555] 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1592] 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0
## [1629] 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 1 1 1 1
## [1666] 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 1 0
## [1703] 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
## [1740] 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0
## [1777] 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1
## [1814] 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0
## [1851] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
## [1888] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1925] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1
## [1962] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 1
## [1999] 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1
## [2036] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0
## [2073] 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 1 1 1 1 1 0 0 0 1 0 0
## [2110] 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
testdf_predictions$TARGET_AMT
```

```
##     [1]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##     [9]    0.000    0.000    0.000    0.000 4346.342    0.000    0.000 3360.126
##    [17] 3414.275    0.000 2515.321    0.000    0.000    0.000    0.000    0.000
##    [25]    0.000    0.000    0.000 3309.341    0.000    0.000    0.000    0.000
##    [33]    0.000    0.000    0.000    0.000    0.000    0.000    0.000 3084.804
##    [41]    0.000    0.000    0.000 3423.893    0.000    0.000    0.000    0.000
##    [49]    0.000 2740.136    0.000    0.000 4109.158    0.000    0.000    0.000
##    [57]    0.000    0.000    0.000 2287.818    0.000    0.000    0.000    0.000
##    [65]    0.000 1574.911 4013.430 3710.800    0.000    0.000    0.000    0.000
##    [73] 3318.909    0.000 3345.358    0.000    0.000    0.000    0.000    0.000
##    [81] 2357.352    0.000    0.000    0.000    0.000 3430.573    0.000    0.000
##    [89]    0.000 3934.113    0.000    0.000    0.000    0.000    0.000    0.000
##    [97]    0.000    0.000    0.000    0.000    0.000    0.000 3310.991 3535.094
##   [105]    0.000    0.000    0.000    0.000    0.000    0.000 2956.188    0.000
##   [113]    0.000    0.000 3069.187    0.000    0.000 3102.105 1840.159    0.000
##   [121]    0.000 3459.670 2666.295    0.000    0.000    0.000    0.000    0.000
##   [129]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##   [137] 4119.037 3553.831    0.000    0.000    0.000 3699.183    0.000    0.000
##   [145]    0.000 2903.229    0.000    0.000    0.000    0.000    0.000    0.000
##   [153] 4322.071    0.000    0.000 2189.982    0.000    0.000 2265.982    0.000
##   [161]    0.000    0.000    0.000    0.000 4904.801    0.000    0.000    0.000
##   [169]    0.000    0.000    0.000 3222.693    0.000 4176.113    0.000    0.000
##   [177] 3528.219 3174.611 4146.255 3050.563 3254.074    0.000    0.000    0.000
##   [185]    0.000    0.000    0.000    0.000    0.000    0.000 2164.372 3248.140
##   [193]    0.000    0.000    0.000 2341.528    0.000    0.000    0.000    0.000
##   [201]    0.000    0.000    0.000    0.000    0.000    0.000 4190.535    0.000
##   [209]    0.000    0.000    0.000    0.000 2977.801 2955.963    0.000    0.000
##   [217]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##   [225]    0.000    0.000 3358.296    0.000    0.000    0.000    0.000    0.000
##   [233] 3694.071    0.000    0.000    0.000    0.000    0.000    0.000 3818.680
##   [241]    0.000    0.000 3959.213    0.000    0.000    0.000    0.000    0.000
##   [249]    0.000 2488.343    0.000 2668.967    0.000    0.000 1694.036    0.000
##   [257]    0.000    0.000 2629.215    0.000    0.000    0.000    0.000    0.000
```

```
## [265]    0.000    0.000    0.000    0.000 4793.566    0.000 2521.215    0.000
## [273]    0.000 2716.135    0.000    0.000 3951.662    0.000    0.000    0.000
## [281]    0.000    0.000    0.000    0.000    0.000 2266.173    0.000    0.000
## [289] 2008.997 3633.614    0.000    0.000    0.000    0.000    0.000    0.000
## [297]    0.000 3154.918    0.000    0.000    0.000    0.000    0.000    0.000
## [305]    0.000    0.000    0.000 2165.967    0.000    0.000 4210.164    0.000
## [313]    0.000 4742.237    0.000    0.000    0.000    0.000 2767.508    0.000
## [321]    0.000 2739.156    0.000    0.000 3620.464    0.000    0.000    0.000
## [329]    0.000    0.000    0.000    0.000 3867.975    0.000    0.000    0.000
## [337]    0.000    0.000    0.000    0.000 2628.751    0.000    0.000 3237.668
## [345]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [353] 4182.229 4268.812    0.000 2683.003    0.000    0.000    0.000    0.000
## [361] 3072.300    0.000    0.000 2598.351    0.000    0.000    0.000    0.000
## [369]    0.000    0.000    0.000    0.000    0.000    0.000    0.000 2710.464
## [377]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [385]    0.000    0.000    0.000    0.000    0.000 2344.299    0.000    0.000
## [393]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [401]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [409]    0.000    0.000    0.000 3184.381    0.000    0.000    0.000    0.000
## [417]    0.000    0.000    0.000    0.000 2902.198 3769.647 4420.564    0.000
## [425]    0.000    0.000    0.000    0.000 2733.449    0.000    0.000    0.000
## [433]    0.000    0.000    0.000 4091.925    0.000    0.000    0.000    0.000
## [441]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [449] 3073.991 3877.864    0.000    0.000    0.000    0.000    0.000 2741.662
## [457]    0.000 4700.213    0.000    0.000    0.000    0.000    0.000    0.000
## [465]    0.000    0.000 4204.309 2240.163    0.000    0.000    0.000 3412.091
## [473]    0.000    0.000    0.000    0.000 3493.719 4141.819    0.000    0.000
## [481]    0.000    0.000    0.000    0.000 3520.683 2768.192    0.000    0.000
## [489]    0.000 3978.810 2991.445    0.000    0.000    0.000    0.000    0.000
## [497]    0.000    0.000    0.000    0.000    0.000    0.000 3420.266    0.000
## [505] 3883.747    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [513]    0.000    0.000    0.000    0.000 3340.964 3040.315    0.000 3010.149
## [521]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [529]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [537]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [545]    0.000    0.000    0.000 4687.548    0.000    0.000    0.000    0.000
## [553]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [561]    0.000    0.000    0.000    0.000    0.000    0.000 5514.220    0.000
## [569]    0.000 3139.104    0.000    0.000    0.000    0.000    0.000    0.000
## [577]    0.000    0.000    0.000    0.000    0.000 2705.559    0.000 3216.993
## [585]    0.000    0.000    0.000    0.000 3616.528    0.000    0.000    0.000
## [593]    0.000    0.000 2614.935 2458.293 2877.629    0.000    0.000 2138.643
## [601] 2722.592    0.000    0.000    0.000    0.000    0.000 3791.845    0.000
## [609]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [617]    0.000    0.000    0.000 3657.614    0.000    0.000    0.000    0.000
## [625]    0.000 4320.423 3039.101    0.000    0.000 3143.036    0.000    0.000
## [633]    0.000    0.000    0.000    0.000    0.000 2975.074    0.000    0.000
## [641]    0.000    0.000    0.000    0.000    0.000 2980.049    0.000    0.000
## [649]    0.000    0.000    0.000    0.000 3441.578    0.000    0.000    0.000
## [657]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [665]    0.000    0.000    0.000    0.000    0.000    0.000    0.000 2821.537
## [673] 2947.197    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [681]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [689]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
```

```
##  [697]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [705]    0.000    0.000    0.000 4107.645    0.000    0.000    0.000    0.000
##  [713]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [721] 2515.978    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [729]    0.000    0.000 2347.332 3011.205    0.000    0.000    0.000    0.000
##  [737]    0.000    0.000    0.000    0.000 2975.883    0.000 3233.931    0.000
##  [745]    0.000    0.000 3440.390    0.000    0.000    0.000    0.000    0.000
##  [753]    0.000 2830.301    0.000    0.000    0.000    0.000    0.000    0.000
##  [761]    0.000 3955.387    0.000 3108.785 2442.731 3185.649    0.000    0.000
##  [769]    0.000    0.000    0.000 2704.319    0.000 3712.839    0.000    0.000
##  [777]    0.000    0.000    0.000    0.000    0.000 4219.961    0.000    0.000
##  [785]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [793]    0.000    0.000    0.000    0.000    0.000 2769.601 3412.353    0.000
##  [801]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [809]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [817]    0.000 3367.857 3093.542    0.000 2776.364    0.000    0.000    0.000
##  [825] 2940.732    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [833]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [841]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [849] 3961.017 3381.752 3231.657    0.000    0.000    0.000 2516.375    0.000
##  [857]    0.000    0.000 4049.653    0.000    0.000 3051.534    0.000    0.000
##  [865]    0.000    0.000 3838.369    0.000    0.000 3284.558    0.000 3452.100
##  [873]    0.000 4486.066    0.000    0.000    0.000    0.000    0.000    0.000
##  [881]    0.000    0.000    0.000    0.000 2919.781    0.000 3942.148    0.000
##  [889]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [897]    0.000    0.000 2231.554    0.000    0.000    0.000 3997.104    0.000
##  [905]    0.000    0.000 3275.490    0.000    0.000    0.000 3948.241    0.000
##  [913]    0.000    0.000    0.000    0.000 4696.655 2603.120    0.000    0.000
##  [921]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [929] 3121.541    0.000    0.000 4391.628    0.000    0.000    0.000    0.000
##  [937]    0.000    0.000    0.000    0.000 3807.007    0.000 2112.229    0.000
##  [945] 3035.707    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [953]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
##  [961] 2181.017    0.000    0.000    0.000    0.000 3541.376    0.000    0.000
##  [969]    0.000    0.000 3164.032    0.000    0.000    0.000    0.000    0.000
##  [977]    0.000    0.000    0.000    0.000    0.000    0.000 3593.193 2637.747
##  [985] 3134.431 2692.620    0.000    0.000    0.000    0.000    0.000    0.000
##  [993]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1001]    0.000 4381.639 2982.583    0.000    0.000    0.000    0.000    0.000
## [1009]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1017]    0.000    0.000    0.000    0.000    0.000    0.000 3587.903    0.000
## [1025] 3409.510 3295.644 2021.219    0.000    0.000    0.000    0.000    0.000
## [1033]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1041]    0.000 1886.266 1690.410    0.000 3709.535    0.000    0.000    0.000
## [1049] 2451.514    0.000    0.000    0.000 3006.525    0.000    0.000    0.000
## [1057]    0.000    0.000 2371.307 2903.436    0.000 3185.622    0.000    0.000
## [1065]    0.000    0.000    0.000    0.000    0.000 2304.252    0.000    0.000
## [1073]    0.000 4290.718    0.000    0.000    0.000    0.000 3027.488    0.000
## [1081] 3596.314 3078.302    0.000    0.000 2247.397 4343.952    0.000    0.000
## [1089]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1097]    0.000    0.000    0.000 3346.956    0.000    0.000    0.000    0.000
## [1105]    0.000    0.000    0.000    0.000    0.000 1601.949 2790.555    0.000
## [1113]    0.000    0.000    0.000    0.000    0.000 4471.567    0.000    0.000
## [1121]    0.000 1988.596    0.000    0.000    0.000    0.000 4673.711    0.000
```

```
## [1129]    0.000    0.000    0.000    0.000 3295.311    0.000    0.000    0.000
## [1137]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1145] 3025.576    0.000    0.000 3973.971    0.000    0.000    0.000 4317.051
## [1153]    0.000    0.000 3562.331    0.000    0.000    0.000    0.000    0.000
## [1161]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1169]    0.000    0.000 1913.090    0.000    0.000 3522.103    0.000    0.000
## [1177]    0.000    0.000    0.000 3519.685    0.000 3685.170    0.000    0.000
## [1185] 3734.497    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1193]    0.000 2915.425    0.000    0.000    0.000    0.000 2519.349    0.000
## [1201]    0.000    0.000    0.000    0.000    0.000    0.000 2477.787    0.000
## [1209]    0.000    0.000    0.000    0.000 3874.528    0.000    0.000    0.000
## [1217] 3073.224    0.000    0.000    0.000    0.000    0.000 3977.356    0.000
## [1225] 2601.493    0.000    0.000    0.000    0.000 3734.712    0.000    0.000
## [1233]    0.000 3402.848    0.000    0.000    0.000 3784.917    0.000    0.000
## [1241]    0.000    0.000    0.000    0.000    0.000 4391.160    0.000    0.000
## [1249]    0.000    0.000    0.000 3790.210    0.000    0.000    0.000    0.000
## [1257] 3721.544    0.000    0.000    0.000    0.000    0.000    0.000 3537.800
## [1265]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1273]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1281] 3158.783    0.000    0.000    0.000    0.000    0.000 2408.281    0.000
## [1289]    0.000    0.000 3556.992    0.000    0.000    0.000    0.000    0.000
## [1297]    0.000    0.000 3330.694    0.000    0.000    0.000    0.000    0.000
## [1305]    0.000    0.000 2942.387    0.000    0.000 3307.941    0.000    0.000
## [1313] 4462.081    0.000    0.000    0.000    0.000    0.000    0.000 3811.096
## [1321]    0.000    0.000 2946.312    0.000    0.000 2808.463    0.000    0.000
## [1329]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1337]    0.000    0.000    0.000    0.000    0.000 2524.099    0.000    0.000
## [1345] 3786.703    0.000    0.000    0.000    0.000    0.000    0.000 2604.646
## [1353]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1361]    0.000    0.000 4059.667    0.000    0.000    0.000 2962.060 2500.707
## [1369]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1377]    0.000    0.000    0.000    0.000    0.000 3368.029    0.000    0.000
## [1385]    0.000    0.000    0.000    0.000    0.000    0.000 3210.848 2645.135
## [1393]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1401]    0.000    0.000 3853.971 2206.858    0.000    0.000    0.000    0.000
## [1409]    0.000 3189.211    0.000    0.000    0.000    0.000    0.000    0.000
## [1417]    0.000    0.000    0.000    0.000    0.000 4638.128 2231.716    0.000
## [1425]    0.000 4747.953    0.000    0.000 2179.216    0.000    0.000    0.000
## [1433]    0.000    0.000    0.000    0.000    0.000 2428.047    0.000    0.000
## [1441]    0.000 3510.759    0.000    0.000    0.000 2829.904    0.000    0.000
## [1449]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1457] 3555.280    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1465]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1473]    0.000    0.000    0.000    0.000 2178.332    0.000    0.000    0.000
## [1481]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1489] 3501.311    0.000    0.000 3161.067    0.000    0.000 3711.656    0.000
## [1497]    0.000    0.000    0.000    0.000    0.000    0.000    0.000 3597.845
## [1505]    0.000    0.000 3137.129    0.000    0.000    0.000    0.000    0.000
## [1513]    0.000    0.000    0.000    0.000    0.000 1817.183    0.000    0.000
## [1521]    0.000    0.000    0.000 2016.923    0.000    0.000    0.000    0.000
## [1529] 2657.317 3161.089    0.000    0.000    0.000    0.000    0.000    0.000
## [1537]    0.000    0.000 3619.778 3945.420    0.000    0.000    0.000    0.000
## [1545] 3103.711    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1553]    0.000 3033.896    0.000    0.000    0.000    0.000    0.000    0.000
```

```
## [1561]    0.000 3062.007    0.000    0.000 4534.045    0.000    0.000    0.000
## [1569]    0.000    0.000 2779.412    0.000    0.000    0.000    0.000    0.000
## [1577]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1585]    0.000    0.000    0.000    0.000    0.000    0.000    0.000 4004.899
## [1593]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1601]    0.000    0.000    0.000 3255.657    0.000 4195.879    0.000    0.000
## [1609]    0.000 4081.018    0.000    0.000    0.000    0.000    0.000 2691.572
## [1617]    0.000 2999.243    0.000    0.000 2693.528    0.000 3383.426    0.000
## [1625]    0.000    0.000    0.000    0.000    0.000 4419.178    0.000 3129.360
## [1633]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1641]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1649]    0.000 2455.384    0.000    0.000    0.000    0.000 3177.257    0.000
## [1657]    0.000    0.000    0.000    0.000 2336.781 3839.072 2756.927 3871.550
## [1665] 5293.727    0.000    0.000    0.000 4162.510    0.000 2933.068    0.000
## [1673] 3520.379    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1681]    0.000 3693.396 3202.301    0.000    0.000    0.000    0.000    0.000
## [1689]    0.000    0.000    0.000    0.000    0.000    0.000    0.000 3065.616
## [1697]    0.000 2756.420 4134.429    0.000 2580.138    0.000    0.000 1361.694
## [1705]    0.000    0.000 3332.846    0.000 4243.905    0.000    0.000    0.000
## [1713]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1721]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1729] 3637.484    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1737]    0.000    0.000    0.000    0.000 4570.431    0.000    0.000    0.000
## [1745]    0.000    0.000    0.000    0.000 2629.770    0.000    0.000    0.000
## [1753]    0.000 3039.880    0.000    0.000    0.000    0.000    0.000    0.000
## [1761] 3852.260    0.000    0.000    0.000    0.000 3742.656    0.000    0.000
## [1769]    0.000    0.000    0.000    0.000    0.000 4178.757    0.000    0.000
## [1777] 2585.026    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1785]    0.000    0.000    0.000    0.000 3280.636    0.000    0.000    0.000
## [1793]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1801]    0.000    0.000    0.000    0.000    0.000    0.000 3155.765 3387.546
## [1809]    0.000    0.000    0.000    0.000 2351.594    0.000 1871.423    0.000
## [1817]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1825]    0.000 3632.162    0.000    0.000    0.000    0.000    0.000    0.000
## [1833]    0.000    0.000    0.000    0.000    0.000 4363.327    0.000 3431.554
## [1841]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1849]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1857]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1865]    0.000    0.000    0.000    0.000    0.000    0.000 3068.884    0.000
## [1873]    0.000    0.000    0.000    0.000    0.000 3668.838    0.000    0.000
## [1881]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1889]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1897]    0.000    0.000    0.000 2510.126    0.000    0.000    0.000    0.000
## [1905]    0.000    0.000    0.000    0.000    0.000 3953.076    0.000    0.000
## [1913]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1921]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1929]    0.000    0.000    0.000    0.000    0.000 3472.979    0.000    0.000
## [1937]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1945]    0.000    0.000 3043.451    0.000 2573.287    0.000    0.000    0.000
## [1953]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1961] 3855.575    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1969]    0.000    0.000    0.000    0.000    0.000 3252.756    0.000    0.000
## [1977]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [1985]    0.000    0.000    0.000    0.000    0.000 2403.709    0.000 3311.629
```

```
## [1993] 3486.338 3299.985    0.000    0.000    0.000 4390.028    0.000    0.000
## [2001] 2770.952    0.000 3559.453    0.000 3006.626    0.000    0.000    0.000
## [2009]    0.000    0.000 2095.736    0.000 5387.863    0.000    0.000    0.000
## [2017]    0.000 3104.122    0.000    0.000    0.000    0.000    0.000    0.000
## [2025]    0.000    0.000    0.000    0.000    0.000 2806.144    0.000    0.000
## [2033]    0.000    0.000 2950.330    0.000    0.000    0.000 2551.447    0.000
## [2041]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [2049]    0.000    0.000    0.000    0.000    0.000    0.000 3217.130    0.000
## [2057]    0.000    0.000    0.000    0.000    0.000    0.000 2067.641    0.000
## [2065]    0.000    0.000    0.000 4185.271    0.000    0.000    0.000    0.000
## [2073] 3413.971    0.000    0.000    0.000 3636.778    0.000    0.000 4914.937
## [2081]    0.000    0.000    0.000    0.000    0.000    0.000    0.000 2776.496
## [2089]    0.000 2855.831    0.000    0.000    0.000    0.000    0.000 3021.110
## [2097]    0.000    0.000 3832.960 2675.241 3246.550 1560.755 4739.731    0.000
## [2105]    0.000    0.000 2799.317    0.000    0.000    0.000 4245.857    0.000
## [2113]    0.000    0.000    0.000    0.000    0.000    0.000 5148.513    0.000
## [2121]    0.000    0.000 3880.107    0.000    0.000    0.000    0.000    0.000
## [2129]    0.000    0.000    0.000    0.000    0.000    0.000    0.000    0.000
## [2137]    0.000    0.000    0.000    0.000    0.000
```

```r
# writing predictions to csv

write.csv(testdf_predictions, "Homework 4 Predictions")
```