

# DATA 621 - Homework 4

2022-11-16

## Problem Statement and Goals

In this report, we generate two different models; a multiple linear regression model and a binary logistic regression model. The multiple linear regression model contains a target variable called **TARGET\_AMT**, which is the amount of money it will cost if the person crashes their car. The binary logistic regression model target variable, **TARGET\_FLAG** consists of 0's and 1's. 1 represents that the person was in a car crash, and zero indicates that the person was not in a car crash. The analysis detailed in this report shows the testing of several models from which a best multiple linear regression model and a best binary logistic regression model were selected based on model performance and various metrics.

## Data Exploration

The following is a summary of the variables provided within the data to generate the binary logistic regression and multiple linear regression models.

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

Variable Name	Definition	Theoretical 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 than 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

A summary of the variables is shown below. The INDEX variable has been removed. The summary below reveals that AGE, VOJ, INCOME, HOME\_VAL, and CAR\_AGE have missing values.

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS
0:6008	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000
1:2153	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000
	Median : 0	Median :0.0000	Median :45.00	Median :0.0000
	Mean : 1504	Mean :0.1711	Mean :44.79	Mean :0.7212
	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000
	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000
			NA's :6	
YOJ	INCOME	PARENT1	HOME_VAL	MSTATUS
Min. : 0.0	Min. : 0	No :7084	Min. : 0	No :3267
1st Qu.: 9.0	1st Qu.: 28097	Yes:1077	1st Qu.: 0	Yes:4894
Median :11.0	Median : 54028		Median :161160	
Mean :10.5	Mean : 61898		Mean :154867	
3rd Qu.:13.0	3rd Qu.: 85986		3rd Qu.:238724	

Max.	:23.0	Max.	:367030	Max.	:885282
NA's	:454	NA's	:445	NA's	:464
SEX		EDUCATION		JOB	
F:4375	<High School:1203	Blue Collar	:1825	Min.	: 5.00
M:3786	Bachelors :2242	Clerical	:1271	1st Qu.:	22.00
	High School :2330	Professional:	1117	Median	: 33.00
	Masters :1658	Manager	: 988	Mean	: 33.49
	PhD : 728	Lawyer	: 835	3rd Qu.:	44.00
		Student	: 712	Max.	:142.00
		(Other)	:1413		
CAR_USE		BLUEBOOK		TIF	
Commercial:3029	Min. : 1500	Min.	: 1.000	Minivan	:2145
Private :5132	1st Qu.: 9280	1st Qu.:	1.000	Panel Truck:	676
	Median :14440	Median	: 4.000	Pickup	:1389
	Mean :15710	Mean	: 5.351	Sports Car	: 907
	3rd Qu.:20850	3rd Qu.:	7.000	SUV	:2294
	Max. :69740	Max.	:25.000	Van	: 750
RED_CAR		OLDCLAIM		CLM_FREQ	
no :5783	Min. : 0	Min.	:0.0000	REVOKED	
yes:2378	1st Qu.: 0	1st Qu.:	0.0000	No :7161	Min. : 0.000
	Median : 0	Median	:0.0000	Yes:1000	1st Qu.: 0.000
	Mean : 4037	Mean	:0.7986		Median : 1.000
	3rd Qu.: 4636	3rd Qu.:	2.0000		Mean : 1.696
	Max. :57037	Max.	:5.0000		3rd Qu.: 3.000
					Max. :13.000
CAR_AGE		URBANICITY			
Min. : -3.000	Highly Rural/	Rural:	1669		
1st Qu.: 1.000	Highly Urban/	Urban:	6492		
Median : 8.000					
Mean : 8.328					
3rd Qu.:12.000					
Max. :28.000					
NA's :510					

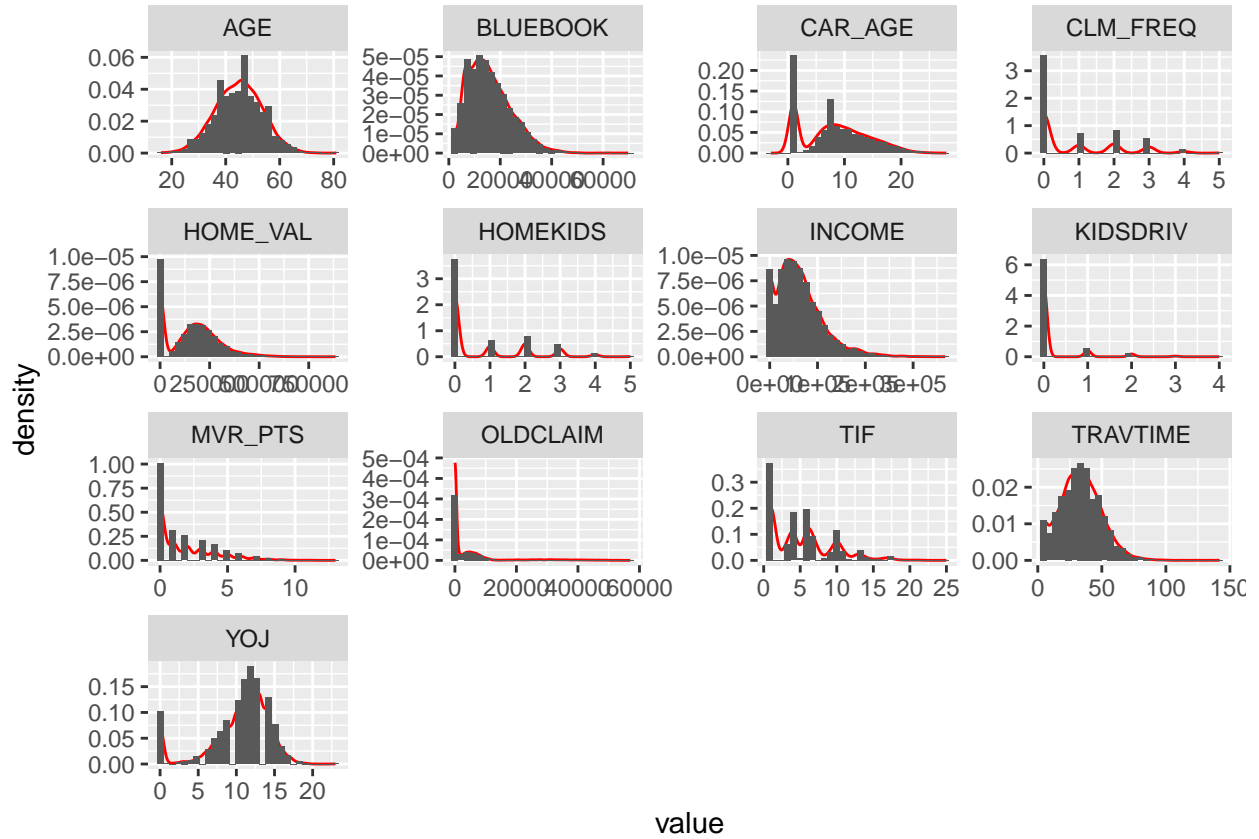


Figure XX: Histograms for all the variables.

The density plots above show that BLUEBOOK, INCOME, and TRAVTIME could be transformed in order to fit the normal distribution assumption of a linear regression model. The variables with a bimodal distribution were dealt with and an explanation of the process is provided in the “Dealing with Bimodal Variables” section.

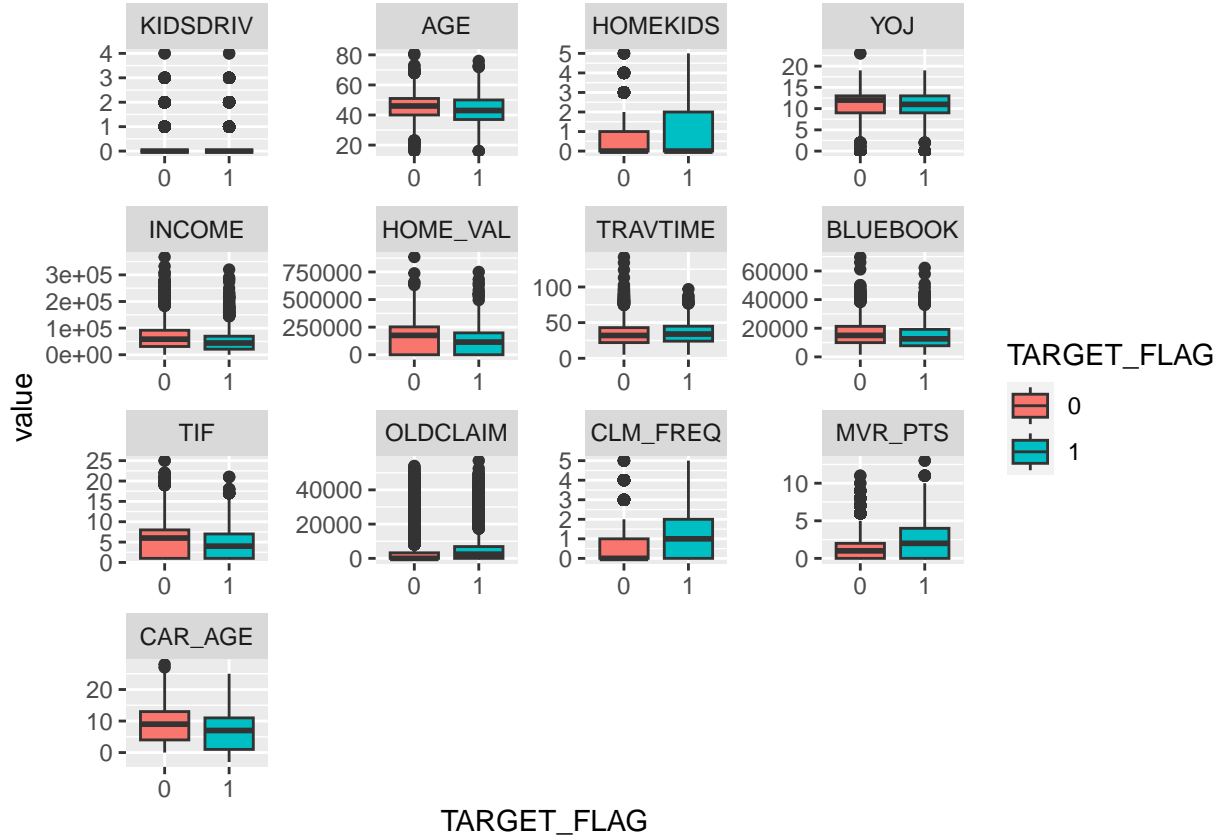


Figure XX: Boxplots for the dataset

We can see some findings that support the theoretical effects for some of the variables using the boxplots in Figure xx. It seems that younger cars are more likely to get into crashes as opposed to older cars as shown in the `CAR_AGE` boxplot. The theoretical effect of the `CLM_FREQ` (The more claims you filed in the past, the more you are likely to file in the future) is supported by the `CLM_FREQ` boxplot. The theoretical effect of `MVR_PTS` (If you get lots of traffic tickets, you tend to get into more crashes) is supported by the `MVR_PTS` boxplot. It would also seem that the theoretical effects of `INCOME` and `TIF` are also supported by the data.

### Examining Feature Multicollinearity

Finally, it is imperative to understand which features are correlated with each other in order to address and avoid multicollinearity within our models. By using a correlation plot, we can visualize the relationships between certain features. The correlation plot is only able to determine the correlation for continuous variables. There are methodologies to determine correlations for categorical variables (tetrachoric correlation). However there is only one binary predictor variable which is why the multicollinearity will only be considered for the continuous variables.

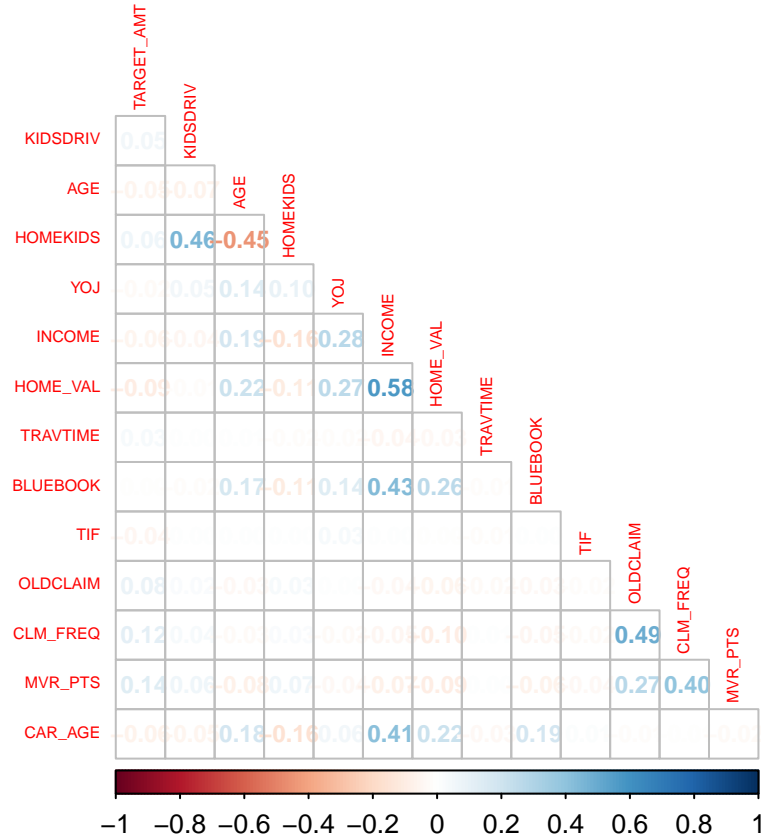


Figure xx: Multicollinearity plot for continuous predictor variables

The figure above shows that there isn't much multicollinearity between the variables. There is a moderately positive correlation of 0.58 between INCOME and HOME\_VAL.

## NA exploration

As can be seen below, some of the columns have missing values. Contextually, this can be possible because not every metric must have a value- for example it is possible that an entire season can be played without a batter being hit by the pitch. However it is less likely that an entire season can be played without any strikeouts by batters. We did some research and came up with ways to address each of these issues- more on that later.

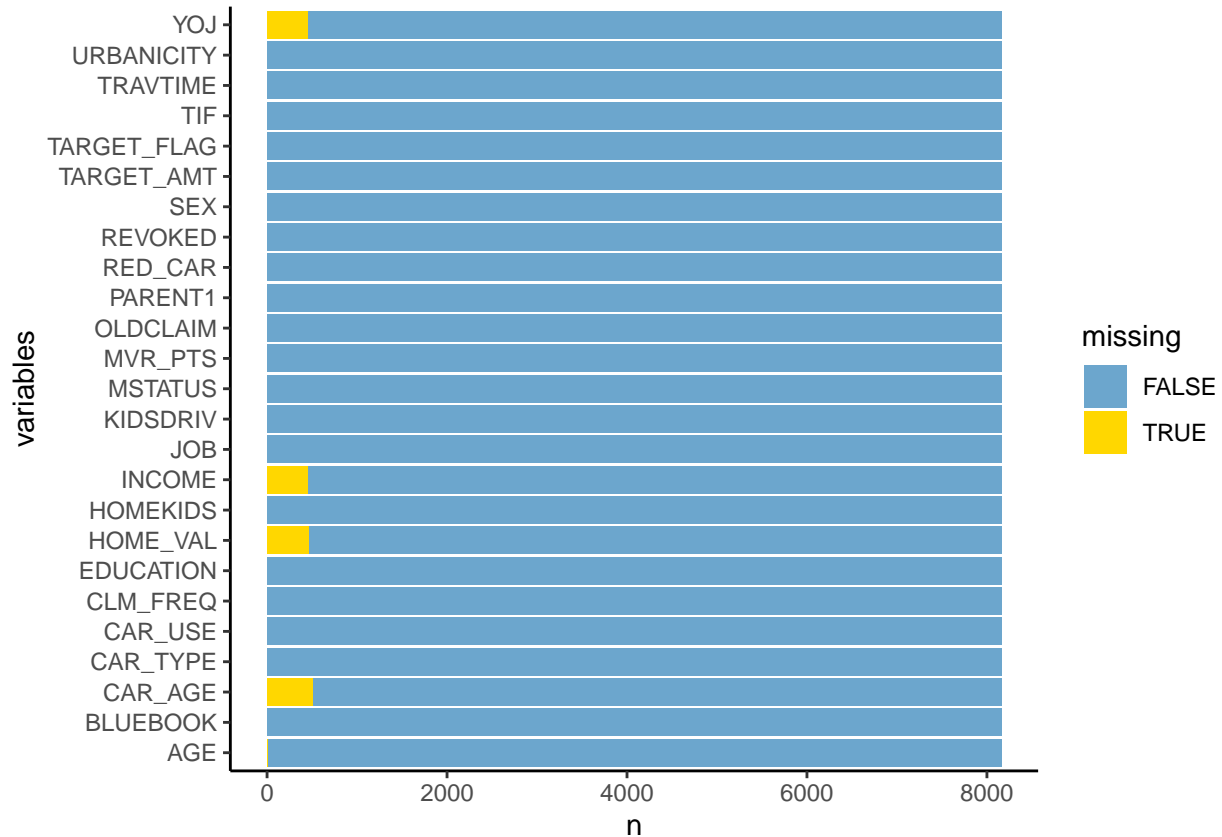


Figure xx: Barplot of number of missing values for each predictor.

The barplot above shows that YOJ, INCOME, HOME\_VAL, AGE, and CAR\_AGE were missing some data values. However, the amount of missing data for each variable is less than 10%. Therefore, imputing can be done on the missing data.

## Data Preparation

### Dealing with Missing Values

In general, imputations by the means/medians is acceptable if the missing values only account for 5% of the sample. Peng et al.(2006) However, should the degree of missing values exceed 20% then using these simple imputation approaches will result in an artificial reduction in variability due to the fact that values are being imputed at the center of the variable's distribution.

Our team decided to employ another technique to handle the missing values: Multiple Regression Imputation using the MICE package.

The MICE package in R implements a methodology where each incomplete variable is imputed by a separate model. Alice points out that plausible values are drawn from a distribution specifically designed for each missing datapoint. Many imputation methods can be used within the package. The one that was selected for the data being analyzed in this report is PMM (Predictive Mean Matching), which is used for quantitative data.

Van Buuren explains that PMM works by selecting values from the observed/already existing data that would most likely belong to the variable in the observation with the missing value. The advantage of this is that it selects values that must exist from the observed data, so no negative values will be used to impute missing data. Not only that, it circumvents the shrinking of errors by using multiple regression models. The variability between the different imputed values gives a wider, but more correct standard error. Uncertainty

is inherent in imputation which is why having multiple imputed values is important. Not only that. Marshall et al. 2010 points out that:

“Another simulation study that addressed skewed data concluded that predictive mean matching ‘may be the preferred approach provided that less than 50% of the cases have missing data...’

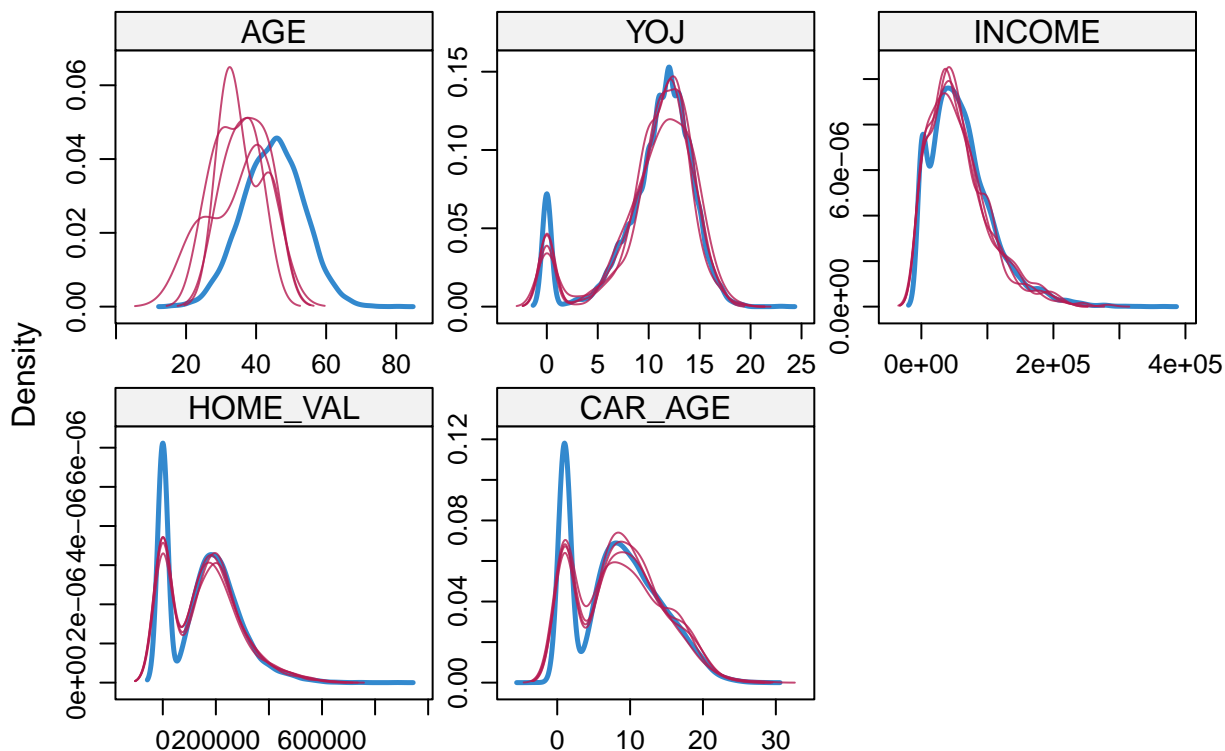


Figure xx: Density plots for variables containing missing data. The number of multiple imputations was set to 4. Each of the red lines represents the distribution for each imputation.

The blue lines for each of the graphs above represent the distributions the non-missing data for each of the variables while the red lines represent the distributions for the imputed data. Note that the distributions for the imputed data for each of the iterations closely matches the distributions for the non-missing data, which is ideal. If the distributions did not match so well, than another imputing method would have had to have been used.

### Feature Manipulation based on Multicollinearity Plot

There is a significant amount of observations for INCOME with a value of 0. Therefore, we reasoned that we could create a new dummy variable based on the INCOME, called EMPLOYMENT, where 0 was unemployed and any positive value for income would be employed. Then, we could effectively be rid of the INCOME variable while still having some sort of distinction that represents this variable that does not have a high correlation with any of the other variables.

### Dealing with Bimodal Variables

Bimodal distributions in data are interesting, in that they represent features which actually contain multiple (2) inherent systems resulting in separated distributional peaks. While a Box-Cox transformation could have been undertaken in order to transform the bimodal variables to a normal distribution. However, this throws away important information that is inherent in the bimodal variable itself. The fact that the variable is



bimodal in the first place is essentially ignored, and the predicted values in the linear multiple regression model will not reflect this bimodality.

For variables that displayed bimodality, new variables were created; `bi_CAR_AGE`, `bi_CLM_FREQ`, `bi_HOME_VAL`, `bi_KIDSDRIV`, `bi_YOJ`. For many of these variables, there are a significant number of 0 values, which results in the bimodal distributions shown above, so 0 will represent observations with a value of 0 and 1 will represent any observations with a value greater than 0. For `CAR_AGE`, many cars are 1 years old, so 0 represents observations where the `CAR_AGE` is 1, while 1 represents any observations with a value greater than 1.

### Box-Cox Transformation for Skewed Variables

Based on the previous distribution plot (using histograms) we noticed that a select group of columns exhibited non-normal skew. In order to address this skewness and attempt to normalize these features for future modeling, we will employ box-cox transformations. Because some of these values include 0, we will need to replace any zero values with infinitesimally small, non-zero values.

The  $\lambda$ 's that were used to transform the skewed variables are shown on Table 2.

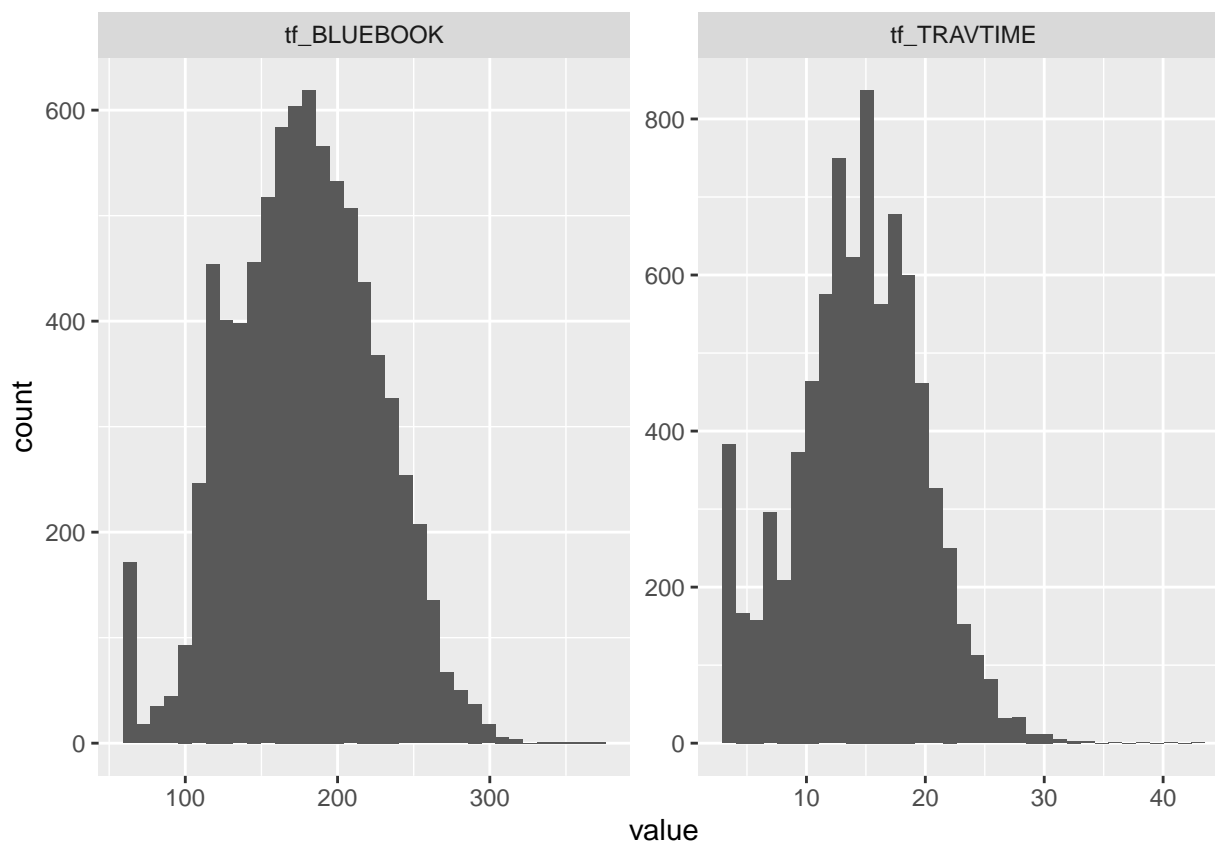


Figure 9: Histograms for transformed variables.

Column Name	$\lambda$
BLUEBOOK	0.461
TRAVTIME	0.687

Table 2:  $\lambda$ 's for skewed variables.

## Split Data Into Testing and Training

The data was into testing and training subsets such that 60% of it will be used to train, and 40% to test. The first row shows the split for the testing data while the second row shows the split for the training data.

```
0    1
1202 431
```

```
0    1
4806 1722
```

## Build Models

### Binary Logistic Regression Model with Original Variables

Call:

```
glm(formula = TARGET_FLAG ~ ., family = binomial(link = "logit"),
     data = original_train_no_target_amt)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4277	-0.7086	-0.3954	0.6203	3.1819

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.973e+00	5.286e-01	-3.733	0.000189 ***
KIDSDRIV	2.189e-01	1.373e-01	1.594	0.110912
AGE	-5.668e-03	4.590e-03	-1.235	0.216890
HOMEKIDS	2.281e-02	4.243e-02	0.538	0.590909
YOJ	2.059e-02	1.359e-02	1.515	0.129760
PARENT1Yes	3.938e-01	1.242e-01	3.170	0.001523 **
HOME_VAL	-1.240e-06	5.466e-07	-2.269	0.023289 *
MSTATUSYes	-3.250e-01	9.920e-02	-3.277	0.001051 **
SEX	1.582e-01	1.247e-01	1.269	0.204489
EDUCATIONBachelors	-4.013e-01	1.306e-01	-3.072	0.002128 **
EDUCATIONHigh School	-2.952e-02	1.075e-01	-0.275	0.783623
EDUCATIONMasters	-3.444e-01	2.063e-01	-1.670	0.094981 .
EDUCATIONPhD	-3.548e-01	2.415e-01	-1.469	0.141857
JOBBlue Collar	4.464e-01	2.076e-01	2.150	0.031554 *
JOB Clerical	5.857e-01	2.186e-01	2.680	0.007368 **
JOB Doctor	-3.209e-01	2.940e-01	-1.091	0.275165
JOB Home Maker	3.189e-01	2.401e-01	1.328	0.184098
JOB Lawyer	3.097e-01	1.893e-01	1.636	0.101936
JOB Manager	-4.536e-01	1.924e-01	-2.357	0.018419 *
JOB Professional	2.475e-01	2.000e-01	1.238	0.215807
JOB Student	1.859e-01	2.460e-01	0.756	0.449863
TRAVTIME	-2.167e-02	1.947e-02	-1.113	0.265651
CAR_USEPrivate	-7.367e-01	1.032e-01	-7.140	9.35e-13 ***
BLUEBOOK	4.036e-05	2.198e-05	1.836	0.066352 .
TIF	-4.863e-02	8.267e-03	-5.883	4.03e-09 ***
CAR_TYPEPanel Truck	3.875e-01	1.863e-01	2.080	0.037517 *
CAR_TYPEPickup	5.476e-01	1.120e-01	4.889	1.01e-06 ***

CAR_TYPESports Car	1.041e+00	1.447e-01	7.198	6.11e-13	***
CAR_TYPE SUV	8.011e-01	1.244e-01	6.441	1.19e-10	***
CAR_TYPEVan	6.429e-01	1.413e-01	4.549	5.39e-06	***
RED_CARyes	-5.252e-02	9.669e-02	-0.543	0.587050	
OLDCLAIM	-2.374e-05	4.712e-06	-5.038	4.69e-07	***
CLM_FREQ	6.644e-02	4.973e-02	1.336	0.181483	
REVOKEDYes	9.804e-01	1.032e-01	9.495	< 2e-16	***
MVR_PTS	9.301e-02	1.581e-02	5.883	4.02e-09	***
CAR_AGE	3.058e-03	1.220e-02	0.251	0.802027	
URBANICITYHighly Urban/ Urban	2.330e+00	1.289e-01	18.070	< 2e-16	***
EMPLOYMENT1	-7.136e-01	3.098e-01	-2.303	0.021281	*
bi_CAR_AGE1	-1.041e-01	1.205e-01	-0.864	0.387484	
bi_CLM_FREQ1	5.972e-01	1.361e-01	4.387	1.15e-05	***
bi_HOME_VAL1	-1.178e-01	1.466e-01	-0.804	0.421567	
bi_KIDSDRIV1	3.300e-01	2.170e-01	1.521	0.128360	
bi_YOJ1	-6.203e-02	3.425e-01	-0.181	0.856262	
tf_BLUEBOOK	-1.092e-02	3.669e-03	-2.975	0.002929	**
tf_TRAVTIME	1.042e-01	5.684e-02	1.834	0.066683	.

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 7533.1 on 6527 degrees of freedom  
Residual deviance: 5818.0 on 6483 degrees of freedom  
AIC: 5908

Number of Fisher Scoring iterations: 5

Setting levels: control = 0, case = 1

Setting direction: controls < cases

### Confusion Matrix for Binary Logistic Regression Model with Original Variables

The confusion matrix for the binary logistic regression model with original variables is provided below.

	Predicted	
Actual	0	1
0	1119	83
1	249	182

### Step-AIC Binary Logistic Regression Model

Call:

```
glm(formula = TARGET_FLAG ~ KIDSDRIV + YOJ + PARENT1 + HOME_VAL +
    MSTATUS + EDUCATION + JOB + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
    OLDCLAIM + REVOKED + MVR_PTS + URBANICITY + EMPLOYMENT +
    bi_CLM_FREQ + bi_KIDSDRIV + tf_BLUEBOOK + tf_TRAVTIME, family = binomial(link = "logit"),
    data = original_train_no_target_amt)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.4416	-0.7139	-0.3948	0.6362	3.1713

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-1.978e+00	4.569e-01	-4.329	1.50e-05	***
KIDSDRIV	2.336e-01	1.341e-01	1.742	0.081595	.
YOJ	1.895e-02	1.178e-02	1.608	0.107858	
PARENT1Yes	4.665e-01	1.070e-01	4.361	1.29e-05	***
HOME_VAL	-1.615e-06	3.539e-07	-4.565	5.00e-06	***
MSTATUSYes	-3.264e-01	9.083e-02	-3.594	0.000326	***
EDUCATIONBachelors	-4.207e-01	1.216e-01	-3.460	0.000540	***
EDUCATIONHigh School	-2.625e-02	1.067e-01	-0.246	0.805678	
EDUCATIONMasters	-3.533e-01	1.795e-01	-1.968	0.049101	*
EDUCATIONPhD	-3.612e-01	2.172e-01	-1.663	0.096324	.
JOBBlue Collar	4.352e-01	2.072e-01	2.101	0.035648	*
JOB Clerical	5.710e-01	2.172e-01	2.629	0.008565	**
JOB Doctor	-3.356e-01	2.933e-01	-1.144	0.252528	
JOB Home Maker	2.438e-01	2.335e-01	1.044	0.296575	
JOB Lawyer	2.766e-01	1.887e-01	1.466	0.142560	
JOB Manager	-4.763e-01	1.918e-01	-2.483	0.013022	*
JOB Professional	2.269e-01	1.994e-01	1.138	0.255136	
JOB Student	2.023e-01	2.428e-01	0.833	0.404756	
CAR_USEPrivate	-7.305e-01	1.029e-01	-7.102	1.23e-12	***
BLUEBOOK	3.745e-05	2.142e-05	1.748	0.080474	.
TIF	-4.819e-02	8.245e-03	-5.844	5.09e-09	***
CAR_TYPEPanel Truck	4.675e-01	1.738e-01	2.691	0.007126	**
CAR_TYPEPickup	5.408e-01	1.117e-01	4.843	1.28e-06	***
CAR_TYPESports Car	9.453e-01	1.200e-01	7.875	3.41e-15	***
CAR_TYPESUV	7.106e-01	9.576e-02	7.420	1.17e-13	***
CAR_TYPEVan	6.836e-01	1.372e-01	4.984	6.22e-07	***
OLDCLAIM	-2.382e-05	4.698e-06	-5.070	3.97e-07	***
REVOKEDYes	9.827e-01	1.031e-01	9.535	< 2e-16	***
MVR_PTS	9.336e-02	1.576e-02	5.923	3.16e-09	***
URBANICITYHighly Urban/ Urban	2.342e+00	1.287e-01	18.197	< 2e-16	***
EMPLOYMENT1	-7.599e-01	1.935e-01	-3.928	8.55e-05	***
bi_CLM_FREQ1	7.357e-01	8.797e-02	8.363	< 2e-16	***
bi_KIDSDRIV1	3.303e-01	2.166e-01	1.525	0.127193	
tf_BLUEBOOK	-1.093e-02	3.639e-03	-3.003	0.002675	**
tf_TRAVTIME	4.141e-02	6.217e-03	6.661	2.72e-11	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 7533.1 on 6527 degrees of freedom  
 Residual deviance: 5826.5 on 6493 degrees of freedom  
 AIC: 5896.5

Number of Fisher Scoring iterations: 5

Setting levels: control = 0, case = 1

Setting direction: controls < cases

## Confusion Matrix for Step-AIC Binary Logistic Regression Model

The confusion matrix for the Step-AIC binary logistic regression model with original variables is provided below.

		Predicted	
Actual	0	1	
	0	1121	81
1	251	180	

## Multiple Linear Regression Model with Original Variables

Call:

```
lm(formula = TARGET_AMT ~ ., data = original_train_no_target_flag)
```

Residuals:

Min	1Q	Median	3Q	Max
-6122	-1731	-758	383	103578

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.035e+03	9.389e+02	-1.102	0.270306
KIDSDRIV	-2.131e+02	2.666e+02	-0.799	0.424235
AGE	6.765e+00	8.199e+00	0.825	0.409327
HOMEKIDS	9.948e+01	7.672e+01	1.297	0.194804
YOJ	8.655e+00	2.358e+01	0.367	0.713527
PARENT1Yes	6.643e+02	2.331e+02	2.849	0.004395 **
HOME_VAL	-1.326e-03	9.005e-04	-1.473	0.140806
MSTATUSYes	-4.208e+02	1.746e+02	-2.411	0.015953 *
SEX_M	3.526e+02	2.096e+02	1.682	0.092521 .
EDUCATIONBachelors	-4.672e+02	2.373e+02	-1.969	0.048972 *
EDUCATIONHigh School	-2.185e+02	1.986e+02	-1.100	0.271312
EDUCATIONMasters	-1.432e+02	3.527e+02	-0.406	0.684856
EDUCATIONPhD	-1.521e+01	4.111e+02	-0.037	0.970497
JOBBlue Collar	3.979e+02	3.677e+02	1.082	0.279223
JOB_Clerical	5.630e+02	3.877e+02	1.452	0.146492
JOBDoctor	-4.532e+02	4.652e+02	-0.974	0.329982
JOBHome Maker	5.003e+02	4.229e+02	1.183	0.236858
JOBLawyer	3.994e+02	3.390e+02	1.178	0.238724
JOBManager	-5.411e+02	3.307e+02	-1.636	0.101841
JOBProfessional	4.261e+02	3.529e+02	1.208	0.227252
JOBStudent	3.026e+02	4.398e+02	0.688	0.491432
TRAVTIME	-8.385e+00	3.223e+01	-0.260	0.794775
CAR_USEPrivate	-8.551e+02	1.883e+02	-4.541	5.69e-06 ***
BLUEBOOK	-1.745e-02	3.776e-02	-0.462	0.644096
TIF	-4.595e+01	1.410e+01	-3.259	0.001123 **
CAR_TYPEPanel Truck	3.053e+02	3.262e+02	0.936	0.349320
CAR_TYPEPickup	3.410e+02	1.946e+02	1.752	0.079750 .
CAR_TYPESports Car	8.628e+02	2.480e+02	3.479	0.000507 ***
CAR_TYPESUV	7.592e+02	2.051e+02	3.701	0.000216 ***
CAR_TYPEVan	3.924e+02	2.449e+02	1.602	0.109204
RED_CARYes	2.414e+01	1.710e+02	0.141	0.887760
OLDCLAIM	-1.662e-02	9.096e-03	-1.827	0.067734 .
CLM_FREQ	1.041e+02	1.013e+02	1.028	0.304105

REVOKEDYes	6.170e+02	2.005e+02	3.078	0.002093	**
MVR_PTS	1.857e+02	3.081e+01	6.027	1.76e-09	***
CAR_AGE	-3.576e+01	2.075e+01	-1.724	0.084831	.
URBANICITYHighly Urban/ Urban	1.670e+03	1.624e+02	10.283	< 2e-16	***
EMPLOYMENT1	-8.873e+02	5.458e+02	-1.626	0.104070	
bi_CAR_AGE1	1.362e+02	2.138e+02	0.637	0.523994	
bi_CLM_FREQ1	2.563e+02	2.702e+02	0.949	0.342846	
bi_HOME_VAL1	1.268e+02	2.587e+02	0.490	0.623971	
bi_KIDSDRIV1	9.308e+02	4.197e+02	2.218	0.026605	*
bi_YOJ1	4.712e+02	6.054e+02	0.778	0.436442	
tf_BLUEBOOK	4.971e+00	6.455e+00	0.770	0.441217	
tf_TRAVTIME	6.028e+01	9.394e+01	0.642	0.521102	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4658 on 6483 degrees of freedom  
Multiple R-squared: 0.07341, Adjusted R-squared: 0.06713  
F-statistic: 11.67 on 44 and 6483 DF, p-value: < 2.2e-16

### Step-AIC Multiple Linear Regression Model

Call:

```
lm(formula = TARGET_AMT ~ PARENT1 + HOME_VAL + MSTATUS + SEX +
    JOB + CAR_USE + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
    MVR_PTS + CAR_AGE + URBANICITY + EMPLOYMENT + bi_KIDSDRIV +
    tf_TRAVTIME, data = original_train_no_target_flag)
```

Residuals:

Min	1Q	Median	3Q	Max
-6253	-1733	-771	335	103540

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.485e+00	5.295e+02	-0.008	0.993242
PARENT1Yes	7.673e+02	2.039e+02	3.762	0.000170 ***
HOME_VAL	-1.052e-03	5.966e-04	-1.764	0.077810 .
MSTATUSYes	-3.213e+02	1.564e+02	-2.055	0.039932 *
SEX	2.913e+02	1.667e+02	1.748	0.080522 .
JOBBlue Collar	2.868e+02	3.025e+02	0.948	0.343066
JOB Clerical	4.688e+02	3.264e+02	1.436	0.150999
JOB Doctor	-3.842e+02	4.256e+02	-0.903	0.366691
JOB Home Maker	2.998e+02	3.838e+02	0.781	0.434837
JOB Lawyer	3.539e+02	3.280e+02	1.079	0.280718
JOB Manager	-7.279e+02	3.062e+02	-2.377	0.017488 *
JOB Professional	1.619e+02	3.032e+02	0.534	0.593342
JOB Student	1.048e+02	3.815e+02	0.275	0.783478
CAR_USEPrivate	-7.628e+02	1.789e+02	-4.263	2.04e-05 ***
TIF	-4.476e+01	1.406e+01	-3.184	0.001461 **
CAR_TYPEPanel Truck	4.907e+02	2.857e+02	1.718	0.085933 .
CAR_TYPEPickup	3.621e+02	1.915e+02	1.891	0.058725 .
CAR_TYPESports Car	7.843e+02	2.320e+02	3.380	0.000730 ***
CAR_TYPESUV	7.012e+02	1.886e+02	3.718	0.000202 ***
CAR_TYPEVan	4.907e+02	2.368e+02	2.072	0.038292 *

OLDCLAIM	-1.314e-02	8.509e-03	-1.544	0.122649	
CLM_FREQ	1.793e+02	6.337e+01	2.829	0.004688	**
REVOKEDYes	5.837e+02	1.981e+02	2.946	0.003229	**
MVR_PTS	1.905e+02	2.974e+01	6.406	1.60e-10	***
CAR_AGE	-3.094e+01	1.259e+01	-2.457	0.014050	*
URBANICITYHighly Urban/ Urban	1.694e+03	1.603e+02	10.573	< 2e-16	***
EMPLOYMENT1	-3.937e+02	2.777e+02	-1.418	0.156290	
bi_KIDSDRIV1	7.462e+02	1.855e+02	4.023	5.82e-05	***
tf_TRAVTIME	3.551e+01	1.080e+01	3.288	0.001014	**

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4657 on 6499 degrees of freedom  
Multiple R-squared: 0.07155, Adjusted R-squared: 0.06755  
F-statistic: 17.89 on 28 and 6499 DF, p-value: < 2.2e-16

### Parsed Step-AIC Multiple Linear Regression Model

In this model, we selected the variables from the original Step-AIC Model that had p-values that were less than 0.05

Call:

```
lm(formula = TARGET_AMT ~ PARENT1 + MSTATUS + JOB + CAR_USE +
    TIF + CAR_TYPE + CLM_FREQ + MVR_PTS + CAR_AGE + URBANICITY +
    bi_KIDSDRIV + tf_TRAVTIME, data = original_train_no_target_flag)
```

Residuals:

Min	1Q	Median	3Q	Max
-5777	-1728	-777	338	103635

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-333.30	432.62	-0.770	0.441077	
PARENT1Yes	760.23	203.85	3.729	0.000194	***
MSTATUSYes	-469.46	137.50	-3.414	0.000643	***
JOBBlue Collar	359.24	300.01	1.197	0.231173	
JOB Clerical	577.97	320.89	1.801	0.071728	.
JOB Doctor	-416.16	425.36	-0.978	0.327926	
JOB Home Maker	553.57	352.70	1.570	0.116576	
JOB Lawyer	387.28	328.01	1.181	0.237767	
JOB Manager	-714.36	306.05	-2.334	0.019621	*
JOB Professional	192.81	302.85	0.637	0.524367	
JOB Student	530.98	342.05	1.552	0.120630	
CAR_USEPrivate	-749.20	178.88	-4.188	2.85e-05	***
TIF	-45.71	14.06	-3.251	0.001157	**
CAR_TYPEPanel Truck	553.79	280.90	1.971	0.048715	*
CAR_TYPEPickup	388.20	191.55	2.027	0.042741	*
CAR_TYPESports Car	617.53	207.97	2.969	0.002996	**
CAR_TYPESUV	544.39	158.81	3.428	0.000612	***
CAR_TYPEVan	543.40	234.50	2.317	0.020516	*
CLM_FREQ	145.00	56.15	2.583	0.009829	**
MVR_PTS	191.92	29.62	6.479	9.91e-11	***
CAR_AGE	-33.16	12.59	-2.634	0.008460	**
URBANICITYHighly Urban/ Urban	1723.57	159.76	10.789	< 2e-16	***

```

bi_KIDSDRIV1          756.44      185.49    4.078 4.59e-05 ***
tf_TRAVTIME           36.38       10.80    3.367 0.000763 ***
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4662 on 6504 degrees of freedom  
Multiple R-squared: 0.06895, Adjusted R-squared: 0.06566  
F-statistic: 20.94 on 23 and 6504 DF, p-value: < 2.2e-16

## Model Selection

### Binary Logistic Regression Models

Model	Precision	Recall	AIC	AUC	F-score	Accuracy	Error
Simple Log Reg	0.69	0.42	5908.03	0.82	0.523	0.8	0.2
Step AIC Log Reg	0.69	0.42	5896.48	0.82	0.52	0.8	0.2

Table 1: Model metrics

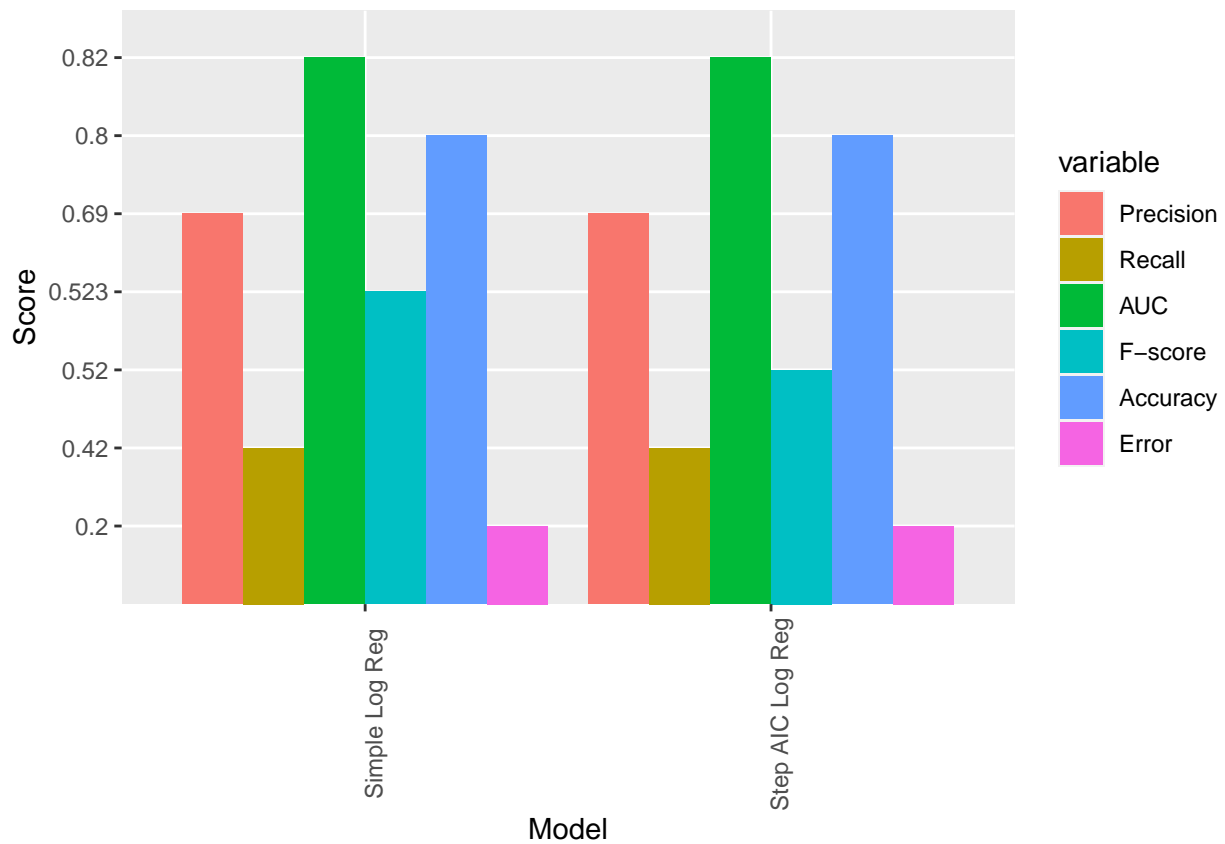


Figure 7: Bar chart of metrics for all 5 models

For this assignment, we will be choosing the Simple Log Reg for our binary logistic regression model. Between the two models, the simple binary logistic regression model has a higher f-score than the Step-AIC Logistic Regression model.



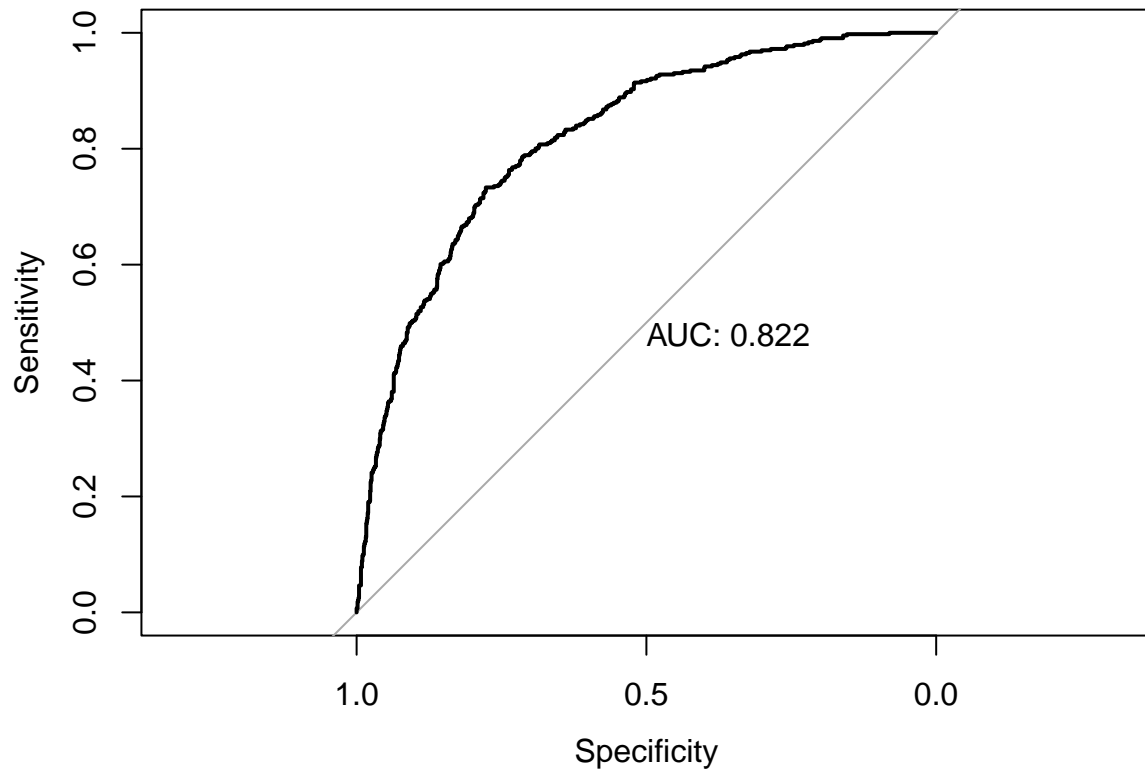


Figure xx: ROC Curve for selected model (Simple Model)

As we see on Figure xx, our model performs really well with an AUC of 0.822.

### Multiple Linear Regression Models

Model	MSE	R-Squared	Adjusted R-Squared	F-Statistic
Simple Linear	16581656.19	0.073	0.067	11.67
Step-AIC Linear	16559343.49	0.072	0.068	17.89
Parsed Step-AIC Linear	16612545.4	0.069	0.066	20.94

Table 1: Model metrics

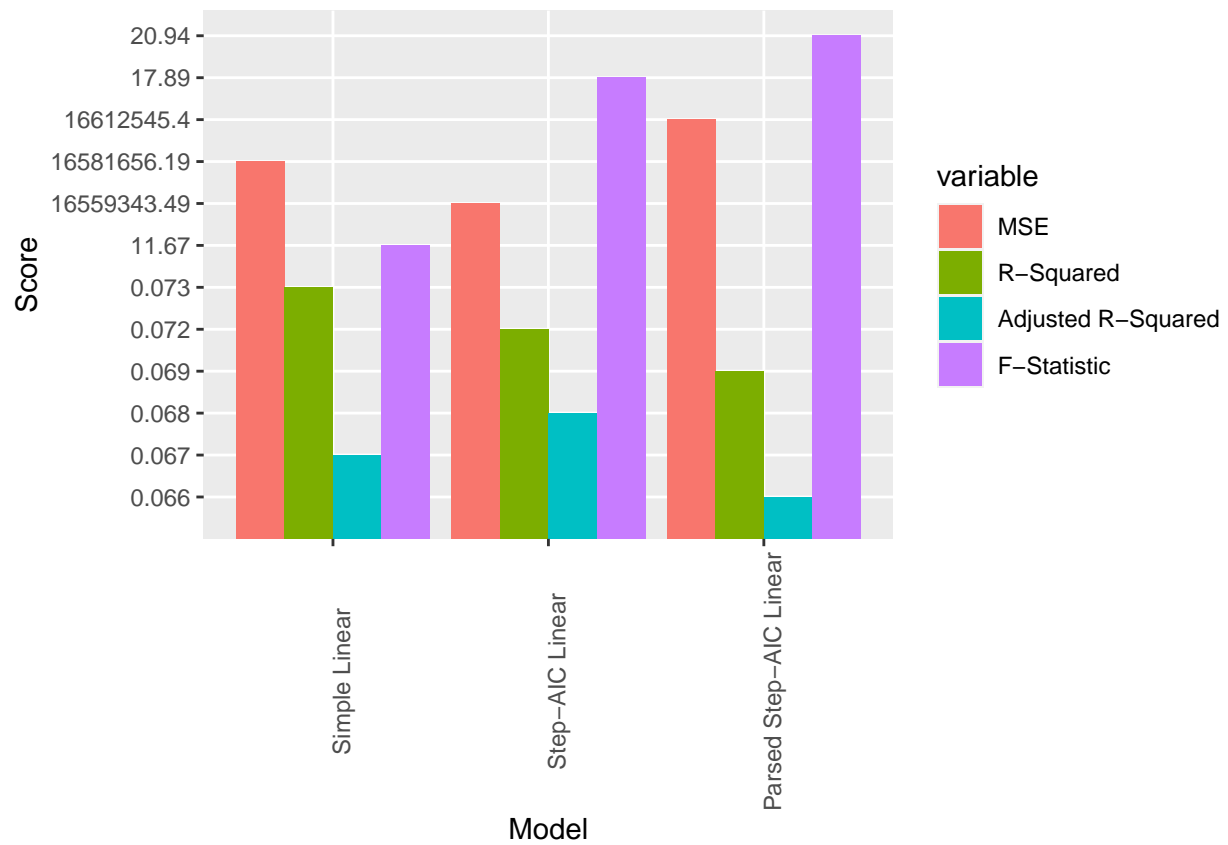


Figure xx: Metrics bar chart for multiple linear regression models

For this assignment, we will be choosing the Step-AIC Linear model for our multiple linear regression model. Between the three models, the multiple linear regression model has the highest adjusted R-squared and the lowest MSE.

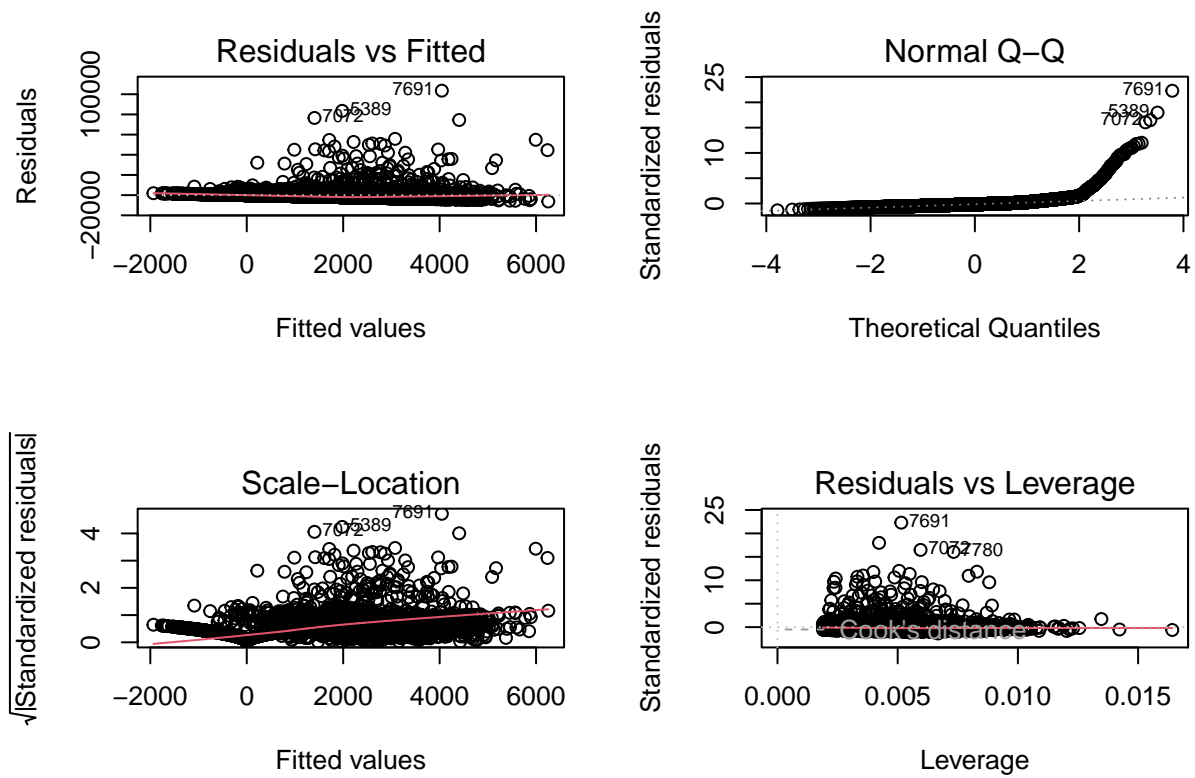


Figure xx: Residual Plots for Step-AIC Linear Model