Critical Thinking Group 4 - HW4 - Auto Insurance

Sreejaya, Suman, Vuthy

November 7, 2016

Overview

The purpose of this project is to predict the probability that a person will crash their car and the amount of money it will cost if the person does crash their car using multiple linear regression and binary logistic regression models.

Dataset
Insurance - Training data
Insurance - Evaluation Data

Below is a short description of the variables in the dataset.

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 Clamins (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

Data Exploration

The dataset contains roughly 8000 observations and 26 variables. Each record has two response variables.

- TARGET_FLAG: Was Car in a crash? 1=YES 0=NO

- TARGET_AMT: If car was in a crash, what was the cost

Running the glimpse() function on the dataset reveals a few issues with the data.

```
## Observations: 8,161
## Variables: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 1...
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0,...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000...
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55...
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0,...
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, ...
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "...
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "...
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,92...
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes"...
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z...
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High Sc...
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Bl...
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25,...
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Pr...
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,00...
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6...
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV",...
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes"...
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", ...
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0,...
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", ...
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0...
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5,...
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Hig...
```

Data Clean Up

Identify Data Elements For Clean up

- INCOME: Need to remove characters like '$', ','
- HOME_VAL: Need to remove characters like '$', ','
- MSTATUS: Remove prefix 'z_'
- SEX: Remove prefix 'z_'
- EDUCATION: Remove prefix 'z_'
- JOB: Remove prefix 'z_'
- BLUEBOOK: Need to remove characters like '$', ','
- CAR_TYPE: Remove prefix 'z_'
- OLDCLAIM: Need to remove characters like '$', ','

Factorization

A few variables need to be converted to factors

- TARGET_FLAG

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

Transformations

Identify variables need to be converted to numeric type & need transformations

- INCOME
- HOME_VAL
- BLUEBOOK
- OLDCLAIM - This is for 5 years, we would need to mutate the dataset to have an average per year.

As with the glimpse() function, the summary() function reveals a few variables need to be coerced as well as a few missing values (NA).

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ          INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS
##  Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      SEX             EDUCATION             JOB               TRAVTIME
##  Length:8161        Length:8161        Length:8161        Min.   :  5.00
##  Class :character   Class :character   Class :character   1st Qu.: 22.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 33.00
##                                                           Mean   : 33.49
##                                                           3rd Qu.: 44.00
```

```
##                                                       Max.   :142.00
##
##    CAR_USE            BLUEBOOK              TIF           CAR_TYPE
##  Length:8161        Length:8161        Min.   : 1.000   Length:8161
##  Class :character   Class :character   1st Qu.: 1.000   Class :character
##  Mode  :character   Mode  :character   Median : 4.000   Mode  :character
##                                        Mean   : 5.351
##                                        3rd Qu.: 7.000
##                                        Max.   :25.000
##
##    RED_CAR            OLDCLAIM            CLM_FREQ        REVOKED
##  Length:8161        Length:8161        Min.   :0.0000   Length:8161
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.7986
##                                        3rd Qu.:2.0000
##                                        Max.   :5.0000
##
##     MVR_PTS          CAR_AGE        URBANICITY
##  Min.   : 0.000   Min.   :-3.000   Length:8161
##  1st Qu.: 0.000   1st Qu.: 1.000   Class :character
##  Median : 1.000   Median : 8.000   Mode  :character
##  Mean   : 1.696   Mean   : 8.328
##  3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :13.000   Max.   :28.000
##                   NA's   :510
```

MISSING VALUES

The following variables have missing values

Variable Name	# NA
AGE	6
CAR_AGE	510
YOJ	454

Dummy Variables

These variables are identified as categorical, and would need to be converted into Dummy Variables for Model buidling

- CAR USE: Commercial vehicles are driven more, so might increase probability of collision
- EDUCATION: In theory more educated people tend to drive more safely
- JOB CATEGORY: 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
- RED_CAR: Urban legend says that red cars (especially red sports cars) are more risky
- 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.
- URBANICITY: Urban Vs Rural

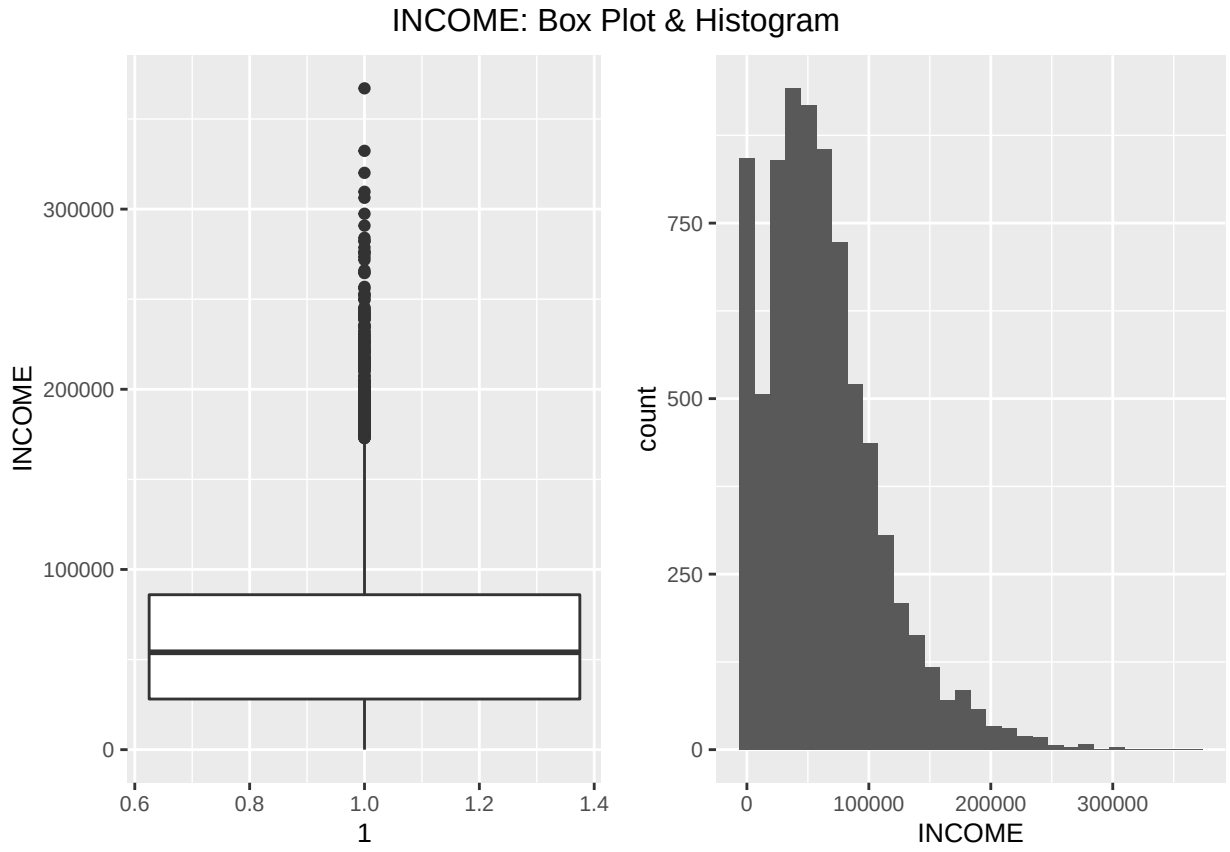
Normality

Lets check the normality in the given numeric predictors

Since the amounts got dollar signs and commas, lets perform a quick clean up here, before we start exploring the data:

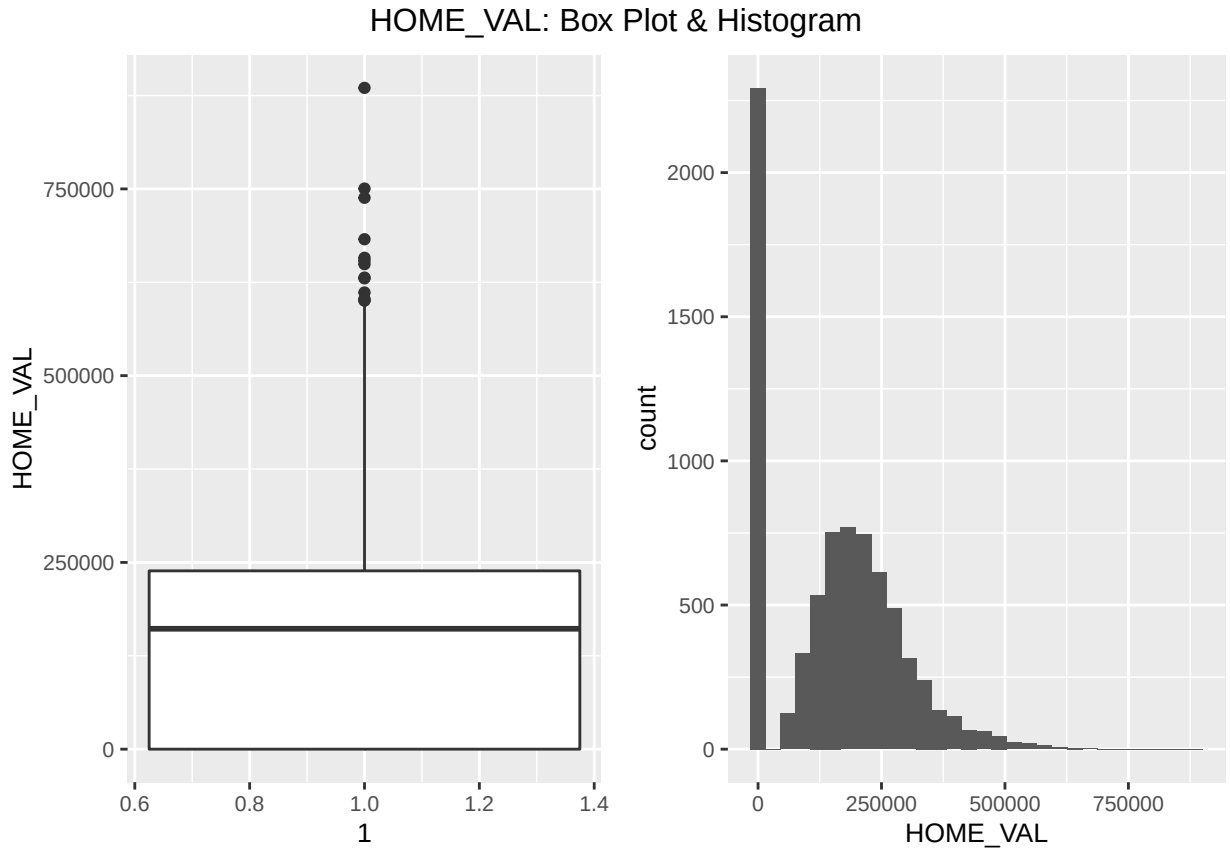
Clean up the variables: **INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM**

INCOME :

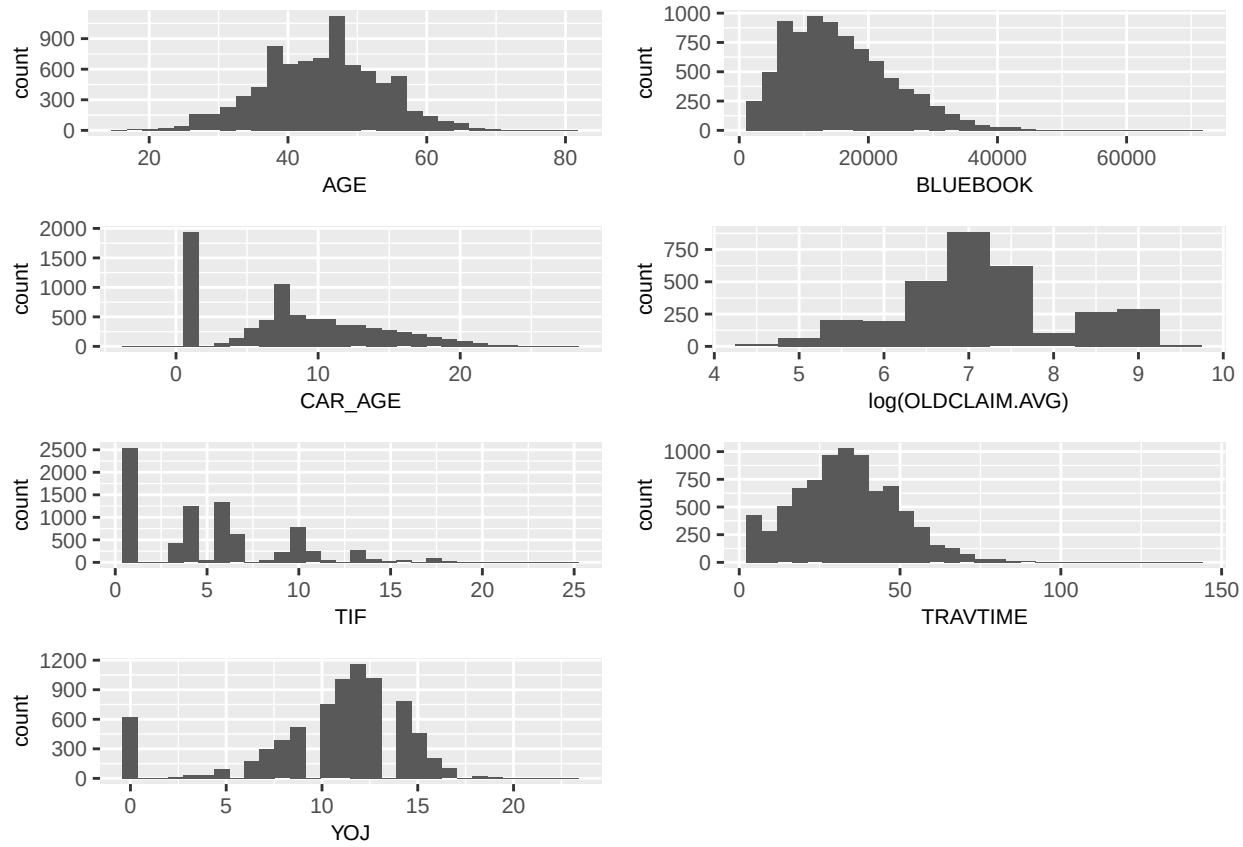


The above box plot and histogram clearly indicates that the INCOME data is positively skewed. We need to address this in *DATA PREPARATION* section.

HOME_VAL :



The above box plot and histogram clearly indicates that the HOME_VAL data is positively skewed. We need to address this in *DATA PREPARATION* section.



From the above histogrames, the 'BLUE BOOK', 'TIF' distributions appears to be positively skewed. We modified the OLDCLAIM variable be the average per year and applied log to make the data distribution nearly normal. We still could see a slight bi-model in it, but we will see how it affect our models.

Data Preparation

Remove unnecessory variables

"INDEX" is nothing more than an arbitrary identifier that provides no predictive value. It will be removed from the data set.

```
insurance.trn$INDEX <- NULL
```

Clean up data

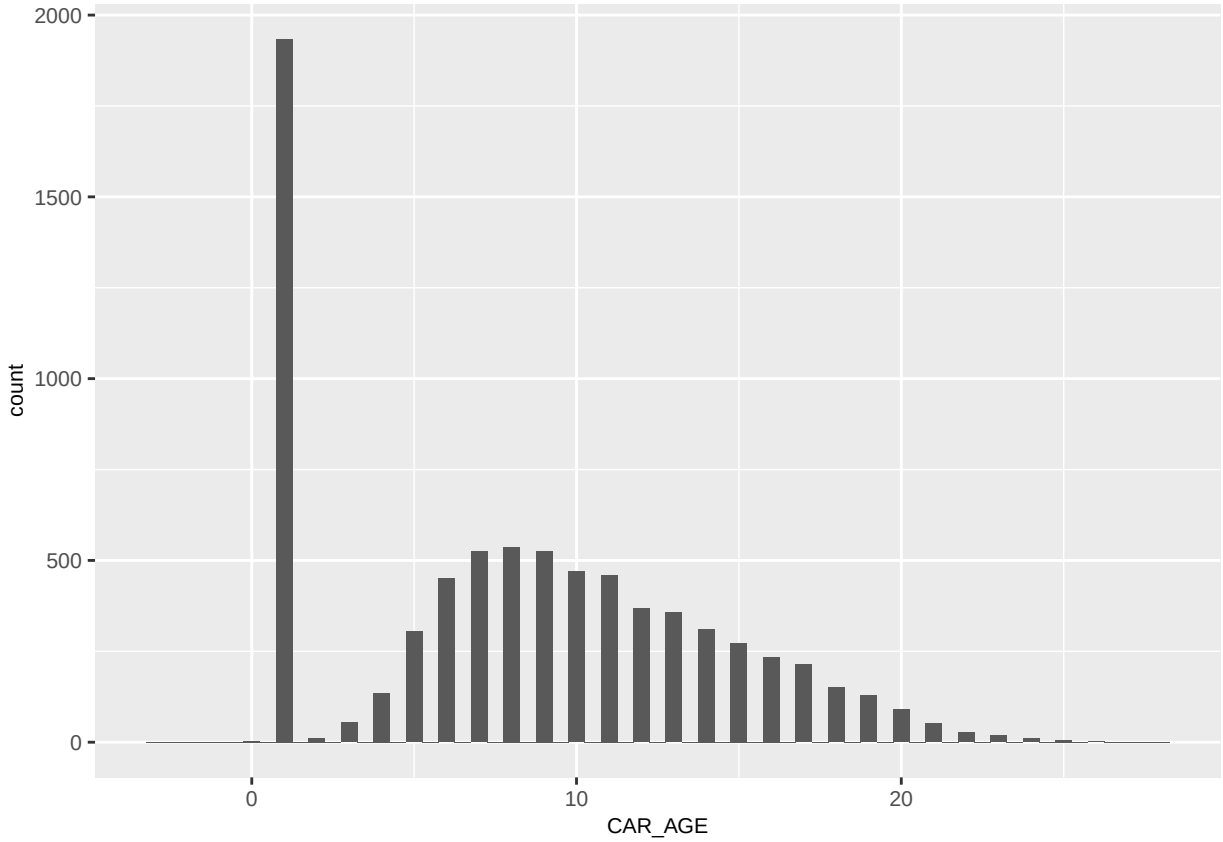
'z_' will be removed from **MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE**

Factorize Variables:

Convert the **TARGET_FLAG, PARENT1, SEX, MSTATUS, EDUCATION, JOB, CAR_USE, CAR_TYPE, RED_CAR, REVOKED, URBANICITY** variables into factors:

Handling missing data

1. CAR_AGE - There are 510 NA's in CAR_AGE, so it is not a good idea to throw these records away.



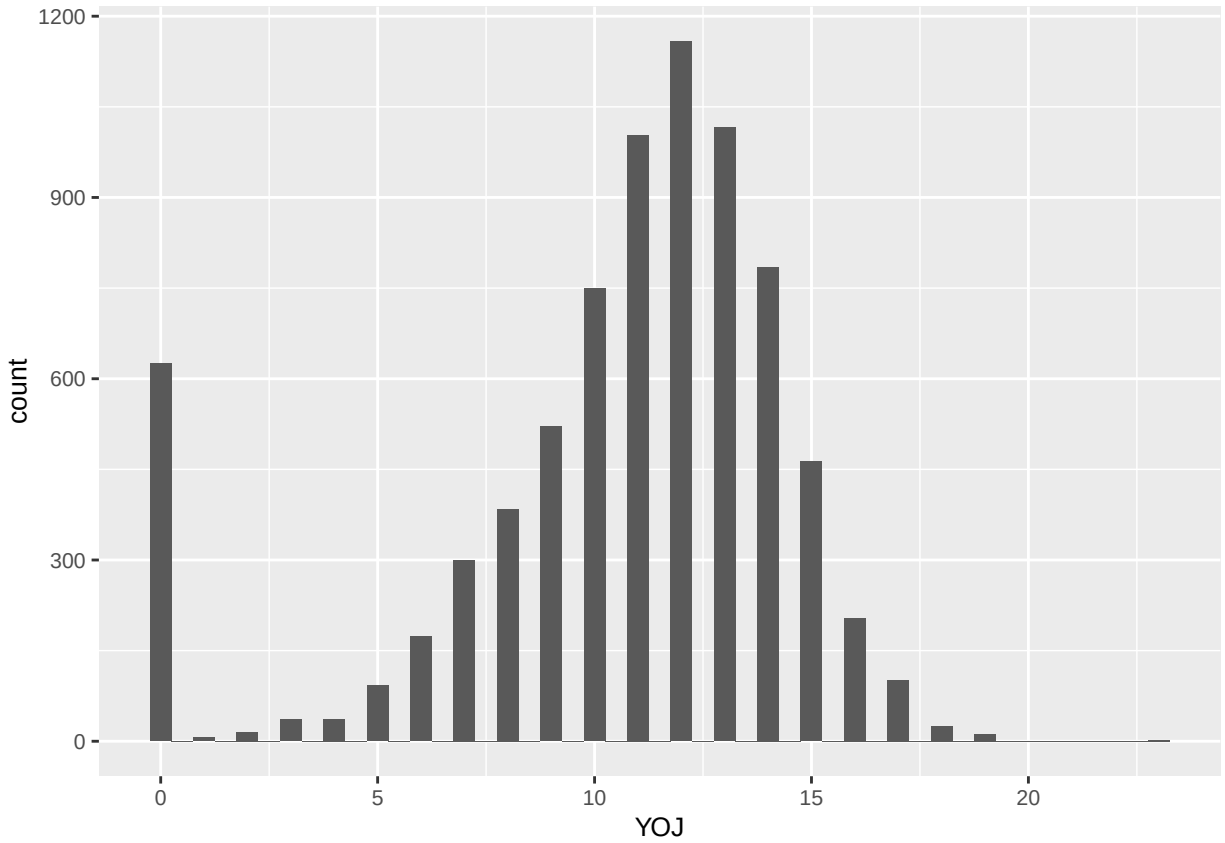
From the histogram, we see that CAR_AGE between 5 and 15 are the most common, But there is one record with negative value, which should be replaced.

so filling in NA's/negative value with median(8) or mean(8.33) would be entirely reasonable. Let's fill in the NA's/negative value with the median value of 8:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   8.000   8.312  12.000  28.000
```

So, as shown above, after we imputing the CAR_AGE, the mean value is consistent/close to the original mean.

2. YOJ - There are 454 NA's in YOJ, so it is not a good idea to throw these records away.

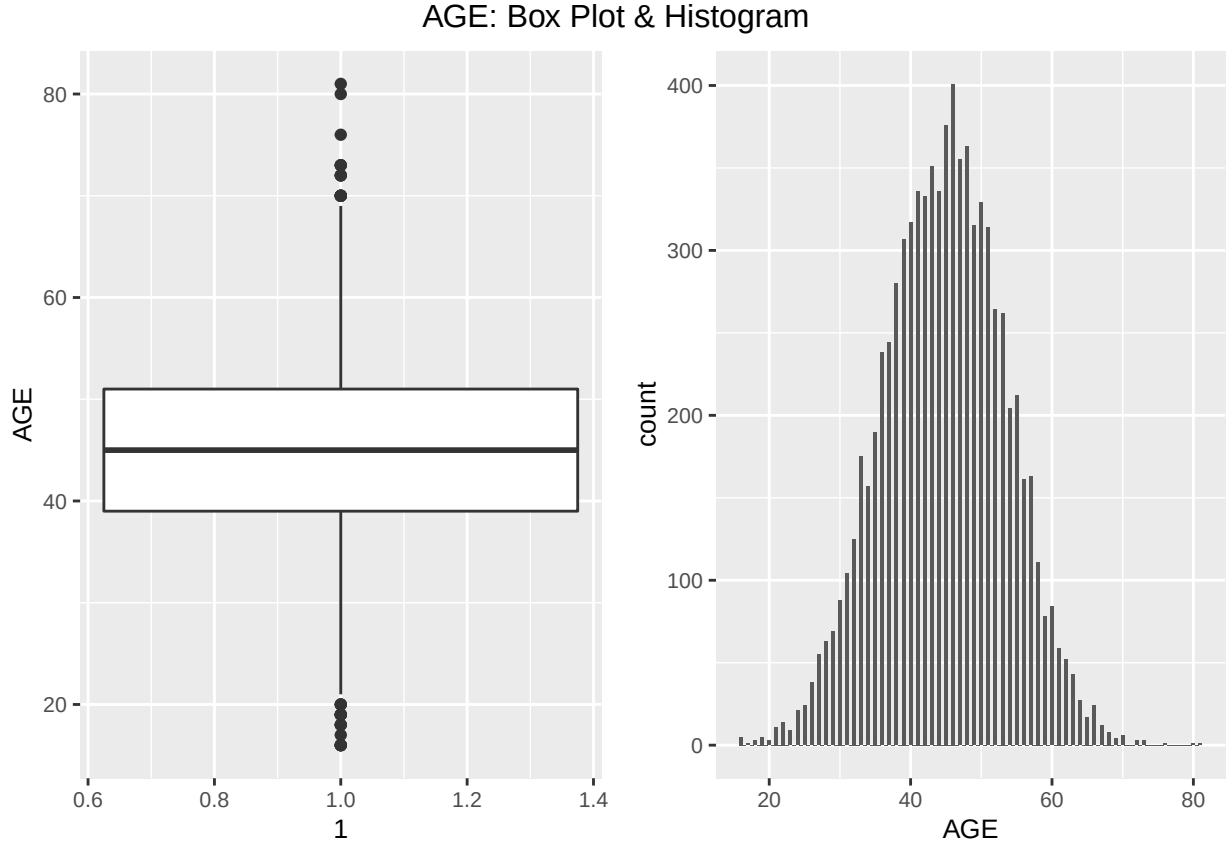


From the histogram, we see that YOJ between 8 and 12 are the most common, so filling in NA's with mean(10.5) or median(11) would be entirely reasonable.

Let's fill in the NA's with the median 11:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.53   13.00   23.00
```

3. AGE - Let us consider outliars for AGE before removing NA's



Since the mean, and median are approximately equal to each other, data is approximately symmetrical and there are no outliars. Mean=44.79 and Median =45. So let's fill in the NA's with the median 45:

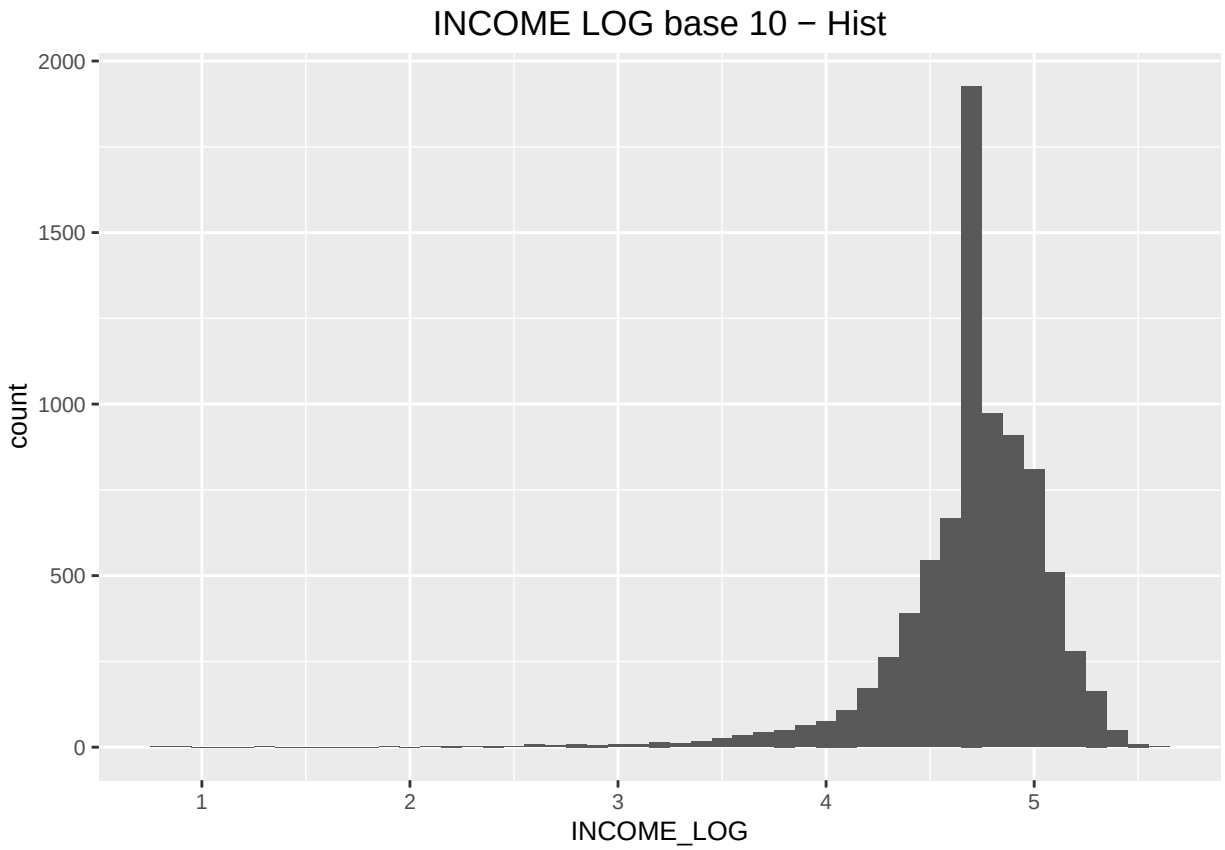
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   39.00   45.00   44.79   51.00   81.00
```

4. INCOME - Transformation

From the DATA EXPLORATION, its clear that the INCOME distribution is positively skewed, So, lets consider transforming it using log base 10.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.365   4.709   4.091   4.921   5.565
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7782  4.5780  4.7090  4.7030  4.9210  5.5650
```



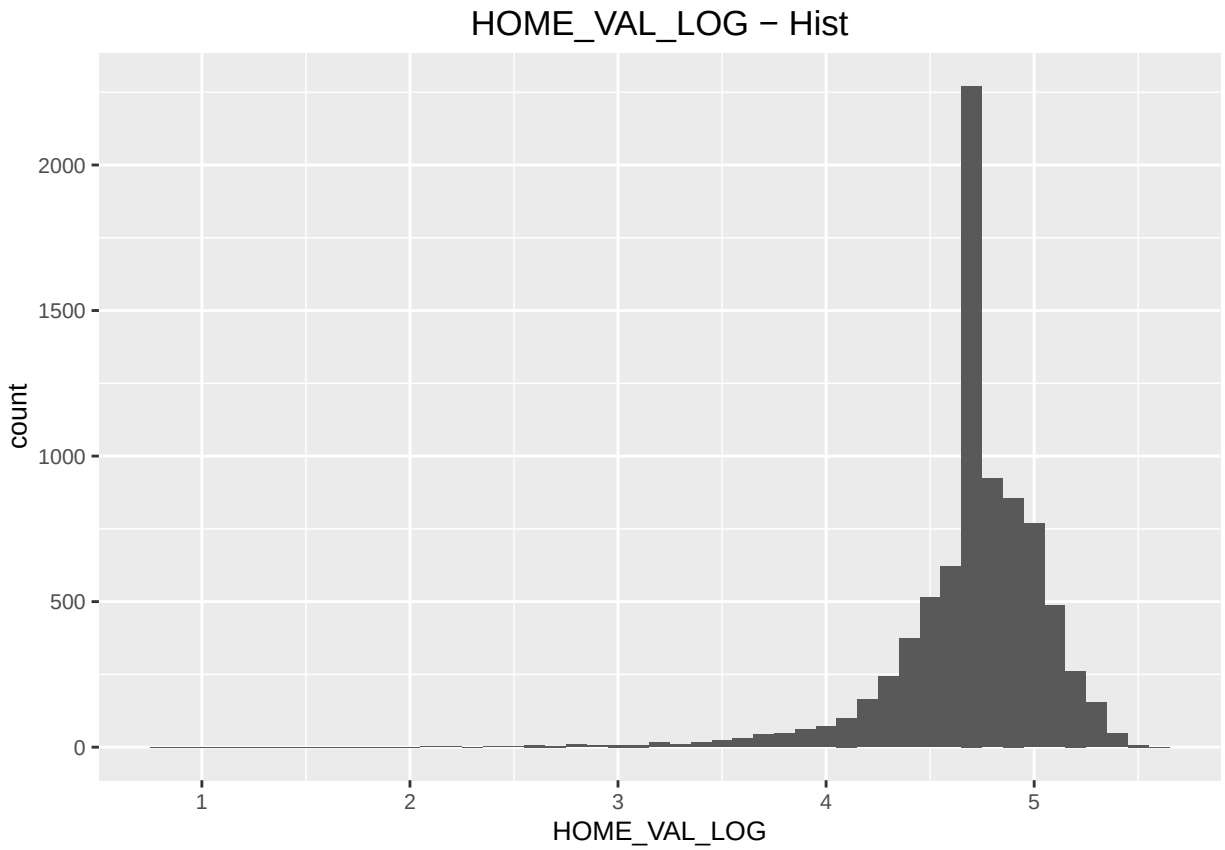
Though there is a slight negative skewness, overall this transformation looks better (nearly normal) than the original distribution.

5. HOME_VAL - Transformation

From the DATA EXPLORATION, its clear that the HOME_VAL distribution is slightly positively skewed, So, lets consider transforming it using log base 10.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.223   4.681   3.864   4.907   5.565
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7782  4.5960  4.6810  4.6980  4.9070  5.5650
```



Though there is a slight negative skewness, overall this transformation looks better (nearly normal) than the original distribution.

Dummy Variables

Look for dummy variables in the categorical variables:

We can chose to make n-1 dummy varaibles, Or, check if response behaviour is similar across frew of these categories and merge them.

CAR USE - Commercial vehicles are driven more, so might increase probability of collision

```
##
##                0    1
##   Commercial 0.65 0.35
##   Private    0.78 0.22
```

EDUCATION - In theory more educated people tend to drive more safely

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

```
##
##                   0    1
##   <High School 0.68 0.32
```

```
##   Bachelors    0.77 0.23
##   High School  0.66 0.34
##   Masters      0.80 0.20
##   PhD          0.83 0.17
```

Lets make a categorical variable of HIGH EDUCATION Vs LOW EDUCATION LEVELS.

JOB CATEGORY - In theory, white collar jobs tend to be safer

Lets check the *JOB* categorical variable:

```
##
##                  0    1
##                0.74 0.26
##   Blue Collar  0.65 0.35
##   Clerical     0.71 0.29
##   Doctor       0.88 0.12
##   Home Maker   0.72 0.28
##   Lawyer       0.82 0.18
##   Manager      0.86 0.14
##   Professional 0.78 0.22
##   Student      0.63 0.37
```

Lets make a categorical variable of WHITE-COLLAR Vs Non-WHITE-COLLAR.

KIDSDRIV - When teenagers drive your car, you are more likely to get into crashes

```
##
##        0    1
##   0 0.75 0.25
##   1 0.63 0.37
##   2 0.60 0.40
##   3 0.50 0.50
##   4 0.50 0.50
```

So,lets split the one or 2 teenagers Vs 3 or 4 teenagers, that drive the car.

MSTATUS - In theory, married people drive more safely

```
##
##          0    1
##   No  0.66 0.34
##   Yes 0.78 0.22
```

RED_CAR - Urban legend says that red cars (especially red sports cars) are more risky. Is that true?

```
##
##          0    1
##   no  0.73 0.27
##   yes 0.74 0.26
```

```
##
##                 0    1
##   Minivan     0.84 0.16
```

```
##   Panel Truck 0.74 0.26
##   Pickup      0.68 0.32
##   Sports Car  0.66 0.34
##   SUV         0.70 0.30
##   Van         0.73 0.27
```

Lets have a category of red sports car to verify the above Urban legend statement.

```
##
##        0    1
##   0 0.74 0.26
##   1 0.57 0.43
```

REVOKED - If your license was revoked in the past 7 years, you probably are a more risky driver.

```
##
##          0    1
##   No  0.76 0.24
##   Yes 0.56 0.44
```

SEX - Urban legend says that women have less crashes then men. Is that true?

Lets just consider adding a dummy variable 'MALE' or not.

```
##
##        0    1
##   F 0.73 0.27
##   M 0.75 0.25
```

The other variables - BLUEBOOK,CAR_AGE,CAR_TYPE , has 'Unknown effect on probability of collision, but probably effect the payout if there is a crash', per data_desc

We will leave the numeric variables BLUEBOOK, CAR_AGE 'as it is', and create the dummy variables for CAR_TYPE, so we would have the CAR_TYPE also numeric.

```
##
##                  0    1
##   Minivan     0.84 0.16
##   Panel Truck 0.74 0.26
##   Pickup      0.68 0.32
##   Sports Car  0.66 0.34
##   SUV         0.70 0.30
##   Van         0.73 0.27
```

Single Parent - Unknown effect

Lets make it numeric.

```
##
##          0    1
##   No  0.76 0.24
##   Yes 0.56 0.44
```

URBANICITY - Home/Work Area

Lets make it numeric.

```
##
##                             0    1
##   Highly Urban/ Urban    0.69 0.31
##   z_Highly Rural/ Rural 0.93 0.07
```

After adding the dummy variables, here's a glimpse of our final dataset after dummy variables.

```
## Observations: 8,161
## Variables: 32
## $ TARGET_FLAG       <fctr> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1...
## $ TARGET_AMT        <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000,...
## $ AGE               <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, ...
## $ HOMEKIDS          <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0,...
## $ YOJ               <int> 11, 11, 10, 14, 11, 12, 11, 11, 10, 7, 14, 5...
## $ INCOME            <dbl> 67349, 91449, 16039, 0, 114986, 125301, 1875...
## $ PARENT1           <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ HOME_VAL          <dbl> 67349, 91449, 16039, 0, 114986, 125301, 0, 1...
## $ TRAVTIME          <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 3...
## $ BLUEBOOK          <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 878...
## $ TIF               <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1...
## $ CLM_FREQ          <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0,...
## $ REVOKED           <dbl> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,...
## $ MVR_PTS           <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3...
## $ CAR_AGE           <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, ...
## $ OLDCLAIM_YRLY_AVG <dbl> 892.2, 0.0, 7738.0, 0.0, 3843.4, 0.0, 0.0, 4...
## $ INCOME_LOG        <dbl> 4.828338, 4.961184, 4.205204, 4.708565, 5.06...
## $ HOME_VAL_LOG      <dbl> 4.828338, 4.961184, 4.205204, 4.680798, 5.06...
## $ CARUSE_COMMERCIAL <dbl> 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0,...
## $ EDU_BACH_MAST_PHD <dbl> 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0,...
## $ JOB_WHITECOLLAR   <dbl> 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0,...
## $ KIDSDRIV_2        <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...
## $ KIDSDRIV_4        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ MARRIED           <dbl> 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1,...
## $ RED_SPORTS_CAR    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ MALE              <dbl> 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0,...
## $ CAR_TYPE_vAN      <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0,...
## $ CAR_TYPE_SUV      <dbl> 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1,...
## $ CAR_TYPE_SPORTS   <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ CAR_TYPE_PICKUP   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CAR_TYPE_PANTRUCK <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,...
## $ URBAN             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,...
```

Now , all of our data has got all numeric values and the dataset is ready for modelling process.

Check for Multicolinearity in the predictors:

Check for Multicolinearity among the predictor variables and remove those with excessive correlation among the explanatory variables.

	Multicolinearity score
INCOME_LOG	22.545583
HOME_VAL_LOG	22.008005
INCOME	10.446977
HOME_VAL	9.318376
CAR_TYPE_SUV	3.099313
MALE	2.951051
CAR_TYPE_SPORTS	2.272462
CAR_TYPE_PANTRUCK	2.247069
BLUEBOOK	2.195422
EDU_BACH_MAST_PHD	2.099079
HOMEKIDS	2.010498
PARENT1	1.947333
CAR_TYPE_PICKUP	1.754282
CAR_AGE	1.698188
JOB_WHITECOLLAR	1.669718
CARUSE_COMMERCIAL	1.652439
OLDCLAIM_YRLY_AVG	1.648866
CAR_TYPE_vAN	1.600807
MARRIED	1.579710
KIDSDRIV_2	1.571462
CLM_FREQ	1.465373
AGE	1.423254
KIDSDRIV_4	1.411280
YOJ	1.338544
REVOKED	1.313804
MVR_PTS	1.160176
URBAN	1.129745
RED_SPORTS_CAR	1.060104
TRAVTIME	1.037033
TIF	1.008651

From the above table, we do see multi-colinearity (with VIF $>$ 10) among the predictors - INCOME_LOG,HOME_VAL_LOG variables.

Lets eliminate INCOME_LOG and see first:

	Multicolinearity score
INCOME	8.630156
HOME_VAL	7.881240
CAR_TYPE_SUV	3.100068
MALE	2.951401
CAR_TYPE_SPORTS	2.273086
CAR_TYPE_PANTRUCK	2.247096
BLUEBOOK	2.194874
HOME_VAL_LOG	2.193388
EDU_BACH_MAST_PHD	2.096837
HOMEKIDS	2.010188
PARENT1	1.948250
CAR_TYPE_PICKUP	1.753911
CAR_AGE	1.697948
JOB_WHITECOLLAR	1.667259
CARUSE_COMMERCIAL	1.651299

	Multicolinearity score
OLDCLAIM_YRLY_AVG	1.648527
CAR_TYPE_vAN	1.601194
MARRIED	1.579884
KIDSDRIV_2	1.569796
CLM_FREQ	1.465235
AGE	1.422441
KIDSDRIV_4	1.410903
REVOKED	1.313710
YOJ	1.311879
MVR_PTS	1.160096
URBAN	1.129823
RED_SPORTS_CAR	1.060066
TRAVTIME	1.036874
TIF	1.008663

Perfect, now all the VIF values are below 10, and we are good to proceed.

Split the dataset into training and test:

We will randomly split our dataset into training (75%) and test (25%).

```
set.seed(2)
s = sample(1:nrow(insurance.trn), 0.75 * nrow(insurance.trn))
insurance.training = insurance.trn[s, ]
insurance.test = insurance.trn[-s, ]
```

Number of observations in *training* dataset is 6120

Number of observations in *test* dataset is 2041

Build Models

Lets start building models to predict the target variables - *TARGET_FLAG* and *TARGET_AMT*

BINARY LOGISTIC REGRESSION MODELS TO PREDICT TARGET_FLAG

Build different logistic models to predict the probability of TARGET_FLAG

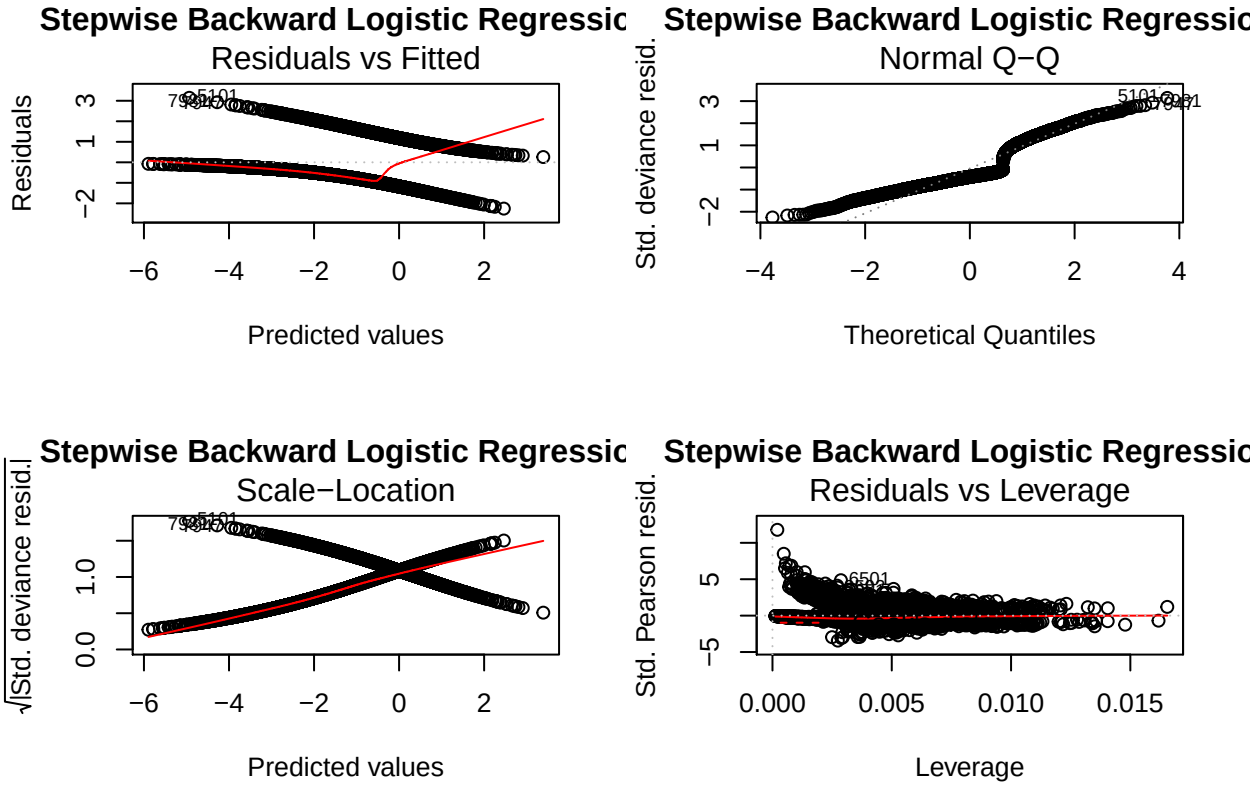
Stepwise Backward Logistic Regression:

```
##
## Call:
## stats::glm(formula = TARGET_FLAG ~ HOMEKIDS + YOJ + INCOME +
##     PARENT1 + TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + REVOKED +
##     MVR_PTS + OLDCLAIM_YRLY_AVG + CARUSE_COMMERCIAL + EDU_BACH_MAST_PHD +
##     JOB_WHITECOLLAR + KIDSDRIV_2 + MARRIED + CAR_TYPE_vAN + CAR_TYPE_SUV +
##     CAR_TYPE_SPORTS + CAR_TYPE_PICKUP + CAR_TYPE_PANTRUCK + URBAN,
##     family = binomial(), data = insurance.training)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2559  -0.7273  -0.4224   0.6340   3.1436
##
## Coefficients:
##                        Estimate    Std. Error z value             Pr(>|z|)
## (Intercept)       -3.2100392684  0.2024774663 -15.854 < 0.0000000000000002
## HOMEKIDS           0.0892842630  0.0380749417   2.345             0.019029
## YOJ               -0.0158848101  0.0087281464  -1.820             0.068766
## INCOME            -0.0000048813  0.0000009331  -5.231 0.000000168481326405
## PARENT1            0.3291867254  0.1261239121   2.610             0.009054
## TRAVTIME           0.0158874112  0.0021412263   7.420 0.000000000000117322
## BLUEBOOK          -0.0000189876  0.0000053380  -3.557             0.000375
## TIF               -0.0573464048  0.0084039051  -6.824 0.000000000008867528
## CLM_FREQ           0.1722253460  0.0326775137   5.270 0.000000136086428352
## REVOKED            0.8955316674  0.1021436579   8.767 < 0.0000000000000002
## MVR_PTS            0.1345549220  0.0158419144   8.494 < 0.0000000000000002
## OLDCLAIM_YRLY_AVG -0.0000570833  0.0000221720  -2.575             0.010037
## CARUSE_COMMERCIAL  0.7029473482  0.0863119183   8.144 0.000000000000000382
## EDU_BACH_MAST_PHD -0.3782491273  0.0828921629  -4.563 0.000005039249700364
## JOB_WHITECOLLAR   -0.3171436585  0.0908748850  -3.490             0.000483
## KIDSDRIV_2         0.5737507577  0.1101710772   5.208 0.000000191075706769
## MARRIED           -0.6731983582  0.0828973941  -8.121 0.000000000000000463
## CAR_TYPE_vAN       0.5525456762  0.1356840972   4.072 0.000046552158493901
## CAR_TYPE_SUV       0.6830475772  0.0974388760   7.010 0.000000000002382997
## CAR_TYPE_SPORTS    0.8956765892  0.1218910877   7.348 0.000000000000200937
## CAR_TYPE_PICKUP    0.4578519177  0.1118010922   4.095 0.000042173758109571
## CAR_TYPE_PANTRUCK  0.4821113866  0.1596527251   3.020             0.002530
## URBAN              2.2240141351  0.1256257322  17.703 < 0.0000000000000002
##
## (Intercept)       ***
## HOMEKIDS          *
## YOJ               .
## INCOME            ***
## PARENT1           **
## TRAVTIME          ***
## BLUEBOOK          ***
## TIF               ***
## CLM_FREQ          ***
## REVOKED           ***
## MVR_PTS           ***
## OLDCLAIM_YRLY_AVG *
## CARUSE_COMMERCIAL ***
## EDU_BACH_MAST_PHD ***
## JOB_WHITECOLLAR   ***
## KIDSDRIV_2        ***
## MARRIED           ***
## CAR_TYPE_vAN      ***
## CAR_TYPE_SUV      ***
## CAR_TYPE_SPORTS   ***
## CAR_TYPE_PICKUP   ***
## CAR_TYPE_PANTRUCK **
## URBAN             ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7053.2  on 6119  degrees of freedom
## Residual deviance: 5561.6  on 6097  degrees of freedom
## AIC: 5607.6
##
## Number of Fisher Scoring iterations: 5

## TARGET_FLAG ~ HOMEKIDS + YOJ + INCOME + PARENT1 + TRAVTIME +
##     BLUEBOOK + TIF + CLM_FREQ + REVOKED + MVR_PTS + OLDCLAIM_YRLY_AVG +
##     CARUSE_COMMERCIAL + EDU_BACH_MAST_PHD + JOB_WHITECOLLAR +
##     KIDSDRIV_2 + MARRIED + CAR_TYPE_vAN + CAR_TYPE_SUV + CAR_TYPE_SPORTS +
##     CAR_TYPE_PICKUP + CAR_TYPE_PANTRUCK + URBAN
```



From the above residual plots, shows that the residuals are not fully linear and the residual variance is not constant. The Normal Q-Q graph indicates that the most of the residuals are on the straight line.However, the Residual Vs Leverage plot has the redline alligned with gray dotted line, which indicates that the assumption of standardized residuals centered around zero is some what true here.

The significant variables include:

INCOME, TRAVTIME, BLUEBOOK, TIF,CLM_FREQ,REVOKED,MVR_PTS,CARUSE_COMMERCIAL EDU_BACH_MAST_PHD,JOB_WHITECOLLAR,KIDSDRIV_2 and ,MARRIED,CAR_TYPE,URBAN

For a unit increase in REVOKED there is a chance of 89% of crash.

And also commercial vehicles tend to be resulting in increased chance of a crash (approx. 70% per unit increase in commericial car use).

Sports vehicles have 89% chance of crash, and its interested to see that the marital status decreases the probability of the crash by 67%.

Higher education results in 37% decrease in crash. And a white collar job decreases the crash chance by 31%.

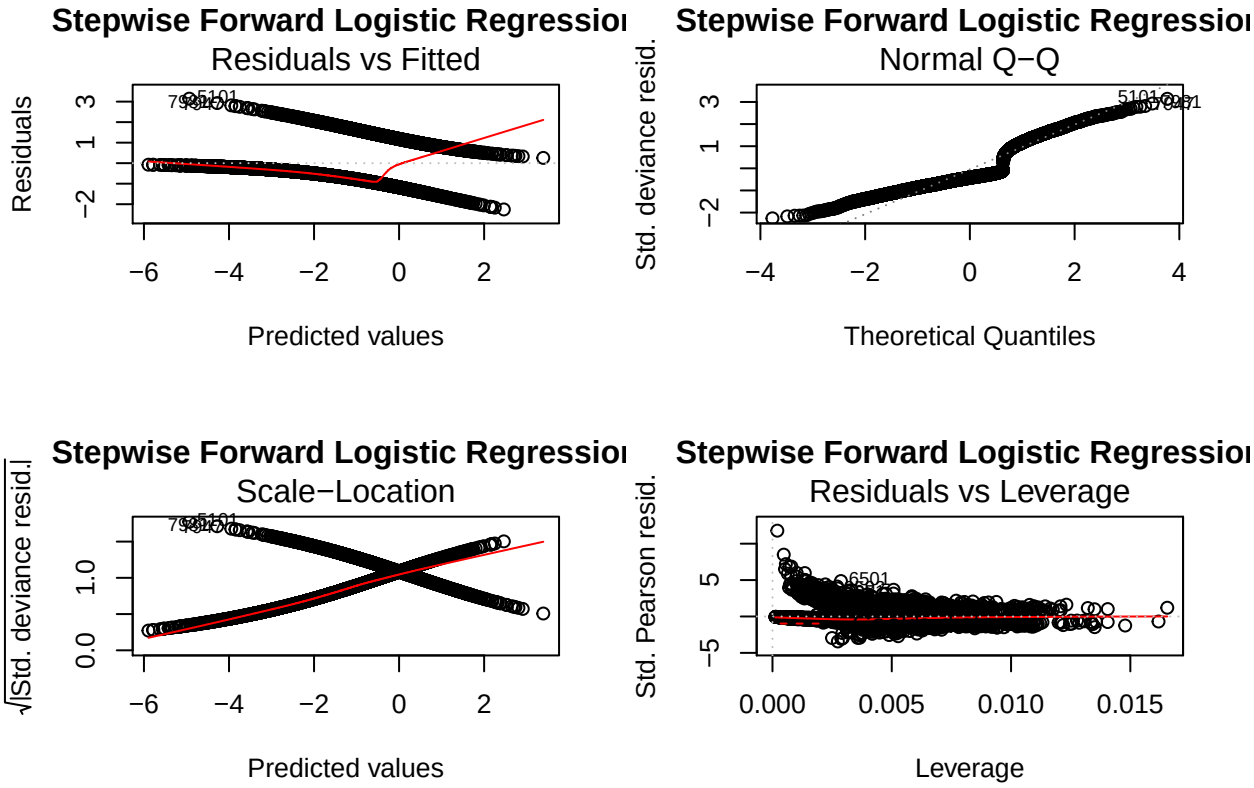
Also, the URBAN areas results in increased chance for car crash.

Stepwise Forward Logistic Regression:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ URBAN + JOB_WHITECOLLAR + MVR_PTS +
##     PARENT1 + REVOKED + INCOME + CARUSE_COMMERCIAL + TRAVTIME +
##     TIF + MARRIED + KIDSDRIV_2 + BLUEBOOK + CLM_FREQ + CAR_TYPE_SPORTS +
##     CAR_TYPE_SUV + EDU_BACH_MAST_PHD + OLDCLAIM_YRLY_AVG + CAR_TYPE_PICKUP +
##     CAR_TYPE_vAN + CAR_TYPE_PANTRUCK + HOMEKIDS + YOJ, family = binomial,
##     data = na.omit(insurance.training))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2559  -0.7273  -0.4224   0.6340   3.1436
##
## Coefficients:
##                         Estimate    Std. Error z value             Pr(>|z|)
## (Intercept)       -3.2100392684  0.2024774663 -15.854 < 0.0000000000000002
## URBAN              2.2240141351  0.1256257322  17.703 < 0.0000000000000002
## JOB_WHITECOLLAR   -0.3171436585  0.0908748850  -3.490             0.000483
## MVR_PTS            0.1345549220  0.0158419144   8.494 < 0.0000000000000002
## PARENT1            0.3291867254  0.1261239121   2.610             0.009054
## REVOKED            0.8955316674  0.1021436579   8.767 < 0.0000000000000002
## INCOME            -0.0000048813  0.0000009331  -5.231 0.000000168481326405
## CARUSE_COMMERCIAL  0.7029473482  0.0863119183   8.144 0.000000000000000382
## TRAVTIME           0.0158874112  0.0021412263   7.420 0.000000000000117322
## TIF               -0.0573464048  0.0084039051  -6.824 0.000000000008867528
## MARRIED           -0.6731983582  0.0828973941  -8.121 0.000000000000000463
## KIDSDRIV_2         0.5737507577  0.1101710772   5.208 0.000000191075706769
## BLUEBOOK          -0.0000189876  0.0000053380  -3.557             0.000375
## CLM_FREQ           0.1722253460  0.0326775137   5.270 0.000000136086428352
## CAR_TYPE_SPORTS    0.8956765892  0.1218910877   7.348 0.000000000000200937
## CAR_TYPE_SUV       0.6830475772  0.0974388760   7.010 0.000000000002382997
## EDU_BACH_MAST_PHD -0.3782491273  0.0828921629  -4.563 0.000005039249700364
## OLDCLAIM_YRLY_AVG -0.0000570833  0.0000221720  -2.575             0.010037
## CAR_TYPE_PICKUP    0.4578519177  0.1118010922   4.095 0.000042173758109571
## CAR_TYPE_vAN       0.5525456762  0.1356840972   4.072 0.000046552158493904
## CAR_TYPE_PANTRUCK  0.4821113866  0.1596527251   3.020             0.002530
## HOMEKIDS           0.0892842630  0.0380749417   2.345             0.019029
## YOJ               -0.0158848101  0.0087281464  -1.820             0.068766
##
## (Intercept)       ***
## URBAN             ***
## JOB_WHITECOLLAR   ***
## MVR_PTS           ***
```

```
## PARENT1           **
## REVOKED           ***
## INCOME            ***
## CARUSE_COMMERCIAL ***
## TRAVTIME          ***
## TIF               ***
## MARRIED           ***
## KIDSDRIV_2        ***
## BLUEBOOK          ***
## CLM_FREQ          ***
## CAR_TYPE_SPORTS   ***
## CAR_TYPE_SUV      ***
## EDU_BACH_MAST_PHD ***
## OLDCLAIM_YRLY_AVG *
## CAR_TYPE_PICKUP   ***
## CAR_TYPE_vAN      ***
## CAR_TYPE_PANTRUCK **
## HOMEKIDS          *
## YOJ               .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7053.2  on 6119  degrees of freedom
## Residual deviance: 5561.6  on 6097  degrees of freedom
## AIC: 5607.6
##
## Number of Fisher Scoring iterations: 5

## TARGET_FLAG ~ URBAN + JOB_WHITECOLLAR + MVR_PTS + PARENT1 + REVOKED +
##     INCOME + CARUSE_COMMERCIAL + TRAVTIME + TIF + MARRIED + KIDSDRIV_2 +
##     BLUEBOOK + CLM_FREQ + CAR_TYPE_SPORTS + CAR_TYPE_SUV + EDU_BACH_MAST_PHD +
##     OLDCLAIM_YRLY_AVG + CAR_TYPE_PICKUP + CAR_TYPE_vAN + CAR_TYPE_PANTRUCK +
##     HOMEKIDS + YOJ
```



The resuls are pretty much similar to what we have seen in the backward regression above.

The residual plots, shows that the residuals are not fully linear and the residual variance is not constant. The Normal Q-Q graph indicates that the most of the residuals are on the straight line.However, the Residual Vs Leverage plot has the redline alligned with gray dotted line, which indicates that the assumption of standardized residuals centered around zero is some what true here.

For a unit increase in REVOKED there is a chance of 89% of crash.

And also commercial vehicles tend to be resulting in incrased chances of a crash (approx. 70% per unit increase in commericial car use).

Sports vehicles have 89% chance of crash, and its interested to see that the marital status decreases the probability of the crash by 67%.

Higher education results in 37% decrease in crash. And a white collar job decreases the crash chance by 31%.

Also, the URBAN areas results in increased chance for car crash.

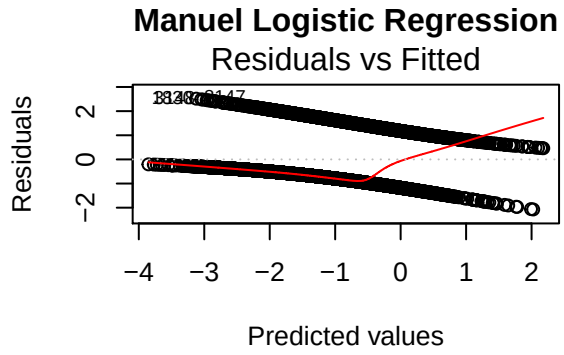
Manual Logistic Regression:

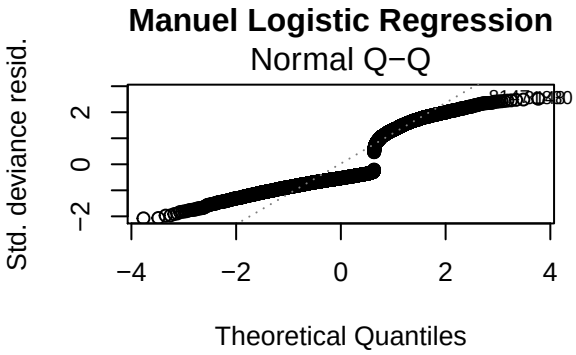
Based on the data descriptions, and the significant predictors noticed above, lets create a manual logistic model.

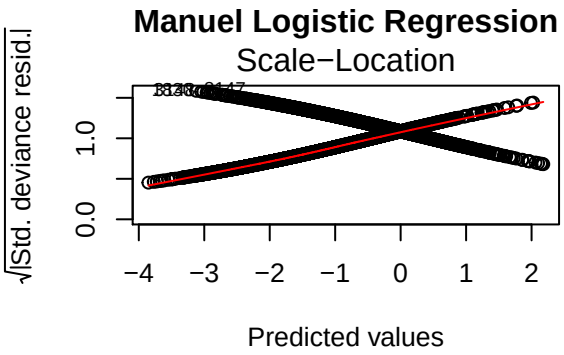
```
##
## Call:
## stats::glm(formula = TARGET_FLAG ~ +AGE + CARUSE_COMMERCIAL +
##     CLM_FREQ + HOME_VAL_LOG + JOB_WHITECOLLAR + KIDSDRIV_2 +
##     KIDSDRIV_4 + MARRIED + MVR_PTS + OLDCLAIM_YRLY_AVG + REVOKED +
```

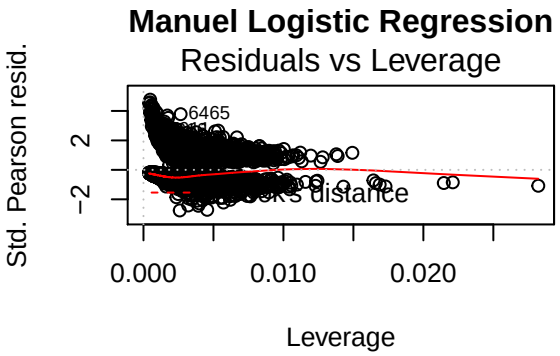
```
##     MALE + TIF + TRAVTIME + YOJ - TARGET_AMT, family = binomial(),
##     data = insurance.training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0732  -0.7574  -0.5324   0.8129   2.5162
##
## Coefficients:
##                       Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)        1.30843774  0.44754777   2.924              0.00346 **
## AGE               -0.01044244  0.00379172  -2.754              0.00589 **
## CARUSE_COMMERCIAL  0.57655739  0.07348418   7.846  0.00000000000000429 ***
## CLM_FREQ           0.30064971  0.03088331   9.735 < 0.0000000000000002 ***
## HOME_VAL_LOG      -0.43851393  0.09119407  -4.809  0.00000152006498977 ***
## JOB_WHITECOLLAR   -0.36562507  0.08034765  -4.551  0.00000535088759291 ***
## KIDSDRIV_2         0.55664853  0.11054573   5.035  0.00000047670470241 ***
## KIDSDRIV_4         0.26340148  0.18364838   1.434              0.15150
## MARRIED           -0.61204275  0.06429805  -9.519 < 0.0000000000000002 ***
## MVR_PTS            0.15631366  0.01510519  10.348 < 0.0000000000000002 ***
## OLDCLAIM_YRLY_AVG -0.00005866  0.00002124  -2.762              0.00574 **
## REVOKED            1.02885215  0.09696378  10.611 < 0.0000000000000002 ***
## MALE              -0.21333734  0.06753973  -3.159              0.00158 **
## TIF               -0.04981398  0.00800792  -6.221  0.00000000049529219 ***
## TRAVTIME           0.00747915  0.00196377   3.809              0.00014 ***
## YOJ               -0.02218021  0.00799513  -2.774              0.00553 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7053.2  on 6119  degrees of freedom
## Residual deviance: 6120.7  on 6104  degrees of freedom
## AIC: 6152.7
##
## Number of Fisher Scoring iterations: 4

## TARGET_FLAG ~ +AGE + CARUSE_COMMERCIAL + CLM_FREQ + HOME_VAL_LOG +
##     JOB_WHITECOLLAR + KIDSDRIV_2 + KIDSDRIV_4 + MARRIED + MVR_PTS +
##     OLDCLAIM_YRLY_AVG + REVOKED + MALE + TIF + TRAVTIME + YOJ -
##     TARGET_AMT
```









The resuls are some what similar to the above 2, however, the normality in residuals is better in stepwise models than our manual model.

In our manual model, we noticed that REVOKED increases the crash chance by almost 100%. And HOME_VAL decreases the decreases the chances of crash significantly. And the families with 1 or 2 teenage kids shows higher chances of crash. And the commercial car usage results in 57% crash increase.

MULTIPLE LINEAR REGRESSION MODELS TO PREDICT TARGET_AMT

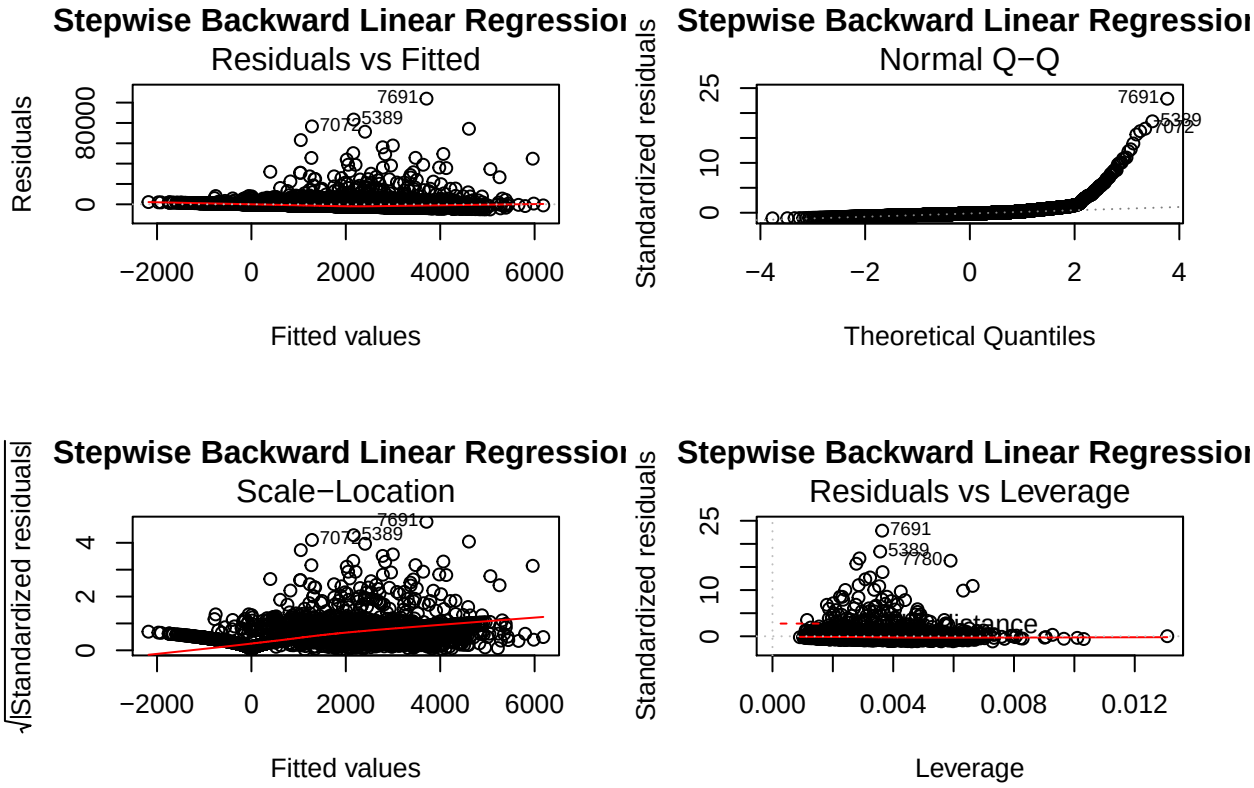
Build different linear models to predict the amount for TARGET_AMT

Stepwise Backward Linear Regression:

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + PARENT1 + TRAVTIME + BLUEBOOK +
##     TIF + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + CARUSE_COMMERCIAL +
##     KIDSDRIV_2 + MARRIED + MALE + CAR_TYPE_vAN + CAR_TYPE_SUV +
##     CAR_TYPE_SPORTS + CAR_TYPE_PICKUP + URBAN, data = insurance.training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5057  -1679   -772    323 103877
##
## Coefficients:
```

```
##                       Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)       -957.561207  332.370253  -2.881             0.003978 **
## INCOME              -0.006816    0.001415  -4.819       0.000001479213 ***
## PARENT1            722.896690  206.688768   3.498             0.000473 ***
## TRAVTIME            12.822222    3.708495   3.458             0.000549 ***
## BLUEBOOK             0.029377    0.008645   3.398             0.000683 ***
## TIF                -47.976764   13.987590  -3.430             0.000608 ***
## CLM_FREQ           130.543645   56.235036   2.321             0.020298 *
## REVOKED            628.154331  177.814607   3.533             0.000414 ***
## MVR_PTS            185.865415   30.114948   6.172       0.000000000718 ***
## CAR_AGE            -28.347441   11.541463  -2.456             0.014072 *
## CARUSE_COMMERCIAL  845.367053  134.270163   6.296       0.000000000327 ***
## KIDSDRIV_2         757.511021  194.997537   3.885             0.000104 ***
## MARRIED           -583.463995  137.093862  -4.256       0.000021128635 ***
## MALE               533.010203  172.258255   3.094             0.001982 **
## CAR_TYPE_vAN       315.633249  219.223098   1.440             0.149981
## CAR_TYPE_SUV       893.272717  205.194758   4.353       0.000013629967 ***
## CAR_TYPE_SPORTS   1199.943619  248.531650   4.828       0.000001411796 ***
## CAR_TYPE_PICKUP    314.963096  184.156064   1.710             0.087260 .
## URBAN             1393.536244  157.274464   8.861 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4553 on 6101 degrees of freedom
## Multiple R-squared:  0.06791,    Adjusted R-squared:  0.06516
## F-statistic: 24.69 on 18 and 6101 DF,  p-value: < 0.00000000000000022
```

```
## TARGET_AMT ~ INCOME + PARENT1 + TRAVTIME + BLUEBOOK + TIF + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + CARUSE_COMMERCIAL + KIDSDRIV_2 +
##     MARRIED + MALE + CAR_TYPE_vAN + CAR_TYPE_SUV + CAR_TYPE_SPORTS +
##     CAR_TYPE_PICKUP + URBAN
```



In the residuals Vs Fitted graph, the red line is about flat, which indicates the linearity in residuals is good. However the residual variance is NOT constant. The Normal Q-Q graph indicates some of the residuals are on the straight line. However, the Residual Vs Leverage plot has the redline slightly alligned with gray dotted line, this indicates that the assumption of standardized residuals centered around zero is true here, though there are few outliers.

The significant predictors in the above are:

INCOME,BLUEBOOK,TRAVTIME,TIF,REVOKED,MVR_PTS,CARUSE_COMMERCIAL KIDSDRIV_2,MARRIED,CAR_TYPE_SUV,CAR_TYPE_SPORTS and URBAN

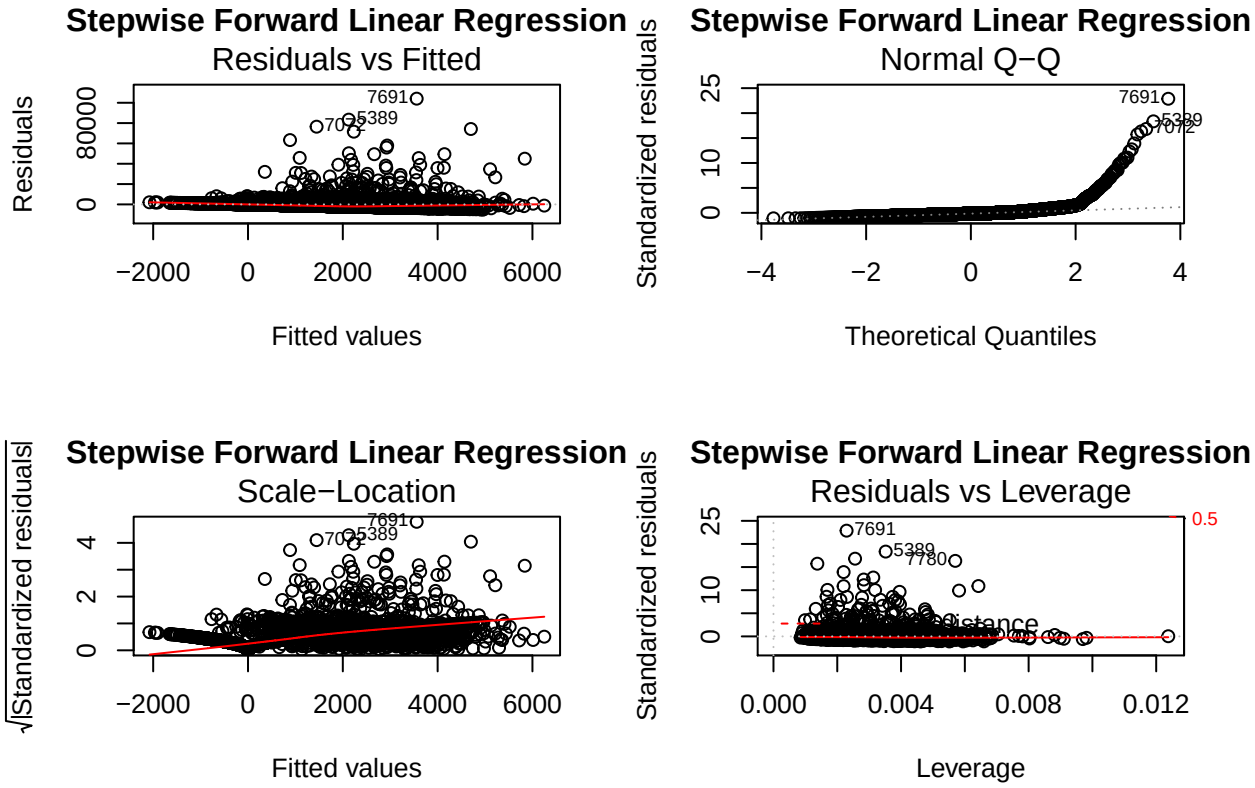
Per unit increase in the URBAN , the target amount increases by 1340. And SPORTS CARs causes in increase of 1199 target amount. Similary we can see that the *MARRIED, WHITE COLLAR, TIF, HOME_VAL_LOG* are actually causes decrease in the CLAIM AMOUNT.

Stepwise Forward Linear Regression:

```
##
## Call:
## lm(formula = TARGET_AMT ~ MVR_PTS + PARENT1 + URBAN + CARUSE_COMMERCIAL +
##     INCOME + REVOKED + MARRIED + KIDSDRIV_2 + TRAVTIME + TIF +
##     CAR_TYPE_SPORTS + CAR_AGE + CLM_FREQ + CAR_TYPE_SUV + BLUEBOOK +
##     MALE, data = insurance.training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4965  -1685   -771    304 104028
```

```
##
## Coefficients:
##                       Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)       -792.179721  315.434772  -2.511             0.012051 *
## MVR_PTS            187.316533   30.107289   6.222     0.00000000052465 ***
## PARENT1            722.541584  206.723477   3.495             0.000477 ***
## URBAN             1395.838477  157.293992   8.874 < 0.0000000000000002 ***
## CARUSE_COMMERCIAL  900.376457  131.201115   6.863     0.00000000000743 ***
## INCOME              -0.006771    0.001414  -4.787     0.00000173441157 ***
## REVOKED            639.524835  177.754523   3.598             0.000323 ***
## MARRIED           -580.534399  137.109227  -4.234     0.00002328369513 ***
## KIDSDRIV_2         754.851348  195.024985   3.871             0.000110 ***
## TRAVTIME            12.745093    3.708922   3.436             0.000594 ***
## TIF                -47.125090   13.983327  -3.370             0.000756 ***
## CAR_TYPE_SPORTS   1056.026153  235.071766   4.492     0.00000717380664 ***
## CAR_AGE            -28.636372   11.542509  -2.481             0.013130 *
## CLM_FREQ           130.779469   56.241184   2.325             0.020087 *
## CAR_TYPE_SUV       748.945103  188.802678   3.967     0.00007367391065 ***
## BLUEBOOK             0.026125    0.008206   3.184             0.001461 **
## MALE               519.965247  168.984162   3.077             0.002100 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4553 on 6103 degrees of freedom
## Multiple R-squared:  0.06729,    Adjusted R-squared:  0.06484
## F-statistic: 27.52 on 16 and 6103 DF,  p-value: < 0.00000000000000022

## TARGET_AMT ~ MVR_PTS + PARENT1 + URBAN + CARUSE_COMMERCIAL +
##     INCOME + REVOKED + MARRIED + KIDSDRIV_2 + TRAVTIME + TIF +
##     CAR_TYPE_SPORTS + CAR_AGE + CLM_FREQ + CAR_TYPE_SUV + BLUEBOOK +
##     MALE
```



The resuls are VERY similar to the Stepwise backward linear regression model.

In the residuals Vs Fitted graph, the red line is about flat, which indicates the linearity in residuals is good. However the residual variance is NOT constant. The Normal Q-Q graph indicates some of the residuals are on the straight line. However, the Residual Vs Leverage plot has the redline slightly alligned with gray dotted line, this indicates that the assumption of standardized residuals centered around zero is true here, though there are few outliers.

The significant predictors in the above are:

MVR_PTS,PARENT1,TRAVTIME,TIF,REVOKED,MVR_PTS,INCOME,CARUSE_COMMERCIAL ,MARRIED,CAR_TYPE_SUV,CAR_TYPE_SPORTS,KIDSDRIV_2,TRAVTIME,TIF and URBAN

Per unit increase in the URBAN , the target amount increases by 1395 And SPORTS CARs causes in increase of 1056 target amount. Similary we can see that the *MARRIED, TIF* are actually causes decrease in the CLAIM AMOUNT.

Also, we notice that the MALE gender increases the claim amount by 519.

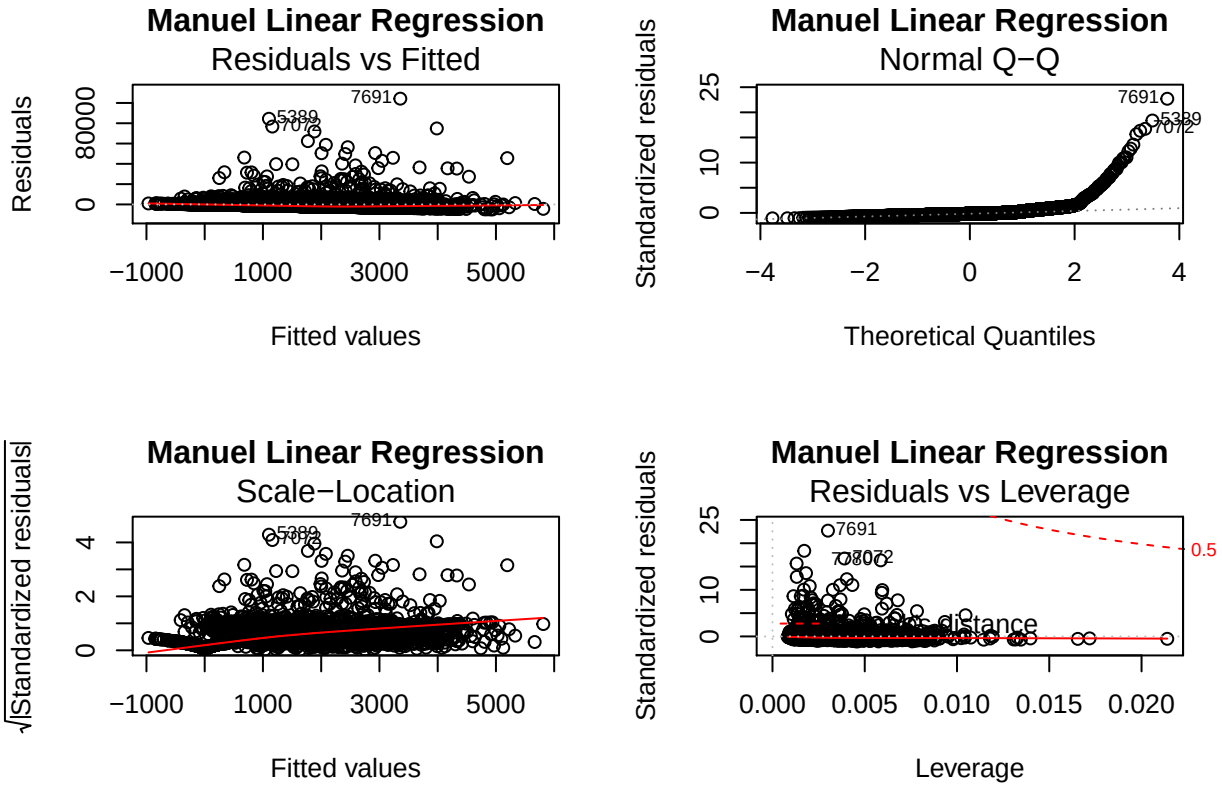
Manual Linear Regression:

Based on the data descriptions, and the significant predictors noticed above, lets create a manual linear model.

```
##
## Call:
## stats::lm(formula = TARGET_AMT ~ +AGE + CARUSE_COMMERCIAL + CLM_FREQ +
##     HOME_VAL_LOG + JOB_WHITECOLLAR + KIDSDRIV_2 + KIDSDRIV_4 +
```

```
##     MARRIED + MVR_PTS + OLDCLAIM_YRLY_AVG + REVOKED + MALE +
##     TIF + TRAVTIME + YOJ - TARGET_FLAG, data = insurance.training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4940  -1641   -844     74 104228
##
## Coefficients:
##                      Estimate Std. Error t value         Pr(>|t|)
## (Intercept)       3072.87992  887.74497   3.461         0.000541 ***
## AGE                 -4.88062    7.14154  -0.683         0.494372
## CARUSE_COMMERCIAL  801.66060  140.62291   5.701 0.00000001248389 ***
## CLM_FREQ           272.62071   62.74986   4.345 0.00001418149823 ***
## HOME_VAL_LOG      -382.67851  180.92128  -2.115         0.034456 *
## JOB_WHITECOLLAR   -120.98786  147.33283  -0.821         0.411572
## KIDSDRIV_2         871.12189  223.59639   3.896 0.00009885766863 ***
## KIDSDRIV_4          12.54029  382.44403   0.033         0.973843
## MARRIED           -741.93252  122.14159  -6.074 0.00000000131981 ***
## MVR_PTS            213.64342   30.46301   7.013 0.00000000000258 ***
## OLDCLAIM_YRLY_AVG   -0.03103    0.04323  -0.718         0.472878
## REVOKED            872.93275  198.90803   4.389 0.00001159913645 ***
## MALE                 8.24076  124.36021   0.066         0.947169
## TIF                -45.19158   14.11999  -3.201         0.001379 **
## TRAVTIME             7.31684    3.69968   1.978         0.048008 *
## YOJ                -11.85130   15.38504  -0.770         0.441144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4599 on 6104 degrees of freedom
## Multiple R-squared:  0.04828,    Adjusted R-squared:  0.04594
## F-statistic: 20.64 on 15 and 6104 DF,  p-value: < 0.00000000000000022

## TARGET_AMT ~ +AGE + CARUSE_COMMERCIAL + CLM_FREQ + HOME_VAL_LOG +
##     JOB_WHITECOLLAR + KIDSDRIV_2 + KIDSDRIV_4 + MARRIED + MVR_PTS +
##     OLDCLAIM_YRLY_AVG + REVOKED + MALE + TIF + TRAVTIME + YOJ -
##     TARGET_FLAG
```



The results are similar to the Stepwise linear regression models.

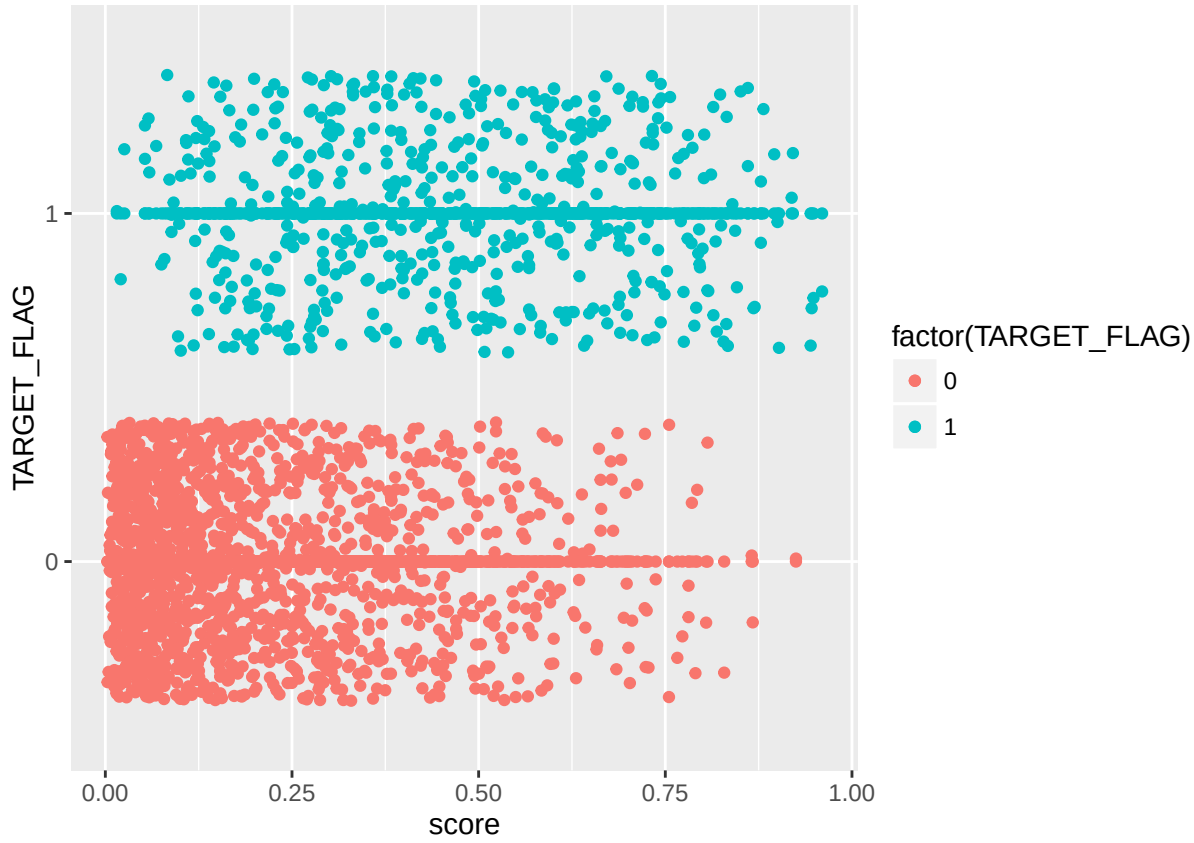
Model Selection

Logistic Regression:

Measure performance among different models and select one:

Finding Score & performance based on test data:

Stepwise Backward:

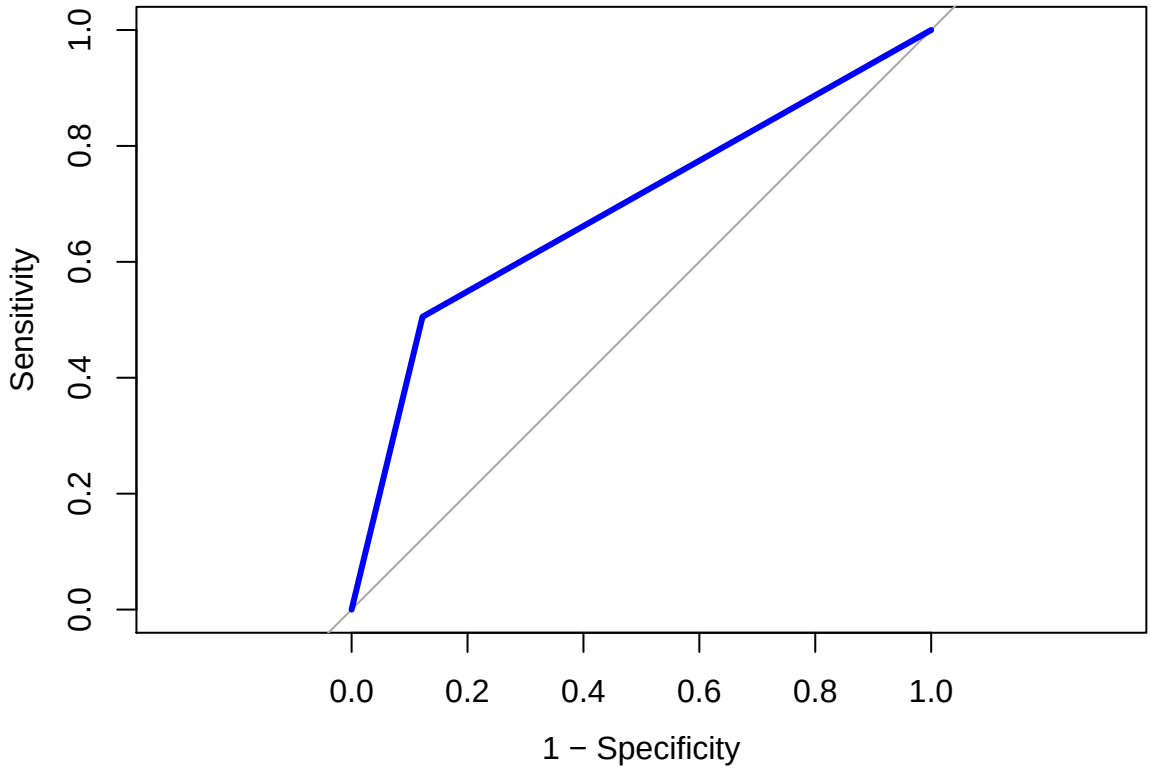


Based on our visualization, it appears like 0.30 could be a better cutoff (?).

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1119  156
##          1  379  387
##
##                Accuracy : 0.7379
##                  95% CI : (0.7182, 0.7568)
##     No Information Rate : 0.734
##     P-Value [Acc > NIR] : 0.3548
##
##                   Kappa : 0.4065
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.7127
##             Specificity : 0.7470
##          Pos Pred Value : 0.5052
##          Neg Pred Value : 0.8776
##              Prevalence : 0.2660
##          Detection Rate : 0.1896
##    Detection Prevalence : 0.3753
##       Balanced Accuracy : 0.7299
##
```

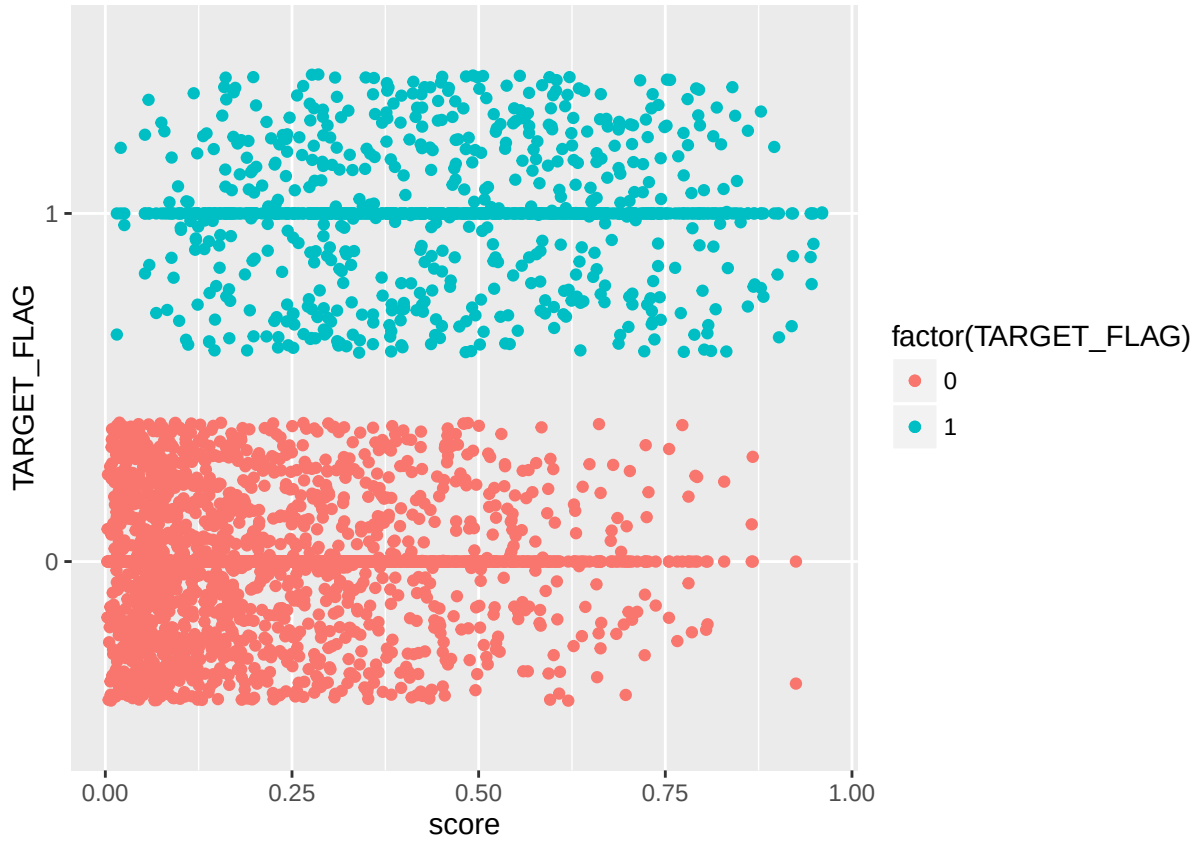
```
##        'Positive' Class : 1
##
```

AUC - Backward




```
##
## Call:
## roc.formula(formula = factor(predicted) ~ as.numeric(TARGET_FLAG),     data = insurance.test, plot =
##
## Data: as.numeric(TARGET_FLAG) in 1275 controls (factor(predicted) 0) < 766 cases (factor(predicted)
## Area under the curve: 0.6914
## 95% CI: 0.6716-0.7113 (DeLong)
```

Stepwise Forward:

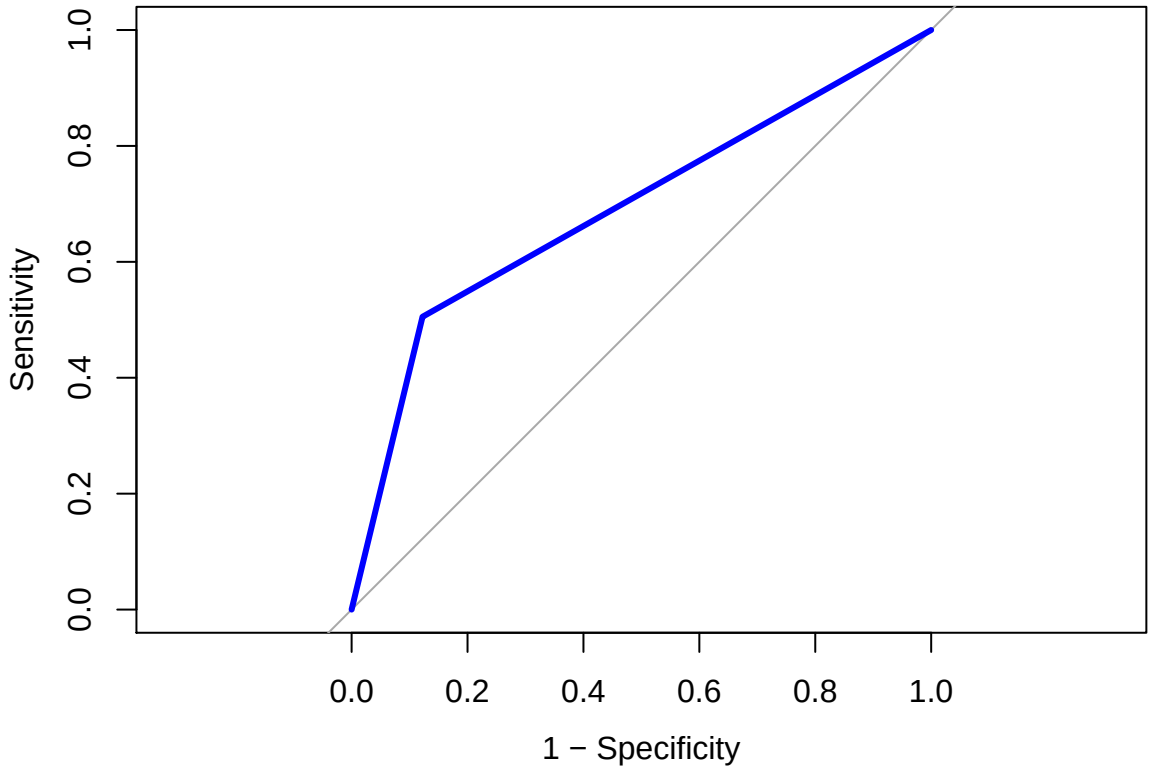


Based on our visualization, it appears like 0.30 could be a better cutoff (?).

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1119  156
##          1  379  387
##
##                Accuracy : 0.7379
##                  95% CI : (0.7182, 0.7568)
##     No Information Rate : 0.734
##     P-Value [Acc > NIR] : 0.3548
##
##                   Kappa : 0.4065
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.7127
##             Specificity : 0.7470
##          Pos Pred Value : 0.5052
##          Neg Pred Value : 0.8776
##              Prevalence : 0.2660
##          Detection Rate : 0.1896
##    Detection Prevalence : 0.3753
##       Balanced Accuracy : 0.7299
##
```

```
##        'Positive' Class : 1
##
```

AUC - Forward




```
##
## Call:
## roc.formula(formula = factor(predicted) ~ as.numeric(TARGET_FLAG),     data = insurance.test, plot =
##
## Data: as.numeric(TARGET_FLAG) in 1275 controls (factor(predicted) 0) < 766 cases (factor(predicted)
## Area under the curve: 0.6914
## 95% CI: 0.6716-0.7113 (DeLong)
```

Based on the regression plots and model summary, we are not considering the manual model in our selection process.

Compare Results:

Method	Sn	Sp	Accuracy	AUC
Step wise Backward	0.7127	0.747	0.7379	0.6914
Step wise Forward	0.7127	0.747	0.7379	0.6914

From the above, both the models result the same in predicting the target variable of the given dataset.

Mulitple Linear Regression:

Measure performance among Mulitple Linear Regression models and select one

Stepwise Backward Linear Regression Performance:

Performance on the test data:

R squared -proportion of variation in the dependent (response) variable R-Squared Higher the better (> 0.70) Adj R-Squared Higher the better MSE (Mean squared error) Lower the better

We notice that the adjusted R square is 6.3% which is very less. So, this suggests that the model may not be able to predict the target amount unlike the target flag.

Compare Results:

Method	Adj R Squared	RMSE
Step wise Backward	0.0652	4550.8815
Step wise Forward	0.0648	4553.7836

From the above, both the models result in poor performance in predicting the target variable of amount from the given dataset. So, predicting the target crash amount would be difficult with our above assumptions.

However, the *Step wise Backward* method performed better in multiple linear regression case. So, we will finalize backward model for *target amount* prediction.

Evaluation:

Lets apply the final models on the evaluation dataset. We will first tidy the evaluation dataset and then predit the *TARGET_FLAG* and *TARGET_AMT*

'z_' will be removed from **MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE**

Factorize the variables from the evaluation data set:

Data imputing for evaluation data set:

Final Results:

Here's our final results from the above models:

Though it was easier to predict the *target flag* - which indicates 'if someone gets into a crash or not', however it was not easier to predict the *TARGET AMOUNT* which indicates 'how much amount will be claimed'. So, the target claim amount from a crash seems to be more random.

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
3	0	1943	1	0	1
9	1	3114	1	0	0
10	0	1002	0	0	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
18	0	1916	1	0	0
21	0	2331	1	0	0
30	0	2189	1	1	1
31	1	2100	0	1	0
37	1	2566	1	1	0
39	0	895	0	0	1
47	0	2117	0	0	1
60	0	485	1	1	0
62	1	2759	0	0	0
63	1	4613	0	0	0
64	0	402	1	1	1
68	0	11	1	1	0
75	1	3280	0	0	0
76	1	3088	0	1	0
83	0	1423	1	1	1
87	1	2646	0	1	0
92	1	1878	0	0	1
98	0	1732	1	1	0
106	1	1352	0	1	0
107	0	999	1	1	1
113	1	1716	0	0	0
120	1	1636	0	1	1
123	1	2577	1	1	0
125	1	1908	1	0	1
126	1	3075	0	0	1
128	0	702	0	1	0
129	0	980	1	1	1
131	0	756	0	1	1
135	1	2686	1	0	1
141	0	575	1	1	1
147	0	1479	1	1	1
148	0	1170	0	0	1
151	0	705	0	1	0
156	0	1137	0	1	0
157	0	1780	0	1	1
174	0	130	1	1	1
186	1	3187	0	1	0
193	1	1458	1	1	1
195	1	2614	0	1	0
212	0	-74	1	0	1
213	1	3274	0	0	0
217	0	-1465	0	1	1
223	0	925	1	1	1
226	0	571	0	1	1
228	1	2409	0	1	1
230	0	-540	0	0	1
241	1	2395	0	0	0
243	0	286	1	1	1
249	0	2802	0	0	1
281	1	4286	1	1	0
288	0	770	1	0	1
294	1	2475	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
295	0	2207	1	0	1
300	1	1600	1	0	1
302	1	2339	0	0	1
303	0	588	1	1	1
308	1	1983	0	1	0
319	0	-934	0	1	1
320	0	527	0	1	1
324	1	1985	1	0	1
331	0	1488	0	0	1
343	0	157	1	1	1
347	1	1260	0	1	1
348	1	3786	0	1	0
350	1	3507	0	0	0
357	0	1356	0	0	1
358	0	1514	0	0	1
360	0	-242	0	1	1
366	0	1692	1	0	1
367	1	3213	0	0	0
368	0	1503	1	1	1
376	1	3048	1	1	0
380	1	1922	0	0	1
388	1	1488	0	1	0
396	0	1669	1	0	1
398	0	781	1	1	0
403	0	333	0	1	0
410	1	2055	0	1	0
412	1	2435	0	0	1
420	1	2009	0	1	1
434	0	704	1	1	0
440	1	2006	0	1	0
450	1	3171	1	1	0
453	1	1485	1	1	0
464	0	1459	0	0	1
465	0	-221	0	1	1
466	1	3766	0	1	0
473	0	933	1	1	1
476	0	1020	0	1	0
478	0	-433	1	1	1
479	0	1696	0	1	1
493	0	480	1	1	0
497	0	2805	1	1	1
503	0	-747	1	1	1
504	1	2374	0	0	0
505	1	2325	1	0	1
507	0	1306	1	1	0
513	1	1079	0	1	0
519	1	3117	1	0	0
521	1	3078	1	1	0
522	1	3330	1	0	0
545	0	1832	0	1	1
549	0	694	0	0	1
551	0	1388	1	0	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
556	0	635	1	0	0
557	1	2641	0	0	1
559	0	1431	0	1	0
560	1	2950	0	0	0
566	0	-473	1	0	1
569	0	1078	0	1	0
573	1	1743	0	1	0
578	1	2998	0	0	0
579	0	2	0	1	0
582	0	-644	1	1	1
596	1	3107	0	1	0
598	1	1877	0	1	0
599	0	1370	0	0	1
602	0	1654	0	0	0
605	1	3844	1	1	0
617	1	2614	0	0	0
619	1	2012	0	0	0
630	0	1563	1	1	1
634	1	2314	1	1	0
643	1	1815	1	1	0
645	0	1296	0	1	1
647	1	1844	0	1	0
649	0	294	0	0	1
656	0	305	0	0	1
657	0	1065	0	0	1
658	0	884	1	1	1
667	0	568	1	1	0
692	0	2179	1	1	1
693	1	1979	0	1	0
698	1	3883	1	0	0
699	1	4271	1	0	1
700	0	132	1	0	0
704	0	629	0	0	0
707	0	377	1	1	1
708	1	3471	0	0	0
709	0	639	1	0	1
713	0	373	1	1	1
714	0	348	1	1	1
716	1	3500	0	1	1
718	0	534	0	0	1
722	0	1008	1	0	1
729	1	3184	1	0	1
731	0	718	1	1	1
733	1	3259	0	0	1
746	1	2596	1	1	1
747	1	4138	0	0	0
748	1	1890	0	1	1
753	0	2270	0	0	1
757	1	2311	1	0	1
763	0	398	0	0	0
767	0	614	1	1	1
774	1	1758	0	0	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
776	1	1907	0	0	0
788	0	1557	0	1	1
794	0	719	1	1	1
799	0	1621	0	0	1
803	0	670	1	1	0
806	1	3858	1	0	1
807	0	1490	0	0	1
811	0	1096	1	1	1
816	0	1125	0	1	1
818	1	2121	0	1	0
819	0	1058	1	1	1
831	0	1399	1	1	0
835	1	3352	1	1	0
837	0	1583	1	1	1
841	1	4490	1	0	1
846	1	2055	1	0	1
856	1	1560	0	1	1
861	1	3561	1	1	0
862	1	2902	1	1	1
863	1	4699	0	0	0
865	1	2701	0	1	0
871	1	3021	1	0	0
879	0	2961	1	1	1
880	0	430	1	1	1
881	1	1813	0	1	1
885	0	1642	1	0	1
887	0	1778	0	0	1
892	0	413	1	1	1
898	0	-261	1	1	1
900	0	935	0	0	0
904	0	2102	0	0	1
906	1	1995	0	1	0
910	1	2892	0	1	0
912	1	3499	1	0	1
913	1	845	0	1	0
919	0	458	0	1	1
924	1	2129	0	0	0
925	1	2396	0	0	0
930	0	1582	1	1	1
940	0	1572	0	1	1
941	0	1755	0	1	0
946	0	1513	0	1	1
949	1	2053	0	1	1
951	0	961	0	1	1
962	0	887	1	1	0
966	0	463	1	1	0
967	0	-566	0	1	1
971	1	4709	1	0	0
981	0	812	0	1	1
982	0	1151	1	0	0
983	0	852	0	1	1
984	0	-417	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
989	0	1667	1	0	1
990	1	2550	0	0	0
992	1	2650	0	1	1
995	0	-161	0	0	0
996	0	2046	1	0	1
998	1	3537	0	0	1
1001	0	279	1	0	0
1007	0	1346	1	1	1
1008	0	63	1	1	1
1016	0	397	1	0	0
1022	0	226	1	1	1
1027	1	2341	1	1	0
1032	1	2022	0	0	1
1033	0	1682	1	0	1
1041	0	1177	0	0	1
1065	1	3341	1	0	0
1074	1	2322	0	1	1
1075	1	2123	1	1	0
1081	0	1626	0	1	1
1094	0	-319	1	1	1
1099	0	1412	1	0	1
1105	1	3962	1	0	1
1123	0	1025	0	1	1
1135	0	-366	0	1	1
1142	0	1065	1	1	0
1155	0	670	1	0	1
1169	0	-169	0	0	1
1176	0	106	1	1	1
1178	1	4084	0	1	1
1180	0	-249	1	0	0
1184	0	1066	0	1	1
1185	1	3751	1	0	1
1193	0	2143	1	1	1
1196	0	863	1	0	1
1199	1	1955	0	0	0
1203	0	1393	0	1	1
1205	1	1937	0	1	0
1207	0	-152	0	0	0
1208	1	2554	1	0	0
1212	1	2283	0	1	0
1213	1	2549	1	0	0
1222	0	1200	0	1	1
1223	0	1859	1	1	1
1226	1	1345	0	0	1
1227	1	2409	0	1	1
1229	0	566	1	1	0
1230	0	2017	0	0	1
1231	1	2899	0	1	1
1241	0	1472	1	0	0
1243	0	295	0	0	0
1244	0	1360	1	1	0
1246	0	688	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
1248	0	1159	0	1	0
1249	0	1092	0	1	1
1252	0	436	1	1	1
1261	0	700	0	1	1
1275	0	410	0	1	0
1281	1	4612	1	0	0
1285	1	1883	0	1	0
1288	1	2229	0	0	0
1290	0	318	0	0	1
1291	1	1748	0	0	0
1304	1	4207	1	0	1
1305	0	1045	1	0	1
1323	1	2359	1	0	1
1342	1	4257	0	0	1
1348	0	1326	0	1	0
1353	0	1830	0	1	0
1363	0	851	0	0	0
1371	0	2002	1	0	1
1372	0	1856	1	0	1
1378	0	1335	0	1	0
1381	1	1678	0	1	0
1382	0	1860	0	1	1
1393	1	2453	1	1	0
1394	0	604	1	1	1
1398	0	1710	0	1	0
1404	1	1641	1	0	1
1405	1	3379	0	1	0
1419	1	2762	0	1	1
1421	0	595	1	1	0
1426	0	473	0	1	0
1431	1	2014	1	1	0
1435	0	1184	0	0	1
1437	1	1996	0	1	0
1438	0	1443	1	1	1
1442	1	3170	0	0	0
1464	0	843	1	1	1
1471	1	1759	0	0	1
1473	0	1679	1	1	0
1476	0	397	0	0	1
1478	0	1312	1	1	0
1479	1	4874	0	0	1
1487	1	2244	0	0	0
1492	1	2301	1	1	1
1496	0	759	1	0	1
1497	1	2017	0	1	0
1515	0	-702	0	1	0
1519	0	1789	1	0	1
1522	1	4024	1	0	0
1526	1	1007	0	1	0
1537	0	-319	0	1	1
1538	1	4537	1	0	0
1540	0	1947	0	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
1543	0	626	0	1	1
1548	0	1982	1	1	1
1549	0	650	1	1	0
1556	1	2982	1	0	0
1564	0	-207	1	1	1
1570	0	1224	1	1	1
1577	1	2778	1	1	0
1585	0	1634	1	1	1
1590	0	1344	1	1	1
1592	1	3645	0	1	0
1594	1	2343	0	0	1
1596	1	3271	0	0	1
1598	0	766	0	1	1
1603	1	2343	0	0	0
1607	0	856	0	1	1
1612	0	1012	0	1	1
1627	0	228	1	1	0
1629	1	3793	0	1	0
1630	0	539	1	1	0
1640	1	3096	1	0	1
1641	0	2273	1	1	1
1646	0	2255	0	0	1
1662	1	1907	0	0	0
1668	0	1453	1	0	1
1671	0	556	0	0	1
1672	1	2596	1	1	1
1673	1	3389	0	1	0
1686	1	2086	0	1	0
1688	1	3146	0	1	0
1696	0	-177	1	1	1
1701	0	-37	1	1	0
1707	0	1117	1	1	1
1708	0	707	0	1	1
1713	0	383	0	0	0
1715	0	1035	0	1	1
1717	0	-285	0	1	0
1721	0	1864	1	0	0
1724	1	4523	0	1	0
1725	1	4570	0	0	1
1730	0	1286	0	1	0
1731	1	2515	0	1	0
1734	1	2102	0	1	1
1740	0	906	1	0	1
1748	0	-458	0	1	1
1749	0	465	0	1	1
1750	1	2823	0	0	0
1763	0	1208	0	0	1
1768	0	58	0	1	1
1773	1	2177	1	1	0
1777	0	649	1	0	1
1778	1	3229	1	0	0
1780	0	328	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
1782	1	1097	0	1	0
1784	0	764	0	1	0
1786	0	1037	0	1	0
1787	0	1155	0	1	1
1792	0	1642	1	0	1
1800	1	3525	0	0	1
1801	0	1136	1	0	0
1803	0	397	0	1	1
1804	1	2558	1	0	0
1807	0	-453	1	1	1
1818	1	2847	1	0	1
1821	0	307	1	1	0
1822	0	1696	0	1	0
1828	0	813	0	1	1
1833	0	1479	0	0	0
1844	0	2118	1	1	0
1847	0	1294	1	1	0
1850	0	689	1	1	1
1854	1	3236	0	1	1
1858	1	2642	1	1	1
1864	0	647	0	1	1
1867	0	-438	0	1	1
1876	1	2387	1	1	0
1880	0	1111	1	0	1
1881	0	607	0	1	0
1891	0	662	0	1	1
1894	0	1628	1	1	0
1895	0	789	0	1	0
1901	1	2129	0	1	1
1905	0	758	0	1	1
1912	1	1983	0	1	1
1918	1	2399	1	1	0
1921	1	2754	1	0	0
1923	1	2412	0	1	1
1924	0	2205	0	1	1
1931	0	190	0	1	1
1941	0	732	1	1	0
1950	0	350	0	1	1
1951	0	1107	0	1	1
1954	0	-1205	0	0	1
1961	1	3473	1	0	1
1966	0	-606	1	1	1
1979	0	503	0	1	1
1982	0	881	1	0	1
1987	1	2944	0	0	0
1997	1	2553	1	0	1
2004	0	-171	1	1	1
2011	1	2391	1	1	0
2015	1	1117	0	1	1
2025	0	-296	0	1	1
2033	1	1703	0	0	1
2034	0	-77	1	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
2035	0	2995	1	1	1
2036	1	2667	0	0	1
2053	1	3581	1	1	0
2059	1	4318	0	0	1
2060	0	771	0	1	0
2073	0	693	0	1	0
2084	1	1909	0	1	0
2089	0	1134	0	0	1
2092	0	1186	0	1	0
2109	1	2346	1	0	1
2129	1	2938	0	1	0
2134	1	1484	0	1	0
2135	0	-481	1	1	1
2148	0	-381	0	1	1
2149	0	78	1	1	0
2150	0	1881	1	1	1
2165	1	4364	1	0	1
2166	0	-359	1	1	1
2168	0	751	1	0	1
2170	0	1845	0	1	1
2171	0	999	0	0	1
2172	0	-247	0	1	1
2176	0	1336	1	1	1
2182	0	-180	0	1	1
2189	0	1668	0	1	0
2191	0	-205	1	0	0
2197	0	-603	0	1	0
2202	0	1585	0	0	0
2203	0	1340	0	0	1
2204	1	2877	0	0	0
2206	1	4001	0	0	0
2218	0	595	0	1	1
2219	0	1665	1	1	1
2221	1	2169	1	0	0
2226	0	1142	0	1	1
2228	1	2339	0	0	1
2232	1	2884	1	0	0
2236	1	2049	0	0	1
2241	1	4391	0	0	0
2245	0	306	1	0	1
2251	1	1227	0	1	0
2255	0	100	1	1	0
2256	0	89	0	1	1
2259	0	-972	0	0	1
2263	0	1282	1	1	1
2264	0	1039	1	1	1
2267	0	655	1	0	1
2273	1	4323	1	0	0
2277	1	2155	0	1	0
2287	0	857	1	1	1
2289	1	2925	0	0	0
2291	0	-205	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
2296	1	3228	1	0	0
2299	0	887	0	1	0
2306	0	-137	0	0	0
2314	0	2473	0	0	1
2317	0	251	0	1	1
2318	1	2071	1	1	0
2321	1	4109	0	0	0
2324	0	-96	0	1	0
2340	1	2021	0	1	0
2343	0	-387	0	1	1
2349	0	686	1	0	1
2352	0	2528	1	0	1
2353	1	1697	1	1	0
2365	1	3397	0	0	0
2370	1	2572	0	1	0
2378	0	1471	0	0	1
2390	1	2408	0	0	1
2399	0	1295	0	1	0
2402	1	3601	0	0	1
2403	1	3013	0	0	0
2404	0	257	0	1	0
2414	0	865	1	1	1
2422	0	1228	0	1	0
2424	0	2222	1	1	1
2430	1	2386	0	0	1
2435	1	2188	1	1	0
2439	0	624	1	0	1
2442	1	2376	1	0	1
2445	0	2599	1	0	1
2449	0	1385	0	1	1
2451	1	1259	0	0	0
2461	1	3814	0	0	0
2464	0	2056	1	0	1
2465	1	3651	0	0	1
2472	0	461	1	1	0
2476	1	2084	0	1	1
2482	0	1287	0	1	0
2487	0	2718	0	0	0
2498	0	2056	0	1	1
2501	0	594	0	1	1
2504	0	1565	0	1	1
2511	1	2129	0	0	0
2518	0	-687	1	1	0
2521	0	1132	0	0	1
2530	0	1160	1	0	1
2543	1	3055	1	1	0
2545	1	3103	0	0	1
2561	0	1900	1	0	0
2566	1	2775	1	0	0
2572	0	1144	1	1	0
2577	0	663	0	0	1
2578	0	2192	1	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
2580	0	1924	1	0	1
2581	0	1508	1	0	1
2582	0	667	1	1	1
2584	0	81	1	1	1
2590	0	131	1	1	1
2598	0	-1522	0	1	0
2602	0	984	0	1	0
2605	0	-709	0	1	1
2616	0	482	0	0	1
2618	0	1597	0	0	1
2619	1	2018	1	1	0
2624	0	-68	1	1	1
2632	0	1480	1	1	1
2640	0	813	0	0	1
2646	0	-569	0	1	1
2651	0	905	1	0	1
2660	0	43	0	1	1
2661	0	168	0	1	1
2668	0	606	0	0	0
2670	1	1597	0	0	1
2680	1	2362	1	0	1
2681	0	-495	1	1	0
2689	0	696	1	1	0
2694	0	704	1	0	1
2695	1	4755	1	0	0
2696	1	2966	1	0	1
2702	0	5	0	1	0
2704	0	758	1	0	0
2708	0	113	1	1	1
2709	0	-19	0	1	1
2714	1	1854	1	0	0
2716	0	306	0	0	1
2723	0	724	1	1	0
2725	1	1852	0	1	0
2738	0	735	0	0	0
2750	1	2667	1	1	0
2756	0	823	0	1	0
2758	0	-276	0	1	1
2766	0	1366	0	1	1
2767	1	1581	0	1	0
2771	0	1105	1	0	1
2775	1	1926	0	1	1
2776	0	1744	1	1	1
2779	1	5557	0	0	0
2780	1	1835	1	1	0
2781	1	2923	0	0	1
2782	1	3473	0	0	1
2783	0	112	0	1	0
2796	0	1348	1	1	0
2798	1	2106	1	0	1
2800	0	342	1	1	1
2803	0	957	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
2806	0	-1437	0	1	1
2813	0	1396	0	0	0
2818	0	1127	0	1	1
2821	1	2578	0	1	1
2825	0	1305	1	0	1
2829	0	156	0	1	0
2830	1	2624	0	1	0
2833	0	474	0	0	1
2839	1	3161	1	1	0
2843	0	862	1	1	0
2846	0	851	1	1	1
2847	0	954	0	1	1
2848	0	1882	0	1	1
2856	1	3665	1	0	0
2863	1	2631	0	0	0
2867	1	1918	0	1	1
2869	0	1155	0	0	1
2873	0	-1274	0	1	0
2874	1	1487	0	1	1
2875	1	2511	0	1	0
2880	1	2527	1	1	0
2886	1	3051	0	0	1
2887	1	1545	1	1	0
2888	0	880	0	1	0
2889	1	1958	0	1	0
2890	1	2424	0	0	0
2892	1	2457	1	0	0
2901	0	1294	0	1	1
2902	0	1226	1	0	1
2905	1	2501	1	1	1
2917	1	2649	0	1	0
2922	1	3563	1	1	1
2924	0	831	0	1	0
2930	0	1650	1	0	1
2931	0	1307	1	1	1
2946	0	1817	1	0	0
2955	1	2217	0	1	0
2962	0	-1000	0	1	0
2964	0	297	1	0	1
2965	1	2029	0	1	1
2967	0	-27	1	0	1
2970	0	-433	1	1	1
2973	1	1863	1	1	0
2974	0	1664	1	1	1
2976	1	3502	0	1	0
2977	1	2692	0	1	0
2978	0	970	0	1	1
2986	0	1056	0	1	1
2988	1	2103	0	1	1
2989	0	1817	1	1	1
2995	1	4068	1	0	0
3005	1	3030	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
3011	0	1096	0	1	1
3013	0	398	1	1	1
3019	1	3403	1	1	1
3021	0	897	1	1	0
3022	1	2065	0	0	1
3029	0	1274	0	1	0
3037	0	1208	0	0	1
3042	0	137	1	1	1
3043	0	1547	1	1	1
3049	0	1393	1	0	1
3050	1	3015	0	1	0
3053	0	1315	1	1	1
3058	0	2182	0	1	1
3062	0	1195	1	1	0
3063	0	3042	1	0	1
3065	0	-185	0	1	0
3080	0	72	0	1	1
3088	0	2674	1	1	1
3093	1	3805	0	0	1
3096	1	2256	1	1	0
3101	0	2260	1	1	0
3103	0	1138	1	1	1
3107	0	1347	0	0	1
3109	0	773	1	1	1
3111	0	197	1	1	0
3113	1	3305	1	0	0
3116	0	-1669	0	1	1
3132	0	920	1	1	1
3141	0	1293	0	1	1
3153	1	2388	1	1	1
3154	0	668	1	1	1
3160	0	1232	1	1	0
3167	0	269	1	0	0
3170	1	2559	0	1	1
3173	1	1537	0	1	0
3174	1	1694	0	0	1
3177	0	1526	1	0	1
3179	0	272	0	0	0
3184	1	1897	1	1	0
3190	1	2247	1	1	1
3193	0	375	0	1	1
3199	0	1436	0	1	0
3201	0	1392	0	1	1
3202	0	1380	1	1	0
3203	1	2564	0	1	1
3206	1	2847	1	0	0
3209	0	12	0	0	0
3210	1	2188	1	1	0
3217	1	1612	0	1	1
3220	0	2002	0	0	1
3228	1	2383	1	1	1
3232	0	-120	1	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
3239	0	1552	1	1	1
3243	1	3344	0	0	1
3245	0	1307	1	1	1
3246	1	2017	0	1	0
3251	0	-563	0	1	1
3253	0	997	1	1	0
3257	0	864	0	1	0
3260	0	-450	1	1	0
3261	0	1623	1	1	1
3263	0	1906	0	1	0
3278	0	2103	1	0	1
3281	0	1695	1	1	0
3283	0	668	1	1	1
3290	0	369	0	1	0
3297	0	1683	1	1	1
3304	0	-16	0	1	1
3305	1	1893	0	0	0
3307	0	943	0	1	1
3308	1	2595	1	1	1
3313	0	1353	0	0	0
3314	0	1883	1	1	0
3317	0	1360	0	1	1
3348	0	454	0	1	1
3350	1	2130	0	1	1
3359	0	-400	0	0	1
3367	0	822	1	1	0
3376	0	1592	1	0	1
3378	1	1948	1	1	0
3384	1	4224	0	0	0
3386	0	1740	1	0	0
3387	0	503	1	0	1
3388	0	1476	0	0	0
3390	0	171	0	1	0
3391	1	2130	0	0	1
3396	1	2426	1	0	1
3398	0	-53	0	1	1
3404	0	178	1	1	1
3406	0	451	1	1	1
3407	0	290	1	0	0
3414	0	228	1	0	1
3419	0	759	0	1	1
3423	1	2232	0	0	0
3427	0	-55	1	1	0
3432	0	708	1	0	1
3434	0	398	1	1	1
3438	0	625	1	1	1
3442	0	1386	1	1	1
3443	0	423	0	0	0
3448	0	1051	0	1	1
3456	0	574	1	0	0
3464	0	463	0	1	0
3470	1	1974	0	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
3475	1	2852	0	0	0
3477	1	1775	0	0	1
3490	0	640	0	1	1
3493	1	2395	0	1	0
3502	1	2203	1	0	0
3508	0	1469	1	1	0
3516	0	351	0	1	0
3517	0	1410	0	1	1
3525	1	1630	1	0	1
3532	1	2907	0	0	0
3535	1	1443	0	1	1
3536	1	2822	0	1	0
3540	0	2137	1	1	1
3547	0	2082	1	0	1
3550	1	2432	0	1	0
3557	1	3243	1	0	0
3562	0	1879	1	0	0
3563	0	799	0	1	1
3564	1	1643	0	0	1
3570	0	851	1	0	0
3573	0	1692	1	1	0
3577	1	2339	0	1	0
3579	1	2424	0	0	0
3581	0	155	1	1	0
3587	1	1686	0	0	1
3602	1	1828	0	1	0
3609	1	2690	1	1	0
3612	0	1180	1	1	1
3621	1	1061	0	1	0
3642	0	1294	1	0	1
3647	1	4152	0	0	0
3649	1	3340	1	0	0
3654	1	2904	0	0	0
3660	1	2144	1	1	0
3665	1	3126	1	1	0
3669	0	1113	0	0	1
3673	0	2164	0	0	1
3675	1	3159	1	0	1
3678	0	22	0	1	0
3680	1	2027	0	1	1
3686	1	2734	1	1	0
3693	0	1793	1	0	1
3710	1	3697	1	0	1
3713	0	-230	0	1	0
3718	1	2341	0	0	0
3725	0	511	0	1	0
3726	1	1566	0	1	0
3747	0	1295	1	1	1
3753	0	-567	0	1	0
3754	0	1740	0	1	1
3760	1	4029	1	0	0
3763	0	67	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
3765	1	3052	1	0	1
3769	0	868	1	1	0
3771	1	3324	1	0	0
3784	0	1375	0	1	1
3787	0	1823	1	1	1
3794	1	1726	0	0	0
3796	0	1055	1	1	0
3798	0	157	1	1	1
3809	0	1143	1	1	0
3812	1	3049	1	0	1
3819	1	1891	0	0	0
3828	0	1035	0	1	1
3831	0	1233	0	1	0
3833	0	809	1	1	1
3837	1	2576	0	1	0
3839	1	4203	1	0	0
3843	1	2069	0	0	0
3846	0	1445	1	0	0
3854	0	-251	1	1	0
3861	0	459	0	0	1
3864	1	1821	1	1	0
3868	0	-93	0	0	1
3869	0	657	0	1	0
3870	0	179	0	1	0
3883	0	1031	1	1	0
3886	0	-242	0	1	0
3889	1	2549	0	1	1
3894	1	2271	1	1	0
3907	0	228	0	1	1
3910	0	1729	0	1	1
3913	0	-751	0	1	1
3914	0	2371	0	0	1
3921	0	1050	1	1	0
3923	0	-181	0	0	1
3929	1	3269	0	0	1
3931	1	3268	0	1	0
3932	1	2103	0	0	0
3937	1	2572	0	0	1
3943	0	718	1	1	0
3956	1	2970	1	0	1
3957	1	1703	0	0	0
3961	1	2624	1	0	1
3971	0	1273	0	0	1
4004	0	1003	0	0	1
4005	0	995	1	1	1
4006	0	-220	1	1	1
4011	0	854	1	1	0
4013	0	1205	0	0	1
4014	0	603	0	1	0
4016	1	2529	1	1	0
4017	0	-216	0	1	1
4020	0	962	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
4022	0	1501	0	1	1
4026	0	1266	0	0	1
4032	0	1400	1	1	1
4043	0	938	0	1	1
4045	0	1405	1	1	1
4048	0	1274	1	0	1
4051	0	1229	0	1	1
4052	1	2648	0	0	1
4056	0	21	0	1	0
4059	0	-221	0	0	1
4069	0	-58	0	1	0
4074	1	1760	0	1	0
4076	0	2410	1	0	1
4077	1	3773	1	0	1
4079	1	3308	1	0	0
4081	1	3484	0	0	0
4088	0	557	0	1	0
4105	0	1736	1	0	1
4125	0	1189	1	1	1
4134	1	2243	1	0	0
4139	0	-1471	0	1	1
4146	0	1302	0	1	0
4149	0	950	1	1	1
4151	1	3912	0	0	1
4155	0	23	0	0	0
4157	0	1070	1	1	1
4168	1	2965	0	0	0
4170	0	1450	1	1	1
4174	0	414	1	1	1
4179	1	2372	0	0	1
4185	0	-213	0	1	1
4199	1	3697	0	0	1
4205	0	499	0	1	1
4208	0	-458	0	1	0
4211	1	2981	0	0	0
4212	0	374	1	1	1
4215	1	3172	1	0	0
4217	0	1507	0	1	1
4219	1	4195	0	0	0
4226	1	1447	0	1	0
4227	1	1324	0	1	1
4229	0	-24	1	1	1
4231	0	1560	1	1	1
4233	0	-962	0	1	1
4237	1	2253	0	1	1
4243	1	2846	0	0	0
4248	0	2268	0	0	0
4255	0	1591	0	1	1
4262	0	1154	1	0	0
4266	1	2828	0	0	0
4268	1	1872	0	1	0
4270	1	4018	0	0	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
4273	0	-46	0	1	1
4276	0	1039	1	1	1
4277	0	1495	1	1	1
4279	1	1975	1	1	1
4299	0	534	1	0	0
4313	0	394	1	0	0
4322	0	615	0	1	1
4324	0	683	1	1	0
4328	1	2248	0	1	0
4331	1	1856	1	0	0
4335	0	620	1	1	1
4337	1	1959	1	0	0
4338	1	2169	1	0	1
4343	0	533	0	0	0
4347	0	944	1	1	1
4355	1	3912	0	0	0
4357	0	-950	0	1	1
4359	0	216	1	1	1
4362	0	2222	0	0	0
4368	1	2967	0	0	0
4374	0	210	1	1	1
4375	1	2033	1	0	1
4378	1	2031	0	1	1
4381	1	3830	0	0	1
4387	0	1425	0	1	1
4400	0	-742	1	1	1
4423	0	749	0	1	1
4424	0	145	0	1	0
4428	1	1593	0	0	1
4433	1	4551	0	0	1
4436	1	2249	0	1	0
4437	0	1021	1	0	0
4439	1	1686	1	0	0
4449	0	1162	0	1	1
4456	0	434	1	1	1
4463	0	905	1	1	1
4467	0	424	1	0	1
4468	0	525	0	0	1
4469	0	337	1	1	1
4472	0	2250	1	1	1
4473	0	159	1	1	0
4476	1	3024	0	0	0
4500	0	-346	0	1	1
4509	0	1191	0	1	0
4513	1	4754	0	1	0
4521	0	2182	1	1	1
4527	1	3462	0	1	1
4530	1	1752	0	1	0
4532	1	2238	0	1	0
4533	0	621	0	1	1
4535	1	1919	0	1	0
4536	1	3066	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
4542	1	1713	1	1	0
4551	1	3871	0	1	0
4554	0	655	0	1	1
4555	1	1689	0	0	0
4564	0	1782	0	0	1
4572	1	3457	0	1	1
4573	0	2325	1	1	1
4577	0	2087	0	0	1
4579	1	1753	1	1	0
4583	0	-876	1	0	1
4584	1	2004	0	1	0
4596	0	-71	1	0	0
4599	0	1586	0	1	1
4607	0	2538	0	0	1
4609	1	2608	0	1	0
4610	0	-113	1	0	1
4616	1	2887	0	0	1
4617	0	1113	0	0	1
4633	0	1230	0	1	0
4638	1	2066	1	1	0
4641	0	192	1	0	1
4653	1	2128	0	1	0
4655	1	2538	1	1	1
4659	1	1635	0	1	0
4669	0	722	1	1	0
4678	0	1165	1	1	0
4685	1	3219	1	0	0
4686	1	2553	0	0	1
4691	0	1810	1	1	0
4695	0	1642	1	0	0
4698	0	1876	0	1	0
4700	1	2999	0	1	0
4711	0	2237	1	0	1
4722	0	552	0	1	1
4727	1	1775	0	1	0
4756	0	-1117	1	1	0
4762	0	1537	1	1	1
4763	1	2091	1	0	0
4766	0	632	0	1	1
4770	0	795	1	0	1
4784	1	2369	1	1	1
4791	0	914	0	0	0
4795	0	313	1	1	1
4799	1	3938	0	0	1
4802	1	2499	0	0	1
4805	1	3252	1	0	1
4814	1	2344	1	0	0
4816	1	2537	1	1	0
4817	0	369	1	1	1
4822	1	1553	1	1	0
4827	1	2792	1	0	1
4833	0	1629	1	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
4836	0	122	1	1	1
4842	0	3044	0	0	1
4844	0	1155	1	0	0
4845	1	1706	1	1	0
4849	0	436	0	1	1
4850	0	1681	0	0	1
4860	0	-397	1	0	0
4863	0	1833	0	0	0
4871	0	2105	0	0	0
4878	0	990	0	1	0
4881	1	3296	0	0	1
4888	1	2694	0	0	0
4900	0	813	0	1	1
4906	1	2598	0	1	0
4909	0	24	1	1	1
4916	0	682	1	1	1
4918	0	2122	1	1	1
4926	1	1677	0	1	0
4928	0	1539	0	1	1
4941	1	1693	0	0	0
4946	0	2198	0	0	1
4949	0	679	1	1	1
4956	0	122	1	1	0
4966	0	326	0	1	1
4969	1	1723	0	1	0
4973	0	810	1	1	1
4978	1	2092	0	1	1
4982	0	1510	0	1	1
4985	0	350	1	1	1
4991	0	1417	1	0	1
4998	0	-407	0	0	0
5000	1	3382	1	1	0
5004	1	2449	1	0	1
5005	1	3370	0	0	1
5011	1	3047	0	1	0
5016	1	1698	0	0	1
5018	0	313	0	1	1
5034	0	1267	0	1	0
5038	0	-628	1	0	1
5042	0	845	1	0	0
5046	0	563	1	1	0
5051	0	1408	0	0	1
5054	0	1296	0	1	1
5057	1	2075	0	1	1
5062	0	94	0	1	1
5063	0	-18	1	0	1
5065	0	955	1	1	0
5066	0	1530	1	1	1
5076	0	1634	1	0	1
5089	0	1001	0	1	0
5092	1	2789	0	0	1
5093	1	1517	0	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
5094	0	-281	0	1	0
5098	1	3960	0	0	1
5102	0	-71	0	1	0
5112	1	2667	1	0	1
5117	1	2711	1	1	0
5127	1	2304	1	1	1
5130	1	2915	1	1	1
5131	1	2414	0	0	1
5132	1	2720	1	0	1
5135	1	2882	0	1	0
5136	0	-700	1	1	1
5147	1	2658	1	1	1
5157	0	903	1	0	0
5160	0	1445	0	1	0
5165	0	-501	1	0	1
5166	1	2756	0	1	0
5172	1	3118	1	0	0
5173	0	876	0	1	0
5179	1	3351	1	1	0
5184	1	2617	1	1	0
5187	0	1884	0	0	1
5191	0	895	0	0	1
5193	0	1333	0	1	1
5194	0	1647	1	1	1
5199	0	1806	1	1	0
5212	0	-942	0	1	1
5213	1	2012	1	1	0
5224	1	1653	0	0	0
5226	0	962	0	1	1
5239	0	1412	1	1	0
5252	1	4196	0	0	1
5264	0	1894	0	0	0
5266	0	-794	0	1	0
5271	0	-39	0	0	1
5273	0	-292	0	1	0
5276	1	2701	1	0	0
5278	0	334	1	1	1
5281	1	3641	0	0	0
5283	1	3248	1	1	0
5291	0	579	0	1	1
5294	0	1678	0	0	1
5296	1	2243	0	1	0
5297	1	4417	0	0	0
5313	0	-326	0	0	0
5314	1	1885	1	1	1
5321	0	1633	0	0	1
5325	0	-1208	0	1	1
5326	0	1756	1	0	1
5328	0	109	1	1	1
5334	0	1178	1	1	0
5338	1	1803	0	1	0
5344	0	2334	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
5348	0	2064	1	0	1
5352	0	1278	1	1	0
5353	0	1026	0	1	0
5354	1	2001	0	0	0
5361	1	3607	1	1	0
5364	0	94	1	1	1
5365	0	457	0	1	0
5367	1	1884	0	1	0
5379	1	2425	0	1	1
5382	1	2945	1	0	0
5386	1	1194	0	0	1
5395	0	1224	1	1	1
5410	1	2787	1	0	1
5411	0	1005	0	1	0
5416	1	1336	0	0	1
5424	1	2649	1	1	0
5426	1	3181	1	1	1
5428	0	1430	0	1	1
5430	1	1552	0	0	1
5433	0	1176	0	1	1
5437	0	-570	1	0	1
5440	0	2061	0	0	1
5442	1	4527	1	0	0
5445	1	2535	0	1	0
5449	0	1026	0	1	1
5452	1	1238	0	1	0
5460	1	1737	1	0	1
5461	0	-544	1	1	0
5465	0	2233	1	0	0
5467	0	284	0	1	1
5471	1	1675	1	1	1
5474	1	4717	1	0	1
5475	0	-681	1	1	1
5480	0	188	1	1	1
5481	0	1808	1	1	0
5484	0	596	0	1	1
5494	0	1300	1	0	1
5495	1	3207	1	0	0
5497	0	526	0	1	0
5499	1	3112	1	0	1
5507	0	650	1	1	0
5510	0	1266	0	0	0
5515	0	1168	0	1	0
5516	0	34	1	1	1
5517	0	1501	1	0	1
5524	0	-223	1	0	1
5530	0	1708	0	0	0
5534	1	1680	0	1	0
5543	1	2259	1	1	0
5545	1	2715	1	1	0
5558	0	1121	0	0	1
5562	0	2102	1	0	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
5573	1	3920	1	1	0
5581	0	1773	0	1	0
5583	1	2886	1	1	0
5587	1	1760	1	0	0
5589	1	4308	0	0	1
5591	0	1879	0	0	1
5596	1	2483	0	1	1
5606	1	3587	1	0	0
5608	0	2153	1	0	1
5611	0	-103	0	1	1
5612	0	1752	1	1	1
5614	0	920	1	0	0
5620	0	88	0	1	1
5623	0	-81	0	1	0
5624	0	1809	1	1	1
5626	1	1990	0	1	1
5633	0	1689	0	0	1
5635	0	688	0	0	1
5640	1	2178	0	0	1
5643	0	1416	0	0	1
5644	1	1756	0	0	0
5653	1	1778	1	1	0
5663	0	-109	0	1	0
5664	1	1972	1	0	0
5667	1	2787	1	1	0
5671	1	2852	1	1	0
5673	1	3277	1	1	0
5676	0	431	0	1	1
5678	0	1039	0	1	1
5698	1	2561	1	1	1
5700	0	453	1	1	1
5705	1	1670	0	0	0
5706	1	3396	0	0	0
5711	0	40	0	1	1
5712	1	3754	0	1	1
5716	0	1269	1	0	1
5719	1	2409	1	1	1
5725	1	3479	0	0	0
5728	0	-337	0	1	1
5734	0	558	0	1	1
5735	0	578	0	1	0
5743	0	2425	1	1	1
5754	0	1316	0	1	0
5755	0	2042	1	0	1
5756	0	1109	0	1	1
5766	0	402	0	1	1
5770	1	2644	0	0	1
5774	0	1077	1	0	1
5775	0	-288	0	0	0
5776	0	1000	0	1	1
5778	0	32	0	0	0
5786	1	2989	0	0	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
5787	1	2115	0	0	0
5791	0	1981	1	1	1
5794	0	1462	1	1	1
5803	0	1476	0	1	1
5804	0	780	0	1	0
5808	0	1357	0	1	0
5810	0	-295	1	1	0
5813	1	2340	0	1	1
5828	0	1975	1	0	0
5839	1	2371	0	0	1
5842	1	2199	1	0	1
5843	0	172	0	1	1
5844	0	1782	1	1	1
5847	1	3895	1	0	1
5851	0	324	1	0	1
5854	0	2344	0	0	1
5857	0	-508	1	1	1
5866	1	2908	0	1	0
5874	1	1009	0	1	0
5886	0	1547	0	0	0
5895	0	114	0	1	0
5897	0	-486	0	1	1
5898	0	932	0	1	0
5900	1	4288	0	0	1
5902	1	2460	1	1	0
5908	1	2771	0	0	0
5909	0	-144	0	1	0
5912	0	117	0	1	1
5913	0	1355	1	1	1
5917	1	2342	1	0	0
5918	1	3563	0	0	0
5921	0	912	1	1	1
5931	0	868	0	1	0
5942	1	1683	0	0	0
5943	1	3491	1	0	1
5950	0	381	1	0	1
5954	0	-656	0	1	1
5983	0	-687	0	1	1
5995	1	3521	0	0	0
6002	0	1388	0	0	1
6005	0	67	0	1	0
6009	1	2378	0	1	1
6011	0	-617	0	1	1
6012	0	-722	1	0	1
6019	0	887	0	1	1
6021	0	1370	1	1	0
6029	1	4323	0	0	0
6036	1	2261	0	0	0
6037	0	-915	1	1	1
6038	0	-218	0	1	1
6043	0	412	1	1	0
6045	0	1079	1	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
6047	1	3552	0	0	1
6048	0	-214	0	1	1
6061	1	2319	1	1	1
6063	0	1518	0	0	1
6064	0	1178	0	0	0
6068	1	3770	0	0	0
6069	0	-133	1	1	1
6070	1	2539	1	1	0
6071	0	1346	0	1	0
6074	1	2384	0	0	0
6079	0	1789	0	0	1
6082	0	544	1	1	1
6088	1	3322	1	0	0
6094	0	1492	1	1	1
6095	0	1044	0	1	0
6098	1	2627	0	1	1
6102	0	-109	0	1	1
6105	1	2616	0	1	1
6113	0	1008	1	0	1
6116	0	1206	1	1	0
6120	1	2708	1	0	0
6121	1	1278	1	0	1
6126	0	1969	1	0	1
6144	0	1019	1	1	1
6145	0	396	0	0	1
6153	0	2915	1	0	1
6156	0	1438	1	0	1
6159	0	1203	1	0	0
6162	0	1797	1	1	0
6184	1	3076	0	1	0
6188	1	2379	1	0	1
6189	0	2512	1	1	0
6191	1	2908	0	0	1
6211	1	2812	0	0	1
6216	0	1851	0	0	0
6218	1	2632	0	1	0
6222	0	1202	1	1	0
6235	0	300	1	0	0
6245	0	1135	0	0	1
6248	1	4049	0	1	0
6253	0	1444	1	1	0
6256	0	-202	1	0	1
6257	1	2165	1	0	0
6259	1	2444	1	1	1
6266	0	1721	1	1	1
6268	1	1027	0	1	0
6275	0	2162	1	0	1
6280	1	3212	1	0	1
6283	1	2787	1	0	1
6288	0	528	1	1	0
6289	0	896	0	1	0
6301	0	1044	1	0	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
6308	0	1877	1	1	1
6314	0	213	1	0	1
6315	0	1124	0	0	1
6316	1	3072	0	0	1
6317	1	2665	0	0	1
6318	0	-44	1	1	1
6323	1	3238	0	0	0
6329	1	3052	0	1	0
6336	1	1796	0	0	0
6341	1	4818	0	0	0
6348	0	2109	1	0	0
6349	0	701	1	1	0
6365	0	575	1	1	1
6372	0	0	0	1	1
6376	0	98	0	0	1
6378	0	374	0	1	1
6379	1	3428	0	0	0
6382	0	754	0	1	0
6383	1	3445	1	1	1
6389	1	2763	0	0	0
6390	0	133	1	1	1
6392	0	996	0	1	0
6394	1	2773	1	0	1
6402	0	746	0	1	1
6404	1	1405	0	1	0
6405	0	-775	0	0	1
6406	0	1000	0	1	0
6409	0	1341	1	1	1
6410	0	176	0	1	1
6411	0	1119	0	1	1
6421	0	298	1	1	0
6428	0	1841	0	0	1
6429	0	1291	0	1	1
6432	0	682	0	1	1
6436	0	-902	0	1	1
6437	0	1710	0	1	1
6438	0	1646	0	1	1
6445	0	465	0	0	1
6447	1	2129	0	0	0
6450	0	117	1	1	1
6462	0	1533	1	0	1
6467	1	3756	1	1	0
6478	0	-311	0	1	1
6484	0	1572	1	0	1
6492	1	2069	1	0	1
6497	0	-132	0	1	1
6504	0	1256	0	0	1
6505	0	1248	1	1	1
6513	1	2401	0	1	0
6525	0	1668	0	1	1
6526	1	2644	1	1	1
6528	0	116	1	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
6540	0	-1284	0	1	1
6542	0	945	1	0	1
6544	1	2303	0	0	0
6548	0	1103	1	1	1
6552	0	1826	1	0	1
6558	0	-222	1	1	1
6567	0	244	1	1	1
6569	1	4028	0	0	1
6572	0	165	0	0	1
6577	0	1341	0	0	1
6581	1	2511	1	1	1
6588	1	3386	1	0	0
6591	1	3130	0	1	0
6594	1	2230	1	1	1
6600	1	3635	0	0	1
6602	1	2734	1	0	1
6604	0	201	1	0	1
6605	0	-210	0	1	1
6614	0	2259	1	1	1
6616	1	2323	1	0	0
6621	1	2258	1	1	1
6640	1	2219	1	1	1
6641	1	2431	0	0	0
6643	0	818	1	1	1
6644	0	1572	0	0	1
6649	1	3363	0	0	0
6650	1	3195	0	0	0
6655	1	2161	1	1	0
6661	0	-776	0	1	0
6672	0	2149	1	1	1
6677	0	119	0	1	1
6688	0	679	1	0	1
6689	0	615	0	1	1
6691	0	-243	0	1	1
6692	1	1365	0	1	0
6694	1	3378	1	0	0
6702	1	2506	0	0	1
6714	0	-37	0	0	0
6716	1	2459	0	1	1
6724	0	1072	0	1	0
6725	0	1009	0	1	1
6730	0	1899	1	0	1
6735	1	2566	1	1	0
6738	1	2487	0	0	1
6739	0	1687	1	1	1
6743	0	1159	1	0	1
6747	0	1346	1	1	1
6750	1	3606	0	0	0
6751	1	2307	1	0	1
6753	1	1441	1	1	0
6754	1	2619	1	1	0
6755	0	942	0	0	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
6762	0	1169	0	1	1
6764	0	566	0	1	1
6772	1	2856	1	0	0
6774	0	741	0	1	1
6787	0	1426	1	0	0
6789	0	261	0	1	0
6793	0	1969	0	1	1
6798	0	-668	1	1	0
6799	0	-650	1	1	0
6800	0	738	0	1	1
6802	0	-45	1	1	0
6808	1	1317	0	0	0
6809	0	749	1	1	1
6812	0	-628	0	1	0
6814	1	4830	0	0	0
6816	1	1965	1	1	0
6822	0	1842	1	1	0
6829	1	2503	0	1	1
6834	1	4649	1	0	0
6836	0	-533	1	1	0
6839	0	1609	0	0	1
6840	1	1931	0	1	1
6843	0	1300	1	0	0
6846	1	2410	0	0	0
6848	0	-131	1	1	1
6852	0	54	1	1	1
6856	0	644	0	1	1
6860	0	819	0	1	1
6866	1	1697	0	1	0
6870	1	1270	1	1	0
6878	1	2077	0	1	0
6880	0	1560	1	1	1
6885	0	-596	1	1	1
6897	0	199	1	1	0
6902	1	3218	1	0	0
6904	1	2476	1	0	1
6907	0	-32	0	1	1
6909	0	2127	1	1	1
6914	1	2819	0	0	0
6915	1	2719	1	1	0
6922	0	1015	1	1	0
6924	0	1150	1	1	1
6933	0	-338	0	1	1
6934	0	714	0	1	0
6941	0	1468	0	0	1
6957	0	1229	0	0	0
6960	0	736	1	1	1
6969	0	630	1	0	1
6975	0	1312	1	0	1
6980	1	3653	1	0	0
6983	0	1400	0	1	1
6987	0	1077	0	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
6994	0	353	0	1	1
6997	0	-1659	0	1	1
7002	0	1061	1	1	1
7010	0	-107	1	1	0
7015	1	2488	0	0	1
7019	0	908	1	0	1
7022	1	3209	0	1	1
7025	0	445	0	1	1
7029	0	742	1	0	1
7031	0	1330	1	1	1
7037	1	1957	0	1	0
7038	0	1662	1	1	0
7043	0	617	0	1	1
7049	0	1197	1	1	1
7052	0	982	0	1	1
7053	1	1756	0	0	0
7056	0	-1203	0	1	1
7057	1	2192	0	1	0
7080	0	1829	1	1	1
7086	0	2062	1	0	1
7087	0	96	0	0	1
7105	1	1787	0	1	0
7108	0	-840	1	1	1
7121	1	1476	0	1	0
7122	0	1757	0	1	0
7125	1	2463	0	1	0
7132	1	1860	0	0	1
7134	0	1342	0	0	0
7151	0	1354	0	1	0
7152	1	3139	1	0	0
7157	0	1311	0	1	1
7159	0	1940	1	1	1
7166	1	3517	0	0	1
7167	0	507	0	1	1
7177	0	-195	1	1	0
7179	1	3746	0	1	0
7181	0	2101	1	0	1
7183	0	2095	1	0	1
7186	0	-215	1	1	1
7193	0	884	1	0	1
7205	0	-428	0	1	1
7207	0	-683	1	0	1
7209	1	2183	0	1	0
7216	0	1851	1	0	1
7232	1	4041	0	0	0
7235	0	964	0	1	0
7238	1	1845	0	0	0
7240	1	1673	0	1	0
7243	0	1355	0	1	0
7252	1	2686	1	0	1
7269	0	2225	1	0	0
7275	0	-166	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
7281	0	1540	1	1	1
7283	0	1369	0	0	0
7287	0	2097	1	0	1
7289	0	2389	1	0	1
7291	1	2612	1	1	0
7294	0	-343	1	0	0
7304	1	1617	1	1	1
7308	0	1729	1	1	1
7313	0	663	1	0	0
7319	1	1884	0	1	1
7325	0	917	1	0	1
7326	0	1101	1	0	1
7330	1	1873	0	0	0
7332	0	-525	0	1	1
7337	1	1940	0	1	1
7341	0	1255	0	0	1
7346	1	3994	0	0	1
7353	1	3190	1	1	1
7354	1	2960	0	1	0
7361	1	2643	0	0	1
7366	1	2788	1	0	0
7368	0	-539	0	1	0
7372	0	-647	1	1	1
7375	1	1807	0	1	0
7377	1	1231	0	1	0
7380	0	1279	0	1	1
7382	1	2628	0	1	1
7385	1	3705	0	0	1
7392	1	3917	0	0	0
7395	0	753	1	1	1
7397	0	1282	0	1	1
7403	0	126	0	0	1
7406	1	3434	0	1	1
7409	1	2962	1	0	0
7410	0	1178	0	0	1
7412	0	876	1	1	1
7419	0	1195	0	0	0
7425	0	-444	0	1	1
7435	0	1323	0	0	1
7438	0	1744	1	1	0
7440	0	1828	1	1	1
7447	0	435	0	0	1
7449	1	2668	0	0	0
7456	0	1412	1	0	0
7464	0	1123	0	1	1
7478	0	2266	1	1	1
7480	0	177	0	1	1
7481	1	2381	0	0	0
7483	0	1661	0	1	1
7484	0	1971	0	0	1
7491	1	3186	1	0	1
7494	1	2614	1	0	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
7501	1	2376	0	1	1
7503	1	4694	1	0	0
7509	1	2458	0	1	1
7517	0	864	1	1	1
7518	0	1474	0	1	0
7519	1	2569	1	1	0
7521	1	1661	0	1	0
7522	1	2840	0	0	0
7536	0	1327	0	0	0
7539	0	603	1	1	1
7547	1	2734	1	1	0
7549	0	-92	1	1	1
7552	1	1697	0	1	0
7554	1	2316	1	1	0
7556	0	965	0	0	0
7564	0	857	0	1	1
7566	0	1856	1	0	1
7570	0	1739	1	0	0
7571	0	-189	1	1	1
7572	0	1262	0	0	0
7575	0	1514	1	1	1
7586	0	121	1	0	1
7589	0	1242	1	1	1
7590	0	87	0	0	1
7597	1	1698	0	1	1
7602	0	416	1	1	1
7604	1	2446	1	1	0
7605	1	1648	1	1	1
7612	1	3745	1	0	0
7615	0	714	0	0	1
7617	0	1463	1	1	0
7624	0	861	1	0	0
7632	0	985	1	1	1
7639	0	1284	1	0	1
7642	1	3059	1	1	1
7643	0	1080	0	0	1
7649	1	1617	0	0	1
7650	1	2428	0	1	0
7653	0	2252	1	0	1
7654	1	1460	0	1	0
7657	1	3185	0	0	1
7662	0	1464	1	1	0
7669	1	4164	1	0	0
7671	0	-1304	0	1	1
7675	0	237	1	1	1
7678	0	2066	0	0	1
7682	1	4044	1	0	1
7688	1	1767	0	1	0
7689	0	1557	1	0	1
7690	0	1735	1	1	1
7692	1	2470	0	1	0
7699	1	2119	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
7705	1	2757	1	0	0
7712	0	911	0	1	0
7726	1	2698	1	0	0
7728	0	1638	0	1	1
7735	1	1625	0	1	0
7737	1	2594	1	1	0
7739	0	328	0	1	0
7743	1	3452	0	0	0
7744	0	1629	0	0	1
7746	0	675	1	1	0
7749	0	1910	1	0	1
7750	1	1439	0	1	0
7752	0	539	0	1	0
7755	0	964	0	1	1
7756	1	4167	1	0	0
7762	0	1900	1	1	0
7764	1	2838	0	0	0
7769	0	642	0	0	0
7770	1	2051	1	0	1
7776	0	144	0	1	1
7778	0	937	1	1	0
7784	1	2809	0	0	1
7786	0	1644	0	1	0
7789	0	1255	0	0	1
7793	1	1822	0	0	1
7794	0	1015	1	1	1
7804	0	1448	0	0	1
7811	0	899	1	1	1
7813	0	1202	1	1	0
7815	1	1612	0	1	0
7817	0	-898	1	1	0
7818	0	3510	1	0	1
7821	0	1125	1	1	1
7825	0	-913	0	1	1
7830	1	2087	0	1	0
7832	0	1033	1	1	1
7835	0	-960	1	1	0
7839	0	2023	1	1	1
7842	0	359	1	1	1
7849	1	3340	1	0	0
7856	1	2458	1	1	0
7857	0	-1395	1	1	1
7863	0	197	1	1	1
7866	0	1000	0	1	1
7871	0	-36	0	1	0
7875	1	2182	0	0	0
7882	1	4656	1	0	0
7887	1	3121	1	0	1
7888	1	3953	1	0	0
7891	1	5439	0	0	1
7895	0	183	1	1	0
7901	1	2356	1	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
7906	0	1687	0	0	1
7908	1	4330	0	0	0
7917	0	1038	0	0	1
7924	1	2860	0	0	0
7948	0	2294	1	1	1
7950	1	3286	1	1	0
7955	0	1259	0	0	1
7957	0	-296	0	1	1
7959	0	1167	1	1	0
7967	0	70	0	0	1
7969	0	1394	1	1	1
7971	0	1766	1	1	0
7974	0	1876	1	0	1
7976	0	250	1	1	1
7986	1	3580	0	0	0
7987	1	2899	0	0	0
7993	0	1163	0	1	0
7996	1	2763	0	1	0
7998	0	1249	0	0	1
8018	0	-31	0	0	1
8019	1	2335	1	1	1
8027	0	-1116	0	1	1
8036	0	329	0	0	1
8040	0	566	0	1	1
8044	0	1272	1	1	0
8050	0	875	0	0	1
8052	1	3204	1	0	1
8054	1	1106	0	1	1
8057	1	2837	0	0	1
8058	0	939	0	0	1
8059	1	2498	0	0	1
8066	1	4672	0	0	1
8070	0	831	1	1	1
8072	1	2986	1	1	1
8078	0	-415	1	1	0
8079	0	1117	1	1	1
8080	1	1128	0	1	1
8081	0	1523	1	1	1
8088	0	1083	1	1	1
8091	1	3222	0	0	0
8094	0	1123	0	0	1
8095	1	4206	0	1	1
8099	0	675	0	1	1
8101	0	1755	1	1	1
8102	0	-100	0	0	1
8116	1	1567	0	0	0
8125	0	2130	1	1	0
8134	1	2455	0	1	0
8139	0	272	1	1	1
8141	0	1099	1	1	0
8147	0	500	1	1	0
8158	1	1753	0	0	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
8160	0	422	1	0	1
8165	1	1168	1	1	0
8187	0	2025	1	1	1
8205	1	2655	1	1	1
8209	0	2558	1	0	1
8211	1	2201	0	1	1
8232	0	-904	1	1	1
8236	0	443	1	1	1
8237	0	1896	0	0	1
8238	1	3388	1	0	0
8245	1	2106	0	0	1
8256	0	1053	0	0	1
8268	0	228	1	1	1
8269	0	-625	0	1	0
8270	1	1602	1	1	0
8286	0	502	0	1	1
8289	0	291	1	1	1
8301	1	2168	1	1	0
8305	0	704	0	1	1
8310	1	2135	1	0	1
8312	0	716	1	0	1
8318	1	4678	0	0	0
8321	0	813	0	0	1
8328	0	30	0	1	1
8331	0	124	0	1	1
8334	1	1263	0	1	0
8344	0	2345	1	0	1
8345	0	2247	0	0	0
8352	0	1201	0	1	0
8358	1	2622	1	1	1
8359	0	1131	1	1	0
8360	0	1510	0	0	0
8365	0	2972	0	0	0
8366	0	893	1	1	0
8369	1	1772	0	1	0
8373	0	565	0	1	0
8378	0	2038	1	1	1
8392	0	1012	0	0	1
8397	1	2793	0	1	1
8399	0	526	0	1	0
8400	0	984	1	1	1
8405	1	4168	1	0	1
8406	0	114	0	1	1
8410	0	622	1	1	1
8413	0	276	1	0	0
8414	1	2629	1	1	1
8416	1	3956	1	0	1
8426	0	107	0	1	0
8434	0	1740	0	1	0
8439	0	961	0	1	1
8440	0	416	0	1	1
8475	0	-201	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
8480	0	1282	0	0	0
8497	0	1631	1	1	1
8499	1	4070	1	0	1
8500	1	1842	1	0	0
8501	0	382	0	0	1
8502	1	2641	0	1	0
8518	1	1422	0	1	0
8520	1	2447	1	0	0
8523	1	2175	1	0	0
8525	0	2171	1	1	1
8532	0	1237	1	1	1
8535	0	2451	1	1	0
8543	0	807	0	1	1
8554	0	1107	1	1	0
8560	0	1007	0	1	1
8561	0	1893	1	1	1
8563	0	123	1	0	1
8566	1	3222	0	0	0
8570	1	2029	0	0	0
8572	0	62	1	1	0
8582	0	571	1	1	1
8583	0	1578	0	1	1
8587	0	1865	0	0	0
8592	0	1163	0	1	0
8593	1	1964	0	0	0
8607	0	-423	0	1	0
8609	0	1731	1	0	1
8610	0	899	1	0	0
8614	1	2358	0	1	1
8616	1	1986	0	0	0
8622	0	747	1	1	0
8623	0	110	0	1	0
8624	0	2076	1	0	1
8633	0	1882	0	1	1
8641	1	2566	0	0	1
8644	1	3034	0	0	0
8649	1	3133	1	0	0
8653	0	880	1	1	1
8657	0	1174	1	0	1
8658	0	254	0	1	1
8663	0	172	0	0	1
8672	1	2434	0	1	0
8680	1	2178	0	0	0
8684	1	1279	0	0	0
8687	0	944	0	1	1
8688	0	-349	1	1	1
8690	0	1211	0	1	0
8712	0	1013	1	0	0
8717	0	1595	0	0	1
8730	0	342	1	1	1
8739	0	1538	1	0	1
8744	0	150	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
8747	1	1518	0	1	0
8748	0	1520	1	0	1
8751	1	3635	0	1	0
8758	0	1908	1	0	1
8761	1	2299	1	0	1
8763	0	-1153	0	1	0
8764	0	2002	0	0	1
8765	0	1071	0	1	0
8773	0	130	0	1	1
8780	0	1305	1	1	1
8781	0	767	0	1	1
8782	1	2009	0	0	1
8785	0	1439	1	0	1
8786	0	577	1	1	0
8797	1	4227	0	0	0
8799	0	477	0	1	1
8807	1	3146	1	1	0
8816	0	233	0	0	0
8817	0	1291	1	0	1
8826	0	1714	0	1	0
8833	0	1202	1	1	1
8834	0	1053	1	1	1
8835	0	1616	1	1	1
8840	0	812	1	1	1
8843	0	544	1	1	1
8849	0	1533	0	1	0
8855	0	1693	0	0	0
8861	0	1693	0	1	1
8862	0	748	1	0	1
8865	0	1967	1	1	0
8868	0	-1538	0	1	1
8870	0	-528	0	1	1
8880	1	1365	0	1	0
8885	0	-283	1	0	1
8894	0	1416	0	0	1
8895	0	321	0	1	1
8899	0	-495	0	1	0
8912	1	2255	1	0	1
8922	0	-760	0	1	1
8924	0	1102	0	1	1
8928	0	976	1	1	0
8932	1	1966	0	1	0
8943	0	905	0	0	1
8945	0	714	1	1	1
8946	0	269	0	0	0
8954	1	2226	1	0	1
8958	1	1723	0	1	0
8960	1	3084	0	0	0
8965	0	1674	0	0	1
8966	0	-235	0	1	0
8967	0	913	1	1	1
8969	1	1597	0	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
8980	0	1371	0	0	0
8984	0	-683	1	1	0
8985	1	3624	1	0	0
8988	1	2519	0	1	1
8989	1	2602	0	1	0
8995	0	841	0	1	0
9004	0	1318	1	1	0
9010	0	-444	0	1	1
9012	0	1867	1	0	1
9018	1	2340	0	0	0
9036	0	860	0	1	0
9037	0	1499	1	1	1
9040	0	788	0	1	1
9041	1	2257	1	0	0
9044	1	3184	1	0	1
9045	0	978	0	1	1
9047	1	2165	0	1	0
9049	0	-576	0	1	0
9061	0	-480	1	1	0
9062	1	2381	1	1	1
9076	1	2541	0	0	1
9079	0	340	0	1	0
9081	0	1372	1	0	1
9082	0	1428	1	1	0
9089	1	2394	0	1	0
9092	0	2131	1	1	1
9094	1	3299	1	0	1
9115	0	-246	0	0	1
9117	1	1372	0	0	0
9118	0	2599	0	0	0
9120	0	59	1	1	0
9124	0	-81	0	0	1
9128	0	2241	1	1	1
9135	1	2793	1	1	0
9136	1	3605	0	0	0
9138	0	1261	0	1	1
9157	1	2052	0	0	0
9176	0	-212	1	1	1
9183	0	1903	1	1	1
9187	1	2241	0	0	0
9188	1	1900	0	1	1
9190	0	1072	0	0	0
9197	0	-281	1	1	0
9200	0	55	0	1	0
9201	0	1182	0	1	1
9203	0	-298	1	1	0
9212	1	1531	0	1	1
9213	0	388	1	1	1
9214	0	1467	0	1	0
9217	0	1197	0	1	1
9219	0	-1777	0	1	1
9220	0	1296	0	1	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
9221	0	1074	0	1	1
9237	0	-460	0	1	1
9240	0	459	0	0	1
9241	0	-960	1	1	1
9248	0	1185	0	1	0
9253	1	2718	1	0	0
9259	1	3513	1	1	0
9267	0	1532	1	1	1
9271	1	2693	0	0	1
9273	0	945	0	1	0
9285	0	128	1	1	1
9290	0	2064	1	0	1
9291	0	1378	0	0	1
9293	0	1504	0	1	1
9294	0	248	1	1	1
9301	0	1440	1	0	0
9302	0	395	0	1	1
9312	0	-920	0	0	1
9316	1	2335	0	1	0
9319	1	3280	0	0	1
9328	0	1273	1	1	1
9331	1	2973	0	1	0
9338	0	370	1	1	1
9350	0	1628	0	0	1
9356	0	1139	1	0	1
9359	1	2164	0	1	0
9362	1	2675	1	0	1
9364	0	1258	1	1	1
9370	1	1337	0	1	0
9380	0	282	1	1	1
9386	0	864	0	1	0
9394	1	2309	0	0	0
9407	1	2017	1	1	0
9411	1	3914	0	0	1
9422	1	2540	1	0	1
9423	1	1665	0	1	1
9429	1	1345	0	1	0
9433	0	1489	1	0	1
9439	0	1166	1	1	1
9451	0	1157	0	1	1
9452	1	2695	0	0	0
9453	0	-1609	0	1	0
9460	0	-313	1	0	1
9465	0	204	1	1	0
9470	0	-42	0	0	0
9476	1	1863	0	0	0
9485	1	3079	1	0	0
9486	0	-107	0	1	1
9488	0	1710	1	1	0
9507	0	-1292	0	1	1
9508	1	3038	0	0	1
9517	1	1650	1	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
9521	0	2564	0	1	1
9528	0	1182	0	1	1
9532	1	2490	1	0	1
9536	0	1709	1	0	1
9540	0	1586	0	1	0
9542	0	1028	1	1	1
9546	0	2013	1	1	1
9548	0	1561	0	1	1
9549	0	1381	1	1	0
9554	0	1526	0	1	1
9555	1	2466	1	1	0
9558	0	886	1	0	1
9573	1	3338	0	1	0
9575	1	3660	1	0	0
9584	1	2960	1	1	0
9586	0	1854	0	0	0
9588	0	2054	0	0	0
9591	1	1834	1	0	1
9592	1	4230	0	1	0
9597	1	1886	0	1	1
9600	0	-24	0	1	1
9603	1	2381	0	0	0
9605	1	2743	0	1	1
9614	1	4190	0	1	1
9616	0	-587	0	1	0
9622	1	3055	1	0	0
9624	0	532	0	1	1
9629	1	4130	1	1	1
9633	0	673	0	1	1
9640	1	1968	0	0	1
9644	1	2376	0	0	1
9645	1	1919	1	1	0
9646	0	332	0	0	1
9648	1	4100	0	0	0
9649	0	-208	1	1	1
9660	0	1736	1	1	1
9664	1	2486	1	0	0
9675	0	142	1	1	1
9679	1	3112	0	1	0
9680	1	2601	1	1	0
9682	0	699	1	0	1
9697	0	-255	0	1	1
9701	0	1408	1	1	1
9704	1	1130	0	1	0
9705	0	2349	0	0	1
9707	1	1953	0	1	1
9714	0	248	0	1	1
9718	0	1049	0	1	1
9722	0	1118	1	1	1
9739	0	916	1	0	1
9747	1	2711	0	1	1
9751	0	734	0	1	0

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
9757	0	1030	0	1	1
9759	0	-791	0	1	0
9760	0	1018	0	1	0
9764	1	2476	0	0	1
9776	1	2255	0	0	1
9778	0	1221	0	1	1
9786	0	873	0	0	0
9803	1	2943	0	1	0
9804	0	1194	0	1	1
9815	0	609	0	1	1
9824	0	208	1	1	1
9825	0	1781	1	1	1
9826	0	2548	1	1	1
9827	0	-103	0	1	1
9833	0	661	1	1	1
9835	0	294	0	1	1
9860	1	2142	0	1	0
9865	0	1756	1	1	0
9871	0	2092	1	0	1
9874	1	1555	1	0	0
9880	0	1309	0	1	0
9882	1	3278	1	0	1
9885	0	1168	1	1	0
9888	1	3385	1	1	0
9892	0	368	0	1	1
9893	0	2259	1	0	1
9896	0	2052	1	0	1
9902	0	82	1	1	1
9906	0	1078	0	0	1
9910	1	1618	0	1	0
9914	1	1557	0	1	1
9918	1	1796	0	1	0
9920	0	1169	1	1	0
9926	0	1102	0	0	1
9931	0	568	1	1	0
9935	1	2289	1	0	0
9945	1	4276	0	0	0
9953	1	1772	0	0	0
9957	0	-441	1	1	0
9963	0	927	1	1	1
9972	1	1681	0	1	1
9976	1	3151	1	0	1
9979	1	2510	1	0	0
9980	0	-294	1	1	1
9982	0	1217	1	1	1
9991	1	3582	0	0	0
10000	0	1502	0	1	0
10003	0	1273	0	0	1
10005	1	4708	0	0	0
10014	0	-1068	0	1	1
10032	1	2352	1	1	0
10034	1	2429	0	0	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
10041	0	-607	0	1	1
10042	0	-620	0	1	0
10044	0	-107	0	1	0
10045	0	1312	0	1	0
10054	1	2959	0	0	1
10061	0	1999	0	0	1
10062	1	2623	0	1	1
10073	0	1193	1	1	1
10081	0	-71	0	0	0
10084	1	2179	0	0	1
10086	0	1233	1	0	1
10093	1	1842	1	0	0
10101	1	3195	0	1	0
10105	1	2765	0	1	0
10110	1	1591	0	0	0
10113	1	4184	1	1	0
10115	1	2447	0	1	1
10119	1	2959	1	1	1
10121	1	1199	0	1	0
10124	1	4781	0	0	1
10126	0	1776	0	1	1
10127	0	699	1	0	1
10145	0	859	0	1	1
10147	1	2452	0	0	0
10148	0	-398	0	1	1
10162	1	872	1	1	0
10163	0	439	1	1	0
10166	1	4260	0	0	1
10172	0	-48	1	1	0
10173	1	2344	1	1	0
10175	0	-755	0	1	0
10180	0	688	0	1	1
10186	0	281	0	1	1
10192	1	2822	0	0	0
10199	0	1455	0	0	0
10209	1	5572	1	0	1
10210	0	1172	0	1	1
10214	0	-22	0	1	1
10215	0	2170	1	1	0
10216	1	3664	0	0	0
10232	1	2523	1	1	1
10239	0	985	0	0	0
10249	0	500	0	1	1
10253	1	2393	1	1	1
10255	0	1693	0	1	1
10262	0	699	0	1	1
10264	0	-427	0	1	0
10266	0	1291	1	0	1
10268	0	2016	1	1	1
10271	0	2251	0	0	0
10272	0	1346	0	1	1
10276	1	2755	0	0	1

INDEX	TARGET_FLAG	TARGET_AMT	MALE	MARRIED	EDU_BACH_MAST_PHD
10277	0	-214	1	1	1
10279	0	1760	0	1	0
10281	0	137	1	1	0
10285	0	-1247	0	1	1
10294	0	2002	1	0	1
10300	0	670	0	1	0

Appendix

```
 library(dplyr)
 library(ggplot2)
 library(gridExtra)
 library(car)
 library(recommenderlab)
 library(knitr)
 library(caret)
 library(pROC)

desc <- 'data_desc.csv'

data_desc <- read.csv(desc)
kable(data_desc[,1:2])

insurance.trn  <- read.csv("https://raw.githubusercontent.com/Nguyver/DATA621-HW/master/HW4/insurance_t

insurance.evl  <- read.csv("https://raw.githubusercontent.com/Nguyver/DATA621-HW/master/HW4/insurance-e
glimpse(insurance.trn)
summary(insurance.trn)
cleanUpAmounts <- function(amount) {
  return(as.numeric(gsub("\\$*|,", "", amount)))
}
insurance.trn$INCOME <- sapply(insurance.trn$INCOME, cleanUpAmounts)
insurance.trn$HOME_VAL <- sapply(insurance.trn$HOME_VAL, cleanUpAmounts)
insurance.trn$BLUEBOOK <- sapply(insurance.trn$BLUEBOOK, cleanUpAmounts)
insurance.trn$OLDCLAIM <- sapply(insurance.trn$OLDCLAIM, cleanUpAmounts)
boxIncome <- ggplot(data = insurance.trn) + geom_boxplot(aes(x=1, y=INCOME)) + theme(axis.text=element_

histIncome <- ggplot(data = insurance.trn) + geom_histogram(aes(x=INCOME)) + theme(axis.text=element_te

grid.arrange(boxIncome, histIncome, ncol=2,  top = "INCOME: Box Plot & Histogram")
boxHomeVal <- ggplot(data = insurance.trn) + geom_boxplot(aes(x=1, y=HOME_VAL)) + theme(axis.text=eleme

histHomeVal <- ggplot(data = insurance.trn) + geom_histogram(aes(x=HOME_VAL)) + theme(axis.text=element

grid.arrange(boxHomeVal, histHomeVal, ncol=2,  top = "HOME_VAL: Box Plot & Histogram")

gg.age <- ggplot(data = insurance.trn) + geom_histogram(aes(x=AGE)) +  theme(axis.text=element_text(siz

gg.bluebook <- ggplot(data = insurance.trn) + geom_histogram(aes(x=BLUEBOOK)) +  theme(axis.text=elemen
```

```
gg.carage <- ggplot(data = insurance.trn) + geom_histogram(aes(x=CAR_AGE)) +  theme(axis.text=element_te

#Total Clamins (Past 5 Years) , Lets take average and draw the histogram.
insurance.trn$OLDCLAIM.AVG <- insurance.trn$OLDCLAIM/5
gg.oldclaimavg <- ggplot(data = insurance.trn) + geom_histogram(aes(x=log(OLDCLAIM.AVG)), binwidth = 0.5
insurance.trn$OLDCLAIM.AVG <- NULL
#People who have been customers for a long time are usually more safe.
gg.tif <- ggplot(data = insurance.trn) + geom_histogram(aes(x=TIF)) +  theme(axis.text=element_text(size

#TRAVTIME - Long drives to work usually suggest greater risk
gg.traveltm <- ggplot(data = insurance.trn) + geom_histogram(aes(x=TRAVTIME)) +  theme(axis.text=elemen

#YOJ - People who stay at a job for a long time are usually more safe
gg.yoj <- ggplot(data = insurance.trn) + geom_histogram(aes(x=YOJ)) +  theme(axis.text=element_text(size

grid.arrange(gg.age, gg.bluebook, gg.carage, gg.oldclaimavg,gg.tif,gg.traveltm,gg.yoj, ncol=2)
insurance.trn$INDEX <- NULL

cleanUpPrefix <- function(stringVal) {
  return(gsub('z_', '', stringVal))
}

insurance.trn$MSTATUS <- sapply(insurance.trn$MSTATUS, cleanUpPrefix)
insurance.trn$SEX <- sapply(insurance.trn$SEX, cleanUpPrefix)
insurance.trn$EDUCATION <- sapply(insurance.trn$EDUCATION, cleanUpPrefix)
insurance.trn$JOB <- sapply(insurance.trn$JOB, cleanUpPrefix)
insurance.trn$CAR_TYPE <- sapply(insurance.trn$CAR_TYPE, cleanUpPrefix)
insurance.trn$TARGET_FLAG <- as.factor(insurance.trn$TARGET_FLAG)
insurance.trn$PARENT1 <- as.factor(insurance.trn$PARENT1)
insurance.trn$SEX <- as.factor(insurance.trn$SEX)
insurance.trn$MSTATUS <- as.factor(insurance.trn$MSTATUS)
insurance.trn$EDUCATION <- as.factor(insurance.trn$EDUCATION)
insurance.trn$JOB <- as.factor(insurance.trn$JOB)
insurance.trn$CAR_USE <- as.factor(insurance.trn$CAR_USE)
insurance.trn$CAR_TYPE <- as.factor(insurance.trn$CAR_TYPE)
insurance.trn$RED_CAR <- as.factor(insurance.trn$RED_CAR)
insurance.trn$REVOKED <- as.factor(insurance.trn$REVOKED)
insurance.trn$URBANICITY <- as.factor(insurance.trn$URBANICITY)
insurance.trn$OLDCLAIM_YRLY_AVG <- insurance.trn$OLDCLAIM / 5
insurance.trn$OLDCLAIM <- NULL
ggplot(data = insurance.trn) + geom_histogram(aes(x=CAR_AGE), binwidth = 0.5) + theme(axis.text=element_

meanVal <- mean(insurance.trn$CAR_AGE, na.rm=TRUE)
medianVal <- median(insurance.trn$CAR_AGE, na.rm=TRUE)

insurance.trn$CAR_AGE<- ifelse( insurance.trn$CAR_AGE <=0 , medianVal, insurance.trn$CAR_AGE)
insurance.trn$CAR_AGE<- ifelse(is.na( insurance.trn$CAR_AGE )==TRUE, medianVal, insurance.trn$CAR_AGE)

summary(insurance.trn$CAR_AGE)

ggplot(data = insurance.trn) + geom_histogram(aes(x=YOJ), binwidth = 0.5) + theme(axis.text=element_tex

medianVal <- median(insurance.trn$YOJ, na.rm=TRUE)
```

```r
insurance.trn$YOJ<- ifelse(is.na( insurance.trn$YOJ )==TRUE, medianVal, insurance.trn$YOJ)

summary(insurance.trn$YOJ)

boxAge <- ggplot(data = insurance.trn) + geom_boxplot(aes(x=1, y=AGE)) + theme(axis.text=element_text(si
histAge <- ggplot(data = insurance.trn) + geom_histogram(aes(x=AGE), binwidth = 0.5) + theme(axis.text=e
grid.arrange(boxAge, histAge, ncol=2,  top = "AGE: Box Plot & Histogram")

medianVal <- median(insurance.trn$AGE, na.rm=TRUE)

insurance.trn$AGE<- ifelse(is.na( insurance.trn$AGE )==TRUE, medianVal, insurance.trn$AGE)

summary(insurance.trn$AGE)

#reference: https://www.youtube.com/watch?v=_c3dVTRIK9c
Xfrmlog10 = function(x) {
   return (logb(x+1, 10))
}

#First lets put zeroes where its NA & then transform using signedlog10
insurance.trn$INCOME <- ifelse(is.na( insurance.trn$INCOME)==TRUE, 0, insurance.trn$INCOME)
insurance.trn$INCOME_LOG <- sapply(insurance.trn$INCOME, Xfrmlog10)

summary(insurance.trn$INCOME_LOG)

#Now impute the zeros with median
insurance.trn$INCOME_LOG<- ifelse(insurance.trn$INCOME_LOG ==0, median(insurance.trn$INCOME_LOG), insura

summary(insurance.trn$INCOME_LOG)

ggplot(data = insurance.trn) + geom_histogram(aes(x=INCOME_LOG), binwidth =0.1 ) + ggtitle("INCOME LOG

#reference: https://www.youtube.com/watch?v=_c3dVTRIK9c

#First lets put zeroes where its NA & then transform using signedlog10
insurance.trn$HOME_VAL <- ifelse(is.na( insurance.trn$HOME_VAL)==TRUE, 0, insurance.trn$INCOME)
insurance.trn$HOME_VAL_LOG <- sapply(insurance.trn$HOME_VAL, Xfrmlog10)

summary(insurance.trn$HOME_VAL_LOG)

#Now impute the zeros with median
insurance.trn$HOME_VAL_LOG<- ifelse(insurance.trn$HOME_VAL_LOG ==0, median(insurance.trn$HOME_VAL_LOG),

summary(insurance.trn$HOME_VAL_LOG)

ggplot(data = insurance.trn) + geom_histogram(aes(x=HOME_VAL_LOG), binwidth =0.1 ) + ggtitle("HOME_VAL_

round(prop.table(table(insurance.trn$CAR_USE,insurance.trn$TARGET_FLAG),1),2)
insurance.trn = insurance.trn %>% mutate(CARUSE_COMMERCIAL=as.numeric(CAR_USE=="Commercial"))  %>% sele

table(insurance.trn$EDUCATION)
round(prop.table(table(insurance.trn$EDUCATION,insurance.trn$TARGET_FLAG),1),2)
```

```
insurance.trn = insurance.trn %>% mutate(EDU_BACH_MAST_PHD=as.numeric(EDUCATION %in% c("PhD", "Masters"

round(prop.table(table(insurance.trn$JOB,insurance.trn$TARGET_FLAG),1),2)

insurance.trn = insurance.trn %>% mutate(JOB_WHITECOLLAR=as.numeric(JOB %in% c("Manager", "Lawyer", "Do

#KIDSDRIV is "number of driving children""
round(prop.table(table(insurance.trn$KIDSDRIV,insurance.trn$TARGET_FLAG),1),2)

insurance.trn = insurance.trn %>% mutate(KIDSDRIV_2=as.numeric(KIDSDRIV %in% c(1,2)),
                                         KIDSDRIV_4=as.numeric(KIDSDRIV %in% c(2,4))) %>%
                                  select(-KIDSDRIV)

round(prop.table(table(insurance.trn$MSTATUS,insurance.trn$TARGET_FLAG),1),2)

insurance.trn = insurance.trn %>% mutate(MARRIED=as.numeric(MSTATUS  == "Yes")) %>%
                                  select(-MSTATUS)

round(prop.table(table(insurance.trn$RED_CAR,insurance.trn$TARGET_FLAG),1),2)
round(prop.table(table(insurance.trn$CAR_TYPE,insurance.trn$TARGET_FLAG),1),2)

insurance.trn = insurance.trn %>% mutate(RED_SPORTS_CAR=as.numeric(RED_CAR  == "yes" & CAR_TYPE == 'Spo
                                  select(-RED_CAR)

round(prop.table(table(insurance.trn$RED_SPORTS_CAR,insurance.trn$TARGET_FLAG),1),2)

round(prop.table(table(insurance.trn$REVOKED,insurance.trn$TARGET_FLAG),1),2)

insurance.trn = insurance.trn %>% mutate(REVOKED=as.numeric(REVOKED  == "Yes" ))


round(prop.table(table(insurance.trn$SEX,insurance.trn$TARGET_FLAG),1),2)
insurance.trn = insurance.trn %>% mutate(MALE=as.numeric(SEX  == "M" ))  %>% select(-SEX)

round(prop.table(table(insurance.trn$CAR_TYPE,insurance.trn$TARGET_FLAG),1),2)

insurance.trn = insurance.trn %>% mutate(CAR_TYPE_vAN=as.numeric(CAR_TYPE == "Van"),
                                         CAR_TYPE_SUV=as.numeric(CAR_TYPE == "SUV"),
                                         CAR_TYPE_SPORTS=as.numeric(CAR_TYPE == "Sports Car"),
                                         CAR_TYPE_PICKUP=as.numeric(CAR_TYPE == "Pickup"),
                                         CAR_TYPE_PANTRUCK=as.numeric(CAR_TYPE == "Panel Truck")) %>%
                                  select(-CAR_TYPE)

round(prop.table(table(insurance.trn$PARENT1,insurance.trn$TARGET_FLAG),1),2)
insurance.trn = insurance.trn %>% mutate(PARENT1=as.numeric(PARENT1  == "Yes" ))

round(prop.table(table(insurance.trn$URBANICITY,insurance.trn$TARGET_FLAG),1),2)
insurance.trn = insurance.trn %>% mutate(URBAN=as.numeric(URBANICITY  == "Highly Urban/ Urban")) %>%
                                  select(-URBANICITY)

glimpse(insurance.trn)
summary(insurance.trn)
```

```r
#alias(glm(TARGET_FLAG ~ .- TARGET_AMT, data=insurance.trn, family = "binomial"))
fit <- glm(TARGET_FLAG ~ .- TARGET_AMT, data=insurance.trn, family = "binomial")

#Lets check for Multi-Collinearity - lets find vif value and drop those that has
vifFit1 <- vif(fit)

#sort by descending
vif.df <- as.data.frame(sort(vifFit1, decreasing = T))
names(vif.df) <- c('Multicolinearity score')
kable(vif.df)

insurance.trn$INCOME_LOG <- NULL
fit <- glm(TARGET_FLAG ~ .- TARGET_AMT, data=insurance.trn, family = "binomial")

#Lets check for Multi-Collinearity - lets find vif value and drop those that has
vifFit1 <- vif(fit)

#sort by descending
vif.df <- as.data.frame(sort(vifFit1, decreasing = T))
names(vif.df) <- c('Multicolinearity score')
kable(vif.df)

set.seed(2)
s=sample(1:nrow(insurance.trn),0.75*nrow(insurance.trn))
insurance.training=insurance.trn[s,]
insurance.test=insurance.trn[-s,]

fit.log <- stats::glm(TARGET_FLAG ~ .-TARGET_AMT , family=binomial(), data =insurance.training)
model.backward.step.log = step(fit.log , trace = FALSE)

summary(model.backward.step.log)
formula(model.backward.step.log)

par(mfrow=c(2, 2))
graphics::plot(model.backward.step.log, main="Stepwise Backward Logistic Regression")

#### Stepwise Forward Logistic Regression:

nothing.mod.lgr <- glm(TARGET_FLAG ~ 1,family=binomial, data=na.omit(insurance.training))
model.forward.step.log  = step(nothing.mod.lgr,
    scope=list(lower=formula(nothing.mod.lgr),upper=formula(fit.log)), direction="forward", trace = FALS

summary(model.forward.step.log)
formula(model.forward.step.log)

par(mfrow=c(2, 2))
graphics::plot(model.forward.step.log, main="Stepwise Forward Logistic Regression")

Based on the data descriptions, and the significant predictors noticed above, lets create a manual logis

model.manual.fit.log <- stats::glm(TARGET_FLAG ~ +AGE +CARUSE_COMMERCIAL +CLM_FREQ +HOME_VAL_LOG +JOB_W

summary(model.manual.fit.log)
```

```r
formula(model.manual.fit.log)

par(mfrow=c(2, 2))
graphics::plot(model.manual.fit.log, main="Manuel Logistic Regression")

Build different linear models to predict the amount for TARGET_AMT

fit.mlr <- lm(formula = TARGET_AMT ~. -TARGET_FLAG , data = insurance.training)
model.backward.step.mlr <- step(fit.mlr, trace=FALSE)

summary(model.backward.step.mlr)
formula(model.backward.step.mlr)


par(mfrow=c(2, 2))
graphics::plot(model.backward.step.mlr, main="Stepwise Backward Linear Regression")

nothing.mod.lnr <- lm(TARGET_AMT ~ 1, data=insurance.training)
model.forward.step.mlr <- step(nothing.mod.lnr, scope=list(lower=formula(nothing.mod.lnr),upper=formula

summary(model.forward.step.mlr)
formula(model.forward.step.mlr)

par(mfrow=c(2, 2))
graphics::plot(model.forward.step.mlr, main="Stepwise Forward Linear Regression")
model.manual.fit.mlr <- stats::lm(TARGET_AMT ~ +AGE +CARUSE_COMMERCIAL +CLM_FREQ +HOME_VAL_LOG +JOB_WHI

summary(model.manual.fit.mlr)
formula(model.manual.fit.mlr)

par(mfrow=c(2, 2))
graphics::plot(model.manual.fit.mlr, main="Manuel Linear Regression")

#Storage to keep performance measures that would aid us in selecting the best model:
#performance.results <- c( model = character(),  Sn = numeric(), Sp = numeric, Accuracy = numeric())
performance.results1 <- vector()
insurance.test$score=predict(model.backward.step.log,insurance.test,type="response")
ggplot(insurance.test,aes(x=score,y=TARGET_FLAG,color=factor(TARGET_FLAG))) + geom_point() + geom_jitte
cutoff=0.3
insurance.test$predicted=as.numeric(insurance.test$score>cutoff)
TP=sum(insurance.test$predicted==1 & insurance.test$TARGET_FLAG==1)
FP=sum(insurance.test$predicted==1 & insurance.test$TARGET_FLAG==0)
FN=sum(insurance.test$predicted==0 & insurance.test$TARGET_FLAG==1)
TN=sum(insurance.test$predicted==0 & insurance.test$TARGET_FLAG==0)

# lets also calculate total number of real positives and negatives in the data
P=TP+FN
N=TN+FP
total = P + N
confusionMatrix(factor(insurance.test$predicted), factor(insurance.test$TARGET_FLAG), positive = "1")

sensitivity <- round(sensitivity(factor(insurance.test$predicted),insurance.test$TARGET_FLAG, positive=
```

```
specificity <- round(specificity(factor(insurance.test$predicted),insurance.test$TARGET_FLAG, negative=

#accuracy = (TP+TN)/(P+N)
accuracy <- round( ( (TP + TN) / (P + N) ) , 4)

cnfMtx <- confusionMatrix(insurance.test$predicted, insurance.test$TARGET_FLAG, positive = "1")
roc <- roc(factor(predicted)~as.numeric(TARGET_FLAG),data=insurance.test, plot=FALSE, ci=TRUE)
graphics::plot(roc, legacy.axes = TRUE, col="blue", lwd=3)
auc <- round(auc(factor(predicted)~as.numeric(TARGET_FLAG),insurance.test),4)
performance.results1 <- rbind(performance.results1, c("Step wise Backward",sensitivity, specificity,  a
insurance.test$score=predict(model.forward.step.log,insurance.test,type="response")
ggplot(insurance.test,aes(x=score,y=TARGET_FLAG,color=factor(TARGET_FLAG))) + geom_point() + geom_jitte

cutoff=0.3
insurance.test$predicted=as.numeric(insurance.test$score>cutoff)
TP=sum(insurance.test$predicted==1 & insurance.test$TARGET_FLAG==1)
FP=sum(insurance.test$predicted==1 & insurance.test$TARGET_FLAG==0)
FN=sum(insurance.test$predicted==0 & insurance.test$TARGET_FLAG==1)
TN=sum(insurance.test$predicted==0 & insurance.test$TARGET_FLAG==0)

# lets also calculate total number of real positives and negatives in the data
P=TP+FN
N=TN+FP
total = P + N

confusionMatrix(factor(insurance.test$predicted), factor(insurance.test$TARGET_FLAG), positive = "1")

sensitivity <- round(sensitivity(factor(insurance.test$predicted),insurance.test$TARGET_FLAG, positive=

specificity <- round(specificity(factor(insurance.test$predicted),insurance.test$TARGET_FLAG, negative=

#accuracy = (TP+TN)/(P+N)
accuracy <- round( ( (TP + TN) / (P + N) ) , 4)

cnfMtx <- confusionMatrix(insurance.test$predicted, insurance.test$TARGET_FLAG, positive = "1")
### AUC - Forward

roc <- roc(factor(predicted)~as.numeric(TARGET_FLAG),data=insurance.test, plot=FALSE, ci=TRUE)
graphics::plot(roc, legacy.axes = TRUE, col="blue", lwd=3)
auc <- round(auc(factor(predicted)~as.numeric(TARGET_FLAG),insurance.test),4)
performance.results1 <- rbind(performance.results1, c("Step wise Forward",sensitivity, specificity,  ac

### Compare Results:

results <- as.data.frame(performance.results1);
colnames(results) <- c("Method", "Sn", "Sp", "Accuracy", "AUC")
kable(results)

From the above, both the models result the same in predicting the target variable of the given dataset.

### Mulitple Linear Regression:

Measure performance among Mulitple Linear Regression models and select one
```

```
#### Stepwise Backward Linear Regression Performance:

Performance on the test data:


performance.results2 <- c()

#Adjusted R squared
ar2 <-  summary(model.backward.step.mlr)$adj.r.squared

#Root mean sqaure error:
rmse=sqrt(mean((predict(model.backward.step.mlr,insurance.test)-insurance.test$TARGET_AMT)**2))

#plot(insurance.test$TARGET_AMT,predict(model.backward.step.mlr,insurance.test))
#hist(model.step.mlr$residuals,breaks = 20)

performance.results2 <- rbind(performance.results2, c("Step wise Backward",ar2, rmse))


#Adjusted R squared
ar2 <-  summary(model.forward.step.mlr)$adj.r.squared

#Root mean sqaure error:
rmse=sqrt(mean((predict(model.forward.step.mlr,insurance.test)-insurance.test$TARGET_AMT)**2))

#plot(insurance.test$TARGET_AMT,predict(model.forward.step.mlr,insurance.test))
#hist(model.step.mlr$residuals,breaks = 20)

performance.results2 <- rbind(performance.results2, c("Step wise Forward",ar2, rmse))


###Compare Results:

results <- as.data.frame(performance.results2);
colnames(results) <- c("Method", "Adj R Squared", "RMSE")
kable(results)

From the above, both the models result in poor performance in predicting the target variable of amount i

However, the *Step wise Backward* method performed better in multiple linear regression case. So, we wil


## Evaluation:

 Lets apply the final models on the evaluation dataset. We will first tidy the evaluation dataset and th

insurance.evl$INCOME <- sapply(insurance.evl$INCOME, cleanUpAmounts)
insurance.evl$HOME_VAL <- sapply(insurance.evl$HOME_VAL, cleanUpAmounts)
insurance.evl$BLUEBOOK <- sapply(insurance.evl$BLUEBOOK, cleanUpAmounts)
insurance.evl$OLDCLAIM <- sapply(insurance.evl$OLDCLAIM, cleanUpAmounts)

'z_' will be removed from **MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE**
```

```
insurance.evl$MSTATUS <- sapply(insurance.evl$MSTATUS, cleanUpPrefix)
insurance.evl$SEX <- sapply(insurance.evl$SEX, cleanUpPrefix)
insurance.evl$EDUCATION <- sapply(insurance.evl$EDUCATION, cleanUpPrefix)
insurance.evl$JOB <- sapply(insurance.evl$JOB, cleanUpPrefix)
insurance.evl$CAR_TYPE <- sapply(insurance.evl$CAR_TYPE, cleanUpPrefix)

### Factorize the variables from the evaluation data set:
insurance.evl$TARGET_FLAG <- as.factor(insurance.evl$TARGET_FLAG)
insurance.evl$PARENT1 <- as.factor(insurance.evl$PARENT1)
insurance.evl$SEX <- as.factor(insurance.evl$SEX)
insurance.evl$MSTATUS <- as.factor(insurance.evl$MSTATUS)
insurance.evl$EDUCATION <- as.factor(insurance.evl$EDUCATION)
insurance.evl$JOB <- as.factor(insurance.evl$JOB)
insurance.evl$CAR_USE <- as.factor(insurance.evl$CAR_USE)
insurance.evl$CAR_TYPE <- as.factor(insurance.evl$CAR_TYPE)
insurance.evl$RED_CAR <- as.factor(insurance.evl$RED_CAR)
insurance.evl$REVOKED <- as.factor(insurance.evl$REVOKED)
insurance.evl$URBANICITY <- as.factor(insurance.evl$URBANICITY)

### Data imputing for evaluation data set:
meanVal <- mean(insurance.evl$CAR_AGE, na.rm=TRUE)
medianVal <- median(insurance.evl$CAR_AGE, na.rm=TRUE)

insurance.evl$CAR_AGE<- ifelse( insurance.evl$CAR_AGE <=0 , medianVal, insurance.evl$CAR_AGE)
insurance.evl$CAR_AGE<- ifelse(is.na( insurance.evl$CAR_AGE )==TRUE, medianVal, insurance.evl$CAR_AGE)

medianVal <- median(insurance.evl$YOJ, na.rm=TRUE)
insurance.evl$YOJ<- ifelse(is.na( insurance.evl$YOJ )==TRUE, medianVal, insurance.evl$YOJ)

medianVal <- median(insurance.evl$AGE, na.rm=TRUE)
insurance.evl$AGE<- ifelse(is.na( insurance.evl$AGE )==TRUE, medianVal, insurance.evl$AGE)

insurance.evl$INCOME <- ifelse(is.na( insurance.evl$INCOME)==TRUE, 0, insurance.evl$INCOME)
insurance.evl$INCOME_LOG <- sapply(insurance.evl$INCOME, Xfrmlog10)
insurance.evl$INCOME_LOG<- ifelse(insurance.evl$INCOME_LOG ==0, median(insurance.evl$INCOME_LOG), insura

insurance.evl$HOME_VAL <- ifelse(is.na( insurance.evl$HOME_VAL)==TRUE, 0, insurance.evl$INCOME)
insurance.evl$HOME_VAL_LOG <- sapply(insurance.evl$HOME_VAL, Xfrmlog10)
insurance.evl$HOME_VAL_LOG<- ifelse(insurance.evl$HOME_VAL_LOG ==0, median(insurance.evl$HOME_VAL_LOG),

insurance.evl = insurance.evl %>% mutate(CARUSE_COMMERCIAL=as.numeric(CAR_USE=="Commercial"))  %>% sele

insurance.evl = insurance.evl %>% mutate(EDU_BACH_MAST_PHD=as.numeric(EDUCATION %in% c("PhD", "Masters"

insurance.evl = insurance.evl %>% mutate(JOB_WHITECOLLAR=as.numeric(JOB %in% c("Manager", "Lawyer", "Do


insurance.evl = insurance.evl %>% mutate(KIDSDRIV_2=as.numeric(KIDSDRIV %in% c(1,2)),
                                         KIDSDRIV_4=as.numeric(KIDSDRIV %in% c(2,4))) %>%
                                  select(-KIDSDRIV)

insurance.evl = insurance.evl %>% mutate(MARRIED=as.numeric(MSTATUS  == "Yes")) %>%
                                  select(-MSTATUS)
```

```r
insurance.evl = insurance.evl %>% mutate(RED_SPORTS_CAR=as.numeric(RED_CAR  == "yes" & CAR_TYPE == 'Spo
                                  select(-RED_CAR)

insurance.evl = insurance.evl %>% mutate(REVOKED=as.numeric(REVOKED  == "Yes" ))

insurance.evl = insurance.evl %>% mutate(MALE=as.numeric(SEX  == "M" ))  %>% select(-SEX)

insurance.evl = insurance.evl %>% mutate(CAR_TYPE_vAN=as.numeric(CAR_TYPE == "Van"),
                                         CAR_TYPE_SUV=as.numeric(CAR_TYPE == "SUV"),
                                         CAR_TYPE_SPORTS=as.numeric(CAR_TYPE == "Sports Car"),
                                         CAR_TYPE_PICKUP=as.numeric(CAR_TYPE == "Pickup"),
                                         CAR_TYPE_PANTRUCK=as.numeric(CAR_TYPE == "Panel Truck")) %>%
                                  select(-CAR_TYPE)

insurance.evl = insurance.evl %>% mutate(PARENT1=as.numeric(PARENT1  == "Yes" ))


insurance.evl = insurance.evl %>% mutate(URBAN=as.numeric(URBANICITY  == "Highly Urban/ Urban")) %>%
                                  select(-URBANICITY)

insurance.evl$OLDCLAIM_YRLY_AVG <- insurance.evl$OLDCLAIM / 5
insurance.evl$OLDCLAIM <- NULL


#Final Results:

#Here's our final results from the above models:

#Though it was easier to predict the *target flag* - which indicates 'if someone gets into a crash or n


#Predict Target Flag using Logistic Regression
insurance.target.flag.prd <- predict(model.backward.step.log, newdata=subset(insurance.evl,select=c(HOM
    EDU_BACH_MAST_PHD , JOB_WHITECOLLAR , KIDSDRIV_2 , MARRIED ,
    CAR_TYPE_vAN , CAR_TYPE_SUV , CAR_TYPE_SPORTS , CAR_TYPE_PICKUP ,
    CAR_TYPE_PANTRUCK , URBAN, INCOME)),type='response')

#Since our cut off is 0.3.
insurance.target.flag.prd <- ifelse(insurance.target.flag.prd > 0.3, 1, 0)
insurance.evl$predicted <- insurance.target.flag.prd
insurance.evl$TARGET_FLAG <- factor(insurance.evl$predicted)


#Predict Amount using Multiple Linear Regression
insurance.evl$TARGET_AMT<- round(predict(model.backward.step.mlr, newdata = insurance.evl))

#Final Results.
kable(subset(insurance.evl, select=c(INDEX, TARGET_FLAG,TARGET_AMT, MALE, MARRIED, EDU_BACH_MAST_PHD)))
#, JOB_WHITECOLLAR, KIDSDRIV_2, URBAN, RED_SPORTS_CAR
```