INDEX

INTRODUCTION

Machine Learning (ML) is currently trending in Information Technology (IT) sectors, with a growing recognition that ML can play a key role in a wide range of critical applications, such as data mining, natural language processing, image recognition, and expert systems and can also provide potential solutions for all these domains and more, can also be a pillar of our future civilization.

Predicting the price for new vehicles (or cars in this project) is more interesting and a challenging problem for many users, estimating the price of the car based on its features such as brand, horsepower, mileage and other important factors, it's not a simple task, to find a number that is beneficial for both seller and customer is quite challenging. We can use Machine Learning to overcome these problems.

In this project a small dataset of Automobile sales data is considered, the dataset contains essential attributes that influence retail price of cars. Price prediction can be carried out using supervised machine learning techniques by employing various regression algorithms to find the best price prediction equation suitable for the dataset.

Even though this approach might be interesting to calculate the average price of a cohort of similar vehicles, it does not take into account the differences among similar cars.

Problem Statement

Automobile sales are on a global increase, so there is a need for car price prediction system to effectively determine the worthiness of the car using a variety of features.

Objective

To predict the retail price of a car by training different machine learning regression models on the given dataset and evaluate the performance of individual models to select the one with high accuracy to maximize the efficiency of price prediction system.

Requirement Specifications

Hardware Requirements :

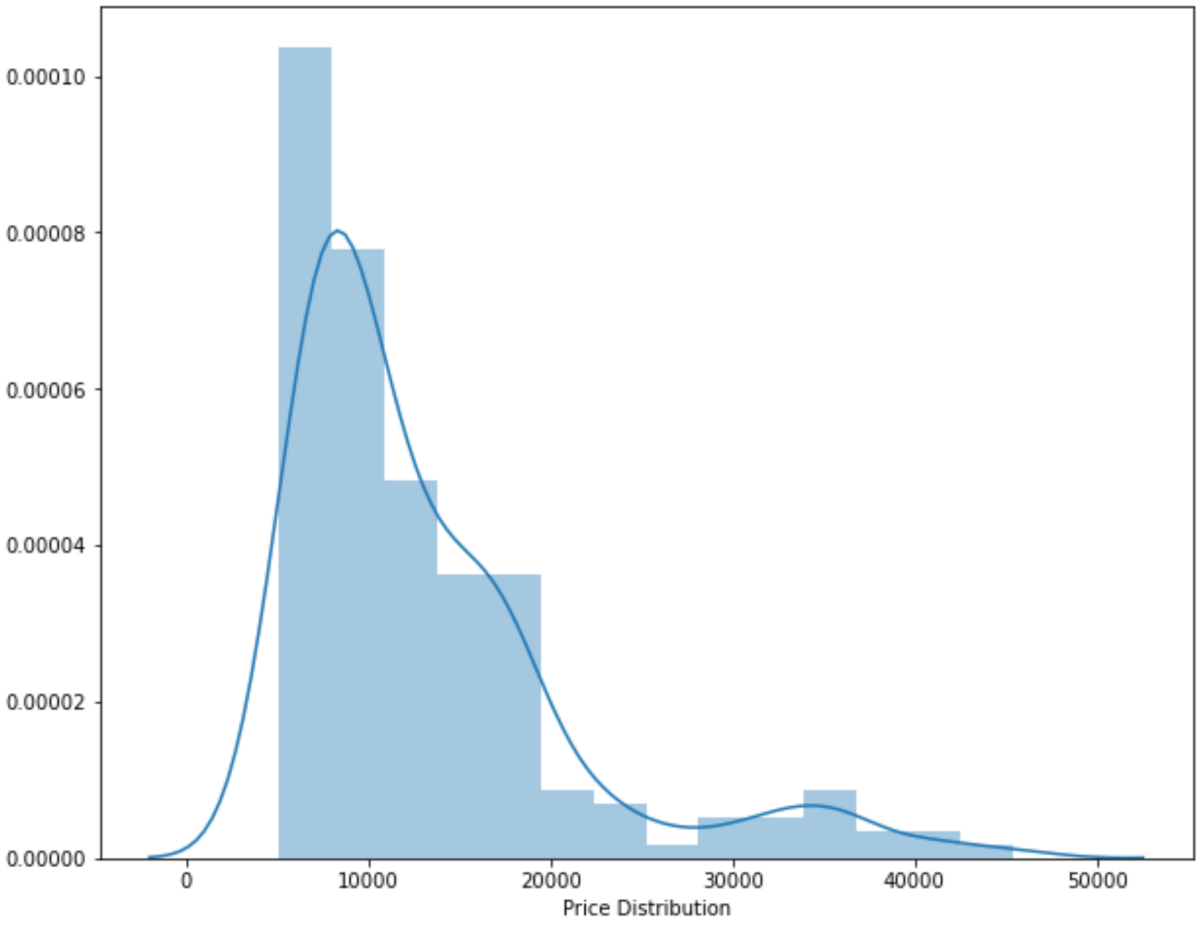
- Hard disk : 40GB Hard disk or higher.
- Processor : Intel i3 core/AMD Athlon or higher.
- Memory : 1 GB or higher.

Software Requirements :

- Windows OS XP/7/8/10.
- Python 3.6 and above.
- Jupyter notebook (*).

Exploratory Data Analysis

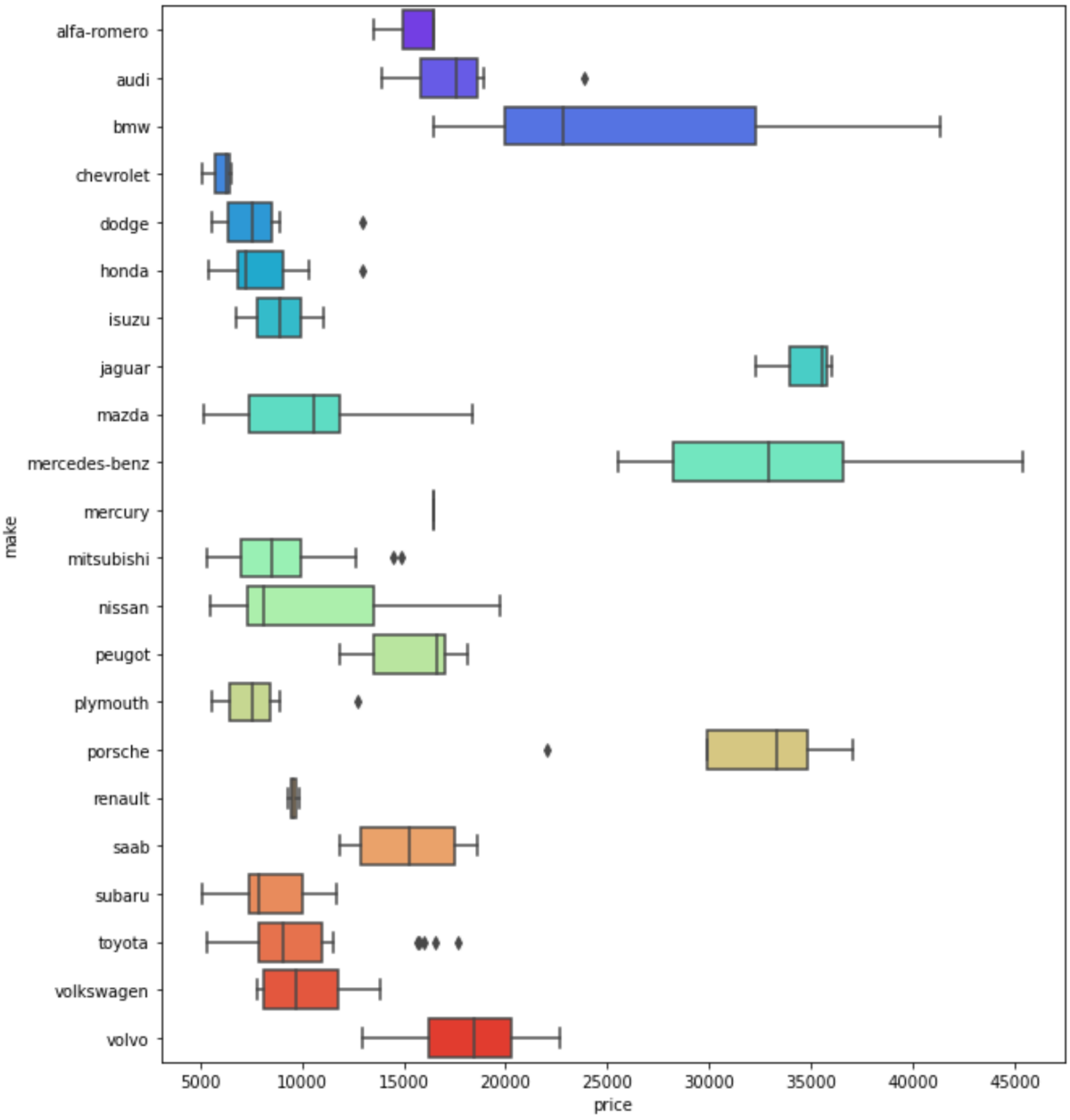
Distplot :



This plot shows car price distribution, we can observe that majority of the cars price ranges around 5000 to 18000 and we can also notice there are barely any cars above 45000 price range.

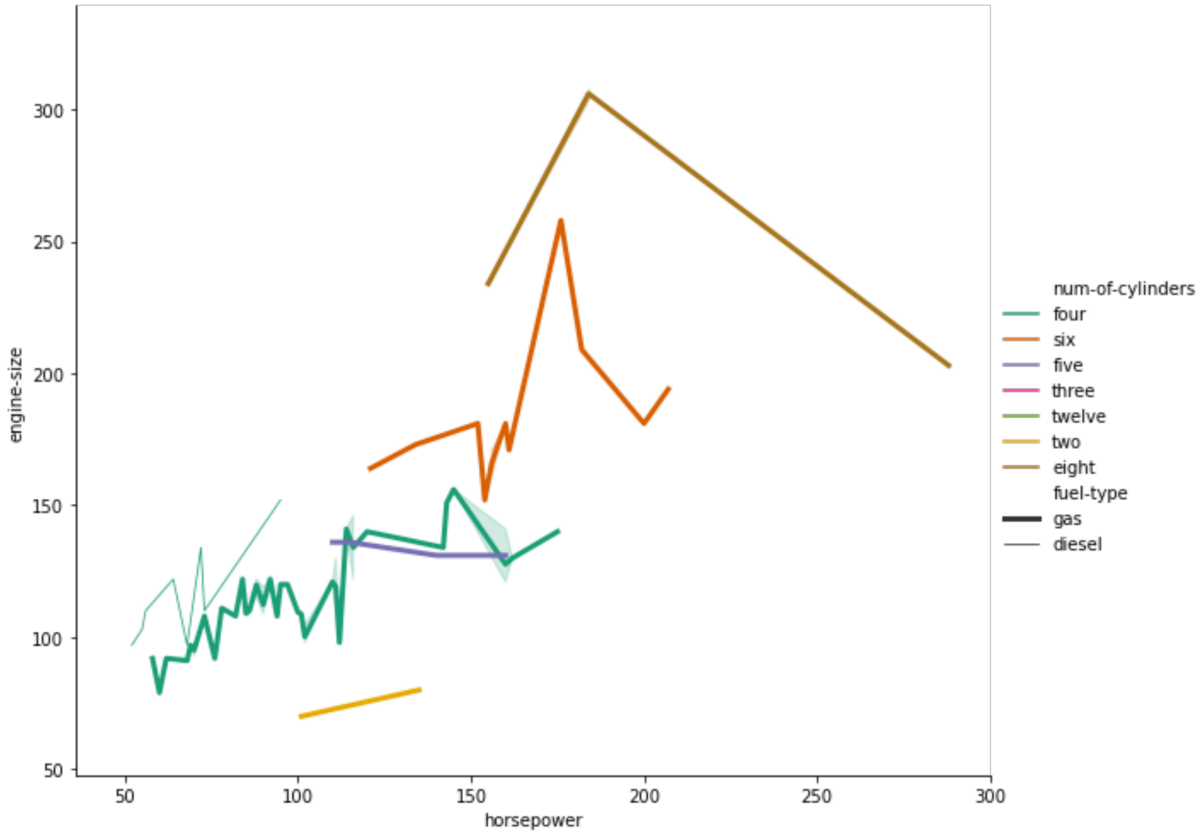
From this plot we can also find minimum car price and maximum car price that is 5000 and 45000 respectively for given dataset.

Boxplot :



From this plot we can observe the total price range of cars by different makers (manufacturers) and we can also classify cars as expensive or economical by their manufacturing companies. For example we can say porshe makes expensive cars and targeted towards smaller consumer base.

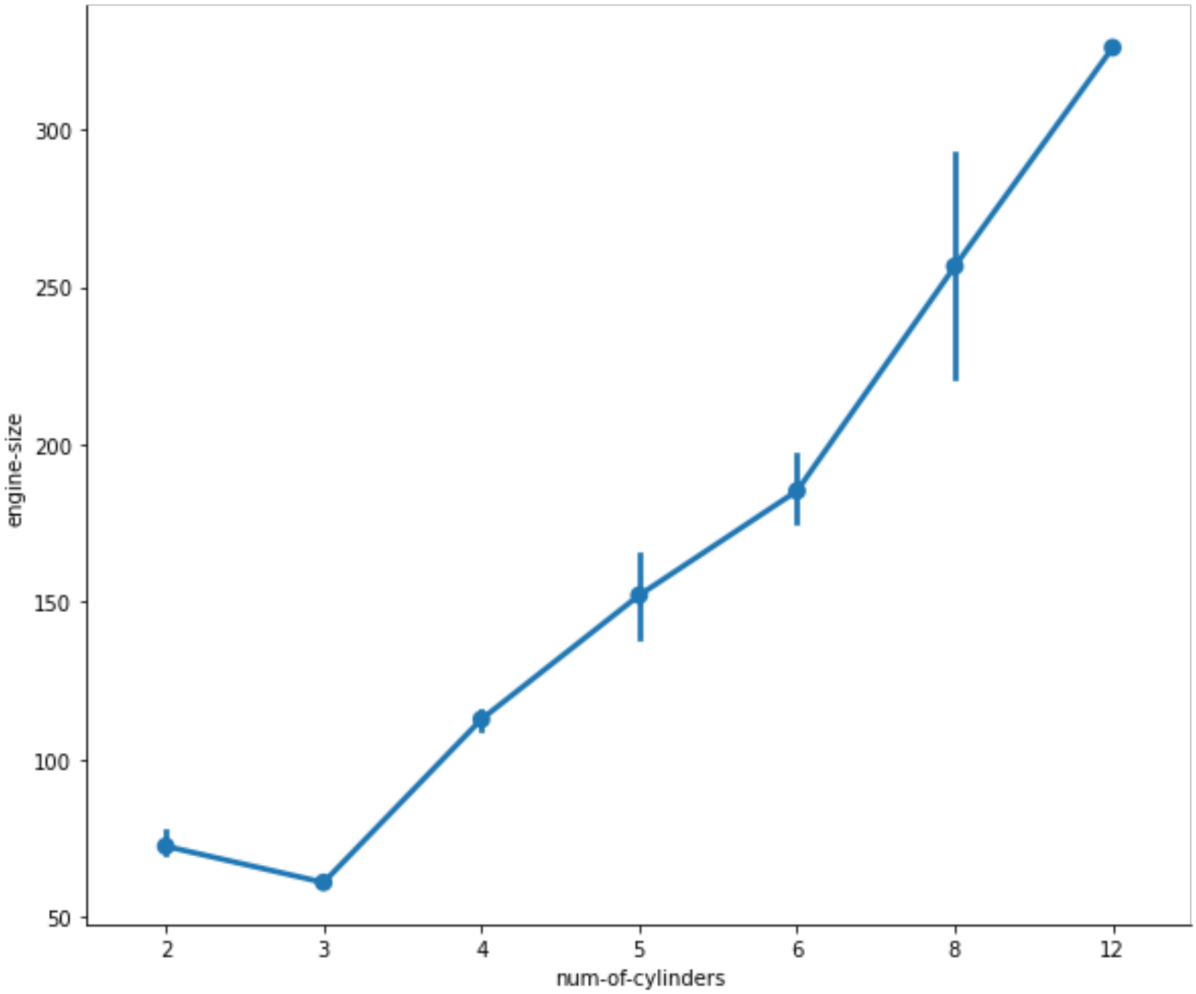
Relplot :



This plot shows the relation between horsepower and engine size of a car and how it varies with number of cylinders along with fuel type, generally we can say more horse power requires more engine size but that's not true all the time, consider number of cylinders as eight (light brown) after certain point the horse power increases when engine size is decreased.

We can also observe that cars using gas as their fuel type can produce more horsepower than diesel if engine size is kept constant, hence we can see more number of cars with gas as their main fuel type rather than diesel.

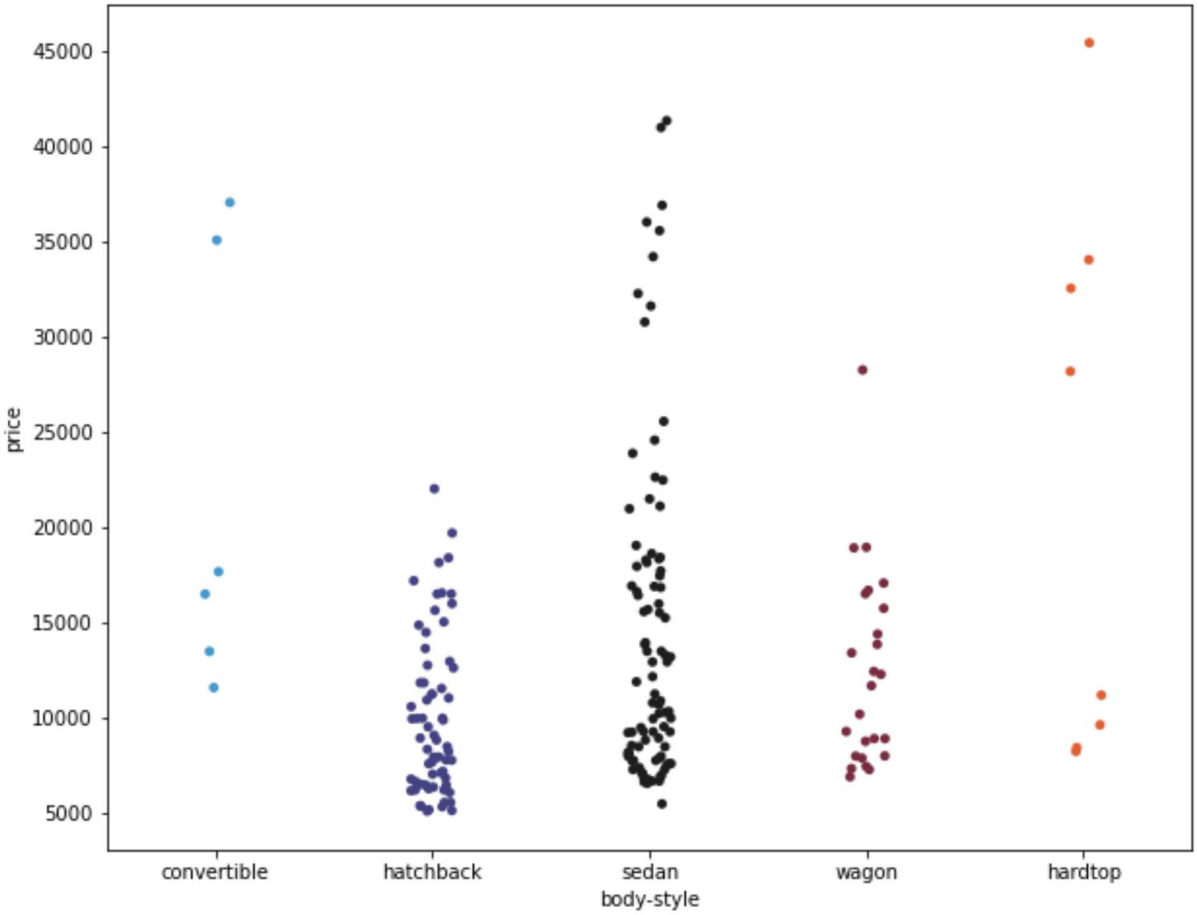
Catplot :



The above graph shows how engine size is associated with number of cylinders of a car.

The dots on the graph represent the mean of that column (number of cylinders) with respect to engine size and we can conclude that more number of engine cylinders results in increase of engine size.

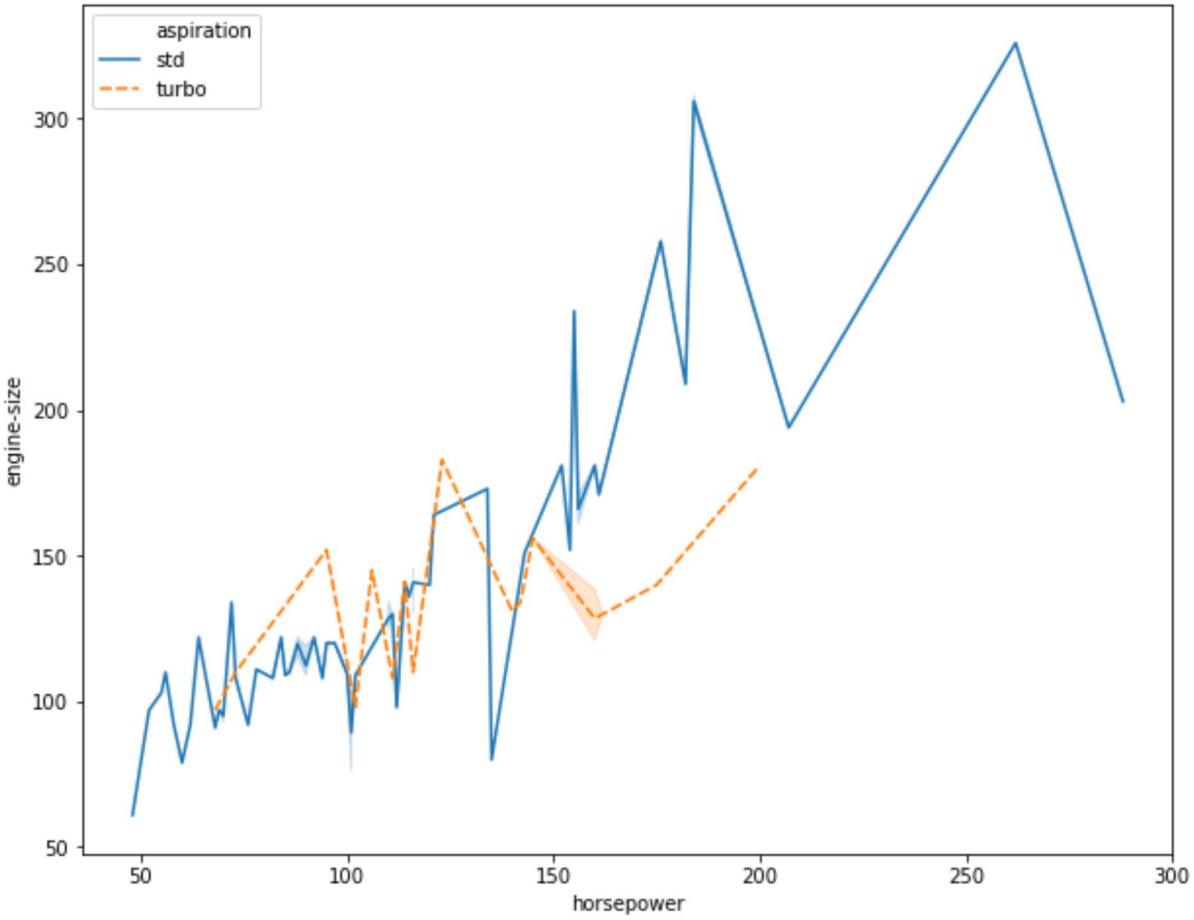
Stripplot :



The above plot shows what kind of body style manufacturers choose and how price of the car is associated with it. By looking at the graph we can say that most cars comes with either hatchback or sedan body style and some cars wagon body style.

Cars with convertible or hard top body style are mostly expensive, hence these cars are rare. Economical cars will have hatchback or sedan or wagon body style.

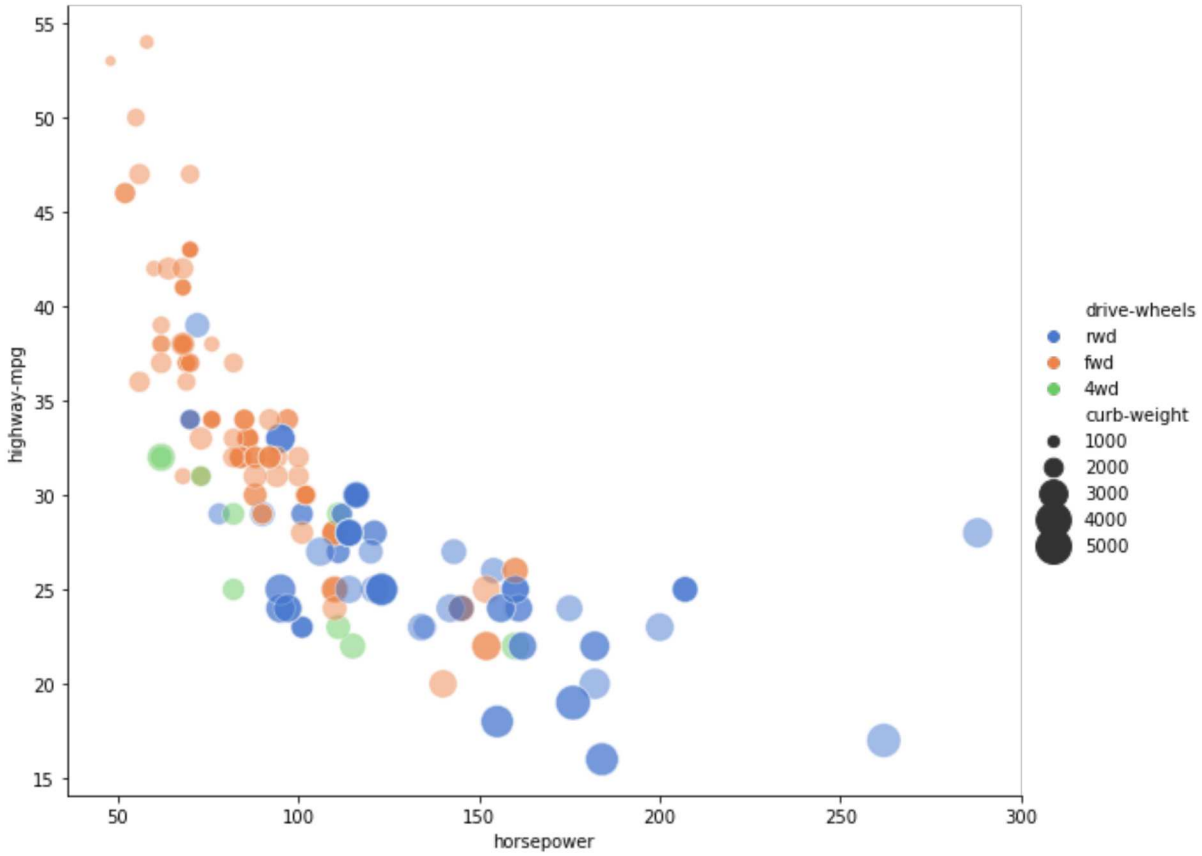
Lineplot :



This plot shows how a car's engine size and horse power are related to each other with respect to its aspiration.

We can say that most cars have standard aspiration, cars with turbo aspiration produces less horsepower and engine size should be larger to obtain same horse power as that of standard aspiration, but in horsepower range 110 to 150 the relation becomes inverse and turbo aspiration type is more efficient.

Relplot :



This plot shows relation between horsepower and highway-mpg and how it changes according to drive wheels and curb weight of a car.

Rear wheel drive (rwd) cars produce more horsepower but gives less highway-mpg whereas Front wheel drive (fwd) cars give more highway-mpg by compromising horsepower while Four wheel drive (4wd) cars are balanced somewhere in middle.

Curb weight of the car gradually increases along with horse power of the car.

Preparing Machine Learning model

Estimating the price of car depends on determining the association and character of the relationship between one dependent variable (‘Price’) and series of other independent variables (‘make’, ’num-of-doors’, ’fuel-type’, ’horsepower’ etc.). To achieve this we use Regression approach.

1. Linear Regression :

```
from sklearn.linear_model import LinearRegression

model1 = LinearRegression()

model1.fit(X_train, y_train)

pred1 = model1.predict(X_test)


from sklearn.metrics import r2_score

r2_score = round(r2_score(y_test, pred1),4)
```

0.8993

2. Extra Tree Regressor :

```
from sklearn.tree import ExtraTreeRegressor

model6 = ExtraTreeRegressor()

model6.fit(X_train,y_train)

pred6 = model6.predict(X_test)


from sklearn.metrics import r2_score

r2_score = round(r2_score(y_test, pred6),4)
```

0.9

3. Decision Tree Regressor :

```
from sklearn.tree import DecisionTreeRegressor

model2 = DecisionTreeRegressor()

model2.fit(X_train, y_train)

pred2 = model2.predict(X_test)


from sklearn.metrics import r2_score

r2_score = round(r2_score(y_test, pred2),4)
```

0.9049

4. Gradient Boosting Regressor :

```
from sklearn.ensemble import GradientBoostingRegressor

model4 = GradientBoostingRegressor()

model4.fit(X_train,y_train)

pred4 = model4.predict(X_test)


from sklearn.metrics import r2_score

r2_score = round(r2_score(y_test, pred4),4)
```

0.9053

5. Random Forest Regressor :

```
from sklearn.ensemble import RandomForestRegressor

model5 = RandomForestRegressor()

model5.fit(X_train, y_train)

pred5 = model5.predict(X_test)


from sklearn.metrics import r2_score

r2_score = round(r2_score(y_test, pred5),4)
```

0.9217

ML model chart

Sl. No	Algorithm Name	Metric used
1.	Random Forest Regressor	r2_score - 0.9217
2.	Gradient Boosting Regressor	r2_score - 0.9053
3.	Decision Tree Regressor	r2_score - 0.9049
4.	Extra Tree Regressor	r2_score - 0.9
5.	Linear Regression	r2_score - 0.8993

Conclusion

In this project, five different machine learning regression algorithm have been used to forecast the price of cars. All these techniques yield with approximately 90 percent accuracy and the results are acceptable.

Among these five algorithms Random Forest (Regression Version) algorithm gave higher accurate results and can be considered as the best fitting model for the given dataset.

Although this is a good project but we still haven't used Machine Learning to its full potential and the dataset used in this project is very small, for future enhancement a large dataset can be used to apply deep learning and obtain even more accurate results by applying concepts of neural networks.

Bibliography

Websites :

1. https://towardsdatascience.com
2. https://seaborn.pydata.org
3. https://jupyter.org

Books :

1. Hands on Machine Learning with Scikit-Learn.