Data 621 Homework 3: Insurance

Tommy Jenkins, Violeta Stoyanova, Todd Weigel, Peter Kowalchuk, Eleanor R-Secoquian, Anthony Pagan

November 6, 2019

OVERVIEW

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

Objective:

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car.

DATA EXPLORATION

Data Summary

The dataset consists of two data files: training and evaluation. The training dataset contains 26 columns, while the evaluation dataset contains 24. The evaluation dataset is missing columns TARGET_FLAG, TARGET_AMT which represend our responce variables, respectively whether the person was in a car crash and the cost of the car crash if the person was in an accident. We will start by exploring the training data set since it will be the one used to generate the models.

The columns in the data set are:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

Missing Data

An important aspect of any dataset is to determine how much, if any, data is missing. We look at all the variables to see which if any have missing data. We look at the basic descriptive statistics as well as the missing data and their percentages.

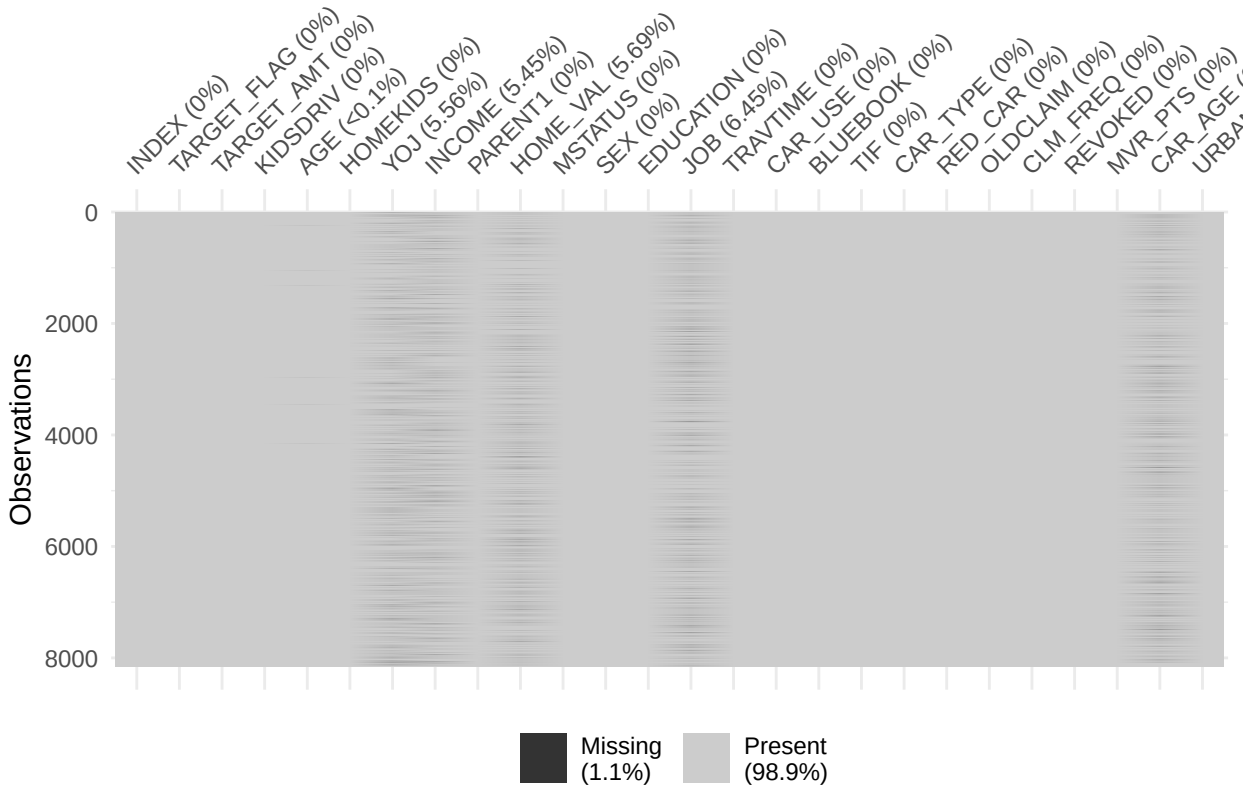
We start by looking at the dataset as awhole and determine how many complete rows, that is rows with data for all predictors, do we have.

```
##    Mode   FALSE    TRUE
## logical    2116    6045
```

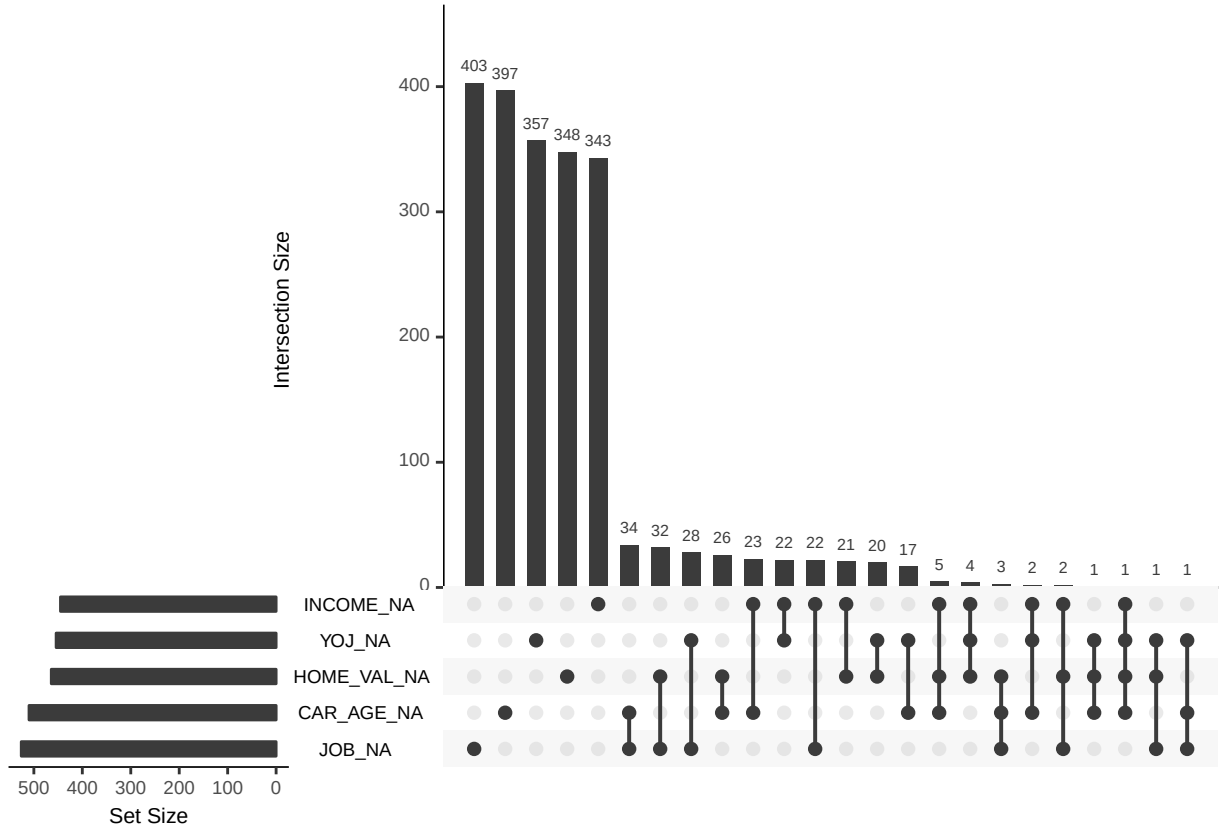
With these results, if we remove all rows with incomplete rows, there will be a total of 6045 rows out of 8161 .If we eliminate all non-complete rows and keep only rows with data for all the predictors in the dataset, our new dataset will results in 74% of the total dataset. We create a subset of data with complete cases only to use later in our analysis.

```
## Observations: 6,045
## Variables: 26
## $ INDEX       <int> 1, 2, 4, 7, 12, 13, 14, 15, 16, 19, 20, 22, 23, 24...
## $ TARGET_FLAG <int> 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 2946.000, 2501.000, 0.000, 60...
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ AGE         <int> 60, 43, 35, 34, 34, 50, 53, 43, 55, 45, 39, 42, 34...
## $ HOMEKIDS    <int> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1, 0, 0,...
## $ YOJ         <int> 11, 11, 10, 12, 10, 7, 14, 5, 11, 0, 12, 11, 13, 1...
## $ INCOME      <fct> "$67,349", "$91,449", "$16,039", "$125,301", "$62,...
## $ PARENT1     <fct> No, No, No, Yes, No, No, No, No, No, No, Yes, No, ...
## $ HOME_VAL    <fct> "$0", "$257,252", "$124,191", "$0", "$0", "$0", "$...
## $ MSTATUS     <fct> z_No, z_No, Yes, z_No, z_No, z_No, z_No, Yes, Yes,...
## $ SEX         <fct> M, M, z_F, z_F, z_F, M, z_F, z_F, M, z_F, z_F, M, ...
## $ EDUCATION   <fct> PhD, z_High School, z_High School, Bachelors, Bach...
## $ JOB         <fct> Professional, z_Blue Collar, Clerical, z_Blue Coll...
## $ TRAVTIME    <int> 14, 22, 5, 46, 34, 48, 15, 36, 25, 48, 43, 42, 27,...
## $ CAR_USE     <fct> Private, Commercial, Private, Commercial, Private,...
## $ BLUEBOOK    <fct> "$14,230", "$14,940", "$4,010", "$17,430", "$11,20...
## $ TIF         <int> 11, 1, 4, 1, 1, 7, 1, 7, 7, 1, 6, 6, 7, 4, 6, 6, 1...
## $ CAR_TYPE    <fct> Minivan, Minivan, z_SUV, Sports Car, z_SUV, Van, S...
## $ RED_CAR     <fct> yes, yes, no, no, no, no, no, no, yes, no, no, no,...
## $ OLDCLAIM    <fct> "$4,461", "$0", "$38,690", "$0", "$0", "$0", "$0",...
## $ CLM_FREQ    <int> 2, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 1, 0,...
## $ REVOKED     <fct> No, No, No, No, No, No, No, No, Yes, No, No, No, N...
## $ MVR_PTS     <int> 3, 0, 3, 0, 0, 1, 0, 0, 3, 3, 0, 0, 0, 0, 0, 5, 1,...
## $ CAR_AGE     <int> 18, 1, 10, 7, 1, 17, 11, 1, 9, 5, 13, 16, 20, 7, 1...
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly U...
```

But we can also look at what specific predictors are missing in our dataset. If we do this we can see how there is much more data available, as we find only 5 predictors with missing data. Data missing for these predictors also only accounts for less than 7% of the respective predictors total.



We look closer at the missing data and look at the intersection of predictors with missing data. We find that the bulk of the missing data is for predictors with no intersection with other missing predictor data.



Having this detail in missing data might be of importance when looking at models. In the next Data Preparation section we will handle these missing cases and build a data set with data for all predictors in all rows.

Data Exploration

Using TARGET_FLAG as response variables we confirm when TARGET_FLAG is 1 TARGET_AMOUNT >0 and when TARGET_FLAG is 0 when TARGET_AMOUNT = 0

```
nrow(subset(InsTrain,TARGET_FLAG == 0))
```

```
## [1] 6008
```

```
nrow(subset(InsTrain,TARGET_AMT == 0))
```

```
## [1] 6008
```

```
nrow(subset(InsTrain,TARGET_FLAG > 0))
```

```
## [1] 2153
```

```
nrow(subset(InsTrain,TARGET_AMT > 0))
```

```
## [1] 2153
```

A glimpse of the data shows that the following columns should be integers and not Factors:

- INCOME
- HOME_VAL
- BLUEBOOK
- OLDCLAIM
- JOB

We display and view data with all cases and only complete cases

```
## 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      <fct> "$67,349", "$91,449", "$16,039", NA, "$114,986", "...
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, N...
## $ HOME_VAL    <fct> "$0", "$257,252", "$124,191", "$306,251", "$243,92...
## $ MSTATUS     <fct> z_No, z_No, Yes, Yes, Yes, z_No, Yes, Yes, z_No, z...
## $ SEX         <fct> M, M, z_F, M, z_F, z_F, z_F, M, z_F, M, z_F, z_F, ...
## $ EDUCATION   <fct> PhD, z_High School, z_High School, <High School, P...
## $ JOB         <fct> Professional, z_Blue Collar, Clerical, z_Blue Coll...
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25,...
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Co...
## $ BLUEBOOK    <fct> "$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    <fct> Minivan, Minivan, z_SUV, Minivan, z_SUV, Sports Ca...
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no...
## $ OLDCLAIM    <fct> "$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     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, ...
## $ 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  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly U...
```

We use Sapply function to review which columns have NA Values. It display columns and percent of values that are missing.

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##         0.0         0.0         0.0         0.0         0.1         0.0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##         5.6         5.5         0.0         5.7         0.0         0.0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##         0.0         6.4         0.0         0.0         0.0         0.0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##         0.0         0.0         0.0         0.0         0.0         0.0
##     CAR_AGE  URBANICITY
##         6.2         0.0
```

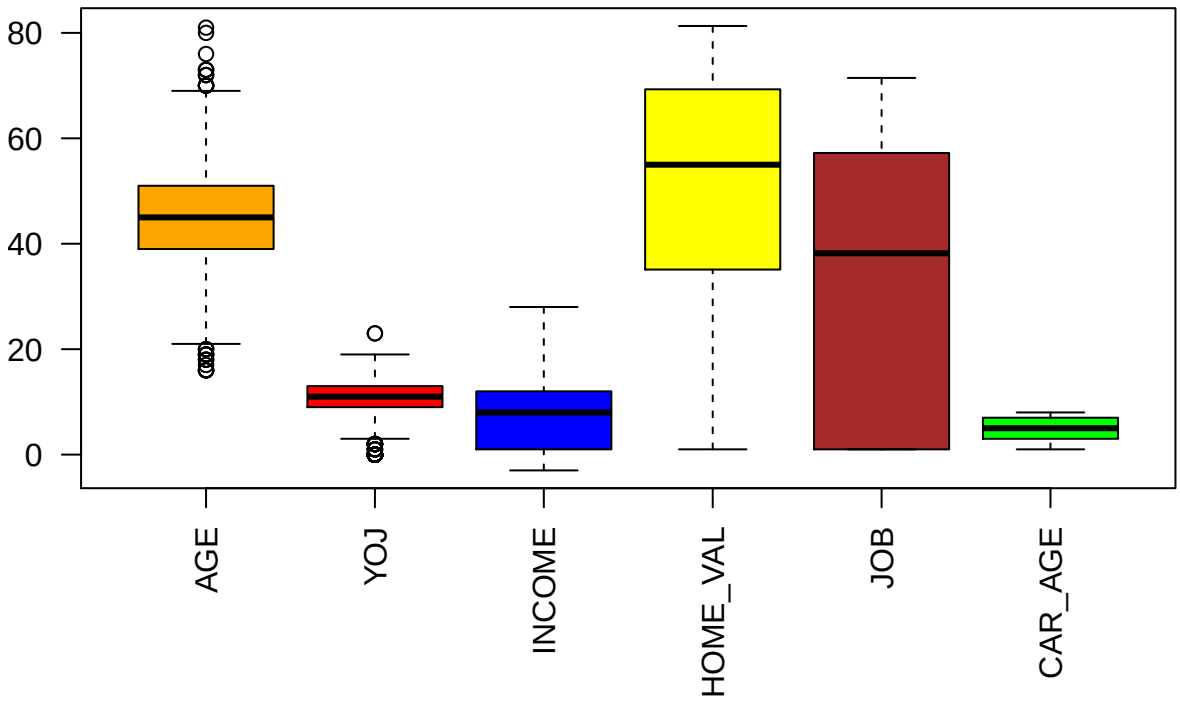
Data Preperation

As revealed earlier there were a list of columns that we factors that should be integers. We start by converting the columns to numeric.

```
## Observations: 8,161
## Variables: 5
## $ INCOME   <fct> "$67,349", "$91,449", "$16,039", NA, "$114,986", "$12...
## $ HOME_VAL <fct> "$0", "$257,252", "$124,191", "$306,251", "$243,925",...
## $ JOB      <fct> Professional, z_Blue Collar, Clerical, z_Blue Collar,...
## $ BLUEBOOK <fct> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000",...
## $ OLDCLAIM <fct> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0...
```

```
## Observations: 8,161
## Variables: 5
## $ INCOME   <dbl> 5032, 6291, 1249, NA, 508, 745, 1487, 314, 4764, 281,...
## $ HOME_VAL <dbl> 1, 3258, 347, 3916, 3033, 1, NA, 4166, 1, 1, 1, 2261,...
## $ JOB      <dbl> 6, 8, 1, 8, 2, 8, 8, 8, 1, 6, 4, 6, 5, NA, 3, 1, 4, 4...
## $ BLUEBOOK <dbl> 434, 503, 2212, 553, 802, 746, 2672, 701, 135, 852, 8...
## $ OLDCLAIM <dbl> 1449, 1, 1311, 1, 432, 1, 1, 510, 1, 1, 1, 1, 1727, 1...
```

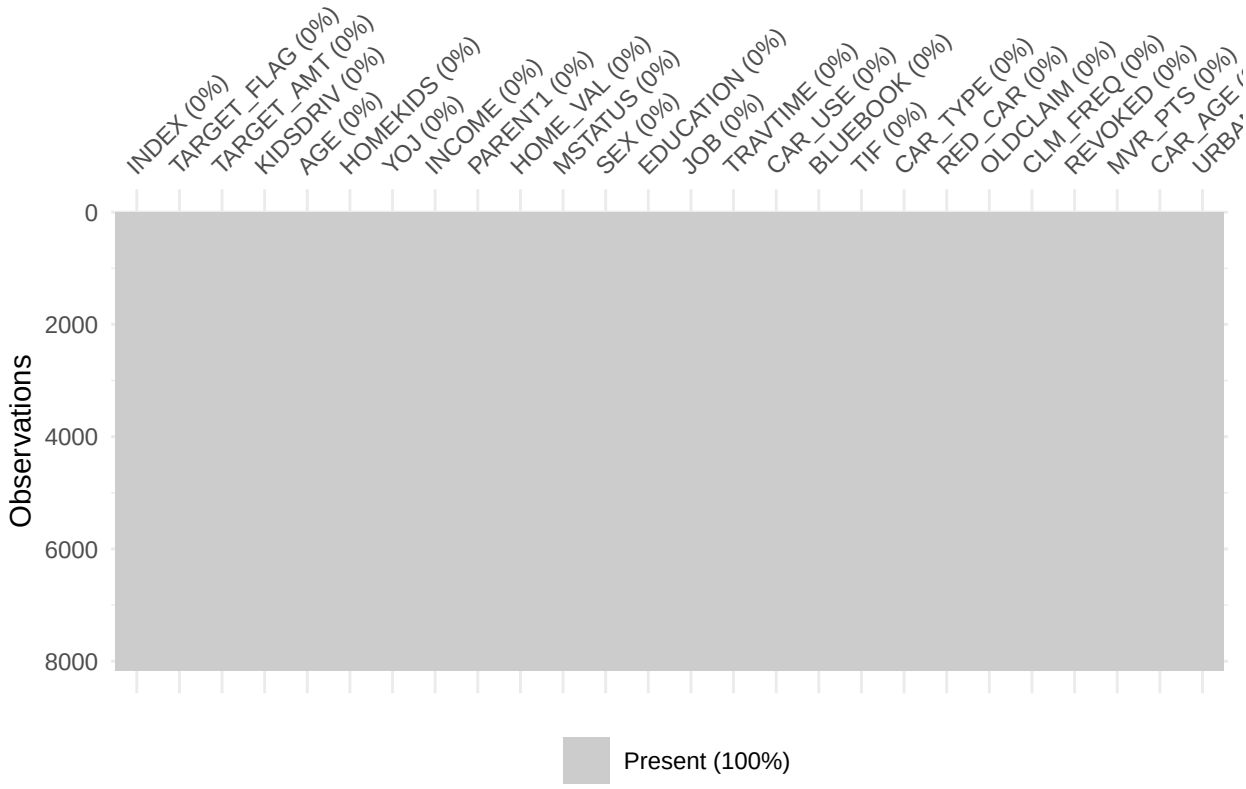
Both boxplot and summary stats with the square root transform of Home_val and Income to confirm we can use median or mean values to replace NA values if we chose.




```
##           vars    n    mean      sd median trimmed     mad min   max range
## AGE          1 8155   44.79    8.63   45.0   44.83    8.90  16    81    65
## YOJ          2 7707   10.50    4.09   11.0   11.07    2.97   0    23    23
## INCOME       3 7716 3040.33 2029.52 3024.5 3018.56 2647.18   1 6612  6611
## HOME_VAL     4 7697 1785.40 1695.15 1459.0 1639.68 2161.63   1 5106  5105
## JOB          5 7635    5.01    2.46    5.0    5.14    2.97   1    8     7
## CAR_AGE      6 7651    8.33    5.70    8.0    7.96    7.41  -3   28    31
##             skew kurtosis    se
## AGE        -0.03    -0.06  0.10
## YOJ        -1.20     1.18  0.05
## INCOME      0.04    -1.24 23.10
## HOME_VAL    0.43    -1.24 19.32
## JOB        -0.34    -1.16  0.03
## CAR_AGE     0.28    -0.75  0.07
```

We next replace all NA values with mean values for cases that are missing values and rerun sapply function to confirm there are no longer any missing values.

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           0           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##           0           0           0           0           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##           0           0
```




```
##          vars    n    mean      sd  median trimmed     mad min  max range
## AGE         1 8161   44.79    8.62   45.00   44.83    8.90  16   81    65
## YOJ         2 8161   10.50    3.98   11.00   11.05    2.97   0   23    23
## INCOME      3 8161 3040.33 1973.41 3040.33 3020.15 2492.74   1 6612  6611
## HOME_VAL    4 8161 1785.40 1646.25 1680.00 1642.31 2489.29   1 5106  5105
## JOB         5 8161    5.01    2.38    5.01    5.14    2.95   1    8     7
## CAR_AGE     6 8161    8.33    5.52    8.33    7.98    5.44  -3   28    31
##            skew kurtosis    se
## AGE       -0.03    -0.06  0.10
## YOJ       -1.24     1.42  0.04
## INCOME     0.05    -1.14 21.84
## HOME_VAL   0.44    -1.14 18.22
## JOB       -0.36    -1.03  0.03
## CAR_AGE    0.29    -0.60  0.06
```

Build Model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - INDEX - TARGET_FLAG - TARGET_AMT,
##     family = binomial(link = "logit"), data = InsTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5922  -0.7276  -0.4209   0.6419   3.1555
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -4.709e-01  2.687e-01  -1.753 0.079675 .
```

```
## KIDSDRIV                           3.641e-01  6.053e-02   6.014 1.81e-09 ***
## AGE                               -4.635e-03  3.910e-03  -1.185 0.235876
## HOMEKIDS                           7.217e-02  3.651e-02   1.977 0.048085 *
## YOJ                               -2.674e-02  7.937e-03  -3.369 0.000755 ***
## INCOME                            -2.314e-05  1.662e-05  -1.392 0.163862
## PARENT1Yes                         3.497e-01  1.082e-01   3.232 0.001229 **
## HOME_VAL                          -9.899e-05  2.047e-05  -4.836 1.33e-06 ***
## MSTATUSz_No                        4.804e-01  7.717e-02   6.226 4.80e-10 ***
## SEXz_F                            -2.579e-01  1.021e-01  -2.526 0.011524 *
## EDUCATIONBachelors                -6.887e-01  1.043e-01  -6.603 4.03e-11 ***
## EDUCATIONMasters                  -7.923e-01  1.291e-01  -6.136 8.47e-10 ***
## EDUCATIONPhD                      -1.026e+00  1.545e-01  -6.644 3.06e-11 ***
## EDUCATIONz_High School            -1.118e-01  9.212e-02  -1.214 0.224900
## JOB                               -2.270e-02  1.406e-02  -1.615 0.106418
## TRAVTIME                           1.517e-02  1.868e-03   8.121 4.62e-16 ***
## CAR_USEPrivate                    -8.701e-01  8.207e-02 -10.602  < 2e-16 ***
## BLUEBOOK                           2.106e-05  3.344e-05   0.630 0.528975
## TIF                               -5.453e-02  7.276e-03  -7.495 6.63e-14 ***
## CAR_TYPEPanel Truck                6.109e-02  1.364e-01   0.448 0.654173
## CAR_TYPEPickup                     5.281e-01  1.006e-01   5.249 1.53e-07 ***
## CAR_TYPESports Car                 1.209e+00  1.217e-01   9.934  < 2e-16 ***
## CAR_TYPEVan                        3.726e-01  1.195e-01   3.119 0.001815 **
## CAR_TYPEz_SUV                      9.618e-01  1.024e-01   9.392  < 2e-16 ***
## RED_CARyes                         1.440e-03  8.558e-02   0.017 0.986572
## OLDCLAIM                           8.744e-05  4.214e-05   2.075 0.037989 *
## CLM_FREQ                           1.161e-01  3.190e-02   3.639 0.000274 ***
## REVOKEDYes                         7.483e-01  7.965e-02   9.395  < 2e-16 ***
## MVR_PTS                            1.101e-01  1.360e-02   8.092 5.86e-16 ***
## CAR_AGE                           -6.893e-04  7.459e-03  -0.092 0.926369
## URBANICITYz_Highly Rural/ Rural -2.281e+00  1.122e-01 -20.331  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7422.7  on 8130  degrees of freedom
## AIC: 7484.7
##
## Number of Fisher Scoring iterations: 5
```

Select Model

Compare Model Statistics

Conclusion

APPENDIX