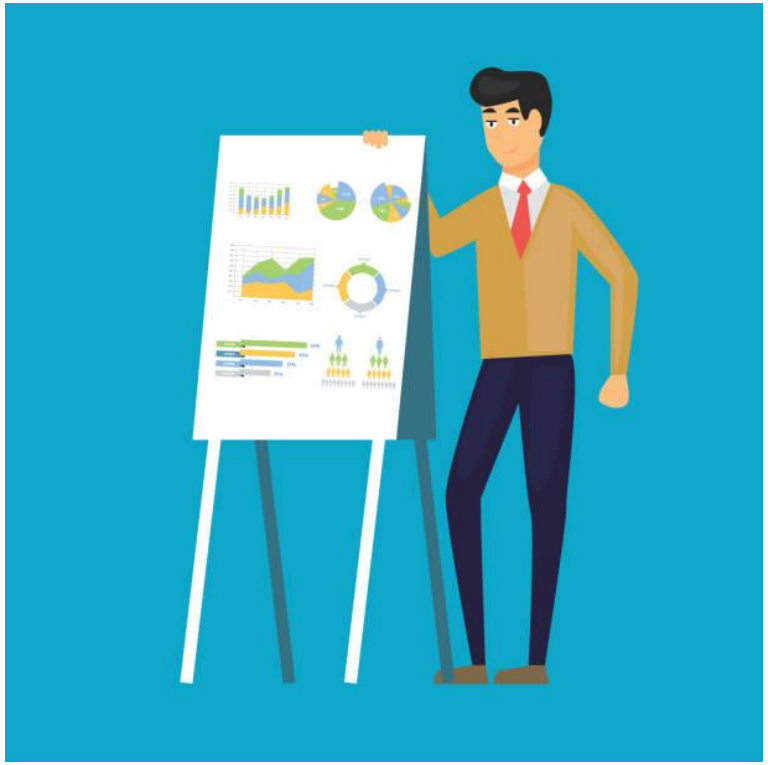
Capstone Project On Second Hand Car Price Prediction

By Mohit Bohra



Agenda

- Data Exploration
- Data insights
- EDA Graphs.
- Graphical Analysis and conclusion on Data
- Data Cleaning & Pre-Processing Steps.
- ML Modeling
- Deployment of ML Models using Streamlit.

Data Exploration

- This data looks like second-hand car data set contains information about used cars that are being sold, and it can be used for various analytical purposes such as price prediction, trend analysis, and market segmentation. It includes important features such as the car make and model, year of manufacture, fuel type, seller type, transmission type, and number of previous owners. This type of dataset can be useful for individuals looking to buy or sell a used car, as well as businesses in the automobile industry looking to gain insights into consumer behavior and market trends. It can also be used by data scientists and machine learning engineers to build predictive models and perform analysis to optimize pricing and marketing strategies.

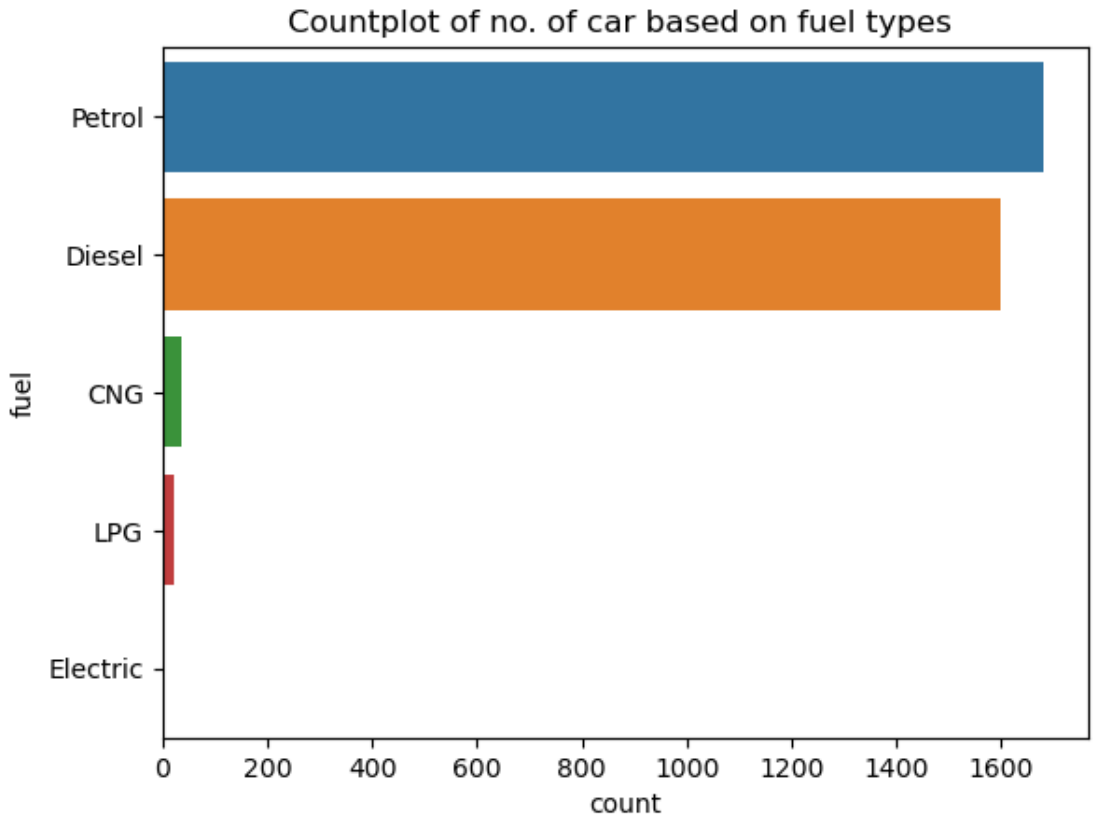
	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner

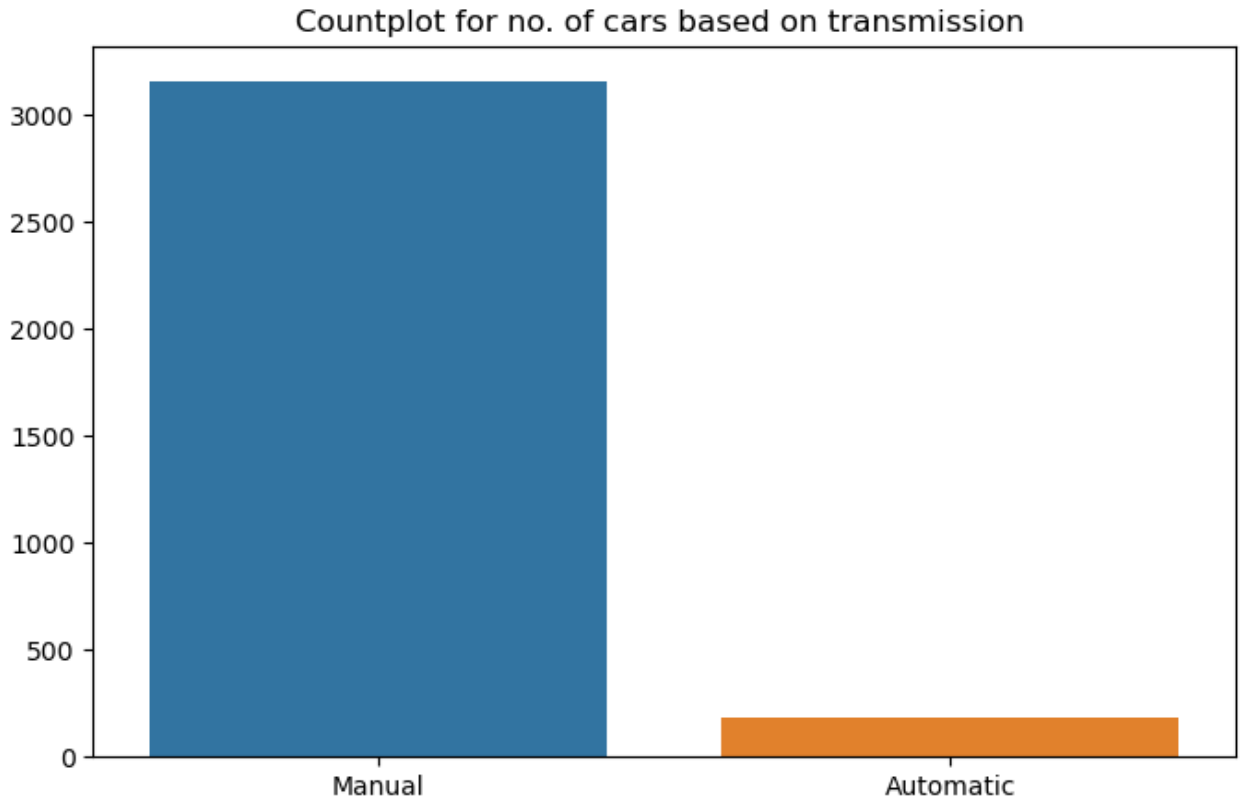
Data Insights

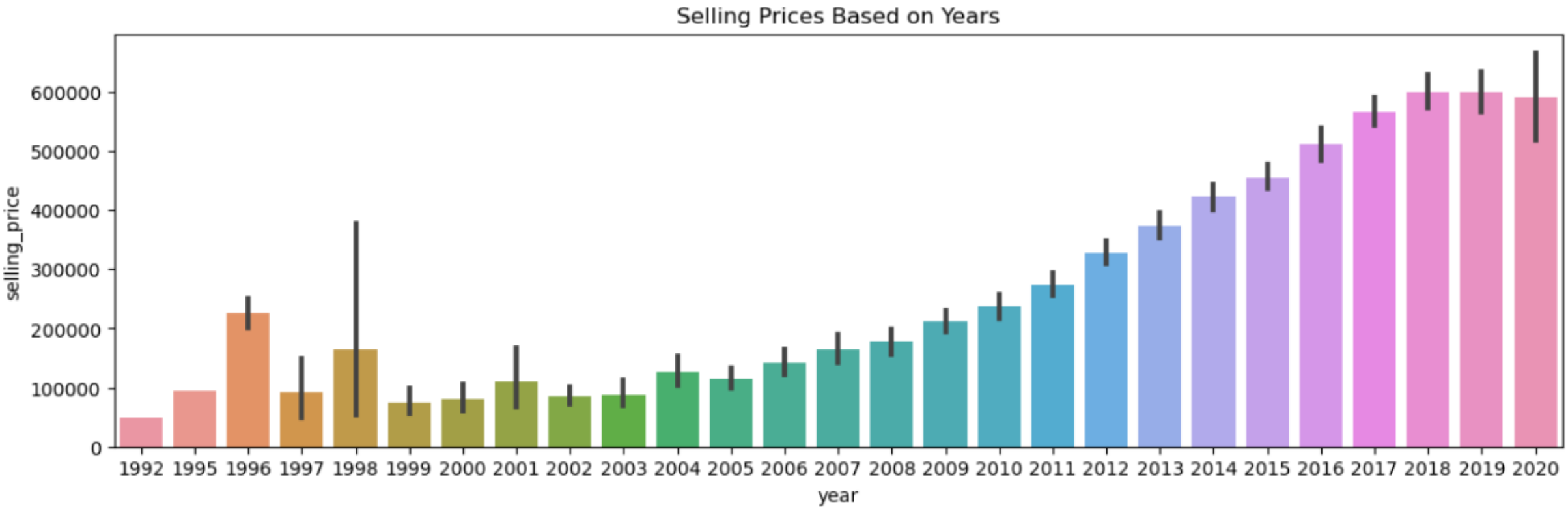
- The Data used in this project it consists 4340 Rows and 8 Columns with no null values.
- Column data consists of independent Features and dependent features.
- The independent Features contains both numerical and categorical data .
- In the data there are duplicate values.

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner

Exploratory Data Analysis(EDA)GRAPHS

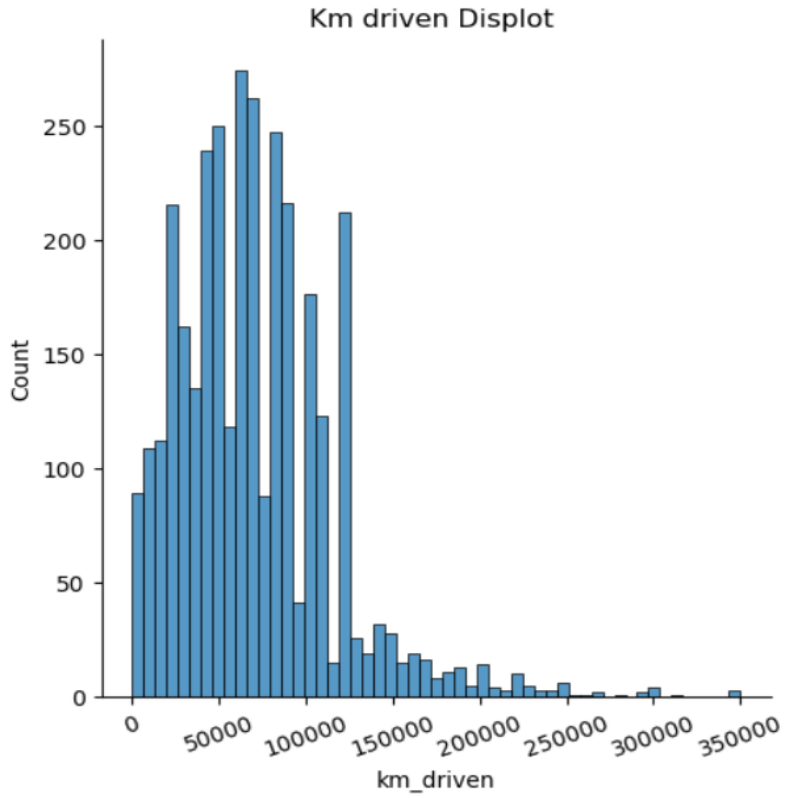




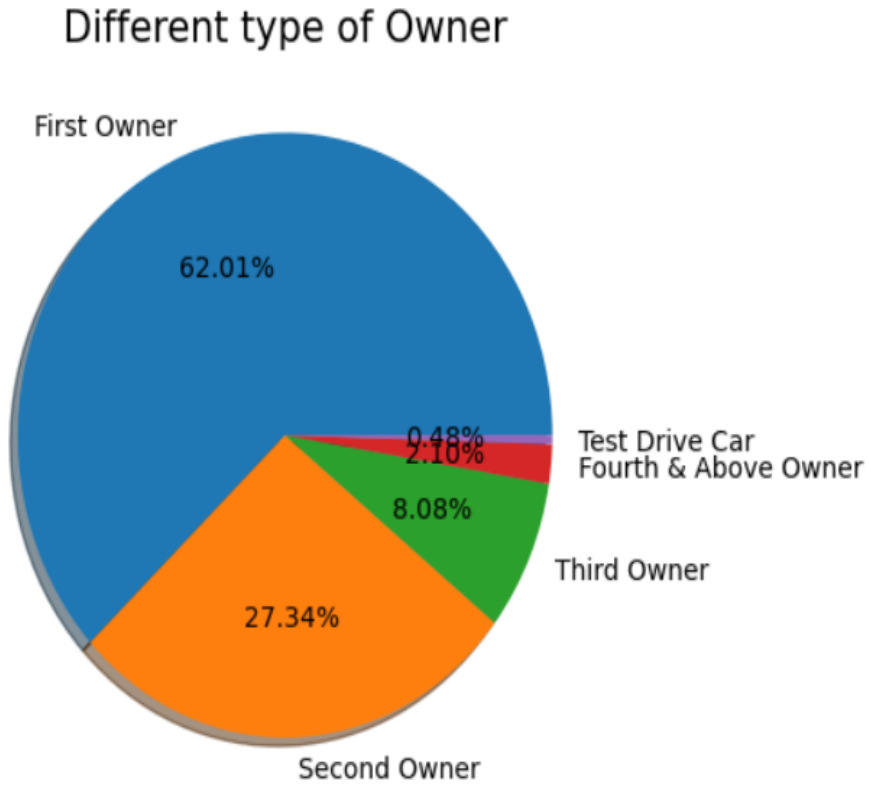


Based on above bar chart we can conclude that prices are decreasing for older cars

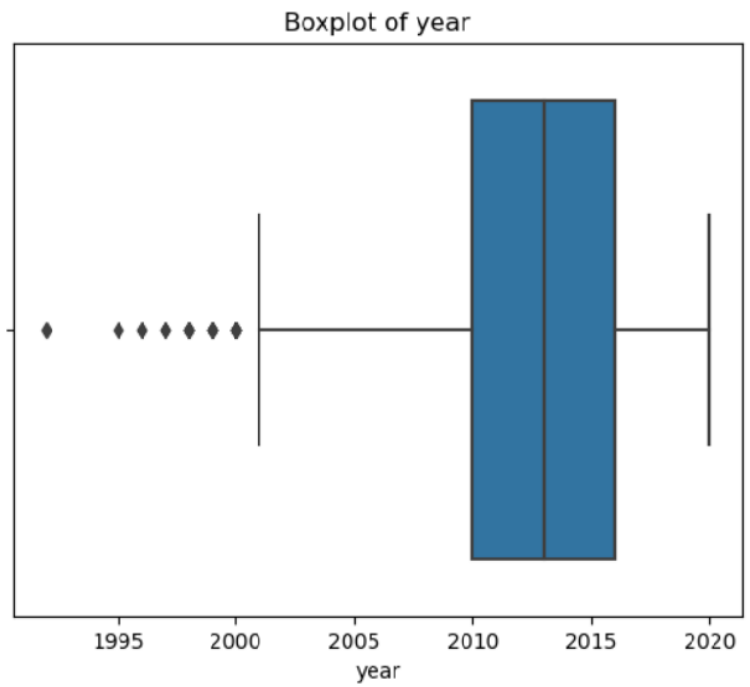
Graphical Analysis and conclusion on Data

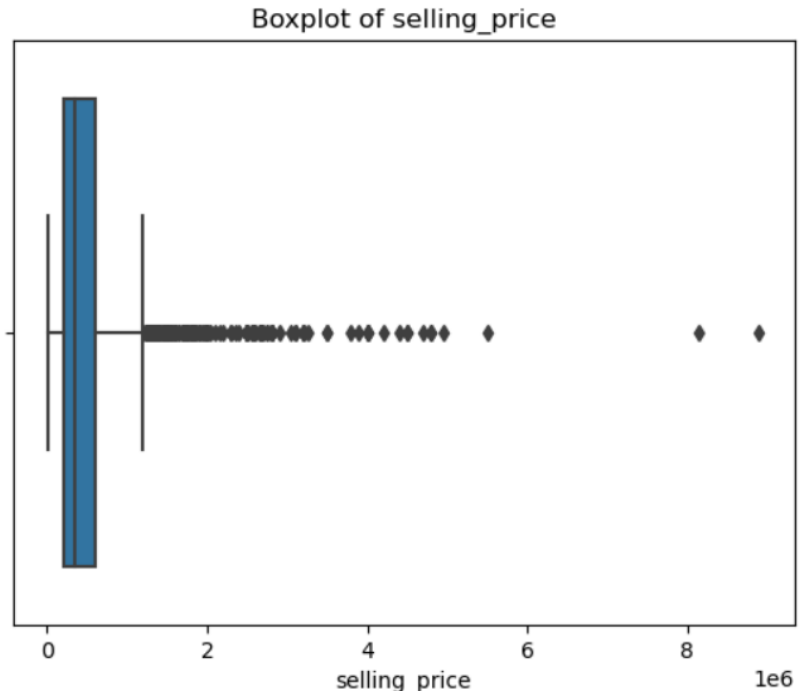


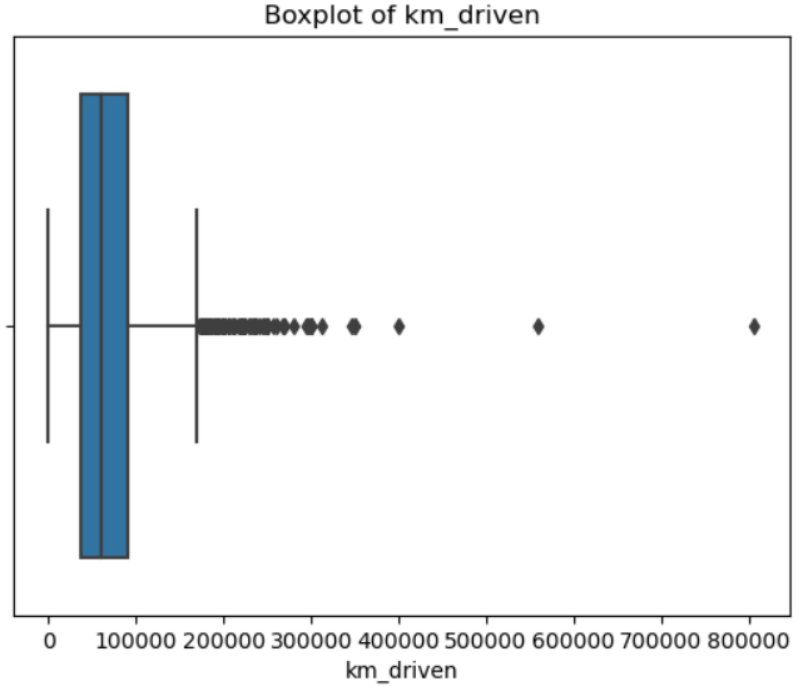
From above displot we can conclude that mostly cars are driven km between 0 to 2lacs km



From above pie chart we can conclude that Most of the car owners are First owners .







Outliers

1) There are outliers in year boxplot , we handle the outliers by removing years below 2000
2) There are outliers in Selling price boxplot, we removed the outlies
3) There are outliers in Km driven boxplot , we removed the outliers.

Data Cleaning and Pre-processing

1. Checked for null values ,Findings there is no null values

```
In [85]: ## Checking Null Values

         car_df.isnull().sum()

Out[85]: name             0
         year             0
         selling_price    0
         km_driven        0
         fuel             0
         seller_type      0
         transmission     0
         owner            0
         dtype: int64
```

1. Check Duplicate values ,Findings 763 Duplicate values
2. Dropped Duplicate values using Drop duplicate function

```
In [87]: #Checking Duplicate values

         car_df.duplicated().sum()

Out[87]: 763
```

There is 763 Duplicate values

```
In [88]: #Dropped the duplicates

         car_df.drop_duplicates(inplace= True)
```

- Created new column brand name by cleaning name column

```
## create new column of car brand name

car_df['brand name']=car_df['name'].str.split(' ').str.slice(0,1).str.join('')
car_df.head()
```

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	brand name
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner	Maruti
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner	Maruti
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner	Hyundai
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner	Datsun
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner	Honda

- Dropped the name column cause its not required

```
#Dropped the name column
car_df.drop('name',axis=1,inplace=True)
car_df.head()
```

	year	selling_price	km_driven	fuel	seller_type	transmission	owner	brand name
0	2007	60000	70000	Petrol	Individual	Manual	First Owner	Maruti
1	2007	135000	50000	Petrol	Individual	Manual	First Owner	Maruti
2	2012	600000	100000	Diesel	Individual	Manual	First Owner	Hyundai
3	2017	250000	46000	Petrol	Individual	Manual	First Owner	Datsun
4	2014	450000	141000	Diesel	Individual	Manual	Second Owner	Honda

- Modified the sequence of column

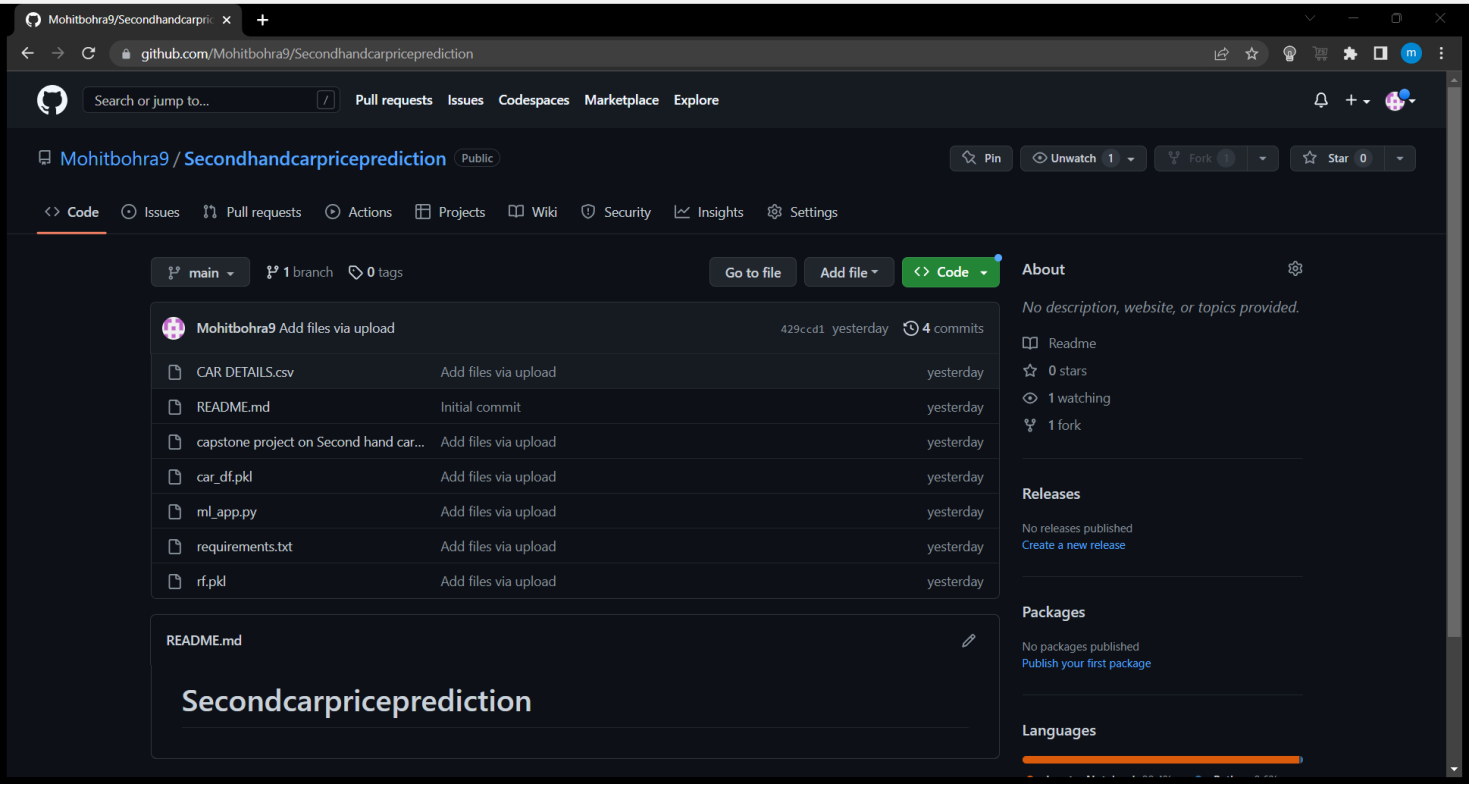
```
## Modify the sequce of columns
car_df=car_df.iloc[:,[7,0,1,2,3,4,5,6]]
car_df.head()
```

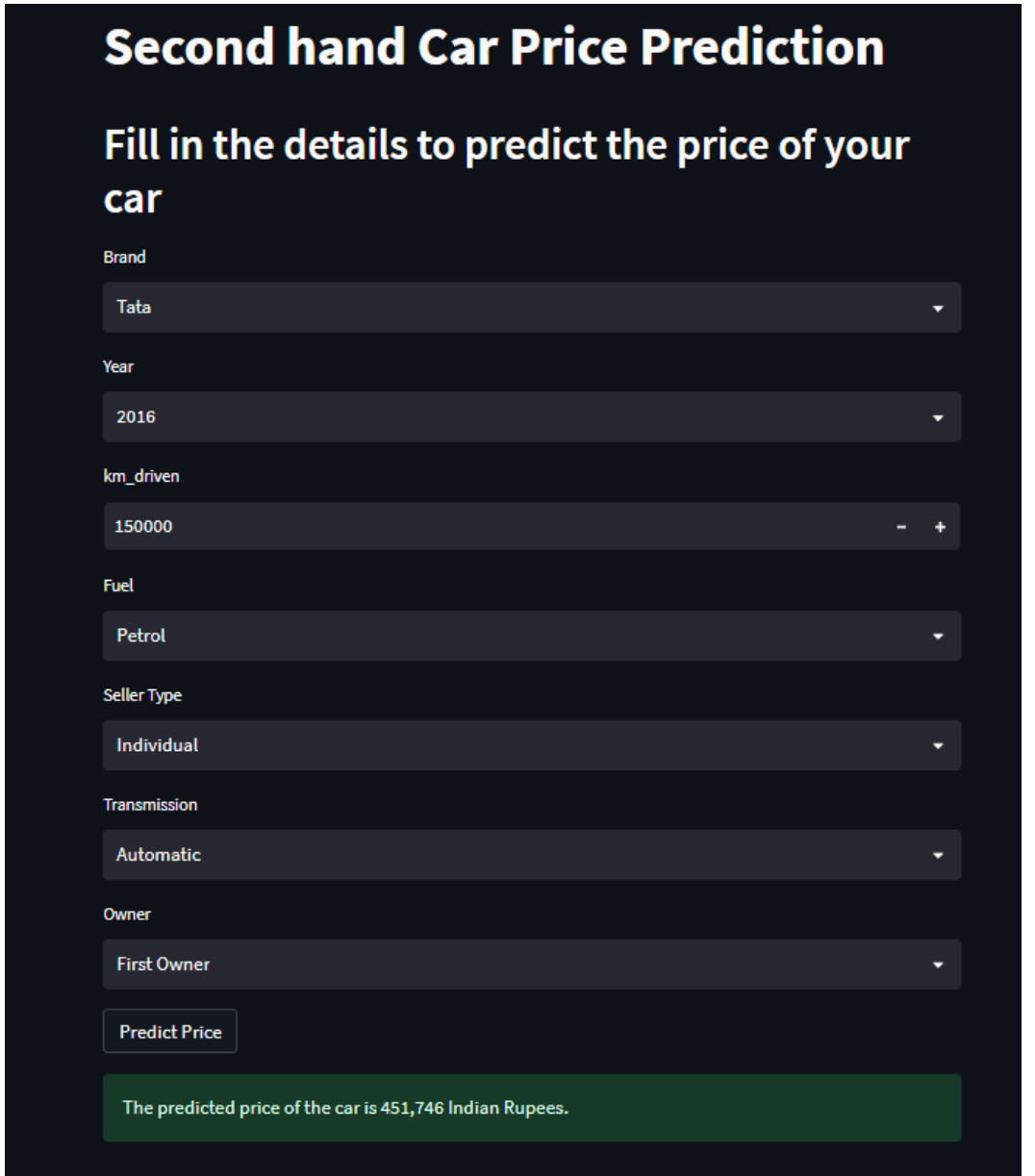
	brand name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	Maruti	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	Hyundai	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	Datsun	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	Honda	2014	450000	141000	Diesel	Individual	Manual	Second Owner

ML Modeling

- I have evaluated six different regression models and calculated their performance metrics, including MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R2 Score.
- Based on the R2 scores alone, we can see that the Random Forest (RF) Regression model had the highest R2 score of 0.6988, which indicates that it explains a relatively high proportion of the variability in the dependent variable.
- The Linear Regression, Ridge Regression, and Lasso Regression models all had similar R2 scores in the mid-0.6 range, indicating that they explain a moderate amount of the variability in the dependent variable.
- The KNN Regression model had a lower R2 score of 0.3982, indicating that it explains a relatively small proportion of the variability in the dependent variable.
- The Decision Tree (DT) Regression model had an R2 score similar to the Ridge and Lasso models, but with a slightly lower MAE and RMSE, suggesting it might be a good choice for this data set.
- However, it's important to note that the choice of the best model depends not only on the R2 score, but also on other factors such as the specific context and goals of the analysis, the interpretability of the model, and the computational resources available.
- **According to analysis the suited model for data base is the Random Forest (RF) Regression model with the highest R2 score of 0.6988**

•Deployment of ML Models using Streamlit.





End Notes

- Reference Links :
- Github link: https://github.com/Mohitbohra9/Secondhandcarpriceprediction.git

- Streamlit App Link:
- https://mohitbohra9-secondhandcarpriceprediction-ml-app-yqnycu.streamlit.app/