Data 621: Homework 4

Car Insurance Data

Aaron Grzasko

April 14, 2018

Introduction

In this report, we analyze insurance data to estimate the following quantities:

- The probability that a specified driver will have a car crash.
- The dollar cost of auto claims, given that the insured was involved in a crash.

In practice, these claim frequency and severity measures are useful for determining appropriate pure premium amounts to charge auto policyholders.

Using the provide training data set, we will build two separate models:

- a binary logistic regression to determine crash probabilities.
- a multiple linear regression model to estimate claim severity.

We will then make predictions on the provided insurance testing data set.

Data Exploration

Variable Overview

The training data set includes 8,161 observations, with 26 variables: 23 predictors, two response variables, and one record identifier.

Below is a brief description of the included variables:

Variable Name	Description	Theoretical Impact
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	In a crash? 1=YES 0=NO	None
TARGET_AMT	Cost of Crash, if applicable	None
KIDSDRIV	# Driving Children	When teenagers drive your car, increased crash risk
AGE	Age of Driver	Young and old drivers might be riskier
HOMEKIDS	# Children at Home	Unknown effect
YOJ	Years on Job	Long-term employees are usually safer
INCOME	Income	In theory, rich have fewer crashes
PARENT1	Single Parent	Unknown impact
HOME_VAL	Home Value	In theory, home owners may drive more responsibly
MSTATUS	Marital Status	In theory, married individuals are less risky
SEX	Gender	Urban legend: females are safer drivers
EDUCATION	Max Education Level	Unknown, but in theory educated people drive more safely
JOB	Job Category	In theory, white collar workers are less risky
TRAVTIME	Commute Distance	Long drives to work usually suggest greater risk
CAR_USE	Vehicle Use	Commercial fleet driven more, may impact collision prob
BLUEBOOK	Value of Vehicle	Unknown impact on collision prob, but impacts crash payout
TIF	Time in Force	Long-term customers are usually safer

Variable Name	Description	Theoretical Impact
CAR_TYPE	Type of Car	Unknown impact on collision prob, but impacts crash payout
RED_CAR	A Red Car	Urban legend: red cars are riskier, particularly sports cars
OLDCLAIM	# Claims (Past 5 Years)	If total payout high, future payouts might be high
CLM_FREQ	Total Claims (Past 5 Years)	Claim count should be positively correlated with future claims
REVOKED	License Revoked (Past 7 Years)	If your license was revoked, you probably are a riskier driver
MVR_PTS	Motor Vehicle Report Points	Traffic ticket counts have postive correlation with crashes
CAR_AGE	Vehicle Age	Unknown impact on collision prob, but impacts crash payout
URBANICITY	Home/Work Area	Unknown impact

There are a couple issues with the raw data file:

- Currency fields were treated as factors due to "$" and "," characters.
- Multiple character field entries included an extraneous "z_" or "<" prefix.

We also rescaled the fields `INCOME`, `HOME_VAL`, `BLUEBOOK`, and `OLDCLAIM` to be so that dollars are expressed in $1,000s,

After cleaning up these minor issues, we're ready to explore data types and sample observations from the training set:

```
'data.frame':   8161 obs. of  26 variables:
 $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
 $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
 $ TARGET_AMT : num  0 0 0 0 0 ...
 $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
 $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
 $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
 $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
 $ INCOME     : num  67.3 91.4 16 NA 115 ...
 $ PARENT1    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ HOME_VAL   : num  0 257 124 306 244 ...
 $ MSTATUS    : Factor w/ 2 levels "No","Yes": 1 1 2 2 2 1 2 2 1 1 ...
 $ SEX        : Factor w/ 2 levels "F","M": 2 2 1 2 1 1 1 2 1 2 ...
 $ EDUCATION  : Ord.factor w/ 4 levels "High School"<..: 4 1 1 1 4 2 1 2 2 2 ...
 $ JOB        : Factor w/ 9 levels "","Blue Collar",..: 8 2 3 2 4 2 2 2 3 8 ...
 $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
 $ CAR_USE    : Factor w/ 2 levels "Commercial","Private": 2 1 2 2 2 1 2 1 2 1 ...
 $ BLUEBOOK   : num  14.23 14.94 4.01 15.44 18 ...
 $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
 $ CAR_TYPE   : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 5 1 5 4 5 6 5 6 ...
 $ RED_CAR    : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 2 1 1 ...
 $ OLDCLAIM   : num  4.46 0 38.69 0 19.22 ...
 $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
 $ REVOKED    : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...
 $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
 $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
 $ URBANICITY : Factor w/ 2 levels "Highly Rural/ Rural",..: 2 2 2 2 2 2 2 2 2 1 ...
```

We ignore the field `INDEX` for modeling purposes, which is used only for identifying observations.

The two response variables, `TARGET_FLAG`and `TARGET_AMT`, contain binary and dollar values, respectively.

The predictors include four discrete count variables:

- KIDSDRIV
- HOMEKIDS
- CLM_FREQ
- MVR_PTS

There are five discrete time measurements:

- AGE
- YOJ
- TRAVTIME
- TIF
- CAR_AGE

There are also seven binary, categorical features:

- PARENT1
- MSTATUS
- SEX
- CAR_USE
- RED_CAR
- REVOKED
- URBANCITY

The data set includes two multinomial, categorical features:

- JOB
- CAR_TYPE

There is one ordinal, categorical predictor:

- EDUCTATION

Finally, there are four predictors that express dollar amounts. These variables are effectively continuous:

- INCOME
- HOME_VAL
- BLUEBOOK
- OLDCLAIM

Now, let's review a high level statistical summary of the variables:

```
     INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV           AGE
 Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000   Min.   :16.00
 1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
 Median : 5133   Median :0.0000   Median :     0   Median :0.0000   Median :45.00
 Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711   Mean   :44.79
 3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
```

```
 Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000   Max.   :81.00
                                                                    NA's   :6
    HOMEKIDS           YOJ            INCOME       PARENT1       HOME_VAL      MSTATUS
 Min.   :0.0000   Min.   : 0.0   Min.   :  0.00   No :7084   Min.   :  0.0   No :3267
 1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28.10   Yes:1077   1st Qu.:  0.0   Yes:4894
 Median :0.0000   Median :11.0   Median : 54.03              Median :161.2
 Mean   :0.7212   Mean   :10.5   Mean   : 61.90              Mean   :154.9
 3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85.99              3rd Qu.:238.7
 Max.   :5.0000   Max.   :23.0   Max.   :367.03              Max.   :885.3
                  NA's   :454    NA's   :445                 NA's   :464
 SEX            EDUCATION             JOB            TRAVTIME            CAR_USE
 F:4375   High School:3533   Blue Collar :1825   Min.   :  5.00   Commercial:3029
 M:3786   Bachelors  :2242   Clerical    :1271   1st Qu.: 22.00   Private   :5132
          Masters    :1658   Professional:1117   Median : 33.00
          PhD        : 728   Manager     : 988   Mean   : 33.49
                             Lawyer      : 835   3rd Qu.: 44.00
                             Student     : 712   Max.   :142.00
                             (Other)     :1413
    BLUEBOOK          TIF                CAR_TYPE    RED_CAR       OLDCLAIM
 Min.   : 1.50   Min.   : 1.000   Minivan    :2145   no :5783   Min.   : 0.000
 1st Qu.: 9.28   1st Qu.: 1.000   Panel Truck: 676   yes:2378   1st Qu.: 0.000
 Median :14.44   Median : 4.000   Pickup     :1389              Median : 0.000
 Mean   :15.71   Mean   : 5.351   Sports Car : 907              Mean   : 4.037
 3rd Qu.:20.85   3rd Qu.: 7.000   SUV        :2294              3rd Qu.: 4.636
 Max.   :69.74   Max.   :25.000   Van        : 750              Max.   :57.037

    CLM_FREQ       REVOKED       MVR_PTS          CAR_AGE                     URBANICITY
 Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000   Highly Rural/ Rural:1669
 1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000   Highly Urban/ Urban:6492
 Median :0.0000              Median : 1.000   Median : 8.000
 Mean   :0.7986              Mean   : 1.696   Mean   : 8.328
 3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
 Max.   :5.0000              Max.   :13.000   Max.   :28.000
                                              NA's   :510
```

We notice a variety of missing values and strange field entries that may reflect data errors. These issues will be address in a later section.

Let's now focus on each variable individually.

TARGET Variables

TARGET_FLAG

The response variable `TARGET_FLAG` has a moderate imbalance, with three-quarters of the observations indicating no crashes.

	0	1	Sum
count	6008.0	2153.0	8161
percent	73.6	26.4	100

TARGET_AMT

The other response, `TARGET_AMT`, exhibits extreme, positive skewness and high kurtosis.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	StdD	Skew	Kurt
0.00	0.00	0.00	1504.32	1036.00	107586.14	4704.03	8.71	115.32

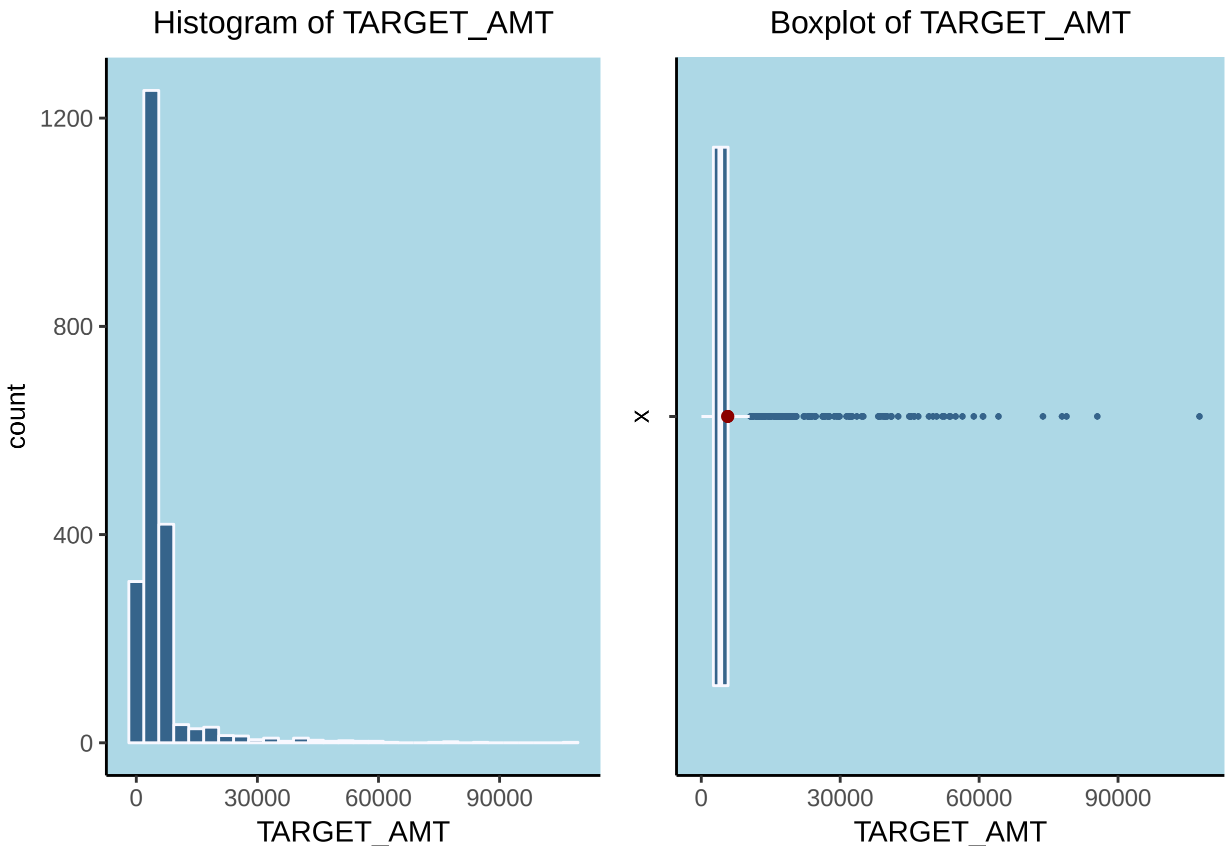
We already noted that almost three quarters of records indicate no car crash. In these cases, the `TARGET_AMT` has a zero value.

For modeling purposes, however, we will only be interested in dollar amounts where a crash occurred. Going forward, when performing calculations and summaries involving `TARGET_AMT`, we will use a subset of the data where zero amounts are filtered out.

Here is a summary of the zero-truncated `TARGET_AMT` variable:

```
  Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      StdD      Skew      Kurt
 30.28   2609.78   4104.00   5702.18   5787.00 107586.14   7743.18      5.64     45.49
```

Even after we remove the zero values, the variable remains highly skewed:



Count Variables

KIDSDRIV

The discrete variable `KIDSDRIV` is right skewed, with 88% of insureds in the training data having no teenage drivers in the household.

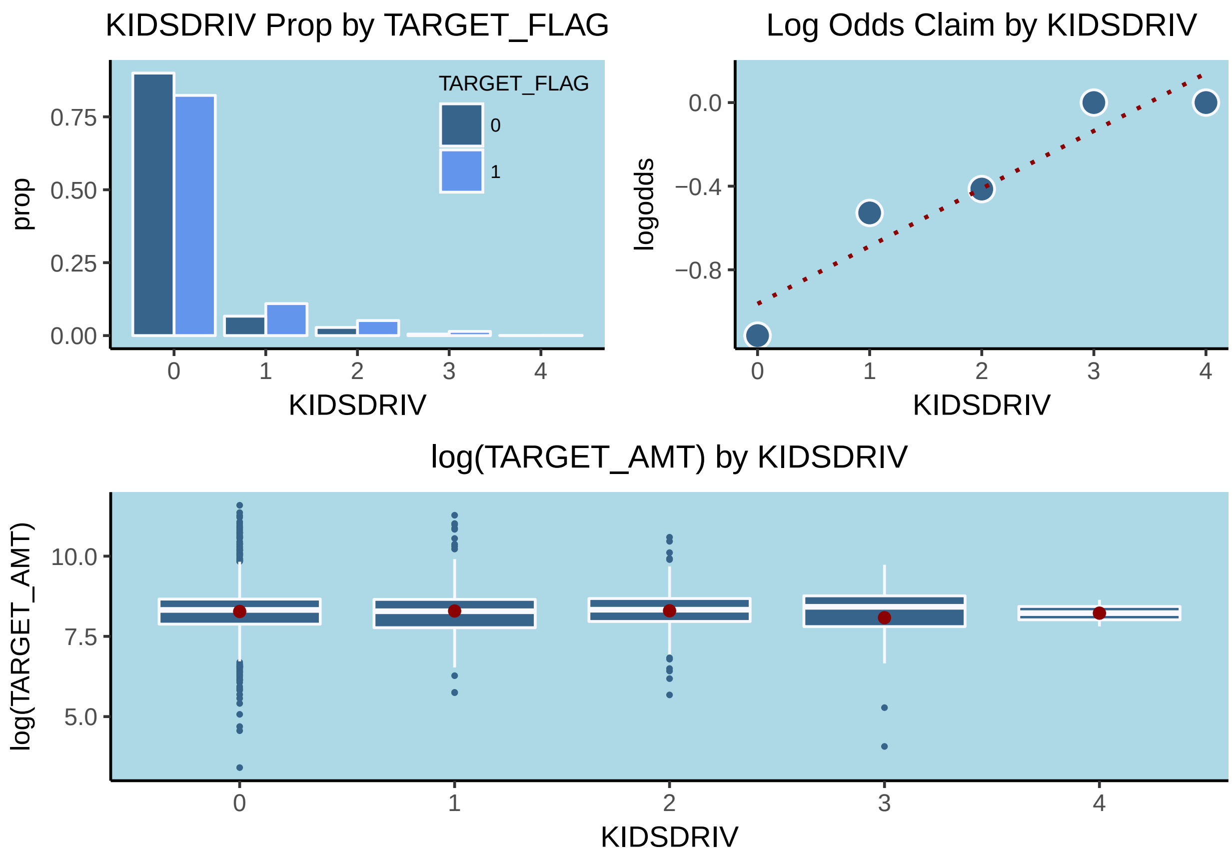
```
           0     1     2    3 4  Sum
count   7180 636.0 279.0 62.0 4 8161
percent   88   7.8   3.4  0.8 0  100
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
  0.00    0.00    0.00    0.17    0.00    4.00    0.51    3.35   14.78
```

In the barplot below, we see different a slightly different distribution of policyholders involved in crashes vis-a-vis the incident-free insureds. Specifically, we see slightly higher concentration of individuals teenage drivers. The scatter plot also indicates a relationship between relationship between number of teenage drivers and the log odds of an auto incident.

We also plot the the log TARGET_AMT–given that a crash occurred–against `KIDSDRIV`. Note: we applied the log transformation because `TARGET_AMT` is highly right skewed, and the directional relationship with `KIDSDRIV` should be clearer in this form.

While not entirely clear in the boxplot, there appears to be a rough, positive relationship between `KIDSDRIV` and the median cost of the crash. The relationship with the mean crash amount, however, is less clear.



KIDSDRIV	ct	mean_cost	median_cost
0	1773	5659	4123
1	236	6220	3958
2	111	5542	4148
3	31	4915	4541
4	2	4054	4054

HOMEKIDS

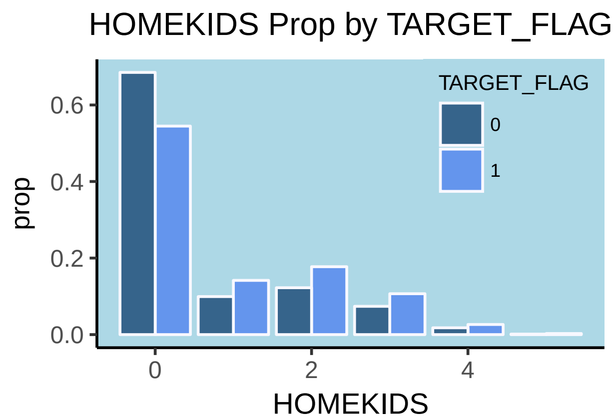
Numerical distribution and statistical summaries are presented below:

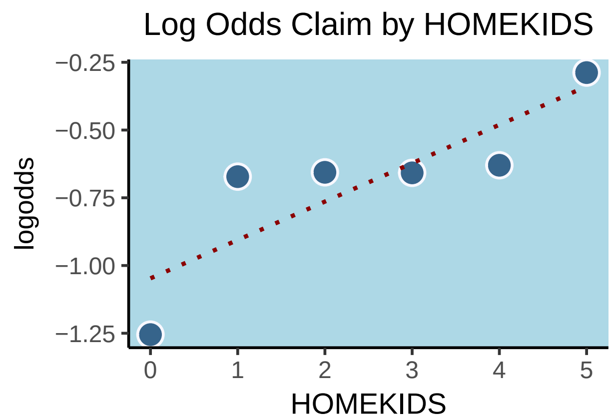
```
              0     1      2     3   4    5  Sum
count    5289.0 902.0 1118.0 674.0 164 14.0 8161
percent    64.8  11.1   13.7   8.3   2  0.2  100
```

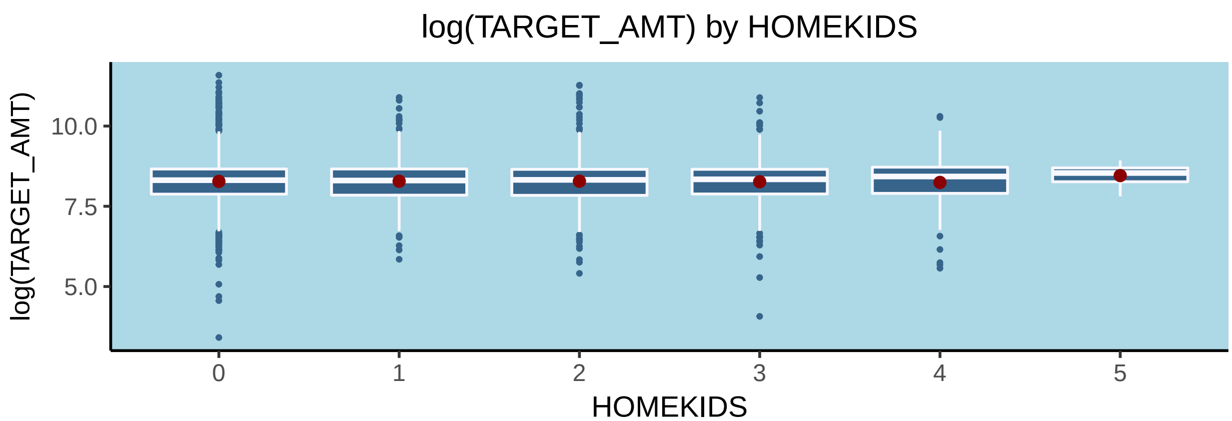
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
   0.00    0.00    0.00    0.72    1.00    5.00    1.12    1.34    3.65
```

The distribution of this discrete variable is right skewed, but not to the same extent as `KIDSRIV`. `HOMEKIDS` contains some of the same information as `KIDSDRIV`: presumably, some of the reported children are also drivers.

Let's look plots relating this predictor to the target variables:







The barplot provides some evidence that policyholders with crashes tend to have more children than than insureds not involved in an auto incident. The scatterplot indicates a significant difference in log odds between policyholders with children vs. insureds without children. We discount the observation associated with five children due to the small sample size.

There appears to be a subtle relationship between `HOMEKIDS` and the median cost of crashes. The relationship with the mean is less clear, given the highly variable and skewed distribution of crash amounts.

HOMEKIDS	ct	mean_cost	median_cost
0	1173	5685	4080
1	305	5522	4036
2	382	6085	4122
3	230	5432	4192
4	57	5610	4575
5	6	5009	5130

CLM_FREQ

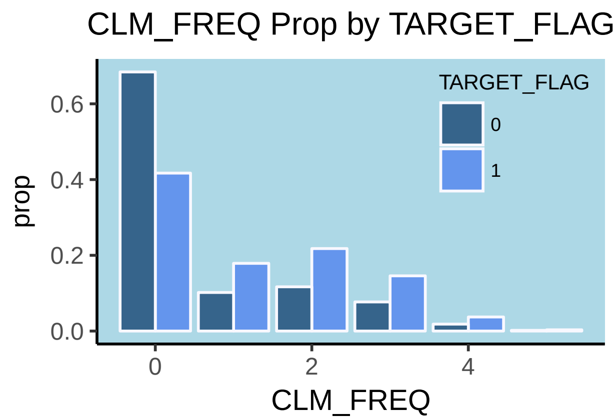
Let's review numerical and statistical summaries for `CLM_FREQ`, another discrete count variable:

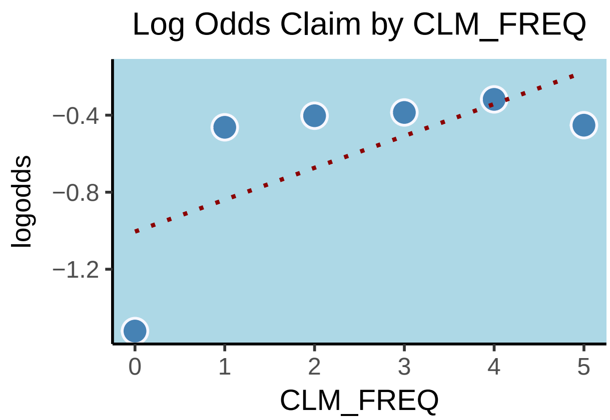
```
          0     1      2     3     4    5   Sum
```

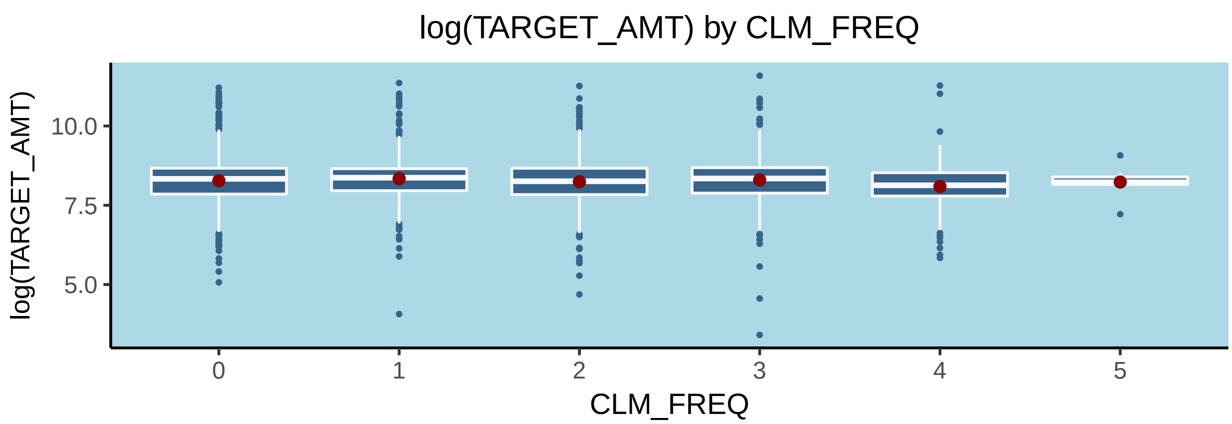
```
count    5009.0 997.0 1171.0 776.0 190.0 18.0 8161
percent    61.4  12.2   14.3   9.5   2.3  0.2  100

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
   0.00    0.00    0.00    0.80    2.00    5.00    1.16    1.21    3.29
```

This variable has a similar skew as the previously reviewed count variable, HOMEKIDS.

Based on the barplot below, there seems to be a significance difference in the distribution of prior claim frequencies for TARGET_FLAG values of zero vs. one. We see roughly 60% of auto claimants have had one or more prior claims in the past five year, while only 40% of non-claimants have had accidents. The scatter plot indicates a nonlinear albeit positive relationship between log odds of a claim and prior claim history. We also do not see a clear pattern between CLM_FREQ and the claim amounts.







CLM_FREQ	ct	mean_cost	median_cost
0	898	5633	4160
1	385	5995	4312
2	469	5463	3858
3	314	5970	4206
4	80	5551	3393
5	7	4247	3737

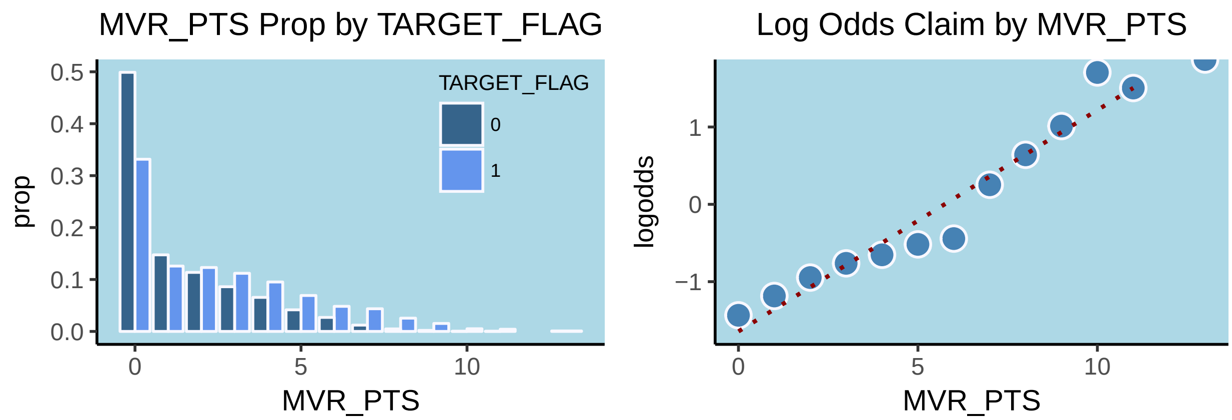
MVR_PTS

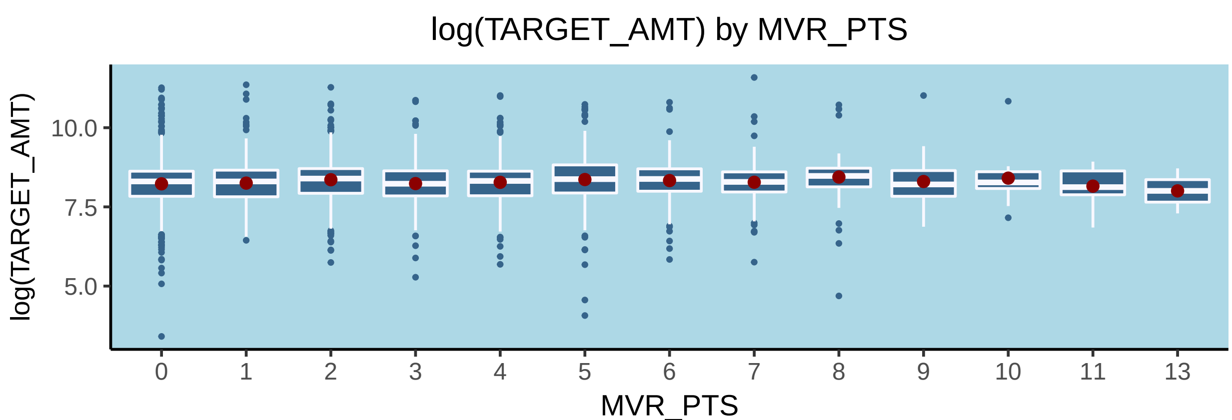
Like the other count variables, MVR_PTS is positively skewed. There seems to be a positive relationship between points and log odds; however the scatter plot below indicates a strange, curved relationship between the two variables. We wonder if the curvature is related to to interaction with another predictor.

We see no straightforward relationship between `TARGET_AMT` and `MVR_PTS`.

	0	1	2	3	4	5	6	7	8	9	10	11	13	Sum
count	3712.0	1157.0	948.0	758.0	599.0	399.0	266.0	167	84	45.0	13.0	11.0	2	8161
percent	45.5	14.2	11.6	9.3	7.3	4.9	3.3	2	1	0.6	0.2	0.1	0	100

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	StdD	Skew	Kurt
0.00	0.00	1.00	1.70	3.00	13.00	2.15	1.35	4.38





BIN_MVR	ct	mean_cost	median_cost
0	985	5379	4037
2	506	5762	4068
4	354	6110	4292
6	198	5896	4194
8	88	6809	4492
10	20	6293	3824
12	2	3786	3786

Time Variables

AGE

Below is table of values of the `AGE` predictor, with ages bucketed into 5 year increments (i.e. [15,20), [20,25), [30,35), etc.])

	15	20	25	30	35	40	45	50	55	60	65	70	75	80	Sum
count	14.0	58.0	249.0	649	1259.0	1673.0	1810.0	1373.0	725.0	265.0	65.0	12.0	1	2	8155

```
percent  0.2  0.7   3.1   8   15.4   20.5   22.2   16.8   8.9   3.2  0.8  0.1  0  0  100
```

Here is a statistical summary:

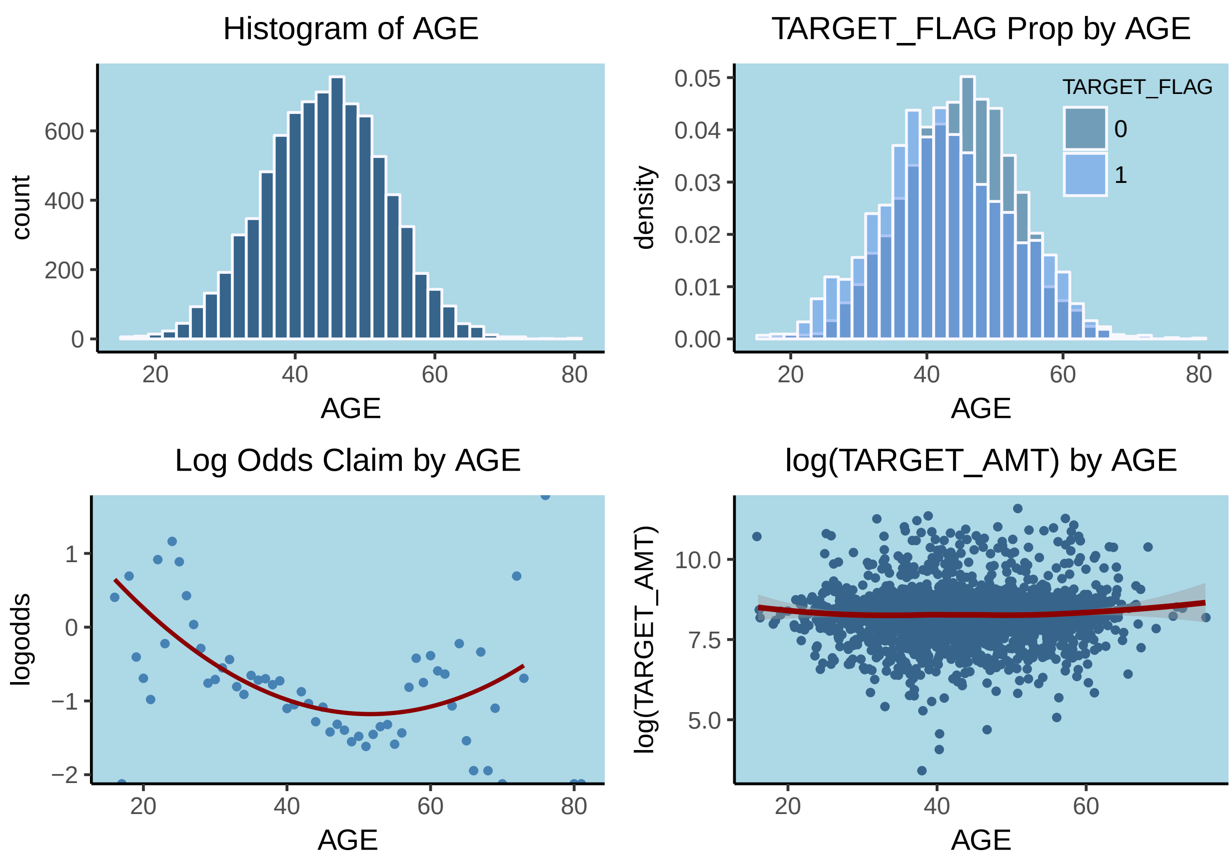
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's    StdD    Skew    Kurt
  16.00   39.00   45.00   44.79   51.00   81.00    6.00    8.63   -0.03    2.94
```

We note six missing values that we'll need to address later.

The distribution of `AGE` is almost perfectly normal. When we break out the data by `TARGET_FLAG` values, the distributions of age by `TARGET_FLAG` are still roughly normal. However, individuals involved in a crash appear to be slightly younger, on average.

In the bottom left scatter plot, we notice a pattern between log odds and age. Specifically, there appears to be a curved relationship where younger ages have a higher odds of a crash. The odds continue to decrease until around age 60, when costs begin to trend upward again.

`TARGET_AMT` appears to be slightly higher for both younger (< 25) and older (>60) drivers, but the differences appear to be subtle.



AGEBIN	ct	mean_cost	median_cost
10	7	9650	3917
20	151	5412	4193
30	627	5601	4144
40	804	5470	3936
50	446	6280	4291
60	109	5907	4460
70	4	4284	4276

YOJ

`YOB` refers to the number of years in the insured's current job.

Here is a numerical summary, binned in 2 year increments:

```
              0     2     4     6     8     10     12     14   16    18  22   Sum
count     631.0  51.0 129.0 473.0 905.0 1752.0 2174.0 1248.0 305  37.0   2  7707
percent     8.2   0.7   1.7   6.1  11.7   22.7   28.2   16.2   4   0.5   0   100
```

Below is the statistical summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's    StdD    Skew    Kurt
   0.00    9.00   11.00   10.50   13.00   23.00  454.00    4.09   -1.20    4.18
```

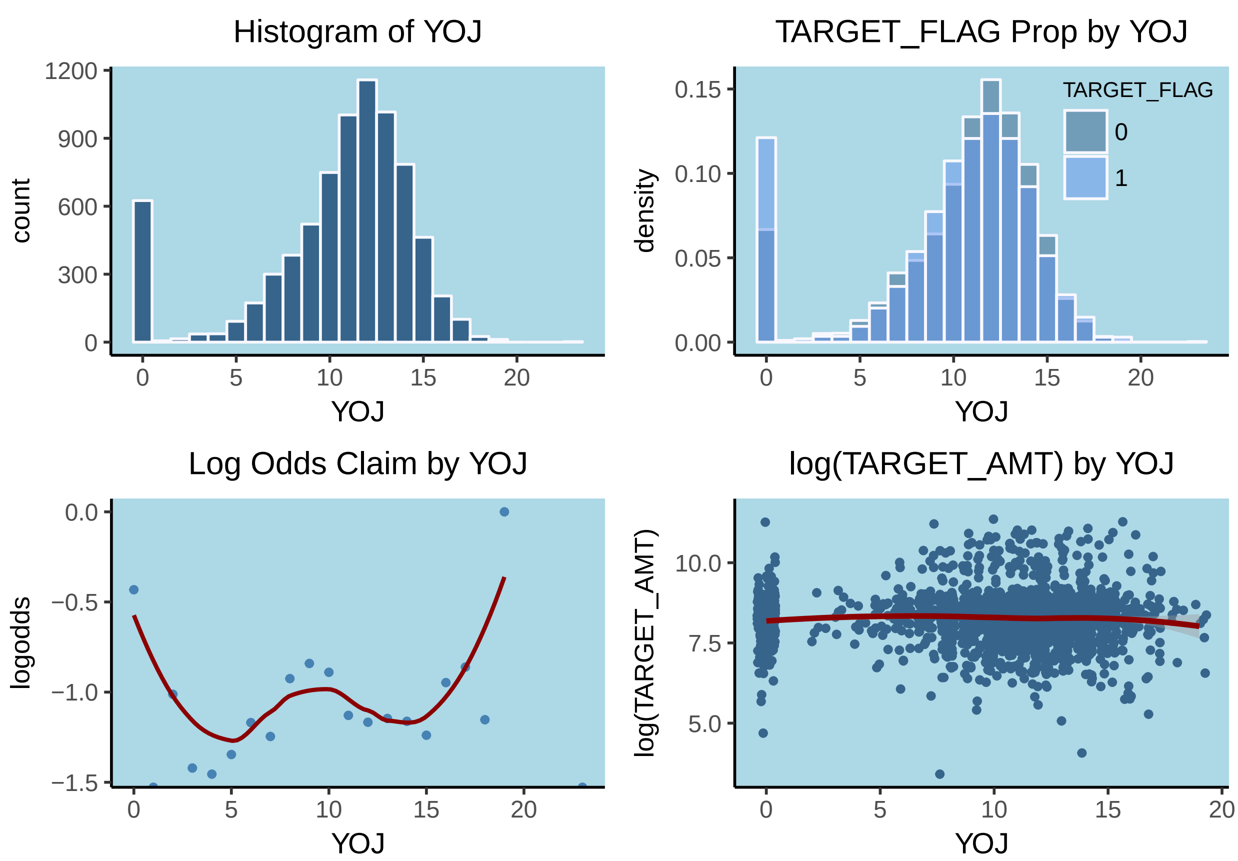
We have a significant number of NA records–454, or 5.6%–that we'll need to address.

The variable would be approximately normally distributed if it weren't for the high percentage of individuals with less than one year on the job.

Insureds with accidents have a relatively high proportion of individuals with less than a year on the job.

The relationship between log odds and `YOB` appears to be complex–see the fitted loess curve below.

The relationship between `YOJ` and `TARGET_AMT` is also not very clear.



YOJBIN	ct	mean_cost	median_cost
0	257	4765	3736
4	134	5993	4512
8	729	5964	4026
12	811	5590	4249

YOJBIN	ct	mean_cost	median_cost
16	99	6001	3969

TRAVTIME

Here is a summary of `TRAVTIME`, the commuting distance to work, in bins of 10 minutes:

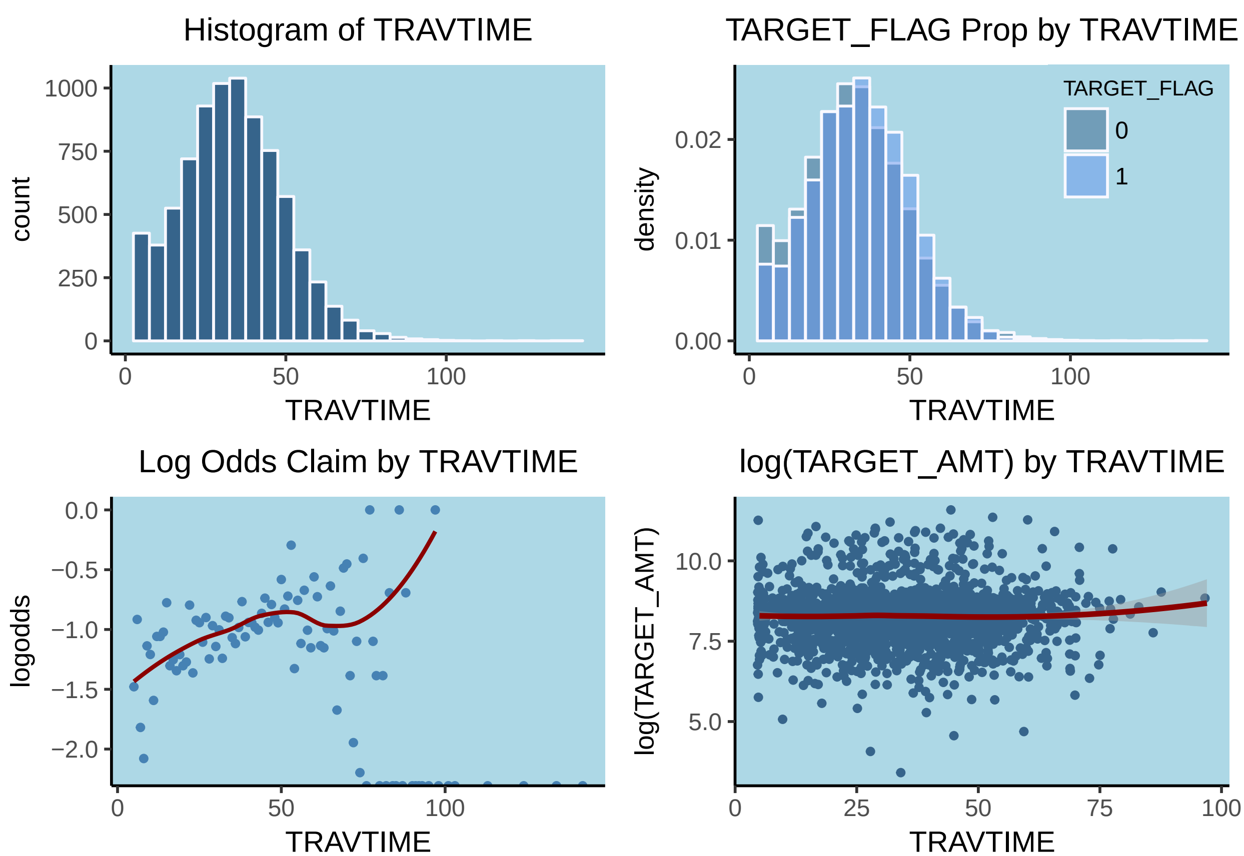
```
            0     10     20     30     40    50    60   70   80   90 100 110 120 130 140  Sum
count   550.0 1036.0 1788.0 2023.0 1531.0 784.0 305.0 94.0 33.0 11.0   2   1   1   1   1 8161
percent   6.7   12.7   21.9   24.8   18.8   9.6   3.7  1.2  0.4  0.1   0   0   0   0   0  100
```

Here is a statistical summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
   5.00   22.00   33.00   33.49   44.00  142.00   15.91    0.45    3.67
```

The distribution has a slight positive skew. The subset of insureds with no accidents have a higher proportion of individuals with short commute times. In the scatterplot below, we notice a generally positive relationship between log odds and `TRAVETIME`. Some of the curvature in the loess curve is likely driven by small sample sizes for long commute times.

Finally, we see a slight, upward curvature in the log of `TARGET_AMT` for long commute times. Judging from the loess curve confidence interval, this upward trend may not be statistically significant.



TRAVBIN	ct	mean_cost	median_cost
0	214	5395	4539
15	596	5846	3998

TRAVBIN	ct	mean_cost	median_cost
30	772	5671	4163
45	451	5516	3964
60	104	6470	4376
75	15	6169	4579
90	1	6909	6909

TIF

Here is a distribution of time in force values in bins of 2 year increments:

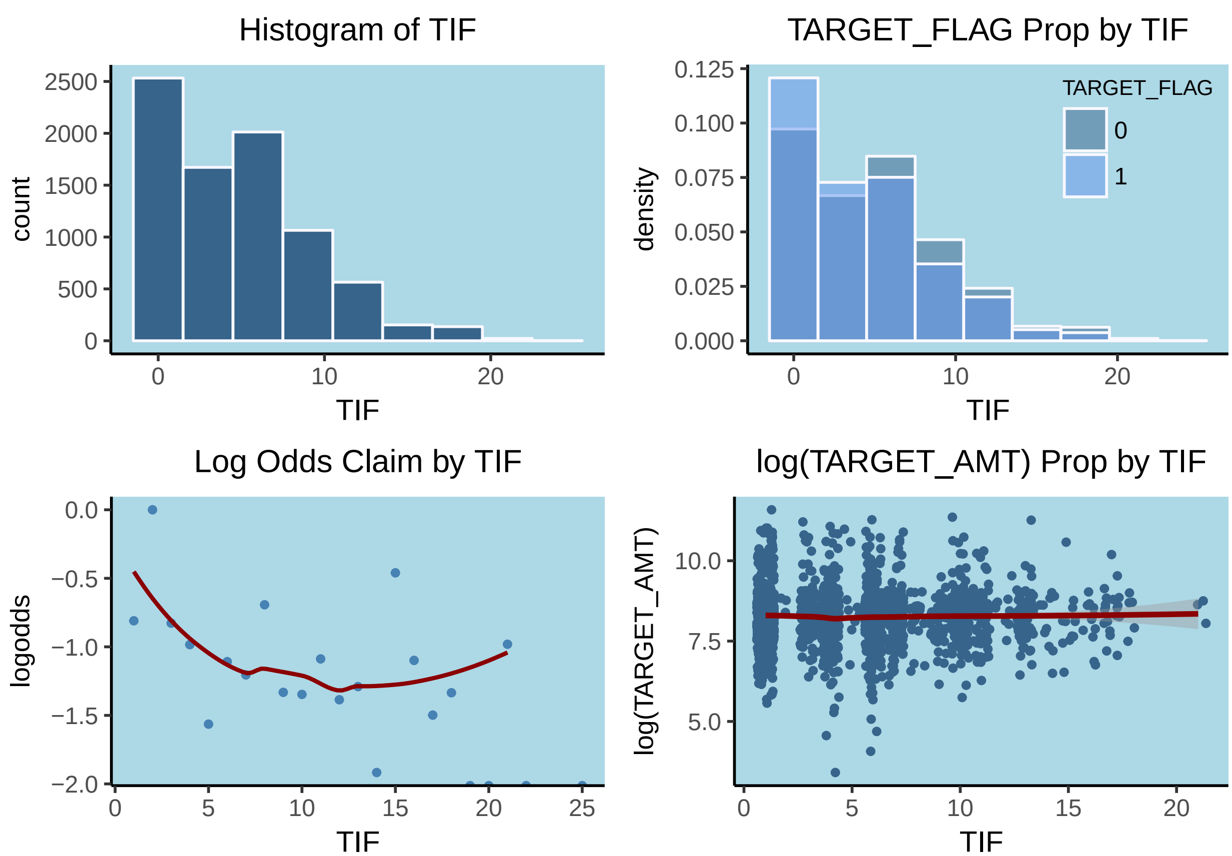
```
           0     2      4    6     8     10  12    14    16   18   20 22 24  Sum
count   2533 430.0 1294.0 1961 285.0 1022.0 323 109.0 148.0 32.0 19.0  3  2 8161
percent   31   5.3   15.9   24   3.5   12.5   4   1.3   1.8  0.4  0.2  0  0  100
```

Here is the statistical summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
   1.00    1.00    4.00    5.35    7.00   25.00    4.15    0.89    3.42
```

The distribution is somewhat positively skewed. We see somewhat small sub-samples for `TIF` values of 2-3 and 8-9. Also, the log odds vs. `TIF` scatterplot suggests a quadratic relationship.

Finally, there does not seem to be a significant relationship between log of `TARGET_AMT` and `TIF`.



TIFBIN	ct	mean_cost	median_cost
0	912	5858	4102

TIFBIN	ct	mean_cost	median_cost
4	823	5609	4049
8	289	5518	4220
12	91	5764	4061
16	35	5257	4868
20	3	5025	5610

CAR_AGE

The predictor CAR_AGE describes the age in years of the insured's car. Below we bin the data in increments of three:

```
        -3      0     3      6    9     12    15    18   21   24 27  Sum
count    1 1949.0 494.0 1512.0 1455 1035.0 720.0 369.0 96.0 18.0  2 7651
percent  0   25.5   6.5   19.8   19   13.5   9.4   4.8  1.3  0.2  0  100
```

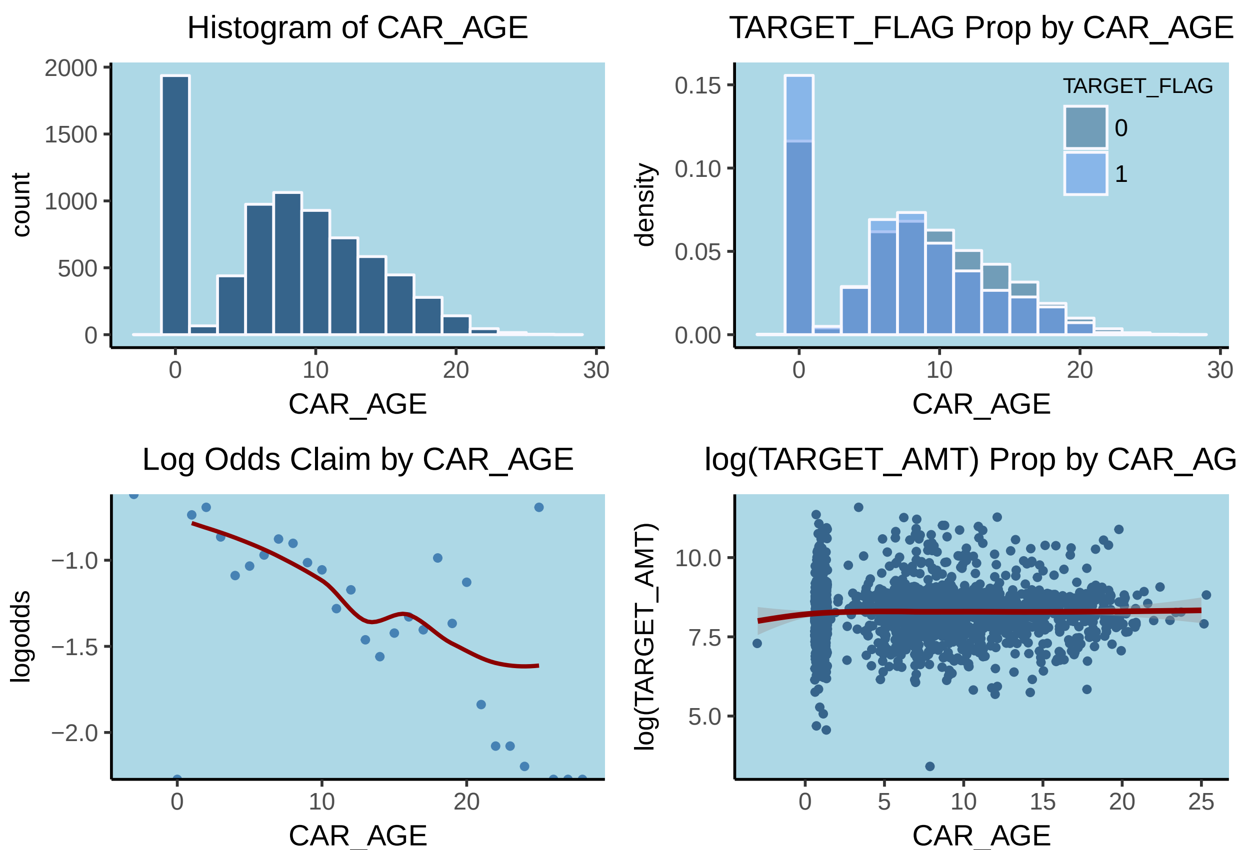
Here is the statistical summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's    StdD    Skew    Kurt
  -3.00    1.00    8.00    8.33   12.00   28.00  510.00    5.70    0.28    2.25
```

There is an observation that indicates a CAR_AGE value of -3. This is clearly an error. Also, there are 510 missing observations, representing 6% of total records. 1934 out of the 8161 (25%) of the records have a value of one.

Surprisingly, CAR_AGE appears to be negatively correlated with the log odds of a claim. Perhaps there is a confounding variable responsible for this result.

Finally, we expected to see a stronger relationship between CAR_AGE and the log of TARGET_AMT.
Our intuition was that payouts go down with the age of the car, as replacement value goes down with age. However, the bottom right scatter plot below does not seem to indicate a significant association.



CAR_AGE	ct	mean_cost	median_cost
-4	1	1469	1469
0	646	5824	4107
4	392	5884	4224
8	516	5789	4088
12	261	5431	4274
16	158	5327	4076
20	34	5438	3150
24	3	4480	3949

Binary Variables

PARENT1

Here is a record summary for `PARENT1`, a binary variable indicating if an insured is a single parent.

```
              No     Yes   Sum
count    7084.0 1077.0 8161
percent    86.8   13.2  100
```

The vast majority (87%) of individuals in the training data are not single parents.

Let's explore the relationship with `TARGET_FLAG`

```
        TARGET_FLAG
PARENT1     0     1   Sum
    No   5407 1677 7084
```

```
    Yes  601  476 1077
    Sum 6008 2153 8161
```

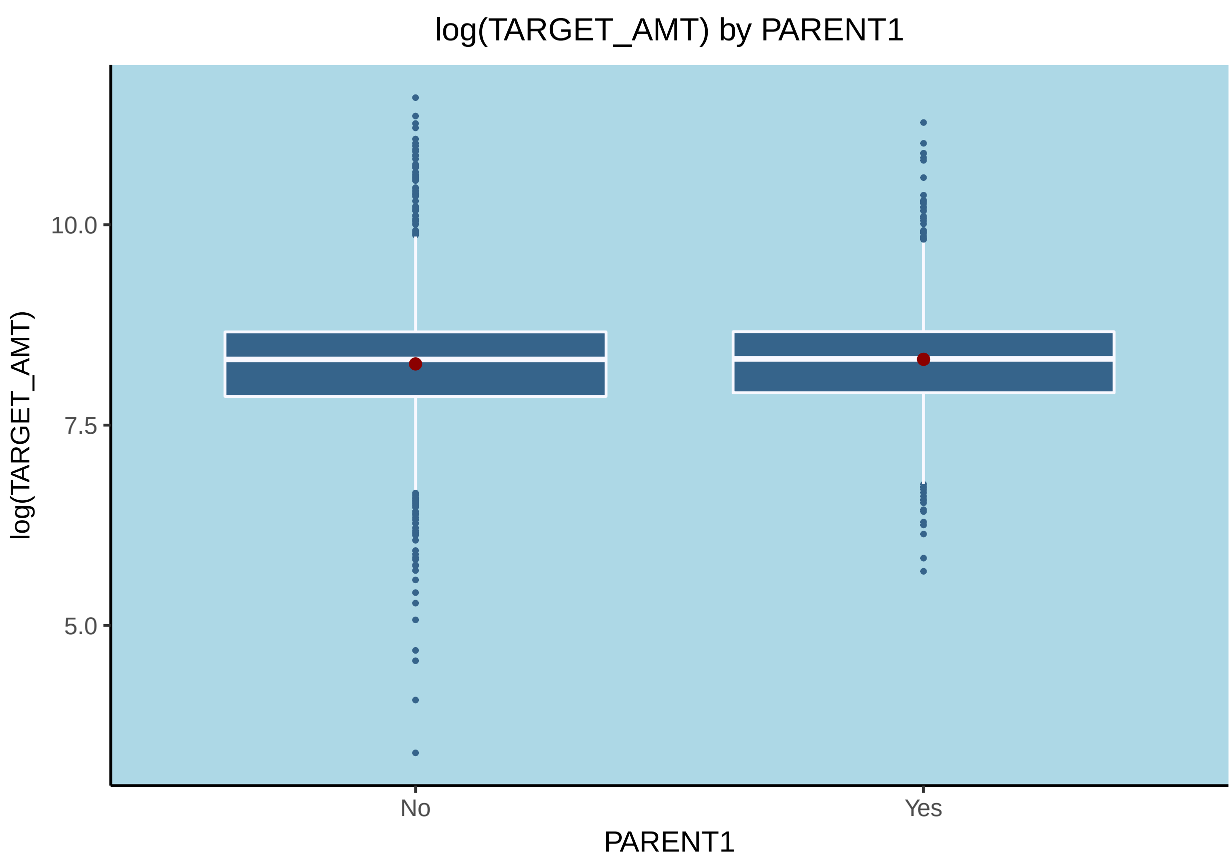
Now let's review the proportions of individuals involved in a crash, given `PARENT1` status:

```
       TARGET_FLAG
PARENT1    0    1
    No  0.76 0.24
    Yes 0.56 0.44
```

The is a 20% difference in the calculated proportions. This difference is statistically significant:

```
    2-sample test for equality of proportions with continuity correction

data:  tbl[1:2, 1:2]
X-squared = 201.7, df = 1, p-value < 0.00000000000000022
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1734351 0.2370404
sample estimates:
   prop 1    prop 2
0.7632693 0.5580316
```

Finally, lets review the relationship of `PARENT1` with `TARGET_AMT`.



PARENT1	ct	mean_cost	median_cost
No	1677	5603	4101

PARENT1	ct	mean_cost	median_cost
Yes	476	6050	4133

The distribution of log amounts is pretty similar for both values of `PARENT1`, with perhaps a very modest increase for single parent insureds.

MSTATUS

The predictor, `MSTATUS`, refers to the married status of the policyholder.

```
          No  Yes  Sum
count   3267 4894 8161
percent   40   60  100
```

There is a fairly balanced split (60/40) between married and single insureds.

Let's look at the relationship with `TARGET_FLAG`

```
       TARGET_FLAG
MSTATUS    0    1  Sum
    No  2167 1100 3267
    Yes 3841 1053 4894
    Sum 6008 2153 8161
```

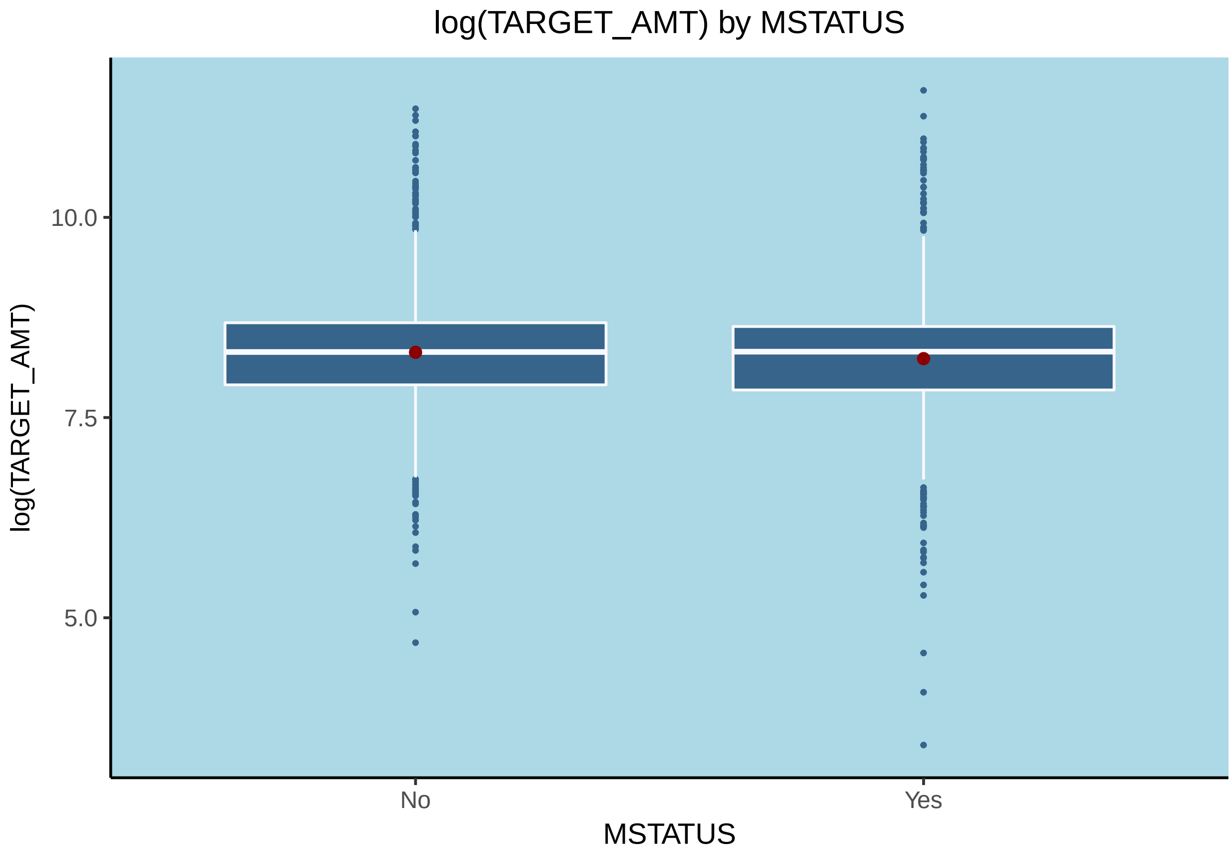
Now let's review the proportions of individuals involved in a crash, given `MSTATUS` status:

```
       TARGET_FLAG
MSTATUS    0    1
    No  0.66 0.34
    Yes 0.78 0.22
```

The 12% difference in proportions is statistically significant:

```
    2-sample test for equality of proportions with continuity correction

data:  tbl[1:2, 1:2]
X-squared = 148.38, df = 1, p-value < 0.00000000000000022
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1416726 -0.1014053
sample estimates:
   prop 1    prop 2
0.6632997 0.7848386
```

Now lets explore the relationship of `MSTATUS` with `TARGET_AMT`.



MSTATUS	ct	mean_cost	median_cost
No	1100	5967	4098
Yes	1053	5426	4117

The distribution of log amounts is very similar across `MSTATUS` type. Non married individuals appear to have a slightly higher average log cost compared to the married cohort. However, median costs are almost identical between the two cohorts.

SEX

The variable, `SEX`, denotes the gender of the insured policyholder.

```
              F      M  Sum
count   4375.0 3786.0 8161
percent   53.6   46.4  100
```

The split between males and females is split almost 50/50.

Let's review the relationship with `TARGET_FLAG`.

```
     TARGET_FLAG
SEX      0    1  Sum
  F   3183 1192 4375
  M   2825  961 3786
  Sum 6008 2153 8161
```

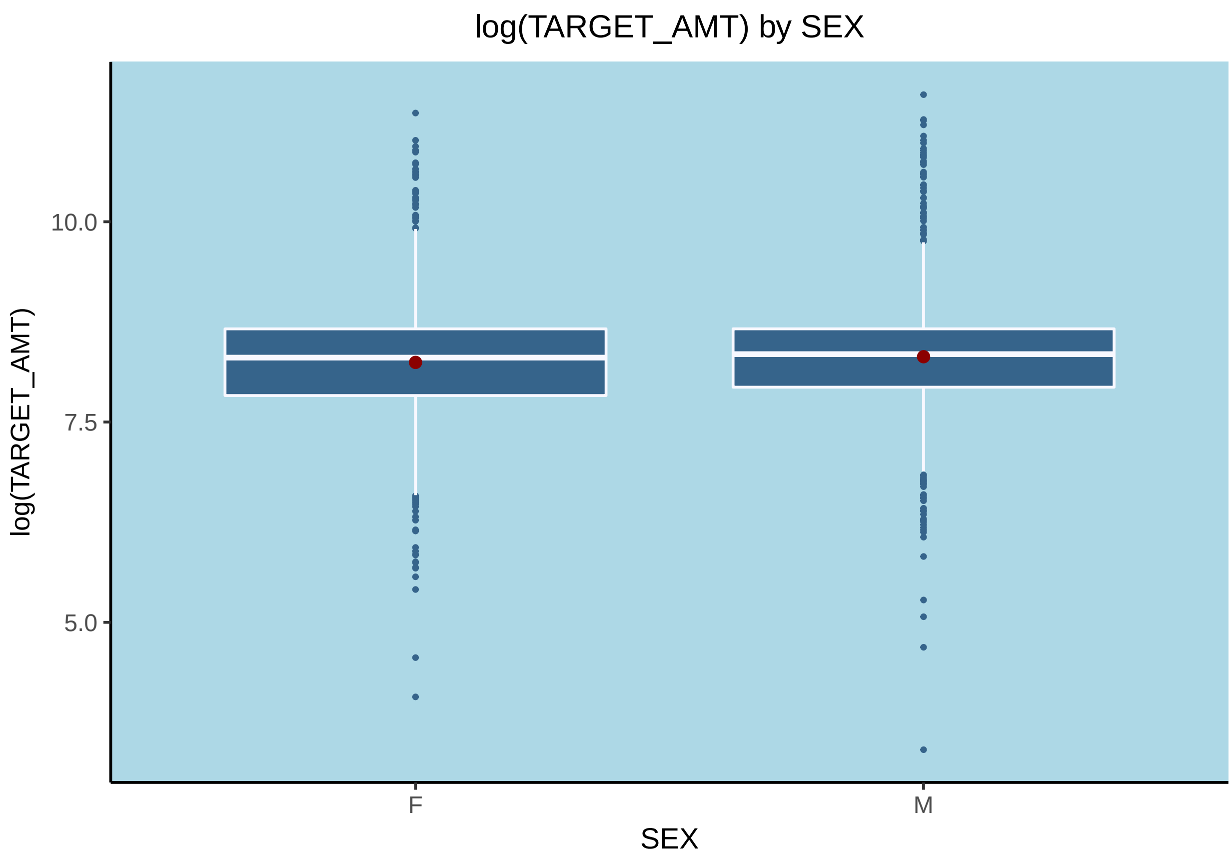
Here are the proportions of individuals involved in a crash, given gender type:

```
   TARGET_FLAG
SEX    0    1
  F 0.73 0.27
  M 0.75 0.25
```

The 2% difference in proportions is not statistically significant:

```
    2-sample test for equality of proportions with continuity correction

data:  tbl[1:2, 1:2]
X-squared = 3.5307, df = 1, p-value = 0.06024
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0380106016  0.0007561151
sample estimates:
   prop 1    prop 2
0.7275429 0.7461701
```

Now lets explore the relationship of `SEX` with `TARGET_AMT`.



SEX	ct	mean_cost	median_cost
F	1192	5344	4039
M	961	6147	4220

The log dollar cost of claims seems to be slightly higher for males, on average.

CAR_USE

CAR_USE is a predictor that indicates how the vehicle is uses (i.e. commercial or private purposes).

```
        Commercial Private  Sum
count       3029.0  5132.0 8161
percent       37.1    62.9  100
```

The majority of the observations involve private use vehicles, at 60%.

Let's explore the relationship with TARGET_FLAG.

```
            TARGET_FLAG
CAR_USE         0    1  Sum
  Commercial 1982 1047 3029
  Private    4026 1106 5132
  Sum        6008 2153 8161
```

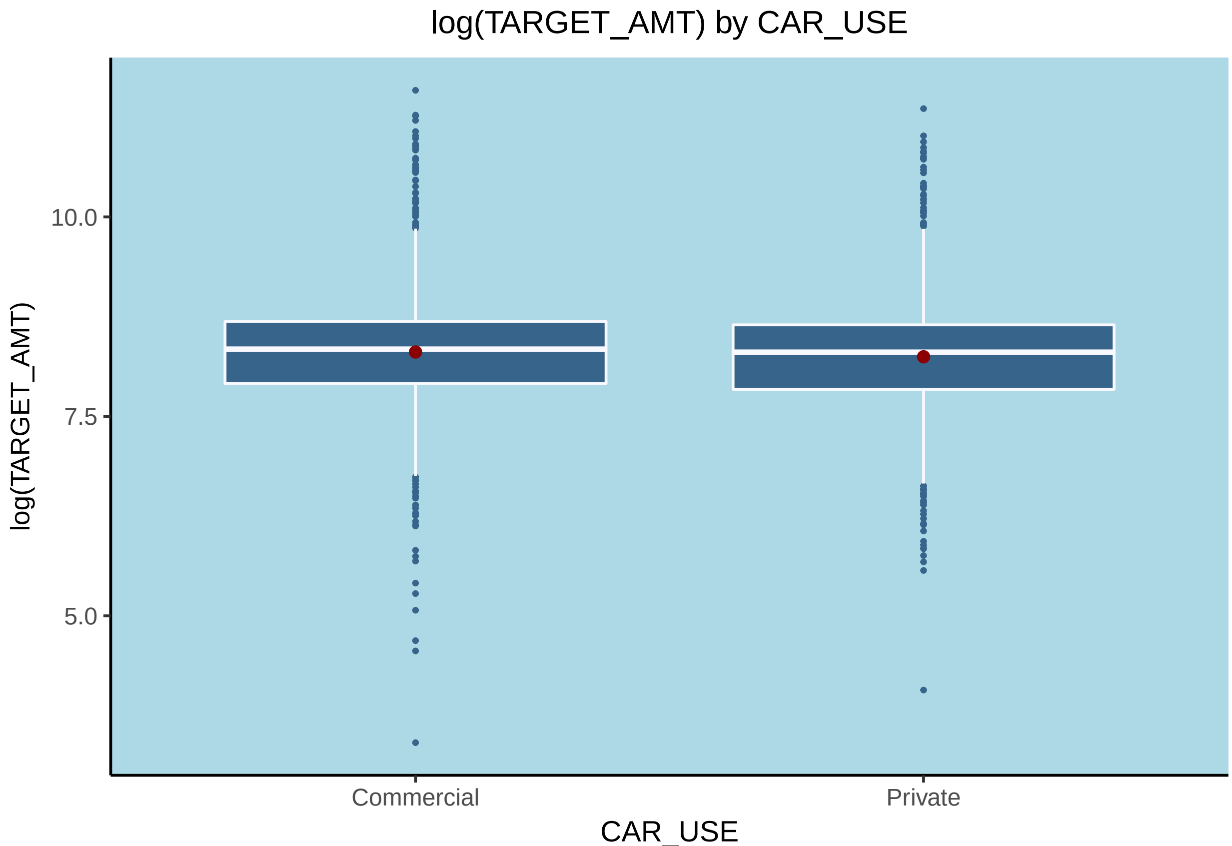
Below are the proportions of individuals involved in a crash, given the CAR_USE indicator:

```
            TARGET_FLAG
CAR_USE         0    1
  Commercial 0.65 0.35
  Private    0.78 0.22
```

There is a 13% difference in proportions between the two cohorts. This result is statistically significant.

```
    2-sample test for equality of proportions with continuity correction

data:  tbl[1:2, 1:2]
X-squared = 165.45, df = 1, p-value < 0.00000000000000022
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1507428 -0.1095535
sample estimates:
   prop 1    prop 2
0.6543414 0.7844895
```

Finally, lets explore the relationship of CAR_USE with TARGET_AMT.



CAR_USE	ct	mean_cost	median_cost
Commercial	1047	6099	4193
Private	1106	5327	4037

The log dollar cost of claims involving commercial transit appears to somewhat higher than costs associated with private transportation.

RED_CAR

`RED_CAR` is a binary predictor indicating whether a care is primarily red in color.

```
            no    yes  Sum
count   5783.0 2378.0 8161
percent   70.9   29.1  100
```

We see only 30% of vehicles in the red category.

Let's look at the relationship with `TARGET_FLAG`.

```
       TARGET_FLAG
RED_CAR    0    1  Sum
    no  4246 1537 5783
    yes 1762  616 2378
    Sum 6008 2153 8161
```

Here are the proportions of individuals involved in a crash, given the `RED_CAR` status:

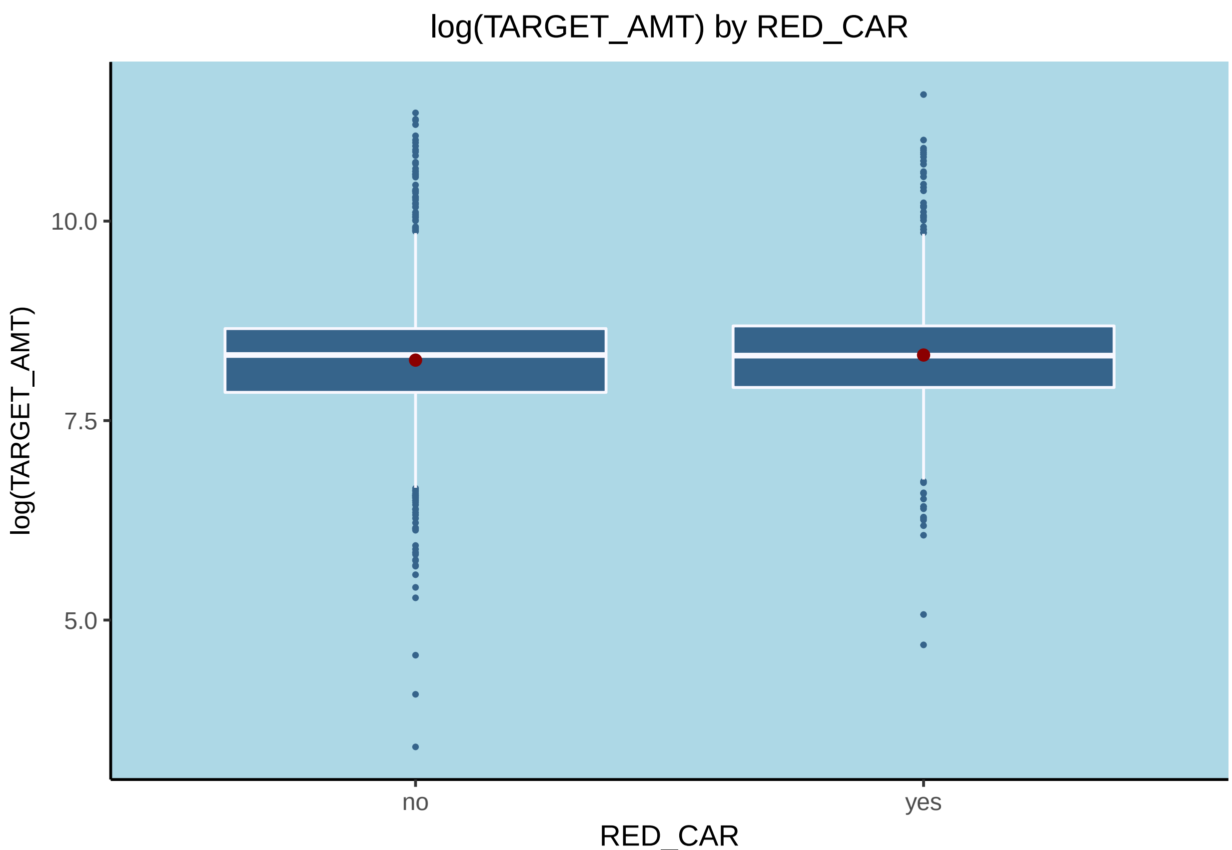
```
       TARGET_FLAG
```

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

There is only 1% difference in proportions between red and non-red cars. The result is not statistically significant.

```
    2-sample test for equality of proportions with continuity correction

data:  tbl[1:2, 1:2]
X-squared = 0.35996, df = 1, p-value = 0.5485
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02800322  0.01452763
sample estimates:
   prop 1    prop 2
0.7342210 0.7409588
```

Now, we'll explore the relationship between `RED_CAR` and `TARGET_AMT`.



RED_CAR	ct	mean_cost	median_cost
no	1537	5568	4112
yes	616	6036	4082

The distribution for each cohort is very similar, with an subtle uptick in average log costs for red car types.

REVOKED

The variable `REVOKED` indicates whether a driver's license has been suspended within the last seven years.

```
            No    Yes  Sum
count   7161.0 1000.0 8161
percent   87.7   12.3  100
```

Only 12% of drivers in the training data have a former license suspension on record.

Here is a look look at the relationship of `REVOKED` with `TARGET_FLAG`.

```
       TARGET_FLAG
REVOKED    0    1  Sum
    No  5451 1710 7161
    Yes  557  443 1000
    Sum 6008 2153 8161
```

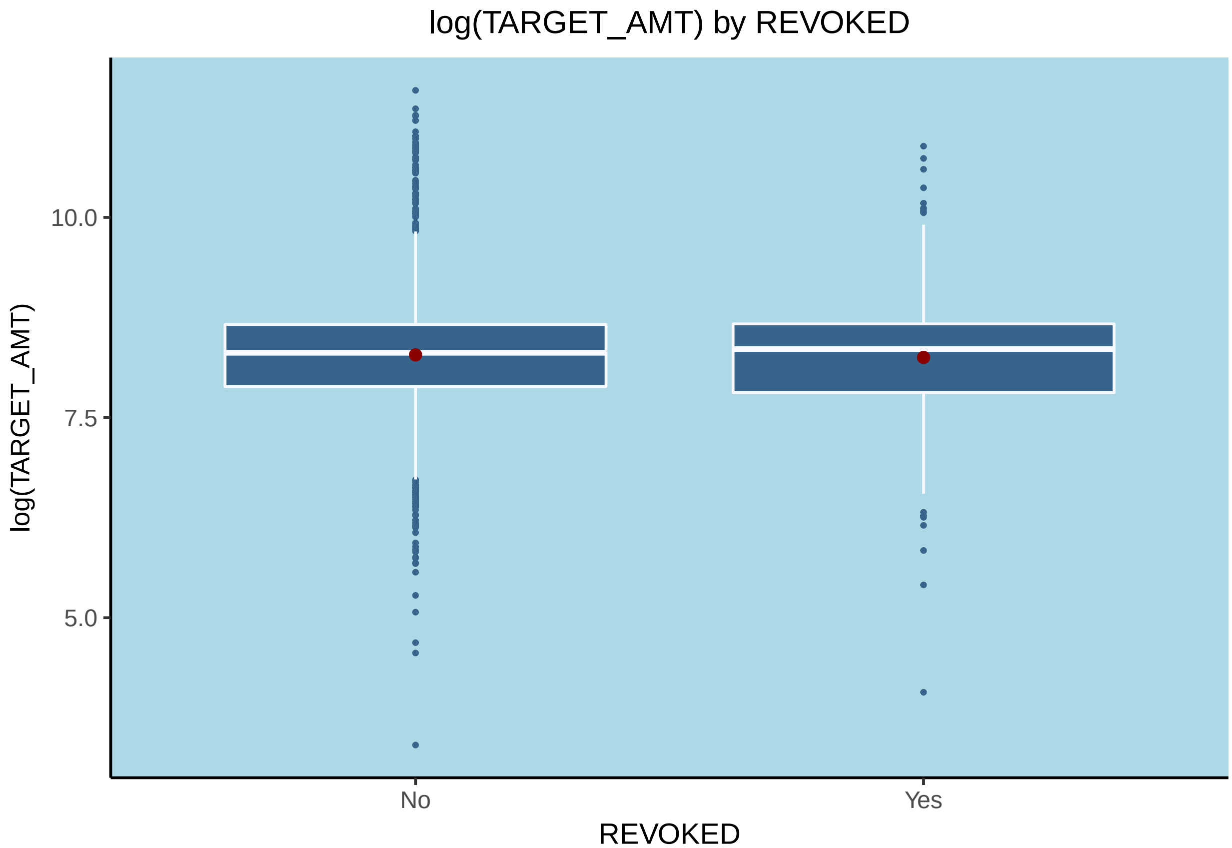
Below are the proportions of individuals involved in a crash, given the `REVOKED` status:

```
       TARGET_FLAG
REVOKED    0    1
    No  0.76 0.24
    Yes 0.56 0.44
```

There is statistically significant difference in proportions by `REVOKED` type, with a 20% observed difference in the training data.

```

    2-sample test for equality of proportions with continuity correction

data:  tbl[1:2, 1:2]
X-squared = 187.35, df = 1, p-value < 0.00000000000000022
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1713042 0.2371089
sample estimates:
   prop 1    prop 2 
0.7612065 0.5570000 
```

Finally, let's explore the relationship between `REVOKED` and `TARGET_AMT`.



REVOKED	ct	mean_cost	median_cost
No	1710	5848	4059
Yes	443	5140	4254

The distribution for each cohort is very similar. In the training data, we very slight increase in average log costs for folks that have not had their license suspended. However, the median log cost seems to be slightly higher for folks with a prior suspended license.

URBANICITY

The predictor `URBANICITY` indicates whether the environment in which the driver primarily drives: urban vs. rural.

```
        Highly Rural/ Rural Highly Urban/ Urban  Sum
count                1669.0              6492.0 8161
percent                20.5                79.5  100
```

Only 30% of drivers in the training data are categorized as being in a rural environment. .

Let's look at the relationship of `URBANICITY` with `TARGET_FLAG`.

```
                     TARGET_FLAG
URBANICITY               0    1  Sum
  Highly Rural/ Rural 1554  115 1669
  Highly Urban/ Urban 4454 2038 6492
  Sum                 6008 2153 8161
```

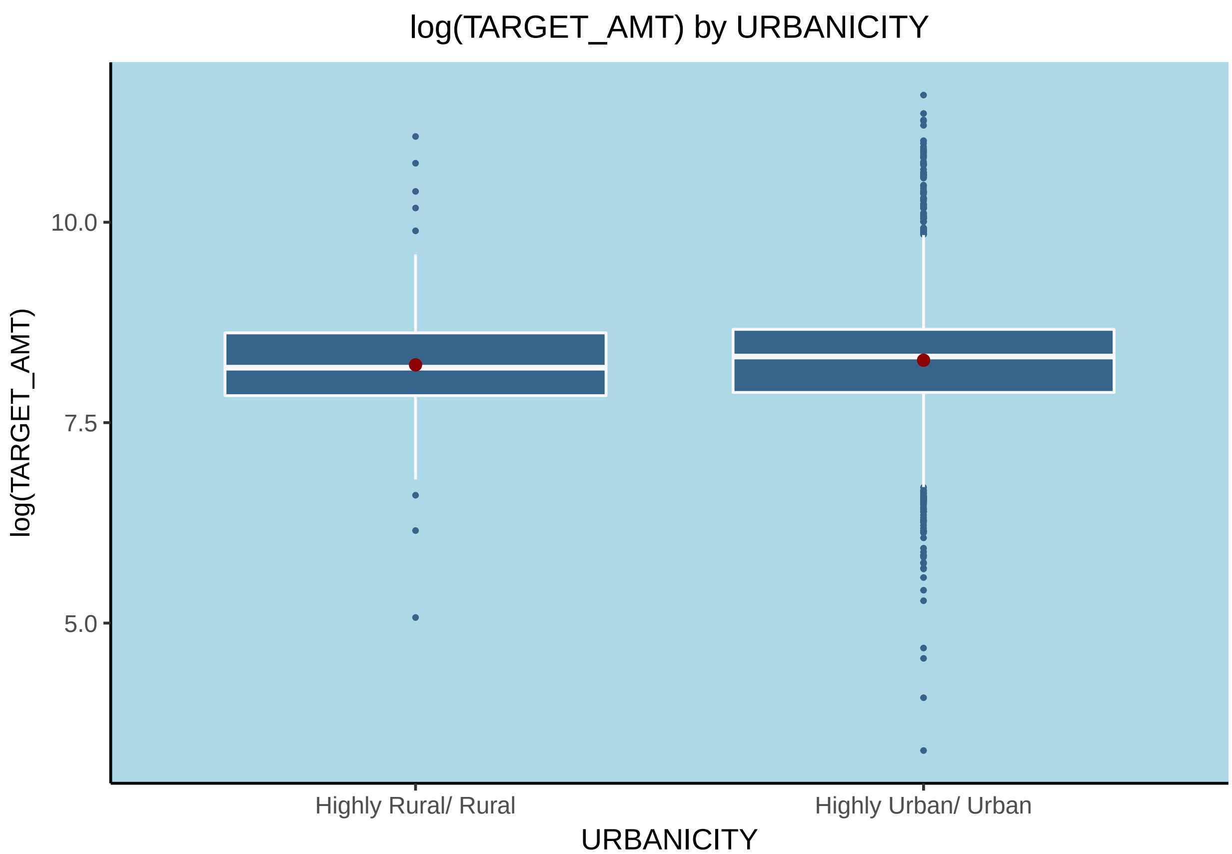
Her are the proportions of individuals involved in a crash, given the `URBANICITY` category :

```
                     TARGET_FLAG
URBANICITY               0    1
  Highly Rural/ Rural 0.93 0.07
  Highly Urban/ Urban 0.69 0.31
```

There is a huge difference, 24%, between the two proportions. Urban drivers appear to be significantly more at risk for being involved in a crash.

```
    2-sample test for equality of proportions with continuity correction

data:  tbl[1:2, 1:2]
X-squared = 409.14, df = 1, p-value < 0.0000000000000022
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2280583 0.2619842
sample estimates:
   prop 1    prop 2
0.9310965 0.6860752
```

Let's review the relationship between `URBANICITY` and `TARGET_AMT`.



URBANICITY	ct	mean_cost	median_cost
Highly Rural/ Rural	115	5545	3589
Highly Urban/ Urban	2038	5711	4125

Urban drivers tend to have somewhat higher claim costs, given a crash. There also appears to be more variability in claim costs for urban drivers compared to their rural counterparts.

Multinomial Categorical Variables

JOB

The variable `JOB` indicates the insured's job category. there is also blank category which we interpret to mean unknown or not working.

		Blue Collar	Clerical	Doctor	Home Maker	Lawyer	Manager	Professional	Student	Sum
count	526.0	1825.0	1271.0	246	641.0	835.0	988.0	1117.0	712.0	8161
percent	6.4	22.4	15.6	3	7.9	10.2	12.1	13.7	8.7	100

Here is a breakdown of job categories by `TARGET_FLAG`:

	TARGET_FLAG		
JOB	0	1	Sum
	390	136	526
Blue Collar	1191	634	1825
Clerical	900	371	1271
Doctor	217	29	246
Home Maker	461	180	641
Lawyer	682	153	835
Manager	851	137	988
Professional	870	247	1117
Student	446	266	712
Sum	6008	2153	8161

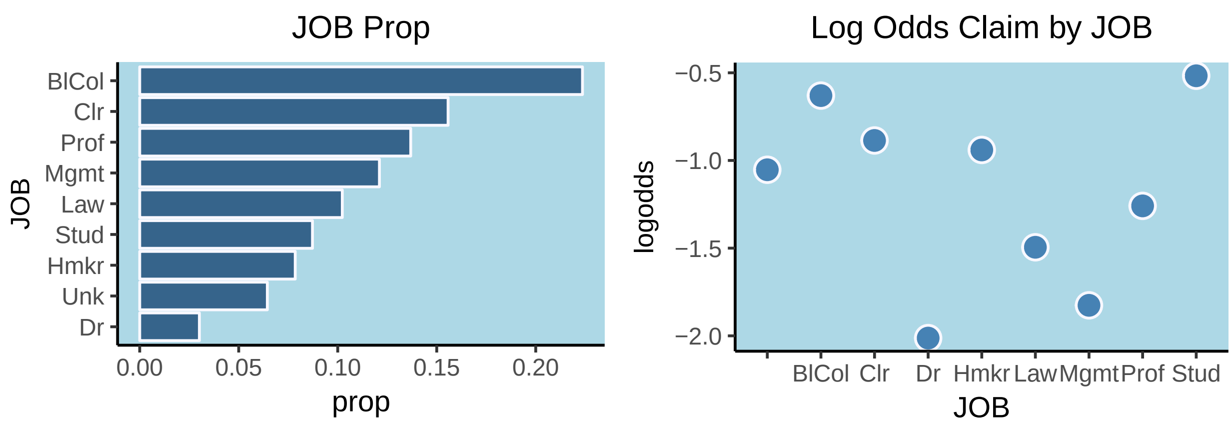
Let's calculate the proportion of observations by job category in each `TARGET_FLAG` indicator:

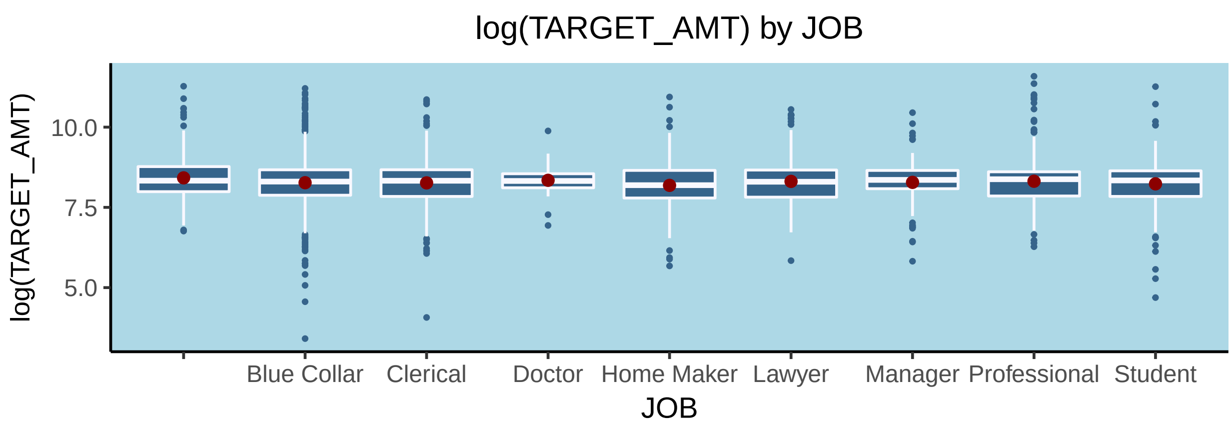
	TARGET_FLAG	
JOB	0	1
	0.74	0.26
Blue Collar	0.65	0.35
Clerical	0.71	0.29
Doctor	0.88	0.12
Home Maker	0.72	0.28
Lawyer	0.82	0.18
Manager	0.86	0.14
Professional	0.78	0.22
Student	0.63	0.37

There appears to be significant variability in the probability of a claim by occupational category, with students and blue collar jobs leading the pack.

We reject the null hypothesis that all proportions are identical:

```
    9-sample test for equality of proportions without continuity correction

data:  tbl[1:9, 1:2]
X-squared = 261.06, df = 8, p-value < 0.00000000000000022
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4    prop 5    prop 6    prop 7    prop 8    prop 9 
0.7414449 0.6526027 0.7081039 0.8821138 0.7191888 0.8167665 0.8613360 0.7788720 0.6264045 
```





JOB	ct	mean_cost	median_cost
	136	6904	4155
Blue Collar	634	5890	4042
Clerical	371	5446	4144
Doctor	29	4896	4117
Home Maker	180	4951	3612
Lawyer	153	5991	4019
Manager	137	4944	4256
Professional	247	6560	4348
Student	266	5021	4188

There is moderate variability in log costs across job categories, with the unknown, student, and professional, and manager categories appearing to have higher median costs than than other five categories. The relationship are somewhat different though when viewing mean costs (or mean log costs) by category.

CAR_TYPE

The predictor `CAR_TYPE` indicates the insured's vehicle type.

```
        Minivan Panel Truck Pickup Sports Car    SUV   Van  Sum
count    2145.0       676.0   1389      907.0 2294.0 750.0 8161
percent    26.3         8.3     17       11.1   28.1   9.2  100
```

Here is a breakdown of CAR_TYPE categories by `TARGET_FLAG`:

```
             TARGET_FLAG
CAR_TYPE         0     1   Sum
```

```
  Minivan     1796  349 2145
  Panel Truck  498  178  676
  Pickup       946  443 1389
  Sports Car   603  304  907
  SUV         1616  678 2294
  Van          549  201  750
  Sum         6008 2153 8161
```

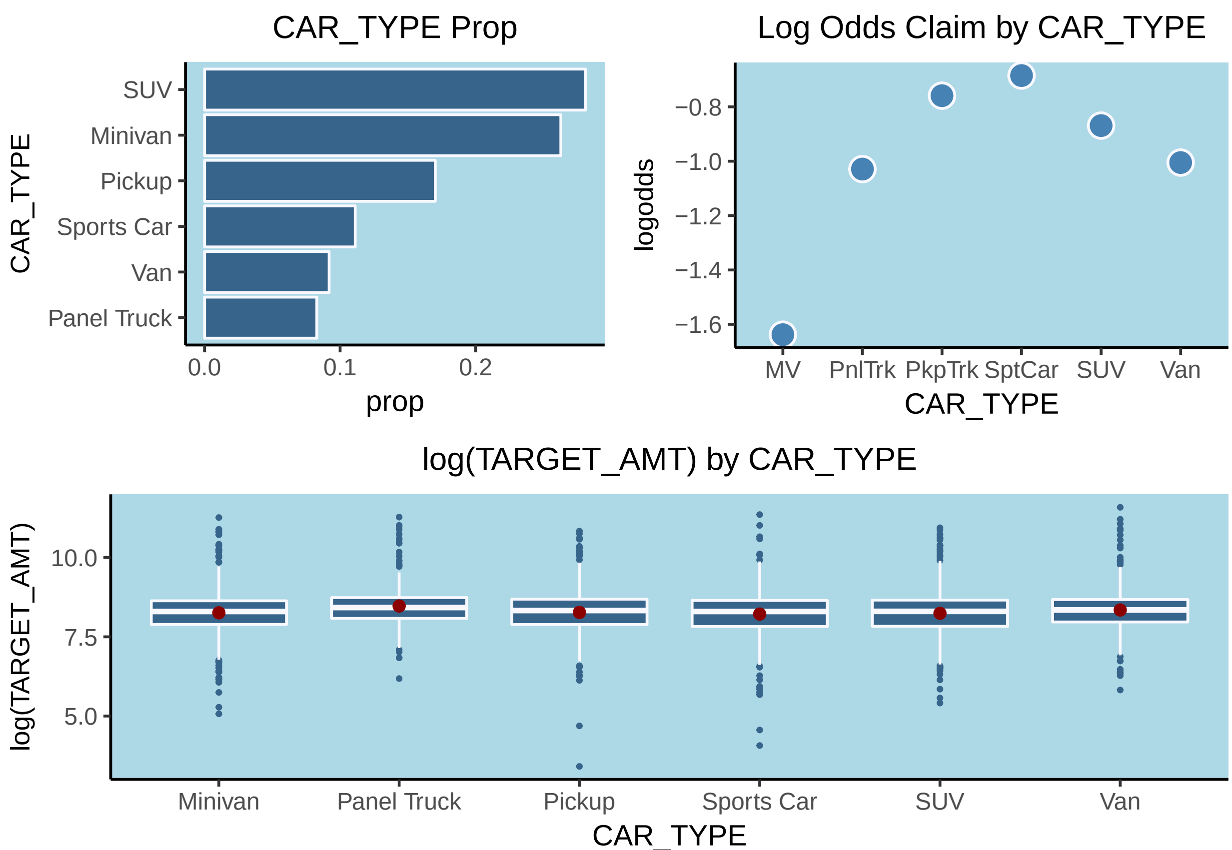
We'll now calculate the proportion of observations by CAR_TYPE category in each `TARGET_FLAG` indicator:

```
             TARGET_FLAG
CAR_TYPE          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
```

There is significant variability in the probability of a claim by occupational category, with sports cars having the highest proportion of accidents, and minivan have the lowest proportion.

We reject the null hypothesis that all proportions are identical:

```
    6-sample test for equality of proportions without continuity correction

data:  tbl[1:6, 1:2]
X-squared = 170.38, df = 5, p-value < 0.00000000000000022
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4    prop 5    prop 6 
0.8372960 0.7366864 0.6810655 0.6648291 0.7044464 0.7320000 
```



CAR_TYPE	ct	mean_cost	median_cost
Minivan	349	5602	4023
Panel Truck	178	7465	4556
Pickup	443	5430	4147
Sports Car	304	5413	4023
SUV	678	5241	4032
Van	201	6909	4220

Panel trucks and vans appear to have significantly higher log claim costs compared to other car types.

Ordinal Categorical Variables

EDUCATION

The variable `EDUCATION` denotes the insured's highest level of education attained.

```
        High School Bachelors Masters   PhD  Sum
count        3533.0    2242.0  1658.0 728.0 8161
percent        43.3      27.5    20.3   8.9  100
```

Below is a breakdown of EDUCATION categories by `TARGET_FLAG`:

```
             TARGET_FLAG
EDUCATION        0    1  Sum
  High School 2355 1178 3533
  Bachelors   1719  523 2242
```

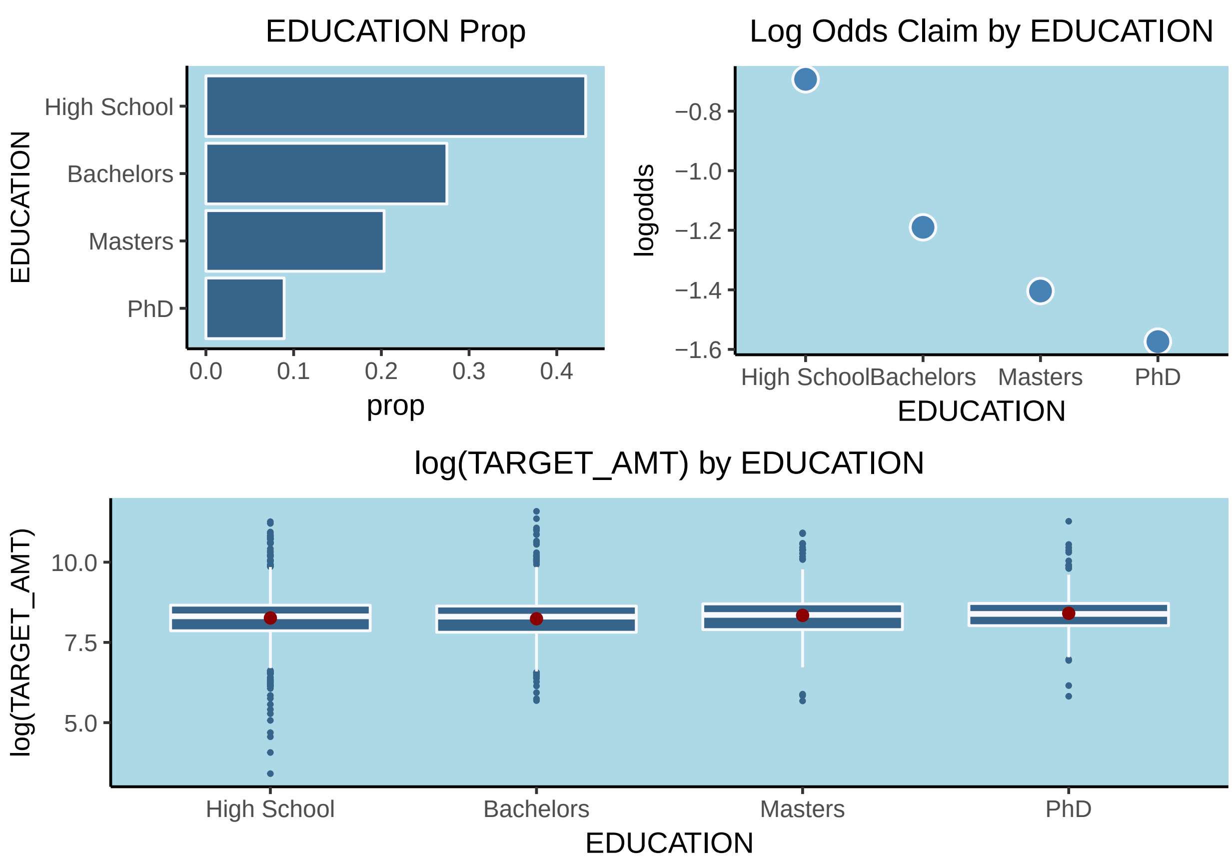
```
  Masters      1331  327 1658
  PhD           603  125  728
  Sum          6008 2153 8161
```

Let's review the proportion of observations by EDUCATION category in each `TARGET_FLAG` indicator:

```
             TARGET_FLAG
EDUCATION        0    1
  High School 0.67 0.33
  Bachelors   0.77 0.23
  Masters     0.80 0.20
  PhD         0.83 0.17
```

There are statistically significant differences between the categories, with higher levels of educational attainment associated with lower probabilities of crashes.

```

    4-sample test for equality of proportions without continuity correction

data:  tbl[1:4, 1:2]
X-squared = 168.58, df = 3, p-value < 0.00000000000000022
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4 
0.6665723 0.7667261 0.8027744 0.8282967 
```



EDUCATION	ct	mean_cost	median_cost
High School	1178	5451	4079

EDUCATION	ct	mean_cost	median_cost
Bachelors	523	5883	4036
Masters	327	5966	4256
PhD	125	6623	4395

Interestingly, the `TARGET_AMT` appears to have a positive association with the level of `EDUCATION`.

Continuous Variables

INCOME

Below is table of values of the `INCOME` predictor, with wage bucketed into $30K increments (i.e. [0,30), [30,60), [60,90), etc.])

```
              0      30      60     90    120    150    180   210   240   270  300  330  360  Sum
count    2059.0  2206.0  1671.0  951.0  417.0  203.0  117.0  56.0  20.0  11.0    3    1    1 7716
percent    26.7    28.6    21.7   12.3    5.4    2.6    1.5   0.7   0.3   0.1    0    0    0  100
```

Here is a statistical summary:

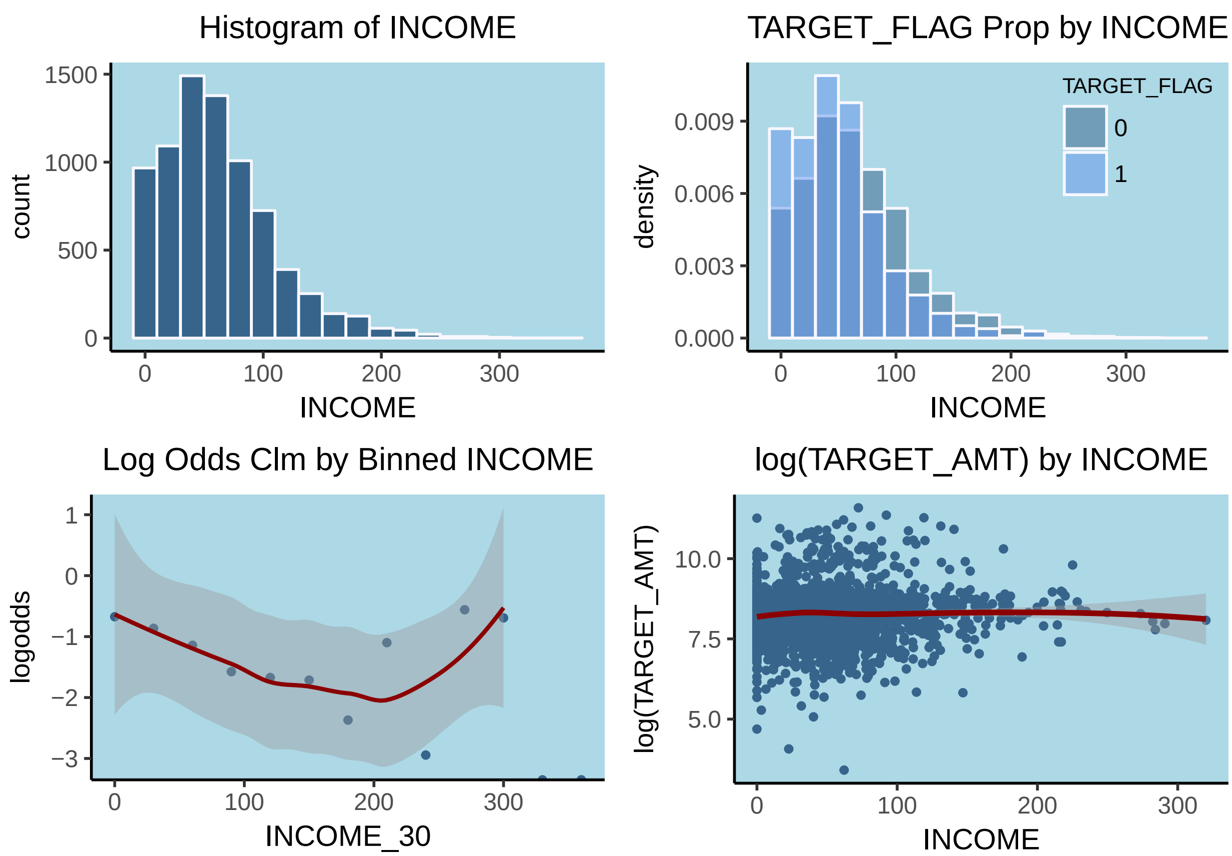
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's    StdD    Skew    Kurt
   0.00   28.10   54.03   61.90   85.99  367.03  445.00   47.57    1.19    5.13
```

There are some missing values, 445 (5.4% of total), that we'll address later.

The distribution of `INCOME` is right skewed, with a significant number of observations indicating $0 in income.

Individuals involved in crashes appear to have a high proportion of low-wage earners. We see the log-odds of a crash decreasing with increases to income–see the lower left scatter plot and loess curve. We see the loess curve starting to bend upward around $210k, although this phenomenon is possibly due to sparse data for high wage earners.

Finally, the relationship between income and log `TARGET_AMT` also does not appear to be very strong.



INCBIN	ct	mean_cost	median_cost
0	695	5102	4045
30	654	5909	4299
60	404	6053	3988
90	163	6603	4254
120	66	6107	3856
150	31	5671	4738
180	10	4091	3940
210	14	6020	5402
240	1	4104	4104
270	4	3100	3007
300	1	3231	3231

HOME_VAL

Below is table of values of the `HOME_VAL` predictor, with home appraisals bucketed into $50K increments.

```
              0    50   100    150    200   250   300   350   400  450  500  550 600 650 700 750 850  Sum
count    2294.0 371.0 902.0 1265.0 1181.0 751.0 435.0 208.0 128.0 85.0 43.0 21.0 7.0   3   1   1   1 7697
percent    29.8   4.8  11.7   16.4   15.3   9.8   5.7   2.7   1.7  1.1  0.6  0.3 0.1   0   0   0   0  100
```

Below is a statistical summary:

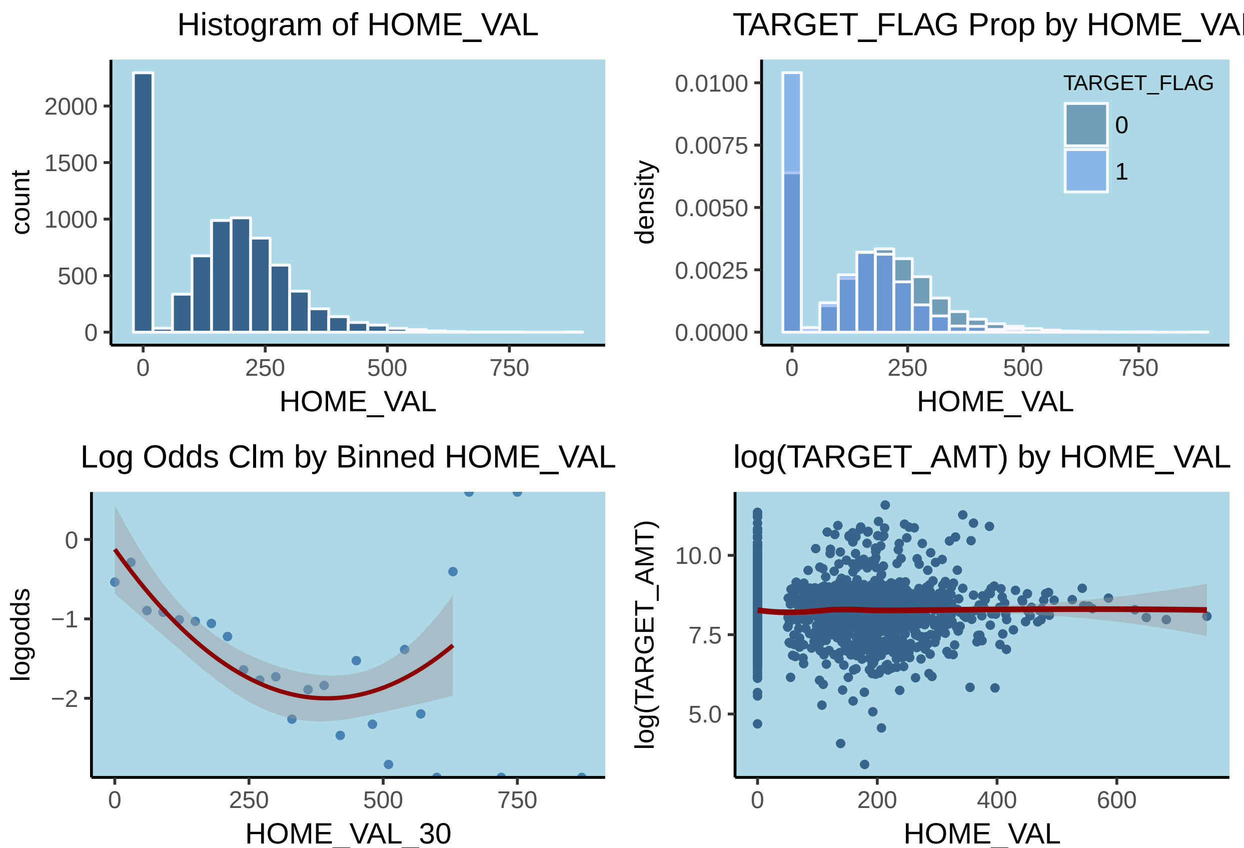
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's    StdD    Skew    Kurt
   0.00    0.00  161.16  154.87  238.72  885.28  464.00  129.12    0.49    2.98
```

There are some missing values, 464 (5.7% of total). We'll address these later.

The distribution of `HOME_VAL` is right skewed, with a large proportion of observations indicating $0 in HOME_VAL–this probably refers to policyholders who are not homeowners.

Individuals involved in crashes have a higher proportion of low home values. The loess curve in the lower left scatterplot indicates a decreasing log odds of claim occurrence with increasing incomes. The upward trend in the curve around $400k is possibly due to the the smaller sample size for higher home values.

Finally, the relationship between `HOME_VAL` and log `TARGET_AMT` also does not appear to be very strong.



HOMEVALBIN	ct	mean_cost	median_cost
0	846	5550	4006
50	111	4591	4400
100	247	5233	4185
150	328	6004	4280
200	264	6086	4015
250	115	5573	4093
300	59	6990	4021
350	24	9735	4071
400	14	4011	3868
450	14	4458	3981
500	3	5917	5455
550	3	4756	4435
600	2	3540	3540
650	1	2908	2908
750	1	3231	3231

BLUEBOOK

Below is table of values of the `BLUEBOOK` predictor, with values bucketed into $4K increments.

	0	4	8	12	16	20	24	28	32	36	40	44	48	56	60	64	68	Sum
count	281.0	1327.0	1503.0	1548	1238.0	903.0	628.0	390.0	190.0	83	45.0	14.0	6.0	1	2	1	1	8161
percent	3.4	16.3	18.4	19	15.2	11.1	7.7	4.8	2.3	1	0.6	0.2	0.1	0	0	0	0	100

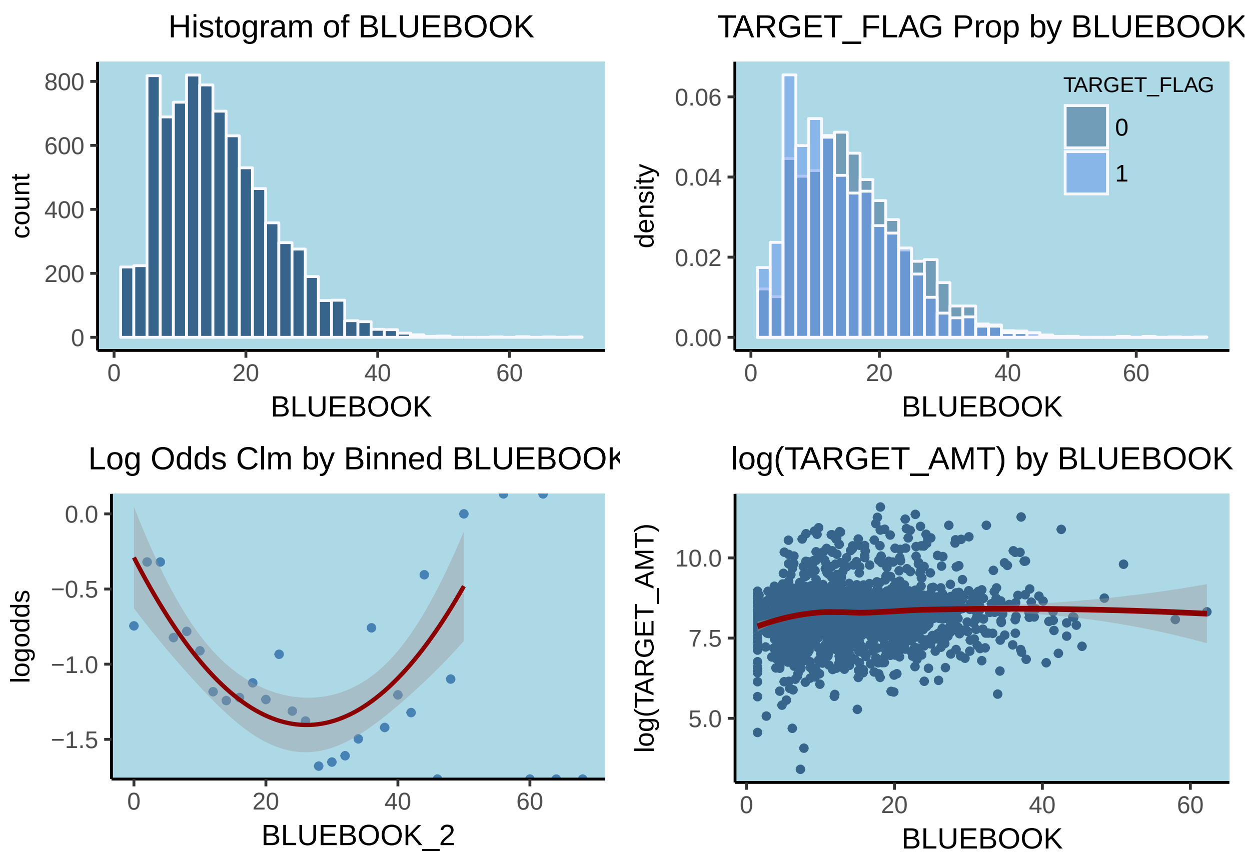
Here is a statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	StdD	Skew	Kurt
1.50	9.28	14.44	15.71	20.85	69.74	8.42	0.79	3.79

The distribution of `BLUEBOOK` is right skewed, like the continuous variables reviewed earlier.

Individuals involved in crashes have a higher proportion of low `BLUEBOOK` values. The loess curve in the lower left scatterplot indicates a decreasing log odds of claim occurrence with increasing incomes. However, there is a prominent bend in the curve around $30k. Perhaps this upward bend indicates risky driving behavior by owners of luxury and sports cars.

Finally, the relationship between BLUEBOOK and log `TARGET_AMT` reveals an increase in payouts with the value of the auto, but the curve flattens out around $20k - $30k.



BLUEBKBIN	ct	mean_cost	median_cost
0	171	3680	3423
5	614	5131	4196
10	504	5463	4051
15	373	6148	4164
20	267	7177	4256

BLUEBKBIN	ct	mean_cost	median_cost
25	123	6224	4369
30	57	6180	4321
35	26	11346	5308
40	13	6798	3041
45	2	3846	3846
50	1	18084	18084
55	1	3231	3231
60	1	4104	4104

OLDCLAIM

Below is table of values of the `OLDCLAIM` variable, with values bucketed into $4K increments.

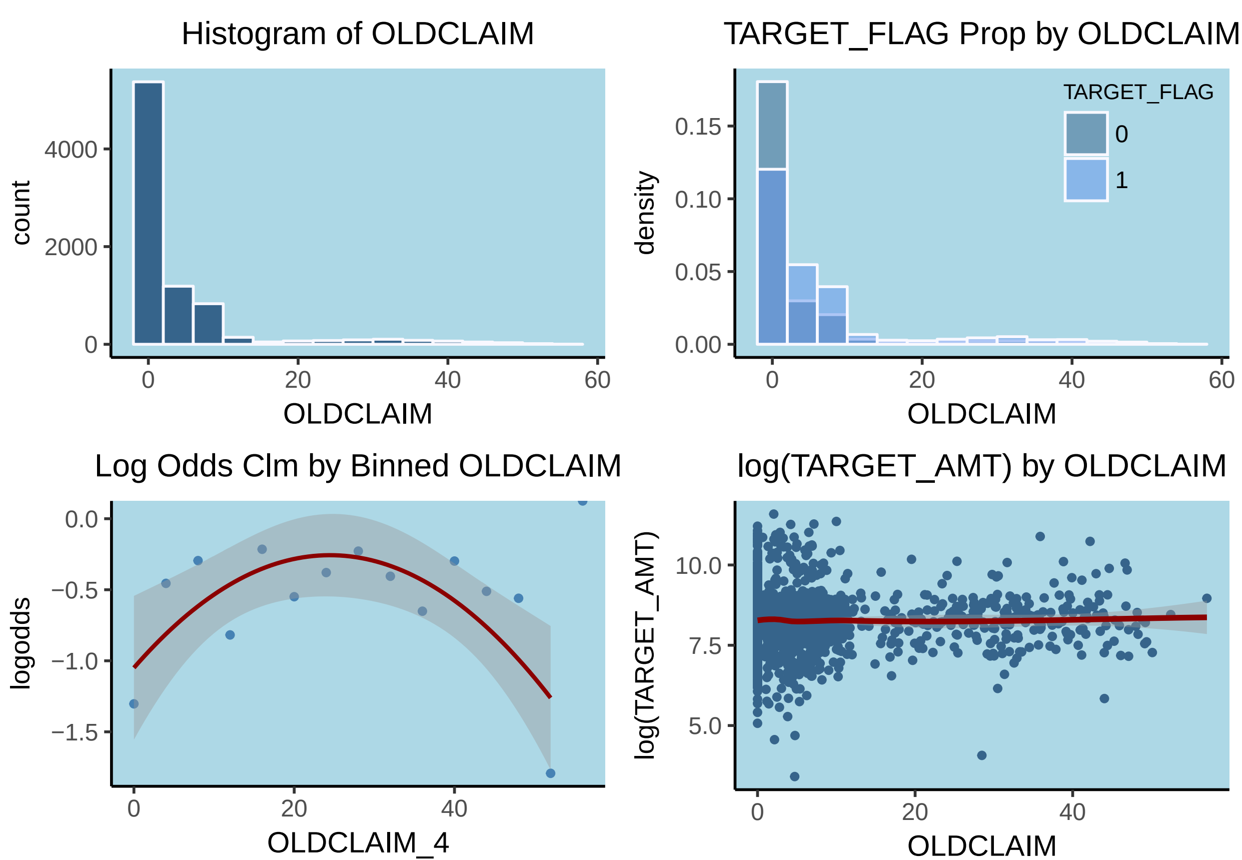
```
           0  Sum
count   8161 8161
percent  100  100
```

Below is a statistical summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
   0.00    0.00    0.00    4.04    4.64   57.04    8.78    3.12   12.86
```

The distribution of `OLDCLAIM` is extremely right skewed.

There does not appear to be a clear relationship between `OLDCLAIM` and log of `TARGET_CLM`. High previous claim amounts seem to be positively associated with the log odds of a future claim. However, this relationship appears to take a quadratic form.



OLD_BIN	ct	mean_cost	median_cost
0	1395	5814	4148
5	452	5409	4025
10	62	6347	4186
15	28	4712	3318
20	33	4865	3764
25	49	5060	4460
30	53	4898	4262
35	33	6875	3779
40	29	6922	4477
45	16	5350	3299
50	2	3070	3070
55	1	7803	7803

Data Preparation

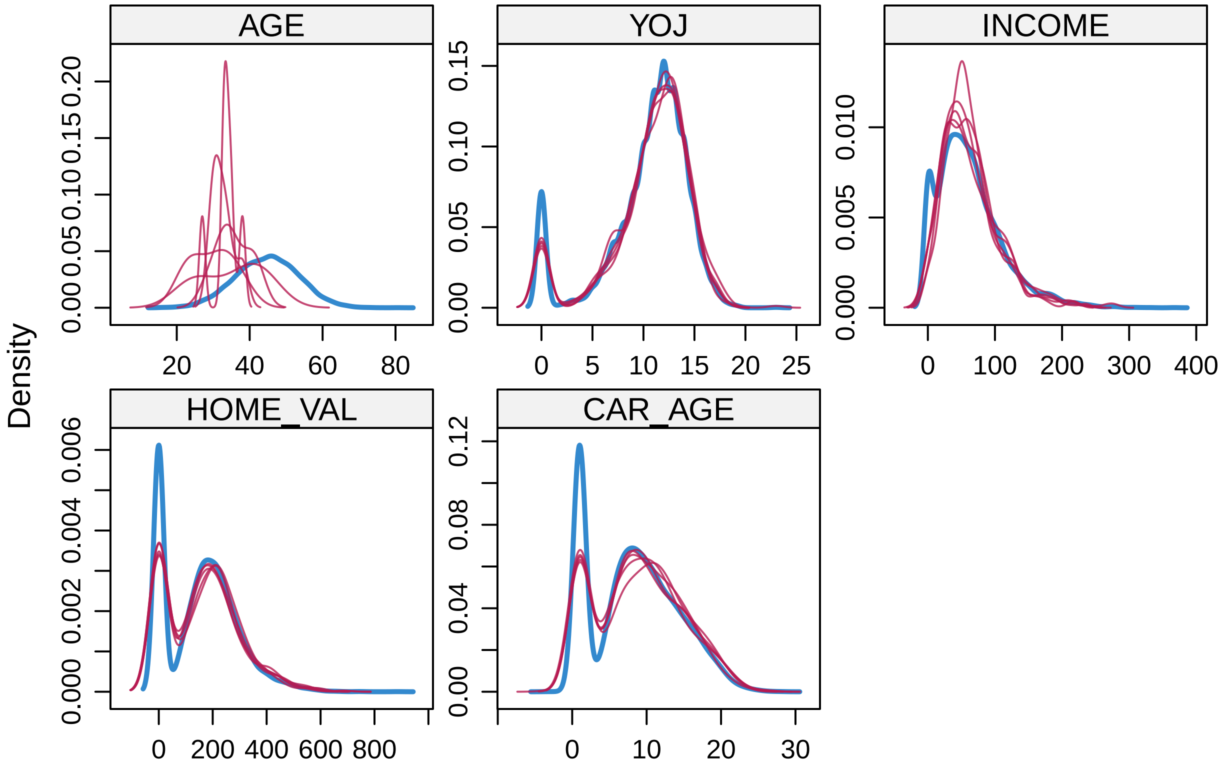
Missing Values

The following five variables have missing variables:

- AGE: 6 missing values (0.07%)
- YOJ: 454 missing values (5.6%)
- CAR_AGE: 510 (6.2%)
- INCOME: 445 (5.4%)
- HOME_VAL: 464 (5.7%)

Because four of the predictors are missing a significant number of records (5%-6%), we will attempt to use a sophisticated imputation procedure from R's MICE package to fill in the missing values. Our goal in using this procedure is to minimize the introduction of bias in our data vis-a-vis simpler methods like mean substitution.

One way to assess the quality of the imputation procedure used in MICE is to compare the distribution of the imputed data to the distribution of the non-missing values. Let's do that now by reviewing density plots:



With the exception of the `AGE` variable, the imputed variables seem to reasonably approximate the original distribution.

we're only missing a handful of values for `AGE`; so we will apply simply mean imputation for this variable.

Finally, we'll assume the -3 value for `CAR_AGE` is actually zero.

Data Transformation

TARGET_FLAG

No changes.

TARGET_AMT

The response variable `TARGET_AMT` is highly right skewed. We will use the box-cox procedure to suggest a reasonable transformation when the `TARGET_AMT` variable is positive:

```
Fitted parameters:
     lambda        beta     sigmasq
0.003172095 8.386645882 0.695511168

Convergence code returned by optim: 0
```

Given this output, We will apply the log transformation.

KIDSDRIV

Given the possible curvature between the log odds and `KIDSDRIV`, we're going to introduce a centered version of this variable to reduce possible collinearity issues.

HOMEKIDS

We will create another centered version of this variable due to a possible quadratic relationship with the response variable, `TARGET_FLAG.` Again, our intent with this transformation is avoid multicollinearity issues.

CLM_FREQ

Once again, we'll center the variable due concerns described earlier.

MVR_PTS

No changes.

AGE

Age appears to have a quadratic relationship with both `TARGET_FLAG` and `TARGET_AMT`. We'll center the variable.

YOJ

The variable `YOJ` has a significant number of zero observations. We'll create a mean centered variable, `YOJ_MOD`, that will be used for years on job of 1 or higher.

TRAVTIME
Well apply the mean-center transformation.

TIF

We'll mean-center the variable to reduce multicollinearity issues if we implement a quadratic term.

CAR_AGE

No changes.

PARENT1

No changes.

MSTATUS

No changes.

SEX

No changes.

CARUSE

No changes.

RED_CAR

No changes.

REVOKED

No changes.

URBANICITY

No changes.

JOB

No changes.

CAR_TYPE

No changes.

EDUCATION

No changes.

INCOME

Income is a positively skewed variable with a significant number zeroes.

We will apply the square root transformation suggested by the box-cox procedure to the original variable to reduce the overall skew.

```
Fitted parameters:
    lambda       beta    sigmasq
 0.4276529 10.9653802 17.4267860

Convergence code returned by optim: 0
```

HOME_VAL

Home values are also moderately right skewed with a significant number of zeroes.
We'll apply a quarter root transformation to the original variable to reduce the overall skew.

```
Fitted parameters:
   lambda      beta   sigmasq
0.1984401 9.4515199 1.5741664

Convergence code returned by optim: 0
```

BLUEBOOK

The `BLUEBOOK` variable has a moderate right skew. We'll apply the square root transformation suggested by the box-cox procedure.

```
Fitted parameters:
   lambda      beta   sigmasq
0.4610754 5.2624962 3.7967126

Convergence code returned by optim: 0
```

OLDCLAIM

`OLDCLAIM` is extremely right skewed. We'll apply a log(x+1) transformation to reduce the overall skew.

```
Fitted parameters:
     lambda        beta     sigmasq
-0.04511237  1.76185840  0.82136055

Convergence code returned by optim: 0
```

Build Models

Binary Logistic Regression

Model 1: Manual Variable Selection, Linear Terms Only

Based on our data exploration, we believe the following variables could be relevant in predicted whether or not a claim occurs:

- `KIDSDRIV`
- `HOMEKIDS`
- `CLM_FREQ`

- MVR_PTS
- AGE
- YOJ
- TRAVTIME
- TIF
- CAR_AGE
- PARENT1
- MSTATUS
- CAR_USE
- REVOKED
- URBANICITY
- JOB
- CAR_TYPE
- EDUCATION
- INCOME
- HOMEVAL
- BLUEBOOK
- OLDCLAIM

Granted, we haven't narrowed down the list much.

Before moving forward, let's calculate variance inflation factors, including all potential predictors in our model.

```
                  GVIF Df GVIF^(1/(2*Df))
PARENT1       1.925904  1        1.387769
MSTATUS       2.270579  1        1.506844
EDUCATION     9.032168  3        1.443107
JOB          29.006656  8        1.234258
CAR_USE       2.228753  1        1.492901
CAR_TYPE      2.446159  5        1.093575
REVOKED       1.147823  1        1.071365
MVR_PTS       1.195296  1        1.093296
CAR_AGE       2.145459  1        1.464739
URBANICITY    1.147586  1        1.071254
KIDSDRIV_MOD  1.347809  1        1.160952
HOMEKIDS_MOD  2.177248  1        1.475550
CLM_FREQ_MOD  2.350962  1        1.533285
```

```
AGE_MOD        1.458599  1        1.207725
YOJ_MOD        1.602499  1        1.265898
TRAVTIME_MOD   1.039607  1        1.019611
TIF_MOD        1.009716  1        1.004846
INCOME_MOD     2.929617  1        1.711612
HOME_VAL_MOD   1.945647  1        1.394864
BLUEBOOK_MOD   1.678511  1        1.295574
OLD_CLAIM_MOD  2.535045  1        1.592183
```

Seeing no major VIF issues–once accounting for degrees of freedom–we will proceed with the model with all suggested predictors.

```
Call:
glm(formula = TARGET_FLAG ~ . - TARGET_AMT - TARGET_AMT_MOD -
    INDEX - KIDSDRIV - HOMEKIDS - CLM_FREQ - AGE - YOJ - TRAVTIME -
    TIF - SEX - RED_CAR - INCOME - HOME_VAL - BLUEBOOK - OLDCLAIM -
    INCOME_30 - HOME_VAL_30 - BLUEBOOK_2 - OLDCLAIM_4, family = "binomial",
    data = auto)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6224  -0.7111  -0.3986   0.6144   3.1620

Coefficients:
               Estimate Std. Error z value             Pr(>|z|)
(Intercept)   -1.713883   0.288210  -5.947  0.00000000273704148 ***
PARENT12       0.380188   0.109715   3.465              0.00053 ***
MSTATUS2      -0.434783   0.087997  -4.941  0.00000077776605299 ***
EDUCATION2    -0.368349   0.088256  -4.174  0.00002997841876851 ***
EDUCATION3    -0.280964   0.160097  -1.755              0.07926 .
EDUCATION4    -0.190026   0.194621  -0.976              0.32887
JOB2           0.317719   0.184819   1.719              0.08560 .
JOB3           0.376967   0.196322   1.920              0.05484 .
JOB4          -0.407352   0.266000  -1.531              0.12567
JOB5           0.042266   0.216159   0.196              0.84498
JOB6           0.132768   0.168627   0.787              0.43108
JOB7          -0.541645   0.170692  -3.173              0.00151 **
JOB8           0.184347   0.177794   1.037              0.29980
JOB9          -0.109212   0.224207  -0.487              0.62618
CAR_USE2      -0.774937   0.087596  -8.847 < 0.0000000000000002 ***
CAR_TYPE2      0.587720   0.146458   4.013  0.00005997836389186 ***
CAR_TYPE3      0.548806   0.100235   5.475  0.00000004370229128 ***
CAR_TYPE4      0.949323   0.108211   8.773 < 0.0000000000000002 ***
CAR_TYPE5      0.716043   0.086037   8.323 < 0.0000000000000002 ***
CAR_TYPE6      0.657848   0.121632   5.409  0.00000006355203919 ***
REVOKED2       0.752452   0.085830   8.767 < 0.0000000000000002 ***
MVR_PTS        0.109599   0.013851   7.913  0.00000000000000252 ***
CAR_AGE       -0.001111   0.007524  -0.148              0.88260
URBANICITY2    2.410890   0.113255  21.287 < 0.0000000000000002 ***
KIDSDRIV_MOD   0.394609   0.061255   6.442  0.00000000011788292 ***
HOMEKIDS_MOD   0.039860   0.037188   1.072              0.28378
CLM_FREQ_MOD   0.173249   0.036264   4.777  0.00000177584154013 ***
AGE_MOD       -0.001762   0.004012  -0.439              0.66044
YOJ_MOD        0.001000   0.008772   0.114              0.90921
```

```
TRAVTIME_MOD   0.014748   0.001885   7.824  0.00000000000000512 ***
TIF_MOD       -0.055066   0.007353  -7.489  0.00000000000006923 ***
INCOME_MOD    -0.085729   0.015072  -5.688  0.00000001286803408 ***
HOME_VAL_MOD  -0.105028   0.022529  -4.662  0.00000313303785010 ***
BLUEBOOK_MOD  -0.185490   0.035090  -5.286  0.00000012490164003 ***
OLD_CLAIM_MOD -0.032814   0.039117  -0.839              0.40153
---
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: 7281.6  on 8126  degrees of freedom
AIC: 7351.6

Number of Fisher Scoring iterations: 5
```

The signs of the coefficients mostly make sense:

- We expect claim probabilities to be higher for single parents.
- Married individuals should be less prone to accidents compared to singles.
- The signs of all the `EDUCATION` levels make sense; however, we expected the Master's and PhD levels to show a greater reduction in log odds relative to the high school reference level. We are not too concerned with this result though, as other variables such as `JOB_TYPE` and `INCOME` help in defining an individual's composite risk profile.
- We didn't have solid a priori expectations about the relationship between different job type. We are not surprised with most of these result though. For instance, we are not surprised by the increased risk associated with blue collar workers relative to doctors.
- The model is consistent with our expectation that private vehicles are less accident prone compared to commercial vehicles.
- The car type signs and magnitudes are fairly consistent with our expectations: Minivans(the reference level) should be safer than the other vehicles. Sports cars should be most likely to be involved in an accident. This is what we see.
- A revoked license is highly indicative of future accident risk.
- MVR points are positively associated with accident risk.
- The model shows a slight negative relationship between car age and log odds of an accident, but the result is not statistically significant.
- Our model shows much greater risk in urban areas compared to rural geographies. This is consistent with our expectation.
- The number of teenage drivers impacts risk unfavorably, as our model shows.
- Our model indicates that more children can adversely impact claims risk. We didn't have firm expectations about this variable's influence. More important, our model does not indicate a statistically significant result.

- Our model reveals claim risk declining with age. This is not necessarily surprising; however, the model coefficient is not statistically significant.
- Our model shows risk increasing slightly with increases to `YOB`. This is maybe a strange result, but our model indicates a very high p-value for our coefficient.
- We expect risk to increase with longer travel times, as our model shows.
- We expect loyal policyholders to be less risky than frequently churning policyholders, as our model indicates.
- Our model indicates decreasing log odds with increases to home value. This is what we would expect.
- We are not surprised to see cars with high `BLUEBOOK` values being associated with reduced risk in our model compared to lower values.
- The coefficient for prior claims cost is opposite of what we would expect. However, the coefficient has a high associated p-value; so we may drop this predictor from our model.

Let's clean up our model by removing some of the statistically insignificant predictors. We will leave all `JOB` levels in our model, as the "Manager" and "Clerical" coefficients are highly significant, and two additional levels have p-values below 10%.

```
Call:
glm(formula = TARGET_FLAG ~ PARENT1 + MSTATUS + EDUCATION + JOB +
    CAR_USE + CAR_TYPE + REVOKED + MVR_PTS + URBANICITY + KIDSDRIV_MOD +
    CLM_FREQ_MOD + TRAVTIME_MOD + TIF_MOD + INCOME_MOD + HOME_VAL_MOD +
    BLUEBOOK_MOD, family = "binomial", data = auto)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6399  -0.7118  -0.3993   0.6121   3.1482

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)  -1.757833   0.282629  -6.220  0.00000000049847552 ***
PARENT12      0.459073   0.094539   4.856  0.00000119837335651 ***
MSTATUS2     -0.402221   0.084254  -4.774  0.00000180697301345 ***
EDUCATION2   -0.375761   0.081055  -4.636  0.00000355440716343 ***
EDUCATION3   -0.300632   0.141531  -2.124              0.03366 *
EDUCATION4   -0.211068   0.179649  -1.175              0.24004
JOB2          0.321791   0.184707   1.742              0.08148 .
JOB3          0.386512   0.196036   1.972              0.04865 *
JOB4         -0.415117   0.265500  -1.564              0.11793
JOB5          0.040669   0.213901   0.190              0.84921
JOB6          0.125024   0.168259   0.743              0.45745
JOB7         -0.550785   0.170509  -3.230              0.00124 **
JOB8          0.179572   0.177647   1.011              0.31209
JOB9         -0.093451   0.222713  -0.420              0.67478
CAR_USE2     -0.771386   0.087427  -8.823 < 0.0000000000000002 ***
CAR_TYPE2     0.591698   0.146233   4.046  0.00005203971882180 ***
CAR_TYPE3     0.548603   0.100178   5.476  0.00000004344038477 ***
CAR_TYPE4     0.945975   0.107917   8.766 < 0.0000000000000002 ***
CAR_TYPE5     0.716538   0.085960   8.336 < 0.0000000000000002 ***
```

```
CAR_TYPE6     0.658220   0.121499   5.418  0.00000006043560923 ***
REVOKED2      0.730901   0.080424   9.088 < 0.0000000000000002 ***
MVR_PTS       0.107957   0.013581   7.949  0.00000000000000188 ***
URBANICITY2   2.408645   0.113142  21.289 < 0.0000000000000002 ***
KIDSDRIV_MOD  0.423779   0.055142   7.685  0.00000000000001527 ***
CLM_FREQ_MOD  0.151294   0.025533   5.926  0.00000000311329229 ***
TRAVTIME_MOD  0.014705   0.001884   7.806  0.00000000000000589 ***
TIF_MOD      -0.054813   0.007347  -7.461  0.00000000000008618 ***
INCOME_MOD   -0.084636   0.014524  -5.827  0.00000000563328460 ***
HOME_VAL_MOD -0.106391   0.022498  -4.729  0.00000225772041669 ***
BLUEBOOK_MOD -0.188871   0.034886  -5.414  0.00000006162739580 ***
---
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: 7284.3  on 8131  degrees of freedom
AIC: 7344.3

Number of Fisher Scoring iterations: 5
```

All predictors are now statistically significant, with the exception of several of the job types. The coefficients in our model also still make sense.

Model 2: Add Quadratic Terms to Model 1

We noted earlier that there appeared to be quadratic relationship between some predictors and log odds.

Starting with our original Model 1–i.e. before we removed the insignificant predictors–we'll add second order polynomial terms for the following variables:

- `KIDSDRIV`
- `HOMEKIDS`
- `CLM_FREQ`
- `AGE`
- `YOJ`
- `TIF`

Here is the summary output from model 2:

```
Call:
glm(formula = TARGET_FLAG ~ PARENT1 + MSTATUS + EDUCATION + JOB +
    CAR_USE + CAR_TYPE + REVOKED + MVR_PTS + CAR_AGE + URBANICITY +
    KIDSDRIV_MOD + HOMEKIDS_MOD + CLM_FREQ_MOD + AGE_MOD + YOJ_MOD +
    TRAVTIME_MOD + TIF_MOD + INCOME_MOD + HOME_VAL_MOD + BLUEBOOK_MOD +
    OLD_CLAIM_MOD + I(KIDSDRIV_MOD^2) + I(HOMEKIDS_MOD^2) + I(CLM_FREQ_MOD^2) +
    I(AGE_MOD^2) + I(YOJ_MOD^2) + I(TIF_MOD^2), family = "binomial",
    data = auto)

Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-2.4272  -0.7068  -0.3906   0.5963   3.0532

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -1.6776996  0.3057439  -5.487 0.000000040819180673 ***
PARENT12           0.2388555  0.1190667   2.006              0.04485 *
MSTATUS2          -0.5167106  0.0910712  -5.674 0.000000013974455379 ***
EDUCATION2        -0.3909598  0.0891435  -4.386 0.000011559497756939 ***
EDUCATION3        -0.2957904  0.1615164  -1.831              0.06705 .
EDUCATION4        -0.2352758  0.1961787  -1.199              0.23041
JOB2               0.3466982  0.1856785   1.867              0.06187 .
JOB3               0.3975150  0.1973076   2.015              0.04394 *
JOB4              -0.4316015  0.2664914  -1.620              0.10532
JOB5               0.0542765  0.2193294   0.247              0.80455
JOB6               0.1352339  0.1692727   0.799              0.42434
JOB7              -0.5374406  0.1709707  -3.143              0.00167 **
JOB8               0.1947748  0.1784243   1.092              0.27499
JOB9              -0.1341038  0.2266228  -0.592              0.55402
CAR_USE2          -0.7793452  0.0882281  -8.833 < 0.0000000000000002 ***
CAR_TYPE2          0.6295135  0.1472723   4.274 0.000019157888812068 ***
CAR_TYPE3          0.5667391  0.1010379   5.609 0.000000020329346729 ***
CAR_TYPE4          0.8874125  0.1094808   8.106 0.000000000000000525 ***
CAR_TYPE5          0.7089577  0.0868127   8.167 0.000000000000000317 ***
CAR_TYPE6          0.6862665  0.1226279   5.596 0.000000021893171964 ***
REVOKED2           0.8321322  0.0887887   9.372 < 0.0000000000000002 ***
MVR_PTS            0.0962345  0.0141337   6.809 0.000000000009838070 ***
CAR_AGE           -0.0016633  0.0075782  -0.219              0.82627
URBANICITY2        2.3948478  0.1136789  21.067 < 0.0000000000000002 ***
KIDSDRIV_MOD       0.7322830  0.1425802   5.136 0.000000280740271759 ***
HOMEKIDS_MOD       0.0511923  0.0685622   0.747              0.45527
CLM_FREQ_MOD       0.3727049  0.0738148   5.049 0.000000443696934759 ***
AGE_MOD           -0.0026885  0.0041431  -0.649              0.51639
YOJ_MOD            0.0059588  0.0110425   0.540              0.58945
TRAVTIME_MOD       0.0148574  0.0018986   7.825 0.000000000000005066 ***
TIF_MOD           -0.0662275  0.0086328  -7.672 0.000000000000016985 ***
INCOME_MOD        -0.0733066  0.0154936  -4.731 0.000002229584044502 ***
HOME_VAL_MOD      -0.1022630  0.0226822  -4.509 0.000006528512402561 ***
BLUEBOOK_MOD      -0.1932071  0.0353434  -5.467 0.000000045883229928 ***
OLD_CLAIM_MOD     -0.1258748  0.0493198  -2.552              0.01070 *
I(KIDSDRIV_MOD^2) -0.1383717  0.0690468  -2.004              0.04507 *
I(HOMEKIDS_MOD^2) -0.0335400  0.0281639  -1.191              0.23370
I(CLM_FREQ_MOD^2) -0.0888912  0.0282768  -3.144              0.00167 **
I(AGE_MOD^2)       0.0022399  0.0002811   7.968 0.000000000000001611 ***
I(YOJ_MOD^2)       0.0020479  0.0015960   1.283              0.19944
I(TIF_MOD^2)       0.0031482  0.0013917   2.262              0.02370 *
---
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: 7195.8  on 8120  degrees of freedom
AIC: 7277.8
```

```
Number of Fisher Scoring iterations: 5
```

Below our key results from this second model:

- The first and second order terms for `KIDSDRIV` are both statistically significant. The negative sign of the second order term makes sense: the unfavorable impact of adding additional children drivers diminishes with each subsequent child.

- Neither the first nor second order terms are significant for `HOMEKIDS`.

- Both `CLM_FREQ` terms are significant. The negative second order term indicates diminishing impact of each additional prior claim.

- The second order term of `AGE` is statistically significant, but the first first order term is not. We'll leave both terms in our model. The coefficients of the terms make sense. There is a primary trend of reduced risk with age–see the negative first order coefficient, but the trend diminishes and potentially reverses for higher ages–as reflected in the positive second order term.

- Neither `YOJ` terms are statistically significant.

- Both first and second order `TIF` terms are significant. The signs of the coefficients have an intuitive explanation: there is a primary effect of risk reduction in risk with increases to `TIF` but the favorable impact diminishes with higher `TIF` values.

- `CAR_AGE` is still insignificant in this model.

Based on the results above, we'll remove all `HOMEKIDS` and `YOJ` terms from our model. Here are the summary results from this modified, second model:

```
Call:
glm(formula = TARGET_FLAG ~ PARENT1 + MSTATUS + EDUCATION + JOB +
    CAR_USE + CAR_TYPE + REVOKED + MVR_PTS + URBANICITY + KIDSDRIV_MOD +
    CLM_FREQ_MOD + AGE_MOD + TRAVTIME_MOD + TIF_MOD + INCOME_MOD +
    HOME_VAL_MOD + BLUEBOOK_MOD + OLD_CLAIM_MOD + I(KIDSDRIV_MOD^2) +
    I(CLM_FREQ_MOD^2) + I(AGE_MOD^2) + I(TIF_MOD^2), family = "binomial",
    data = auto)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4257  -0.7041  -0.3926   0.5955   3.0433

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -1.6759581  0.2954057  -5.673 0.000000013998177240 ***
PARENT12           0.2777259  0.1018441   2.727              0.00639 **
MSTATUS2          -0.4909029  0.0858650  -5.717 0.000000010832584791 ***
EDUCATION2        -0.3919889  0.0818148  -4.791 0.000001658104571861 ***
EDUCATION3        -0.3010222  0.1425353  -2.112              0.03469 *
EDUCATION4        -0.2372596  0.1810645  -1.310              0.19007
JOB2               0.3405307  0.1855451   1.835              0.06646 .
JOB3               0.3855461  0.1969537   1.958              0.05028 .
JOB4              -0.4299822  0.2667154  -1.612              0.10693
JOB5               0.0928347  0.2152047   0.431              0.66619
JOB6               0.1305186  0.1692840   0.771              0.44070
```

```
JOB7              -0.5401499  0.1709242  -3.160              0.00158 **
JOB8               0.1923674  0.1783272   1.079              0.28071
JOB9              -0.1062107  0.2238306  -0.475              0.63513
CAR_USE2          -0.7809147  0.0880529  -8.869 < 0.0000000000000002 ***
CAR_TYPE2          0.6263114  0.1472346   4.254 0.000021014260357950 ***
CAR_TYPE3          0.5627747  0.1009482   5.575 0.000000024769216593 ***
CAR_TYPE4          0.8878065  0.1093148   8.122 0.000000000000000460 ***
CAR_TYPE5          0.7074706  0.0867202   8.158 0.000000000000000340 ***
CAR_TYPE6          0.6832450  0.1225326   5.576 0.000000024607726121 ***
REVOKED2           0.8285955  0.0887177   9.340 < 0.0000000000000002 ***
MVR_PTS            0.0972327  0.0141100   6.891 0.000000000005538170 ***
URBANICITY2        2.3915266  0.1135807  21.056 < 0.0000000000000002 ***
KIDSDRIV_MOD       0.7844472  0.1289314   6.084 0.000000001170576243 ***
CLM_FREQ_MOD       0.3686980  0.0737462   5.000 0.000000574642059029 ***
AGE_MOD           -0.0023818  0.0036156  -0.659              0.51006
TRAVTIME_MOD       0.0148220  0.0018979   7.810 0.000000000000005729 ***
TIF_MOD           -0.0662577  0.0086283  -7.679 0.000000000000016025 ***
INCOME_MOD        -0.0784599  0.0146356  -5.361 0.000000082807106707 ***
HOME_VAL_MOD      -0.1023873  0.0226702  -4.516 0.000006290735752587 ***
BLUEBOOK_MOD      -0.1936529  0.0353362  -5.480 0.000000042461339546 ***
OLD_CLAIM_MOD     -0.1237745  0.0492927  -2.511              0.01204 *
I(KIDSDRIV_MOD^2) -0.1634110  0.0651681  -2.508              0.01216 *
I(CLM_FREQ_MOD^2) -0.0870421  0.0282286  -3.083              0.00205 **
I(AGE_MOD^2)       0.0022712  0.0002767   8.207 0.000000000000000226 ***
I(TIF_MOD^2)       0.0031561  0.0013905   2.270              0.02322 *
---
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: 7199.1  on 8125  degrees of freedom
AIC: 7271.1

Number of Fisher Scoring iterations: 5
```

Model 3: Stepwise Regression

For our third model, we'll implement stepwise regression, with variable selecting occurring in both directions. We'll include all predictors (transformed versions where applicable) in our potential universe of candidates.

For simplicity, we'll only include first order terms.

```
Call:
glm(formula = TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + HOME_VAL_MOD +
    CAR_TYPE + REVOKED + PARENT1 + CAR_USE + TRAVTIME_MOD + INCOME_MOD +
    TIF_MOD + KIDSDRIV_MOD + CLM_FREQ_MOD + BLUEBOOK_MOD + MSTATUS +
    EDUCATION, family = "binomial", data = auto_redux)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6399  -0.7118  -0.3993   0.6121   3.1482

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
```

```
(Intercept)  -1.757833   0.282629  -6.220  0.00000000049847552 ***
URBANICITY2   2.408645   0.113142  21.289 < 0.0000000000000002 ***
JOB2          0.321791   0.184707   1.742              0.08148 .
JOB3          0.386512   0.196036   1.972              0.04865 *
JOB4         -0.415117   0.265500  -1.564              0.11793
JOB5          0.040669   0.213901   0.190              0.84921
JOB6          0.125024   0.168259   0.743              0.45745
JOB7         -0.550785   0.170509  -3.230              0.00124 **
JOB8          0.179572   0.177647   1.011              0.31209
JOB9         -0.093451   0.222713  -0.420              0.67478
MVR_PTS       0.107957   0.013581   7.949  0.00000000000000188 ***
HOME_VAL_MOD -0.106391   0.022498  -4.729  0.00000225772041669 ***
CAR_TYPE2     0.591698   0.146233   4.046  0.00005203971882181 ***
CAR_TYPE3     0.548603   0.100178   5.476  0.00000004344038477 ***
CAR_TYPE4     0.945975   0.107917   8.766 < 0.0000000000000002 ***
CAR_TYPE5     0.716538   0.085960   8.336 < 0.0000000000000002 ***
CAR_TYPE6     0.658220   0.121499   5.418  0.00000006043560923 ***
REVOKED2      0.730901   0.080424   9.088 < 0.0000000000000002 ***
PARENT12      0.459073   0.094539   4.856  0.00000119837335651 ***
CAR_USE2     -0.771386   0.087427  -8.823 < 0.0000000000000002 ***
TRAVTIME_MOD  0.014705   0.001884   7.806  0.00000000000000589 ***
INCOME_MOD   -0.084636   0.014524  -5.827  0.00000000563328460 ***
TIF_MOD      -0.054813   0.007347  -7.461  0.00000000000008618 ***
KIDSDRIV_MOD  0.423779   0.055142   7.685  0.00000000000001527 ***
CLM_FREQ_MOD  0.151294   0.025533   5.926  0.00000000311329229 ***
BLUEBOOK_MOD -0.188871   0.034886  -5.414  0.00000006162739580 ***
MSTATUS2     -0.402221   0.084254  -4.774  0.00000180697301345 ***
EDUCATION2   -0.375761   0.081055  -4.636  0.00000355440716343 ***
EDUCATION3   -0.300632   0.141531  -2.124              0.03366 *
EDUCATION4   -0.211068   0.179649  -1.175              0.24004
---
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: 7284.3  on 8131  degrees of freedom
AIC: 7344.3

Number of Fisher Scoring iterations: 5
```

```
 [1] "PARENT1"      "MSTATUS"      "SEX"          "EDUCATION"    "JOB"          "CAR_USE"
 [7] "CAR_TYPE"     "RED_CAR"      "REVOKED"      "MVR_PTS"      "CAR_AGE"      "URBANICITY"
[13] "KIDSDRIV_MOD" "HOMEKIDS_MOD" "CLM_FREQ_MOD" "AGE_MOD"      "YOJ_MOD"      "TRAVTIME_MOD"
[19] "TIF_MOD"      "INCOME_MOD"   "HOME_VAL_MOD" "BLUEBOOK_MOD" "OLD_CLAIM_MOD" "TARGET_FLAG"
```

Surprisingly, the stepwise procedure produced a model that is identical to our modified Model 1!

Multiple Linear Regression

Now We'll model claim costs using the subset of the training data where claim costs were greater than one.

Before we build our models, let's check for multicollinearity uses by reviewing variance inflation factors for a linear model that includes all predictors.

```
                   GVIF Df GVIF^(1/(2*Df))
```

```
PARENT1        2.161870  1        1.470330
MSTATUS        2.424138  1        1.556964
SEX            3.759011  1        1.938817
EDUCATION      9.422433  3        1.453318
JOB           37.076989  8        1.253340
CAR_USE        2.257139  1        1.502378
CAR_TYPE       6.333798  5        1.202725
RED_CAR        1.833205  1        1.353959
REVOKED        1.266221  1        1.125265
MVR_PTS        1.162829  1        1.078345
CAR_AGE        2.117086  1        1.455021
URBANICITY     1.052890  1        1.026104
KIDSDRIV_MOD   1.435451  1        1.198103
HOMEKIDS_MOD   2.245895  1        1.498631
CLM_FREQ_MOD   2.233772  1        1.494581
AGE_MOD        1.515101  1        1.230894
YOJ_MOD        1.914315  1        1.383588
TRAVTIME_MOD   1.030286  1        1.015030
TIF_MOD        1.017536  1        1.008730
INCOME_MOD     3.484163  1        1.866591
HOME_VAL_MOD   1.993510  1        1.411917
BLUEBOOK_MOD   2.070495  1        1.438921
OLD_CLAIM_MOD  2.485622  1        1.576585
```

There does not appear to be an issue with mulicollinearity.

Model 4: Manual Variable Selection, Linear Terms Only

Choosing relevant predictors manually is a challenging exercise as most predictors seemed to have only a subtle influence–if any–on claim costs.

Based on our exploratory work, we believe the following variables may be relevant:

- `AGE`
- `KIDSDRIV`
- `HOMEKIDS`
- `TRAVTIME`
- `SEX`
- `CAR_USE`
- `RED_CAR`
- `UBANICITY`
- `JOB`
- `CARTYPE`
- `EDUCATION`
- `BLUEBOOK`

Let's look at a preliminary model using all 12 of our proposed predictors:

```
Call:
lm(formula = TARGET_AMT_MOD ~ KIDSDRIV_MOD + HOMEKIDS_MOD + AGE_MOD +
    TRAVTIME_MOD + SEX + CAR_USE + RED_CAR + URBANICITY + JOB +
    CAR_TYPE + EDUCATION + BLUEBOOK_MOD, data = auto_clm)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7978 -0.4031  0.0403  0.4041  3.2713

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)   7.7453078  0.1842369  42.040 < 0.0000000000000002 ***
KIDSDRIV_MOD -0.0359514  0.0331619  -1.084                0.278
HOMEKIDS_MOD  0.0202516  0.0190526   1.063                0.288
AGE_MOD       0.0003962  0.0021644   0.183                0.855
TRAVTIME_MOD -0.0005149  0.0011623  -0.443                0.658
SEX2          0.0942848  0.0678354   1.390                0.165
CAR_USE2     -0.0042440  0.0519132  -0.082                0.935
RED_CAR2      0.0270291  0.0521879   0.518                0.605
URBANICITY2   0.0274335  0.0789864   0.347                0.728
JOB2          0.0579112  0.1194467   0.485                0.628
JOB3          0.0768355  0.1242451   0.618                0.536
JOB4         -0.0420891  0.1844233  -0.228                0.819
JOB5          0.0186923  0.1219753   0.153                0.878
JOB6         -0.0039060  0.1076219  -0.036                0.971
JOB7          0.0143125  0.1117662   0.128                0.898
JOB8          0.0966539  0.1180372   0.819                0.413
JOB9          0.0707929  0.1262240   0.561                0.575
CAR_TYPE2    -0.0025530  0.0960169  -0.027                0.979
CAR_TYPE3     0.0298786  0.0622815   0.480                0.631
CAR_TYPE4     0.0722865  0.0781345   0.925                0.355
CAR_TYPE5     0.0922557  0.0689352   1.338                0.181
CAR_TYPE6    -0.0279082  0.0804533  -0.347                0.729
EDUCATION2   -0.0537016  0.0471957  -1.138                0.255
EDUCATION3    0.0881450  0.0939857   0.938                0.348
EDUCATION4    0.1428179  0.1147912   1.244                0.214
BLUEBOOK_MOD  0.0980000  0.0227053   4.316            0.0000166 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8092 on 2127 degrees of freedom
Multiple R-squared:  0.01979,   Adjusted R-squared:  0.008271
F-statistic: 1.718 on 25 and 2127 DF,  p-value: 0.01491
```

Only one of our predictors, `BLUEBOOK` appears to be significant in our model.

Here is our modified model, with `BLUEBOOK` as the sole predictor:

```
Call:
lm(formula = TARGET_AMT_MOD ~ BLUEBOOK_MOD, data = auto_clm)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-4.7842 -0.3929  0.0404  0.3952  3.2528

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)   7.95359    0.06005 132.445 < 0.0000000000000002 ***
BLUEBOOK_MOD  0.08921    0.01590   5.609          0.000000023 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8069 on 2151 degrees of freedom
Multiple R-squared:  0.01441,   Adjusted R-squared:  0.01396
F-statistic: 31.46 on 1 and 2151 DF,  p-value: 0.00000002298
```

The positive coefficient for `Bluebook` makes sense: we expect the replacement cost and/or repairs for a high-valued car to be more expensive than auto with a low replacement cost.

Model 5: Add Quadratic Terms to Model 4

In the exploratory section, we noted potential curved relationship between some predictors and the log of `TARGET_AMT`.

Those predictors were `AGE` and `TRAVTIME`. Let's include squared terms for these two predictors and also for for `BLUEBOOK`:

```
Call:
lm(formula = TARGET_AMT_MOD ~ BLUEBOOK_MOD + I(BLUEBOOK_MOD^2) +
    TRAVTIME_MOD + I(TRAVTIME_MOD^2) + AGE_MOD + I(AGE_MOD^2),
    data = auto_clm)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7676 -0.3878  0.0346  0.3986  3.2456

Coefficients:
                    Estimate  Std. Error t value             Pr(>|t|)
(Intercept)        7.56281355  0.15338179  49.307 < 0.0000000000000002 ***
BLUEBOOK_MOD       0.29957271  0.08325430   3.598             0.000328 ***
I(BLUEBOOK_MOD^2) -0.02796890  0.01094904  -2.554             0.010704 *
TRAVTIME_MOD      -0.00057023  0.00119024  -0.479             0.631926
I(TRAVTIME_MOD^2)  0.00001846  0.00005609   0.329             0.742148
AGE_MOD            0.00038530  0.00186400   0.207             0.836260
I(AGE_MOD^2)       0.00027981  0.00014693   1.904             0.057004 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8058 on 2146 degrees of freedom
Multiple R-squared:  0.0192,    Adjusted R-squared:  0.01646
F-statistic: 7.002 on 6 and 2146 DF,  p-value: 0.0000002163
```

The second order term for `BLUEBOOK` is statistically significant. The second order term for `AGE` is borderline significant; so we include leave both age variables in our model; but remove terms related to `TRAVTIME`.

```
Call:
lm(formula = TARGET_AMT_MOD ~ BLUEBOOK_MOD + I(BLUEBOOK_MOD^2) +
```

```
    AGE_MOD + I(AGE_MOD^2), data = auto_clm)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7718 -0.3864  0.0345  0.3947  3.2387 

Coefficients:
                    Estimate Std. Error t value            Pr(>|t|)    
(Intercept)        7.5676524  0.1528591  49.507 < 0.0000000000000002 ***
BLUEBOOK_MOD       0.2990652  0.0831884   3.595            0.000332 ***
I(BLUEBOOK_MOD^2) -0.0279518  0.0109400  -2.555            0.010687 *  
AGE_MOD            0.0003620  0.0018602   0.195            0.845717    
I(AGE_MOD^2)       0.0002827  0.0001467   1.927            0.054140 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8055 on 2148 degrees of freedom
Multiple R-squared:  0.01908,   Adjusted R-squared:  0.01725 
F-statistic: 10.44 on 4 and 2148 DF,  p-value: 0.00000002233
```

This model indicates that log costs increase quadratically with Age. This result seems possible.

Model 6: Stepwise Regression

Finally, we'll perform a basic stepwise regression with variable selection performed in both directions. For simplicity, we'll only include linear terms.

```
Call:
lm(formula = TARGET_AMT_MOD ~ BLUEBOOK_MOD + MSTATUS + MVR_PTS + 
    SEX + CLM_FREQ_MOD, data = auto_clm)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6968 -0.4038  0.0391  0.4093  3.2236 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept)   7.940083   0.066264 119.825 < 0.0000000000000002 ***
BLUEBOOK_MOD  0.086861   0.015942   5.449         0.0000000566 ***
MSTATUS2     -0.073598   0.034727  -2.119               0.0342 *  
MVR_PTS       0.017181   0.007053   2.436               0.0149 *  
SEX2          0.055410   0.035037   1.581               0.1139    
CLM_FREQ_MOD -0.022434   0.014567  -1.540               0.1237    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8051 on 2147 degrees of freedom
Multiple R-squared:  0.02064,   Adjusted R-squared:  0.01836 
F-statistic: 9.051 on 5 and 2147 DF,  p-value: 0.00000001583
```

The stepwise procedure included four additional predictors in addition to `BLUEBOOK`:
* `MSTATUS`: the negative coefficient indicates the married individuals are less expensive than singles. This seems reasonable.
* `MVR_PTS`: the model indicates a positive association between `MVR_PTS` and claim costs. This also seems reasonable.

* `SEX`: According to the model, males are more expensive than females. This is consistent with our earlier exploratory work.
* `CLM_FREQ_MOD1`: The coefficient is negative. This result is counterintuitive. We would expect folks with a high incidence accident rate to potentially be at risk for higher cost accidents. For now, we'll leave this predictor in, but we may want do additional analysis.

SELECT MODELS

Binary Logistic Regression Models

Let's compare model fits for all of our models:

```
            AIC     AICc      BIC    loglik
m1     7351.575 7351.885 7596.824 -3640.787
m1_mod 7344.333 7344.562 7554.547 -3642.167
m2     7277.850 7278.274 7565.142 -3597.925
m2_mod 7271.090 7271.417 7523.346 -3599.545
m3     7344.333 7344.562 7554.547 -3642.167
```

Based on the various model evaluation criteria, model m2_mod appears to be the clear winner. This model is the binary logistic regression model that included multiple quadratic terms, but removed statistically insignificant predictors from the original model 2 formulation.

Model 2 is superior in that the AIC, AIC_c, and BIC measures are lower than all other evaluated models. The log likelihood is not quite as the original model 2, but this measure does not account for model complexity as the other models do.

Let's also compare the AUC measure for all models:

```
[1] "Model 1: 0.815"
```

```
[1] "Model 1 mod: 0.814"
```

```
[1] "Model 2: 0.82"
```

```
[1] "Model 2 mod: 0.82"
```

```
[1] "Model 3: 0.814"
```

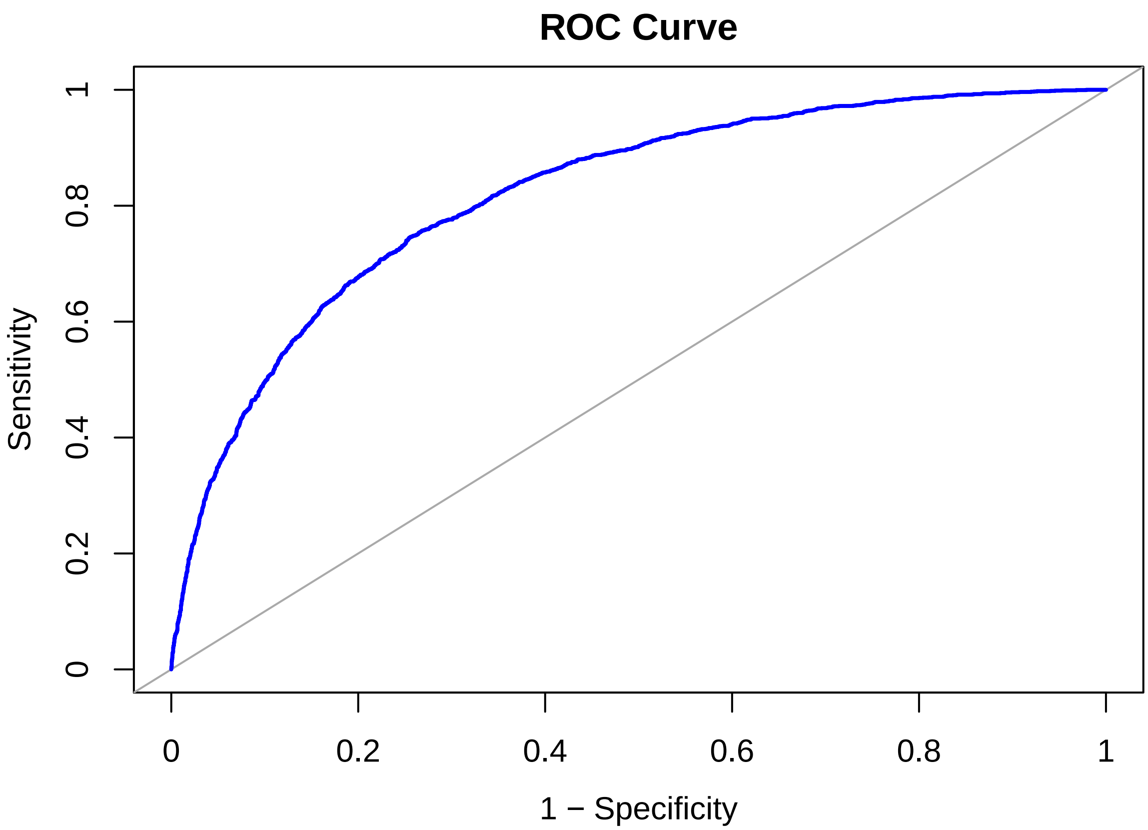
Models 2 and Model 2 mod are tied and have the highest AUC measures. Given that Model 2 mod has fewer parameters than Model 2, the AUC measure supports our contention that Model 2 mod is superior to the other models.

Let's explore a summary of our model predictions using the training data:

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002396 0.075235 0.195732 0.263816 0.401488 0.959960
```

We now need to choose an appropriate probability cutoff measure for predicting whether or not an individual will have a claim.

We will use Youden's index to determine this optimal cutoff.




```
[1] 0.2760209
```

Using Youden's index, we select a relatively low cutoff measure of 0.276. With such a low cutoff, we will inevitably sacrifice specificity for gains in sensitivity compared to a traditional 0.5 cutoff. However, this lower cutoff provides a better balance between precision and recall.

```
          Reference
Prediction    0    1
         0 4470  546
         1 1538 1607
```

```
$accuracy
[1] 0.7446391

$error_rt
[1] 0.2553609

$precision
[1] 0.5109698

$sensitivity
[1] 0.7464004

$specificity
[1] 0.744008

$F1
[1] 0.606644
```

For comparison purposes, here is the confusion matrix and related classification metrics with a 0.5 cutoff.

```
          Reference
Prediction    0    1
         0 5547 1212
         1  461  941

$accuracy
[1] 0.7950006

$error_rt
[1] 0.2049994

$precision
[1] 0.671184

$sensitivity
[1] 0.4370646

$specificity
[1] 0.923269

$F1
[1] 0.5293952
```

We see that our 0.276 cutoff also results in lower accuracy vis-a-vis the 0.5 threshold. But our lower threshold also results in better balance in precision vs. recall, as indicated by the improved F1 measure.

Multiple Regression Models

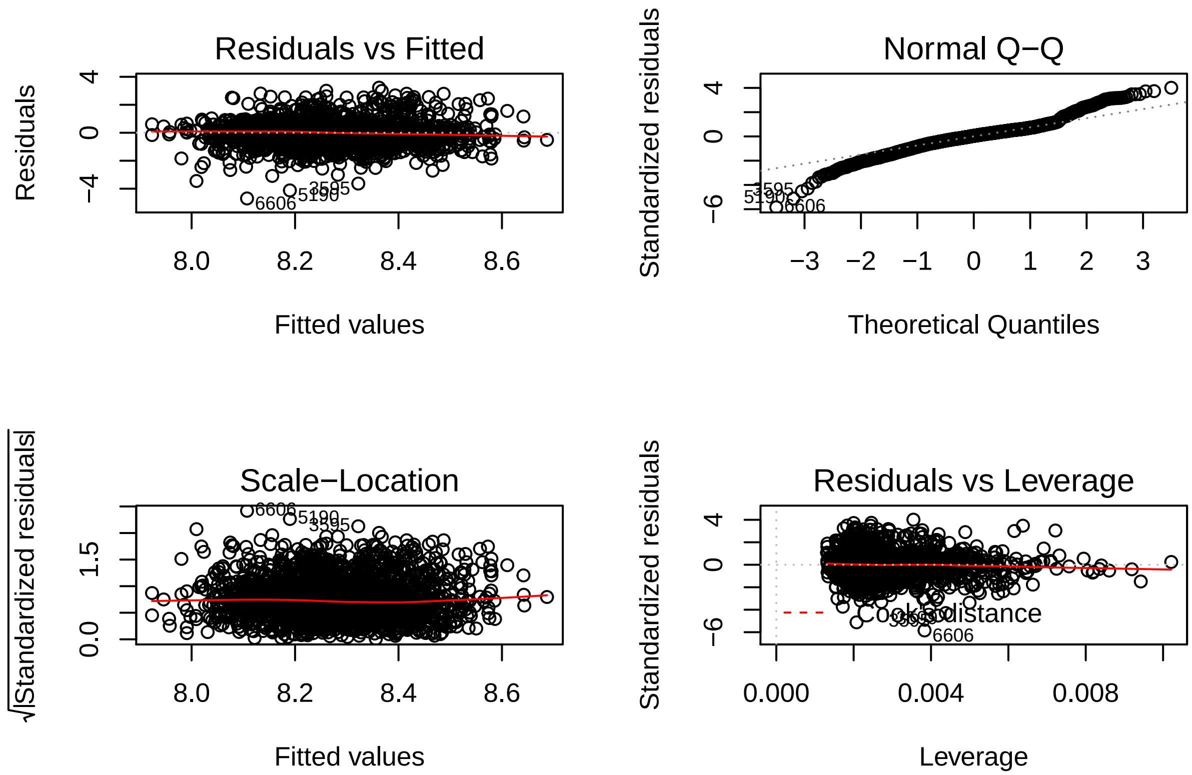
We'll review five different measures for assessing our multiple regression models:

- R-Squared
- Adjusted R-Squared
- Root Mean Squared Error
- AIC
- Corrected AIC
- BIC

```
          R_2 adj_R_2     rmse      AIC     AICc      BIC
m4     0.0198  0.0083 7891.530 5226.140 5226.852 5379.355
m4_mod 0.0144  0.0140 7901.013 5189.919 5189.931 5206.943
m5     0.0192  0.0165 7897.178 5189.442 5189.509 5234.839
m5_mod 0.0191  0.0172 7897.228 5185.715 5185.754 5219.763
m6     0.0206  0.0184 7885.800 5184.273 5184.325 5223.995
```

Based on the table above, model 6 appears to be the superior model.
It has superior measures in all categories except for BIC.

Let's now do some quick model diagnostics for our selected model, model 6:



The residuals in our model appear to have a relatively constant variance across all fitted values. The qq plot indicates standardized residuals that are fairly well behaved, with only minor departures from normality. Finally, there are only a couple outliers in our data–none appear to be high leverage points.

Make Predictions

Let's wrap up by scrubbing the test data set and make predictions. Please refer to the Github account in the Appendix to access the prediction file.

Appendix

- Link to full code: https://github.com/spitakiss/Data621/tree/master/Homework4/Grzasko_HW4.Rmd
- Prediction file: https://github.com/spitakiss/Data621/blob/master/Homework4/evaluation_data_w_predictions.csv