Mahesh Babu Vusa

Predicting the price of pre-owned cars

Abstract

In this project I built a simple machine learning model called linear regression model and random forest regression model in order to predict the price of pre-owned cars for a given data set. The price of the car varies on different features like year of registration; power, kilometers travelled, model of the car, fuel type etc.

Introduction: The second hand cars in the market are abundant and not everyone are able to buy first hand cars, a company wanted to sell the pre-owned cars and wanted an website for the customers who wanted to sell and buy car needed an machine learning model to predict the price to sell and buy depending on the different features like age, power, vehicle type, kilometers travelled, model of the car, type of gear fuel type etc.

Target set: price of the car

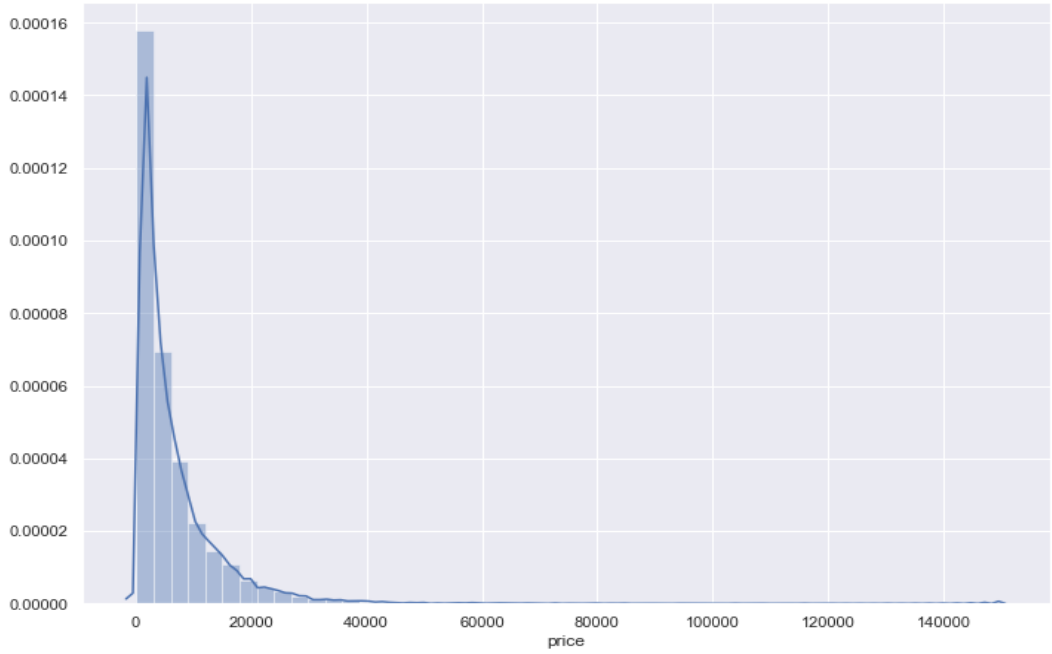
Features set: year of registration, power, Model of the car, gear box type, vehicle type, fuel type and kilometers travelled etc.

Project Objective: The goal of this project is to use techniques of data science to predict the price of a pre-owned car depending up on different features.

Exploratory Data Analysis: The data set was important in Python and calculations were performed using python (jupyter Notebook). We plot the following figures:

Variable: price

Histogram:

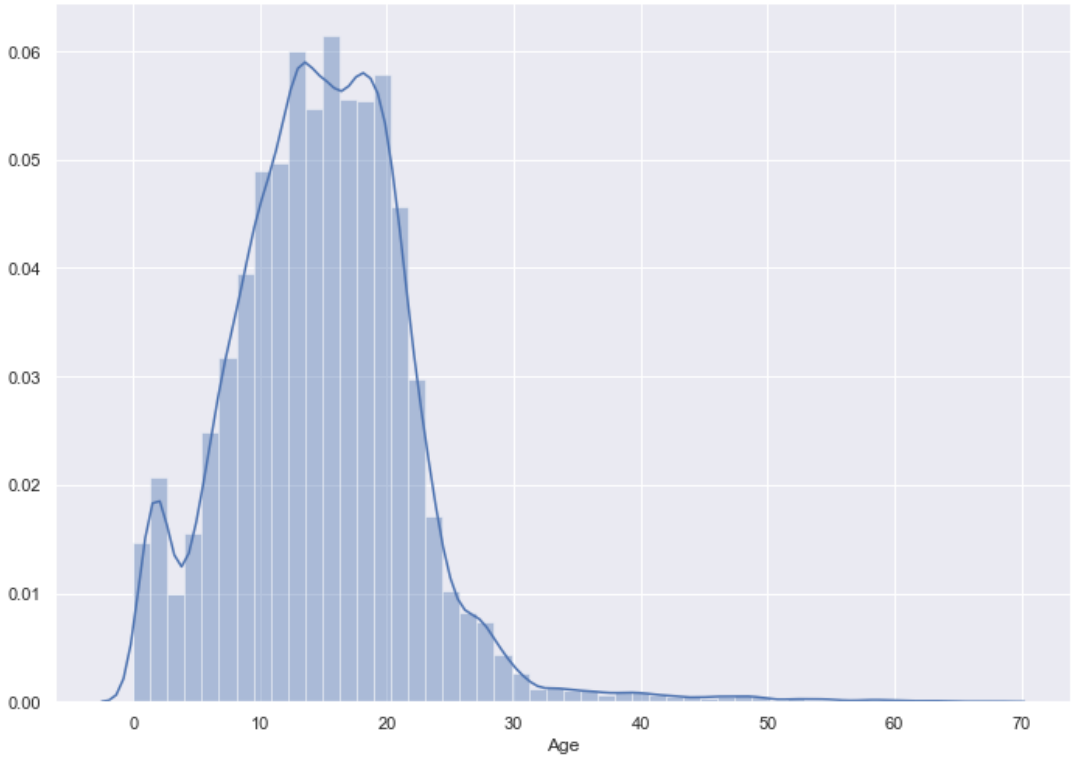


Box plot: price

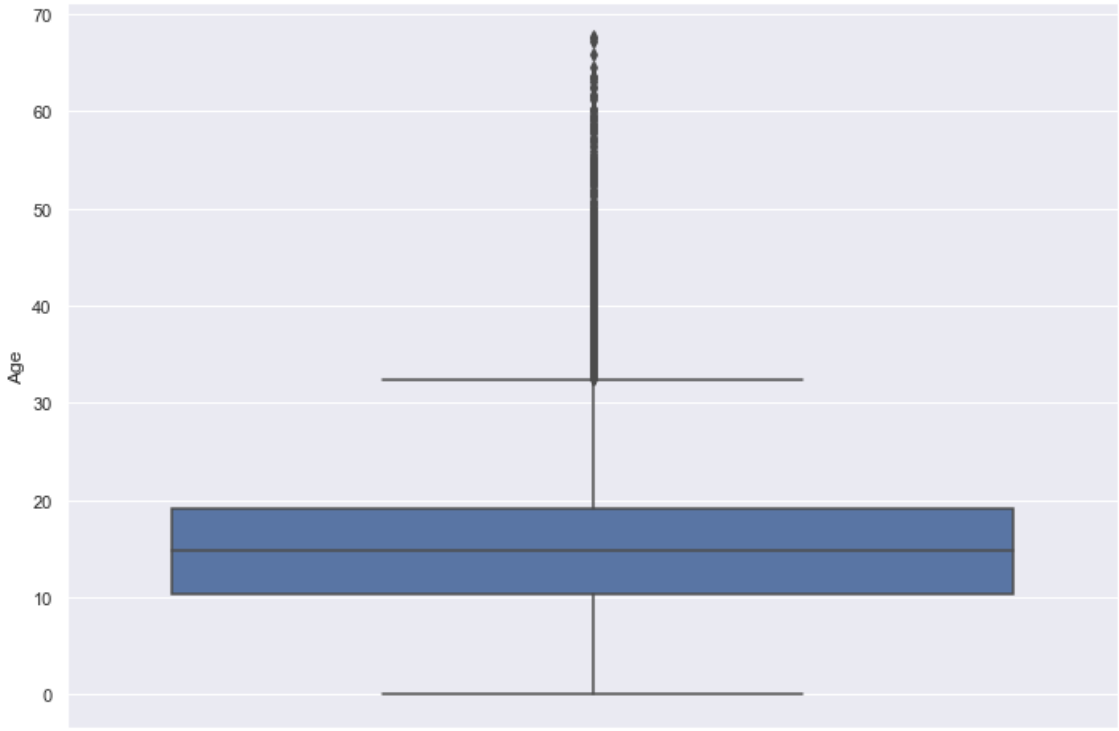
Variable: **Age**

It is obtained from manipulating year of registration and month of registration.

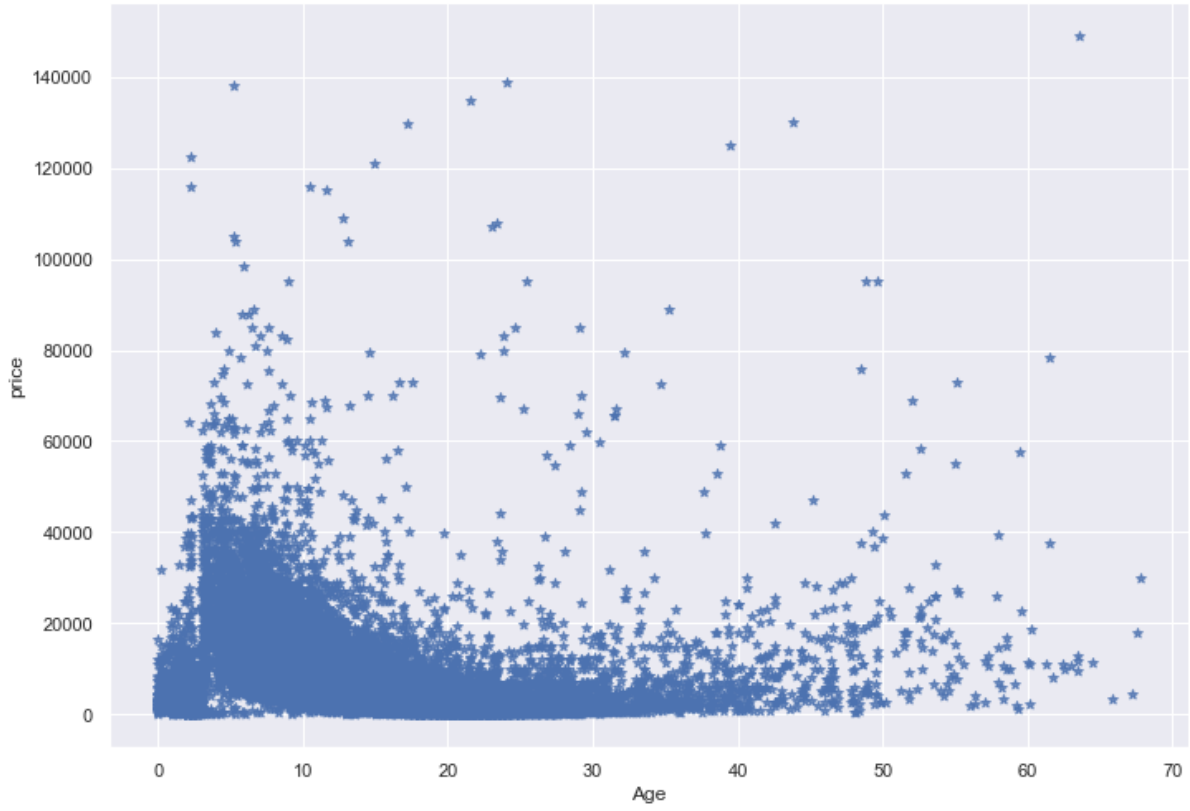
Histogram:



Boxplot:

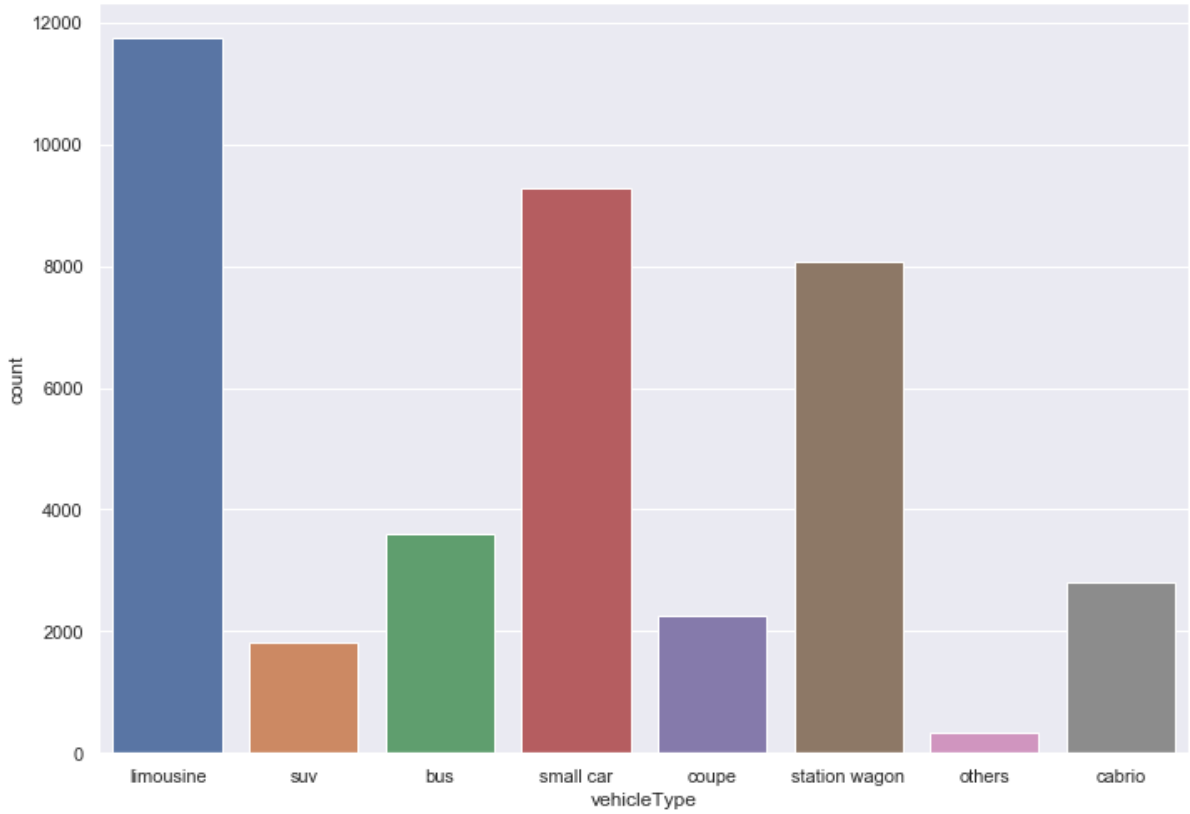


Scatter plot: Age vs price

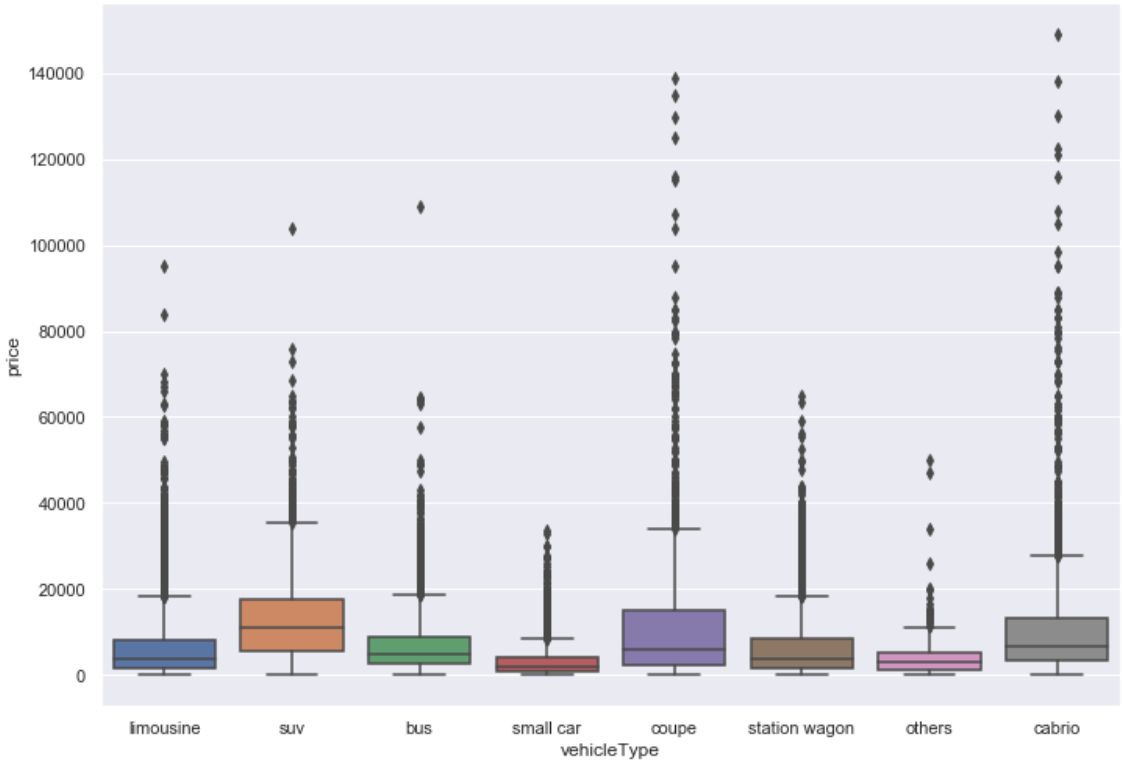


Variable: vehicle type

Histogram:

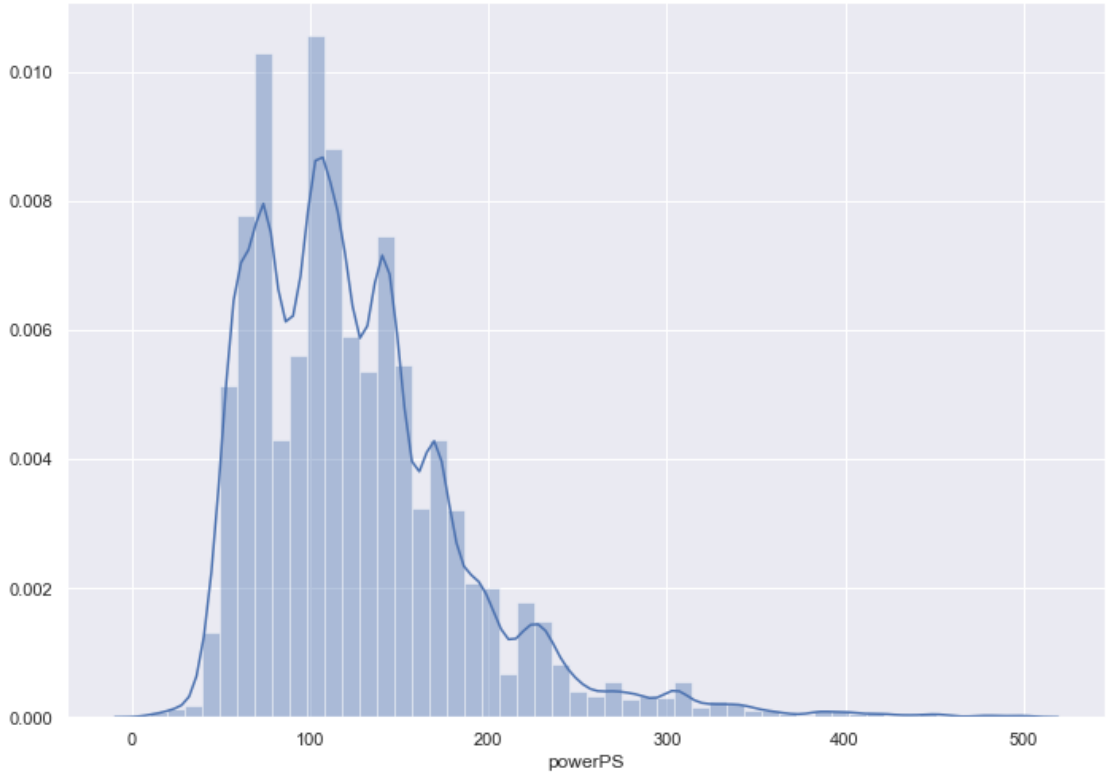


Boxplot: vehicle type vs price

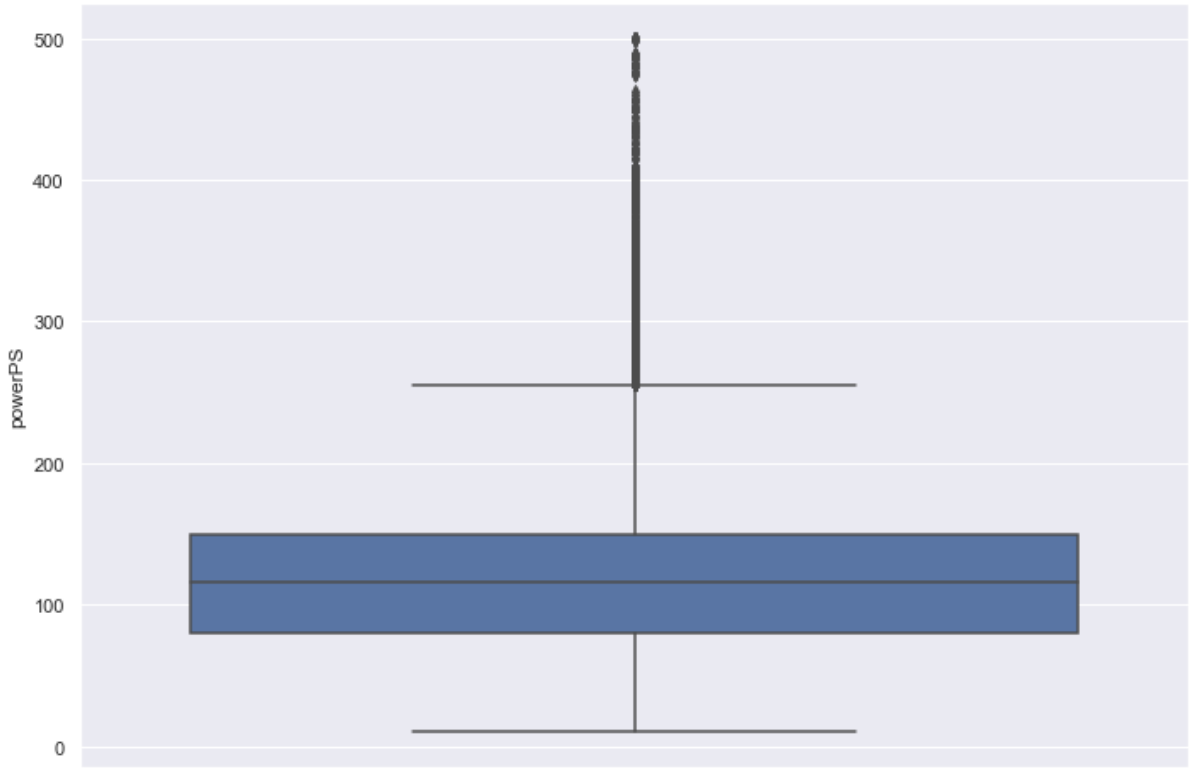


Variable: Power ps

Histogram:

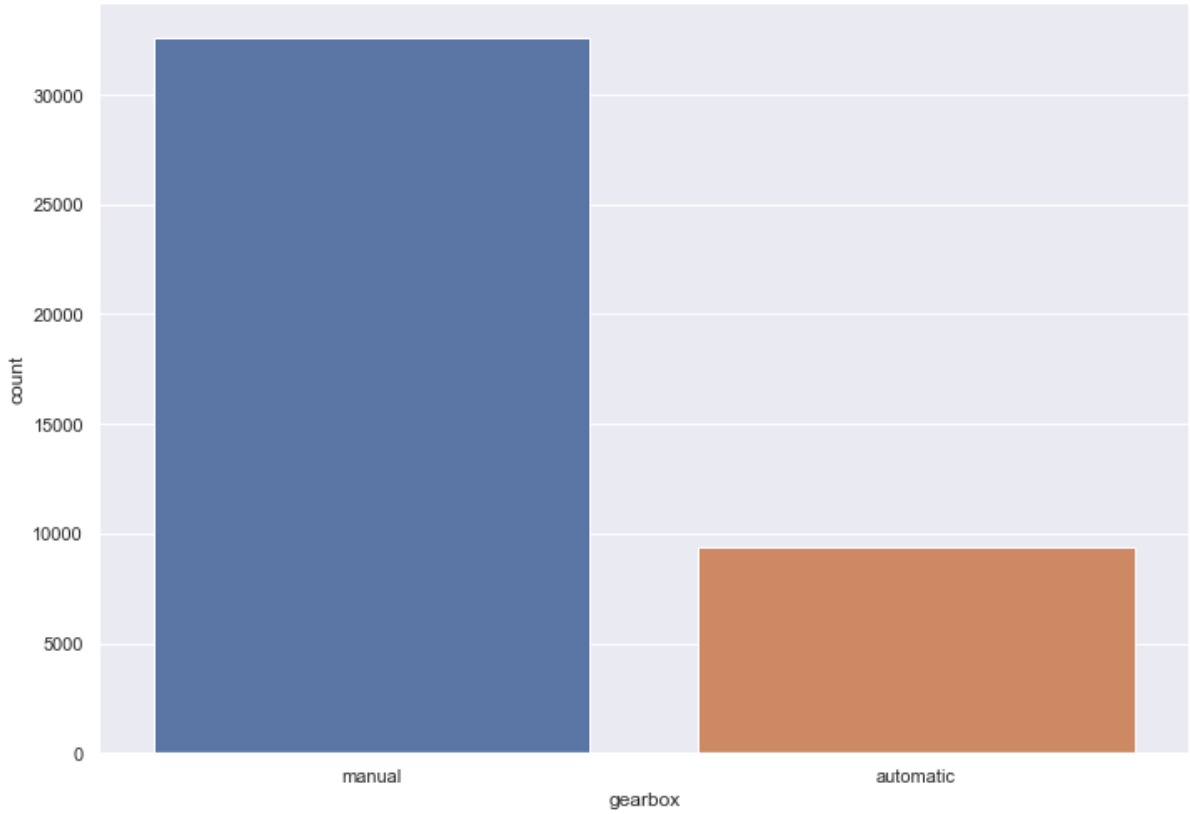


Boxplot :

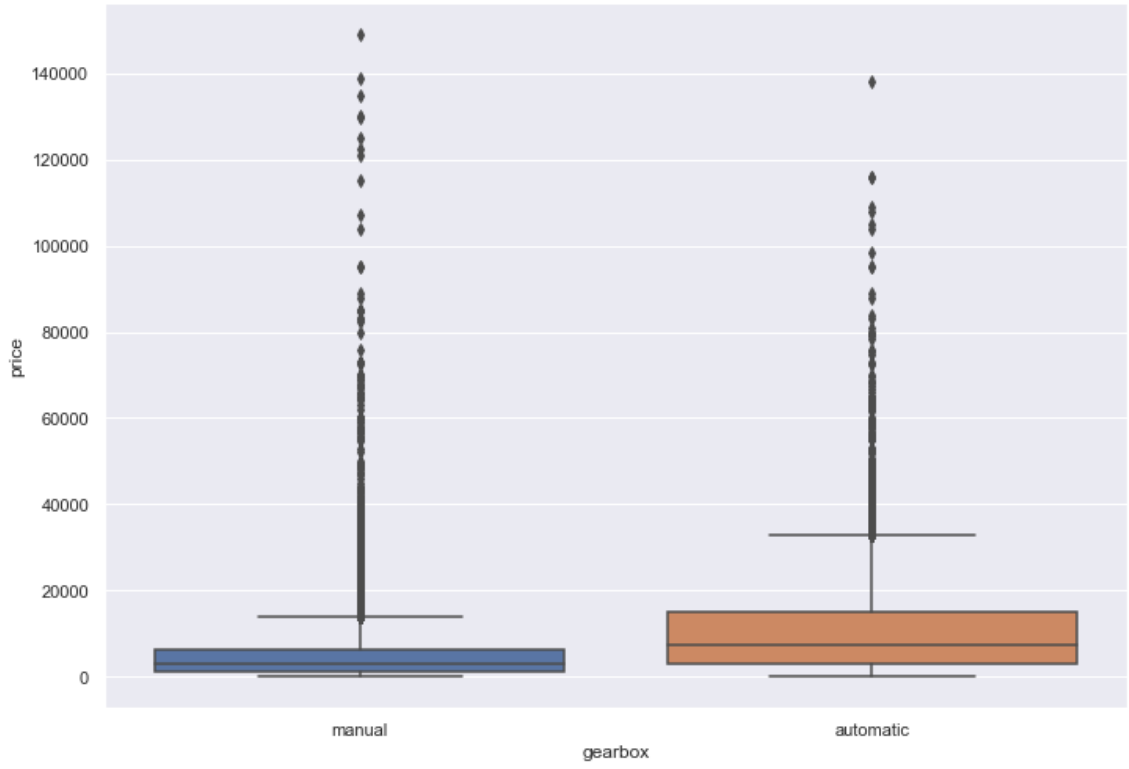


Variable: Gear box

Histogram:

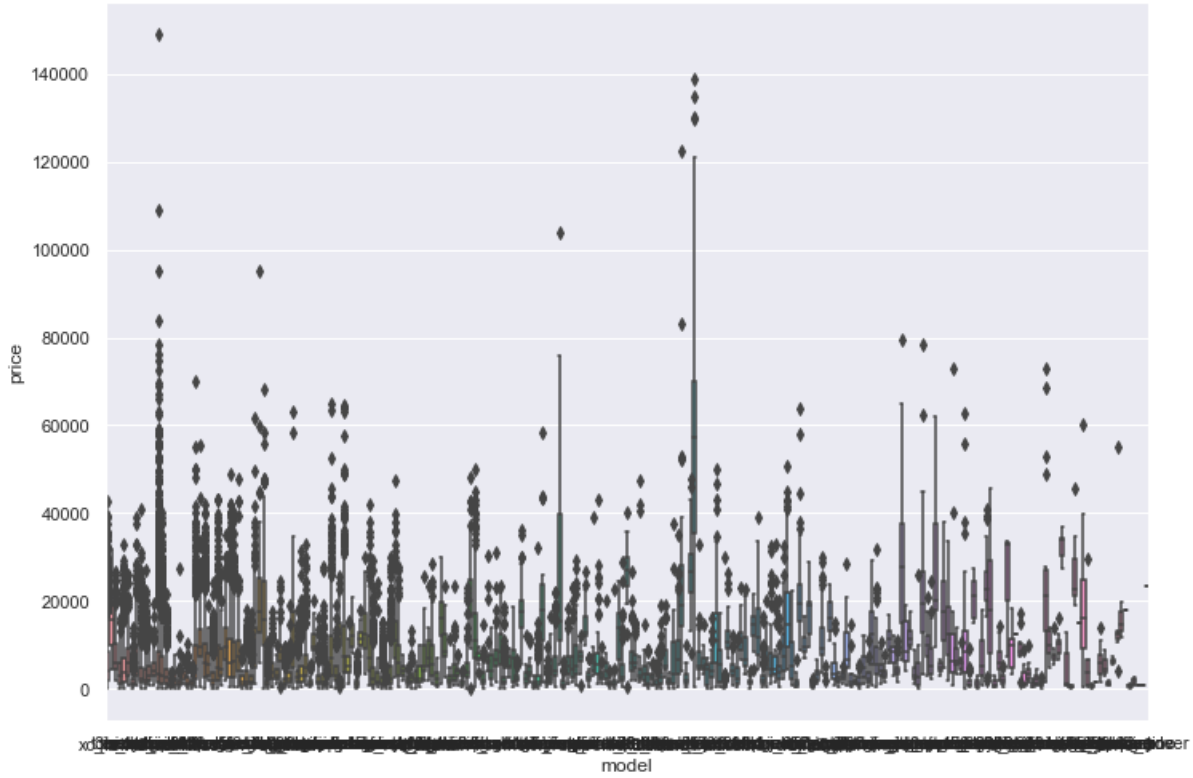


Box plot: Gear box vs price



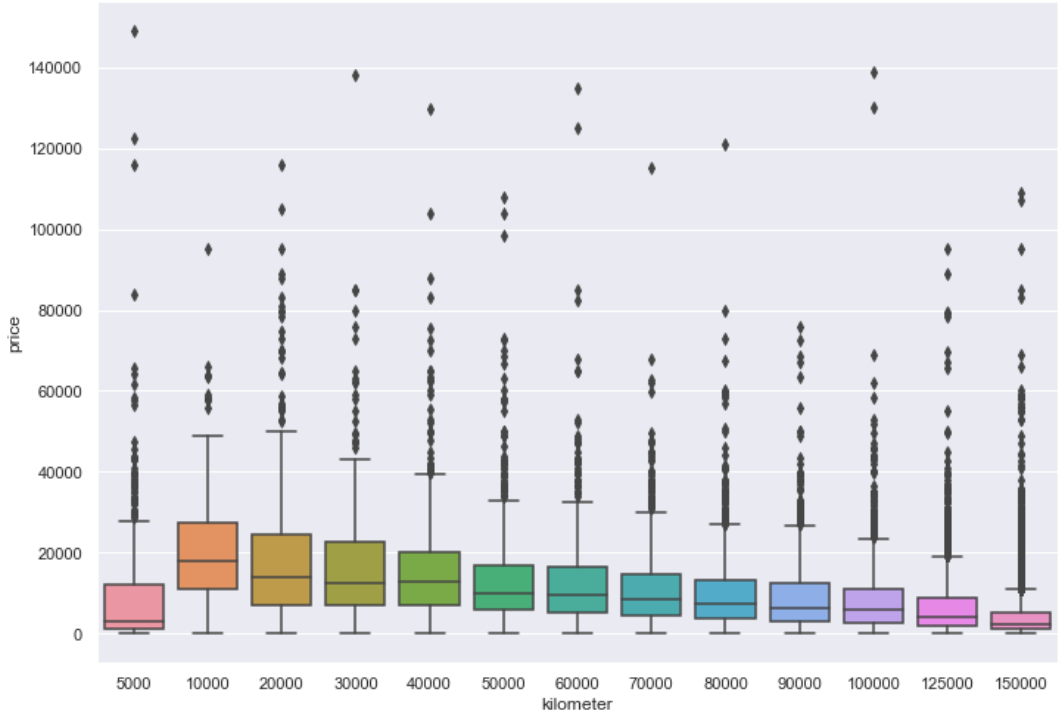
Variable: Model

Boxplot: model Vs price



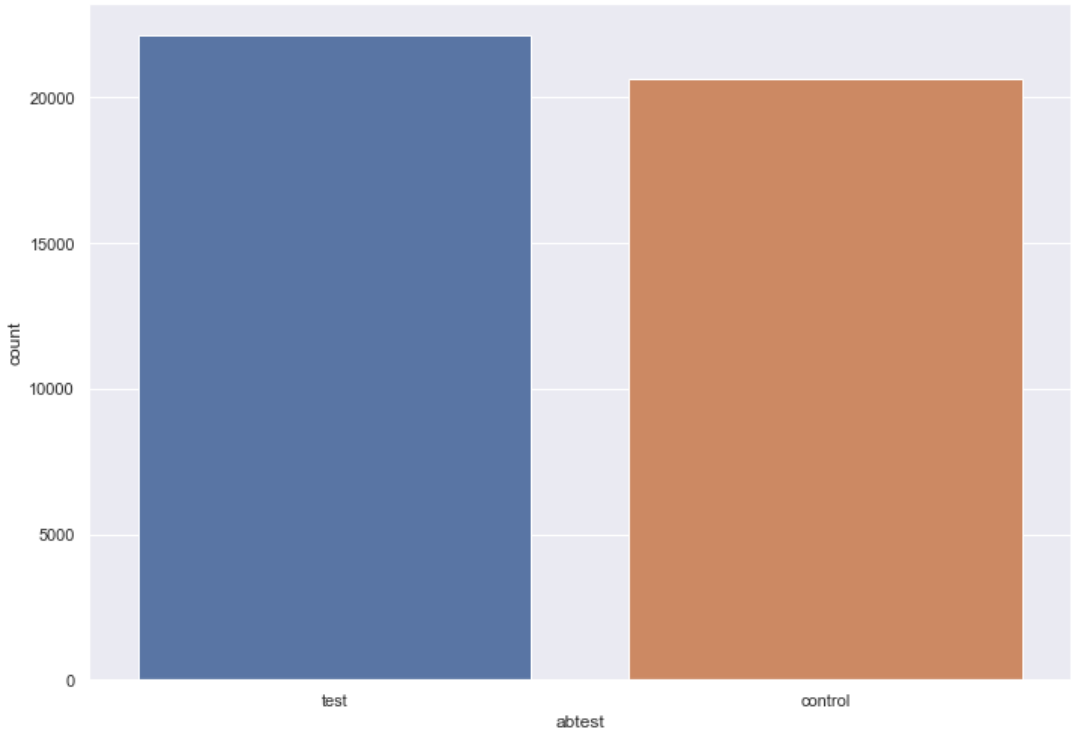
Variable: Kilometre

Boxplot:

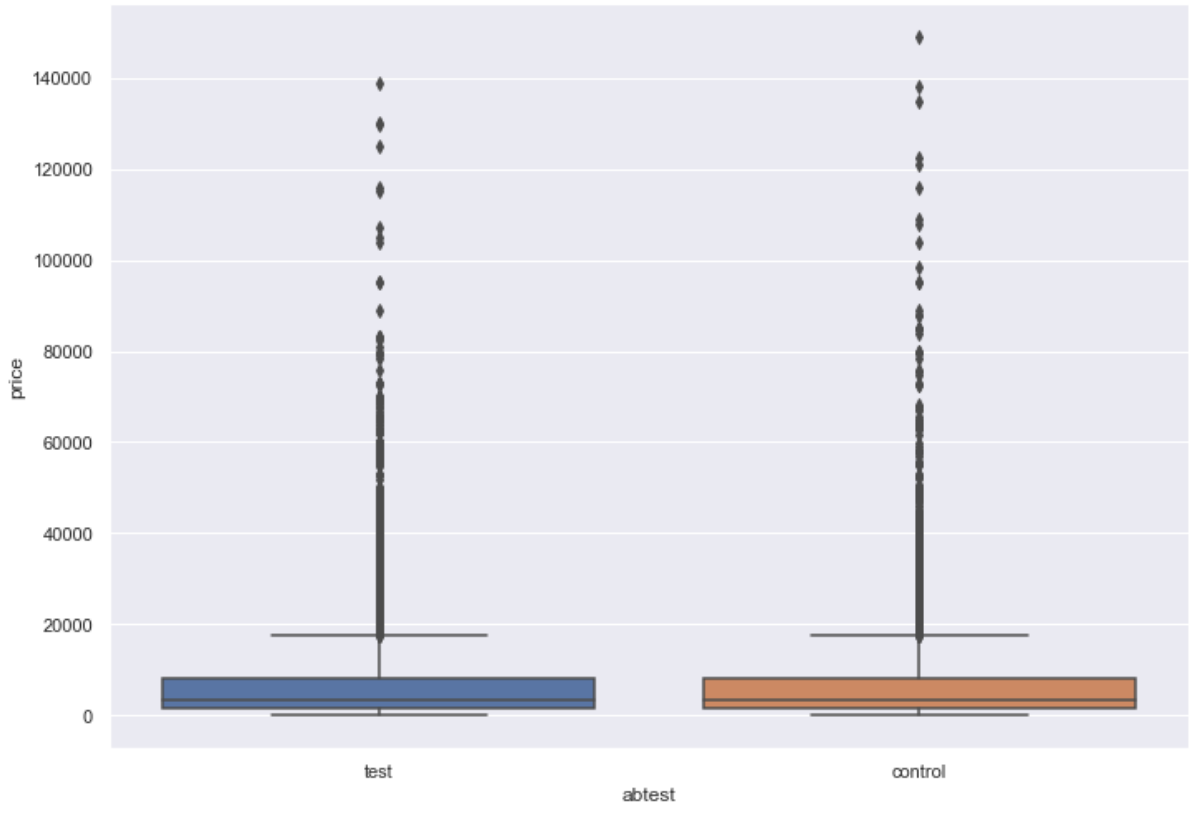


Variable: abtest

Histogram:

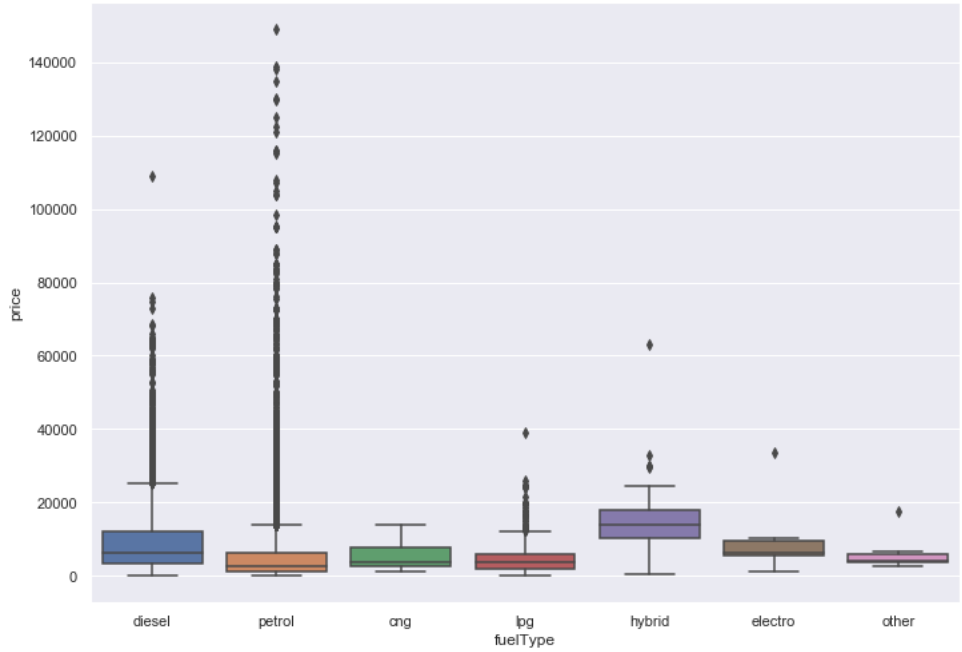


Boxplot:



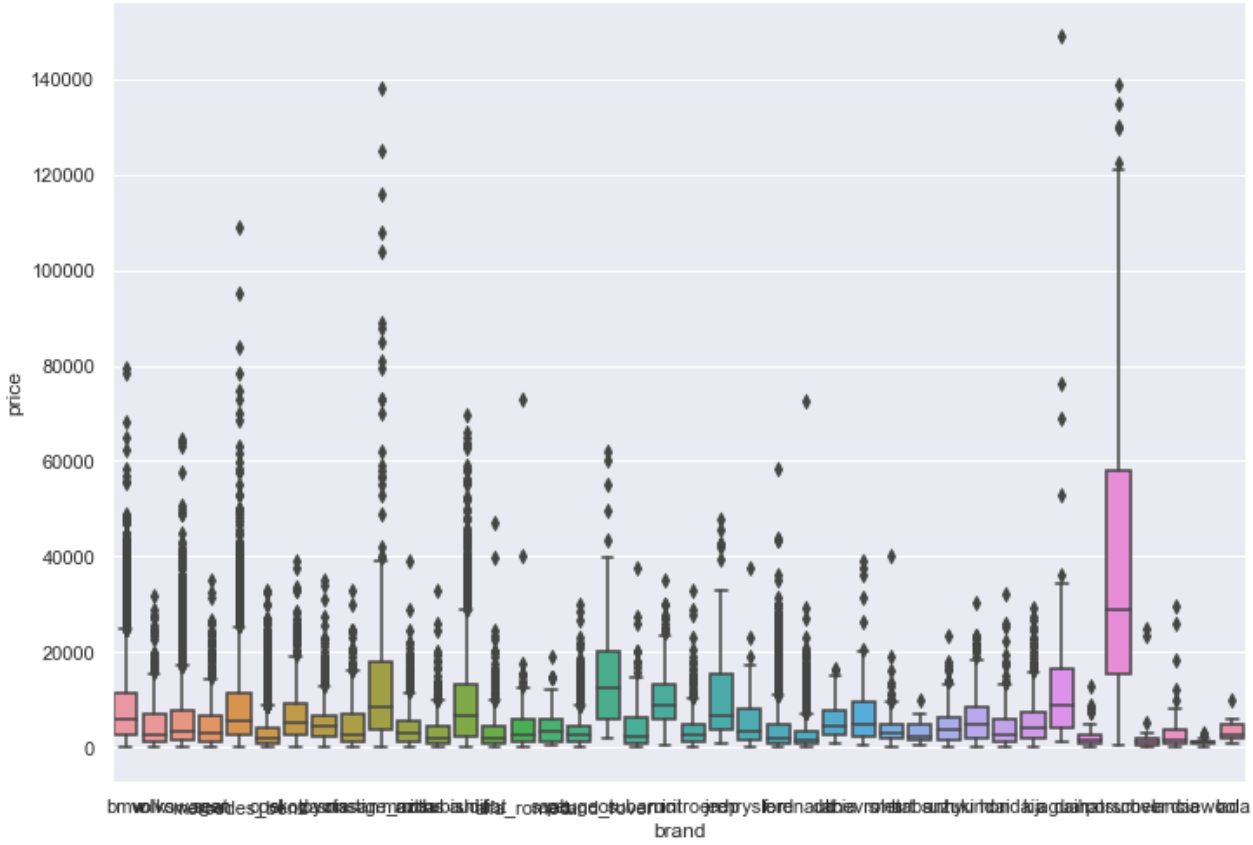
Variable: fuel type

Boxplot: fuel type vs price



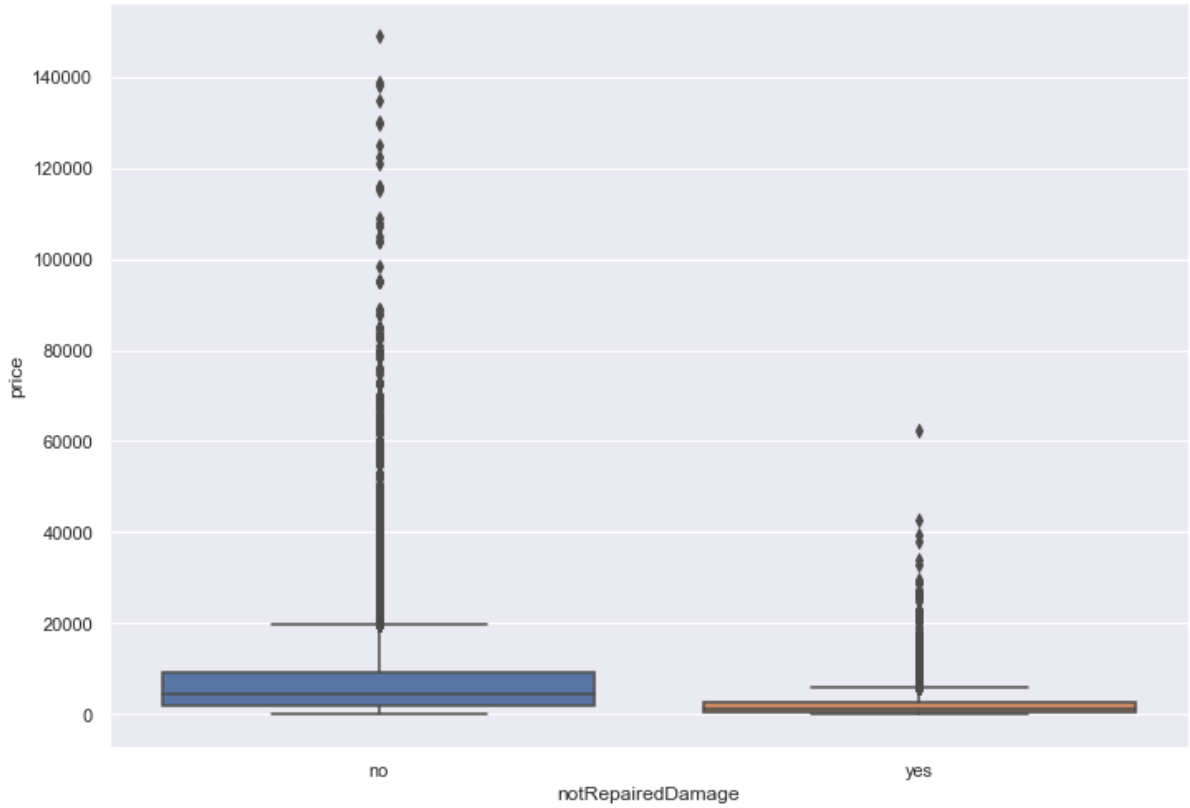
Variable: Brand

Boxplot: Brand vs price



Variable: notRepairedDamage

Boxplot: notRepairedDamage vs price



All the above features are varying with the price, so they are considered in the modelling.

Model selection: In the given data set by observing the boxplots and scatter plot from the above figures features showed the linear relation with the target set, we therefore choose linear regression model and Random forest regression for predicting the price of pre-owned cars for given features.

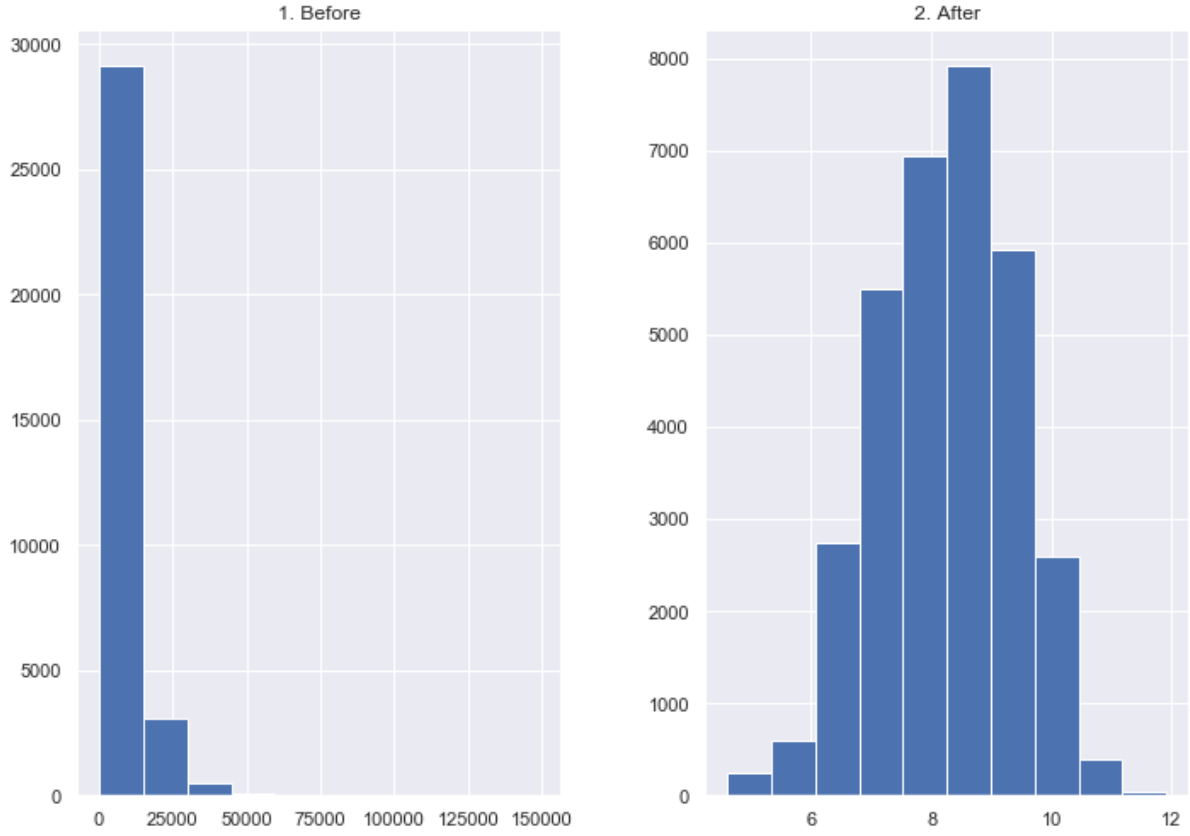
Model Calculations:

Correlation:

	price	power PS	kilometre	Age
price	1	0.575	-0.440	-0.336
power PS	0.575	1	-0.016	-0.151
kilometre	-0.440	-0.016	1	0.292
Age	-0.336	-0.151	0.292	1

```
#separating input and output features
x1 = cars_omit.drop(['price'],axis='columns',inplace=False)
y1 = cars_omit['price']
```

#plotting the variable price
prices = pd.DataFrame({"1. Before":y1, "2. After":np.log(y1)})



#transforming price as a logarithmic value
y1= np.log (y1)

 # LINEAR REGRESSION:

Root of mean square error(RMSE)	0.55	
R^2	test	train
	0.765	0.780

 # RANDOM FOREST REGRESSION:

Root of mean square error(RMSE)	0.436	
R^2	test	train
	0.85	0.92

Conclusion: I have presented a simple machine learning models which predict the price of pre-owned cars considering the impact of features on the price by linear regression and random forest regression.

The above table clearly tells us that random forest regression model is more suitable than linear regression model**.**