VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belgaum: 590018

Internship report on

"PREDICTING THE PRICE OF USED CARS AND BIKES"

A Dissertation work submitted in partial fulfillment of the requirement for the award of the degree of

Bachelor of Engineering
in
Computer Science and Engineering

by

Rachana J M 1AY17CS069

Under the guidance of
Prof. Vani K S
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
ACHARYA INSTITUTE OF TECHNOLOGY
(Affiliated to Visvesvaraya Technological University, Belgaum)

2020-2021

ACHARYA INSTITUTE OF TECHNOLOGY

Acharya Dr. Sarvepalli Radhakrishnan Road, Soladevanahalli, Bangalore – 560107
(Affiliated to Visvesvaraya Technological University, Belgaum)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Certificate

Certified that the internship entitled **"Predicting the price of used cars and bikes"** is a bonafide work carried out by **Rachana JM (1AY17CS069)** in partial fulfillment for the award of degree of **Bachelor of Engineering in Computer Science & Engineering of Visvesvaraya Technological University**, Belgaum during the year **2020-2021.** It is certified that all corrections/ suggestions indicated for internal assessments have been incorporated in the Report deposited in the departmental library. The Internship report has been approved as it satisfies the academic requirements in respect of Internship prescribed for the **Bachelor of Engineering Degree**.

Signature of Guide	**Signature of H.O.D**	**Signature of Principal**
Prof. Vani K S	Dr. Prashanth C M	Principal
Assistant professor	Head of the Department	

Signature of Coordinator
Dr. V Nagaveni
Professor

External Viva

Name of the Examiners	**Signature with Date**
1.	
2.	

Company certificate

TEQUED LABS
INVENT - INNOVATE - ITERATE
RESEARCH AND INNOVATION HUB
No 10 Anjaneya Nagar Banashankari 3rd Stage Bangalore 85

CERTIFICATE ID: TQL20AIIB1124

CERTIFICATE OF COMPLETION

IS PROUDLY PRESENTED TO

RACHANA JM

FOR SUCCESSFULLY COMPLETING THE
ONE MONTH INTERNSHIP ON
ARTIFICIAL INTELLIGENCE
CONDUCTED BY TEQUED LABS
FROM 08-08-2020 TO 08-09-2020

Aditya S.K.
DIRECTOR
TEQUED LABS

CEO
TEQUED LABS

ACKNOWLEDGEMENT

I express my gratitude to our institution and management for providing us with good infrastructure, laboratory, facilities and inspiring staff, and whose gratitude was of immense help in completion of this report successfully.

I deeply indebted to **Dr. Prakash M R,** Principal, Acharya Institute of Technology, Bangalore, who has been a constant source of enthusiastic inspiration to steer us forward.

A hearty thank to **Dr. Prashanth C M**, Head of the Department, Department of Computer Science and Engineering, Acharya Institute of Technology Bangalore, for his valuable support and for rendering us resources for this Internship work.

I specially thank **Prof. Vani K S,** Assistant Professor, Department of Computer Science and Engineering who guided me with valuable suggestions in completing this Internship at every stage.

Also, I wish to express deep sense of gratitude for Internship coordinator Dr. **V Nagaveni**, Professor, Department of Computer Science and Engineering, Acharya Institute of Technology for her support and advice during the course of this final year internship.

I would also like to express sincere thanks and heartfelt gratitude to beloved Parents, Respected Professors, Classmates, Friends, Juniors for their indispensable help at all times.

Last but not the least a respectful thanks to the Almighty.

Rachana J M (1AY17CS069)

ABSTRACT

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. The predictions are based on historical data collected from the daily Kaggle website. Different techniques like multiple linear regression analysis have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. In the future, we intend to use more sophisticated algorithms to make the predictions.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

About the Industry

CHAPTER 1

ABOUT THE INDUSTRY

1.1 Who are we?

- Tequed labs is an Innovation and research hub based in Bangalore. It is focused on providing quality education to students regarding the various latest technologies developing all over the world everyday.
- It also focuses on creating innovative products to the society and also mentors various potential startups and ideas

1.2 Goals and achievement:

- Tequed Labs is a research and development center and educational institute based in Bangalore started by Mr Aditya S K and Mr Supreeth Y S. They have focused on providing quality education on latest technologies and develop products which are of great need to the society. They run a project consultancy where they undertake various projects from wide range of companies and assist them technically and build products and provide services to them. They are continuously involved in research about futuristic technologies and finding ways to simplify them for students.
- Tequed Labs has collaborated with 40 Institutions, and has trained over 9655 students in 30 different domains.

Chapter 2

Training

/

Technology

CHAPTER 2

TRAINING/TECHNOLOGY

Internships are available at Tequed Labs for undergraduate and graduate students who are actively pursuing a degree. Internships are available in many fields like testing, development, etc. Student internship lasts 4-8 weeks, and is a great way to enhance our skills, gain experience and help unlock our professional potential. The working environment is amazing and there are mentors who guides us during the internship.

2.1 Machine Learning:

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data and it is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

2.2 Python Programming Language:

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

2.3 Technique Used:

- Predicting the price of used cars and bikes is both an important and interesting problem. According to data obtained, the number of cars and bikes registered between 2007 and 2017 has witnessed a spectacular increase of 70%. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars and bikes will increase. It is reported that the sales of new cars has registered a decrease of 8% in 2016.

- Deciding whether a used car and bike is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car and bike. From the perspective of a seller, it is also a dilemma to price a used car appropriately.
- Linear Regression, is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out the cause and effect relationship between variables.
- Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Chapter 3

Software Design

CHAPTER 3

SOFTWARE DESIGN

3.1 Dataset:

For this project, we are using the dataset on used car and bikes sales available on Kaggle website. The features available in this dataset are name of the vehicle, vehicle type, year, odometer, fuel, owner, selling price.

vehicle_type	name	year	odometer	fuel	owner	selling_price
bike	bajaj pulsar 220c	2016	36922	petrol	FirstOwner	44400
car	Maruti 800 AC	2007	70000	petrol	FirstOwner	60000
bike	bajaj avenger 22(	2013	30000	petrol	FirstOwner	45000
car	Maruti Wagon R	2007	50000	petrol	FirstOwner	135000
bike	yamaha fz25 250	2017	10000	petrol	SecondOwner	71000
car	Hyundai Verna 1.	2012	100000	Diesel	FirstOwner	600000
bike	yamaha fz v2.0 1	2016	15366	petrol	FirstOwner	56000
car	Datsun RediGO T	2017	46000	petrol	FirstOwner	250000
bike	honda x blade 16	2019	9100	petrol	FirstOwner	81200
car	Honda Amaze V)	2014	141000	Diesel	SecondOwner	450000
bike	ktm rc 390cc	2015	37000	petrol	FirstOwner	98000
car	Maruti Alto LX BS	2007	125000	petrol	FirstOwner	140000
bike	bajaj pulsarns200	2018	11574	petrol	FirstOwner	88000
car	Hyundai Xcent 1.	2016	25000	petrol	FirstOwner	550000
bike	yamaha fZs	2015	19073	petrol	FirstOwner	46000
car	Tata Indigo Gran	2014	60000	petrol	SecondOwner	240000
bike	royal enfield clas:	2018	13700	petrol	FirstOwner	150000
car	Hyundai Creta 1.(	2015	25000	petrol	FirstOwner	850000
bike	bajaj avenger 22(	2015	24294	petrol	SecondOwner	29000

Figure 3.1: Sample data

3.2 Methodology:

The workflow of the machine learning model includes the following steps:

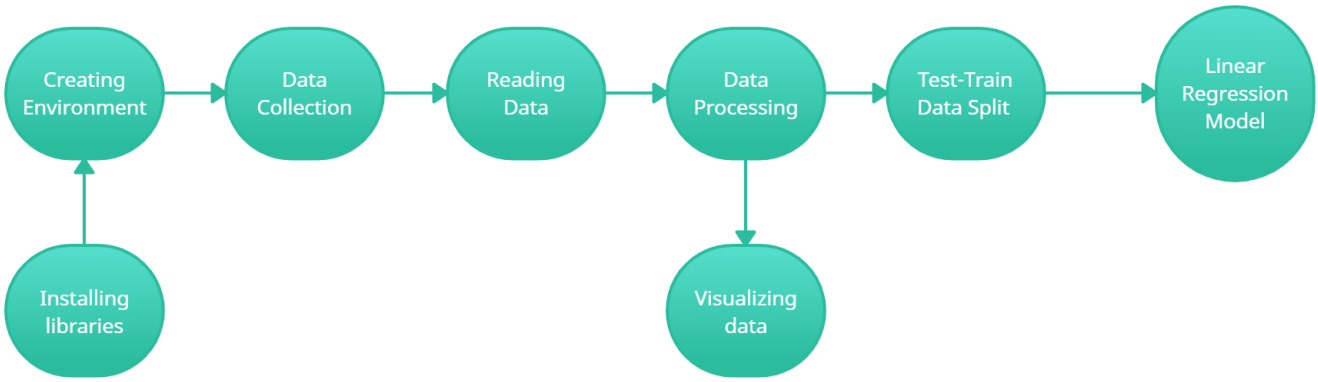

Figure 3.2: Model Flow

3.3 Model Design:

- This project aims to develop a good regression model to offer accurate prediction of car price. Based on existing data, we have used linear regression techniques to develop models for predicting used car and bike prices.

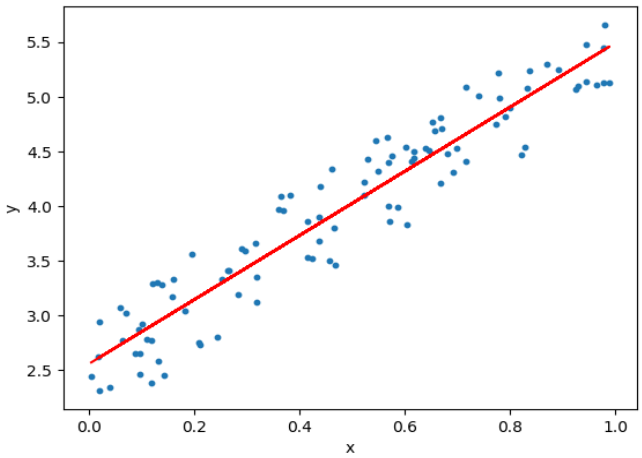

Figure 3.3: Linear Regression

- Linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the

independent(x) and dependent(y) variable.

- The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. So, this Linear regression technique finds out a linear relationship between x (input) and y (output).
- In our model we have used Multiple linear regression based on supervised learning which is an extension of linear regression, where in the number of independent variables is more than one.
- Once the data collection was over, we process the data using multiple linear regression technique for price prediction.

Chapter 4

Algorithms

CHAPTER 4

ALGORITHMS

4.1 Algorithm used for predicting the costs of used cars and bikes:

Linear regression is the algorithm employed to find out the extent up to which there is a linear relationship between some dependent and one or more independent variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Step 1: Reading and Understanding data:

- The dataset on used car and bikes sales are collected from Kaggle.
- The file can be saved in any of the formats (like csv, pdf, excel, txt etc).

Step 2: Importing Packages:

- Importing data using the pandas library.
- Understanding the structure of the data

Step 3: Testing the data:

To look at the data types to see which columns need to be encoded.

Step 4: Splitting the data:

To Split the dataset into independent(X) and dependent(Y) datasets.

Step 5: Building Regular Expression:

This is the most important step. According to the data, Regular Expression need to be constructed properly.

Chapter 5

Implementation

CHAPTER 5

IMPLEMENTATION DETAILS

5.1 Load libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

5.2 Load Data

```
dataset=pd.read_csv("used cars and bikes.csv")
```

5.3 Data Inspection

Checking for null values:

```
sns.heatmap(dataset.isnull())

plt.show()
```

5.4 Data Visualization

```
plt.figure(figsize=(10, 20))
plt.subplot(4,2,1)
sns.boxplot(x = 'vehicle_type', y = 'selling_price', data = dataset)
plt.subplot(4,2,2)
sns.boxplot(x = 'fuel', y = 'selling_price', data = dataset)
plt.subplot(4,2,3)
sns.boxplot(x = 'owner', y = 'selling_price', data = dataset)
plt.tight_layout()
plt.show()
```

5.5 Creating Dummies for categorical columns:

Categorical Variables are converted into Numerical Variables with the help of Dummy Variable as machine learning algorithm only understands numeric values.

```
vehicle=pd.get_dummies(dataset['vehicle_type'],drop_first=True)

print(vehicle)

Fuel=pd.get_dummies(dataset['fuel'],drop_first=True)
print(Fuel)

Owner=pd.get_dummies(dataset['owner'],drop_first=True)
print(Owner)
```

Now that we have created the dummy variables we drop the categorical variables and add the dummy variables to the dataset and view the new dataset with only numeric values.

```
dataset.drop(['vehicle_type','name','fuel','owner'],axis=1,inplace=True)
dataset=pd.concat([dataset,vehicle,Fuel,Owner],axis=1)

print(dataset)
```

5.6 Split the Data into Training and Testing sets:

Here, the dataset is split into 20% test data and 80% train data:

```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(dataset.drop('selling_price',axis=1),dataset['s
elling_price'],test_size=0.20,random_state=50)
```

5.7 Train the model using Linear Regression:

We train the model using multiple linear regression algorithm and also predict the selling price for the given values of input.

```
from sklearn.linear_model import LinearRegression
linreg=LinearRegression()
linreg.fit(x_train,y_train)
y_pred=linreg.predict(x_test)
print(y_pred)
y_pred1=linreg.predict([[2014,28000,1,0,0]])
print(y_pred1)
```

5.8 Calculating the mean squared error and model accuracy:

The mean squared error and the model accuracy is calculated.

```
from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test,y_pred))
print(linreg.score(x_test, y_test)*100,'% Prediction Accuracy')
```

Chapter 6

Results

CHAPTER 6

RESULTS

6.1 Reading and printing the dataset:

```
   vehicle_type                     name  year  odometer    fuel  \
0          bike        bajaj pulsar 220cc  2016     36922  petrol
1           car             Maruti 800 AC  2007     70000  petrol
2          bike       bajaj avenger 220cc  2013     30000  petrol
3           car  Maruti Wagon R LXI Minor  2007     50000  petrol
4          bike         yamaha fz25 250cc  2017     10000  petrol
..          ...                       ...   ...       ...     ...
75          car                       NaN  2015     23000  Diesel
76         bike                       NaN  2013     38000  petrol
77         bike                       NaN  2016     12000  petrol
78          car                       NaN  2017     17999  petrol
79         bike                       NaN  2013     40000  petrol

          owner  selling_price
0    FirstOwner          44400
1    FirstOwner          60000
2    FirstOwner          45000
3    FirstOwner         135000
4   SecondOwner          71000
..          ...            ...
75   FirstOwner         643000
76  SecondOwner          61000
77   FirstOwner          89000
78   FirstOwner         810000
```

Figure 6.1: Dataset

6.2 The data types to see which columns need to be encoded:

```
[80 rows x 7 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 80 entries, 0 to 79
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   vehicle_type   80 non-null     object
 1   name           60 non-null     object
 2   year           80 non-null     int64
 3   odometer       80 non-null     int64
 4   fuel           80 non-null     object
 5   owner          80 non-null     object
 6   selling_price  80 non-null     int64
dtypes: int64(3), object(4)
memory usage: 3.2+ KB
None
```

Figure 6.2: Data type to be encoded

6.3 To check for any null values:

Fig 6.3 heatmap indicates that we do not have any null values in the dataset.

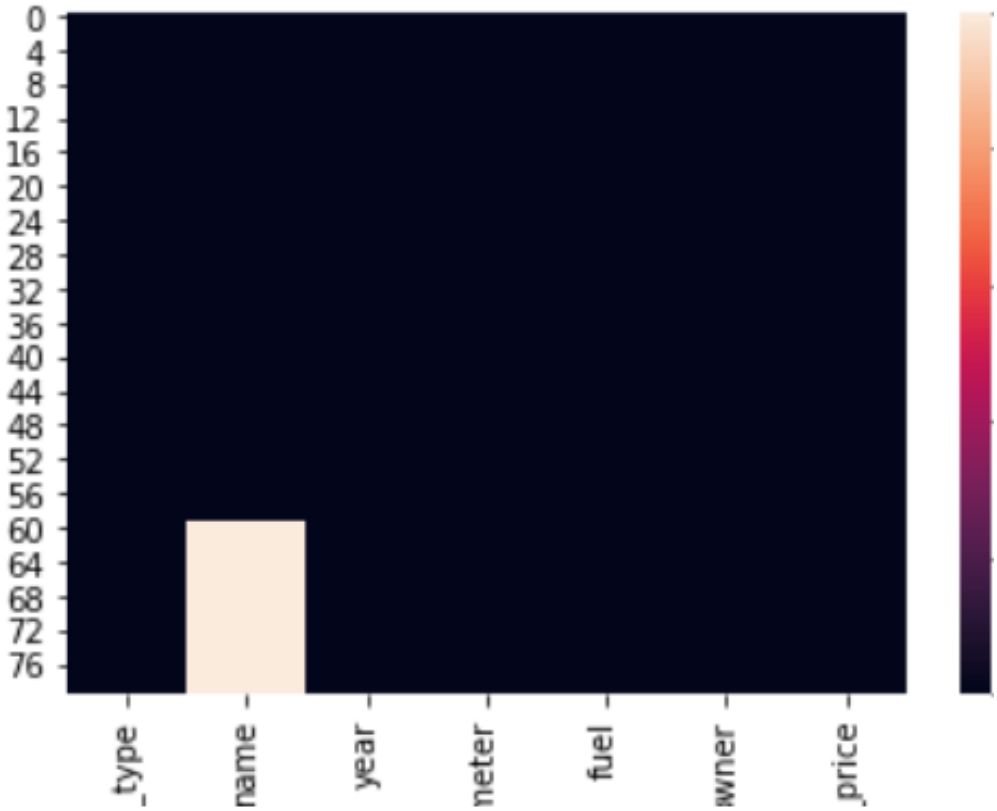

Figure 6.3: To check for any null values in the dataset

6.4 Graph against vehicle type versus price:

- Price column is the dependent value and rest of the column is independent values.
- So we plot a graph for each independent columns versus price column

Fig 6.4: Graph against vehicle type versus price

6.5 Graph against manufacturing year versus price:

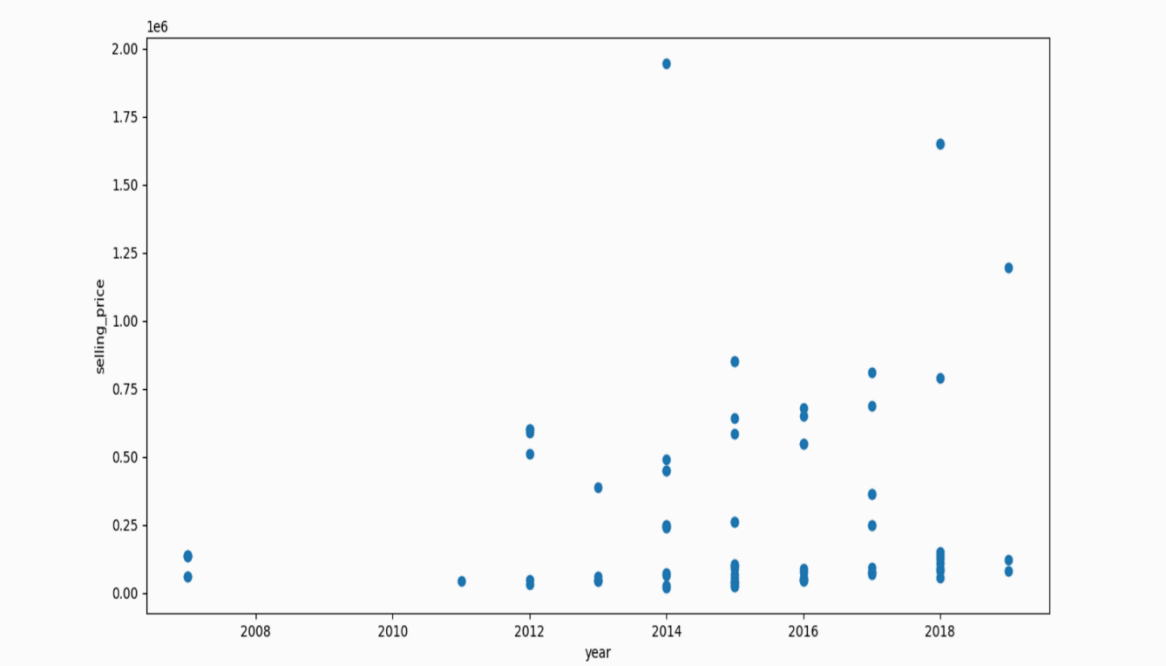

Fig 6.5: Graph against manufacturing year versus price

6.6 Graph against odometer versus price:

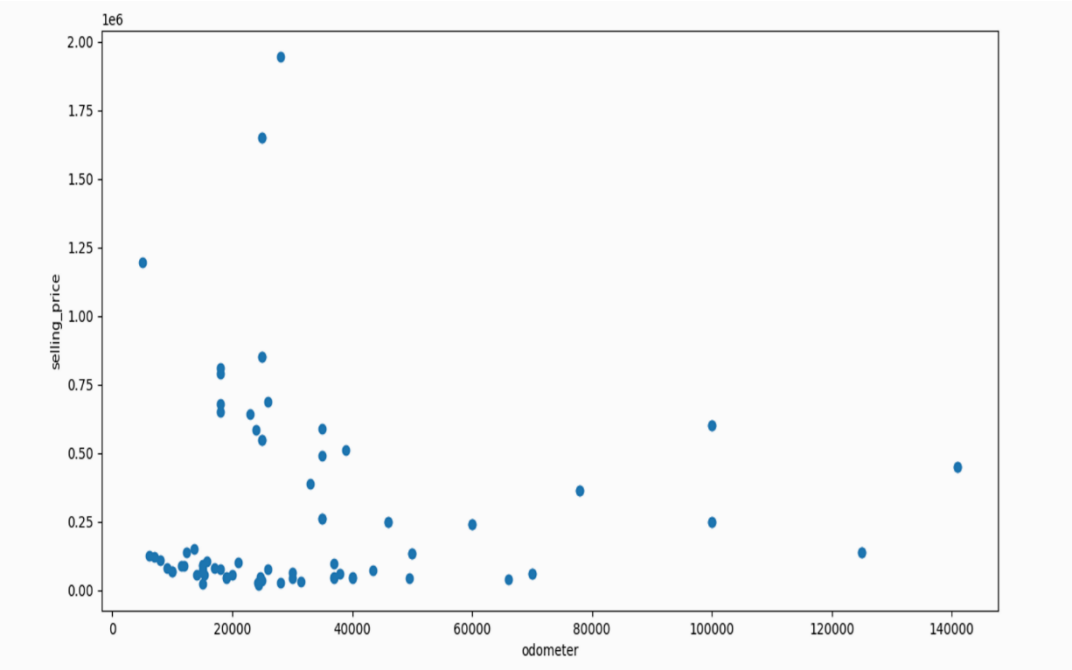

Fig 6.6: Graph against odometer versus price

6.7 Graph against fuel type versus price:

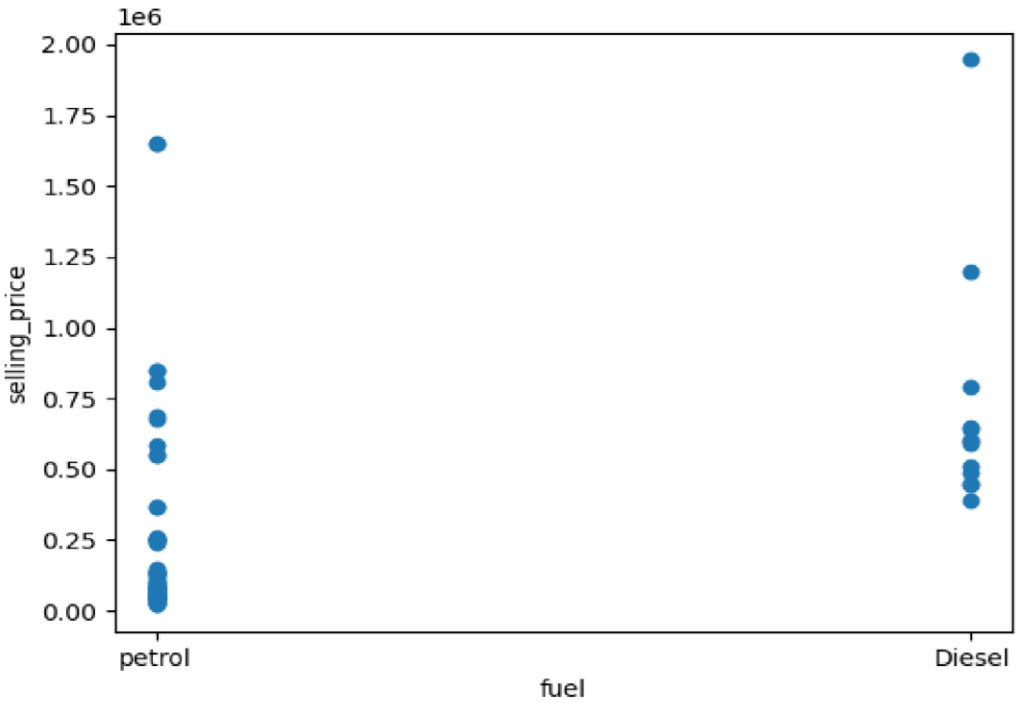

Fig 6.7: Graph against fuel type versus price

6.8 Graph against owner versus price:

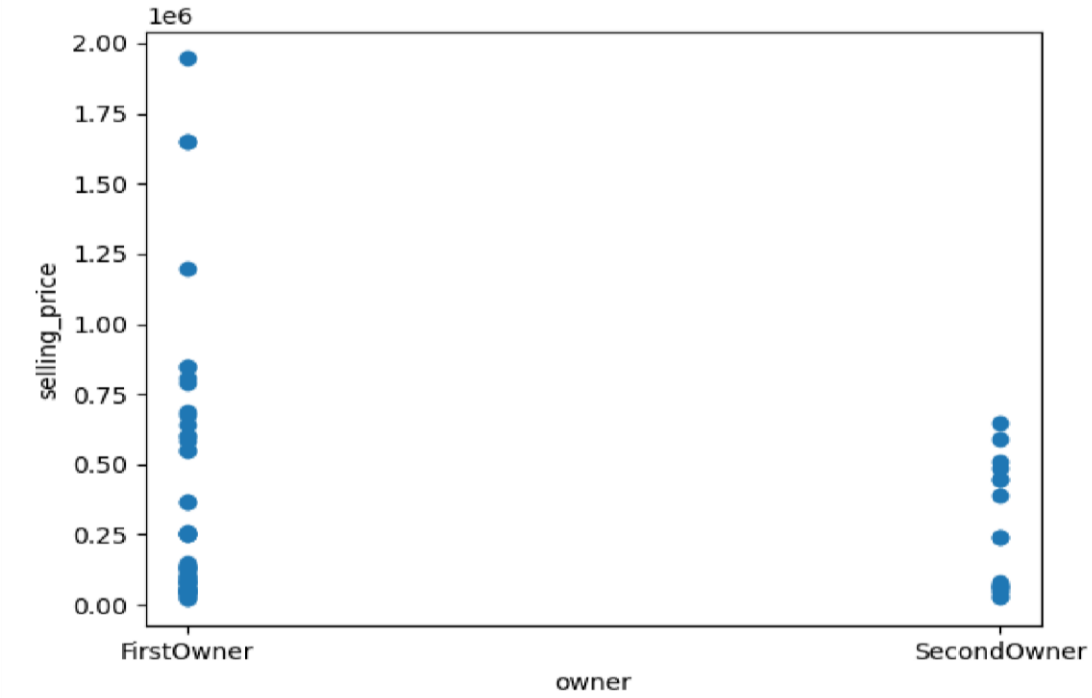

Fig 6.8: Graph against owner versus price

6.9 Visualizing numeric variables:

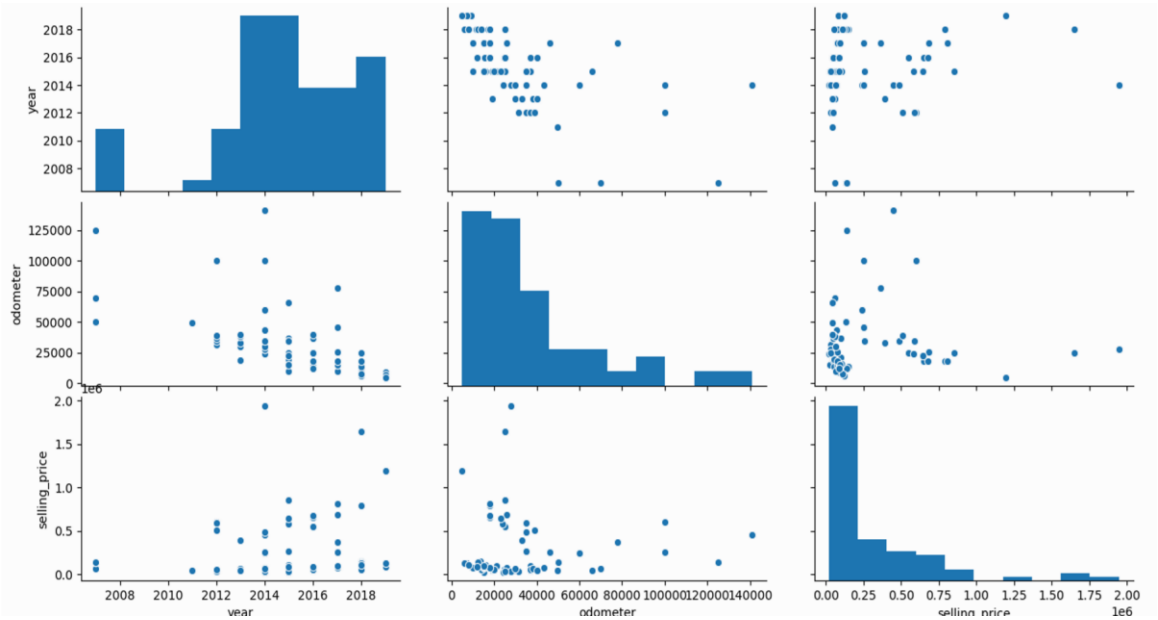

Fig 6.9: Graph visualizing numeric variables

6.10 Visualizing categorical variables:

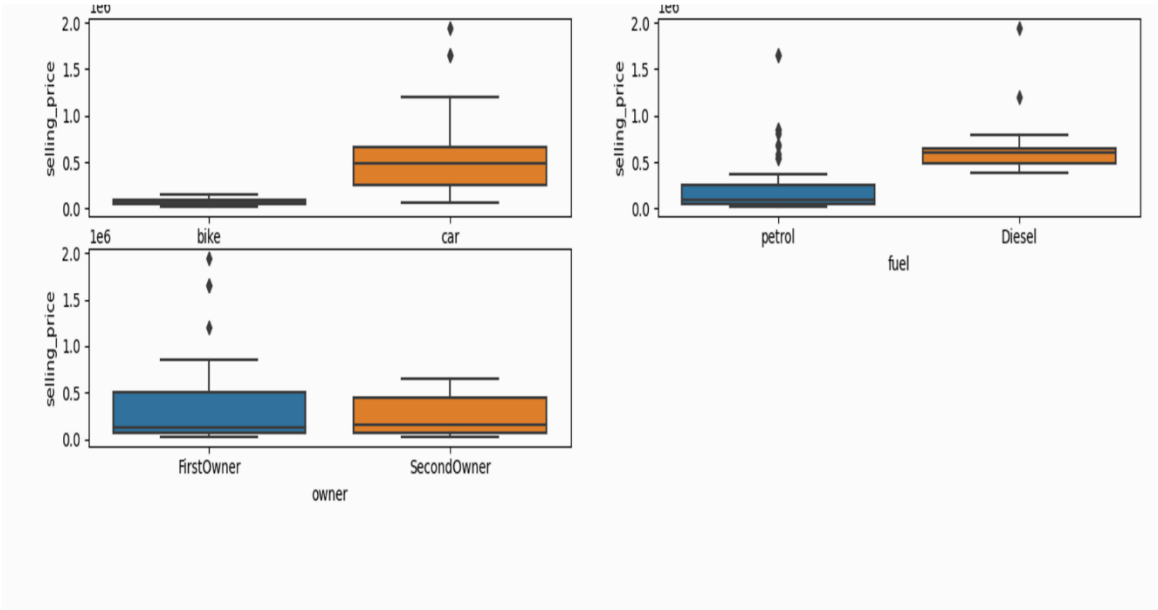

Fig 6.10: Graph visualizing categorical variables

6.11 The mean squared error and model accuracy:

Fig 6.11 shows the model accuracy of Linear Regression is 0.84 which is good and predictions were quite close to the original selling prices.

```
[80 rows x 6 columns]
[-1.70769782e+05  6.01120307e+05 -3.38140867e+02  5.33096962e+05
  2.27468187e+05  1.10590337e+05  6.72685792e+05  2.76941372e+05
  1.12603638e+05 -5.43587009e+04  4.05814872e+04  1.09594874e+06
  2.47601191e+05  3.62665320e+05  2.82578613e+05  1.86853523e+05]
[848582.81173451]
17114378203.607681
84.68881864383529 % Prediction Accuracy
```

Fig 6.11: Model Accuracy

Chapter 7

Learning Outcomes

CHAPTER 7

LEARNING OUTCOMES

VALUE GENERATED

A different exposure to the advanced technologies used in industry. Learnt how to build regular expression and how it is useful in real time. Understood how necessary and important to know the price of the vehicle we buy today would be tomorrow.

CHALLENGES AND DIFFICULTIES

- Data Collection is difficult: For some of the unique identification numbers, data collection is very difficult. So it will not lead to proper testing results.
- Building Regular Expression: For some of the data, the identification number is way too long and does not follow any proper pattern. So building regular expression is difficult.

FUTURE ENHANCEMENTS

- For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.
- Additional effectiveness could be achieved by: Continuing to increase the size of the dataset and by focusing on a defined subset of cars on the market.

CONCLUSION

- By the help of AI/ML code using python we can analyse any dataset given as per the topic chosen about used Car and bikes dataset.
- By using Linear Regression we trained our dataset, to predict the price with a fair amount of accuracy.

STIPEND

STIPEND RECEIVED DETAILS

This was an unpaid internship.