HW1 Proposal

RJ Cass

1. Understanding the Problem

Background:

Kelley Blue Book (KBB) has extensive car selling data: car conditions, mileage, packages, etc., associated with the selling price. I want to use this data to accurately guide consumers on what their car is worth, and how much they could expect to sell their car for.

Goals:

In this analysis, I want to address several key questions:

i. What factors lead to higher/lower resale values?

- I will perform variable selection to determine which factors or impactful to price.

ii. Are there other factors not included in this dataset that likely explain how much a car is worth? If so, what other factors do you think explain resale value?

- I will determine what other factors may be impactful by consulting expertise (my own experience, search online) as well as viewing the R-Squared value from my model (if RSquared is low, it indicates there are variables missing that would help better explain the relationship to price).

iii. Generally, as mileage increases, the price should decrease. But, does the amount of de-crease in value from additional mileage differ depending upon the make of the car? If so,which makes hold the value better with more miles?

- To determine how the make of the car impacts the effect of mileage on the price of the car, I will need to see if the interaction of Make and Mileage is significant. It may not end up in the final model, but I do need to check if it's significant.

iv. Which car (and with what characteristics) has the highest resale value at 15000 miles?

- To see which car has the highest resale value at 15k miles, I will build a predictive model, plug in a value of 15k miles, and see what values for the other variables give the highest price.

v. What is a reasonable resale value for the following vehicle: Cadillac CTS 4D Sedan with 17,000 miles, 6 cylinder, 2.8 liter engine, cruise control, upgraded speakers and leather seats?

- To identify what is a reasonable price for a car with the above values, I will create a predictive model in which I can enter the relevant values, and it will predict a price range that the car is worth. This will require that my model by able to accurately predict price given the variable inputs.

2. Exploratory Data Analysis

In my initial exploration, the first things that stick out about the data is that price is pretty heavily skewed with a long right tail. I will need to be careful in any assumptions I make about this dataset, particularly that price is normally distributed (I may need to perform a transformation).

THINGS TO LOOK AT: collinearity, responses non-normal, etc.

I did note that none of the factors have any missing data (no NULLs or empty fields). The values of each field appear to be reasonable (ie. mileage is within normal range for cars that are 1 year old), and there don't appear to be any misspelled values (ie. values that should be the same but are currently counting as separate values). I do not believe I need to clean/perpare the data at all prior to performing analysis.

3. Desired Attributes

The model used in this analysis most function as a predictor of price: that is, given certain inputs for the other variables, it must be able to provide a range of price values that are reasonable given the condition of the car. In particular, to answer the final question, it must account for Make, Model, Type, Mileage, Cylinders, Liters, Doors, Cruise Control Speakers, and Trim (if some of these inputs is deemed not impactful on output, is may be excluded from the model itself).

The research goals require an interpretable model that gives predictions, R^2, and that can include Mileage*Make interactions

(b) FINISH

Data shows we may need to address collinearity and non-normality

(c) FINISH

I do not anticipate any other issues in the data.

(d) FINISH

Without meeting the requirements of the research questions, I will not be able to answer those questions. If I don't address collinearity, my coefficient estimates won't be true. If I don't address non-normality, standard models' likelihoods won't be appropriate.

4. Proposed Method

In this analysis, a linear regression model is appropriate. Given the continoues numerical output (price), appears to trend linearly with the only numerical input (mileage), using a linear model is a good fit.

(b)

I will use OLS regression using the format $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$

The model I will use will have mileage as a continous input variable (ie. a 1 mile increase in mileage causes price to drop by x dollars) while the rest of the inputs will be factors (ie. having cruise control vs. not having cruise control increases price by x dollars). I will also be including interaction between mileage and make.

(c) FINISH

i. OLS can have transformations to account for non-normality. Stepwise selection in ELS can help with collinearity.
ii. For the goals of the analysis, regression can have interactions and predict.
iii. Answer questions by 1) State include variables, 2) State RSquared and ideas for non-included variables, 3) Report significance of interactions 4) Predict appropriately 5) Predict appropriately

(d) FINISH

Linearity, independence, normality, equal variance (LINE). Either these will be ok to assume, or we will have tools to address them.

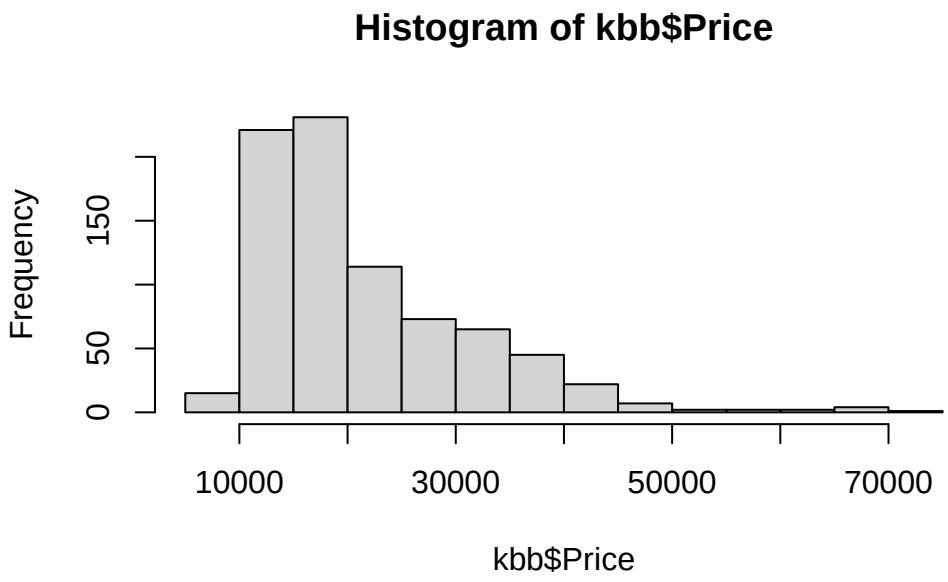
```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

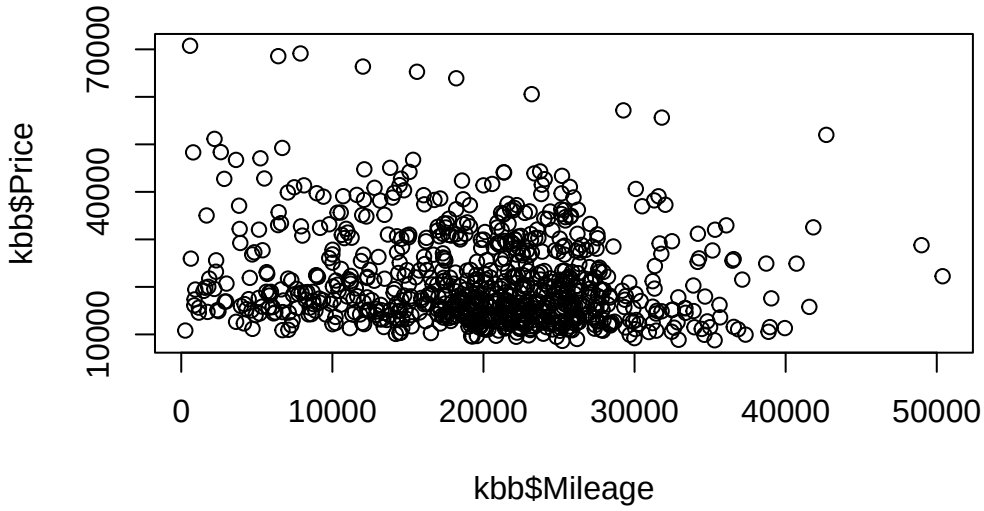
```r
kbb <- read.csv('KBB.csv', header = TRUE)

print(unique(kbb$Model))
```

```
 [1] "Century"     "Lacrosse"    "Lesabre"     "Park Avenue" "CST-V"
 [6] "CTS"         "Deville"     "STS-V6"      "STS-V8"      "XLR-V8"
[11] "AVEO"        "Cavalier"    "Classic"     "Cobalt"      "Corvette"
[16] "Impala"      "Malibu"      "Monte Carlo" "Bonneville"  "G6"
[21] "Grand Am"    "Grand Prix"  "GTO"         "Sunfire"     "Vibe"
[26] "9_3"         "9_3 HO"      "9_5"         "9_5 HO"      "9-2X AWD"
[31] "Ion"         "L Series"
```

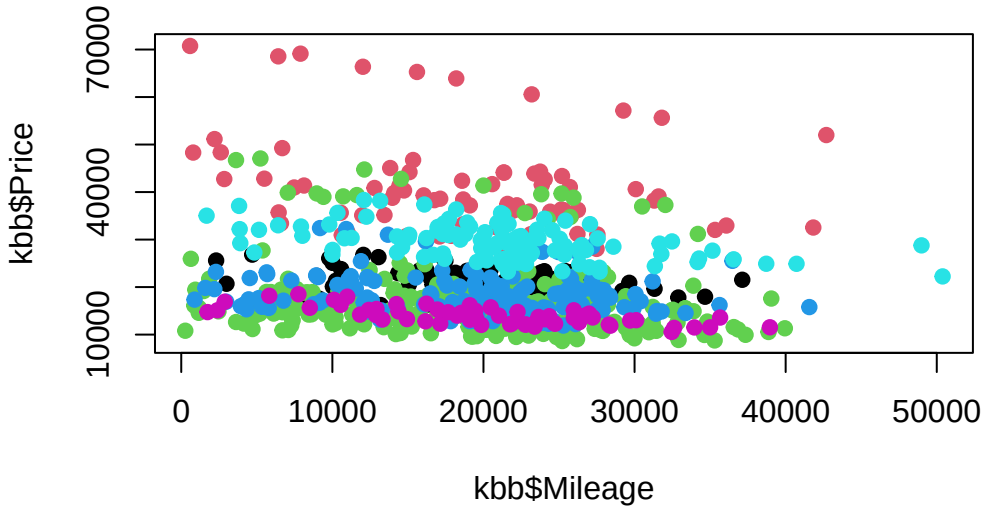
```r
#head(data, 10)

#sum <- summary(data)

#library(skimr)
#skim <- skim(data)

hist(kbb$Price)
```




```
plot(kbb$Mileage, kbb$Price)
```



```r
plot(kbb$Mileage, kbb$Price, col = factor(kbb$Make), pch = 19)
```




```r
#mod = lm(Price ~ ., data = kbb)
#plot(mod)



# THINGS TO EXPLORE
# Pairs Plot (useful comparing continuous to continous)
# Box Plots (useful comparing categoricals)
# Fit a regression using the lm function
```