Abstract:
For this research, we conducted a comparative study on performance of regression based on supervised machine learning models. Each model is trained using data of used car data collected from https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data.. As a result, Multiple linear regression gives the best performance with mean absolute error (MSE) = 1.2 x10^7 and R^2 = 0.7644760964608249

Introduction:
Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, horsepower, body style, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices

Data Source:
The data set which I use for this project is "Automobile Data Set". You can find the data by clicking on a link above.

Fields including:

- **normalized-losses:** about normalize losses
- **make:** Who make car.
- **fuel-type:** Fuel type **(**gas Or diesel)
- **aspiration:** About aspiration
- **num-of-doors:** How many doors car have(Two or Four)
- **body-style:** About body style
- **drive-wheels:** Front wheel drive(fwd) Or Rear wheel drive(rwd) Or Four wheel drive(fwd)
- **engine-location:** Engine location (Front OR Rear)
- **wheel-base:** car base wheel
- **length:** Length of the cars
- **width:** Width of the Cars
- **height:** Heights of the Car
- **curb-weight:** The total mass of a vehicle with standard equipment and all necessary operating consumables
- **engine-type:** What type of engine Car have
- **num-of-cylinders:** How many cylinders Car have.
- **engine-size:** size of engine
- **fuel-system:** type of fuel system Car have.
- **Bore:** In a piston engine, the bore (or cylinder bore) is the diameter of each cylinder
- **Stroke:** about stroke
- **compression-ratio:** degree to which the fuel mixture is compressed before ignition. **horsepower:** How much horsepower Car have.
- **peak-rpm:** revolutions per minute
- **city-mpg:** refers to driving with occasional stopping and braking, simulating the conditions you're likely to run into while driving on city streets
- **highway-mpg:** is based on more continuous acceleration, which usually yields a higher figure because it's a more efficient use of the engine.
- **price:** Price of the Car
- **city-L/100km:** Fuel consumption while drive in a city.

Libraries we use:

The Libraries I use for this Project are following:

I. **Pandas:** For importing data from data source to Juptyer notebook, and also for manipulating data.
II. **Numpy:** For handling missing or null values.
III. **Matplotlib:** For visualizing plots
IV. **Seaborn:** For plotting regression plots
V. **Sklearn:** For applying multiple linear regression.

Steps Followed For This Project:

People use different steps to solve a problem, according to their skills and need.

The steps which I followed to complete this task are:

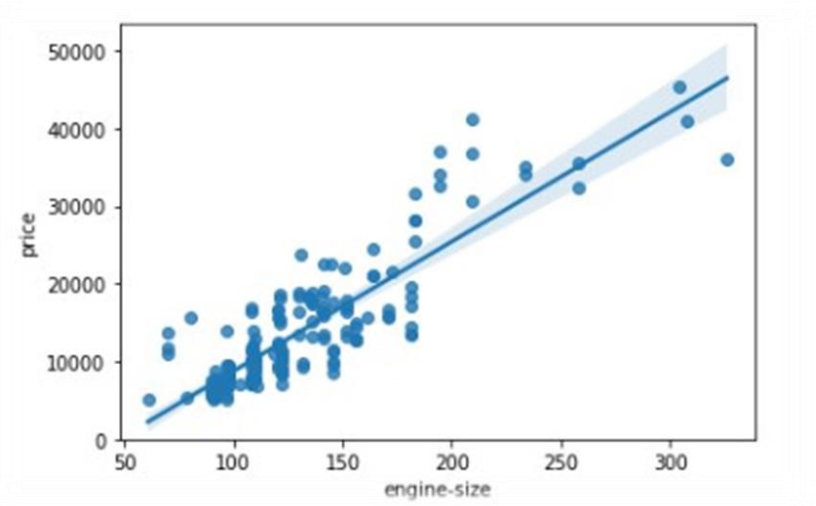
I. Importing data
II. Finding null values
III. Exploratory data analysis
IV. Descriptive statistics
V. Find Correlations
VI. Extract useful attributes
VII. Apply regression

Methodology and Results:

First of all I import data in to Jupyter notebook. The data sets had no columns names or headers. So I gave headers. After that, I replace "?" with NaN (Not a Number), which is Python's default missing value marker, for reasons of computational speed and convenience. Then, I drop NaN rows in some places, and in some places I **replace Nan** with mean of that column where they are found. The next step is to convert each variable into **correct data type**. For instance, the price variable had object data type, but it should be numeric. Because, the data we have in the price variable are all numbers with decimal point. When visualizing individual variables, it is important to first understand what type of variable you are dealing with. This will help us find the **right visualization** method for that variable.

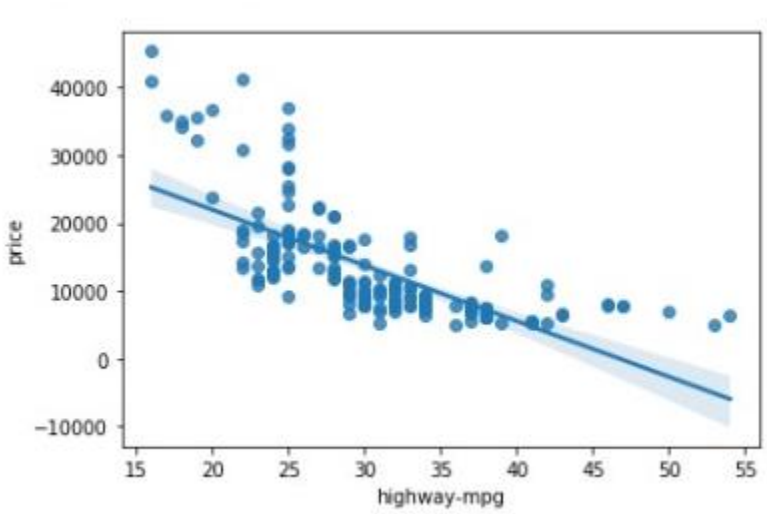
Furthermore, my goal is to predict Car price on the basic of its features. That's why, I have to check which **features** effects the Car price most. In order to start understanding the **(linear)** relationship between an individual variable and the price. We can do this by using "regplot", which plots the scatter plot plus the fitted regression line for the data.

So what I found is, **engine-size** goes up, the **price** goes up: this indicates a positive direct correlation between these two variables. Engine size seems like a pretty good predictor of price since the regression line is almost a perfect diagonal line.



And the **correlation** between 'engine-size' and 'price' and see it's **approximately 0.87**.

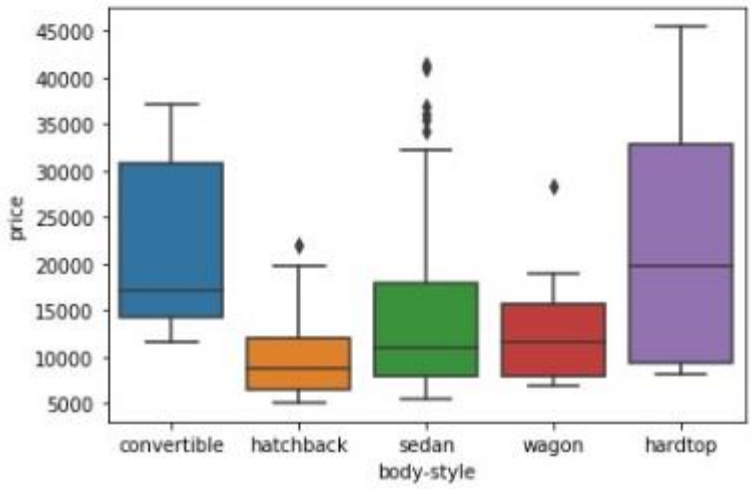
Similarly, the **correlation** between **'highway-mpg'** and **'price'** and see it's **approximately -0.704**



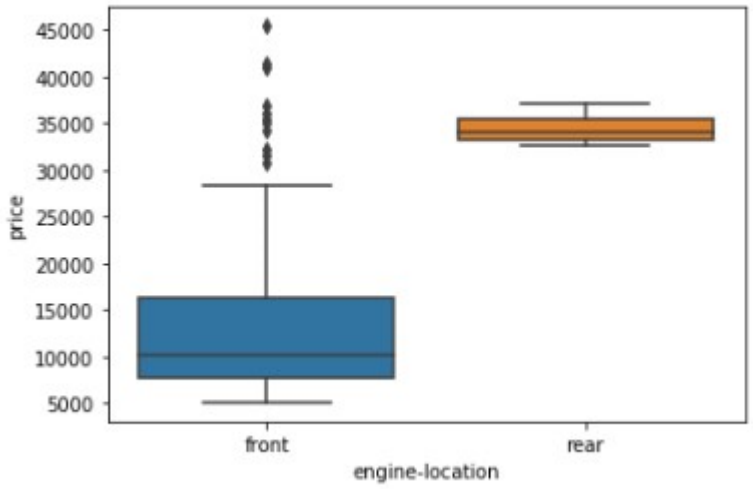
It's indicates an inverse/**negative relationship** between these two variables, because when the highway-mpg goes up, the price goes down. So, Highway mpg could potentially be a predictor of price.
Note: We use regression and scatter plots only when we have **continuous data**.

Moreover, we have **categorical variables** as well. They are variables that describe a 'characteristic' of a data unit, and are selected from a small group of categories. The categorical variables can have the type "object" or "int64". So, the best way to visualize categorical variables is by using **box plots.**

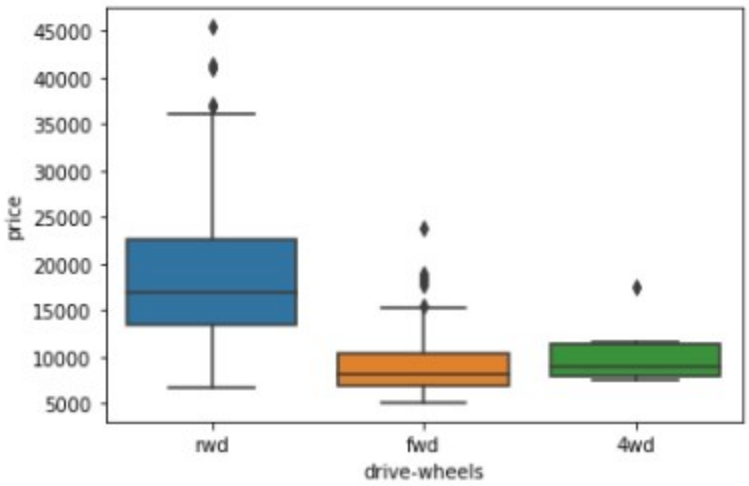
So, after examining relationship between "body-style" and "price" using box plot.



I able to see that the distributions of price between the different body-style categories have a significant overlap, and so **body-style** would not be a good predictor of price.



Similarly, I see that the distribution of price between these two **engine-location** categories, front and rear, are distinct enough to take engine-location as a potential good predictor of price. Also, I found that the distribution of price between the different drive-wheels categories differs; as such drive-wheels could potentially be a predictor of price.



Also I use **Pearson Correlation** measures the linear dependence between two variables, and the results are given blow.

```
1- The Pearson Correlation Coefficient between 'wheel-base' and 'price'  is 0.5846418222655083  with a P-value of P = 8.0764882
70732873e-20
2- The Pearson Correlation Coefficient between 'horsepower' and 'price' is 0.8096068016571052  with a P-value of P =  6.2735362
70650862e-48
3- The Pearson Correlation Coefficient betwoon 'engine-size' and 'price' is 0.8723351674455185  with a P-value of P = 9.2654916
22198389e-64
4- The Pearson Correlation Coefficient between 'bore' and 'price' is 0.5431537659807725  with a P-value of P =   8.051208825441
932e-17
5- The Pearson Correlation Coefficient between 'highway-mpg' and 'price' is -0.7046922650589533  with a P-value of P =  1.74954
71144474617e-31
```

Conclusion:

1- Since the p-value is < 0.001, the correlation between wheel-base and price is statistically significant, although the linear relationship isn't extremely strong (~0.585)

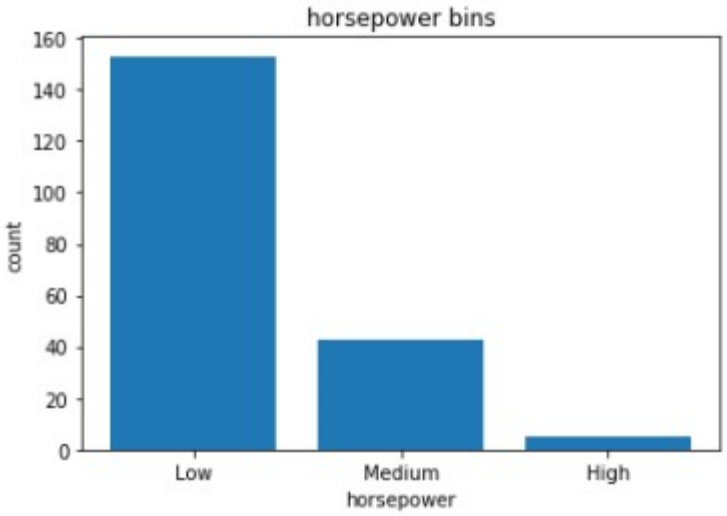
2- Since the p-value is < 0.001, the correlation between horsepower and price is statistically significant, and the linear relationship is quite strong (~0.809, close to 1)

3- Since the p-value is < 0.001, the correlation between engine-size and price is statistically significant, and the linear relationship is very strong (~0.872).

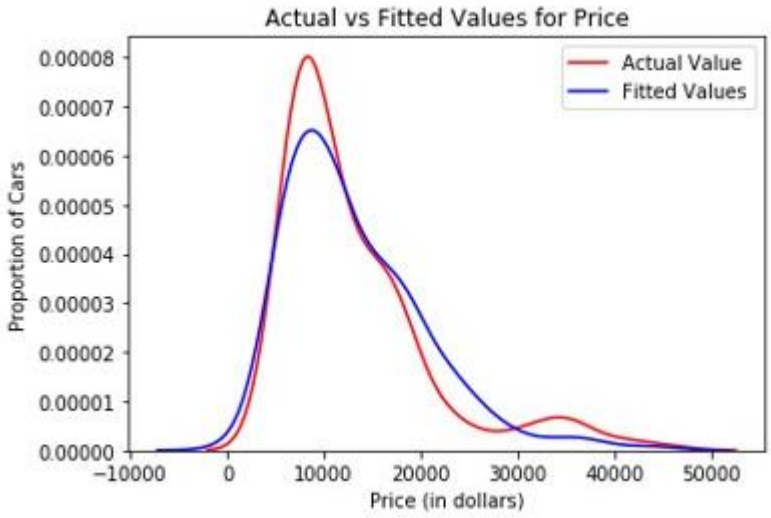
4- Since the p-value is < 0.001, the correlation between bore and price is statistically significant, but the linear relationship is only moderate (~0.521).

5- Since the p-value is < 0.001, the correlation between highway-mpg and price is statistically significant, and the coefficient of ~ -0.705 shows that the relationship is negative and moderately strong.

In our dataset, **"horsepower"** is a real valued variable ranging from 48 to 288, it has 57 unique values. What if we only care about the price difference between cars with high horsepower, medium horsepower, and little horsepower (3 types)?. For this I use **Binning.**



At last, I **fit** best features which I find into **multiple linear regression** model, and use **distribution** plot for visualization.



You can look at the distribution of the fitted values that result from the model and compare it to the distribution of the actual values.

Conclusion: Important Variables:

We now have a better idea of what our data looks like and which variables are important to take into account when predicting the car price. We have narrowed it down to the following variables:

Continuous numerical variables:

- Engine-size
- Horsepower
- City-mpg
- Highway-mpg
- Wheel-base
- Bore

Categorical variables:

- Drive-wheels