Final Project Proposal

Isaac Plotkin and Justin Chan

2/8/2022

Narrative

Research Question

Background

One of the hardest things for investments is to decide whether the amount to pay for that investment will have a great return. Often times, people struggle choosing what is the best car for them and if they are paying a fair price for that type of car. However, it can be hard to estimate a fair price for a car due to a number of various features such as car brand, car model, engine type, horsepower, car size and many other features that can be hard to compare car prices depending on different features.

Motivation

In this project, we want to make it easier for others to be able to actually determine if they are receiving fair price for the car they are purchasing. Many salesmen try to upsell a car to make more commissions on a car where some may be overpaying for a car when comparing to its market price. To be able to determine people are getting a fair price for a car, we hope to create a linear regression model that can accurate predict car price based on these various car features.

Hypothesis

When making the linear regression model, we hypothesize that car brand (carName), type of car body (carBody), and fuel type (fueltype) will be the largest factors when determining a car's price. Furthermore, we believe that a linear regression model will be the ideal model to use to predict car prices.

Dataset

Obervations

The obervations of the dataset are pretty straight forward where each obervation or row is a car with the columns being various features of the car. The dataset includes 26 columns where one column is an obervation index and another column is car price which is the variable we are trying to predict so we have 24 input or car features for 205 observations/cars that we can use our linear regression model.

General Description of Variables

The following is the data dictionary of our dataset that gives a clear description of each variable and the type of variable (categorical, continuous/numeric) for each of our possible inputs to use in the linear model from our dataset.

```
print(car_data_dictionary, n = 25)
```

```
## # A tibble: 25 x 2
##    Car_ID                    'Unique id of each observation (Interger)'
##    <chr>                     <chr>
##  1 Symboling                 Its assigned insurance risk rating, A value of +3 ~
##  2 carCompany                Name of car company (Categorical)
##  3 fueltype                  Car fuel type i.e gas or diesel (Categorical)
##  4 aspiration                Aspiration used in a car (Categorical)
##  5 doornumber                Number of doors in a car (Categorical)
##  6 carbody                   body of car (Categorical)
##  7 drivewheel                type of drive wheel (Categorical)
##  8 enginelocation            Location of car engine (Categorical)
##  9 wheelbase                 Weelbase of car (Numeric)
## 10 carlength                 Length of car (Numeric)
## 11 carwidth                  Width of car (Numeric)
## 12 carheight                 height of car (Numeric)
## 13 curbweight                The weight of a car without occupants or baggage. ~
## 14 enginetype                Type of engine. (Categorical)
## 15 cylindernumber            cylinder placed in the car (Categorical)
## 16 enginesize                Size of car (Numeric)
## 17 fuelsystem                Fuel system of car (Categorical)
## 18 boreratio                 Boreratio of car (Numeric)
## 19 stroke                    Stroke or volume inside the engine (Numeric)
## 20 compressionratio          compression ratio of car (Numeric)
## 21 horsepower                Horsepower (Numeric)
## 22 peakrpm                   car peak rpm (Numeric)
## 23 citympg                   Mileage in city (Numeric)
## 24 highwaympg                Mileage on highway (Numeric)
## 25 price(Dependent variable) Price of car (Numeric)
```

Data Collection

The data was originally collected from various market surveys of different types of cars across the United States market around 1987 to learn how to price cars in China depending on the American market. There is an assumption that the cars in the data set have been randomly chosen from the set of cars in the various market surveys. Link to the dataset: https://www.kaggle.com/hellbuoy/car-price-prediction

The Data

Summary

As mentioned above, the data contains 205 rows or observations that represent a car and 26 columns which are 26 different features of each car including its sales price and index in the dataset.

Glimpse

To see a summary of the car dataset, glimpse will be called on the datset to have a better sense of how the data looks in code.

```
glimpse(car_data)
```

```
## Rows: 205
## Columns: 26
## $ car_ID           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16~
## $ symboling        <dbl> 3, 3, 1, 2, 2, 2, 1, 1, 1, 0, 2, 0, 0, 0, 1, 0, 0, 0,~
## $ CarName          <chr> "alfa-romero giulia", "alfa-romero stelvio", "alfa-ro~
## $ fueltype         <chr> "gas", "gas", "gas", "gas", "gas", "gas", "gas", "gas~
## $ aspiration       <chr> "std", "std", "std", "std", "std", "std", "std", "std~
## $ doornumber       <chr> "two", "two", "two", "four", "four", "two", "four", "~
## $ carbody          <chr> "convertible", "convertible", "hatchback", "sedan", "~
## $ drivewheel       <chr> "rwd", "rwd", "rwd", "fwd", "4wd", "fwd", "fwd", "fwd~
## $ enginelocation   <chr> "front", "front", "front", "front", "front", "front",~
## $ wheelbase        <dbl> 88.6, 88.6, 94.5, 99.8, 99.4, 99.8, 105.8, 105.8, 105~
## $ carlength        <dbl> 168.8, 168.8, 171.2, 176.6, 176.6, 177.3, 192.7, 192.~
## $ carwidth         <dbl> 64.1, 64.1, 65.5, 66.2, 66.4, 66.3, 71.4, 71.4, 71.4,~
## $ carheight        <dbl> 48.8, 48.8, 52.4, 54.3, 54.3, 53.1, 55.7, 55.7, 55.9,~
## $ curbweight       <dbl> 2548, 2548, 2823, 2337, 2824, 2507, 2844, 2954, 3086,~
## $ enginetype       <chr> "dohc", "dohc", "ohcv", "ohc", "ohc", "ohc", "ohc", "~
## $ cylindernumber   <chr> "four", "four", "six", "four", "five", "five", "five"~
## $ enginesize       <dbl> 130, 130, 152, 109, 136, 136, 136, 136, 131, 131, 108~
## $ fuelsystem       <chr> "mpfi", "mpfi", "mpfi", "mpfi", "mpfi", "mpfi", "mpfi~
## $ boreratio        <dbl> 3.47, 3.47, 2.68, 3.19, 3.19, 3.19, 3.19, 3.19, 3.13,~
## $ stroke           <dbl> 2.68, 2.68, 3.47, 3.40, 3.40, 3.40, 3.40, 3.40, 3.40,~
## $ compressionratio <dbl> 9.00, 9.00, 9.00, 10.00, 8.00, 8.50, 8.50, 8.50, 8.30~
## $ horsepower       <dbl> 111, 111, 154, 102, 115, 110, 110, 110, 140, 160, 101~
## $ peakrpm          <dbl> 5000, 5000, 5000, 5500, 5500, 5500, 5500, 5500, 5500,~
## $ citympg          <dbl> 21, 21, 19, 24, 18, 19, 19, 19, 17, 16, 23, 23, 21, 2~
## $ highwaympg       <dbl> 27, 27, 26, 30, 22, 25, 25, 25, 20, 22, 29, 29, 28, 2~
## $ price            <dbl> 13495.00, 16500.00, 16500.00, 13950.00, 17450.00, 152~
```