412/612 Project Proposal

Anna Livingstone & Philip Eigen

2023-04-09

For this project, we are looking at how a number of variables affect sale price of a used car between individuals. The original data for this project can be found here.

```
library(tidyverse)
library(fastDummies)
data <- read.csv("cardetailsv4.csv")
```

We have to do a few things to clean the data and extract the information we want. First we get rid of all NA values. Next, we select our variables. We will be using price (in dollars) as out response variable, and year, kilometers (on the car), engine (size in cc), seating capacity, fuel tank capacity (in liters), transmission (manual vs. automatic), drivetrain (FWD, RWD, & AWD), height (in mm), and width (in mm). Next, we must reformat Engine from a character vector presented with " cc" at the end to a numerical vector.

```
data %>%
  na.omit() %>%
  filter(Seller.Type=="Individual", Engine!="", Drivetrain!="") %>%
  select(Price, Year, Kilometer, Engine, Seating.Capacity, Fuel.Tank.Capacity, Transmission, Drivetrain
  mutate(Engine=as.numeric(str_replace_all(Engine, "[ c]", "")))-> data2
```

Next, because we have two categorical variables which cannot be represented numerically, we must one-hot encode them. Those two variables are Transmission and Drivetrain. This process will automatically remove the existing Transmission and Drivetrain columns. In order to be able to calculate the regression, we will have to pick a baseline for each variable and remove it from our data. In our case, we will be setting automatic transmissions and front-wheel drive as our baselines, hence why we drop them with the subset function.

```
data3 <- dummy_cols(.data=data2, select_columns=c("Transmission","Drivetrain"), remove_selected_columns=
data3 = subset(data3, select=-c(Transmission_Automatic, Drivetrain_FWD))
names(data3)
```

```
##  [1] "Price"               "Year"                "Kilometer"
##  [4] "Engine"              "Seating.Capacity"    "Fuel.Tank.Capacity"
##  [7] "Height"              "Width"               "Transmission_Manual"
## [10] "Drivetrain_AWD"      "Drivetrain_RWD"
```

With all of our cleaning said and done, we can see our final columns above. Our final dataset has dimensions of 1816, 11.