

# Auto Insurance Claims

John Cruz, Noori Selina, Shaya Engelman, Daniel Craig, Gavriel Steinmetz-Silber

2024-04-02

## Required Libraries

```
library(tidyverse)
library(janitor)
library(kableExtra)
library(latex2exp)
library(psych)
library(scales)
library(stringr)
library(ggcorrplot)
```

## Introduction

We will explore, analyze and model a dataset containing approximately 8000 records representing customers at an auto insurance company. Each record has two response variables. The first response variable, **TARGET\_FLAG**, is a 1 or a 0 (zero). A 1 means that the person was in a car crash. A 0 means that the person was not in a car crash. The second response variable is **TARGET\_AMT**. This value is zero if the person did not crash their car, however, if they did crash their car, this number will be a value greater than zero.

VARIABLE	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly

VARIABLE	DEFINITION	THEORETICAL EFFECT
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

## Data Exploration

### Import Data

When we import the training and evaluation dataset, we have 26 columns representing each variable we have defined above. We also have 8,161 total rows for the training set and 2,141 rows for the evaluation set. As we glance through the values in each column, we can see there is some data wrangling that will need to be performed prior to evaluating any summary statistics.

Table 2: Training Set

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	
1	0	0	0	60	0	11	\$67,349	No	
2	0	0	0	43	0	11	\$91,449	No	
4	0	0	0	35	1	10	\$16,039	No	
5	0	0	0	51	0	14		No	
6	0	0	0	50	0	NA	\$114,986	No	
7	1	2946	0	34	1	12	\$125,301	Yes	

*Dimensions:*

8161 x 26

Table 3: Evaluation Set

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1
3	NA	NA	0	48	0	11	\$52,881	No
9	NA	NA	1	40	1	11	\$50,815	Yes
10	NA	NA	0	44	2	12	\$43,486	Yes
18	NA	NA	0	35	2	NA	\$21,204	Yes
21	NA	NA	0	59	0	12	\$87,460	No
30	NA	NA	0	46	0	14		No

*Dimensions:*

2141 x 26

### Data Wrangling

- We can drop the INDEX column as it provides no value to our analysis. **Any changes applied to the training set will be similiary applied to the evaluation set, unless otherwise noted.**

Table 4: Training Set

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL
0	0	0	60	0	11	\$67,349	No	\$0
0	0	0	43	0	11	\$91,449	No	\$257,252
0	0	0	35	1	10	\$16,039	No	\$124,191
0	0	0	51	0	14		No	\$306,251
0	0	0	50	0	NA	\$114,986	No	\$243,925
1	2946	0	34	1	12	\$125,301	Yes	\$0

*Note:*

Dropped 'INDEX' column:

- The INCOME, HOME\_VAL, BLUEBOOK and OLDCLAIM columns are in a currency string format and needs to be changed to a numeric value we can work with.

Table 5: Training Set: Before

INCOME	HOME_VAL	BLUEBOOK	OLDCLAIM
\$67,349	\$0	\$14,230	\$4,461
\$91,449	\$257,252	\$14,940	\$0
\$16,039	\$124,191	\$4,010	\$38,690
	\$306,251	\$15,440	\$0
\$114,986	\$243,925	\$18,000	\$19,217
\$125,301	\$0	\$17,430	\$0

Table 6: Training Set: After

INCOME	HOME_VAL	BLUEBOOK	OLDCLAIM
67349	0	14230	4461
91449	257252	14940	0
16039	124191	4010	38690
NA	306251	15440	0
114986	243925	18000	19217
125301	0	17430	0

- **MSTATUS**, **SEX**, **EDUCATION**, **JOB**, **CAR\_TYPE**, **URBANICITY** has extra characters **z\_** that need to be removed from their binary (No) or categorical values (ex. SUV). We also have **EDUCATION** having the < within it as well.

Table 7: Training Set: Before

MSTATUS	SEX	EDUCATION	JOB	CAR_TYPE	URBANICITY
z_No	M	PhD	Professional	Minivan	Highly Urban/ Urban
z_No	M	z_High School	z_Blue Collar	Minivan	Highly Urban/ Urban
Yes	z_F	z_High School	Clerical	z_SUV	Highly Urban/ Urban
Yes	M	<High School	z_Blue Collar	Minivan	Highly Urban/ Urban
Yes	z_F	PhD	Doctor	z_SUV	Highly Urban/ Urban
z_No	z_F	Bachelors	z_Blue Collar	Sports Car	Highly Urban/ Urban

Table 8: Training Set: After

MSTATUS	SEX	EDUCATION	JOB	CAR_TYPE	URBANICITY
No	M	PhD	Professional	Minivan	Highly Urban/ Urban
No	M	High School	Blue Collar	Minivan	Highly Urban/ Urban
Yes	F	High School	Clerical	SUV	Highly Urban/ Urban
Yes	M	High School	Blue Collar	Minivan	Highly Urban/ Urban
Yes	F	PhD	Doctor	SUV	Highly Urban/ Urban
No	F	Bachelors	Blue Collar	Sports Car	Highly Urban/ Urban

- The **URBANICITY** has two values within it as noted in our definitions above. The first value is their home area and the second is their work area. So a person could live in a highly rural area, but works in a rural area. We will separate this column into two new columns, while retaining the original one for flexibility later on.

Table 9: Training Set: Before

URBANICITY
Highly Urban/ Urban
Highly Urban/ Urban
Highly Urban/ Urban
Highly Urban/ Urban
Highly Urban/ Urban
Highly Urban/ Urban

Table 10: Training Set: After

URBANICITY	HOME_AREA	WORK_AREA
Highly Urban/ Urban	Highly Urban	Urban
Highly Urban/ Urban	Highly Urban	Urban
Highly Urban/ Urban	Highly Urban	Urban
Highly Urban/ Urban	Highly Urban	Urban
Highly Urban/ Urban	Highly Urban	Urban
Highly Urban/ Urban	Highly Urban	Urban

- Here we will change some of our variables' values into factors.
  - PARENT1: Yes/No
  - MSTATUS: Yes/No
  - SEX: M/F
  - RED\_CAR: Yes/No (Fix capital punctuation of these values)
  - REVOKED: Yes/No
  - EDUCATION: High School, Bachelors, Masters, PhD (Ordered Factor as each level has an ordered precedence of completing it.)

Table 11: Training Set: Before

PARENT1	MSTATUS	SEX	RED_CAR	REVOKED	EDUCATION
No	No	M	yes	No	PhD
No	No	M	yes	No	High School
No	Yes	F	no	No	High School
No	Yes	M	yes	No	High School
No	Yes	F	no	Yes	PhD
Yes	No	F	no	No	Bachelors

Table 12: Training Set: After

PARENT1	MSTATUS	SEX	RED_CAR	REVOKED	EDUCATION
No	No	M	Yes	No	PhD
No	No	M	Yes	No	High School
No	Yes	F	No	No	High School
No	Yes	M	Yes	No	High School
No	Yes	F	No	Yes	PhD
Yes	No	F	No	No	Bachelors

### Summary Statistics

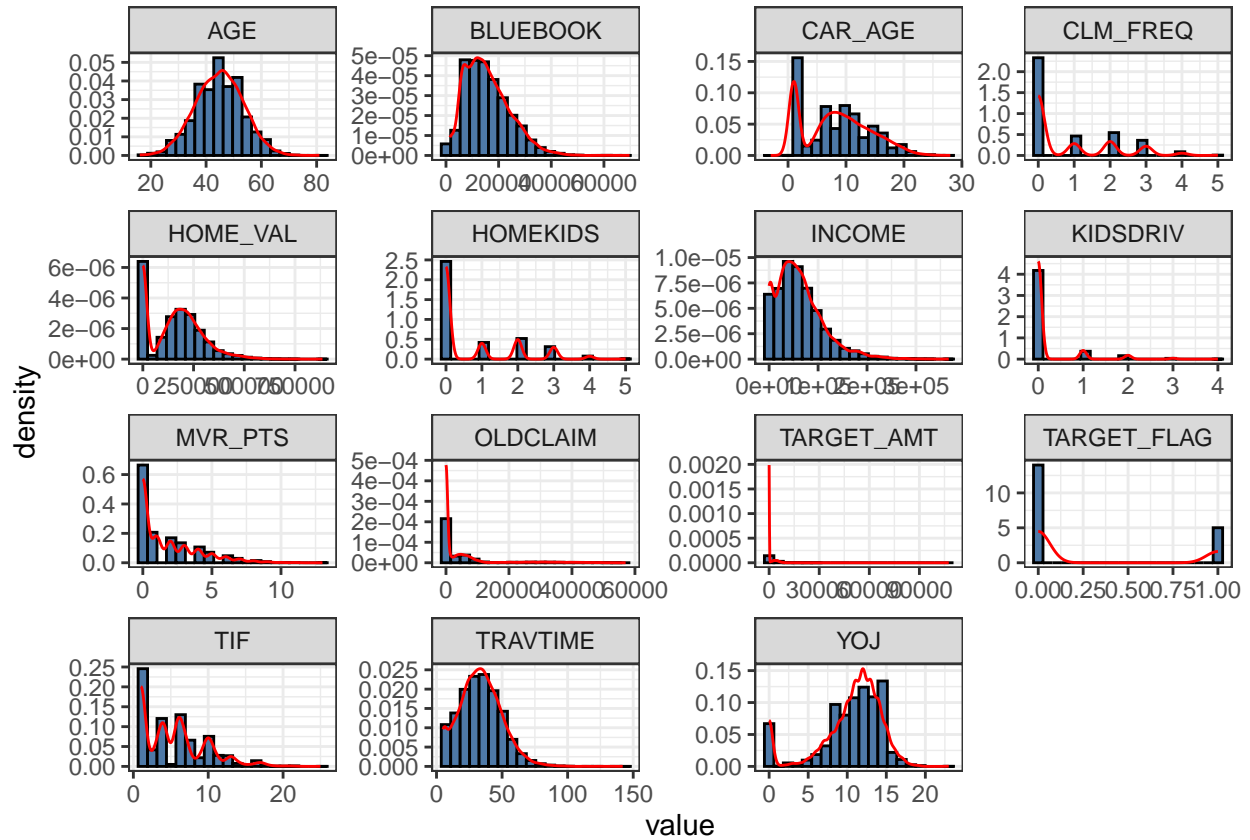
We have an average customer age of 44.79. Their average income is almost \$62k while their home value is approximately \$155k. For cars in a crash there is an average cost of \$1500.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
TARGET_FLAG	1	8161	0.26	0.44	0	0.20	0.00	0	1.0	1.0	1.07	-0.85	0.00
TARGET_AMT	2	8161	1504.32	4704.03	0	593.71	0.00	0	107586.1	107586.1	8.71	112.29	52.07
KIDSDRIV	3	8161	0.17	0.51	0	0.03	0.00	0	4.0	4.0	3.35	11.78	0.01
AGE	4	8155	44.79	8.63	45	44.83	8.90	16	81.0	65.0	-0.03	-0.06	0.10
HOMEKIDS	5	8161	0.72	1.12	0	0.50	0.00	0	5.0	5.0	1.34	0.65	0.01
YOJ	6	7707	10.50	4.09	11	11.07	2.97	0	23.0	23.0	-1.20	1.18	0.05
INCOME	7	7716	61898.09	47572.68	54028	56840.98	41792.27	0	367030.0	367030.0	1.19	2.13	541.58
HOME_VAL	9	7697	154867.29	129123.77	161160	144032.07	147867.11	0	885282.0	885282.0	0.49	-0.02	1471.79
TRAVTIME	14	8161	33.49	15.91	33	33.00	16.31	5	142.0	137.0	0.45	0.66	0.18
BLUEBOOK	16	8161	15709.90	8419.73	14440	15036.89	8450.82	1500	69740.0	68240.0	0.79	0.79	93.20
TIF	17	8161	5.35	4.15	4	4.84	4.45	1	25.0	24.0	0.89	0.42	0.05
OLDCLAIM	20	8161	4037.08	8777.14	0	1719.29	0.00	0	57037.0	57037.0	3.12	9.86	97.16
CLM_FREQ	21	8161	0.80	1.16	0	0.59	0.00	0	5.0	5.0	1.21	0.28	0.01
MVR_PTS	23	8161	1.70	2.15	1	1.31	1.48	0	13.0	13.0	1.35	1.38	0.02
CAR_AGE	24	7651	8.33	5.70	8	7.96	7.41	-3	28.0	31.0	0.28	-0.75	0.07

## Visualizations

### Density

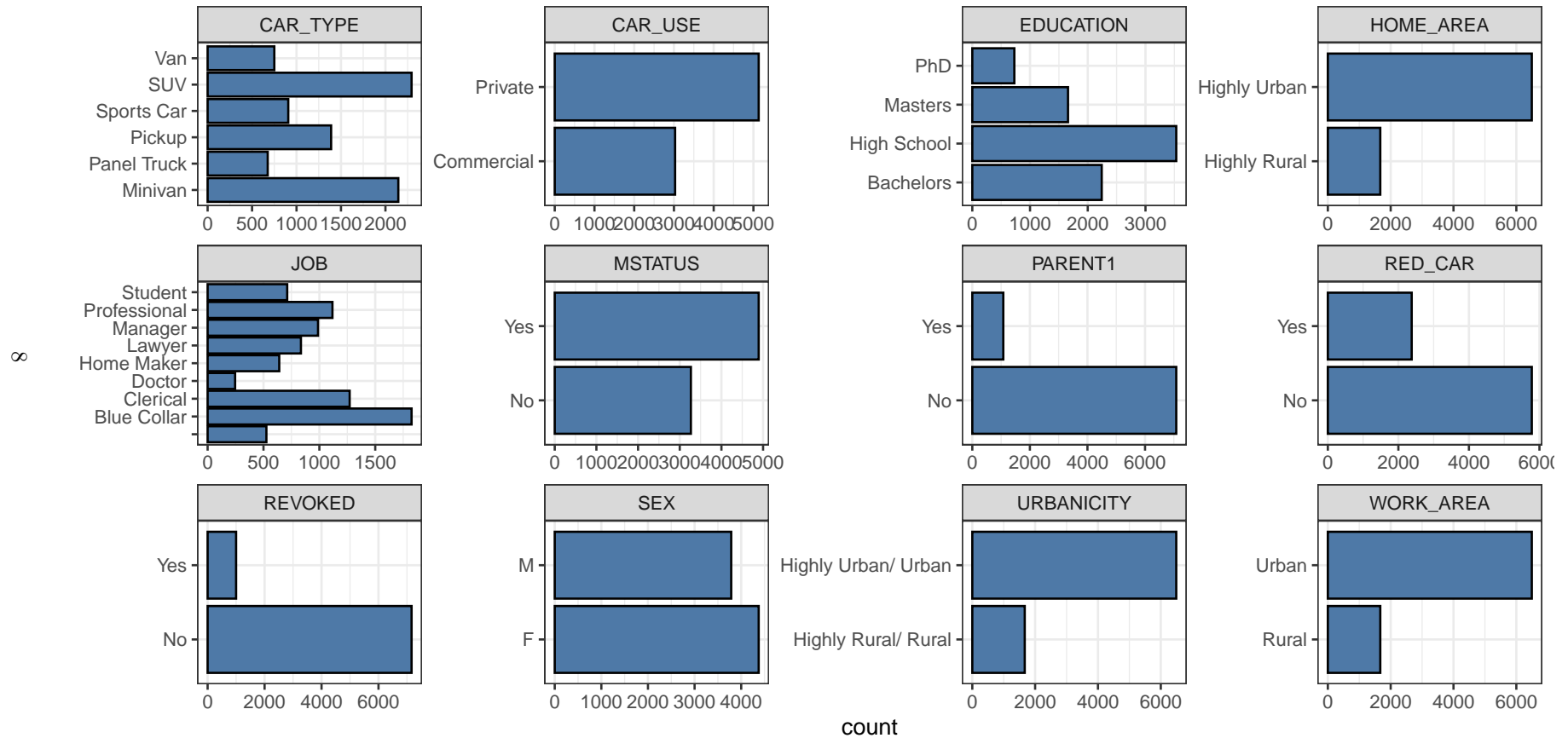
We can get a better idea of the distributions and skewness by plotting our variables. We have a normal distribution for **AGE**. As for our response variable **TARGET\_FLAG**, it clearly shows the logit function between zero and one. Other plots show significant right skewness for **BLUEBOOK**, **INCOME**, **MVR\_PTS**, **OLDCLAIM**, **TARGET\_AMT**, **TIF** and **TRAVTIME**. We also have some bimodal distributions for **CAR\_AGE**, **HOME\_VAL** and **YOJ**. We would need to perform some transformations on these variables, and possibly consider grouping the bimodal variables.



## Bar Plots

Our bar plots show us how our categorical data is divided up.

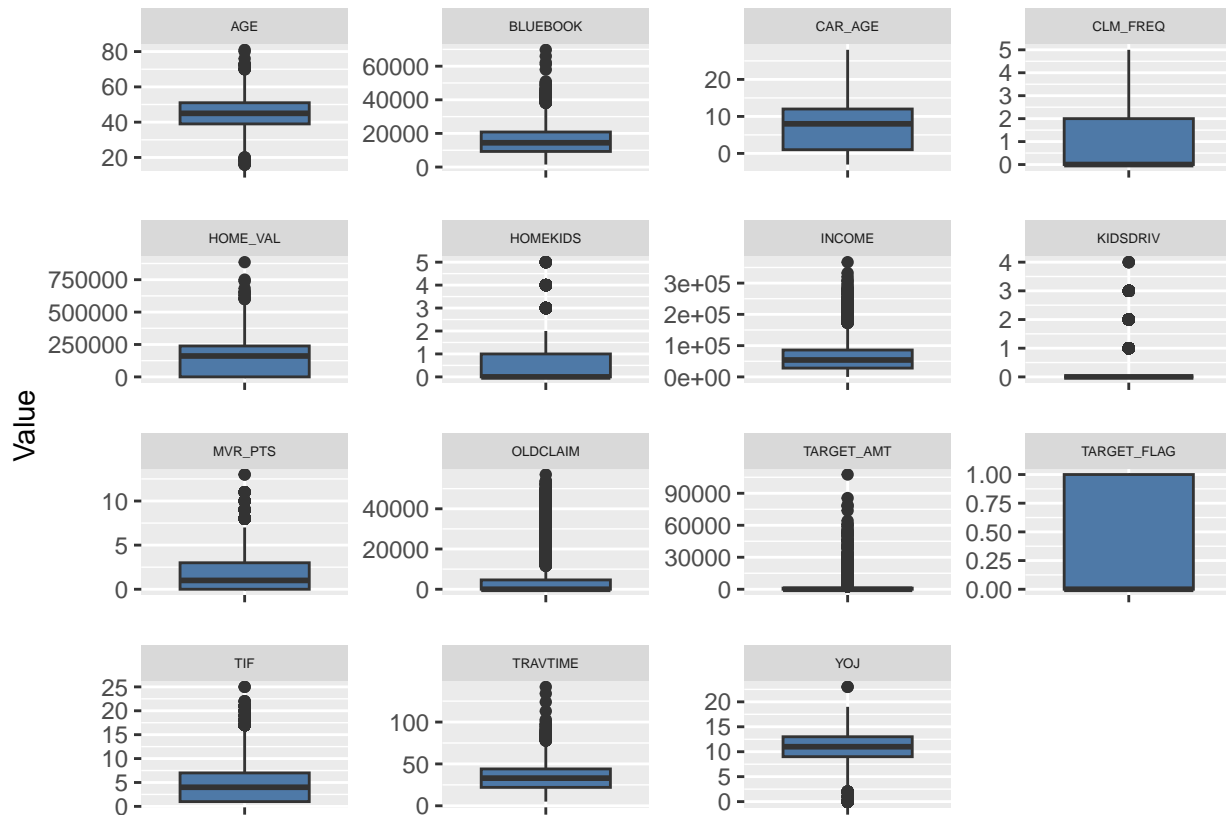
- Most of the car types we have are either SUV or Minivan.
- We see that most drivers highest education is High School or Bachelors
- The drivers predominately live/work in Highly Urban/Urban areas.





## Box Plots

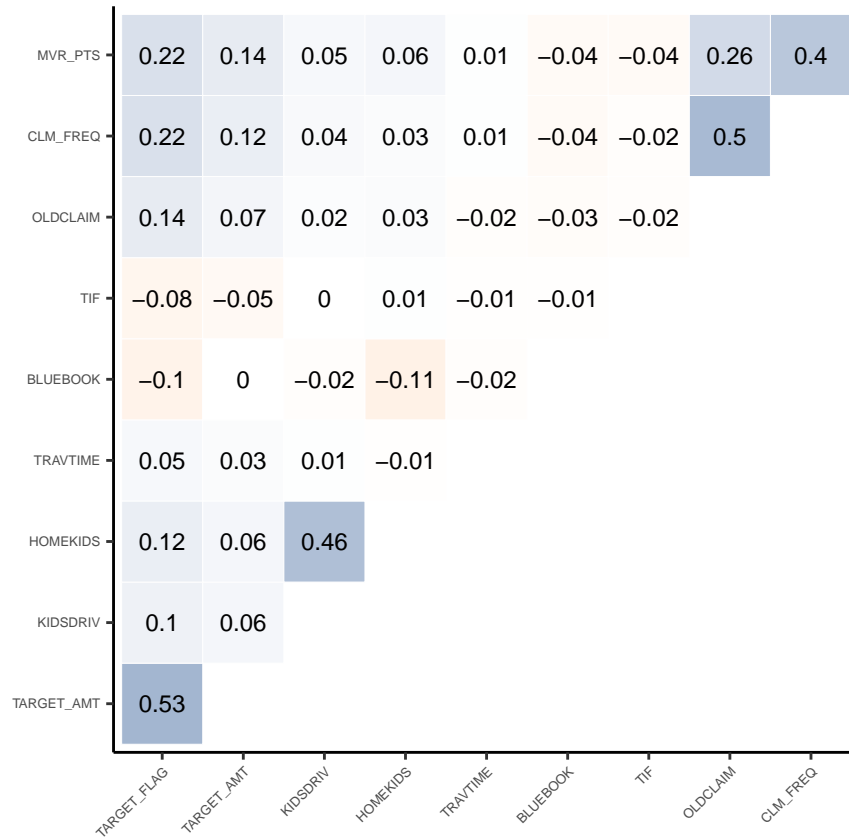
Our box plots show us there are some outliers to be dealt with. We can see the **BLUEBOOK** value of cars have some quite pricey vehicles being insured. We also see how some of our variables where they are countable numbers such as **HOMEKIDS** and **KIDSDRIV** where parents have a child, but they are not driving yet.



## Correlation Matrix

We have some moderately strong correlations between our variables. This will have to be addressed with when we build our models.

- **KIDSDRIV** and **HOMEKIDS**: They should have some multicollinearity as if you have children, they may be of age to drive already
- **MVR\_PTS** and **CLM\_FREQ**: This association should have multicollinearity as if you have higher motor vehicle points accumulated from negative driving habits, you may be more likely to have accidents and require to file more claims than the average driver.
- **CLM\_FREQ** and **OLDCLAIM**: There would be some multicollinearity as when you have more claims filed, you should have an older claim amount as a value.
- **TARGET\_AMT** and **TARGET\_FLAG**: If you were in a crash, you should have how much that accident was valued at.



## Missing Values

We can see we have some columns missing values.

- **AGE:** Only missing a few values and given that it is a normally distributed variable, we have many options to impute them
- **YOJ:** We are missing a lot of values for how many year people have been at their job
- **INCOME:** We don't have how much money they are making in a year. It could be that they are not working.
- **HOME\_VAL:** These missing values may be under the assumption they don't own a home and possibly renting
- **CAR\_AGE:** The highest amount of values we don't have is how old the car is.

AGE	YOJ	INCOME	HOME_VAL	CAR_AGE
6	454	445	464	510