**Predicting Price for Used Cars using**

**Linear Regression**

**Artificial Intelligence CS-617-A**

**Avalons**

****

**Sacred Heart University**

School of Computer Science & Engineering

The Jack Welch College of Business & Technology

Submitted To:

**Dr. Reza Sadeghi**

**Fall 2022**

**Project Report of Predicting Price for used car using**

**Linear Regression**

**Team Name**

Name of the Team **Avalons**

**Team Members**

1. Sambasiva Rao Chennamsetty [chennamsettys@mail.sacredheart.edu](mailto:chennamsettys@mail.sacredheart.edu)(Team Head)
2. Arif Pasha Shaik [shaiks11@mail.sacredheart.edu](mailto:shaiks11@mail.sacredheart.edu)(TM)
3. Jagadishwar Reddy Velma [velmaj@mail.sacredheart.edu](mailto:velmaj@sacredheart.edu) (TM)
4. Sai Hrithik Peddi [peddis4@mail.sacredheart.edu](mailto:peddis4@mail.sacredheart.edu) (TM)
5. Vamsi Kiran Kakkera [kakkerav@mail.sacredheart.edu](mailto:ravulat2@mail.sacredheart.edu) (TM)
6. Kaki Rohit Reddy [kakir@mail.sacredheart.edu](mailto:kakir@mail.sacredheart.edu) (TM)

**Description of Team Members**

1. **Sambasiva Rao Chennamsetty**

I completed my Bachelor’s in Information Technology. I had 3+ years of experience as a full-stack developer with Java programming as a backend. I like to work with a team with more commitment to work.

1. **Arif Pasha Shaik**

I have completed my Bachelor’s in Information Technology, I have done a couple of internships on Visual Basic .net, and Business Analytics: Data mining and Data warehousing.And I love working in a team that has its full dedication.

1. **Jagadishwar Reddy Velma**

I hold 7+ years of experience in SQL Database Administration. I am here to learn and improve better development skills which help me to become an extensive experienced Core Developer.

1. **Sai Hrithik Peddi**

I am a graduate student at sacred Heart University. I have completed my Undergraduate in Computer Science. After, I worked as an Android Developer at Sensorise Digital services for 6 months. I’m very passionate about my work role.

1. **Vamsi Kiran Kakkera**

I have done my Bachelor's degree in the stream of computer science. I'm having work Experience of 2.5 years in the AWS cloud as an Associate Developer. I've chosen this team as they are very coordinative and discuss everything with the team members.

1. **Kaki Rohit Reddy**

I pursued my Bachelor's in Electronics and Communication Engineering then started working as a .net full stack developer in a reputed organization after that to gain more insight and upgrade my skill set and change my career track I came to the United States to pursue a master's in data science in Sacred Heart University.

**Table of Contents**

[**1 Introduction 6**](#_Toc119725123)

[**1.1 Research Questions 6**](#_Toc119725124)

[**1.2 GitHub Repository 6**](#_Toc119725125)

[**2 Dataset Description 6**](#_Toc119725126)

[**2.1 URL of Dataset 6**](#_Toc119725127)

[**2.2 Dataset Explanation 6**](#_Toc119725128)

[**2.3 Features of Dataset 6**](#_Toc119725129)

[**3 Related Work 7**](#_Toc119725130)

[**3.1 Pro’s 7**](#_Toc119725131)

[**3.2 Con’s 7**](#_Toc119725132)

[**4 Project Plan 7**](#_Toc119725133)

[**5 Data Exploration 8**](#_Toc119725134)

[**5.1.1 Univariate Analysis: 8**](#_Toc119725135)

[**5.2 Bivariate Analysis: 10**](#_Toc119725136)

[**5.2.1 Pearson Correlation: 11**](#_Toc119725137)

[**5.2.2 Correlation Matrix: 12**](#_Toc119725138)

[**5.2.3 Pair Plot: 13**](#_Toc119725139)

[**6 Data Modeling 14**](#_Toc119725140)

[**6.1 Pre-Processing 14**](#_Toc119725141)

[**6.2 Data Splitting 15**](#_Toc119725142)

[**6.3 Data Fitting 18**](#_Toc119725143)

[**6.4 Measuring Performance 19**](#_Toc119725144)

[**6.4.1 Plot ROC curve 19**](#_Toc119725145)

[**6.4.2 Model 1 20**](#_Toc119725146)

[**6.4.3 Model 2 20**](#_Toc119725147)

[**6.4.4 Model 3 21**](#_Toc119725148)

[**6.4.5 Model 4: 21**](#_Toc119725149)

[**7 Residual Analysis 22**](#_Toc119725150)

[**8 Evolution 22**](#_Toc119725151)

[**9 GitHub Repository 25**](#_Toc119725152)

[**10 References 25**](#_Toc119725153)

**List of Figures**

[**Figure 1: Data Visualization 8**](#_Toc119724494)

[**Figure 2: Data Visualization using Univariate Analysis. 9**](#_Toc119724495)

[**Figure 3: Box Plots 10**](#_Toc119724496)

[**Figure 4: Bivariate Analysis 11**](#_Toc119724497)

[**Figure 5: Pearson Correlation 12**](#_Toc119724498)

[**Figure 6: Correlation Matrix - HeatMap 13**](#_Toc119724499)

[**Figure 7: Pair Plot 14**](#_Toc119724500)

[**Figure 8: Data Reading 15**](#_Toc119724501)

[**Figure 9: Info and Columns of Dataset 15**](#_Toc119724502)

[**Figure 10: Data Cleaning and Preparation 16**](#_Toc119724503)

[**Figure 11: Checking Duplicates and dropping 16**](#_Toc119724504)

[**Figure 12: Identifying junk values 17**](#_Toc119724505)

[**Figure 13: Splitting Data – OLS 18**](#_Toc119724506)

[**Figure 14: Measuring performance using ROC Curve 19**](#_Toc119724507)

[**Figure 15: ROC Graph 19**](#_Toc119724508)

[**Figure 16: Model 1 20**](#_Toc119724509)

[**Figure 17: Model 2 20**](#_Toc119724510)

[**Figure 18: Model 3 21**](#_Toc119724511)

[**Figure 19: Model 4 21**](#_Toc119724512)

[**Figure 20: Residual Error 22**](#_Toc119724513)

[**Figure 21: Model Evalution 23**](#_Toc119724514)

[**Figure 22: Adjusted R^2 23**](#_Toc119724515)

[**Figure 23: Final Summary 24**](#_Toc119724516)

# Introduction

As the world evolving in all directions significantly, the economic gaps between the people are still exist. The livelihood of different people from different financial backgrounds are changing a lot. When it comes to the comfortable travel the cars are playing a vital role. Also, considering the COVID pandemic, most of the lower- and middle-income group of people also attracting to travel in a safe environment and not willing to choose public transport.

* At the same time the car manufacturers also increased the price of the new cars, which is directly affecting the buying capability of low-income group people.
* Hence, most of the people are looking at the used cars now.
* There are few people who cannot afford to buy new luxury car, but they wish to travel in it. For those, this used cars are the sunlight in dark. [1]
* This used cars has become an opportunity for the business. And it's going to generate a decent revenue for business as well.

## Research Questions

* Which variables are significant in predicting the price of a used car?
* How well those variables describe the price of a car?

## GitHub Repository

[**https://github.com/samba-chennamsetty/used-car-selling-price-linear-regression**](https://github.com/samba-chennamsetty/used-car-selling-price-linear-regression)

# Dataset Description

## URL of Dataset

[Old Car Selling Price with Linear Regression | Kaggle](https://www.kaggle.com/code/gauravduttakiit/old-car-selling-price-with-linear-regression/data?select=car+data.csv) [2]

## Dataset Explanation

* This dataset contains information about used cars listed on [www.cardekho.com](http://www.cardekho.com) [3]
* This data can be used for a lot of purposes such as price prediction to exemplify the use of linear regression in Machine Learning.

## Features of Dataset

The columns are in the given dataset is as follows:

* + - 1. **Car\_Name:** This column consists of the name of the cars.
      2. **Year:** This column has the year in which the car was bought.
      3. **Selling\_Price:** This column has the price the owner wants to sell the car at.
      4. **Present\_Price:** This is the current ex-showroom price of the car.
      5. **Kms\_Driven:** This is the distance completed by the car in km.
      6. **Fuel\_Type:** Fuel type of the car.
      7. **Seller\_Type:** Defines whether the seller is a dealer or an individual.
      8. **Transmission:** Defines whether the car is manual or automatic.
      9. **Owner:** Defines the number of owners the car previously had.

# Related Work

We are comparing this model with car prediction prices [4] which has multi linear regression with less no of dependencies and models.

## Pro’s

The advantages we have over the other related works are

* Using linear regression model allows us to make our analysis simple.
* Providing a variety of visual representations of impact with each feature.
* It is planned to build multiple models based on type of company.
* Considering the best prediction relational fields.

## Con’s

* Based on our source project referred there is no multiple regression, which is also based on many features of the referred project.

# Project Plan

The project plan has the below steps in it.

1. Data-preprocessing
2. Model building
3. Optimizing Model
4. Model Evaluation

# Data Exploration

### Univariate Analysis:

Univariate analyses are used extensively in quality-of-life research. Univariate analysis is defined as analysis carried out on only one (“uni”) variable (“variate”) to summarize or describe the variable. However, another use of the term “univariate analysis” exists and refers to statistical analyses that involve only one dependent variable and which are used to test hypotheses and draw inferences about populations based on samples, also referred to as univariate.

We find the univariate using distplot and boxplot graphs with below code. Here we’re using only uni one feature for the analysis.

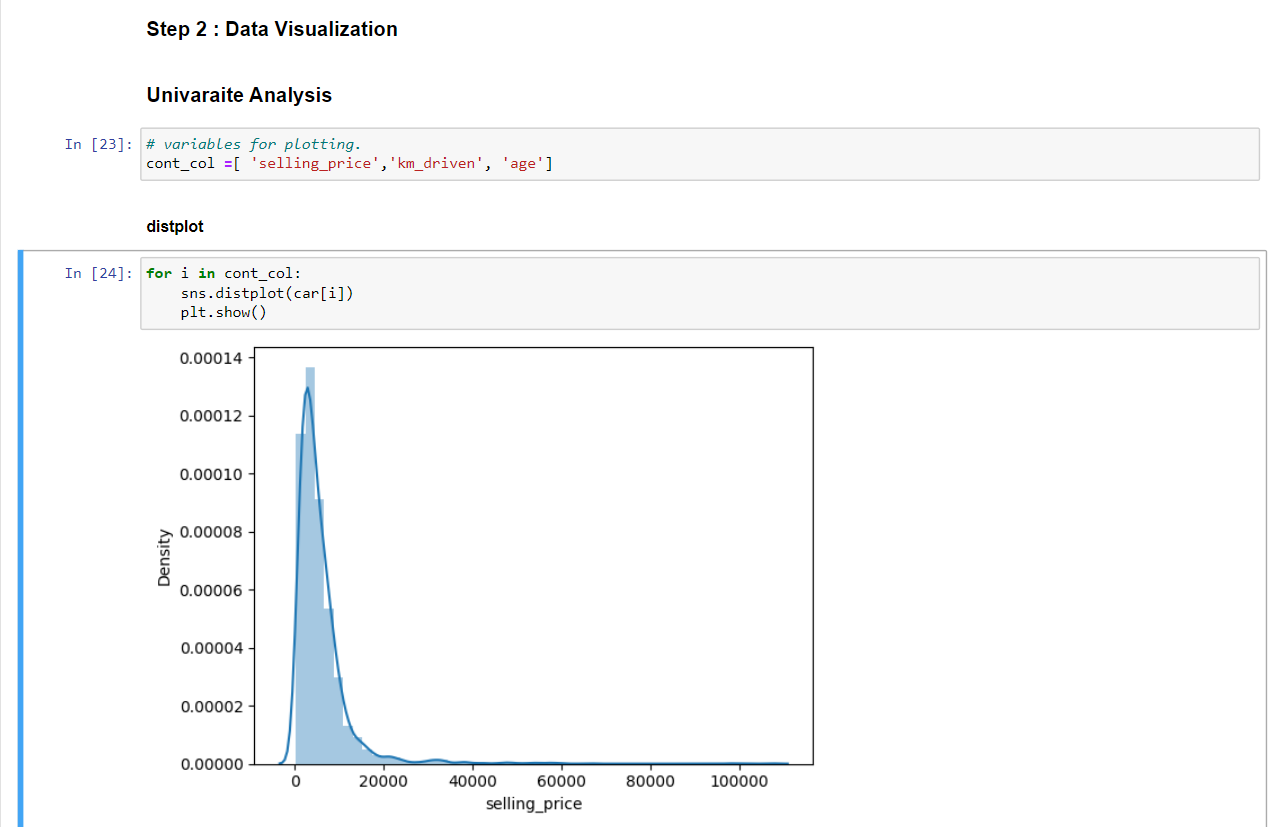
****

Figure 1: Data Visualization

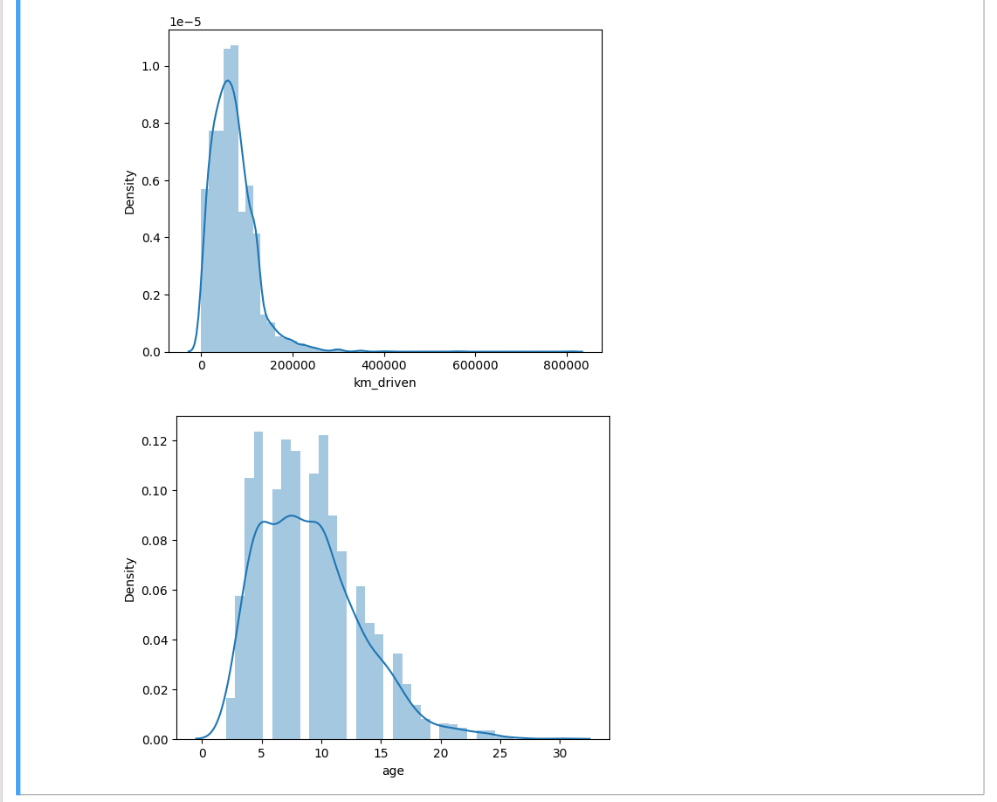
****

Figure 2: Data Visualization using Univariate Analysis.

Using distplot function Univariate Analysis has been made which gives a similar kind of distribution, some features are showing nearby normal distribution while some are skewed.

**Boxplot**

* Boxplot - A graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum.
* In the fig 7 the observation states that the first quartile lies at the lower end of the box and the upper end is the 3rd quartile. The box indicates the range in which the middle 50% of all the data lies.
* The line in between the first quartile and the third quartile lies the median. (Median in the boxplot represents with solid line and the Mean in the boxplot represents with dashed line)

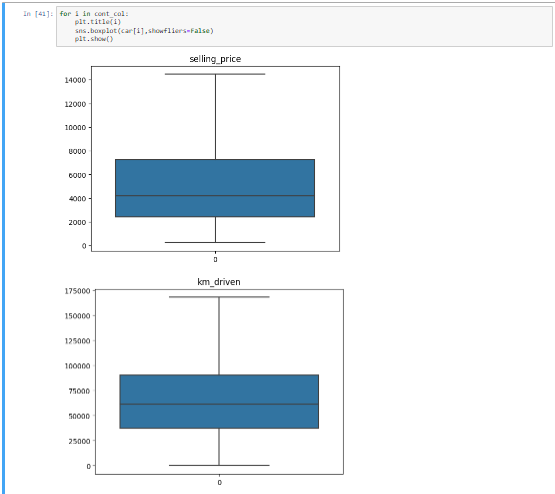


Figure 3: Box Plots

## Bivariate Analysis:

* Bivariate analysis refers to the analysis of two variables to determine relationships between them. Bivariate analyses are often reported in quality-of-life research. For an excellent example of research that utilizes bivariate analyses and demonstrates how the results of bivariate analyses can be used to inform furthermore complex analyses.
* We find the relation between Selling Price and Car age which is bi with scatter plotting.

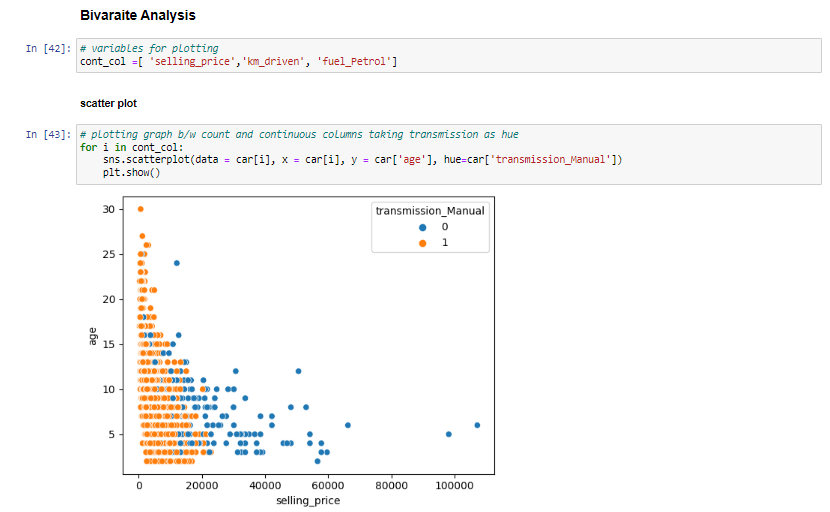


Figure 4: Bivariate Analysis

* Here we can notice that as the age goes up the selling prices decreases.
* And most of the manual transmission cars are under the age 15 and price range of 20000.

### Pearson Correlation:

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

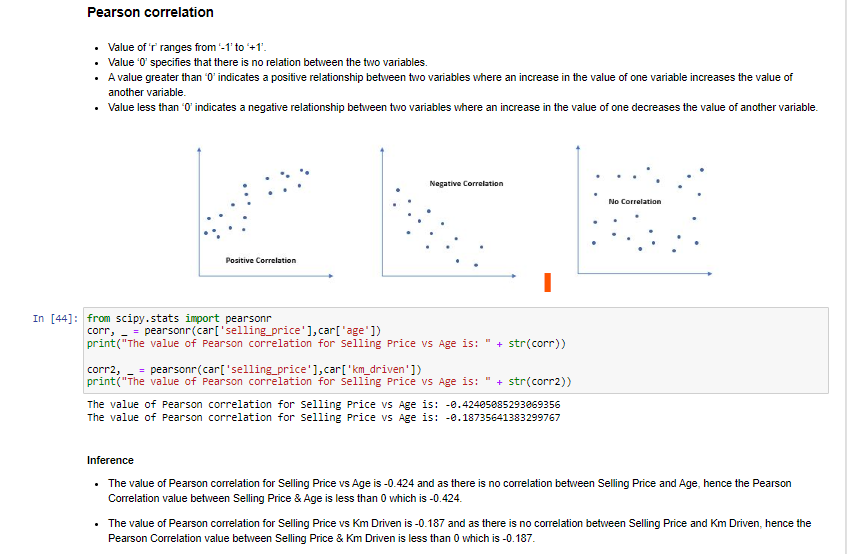
****

Figure 5: Pearson Correlation

### Correlation Matrix:

A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1. Correlation between two random variables or bivariate data does not necessarily imply a causal relationship.

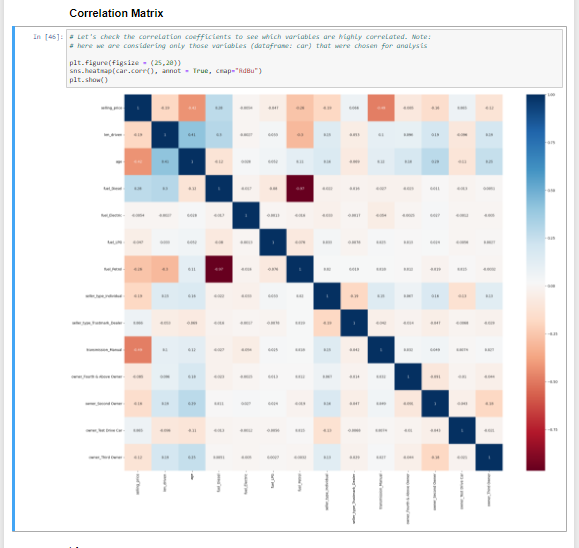
****

Figure 6: Correlation Matrix - HeatMap

### Pair Plot:

Here we took selling price and compare it with km driven and age of car.

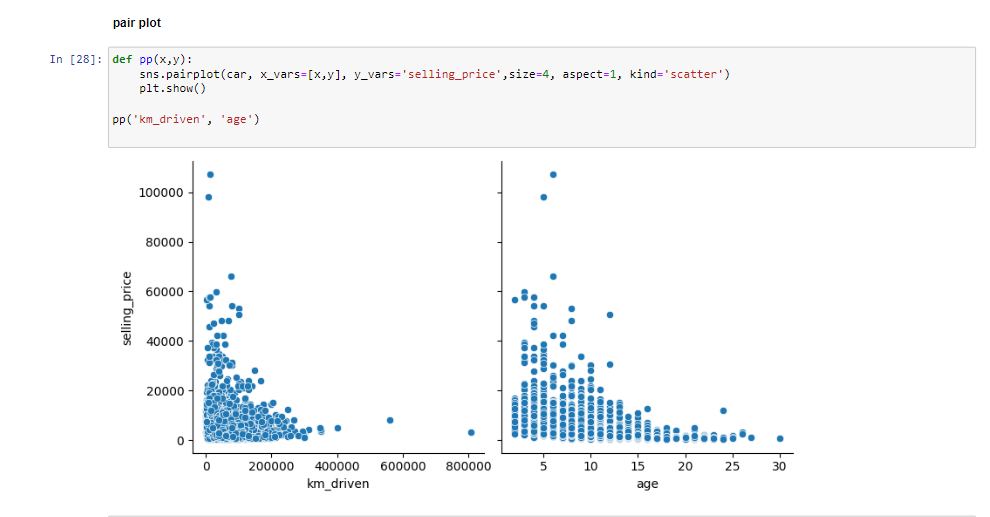


Figure 7: Pair Plot

* Selling price is compared with kms driven and age of the car.
* The pair plot graphs portray the comparison between selling price and kms driven along with age.
* Here, the observation states that the increase in kms driven makes the selling price decrease, vice-versa i.e., data shown in fig9. The car that has driven 800000 kms has a selling price of 0. Whereas highest selling price which is more than 100000 has only driven very less (near to 0 or 10000)

# Data Modeling

## Pre-Processing

* We import the dataset using the function pd.read\_csv(“UsedCarDetails.csv”) and we are looking for the head rows in the dataset.

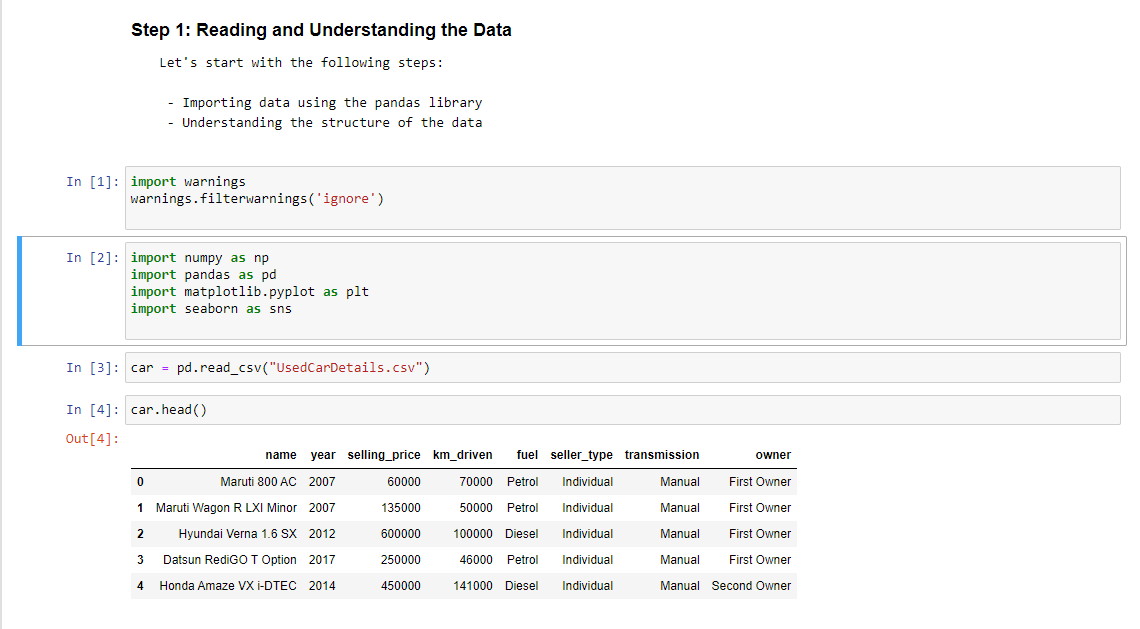


Figure 8: Data Reading

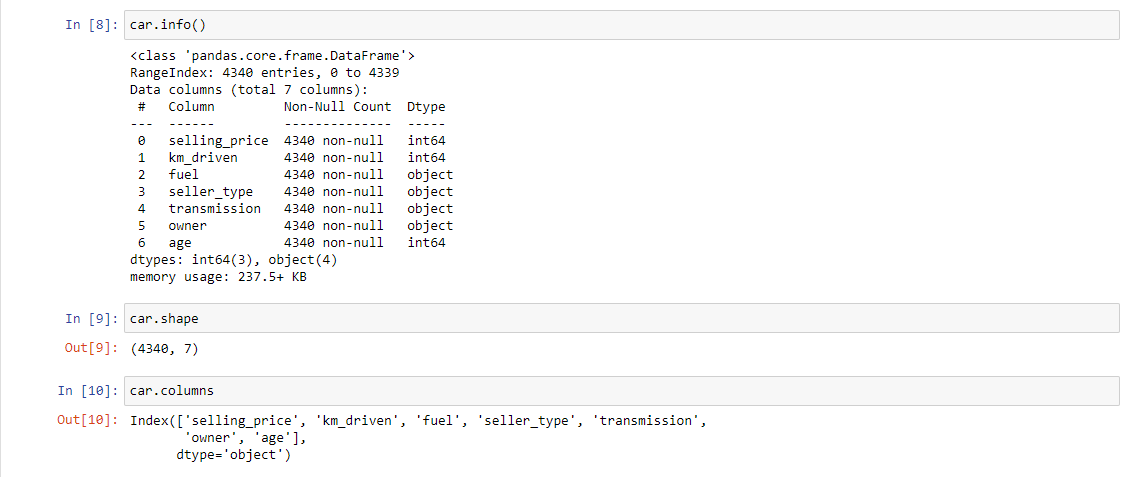


Figure 9: Info and Columns of Dataset

Here we are,

* Using the .shape function, to find the number of rows and columns in the dataset.
* Using .columns function to view the columns in the function.
* Using .info function to know all the details of the car data set with their datatype.

## Data Splitting

* Adding a new variable for calculating the age of the car.
* As part of this we clean the unwanted data and make the data right and good for the model with error free.

A picture containing graphical user interface

Description automatically generated

Figure 10: Data Cleaning and Preparation

**Duplicate Data Check**

Checking if there is any duplicate data and dropping the entire duplicate row if any

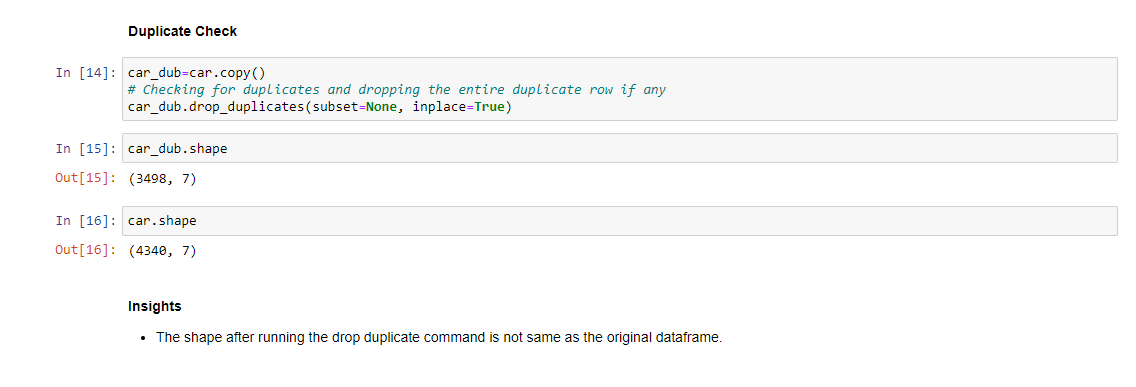


Figure 11: Checking Duplicates and dropping

**Identifying junk values:**

**Graphical user interface, text, application

Description automatically generated**

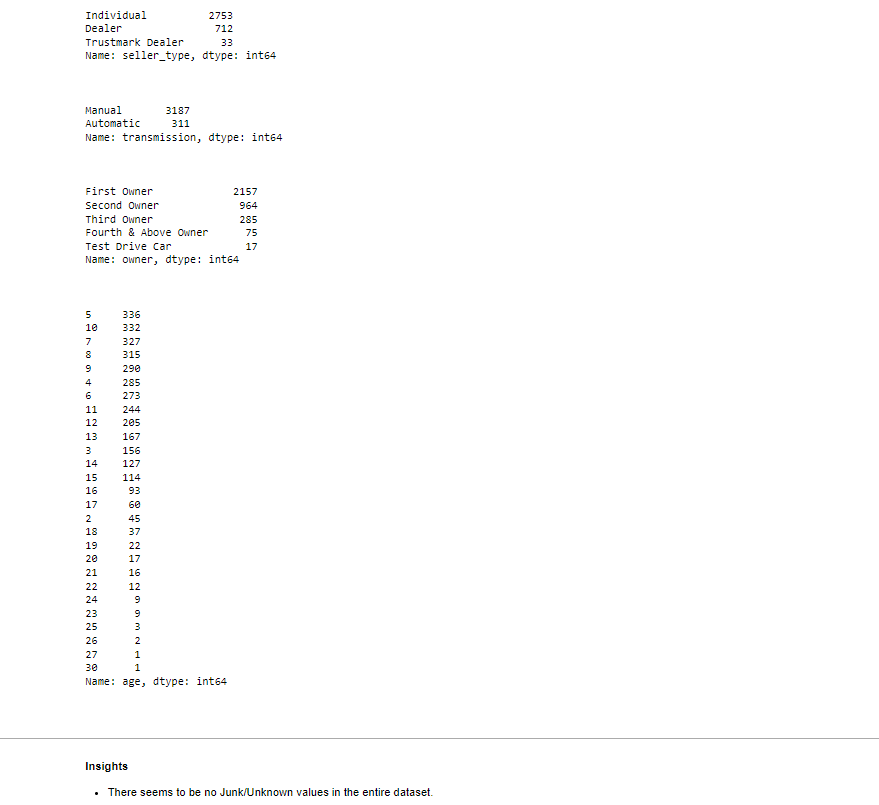
****

Figure 12: Identifying junk values

* **Junk Values** – Data that doesn’t serve any real purpose.
* We found that there is no Junk or Unknown values exists in the data set.
* The above figure states that the .value\_count function gives the count of each column values. For example, 300000 repeats 122 times in the entire dataset and same for the rest.

## Data Fitting

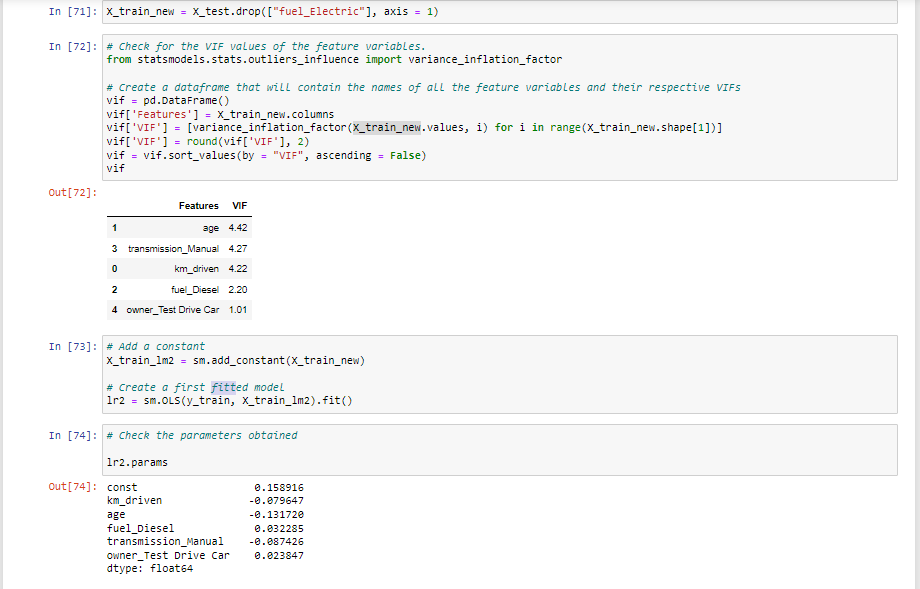


Figure 13: Splitting Data – OLS

## Measuring Performance

### Plot ROC curve

* ROC curves in logistic regression are used for determining the best cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1).

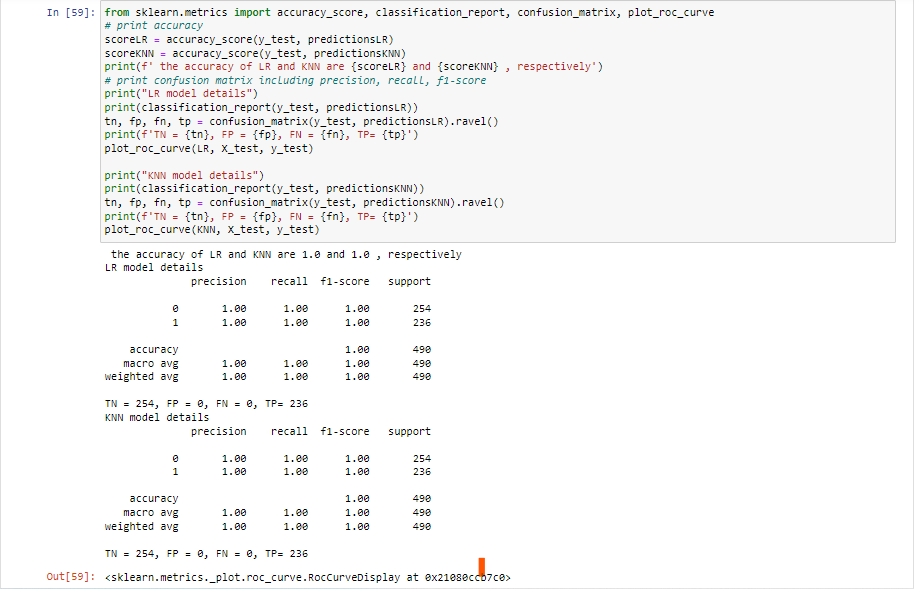


Figure 14: Measuring performance using ROC Curve

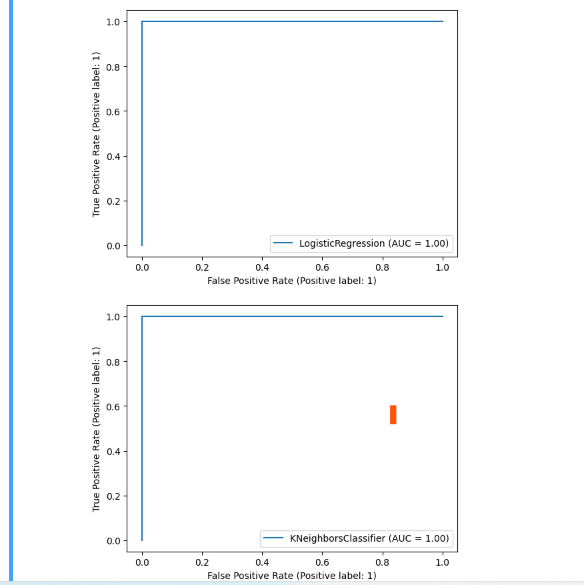


Figure 15: ROC Graph

### Model 1

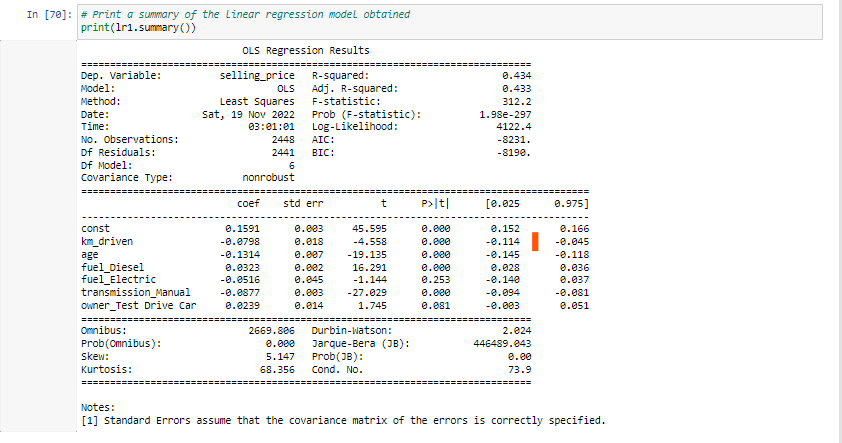


Figure 16: Model 1

### Model 2

* Removing the variable 'fuel\_Electric' based on its High p-value

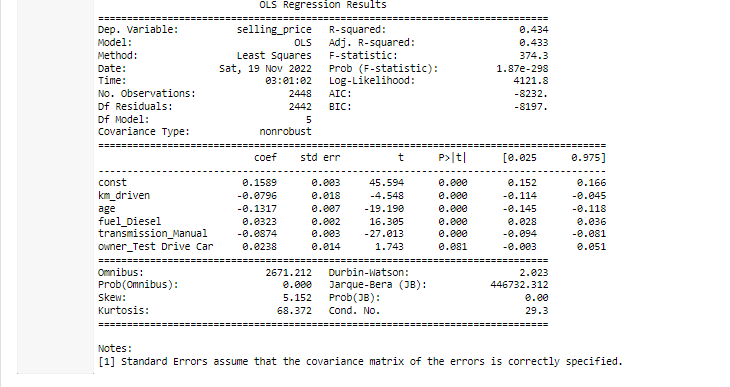


Figure 17: Model 2

### Model 3

* Removing the variable 'owner\_Test Drive Car' based on its High p-value

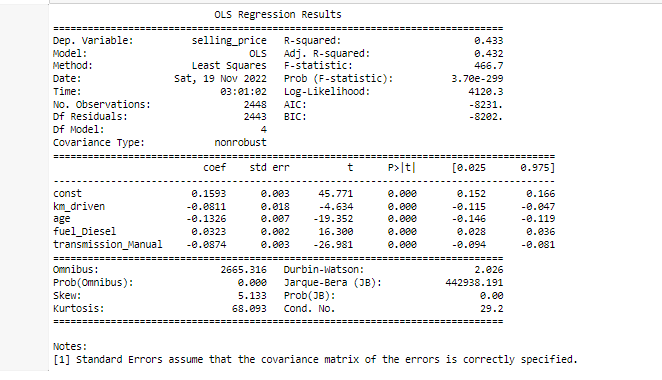


Figure 18: Model 3

### Model 4:

* Removing the variable 'transmission\_Manual' based on its High p-value

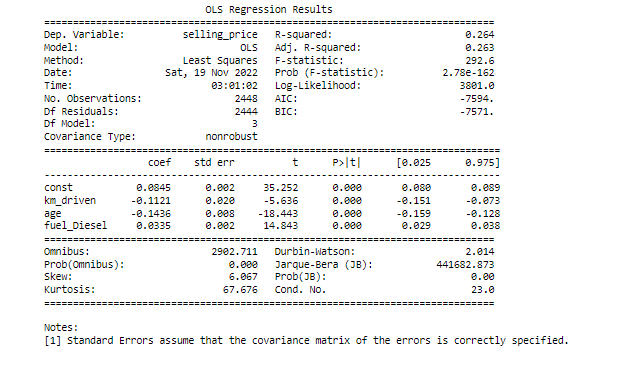


Figure 19: Model 4

# Residual Analysis

We plot the graph to find the error terms of the model w.r.t prediction value of price. The error graph shows the increase in the density and drops down when error is increased. So, at o the density is high and it is distributed normally.

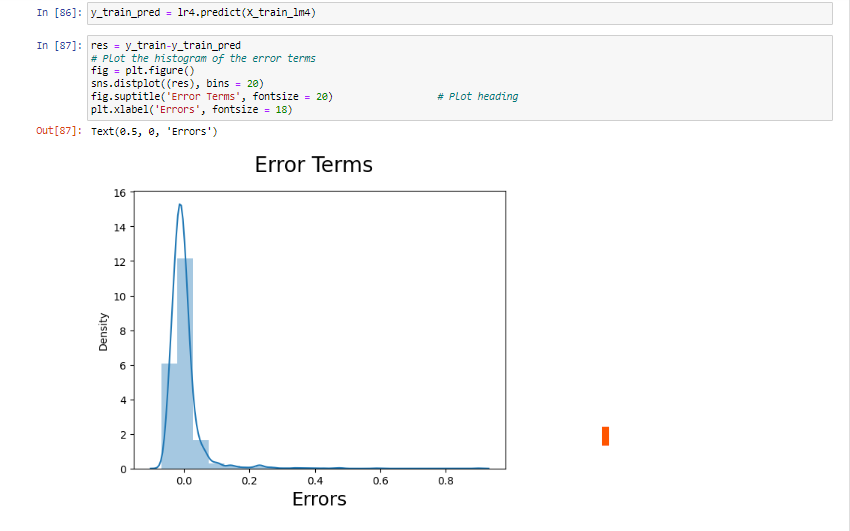


Figure 20: Residual Error

* From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

# Evolution

* Evaluate the actual price and predicted price with the results obtained by plotting the graph with graphical representation.
* We can observe how the actual and predicted prices has variance we can see a few outliers on the top right with high variance.

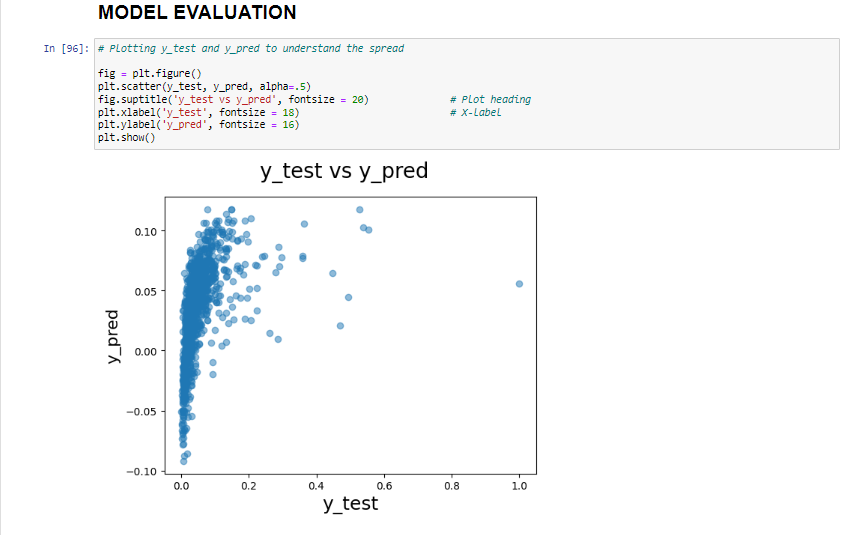


Figure : Model Evalution

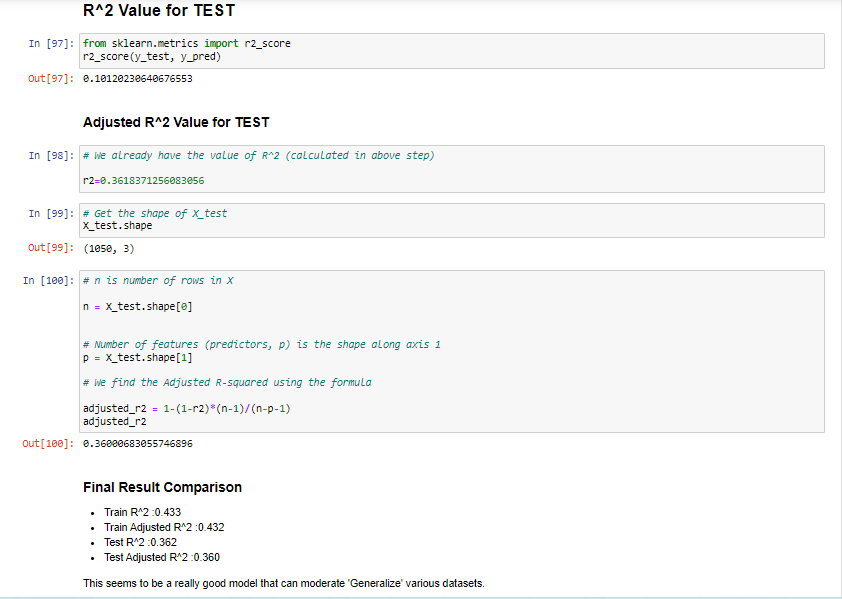


Figure : Adjusted R^2

**Final Result Comparison:**

* Train R^2 :0.433
* Train Adjusted R^2 :0.432
* Test R^2 :0.362
* Test Adjusted R^2 :0.360

As per our final Model, the top predictor variables that influences the selling\_price are:

* km\_driven: A coefficient value of ‘0.081104’ indicated that a unit increase in km\_driven variable, decreases the selling\_price numbers by 0.081104 units.
* age: A coefficient value of ‘-0.132559’ indicated that, a unit increase in age variable, decreases the selling\_price numbers by 0.132559 units.
* fuel\_Diesel: A coefficient value of ‘0.032289’ indicated that w.r.t Petrol, a unit increase in fuel\_Diesel variable increases the selling\_price numbers by 0.032289 units.
* transmission\_Manual: A coefficient value of ‘-0.087353’ indicated that w.r.t Automatic, a unit increase in transmission\_Manual variable decreases the selling\_price numbers by 0.087353 units.

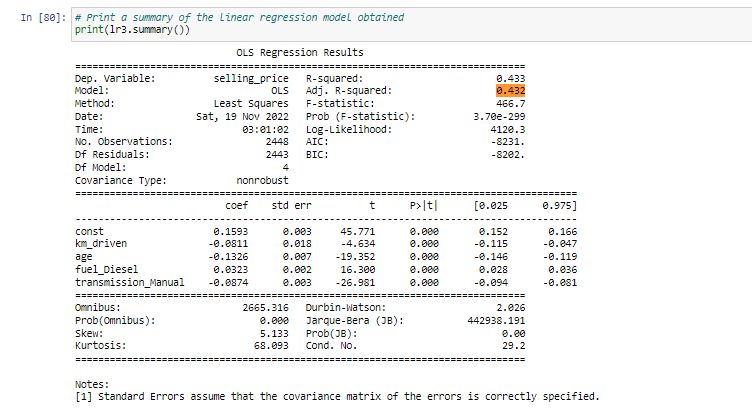


Figure : Final Summary

# GitHub Repository

**<https://github.com/samba-chennamsetty/used-car-selling-price-linear-regression>**

# References

[1]<https://www.kaggle.com/code/gauravduttakiit/old-car-selling-price-with-linear-regression>

[2] <https://www.kaggle.com/code/gauravduttakiit/old-car-selling-price-with-linear-regression/data?select=car+data.csv>

[3] [www.cardekho.com](file:///C:\Users\Arif%20Pasha%20Shaik\Downloads\www.cardekho.com)