WICHITA STATE UNIVERSITY
W. FRANK BARTON SCHOOL OF BUSINESS

INTRODUCTION TO BUSINESS ANALYTICS
BSAN 775, CRN: 17395
Fall-2022

Group -02

Time: 14.00-3.15 PM (Tuesday & Thursday)

Eswar Hemant Majeti- G233R756

Analysis of Pre-Owned Car market for Car Traders using Regression

Table of contents

1. ***Abstract:***

The purpose of this study was to research how the selling price of the can be estimated considering various attributes that are highly correlated with the selling price of the car. We use the multiple regression to predict the selling price of the used car. Null, redundant, and missing values were removed from the dataset during pre-processing. Three cases of multiple regression is performed out of which the best fit model is identified for predicting the selling price of the used cars. The research for this project anticipates that in the near future, the most sophisticated algorithm is used for making predictions.

2. ***Introduction:***

Determining if the quoted price is accurate of the used cars is a difficult task, as various elements that influence a used car's market pricing. With the increase in the car production and the price of new cars, the demand for the pre-owned cars has increased. So companies need to take care of few things in perspective of the customer to drive their business of selling the pre owned cars successfully with good profits. Companies always need to keep a track of their sales and check if they are able to get more customers.

The process which comprises of processes, tools and techniques of data analysis and management that includes collection, storage and organizing the data by which we can analyze the raw data and generate actionable insights is known as Data Analytics. In order to improve the performance and gain the competitive advantage many companies perform data analytics. Data analytics is performed on a variety of big data sets, like transactions, server logs, electronic health records, click streams, insurance claims, etc.

The data analytical techniques are:

- **Descriptive analytics** – In order to understand some past events and performance summarizing the data is required
- **Diagnostic analytics** – Identify the root cause of some certain events
- **Predictive analytics** – In order to plan in the future Predicting the future is required
- **Prescriptive analytics** - Advising on the best course of action and considering the optimal outcomes

Here, we will be using the descriptive statistics in order to generate insights from the data.

People need to take care of lot of things while selecting a used car. With the emergence of internet and mobile phones, there is a very good chance to find good deals on the cars based on customers' requirements.

Companies like Car Trade are car dealers having a good online platform to search pre-owned cars that help customers buy cars as per their requirements.

Here, a research is conducted in the area of Pre-owned car selling market. An observation is made on how Car Trade Company is handling their pre owned car sales market and check if the selling price set by the company is appropriate or not. As the selling price of the car is key factor to target the customers in the market. The goal of this research is to create machine learning models that can precisely predict a used car's price based on its attributes so that buyers can make educated decisions. Here we apply linear regression on selling price of pre-owned or used cars. As Linear regression is a statistical analysis to determine the character and strength of the association between a selling price of the car and other independent variables. It also forms a relationship of linear equations explicitly. The estimated linear regression equation is done by the

least squares method, but it is not practical to do this by hand. This will help to find the car sales accuracy of the company and also to check how the independent variables affect the selling price and accuracy of sales.

Keywords: Linear Regression, Multiple Linear Regression, Ridge Regression and Lasso Regression, Used car pricing.

Software Requirements: IBM SPSS Statistics 26, Microsoft Excel

3. Problem Statement:

It can be challenging to decide whether a used car is worth the asking price while viewing listings online.

The actual value of an automobile might vary depending on a number of factors, such as mileage, make, model, year, etc. Pricing a used car fairly is a challenge from the seller's point of view.

Car Trade Company has come up with a Pre-owned car sales platform. It wants to know the opportunity it has in this market against other companies and new car sales. It also wants to know what strategies can be applied in order to attract the customer more affectively.

The dealers in the pre-owned car market it is important to learn about the market and Target Customers

- To know how Car Trade company can handle the 'pre-Owned' market?
- To know how to attract the customers more affectively.
- To know how to advertise their market in a better way.

The Car Trade Company decides the selling price based on various factors like car model, engine, Car mileage etc. So, in order to attract more customers, the company needs to set a reasonable selling price, as the prospective customer's ability to purchase a new car is not too strong, so they decide to used cars as an option. The car trade company wants to know the accuracy of their car pricing and also check if their pricing is correct or not based on considering the different factors. It is difficult to calculate manually the selling price of each car. With data, the goal is to create models for forecasting used car prices using machine learning algorithms. So a linear regression is used to estimate the equation of the dependent variable i.e selling price of the used car based on different independent variables. Using regression model we can also estimate that whether the model developed is a good fit for analyzing the used car selling price or not.

4. Objectives:

To know and analyze the effectiveness of pre-owned car sales in the market by Car trade company.

To estimate the equation of selling prince of used cars and check if the selling price is accurate, also to check the estimated linear regression model is a good fit or not.

The selling price of different used cars depends on various factors, and it changes with the changes in those variables.

Sub Objectives:

- To know the nature of used car sales market.
- To enumerate the various used cars in the market.

- To enumerate the various reasons for customers opting a used car.
- To know the impact of promotional activities on the sale of used car.
- To know the accuracy of the selling price set for used cars
- To know the factors affecting the selling price of used cars.
- To know the impact of selling price of used cars in the market

Here, we use the regression model to check if the estimated regression model for the sales of the used cars is a good fit or not to achieve the mentioned objectives.

5. Preliminary Literature Review:

As we know several studies related to works have been conducted in the past to forecast used cars, and price prediction using different methodologies and approaches with an accuracy of 20% to 90%. Several methodologies are used to determine the accuracy of the prices of used cars. It is important to develop strategies according to customer preferences, with the help of data it is easy to determine the mitigations and future disruptions to overcome in today's life.

Every one dream is to own a car in their life, Due to many drastic changes and many mitigations that have been occurring over many years the prices of new cars have been raised due to many changes so there another option to buy a car is the Reused car, so it is important to know the price of the Cars, so it is easy to choose the cars in the market. Many researchers have done Price prediction using different methodologies and different software.

The topic we have chosen and gathered data from the Kaggle "Vehicle data set "(Price prediction on used cars) using regression. There are many methodologies used for the prediction of the price, so using linear regression we can get to know the accuracy and perfect price of the cars in

the market, which is the most suitable model to buy, so we decided to run the Regression. The Regression model helps us to find out the relationship between the variables. As we are predicting the data using the different attributes which are provided in the data set. So, using the SPSS software can give us the perfect analysis on the data and the outcome of the result is perfect. So now we have decided to use the SPSS software to predict the on-time price output. It can enhance business sales, as well as trust in the companies, will be increased in society. The data set provides the attributes of Car models, Kilometers driven, year of manufacture, owners, transmission, and dealer.

5.1. Literature Review:

Prediction of prices of Secondhand Cars (Ozer Celik and U Omer Osmanogula): The article provides information about the prediction of car values using machine learning techniques and predicting the R2 value how it is important and makes talks about the importance of second-hand markets in today's world and strategies to develop.

https://dergipark.org.tr/tr/download/article-file/728216

Used car's price prediction and valuation using Data mining Techniques (Abdulla Arshareed) Rochester institute of technology, Dubai: The article briefly talks about the price prediction and valuation of used cars in the market using data mining Techniques. The best price strategies and better decisions in the business to get the perfect outcome of the output of this are explained and classified and explained the comparing different attributes.

https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses

Predicting used cars Prices (Kshitij Kumbar, Pranav Gadre, Varun Nayak): The article talks about the price prediction of the used cars in the USA which have different types of attributes Random forest model is a machine learning technique that can predict the data accurately to find out the linear regression.

http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf

End to End price prediction on used cars using the Python Tool: The article describes the python techniques used to predict the price of the cars in the present market with different variables which it is interpreted in the python model.

https://medium.datadriveninvestor.com/end-to-end-project-on-used-car-price-prediction-3dc412d24aa0

6. Methodology:

The data set, the method of pre-processing the data and the methodology used to create and identify the most accurate model are briefly discussed in the sections that follow.

We used the data from the Kaggle site, which sells old vehicles, to construct the flexible regression model. People mostly take the selling price into account.

Here, we consider all other factors as independent and the selling price as the dependent variable.

Using a linear regression model in SPSS, we were able to determine the relationship between the selling price of a used car and all other variables.

The Data set is obtained from Kaggle:

Reference link:

https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv

6.1. Flow chart of the methodology is shown below:

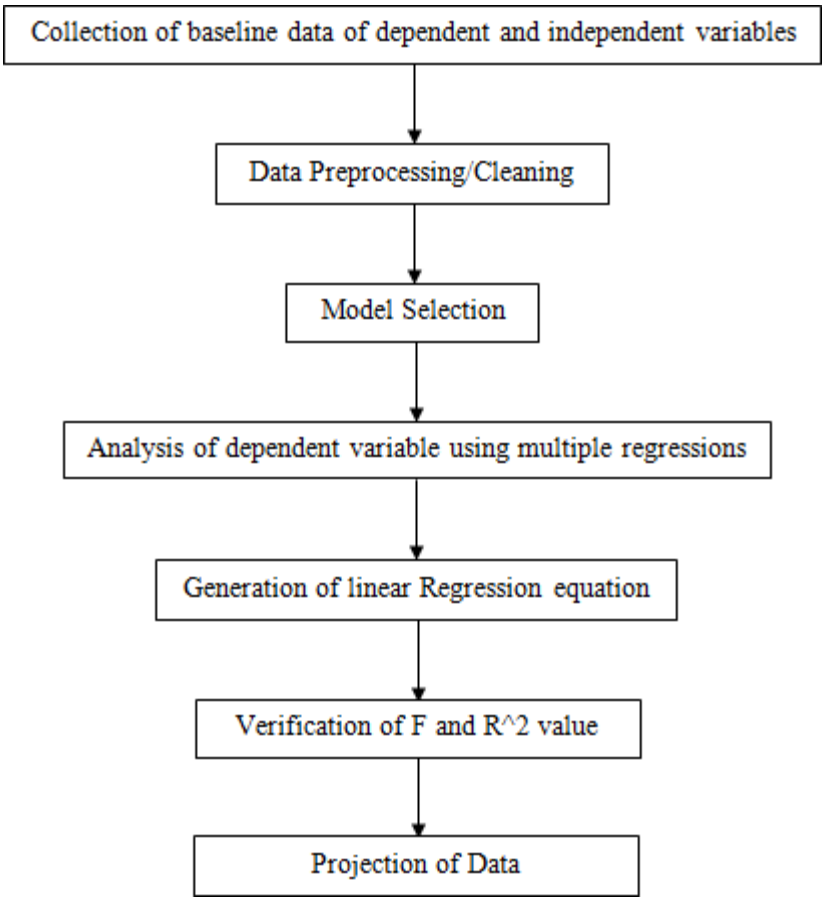

Figure1: Flow chart of the methodology

Step 1: In this step we collect the required data or access the collected dataset from various trading companies and surveys from the customers regarding pre-owned cars. Here, an appropriate algorithm is selected for prediction of dependent variable based on business understanding. In this study a public Dataset was accessed from Kaggle dataset platform. This

project aims to determine the Selling price of used cars based on various factors. The main goal is to create a machine learning model for prediction of accurate selling price of a used car.

Step 2: In this Step, data preparation entails comprehending the data and being prepared to segment, develop features, manipulate data, reduce dimensionality, and alter it.

Step 3: We need to select an appropriate model for getting accurate result. So, we use a multiple regression model for prediction of selling price of the car.

Step 4: We run multiple regressions in this step by selecting various dependent and independent variables in the dataset in order to obtain an accurate and good fit model for predicting the selling price of a car.

Step 5: We determine a good fit model for predicting the selling price of car by selecting various dependent and independent variables from the data set and generate the equations using multiple regressions.

Step 6: This step involves checking and verifying the R squared value. The R square value aids in calculating the proportion of the response variable's variance that can be explained by the regression model.

Step 7: Data projection is the final step. Here, we compile the data from all the models and produce the results report. The report aids management in creating a strategy to address it by detailing all the processes and results in a step-by-step fashion.

7. Data Description:

The Data set is obtained from Kaggle:

Reference link:

https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv

The data set consists of sales data of all the sold cars during 1983 to 2020. To expand the company's business, gain and retain customers we are going to analyze this dataset and help them to stand out the competitions they face

The data set has 8128 data points with 13 features in it related to :

- **Car Details -** *Car name, transmission, fuel type, number of seats, year of manufacturing.*
- **Engine Details -** *Mileage, Engine type, Torque, Maximum power in BHP.*
- **Sale Details -** *Selling price, kilometers driven by the car.*

To understand the patterns in the dataset and achieve the high quality, consistent results we will implement learning algorithms and modeling techniques.

8. Data Cleaning and preprocessing:

Here, in data cleaning process we identify incomplete, inaccurate, incomplete, or irrelevant parts of data and then replace, modify or delete the unwanted data.

Remove unit measurements from the 'Mileage' column:

The attribute "Mileage" consists of categorical values suffixed with unit measurements of the fuel efficiency and these units need to be removed in the process of cleaning data (like Kmpl, Km/Kg, etc) and convert the data type to numeric for further utilization in analysis.

Remove unit measurements from the 'Engine' column:

The attribute "Engine" consists of categorical values suffixed with unit measurements of the engine capacity and these units need to be removed in the process of cleaning data (like CC) and convert the data type to numeric for further utilization in analysis.

Remove unit measurements from the 'Max Power' column:

The attribute 'Max_power' contains nominal categorical values suffixed with the units of the Break Horsepower (BHP). We need to change this variable from categorical to numerical by removing the measurement units (like bhp) from the data. This step needs to be performed in order to utilize the data in further analysis.

Dropping the 'Torque' column:

The attribute 'Torque' contains nominal categorical values having two values suffixed with imperial measurements of Newton meters along with 'Revolutions per minute' (like 190Nm@ 2000rpm) and in other cases we have been given Kgm data (like 12.7@ 2,700(kgm@ rpm) and 22.4 kgm at 1750-2750rpm) instead of Nm. Hence removing these measurements is quite a

complex as data is given in different forms in each cell. So, here we drop the column Torque and do not utilize the data in further analysis.

Identify blank and null values in the data:

We checked if the data set has any NA and blank values in all of its features and dropped 222 rows having NA and blank values.

Remove unnecessary columns in the dataset:

We removed the unnecessary columns in the data set because either informative values were extracted from the dataset columns or were of no use.

Creation of Dummy variables:

We created 11 dummy variables columns so that categorical variables are coded numerically. We changed the trustmark dealer to dealer so that we can have only two categories of seller type

Fuel: Petrol, Diesel, CNG and LPG.

seller_type: Seller (It has 2 categories Individual and dealer)

transmission: Trans (It has 2 categories Manual and Automatic)

Owner: First owner, Second owner, Third owner, Fourth & Above Owner, Test Drive Car.

9. Model Selection:

1) Simple Linear Regression: In simple linear regression model, a simple linear relationship is drawn between the dependent and the independent variables

$Y = a + bX + e$

(a is intercept and b is slope and e is error)

2) Multiple Linear Regression: Like the linear regression model, in multiple regression model based on the correlation strength between the dependent and independent variables statistical analysis is made. Here, in multiple regression model we consider only one dependent variable and many independent variables to predict a perfect model for the analysis. The value of the dependent variable (Y) is now determined based on the values of the predictor variables.

$Y = a+ b1X1 + b2X2 + b3X3 + ... + bnXn +e$

(b1, b2, b3,. , bn > are the regression coefficients)

3) Polynomial Regression: Polynomial regression is a kind of regression model where a curve is fitted against the dependent and independent variables. Unlike linear regression this regression model does not have a straight line instead it consists of a curved line. It is done by establishing a curvilinear link between the variables using a polynomial equation of degree 'n' that is fitted to the non- linear data.

$Y= a+ b1X^1 + b2X^2 +.....+ bn X^ n +e$

4) Ridge Regression: Ridge regression model is a common tuning procedure to analyze data with multicollinearity. This strategy is to approximate the regression model's coefficients when

the data suffers from multicollinearity. The basic goal of Ridge Regression is to consider the dataset and fit a new line into it without over fitting the model.

5) Lasso Regression: Least Absolute Shrinkage and Selection Operator (LASSO) is a type of regression that penalizes the regression model, comparable to ridge regression. In this model we use L1 regression. This model has higher prediction accuracy when compared to the other models mentioned above. In Lasso regression model there will be a reduction in the variance of the model by yielding a line with small amount of bias added to it.

From the above-mentioned models, we use the multiple regression model and predict the selling price of a car.

10. Data Analysis:

The Multiple Linear regression analysis is used to predict the value of the selling price variable based on other variable values.

The generalized multiple regression equation will be as follows:

$$Y = a_1 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + ... + b_nX_n$$

Where,

a_1 is a constant

X_i is the variables and i=1,2,3….n

b_i is the coefficient of slope and i=1,2,3…..n

Collectively the constant and the coefficient together are called regression coefficients

When X increases by a unit and other Xs (other variables) remain constant then each slope coefficient is the expected change in Y.

Our data set consist of 8128 entries with one dependent variable and independent variables. The Year ranges from 1983 to 2020. Where the Categorical variables are seller type, Transmission and fuel type, owner, number of seats, while the remaining variables are independent variables. Later we have created dummy variables.

The explanatory variables in the dataset are categorical and cannot be measured on a quantitative scale. All these categorical variables need to be included in the regression equation as they are often related to the dependent variables.

In order to include all the categorical variables, we include the dummy variables.

A dummy variable has the value of 1 or 0. If the value is equals to 1 then the observation is in a particular category else if we get 0 it is not an observation of that particular variable.

We check for the R value. When the R value is near to one then the model is a good fit. Here the value of R is close to one, this estimated multiple linear regression model for selling price of house, size of the house, age number of rooms and indication of whether the house has an attached garage is a good fit.

We perform the Validation of Fit to check whether the equation obtained is a good fit and will predict well for the new data.

Firstly, Data exploration is required to from the correlation and scatter plots, then we have identified the relation between the attributes which will be useful for the prediction of the selling price of the car.

After observing the relation between the attributes multiple linear regression is performed on the selective attributes to obtain a good fit model for the prediction of the prices.

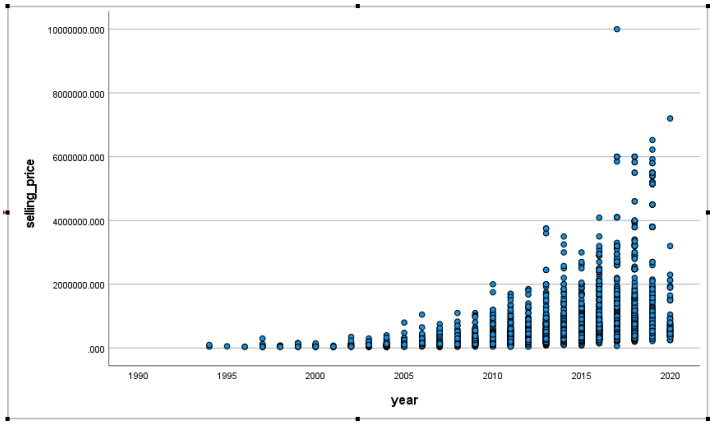

Figure2: Scatter plot for Year and Selling price

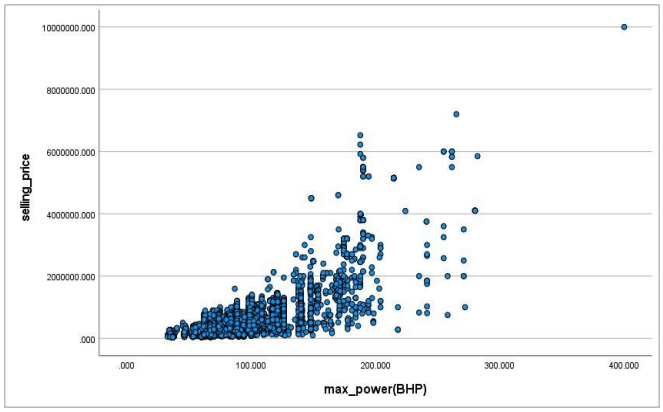

Figure3: Scatter plot for max_power and Selling price

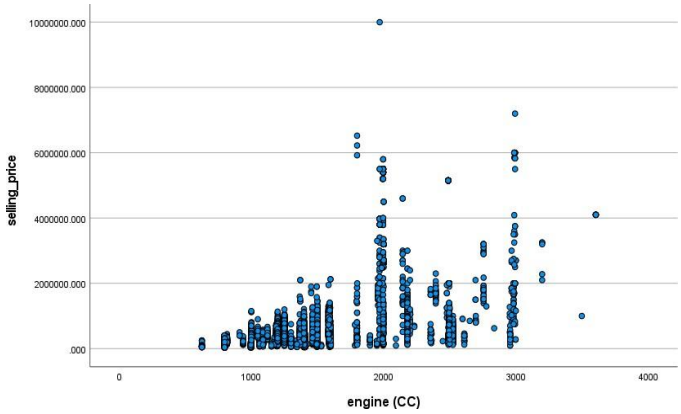

Figure4: Scatter plot for engine and Selling price

From the above scatter plots Engine vs Selling price, Max_power vs Selling price and years vs selling price, we can observe that with a gradual increase of years, max_power and engine capacity the selling price of the car is also increasing. So, we can conclude that the Selling price of the car increases with the increase in years, max_power of car and engine capacity.

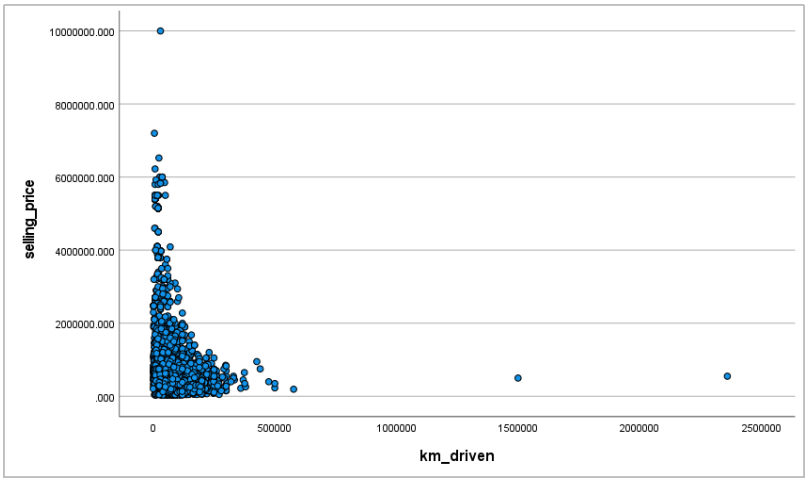

Figure5: Scatter plot for Km_driven and Selling price

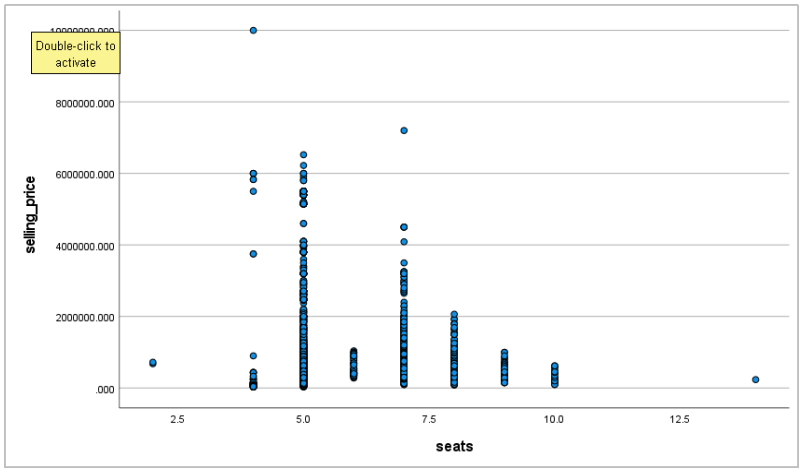

Figure6: Scatter plot for seats and Selling price

From the above scatter plots Number of seats vs Selling price, KM_Driven vs Selling price, we can observe that with the increase in the number of seats in the car and the km driven by the car

the selling price of the decreases gradually. So, we can conclude that the selling price of car decreases with increase in number of seats in car and the km driven by the car.

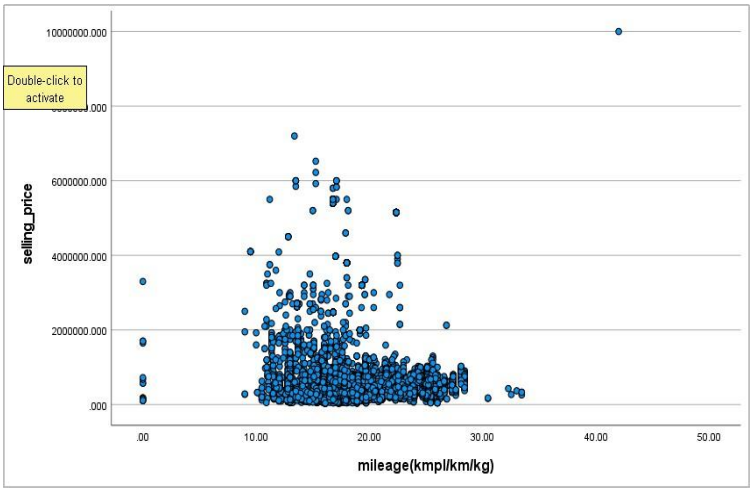

Figure7: Scatter plot for mileage and Selling price

From the above scatter plot Mileage vs Selling price, we can observe that the selling price of the car is high when the car gives a mileage between 15-25 km, whereas with the increase of the mileage of the car the selling price of the car also decreases. Hence, we can conclude that the selling price of the car decreases with the increase in the mileage of the pre-owned cars.

Table1: Case Processing Summary for Fuel and selling price

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	fuel	N	Percent	N	Percent	N	Percent
selling_price	CNG	52	100.0%	0	0.0%	52	100.0%
	Diesel	4299	100.0%	0	0.0%	4299	100.0%
	LPG	35	100.0%	0	0.0%	35	100.0%
	Petrol	3520	100.0%	0	0.0%	3520	100.0%

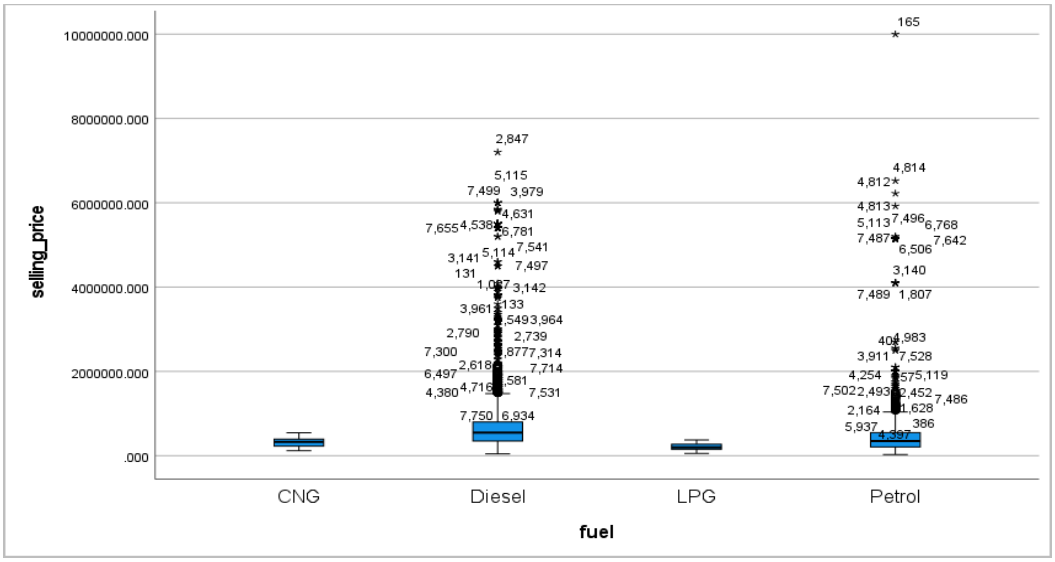

Figure8: Box plot for fuel and Selling price

From the above box plot which is plotted for Fuel type of the car and the selling price, it can be observed that the selling price of the car with CNG and LPG are almost same, and the selling price of the petrol type car is bit higher than Gas type cars. Here Diesel type cars have high selling price when compared to other types of cars. So, we can conclude that the selling price of the diesel cars are more when compared to petrol, CNG and LPG type cars.

Table2: Case Processing Summary for Seller type and selling price

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	seller_type	N	Percent	N	Percent	N	Percent
selling_price	Dealer	1343	100.0%	0	0.0%	1343	100.0%
	Individu	6563	100.0%	0	0.0%	6563	100.0%

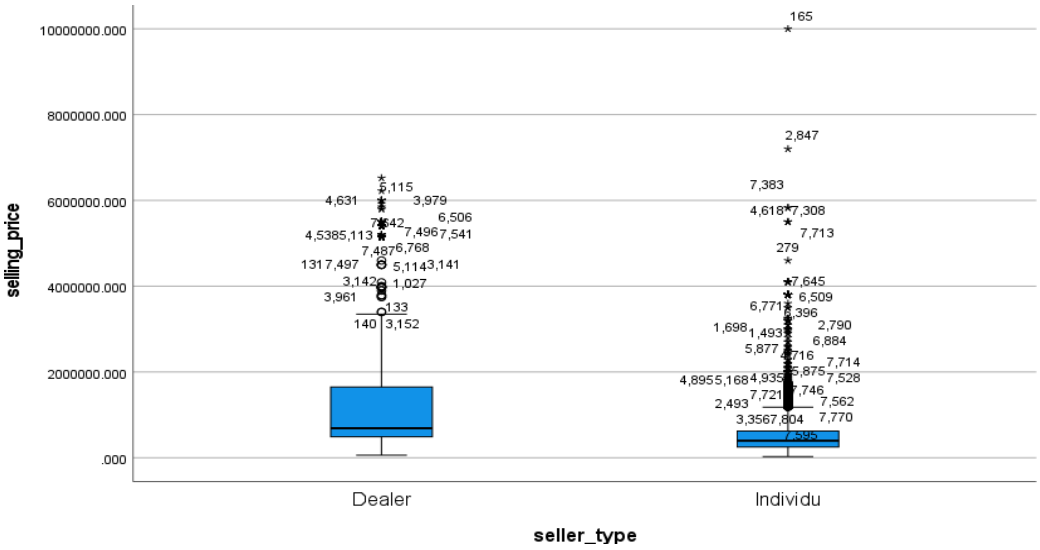

Figure9: Box plot for Seller_type and Selling price

From the above box plot which is plotted for Seller type and the selling price, it can be observed that the selling price of the car which is sold by the Dealer is more when compared to the cars sold by the individual. Hence, we can conclude that the selling price of the car is high when dealer is selling the car when compared to the car sold by the individual.

Table3: Case Processing Summary for transmission and selling price

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	transmission	N	Percent	N	Percent	N	Percent
selling_price	Automati	1041	100.0%	0	0.0%	1041	100.0%
	Manual	6865	100.0%	0	0.0%	6865	100.0%

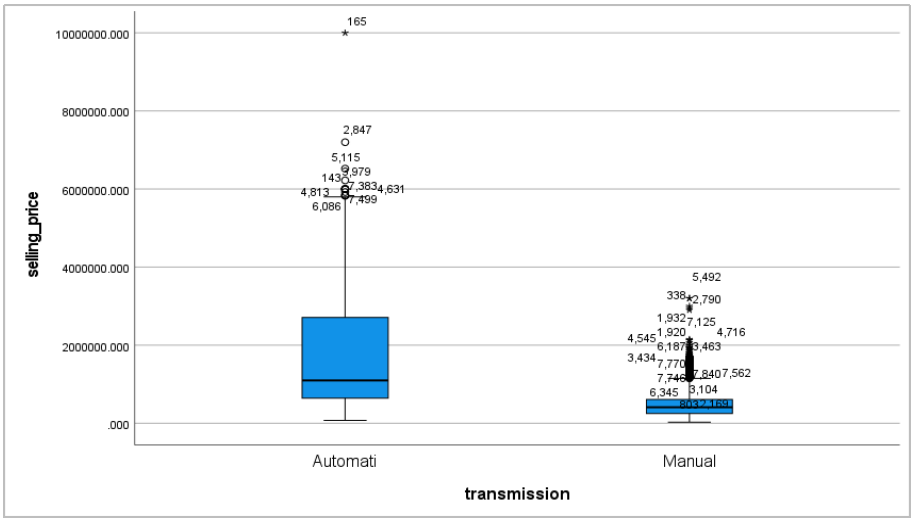

Figure10: Box plot for transmission and Selling price

From the above box plot which is plotted for transmission type and the selling price, it can be observed that the selling price of the car which is Automatic is more when compared to the cars with manual transmission. Hence, we can conclude that the selling price of the car is high when the car has a automatic transmission system when compared to the car with manual transmission system.

Table4: Case Processing Summary for owner and selling price

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	owner	N	Percent	N	Percent	N	Percent
selling_price	First Ow	5215	100.0%	0	0.0%	5215	100.0%
	Fourth &	160	100.0%	0	0.0%	160	100.0%
	Second O	2016	100.0%	0	0.0%	2016	100.0%
	Test Dri	5	100.0%	0	0.0%	5	100.0%
	Third Ow	510	100.0%	0	0.0%	510	100.0%

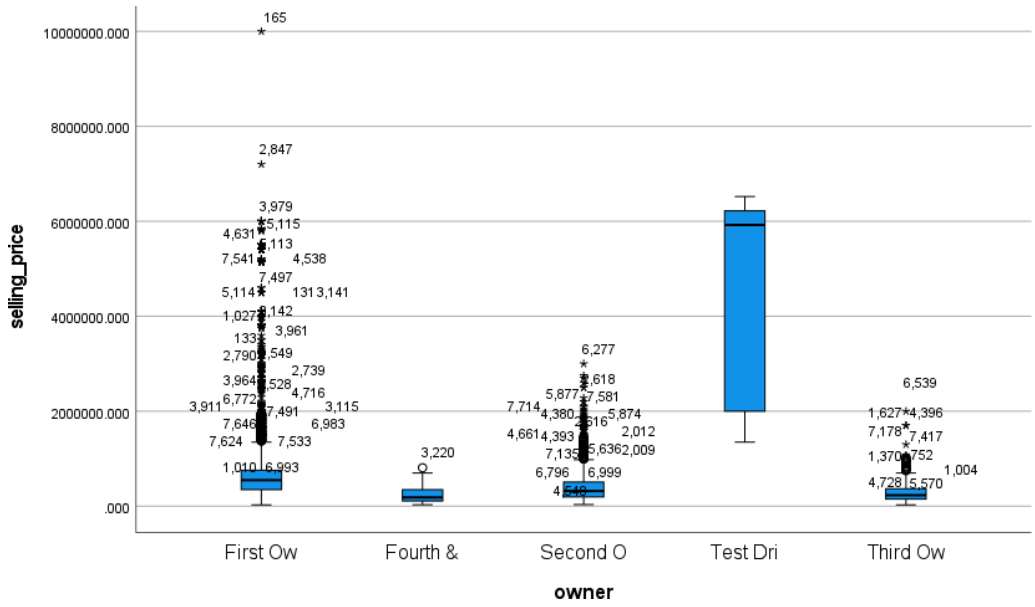

Figure11: Box plot for owner and Selling price

From the above box plot which is plotted for owner of the car and the selling price, it can be observed that the selling price of the car is more for the test drive cars. As the owner changes from first to the second and so on the selling price of the car is also decreasing gradually. Hence, we can conclude that the selling price of the car decreases with the increase in the number of owners for the car and the selling price of the test drive car is the highest.

Table5: Number of outliners for the attributes

selling_price	600
km_driven	168
mileage	18
engine	1186
max_power	573
torque	405
rpm	0
age	78

Note: RPM= Rotation per minute, KM= Kilometers

The count of the number of outliers for each attribute is given in the table above. The count of outliers for selling price attribute is 600 and the engine is more than 1000.

Table6: Descriptive Statistics for the attributes

Descriptive Statistics

	Mean	Std. Deviation	N
selling_price	649813.7208	813582.7484	7906
year	2013.98	3.864	7906
km_driven	69188.66	56792.296	7906
mileage(kmpl/km/kg)	19.4199	4.03626	7906
engine (CC)	1458.71	503.893	7906
max_power(BHP)	91.58737	35.747216	7906
seats	5.42	.959	7906
Petrol	.45	.497	7906
Diesel	.54	.498	7906
CNG	.01	.081	7906
LPG	.00	.066	7906
Seller	.83	.376	7906
Trans	.87	.338	7906
First owner	.66	.474	7906
Second owner	.25	.436	7906
Third owner	.06	.246	7906
Fourth & Above Owner	.02	.141	7906
Test Drive Car	.00	.025	7906

Note: LPG= Liquified petroleum gas, CNG=Compressed Natural gas, N= Number of data count used, CC= Cubic centimeter

From the above description statistics, we can observe the mean, standard deviation and the number of data count used in the dataset for various variables used in the dataset like Selling_price, year, seats etc. These statistics are obtained for the entire data used for the prediction of the selling price of the car.

Table7: Correlation matrix for the all the attributes

Correlations

		selling_price	year	km_driven	mileage(kmpl/km/kg)	engine (CC)	max_power (BHP)	seats	Petrol	Diesel	CNG	LPG	Seller	Trans	First owner	Second owner	Third owner	Fourth & Above Owner	Test Drive Car
Pearson Correlation	selling_price	1.000	.412	-.222	-.126	.456	.750	.042	-.195	.205	-.033	-.036	-.386	-.590	.240	-.179	-.115	-.074	.116
	year	.412	1.000	-.429	.329	.018	.227	-.008	-.034	.038	.029	-.060	-.244	-.249	.492	-.317	-.271	-.206	.033
	km_driven	-.222	-.429	1.000	-.173	.206	-.038	.227	-.274	.272	-.005	.023	.203	.201	-.295	.210	.149	.089	-.024
	mileage(kmpl/km/kg)	-.126	.329	-.173	1.000	-.576	-.375	-.452	-.075	.060	.101	-.014	.012	.179	.166	-.102	-.097	-.072	-.016
	engine (CC)	.456	.018	.206	-.576	1.000	.704	.611	-.491	.507	-.060	-.057	-.132	-.283	-.016	.021	-.005	-.006	.014
	max_power(BHP)	.750	.227	-.038	-.375	.704	1.000	.192	-.286	.305	-.070	-.060	-.286	-.542	.115	-.085	-.054	-.037	.050
	seats	.042	-.008	.227	-.452	.611	.192	1.000	-.345	.355	-.039	-.029	.081	.073	-.035	.034	.005	.008	-.011
	Petrol	-.195	-.034	-.274	-.075	-.491	-.286	-.345	1.000	-.978	-.073	-.060	-.013	-.034	.044	-.055	.006	.009	.018
	Diesel	.205	.038	.272	.060	.507	.305	.355	-.978	1.000	-.089	-.073	.004	.026	-.040	.052	-.008	-.009	-.017
	CNG	-.033	.029	-.005	.101	-.060	-.070	-.039	-.073	-.089	1.000	-.005	.037	.032	-.004	.010	-.009	-.001	-.002
	LPG	-.036	-.060	.023	-.014	-.057	-.060	-.029	-.060	-.073	-.005	1.000	.025	.026	-.024	.013	.021	.004	-.002
	Seller	-.386	-.244	.203	.012	-.132	-.286	.081	-.013	.004	.037	.025	1.000	.379	-.230	.168	.113	.065	-.056
	Trans	-.590	-.249	.201	.179	-.283	-.542	.073	-.034	.026	.032	.026	.379	1.000	-.159	.122	.072	.040	-.050
	First owner	.240	.492	-.295	.166	-.016	.115	-.035	.044	-.040	-.004	-.024	-.230	-.159	1.000	-.814	-.366	-.200	-.035
	Second owner	-.179	-.317	.210	-.102	.021	-.085	.034	-.055	.052	.010	.013	.168	.122	-.814	1.000	-.154	-.084	-.015
	Third owner	-.115	-.271	.149	-.097	-.005	-.054	.005	.006	-.008	-.009	.021	.113	.072	-.366	-.154	1.000	-.038	-.007
	Fourth & Above Owner	-.074	-.206	.089	-.072	-.006	-.037	.008	.009	-.009	-.001	.004	.065	.040	-.200	-.084	-.038	1.000	-.004
	Test Drive Car	.116	.033	-.024	-.016	.014	.050	-.011	.018	-.017	-.002	-.002	-.056	-.050	-.035	-.015	-.007	-.004	1.000

The above table is the correlation matrix which describes the correlation between various attributes used in the dataset. From the above correlation matrix, it can be observed that the attributes year, engine, Max_power and transmission has high correlation with the selling price of the car when compared to the other variables. Then we have the attributes Diesel, KM_driven, Seller and First owner with some less correlation with the selling price of the car when compared to attributes year, engine, Max_power and transmission. So, here we perform the multiple regressions based on the correlation of the attributes with the selling price of the car.

11. Results and Discussions:

From the above scatter plots various relations between the attributes can be observed, where year, engine, Max_power and transmission attributes have high correlation with the selling price.

From the above correlation matrix the correlation between various attributes can be observed and we perform multiple regressions considering the different variables with the different correlations.

We utilized SPSS to run multiple regressions to estimate the dependent variable model and examine the relationship of the independent variables

Case 1:

From the observations made from the correlation matrix above. Let us consider year, engine, Max_power and transmission attributes which have high correlation with the selling price of the car. Using these attributes, we create a regression model where selling price of the car is the dependent variable and the attributes year, engine, Max_power and transmission as independent variables.

Then the general multiple regression model including all the three explanatory variables is given as:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

Where, the intercept a is the predicted value of Y when the entire $X_i's$ are equal to zero and $b_i's$ are the slope coefficient.

Table8: Coefficients table for attributes in case 1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-93827495.0	2950633.026		-31.799	<.001	-99611515.5	-88043474.4					
	year	46548.372	1464.262	.221	31.790	<.001	43678.032	49418.712	.412	.337	.209	.895	1.117
	engine (CC)	-62.920	15.423	-.039	-4.080	<.001	-93.154	-32.686	.456	-.046	-.027	.474	2.109
	max_power(BHP)	13887.182	249.280	.610	55.709	.000	13398.526	14375.837	.750	.531	.366	.361	2.772
	Trans	-518782.681	19242.302	-.216	-26.961	<.001	-556502.678	-481062.684	-.590	-.290	-.177	.677	1.478

a. Dependent Variable: selling_price

Note: VIF= Variation Inflation factor, Sig.= Significance, std. error= Standard error

The required multiple regression equation is:

Selling Price =-93827494.964+(46548.372) year +(-62.920) engine (CC)+(13887.182) max_power(BHP)+(-518782.681) Trans

Interpretation of the equation:

The equation above describes that the selling price of the car will be 93827494.964, when we consider all other variables as constants. The Selling prices increases by 46548.372 when we consider only year and consider other variables as constant.

The selling price of car decreases by 62.920, when we consider engine only and other variables as constant.

The selling price of the car increase by 13887.182, when we consider max_power and all other variables as constant.

The selling price of the car decreases by 518782.681, when we consider transmission and other variables as constant.

Table9: Model summary for attributes in case 1

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.811[a]	.658	.658	475844.4486	.658	3801.922	4	7901	.000	1.613

a. Predictors: (Constant), Trans, year, engine (CC), max_power(BHP)

b. Dependent Variable: selling_price

Note: df= Degrees of freedom, Sig.= Significance

$R^2\ Linear = 0.658$

The value of R indicates the variation of 65.8% in selling price of the car when we consider the highly correlated variables.

When the R value is near to one then the model is a good fit. Here the value of R is close to one, this estimated linear regression model for selling price of the car is a good fit.

Table10: ANOVA table for attributes in case 1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.443E+15	4	8.609E+14	3801.922	.000[b]
	Residual	1.789E+15	7901	2.264E+11		
	Total	5.232E+15	7905			

a. Dependent Variable: selling_price

b. Predictors: (Constant), Trans, year, engine (CC), max_power(BHP)

Note: df= Degrees of freedom, Sig.= Significance

As the F value is very high i.e 3801.922 as observed in the ANOVA table above the significance of the model will be very less. So, consider the other variables with a less correlation to obtain a better regression model.

Case 2:

Let us consider year, engine, Max_power, transmission, Diesel, KM_driven, Seller and First owner attributes which have good correlation with the selling price of the car. Using these attributes, we create a regression model where selling price of the car is the dependent variable and the attributes year, engine, Max_power, transmission, Diesel, KM_driven, Seller and First owner as independent variables.

Then the general multiple regression model including all the three explanatory variables is given as:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + ... + b_8X_8$$

Where, the intercept a is the predicted value of Y when the entire $X_i's$ are equal to zero and $b_i's$ are the slope coefficient.

Table11: Coefficients table for attributes in case 2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-69981995.9	3468535.888		-20.176	<.001	-76781243.4	-63182748.4					
	year	34793.742	1722.529	.165	20.199	<.001	31417.130	38170.354	.412	.222	.130	.620	1.613
	engine (CC)	-79.080	16.831	-.049	-4.698	<.001	-112.073	-46.087	.456	-.053	-.030	.382	2.620
	max_power(BHP)	13424.897	245.544	.590	54.674	.000	12943.565	13906.228	.750	.524	.352	.356	2.806
	Trans	-463250.407	19850.038	-.193	-23.338	<.001	-502161.731	-424339.084	-.590	-.254	-.150	.609	1.641
	Diesel	114434.662	12800.540	.070	8.940	<.001	89342.218	139527.106	.205	.100	.058	.675	1.481
	km_driven	-1.073	.110	-.075	-9.720	<.001	-1.289	-.857	-.222	-.109	-.063	.698	1.432
	Seller	-197367.152	15524.386	-.091	-12.713	<.001	-227799.054	-166935.251	-.386	-.142	-.082	.808	1.238
	First owner	32878.203	12877.057	.019	2.553	.011	7635.766	58120.641	.240	.029	.016	.737	1.356

a. Dependent Variable: selling_price

Note: VIF= Variation Inflation factor, Sig.= Significance, std. error= Standard error

The required multiple regression equation is:

Selling Price=-69981995.898+(34793.742) year+(-79.080) engine (CC)+(13424.897) max_power(BHP)+(-463250.407) Trans+(114434.662) Diesel+(-1.073) km_driven+ (-197367.152) Seller+(32878.203) First owner

Interpretation of the equation:

The equation above describes that the selling price of the car will be 69981995.898, when we consider all other variables as constants. The Selling prices increases by 34793.742 when we consider only year and consider other variables as constant.

The selling price of car decreases by 79.080, when we consider engine only and other variables as constant.

The selling price of the car increase by 13424.897, when we consider max_power and all other variables as constant.

The selling price of the car decreases by 463250.407, when we consider transmission and other variables as constant.

The Selling price of the car increases by 114434.662, when we consider Diesel and all other variables as constant.

The selling price of the car decreases by 1.073, when we consider KM_driven and other variables as constant.

The selling price of the car decreases by 197367.152, when we consider Seller and other variables as constant.

The Selling price of the car increases by 32878.203, when we consider First owner and all other variables as constant.

Table12: Model Summary for attributes in case 2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.820[a]	.672	.672	465879.1739	.672	2026.358	8	7897	.000	1.613

a. Predictors: (Constant), First owner, engine (CC), Seller, km_driven, Trans, Diesel, year, max_power(BHP)

b. Dependent Variable: selling_price

Note: df= Degrees of freedom, Sig.= Significance

$R^2\ Linear = 0.672$

The value of R indicates the variation of 67.2% in selling price of the car when we consider the highly correlated variables.

When the R value is near to one then the model is a good fit. Here the value of R is close to one, this estimated linear regression model for Selling price of the car is a good fit when compared to case 1.

Table13: ANOVA table for attributes in case 2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.518E+15	8	4.398E+14	2026.358	.000[b]
	Residual	1.714E+15	7897	2.170E+11		
	Total	5.232E+15	7905			

a. Dependent Variable: selling_price

b. Predictors: (Constant), First owner, engine (CC), Seller, km_driven, Trans, Diesel, year, max_power(BHP)

Note: df= Degrees of freedom, Sig.= Significance

As the F value is very high i.e 2026.358 as observed in the ANOVA table above the significance of the model will be more less when compared to the case 1. So, consider the all other variables to obtain a better regression model.

Case 3:

Let us consider year, engine, Max_power, transmission, Diesel, KM_driven, Seller and First owner, mileage, seats, petrol, CNG, LPG, second owner, Third owner, fourth owner & anbove, Test drive car attributes which have good correlation with the selling price of the car. Using these attributes, we create a regression model where selling price of the car is the dependent variable and the remaining attributes as independent variables.

Then the general multiple regression model including all the three explanatory variables is given as:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + ... + b_iX_i$$

Where, the intercept a is the predicted value of Y when the entire $X_i's$ are equal to zero and $b_i's$ are the slope coefficient.

Table14: Coefficients table for attributes in case 3

Coefficients[a]

el	Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-64025584.8	3841088.857		-16.669	<.001	-71555135.7	-56496033.9					
year	31789.781	1920.664	.151	16.551	<.001	28024.771	35554.790	.412	.183	.105	.487	2.054
km_driven	-1.070	.109	-.075	-9.781	<.001	-1.284	-.855	-.222	-.109	-.062	.695	1.439
mileage(kmpl/km/kg)	11464.317	2149.182	.057	5.334	<.001	7251.350	15677.283	-.126	.060	.034	.356	2.806
engine (CC)	56.823	23.759	.035	2.392	.017	10.250	103.397	.456	.027	.015	.187	5.345
max_power(BHP)	13024.191	259.335	.572	50.221	.000	12515.825	13532.557	.750	.492	.320	.312	3.205
seats	-33695.489	8038.990	-.040	-4.192	<.001	-49454.038	-17936.940	.042	-.047	-.027	.451	2.218
Petrol	-78315.781	15030.286	-.048	-5.211	<.001	-107779.120	-48852.442	-.195	-.059	-.033	.480	2.081
CNG	27100.803	64773.527	.003	.418	.676	-99872.455	154074.060	-.033	.005	.003	.978	1.023
LPG	191251.665	79236.002	.016	2.414	.016	35928.126	346575.203	-.036	.027	.015	.969	1.032
Seller	-182346.072	15427.641	-.084	-11.819	<.001	-212588.332	-152103.813	-.386	-.132	-.075	.799	1.252
Trans	-452906.946	19819.731	-.188	-22.851	<.001	-491758.865	-414055.027	-.590	-.249	-.145	.597	1.675
Second owner	-47902.091	13412.758	-.026	-3.571	<.001	-74194.647	-21609.535	-.179	-.040	-.023	.784	1.275
Third owner	-21504.271	23085.798	-.006	-.931	.352	-66758.546	23750.004	-.115	-.010	-.006	.834	1.200
Fourth & Above Owner	7405.806	38567.214	.001	.192	.848	-68196.143	83007.754	-.074	.002	.001	.909	1.100
Test Drive Car	2177296.456	206552.316	.067	10.541	<.001	1772399.243	2582193.669	.116	.118	.067	.994	1.006

Dependent Variable: selling_price

Note: VIF= Variation Inflation factor, Sig.= Significance, std. error= Standard error

Table15: Excluded variables table for case 3

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Diesel	.[b]	.	.	.	.000	.	.000
	First owner	.000[b]	.000	1.000	.000	5.163E-14	1.937E+13	5.163E-14

a. Dependent Variable: selling_price

b. Predictors in the Model: (Constant), Test Drive Car, LPG, Fourth & Above Owner, CNG, Third owner, seats, Trans, Second owner, Petrol, Seller, km_driven, mileage(kmpl/km/kg), year, max_power(BHP), engine (CC)

Note: VIF= Variation Inflation factor, Sig.= Significance, std. error= Standard error

The required multiple regression equation is:

Selling Price = -64025584.833+(31789.781) year+(-1.070) km_driven+(11464.317) mileage(kmpl/km/kg)+(56.823) engine (CC)+(13024.191) max_power(BHP)+(-33695.489) seats+(-78315.781) Petrol+(27100.803) CNG+(191251.665) LPG+(-182346.072) Seller+

(-452906.946) Trans+(-47902.091) Second owner+(-21504.271) Third owner+(7405.806) Fourth & Above Owner+(2177296.456) Test Drive Car

Interpretation of the equation:

The equation above describes that the selling price of the car will be 64025584.833, when we consider all other variables as constants. The Selling prices increases by 31789.781when we consider only year and consider other variables as constant.

The selling price of car increases by 56.823 when we consider engine only and other variables as constant.

The selling price of the car increase by 13024.191 when we consider max_power and all other variables as constant.

The selling price of the car decreases by 452906.946 when we consider transmission and other variables as constant.

The selling price of the car decreases by 1.070 when we consider KM_driven and other variables as constant.

The selling price of the car decreases by 182346.072 when we consider Seller and other variables as constant.

The Selling price of the car increases by 11464.317 when we consider mileage and all other variables as constant.

The selling price of the car decreases by 33695.489 when we consider Seats and other variables as constant.

The selling price of the car decreases by 78315.781 when we consider Petrol type and other variables as constant.

The Selling price of the car increases by 27100.803 when we consider CNG type and all other variables as constant.

The Selling price of the car increases by 191251.665 when we consider LPG type and all other variables as constant.

The selling price of the car decreases by 47902.091 when we consider Second owner and other variables as constant.

The selling price of the car decreases by 21504.271 when we consider Third owner and other variables as constant.

The Selling price of the car increases by 7405.806 when we consider Fourth & above owner and all other variables as constant.

The Selling price of the car increases by 2177296.456 when we consider Test Drive car and all other variables as constant.

The variables First owner and Diesel are excluded in the multiple regression as they are redundant with other variables that are in the model.

Table16: Model Summary for attributes in case 1

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.825[a]	.680	.680	460385.8400	.680	1119.774	15	7890	.000	1.622

a. Predictors: (Constant), Test Drive Car, LPG, Fourth & Above Owner, CNG, Third owner, seats, Trans, Second owner, Petrol, Seller, km_driven, mileage(kmpl/km/kg), year, max_power(BHP), engine (CC)

b. Dependent Variable: selling_price

Note: df= Degrees of freedom, Sig.= Significance

$R^2\ Linear = 0.680$

The value of R indicates the variation of 68% in selling price of the car when we consider the highly correlated variables.

When the R value is near to one then the model is a good fit. Here the value of R is close to one, this estimated linear regression model for Selling price of the car is a good fit when compared to case 1 and case 2.

Table17: ANOVA table for attributes in case 1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.560E+15	15	2.373E+14	1119.774	.000[b]
	Residual	1.672E+15	7890	2.120E+11		
	Total	5.232E+15	7905			

a. Dependent Variable: selling_price

b. Predictors: (Constant), Test Drive Car, LPG, Fourth & Above Owner, CNG, Third owner, seats, Trans, Second owner, Petrol, Seller, km_driven, mileage(kmpl/km/kg), year, max_power(BHP), engine (CC)

Note: df= Degrees of freedom, Sig.= Significance

As the F value is very high i.e. 1119.774 as observed in the ANOVA table above the significance of the model will be more less when compared to the case 1 and case 2. So, we can conclude that case 3 is the better model for prediction of the selling price of the car.

12. Hypothesis Testing:

To validate an assumption regarding the selling price of the car we use Hypothesis testing. Based on Sample known statistics, samples and distribution, different types of hypothesis can be employed.

Steps:

1. Draft the null and alternate hypothesis.
2. Determine the essential value and plan the test to be run
3. Conduct the experiment and get the test statistics.
4. Reject the NULL hypothesis or assert that it is plausible.

One sample t-test of the kilometers traveled by the seller-type "Individual"

NULL Hupothesis, H0: The actual number of kilometers driven by sellers who identify as "Individuals" is the same as the average for all sellers.

Alternate Hypothesis, H1: The True Mean of Kilometers Traveled by Seller Type of 'Individual' is Greater than the Average.

Alternative: Greater

We reject the null hypothesis of the test since the p-value of our one-sample t-test is smaller than our alpha = 0.05. The True Mean of the kilometers driven by the 'Individual' seller-type in the data set is therefore bigger than the True Mean of kilometers driven by other seller-types in the data set.

One sample t-test of the selling price of automobiles sold by the seller-type of the 'Dealer'

The null hypothesis H0: the true average selling price of cars sold by sellers who identify as "Dealers" is the same as the overall average.

Alternate Hypothesis, H1: The true mean selling price of vehicles sold by seller types classified as "Dealers" is higher than the average of all vehicles.

Alternative: Greater

We reject the null hypothesis of the test since the p-value for our one-sample t-test is smaller than our alpha = 0.05. Thus, we have enough data to conclude that the True Mean of the selling price of automobiles sold by the seller type "Dealer" in the data set is higher than the True Mean of the selling price of vehicles sold by all seller types in the data set.

One sample t-Test showing the kilometers logged by the vehicles' "First Owners"

NULL Hypothesis, H0: The actual kilometers driven by ‘First Owner’ owner-type are equal to the overall average.

Alternate Hypothesis, H1: The actual kilometers driven by ‘First Owner’ owner-type are less than the overall average.

Alternative: Less

We reject the null hypothesis of the test since the p-value for our one-sample t-test is smaller than our alpha = 0.05. We conclude that the actual of kilometers driven by the "First Owner" owner-type in the data set is lower than the actual kilometers driven by all other "First Owner" owner-types in the data set.

One sample t-test of the selling price of automobiles sold that were owned by their "First Owners"

NULL Hypothesis, H0 : The actual of selling price of cars sold by ‘First Owner’ owner-type is equal to the overall average.

Alternate Hypothesis, H1 : The actual of selling price of cars sold by ‘First Owner’ owner-type is greater than the overall average.

Alternative : Greater

We reject the null hypothesis of the test since the p-value of our one-sample t-test is less than our alpha = 0.05. This indicates that the selling price of vehicles sold by the owner-type "First Owner" in the data set is higher than the selling price of cars sold by all owner-type in the data set.

Chi Square test to see whether "Owner Type & Fuel" are associated

NULL Hypothesis, H0: There is no relationship between Owner-Type and Fuel-Type of cars. The variables are independent.

Alternate Hypothesis, H1: There exists some relationship between Owner-Type and Fuel-Type of cars. The variables are dependent.

One-way ANOVA between the continuous variables Owner and Fuel and the category Owner

Categorical variables:

- fuel
- owner

Continuous variables:

- max_power
- engine
- mileage
- km_driven

NULL Hypothesis, H0: The Fuel and Max-Power groups are same.

Alternate Hypothesis, H1: The Fuel and Max-Power groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. This indicates that the Fuel and Max-Power groups' True Means are not the same in the data set.

NULL Hypothesis, H0: The Fuel and Engine groups are same.

Alternate Hypothesis, H1: The Fuel and Engine groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. As a result, we can conclude that the data set's True Means for the Fuel and Engine groups are different.

NULL Hypothesis, H0: The Fuel and Mileage groups are same.

Alternate Hypothesis, H1: The Fuel and Mileage groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. This indicates that the Fuel and Mileage groups' True Means are not the same in the data set.

NULL Hypothesis, H0: The Fuel and Kilometers-Driven groups are same.

Alternate Hypothesis, H1: The Fuel and Kilometers-Driven groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. As a result, we can conclude that the data set's True Means for the Fuel and Kilometers-Driven groups are different.

NULL Hypothesis, H0: The Owner and Max-Power groups are same.

Alternate Hypothesis, H1: The Owner and Max-Power groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. This indicates that the Max-Power and Owner groups' True Means are not the same in the data set.

NULL Hypothesis, H0: The Owner and Engine groups are same.

Alternate Hypothesis, H1: The Owner and Engine groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. This signifies that the True Means of the Owner and Engine groups in the data set are not the same.

NULL Hypothesis, H0: The Owner and Mileage groups are same.

Alternate Hypothesis, H1: The Owner and Mileage groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. This indicates that the Owner and Mileage groups' True Means are not the same in the data set.

NULL Hypothesis, H0: The Owner and Kilometers-Driven groups are same.

Alternate Hypothesis, H1: The Owner and Kilometers-Driven groups are not same.

We may reject the null hypothesis of the test because the p-value for our ANOVA test is smaller than our alpha value of 0.05. This indicates that the Owner and Kilometers-Driven groups' True Means are not the same in the data set.

13. Limitations:

In the past few years, the world of automobiles is facing a drastic change with the shortage of semiconductors after the pandemic, which led to increase in the prices of the used cars. Hence, the changes in the car prices will affect the actual price prediction in the future. As the car sales dataset will undervalue the cars in the future market. Also, the dataset used in this research has limited information which was not sufficient to predict the price 100% accurately, we have achieved only 68% accuracy based on the present data available in the dataset. If more attributes were provided then the accuracy of the prediction of the selling price will increase. Therefore, a model built on real time data will be useful to predict the selling price of the car accurately. Hence the multiple regression model is not enough to predict the selling price with 100% accuracy due to data insufficiency.

14. Conclusion:

Various insights about the types of cars sold in the industry and the relations between them have been provided in the dataset of car sales. The data set contains 8128 data points along with 13 features related to car details.

Data cleaning and data preprocessing is done so that data can be used efficiently. Unnecessary data is dropped from the dataset and Null or blank spaces were also removed. Dummy variables were created to make categorical variables to numerical. Once data cleaning and exploratory analysis is performed, hypothesis testing is done to validate the assumptions on the given data.

By developing regression models and contrasting them with one another, we were able to accurately estimate the Selling price of the used car. Selling price of the car differ significantly different attributes in the data. The sales data shows that there will be influence on the selling price of the used cars by the factors, such as year, engine, Max_power, transmission, Diesel, KM_driven, Seller and First owner, mileage, seats, petrol, CNG, LPG, second owner, Third owner, fourth owner & anbove, Test drive car. In actual sales, additional considerations like car sales history, car tax, rebuild information etc. are considered. We were unable to deal with such data because the dataset lacked such information in the data set used in this research.

15. References:

Text Book: Business analytics data analysis and decision making by S. Christian Albright Wayne L. Winston.

1. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), 753-764.

2. Ackerman, S. R. (1973). Used cars as a depreciating asset. *Economic Inquiry*, *11*(4), 463.

3. Andrews, T., & Benzing, C. (2007). The determinants of price in internet auctions of used cars. *Atlantic Economic Journal*, *35*(1), 43-57.

4. Grubel, H. G. (1980). International trade in used cars and problems of economic development. *World Development*, *8*(10), 781-788.

5. Emons, W., & Sheldon, G. (2009). The market for used cars: new evidence of the lemons phenomenon. *Applied economics*, *41*(22), 2867-2885.

6. Berkovec, J. (1985). New car sales and used car stocks: A model of the automobile market. *The Rand Journal of Economics*, 195-214.

7. Asghar, M., Mehmood, K., Yasin, S., & Khan, Z. M. (2021). Used Cars Price Prediction using Machine Learning with Optimal Features. *Pakistan Journal of Engineering and Technology*, *4*(2), 113-119.

8. Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. *TEM Journal*, *8*(1), 113.

9. Venkatasubbu, P., & Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. *Int. J. Eng. Adv. Technol.(IJEAT)*, *9*(1S3).

10. Benabbou, F., Sael, N., & Herchy, I. (2022). Machine Learning for Used Cars Price Prediction: Moroccan Use Case. In *International Conference On Big Data and Internet of Things* (pp. 332-346). Springer, Cham.

11. Jin, C. (2021, November). Price Prediction of Used Cars Using Machine Learning. In *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)* (pp. 223-230). IEEE.

12. Pandey, A., Rastogi, V., & Singh, S. (2020, March). Car's selling price prediction using random forest machine learning algorithm. In *5th International Conference on Next Generation Computing Technologies (NGCT-2019)*.

13. Jayadeera, T. P., & Jayamanne, D. J. (2019). Fair price prediction system for used cars in Sri Lanka using machine learning and robotic process automation. In *2019 International Conference On Business Innovation (ICOBI), Colombo*.

14. Reddy, A., & Kamalraj, R. (2021). Old/Used Cars Price Prediction using Machine Learning Algorithms. *IITM Journal of Management and IT*, *12*(1), 32-35.

15. Das Adhikary, D. R., Sahu, R., & Pragyna Panda, S. (2022). Prediction of Used Car Prices Using Machine Learning. In *Biologically Inspired Techniques in Many Criteria Decision Making* (pp. 131-140). Springer, Singapore.