CAR PRICE PREDICTION PROJECT

Submitted by:
ANISH ANTONY

ACKNOWLEDGMENT:

I would like to extend my thanks and appreciation to Datatrained team for their continuous support and guidance during this project, we can also never forget the efforts of all the online tutors that taught me during the Data Analytics program and guided me though the world of data, which was a new realm for me.

Moreover, I would like to extend my gratitude to Flipnwork for his guidance and patience with us during the capstone project.

ABSTRACT:

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model. To predict this model, we need to scrape the used car details from online websites using selenium web driver, and the data have preprocessed, trained and tested using the regression algorithms. Then it is hypertuned and best algorithm with best parameters is obtained and finally the selling price of the car is predicted.

Keywords: Cars ,Selenium ,Data cleaning ,Selling Price ,Regression

CHAPTER I

INTRODUCTION

1.1 Business Problem Framing:

Problem Description:

One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models.

So, they are looking for new machine learning models from new data. We have to make car price valuation model.

They want to understand the factors affecting the pricing of cars in the market, Essentially, the client wants to know:

- Which variables are significant in predicting the price of a car?
- How well those variables describe the price of a car
- Based on used car details, we gathered a large dataset of different types of cars across the different websites.

Business Objectives:
As a Data scientist it is required to apply some data science techniques for the price of cars with the available independent variables. That should help the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

Problem Statement:

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have prepared a car price valuation model. This project contains two phase-

1. Data Collection Phase

In this project we have scrapped close to 5499 used car details of 9 features. The data has been scraped from websites such as Droom, Olx, Cardeko, Carwale, etc. The data has been scrapped from different locations such as Coimbatore, Chennai, Kochi, Hyderabad, Bangalore and Visakhapatnam

The features used in the dataset are:

- Registration Year
- Make
- Model
- KMs Driven
- No of Owners
- Transmission
- Fuel Type
- Mileage
- Location
- Selling Price

2. Model Building Phase:
   After collecting the data, the data is preprocessed. Based on this a machine learning model is tested with different models with different hyper parameters and select the best model.
   Follow the complete life cycle of data science. Include all the steps like.
   - Data Cleaning
   - Exploratory Data Analysis
   - Data Pre-processing
   - Model Building
   - Model Evaluation
   - Selecting the best model

1.2 Conceptual Background of the Domain Problem

Webscraping:

Web scraping is the practice of gathering data through any means other than a program interacting with an API (or, obviously, through a human using a web browser). This is most commonly accomplished by writing an automated program that queries a web server, requests data (usually in the form of HTML

and other files that compose web pages), and then parses that data to extract needed information.

There are several methods of webscraping such as BeautifulSoup, Selenium, Scrappy, etc. Since we need to iterate many individual pages , we use Selenium for webscraping used car details.

Selenium Python bindings provides a simple API to write functional/acceptance tests using Selenium WebDriver. Through Selenium Python API you can access all functionalities of Selenium WebDriver in an intuitive way. We use Chrome WebDriver.

Used Cars:

India's used-car market is booming and startups are capitalising like never before. Riding the digital wave, India's used car market is set to grow at a compounded annual growth of 11% and likely to touch sales of up to 8.3 million units by FY26 as more people have been opting for pre-owned cars for personal mobility in the amid the ongoing supply shortages for manufacturing new cars. This growth is driven by increased sales of in metro cities and a rise in online sales platforms, such as CarDekho, Cars24, and Droom.

Machine Learning:

A machine learning model is a file that is trained to identify multiple relationships in a dataset. Usually, we train a model using a machine learning algorithm and use it for further predictions. Many powerful machine learning models are trained and made available for our use in the form of libraries

The price of a car depends on a lot of factors like the goodwill of the brand of the car, features of the car, horsepower and the mileage it gives and many more. Car price prediction is one of the major research areas in machine learning. We can train a machine learning model for the task of predicting car prices by using the Python programming language. It is a major research topic in machine learning because the price of a car depends on many factors.

1.3 Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

Many studies and related articles have been done previously to predict used car prices using different methodologies and approaches, with varying results of accuracy from 50% to 90%. In (Pudaruth, 2014) the researcher proposed to

predict used car prices in Mauritius, where he applied different machine learning techniques to predict his results with algorithms such as decision tree, K-nearest neighbours, Multiple Regression and Naïve Bayes algorithms to predict the used cars prices, based on historical data gathered from the newspaper.

The achieved results ranged from accuracy of 60-70 percent; the author suggested that using more sophisticated models. The main weakness of the decision tree and naïve Bayes is that it is required to discretize the price and classify it but, it led to more inaccuracies. Moreover, he suggested a larger set of data of data to train the models hence the data gathered was not sufficient.

(Monburinon, et al., 2018) collected data from a German e-commerce site that totalled to 304,133 rows and 11 attributes to predict the prices of used car using different techniques and measured their results using Mean Absolute Error (MEA) to compare their results. He tested the dataset using different models and hyper parameters. Highest results achieved was by using gradient boosted regression tree with a MAE of 0.28, and MEA of 0.35 and 0.55 for mean absolute error and multiple linear regression respectively. Authors suggested to adjust the parameters in future to yield better results, as well as using one hot encoding instead of label encoding for more realistic data interpretations on categorical data.

1.4 Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

Used car price prediction is very important for both the customer and the seller. The price can be predicted based on the features such as model, brand, no of years, mileage, performance, etc. Based on data obtained various websites, the aim is to use machine learning algorithms to develop models for predicting used car prices.

CHAPTER II

Analytical Problem Framing

2.1 Mathematical/ Analytical Modeling of the Problem:

Machine Learning:

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists chooses to use depends on what type of data they want to predict.

Supervised learning:

In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Supervised learning can be separated into two types of problems when data mining—classification and regression

- Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, etc.
- Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

Unsupervised learning:

This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

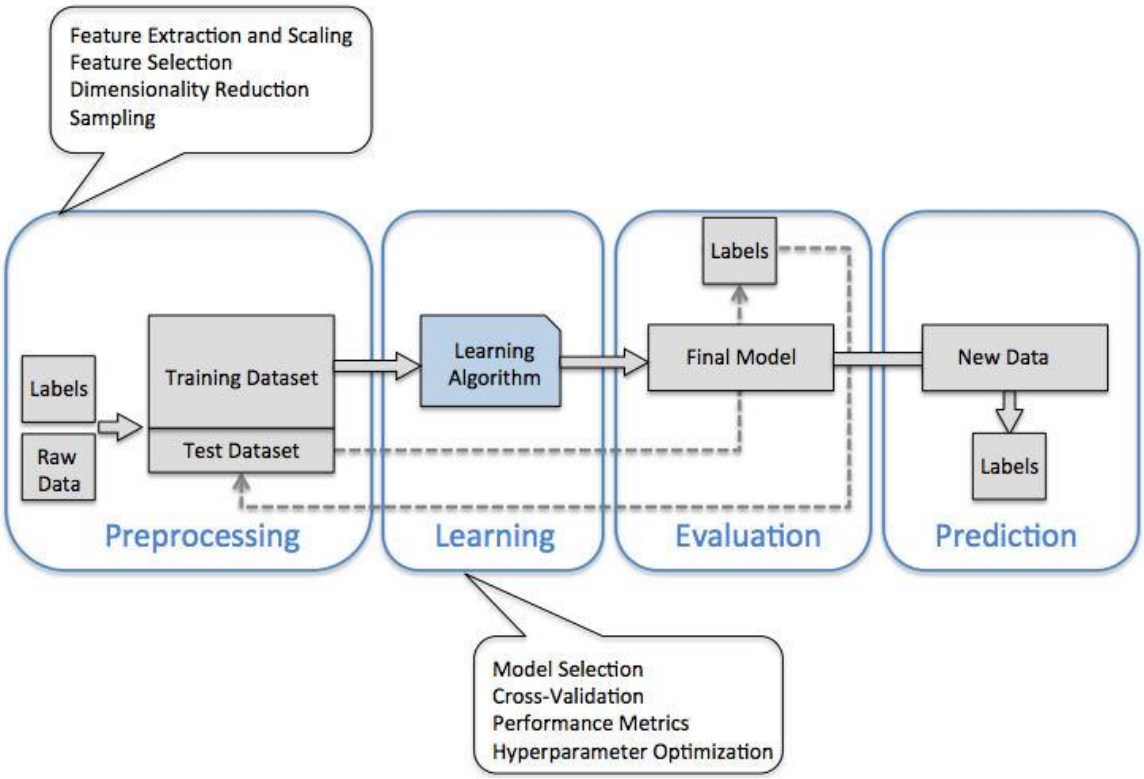
Popular un-supervised algorithms are K-means clustering, affinity propagation etc.

Semi-supervised learning:

This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

Reinforcement learning:

Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.



Linear regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

Random forest Algorithm:

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type i.e., multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

Statistical Analysis:

Statistical analysis, or statistics, involves collecting, organizing and analysing data based on established principles to identify patterns and trends.

Predictive analysis:

Predictive analysis uses powerful statistical algorithms and machine learning tools to predict future events and behaviour based on new and historical data trends. It is important to note that predictive analysis can only make hypothetical forecasts and the quality of the predictions depends on the accuracy of the underlying data sets.

The terms used for statistical analysis are:

Mean	$\bar{x} = \frac{\sum x}{n}$	x = Observations given n = Total number of observations

Median	If n is odd, then
	$M = \frac{n+1}{2th}$ term
	If n is even, then
	$M = \frac{\left(\frac{n}{2}\right)th\ term + \left(\frac{n}{2}+1\right)th\ term}{2}$
		n = Total number of observations

Mode	The value which occurs most frequently	
Variance	$= \sigma^2 = \sum \frac{(x-x)^2}{n}$	x = Observations given = Mean n = Total number of observations

Standard Deviation	$S = \sigma = \sqrt{\sum \frac{(x-x)^2}{n}}$	x = Observations given $\bar{x}$ = Mean n = Total number of observations

$$\text{Z score} = \frac{x-\bar{x}}{\sigma}$$

Where,

x = Standardized random variable

$\bar{x}$ = Mean

σ = Standard deviation.

Quartile Formula:

When the set of observation is arranged in an ascending order, then the 25th percentile is given as:

$$Q_1 = (\frac{n+1}{4})^{th} \text{ term}$$

The second quartile or the 50th percentile or the Median is given as:

$$Q_2 = (\frac{n+1}{2})^{th} \text{ term}$$

The third Quartile of the 75th Percentile (Q3) is given as:

$$Q_2 = (\frac{3(n+1)}{4})^{th} \text{ term}$$

$$\text{IQR} = \text{Upper Quartile} - \text{Lower Quartile}$$

Regression:

Regression is a statistical technique used to find a relationship between a dependent variable and an independent variable. It helps track how changes in one variable affect changes in another or the effect of one on the other. Regression can show whether the relationship between two variables is weak, strong or varies over a time interval. The regression formula is:

$$Y = a + b(x)$$

Y represents the independent variable, or the data used to predict the dependent variable

x represents the dependent variable which is the variable you want to measure

a represents the y-intercept or the value of y when x equals zero

b represents the slope of the regression graph

Hypothesis testing:

Hypothesis testing is used to test if a conclusion is valid for a specific data set by comparing the data against a certain assumption. The result of the test can nullify the hypothesis, where it is called the null hypothesis or hypothesis 0. Anything that violates the null hypothesis is called the first hypothesis or hypothesis 1.

2.2 Data Sources and their formats:

1. Data Collection Phase

In this project we have scrapped close to 5499 used car details of 10 features. Initially, we surfed the internet for used cars. We searched through the shopping websites. The data has been scraped from websites such as Droom, Cardeko, Carwale, etc. The data has been scrapped from different locations such as Coimbatore, Chennai, Kochi, Hyderabad, Bangalore and Visakhapatnam.

Scrape the necessary details and transform it into a dataframe using python in the jupyter notebook. Combine all individual dataframes and merge into a single dataframe

2. Data Cleaning:

We need to need to clean the dataset since the data is collected from different websites.

For use of access, we extract all the 10 features into 3 columns, because since webscraping for 5499 data takes much more time. These three columns are then split into 15 features that’s the way the data is available

Name column contains the Model, Brand Year and name of the vehicle and it is split into the respective features.

The features used in the dataset are:

- Registration Year – (Converted to object)
- Make – (Extracted from Name)
- Model – (Extracted from Name)
- KMs Driven – (Removed the string and converted to integer)
- No of Owners - (Made as unique variables for all websites)
- Transmission – If not available in some websites, then give the data as not available
- Fuel Type - (Made as unique variables for all websites)
- Registration – If not available in some websites, then give the data as not available
- Location – (Based on the location give the location)
- Selling Price – (Removed the string and converted to integer)

The categorical features were standardized and made uniform for all data

Dataset Description:

S.NO	Columns	Datatype	Unique values	Mode/Mean
1	Registration Year	int64	28	2013.56283
2	Make	object	29	Maruti
3	Model	object	1155	Maruti Suzuki Swift VXi
4	KMs Driven	int64	629	77207.429351
5	No of Owners	object	4	First
6	Transmission	object	3	Manual
7	Fuel Type	object	9	Petrol
8	Location	object	6	Coimbatore

9	Selling Price	int32	805	607882.439353

Some of the features were redundant, when comparing different websites, the features were fixed based on the requirement and importance and others were removed.

2.3 Data Preprocessing Done:

Exploratory Data Analysis (EDA):

Some of the features have 'Not available' as feature values. It has to replace with mode or mean based on the dataset datatype.

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

Assumptions for data collections:

Since the data extracted needed to cover most major features, different locations and websites. The data is standardized with respect to Droom websites since it most features and different location and other websites data has been benchmarked with these websites. Due to this some of the data has been removed. This data is taken from cities of southern states and prices will be dependent on the location.

2.5 Hardware and Software Requirements and Tools Used:

Hardware – PC Windows 10, 4 GB Ram

Software – Google chrome, MS Excel, Python, Selenium webdriver

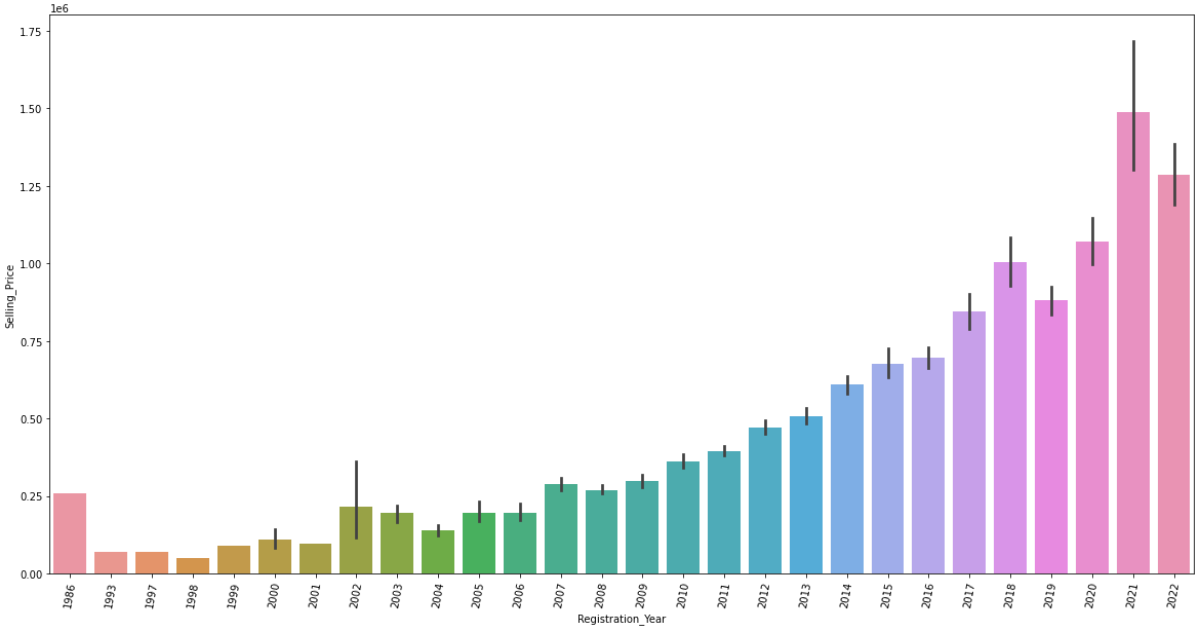
Libraries – Pandas, NumPy, Matplotlib, Seaborn, sklearn, SciPy. Stats

- Browsing – Google Chrome
- Webscraping – Python, Selenium webdriver
- Data cleaning – Python, Pandas, NumPy & SciPy. Stats
- Data visualization – Matplotlib & Seaborn
- Machine learning – Sklearn

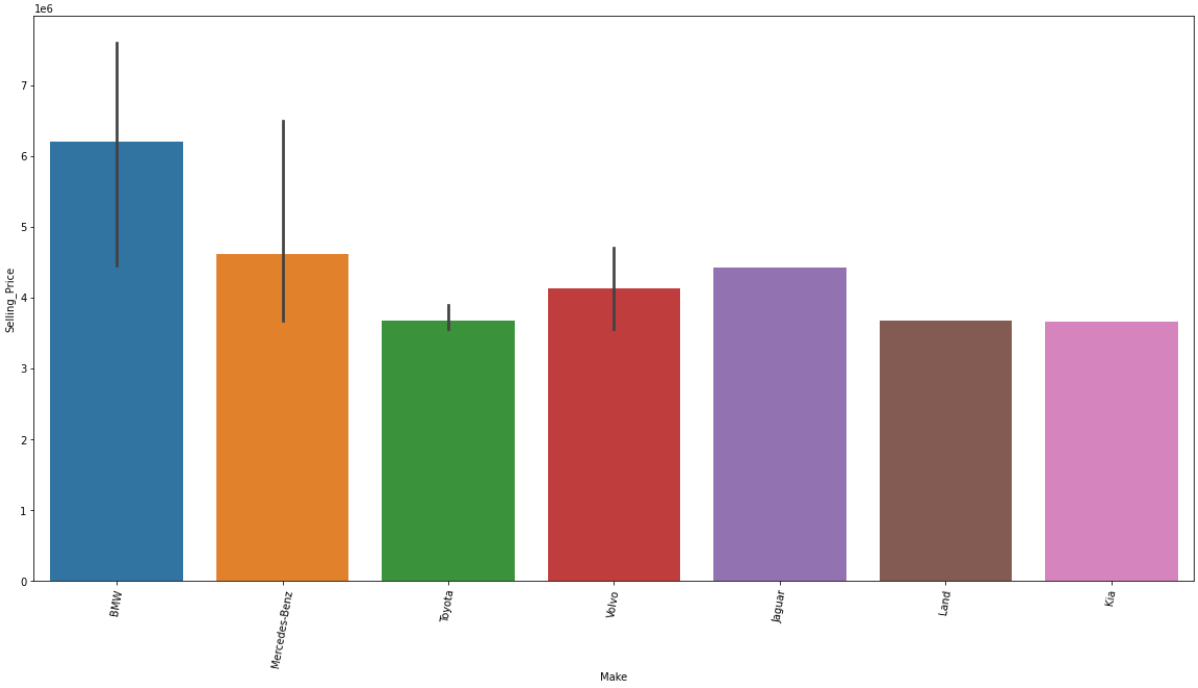
2.6 Data Inputs- Logic- Output Relationships:

1. Registration Year:

In general, the selling price decreases as the car ages,



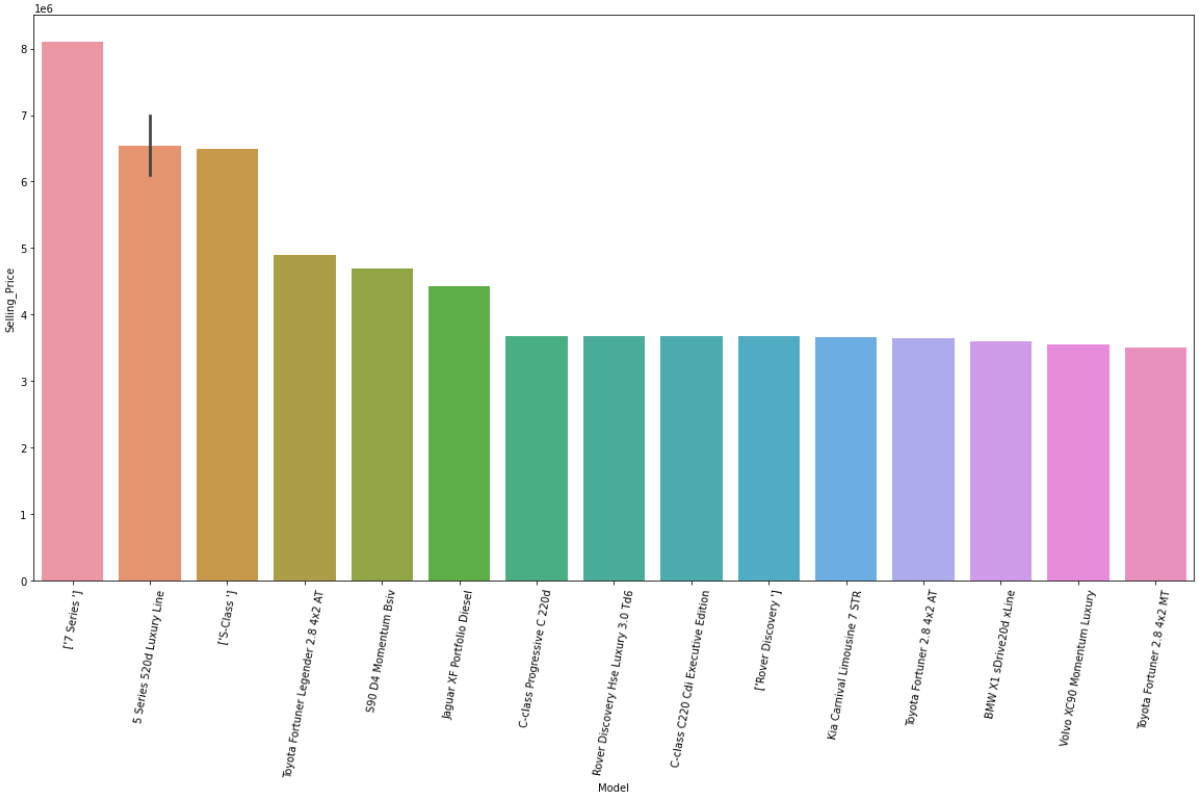
3. Make:



From the graph we find that BMW, Mercendez Benz are the costliest brands

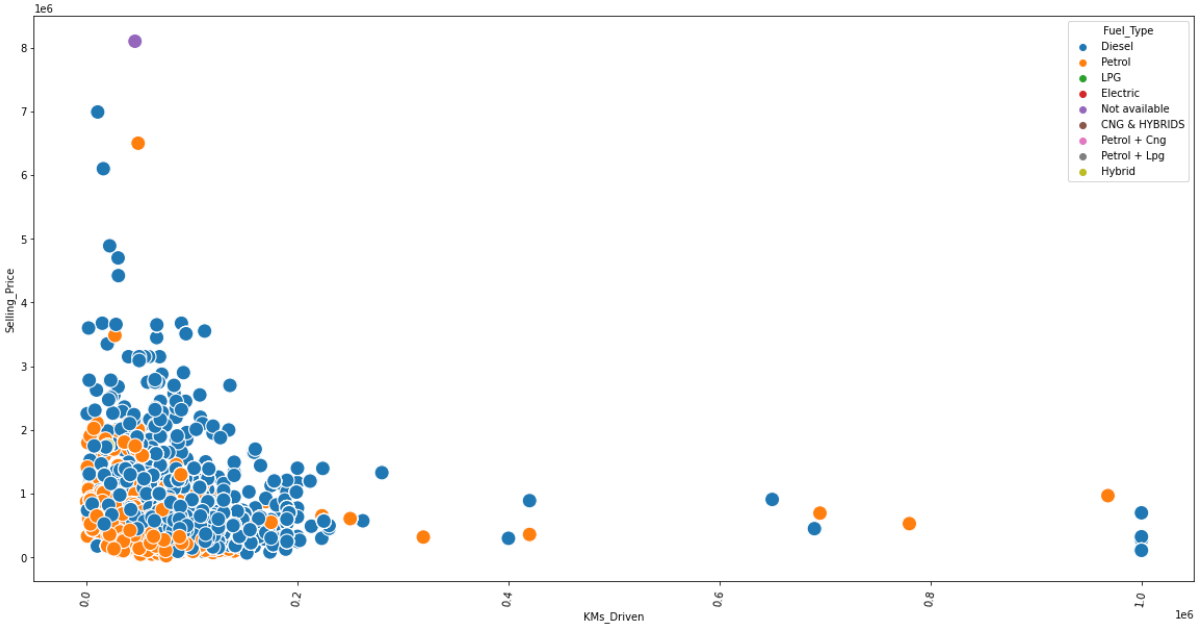
4. Model:

From the graph we find that 7 Series, 5 Series, S-class, Fortuner are the costliest model



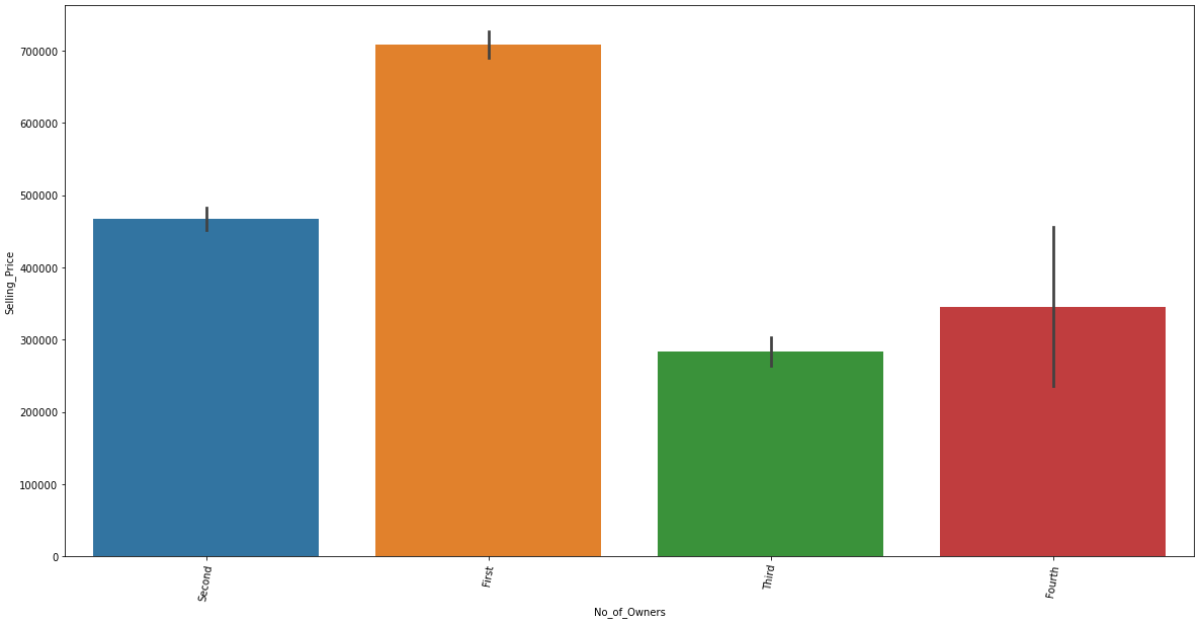
5. KM's driven:

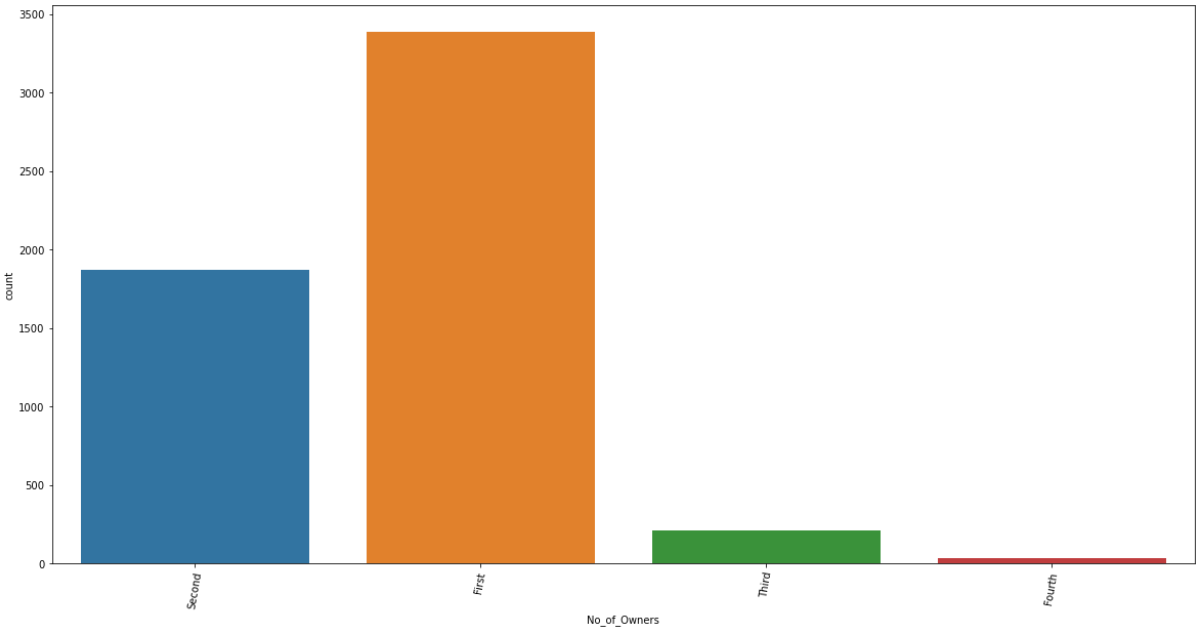
From the graph we find that most cars are either petrol or diesel and are within 20000 km's .Cars which are driven less are more costly.



6.No of owners:

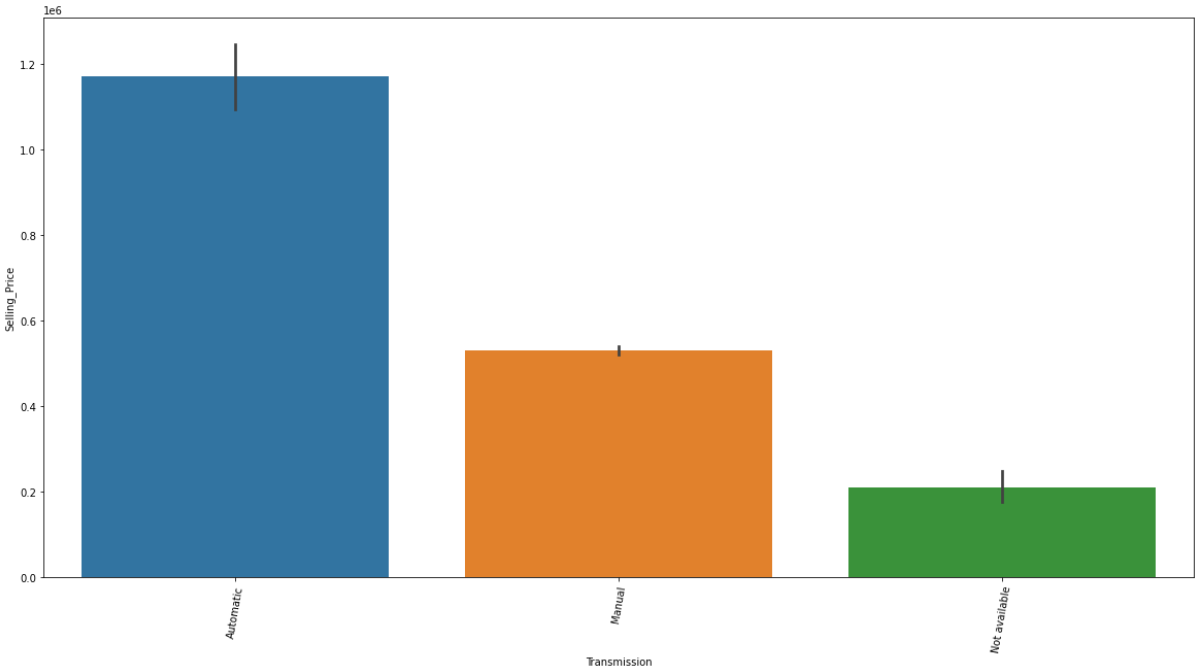
From the graph, we find that as no of owners increase, the decreases and the price decreases. Moreover first owners are more in number.

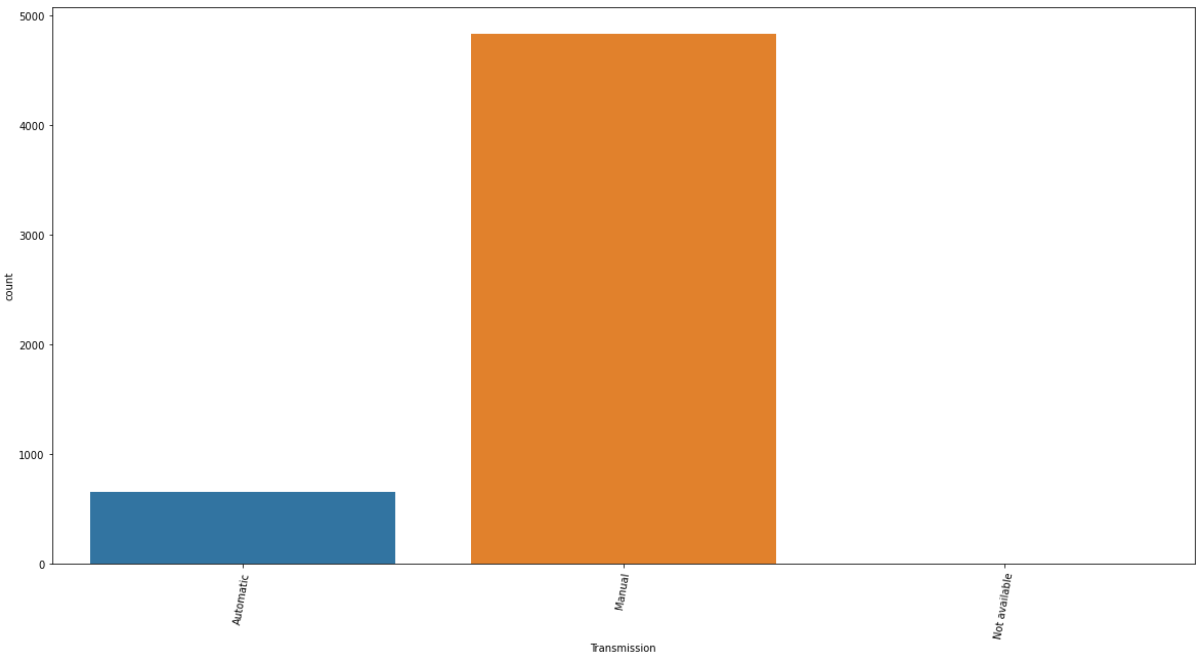




7. Transmission:

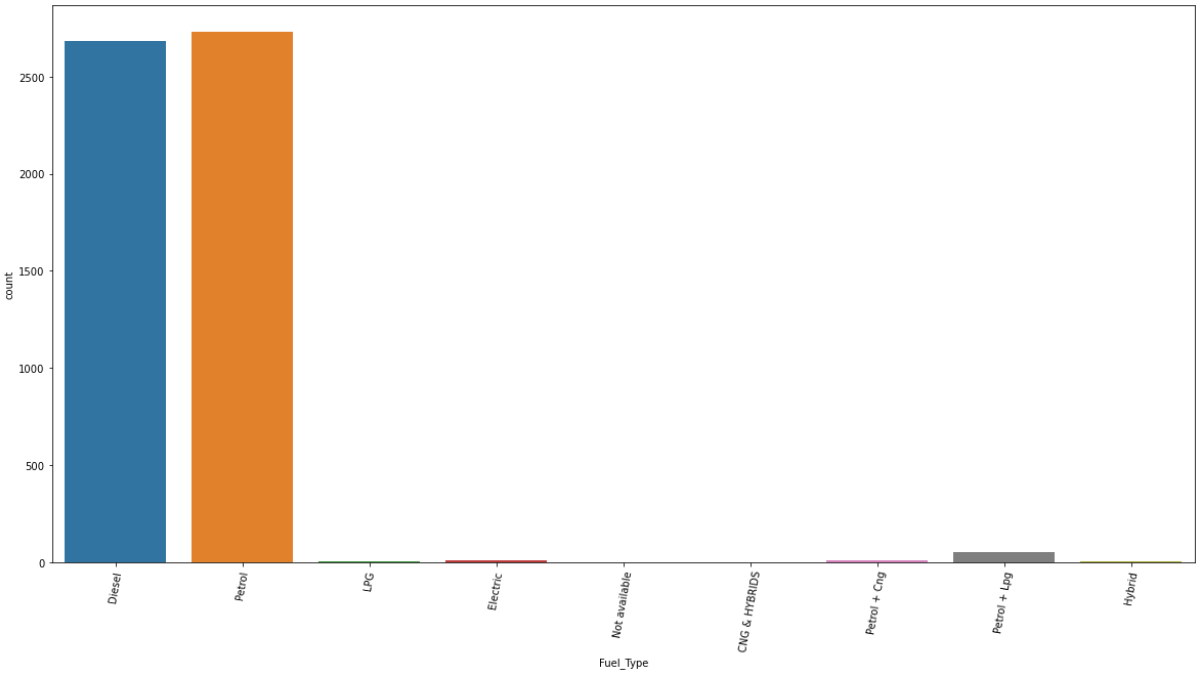
From the plots we find that automatic transmission is costly and at the same time Manual transmission is mostly available.





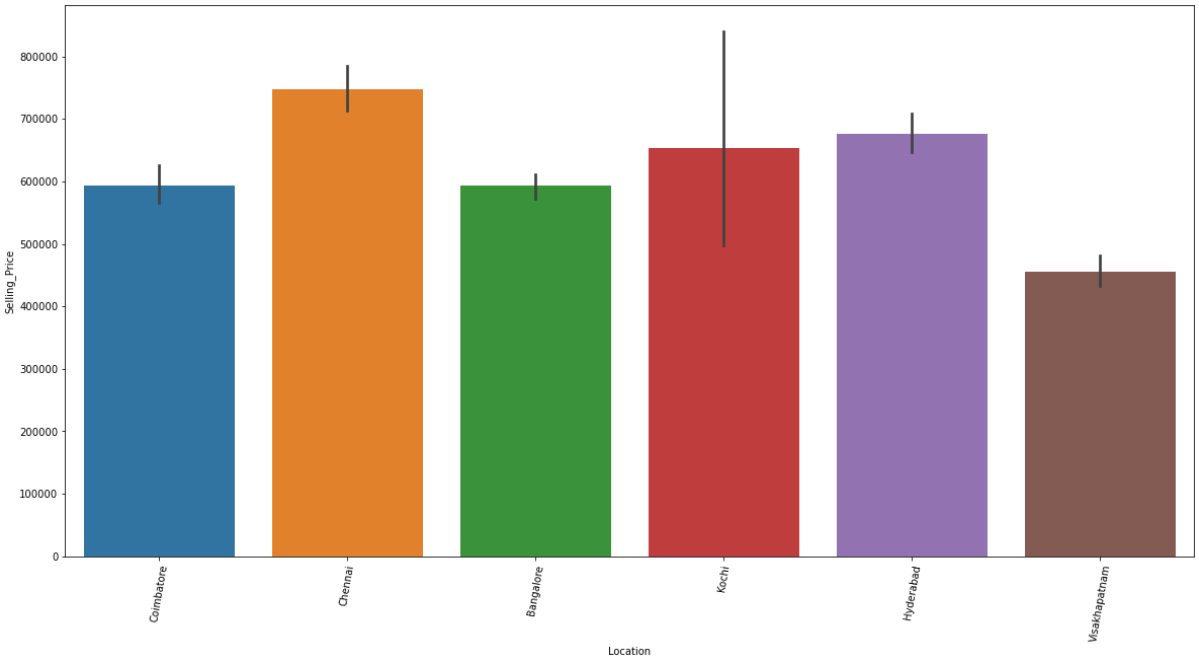
8. Fuel Type:

From the plots we find that diesel cars are costly and at the same time, when compared to other fuel types, only petrol and diesel are mostly available.

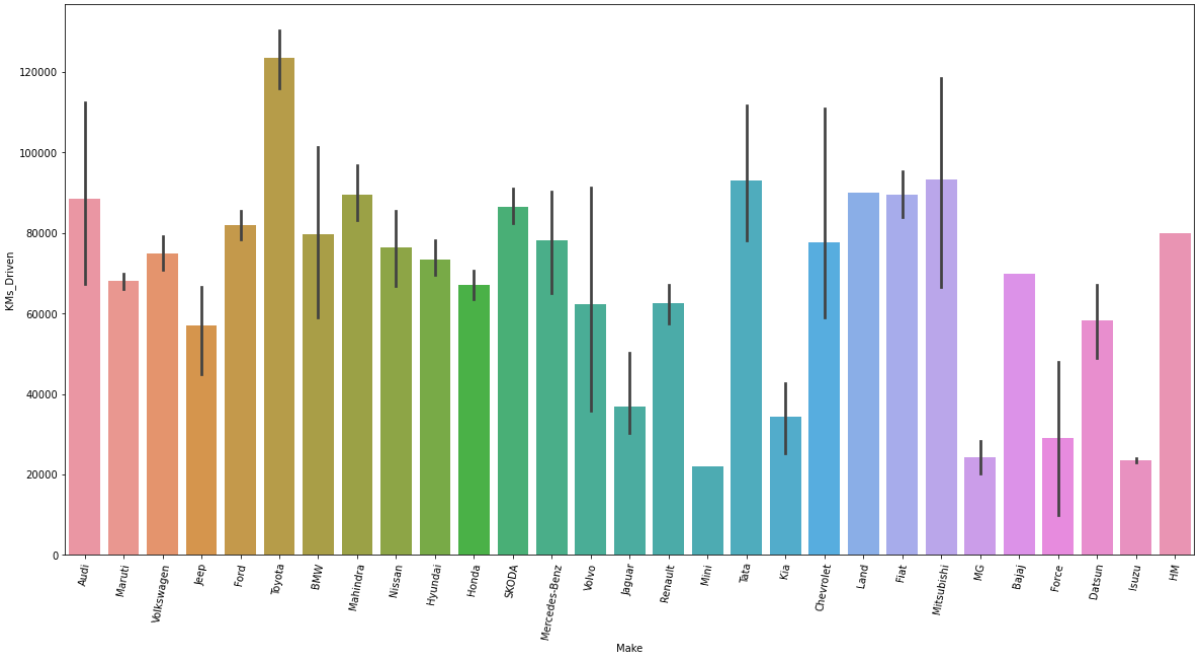


9. Location:

From the location plot, we find that Chennai & Hyderabad cars are costly compared to other states .

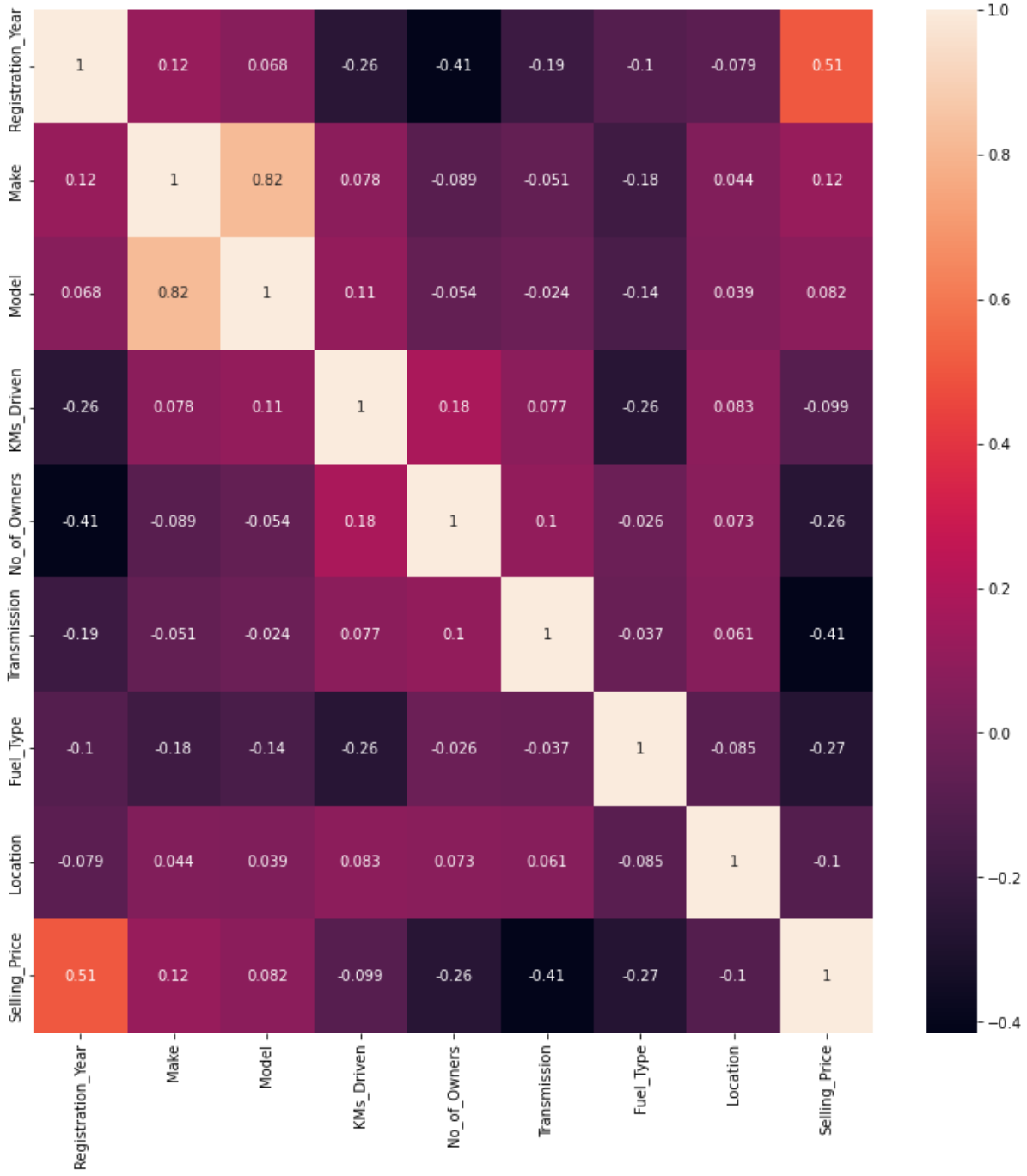


KM’s Driven vs Make



Once preprocessing is completed, the categorical data is converted to continuous data using label encoding.

Now the dataframe is checked for correlation and heatmap is shown below:



Now the dataset is checked for presence of outliers. Here we use two methods

1. IQR Method

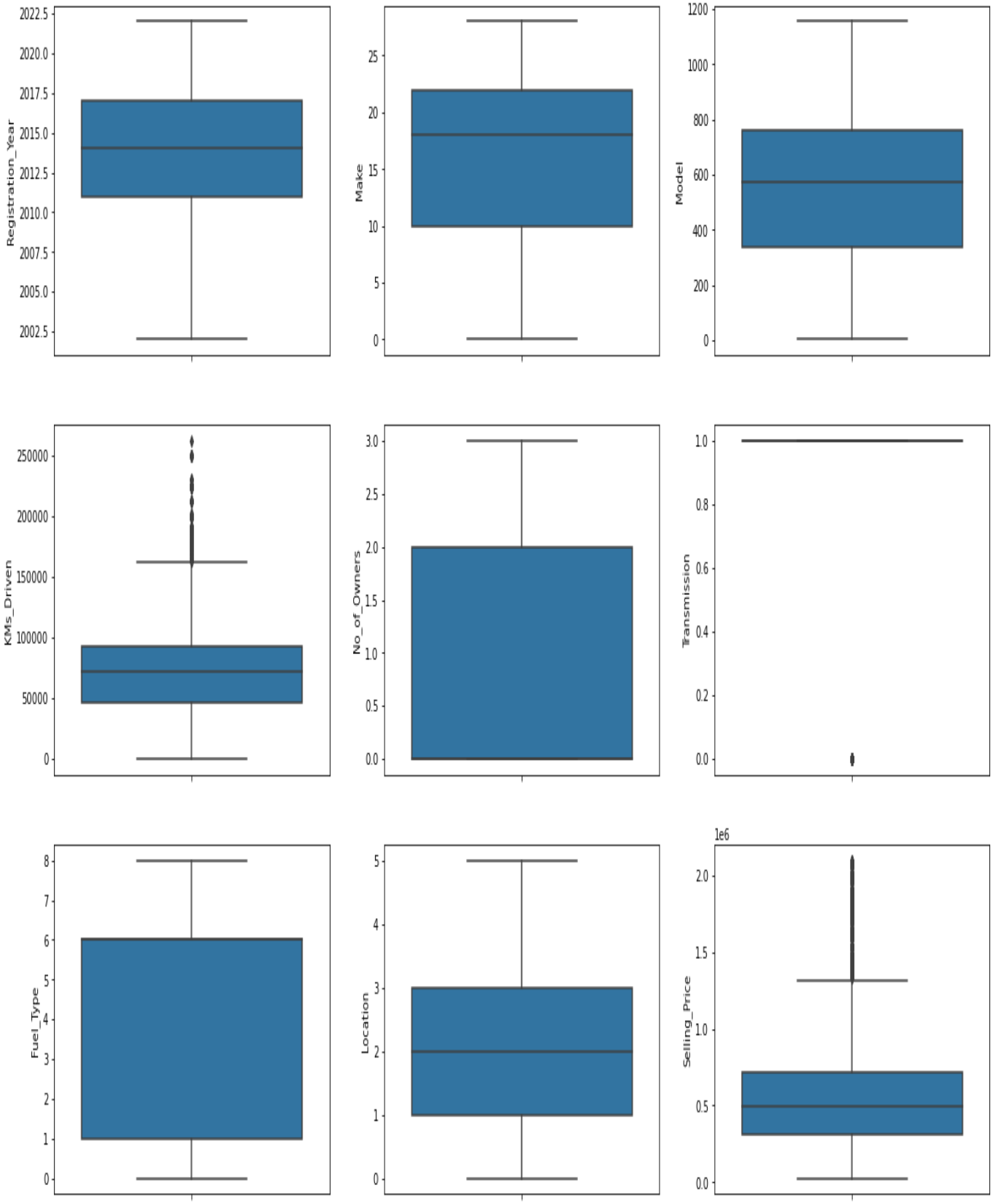
From this method we couldn't find any outliers

2. Z-Score Method

Using Z-Score method we detected some outliers and new dataframe is used further.

The new dataframe has the shape of 5341 rows and 9 columns

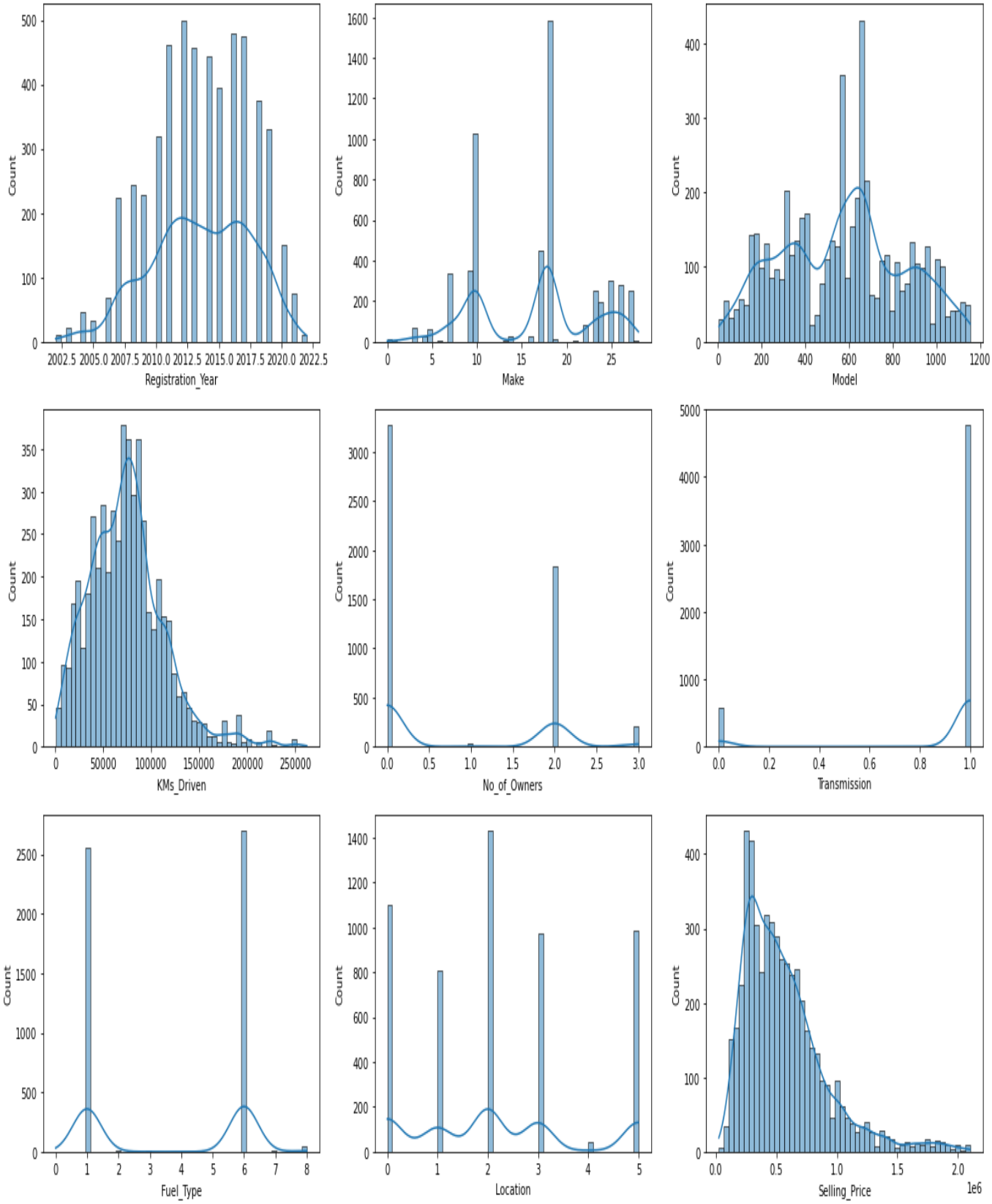
After removing outliers, we plotted the box-plot



Skewness:

Histogram:

A histogram is plot to check whether the features are normally distributed or not.



CHAPTER III

Model/s Development and Evaluation

3.1 Identification of possible problem-solving approaches (methods)

Basic Parameters:

1. Standardization:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

2. Train Test data:

Train/Test is a method to measure the accuracy of your model. It is called Train/Test because we split the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing. We train the model using the training set. We test the model using the testing set.

3. Linear regression

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.

5. Random Forest Regressor

Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees. It executes by constructing a different number of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

3.2 Testing of Identified Approaches (Algorithms):

Listing down all the algorithms used for the training and testing.

- Linear Regression
- Gradient Boosting Regressor
- AdaBoost Regressor
- Decision Tree Regressor
- KNeighbors Regressor
- Extra Trees Regressor
- Random Forest Regressor

3.3 Run and evaluate selected models:

From the above, the model is scaled using standard scaler, looped with the above methods and best model is obtained.

From this the best model is Random Forest Regressor from the Random state 67. Now this model is hypertuned and best parameters is obtained

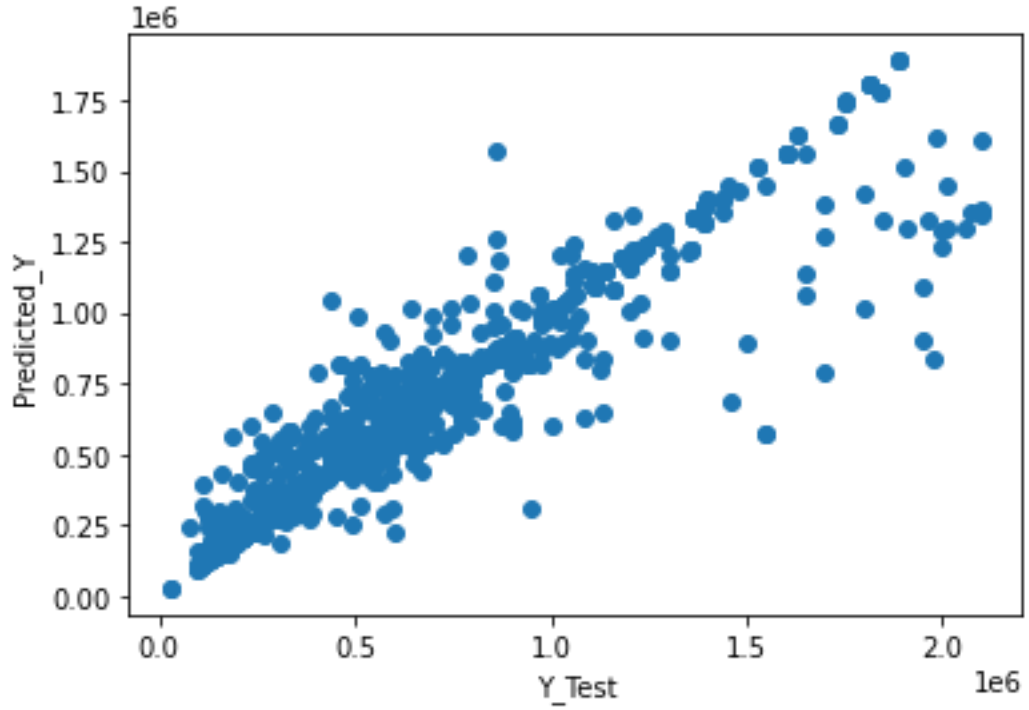
3.4 Key Metrics for success in solving problem under consideration:

Accuracy Parameter:

- R2 Score: 88.5718027742514
- Mean Absolute Error: 43667.17472188806
- Mean squared Error: 14141070734.25445
- Root Mean Absolute Error: 208.96692255447527

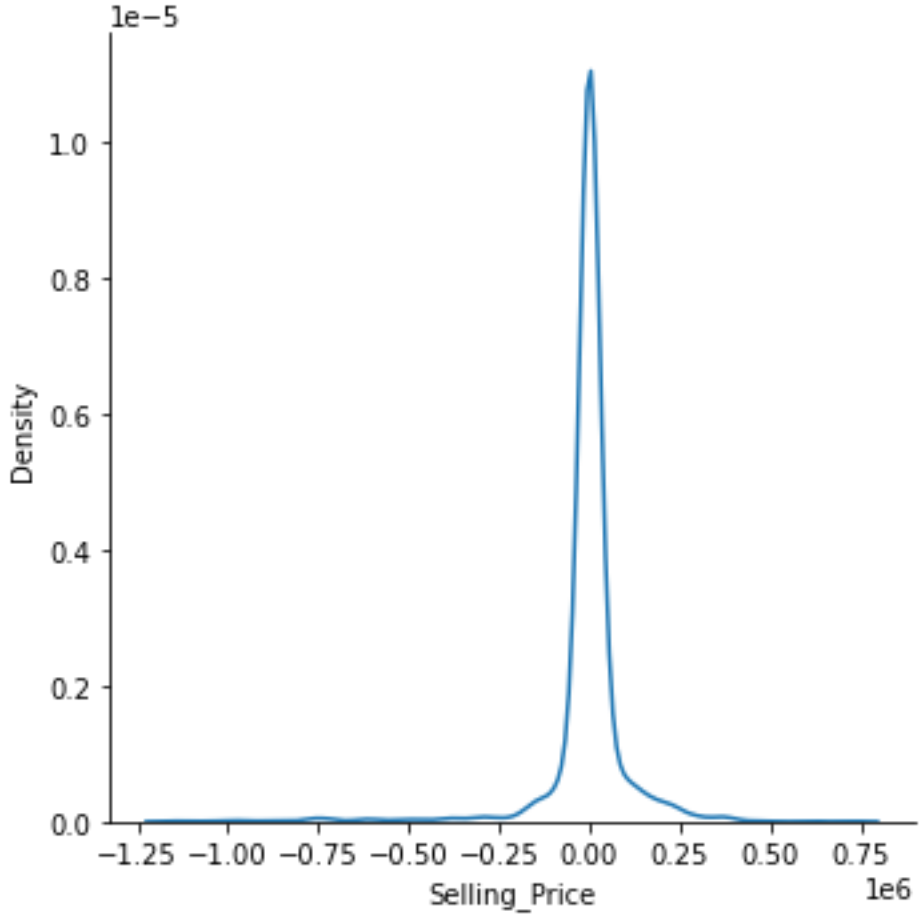
The best model is obtained by Hypertuning the existing models.

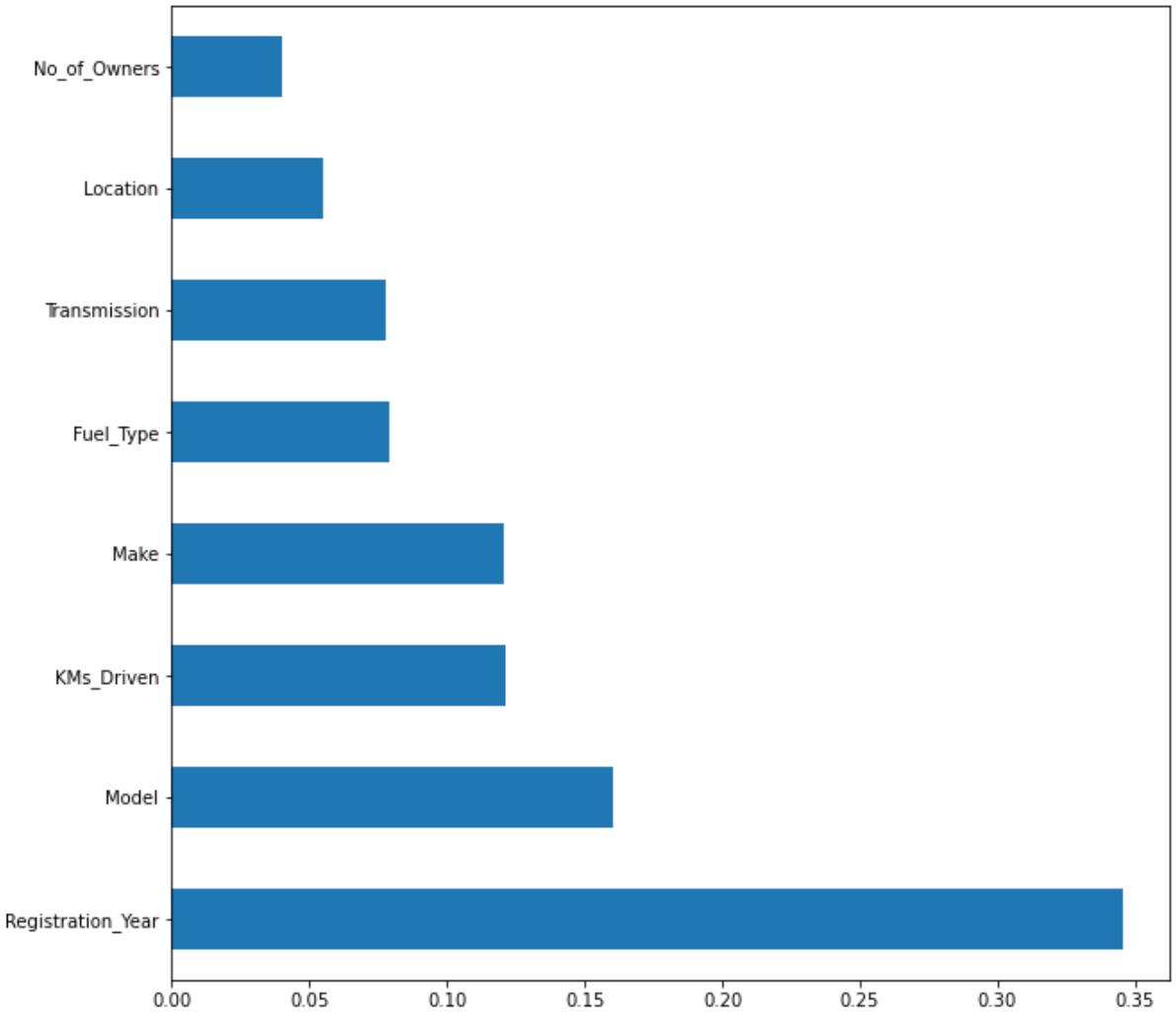
3.5 Visualizations:



We get almost a linear plot

Distribution of Predicted Values:





Regarding the feature importance's, registration year dominates the selling price more.

3.6 Interpretation of the Results

- From the results, we find that this problem can be solved by a regression method and selling price can be predicted.
- The skewness is not reduced as most of the data is categorical
- Registration year dominates the selling price more.

CHAPTER IV
CONCLUSION

4.1 Key Findings and Conclusions of the Study:

- This dataset has been taken from 3 websites of this, Droom website constitutes the majority of data
- Since the target feature is continuous data, this problem can be solved by regression algorithms
- Random Forest Regression gives an R2 score 0.88
- Registration year dominates the selling price more.

4.2 Learning Outcomes of the Study in respect of Data Science:

- It gives a deep learning of Selenium webscraping
- It emphasizes the importance of data cleaning
- Data uniformity and the importance of features required
- How data collection affects the results
- Role of data cleaning, feature selection, etc.

4.3 Limitations of this work and Scope for Future Work:

- Initially 4 websites have been web scrapped, but due to the non-uniformity of data, and the features, some of the data have been removed
- Through web scrapping, keyboard interrupt has been used, to stop the iteration.
- Skewness have not been reduced

In feature, data will be gathered from more websites and more algorithms will be used along with different methods of scaling.

In future, more features will be added in predicting the selling price.