CAR PRICE PREDICTION

Submitted by:

RAJ

ACKNOWLEDGMENT

This is to acknowledge that: The entire data used for this project is collected from the website **CarsDekho**.

INTRODUCTION

- Business Problem Framing

  There are lakhs of cars of different brands and models available around us. If one wants to sell their used car or individual wants to buy they should know what the price range they can sell their car at or purchase at. There are different factors that affect the car price. So we made a Machine Learning model to predict the price range at which it can be sold or bought.

- Conceptual Background of the Domain Problem

  While selling or buying a used car there are many factors to be considered that determines the price of car.

  For instance, how many years old the car is, older the car price depreciates more. Based on the km driven in the car the more distance the vehicle travelled there will be more price depreciates , since wear and tear of the car parts and need to be changed if the parts are replaced recently then the price can be increased a bit. In general the parts that to be changed based on some distance travelled factor are like Wheels, Engine Oil, Break Pad, Break oil, Clutch Plates and many more.

  Price also depends on the colour, variant, type of car i.e. SUV or Sedan or Hatchback, version of car w.r.t to brand, model. Price will be reduced if there are any damages on the body of car, car is encountered with accident and insurance claims etc.

  Distance travelled and years old are generally correlated where in general older the car more distance might be driven. Even the same care based on the place their price varies due to different tax rates etc.

If the cars are older than 20 years or more they might be considered as scrap and can be sold or bought only at scrap value. Which doesn’t cost more than 25 to 30k.

- Motivation for the Problem Undertaken

  My hunch and interest towards cars and also to improve the efficiency or create awareness to seller / buyer of the fair price of the used car.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Our target is to predict the price of car. The price is a continuous variable. So we use regression models to predict the target.

  Below the list of models used their accuracy, cross validated mean and difference between them.

Algorithm	R2 Score	Mean CV Score	Difference
Linear Regression	41.3	38.5	2.8
SVR	53.1	52.1	1.0
Random Forest	88.7	83.0	4.7
Gradient Boosting	78.0	76.0	2.0

  From the above we can see the algorithms Random Forest and Gradient Boosting are giving more accuracy. When compared with Cross Validation Score Gradient Boosting algorithm is better, since less difference.

  So we used this algorithm as our final model for production.

- Data Sources and their formats

  The complete data we used in this project is collected from the website: CarDekho

  The details scraped from their website are stored in the format of ".csv ".

  In total the dataset have: 12070 rows and 11 columns

	Unnamed: 0	car_type	location	brand	model	version	year	fuel_type	transmission_type	distance_driven	price
0	0	-	Chennai	Maruti	Ertiga	VDI	2012	Diesel	Manual	64,523 kms	6.1 Lakh
1	1	-	Chennai	Maruti	Wagon R	AMT VXI	2018	Petrol	Automatic	29,000 kms	5.15 Lakh
2	2	-	Chennai	Hyundai	Verna	CRDi 1.4 EX	2019	Diesel	Manual	38,000 kms	9.55 Lakh
3	3	-	Chennai	Maruti	Alto	LXI	2006	Petrol	Manual	91,500 kms	1.4 Lakh
4	4	-	Chennai	Ford	Ecosport	1.5 DV5 MT Titanium	2014	Diesel	Manual	45,000 kms	6.25 Lakh

The columns and their details:

Unnamed: 0 – Index values

car_type: The body type of car Hatch back/Sedan/SUV etc.

String format – Categorical column -

object Location: Name of the city / location

String format – Categorical column -

object Brand: Manufacturing company name

String format – Categorical column -

object Model: The name of car

String format – Categorical column -

object Version: Variant of the car w.r.t to

model

String format – Categorical column - object

Year: The year when the car is manufactured or bought by the individual.

Numerical format – Categorical column -integer

Fuel_type: This describes the type of engine of car like

diesel/petrol String format – Categorical column -

object

Transmission_type: This describes what kind of gearing mechanism manual/automatic

String format – Categorical column - object

Distance_driven: This is how many km car has been

driven

Numerical format – Continuous data -

int Price: The selling cost listed on the website

Numerical format – Continuous data - Float

- Data Preprocessing Done
  - Loaded dataset
  - Checked their data types
  - Changed the data types into correct where there are faults by observing, modifying the data into correct format.
  - Removed unnecessary columns
  - Got insights of each column and observed trends
  - Removed outliers
  - Reduced the skewness which is found in the column distance driven.
  - Encoded the categorical columns and found correlation
  - Separated the target from features
  - Scaled the features
  - Split the data into train and test to make the model
  - Then the data is trained and tested with various algorithms
  - Found the best model and applied tuning to improve accuracy.
- Data Inputs- Logic- Output Relationships

  Year: Number of years old the car is, the price reduces. The depreciation rate will increase by how older the car is.

  Distance Driven: More the distance the vehicle travelled, more the wear and tear on the parts of vehicle, so the cost will reduce.

  These are two major factors in evaluating the car price.

  Brand: Categorical type – There are lot of brands which manufacture cars, some cars based on brand, features of the car and the price varies. This variation due to lot of reasons like their standards, materials used, if they are imported from other countries due to import or export duty and of local manufacturing so there will be difference of car price. So even when resale this makes little impact to price.

  Location: Categorical type – The price of car varies based on its location since variation in tax rates, or transportation costs occurred, while resale /purchase there will be transfer costs occurred etc.

Model and Version: There are different models and versions of car manufactured by brand, like based on seating capacity, kind of engine, volume of engine, and other this model plays some impact in price. If it is a good model, with top version price might be bit higher.

There might be few failure models, where the price will be drastically less.

Fuel Type: Based on the fuel used, there will be changes in engine which make them costlier or less price. In general diesel engines are more robust in build which makes diesel cars are costlier to other fuel type. Each engine has its own pros and cons.

Transmission type: Based on the type of transmission the cost of manufacturing and some other extra things involved in the car. So automatic car prices are bit more than manual.

- State the set of assumptions (if any) related to the problem under consideration

Assumed the version of car plays minute role in determining the price. So dropped that column.

- Hardware and Software Requirements and Tools Used Minimum Hardware Requirements: I3 processor, 4 GB ram Software Requirements:

  Numpy, Pandas, Scikitlearn, Seaborn, Ensemble, Scipy modules Tools Used:

  Selenium for scraping the data

  Python, Anaconda Framework, Jupyter Notebook

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  From problem we need to find the variable price which is continuous. So we use the Regression model to predict.

  The approach I took is, the price of a car varies on various factors.

  From the available data, first I want to know the location, brand of the car, then we will consider the model of car. These are my first few inputs then what kind of engine, transmission type. Distance driven, years old are my two major factors that affects the price of car.

  Considering all these factors, I tried different regression models on the cleaned and encoded data.

  Methods implemented:

  Linear

  Regression

  SVR,

  Decision Tree Regression,

  Gradient Boosting

  Regression

- Testing of Identified Approaches (Algorithms)

  I tested the same trained data with the algorithms chosen.

Algorithm	Mean Squared Error	R2 Score	CV Score	Difference
Linear Regression	42.7	41.3	38.5	2.8
SVR	34.1	53.1	52.1	1.0
Random Forest	8.1	88.7	83.0	4.7
Gradient Boosting	16.0	78.0	76.0	2.0

- Run and Evaluate selected models

```
: from sklearn.model_selection import cross_val_score
ml_models=[LinearRegression(),SVR(),RandomForestRegressor(),GradientBoostingRegressor()]
for m in ml_models:
    m.fit(x_train,y_train)
    predm=m.predict(x_test)
    mse=mean_squared_error(y_test,predm)
    mae=mean_absolute_error(y_test,predm)
    r2=r2_score(y_test,predm)
    print(f'metrics of {m}:')
    print(f' mean_absolute_error: {mae}\n mean_squared_error: {mse}\n r2_score: {r2} ')
    score=cross_val_score(m,x_scaled,y, cv=5)
    print(' mean cv score:',score.mean())
    print('\n\n')
```

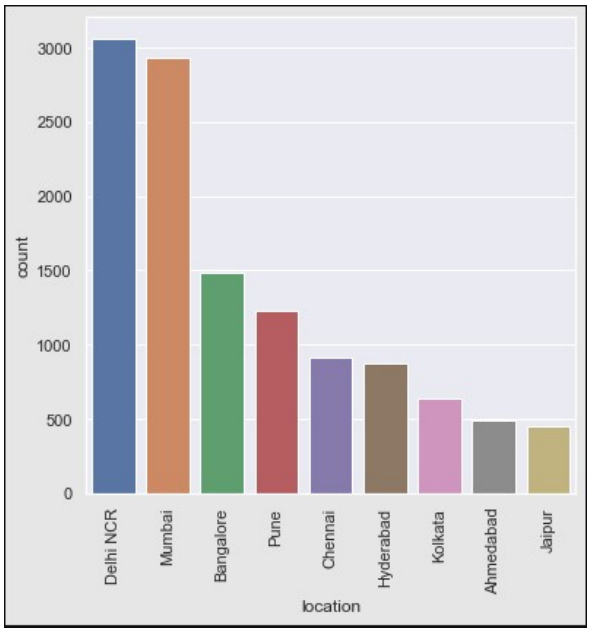
```
metrics of LinearRegression():
 mean_absolute_error: 3.8183894948676045
 mean_squared_error: 42.690310110761274
 r2_score: 0.413367102816257
 mean cv score: 0.38477066252901365


metrics of SVR():
 mean_absolute_error: 2.611479700854225
 mean_squared_error: 34.1094426112222
 r2_score: 0.5312818977321063
 mean cv score: 0.5216519097238044


metrics of RandomForestRegressor():
 mean_absolute_error: 1.0938853562967288
 mean_squared_error: 8.178780219524953
 r2_score: 0.8876105251247739
 mean cv score: 0.82945447534738


metrics of GradientBoostingRegressor():
 mean_absolute_error: 2.0392400106013295
 mean_squared_error: 16.03248300890133
 r2_score: 0.7796881322211316
 mean cv score: 0.7610368671696015
```

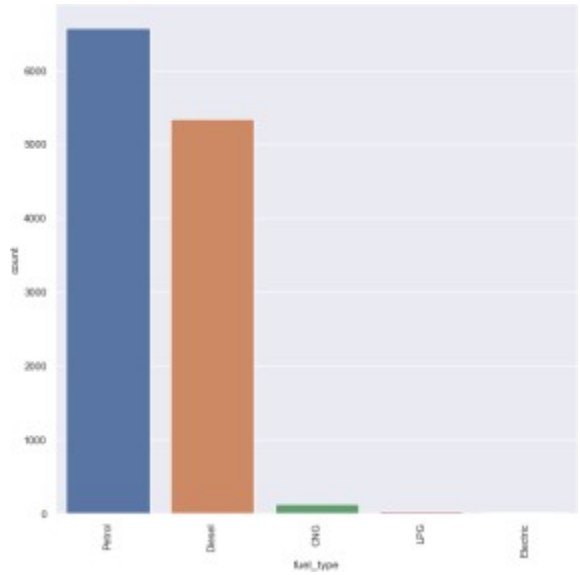
- Visualizations



We can see how many cars data from each city we have.

More car details are from cities Delhi and Mumbai.

Least car details are from cities Jaipur and Ahmedabad

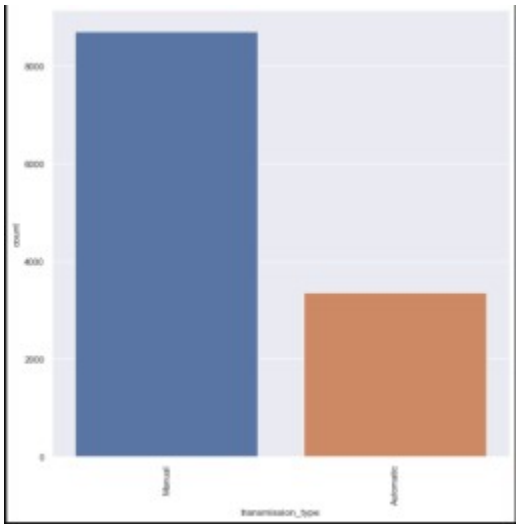


We can see how many cars data of different fuel types.

More cars are petrol type and least electric

fuel_type	year	distance_driven	price
CNG	6.277778	60470.666667	3.768016
Diesel	6.301379	70725.855961	10.440471
LPG	11.423077	67726.615385	2.952000
Petrol	6.493144	47892.028730	5.485574

Average price, Year and distance based on fuel types

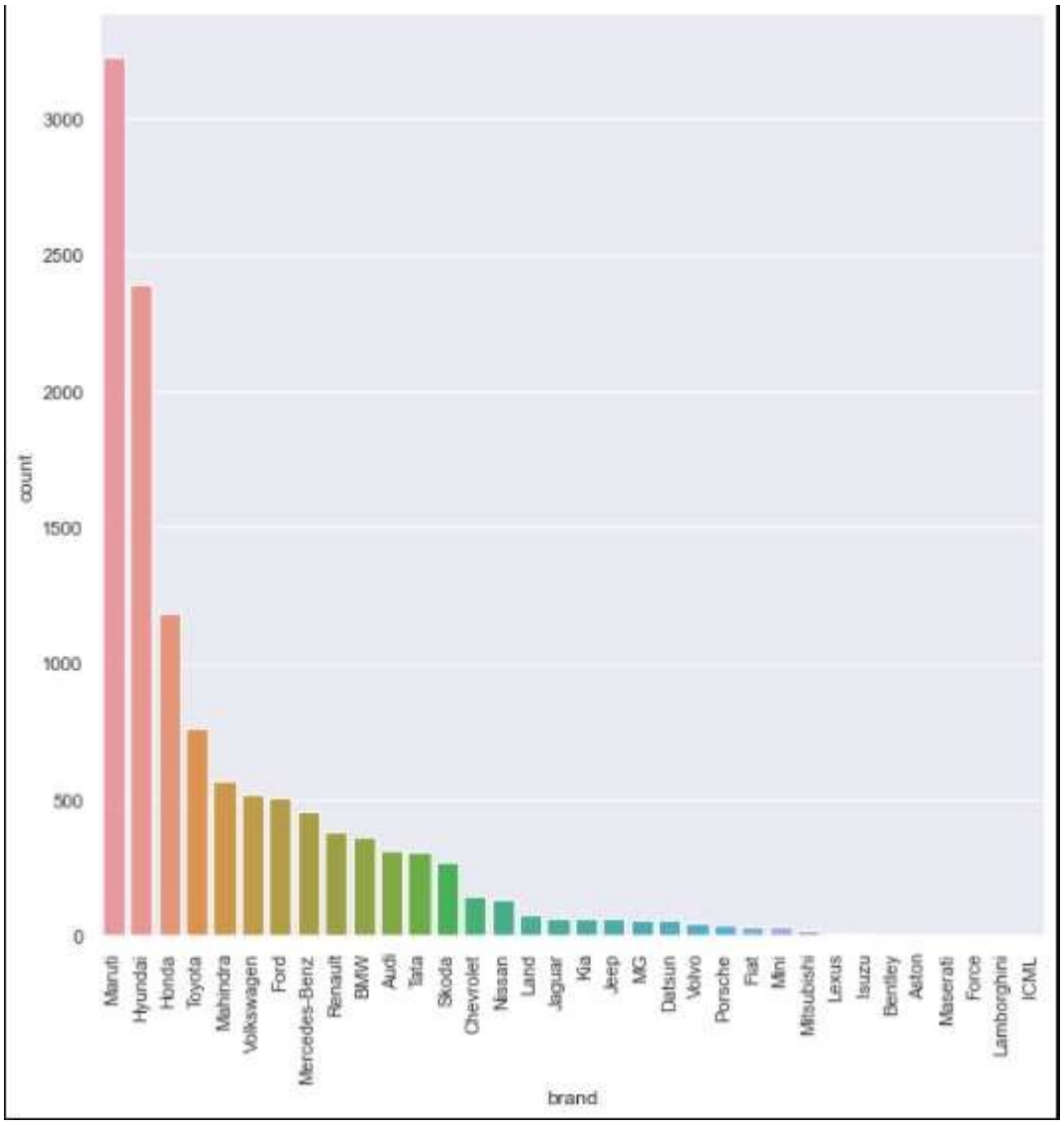


Majority of the cars are manual type in our data.

Manual

Automati

c



Different branded cars and their count in our data

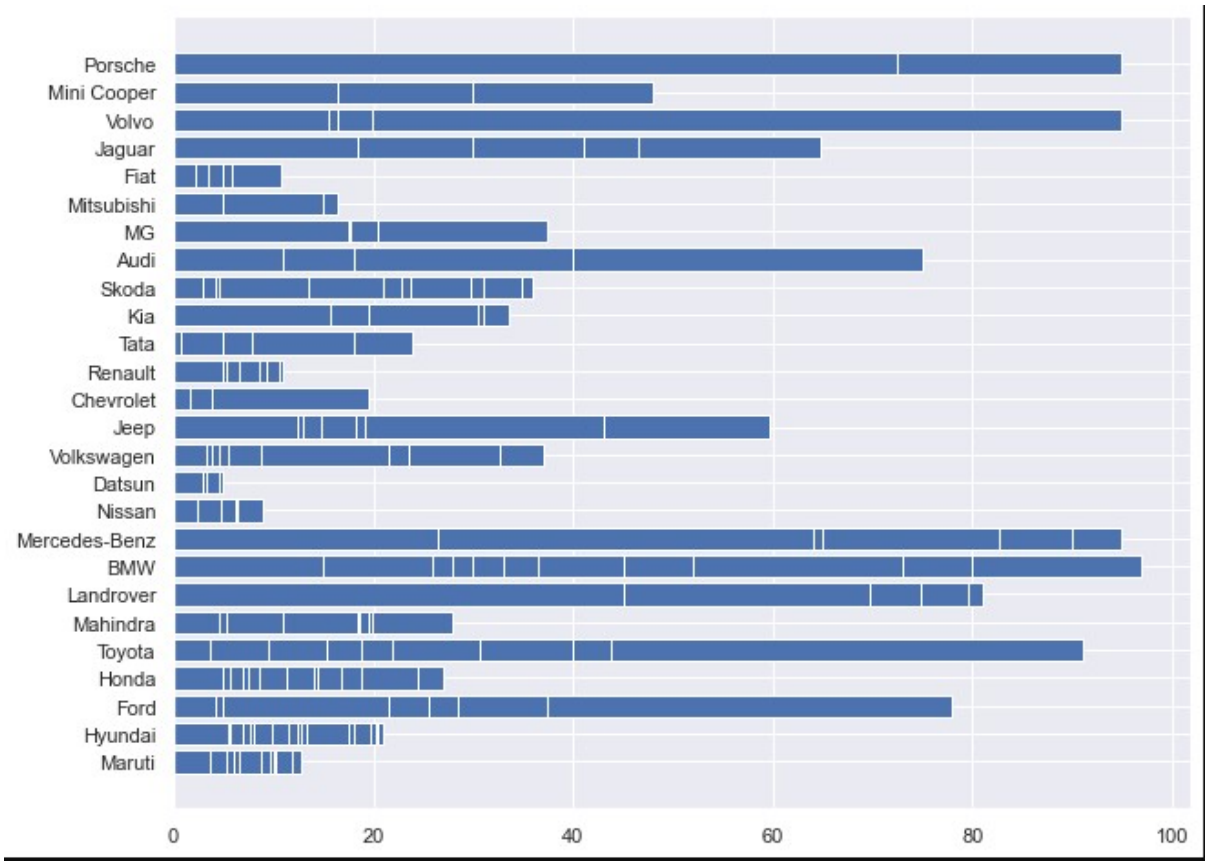
There are 34 different branded cars in our

dataset. Majority of the cars are of brand Maruthi

and Hyundai

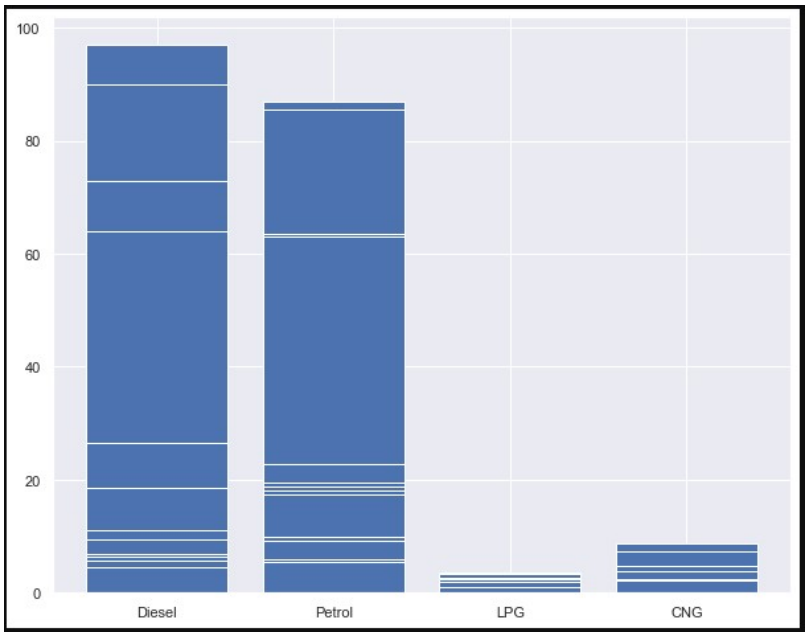
There is a brand named Land, Mini instead of Land Rover, Mini Cooper so we will make it correct.

There are few Luxury / High premium cars like lexus, Isuzu, etc where very minute data. So let us drop those rows with data < 10 since that data is not useful to predict.

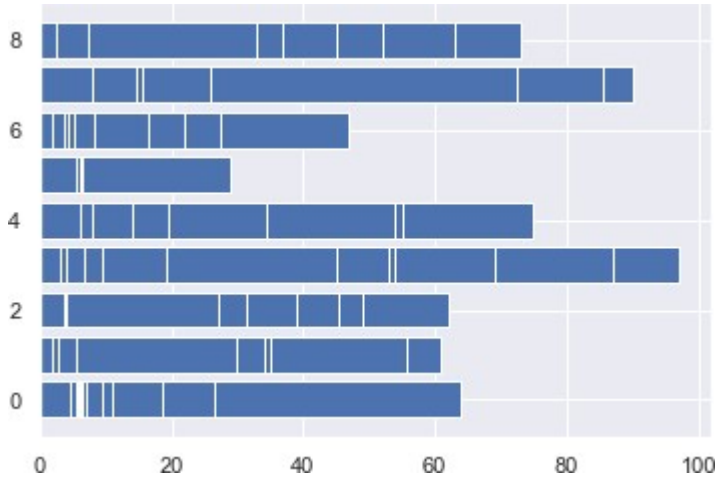


Price range of car with respect to brands.

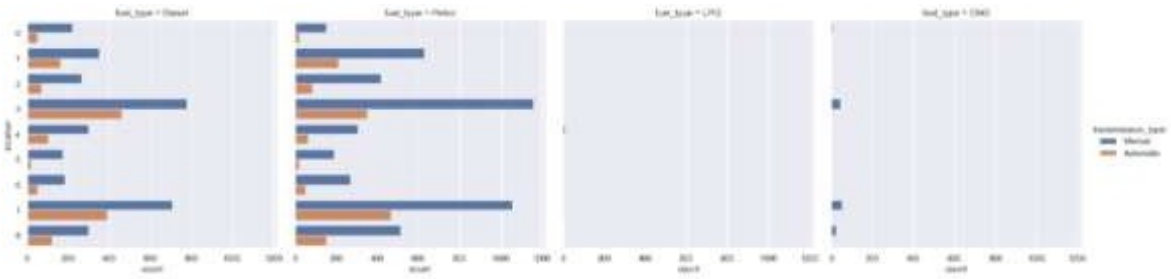
BMW has the highest price and Datsun has the least price



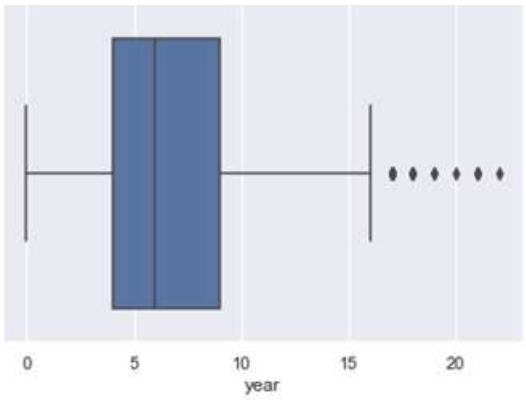
Price of Diesel cars is more compared to that of petrol



Prices based on location



No. of automatic or manual cars in each location based on the fuel type.



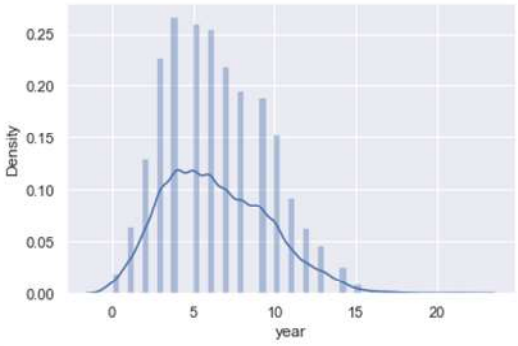
Few outliers are present in this column

Skewness In the column year normally distributed.

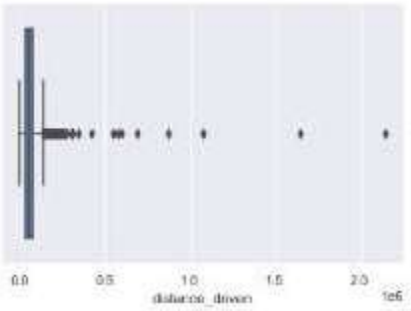
Skewness is 0.47, acceptable range.

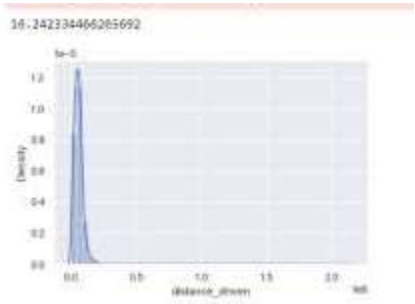
```
In [54]: sns.distplot(df_new['year'])
         print(df_new['year'].skew())
           warnings.warn(msg, FutureWarning)

0.4726957214279264
```

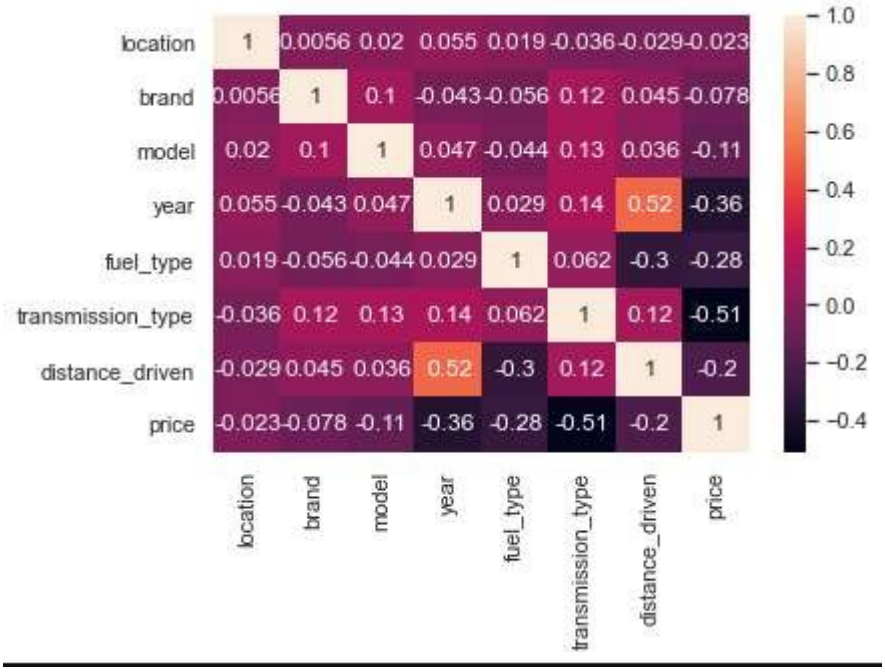


Distance driven: Many outliers in this column, also the skewness is more.





Correlation between different columns w.r.t to each other.

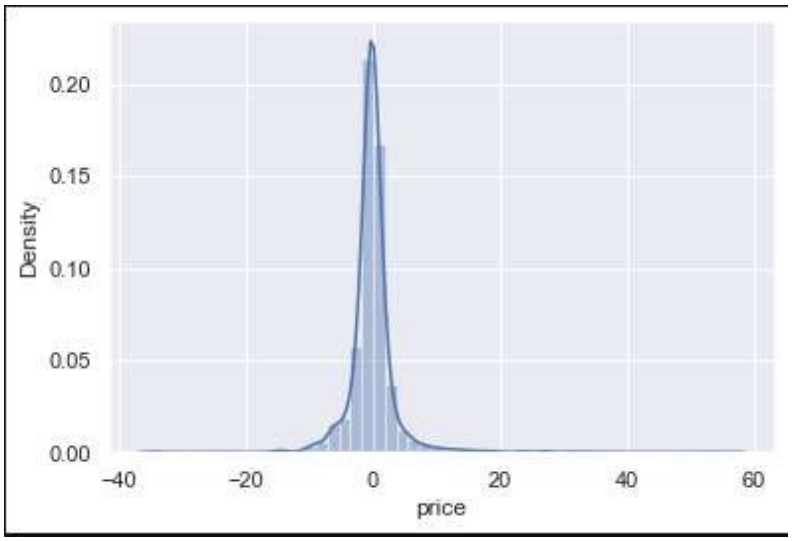


Observation:

We can see year and distance driven are correlated with each other. Price is highly –vely correlated with distance driven and year.

Location and brand have very less / no relation with price.

Distribution plot of original vs predicted price :



The output graph is nearly having Gaussian distribution.

CONCLUSION

- Key Findings and Conclusions of the Study

  ⇨ Diesel cars are more costly than petrol cars.
  ⇨ Automatic cars are costly than manual gear
  ⇨ Based on different location average price of same vehicle changed
  ⇨ LPG and CNG vehicles are very less available.
  ⇨ Older the car, lesser the price.
  ⇨ More the distance travelled, price of vehicle decreases.
  ⇨ The output price is in terms of lakhs.
  ⇨ The model we created is 78% accurate in predicting the price.

- Limitations of this work and Scope for Future Work

  Due to the unavailability of some details, and more data of some cars w.r.t variants, versions column and car_type column were dropped. For future work we can get the data of those columns also from other cities and verified listings / reliable data helps us to determine the price more accurately