A project report

On

"Used Car Price Prediction Using Multiple Linear Regression"

By

Amarendra Pratap Deo
MSc(AI & ML)

Table of Contents

Abstract.....2

Introduction.....3

Literature Survey.....5

Problem Statement.....9

Methodology.....10

Data Set.....11

Implementation Details.....13

Results.....31

Conclusion.....34

Suggestions for Further Study.....34

Bibliography.....36

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Abstract

In this project we have developed a model for predicting the selling price of used cars in India. We have implemented Multiple Linear Regression under supervised machine learning. The dataset is retrieved from Kaggle and applied for Cardekho which is India's leading online car marketplace that helps users buy and sell cars. We have developed the model using Scikit Learn Library in Python. The model developed in the project has an average percentage error of 16% and an r squared value of 0.84

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Introduction

The purpose of this research is to analyse the used car market in India and build an effective model for predicting the car prices of used cars over the past years, and verify whether the regression analysis methods and models can effectively predict the prices of used cars on listed on online aggregator website like Cardekho [1].

The used car market in India is growing at a phenomenal rate of 15% (CAGR). In 2020 used car sales was estimated to be 4.4 million whereas new car sales were only around 2.7 million. The rise in the demand for used cars has led to the emergence of online aggregators like Cardekho. The aim is to develop a model for customers, so that they can get an idea of the fair price for their used car before putting up the advertisement. The rise in the demand for used cars. We wanted to develop a model for customers, so that they can get an idea of the fair price for their used car before putting up the advertisement.

CarDekho operates as a platform for business-to-business (B2B) and business-to-consumer (B2C) operations. CarDekho has footprints in India, Indonesia and the Philippines. It’s headquartered in Gurugram India. It’s website and app carry rich automotive content such as expert reviews, detailed specs and prices,

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

comparison as well as videos and pictures of all car brands and models available in India. The company has tie-ups with many automotive manufacturers, more than 4000 car dealers and numerous financial institutions to facilitate the purchase of vehicles[1]. These include app for dealer sales executives to manage leads, cloud services for tracking sale performance, call tracker solutions, digital marketing support, virtual online showroom, and outsourced lead management operational process for taking consumers from enquiry to sale. Their vision is to construct a complete ecosystem for consumers and car manufacturers, dealers and related businesses such that consumers have easy and complete access to not only buying and selling cars, but also manage their entire ownership experience, be it accessories, tires, batteries, insurance or roadside assistance.

Pre-Owned Car market :

The pre-owned vehicle segment which accounts for 18 percent of the market share in India, in the financial year 2019-20 it registered an estimated sales of 44 lakhs units. The used car market in India is growing at a phenomenal rate of 15% (CAGR). In 2020 used car sales was estimated to be 4.4 million whereas new car sales were only around 2.7 million. Currently the new-car to used-car ration is around 1:2, meaning for every new car that is sold, there are two old

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

cars that are sold, auto consultants postulate that post COVID the ration will be more towards the 1:3 mark [2]

Literature Survey

In the research paper Introduction to Multiple Regression: How Much Is Your Car Worth Shonda Kuiper postulates that multivariate regression model can be effective in classifying and predicting values of numeric type, Kuiper used Multiple Linear regression to model to predict price of 2005 General Motor (GM) cars. Kuiper made car price prediction and introduced variable selection techniques which helped in finding which variables can be included to improve the model accuracy. The research was carried out using Minitab software[3].

In the research paper Support Vector Regression Analysis for Price Prediction in a Car Leasing Application Mariana Listiani uses Support Vector Machines (SVM) to predict the prices of leased cars in Germany. This research showed that SVM is far more accurate in predicting prices as compared to the multiple linear regression in that SVM has a lower root mean squared error as compared to multiple linear regression when a very large dataset is available. SVM also handles high dimensional data better and avoids both the under-fitting and over-fitting issues which is prevalent in regression[4].

Regression analysis:

Regression analysis is a technique used in statistics for investigating and modelling the relationship between variables[5] .

Simple linear regression:

Simple linear regression is a model with a single predictor or regressor variable x that has a relationship with a response variable y that is a straight line[5]. It is a linear approximation of a causal relationship between a single variable x with a predictor variable y. This simple linear regression model can be expressed as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where the intercept β0 and the slope β1 are unknown constants and ε is a random error component, which is the difference between the observed value of *y* and the straight line $y = \beta_0 + \beta_1 x$ (Douglas Montgomery, Peck, & Vinning, 2012)[5] .

Multiple linear regression:

If there is more than one regressor, it is called **multiple linear regression.** In general, the response variable y may be related to k regressors, $x_1, x_2, \ldots, x_k$, so that

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

The parameters $\beta j \;\; j = 0, 1, \ldots, k,$ are called the **regression coefficients**. This model describes a hyper- plane in the *k*-dimensional space of the regressor variables x_j. The parameter β_j represents the expected change in the response *y* per unit change in x_j **when all of the remaining regressor variables are held constant**[3]. (Douglas Montgomery, Peck, & Vinning, 2012).

Least Squares Estimation:

The objective of regression analysis is to **estimate the unknown parameters** in the regression model. This process is also called fitting the model to the data. One of the method used is Least Square estimation, the method of least squares is used to estimate $\beta_0, \beta_1, \ldots \beta_k$. That is, we

estimate β_0 and β_1 so that the sum of the squares of the differences between the observations $\hat{y}_i$ and the straight line is a minimum[5]

R-squared:

R-squared is a measure in statistics of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for multiple regressions[6]. It is the percentage of the response variable variation that is explained by a linear model.

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

$$R - Squared = \frac{\text{Explained variation}}{\text{Total variation}}$$

R-squared is always between 0 and 1. 0 means the model explains none of the variability of the response data around its mean. 1 indicates that the model explains all the variability of the response data around its mean. Generally, the higher the R-squared, the better the model fits the data[6].

Scikit Learn Library:

Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data pre-processing, model selection and evaluation, and many other utilities[7].

Mean Absolute Error

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

The mean absolute error is a metric used to evaluate regression models, it is calculated using the above formula, wherein y_i is the predicted value, x_i is the true value and n is the total number of data points[8].

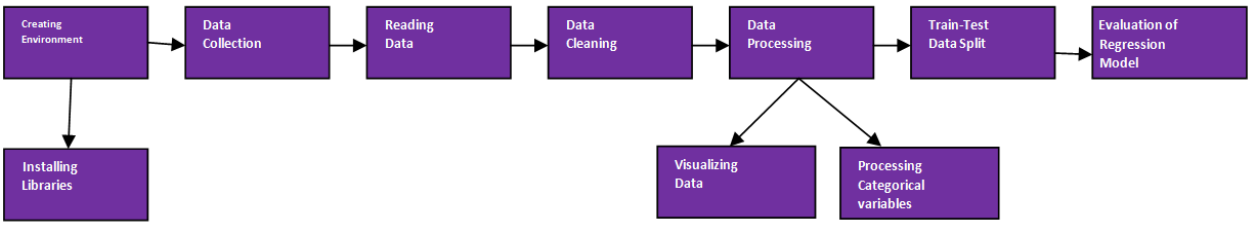
Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Problem Statement

The consumers who are planning to sell their used car online need an indication of the fair price or the ideal price of their car based on its parameters like brand, kilometres driven, transmission type, number of seats, engine power, etc. Thus providing an ideal price based on past data will allow the customer get a better understanding of the value of their car. The used car price prediction model is a model which is used to predict the price of a used car by using Multiple linear regression. This model will help customers to predict the prices of used cars and get the best value of their asset.

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Methodology



1. Data Collection (Kaggle)
    - Data Set preparation
2. Data pre-processing (Data Cleaning, outlier detection and removal)
3. Exploratory Data Analysis
4. Data Visualization
5. Outlier Detection
6. By using IQR (Interquartile Range)
7. Data Processing
    - Converting Categorical Variables to Numerical One Hot Encoding
8. Train Test Split (70:30)
9. Apply Multiple Linear Regression
10. Model Evaluation

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Data Set

The dataset contains the following 14 columns:-

Id, Car Name, Brand, Model, Vehicle age, Kilometres driven, Seller type: Dealer or Individual, Fuel Type: Petrol, Diesel, CNG (Compressed Natural Gas) or Electric, Transmission type: Automatic or Manual, Mileage, Engine, Max Power, Seats, Selling Price, Current Year and Purchased Year.

There are in total 19542 rows

```
data.head(10)
```

	id	car_name	brand	model	vehicle_age	km_driven	seller_type	fuel_type	transmission_type	mileage	engine	max_power	seats	selling_price
0	0	Maruti Alto	Maruti	Alto	9	120000	Individual	Petrol	Manual	19.70	796	46.30	5	120000
1	1	Hyundai Grand	Hyundai	Grand	5	20000	Individual	Petrol	Manual	18.90	1197	82.00	5	550000
2	2	Hyundai i20	Hyundai	i20	11	60000	Individual	Petrol	Manual	17.00	1197	80.00	5	215000
3	3	Maruti Alto	Maruti	Alto	9	37000	Individual	Petrol	Manual	20.92	998	67.10	5	226000
4	4	Ford Ecosport	Ford	Ecosport	6	30000	Dealer	Diesel	Manual	22.77	1498	98.59	5	570000
5	5	Maruti Wagon R	Maruti	Wagon R	8	35000	Individual	Petrol	Manual	18.90	998	67.10	5	350000
6	6	Hyundai i10	Hyundai	i10	8	40000	Dealer	Petrol	Manual	20.36	1197	78.90	5	315000
7	7	Maruti Wagon R	Maruti	Wagon R	3	17512	Dealer	Petrol	Manual	20.51	998	67.04	5	410000
8	8	Hyundai Venue	Hyundai	Venue	2	20000	Individual	Petrol	Automatic	18.15	998	118.35	5	1050000
9	9	Mahindra TUV	Mahindra	TUV	4	70000	Dealer	Diesel	Manual	18.49	1493	100.00	7	575000

1. Id - It represents a unique Identification of Cars.
2. Car Name - It represents cars name(e.g. 'Toyota Innova', 'Maruti Baleno’)
3. Brand:- It represents Cars brand(e.g. 'Toyota', 'Maruti’, ‘BMW’, etc)
4. Model:- It is used to identify and describe the cars(e.g. Hyundai -> ‘i20’)
5. Vehicle age:- It means the numerical difference between the current calendar year and the vehicle model year. (e.g. 5 Years)

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

6. KM driven:- It is a unit of length, the common measure of distances(e.g. ‘12000 KM’)
7. Seller type:- It contains the information about seller (e.g. ‘Dealer’)
8. Fuel type:- Its uniquely identifies the type of fuel(e.g. ‘Petrol’, ‘Electric’)
9. Transmission type:- It consists of ‘Automatic’ or ‘Manual’ .
10. Mileage:- It indicates the distance that a vehicle can travel with a specific amount of fuel(e.g. 20 km/h)
11. Engine:- The engine which measures the volume in CC (Cubic Centimetres). Thus CC is a unit of volume(e.g. 1000CC)
12. Max power:- It is useful for calculating maximum speeds (e.g. 135 km/h)
13. Seats:- It represents how many number of Persons are allowed to be seated (e.g. 5)
14. Selling price:- It is the amount a buyer pays for a car (e.g. ₹50100)

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Implementation Details

1. Importing Python Libraries

a) Pandas :-

It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analysing data much easier[9].

b) NumPy :-

NumPy is the fundamental package for scientific computing in Python. It provides a multidimensional array object, various derived objects (such as masked arrays and matrices)[10].

And an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more[10].

c) Matplotlib:-

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays.

d) Seaborn:-

Seaborn is a visualization library in Python. To analyse a set of data using Python, we make use of Matplotlib, a widely implemented 2D plotting library Likewise. It is built on top of Matplotlib [11].

Visualizing univariate and bivariate data.

Plotting statistical time series data.

Seaborn works well with NumPy and Pandas data structures.

e) Scikit-Learn:-

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python[7]. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python[7].

Predicting Price of Used Cars from Cardekho ¶

Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

2. Importing Data

The raw data was obtained from Kaggle and was in coma separated values (CSV) form,

Importing Data

```
data=pd.read_csv('cardekho_pre.csv')
```

```
data.head(10)
```

	id	car_name	brand	model	vehicle_age	km_driven	seller_type	fuel_type	transmission_type	mileage	engine	max_power	seats	selling_price
0	0	Maruti Alto	Maruti	Alto	9	120000	Individual	Petrol	Manual	19.70	796	46.30	5	120000
1	1	Hyundai Grand	Hyundai	Grand	5	20000	Individual	Petrol	Manual	18.90	1197	82.00	5	550000
2	2	Hyundai i20	Hyundai	i20	11	60000	Individual	Petrol	Manual	17.00	1197	80.00	5	215000
3	3	Maruti Alto	Maruti	Alto	9	37000	Individual	Petrol	Manual	20.92	998	67.10	5	226000
4	4	Ford Ecosport	Ford	Ecosport	6	30000	Dealer	Diesel	Manual	22.77	1498	98.59	5	570000
5	5	Maruti Wagon R	Maruti	Wagon R	8	35000	Individual	Petrol	Manual	18.90	998	67.10	5	350000
6	6	Hyundai i10	Hyundai	i10	8	40000	Dealer	Petrol	Manual	20.36	1197	78.90	5	315000
7	7	Maruti Wagon R	Maruti	Wagon R	3	17512	Dealer	Petrol	Manual	20.51	998	67.04	5	410000
8	8	Hyundai Venue	Hyundai	Venue	2	20000	Individual	Petrol	Automatic	18.15	998	118.35	5	1050000
9	9	Mahindra TUV	Mahindra	TUV	4	70000	Dealer	Diesel	Manual	18.49	1493	100.00	7	575000

The data was imported using the **pandas read_csv** method[9]. The top 10 rows are displayed above

3. Data Summary

Once the data was imported and loaded in pandas dataframe, the null values in the data was checked[9].

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Data Summary

Checking for NA ¶

```
print(data.isnull().sum())
```

```
id                  0
car_name            0
brand               0
model               0
vehicle_age         0
km_driven           0
seller_type         0
fuel_type           0
transmission_type   0
mileage             0
engine              0
max_power           0
seats               0
selling_price       0
dtype: int64
```

In the columns there were no null or NA values that was found.

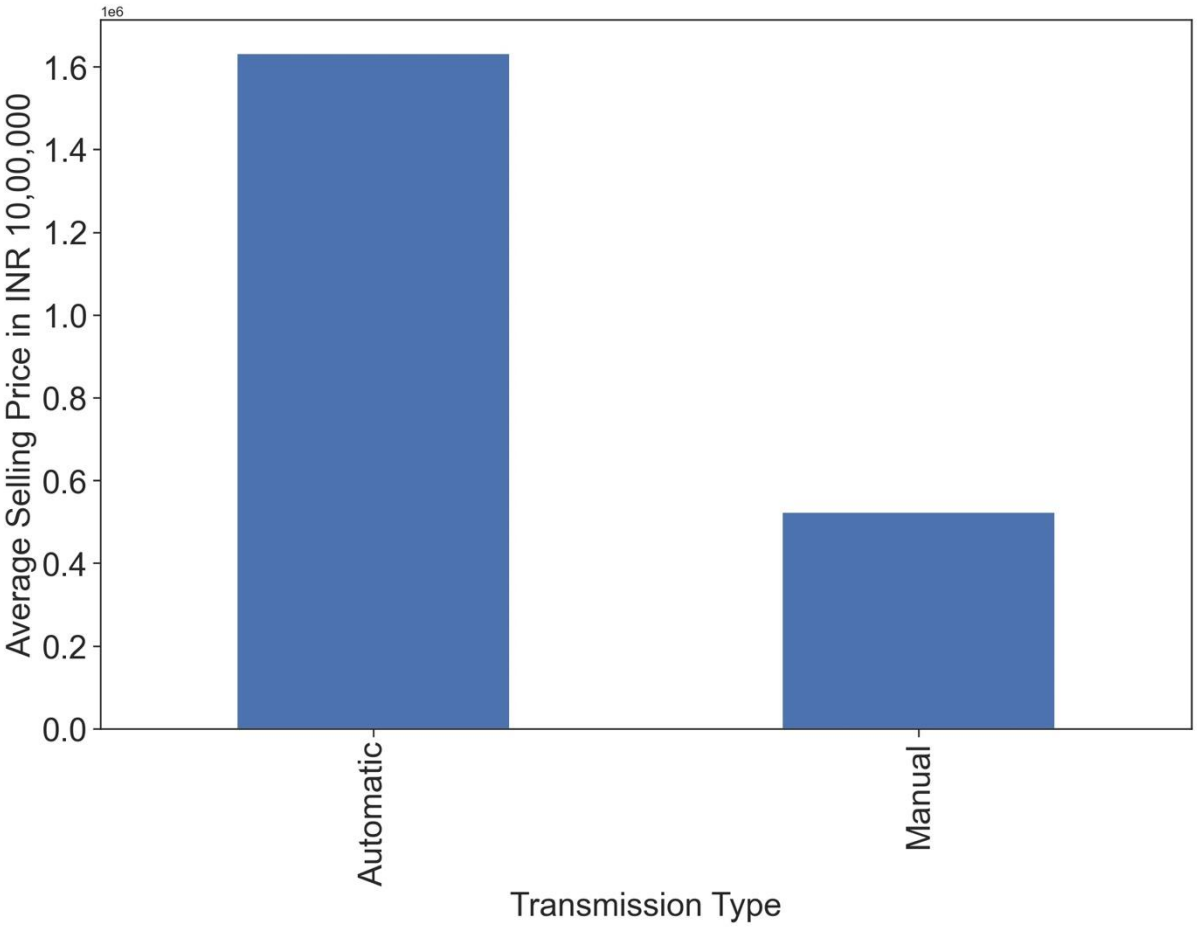
```
data.info()
```

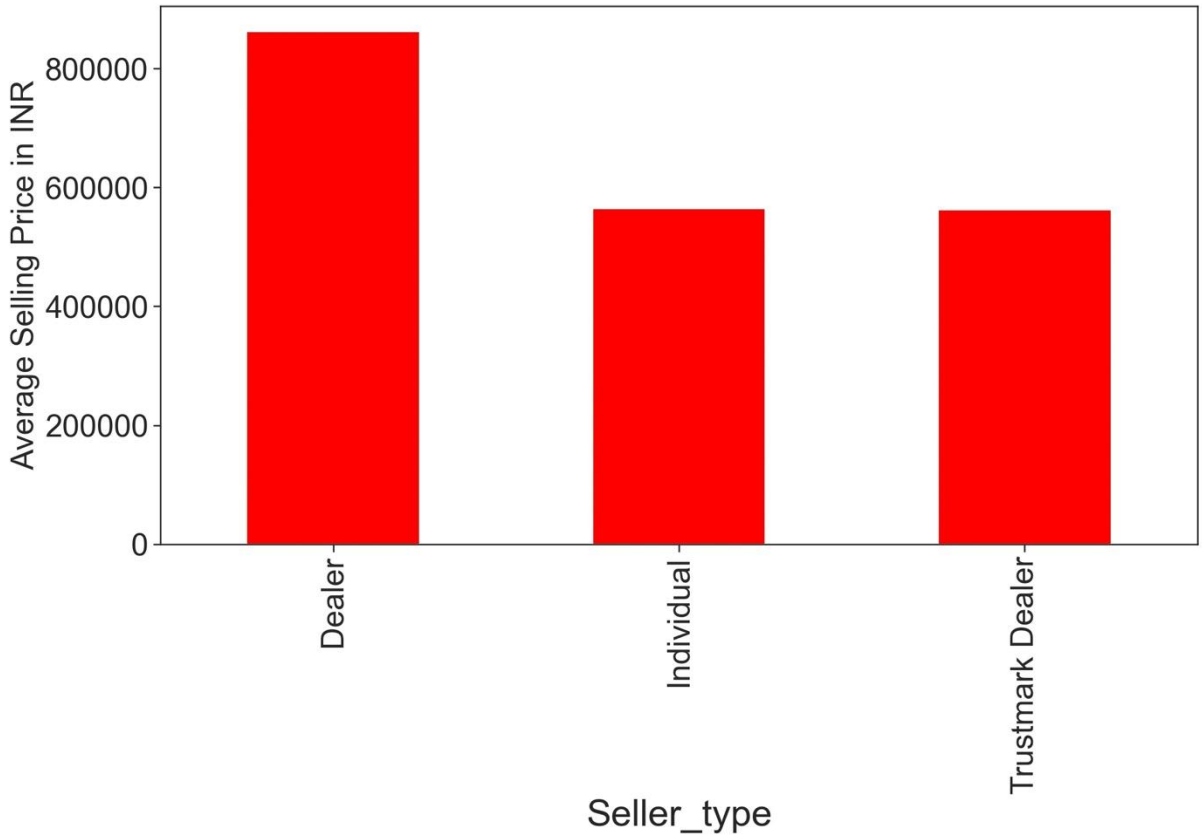
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19542 entries, 0 to 19541
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 19542 non-null  int64  
 1   car_name           19542 non-null  object 
 2   brand              19542 non-null  object 
 3   model              19542 non-null  object 
 4   vehicle_age        19542 non-null  int64  
 5   km_driven          19542 non-null  int64  
 6   seller_type        19542 non-null  object 
 7   fuel_type          19542 non-null  object 
 8   transmission_type  19542 non-null  object 
 9   mileage            19542 non-null  float64
 10  engine             19542 non-null  int64  
 11  max_power          19542 non-null  float64
 12  seats              19542 non-null  int64  
 13  selling_price      19542 non-null  int64  
dtypes: float64(2), int64(6), object(6)
memory usage: 2.1+ MB
```


Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

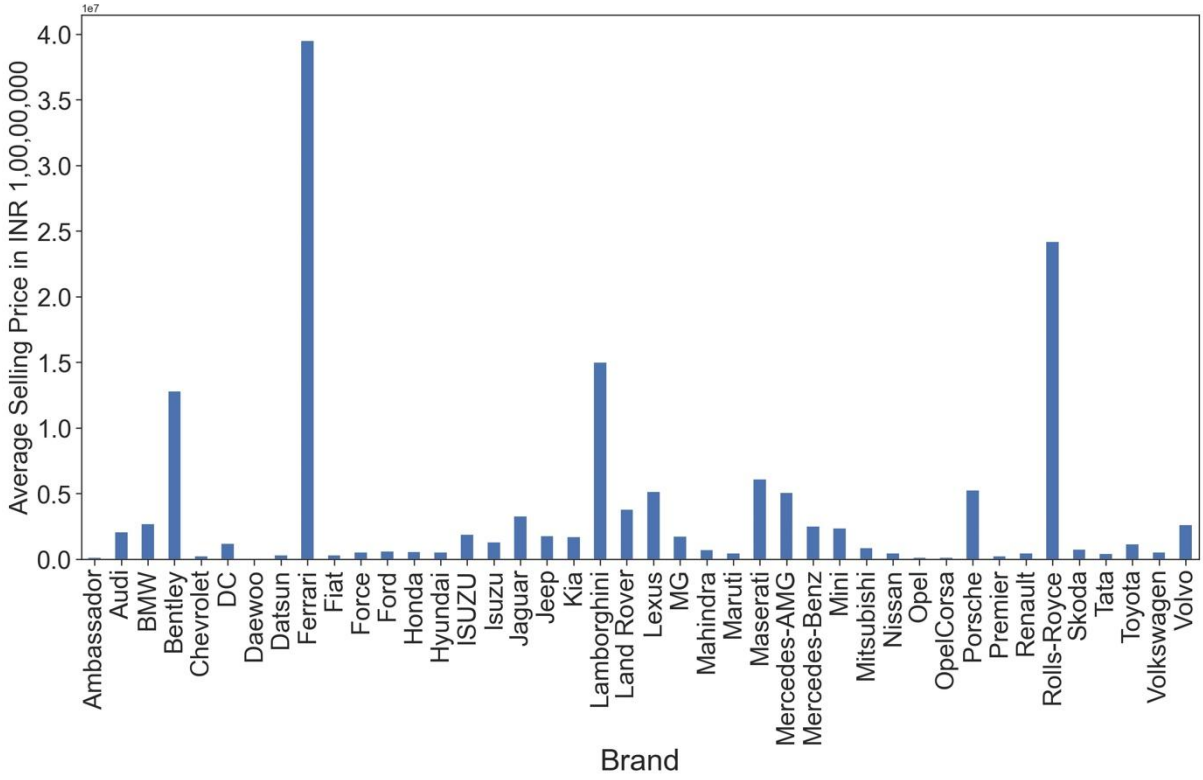
4. Exploratory Data Analysis

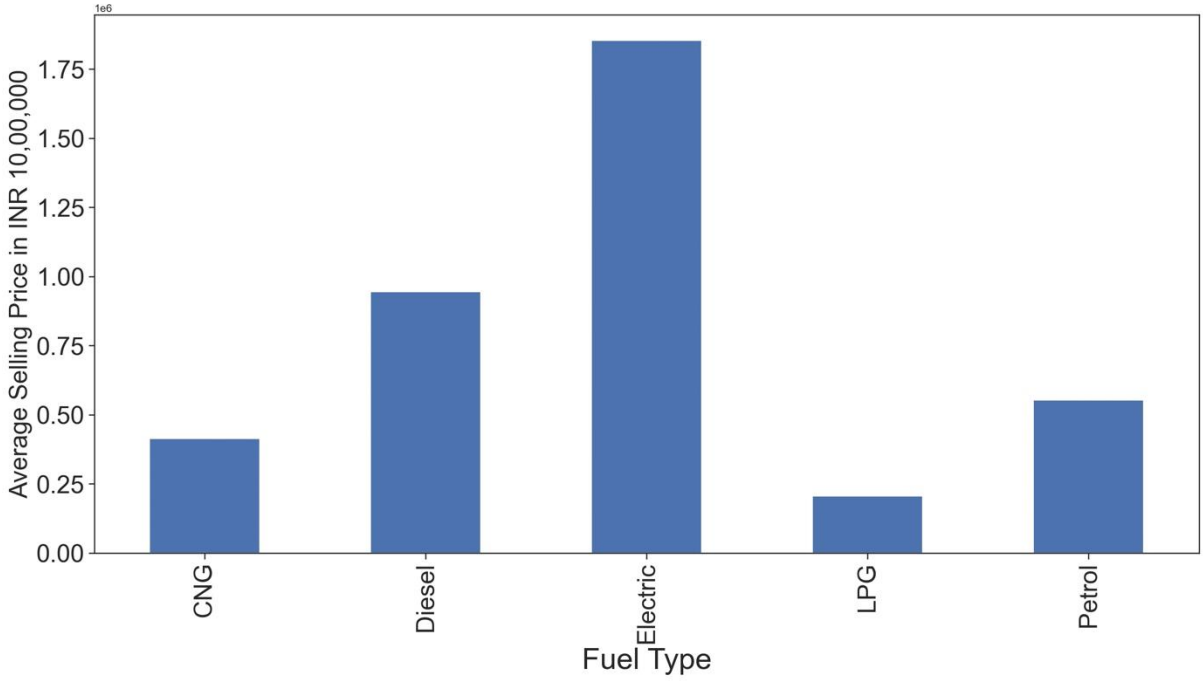
Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics, often plotting them visually[12]. This is a very important step especially when we arrive at modelling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter, etc.





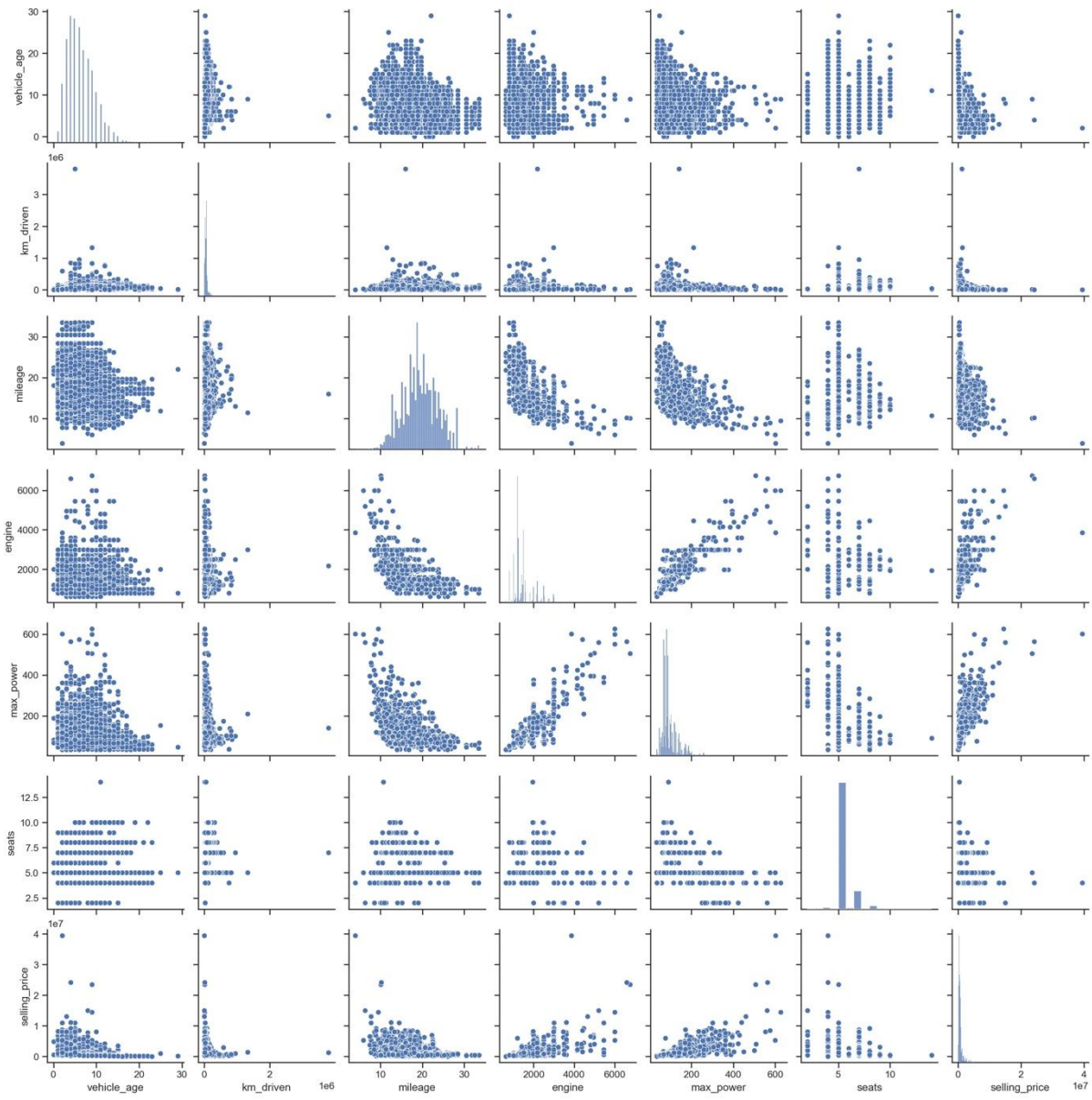
Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)





Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

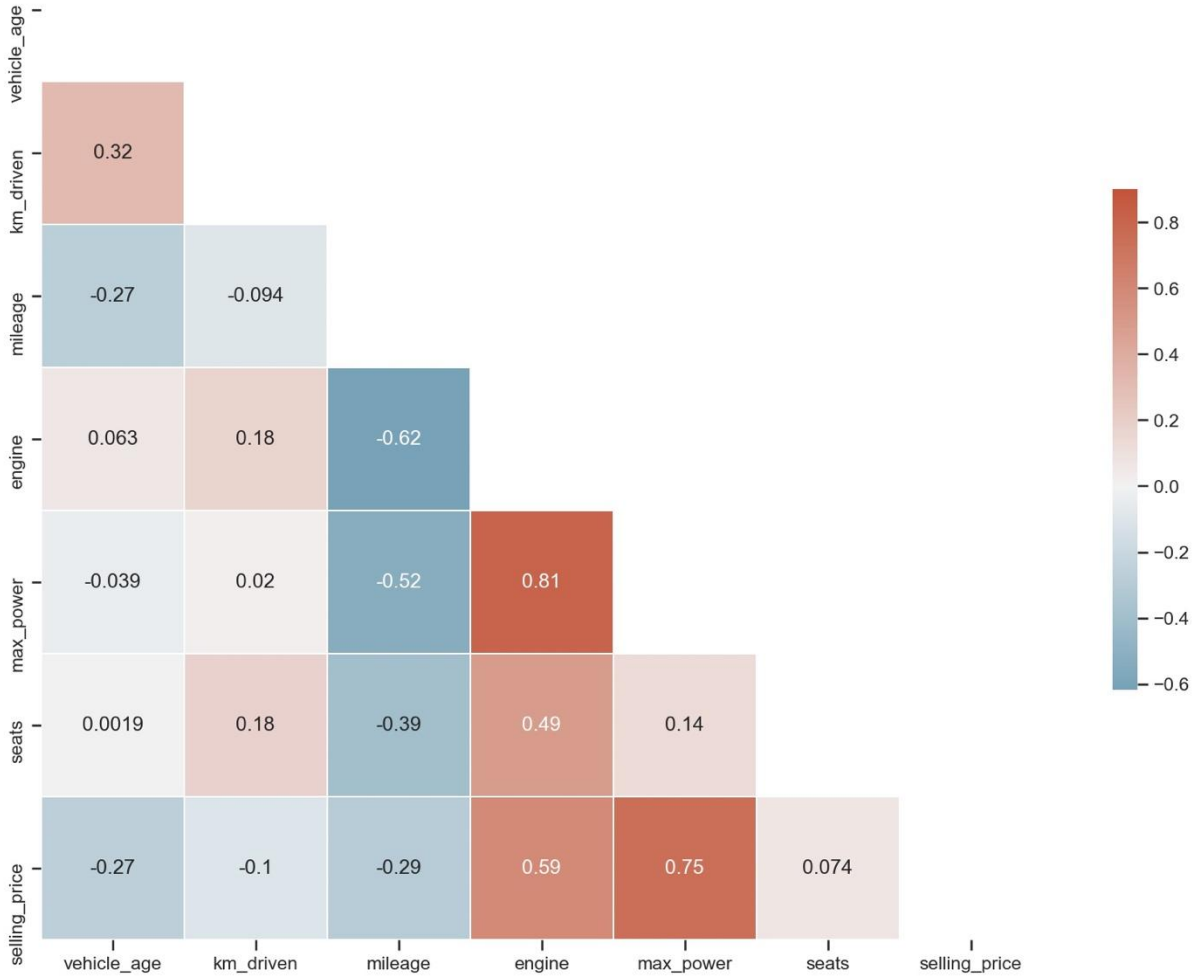
Pairplot of the features in the dataset



This pairplot helps in identifying relationship of various features with the target variable which

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Correlation Matrix



The correlation matrix helps identify the predictor variables and their correlation with the target variables, it also indicates the correlation present among the features variables[13].

5. Outlier Detection and Removal

The dataset contained outliers which affects the performance of regression model and negatively affects the mean absolute error, thus outliers were detected first by plotting a box and whiskers plot and then using the IQR calculation to detect the outliers[12].

Outlier Detection and Removal

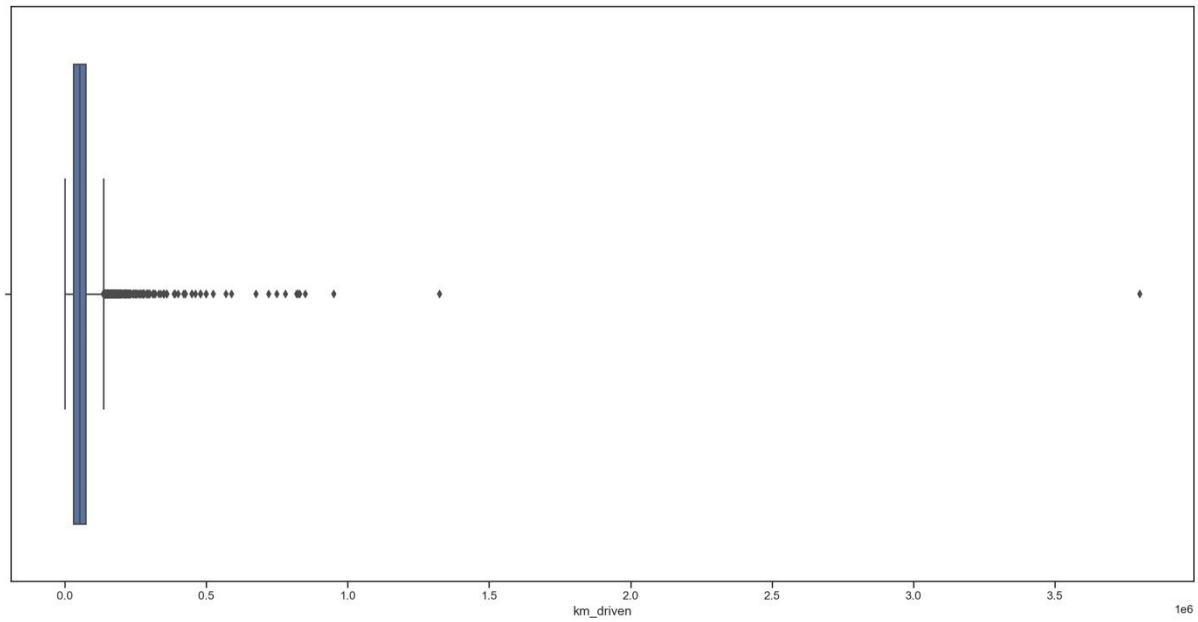
[14]:
```
# Data Summary of the data, here the include all keyword will give the summary of all the columns
data.describe(include='all')
```

[14]:

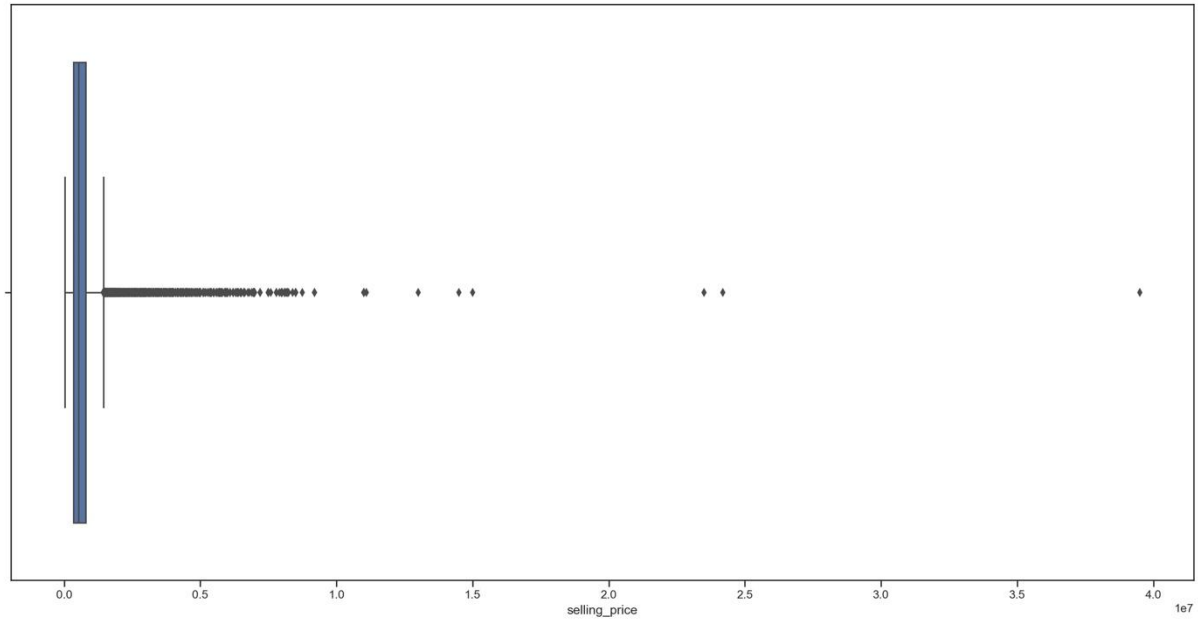
	id	car_name	brand	model	vehicle_age	km_driven	seller_type	fuel_type	transmission_type	mileage	engine	max_power	
count	19542.000000	19542	19542	19542	19542.000000	1.954200e+04	19542	19542	19542	19542.000000	19542.000000	19542.000000	1954
unique	NaN	262	42	258	NaN	NaN	3	5	2	NaN	NaN	NaN	
top	NaN	Hyundai i20	Maruti	i20	NaN	NaN	Dealer	Diesel	Manual	NaN	NaN	NaN	
freq	NaN	906	5547	906	NaN	NaN	11724	9641	15674	NaN	NaN	NaN	
mean	9770.500000	NaN	NaN	NaN	6.380309	5.787375e+04	NaN	NaN	NaN	19.508373	1478.475591	99.405598	
std	5641.433816	NaN	NaN	NaN	3.152225	5.054321e+04	NaN	NaN	NaN	4.075127	519.077385	43.717336	
min	0.000000	NaN	NaN	NaN	0.000000	1.000000e+02	NaN	NaN	NaN	4.000000	624.000000	34.200000	
25%	4885.250000	NaN	NaN	NaN	4.000000	3.100000e+04	NaN	NaN	NaN	16.950000	1197.000000	73.940000	
50%	9770.500000	NaN	NaN	NaN	6.000000	5.150000e+04	NaN	NaN	NaN	19.300000	1248.000000	86.800000	
75%	14655.750000	NaN	NaN	NaN	8.000000	7.300100e+04	NaN	NaN	NaN	22.320000	1582.000000	113.420000	
max	19541.000000	NaN	NaN	NaN	29.000000	3.800000e+06	NaN	NaN	NaN	33.540000	6752.000000	626.000000	1

The standard deviation in the case of Kilometer Driven is around 50500 KM, the standard deviation of the selling price is around INR 9,00,000 both standard deviations are high indicating outliers

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)



The above boxplot clearly shows that there are outliers in the Kilometres driven column



The above boxplot clearly shows that there are outliers in the selling price column

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Calculation for Outliers

$$IQR = Q_3 - Q_1$$

Calculation for Upper Bound and Lower Bound

$$Upper\ Bound = Q_3 + 0.5 \times IQR$$

$$Lower\ Bound = Q_1 - 0.5 \times IQR$$

Removing Outliers using IQR

```
#Removing Outliers of Selling Price and Kilometers Driven
z=data['selling_price']
Q1 =  z.quantile(0.25)
Q3 =  z.quantile(0.75)
IQR = Q3 - Q1
up_b=Q3 + (0.5 * IQR)
l_b=Q1 - (0.5 * IQR)

outliers=((z < l_b) | (z > up_b))
np.unique(outliers,return_counts=True)
```

```
(array([False,  True]), array([16070,  3472]))
```

There are 3472 Outliers that are detected and removed from the selling price column

```
data_w_o=data[~outliers]
```

```
#Removing Outliers of Kilometers Driven
y=data_w_o['km_driven']
Q1 =  y.quantile(0.25)
Q3 =  y.quantile(0.75)
IQR = Q3 - Q1
up_b=Q3 + (0.5 * IQR)
l_b=Q1 - (0.5 * IQR)

outliers_km=((y < l_b) | (y > up_b))
np.unique(outliers_km,return_counts=True)
```

```
(array([False,  True]), array([13266,  2804]))
```

There are 2804 Outliers that are detected and removed from the kilometer driven column

```
data_w_o=data_w_o[~outliers]
```

```
data_w_o=data_w_o[~outliers_km]
```

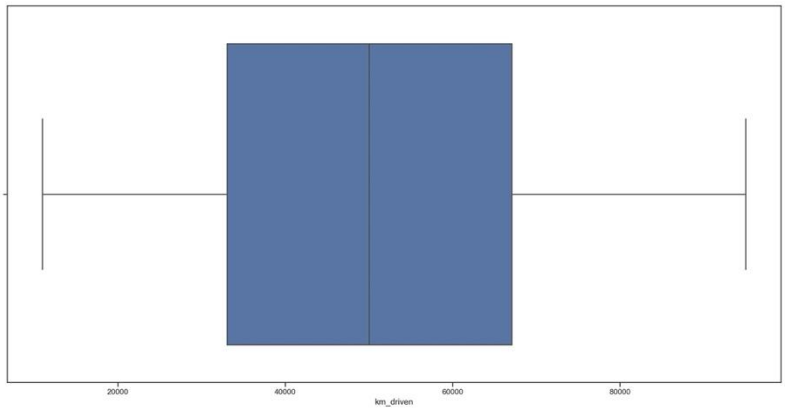
```
data_w_o.describe()
```

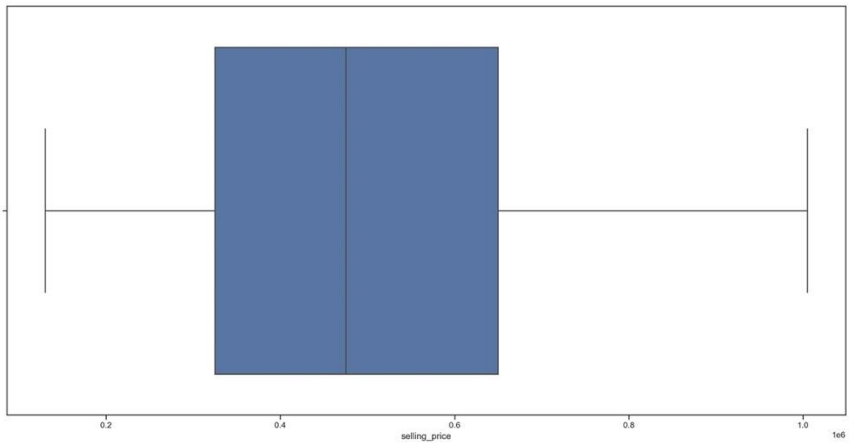
	id	vehicle_age	km_driven	mileage	engine	max_power	seats	selling_price
count	13266.000000	13266.000000	13266.000000	13266.000000	13266.000000	13266.000000	13266.000000	1.326600e+04
mean	9782.004071	6.337479	50321.438263	20.380748	1327.742500	87.004164	5.219207	4.987155e+05
std	5601.821249	2.795561	21290.423845	3.761103	349.365973	24.694924	0.685455	2.098800e+05
min	1.000000	0.000000	11000.000000	8.450000	624.000000	34.200000	2.000000	1.300000e+05
25%	4889.250000	4.000000	33036.750000	17.800000	1197.000000	73.000000	5.000000	3.250000e+05
50%	9838.500000	6.000000	50000.000000	20.360000	1248.000000	82.900000	5.000000	4.750000e+05
75%	14546.500000	8.000000	67051.250000	22.950000	1497.000000	98.600000	5.000000	6.500000e+05
max	19539.000000	25.000000	95000.000000	33.540000	5461.000000	388.000000	14.000000	1.005000e+06

There are **6276** Outliers that are detected and removed.

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

The criteria for Outlier detection and removal was using IQR (Inter-Quartile Range)





The standard deviation of Kilometres driven and Selling Price has reduced significantly, now since the outlier are removed we can apply the Regression Model

From the above Boxplot we can observe that Selling Price and Kilometres Driven have outliers

6. Converting Categorical Values to Numerical Values using one Hot encoding

Converting Categorical Values to Numerical Values using one Hot encoding

```
data_w_o=pd.get_dummies(data_w_o, columns=['brand','seller_type','model', 'fuel_type','transmission_type'],drop_first=True)
data_w_o.head(3)
```

	id	car_name	vehicle_age	km_driven	mileage	engine	max_power	seats	selling_price	brand_Audi	...	model_Yeti	model_Zen	model_Zest	model_i10	model_i20
1	1	Hyundai Grand	5	20000	18.90	1197	82.0	5	550000	0	...	0	0	0	0	0
2	2	Hyundai i20	11	60000	17.00	1197	80.0	5	215000	0	...	0	0	0	0	1
3	3	Maruti Alto	9	37000	20.92	998	67.1	5	226000	0	...	0	0	0	0	0

3 rows × 216 columns

Converting Categorical Values to Numerical Values using One Hot Encoding. The categorical values [Brand, seller_type, model, fuel_type, transmission_type] in our Data are not Ordinal in nature, there is no hierarchy. We used One Hot encoding Using Pandas method *pandas.get_dummies.* It will create a binary variable for each category

transmission_type
Manual
Manual

transmission_type_Manual
1
1

7. Dividing the Data into X(predictor variables) and y(target variable)

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Dividing the Data into X(predictor variables) and y(target variable) ¶

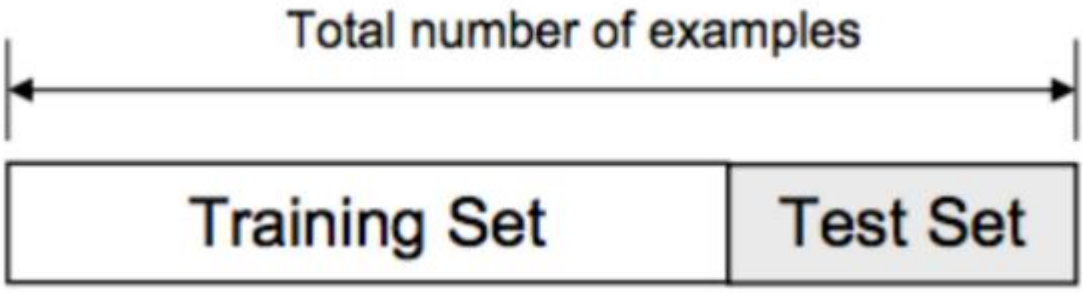
```
data_w_o=data_w_o.reset_index(drop=True)
#drop parameter to avoid the old index being added as a column:
data_w_o.head(2)
```

	id	car_name	vehicle_age	km_driven	mileage	engine	max_power	seats	selling_price	brand_Audi	...	model_Yeti	model_Zen	model_Zest	model_i10	model_i20
0	1	Hyundai Grand	5	20000	18.9	1197	82.0	5	550000	0	...	0	0	0	0	0
1	2	Hyundai i20	11	60000	17.0	1197	80.0	5	215000	0	...	0	0	0	0	1

2 rows × 216 columns

8. Splitting the data into Train set and Test Set

We utilized a 70% - 30% split for the training and test data. The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset[14].



Spliting the data into Train set and Test Set in the ration of 70:30

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=2)
print("x train: ",X_train.shape)
print("x test: ",X_test.shape)
print("y train: ",y_train.shape)
print("y test: ",y_test.shape)
```

```
x train:  (9286, 213)
x test:  (3980, 213)
y train:  (9286,)
y test:  (3980,)
```

9. Applying Multiple Linear Regression

Multiple Linear Regression is a case of linear regression with two or more independent variables[5]. Regression is used to predict continuous numerical values, and since our predictor variable is selling price which is a continuous variable, we have employed Linear Regression. In our model we have multiple independent variables X= [‘vehicle_age’, ‘mileage’, ‘engine’].

Importing LinearRegression Model from Sklearn

```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
```

Applying the linear Regression on the X_train and y_train dataset

```
lr.fit(X_train,y_train)
```

```
LinearRegression()
```

```
y_pred=lr.predict(X_test)
```

```
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
```

```
mean_absolute_error(y_test,y_pred).round()
```

```
65456.0
```


Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

10. Model Evaluation

```
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
```

```
mean_absolute_error(y_test,y_pred).round()
```

```
65456.0
```

The mean absolute Error of the Model is INR 65456

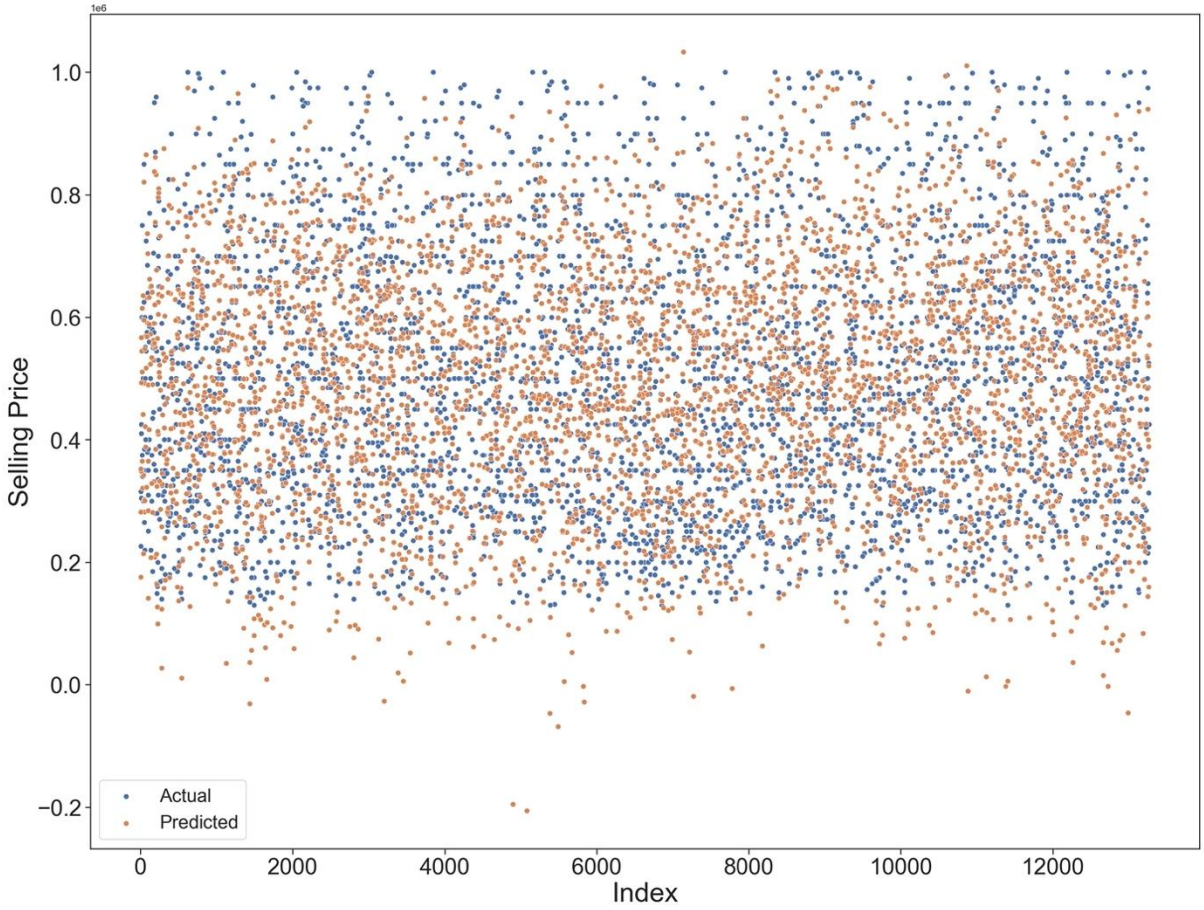
```
r2_score(y_test,y_pred).round(2)
```

```
0.84
```

The R^2 value of the Model is 0.84 meaning 84% of the Variation in the selling price is explained by the model

Paste cells from the clipboard

```
plt.figure(figsize=(22,17))
sns.scatterplot(x=y_test.index, y=y_test)
sns.scatterplot(x=y_test.index, y=y_pred)
plt.xlabel("Index",size=30)
plt.ylabel("Selling Price",size=30)
plt.xticks(fontsize=25)
plt.yticks(fontsize=25)
plt.legend(("Actual", "Predicted"),loc=3, fontsize='xx-large')
plt.savefig("result.jpg", dpi=150, bbox_inches="tight")
plt.show()
```



Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

```
y_pred_X=lr.predict(X)
```

```
X['Y_Pred']=y_pred_X
X['Selling_Price']=data_w_o['selling_price']
```

```
X['Percentage_Error']=((abs(X.Selling_Price-X.Y_Pred))/(X.Selling_Price))*100
```

```
percentage_error=X.Percentage_Error.sum()/(len(X.Percentage_Error)+1)
percentage_error.round()
```

```
16.0
```


Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Results

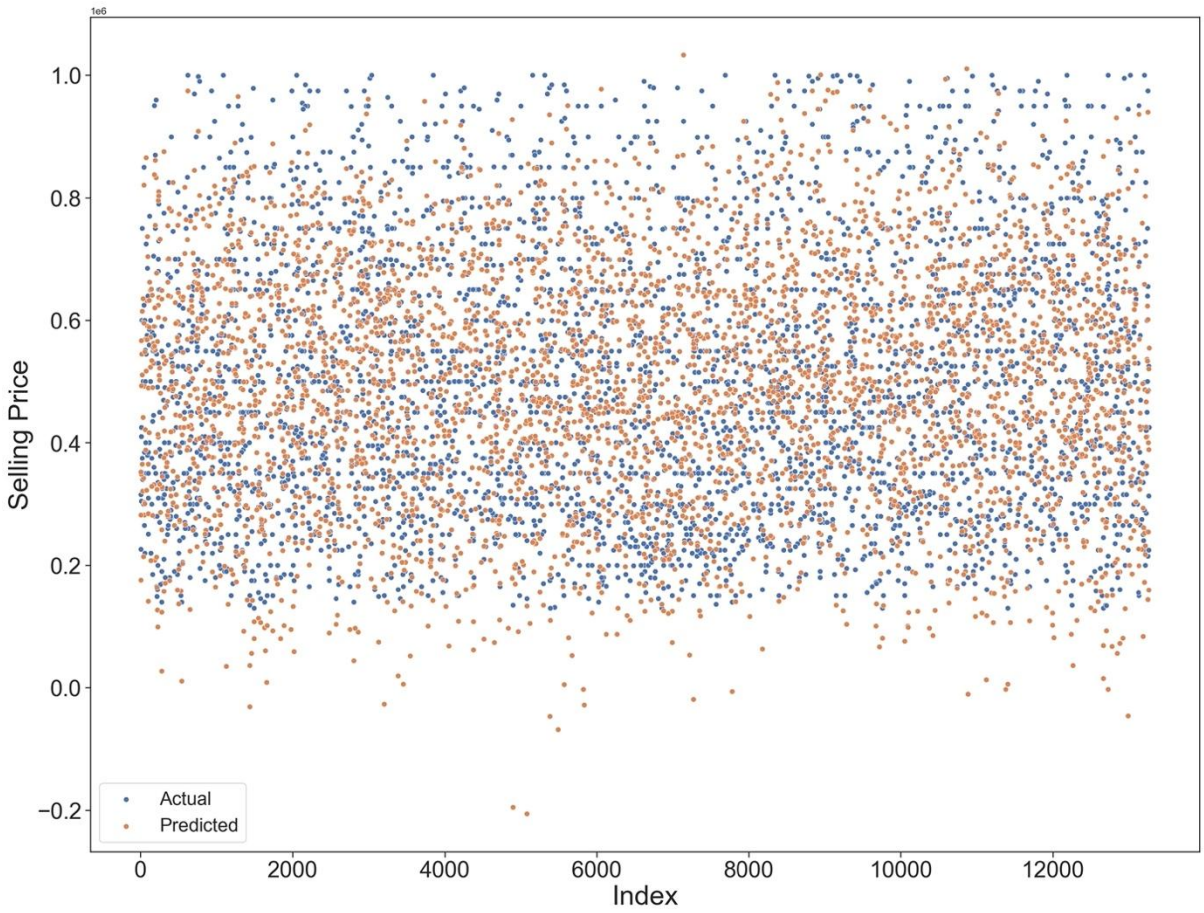
From the model and research, it can be concluded that the maximum power plays the most significant role in the cars selling price. It can be used as a feature to estimate the car's selling price. Based on the study findings, vehicle age, kilometre driven and mileage has an inverse relation with the selling price, whereas engine in cubic centimetres, maximum power and number of seats are positively correlated with the selling price of a used car. In essence the consumers should pay close attention to the max power, number of seats, the kilometre driven and the millage of the particular car they want to sell online.

Applying the Multiple Linear Regression through the Scikit Learn Library on the Car Dekho dataset, we were able to obtain a mean absolute error of INR 65,457, meaning on an average our predicted estimate of the car price was off by, on the positive and the negative side, INR 65,000. The R squared of the regression model was 0.84, implying that 84% of the variation in the response variable that is the second hand car price is explained by our model.

Mean Absolute Error	INR 65457
R Squared	0.84
Average Percentage Error	16 %

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

The above table summarises the results of our Multiple Linear Regression model. The average percentage change is calculated by taking the absolute of the actual selling price of the car and the predicted selling price, and dividing it by the actually selling price and multiplying by 100. The results for each observation is summed and divided by the number of observation to arrive at the average percentage error



The above scatter plot represents the actual price of a second hand car in blue and the price predicted by the our model in orange. The index is used as x-axis

to ensure consistency, thus ensuring the instance of the car is compared with the actual and predicted price.

ID	Y Predicted	Selling Price
0	458154.0	550000
1	295850.0	215000
2	175513.0	226000
3	676172.0	570000
4	282210.0	350000
5	352230.0	315000
6	509102.0	410000
7	710201.0	575000
8	348292.0	305000
9	565298.0	511000
10	493104.0	410000
11	566848.0	425000
12	640876.0	750000
13	679656.0	650000
14	560607.0	627000
15	461199.0	425000
16	544654.0	600000
17	551373.0	575000
18	446128.0	425000
19	236472.0	230000

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

Table 1.1

The above table 1.1 represents the first 20 observation of the second-hand car selling price on the website and the selling price predicted by the regression model.

Conclusion

In this project we have used regression analysis and have predicted the selling price of used cars based on various features of the cars.

- The R Squared value of our regression model is 84%. Meaning that 84% of the variation in the Selling Price is explained by our model.
- The Mean Absolute Error is **65457**. Meaning on average our prediction is off by + or – INR 65457.

From this study we can conclude that the use of multiple linear regression can help in predicting the price of used cars and it will also help customers to get the best price for their cars when they visit the website or application.

Suggestions for Further Study

To get even more accurate performance, we can choose more advanced machine learning algorithms such as decision tree, random forests and Use of Multiple Polynomial Regression which creates multiple decision or regression trees. To correct for overfitting in Multiple Linear Regression, different selections of

Submitted by :- Amarendra Pratap Deo
M.S.c(AI & ML)

features and number of trees will be tested to check for change in performance by using Random Forest and Decision tree.

1) Use Decision Tree for better accuracy:-

- Decision tree requires fewer data pre-processing from the user, for example, there is no need to normalize columns. It can be used for feature engineering such as predicting missing values, suitable for variable selection.

2) Use Random Forest:-

- One major advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems

3) Introduce condition of the car as a predictor variable in the model

- A variable called condition of the car, can be introduced in the model, which is a unique selling point to sell a car as it will predict the condition of the car which means whether the car is in good or bad condition. This variable will help customers to find out whether a particular car is in a good or bad condition and it will make it easier for customers to make their decisions whether buying a particular car is worth it or not.

4) Feature Selection using VIF

- Variance inflation factor (VIF) is a technique to estimate the severity of multicollinearity among independent variables within the context of a regression.

5) Use Multiple Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation.

$$Y = a + bX + cX^2$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

6) T-test and F-test for validation of model

- The T-test is a form of the statistical hypothesis test and its main advantage of the t-test is we can do it directly from the usual regression output, even if one didn't think to prepare the output to test that hypothesis. The F-Test indicates whether a linear regression model provides a better fit to the data than a model that contains no independent variables. It consists of the null and alternate hypothesis and the test statistic helps to prove or disprove the null hypothesis.

Bibliography

[1] G. S. P. Ltd, "About CarDekho," 2021. https://www.cardekho.com/info/about_us.

[2] "INDIA USED CAR MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS." https://www.mordorintelligence.com/industry-reports/india-used-car-market.

[3] S. Kuiper, "Introduction to Multiple Regression: How Much Is Your Car Worth?," *J. Stat. Educ.*, vol. 16, no. 3, 2008, doi: 10.1080/10691898.2008.11889579.

[4] M. Listiani, “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application,” *Technology*, no. March, 2009.

[5] V. A. Barbur, D. C. Montgomery, and E. A. Peck, “Introduction to Linear Regression Analysis.,” *Stat.*, vol. 43, no. 2, p. 339, 1994, doi: 10.2307/2348362.

[6] C. Maklin, “R Squared Linear Regression,” 2019. https://towardsdatascience.com/statistics-for-machine-learning-r-squared-explained-425ddfebf667.

[7] scikit-learn developers (BSD License)., “Model evaluation with scikit-learn,” *2007 - 2020*, 2020. https://scikit-learn.org/stable/getting_started.html.

[8] W. Wang and Y. Lu, “Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 324, no. 1, 2018, doi: 10.1088/1757-899X/324/1/012049.

[9] W. McKinney and P. D. Team, “Pandas - Powerful Python Data Analysis Toolkit,” *Pandas - Powerful Python Data Anal. Toolkit*, p. 1625, 2015.

[10] N. Community, “NumPy User Guide 1.11,” p. 109, 2013.

[11] P. Michael Waskom, “An introduction to seaborn.” https://seaborn.pydata.org/introduction.html#.

[12] Edureka, “Python, The Why And How Of Exploratory Data Analysis In.” https://www.edureka.co/blog/exploratory-data-analysis-in-python/.

[13] P. Waskom Michael, “Plotting a diagonal correlation matrix.” https://seaborn.pydata.org/examples/many_pairwise_correlations.html.

[14] “Training and Test Sets: Splitting Data,” *Google Developers*, 2020. https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data.