FLIP ROBO

USED CAR: PRICE PREDICTION

SUBMITTED BY

TUSHAR KUMAR PATEL

2021

INTRODUCTION

• Business problem framing

In this Covid situation cars are one of the necessary need of each and every person around the globe and therefore new car and used car market is also one of the market which has one of the major contributors in the economy also. It is also very large market same as the market for new car and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in sales and purchase of the used cars also. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for used car companies. Our problem is related to price prediction of the used cars during this covid-19 situation.

- Conceptual Background of the Domain Problem

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.

- Review of Literature

Cars is one of human life's most essential needs that are used for daily commute, along with other fundamental needs such as house, food, water, and many more. Demand for cars has also grown-up rapidly over the years as people's living standards improved, but for last two years it has been more due to covid-19 as it is not safer to commute in a public transport. While there are people who make purchase new car, but the people who can't afford to purchase the new cars they opt for used cars. So, they go for used cars websites or stores, yet most people around the world are buying a car as their mean for daily commute. According to the researchers the demand for cars has been increased but due to chip shortage worldwide many people are opting to go and purchase used vehicles as they need safer means for their commute.

- Motivation for the Problem Undertaken

We have the data of different car prices with independent factors. Our objective is to find the important features that affect the car price and to build a model that predicts the car prices given the

- Independent features are provided. This model will be helpful for people who are looking for used cars in India to estimate their expenditure.
- I have used visualization tool to understand the data in a better way.

- I have also used label encoder technique and converted all the data into numerical form to do the data analysis in an easier way.

- I have done the model building and also applied hyper parameter tuning in regression model to improve my accuracy of my prediction.

Analytical Problem Framing

- **Mathematical / Analytical modeling of the problem**

- In this particular project I need to understand the importance of cars.I have done the exploratory data analysis process and try to figure out the Used Car Price Prediction in a better way. Here I tried to use different regression modeling techniques such as SVR, KNN, linear regression, etc to find the best accuracy

- Data Sources and their formats

The dataset has been scrapped from different used cars websites such as Olx, Cardekho, Car24, etc. The data is scraped and saved in CSV format. The data consist of 5781 rows and 10 columns.

Description of columns:

brand: Brand of the car model:
Model of the car
variant: Variant of the car model
mf_year: Year of manufacture of the car dr_kms:
Driven kilometers
fuel_type: type of fuel used in the car no_of_owners:
How many people owned the car location: Location
of the selling ad is placed transmission:
Automatic/manual
price: Price of the car

- Data Preprocessing Done

The more you preprocess the dataset the more accurate result you will get. Basically, it is the process where we remove some unwanted or not useful, noisy data from the collected data. Also, if we don’t remove any null value or empty field then we cannot get the proper results. So, it is very important process to develop the model.
Since the data is scraped by ourselves, there is not much cleaning was required.

- There were some missing values
- Since the number of missing values was very less, the rows are removed
- Corrected the format of the column and converted into integer (for kilometers and price)
- Plotted the box pot of the numerical values
- There were outliers. But decided to keep it since those were real values
- Checked the skewness of the columns and removed using the square root method
- Encoded the categorical columns
- Scaled the input

- Data Inputs- Logic- Output Relationships

EDA was performed by creating valuable insights using various visualization libraries.

- Most of the vehicles in used car industry are Maruti and Hyundai
- Premium vehicles are very less
- The most expensive cars are ferrari and lamborghini
- Hyundai, Datsun, and Maruti are some of the budget-friendly brands
- lesser the age of the vehicle higher will be the price
- we can see that the purchasing of vehicles started booming around 2010
- from 2018, it has been started to decline
- Most of the vehicles are petrol
- Gas or hybrid vehicles are very less
- No electric vehicles in our dataset
- In used car industry, uncommon fueled vehicles are cheaper(CNG, LPG etc) than petrol and diesel
- Generally we can say that when no. of owners increases, price decreases
- The prices of the car is not much differ from city to city
- Automatic cars are more expensive
- In every fuel type, manual is higher in number than automatic
- Driven kilometers and mf_year are highly -ve correlated

* Hardware and Software Requirements and Tools Used

1) Anaconda is the required software that has jupyter notebook which is the main tool used for the analysis of the dataset.
2) Pandas is open-source library tool which provides high performance data analysis tool by its powerful data structures. It helps to shorten the procedure of handling the data with extensive set of features.
3) NumPy is most used package for scientific computing for multidimensional array of objects.
4) Other than this, as a pre-processing steps, I Imported Standard scaler for scaling the data.
5) In terms of selecting the which model is best, I Imported Train test split where I am splitting the train data and test data and using cross Val score to calculate whether the model is overfitting or under-fitting and RandomizedsearchCV to check improve the accuracy score.
6) Imported f1 score, classification report, confusion matrix, roc curve in terms of metrics to calculate the model score.

- Model/s Development and Evaluation

- Identification of possible problem-solving approaches(methods)

  For understanding the distribution of data and any deviations present, which part of the data contains outliers we used describe as statistical tool and observed the data on the basis of difference between mean and standard deviation, minimum and Q1,Q3 and maximum data. Correlation as used to understand the extent attributes bear impact on our target column 'label'.

  Used scalar to train and test the model and cross_val_score to cross validate the outcome in choose the best, then applied GridsearchCV to choose best parameters for the highest scoring models and metrics

- **Testing of identified Approaches (Algorithms)**

  - SVR
  - KNN
  - RandomForest
  - Linear Regression
  - Ridge

- **Run and Evaluate selected models**

In [90]:
```
models = {"SVR":SVR(),"KNN":KNeighborsRegressor(), "RandomForest":RandomForestRegressor(),
          "LinearRegression":LinearRegression(), "Ridge":Ridge(), "dtr":DecisionTreeRegressor() }
acc = {}
mod_list = []
for i in models:
    mod = i
    mod = models[i]
    #mod = DecisionTreeRegressor()
    mod.fit(x_train, y_train)
    pred = mod.predict(x_test)
    r2_sc = r2_score(y_test,pred)
    acc[i] = r2_sc
    mod_list.append(mod)
print(acc)
```

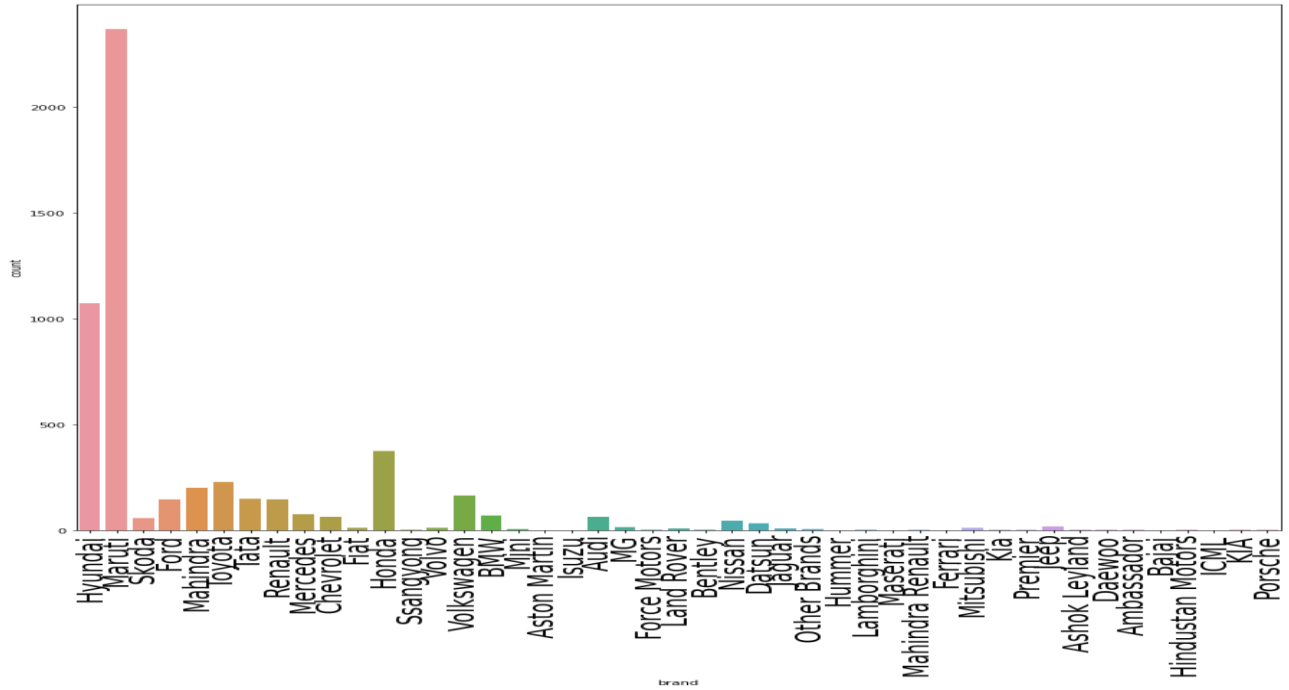
We have used SVR, KNN, RandomForestRegressor, LinearRegression, and ridge. The result of these algorithm as follows

```
{'SVR': -0.06098555762736457, 'KNN': 0.017600004792900026, 'RandomForest': 0.6208230217100357, 'LinearRegression': 0.2268197286
9206657, 'Ridge': 0.22726266082261815, 'dtr': 0.5026174879947103}
```

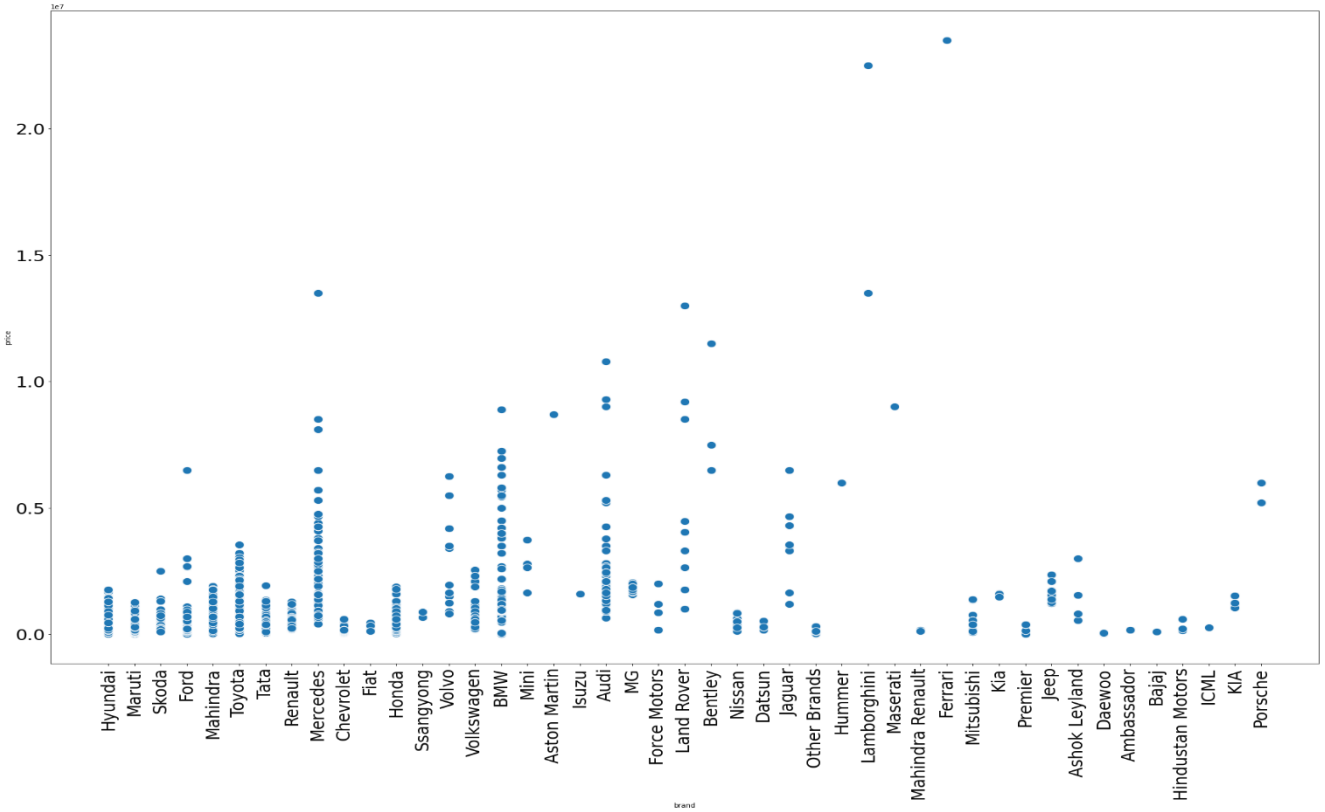
- Key Metrics for success in solving problem under consideration
  Here we have used r2_score since this is a regression problem

- Visualizations

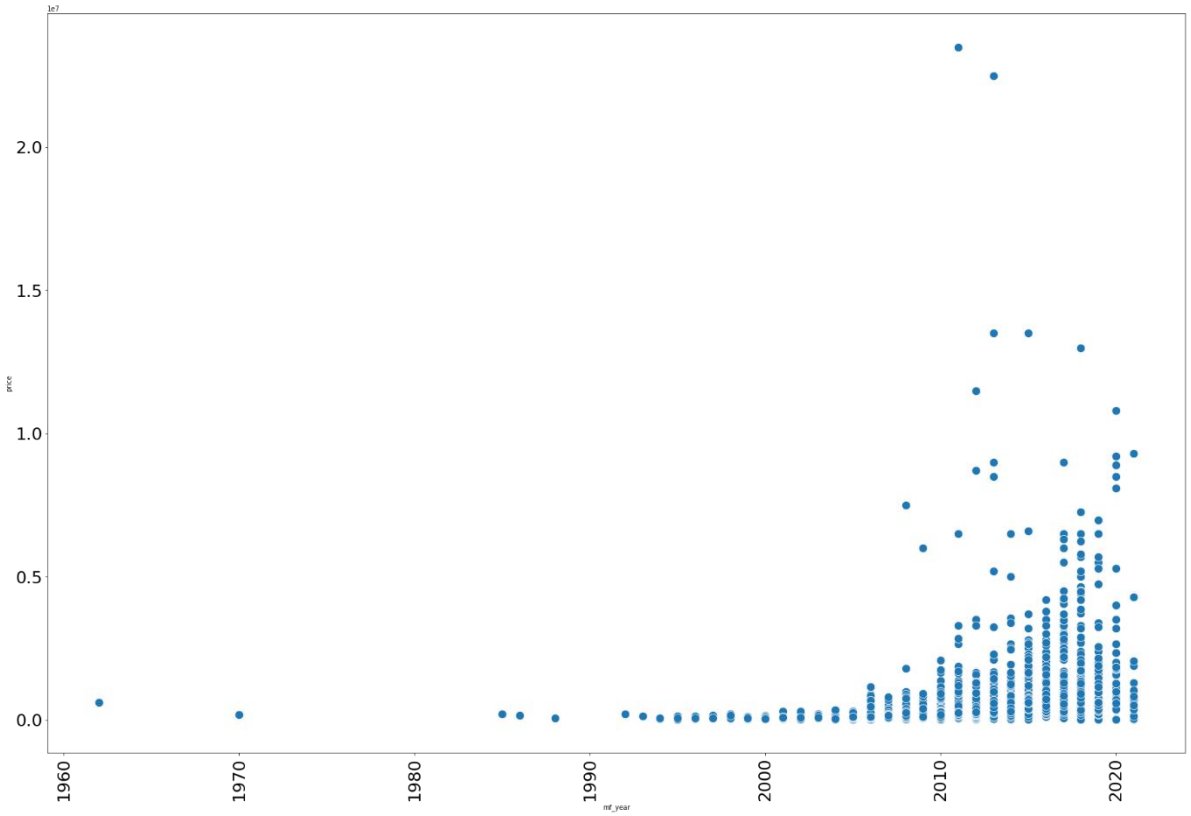
Brands



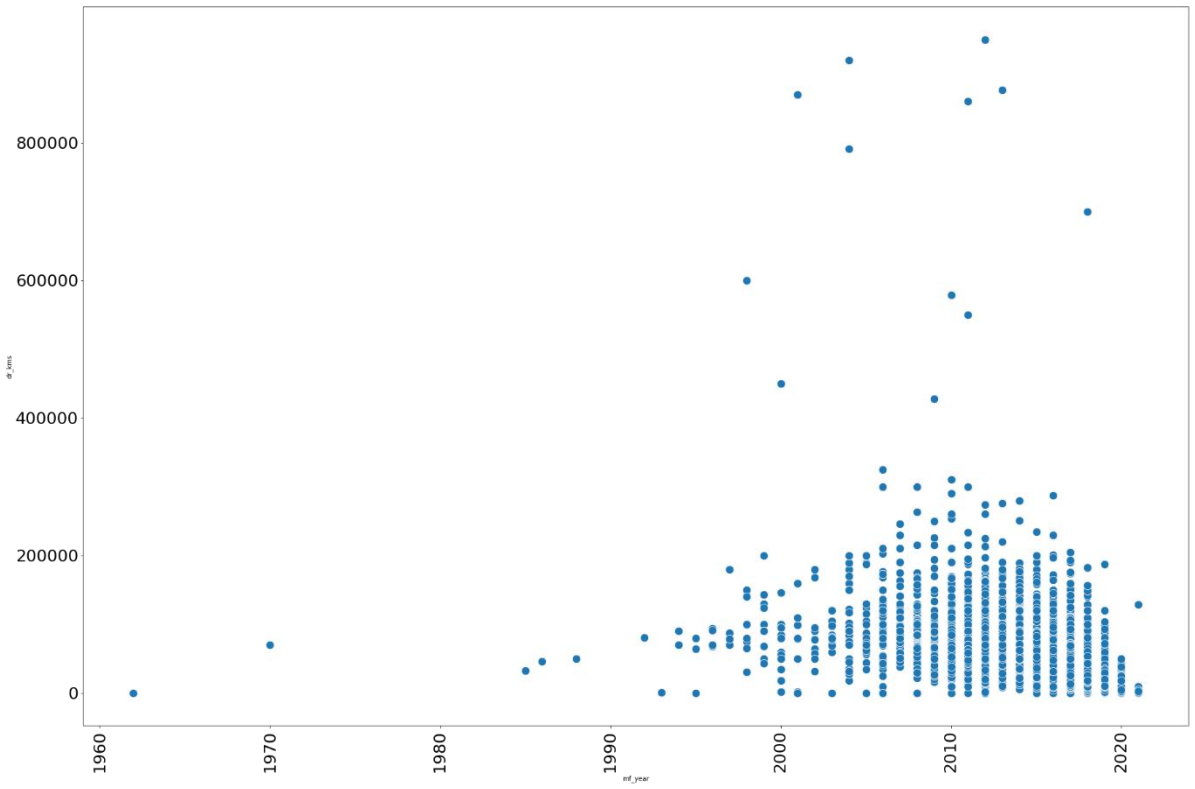
Price



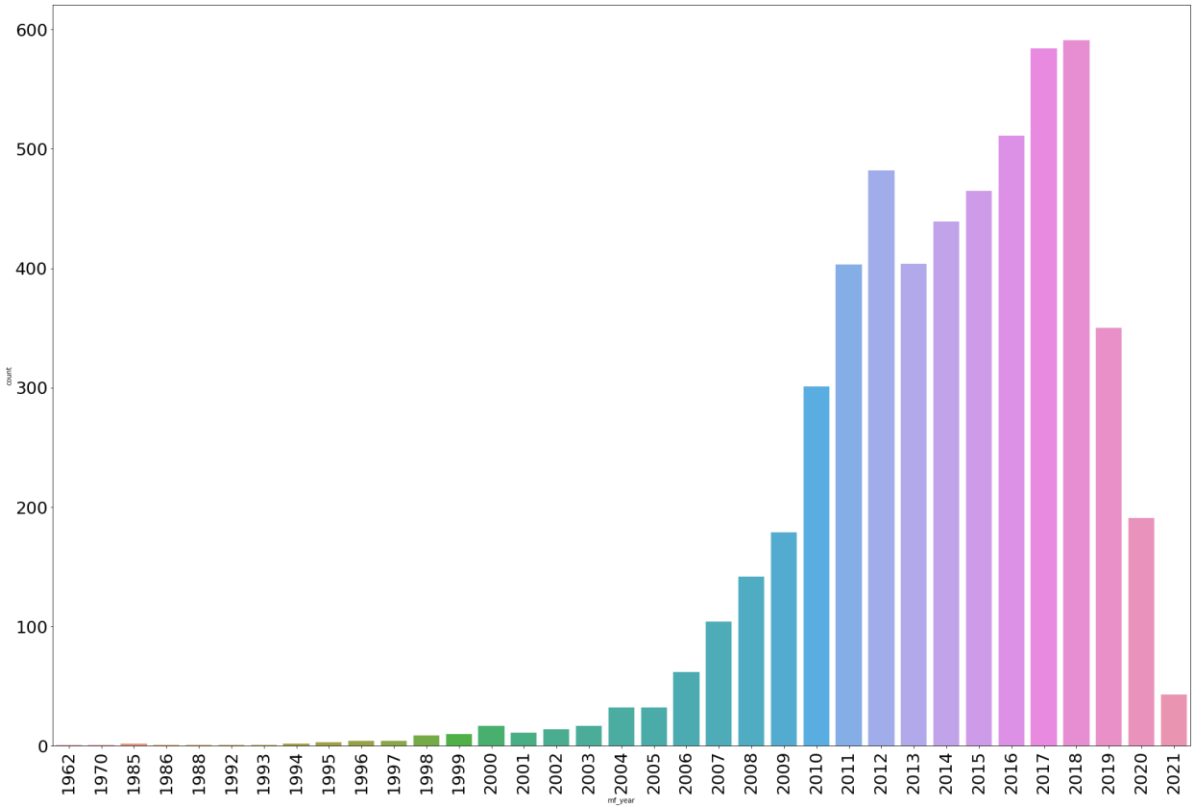
Price year wise



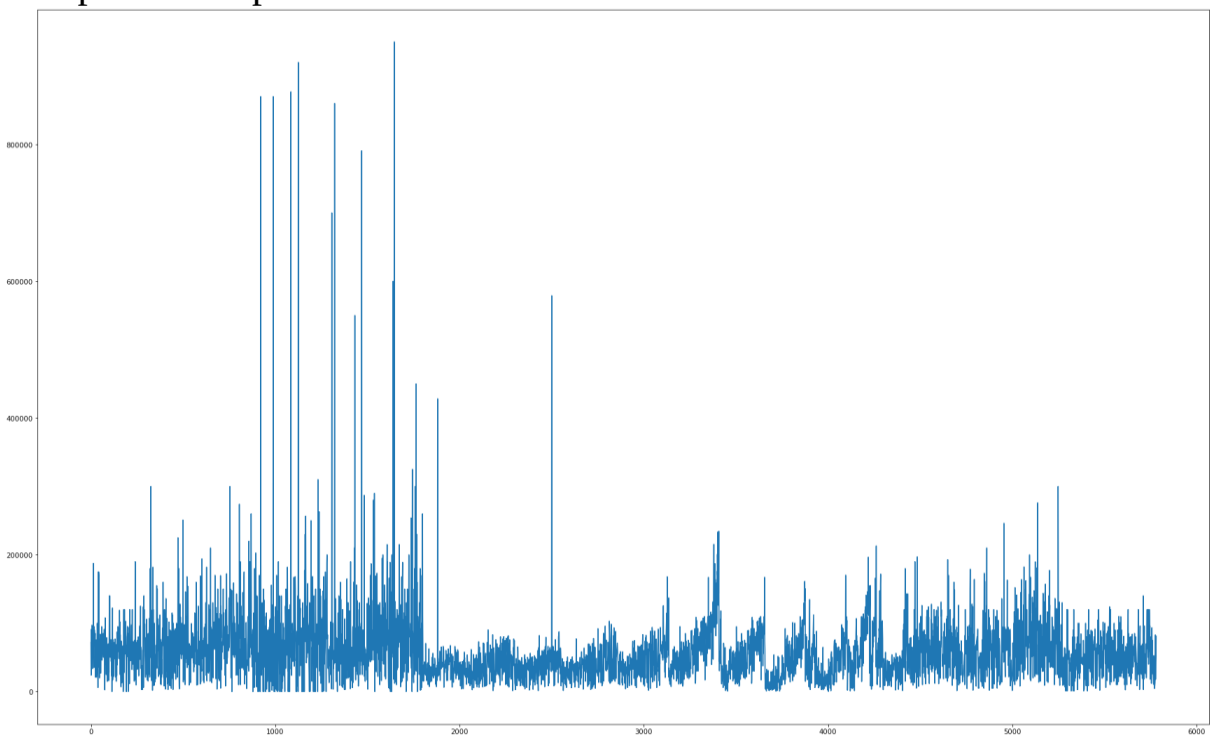
Manufacture year with driven kilometers



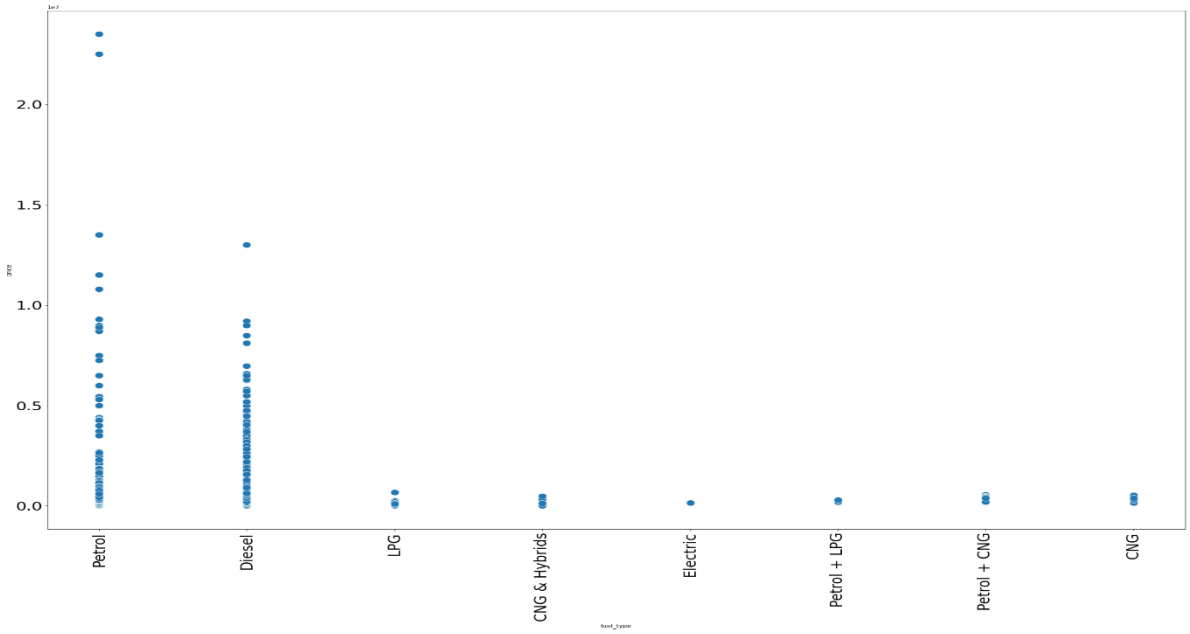
Count with a manufacture year



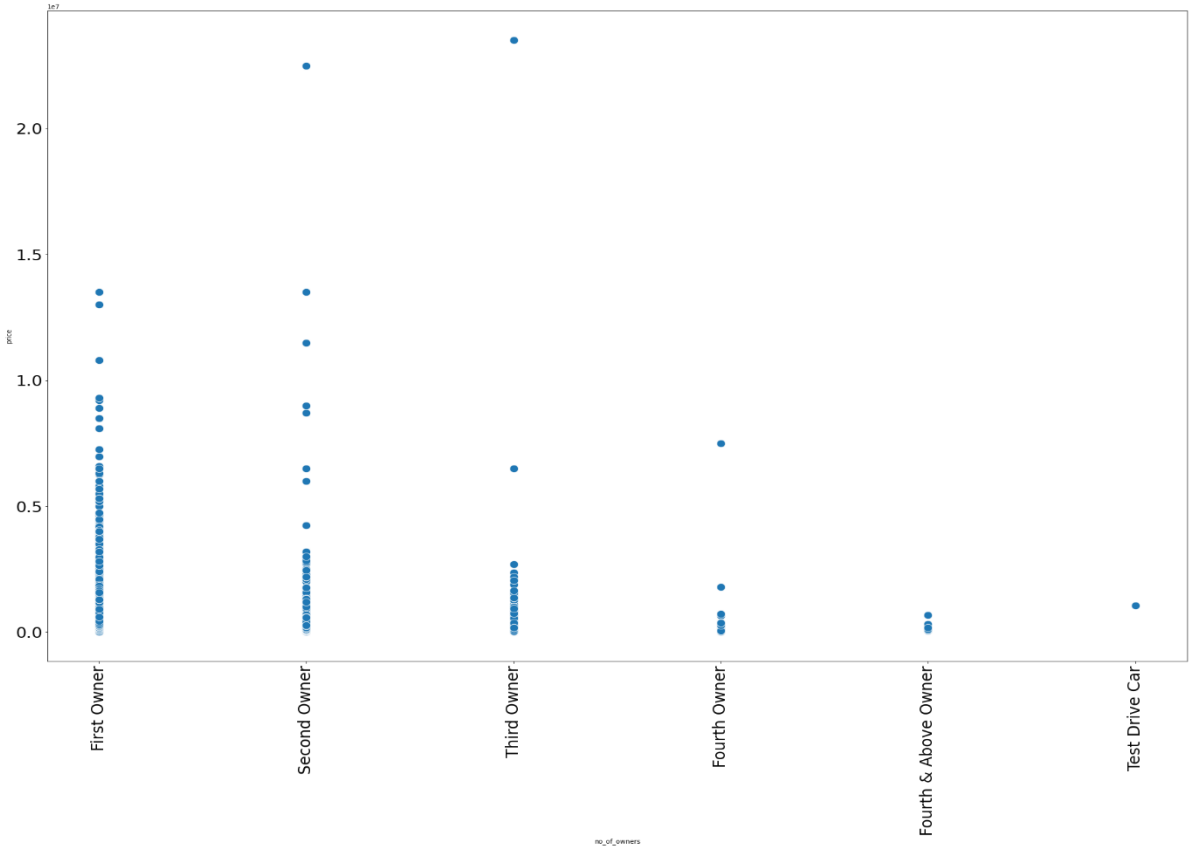
Distplot of the price



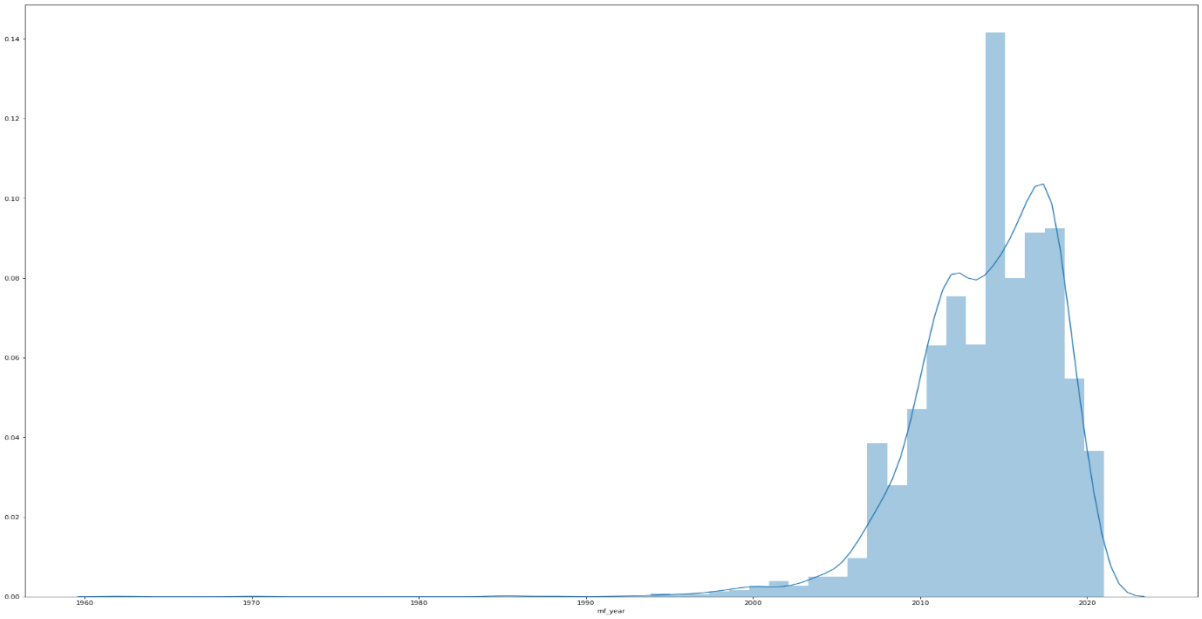
Fuel type with price

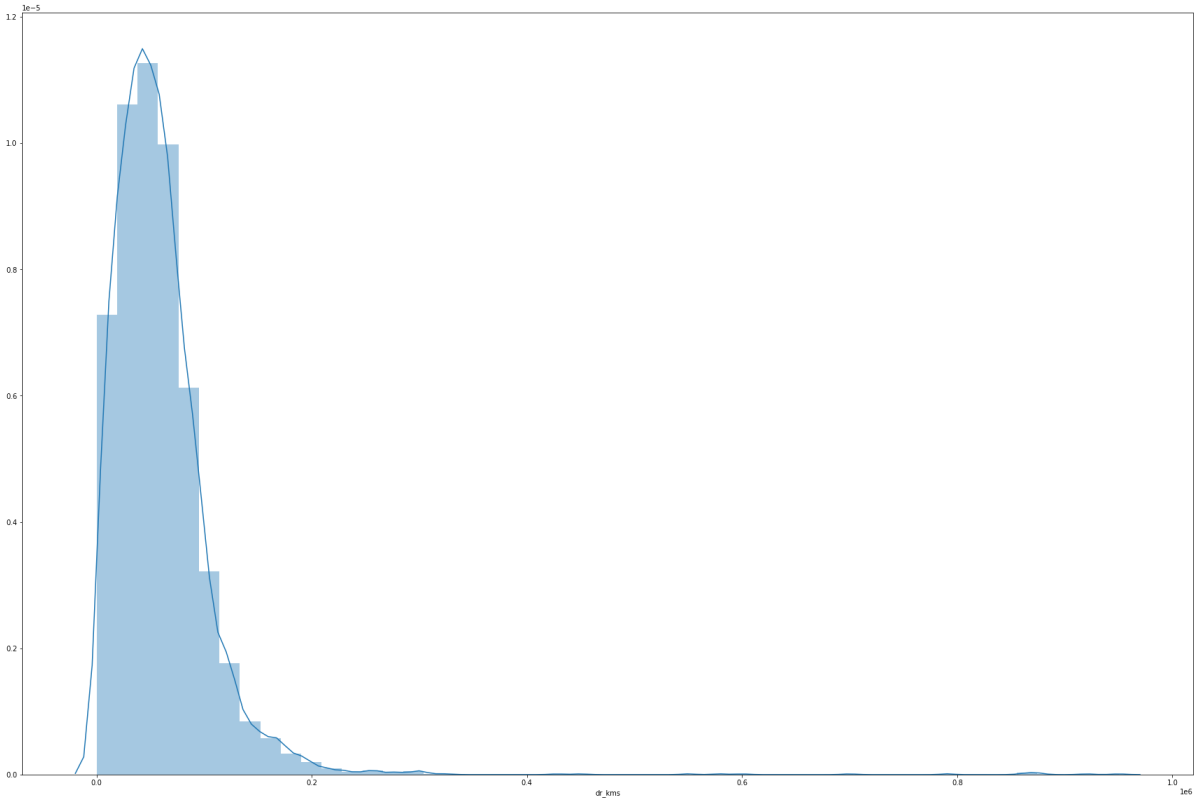


No.of owners with price

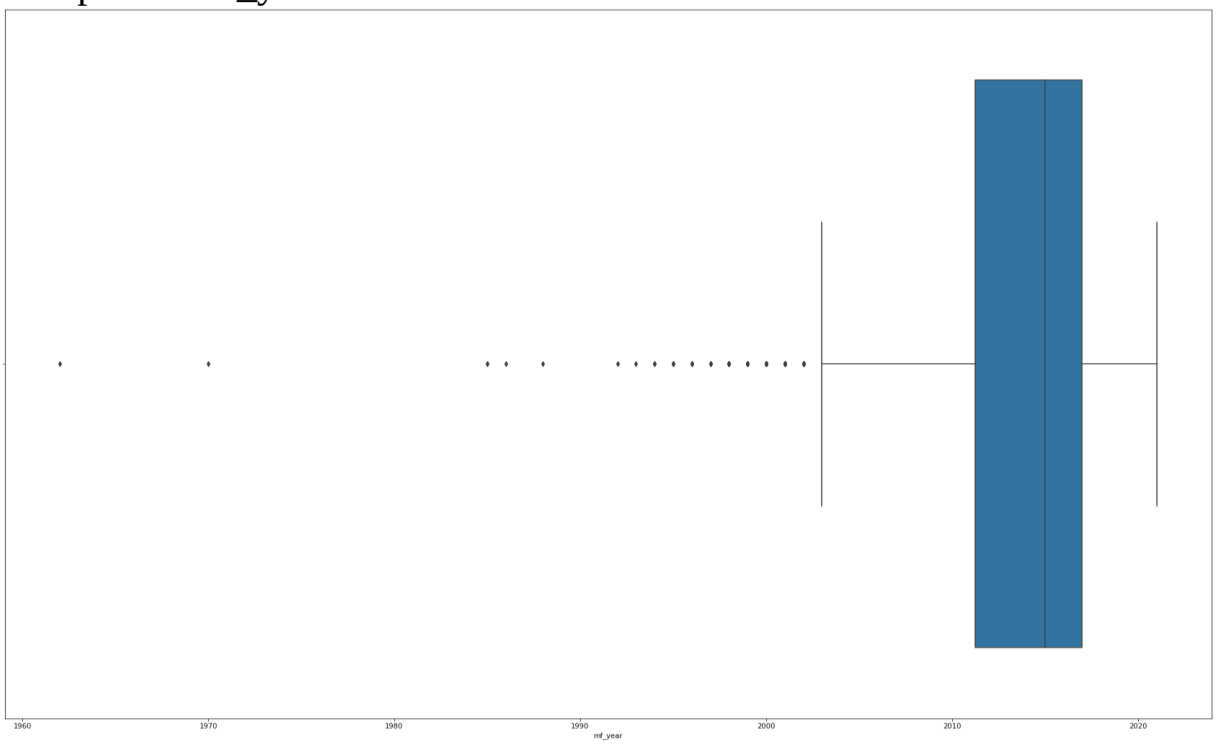


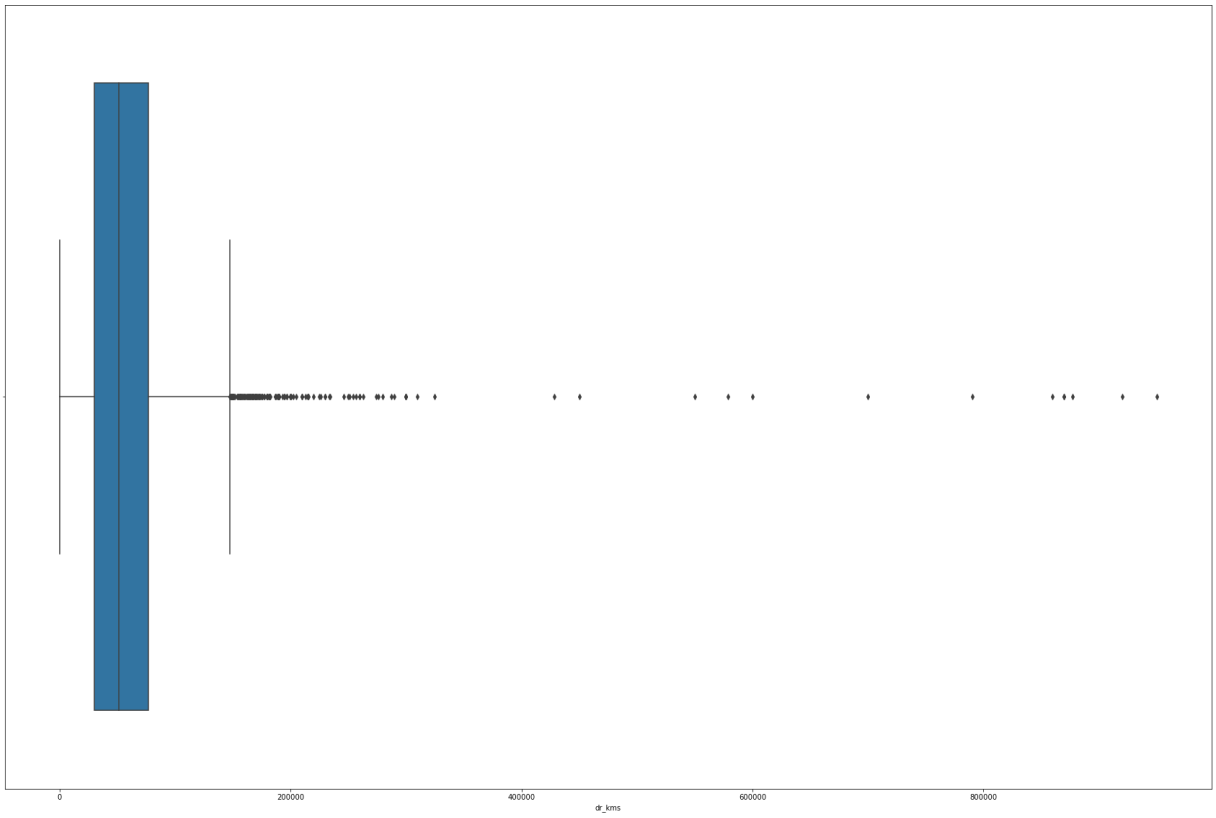
Distplot of mf_year and driven kilometers



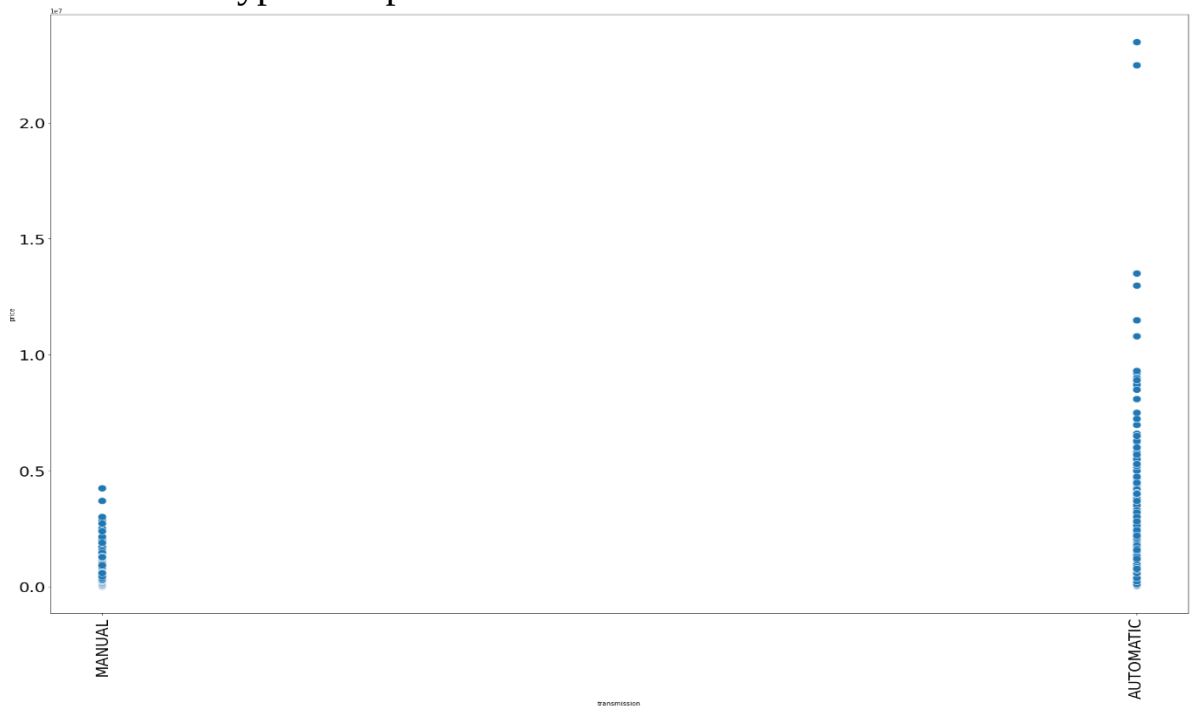


Boxplot of mf_year and driven kilometers

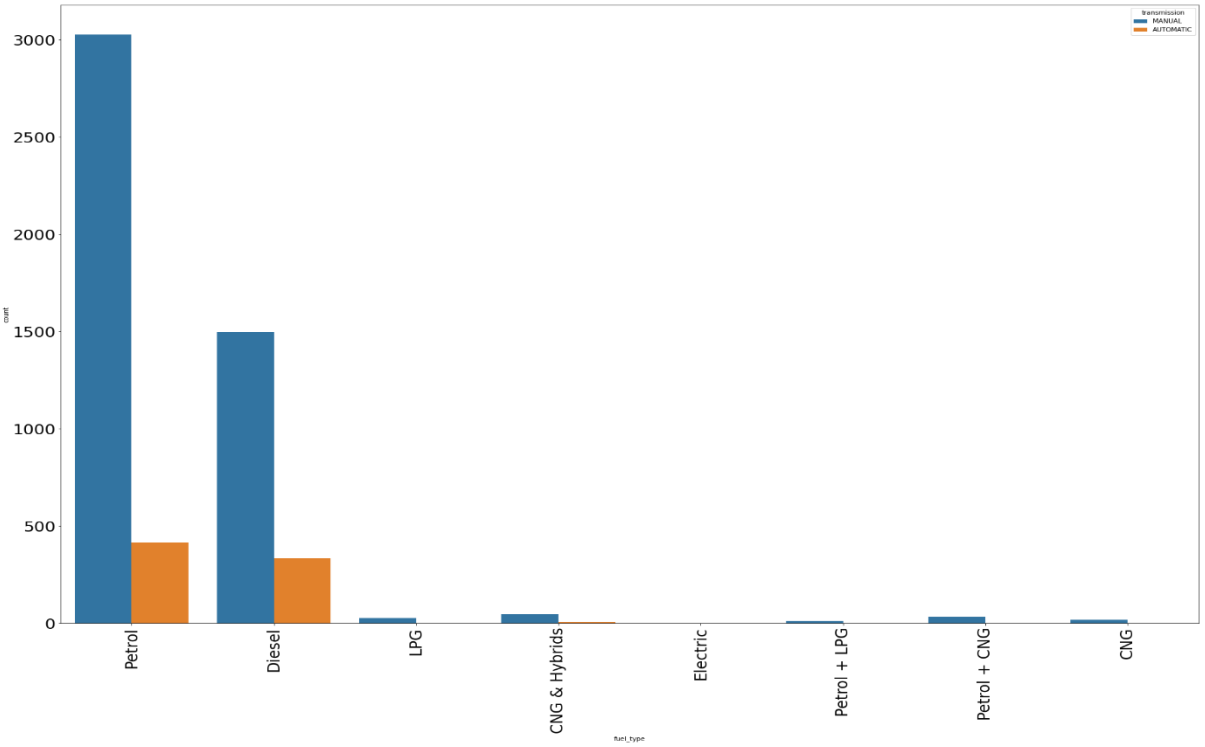




Transmission type with price



Count of vehicles with different fuel types



· INTERPRETATION OF THE RESULTS

- Most of the vehicles in the used car industry are Maruti and Hyundai
- Premium vehicles are very less
- The most expensive cars are Ferrari and Lamborghini
- Hyundai, Datsun, and Maruti are some of the budget-friendly brands
- lesser the age of the vehicle higher will be the price
- we can see that the purchasing of vehicles started booming around 2010
- from 2018, it has been started to decline
- Most of the vehicles are petrol
- Gas or hybrid vehicles are very less
- No electric vehicles in our dataset
- In used car industry, uncommon fueled vehicles are cheaper(CNG, LPG etc) than petrol and diesel
- Generally, we can say that when no. of owners increases, price decreases
- The prices of the car is not much different from city to city
- Automatic cars are more expensive
- In every fuel type, the manual is higher in number than automatic
- driven kilometers and mf_year are highly -ve correlated

· KEY FINDINGS AND CONCLUSIONS OF THE STUDY

Found out the key features that related to the price of a used car and was able to make a machine learning model that predicts the car price.

· LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

Got the opportunity to work with different Ml algorithms like SVR, KNN, etc

· LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

The accuracy of the ML model is low. We got an accuracy of around 54.7%.

- Learning outcomes of the study in respect of Data Science

With the help of exploratory data analysis we could visualize the large data and make interpretation of it and also summary statistics helped us understand how data is spread, where outliers are present to what extend deviation is present and understanding the relationship between target column and other attributes, repeated columns was eliminated to making right prediction.