# Data 621 Assignment 4

Mark Gonsalves, Joshua Hummell, Claire Meyer, Chinedu Onyeka, Rathish Parayil Sasidharan

April 12th, 2022

**Overview**

In this homework assignment, you will 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.
Your objective is 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 and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

**1. Data Exploration**

Describe the size and the variables in the insurance training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.
a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

**Required Libraries**

```
library(tidyverse)
library(Amelia)
library("naniar")
library(pROC)
library(visdat)
library(cowplot)
library(corrplot)
library(kableExtra)
library(Hmisc)
library(caTools)
library(car)
library(caret)
library(lmtest)
library("MASS")
```

**Load the data from github**

```
#url_train <- "https://raw.githubusercontent.com/chinedu2301/data621-business-analytics-data-mining/mai
#url_eval <- "https://raw.githubusercontent.com/chinedu2301/data621-business-analytics-data-mining/main
url_train <- "insurance_training_data .csv"
url_eval  <- "insurance-evaluation-data.csv"
training <- read_csv(url_train)
evaluation <- read_csv(url_eval)
```

**Check the head of training dataset**

```
head(training, 10)
```

```
## # A tibble: 10 x 26
##    INDEX TARGET_FLAG TARGET_AMT KIDSDRIV   AGE HOMEKIDS   YOJ INCOME   PARENT1
##    <dbl>       <dbl>      <dbl>    <dbl> <dbl>    <dbl> <dbl> <chr>    <chr>
##  1     1           0          0        0    60        0    11 $67,349  No
##  2     2           0          0        0    43        0    11 $91,449  No
##  3     4           0          0        0    35        1    10 $16,039  No
##  4     5           0          0        0    51        0    14 <NA>     No
##  5     6           0          0        0    50        0    NA $114,986 No
##  6     7           1       2946        0    34        1    12 $125,301 Yes
##  7     8           0          0        0    54        0    NA $18,755  No
##  8    11           1       4021        1    37        2    NA $107,961 No
##  9    12           1       2501        0    34        0    10 $62,978  No
## 10    13           0          0        0    50        0     7 $106,952 No
## # ... with 17 more variables: HOME_VAL <chr>, MSTATUS <chr>, SEX <chr>,
## #   EDUCATION <chr>, JOB <chr>, TRAVTIME <dbl>, CAR_USE <chr>, BLUEBOOK <chr>,
## #   TIF <dbl>, CAR_TYPE <chr>, RED_CAR <chr>, OLDCLAIM <chr>, CLM_FREQ <dbl>,
## #   REVOKED <chr>, MVR_PTS <dbl>, CAR_AGE <dbl>, URBANICITY <chr>
```

**Get the dimension of the training dataset**

```
dim(training)
```

```
## [1] 8161   26
```

**Get a glimpse of the training dataset**

```
glimpse(training)
```

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <dbl> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <dbl> 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    <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <dbl> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <dbl> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <dbl> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", NA, "$114,986", "$125,301~
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No~
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "$0"~
```

| | INDEX | TARGET_FLAG | TARGET_AMT | KIDSDRIV | AGE | HOMEKIDS | YOJ |
|---|---|---|---|---|---|---|---|
| | Min. : 1 | Min. :0.0000 | Min. : 0 | Min. :0.0000 | Min. :16.00 | Min. :0.0000 | Min. : 0. |
| | 1st Qu.: 2559 | 1st Qu.:0.0000 | 1st Qu.: 0 | 1st Qu.:0.0000 | 1st Qu.:39.00 | 1st Qu.:0.0000 | 1st Qu.: |
| | Median : 5133 | Median :0.0000 | Median : 0 | Median :0.0000 | Median :45.00 | Median :0.0000 | Median : |
| | Mean : 5152 | Mean :0.2638 | Mean : 1504 | Mean :0.1711 | Mean :44.79 | Mean :0.7212 | Mean :10 |
| | 3rd Qu.: 7745 | 3rd Qu.:1.0000 | 3rd Qu.: 1036 | 3rd Qu.:0.0000 | 3rd Qu.:51.00 | 3rd Qu.:1.0000 | 3rd Qu.:1 |
| | Max. :10302 | Max. :1.0000 | Max. :107586 | Max. :4.0000 | Max. :81.00 | Max. :5.0000 | Max. :23 |
| | NA | NA | NA | NA | NA's :6 | NA | NA's :454 |

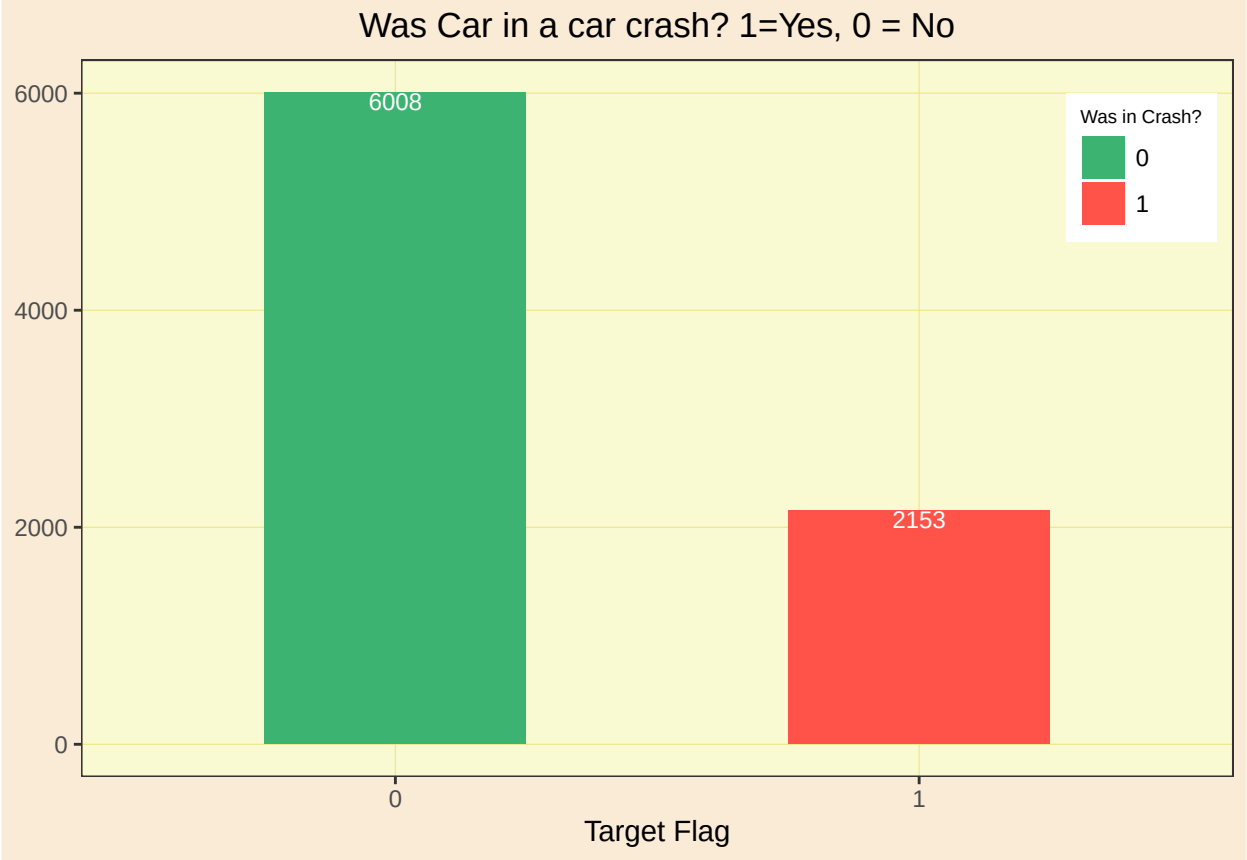
```
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Yes",~
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", "M"~
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School", "~
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Colla~
## $ TRAVTIME    <dbl> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private", ~
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "$17~
## $ TIF         <dbl> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Sports~
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no", ~
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0", "$~
## $ CLM_FREQ    <dbl> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
## $ MVR_PTS     <dbl> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <dbl> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly Urba~
```

**Summary Statistics** - Mean and median of numerical columns are shown below:

```
summary(training) %>% kbl() %>% kable_styling()
```

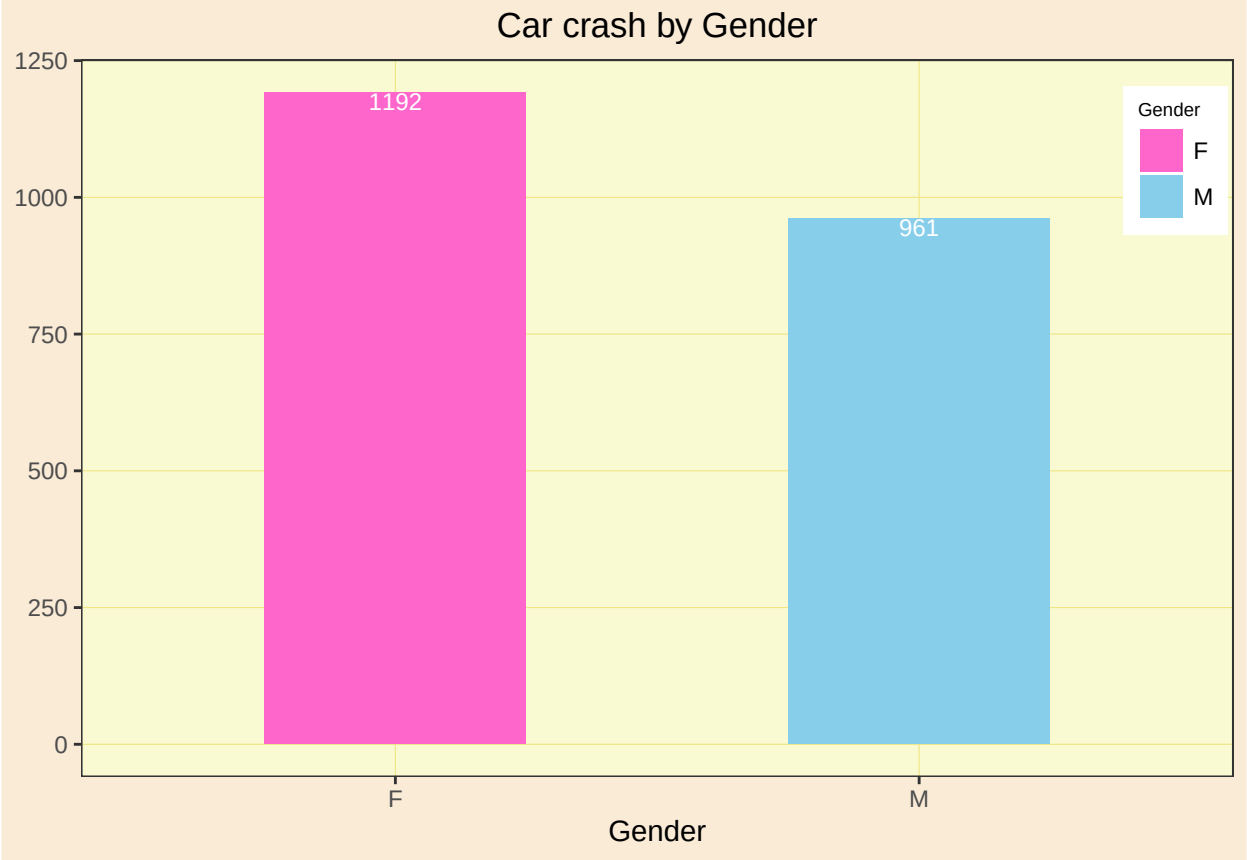
**Understand the data using visuals  Target Flag - Was Car in a crash? 1=YES, 0=NO**

```
# plot a bar chart to show the distribution of car crash
target_flag <- training %>% ggplot(aes(x = as.factor(TARGET_FLAG), fill = as.factor(TARGET_FLAG))) +
    geom_bar(width = 0.5) + theme_bw() +
    geom_text(stat = 'count', aes(label = ..count..), vjust = 1, color = "white", size = 3.2) +
    scale_fill_manual("Was in Crash?", values = c("0" = "#3CB371", "1" = "#FF5349")) +
    theme(panel.grid.major = element_line(colour = "khaki2",
    linetype = "blank"), plot.title = element_text(face = "bold",
    hjust = 0.5), panel.background = element_rect(fill = "lightgoldenrodyellow"),
    plot.background = element_rect(fill = "antiquewhite")) +labs(title = "Was Car in a car crash? 1=Yes
    x = "Target Flag", y = NULL) + theme(panel.grid.minor = element_line(linetype = "blank"),
    plot.title = element_text(face = "plain"),
    legend.title = element_text(size = 7),
    legend.position = c(0.92, 0.85)) + theme(panel.grid.major = element_line(size = 0.2,
    linetype = "solid"))

# display the chart
target_flag
```




**Car crash by Gender**

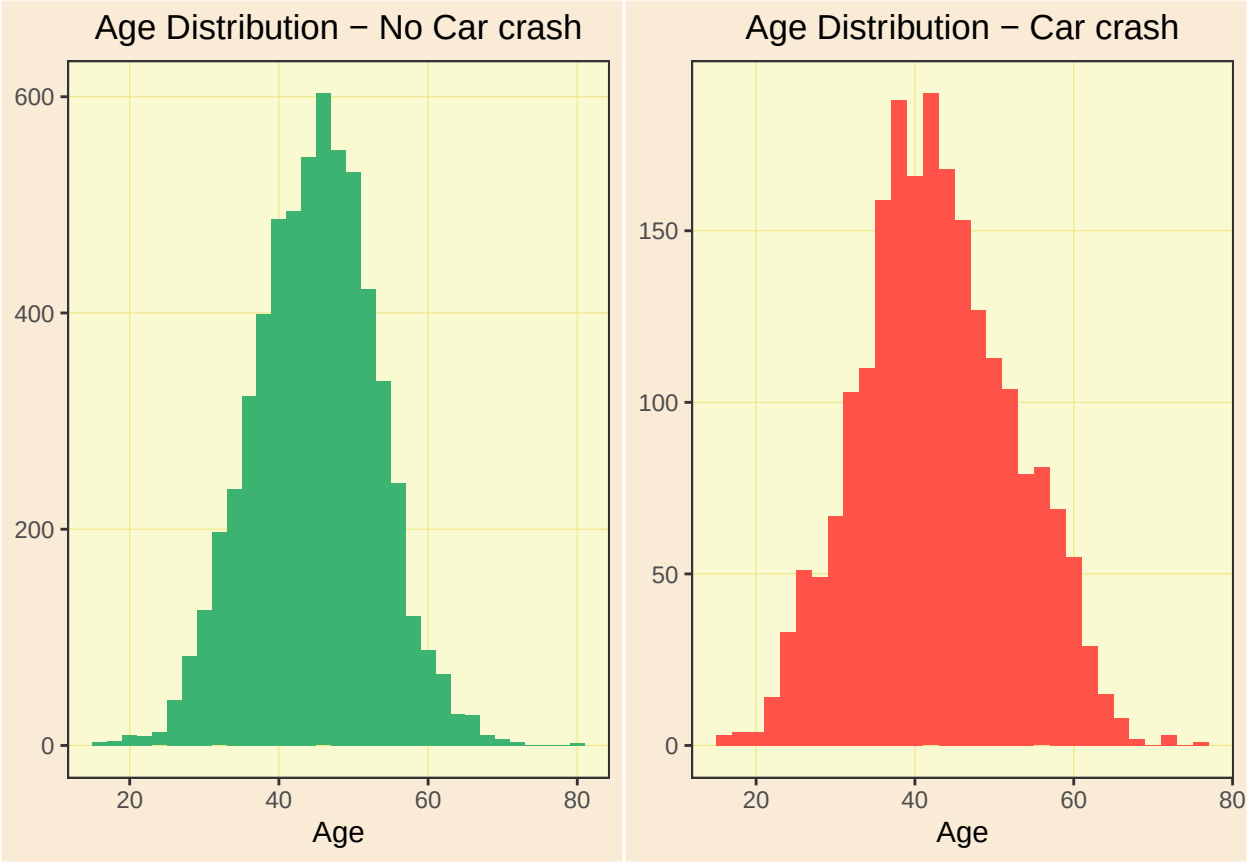
```
# clean the SEX column
training$SEX <- gsub("z_", "", training$SEX)

# plot the gender distribution
car_crash_gender <- training %>% filter(TARGET_FLAG == 1) %>% ggplot(aes(x = as.factor(SEX),
                                                                         fill = as.factor(SEX))) +
    geom_bar(width = 0.5) + theme_bw() +
    geom_text(stat = 'count', aes(label = ..count..), vjust = 1, color = "white", size = 3.2) +
    scale_fill_manual("Gender", values = c("F" = "#FF66CC", "M" = "#87CEEB")) +
    theme(panel.grid.major = element_line(colour = "khaki2",
    linetype = "blank"), plot.title = element_text(face = "bold",
    hjust = 0.5), panel.background = element_rect(fill = "lightgoldenrodyellow"),
    plot.background = element_rect(fill = "antiquewhite")) +labs(title = "Car crash by Gender",
    x = "Gender", y = NULL) + theme(panel.grid.minor = element_line(linetype = "blank"),
    plot.title = element_text(face = "plain"),
    legend.title = element_text(size = 7),
    legend.position = c(0.95, 0.86)) + theme(panel.grid.major = element_line(size = 0.2,
    linetype = "solid"))
car_crash_gender
```




**Age Distribution**

```
# plot the age distribution of individuals involved in a crash
age_crash <- training %>% filter(TARGET_FLAG == 1) %>% ggplot(aes(x = AGE)) +
    geom_histogram(binwidth = 2, fill = "#FF5349") + theme_bw() +
    theme(panel.grid.major = element_line(colour = "khaki2",
    linetype = "blank"), plot.title = element_text(face = "bold",
    hjust = 0.5), panel.background = element_rect(fill = "lightgoldenrodyellow"),
    plot.background = element_rect(fill = "antiquewhite")) +
    labs(title = "Age Distribution - Car crash",
    x = "Age", y = NULL) + theme(panel.grid.minor = element_line(linetype = "blank"),
    plot.title = element_text(face = "plain"),
    legend.title = element_text(size = 7),
    legend.position = c(0.95, 0.86)) + theme(panel.grid.major = element_line(size = 0.2,
    linetype = "solid"))

# plot the age distribution of individuals not involved in a crash
age_no_crash <- training %>% filter(TARGET_FLAG == 0) %>% ggplot(aes(x = AGE)) +
    geom_histogram(binwidth = 2, fill = "#3CB371") + theme_bw() +
    theme(panel.grid.major = element_line(colour = "khaki2",
    linetype = "blank"), plot.title = element_text(face = "bold",
    hjust = 0.5), panel.background = element_rect(fill = "lightgoldenrodyellow"),
    plot.background = element_rect(fill = "antiquewhite")) +
    labs(title = "Age Distribution - No Car crash",
    x = "Age", y = NULL) + theme(panel.grid.minor = element_line(linetype = "blank"),
    plot.title = element_text(face = "plain"),
```

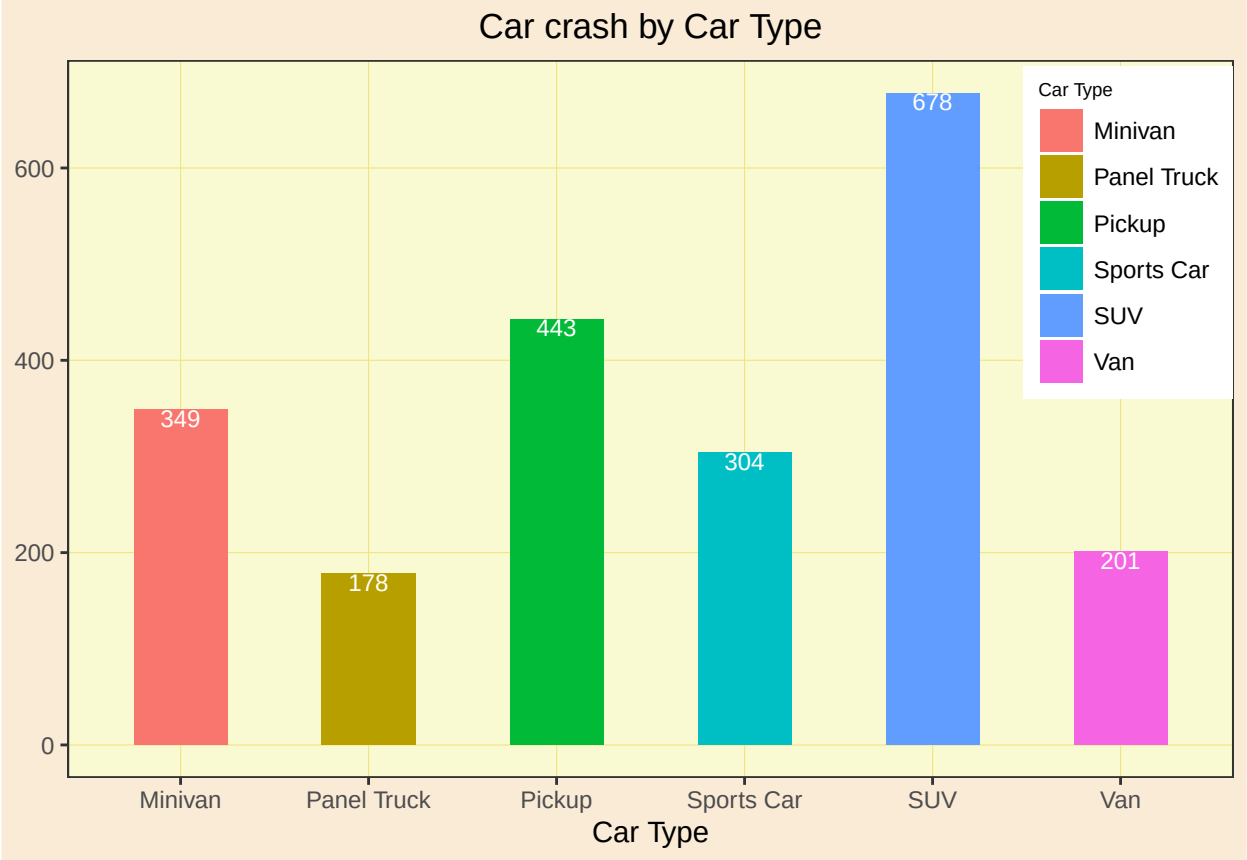
```
    legend.title = element_text(size = 7),
    legend.position = c(0.95, 0.86)) + theme(panel.grid.major = element_line(size = 0.2,
    linetype = "solid"))

# display the plots
plot_grid(age_no_crash, age_crash)
```




**Car crash by Car Type**

```
# clean the CAR_TYPE column
training$CAR_TYPE <- gsub("z_", "", training$CAR_TYPE)

# plot the gender distribution
car_crash_car_type <- training %>% filter(TARGET_FLAG == 1) %>% ggplot(aes(x = as.factor(CAR_TYPE),
                                                                           fill = as.factor(CAR_TYPE))) +
    geom_bar(width = 0.5) + theme_bw() + scale_fill_discrete(name = "Car Type") +
    geom_text(stat = 'count', aes(label = ..count..), vjust = 1, color = "white", size = 3.2) +
    theme(panel.grid.major = element_line(colour = "khaki2",
    linetype = "blank"), plot.title = element_text(face = "bold",
    hjust = 0.5), panel.background = element_rect(fill = "lightgoldenrodyellow"),
    plot.background = element_rect(fill = "antiquewhite")) +labs(title = "Car crash by Car Type",
    x = "Car Type", y = NULL) + theme(panel.grid.minor = element_line(linetype = "blank"),
    plot.title = element_text(face = "plain"),
    legend.title = element_text(size = 7),
    legend.position = c(0.91, 0.76)) + theme(panel.grid.major = element_line(size = 0.2,
```

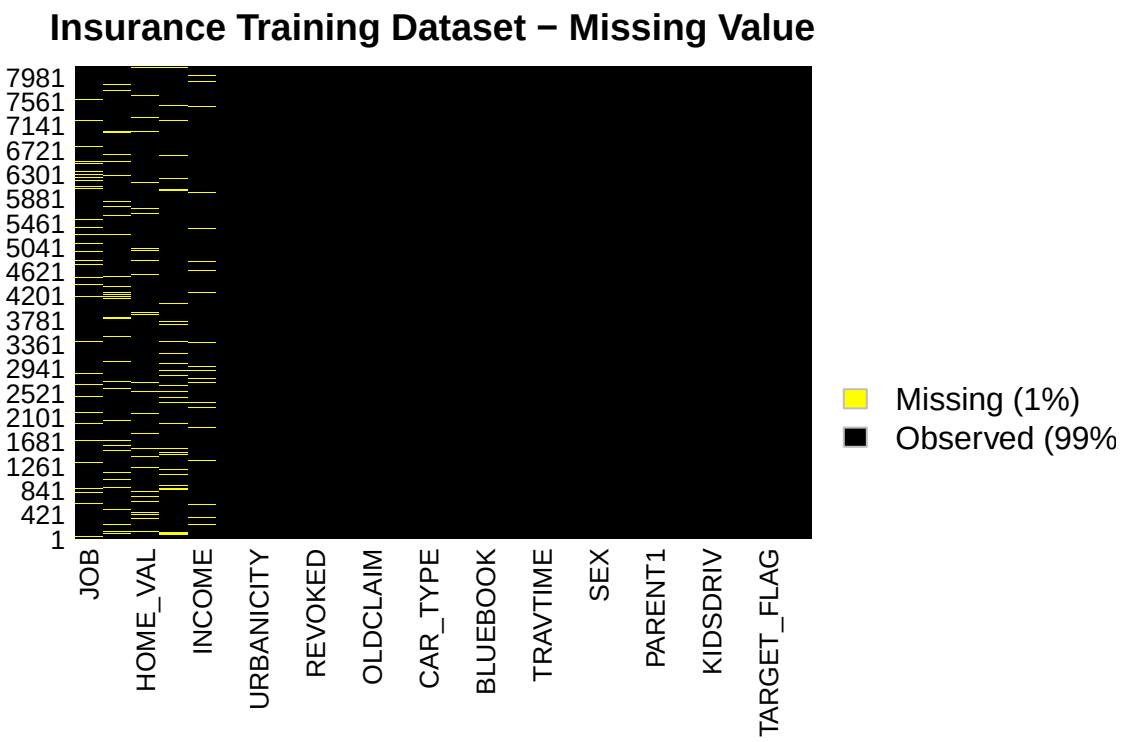
```
    linetype = "solid"))
car_crash_car_type
```




**Missing Values**

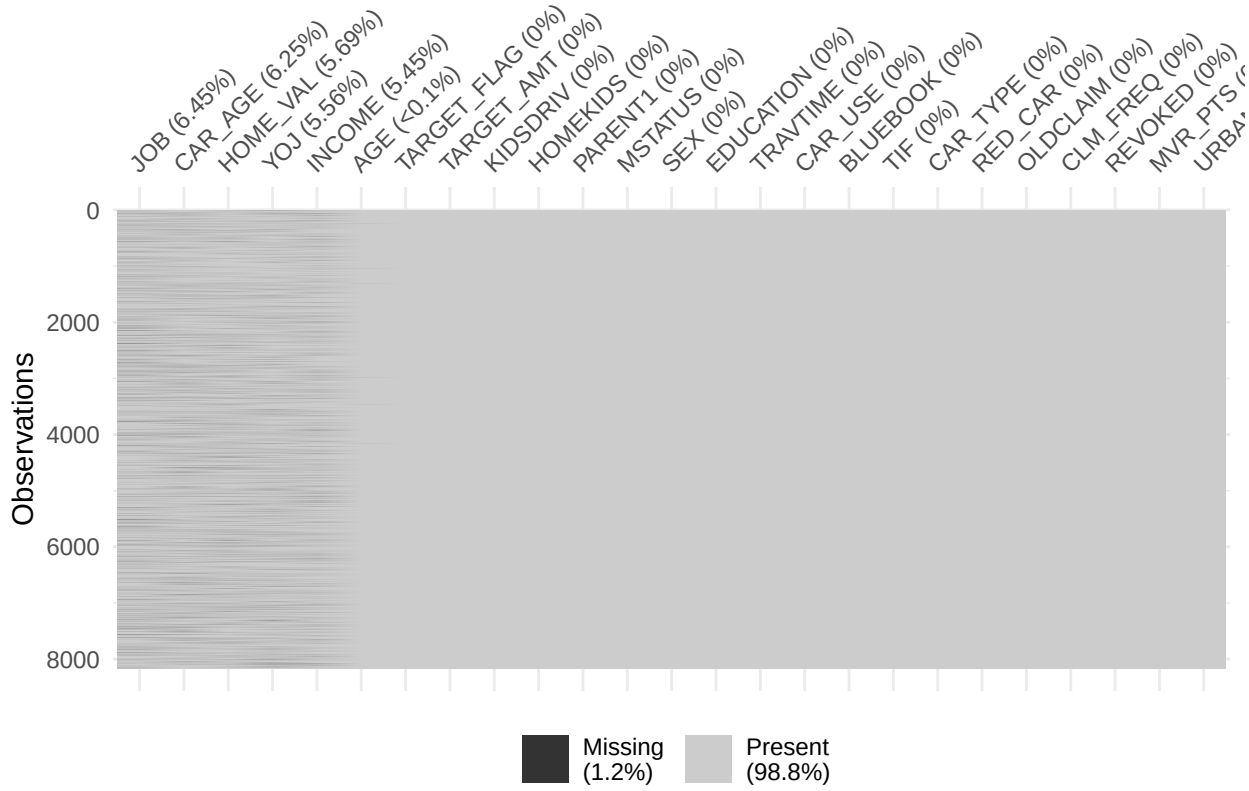
*Visualize the missing data using the Amelia package*

```
# check for NA values using the missmap function from Amelia package
missmap(training, main = "Insurance Training Dataset - Missing Values",
        col = c("yellow", "black"), margins = c(8,5))
```



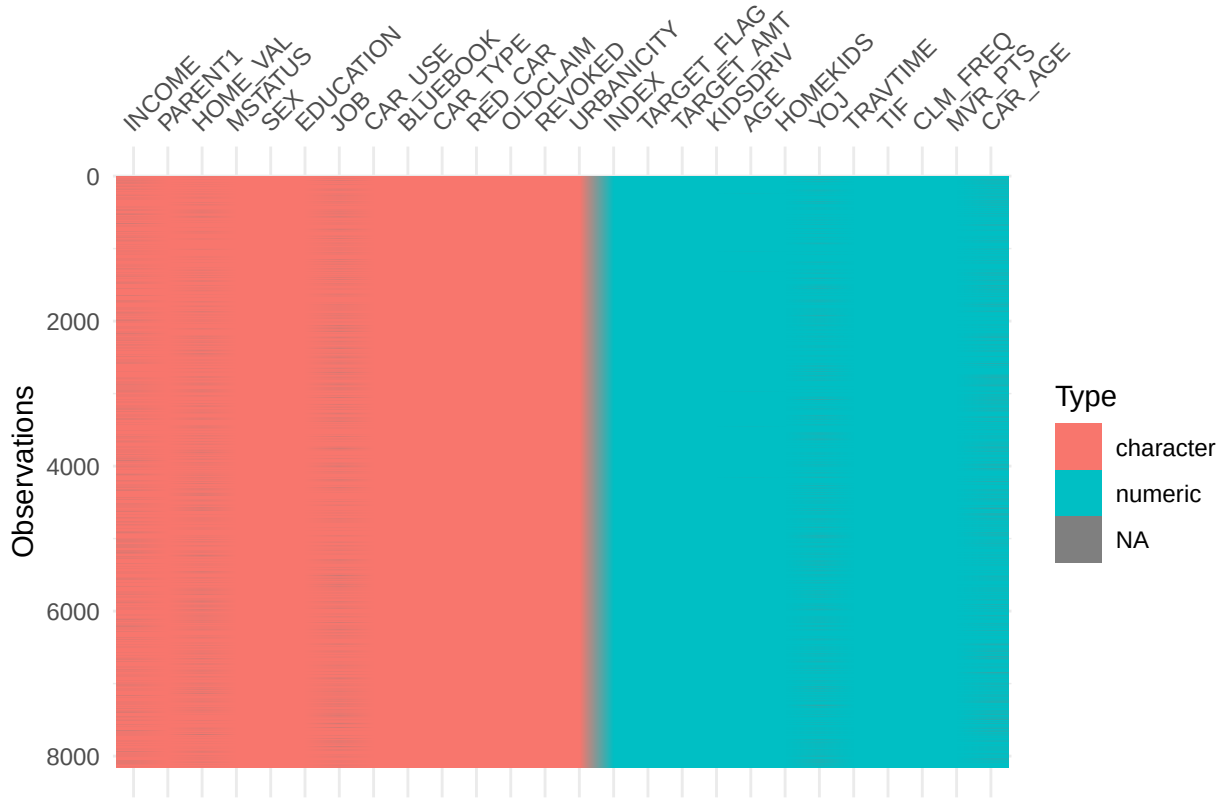

*Visualize the percentage of missingness in each column using the nanair package*

```
# visualize missing values and the percentage of missingness for each column
training %>% dplyr::select(-INDEX) %>% vis_miss(sort_miss = TRUE)
```




*Visualize the missing values using visdat package*

```
# visualize the missing values and their data type using vis_dat
vis_dat(training)
```




**Clean the dataset**

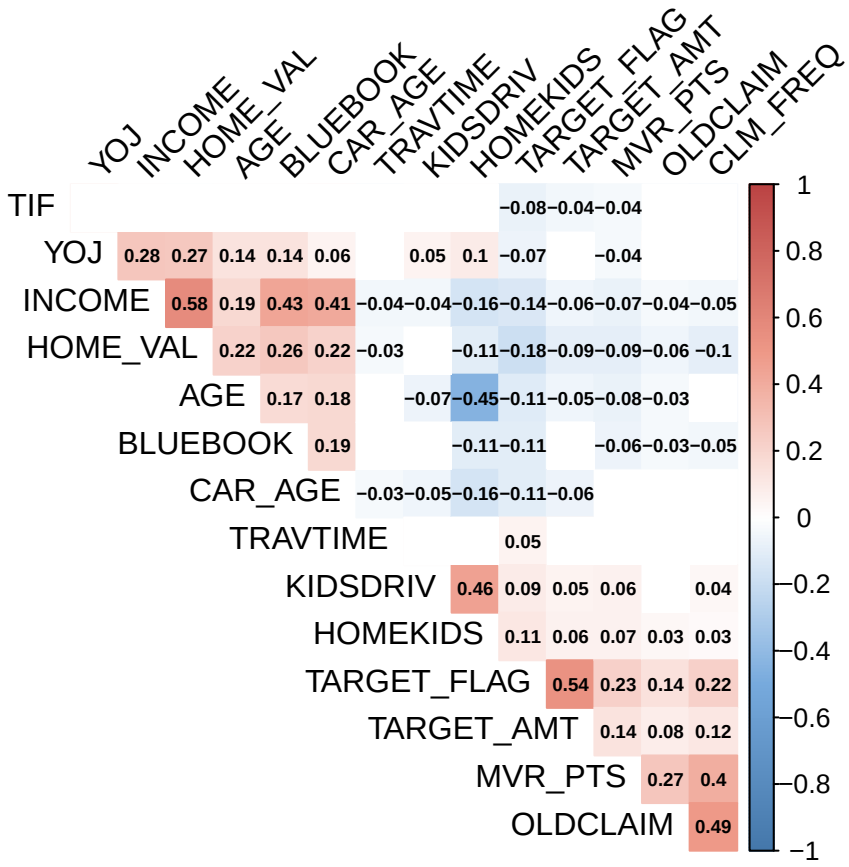
```
# Create a function to remove "$", "," and convert to numeric
clean_money = function(in_col) {
  # this function accepts a currency column and removes any occurrence of "$" and "," and converts it t
  remove_dollar_sign = gsub("\\$", "", in_col)
  remove_comma = gsub(",", "", remove_dollar_sign)
  out_col <- as.numeric(remove_comma)
  return(out_col)
}

# Create a function to remove "z_" and "<"
remove_z = function(in_col){
  # this function accepts a column and removes any occurrence of the strings "z_" or "<" from the colum
  rem_z <- gsub("z_", "", in_col)
  out_col <- gsub("<", "", rem_z)
  return (out_col)
}

# apply the cleaning functions to the applicable columns in the training dataset
training <- training %>% mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"), clean_money) %>%
  mutate_at(c("EDUCATION","JOB","CAR_TYPE","URBANICITY", "MSTATUS"), remove_z)

head(training, 10)
```

```
## # A tibble: 10 x 26
```

```
##    INDEX TARGET_FLAG TARGET_AMT KIDSDRIV   AGE HOMEKIDS   YOJ INCOME PARENT1
##    <dbl>       <dbl>      <dbl>    <dbl> <dbl>    <dbl> <dbl>  <dbl> <chr>
##  1     1           0          0        0    60        0    11  67349 No
##  2     2           0          0        0    43        0    11  91449 No
##  3     4           0          0        0    35        1    10  16039 No
##  4     5           0          0        0    51        0    14     NA No
##  5     6           0          0        0    50        0    NA 114986 No
##  6     7           1       2946        0    34        1    12 125301 Yes
##  7     8           0          0        0    54        0    NA  18755 No
##  8    11           1       4021        1    37        2    NA 107961 No
##  9    12           1       2501        0    34        0    10  62978 No
## 10    13           0          0        0    50        0     7 106952 No
## # ... with 17 more variables: HOME_VAL <dbl>, MSTATUS <chr>, SEX <chr>,
## #   EDUCATION <chr>, JOB <chr>, TRAVTIME <dbl>, CAR_USE <chr>, BLUEBOOK <dbl>,
## #   TIF <dbl>, CAR_TYPE <chr>, RED_CAR <chr>, OLDCLAIM <dbl>, CLM_FREQ <dbl>,
## #   REVOKED <chr>, MVR_PTS <dbl>, CAR_AGE <dbl>, URBANICITY <chr>
```

### Correlation

```r
# select only numeric columns excluding the INDEX column
corr_data <- select_if(training, is.numeric) %>% dplyr::select(-INDEX)
# get the correlation
training_corr = corr_data %>% cor(corr_data, use = "na.or.complete" )
# define the color pallete to use in the correlation plot
col <- colorRampPalette(c("#4477AA", "#77AADD", "#FFFFFF", "#EE9988", "#BB4444"))
# get the matrix of p_values using the rcorr function from the Hmisc package
p_mat <- rcorr(as.matrix(corr_data))$P
# correlation plot
corrplot(training_corr, method="color", col=col(200),
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         number.cex = 0.6,
         tl.col="black", tl.srt=45, #Text label color and rotation
         # Combine with significance
         p.mat = p_mat, sig.level = 0.01, insig = "blank",
         # hide correlation coefficient on the principal diagonal
         diag=FALSE
         )
```




**2. Data Preparation**

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.

c. Transform data by putting it into buckets

d. Mathematical transforms such as log or square root (or use Box-Cox)

e. Combine variables (such as ratios or adding or multiplying) to create new variables

To start, we want to address the missing data. We see there is missing data from 6 fields. 3 job-related fields: Job, YOJ, Income, and 3 others: Age, Car Age, and Home Value. We can replace 'Job' with None, and the Income and YOJ with 0 - it's plausible there are unemployed folks in this dataset. For the second 3, we can use the mean.

We also want to remove Target_AMT, because it is only showing values where TARGET_FLAG = 1, and INDEX because we won't use it.

```
training_clean = training
training <- subset(training, select = -c(INDEX, TARGET_AMT))
training2 <- subset(training_clean, select = -c(INDEX))


training$JOB[is.na(training$JOB)] <- 'None'
training$INCOME[is.na(training$INCOME)] <- 0
```

```
training$YOJ[is.na(training$YOJ)] <- 0

training$HOME_VAL[is.na(training$HOME_VAL)] <- mean(training$HOME_VAL,na.rm=TRUE)
training$AGE[is.na(training$AGE)] <- mean(training$AGE,na.rm=TRUE)
training$CAR_AGE[is.na(training$CAR_AGE)] <- mean(training$CAR_AGE,na.rm=TRUE)


training2$JOB[is.na(training2$JOB)] <- 'None'
training2$INCOME[is.na(training2$INCOME)] <- 0
training2$YOJ[is.na(training2$YOJ)] <- 0

training2$HOME_VAL[is.na(training2$HOME_VAL)] <- mean(training2$HOME_VAL,na.rm=TRUE)
training2$AGE[is.na(training2$AGE)] <- mean(training2$AGE,na.rm=TRUE)
training2$CAR_AGE[is.na(training2$CAR_AGE)] <- mean(training2$CAR_AGE,na.rm=TRUE)

sum(is.na(training))
```

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

```
sum(is.na(training2))
```

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

We don't see evidence of collinearity in the correlation plot, so we can move on to creating a few transformations. We'll create Claims/TIF, Claims/Car Age, and TIF/Car Age.

```
training$claims_tif <- training$CLM_FREQ / training$TIF
training$claims_age <- training$CLM_FREQ / training$CAR_AGE
training$tif_age <- training$TIF / training$CAR_AGE
```

Finally, we'll split the dataset to get ready for model development:

```
set.seed(315)

split <- sample.split(training$TARGET_FLAG, SplitRatio = 0.8)

split2 <- sample.split(training2$TARGET_AMT, SplitRatio = 0.8)

training_set <- subset(training, split == TRUE)
test_set <- subset(training, split == FALSE)


training_set2 <- subset(training2, split2 == TRUE)
test_set2 <- subset(training2, split2 == FALSE)
```

### Logistic Models

### 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.

First, let's build our logistic regression models. To start, we build a version using all fields. AIC is quite high, with several fields with low p-values.

```
logit_1 <- glm(TARGET_FLAG~., family = binomial,data = training_set)

summary(logit_1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4640  -0.7129  -0.3942   0.6161   3.1006
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.170e+00  3.247e-01  -9.762  < 2e-16 ***
## KIDSDRIV                      3.093e-01  6.836e-02   4.525 6.03e-06 ***
## AGE                          -2.334e-03  4.488e-03  -0.520 0.603053
## HOMEKIDS                      6.938e-02  4.100e-02   1.692 0.090595 .
## YOJ                          -1.010e-02  7.889e-03  -1.280 0.200596
## INCOME                       -2.217e-06  1.097e-06  -2.020 0.043367 *
## PARENT1Yes                    4.642e-01  1.229e-01   3.776 0.000159 ***
## HOME_VAL                     -1.402e-06  3.801e-07  -3.690 0.000225 ***
## MSTATUSYes                   -4.081e-01  9.403e-02  -4.340 1.43e-05 ***
## SEXM                          1.581e-01  1.259e-01   1.255 0.209421
## EDUCATIONHigh School          3.461e-01  9.810e-02   3.528 0.000419 ***
## EDUCATIONMasters              5.742e-02  1.606e-01   0.358 0.720656
## EDUCATIONPhD                  1.185e-01  2.030e-01   0.584 0.559483
## JOBClerical                   1.038e-01  1.197e-01   0.867 0.385928
## JOBDoctor                    -8.599e-01  3.252e-01  -2.644 0.008181 **
## JOBHome Maker                -7.925e-02  1.652e-01  -0.480 0.631467
## JOBLawyer                    -3.247e-01  2.118e-01  -1.533 0.125299
## JOBManager                   -9.883e-01  1.554e-01  -6.361 2.01e-10 ***
## JOBNone                      -3.901e-01  2.092e-01  -1.864 0.062278 .
## JOBProfessional              -1.037e-01  1.298e-01  -0.799 0.424224
## JOBStudent                    5.156e-02  1.410e-01   0.366 0.714671
## TRAVTIME                      1.548e-02  2.129e-03   7.272 3.54e-13 ***
## CAR_USEPrivate               -7.233e-01  9.719e-02  -7.442 9.91e-14 ***
## BLUEBOOK                     -1.702e-05  5.899e-06  -2.885 0.003915 **
## TIF                          -6.728e-02  1.175e-02  -5.729 1.01e-08 ***
## CAR_TYPEPanel Truck           5.387e-01  1.798e-01   2.996 0.002732 **
## CAR_TYPEPickup                6.609e-01  1.127e-01   5.864 4.52e-09 ***
## CAR_TYPESports Car            1.092e+00  1.461e-01   7.470 8.00e-14 ***
## CAR_TYPESUV                   7.932e-01  1.261e-01   6.288 3.22e-10 ***
```

```
## CAR_TYPEVan                     5.861e-01  1.411e-01   4.154 3.26e-05 ***
## RED_CARyes                      1.870e-02  9.689e-02   0.193 0.846934
## OLDCLAIM                       -1.536e-05  4.391e-06  -3.499 0.000467 ***
## CLM_FREQ                        2.278e-01  4.779e-02   4.767 1.87e-06 ***
## REVOKEDYes                      9.116e-01  1.023e-01   8.914  < 2e-16 ***
## MVR_PTS                         1.253e-01  1.529e-02   8.194 2.52e-16 ***
## CAR_AGE                         4.992e-04  1.077e-02   0.046 0.963030
## URBANICITYHighly Urban/ Urban   2.378e+00  1.253e-01  18.978  < 2e-16 ***
## claims_tif                     -2.713e-02  6.118e-02  -0.443 0.657454
## claims_age                     -7.073e-02  6.575e-02  -1.076 0.282058
## tif_age                         4.613e-03  1.800e-02   0.256 0.797678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7533.1  on 6527  degrees of freedom
## Residual deviance: 5805.5  on 6488  degrees of freedom
## AIC: 5885.5
##
## Number of Fisher Scoring iterations: 5
```

Let's create a second model using backward selection. This approach leaves us with KIDSDRIV, HOMEKIDS, INCOME, PARENT1, HOME_VAL, MSTATUS, EDUCATION, JOB, TRAVTIME, CAR_USE, BLUEBOOK, TIF, CAR_TYPE, OLDCLAIM, CLM_FREQ, REVOKED, MVR_PTS, and URBANICITY.

```
logit_2 <- step(logit_1, trace = FALSE) # backward selection (if you don't specify anything)
```

```
summary(logit_2)
```

```
##
## Call:
## glm(formula = 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, family = binomial, data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5355  -0.7131  -0.3947   0.6133   3.1060
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.171e+00  2.217e-01 -14.304  < 2e-16 ***
## KIDSDRIV                        3.039e-01  6.711e-02   4.528 5.95e-06 ***
## HOMEKIDS                        6.714e-02  3.759e-02   1.786 0.074082 .
## INCOME                         -2.306e-06  1.092e-06  -2.113 0.034610 *
## PARENT1Yes                      4.708e-01  1.222e-01   3.853 0.000117 ***
## HOME_VAL                       -1.430e-06  3.782e-07  -3.780 0.000157 ***
## MSTATUSYes                     -4.205e-01  9.366e-02  -4.490 7.14e-06 ***
## EDUCATIONHigh School            3.234e-01  9.034e-02   3.579 0.000345 ***
## EDUCATIONMasters                6.153e-02  1.529e-01   0.402 0.687363
```

```
## EDUCATIONPhD                    1.166e-01  1.971e-01   0.591 0.554310
## JOBClerical                     1.050e-01  1.195e-01   0.879 0.379440
## JOBDoctor                      -8.478e-01  3.242e-01  -2.615 0.008913 **
## JOBHome Maker                  -5.808e-02  1.582e-01  -0.367 0.713539
## JOBLawyer                      -3.259e-01  2.111e-01  -1.544 0.122608
## JOBManager                     -9.927e-01  1.551e-01  -6.402 1.54e-10 ***
## JOBNone                        -3.859e-01  2.089e-01  -1.847 0.064685 .
## JOBProfessional                -1.054e-01  1.295e-01  -0.814 0.415707
## JOBStudent                      9.260e-02  1.361e-01   0.680 0.496327
## TRAVTIME                        1.547e-02  2.125e-03   7.283 3.27e-13 ***
## CAR_USEPrivate                 -7.249e-01  9.697e-02  -7.476 7.69e-14 ***
## BLUEBOOK                       -2.132e-05  5.264e-06  -4.050 5.13e-05 ***
## TIF                            -6.374e-02  8.344e-03  -7.639 2.19e-14 ***
## CAR_TYPEPanel Truck             6.395e-01  1.672e-01   3.826 0.000130 ***
## CAR_TYPEPickup                  6.521e-01  1.124e-01   5.800 6.65e-09 ***
## CAR_TYPESports Car              9.667e-01  1.198e-01   8.066 7.26e-16 ***
## CAR_TYPESUV                     6.689e-01  9.691e-02   6.902 5.13e-12 ***
## CAR_TYPEVan                     6.443e-01  1.363e-01   4.727 2.28e-06 ***
## OLDCLAIM                       -1.547e-05  4.390e-06  -3.524 0.000424 ***
## CLM_FREQ                        1.934e-01  3.221e-02   6.005 1.91e-09 ***
## REVOKEDYes                      9.140e-01  1.020e-01   8.957  < 2e-16 ***
## MVR_PTS                         1.256e-01  1.526e-02   8.232  < 2e-16 ***
## URBANICITYHighly Urban/ Urban   2.376e+00  1.253e-01  18.965  < 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: 7533.1  on 6527  degrees of freedom
## Residual deviance: 5811.1  on 6496  degrees of freedom
## AIC: 5875.1
##
## Number of Fisher Scoring iterations: 5
```

One potentially counterintuitive coefficient here is OLDCLAIM - which suggests there is a negative relationship between the $ of a claim and the likelihood to crash. This could be to do with the car value, however.

Lastly, we can try a version that just focuses on the proportional transformation fields. This results in a much higher AIC value, though both `claims_` fields had low p-values.

```
logit_3 <- glm(TARGET_FLAG ~ claims_tif + claims_age + tif_age, family = binomial,data = training_set)
summary(logit_3)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ claims_tif + claims_age + tif_age,
##     family = binomial, data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1889  -0.7055  -0.7045   1.1309   1.7789
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.26361    0.03821 -33.069  < 2e-16 ***
## claims_tif   0.37622    0.03972   9.471  < 2e-16 ***
## claims_age   0.33749    0.04795   7.038 1.95e-12 ***
## tif_age     -0.00492    0.01147  -0.429    0.668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7533.1  on 6527  degrees of freedom
## Residual deviance: 7290.9  on 6524  degrees of freedom
## AIC: 7298.9
##
## Number of Fisher Scoring iterations: 4
```

### 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.

We will evaluate each model as they were created, so the first one we will predict with is the logistic model with all fields.

```
test_clean = test_set


test_set[ ,"probability.all"] <- predict(logit_1, test_clean, type="response")
test_set[ ,"class.all"] <- ifelse(test_set$probability.all < 0.5, 0, 1)
cm1 <- confusionMatrix(as.factor(test_set$class.all), as.factor(test_set$TARGET_FLAG), positive = "1")
cm1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1094  263
##          1  105  168
##
##                Accuracy : 0.7742
##                  95% CI : (0.7531, 0.7943)
##     No Information Rate : 0.7356
##     P-Value [Acc > NIR] : 0.0001823
##
##                   Kappa : 0.3424
##
##  Mcnemar's Test P-Value : 2.741e-16
```

```
##
##             Sensitivity : 0.3898
##             Specificity : 0.9124
##          Pos Pred Value : 0.6154
##          Neg Pred Value : 0.8062
##              Prevalence : 0.2644
##          Detection Rate : 0.1031
##    Detection Prevalence : 0.1675
##       Balanced Accuracy : 0.6511
##
##        'Positive' Class : 1
##
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## [1] "The model using all of the columns with no changes had the following metrics, Accuracy: 0.7742
```

The next is the Logistic model using Backwards Selection.

```
test_set[ ,"probability.back"] <- predict(logit_2, test_clean, type="response")
test_set[ ,"class.back"] <- ifelse(test_set$probability.back < 0.5, 0, 1)
cm1 <- confusionMatrix(as.factor(test_set$class.back), as.factor(test_set$TARGET_FLAG), positive = "1")
cm1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1102  262
##          1  100  169
##
##                Accuracy : 0.7783
##                  95% CI : (0.7574, 0.7983)
##     No Information Rate : 0.7361
##     P-Value [Acc > NIR] : 4.564e-05
##
##                   Kappa : 0.3513
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3921
##             Specificity : 0.9168
##          Pos Pred Value : 0.6283
##          Neg Pred Value : 0.8079
##              Prevalence : 0.2639
##          Detection Rate : 0.1035
##    Detection Prevalence : 0.1647
##       Balanced Accuracy : 0.6545
##
##        'Positive' Class : 1
##
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## [1] "The model using all of the columns with no changes had the following metrics, Accuracy: 0.7783
```

The final is the Logistic model using transformed fields Selection.

```
test_set[ ,"probability.tr"] <- predict(logit_3, test_clean, type="response")
test_set[ ,"class.tr"] <- ifelse(test_set$probability.tr < 0.5, 0, 1)
cm1 <- confusionMatrix(as.factor(test_set$class.tr), as.factor(test_set$TARGET_FLAG), positive = "1")
cm1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1160  389
##          1   39   42
##
##                Accuracy : 0.7374
##                  95% CI : (0.7153, 0.7586)
##     No Information Rate : 0.7356
##     P-Value [Acc > NIR] : 0.4459
##
##                   Kappa : 0.0877
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.09745
##             Specificity : 0.96747
##          Pos Pred Value : 0.51852
##          Neg Pred Value : 0.74887
##              Prevalence : 0.26442
##          Detection Rate : 0.02577
##    Detection Prevalence : 0.04969
##       Balanced Accuracy : 0.53246
##
##        'Positive' Class : 1
##

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## [1] "The model using all of the columns with no changes had the following metrics, Accuracy: 0.7374
```

We can see in these models that the most accurate was the backward selection model. We will use that on the test data after we evaluate the Linear Models.

**Linear Models**

**3. Build Models**

Then we can build our multiple linear regression models. Again, we'll first start with a model using all fields. The Adjusted R-Squared value is quite low at 7%.

```r
mlm_1 <- lm(TARGET_AMT~.,data = training_set2)

summary(mlm_1)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = training_set2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6364   -621    -87    278 101011
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -6.223e+02  4.823e+02  -1.290   0.1970
## TARGET_FLAG                     5.719e+03  1.286e+02  44.455  < 2e-16 ***
## KIDSDRIV                       -3.842e+01  1.148e+02  -0.335   0.7379
## AGE                             7.332e+00  7.187e+00   1.020   0.3077
## HOMEKIDS                        5.539e+01  6.610e+01   0.838   0.4021
## YOJ                             7.711e-01  1.252e+01   0.062   0.9509
## INCOME                         -2.301e-03  1.693e-03  -1.359   0.1741
## PARENT1Yes                      1.599e+02  2.063e+02   0.775   0.4381
## HOME_VAL                        4.419e-04  6.029e-04   0.733   0.4636
## MSTATUSYes                     -1.971e+02  1.484e+02  -1.328   0.1841
## SEXM                            3.477e+02  1.908e+02   1.823   0.0684 .
## EDUCATIONHigh School           -1.916e+02  1.568e+02  -1.222   0.2218
## EDUCATIONMasters                1.930e+02  2.308e+02   0.836   0.4031
## EDUCATIONPhD                    4.622e+02  2.990e+02   1.546   0.1222
## JOBClerical                    -8.006e+01  1.952e+02  -0.410   0.6818
## JOBDoctor                      -4.793e+02  4.561e+02  -1.051   0.2934
## JOBHome Maker                  -2.044e+02  2.665e+02  -0.767   0.4431
## JOBLawyer                      -1.751e+01  3.185e+02  -0.055   0.9562
## JOBManager                     -2.631e+02  2.394e+02  -1.099   0.2718
## JOBNone                        -9.987e+01  3.302e+02  -0.302   0.7623
## JOBProfessional                 9.439e+01  2.136e+02   0.442   0.6586
## JOBStudent                     -2.557e+02  2.333e+02  -1.096   0.2732
## TRAVTIME                        9.168e-01  3.347e+00   0.274   0.7841
## CAR_USEPrivate                 -7.257e+01  1.617e+02  -0.449   0.6536
## BLUEBOOK                        3.536e-02  8.909e-03   3.969 7.29e-05 ***
## TIF                            -3.171e+00  1.262e+01  -0.251   0.8015
## CAR_TYPEPanel Truck            -5.497e+01  2.831e+02  -0.194   0.8461
## CAR_TYPEPickup                 -1.481e+01  1.758e+02  -0.084   0.9329
## CAR_TYPESports Car              2.590e+02  2.246e+02   1.154   0.2487
## CAR_TYPESUV                     1.906e+02  1.862e+02   1.024   0.3061
## CAR_TYPEVan                     1.154e+02  2.187e+02   0.528   0.5976
## RED_CARyes                     -3.176e+01  1.532e+02  -0.207   0.8357
## OLDCLAIM                        3.910e-03  7.577e-03   0.516   0.6058
## CLM_FREQ                       -4.679e+01  5.635e+01  -0.830   0.4064
## REVOKEDYes                     -3.716e+02  1.772e+02  -2.097   0.0361 *
## MVR_PTS                         6.168e+01  2.615e+01   2.359   0.0184 *
## CAR_AGE                        -3.120e+01  1.308e+01  -2.385   0.0171 *
## URBANICITYHighly Urban/ Urban -3.573e+01  1.505e+02  -0.237   0.8123
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4300 on 6906 degrees of freedom
## Multiple R-squared:  0.2799, Adjusted R-squared:  0.2761
## F-statistic: 72.56 on 37 and 6906 DF,  p-value: < 2.2e-16
```

For our final model, we'll repeat our backwards selection process on the Multiple Linear Regression model. Interestingly, the model lands on near identical fields to the logistic regression, with the addition of SEX and YOJ. The inclusion of YOJ is somewhat surprising - relative to other fields, it seems much less likely to influence outcome, though higher YOJ leads to less likelihood to crash.

```
mlm_2 <- step(mlm_1, trace = FALSE)
```

```
summary(mlm_2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ TARGET_FLAG + PARENT1 + SEX + BLUEBOOK +
##     REVOKED + MVR_PTS + CAR_AGE, data = training_set2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6367   -552    -68    239 101265
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.304e+02  1.478e+02  -4.266 2.01e-05 ***
## TARGET_FLAG  5.722e+03  1.179e+02  48.537  < 2e-16 ***
## PARENT1Yes   2.536e+02  1.537e+02   1.650   0.0989 .
## SEXM         2.291e+02  1.039e+02   2.206   0.0274 *
## BLUEBOOK     3.539e-02  6.304e-03   5.615 2.04e-08 ***
## REVOKEDYes  -3.253e+02  1.565e+02  -2.079   0.0377 *
## MVR_PTS      5.771e+01  2.414e+01   2.391   0.0168 *
## CAR_AGE     -1.502e+01  9.574e+00  -1.569   0.1167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4296 on 6936 degrees of freedom
## Multiple R-squared:  0.2781, Adjusted R-squared:  0.2774
## F-statistic: 381.8 on 7 and 6936 DF,  p-value: < 2.2e-16
```

### 4. Select Models

The first is the linear model with everything selected.

We will begin by checking for multicollinearity:

```
dwtest(mlm_1)
```

```
##
##  Durbin-Watson test
##
## data:  mlm_1
## DW = 1.9883, p-value = 0.3134
## alternative hypothesis: true autocorrelation is greater than 0
```

The null hypothesis is that there does not exist autocorrelation (multicollinearity). Since the p-value is large, we fail to reject the null hypothesis.

Mean Square Error and RMSE

```
##   MODEL      MSE     RMSE R.SQUARED ADJ.R.SQUARED    value numdf dendf
## 1 mlm_1 18487638 4299.725 0.2799215     0.2760636 72.55733    37  6906
```

The Mean Squared Error is the square of the RMSE. The benefit of using the RMSE is that it is expressed in the same units as the target variable. For these models, we see that standard error of the mean (RMSE) is 3926 off, signifying that the model needs a bit of work. Part of the issue is that we need to make sure it is not calculating amount unless there is a crash.

$$R^2$$

represents the percent change in

$$Y$$

explained by the predictor variables with

$$R^2$$

1 indicating a perfect model, since ours is .29, it does need some work. Adjusted

$$R^2$$

is more appropriate for this model since it has multiple variables. It incorporates a penalty to account for the decrease in degrees of freedom (from additional variables). The penalty did not improve the evaluation in this case, it is slightly lower.

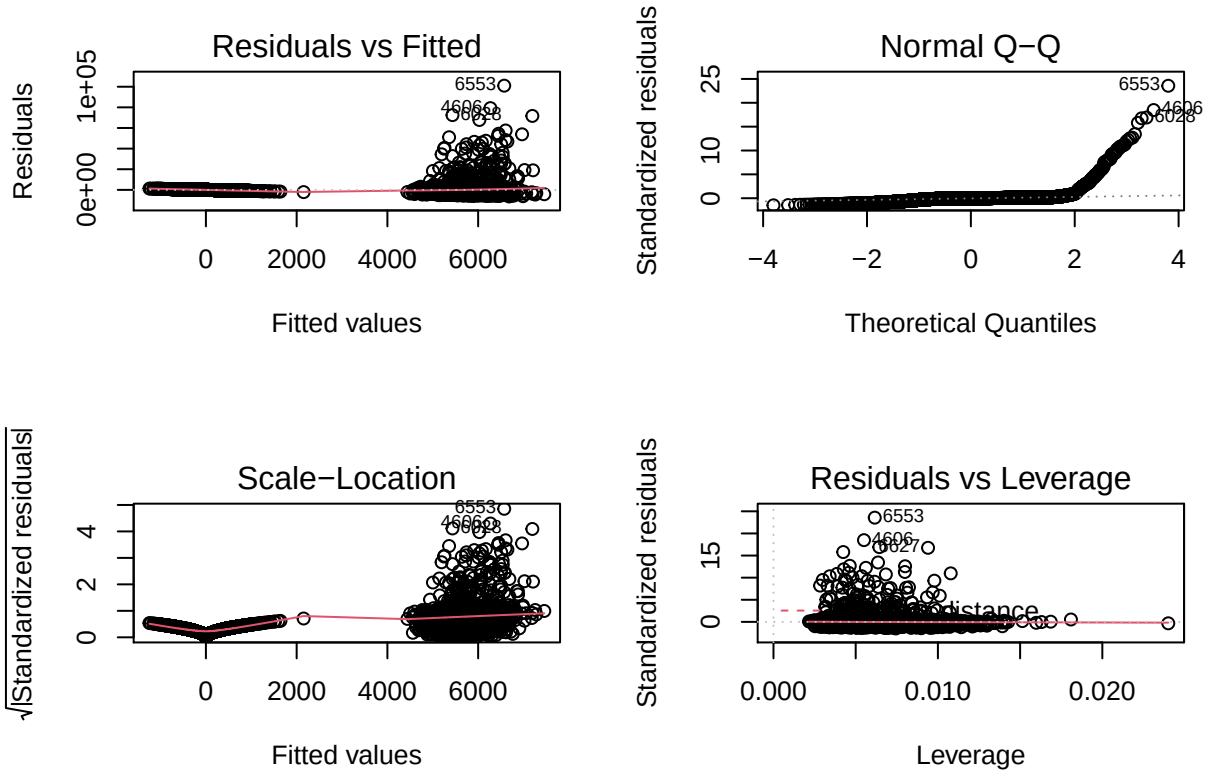
The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one does not. At an

$$\alpha = 2.2e - 16$$

the F-statistic indicates that the model fits the data better than the intercept-only model.

Now let's take a look at the Residuals

```
par(mfrow = c(2,2))
plot(mlm_1)
```




This shows that there is very little error at the lower end of amount and it is due to us filtering for 0s in the test set (based on backward step model from the logistic section). Now let's compare the model with the test data.

```
test_set[ ,"msm1.ALL"] <- predict(mlm_1, newdata = test_clean)

test_set$msm1.ALL[test_set$class.back == 0 ] = 0

paste0("The Root Sqaure Mean Error for model one is: ", round(sqrt(mean((test_set$TARGET_AMT - test_set$
```

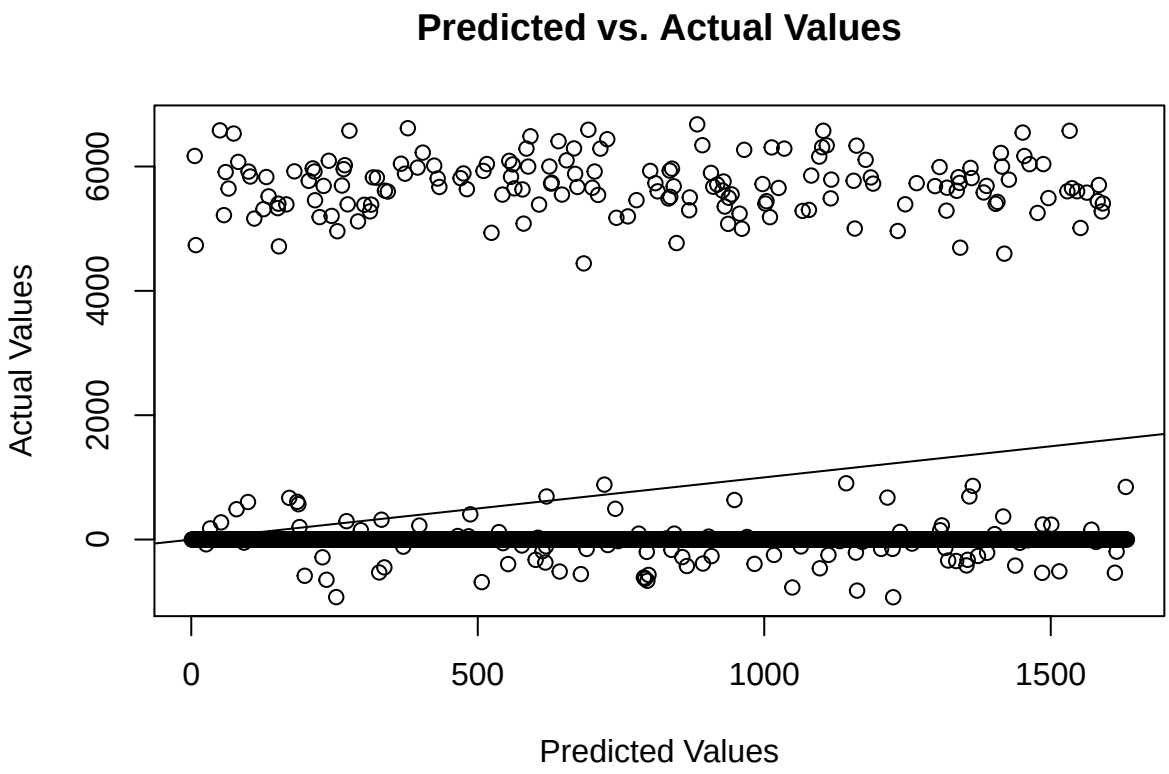
```
## Warning: Unknown or uninitialised column: 'TARGET_AMT'.
```

```
## [1] "The Root Sqaure Mean Error for model one is: NaN"
```

```
plot(x=test_set$msm1.ALL, y= test_set$TARGET_AMT,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Predicted vs. Actual Values')
```

```
## Warning: Unknown or uninitialised column: 'TARGET_AMT'.
```

```
abline(a=0, b=1)
```




Again it shows there is no error at the low end of amount but there is some variation at the higher end. Now let's compare the same with model two

The second and final is the linear model using the backwards selection

We will begin by checking for multicollinearity:

```
dwtest(mlm_2)
```

```
##
##  Durbin-Watson test
##
## data:  mlm_2
## DW = 1.9891, p-value = 0.3241
## alternative hypothesis: true autocorrelation is greater than 0
```

The null hypothesis is that there does not exist autocorrelation (multicollinearity). Since the p-value is large, we fail to reject the null hypothesis.

Mean Square Error and RMSE

```
##   MODEL      MSE     RMSE R.SQUARED ADJ.R.SQUARED   value numdf dendf
## 1 mlm_1 18453570 4295.762 0.2781261     0.2773976 381.761     7  6936
```

The Mean Squared Error is the square of the RMSE. The benefit of using the RMSE is that it is expressed in the same units as the target variable. For these models, we see that standard error of the mean (RMSE)

is 3922, which is slighlt better than model 1, signifying that the model needs a bit of work. Part of the issue is that we need to make sure it is not calculating amount unless there is a crash.

$$R^2$$

represents the percent change in

$$Y$$

explained by the predictor variables with

$$R^2$$

1 indicating a perfect model, since ours is .29, it does need some work. Adjusted

$$R^2$$

is more appropriate for this model since it has multiple variables. It incorporates a penalty to account for the decrease in degrees of freedom (from additional variables). The penalty did not improve the evaluation in this case, it is slightly lower. But, again we see that it better than model 1.

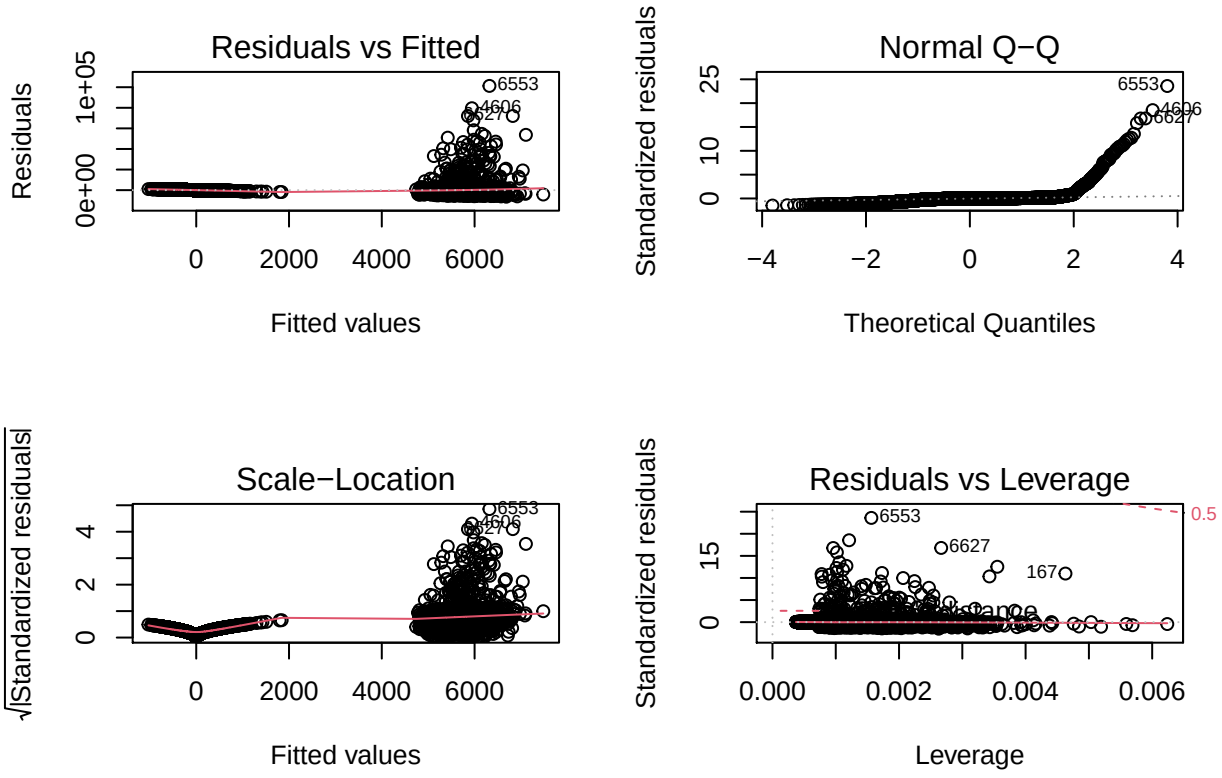
The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one does not. At an

$$\alpha = 2.2e - 16$$

the F-statistic indicates that the model fits the data better than the intercept-only model.

Now let's take a look at the Residuals

```
par(mfrow = c(2,2))
plot(mlm_2)
```




This shows that there is very little error between the residuals and fitted by all accounts. The worry here is that we overfitted the model. So let's compare the model with the test data.

```
test_set[ ,"msm2.ALL"] <- predict(mlm_2, newdata = test_clean)

test_set$msm2.ALL[test_set$class.back == 0 ] = 0

paste0("The Root Sqaure Mean Error for model one is: ", round(sqrt(mean((test_set$TARGET_AMT - test_set$
```

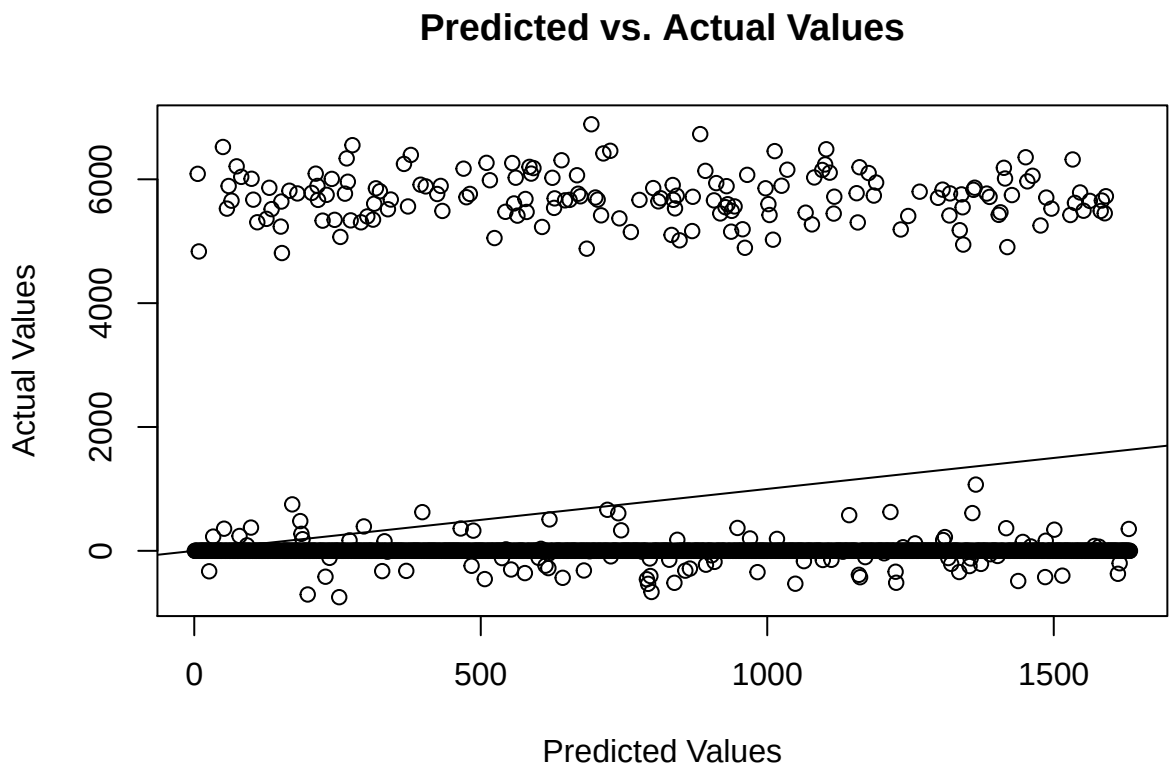
```
## Warning: Unknown or uninitialised column: 'TARGET_AMT'.
```

```
## [1] "The Root Sqaure Mean Error for model one is: NaN"
```

```
plot(x=test_set$msm2.ALL, y= test_set$TARGET_AMT,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Predicted vs. Actual Values')
```

```
## Warning: Unknown or uninitialised column: 'TARGET_AMT'.
```

```
abline(a=0, b=1)
```




The RSME is slightly lower than the model, so we will work with this model, despite there being much of a difference.

**Top Model Evaluation**

**Clean the dataset**

```
# apply the cleaning functions to the applicable columns in the training dataset
evaluation <- evaluation %>% mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"), clean_money) %>%
  mutate_at(c("EDUCATION","JOB","CAR_TYPE","URBANICITY", "MSTATUS"), remove_z)

head(evaluation, 10)
```

```
## # A tibble: 10 x 26
##    INDEX TARGET_FLAG TARGET_AMT KIDSDRIV   AGE HOMEKIDS   YOJ INCOME PARENT1
##    <dbl> <lgl>       <lgl>         <dbl> <dbl>    <dbl> <dbl>  <dbl> <chr>
##  1     3 NA          NA                0    48        0    11  52881 No
##  2     9 NA          NA                1    40        1    11  50815 Yes
##  3    10 NA          NA                0    44        2    12  43486 Yes
##  4    18 NA          NA                0    35        2    NA  21204 Yes
##  5    21 NA          NA                0    59        0    12  87460 No
##  6    30 NA          NA                0    46        0    14     NA No
##  7    31 NA          NA                0    60        0    12  37940 No
##  8    37 NA          NA                0    54        0    12  33212 No
##  9    39 NA          NA                2    36        2    12 130540 Yes
## 10    47 NA          NA                0    50        0     8 167469 No
## # ... with 17 more variables: HOME_VAL <dbl>, MSTATUS <chr>, SEX <chr>,
## #   EDUCATION <chr>, JOB <chr>, TRAVTIME <dbl>, CAR_USE <chr>, BLUEBOOK <dbl>,
## #   TIF <dbl>, CAR_TYPE <chr>, RED_CAR <chr>, OLDCLAIM <dbl>, CLM_FREQ <dbl>,
## #   REVOKED <chr>, MVR_PTS <dbl>, CAR_AGE <dbl>, URBANICITY <chr>
```

```
evaluation <- evaluation %>% dplyr::select(-c(TARGET_AMT, INDEX, TARGET_FLAG))

evaluation$JOB[is.na(evaluation$JOB)] <- 'None'
evaluation$INCOME[is.na(evaluation$INCOME)] <- 0
evaluation$YOJ[is.na(evaluation$YOJ)] <- 0

evaluation$HOME_VAL[is.na(evaluation$HOME_VAL)] <- mean(evaluation$HOME_VAL,na.rm=TRUE)
evaluation$AGE[is.na(evaluation$AGE)] <- mean(evaluation$AGE,na.rm=TRUE)
evaluation$CAR_AGE[is.na(evaluation$CAR_AGE)] <- mean(evaluation$CAR_AGE,na.rm=TRUE)

evaluation$SEX
```

```
##    [1] "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F"
##   [13] "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F"
##   [25] "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F"
##   [37] "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F"
##   [49] "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "z_F"
##   [61] "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"
##   [73] "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"
##   [85] "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"
##   [97] "M"   "z_F" "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "M"   "M"
##  [109] "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F"
##  [121] "z_F" "M"   "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F"
##  [133] "M"   "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"
##  [145] "M"   "z_F" "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"
##  [157] "z_F" "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"   "z_F" "M"   "z_F"
##  [169] "z_F" "M"   "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "M"   "z_F" "z_F"
##  [181] "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F"
##  [193] "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"
```

```
##  [205] "M"   "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"
##  [217] "z_F" "M"   "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "z_F"
##  [229] "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F" "M"   "z_F"
##  [241] "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"
##  [253] "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F"
##  [265] "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"
##  [277] "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F"
##  [289] "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F"
##  [301] "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F"
##  [313] "z_F" "M"   "z_F" "z_F" "M"   "M"   "M"   "M"   "M"   "M"   "M"   "M"
##  [325] "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"
##  [337] "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F"
##  [349] "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F"
##  [361] "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"
##  [373] "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"
##  [385] "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F"
##  [397] "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"
##  [409] "M"   "z_F" "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"
##  [421] "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"
##  [433] "z_F" "M"   "M"   "M"   "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F"
##  [445] "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"
##  [457] "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "M"   "M"   "z_F"
##  [469] "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F"
##  [481] "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F"
##  [493] "M"   "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "z_F" "z_F" "z_F" "M"
##  [505] "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"
##  [517] "M"   "z_F" "M"   "M"   "M"   "z_F" "M"   "M"   "M"   "M"   "M"   "M"
##  [529] "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F" "M"   "z_F"
##  [541] "z_F" "z_F" "z_F" "M"   "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "M"
##  [553] "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"
##  [565] "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F"
##  [577] "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F"
##  [589] "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F"
##  [601] "z_F" "M"   "z_F" "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "M"   "z_F"
##  [613] "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F"
##  [625] "M"   "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"
##  [637] "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "M"   "M"
##  [649] "M"   "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "M"   "M"
##  [661] "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F"
##  [673] "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "z_F"
##  [685] "M"   "z_F" "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "z_F"
##  [697] "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F"
##  [709] "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "z_F"
##  [721] "z_F" "M"   "M"   "M"   "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F"
##  [733] "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"
##  [745] "M"   "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F"
##  [757] "z_F" "M"   "M"   "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F"
##  [769] "M"   "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F"
##  [781] "z_F" "M"   "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "M"
##  [793] "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F" "M"
##  [805] "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"
##  [817] "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "M"   "z_F" "z_F" "M"
##  [829] "M"   "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"
##  [841] "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F"
```

```
##  [853] "M"   "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"   "M"
##  [865] "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F"
##  [877] "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F"
##  [889] "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "M"   "M"
##  [901] "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F"
##  [913] "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "M"   "M"   "M"
##  [925] "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F"
##  [937] "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"
##  [949] "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"
##  [961] "z_F" "M"   "z_F" "M"   "M"   "M"   "z_F" "M"   "M"   "z_F" "z_F" "M"
##  [973] "z_F" "z_F" "M"   "M"   "M"   "z_F" "M"   "M"   "z_F" "M"   "z_F" "z_F"
##  [985] "M"   "M"   "M"   "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "M"   "z_F"
##  [997] "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"
## [1009] "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "M"
## [1021] "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"
## [1033] "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F"
## [1045] "z_F" "z_F" "M"   "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "M"
## [1057] "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"
## [1069] "z_F" "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"
## [1081] "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "M"
## [1093] "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F"
## [1105] "M"   "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"
## [1117] "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "M"   "M"   "M"
## [1129] "M"   "M"   "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"
## [1141] "M"   "z_F" "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F"
## [1153] "z_F" "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F"
## [1165] "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "z_F" "z_F"
## [1177] "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F"
## [1189] "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F"
## [1201] "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "z_F" "M"
## [1213] "M"   "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"
## [1225] "z_F" "z_F" "z_F" "M"   "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F"
## [1237] "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"
## [1249] "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F"
## [1261] "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"
## [1273] "M"   "M"   "M"   "z_F" "M"   "M"   "M"   "M"   "z_F" "M"   "M"   "z_F"
## [1285] "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"   "M"   "M"   "M"
## [1297] "z_F" "M"   "M"   "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F"
## [1309] "M"   "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F"
## [1321] "z_F" "M"   "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F"
## [1333] "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"
## [1345] "M"   "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F"
## [1357] "M"   "z_F" "M"   "M"   "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F"
## [1369] "M"   "z_F" "M"   "M"   "z_F" "M"   "M"   "M"   "M"   "z_F" "M"   "z_F"
## [1381] "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"   "z_F"
## [1393] "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "M"
## [1405] "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"
## [1417] "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F"
## [1429] "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"
## [1441] "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F"
## [1453] "z_F" "M"   "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F"
## [1465] "M"   "z_F" "z_F" "M"   "M"   "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F"
## [1477] "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F"
## [1489] "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "z_F"
```

```
## [1501] "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"   "M"
## [1513] "z_F" "M"   "M"   "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "M"   "z_F"
## [1525] "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F"
## [1537] "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F"
## [1549] "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F"
## [1561] "z_F" "M"   "M"   "z_F" "M"   "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F"
## [1573] "M"   "M"   "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "M"   "z_F" "M"
## [1585] "M"   "M"   "z_F" "z_F" "M"   "M"   "M"   "M"   "z_F" "M"   "M"   "M"
## [1597] "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "M"
## [1609] "z_F" "M"   "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F" "M"   "z_F" "z_F"
## [1621] "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F"
## [1633] "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"
## [1645] "z_F" "M"   "M"   "M"   "z_F" "z_F" "M"   "M"   "M"   "M"   "M"   "M"
## [1657] "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "M"   "M"   "z_F"
## [1669] "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"
## [1681] "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"
## [1693] "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "M"   "M"   "z_F"
## [1705] "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"
## [1717] "M"   "M"   "z_F" "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "M"   "z_F"
## [1729] "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F" "M"   "M"
## [1741] "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F"
## [1753] "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F" "M"   "M"
## [1765] "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"   "M"   "M"   "z_F"
## [1777] "z_F" "z_F" "M"   "M"   "M"   "M"   "M"   "z_F" "M"   "z_F" "M"   "M"
## [1789] "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "M"   "z_F"
## [1801] "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "M"   "z_F" "z_F"
## [1813] "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F"
## [1825] "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"
## [1837] "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"   "M"   "M"   "M"
## [1849] "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F"
## [1861] "M"   "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "z_F"
## [1873] "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"
## [1885] "z_F" "z_F" "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F"
## [1897] "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "M"
## [1909] "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F"
## [1921] "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F"
## [1933] "M"   "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F"
## [1945] "z_F" "z_F" "z_F" "M"   "z_F" "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F"
## [1957] "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F"
## [1969] "z_F" "M"   "M"   "z_F" "z_F" "M"   "z_F" "M"   "z_F" "z_F" "M"   "z_F"
## [1981] "z_F" "M"   "M"   "z_F" "M"   "M"   "z_F" "M"   "z_F" "M"   "M"   "z_F"
## [1993] "M"   "M"   "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F"
## [2005] "M"   "z_F" "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "M"   "M"
## [2017] "M"   "z_F" "M"   "M"   "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"
## [2029] "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F" "z_F"
## [2041] "z_F" "M"   "M"   "M"   "z_F" "M"   "z_F" "z_F" "M"   "M"   "M"   "z_F"
## [2053] "M"   "M"   "M"   "z_F" "M"   "M"   "M"   "z_F" "z_F" "z_F" "z_F" "M"
## [2065] "z_F" "M"   "M"   "z_F" "z_F" "M"   "M"   "z_F" "M"   "M"   "M"   "M"
## [2077] "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "z_F"
## [2089] "z_F" "z_F" "M"   "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"   "z_F"
## [2101] "M"   "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F" "M"
## [2113] "M"   "z_F" "z_F" "z_F" "z_F" "z_F" "M"   "z_F" "z_F" "M"   "z_F" "M"
## [2125] "z_F" "z_F" "M"   "z_F" "z_F" "z_F" "M"   "M"   "z_F" "z_F" "z_F" "M"
## [2137] "z_F" "M"   "z_F" "M"   "z_F"
```

```r
evaluation$SEX <- gsub("z_", "", evaluation$SEX)

sum(is.na(evaluation))
```

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

```r
evaluation$claims_tif <- evaluation$CLM_FREQ / evaluation$TIF
evaluation$claims_age <- evaluation$CLM_FREQ / evaluation$CAR_AGE
evaluation$tif_age <- evaluation$TIF / evaluation$CAR_AGE
```

```r
evaluation$TARGET_perc <- predict(logit_2, evaluation, type="response")
evaluation$TARGET_FLAG <- ifelse(evaluation$TARGET_perc < 0.5, 0, 1)

evaluation <- evaluation %>% dplyr::select(-TARGET_perc)

evaluation$TARGET_AMT <- predict(mlm_2, newdata = evaluation)

evaluation$TARGET_AMT[evaluation$TARGET_FLAG == 0 ] = 0

#write.csv(evaluation,"C:/Users/humme/Downloads/insurance-evaluation-data_final.csv", row.names = FALSE
write.csv(evaluation,"insurance-evaluation-data_final.csv", row.names = FALSE)
```