DATA 621 - HW4

Andrew Bowen, Glen Davis, Shoshana Farber, Joshua Forster, Charles Ugiagbe

2023-10-30

Homework 4 - Binary Logistic Regression & Multiple Linear Regression

Introduction:

We load an auto insurance company dataset containing 8,161 records. Each record represents a customer, and each record has two response variables: `TARGET_FLAG` and `TARGET_AMT`. Below is a short description of all the variables of interest in the data set, including these response variables:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
`INDEX`	Identification Variable	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 and very old people tend to be risky
`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 but possible more educated people tend to drive safer
`HOMEKIDS`	# Children at Home	Unknown
`HOME_VAL`	Home Value	Homeowners tend to drive safer
`INCOME`	Income	Rich people tend to be in fewer crashes
`JOB`	Job Category	White collar jobs tend to be safer

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	Married people driver safer
MVR_PTS	Motor Vehicle Record Points	If you get a lot of traffic tickets, you tend to get into more accidents
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
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver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
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:

We check the classes of our variables to determine whether any of them need to be coerced to numeric or other classes prior to exploratory data analysis.

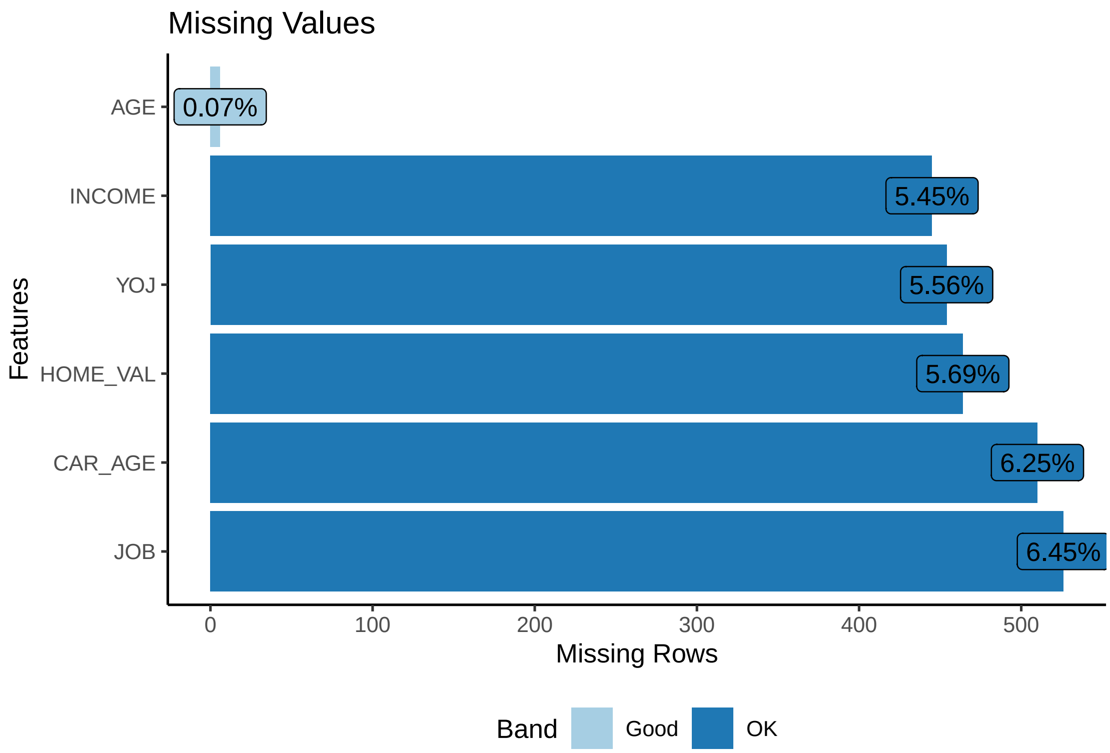
Class	Count	Variables
character	14	BLUEBOOK, CAR_TYPE, CAR_USE, EDUCATION, HOME_VAL, INCOME, JOB, MSTATUS, OLDCLAIM, PARENT1, RED_CAR, REVOKED, SEX, URBANICITY
integer	11	AGE, CAR_AGE, CLM_FREQ, HOMEKIDS, INDEX, KIDSDRIV, MVR_PTS, TARGET_FLAG, TIF, TRAVTIME, YOJ
numeric	1	TARGET_AMT

`INCOME`, `HOME_VAL`, `BLUEBOOK`, and `OLDCLAIM` are all character variables that will need to be coerced to integers after we strip the "$" from their strings. `TARGET_FLAG` and the remaining character variables will all need to be coerced to factors.

We remove the identification variable `INDEX` and take a look at a summary of the dataset's completeness.

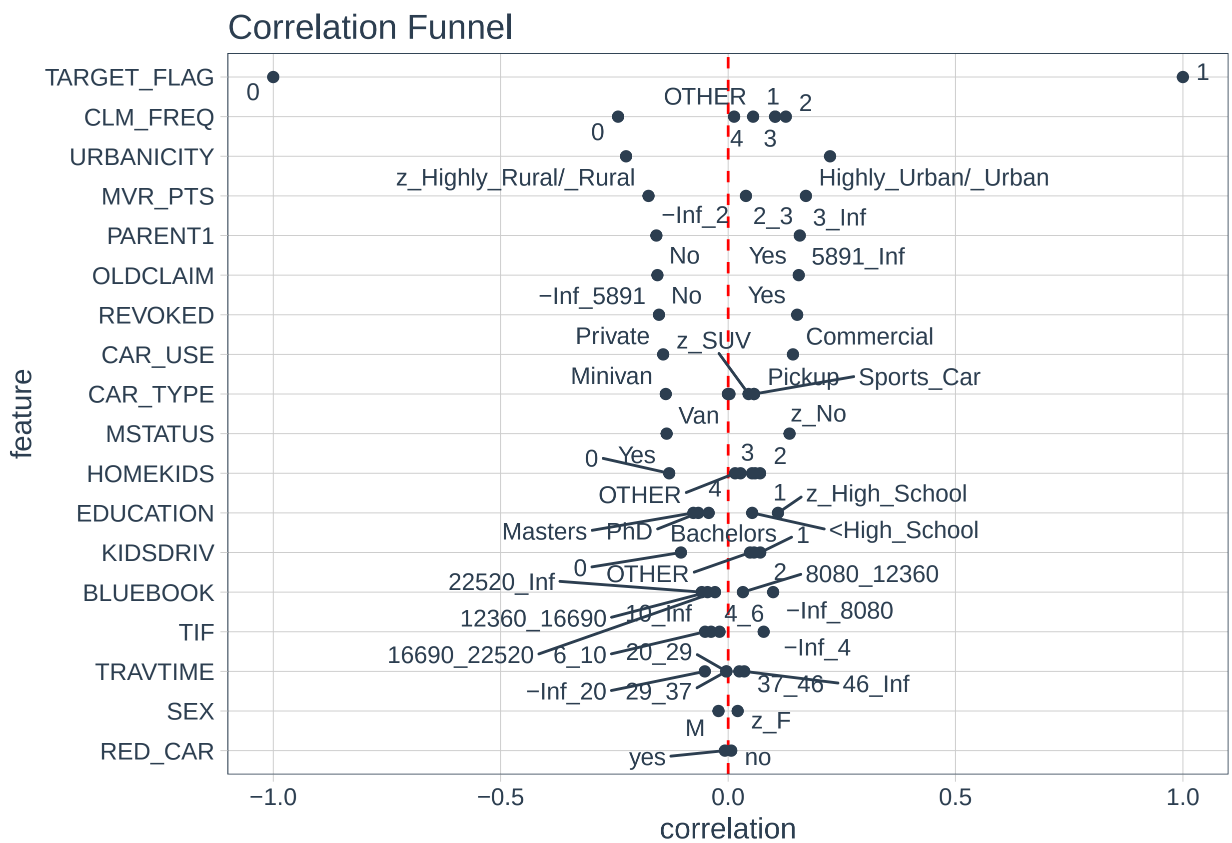
rows	8161
columns	25
all_missing_columns	0
total_missing_values	2405
complete_rows	6045

None of our columns are completely devoid of data. There are 6,045 complete rows in the dataset, which is about 74% of our observations. There are 2,405 total missing values. We take a look at which variables contain these missing values and what the spread is.



A very small percentage of observations contain missing `AGE` values. The `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`, and `JOB` variables are each missing around 5.5 to 6.5 percent of values. There are no variables containing such extreme proportions of missing values that removal would be warranted on that basis alone.

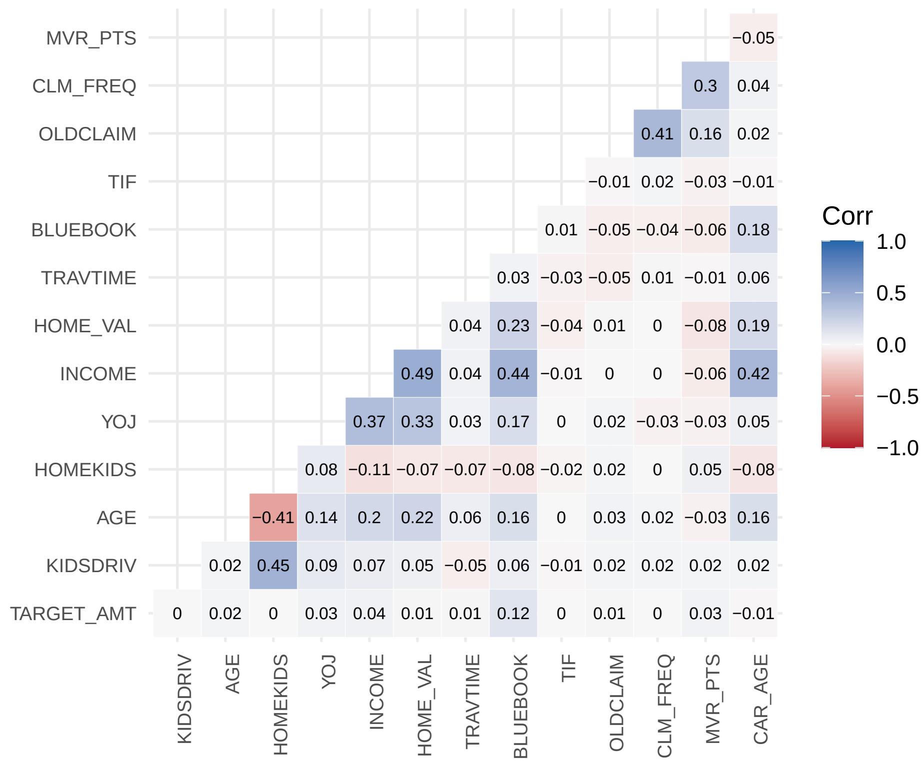
To check whether the predictor variables are correlated with the binary response variable, we produce a correlation funnel that visualizes the strength of the relationships between our predictors and `TARGET_FLAG`. This correlation funnel will not include variables for which there are any missing values.



The predictor variables without missing values that are most correlated with getting into a car crash are `CLM_FREQ`, `URBANICITY`, `MVR_PTS`, `OLDCLAIM`, `PARENT1`, `REVOKED`, and `CAR_USE`. Some of this is unsurprising. Increased claim frequency, increased numbers of traffic tickets, increased past payouts, having your license previously revoked, and using your car commercially all positively correlate with getting into a car crash, as we expected they would. We did not expect `URBANICITY` to be so relevant, but urban areas can often be more difficult to drive through and have more traffic, so that combination could reasonably make urban-dwellers more likely to get into car crashes, as the correlation suggests. We also did not expect `PARENT1` to be so relevant, but the correlation between being a single parent and getting into a car crash is very similar to that of having your license previously revoked and getting into a car crash.

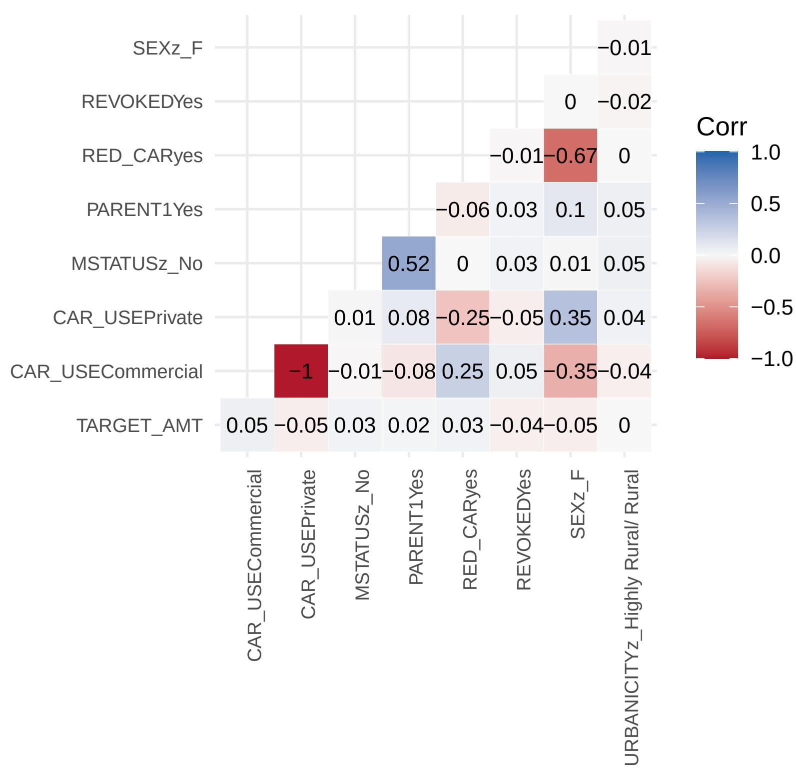
The predictor variables without missing values that are least correlated with getting into a car crash are `SEX` and `RED_CAR`. Being a woman has a very slight positive correlation with getting into a car crash, and driving a red car has a slightly negative correlation with getting into a car crash. These are contrary to urban legend, and more importantly they probably won't be useful when modeling.

To check whether the predictor variables are correlated with the numeric response variable, we produce correlation plots that visualize the strength of the relationships between our predictors and `TARGET_AMT` (only when observations involve a car crash, as otherwise we know `TARGET_AMT` $= 0$). For readability, first we look at numeric predictors only.



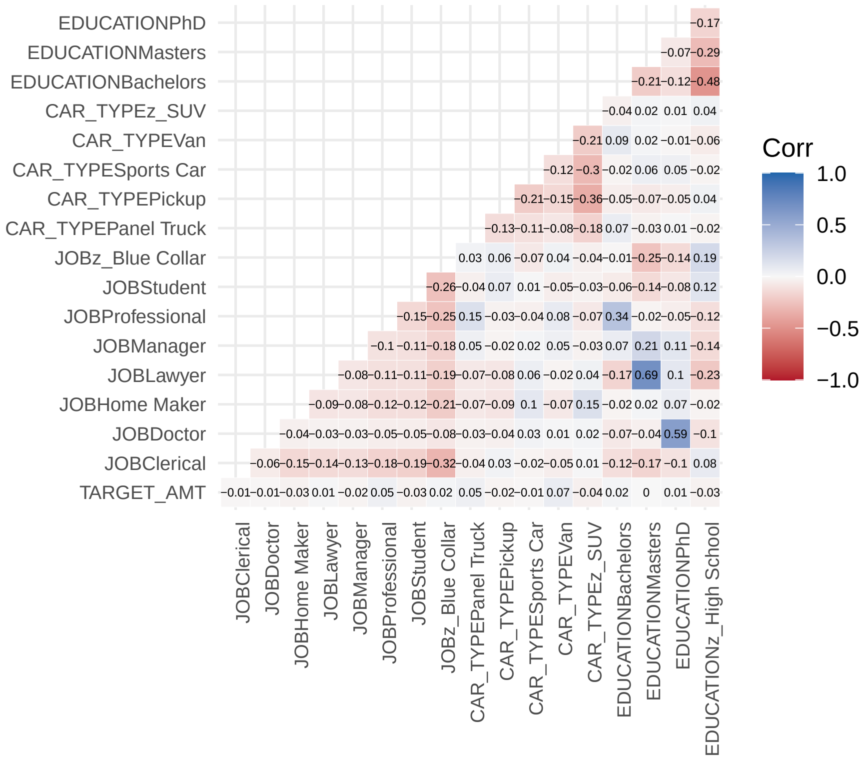
It's worth noting that `BLUEBOOK` is the single numeric variable most correlated with an increased `TARGET_AMT`, which is sensible. Cars that are currently still more valuable can be more expensive to fix. We expected `CAR_AGE` to be more negatively correlated with `TARGET_AMT`.

Next we look at two-level factors.



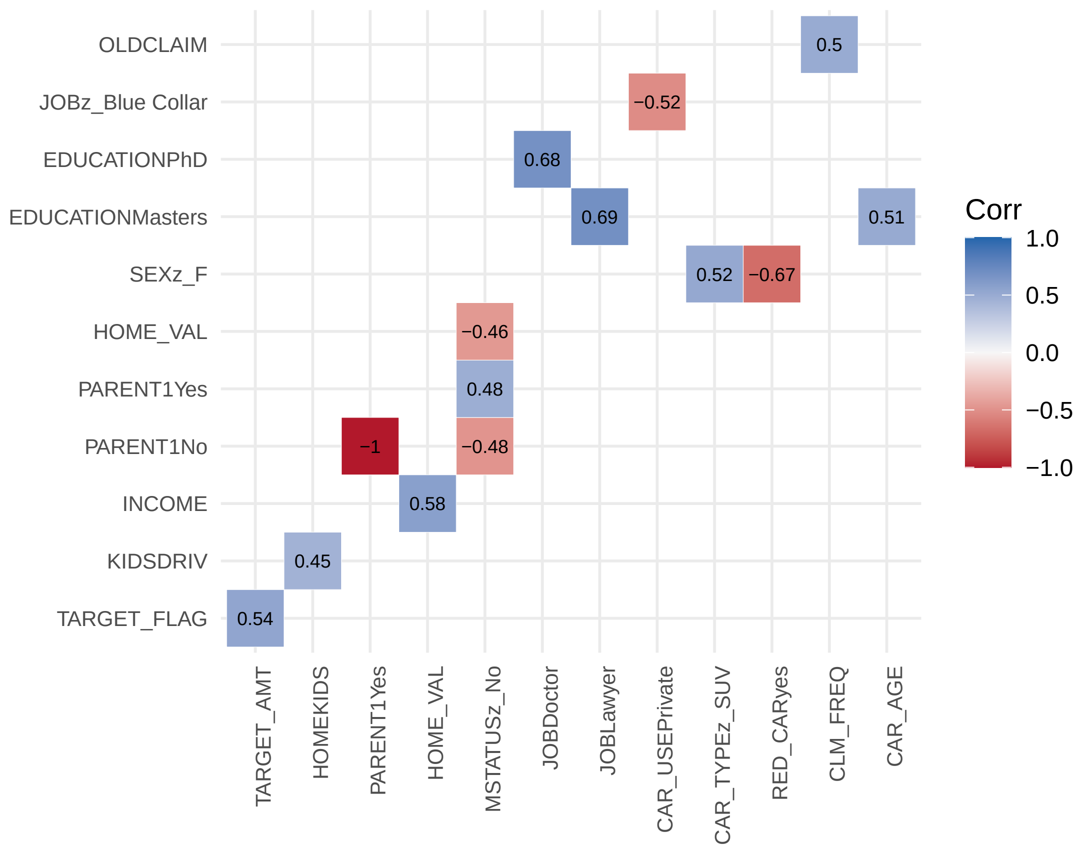
We see a small positive correlation between using your car commercially and `TARGET_AMT`. But we see an equally large negative correlation between being female and `TARGET_AMT`. The former is more logical than the latter, so neither may be a good predictor of `TARGET_AMT` ultimately.

Finally we look at factors with more than two levels.



The various car types don't have as high of a correlation (either positively or negatively) with `TARGET_AMT` as expected, but we still believe `CAR_TYPE` will be somewhat useful for modeling.

Because we have so many variables, it would be difficult to check for and visualize collinearity for our responses and predictors all at the same time without setting a threshold. So we will set a correlation threshold of 0.45 (in absolute value) and only visualize variables with any correlation values at or above that level.



We see some expected collinearity. `KIDSDRIV` and `HOMEKIDS` are moderately positively correlated because teenagers driving your car depends on you having kids at all, but the number of teens driving your car won't always exactly match the number of kids you have. `HOME_VAL` and `INCOME` are pretty positively correlated, as higher incomes lead to the ability to purchase higher valued homes. Not being married is also moderately negatively correlated with `HOME_VAL`, likely because married people often have two incomes instead of one and can therefore purchase higher valued homes. Having a PhD is equally correlated with being a doctor or lawyer, which makes sense because those jobs require them. Working a blue collar job is logically pretty negatively correlated with driving your car privately since driving your car commercially is itself a blue collar job. Being a woman is very negatively correlated with driving a red car. Lastly of note, claim frequency is moderately correlated with higher past payouts, which adds up.

We have 14 numeric variables and 11 categorical variables (including the dummy variable `TARGET_FLAG`). We list the possible ranges or values for each variable in the breakdown below:

Variable	Type	Values
AGE	Numeric	16 - 81
BLUEBOOK	Numeric	1500 - 69740
CAR_AGE	Numeric	-3 - 28
CLM_FREQ	Numeric	0 - 5
HOME_VAL	Numeric	0 - 885282
HOMEKIDS	Numeric	0 - 5
INCOME	Numeric	0 - 367030
KIDSDRIV	Numeric	0 - 4
MVR_PTS	Numeric	0 - 13
OLDCLAIM	Numeric	0 - 57037
TARGET_AMT	Numeric	0 - 107586.1
TIF	Numeric	1 - 25
TRAVTIME	Numeric	5 - 142
YOJ	Numeric	0 - 23
CAR_TYPE	Categorical	Minivan, Panel Truck, Pickup, Sports Car, Van, z_SUV
CAR_USE	Categorical	Commercial, Private
EDUCATION	Categorical	<High School, Bachelors, Masters, PhD, z_High School
JOB	Categorical	Clerical, Doctor, Home Maker, Lawyer, Manager, Professional, Student, z_Blue Collar
MSTATUS	Categorical	Yes, z_No
PARENT1	Categorical	No, Yes
RED_CAR	Categorical	no, yes
REVOKED	Categorical	No, Yes
SEX	Categorical	M, z_F
TARGET_FLAG	Categorical	0, 1
URBANICITY	Categorical	Highly Urban/ Urban, z_Highly Rural/ Rural

The ranges for `TARGET_AMT`, `HOME_VAL`, `INCOME`, `KIDSDRIV`, and `HOMEKIDS` all include zero, and recoding these zero values as `NA` will make analyzing summary statistics for these variables more meaningful than if we included zeroes in their calculations. (We will maintain a separate copy of the data, in which we do not introduce additional `NA` values, for later use when creating the fully imputed dataset that some of our models will rely on for completeness.)

The range for `CAR_AGE` includes -3. Since the variable can only take positive or zero values logically, and only one observation in the dataset has a negative sign, we make the assumption that the age of 3 years is correct for this observation, and the sign is simply a data entry error. We fix this observation.

Some of the factor levels are named inconsistently, so we will rename and relevel them in the next section.

Let's take a look at the summary statistics for each variable.

```
##  TARGET_FLAG   TARGET_AMT          KIDSDRIV          AGE
##  0:6008      Min.   :    30.28   Min.   :1.000   Min.   :16.00
##  1:2153      1st Qu.:  2609.78   1st Qu.:1.000   1st Qu.:39.00
##              Median :  4104.00   Median :1.000   Median :45.00
##              Mean   :  5702.18   Mean   :1.423   Mean   :44.79
##              3rd Qu.:  5787.00   3rd Qu.:2.000   3rd Qu.:51.00
##              Max.   :107586.14   Max.   :4.000   Max.   :81.00
##              NA's   :6008        NA's   :7180    NA's   :6
```

```
##     HOMEKIDS          YOJ           INCOME       PARENT1       HOME_VAL
##  Min.   :1.000   Min.   : 0.0   Min.   :      5   No :7084   Min.   : 50223
##  1st Qu.:1.000   1st Qu.: 9.0   1st Qu.: 34135   Yes:1077   1st Qu.:153074
##  Median :2.000   Median :11.0   Median : 58438              Median :206692
##  Mean   :2.049   Mean   :10.5   Mean   : 67259              Mean   :220621
##  3rd Qu.:3.000   3rd Qu.:13.0   3rd Qu.: 90053              3rd Qu.:270023
##  Max.   :5.000   Max.   :23.0   Max.   :367030              Max.   :885282
##  NA's   :5289    NA's   :454    NA's   :1060                NA's   :2758
##  MSTATUS      SEX               EDUCATION              JOB
##  Yes :4894   M  :3786   <High School :1203   z_Blue Collar:1825
##  z_No:3267   z_F:4375   Bachelors    :2242   Clerical     :1271
##                         Masters      :1658   Professional :1117
##                         PhD          : 728   Manager      : 988
##                         z_High School:2330   Lawyer       : 835
##                                              (Other)      :1599
##                                              NA's         : 526
##     TRAVTIME            CAR_USE        BLUEBOOK          TIF
##  Min.   :  5.00   Commercial:3029   Min.   : 1500   Min.   : 1.000
##  1st Qu.: 22.00   Private   :5132   1st Qu.: 9280   1st Qu.: 1.000
##  Median : 33.00                     Median :14440   Median : 4.000
##  Mean   : 33.49                     Mean   :15710   Mean   : 5.351
##  3rd Qu.: 44.00                     3rd Qu.:20850   3rd Qu.: 7.000
##  Max.   :142.00                     Max.   :69740   Max.   :25.000
##
##         CAR_TYPE    RED_CAR       OLDCLAIM        CLM_FREQ      REVOKED
##  Minivan     :2145   no :5783   Min.   :    0   Min.   :0.0000   No :7161
##  Panel Truck: 676   yes:2378   1st Qu.:    0   1st Qu.:0.0000   Yes:1000
##  Pickup     :1389              Median :    0   Median :0.0000
##  Sports Car : 907              Mean   : 4037   Mean   :0.7986
##  Van        : 750              3rd Qu.: 4636   3rd Qu.:2.0000
##  z_SUV      :2294              Max.   :57037   Max.   :5.0000
##
##     MVR_PTS          CAR_AGE                    URBANICITY
##  Min.   : 0.000   Min.   : 0.000   Highly Urban/ Urban  :6492
##  1st Qu.: 0.000   1st Qu.: 1.000   z_Highly Rural/ Rural:1669
##  Median : 1.000   Median : 8.000
##  Mean   : 1.696   Mean   : 8.329
##  3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :13.000   Max.   :28.000
##                   NA's   :510
```

The majority of observations live/work in a highly urban or urban area. There are more married than unmarried observations, and there are also more female than male observations. The average observation has a median age of 45 years old, has been in their job for a median of 11 years, and has a median income of roughly $58,500.00. Most cars in the dataset are driven for private use rather than commercially, and the median car age is 8 years.

6,008 observations, which is the majority of observations, do not involve car crashes, and we now correctly record 6,008 `NA` observations for `TARGET_AMT`. (Since we introduced `NA` values for `TARGET_AMT` on purpose, we will not consider imputing them.)

There are 6 `NA` values in `AGE`, 510 in `CAR_AGE`, 454 in `YOJ`, 1,060 in `INCOME`, 2,758 in `HOME_VAL`, and 526 in `JOB`. In the next section, we will impute all these missing values in an alternate version of our dataset, as we mentioned earlier, and in the main version of our dataset, we will only impute the variables if we determine their data is at least Missing at Random (MAR), and there's no other evidence we should exclude them from imputation.

We check whether there is evidence that the data are Missing Completely at Random (MCAR), a higher standard than MAR, using the `mcar_test` function from the `naniar` package. Meeting this standard is unlikely with real data, but still worth checking.

statistic	df	p.value	missing.patterns
24554.84	2797	0	131

The low p-value provides evidence that missing data on these variables are **not** MCAR.

Excluding `AGE` since the number of missing values is so small for that variable, and we plan to impute it anyway, let's check whether missingness in any of the others is associated with any of the other predictors or the response variables using the `missing_compare` function from the `finalfit` package. Due to the large number of variables, we exclude any observed variables that could not account for a variable's missingness in the output by setting a p-value threshold of 0.05.

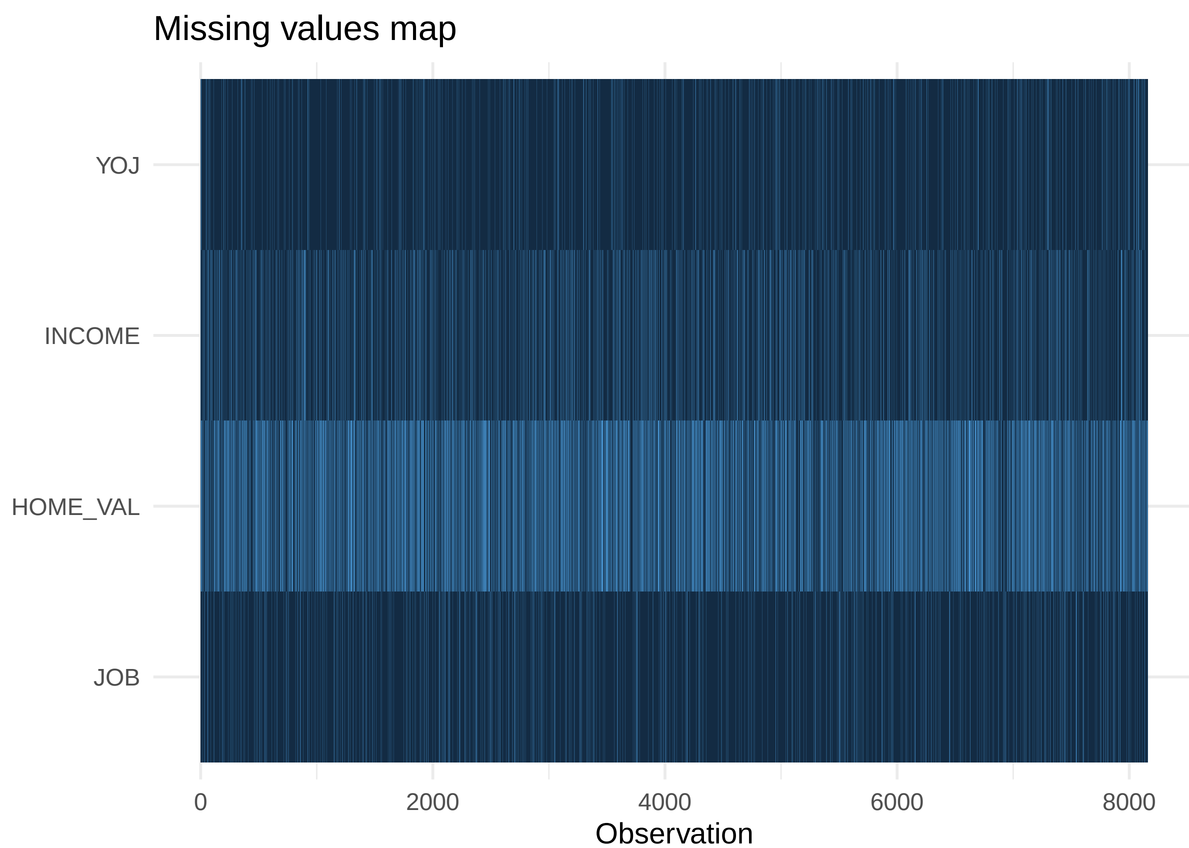
Dependant	Explanatory	Ref	Not Missing	Missing	p
INCOME	TARGET_FLAG	0	5308 (88.3)	700 (11.7)	0.001
INCOME	TARGET_FLAG	1	1793 (83.3)	360 (16.7)	NA
INCOME	AGE	Mean (SD)	44.9 (8.6)	43.8 (9.0)	0.001
INCOME	HOMEKIDS	Mean (SD)	2.0 (0.9)	2.1 (1.0)	0.014
INCOME	YOJ	Mean (SD)	11.4 (2.8)	4.3 (5.8)	0.001
INCOME	PARENT1	No	6188 (87.4)	896 (12.6)	0.022
INCOME	PARENT1	Yes	913 (84.8)	164 (15.2)	NA
INCOME	HOME_VAL	Mean (SD)	227842.0 (93771.4)	155319.7 (92741.6)	0.001
INCOME	SEX	M	3420 (90.3)	366 (9.7)	0.001
INCOME	SEX	z_F	3681 (84.1)	694 (15.9)	NA
INCOME	EDUCATION	<High School	982 (81.6)	221 (18.4)	0.001
INCOME	EDUCATION	Bachelors	1972 (88.0)	270 (12.0)	NA
INCOME	EDUCATION	Masters	1547 (93.3)	111 (6.7)	NA
INCOME	EDUCATION	PhD	652 (89.6)	76 (10.4)	NA
INCOME	EDUCATION	z_High School	1948 (83.6)	382 (16.4)	NA
INCOME	JOB	Clerical	1198 (94.3)	73 (5.7)	0.001
INCOME	JOB	Doctor	232 (94.3)	14 (5.7)	NA
INCOME	JOB	Home Maker	308 (48.0)	333 (52.0)	NA
INCOME	JOB	Lawyer	792 (94.9)	43 (5.1)	NA
INCOME	JOB	Manager	937 (94.8)	51 (5.2)	NA
INCOME	JOB	Professional	1055 (94.4)	62 (5.6)	NA
INCOME	JOB	Student	350 (49.2)	362 (50.8)	NA
INCOME	JOB	z_Blue Collar	1727 (94.6)	98 (5.4)	NA
INCOME	CAR_USE	Commercial	2675 (88.3)	354 (11.7)	0.008
INCOME	CAR_USE	Private	4426 (86.2)	706 (13.8)	NA
INCOME	BLUEBOOK	Mean (SD)	16199.2 (8430.5)	12432.0 (7574.9)	0.001
INCOME	TIF	Mean (SD)	5.4 (4.2)	5.1 (4.0)	0.045
INCOME	CAR_TYPE	Minivan	1922 (89.6)	223 (10.4)	0.001
INCOME	CAR_TYPE	Panel Truck	632 (93.5)	44 (6.5)	NA
INCOME	CAR_TYPE	Pickup	1225 (88.2)	164 (11.8)	NA
INCOME	CAR_TYPE	Sports Car	729 (80.4)	178 (19.6)	NA
INCOME	CAR_TYPE	Van	683 (91.1)	67 (8.9)	NA
INCOME	CAR_TYPE	z_SUV	1910 (83.3)	384 (16.7)	NA
INCOME	RED_CAR	no	4974 (86.0)	809 (14.0)	0.001
INCOME	RED_CAR	yes	2127 (89.4)	251 (10.6)	NA

Dependant	Explanatory	Ref	Not Missing	Missing	p
INCOME	CLM_FREQ	Mean (SD)	0.8 (1.2)	0.9 (1.2)	0.048
INCOME	CAR_AGE	Mean (SD)	8.5 (5.7)	7.2 (5.3)	0.001
INCOME	URBANICITY	Highly Urban/ Urban	5753 (88.6)	739 (11.4)	0.001
INCOME	URBANICITY	z_Highly Rural/ Rural	1348 (80.8)	321 (19.2)	NA
HOME_VAL	TARGET_FLAG	0	4217 (70.2)	1791 (29.8)	0.001
HOME_VAL	TARGET_FLAG	1	1186 (55.1)	967 (44.9)	NA
HOME_VAL	AGE	Mean (SD)	45.4 (8.5)	43.5 (8.7)	0.001
HOME_VAL	YOJ	Mean (SD)	11.1 (3.7)	9.3 (4.6)	0.001
HOME_VAL	INCOME	Mean (SD)	68771.2 (44434.0)	63968.7 (48518.0)	0.001
HOME_VAL	PARENT1	No	5055 (71.4)	2029 (28.6)	0.001
HOME_VAL	PARENT1	Yes	348 (32.3)	729 (67.7)	NA
HOME_VAL	MSTATUS	Yes	4267 (87.2)	627 (12.8)	0.001
HOME_VAL	MSTATUS	z_No	1136 (34.8)	2131 (65.2)	NA
HOME_VAL	EDUCATION	<High School	729 (60.6)	474 (39.4)	0.001
HOME_VAL	EDUCATION	Bachelors	1545 (68.9)	697 (31.1)	NA
HOME_VAL	EDUCATION	Masters	1166 (70.3)	492 (29.7)	NA
HOME_VAL	EDUCATION	PhD	474 (65.1)	254 (34.9)	NA
HOME_VAL	EDUCATION	z_High School	1489 (63.9)	841 (36.1)	NA
HOME_VAL	JOB	Clerical	913 (71.8)	358 (28.2)	0.001
HOME_VAL	JOB	Doctor	154 (62.6)	92 (37.4)	NA
HOME_VAL	JOB	Home Maker	456 (71.1)	185 (28.9)	NA
HOME_VAL	JOB	Lawyer	596 (71.4)	239 (28.6)	NA
HOME_VAL	JOB	Manager	703 (71.2)	285 (28.8)	NA
HOME_VAL	JOB	Professional	817 (73.1)	300 (26.9)	NA
HOME_VAL	JOB	Student	100 (14.0)	612 (86.0)	NA
HOME_VAL	JOB	z_Blue Collar	1309 (71.7)	516 (28.3)	NA
HOME_VAL	CAR_USE	Commercial	1942 (64.1)	1087 (35.9)	0.002
HOME_VAL	CAR_USE	Private	3461 (67.4)	1671 (32.6)	NA
HOME_VAL	BLUEBOOK	Mean (SD)	16073.5 (8388.1)	14997.6 (8437.5)	0.001
HOME_VAL	OLDCLAIM	Mean (SD)	3726.1 (8512.2)	4646.3 (9245.6)	0.001
HOME_VAL	CLM_FREQ	Mean (SD)	0.7 (1.1)	0.9 (1.2)	0.001
HOME_VAL	REVOKED	No	4801 (67.0)	2360 (33.0)	0.001
HOME_VAL	REVOKED	Yes	602 (60.2)	398 (39.8)	NA
HOME_VAL	MVR_PTS	Mean (SD)	1.6 (2.0)	1.9 (2.3)	0.001
HOME_VAL	CAR_AGE	Mean (SD)	8.4 (5.7)	8.1 (5.7)	0.012
HOME_VAL	URBANICITY	Highly Urban/ Urban	4345 (66.9)	2147 (33.1)	0.007
HOME_VAL	URBANICITY	z_Highly Rural/ Rural	1058 (63.4)	611 (36.6)	NA
JOB	AGE	Mean (SD)	44.7 (8.7)	46.5 (8.0)	0.001
JOB	HOMEKIDS	Mean (SD)	2.1 (0.9)	1.9 (0.9)	0.040
JOB	YOJ	Mean (SD)	10.4 (4.2)	11.3 (2.7)	0.001
JOB	INCOME	Mean (SD)	63334.1 (42157.2)	118852.9 (58861.8)	0.001
JOB	PARENT1	No	6601 (93.2)	483 (6.8)	0.001
JOB	PARENT1	Yes	1034 (96.0)	43 (4.0)	NA
JOB	HOME_VAL	Mean (SD)	213485.5 (89924.5)	322080.5 (121344.9)	0.001
JOB	SEX	M	3365 (88.9)	421 (11.1)	0.001
JOB	SEX	z_F	4270 (97.6)	105 (2.4)	NA
JOB	EDUCATION	<High School	1203 (100.0)	0 (0.0)	0.001
JOB	EDUCATION	Bachelors	2242 (100.0)	0 (0.0)	NA
JOB	EDUCATION	Masters	1330 (80.2)	328 (19.8)	NA
JOB	EDUCATION	PhD	530 (72.8)	198 (27.2)	NA
JOB	EDUCATION	z_High School	2330 (100.0)	0 (0.0)	NA
JOB	CAR_USE	Commercial	2557 (84.4)	472 (15.6)	0.001

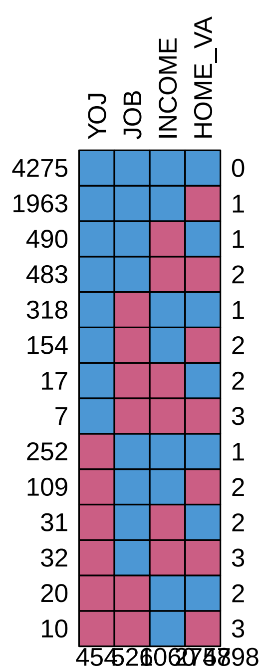
Dependant	Explanatory	Ref	Not Missing	Missing	p
JOB	CAR_USE	Private	5078 (98.9)	54 (1.1)	NA
JOB	BLUEBOOK	Mean (SD)	15161.5 (8018.6)	23669.5 (9952.7)	0.001
JOB	CAR_TYPE	Minivan	2123 (99.0)	22 (1.0)	0.001
JOB	CAR_TYPE	Panel Truck	435 (64.3)	241 (35.7)	NA
JOB	CAR_TYPE	Pickup	1265 (91.1)	124 (8.9)	NA
JOB	CAR_TYPE	Sports Car	902 (99.4)	5 (0.6)	NA
JOB	CAR_TYPE	Van	634 (84.5)	116 (15.5)	NA
JOB	CAR_TYPE	z_SUV	2276 (99.2)	18 (0.8)	NA
JOB	RED_CAR	no	5510 (95.3)	273 (4.7)	0.001
JOB	RED_CAR	yes	2125 (89.4)	253 (10.6)	NA
JOB	OLDCLAIM	Mean (SD)	3980.4 (8722.8)	4859.5 (9501.7)	0.026
JOB	CLM_FREQ	Mean (SD)	0.8 (1.2)	1.0 (1.3)	0.001
JOB	CAR_AGE	Mean (SD)	7.9 (5.6)	14.0 (4.6)	0.001
JOB	URBANICITY	Highly Urban/ Urban	5987 (92.2)	505 (7.8)	0.001
JOB	URBANICITY	z_Highly Rural/ Rural	1648 (98.7)	21 (1.3)	NA

There is evidence that some of the missingness for `INCOME`, `HOME_VAL`, and `JOB` can be explained by other observed information, so they could be considered Missing at Random (MAR). There is no evidence missing values for `CAR_AGE` or `YOJ` can be explained by other observed information, so we will no longer consider imputing them in the main version of our dataset.

It's reasonable to assume that the missing values in `YOJ`, `HOME_VAL`, `INCOME` and `JOB` might all be related because money, employment, and assets are interconnected. Therefore the missingness of one or more of these variables might be dependent on the missingness of one or more of the others. Let's look at the overlap of observations with missing values for these variables using the `missing_plot` function from the `finalfit` package.

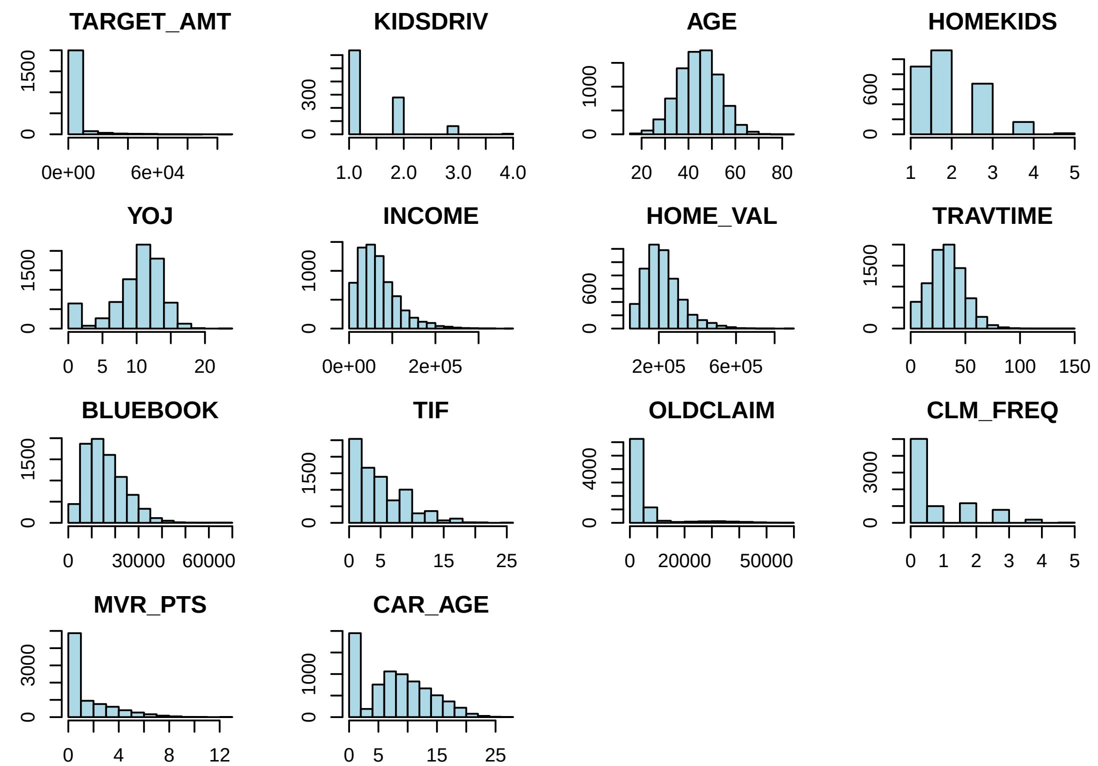


We do see some overlap in the observations that have missing values for these variables, but it's hard to detect anything more conclusive from this plot. To take a closer look at the patterns of missingness between these variables, we can use the `missing_pattern` function from the `finalfit` package.



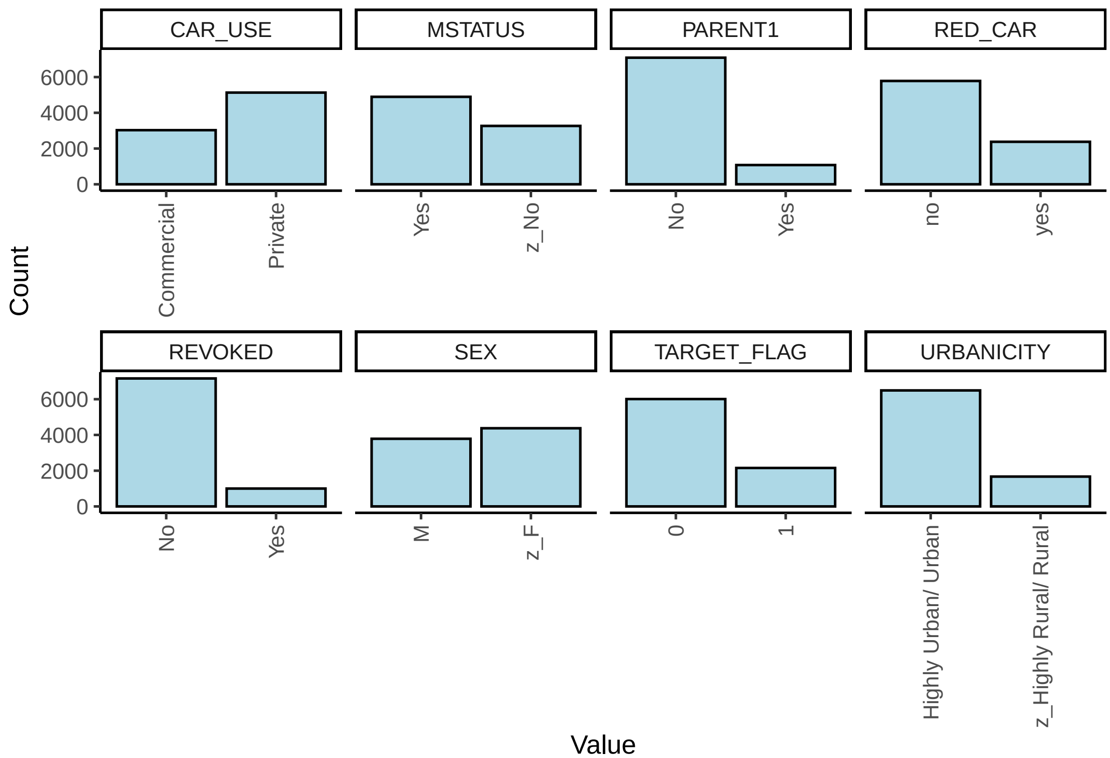
Here, we see several patterns of missingness worth noting. 814 observations are missing two out of these four variables, and 49 observations are missing three. Of the observations that are missing `HOME_VAL`, 483 are also missing `INCOME`, 154 are also missing `JOB`, and 109 are also missing `YOJ`. Due to these patterns of related missingness, we will no longer consider imputing these variables in the main version of our dataset.

Let's take a look at the distributions of the numeric variables.



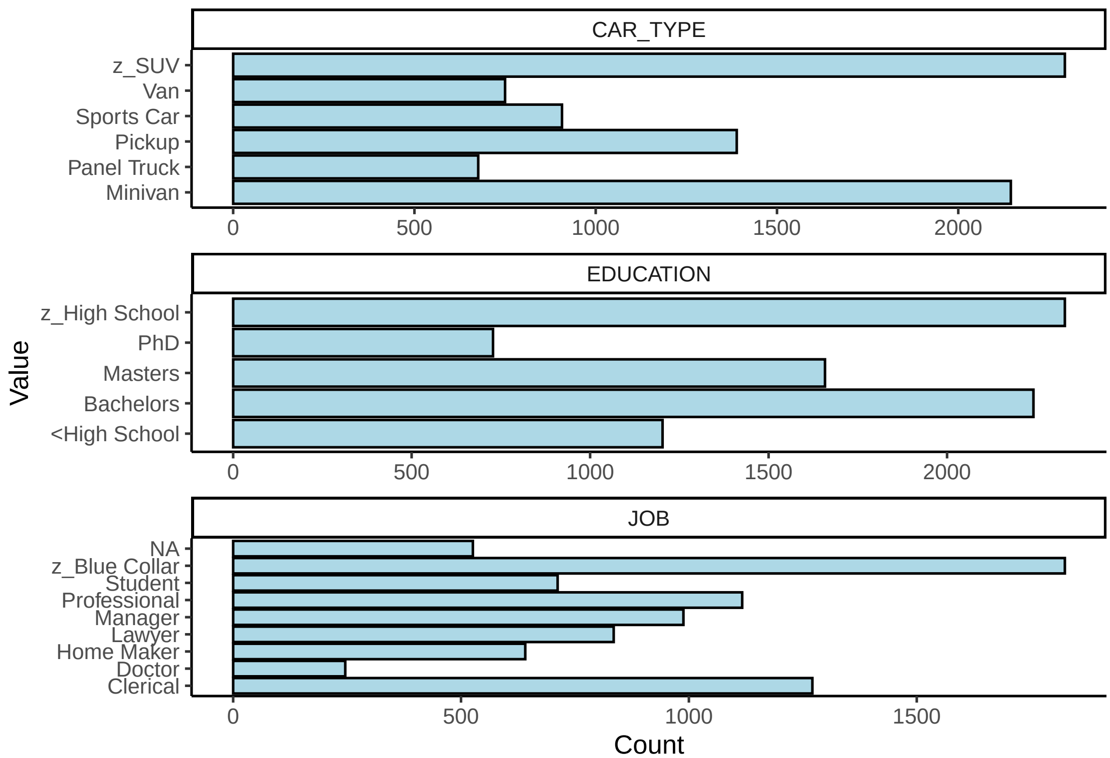
The distribution for `AGE` is approximately normal. The distribution for `YOJ` is left-skewed. The distributions for `TARGET_AMT`, `KIDSDRIV`, `HOMEKIDS`, `INCOME`, `HOME_VAL`, `TRAVTIME`, `BLUEBOOK`, `TIF`, `OLDCLAIM`, `CLM_FREQ`, `MVR_PTS`, and `CAR_AGE` are all right-skewed. 75% of observations for `TARGET_AMT` are at or below $5,787.00, but the maximum value recorded is $107,586.14.

Let's also take a look at the distributions of the categorical variables. First, we look at the distributions for categorical variables with only two levels.



Looking at `PARENT1` and `REVOKED`, we can see that single parents represent relatively few observations in the dataset, as do people whose licenses were revoked in the past seven years. `MSTATUS` and `SEX` are the most evenly split categorical variables with two levels in the dataset.

Next we look at the distributions for the categorical variables with more than two levels.



The most common profession represented in the observations is blue collar, and the most commonly represented cars are the SUV and the minivan. The number of observations with high school diplomas and bachelor's degrees are fairly similar. Having less or more education is less common.

Data Preparation

First, we rename and relevel the inconsistently named and leveled factor variables we noted earlier. A summary of only the factors we changed the levels for is below, with the first level in each list always being the reference level. For variables which have "Yes" and "No" values, we will replace these with 1/0 (1 = "Yes", 0 = "No").

Factor	New Levels
CAR_TYPE	Minivan, Panel Truck, Pickup, Sports Car, SUV, Van
EDUCATION	<High School, High School, Bachelors, Masters, PhD
JOB	Blue Collar, Clerical, Doctor, Home Maker, Lawyer, Manager, Professional, Student
PARENT1	0, 1
MSTATUS	0, 1
RED_CAR	0, 1
REVOKED	0, 1
SEX	Male, Female
URBANICITY	0, 1

We reduce the scale of the `INCOME` and `HOME_VAL` variables to thousands of dollars so the figures will be more readable when visualized. The replacement variables are `INCOME_THOU` and `HOME_VAL_THOU`.

Some observations list `Student` as their occupation as well as a value for `YOJ`. We recode these values as `NA`. The most likely interpretation is that people incorrectly listed how many years they've been in school here, which will not be useful to our analysis.

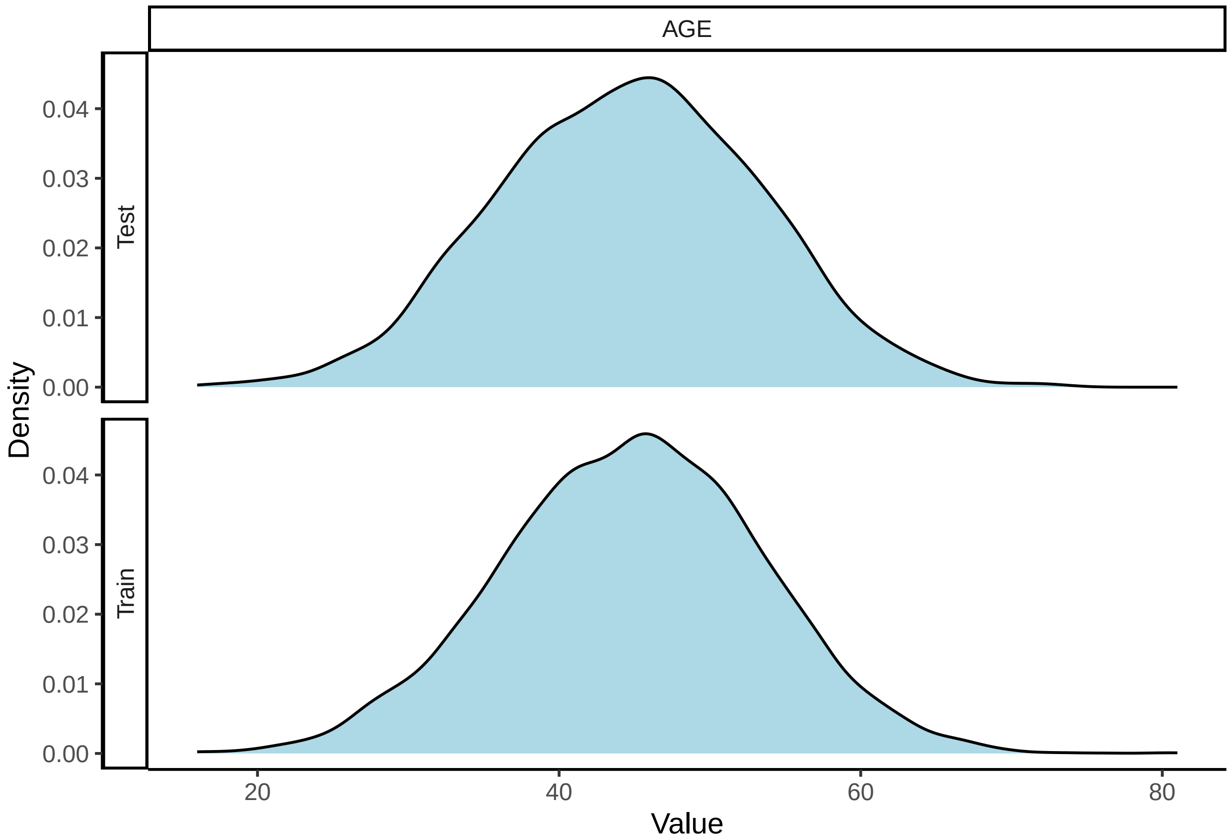
Based on the descriptions of some of the variables and their theoretical effects on the target variables, and to handle the variables that have missing data that we chose not to impute, including those for which we replaced zero or incorrect values with `NA` values, we create several factors that we believe will be helpful when building models:

- `HOME_VAL_CAT` (Levels based on `HOME_VAL_THOU` = "<=250K", "251-500K", "501-750K", "751K+", "")
- `HOMEOWNER` (1 = `HOME_VAL_THOU` not NA)
- `INCOME_CAT` (Levels based on `INCOME` = "<=50K", "51-100K", "101-150K", "151K+", "")
- `INCOME_FLAG` (1 = `INCOME_THOU` not NA)
- `KIDSDRIV_FLAG` (1 = `KIDSDRIV` number of children not `NA`)
- `HOMEKIDS_FLAG` (1 = `HOMEKIDS` number of children not `NA`)
- `EMPLOYED` (1 = `JOB` not NA/Student/Home Maker)
- `CAR_AGE_CAT` (Levels based on `CAR_AGE` = "<=4", "5-8", "9-12", "13+", "")
- `WHITE_COLLAR` (1 = `JOB` not NA/Student/Home Maker/Blue Collar)

We then split both the main version of our dataset and the alternate version we created earlier into train and test sets. The main version will have all the derived variables we just created, imputed values for the `AGE` variable, and any transformations we make. The alternate version will not include any derived variables or transformations, but it will include imputed values for all variables with missing values.

We impute missing data in the main train and test sets for one numeric variable, `AGE`, using the mean value since it is normally distributed.

We take a look at the distributions for our imputed variable to see if the distributions of this variable in the train and test sets differ from what we originally observed or between sets.



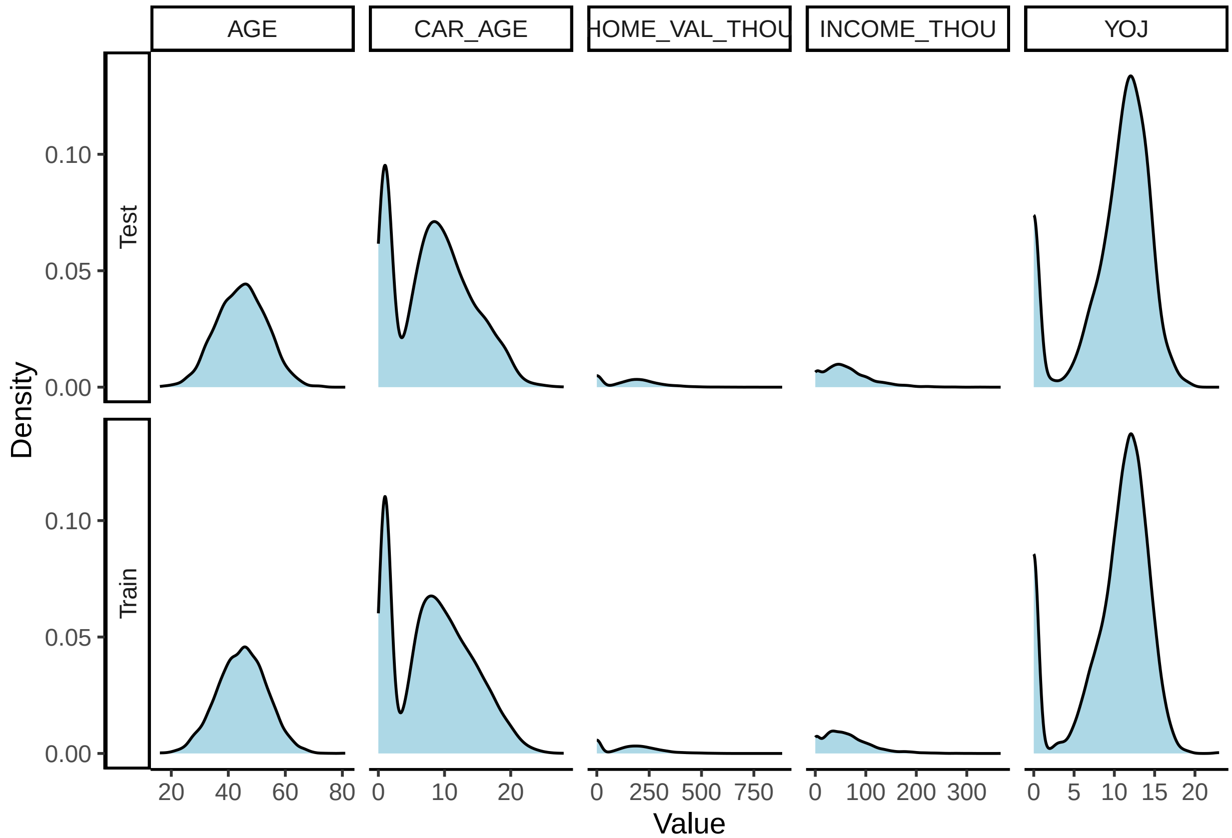
The distributions in the train and test sets for `AGE` are similar to each other and to its original distribution.

We impute missing data in the alternate train and test sets for all variables with missing values using the `mice` package.

We confirm there are no longer any missing values in the alternate train or test datasets.

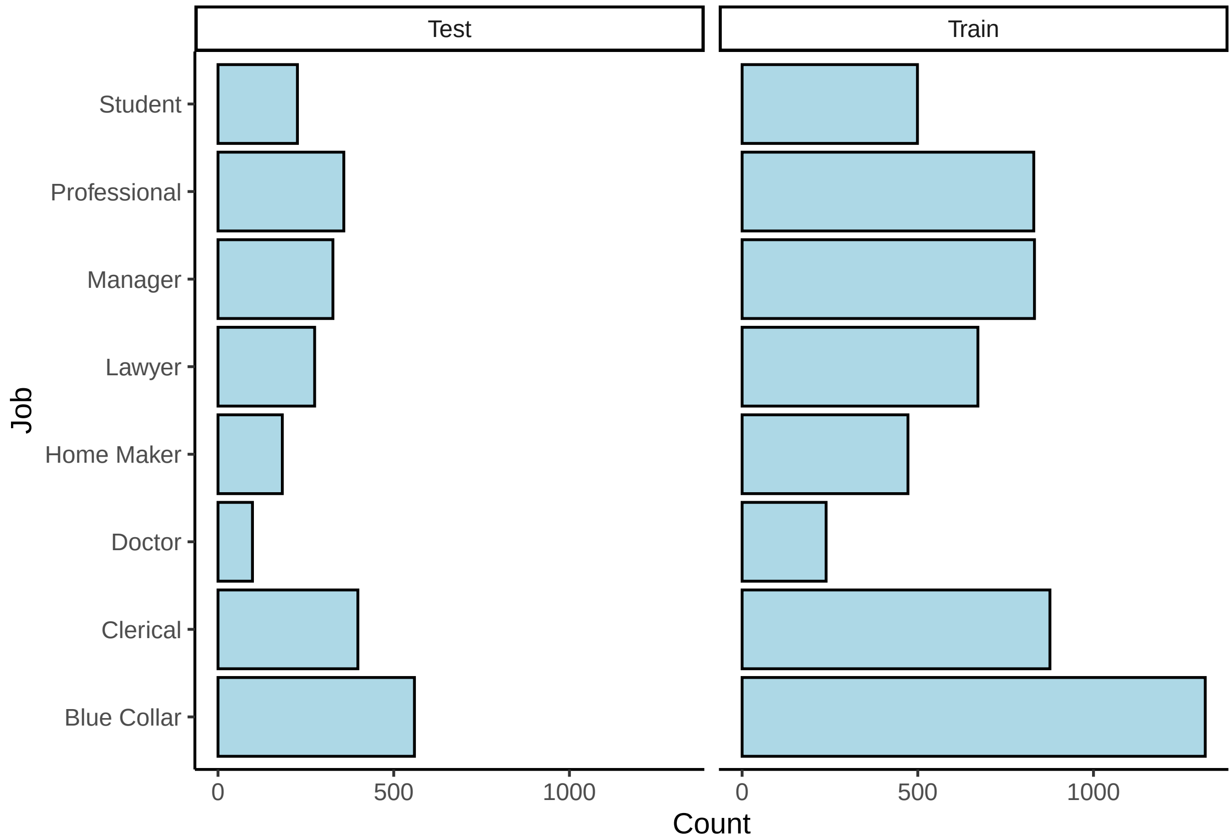
```
## [1] TRUE
```

We take a look at the distributions for the imputed numeric variables to see if their distributions in the alternate train and test sets differ from what we originally observed or between sets.



The distributions for the imputed numeric variables don't differ between the alternate train and test sets or from what we originally observed.

We also perform the same check for the single categorical variable we imputed in the alternate train and test sets: `JOB`.



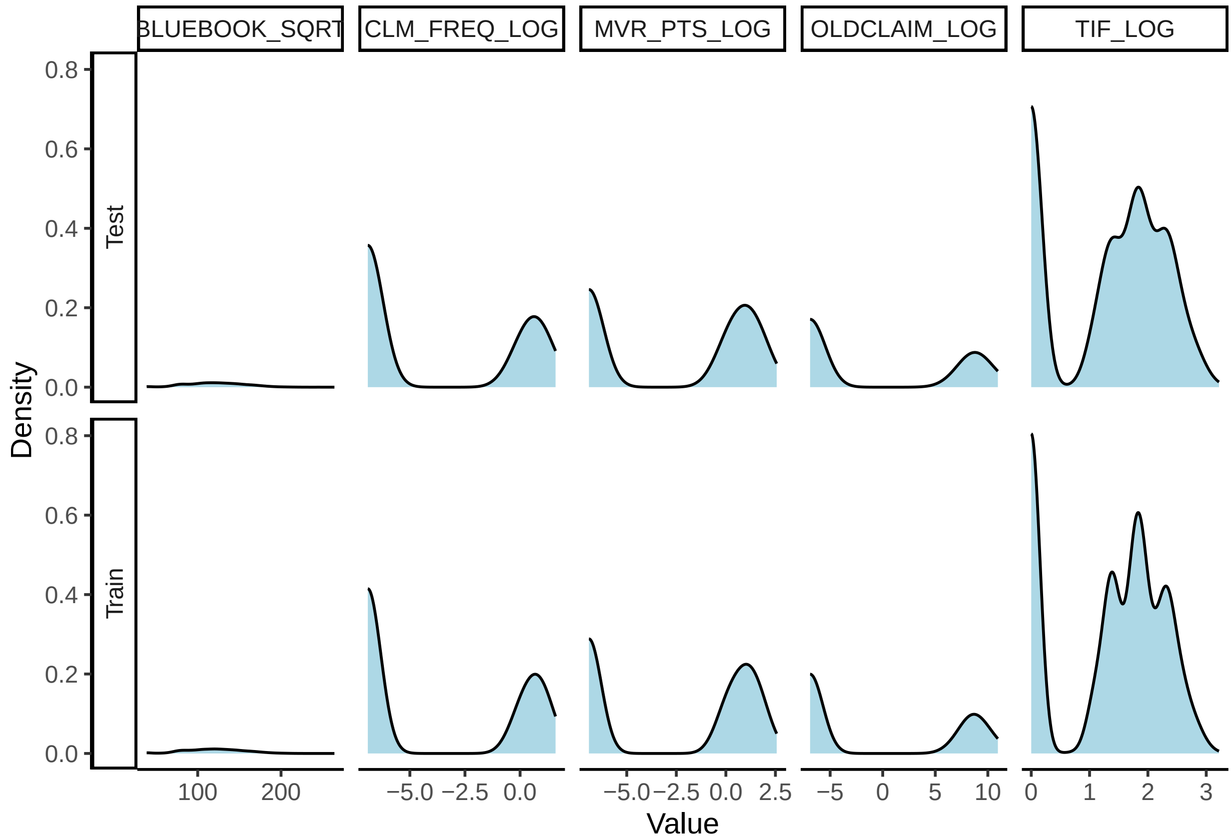
The distributions in the alternate train and test sets for the single imputed categorical variable, `JOB`, are similar to each other, and the rankings of most frequent to least frequent occupation here are similar to the rankings of the original distribution. We note that the "Professional" and "Manager" occupations are more tied in the rankings here than they were in the original distribution, however.

Since the distributions of some of our numeric variables are skewed, we transform the data for some of them in the main dataset, but not the alternate dataset. We exclude any numeric variables with missing values that we decided not to impute in the main dataset and for which we have already created factors. We exclude the response variable `TARGET_AMT` as well.

Below is a breakdown of the variables, the ideal labmdas proposed by Box-Cox, and the reasonable alternative transformations we have chosen to make instead:

Skewed Variable	Ideal Lambda Proposed by Box-Cox	Reasonable Alternative Transformation
TRAVTIME	0.7	no transformation
BLUEBOOK	0.45	square root
TIF	0.25	log
OLDCLAIM	-0.0999999999999999	log
CLM_FREQ	-0.2	log
MVR_PTS	0.0500000000000003	log

We check whether the distributions of the transformed variables now differ between the train and test sets.



They do not.

Build Models

Binary Logistic Regression Models

Model BLR:1 - Full Model Using Original, Untransformed Variables, with All Missing Values Imputed We create Model BLR:1, our baseline binary logistic regression model based on all the original, untransformed variables, with all missing values imputed so that no observations or predictors have to be excluded from the model. That way, it can truly be considered the full binary logistic regression model.

A summary of Model BLR:1 is below:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = dat)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.612e+00  3.413e-01  -7.654 1.95e-14 ***
## KIDSDRIV              4.282e-01  7.277e-02   5.885 3.98e-09 ***
## AGE                  -3.843e-03  4.854e-03  -0.792  0.42846
## HOMEKIDS              4.446e-02  4.432e-02   1.003  0.31577
## YOJ                   7.564e-03  1.180e-02   0.641  0.52150
## PARENT11              3.342e-01  1.311e-01   2.550  0.01077 *
```

```
## MSTATUS1            -5.305e-01  1.024e-01  -5.183 2.19e-07 ***
## SEXFemale            4.383e-03  1.356e-01   0.032  0.97422
## EDUCATIONHigh School -2.892e-01  1.384e-01  -2.090  0.03664 *
## EDUCATIONBachelors   -3.307e-02  1.132e-01  -0.292  0.77015
## EDUCATIONMasters     -2.037e-01  2.059e-01  -0.989  0.32243
## EDUCATIONPhD          1.065e-01  2.457e-01   0.433  0.66479
## JOBClerical           1.228e-01  1.268e-01   0.969  0.33260
## JOBDoctor            -7.597e-01  2.862e-01  -2.654  0.00795 **
## JOBHome Maker         4.684e-02  1.813e-01   0.258  0.79609
## JOBLawyer            -5.931e-02  1.987e-01  -0.299  0.76529
## JOBManager           -8.115e-01  1.541e-01  -5.266 1.40e-07 ***
## JOBProfessional      -2.022e-01  1.385e-01  -1.460  0.14420
## JOBStudent           -1.286e-01  1.954e-01  -0.658  0.51052
## TRAVTIME              1.643e-02  2.251e-03   7.298 2.92e-13 ***
## CAR_USEPrivate       -7.838e-01  1.051e-01  -7.458 8.77e-14 ***
## BLUEBOOK             -1.735e-05  6.289e-06  -2.760  0.00579 **
## TIF                  -5.548e-02  8.950e-03  -6.199 5.67e-10 ***
## CAR_TYPEPanel Truck   5.273e-01  1.936e-01   2.723  0.00646 **
## CAR_TYPEPickup        5.281e-01  1.201e-01   4.397 1.10e-05 ***
## CAR_TYPESports Car    1.008e+00  1.546e-01   6.521 6.99e-11 ***
## CAR_TYPESUV           7.263e-01  1.342e-01   5.413 6.20e-08 ***
## CAR_TYPEVan           6.445e-01  1.502e-01   4.290 1.79e-05 ***
## RED_CAR1              7.780e-02  1.048e-01   0.743  0.45777
## OLDCLAIM             -1.504e-05  4.691e-06  -3.207  0.00134 **
## CLM_FREQ              1.841e-01  3.387e-02   5.434 5.50e-08 ***
## REVOKED1              9.243e-01  1.105e-01   8.362  < 2e-16 ***
## MVR_PTS               1.310e-01  1.628e-02   8.046 8.56e-16 ***
## CAR_AGE              -1.419e-02  9.023e-03  -1.573  0.11583
## URBANICITY1           2.383e+00  1.337e-01  17.831  < 2e-16 ***
## INCOME_THOU          -4.364e-03  1.343e-03  -3.250  0.00115 **
## HOME_VAL_THOU        -1.368e-03  4.171e-04  -3.280  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6640.0  on 5736  degrees of freedom
## Residual deviance: 5105.5  on 5700  degrees of freedom
## AIC: 5179.5
##
## Number of Fisher Scoring iterations: 5
```

The AIC of Model BLR:1 is 5179.5.

Model BLR:2 - Select Model Using Original & Derived, but Untransformed Variables, with Only `AGE` Values Imputed We create Model BLR:2, a second binary logistic regression model based on variables we believe will be the best predictors of `TARGET_FLAG`, including some original variables and some variables we derived from other variables, but no transformed variables. The only value we've imputed for this model is `AGE`.

A summary of Model BLR:2 is below:

```
##
```

```
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CLM_FREQ + HOMEOWNER + INCOME_FLAG +
##     EMPLOYED + WHITE_COLLAR + MSTATUS + PARENT1 + REVOKED + SEX +
##     TRAVTIME, family = "binomial", data = train_df_imputed)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.593521   0.224284  -2.646 0.008138 **
## AGE           -0.014439   0.003959  -3.647 0.000266 ***
## CLM_FREQ       0.374499   0.025589  14.635  < 2e-16 ***
## HOMEOWNER1    -0.262771   0.081438  -3.227 0.001253 **
## INCOME_FLAG1  -0.478040   0.092563  -5.164 2.41e-07 ***
## EMPLOYED1      0.473987   0.107889   4.393 1.12e-05 ***
## WHITE_COLLAR1 -0.644178   0.075688  -8.511  < 2e-16 ***
## MSTATUS1      -0.241994   0.084976  -2.848 0.004402 **
## PARENT11       0.487306   0.103513   4.708 2.51e-06 ***
## REVOKED1       0.904666   0.087874  10.295  < 2e-16 ***
## SEXFemale      0.110879   0.065567   1.691 0.090822 .
## TRAVTIME       0.008946   0.001991   4.494 6.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6640  on 5736  degrees of freedom
## Residual deviance: 5993  on 5725  degrees of freedom
## AIC: 6017
##
## Number of Fisher Scoring iterations: 4
```

The AIC of Model BLR:2 is 6017.

Model BLR:3 - Select Model Using Original, Derived, & Transformed Variables, with Only `AGE` Values Imputed We create Model BLR:3, a third binary logistic regression model based on variables we believe will be the best predictors of `TARGET_FLAG`, including some original variables, some variables we derived from other variables, and some variables we transformed. The only value we've imputed for this model is `AGE`.

```
##  [1] "AGE"           "CLM_FREQ_LOG"  "URBANICITY"    "MVR_PTS_LOG"
##  [5] "OLDCLAIM_LOG"  "PARENT1"       "REVOKED"       "CAR_USE"
##  [9] "CAR_TYPE"      "MSTATUS"       "EDUCATION"     "KIDSDRIV_FLAG"
## [13] "INCOME_CAT"    "EMPLOYED"
```

In choosing some of these variables, we excluded others for which collinearity might be an issue. That is, our factor describing income was chosen over the home value factor, the kids driving factor was chosen over the kids at home factor, and the education factor was chosen over the job factor.

A summary of Model BLR:3 is below:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = dat)
##
```

```
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.837748   0.375976  -2.228 0.025867 *
## AGE                   -0.009020   0.004273  -2.111 0.034759 *
## CLM_FREQ_LOG           0.248147   0.073986   3.354 0.000797 ***
## URBANICITY1            2.083986   0.130205  16.005  < 2e-16 ***
## MVR_PTS_LOG            0.062018   0.009233   6.717 1.86e-11 ***
## OLDCLAIM_LOG          -0.088232   0.035689  -2.472 0.013428 *
## PARENT11               0.309247   0.115743   2.672 0.007544 **
## REVOKED1               0.862490   0.099081   8.705  < 2e-16 ***
## CAR_USEPrivate        -0.796227   0.087558  -9.094  < 2e-16 ***
## CAR_TYPEPanel Truck    0.086566   0.155030   0.558 0.576585
## CAR_TYPEPickup         0.495521   0.114622   4.323 1.54e-05 ***
## CAR_TYPESports Car     1.013510   0.122824   8.252  < 2e-16 ***
## CAR_TYPESUV            0.756772   0.099014   7.643 2.12e-14 ***
## CAR_TYPEVan            0.471826   0.138083   3.417 0.000633 ***
## MSTATUS1              -0.630560   0.080548  -7.828 4.94e-15 ***
## EDUCATIONHigh School -0.214308   0.107842  -1.987 0.046895 *
## EDUCATIONBachelors    -0.767253   0.118015  -6.501 7.96e-11 ***
## EDUCATIONMasters      -0.858061   0.129548  -6.623 3.51e-11 ***
## EDUCATIONPhD          -1.165922   0.162958  -7.155 8.38e-13 ***
## KIDSDRIV_FLAG1         0.732433   0.101027   7.250 4.17e-13 ***
## INCOME_CAT.L           0.098315   0.076196   1.290 0.196948
## INCOME_CAT.Q           0.362249   0.087949   4.119 3.81e-05 ***
## INCOME_CAT.C           0.240618   0.086842   2.771 0.005593 **
## EMPLOYED1              0.007841   0.102295   0.077 0.938902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6640.0  on 5736  degrees of freedom
## Residual deviance: 5340.5  on 5713  degrees of freedom
## AIC: 5388.5
##
## Number of Fisher Scoring iterations: 5
```

We remove the only statistically insignificant variable, `EMPLOYED`, and print a new summary.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CLM_FREQ_LOG + URBANICITY +
##     MVR_PTS_LOG + OLDCLAIM_LOG + PARENT1 + REVOKED + CAR_USE +
##     CAR_TYPE + MSTATUS + EDUCATION + KIDSDRIV_FLAG + INCOME_CAT,
##     family = "binomial", data = dat)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.833765   0.372358  -2.239 0.025146 *
## AGE                   -0.008989   0.004253  -2.113 0.034565 *
## CLM_FREQ_LOG           0.248118   0.073983   3.354 0.000797 ***
## URBANICITY1            2.084292   0.130147  16.015  < 2e-16 ***
## MVR_PTS_LOG            0.062019   0.009233   6.717 1.86e-11 ***
## OLDCLAIM_LOG          -0.088226   0.035688  -2.472 0.013431 *
```

```
## PARENT11              0.309041   0.115712   2.671 0.007568 **
## REVOKED1              0.862381   0.099071   8.705  < 2e-16 ***
## CAR_USEPrivate       -0.794845   0.085678  -9.277  < 2e-16 ***
## CAR_TYPEPanel Truck   0.085585   0.154503   0.554 0.579621
## CAR_TYPEPickup        0.495262   0.114573   4.323 1.54e-05 ***
## CAR_TYPESports Car    1.013516   0.122825   8.252  < 2e-16 ***
## CAR_TYPESUV           0.756753   0.099015   7.643 2.13e-14 ***
## CAR_TYPEVan           0.471640   0.138060   3.416 0.000635 ***
## MSTATUS1             -0.630675   0.080533  -7.831 4.83e-15 ***
## EDUCATIONHigh School -0.213838   0.107666  -1.986 0.047018 *
## EDUCATIONBachelors   -0.766369   0.117448  -6.525 6.79e-11 ***
## EDUCATIONMasters     -0.858853   0.129140  -6.651 2.92e-11 ***
## EDUCATIONPhD         -1.166704   0.162646  -7.173 7.32e-13 ***
## KIDSDRIV_FLAG1        0.732549   0.101016   7.252 4.11e-13 ***
## INCOME_CAT.L          0.097117   0.074578   1.302 0.192837
## INCOME_CAT.Q          0.361047   0.086543   4.172 3.02e-05 ***
## INCOME_CAT.C          0.240450   0.086813   2.770 0.005610 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6640.0  on 5736  degrees of freedom
## Residual deviance: 5340.5  on 5714  degrees of freedom
## AIC: 5386.5
##
## Number of Fisher Scoring iterations: 5
```

The AIC of Model BLR:3 is 5386.5.

Multiple Linear Regression Models

Model MLR:1 - Full Model Using Original, Untransformed Variables, with All Missing Values Imputed We create Model MLR:1, our baseline multiple linear regression model based on all the original, untransformed variables, with all missing values imputed so that no observations or predictors have to be excluded from the model. That way, it can truly be considered the full multiple linear regression model.

A summary of Model MLR:1 is below:

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5380  -1719   -771    364 103426
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -8.203e+01  5.872e+02  -0.140 0.888897
## KIDSDRIV              4.166e+02  1.412e+02   2.950 0.003189 **
## AGE                   4.046e+00  8.836e+00   0.458 0.647081
## HOMEKIDS              5.517e+01  8.154e+01   0.677 0.498686
```

```
## YOJ                  2.623e+01  2.117e+01   1.239 0.215447
## PARENT11             4.698e+02  2.511e+02   1.871 0.061440 .
## MSTATUS1            -6.204e+02  1.839e+02  -3.374 0.000746 ***
## SEXFemale            4.234e+02  2.316e+02   1.828 0.067592 .
## EDUCATIONHigh School -1.456e+02  2.560e+02  -0.569 0.569424
## EDUCATIONBachelors  -9.881e+01  2.134e+02  -0.463 0.643330
## EDUCATIONMasters     1.638e+01  3.670e+02   0.045 0.964411
## EDUCATIONPhD         4.306e+02  4.360e+02   0.988 0.323392
## JOBClerical         -1.963e+02  2.377e+02  -0.826 0.408949
## JOBDoctor           -7.119e+02  4.729e+02  -1.505 0.132309
## JOBHome Maker       -5.808e+01  3.318e+02  -0.175 0.861075
## JOBLawyer           -8.640e+01  3.554e+02  -0.243 0.807958
## JOBManager          -9.957e+02  2.739e+02  -3.636 0.000280 ***
## JOBProfessional      2.915e+01  2.569e+02   0.113 0.909653
## JOBStudent          -7.581e+01  3.613e+02  -0.210 0.833827
## TRAVTIME             1.491e+01  4.020e+00   3.710 0.000209 ***
## CAR_USEPrivate      -8.743e+02  1.957e+02  -4.467 8.08e-06 ***
## BLUEBOOK             1.735e-02  1.084e-02   1.600 0.109563
## TIF                 -3.873e+01  1.540e+01  -2.515 0.011931 *
## CAR_TYPEPanel Truck  1.570e+02  3.456e+02   0.454 0.649537
## CAR_TYPEPickup       3.191e+02  2.119e+02   1.506 0.132124
## CAR_TYPESports Car   9.492e+02  2.714e+02   3.497 0.000474 ***
## CAR_TYPESUV          6.837e+02  2.268e+02   3.014 0.002588 **
## CAR_TYPEVan          4.372e+02  2.637e+02   1.658 0.097366 .
## RED_CAR1            -1.423e+02  1.867e+02  -0.762 0.445856
## OLDCLAIM            -1.399e-02  9.219e-03  -1.517 0.129307
## CLM_FREQ             1.439e+02  6.787e+01   2.120 0.034091 *
## REVOKED1             4.954e+02  2.181e+02   2.271 0.023154 *
## MVR_PTS              1.888e+02  3.205e+01   5.889 4.10e-09 ***
## CAR_AGE             -3.505e+01  1.600e+01  -2.192 0.028452 *
## URBANICITY1          1.683e+03  1.737e+02   9.693  < 2e-16 ***
## INCOME_THOU         -5.338e+00  2.342e+00  -2.279 0.022682 *
## HOME_VAL_THOU       -8.745e-01  7.551e-01  -1.158 0.246855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4740 on 5700 degrees of freedom
## Multiple R-squared:  0.06957,    Adjusted R-squared:  0.06369
## F-statistic: 11.84 on 36 and 5700 DF,  p-value: < 2.2e-16
```

Model MLR:2 - Select Model Using Original & Derived, but Untransformed Variables, with Only `AGE` Values Imputed We create Model MLR:2, a second multiple linear regression model based on variables we believe will be the best predictors of `TARGET_AMT` due to their definitions and theories regarding their impact, including some original variables, some variables we derived from other variables, but no transformed or imputed variables.

A summary of Model MLR:2 is below:

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK + MVR_PTS + REVOKED, data = train_df_imputed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
##  -8517  -3188  -1647    349 100664
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3860.49477  487.87936   7.913 4.81e-15 ***
## BLUEBOOK       0.12056    0.02556   4.716 2.62e-06 ***
## MVR_PTS      125.50920   80.16877   1.566   0.1177
## REVOKED1    -896.59669  511.47017  -1.753   0.0798 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8087 on 1519 degrees of freedom
##   (4214 observations deleted due to missingness)
## Multiple R-squared:  0.01807,    Adjusted R-squared:  0.01613
## F-statistic: 9.319 on 3 and 1519 DF,  p-value: 4.177e-06
```

Model MLR:3 - Select Model Using Original, Derived, & Transformed Variables, with No Imputed Variables We create Model MLR:3, a third multiple linear regression model based on variables we believe will be the best predictors of `TARGET_AMT`, including some original variables, some variables we derived from other variables, some variables we transformed, and some variables we imputed.

```
## [1] "BLUEBOOK_SQRT" "CAR_AGE_CAT"   "CAR_TYPE"      "CAR_USE"
## [5] "OLDCLAIM_LOG"
```

A summary of Model MLR:3 is below:

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7700  -3160  -1555    366 100643
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3255.858   1082.572   3.008 0.002677 **
## BLUEBOOK_SQRT         26.304      7.509   3.503 0.000473 ***
## CAR_AGE_CAT.L       -436.604    410.182  -1.064 0.287311
## CAR_AGE_CAT.Q       -213.218    424.603  -0.502 0.615628
## CAR_AGE_CAT.C        118.125    440.663   0.268 0.788689
## CAR_TYPEPanel Truck  -13.105   1051.173  -0.012 0.990054
## CAR_TYPEPickup      -384.446    734.476  -0.523 0.600753
## CAR_TYPESports Car  -204.373    763.878  -0.268 0.789085
## CAR_TYPESUV         -377.252    647.651  -0.582 0.560322
## CAR_TYPEVan            7.631    898.664   0.008 0.993226
## CAR_USEPrivate      -642.251    495.798  -1.295 0.195384
## OLDCLAIM_LOG          -2.672     26.881  -0.099 0.920846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8108 on 1511 degrees of freedom
## Multiple R-squared:  0.01815,    Adjusted R-squared:  0.01101
## F-statistic:  2.54 on 11 and 1511 DF,  p-value: 0.003515
```

BLUEBOOK_SQRT might be the only significant predictor of TARGET_AMT. We perform stepwise backward selection to confirm and reduce the model.

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK_SQRT + CAR_USE, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7341  -3157  -1590    334 100968
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2971.738    815.357   3.645 0.000277 ***
## BLUEBOOK_SQRT     27.094      6.205   4.367 1.35e-05 ***
## CAR_USEPrivate  -690.752    425.240  -1.624 0.104502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8090 on 1520 degrees of freedom
## Multiple R-squared:  0.01685,    Adjusted R-squared:  0.01556
## F-statistic: 13.03 on 2 and 1520 DF,  p-value: 2.458e-06
```

CAR_USE is close to what we generally consider significant to the model, so we will leave it in. Still, this model explains very little of the variation in TARGET_AMT.

Select Models

Appendix: Report Code

Below is the code for this report to generate the models and charts above.

```r
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(DataExplorer)
library(knitr)
library(cowplot)
library(finalfit)
library(correlationfunnel)
library(ggcorrplot)
library(RColorBrewer)
library(naniar)
library(mice)
library(MASS)
select <- dplyr::select
library(kableExtra)

cur_theme <- theme_set(theme_classic())

my_url <- "https://raw.githubusercontent.com/andrewbowen19/businessAnalyticsDataMiningDATA621/main/data
main_df <- read.csv(my_url, na.strings = "")

classes <- as.data.frame(unlist(lapply(main_df, class))) |>
```

```
    rownames_to_column()
cols <- c("Variable", "Class")
colnames(classes) <- cols
classes_summary <- classes |>
    group_by(Class) |>
    summarize(Count = n(),
              Variables = paste(sort(unique(Variable)),collapse=", "))
kable(classes_summary, "latex", booktabs = T) |>
  kableExtra::column_spec(2:3, width = "7cm")

vars <- c("INCOME", "HOME_VAL", "BLUEBOOK", "OLDCLAIM")
main_df <- main_df |>
    mutate(across(all_of(vars), ~gsub("\\$|,", "", .) |> as.integer()))

main_df <- main_df |>
    select(-INDEX)
remove <- c("discrete_columns", "continuous_columns",
            "total_observations", "memory_usage")
completeness <- introduce(main_df) |>
    select(-all_of(remove))
knitr::kable(t(completeness), format = "simple")

p1 <- plot_missing(main_df, missing_only = TRUE,
                   ggtheme = theme_classic(), title = "Missing Values")

p1 <- p1 +
    scale_fill_brewer(palette = "Paired")
p1

exclude <- c("TARGET_AMT", "AGE", "INCOME", "YOJ", "HOME_VAL", "CAR_AGE", "JOB")
main_df_binarized <- main_df |>
    select(-all_of(exclude)) |>
    binarize(n_bins = 5, thresh_infreq = 0.01, name_infreq = "OTHER",
           one_hot = TRUE)
main_df_corr <- main_df_binarized |>
    correlate(TARGET_FLAG__1)
main_df_corr |>
    plot_correlation_funnel()

palette <- brewer.pal(n = 7, name = "RdBu")[c(1, 4, 7)]
excl <- c("TARGET_FLAG", "JOB", "CAR_TYPE", "CAR_USE", "EDUCATION",
             "MSTATUS", "PARENT1", "RED_CAR", "REVOKED", "SEX", "URBANICITY")
model.matrix(~0+., data = main_df |> filter(TARGET_FLAG == 1) |>select(-all_of(excl))) |>
    cor(use = "pairwise.complete.obs") |>
    ggcorrplot(show.diag = FALSE, type = "lower", lab = TRUE, lab_size = 2.5,
               tl.cex = 8, tl.srt = 90,
               colors = palette, outline.color = "white")

incl <- c("TARGET_AMT", "CAR_USE", "MSTATUS", "PARENT1", "RED_CAR",
             "REVOKED", "SEX", "URBANICITY")
model.matrix(~0+., data = main_df |> filter(TARGET_FLAG == 1) |> select(all_of(incl))) |>
    cor(use = "pairwise.complete.obs") |>
    ggcorrplot(show.diag = FALSE, type = "lower", lab = TRUE, lab_size = 3,
```

```r
               tl.cex = 8, tl.srt = 90,
               colors = palette, outline.color = "white")

incl <- c("TARGET_AMT", "JOB", "CAR_TYPE", "EDUCATION")
model.matrix(~0+., data = main_df |> filter(TARGET_FLAG == 1) |> select(all_of(incl))) |>
    cor(use = "pairwise.complete.obs") |>
    ggcorrplot(show.diag = FALSE, type = "lower", lab = TRUE, lab_size = 1.75,
               tl.cex = 8, tl.srt = 90,
               colors = palette, outline.color = "white")

r <- model.matrix(~0+., data = main_df) |>
    cor(use = "pairwise.complete.obs")
is.na(r) <- abs(r) < 0.45
r |>
    ggcorrplot(show.diag = FALSE, type = "lower", lab = TRUE, lab_size = 2.5,
               tl.cex = 8, tl.srt = 90,
               colors = palette, outline.color = "white")

output <- split_columns(main_df, binary_as_factor = TRUE)
num <- data.frame(Variable = names(output$continuous),
                  Type = rep("Numeric", ncol(output$continuous)))
cat <- data.frame(Variable = names(output$discrete),
                  Type = rep("Categorical", ncol(output$discrete)))
ranges <- as.data.frame(t(sapply(main_df |> select(-names(output$discrete)),
                                 range, na.rm = TRUE)))
factors <- names(output$discrete)
main_df <- main_df |>
    mutate(across(all_of(factors), ~as.factor(.)))
values <- as.data.frame(t(sapply(main_df |> select(all_of(factors)),
                                 levels)))
values <- values |>
    mutate(across(all_of(factors), ~toString(unlist(.))))
values <- as.data.frame(t(values)) |>
    rownames_to_column()
cols <- c("Variable", "Values")
colnames(values) <- cols
remove <- c("V1", "V2")
ranges <- ranges |>
    rownames_to_column() |>
    group_by(rowname) |>
    mutate(Values = toString(c(V1, " - ", round(V2, 1))),
           Values = str_replace_all(Values, ",", "")) |>
    select(-all_of(remove))
colnames(ranges) <- cols
num <- num |>
    merge(ranges)
cat <- cat |>
    merge(values)
num_vs_cat <- num |>
    bind_rows(cat)
knitr::kable(num_vs_cat, "latex", booktabs = T)|>
  kableExtra::column_spec(2:3, width = "6cm")
```

```
alt_df <- main_df
main_df <- main_df |>
    mutate(TARGET_AMT = case_when(as.numeric(as.character(TARGET_FLAG)) < 1 ~ NA,
                                  TRUE ~ TARGET_AMT),
           HOME_VAL = case_when(HOME_VAL < 1 ~ NA,
                                  TRUE ~ HOME_VAL),
           INCOME = case_when(INCOME < 1 ~ NA,
                                 TRUE ~ INCOME),
           KIDSDRIV = case_when(KIDSDRIV < 1 ~ NA,
                                 TRUE ~ KIDSDRIV),
           HOMEKIDS = case_when(HOMEKIDS < 1 ~ NA,
                                 TRUE ~ HOMEKIDS))

main_df <- main_df |>
    mutate(CAR_AGE = case_when(CAR_AGE < 0 ~ CAR_AGE * -1,
                                  TRUE ~ CAR_AGE))
alt_df <- alt_df |>
    mutate(CAR_AGE = case_when(CAR_AGE < 0 ~ CAR_AGE * -1,
                                  TRUE ~ CAR_AGE))

summary(main_df)

littles_test <- main_df |>
    mcar_test()
knitr::kable(littles_test, format = "simple")

x <- colnames(main_df)
dep = c("CAR_AGE")
exp = x[!x %in% dep]
missing_comp1 <- main_df |>
    missing_compare(explanatory = exp, dependent = dep) |>
    mutate(p = as.numeric(case_when(p == "<0.001" ~ "0.001",
                                       TRUE ~ p))) |>
    mutate(Dependant = dep)
colnames(missing_comp1) <- c("Explanatory", "Ref", "Not Missing", "Missing", "p",
                              "Dependant")
dep = c("YOJ")
exp = x[!x %in% dep]
missing_comp2 <- main_df |>
    missing_compare(explanatory = exp, dependent = dep) |>
    mutate(p = as.numeric(case_when(p == "<0.001" ~ "0.001",
                                       TRUE ~ p))) |>
    mutate(Dependant = dep)
colnames(missing_comp2) <- c("Explanatory", "Ref", "Not Missing", "Missing", "p",
                              "Dependant")
dep = c("INCOME")
exp = x[!x %in% dep]
missing_comp3 <- main_df |>
    missing_compare(explanatory = exp, dependent = dep) |>
    mutate(p = as.numeric(case_when(p == "<0.001" ~ "0.001",
                                       TRUE ~ p))) |>
    mutate(Dependant = dep)
colnames(missing_comp3) <- c("Explanatory", "Ref", "Not Missing", "Missing", "p",
```

```
                              "Dependant")
dep = c("HOME_VAL")
exp = x[!x %in% dep]
missing_comp4 <- main_df |>
    missing_compare(explanatory = exp, dependent = dep) |>
    mutate(p = as.numeric(case_when(p == "<0.001" ~ "0.001",
                                    TRUE ~ p))) |>
    mutate(Dependant = dep)
colnames(missing_comp4) <- c("Explanatory", "Ref", "Not Missing", "Missing", "p",
                              "Dependant")
dep = c("JOB")
exp = x[!x %in% dep]
missing_comp5 <- main_df |>
    missing_compare(explanatory = exp, dependent = dep) |>
    mutate(p = as.numeric(case_when(p == "<0.001" ~ "0.001",
                                    TRUE ~ p))) |>
    mutate(Dependant = dep)
colnames(missing_comp5) <- c("Explanatory", "Ref", "Not Missing", "Missing", "p",
                              "Dependant")
missing_comp <- missing_comp1 |>
    bind_rows(missing_comp2, missing_comp3, missing_comp4, missing_comp5) |>
    mutate(Explanatory = case_when(is.na(p) ~ NA,
                                   TRUE ~ Explanatory)) |>
    fill(Explanatory, .direction = "down") |>
    group_by(Dependant, Explanatory) |>
    filter(any(p < 0.05)) |>
    select(Dependant, everything())
knitr::kable(missing_comp, format = "simple")

show <- c("YOJ", "INCOME", "HOME_VAL", "JOB")
p2 <- main_df |>
    select(all_of(show)) |>
    missing_plot()
p2

explanatory = c("JOB", "INCOME", "YOJ")
dependent = "HOME_VAL"
p3 <- main_df |>
    select(all_of(show)) |>
    missing_pattern(dependent, explanatory)

# just numeric variables
numeric_train <- main_df[,sapply(main_df, is.numeric)]
par(mfrow=c(4,4))
par(mai=c(.3,.3,.3,.3))
variables <- names(numeric_train)
for (i in 1:(length(variables))) {
  hist(numeric_train[[variables[i]]], main = variables[i], col = "lightblue")
}

cat_pivot <- main_df |>
    select(all_of(factors)) |>
    pivot_longer(cols = all_of(factors),
```

```
                 names_to = "Variable",
                 values_to = "Value") |>
    group_by(Variable, Value) |>
    summarize(Count = n()) |>
    group_by(Variable) |>
    mutate(Levels = n()) |>
    ungroup()
p4 <- cat_pivot |>
    filter(Levels == 2) |>
    ggplot(aes(x = Value, y = Count)) +
    geom_col(fill = "lightblue", color = "black") +
    facet_wrap(vars(Variable), ncol = 4, scales = "free_x") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
p4

p5 <- cat_pivot |>
    filter(Levels > 2) |>
    ggplot(aes(x = Value, y = Count)) +
    geom_col(fill = "lightblue", color = "black") +
    coord_flip() +
    facet_wrap(vars(Variable), ncol = 1, scales = "free")
p5

# car type
x <- main_df$CAR_TYPE
main_df$CAR_TYPE <- case_match(x, "z_SUV" ~ "SUV", .default = x)
main_df$CAR_TYPE <- factor(main_df$CAR_TYPE,
                           levels = c("Minivan", "Panel Truck",
                                      "Pickup", "Sports Car", "SUV", "Van"))
x <- alt_df$CAR_TYPE
alt_df$CAR_TYPE <- case_match(x, "z_SUV" ~ "SUV", .default = x)
alt_df$CAR_TYPE <- factor(alt_df$CAR_TYPE,
                           levels = c("Minivan", "Panel Truck",
                                      "Pickup", "Sports Car", "SUV", "Van"))

# education
x <- main_df$EDUCATION
main_df$EDUCATION <- case_match(x, "z_High School" ~ "High School", .default = x)
main_df$EDUCATION <- factor(main_df$EDUCATION,
                             levels = c("<High School", "High School",
                                        "Bachelors", "Masters", "PhD"))
x <- alt_df$EDUCATION
alt_df$EDUCATION <- case_match(x, "z_High School" ~ "High School", .default = x)
alt_df$EDUCATION <- factor(alt_df$EDUCATION,
                             levels = c("<High School", "High School",
                                        "Bachelors", "Masters", "PhD"))

# job
x <- main_df$JOB
main_df$JOB <- case_match(x, "z_Blue Collar" ~ "Blue Collar", .default = x)
main_df$JOB <- factor(main_df$JOB, levels = c("Blue Collar", "Clerical",
                                             "Doctor", "Home Maker","Lawyer",
                                             "Manager", "Professional", "Student"))
```

```r
x <- alt_df$JOB
alt_df$JOB <- case_match(x, "z_Blue Collar" ~ "Blue Collar", .default = x)
alt_df$JOB <- factor(alt_df$JOB, levels = c("Blue Collar", "Clerical",
                                              "Doctor", "Home Maker","Lawyer",
                                              "Manager", "Professional", "Student"))

# single parent
main_df <- main_df |>
  mutate(PARENT1 = as.factor(ifelse(PARENT1 == "Yes", 1, 0)))
alt_df <- alt_df |>
  mutate(PARENT1 = as.factor(ifelse(PARENT1 == "Yes", 1, 0)))

# marital status
x <- main_df$MSTATUS
main_df$MSTATUS <- case_match(x, "z_No" ~ "No", .default = x)
main_df <- main_df |>
  mutate(MSTATUS = as.factor(ifelse(MSTATUS == "Yes", 1, 0)))
x <- alt_df$MSTATUS
alt_df$MSTATUS <- case_match(x, "z_No" ~ "No", .default = x)
alt_df <- alt_df |>
  mutate(MSTATUS = as.factor(ifelse(MSTATUS == "Yes", 1, 0)))

# red car
x <- main_df$RED_CAR
main_df$RED_CAR <- case_match(x, "no" ~ "No", "yes" ~ "Yes", .default = x)
main_df <- main_df |>
  mutate(RED_CAR = as.factor(ifelse(RED_CAR == "Yes", 1, 0)))
x <- alt_df$RED_CAR
alt_df$RED_CAR <- case_match(x, "no" ~ "No", "yes" ~ "Yes", .default = x)
alt_df <- alt_df |>
  mutate(RED_CAR = as.factor(ifelse(RED_CAR == "Yes", 1, 0)))

# revoked
main_df <- main_df |>
  mutate(REVOKED = as.factor(ifelse(REVOKED == "Yes", 1, 0)))
alt_df <- alt_df |>
  mutate(REVOKED = as.factor(ifelse(REVOKED == "Yes", 1, 0)))

# sex
x <- main_df$SEX
main_df$SEX <- case_match(x, "M" ~ "Male", "z_F" ~ "Female", .default = x)
main_df$SEX <- factor(main_df$SEX, levels = c("Male", "Female"))
x <- alt_df$SEX
alt_df$SEX <- case_match(x, "M" ~ "Male", "z_F" ~ "Female", .default = x)
alt_df$SEX <- factor(alt_df$SEX, levels = c("Male", "Female"))

# urban city - 1 if urban, 0 if rural
x <- main_df$URBANICITY
main_df$URBANICITY <- case_match(x, "Highly Urban/ Urban" ~ "Urban",
                                 "z_Highly Rural/ Rural" ~ "Rural", .default = x)
main_df <- main_df |>
  mutate(URBANICITY = as.factor(ifelse(URBANICITY == "Urban", 1, 0)))
x <- alt_df$URBANICITY
```

```r
alt_df$URBANICITY <- case_match(x, "Highly Urban/ Urban" ~ "Urban",
                                "z_Highly Rural/ Rural" ~ "Rural", .default = x)
alt_df <- alt_df |>
  mutate(URBANICITY = as.factor(ifelse(URBANICITY == "Urban", 1, 0)))

vars <- c("CAR_TYPE", "EDUCATION", "JOB", "PARENT1", "MSTATUS", "RED_CAR",
          "REVOKED", "SEX", "URBANICITY")

levs <- c("Minivan, Panel Truck, Pickup, Sports Car, SUV, Van",
          "<High School, High School, Bachelors, Masters, PhD",
          "Blue Collar, Clerical, Doctor, Home Maker, Lawyer, Manager, Professional, Student",
          "0, 1",
          "0, 1",
          "0, 1",
          "0, 1",
          "Male, Female",
          "0, 1")

vars_levs <- as.data.frame(cbind(vars, levs))
colnames(vars_levs) <- c("Factor", "New Levels")
knitr::kable(vars_levs, format = "simple")

drop <- c("INCOME", "HOME_VAL")
main_df <- main_df |>
    mutate(INCOME_THOU = INCOME / 1000,
           HOME_VAL_THOU = HOME_VAL / 1000) |>
    select(-all_of(drop))
alt_df <- alt_df |>
    mutate(INCOME_THOU = INCOME / 1000,
           HOME_VAL_THOU = HOME_VAL / 1000) |>
    select(-all_of(drop))

main_df <- main_df |>
    mutate(YOJ = case_when(JOB == "Student" ~ NA,
                           TRUE ~ YOJ))
alt_df <- alt_df |>
    mutate(YOJ = case_when(JOB == "Student" ~ 0,
                           TRUE ~ YOJ))

exclude1 <- c("Student", "Homemaker")
exclude2 <- c(exclude1, "Blue Collar")
main_df <- main_df |>
    mutate(HOME_VAL_CAT = factor(case_when(HOME_VAL_THOU < 251 ~ "<=250K",
                                        HOME_VAL_THOU < 501 ~ "251-500K",
                                        HOME_VAL_THOU < 751 ~ "501-750K",
                                        TRUE ~ "751K+"),
                              ordered = TRUE,
                              levels = c("<=250K", "251-500K", "501-750K", "751K+"),
                              exclude = NULL),
           HOMEOWNER = as.factor(ifelse(is.na(HOME_VAL_THOU), 0, 1)),
           INCOME_CAT = factor(case_when(INCOME_THOU < 51 ~ "<=50K",
                                          INCOME_THOU < 101 ~ "51-100K",
                                          INCOME_THOU < 151 ~ "101-150K",
```

```
                                        TRUE ~ "151K+"),
                              ordered = TRUE,
                              levels = c("<=50K", "51-100K", "101-150K", "151K+"),
                              exclude = NULL),
         INCOME_FLAG = as.factor(ifelse(is.na(INCOME_THOU), 0, 1)),
         KIDSDRIV_FLAG = as.factor(case_when(!is.na(KIDSDRIV) ~ 1,
                                             TRUE ~ 0)),
         HOMEKIDS_FLAG = as.factor(case_when(!is.na(HOMEKIDS) ~ 1,
                                             TRUE ~ 0)),
         EMPLOYED = as.factor(ifelse(JOB %in% exclude1 | is.na(JOB),
                                     0, 1)),
         CAR_AGE_CAT = factor(case_when(CAR_AGE < 5 ~ "<=4",
                                        CAR_AGE < 9 ~ "5-8",
                                        CAR_AGE < 13 ~ "9-12",
                                        TRUE ~ "13+"),
                               ordered = TRUE,
                               levels = c("<=4", "5-8", "9-12", "13+"),
                               exclude = NULL),
         WHITE_COLLAR = as.factor(ifelse(JOB %in% exclude2 | is.na(JOB),
                                         0, 1)))
main_df$JOB <- factor(main_df$JOB, exclude = NULL)

set.seed(202)
rows <- sample(nrow(main_df))
main_df <- main_df[rows, ]
alt_df <- alt_df[rows, ]
sample <- sample(c(TRUE, FALSE), nrow(main_df), replace=TRUE,
                 prob=c(0.7,0.3))
train_df <- main_df[sample, ]
test_df <- main_df[!sample, ]
alt_train_df <- alt_df[sample, ]
alt_test_df <- alt_df[!sample, ]

train_df_imputed <- train_df |>
    mutate(AGE = case_when(is.na(AGE) ~ mean(AGE, na.rm = TRUE),
                           TRUE ~ AGE))
test_df_imputed <- test_df |>
    mutate(AGE = case_when(is.na(AGE) ~ mean(AGE, na.rm = TRUE),
                           TRUE ~ AGE))

missing <- c("AGE")
imp_train_num <- train_df_imputed |>
    select(all_of(missing)) |>
    mutate(Set = "Train")
imp_test_num <- test_df_imputed |>
    select(all_of(missing)) |>
    mutate(Set = "Test")
imp_num <- imp_train_num |>
    bind_rows(imp_test_num)
imp_num_pivot <- imp_num |>
    pivot_longer(!Set, names_to = "Variable", values_to = "Value")
p6 <- imp_num_pivot |>
    ggplot(aes(x = Value)) +
```

```r
    geom_density(fill = "lightblue", color = "black") +
    labs(y = "Density") +
    facet_grid(rows = vars(Set), cols = vars(Variable),
               switch = "y", scales = "free_x")
p6

col_classes <- unlist(lapply(alt_train_df, class))
missing <- c("AGE", "INCOME_THOU", "YOJ", "HOME_VAL_THOU", "CAR_AGE", "JOB")
x <- names(col_classes)
not_missing <- x[!x %in% missing]
#Since the imputation process is a little slow, we only do the imputations once, save the results as .c
if (file.exists("alt_train_df_imputed.csv") & file.exists("alt_test_df_imputed.csv")){
    alt_train_df_imputed <- read.csv("alt_train_df_imputed.csv", na.strings = "",
                                     colClasses = col_classes)
    alt_test_df_imputed <- read.csv("alt_test_df_imputed.csv", na.strings = "",
                                     colClasses = col_classes)
}else{
    #Start with alt_train_df
    init = mice(alt_train_df, maxit=0)
    meth = init$method
    predM = init$predictorMatrix

    #Skip variables without missing data
    meth[not_missing] = ""

    #Set different imputation methods for each of the variables with missing data
    meth[c("AGE")] = "pmm" #Predictive mean matching
    meth[c("INCOME_THOU")] = "pmm"
    meth[c("YOJ")] = "pmm"
    meth[c("HOME_VAL_THOU")] = "pmm"
    meth[c("CAR_AGE")] = "pmm"
    meth[c("JOB")] = "polyreg" #Polytomous (multinomial) logistic regression

    #Impute
    imputed = mice(alt_train_df, method=meth, predictorMatrix=predM, m=5,
                   printFlag = FALSE)
    alt_train_df_imputed <- complete(imputed)
    write.csv(alt_train_df_imputed, "alt_train_df_imputed.csv", row.names = FALSE,
              fileEncoding = "UTF-8")

    #Repeat for alt_test_df
    init = mice(alt_test_df, maxit=0)
    meth = init$method
    predM = init$predictorMatrix
    meth[not_missing] = ""
    meth[c("AGE")] = "pmm"
    meth[c("INCOME_THOU")] = "pmm"
    meth[c("YOJ")] = "pmm"
    meth[c("HOME_VAL_THOU")] = "pmm"
    meth[c("CAR_AGE")] = "pmm"
    meth[c("JOB")] = "polyreg"
    imputed = mice(alt_test_df, method=meth, predictorMatrix=predM, m=5,
                   printFlag = FALSE)
```

```r
    alt_test_df_imputed <- complete(imputed)
    write.csv(alt_test_df_imputed, "alt_test_df_imputed.csv", row.names = FALSE,
              fileEncoding = "UTF-8")
}

#Make sure the levels stay the same
levels(alt_train_df_imputed$CAR_TYPE) <- levels(main_df$CAR_TYPE)
levels(alt_train_df_imputed$EDUCATION) <- levels(main_df$EDUCATION)
levels(alt_train_df_imputed$JOB) <- levels(main_df$JOB)
levels(alt_train_df_imputed$SEX) <- levels(main_df$SEX)
levels(alt_test_df_imputed$CAR_TYPE) <- levels(main_df$CAR_TYPE)
levels(alt_test_df_imputed$EDUCATION) <- levels(main_df$EDUCATION)
levels(alt_test_df_imputed$JOB) <- levels(main_df$JOB)
levels(alt_test_df_imputed$SEX) <- levels(main_df$SEX)

x <- sapply(alt_train_df_imputed, function(x) sum(is.na(x)))
y <- sapply(alt_test_df_imputed, function(x) sum(is.na(x)))
sum(x, y) == 0

missing_num <- c("AGE", "INCOME_THOU", "YOJ", "HOME_VAL_THOU", "CAR_AGE")
imp_alt_train_num <- alt_train_df_imputed |>
    select(all_of(missing_num)) |>
    mutate(Set = "Train")
imp_alt_test_num <- alt_test_df_imputed |>
    select(all_of(missing_num)) |>
    mutate(Set = "Test")
imp_alt_num <- imp_alt_train_num |>
    bind_rows(imp_alt_test_num)
imp_alt_num_pivot <- imp_alt_num |>
    pivot_longer(!Set, names_to = "Variable", values_to = "Value")
p7 <- imp_alt_num_pivot |>
    ggplot(aes(x = Value)) +
    geom_density(fill = "lightblue", color = "black") +
    labs(y = "Density") +
    facet_grid(rows = vars(Set), cols = vars(Variable),
               switch = "y", scales = "free_x")
p7

missing_cat <- c("JOB")
imp_alt_train_cat <- alt_train_df_imputed |>
    select(all_of(missing_cat)) |>
    pivot_longer(cols = all_of(missing_cat),
                 names_to = "Variable",
                 values_to = "Value") |>
    group_by(Variable, Value) |>
    summarize(Count = n()) |>
    mutate(Set = "Train")
imp_alt_test_cat <- alt_test_df_imputed |>
    select(all_of(missing_cat)) |>
    pivot_longer(cols = all_of(missing_cat),
                 names_to = "Variable",
                 values_to = "Value") |>
    group_by(Variable, Value) |>
```

```
    summarize(Count = n()) |>
    mutate(Set = "Test")
imp_alt_pivot_cat <- imp_alt_train_cat |>
    bind_rows(imp_alt_test_cat)
p8 <- imp_alt_pivot_cat |>
    ggplot(aes(x = Value, y = Count)) +
    geom_col(fill = "lightblue", color = "black") +
    labs(x = "Job") +
    coord_flip() +
    facet_wrap(vars(Set), ncol = 2)
p8

skewed <- c("TRAVTIME", "BLUEBOOK", "TIF", "OLDCLAIM", "CLM_FREQ", "MVR_PTS")
train_df_trans <- train_df_imputed
for (i in 1:(length(skewed))){
    #Add a small constant to columns with any 0 values
    if (sum(train_df_trans[[skewed[i]]] == 0) > 0){
        train_df_trans[[skewed[i]]] <-
            train_df_trans[[skewed[i]]] + 0.001
    }
}
for (i in 1:(length(skewed))){
    if (i == 1){
        lambdas <- c()
    }
    bc <- boxcox(lm(train_df_trans[[skewed[i]]] ~ 1),
                 lambda = seq(-2, 2, length.out = 81),
                 plotit = FALSE)
    lambda <- bc$x[which.max(bc$y)]
    lambdas <- append(lambdas, lambda)
}
lambdas <- as.data.frame(cbind(skewed, lambdas))
adj <- c("no transformation", "square root", "log", "log", "log", "log")
lambdas <- cbind(lambdas, adj)
cols <- c("Skewed Variable", "Ideal Lambda Proposed by Box-Cox", "Reasonable Alternative Transformation
colnames(lambdas) <- cols
knitr::kable(lambdas, format = "simple")

remove <- c("BLUEBOOK", "TIF", "OLDCLAIM", "CLM_FREQ", "MVR_PTS")
train_df_trans <- train_df_trans |>
    mutate(BLUEBOOK_SQRT = BLUEBOOK^0.5,
           TIF_LOG = log(TIF),
           OLDCLAIM_LOG = log(OLDCLAIM),
           CLM_FREQ_LOG = log(CLM_FREQ),
           MVR_PTS_LOG = log(MVR_PTS)) |>
    select(-all_of(remove))
test_df_trans <- test_df_imputed
for (i in 1:(length(skewed))){
    #Add a small constant to columns with any 0 values
    if (sum(test_df_trans[[skewed[i]]] == 0) > 0){
        test_df_trans[[skewed[i]]] <-
            test_df_trans[[skewed[i]]] + 0.001
    }
```

```
}
test_df_trans <- test_df_trans |>
    mutate(BLUEBOOK_SQRT = BLUEBOOK^0.5,
           TIF_LOG = log(TIF),
           OLDCLAIM_LOG = log(OLDCLAIM),
           CLM_FREQ_LOG = log(CLM_FREQ),
           MVR_PTS_LOG = log(MVR_PTS)) |>
    select(-all_of(remove))

transformed <- c("BLUEBOOK_SQRT", "TIF_LOG", "OLDCLAIM_LOG", "CLM_FREQ_LOG",
                 "MVR_PTS_LOG")
train_df_trans_set <- train_df_trans |>
    select(all_of(transformed)) |>
    mutate(Set = "Train")
test_df_trans_set <- test_df_trans |>
    select(all_of(transformed)) |>
    mutate(Set = "Test")
trans_sets <- train_df_trans_set |>
    bind_rows(test_df_trans_set)
trans_sets_pivot <- trans_sets |>
    pivot_longer(!Set, names_to = "Variable", values_to = "Value")
p9 <- trans_sets_pivot |>
    ggplot(aes(x = Value)) +
    geom_density(fill = "lightblue", color = "black") +
    labs(y = "Density") +
    facet_grid(rows = vars(Set), cols = vars(Variable),
               switch = "y", scales = "free_x")
p9

dat <- alt_train_df_imputed |>
    select(-TARGET_AMT)
model_blr_1 <- glm(TARGET_FLAG ~ ., family = 'binomial', data = dat)
summary(model_blr_1)

model_blr_2 <- glm(TARGET_FLAG ~ AGE + CLM_FREQ + HOMEOWNER + INCOME_FLAG + EMPLOYED + WHITE_COLLAR + MS
                   data=train_df_imputed, family='binomial')
model_blr_2 <- step(model_blr_2, trace=0)
summary(model_blr_2)

choices <- c("AGE", "CLM_FREQ_LOG", "URBANICITY", "MVR_PTS_LOG", "OLDCLAIM_LOG", "PARENT1", "REVOKED",
print(choices)

sel <- c("TARGET_FLAG", choices)
dat <- train_df_trans |>
    select(all_of(sel))
model_blr_3 <- glm(TARGET_FLAG ~ ., family = 'binomial', data = dat)
summary(model_blr_3)

model_blr_3 <- update(model_blr_3, ~ . - EMPLOYED)
summary(model_blr_3)

dat <- alt_train_df_imputed |>
    select(-TARGET_FLAG)
```

```
model_mlr_1 <- lm(TARGET_AMT ~ ., data = dat)
summary(model_mlr_1)

model_mlr_2 <- lm(TARGET_AMT ~ BLUEBOOK + CAR_AGE_CAT + CAR_TYPE + INCOME_FLAG + RED_CAR + URBANICITY +
model_mlr_2 <- step(model_mlr_2, trace=0)
summary(model_mlr_2)

choices <- c("BLUEBOOK_SQRT", "CAR_AGE_CAT", "CAR_TYPE", "CAR_USE", "OLDCLAIM_LOG")
print(choices)

sel <- c("TARGET_AMT", choices)
dat <- train_df_trans |>
    filter(!is.na(TARGET_AMT)) |>
    select(all_of(sel))
model_mlr_3 <- lm(TARGET_AMT ~ ., data = dat)
summary(model_mlr_3)

model_mlr_3 <- step(model_mlr_3, direction="backward", trace = 0)
summary(model_mlr_3)
```