CUNY School of Professional Studies

Education That Works As Hard As You Do®

# Program Data Analyst: Energy Management follow up

*Duubar Villalobos Jimenez*

*May 27, 2019*

## Contents


## Requirements

A work sample report should be redacted, reflect your individual work effort, and illustrate your capability as a data analyst. Please include a brief summary that identifies the project goals, methodology, data sample, tools, etc. You are requested to submit the document in PDF format to us no later than noon on Thursday May 30th.

## Summary

This work sample will be created using a tool called R. R Core Team (2016) is a language and environment for statistical computing and graphics that is rich for statistical and data analysis and for sharing results in various forms.

This sample, will encompass a total of two different projects, one involving time series; the other involving a more methodical approach to a given data set.

# Work Samples

## Example 1

### Overview

Example 1 consists of a simple data set of residential power usage for January 1998 until December 2013. The data is given in a single file. The variable *"KWH"* is power consumption in Kilowatt hours, the rest is straight forward.

### Objective

The objective is to model the data and to perform a monthly forecast for 2014.

### Procedure

**First**, let's have a small idea of how the data look like:

| CaseSequence | YYYY-MMM | KWH |
|---|---|---|
| 733 | 1998-Jan | 6862583 |
| 734 | 1998-Feb | 5838198 |
| 735 | 1998-Mar | 5420658 |
| 736 | 1998-Apr | 5010364 |
| 737 | 1998-May | 4665377 |
| 738 | 1998-Jun | 6467147 |

From above, we notice 3 columns as follows:

**CaseSequence**: Indicate the Sequence of the readings.

**YYYY-MMM** : Indicate the date of the reading.

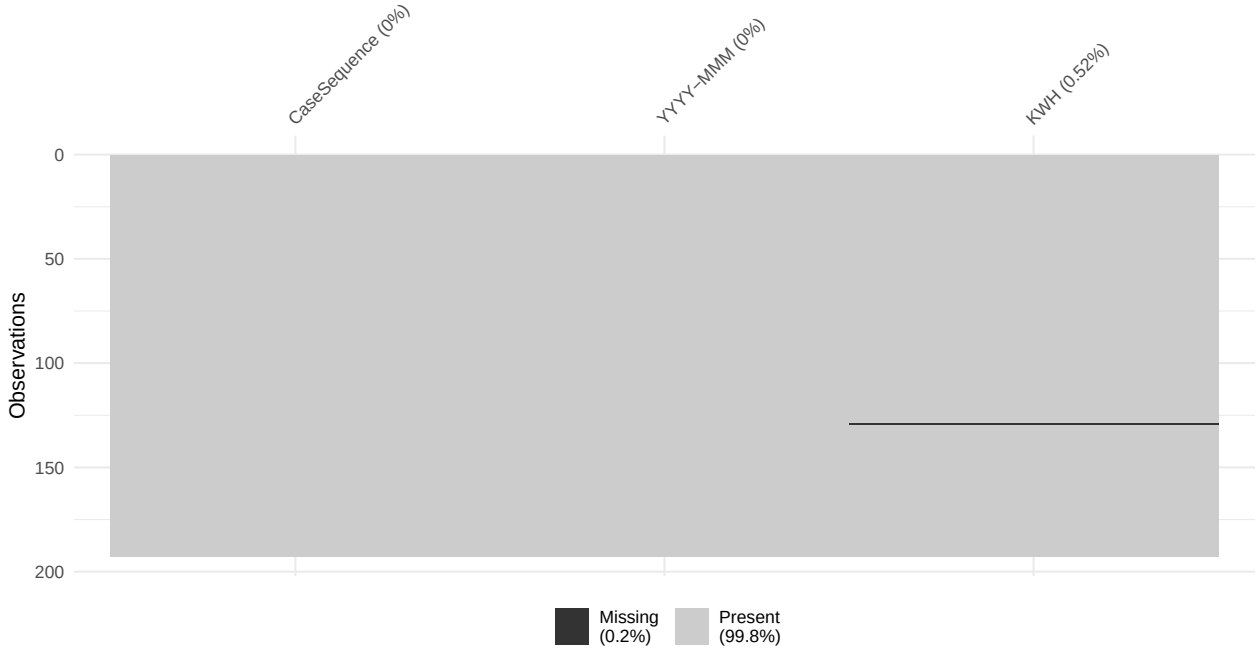
**KWH**: Indicate the value of the reading in KWH.

**Second**, let's have a description of the data:

```
##   CaseSequence     YYYY-MMM              KWH
##  Min.   :733.0   Length:192         Min.   :  770523
##  1st Qu.:780.8   Class :character   1st Qu.: 5429912
##  Median :828.5   Mode  :character   Median : 6283324
##  Mean   :828.5                      Mean   : 6502475
##  3rd Qu.:876.2                      3rd Qu.: 7620524
##  Max.   :924.0                      Max.   :10655730
##                                     NA's   :1
```

From above, there seems to be a missing value as reported in the summary table under NA's.

**Third**, I would like to have a visualization of the missing data since there's an indication of `NA`. For this purpose, I will make use of the function `vis_miss()` from the library `naniar`.




Let's have a better understanding of the missing data.

| CaseSequence | YYYY-MMM | KWH |
|---|---|---|
| 861 | 2008-Sep | NA |

Currently, we are not sure why there's a missing value for the month of September of 2008. At this current point in time I am not sure if I should just remove the missing value or replace it with a more meaningful reading perhaps the mean value for all months representing September. I will come back to this issue as I go further.

**Fourth:** Let's create a time series object.

Let's have a better understanding of the time series.

```
##          Jan      Feb      Mar      Apr      May      Jun      Jul
## 1998 6862583  5838198  5420658  5010364  4665377  6467147  8914755
## 1999 7183759  5759262  4847656  5306592  4426794  5500901  7444416
## 2000 7068296  5876083  4807961  4873080  5050891  7092865  6862662
## 2001 7538529  6602448  5779180  4835210  4787904  6283324  7855129
## 2002 7099063  6413429  5839514  5371604  5439166  5850383  7039702
## 2003 7256079  6190517  6120626  4885643  5296096  6051571  6900676
## 2004 7584596  6560742  6526586  4831688  4878262  6421614  7307931
## 2005 8225477  6564338  5581725  5563071  4453983  5900212  8337998
## 2006 7793358  5914945  5819734  5255988  4740588  7052275  7945564
## 2007 8031295  7928337  6443170  4841979  4862847  5022647  6426220
## 2008 7964293  7597060  6085644  5352359  4608528  6548439  7643987
## 2009 8072330  6976800  5691452  5531616  5264439  5804433  7713260
## 2010 9397357  8390677  7347915  5776131  4919289  6696292   770523
## 2011 8394747  8898062  6356903  5685227  5506308  8037779 10093343
```

```
## 2012  8991267  7952204  6356961  5569828  5783598  7926956  8886851
## 2013 10655730  7681798  6517514  6105359  5940475  7920627  8415321
##           Aug      Sep      Oct      Nov      Dec
## 1998  8607428  6989888  6345620  4640410  4693479
## 1999  7564391  7899368  5358314  4436269  4419229
## 2000  7517830  8912169  5844352  5041769  6220334
## 2001  8450717  7112069  5242535  4461979  5240995
## 2002  8058748  8245227  5865014  4908979  5779958
## 2003  8476499  7791791  5344613  4913707  5756193
## 2004  7309774  6690366  5444948  4824940  5791208
## 2005  7786659  7057213  6694523  4313019  6181548
## 2006  8241110  7296355  5104799  4458429  6226214
## 2007  7447146  7666970  5785964  4907057  6047292
## 2008  8037137       NA  5101803  4555602  6442746
## 2009  8350517  7583146  5566075  5339890  7089880
## 2010  7922701  7819472  5875917  4800733  6152583
## 2011 10308076  8943599  5603920  6154138  8273142
## 2012  9612423  7559148  5576852  5731899  6609694
## 2013  9080226  7968220  5759367  5769083  9606304
```

From the above table, it is evident that we need to replace the NA with a more "meaningful" value, it is not recommended eliminate such value; my approach will be to calculate the mean of all readings for all years for the month of September and replace the NA with such value.

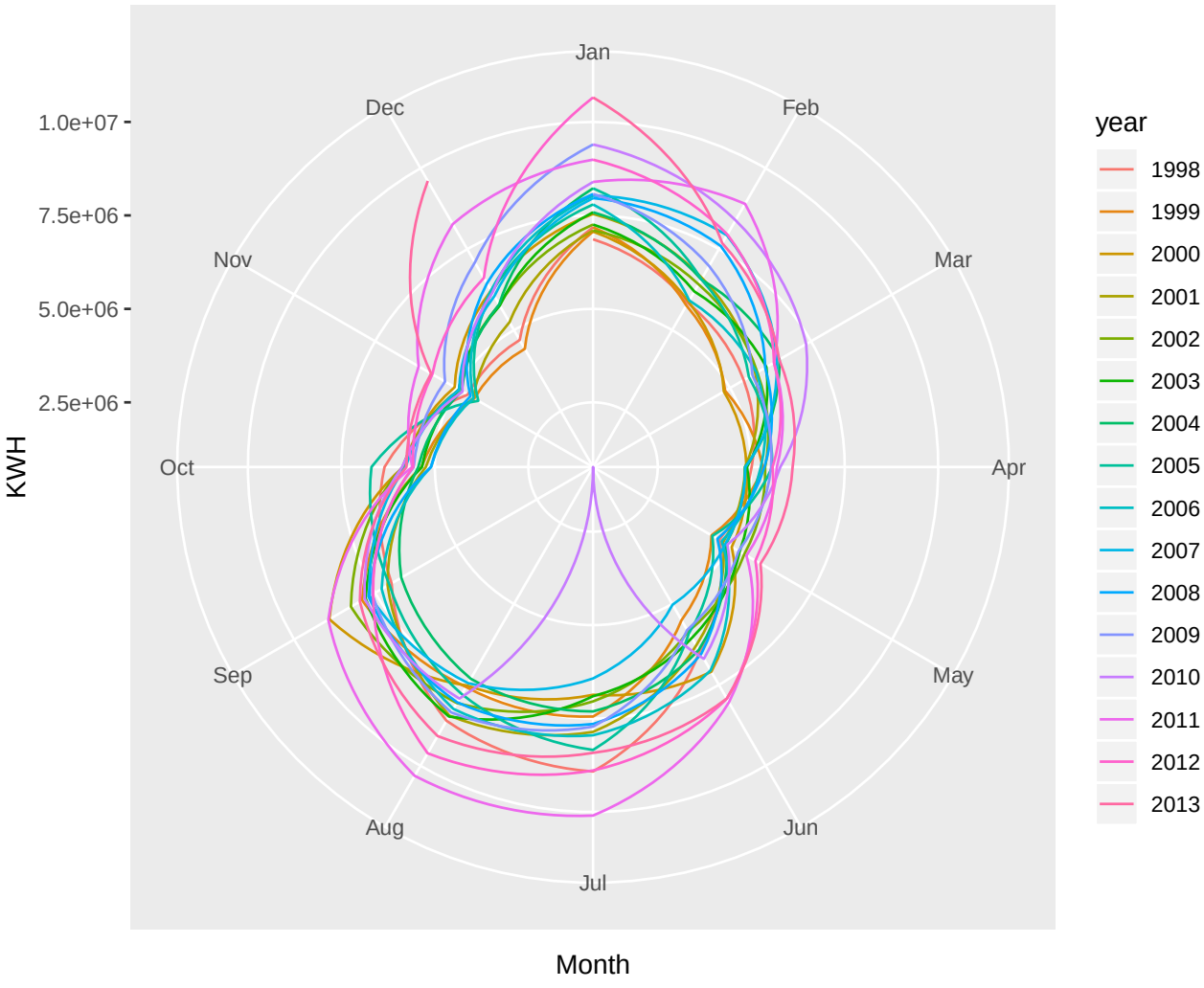
Time series after replacement of missing data with the mean for the respective month, in this case it was for Sep, 2008; it got replaced for 7702333.

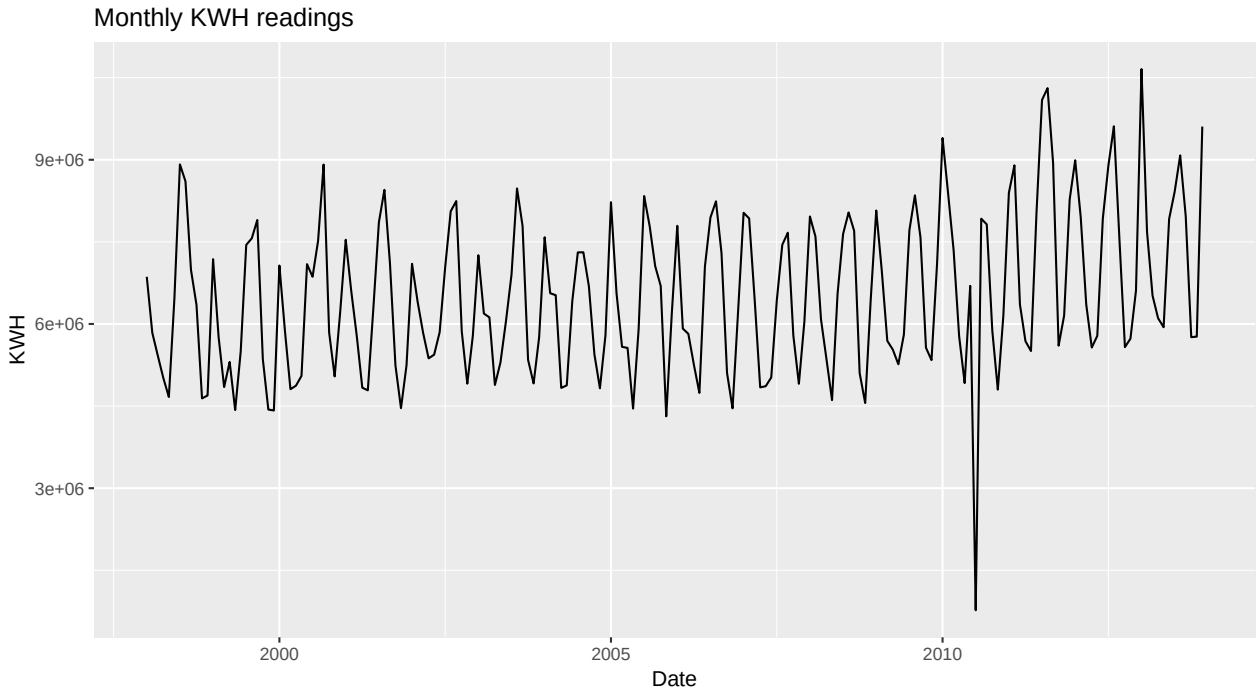
```
##           Jan      Feb      Mar      Apr      May      Jun      Jul
## 1998  6862583  5838198  5420658  5010364  4665377  6467147  8914755
## 1999  7183759  5759262  4847656  5306592  4426794  5500901  7444416
## 2000  7068296  5876083  4807961  4873080  5050891  7092865  6862662
## 2001  7538529  6602448  5779180  4835210  4787904  6283324  7855129
## 2002  7099063  6413429  5839514  5371604  5439166  5850383  7039702
## 2003  7256079  6190517  6120626  4885643  5296096  6051571  6900676
## 2004  7584596  6560742  6526586  4831688  4878262  6421614  7307931
## 2005  8225477  6564338  5581725  5563071  4453983  5900212  8337998
## 2006  7793358  5914945  5819734  5255988  4740588  7052275  7945564
## 2007  8031295  7928337  6443170  4841979  4862847  5022647  6426220
## 2008  7964293  7597060  6085644  5352359  4608528  6548439  7643987
## 2009  8072330  6976800  5691452  5531616  5264439  5804433  7713260
## 2010  9397357  8390677  7347915  5776131  4919289  6696292   770523
## 2011  8394747  8898062  6356903  5685227  5506308  8037779 10093343
## 2012  8991267  7952204  6356961  5569828  5783598  7926956  8886851
## 2013 10655730  7681798  6517514  6105359  5940475  7920627  8415321
##           Aug      Sep      Oct      Nov      Dec
## 1998  8607428  6989888  6345620  4640410  4693479
## 1999  7564391  7899368  5358314  4436269  4419229
## 2000  7517830  8912169  5844352  5041769  6220334
## 2001  8450717  7112069  5242535  4461979  5240995
```

```
## 2002   8058748   8245227   5865014   4908979   5779958
## 2003   8476499   7791791   5344613   4913707   5756193
## 2004   7309774   6690366   5444948   4824940   5791208
## 2005   7786659   7057213   6694523   4313019   6181548
## 2006   8241110   7296355   5104799   4458429   6226214
## 2007   7447146   7666970   5785964   4907057   6047292
## 2008   8037137   7702333   5101803   4555602   6442746
## 2009   8350517   7583146   5566075   5339890   7089880
## 2010   7922701   7819472   5875917   4800733   6152583
## 2011  10308076   8943599   5603920   6154138   8273142
## 2012   9612423   7559148   5576852   5731899   6609694
## 2013   9080226   7968220   5759367   5769083   9606304
```

Let's visualize our data.

Polar seasonal plot: Monthly KWH readings.

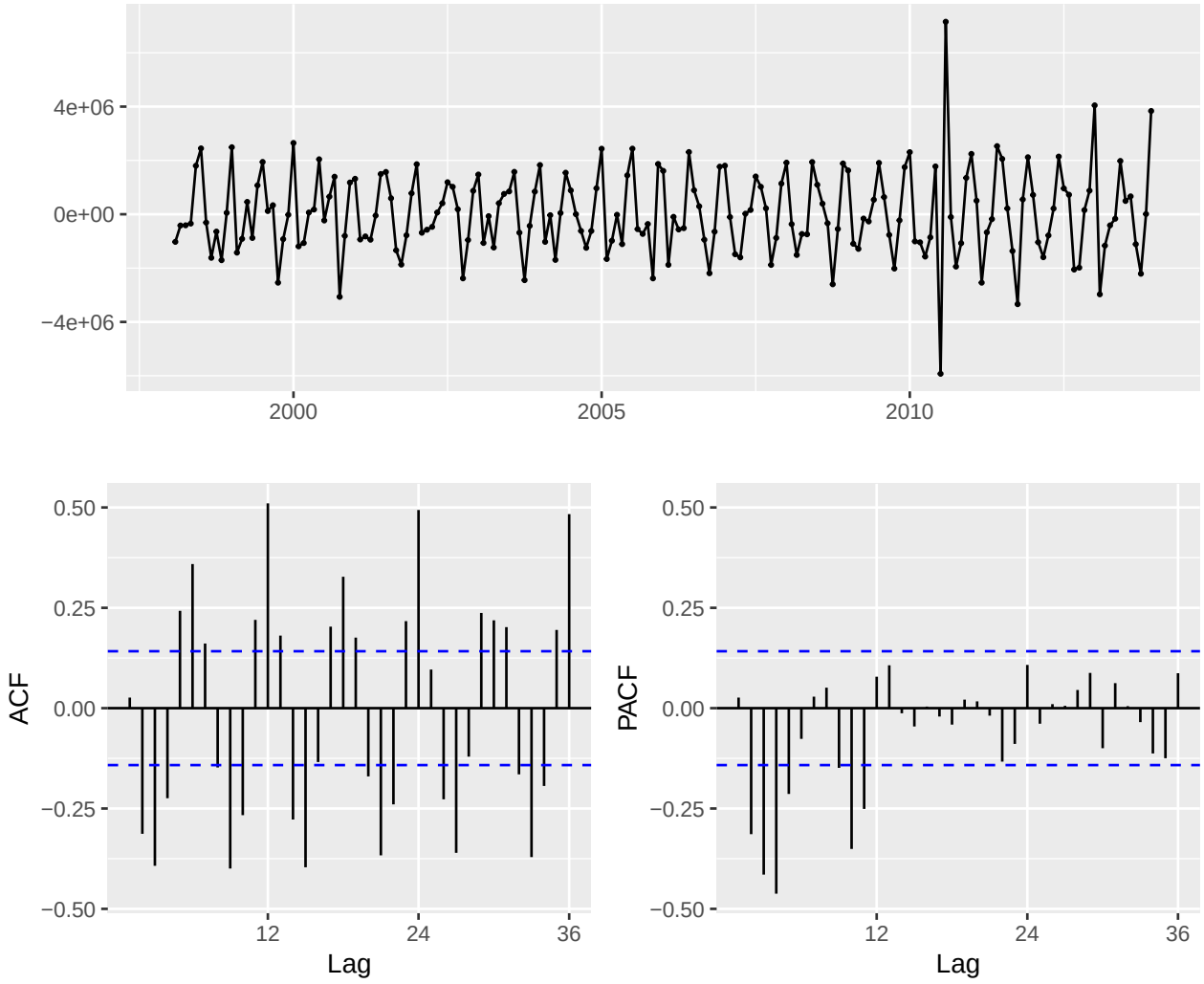





In this particular case, I am not sure as to why there is a very low reading for July, 2010,it is currently showing unusually low (corresponding to some large values in the remainder time series). Some possibilities could point to be a power outage during summer time; this seems to be a good possibility. I did some research and since there's no reference as to the geographical area for the data set, I could not confirm such thing. I will assume this to be the cause, I will consider this to be an accurate reading and I will not change that value.

Also, the data are clearly non-stationary, as the series wanders up and down for some periods. Consequently, we will take a first difference of the data. The difference data are shown below.

Let's have a visualization of the differences.




In the above plot, we notice some auto correlations in the lag, the PACF suggest a AR(3) model. So an initial candidate model is **ARIMA(3,1,0)**.

**Training/test:** In this section, I will split the given data into Train/Test data. This will be used in order to determine the accuracy of the model.

```
power.train <- window(power.ts, end=c(2012,12))
power.test <- window(power.ts, start=c(2013,1))
```

**ARIMA** Let's find an arima model.

**Regular fit.** No transformation, the reason why, is because there seems to be no evidence of changing variance.

```
power.fit.manual.arima <- Arima(power.train, c(3,1,0))
power.fit.auto.arima <- auto.arima(power.train, seasonal=FALSE, stepwise=FALSE, approximation=FA
```

Let's see the results:

**Manual Arima fit**

```
## Series: power.train
```

```
## ARIMA(3,1,0)
##
## Coefficients:
##           ar1      ar2      ar3
##       -0.0852  -0.2971  -0.4079
## s.e.   0.0681   0.0643   0.0678
##
## sigma^2 estimated as 1.707e+12:  log likelihood=-2773.71
## AIC=5555.43   AICc=5555.66   BIC=5568.18
##
## Training set error measures:
##                      ME    RMSE      MAE       MPE     MAPE     MASE
## Training set -185.2833 1292085 954369.9 -6.655492 19.56656 1.381861
##                     ACF1
## Training set -0.1858258
```

**Auto Arima fit**

```
## Series: power.train
## ARIMA(3,1,1) with drift
##
## Coefficients:
##          ar1      ar2      ar3      ma1     drift
##       0.4654  -0.3206  -0.2597  -0.9759  6359.358
## s.e.  0.0782   0.0759   0.0789   0.0527  2604.902
##
## sigma^2 estimated as 1.191e+12:  log likelihood=-2742.1
## AIC=5496.2   AICc=5496.68   BIC=5515.32
##
## Training set error measures:
##                      ME    RMSE      MAE       MPE    MAPE     MASE
## Training set -28105.18 1072783 791147.6 -7.011082 16.8035 1.145527
##                     ACF1
## Training set -0.01324442
```

If we compare both models, we notice how the RMSE value is by far a better value in the Auto Arima model, also, another indication is the AICc value, in this case the Auto Arima model has a better value compared to our manually selected model. Hence, I will pick the Automated Arima model.

**Accuracy** Let's find how accurate the models are:

**Manual Arima(3,1,0)**

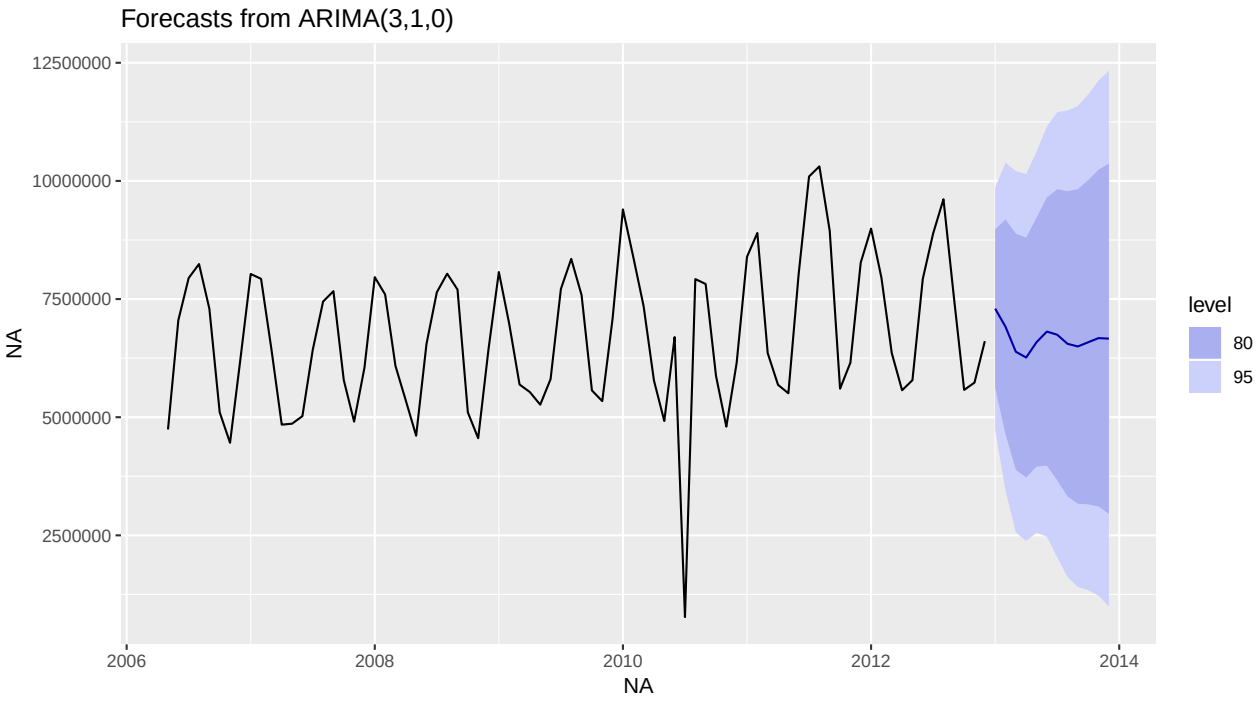
Let's visualize the manually selected Arima(3,1,0) model forecast results:

| | Point.Forecast | Lo.80 | Hi.80 | Lo.95 | Hi.95 |
|---|---|---|---|---|---|
| Jan 2013 | 7297359 | 5622774 | 8971944 | 4736302.5 | 9858415 |
| Feb 2013 | 6914690 | 4645149 | 9184231 | 3443725.9 | 10385654 |
| Mar 2013 | 6384942 | 3885764 | 8884120 | 2562779.0 | 10207105 |
| Apr 2013 | 6263306 | 3724433 | 8802180 | 2380434.5 | 10146178 |
| May 2013 | 6587161 | 3953365 | 9220957 | 2559117.4 | 10615204 |
| Jun 2013 | 6811776 | 3974484 | 9649068 | 2472511.9 | 11151040 |
| Jul 2013 | 6746019 | 3667713 | 9824325 | 2038155.9 | 11453882 |
| Aug 2013 | 6552789 | 3324221 | 9781356 | 1615121.1 | 11490457 |
| Sep 2013 | 6497179 | 3169416 | 9824942 | 1407804.7 | 11586554 |
| Oct 2013 | 6586154 | 3156549 | 10015759 | 1341026.3 | 11831282 |
| Nov 2013 | 6673909 | 3110534 | 10237285 | 1224196.9 | 12123622 |
| Dec 2013 | 6662675 | 2957089 | 10368262 | 995469.7 | 12329881 |

Let's take a look at the accuracy table and let's focus on the RMSE results for the manually selected Arima model. In this particular case, the test set results are not very promising.

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil.s.U |
|---|---|---|---|---|---|---|---|---|
| Training set | -185.2833 | 1292085 | 954369.9 | -6.655492 | 19.56656 | 1.381861 | -0.1858258 | NA |
| Test set | 953505.3038 | 1709364 | 1376213.1 | 9.233251 | 16.48537 | 1.992661 | 0.1195800 | 0.817616 |

Let's have a visualization of the manually selected Arima model forecasts.




In effect, the curve seems not to follow the pattern of the data.

**Autmated Arima(3,1,1)**

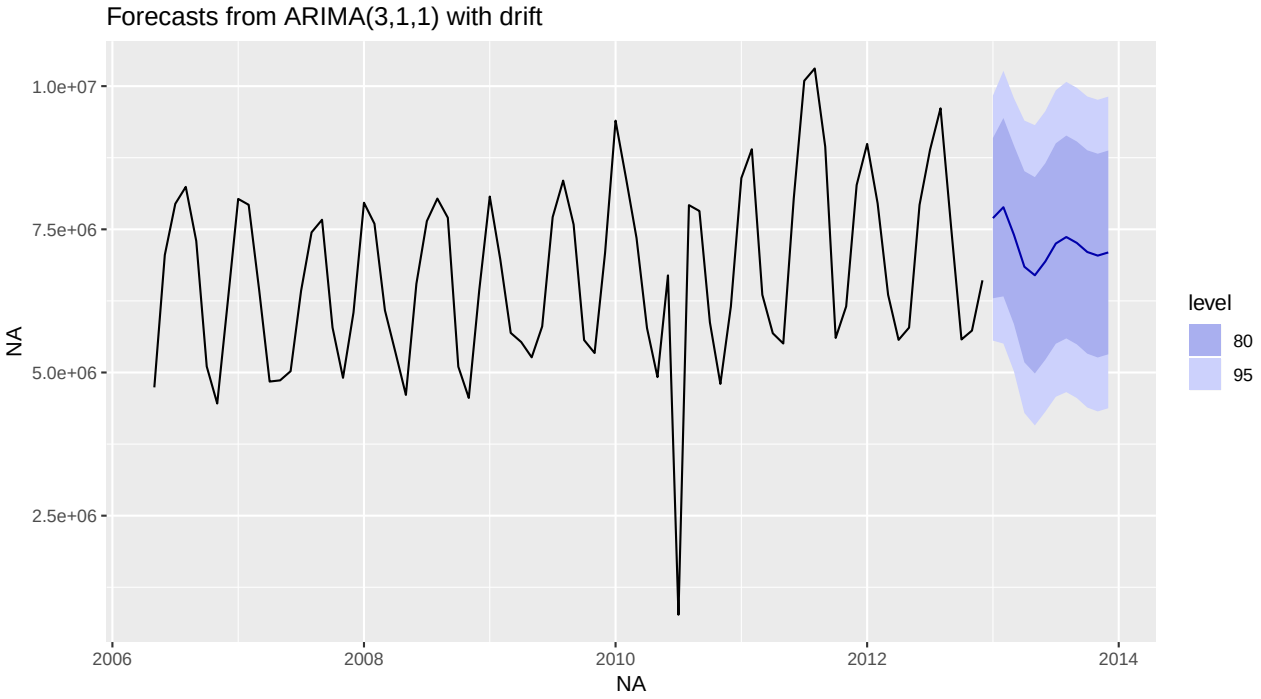
Let's visualize the automatically selected Arima(3,1,1) with drift model forecast results:

| | Point.Forecast | Lo.80 | Hi.80 | Lo.95 | Hi.95 |
|---|---|---|---|---|---|
| Jan 2013 | 7695337 | 6297001 | 9093673 | 5556767 | 9833907 |
| Feb 2013 | 7886045 | 6329165 | 9442925 | 5505002 | 10267088 |
| Mar 2013 | 7405840 | 5845999 | 8965681 | 5020270 | 9791411 |
| Apr 2013 | 6846288 | 5177271 | 8515305 | 4293747 | 9398829 |
| May 2013 | 6697353 | 4983444 | 8411263 | 4076155 | 9318552 |
| Jun 2013 | 6939245 | 5224004 | 8654485 | 4316011 | 9562479 |
| Jul 2013 | 7252012 | 5502568 | 9001456 | 4576469 | 9927556 |
| Aug 2013 | 7365814 | 5595131 | 9136498 | 4657787 | 10073841 |
| Sep 2013 | 7262770 | 5491779 | 9033761 | 4554273 | 9971267 |
| Oct 2013 | 7104174 | 5328558 | 8879790 | 4388604 | 9819744 |
| Nov 2013 | 7040922 | 5262050 | 8819793 | 4320373 | 9761470 |
| Dec 2013 | 7096182 | 5317239 | 8875126 | 4375523 | 9816842 |

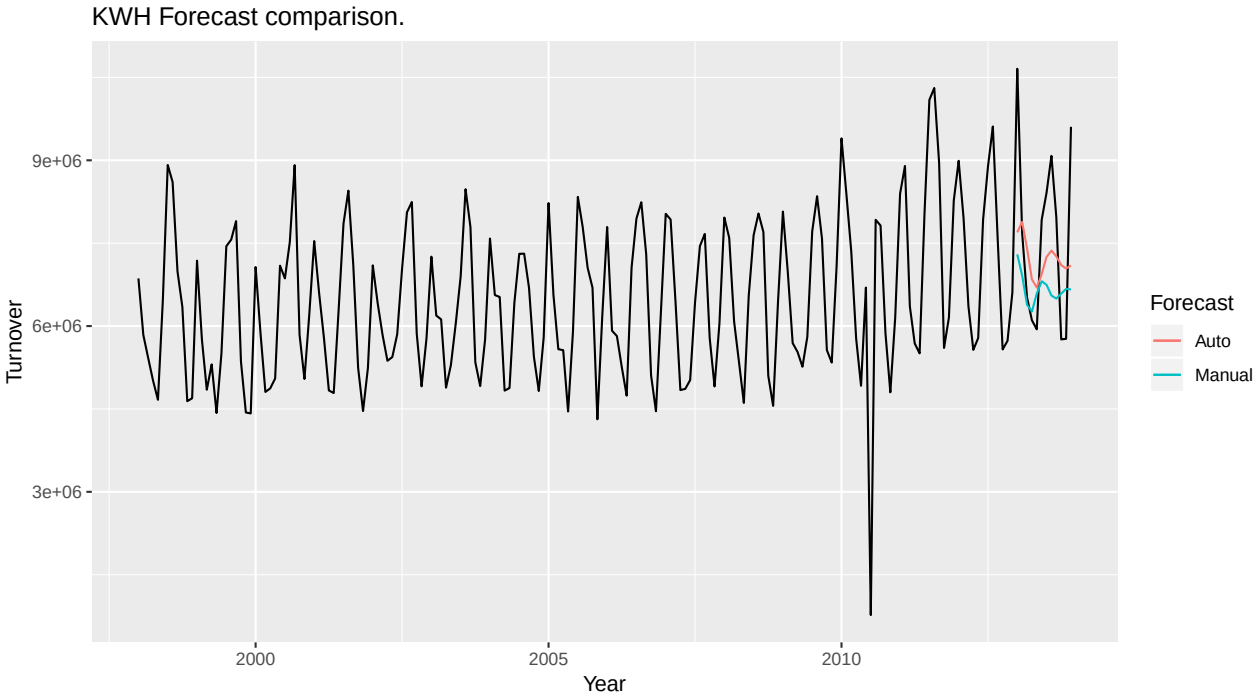
Let's take a look at the accuracy table and let's focus on the RMSE results for the automatically selected Arima model. In this particular case, the test set results are not very promising. Now, comparing the RMSE values to our manually selected model, there's an improvement but still, it seems that the forecasts are not very accurate with this model.

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil.s.U |
|---|---|---|---|---|---|---|---|---|
| Training set | -28105.18 | 1072783 | 791147.6 | -7.011082 | 16.80350 | 1.145527 | -0.0132444 | NA |
| Test set | 402336.77 | 1477517 | 1270174.5 | 1.774910 | 16.20176 | 1.839124 | 0.0851104 | 0.7294665 |

Let's have a visualization of the automatically selected Arima model forecasts.



Let's compare side by side the test forecasts, compared to our test data.




In effect, the forecasts are not very accurate and perhaps another model should be selected.

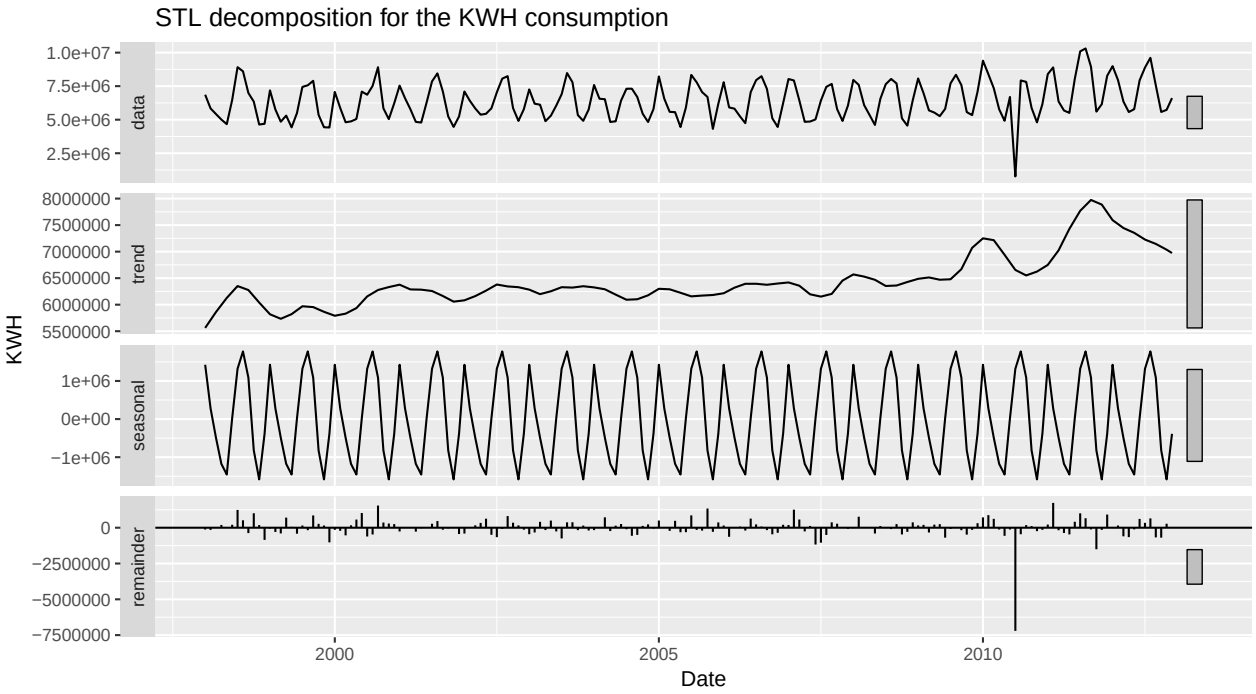
**STL model:** Based on the previous results, I will focus on the STL model.

STL is a versatile and robust method for decomposing time series. STL is an acronym for "Seasonal and Trend decomposition using Loess".

STL has several advantages over the classical, SEATS and X11 decomposition methods:

- Unlike SEATS and X11, STL will handle any type of seasonality, not only monthly and quarterly data.
- The seasonal component is allowed to change over time, and the rate of change can be controlled by the user.
- The smoothness of the trend-cycle can also be controlled by the user.
- It can be robust to outliers (i.e., the user can specify a robust decomposition), so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component.

Let's visualize the STL decomposition.




Let's forecast with the **naive** and **snaive** method.

```
# Calculating forecasts for naive and snaive
power.fit.naive <- forecast(power.fit.stl, method="naive", h =12)
power.fit.snaive <- snaive(power.train[,1], h =12)

# Calculating accuracy
power.accuracy.naive <- accuracy(power.fit.naive, power.test)
power.accuracy.snaive <- accuracy(power.fit.snaive, power.test)
```

Let's see the respective accuracy results for both models.

**Naive** Forecast accuracy results.

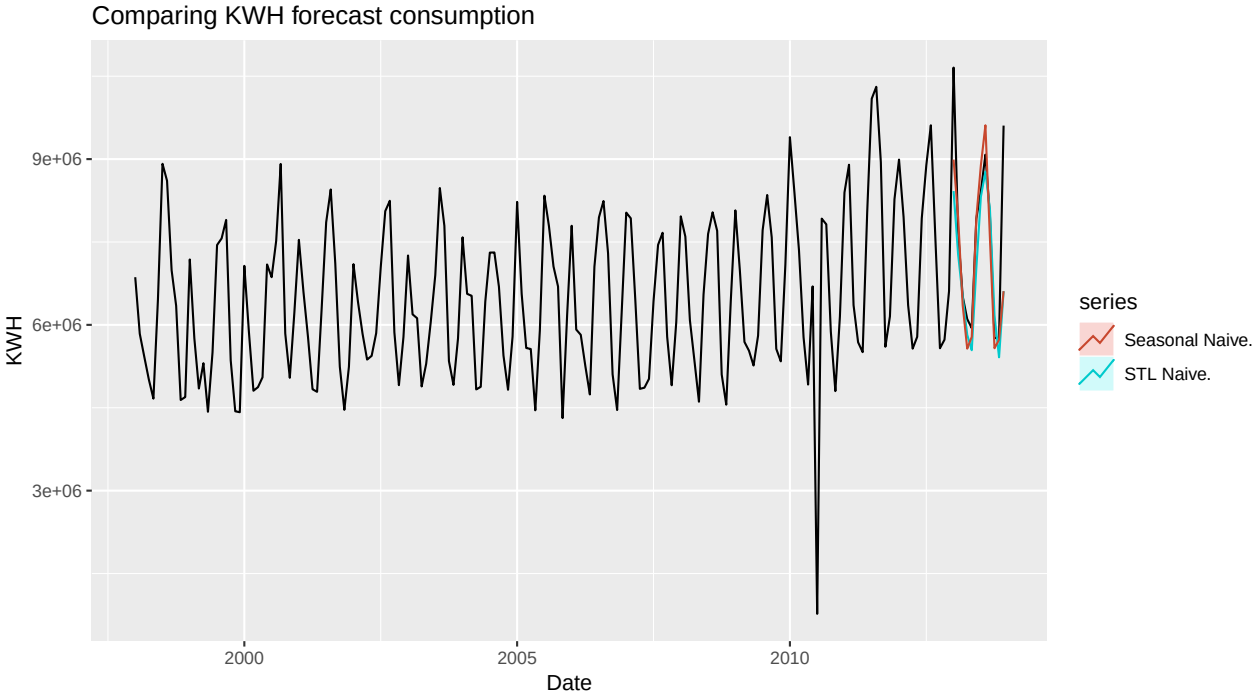
| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil.s.U |
|---|---|---|---|---|---|---|---|---|
| Training set | 8710.895 | 984708.5 | 579079.1 | -5.226505 | 13.670208 | 0.8384662 | -0.3862982 | NA |
| Test set | 621948.397 | 1140409.9 | 710563.5 | 6.869135 | 8.314329 | 1.0288465 | 0.0602833 | 0.6368615 |

**SNaive** Forecast accuracy results.

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil.s.U |
|---|---|---|---|---|---|---|---|---|
| Training set | 72034.37 | 1182065 | 690640.9 | -4.526757 | 14.989184 | 1.0000000 | 0.2550212 | NA |
| Test set | 405195.25 | 1035538 | 618605.6 | 4.547778 | 7.058492 | 0.8956979 | -0.0313027 | 0.6156093 |

In this particular case, the **snaive** method offers a much better RMSE value. Making this model the most accurate of them all.

Let's visualize the results.

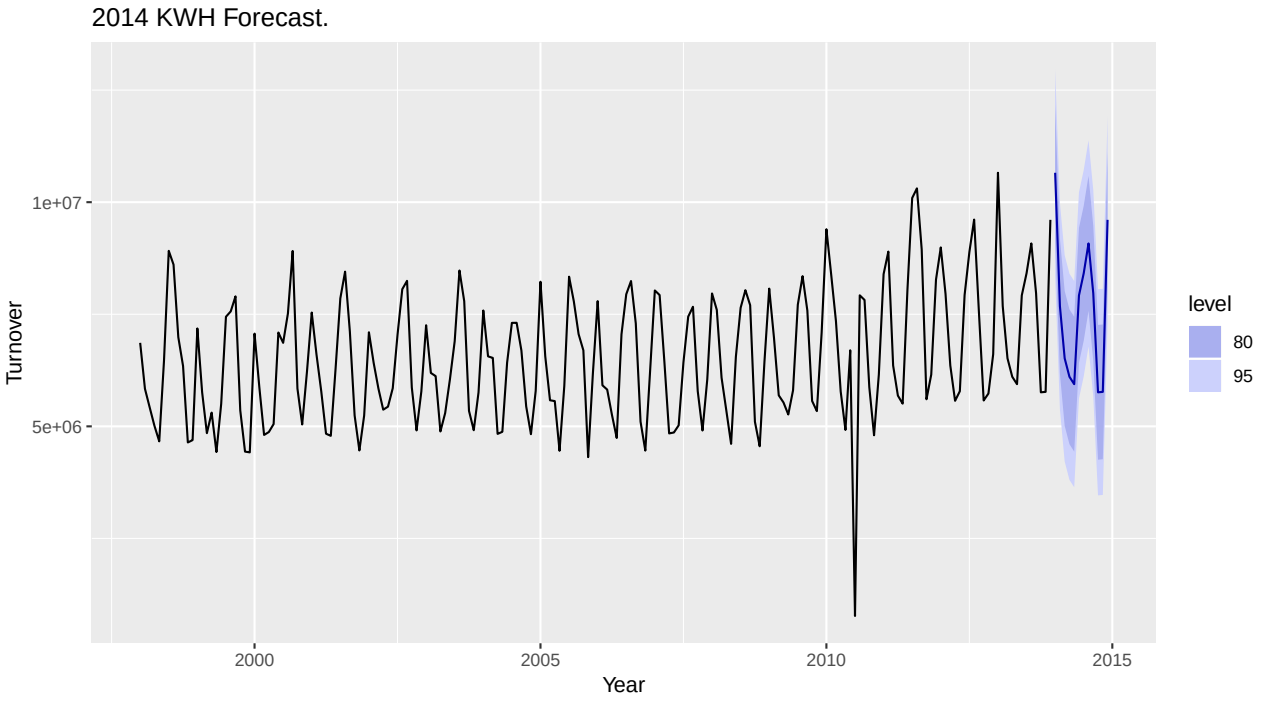



**Forecasting 2014** Employing snaive model.

**Forecast results**

| | Point.Forecast | Lo.80 | Hi.80 | Lo.95 | Hi.95 |
|---|---|---|---|---|---|
| Jan 2014 | 10655730 | 9152641 | 12158819 | 8356954 | 12954506 |
| Feb 2014 | 7681798 | 6178709 | 9184887 | 5383022 | 9980574 |
| Mar 2014 | 6517514 | 5014425 | 8020603 | 4218738 | 8816290 |
| Apr 2014 | 6105359 | 4602270 | 7608448 | 3806583 | 8404135 |
| May 2014 | 5940475 | 4437386 | 7443564 | 3641699 | 8239251 |
| Jun 2014 | 7920627 | 6417538 | 9423716 | 5621851 | 10219403 |
| Jul 2014 | 8415321 | 6912232 | 9918410 | 6116545 | 10714097 |
| Aug 2014 | 9080226 | 7577137 | 10583315 | 6781450 | 11379002 |
| Sep 2014 | 7968220 | 6465131 | 9471309 | 5669444 | 10266996 |
| Oct 2014 | 5759367 | 4256278 | 7262456 | 3460591 | 8058143 |
| Nov 2014 | 5769083 | 4265994 | 7272172 | 3470307 | 8067859 |
| Dec 2014 | 9606304 | 8103215 | 11109393 | 7307528 | 11905080 |

**Forecast visualization**




**Conclusion**

From the above analysis, we could conclude that a good prediction model will be the STL employing the snaive method. Thus due to the similar pattern followed for the testing data and the predicted future values.

## Example 2

### Overview

Example 2 consist as follows: to explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, **TARGET_FLAG**, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is **TARGET_AMT**. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

### Objective

The objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. I can only use the variables given (or variables that I derive from the variables provided).

### Procedure

In this example, I will make use of a version and collaboration tool called github, alongside R. GitHub, a subsidiary of Microsoft, is an American web-based hosting service for version control using Git. It is mostly used for computer code. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features.

It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for every project.

### Dataset description

Let's start by taking a look at the data and it's respective dictionary.

### Variable definitions

The below list represent the definitions for each given variable.

| VARIABLE_NAME | DEFINITION |
|---|---|
| INDEX | Identification Variable (do not use) |
| TARGET_FLAG | Was Car in a crash? 1=YES 0=NO |
| TARGET_AMT | If car was in a crash, what was the cost |
| AGE | Age of Driver |
| BLUEBOOK | Value of Vehicle |
| CAR_AGE | Vehicle Age |
| CAR_TYPE | Type of Car |
| CAR_USE | Vehicle Use |
| CLM_FREQ | # Claims (Past 5 Years) |
| EDUCATION | Max Education Level |
| HOMEKIDS | # Children at Home |
| HOME_VAL | Home Value |
| INCOME | Income |
| JOB | Job Category |
| KIDSDRIV | # Driving Children |
| MSTATUS | Marital Status |
| MVR_PTS | Motor Vehicle Record Points |
| OLDCLAIM | Total Claims (Past 5 Years) |
| PARENT1 | Single Parent |
| RED_CAR | A Red Car |
| REVOKED | License Revoked (Past 7 Years) |
| SEX | Gender |
| TIF | Time in Force |
| TRAVTIME | Distance to Work |
| URBANICITY | Home/Work Area |
| YOJ | Years on Job |

**Theoretical effect of variables**

The below list represent the theoretical effects for each given variable.

| VARIABLE_NAME | THEORETICAL_EFFECT |
|---|---|
| INDEX | None |
| TARGET_FLAG | None |
| TARGET_AMT | None |
| AGE | Very young people tend to be risky. Maybe very old people also. |
| BLUEBOOK | Unknown effect on probability of collision, but probably effect the payout if there is a cra |
| CAR_AGE | Unknown effect on probability of collision, but probably effect the payout if there is a cra |
| CAR_TYPE | Unknown effect on probability of collision, but probably effect the payout if there is a cra |
| CAR_USE | Commercial vehicles are driven more, so might increase probability of collision |
| CLM_FREQ | The more claims you filed in the past, the more you are likely to file in the future |
| EDUCATION | Unknown effect, but in theory more educated people tend to drive more safely |
| HOMEKIDS | Unknown effect |
| HOME_VAL | In theory, home owners tend to drive more responsibly |
| INCOME | In theory, rich people tend to get into fewer crashes |
| JOB | In theory, white collar jobs tend to be safer |
| KIDSDRIV | When teenagers drive your car, you are more likely to get into crashes |
| MSTATUS | In theory, married people drive more safely |
| MVR_PTS | If you get lots of traffic tickets, you tend to get into more crashes |
| OLDCLAIM | If your total payout over the past five years was high, this suggests future payouts will b |
| PARENT1 | Unknown effect |
| RED_CAR | Urban legend says that red cars (especially red sports cars) are more risky. Is that true? |
| REVOKED | If your license was revoked in the past 7 years, you probably are a more risky driver. |
| SEX | Urban legend says that women have less crashes then men. Is that true? |
| TIF | People who have been customers for a long time are usually more safe. |
| TRAVTIME | Long drives to work usually suggest greater risk |
| URBANICITY | Unknown |
| YOJ | People who stay at a job for a long time are usually more safe |

**Data Exploration**

Let's take a look at the hidden layers and some composition of the data set.

**Data acquisition**

For reproducibility purposes, I have included the original data sets in my Git Hub account, I will read it as a data frame from that location.

```
data.train <- get_data(git_user, git_dir, 'insurance_training_data.csv')
data.eval <- get_data(git_user, git_dir, 'insurance-evaluation-data.csv')
```

**General exploration**

The below process will help us obtain insights from our given data.

**Dimensions**

Let's see the dimensions of our training data set.

| Records | Variables |
|---|---|
| 8161 | 26 |

From the above table, we can see how the training data set has a total of 8161 different records and 26 variables including **INDEX, TARGET_FLAG** and **TARGET_AMT**. These variables do not represent much of the initial insights since they correspond to our response variables.

For simplicity reasons, I will discard the **INDEX** column.

```
remove_cols <- names(data.train) %in% c('INDEX')
data.train <- data.train[!remove_cols]
```

**Structure**

The below structure is currently present in the data, for simplicity reasons, I have previously loaded and treated this data set as a data frame in which all the variables with decimals are numeric.

| variable | class | levels |
|---|---|---|
| TARGET_FLAG | integer | NA |
| TARGET_AMT | numeric | NA |
| KIDSDRIV | integer | NA |
| AGE | integer | NA |
| HOMEKIDS | integer | NA |
| YOJ | integer | NA |
| INCOME | Factor w/ 6613 levels | "", "$0", "$1,007", "$1,022", ... |
| PARENT1 | Factor w/ 2 levels | "No", "Yes" |
| HOME_VAL | Factor w/ 5107 levels | "", "$0", "$100,093", "$100,123", ... |
| MSTATUS | Factor w/ 2 levels | "Yes", "z_No" |
| SEX | Factor w/ 2 levels | "M", "z_F" |
| EDUCATION | Factor w/ 5 levels | "<High School", "Bachelors", "Masters", "PhD", ... |
| JOB | Factor w/ 9 levels | "", "Clerical", "Doctor", "Home Maker", ... |
| TRAVTIME | integer | NA |
| CAR_USE | Factor w/ 2 levels | "Commercial", "Private" |
| BLUEBOOK | Factor w/ 2789 levels | "$1,500", "$1,520", "$1,530", "$1,540", ... |
| TIF | integer | NA |
| CAR_TYPE | Factor w/ 6 levels | "Minivan", "Panel Truck", "Pickup", "Sports Car", ... |
| RED_CAR | Factor w/ 2 levels | "no", "yes" |
| OLDCLAIM | Factor w/ 2857 levels | "$0", "$1,000", "$1,008", "$1,011", ... |
| CLM_FREQ | integer | NA |
| REVOKED | Factor w/ 2 levels | "No", "Yes" |
| MVR_PTS | integer | NA |
| CAR_AGE | integer | NA |
| URBANICITY | Factor w/ 2 levels | "Highly Urban/ Urban", "z_Highly Rural/ Rural" |

From the above table, we can notice how we need to take care of certain strings that are seeing as factors but in reality they are representing numbers and should not be seeing as factors. This will be addressed in more detail as we advance.

**Summaries**

Let's find some summary statistics about our given data, for that; I will get a little bit more insights for all the columns including the **TARGET_FLAG** and **TARGET_AMT** variables.

### Combined Summary

In this section, we will explore the combined results as introductory insights.

```
data.train.summary <- get_df_summary(data.train)
data.train.summary
```

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max | Other |
|---|---|---|---|---|---|---|---|
| TARGET_FLAG | 0 | 0 | 0 | 0.2638 | 1 | 1 | |
| TARGET_AMT | 0 | 0 | 0 | 1504 | 1036 | 107586 | |
| KIDSDRIV | 0 | 0 | 0 | 0.1711 | 0 | 4 | |
| AGE | 16 | 39 | 45 | 44.79 | 51 | 81 | NA's :6 |
| HOMEKIDS | 0 | 0 | 0 | 0.7212 | 1 | 5 | |
| YOJ | 0 | 9 | 11 | 10.5 | 13 | 23 | NA's :454 |
| INCOME | | | | | | | (Other) :7086 |
| PARENT1 | | | | | | | |
| HOME_VAL | | | | | | | (Other) :5391 |
| MSTATUS | | | | | | | |
| SEX | | | | | | | |
| EDUCATION | | | | | | | |
| JOB | | | | | | | (Other) :1413 |
| TRAVTIME | 5 | 22 | 33 | 33.49 | 44 | 142 | |
| CAR_USE | | | | | | | |
| BLUEBOOK | | | | | | | (Other):7843 |
| TIF | 1 | 1 | 4 | 5.351 | 7 | 25 | |
| CAR_TYPE | | | | | | | |
| RED_CAR | | | | | | | |
| OLDCLAIM | | | | | | | (Other):3134 |
| CLM_FREQ | 0 | 0 | 0 | 0.7986 | 2 | 5 | |
| REVOKED | | | | | | | |
| MVR_PTS | 0 | 0 | 1 | 1.696 | 3 | 13 | |
| CAR_AGE | -3 | 1 | 8 | 8.328 | 12 | 28 | NA's :510 |
| URBANICITY | | | | | | | |

Please note that this is for introductory insights and should not be considered as complete results.

### TARGET_FLAG Summaries

In the mean time I will split the data into two data-sets depending on the **TARGET_FLAG**. Let's see some summaries for each group.

```
TARGET_FLAG_0 <- data.train[data.train$TARGET_FLAG == 0,]
TARGET_FLAG_1 <- data.train[data.train$TARGET_FLAG == 1,]
```

### Number of records by group

The below table shows how many records each group has.

| | Records | Percentage |
|---|---|---|
| TARGET_FLAG_0 | 6008 | 73.62 |
| TARGET_FLAG_1 | 2153 | 26.38 |
| TOTAL | 8161 | 100.00 |

**TARGET_FLAG = 0**

Let's have a better look at the individualized summaries by having TARGET_FLAG = 0.

```
data.train.summary_0 <- get_df_summary(TARGET_FLAG_0)
data.train.summary_0
```

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max | Other |
|---|---|---|---|---|---|---|---|
| TARGET_FLAG | 0 | 0 | 0 | 0 | 0 | 0 | |
| TARGET_AMT | 0 | 0 | 0 | 0 | 0 | 0 | |
| KIDSDRIV | 0 | 0 | 0 | 0.1393 | 0 | 4 | |
| AGE | 16 | 40 | 46 | 45.32 | 51 | 81 | NA's :1 |
| HOMEKIDS | 0 | 0 | 0 | 0.644 | 1 | 5 | |
| YOJ | 0 | 9 | 12 | 10.67 | 13 | 23 | NA's :331 |
| INCOME | | | | | | | (Other) :5294 |
| PARENT1 | | | | | | | |
| HOME_VAL | | | | | | | (Other) :4205 |
| MSTATUS | | | | | | | |
| SEX | | | | | | | |
| EDUCATION | | | | | | | |
| JOB | | | | | | | (Other) :1053 |
| TRAVTIME | 5 | 22 | 32 | 33.03 | 43 | 142 | |
| CAR_USE | | | | | | | |
| BLUEBOOK | | | | | | | (Other):5789 |
| TIF | 1 | 1 | 6 | 5.556 | 8 | 25 | |
| CAR_TYPE | | | | | | | |
| RED_CAR | | | | | | | |
| OLDCLAIM | | | | | | | (Other):1885 |
| CLM_FREQ | 0 | 0 | 0 | 0.6486 | 1 | 5 | |
| REVOKED | | | | | | | |
| MVR_PTS | 0 | 0 | 1 | 1.414 | 2 | 11 | |
| CAR_AGE | 0 | 4 | 9 | 8.671 | 13 | 28 | NA's :368 |
| URBANICITY | | | | | | | |

**TARGET_FLAG = 1**

Let's have a better look at the individualized summaries by having TARGET_FLAG = 1.

```
data.train.summary_1 <- get_df_summary(TARGET_FLAG_1)
data.train.summary_1
```

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max | Other |
|---|---|---|---|---|---|---|---|
| TARGET_FLAG | 1 | 1 | 1 | 1 | 1 | 1 | |
| TARGET_AMT | 30.28 | 2609.78 | 4104 | 5702.18 | 5787 | 107586.14 | |
| KIDSDRIV | 0 | 0 | 0 | 0.2596 | 0 | 4 | |
| AGE | 16 | 37 | 43 | 43.3 | 50 | 76 | NA's :5 |
| HOMEKIDS | 0 | 0 | 0 | 0.9368 | 2 | 5 | |
| YOJ | 0 | 9 | 11 | 10.02 | 13 | 19 | NA's :123 |
| INCOME | | | | | | | (Other) :1781 |
| PARENT1 | | | | | | | |
| HOME_VAL | | | | | | | (Other) :1178 |
| MSTATUS | | | | | | | |
| SEX | | | | | | | |
| EDUCATION | | | | | | | |
| JOB | | | | | | | (Other) :302 |
| TRAVTIME | 5 | 24 | 34 | 34.77 | 45 | 97 | |
| CAR_USE | | | | | | | |
| BLUEBOOK | | | | | | | (Other):2040 |
| TIF | 1 | 1 | 4 | 4.781 | 7 | 21 | |
| CAR_TYPE | | | | | | | |
| RED_CAR | | | | | | | |
| OLDCLAIM | | | | | | | (Other):1244 |
| CLM_FREQ | 0 | 0 | 1 | 1.217 | 2 | 5 | |
| REVOKED | | | | | | | |
| MVR_PTS | 0 | 0 | 2 | 2.482 | 4 | 13 | |
| CAR_AGE | -3 | 1 | 7 | 7.367 | 11 | 25 | NA's :142 |
| URBANICITY | | | | | | | |

From the above reports, we can notice how we need to "transform" our data in order to make it more workable.

In order to do so, I will start to prep the data. Graphical visualizations and correlations will be provided later on since we need to address a few things in order to make our data more workable.

**Findings**

From the above tables, is interesting to note as follows:

- The training data-set shows the presence of missing values or **NAs** in some columns; that can be seeing in the **Other** column. This will be addressed as we prepare our data down the road.
- **"(Other)"** means that there are factor values that could not be grouped accordingly.
- Interesting to see that **CAR_AGE** shows a minimum value of -3. This needs to be investigated since it seems that it's not accurate.
- The Maximum value for **TARGET_AMT** seems to be very far away from the mean and the median value. This needs to be evaluated and find out if this is accurate.

**Data Preparation**

In this section, I will prepare our given data-set. For that I will need to address a few things, like factors and missing data.

**Data conversion**

In this section, I will describe the conversion of the data that is required in order to have a more manageable understanding of it.

**FACTOR to NUMERIC**

This section explains the conversions of currency values in which the system interpreted as factors when in reality these should have been treated as numeric type.

The variables that need to be converted are: **INCOME, HOME_VAL, BLUEBOOK** and **OLDCLAIM**.

**FACTOR to "Dummy"**

In this section, I will transform the remaining factor variables into various binary "Dummy" variables with values 1 for "Yes" and 0 for "No".

The variables that need to be converted are: **PARENT1, MSTATUS, SEX, EDUCATION, JOB, CAR_USE, CAR_TYPE, RED_CAR, REVOKED** and **URBANICITY**.

Please note that there will be a a new set of variables summarizing the above data as follows:

- **IS_SINGLE_PARENT** will represent if **PARENT1** = "Yes" with a value of 1; 0 otherwise.
- **IS_MARRIED** will represent if **MSTATUS** = "Yes"" with a value of 1; 0 otherwise.
- **IS_FEMALE** will represent if **SEX** = "z_F" with a value of 1; 0 otherwise.
- **EDUCATION** will represent diverse **EDUCATION** levels, "<High School" is the default and this column was not included.
- **JOB** will represent diverse **JOB** levels, "Blank" is the default and this column was not included.
- **IS_CAR_PRIVATE_USE** will represent if **CAR_USE** = "Private" with a value of 1; 0 otherwise.
- **CAR_TYPE** will represent diverse **CAR_TYPE** levels, "Minivan" is the default and this column was not included.
- **IS_CAR_RED** will represent if **RED_CAR** = "Yes" with a value of 1; 0 otherwise.
- **IS_LIC_REVOKED** will represent if **REVOKED** = "Yes" with a value of 1; 0 otherwise.
- **IS_URBAN** will represent if **URBANICITY** = "Highly Urban/ Urban" with a value of 1; 0 otherwise.

Let's see our resulting table.

| Column_Names |
|---|
| TARGET_FLAG |
| TARGET_AMT |
| KIDSDRIV |
| AGE |
| HOMEKIDS |
| YOJ |
| INCOME |
| HOME_VAL |
| TRAVTIME |
| BLUEBOOK |
| TIF |
| OLDCLAIM |
| CLM_FREQ |
| MVR_PTS |
| CAR_AGE |
| IS_SINGLE_PARENT |
| IS_MARRIED |
| IS_FEMALE |
| EDUCATIONBachelors |
| EDUCATIONMasters |
| EDUCATIONPhD |
| EDUCATIONz_High.School |
| JOBClerical |
| JOBDoctor |
| JOBHome.Maker |
| JOBLawyer |
| JOBManager |
| JOBProfessional |
| JOBStudent |
| JOBz_Blue.Collar |
| IS_CAR_PRIVATE_USE |
| CAR_TYPEPanel.Truck |
| CAR_TYPEPickup |
| CAR_TYPESports.Car |
| CAR_TYPEVan |
| CAR_TYPEz_SUV |
| IS_CAR_RED |
| IS_LIC_REVOKED |
| IS_URBAN |

**NAs prep**

First, let's see how our values are represented after the above transformations.

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max | Other |
|---|---|---|---|---|---|---|---|
| TARGET_FLAG | 0 | 0 | 0 | 2.63800e-01 | 1 | 1 | |
| TARGET_AMT | 0 | 0 | 0 | 1.50400e+03 | 1036 | 107586 | |
| KIDSDRIV | 0 | 0 | 0 | 1.71100e-01 | 0 | 4 | |
| AGE | 16 | 39 | 45 | 4.47900e+01 | 51 | 81 | NA's :6 |
| HOMEKIDS | 0 | 0 | 0 | 7.21200e-01 | 1 | 5 | |
| YOJ | 0 | 9 | 11 | 1.05000e+01 | 13 | 23 | NA's :454 |
| INCOME | 0 | 28097 | 54028 | 6.18980e+04 | 85986 | 367030 | NA's :445 |
| HOME_VAL | 0 | 0 | 161160 | 1.54867e+05 | 238724 | 885282 | NA's :464 |
| TRAVTIME | 5 | 22 | 33 | 3.34900e+01 | 44 | 142 | |
| BLUEBOOK | 1500 | 9280 | 14440 | 1.57100e+04 | 20850 | 69740 | |
| TIF | 1 | 1 | 4 | 5.35100e+00 | 7 | 25 | |
| OLDCLAIM | 0 | 0 | 0 | 4.03700e+03 | 4636 | 57037 | |
| CLM_FREQ | 0 | 0 | 0 | 7.98600e-01 | 2 | 5 | |
| MVR_PTS | 0 | 0 | 1 | 1.69600e+00 | 3 | 13 | |
| CAR_AGE | -3 | 1 | 8 | 8.32800e+00 | 12 | 28 | NA's :510 |
| IS_SINGLE_PARENT | 0 | 0 | 0 | 1.32000e-01 | 0 | 1 | |
| IS_MARRIED | 0 | 0 | 1 | 5.99700e-01 | 1 | 1 | |
| IS_FEMALE | 0 | 0 | 1 | 5.36100e-01 | 1 | 1 | |
| EDUCATIONBachelors | 0 | 0 | 0 | 2.74700e-01 | 1 | 1 | |
| EDUCATIONMasters | 0 | 0 | 0 | 2.03200e-01 | 0 | 1 | |
| EDUCATIONPhD | 0 | 0 | 0 | 8.92000e-02 | 0 | 1 | |
| EDUCATIONz_High.School | 0 | 0 | 0 | 2.85500e-01 | 1 | 1 | |
| JOBClerical | 0 | 0 | 0 | 1.55700e-01 | 0 | 1 | |
| JOBDoctor | 0 | 0 | 0 | 3.01400e-02 | 0 | 1 | |
| JOBHome.Maker | 0 | 0 | 0 | 7.85400e-02 | 0 | 1 | |
| JOBLawyer | 0 | 0 | 0 | 1.02300e-01 | 0 | 1 | |
| JOBManager | 0 | 0 | 0 | 1.21100e-01 | 0 | 1 | |
| JOBProfessional | 0 | 0 | 0 | 1.36900e-01 | 0 | 1 | |
| JOBStudent | 0 | 0 | 0 | 8.72400e-02 | 0 | 1 | |
| JOBz_Blue.Collar | 0 | 0 | 0 | 2.23600e-01 | 0 | 1 | |
| IS_CAR_PRIVATE_USE | 0 | 0 | 1 | 6.28800e-01 | 1 | 1 | |
| CAR_TYPEPanel.Truck | 0 | 0 | 0 | 8.28300e-02 | 0 | 1 | |
| CAR_TYPEPickup | 0 | 0 | 0 | 1.70200e-01 | 0 | 1 | |
| CAR_TYPESports.Car | 0 | 0 | 0 | 1.11100e-01 | 0 | 1 | |
| CAR_TYPEVan | 0 | 0 | 0 | 9.19000e-02 | 0 | 1 | |
| CAR_TYPEz_SUV | 0 | 0 | 0 | 2.81100e-01 | 1 | 1 | |
| IS_CAR_RED | 0 | 0 | 0 | 2.91400e-01 | 1 | 1 | |
| IS_LIC_REVOKED | 0 | 0 | 0 | 1.22500e-01 | 0 | 1 | |
| IS_URBAN | 0 | 1 | 1 | 7.95500e-01 | 1 | 1 | |

In order to work the missing values, I will proceed as follows.

**Proportion findings**

Let's calculate the proportion of missing values in order to determine the best approach for these variables.

The below list display the combined missing percentage values for each variable.

| | % Total missing |
|---|---|
| AGE | 0.07 |
| YOJ | 5.56 |
| INCOME | 5.45 |
| HOME_VAL | 5.69 |
| CAR_AGE | 6.25 |

**NAs by TARGET_FLAG group**

For this, let's see how many records each group has.

| | % Total missing | % TARGET_FLAG = 0 | % TARGET_FLAG = 1 |
|---|---|---|---|
| AGE | 0.07 | 0.02 | 0.23 |
| YOJ | 5.56 | 5.51 | 5.71 |
| INCOME | 5.45 | 5.58 | 5.11 |
| HOME_VAL | 5.69 | 5.71 | 5.62 |
| CAR_AGE | 6.25 | 6.13 | 6.60 |

Since those values are considered low percentages compared to our data set; I will replace the missing NA values with randomly selected values in between the respective Min and Max value with the exception of CAR_AGE since it shows a negative value, hence I will select 0 to be the minimum value for that particular variable.

**New Structure**

In order to visualize our new structure, I will put together the new set of variables with the transformations. Let's see our structure once again, but this time after the transformation of the data.

| variable | class | levels |
|---|---|---|
| TARGET_FLAG | integer | NA |
| TARGET_AMT | numeric | NA |
| KIDSDRIV | integer | NA |
| AGE | integer | NA |
| HOMEKIDS | integer | NA |
| YOJ | integer | NA |
| INCOME | numeric | NA |
| HOME_VAL | numeric | NA |
| TRAVTIME | integer | NA |
| BLUEBOOK | numeric | NA |
| TIF | integer | NA |
| OLDCLAIM | numeric | NA |
| CLM_FREQ | integer | NA |
| MVR_PTS | integer | NA |
| CAR_AGE | integer | NA |
| IS_SINGLE_PARENT | numeric | NA |
| IS_MARRIED | numeric | NA |
| IS_FEMALE | numeric | NA |
| EDUCATIONBachelors | numeric | NA |
| EDUCATIONMasters | numeric | NA |
| EDUCATIONPhD | numeric | NA |
| EDUCATIONz_High.School | numeric | NA |
| JOBClerical | numeric | NA |
| JOBDoctor | numeric | NA |
| JOBHome.Maker | numeric | NA |
| JOBLawyer | numeric | NA |
| JOBManager | numeric | NA |
| JOBProfessional | numeric | NA |
| JOBStudent | numeric | NA |
| JOBz_Blue.Collar | numeric | NA |
| IS_CAR_PRIVATE_USE | numeric | NA |
| CAR_TYPEPanel.Truck | numeric | NA |
| CAR_TYPEPickup | numeric | NA |
| CAR_TYPESports.Car | numeric | NA |
| CAR_TYPEVan | numeric | NA |
| CAR_TYPEz_SUV | numeric | NA |
| IS_CAR_RED | numeric | NA |
| IS_LIC_REVOKED | numeric | NA |
| IS_URBAN | numeric | NA |

**CAR_AGE investigation**

Let's find out why CAR_AGE has a minimum value of $-3$ which seems to be incorrect, for this I will select the records for **CAR_AGE < 0**, with the goal of identifying more possible unrealistic values.

| | Values |
|---|---|
| TARGET_FLAG | 1 |
| TARGET_AMT | 1469 |
| KIDSDRIV | 0 |
| AGE | 47 |
| HOMEKIDS | 0 |
| YOJ | 12 |
| INCOME | 48696 |
| HOME_VAL | 212014 |
| TRAVTIME | 46 |
| BLUEBOOK | 15390 |
| TIF | 4 |
| OLDCLAIM | 33521 |
| CLM_FREQ | 3 |
| MVR_PTS | 1 |
| CAR_AGE | -3 |
| IS_SINGLE_PARENT | 0 |
| IS_MARRIED | 0 |
| IS_FEMALE | 1 |
| EDUCATIONBachelors | 1 |
| EDUCATIONMasters | 0 |
| EDUCATIONPhD | 0 |
| EDUCATIONz_High.School | 0 |
| JOBClerical | 0 |
| JOBDoctor | 0 |
| JOBHome.Maker | 0 |
| JOBLawyer | 0 |
| JOBManager | 0 |
| JOBProfessional | 1 |
| JOBStudent | 0 |
| JOBz_Blue.Collar | 0 |
| IS_CAR_PRIVATE_USE | 1 |
| CAR_TYPEPanel.Truck | 0 |
| CAR_TYPEPickup | 1 |
| CAR_TYPESports.Car | 0 |
| CAR_TYPEVan | 0 |
| CAR_TYPEz_SUV | 0 |
| IS_CAR_RED | 0 |
| IS_LIC_REVOKED | 1 |
| IS_URBAN | 1 |

From the above results, there seems to be no apparent reason as to why this value was entered. A possible reason could be that the person who typed the record, entered a wrong number; it could be probably 3 or 0 or any other value. In order to keep data integrity, I will remove that record from our data set.

**Visualizations**

From previous summary tables, we established that TARGET_FLAG groups have diverse values as means, from which we could start creating some hypothesis such as:

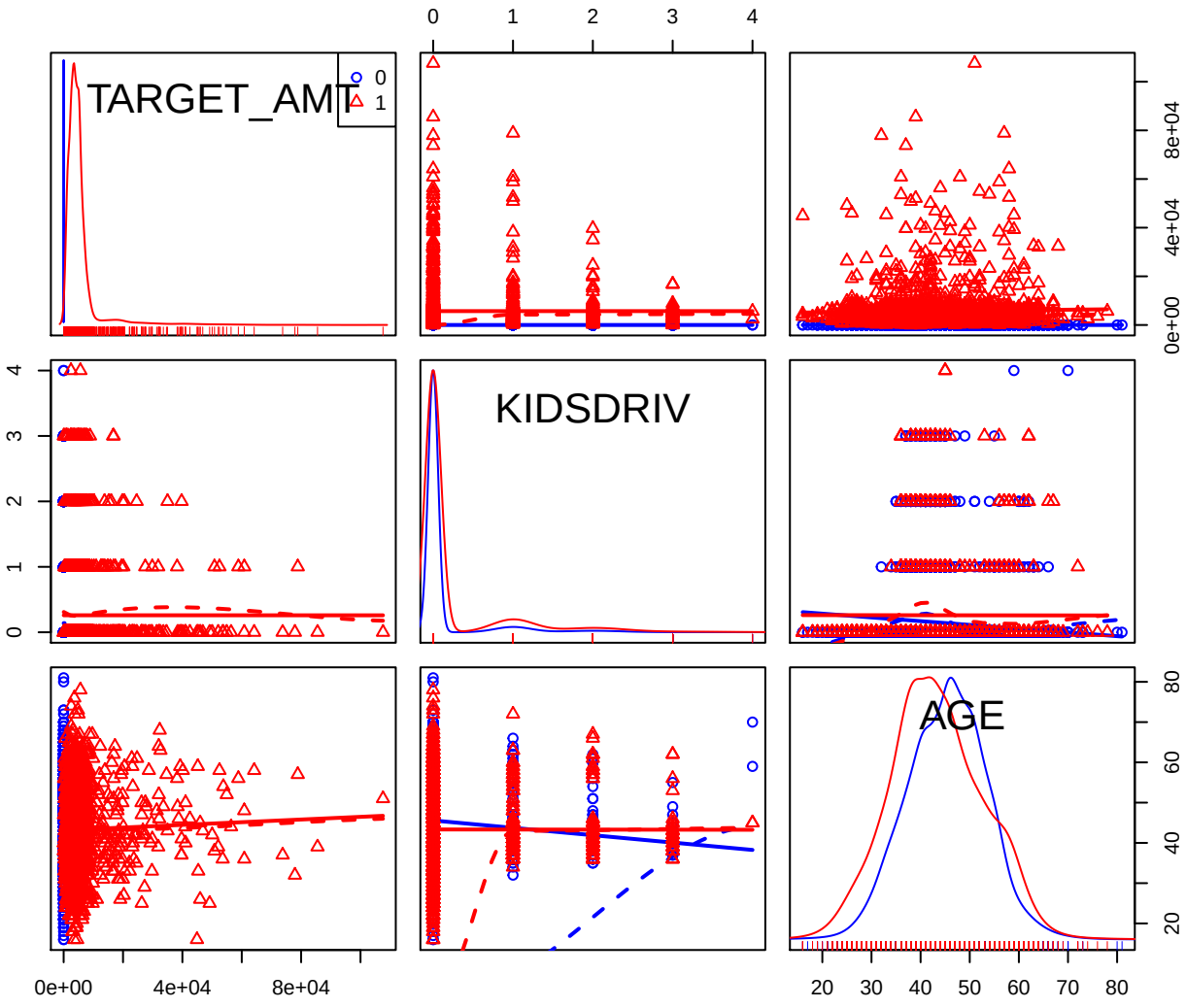
$H_0$: The means for the divided data-set in which TARGET_FLAG = 1 and TARGET_FLAG = 0 are the same.

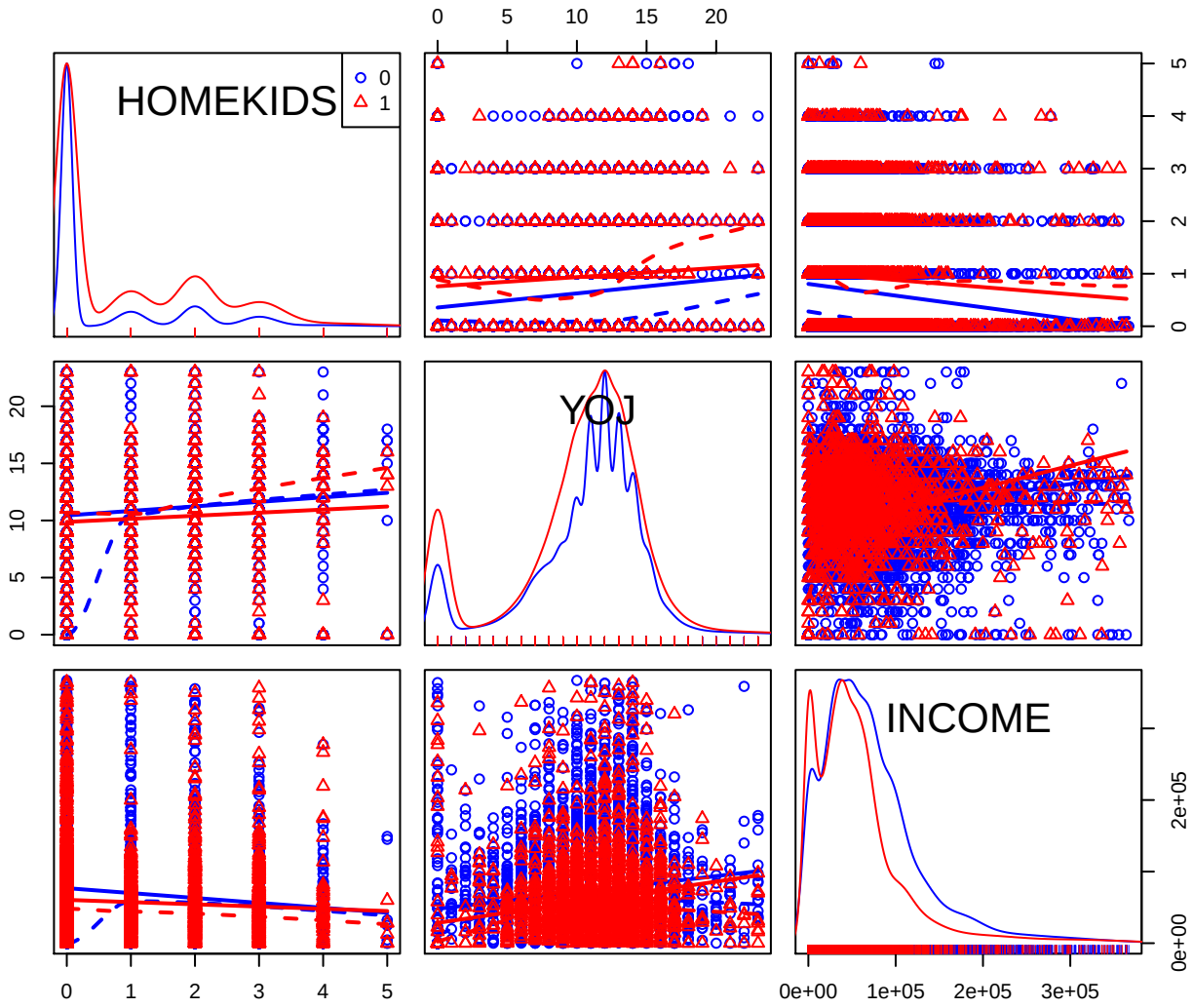
$H_1$: The means for the divided data-set in which TARGET_FLAG = 1 and TARGET_FLAG = 0 are not the same.

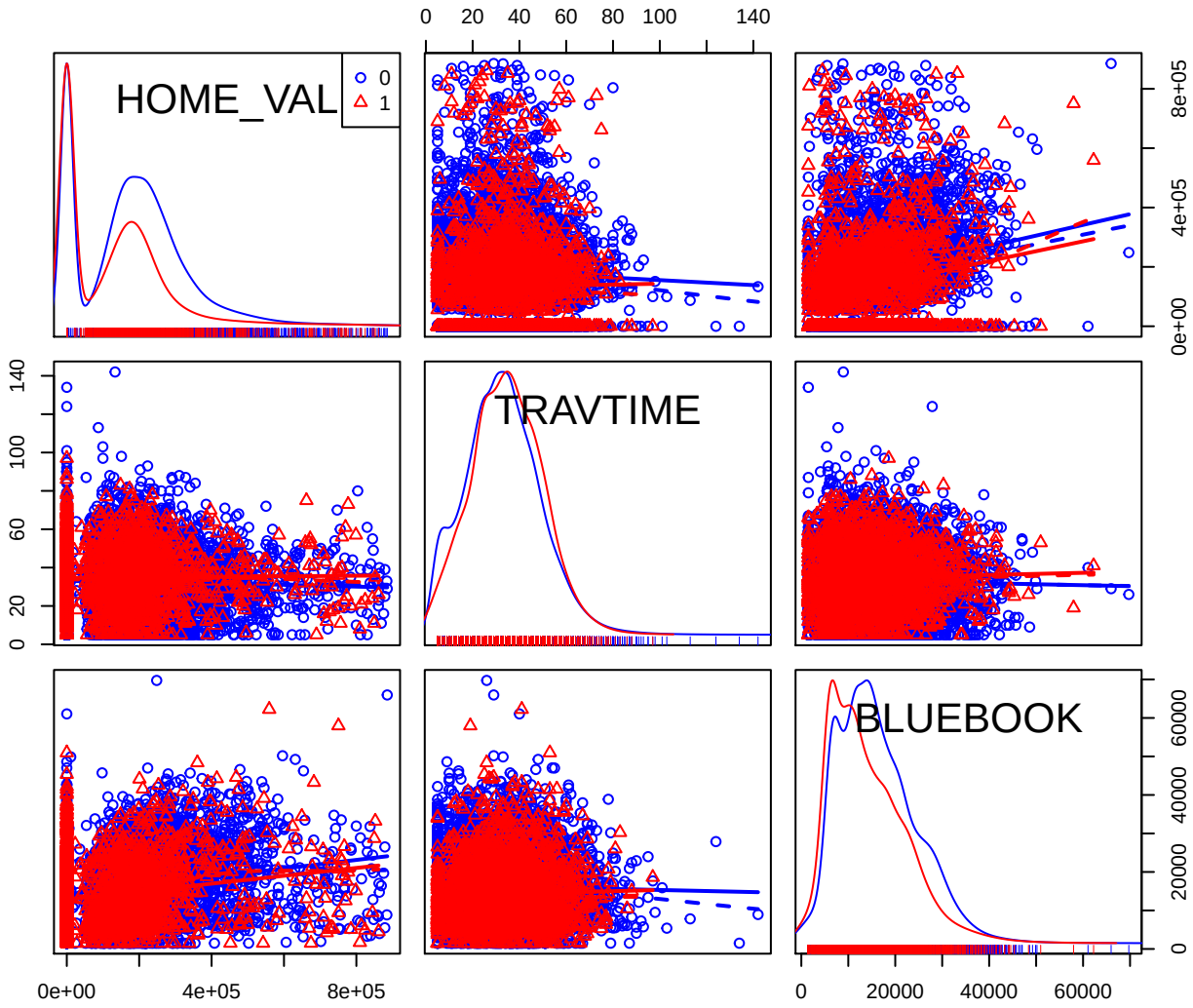
Let's create some visualizations and see how this data behave.

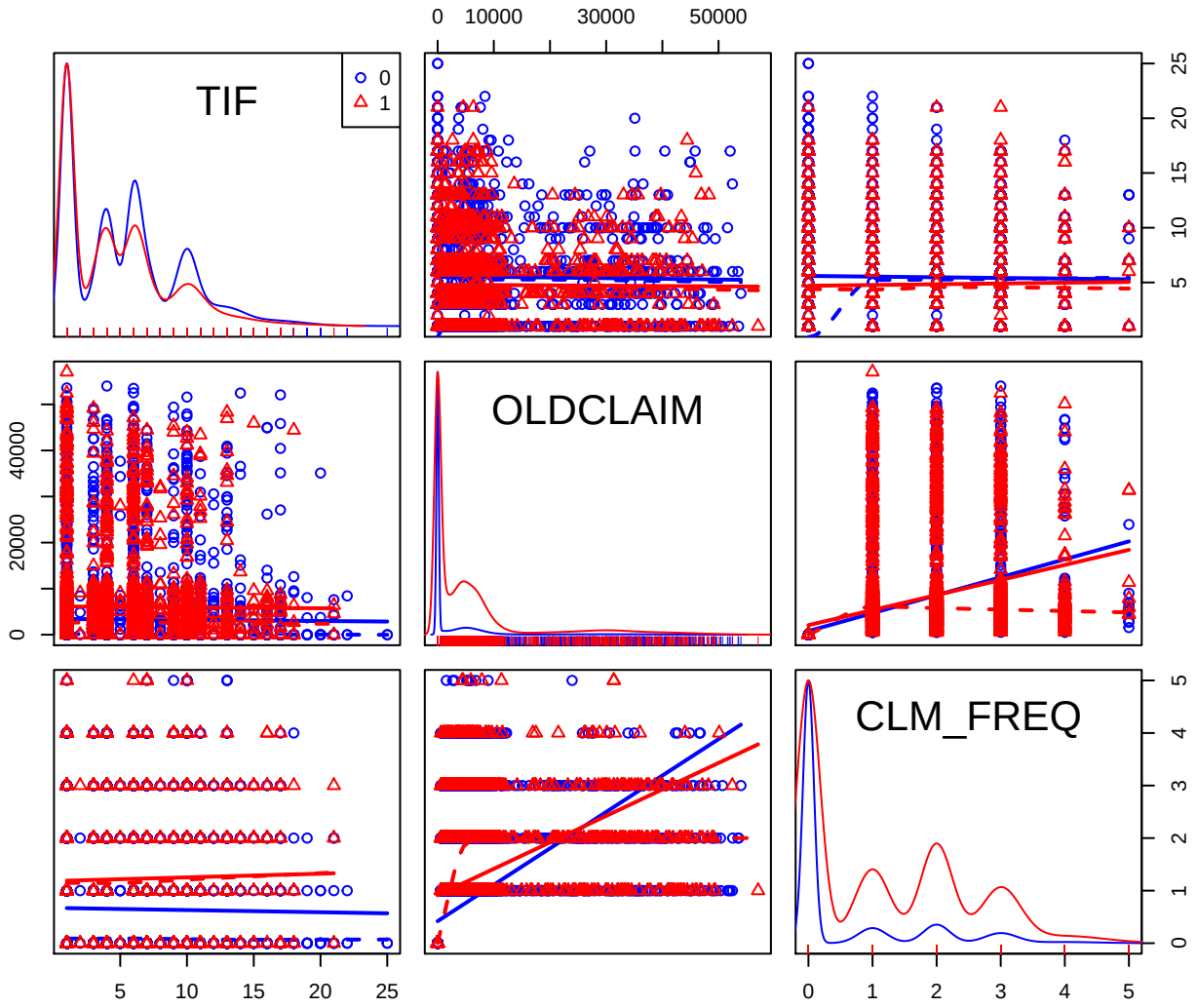
**TARGET_FLAG vs other variables.**

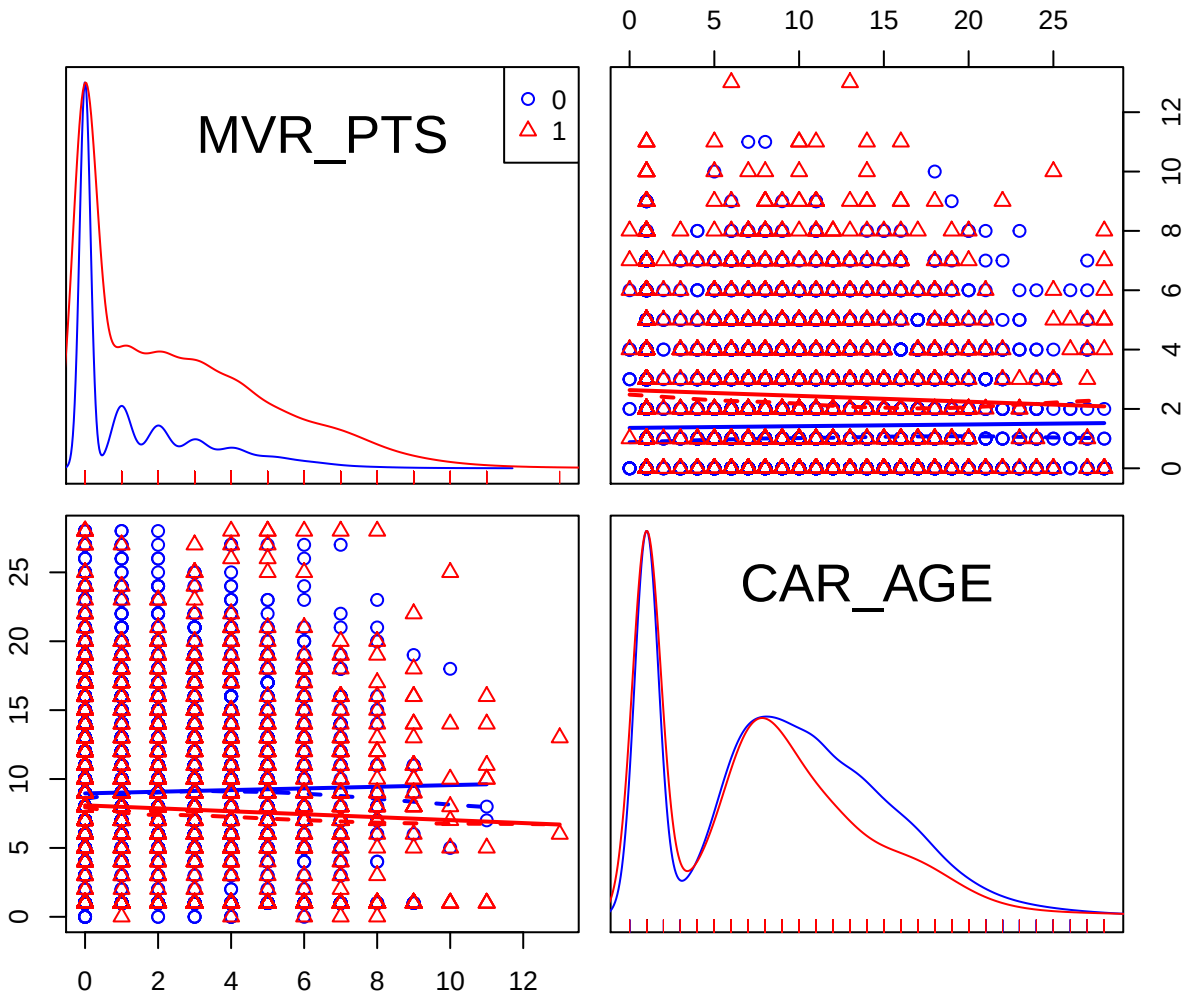
In this case, I will compare our data by separating our **TARGET_FLAG** values. That is, the light-green color represent "0" and the color red represent "1", meaning that red was involved in a Car accident while blue or light green was not.




HOMEKIDS
0
1
YOJ
INCOME
0
5
10
15
20
0
1
2
3
4
5
0e+00
1e+05
2e+05
3e+05

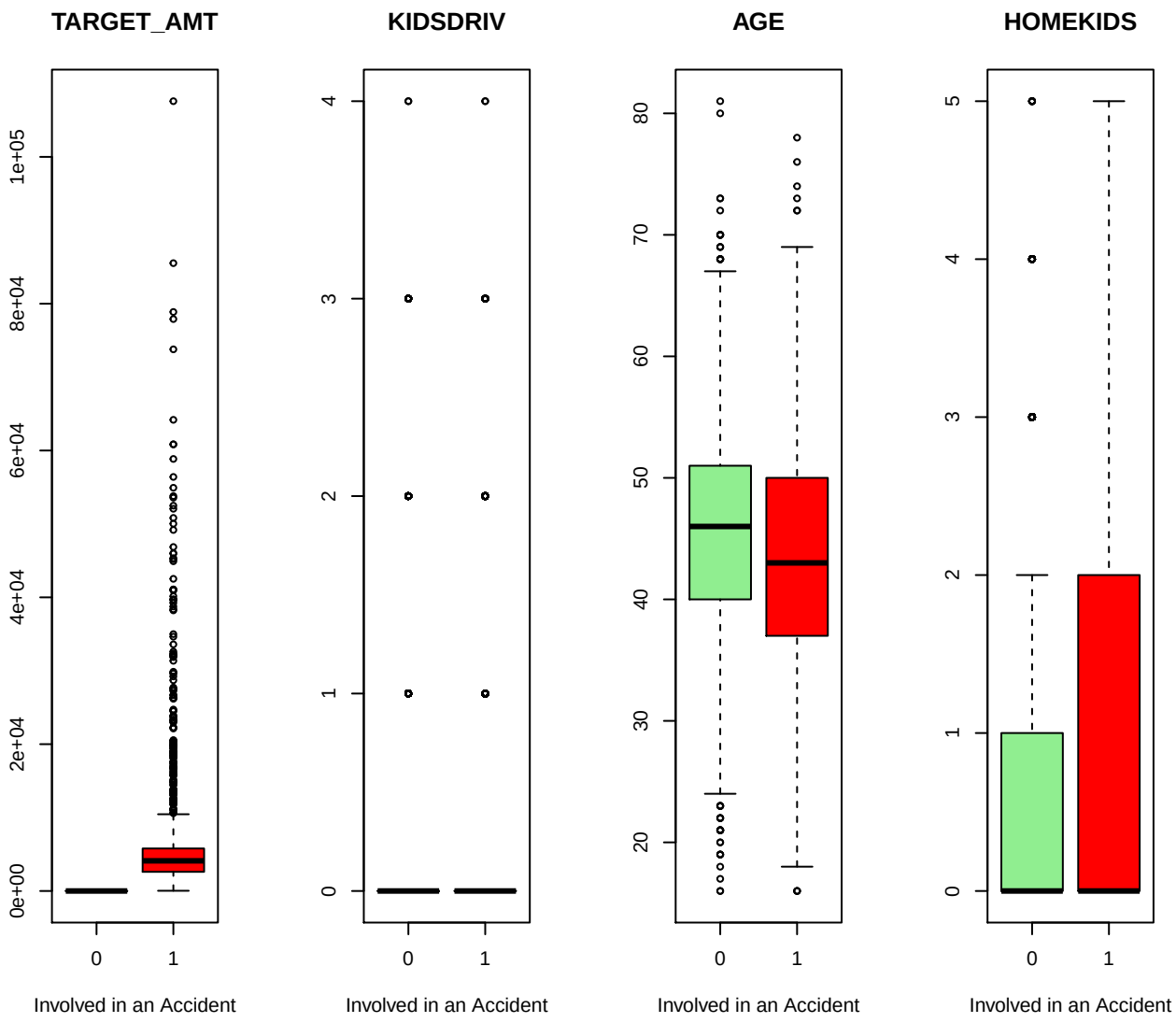

HOME_VAL
TRAVTIME
BLUEBOOK
0
1
0 20 40 60 80 100 140
0e+00 4e+05 8e+05
0 20 60 100 140
0 20000 40000 60000

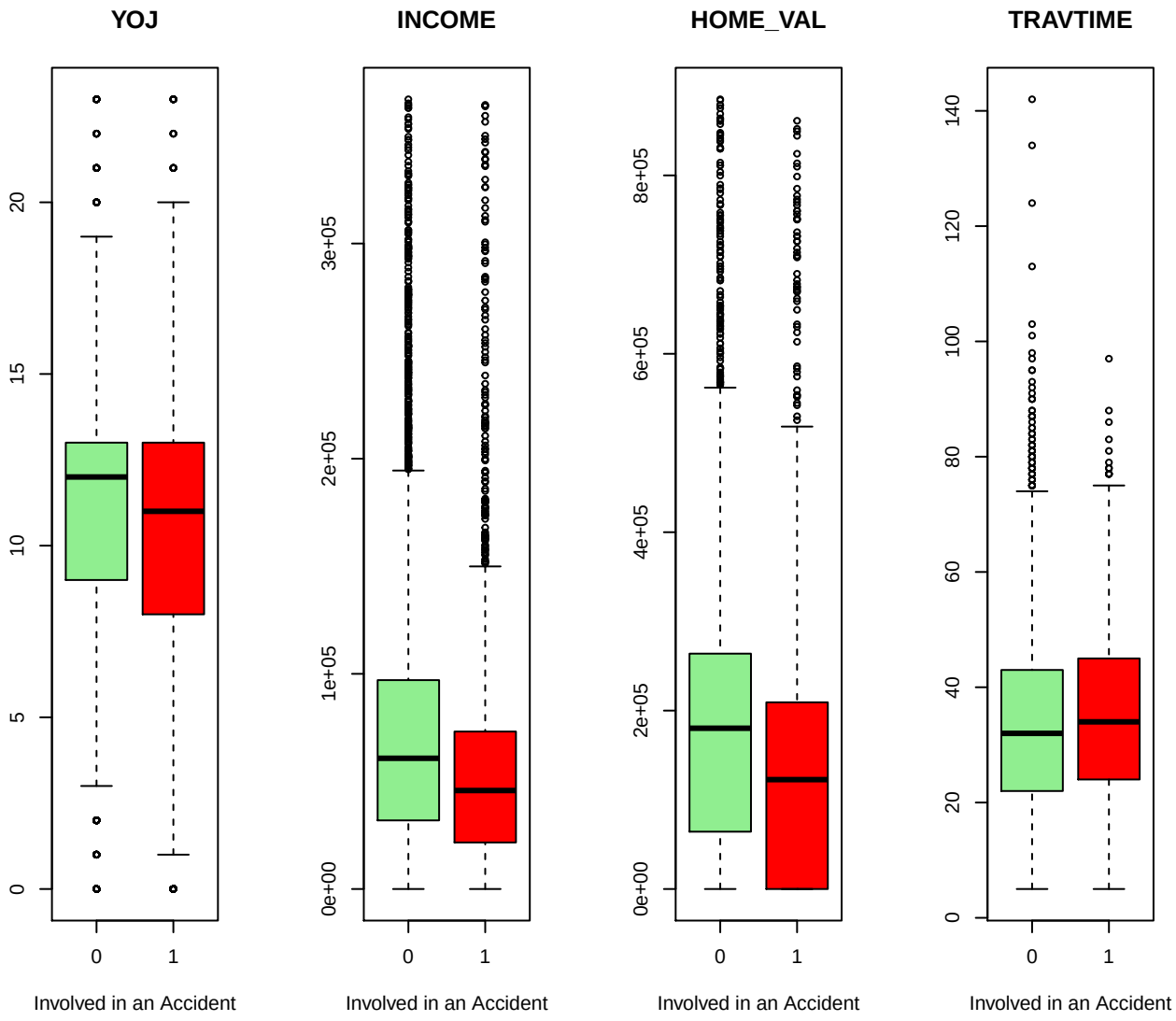

TIF
0
1
OLDCLAIM
CLM_FREQ
0
10000
30000
50000
0
20000
40000
5
10
15
20
25
0
1
2
3
4
5

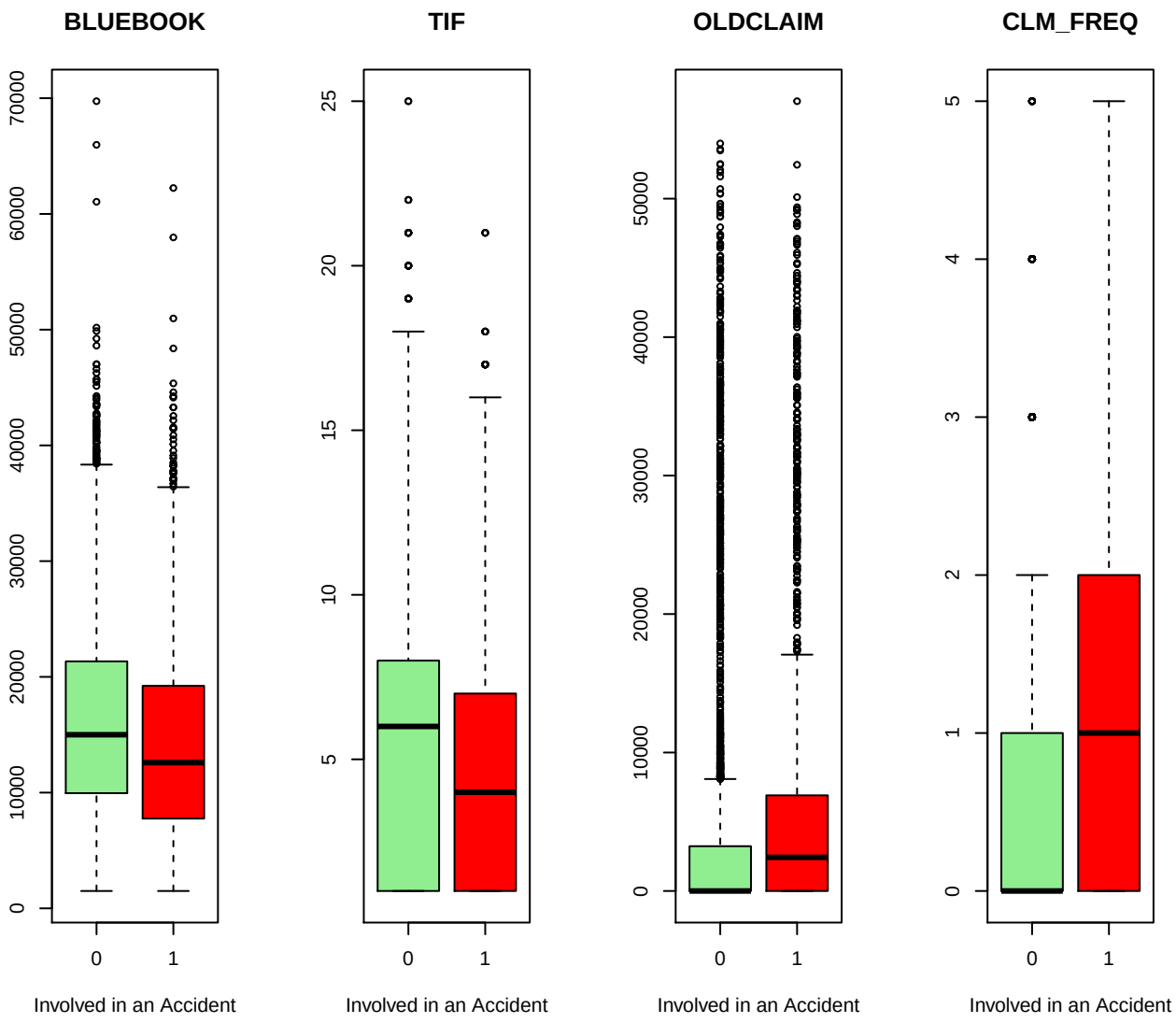


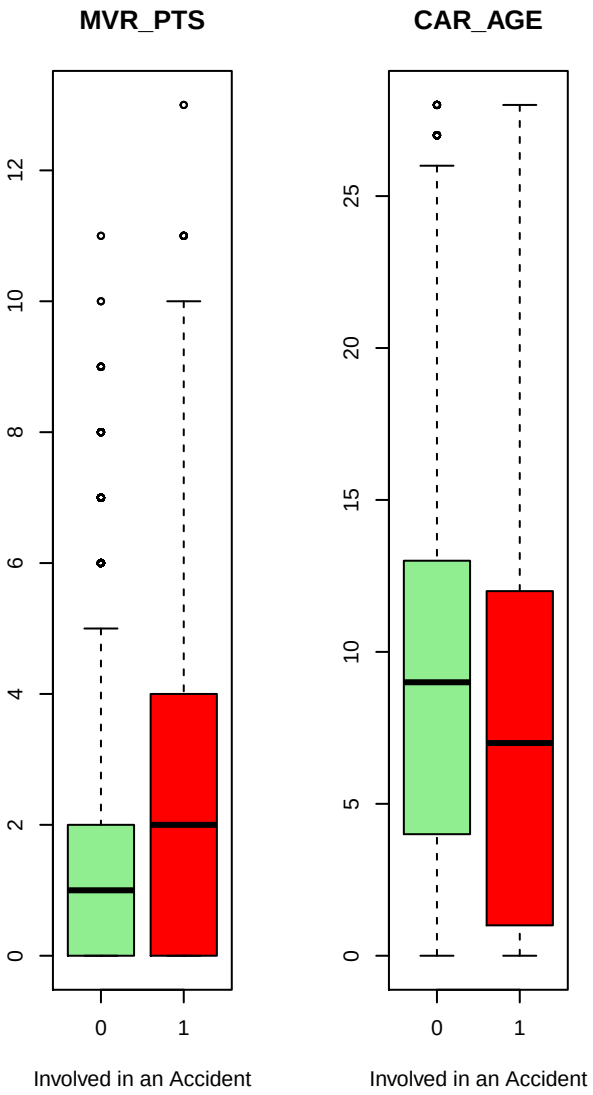

**Box plots**

From the previous visualizations, we can notice how some sort of relationship exist in between some of the variables. In order to create better understanding, let's visualize their behavior by analyzing individual cases.


TARGET_AMT
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
0
1
Involved in an Accident
KIDSDRIV
0
1
2
3
4
0
1
Involved in an Accident
AGE
20
30
40
50
60
70
80
0
1
Involved in an Accident
HOMEKIDS
0
1
2
3
4
5
0
1
Involved in an Accident


YOJ
INCOME
HOME_VAL
TRAVTIME
0
5
10
15
20
0e+00
1e+05
2e+05
3e+05
0e+00
2e+05
4e+05
6e+05
8e+05
0
20
40
60
80
100
120
140
0
1
Involved in an Accident


BLUEBOOK
0
10000
20000
30000
40000
50000
60000
70000
0
1
Involved in an Accident
TIF
5
10
15
20
25
0
1
Involved in an Accident
OLDCLAIM
0
10000
20000
30000
40000
50000
0
1
Involved in an Accident
CLM_FREQ
0
1
2
3
4
5
0
1
Involved in an Accident




Now, If we compare our previous plots and compare them to our given theoretical effect, we can see as follows:

- **AGE**: Definitely AGE seems to be an important factor, we can notice how the mean value on the data-set that had an accident has a considerable age difference, implying that younger people are more risky.
- **BLUEBOOK**: Definitely, the values are considerable different for the means in terms of records having accidents vs the ones who do not. The BLUEBOOK mean value is lower when there's an accident, thus making sense for the data, since the car should have less value after an accident happens.
- **CAR_AGE**: This is very interesting, it seems that newer cars are more involved in accidents than older cars.
- **INCOME**: The provided data agree with the theoretical effect. That is, the mean income value of the people who are involved in accidents is lower than those who are not. From my perspective, this is something very interesting that could be studied in more detail.Perhaps this is a factor for economic growth, lower income individuals tend to have more accidents,

hence limiting their income growth due to repair expenses, fees, fine, loss of time and insurance premiums hikes.

- **MVR_PTS**: The data agree with this theoretical effect, it is noted how the mean of MVR_PTS is higher when accidents are reported.
- **TIF**: This seems to be true, the data report a higher mean for those who do not have an accident.
- **TRAVTIME**: Not much of a difference but the data seems to agree.

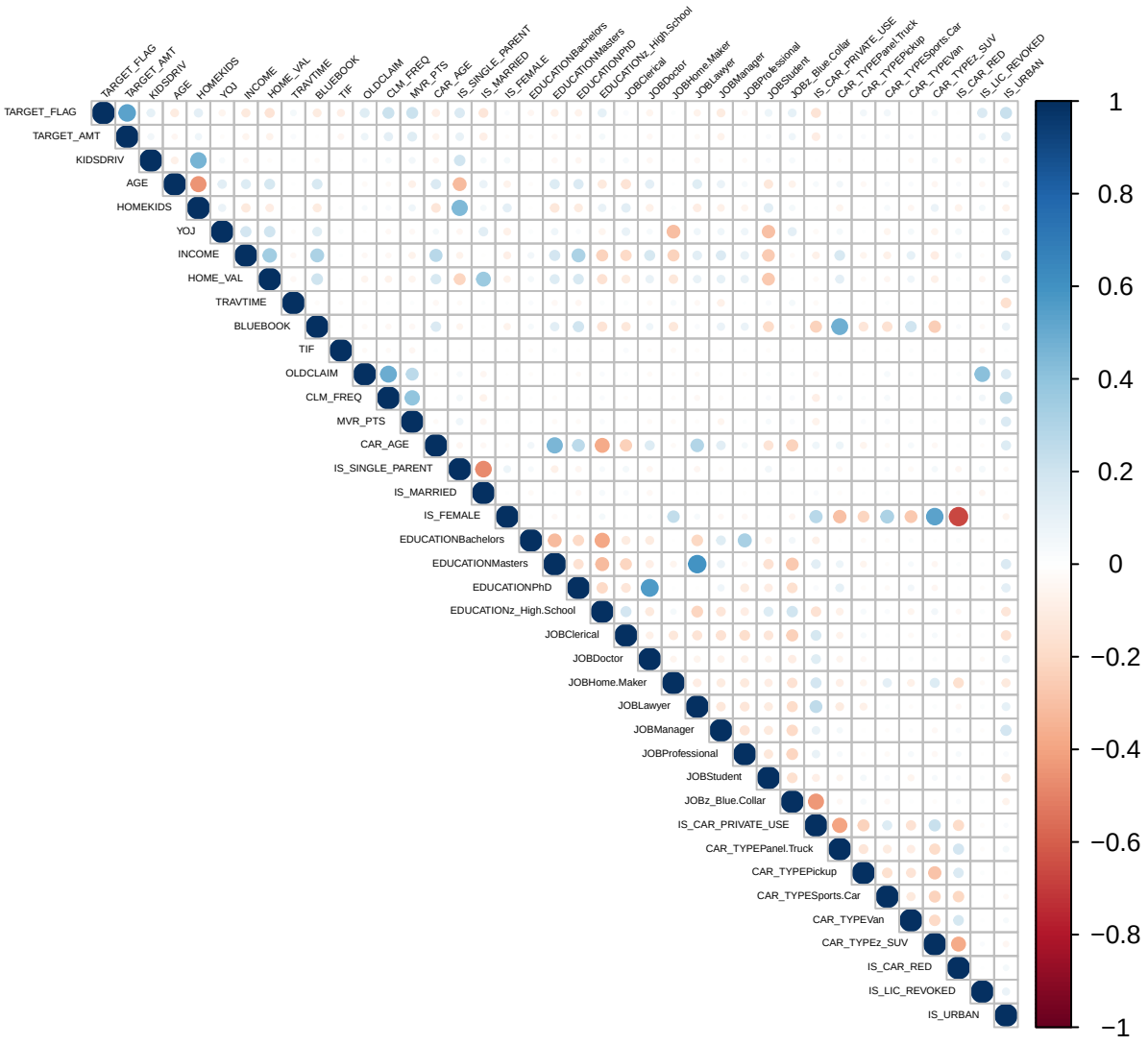
### Correlations

Let's create some visualizations for the correlation matrix.

Let's start with a combined correlation, that is, no difference in between TARGET_FLAG.

### Combined Graphical visualization

First, let's create a visual representation of correlations with a heat map as a guide.



Something interesting to note from the above graph is the existing moderate negative correlation in between IS_FEMALE and IS_RED_CAR, the value for this correlation is: -0.6666. Now, at this point we should not make any inference from this data since IS_FEMALE means either "MALE" or "FEMALE" and IS_RED_CAR means either "Yes" or "No". Further analysis needs to be performed to attain any conclusion related to those two variables.

Also, we can notice some moderate strong correlations from our given data set such as the relation ship in between EDUCATIONPhD and JOBDoctor with a correlation value of 0.5633. Another moderate correlation noticed in the data set is between EDUCATIONMasters and JOBLawyer with a correlation value of 0.5993.

**Combined Numerical visualization**

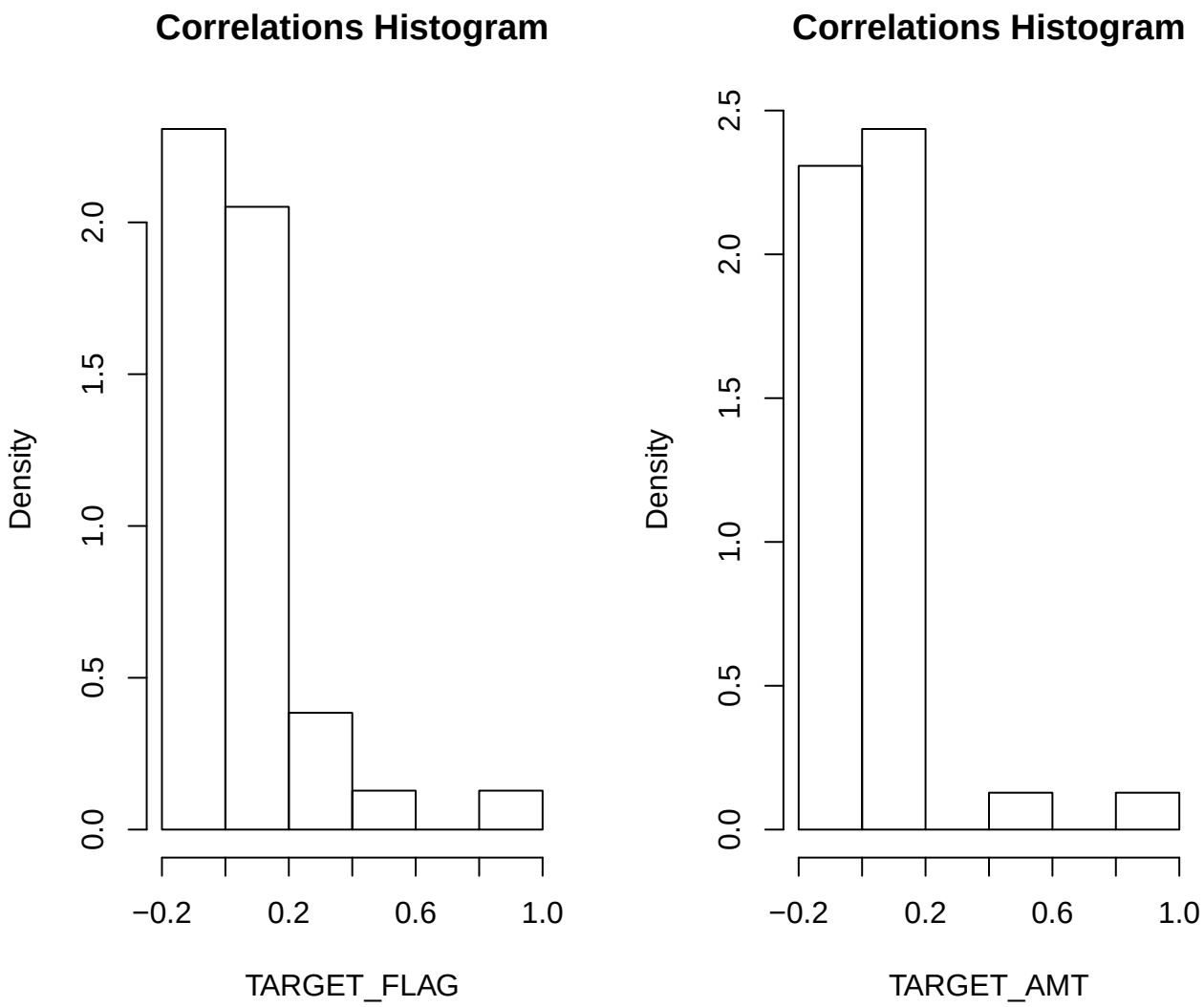
From the above graph, we can easily identify some sort of correlations in between the response variables TARGET_FLAG and TARGET_AMT and other variables.

Let's read our correlations table to gain extra insights.

| | TARGET_FLAG | TARGET_AMT |
|---|---|---|
| TARGET_FLAG | 1.0000000 | 0.5343390 |
| TARGET_AMT | 0.5343390 | 1.0000000 |
| KIDSDRIV | 0.1037552 | 0.0553942 |
| AGE | -0.1009939 | -0.0410444 |
| HOMEKIDS | 0.1157760 | 0.0619890 |
| YOJ | -0.0583146 | -0.0143049 |
| INCOME | -0.1102605 | -0.0539195 |
| HOME_VAL | -0.1437383 | -0.0633940 |
| TRAVTIME | 0.0482173 | 0.0279888 |
| BLUEBOOK | -0.1033931 | -0.0046996 |
| TIF | -0.0823179 | -0.0464814 |
| OLDCLAIM | 0.1375147 | 0.0710055 |
| CLM_FREQ | 0.2158917 | 0.1164467 |
| MVR_PTS | 0.2193023 | 0.1378661 |
| CAR_AGE | -0.0880951 | -0.0541220 |
| IS_SINGLE_PARENT | 0.1577305 | 0.0969660 |
| IS_MARRIED | -0.1349096 | -0.0876704 |
| IS_FEMALE | 0.0208928 | -0.0110533 |
| EDUCATIONBachelors | -0.0429995 | -0.0172792 |
| EDUCATIONMasters | -0.0762068 | -0.0351720 |
| EDUCATIONPhD | -0.0653596 | -0.0244249 |
| EDUCATIONz_High.School | 0.1099203 | 0.0420985 |
| JOBClerical | 0.0274597 | 0.0078049 |
| JOBDoctor | -0.0583510 | -0.0347507 |
| JOBHome.Maker | 0.0113211 | -0.0070821 |
| JOBLawyer | -0.0616731 | -0.0291860 |
| JOBManager | -0.1053383 | -0.0646074 |
| JOBProfessional | -0.0391082 | -0.0045465 |
| JOBStudent | 0.0770910 | 0.0244097 |
| JOBz_Blue.Collar | 0.1019157 | 0.0618307 |
| IS_CAR_PRIVATE_USE | -0.1428605 | -0.0986167 |
| CAR_TYPEPanel.Truck | -0.0002809 | 0.0294682 |
| CAR_TYPEPickup | 0.0562177 | 0.0219152 |
| CAR_TYPESports.Car | 0.0573354 | 0.0232939 |
| CAR_TYPEVan | 0.0030861 | 0.0234793 |
| CAR_TYPEz_SUV | 0.0451690 | 0.0059422 |
| IS_CAR_RED | -0.0068173 | 0.0080916 |
| IS_LIC_REVOKED | 0.1514836 | 0.0614149 |
| IS_URBAN | 0.2241890 | 0.1209762 |

**Combined Correlations histogram**

Something very interesting to note from the above table, is that the correlations in between the data-sets seems to be very low. We can notice that in the below density distributions for the respective correlations.



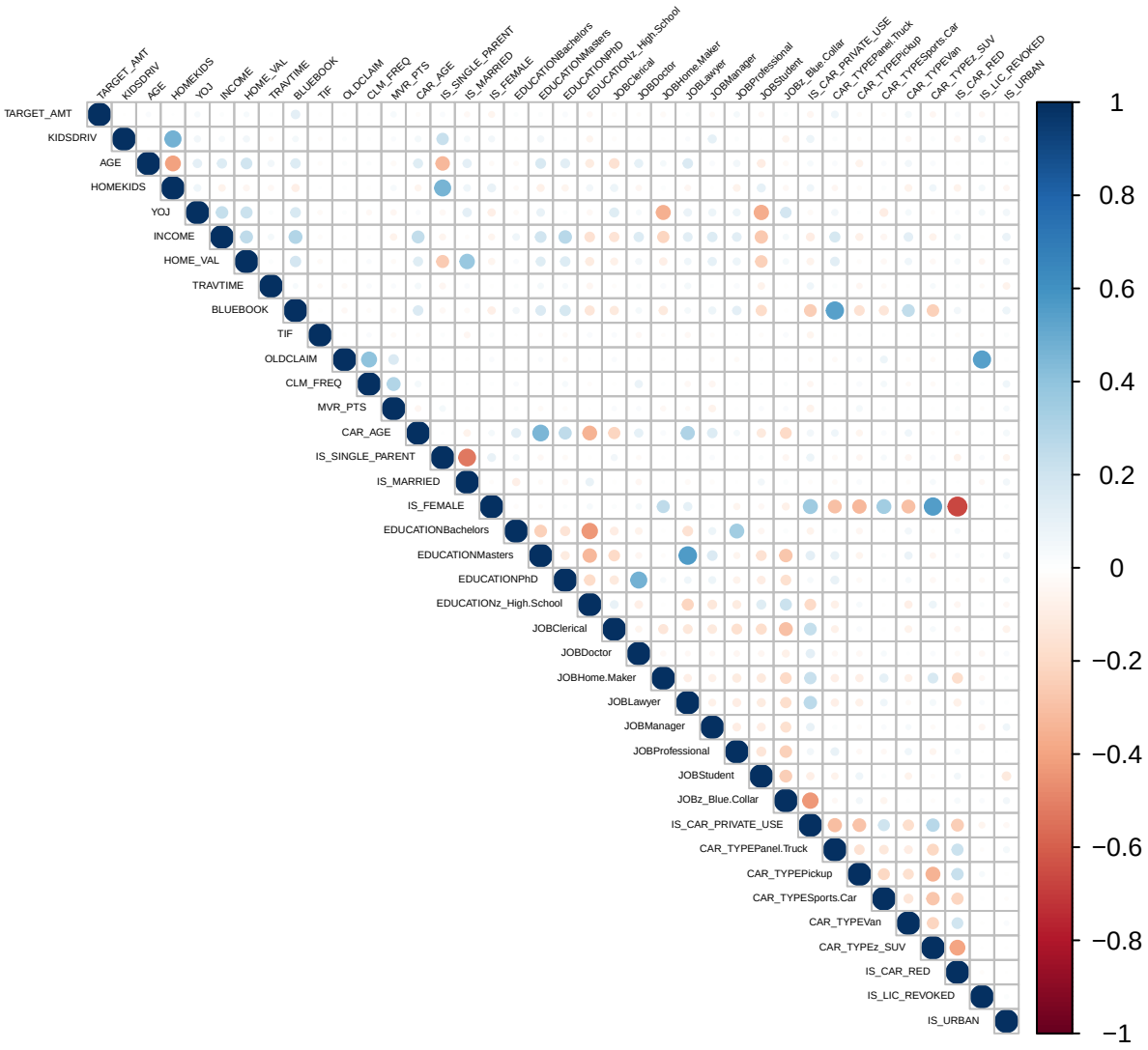

**TARGET_FLAG = Accidents**

Let's do a correlation analysis for the data set in which `TARGET_FLAG` = 1 (meaning it had an accident).

**Accidents Graphical visualization**

Let's create a visual representation of correlations with a heat map as a guide for all records in which `TARGET_FLAG` = 1.

Please note that the row named `TARGET_FLAG` is not included since all records have `TARGET_FLAG` value of 1, making it redundant.




Same as before, is interesting to note from the above graph, the existing moderate negative correlation in between `IS_FEMALE` and `IS_RED_CAR`, the value for this correlation is: -0.6678. Now, at this point we should not make any inference from this data since `IS_FEMALE` means either "MALE" or "FEMALE" and `IS_RED_CAR` means either "Yes" or "No". Further analysis needs to be performed to attain any conclusion related to those two variables.

Also, we can notice how new moderate correlations appeared that were not present before the split; that is:

- `IS_FEMALE` seems to be moderately correlated to `CAR_TYPEz_SUV` with a value of 0.5529.
- `BLUEBOOK` seems to be moderately correlated to `CAR_TYPEPanel.Truck` with a value of 0.5484.
- `OLDCLAIM` seems to be moderately correlated to `IS_LIC_REVOKED` with a value of 0.5427.

Now, if we think about the data and their correlations, some data points seem to make sense. I will not extrapolate too much into this since our main goal is to create a Model in which we could predict the probability that a person will crash their car and also how much money it will cost if the person does crash the car. Hence, I will continue but will keep this correlations in mind.

**Accidents Numerical visualization**

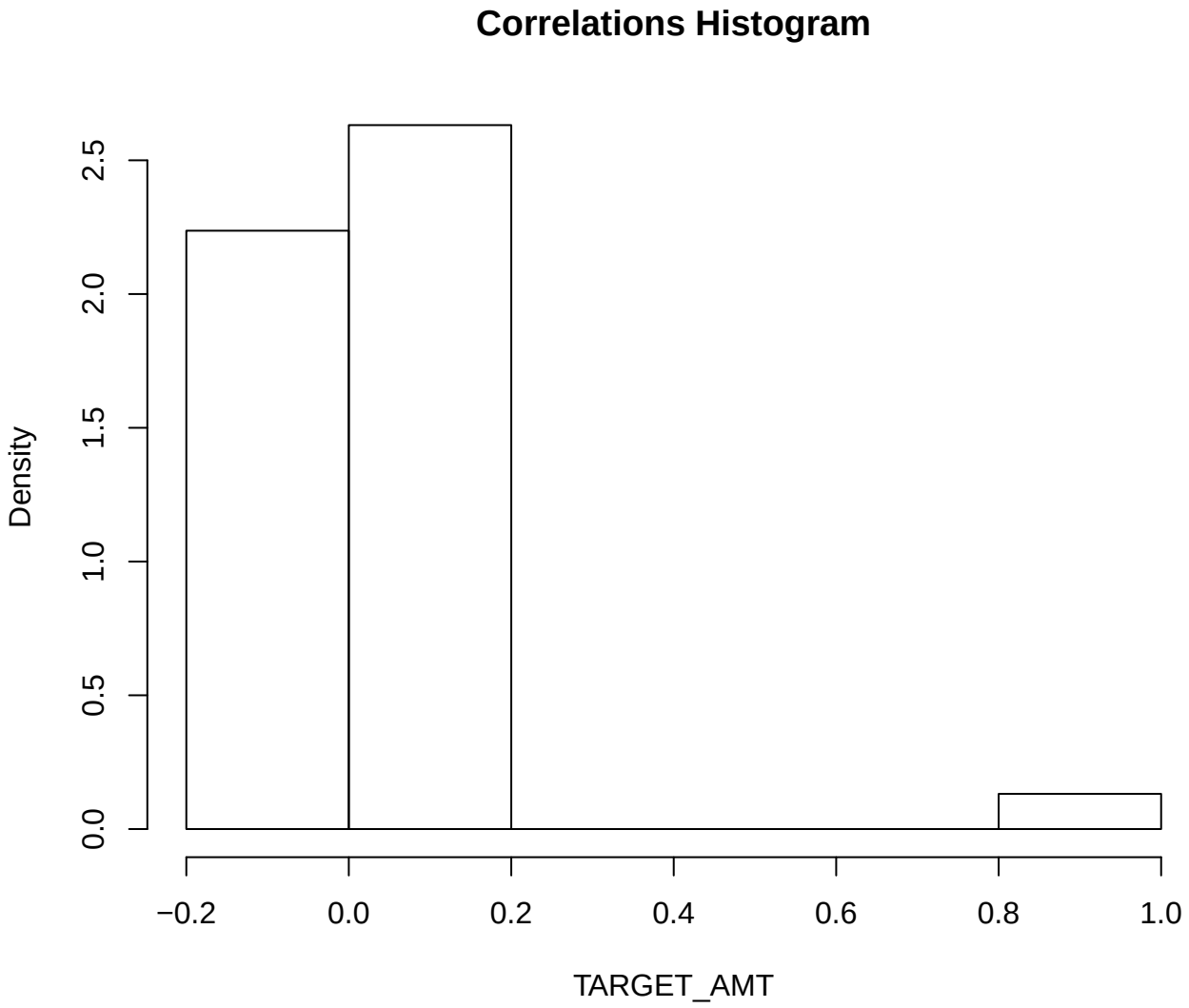
From the above graph, we can easily identify some sort of correlations in between the response variables `TARGET_AMT` and the other variables.

Let's read our correlations table to gain extra insights.

| | TARGET_AMT |
|---|---|
| TARGET_AMT | 1.0000000 |
| KIDSDRIV | -0.0000869 |
| AGE | 0.0267510 |
| HOMEKIDS | 0.0002698 |
| YOJ | 0.0355627 |
| INCOME | 0.0119340 |
| HOME_VAL | 0.0313537 |
| TRAVTIME | 0.0053657 |
| BLUEBOOK | 0.1181297 |
| TIF | -0.0060620 |
| OLDCLAIM | -0.0049723 |
| CLM_FREQ | 0.0023251 |
| MVR_PTS | 0.0396710 |
| CAR_AGE | -0.0164264 |
| IS_SINGLE_PARENT | 0.0238302 |
| IS_MARRIED | -0.0351848 |
| IS_FEMALE | -0.0513430 |
| EDUCATIONBachelors | 0.0136662 |
| EDUCATIONMasters | 0.0143267 |
| EDUCATIONPhD | 0.0294767 |
| EDUCATIONz_High.School | -0.0358840 |
| JOBClerical | -0.0151891 |
| JOBDoctor | -0.0122018 |
| JOBHome.Maker | -0.0293974 |
| JOBLawyer | 0.0102382 |
| JOBManager | -0.0256129 |
| JOBProfessional | 0.0406747 |
| JOBStudent | -0.0331511 |
| JOBz_Blue.Collar | 0.0155259 |
| IS_CAR_PRIVATE_USE | -0.0496142 |
| CAR_TYPEPanel.Truck | 0.0682806 |
| CAR_TYPEPickup | -0.0174060 |
| CAR_TYPESports.Car | -0.0152654 |
| CAR_TYPEVan | 0.0499290 |
| CAR_TYPEz_SUV | -0.0405600 |
| IS_CAR_RED | 0.0271768 |
| IS_LIC_REVOKED | -0.0365018 |
| IS_URBAN | 0.0048888 |

**Accidents Correlations histogram**

Something very interesting to note from the above table, is that the correlations in between the data-sets seems to be very low. We can notice that behavior in the below density distribution for the respective correlations.




**Comparing Means**

The below table, compare the means for both records indicated in the `TARGET_FLAG` variable, that is 1 = Accident vs 0 = No Accident.

| | No Accident | Accident | Insights | Obs | Level |
|---|---|---|---|---|---|
| TARGET_FLAG | 0.00 | 1.00 | Inf % higher | - | *** |
| TARGET_AMT | 0.00 | 5704.15 | Inf % higher | - | *** |
| KIDSDRIV | 0.14 | 0.26 | 86 % higher | - | ** |
| AGE | 45.32 | 43.34 | 4 % lower | + | . |
| HOMEKIDS | 0.64 | 0.94 | 47 % higher | - | * |
| YOJ | 10.70 | 10.13 | 5 % lower | + | . |
| INCOME | 72457.00 | 57593.00 | 21 % lower | + | . |
| HOME_VAL | 183495.00 | 133802.00 | 27 % lower | + | * |
| TRAVTIME | 33.03 | 34.77 | 5 % higher | - | . |
| BLUEBOOK | 16231.00 | 14255.00 | 12 % lower | + | . |
| TIF | 5.56 | 4.78 | 14 % lower | + | . |
| OLDCLAIM | 3312.00 | 6049.00 | 83 % higher | - | ** |
| CLM_FREQ | 0.65 | 1.22 | 88 % higher | - | ** |
| MVR_PTS | 1.41 | 2.48 | 76 % higher | - | ** |
| CAR_AGE | 9.04 | 7.82 | 13 % lower | + | . |
| IS_SINGLE_PARENT | 0.10 | 0.22 | 120 % higher | - | *** |
| IS_MARRIED | 0.64 | 0.49 | 23 % lower | + | . |
| IS_FEMALE | 0.53 | 0.55 | 4 % higher | - | . |
| EDUCATIONBachelors | 0.29 | 0.24 | 17 % lower | + | . |
| EDUCATIONMasters | 0.22 | 0.15 | 32 % lower | + | * |
| EDUCATIONPhD | 0.10 | 0.06 | 40 % lower | + | * |
| EDUCATIONz_High.School | 0.26 | 0.37 | 42 % higher | - | * |
| JOBClerical | 0.15 | 0.17 | 13 % higher | - | . |
| JOBDoctor | 0.04 | 0.01 | 75 % lower | + | ** |
| JOBHome.Maker | 0.08 | 0.08 | No insights | | |
| JOBLawyer | 0.11 | 0.07 | 36 % lower | + | * |
| JOBManager | 0.14 | 0.06 | 57 % lower | + | * |
| JOBProfessional | 0.14 | 0.11 | 21 % lower | + | . |
| JOBStudent | 0.07 | 0.12 | 71 % higher | - | * |
| JOBz_Blue.Collar | 0.20 | 0.29 | 45 % higher | - | * |
| IS_CAR_PRIVATE_USE | 0.67 | 0.51 | 24 % lower | + | . |
| CAR_TYPEPanel.Truck | 0.08 | 0.08 | No insights | | |
| CAR_TYPEPickup | 0.16 | 0.21 | 31 % higher | - | * |
| CAR_TYPESports.Car | 0.10 | 0.14 | 40 % higher | - | * |
| CAR_TYPEVan | 0.09 | 0.09 | No insights | | |
| CAR_TYPEz_SUV | 0.27 | 0.32 | 19 % higher | - | . |
| IS_CAR_RED | 0.29 | 0.29 | No insights | | |
| IS_LIC_REVOKED | 0.09 | 0.21 | 133 % higher | - | *** |
| IS_URBAN | 0.74 | 0.95 | 28 % higher | - | * |

It is very important to note how some insights are taken from the two data sets by comparing side by side. In order to identify those results, I have created two extra columns labeled `Obs` in which denotes how some variables could imply positive or negative outcomes related to accidents, the `Level` has four different indicators depending on the percentage increase as follows:

- **Pos**: Two options "+" or "-" meaning, the data shows an increase or decrease in between

comparisons.

- **Level**: Has four level as follows:

– **"."**: Percentage of difference is less than 25%.

– **"*"**: Percentage of difference is between $[25, 75[$%.

– **"**"**: Percentage of difference is between $[75, 100[$%.

– **"***"**: Percentage of difference is more or equal than 100%.

Also, is important to note that some variables seem to have a beneficial role in avoiding accidents, and that can be seeing in the above table as well.

Something worthy of mentioning is that our theoretical effect mentions: *"urban leyend says that women have less crashes than men"*. By looking at the above table, this legend could be answered as to be false, we could see a slight increase of FEMALES involved in car accidents in about 3% to 4% higher than not having accidents, that is an increase from about 0.53 to about 0.55. Also, we should expect a rate of about 50% since is considered even for insured drivers in America.

**BUILD MODELS**

At this point, we are getting ready to start building models, however I would like to point out that in this case is a little bit difficult to determine what data transformation could be used in order to refine our models.

**Binary Logistic Regression Models**

I would like to point that since this work requires **Binary Logistic Regression**, we are going to be using the **logit** function as our Likelihood link function for Logistic Regression by assuming that it follows a binomial distribution as follows:

$$y_i|x_i \sim Bin(m_i, \theta(x_i))$$

so that,

$$P(Y_i = y_i|x_i) = \binom{m_i}{y_i}\theta(x_i)^{y_i}(1-\theta(x_i))^{m_i-y_i}$$

Now, in order to solve our problem, we need to build a linear predictor model in which the individual predictors that compose the response $Y_i$ are all subject to the same $q$ predictors $(x_{i1}, \ldots, x_{iq})$. Please note that the group of predictors, are commonly known as **covariate classess**. In this case, we need a model that describes the relationship of $x_1, \ldots, x_q$ to $p$. In order to solve this problem, we will construct a linear predictor model as follows:

$$\mathfrak{N}_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}$$

**Logit link function**

In this case, since we need to set $\mathfrak{N}_i = p_i$; with $0 \leq p_i \leq 1$, I will use the *link function* $g$ such that $\mathfrak{N}_i = g(p_i)$ with $0 \leq g^{-1}(\mathfrak{N}) \leq 1$ for any $\mathfrak{N}$. In order to do so, I will pick the **Logit** link function $\mathfrak{N} = log(p/(1-p))$.

An alternate way will be by employing the $\chi^2$ Chi square distribution; for the purposes of this project, I will employ the use of the binomial distribution or the $\chi^2$ depending on which one is a better choice, also I will assume that all $Y_i$ are all independent of each other.

**Binomial NULL Model**

In this section I will build a **Binary Logistic Regression** Null model utilizing all the variables and data, please note that I won't do any transformations. This model will be considered to be valid and will be considered as we advance. In order to build this model, I will not include the **TARGET_AMT** variable since that will be employed in the next model build up.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ 1, family = binomial(link = "logit"),
##     data = data.train.bin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7825  -0.7825  -0.7825   1.6327   1.6327
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.02669    0.02512  -40.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 9415.3  on 8159  degrees of freedom
## AIC: 9417.3
##
## Number of Fisher Scoring iterations: 4
```

I will assume that this to be a valid model.

**Binomial FULL Model**

In this section I will build a **Binary Logistic Regression** Full model utilizing all the variables and data, please note that I won't do any transformations. This model will be considered to be valid and will be considered as we advance.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial(link = "logit"),
```

```
##     data = data.train.bin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5767  -0.7119  -0.4036   0.6227   3.1503
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.994e+00  3.314e-01  -9.032  < 2e-16 ***
## KIDSDRIV                3.840e-01  6.107e-02   6.287 3.23e-10 ***
## AGE                    -7.555e-04  3.975e-03  -0.190 0.849257
## HOMEKIDS                4.966e-02  3.692e-02   1.345 0.178586
## YOJ                    -7.653e-03  7.788e-03  -0.983 0.325777
## INCOME                 -1.625e-06  6.267e-07  -2.593 0.009526 **
## HOME_VAL               -6.858e-07  2.348e-07  -2.921 0.003489 **
## TRAVTIME                1.441e-02  1.881e-03   7.660 1.85e-14 ***
## BLUEBOOK               -2.297e-05  5.212e-06  -4.407 1.05e-05 ***
## TIF                    -5.522e-02  7.330e-03  -7.534 4.92e-14 ***
## OLDCLAIM               -1.422e-05  3.907e-06  -3.639 0.000274 ***
## CLM_FREQ                1.999e-01  2.849e-02   7.014 2.31e-12 ***
## MVR_PTS                 1.153e-01  1.360e-02   8.477  < 2e-16 ***
## CAR_AGE                -7.370e-04  6.266e-03  -0.118 0.906362
## IS_SINGLE_PARENT        3.825e-01  1.094e-01   3.497 0.000471 ***
## IS_MARRIED             -5.582e-01  7.783e-02  -7.172 7.39e-13 ***
## IS_FEMALE              -8.132e-02  1.118e-01  -0.727 0.466943
## EDUCATIONBachelors     -4.376e-01  1.128e-01  -3.881 0.000104 ***
## EDUCATIONMasters       -3.537e-01  1.726e-01  -2.049 0.040433 *
## EDUCATIONPhD           -3.174e-01  2.051e-01  -1.548 0.121597
## EDUCATIONz_High.School -3.047e-03  9.486e-02  -0.032 0.974379
## JOBClerical             5.050e-01  1.942e-01   2.601 0.009297 **
## JOBDoctor              -4.103e-01  2.654e-01  -1.546 0.122087
## JOBHome.Maker           4.230e-01  2.036e-01   2.078 0.037726 *
## JOBLawyer               1.266e-01  1.686e-01   0.751 0.452621
## JOBManager             -5.246e-01  1.705e-01  -3.077 0.002089 **
## JOBProfessional         1.895e-01  1.775e-01   1.068 0.285511
## JOBStudent              4.288e-01  2.087e-01   2.055 0.039873 *
## JOBz_Blue.Collar        3.499e-01  1.845e-01   1.896 0.057940 .
## IS_CAR_PRIVATE_USE     -7.639e-01  9.170e-02  -8.330  < 2e-16 ***
## CAR_TYPEPanel.Truck     5.505e-01  1.614e-01   3.411 0.000647 ***
## CAR_TYPEPickup          5.510e-01  1.007e-01   5.472 4.46e-08 ***
## CAR_TYPESports.Car      1.029e+00  1.296e-01   7.941 2.01e-15 ***
## CAR_TYPEVan             6.114e-01  1.261e-01   4.847 1.25e-06 ***
## CAR_TYPEz_SUV           7.703e-01  1.111e-01   6.933 4.12e-12 ***
## IS_CAR_RED             -2.506e-03  8.618e-02  -0.029 0.976805
## IS_LIC_REVOKED          8.894e-01  9.117e-02   9.756  < 2e-16 ***
## IS_URBAN                2.387e+00  1.129e-01  21.147  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7316.8  on 8122  degrees of freedom
## AIC: 7392.8
##
## Number of Fisher Scoring iterations: 5
```

In this particular case, we notice how some variables are not statistically significant; for study purposes, I will assume that this is a valid model.

**Binomial STEP Model**

In this case, I will create multiple models using the STEP function from R.

```
bin_Model_STEP <- step(bin_Model_NULL,
                       scope = list(upper=bin_Model_FULL),
                       direction="both",
                       test="Chisq",
                       data=data.train.bin)
```

For simplicity reasons, I have decided not to include the automatic responses; instead I will present the final model results.

**ANOVA results**

Let's check an ANOVA table based on the above testing results.

| Step | Df | Deviance | Resid. Df | Resid. Dev | AIC |
|---|---|---|---|---|---|
| | NA | NA | 8159 | 9415.297 | 9417.297 |
| + IS_URBAN | -1 | 501.7552780 | 8158 | 8913.542 | 8917.542 |
| + MVR_PTS | -1 | 271.5277948 | 8157 | 8642.014 | 8648.014 |
| + HOME_VAL | -1 | 219.1676131 | 8156 | 8422.846 | 8430.846 |
| + IS_CAR_PRIVATE_USE | -1 | 174.3281407 | 8155 | 8248.518 | 8258.518 |
| + BLUEBOOK | -1 | 138.9556413 | 8154 | 8109.562 | 8121.562 |
| + IS_SINGLE_PARENT | -1 | 112.1201022 | 8153 | 7997.442 | 8011.442 |
| + IS_LIC_REVOKED | -1 | 95.2189847 | 8152 | 7902.223 | 7918.223 |
| + JOBManager | -1 | 88.0830361 | 8151 | 7814.140 | 7832.140 |
| + TRAVTIME | -1 | 58.2595752 | 8150 | 7755.881 | 7775.881 |
| + TIF | -1 | 57.0485376 | 8149 | 7698.832 | 7720.832 |
| + KIDSDRIV | -1 | 53.2561495 | 8148 | 7645.576 | 7669.576 |
| + CLM_FREQ | -1 | 43.2287532 | 8147 | 7602.347 | 7628.347 |
| + EDUCATIONz_High.School | -1 | 33.8405122 | 8146 | 7568.507 | 7596.507 |
| + CAR_TYPESports.Car | -1 | 28.2833549 | 8145 | 7540.223 | 7570.223 |
| + CAR_TYPEz_SUV | -1 | 35.1767133 | 8144 | 7505.047 | 7537.047 |
| + IS_MARRIED | -1 | 30.4554129 | 8143 | 7474.591 | 7508.591 |
| + INCOME | -1 | 32.1572427 | 8142 | 7442.434 | 7478.434 |
| + JOBClerical | -1 | 17.1378050 | 8141 | 7425.296 | 7463.296 |
| + OLDCLAIM | -1 | 13.3590688 | 8140 | 7411.937 | 7451.937 |
| + CAR_TYPEPickup | -1 | 11.8937289 | 8139 | 7400.043 | 7442.043 |
| + CAR_TYPEVan | -1 | 14.2678147 | 8138 | 7385.776 | 7429.776 |
| + CAR_TYPEPanel.Truck | -1 | 16.0797833 | 8137 | 7369.696 | 7415.696 |
| + JOBDoctor | -1 | 8.4824840 | 8136 | 7361.213 | 7409.213 |
| + CAR_AGE | -1 | 6.6048430 | 8135 | 7354.609 | 7404.609 |
| + EDUCATIONBachelors | -1 | 5.3370024 | 8134 | 7349.272 | 7401.272 |
| + EDUCATIONMasters | -1 | 11.3304398 | 8133 | 7337.941 | 7391.941 |
| + EDUCATIONPhD | -1 | 9.4981966 | 8132 | 7328.443 | 7384.443 |
| - CAR_AGE | 1 | 0.0047192 | 8133 | 7328.448 | 7382.448 |
| - EDUCATIONz_High.School | 1 | 0.0934403 | 8134 | 7328.541 | 7380.541 |
| + HOMEKIDS | -1 | 2.2554506 | 8133 | 7326.286 | 7380.286 |
| + YOJ | -1 | 2.7828687 | 8132 | 7323.503 | 7379.503 |

From the above results and calculations, it was concluded that the best model is as follows:

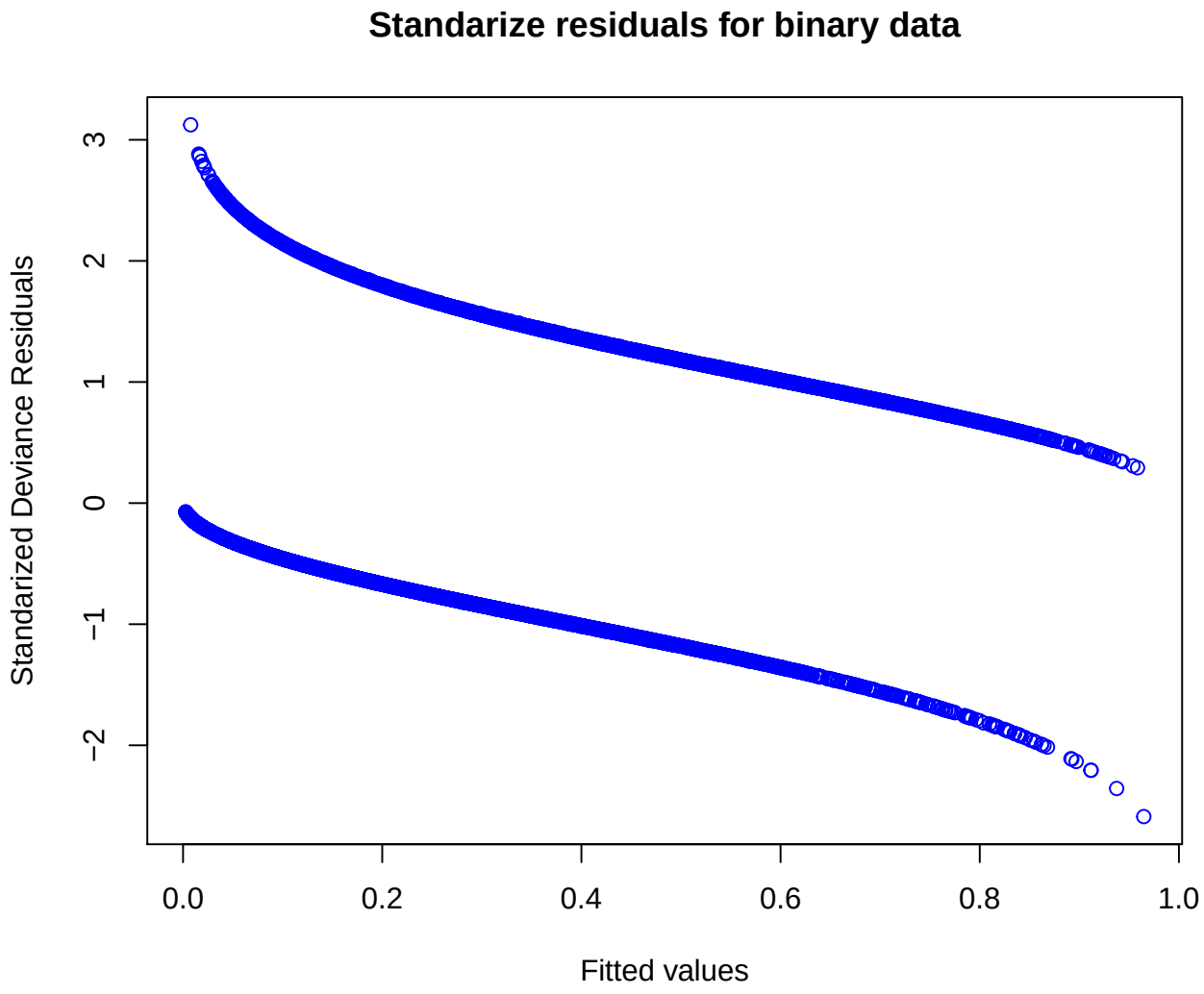
```
##
## Call:
## glm(formula = TARGET_FLAG ~ IS_URBAN + MVR_PTS + HOME_VAL + IS_CAR_PRIVATE_USE +
##     BLUEBOOK + IS_SINGLE_PARENT + IS_LIC_REVOKED + JOBManager +
##     TRAVTIME + TIF + KIDSDRIV + CLM_FREQ + CAR_TYPESports.Car +
##     CAR_TYPEz_SUV + IS_MARRIED + INCOME + JOBClerical + OLDCLAIM +
##     CAR_TYPEPickup + CAR_TYPEVan + CAR_TYPEPanel.Truck + JOBDoctor +
##     EDUCATIONBachelors + EDUCATIONMasters + EDUCATIONPhD + HOMEKIDS +
##     YOJ, family = binomial(link = "logit"), data = data.train.bin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.5863  -0.7155  -0.4047   0.6283   3.1234
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.592e+00  1.839e-01 -14.095  < 2e-16 ***
## IS_URBAN             2.372e+00  1.124e-01  21.092  < 2e-16 ***
## MVR_PTS              1.141e-01  1.357e-02   8.408  < 2e-16 ***
## HOME_VAL            -7.446e-07  2.309e-07  -3.225 0.001260 **
## IS_CAR_PRIVATE_USE  -7.864e-01  7.499e-02 -10.487  < 2e-16 ***
## BLUEBOOK            -2.527e-05  4.647e-06  -5.437 5.41e-08 ***
## IS_SINGLE_PARENT     3.760e-01  1.087e-01   3.458 0.000544 ***
## IS_LIC_REVOKED       8.868e-01  9.105e-02   9.740  < 2e-16 ***
## JOBManager          -7.463e-01  1.072e-01  -6.961 3.39e-12 ***
## TRAVTIME             1.449e-02  1.879e-03   7.714 1.21e-14 ***
## TIF                 -5.541e-02  7.323e-03  -7.567 3.82e-14 ***
## KIDSDRIV             3.820e-01  5.997e-02   6.369 1.91e-10 ***
## CLM_FREQ             1.997e-01  2.845e-02   7.019 2.23e-12 ***
## CAR_TYPESports.Car   9.832e-01  1.065e-01   9.232  < 2e-16 ***
## CAR_TYPEz_SUV        7.235e-01  8.523e-02   8.489  < 2e-16 ***
## IS_MARRIED          -5.487e-01  7.748e-02  -7.081 1.43e-12 ***
## INCOME              -1.924e-06  6.116e-07  -3.146 0.001653 **
## JOBClerical          1.720e-01  8.898e-02   1.934 0.053174 .
## OLDCLAIM            -1.413e-05  3.902e-06  -3.622 0.000293 ***
## CAR_TYPEPickup       5.293e-01  9.852e-02   5.373 7.76e-08 ***
## CAR_TYPEVan          6.065e-01  1.196e-01   5.069 3.99e-07 ***
## CAR_TYPEPanel.Truck  5.395e-01  1.431e-01   3.769 0.000164 ***
## JOBDoctor           -5.278e-01  2.494e-01  -2.117 0.034268 *
## EDUCATIONBachelors  -4.805e-01  7.559e-02  -6.356 2.07e-10 ***
## EDUCATIONMasters    -5.420e-01  9.198e-02  -5.892 3.80e-09 ***
## EDUCATIONPhD        -5.061e-01  1.458e-01  -3.472 0.000517 ***
## HOMEKIDS             5.740e-02  3.391e-02   1.693 0.090442 .
## YOJ                 -1.185e-02  7.098e-03  -1.669 0.095071 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7323.5  on 8132  degrees of freedom
## AIC: 7379.5
##
## Number of Fisher Scoring iterations: 5
```

From the above results, we noticed how all the predictors are statistical significant, also, we notice how from the above model, the HOME_VAL and the INCOME are not as statistical significant compared to other variables.

**Plot of standardized residuals**

The below plot shows our fitted models vs the deviance standardized residuals.




**Confusion Matrix**

Let's start by building a confusion matrix in order to obtain valuable insights.

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction    0    1
##          0 5550 1257
##          1  458  895
##
##                Accuracy : 0.7898
##                  95% CI : (0.7808, 0.7986)
##     No Information Rate : 0.7363
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3856
##
```

```
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4159
##             Specificity : 0.9238
##          Pos Pred Value : 0.6615
##          Neg Pred Value : 0.8153
##              Prevalence : 0.2637
##          Detection Rate : 0.1097
##    Detection Prevalence : 0.1658
##       Balanced Accuracy : 0.6698
##
##        'Positive' Class : 1
##
```

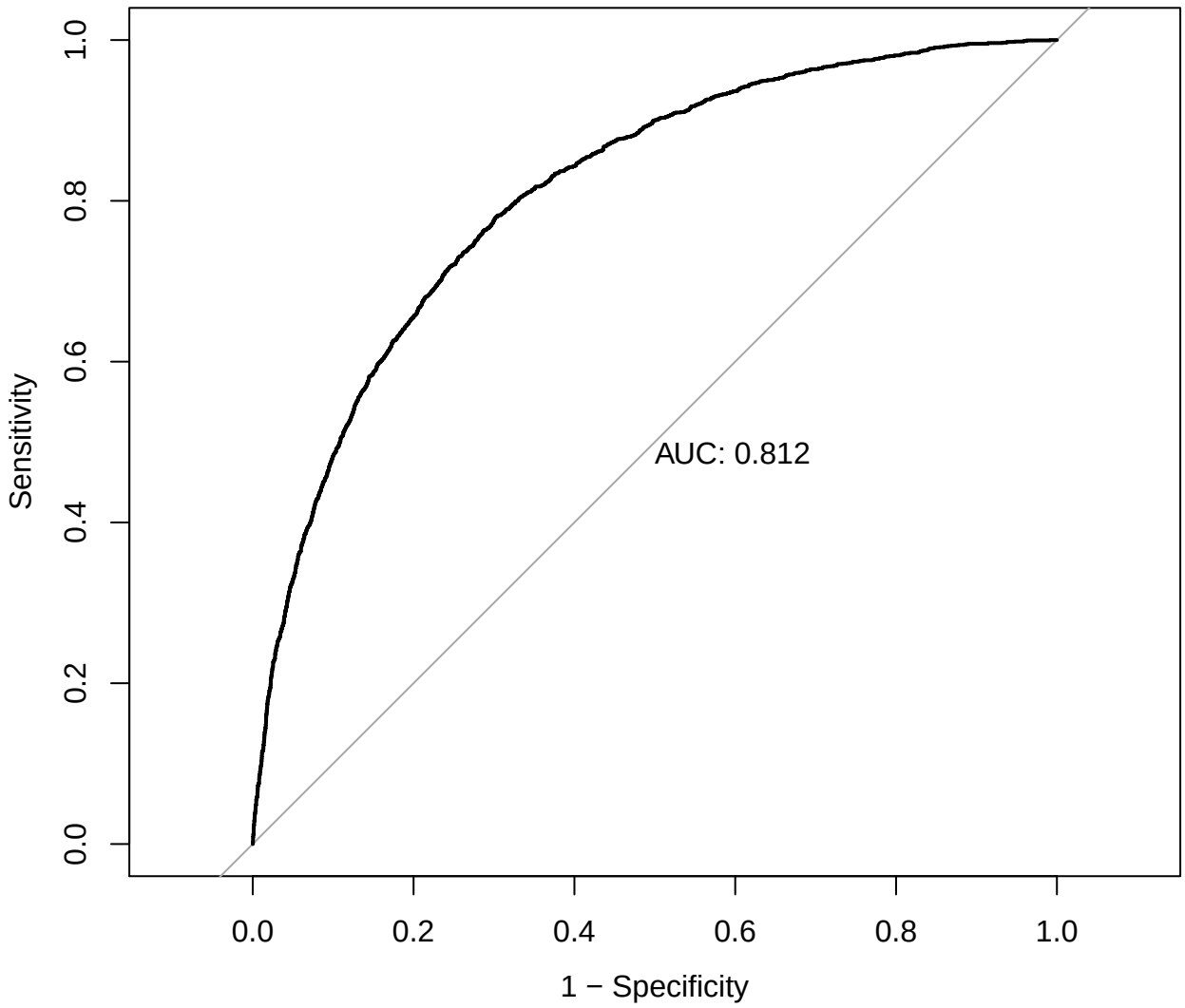
Is interesting to note that the reported Accuracy is 0.7898284.

From the above results, we obtain as follows:

| | Value |
|---|---|
| Sensitivity | 0.4158922 |
| Specificity | 0.9237683 |
| Pos Pred Value | 0.6614930 |
| Neg Pred Value | 0.8153372 |
| Precision | 0.6614930 |
| Recall | 0.4158922 |
| F1 | 0.5106990 |
| Prevalence | 0.2637255 |
| Detection Rate | 0.1096814 |
| Detection Prevalence | 0.1658088 |
| Balanced Accuracy | 0.6698303 |

**ROC and AUC**

As we know, the **Receiver Operating Characteristic Curves** (ROC) is a great quantitative assessment tool of the model. In order to quantify our model, I will employ as follows:




Let's see our confidence intervals for the area under the curve.

| | AUC |
|---|---|
| Lower bound | 0.8015396 |
| Estimated value | 0.8118504 |
| Higher bound | 0.8221612 |

**Binary STEP MODIFIED Model**

In this case, I will add 1 to the following variables and then I will calculate the log, thus to avoid errors since some entries reported 0 and log(0) will produce errors, also I will remove the variable HOMEKIDS; let's take a look as follows:

- Log(1 + INCOME)
- Log(1 + HOME_VAL)
- Log(1 + BLUEBOOK)
- Log(1 + OLDCLAIM)
- HOMEKIDS <– Remove

Let's see the results:

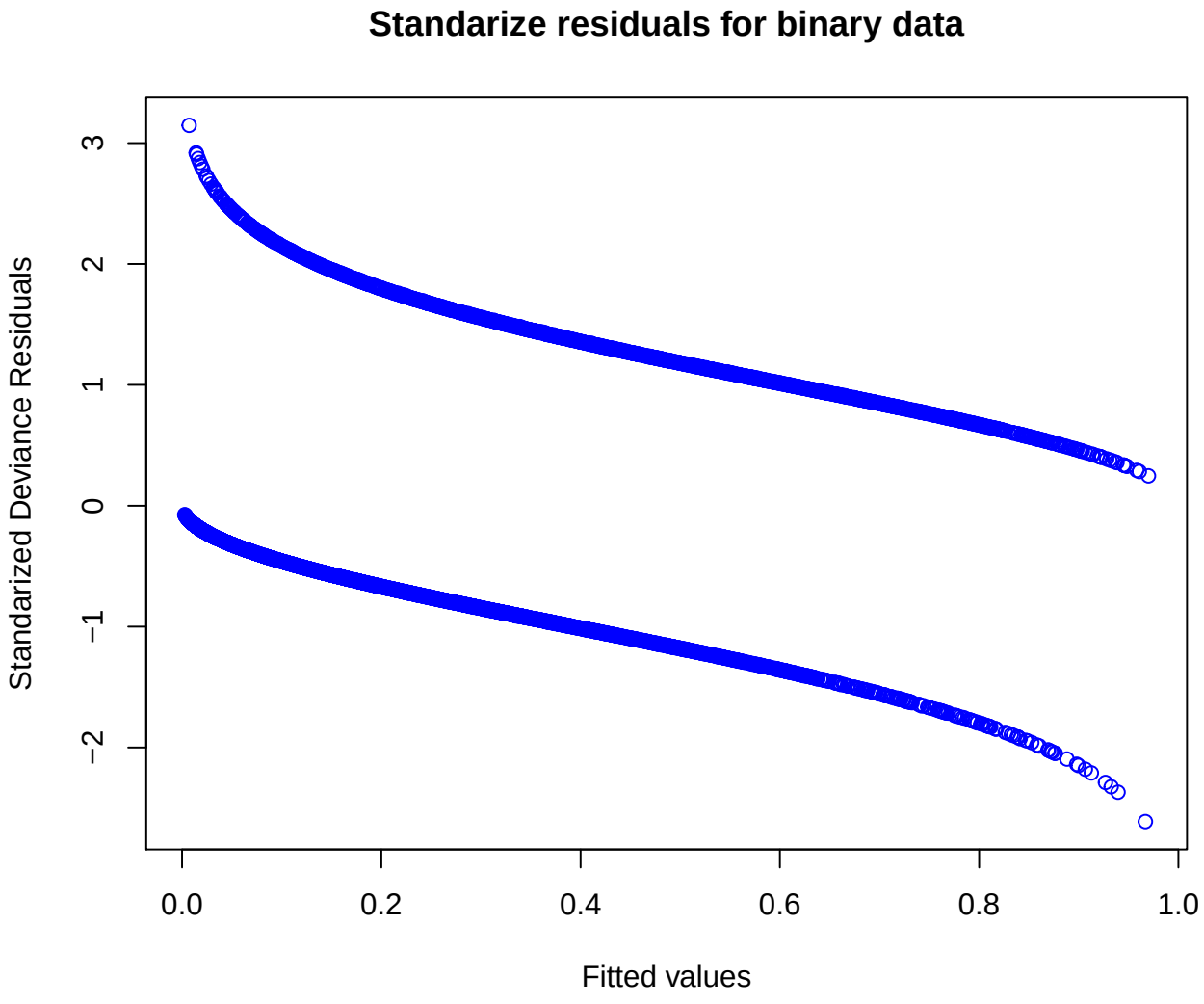
```
##
## Call:
## glm(formula = TARGET_FLAG ~ IS_URBAN + MVR_PTS + log(1 + HOME_VAL) +
##     IS_CAR_PRIVATE_USE + log(1 + BLUEBOOK) + IS_SINGLE_PARENT +
##     IS_LIC_REVOKED + JOBManager + TRAVTIME + TIF + KIDSDRIV +
##     CLM_FREQ + CAR_TYPESports.Car + CAR_TYPEz_SUV + IS_MARRIED +
##     log(1 + INCOME) + JOBClerical + log(1 + OLDCLAIM) + CAR_TYPEPickup +
##     CAR_TYPEVan + CAR_TYPEPanel.Truck + JOBDoctor + EDUCATIONBachelors +
##     EDUCATIONMasters + EDUCATIONPhD + YOJ, family = binomial(link = "logit"),
##     data = data.train.bin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6106  -0.7169  -0.4012   0.6251   3.1464
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.449796   0.525004   0.857  0.39158
## IS_URBAN             2.408764   0.113906  21.147  < 2e-16 ***
## MVR_PTS              0.103476   0.013992   7.395 1.41e-13 ***
## log(1 + HOME_VAL)   -0.023995   0.006356  -3.775  0.00016 ***
## IS_CAR_PRIVATE_USE  -0.807898   0.075476 -10.704  < 2e-16 ***
## log(1 + BLUEBOOK)   -0.328587   0.054484  -6.031 1.63e-09 ***
## IS_SINGLE_PARENT     0.432665   0.094510   4.578 4.69e-06 ***
## IS_LIC_REVOKED       0.700961   0.081276   8.624  < 2e-16 ***
## JOBManager          -0.722699   0.107335  -6.733 1.66e-11 ***
## TRAVTIME             0.014845   0.001883   7.885 3.14e-15 ***
## TIF                 -0.054347   0.007329  -7.415 1.21e-13 ***
## KIDSDRIV             0.423039   0.054971   7.696 1.41e-14 ***
## CLM_FREQ             0.091849   0.043361   2.118  0.03416 *
## CAR_TYPESports.Car   0.945197   0.107131   8.823  < 2e-16 ***
## CAR_TYPEz_SUV        0.732312   0.085011   8.614  < 2e-16 ***
## IS_MARRIED          -0.491494   0.081374  -6.040 1.54e-09 ***
## log(1 + INCOME)     -0.072339   0.013133  -5.508 3.63e-08 ***
## JOBClerical          0.275260   0.090260   3.050  0.00229 **
## log(1 + OLDCLAIM)    0.022103   0.012436   1.777  0.07552 .
## CAR_TYPEPickup       0.535625   0.098545   5.435 5.47e-08 ***
## CAR_TYPEVan          0.595850   0.119427   4.989 6.06e-07 ***
## CAR_TYPEPanel.Truck  0.418935   0.135944   3.082  0.00206 **
## JOBDoctor           -0.484072   0.248086  -1.951  0.05103 .
## EDUCATIONBachelors  -0.469375   0.075173  -6.244 4.27e-10 ***
## EDUCATIONMasters    -0.548683   0.090087  -6.091 1.13e-09 ***
## EDUCATIONPhD        -0.628394   0.139306  -4.511 6.46e-06 ***
## YOJ                  0.017980   0.008905   2.019  0.04348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7301.2  on 8133  degrees of freedom
## AIC: 7355.2
##
## Number of Fisher Scoring iterations: 5
```

As we can see, this transformation produced similar entries compared to our automatically selected model but it seems to be slightly better since the AIC is lower than the automatically selected model by the STEP procedure.

**Plot of standardized residuals**

The below plot shows our fitted models vs the deviance standardized residuals.




**Confusion Matrix**

Let's start by building a confusion matrix in order to obtain valuable insights.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5561 1268
##          1  447  884
##
##                Accuracy : 0.7898
##                  95% CI : (0.7808, 0.7986)
##     No Information Rate : 0.7363
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3833
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4108
##             Specificity : 0.9256
##          Pos Pred Value : 0.6642
##          Neg Pred Value : 0.8143
##              Prevalence : 0.2637
##          Detection Rate : 0.1083
##    Detection Prevalence : 0.1631
##       Balanced Accuracy : 0.6682
##
##        'Positive' Class : 1
##
```

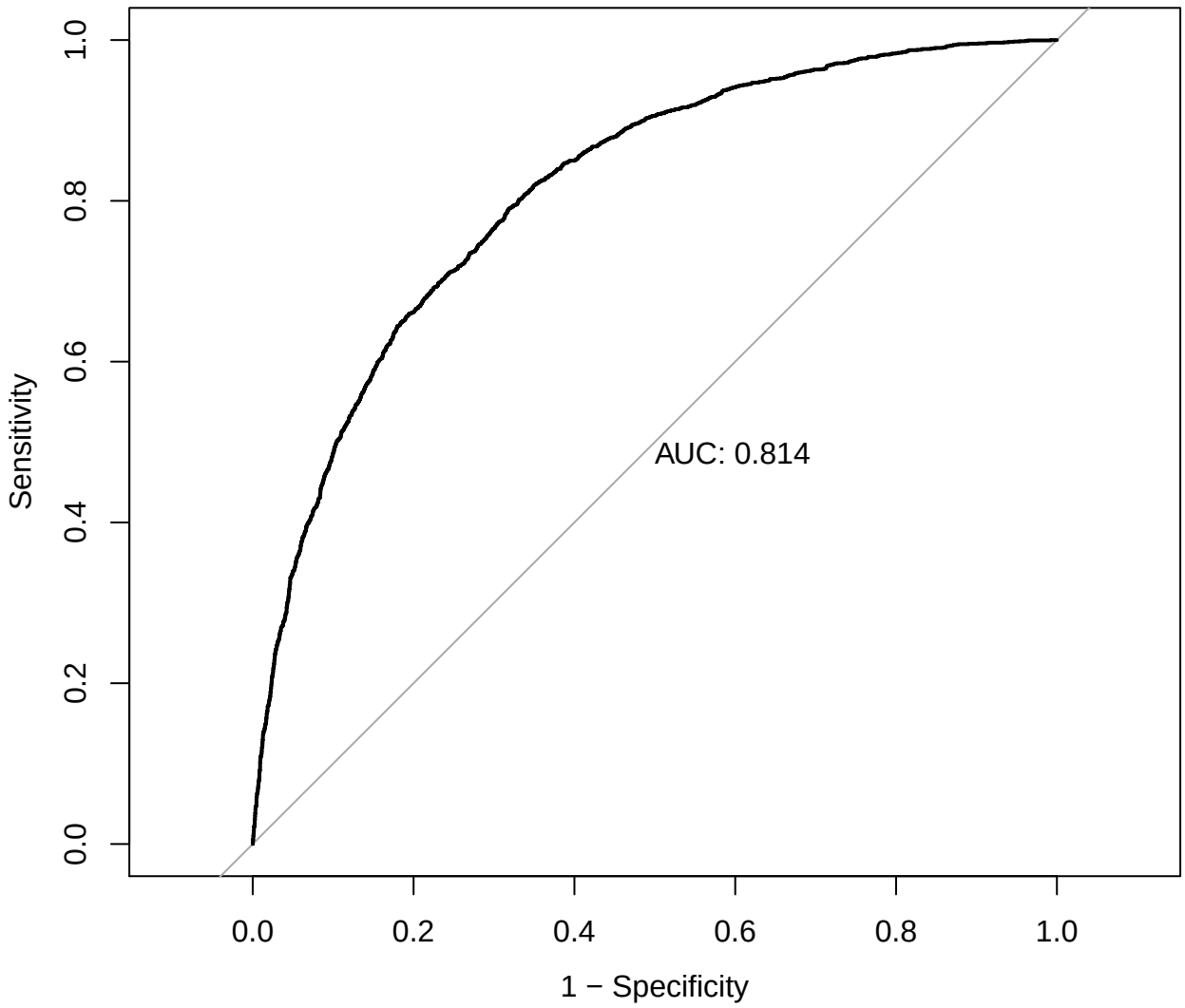
Is interesting to note that the reported Accuracy is 0.7898284.

From the above results, we obtain as follows:

| | Value |
|---|---|
| Sensitivity | 0.4107807 |
| Specificity | 0.9255992 |
| Pos Pred Value | 0.6641623 |
| Neg Pred Value | 0.8143213 |
| Precision | 0.6641623 |
| Recall | 0.4107807 |
| F1 | 0.5076084 |
| Prevalence | 0.2637255 |
| Detection Rate | 0.1083333 |
| Detection Prevalence | 0.1631127 |
| Balanced Accuracy | 0.6681899 |

**ROC and AUC**

As we know, the **Receiver Operating Characteristic Curves** (ROC) is a great quantitative assessment tool of the model. In order to quantify our model, I will employ as follows:




Let's see our confidence intervals for the area under the curve.

| | AUC |
|---|---|
| Lower bound | 0.8033390 |
| Estimated value | 0.8135629 |
| Higher bound | 0.8237867 |

**IMPORTANT**

If we check our theory, the **AIC** defines as follows: *the smaller the value for AIC the better the model*; in this case, we can easily observe how the by adding certain variables, our AIC values decrease making it a better model.

**Binary MODEL SELECTION**

In this case, I will select the model returned in the STEP procedure, that is:

```
bin_Model_FINAL <- bin_Model_STEP
```

The reasons are explained below:

- This model returned the second lowest **Akaike's Information Criterion** AIC.
- This model returned the second nearest to zero median value.
- This model displayed the smallest standard errors for the considered predictor variables.
- This model present the smallest rate of change for all predictor variables.
- This model returned the second lowest residual deviance.
- From the below table we can see how the probability of being higher than the $\chi^2$ are very low.

| | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|
| IS_URBAN | 1 | 444.877135 | 0.0000000 |
| MVR_PTS | 1 | 70.699251 | 0.0000000 |
| HOME_VAL | 1 | 10.400387 | 0.0012599 |
| IS_CAR_PRIVATE_USE | 1 | 109.982078 | 0.0000000 |
| BLUEBOOK | 1 | 29.562730 | 0.0000001 |
| IS_SINGLE_PARENT | 1 | 11.959901 | 0.0005436 |
| IS_LIC_REVOKED | 1 | 94.871581 | 0.0000000 |
| JOBManager | 1 | 48.451199 | 0.0000000 |
| TRAVTIME | 1 | 59.513042 | 0.0000000 |
| TIF | 1 | 57.257133 | 0.0000000 |
| KIDSDRIV | 1 | 40.559335 | 0.0000000 |
| CLM_FREQ | 1 | 49.266736 | 0.0000000 |
| CAR_TYPESports.Car | 1 | 85.227207 | 0.0000000 |
| CAR_TYPEz_SUV | 1 | 72.061019 | 0.0000000 |
| IS_MARRIED | 1 | 50.145391 | 0.0000000 |
| INCOME | 1 | 9.899663 | 0.0016531 |
| JOBClerical | 1 | 3.738453 | 0.0531737 |
| OLDCLAIM | 1 | 13.115512 | 0.0002929 |
| CAR_TYPEPickup | 1 | 28.866084 | 0.0000001 |
| CAR_TYPEVan | 1 | 25.698922 | 0.0000004 |
| CAR_TYPEPanel.Truck | 1 | 14.206252 | 0.0001638 |
| JOBDoctor | 1 | 4.481264 | 0.0342684 |
| EDUCATIONBachelors | 1 | 40.398578 | 0.0000000 |
| EDUCATIONMasters | 1 | 34.721076 | 0.0000000 |
| EDUCATIONPhD | 1 | 12.053972 | 0.0005168 |
| HOMEKIDS | 1 | 2.866491 | 0.0904419 |
| YOJ | 1 | 2.786344 | 0.0950709 |

**Test Binary model**

From the above chosen model, I will create a reduced data frame containing only the variables needed in order to run our model. The selected variables are:

| vars |
|---|
| TARGET_FLAG |
| IS_URBAN |
| MVR_PTS |
| HOME_VAL |
| IS_CAR_PRIVATE_USE |
| BLUEBOOK |
| IS_SINGLE_PARENT |
| IS_LIC_REVOKED |
| JOBManager |
| TRAVTIME |
| TIF |
| KIDSDRIV |
| CLM_FREQ |
| CAR_TYPESports.Car |
| CAR_TYPEz_SUV |
| IS_MARRIED |
| INCOME |
| JOBClerical |
| OLDCLAIM |
| CAR_TYPEPickup |
| CAR_TYPEVan |
| CAR_TYPEPanel.Truck |
| JOBDoctor |
| EDUCATIONBachelors |
| EDUCATIONMasters |
| EDUCATIONPhD |
| HOMEKIDS |
| YOJ |

**Final Model Comparisons**

From here, I will define a NULL model with the chosen variables in order to compare results with the FINAL model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ 1, family = binomial(link = "logit"),
##     data = data.train.final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7825  -0.7825  -0.7825   1.6327   1.6327
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.02669    0.02512  -40.87   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 9415.3  on 8159  degrees of freedom
## AIC: 9417.3
##
## Number of Fisher Scoring iterations: 4
```

**Analysis of Deviance Table**

The below table, will display a Deviance analysis by employing the $\chi^2$ test.

| Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) |
|---|---|---|---|---|
| 8132 | 7323.503 | NA | NA | NA |
| 8159 | 9415.297 | -27 | -2091.794 | 0 |

In the above results, we can easily compare our Residual Deviance in which our model has better results compared to the null model since the null model's deviance will change in -2091.7941687 units compared to our final model.

**Multiple Linear Regression Model**

Now, that we have build our Binary Model, I will proceed to build a Multiple Linear Regression Model in order to predict the `TARGET_AMT` for the records indicating that an accident happened.

**Summaries**

In order to have a better understanding of our current data, I will present a summary table for all the records that have accidents only.

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max |
|---|---|---|---|---|---|---|
| TARGET_FLAG | 1.00 | 1.00 | 1.00 | 1.00000e+00 | 1 | 1.0 |
| TARGET_AMT | 30.28 | 2610.69 | 4108.07 | 5.70415e+03 | 5788 | 107586.1 |
| KIDSDRIV | 0.00 | 0.00 | 0.00 | 2.59800e-01 | 0 | 4.0 |
| AGE | 16.00 | 37.00 | 43.00 | 4.33400e+01 | 50 | 78.0 |
| HOMEKIDS | 0.00 | 0.00 | 0.00 | 9.37300e-01 | 2 | 5.0 |
| YOJ | 0.00 | 8.00 | 11.00 | 1.01300e+01 | 13 | 23.0 |
| INCOME | 0.00 | 21636.00 | 45776.00 | 5.75930e+04 | 73151 | 364441.0 |
| HOME_VAL | 0.00 | 0.00 | 122628.00 | 1.33802e+05 | 209151 | 861206.0 |
| TRAVTIME | 5.00 | 24.00 | 34.00 | 3.47700e+01 | 45 | 97.0 |
| BLUEBOOK | 1500.00 | 7758.00 | 12590.00 | 1.42550e+04 | 19218 | 62240.0 |
| TIF | 1.00 | 1.00 | 4.00 | 4.78100e+00 | 7 | 21.0 |
| OLDCLAIM | 0.00 | 0.00 | 2432.00 | 6.04900e+03 | 6905 | 57037.0 |
| CLM_FREQ | 0.00 | 0.00 | 1.00 | 1.21600e+00 | 2 | 5.0 |
| MVR_PTS | 0.00 | 0.00 | 2.00 | 2.48200e+00 | 4 | 13.0 |
| CAR_AGE | 0.00 | 1.00 | 7.00 | 7.82200e+00 | 12 | 28.0 |
| IS_SINGLE_PARENT | 0.00 | 0.00 | 0.00 | 2.21200e-01 | 0 | 1.0 |
| IS_MARRIED | 0.00 | 0.00 | 0.00 | 4.89300e-01 | 1 | 1.0 |
| IS_FEMALE | 0.00 | 0.00 | 1.00 | 5.53400e-01 | 1 | 1.0 |
| EDUCATIONBachelors | 0.00 | 0.00 | 0.00 | 2.42600e-01 | 0 | 1.0 |
| EDUCATIONMasters | 0.00 | 0.00 | 0.00 | 1.52000e-01 | 0 | 1.0 |
| EDUCATIONPhD | 0.00 | 0.00 | 0.00 | 5.80900e-02 | 0 | 1.0 |
| EDUCATIONz_High.School | 0.00 | 0.00 | 0.00 | 3.68500e-01 | 1 | 1.0 |
| JOBClerical | 0.00 | 0.00 | 0.00 | 1.72400e-01 | 0 | 1.0 |
| JOBDoctor | 0.00 | 0.00 | 0.00 | 1.34800e-02 | 0 | 1.0 |
| JOBHome.Maker | 0.00 | 0.00 | 0.00 | 8.36400e-02 | 0 | 1.0 |
| JOBLawyer | 0.00 | 0.00 | 0.00 | 7.11000e-02 | 0 | 1.0 |
| JOBManager | 0.00 | 0.00 | 0.00 | 6.36600e-02 | 0 | 1.0 |
| JOBProfessional | 0.00 | 0.00 | 0.00 | 1.14300e-01 | 0 | 1.0 |
| JOBStudent | 0.00 | 0.00 | 0.00 | 1.23600e-01 | 0 | 1.0 |
| JOBz_Blue.Collar | 0.00 | 0.00 | 0.00 | 2.94600e-01 | 1 | 1.0 |
| IS_CAR_PRIVATE_USE | 0.00 | 0.00 | 1.00 | 5.13500e-01 | 1 | 1.0 |
| CAR_TYPEPanel.Truck | 0.00 | 0.00 | 0.00 | 8.27100e-02 | 0 | 1.0 |
| CAR_TYPEPickup | 0.00 | 0.00 | 0.00 | 2.05400e-01 | 0 | 1.0 |
| CAR_TYPESports.Car | 0.00 | 0.00 | 0.00 | 1.41300e-01 | 0 | 1.0 |
| CAR_TYPEVan | 0.00 | 0.00 | 0.00 | 9.34000e-02 | 0 | 1.0 |
| CAR_TYPEz_SUV | 0.00 | 0.00 | 0.00 | 3.15100e-01 | 1 | 1.0 |
| IS_CAR_RED | 0.00 | 0.00 | 0.00 | 2.86200e-01 | 1 | 1.0 |
| IS_LIC_REVOKED | 0.00 | 0.00 | 0.00 | 2.05400e-01 | 0 | 1.0 |
| IS_URBAN | 0.00 | 1.00 | 1.00 | 9.46600e-01 | 1 | 1.0 |

**Transformations**

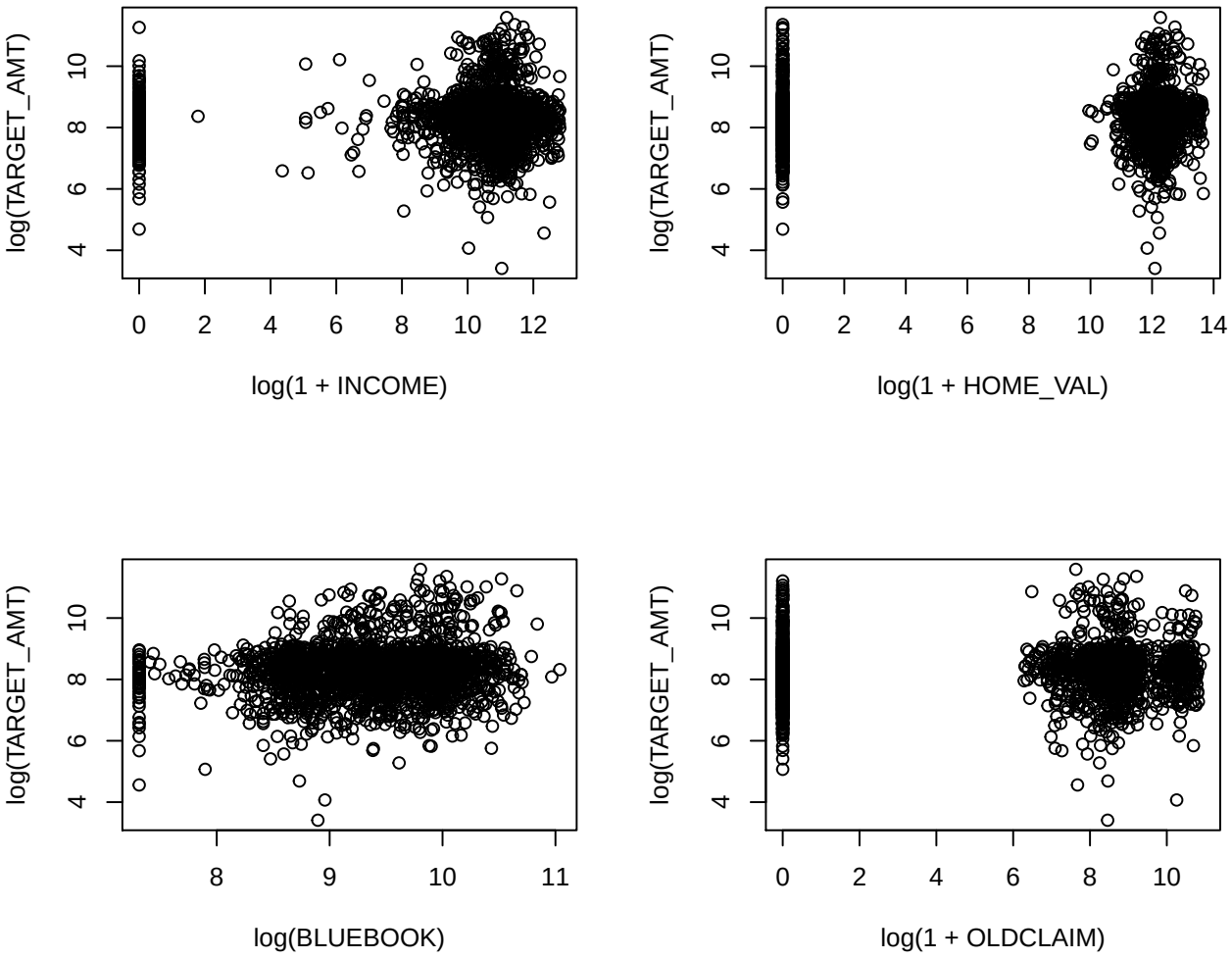
Notice how some variables present 0 as minimum input; that is **INCOME, HOME_VAL** and **OLDCLAIM**. This is very important since I will perform some transformations as follows:

- **TARGET_AMT**: will be transformed to **log(TARGET_AMT)**.

- **INCOME**: will be transformed to **log(1 + INCOME)** <- To avoid $log(0)$ problem.
- **HOME_VAL**: will be transformed to **log(1 + HOME_VAL)** <- To avoid $log(0)$ problem.
- **BLUEBOOK**: will be transformed to **log(BLUEBOOK)**.
- **OLDCLAIM**: will be transformed to **log(1 + OLDCLAIM)** <- To avoid $log(0)$ problem.

**Visualizations**

Let's see if could find some linear relationships in terms of linearity among `TARGET_AMT` vs other variables.




From the above graphs, we could seems to identify some sort of linearity in the given data set, also, we notice how the **log(1 + VARIABLE)** has some effect in the plots.

**Leverage and Outliers**

In this section, I will try to identify and build a list of Leverage points alongside outliers.

**Multiple Regression NULL Model**

Let's start with a null model in order to start having a better understanding. This model will be considered to be valid and will be considered as we advance.

```
##
## Call:
## lm(formula = log_TARGET_AMT ~ 1, data = TARGET_FLAG_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8660 -0.4091  0.0443  0.3871  3.3096
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.27644    0.01751   472.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8125 on 2151 degrees of freedom
```

**Multiple Regression FULL Model**

In this section, I will build a FULL model, thus in order to keep having a better understanding of the model. This model will be considered to be valid and will be considered as we advance.

```
##
## Call:
## lm(formula = log_TARGET_AMT ~ ., data = TARGET_FLAG_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0505 -0.3791  0.0624  0.4248  2.9125
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.6219045  0.3637622  18.204  < 2e-16 ***
## TARGET_FLAG                    NA         NA      NA       NA
## KIDSDRIV               -0.0301002  0.0329426  -0.914 0.360970
## AGE                     0.0008240  0.0021749   0.379 0.704840
## HOMEKIDS                0.0136086  0.0215630   0.631 0.528037
## YOJ                    -0.0001170  0.0053961  -0.022 0.982707
## TRAVTIME               -0.0003177  0.0011534  -0.275 0.783026
## TIF                    -0.0011951  0.0044208  -0.270 0.786927
## CLM_FREQ               -0.0530680  0.0242192  -2.191 0.028549 *
## MVR_PTS                 0.0139056  0.0072827   1.909 0.056345 .
## CAR_AGE                -0.0007791  0.0037051  -0.210 0.833478
## IS_SINGLE_PARENT        0.0232964  0.0611578   0.381 0.703299
## IS_MARRIED             -0.1201347  0.0526775  -2.281 0.022673 *
## IS_FEMALE              -0.0859461  0.0658086  -1.306 0.191694
```

```
## EDUCATIONBachelors     -0.0649046  0.0647233  -1.003 0.316072
## EDUCATIONMasters        0.1029520  0.1090695   0.944 0.345323
## EDUCATIONPhD            0.1537568  0.1269872   1.211 0.226106
## EDUCATIONz_High.School  0.0059409  0.0535503   0.111 0.911673
## JOBClerical             0.0743435  0.1231961   0.603 0.546270
## JOBDoctor              -0.0131399  0.1839070  -0.071 0.943047
## JOBHome.Maker           0.0303227  0.1325648   0.229 0.819094
## JOBLawyer              -0.0354852  0.1069424  -0.332 0.740061
## JOBManager              0.0208008  0.1108565   0.188 0.851179
## JOBProfessional         0.1130966  0.1170222   0.966 0.333928
## JOBStudent              0.1260001  0.1361558   0.925 0.354858
## JOBz_Blue.Collar        0.0564755  0.1186269   0.476 0.634069
## IS_CAR_PRIVATE_USE      0.0136397  0.0543605   0.251 0.801907
## CAR_TYPEPanel.Truck     0.0368181  0.0916320   0.402 0.687869
## CAR_TYPEPickup          0.0376195  0.0620571   0.606 0.544441
## CAR_TYPESports.Car      0.0718072  0.0765241   0.938 0.348167
## CAR_TYPEVan            -0.0187183  0.0791974  -0.236 0.813184
## CAR_TYPEz_SUV           0.0954950  0.0669215   1.427 0.153736
## IS_CAR_RED              0.0359232  0.0516642   0.695 0.486931
## IS_LIC_REVOKED         -0.0337374  0.0440626  -0.766 0.443958
## IS_URBAN                0.0473326  0.0785668   0.602 0.546939
## log_INCOME             -0.0011698  0.0090670  -0.129 0.897352
## log_HOME_VAL            0.0056977  0.0039472   1.443 0.149033
## log_BLUEBOOK            0.1578179  0.0342402   4.609 4.28e-06 ***
## log_OLDCLAIM            0.0114251  0.0070581   1.619 0.105657
## INCOME_Outlier         -1.2126046  0.4517518  -2.684 0.007327 **
## HOME_VAL_Outlier        0.7621985  0.6101766   1.249 0.211751
## BLUEBOOK_Outlier        0.7088887  0.1902301   3.726 0.000199 ***
## OLDCLAIM_Outlier        0.1474221  0.6087918   0.242 0.808683
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7996 on 2110 degrees of freedom
## Multiple R-squared:  0.0499, Adjusted R-squared:  0.03144
## F-statistic: 2.703 on 41 and 2110 DF,  p-value: 4.402e-08
```

Interesting to see only a few statistically significant variables while the $R^2$ shows to be low. The p-value is very low and the median is considered to be near zero.

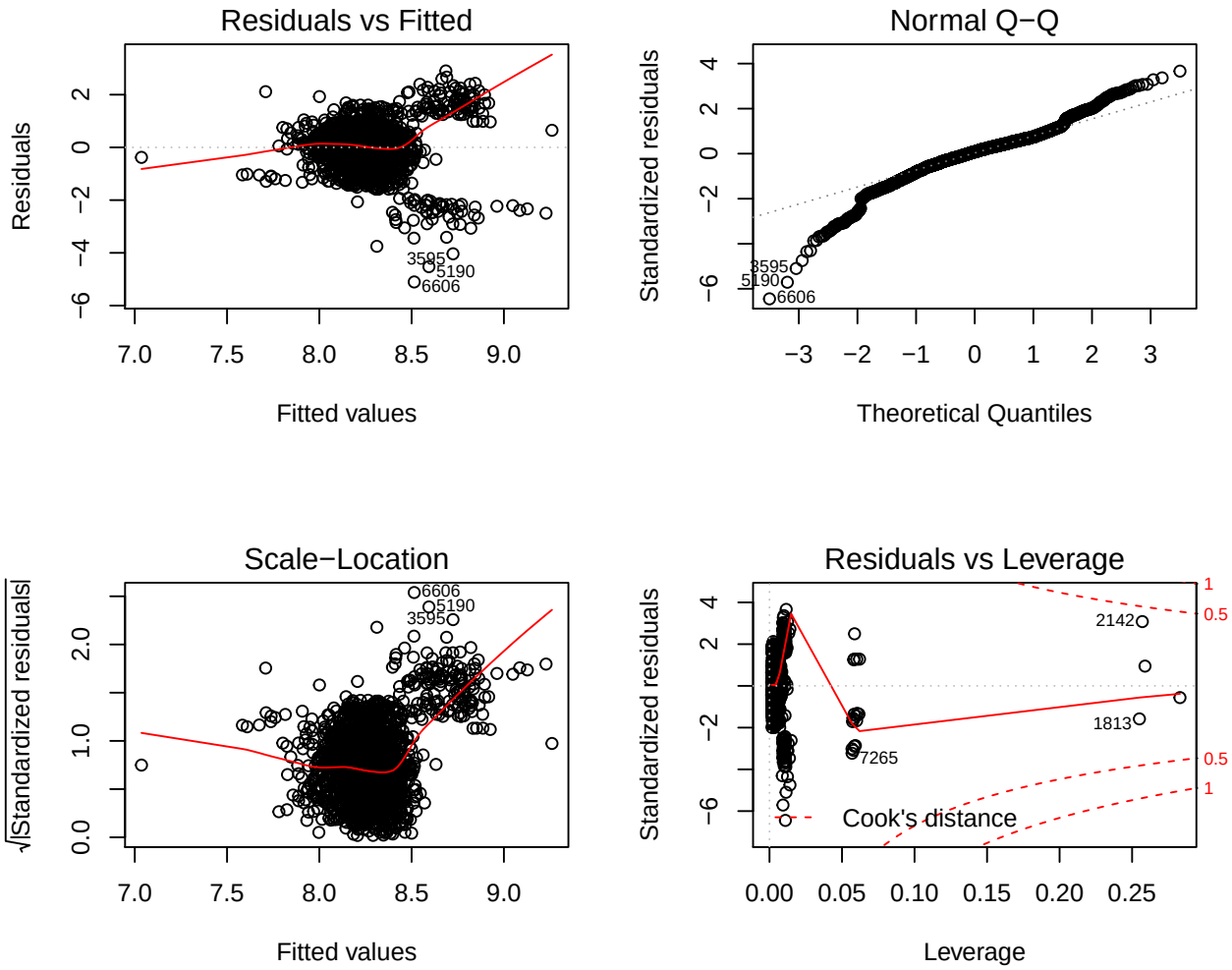
### Multiple Regression STEP Model

In this section, I will build a model by employing the STEP function from R, thus in order to keep having a better understanding of the model. This model will be considered to be valid and will be considered as we advance.

```
##
## Call:
## lm(formula = log_TARGET_AMT ~ BLUEBOOK_Outlier + log_BLUEBOOK +
```

```
##     IS_MARRIED + INCOME_Outlier + HOME_VAL_Outlier + MVR_PTS +
##     EDUCATIONBachelors + IS_FEMALE + CLM_FREQ, data = TARGET_FLAG_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1036 -0.3798  0.0669  0.4300  2.9004
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.955403   0.250897  27.722  < 2e-16 ***
## BLUEBOOK_Outlier     0.712254   0.188213   3.784 0.000158 ***
## log_BLUEBOOK         0.145953   0.026226   5.565 2.95e-08 ***
## IS_MARRIED          -0.081364   0.034463  -2.361 0.018319 *
## INCOME_Outlier      -1.182167   0.422567  -2.798 0.005195 **
## HOME_VAL_Outlier     0.874650   0.403540   2.167 0.030311 *
## MVR_PTS              0.016472   0.006979   2.360 0.018352 *
## EDUCATIONBachelors -0.076328   0.040307  -1.894 0.058403 .
## IS_FEMALE           -0.053820   0.034684  -1.552 0.120874
## CLM_FREQ            -0.021031   0.014426  -1.458 0.145046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7961 on 2142 degrees of freedom
## Multiple R-squared:  0.04399,    Adjusted R-squared:  0.03997
## F-statistic: 10.95 on 9 and 2142 DF,  p-value: < 2.2e-16
```

Something interesting from the above results is that it shows non statistical significant values as part of the model. That is IS_FEMALE and CLM_FREQ are not statistically significant.




From the above graphs, we can quickly identify the **lobster shape** figure, which indicates that this model seems to be some how appropriate due to being some how homoscedastic, plus the Normal Q-Q line seems to differ on the lower end but not by much on the upper end which will be more problematic due to the nature of paying out insurance money.

**ANOVA results**

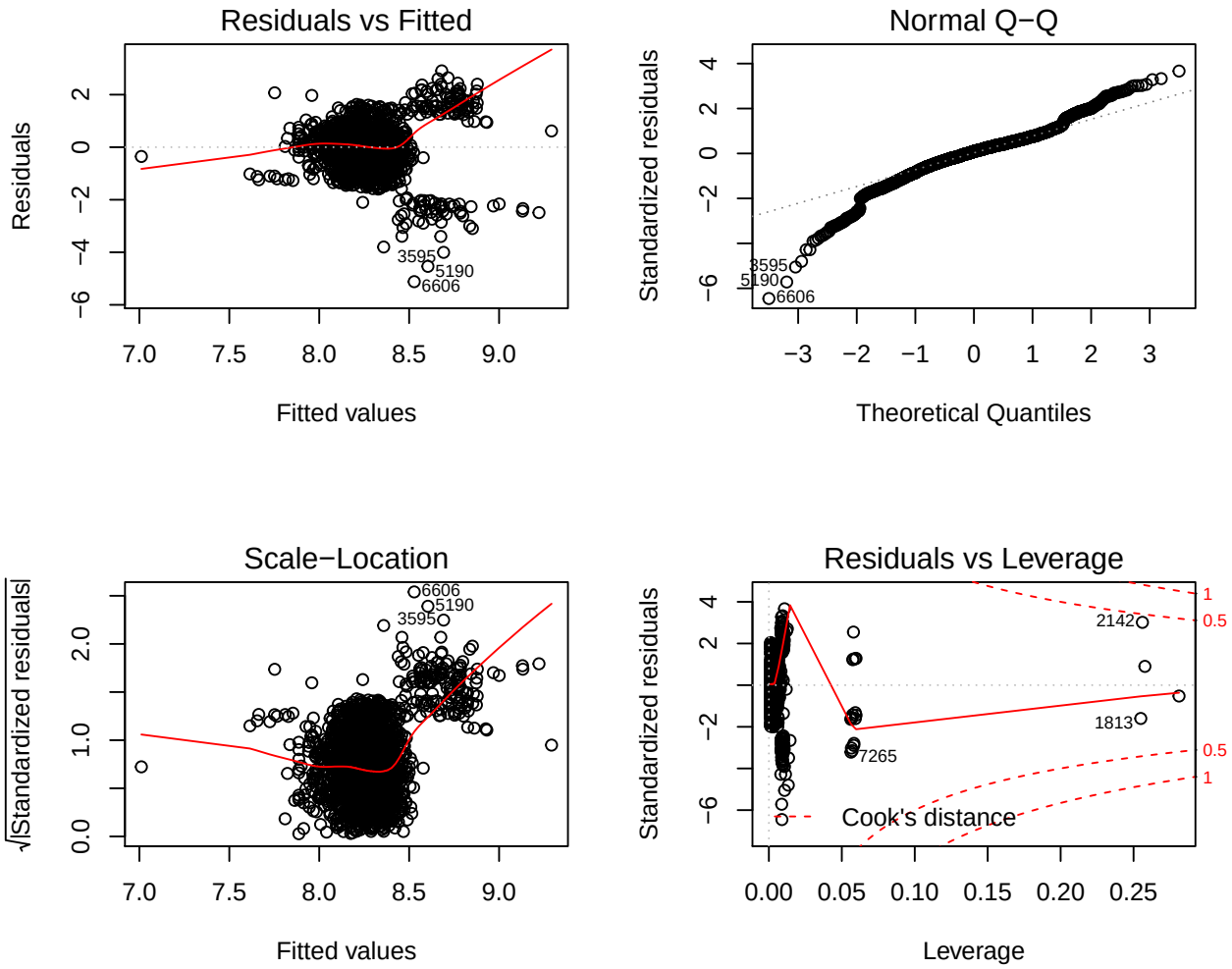
Let's check an ANOVA table based on the above testing results.

| Step | Df | Deviance | Resid. Df | Resid. Dev | AIC |
|---|---|---|---|---|---|
| | NA | NA | 2151 | 1419.904 | -892.8199 |
| + BLUEBOOK_Outlier | -1 | 24.241941 | 2150 | 1395.662 | -927.8781 |
| + log_BLUEBOOK | -1 | 21.713173 | 2149 | 1373.949 | -959.6212 |
| + IS_MARRIED | -1 | 2.955640 | 2148 | 1370.993 | -962.2556 |
| + INCOME_Outlier | -1 | 2.577242 | 2147 | 1368.416 | -964.3048 |
| + HOME_VAL_Outlier | -1 | 3.299849 | 2146 | 1365.116 | -967.5005 |
| + MVR_PTS | -1 | 2.493846 | 2145 | 1362.622 | -969.4355 |
| + EDUCATIONBachelors | -1 | 2.281735 | 2144 | 1360.340 | -971.0420 |
| + IS_FEMALE | -1 | 1.546264 | 2143 | 1358.794 | -971.4896 |
| + CLM_FREQ | -1 | 1.346765 | 2142 | 1357.447 | -971.6236 |

**Multiple Regression STEP MODIFIED Model**

In this section, I will create a manual model in order to try to overcome the previous identify problems. In this case, I will add an iteration `AGE:IS_FEMALE` and I will keep the statistically significant values provided above.

```
##
## Call:
## lm(formula = log_TARGET_AMT ~ BLUEBOOK_Outlier + log_BLUEBOOK +
##     IS_MARRIED + INCOME_Outlier + HOME_VAL_Outlier + MVR_PTS +
##     EDUCATIONBachelors, data = TARGET_FLAG_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1177 -0.3756  0.0705  0.4232  2.9049
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.873667   0.247386  27.785  < 2e-16 ***
## BLUEBOOK_Outlier    0.707848   0.188193   3.761 0.000174 ***
## log_BLUEBOOK        0.149567   0.026142   5.721 1.21e-08 ***
## IS_MARRIED         -0.080899   0.034480  -2.346 0.019052 *
## INCOME_Outlier     -1.206908   0.422566  -2.856 0.004330 **
## HOME_VAL_Outlier    0.903923   0.403548   2.240 0.025197 *
## MVR_PTS             0.013377   0.006664   2.008 0.044817 *
## EDUCATIONBachelors -0.076464   0.040322  -1.896 0.058047 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7965 on 2144 degrees of freedom
## Multiple R-squared:  0.04195,    Adjusted R-squared:  0.03882
## F-statistic: 13.41 on 7 and 2144 DF,  p-value: < 2.2e-16
```




Something interesting to note is that in this case, all predictors are statistically significant, the Normal Q-Q plot seems to follow a line for almost all the values with only a slight overpriced on the top but with an under prediction on the bottom left. The Residuals vs the Fitted lobster shape figure seems to be homoscedastic, my only concern will be the Multiple $R^2$ which is very low but I will consider this to be OK due to previous correlations showed that the correlations are very low as well.

### Multiple Linear Regression MODEL SELECTION

Below, I will describe why I have chosen the **Multiple Regression STEP MODIFIED Model** to be my selected model for the multiple linear regression model.

```
lm_Model_FINAL <- lm_Model_STEP_Modified
```

- The generated ANOVA table shows this combination of variables to be have lowest AIC.
- The coefficients make sense.
- The Median is near Zero.
- The coefficients are considered low, alongside the standard errors as well.
- The Residuals vs the Fitted values seems to be homoscedastic.

- The residuals and the normal Q-Q plot also make sense and follow "good" standards for data analysis.
- My only concerns will be the $R^2$ to be too low but it was noted the low correlation among variables.

**Predictions**

In this section, I will proceed to predict values from the evaluation data set.

**Evaluation data transformations**

In this section I will transform our evaluation data same as our original data has.

**Predict TARGET_FLAG**

In this section, I will predict the probability of having an accident or no accident and categorize it in the TARGET_FLAG variable.

**Accident Predictions Table**

In this section, I will predict the values on the **evaluation** data set employing the **training** data set.

Let's see a table for the first 20 records.

| INDEX | TARGET_FLAG |
|---|---|
| 3 | 0 |
| 9 | 0 |
| 10 | 0 |
| 18 | 0 |
| 21 | 0 |
| 30 | 0 |
| 31 | 0 |
| 37 | 0 |
| 39 | 0 |
| 47 | 0 |
| 60 | 0 |
| 62 | 1 |
| 63 | 1 |
| 64 | 0 |
| 68 | 0 |
| 75 | 1 |
| 76 | 1 |
| 83 | 0 |
| 87 | 1 |
| 92 | 0 |

**Predict TARGET_AMT**

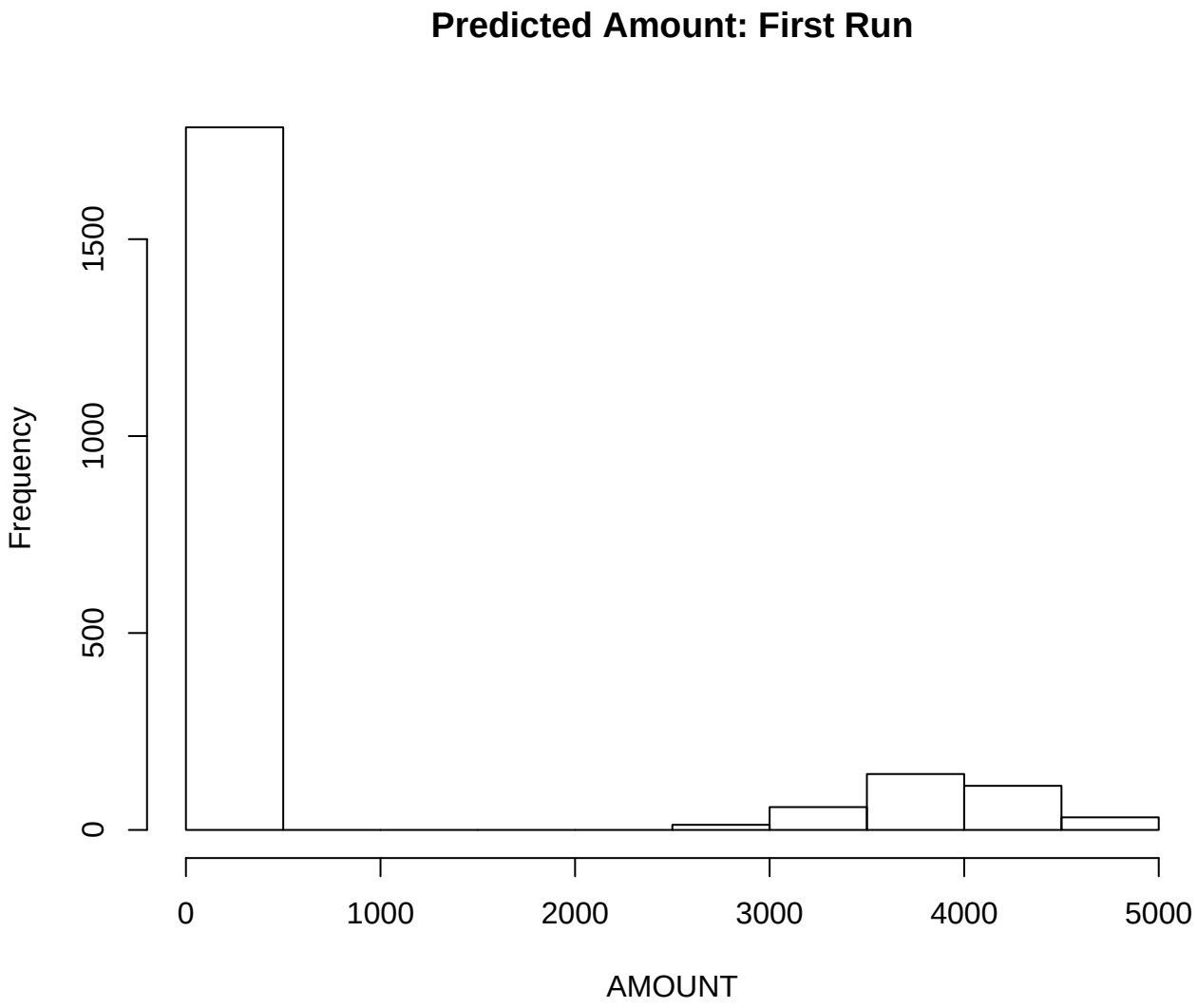
In this section, I will predict the amount based on the final linear model selected. In order to accomplish this goal, I need to do as follows:

Since there's no way to indicate if the new values are considered outliers or not, I will assign a zero instead; then I will re-evaluate with the given values.

- data.eval$BLUEBOOK_Outlier <- 0
- data.eval$INCOME_Outlier <- 0
- data.eval$HOME_VAL_Outlier <- 0

Let's see the first 20 records of the first run for the predicted data set.

| INDEX | TARGET_FLAG | TARGET_AMT |
|---|---|---|
| 3 | 0 | 0.00 |
| 9 | 0 | 0.00 |
| 10 | 0 | 0.00 |
| 18 | 0 | 0.00 |
| 21 | 0 | 0.00 |
| 30 | 0 | 0.00 |
| 31 | 0 | 0.00 |
| 37 | 0 | 0.00 |
| 39 | 0 | 0.00 |
| 47 | 0 | 0.00 |
| 60 | 0 | 0.00 |
| 62 | 1 | 4160.33 |
| 63 | 1 | 3794.61 |
| 64 | 0 | 0.00 |
| 68 | 0 | 0.00 |
| 75 | 1 | 4026.46 |
| 76 | 1 | 2961.95 |
| 83 | 0 | 0.00 |
| 87 | 1 | 3583.60 |
| 92 | 0 | 0.00 |

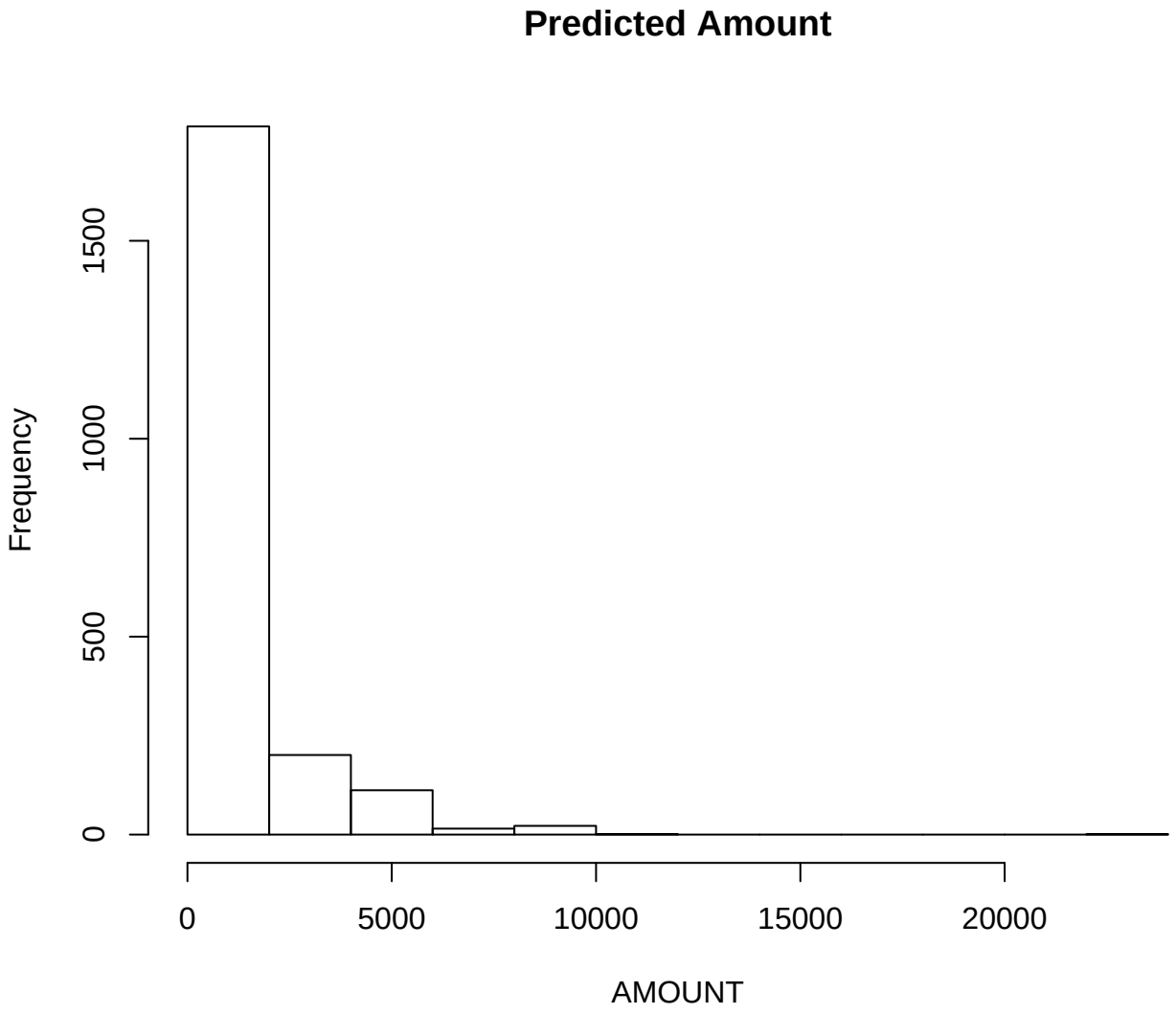



Now, the above values were calculated assuming that all outliers were ZERO. Let's recalculate and see if new Outliers can be found in order to refine our values.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   646.2     0.0  4932.3
```

Let's see the first 40 records of the final predicted data set.

| INDEX | TARGET_FLAG | TARGET_AMT |
|---|---|---|
| 3 | 0 | 0.00 |
| 9 | 0 | 0.00 |
| 10 | 0 | 0.00 |
| 18 | 0 | 0.00 |
| 21 | 0 | 0.00 |
| 30 | 0 | 0.00 |
| 31 | 0 | 0.00 |
| 37 | 0 | 0.00 |
| 39 | 0 | 0.00 |
| 47 | 0 | 0.00 |
| 60 | 0 | 0.00 |
| 62 | 1 | 4160.33 |
| 63 | 1 | 3794.61 |
| 64 | 0 | 0.00 |
| 68 | 0 | 0.00 |
| 75 | 1 | 4026.46 |
| 76 | 1 | 2187.72 |
| 83 | 0 | 0.00 |
| 87 | 1 | 3583.60 |
| 92 | 0 | 0.00 |
| 98 | 0 | 0.00 |
| 106 | 0 | 0.00 |
| 107 | 0 | 0.00 |
| 113 | 0 | 0.00 |
| 120 | 0 | 0.00 |
| 123 | 0 | 0.00 |
| 125 | 0 | 0.00 |
| 126 | 0 | 0.00 |
| 128 | 0 | 0.00 |
| 129 | 0 | 0.00 |
| 131 | 0 | 0.00 |
| 135 | 0 | 0.00 |
| 141 | 0 | 0.00 |
| 147 | 0 | 0.00 |
| 148 | 0 | 0.00 |
| 151 | 0 | 0.00 |
| 156 | 0 | 0.00 |
| 157 | 0 | 0.00 |
| 174 | 0 | 0.00 |
| 186 | 1 | 3918.28 |




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   719.6     0.0 23918.2
```

**Export file**

In order to provide a csv output for the predictions table.

```
write.csv(data.eval_ORIGINAL, file = "insurance-my-evaluated-data.csv",row.names=FALSE)
```

**Conclusion**

In the above example, we can comprehend the importance of understanding the data in order to provide meaningful results. Not all data sets are alike and different approaches need to be taken in order to extract valuable information out of it.

## References

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.