FLIP ROBO

# CAR PRICE PREDICTION

Submitted by:

SANDES P

# ACKNOWLEDGMENT

I used informative tutorials to follow some steps in the task from websites like geeksforgeeks, stackoverflow, etc.

# INTRODUCTION

- Business Problem Framing

  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.

- Conceptual Background of the Domain Problem

  There are a lot of used car-selling websites out there. Our goal is to make a model that predicts the car pricing given the details.

- Review of Literature

  Used car price depends on a lot of factors like age of the car, driven kilometers, etc. Most of the cars in the used car realm are budget cars made by Maruti, Mahindra, Hyundai, etc. Online sales (not price) have boomed during this pandemic.

- 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.

## Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Here we 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 data required to build this model is scraped from different used-car selling websites 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

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

- 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

- 8GB RAM
- i5 7th gen processor

Softer requirements

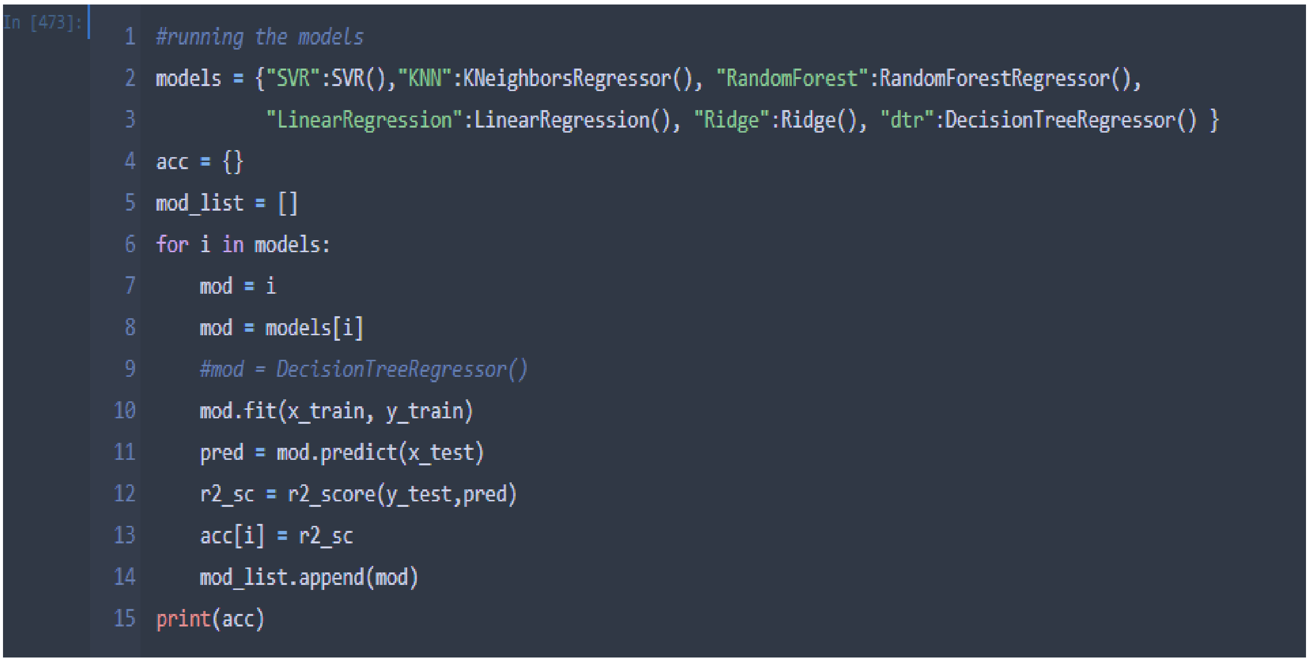
- Python
- Jupyter notebook

Libraries

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Re
- Sklearn
- joblib

# Model/s Development and Evaluation

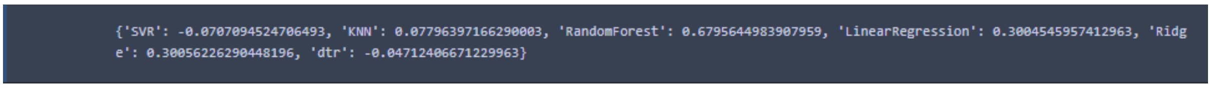
- Identification of possible problem-solving approaches (methods)
    - Treating outliers, missing values
    - EDA for finding the relationship
    - Encoding
    - Scaling
- Testing of Identified Approaches (Algorithms)
    - SVR
    - KNN
    - RandomForest
    - Linear Regression
    - Ridge
- Run and Evaluate selected models



In [473]:

```
#running the models
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



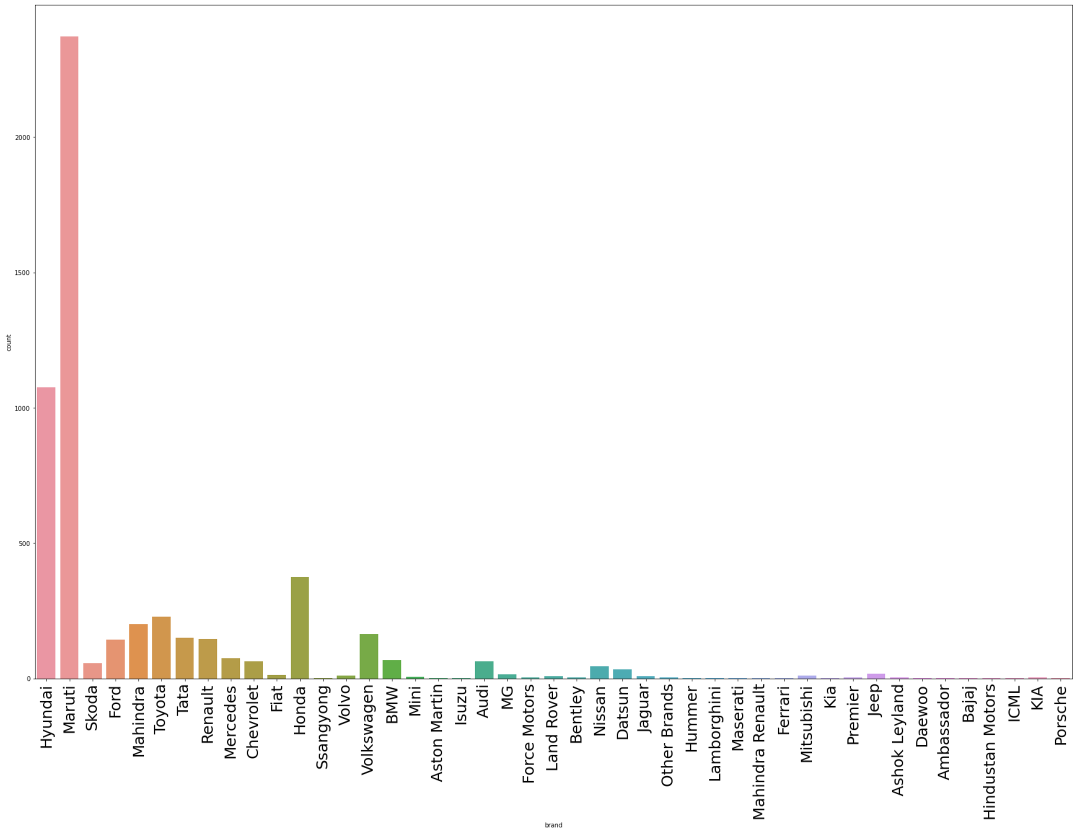
```
{'SVR': -0.0707094524706493, 'KNN': 0.07796397166290003, 'RandomForest': 0.6795644983907959, 'LinearRegression': 0.3004545957412963, 'Ridg
e': 0.30056226290448196, 'dtr': -0.04712406671229963}
```

- Key Metrics for success in solving problem under consideration

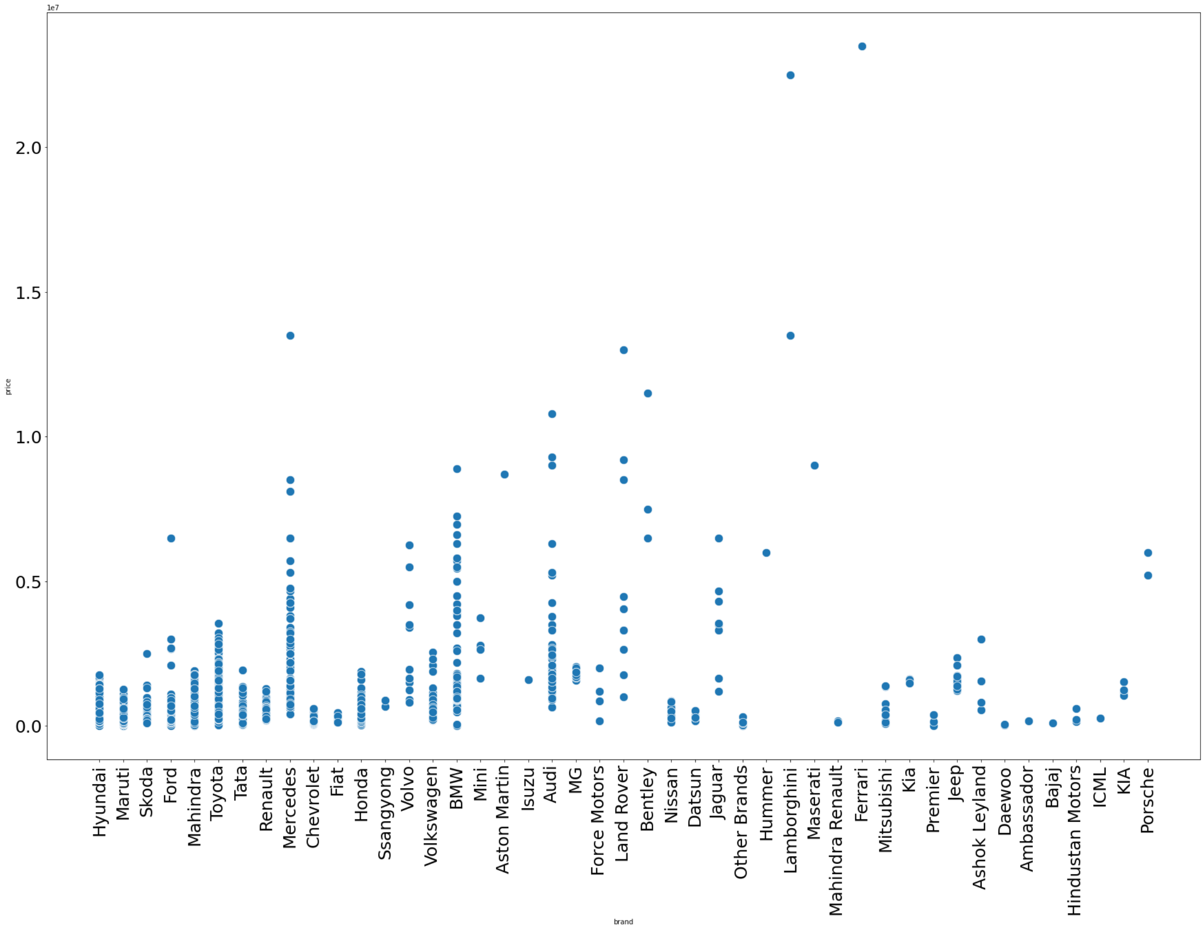
  Here we used r2score since this is a regression problem

- Visualizations

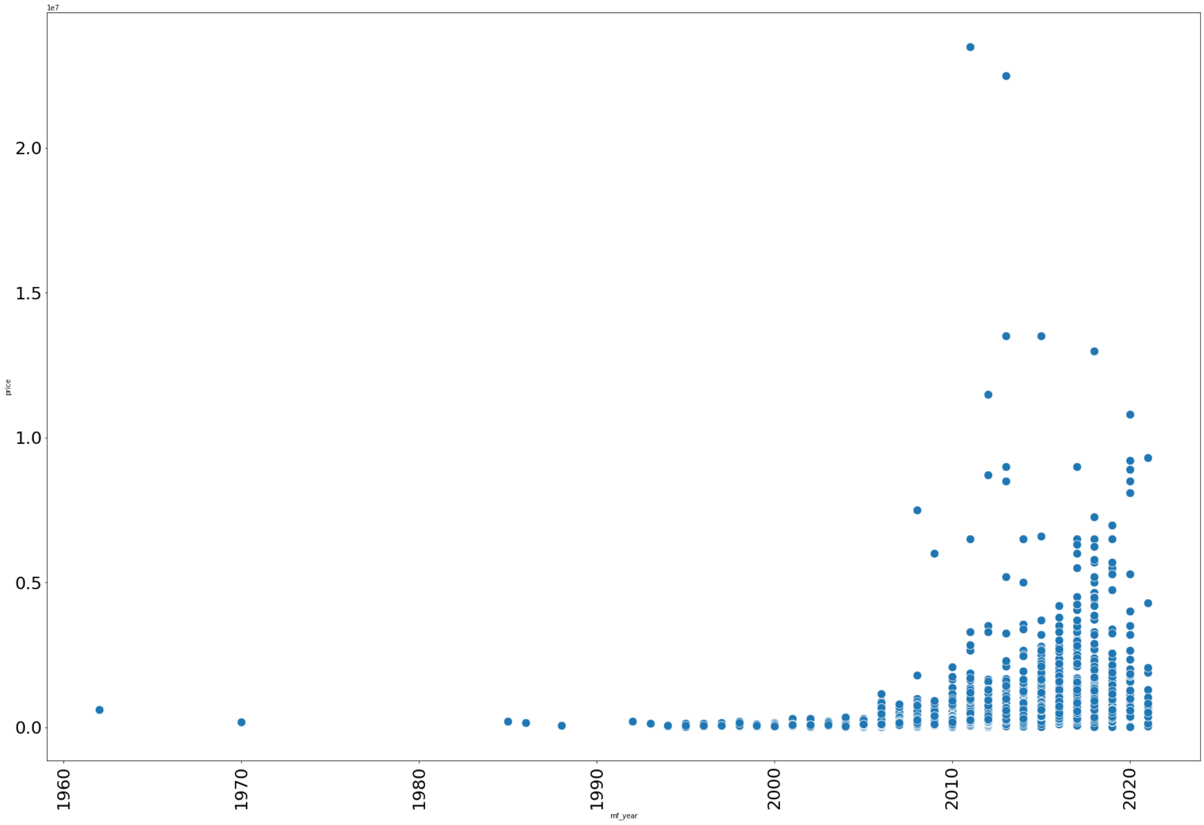
  Different brands



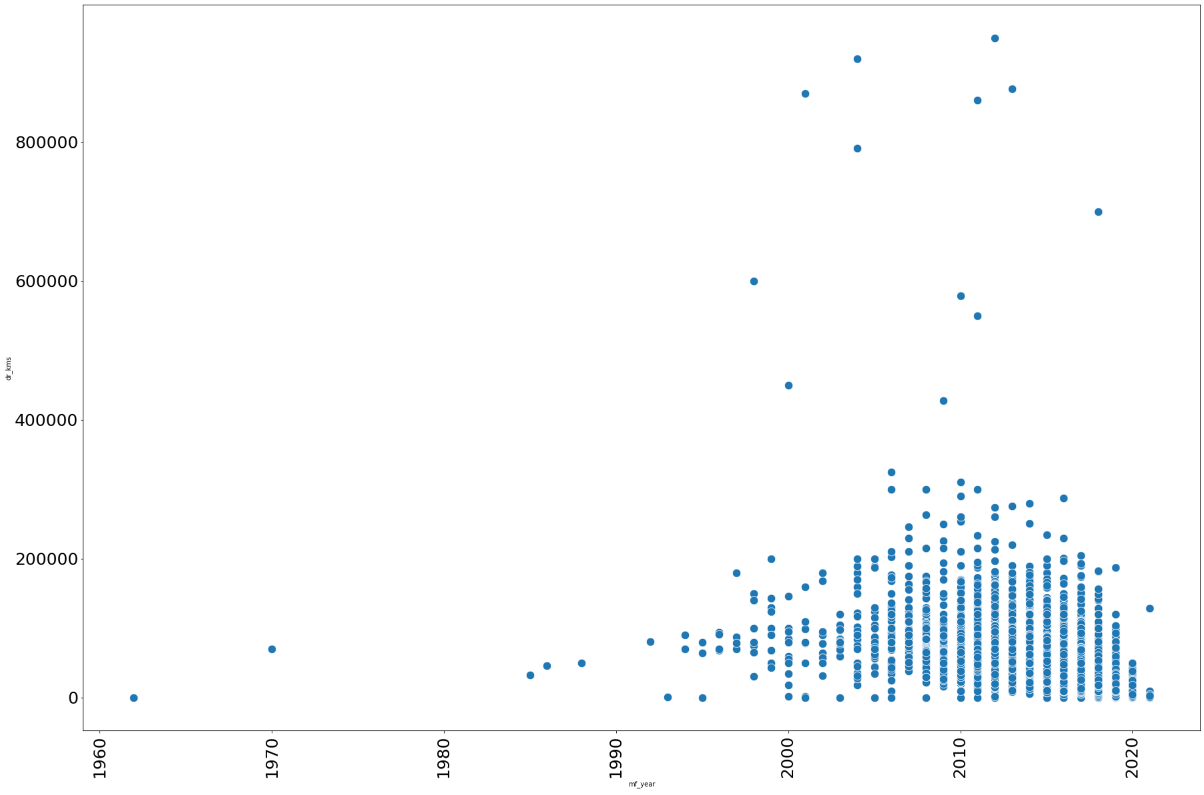
  Brands with price



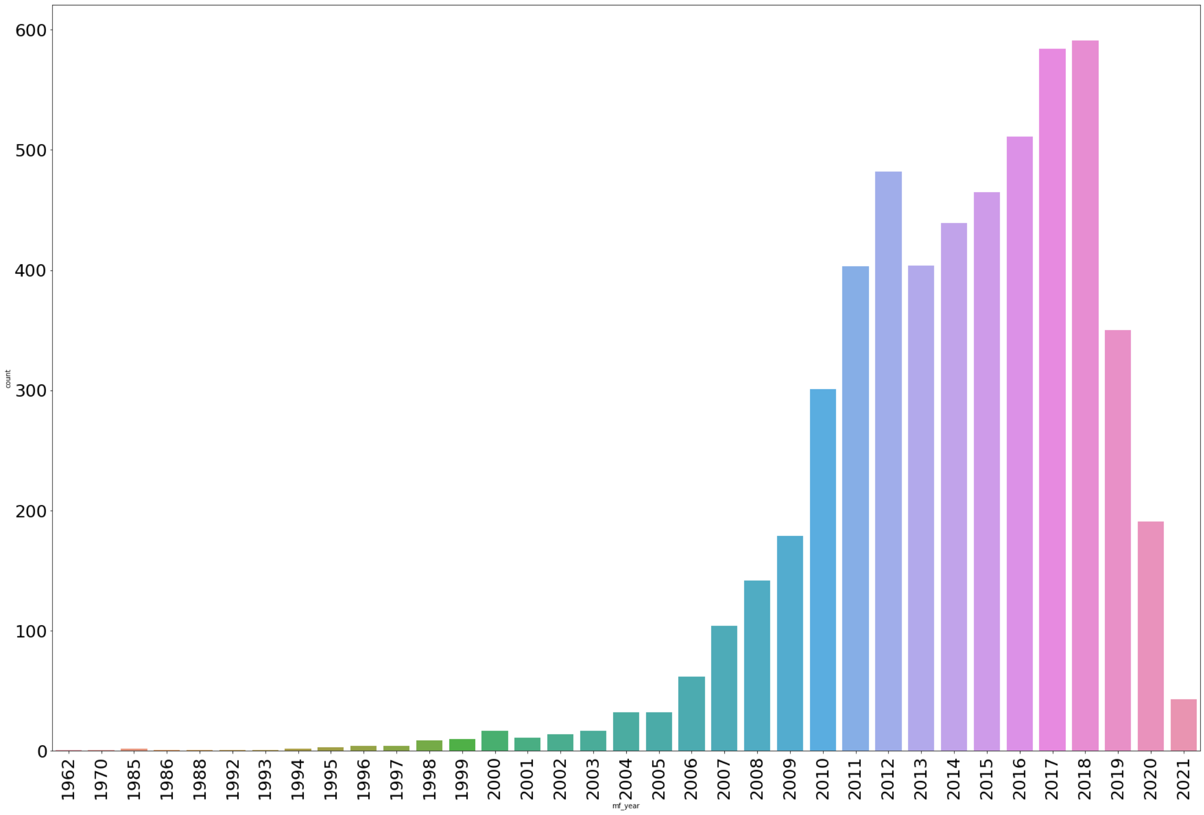
# Price with year



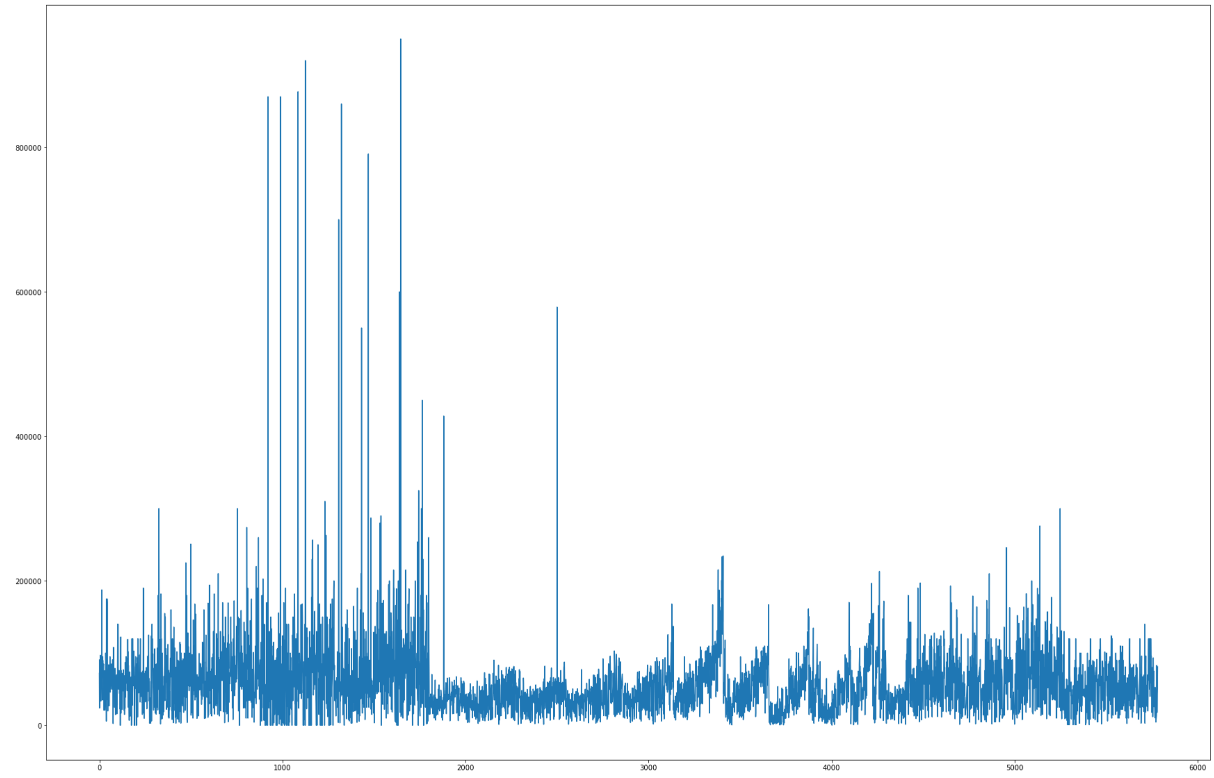
# Manuacture year with driven kilometers



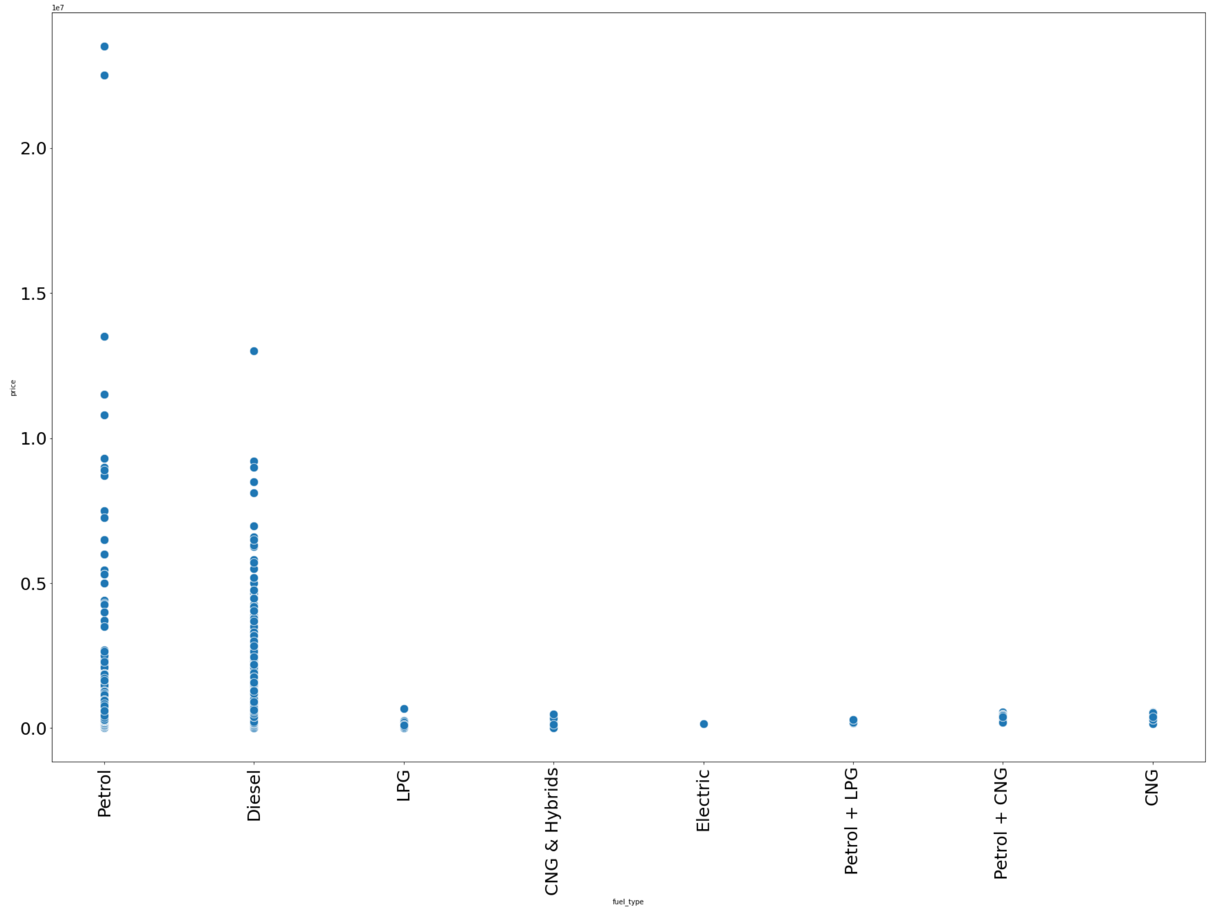
# Count with a manufacture year



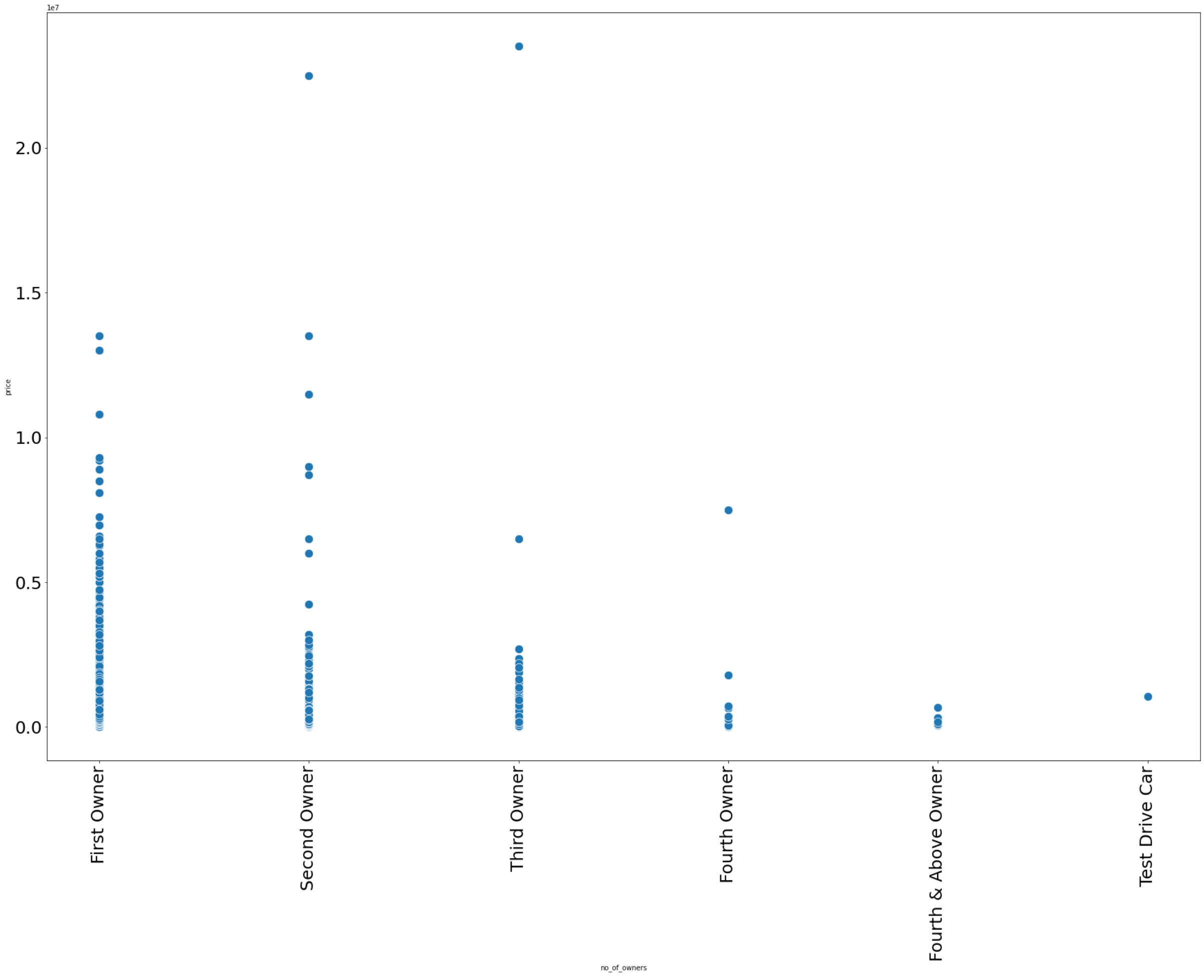
# Distplot of the price



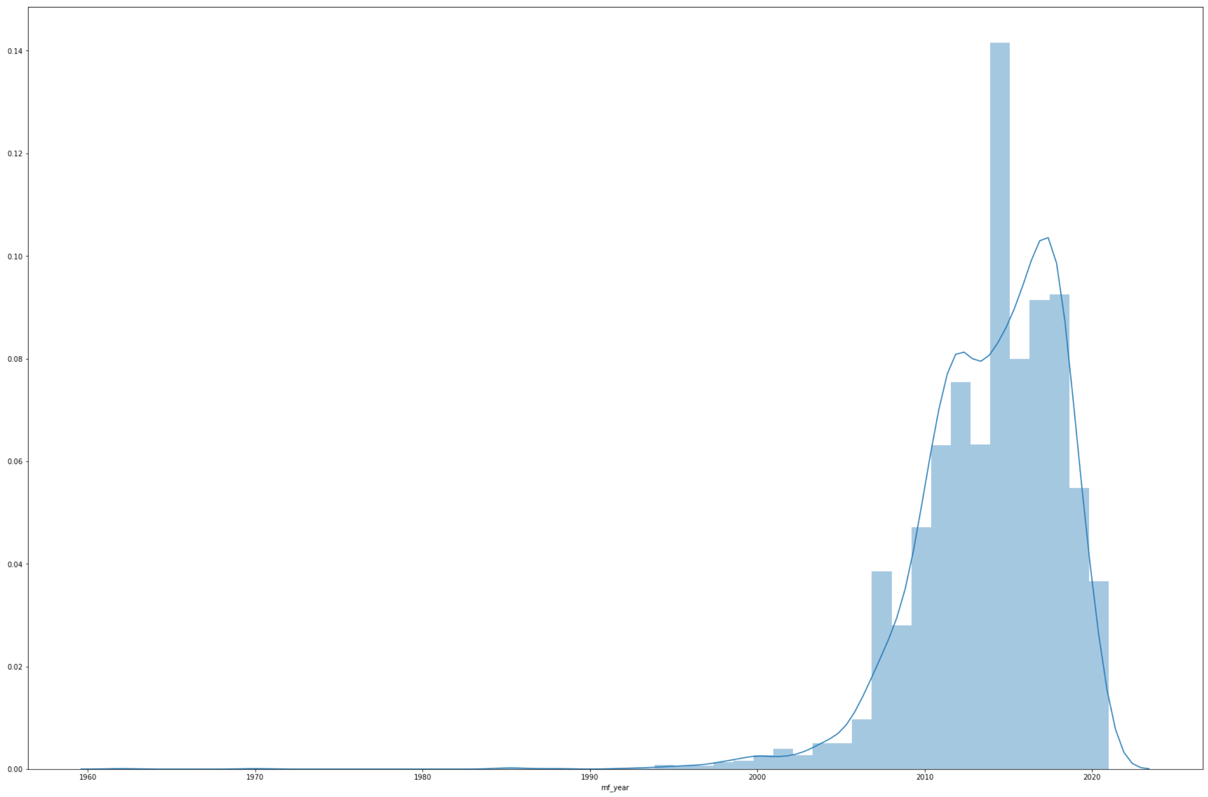
# Fuel type with price

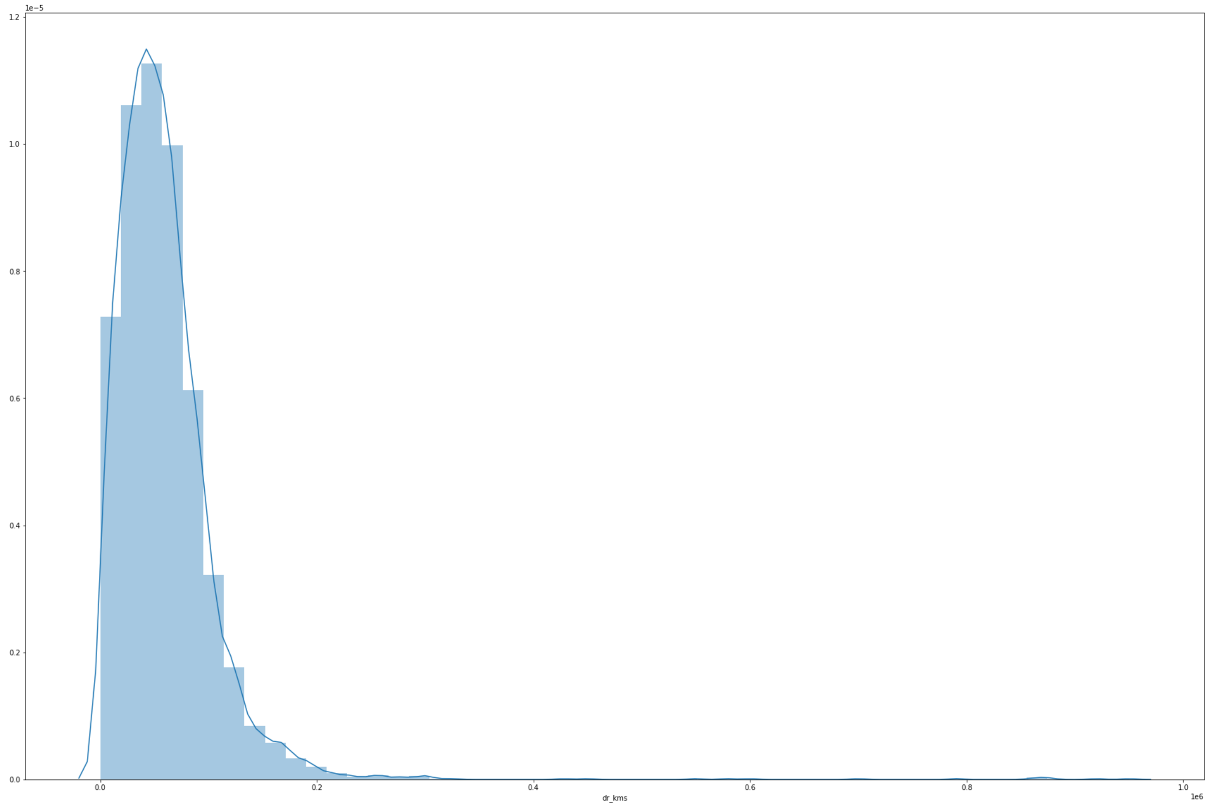


# No.of owners with price

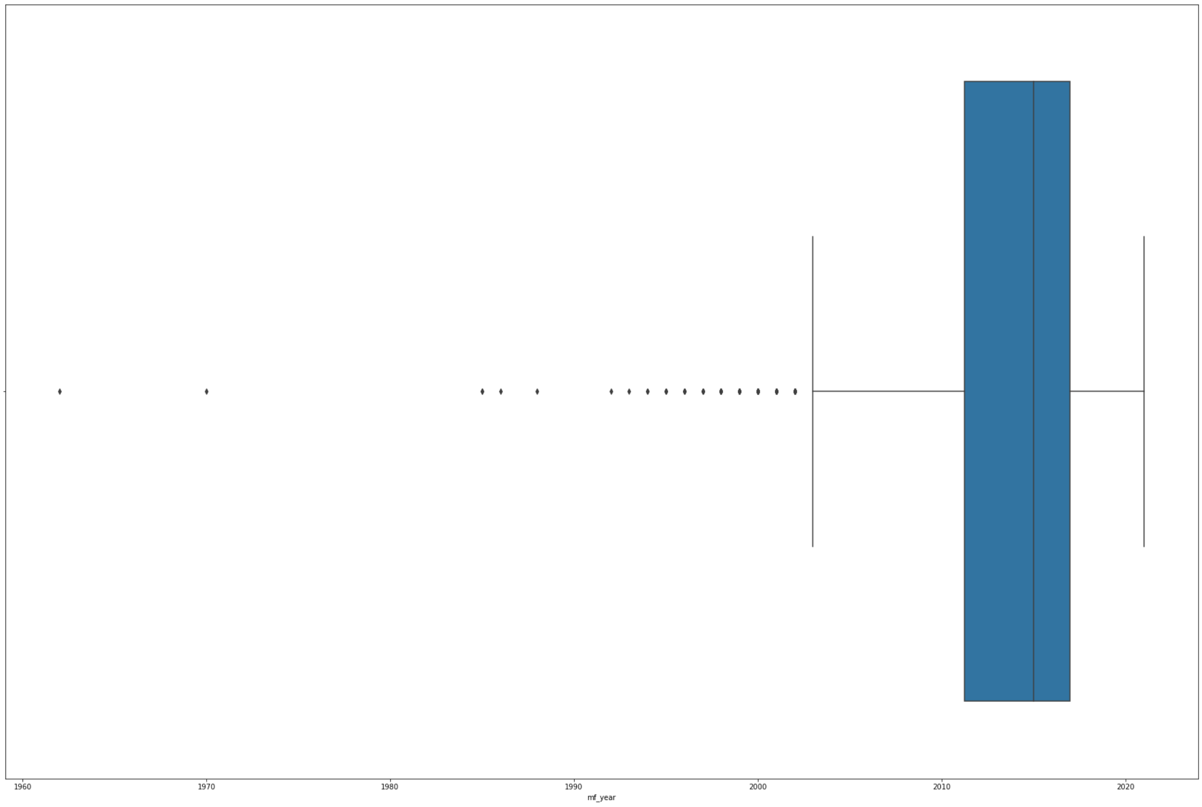


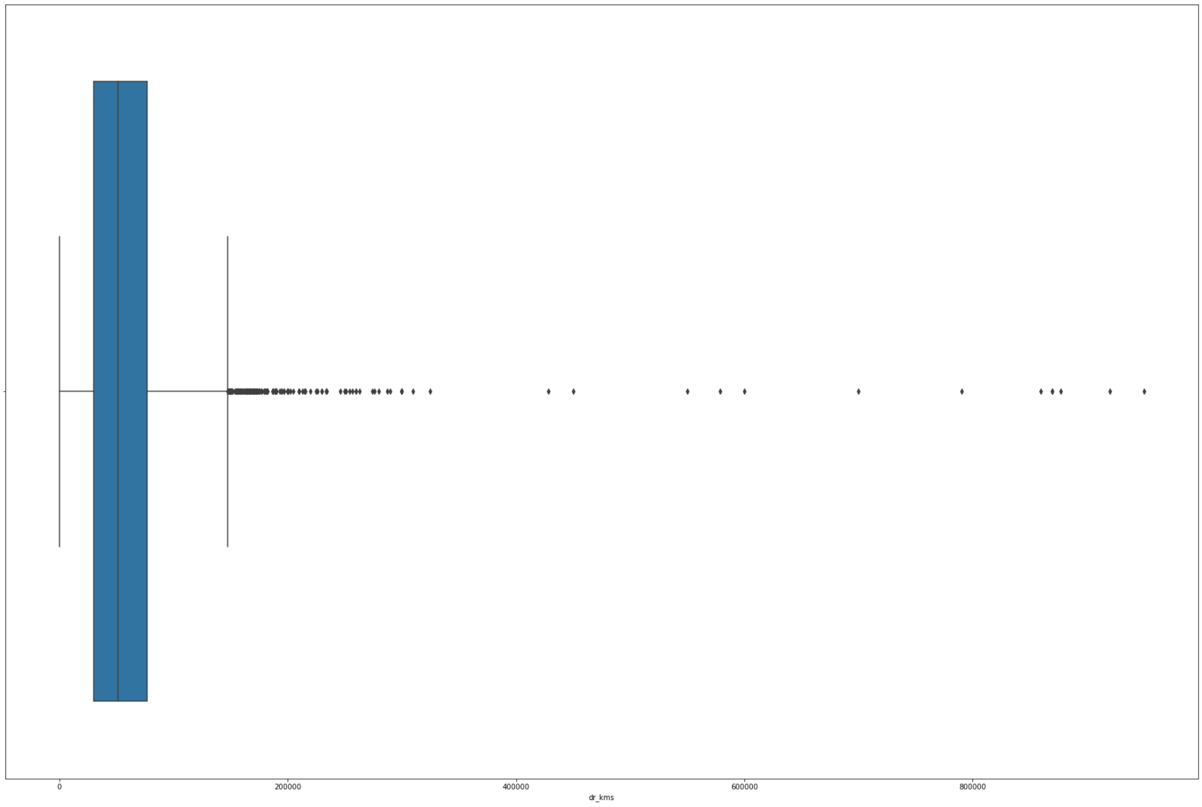
# Distplot of mf_yaer and driven kilometers



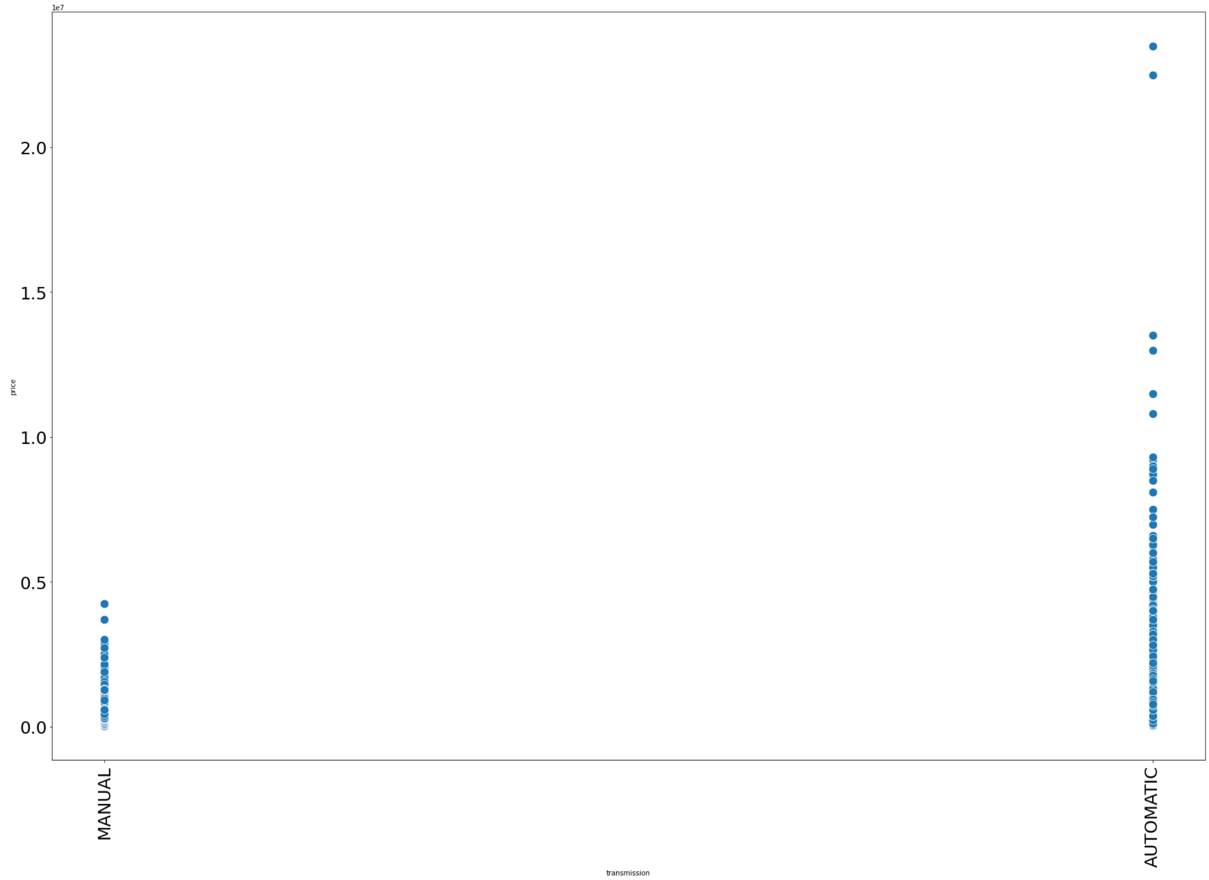


# Boxplot of mf_year and driven kilometers

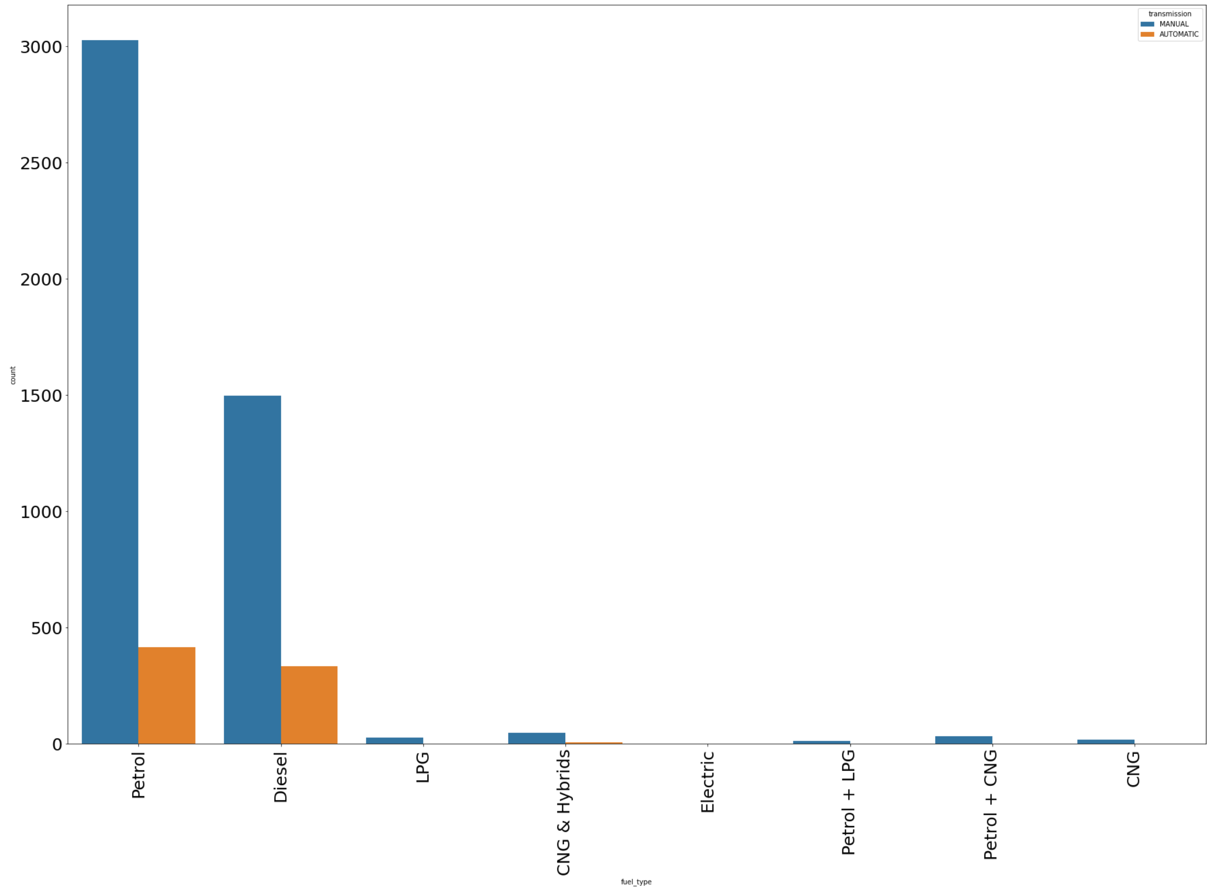




## Transmission type with price



# 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

# CONCLUSION

- 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 65%.