Insurance Quotes

Anthony Arroyo, Seung min song, Alice A. Friedman

05/07/2023

Abstract

We discuss the challenges from evaluating insurance data, assessing whether a car is involved in a crash and then how expensive the claim would be. Our goal is to create a logistic regression model, which predicts the binary value of whether or not a car has been in a crash (TARGET_FLAG), as well as a linear regression, which predicts the amount of the claim (TARGET_AMT). Our linear regression models pull some insights regarding the effects of marital status and others on claim costs. MORE TK.

Part 1. Data Exploration

The data has 23 features and 2 target variables that we wish to predict. As part of our predictive analytics, first we will decide if the car that belongs to a license plate has been in a crash before. Second, we will determine how much the cost of the crash was if they were.

- TARGET_FLAG: Was car in a crash? 1=YES 0=NO None
- TARGET_AMT: If car was in a crash, what was the cost of the damage?

There are some general myths related to driving that may skew our view about what we expect. How much of it is true? Provided below is a quick data dictionary of the features we will be measuring.

- AGE: Age of Driver
- BLUEBOOK: Value of Vehicle
- CAR_AGE: Vehicle Age
- CAR_TYPE: Type of Car
- CAR_USE: Vehicle Use
- CLM_FREQ: # Claims (Past 5 Years)
- EDUCATION: Max Education Level
- HOMEKIDS: # Children at Home
- HOME_VAL: Home Value
- INCOME: Income
- JOB: Job Category
- KIDSDRIV: # Driving Children
- MSTATUS: Marital Status
- MVR_PTS: Motor Vehicle Record Points
- OLDCLAIM: Total Claims (Past 5 Years)
- PARENT1: Single Parent
- RED_CAR: A Red Car
- REVOKED: License Revoked (Past 7 Years)

- SEX: Gender
- TIF: Time in Force
- TRAVTIME: Distance to Work
- URBANICITY: Home/Work Area
- YOJ: Years on Job

The top of the dataset is shown below, where we can see a smattering of isues such as untidy data and even a missing value in one of the income observations.

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ  INCOME PARENT1
## 1     1           0          0        0  60        0  11 $67,349       No
## 2     2           0          0        0  43        0  11 $91,449       No
## 3     4           0          0        0  35        1  10 $16,039       No
## 4     5           0          0        0  51        0  14                No
##   HOME_VAL MSTATUS SEX       EDUCATION          JOB TRAVTIME     CAR_USE BLUEBOOK
## 1       $0   z_No   M             PhD  Professional       14     Private
$14,230
## 2 $257,252   z_No   M z_High School z_Blue Collar       22  Commercial
$14,940
## 3 $124,191    Yes z_F z_High School      Clerical        5     Private
$4,010
## 4 $306,251    Yes   M <High School z_Blue Collar       32     Private
$15,440
##   TIF CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11  Minivan     yes   $4,461        2      No       3      18
## 2   1  Minivan     yes       $0        0      No       0       1
## 3   4    z_SUV      no  $38,690        2      No       3      10
## 4   7  Minivan     yes       $0        0      No       0       6
##              URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Urban/ Urban
## 4 Highly Urban/ Urban
```

Tidy Data

After carefully examining the data in the raw data frame, it was found that certain variables contained unnecessary characters such as dollar signs. To address this issue, regular expression (regex) was used to remove these unwanted characters from the variables. By doing so, the data is now cleaner and more suitable for analysis.

Additionally, there are several nominal categories that are yes and no questions. We reduced this to binary where 1=yes and 0=no using regex again to further simplify the process.

Finally, we convert variables to type factor or numeric, as appropriate.

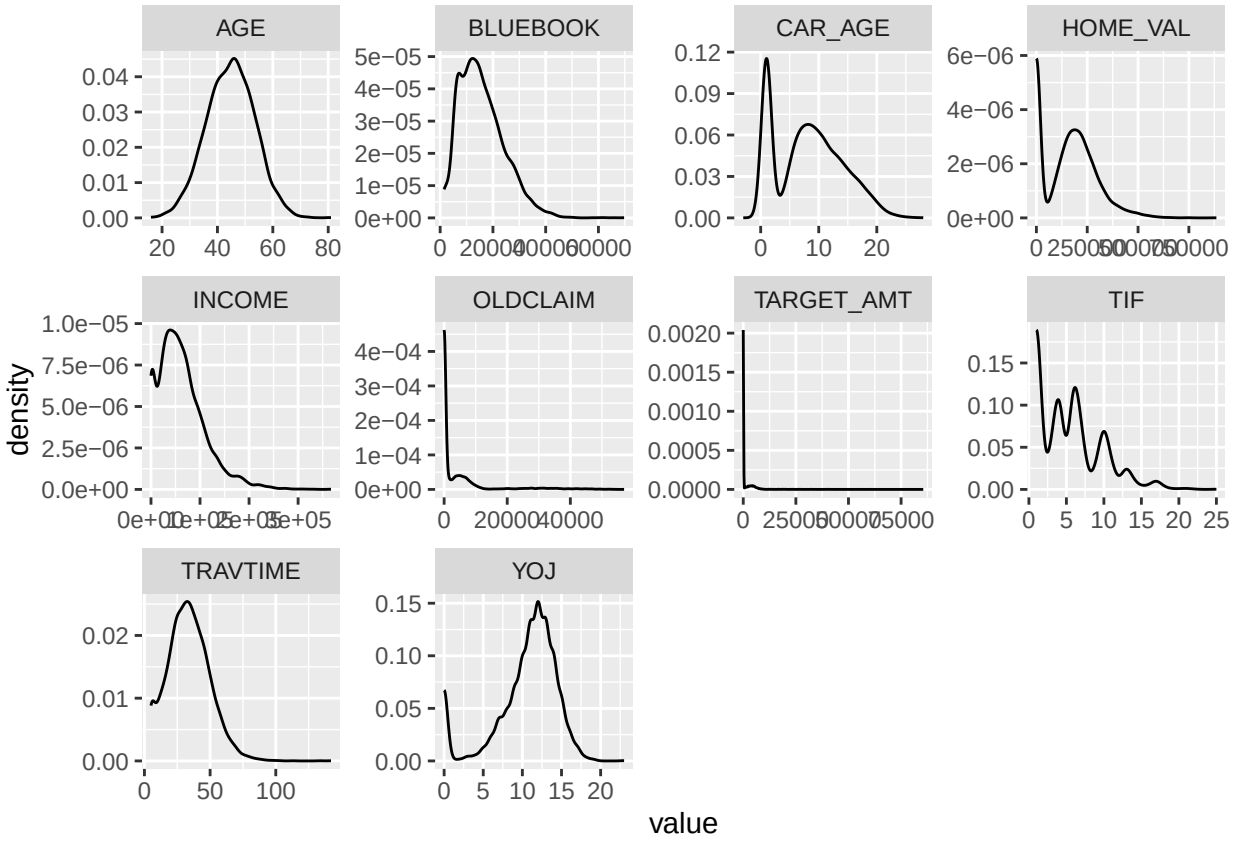
```
##       INDEX          TARGET_FLAG       TARGET_AMT         KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE            HOMEKIDS           YOJ             INCOME
```

```
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Min.   :     0
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097
##  Median :45.00   Median :0.0000   Median :11.0   Median : 54028
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   Mean   : 61898
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
##  NA's   :6                        NA's   :454    NA's   :445
##     PARENT1          HOME_VAL         MSTATUS       SEX            EDUCATION
##  Min.   :0.000   Min.   :     0   Min.   :0.0000   F:4375   Bachelors
:2242
##  1st Qu.:0.000   1st Qu.:     0   1st Qu.:0.0000   M:3786   High School:3533
##  Median :0.000   Median :161160   Median :1.0000            Masters
:1658
##  Mean   :0.132   Mean   :154867   Mean   :0.5997            PhD
: 728
##  3rd Qu.:0.000   3rd Qu.:238724   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :885282   Max.   :1.0000
##                  NA's   :464
##             JOB           TRAVTIME         CAR_USE        BLUEBOOK
##  Blue Collar :1825   Min.   : 5.00   Commercial:3029   Min.   : 1500
##  Clerical    :1271   1st Qu.: 22.00   Private   :5132   1st Qu.: 9280
##  Professional:1117   Median : 33.00                     Median :14440
##  Manager     : 988   Mean   : 33.49                     Mean   :15710
##  Lawyer      : 835   3rd Qu.: 44.00                     3rd Qu.:20850
##  Student     : 712   Max.   :142.00                     Max.   :69740
##  (Other)     :1413
##       TIF             CAR_TYPE       RED_CAR          OLDCLAIM
##  Min.   : 1.000   Minivan    :2145   Min.   :0.0000   Min.   :     0
##  1st Qu.: 1.000   Panel Truck: 676   1st Qu.:0.0000   1st Qu.:     0
##  Median : 4.000   Pickup     :1389   Median :0.0000   Median :     0
##  Mean   : 5.351   Sports Car : 907   Mean   :0.2914   Mean   : 4037
##  3rd Qu.: 7.000   SUV        :2294   3rd Qu.:1.0000   3rd Qu.: 4636
##  Max.   :25.000   Van        : 750   Max.   :1.0000   Max.   :57037
##
##     CLM_FREQ        REVOKED         MVR_PTS         CAR_AGE
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000   Median :0.0000   Median : 1.000   Median : 8.000
##  Mean   :0.7986   Mean   :0.1225   Mean   : 1.696   Mean   : 8.328
##  3rd Qu.:2.0000   3rd Qu.:0.0000   3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000   Max.   :1.0000   Max.   :13.000   Max.   :28.000
##                                                     NA's   :510
##              URBANICITY
##  Highly Rural/ Rural:1669
##  Highly Urban/ Urban:6492
##
##
##
##
##
##
##    INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1
## 1      1           0          0        0  60         0  11  67349       0
## 2      2           0          0        0  43         0  11  91449       0
```

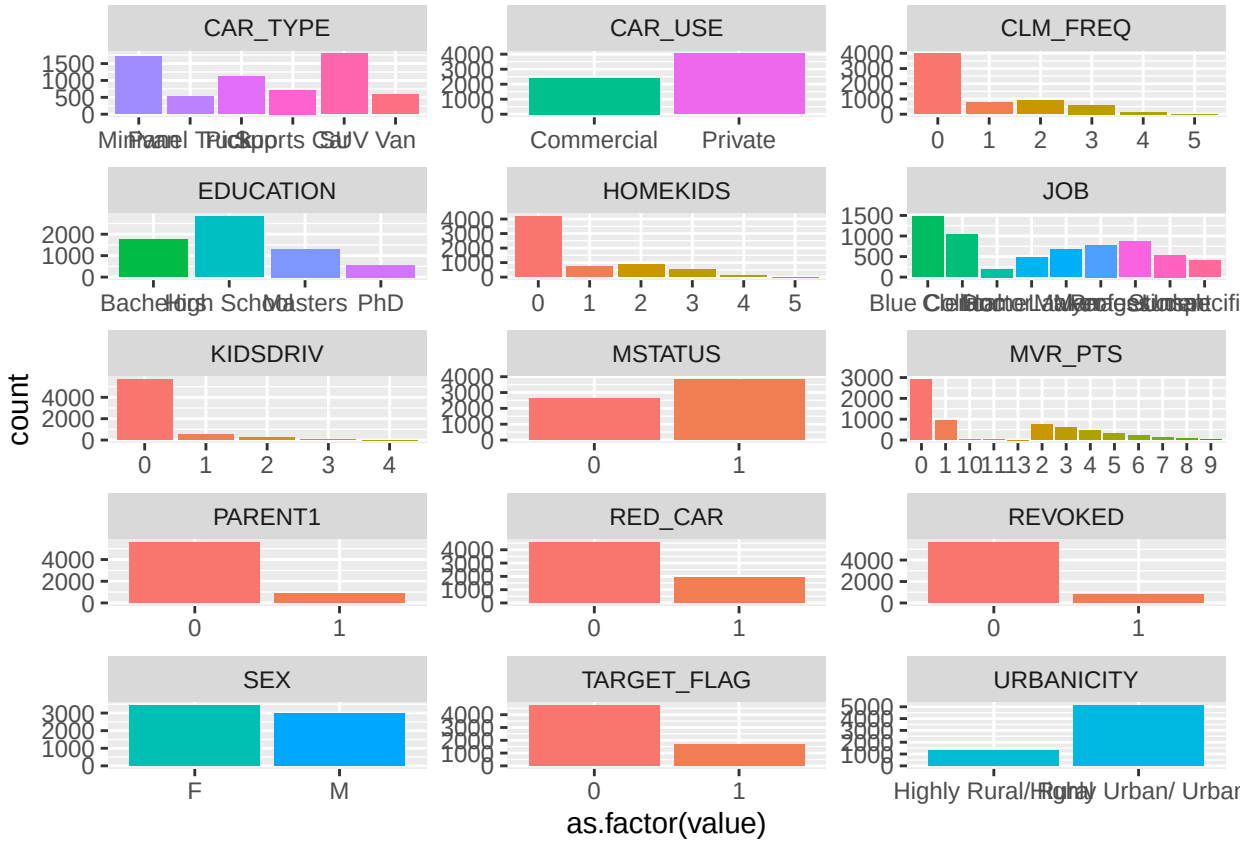
```
## 3      4             0           0          0  35          1  10  16039          0
## 4      5             0           0          0  51          0  14     NA          0
## 5      6             0           0          0  50          0  NA 114986          0
## 6      7             1        2946          0  34          1  12 125301          1
##   HOME_VAL MSTATUS SEX    EDUCATION          JOB TRAVTIME     CAR_USE BLUEBOOK
## 1        0       0   M          PhD Professional        14     Private
14230
## 2   257252       0   M High School  Blue Collar        22 Commercial
14940
## 3   124191       1   F High School     Clerical         5     Private
4010
## 4   306251       1   M High School  Blue Collar        32     Private
15440
## 5   243925       1   F          PhD       Doctor        36     Private
18000
## 6        0       0   F    Bachelors  Blue Collar        46 Commercial
17430
##   TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11     Minivan       1     4461        2       0       3      18
## 2   1     Minivan       1        0        0       0       0       1
## 3   4         SUV       0    38690        2       0       3      10
## 4   7     Minivan       1        0        0       0       0       6
## 5   1         SUV       0    19217        2       1       3      17
## 6   1 Sports Car       0        0        0       0       0       7
##              URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Urban/ Urban
## 4 Highly Urban/ Urban
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

Visulization of the data set

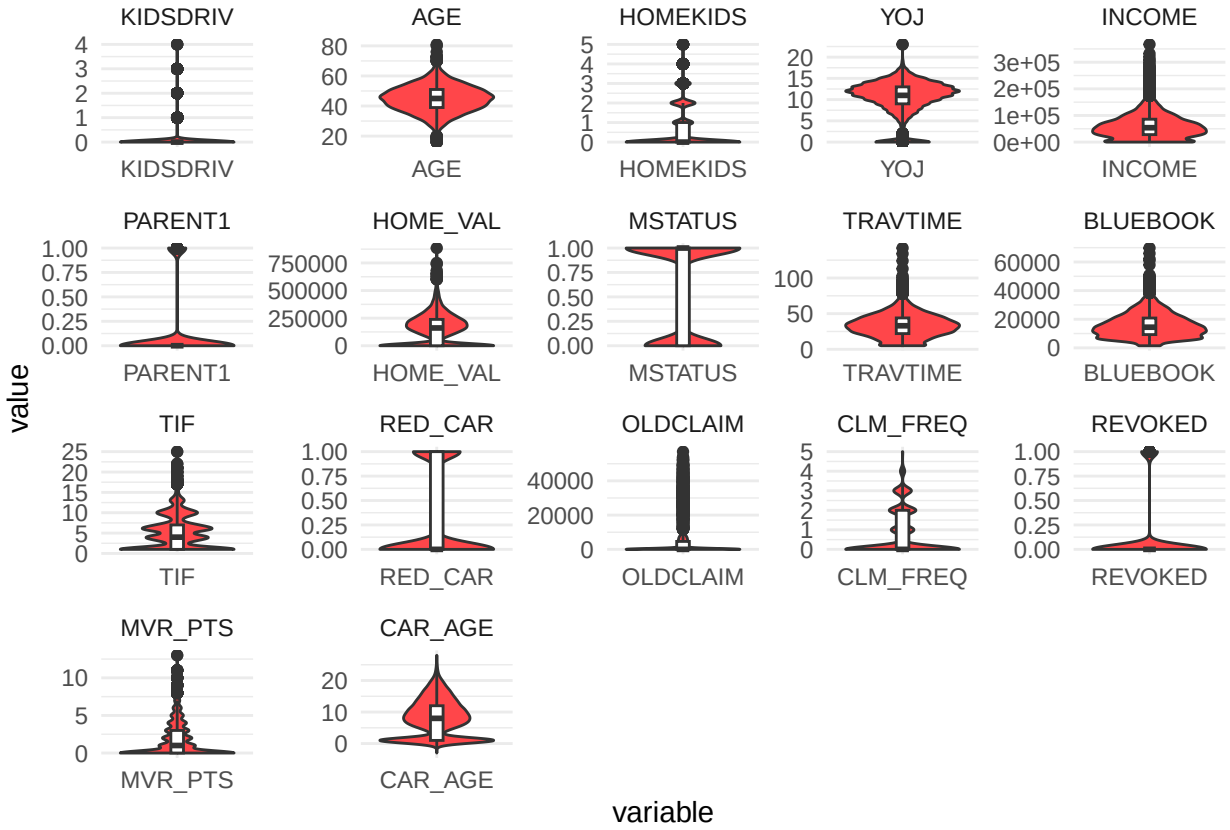
Our features span a multitude of different distributions. None of them are strictly normal unless 0 values are excluded, although AGE comes closest. It is unclear in the data if 0 is equivalent to missing data, or whether some car owners have 0 income (many of whom are coded as students or homemakers) or a 0 home value (perhaps renters?). Several are categorical(e.g., JOB), ordinal (e.g., EDUCATION) or binomial (e.g., RED_CAR).



Warning: attributes are not identical across measure variables; they will be
dropped




```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation ideoms with `aes()`
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Correlation

Several of the continuous variables are show to correlated. We see a strong correlation between INCOME, HOME_VAL, BLUEBOOK, and CAR_AGE, which intuitively makes sense as higher income people are more likley to have more expensive houses, and newer, more expensive cars.

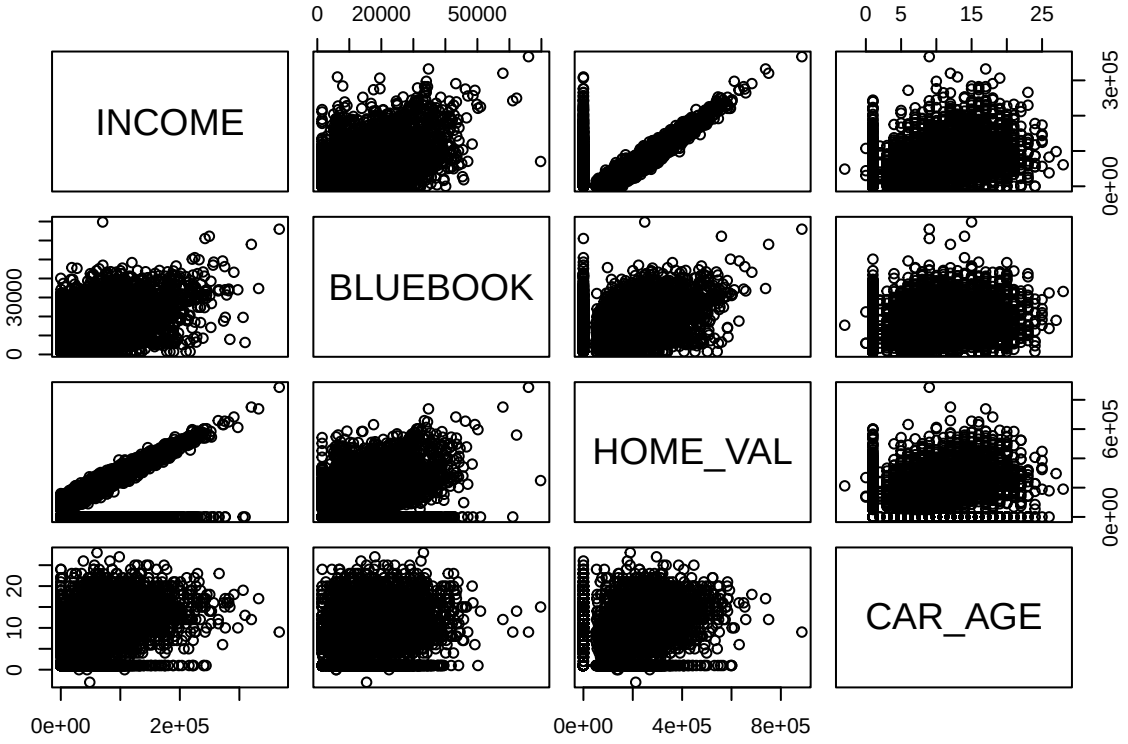
INCOME and HOME_VAL have the strongest correlation, with a correlation coefficient of 0.58. It is unlikely that both of these features should be used in the final model.

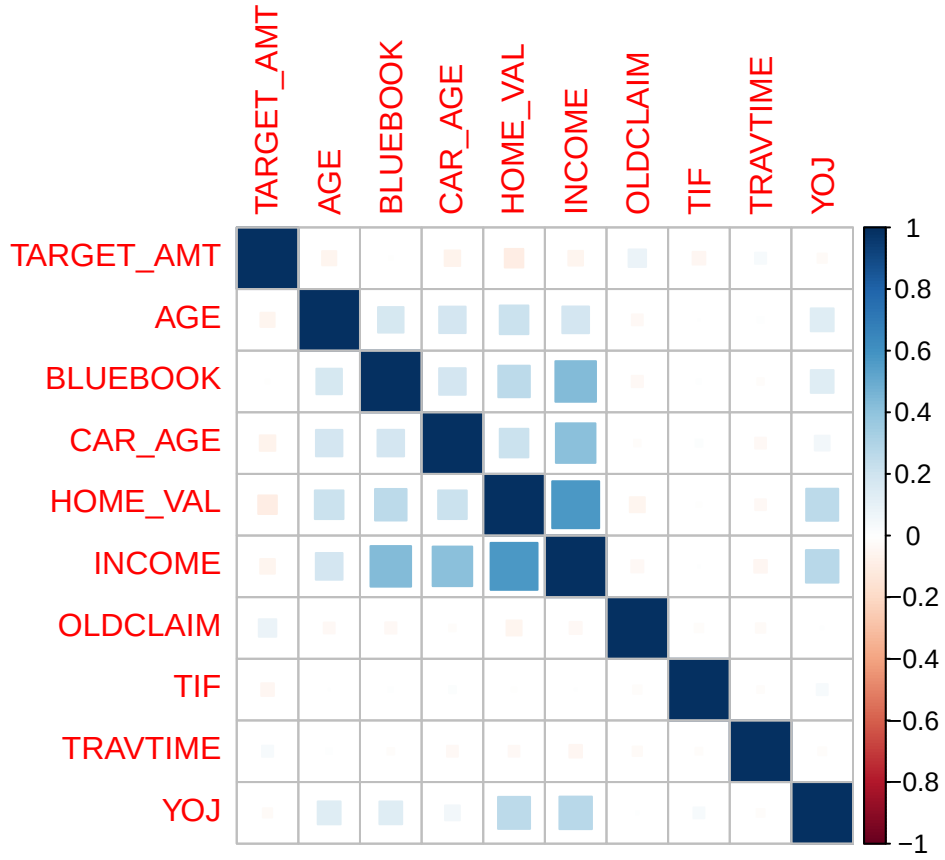
	TARGET_AMT	AGE	BLUEBOOK	CAR_AGE	HOME_VAL	INCOME	OLDCLAIM	TIF
TARGET_AMT	1.00	−0.05	0.00	−0.06	−0.09	−0.06	0.08	−0.04
AGE	−0.05	1.00	0.17	0.18	0.22	0.19	−0.03	0.00
BLUEBOOK	0.00	0.17	1.00	0.19	0.26	0.43	−0.03	0.00
CAR_AGE	−0.06	0.18	0.19	1.00	0.22	0.41	−0.01	0.01
HOME_VAL	−0.09	0.22	0.26	0.22	1.00	0.58	−0.06	0.00
INCOME	−0.06	0.19	0.43	0.41	0.58	1.00	−0.04	0.00
OLDCLAIM	0.08	−0.03	−0.03	−0.01	−0.06	−0.04	1.00	−0.02
TIF	−0.04	0.00	0.00	0.01	0.00	0.00	−0.02	1.00
TRAVTIME	0.03	0.01	−0.01	−0.03	−0.03	−0.04	−0.02	−0.01
YOJ	−0.02	0.14	0.14	0.06	0.27	0.28	0.00	0.03

	TRAVTIME	YOJ
TARGET_AMT	0.03	−0.02
AGE	0.01	0.14
BLUEBOOK	−0.01	0.14
CAR_AGE	−0.03	0.06
HOME_VAL	−0.03	0.27

```
## INCOME          -0.04   0.28
## OLDCLAIM        -0.02   0.00
## TIF             -0.01   0.03
## TRAVTIME         1.00  -0.02
## YOJ             -0.02   1.00
```

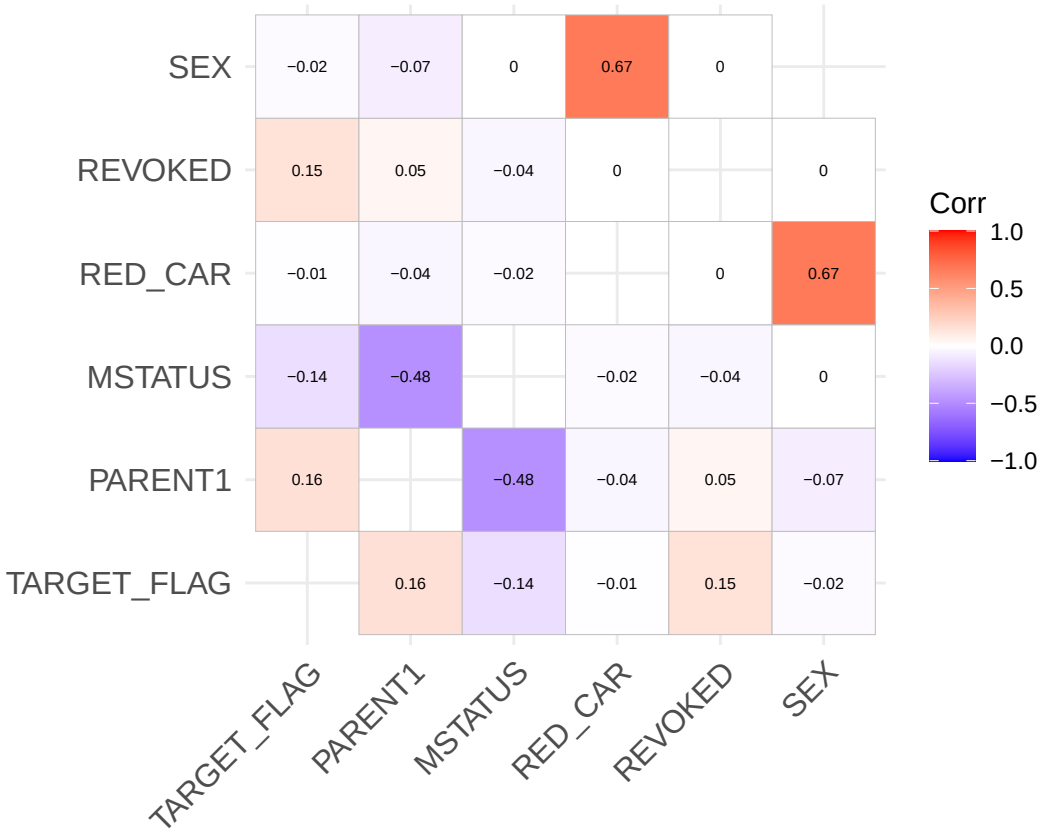
```
##             INCOME BLUEBOOK HOME_VAL CAR_AGE
## INCOME        1.00     0.43     0.58    0.41
## BLUEBOOK      0.43     1.00     0.26    0.19
## HOME_VAL      0.58     0.26     1.00    0.22
## CAR_AGE       0.41     0.19     0.22    1.00
```





Categorical data correlations must be calculated differently. Purely categorical data correlation can be assessed using different methods. One method to compare correlation between a continuous and binomial feauture is to determine if a logistic model can be developed with a significant P value.

As shown below, many of the categorical variables are highly correlated. For example, RED_CAR is highly predictive of SEX == M. Martial status and being a single parent are (of course) mechanistically connected (all single parents are unmarried).


SEX
REVOKED
RED_CAR
MSTATUS
PARENT1
TARGET_FLAG
−0.02
−0.07
0
0.67
0
0.15
0.05
−0.04
0
0
−0.01
−0.04
−0.02
0
0.67
−0.14
−0.48
−0.02
−0.04
0
0.16
−0.48
−0.04
0.05
−0.07
0.16
−0.14
−0.01
0.15
−0.02
TARGET_FLAG
PARENT1
MSTATUS
RED_CAR
REVOKED
SEX
Corr
1.0
0.5
0.0
−0.5
−1.0