# 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| REVOKED | License Revoked (Past 7 Years) | If your license was revoked in the past 7 years, you probably are a more risky driver |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.

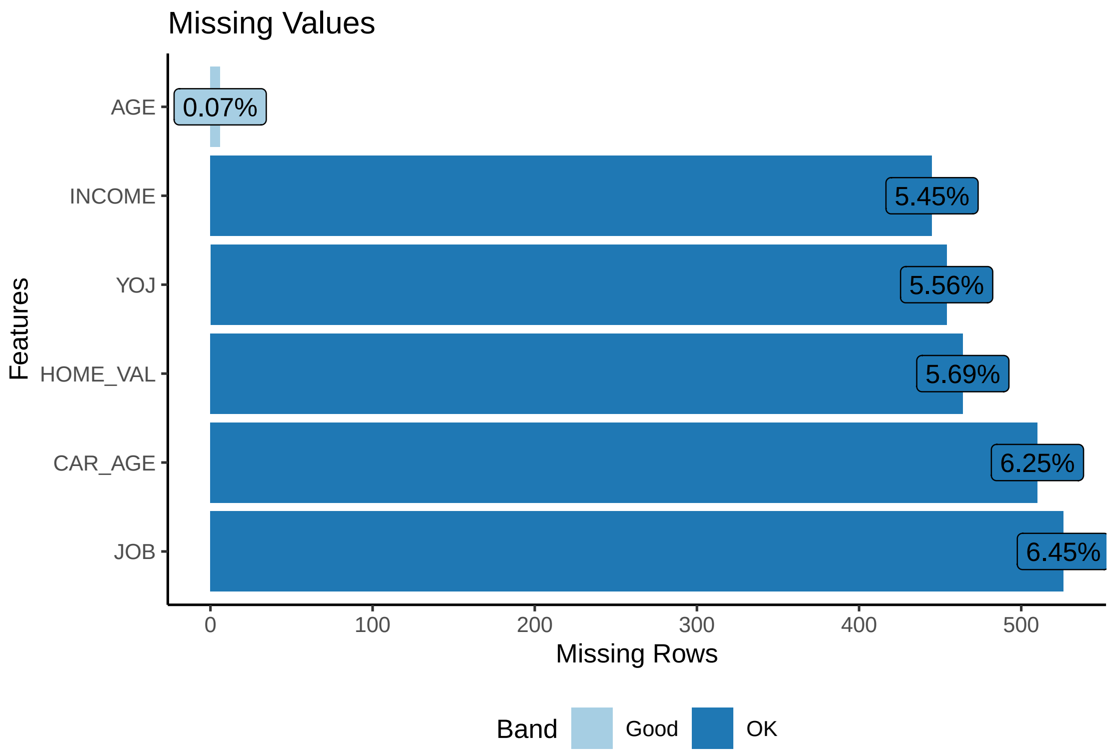
| Class | Count | Variables |
|---|---|---|
| character | 14 | BLUEBOOK, CAR_TYPE, CAR_USE, EDUCATION, HOME_VAL, INCOME, JOB, MSTATUS, OLDCLAIM, PARENT1, RED_CAR, REVOKED, SEX, URBANICITY |
| integer | 11 | AGE, CAR_AGE, CLM_FREQ, HOMEKIDS, INDEX, KIDSDRIV, MVR_PTS, TARGET_FLAG, TIF, TRAVTIME, YOJ |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

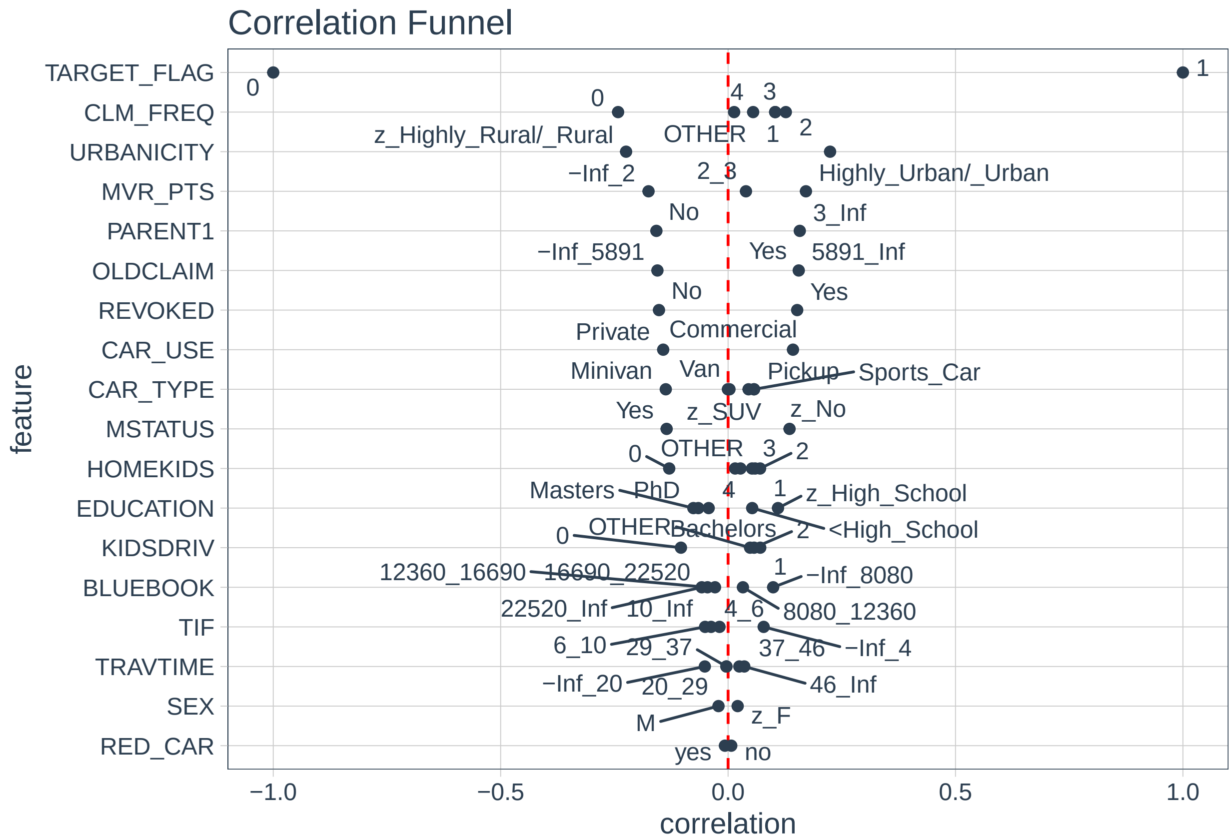



A very small percentage of observations contain missing `AGE` values. The `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`, and `JOB` variables are each missing around 5.5 to 6.5 percent of values. There are no variables containing such extreme proportions of missing values that removal would be warranted on that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

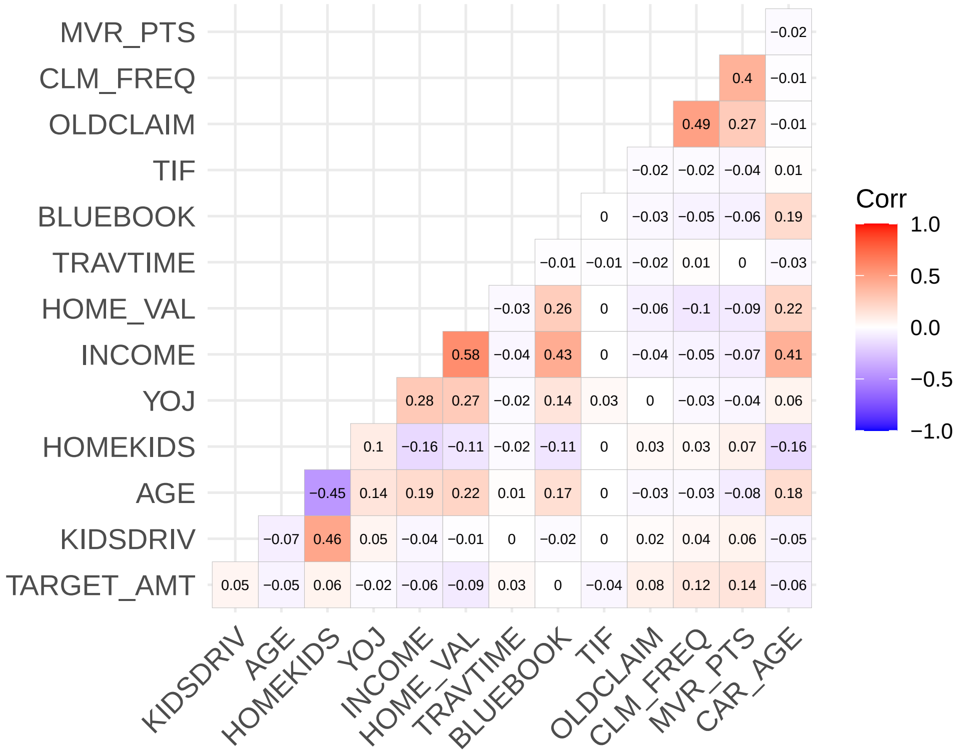
To check whether the predictor variables are correlated with the binary response variable, we produce a correlation funnel that visualizes the strength of the relationships between our predictors and `TARGET_FLAG`. This correlation funnel will not include variables for which there are any missing values.

**This plot needs to be improved. Data point overlap issues.**

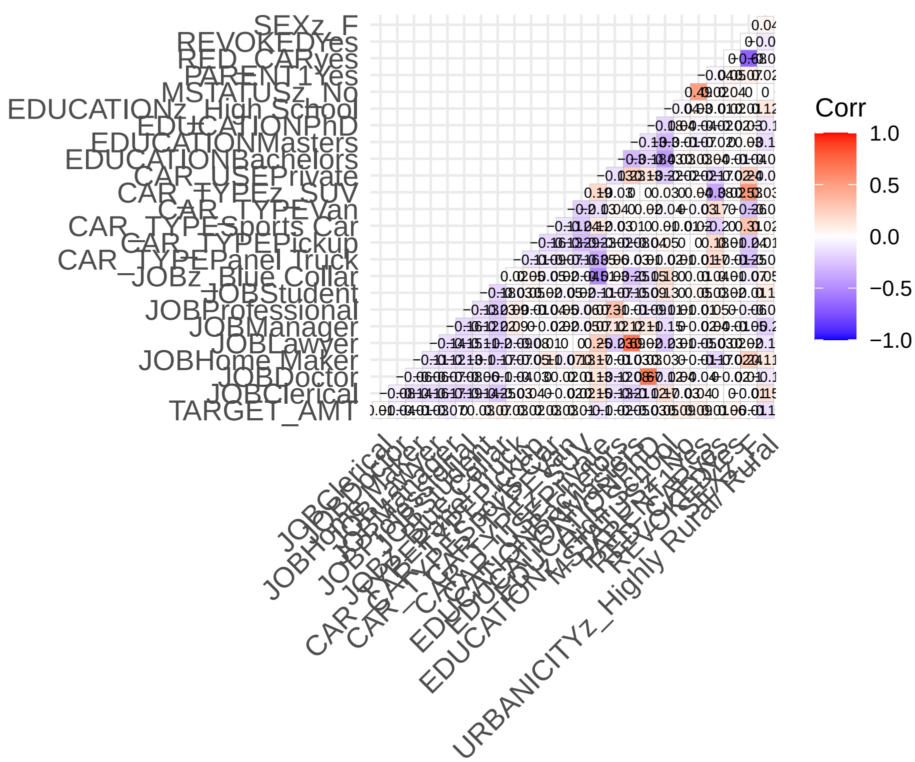
```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




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`. First we look at numeric predictors only, and then we look at non-numeric predictors only.




**This plot will have to be improved, probably by splitting factors with two levels into one plot and factors with more than two levels into another plot.**




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`, `HOMEKIDS`, and `OLDCLAIM` 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.

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 them in the next section. We will also set the reference level for each factor to be the level that we assume increases the risk of getting into a car crash the most. That way, no matter what factor we're looking at later when we're modeling, we should expect negative coefficients for all levels other than the reference level. If we assume nothing regarding how the factor affects the risk of getting into a car crash, then the reference level for that factor will simply be the first level alphabetically after any renaming we do.

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
```

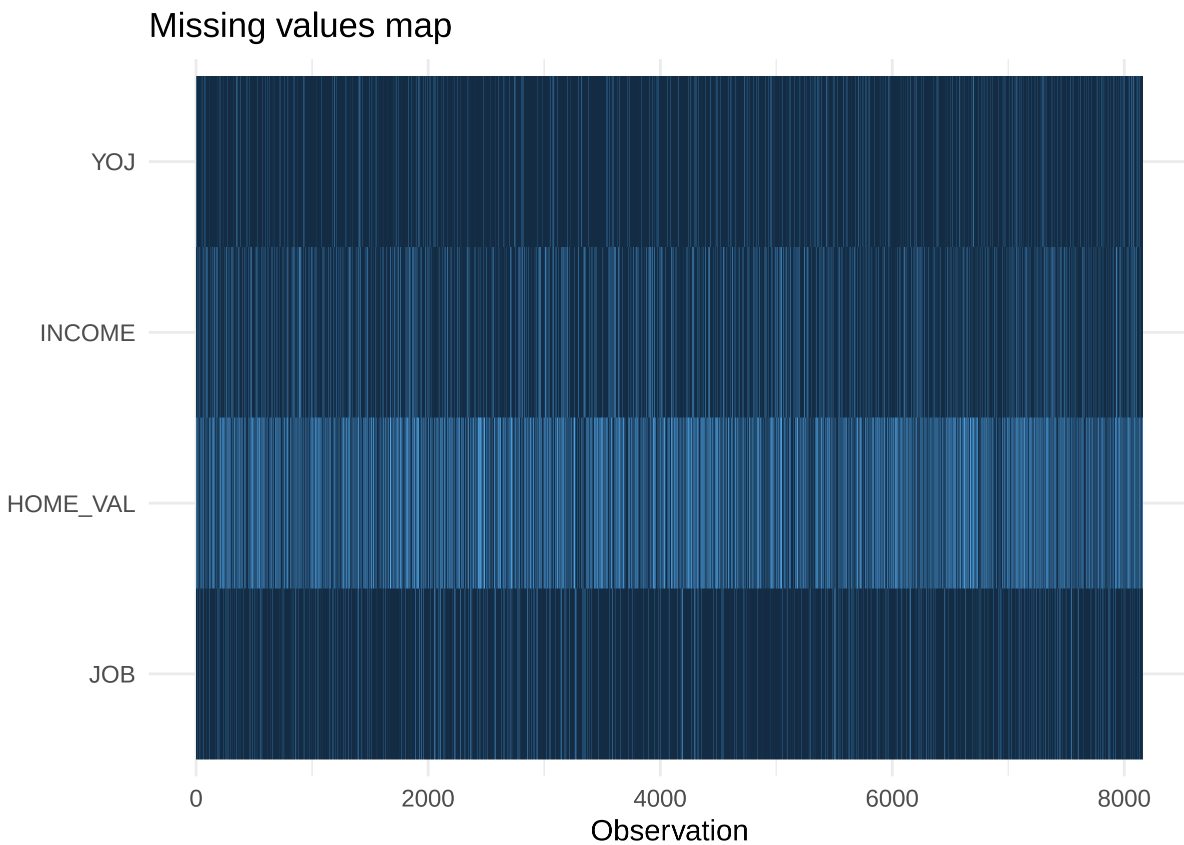
```
##             Median :  4104.00   Median :1.000   Median :45.00  
##             Mean   :  5702.18   Mean   :1.423   Mean   :44.79  
##             3rd Qu.:  5787.00   3rd Qu.:2.000   3rd Qu.:51.00  
##             Max.   :107586.14   Max.   :4.000   Max.   :81.00  
##             NA's   :6008        NA's   :7180    NA's   :6      
##    HOMEKIDS          YOJ           INCIME        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.   :  502   Min.   :0.0000   No :7161  
##  Panel Truck: 676   yes:2378   1st Qu.: 3663   1st Qu.:0.0000   Yes:1000  
##  Pickup     :1389              Median : 6052   Median :0.0000             
##  Sports Car : 907              Mean   :10453   Mean   :0.7986             
##  Van        : 750              3rd Qu.: 9866   3rd Qu.:2.0000             
##  z_SUV      :2294              Max.   :57037   Max.   :5.0000             
##                                NA's   :5009                               
##     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.

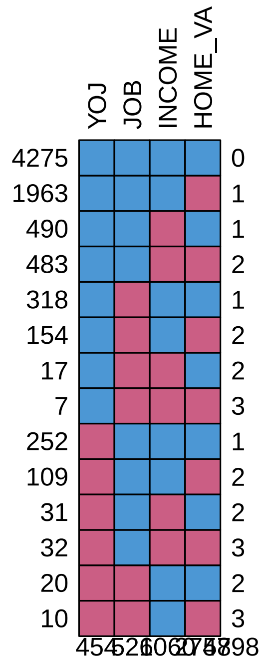
6,008 observations, which is the majority of observations, do not involve car crashes, and we now correctly record 6,008 `NA` observations for `TARGET_AMT`. (Since we introduced `NA` values for `TARGET_AMT` on purpose, we will not impute them in the next section.)

There are 6 `NA` values in `AGE` and 510 in `CAR_AGE` that we can consider Missing at Random (MAR), and we will impute them in the next section.

There are 454 `NA` values in `YOJ`, 1,060 in `INCOME`, 2,758 in `HOME_VAL`, and 526 in `JOB` that we cannot necessarily consider MAR. 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.

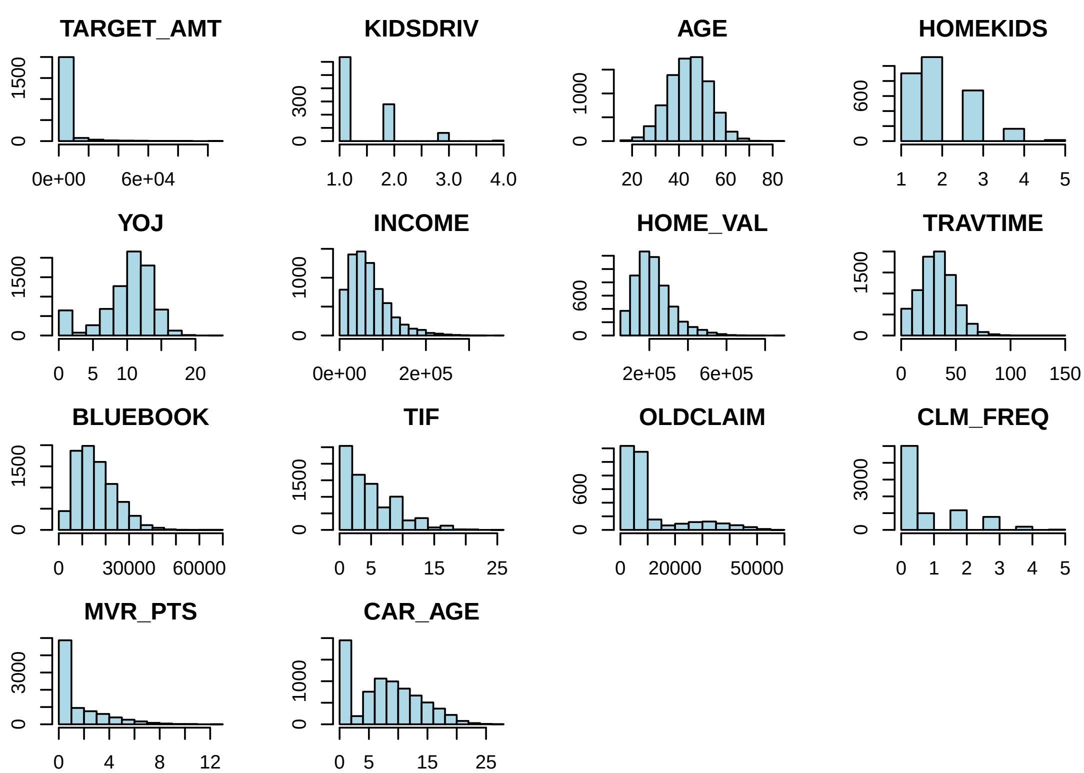



We do see some overlap in the observations that have missing values for these variables, but it's hard to detect anything more conclusive from this plot. To take a closer look at the patterns of missingness between these variables, we can use the `missing_pattern` function from the `finalfit` package.



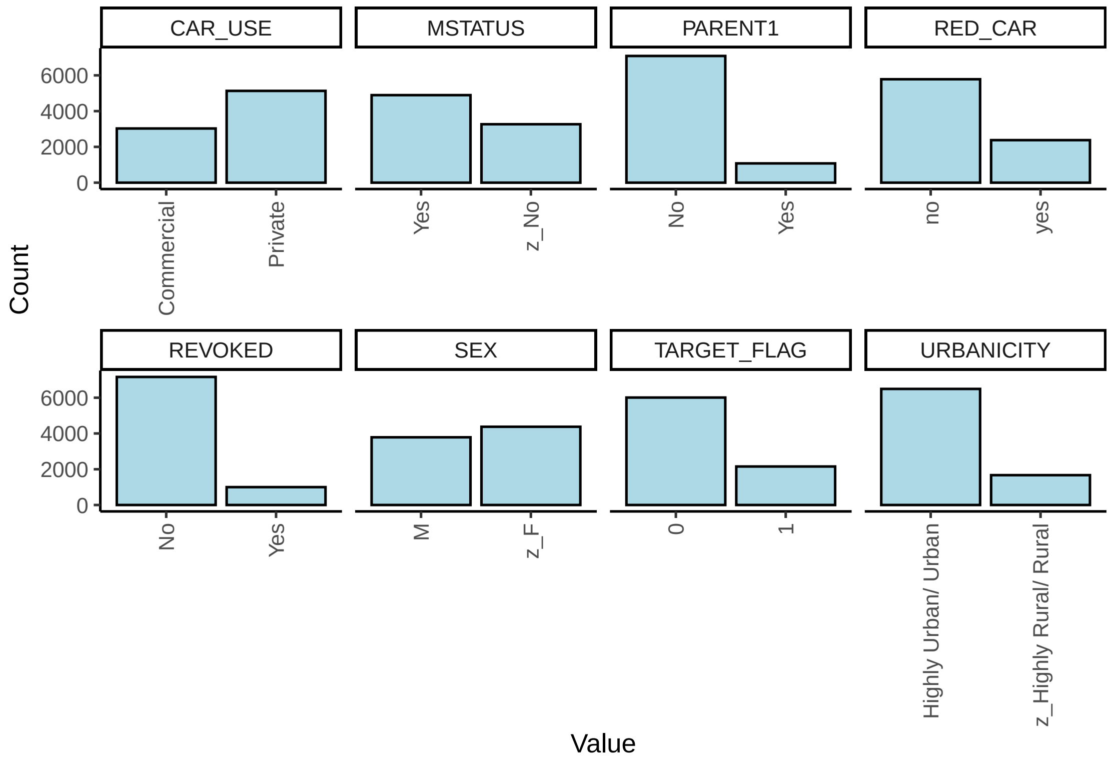

Here, we see several patterns of missingness worth noting. 814 observations are missing two out of these four variables, and 49 observations are missing three. Of the observations that are missing `HOME_VAL`, 483 are also missing `INCOME`, 154 are also missing `JOB`, and 109 are also missing `YOJ`. Due to these patterns of related missingness, we choose not to impute these variables. Doing so would introduce bias.

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



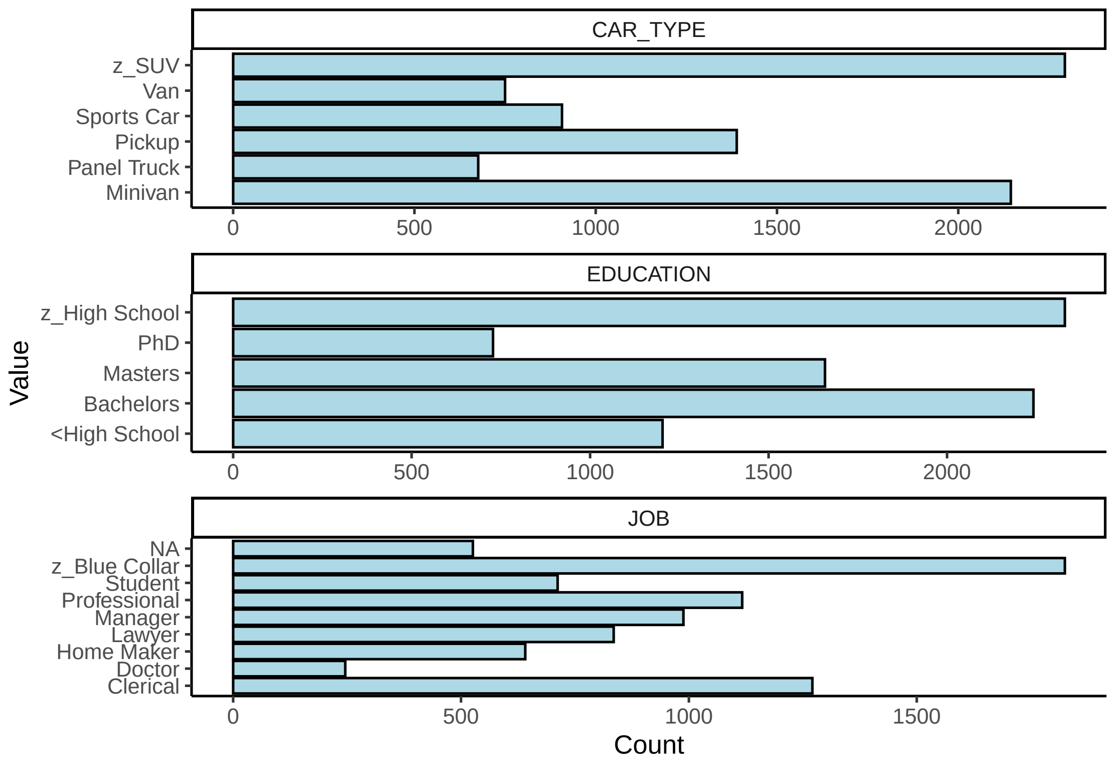
The distributions for AGE is approximately normal. The distribution for YOJ is left-skewed. The distributions for TARGET_AMT, KIDSDRIV, HOMEKIDS, INCOME, HOME_VAL, TRAVTIME, BLUEBOOK, TIF, OLDCLAIM, CLM_FREQ, MVR_PTS, and CAR_AGE are all right-skewed. 75% of observations for TARGET_AMT are at or below $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" or "No", we will replace with a dummy variable 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 |
| MSTATUS | 1, 0 |
| RED_CAR | 1, 0 |
| REVOKED | 1, 0 |
| SEX | 1, 0 |
| URBANICITY | 1, 0 |

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.

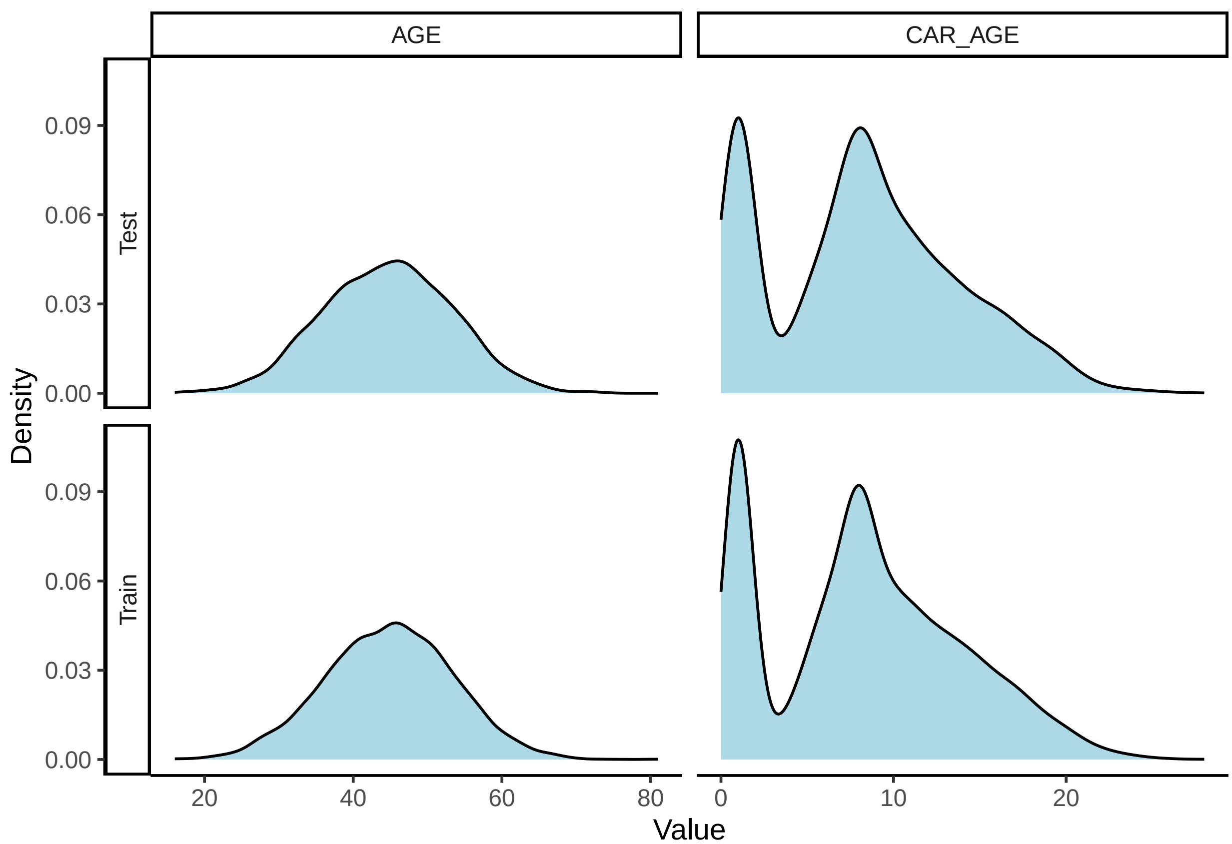
Based on the descriptions of some of the variables and their theoretical effects on the target variables, and to handle the variables that have missing data that we chose not to impute, including those for which we replaced zero or incorrect values with `NA` values, we create several dummy variables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

- `HOMEOWNER` (1 = `HOME_VAL_THOU` $ amount not NA)
- `NUM_KIDSDRIV` (new variable for `KIDSDRIV`; fill in with 0 if NA)
- `KIDSDRIV` (1 = `NUM_KIDSDRIV` number of children not 0)
- `EMPLOYED` (1 = `JOB` neither NA nor Student nor Home Maker or `YOJ` greater than 0/not NA)
- `WHITE_COLLAR` (1 = `JOB` not NA nor Student nor Home Maker nor Blue Collar)

We then split the data into a train and test set.

We impute missing data in the train and test sets for two numeric variables. For `AGE`, we replace `NA` values with the mean value since it is normally distributed. For `CAR_AGE`, we replace `NA` values with the median value since its distribution is left-skewed.

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




The distributions in the train and test sets for are similar to each other, and neither of them are dissimilar from the distributions of the original data.

### Build Models

**Linear Models** ***Full Model***:

**Model with all variables?**

***Select Model***:

Based on the definitions and theories of some of the variables, let's create a linear model using select variables from the dataset.

```
##
## 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
```

**Logistic Models** ***Select Model***:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CLM_FREQ + HOMEOWNER + INCOME +
##     EMPLOYED + WHITE_COLLAR + MSTATUS + PARENT1 + REVOKED + TRAVTIME,
##     family = "binomial", data = train_df_imputed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1532  -0.7686  -0.5735   0.9125   2.2550
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.551171   0.216706  -2.543 0.010978 *
## AGE           -0.014624   0.003953  -3.699 0.000216 ***
## CLM_FREQ       0.373976   0.025568  14.627  < 2e-16 ***
## HOMEOWNER1    -0.265393   0.079453  -3.340 0.000837 ***
## INCOME1       -0.418515   0.098154  -4.264 2.01e-05 ***
## EMPLOYED1      0.438276   0.097062   4.515 6.32e-06 ***
## WHITE_COLLAR1 -0.670359   0.077206  -8.683  < 2e-16 ***
```

```
## MSTATUS1      -0.234082   0.084154  -2.782 0.005410 **
## PARENT11       0.503889   0.103261   4.880 1.06e-06 ***
## REVOKED1       0.908068   0.087895  10.331  < 2e-16 ***
## TRAVTIME       0.008825   0.001991   4.431 9.37e-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: 5991  on 5726  degrees of freedom
## AIC: 6013
##
## Number of Fisher Scoring iterations: 4
```

### 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)

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()

exclude <- c("TARGET_FLAG", "JOB", "CAR_TYPE", "CAR_USE", "EDUCATION",
             "MSTATUS", "PARENT1", "RED_CAR", "REVOKED", "SEX", "URBANICITY")
model.matrix(~0+., data = main_df |> select(-all_of(exclude))) |>
    cor(use = "pairwise.complete.obs") |>
    ggcorrplot(show.diag = FALSE, type = "lower", lab = TRUE, lab_size = 2)

include <- c("TARGET_AMT", "JOB", "CAR_TYPE", "CAR_USE", "EDUCATION",
             "MSTATUS", "PARENT1", "RED_CAR", "REVOKED", "SEX", "URBANICITY")
model.matrix(~0+., data = main_df |> select(all_of(include))) |>
    cor(use = "pairwise.complete.obs") |>
    ggcorrplot(show.diag = FALSE, type = "lower", lab = TRUE, lab_size = 2)

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")

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),
           OLDCLAIM = case_when(OLDCLAIM < 1 ~ NA,
                                 TRUE ~ OLDCLAIM))

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

summary(main_df)

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)
```

```r
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"))

# 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"))

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

```r
# 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)))

# 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)))

# revoked
main_df <- main_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"))

# 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)))

vars <- c("CAR_TYPE", "EDUCATION", "JOB", "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",
          "1, 0",
          "1, 0",
          "1, 0",
          "1, 0",
          "1, 0")

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))

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

```
main_df <- main_df |>
  mutate('EMPLOYED' = as.factor(ifelse(JOB %in% c('Student', 'Home Maker') | is.na(JOB), 0, 1)),
         'WHITE_COLLAR' = as.factor(ifelse(JOB %in% c('Student', 'Home Maker', 'Blue Collar') | is.na(J
         'HOMEOWNER' = as.factor(ifelse(is.na(HOME_VAL_THOU), 0, 1)),
         'NUM_KIDSDRIV' = ifelse(is.na(KIDSDRIV), 0, KIDSDRIV),
         'KIDSDRIV' = as.factor(ifelse(NUM_KIDSDRIV != 0, 0, 1)),
         'INCOME' = as.factor(ifelse(is.na(INCOME_THOU), 0, 1)))

set.seed(202)
rows <- sample(nrow(main_df))
main_df <- main_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, ]

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

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

missing <- c("AGE", "CAR_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)) +
    geom_density(fill = "lightblue", color = "black") +
    labs(y = "Density") +
    facet_grid(rows = vars(Set), cols = vars(Variable),
               switch = "y", scales = "free_x")
p6

lm2 <- lm(TARGET_AMT ~ BLUEBOOK + CAR_AGE + CAR_TYPE + INCOME + RED_CAR + URBANICITY + MVR_PTS + REVOKE
lm2 <- step(lm2, trace=0)
summary(lm2)

glm2 <- glm(TARGET_FLAG ~ AGE + CLM_FREQ + HOMEOWNER + INCOME + EMPLOYED + WHITE_COLLAR + MSTATUS + PAR
glm2 <- step(glm2, trace=0)
```

```r
summary(glm2)
```