

Car Price Prediction

Submitted by:

Kartik Krishna

**ACKNOWLEDGMENT**

I would like to thank Shubham Yadav Sir and Sajid Choudhary Sir, who helped me in this project by clearing my doubts in doubt clearing sessions and providing me with learning material that proved to be helpful for my project. Also, DataTrained live sessions and Krish Naik youtube channels helped me understand concepts that were alien to me.

**INTRODUCTION**

With the covid 19 impact in the market, we have seen 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. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We used histogram, count plots and distplots to see the data distribution in each column, we also used stripplots and scatterplots for studying the correlations present between the target variable and other features in the dataset. We used log and square root transformation to treat the skewness of the dataset.

* Data Sources and their formats

The data comprises 2240 rows and 11 columns. This data was collected from olx.com.

* Data Preprocessing Done

All the null values were dropped from the dataset, certain columns with numeric values were stored in the form of strings, they were converted into float values, skewness present in numeric columns were treated where ever necessary.

* Data Inputs- Logic- Output Relationships

The model was build on the relation between different features like location of car, brand of car, km driven by car, number of previous of owners of car, etc. with the price of the car, how certain variables impact the car price.

* Hardware and Software Requirements and Tools Used

All of the work in this project was done on Jupyter notebook. We used pandas and NumPy for working on data and using all the basic mathematical functions on it. We also used matplotlib.pyplot and seaborn libraries for data visualization. We used sci-kit learn library for data preprocessing like power transform, standard scaler, and importing different models like Linear Regression, Ridge and Lasso, Decision Tree Regressor, Random Forest Regressor, etc.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

For treatment of skewness both log transformation and square root transformation were used but since square root transformation gave more favourable results, this approach was preferred over log transformation.

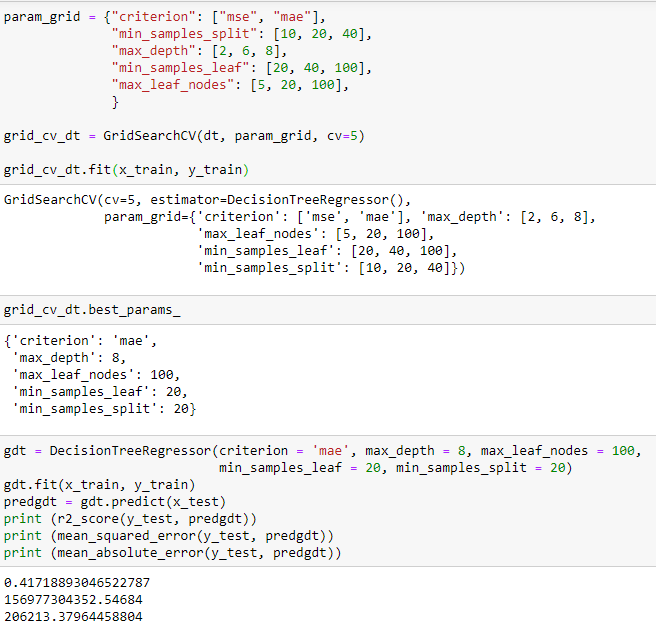
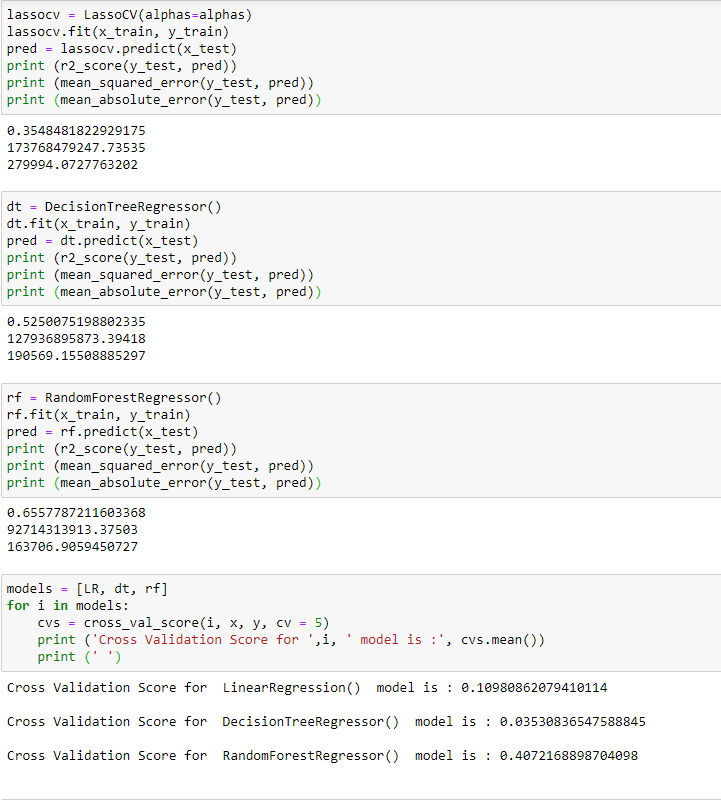
* Testing of Identified Approaches (Algorithms)

We used Linear Regression, Ridge and Lasso, Decision Tree Regressor and Random Forest Regressor models for building our model. Then we used gridsearchCV for hypertuning and cross-validating the models..

* Run and Evaluate selected models

We used Linear Regression, Ridge and Lasso, Decision Tree Regressor and Random Forest Regressor models for building our model. Then we used gridsearchCV for hypertuning and cross-validating the models, below is the snapshot of code that we used to implement our model.



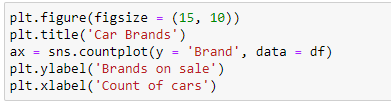


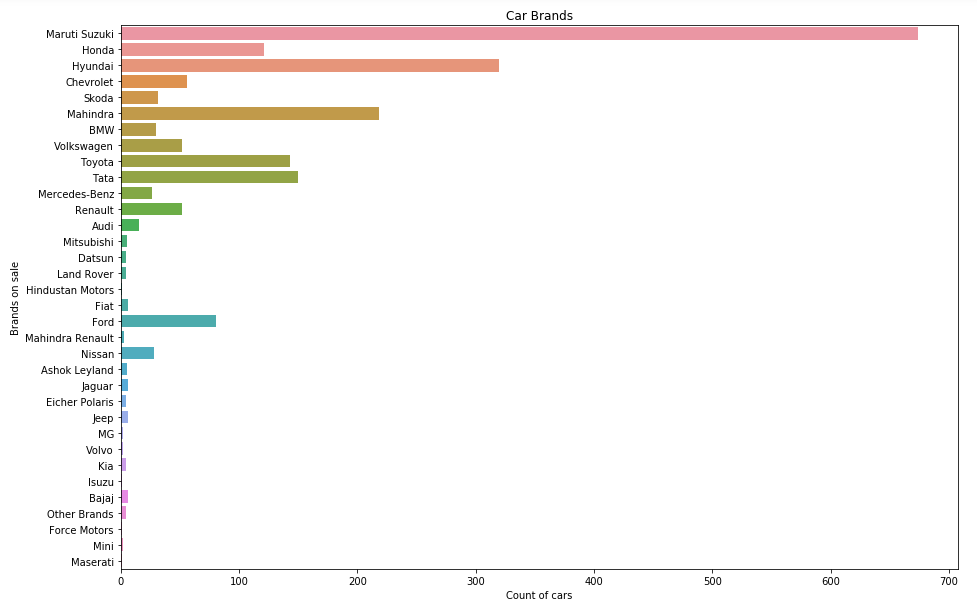
* Key Metrics for success in solving problem under consideration

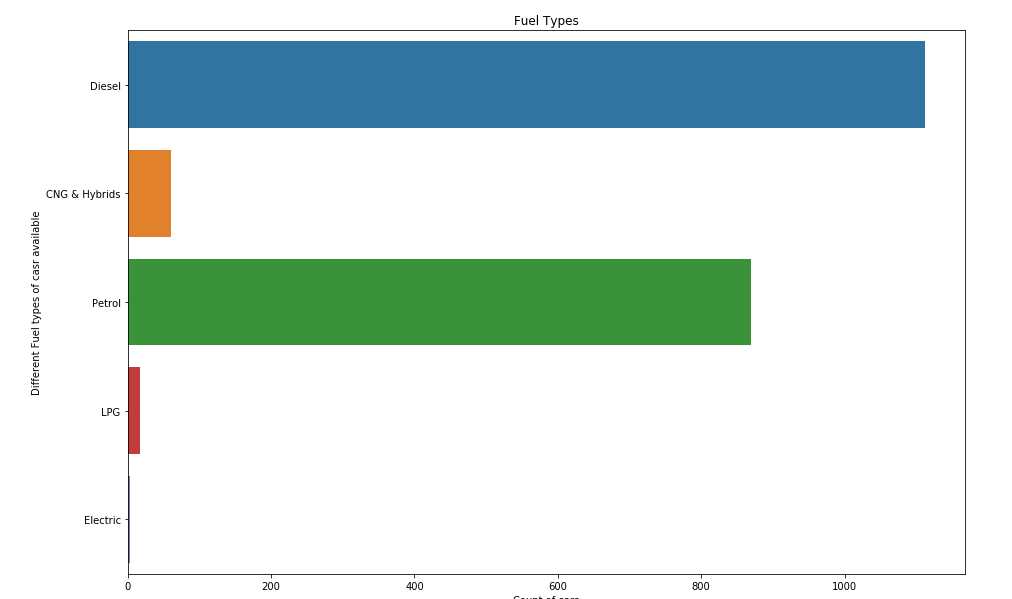
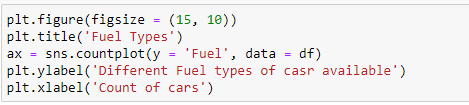
For evaluation of our models, we used R2 score, mean squared error and mean absolute error.

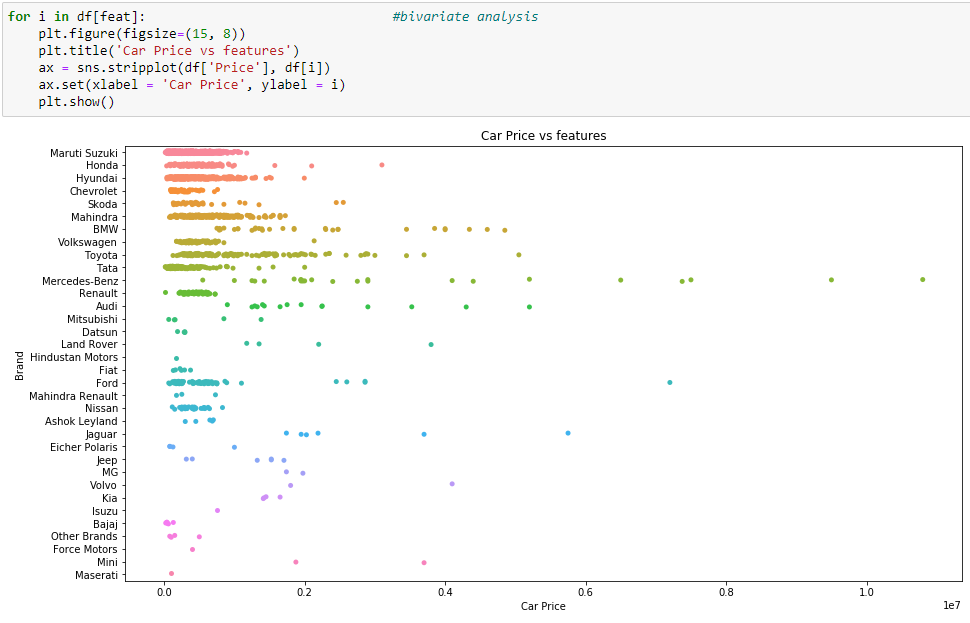
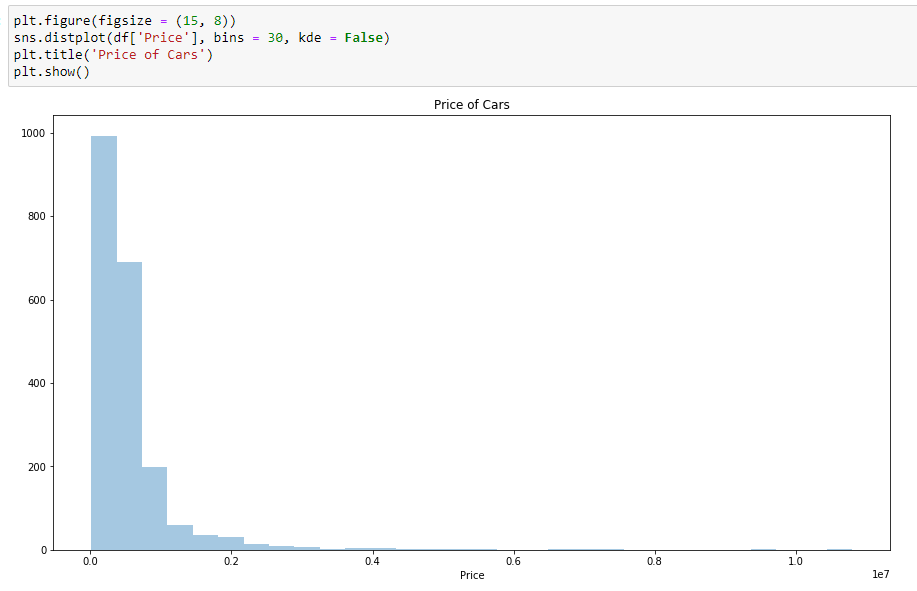
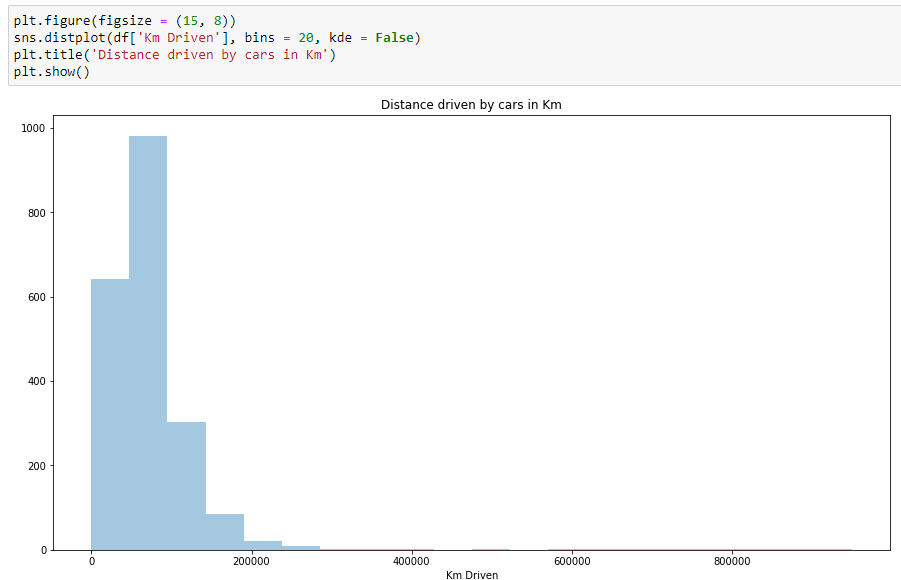
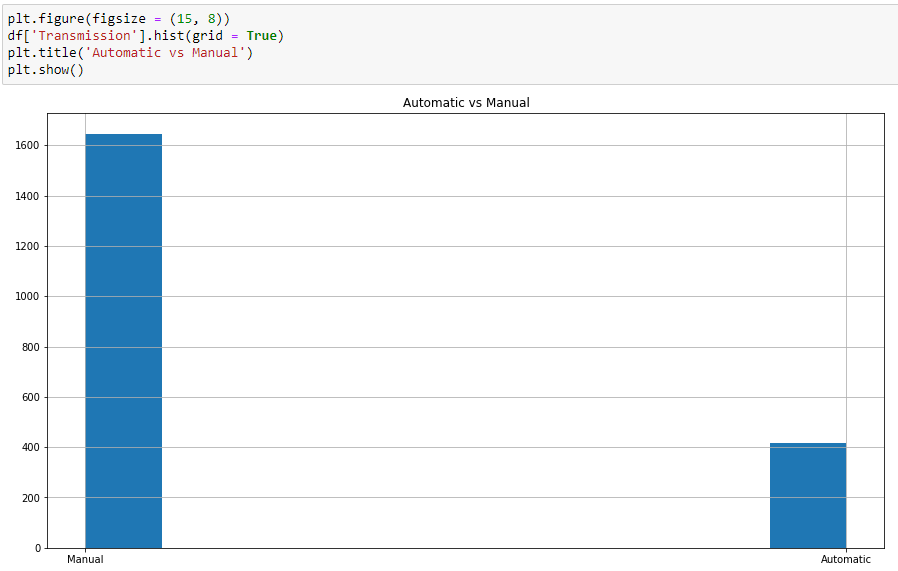
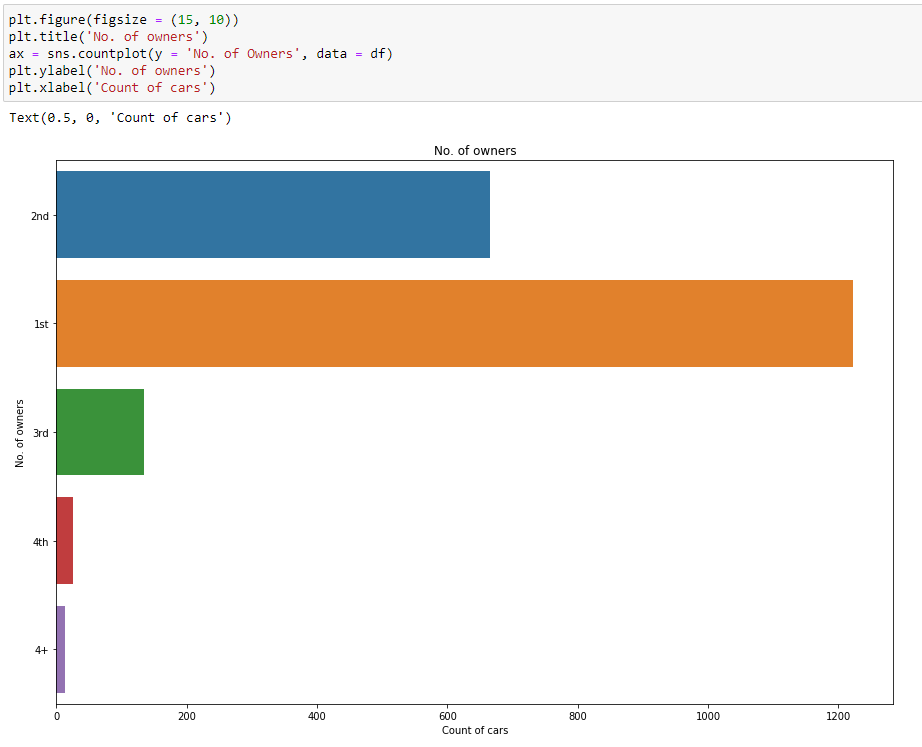
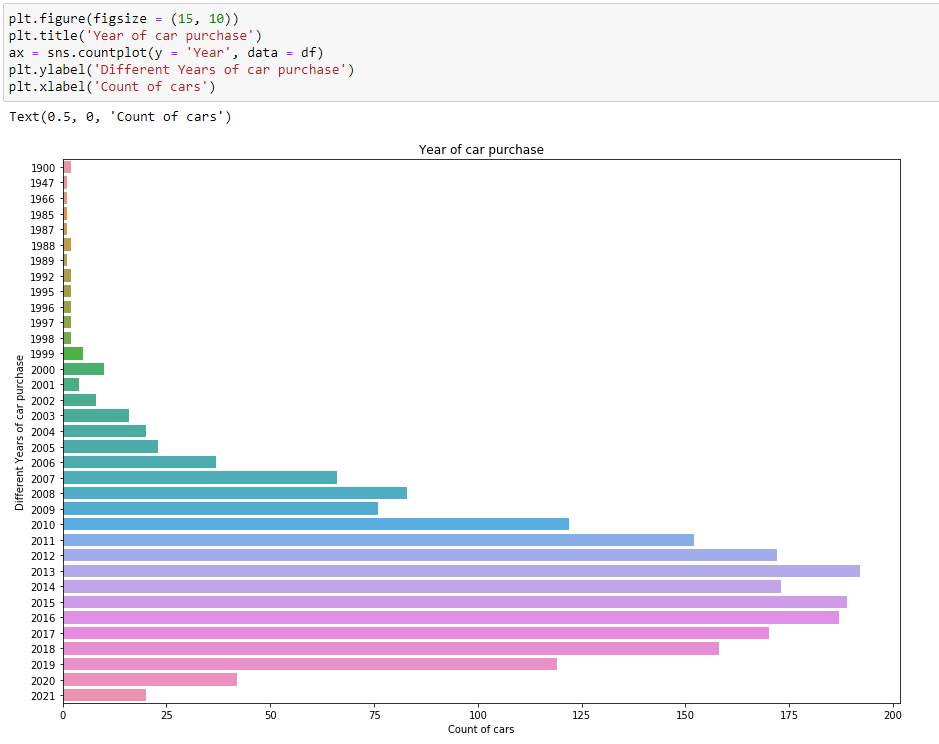
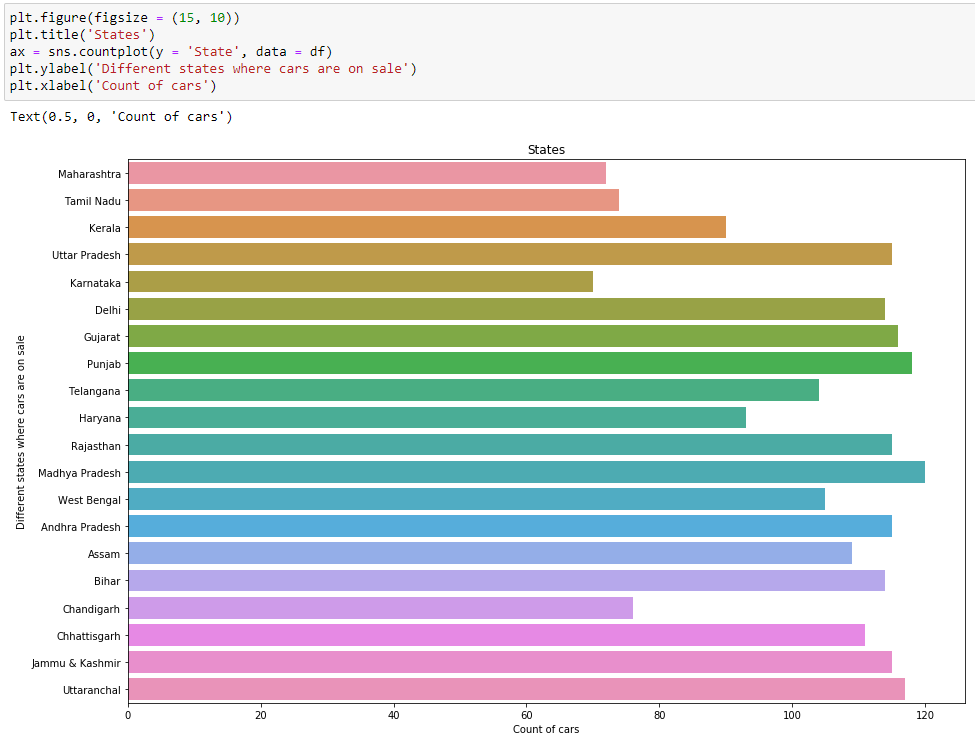
* Visualizations

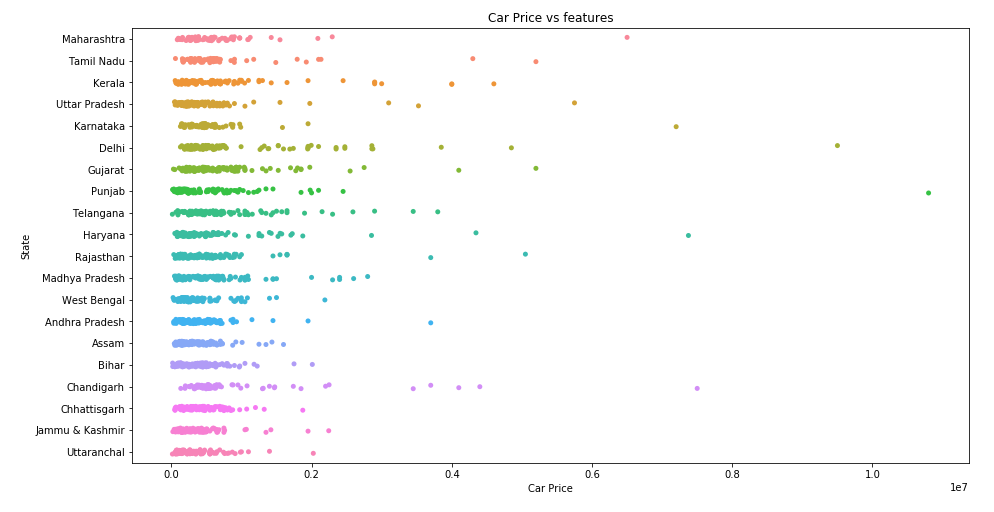
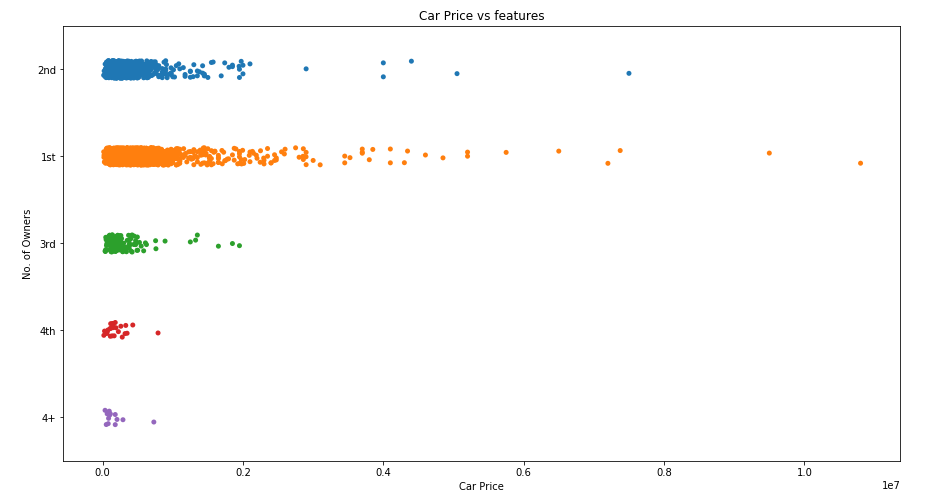
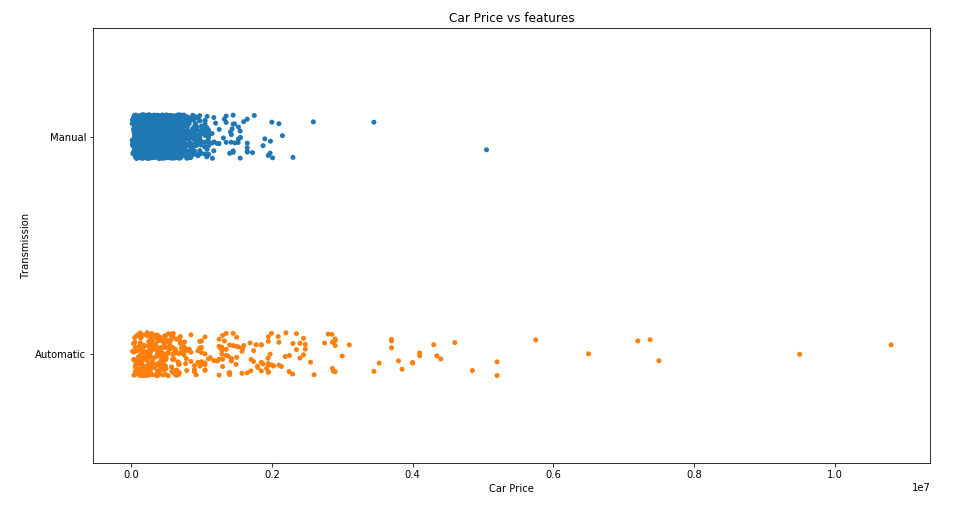
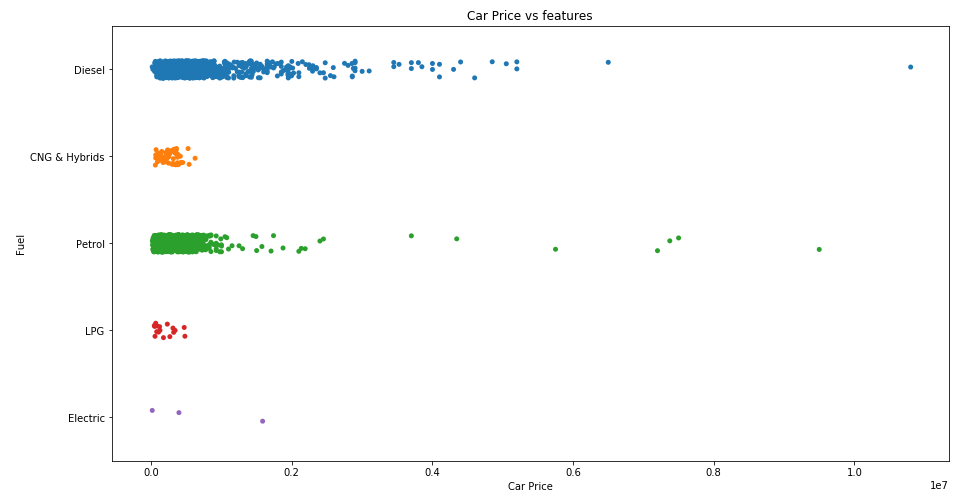
We used histogram, count plots and distplots to see the data distribution in each column, we also used stripplots and scatterplots for studying the correlations present between the target variable and other features in the dataset.











* Interpretation of the Results

We observed from the distplot that the target variable which is continuous data is right side skewed distribution. From the stripplot, we can see how categorical columns are correlated with the target variable.

After preprocessing we got rid of rows with null values, also we got rid of some outliers present after we also got rid of the skewness present in continuous data.

After building and hypertuning our model on the dataset our R2 scores were Linear Regression - 0.351, Ridge Regression - 0.353, Lasso Regression - 0.354, Decision Tree Regressor - 0.417, and Random Forest Regressor - 0.653, hence we decided to finalize our Random Forest model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project, we learned about the car resale market.

Using predictive modeling and EDA we can determine which factors impact the market and which factors can help determine the prices.

After treating outliers, getting rid of redundant columns, and other preprocessing, we manage to build a model which will give us fairly accurate results.

* Learning Outcomes of the Study in respect of Data Science

While working on this project I learned about the car resale market and how data science and machine learning can be so helpful in this field.

* Limitations of this work and Scope for Future Work

The accuracy of the model is not as good as expected, maybe with more data and more different columns the model performance could be improved significantly.