homework4

Alice Ding

2023-11-28

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

Overview

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

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

- `INDEX`: Identification Variable (do not use)
- `TARGET_FLAG`: Was Car in a crash? 1=YES 0=NO
- `TARGET_AMT`: If car was in a crash, what was the cost
- `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

Data Exploration

First, we'll view the summary and then we'll check if there are data points missing. Then, we'll clean the fields up to make sure they're ready for analysis.

```
training <- read.csv('https://raw.githubusercontent.com/addsding/data621/main/homework4/insurance_train
evaluation <- read.csv('https://raw.githubusercontent.com/addsding/data621/main/homework4/insurance-eva

summary <- as.data.frame(describe(training))
nulls <- 8161 - summary['n']
nulls_pct <- nulls / 8161
summary['nulls'] <- nulls
summary['nulls_pct'] <- nulls_pct
kable(summary, digits=2) |>
  kable_styling(c("striped", "scale_down")) |>
  scroll_box(width = "100%")
```

The data has 26 variables with 8161 observations

It looks like the only fields with nulls are `YOJ`, `AGE`, and `CAR_AGE` so good to know there won't be much cleaning there.

It appears we also have a few highly skewed variables due to many medians being quite different from the means. Some examples include the variables `OLDCLAIM` and potentially `HOME_VAL`.

What types of fields are each of our variables?

```
summary(training)
```

```
##      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
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
INDEX	1	8161	5151.87	2978.89	5133	5151.93	3841.42	1	10302.0	10301.0	0.00
TARGET_FLAG	2	8161	0.26	0.44	0	0.20	0.00	0	1.0	1.0	1.07
TARGET_AMT	3	8161	1504.32	4704.03	0	593.71	0.00	0	107586.1	107586.1	8.71
KIDSDRIV	4	8161	0.17	0.51	0	0.03	0.00	0	4.0	4.0	3.35
AGE	5	8155	44.79	8.63	45	44.83	8.90	16	81.0	65.0	-0.03
HOMEKIDS	6	8161	0.72	1.12	0	0.50	0.00	0	5.0	5.0	1.34
YOJ	7	7707	10.50	4.09	11	11.07	2.97	0	23.0	23.0	-1.20
INCOME*	8	8161	2875.55	2090.68	2817	2816.95	2799.15	1	6613.0	6612.0	0.11
PARENT1*	9	8161	1.13	0.34	1	1.04	0.00	1	2.0	1.0	2.17
HOME_VAL*	10	8161	1684.89	1697.38	1245	1516.50	1842.87	1	5107.0	5106.0	0.52
MSTATUS*	11	8161	1.40	0.49	1	1.38	0.00	1	2.0	1.0	0.41
SEX*	12	8161	1.54	0.50	2	1.55	0.00	1	2.0	1.0	-0.14
EDUCATION*	13	8161	3.09	1.44	3	3.11	1.48	1	5.0	4.0	0.12
JOB*	14	8161	5.69	2.68	6	5.81	2.97	1	9.0	8.0	-0.31
TRAVTIME	15	8161	33.49	15.91	33	33.00	16.31	5	142.0	137.0	0.45
CAR_USE*	16	8161	1.63	0.48	2	1.66	0.00	1	2.0	1.0	-0.53
BLUEBOOK*	17	8161	1283.62	893.51	1124	1259.57	1132.71	1	2789.0	2788.0	0.25
TIF	18	8161	5.35	4.15	4	4.84	4.45	1	25.0	24.0	0.89
CAR_TYPE*	19	8161	3.53	1.97	3	3.54	2.97	1	6.0	5.0	0.00
RED_CAR*	20	8161	1.29	0.45	1	1.24	0.00	1	2.0	1.0	0.92
OLDCLAIM*	21	8161	552.27	862.20	1	380.32	0.00	1	2857.0	2856.0	1.31
CLM_FREQ	22	8161	0.80	1.16	0	0.59	0.00	0	5.0	5.0	1.21
REVOKED*	23	8161	1.12	0.33	1	1.03	0.00	1	2.0	1.0	2.30
MVR_PTS	24	8161	1.70	2.15	1	1.31	1.48	0	13.0	13.0	1.35
CAR_AGE	25	7651	8.33	5.70	8	7.96	7.41	-3	28.0	31.0	0.28
URBANICITY*	26	8161	1.20	0.40	1	1.13	0.00	1	2.0	1.0	1.46

```
##  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
```

It looks like we have 12 continuous variables and the rest are characters. We however see that certain fields should be numerical, however they're formatted in a way that makes them characters. These fields are:

- INCOME
- HOME_VAL
- BLUEBOOK
- OLDCLAIM

We can clean these up next.

Data Cleaning

Formatting

It looks as though a few of these numerical values have $ and commas, signalling they would be difficult to interpret when charting or trying to visualize them. Let's try to address these fields and fix them.

```
clean <- function(x){
  x <- as.character(x)
  x <- gsub(",", "", x)
  x <- gsub("\\$", "", x)
  as.numeric(x)
}

training$INCOME_CLEAN <- clean(training$INCOME)
training$HOME_VAL_CLEAN <- clean(training$HOME_VAL)
training$BLUEBOOK_CLEAN <- clean(training$BLUEBOOK)
training$OLDCLAIM_CLEAN <- clean(training$OLDCLAIM)

summary(training)
```

```
##      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         INCOME_CLEAN    HOME_VAL_CLEAN
##  Min.   :-3.000   Length:8161        Min.   :     0   Min.   :     0
##  1st Qu.: 1.000   Class :character   1st Qu.: 28097   1st Qu.:     0
##  Median : 8.000   Mode  :character   Median : 54028   Median :161160
##  Mean   : 8.328                      Mean   : 61898   Mean   :154867
##  3rd Qu.:12.000                      3rd Qu.: 85986   3rd Qu.:238724
##  Max.   :28.000                      Max.   :367030   Max.   :885282
##  NA's   :510                         NA's   :445      NA's   :464
##  BLUEBOOK_CLEAN  OLDCLAIM_CLEAN
##  Min.   : 1500   Min.   :    0
##  1st Qu.: 9280   1st Qu.:    0
##  Median :14440   Median :    0
##  Mean   :15710   Mean   : 4037
##  3rd Qu.:20850   3rd Qu.: 4636
##  Max.   :69740   Max.   :57037
##
```

Data Types

It seems as though a few fields should be identified as factors rather than characters – these include:

- TARGET_FLAG
- CAR_USE
- CAR_TYPE
- EDUCATION
- JOB
- MSTATUS
- RED_CAR
- PARENT1
- REVOKED
- SEX
- URBANICITY

```
training$TARGET_FLAG <- as.factor(training$TARGET_FLAG)
training$CAR_USE <- as.factor(training$CAR_USE)
training$CAR_TYPE <- as.factor(training$CAR_TYPE)
training$JOB <- as.factor(training$JOB)
training$MSTATUS <- as.factor(training$MSTATUS)
training$RED_CAR <- as.factor(training$RED_CAR)
training$PARENT1 <- as.factor(training$PARENT1)
```

```r
training$REVOKED <- as.factor(training$REVOKED)
training$SEX <- as.factor(training$SEX)
training$URBANICITY <- as.factor(training$URBANICITY)
summary(training)
```

```
##      INDEX       TARGET_FLAG   TARGET_AMT        KIDSDRIV           AGE
##  Min.   :    1   0:6008      Min.   :     0   Min.   :0.0000   Min.   :16.00
##  1st Qu.: 2559   1:2153      1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
##  Median : 5133               Median :     0   Median :0.0000   Median :45.00
##  Mean   : 5152               Mean   :  1504   Mean   :0.1711   Mean   :44.79
##  3rd Qu.: 7745               3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :10302               Max.   :107586   Max.   :4.0000   Max.   :81.00
##                                                                NA's   :6
##     HOMEKIDS           YOJ          INCOME          PARENT1
##  Min.   :0.0000   Min.   : 0.0   Length:8161        No :7084
##  1st Qu.:0.0000   1st Qu.: 9.0   Class :character   Yes:1077
##  Median :0.0000   Median :11.0   Mode  :character
##  Mean   :0.7212   Mean   :10.5
##  3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :5.0000   Max.   :23.0
##                   NA's   :454
##    HOME_VAL         MSTATUS      SEX        EDUCATION
##  Length:8161        Yes :4894   M  :3786   Length:8161
##  Class :character   z_No:3267   z_F:4375   Class :character
##  Mode  :character                          Mode  :character
##
##
##
##
##              JOB          TRAVTIME            CAR_USE       BLUEBOOK
##  z_Blue Collar:1825   Min.   :  5.00   Commercial:3029   Length:8161
##  Clerical     :1271   1st Qu.: 22.00   Private   :5132   Class :character
##  Professional :1117   Median : 33.00                     Mode  :character
##  Manager      : 988   Mean   : 33.49
##  Lawyer       : 835   3rd Qu.: 44.00
##  Student      : 712   Max.   :142.00
##  (Other)      :1413
##       TIF               CAR_TYPE    RED_CAR      OLDCLAIM
##  Min.   : 1.000   Minivan    :2145   no :5783   Length:8161
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   Class :character
##  Median : 4.000   Pickup     :1389              Mode  :character
##  Mean   : 5.351   Sports Car : 907
##  3rd Qu.: 7.000   Van        : 750
##  Max.   :25.000   z_SUV      :2294
##
##     CLM_FREQ       REVOKED        MVR_PTS          CAR_AGE
##  Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000              Median : 1.000   Median : 8.000
##  Mean   :0.7986              Mean   : 1.696   Mean   : 8.328
##  3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000              Max.   :13.000   Max.   :28.000
##                                               NA's   :510
```

```
##                   URBANICITY    INCOME_CLEAN    HOME_VAL_CLEAN   BLUEBOOK_CLEAN
##  Highly Urban/ Urban  :6492   Min.   :     0   Min.   :     0   Min.   : 1500
##  z_Highly Rural/ Rural:1669   1st Qu.: 28097   1st Qu.:     0   1st Qu.: 9280
##                               Median : 54028   Median :161160   Median :14440
##                               Mean   : 61898   Mean   :154867   Mean   :15710
##                               3rd Qu.: 85986   3rd Qu.:238724   3rd Qu.:20850
##                               Max.   :367030   Max.   :885282   Max.   :69740
##                               NA's   :445      NA's   :464
##  OLDCLAIM_CLEAN
##  Min.   :    0
##  1st Qu.:    0
##  Median :    0
##  Mean   : 4037
##  3rd Qu.: 4636
##  Max.   :57037
##
```

```
summary <- as.data.frame(describe(training))
nulls <- 8161 - summary['n']
nulls_pct <- nulls / 8161
summary['nulls'] <- nulls
summary['nulls_pct'] <- nulls_pct
kable(summary, digits=2) |>
  kable_styling(c("striped", "scale_down")) |>
  scroll_box(width = "100%")
```

Class Bias Check

For our binary logistic regression model, we only have two target values: 0 and 1. We ideally want an equal representation of both as if imbalance were to deviate, our model performance would suffer both from effects of differential variance between the classes and bias, thus picking the more represented class. For logistic regression, if we see a strong imbalance, we can:

- up-sample the smaller group (e.g. bootstrapping),
- down-sample the larger group (e.g. sampling or bootstrapping)
- adjust our threshold for assigning the predicted value away from 0.5.

What is the exact distribution of `TARGET_FLAG`?

```
table(training$TARGET_FLAG)
```

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

This unfortunately does not look like an even split as 0 is represented more heavily here and thus this would affect our model. To help alleviate this bias, we'll be up-sampling the smaller group.

```
set.seed(123)
training_fixed <- upSample(x=training[, -ncol(training)],
                     y=as.factor(training$TARGET_FLAG))
table(training_fixed$TARGET_FLAG)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	
INDEX	1	8161	5151.87	2978.89	5133	5151.93	3841.42	1	10302.0	10
TARGET_FLAG*	2	8161	1.26	0.44	1	1.20	0.00	1	2.0	
TARGET_AMT	3	8161	1504.32	4704.03	0	593.71	0.00	0	107586.1	107
KIDSDRIV	4	8161	0.17	0.51	0	0.03	0.00	0	4.0	
AGE	5	8155	44.79	8.63	45	44.83	8.90	16	81.0	
HOMEKIDS	6	8161	0.72	1.12	0	0.50	0.00	0	5.0	
YOJ	7	7707	10.50	4.09	11	11.07	2.97	0	23.0	
INCOME*	8	8161	2875.55	2090.68	2817	2816.95	2799.15	1	6613.0	6
PARENT1*	9	8161	1.13	0.34	1	1.04	0.00	1	2.0	
HOME_VAL*	10	8161	1684.89	1697.38	1245	1516.50	1842.87	1	5107.0	5
MSTATUS*	11	8161	1.40	0.49	1	1.38	0.00	1	2.0	
SEX*	12	8161	1.54	0.50	2	1.55	0.00	1	2.0	
EDUCATION*	13	8161	3.09	1.44	3	3.11	1.48	1	5.0	
JOB*	14	8161	5.69	2.68	6	5.81	2.97	1	9.0	
TRAVTIME	15	8161	33.49	15.91	33	33.00	16.31	5	142.0	
CAR_USE*	16	8161	1.63	0.48	2	1.66	0.00	1	2.0	
BLUEBOOK*	17	8161	1283.62	893.51	1124	1259.57	1132.71	1	2789.0	2
TIF	18	8161	5.35	4.15	4	4.84	4.45	1	25.0	
CAR_TYPE*	19	8161	3.53	1.97	3	3.54	2.97	1	6.0	
RED_CAR*	20	8161	1.29	0.45	1	1.24	0.00	1	2.0	
OLDCLAIM*	21	8161	552.27	862.20	1	380.32	0.00	1	2857.0	2
CLM_FREQ	22	8161	0.80	1.16	0	0.59	0.00	0	5.0	
REVOKED*	23	8161	1.12	0.33	1	1.03	0.00	1	2.0	
MVR_PTS	24	8161	1.70	2.15	1	1.31	1.48	0	13.0	
CAR_AGE	25	7651	8.33	5.70	8	7.96	7.41	-3	28.0	
URBANICITY*	26	8161	1.20	0.40	1	1.13	0.00	1	2.0	
INCOME_CLEAN	27	7716	61898.09	47572.68	54028	56840.98	41792.27	0	367030.0	367
HOME_VAL_CLEAN	28	7697	154867.29	129123.77	161160	144032.07	147867.11	0	885282.0	885
BLUEBOOK_CLEAN	29	8161	15709.90	8419.73	14440	15036.89	8450.82	1500	69740.0	68
OLDCLAIM_CLEAN	30	8161	4037.08	8777.14	0	1719.29	0.00	0	57037.0	57

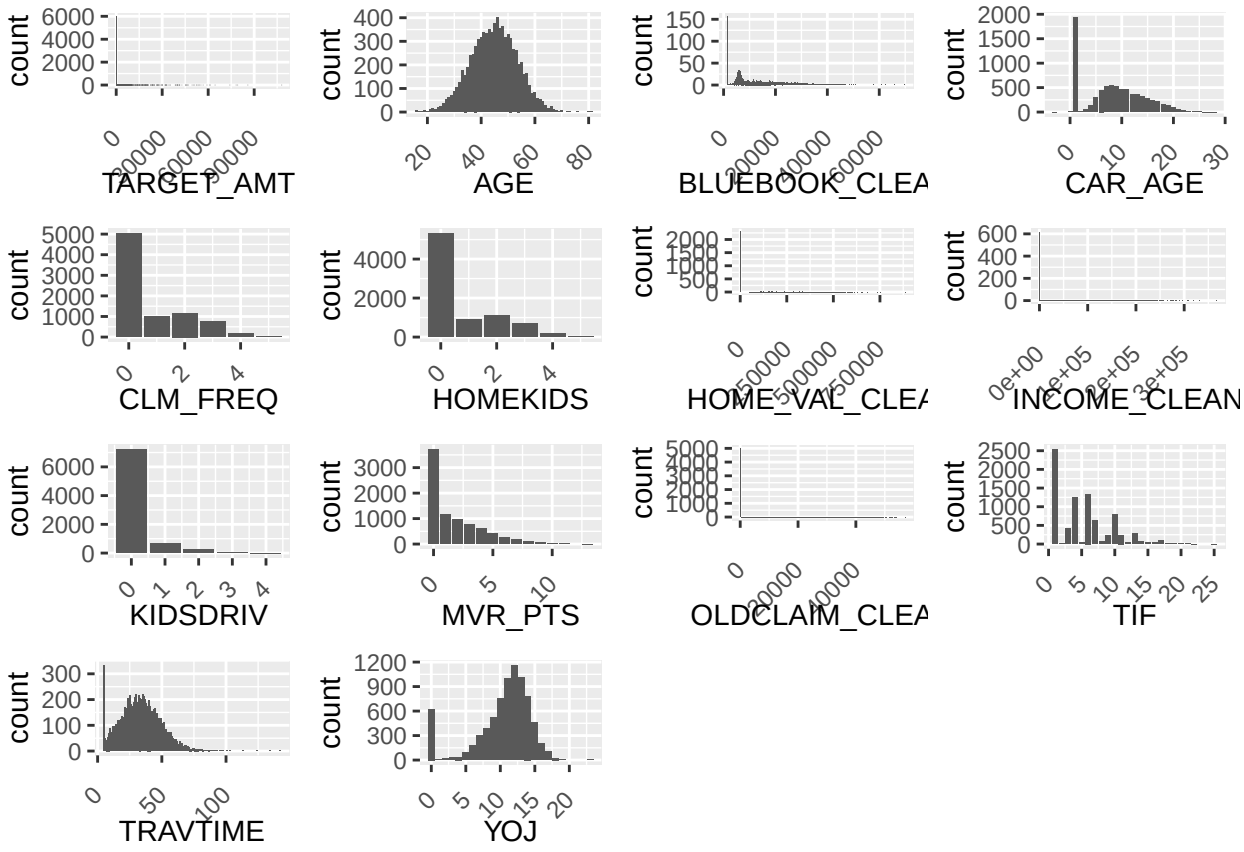
```
##
##    0    1
## 6008 6008
```

Perfect 50/50 split!

Distributions

Numerical Fields

Let's see what all of the numerical fields look like distribution wise.

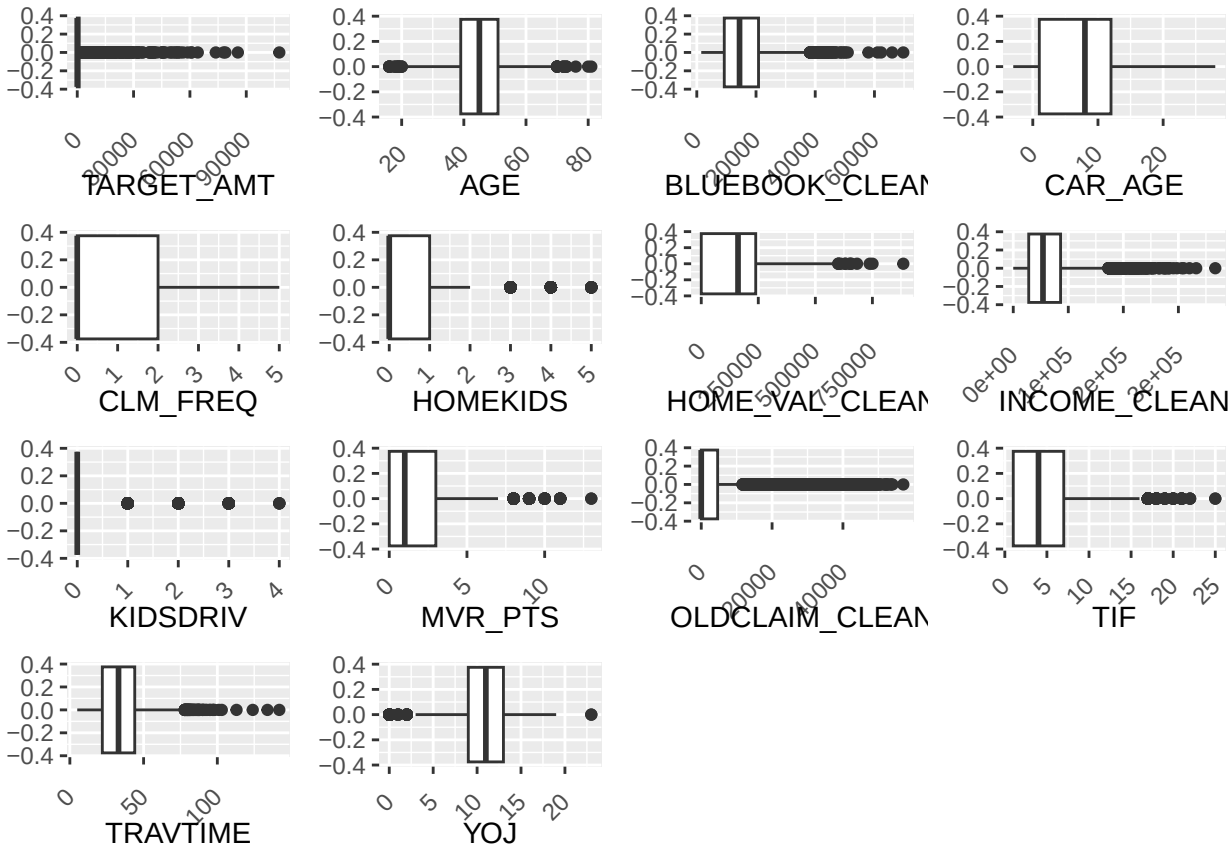


At first glance, it looks like these fields are relatively normal or have a good curve:

- `AGE`
- `CAR_AGE`
- `TRAVTIME`
- `YOJ`

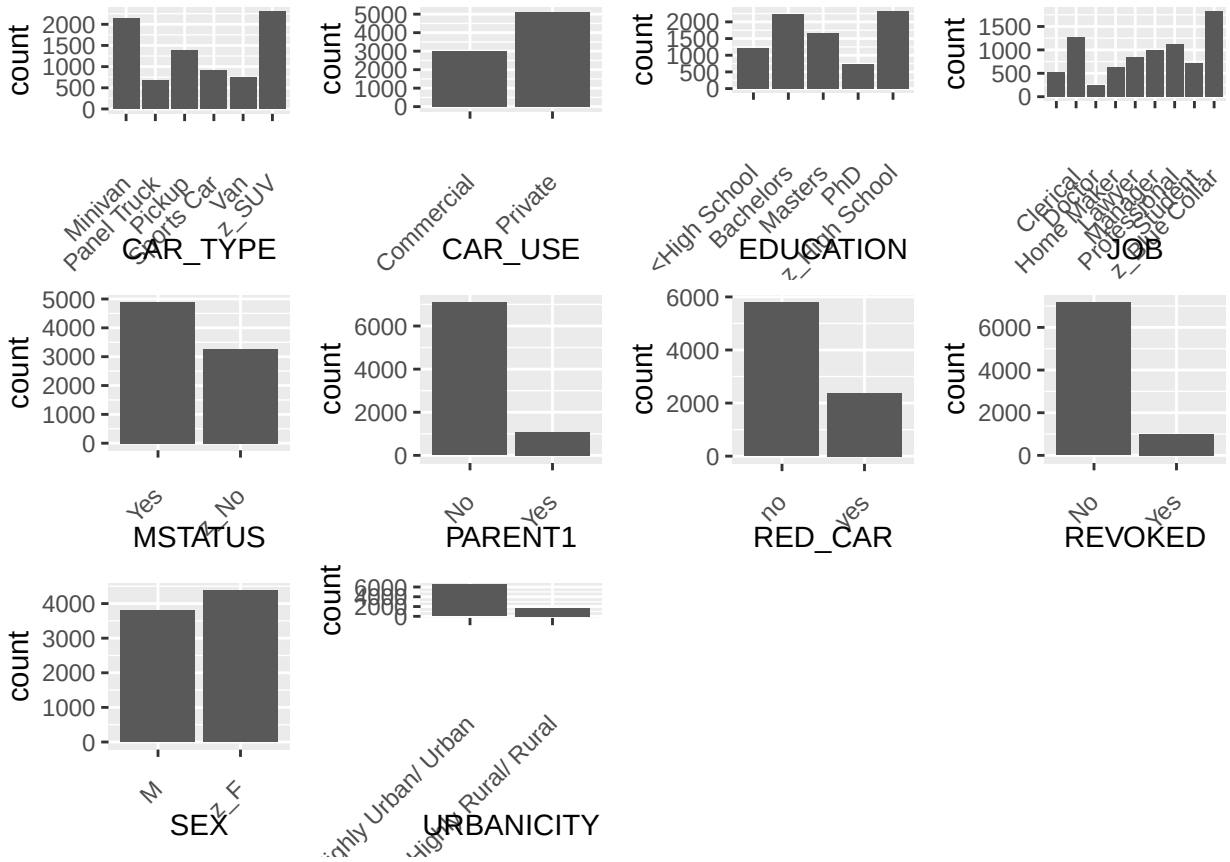
The rest either are pretty skewed in either direction or have no pattern really at all.

How do these look as boxplots?



In most of these fields, it looks like there are quite a few outliers except in `CAR_AGE` and `CLM_FREQ`.

Categorical Fields



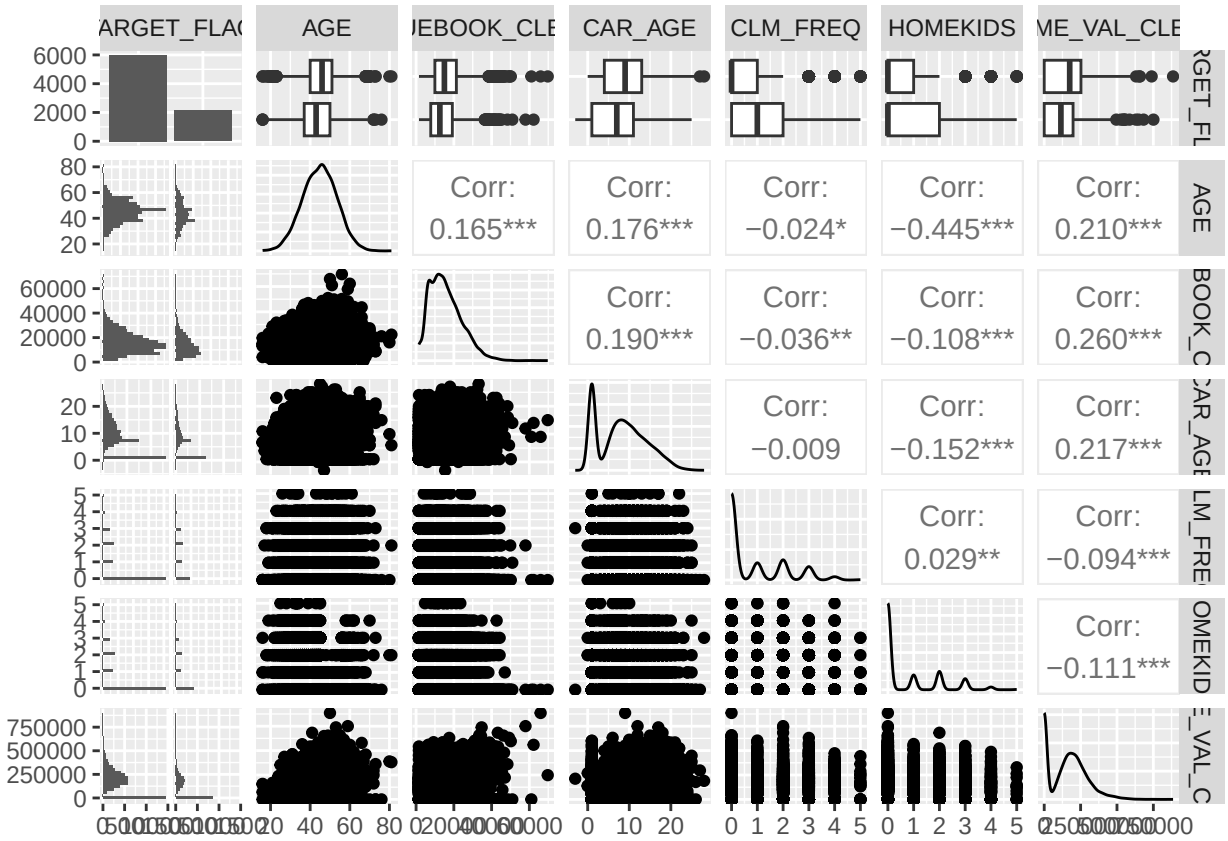
It looks like some of these fields are more heavily skewed towards one category vs. the others, in particular `PARENT1` and `REVOKED`.

Now that we have a sense of how the data is distributed, what do the relationships between the variables as well as with our target look like?

Correlations

Target Flag

Let's see how each of the numerical fields correlate with `TARGET_FLAG` – we'll start with the first six fields.



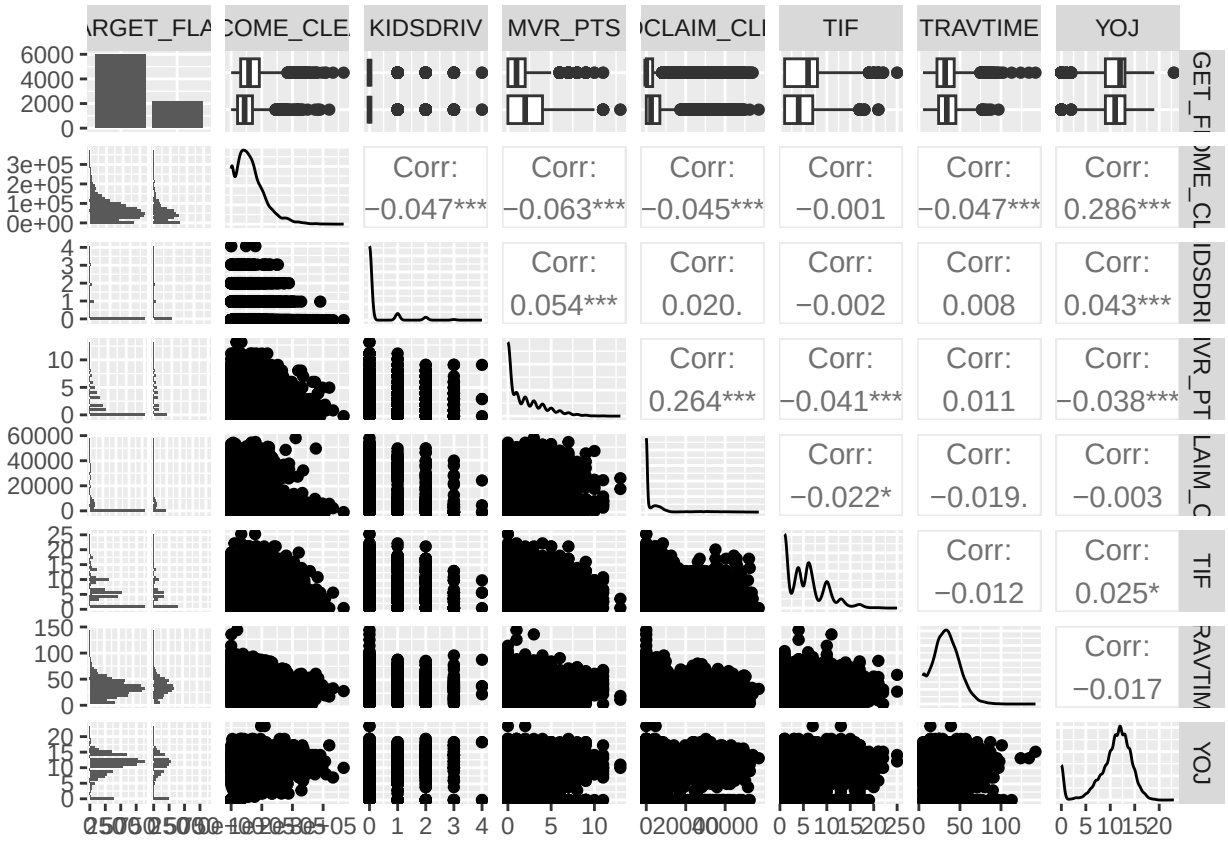
Interestingly, it seems that every field is significantly correlated except for `CAR_AGE`. The one with negative correlation is `HOMEKIDS` while `AGE`, `BLUEBOOK_CLEAN`, and `CLM_FREQ` are positively correlated.

Implication wise:

- `CAR_AGE` - not correlated - makes sense as this was theoretically known to have an unknown effect and moreso would be used for the `TARGET_AMT` field
- `HOMEKIDS` - negative effect - if you have kids, you could be more likely to be a responsible driver
- `AGE` - positive effect - could make sense as older people are theoretically more likely to be risky
- `BLUEBOOK_CLEAN` - positive effect - this is interesting as this seemed to be unknown
- `CLM_FREQ` - positive effect - the fact that they've had claims previously is maybe indicative of future accidents

These fields are also pretty correlated with one another for the most part which may serve as an issue for our model.

What do the relationships look like for the rest of the fields?

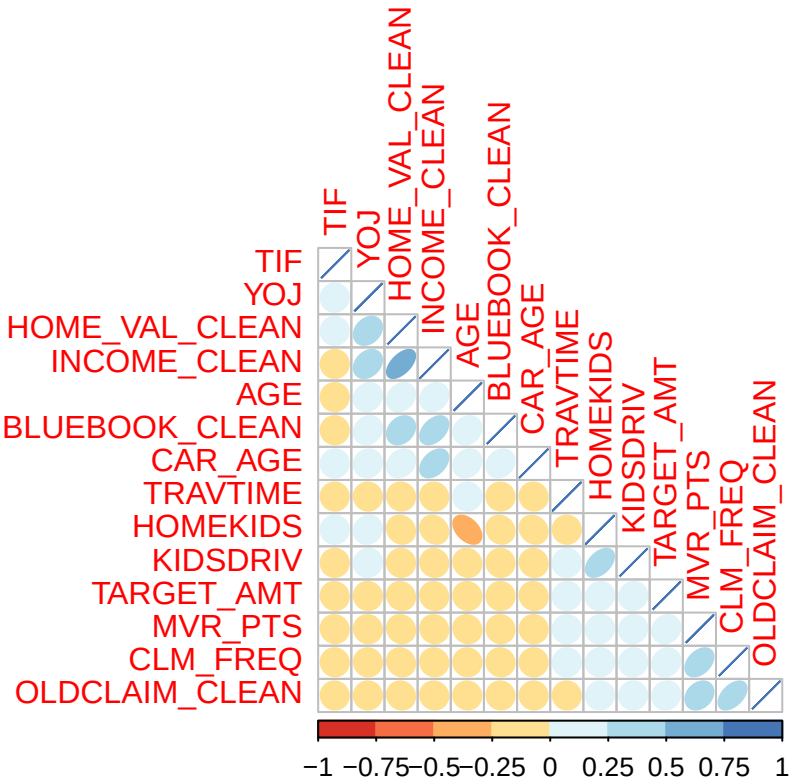


Interestingly, it seems that every field is significantly correlated except for `TIF` and `URBANICITY`. The ones with negative correlation are `INCOME_CLEAN` and `OLDCLAIM_CLEAN` while `KIDSDRIV`, and `MVR_PTS` are positively correlated.

Implication wise:

- `TIF` - not correlated - interesting how this is not correlated as there was a theory that this would indicate people are safer the longer they've been customers
- `URBANICITY` - not correlated - there was no expectation for this to be correlated so this is not surprising
- `INCOME_CLEAN` - negative effect - this makes sense as in theory, those with a higher income tend to get in fewer crashes
- `OLDCLAIM_CLEAN` - negative effect - there was no expectation for this to be correlated, so it's interesting how it's super significant
- `KIDSDRIV` - positive effect - this makes sense as a positive impact given teenagers/younger people are more likely to be in a crash
- `MVR_PTS` - positive effect - this makes sense as a positive impact given the more crashes you have, the more wreckless you are
- `CLM_FREQ` - positive effect - there was no expectation for this to be correlated so this is not surprising

There are some correlated fields here, but let's see if they're also correlated with each other beyond just the seven displayed here.



Given none of these are extremely correlated, we probably won't have to remove fields in the preparation phase!

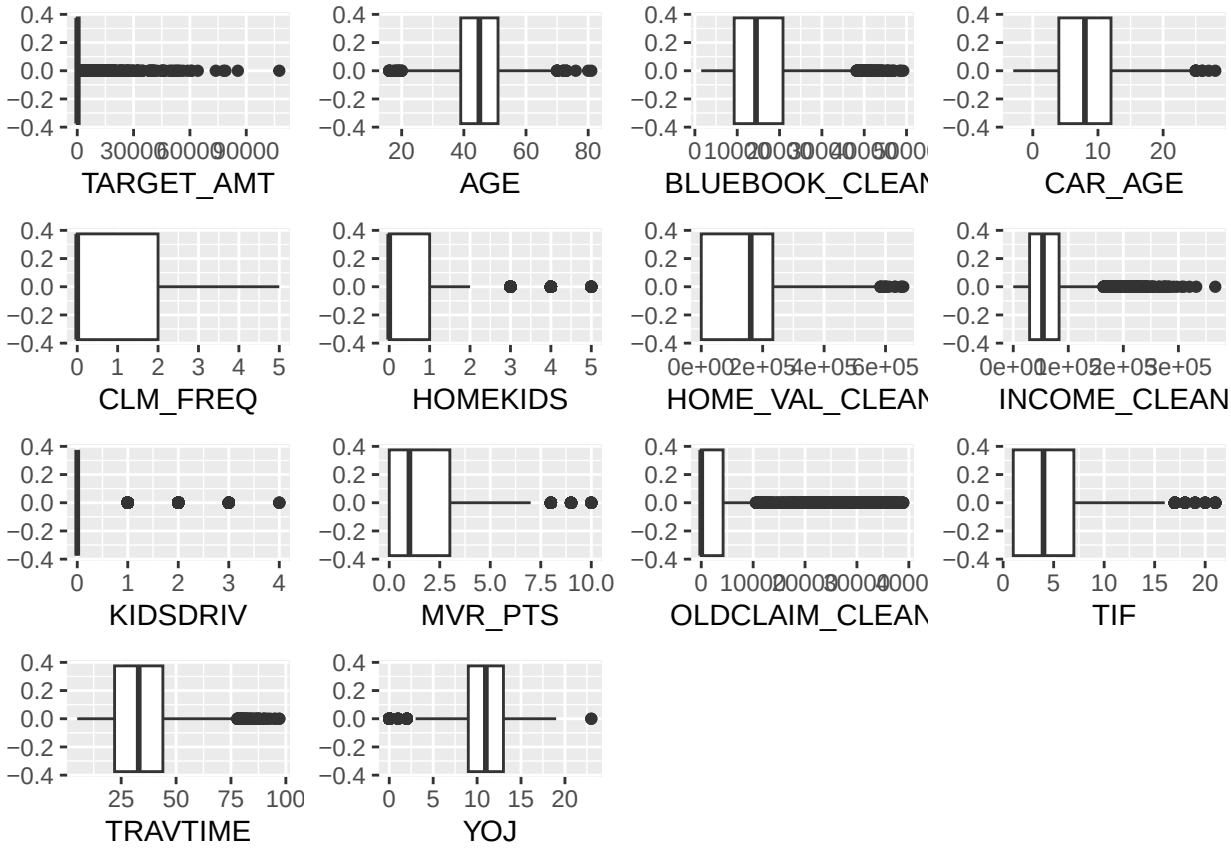
Data Preparation

Outliers & Nulls

There are some pretty extreme outliers and some nulls scattered throughout various numerical fields (see the boxplot in the previous section). To account for these, we will use the median of the data again to replace these outliers if they are more than four standard deviations from the mean or if they're null.

What we decided to do with each field:

- `AGE`: Not replacing outliers as age is not something we should change
- `BLUEBOOK_CLEAN`: Replacing 8 outliers with the median
- `CAR_AGE`: Imputing 510 nulls with the median
- `CLM_FREQ`: No outliers
- `HOMEKIDS`: No outliers
- `HOME_VAL_CLEAN`: Imputing 468 total fields with the median, 464 nulls and 4 outliers
- `INCOME_CLEAN`: Imputing 445 nulls with the median
- `KIDSDRIV`: Not replacing outliers as they're only values 3-4
- `MVR_PTS`: Replacing 13 outliers with the median
- `OLDCLAIM_CLEAN`: Replacing 144 outliers with the median
- `TIF`: Replacing 5 outliers with the median
- `TRAVTIME`: Replacing 7 outliers with the median
- `YOJ`: Imputing 454 nulls with the median



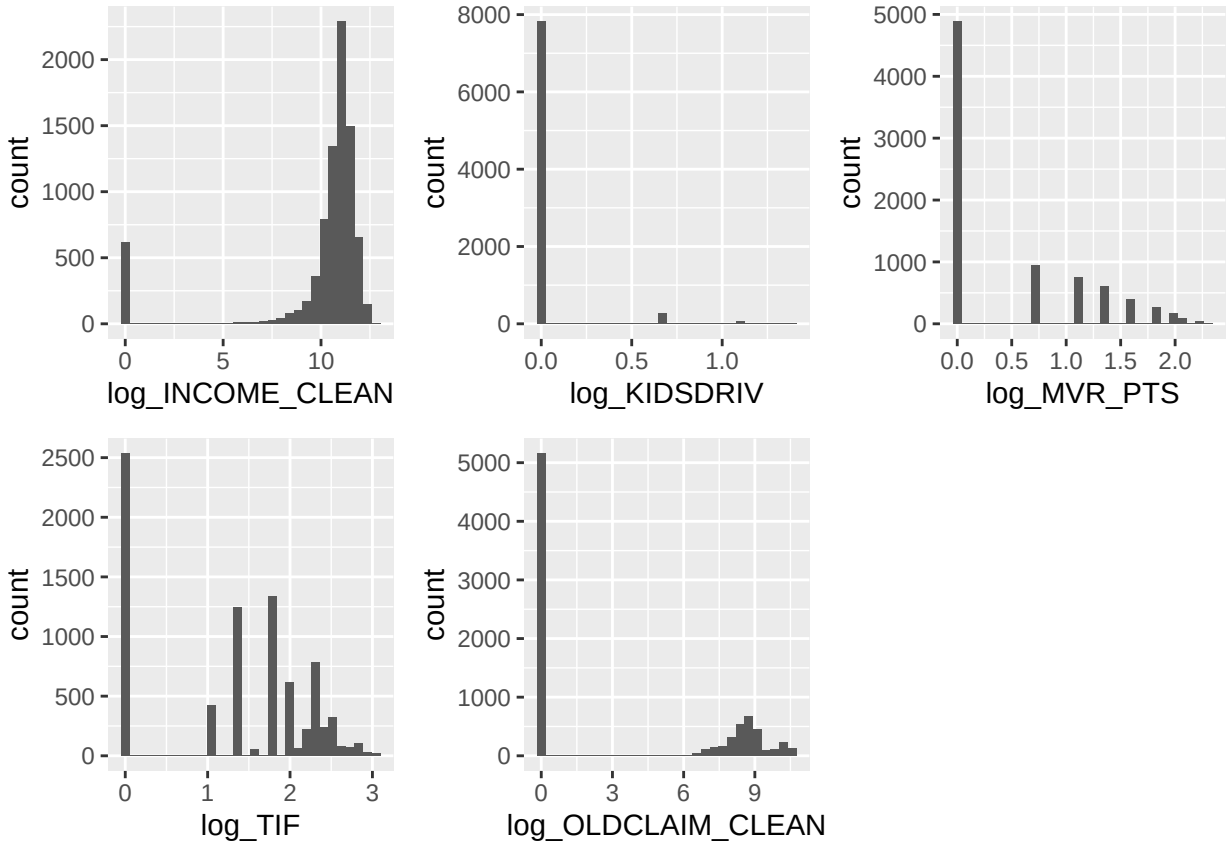
When comparing these boxplots to the original, it's definitely looking a bit cleaner and with less outliers than before!

Transform Non-Normal Variables

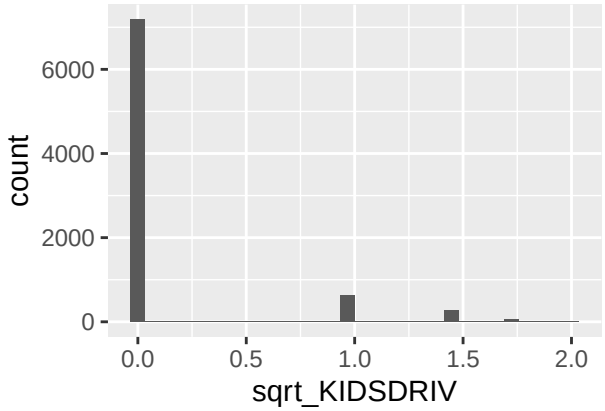
The last alteration before modeling is ensuring that our variables are normal by transforming the ones that don't seem to have much of normal distribution. The fields with distributions that aren't normal are:

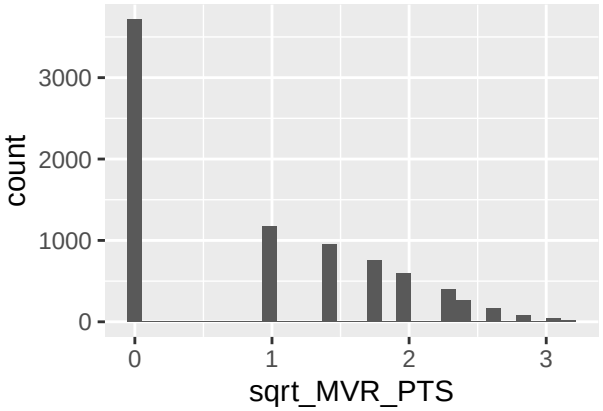
- `INCOME_CLEAN`
- `KIDSDRIV`
- `MVR_PTS`
- `TIF`
- `OLDCLAIM_CLEAN`

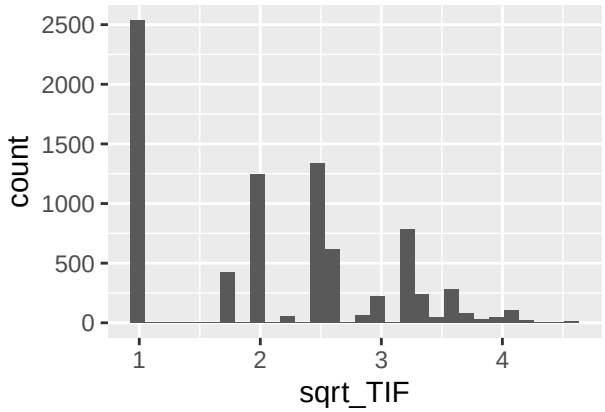
We'll try transforming these with `log` first and if that doesn't work, then we'll `sqrt` it.



The two clean fields (`INCOME_CLEAN` and `OLDCLAIM_CLEAN`) look a bit more normal, but the other three are still not as well-distributed. Let's try `sqrt`.







`TIF` looks a little better here so we'll opt to use this version of the field. The rest of the transformations don't seem to be super helpful.

Given not every piece of data can be normalized, we'll opt to use these versions of the three above variables:

- `INCOME_CLEAN` -> `log_INCOME_CLEAN`
- `OLDCLAIM_CLEAN` -> `log_OLDCLAIM_CLEAN`
- `TIF` -> `sqrt_TIF`

Build Models

Before doing anything, we will split the data into training and test sets with a 70/30 split.

We'll go through two sets of models:

- Model 1: Start from using all the coefficients as is and only use the transformed ones if they don't seem to have a solid impact on the model
- Model 2: Start with all normalized (to the best of our ability) variables and select from there

Let's begin with the binary models.

Binary Models

Model 1A

This first model will use all the fields pre-transformed ones.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + BLUEBOOK_CLEAN + CAR_AGE +
##     CAR_TYPE + CAR_USE + CLM_FREQ + EDUCATION + HOMEKIDS + HOME_VAL_CLEAN +
##     INCOME_CLEAN + JOB + KIDSDRIV + MSTATUS + MVR_PTS + OLDCLAIM_CLEAN +
##     PARENT1 + RED_CAR + REVOKED + SEX + TIF + TRAVTIME + URBANICITY +
##     YOJ, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6835  -0.7108  -0.3886   0.6088   2.9855
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.410e+00  3.857e-01  -3.656 0.000256 ***
## AGE                              1.681e-03  4.874e-03   0.345 0.730151
## BLUEBOOK_CLEAN                  -2.002e-05  6.271e-06  -3.192 0.001411 **
## CAR_AGE                          3.302e-03  9.124e-03   0.362 0.717384
## CAR_TYPEPanel Truck              5.053e-01  1.927e-01   2.623 0.008719 **
## CAR_TYPEPickup                   5.540e-01  1.208e-01   4.584 4.55e-06 ***
## CAR_TYPESports Car               1.027e+00  1.555e-01   6.607 3.93e-11 ***
## CAR_TYPEVan                      6.221e-01  1.536e-01   4.051 5.09e-05 ***
## CAR_TYPEz_SUV                    7.635e-01  1.326e-01   5.759 8.44e-09 ***
## CAR_USEPrivate                  -8.183e-01  1.101e-01  -7.430 1.09e-13 ***
## CLM_FREQ                         2.032e-01  3.403e-02   5.973 2.33e-09 ***
## EDUCATIONBachelors              -4.352e-01  1.391e-01  -3.129 0.001755 **
## EDUCATIONMasters                -2.237e-01  2.107e-01  -1.061 0.288503
## EDUCATIONPhD                    -2.230e-01  2.548e-01  -0.875 0.381459
## EDUCATIONz_High School          -3.779e-02  1.144e-01  -0.330 0.741062
## HOMEKIDS                         4.477e-02  4.524e-02   0.990 0.322319
## HOME_VAL_CLEAN                  -1.365e-06  4.138e-07  -3.300 0.000969 ***
## INCOME_CLEAN                    -3.334e-06  1.276e-06  -2.613 0.008987 **
## JOBClerical                      6.465e-01  2.332e-01   2.772 0.005564 **
## JOBDoctor                       -2.599e-01  3.208e-01  -0.810 0.417863
## JOBHome Maker                    3.721e-01  2.494e-01   1.492 0.135652
## JOBLawyer                        8.466e-02  2.064e-01   0.410 0.681642
## JOBManager                      -3.426e-01  2.034e-01  -1.684 0.092171 .
## JOBProfessional                  3.334e-01  2.099e-01   1.589 0.112112
## JOBStudent                       2.718e-01  2.548e-01   1.067 0.286091
## JOBz_Blue Collar                 5.418e-01  2.197e-01   2.466 0.013666 *
## KIDSDRIV                         3.918e-01  7.435e-02   5.269 1.37e-07 ***
## MSTATUSz_No                      5.472e-01  1.003e-01   5.458 4.82e-08 ***
## MVR_PTS                          1.082e-01  1.660e-02   6.516 7.21e-11 ***
## OLDCLAIM_CLEAN                  -1.060e-05  5.835e-06  -1.817 0.069274 .
## PARENT1Yes                       4.099e-01  1.320e-01   3.105 0.001903 **
## RED_CARyes                      -1.350e-02  1.041e-01  -0.130 0.896861
## REVOKEDYes                       7.908e-01  1.035e-01   7.642 2.14e-14 ***
## SEXz_F                          -6.783e-02  1.337e-01  -0.508 0.611801
## TIF                             -5.218e-02  8.826e-03  -5.913 3.37e-09 ***
## TRAVTIME                         1.659e-02  2.290e-03   7.245 4.34e-13 ***
## URBANICITYz_Highly Rural/ Rural -2.490e+00  1.400e-01 -17.787  < 2e-16 ***
## YOJ                             -1.203e-03  1.027e-02  -0.117 0.906797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6579.8  on 5708  degrees of freedom
## Residual deviance: 5058.9  on 5671  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 5134.9
##
## Number of Fisher Scoring iterations: 5
```

Coefficient Evaluation

Looking at the model's coefficients, these had negative values:

- AGE
- BLUEBOOK_CLEAN
- CAR_AGE
- CAR_USEPrivate
- All EDUCATION values except for High School
- HOME_VAL_CLEAN
- INCOME_CLEAN
- JOBDoctor
- JOBManager
- OLDCLAIM_CLEAN
- RED_CARyes
- SEXz_F
- URBANICITYz_Higlhly Rural/ Rural
- YOJ
- Intercept

These negative coefficients imply that the higher these values (or the presence of them for the categorical values), the less likely these people will get in a crash. For example, just using some of the variables, we can interpret here that older people with more education and certain professions (Doctor/Manager) as well as higher home value and income are less likely to get in a crash.

For the positive values:

- All CAR_TYPE values
- CLM_FREQ
- EDUCATIONz_High School
- HOMEKIDS
- JOBClerical
- JOBHome Maker
- JOBLawyer
- JOBProfessional
- JOBStudent
- JOBz_Blue Collar
- KIDSDRIV
- MSTATUSz_NO
- MVR_PTS
- REVOKEDYes
- TRAVTIME

This would imply that for these fields, if they are higher in value or are present for the categorical ones, this would mean that they're more likely to crash their car. For example, just using some of the variables, blue

collar workers with kids who drive and have had their license revoked are more likely to get in a crash than those who don't have these variables.

Significance Evaluation & Performance

A good amount of these variables were significant – we'll opt to discard those that aren't significant though. That includes:

- AGE
- CAR_AGE
- HOMEKIDS
- OLDCLAIM_CLEAN
- RED_CAR
- SEX
- YOJ

With an AIC of 5134.9 and residual deviance of 5058.9, we'll use this as a baseline to compare to as we iterate on the model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ BLUEBOOK_CLEAN + CAR_TYPE + CAR_USE +
##     CLM_FREQ + EDUCATION + HOME_VAL_CLEAN + INCOME_CLEAN + JOB +
##     KIDSDRIV + MSTATUS + MVR_PTS + PARENT1 + REVOKED + TIF +
##     TRAVTIME + URBANICITY, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7011  -0.7118  -0.3899   0.6199   3.0126
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.323e+00  3.045e-01  -4.343 1.40e-05 ***
## BLUEBOOK_CLEAN                -2.150e-05  5.641e-06  -3.811 0.000138 ***
## CAR_TYPEPanel Truck            5.463e-01  1.791e-01   3.050 0.002289 **
## CAR_TYPEPickup                 5.542e-01  1.206e-01   4.595 4.33e-06 ***
## CAR_TYPESports Car             9.909e-01  1.301e-01   7.617 2.60e-14 ***
## CAR_TYPEVan                    6.506e-01  1.479e-01   4.399 1.09e-05 ***
## CAR_TYPEz_SUV                  7.220e-01  1.037e-01   6.964 3.30e-12 ***
## CAR_USEPrivate                -8.166e-01  1.098e-01  -7.437 1.03e-13 ***
## CLM_FREQ                       1.754e-01  3.045e-02   5.760 8.43e-09 ***
## EDUCATIONBachelors            -4.240e-01  1.310e-01  -3.238 0.001204 **
## EDUCATIONMasters              -1.961e-01  1.899e-01  -1.033 0.301579
## EDUCATIONPhD                  -1.938e-01  2.387e-01  -0.812 0.416841
## EDUCATIONz_High School        -3.695e-02  1.136e-01  -0.325 0.745096
## HOME_VAL_CLEAN                -1.366e-06  4.108e-07  -3.324 0.000886 ***
## INCOME_CLEAN                  -3.349e-06  1.268e-06  -2.642 0.008243 **
## JOBClerical                    6.515e-01  2.327e-01   2.800 0.005111 **
## JOBDoctor                     -2.355e-01  3.195e-01  -0.737 0.461161
## JOBHome Maker                  3.922e-01  2.426e-01   1.617 0.105854
## JOBLawyer                      9.470e-02  2.057e-01   0.460 0.645258
## JOBManager                    -3.372e-01  2.029e-01  -1.662 0.096533 .
## JOBProfessional                3.398e-01  2.093e-01   1.623 0.104555
## JOBStudent                     3.043e-01  2.498e-01   1.218 0.223083
```

```
## JOBz_Blue Collar                 5.513e-01  2.193e-01   2.514 0.011944 *
## KIDSDRIV                         4.243e-01  6.658e-02   6.373 1.85e-10 ***
## MSTATUSz_No                      5.226e-01  9.532e-02   5.482 4.20e-08 ***
## MVR_PTS                          1.058e-01  1.650e-02   6.412 1.44e-10 ***
## PARENT1Yes                       4.709e-01  1.126e-01   4.183 2.88e-05 ***
## REVOKEDYes                       7.259e-01  9.620e-02   7.546 4.50e-14 ***
## TIF                             -5.197e-02  8.807e-03  -5.901 3.61e-09 ***
## TRAVTIME                         1.671e-02  2.286e-03   7.310 2.67e-13 ***
## URBANICITYz_Highly Rural/ Rural -2.493e+00  1.399e-01 -17.825  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6588.4  on 5712  degrees of freedom
## Residual deviance: 5068.4  on 5682  degrees of freedom
## AIC: 5130.4
##
## Number of Fisher Scoring iterations: 5
```

Coefficient Evaluation

In comparison with binary Model 1A, not much has changed here – all variables that were negative stayed negative and vice versa for positive. A bit of the magnitude has adjusted for some of the fields, but overall, there wasn't too much change between this and the previous model.

Significance Evaluation & Performance

Some variables increased in significance which is good to see.

Fortunately, our AIC and residual deviance both decreased, signalling this model is a better fit than the first iteration.

Model 2A

This model will use all the fields, defaulting to the ones that are normalized/transformed.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + BLUEBOOK_CLEAN + CAR_AGE +
##     CAR_TYPE + CAR_USE + CLM_FREQ + EDUCATION + HOMEKIDS + HOME_VAL_CLEAN +
##     log_INCOME_CLEAN + JOB + KIDSDRIV + MSTATUS + MVR_PTS + log_OLDCLAIM_CLEAN +
##     PARENT1 + RED_CAR + REVOKED + SEX + sqrt_TIF + TRAVTIME +
##     URBANICITY + YOJ, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6963  -0.7070  -0.3864   0.5977   2.9496
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -8.144e-01  4.172e-01  -1.952 0.050935 .
## AGE                            3.556e-04  4.908e-03   0.072 0.942240
## BLUEBOOK_CLEAN                -2.100e-05  6.220e-06  -3.376 0.000735 ***
## CAR_AGE                        2.540e-03  9.118e-03   0.279 0.780572
```

```
## CAR_TYPEPanel Truck              5.041e-01  1.923e-01   2.621 0.008761 **
## CAR_TYPEPickup                   5.772e-01  1.210e-01   4.772 1.82e-06 ***
## CAR_TYPESports Car               1.050e+00  1.556e-01   6.744 1.54e-11 ***
## CAR_TYPEVan                      6.249e-01  1.535e-01   4.070 4.70e-05 ***
## CAR_TYPEz_SUV                    7.803e-01  1.328e-01   5.877 4.17e-09 ***
## CAR_USEPrivate                  -8.118e-01  1.104e-01  -7.353 1.94e-13 ***
## CLM_FREQ                         9.331e-02  4.772e-02   1.955 0.050560 .
## EDUCATIONBachelors              -4.570e-01  1.381e-01  -3.310 0.000932 ***
## EDUCATIONMasters                -2.584e-01  2.090e-01  -1.236 0.216302
## EDUCATIONPhD                    -3.324e-01  2.477e-01  -1.342 0.179634
## EDUCATIONz_High School          -3.219e-02  1.145e-01  -0.281 0.778573
## HOMEKIDS                         2.352e-02  4.553e-02   0.517 0.605406
## HOME_VAL_CLEAN                  -1.516e-06  3.943e-07  -3.845 0.000121 ***
## log_INCOME_CLEAN                -7.383e-02  2.116e-02  -3.490 0.000484 ***
## JOBClerical                      7.044e-01  2.310e-01   3.050 0.002289 **
## JOBDoctor                       -2.535e-01  3.200e-01  -0.792 0.428187
## JOBHome Maker                    2.442e-01  2.581e-01   0.946 0.344195
## JOBLawyer                        1.096e-01  2.058e-01   0.532 0.594482
## JOBManager                      -3.217e-01  2.029e-01  -1.586 0.112800
## JOBProfessional                  3.640e-01  2.093e-01   1.739 0.082102 .
## JOBStudent                       1.355e-01  2.634e-01   0.514 0.606986
## JOBz_Blue Collar                 5.864e-01  2.193e-01   2.674 0.007491 **
## KIDSDRIV                         4.036e-01  7.449e-02   5.418 6.02e-08 ***
## MSTATUSz_No                      5.491e-01  9.946e-02   5.521 3.37e-08 ***
## MVR_PTS                          9.576e-02  1.700e-02   5.634 1.76e-08 ***
## log_OLDCLAIM_CLEAN               3.205e-02  1.385e-02   2.314 0.020652 *
## PARENT1Yes                       4.129e-01  1.321e-01   3.127 0.001765 **
## RED_CARyes                      -1.239e-02  1.042e-01  -0.119 0.905378
## REVOKEDYes                       7.224e-01  9.657e-02   7.480 7.40e-14 ***
## SEXz_F                          -8.246e-02  1.336e-01  -0.617 0.537052
## sqrt_TIF                        -2.348e-01  3.885e-02  -6.043 1.51e-09 ***
## TRAVTIME                         1.685e-02  2.295e-03   7.344 2.07e-13 ***
## URBANICITYz_Highly Rural/ Rural -2.479e+00  1.406e-01 -17.638  < 2e-16 ***
## YOJ                              2.264e-02  1.284e-02   1.764 0.077808 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6579.8  on 5708  degrees of freedom
## Residual deviance: 5049.6  on 5671  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 5125.6
##
## Number of Fisher Scoring iterations: 5
```

Looking at the model's coefficients, these had negative values:

- AGE
- BLUEBOOK_CLEAN
- CAR_AGE
- CAR_USEPrivate
- All EDUCATION values except for High School
- HOME_VAL_CLEAN

- log_INCOME_CLEAN
- JOBDoctor
- JOBManager
- JOBStudent
- OLDCLAIM_CLEAN
- RED_CARyes
- SEXz_F
- sqrt_TIF
- URBANICITYz_Highlhly Rural/ Rural
- Intercept

For the positive values:

- All CAR_TYPE values
- CLM_FREQ
- EDUCATIONz_High School
- HOMEKIDS
- JOBClerical
- JOBHome Maker
- JOBLawyer
- JOBProfessional
- JOBz_Blue Collar
- KIDSDRIV
- MSTATUSz_NO
- MVR_PTS
- REVOKEDYes
- TRAVTIME

Compared to Model 1, a few of these variables have switched from positive to negative or vice versa (specifically JOBStudent and OLDCLAIM_CLEAN). A person being a student theoretically should be a positive thing if we base it off of age, but then again, it could also depend on what type of student they are. For OLDCLAIM, this isn't a field we would expect to have a particular effect, so this flip-flopping isn't something we need to dwell on.

Significance Evaluation & Performance

A good amount of these variables were significant – we'll opt to discard those that aren't significant though. That includes:

- AGE
- CAR_AGE
- CLM_FREQ
- 'HOMEKIDS"
- RED_CAR
- SEX
- YOJ

With an AIC of 5125.6 and residual deviance of 5049.6 we'll use this as a baseline to compare to as we iterate on the model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ BLUEBOOK_CLEAN + CAR_TYPE + CAR_USE +
```

```
##     EDUCATION + HOME_VAL_CLEAN + log_INCOME_CLEAN + JOB + KIDSDRIV +
##     MSTATUS + MVR_PTS + log_OLDCLAIM_CLEAN + PARENT1 + REVOKED +
##     sqrt_TIF + TRAVTIME + URBANICITY, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6691  -0.7101  -0.3890   0.6032   2.9533
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -7.546e-01  3.487e-01  -2.164 0.030471 *
## BLUEBOOK_CLEAN                  -2.277e-05  5.586e-06  -4.076 4.59e-05 ***
## CAR_TYPEPanel Truck              5.479e-01  1.788e-01   3.063 0.002188 **
## CAR_TYPEPickup                   5.759e-01  1.207e-01   4.772 1.82e-06 ***
## CAR_TYPESports Car               1.003e+00  1.301e-01   7.708 1.28e-14 ***
## CAR_TYPEVan                      6.532e-01  1.480e-01   4.414 1.01e-05 ***
## CAR_TYPEz_SUV                    7.275e-01  1.040e-01   6.997 2.61e-12 ***
## CAR_USEPrivate                  -8.087e-01  1.100e-01  -7.349 1.99e-13 ***
## EDUCATIONBachelors              -4.610e-01  1.295e-01  -3.559 0.000372 ***
## EDUCATIONMasters                -2.528e-01  1.875e-01  -1.348 0.177604
## EDUCATIONPhD                    -3.256e-01  2.307e-01  -1.411 0.158236
## EDUCATIONz_High School          -4.107e-02  1.136e-01  -0.362 0.717668
## HOME_VAL_CLEAN                  -1.521e-06  3.917e-07  -3.884 0.000103 ***
## log_INCOME_CLEAN                -5.045e-02  1.684e-02  -2.996 0.002735 **
## JOBClerical                      7.103e-01  2.304e-01   3.083 0.002048 **
## JOBDoctor                       -2.347e-01  3.181e-01  -0.738 0.460768
## JOBHome Maker                    2.606e-01  2.563e-01   1.017 0.309223
## JOBLawyer                        1.098e-01  2.052e-01   0.535 0.592567
## JOBManager                      -3.216e-01  2.023e-01  -1.590 0.111837
## JOBProfessional                  3.580e-01  2.090e-01   1.713 0.086675 .
## JOBStudent                       1.700e-01  2.618e-01   0.649 0.516091
## JOBz_Blue Collar                 5.916e-01  2.189e-01   2.703 0.006872 **
## KIDSDRIV                         4.203e-01  6.654e-02   6.317 2.67e-10 ***
## MSTATUSz_No                      5.047e-01  9.396e-02   5.371 7.81e-08 ***
## MVR_PTS                          9.763e-02  1.695e-02   5.760 8.43e-09 ***
## log_OLDCLAIM_CLEAN               5.273e-02  8.860e-03   5.951 2.66e-09 ***
## PARENT1Yes                       4.605e-01  1.127e-01   4.087 4.37e-05 ***
## REVOKEDYes                       7.314e-01  9.624e-02   7.600 2.96e-14 ***
## sqrt_TIF                        -2.332e-01  3.875e-02  -6.017 1.77e-09 ***
## TRAVTIME                         1.691e-02  2.290e-03   7.387 1.50e-13 ***
## URBANICITYz_Highly Rural/ Rural -2.489e+00  1.404e-01 -17.735  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6588.4  on 5712  degrees of freedom
## Residual deviance: 5062.7  on 5682  degrees of freedom
## AIC: 5124.7
##
## Number of Fisher Scoring iterations: 5
```

Coefficient Evaluation

In comparison with binary Model 2A, not much has changed here – all variables that were negative stayed negative and vice versa for positive. A bit of the magnitude has adjusted for some of the fields, but overall, there wasn't too much change between this and the previous model.

Significance Evaluation & Performance

Some variables increased in significance which is good to see.

Unfortunately, our AIC and residual deviance both increased, signalling this model wasn't better than the first iteration. It's interesting how Model 2B and 1B are both similar in their comparisons to 2A and 1A respectively with similar changes in both iterations.

Multiple Linear Regression Models

Model 1A

This first model will use all the fields pre-transformed ones.

```
##
## Call:
## lm(formula = TARGET_AMT ~ AGE + BLUEBOOK_CLEAN + CAR_AGE + CAR_TYPE +
##     CAR_USE + CLM_FREQ + EDUCATION + HOMEKIDS + HOME_VAL_CLEAN +
##     INCOME_CLEAN + JOB + KIDSDRIV + MSTATUS + MVR_PTS + OLDCLAIM_CLEAN +
##     PARENT1 + RED_CAR + REVOKED + SEX + TIF + TRAVTIME + URBANICITY +
##     YOJ, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6161  -1699   -734    367  83147
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.999e+02  6.653e+02   0.601 0.547783
## AGE                            9.174e+00  8.533e+00   1.075 0.282367
## BLUEBOOK_CLEAN                 1.781e-02  1.028e-02   1.733 0.083187 .
## CAR_AGE                       -2.029e+01  1.537e+01  -1.320 0.186728
## CAR_TYPEPanel Truck            3.201e+02  3.302e+02   0.969 0.332468
## CAR_TYPEPickup                 3.042e+02  2.033e+02   1.496 0.134740
## CAR_TYPESports Car             1.132e+03  2.596e+02   4.361 1.32e-05 ***
## CAR_TYPEVan                    3.610e+02  2.569e+02   1.405 0.160026
## CAR_TYPEz_SUV                  5.665e+02  2.132e+02   2.657 0.007904 **
## CAR_USEPrivate                -7.143e+02  1.961e+02  -3.643 0.000272 ***
## CLM_FREQ                       1.123e+02  6.585e+01   1.705 0.088283 .
## EDUCATIONBachelors            -3.178e+02  2.453e+02  -1.296 0.195184
## EDUCATIONMasters              -4.110e+01  3.568e+02  -0.115 0.908307
## EDUCATIONPhD                   2.642e+02  4.223e+02   0.626 0.531540
## EDUCATIONz_High School        -1.284e+02  2.053e+02  -0.626 0.531615
## HOMEKIDS                       7.195e+01  7.918e+01   0.909 0.363543
## HOME_VAL_CLEAN                -4.147e-04  7.106e-04  -0.584 0.559497
## INCOME_CLEAN                  -4.297e-03  2.112e-03  -2.035 0.041948 *
## JOBClerical                    8.162e+02  4.073e+02   2.004 0.045151 *
## JOBDoctor                     -3.387e+02  4.952e+02  -0.684 0.494012
## JOBHome Maker                  6.459e+02  4.339e+02   1.488 0.136702
## JOBLawyer                      2.857e+02  3.548e+02   0.805 0.420767
```

```
## JOBManager                       -2.260e+02  3.463e+02  -0.653 0.513986
## JOBProfessional                   6.854e+02  3.671e+02   1.867 0.061944 .
## JOBStudent                        3.733e+02  4.465e+02   0.836 0.403192
## JOBz_Blue Collar                  8.222e+02  3.843e+02   2.139 0.032443 *
## KIDSDRIV                          2.901e+02  1.354e+02   2.143 0.032163 *
## MSTATUSz_No                       7.686e+02  1.733e+02   4.436 9.35e-06 ***
## MVR_PTS                           1.749e+02  3.157e+01   5.540 3.16e-08 ***
## OLDCLAIM_CLEAN                   -6.669e-03  1.114e-02  -0.598 0.549566
## PARENT1Yes                        7.026e+02  2.421e+02   2.903 0.003715 **
## RED_CARyes                       -1.386e+02  1.788e+02  -0.775 0.438270
## REVOKEDYes                        5.591e+02  1.977e+02   2.828 0.004698 **
## SEXz_F                           -4.065e+02  2.187e+02  -1.858 0.063170 .
## TIF                              -6.020e+01  1.470e+01  -4.096 4.26e-05 ***
## TRAVTIME                          1.455e+01  3.905e+00   3.726 0.000197 ***
## URBANICITYz_Highly Rural/ Rural -1.650e+03  1.671e+02  -9.878  < 2e-16 ***
## YOJ                               9.167e+00  1.790e+01   0.512 0.608500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4545 on 5671 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.07509,    Adjusted R-squared:  0.06905
## F-statistic: 12.44 on 37 and 5671 DF,  p-value: < 2.2e-16
```

Coefficient Evaluation

Looking at the model's coefficients, these had negative values:

- CAR_AGE
- CAR_USEPrivate
- All EDUCATION values except for PhD
- HOME_VAL_CLEAN
- INCOME_CLEAN
- JOBDoctor
- JOBManager
- OLDCLAIM_CLEAN
- RED_CARyes
- SEXz_F
- TIF
- URBANICITYz_Higlhly Rural/ Rural
- YOJ

This is indicating that for example, the older the car and the fact that the driver is a woman as well as a doctor or manager, this means that the payout would be lower.

For the positive values:

- AGE
- BLUEBOOK_CLEAN
- All CAR_TYPE values
- CLM_FREQ
- EDUCATIONz_PhD
- HOMEKIDS
- JOBClerical

- JOBHome Maker
- JOBLawyer
- JOBProfessional
- JOBz_Blue Collar
- JOBStudent
- KIDSDRIV
- MSTATUSz_NO
- MVR_PTS
- PARENT1Yes
- REVOKEDYes
- TRAVTIME
- Intercept

This is indicating that for example, a lawyer who is not married and has children who drive would have a higher payout than vice versa.

Significance Evaluation & Performance

A majority of our fields weren't significant, but we'll still try to drop them and see how this affects the model. These dropped fields include:

- AGE
- BLUEBOOK_CLEAN
- EDUCATION
- HOMEKIDS
- HOME_VAL_CLEAN
- JOB
- OLDCLAIM_CLEAN
- YOJ

With an adjusted R-squared of 0.06905, this really isn't a very accurate model so let's hope to improve this moving forward.

Model 1B

```
##
## Call:
## lm(formula = TARGET_AMT ~ CAR_AGE + CAR_TYPE + CAR_USE + CLM_FREQ +
##     INCOME_CLEAN + KIDSDRIV + MSTATUS + MVR_PTS + PARENT1 + RED_CAR +
##     REVOKED + SEX + TIF + TRAVTIME + URBANICITY, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5861  -1689   -769    310  83175
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.682e+03  2.920e+02   5.762 8.75e-09 ***
## CAR_AGE                      -3.686e+01  1.202e+01  -3.066 0.002179 **
## CAR_TYPEPanel Truck           4.673e+02  2.756e+02   1.696 0.090028 .
## CAR_TYPEPickup                2.396e+02  1.971e+02   1.216 0.224148
## CAR_TYPESports Car            9.798e+02  2.440e+02   4.015 6.02e-05 ***
```

```
## CAR_TYPEVan                       4.346e+02  2.436e+02   1.784 0.074508 .
## CAR_TYPEz_SUV                     4.203e+02  1.977e+02   2.127 0.033496 *
## CAR_USEPrivate                   -7.941e+02  1.508e+02  -5.267 1.44e-07 ***
## CLM_FREQ                          9.720e+01  5.818e+01   1.671 0.094831 .
## INCOME_CLEAN                     -5.242e-03  1.494e-03  -3.509 0.000453 ***
## KIDSDRIV                          3.524e+02  1.214e+02   2.902 0.003723 **
## MSTATUSz_No                       7.679e+02  1.430e+02   5.369 8.25e-08 ***
## MVR_PTS                           1.810e+02  3.140e+01   5.766 8.55e-09 ***
## PARENT1Yes                        7.316e+02  2.099e+02   3.485 0.000495 ***
## RED_CARyes                       -1.743e+02  1.780e+02  -0.979 0.327553
## REVOKEDYes                        5.282e+02  1.854e+02   2.849 0.004406 **
## SEXz_F                           -2.728e+02  2.030e+02  -1.344 0.179113
## TIF                              -5.913e+01  1.467e+01  -4.029 5.66e-05 ***
## TRAVTIME                          1.546e+01  3.899e+00   3.964 7.45e-05 ***
## URBANICITYz_Highly Rural/ Rural -1.503e+03  1.626e+02  -9.243  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4551 on 5693 degrees of freedom
## Multiple R-squared:  0.06933,    Adjusted R-squared:  0.06623
## F-statistic: 22.32 on 19 and 5693 DF,  p-value: < 2.2e-16
```

Coefficient Evaluation

In comparison with binary Model 1A, not much has changed here – all variables that were negative stayed negative and vice versa for positive. A bit of the magnitude has adjusted for some of the fields, but overall, there wasn't too much change between this and the previous model.

Significance Evaluation & Performance

A majority of these variables have increased in significance which is good to see.

Our adjusted R-squared went down slightly to 0.06623 though which isn't great, but it also isn't a huge drop so overall, both models perform around the same in terms of fit.

Model 2A

This model will begin using normalized variables and transformed versions if their original forms aren't normal.

```
##
## Call:
## lm(formula = TARGET_AMT ~ AGE + BLUEBOOK_CLEAN + CAR_AGE + CAR_TYPE +
##     CAR_USE + CLM_FREQ + EDUCATION + HOMEKIDS + HOME_VAL_CLEAN +
##     log_INCOME_CLEAN + JOB + KIDSDRIV + MSTATUS + MVR_PTS + log_OLDCLAIM_CLEAN +
##     PARENT1 + RED_CAR + REVOKED + SEX + sqrt_TIF + TRAVTIME +
##     URBANICITY + YOJ, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6188  -1684   -742    374  82930
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     7.461e+02  7.247e+02   1.030 0.303232
```

```
## AGE                                9.300e+00  8.583e+00   1.083 0.278638
## BLUEBOOK_CLEAN                     1.523e-02  1.019e-02   1.494 0.135147
## CAR_AGE                           -2.089e+01  1.537e+01  -1.359 0.174152
## CAR_TYPEPanel Truck                3.082e+02  3.303e+02   0.933 0.350776
## CAR_TYPEPickup                     3.170e+02  2.034e+02   1.558 0.119237
## CAR_TYPESports Car                 1.140e+03  2.597e+02   4.390 1.16e-05 ***
## CAR_TYPEVan                        3.589e+02  2.570e+02   1.396 0.162690
## CAR_TYPEz_SUV                      5.769e+02  2.133e+02   2.705 0.006858 **
## CAR_USEPrivate                    -7.155e+02  1.962e+02  -3.646 0.000268 ***
## CLM_FREQ                           6.366e+01  9.457e+01   0.673 0.500901
## EDUCATIONBachelors                -3.758e+02  2.437e+02  -1.542 0.123130
## EDUCATIONMasters                  -1.314e+02  3.538e+02  -0.372 0.710257
## EDUCATIONPhD                       7.196e+01  4.096e+02   0.176 0.860540
## EDUCATIONz_High School            -1.493e+02  2.052e+02  -0.728 0.466862
## HOMEKIDS                           6.030e+01  7.976e+01   0.756 0.449688
## HOME_VAL_CLEAN                    -8.880e-04  6.624e-04  -1.341 0.180105
## log_INCOME_CLEAN                  -3.144e+01  3.682e+01  -0.854 0.393092
## JOBClerical                        9.220e+02  4.035e+02   2.285 0.022342 *
## JOBDoctor                         -3.337e+02  4.953e+02  -0.674 0.500538
## JOBHome Maker                      7.430e+02  4.440e+02   1.673 0.094293 .
## JOBLawyer                          3.302e+02  3.543e+02   0.932 0.351474
## JOBManager                        -1.961e+02  3.462e+02  -0.566 0.571083
## JOBProfessional                    7.311e+02  3.665e+02   1.995 0.046127 *
## JOBStudent                         4.146e+02  4.609e+02   0.899 0.368473
## JOBz_Blue Collar                   8.783e+02  3.834e+02   2.290 0.022030 *
## KIDSDRIV                           2.893e+02  1.355e+02   2.135 0.032822 *
## MSTATUSz_No                        7.096e+02  1.707e+02   4.157 3.27e-05 ***
## MVR_PTS                            1.720e+02  3.227e+01   5.331 1.02e-07 ***
## log_OLDCLAIM_CLEAN                 1.179e+01  2.700e+01   0.437 0.662297
## PARENT1Yes                         7.245e+02  2.420e+02   2.994 0.002765 **
## RED_CARyes                        -1.392e+02  1.789e+02  -0.778 0.436436
## REVOKEDYes                         5.173e+02  1.856e+02   2.787 0.005345 **
## SEXz_F                            -4.164e+02  2.187e+02  -1.904 0.057021 .
## sqrt_TIF                          -2.660e+02  6.619e+01  -4.019 5.91e-05 ***
## TRAVTIME                           1.454e+01  3.905e+00   3.722 0.000199 ***
## URBANICITYz_Highly Rural/ Rural -1.638e+03  1.682e+02  -9.741  < 2e-16 ***
## YOJ                                1.710e+01  2.208e+01   0.774 0.438738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4547 on 5671 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.07447,    Adjusted R-squared:  0.06843
## F-statistic: 12.33 on 37 and 5671 DF,  p-value: < 2.2e-16
```

Coefficient Evaluation

Looking at the model's coefficients, these had negative values:

- AGE
- CAR_AGE
- CAR_USEPrivate
- CLM_FREQ
- All EDUCATION values except for PhD

- HOME_VAL_CLEAN
- INCOME_CLEAN
- JOBDoctor
- JOBManager
- RED_CARyes
- SEXz_F
- sqrt_TIF
- URBANICITYz_Higlhly Rural/ Rural
- YOJ

For the positive values:

- AGE
- BLUEBOOK_CLEAN
- All CAR_TYPE values
- EDUCATIONz_PhD
- HOMEKIDS
- JOBClerical
- JOBHome Maker
- JOBLawyer
- JOBProfessional
- JOBz_Blue Collar
- JOBStudent
- KIDSDRIV
- MSTATUSz_NO
- MVR_PTS
- OLDCLAIM_CLEAN
- PARENT1Yes
- REVOKEDYes
- TRAVTIME
- Intercept

Compared to Model 1, a few of these variables have switched from positive to negative or vice versa (specifically AGE, OLDCLAIM_CLEAN, and CLM_FREQ). Age might not be indicative of a payout value as older claims though – theoretically, the higher that value, the higher your future payouts. The fact that this was negative in the previous iteration is a little confusing, but it seems like it'd be a good thing here now that it's positive and makes more sense.

Regardless, none of these fields are statistically significant at least.

Significance Evaluation & Performance

A majority of our fields weren't significant, but we'll still try to drop them and see how this affects the model. These dropped fields include:

- AGE
- BLUEBOOK_CLEAN
- CLM_FREQ
- HOMEKIDS
- log_INCOME_CLEAN
- log_OLDCLAIM_CLEAN
- RED_CAR
- YOJ

With an adjusted R-squared of 0.06843, this really isn't a very accurate model so let's hope to improve this moving forward.

```
##
## Call:
## lm(formula = TARGET_AMT ~ CAR_AGE + CAR_TYPE + CAR_USE + EDUCATION +
##     HOME_VAL_CLEAN + JOB + KIDSDRIV + MSTATUS + MVR_PTS + PARENT1 +
##     REVOKED + SEX + sqrt_TIF + TRAVTIME + URBANICITY, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6037  -1672   -760    352  83038
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.244e+03  5.243e+02   2.372 0.017717 *
## CAR_AGE                        -2.138e+01  1.535e+01  -1.392 0.163916
## CAR_TYPEPanel Truck             5.354e+02  2.965e+02   1.805 0.071056 .
## CAR_TYPEPickup                  2.827e+02  2.020e+02   1.400 0.161570
## CAR_TYPESports Car              1.072e+03  2.430e+02   4.412 1.04e-05 ***
## CAR_TYPEVan                     4.648e+02  2.493e+02   1.865 0.062289 .
## CAR_TYPEz_SUV                   4.982e+02  1.967e+02   2.533 0.011327 *
## CAR_USEPrivate                 -7.385e+02  1.956e+02  -3.775 0.000162 ***
## EDUCATIONBachelors             -3.784e+02  2.423e+02  -1.561 0.118476
## EDUCATIONMasters               -1.392e+02  3.525e+02  -0.395 0.692867
## EDUCATIONPhD                    9.796e+01  4.078e+02   0.240 0.810155
## EDUCATIONz_High School         -1.550e+02  2.044e+02  -0.758 0.448466
## HOME_VAL_CLEAN                 -8.611e-04  6.489e-04  -1.327 0.184538
## JOBClerical                     9.146e+02  4.026e+02   2.272 0.023123 *
## JOBDoctor                      -2.880e+02  4.940e+02  -0.583 0.559896
## JOBHome Maker                   8.006e+02  4.092e+02   1.957 0.050452 .
## JOBLawyer                       3.620e+02  3.530e+02   1.025 0.305231
## JOBManager                     -1.951e+02  3.452e+02  -0.565 0.571870
## JOBProfessional                 7.388e+02  3.658e+02   2.019 0.043498 *
## JOBStudent                      4.664e+02  4.302e+02   1.084 0.278335
## JOBz_Blue Collar                8.728e+02  3.831e+02   2.278 0.022742 *
## KIDSDRIV                        3.470e+02  1.214e+02   2.858 0.004285 **
## MSTATUSz_No                     6.743e+02  1.623e+02   4.155 3.30e-05 ***
## MVR_PTS                         1.927e+02  2.939e+01   6.554 6.08e-11 ***
## PARENT1Yes                      7.556e+02  2.105e+02   3.589 0.000334 ***
## REVOKEDYes                      5.156e+02  1.854e+02   2.781 0.005434 **
## SEXz_F                         -2.377e+02  1.747e+02  -1.361 0.173605
## sqrt_TIF                       -2.665e+02  6.609e+01  -4.033 5.59e-05 ***
## TRAVTIME                        1.476e+01  3.899e+00   3.787 0.000154 ***
## URBANICITYz_Highly Rural/ Rural -1.696e+03  1.630e+02 -10.404  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4546 on 5683 degrees of freedom
## Multiple R-squared:  0.0731, Adjusted R-squared:  0.06837
## F-statistic: 15.45 on 29 and 5683 DF,  p-value: < 2.2e-16
```

Coefficient Evaluation

In comparison with binary Model 2A, not much has changed here – all variables that were negative stayed

negative and vice versa for positive. A bit of the magnitude has adjusted for some of the fields, but overall, there wasn't too much change between this and the previous model.

Significance Evaluation & Performance

A few of these variables have increased in significance which is good to see.

Our adjusted R-squared went down slightly to 0.06837 unfortunately though, signalling that this isn't a better fit, but it's such a slight decrease that it isn't too significant.

Select Models

Binary Models

Confusion Matrices

First, we'll take a look at confusion matrices for each of the models.

```
# if the prediction is >= 0.5, then we would predict 1 for that row, otherwise 0
test$model1a_binary <- ifelse(predict.glm(model1a_binary, test, "response") >= 0.5, 1, 0)

# create the confusion matrix
cm1a <- confusionMatrix(factor(test$model1a_binary), factor(test$TARGET_FLAG), "1")
results <- tibble(Model = "Model #1A", Accuracy=cm1a$byClass[11], F1 = cm1a$byClass[7],
                  Deviance= model1a_binary$deviance,
                  R2 = 1 - model1a_binary$deviance / model1a_binary$null.deviance,
                  AIC = model1a_binary$aic)
cm1a
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1644  387
##          1  156  259
##
##                Accuracy : 0.778
##                  95% CI : (0.761, 0.7943)
##     No Information Rate : 0.7359
##     P-Value [Acc > NIR] : 8.57e-07
##
##                   Kappa : 0.3549
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4009
##             Specificity : 0.9133
##          Pos Pred Value : 0.6241
##          Neg Pred Value : 0.8095
##              Prevalence : 0.2641
##          Detection Rate : 0.1059
##    Detection Prevalence : 0.1697
##       Balanced Accuracy : 0.6571
##
##        'Positive' Class : 1
```

```
##
```

```
# if the prediction is >= 0.5, then we would predict 1 for that row, otherwise 0
test$model1b_binary <- ifelse(predict.glm(model1b_binary, test, "response") >= 0.5, 1, 0)

# create the confusion matrix
cm1b <- confusionMatrix(factor(test$model1b_binary), factor(test$TARGET_FLAG), "1")
results <- tibble(Model = "Model #1B", Accuracy=cm1b$byClass[11], F1 = cm1b$byClass[7],
                  Deviance= model1b_binary$deviance,
                  R2 = 1 - model1b_binary$deviance / model1b_binary$null.deviance,
                  AIC= model1b_binary$aic)
cm1b
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1649  388
##          1  151  260
##
##                Accuracy : 0.7798
##                  95% CI : (0.7629, 0.7961)
##     No Information Rate : 0.7353
##     P-Value [Acc > NIR] : 2.047e-07
##
##                   Kappa : 0.3594
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4012
##             Specificity : 0.9161
##          Pos Pred Value : 0.6326
##          Neg Pred Value : 0.8095
##              Prevalence : 0.2647
##          Detection Rate : 0.1062
##    Detection Prevalence : 0.1679
##       Balanced Accuracy : 0.6587
##
##        'Positive' Class : 1
##
```

```
# if the prediction is >= 0.5, then we would predict 1 for that row, otherwise 0
test$model2a_binary <- ifelse(predict.glm(model2a_binary, test, "response") >= 0.5, 1, 0)

# create the confusion matrix
cm2a <- confusionMatrix(factor(test$model2a_binary), factor(test$TARGET_FLAG), "1")
results <- tibble(Model = "Model #2A", Accuracy=cm2a$byClass[11], F1 = cm2a$byClass[7],
                  Deviance= model2a_binary$deviance,
                  R2 = 1 - model2a_binary$deviance / model2a_binary$null.deviance,
                  AIC= model2a_binary$aic)
cm2a
```

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

```
##
##            Reference
## Prediction    0    1
##          0 1643  384
##          1  157  262
##
##                Accuracy : 0.7788
##                  95% CI : (0.7618, 0.7951)
##     No Information Rate : 0.7359
##     P-Value [Acc > NIR] : 5.309e-07
##
##                   Kappa : 0.3588
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4056
##             Specificity : 0.9128
##          Pos Pred Value : 0.6253
##          Neg Pred Value : 0.8106
##              Prevalence : 0.2641
##          Detection Rate : 0.1071
##    Detection Prevalence : 0.1713
##       Balanced Accuracy : 0.6592
##
##        'Positive' Class : 1
##
```

```
# if the prediction is >= 0.5, then we would predict 1 for that row, otherwise 0
test$model2b_binary <- ifelse(predict.glm(model2b_binary, test, "response") >= 0.5, 1, 0)

# create the confusion matrix
cm2b <- confusionMatrix(factor(test$model2b_binary), factor(test$TARGET_FLAG), "1")
results <- tibble(Model = "Model #2B", Accuracy=cm2b$byClass[11], F1 = cm2b$byClass[7],
                  Deviance= model2b_binary$deviance,
                  R2 = 1 - model2b_binary$deviance / model2b_binary$null.deviance,
                  AIC= model2b_binary$aic)
cm2b
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction    0    1
##          0 1656  379
##          1  144  269
##
##                Accuracy : 0.7864
##                  95% CI : (0.7696, 0.8024)
##     No Information Rate : 0.7353
##     P-Value [Acc > NIR] : 2.766e-09
##
##                   Kappa : 0.3791
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
```

```
##             Sensitivity : 0.4151
##             Specificity : 0.9200
##          Pos Pred Value : 0.6513
##          Neg Pred Value : 0.8138
##              Prevalence : 0.2647
##          Detection Rate : 0.1099
##    Detection Prevalence : 0.1687
##       Balanced Accuracy : 0.6676
##
##        'Positive' Class : 1
##
```

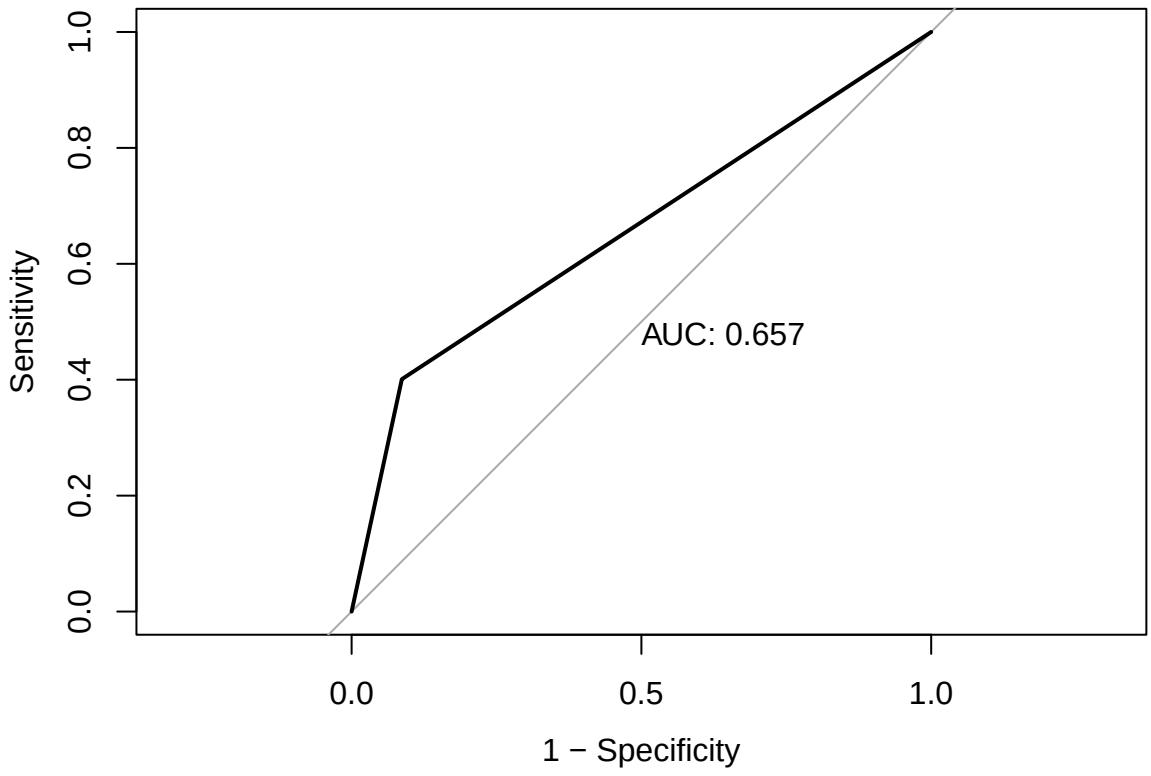
ROC

Now with all of these matrices, we'll look at ROC curves.

```
print('Model 1A ROC Curve')
```

```
## [1] "Model 1A ROC Curve"
```

```
roc(test[["TARGET_FLAG"]], test[["model1a_binary"]], plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```



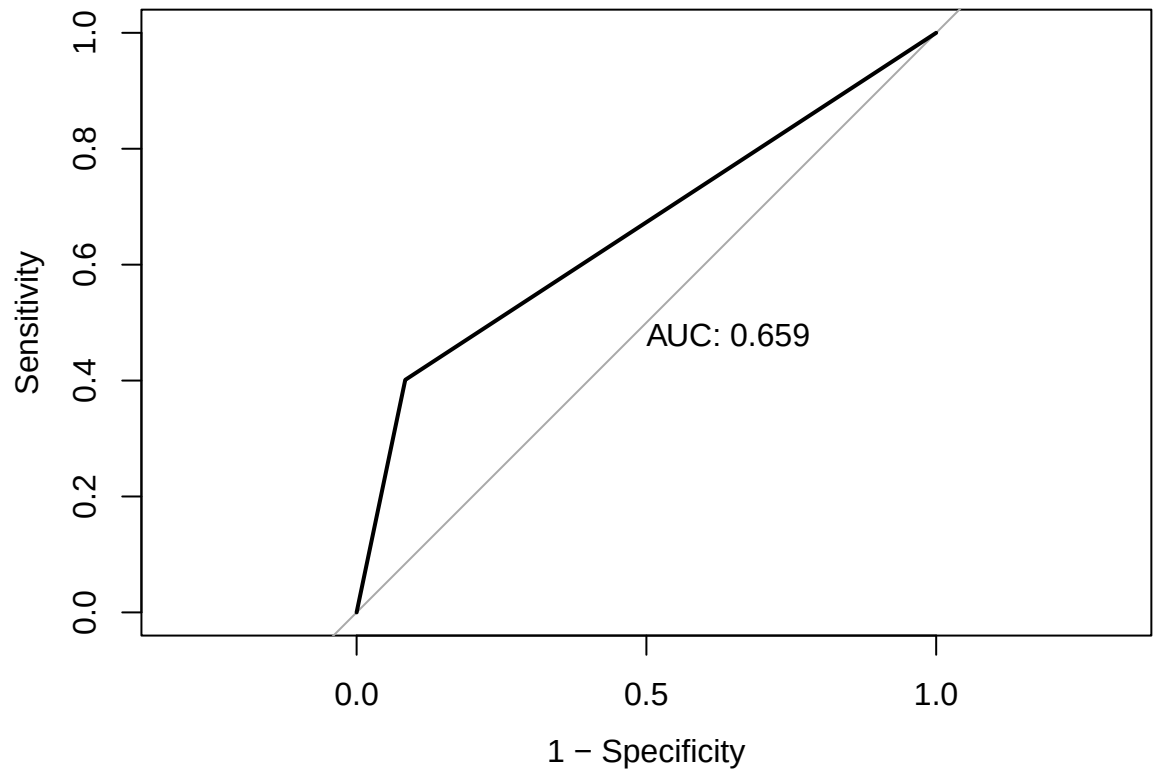

```
##
## Call:
## roc.default(response = test[["TARGET_FLAG"]], predictor = test[["model1a_binary"]],     plot = TRUE,
```

```
##
## Data: test[["model1a_binary"]] in 1800 controls (test[["TARGET_FLAG"]] 0) < 646 cases (test[["TARGET
## Area under the curve: 0.6571
```

```
print('Model 1B ROC Curve')
```

```
## [1] "Model 1B ROC Curve"
```

```
roc(test[["TARGET_FLAG"]], test[["model1b_binary"]], plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```

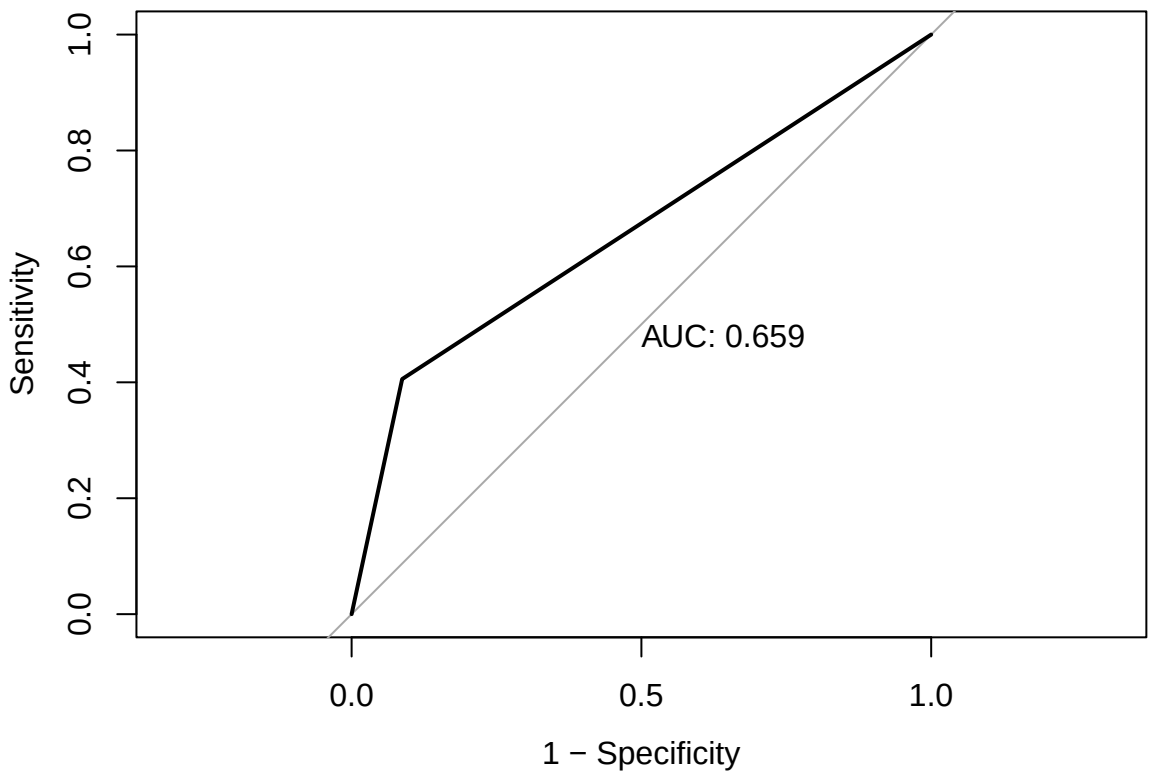



```
##
## Call:
## roc.default(response = test[["TARGET_FLAG"]], predictor = test[["model1b_binary"]],     plot = TRUE,
##
## Data: test[["model1b_binary"]] in 1800 controls (test[["TARGET_FLAG"]] 0) < 648 cases (test[["TARGET
## Area under the curve: 0.6587
```

```
print('Model 2A ROC Curve')
```

```
## [1] "Model 2A ROC Curve"
```

```
roc(test[["TARGET_FLAG"]], test[["model2a_binary"]], plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```

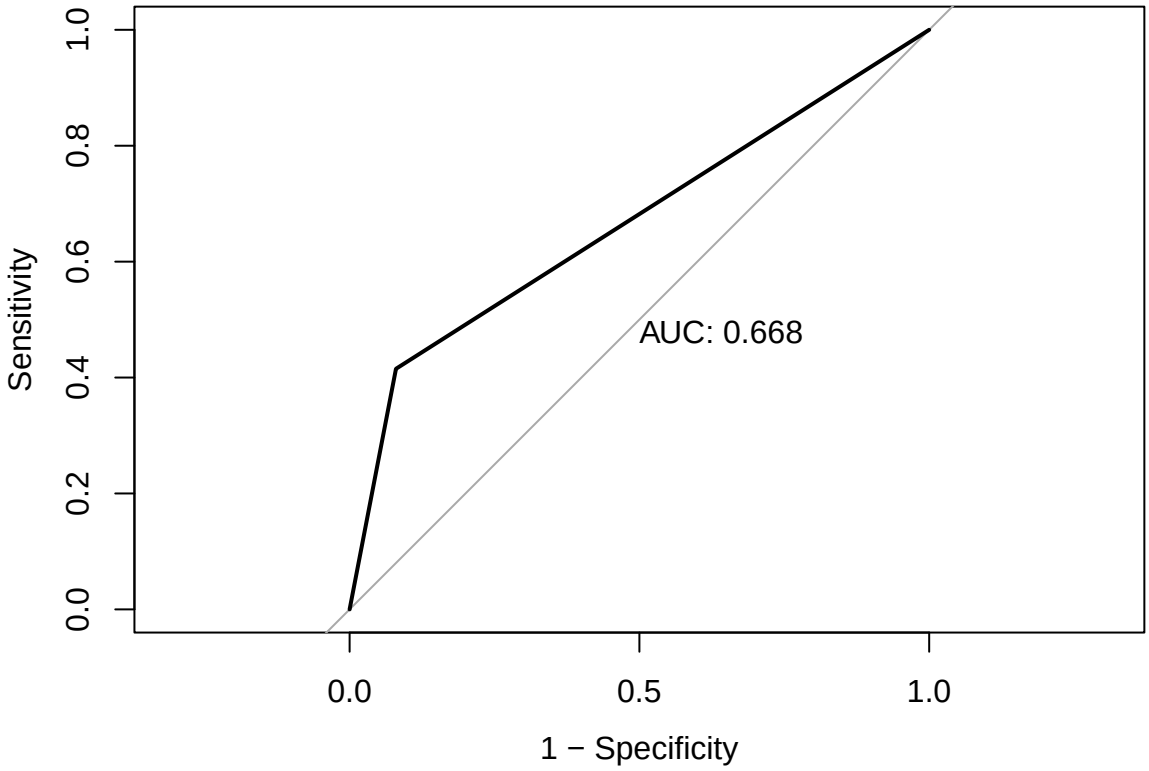



```
##
## Call:
## roc.default(response = test[["TARGET_FLAG"]], predictor = test[["model2a_binary"]],     plot = TRUE,
##
## Data: test[["model2a_binary"]] in 1800 controls (test[["TARGET_FLAG"]] 0) < 646 cases (test[["TARGET
## Area under the curve: 0.6592
```

```
print('Model 2B ROC Curve')
```

```
## [1] "Model 2B ROC Curve"
```

```
roc(test[["TARGET_FLAG"]], test[["model2b_binary"]], plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```




```
##
## Call:
## roc.default(response = test[["TARGET_FLAG"]], predictor = test[["model2b_binary"]],     plot = TRUE,
##
## Data: test[["model2b_binary"]] in 1800 controls (test[["TARGET_FLAG"]] 0) < 648 cases (test[["TARGET_
## Area under the curve: 0.6676
```

Overall Comparisons

```
##          Residual.Deviance      AIC  Accuracy        F1        R2
## Model 1A          5058.918 5134.918 0.6571311 0.4882187 0.2311470
## Model 1B          5068.446 5130.446 0.6586728 0.4910293 0.2307066
## Model 2A          5049.561 5125.561 0.6591753 0.4920188 0.2325690
## Model 2B          5062.698 5124.698 0.6675617 0.5070688 0.2315790
```

Based on the above output:

- Residual Deviance: Model 1B had the lowest Residual Deviance byfar
- AIC: Model 1B had the best AIC as well
- Accuracy: Model 2B had the best Accuracy, but not by much
- F1: Model 2B had the best F1 statistic
- R^2: Model 2A had the highest R^2, although this was just barely

Binary Model Conclusion

With this information, Model 1B seems to outperform the other models due to its better Residual Deviance and AIC, Accuracy, F1, and R^2 are all so close between the other models, so given 1B outperforms so much more for RD and AIC, we believe this would be the best fitting model.

Linear Models

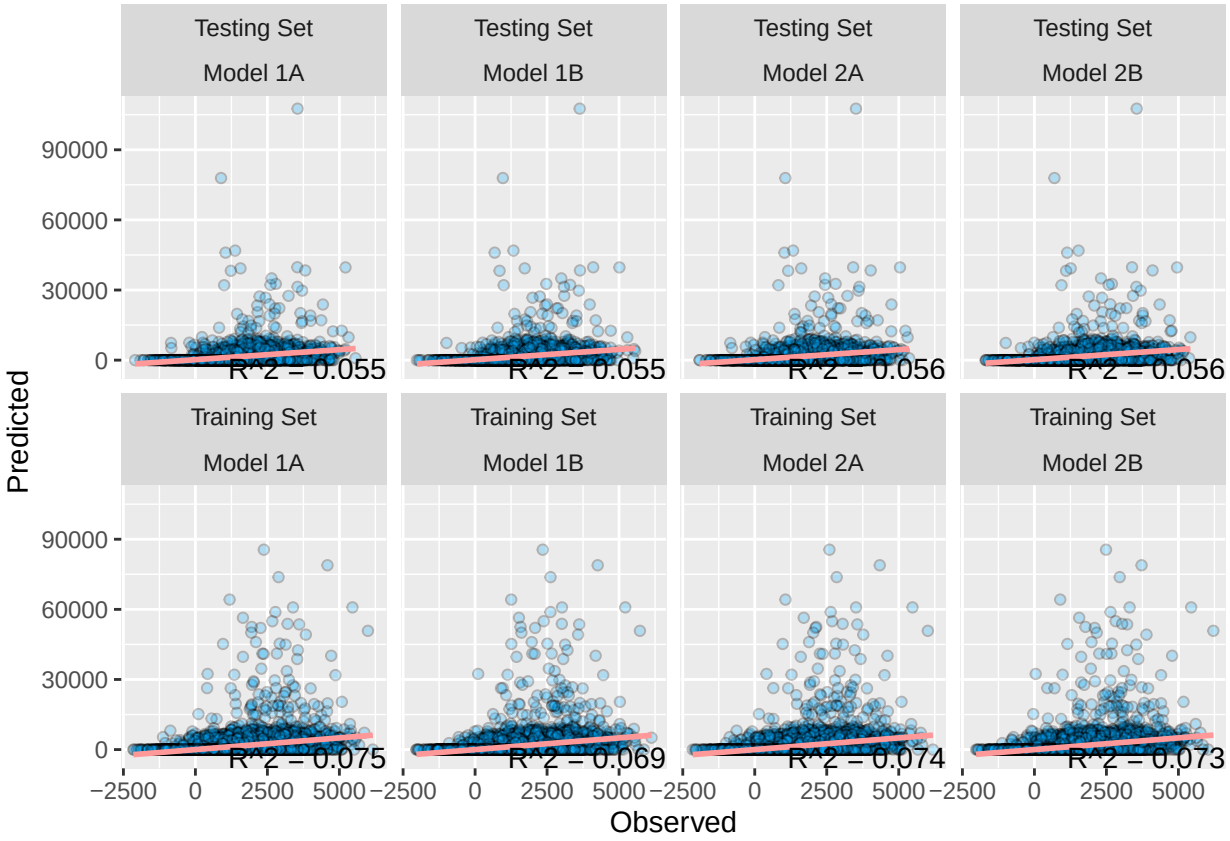
We'll start by looking at mean squared error, adjusted r-squared, and F-statistics before plotting residuals after.

```
##                MSE Adjusted.R.Squared F.Statistic    F.p.value
## Model 1A 20519571         0.06905167    12.44276 3.211104e-71
## Model 1B 20636709         0.06622591    22.32165 4.277568e-75
## Model 2A 20533284         0.06842951    12.33209 1.848638e-70
## Model 2B 20553235         0.06836651    15.45402 1.960997e-73
```

Based on the above output:

- MSE: Model 1A had the best MSE as it's slightly lower than the other values
- Adjusted R-Squared: Model 1A once again is doing slightly better here
- F Statistic: Model 1B does the best here at 22.32

Next, we will compare how these models do with the test dataset and compare residuals.



Interestingly, it seems that the second set of models performed slightly better when using the test dataset. The best performing model looking at the training set was the first one which was just using all features

in the state they're provided (so untransformed). Visually, the residual plots don't seem to vary too much; they all are around the same r-squared so the change between them isn't too apparent.

Linear Model Conclusion

Based on all of the factors shown above, Model 2B seems to be the most viable. It performs solidly when we looked the MSE, adjusted r-squared, and F-statistic, while also was one of the slightly better performing one out of the second round of models when using the test dataset. It uses transformed/more normalized variables while also filtering out the statistically insignificant features as well, resulting in a well-performing model with relevant features.