Report on the dataset used for [“Used Car Price Prediction” project]

We have used two datasets obtained from the Kaggle website for this project. The datasets are the train and the test datasets in the csv format. We have used different preprocessing and feature engineering methods to obtain the final dataset used for training the model.

The initial datasets combined -

	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1229	Volkswagen Vento Diesel Trendline	Hyderabad	2011	89411	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5.0	NaN	NaN
1230	Volkswagen Polo GT TSI	Mumbai	2015	59000	Petrol	Automatic	First	17.21 kmpl	1197 CC	103.6 bhp	5.0	NaN	NaN
1231	Nissan Micra Diesel XV	Kolkata	2012	28000	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5.0	NaN	NaN
1232	Volkswagen Polo GT TSI	Pune	2013	52262	Petrol	Automatic	Third	17.2 kmpl	1197 CC	103.6 bhp	5.0	NaN	NaN
1233	Mercedes-Benz E-Class 2009-2013 E 220 CDI Avan...	Kochi	2014	72443	Diesel	Automatic	First	10.0 kmpl	2148 CC	170 bhp	5.0	NaN	NaN

7253 rows × 13 columns

```
<class 'pandas.core.frame.DataFrame'>
Index: 7253 entries, 0 to 1233
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Name               7253 non-null   object
 1   Location           7253 non-null   object
 2   Year               7253 non-null   int64
 3   Kilometers_Driven  7253 non-null   int64
 4   Fuel_Type          7253 non-null   object
 5   Transmission       7253 non-null   object
 6   Owner_Type         7253 non-null   object
 7   Mileage            7251 non-null   object
 8   Engine             7207 non-null   object
 9   Power              7207 non-null   object
 10  Seats              7200 non-null   float64
 11  New_Price          1006 non-null   object
 12  Price              6019 non-null   float64
dtypes: float64(2), int64(2), object(9)
memory usage: 793.3+ KB
```

The dataset structure is shown

We perform the following steps -

1. Extracting “Brand” from “Name” feature. (in preprocessing.py)
2. Extracting “Model” from “Name” feature.
3. Converting “Year” into “Age” (Age of the car).
4. Removing the units from “Power”, “Mileage”, “Engine”.
5. Converting 0 values into NaN values.
6. Modifying the “New_Price” values and creating a new feature “new_price_num”
7. Replacing 0 seats with NaN values

The DataFrame obtained is -

	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price	Brand	Model	Age	new_price_num
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	3	21.01	998.0	58.16	5.0	NaN	1.75	Maruti	Wagon R	10	NaN
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	3	19.67	1582.0	126.20	5.0	NaN	12.50	Hyundai	Creta 1.6	5	NaN
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	3	18.20	1199.0	88.70	5.0	8.61 Lakh	4.50	Honda	Jazz V	9	8.61
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	3	20.77	1248.0	88.76	7.0	NaN	6.00	Maruti	Ertiga VDI	8	NaN
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	2	15.20	1968.0	140.80	5.0	NaN	17.74	Audi	A4 New	7	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1229	Volkswagen Vento Diesel Trendline	Hyderabad	2011	89411	Diesel	Manual	3	20.54	1598.0	103.60	5.0	NaN	NaN	Volkswagen	Vento Diesel	9	NaN
1230	Volkswagen Polo GT TSI	Mumbai	2015	59000	Petrol	Automatic	3	17.21	1197.0	103.60	5.0	NaN	NaN	Volkswagen	Polo GT	5	NaN
1231	Nissan Micra Diesel XV	Kolkata	2012	28000	Diesel	Manual	3	23.08	1461.0	63.10	5.0	NaN	NaN	Nissan	Micra Diesel	8	NaN
1232	Volkswagen Polo GT TSI	Pune	2013	52262	Petrol	Automatic	1	17.20	1197.0	103.60	5.0	NaN	NaN	Volkswagen	Polo GT	7	NaN
1233	Mercedes-Benz E-Class 2009-2013 E 220 CDI Avan...	Kochi	2014	72443	Diesel	Automatic	3	10.00	2148.0	170.00	5.0	NaN	NaN	Mercedes-Benz	E-Class 2009-2013	6	NaN

7253 rows × 17 columns

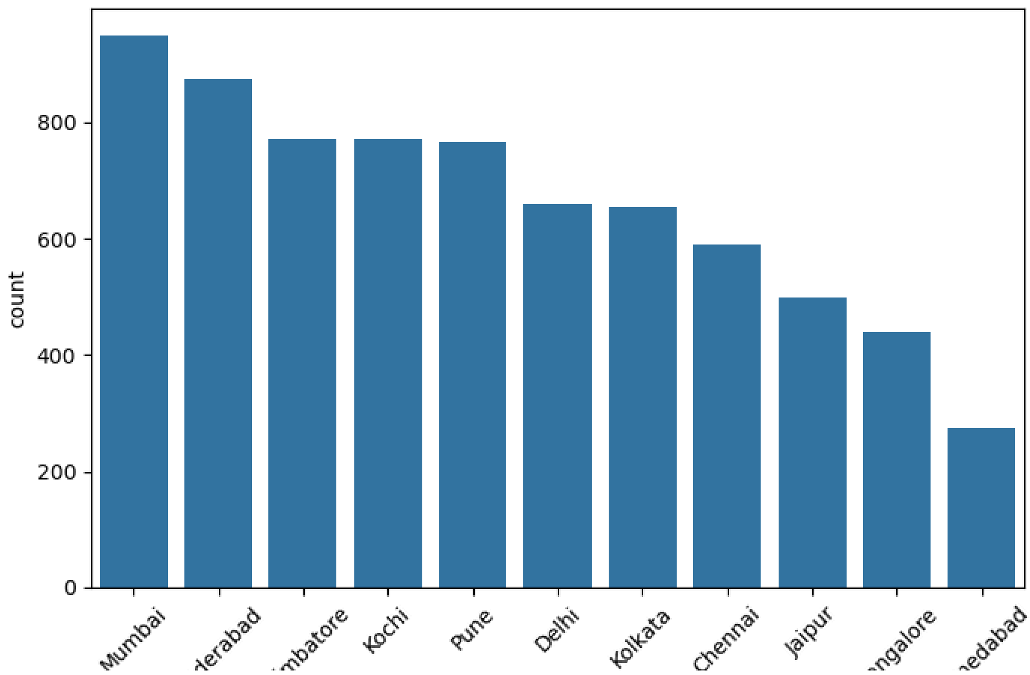
After properly indexing the DataFrame, we get this DataFrame used for EDA.

PTO.

S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price	Brand	Model	Age	new_price_num
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	3	21.01	998.0	58.16	5.0	1.75	Maruti	Wagon R	10	NaN
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	3	19.67	1582.0	126.20	5.0	12.50	Hyundai	Creta 1.6	5	NaN
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	3	18.20	1199.0	88.70	5.0	4.50	Honda	Jazz V	9	8.61
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	3	20.77	1248.0	88.76	7.0	6.00	Maruti	Ertiga VDI	8	NaN
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	2	15.20	1968.0	140.80	5.0	17.74	Audi	A4 New	7	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7248	Volkswagen Vento Diesel Trendline	Hyderabad	2011	89411	Diesel	Manual	3	20.54	1598.0	103.60	5.0	NaN	Volkswagen	Vento Diesel	9	NaN
7249	Volkswagen Polo GT TSI	Mumbai	2015	59000	Petrol	Automatic	3	17.21	1197.0	103.60	5.0	NaN	Volkswagen	Polo GT	5	NaN
7250	Nissan Micra Diesel XV	Kolkata	2012	28000	Diesel	Manual	3	23.08	1461.0	63.10	5.0	NaN	Nissan	Micra Diesel	8	NaN
7251	Volkswagen Polo GT TSI	Pune	2013	52262	Petrol	Automatic	1	17.20	1197.0	103.60	5.0	NaN	Volkswagen	Polo GT	7	NaN
7252	Mercedes-Benz E-Class 2009-2013 E 220 CDI Avan...	Kochi	2014	72443	Diesel	Automatic	3	10.00	2148.0	170.00	5.0	NaN	Mercedes-Benz	E-Class 2009-2013	6	NaN

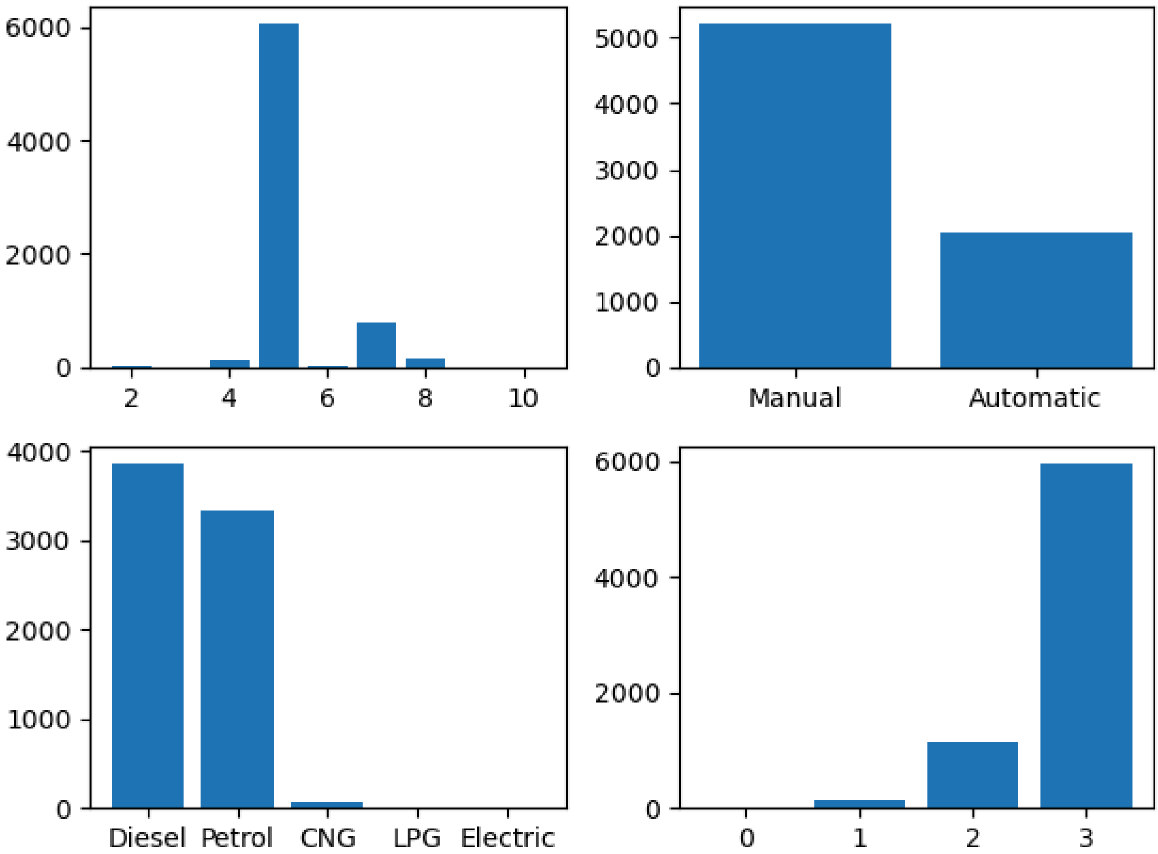
Exploratory Data Analysis (EDA) -

1) Making a decreasing order countplot for the no. of cars in a city.



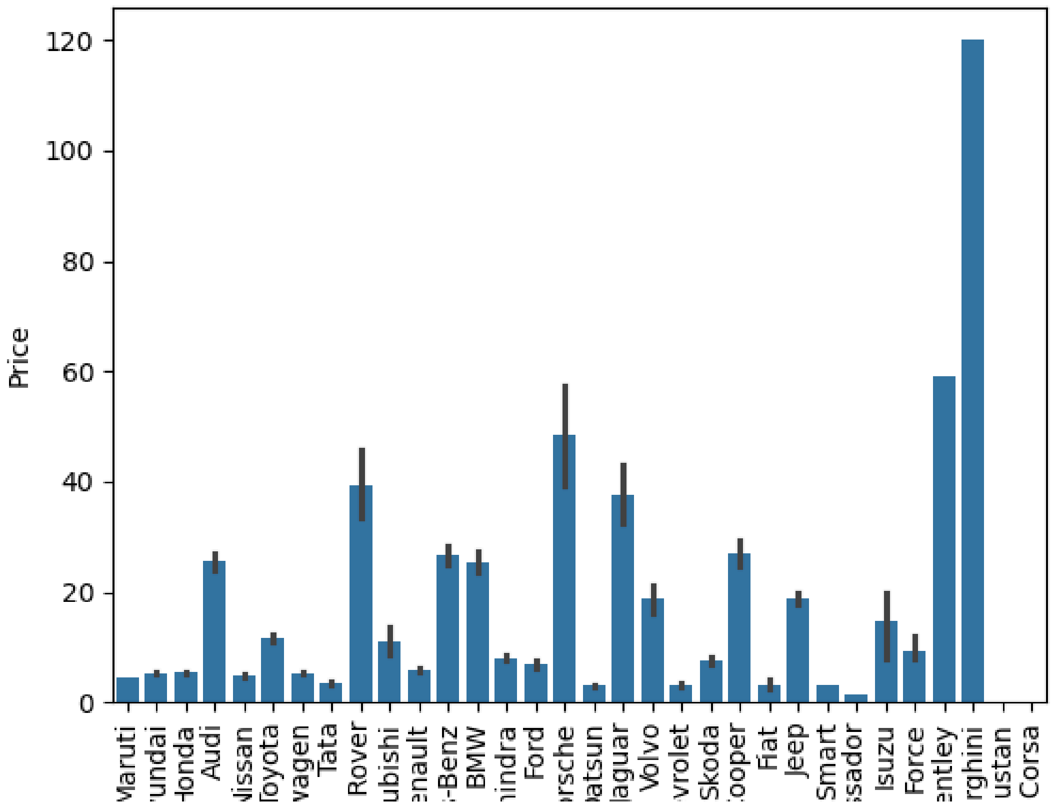
It is quite visible that Mumbai leads in the no. of used cars being sold, followed by Hyderabad.

2) Then we examine the rest of the count plots for categorical data type features.



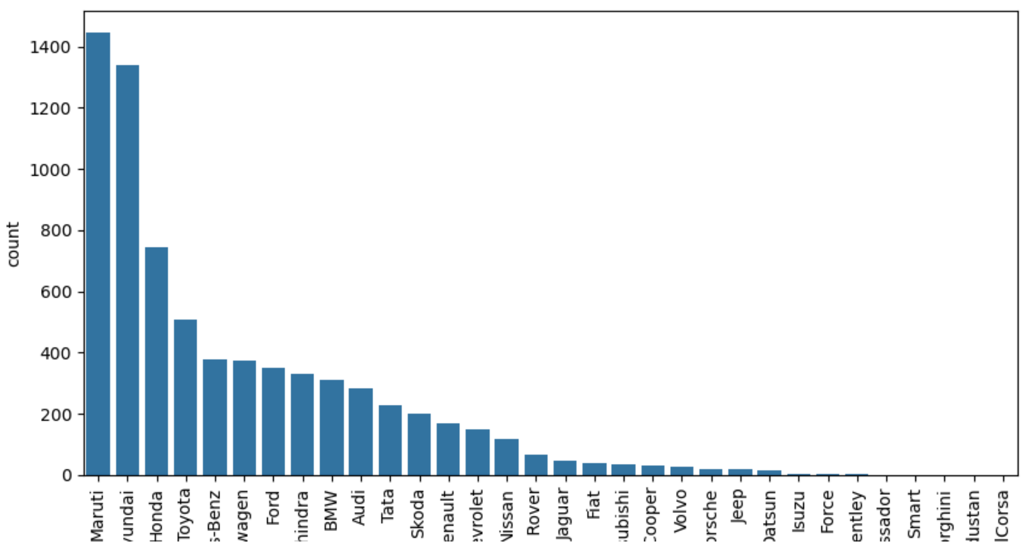
From these graphs, we can conclude that most of the cars are either 5 seaters or 7 seaters. Around 70% of the cars being sold have manual transmission. Diesel is the most common fuel type, just followed by Petrol. This shows that the used car market for Diesel is more than Petrol which might be due to commercial vehicles such as cabs. The last plot shows the owner type, 4 indicating first owner, 3 indicating second owner and so on. The cars are mostly first owner type of second owner type.

3) Next we examine the average price of the car corresponding to a car company.



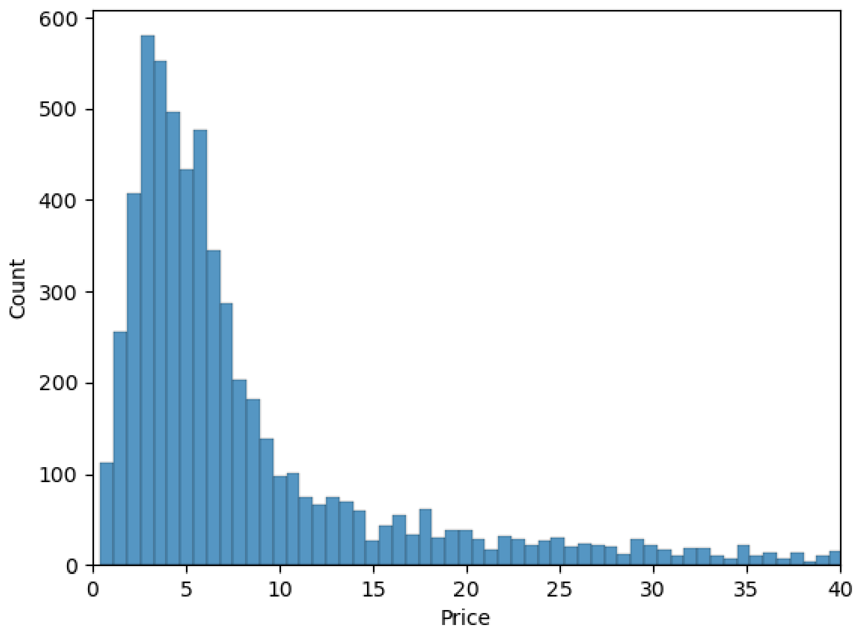
We see that we can segment the cars into 2 segments. One being the high value car brand, and the other to be low value car brand. We take the separating brand to be Isuzu.

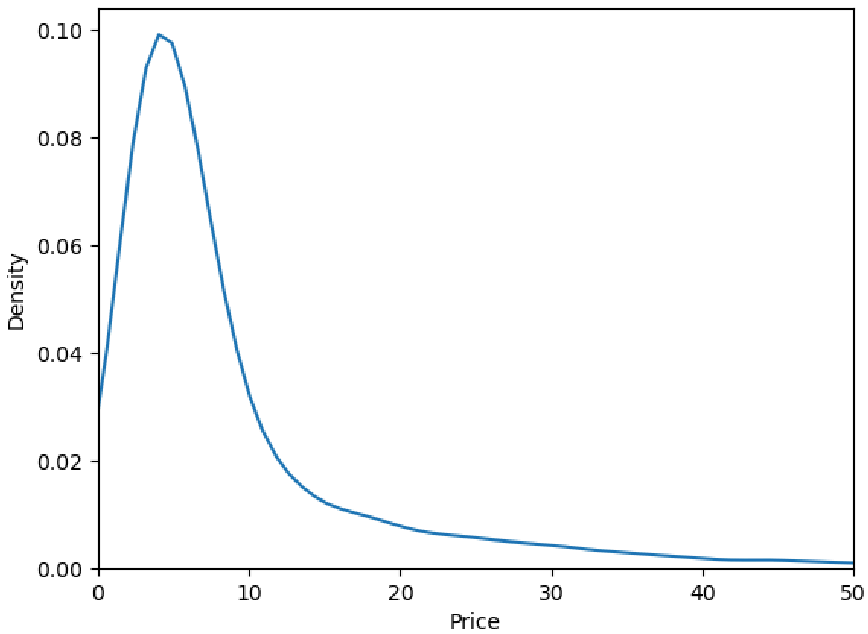
4) Next we make a countplot for the car brands to see their market share.



Maruti, Hyundai, Honda, Toyota are the major players in the market share.

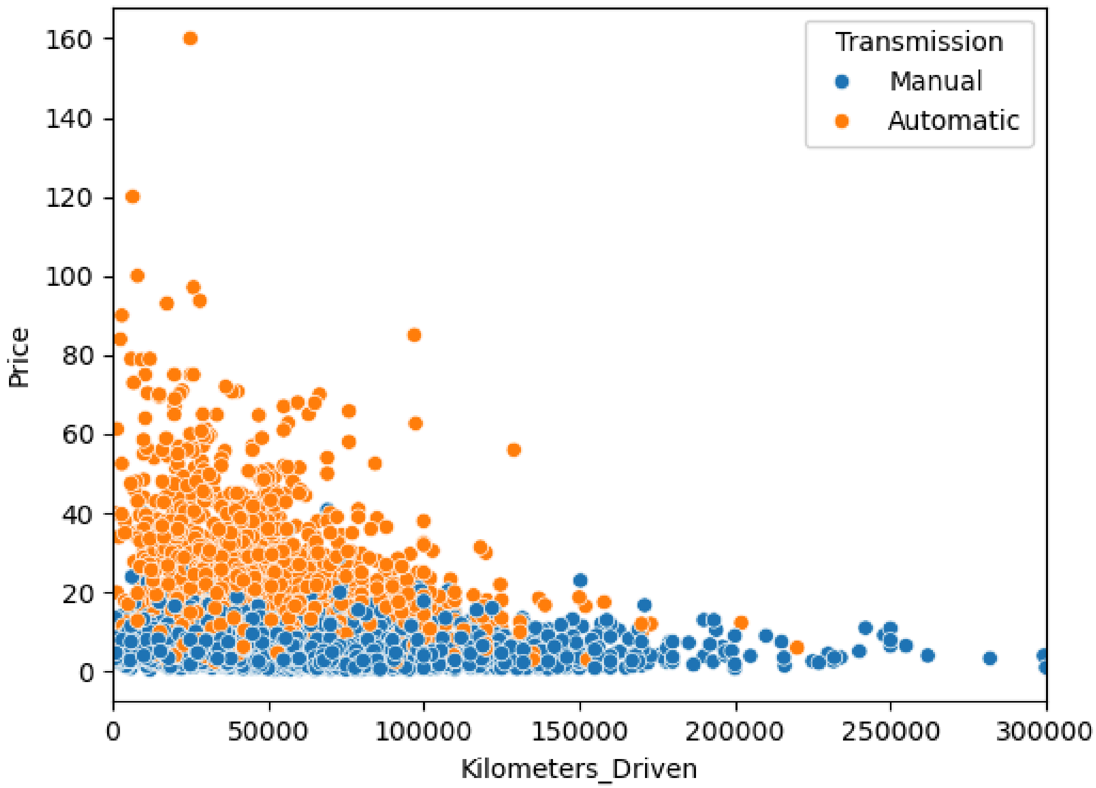
5) A histogram and KDE depicting the price distribution of the used cars -





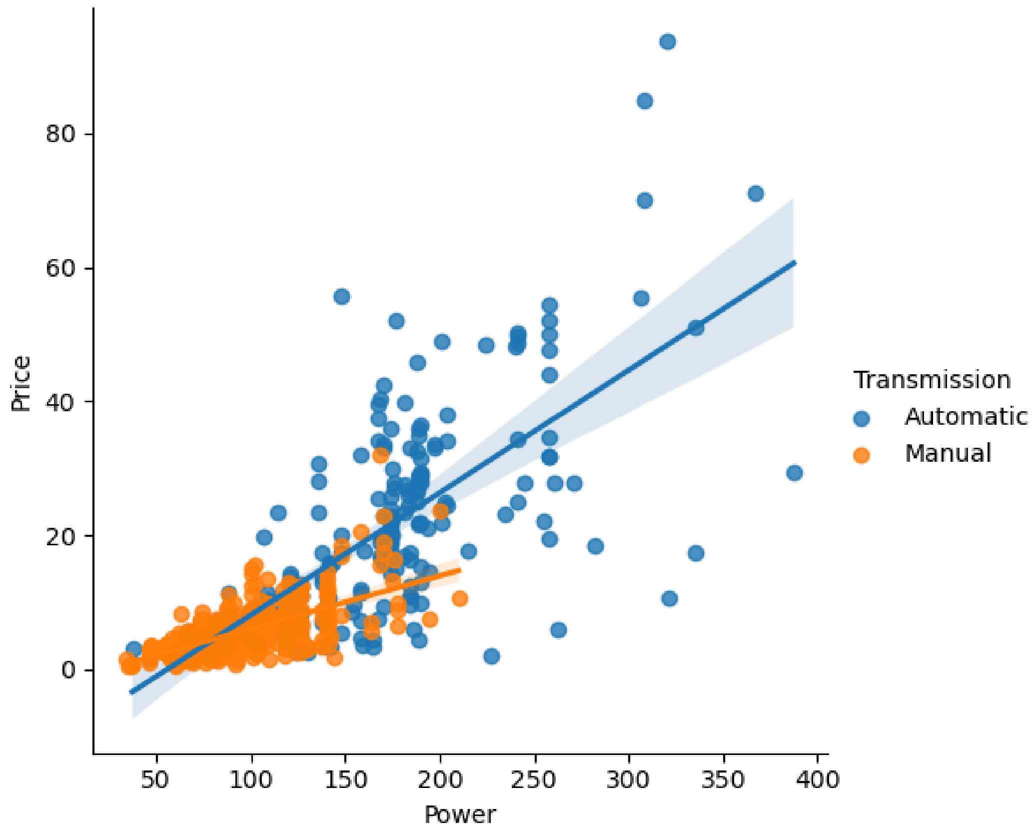
The price is depicted in INR Lakhs. So most of the distribution lies within 15 Lakhs. The distribution is right skewed.

6) Next we try to find a relation between the kilometres driven and the price of the car.



There seems to be a mild positive correlation between the kilometres driven and the price of the car.

7) We examine the relation between the power of the car and the price of the car.



There is a good amount of positive correlation between the two quantities. We have depicted the relation with a linear regression line with Transmission hue as well.

Final Correlation values with Price.

```
Mileage              -0.329418
Age                  -0.305065
Kilometers_Driven    -0.168299
Seats                 0.053645
Owner_Type            0.097392
Year                  0.305065
Engine                0.658102
Power                 0.772383
new_price_num         0.871847
Price                 1.000000
Name: Price, dtype: float64
```

Model Selection and final database -

We fill in the null values using median values of similar car types in the feature_engineering.py file. The module null_values helps us fill up missing Power, Engine, New Price values. We remove the null values that cannot be filled up and we drop the other irrelevant features. We obtain this database -

S.No.	Location	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price	Brand	Model	Age	new_price_num
0	Mumbai	72000	CNG	Manual	3	21.01	998.0	58.16	5.0	1.75	Maruti	Wagon R	10	5.29
1	Pune	41000	Diesel	Manual	3	19.67	1582.0	126.20	5.0	12.50	Hyundai	Creta 1.6	5	16.06
2	Chennai	46000	Petrol	Manual	3	18.20	1199.0	88.70	5.0	4.50	Honda	Jazz V	9	8.61
3	Chennai	87000	Diesel	Manual	3	20.77	1248.0	88.76	7.0	6.00	Maruti	Ertiga VDI	8	11.27
4	Coimbatore	40670	Diesel	Automatic	2	15.20	1968.0	140.80	5.0	17.74	Audi	A4 New	7	53.14

After thorough examination of the models, we see that the features Owner_Type, Seats and Mileage do not affect the price of the car, hence we drop them.
Next we perform one hot encoding for the categorical data types.
The database obtained is -

S.No.	Kilometers_Driven	Engine	Power	Price	Age	new_price_num	Location_Bangalore	Location_Chennai	Location_Coimbatore	Location_Delhi	...	Model_
0	72000	998.0	58.16	1.75	10	5.290	0.0	0.0	0.0	0.0	...	
1	41000	1582.0	126.20	12.50	5	16.060	0.0	0.0	0.0	0.0	...	
2	46000	1199.0	88.70	4.50	9	8.610	0.0	1.0	0.0	0.0	...	
3	87000	1248.0	88.76	6.00	8	11.270	0.0	1.0	0.0	0.0	...	
4	40670	1968.0	140.80	17.74	7	53.140	0.0	0.0	1.0	0.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
7248	89411	1598.0	103.60	NaN	9	10.940	0.0	0.0	0.0	0.0	...	
7249	59000	1197.0	103.60	NaN	5	10.830	0.0	0.0	0.0	0.0	...	
7250	28000	1461.0	63.10	NaN	8	15.060	0.0	0.0	0.0	0.0	...	
7251	52262	1197.0	103.60	NaN	7	11.045	0.0	0.0	0.0	0.0	...	
7252	72443	2148.0	170.00	NaN	6	49.490	0.0	0.0	0.0	0.0	...	

7024 rows × 710 columns

The shape of the database is 7024 rows x 710 columns. This can be attributed to the large no. of models that are considered for this analysis.

We scale our data and perform grid search for multiple regression models such as Linear Regression, Elastic Net Regression, Random Forests

Regression, XGBoost Regression. After the analysis, we find that XGBoost is the best performing model with the following hyper-parameters - (learning_rate=0.1,max_depth=5,n_estimators=300).

We compare models on the basis of Mean Absolute Error and Root Mean Squared Error. We get the following errors for different combinations of feature selection. We perform a train test split to find the errors with (10% test size).

MAE -
#mean_absolute_error(y_test,y_pred)
without owner, seats - 1.3612876790557822
without owner, seats, mileage, model - 1.3755960163065832
without owner, seats, model - 1.3510027499394874
without owner, seats, model, location - 1.6612586461415846
without owner, seats, mileage - 1.342713504907203
without owner, seats, mileage, location - 1.593179432178197

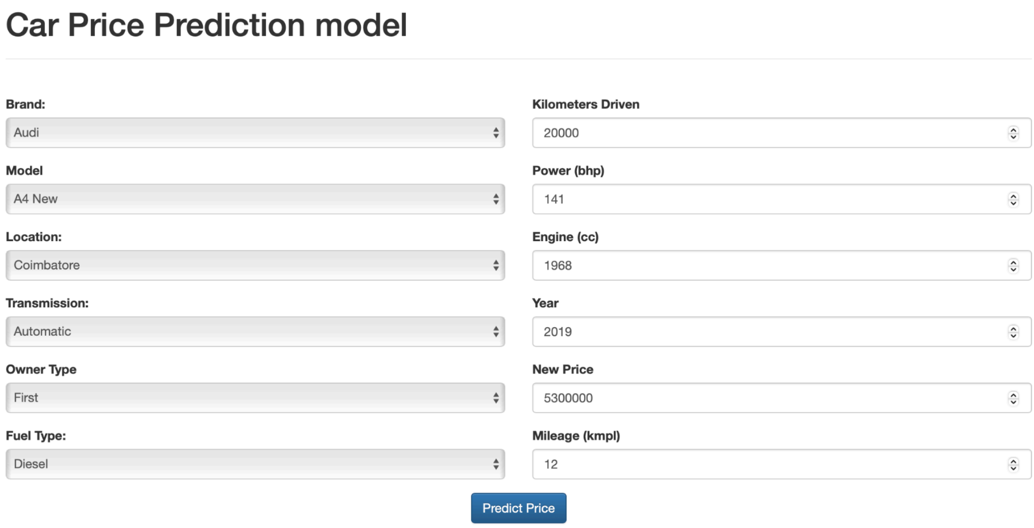
RMSE -
#np.sqrt(mean_squared_error(y_test,y_pred))
without owner, seats - 4.251735228502176
without owner, seats, mileage, model - 3.8843320467104454
without owner, seats, model - 4.2311705631846825
without owner, seats, model, location - 5.277752116467163
without owner, seats, mileage - 3.925457564011466
without owner, seats, mileage, location - 4.312418092486159

After this, we shall train our model on the full training dataset.
The model score is 0.9861

We have created the predict module to help predict prices when we are using the Flask API.

Next we created the Flask API, HTML pages and then Javascript pages to complete the User Interface of the project. Sample outputs are shown -

The input page -



Car Price Prediction model

Brand: Audi

Model A4 New

Location: Coimbatore

Transmission: Automatic

Owner Type First

Fuel Type: Diesel

Kilometers Driven 20000

Power (bhp) 141

Engine (cc) 1968

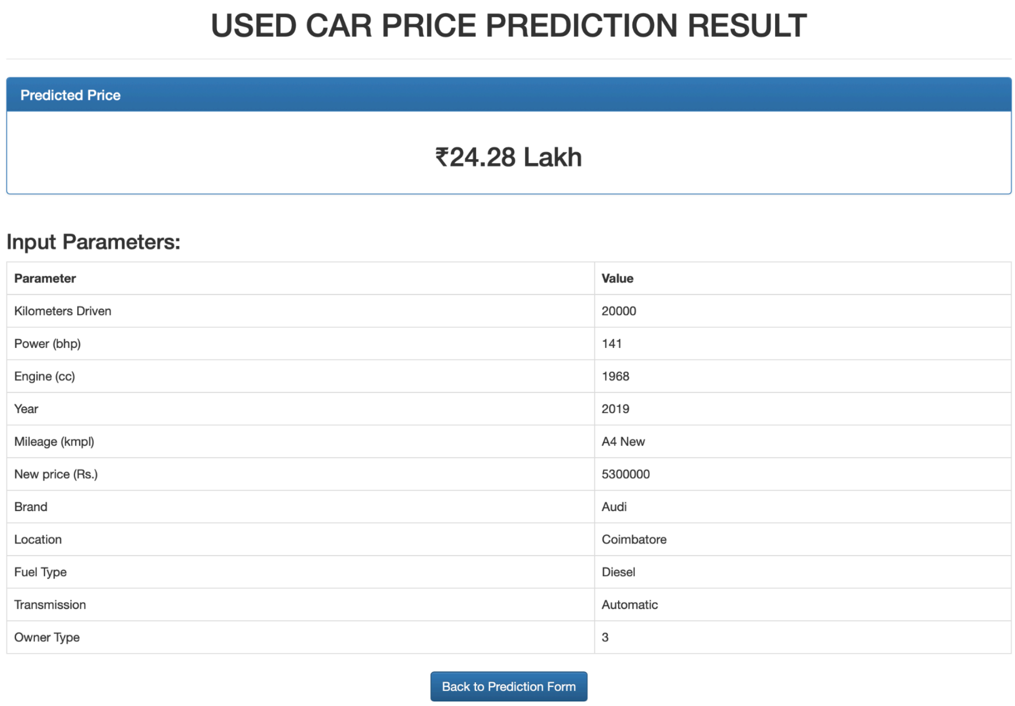
Year 2019

New Price 5300000

Mileage (kmpl) 12

Predict Price

The output page -



USED CAR PRICE PREDICTION RESULT

Predicted Price

₹24.28 Lakh

Input Parameters:

Parameter	Value
Kilometers Driven	20000
Power (bhp)	141
Engine (cc)	1968
Year	2019
Mileage (kmpl)	A4 New
New price (Rs.)	5300000
Brand	Audi
Location	Coimbatore
Fuel Type	Diesel
Transmission	Automatic
Owner Type	3

Back to Prediction Form

We have saved all the modules and other pkl files which might be needed later for scaling and other prediction purposes.

Sources of Error -

The major challenge of analysing this dataset and using it for building an ML model was the absence of clean data. The dataset available is fairly small, and had many missing values for key model features which had to be filled in using median values. This is the main source of error in this model.