Home Work Assignment - 04

Critical Thinking Group 5
Arindam Barman
Mohamed Elmoudni
Shazia Khan
Kishore Prasad

Contents


```
## NULL
```

1 Overview

The data set contains approximately 8161 records. Each record represents a customer profile 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 is the amount spent on repairs if there was a crash. 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.

We will be exploring, analyzing, and modeling the training data. Since there are 2 different predictions, we will deal with each prediction independently. The following are the 2 predictions we will be modeling for:

1. TARGET_FLAG - This dependent variable tells whether there was a crash or not. This is a binary variable and as such we will be using a Logistic Regression Model.
2. TARGET_AMT - This dependent variable gives the amount / cost of repairs if there was a crash. This is a continuous variable and we will be using a Linear Regression.

Each of the above models will be built and evaluated separately. In the first section of this document we will deal with the Logistic Model for TARGET_FLAG and in the second section we will deal with Linear Model for the TARGET_AMT.

Out of the many models for each task, we will shortlist one model that works best. We will then use these models (one for each task) on the test / evaluation data.

To attain our objective, we will be follow the below steps for each modeling exercise:

1 -Data Exploration
2 -Data Preparation
3 -Build Models
4 -Select Models

Please note that for Model Selection, as strategy, we will split the train dataset into 2 parts: TRAIN and VALID. In the VALID dataset, we will hold out some values to validate how well the model is trained using the TRAIN dataset. Then we will use the Model that performs best on the EVALUATION data to give the required output. We will split the TRAIN / VALID data after the Data Exploration / Preparation before the Build Models secion.

Please Note:

- There are some common clean-up and transformations that we will carry out initially that will serve all the models.
- While working on the Linear Models for the TARGET_AMT, we will be using only a subset of the data where the TARGET_FLAG = 1. This will give us all the records where there was a crash and subsequently a repair amount.
- While Predicting the TARGET_AMT with the given Evaluation dataset, We will take the output of the TARGET_FLAG predictions on the Evaluation dataset and use only those rows that were classified as a "Crash" and use it as the input to the TARGET_AMT prediction. So this is a two step prediction, one for the TARGET_FLAG and using the output to predict TARGET_AMT.

2 Data Exploration and Cleanup / Common Transformations

In this section we go ahead and perform some common cleanup and create additional variables that will be used for modeling both the logistic as well as the linear regressions. We will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

- Variable Identification / Relationships
- Data Clean-up
- Common Transformations
- Create Missing Flags / Impute Missing Values

2.1 Variable Identification

First let's display and examine the data dictionary or the data columns as shown in below table:

Table 1: Variable Description

VARIABLE	DEFINITION	THEORETICAL_EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes

VARIABLE	DEFINITION	THEORETICAL_EFFECT
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

We notice that there are 2 dependent variables - TARGET_FLAG and TARGET_AMT. Apart from these 2 dependent variables, we have 23 independent or predictor variables.

2.2 Data Cleanup

From the Variable Level table below we can make the following observations:

- Some of the variables like MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE, URBANICITY have some of the values encoded with "z_". Not that this will impact the analysis, but it will look a bit odd. So we will be fixing this.
- EDUCATION has 2 "High School" values - one starting with "<" and another starting with "z_". It is assumed that both these values are to be converted to "HIGH School".
- JOB has a "" value. This would indicate that the job is unknown or is not coded. Hence, we will replace this with "Unknown".

Table 2: Variable Levels

MSTATUS	SEX	EDUCATION	CAR_TYPE	URBANICITY	CAR_USE	REVOKED	JOB
Yes	M	<High School	Minivan	Highly Urban/ Urban	Commercial	No	
z_No	z_F	Bachelors	Panel Truck	z_Highly Rural/ Rural	Private	Yes	Clerical
Yes	M	Masters	Pickup	Highly Urban/ Urban	Commercial	No	Doctor
z_No	z_F	PhD	Sports Car	z_Highly Rural/ Rural	Private	Yes	Home Maker
Yes	M	z_High School	Van	Highly Urban/ Urban	Commercial	No	Lawyer
z_No	z_F	<High School	z_SUV	z_Highly Rural/ Rural	Private	Yes	Manager

MSTATUS	SEX	EDUCATION	CAR_TYPE	URBANICITY	CAR_USE	REVOKED	JOB
Yes	M	Bachelors	Minivan	Highly Urban/ Urban	Commercial	No	Professional
z_No	z_F	Masters	Panel Truck	z_Highly Rural/ Rural	Private	Yes	Student
Yes	M	PhD	Pickup	Highly Urban/ Urban	Commercial	No	z_Blue Collar

In addition, from the summary output, some numeric variables like INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM have been converted to Factor variables which need to be rectified.
In addition, there are records where CAR_AGE is negative or zero, which is improbable. Upon investigation, we find that there are 4 records that are affected. We will remove these records.

2.3 Common Transformations

In this section, we will create dummy variables for all the factors. Below is a summary of the old and new variable transformation:

Table 3: Variable Transformaton

Old.Variable	Old.Value	New.Variable	New.Value
CAR_USE	Commercial	CAR_USE_Commercial	1
	private		0
MSTATUS	Yes	MSTATUS_Yes	1
	No		0
PARENT1	Yes	PARENT1_Yes	1
	No		0
RED_CAR	Yes	RED_CAR_yes	1
	No		0
SEX	M	SEX_M	1
	else		0
URBANICITY	Highly Rural/ Rural	URBANICITY_Rura	1
	else		0

- Please note that we will not be using INDEX variable as it serves as just an identifier for each row. And has no relationships to other variables.

Making the above fixes to the data, we now have a "clean" dataset which can be explored further.

2.4 Create Missing Flags / Impute Missing Values

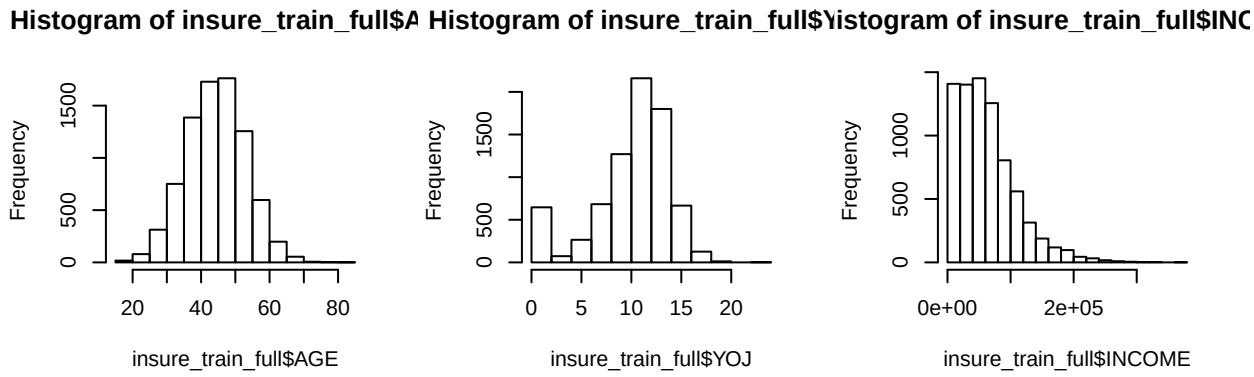
Based on the missing data from the below table, we can see that there are a few missing values for AGE, YOJ, INCOME, HOME_VAL, CAR_AGE variables. We will create flags to indicate that there are missing values in some of the variables.

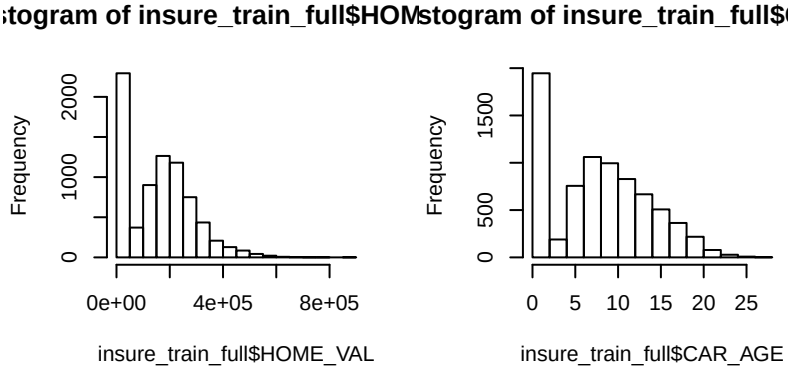
Table 4: Missing Values

	missings
TARGET_FLAG	0
TARGET_AMT	0
KIDSDRIV	0
AGE	6

	missings
HOMEKIDS	0
YOJ	454
INCOME	445
HOME_VAL	464
TRAVTIME	0
BLUEBOOK	0
TIF	0
OLDCLAIM	0
CLM_FREQ	0
MVR_PTS	0
CAR_AGE	510
CAR_USE_Commercial	0
MSTATUS_Yes	0
PARENT1_Yes	0
RED_CAR_yes	0
REVOKED_Yes	0
SEX_M	0
URBANICITY_Rural	0
EDUCATION_Bachelors	0
EDUCATION_High.School	0
EDUCATION_Masters	0
EDUCATION_PhD	0
JOB_Blue.Collar	0
JOB_Clerical	0
JOB_Doctor	0
JOB_Home.Maker	0
JOB_Lawyer	0
JOB_Manager	0
JOB_Professional	0
JOB_Student	0
JOB_Unknown	0
CAR_TYPE_Minivan	0
CAR_TYPE_Panel.Truck	0
CAR_TYPE_Pickup	0
CAR_TYPE_Sports.Car	0
CAR_TYPE_SUV	0
CAR_TYPE_Van	0

We now impute values to AGE, YOJ, INCOME, HOME_VAL, CAR_AGE. However, while doing the impute, we will impute to a new variable so as not to impact the original variables. We will look at the distributions for each of the variable to determine the value to use to impute. Given that Age and YOJ look to be somewhat normally distributed, we can go ahead and use the mean to impute the missing values for these variables. For INCOME, HOME_VAL and CAR_AGE the median seems to be a better value to impute since there are strong right skews. We will carry out these transformation while data preparation.





Now that we are done with the common clean-up and transformations, we can proceed to each specific model as below.

3 Logistic Regression for TARGET_FLAG

In this section we will use Logistic regression to model the TARGET_FLAG. We will first start with the Data Summary and Correlation.

3.1.1 Data Summary

In this section, we will create summary data to better understand the relationship each of the variables have with our dependent variables using correlation, central tendency, and dispersion as shown below:

Table 5: Data Summary

	vars	n	mean	sd	median	trimmed	mad
TARGET_FLAG	1	8157	2.638225e-01	4.407312e-01	0	2.048414e-01	0.0000
TARGET_AMT	2	8157	1.504882e+03	4.705092e+03	0	5.944205e+02	0.0000
KIDSDRIV	3	8157	1.708962e-01	5.112480e-01	0	2.527960e-02	0.0000
AGE	4	8157	4.479021e+01	8.626488e+00	45	4.483051e+01	8.8956
HOMEKIDS	5	8157	7.207307e-01	1.116104e+00	0	4.967060e-01	0.0000
YOJ	6	8157	1.049825e+01	3.977187e+00	11	1.104615e+01	2.9652
INCOME	7	8157	6.147181e+04	4.629838e+04	54046	5.655987e+04	38967.1758
HOME_VAL	8	8157	1.552083e+05	1.254299e+05	161160	1.450380e+05	131595.5760
TRAVTIME	9	8157	3.348903e+01	1.590913e+01	33	3.299908e+01	16.3086
BLUEBOOK	10	8157	1.571152e+04	8.420070e+03	14440	1.503899e+04	8450.8200
TIF	11	8157	5.350129e+00	4.145349e+00	4	4.839589e+00	4.4478
OLDCLAIM	12	8157	4.030096e+03	8.767493e+03	0	1.717216e+03	0.0000
CLM_FREQ	13	8157	7.982101e-01	1.158368e+00	0	5.883254e-01	0.0000
MVR_PTS	14	8157	1.695844e+00	2.147412e+00	1	1.314386e+00	1.4826
CAR_AGE	15	8157	8.312247e+00	5.517924e+00	8	7.962157e+00	5.9304
CAR_USE_Commercial	16	8157	3.713375e-01	4.831921e-01	0	3.392064e-01	0.0000
MSTATUS_Yes	17	8157	5.996077e-01	4.900079e-01	1	6.244829e-01	0.0000
PARENT1_Yes	18	8157	1.320338e-01	3.385483e-01	0	4.014100e-02	0.0000
RED_CAR_yes	19	8157	2.915287e-01	4.544943e-01	0	2.394668e-01	0.0000
REVOKED_Yes	20	8157	1.223489e-01	3.277084e-01	0	2.803740e-02	0.0000
SEX_M	21	8157	4.637734e-01	4.987165e-01	0	4.547265e-01	0.0000
URBANICITY_Rural	22	8157	2.044869e-01	4.033509e-01	0	1.306879e-01	0.0000
EDUCATION_Bachelors	23	8157	2.747334e-01	4.464072e-01	0	2.184771e-01	0.0000
EDUCATION_High.School	24	8157	4.330023e-01	4.955214e-01	0	4.162709e-01	0.0000
EDUCATION_Masters	25	8157	2.031384e-01	4.023593e-01	0	1.290026e-01	0.0000
EDUCATION_PhD	26	8157	8.912590e-02	2.849429e-01	0	0.000000e+00	0.0000
JOB_Blue.Collar	27	8157	2.237342e-01	4.167715e-01	0	1.547418e-01	0.0000
JOB_Clerical	28	8157	1.556945e-01	3.625877e-01	0	6.971040e-02	0.0000
JOB_Doctor	29	8157	3.003560e-02	1.706956e-01	0	0.000000e+00	0.0000
JOB_Home.Maker	30	8157	7.858280e-02	2.691030e-01	0	0.000000e+00	0.0000
JOB_Lawyer	31	8157	1.023661e-01	3.031477e-01	0	3.064200e-03	0.0000
JOB_Manager	32	8157	1.210004e-01	3.261477e-01	0	2.635210e-02	0.0000
JOB_Professional	33	8157	1.368150e-01	3.436730e-01	0	4.611610e-02	0.0000
JOB_Student	34	8157	8.728700e-02	2.822725e-01	0	0.000000e+00	0.0000
JOB_Unknown	35	8157	6.448450e-02	2.456291e-01	0	0.000000e+00	0.0000
CAR_TYPE_Minivan	36	8157	2.627191e-01	4.401381e-01	0	2.034625e-01	0.0000
CAR_TYPE_Panel.Truck	37	8157	8.287360e-02	2.757080e-01	0	0.000000e+00	0.0000
CAR_TYPE_Pickup	38	8157	1.700380e-01	3.756892e-01	0	8.763600e-02	0.0000
CAR_TYPE_Sports.Car	39	8157	1.111928e-01	3.143901e-01	0	1.409530e-02	0.0000
CAR_TYPE_SUV	40	8157	2.812308e-01	4.496274e-01	0	2.265972e-01	0.0000
CAR_TYPE_Van	41	8157	9.194560e-02	2.889668e-01	0	0.000000e+00	0.0000
YOJ_MISS	42	8157	5.565770e-02	2.292736e-01	0	0.000000e+00	0.0000
INCOME_MISS	43	8157	5.455440e-02	2.271222e-01	0	0.000000e+00	0.0000
HOME_VAL_MISS	44	8157	5.688370e-02	2.316344e-01	0	0.000000e+00	0.0000
CAR_AGE_MISS	45	8157	6.252300e-02	2.421178e-01	0	0.000000e+00	0.0000

Table 6: Data Summary (Cont)

	min	max	range	skew	kurtosis	se
TARGET_FLAG	0	1.0	1.0	1.0716217	-0.8517313	0.0048799
TARGET_AMT	0	107586.1	107586.1	8.7043164	112.2364581	52.0958190
KIDSDRIV	0	4.0	4.0	3.3544611	11.8047140	0.0056607
AGE	16	81.0	65.0	-0.0289590	-0.0609233	0.0955144
HOMEKIDS	0	5.0	5.0	1.3425541	0.6535085	0.0123577
YOJ	0	23.0	23.0	-1.2379372	1.4229728	0.0440363
INCOME	0	367030.0	367030.0	1.2446991	2.4535676	512.6259028
HOME_VAL	0	885282.0	885282.0	0.4947333	0.1567474	1388.7880484
TRAVTIME	5	142.0	137.0	0.4466705	0.6644712	0.1761494
BLUEBOOK	1500	69740.0	68240.0	0.7942563	0.7914970	93.2288765
TIF	1	25.0	24.0	0.8908250	0.4237528	0.0458982
OLDCLAIM	0	57037.0	57037.0	3.1245638	9.9049577	97.0756265
CLM_FREQ	0	5.0	5.0	1.2096086	0.2865283	0.0128257
MVR_PTS	0	13.0	13.0	1.3475573	1.3742211	0.0237766
CAR_AGE	1	28.0	27.0	0.3026810	-0.5951905	0.0610957
CAR_USE_Commercial	0	1.0	1.0	0.5324869	-1.7166681	0.0053500
MSTATUS_Yes	0	1.0	1.0	-0.4065056	-1.8349781	0.0054255
PARENT1_Yes	0	1.0	1.0	2.1735218	2.7245310	0.0037485
RED_CAR_yes	0	1.0	1.0	0.9172643	-1.1587682	0.0050323
REVOKED_Yes	0	1.0	1.0	2.3045169	3.3112039	0.0036285
SEX_M	0	1.0	1.0	0.1452613	-1.9791417	0.0055219
URBANICITY_Rural	0	1.0	1.0	1.4651104	0.1465665	0.0044660
EDUCATION_Bachelors	0	1.0	1.0	1.0091192	-0.9817988	0.0049427
EDUCATION_High.School	0	1.0	1.0	0.2703797	-1.9271310	0.0054865
EDUCATION_Masters	0	1.0	1.0	1.4754234	0.1768960	0.0044550
EDUCATION_PhD	0	1.0	1.0	2.8835517	6.3156446	0.0031550
JOB_Blue.Collar	0	1.0	1.0	1.3255796	-0.2428685	0.0046146
JOB_Clerical	0	1.0	1.0	1.8989243	1.6061105	0.0040147
JOB_Doctor	0	1.0	1.0	5.5057871	28.3171632	0.0018900
JOB_Home.Maker	0	1.0	1.0	3.1316299	7.8080632	0.0029796
JOB_Lawyer	0	1.0	1.0	2.6230461	4.8809695	0.0033565
JOB_Manager	0	1.0	1.0	2.3238133	3.4005250	0.0036112
JOB_Professional	0	1.0	1.0	2.1132906	2.4662995	0.0038052
JOB_Student	0	1.0	1.0	2.9238588	6.5497534	0.0031254
JOB_Unknown	0	1.0	1.0	3.5456888	10.5732053	0.0027197
CAR_TYPE_Minivan	0	1.0	1.0	1.0780788	-0.8378488	0.0048733
CAR_TYPE_Panel.Truck	0	1.0	1.0	3.0254856	7.1544401	0.0030527
CAR_TYPE_Pickup	0	1.0	1.0	1.7563536	1.0849109	0.0041597
CAR_TYPE_Sports.Car	0	1.0	1.0	2.4731030	4.1167430	0.0034810
CAR_TYPE_SUV	0	1.0	1.0	0.9729937	-1.0534124	0.0049784
CAR_TYPE_Van	0	1.0	1.0	2.8238842	5.9750546	0.0031995
YOJ_MISS	0	1.0	1.0	3.8756126	13.0219693	0.0025386
INCOME_MISS	0	1.0	1.0	3.9220378	13.3840215	0.0025147
HOME_VAL_MISS	0	1.0	1.0	3.8255290	12.6362214	0.0025647
CAR_AGE_MISS	0	1.0	1.0	3.6133098	11.0573631	0.0026808

3.1.2 Correlations

Now we will produce the correlation table between the independent variables and the dependent variable - TARGET_FLAG

Table 7: Correlation between TARGET_FLAG and predictor variables

	Correlation_TARGET_FLAG
TARGET_FLAG	1.0000000
TARGET_AMT	0.5343138
MVR_PTS	0.2192671
CLM_FREQ	0.2159652
PARENT1_Yes	0.1576594
REVOKED_Yes	0.1517045
CAR_USE_Commercial	0.1427163
EDUCATION_High.School	0.1382094
OLDCLAIM	0.1378435
HOMEKIDS	0.1161499
KIDSDRIV	0.1040583
JOB_Blue.Collar	0.1018097
JOB_Student	0.0770293
CAR_TYPE_Sports.Car	0.0572627
CAR_TYPE_Pickup	0.0563353
TRAVTIME	0.0480461
CAR_TYPE_SUV	0.0450376
JOB_Clerical	0.0275791
JOB_Home.Maker	0.0112577
CAR_AGE_MISS	0.0085607
YOJ_MISS	0.0039126
CAR_TYPE_Van	0.0030163
CAR_TYPE_Panel.Truck	-0.0003471
HOME_VAL_MISS	-0.0016978
JOB_Unknown	-0.0031380
RED_CAR_yes	-0.0069595
INCOME_MISS	-0.0090653
SEX_M	-0.0206620
JOB_Professional	-0.0391996
EDUCATION_Bachelors	-0.0431408
JOB_Doctor	-0.0580794
JOB_Lawyer	-0.0617528
EDUCATION_PhD	-0.0652170
YOJ	-0.0684748
EDUCATION_Masters	-0.0761613
TIF	-0.0821748
CAR_AGE	-0.0974530
AGE	-0.1032152
BLUEBOOK	-0.1035337
JOB_Manager	-0.1052506
MSTATUS_Yes	-0.1347552
CAR_TYPE_Minivan	-0.1367604
INCOME	-0.1377852
HOME_VAL	-0.1785848
URBANICITY_Rural	-0.2241940

The above table suggests that none of the variables seem to have a very strong correlation with TARGET_FLAG. However, CAR_TYPE_Van, RED_CAR_no, JOB_Home.Maker, SEX_F, JOB_Clerical, CAR_TYPE_SUV, TRAVTIME, CAR_TYPE_Pickup, CAR_TYPE_Sports.Car, JOB_Student, JOB_Blue.Collar, KIDSDRIV, HOMEKIDS, MSTATUS_No, OLDCLAIM, EDUCATION_High.School, CAR_USE_Commercial, REVOKED_Yes, PARENT1_Yes, CLM_FREQ, MVR_PTS, URBANICITY_Highly.Urban..Urban have a positive correlation.

Similarly, URBANICITY_Highly.Rural..Rural, HOME_VAL, PARENT1_No, REVOKED_No, CAR_USE_Private, INCOME, CAR_TYPE_Minivan, MSTATUS_Yes, JOB_Manager, BLUEBOOK, AGE, CAR_AGE, TIF, EDUCATION_Masters, YOJ, EDUCATION_PhD, JOB_Lawyer, JOB_Doctor, EDUCATION_Bachelors, JOB_Professional, SEX_M, RED_CAR_yes, CAR_TYPE_Panel.Truck have a negative correlation.

Lets now see how values in some of the variable affects the correlation:

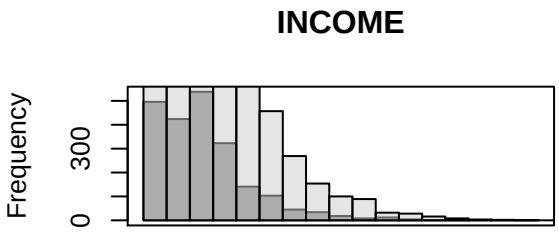
CAR_TYPE - If you drive Minivans and Panel Trucks you have lesser chance of being in a crash as against Pickups, Sports, SUVs and Vans. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation.

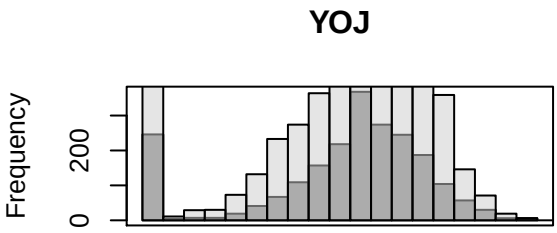
EDUCATION - If you have only a high school education then you are more likely to crash than if you have a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation.

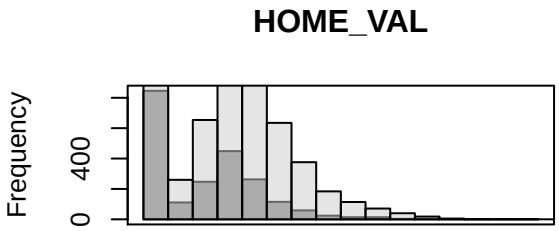
JOB - If you are a Student, Homemaker, or in a Blue Collar or Clerical job, you are more likely to be in a crash against Doctor, Lawyer, Manager, professional or Unknown job. Again binning this variable will strengthen the correlation.

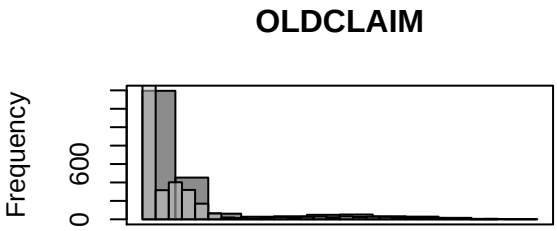
3.1.3 Binning of Variables

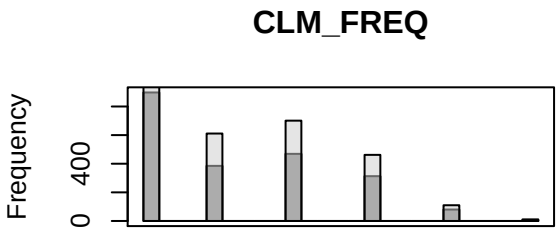
Lets have a look at the following numeric variables to see how they are distributed vis-a-vis TARGET_FLAG: INCOME, YOJ, HOME_VAL, OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, AGE, BLUEBOOK, TIF, TRAVTIME. The goal here is to see if we can bin these variables into zero and non-zero bin values and check the correlations. While doing that we will also see how the variables are distributed vis-a-vis TARGET_FLAG.

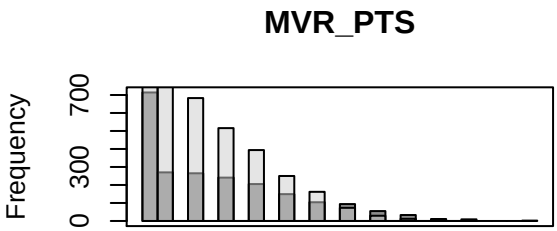

INCOME
Frequency
0
300

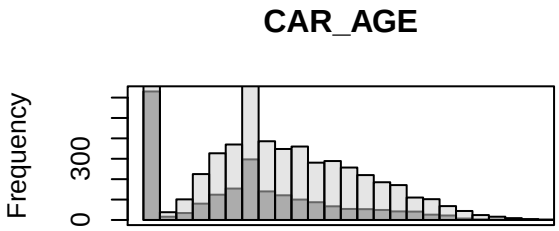

YOJ
Frequency
0
200

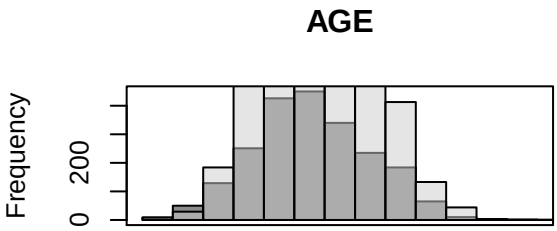

HOME_VAL
Frequency
0
400

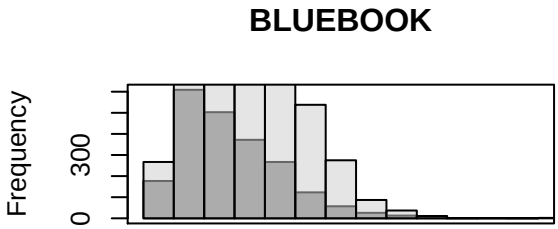

OLDCLAIM
Frequency
0
600

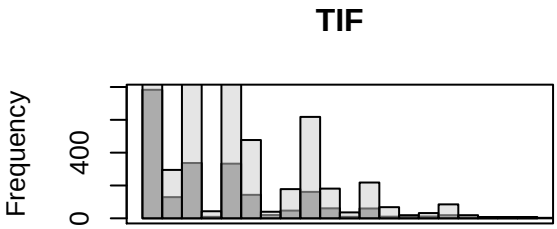

CLM_FREQ
Frequency
0
400

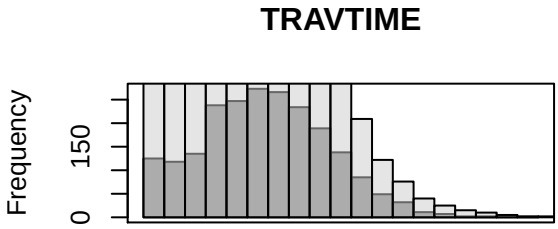

MVR_PTS
Frequency
0
300
700


CAR_AGE
Frequency
0
300


AGE
Frequency
0
200







From the outputs above, we can come to the following conclusions:

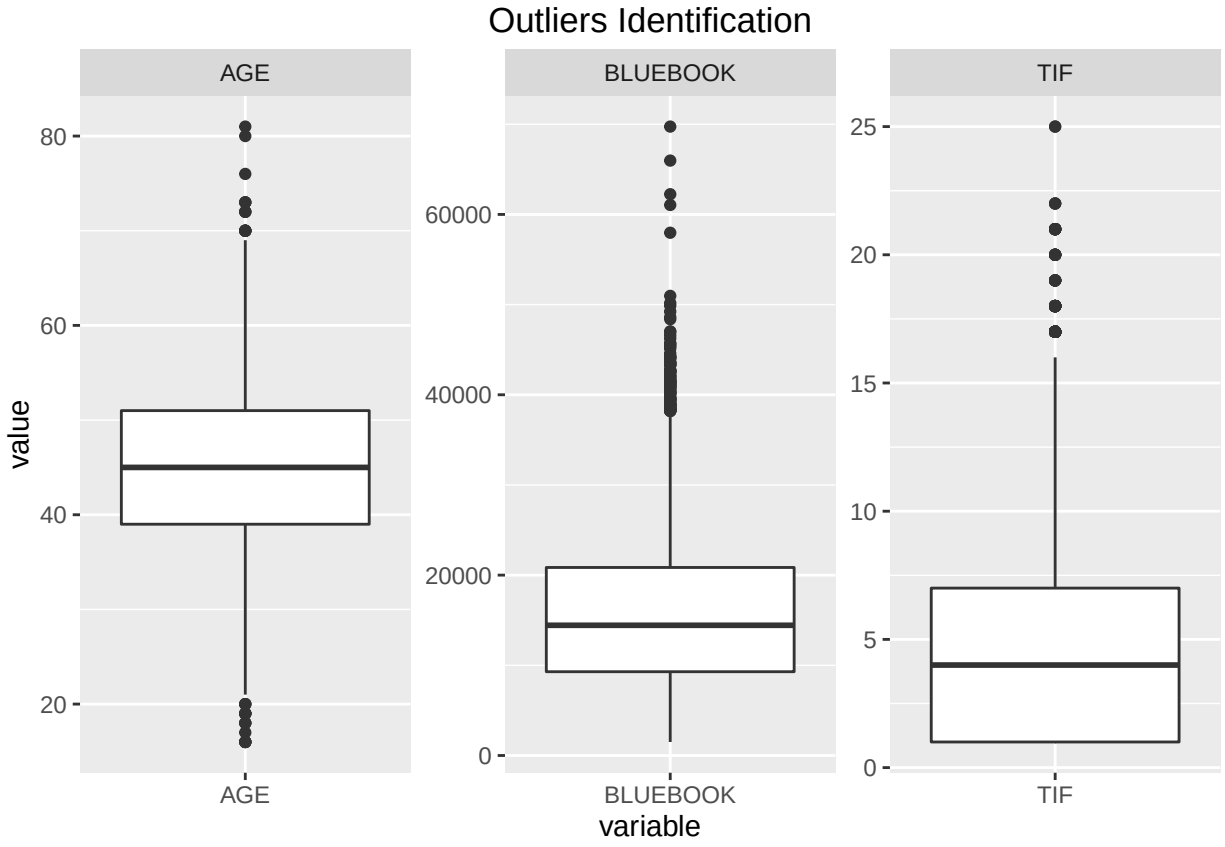
- INCOME - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this at zero value.
- YOJ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.
- HOME_VAL - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.
- OLDCLAIM- There is a huge difference in the coorrelation when we transform this vatiable. Binning this variable seems like a good idea.
- CLM_FREQ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.
- MVR_PTS - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.
- AGE - There is no specific pattern that emerges. We will retain this variable as is.
- BLUEBOOK - There is no specific pattern that emerges. We will retain the variable as is.
- TIF - Looking at the plots, values and the correlations with TARGET_FLAG, we can conclude that this is not a good variable for binning. We will retain this variable as is.

- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

We will carry out the above transformations in the Data Preparation phase.

3.1.4 Outliers identification

In this sub-section, we will look at the boxplots and determine the outliers in variables and decide on whether to act on the outliers. We will do the outliers only on some of the currency and few other variables. Below are the plots:



From the "Outliers identification" plot above, we see that we have few outliers that we need to treat. We will treat the outliers in this variable when we do the data preparation for modeling the TARGET_FLAG.

3.1.5 Analysis of the link function

In this section, we will investigate how our initial data aligns with a typical logistic model plot.

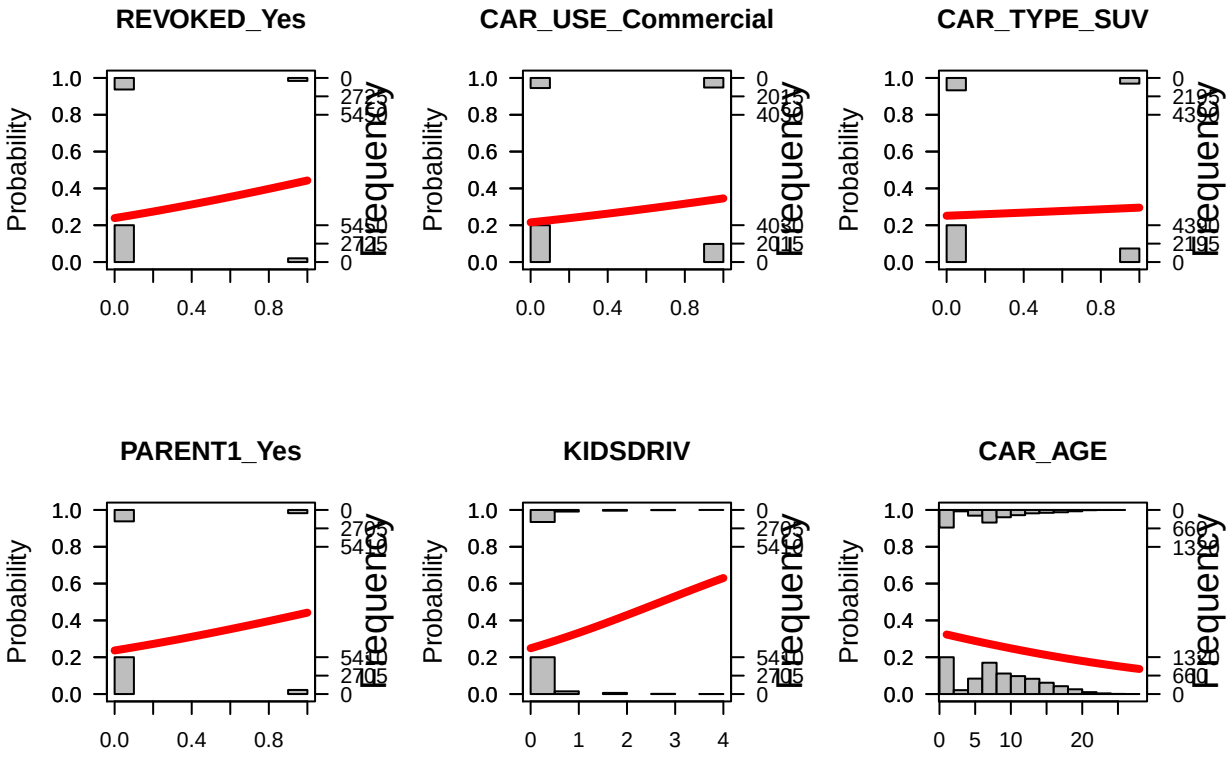
Recall the Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). The fundamental equation of generalized linear model is:

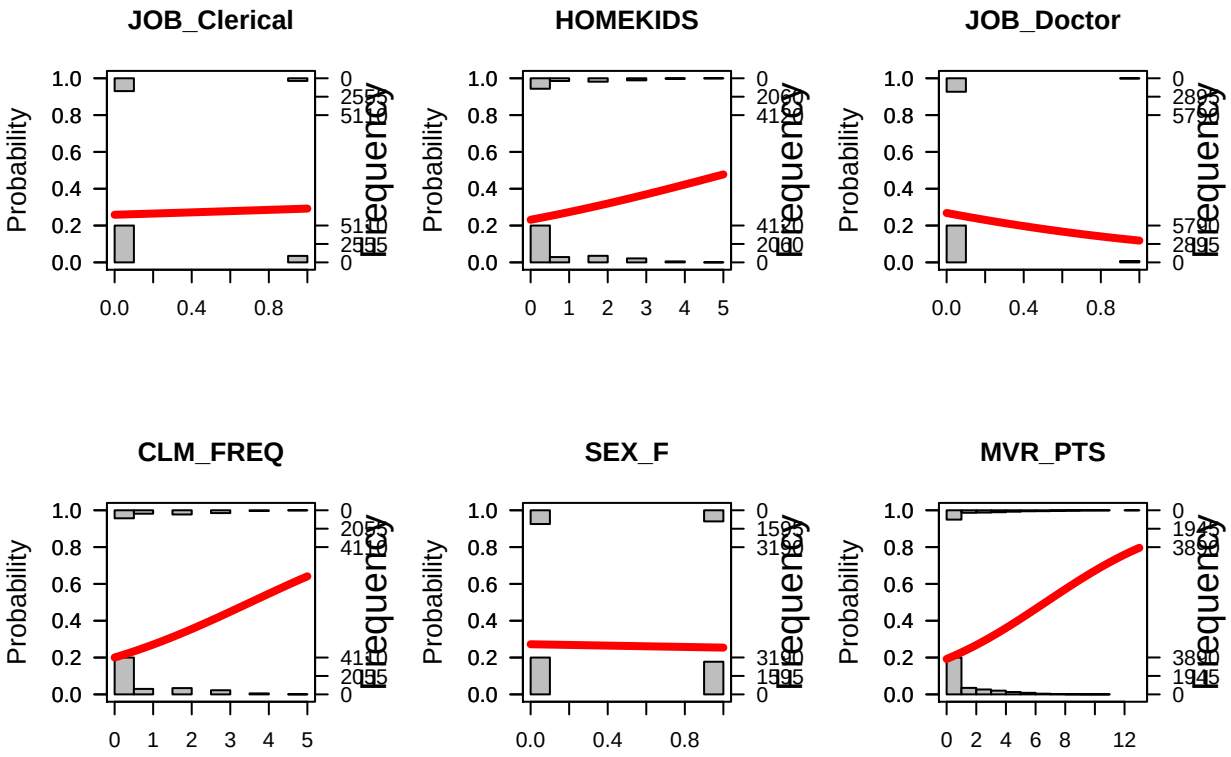
$g(E(y)) = a + Bx_1 + B_2x_2 + B_3x_3 + ...$

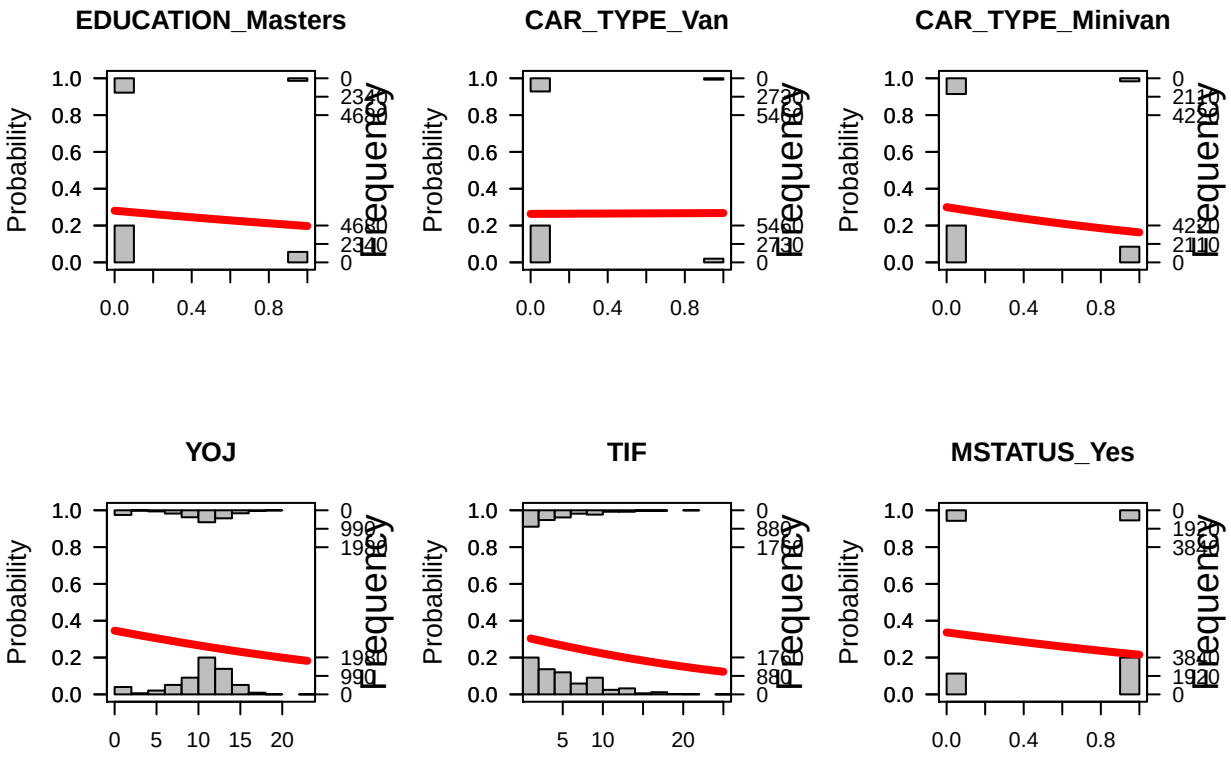
where, g() is the link function, E(y) is the expectation of target variable and $B_0 + B_1x_1 + B_2x_2 + B_3x_3$ is the linear predictor (B_0, B_1, B_2, B_3 to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

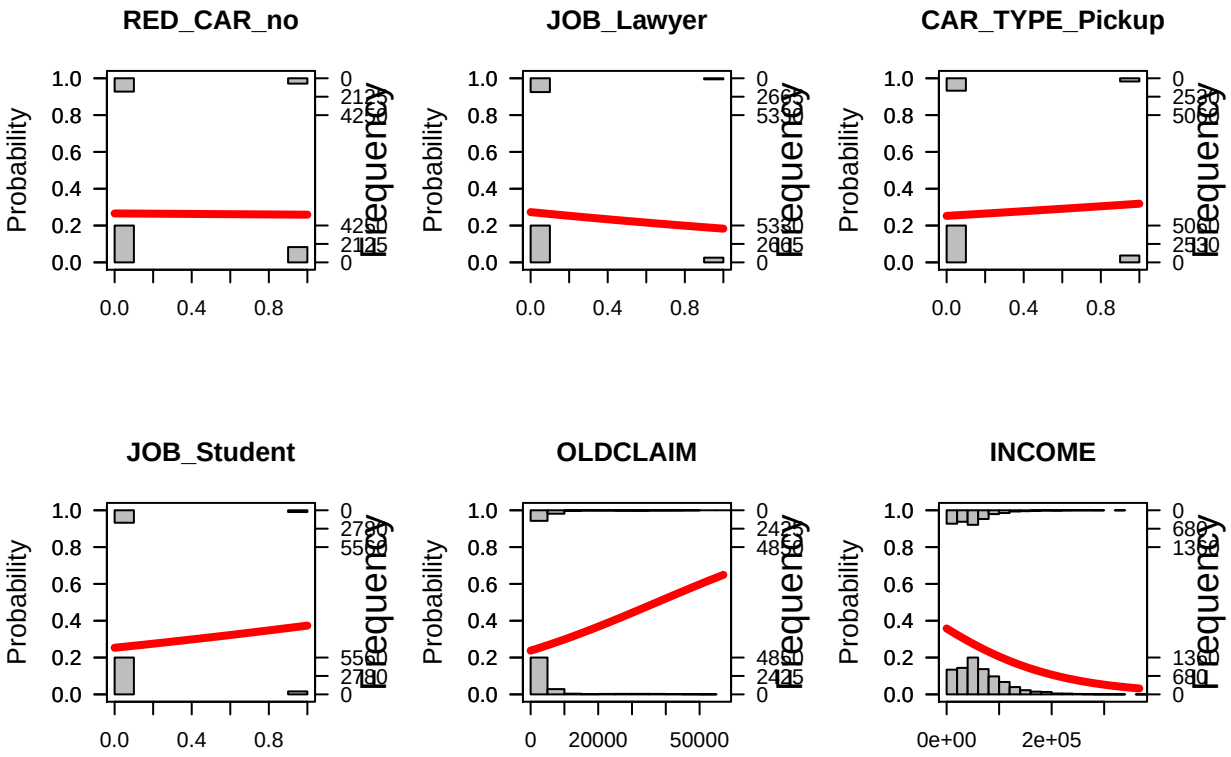
In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). As described above, g() is the link function. This function is established using two things: Probability of Success (p) and Probability of Failure (1-p). p should meet following criteria: It must always be positive (since p >= 0) It must always be less than equals to 1 (since p <= 1).

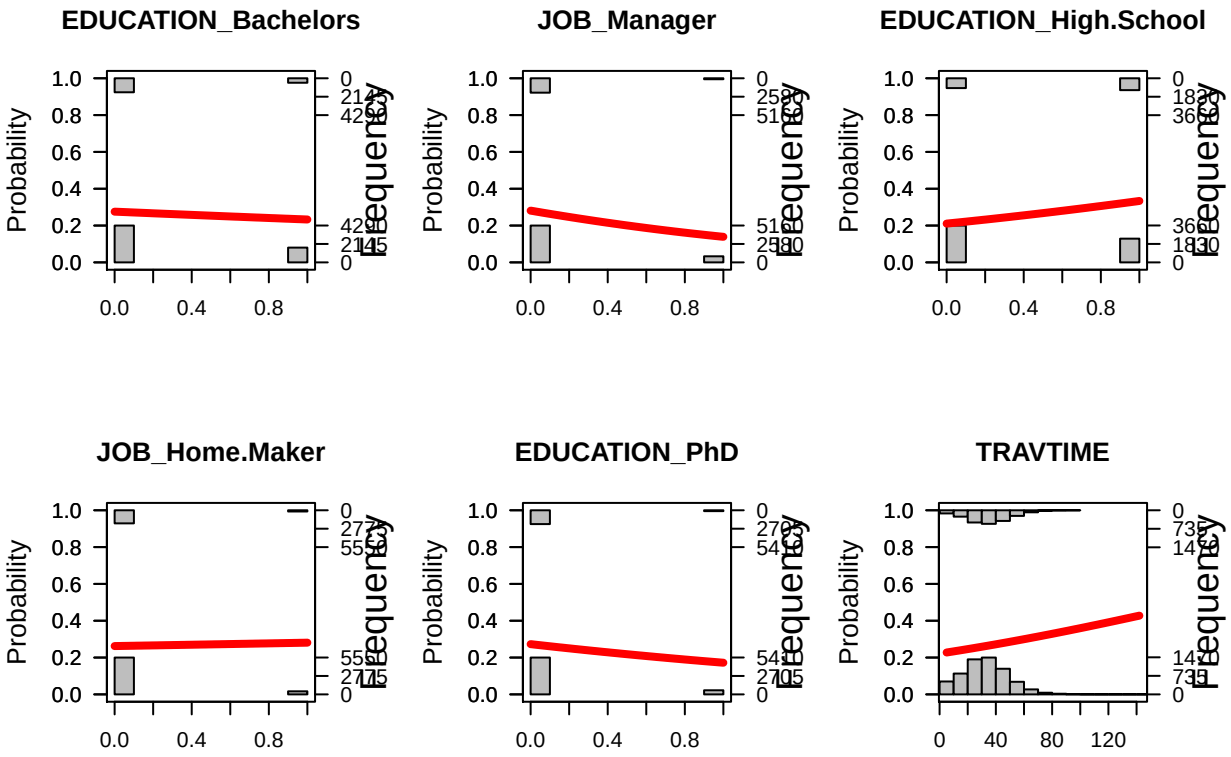
Now let's investigate how our initial data model aligns with the above criteria. In other words, we will plot regression model plots for each variable and compare it to a typical logistic model plot:

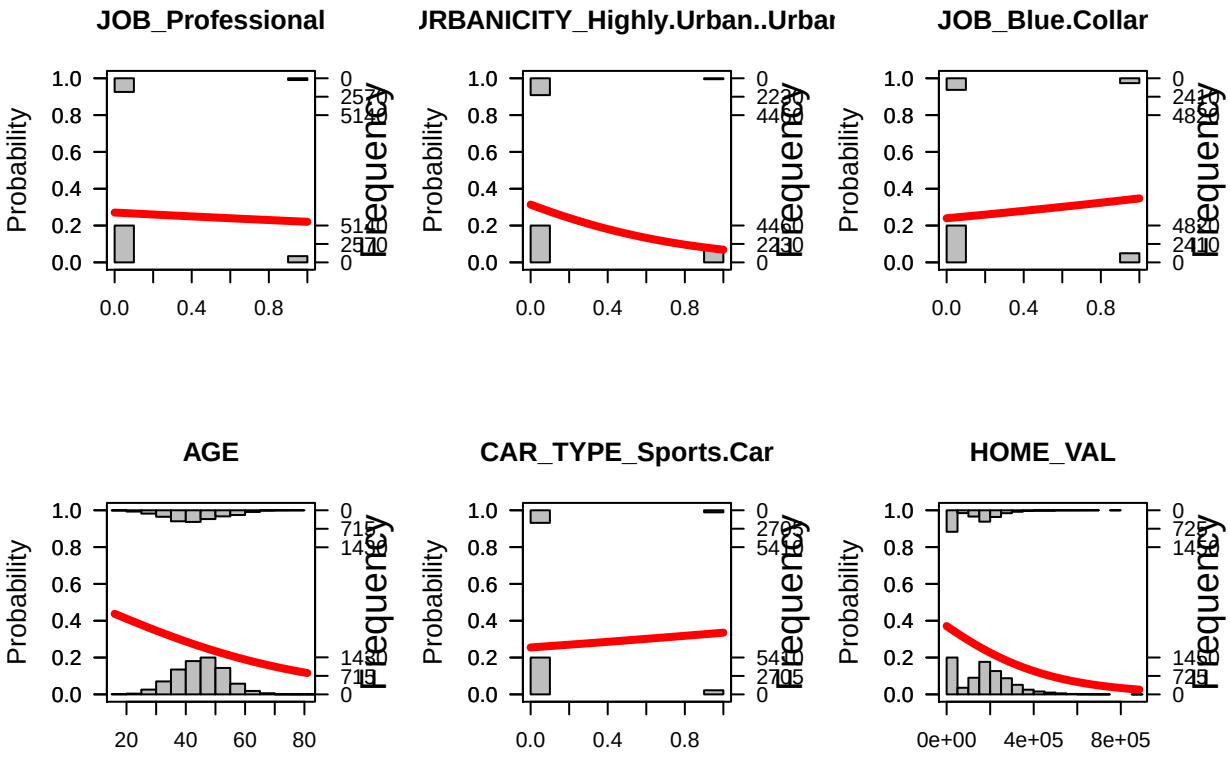


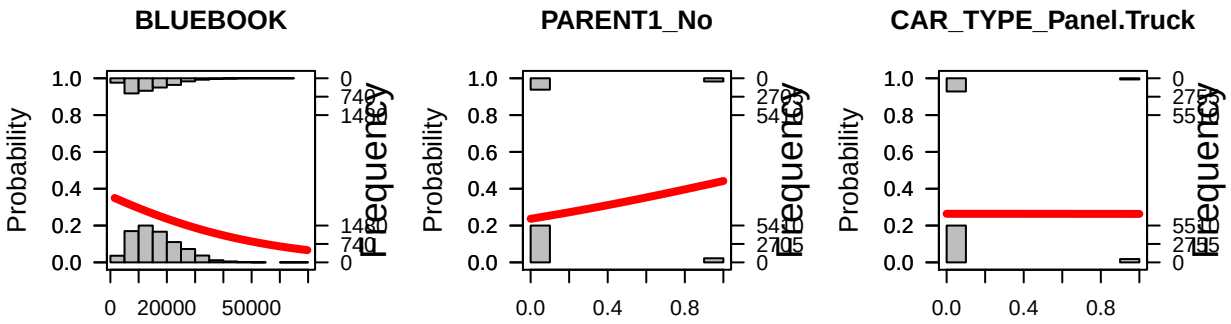

JOB_Clerical
HOMEKIDS
JOB_Doctor
CLM_FREQ
SEX_F
MVR_PTS
Probability
Frequency


EDUCATION_Masters
CAR_TYPE_Van
CAR_TYPE_Minivan
YOJ
TIF
MSTATUS_Yes
Probability
Frequency


RED_CAR_no
JOB_Lawyer
CAR_TYPE_Pickup
JOB_Student
OLDCLAIM
INCOME
Probability
Frequency


EDUCATION_Bachelors
JOB_Manager
EDUCATION_High.School
JOB_Home.Maker
EDUCATION_PhD
TRAVTIME
Probability
Frequency


JOB_Professional
JRBANICITY_Highly.Urban..Urbar
JOB_Blue.Collar
AGE
CAR_TYPE_Sports.Car
HOME_VAL
Probability
Frequency



3.1.5.1 Interpretation

You can see that the probability of crashing increases as we get closer to the "1" classification for the CAR_TYPE_Van, RED_CAR_no, JOB_Home.Maker, SEX_F, JOB_Clerical, CAR_TYPE_SUV, TRAVTIME, BLUEBOOK, CAR_TYPE_Pickup, CAR_TYPE_Sports.Car, JOB_Student, KIDSDRIV, JOB_Blue.Collar, HOMEKIDS, MSTATUS_No, EDUCATION_High.School, CAR_USE_Commercial, REVOKED_Yes, PARENT1_Yes, OLDCLAIM, CLM_FREQ, MVR_PTS, URBANICITY_Highly.Urban..Urban variables.

You can see that the probability of crashing decreases as we get closer to the "1" classification for the HOME_VAL, CAR_TYPE_Minivan, MSTATUS_Yes, JOB_Manager, AGE, CAR_AGE, TIF, EDUCATION_Masters, YOJ, EDUCATION_PhD, JOB_Lawyer, JOB_Doctor, EDUCATION_Bachelors, JOB_Professional, INCOME, SEX_M, RED_CAR_yes, CAR_TYPE_Panel.Truck variables.

3.2 Data Preparation

Now that we have completed the data exploration / analysis, we will be transforming the data for use in analysis and modeling.

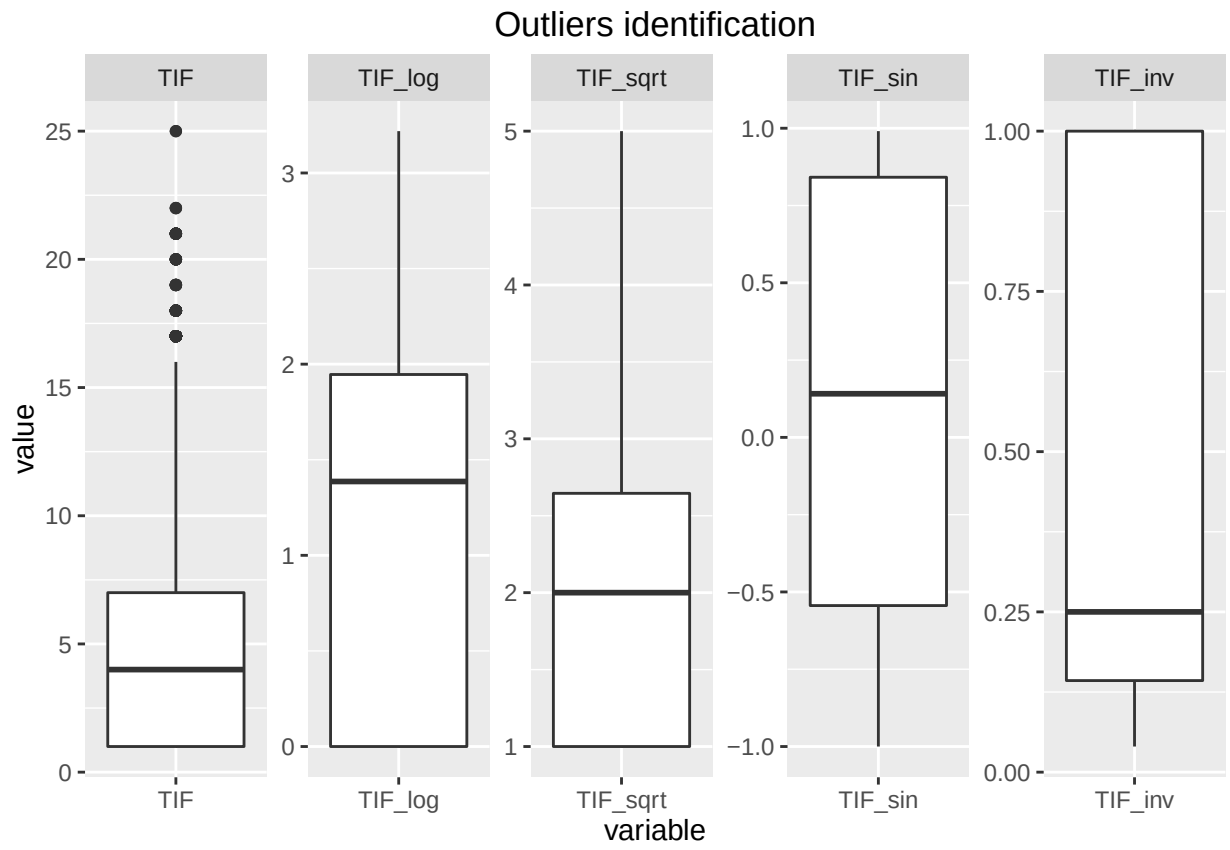
We will be following the below steps as guidelines: - Outliers treatment - Adding New Variables

3.2.1 Outliers treatment

In this sub-section, we will check different transformations for AGE, BLUEBOOK and TIF to create the appropriate outlier-handled / transformed variables.

- Transformations for TIF

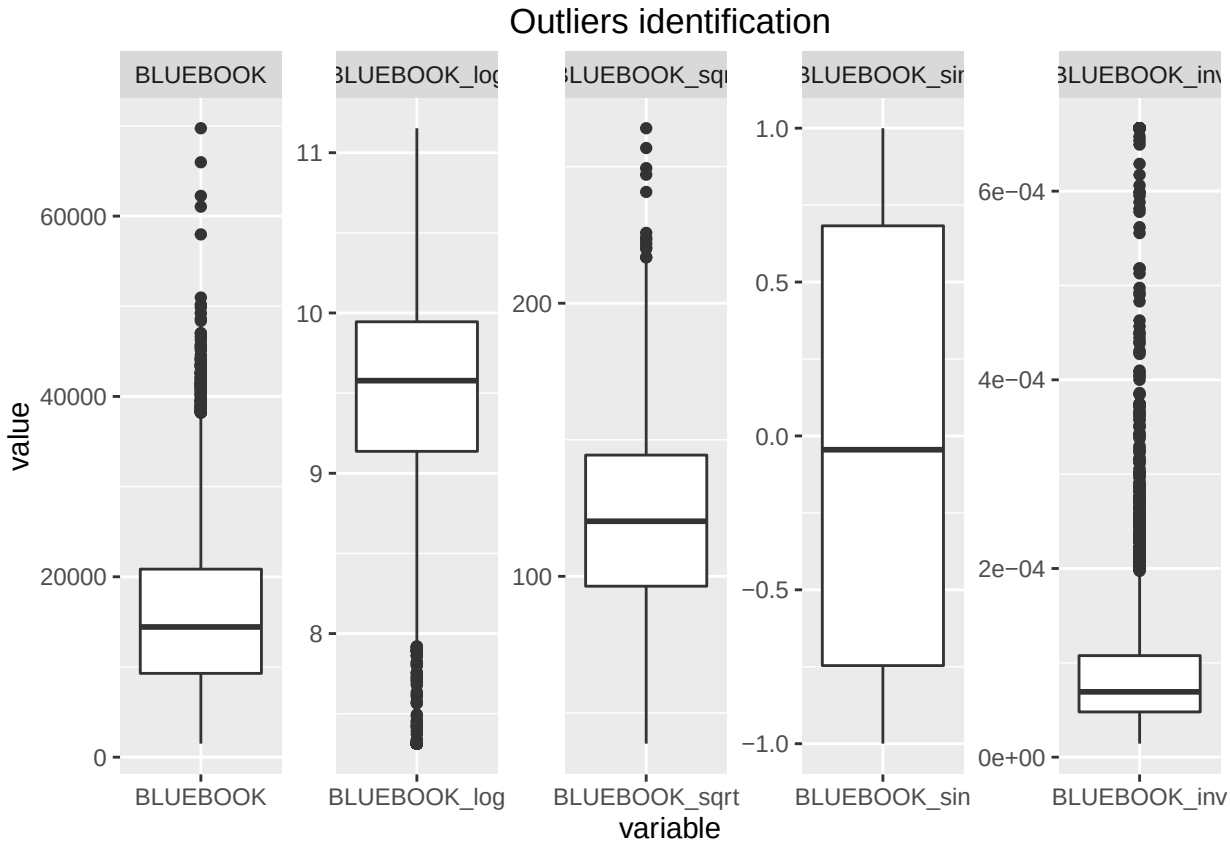
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    4.00    5.35    7.00   25.00
```



From the above charts we can see that a log, sqrt, sin or an inverse transformation works well for TIF. However, the sin transformation seems to be better distributed. Hence, We will create this variable.

- Transformations for BLUEBOOK

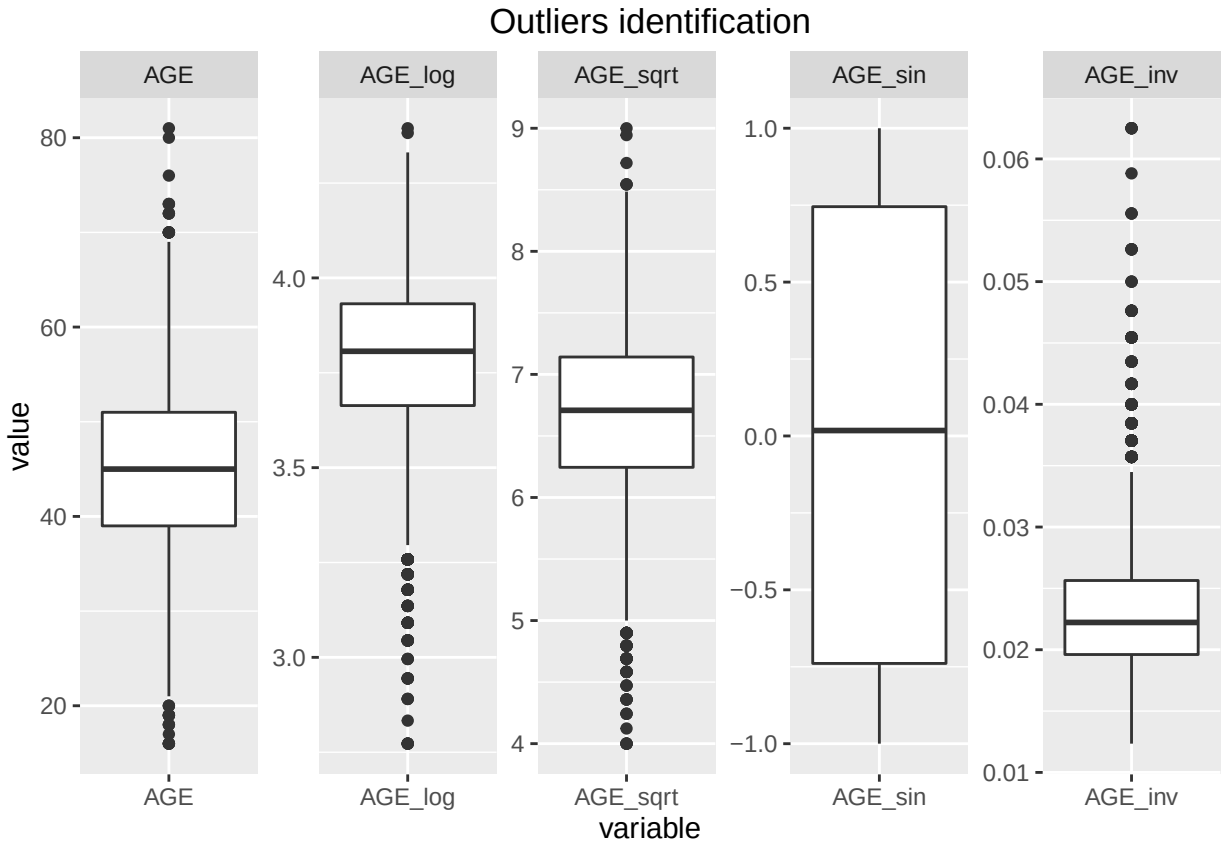
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1500    9290   14440   15710   20850   69740
```



From the above charts we can see that a sin transformation works well. Hence, We will create this variable.

- Transformations for AGE

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



From the above charts we can see that a sin works well for AGE.Hence, We will create this variable.

3.2.2 Adding New Variables

In this section, we generate some additional variables that we feel will help the correlations. The following were some of the observations we made during the data exploration phase for TARGET_FLAG

CAR_TYPE - If you drive Minivans and Panel Trucks you have lesser chance of being in a crash as against Pickups, Sports, SUVs and Vans. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation. Accordingly, we will bin this variable as below:

CAR_TYPE_FLAG_BIN :

- 1 : if CAR_TYPE is Minivans or Panel Trucks
- 0 : if CAR_TYPE is Pickups, Sports, SUVs or Vans

EDUCATION - If you have only a high school education then you are more likely to crash than if you have a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation:

EDUCATION_FLAG_BIN :

- 0 : if EDUCATION is High School
- 1 : if EDUCATION is Bachelors, Masters or Phd

JOB - If you are a Student, Homemaker, or in a Blue Collar or Clerical job, you are more likely to be in a crash against Doctor, Lawyer, Manager or professional. Again binning this variable will strengthen the correlation:

JOB_TYPE_FLAG_BIN :

- 1 : if JOB_TYPE is Student, Homemaker, or in a Blue Collar or Clerical
- 0 : if JOB_TYPE is Doctor, Lawyer, Manager, professional, Unknown
- INCOME - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this at zero value.

INCOME_FLAG_BIN :

- 1 : if INCOME <= 0
- 0 : if INCOME > 0
- YOJ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

YOJ_FLAG_BIN :

- 1 : if YOJ <= 0
- 0 : if YOJ > 0
- HOME_VAL - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

HOME_VAL_FLAG_BIN :

- 1 : if HOME_VAL <= 0
- 0 : if HOME_VAL > 0
- OLDCLAIM- There is a huge difference in the coorrelation when we transform this vatiable. Binning this variable seems like a good idea.

OLDCLAIM_FLAG_BIN :

- 1 : if OLDCLAIM <= 0
- 0 : if OLDCLAIM > 0
- CLM_FREQ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

CLM_FREQ_FLAG_BIN :

- 1 : if CLM_FREQ <= 0
- 0 : if CLM_FREQ > 0
- MVR_PTS - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

MVR_PTS_FLAG_BIN :

- 1 : if MVR_PTS <= 0
- 0 : if MVR_PTS > 0
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.

CAR_AGE_FLAG_BIN :

- 1 : if CAR_AGE <= 1
- 0 : if CAR_AGE > 0
- AGE - There is no specific pattern that emerges. We will retain this variable as is.
- BLUEBOOK - There is no specific pattern that emerges. We will retain this variable as is.
- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

TRAVTIME_FLAG_BIN :

- 1 : if TRAVTIME <= 20
- 0 : if TRAVTIME > 0

3.2.3 Additional Binned Variables

After having prepared the data, we will go ahead and drop some of the variables.

```
## 'data.frame':    8157 obs. of  39 variables:
##  $ TARGET_FLAG       : int  0 0 0 0 0 1 0 1 1 0 ...
##  $ KIDSDRIV          : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE               : num  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS          : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ               : num  11 11 10 14 10.5 ...
##  $ INCOME            : num  67349 91449 16039 54046 114986 ...
##  $ HOME_VAL          : num  0 257252 124191 306251 243925 ...
##  $ TRAVTIME          : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ BLUEBOOK          : num  14230 14940 4010 15440 18000 ...
##  $ TIF               : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ OLDCLAIM          : num  4461 0 38690 0 19217 ...
##  $ CLM_FREQ          : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ MVR_PTS           : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE           : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ CAR_USE_Commercial: num  0 1 0 0 0 1 0 1 0 1 ...
##  $ MSTATUS_Yes       : num  0 0 1 1 1 0 1 1 0 0 ...
##  $ PARENT1_Yes       : num  0 0 0 0 0 1 0 0 0 0 ...
##  $ RED_CAR_yes       : num  1 1 0 1 0 0 0 1 0 0 ...
##  $ REVOKED_Yes       : num  0 0 0 0 1 0 0 1 0 0 ...
##  $ SEX_M             : num  1 1 0 1 0 0 0 1 0 1 ...
##  $ URBANICITY_Rural  : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ YOJ_MISS          : num  0 0 0 0 1 0 1 1 0 0 ...
##  $ INCOME_MISS       : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ HOME_VAL_MISS     : num  0 0 0 0 0 0 1 0 0 0 ...
```

```
##  $ CAR_AGE_MISS      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TIF_sin           : num  -1 0.841 -0.757 0.657 0.841 ...
##  $ BLUEBOOK_sin      : num  -0.988 -0.988 0.971 0.8 -0.97 ...
##  $ AGE_sin           : num  -0.305 -0.832 -0.428 0.67 -0.262 ...
##  $ CAR_TYPE_FLAG_BIN : num  1 1 0 1 0 0 0 0 0 0 ...
##  $ EDUCATION_FLAG_BIN: num  1 0 0 0 1 1 0 1 1 1 ...
##  $ JOB_TYPE_FLAG_BIN : num  0 1 1 1 0 1 1 1 1 0 ...
##  $ INCOME_FLAG_BIN   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ YOJ_FLAG_BIN      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HOME_VAL_FLAG_BIN : num  1 0 0 0 0 1 0 0 1 1 ...
##  $ OLDCLAIM_FLAG_BIN : num  0 1 0 1 0 1 1 0 1 1 ...
##  $ CLM_FREQ_FLAG_BIN : num  0 1 0 1 0 1 1 0 1 1 ...
##  $ MVR_PTS_FLAG_BIN  : num  0 1 0 1 0 1 1 0 1 0 ...
##  $ CAR_AGE_FLAG_BIN  : num  0 1 0 0 0 0 1 0 1 0 ...
##  $ TRAVTIME_FLAG_BIN : num  1 0 1 0 0 0 0 0 0 0 ...
```

3.3 Build Models

In this section, we will create 3 models. Aside from using original and transformed data, we will also using different methods and functions such as Linear Discriminant Analysis, step function, and logit function to enhance our models. newline newline Below is our model definition:
-Model 1- This model will be created using all the variables in train data set with logit function GLM.
-Model 2: This model step function will be used to enhance the model 1.
-Model 3- This model will be created using calssification and regression tree.

3.3.1 Prepare TRAIN and VALID datasets

However, prior to that, we hold out a subset of data as a validation dataset to check model performance. This will be useful when we select a model.

3.3.2 Model 1 and enahncement of Model 1 with step function (Model 2)

In this model, we will be using all the given variables in train data set. We will create model using logit function. We will then step thru the model to remove unnecessary variables and generate the refined model. We will highlight the summary of the refined model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = na.omit(DS_TARGET_FLAG_TRAIN))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4903  -0.7193  -0.4098   0.6561   3.1494
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.988e-01  3.403e-01  -1.172 0.241181
## KIDSDRIV            3.979e-01  6.763e-02   5.883 4.03e-09 ***
## AGE                -4.759e-03  4.458e-03  -1.067 0.285793
## HOMEKIDS            1.818e-02  4.169e-02   0.436 0.662766
## YOJ                 1.598e-02  1.378e-02   1.159 0.246305
## INCOME             -3.058e-06  1.260e-06  -2.427 0.015231 *
```

```
## HOME_VAL           -8.986e-07  6.553e-07  -1.371 0.170277
## TRAVTIME            1.334e-02  2.786e-03   4.786 1.70e-06 ***
## BLUEBOOK           -1.680e-05  4.762e-06  -3.527 0.000420 ***
## TIF                -5.153e-02  9.256e-03  -5.567 2.59e-08 ***
## OLDCLAIM           -2.124e-05  4.735e-06  -4.487 7.23e-06 ***
## CLM_FREQ            7.001e-02  4.953e-02   1.413 0.157567
## MVR_PTS             1.042e-01  2.119e-02   4.918 8.73e-07 ***
## CAR_AGE             5.177e-03  1.077e-02   0.481 0.630837
## CAR_USE_Commercial  7.512e-01  7.643e-02   9.828  < 2e-16 ***
## MSTATUS_Yes        -5.337e-01  9.590e-02  -5.565 2.62e-08 ***
## PARENT1_Yes         3.787e-01  1.219e-01   3.106 0.001898 **
## RED_CAR_yes        -1.602e-02  9.656e-02  -0.166 0.868237
## REVOKED_Yes         1.052e+00  1.032e-01  10.194  < 2e-16 ***
## SEX_M              -7.591e-03  9.676e-02  -0.078 0.937467
## URBANICITY_Rural   -2.313e+00  1.254e-01 -18.453  < 2e-16 ***
## YOJ_MISS           -9.088e-02  1.503e-01  -0.605 0.545393
## INCOME_MISS        -8.443e-02  1.491e-01  -0.566 0.571275
## HOME_VAL_MISS      -1.280e-02  1.414e-01  -0.091 0.927878
## CAR_AGE_MISS        2.667e-01  1.351e-01   1.975 0.048291 *
## TIF_sin             2.893e-02  5.475e-02   0.528 0.597165
## BLUEBOOK_sin       -2.722e-02  4.562e-02  -0.597 0.550670
## AGE_sin             1.864e-02  4.599e-02   0.405 0.685238
## CAR_TYPE_FLAG_BIN  -5.584e-01  8.259e-02  -6.760 1.38e-11 ***
## EDUCATION_FLAG_BIN -3.764e-01  9.592e-02  -3.923 8.73e-05 ***
## JOB_TYPE_FLAG_BIN   3.225e-01  9.760e-02   3.304 0.000953 ***
## INCOME_FLAG_BIN     4.796e-01  3.508e-01   1.367 0.171569
## YOJ_FLAG_BIN        8.043e-02  3.797e-01   0.212 0.832256
## HOME_VAL_FLAG_BIN   1.107e-02  1.552e-01   0.071 0.943148
## OLDCLAIM_FLAG_BIN  -4.899e-01  1.371e-01  -3.572 0.000354 ***
## CLM_FREQ_FLAG_BIN          NA         NA      NA       NA
## MVR_PTS_FLAG_BIN    2.785e-02  9.490e-02   0.293 0.769170
## CAR_AGE_FLAG_BIN    8.810e-02  1.170e-01   0.753 0.451457
## TRAVTIME_FLAG_BIN  -9.989e-02  1.104e-01  -0.905 0.365589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7551.7  on 6524  degrees of freedom
## Residual deviance: 5886.1  on 6487  degrees of freedom
## AIC: 5962.1
##
## Number of Fisher Scoring iterations: 5
```

Interpretation for the TF_Model1 and TF_Model1_ref TF_Model1:

From model 1 summary we can find following important points-

(i) Variable URBANICITY_Rural has most significant association with lowest p value. negative value of log odd function indicates that chances of accidents ae higher in Urbancity areas compare to rural area.

(ii) For MSTATUS_Yes variable log odd is negative which indicates married people tend to drive slowly and have less number of accidents.

(iii) Sex variable has no significant association which means driving patters does not depend on men and women.

(iv) variable REVOKED_Yes has strong association which indicates if person's license has been revoked in last 7 years then chane of end up in accidents are much higher with log odds value of 0.809090.

(v) If person has a claim in last 5 years then chances of more claims are higher. Variable OLDCLAIM_FLAG_BI indicates that with negative log odds value(1 is here no claim -0.559409).

(vi) AIC value of the model is AIC: 6078.7 and number of iteration was 5.

TF_Model1_ref:

Model 1 has AIC value 6078.7 and enhaned model TF_Model1_ref has AIC value 6064.9 slightly better compared to model1. We will look into more details on model1_ref below-

(i) Based on the outcome from model_ref, it can be seen that following variables KIDSDRIV,PARENT1_Yes, MSTATUS_Yes,CAR_USE_Commercial, REVOKED_Yes, TIF_sin , CAR_TYPE_FLAG_BIN , EDUCATION_FLAG_BIN, JOB_TYPE_FLAG_BIN, INCOME_FLAG_BIN, HOME_VAL_FLAG_BIN,OLDCLAIM_FLAG_BIN,MVR_PTS_FLAG_BIN, TRAVTIME_FLAG_BIN , URBANICITY_Rural are only statistically significant. Most of the variables are having similar association as above model 1.

(ii)As for the statistically significant variables, URBANICITY_Rural has the lowest p-value suggesting a strong association of the URBANICITY_Rural to the target variable.Implication is also same negative value indicate lower chances of accidents in rural areas.

(iii) One interesting outcome is when childrens are dirivng your car then more chances of accidents with log odd value of 0.41327 for variable KIDSDRIV.

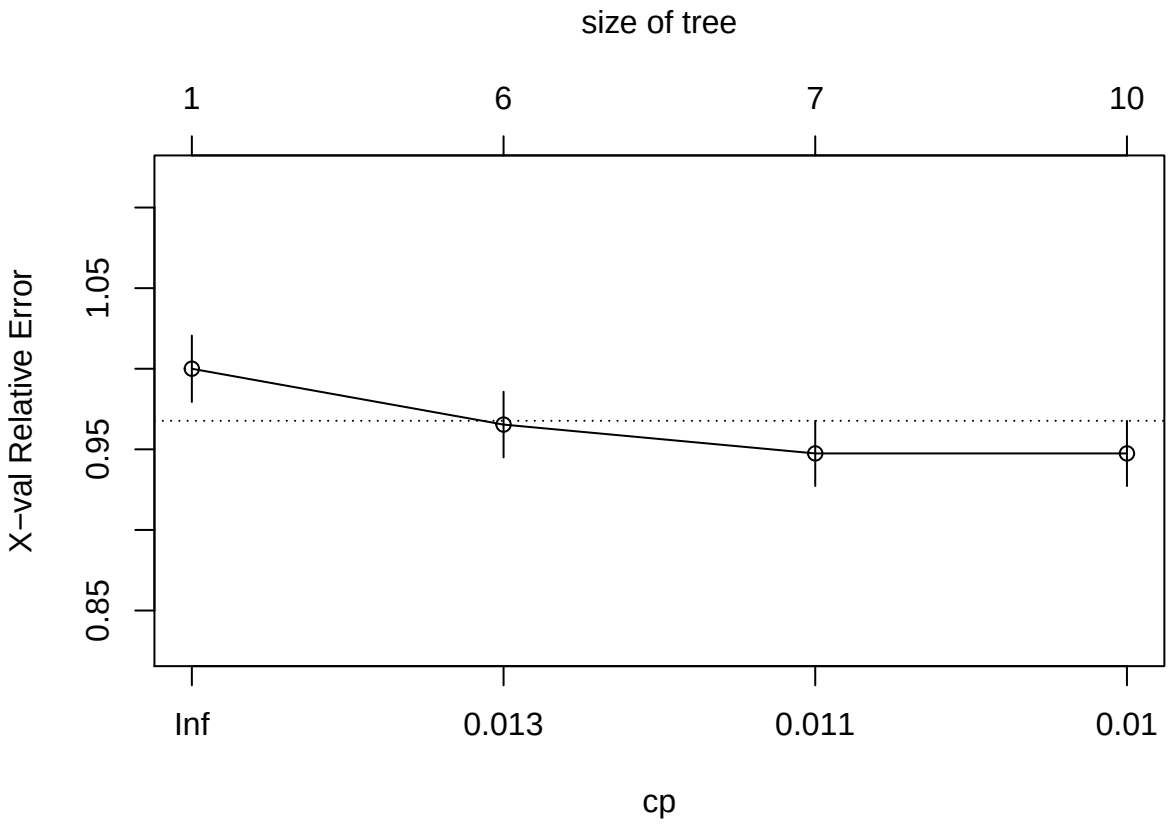
(iv) For variable CAR_TYPE_FLAG_BIN there is high negative corelation is there with log odds value of -0.65867 that means Mnivan and Panel truck has higher chance of getting into an accident.

(v) Variable EDUCATION_FLAG_BIN has negative log odds value of -0.46755 indicaating that people with higher education above high school has less chance of an accident comared to the other group.

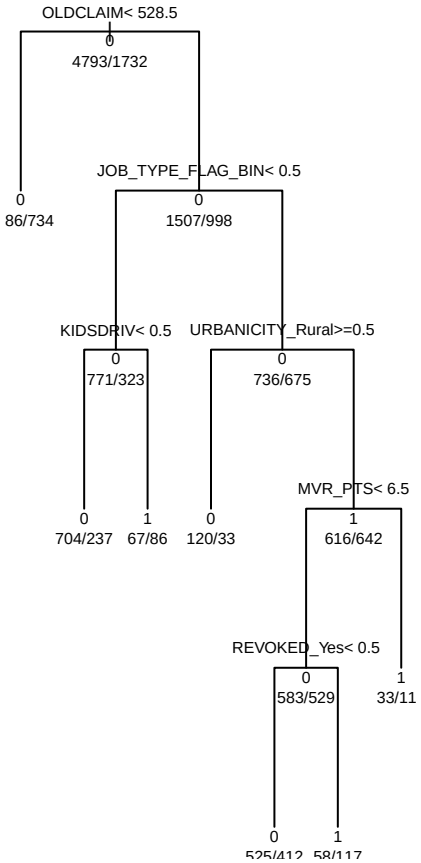
(vi) No. of iterations are 5 before lowest value of AIC was derived for this model.

3.3.3 Model 3

In this model, we will be using original variables; however we use the CART (Classification and Regression Trees) algorithm to train the model. We will then Prune the tree and have a look at the summary of this pruned model.


size of tree
1
6
7
10
X−val Relative Error
0.85
0.95
1.05
Inf
0.013
0.011
0.01
cp

*Pruned Classification Tree for TARGET_FLAG



Interpretation for Model 3

Following analysis can be drawn from this model:

(i)The following variables have been used for classification - OLD_CLAIM, JOB_TYPE_FLAG_BIN, URBANICITY_Rural, KIDSDRIV, MVR_PTS, REVOKED_Yes.

(ii)lowest Cp value and Xerror occured on split 7.

(iii) OLDCLAIM_FLAG_BIN is the first variable used to split the classification based on its value 0 and 1. When there is claim (1 in abobe variable) branch is further split to other branches by variable JOB_TYPE_FLAG_BIN (based on value 0 and 1). Based on value of JOB_TYPE_FLAG_BIN (0 and 1) there is two different routes in calssification. one Split (774/323) is based on variable KIDSDRIV and the other one (738/675) is based on URBANICITY_Rural variable. Using the above variable total 7 splits have been performed for classification.

3.4 Model Evaluation Using VALID Data

Lets go ahead and apply the above models to the VALID dataset that we had held out. Below is the table of predictions for each of the models:

3.4.1 Evaluation of Model 1

Table 8: Model 1 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.7947304	0.2052696	0.447619	0.6460481	0.8269948	0.3926341	0.8079817

Model 1 has good accuracy value close to 78.3%. sensitivity value is lower than the specificity value.

3.4.2 Evaluation of Model 2

Table 9: Model 2 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
2	0.7830882	0.2169118	0.2690476	0.70625	0.7914402	0.2537439	0.8079817

Model 2 has good accuracy value close to 77.3% and very close to model1. sensitivity value is lower than the specificity value.

3.4.3 Evaluation of Model 3

Table 10: Model 3 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.7947304	0.2052696	0.4476190	0.6460481	0.8269948	0.3926341	0.8079817
2	0.7830882	0.2169118	0.2690476	0.7062500	0.7914402	0.2537439	0.8079817
3	0.7561275	0.2438725	0.1666667	0.5932203	0.7688243	0.1451789	0.6733272

This model has accuracy value of 75.4%. AUC for this model is 67.4 % and less compared to the other two models.

3.5 Final Logistic Model Selection Summary

Following is the comparison of various metrics for above 3 models

Table 11: Model Performance Metrics Comparison

Model_No	Accuracy	Error_Rate	AUC	Precision	sensitivity	specificity	F1_Score
1	0.7947304	0.2052696	0.8079817	0.4476190	0.6460481	0.8269948	0.3926341
2	0.7830882	0.2169118	0.8079817	0.2690476	0.7062500	0.7914402	0.2537439
3	0.7561275	0.2438725	0.6733272	0.1666667	0.5932203	0.7688243	0.1451789

From the comparison table, we see that Model 1 is quite superior from the accuracy and AUC perspective.

The AUC provides the best score on probability of correctly identifying the patterns at various cut off values. The Accuracy, on the other hand, is calculated as specific cut off value. For this assignment we will go with

cut off value of 0.5 and choose the Model 1 based on Accuracy value for further prediction on evaluation data set.

3.5.1 Detailed Inference for Final Model

The following analysis will be carried out on the final model:

(i) Relevant variables in the model
(ii) Estimate confidence interval for coefficient
(iii) odds ratios and 95% CI
(iv) AUC curve
(v) Distribution of prediction

3.5.2 Most important variables in the model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = na.omit(DS_TARGET_FLAG_TRAIN))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4903  -0.7193  -0.4098   0.6561   3.1494
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.988e-01  3.403e-01  -1.172 0.241181
## KIDSDRIV            3.979e-01  6.763e-02   5.883 4.03e-09 ***
## AGE                -4.759e-03  4.458e-03  -1.067 0.285793
## HOMEKIDS            1.818e-02  4.169e-02   0.436 0.662766
## YOJ                 1.598e-02  1.378e-02   1.159 0.246305
## INCOME             -3.058e-06  1.260e-06  -2.427 0.015231 *
## HOME_VAL           -8.986e-07  6.553e-07  -1.371 0.170277
## TRAVTIME            1.334e-02  2.786e-03   4.786 1.70e-06 ***
## BLUEBOOK           -1.680e-05  4.762e-06  -3.527 0.000420 ***
## TIF                -5.153e-02  9.256e-03  -5.567 2.59e-08 ***
## OLDCLAIM           -2.124e-05  4.735e-06  -4.487 7.23e-06 ***
## CLM_FREQ            7.001e-02  4.953e-02   1.413 0.157567
## MVR_PTS             1.042e-01  2.119e-02   4.918 8.73e-07 ***
## CAR_AGE             5.177e-03  1.077e-02   0.481 0.630837
## CAR_USE_Commercial  7.512e-01  7.643e-02   9.828  < 2e-16 ***
## MSTATUS_Yes        -5.337e-01  9.590e-02  -5.565 2.62e-08 ***
## PARENT1_Yes         3.787e-01  1.219e-01   3.106 0.001898 **
## RED_CAR_yes        -1.602e-02  9.656e-02  -0.166 0.868237
## REVOKED_Yes         1.052e+00  1.032e-01  10.194  < 2e-16 ***
## SEX_M              -7.591e-03  9.676e-02  -0.078 0.937467
## URBANICITY_Rural   -2.313e+00  1.254e-01 -18.453  < 2e-16 ***
## YOJ_MISS           -9.088e-02  1.503e-01  -0.605 0.545393
## INCOME_MISS        -8.443e-02  1.491e-01  -0.566 0.571275
## HOME_VAL_MISS      -1.280e-02  1.414e-01  -0.091 0.927878
## CAR_AGE_MISS        2.667e-01  1.351e-01   1.975 0.048291 *
## TIF_sin             2.893e-02  5.475e-02   0.528 0.597165
## BLUEBOOK_sin       -2.722e-02  4.562e-02  -0.597 0.550670
```

```
## AGE_sin             1.864e-02  4.599e-02   0.405 0.685238
## CAR_TYPE_FLAG_BIN  -5.584e-01  8.259e-02  -6.760 1.38e-11 ***
## EDUCATION_FLAG_BIN -3.764e-01  9.592e-02  -3.923 8.73e-05 ***
## JOB_TYPE_FLAG_BIN   3.225e-01  9.760e-02   3.304 0.000953 ***
## INCOME_FLAG_BIN     4.796e-01  3.508e-01   1.367 0.171569
## YOJ_FLAG_BIN        8.043e-02  3.797e-01   0.212 0.832256
## HOME_VAL_FLAG_BIN   1.107e-02  1.552e-01   0.071 0.943148
## OLDCLAIM_FLAG_BIN  -4.899e-01  1.371e-01  -3.572 0.000354 ***
## CLM_FREQ_FLAG_BIN          NA         NA      NA       NA
## MVR_PTS_FLAG_BIN    2.785e-02  9.490e-02   0.293 0.769170
## CAR_AGE_FLAG_BIN    8.810e-02  1.170e-01   0.753 0.451457
## TRAVTIME_FLAG_BIN  -9.989e-02  1.104e-01  -0.905 0.365589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7551.7  on 6524  degrees of freedom
## Residual deviance: 5886.1  on 6487  degrees of freedom
## AIC: 5962.1
##
## Number of Fisher Scoring iterations: 5
```

Following are the most relevant variables for the model:
CAR_USE_Commercial, REVOKED_Yes, URBANICITY_Rural, CAR_TYPE_FLAG_BIN,TRAVTIME, OLDCLAIM_FLAG_BIN, MVR_PTS, TIF_SIN, KIDSDRIV, PARENT1_Yes, EDUCATION_FLAG_BIN, MSTATUS_Yes, BLUEBOOK, OLDCLAIM, JOB_TYPE_FLAG_BIN, HOME_VAL, INCOME_FLAG_BIN.

We can write the equation of the Model 1 as:

$$log(y) = -0.4015 + 0.3431 * KIDSDRIV - 0.000001027 * HOME_VAL + 0.01557 * TRAVTIME - 0.00001858 * BLUEBOOK - 0.05103 * TIF - 0.00001873 * OLDCLAIM + 0.1043 * MVR_PTS + 0.7632 * CAR_USE_Commercial - 0.4066 * MSTATUS_Yes + 0.5246 * PARENT1_Yes + 1.025 * REVOKED_Yes - 2.221 * URBANICITY_Rural - 0.5662 * CAR_TYPE_FLAG_BIN - 0.3996 * EDUCATION_FLAG_BIN + 0.3584 * JOB_TYPE_FLAG_BIN + 0.311 * INCOME_FLAG_BIN - 0.627 * OLDCLAIM_FLAG_BIN$$

3.5.3 Analysis of odds ratios of variables 95% CI

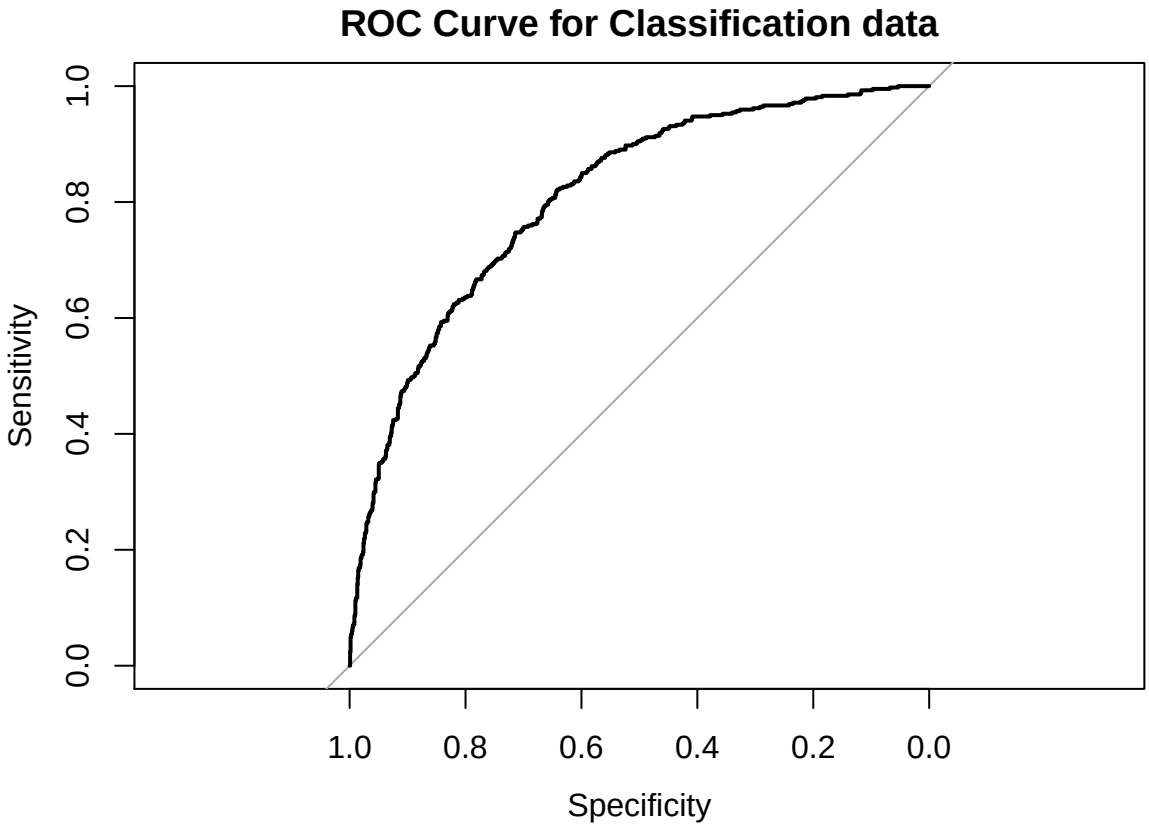
```
##                          OR      2.5 %    97.5 %
## (Intercept)       0.67113058 0.34449416 1.3074714
## KIDSDRIV          1.48862259 1.30382382 1.6996140
## AGE               0.99525237 0.98659334 1.0039874
## HOMEKIDS          1.01834688 0.93844701 1.1050495
## YOJ               1.01610813 0.98902600 1.0439318
## INCOME            0.99999694 0.99999447 0.9999994
## HOME_VAL          0.99999910 0.99999782 1.0000004
## TRAVTIME          1.01342437 1.00790509 1.0189739
## BLUEBOOK          0.99998320 0.99997387 0.9999925
## TIF               0.94977489 0.93269993 0.9671624
## OLDCLAIM          0.99997876 0.99996948 0.9999880
## CLM_FREQ          1.07251612 0.97328473 1.1818647
## MVR_PTS           1.10986959 1.06470918 1.1569455
## CAR_AGE           1.00519015 0.98418928 1.0266391
```

```
## CAR_USE_Commercial 2.11945792 1.82459644 2.4619701
## MSTATUS_Yes        0.58643036 0.48593986 0.7077019
## PARENT1_Yes        1.46035120 1.14992968 1.8545705
## RED_CAR_yes        0.98410869 0.81442639 1.1891436
## REVOKED_Yes        2.86451319 2.33977796 3.5069293
## SEX_M              0.99243744 0.82099263 1.1996844
## URBANICITY_Rural   0.09894387 0.07738989 0.1265009
## YOJ_MISS           0.91313153 0.68015614 1.2259085
## INCOME_MISS        0.91903450 0.68611221 1.2310296
## HOME_VAL_MISS      0.98728640 0.74837344 1.3024707
## CAR_AGE_MISS       1.30565907 1.00200553 1.7013335
## TIF_sin            1.02935564 0.92462226 1.1459523
## BLUEBOOK_sin       0.97314416 0.88991133 1.0641617
## AGE_sin            1.01881446 0.93100276 1.1149085
## CAR_TYPE_FLAG_BIN  0.57214143 0.48662962 0.6726796
## EDUCATION_FLAG_BIN 0.68635498 0.56871853 0.8283239
## JOB_TYPE_FLAG_BIN  1.38053840 1.14018363 1.6715608
## INCOME_FLAG_BIN    1.61542333 0.81225230 3.2127856
## YOJ_FLAG_BIN       1.08375386 0.51487406 2.2811839
## HOME_VAL_FLAG_BIN  1.01113208 0.74588194 1.3707103
## OLDCLAIM_FLAG_BIN  0.61268865 0.46828247 0.8016259
## CLM_FREQ_FLAG_BIN          NA         NA        NA
## MVR_PTS_FLAG_BIN   1.02824027 0.85372369 1.2384312
## CAR_AGE_FLAG_BIN   1.09209197 0.86830795 1.3735506
## TRAVTIME_FLAG_BIN  0.90493543 0.72885274 1.1235577
```

The following points can be made for the important variables in the model:

In keeping all other variables same, the odds of an accident increases as follow: 1.8449962 for per unit change in CAR_USE_Commercial, 2.2458626 per unit change in REVOKED_Yes, 0.1104633 for per unit change in URBANICITY_Rural, etc. Any value which is less than 1, it means that there is less chance of an event with the per unit increase of the variable.

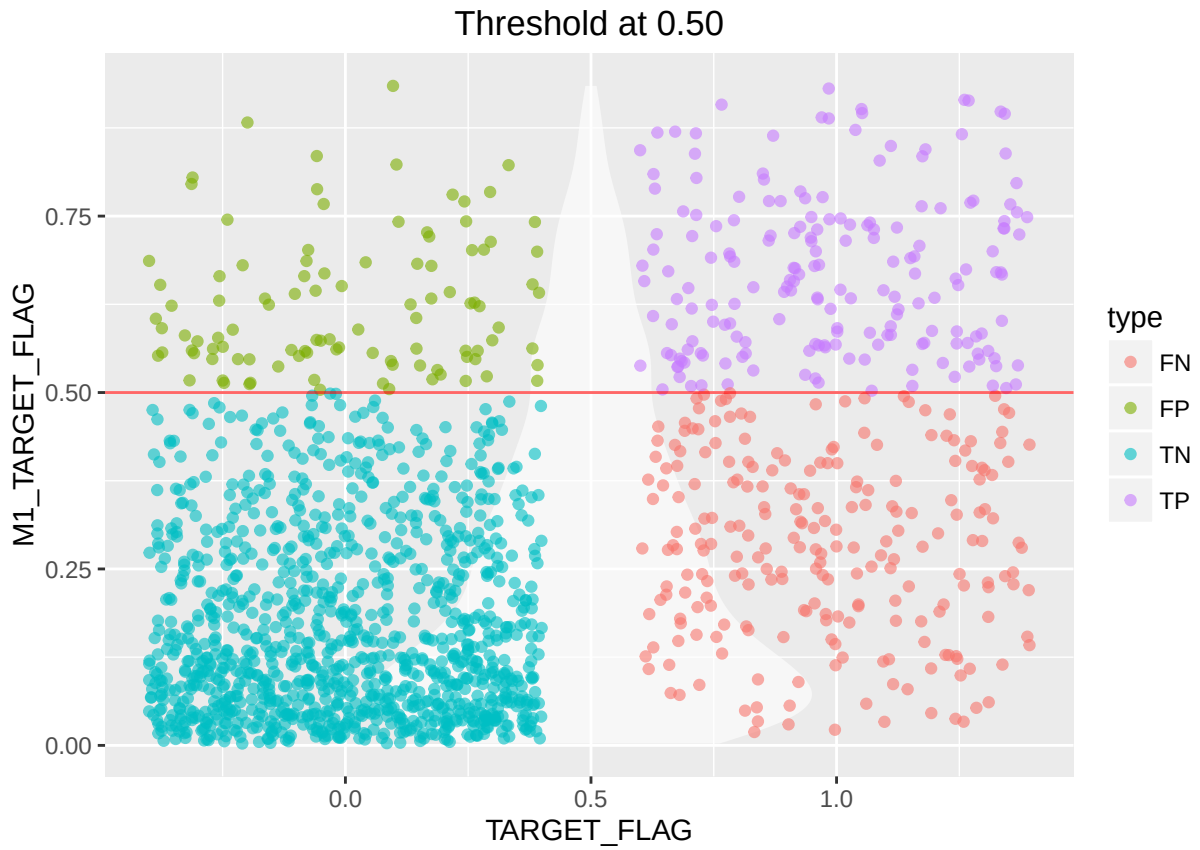
3.5.4 ROC curve for the selected model




```
##
## Call:
## roc.formula(formula = DS_TARGET_FLAG_VALID$TARGET_FLAG ~ DS_TARGET_FLAG_VALID$M1_TARGET_FLAG,     da
##
## Data: DS_TARGET_FLAG_VALID$M1_TARGET_FLAG in 1212 controls (DS_TARGET_FLAG_VALID$TARGET_FLAG 0) < 42
## Area under the curve: 0.808
```

3.5.5 Distribution of the Predictions

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```



Considering the target has value 1 (accident occurs) and 0 when no accident, then the above plot illustrates the tradeoff of choosing a reasonable threshold. In other words, if the threshold is increased, the number of false positive (FP) results is lowered; while the number of false negative (FN) results increases.

4 Linear Regression for TARGET_AMT

In this section we will use Linear regression to model the TARGET_AMT. We will first start with the Data Exploration. We will be using only those records where the TARGET_FLAG is 1. This indicates that the vehicle crashed. In such a scenario, we will be modeling the cost of repair using Linear Regression. First, lets create the required data set for the "Crashed" data from the existing "clean" full data and look at the structure of the resulting dataset. We will remove from the new "crashed" dataset all those variables that were created specifically for predicting TARGET_FLAG. We will be creating these variables separately for predicting TARGET_AMT.

We notice that the dependent variable here is TARGET_AMT. Apart from the dependent variables, we have 49 independent or predictor variables.

Also, since we created this dataset from the "Clean" full dataset, we already have taken care of the missing values.

However, we may need to look into the outliers and correlations again since we have a new target variable to correlate against.

4.1 Data Summary and Correlation Analysis

4.1.1 Data Summary

In this section, we will create summary data to better understand the relationship each of the variables have with our dependent variables using correlation, central tendency, and dispersion as shown below:

4.1.2 Correlations

Now we will produce the correlation table between the independent variables and the dependent variable - TARGET_AMT

Table 12: Correlation between TARGET_AMT and predictor variables

	Correlation_TARGET_AMT
TARGET_AMT	1.0000000
BLUEBOOK	0.1181297
CAR_TYPE_Panel.Truck	0.0682806
SEX_M	0.0513430
CAR_TYPE_Van	0.0499290
CAR_USE_Commercial	0.0496142
INCOME	0.0454604
JOB_Professional	0.0406747
JOB_Unknown	0.0402613
MVR_PTS	0.0396710
YOJ	0.0343807
EDUCATION_PhD	0.0294767
HOME_VAL	0.0291859
AGE	0.0279828
RED_CAR_yes	0.0271768
PARENT1_Yes	0.0238302
JOB_Blue.Collar	0.0155259
EDUCATION_Masters	0.0143267
EDUCATION_Bachelors	0.0136662
JOB_Lawyer	0.0102382
TRAVTIME	0.0053657
CLM_FREQ	0.0023251
INDEX	0.0010044
HOMEKIDS	0.0002698
KIDSDRIV	-0.0000869
URBANICITY_Rural	-0.0048888
OLDCLAIM	-0.0049723
CAR_TYPE_Minivan	-0.0058234
TIF	-0.0060620
JOB_Doctor	-0.0122018
CAR_AGE	-0.0135348
JOB_Clerical	-0.0151891
CAR_TYPE_Sports.Car	-0.0152654
CAR_TYPE_Pickup	-0.0174060
JOB_Manager	-0.0256129
JOB_Home.Maker	-0.0293974
JOB_Student	-0.0331511

	Correlation_TARGET_AMT
MSTATUS_Yes	-0.0351848
EDUCATION_High.School	-0.0359529
REVOKED_Yes	-0.0365018
CAR_TYPE_SUV	-0.0405600

The above table suggests that none of the variables seem to have a very strong correlation with TARGET_AMT.

However, BLUEBOOK, CAR_TYPE_Panel.Truck, SEX_M, CAR_TYPE_Van, CAR_USE_Commercial, INCOME, INCOME_IMPUTE, JOB_Professional, JOB_Unknown, MVR_PTS, YOJ, YOJ_IMPUTE, HOME_VAL_IMPUTE, EDUCATION_PhD, HOME_VAL, AGE, AGE_IMPUTE, RED_CAR_yes, PARENT1_Yes, YOJ_MISS, HOME_VAL_MISS, JOB_Blue.Collar, EDUCATION_Masters, EDUCATION_Bachelors, JOB_Lawyer, TRAVTIME, CLM_FREQ, HOMEKIDS have a positive correlation.

Similarly, KIDSDRIV, TRAVTIME_FLAG_BIN, INCOME_MISS, URBANICITY_Rural, OLDCLAIM, CAR_TYPE_Minivan, TIF, CAR_AGE_MISS, JOB_Doctor, CAR_AGE, CAR_AGE_IMPUTE, JOB_Clerical, CAR_TYPE_Sports.Car, CAR_TYPE_Pickup, JOB_Manager, JOB_Home.Maker, JOB_Student, MSTATUS_Yes, EDUCATION_High.School, REVOKED_Yes, CAR_TYPE_SUV have a negative correlation.

Lets now see how values in some of the variable affects the correlation:

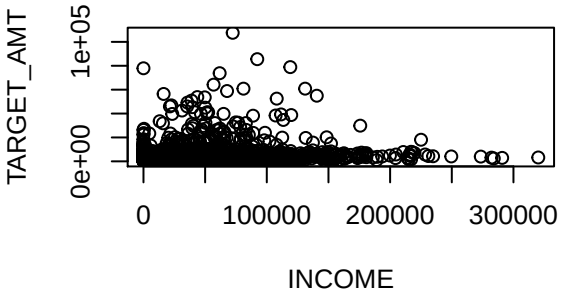
CAR_TYPE - If you drive Vans or Panel Trucks your cost of repair seems to increase as against Minivan, Pickup, Sports.Car, SUV. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation.

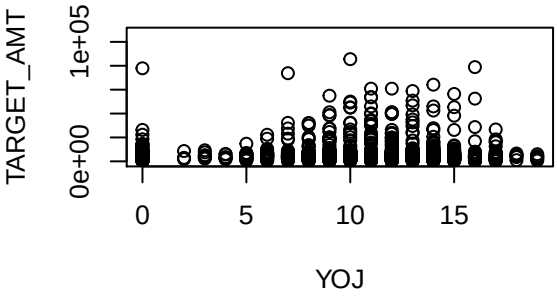
EDUCATION - If you have only a high school education then your cost of repair is less compared to a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation.

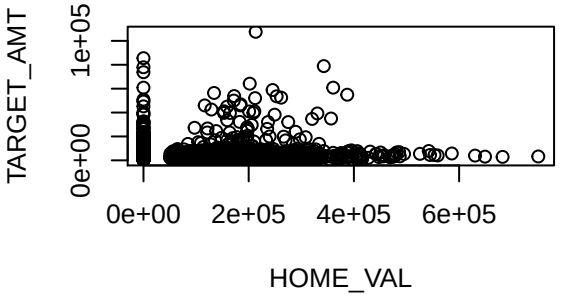
JOB - If you are a Lawyer, Professional, in a Blue Collar job or the job is unknown, you spend more on repairs as compared to a Doctor, Manager, Home Maker, Student, or Clerical job. Again binning this variable will strengthen the correlation.

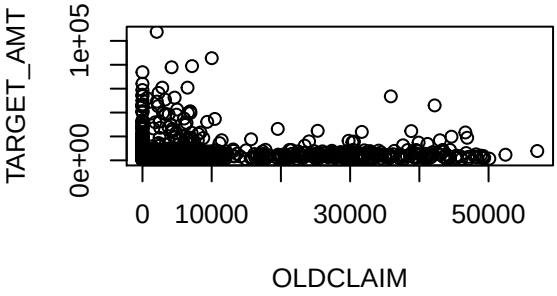
4.1.3 Binning of Variables

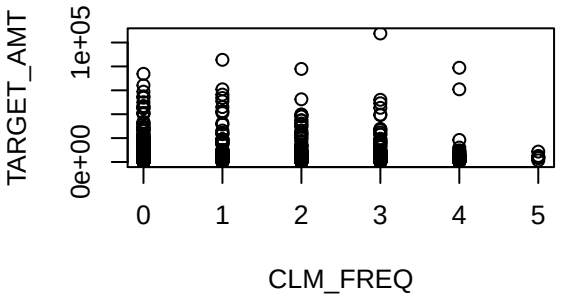
Lets have a look at the following numeric variables that have 0 as one of their values: INCOME, YOJ, HOME_VAL, OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, AGE, BLUEBOOK, TIF, TRAVTIME. The goal here is to see if we can bin these variables into zero and non-zero bin values and check the correlations. While doing that we will also see how the variables are distributed vis-a-vis TARGET_AMT.

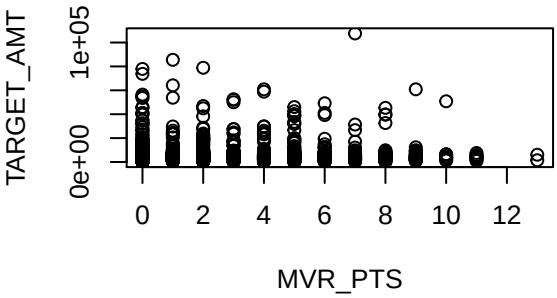

TARGET_AMT
0e+00
1e+05
0
100000
200000
300000
INCOME

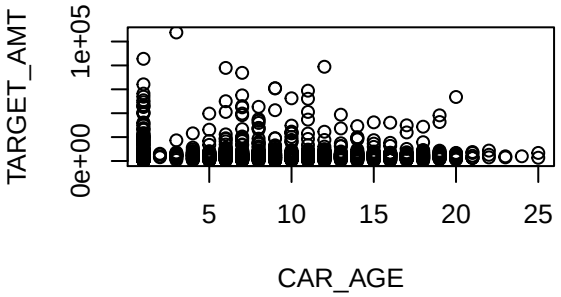

TARGET_AMT
0e+00
1e+05
0
5
10
15
YOJ

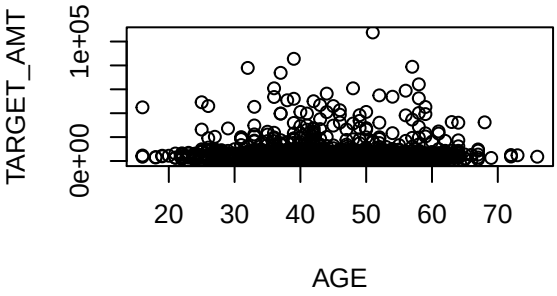

TARGET_AMT
0e+00
1e+05
0e+00
2e+05
4e+05
6e+05
HOME_VAL

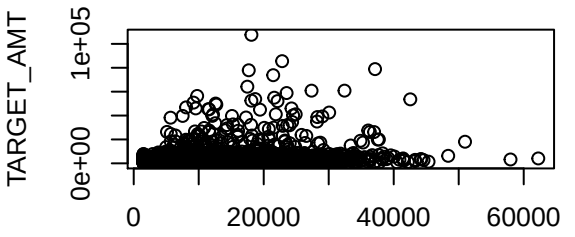

TARGET_AMT
0e+00
1e+05
0
10000
30000
50000
OLDCLAIM

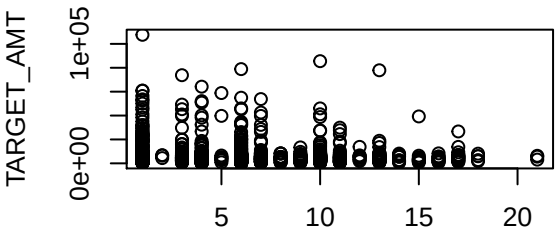

TARGET_AMT
0e+00
1e+05
0
1
2
3
4
5
CLM_FREQ

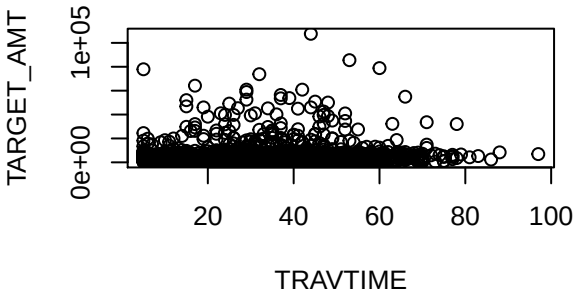

TARGET_AMT
0e+00
1e+05
0
2
4
6
8
10
12
MVR_PTS


TARGET_AMT
0e+00
1e+05
5
10
15
20
25
CAR_AGE


TARGET_AMT
0e+00
1e+05
20
30
40
50
60
70
AGE







From the outputs above, we can come to the following conclusions:

- INCOME - From the plot we can see that there is a marked difference in the chart at around 125000. We will use this value to bin this variable.
- YOJ - We can see that from 7 - 17 years, there is a visible change in the TARGET_AMT. We will use this bound to create the binned variable.
- HOME_VAL - We see from the plot 3 distinct segments - Between 0-10000, 60000-400000 and the rest. We will use these values to create 2 bins.
- OLDCLAIM- We can visualize 3 clusters in the data - 0-2000, 2000-10000, $>$ 10000, We will use these values to create 2 bins.
- CLM_FREQ - Values less than 4 seem to have a positive correlation. We will use this value for binning.
- MVR_PTS - We can see from the plot that after 2, the TARGET_AMT starts decreasing. We will use this value for binning.
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.
- AGE - There is no specific pattern that emerges in AGE. We will retain the variable as is.
- BLUEBOOK - There is no specific pattern that emerges. We will retain the variable as is.
- TIF - Looking at the plot we can conclude that this is not a good variable for binning. We will retain this variable as is.

- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

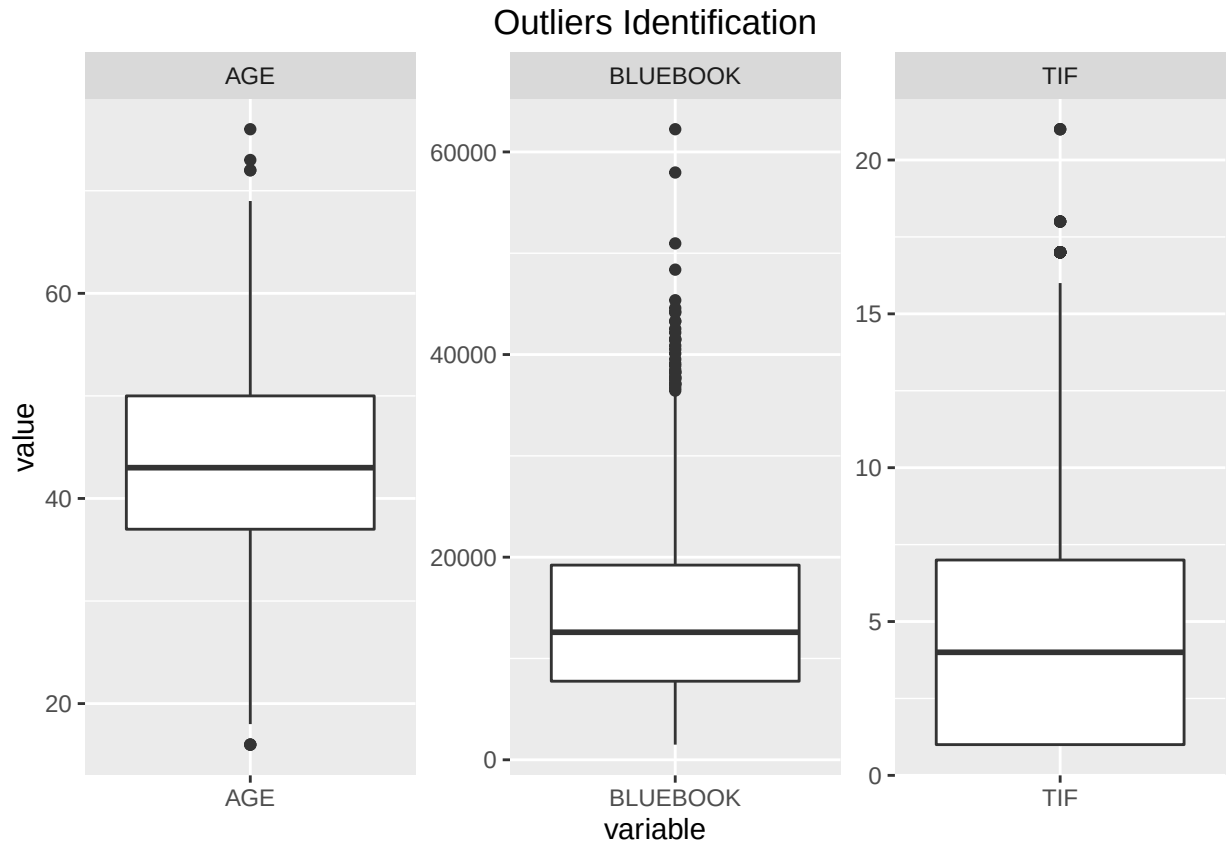
We will carry out the above transformations in the Data Preparation phase.

4.1.4 Outliers identification

In this sub-section, we will look at the boxplots and determine the outliers in variables and decide on whether to act on the outliers.

We will do the outliers only on the numeric variables: AGE, BLUEBOOK and TIF. The other variables will be binned and would not beed outlier handling.

Below are the plots:



From the "Outliers identification" plot above, we see that we have few outliers that we need to treat.

We see thatall the 3 variables need to be treated when we do the data preparation for modeling the TARGET_AMT.

4.2 Data Preparation

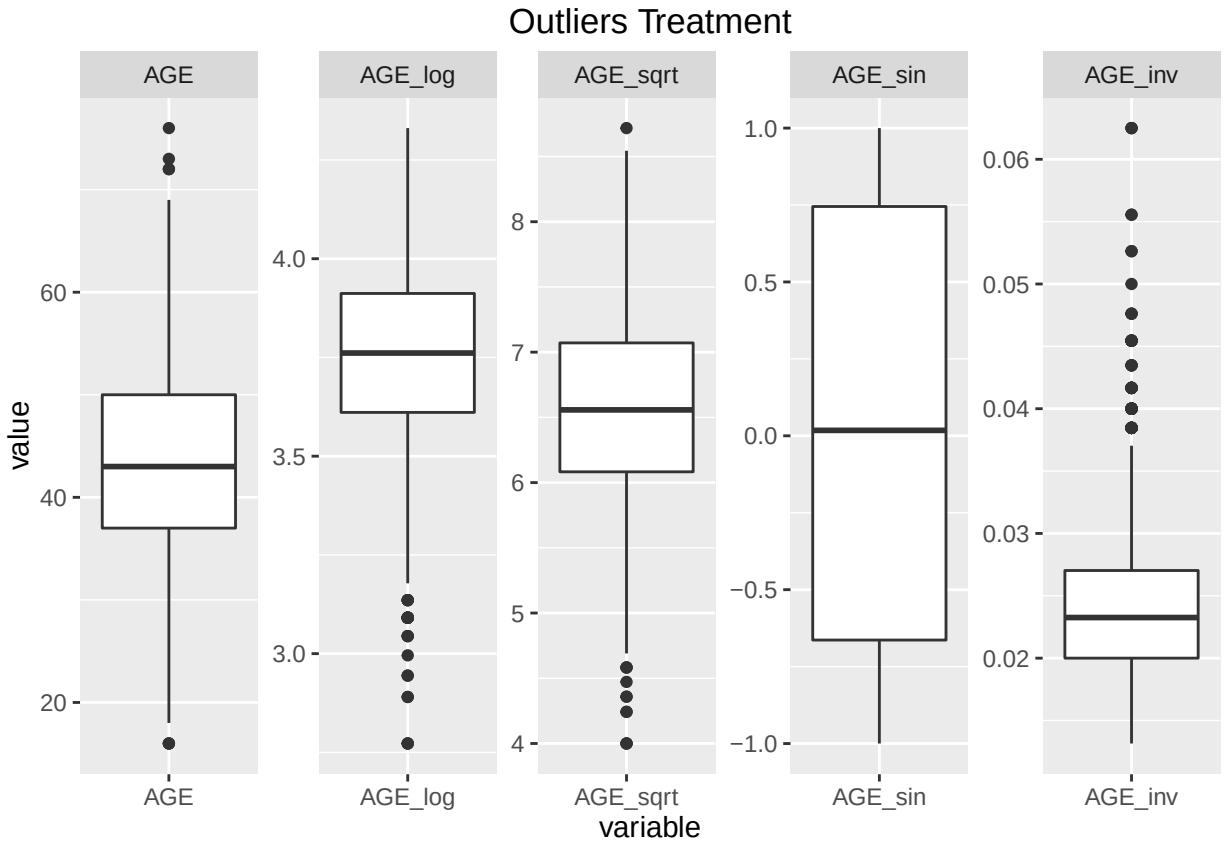
Now that we have completed the data exploration / analysis, we will be transforming the data for use in analysis and modeling.

We will be following the below steps as guidelines: - Outliers treatment - Adding New Variables

4.2.1 Outliers treatment

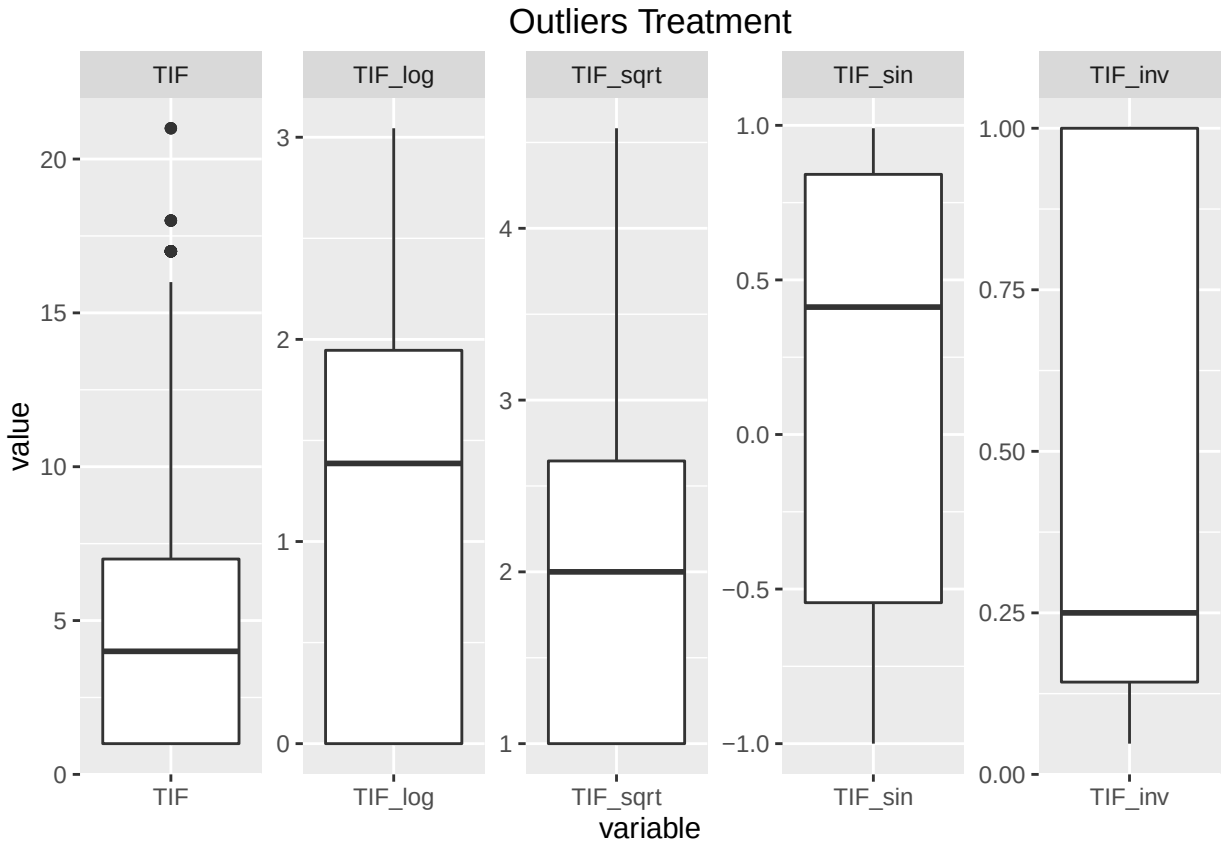
In this sub-section, we will check different transformations for each of the variables - AGE, BLUEBOOK, TIF - and create the appropriate outlier-handled / transformed variables.

Transformations for AGE



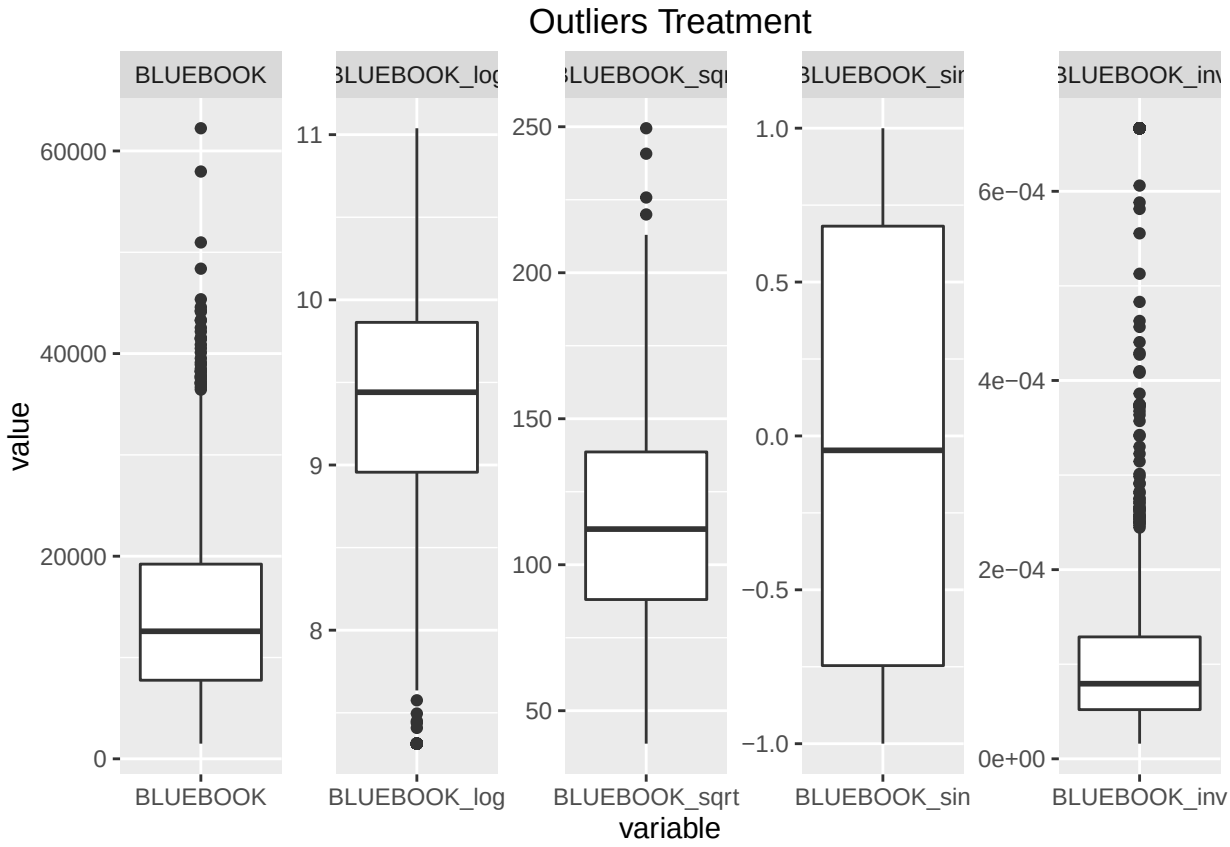
From the above charts we can see that a sin transformation works well for AGE. We will create this variable.

Transformations for TIF



From the above charts we can see that a log, sqrt, sin or an inverse transformation works well for TIF. However, a sin transformation seems to be more appropriate as it is well centered. Hence, We will create these variables.

Transformations for BLUEBOOK



From the above charts we can see that a sin transformation works well for BLUEBOOK. We will create these variables.

4.2.2 Adding New Variables

In this section, we generate some additional variables that we feel will help the correlations. The following were some of the observations we made during the data exploration phase for TARGET_AMT.

The following were some of the observations we made during the data exploration phase for TARGET_AMT

CAR_TYPE - If you drive Vans or Panel Trucks your cost of repair seems to increase as against Minivan, Pickup, Sports.Car, SUV. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation. Accordingly, we will bin these variables as below:

CAR_TYPE_AMT_BIN :

- 1 : if CAR_TYPE is Vans or Panel Trucks
- 0 : if CAR_TYPE is Pickups, Sports, SUVs or Minivans

EDUCATION - If you have only a high school education then your cost of repair is less compared to a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation. Accordingly, we will bin these variables as below:

EDUCATION_AMT_BIN :

- 1 : if EDUCATION is High School
- 0 : if EDUCATION is Bachelors, Masters or Phd

JOB - If you are a Lawyer, Professional, in a Blue Collar job or the job is unknown, you spend more on repairs as compared to a Doctor, Manager, Home Maker, Student, or Clerical job. Again binning this variable will strengthen the correlation. Accordingly, we will bin these variables as below:

JOB_TYPE_AMT_BIN :

- 1 : if JOB_TYPE is Lawyer, Professional, Unknown or in a Blue Collar
- 0 : if JOB_TYPE is Doctor, Manager, Home Maker, Student, or Clerical

INCOME - From the plot we can see that there is a marked difference in the chart at around 125000. We will use this value to bin this variable.

INCOME_AMT_BIN :

- 1 : if INCOME <= 125000
- 0 : if INCOME > 125000
- YOJ - We can see that from 7 - 17 years, there is a visible change in the TARGET_AMT. We will use this bound to create the binned variable.

YOJ_AMT_BIN :

- 1 : if YOJ >=7 and YOJ<= 17
- 0 : ELSE 0
- HOME_VAL - We see from the plot 3 distinct segments - Between 0-10000, 60000-400000 and the rest. We will use these values to create 2 bins.

HOME_VAL_AMT_0_10K_BIN :

- 1 : if HOME_VAL >=0 and HOME_VAL<= 10000
- 0 : ELSE 0

HOME_VAL_AMT_60K_400K_BIN :

- 1 : if HOME_VAL >=60000 and HOME_VAL<= 400000
- 0 : ELSE 0

OLDCLAIM- We can visualize 3 clusters in the data - 0-2000, 2000-10000, > 10000, We will use these values to create 2 bins.

OLDCLAIM_AMT_0_2K_BIN :

- 1 : if OLDCLAIM >=0 and OLDCLAIM<= 2000
- 0 : ELSE 0

OLDCLAIM_AMT_2K_10K_BIN :

- 1 : if OLDCLAIM >=2000 and OLDCLAIM<= 10000
- 0 : ELSE 0

- CLM_FREQ - Values less than 4 seem to have a positive correlation. We will use this value for binning.

CLM_FREQ_AMT_BIN :

- 1 : if CLM_FREQ < 4
- 0 : if CLM_FREQ >= 4
- MVR_PTS - We can see from the plot that after 2, the TARGET_AMT starts decreasing. We will use this value for binning.

MVR_PTS_AMT_BIN :

- 1 : if MVR_PTS <=2
- 0 : if MVR_PTS > 0
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.

CAR_AGE_AMT_BIN :

- 1 : if CAR_AGE <= 1
- 0 : if CAR_AGE > 0
- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

TRAVTIME_AMT_BIN :

- 1 : if TRAVTIME <= 20
- 0 : if TRAVTIME > 0

4.3 Build Models

Now that we have the dataset in a shape that can be modeled, we will go ahead and train the model for TARGET_AMT. We will train 2 models and select the best among these 2 models. The following will be the model specifications:

- Model1 (All Variables in Linear Dataset) - This will use the standard lm for building the model. We will use all available variables.
- Model2 - (A few selected variables) - This will use the standard lm for building the model. However, we will use only a few selected variables that seemed to have a good correlation with TARGET_AMT.

4.3.1 Model 1

In this model, we will be using the standard lm modeling technique. We will use the entire set of variables from the Linear dataset.

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = na.omit(DS_TARGET_AMT))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9677  -3249  -1387    792  75391
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.005e+03  6.472e+03   0.310  0.75674
## KIDSDRIV                   -1.683e+01  1.782e+03  -0.009  0.99247
## AGE                         1.114e+01  2.565e+01   0.434  0.66415
## HOMEKIDS                    3.102e+02  4.101e+02   0.756  0.44956
## YOJ                        -1.288e+02  1.189e+02  -1.083  0.27895
## INCOME                     -1.149e-02  1.083e-02  -1.061  0.28897
## HOME_VAL                    1.941e-03  4.389e-03   0.442  0.65832
## TRAVTIME                    5.253e+00  1.641e+01   0.320  0.74890
## BLUEBOOK                    1.528e-01  3.399e-02   4.496 7.41e-06 ***
## TIF                        -4.696e+01  1.116e+02  -0.421  0.67391
## OLDCLAIM                    1.300e-02  4.408e-02   0.295  0.76809
## CLM_FREQ                   -5.938e+02  3.286e+02  -1.807  0.07095 .
## MVR_PTS                     6.051e+02  2.594e+02   2.332  0.01980 *
## CAR_AGE                    -1.689e+02  7.621e+01  -2.217  0.02678 *
## CAR_USE_Commercial          6.193e+01  5.625e+02   0.110  0.91235
## MSTATUS_Yes                -1.815e+03  6.199e+02  -2.928  0.00346 **
## PARENT1_Yes                -7.486e+02  7.574e+02  -0.988  0.32309
## RED_CAR_yes                -1.314e+02  5.557e+02  -0.236  0.81308
## REVOKED_Yes                -1.130e+03  6.178e+02  -1.830  0.06749 .
## SEX_M                       1.924e+03  7.219e+02   2.666  0.00776 **
## URBANICITY_Rural           -2.500e+02  8.220e+02  -0.304  0.76106
## EDUCATION_Bachelors        -3.466e+03  1.323e+03  -2.620  0.00887 **
## EDUCATION_High.School      -4.125e+03  1.420e+03  -2.904  0.00373 **
## EDUCATION_Masters          -2.236e+03  1.083e+03  -2.065  0.03911 *
## JOB_Blue.Collar             8.884e+02  1.311e+03   0.677  0.49824
## JOB_Clerical               -1.907e+02  1.376e+03  -0.139  0.88981
## JOB_Doctor                 -3.153e+03  1.885e+03  -1.673  0.09460 .
## JOB_Home.Maker             -1.200e+02  1.503e+03  -0.080  0.93638
## JOB_Lawyer                 -3.472e+02  1.156e+03  -0.300  0.76396
## JOB_Manager                -9.609e+02  1.227e+03  -0.783  0.43367
## JOB_Professional            1.427e+03  1.289e+03   1.107  0.26827
## JOB_Student                 2.542e+01  1.563e+03   0.016  0.98703
## CAR_TYPE_Minivan            1.140e+02  8.588e+02   0.133  0.89438
## CAR_TYPE_Panel.Truck       -6.953e+01  9.576e+02  -0.073  0.94213
## CAR_TYPE_Pickup             5.865e+02  8.258e+02   0.710  0.47769
## CAR_TYPE_Sports.Car         2.203e+03  1.115e+03   1.975  0.04839 *
## CAR_TYPE_SUV                1.972e+03  1.047e+03   1.882  0.05998 .
## AGE_sin                    -1.392e+01  2.635e+02  -0.053  0.95788
## TIF_sin                    -2.987e+02  3.791e+02  -0.788  0.43086
```

```
## BLUEBOOK_sin               1.310e+02  2.608e+02   0.502  0.61557
## INCOME_AMT_BIN            -5.870e+02  1.278e+03  -0.459  0.64598
## YOJ_AMT_BIN                1.765e+03  1.017e+03   1.735  0.08296 .
## HOME_VAL_AMT_0_10K_BIN     2.353e+03  1.959e+03   1.201  0.22985
## HOME_VAL_AMT_60K_400K_BIN  2.907e+03  1.496e+03   1.944  0.05212 .
## OLDCLAIM_AMT_0_2K_BIN     -2.182e+03  1.453e+03  -1.501  0.13344
## OLDCLAIM_AMT_2K_10K_BIN   -4.936e+02  1.139e+03  -0.434  0.66466
## CLM_FREQ_AMT_BIN          -1.331e+03  1.183e+03  -1.125  0.26082
## MVR_PTS_AMT_BIN            1.740e+03  7.970e+02   2.183  0.02918 *
## CAR_AGE_AMT_BIN           -6.829e+02  7.036e+02  -0.971  0.33190
## TRAVTIME_AMT_BIN           7.007e+01  6.639e+02   0.106  0.91596
## KIDSDRIV_AMT_BIN_0        -9.347e+02  1.928e+03  -0.485  0.62787
## KIDSDRIV_AMT_BIN_1         2.079e+03  2.355e+03   0.883  0.37742
## HOMEKIDS_AMT_BIN_0        -5.618e+02  9.971e+02  -0.563  0.57318
## HOMEKIDS_AMT_BIN_3         8.051e+02  1.428e+03   0.564  0.57285
## YOJ_AMT_BIN_0_AND_9To14    6.615e+02  7.114e+02   0.930  0.35259
## INCOME_AMT_BIN_MISS_0     -2.171e+02  1.778e+03  -0.122  0.90283
## TIF_AMT_BIN_6             -1.557e+02  8.834e+02  -0.176  0.86015
## OLDCLAIM_AMT_BIN_MISS_0    4.083e+02  1.026e+03   0.398  0.69059
## MVR_PTS_AMT_BIN_0          4.155e+02  6.288e+02   0.661  0.50888
## MVR_PTS_AMT_BIN_5          1.819e+03  1.083e+03   1.680  0.09312 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7577 on 1642 degrees of freedom
## Multiple R-squared:  0.05812,    Adjusted R-squared:  0.02428
## F-statistic: 1.717 on 59 and 1642 DF,  p-value: 0.0006848
```

Interpretation of the Model

Based on the Model output, below are the characteristics of the refined model :

- The Residual standard error is 7577
- Multiple R-squared: 0.05812
- Adjusted R-squared: 0.02428
- F-statistic: 1.717 on 59 and 1642 DF
- p-value: < 0.0006848

4.3.2 Model 2

In this model, we will be using the standard lm modeling technique. We will use only those variables that seemed to have a good correlation with TARGET_AMT.

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = DS_SELECTED_VARS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9665  -3182  -1427    758  75206
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)               -8.686e+02  6.126e+03  -0.142  0.88725
## AGE                        1.359e+01  2.550e+01   0.533  0.59414
## BLUEBOOK                   1.533e-01  3.261e-02   4.702  2.8e-06 ***
## CAR_AGE                   -1.077e+02  7.287e+01  -1.478  0.13958
## CAR_AGE_AMT_BIN           -2.792e+02  6.877e+02  -0.406  0.68485
## CAR_USE_Commercial        -1.467e+01  5.460e+02  -0.027  0.97857
## CLM_FREQ                  -6.803e+02  2.613e+02  -2.603  0.00932 **
## CLM_FREQ_AMT_BIN          -1.586e+03  1.128e+03  -1.406  0.15978
## HOME_VAL                   1.886e-03  4.376e-03   0.431  0.66661
## HOME_VAL_AMT_0_10K_BIN     2.213e+03  1.955e+03   1.132  0.25778
## HOME_VAL_AMT_60K_400K_BIN  2.762e+03  1.492e+03   1.851  0.06430 .
## HOMEKIDS                   2.804e+02  4.087e+02   0.686  0.49275
## HOMEKIDS_AMT_BIN_0        -5.455e+02  9.930e+02  -0.549  0.58280
## HOMEKIDS_AMT_BIN_3         6.281e+02  1.424e+03   0.441  0.65922
## INCOME                    -8.304e-03  1.067e-02  -0.778  0.43654
## INCOME_AMT_BIN            -1.080e+03  1.259e+03  -0.858  0.39089
## KIDSDRIV                   8.219e+01  1.779e+03   0.046  0.96315
## KIDSDRIV_AMT_BIN_0        -8.472e+02  1.925e+03  -0.440  0.65983
## KIDSDRIV_AMT_BIN_1         2.178e+03  2.351e+03   0.926  0.35447
## MSTATUS_Yes               -1.760e+03  6.180e+02  -2.848  0.00445 **
## MVR_PTS                    6.149e+02  2.590e+02   2.374  0.01769 *
## MVR_PTS_AMT_BIN            1.798e+03  7.952e+02   2.261  0.02387 *
## MVR_PTS_AMT_BIN_0          3.765e+02  6.280e+02   0.600  0.54890
## MVR_PTS_AMT_BIN_5          1.901e+03  1.082e+03   1.757  0.07910 .
## OLDCLAIM                   6.387e-03  4.365e-02   0.146  0.88368
## OLDCLAIM_AMT_0_2K_BIN     -2.130e+03  1.385e+03  -1.538  0.12431
## OLDCLAIM_AMT_2K_10K_BIN   -6.152e+02  1.135e+03  -0.542  0.58796
## PARENT1_Yes               -6.825e+02  7.537e+02  -0.905  0.36538
## RED_CAR_yes               -1.771e+02  5.530e+02  -0.320  0.74888
## REVOKED_Yes               -1.065e+03  6.139e+02  -1.734  0.08303 .
## SEX_M                      1.898e+03  7.037e+02   2.697  0.00707 **
## TIF                        2.202e+01  8.168e+01   0.270  0.78754
## TIF_AMT_BIN_6              3.152e+02  7.370e+02   0.428  0.66892
## TRAVTIME                   6.230e+00  1.640e+01   0.380  0.70399
## TRAVTIME_AMT_BIN           5.334e+01  6.636e+02   0.080  0.93594
## URBANICITY_Rural          -1.737e+02  8.191e+02  -0.212  0.83210
## YOJ                       -1.253e+02  8.558e+01  -1.464  0.14351
## YOJ_AMT_BIN                1.847e+03  1.012e+03   1.824  0.06828 .
## YOJ_AMT_BIN_0_AND_9To14    6.231e+02  5.531e+02   1.127  0.26003
## EDUCATION_Masters         -2.763e+02  7.656e+02  -0.361  0.71822
## EDUCATION_High.School     -7.452e+02  5.639e+02  -1.322  0.18646
## JOB_Clerical              -1.080e+03  6.458e+02  -1.672  0.09478 .
## JOB_Doctor                -1.692e+03  1.709e+03  -0.990  0.32210
## JOB_Home.Maker            -6.548e+02  1.021e+03  -0.641  0.52151
## JOB_Lawyer                -2.863e+02  1.029e+03  -0.278  0.78096
## JOB_Professional           3.888e+02  7.117e+02   0.546  0.58492
## JOB_Manager               -1.396e+03  9.149e+02  -1.526  0.12711
## JOB_Student               -7.835e+02  9.216e+02  -0.850  0.39537
## CAR_TYPE_Panel.Truck      -2.730e+01  8.979e+02  -0.030  0.97575
## CAR_TYPE_Pickup            5.834e+02  5.965e+02   0.978  0.32815
## CAR_TYPE_Sports.Car        2.118e+03  8.203e+02   2.582  0.00992 **
## CAR_TYPE_SUV               1.814e+03  7.294e+02   2.487  0.01297 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7577 on 1650 degrees of freedom
##   (450 observations deleted due to missingness)
## Multiple R-squared:  0.05333,    Adjusted R-squared:  0.02407
## F-statistic: 1.823 on 51 and 1650 DF,  p-value: 0.0004045
```

Interpretation of the Model

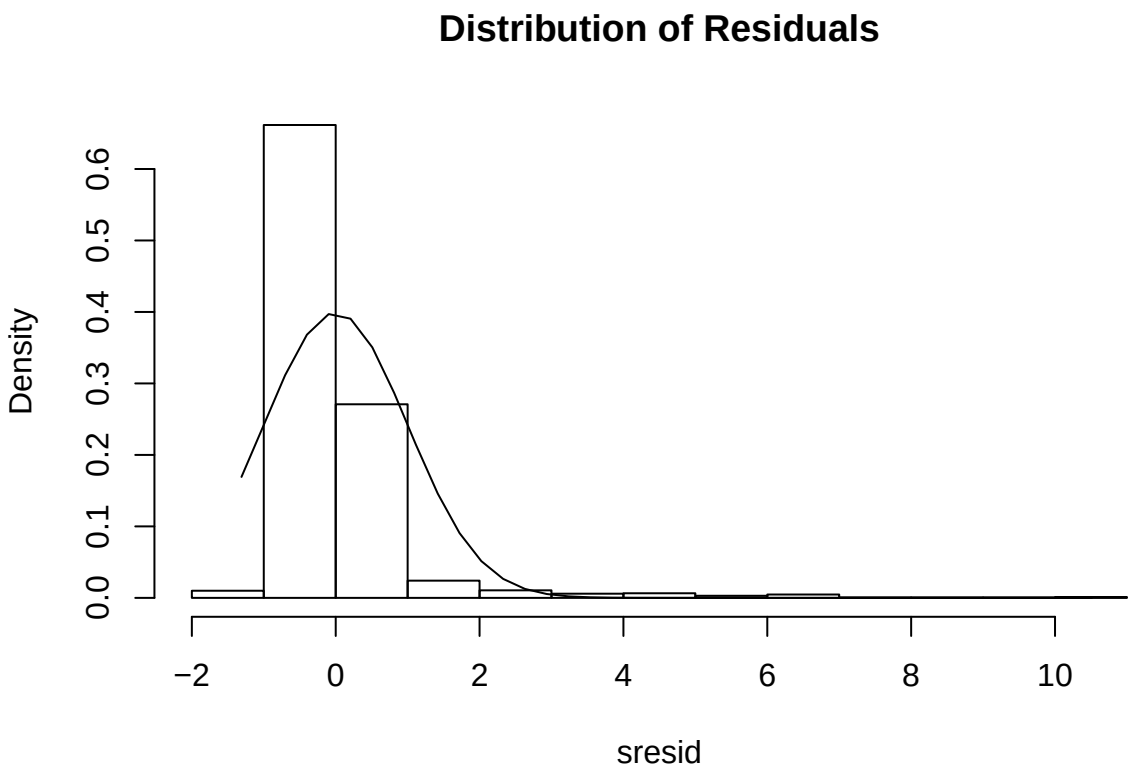
Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 7577
- Multiple R-squared: 0.05333
- Adjusted R-squared: 0.02407
- F-statistic: 1.823 on 51 and 1650 DF
- p-value: < 0.0004045

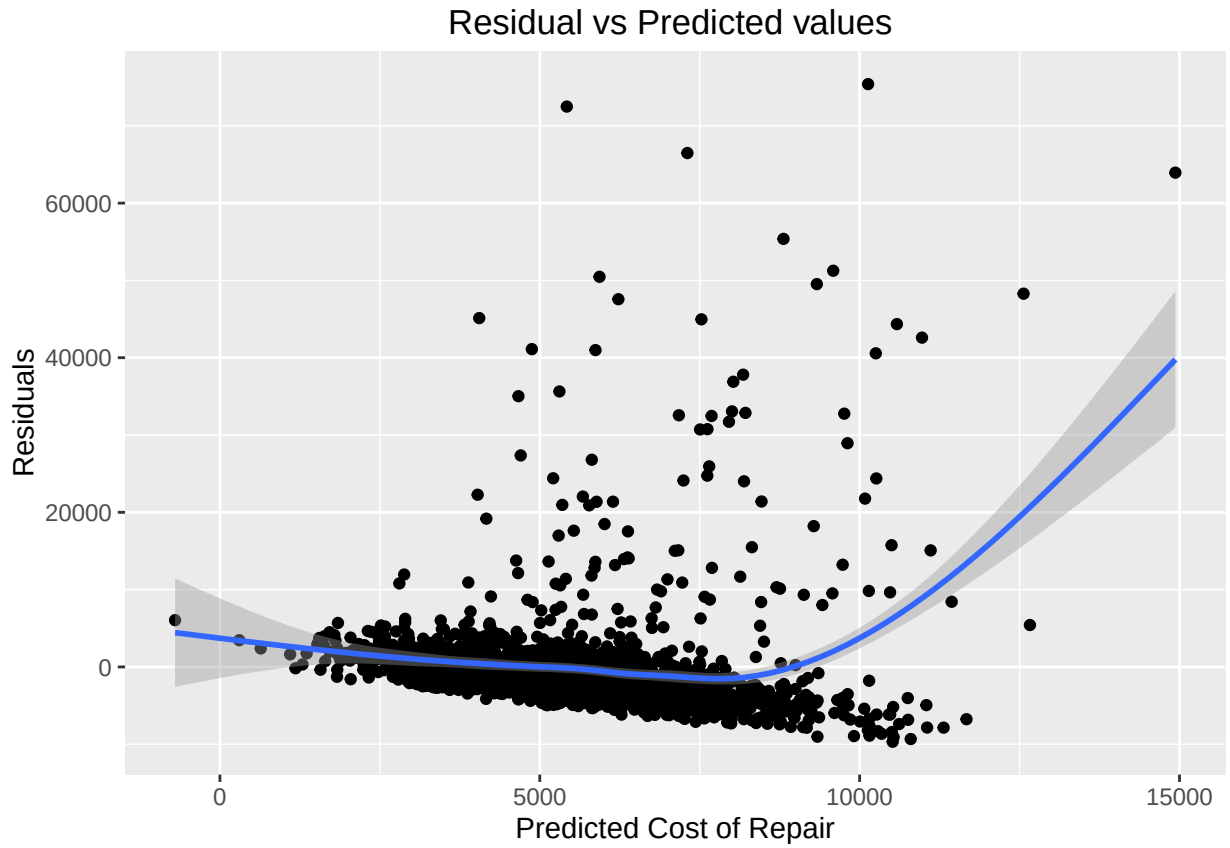
4.4 Final Linear Model Selection Summary

Based on the Models above, it is clear that Model 1 performs slightly better than Model 2. We will now use this model to validate further. We will create plots to validate the assumption of Linear Regression:

Normality check of residual values:



Based on the normality plot it appears that residual distribution is nearly normal. This indicates the mean of the difference between our predictions.



Distribution of residual values are random around base line and do not show any pattern around base line.

Evaluate homoscedasticity:

The test confirms the non-constant error variance test. It also has a p-value higher than a significance level of 0.05.

Analysis of collinearity:

```
##                KIDSDRIV                     AGE
##                5.929505                1.359778
##                HOMEKIDS                     YOJ
##                2.659453                2.901429
##                  INCOME                HOME_VAL
##                2.532339                2.860978
##                TRAVTIME                BLUEBOOK
##                1.354146                1.549806
##                     TIF                OLDCLAIM
##                2.404457                2.435889
##                CLM_FREQ                 MVR_PTS
##                2.240646                3.678998
##                 CAR_AGE       CAR_USE_Commercial
```

```
##                 2.285783                 1.531233
##              MSTATUS_Yes              PARENT1_Yes
##                 1.687018                 1.725320
##              RED_CAR_yes              REVOKED_Yes
##                 1.358885                 1.350861
##                    SEX_M          URBANICITY_Rural
##                 1.954711                 1.032607
##       EDUCATION_Bachelors     EDUCATION_High.School
##                 3.056577                 3.841265
##         EDUCATION_Masters           JOB_Blue.Collar
##                 2.083557                 3.278455
##             JOB_Clerical               JOB_Doctor
##                 2.898793                 1.282385
##            JOB_Home.Maker               JOB_Lawyer
##                 2.196223                 1.654125
##              JOB_Manager          JOB_Professional
##                 1.563833                 2.204776
##              JOB_Student          CAR_TYPE_Minivan
##                 2.740883                 1.731236
##      CAR_TYPE_Panel.Truck           CAR_TYPE_Pickup
##                 1.446472                 1.813465
##       CAR_TYPE_Sports.Car              CAR_TYPE_SUV
##                 2.134942                 2.646477
##                  AGE_sin                  TIF_sin
##                 1.019210                 1.398977
##              BLUEBOOK_sin            INCOME_AMT_BIN
##                 1.015772                 1.643573
##               YOJ_AMT_BIN    HOME_VAL_AMT_0_10K_BIN
##                 2.047651                 5.255209
## HOME_VAL_AMT_60K_400K_BIN     OLDCLAIM_AMT_0_2K_BIN
##                 4.044083                 3.953793
##   OLDCLAIM_AMT_2K_10K_BIN          CLM_FREQ_AMT_BIN
##                 3.007588                 1.279461
##           MVR_PTS_AMT_BIN          CAR_AGE_AMT_BIN
##                 2.144426                 1.784942
##          TRAVTIME_AMT_BIN        KIDSDRIV_AMT_BIN_0
##                 1.353599                 3.930334
##        KIDSDRIV_AMT_BIN_1        HOMEKIDS_AMT_BIN_0
##                 3.165772                 2.705844
##        HOMEKIDS_AMT_BIN_3   YOJ_AMT_BIN_0_AND_9To14
##                 1.299757                 1.793634
##     INCOME_AMT_BIN_MISS_0             TIF_AMT_BIN_6
##                 3.111211                 2.116778
##   OLDCLAIM_AMT_BIN_MISS_0         MVR_PTS_AMT_BIN_0
##                 2.753036                 1.608030
##         MVR_PTS_AMT_BIN_5
##                 2.083625
```

Variables have been tested with variance inflation factors (VIF). If any variable has value which is greater than 3 then the highest value variable been removed from model and model performance has been evaluated. Following are the out comes from this assessment steps-

Pass 1- Based on that variance inflation factors (VIF) following variable "KIDSDRIV" has highest value > 3 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value has changed

to 0.02487 due to removal of this variable. Hence this variable is not adding lot of value to the model and can be removed.

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - KIDSDRIV, data = na.omit(DS_TARGET_AMT))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9677  -3249  -1393    792  75391
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.968e+03  5.166e+03   0.381  0.70325
## AGE                        1.114e+01  2.564e+01   0.435  0.66392
## HOMEKIDS                   3.099e+02  4.091e+02   0.758  0.44879
## YOJ                       -1.288e+02  1.189e+02  -1.084  0.27868
## INCOME                    -1.149e-02  1.083e-02  -1.061  0.28876
## HOME_VAL                   1.941e-03  4.388e-03   0.442  0.65822
## TRAVTIME                   5.252e+00  1.640e+01   0.320  0.74884
## BLUEBOOK                   1.528e-01  3.398e-02   4.498 7.33e-06 ***
## TIF                       -4.693e+01  1.115e+02  -0.421  0.67389
## OLDCLAIM                   1.299e-02  4.406e-02   0.295  0.76811
## CLM_FREQ                  -5.938e+02  3.285e+02  -1.808  0.07083 .
## MVR_PTS                    6.051e+02  2.593e+02   2.333  0.01976 *
## CAR_AGE                   -1.689e+02  7.617e+01  -2.218  0.02670 *
## CAR_USE_Commercial         6.238e+01  5.604e+02   0.111  0.91138
## MSTATUS_Yes               -1.815e+03  6.197e+02  -2.929  0.00345 **
## PARENT1_Yes               -7.485e+02  7.570e+02  -0.989  0.32296
## RED_CAR_yes               -1.314e+02  5.555e+02  -0.237  0.81302
## REVOKED_Yes               -1.130e+03  6.175e+02  -1.831  0.06734 .
## SEX_M                      1.924e+03  7.216e+02   2.667  0.00773 **
## URBANICITY_Rural          -2.499e+02  8.217e+02  -0.304  0.76107
## EDUCATION_Bachelors       -3.466e+03  1.322e+03  -2.621  0.00884 **
## EDUCATION_High.School     -4.125e+03  1.420e+03  -2.905  0.00372 **
## EDUCATION_Masters         -2.236e+03  1.082e+03  -2.065  0.03905 *
## JOB_Blue.Collar            8.878e+02  1.309e+03   0.678  0.49790
## JOB_Clerical              -1.910e+02  1.375e+03  -0.139  0.88957
## JOB_Doctor                -3.153e+03  1.884e+03  -1.673  0.09450 .
## JOB_Home.Maker            -1.202e+02  1.502e+03  -0.080  0.93624
## JOB_Lawyer                -3.472e+02  1.156e+03  -0.300  0.76392
## JOB_Manager               -9.610e+02  1.227e+03  -0.783  0.43348
## JOB_Professional           1.427e+03  1.287e+03   1.108  0.26786
## JOB_Student                2.459e+01  1.560e+03   0.016  0.98743
## CAR_TYPE_Minivan           1.142e+02  8.585e+02   0.133  0.89422
## CAR_TYPE_Panel.Truck      -6.954e+01  9.573e+02  -0.073  0.94210
## CAR_TYPE_Pickup            5.863e+02  8.254e+02   0.710  0.47760
## CAR_TYPE_Sports.Car        2.203e+03  1.115e+03   1.976  0.04829 *
## CAR_TYPE_SUV               1.972e+03  1.047e+03   1.883  0.05990 .
## AGE_sin                   -1.383e+01  2.633e+02  -0.053  0.95810
## TIF_sin                   -2.986e+02  3.789e+02  -0.788  0.43072
## BLUEBOOK_sin               1.310e+02  2.608e+02   0.502  0.61551
## INCOME_AMT_BIN            -5.872e+02  1.277e+03  -0.460  0.64574
## YOJ_AMT_BIN                1.765e+03  1.016e+03   1.737  0.08265 .
```

```
## HOME_VAL_AMT_0_10K_BIN     2.352e+03  1.958e+03   1.202  0.22971
## HOME_VAL_AMT_60K_400K_BIN  2.906e+03  1.495e+03   1.944  0.05202 .
## OLDCLAIM_AMT_0_2K_BIN     -2.182e+03  1.453e+03  -1.502  0.13329
## OLDCLAIM_AMT_2K_10K_BIN   -4.937e+02  1.138e+03  -0.434  0.66450
## CLM_FREQ_AMT_BIN          -1.331e+03  1.183e+03  -1.125  0.26068
## MVR_PTS_AMT_BIN            1.740e+03  7.962e+02   2.186  0.02899 *
## CAR_AGE_AMT_BIN           -6.830e+02  7.034e+02  -0.971  0.33172
## TRAVTIME_AMT_BIN           7.013e+01  6.637e+02   0.106  0.91587
## KIDSDRIV_AMT_BIN_0        -9.177e+02  7.034e+02  -1.305  0.19217
## KIDSDRIV_AMT_BIN_1         2.099e+03  9.813e+02   2.139  0.03257 *
## HOMEKIDS_AMT_BIN_0        -5.624e+02  9.949e+02  -0.565  0.57194
## HOMEKIDS_AMT_BIN_3         8.055e+02  1.426e+03   0.565  0.57237
## YOJ_AMT_BIN_0_AND_9To14    6.616e+02  7.110e+02   0.930  0.35228
## INCOME_AMT_BIN_MISS_0     -2.171e+02  1.778e+03  -0.122  0.90281
## TIF_AMT_BIN_6             -1.556e+02  8.831e+02  -0.176  0.86016
## OLDCLAIM_AMT_BIN_MISS_0    4.081e+02  1.025e+03   0.398  0.69058
## MVR_PTS_AMT_BIN_0          4.154e+02  6.286e+02   0.661  0.50879
## MVR_PTS_AMT_BIN_5          1.819e+03  1.082e+03   1.681  0.09302 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7574 on 1643 degrees of freedom
## Multiple R-squared:  0.05812,    Adjusted R-squared:  0.02487
## F-statistic: 1.748 on 58 and 1643 DF,  p-value: 0.0005054
```

Pass 2- Based on that variance inflation factors (VIF) following variable "HOME_VAL_AMT_0_10K_BIN" has highest value < 3 and is removed from model, and model is evaluated without that variable. Adjusted R^2 values changed to 0.02461. Hence this variable is not adding lot of value to the model and can be removed. We stopped at this point as further removal of variables led to a rapid detoriation of Adjusted R^2.

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - KIDSDRIV - HOME_VAL_AMT_0_10K_BIN,
##     data = na.omit(DS_TARGET_AMT))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9838  -3162  -1404    806  75299
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.867e+03  4.919e+03   0.786  0.43182
## AGE                       1.254e+01  2.562e+01   0.489  0.62457
## HOMEKIDS                  2.883e+02  4.087e+02   0.705  0.48064
## YOJ                      -1.294e+02  1.189e+02  -1.088  0.27662
## INCOME                   -8.605e-03  1.056e-02  -0.815  0.41525
## HOME_VAL                 -1.617e-03  3.238e-03  -0.499  0.61753
## TRAVTIME                  4.999e+00  1.640e+01   0.305  0.76061
## BLUEBOOK                  1.532e-01  3.398e-02   4.510 6.96e-06 ***
## TIF                      -5.043e+01  1.115e+02  -0.452  0.65107
## OLDCLAIM                  1.527e-02  4.403e-02   0.347  0.72881
## CLM_FREQ                 -5.833e+02  3.284e+02  -1.776  0.07589 .
## MVR_PTS                   5.892e+02  2.590e+02   2.275  0.02306 *
## CAR_AGE                  -1.687e+02  7.618e+01  -2.214  0.02698 *
```

```
## CAR_USE_Commercial         6.536e+01  5.604e+02   0.117  0.90718
## MSTATUS_Yes               -1.890e+03  6.167e+02  -3.064  0.00222 **
## PARENT1_Yes               -6.960e+02  7.559e+02  -0.921  0.35732
## RED_CAR_yes               -1.426e+02  5.555e+02  -0.257  0.79743
## REVOKED_Yes               -1.119e+03  6.175e+02  -1.812  0.07010 .
## SEX_M                      1.930e+03  7.216e+02   2.675  0.00755 **
## URBANICITY_Rural          -2.570e+02  8.218e+02  -0.313  0.75456
## EDUCATION_Bachelors       -3.421e+03  1.322e+03  -2.588  0.00973 **
## EDUCATION_High.School     -4.093e+03  1.420e+03  -2.883  0.00398 **
## EDUCATION_Masters         -2.156e+03  1.081e+03  -1.996  0.04613 *
## JOB_Blue.Collar            9.015e+02  1.310e+03   0.688  0.49133
## JOB_Clerical              -2.158e+02  1.375e+03  -0.157  0.87530
## JOB_Doctor                -3.241e+03  1.883e+03  -1.721  0.08539 .
## JOB_Home.Maker            -2.327e+02  1.499e+03  -0.155  0.87670
## JOB_Lawyer                -3.148e+02  1.156e+03  -0.272  0.78533
## JOB_Manager               -9.840e+02  1.227e+03  -0.802  0.42254
## JOB_Professional           1.463e+03  1.287e+03   1.137  0.25571
## JOB_Student                2.080e+02  1.553e+03   0.134  0.89348
## CAR_TYPE_Minivan           1.520e+02  8.580e+02   0.177  0.85939
## CAR_TYPE_Panel.Truck      -2.704e+01  9.568e+02  -0.028  0.97745
## CAR_TYPE_Pickup            6.040e+02  8.254e+02   0.732  0.46439
## CAR_TYPE_Sports.Car        2.226e+03  1.115e+03   1.997  0.04599 *
## CAR_TYPE_SUV               1.999e+03  1.047e+03   1.909  0.05641 .
## AGE_sin                   -7.305e+00  2.632e+02  -0.028  0.97787
## TIF_sin                   -3.045e+02  3.789e+02  -0.804  0.42164
## BLUEBOOK_sin               1.315e+02  2.608e+02   0.504  0.61430
## INCOME_AMT_BIN            -2.323e+02  1.243e+03  -0.187  0.85172
## YOJ_AMT_BIN                1.773e+03  1.017e+03   1.744  0.08132 .
## HOME_VAL_AMT_60K_400K_BIN  1.341e+03  7.330e+02   1.830  0.06751 .
## OLDCLAIM_AMT_0_2K_BIN     -2.134e+03  1.452e+03  -1.470  0.14185
## OLDCLAIM_AMT_2K_10K_BIN   -4.602e+02  1.138e+03  -0.404  0.68598
## CLM_FREQ_AMT_BIN          -1.371e+03  1.182e+03  -1.159  0.24658
## MVR_PTS_AMT_BIN            1.703e+03  7.957e+02   2.140  0.03253 *
## CAR_AGE_AMT_BIN           -6.791e+02  7.035e+02  -0.965  0.33450
## TRAVTIME_AMT_BIN           5.845e+01  6.637e+02   0.088  0.92984
## KIDSDRIV_AMT_BIN_0        -9.284e+02  7.034e+02  -1.320  0.18707
## KIDSDRIV_AMT_BIN_1         2.062e+03  9.810e+02   2.102  0.03570 *
## HOMEKIDS_AMT_BIN_0        -5.557e+02  9.950e+02  -0.559  0.57656
## HOMEKIDS_AMT_BIN_3         7.544e+02  1.426e+03   0.529  0.59684
## YOJ_AMT_BIN_0_AND_9To14    6.725e+02  7.111e+02   0.946  0.34438
## INCOME_AMT_BIN_MISS_0     -2.521e+02  1.778e+03  -0.142  0.88723
## TIF_AMT_BIN_6             -1.820e+02  8.830e+02  -0.206  0.83674
## OLDCLAIM_AMT_BIN_MISS_0    4.252e+02  1.025e+03   0.415  0.67841
## MVR_PTS_AMT_BIN_0          3.963e+02  6.285e+02   0.631  0.52840
## MVR_PTS_AMT_BIN_5          1.761e+03  1.082e+03   1.628  0.10374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7575 on 1644 degrees of freedom
## Multiple R-squared:  0.05729,    Adjusted R-squared:  0.02461
## F-statistic: 1.753 on 57 and 1644 DF,  p-value: 0.000521
```

Final model was derived after number of iterations of variable eliminations were carried out. VIF values in the final model among variables < 3. In this scenario a model with slightly less performance was selected to

avoid collinearity effect among variables and reduced complexity.

5 Prediction Using Evaluation Data

Now that we have selected the final models for both the TARGET_FLAG and the TARGET_AMT, we will go ahead and use these models to predict the results for the evaluation dataset. After transforming the data to meet the needs of the trained models, we will apply the models in 2 steps.

Step 1 - Here we use the transformed evaluation dataset to predict for the TARGET_FLAG using the requisite predictors.

Step 2 - Once we have the prediction for the TARGET_FLAG, we will filter this data for only those rows that were predicted for a CRASH. We then use this smaller dataset to precict for the TARGET_AMT.

5.1 Tranformation of Evaluation Data

First we need to transform the evaluation dataset to account for all the predictors that were used in both the models.

5.2 Model Output for Logistic Regression

We now apply the final Logistic regression model that was trained for predicting the TARGET_FLAG. Below is a table of predictions.

Count for Crash / No Crash

Table 13: Predicted Crash Counts

Crash Predicted?	Counts
0	1410
1	304

5.3 Model Output for Linear Regression

Next we filter for the "predicted" crashes and we apply the final linear model to this smaller dataset to predict the TARGET_AMT. Below are the results for ONLY the "Crashed" records.

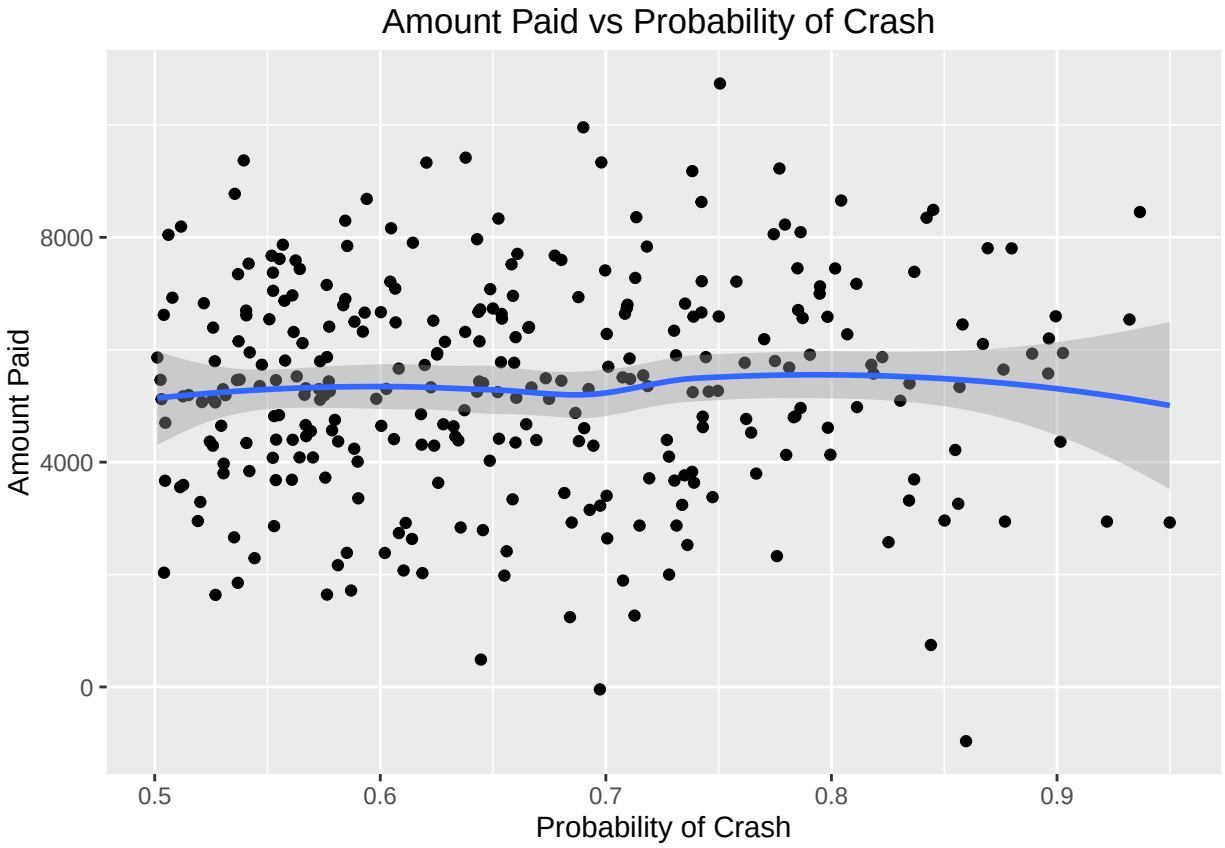
Top 10 Records by TARGET_AMT for TARGET_FLAG = 1

Table 14: Linear Regression Results

	INDEX	TARGET_FLAG	TARGET_FLAG_prob	TARGET_AMT
3	598	1	0.6446238	486.9448
4	5361	1	0.8441118	746.9066
5	2889	1	0.6840995	1241.5246
6	8672	1	0.7127093	1269.6456
7	6447	1	0.5269622	1638.1470

	INDEX	TARGET_FLAG	TARGET_FLAG_prob	TARGET_AMT
8	748	1	0.5763727	1643.4743
9	6329	1	0.5870784	1715.8916
10	4584	1	0.5368640	1852.6016
11	865	1	0.7076172	1893.0028
12	10115	1	0.6550164	1980.4875

5.4 Conclusion



Outcome from regression model and outcome from linear model was plotted in the chart above.It can be seen from the chart above that probability associated with classification and predicted amount from linear model does not show any specific patterns. From insurance business standpoint cases where probability of incident and repair expense amount is high will be the focus area top right side corner of the chart.

Appendix A: DATA621 Homework 04 R Code

```
if (!require("ggplot2",character.only = TRUE)) (install.packages("ggplot2",repos = "http://cran.us.r-pro
if (!require("MASS",character.only = TRUE)) (install.packages("MASS",repos = "http://cran.us.r-project.
if (!require("knitr",character.only = TRUE)) (install.packages("knitr",repos = "http://cran.us.r-projec
if (!require("xtable",character.only = TRUE)) (install.packages("xtable",repos = "http://cran.us.r-proj
if (!require("dplyr",character.only = TRUE)) (install.packages("dplyr",repos = "http://cran.us.r-projec
if (!require("psych",character.only = TRUE)) (install.packages("psych",repos = "http://cran.us.r-projec
if (!require("stringr",character.only = TRUE)) (install.packages("stringr",repos = "http://cran.us.r-pro
if (!require("car",character.only = TRUE)) (install.packages("car",repos = "http://cran.us.r-project.org
if (!require("faraway",character.only = TRUE)) (install.packages("faraway",repos = "http://cran.us.r-pro
if (!require("aod",character.only = TRUE)) (install.packages("aod",repos = "http://cran.us.r-project.org
if (!require("ISLR",character.only = TRUE)) (install.packages("ISLR",repos = "http://cran.us.r-project.
if (!require("AUC",character.only = TRUE)) (install.packages("AUC",repos = "http://cran.us.r-project.org
if (!require("ROCR",character.only = TRUE)) (install.packages("ROCR",repos = "http://cran.us.r-project.
if (!require("leaps",character.only = TRUE)) (install.packages("leaps",repos = "http://cran.us.r-projec
if (!require("pander",character.only = TRUE)) (install.packages("pander",repos = "http://cran.us.r-proj

if (!require("dummy",character.only = TRUE)) (install.packages("dummy",repos = "http://cran.us.r-projec
if (!require("reshape2",character.only = TRUE)) (install.packages("reshape2",repos = "http://cran.us.r-p
if (!require("popbio",character.only = TRUE)) (install.packages("popbio",repos = "http://cran.us.r-proj
if (!require("rpart",character.only = TRUE)) (install.packages("rpart",repos = "http://cran.us.r-projec
if (!require("pROC)",character.only = TRUE)) (install.packages("pROC)",repos = "http://cran.us.r-projec


library(ggplot2)
library(MASS)
library(knitr)
library(xtable)
library(dplyr)
library(psych)
library(stringr)
library(car)
library(faraway)
library(dummy)
library(reshape2)
library(popbio)
library(rpart)
library(pROC)
library(pander)


insure_train_full <- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW4/insuran

#kable(read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW4/insurevars.csv"), cap

variables<- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW4/insurevars.csv")
pander::pander(variables, split.cells = c(60, 20, 30), split.table = Inf, justify = 'left', caption = "



#str(insure_train_full)
```

```r
#levels(insure_train_full$MSTATUS)
#levels(insure_train_full$SEX)
#levels(insure_train_full$EDUCATION)
#levels(insure_train_full$JOB)
#levels(insure_train_full$CAR_TYPE)
#levels(insure_train_full$URBANICITY)
#levels(insure_train_full$REVOKED)

#summary(insure_train_full)




MSTATUS<-levels(insure_train_full$MSTATUS)
SEX<- levels(insure_train_full$SEX)
EDUCATION<- levels(insure_train_full$EDUCATION)
JOB<- levels(insure_train_full$JOB)
CAR_TYPE<- levels(insure_train_full$CAR_TYPE)
URBANICITY<- levels(insure_train_full$URBANICITY)
REVOKED<- levels(insure_train_full$REVOKED)
CAR_USE<- levels(insure_train_full$CAR_USE)


levels1<- (data.frame(cbind(MSTATUS,SEX, EDUCATION, CAR_TYPE, URBANICITY, CAR_USE, REVOKED, JOB)))


kable(levels1, caption = "Variable Levels")

insure_train_full$INCOME <- as.numeric(str_replace_all(insure_train_full$INCOME, pattern =  "[$*,]", re
insure_train_full$HOME_VAL <- as.numeric(str_replace_all(insure_train_full$HOME_VAL, pattern =  "[$*,]"
insure_train_full$BLUEBOOK <- as.numeric(str_replace_all(insure_train_full$BLUEBOOK, pattern =  "[$*,]"
insure_train_full$OLDCLAIM <- as.numeric(str_replace_all(insure_train_full$OLDCLAIM, pattern =  "[$*,]"

insure_train_full$MSTATUS <- as.factor(str_replace_all(insure_train_full$MSTATUS, "z_", ""))
insure_train_full$SEX <- as.factor(str_replace_all(insure_train_full$SEX, "z_", ""))
insure_train_full$EDUCATION <- as.factor(str_replace_all(insure_train_full$EDUCATION, "z_", ""))
insure_train_full$EDUCATION <- as.factor(str_replace_all(insure_train_full$EDUCATION, "<", ""))
insure_train_full$CAR_TYPE <- as.factor(str_replace_all(insure_train_full$CAR_TYPE, "z_", ""))
insure_train_full$URBANICITY <- as.factor(str_replace_all(insure_train_full$URBANICITY, "z_", ""))

insure_train_full$JOB <- as.character(insure_train_full$JOB)
insure_train_full$JOB[insure_train_full$JOB==""] <- "Unknown"
insure_train_full$JOB <- as.factor(str_replace_all(insure_train_full$JOB, "z_", ""))

insure_train_full <- insure_train_full[ -which( insure_train_full$CAR_AGE == -3 | insure_train_full$CAR_

trans<- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW4/trans.csv")
kable(trans, caption = "Variable Transformaton")

# Create Dummy Variable for 2 factor variables
insure_train_full$CAR_USE_Commercial <- ifelse(insure_train_full$CAR_USE=="Commercial", 1, 0)
insure_train_full$MSTATUS_Yes <- ifelse(insure_train_full$MSTATUS=="Yes", 1, 0)
```

```r
insure_train_full$PARENT1_Yes <- ifelse(insure_train_full$PARENT1=="Yes", 1, 0)
insure_train_full$RED_CAR_yes <- ifelse(insure_train_full$RED_CAR=="yes", 1, 0)
insure_train_full$REVOKED_Yes <- ifelse(insure_train_full$REVOKED=="Yes", 1, 0)
insure_train_full$SEX_M <- ifelse(insure_train_full$SEX=="M", 1, 0)
insure_train_full$URBANICITY_Rural <- ifelse(insure_train_full$URBANICITY=="Highly Rural/ Rural", 1, 0)

# remove original variables
insure_train_full <- select(insure_train_full, -CAR_USE, -MSTATUS, -PARENT1, -RED_CAR, -REVOKED, -SEX, -

insure_without_dummy <- insure_train_full
#- We will also create dummy variables for all the factors and drop the original variables.
dummy_vars<-as.data.frame(sapply(dummy(insure_train_full), FUN = as.numeric))
dummy_vars <- dummy_vars-1

# remove original variables
insure_train_full <- select(insure_train_full, -EDUCATION, -JOB, -CAR_TYPE)

insure_train_full <- cbind(insure_train_full, dummy_vars)
insure_train_full <- select(insure_train_full, -INDEX)

missings<- sapply(insure_train_full,function(x) sum(is.na(x)))
kable(data.frame(missings), caption = "Missing Values")


par(mfrow=c(2,3))
hist(insure_train_full$AGE)
hist(insure_train_full$YOJ)
hist(insure_train_full$INCOME)
hist(insure_train_full$HOME_VAL)
hist(insure_train_full$CAR_AGE)
# Missing Flags

insure_train_full$YOJ_MISS  <- ifelse(is.na(insure_train_full$YOJ), 1, 0)
insure_train_full$INCOME_MISS  <- ifelse(is.na(insure_train_full$INCOME), 1, 0)
insure_train_full$HOME_VAL_MISS  <- ifelse(is.na(insure_train_full$HOME_VAL), 1, 0)
insure_train_full$CAR_AGE_MISS  <- ifelse(is.na(insure_train_full$CAR_AGE), 1, 0)

# Missing Impute

# insure_train_full$AGE_IMPUTE <- insure_train_full$AGE
# insure_train_full$AGE_IMPUTE[is.na(insure_train_full$AGE_IMPUTE)] <- mean(insure_train_full$AGE_IMPUT
#
# insure_train_full$YOJ_IMPUTE <- insure_train_full$YOJ
# insure_train_full$YOJ_IMPUTE[is.na(insure_train_full$YOJ_IMPUTE)] <- mean(insure_train_full$YOJ_IMPUT
#
# insure_train_full$INCOME_IMPUTE <- insure_train_full$INCOME
# insure_train_full$INCOME_IMPUTE[is.na(insure_train_full$INCOME_IMPUTE)] <- median(insure_train_full$I
#
# insure_train_full$HOME_VAL_IMPUTE <- insure_train_full$HOME_VAL
# insure_train_full$HOME_VAL_IMPUTE[is.na(insure_train_full$HOME_VAL_IMPUTE)] <- median(insure_train_fu
#
# insure_train_full$CAR_AGE_IMPUTE <- insure_train_full$CAR_AGE
# insure_train_full$CAR_AGE_IMPUTE[is.na(insure_train_full$CAR_AGE_IMPUTE)] <- median(insure_train_full
```

```r
# Direct Impute

insure_train_full$AGE[is.na(insure_train_full$AGE)] <- mean(insure_train_full$AGE, na.rm = T)
insure_train_full$YOJ[is.na(insure_train_full$YOJ)] <- mean(insure_train_full$YOJ, na.rm = T)
insure_train_full$INCOME[is.na(insure_train_full$INCOME)] <- median(insure_train_full$INCOME, na.rm = T
insure_train_full$HOME_VAL[is.na(insure_train_full$HOME_VAL)] <- median(insure_train_full$HOME_VAL, na.
insure_train_full$CAR_AGE[is.na(insure_train_full$CAR_AGE)] <- median(insure_train_full$CAR_AGE, na.rm =

# Save point for Original data set with dummies created
insure_orig <- insure_train_full


ds_stats <- psych::describe(insure_train_full, skew = TRUE, na.rm = TRUE)
#ds_stats
kable(ds_stats[1:7], caption= "Data Summary")
kable(ds_stats[8:13], caption= "Data Summary (Cont)")

fun1 <- function(a, y) cor(y, a , use = 'na.or.complete')
Correlation_TARGET_FLAG <- sapply(insure_train_full, FUN = fun1, y=insure_train_full$TARGET_FLAG)
Correlation_TARGET_FLAG <- sort(Correlation_TARGET_FLAG, decreasing = TRUE)
kable(data.frame(Correlation_TARGET_FLAG), caption = "Correlation between TARGET_FLAG and predictor var


show_hist <- function(var, t) {

    col_x <- which(colnames(insure_train_full)==var)
    h0 <- select(insure_train_full[insure_train_full$TARGET_FLAG==1,], col_x)
    h1 <- select(insure_train_full[insure_train_full$TARGET_FLAG==0,], col_x)

    min_x <- min(select(insure_train_full, col_x), na.rm = TRUE)
    max_x <- max(select(insure_train_full, col_x), na.rm = TRUE)
    by_x <- (max_x - min_x) / 20


    #hist(h0[,1], breaks = 20, col=rgb(1,0,0,0.5), main=t, xlab = NA, xaxt = "n")
    hist(h0[,1], breaks = 20, col=rgb(0.1,0.1,0.1,0.5), main=t, xlab = NA, xaxt = "n")
    #axis(1, at = seq(min_x, max_x, by = by_x), las=2)

    hist(h1[,1], breaks = 20, col=rgb(0.8,0.8,0.8,0.5), add=T) #
#    axis(1, at = seq(min_x, max_x, by = by_x), las=2)

    box()
}

check_bins <- function(var, thresholds) {

    col_x <- which(colnames(insure_train_full)==var)
    old_x <- select(insure_train_full, col_x)
    cor_old <- cor(old_x, insure_train_full$TARGET_FLAG,use = 'na.or.complete')
    ds <- data.frame("Item" = "Original", "Correlation"= round(cor_old, 5))


    for(i in 1:length(thresholds)) {
```

```
        New_x <- ifelse(select(insure_train_full, col_x)<=thresholds[i],0,1)
        cor_new <- cor(New_x, insure_train_full$TARGET_FLAG,use = 'na.or.complete')
        ds_1 <- data.frame("Item" = as.character(thresholds[i]), "Correlation"= round(cor_new, 5))
        ds <- rbind(ds, ds_1)
    }
    return (ds)
}

par(mfrow=c(2,2))

show_hist("INCOME", "INCOME")
#check_bins("INCOME", c(0, 20000, 90000, 130000))

show_hist("YOJ", "YOJ")
#check_bins("YOJ", c(0, 4, 8, 15))

show_hist("HOME_VAL", "HOME_VAL")
#check_bins("HOME_VAL", c(0, 20000, 90000, 130000))

show_hist("OLDCLAIM", "OLDCLAIM")
#check_bins("OLDCLAIM", c(0, 5000, 10000, 15000, 20000, 40000))

show_hist("CLM_FREQ", "CLM_FREQ")
#check_bins("CLM_FREQ", c(0, 1, 2, 3, 4))

#table(insure_train_full$MVR_PTS)
show_hist("MVR_PTS", "MVR_PTS")
#check_bins("MVR_PTS", c(0:12))

#table(insure_train_full$CAR_AGE)
show_hist("CAR_AGE", "CAR_AGE")
#check_bins("CAR_AGE", c(1:27))

#table(insure_train_full$AGE)
show_hist("AGE", "AGE")
#check_bins("AGE", c(16:80))

#table(insure_train_full$BLUEBOOK)
show_hist("BLUEBOOK", "BLUEBOOK")
#check_bins("BLUEBOOK", c(11000, 41000, 41050, 57500, 58000))

#table(insure_train_full$TIF)
show_hist("TIF", "TIF")
#check_bins("TIF", c(1, 4, 6, 10, 24))

#table(insure_train_full$TRAVTIME)
show_hist("TRAVTIME", "TRAVTIME")
#check_bins("TRAVTIME", c(21, 59, 120))




#
```

```
mdata<- select(insure_train_full, AGE, BLUEBOOK, TIF)
mdata2 <- melt(mdata)
# Output the boxplot
par(mfrow=c(1,1))
p <- ggplot(data = mdata2, aes(x=variable, y=value)) +
  geom_boxplot() + ggtitle("Outliers Identification")
p + facet_wrap( ~ variable, scales="free")

par(mfrow=c(2,3))


x <- select(insure_train_full, -TARGET_AMT)
x <- x[complete.cases(x),]

# sapply(x, FUN = show_chart_logi.hist, y=x$TARGET_FLAG)

logi.hist.plot(x$REVOKED_Yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CAR_USE_Commercial,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mai
logi.hist.plot(x$CAR_TYPE_SUV,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$PARENT1_Yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$KIDSDRIV,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'K
logi.hist.plot(x$CAR_AGE,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'CA
logi.hist.plot(x$JOB_Clerical,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$HOMEKIDS,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'H
logi.hist.plot(x$JOB_Doctor,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CLM_FREQ,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'C
logi.hist.plot(x$SEX_M,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'SEX_
logi.hist.plot(x$MVR_PTS,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'MV
logi.hist.plot(x$EDUCATION_Masters,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainl
logi.hist.plot(x$CAR_TYPE_Van,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$CAR_TYPE_Minivan,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainla
logi.hist.plot(x$YOJ,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'YOJ')
logi.hist.plot(x$TIF,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'TIF')
logi.hist.plot(x$MSTATUS_Yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$RED_CAR_yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$JOB_Lawyer,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CAR_TYPE_Pickup,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainla
logi.hist.plot(x$JOB_Student,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$OLDCLAIM,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'O
logi.hist.plot(x$INCOME,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'INC
logi.hist.plot(x$EDUCATION_Bachelors,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', ma
logi.hist.plot(x$JOB_Manager,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$EDUCATION_High.School,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', m
logi.hist.plot(x$JOB_Home.Maker,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabe
logi.hist.plot(x$EDUCATION_PhD,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$TRAVTIME,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'T
logi.hist.plot(x$JOB_Professional,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainla
logi.hist.plot(x$URBANICITY_Rural,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainla
logi.hist.plot(x$JOB_Blue.Collar,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlab
logi.hist.plot(x$AGE,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'AGE')
logi.hist.plot(x$CAR_TYPE_Sports.Car,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', ma
logi.hist.plot(x$HOME_VAL,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'H
logi.hist.plot(x$BLUEBOOK,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'B
```

```
logi.hist.plot(x$PARENT1_Yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CAR_TYPE_Panel.Truck,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', ma

show_charts <- function(x, varlab, ...) {
    xlabel <- varlab
    xlab_log <- paste0(xlabel, '_log')
    xlab_sqrt <- paste0(xlabel, '_sqrt')
    xlab_sin <- paste0(xlabel, '_sin')
    xlab_inv <- paste0(xlabel, '_inv')

    mdata <- cbind(x, log(x), sqrt(x), sin(x), 1/x)
    colnames(mdata) <- c(xlabel, xlab_log, xlab_sqrt, xlab_sin, xlab_inv)
    mdata2 <- melt(mdata)
    mdata2 <- mdata2[, c(2:3)]
    names(mdata2) <- c("variable", "value")

    # Output the boxplot
    p <- ggplot(data = mdata2, aes(x=variable, y=value)) + geom_boxplot() + ggtitle("Outliers identifica
    p + facet_wrap( ~ variable, scales="free", ncol=5)
}

summary(insure_train_full$TIF)
show_charts(insure_train_full$TIF, 'TIF')
insure_train_full$TIF_sin <- sin(insure_train_full$TIF)
summary(insure_train_full$BLUEBOOK)
show_charts(insure_train_full$BLUEBOOK, 'BLUEBOOK')
insure_train_full$BLUEBOOK_sin <- sin(insure_train_full$BLUEBOOK)
summary(insure_train_full$AGE)
show_charts(insure_train_full$AGE, 'AGE')
insure_train_full$AGE_sin <- sin(insure_train_full$AGE)

insure_train_full$CAR_TYPE_FLAG_BIN <- ifelse(insure_train_full$CAR_TYPE_Minivan | insure_train_full$CA


insure_train_full$EDUCATION_FLAG_BIN <- ifelse(insure_train_full$EDUCATION_High.School, 0, 1)


insure_train_full$JOB_TYPE_FLAG_BIN <- ifelse(insure_train_full$JOB_Student |  insure_train_full$JOB_Ho


insure_train_full$INCOME_FLAG_BIN <- ifelse(insure_train_full$INCOME <=0, 1, 0)


insure_train_full$YOJ_FLAG_BIN <- ifelse(insure_train_full$YOJ <=0, 1, 0)


insure_train_full$HOME_VAL_FLAG_BIN <- ifelse(insure_train_full$HOME_VAL <=0, 1, 0)

insure_train_full$OLDCLAIM_FLAG_BIN <- ifelse(insure_train_full$OLDCLAIM <=0, 1, 0)
insure_train_full$CLM_FREQ_FLAG_BIN <- ifelse(insure_train_full$CLM_FREQ <=0, 1, 0)
insure_train_full$MVR_PTS_FLAG_BIN <- ifelse(insure_train_full$MVR_PTS <=0, 1, 0)

insure_train_full$CAR_AGE_FLAG_BIN <- ifelse(insure_train_full$CAR_AGE <=1, 1, 0)
```

```
insure_train_full$TRAVTIME_FLAG_BIN <- ifelse(insure_train_full$TRAVTIME <=20, 1, 0)



#write.csv(insure_train_full, file = "D:/CUNY/Courses/Business Analytics and Data Mining/Assignments/da

#DS_TARGET_FLAG <- insure_train_full
DS_TARGET_FLAG <- select(insure_train_full, -TARGET_AMT, -JOB_Blue.Collar, -JOB_Clerical, -JOB_Doctor, 

# New Additional Variables.
#-AGE, -AGE_IMPUTE, -BLUEBOOK,-CAR_AGE_IMPUTE, -CAR_AGE_MISS, -CLM_FREQ, -HOME_VAL, -HOME_VAL_IMPUTE, -

str(DS_TARGET_FLAG)


smp_size <- floor(0.80 * nrow(DS_TARGET_FLAG))

## set the seed to make your partition reproductible
set.seed(123)

train_index <- sample(seq_len(nrow(DS_TARGET_FLAG)), size = smp_size)

DS_TARGET_FLAG_TRAIN<- DS_TARGET_FLAG[train_index, ]
DS_TARGET_FLAG_VALID <- DS_TARGET_FLAG[-train_index, ]

TF_Model1 <- glm(TARGET_FLAG ~ ., data = na.omit(DS_TARGET_FLAG_TRAIN), family = "binomial")
TF_Model1_ref<- TF_Model1
#TF_Model1_ref<- step(TF_Model1, direction="backward")
summary(TF_Model1_ref)


 # grow tree
TF_Model2 <- rpart(TARGET_FLAG~., data=DS_TARGET_FLAG_TRAIN, method = "class")
plotcp(TF_Model2)

# plot tree
# plot(model2, uniform=TRUE, main="Classification Tree for TARGET_FLAG")
# text(model2, use.n=TRUE, all=TRUE, cex=.8)

# create attractive postscript plot of tree
#post(fit, file = "c:/tree.ps", title = "Classification Tree for Kyphosis")

TF_Model2_ref <- prune(TF_Model2, cp = TF_Model2$cptable[which.min(TF_Model2$cptable[,"xerror"]),"CP"])

#par(mfrow=c(1,1))
#plot(TF_Model2_ref, uniform=TRUE, main="P")
#text(TF_Model2_ref, use.n=TRUE, all=TRUE, cex=.8)

#printcp(TF_Model2_ref) # display the results


#post(pfit, file = "c:/ptree.ps",    title = "Pruned Classification Tree for Kyphosis")
```

```r
# plot the pruned tree

par(mfrow = c(1, 3), mar = rep(0.1, 4))
#par(mfrow=c(1,1))
plot(TF_Model2_ref, uniform=TRUE, main=NA)
text(TF_Model2_ref, use.n=TRUE, all=TRUE, cex=.8)
#plotcp(TF_Model2_ref)



#Following function Eval() will be used to calculate various metrics related to the model like Accuracy

Eval<-function(x){
    TP<-x$Freq[x$metrics=="TRUE_1"]
    FP<-x$Freq[x$metrics=="FALSE_1"]
    TN<-x$Freq[x$metrics=="FALSE_0"]
    FN<-x$Freq[x$metrics=="TRUE_0"]
    Accuracy <-(TP+TN)/(TP+TN+FP+FN)
    Error_Rate<-(FP+FN)/(TP+TN+FP+FN)
    Precision<-TP/(TP+FP)
    sensitivity<-TP/(TP+FN)
    specificity<-TN/(TN+FP)
    F1_Score=2*Precision*sensitivity/(sensitivity+specificity)
    eval_result<-data.frame(Accuracy=c(0),Error_Rate=c(0),Precision=c(0),sensitivity=c(0),specificity=c

    eval_result[1,1]<-Accuracy
    eval_result[1,2]<-Error_Rate
    eval_result[1,3]<- Precision
    eval_result[1,4]<-sensitivity
    eval_result[1,5]<-specificity
    eval_result[1,6]<-F1_Score
    eval_result
}

model_comparison<-data.frame(Accuracy=c(0),Error_Rate=c(0),Precision=c(0),sensitivity=c(0),specificity=

#confusion matrix

DS_TARGET_FLAG_VALID$M1_TARGET_FLAG <- predict(TF_Model1, newdata=DS_TARGET_FLAG_VALID, type="response"
df_pre_train1<-as.data.frame(table(DS_TARGET_FLAG_VALID$M1_TARGET_FLAG>0.5,DS_TARGET_FLAG_VALID$TARGET_
df_pre_train1$metrics <- paste(df_pre_train1$Var1,df_pre_train1$Var2, sep = '_')
model_comparison[1,]<-Eval(df_pre_train1)
model_comparison[1,c("AUC")]<-c(auc(DS_TARGET_FLAG_VALID$TARGET_FLAG, DS_TARGET_FLAG_VALID$M1_TARGET_FLA
kable(model_comparison[1,],row.names = TRUE, caption = " Model 1 evaluation KPIs")


DS_TARGET_FLAG_VALID$M2_TARGET_FLAG <- predict(TF_Model1_ref,newdata=DS_TARGET_FLAG_VALID)
df_pre_train1<-as.data.frame(table(DS_TARGET_FLAG_VALID$M2_TARGET_FLAG>0.5,DS_TARGET_FLAG_VALID$TARGET_
df_pre_train1$metrics <- paste(df_pre_train1$Var1,df_pre_train1$Var2, sep = '_')
model_comparison[2,]<-Eval(df_pre_train1)
model_comparison[2,c("AUC")]<-c(auc(DS_TARGET_FLAG_VALID$TARGET_FLAG, DS_TARGET_FLAG_VALID$M2_TARGET_FLA
kable(model_comparison[2,],row.names = TRUE, caption = " Model 2 evaluation KPIs")
```

```
#ds <- select(DS_TARGET_FLAG_VALID, -AGE, -AGE_IMPUTE, -BLUEBOOK, -CAR_AGE_IMPUTE, -CAR_AGE_MISS, -CLM_

DS_TARGET_FLAG_VALID$M2_TARGET_FLAG <- predict(TF_Model2_ref,newdata=DS_TARGET_FLAG_VALID)[,2]
df_pre_train1<-as.data.frame(table(DS_TARGET_FLAG_VALID$M2_TARGET_FLAG>0.5,DS_TARGET_FLAG_VALID$TARGET_
df_pre_train1$metrics <- paste(df_pre_train1$Var1,df_pre_train1$Var2, sep = '_')
model_comparison[3,]<-Eval(df_pre_train1)
model_comparison[3,c("AUC")]<-c(auc(DS_TARGET_FLAG_VALID$TARGET_FLAG, DS_TARGET_FLAG_VALID$M2_TARGET_FLA
kable(model_comparison,row.names = TRUE, caption = " Model 3 evaluation KPIs")

model_comparison$Model_No<-c(1:3)
kable(model_comparison[,c("Model_No","Accuracy","Error_Rate","AUC","Precision","sensitivity","specificit


summary(TF_Model1)


exp(cbind(OR = coef(TF_Model1), confint.default(TF_Model1)))

#AUC

myRoc <- roc(DS_TARGET_FLAG_VALID$TARGET_FLAG~DS_TARGET_FLAG_VALID$M1_TARGET_FLAG, DS_TARGET_FLAG_VALID
plot(myRoc, main="ROC Curve for Classification data")

# pred <- prediction(TF_Model1_ref, DS_TARGET_FLAG_VALID$TARGET_FLAG)
# perf <- performance(pred, measure = "tpr", x.measure = "fpr")
#
# auc <- performance(pred, measure = "auc")
# auc <- auc@y.values[[1]]
#
# roc.data <- data.frame(fpr=unlist(perf@x.values),
#                        tpr=unlist(perf@y.values),
#                        model="GLM")
# ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
#     geom_ribbon(alpha=0.2) +
#     geom_line(aes(y=tpr)) +
#     ggtitle(paste0("ROC Curve w/ AUC=", auc))

plot_pred_type_distribution <- function(df, threshold) {
  v <- rep(NA, nrow(df))
  v <- ifelse(DS_TARGET_FLAG_VALID$M1_TARGET_FLAG >= threshold & DS_TARGET_FLAG_VALID$TARGET_FLAG == 1,
  v <- ifelse(DS_TARGET_FLAG_VALID$M1_TARGET_FLAG >= threshold & DS_TARGET_FLAG_VALID$TARGET_FLAG == 0,
  v <- ifelse(DS_TARGET_FLAG_VALID$M1_TARGET_FLAG < threshold & DS_TARGET_FLAG_VALID$TARGET_FLAG == 1,
  v <- ifelse(DS_TARGET_FLAG_VALID$M1_TARGET_FLAG < threshold & DS_TARGET_FLAG_VALID$TARGET_FLAG == 0,

  DS_TARGET_FLAG_VALID$pred_type <- v

  ggplot(data=DS_TARGET_FLAG_VALID, aes(x=TARGET_FLAG, y=M1_TARGET_FLAG)) +
    geom_violin(fill=rgb(1,1,1,alpha=0.6), color=NA) +
    geom_jitter(aes(color=pred_type), alpha=0.6) +
    geom_hline(yintercept=threshold, color="red", alpha=0.6) +
    scale_color_discrete(name = "type") +
    labs(title=sprintf("Threshold at %.2f", threshold))
}
```

```
DS_TARGET_FLAG_VALID$M1_TARGET_FLAG <- predict(TF_Model1_ref, newdata=DS_TARGET_FLAG_VALID, type="respor

plot_pred_type_distribution (DS_TARGET_FLAG_VALID,0.5)

insure_train_crash <- insure_without_dummy[insure_without_dummy$TARGET_FLAG==1,]

dummy_vars<-as.data.frame(sapply(dummy(insure_train_crash), FUN = as.numeric))
dummy_vars <- dummy_vars-1

insure_train_crash <- cbind(insure_train_crash, dummy_vars)



#write.csv(insure_train_crash, file = "D:/CUNY/Courses/Business Analytics and Data Mining/Assignments/d


#insure_train_crash <- select(insure_train_crash, -CAR_AGE_FLAG_BIN, -CAR_TYPE_FLAG_BIN, -CLM_FREQ_FLAG_

#str(insure_train_crash)

ds_stats <- psych::describe(insure_train_crash, skew = TRUE, na.rm = TRUE)
#ds_stats
kable(ds_stats[1:7], caption= "Data Summary")
kable(ds_stats[8:13], caption= "Data Summary (Cont)")

fun1 <- function(a, y) cor(y, a , use = 'na.or.complete')
x<-select(insure_train_crash, -EDUCATION, -JOB, -CAR_TYPE)
Correlation_TARGET_AMT <- sapply(x, FUN = fun1, y=insure_train_crash$TARGET_AMT)
Correlation_TARGET_AMT <- sort(Correlation_TARGET_AMT, decreasing = TRUE)
kable(data.frame(Correlation_TARGET_AMT), caption = "Correlation between TARGET_AMT and predictor variab


check_bins <- function(var, thresholds) {
    col_x <- which(colnames(insure_train_crash)==var)
    old_x <- select(insure_train_crash, col_x)
    cor_old <- cor(old_x, insure_train_crash$TARGET_AMT,use = 'na.or.complete')
    ds <- data.frame("Item" = "Original", "Correlation"= round(cor_old, 5))

    old_tresh <- 0
    for(i in 1:length(thresholds)) {
        New_x <- ifelse((select(insure_train_crash, col_x) >= old_tresh & select(insure_train_crash, col

        cor_new <- cor(New_x, insure_train_crash$TARGET_AMT,use = 'na.or.complete')

        ds_1 <- data.frame("Item" = as.character(thresholds[i]), "Correlation"= round(cor_new, 5))

        ds <- rbind(ds, ds_1)
        old_tresh <- thresholds[i]
    }

    return (ds)
}
```

```
par(mfrow=c(2,2))
plot(insure_train_crash$INCOME, insure_train_crash$TARGET_AMT, xlab = "INCOME", ylab = "TARGET_AMT")
#check_bins("INCOME", c(0, 50000, 125000, 200000))

plot(insure_train_crash$YOJ, insure_train_crash$TARGET_AMT, xlab = "YOJ", ylab = "TARGET_AMT")
#check_bins("YOJ", c(0:19))

plot(insure_train_crash$HOME_VAL, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "HOME_VAL")
#check_bins("HOME_VAL", c(seq(0, 600000, 10000)))

plot(insure_train_crash$OLDCLAIM, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "OLDCLAIM")
#hist(insure_train_crash$OLDCLAIM, breaks=50)
#check_bins("OLDCLAIM", c(seq(0, 50000, 1000)))

#show_hist("CLM_FREQ")
plot(insure_train_crash$CLM_FREQ, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "CLM_FREQ")
#check_bins("CLM_FREQ", c(0, 1, 2, 3, 4))

#table(insure_train_full$MVR_PTS)
plot(insure_train_crash$MVR_PTS, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "MVR_PTS")
#check_bins("MVR_PTS", c(0:12))

#table(insure_train_full$CAR_AGE)
#show_hist("CAR_AGE")
plot(insure_train_crash$CAR_AGE, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "CAR_AGE")
#check_bins("CAR_AGE", c(1:27))

#table(insure_train_full$AGE)
plot(insure_train_crash$AGE, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "AGE")
#show_hist("AGE")
#check_bins("AGE", c(16:80))

#table(insure_train_full$BLUEBOOK)
plot(insure_train_crash$BLUEBOOK, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "BLUEBOOK")
#show_hist("BLUEBOOK")
#check_bins("BLUEBOOK", c(5000, 10000, 20000, 30000, 45000, 57500, 58000))

# table(insure_train_full$TIF)
# show_hist("TIF")
plot(insure_train_crash$TIF, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "TIF")
# check_bins("TIF", c(1, 4, 6, 10, 24))

#table(insure_train_full$TRAVTIME)
plot(insure_train_crash$TRAVTIME, insure_train_crash$TARGET_AMT, ylab = "TARGET_AMT", xlab = "TRAVTIME")
#show_hist("TRAVTIME")
#check_bins("TRAVTIME", c(21, 59, 120))


# AGE, BLUEBOOK, CAR_AGE, CLM_FREQ, HOME_VAL, HOMEKIDS, INCOME, KIDSDRIV, MVR_PTS, OLDCLAIM, TIF, TRAVT

mdata<- select(insure_train_crash, AGE, BLUEBOOK, TIF)
mdata2 <- melt(mdata)
# Output the boxplot
```

```
p <- ggplot(data = mdata2, aes(x=variable, y=value)) +
  geom_boxplot() + ggtitle("Outliers Identification")
p + facet_wrap( ~ variable, scales="free", ncol=5)

show_charts <- function(x, varlab, ...) {
    xlabel <- varlab
    xlab_log <- paste0(xlabel, '_log')
    xlab_sqrt <- paste0(xlabel, '_sqrt')
    xlab_sin <- paste0(xlabel, '_sin')
    xlab_inv <- paste0(xlabel, '_inv')

    mdata <- cbind(x, log(x), sqrt(x), sin(x), 1/x)
    colnames(mdata) <- c(xlabel, xlab_log, xlab_sqrt, xlab_sin, xlab_inv)
    mdata2 <- melt(mdata)
    mdata2 <- mdata2[, c(2:3)]
    names(mdata2) <- c("variable", "value")

    # Output the boxplot
    p <- ggplot(data = mdata2, aes(x=variable, y=value)) + geom_boxplot() + ggtitle("Outliers Treatment
    p + facet_wrap( ~ variable, scales="free", ncol=5)
}

#KIDSDRIV, AGE, CAR_AGE, MVR_PTS, TIF, TRAVTIME and YOJ
show_charts(insure_train_crash$AGE, 'AGE')
insure_train_crash$AGE_sin <- sin(insure_train_crash$AGE)
show_charts(insure_train_crash$TIF, 'TIF')
insure_train_crash$TIF_sin <- sin(insure_train_crash$TIF)
show_charts(insure_train_crash$BLUEBOOK, 'BLUEBOOK')
insure_train_crash$BLUEBOOK_sin <- sin(insure_train_crash$BLUEBOOK)
insure_train_crash$CAR_TYPE_AMT_BIN <- ifelse(insure_train_crash$CAR_TYPE_Van | insure_train_crash$CAR_


insure_train_crash$EDUCATION_AMT_BIN <- ifelse(insure_train_crash$EDUCATION_High.School, 1, 0)


insure_train_crash$JOB_TYPE_AMT_BIN <- ifelse(insure_train_crash$JOB_Lawyer |  insure_train_crash$JOB_P


insure_train_crash$INCOME_AMT_BIN <- ifelse(insure_train_crash$INCOME <=125000, 1, 0)


insure_train_crash$YOJ_AMT_BIN <- ifelse((insure_train_crash$YOJ>=7 & insure_train_crash$YOJ<=17), 1, 0


insure_train_crash$HOME_VAL_AMT_0_10K_BIN <- ifelse((insure_train_crash$HOME_VAL>=0 & insure_train_crash

insure_train_crash$HOME_VAL_AMT_60K_400K_BIN <- ifelse((insure_train_crash$HOME_VAL>=60000 & insure_tra


insure_train_crash$OLDCLAIM_AMT_0_2K_BIN <- ifelse((insure_train_crash$OLDCLAIM>=0 & insure_train_crash

insure_train_crash$OLDCLAIM_AMT_2K_10K_BIN <- ifelse((insure_train_crash$OLDCLAIM>=2001 & insure_train_
```

```
insure_train_crash$CLM_FREQ_AMT_BIN <- ifelse(insure_train_crash$CLM_FREQ <4, 1, 0)
insure_train_crash$MVR_PTS_AMT_BIN <- ifelse(insure_train_crash$MVR_PTS <=2, 1, 0)
insure_train_crash$CAR_AGE_AMT_BIN <- ifelse(insure_train_crash$CAR_AGE <=1, 1, 0)
insure_train_crash$TRAVTIME_AMT_BIN <- ifelse(insure_train_crash$TRAVTIME <=20, 1, 0)


insure_train_crash$KIDSDRIV_AMT_BIN_0  <- ifelse(insure_train_crash$KIDSDRIV <=0, 1, 0)
insure_train_crash$KIDSDRIV_AMT_BIN_1  <- ifelse(insure_train_crash$KIDSDRIV <=1, 1, 0)

insure_train_crash$HOMEKIDS_AMT_BIN_0  <- ifelse(insure_train_crash$HOMEKIDS <=0, 1, 0)
insure_train_crash$HOMEKIDS_AMT_BIN_3  <- ifelse(insure_train_crash$HOMEKIDS <=3, 1, 0)

insure_train_crash$YOJ_AMT_BIN_0_AND_9To14  <- ifelse((insure_train_crash$YOJ ==0 | (insure_train_crash$

insure_train_crash$INCOME_AMT_BIN_MISS_0  <- ifelse((is.na(insure_train_crash$INCOME) |  insure_train_c

insure_train_crash$HOME_VAL_AMT_BIN_MISS_0  <- ifelse((is.na(insure_train_crash$HOME_VAL) |  insure_tra

insure_train_crash$EDUCATION_AMT_BIN_HS  <- ifelse(insure_train_crash$EDUCATION_High.School==1, 1, 0)
insure_train_crash$EDUCATION_AMT_BIN_HS_B  <- ifelse((insure_train_crash$EDUCATION_Bachelors |  insure_

insure_train_crash$JOB_AMT_BIN_CPSB  <- ifelse((insure_train_crash$JOB_Clerical |  insure_train_crash$JO

insure_train_crash$TIF_AMT_BIN_6  <- ifelse(insure_train_crash$TIF <=6, 1, 0)

insure_train_crash$CAR_TYPE_AMT_BIN_V_PT_MV  <- ifelse((insure_train_crash$CAR_TYPE_Van |  insure_train

insure_train_crash$OLDCLAIM_AMT_BIN_MISS_0  <- ifelse((is.na(insure_train_crash$OLDCLAIM) |  insure_tra

insure_train_crash$CLM_FREQ_AMT_BIN_0  <- ifelse(insure_train_crash$CLM_FREQ<=0, 1, 0)
insure_train_crash$CLM_FREQ_AMT_BIN_3  <- ifelse(insure_train_crash$CLM_FREQ<=3, 1, 0)

insure_train_crash$MVR_PTS_AMT_BIN_0  <- ifelse(insure_train_crash$MVR_PTS<=0, 1, 0)
insure_train_crash$MVR_PTS_AMT_BIN_5  <- ifelse(insure_train_crash$MVR_PTS<=5, 1, 0)

#write.csv(ds, file = "D:/CUNY/Courses/Business Analytics and Data Mining/Assignments/data621-ctg5/HW4/

insure_train_crash <- select(insure_train_crash, -TARGET_FLAG, -INDEX)

DS_TARGET_AMT <- insure_train_crash

#DS_TARGET_AMT <- select(insure_train_crash, -AGE, -BLUEBOOK, -CAR_AGE, -CAR_TYPE_Minivan, -CAR_TYPE_Pa

DS_TARGET_AMT <- select(insure_train_crash, -EDUCATION, -JOB, -CAR_TYPE, -EDUCATION_AMT_BIN_HS_B, -EDUC


#origvars <- c(3, 9, 14, 44, 35, 36, 37, 38, 39, 40, 15, 12, 22, 23, 24, 25, 7, 43, 4, 6, 42, 26, 27, 2

#DS_TARGET_AMT <- select(DS_TARGET_AMT, )
# TA_Model1 <- lm(TARGET_AMT~.-EDUCATION-JOB-CAR_TYPE-EDUCATION_AMT_BIN_HS_B-EDUCATION_AMT_BIN_HS-EDUCA

TA_Model1 <- lm(TARGET_AMT~., data=na.omit(DS_TARGET_AMT))
summary(TA_Model1)
```

```
# TA_Model1_ref<-step(TA_Model1,direction="backward",test="F")
# summary(TA_Model1_ref)


DS_SELECTED_VARS <- select(DS_TARGET_AMT, TARGET_AMT, AGE, BLUEBOOK, CAR_AGE, CAR_AGE_AMT_BIN, CAR_USE_C

TA_Model2<- lm(TARGET_AMT~., data=DS_SELECTED_VARS)
summary(TA_Model2)

# #DS_TARGET_AMT_TRAIN_ORIG <- select(DS_TARGET_AMT_TRAIN, origvars)
# TA_Model2 <- lm(TARGET_AMT~., data=na.omit(DS_TARGET_AMT))
# summary(TA_Model2)

# TA_Model2_ref<-step(TA_Model2,direction="backward",test="F")
# summary(TA_Model2_ref)


# TA_model2 <- lm(TARGET_AMT~.-EDUCATION_PhD-JOB_Unknown-CAR_TYPE_Van-CAR_TYPE_AMT_BIN-EDUCATION_AMT_BI
#
#
# TA_model2 <- lm(TARGET_AMT~BLUEBOOK+CAR_AGE+REVOKED_Yes+SEX_M+EDUCATION_Bachelors+EDUCATION_High.Scho


#  Analysis of plot on residuals to verify normal distribution of residuals

library(MASS)
sresid <- studres(TA_Model1)
hist(sresid, freq=FALSE,
     main="Distribution of Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)

step4.res <- resid(TA_Model1)
score<-predict(TA_Model1,type="response")

plot(score, step4.res,  ylab="Residuals", xlab="Predicted Cost of Repair", main="Residual vs Predicted
abline(0, 0)
##
## library(car)
##
## ncvTest(TA_Model1)
## # plot studentized residuals vs. fitted values
## spreadLevelPlot(TA_Model1)
##
##
##

library(faraway)

# Evaluate Collinearity of the variables in model "step4" vif(step4) # variance inflation factors
#kable(sqrt(vif(TA_Model1)), caption = 'Analysis of collinearity')
```

```r
sqrt(vif(TA_Model1))

TA_Model1 <- lm(TARGET_AMT~.-KIDSDRIV, data=na.omit(DS_TARGET_AMT))
#summary(TA_Model1)
TA_Model1 <- lm(TARGET_AMT~.-KIDSDRIV, data=na.omit(DS_TARGET_AMT))
summary(TA_Model1)

# Adjusted R2 = 0.02461
TA_Model1 <- lm(TARGET_AMT~.-KIDSDRIV-HOME_VAL_AMT_0_10K_BIN, data=na.omit(DS_TARGET_AMT))
summary(TA_Model1)

# # Adjusted R2 = 0.01871
# TA_Model1 <- lm(TARGET_AMT~.-KIDSDRIV-HOME_VAL_AMT_0_10K_BIN-OLDCLAIM_AMT_2K_10K_BIN-EDUCATION_High.S
# summary(TA_Model1)



eval_ds <- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW4/insurance-evaluati

eval_ds$INCOME <- as.numeric(str_replace_all(eval_ds$INCOME, pattern =  "[$*,]", replacement = ""))
eval_ds$HOME_VAL <- as.numeric(str_replace_all(eval_ds$HOME_VAL, pattern =  "[$*,]", replacement = ""))
eval_ds$BLUEBOOK <- as.numeric(str_replace_all(eval_ds$BLUEBOOK, pattern =  "[$*,]", replacement = ""))
eval_ds$OLDCLAIM <- as.numeric(str_replace_all(eval_ds$OLDCLAIM, pattern =  "[$*,]", replacement = ""))

eval_ds$MSTATUS <- as.factor(str_replace_all(eval_ds$MSTATUS, "z_", ""))
eval_ds$SEX <- as.factor(str_replace_all(eval_ds$SEX, "z_", ""))
eval_ds$EDUCATION <- as.factor(str_replace_all(eval_ds$EDUCATION, "z_", ""))
eval_ds$EDUCATION <- as.factor(str_replace_all(eval_ds$EDUCATION, "<", ""))
eval_ds$CAR_TYPE <- as.factor(str_replace_all(eval_ds$CAR_TYPE, "z_", ""))
eval_ds$URBANICITY <- as.factor(str_replace_all(eval_ds$URBANICITY, "z_", ""))

eval_ds$JOB <- as.character(eval_ds$JOB)
eval_ds$JOB[eval_ds$JOB==""] <- "Unknown"
eval_ds$JOB <- as.factor(str_replace_all(eval_ds$JOB, "z_", ""))

eval_ds <- eval_ds[ -which( eval_ds$CAR_AGE == -3 | eval_ds$CAR_AGE == 0 ) , ]

# Create Dummy Variable for 2 factor variables
eval_ds$CAR_USE_Commercial <- ifelse(eval_ds$CAR_USE=="Commercial", 1, 0)
eval_ds$MSTATUS_Yes <- ifelse(eval_ds$MSTATUS=="Yes", 1, 0)
eval_ds$PARENT1_Yes <- ifelse(eval_ds$PARENT1=="Yes", 1, 0)
eval_ds$RED_CAR_yes <- ifelse(eval_ds$RED_CAR=="yes", 1, 0)
eval_ds$REVOKED_Yes <- ifelse(eval_ds$REVOKED=="Yes", 1, 0)
eval_ds$SEX_M <- ifelse(eval_ds$SEX=="M", 1, 0)
eval_ds$URBANICITY_Rural <- ifelse(eval_ds$URBANICITY=="Highly Rural/ Rural", 1, 0)

# remove original variables
eval_ds <- select(eval_ds, -CAR_USE, -MSTATUS, -PARENT1, -RED_CAR, -REVOKED, -SEX, -URBANICITY)

#- We will also create dummy variables for all the factors and drop the original variables.
dummy_vars<-as.data.frame(sapply(dummy(eval_ds), FUN = as.numeric))
dummy_vars <- dummy_vars-1
```

```r
# remove original variables
#eval_ds <- select(eval_ds, -EDUCATION, -JOB, -CAR_TYPE)

eval_ds <- cbind(eval_ds, dummy_vars)

eval_ds$YOJ_MISS  <- ifelse(is.na(eval_ds$YOJ), 1, 0)
eval_ds$INCOME_MISS  <- ifelse(is.na(eval_ds$INCOME), 1, 0)
eval_ds$HOME_VAL_MISS  <- ifelse(is.na(eval_ds$HOME_VAL), 1, 0)
eval_ds$CAR_AGE_MISS  <- ifelse(is.na(eval_ds$CAR_AGE), 1, 0)

# Direct Impute

insure_train_full$AGE[is.na(insure_train_full$AGE)] <- mean(insure_train_full$AGE, na.rm = T)
insure_train_full$YOJ[is.na(insure_train_full$YOJ)] <- mean(insure_train_full$YOJ, na.rm = T)
insure_train_full$INCOME[is.na(insure_train_full$INCOME)] <- median(insure_train_full$INCOME, na.rm = T
insure_train_full$HOME_VAL[is.na(insure_train_full$HOME_VAL)] <- median(insure_train_full$HOME_VAL, na.
insure_train_full$CAR_AGE[is.na(insure_train_full$CAR_AGE)] <- median(insure_train_full$CAR_AGE, na.rm =

eval_ds$TIF_sin <- sin(eval_ds$TIF)
eval_ds$BLUEBOOK_sin <- sin(eval_ds$BLUEBOOK)
eval_ds$AGE_sin <- sin(eval_ds$AGE)

eval_ds$CAR_TYPE_FLAG_BIN <- ifelse(eval_ds$CAR_TYPE_Minivan | eval_ds$CAR_TYPE_Panel.Truck, 1, 0)

eval_ds$EDUCATION_FLAG_BIN <- ifelse(eval_ds$EDUCATION_High.School, 0, 1)

eval_ds$JOB_TYPE_FLAG_BIN <- ifelse(eval_ds$JOB_Student |  eval_ds$JOB_Home.Maker | eval_ds$JOB_Clerical


eval_ds$INCOME_FLAG_BIN <- ifelse(eval_ds$INCOME <=0, 1, 0)

eval_ds$YOJ_FLAG_BIN <- ifelse(eval_ds$YOJ <=0, 1, 0)

eval_ds$HOME_VAL_FLAG_BIN <- ifelse(eval_ds$HOME_VAL <=0, 1, 0)

eval_ds$OLDCLAIM_FLAG_BIN <- ifelse(eval_ds$OLDCLAIM <=0, 1, 0)

eval_ds$CLM_FREQ_FLAG_BIN <- ifelse(eval_ds$CLM_FREQ <=0, 1, 0)


eval_ds$MVR_PTS_FLAG_BIN <- ifelse(eval_ds$MVR_PTS <=0, 1, 0)

eval_ds$CAR_AGE_FLAG_BIN <- ifelse(eval_ds$CAR_AGE <=1, 1, 0)

eval_ds$TRAVTIME_FLAG_BIN <- ifelse(eval_ds$TRAVTIME <=20, 1, 0)

new_ds_full <- eval_ds

#eval_ds <- select(eval_ds, -JOB_Blue.Collar, -JOB_Clerical, -JOB_Doctor, -JOB_Home.Maker, -JOB_Lawyer,

## Create Variables for Linear Regression

eval_ds$AGE_sin <- sin(eval_ds$AGE)
```

```
eval_ds$TIF_sin <- sin(eval_ds$TIF)


eval_ds$BLUEBOOK_sin <- sin(eval_ds$BLUEBOOK)


eval_ds$CAR_TYPE_AMT_BIN <- ifelse(eval_ds$CAR_TYPE_Van | eval_ds$CAR_TYPE_Panel.Truck, 1, 0)

eval_ds$EDUCATION_AMT_BIN <- ifelse(eval_ds$EDUCATION_High.School, 1, 0)

eval_ds$JOB_TYPE_AMT_BIN <- ifelse(eval_ds$JOB_Lawyer |  eval_ds$JOB_Professional | eval_ds$JOB_Blue.Co

eval_ds$INCOME_AMT_BIN <- ifelse(eval_ds$INCOME <=125000, 1, 0)

eval_ds$YOJ_AMT_BIN <- ifelse((eval_ds$YOJ>=7 & eval_ds$YOJ<=17), 1, 0)

eval_ds$HOME_VAL_AMT_0_10K_BIN <- ifelse((eval_ds$HOME_VAL>=0 & eval_ds$HOME_VAL<=10000), 1, 0)

eval_ds$HOME_VAL_AMT_60K_400K_BIN <- ifelse((eval_ds$HOME_VAL>=60000 & eval_ds$HOME_VAL<=400000), 1, 0)


eval_ds$OLDCLAIM_AMT_0_2K_BIN <- ifelse((eval_ds$OLDCLAIM>=0 & eval_ds$OLDCLAIM<=2000), 1, 0)

eval_ds$OLDCLAIM_AMT_2K_10K_BIN <- ifelse((eval_ds$OLDCLAIM>=2001 & eval_ds$OLDCLAIM<=10000), 1, 0)


eval_ds$CLM_FREQ_AMT_BIN <- ifelse(eval_ds$CLM_FREQ <4, 1, 0)


eval_ds$MVR_PTS_AMT_BIN <- ifelse(eval_ds$MVR_PTS <=2, 1, 0)


eval_ds$CAR_AGE_AMT_BIN <- ifelse(eval_ds$CAR_AGE <=1, 1, 0)

eval_ds$TRAVTIME_AMT_BIN <- ifelse(eval_ds$TRAVTIME <=20, 1, 0)

eval_ds$KIDSDRIV_AMT_BIN_0  <- ifelse(eval_ds$KIDSDRIV <=0, 1, 0)
eval_ds$KIDSDRIV_AMT_BIN_1  <- ifelse(eval_ds$KIDSDRIV <=1, 1, 0)

eval_ds$HOMEKIDS_AMT_BIN_0  <- ifelse(eval_ds$HOMEKIDS <=0, 1, 0)
eval_ds$HOMEKIDS_AMT_BIN_3  <- ifelse(eval_ds$HOMEKIDS <=3, 1, 0)

eval_ds$YOJ_AMT_BIN_0_AND_9To14  <- ifelse((eval_ds$YOJ ==0 | (eval_ds$YOJ>=9 & eval_ds$YOJ>=14)), 1, 0

eval_ds$INCOME_AMT_BIN_MISS_0  <- ifelse((is.na(eval_ds$INCOME) |  eval_ds$INCOME<=0), 1, 0)

eval_ds$HOME_VAL_AMT_BIN_MISS_0  <- ifelse((is.na(eval_ds$HOME_VAL) |  eval_ds$HOME_VAL<=0), 1, 0)

eval_ds$EDUCATION_AMT_BIN_HS  <- ifelse(eval_ds$EDUCATION_High.School==1, 1, 0)
eval_ds$EDUCATION_AMT_BIN_HS_B  <- ifelse((eval_ds$EDUCATION_Bachelors |  eval_ds$EDUCATION_High.School

eval_ds$JOB_AMT_BIN_CPSB  <- ifelse((eval_ds$JOB_Clerical |  eval_ds$JOB_Blue.Collar | eval_ds$JOB_Profe
```

```
eval_ds$TIF_AMT_BIN_6  <- ifelse(eval_ds$TIF <=6, 1, 0)

eval_ds$CAR_TYPE_AMT_BIN_V_PT_MV  <- ifelse((eval_ds$CAR_TYPE_Van |  eval_ds$CAR_TYPE_Panel.Truck | eval

eval_ds$OLDCLAIM_AMT_BIN_MISS_0  <- ifelse((is.na(eval_ds$OLDCLAIM) |  eval_ds$OLDCLAIM<=0), 1, 0)

eval_ds$CLM_FREQ_AMT_BIN_0  <- ifelse(eval_ds$CLM_FREQ<=0, 1, 0)
eval_ds$CLM_FREQ_AMT_BIN_3  <- ifelse(eval_ds$CLM_FREQ<=3, 1, 0)

eval_ds$MVR_PTS_AMT_BIN_0  <- ifelse(eval_ds$MVR_PTS<=0, 1, 0)
eval_ds$MVR_PTS_AMT_BIN_5  <- ifelse(eval_ds$MVR_PTS<=5, 1, 0)









eval_ds$TARGET_FLAG_prob <- unlist(data.frame(predict(TF_Model1, type ="response", newdata=eval_ds)))
eval_ds$TARGET_FLAG<-ifelse(eval_ds$TARGET_FLAG_prob>0.5,1,0)

# eval_ds$TARGET_FLAG <- ifelse(eval_ds$TARGET_FLAG_prob>0.5, 1, 0)
x <- as.data.frame(table(eval_ds$TARGET_FLAG))

names(x) <- c("Crash Predicted?", "Counts")
# x[1,1] <- FALSE
# x[2,1] <- TRUE

kable(x, caption="Predicted Crash Counts")


###Top 10 Records by Index
#kable(head(eval_ds)[,c(1,2,94)], caption="Logistic Regression Results")


# eval_ds$TARGET_FLAG <- ifelse(eval_ds$TARGET_FLAG_prob>0.5, 1, 0)
#table(eval_ds$TARGET_FLAG)
#kable(eval_ds[eval_ds$TARGET_FLAG_prob="NA",c(1,61,62)], caption="Outcome on evaluation data set")


eval_ds$TARGET_AMT <- 0

eval_ds_TA <- filter(eval_ds, TARGET_FLAG == 1)
eval_ds_TA$TARGET_FLAG<-as.numeric(eval_ds_TA$TARGET_FLAG)

eval_ds_TA$TARGET_AMT <- predict(TA_Model1, newdata=eval_ds_TA)

x<-arrange(eval_ds_TA, (TARGET_AMT))
x<-x[-c(1:2),]
kable(x[1:10,c(1,2,94,3)], caption="Linear Regression Results")
```

```r
Class_Expense<-lm(TARGET_AMT~TARGET_FLAG_prob,data=eval_ds_TA)

plot(eval_ds_TA$TARGET_FLAG_prob,eval_ds_TA$TARGET_AMT)
abline(Class_Expense,col="red")



##
```