

Used Car Price Prediction

Submitted by:

Gaurav More

# ACKNOWLEDGMENT

I express my sincere gratitude to Flip Robo Technologies for giving me the opportunity to work on this project on Used Car Price Prediction using machine learning algorithms. I acknowledge my indebtedness to the authors of online articles titled: “Factors which affect used car valuation price” and “Just What Factors Into The Value Of Your Used Car?” for providing me with invaluable insights and knowledge of the various factors that determine the price of a used car.

# INTRODUCTION

### Business Problem Framing

With the covid 19 impact in the market, there have been lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. With the change in market due to covid 19 impact, our client there is a need for new machine learning models from new data.

Therefore, new car price valuation model is required to be made.

### Conceptual Background of the Domain Problem

Predictive modelling, Regression algorithms are some of the machine learning techniques used for predicting Used Car prices. Identifying various relevant features of a car, its present condition, ownership and total usage by previous owner(s) are crucial for working on the project as they determine its valuation.

### Review of Literature

3 Online Articles, namely: “Factors which affect used car valuation price” and “Just What Factors Into The Value Of Your Used Car?” were reviewed and studied to gain insights into all the attributes that contribute to the price of a used car.

It is learnt that Economic Factors, Vehicle Make, Vehicle Class and Body Style, Mileage, Transmission Type and technology are some the most important factors that determine a used car’s valuation.

* [https://www.moneycrashers.com/factors-affect-used-cars-resale- value/](https://www.moneycrashers.com/factors-affect-used-cars-resale-value/)
* [https://www.investopedia.com/articles/investing/090314/just-what- factors-value-your-used-car.asp](https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp)
* [https://autoportal.com/articles/factors-which-affect-used-car- valuation-price-6446.html](https://autoportal.com/articles/factors-which-affect-used-car-valuation-price-6446.html)

### Motivation for the Problem Undertaken

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now there is a boom in demand for cars in the market, hence making them costly, while some are not in demand hence cheaper. There is a need for building machine learning models from new data that would help accurately predict the valuation of used cars based on various attributes.

# Analytical Problem Framing

### Mathematical/ Analytical Modeling of the Problem

Various Regression analysis techniques were used to build predictive models to understand the relationships that exist between used car price and features and attributes of the car. The Regression analysis models were used to predict the car price value for changes in car attributes. Regression modelling techniques were used in this Problem since Car Price data distribution is continuous in nature.

In order to forecast car price, predictive models such as ridge regression Model, Random Forest Regression model, Decision tree Regression Model, Support Vector Machine Regression model, Extreme Gradient Boost Regression and K Nearest Neighbours model were used to describe how the values of Car Price depended on the independent variables of various Car attributes.

### Data Sources and their formats

The Dataset was compiled by scraping Data for various Car attributes and Price from <https://droom.in/>.

The data was converted into a Pandas Dataframe under various Feature and Label columns and saved as a .csv file.



#### Dataset Description

###### The Independent Feature columns are:

* + Brand Name: Name of the Car Brand
  + Model Name: Name of the specific Car Model of a Brand
  + Year: Year of Manufacture
  + Type: Car Model Type
  + Total Kilometers Driven: Total Kilometers,for which the car has been so far driven.
  + Fuel Type: Type of fuel used
  + Owner: The ordinal number of previous owner
  + Location: Availability Location of Car
  + Transmission: Power Transmission type of the car
  + Seating Capacity: Total number of passengers that the car can accomodate
  + Mileage: Mileage of the Car

###### Target / Label Column:

* + Price: Selling Price of the Car

### Data Preprocessing Done

* + Rows containing incorrectly entered data were removed first.
  + Data in columns: Seating Capacity,Mileage,Price,Total Kilometers Driven,Year were converted to int64 / float data type.
  + There were rows containing a character: ‘-‘ as data. These entries were first converted into NaN values and then imputed with data using most frequently occurring value imputation and KNN imputation techniques.
  + Duplicate data elements which had their starting letters in upper case and lower case were converted to data elements starting with uppercase letters.
  + Columns: Unnamed: 0(just a series of numbers) was dropped since it doesn't contribute to building a good model for predicting the target variable values.

#### Feature Engineering:

* + In order to better understand the relationships between Car price and Car attributes, ‘Brand’,’Model’ and ‘Variant’ columns were created based on data of existing columns: ‘Brand Name’ and ‘Model Name'.
  + Column ‘Car Age’ was created based on data from ‘Year’.

### Data Inputs- Logic- Output Relationships

* The Datasets consist mainly of Float and Object data type variables. The relationships between the independent variables and dependent variable were analysed.

### Hardware and Software Requirements and Tools Used

#### Hardware Used:

* Processor: AMD Ryzen 9 5900HX(8 Cores 16 Logical Processors)
* Physical Memory: 16.0GB (3200MHz)
* GPU: Nvidia RTX 3060 (192 bits), 6GB DDR6 VRAM, 3840 CUDA cores.

#### Software Used:

* Windows 10 Operating System
* Anaconda Package and Environment Manager: Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows and provides a host of tools and environment for conducting Data Analytical and Scientific works. Anaconda provides all the necessary Python packages and libraries for Machine learning projects.
* Jupyter Notebook: The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.
* Python3: It is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best languages used for Data Analytics And Data science projects/application. Python provides numerous libraries to deal with mathematics, statistics and scientific function.
* Python Libraries used:
* Pandas: For carrying out Data Analysis, Data Manipulation,

Data Cleaning etc

* Numpy: For performing a variety of operations on the datasets.
* matplotlib.pyplot, Seaborn: For visualizing Data and various relationships between Feature and Label Columns
* Scipy: For performing operations on the datasets
* Statsmodels: For performing statistical analysis
* sklearn for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation,Data Scaling, Component analysis,Feature selection etc.

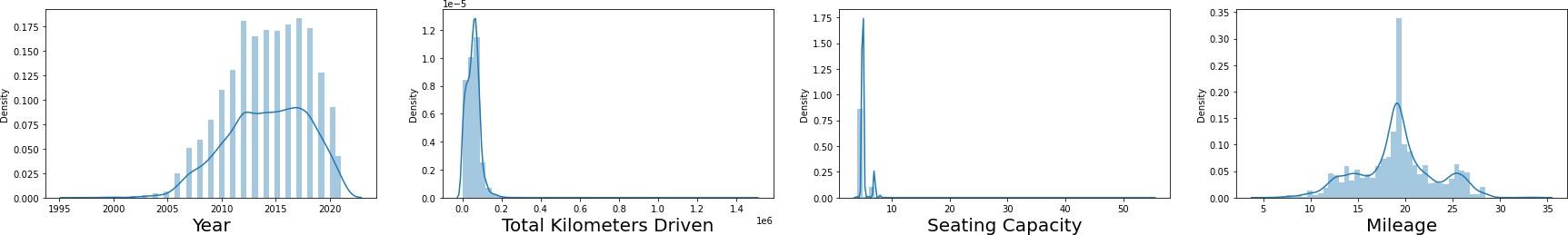
### Exploratory Data Analysis

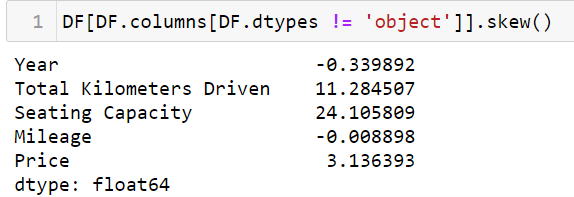
**Visualizations** Barplots,Distplots,Boxplots,Countplots,lineplots were used to visualise the data of all the columns and their relationships with Target variable.

#### Univariate Analysis Analyzing the Target Class

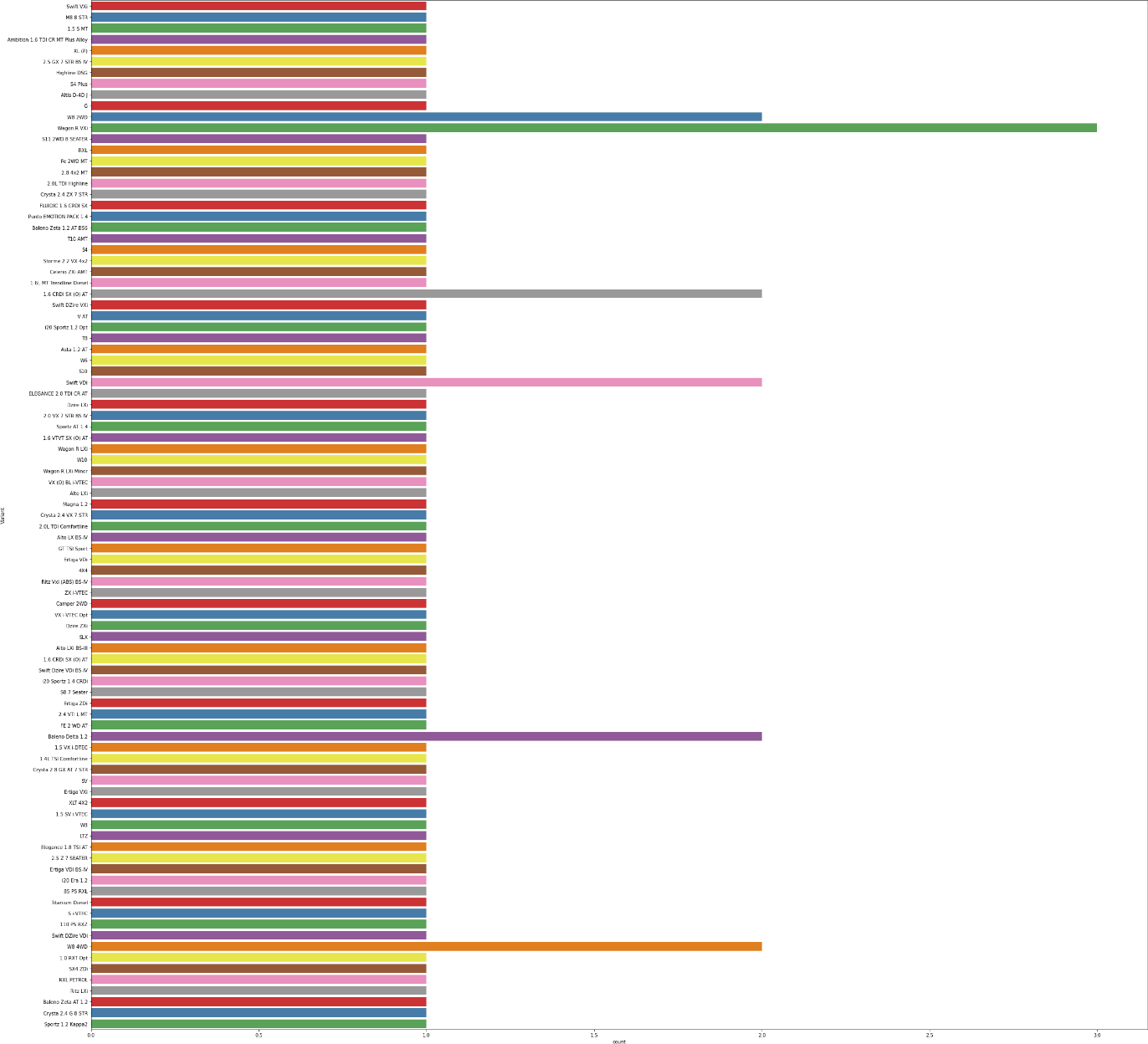
From the graph above it is observed that the Price data forms a continuous distribution and the distribution is skewed.

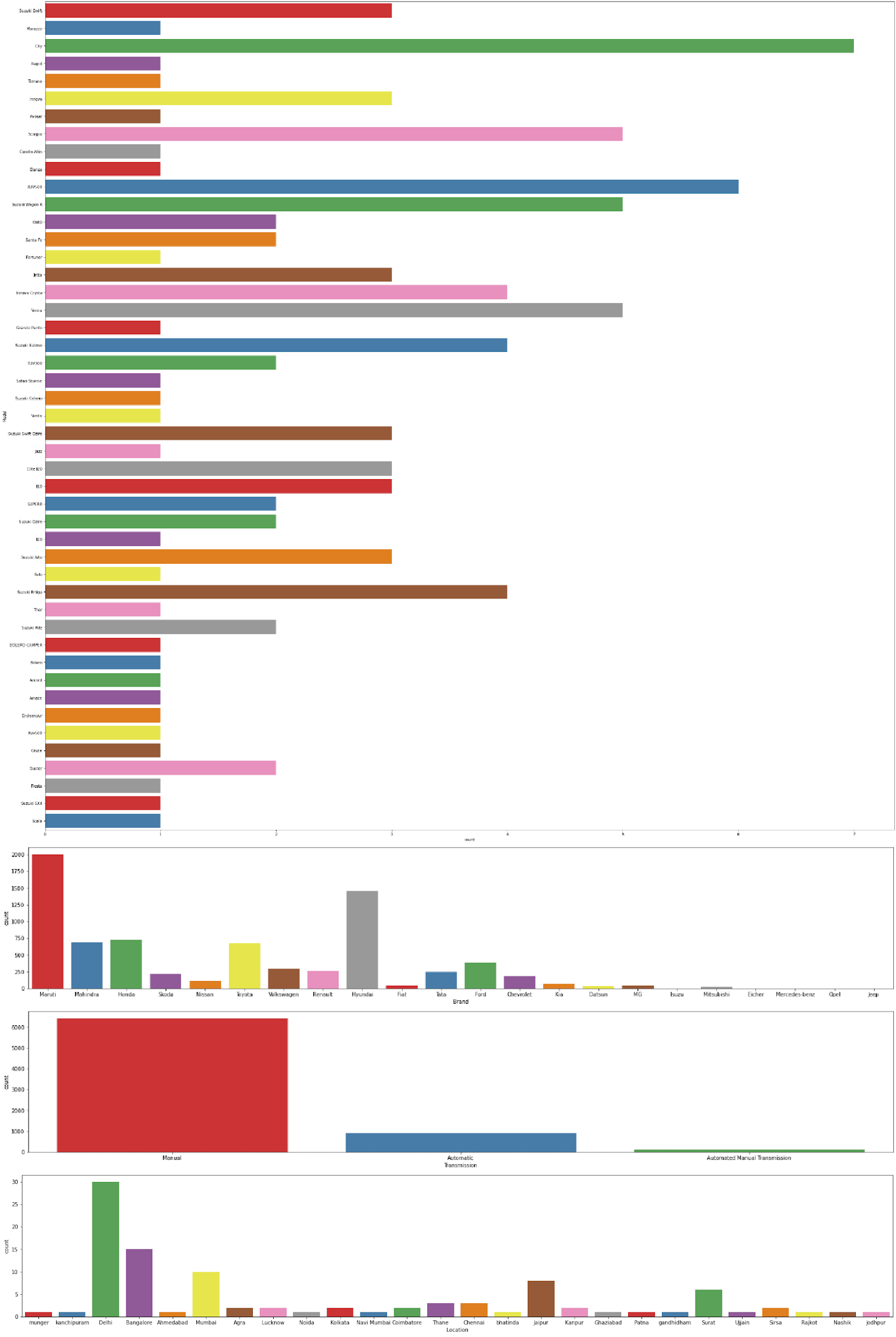
#### Analyzing the Feature Columns

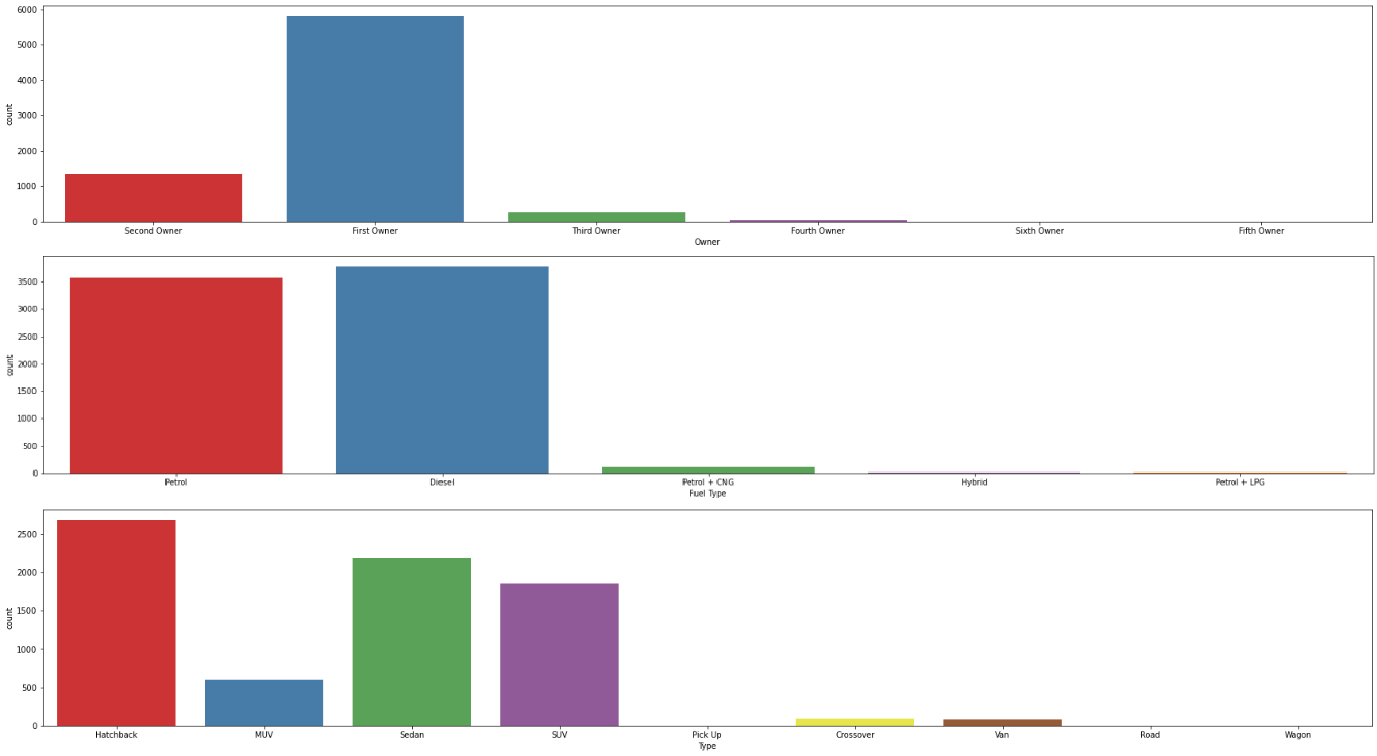




Year and Mileage columns look normally distributed, while Total kilometers Driven and Seating Capacity are skewed



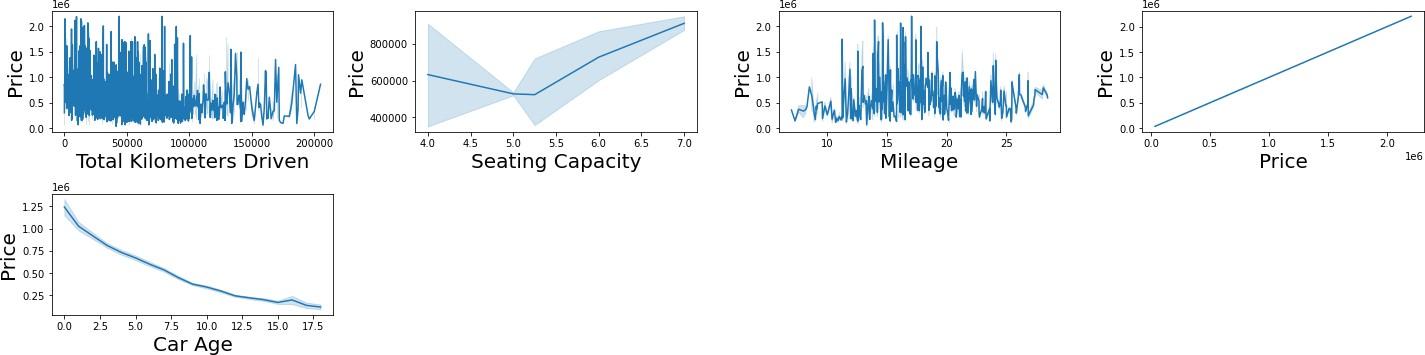




From the above graphs it is observed that:

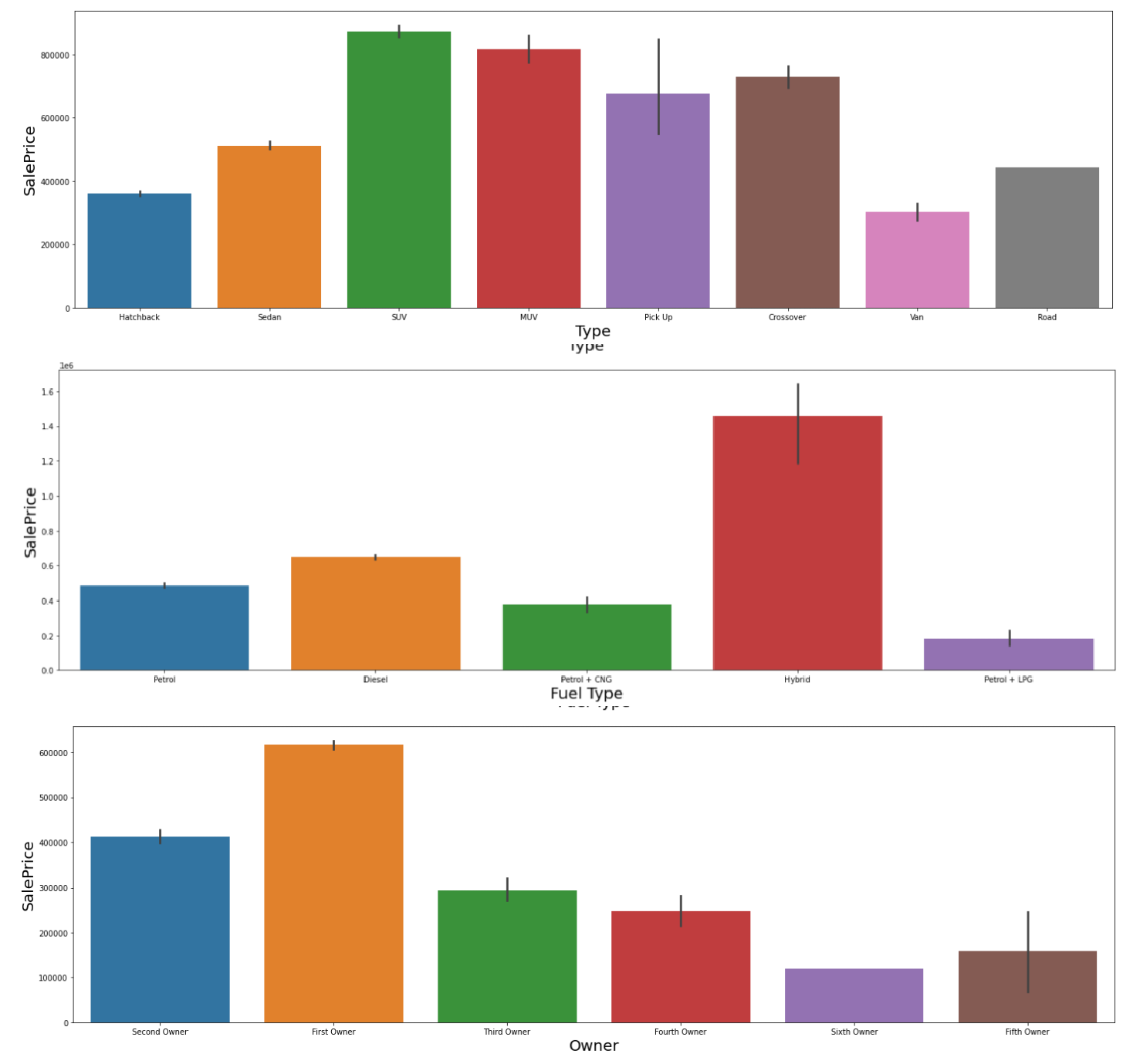
* Swift VDi,Swift DZire VDi,Swift VXi,Wagon R LXi,Alto LXi are the most common used cars on sale
* Maruti and Hyundai are the most common brands of used cars
* Manual Transmission is the most common amongst cars
* Most used cars on sale are located in Delhi and Bangalore
* Most used cars on sale have only had 1 owner before
* Petrol and Diesel are the most common fuel types
* Hatchback,Sedan and SUV are the most common Car types available

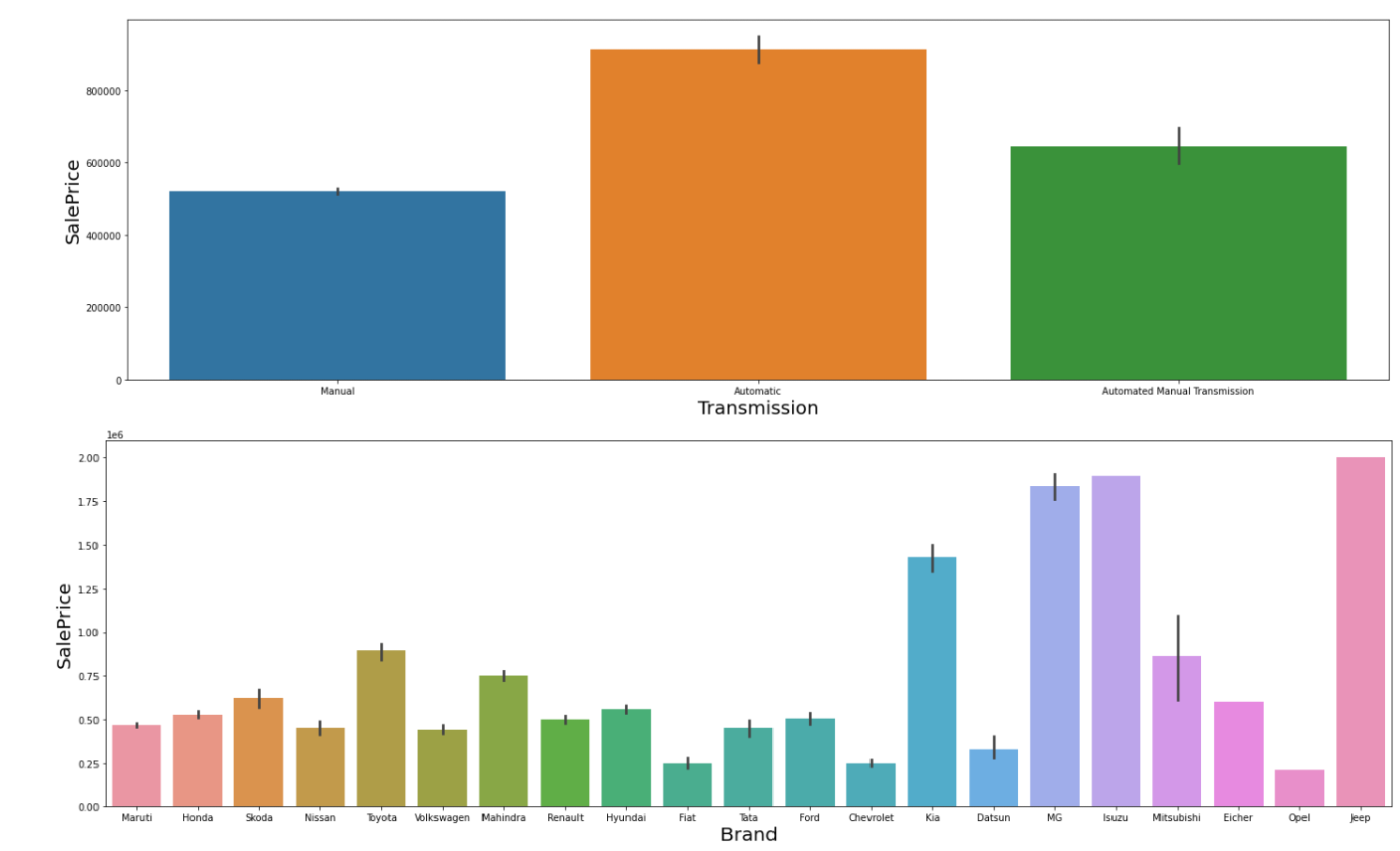
#### Bivariate Analysis

**Interpreting Relationship between Dependent Variable and Independent Variable Columns**

Following Observations are made from graphs above:

* + There is a negative correlation between Total Kilometers Driven and Price
  + Cars with Mileage between 14 km/l and 19 km/l have the highest prices
  + There is a negative correlation between Car Age and Price
  + There is a positive correlation between Seating capacity and Price

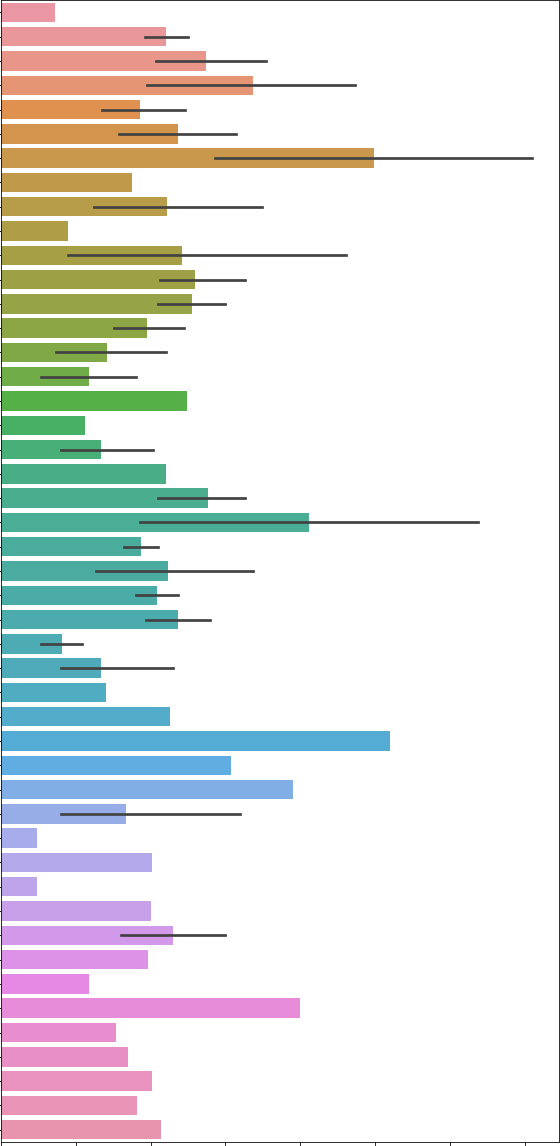








Ahmedabad

Thane Dennai Lhatinda

Kanpur

Patna

gandhidh am

gjain

]odhp ur Faridabad

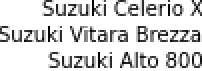
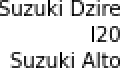
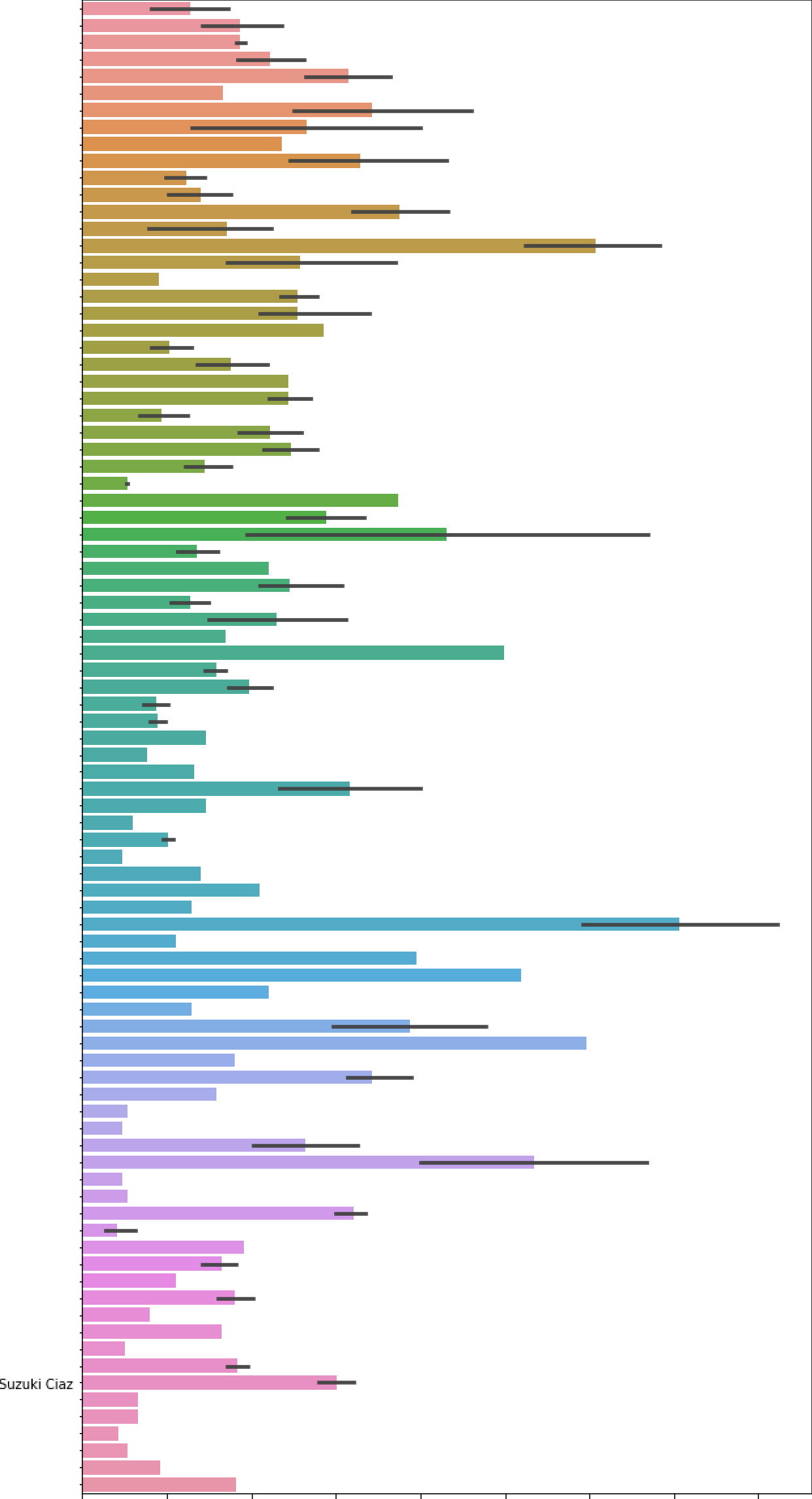


Pune angul

D andigarh

Haflong Nagaon Kanpur Nagar De hradun

000 025 050 075 100 125 150 175

Rapid

1 n nova

XUV500

Santa Fe 1 n nova Crysta

trna

BUZ IN k 1 BE IP FI O

s 00

safari ,

Into

Suz u k i Swift Dzire

Elite 12 0

Suzuki SX4  Suzuki Sx4

Avventura Fluence Suzuki A-Star

Suzuki Eeco

Venue

Suz u k i Omni

Santro King Suz u k i Alto f'l0

Scent lane a

Suzuki Y/agon R Duo

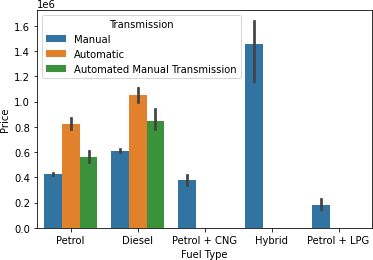
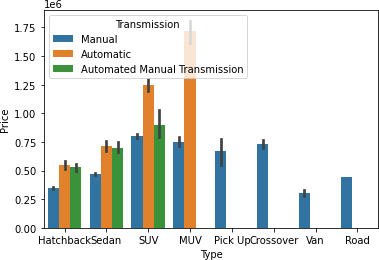
Punto Beat

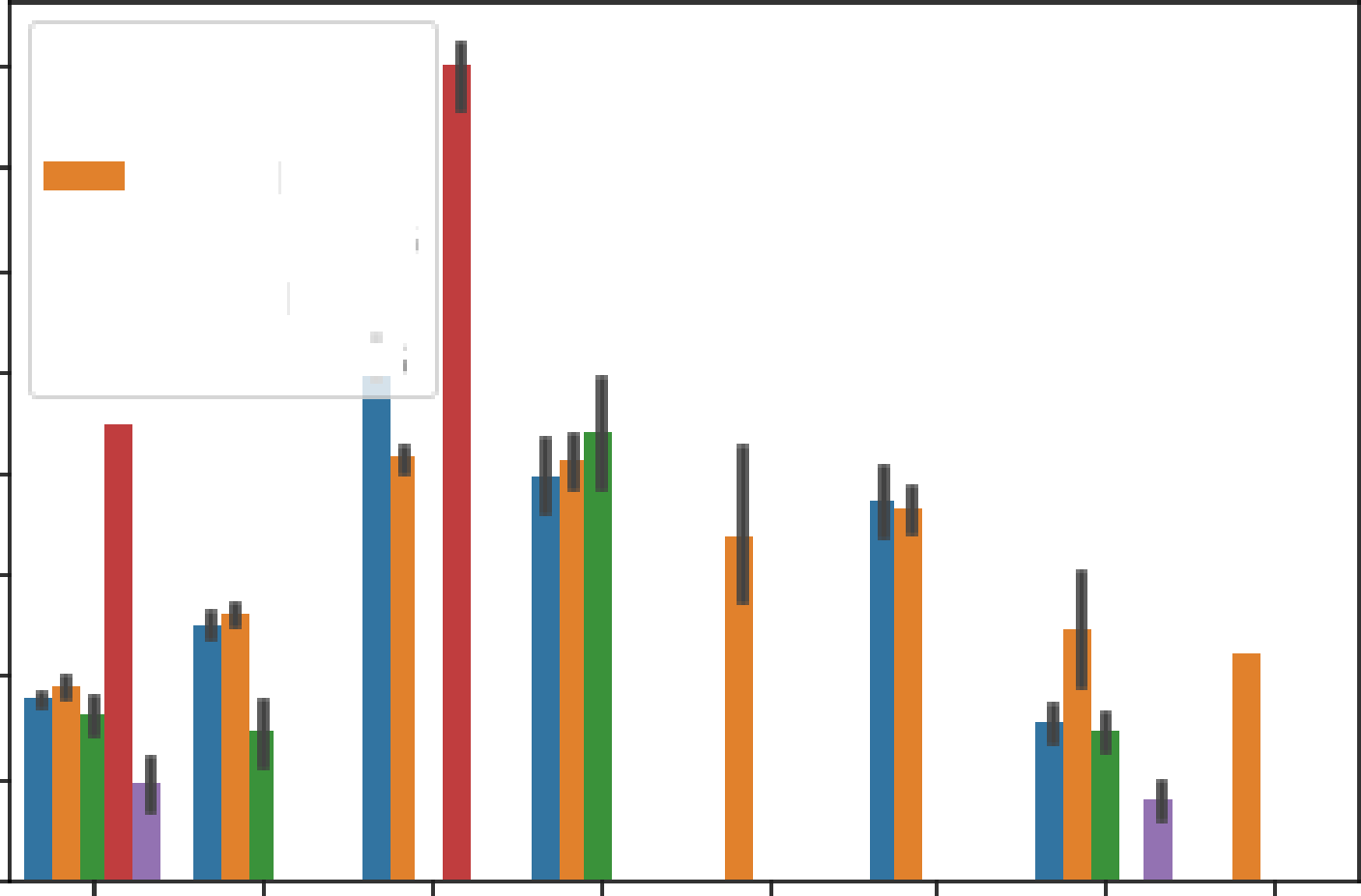
0 00 025 050 075 1 00 125 150 175 2 00

Following Observations are made from graphs above:

* SUV,MUV,Pickup and Crossover type Cars have the highest Prices
* Hybrid Fuel Type Cars are the costliest
* As the Number of previous owners increases, the price of used car decreases, so there is a negative correlation between ownership and Car Price
* Automatic Cars have the highest prices
* Kia,MG,Isuzu and Jeep are amongst the most expensive Car Brands while Maruti,Volkswagen,Chevrolet,Opel,Tata,Honda,Fiat and Ford are the most affordable Car Brands
* Car Prices are highest in Lucknow,Rajkot,Bhubhaneshwar,Kohlapur,Ahmedabad and Haflong
* Creta,Fortuner,Tucson,Seltos,Xuv500,Thar,Innova Crysta are amongst the most expensive car models, while Beat,Punto,Wagon R,Santro Xing,Alto,Fiesta are amongst the most affordable

#### Multivariate Analysis



Ie6



Fuel Type

Petrol

1 4 Diesel

Petrol + CNG

- Hybrid

08

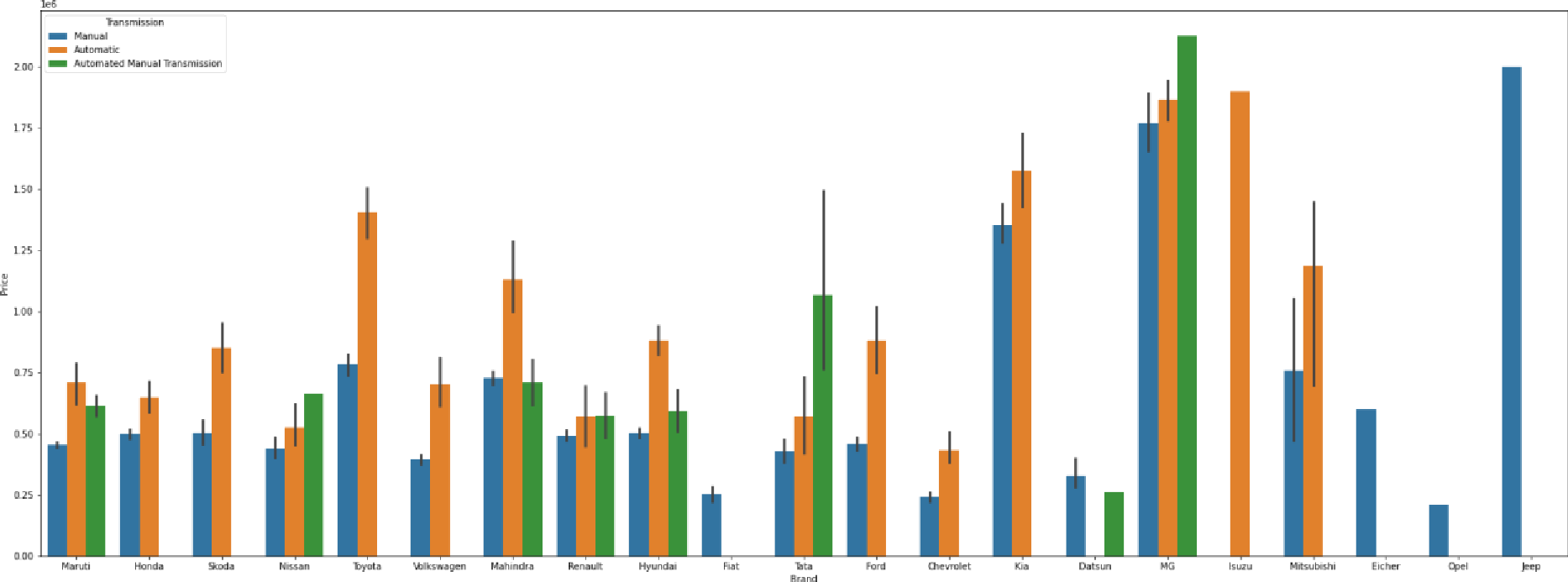
06

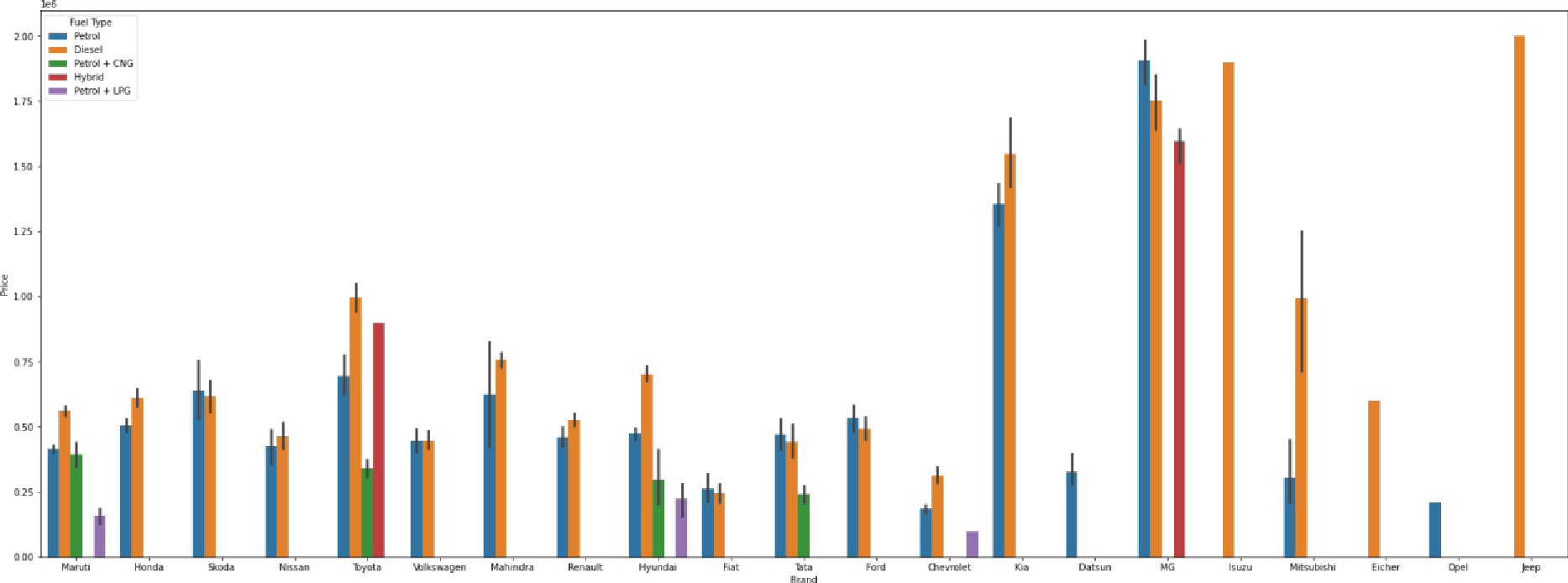
04

02

00 Hatchback5edan SUV MUV Pick UpCrossover Van Road

Type

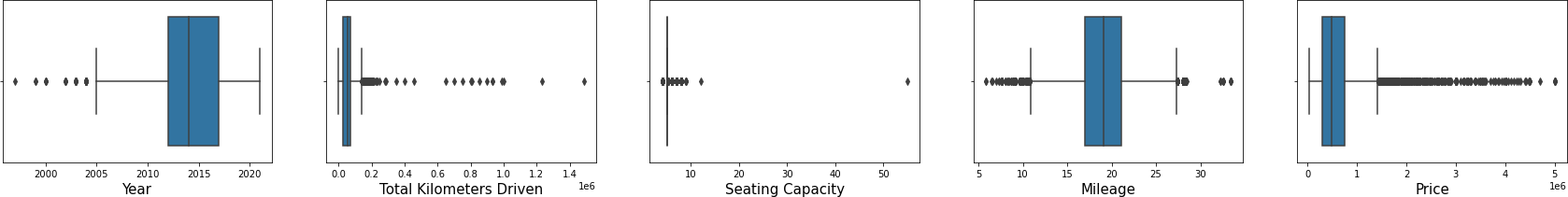




Following Observations are made from graphs above:

* SUV and MUV cars with Automatic Transmission are the costliest
* Hybrid Cars with Manual Transmission have the highest price followed by Diesel and Petrol cars with Automatic Transmission
* Hybrid Type Hatchback and SUV Cars are the most expensive
* Automatic Transmission Variants are the most expensive cars of most car brands
* Diesel Variants,followed by Petrol Variants are the most expensive cars of most car brands

#### Checking for Outliers



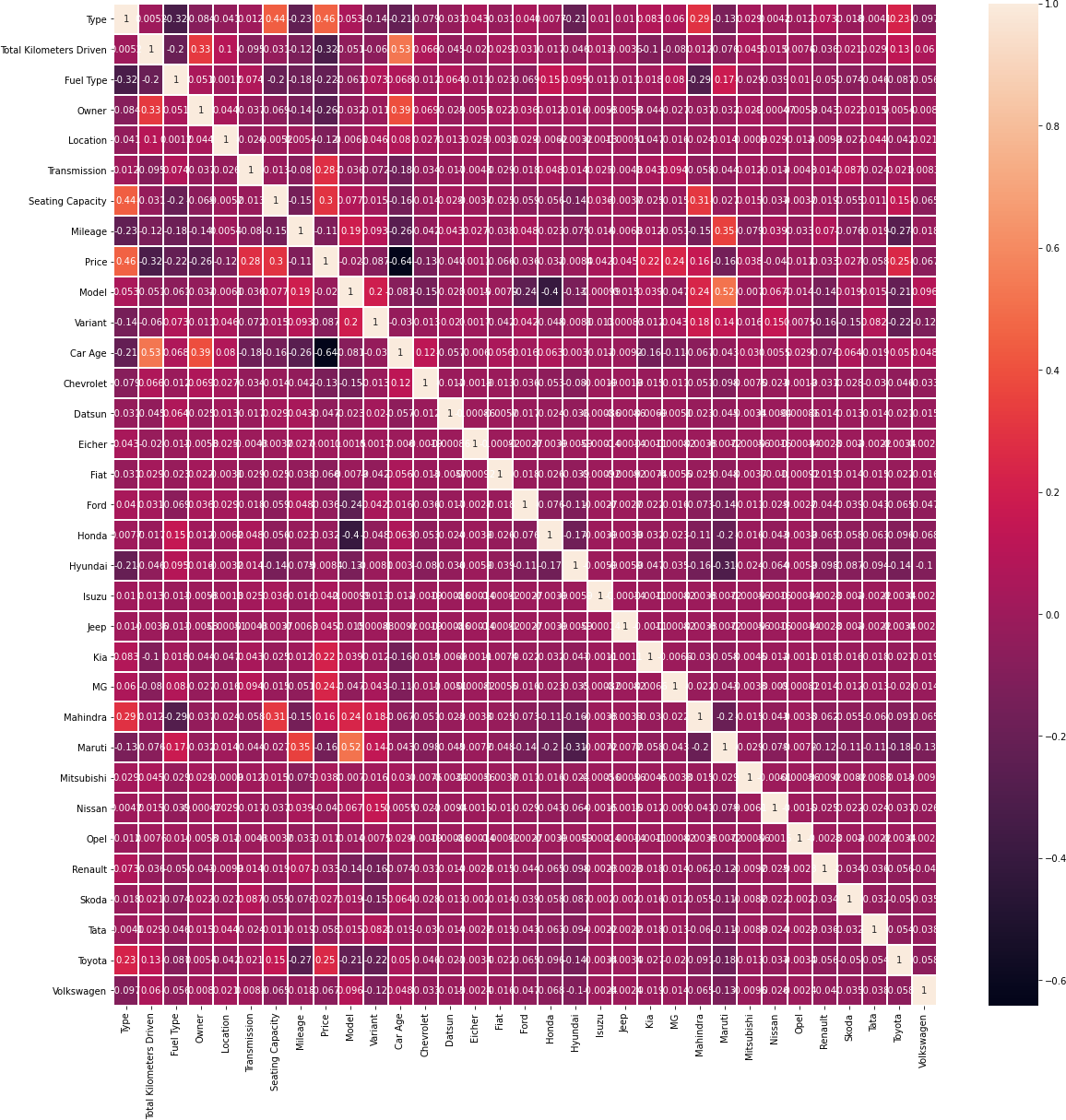
There are considerable outliers in the columns.

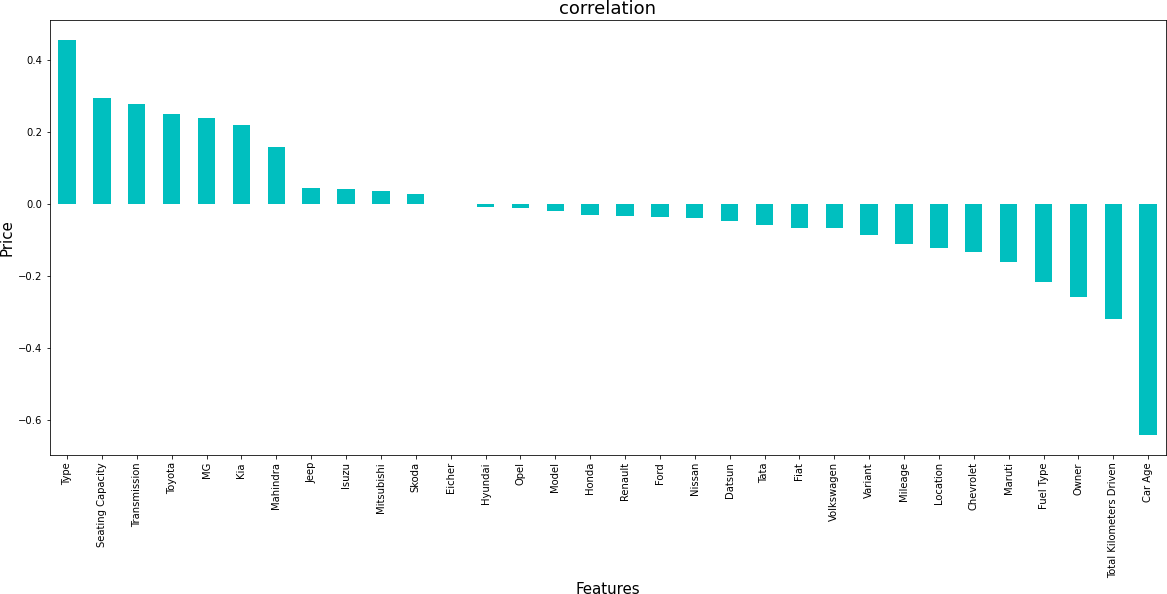
Outliers were Removed using Z score method which resulted in a total data loss of 3.82%, which is within acceptable range.

**Encoding Categorical Columns**

Categorical Columns were encoded using Label Encoding technique and get\_dummies() technique.

## Finding Correlation between Feature and Target columns





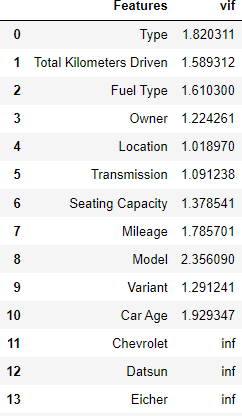
Type, Seating Capacity have the strongest positive correlation with Price while Car Age,Total Kilometers Driven,Owner and Fuel Type have the strongest negative correlation with Price.

# Model/s Development and Evaluation

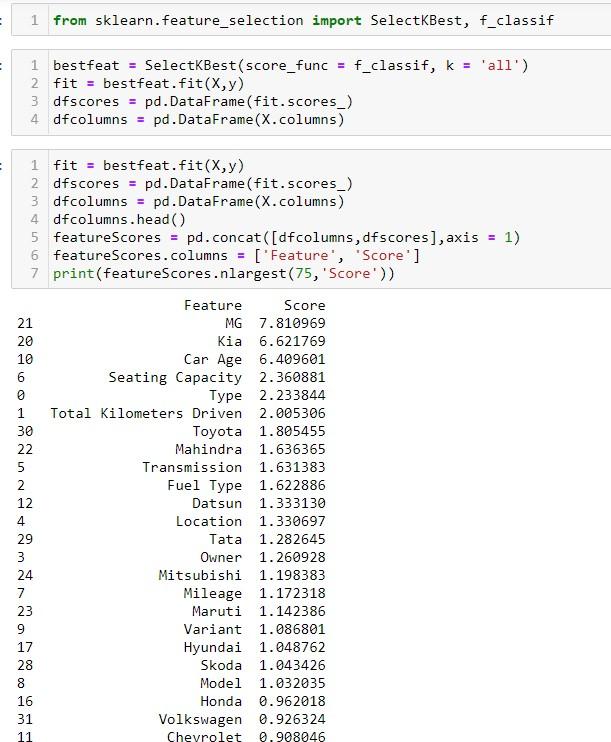
#### Feature Selection

Features were first checked for presence of multicollinearity and then based on respective ANOVA f-score values, the feature columns were

selected that would best predict the Target variable, to train and test machine learning models.



There is no Multicollinearity among the columns



Using SelectKBest and f\_classif for measuring the respective ANOVA f-score values of the columns, the best features were selected. Using StandardScaler, the features were scaled by resizing the distribution values so that mean of the observed values in each feature column is 0 and standard deviation is 1. From sklearn.model\_selection’s train\_test\_split, the data was divided into train and test data. Training data comprised 75% of total data where as test data comprised 25% based on the best random state that would result in best model accuracy.

##### The model algorithms used were as follows:

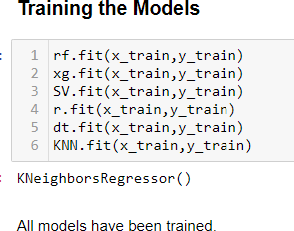
* Ridge: Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. Since the features have multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be

far away from the actual values. Ridge shrinks the parameters. Therefore, it is used to prevent multicollinearity.

* DecisionTreeRegressor: Decision Tree solves the problem of machine learning by transforming the data into a tree representation. Each internal node of the tree representation denotes an attribute and each leaf node denotes a class label. A decision tree does not require normalization of data. A decision tree does not require normalization of data.
* XGBRegressor: XGBoost uses decision trees as base learners; combining many weak learners to make a strong learner. As a result it is referred to as an ensemble learning method since it uses the output of many models in the final prediction. It uses the power of parallel processing,supports regularization, and works well in small to medium dataset.
* RandomForestRegressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A random forest produces good predictions that can be understood easily. It reduces overfitting and can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.
* Support Vector Regressor: SVR works on the principle of SVM with few minor differences. Given data points, it tries to find the curve. But since it is a regression algorithm instead of using the curve as a decision boundary it uses the curve to find the match between the vector and position of the curve. Support Vectors helps in determining the closest match between the data points and the function which is used to represent them. SVR is robust to the outliers. SVR performs lower computation compared to other regression techniques.
* K-Nearest Neighbor Regressor: It is a lazy learning, non- parametric algorithm. It uses data with several classes to predict the classification of the new sample point. KNN is non-parametric since it doesn’t make any assumptions on the data being studied. The Training phase is fast. KNN Regressor keeps all training data since they are needed

during testing phase. KNN algorithm fairs across all parameters of considerations. But mostly, it is used due to its ease of interpretation and low calculation time.

**Regression Model Building**



Ridge Regression Model

y\_r pred : r.predict(x test)

R2 Score



*8 .* 703745 3679274228

Mean Squared Error



43531846036.22704

Root Mean Squared Error

np.sqrt(mean\_squared error(y test,y r\_pred))

208642. 8672066865 5

Random Forest Regression Model

y\_rf\_pred : rf. predict(x test)

R2 Score



*8.* 9060733783304681

Mean Squared Error



13801638153.692486

Root Mean Squared Error

np.sqrt(mean\_squared\_error(y\_test,y\_rf\_pred))

117480.37)48294602

XGB Regression Model

y\_x p red = xg . pred1ct(x\_tes I )

R2 Score



8.9263511747648791

Mean Squared Error



1B821858772 . 668312

Root Mean Squared Error

np . sq rt ( mean\_s q ua red\_er rar (y\_test , y\_xg\_pred ) )

1B4828.16336294856

Support Vector Regression Model

y\_s vr\_pr‘ed = IV . p redi ct (x\_test )

R2 Score



. -B.86863993854979955

Mean Squared Error



15702 66 39373. 24176

Root Mean Squared Error

np.sqrt(mean\_squared\_error(y\_test,y\_svr\_pred))

. 396265.869553815g

Decision Tree Regression Model

y\_dt\_p red = dt . pred ict (x test )

R2 Score



0.83616g10751748l2

Mean Squared Error



24073563574.27525

Root Mean Squared Error

np.sqrt(mean\_squared\_error(y\_test,y\_dt\_pred))

155156.5776055764

KNN Regression Model

y\_knn\_pred = KNN.predict(x\_test)

R2 Score

n2\_z c ane (y\_tes I , y\_k nn\_p red )

0.805527649990)519

R2 Score



J.gJJs27F4999035l9

Mean Squared Error

mean\_squared\_error(y\_test,y\_knn\_pred)

28575892095.585197

Root Mean Squared Err<ir

np . sqrt ( nean\_ s qua red er rar (y test, y knn pred ) )

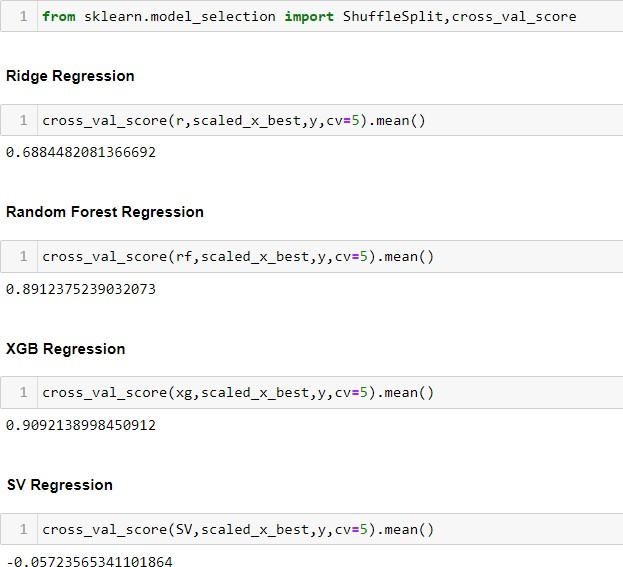
169044. 0537125 9B54

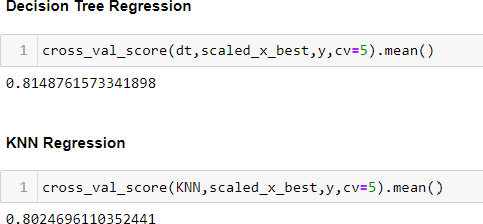
## Analyzing Accuracy of The Models

Mean Squared Error and Root Mean Squared Error metrics were used to evaluate the Model performance. The advantage of MSE and RMSE being that it is easier to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced than smaller error, hence the model can now focus more on the larger errors.

Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

Using cross-validation, there are high chances that we can detect over-fitting with ease. Model Cross Validation scores were then obtained for assessing how the statistical analysis generalises to an independent data set. The models were evaluated by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.



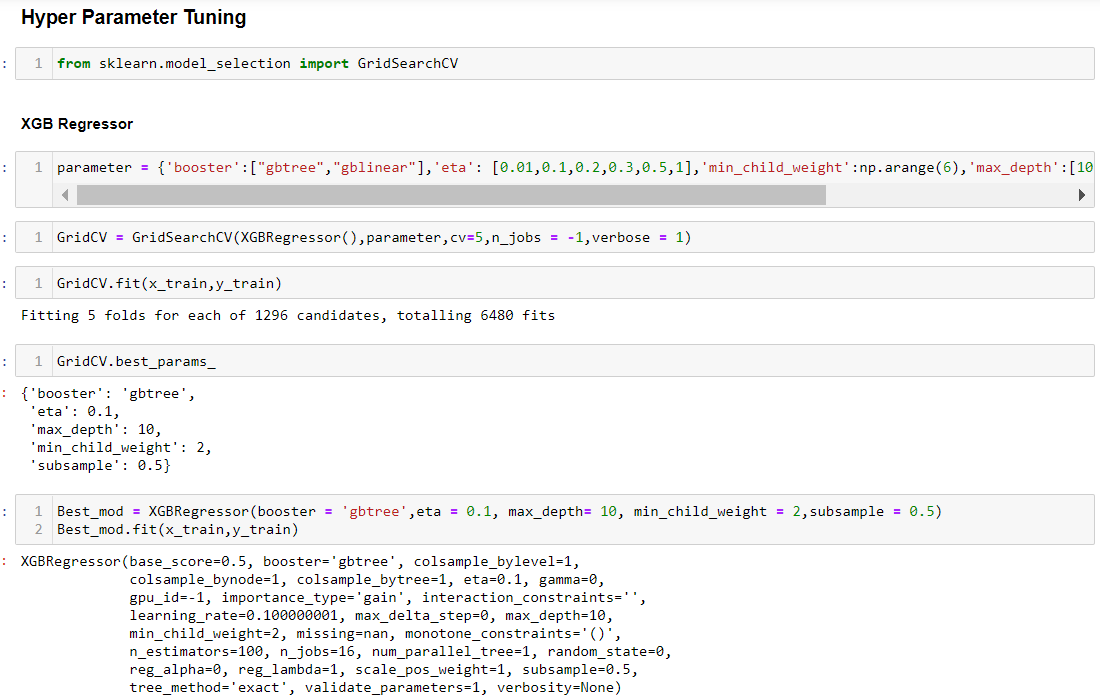


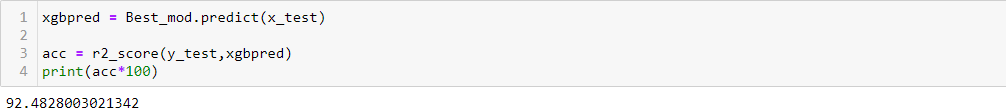
Interpretation of the Results

Based on comparing Accuracy Score results with Cross Validation results, it is determined that XGB Regressor is the best model. It also has the lowest Root Mean Squared Error score.

#### Hyper Parameter Tuning

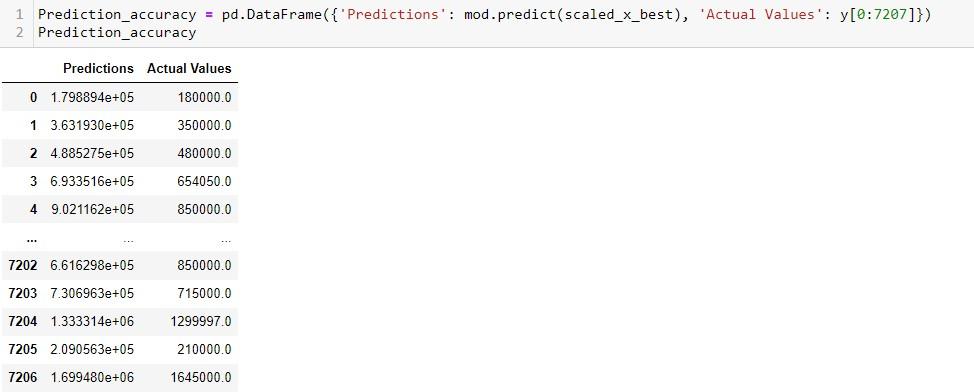
GridSearchCV was used for Hyper Parameter Tuning of the XGB Regressor model.





Based on the input parameter values and after fitting the train datasets The XGB Regressor model was further tuned based on the parameter values yielded from GridsearchCV. The Random Forest Regressor model displayed an accuracy of 92.48%

This model was then tested using a scaled Test Dataset comprising of 7207 entries. The model performed with good amount of accuracy.



In summary, Based on the visualizations of the feature-column relationships, it is determined that, Features like Type, Seating Capacity have the strongest positive correlation with Price while Car Age,Total Kilometers Driven,Owner and Fuel Type have the strongest negative correlation with Price. and are some of the most important features to predict the label values. XGB Regressor Performed the best out of all the models that were tested. It also worked well with the outlier handling.

## CONCLUSION

##### Key Findings and Conclusions of the Study and Learning Outcomes with respect to Data Science

Based on the in-depth analysis of the Housing Project, The Exploratory analysis of the datasets, and the analysis of the Outputs of the models the following observations are made:

* Car attributes like Type,Car Age, Seating Capacity,Total Kilometers Driven,Transmission,Fuel Type,Owner and Mileage etc play a big role in influencing the used car price.
* Brand Name also has a very important role in determining the used car price.
* Various plots like Barplots,Countplots and Lineplots helped in visualising the Feature-label relationships which corroborated

the importance of Car features and attributes for estimating Sale Prices.

* Due to the Training dataset being very small, the outliers had to be retained for proper training of the models.
* Therefore, XGB Regressor, which uses the power of parallel processing,supports regularization, and works well in small to medium dataset performed well despite having to work on small dataset.

##### Learning Outcomes of the Study in respect of Data Science

Data cleaning was a very important step in removing plenty of anomalous data from the huge dataset that was provided.

Visualising data helped identify outliers and the relationships between target and feature columns as well as analysing the strength of correlation that exists between them.

Limitations of this work and Scope for Future Work

A small dataset to posed a challenge in building highly accurate models. The presence of anomalous entries in the

numbers heavily distorted the data distributions and may have had some impact on model learning.

Availability of more features would help build better models.