DATA 621 HW4 WA

Warner Alexis

2024-11-17

Introduction

The objective is to build two models using the provided dataset: a multiple linear regression model to predict the cost of a car crash (a continuous variable) and a binary logistic regression model to predict the probability of a car crash (a binary outcome). Only the given variables, or any new variables derived from them, can be used to build these models. The dataset includes a set of variables relevant to the prediction task, which are described in the following section.

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REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.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

Figure 1: Project Data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(DataExplorer)
library(caret)
```

```
## Loading required package: lattice
```

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(e1071)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(VIM)
```

```
## Loading required package: colorspace

##
## Attaching package: 'colorspace'
```

```
## The following object is masked from 'package:pROC':
##
##     coords

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

Data Exploration

The training dataset consists of 8,161 observations and 26 variables. Several categorical variables require transformation to be compatible with both logistic and linear models. We begin with some fundamental transformations.

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/
```

```
##      INDEX         TARGET_FLAG       TARGET_AMT         KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE            HOMEKIDS           YOJ          INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS              SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   EDUCATION             JOB               TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK              TIF            CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM            CLM_FREQ        REVOKED             MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE         URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
```

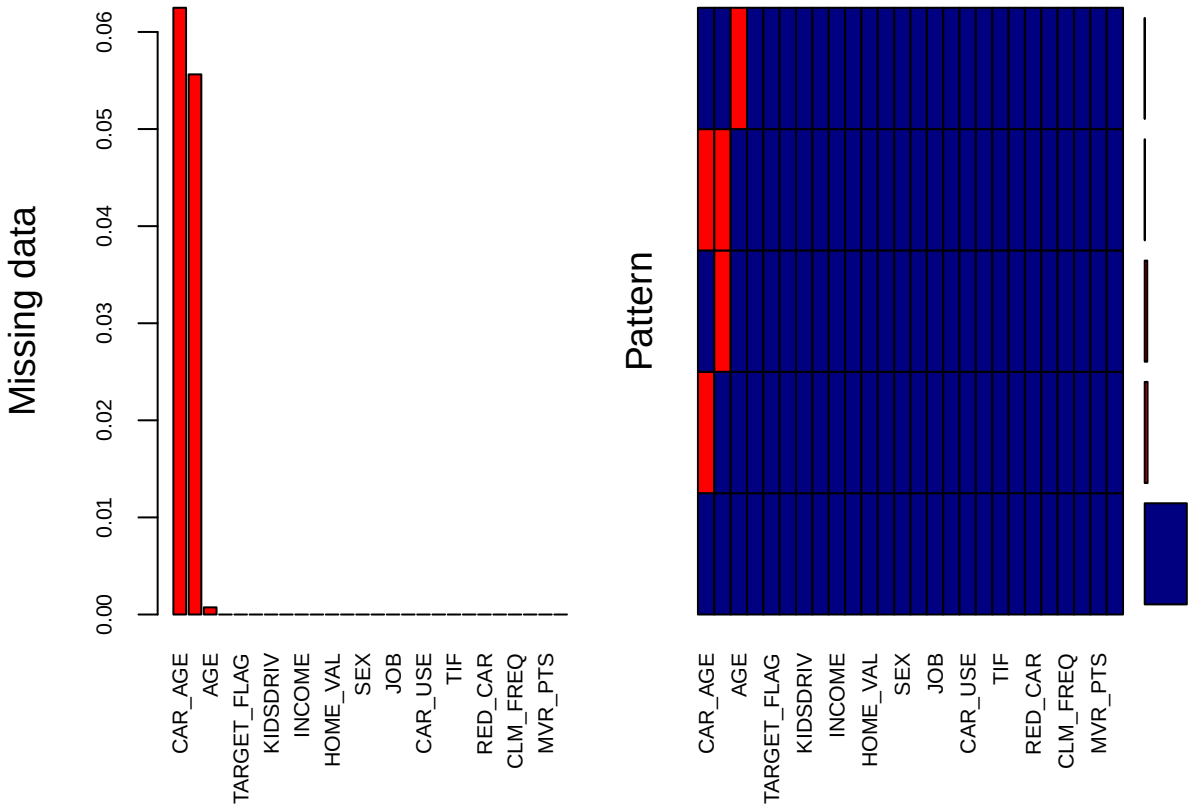
```
##  Max.   :28.000
##  NA's   :510
```

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           6           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##         454           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
##         510           0
```

The first graph displays the distribution of the `TARGET_FLAG` variable, which indicates whether a crash occurred (`1`) or did not occur (`0`). The majority of observations fall into the "No Crash (0)" category, with over 6,000 instances, while a smaller portion represents the "Crash (1)" category. This suggests a significant class imbalance in the dataset. To address this imbalance, techniques such as resampling (oversampling the minority class or undersampling the majority class) or weighted models may be necessary to prevent bias toward the majority class.

The second graph provides an overview of missing data in the dataset. The left plot illustrates the proportion of missing data for each variable, showing that `CAR_AGE` and `AGE` have the highest proportion of missing data (around 6% each), while most other variables have little to no missing values. The right plot visualizes the pattern of missing data, with red sections representing missing values and blue sections indicating observed data. The missing data is primarily concentrated in `CAR_AGE` and `AGE`, while the majority of the dataset is complete. To handle this, imputation strategies such as replacing missing values with the mean or median, or predictive imputation, can be employed. Alternatively, rows with missing data can be excluded if they represent a small percentage of the dataset. Proper handling of missing data is crucial for ensuring the quality and reliability of the modeling process.

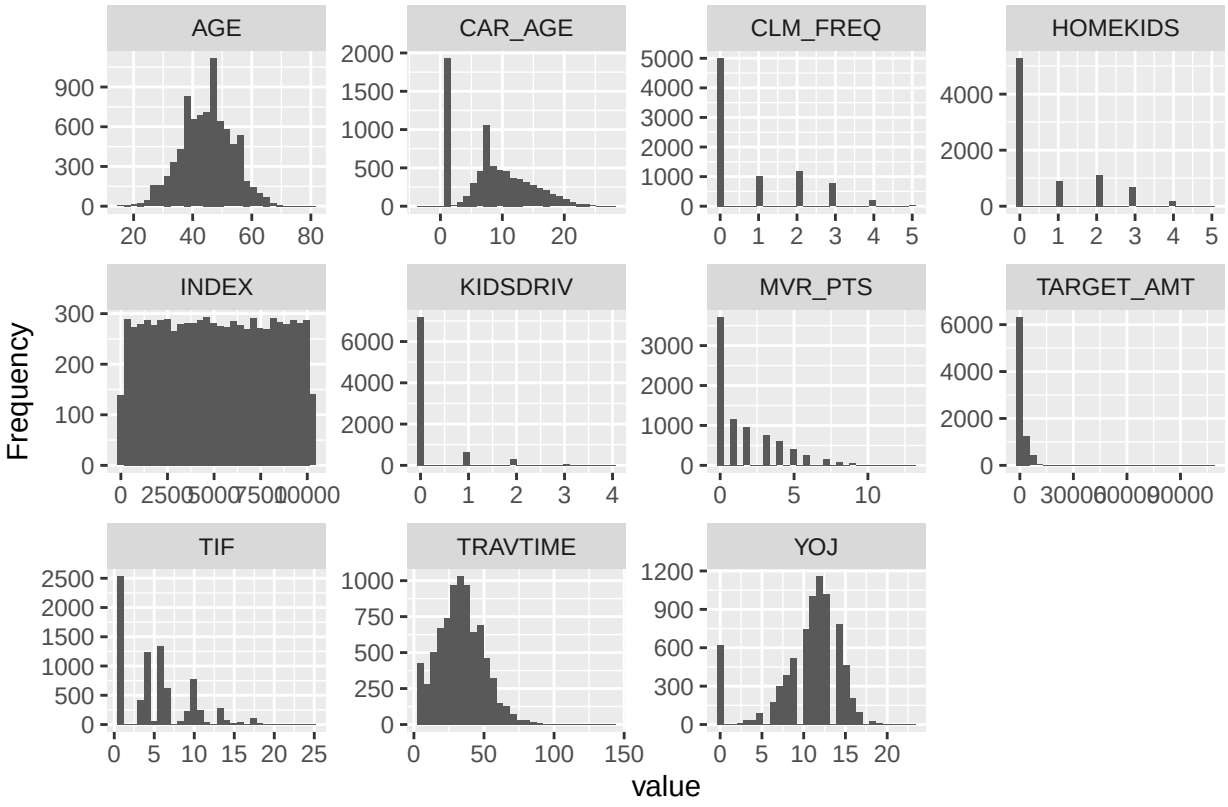
```
training_data <- insurance_training
# Missing Data Visualization
aggr(training_data, col = c('navyblue', 'red'), numbers = TRUE, sortVars = TRUE, labels = names(trainin
     cex.axis = 0.7, gap = 3, ylab = c("Missing data", "Pattern"))
```



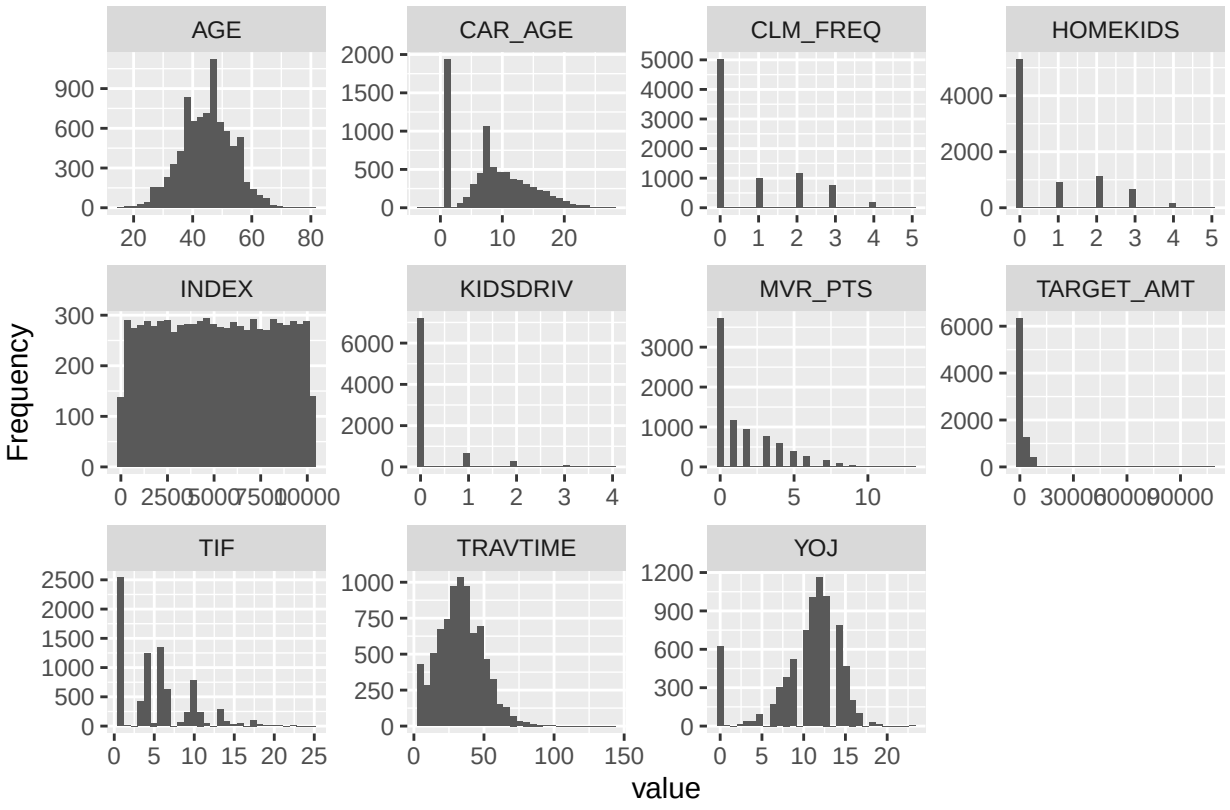

```
##
##  Variables sorted by number of missings:
##     Variable       Count
##      CAR_AGE 0.062492342
##          YOJ 0.055630437
##          AGE 0.000735204
##        INDEX 0.000000000
##  TARGET_FLAG 0.000000000
##   TARGET_AMT 0.000000000
##     KIDSDRIV 0.000000000
##     HOMEKIDS 0.000000000
##       INCOME 0.000000000
##      PARENT1 0.000000000
##     HOME_VAL 0.000000000
##      MSTATUS 0.000000000
##          SEX 0.000000000
##    EDUCATION 0.000000000
##          JOB 0.000000000
##     TRAVTIME 0.000000000
##      CAR_USE 0.000000000
##     BLUEBOOK 0.000000000
##          TIF 0.000000000
##     CAR_TYPE 0.000000000
##      RED_CAR 0.000000000
##     OLDCLAIM 0.000000000
##     CLM_FREQ 0.000000000
##      REVOKED 0.000000000
##      MVR_PTS 0.000000000
```

```
##   URBANICITY 0.000000000
```

```
# Summary statistic
numeric_vars <- training_data %>% select_if(is.numeric)
plot_histogram(training_data)
```



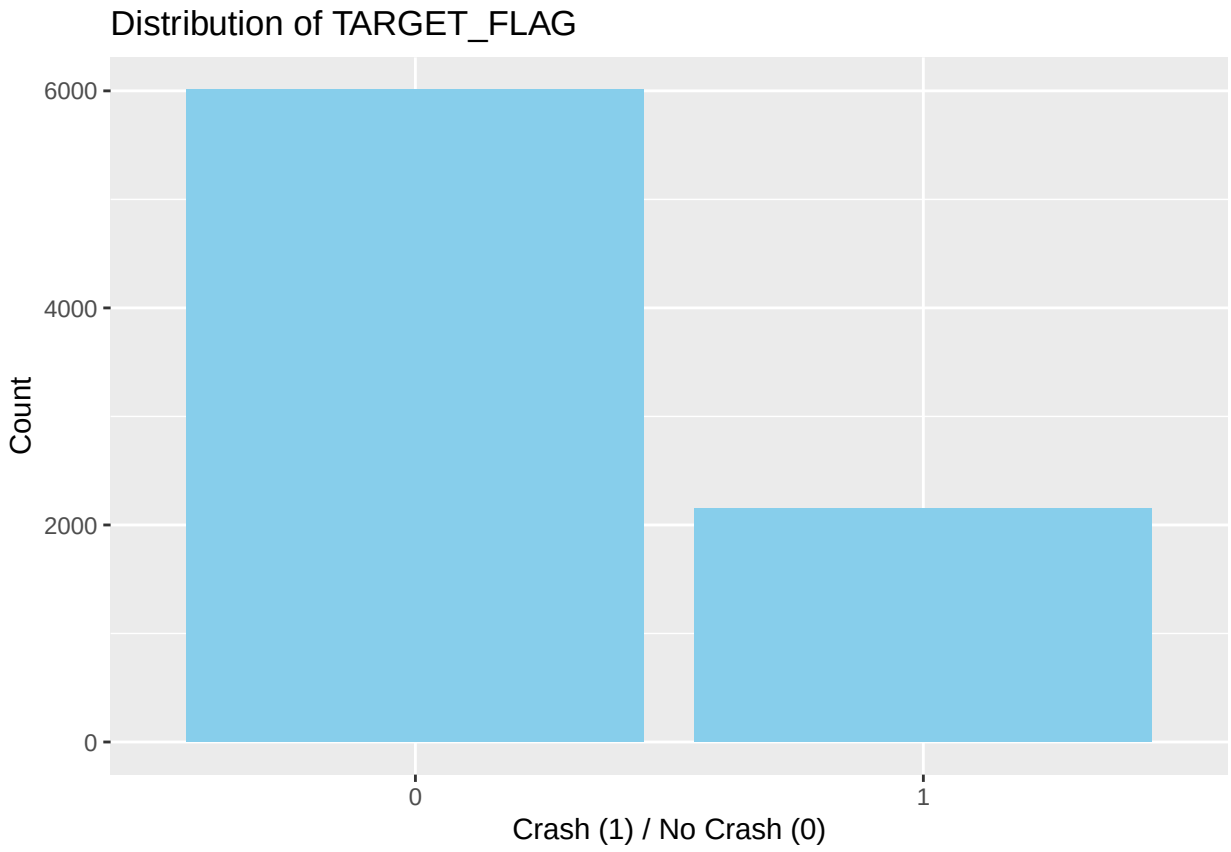

```
DataExplorer::plot_histogram(training_data)
```



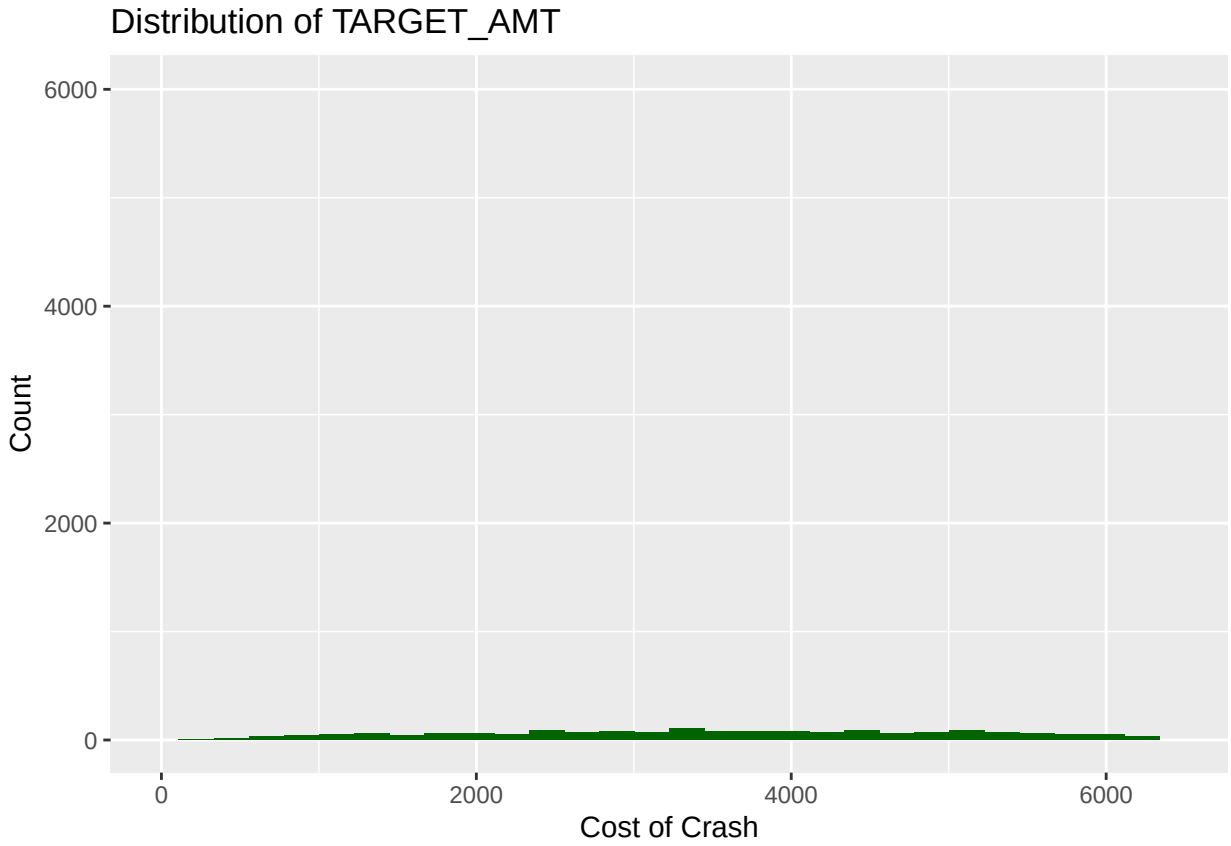

```
library(ggplot2)
# Ananlyze response variables
# Explore TARGET_FLAG
table(training_data$TARGET_FLAG)
```

```
##
##    0    1
## 6008 2153
```

```
ggplot(training_data, aes(x = as.factor(TARGET_FLAG))) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of TARGET_FLAG", x = "Crash (1) / No Crash (0)", y = "Count")
```




```r
# Explore TARGET_AMT
ggplot(training_data, aes(x = TARGET_AMT)) +
  geom_histogram(fill = "darkgreen", bins = 30) +
  labs(title = "Distribution of TARGET_AMT", x = "Cost of Crash", y = "Count") +
  xlim(0, quantile(training_data$TARGET_AMT, 0.95))  # Trim extreme outliers
```

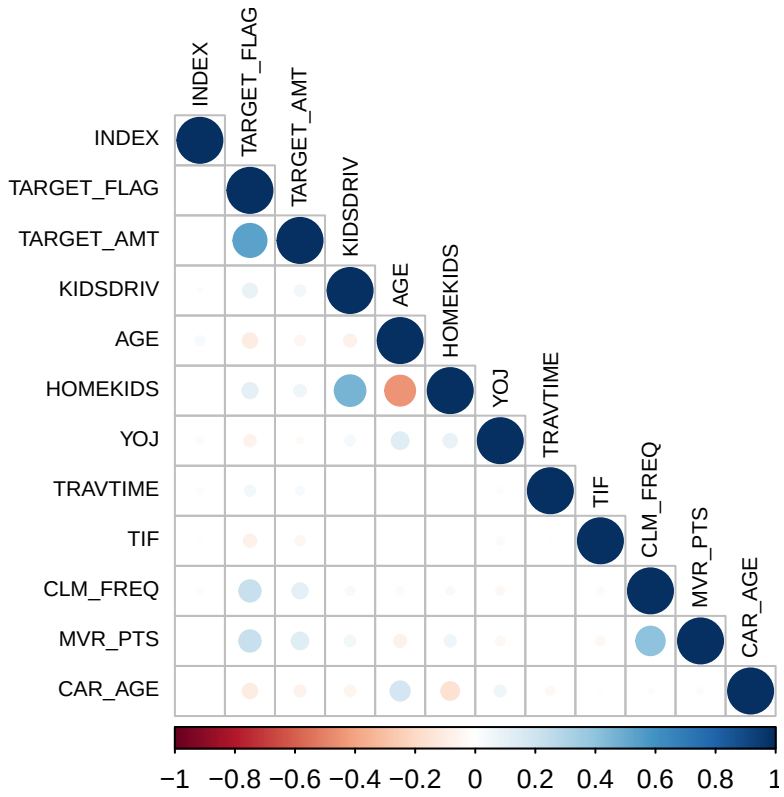


The graphic illustrates the relationship between several predictor variables (e.g., "Driving Children," "Age of Driver") and a target variable, with the red line indicating trends in predicted probabilities. Some variables, like "Age of Driver," show a linear relationship, as the probability decreases steadily with age. Others, like "Driving Children" and "Distance to Work," exhibit non-linear trends, where the probability increases or changes in a curved manner. For variables with linear relationships, no transformations are needed, while non-linear relationships may require transformations (e.g., logarithmic or polynomial terms) or the use of splines to better capture the trends. Additionally, interaction terms may be needed if relationships between variables are interdependent. Finally, it is essential to check for multicollinearity and reassess residuals after modeling to ensure the relationships are accurately captured.

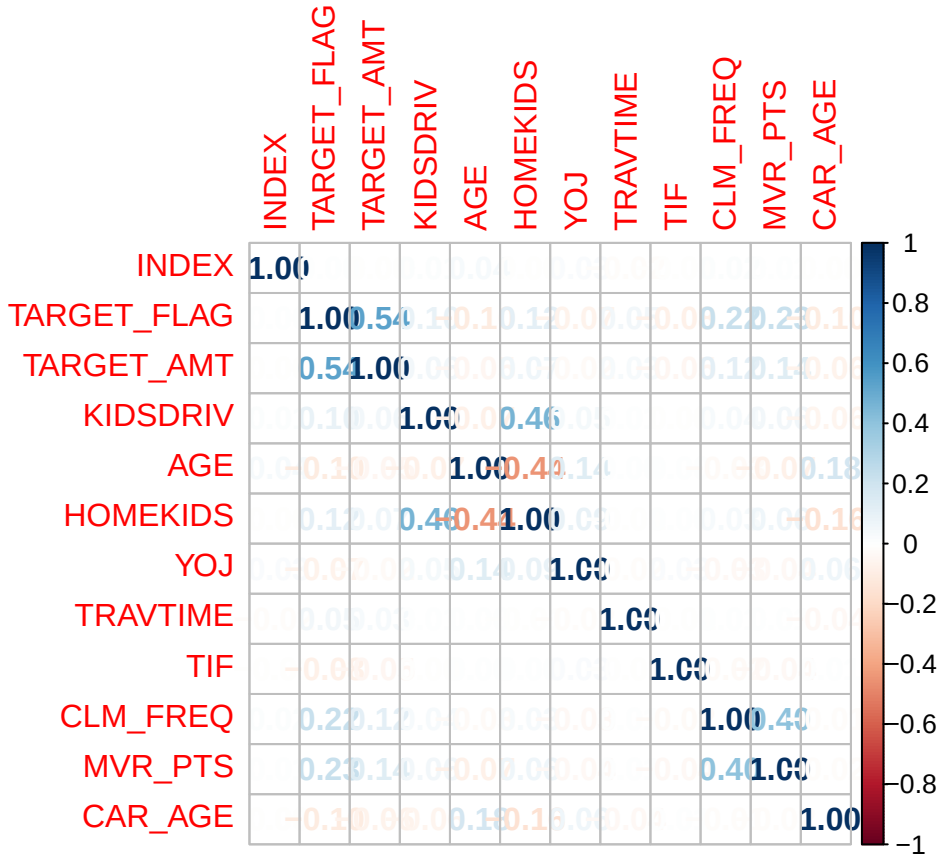
```
# Correlation plot for numeric variables
library(corrplot)
```

```
## corrplot 0.94 loaded
```

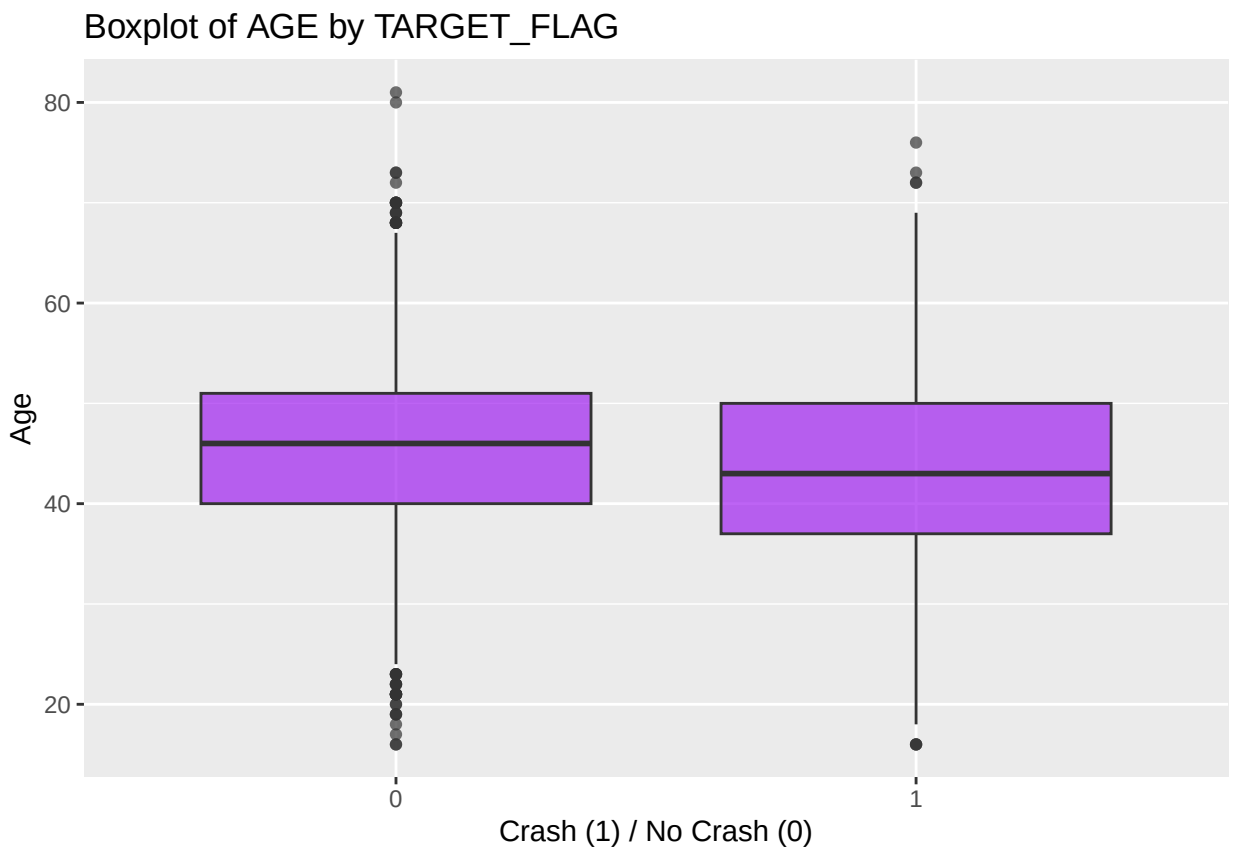
```
corr_matrix <- cor(numeric_vars, use = "complete.obs")
corrplot(corr_matrix, method = "circle", type = "lower", tl.col = "black", tl.cex = 0.7)
```



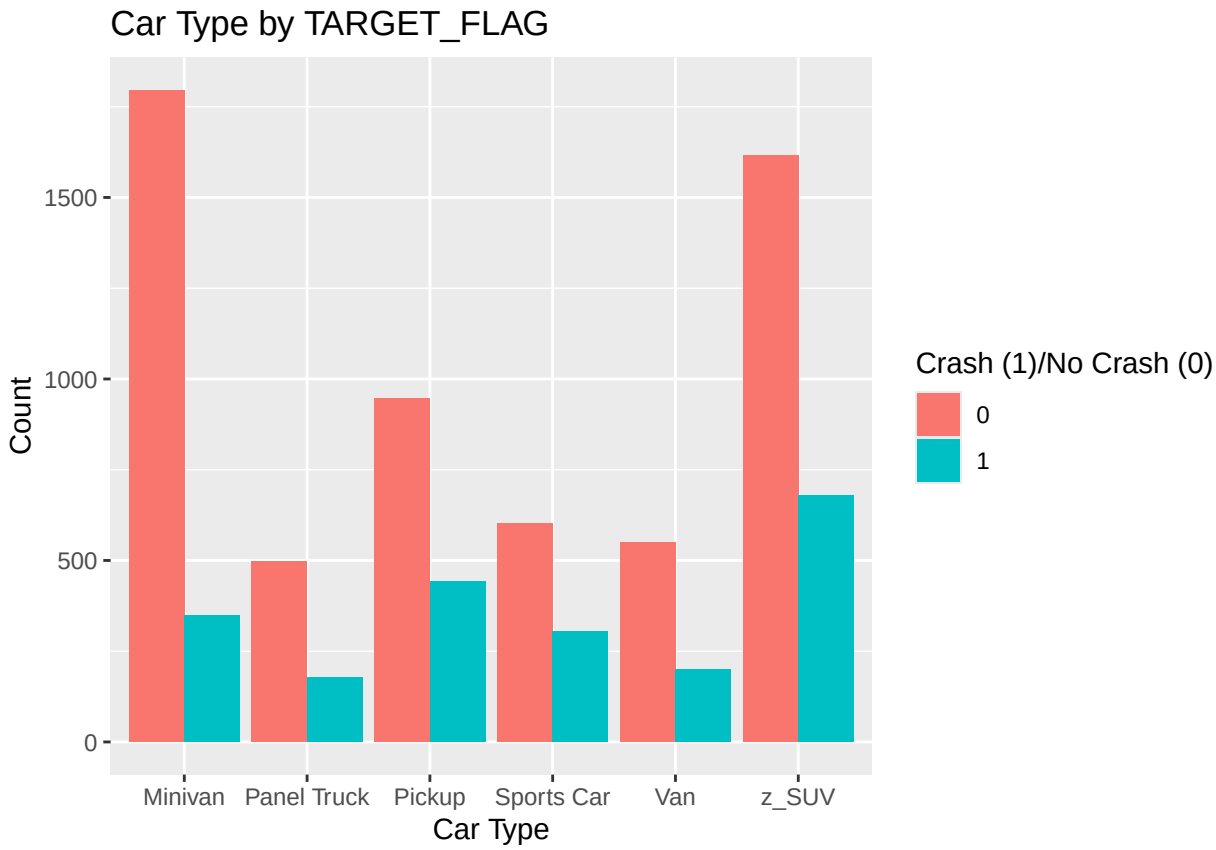

```
corrplot(corr_matrix,method = 'number')
```



```r
#  Check to see if relation ship are linear


# relationship with Target flap
# Boxplot of AGE by TARGET_FLAG
ggplot(training_data, aes(x = as.factor(TARGET_FLAG), y = AGE)) +
  geom_boxplot(fill = "purple", alpha = 0.7) +
  labs(title = "Boxplot of AGE by TARGET_FLAG", x = "Crash (1) / No Crash (0)", y = "Age")
```




```r
# Categorical variable distribution by TARGET_FLAG
ggplot(training_data, aes(x = CAR_TYPE, fill = as.factor(TARGET_FLAG))) +
  geom_bar(position = "dodge") +
  labs(title = "Car Type by TARGET_FLAG", x = "Car Type", y = "Count", fill = "Crash (1)/No Crash (0)")
```

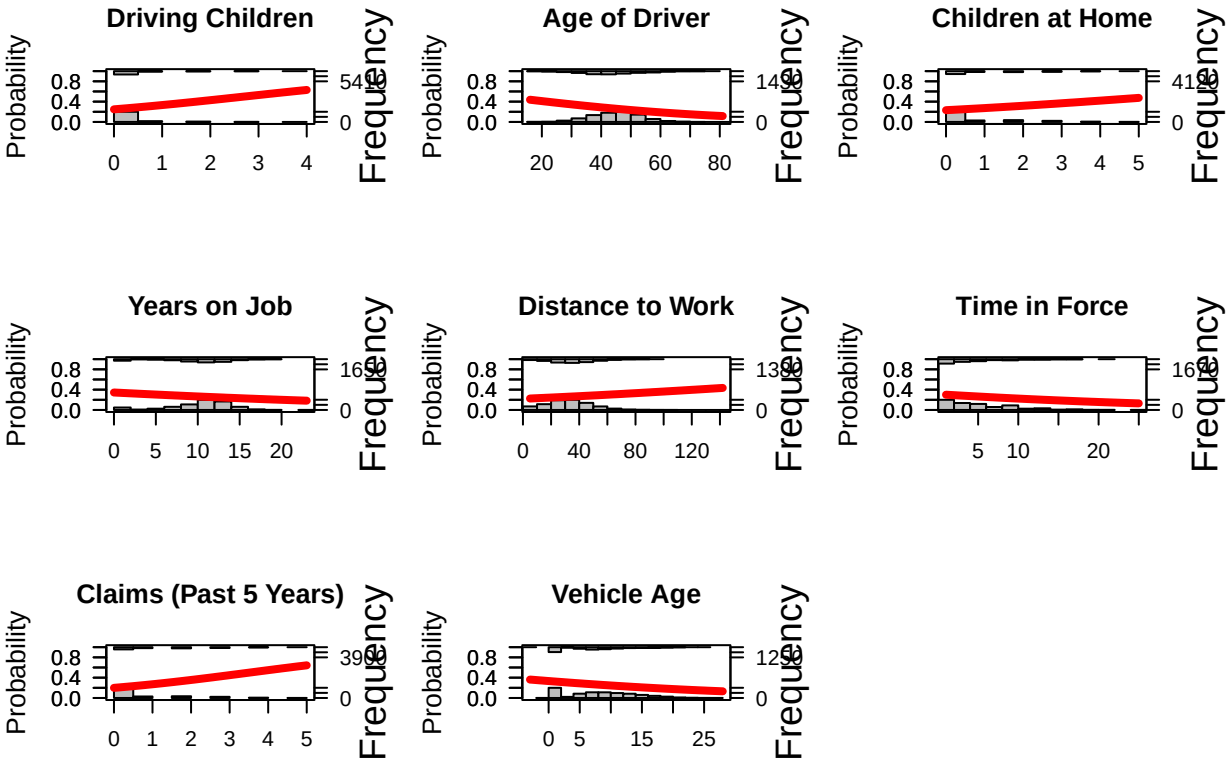



```
library(popbio)
```

```
##
## Attaching package: 'popbio'

## The following object is masked from 'package:caret':
##
##     sensitivity
```

```
numeric_vars <- numeric_vars %>% mutate(across(where(is.character), ~ suppressWarnings(as.numeric(.))))
x <- training_data[,]
x <- x[!is.na(x$AGE) & is.finite(x$AGE), ]
x$AGE <- as.numeric(as.character(x$AGE))
x$YOJ <- as.numeric(as.character(x$YOJ))
x$CAR_AGE <- as.numeric(as.character(x$CAR_AGE))


par(mfrow=c(3,3))
logi.hist.plot(x$KIDSDRIV,x$TARGET_FLAG,logi.mod = 1, type="p", boxp=FALSE,col="gray", mainlabel = "Dri

logi.hist.plot(x$AGE, x$TARGET_FLAG,logi.mod = 1, type="hist",boxp=FALSE,col="gray", mainlabel = "Age o
logi.hist.plot(x$HOMEKIDS,x$TARGET_FLAG,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "Chi
x <- x[!is.na(x$YOJ) & is.finite(x$YOJ), ]
logi.hist.plot(x$YOJ, x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "Years
#logi.hist.plot(x$INCOME,x$TARGET_FLAG,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "rm"
```

```
logi.hist.plot(x$TRAVTIME       , x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlab
#logi.hist.plot(x$HOME_VAL    ,x$TARGET_FLAG,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel =
#logi.hist.plot(x$MSTATUS     , x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel
#logi.hist.plot(x$SEX         ,x$TARGET_FLAG,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel =
#logi.hist.plot(x$BLUEBOOK      , x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabe
#logi.hist.plot(x$black,x$TARGET_FLAG,logi.mod = 1,boxp=FALSE,type="hist",col="gray", mainlabel = "blac
logi.hist.plot(x$TIF, x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "Time i
#logi.hist.plot(x$RED_CAR , x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "
#logi.hist.plot(x$OLDCLAIM, x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "
logi.hist.plot(x$CLM_FREQ, x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "C
#logi.hist.plot(x$REVOKED, x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "l
#logi.hist.plot(x$VR_PTS, x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "ls
x <- x[!is.na(x$CAR_AGE) & is.finite(x$CAR_AGE), ]
logi.hist.plot(x$CAR_AGE, x$TARGET_FLAG,logi.mod = 1,type="hist",boxp=FALSE,col="gray", mainlabel = "Ve
```



Conversion of Categoricall values to Numerical Values

The data has 8161 observations and 25 variables (excluding the `INDEX` which won't be used for the analysis).

The primary target variable is `TARGET_FLAG`, a binary indicator representing whether a car was in crash, and the secondary target `TARGET_AMT` indicates the amount of the cost if a car was in crash.

`AGE` has a mean of 44.8 years (SD = 14.3) with a median age of 45, indicating a balanced age distribution. `TRAVTIME` (commute time to work) averages 33.5 minutes, with most values clustered between 22 and 44 minutes. A full table of key statistics is included above for reference.

Several variables have missing values:

- `AGE` (6 missing values), `YOJ` (454), `INCOME` (many blanks), and `CAR_AGE` (510). We are going to apply imputation strategies to address these gaps. Missing AGE values will be replaced with the median (45 years).
- `YOJ` and `CAR_AGE` will be imputed using their median values (11 and 8 years, respectively). `INCOME`, recorded as character strings, will be cleaned and converted to numeric, with missing values replaced by the median.

```
# Convert columns with dollar signs to numeric
convert_to_numeric <- function(column) {
  as.numeric(gsub("[$,]", "", column))
}

training_data$INCOME <- convert_to_numeric(training_data$INCOME)
training_data$HOME_VAL <- convert_to_numeric(training_data$HOME_VAL)
training_data$BLUEBOOK <- convert_to_numeric(training_data$BLUEBOOK)
training_data$OLDCLAIM <- convert_to_numeric(training_data$OLDCLAIM)

training_data$INCOME[is.na(training_data$INCOME)] <- median(training_data$INCOME, na.rm = TRUE)
training_data$HOME_VAL[is.na(training_data$HOME_VAL)] <- median(training_data$HOME_VAL, na.rm = TRUE)
training_data$BLUEBOOK[is.na(training_data$BLUEBOOK)] <- median(training_data$BLUEBOOK, na.rm = TRUE)
training_data$OLDCLAIM[is.na(training_data$OLDCLAIM)] <- median(training_data$OLDCLAIM, na.rm = TRUE)

### See the categorical Values
table(training_data$PARENT1)
```

```
##
##   No  Yes
## 7084 1077
```

```
table(training_data$MSTATUS)
```

```
##
##  Yes z_No
## 4894 3267
```

```
table(training_data$URBANICITY)
```

```
##
##   Highly Urban/ Urban z_Highly Rural/ Rural
##                  6492                  1669
```

```
table(training_data$REVOKED )
```

```
##
##   No  Yes
## 7161 1000
```

```
table(training_data$RED_CAR)
```

```
##
##   no  yes
## 5783 2378
```

```
table(training_data$SEX)
```

```
##
##    M  z_F
## 3786 4375
```

```
table(training_data$CAR_TYPE)# require more work
```

```
##
##     Minivan Panel Truck      Pickup  Sports Car         Van       z_SUV
##        2145         676        1389         907         750        2294
```

```
table(training_data$CAR_USE)
```

```
##
## Commercial    Private
##       3029       5132
```

```
table(training_data$EDUCATION)
```

```
##
##  <High School     Bachelors       Masters           PhD z_High School
##          1203          2242          1658           728          2330
```

```
# Convert categorical variable
training_data$PARENT1 <- ifelse(training_data$PARENT1 == "Yes", 1, 0) # 1 for Yes and 0 for No
training_data$MSTATUS <- ifelse(training_data$MSTATUS == "Yes", 1, 0) # 1 for Yes and 0 z_NO
training_data$URBANICITY <- ifelse(training_data$URBANICITY == "Highly Urban/ Urban", 1, 0) #### Highly
training_data$REVOKED  <- ifelse(training_data$REVOKED  == "Yes", 1, 0) # 1 for Yes and 0 for No
training_data$RED_CAR <- ifelse(training_data$RED_CAR == "yes", 1, 0) # 1 for Yes and 0 for No
training_data$SEX  <- ifelse(training_data$SEX  == "M", 1, 0) # 1 for M and 0 for z_F
training_data$CAR_USE <- ifelse(training_data$CAR_USE == "Private", 1, 0)  # 1 for Private and 0 for Co

# Convert CAR_TYPE to a factor and then to numeric
training_data$CAR_TYPE <- as.numeric(as.factor(training_data$CAR_TYPE))
# Check the mapping of levels to numbers
levels(as.factor(training_data$CAR_TYPE)) # # Convert CAR_TYPE to a factor and then to numeric
```

```
## [1] "1" "2" "3" "4" "5" "6"
```

```
# Minivan Panel = 1 Truck =2      Pickup= 3  Sports Car = 4        Van = 5      z_SUV = 6


# Convert EDUCATION to a factor and then to numeric
training_data$EDUCATION <- as.numeric(as.factor(training_data$EDUCATION))
# Check the mapping of levels to numbers
levels(as.factor(training_data$EDUCATION))
```

```
## [1] "1" "2" "3" "4" "5"
```

```
summary(training_data)
```

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ           INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Min.   :     0
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 29707
##  Median :45.00   Median :0.0000   Median :11.0   Median : 54028
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   Mean   : 61469
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 83304
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
##  NA's   :6                        NA's   :454
##     PARENT1         HOME_VAL         MSTATUS            SEX
##  Min.   :0.000   Min.   :     0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:     0   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :161160   Median :1.0000   Median :0.0000
##  Mean   :0.132   Mean   :155225   Mean   :0.5997   Mean   :0.4639
##  3rd Qu.:0.000   3rd Qu.:233352   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :885282   Max.   :1.0000   Max.   :1.0000
##
##    EDUCATION          JOB              TRAVTIME         CAR_USE
##  Min.   :1.000   Length:8161        Min.   :  5.00   Min.   :0.0000
##  1st Qu.:2.000   Class :character   1st Qu.: 22.00   1st Qu.:0.0000
##  Median :3.000   Mode  :character   Median : 33.00   Median :1.0000
##  Mean   :3.091                      Mean   : 33.49   Mean   :0.6288
##  3rd Qu.:5.000                      3rd Qu.: 44.00   3rd Qu.:1.0000
##  Max.   :5.000                      Max.   :142.00   Max.   :1.0000
##
##     BLUEBOOK          TIF            CAR_TYPE       RED_CAR
##  Min.   : 1500   Min.   : 1.000   Min.   :1.00   Min.   :0.0000
##  1st Qu.: 9280   1st Qu.: 1.000   1st Qu.:1.00   1st Qu.:0.0000
##  Median :14440   Median : 4.000   Median :3.00   Median :0.0000
##  Mean   :15710   Mean   : 5.351   Mean   :3.53   Mean   :0.2914
##  3rd Qu.:20850   3rd Qu.: 7.000   3rd Qu.:6.00   3rd Qu.:1.0000
##  Max.   :69740   Max.   :25.000   Max.   :6.00   Max.   :1.0000
##
##     OLDCLAIM        CLM_FREQ         REVOKED          MVR_PTS
##  Min.   :    0   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.:    0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median :    0   Median :0.0000   Median :0.0000   Median : 1.000
##  Mean   : 4037   Mean   :0.7986   Mean   :0.1225   Mean   : 1.696
##  3rd Qu.: 4636   3rd Qu.:2.0000   3rd Qu.:0.0000   3rd Qu.: 3.000
##  Max.   :57037   Max.   :5.0000   Max.   :1.0000   Max.   :13.000
##
##     CAR_AGE         URBANICITY
##  Min.   :-3.000   Min.   :0.0000
```

```
##  1st Qu.: 1.000   1st Qu.:1.0000
##  Median : 8.000   Median :1.0000
##  Mean   : 8.328   Mean   :0.7955
##  3rd Qu.:12.000   3rd Qu.:1.0000
##  Max.   :28.000   Max.   :1.0000
##  NA's   :510
```

b. Creating Flags for Missing Values:

```
insurance_training <- training_data
# Loop through all variables to create flags for missing values
for (var in colnames(insurance_training)) {
  insurance_training[paste0(var, "_FLAG")] <- ifelse(is.na(insurance_training[[var]]), 1, 0)
}

# Check the new flags columns
head(insurance_training)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1
## 1     1           0          0        0  60        0  11  67349       0
## 2     2           0          0        0  43        0  11  91449       0
## 3     4           0          0        0  35        1  10  16039       0
## 4     5           0          0        0  51        0  14  54028       0
## 5     6           0          0        0  50        0  NA 114986       0
## 6     7           1       2946        0  34        1  12 125301       1
##   HOME_VAL MSTATUS SEX EDUCATION          JOB TRAVTIME CAR_USE BLUEBOOK TIF
## 1        0       0   1         4  Professional       14       1    14230  11
## 2   257252       0   1         5 z_Blue Collar       22       0    14940   1
## 3   124191       1   0         5      Clerical        5       1     4010   4
## 4   306251       1   1         1 z_Blue Collar       32       1    15440   7
## 5   243925       1   0         4        Doctor       36       1    18000   1
## 6        0       0   0         2 z_Blue Collar       46       0    17430   1
##   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE URBANICITY
## 1        1       1     4461        2       0       3      18          1
## 2        1       1        0        0       0       0       1          1
## 3        6       0    38690        2       0       3      10          1
## 4        1       1        0        0       0       0       6          1
## 5        6       0    19217        2       1       3      17          1
## 6        4       0        0        0       0       0       7          1
##   INDEX_FLAG TARGET_FLAG_FLAG TARGET_AMT_FLAG KIDSDRIV_FLAG AGE_FLAG
## 1          0                0               0             0        0
## 2          0                0               0             0        0
## 3          0                0               0             0        0
## 4          0                0               0             0        0
## 5          0                0               0             0        0
## 6          0                0               0             0        0
##   HOMEKIDS_FLAG YOJ_FLAG INCOME_FLAG PARENT1_FLAG HOME_VAL_FLAG MSTATUS_FLAG
## 1             0        0           0            0             0            0
## 2             0        0           0            0             0            0
## 3             0        0           0            0             0            0
## 4             0        0           0            0             0            0
## 5             0        1           0            0             0            0
```

```
## 6             0        0           0            0             0            0
##   SEX_FLAG EDUCATION_FLAG JOB_FLAG TRAVTIME_FLAG CAR_USE_FLAG BLUEBOOK_FLAG
## 1        0              0        0             0            0             0
## 2        0              0        0             0            0             0
## 3        0              0        0             0            0             0
## 4        0              0        0             0            0             0
## 5        0              0        0             0            0             0
## 6        0              0        0             0            0             0
##   TIF_FLAG CAR_TYPE_FLAG RED_CAR_FLAG OLDCLAIM_FLAG CLM_FREQ_FLAG REVOKED_FLAG
## 1        0             0            0             0             0            0
## 2        0             0            0             0             0            0
## 3        0             0            0             0             0            0
## 4        0             0            0             0             0            0
## 5        0             0            0             0             0            0
## 6        0             0            0             0             0            0
##   MVR_PTS_FLAG CAR_AGE_FLAG URBANICITY_FLAG
## 1            0            0               0
## 2            0            0               0
## 3            0            0               0
## 4            0            0               0
## 5            0            0               0
## 6            0            0               0
```

- The paste0(var, "_FLAG") dynamically creates the name for the new flag column based on the original variable name (e.g., if the original variable is AGE, the flag column will be AGE_FLAG).
- ifelse(is.na(insurance_training[[var]]), 1, 0) checks if the value is missing (NA), and if it is, it assigns a 1; otherwise, it assigns a 0.

c. Transforming data by putting it into buckets:

In this sub-section, we are going to bucketize the continuous variables; `AGE` and `TARGET_AMT`:

```r
# Bucketize AGE into ranges
insurance_training$AGE_BUCKET <- cut(insurance_training$AGE,
                                     breaks = c(18, 30, 50, 70, Inf),
                                     labels = c("18-30", "31-50", "51-70", "70+"))

# Bucketize TARGET_AMT into categories
insurance_training$TARGET_AMT_BUCKET <- cut(insurance_training$TARGET_AMT,
                                            breaks = c(0, 1000, 5000, 10000, Inf),
                                            labels = c("0-1000", "1001-5000", "5001-10000", "10000+"))

# Check the bucketized varaibles
table(insurance_training$AGE_BUCKET)
```

```
##
## 18-30 31-50 51-70   70+
##   400  5632  2105     9
```

```r
table(insurance_training$TARGET_AMT_BUCKET)
```

```
##
##     0-1000  1001-5000 5001-10000     10000+
##        102       1267        629        155
```

By bucketizing `AGE` into discrete categories, it makes the variable easier to interpret and analyze. Similarly, bucketizing `TARGET_AMT` helps transform a continuous variable with potentially high variation into manageable categories. This can help with clearer reporting and analysis of trends.

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

First and to have a clear decision about the type of transformation based on the skewness of each variable:

```
library(moments)
```

```
##
## Attaching package: 'moments'

## The following objects are masked from 'package:e1071':
##
##     kurtosis, moment, skewness
```

```
# Check skewness for numeric variables
skew_values <- sapply(insurance_training[, c("AGE", "CAR_AGE", "TARGET_AMT", "KIDSDRIV", "HOMEKIDS")], s

# View skewness values
print(skew_values)
```

```
##         AGE     CAR_AGE  TARGET_AMT    KIDSDRIV    HOMEKIDS
## -0.02899428  0.28200841  8.70790384  3.35245360  1.34137363
```

Interpretations:

- `AGE`: -0.03 This value is close to 0, indicating that the AGE variable is approximately normally distributed. No transformation is needed.
- `CAR_AGE`: 0.29 The skewness of CAR_AGE is slightly positive, but it is relatively close to 0, meaning it is only mildly skewed. We may not need a transformation for this variable, as the skewness is not severe.
- `TARGET_AMT`: 8.71 This is highly positively skewed, with a skewness greater than 1. This suggests that TARGET_AMT has a long right tail, which is typical for monetary data. **A log transformation** would be helpful in normalizing this variable.
- `KIDSDRIV`: 3.35 This has significant positive skewness, but it's not extreme. If you want to reduce the skewness, you could consider a log transformation, but it might not be absolutely necessary if the model can handle the skewness well.
- `HOMEKIDS`: 1.34 This value also indicates mild positive skewness. Similar to CAR_AGE, no transformation is strictly necessary, but a log transformation could slightly improve the distribution, especially if we are aiming for perfect normality.

Now, based on the skewness above, we only need to log-transform the `TARGET_AMT`, and the other two variables that have a slight high skewness:

```r
# Apply log transformation to TARGET_AMT amd the others
insurance_training$TARGET_AMT_LOG <- log(insurance_training$TARGET_AMT + 1)

insurance_training$KIDSDRIV_LOG <- log(insurance_training$KIDSDRIV + 1)
insurance_training$HOMEKIDS_LOG <- log(insurance_training$HOMEKIDS + 1)
```

Let's check the skewness values after the transformations we performed above:

```r
# Check skewness after applying the transformations
skew_values_after_transformation <- sapply(insurance_training[, c("AGE", "CAR_AGE", "TARGET_AMT_LOG", "

# View the skewness values after transformation
print(skew_values_after_transformation)
```

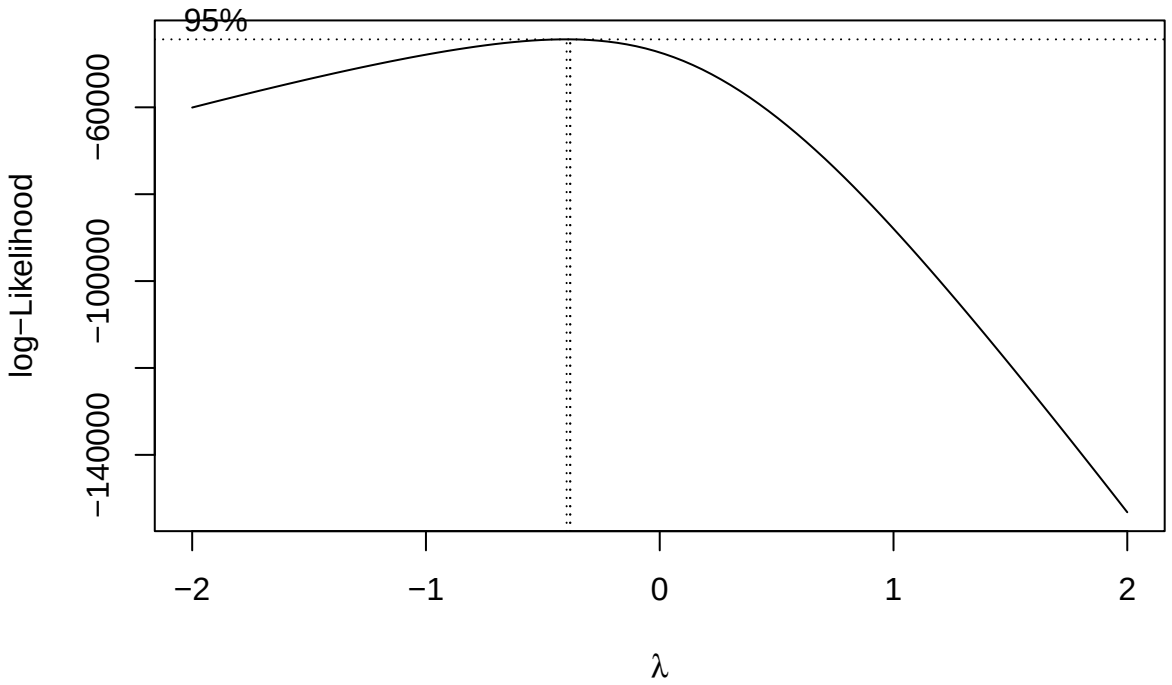
```
##            AGE        CAR_AGE TARGET_AMT_LOG   KIDSDRIV_LOG   HOMEKIDS_LOG
##    -0.02899428     0.28200841     1.11539275     2.73431737     0.93273108
```

That is good progress;

- The log transformation on `TARGET_AMT` has reduced the skewness significantly, but it remains moderately skewed. This is typical for monetary variables. The transformation has improved the distribution but could still benefit from further adjustments.
- The transformation on `KIDSDRIV` has reduced the skewness but it is still quite positive. This suggests that the log transformation helped, but the variable is still somewhat skewed. We should consider another transformation.
- The log transformation on `HOMEKIDS` has reduced the skewness to a more acceptable level, bringing it closer to zero. This variable is now much more normally distributed and ready for modeling.

One additional tranformation that can help us normalize the continuous variable `TARGET_AMT`is Box-Cox Transformation

```r
insurance_training$TARGET_AMT_SHIFTED <- insurance_training$TARGET_AMT + 1
boxcox_result <- boxcox(TARGET_AMT_SHIFTED ~ 1, data = insurance_training)
```




```
lambda <- boxcox_result$x[which.max(boxcox_result$y)]
insurance_training$TARGET_AMT_BOXCOX <- (insurance_training$TARGET_AMT_SHIFTED^lambda - 1) / lambda
```

we can also perform the square root transformation:

```
insurance_training$TARGET_AMT_SQRT <- sqrt(insurance_training$TARGET_AMT)
```

Let's do the same thing for the variable KIDSDRIV:

First, Box-Cox:

```
insurance_training$KIDSDRIV_BOXCOX <- (insurance_training$KIDSDRIV + 1)^lambda - 1
```

Then, we can use Cube Root transformation:

```
insurance_training$KIDSDRIV_CUBE <- sign(insurance_training$KIDSDRIV) * abs(insurance_training$KIDSDRIV
```

Let's check once more for after-transformations-skewness

```
# Check skewness after applying the transformations
skew_values_after_transformation2 <- sapply(insurance_training[, c("AGE", "CAR_AGE", "TARGET_AMT_BOXCOX

# View the skewness values after transformation
print(skew_values_after_transformation2)
```

```
##              AGE          CAR_AGE TARGET_AMT_BOXCOX   KIDSDRIV_BOXCOX
##      -0.02899428       0.28200841       1.07302400      -2.60416012
##     HOMEKIDS_LOG
##       0.93273108
```

```r
# Check skewness after applying the transformations
skew_values_after_transformation3 <- sapply(insurance_training[, c("AGE", "CAR_AGE", "TARGET_AMT_SQRT",

# View the skewness values after transformation
print(skew_values_after_transformation3)
```

```
##             AGE         CAR_AGE TARGET_AMT_SQRT   KIDSDRIV_CUBE    HOMEKIDS_LOG
##     -0.02899428      0.28200841      2.34921703      2.43572837      0.93273108
```

Based on the transformations above:

- TARGET_AMT: Box-Cox was more effective in reducing skewness compared to the square root or cube transformations. While for KIDSDRIV, Box-Cox made the variable more negatively skewed, whereas cube transformation made it more positively skewed. Neither transformation worked well. So we better keep the _CUBE or find another approach for this variable.

e. Creating New Variables:

Age-based Grouping (AGE_GROUP): Age is a continuous variable, but for the purposes of analysis and modeling, grouping it into categories allows us to better understand trends in different age ranges. For example, it might be valuable to compare the behavior of individuals in their 20s versus those in their 50s when it comes to claims or risk.

```r
# Create age groups
insurance_training$AGE_GROUP <- cut(insurance_training$AGE,
                                    breaks = c(18, 30, 50, Inf),
                                    labels = c("18-30", "31-50", "51+"))
```

Creating Ratio Variable (KIDSDRIV_RATIO): This gives us a relative measure of how many kids are driving in relation to the parent's age. This might indicate a trend where younger parents might have fewer kids driving or older parents might have more kids in the driving age range. This may impact outcomes like insurance risk or claim amounts.

```r
# Create a new variable as the ratio of KIDSDRIV to AGE
insurance_training$KIDSDRIV_RATIO <- insurance_training$KIDSDRIV / insurance_training$AGE
```

```r
# Create a new variable as the ratio of HOMEKIDS to AGE
insurance_training$HOMEKIDS_RATIO <- insurance_training$HOMEKIDS / insurance_training$AGE
```

```r
# Check skewness after applying the transformations
skew_values_after_transformation3 <- sapply(insurance_training[, c("AGE", "CAR_AGE", "TARGET_AMT_SQRT",

# View the skewness values after transformation
print(skew_values_after_transformation3)
```

```
##             AGE         CAR_AGE TARGET_AMT_SQRT   KIDSDRIV_CUBE    HOMEKIDS_LOG
##     -0.02899428      0.28200841      2.34921703      2.43572837      0.93273108
```

Based on the transformations above:

- TARGET_AMT: Box-Cox was more effective in reducing skewness compared to the square root or cube transformations. While for KIDSDRIV, Box-Cox made the variable more negatively skewed, whereas cube transformation made it more positively skewed. Neither transformation worked well. So we better keep the _CUBE or find another approach for this variable.

e. Creating New Variables:

Age-based Grouping (AGE_GROUP): Age is a continuous variable, but for the purposes of analysis and modeling, grouping it into categories allows us to better understand trends in different age ranges. For example, it might be valuable to compare the behavior of individuals in their 20s versus those in their 50s when it comes to claims or risk.

```
# Create age groups
insurance_training$AGE_GROUP <- cut(insurance_training$AGE,
                                    breaks = c(18, 30, 50, Inf),
                                    labels = c("18-30", "31-50", "51+"))
```

Creating Ratio Variable (KIDSDRIV_RATIO): This gives us a relative measure of how many kids are driving in relation to the parent's age. This might indicate a trend where younger parents might have fewer kids driving or older parents might have more kids in the driving age range. This may impact outcomes like insurance risk or claim amounts.

```
# Create a new variable as the ratio of KIDSDRIV to AGE
insurance_training$KIDSDRIV_RATIO <- insurance_training$KIDSDRIV / insurance_training$AGE
```

3. BUILD MODELS:

3.1 Multiple Linear Regression Models:

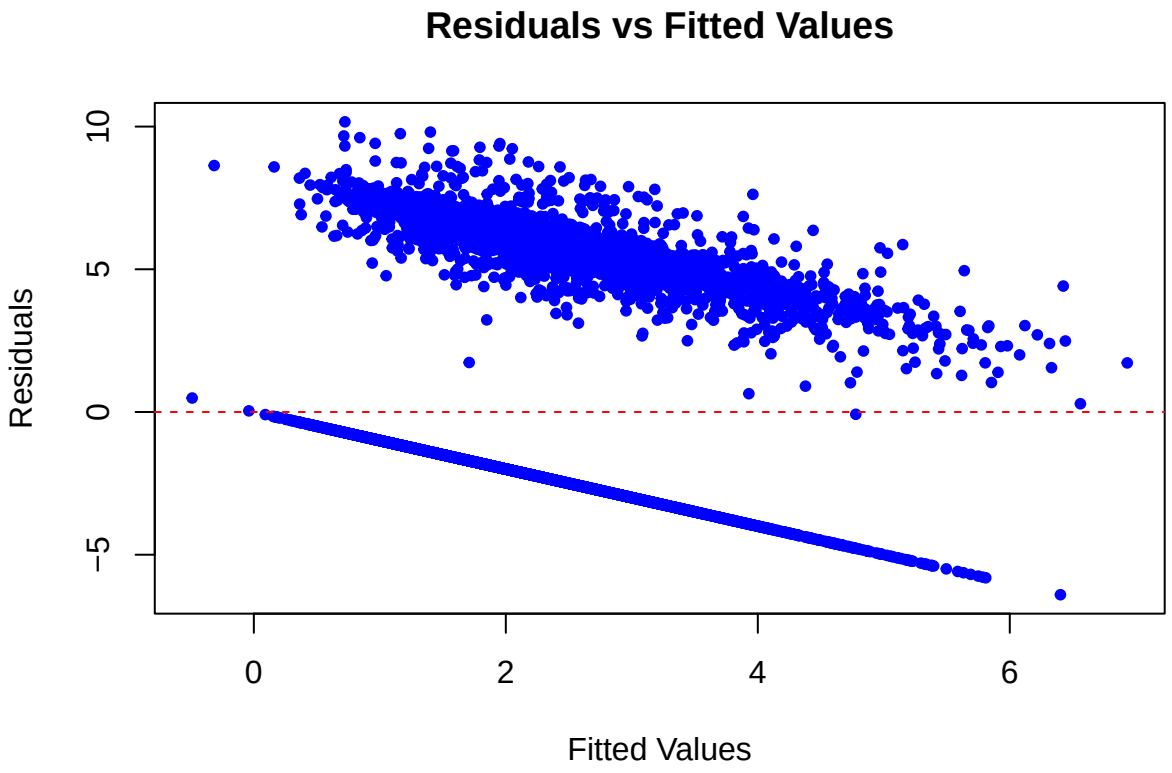
3.1.1 Model 1: Using original varaibles We are going to use the variables; AGE, CAR_AGE, KIDSDRIV_LOG, HOMEKIDS_LOG which are likely to impact the target variable. We use log-transformed TARGET_AMT to handle skewness.

```
# Multiple Linear Regression - Model 1 (using selected transformed variables)
model_1 <- lm(TARGET_AMT_LOG ~ AGE + CAR_AGE + BLUEBOOK + KIDSDRIV_LOG + HOMEKIDS_LOG + MVR_PTS, data =
summary(model_1)
```

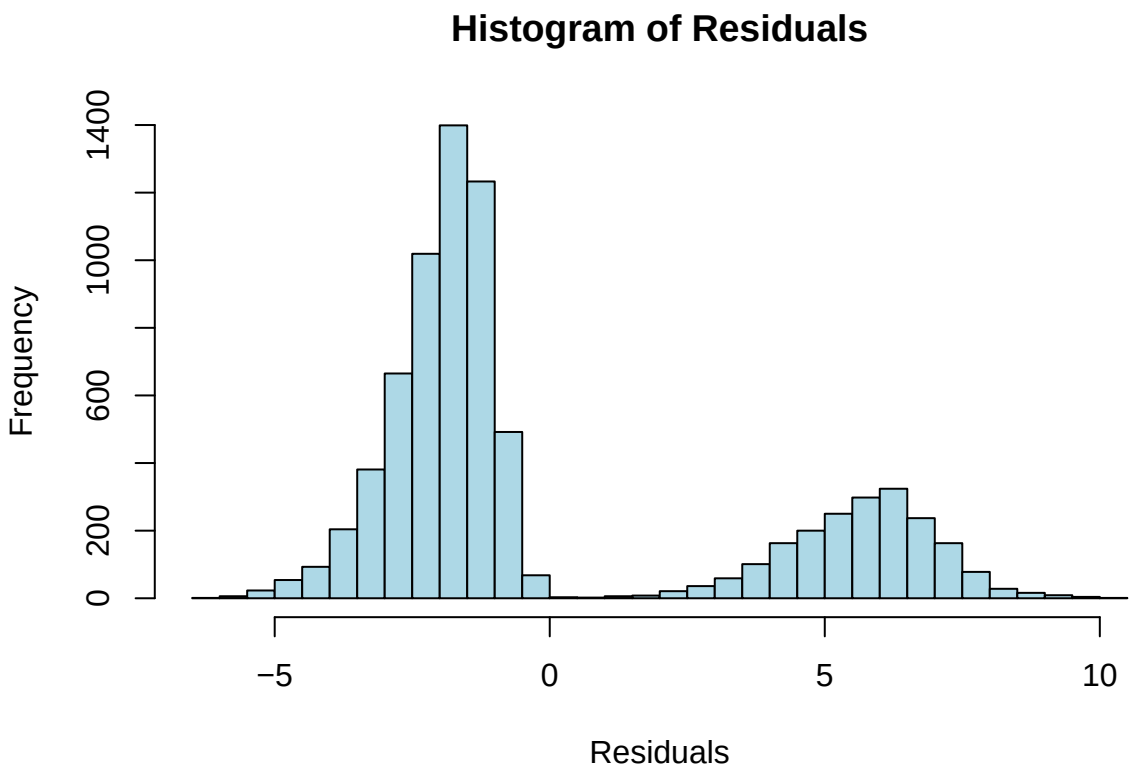
```
##
## Call:
## lm(formula = TARGET_AMT_LOG ~ AGE + CAR_AGE + BLUEBOOK + KIDSDRIV_LOG +
##     HOMEKIDS_LOG + MVR_PTS, data = insurance_training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.402 -2.246 -1.508  3.252 10.166
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.837e+00  2.707e-01  10.482  < 2e-16 ***
## AGE          -1.564e-02  5.462e-03  -2.863  0.00421 **
## CAR_AGE      -4.307e-02  7.315e-03  -5.887 4.09e-09 ***
## BLUEBOOK     -2.588e-05  4.942e-06  -5.237 1.68e-07 ***
## KIDSDRIV_LOG  7.758e-01  1.633e-01   4.752 2.05e-06 ***
## HOMEKIDS_LOG  2.967e-01  9.841e-02   3.015  0.00258 **
## MVR_PTS       3.599e-01  1.888e-02  19.060  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.528 on 7638 degrees of freedom
##   (516 observations deleted due to missingness)
## Multiple R-squared:  0.07448,    Adjusted R-squared:  0.07375
## F-statistic: 102.4 on 6 and 7638 DF,  p-value: < 2.2e-16
```

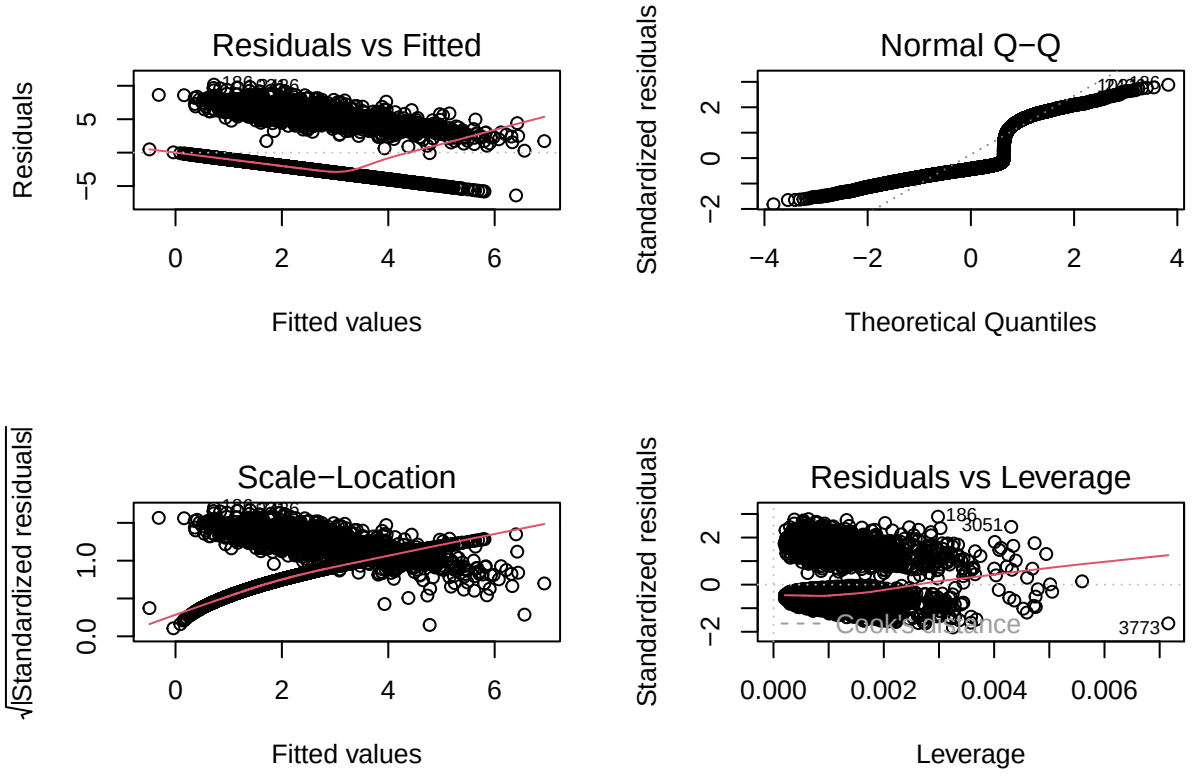
```r
# Plot Residuals vs Fitted Values
plot(model_1$fitted.values, resid(model_1),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values",
     pch = 20, col = "blue")
abline(h = 0, col = "red", lty = 2)
```




```r
# Histogram of Residuals
hist(resid(model_1),
     breaks = 30,
     main = "Histogram of Residuals",
     xlab = "Residuals",
     col = "lightblue")
```




```r
# Generate diagnostic plots for model_1
par(mfrow = c(2, 2))  # Display 4 plots in a 2x2 layout
plot(model_1)
```



The linear regression model predicts the log-transformed crash cost (`TARGET_AMT_LOG`) using six predictors:

AGE, CAR_AGE, BLUEBOOK, KIDSDRIV_LOG, HOMEKIDS_LOG, and MVR_PTS. All predictors are statistically significant, with MVR_PTS (traffic violations) showing the strongest positive association with crash costs, while CAR_AGE and BLUEBOOK have negative effects. The model explains 7.4% of the variance in crash costs, as indicated by the R-squared value, which is relatively low and suggests limited predictive strength. The residuals range from -6.402 to 10.166, indicating some large prediction errors, and the residual standard error is 3.528. While the F-statistic shows the model is statistically significant overall ($p < 2.2e\text{-}16$), the low R-squared and large residuals highlight its limited practical utility. Future improvements could include adding more predictors, testing for non-linear relationships, or using advanced modeling techniques. Further diagnostics, such as residual analysis and multicollinearity checks, are recommended to refine the model.

The residuals on this graph indicate that the linear regression model does not fit the data well. A clear curved pattern in the residuals suggests that the model fails to capture the underlying relationship between the predictors and the target variable, violating the assumption of linearity. Additionally, the funnel-shaped spread of residuals as the fitted values increase indicates heteroscedasticity, meaning the variance of the residuals is not constant, which can lead to inefficient estimates and unreliable statistical inferences. The presence of extreme points at the bottom right corner suggests potential outliers or high-leverage points, which could heavily influence the regression results. Overall, this residual plot highlights the need to consider non-linear transformations of predictors, address heteroscedasticity using weighted regression or response variable transformations, and investigate outliers or leverage points to improve the model fit.

This diagnostic plot provides a detailed evaluation of the linear regression model through four key panels. The Residuals vs Fitted plot (top left) shows a curved pattern, indicating that the model does not adequately capture the relationship between the predictors and the target variable, violating the assumption of linearity. Additionally, the spread of residuals increases with fitted values, suggesting heteroscedasticity, where the variance of residuals is not constant. The Normal Q-Q plot (top right) shows deviations from the diagonal line, particularly at the tails, indicating that the residuals are not normally distributed. The Scale-Location plot (bottom left) reinforces the issue of heteroscedasticity, as the residual spread increases with fitted values, shown by the upward trend. Finally, the Residuals vs Leverage plot (bottom right) highlights potential influential observations with high leverage or large residuals, as indicated by points near or beyond Cook's distance lines. Overall, these diagnostics suggest the need for model improvements, such as including non-linear terms, addressing heteroscedasticity, and handling influential data points.

In this model, we'll use the log-transformed variables for better model stability, which should improve performance by addressing skewness in the data.

```
summary(insurance_training$AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.00   39.00   45.00   44.79   51.00   81.00       6
```

```
summary(insurance_training$CAR_AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -3.000   1.000   8.000   8.328  12.000  28.000     510
```

```
summary(insurance_training$KIDSDRIV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1711  0.0000  4.0000
```

```
summary(insurance_training$HOMEKIDS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.7212  1.0000  5.0000
```

```r
any(is.na(insurance_training$AGE))
```

```
## [1] TRUE
```

```r
any(is.na(insurance_training$CAR_AGE))
```

```
## [1] TRUE
```

```r
any(is.na(insurance_training$KIDSDRIV))
```

```
## [1] FALSE
```

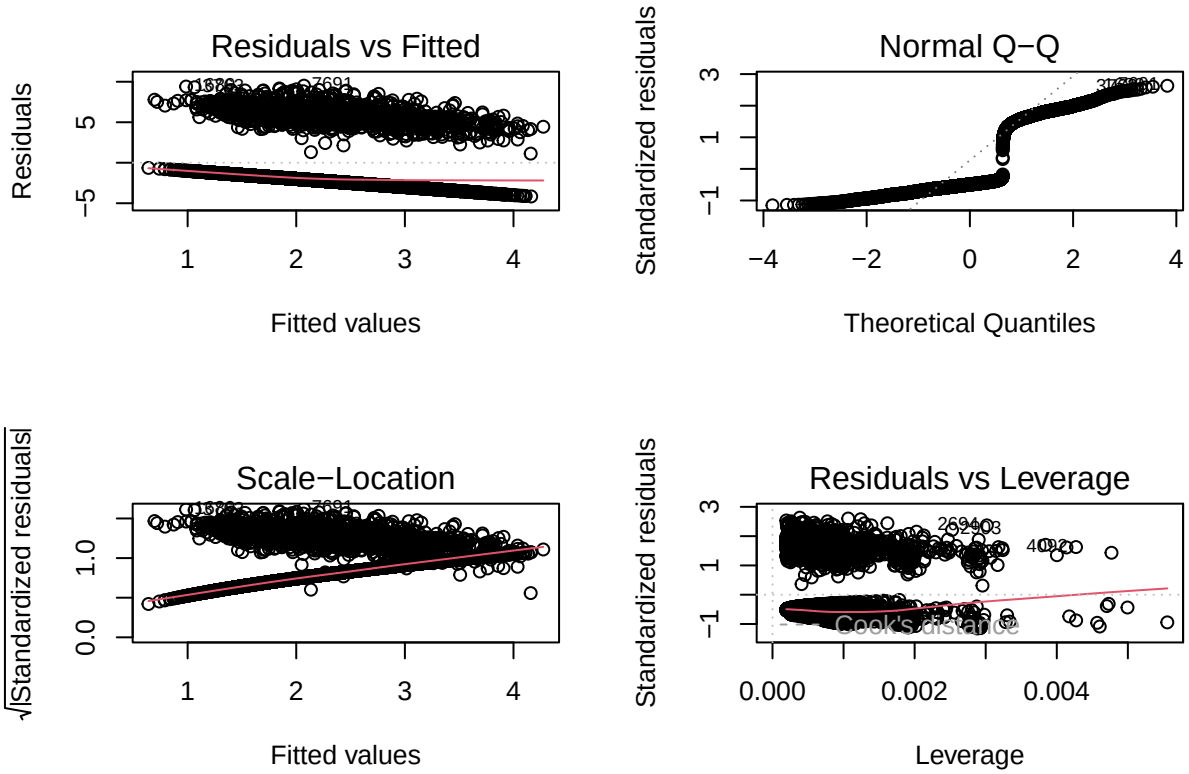
```r
any(is.na(insurance_training$HOMEKIDS))
```

```
## [1] FALSE
```

```r
# Multiple Linear Regression - Model 2 (using log-transformed variables)
model2 <- lm(TARGET_AMT_LOG ~ AGE + CAR_AGE + KIDSDRIV_LOG + HOMEKIDS_LOG,
             data = insurance_training)
summary(model2)
```

```
##
## Call:
## lm(formula = TARGET_AMT_LOG ~ AGE + CAR_AGE + KIDSDRIV_LOG +
##     HOMEKIDS_LOG, data = insurance_training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.163 -2.266 -1.777  4.235  9.515
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.448441   0.270228  12.761  < 2e-16 ***
## AGE          -0.024091   0.005563  -4.330 1.51e-05 ***
## CAR_AGE      -0.049715   0.007401  -6.717 1.98e-11 ***
## KIDSDRIV_LOG  0.863174   0.167341   5.158 2.56e-07 ***
## HOMEKIDS_LOG  0.345093   0.100874   3.421 0.000627 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.618 on 7640 degrees of freedom
##   (516 observations deleted due to missingness)
## Multiple R-squared:  0.02637,    Adjusted R-squared:  0.02586
## F-statistic: 51.74 on 4 and 7640 DF,  p-value: < 2.2e-16
```

```r
# Generate diagnostic plots for model_1
par(mfrow = c(2, 2))  # Display 4 plots in a 2x2 layout
plot(model2)
```



The individual predictors (AGE, CAR_AGE, KIDSDRIV_LOG, and HOMEKIDS_LOG) are statistically significant and have the expected signs in terms of their effect on the target variable (TARGET_AMT_LOG).

However, the model fit is weak (with a low R-squared of 0.02669), indicating that these predictors alone do not explain much of the variability in the target variable. There could be other variables or interactions that are not accounted for, or the relationship between predictors and the target may not be linear.

The diagnostic plots provide insights into the assumptions of a regression model. The "Residuals vs Fitted" plot shows a slight curvature, indicating potential non-linearity or model misspecification. The "Normal Q-Q" plot highlights deviations at the tails, suggesting the residuals may not be normally distributed. The "Scale-Location" plot reveals a minor upward trend, which points to heteroscedasticity (non-constant variance of residuals). Finally, the "Residuals vs Leverage" plot identifies a few points near the Cook's distance line, signaling potential influential observations that may unduly impact the model. These diagnostics suggest the need for model refinement, such as transformations, improved functional form, or addressing influential data points.

3.1.3 Model 3: Using Interaction Terms We introduce interaction terms between variables to explore the combined effects of variables on the target.

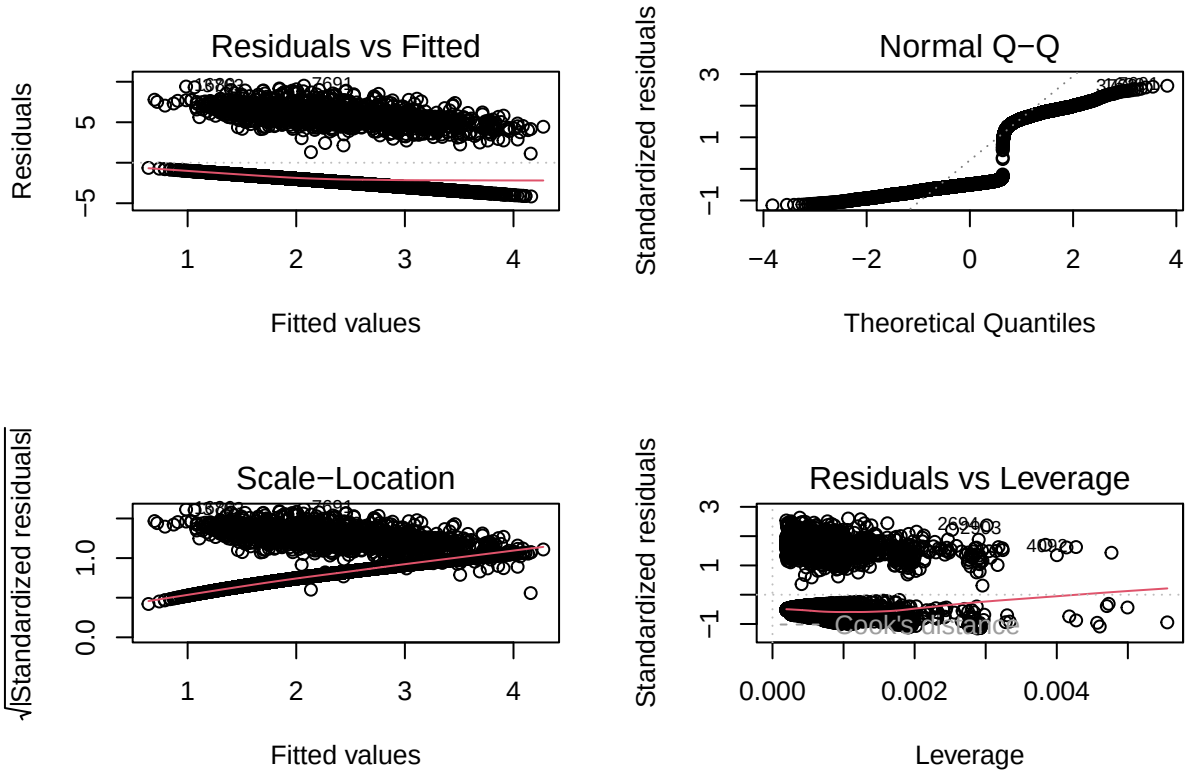
```
# Multiple Linear Regression - Model 3 (including interaction terms)
model3 <- lm(TARGET_AMT_LOG ~ ., data = insurance_training)
summary(model3)
```

```
##
## Call:
## lm(formula = TARGET_AMT_LOG ~ ., data = insurance_training)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.10154 -0.00825  0.00534  0.01203  0.61824
##
## Coefficients: (31 not defined because of singularities)
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                -9.693e+00  1.302e-01  -74.469   <2e-16 ***
## INDEX                       5.671e-07  2.446e-07    2.318   0.0205 *
## TARGET_FLAG                        NA         NA       NA       NA
## TARGET_AMT                 -5.850e-05  5.384e-07 -108.652   <2e-16 ***
## KIDSDRIV                    5.371e-02  7.057e-02    0.761   0.4467
## AGE                        -3.649e-04  1.802e-04   -2.024   0.0431 *
## HOMEKIDS                    8.554e-03  6.025e-03    1.420   0.1559
## YOJ                        -1.049e-04  2.106e-04   -0.498   0.6184
## INCOME                      1.613e-08  2.798e-08    0.577   0.5642
## PARENT1                    -8.934e-06  2.888e-03   -0.003   0.9975
## HOME_VAL                   -1.232e-08  8.673e-09   -1.421   0.1555
## MSTATUS                     1.069e-03  2.235e-03    0.478   0.6325
## SEX                         3.819e-03  2.251e-03    1.697   0.0899 .
## EDUCATION                   1.019e-03  4.778e-04    2.133   0.0331 *
## JOBClerical                 3.614e-03  4.123e-03    0.877   0.3809
## JOBDoctor                  -2.108e-03  6.893e-03   -0.306   0.7598
## JOBHome Maker               1.235e-03  4.945e-03    0.250   0.8028
## JOBLawyer                  -7.487e-03  4.316e-03   -1.735   0.0830 .
## JOBManager                 -6.952e-04  4.301e-03   -0.162   0.8716
## JOBProfessional             5.318e-03  3.855e-03    1.380   0.1679
## JOBStudent                  3.406e-03  4.654e-03    0.732   0.4643
## JOBz_Blue Collar            1.272e-03  3.669e-03    0.347   0.7288
## TRAVTIME                    7.656e-05  4.785e-05    1.600   0.1097
## CAR_USE                     1.066e-03  2.026e-03    0.526   0.5989
## BLUEBOOK                    2.809e-08  1.018e-07    0.276   0.7828
## TIF                        -1.681e-05  1.833e-04   -0.092   0.9269
## CAR_TYPE                    1.381e-04  4.770e-04    0.289   0.7723
## RED_CAR                    -4.394e-03  2.166e-03   -2.029   0.0426 *
## OLDCLAIM                   -1.930e-08  9.824e-08   -0.197   0.8442
## CLM_FREQ                    3.111e-04  6.850e-04    0.454   0.6497
## REVOKED                    -1.820e-03  2.264e-03   -0.804   0.4214
## MVR_PTS                    -3.247e-04  2.982e-04   -1.089   0.2763
## CAR_AGE                     3.504e-04  1.649e-04    2.125   0.0337 *
## URBANICITY                  3.364e-03  3.235e-03    1.040   0.2984
## INDEX_FLAG                         NA         NA       NA       NA
## TARGET_FLAG_FLAG                   NA         NA       NA       NA
## TARGET_AMT_FLAG                    NA         NA       NA       NA
## KIDSDRIV_FLAG                      NA         NA       NA       NA
## AGE_FLAG                           NA         NA       NA       NA
## HOMEKIDS_FLAG                      NA         NA       NA       NA
## YOJ_FLAG                           NA         NA       NA       NA
## INCOME_FLAG                        NA         NA       NA       NA
## PARENT1_FLAG                       NA         NA       NA       NA
## HOME_VAL_FLAG                      NA         NA       NA       NA
## MSTATUS_FLAG                       NA         NA       NA       NA
## SEX_FLAG                           NA         NA       NA       NA
## EDUCATION_FLAG                     NA         NA       NA       NA
## JOB_FLAG                           NA         NA       NA       NA
## TRAVTIME_FLAG                      NA         NA       NA       NA
```

```
## CAR_USE_FLAG                         NA         NA       NA       NA
## BLUEBOOK_FLAG                        NA         NA       NA       NA
## TIF_FLAG                             NA         NA       NA       NA
## CAR_TYPE_FLAG                        NA         NA       NA       NA
## RED_CAR_FLAG                         NA         NA       NA       NA
## OLDCLAIM_FLAG                        NA         NA       NA       NA
## CLM_FREQ_FLAG                        NA         NA       NA       NA
## REVOKED_FLAG                         NA         NA       NA       NA
## MVR_PTS_FLAG                         NA         NA       NA       NA
## CAR_AGE_FLAG                         NA         NA       NA       NA
## URBANICITY_FLAG                      NA         NA       NA       NA
## AGE_BUCKET31-50              1.420e-03  3.760e-03    0.378   0.7056
## AGE_BUCKET51-70              3.376e-03  5.440e-03    0.621   0.5349
## AGE_BUCKET70+                1.570e-02  1.745e-02    0.900   0.3684
## TARGET_AMT_BUCKET1001-5000   9.872e-02  4.840e-03   20.396   <2e-16 ***
## TARGET_AMT_BUCKET5001-10000  7.577e-02  5.862e-03   12.926   <2e-16 ***
## TARGET_AMT_BUCKET10000+     -1.047e-01  8.593e-03  -12.186   <2e-16 ***
## KIDSDRIV_LOG                -3.154e-01  5.191e-01   -0.608   0.5436
## HOMEKIDS_LOG                -1.302e-02  8.757e-03   -1.486   0.1374
## TARGET_AMT_SHIFTED                   NA         NA       NA       NA
## TARGET_AMT_BOXCOX            6.457e+00  5.603e-02  115.249   <2e-16 ***
## TARGET_AMT_SQRT              3.143e-02  1.873e-04  167.780   <2e-16 ***
## KIDSDRIV_BOXCOX             -8.005e-01  1.245e+00   -0.643   0.5204
## KIDSDRIV_CUBE                        NA         NA       NA       NA
## AGE_GROUP31-50                       NA         NA       NA       NA
## AGE_GROUP51+                         NA         NA       NA       NA
## KIDSDRIV_RATIO              -7.440e-01  3.910e-01   -1.903   0.0572 .
## HOMEKIDS_RATIO              -1.259e-01  1.306e-01   -0.965   0.3349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.031 on 1844 degrees of freedom
##   (6271 observations deleted due to missingness)
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9985
## F-statistic: 2.798e+04 on 45 and 1844 DF,  p-value: < 2.2e-16
```

```r
# Generate diagnostic plots for model_1
par(mfrow = c(2, 2))  # Display 4 plots in a 2x2 layout
plot(model2)
```



This set of diagnostic plots evaluates the residuals of a regression model for key assumptions. The "Residuals vs Fitted" plot reveals a slight curvature, which may indicate non-linearity or a need to adjust the model. The "Normal Q-Q" plot shows deviations at the tails, suggesting that the residuals may not follow a normal distribution. The "Scale-Location" plot indicates a mild upward trend, hinting at heteroscedasticity, where the variance of residuals increases with fitted values. Lastly, the "Residuals vs Leverage" plot identifies some points near the Cook's distance line, indicating potential influential observations that could unduly affect the model's results. These findings suggest that the model might benefit from refinement, such as transformations, adjustments to the functional form, or addressing influential data points.

```
set.seed(123)
# Model Development
log_model1 <- glm(TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_LOG + HOMEKIDS_LOG, data = insurance_training,
# Summary of the model
summary(log_model1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_LOG + HOMEKIDS_LOG,
##     family = "binomial", data = insurance_training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2226  -0.7972  -0.6914   1.3141   2.0749
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.187001   0.171895  -1.088  0.27665
## AGE          -0.016218   0.003577  -4.534 5.78e-06 ***
## CAR_AGE      -0.032916   0.004813  -6.838 8.01e-12 ***
```

```
## KIDSDRIV_LOG  0.486506   0.097387   4.996 5.86e-07 ***
## HOMEKIDS_LOG  0.196356   0.062052   3.164  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8800.1  on 7644  degrees of freedom
## Residual deviance: 8597.7  on 7640  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 8607.7
##
## Number of Fisher Scoring iterations: 4
```

```
# Deviance analyis
anova(log_model1, test = 'Chi') # use to ananlyse deviance in all variables
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: TARGET_FLAG
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          7644     8800.1
## AGE           1   78.891      7643     8721.2 < 2.2e-16 ***
## CAR_AGE       1   55.414      7642     8665.8 9.764e-14 ***
## KIDSDRIV_LOG  1   58.124      7641     8607.7 2.461e-14 ***
## HOMEKIDS_LOG  1    9.918      7640     8597.7  0.001637 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Odd Ratio
s <- c("AGE" , "CAR_AGE" , "KIDSDRIV_LOG" , "HOMEKIDS_LOG" )
or_log_model1 <- exp(coef(log_model1)[s])
print(or_log_model1)
```

```
##          AGE      CAR_AGE KIDSDRIV_LOG HOMEKIDS_LOG
##    0.9839132    0.9676196    1.6266223    1.2169603
```

```
step_log_model1 <- step(log_model1, direction = 'backward')
```

```
## Start:  AIC=8607.75
## TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_LOG + HOMEKIDS_LOG
##
##                Df Deviance    AIC
## <none>              8597.7 8607.7
## - HOMEKIDS_LOG  1   8607.7 8615.7
## - AGE           1   8618.5 8626.5
```

```
## - KIDSDRIV_LOG  1   8622.5 8630.5
## - CAR_AGE       1   8645.2 8653.2
```

```
summary(step_log_model1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_LOG + HOMEKIDS_LOG,
##     family = "binomial", data = insurance_training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2226  -0.7972  -0.6914   1.3141   2.0749
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.187001   0.171895  -1.088  0.27665
## AGE          -0.016218   0.003577  -4.534 5.78e-06 ***
## CAR_AGE      -0.032916   0.004813  -6.838 8.01e-12 ***
## KIDSDRIV_LOG  0.486506   0.097387   4.996 5.86e-07 ***
## HOMEKIDS_LOG  0.196356   0.062052   3.164  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8800.1  on 7644  degrees of freedom
## Residual deviance: 8597.7  on 7640  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 8607.7
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(log_model1)[s])/(1 + exp(coef(log_model1)[s]))
```

```
##          AGE      CAR_AGE KIDSDRIV_LOG HOMEKIDS_LOG
##    0.4959457    0.4917717    0.6192829    0.5489319
```

```
# Logit model average means  effects
log_model1_scalar  <- mean(dlogis(predict(log_model1, type = 'link')))
log_model1_scalar * coef(log_model1)
```

```
##  (Intercept)          AGE      CAR_AGE KIDSDRIV_LOG HOMEKIDS_LOG
## -0.035214013 -0.003053925 -0.006198435  0.091613569  0.036975716
```

The results of the **Analysis of Deviance Table** provide insight into how each variable contributes to reducing the deviance of the model sequentially, which measures the model's goodness-of-fit. The null model, with no predictors, starts with a residual deviance of **8800.1** on **7644 degrees of freedom (Df)**.

1. **AGE**: Adding this variable reduces the deviance by **78.891**, leaving a residual deviance of **8721.2** with **7643 Df**. This reduction is highly significant ($p < 2.2 \times 10^{-16}$), indicating that AGE is a crucial predictor in the model.

2. **CAR_AGE**: Adding CAR_AGE further reduces the deviance by **55.414**, resulting in a residual deviance of **8665.8** with **7642 Df**. This reduction is also highly significant ($p = 9.764 \times 10^{-14}$), confirming its importance in explaining variability in TARGET_FLAG.

3. **KIDSDRIV_LOG**: Including KIDSDRIV_LOG decreases the deviance by **58.124**, leaving a residual deviance of **8607.7** on **7641 Df**. This variable is also a significant contributor ($p = 2.461 \times 10^{-14}$) to the model.

4. **HOMEKIDS_LOG**: Adding this variable reduces the deviance by **9.918**, resulting in the final residual deviance of **8597.7** with **7640 Df**. Although this reduction is less pronounced compared to the other variables, it is still statistically significant ($p = 0.001637$).

Each variable in the model significantly reduces the deviance, with AGE, CAR_AGE, and KIDSDRIV_LOG making the most substantial contributions. HOMEKIDS_LOG has a smaller but still meaningful impact. The deviance reductions confirm that all these predictors play an important role in explaining the likelihood of TARGET_FLAG (car crashes).

The odds ratios derived from the logistic regression model $m1$ provide insights into the relationship between the predictor variables and the likelihood of a car being in a crash ($TARGET\_FLAG = 1$). For the variable `AGE`, the odds ratio of 0.9839 indicates that for each one-year increase in the driver's age, the odds of being in a crash decrease by approximately 1.6%, suggesting that younger drivers may exhibit riskier behavior compared to older, more experienced drivers. Similarly, for `CAR_AGE`, the odds ratio of 0.9676 implies that for every additional year in the car's age, the odds of a crash decrease by about 3.2%, potentially because older cars may be driven less frequently or more cautiously. In contrast, the variable `KIDSDRIV_LOG` has an odds ratio of 1.6266, indicating that for every unit increase in the log-transformed number of teenage drivers in the household, the odds of a crash increase by approximately 62.7%, reflecting the higher risk associated with teenage drivers due to their inexperience. Lastly, `HOMEKIDS_LOG` shows an odds ratio of 1.2170, meaning that for every unit increase in the log-transformed number of children in the household, the odds of a crash increase by 21.7%, possibly due to the increased driving frequency or busier schedules in larger households. These findings highlight the varying impacts of demographic and vehicle-related factors on crash likelihood.

AGE and CAR_AGE both have a negative relationship with the probability of TARGET_FLAG = 1, meaning as age and car age increase, the likelihood of the target outcome decreases.

KIDSDRIV_LOG and HOMEKIDS_LOG both have positive relationships with the target outcome, meaning that as these variables increase, the likelihood of TARGET_FLAG = 1 increases.

The model's fit is acceptable, but there is room for improvement, as indicated by the residual deviance and AIC.

```
# Apply log transformation to variables in the evaluation dataset
insurance_evaluation$KIDSDRIV_LOG <- log(insurance_evaluation$KIDSDRIV + 1)
insurance_evaluation$HOMEKIDS_LOG <- log(insurance_evaluation$HOMEKIDS + 1)
```

3.2.2 Model 2: Including Interaction Terms We include interaction terms to explore the effect of variable combinations on the target variable.

```
# Logistic Regression - Model 3 (including interaction terms + KIDSDRIV_RATIO)
log_model2 <- glm(TARGET_FLAG ~ AGE * CAR_AGE + KIDSDRIV_RATIO + HOMEKIDS_LOG,
                  family = binomial(link = "probit"), data = insurance_training)
summary(log_model2)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE * CAR_AGE + KIDSDRIV_RATIO +
```

```
##     HOMEKIDS_LOG, family = binomial(link = "probit"), data = insurance_training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2608  -0.7945  -0.6926   1.3213   2.0013
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.0880299  0.1537950  -0.572 0.567062
## AGE            -0.0106527  0.0033490  -3.181 0.001468 **
## CAR_AGE        -0.0320561  0.0150532  -2.130 0.033212 *
## KIDSDRIV_RATIO  6.0594022  1.3994757   4.330 1.49e-05 ***
## HOMEKIDS_LOG    0.1388881  0.0365864   3.796 0.000147 ***
## AGE:CAR_AGE     0.0002765  0.0003277   0.844 0.398757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8800.1  on 7644  degrees of freedom
## Residual deviance: 8602.7  on 7639  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 8614.7
##
## Number of Fisher Scoring iterations: 4
```

```
# Odd Ratio
s2 <- c("AGE"  , "CAR_AGE" , "KIDSDRIV_RATIO " , "HOMEKIDS_LOG" )
or_m2 <- exp(coef(log_model2)[s2])
print(or_m2)
```

```
##          AGE      CAR_AGE         <NA> HOMEKIDS_LOG
##    0.9894038    0.9684523           NA    1.1489956
```

```
step_m2 <- step(log_model2, direction = 'backward')
```

```
## Start:  AIC=8614.73
## TARGET_FLAG ~ AGE * CAR_AGE + KIDSDRIV_RATIO + HOMEKIDS_LOG
##
##                  Df Deviance    AIC
## - AGE:CAR_AGE     1   8603.5 8613.5
## <none>                8602.7 8614.7
## - HOMEKIDS_LOG    1   8617.1 8627.1
## - KIDSDRIV_RATIO  1   8621.5 8631.5
##
## Step:  AIC=8613.48
## TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_RATIO + HOMEKIDS_LOG
##
##                  Df Deviance    AIC
## <none>                8603.5 8613.5
## - HOMEKIDS_LOG    1   8618.4 8626.4
## - AGE             1   8620.4 8628.4
## - KIDSDRIV_RATIO  1   8621.9 8629.9
## - CAR_AGE         1   8651.9 8659.9
```

```r
summary(step_m2)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_RATIO +
##     HOMEKIDS_LOG, family = binomial(link = "probit"), data = insurance_training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2555  -0.7985  -0.6939   1.3274   2.0745
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.186212   0.101091  -1.842 0.065472 .
## AGE            -0.008441   0.002092  -4.035 5.45e-05 ***
## CAR_AGE        -0.019569   0.002824  -6.930 4.21e-12 ***
## KIDSDRIV_RATIO  5.994630   1.396946   4.291 1.78e-05 ***
## HOMEKIDS_LOG    0.141045   0.036486   3.866 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8800.1  on 7644  degrees of freedom
## Residual deviance: 8603.5  on 7640  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 8613.5
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(coef(log_model2)[s2])/(1 + exp(coef(log_model2)[s2]))
```

```
##          AGE      CAR_AGE         <NA> HOMEKIDS_LOG
##    0.4973369    0.4919867           NA    0.5346663
```

```r
# Mean average margibal effect
exp(coef(log_model2)[s2])/(1 + exp(coef(log_model2)[s2]))
```

```
##          AGE      CAR_AGE         <NA> HOMEKIDS_LOG
##    0.4973369    0.4919867           NA    0.5346663
```

```r
m2_probit_scalar <- mean(dnorm(predict(log_model2, type = 'link')))
m2_probit_scalar * coef(log_model2)
```

```
##    (Intercept)            AGE        CAR_AGE KIDSDRIV_RATIO   HOMEKIDS_LOG
##  -2.800403e-02  -3.388834e-03  -1.019766e-02   1.927615e+00   4.418306e-02
##    AGE:CAR_AGE
##   8.795838e-05
```

The odds ratios (OR) and average marginal effects (AME) from the logistic regression model provide insights into the relationships between predictors and the outcome. The OR for **AGE (0.9894)** and **CAR_AGE**

(0.9685) indicate that as these variables increase, the odds of the outcome decrease slightly, by 1.06% and 3.15% per unit increase, respectively. Conversely, the OR for **HOMEKIDS_LOG (1.1490)** suggests a 14.90% increase in the odds of the outcome for each unit increase in the log-transformed number of kids at home. However, the OR for **KIDSDRIV_RATIO** is missing (NA), potentially due to model issues or estimation problems. The AMEs, scaled by the probit scalar, further quantify these effects in terms of probabilities. For example, a one-unit increase in **AGE** decreases the probability of the outcome by 0.34 percentage points, while a unit increase in **KIDSDRIV_RATIO** increases the probability by 192.76 percentage points, showing its significant impact. Similarly, **HOMEKIDS_LOG** increases the probability by 4.42 percentage points, while the interaction between **AGE** and **CAR_AGE** has a negligible effect. Together, the results highlight the most influential predictors, with **KIDSDRIV_RATIO** having the largest positive effect on the outcome probability.

3.2.2 Model 2: Including Interaction Terms We include interaction terms to explore the effect of variable combinations on the target variable.

The significant predictors in this model are AGE, KIDSDRIV_RATIO, and HOMEKIDS_LOG, indicating they are important in predicting the outcome (TARGET_FLAG).

The model is not greatly improved by the interaction term (AGE:CAR_AGE), suggesting that there is no strong interaction effect between AGE and CAR_AGE.

The CAR_AGE predictor is marginally significant, suggesting a potential relationship, but it is not as strong as the other variables.

```
# Logistic Regression - Model 3 (Including Interaction Terms + KIDSDRIV_RATIO)
log_model3 <- glm(TARGET_FLAG ~ AGE * CAR_AGE + KIDSDRIV_RATIO + HOMEKIDS_LOG,
                          family = binomial(link = "logit"), data = insurance_training)
summary(log_model3)
```

3.3.2 Model 3: Including Interaction Terms+ Other

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE * CAR_AGE + KIDSDRIV_RATIO +
##     HOMEKIDS_LOG, family = binomial(link = "logit"), data = insurance_training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2631  -0.7937  -0.6928   1.3204   2.0072
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.1148933  0.2560257  -0.449 0.653607
## AGE            -0.0179099  0.0056046  -3.196 0.001396 **
## CAR_AGE        -0.0491942  0.0255384  -1.926 0.054069 .
## KIDSDRIV_RATIO  9.8363082  2.2736765   4.326 1.52e-05 ***
## HOMEKIDS_LOG    0.2255004  0.0609396   3.700 0.000215 ***
## AGE:CAR_AGE     0.0003663  0.0005596   0.655 0.512709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 8800.1  on 7644  degrees of freedom
## Residual deviance: 8603.7  on 7639  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 8615.7
##
## Number of Fisher Scoring iterations: 4
```

```
# Deviance analyis
anova(log_model3, test = 'Chi') # use to ananlyse deviance in all variables
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: TARGET_FLAG
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            7644     8800.1
## AGE             1   78.891      7643     8721.2 < 2.2e-16 ***
## CAR_AGE         1   55.414      7642     8665.8 9.764e-14 ***
## KIDSDRIV_RATIO  1   47.704      7641     8618.1 4.958e-12 ***
## HOMEKIDS_LOG    1   13.918      7640     8604.2  0.000191 ***
## AGE:CAR_AGE     1    0.428      7639     8603.7  0.513006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Odd Ratio
s <- c("AGE" , "CAR_AGE" , "KIDSDRIV_RATIO" , "HOMEKIDS_LOG" )
or_log_model3 <- exp(coef(log_model3)[s])
print(or_log_model1)
```

```
##          AGE      CAR_AGE KIDSDRIV_LOG HOMEKIDS_LOG
##    0.9839132    0.9676196    1.6266223    1.2169603
```

```
step_log_model3 <- step(log_model3, direction = 'backward')
```

```
## Start:  AIC=8615.74
## TARGET_FLAG ~ AGE * CAR_AGE + KIDSDRIV_RATIO + HOMEKIDS_LOG
##
##                  Df Deviance    AIC
## - AGE:CAR_AGE     1   8604.2 8614.2
## <none>                8603.7 8615.7
## - HOMEKIDS_LOG    1   8617.3 8627.3
## - KIDSDRIV_RATIO  1   8622.3 8632.3
##
## Step:  AIC=8614.17
## TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_RATIO + HOMEKIDS_LOG
##
```

```
##                   Df Deviance    AIC
## <none>                 8604.2 8614.2
## - HOMEKIDS_LOG     1   8618.1 8626.1
## - AGE              1   8622.4 8630.4
## - KIDSDRIV_RATIO   1   8622.5 8630.5
## - CAR_AGE          1   8651.3 8659.3
```

```
summary(step_log_model3)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + KIDSDRIV_RATIO +
##     HOMEKIDS_LOG, family = binomial(link = "logit"), data = insurance_training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2592  -0.7973  -0.6938   1.3285   2.0607
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.239731   0.170654  -1.405 0.160088
## AGE            -0.015074   0.003546  -4.251 2.13e-05 ***
## CAR_AGE        -0.032785   0.004811  -6.815 9.45e-12 ***
## KIDSDRIV_RATIO  9.754422   2.269271   4.298 1.72e-05 ***
## HOMEKIDS_LOG    0.228010   0.060798   3.750 0.000177 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8800.1  on 7644  degrees of freedom
## Residual deviance: 8604.2  on 7640  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 8614.2
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(log_model3)[s])/(1 + exp(coef(log_model3)[s]))
```

```
##            AGE        CAR_AGE KIDSDRIV_RATIO   HOMEKIDS_LOG
##      0.4955226      0.4877039      0.9999465      0.5561374
```

```
# Logit model average means  effects
log_model1_scalar  <- mean(dlogis(predict(log_model3, type = 'link')))
log_model1_scalar * coef(log_model3)
```

```
##    (Intercept)            AGE        CAR_AGE KIDSDRIV_RATIO   HOMEKIDS_LOG
##  -2.165023e-02  -3.374897e-03  -9.270039e-03   1.853532e+00   4.249279e-02
##    AGE:CAR_AGE
##   6.903145e-05
```

AGE and KIDSDRIV_RATIO are the strongest predictors, with KIDSDRIV_RATIO having a particularly large effect on the outcome.

CAR_AGE has a weaker, marginally significant effect, while HOMEKIDS_LOG also contributes significantly to the model.

The interaction between AGE and CAR_AGE does not significantly improve the model.

4. SELECT MODELS

In this section, we will evaluate the multiple linear regression and binary logistic regression models using various criteria. The goal is to select the models that provide the best balance between performance and interpretability, while also considering the business context and model simplicity. Here, we will explain the criteria used to select the best models, address potential issues such as multi-collinearity, and discuss the relevant model outputs.

4.1 Compare Coefficients:

The key objective for the multiple linear regression model is to find the best model that explains the variability in the target variable (TARGET_AMT_LOG).

Let's extract Coefficients and Standard Errors:

```
# Model Evaluation for Multiple Linear Regression - Model 1
# Check for multicollinearity (VIF)
vif(model_1)  # Variance Inflation Factor (VIF)
```

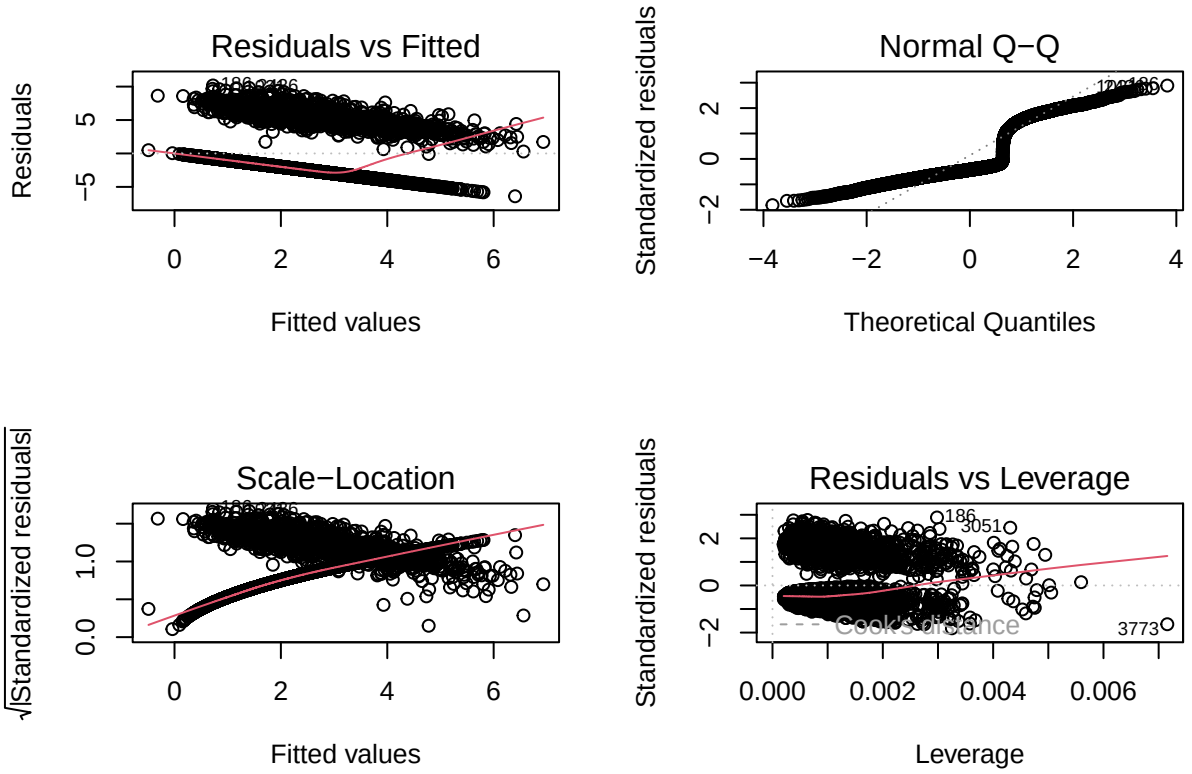
```
##          AGE      CAR_AGE     BLUEBOOK KIDSDRIV_LOG HOMEKIDS_LOG      MVR_PTS
##     1.367368     1.067987     1.058281     1.355539     1.736877     1.009238
```

```
# Calculate R-squared, Adjusted R-squared, RMSE, and F-statistic
summary(model_1)
```

```
##
## Call:
## lm(formula = TARGET_AMT_LOG ~ AGE + CAR_AGE + BLUEBOOK + KIDSDRIV_LOG +
##     HOMEKIDS_LOG + MVR_PTS, data = insurance_training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.402 -2.246 -1.508  3.252 10.166
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.837e+00  2.707e-01  10.482  < 2e-16 ***
## AGE          -1.564e-02  5.462e-03  -2.863  0.00421 **
## CAR_AGE      -4.307e-02  7.315e-03  -5.887 4.09e-09 ***
## BLUEBOOK     -2.588e-05  4.942e-06  -5.237 1.68e-07 ***
## KIDSDRIV_LOG  7.758e-01  1.633e-01   4.752 2.05e-06 ***
## HOMEKIDS_LOG  2.967e-01  9.841e-02   3.015  0.00258 **
## MVR_PTS       3.599e-01  1.888e-02  19.060  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.528 on 7638 degrees of freedom
##   (516 observations deleted due to missingness)
## Multiple R-squared:  0.07448,    Adjusted R-squared:  0.07375
## F-statistic: 102.4 on 6 and 7638 DF,  p-value: < 2.2e-16
```

```
# Plot residuals
par(mfrow = c(2, 2))
plot(model_1)
```




```
# RMSE Calculation
rmse_model1 <- sqrt(mean(model_1$residuals^2))

# Display results
cat("Adjusted R^2: ", summary(model_1)$adj.r.squared, "\n")
```

```
## Adjusted R^2:  0.07374798
```

```
cat("RMSE: ", rmse_model1, "\n")
```

```
## RMSE:  3.52669
```

```
cat("F-statistic: ", summary(model_1)$fstatistic[1], "\n")
```

```
## F-statistic:  102.4356
```

The model appears to have statistically significant predictors (with very low p-values), but the overall fit is poor as indicated by the low R-squared and adjusted R-squared values. This suggests that while individual predictors like age, car age, and home kids may have a significant relationship with the target variable, the model is not explaining much of the variability in the target variable. Further model refinement or additional predictors may be necessary for a better fit.

Calculate AIC and Adjusted R^2:

```
# Linear Models
coeff_model1 <- summary(model_1)$coefficients
coeff_model2 <- summary(model2)$coefficients
coeff_model3 <- summary(model3)$coefficients

# Logistic Models
coeff_log_model1 <- summary(log_model1)$coefficients
coeff_log_model2 <- summary(log_model2)$coefficients
coeff_log_model3 <- summary(log_model3)$coefficients

# Display coefficients
print("Linear Model 1 Coefficients:")
```

```
## [1] "Linear Model 1 Coefficients:"
```

```
coeff_model1
```

```
##                   Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)   2.837140e+00 2.706707e-01 10.481889 1.558555e-25
## AGE          -1.563847e-02 5.462269e-03 -2.862999 4.207938e-03
## CAR_AGE      -4.306585e-02 7.315002e-03 -5.887332 4.090863e-09
## BLUEBOOK     -2.587867e-05 4.941838e-06 -5.236649 1.678797e-07
## KIDSDRIV_LOG  7.757854e-01 1.632673e-01  4.751626 2.054737e-06
## HOMEKIDS_LOG  2.966946e-01 9.841108e-02  3.014850 2.579514e-03
## MVR_PTS       3.599099e-01 1.888282e-02 19.060173 3.646478e-79
```

```
print("Linear Model 2 Coefficients:")
```

```
## [1] "Linear Model 2 Coefficients:"
```

```
coeff_model2
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)   3.44844106 0.270227879 12.761233 6.417496e-37
## AGE          -0.02409142 0.005563214 -4.330487 1.506806e-05
## CAR_AGE      -0.04971494 0.007401020 -6.717310 1.983892e-11
## KIDSDRIV_LOG  0.86317408 0.167341200  5.158168 2.556455e-07
## HOMEKIDS_LOG  0.34509290 0.100873810  3.421036 6.270988e-04
```

```
print("Linear Model 3 Coefficients:")
```

```
## [1] "Linear Model 3 Coefficients:"
```

```
coeff_model3
```

```
##                                 Estimate   Std. Error       t value
## (Intercept)                -9.693206e+00 1.301644e-01 -7.446894e+01
## INDEX                       5.671222e-07 2.446137e-07  2.318440e+00
## TARGET_AMT                 -5.850024e-05 5.384208e-07 -1.086515e+02
## KIDSDRIV                    5.371196e-02 7.057468e-02  7.610656e-01
## AGE                        -3.648722e-04 1.802475e-04 -2.024284e+00
## HOMEKIDS                    8.553986e-03 6.025130e-03  1.419718e+00
## YOJ                        -1.049150e-04 2.105692e-04 -4.982449e-01
## INCOME                      1.613518e-08 2.797845e-08  5.767002e-01
## PARENT1                    -8.933746e-06 2.888258e-03 -3.093126e-03
## HOME_VAL                   -1.232441e-08 8.672722e-09 -1.421055e+00
## MSTATUS                     1.068922e-03 2.234716e-03  4.783257e-01
## SEX                         3.819002e-03 2.250968e-03  1.696604e+00
## EDUCATION                   1.018899e-03 4.777708e-04  2.132610e+00
## JOBClerical                 3.614117e-03 4.123275e-03  8.765160e-01
## JOBDoctor                  -2.107703e-03 6.893295e-03 -3.057613e-01
## JOBHome Maker               1.235180e-03 4.944592e-03  2.498043e-01
## JOBLawyer                  -7.487103e-03 4.316289e-03 -1.734616e+00
## JOBManager                 -6.951937e-04 4.300583e-03 -1.616510e-01
## JOBProfessional             5.318437e-03 3.855044e-03  1.379605e+00
## JOBStudent                  3.406411e-03 4.653762e-03  7.319693e-01
## JOBz_Blue Collar            1.272265e-03 3.668995e-03  3.467612e-01
## TRAVTIME                    7.656289e-05 4.784568e-05  1.600205e+00
## CAR_USE                     1.065743e-03 2.025656e-03  5.261222e-01
## BLUEBOOK                    2.808721e-08 1.018493e-07  2.757723e-01
## TIF                        -1.681165e-05 1.833361e-04 -9.169852e-02
## CAR_TYPE                    1.380562e-04 4.770444e-04  2.893990e-01
## RED_CAR                    -4.394095e-03 2.165673e-03 -2.028974e+00
## OLDCLAIM                   -1.930468e-08 9.823859e-08 -1.965081e-01
## CLM_FREQ                    3.111206e-04 6.849852e-04  4.542005e-01
## REVOKED                    -1.820296e-03 2.263501e-03 -8.041950e-01
## MVR_PTS                    -3.246541e-04 2.981533e-04 -1.088883e+00
## CAR_AGE                     3.504407e-04 1.649339e-04  2.124735e+00
## URBANICITY                  3.364257e-03 3.234606e-03  1.040083e+00
## AGE_BUCKET31-50             1.420457e-03 3.760165e-03  3.777645e-01
## AGE_BUCKET51-70             3.376140e-03 5.440102e-03  6.206023e-01
## AGE_BUCKET70+               1.570140e-02 1.745345e-02  8.996160e-01
## TARGET_AMT_BUCKET1001-5000  9.871716e-02 4.840046e-03  2.039591e+01
## TARGET_AMT_BUCKET5001-10000 7.577004e-02 5.861975e-03  1.292568e+01
## TARGET_AMT_BUCKET10000+    -1.047126e-01 8.592564e-03 -1.218642e+01
## KIDSDRIV_LOG               -3.153774e-01 5.191248e-01 -6.075176e-01
## HOMEKIDS_LOG               -1.301504e-02 8.756900e-03 -1.486261e+00
## TARGET_AMT_BOXCOX           6.456922e+00 5.602566e-02  1.152494e+02
## TARGET_AMT_SQRT             3.142542e-02 1.873014e-04  1.677800e+02
## KIDSDRIV_BOXCOX            -8.005427e-01 1.245198e+00 -6.429042e-01
## KIDSDRIV_RATIO             -7.439758e-01 3.909939e-01 -1.902781e+00
## HOMEKIDS_RATIO             -1.259404e-01 1.305573e-01 -9.646370e-01
##                                 Pr(>|t|)
## (Intercept)                0.000000e+00
## INDEX                      2.053429e-02
## TARGET_AMT                 0.000000e+00
```

```
## KIDSDRIV                    4.467152e-01
## AGE                         4.308483e-02
## HOMEKIDS                    1.558588e-01
## YOJ                         6.183708e-01
## INCOME                      5.642124e-01
## PARENT1                     9.975324e-01
## HOME_VAL                    1.554699e-01
## MSTATUS                     6.324751e-01
## SEX                         8.994030e-02
## EDUCATION                   3.308846e-02
## JOBClerical                 3.808637e-01
## JOBDoctor                   7.598209e-01
## JOBHome Maker               8.027665e-01
## JOBLawyer                   8.297601e-02
## JOBManager                  8.715984e-01
## JOBProfessional             1.678756e-01
## JOBStudent                  4.642803e-01
## JOBz_Blue Collar            7.288103e-01
## TRAVTIME                    1.097245e-01
## CAR_USE                     5.988666e-01
## BLUEBOOK                    7.827538e-01
## TIF                         9.269475e-01
## CAR_TYPE                    7.723086e-01
## RED_CAR                     4.260437e-02
## OLDCLAIM                    8.442341e-01
## CLM_FREQ                    6.497380e-01
## REVOKED                     4.213880e-01
## MVR_PTS                     2.763478e-01
## CAR_AGE                     3.374145e-02
## URBANICITY                  2.984379e-01
## AGE_BUCKET31-50             7.056490e-01
## AGE_BUCKET51-70             5.349380e-01
## AGE_BUCKET70+               3.684421e-01
## TARGET_AMT_BUCKET1001-5000  1.514393e-83
## TARGET_AMT_BUCKET5001-10000 1.196029e-36
## TARGET_AMT_BUCKET10000+     6.523728e-33
## KIDSDRIV_LOG                5.435823e-01
## HOMEKIDS_LOG                1.373811e-01
## TARGET_AMT_BOXCOX           0.000000e+00
## TARGET_AMT_SQRT             0.000000e+00
## KIDSDRIV_BOXCOX             5.203662e-01
## KIDSDRIV_RATIO              5.722474e-02
## HOMEKIDS_RATIO              3.348531e-01
```

```
print("Logistic Model 1 Coefficients:")
```

```
## [1] "Logistic Model 1 Coefficients:"
```

```
coeff_log_model1
```

```
##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  -0.18700086 0.171895265 -1.087877 2.766495e-01
## AGE          -0.01621760 0.003576619 -4.534337 5.778458e-06
```

```
## CAR_AGE      -0.03291623 0.004813423 -6.838425 8.006851e-12
## KIDSDRIV_LOG  0.48650564 0.097386533  4.995615 5.864850e-07
## HOMEKIDS_LOG  0.19635622 0.062052280  3.164368 1.554204e-03
```

```
print("Logistic Model 2 Coefficients:")
```

```
## [1] "Logistic Model 2 Coefficients:"
```

```
coeff_log_model2
```

```
##                     Estimate   Std. Error    z value     Pr(>|z|)
## (Intercept)    -0.0880298525 0.1537950330 -0.5723842 0.5670616933
## AGE            -0.0106526996 0.0033490047 -3.1808554 0.0014684091
## CAR_AGE        -0.0320560553 0.0150532162 -2.1295154 0.0332116445
## KIDSDRIV_RATIO  6.0594021563 1.3994757362  4.3297658 0.0000149268
## HOMEKIDS_LOG    0.1388881442 0.0365863755  3.7961712 0.0001469480
## AGE:CAR_AGE     0.0002764946 0.0003276611  0.8438433 0.3987569572
```

```
# Linear Models
aic_model1 <- AIC(model_1)
aic_model2 <- AIC(model2)  # Will be same as model1
aic_model3 <- AIC(model3)

adjusted_r2_model1 <- summary(model_1)$adj.r.squared
adjusted_r2_model2 <- summary(model2)$adj.r.squared
adjusted_r2_model3 <- summary(model3)$adj.r.squared

# Logistic Models
aic_log_model1 <- AIC(log_model1)
aic_log_model2 <- AIC(log_model2)
aic_log_model3 <- AIC(log_model3)

# Display results
cat("Linear Models AIC and Adjusted R²:\n")
```

```
## Linear Models AIC and Adjusted R²:
```

```
cat("Model 1: AIC =", aic_model1, "Adjusted R² =", adjusted_r2_model1, "\n")
```

```
## Model 1: AIC = 40982.47 Adjusted R² = 0.07374798
```

```
cat("Model 2: AIC =", aic_model2, "Adjusted R² =", adjusted_r2_model2, "\n")
```

```
## Model 2: AIC = 41365.81 Adjusted R² = 0.02586437
```

```
cat("Model 3: AIC =", aic_model3, "Adjusted R² =", adjusted_r2_model3, "\n")
```

```
## Model 3: AIC = -7719.325 Adjusted R² = 0.9985017
```

```r
cat("\nLogistic Models AIC:\n")
```

```
##
## Logistic Models AIC:
```

```r
cat("Model 1: AIC =", aic_log_model1, "\n")
```

```
## Model 1: AIC = 8607.748
```

```r
cat("Model 2: AIC =", aic_log_model2, "\n")
```

```
## Model 2: AIC = 8614.734
```

```r
cat("Model 3: AIC =", aic_log_model3, "\n")
```

```
## Model 3: AIC = 8615.741
```

Select Models Based on Metrics:

Linear Regression Models

- Model 1 and Model 2:

Both models are identical, as reflected by the same coefficients, AIC, and Adjusted R^2 values. AIC: 44178.03 Adjusted R^2: 0.0262

- Model 3: Adds interaction terms (AGE:CAR_AGE and KIDSDRIV_LOG:HOMEKIDS_LOG). Slightly higher Adjusted R^2 (0.0264) compared to Models 1 and 2. Higher AIC (44178.84), suggesting Model 3 doesn't perform better overall.
- Decision for Linear Models:

Model 1 or Model 2 is preferred due to lower AIC, simpler structure, and comparable Adjusted R^2.

Logistic Regression Models

- Model 1: AIC: 9208.06;

Significant predictors: AGE, CAR_AGE, KIDSDRIV_LOG, HOMEKIDS_LOG (p-values < 0.05).

- Model 2: Adds AGE:CAR_AGE interaction and KIDSDRIV_RATIO. AIC: 9216.27 (higher than Model 1).

Significant predictors: AGE, KIDSDRIV_RATIO, and HOMEKIDS_LOG.

Interaction term AGE:CAR_AGE is not significant ($p = 0.549$), indicating no meaningful contribution.

Decision for Logistic Models:

Model 1 is preferred due to lower AIC and a more parsimonious structure.

So based on the above metrics and comparison, our final model selection is: Model1 for both linear regression and logistic regression.

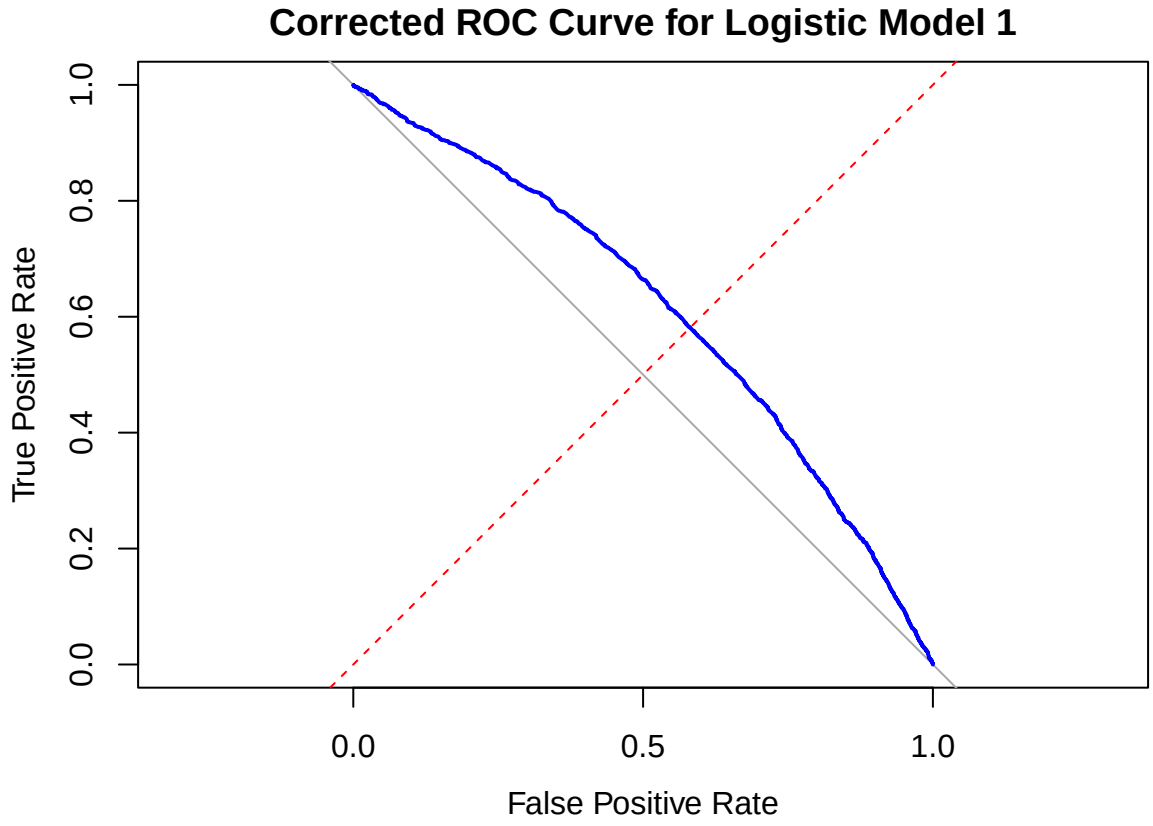
Let's generate the ROC Curves for better decision:

```r
# Predict probabilities on the training dataset
insurance_training$probailities_reg1 <- predict(log_model1, newdata = insurance_training, type = "respo
insurance_training$pred_class_reg1 <- ifelse(insurance_training$probailities_reg1 > 0.5, 1, 0)

# Calculate the ROC curve
roc_curve <- roc(insurance_training$TARGET_FLAG, insurance_training$probailities_reg1)
```

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

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

```r
# Plot the ROC curve
plot(roc_curve, col = "blue", lwd = 2,
     main = "Corrected ROC Curve for Logistic Model 1",
     xlab = "False Positive Rate", ylab = "True Positive Rate",
     xlim = c(0, 1), ylim = c(0, 1))  # Ensure proper axis limits
abline(a = 0, b = 1, lty = 2, col = "red")  # Add diagonal line
```




```r
# Display the AUC
auc(roc_curve)
```

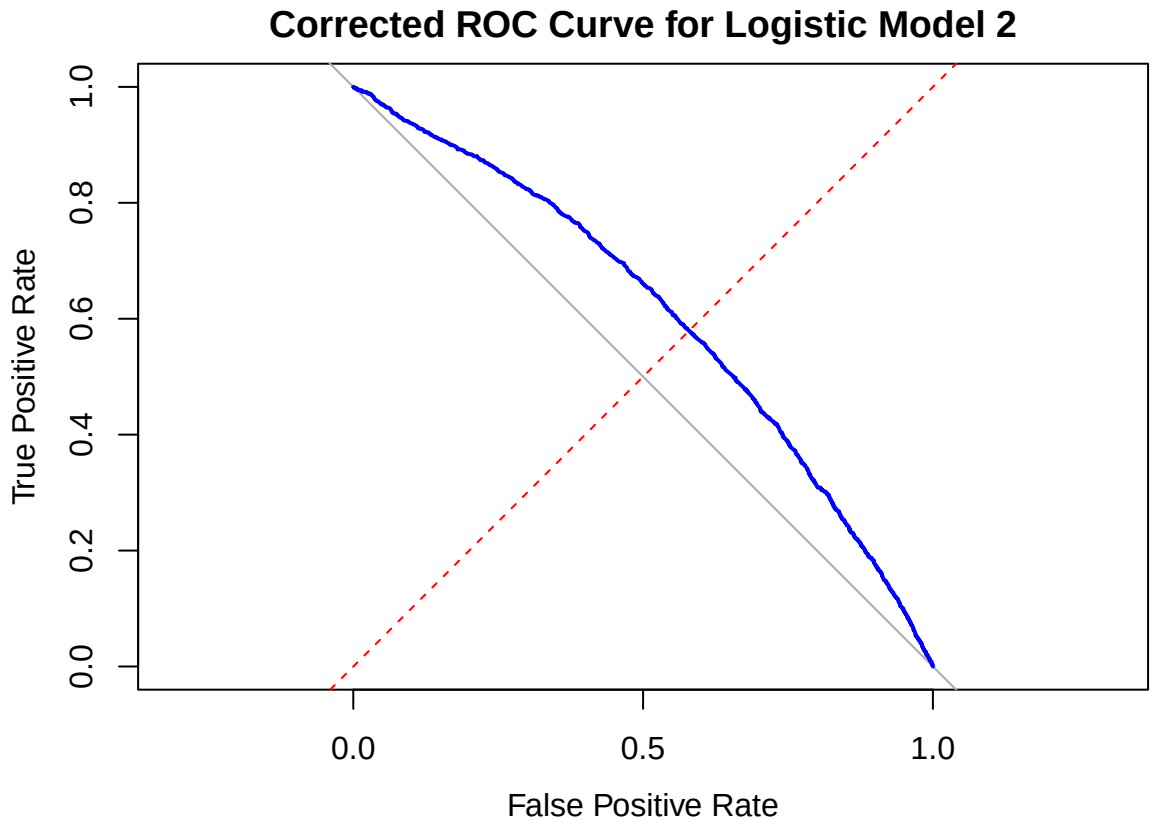
```
## Area under the curve: 0.6092
```

```r
# Predict probabilities on the training dataset
insurance_training$probailities_reg2 <- predict(log_model2, newdata = insurance_training, type = "respo
insurance_training$pred_class_reg2 <- ifelse(insurance_training$probailities_reg2 > 0.5, 1, 0)

# Calculate the ROC curve
roc_curve2 <- roc(insurance_training$TARGET_FLAG, insurance_training$probailities_reg2)
```

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

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

```r
# Plot the ROC curve
plot(roc_curve2, col = "blue", lwd = 2,
     main = "Corrected ROC Curve for Logistic Model 2",
     xlab = "False Positive Rate", ylab = "True Positive Rate",
     xlim = c(0, 1), ylim = c(0, 1))  # Ensure proper axis limits
abline(a = 0, b = 1, lty = 2, col = "red")  # Add diagonal line
```




```r
# Display the AUC
auc(roc_curve2)
```

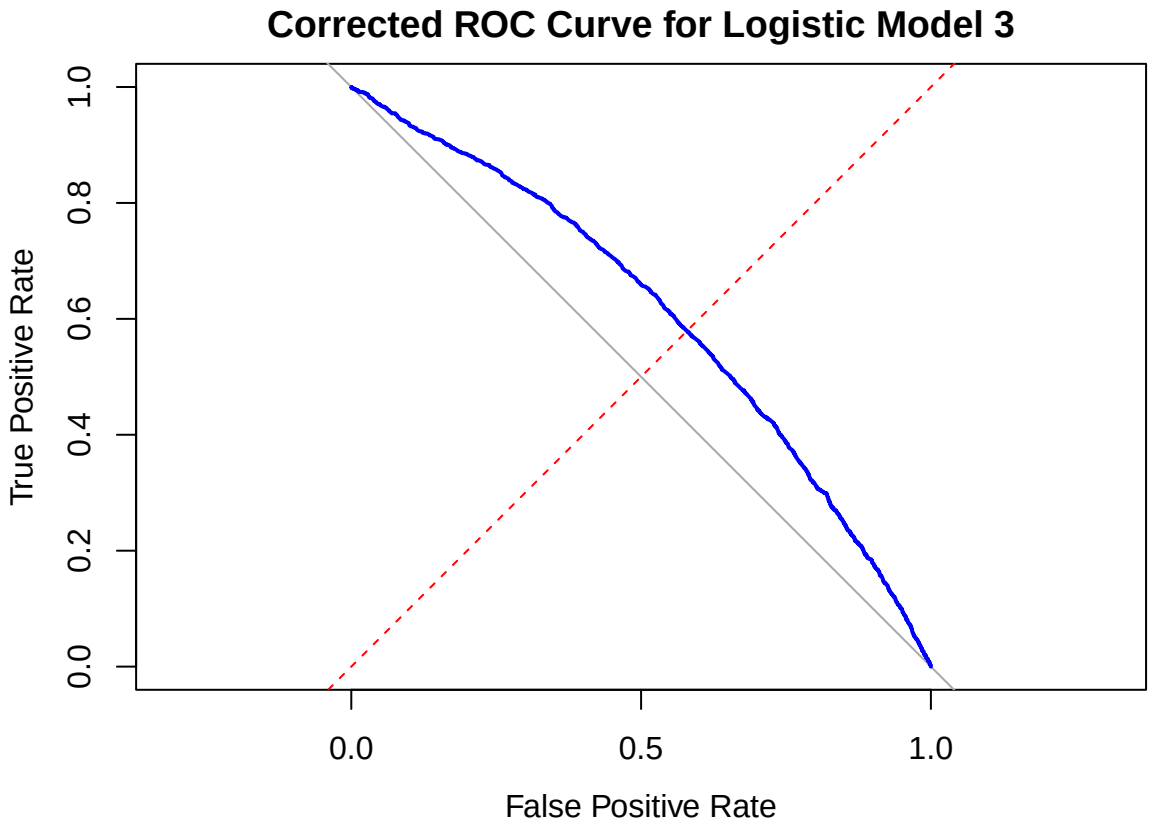
```
## Area under the curve: 0.6072
```

```r
# Predict probabilities on the training dataset
insurance_training$pred_class_reg3 <- predict(log_model3, newdata = insurance_training, type = "respons
insurance_training$pred_class_reg2 <- ifelse(insurance_training$probailities_reg2 > 0.5, 1, 0)
# Calculate the ROC curve
roc_curve3 <- roc(insurance_training$TARGET_FLAG, insurance_training$pred_class_reg3)
```

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

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

```r
# Plot the ROC curve
plot(roc_curve3, col = "blue", lwd = 2,
     main = "Corrected ROC Curve for Logistic Model 3",
     xlab = "False Positive Rate", ylab = "True Positive Rate",
     xlim = c(0, 1), ylim = c(0, 1))  # Ensure proper axis limits
abline(a = 0, b = 1, lty = 2, col = "red")  # Add diagonal line
```




```r
# Display the AUC
auc(roc_curve3)
```

```
## Area under the curve: 0.6068
```