# AI Car Salesman

Rylan Mahany


Dataset
https://www.kaggle.com/datasets/goyalshalini93/car-data?select=CarPrice_Assignment.csv

# Problem

Today is Bender's first day as a car salesman, and he is very excited! Unfortunately, his manager didn't show up to train him. Fortunately, he found the key to the filing cabinet and spent all morning reading previous sale records. Will Bender do a good enough job today to keep his new job?

# Approach to the problem

Bender will first split the sale records into two groups. One to first train himself on, and a second to test himself on after. He will use **Linear Regression** to estimate the prices of the cars. Linear Regression is a machine learning technique that predicts the value of a variable given other variables. He will use **Mean-Absolute-Error** (MAE) to check the accuracy of his work. If he has an MAE of $1,000 and the price of the car is $10,000, Bender will estimate the price of the car to be anything $9,000 - $11,000. This is calculated by adding up the Absolute Values of the difference between the price of the vehicle and Bender's prediction, then dividing by the number of predictions.

$$\text{MAE} = \frac{1}{N} \sum |PriceOfCar - Bender'sPrediction|$$

$N$ = # of predictions made

# Linear Regression

Linear Regression aims to create a Linear Function that uses data points to predict another related data point. This function takes the form:

$$Y = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

Y: The target feature that you are trying to find. In Bender's case, Y is the price of the car.

X: The other features of the vehicle. Quantitative (numerical) X such as MPG, or mileage are fine as is. Qualitative (non-numerical) X such as engine type or brand, must be broken apart into separate X. For example, brand could be broken into brand-Jeep, brand-Honda, and brand-Toyota. Qualitative data can only hold X values of 0 and 1. *n* represents the total # of features.

**[brand-Jeep=1, brand-Honda=0, brand-Toytota=0]**

This would be an example of how Bender sees a Jeep.

W: The coefficient weights applied to all X. This is what the Linear Regression process changes to make more accurate predictions. Consider that Bender thinks $w_{MilesPerGallon} = 40$. This means if a car has $X_{MilesPerGallon} = 25$, then Bender would add 40*25 = $1000 to the total price. For $w_0$, this can be seen as a flat commission fee.

# Results

For my analysis of the Dataset, I looked at the following Features: fueltype (gas/diesel), aspiration (std,turbo), cylindernumber (four, six, other), Horsepower, citympg, highwaympg, and price. This gives the equation:

$$price = w_0 + w_1x_{fueltype} + w_2x_{aspiration} + w_3x_{cylindernumber} + w_4x_{Horsepower} + w_5x_{citympg} + w_6x_{highwaympg}$$

The data looks like:

| | fueltype | aspiration | cylindernumber | horsepower | citympg | highwaympg | price |
|---|---|---|---|---|---|---|---|
| 0 | gas | std | four | 111 | 21 | 27 | 13495.0 |
| 1 | gas | std | four | 111 | 21 | 27 | 16500.0 |
| 2 | gas | std | six | 154 | 19 | 26 | 16500.0 |
| 3 | gas | std | four | 102 | 24 | 30 | 13950.0 |
| 4 | gas | std | five | 115 | 18 | 22 | 17450.0 |

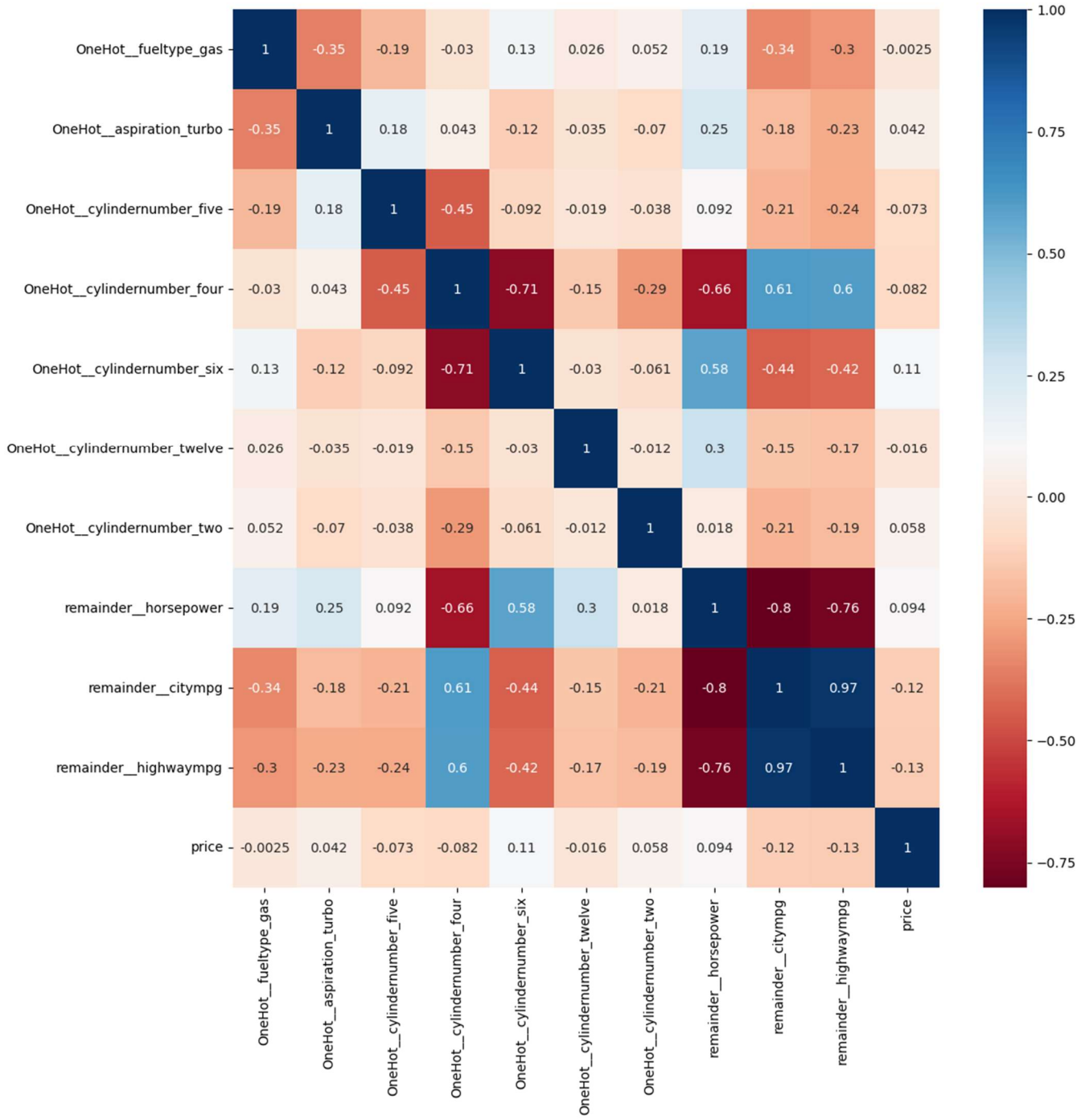
```
#Fitting model and making predictions
LG = LinearRegression().fit(X_train_hot, y_train)
y_predictions = LG.predict(X_test_hot)

y_test = y_test.values

for i in range(len(y_predictions)):
  diff = abs(y_predictions[i] - y_test[i])
MAE = diff / len(y_test)
print('MAE = ', MAE)
```

```
MAE =  13.48255751938366
```

Using these features, Bender was able to estimate the price of each car with an average error of ±$13.48




# Correlation Map of the Data

# Advantages and Limitations

Linear Regression is a powerful model. It's fairly easy to understand, and the results are very easy to interpret. It's also a very flexible model. If Bender's boss was impressed with his work and wanted him to start estimating the MPG of vehicles on the lot, he easily could. It is a linear function, so he could keep his coefficients and solve for the other variables.

However, if the last guy in his position was fired for always undercharging customers, since he learned from his data, Bender may find himself doing the same. Another disadvantage to Bender's approach is that Linear Regression is sensitive to outliers. For example, if some sort of special order, highest-package Honda Civic cost a customer $150,000, Bender may take this outlier and increase the price of all Honda Civics because of it.