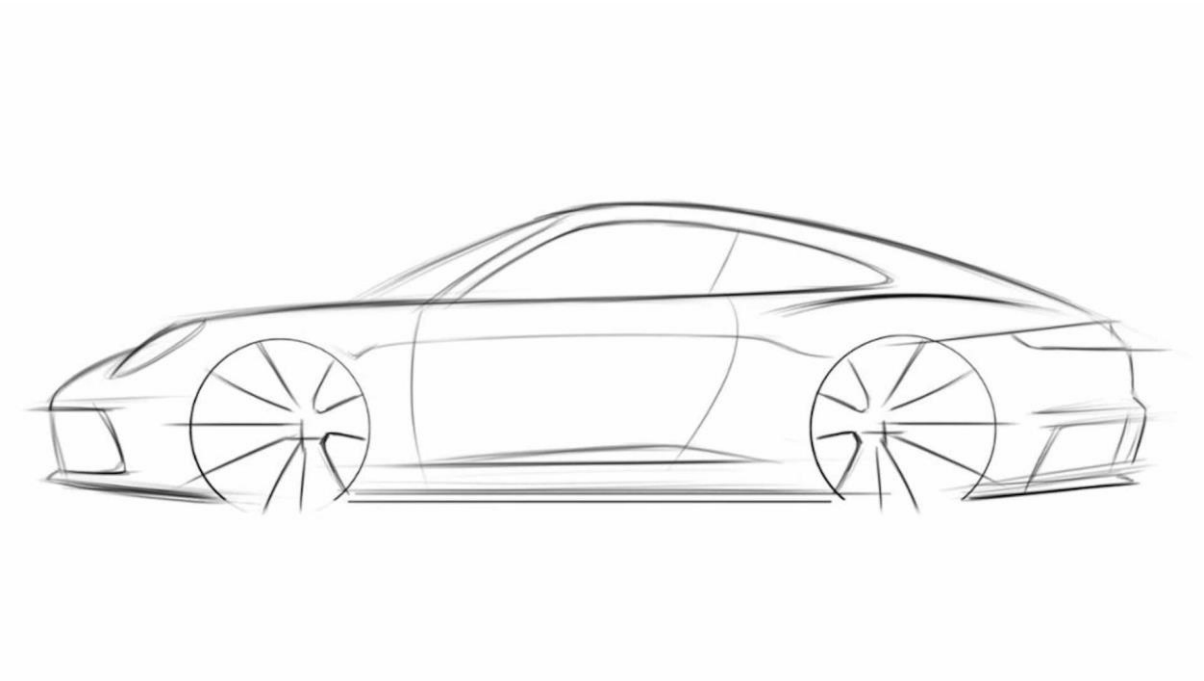
CARS

The used car market in India has been the centre of attention in the slow growing automotive industry in India. In the last year, demand for used cars has soared with over 42 lakh buyers (Source: CRISIL).

A few years ago, the ratio of new cars to used cars was 1:1.2 which is now at 1:2.2. Basically, when 10 new cars are sold, 22 used cars are available for sale in the market.

In 2008-09, the estimated sale of pre-owned cars was at 37 lakhs. For the 2018-19 period, projected sales are at 62 lakhs estimated to be worth Rs.1.62 lakh crores. The average holding time of a new car has come down to just 3 years which was 5-6 years earlier. For the first time aspiring buyer, the used car category has opened up many options.

We will further try to analyse all variables present in the dataset and check whether which feature is the most responsible for decrease in the price of a car over the period of years.



Features – Cars

- Car_Name: Represents name of the vehicle.
- Year: Represents the year in which the car was bought.
- Selling_Price: Represents with the price at which the owner wants to sell the car.
- Present_Price: Represents the current ex-showroom price of the vehicle.
- Kms_Driven: Represents the distance completed by the vehicle in km.
- Fuel_Type: Represents the fuel type of the vehicle.
- Seller_Type: Defines is the seller is an individual or a dealer.
- Transmission: Defines whether the vehicle is automatic or manual.
- Owners: Defines the number of owners the car previously had.

Further, we will look into the data wrangling techniques used for this particular data set.

Data Wrangling Techniques – Cars

Some of the data wrangling techniques used are mentioned below as follows:

- Data Set was acquired from Kaggle open source of Cardekho.
- Null values were looked up and discarded to avoid errors on the dataset.
- Duplicate values were first highlighted, cross checked with other columns for difference and then discarded to avoid errors in statistical modelling.
- Data was then plotted using a scatter plot and a bar chart to check for outliers.
- We then used label encoder to change categorical values to numerical for better correlation amongst each feature.
- We gave an addition to the dataset by providing a column named depreciation which is the difference of the selling price and present price of the vehicle. The column is named as depreciation.

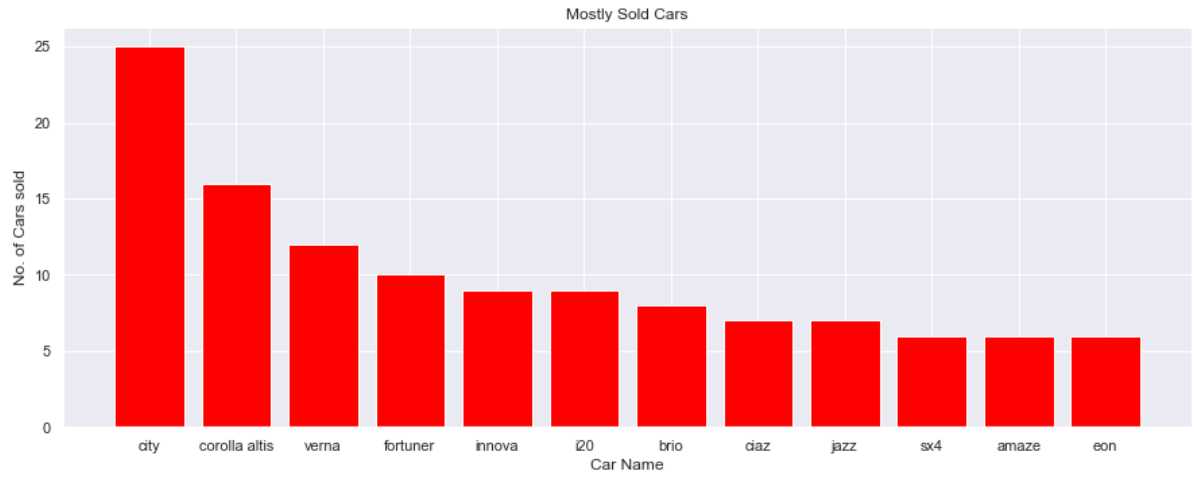
	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner	depreciation
0	ritz	2014	3.35	5.59	27000	2	0	1	0	2.24
1	sx4	2013	4.75	9.54	43000	1	0	1	0	4.79
2	ciaz	2017	7.25	9.85	6900	2	0	1	0	2.60
3	wagon r	2011	2.85	4.15	5200	2	0	1	0	1.30
4	swift	2014	4.60	6.87	42450	1	0	1	0	2.27

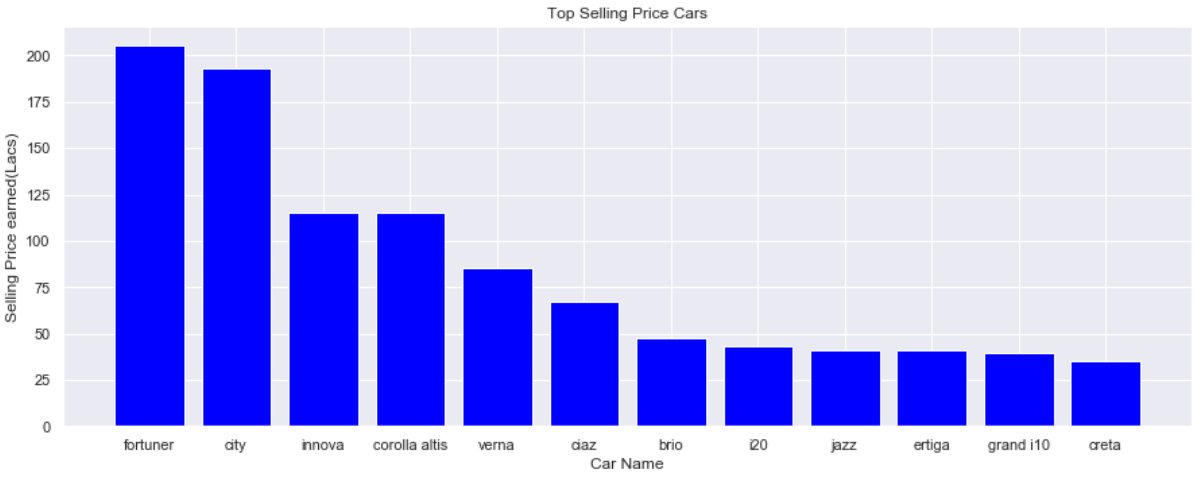
No outliers were found post wrangling method and later values were checked using the .describe() function in the pandas library. Exploratory Data Analysis was done post successful wrangling.

Exploratory Data Analysis (EDA)

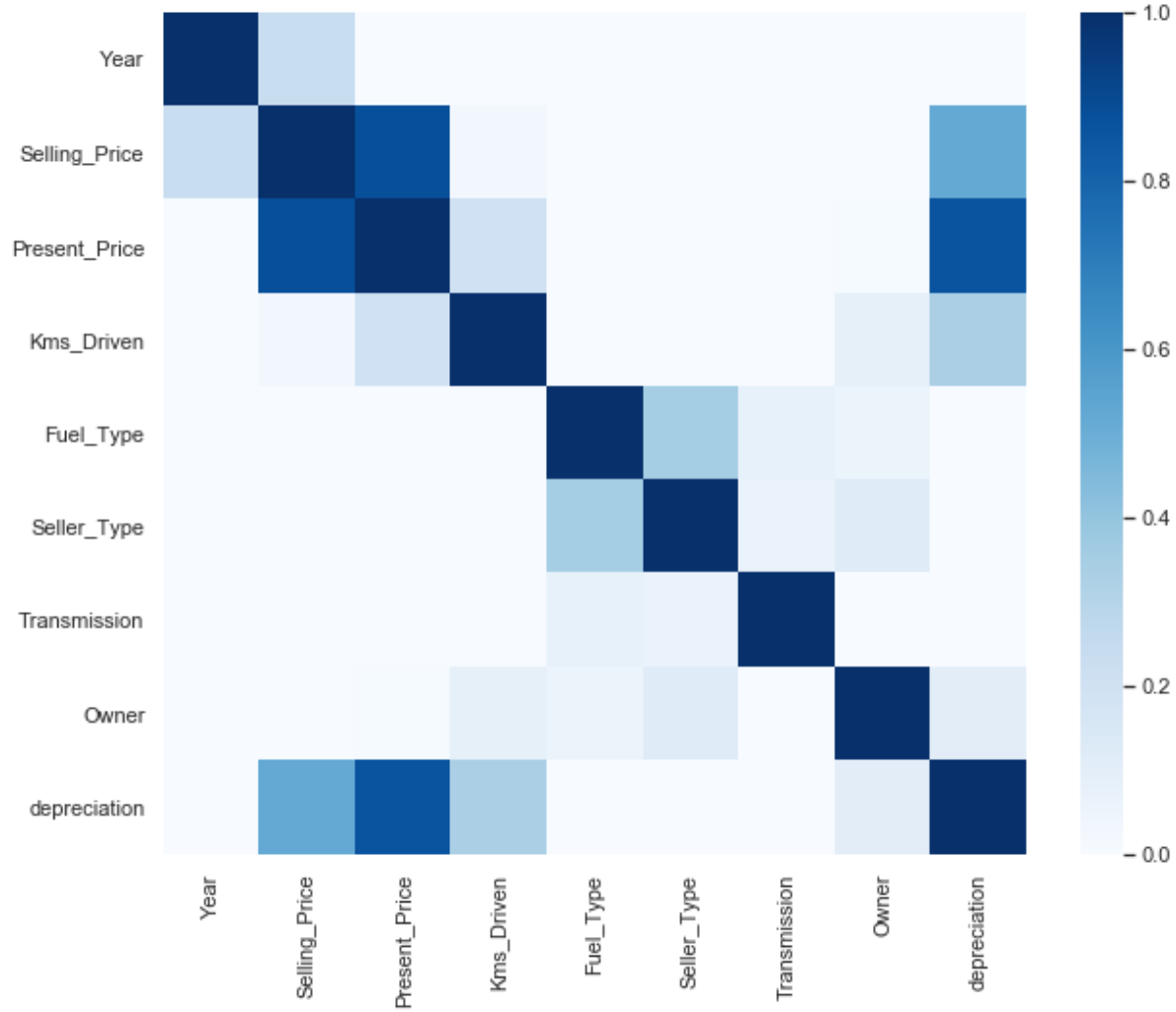
In this section, the various insights produced through descriptive statistics and data visualization is presented.

We plotted the graph to find the most numbers of cars sold on the basis of their count and a separate one showing the highest cost of the vehicle sold.





We further checked the correlation of different variables present in the data set across each other.



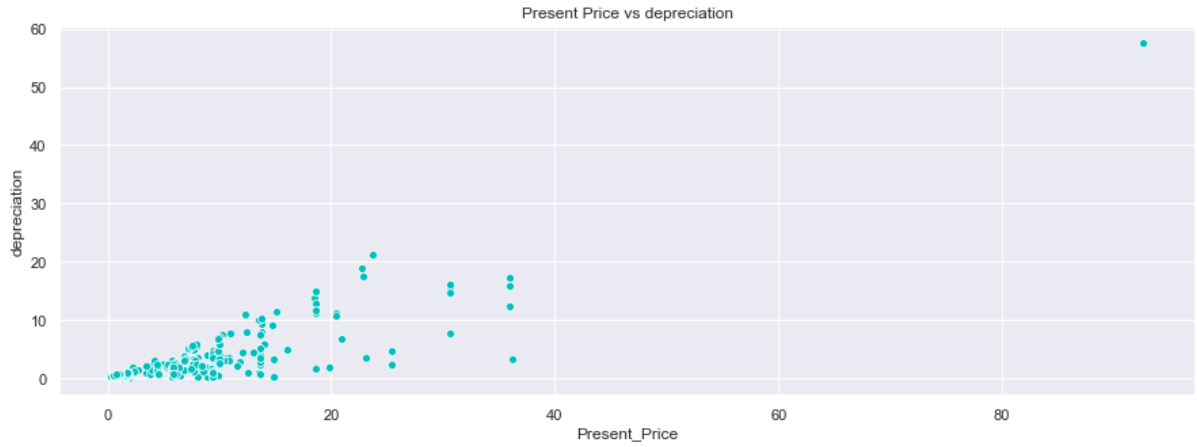
This means there is a strong corelation between Selling Price, Present price of cars and depreciation. Further, a less strong relationship between Present Price- Kilometers driven, selling price-year of buying that car and so on. There is very less correlation of Selling price with Fuel_type,Seller_Type and Owner.

In addition, we can also see that depreciation has a stronger correlation with Present_Price followed by Selling_Price, Kms_Driver and Owner.
Since we know which variables have a higher correlation, let's dive in deep to compare individual variables.

We will further begin to show scatter plots for different relationships.

- Selling_Price vs Present_Price

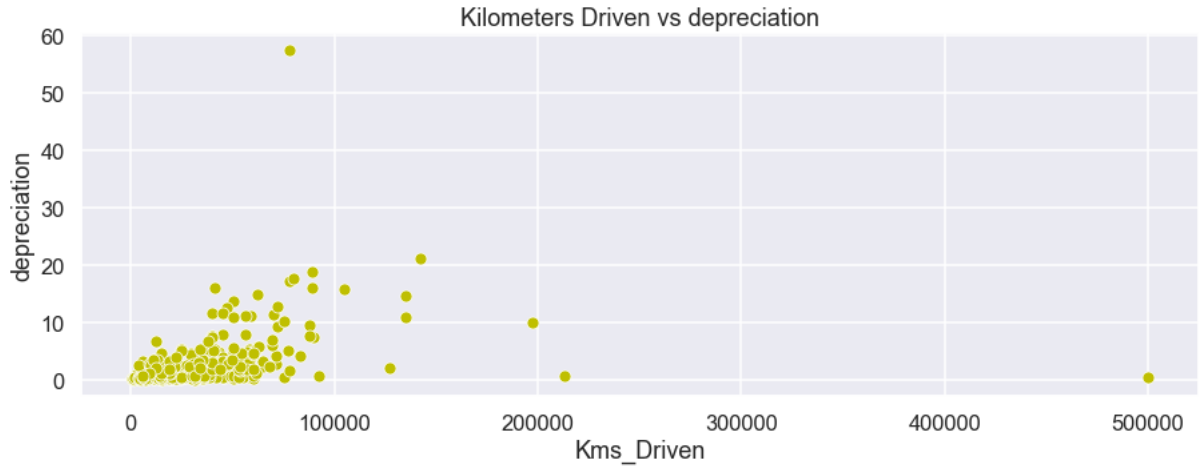
- Present_Price vs depreciation of the vehicle



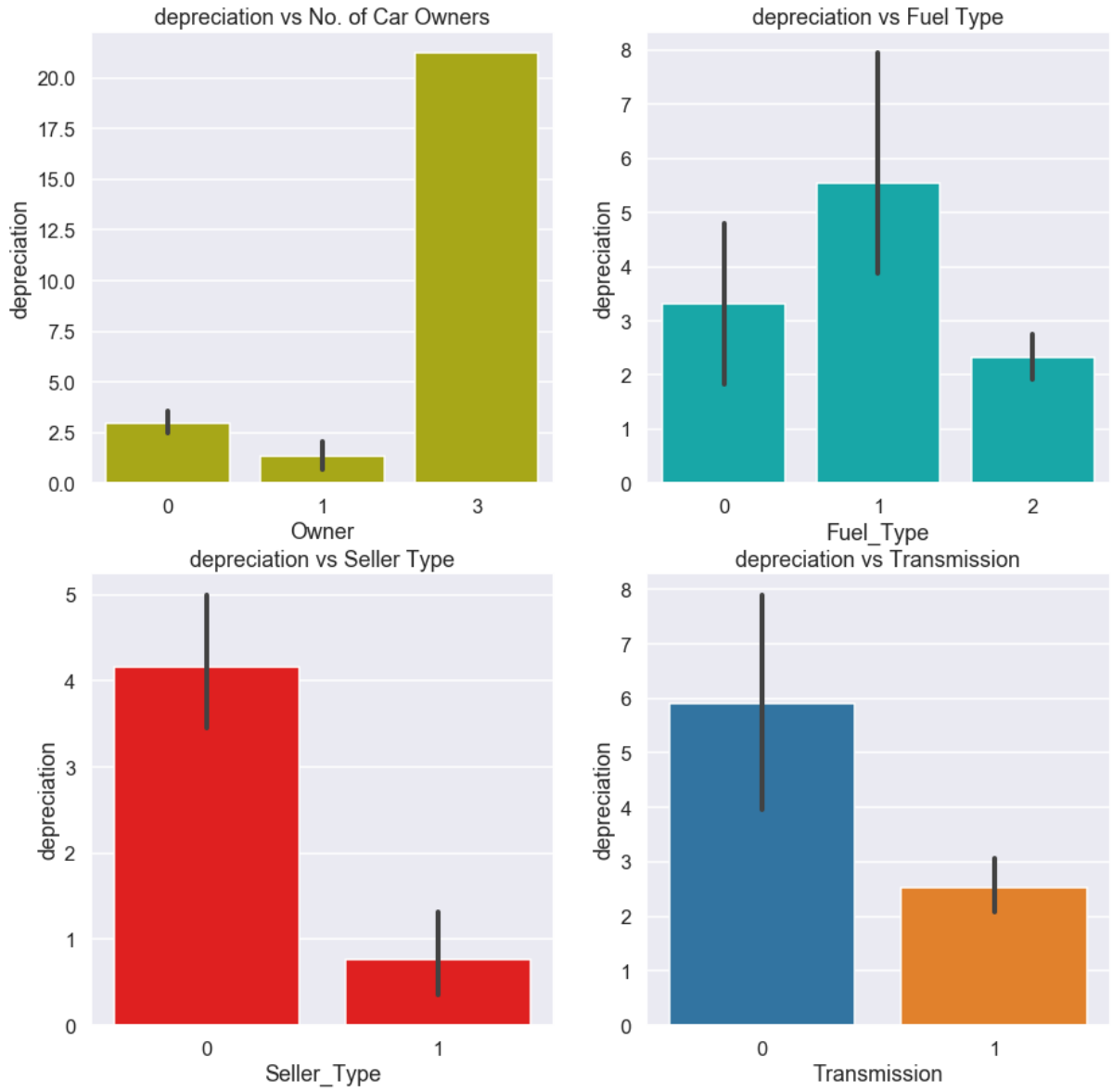
- Selling_Price vs depreciation of the vehicle

We then looked into the relationship between Selling_Price and year of purchase

To further analyse variables, we chose the relationship between depreciation and Kms_Driven



Since, we have the depreciation of the vehicles, now let's look at an overall behaviour of depreciation with other variables



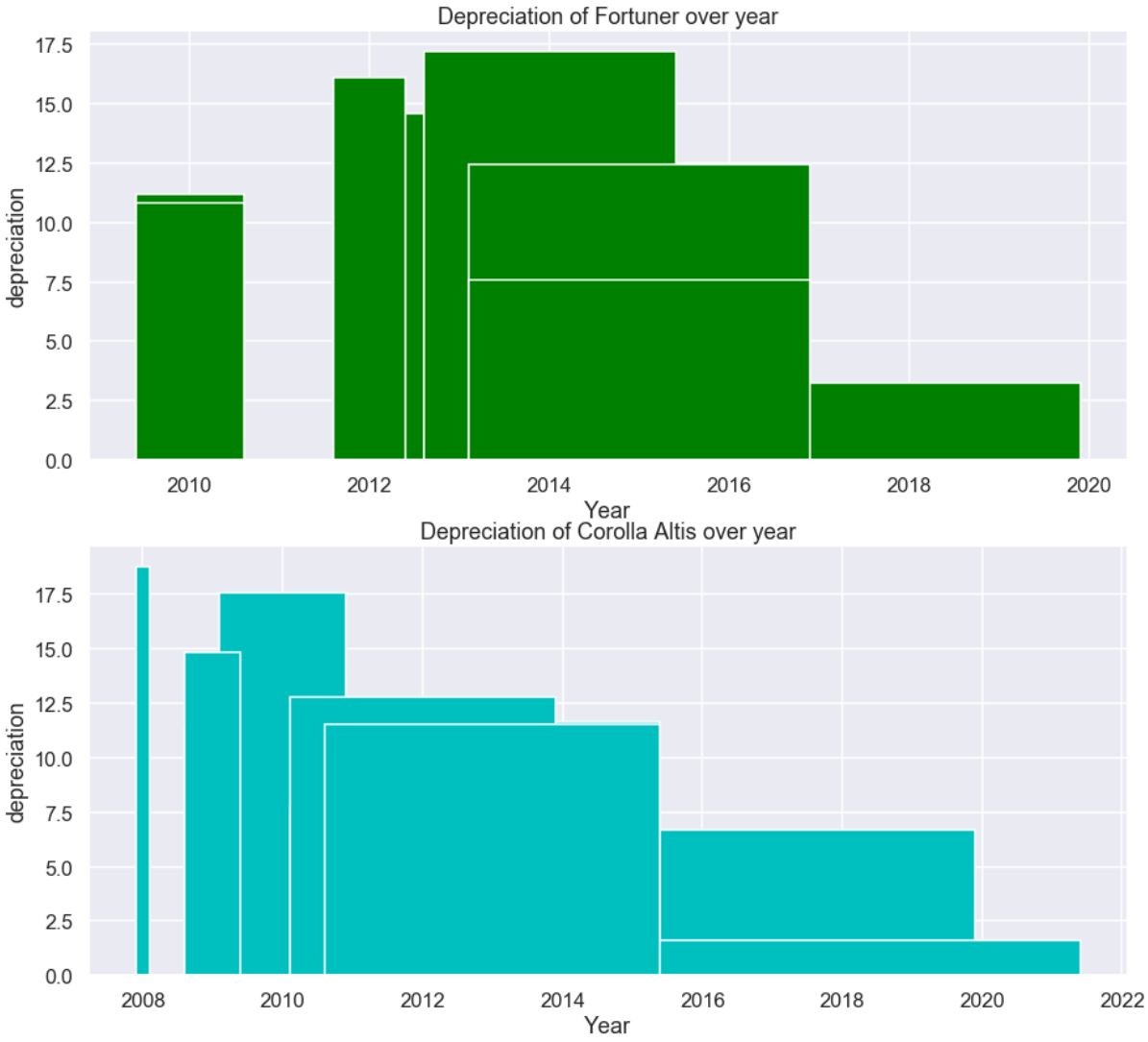
Here, we can find following obervations:

- The cars with a greater number of owners are having large reduction of Selling Price.
- Cars with Fule Type-1(i.e Diesel Cars) have higher depreciaiton in Selling Price.
- Seller Type-0(i.e. Cars sold from Dealers have higher depreciation in Selling Price.
- The cars with Manual transmission (i.e. value=0) have higher depreciation in Selling Price.

Furthermore, we grouped the maximum depreciation of cars by their value.

	depreciation
Car_Name	
land cruiser	57.600000
camry	21.230000
fortuner	12.047273
corolla	10.850000
corolla altis	10.399375
innova	5.068889
sx4	4.908333
city	4.372692
etios cross	3.333333
verna	3.292143
elantra	3.190000
dzire	3.157500
etios gd	3.100000
ertiga	2.848333

Cars like Land Cruiser, Camry, Corolla are having only one record. So, let us analyse if there is any relationship between year and selling price of Fortuner and Corolla Altis.



The above graph shows that variables such as Transmission and Seller_Type do not have much impact on depreciation as compared to Kms_Driver and Year. Hence, we can conclude that Kms_driven and year have a hige impact on depreciation than other factors.