# Machine Learning Peer Graded Assignment

This assignment highlights the steps taken to build several regression models as part of the IBM Machine Learning Professional Certificate.

## The Data

The dataset used will be the Ford Car Price Dataset. This dataset was obtained from Kaggle. Please [click here](https://www.kaggle.com/datasets/adhurimquku/ford-car-price-prediction) to access the full dataset Subsequently, this dataset will be used to create several regression models that will be compared. It is important to note that these models were built to predict the price of a ford vehicle.

## Data Attributes

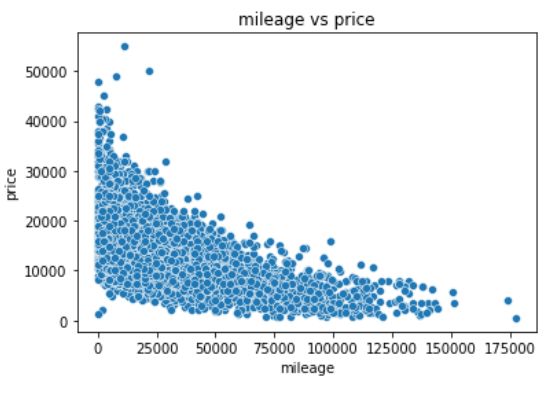
The following list describes what column represents:

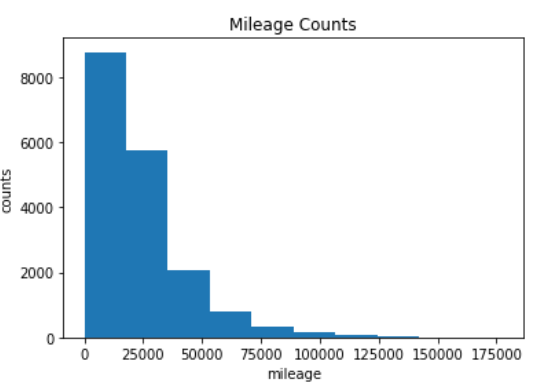
* model - Ford Car Brands
* year - Production Year
* price - Price of car in $
* transmission - Automatic,Manual, Semi-Auto
* mileage - Number of miles traveled
* fuel\_Type - Petrol,Diesel,Hybrid,Electric,Other
* tax - Annual Tax
* mpg - Miles per Gallon
* engineSize - Car's Engine Size

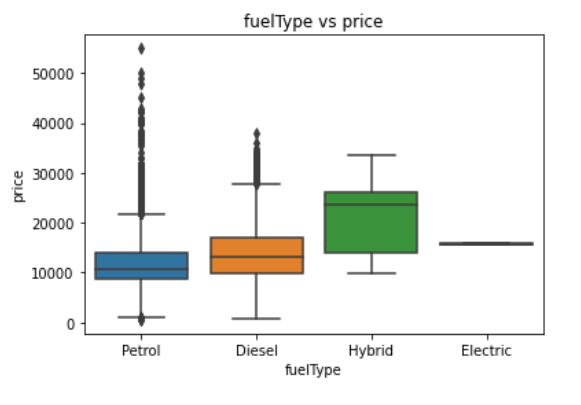
## Preprocessing the Data

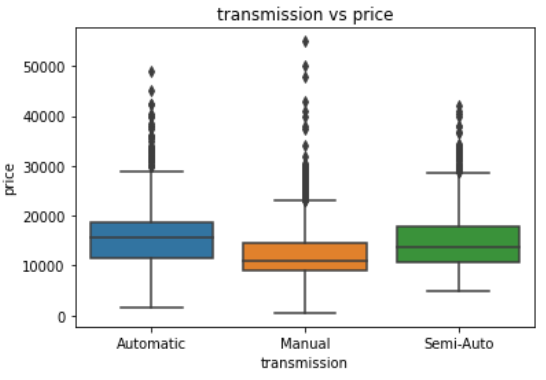
During the preprocessing phase, the data was cleaned by determining how many null values that were in the dataset. Luckily this dataset did not contain any null values. Also, a list of production years was produced. To that end, an observation with a production year of 2060 was present in the data and was subsequently removed. All other columns containing floats and integers contained values that were reasonable, and were left unchanged.

Several charts were created to better understand the data and to choose suitable features to use to train the regression models.









## Regression Model

After completing the preprocessing phase, the following features were selected to build the regression models:

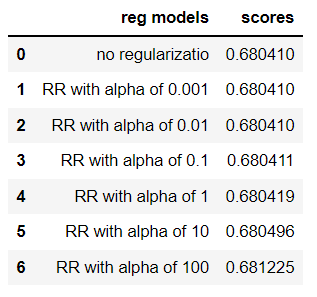
* model
* year
* transmission
* mileage
* fuel type
* mpg
* engine size

These features were used to predict the price of a Ford vehicle. The features were scaled and and then used to train the regression models.

Several regression models were created with the features mentioned above. The first model was a multi variable linear regression model. This model was used as a baseline to gauge the accuracy of an untuned model. Furthermore, six ridge regression models were created with alpha values ranging from 0.001 – 100. These ridge regression models were used to determine how tuning a hyperparameter would improve the accuracy of the predictions.

## Results

The results showed as seen in the figure below that the ridge regression models performed slightly better than the model without any regularization. It was also noted that increasing the value of alpha continued to improve the model’s accuracy.



This suggests that the baseline model (no regularization) was very complex, and reducing the complexity of the mode improves the model’s ability to predict unseen data.

## Conclusion

Several regression models were created to predict the price of a Ford vehicle. The most accurate model was the model that was tuned with an alpha value of 100. It is important to note that this model can be improved by:

* Performing cross validation as doing this will make the model more statistically significant as errors are more generalized across multiple test sets
* Using log transform to transform skewed distributions in the dataset to a normal distribution
* Training the model with more data

More work can be done to improve the model and can be revisited in the future.