INTERIM REPORT

USED CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

submitted in partial fulfillment of the criteria for award
of
POST GRADUATE PROGRAM
in
DATA SCIENCE AND ENGINEERING

by

PRASSANTH E	(VX3WJ2477S)
PRATIK ASARKAR	(45BA1HZ8S1)
PRIYANKA PARLIKAR	(YG5GCUACL6)
RUDRA PRATAP SEN	(G53ZF1NIFJ)
ANN MARIA JOHN	(MYDEW11NAD)

Under the supervision of
Mr. Srikar Muppidi

GREAT LAKES INSTITUTE OF MANAGEMENT
Bangalore – 560 102, INDIA

JULY 2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

S. No.	***Abbreviation***	***Detail***
1.	*DAT*	*Deutsche Automobil Treuhand*
2.	*BS-VI*	*Bharat Stage IV*
3.	*SVM*	*Support Vector Machines*
4.	*k-NN*	*k-nearest neighbors*
5.	*DT*	*Decision Trees*
6.	*NB*	*Naive Bayes*
7.	*ANN*	*Artificial Neural Networks*
8.	*GTV*	*Gross Transaction Value*
9.	*ANOVA*	*Analysis of variance*
10.	*OEM*	*Original Equipment Manufacturer*
11.	*ACF*	*Auto-Correlation Function*
12.	*VIF*	*Variance Inflation Factor*
13.	*Q-Q Plot*	*Quantile-Quantile Plot*

1. INTRODUCTION

1.1.Motivation and Background

There could be one small beacon of light in the economic mess following in the coronavirus pandemic's wake: cheaper cars. According to JPMorgan, used car prices could plunge more than 15% as demand for vehicles plummets. That's even more than the prices that fell in the last recession a decade ago (Graham Rapier, 2020).

Valuation, in general, is the mechanism of finding the present market value of a future income flow. In automobile valuation, the value presented in a valuation report should represent the correct resale value of a vehicle.

When people seek the assistance of a financial institution in order to find the necessary financing for a vehicle, they focus on obtaining the real value of the vehicle. In order to get it, a valuation report is requested. Hence, a proper mechanism to obtain a value for a vehicle is paramount in such a scenario.

Due to the competition in the vehicle leasing industry, it is vital for a leasing company to give a reasonable as well as profitable price to remain a profitable company. A manual mechanism is used in Germany when calculating the second-hand vehicles and mentions that the system in place consists of major limitations in the attributes considered and also there is no proper mechanism to capture the attributes correctly (Mariana Listiani, 2009).

Valuation of vehicles is important to leasing companies in order to avoid losses as well as for people who have bought new vehicles and are willing to sell the vehicle at an appropriate price (Sameerchand, 2014).

With the number of vehicles increasing exponentially, the current process of manual valuation will become a difficult, time consuming and ineffective process. Therefore, there is a clear research gap where a technological solution can be provided to appraise vehicles faster and much more accurately.

This project attempts to provide a used car evaluation service based on the online listing information, which has great potential in business application. The objective of the project is to discuss the features of used-car with regard to its influence on the market price. Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, horsepower and mileage.

The fuel type used in the car as well as fuel consumption per mile highly affect price of a car due to a frequent change in the price of a fuel. Different features like exterior color, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price. In this paper, we applied different methods and techniques in order to achieve higher precision of the used car price prediction (Enis Gegic *et al.*, 2019).

Based on such analysis, a model for price prediction could be constructed in order to provide a real-time used car evaluation service. The final product takes features (brand, model, gearbox, vehicle type, kilometers driven, etc.) of a used car as input and output a prediction of price. Moreover, it should be capable of producing an interval prediction as price range, and it should be able to handle missing features so that the user can make a query even when they are not clear about certain feature.

1.2.Aim and Objective

This project aims to analyze how features of a used-car influence its market price and to predict the price based on the car features in the given data. The final product of the project are machine learning models that can predict market value estimation of a used car given its features.

The objectives of the study are:

i. Perform data cleaning and visualization, in order to reach an elementary understanding of each car feature and its influence on the market price.

ii. Build and evaluate models using machine learning algorithms for price prediction in order to provide a real-time used car evaluation service.

2. INDUSTRY REVIEW

2.1.Current Practices

The global auto industry expects to sell 59.5 million automobiles in 2020, a dramatic decrease of over 20 percent year-on-year. The sector is projected to experience a downward trend on the back of a slowing global economy and the advent of a pandemic in all key economies. Previously, it had been estimated that international car sales were on track to reach 80 million in 2019; however, as economies woes continue, demand for new motor vehicles is expected to be in the negative throughout 2020 (Statista Research Department, 2020).

The global spread of COVID-19, already has car shoppers abandoning showrooms. While states and municipalities only implemented social distancing measures fairly recently and are continuing to do so unevenly, the effect is being felt in showrooms (Stephen Edelstein, 2020). The first two months of the year started off at a healthy sale pace, but the market took a dramatic turn in mid-March as more cities and states began to implement stay-at-home policies due to the coronavirus crisis, and consumers understandably shifted their focus to other things. The whole world is turned upside down right now, and the auto industry is unfortunately not immune to the wide-ranging economic impacts of this unprecedented pandemic (Edmunds Forecasts, 2020).

The game-changer is the surge of 2-to-3-year old leased vehicles pulsing through used car auctions as more consumers are going to start looking at used car models. (Greg Gardner (Forbes), 2018).

As the smartphone and internet penetration in emerging economies, especially in Asia Pacific is increasing, the used car market is getting more organized because used car retailers are using digitalization to make market offerings attractive. Facilities such as enormous number of photos and videos on the online platform and easy online instant finance service is drawing more customers into buying used cars (Mordor Intelligence).

2.2.Background Research

2.2.1. Market Overview

The Global used car market is expected to grow at a CAGR of 12.81%, during the period of 2020-2025. The used car market across the world is growing rapidly. The main reason for the same has been the advent of organized players in the market in developing nations, which has taken care of the trust deficit, plaguing the used car market in those countries for ages (Mordor Intelligence).

Lack of financing or expensive financing options in many countries for used cars is expected to hinder the growth of the market. While finance for new cars is easy to obtain, used cars attract a higher rate of interest and are not sanctioned so easily. However, with the gradual growth in the organized sector, the market situation is changing. Major players, like Toyota, BMW, Maruti Suzuki, etc., have ventured into the used-car space in India. On global level, OEMs such as Volkswagen and Daimler have backed online vehicle sales startup such as Heycar in 2017 and 2018 (Mordor Intelligence).

2.2.2. Key Market Trends

German car brands are known and exported the world over. The United States were the leading importer of motor vehicles from Germany, followed by the United Kingdom and China. Leading importers in Europe include France, Italy and Spain. Car brands such as BMW, Mercedes and Volkswagen, to name just a few, are a key part of Germany's international image and economic standing, belonging to one of the longest-running and most successful industries in the country. In general, cars were the most traded type of goods in the world and Germany has been a driving force in this development. As of recently, Germany exported 18.75 million tons of motor vehicles. Based on ownership share, Volkswagen is the most popular among German vehicle owners and consistently wins the race, followed by Opel and Mercedes-Benz. In terms of car models, the highest number of new registrations was recorded for Volkswagen's Golf, Tiguan and Polo (Evgeniya Koptyug, 2019).

China currently has more than 300 million registered vehicles. This is expected to become a large used vehicle inventory for the world in the near future. The used car market in India is expected to grow at a rate close to 15% over the next five years. Online classified platform OLX expects the used car market to reach USD 25 billion mark by 2023. Millennials are considered as the factor driving the used car market in India. Millennials in India are tech savvy and are focusing on resale value of the vehicle rather than size and brand. As the Indian auto-industry is entering the BS-VI era from April 2020, the value proposition of the used car can grow stronger, as new cars are expected to become expensive due to additional technology costs (Mordor Intelligence).

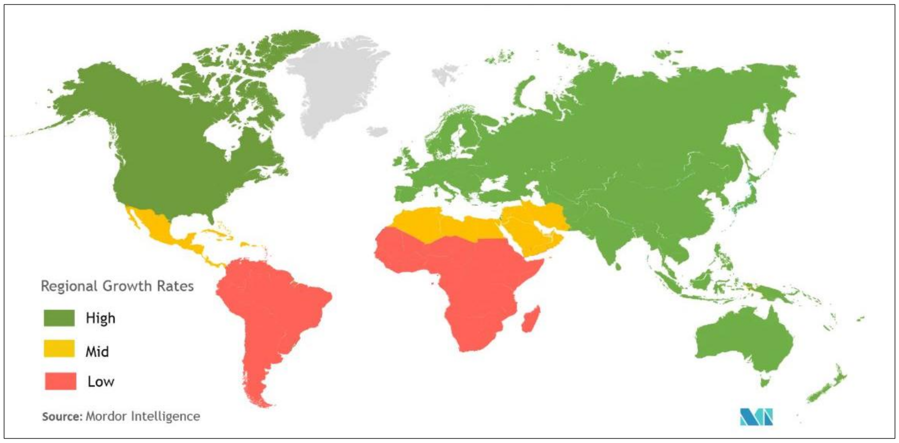


Fig. 1: Used Car Market – Growth Rate by Region (2020 – 2025)

2.2.3. Competitive Landscape

The market for global used cars is fragmented. Major global players have an edge over other smaller players due to their superior business models and increased number of pre-owned car retail outlets. But the market is dominated by small and unorganized regional players who are present in most of the nations. The growing organized and semi-organized sector is expected to bring in more revenue to the bigger organized players towards the latter half of the forecast period, thus moving the market ever slightly towards a consolidated one (Mordor Intelligence).

2.2.4. German Used Car Market Outlook

Germany Used Car Market is expected to reach about EUR 105 Billion in Gross Transaction Value (GTV) by 2023. The used car industry is anticipated to grow in the next few years owing to the fall in the new car sales due to the ban on the diesel and petrol cars. Additionally, the German population is not wholly ready to accept the electric vehicles, which will contribute in the growth of the used cars market. Further, the surge in the adoption of car financing and insurance options at the used car dealerships in Germany will lead to a significant growth in the used car market (Monika Singh, 2019).

The below statistic shows the revenue of the used car market in Germany between 2000 and 2019. In 2019, the revenue amounted to roughly 89.73 billion euros (Evgeniya Koptyug, 2020).

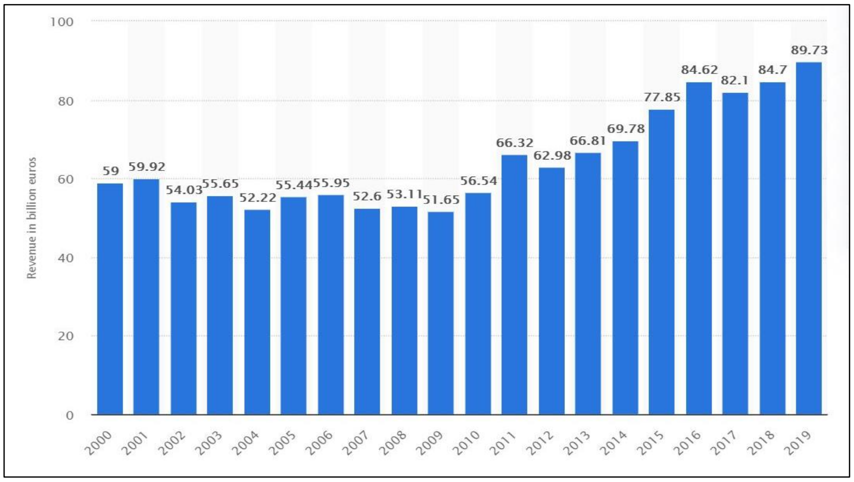


Fig. 2: Revenue of the used car market in Germany between 2000 and 2019

3. LITERATURE REVIEW

In this section, an analysis of the theoretical background and related work regarding the domain of used vehicle valuation have been conducted.

3.1.Theoretical Background

Valuation is a process of finding the present value of a future income stream of a return in a property based on past and current information. Valuation, in its simplest form, is the determination of amount for which the property will transact on a particular date.

Further as per the Royal Institute of Chartered Surveyors, valuation is a hypothetical transaction. The very act of asking for a valuation is equivalent to putting the property on the market. In addition, from definitions from R.T.C.S. Manual, it can be concluded that a valuation is a prediction of the most likely selling price. A calculation of worth is an estimate of what an investment is worth to a particular buyer or seller, and an appraisal is a combination of the above two. "The act or process of estimating value corresponds to the term "valuation appraisal" defined as "an unbiased analysis, opinion, or conclusion that estimates the value of an identified parcel of real estate or real property at a particular point in time" (Amjadh and Kaneeka, 2018).

As it has been indicated in the above definitions, valuation is a process of determining the current worth of an assets or a property. Unbiased valuation for the property is very important (Amjadh and Kaneeka, 2018).

The quantity and usage of vehicles have increased and the complexities of it have increased simultaneously. The necessity of getting the realistic values for these vehicles increased day by day for the various purposes such as insurance, selling purposes, purchasing, accounting purposes, etc. (Amjadh and Kaneeka, 2018).

3.2.Related Work

The automobile industry is globally expanding exponentially. As a result, the used vehicle market has also faced an unexpected growth. In order to cope with the future expansion of vehicle trade, research in computational approaches to used vehicle valuation have been scarce. Given below is an analysis of previous approaches to used vehicle valuation and value prediction.

Listiani, in her MSc thesis, has compared approaches of multiple linear regression and Support Vector Machines (SVM) in estimating the residual price of leased cars. In her thesis, she proves that a high accuracy can be gained by using SVM than simple multiple regression or multivariate regression. SVM is predicting value of used vehicles by utilizing machine learning systems which are more ready to manage high dimensional information (number of features used to anticipate the cost) and can keep away from both overfitting and underfitting. In order to find the optimal parameters for the SVM, a genetic algorithm has been used. This reduces the time taken to generate the SVM. The main downside of this approach is that the superiority of SVM rather than simple multiple linear regression was not depicted in basic measures like mean variance or mean deviation (Mariana Listiani, 2009).

Furthermore, Pudaruth applied various machine learning algorithms, namely: k-nearest neighbors (k-NN), multiple linear regression analysis, Decision Trees (DT) and Naive Bayes (NB) for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in period less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometers, production year, exterior color, transmission type and price. However, the author found out that NB and DT were unable to predict and classify numeric values. Additionally, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%. Pudaruth in his work used four different machine learning techniques to forecast the price of used cars in Mauritius. The mean error with linear regression was about Rs. 51,000 while for k-NN it was about Rs. 27,000 for Nissan cars and about Rs. 45,000 for Toyota cars. J48 and NaiveBayes accuracy dangled between 60-70% for different combinations of parameters. The main weakness of DT and NB is their inability to handle output classes with numeric values. Hence, the price attribute was classified into classes which contained a range of prices but this evidently introduced further grounds for inaccuracies. The main limitation of this study was the low number of records that were used (Sameerchand, 2014).

In Germany, there are well-known lists, namely Schwacke List and Deutsche Automobil Treuhand GmbH (DAT) List, that help public domain to approximate the value of a second hand car, based on its attributes like type, year of manufacture, common equipment and kilometers driven. Nevertheless, one has to note that both lists have a limitation on the attributes scope. Only basic car equipments are included for price calculation, while other features, that may also be influential, are failed to be captured (Mariana Listiani, 2009).

In his work, Enis Gegic *et al.*, has sighted the work of Gongqi, who proposed a model that is built using Artificial Neural Networks (ANN) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision than other linear models (Enis Gegic *et al.*, 2019).

Pudaruth in his paper has reported the work of Richardson (another university thesis), wherein he was working on the hypothesis that car manufacturers are more willing to produce vehicles which do not depreciate rapidly. In particular, by using a multiple regression analysis, he showed that hybrid cars (cars which use two different power sources to propel the car, i.e. they have both an internal combustion engine and an electric motor) are able to keep their value better than traditional vehicles. This is likely due to more environmental concerns about the climate and because of its higher fuel efficiency. The importance of other factors like age, mileage, make and MPG (miles per gallon) were also considered in this study. He collected all his data from various websites (Sameerchand, 2014).

Amjadh and Kaneeka reported the work of Du *et al*. In USA, a large amount of vehicles are sold through leasing. Because most vehicles are being returned at the end of the leasing period, institutions need to appraise the vehicles accurately in order to resell the vehicles which are returned. In order to address this situation, the ODAV (Optimal Distribution of Auction Vehicles) system was developed by Du *et al*. Using a k-NN regression model, the system has the ability of estimating the price of vehicles entered. This solution was created in 2003. Since the United States is a huge country, the state in which the vehicle is present also plays a paramount role in the vehicle price (Amjadh and Kaneeka, 2018).

Enis Gegic *et al.* applied single machine algorithm on the data set and got an accuracy less than 50%. Therefore, an ensemble of multiple machine learning algorithms was proposed and used which gave an accuracy of 92.38%. This was a significant improvement compared to single machine learning method approach. However, the drawback of the proposed system was that it consumed much more computational resources than single machine learning algorithm (Enis Gegic *et al.*, 2019).

Harikrushna’s work examined different Machine Learning techniques for predicting the price of the used car. It was concluded that k-NN or K-star algorithm was the best fit for predicting price of a used car. But, this analysis was performed for only one brand of car which could be extended to multiple brands (Harikrushna, 2018).

According to Shiva Shankar, the pre-owned car market may be transformed into a more organized market with the advent of Indian and global car makers and other major corporate houses of India. Auto majors not only increase their market penetration through the pre-owned car business, but also make a profit out of this venture. Unorganized pre-owned car dealers are trying to match the service standards of organized used car dealers to become more professional in their marketing approach. The general consensus among the industry is that the pre-owned car segment may become almost double of the new car market in another five years as is the case in the developed countries (Shiva Shankar, 2016).

4. PROJECT OVERVIEW

4.1.Problem Statement

The challenging part for used car dealers is to predict the price of used cars with accuracy. The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling.

4.2.Need of Study

The increase in e-commerce usage over the past few years has created potential in the used car market, enabling the used car dealers to reach out to a huge customer base. To be able to predict used cars market value can help both buyers and dealers (sellers in the organized used car market).

4.2.1. Target Business: Online User-Car Dealers

They are one of the biggest target group that can be interested in results of this study. If used car dealers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and market products with better price.

4.2.2. Target Customers

Buyers who would like to purchase used cars via online portal, wherein, it's a big corner to pay too much or sell less then it's market value.

4.2.3. Revenue Generation

E-commerce is being preferred over other businesses as the proportion of revenue investment, which is apparently better than other options and requires less investment. Here, is a list of some revenue generation methods via the online used car dealership:

1. Google Ad-Revenue (based on high user traffic)
2. Target Marketing based on user profile (user details): Target Ad based on the user's profile

3. Subscriptions: For many websites, it's the primary source of revenue generation. Such sites earn revenue by charging a subscription/membership fee from buyer or seller to access/avail/buy/sell the services or products available on the website.

Hence, we can see that estimating the price of used cars is of very high commercial importance.

4.3.Scope of Study

The organized sector is dominating the German used car market due to higher network chains with a record of satisfactory relationship with their customers. This gives the organized dealerships an edge over the unorganized dealers in terms of enhanced quality of documentation process, certified inspection and many other factors. Additionally, the multi-brand dealerships recorded a larger market share in the organized used car market as multi-brand used car dealers have various brands and models available with them and the consumers have the choice of comparing and then purchasing the used cars. Furthermore, the organized used car dealers have a larger geographical presence in the country.

To predict the price of used cars, the business would require pointers about various features pertaining to the used car. The study will aid in price prediction based on the trained model on the used car dataset and give insights on how the price varies depending on the features.

4.4.Complexity Involved

The dataset has some complexity which needs to be resolved in order to get better results of the predicted price. The complexity in the dataset includes:

- The dataset has many outliers and missing values that need to be treated.
- High number of categories for features like model was difficult to handle as data would be spread over a large area, and so encoding techniques were done.

4.5.Data Sources

In order to predict used car prices, the data was retrieved from Kaggle (https://www.kaggle.com/orgesleka/used-cars-database#autos.csv). It contains 20 features with 371528 raw observations, scraped from used car listing on Ebay-Kleinanzeigen (German). Each record belongs to a unique used car.

4.6.Dataset Description

The dataset consists of both numerical and categorical variables.

The below table consists of numerical features and their description.

Feature Name	Feature Description	Min.	Max.	Std. Deviation
Price	The price on the ad to sell the car	0	2147484000	3587954
Year of Registration	Year in which the car was first registered	1000	9999	92.866598
Power (PS)	Power of the car in PS	0	20000	192.139578
Kilometer	How many kilometers the car has driven	5000	150000	40112.337051
Month of Registration	At which month the car was first registered	0	12	3.712412
No. of pictures	Number of pictures in the ad	0	0	0
Postal code	Postal code of the city where the car is available.	1067	99998	25799.08247

Table 1: Numerical features used in the price prediction model

Table 1 shows the seven numerical features along with their statistical parameters. Among these, Price is monetary values in Euros. PowerPS represents the power of the car in PS (Pferdestärke, German) equivalent of horsepower or Torque (1 pferdestarke = 0.9863200706195 horsepower).

The below table consists of categorical features and their description.

Feature Name	Feature Description	Number of Categorical Values
Name	Name of the car	233531
Seller	Private or dealer	2
Offer type	Offer or Application	2
A/B Test	Test or Control	2
Vehicle type	Describes the type of Vehicle	8
Gearbox	Whether the car has manual or automatic gearbox	2

Model	The model of a car is the name used by a manufacturer to market a range of similar cars	251
Fuel type	The Fuel system on which the car runs	7
Brand	The Brand which the car belongs to	40
Not repaired damage	If the car has a damage which is not repaired yet	2

Table 2: Categorical features used in the price prediction model

Table 2 shows the thirteen categorical features along with the number of categorical values for each of them. Further we can divide above features into ordinal and nominal categorical features. The features which have inherent order in it are called ordinal features while features which do not have inherent order are called nominal features. In the context of this dataset, all the features are nominal categorical features.

The below table consists of date-time features and their description.

Feature Name	Feature Description	Unique Values	First Value	Last Value
Date crawled	When this ad was first crawled, all field-values are taken from this date	280500	2016-03-05	2016-04-07
Date created	The date for which the ad at EBay was created	114	2014-03-10	2016-04-07
Last seen online	When the crawler saw this ad last online	182806	2016-03-05	2016-04-07

Table 3: Date-time type features used in the price prediction model

Table 3 shows the three date-time features along with their statistical parameters. These features represent the time duration in the form of date.

4.7.Data Preparation

In the following section, we are going to experiment with the dataset by data cleaning, feature selection, feature extraction, and feature engineering.

4.7.1. Data Cleaning

4.7.1.1. Outlier Treatment

Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. Let’s get a sense of how we handled outliers in our dataset.

i. Price

Looking into the real used car market in Germany, the prices of a vehicle could range from anywhere between 100 euros to 80,000 euros. But we did have our records which had price range starting from 0 euros to as high as 2.147484e+09 euros. These values seemed practically impossible for any used car in the market and seemed to be some sort of data collection error. As these values above practical range (80,000 euros) were very low in numbers, we capped our price range between 100 euros to 1,00,000 euros thus getting rid of major chunk of outliers. We used isolation forest anomaly detection technique and could still see some outliers in our price range which we handled using log transformation technique to normalize the range.

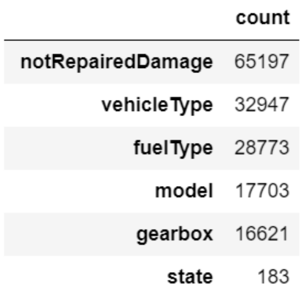
ii. PowerPS

The power of a German car could range from anywhere between 40 to 800 PS. But we did have our records which had power range starting from 0 PS to as high as 20,000 PS. These values seemed practically impossible for any used car in the market and seemed to be some sort of data collection error. As the values above practical range (800 PS) were very low in numbers, we capped our power range between 40 PS to 800 PS thus getting rid of major chunk of outliers. We used isolation forest anomaly detection technique for outlier detection.

4.7.1.2. Missing Value Imputation

A common problem when dealing with real-world data is missing values. These can arise for many reasons and have to be either filled in or removed before we train a machine learning model. First, let’s get a sense of how many missing values are in each column.

Following are the features and their corresponding count of missing values in the dataset after price range was capped:



	count
notRepairedDamage	65197
vehicleType	32947
fuelType	28773
model	17703
gearbox	16621
state	183

Fig. 3: Count of missing values

From the above features, we decided to drop the missing values for the features ‘notRepairedDamage’, ‘fuelType’, ‘model’, ‘gearbox’ and ‘state’. The reason for NOT imputing these values was the fact that all these variables were quite independent of any other features in the dataset, so imputing these values depending on other features would actually result in increase in multicollinearity between independent features. Secondly, as we have a huge dataset at our disposal, we could afford losing some data as it would not cost us much, as far as information loss is concerned.

For the feature ‘vehicleType’, we did KNN imputation using model, brand and price as these feature come very close to predicting the type of vehicle.

4.7.2. Feature Selection

A major chunk of our feature selection was done based on the statistical analysis of the features. All of the features passed the significance test except for abtest for which we got a p-value greater than alpha (0.05).

Apart from that, features like seller and offerType were highly imbalanced and had no significant influence on price prediction due to the extreme imbalance so we kept it out from the final dataset.

Features dateCrawled, noOfpictures practically served no purpose in predicting the price of a used car. So these features were also kept out from the final dataset.

4.7.3. Feature Extraction

i. ageOfVehicle

Considering two categorical columns (yearOfRegistration and monthOfRegistration) seemed not a feasible solution to find out how old the car is. So we extracted a new feature from these two columns called "AgeOfVehicle" which severed as an additional continuous variable which described how old the car is in years. Additionally, we got rid of two variables from the dataset namely, yearOfRegistration and monthOfRegistration.

ii. No_of_days_online

This is one variable which could be quite tricky to understand. Our idea behind introducing this feature was that, as the number of days of a post/advertisement of a certain used car being online increases, there could be chance for an interested buyer to negotiate the price of the car he/she is interested in. So, to let the buyer know exactly the valuation of a certain car after being online for certain number of days, would be a helpful added entity. So using the columns dateCreated and lastSeen, we came up with the new feature No_of_days_online, thus getting rid of the former two features.

iii. State

This feature was extracted from the postalCode feature. It gave an additional support to price prediction depending on in which state the vehicle is available.

iv. CountryOfManufacture

This feature was extracted from the feature brand. Similar to the feature State, CountryOfManufacture also provided an additional support to the price prediction wherein it was estimated if the price of the used car varies depending on the country of manufacture

4.7.4. Feature Engineering

There were many categorical features in our dataset, so finding the best encoding technique for these variables was certainly a challenging task. We decided to divide the categorical features into high and low cardinality features. For low cardinality features, we used One Hot Encoding technique as it did not add too many columns to our dataset. On the other hand, we had high cardinality features, which we could not deal with just using One Hot Encoding as it could add high dimensionality to our data. So we went through the rigorous process for combining OHE for low cardinality features with some other different techniques for handling the high cardinality features.

Depending on the best RMSE scores, we decided to stick with a combination of OHE and k-fold target encoding technique for our final dataset.

5. EXPLORATORY DATA ANALYSIS

In the following section, we are going to experiment with visualization, in order to reach an elementary understanding of each car feature and its influence on the used car market price. The purpose of exploratory data analysis is two-fold:

- To understand the data in terms of price across various independent variables/features.
- Get insights on various features.

5.1.Univariate Analysis

i. Price

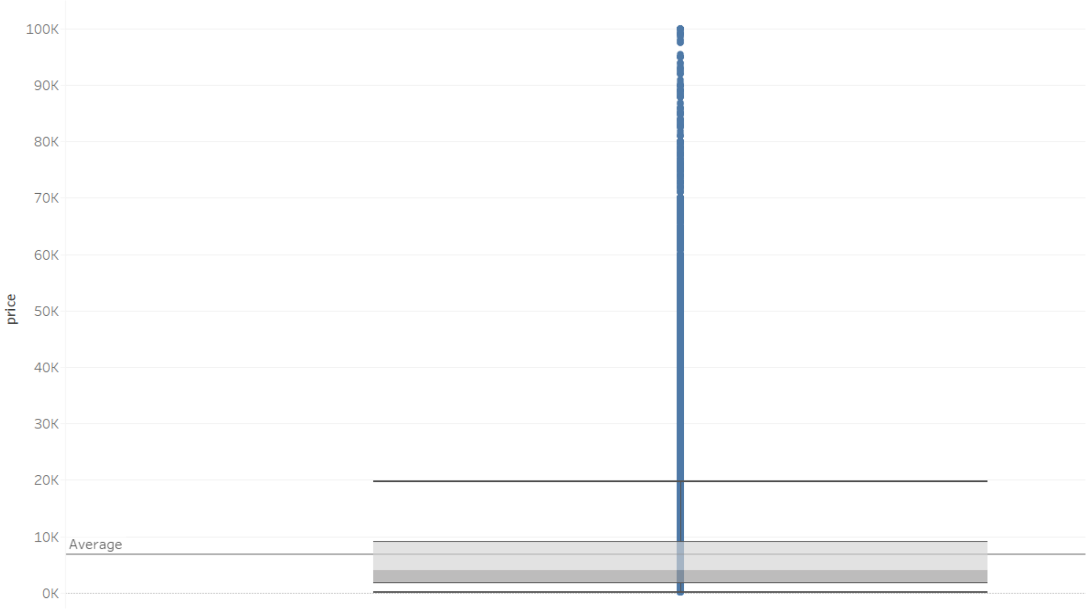


Fig. 4: Box-Whisker Plot: Price

ii. powerPS

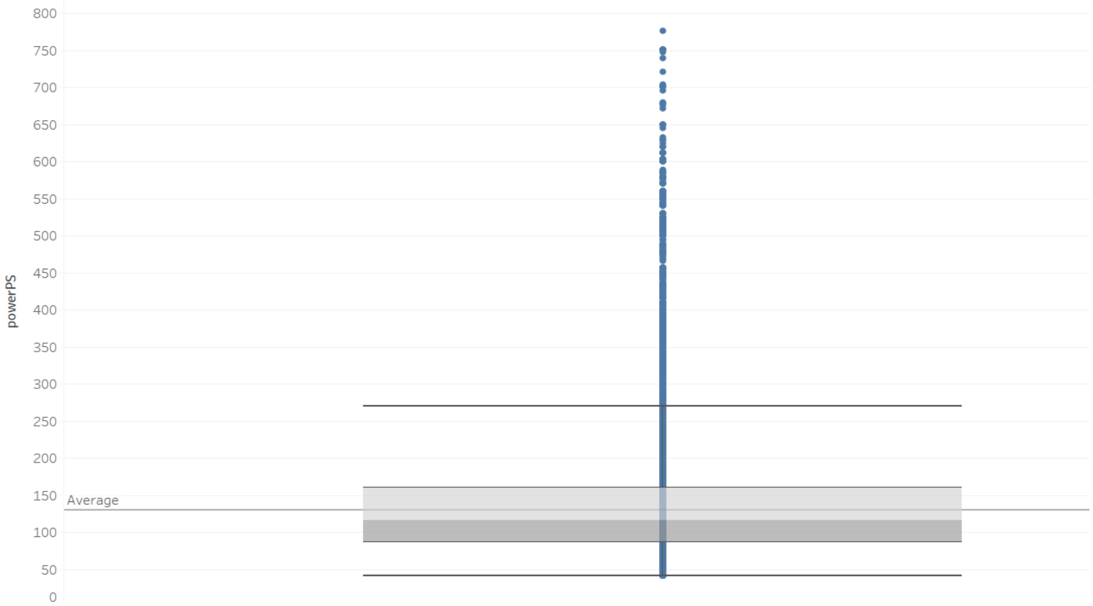


Fig. 5: Box-Whisker Plot: powerPS

Statistical Information	Price	powerPS
25% (lower quartile)	100	41
50% (Median – middle quartile)	4000	116
75% (upper quartile)	19,800	270
Mean	6847	129

Table 4: Statistical Information for Price and powerPS

iii. ageOfVehicle

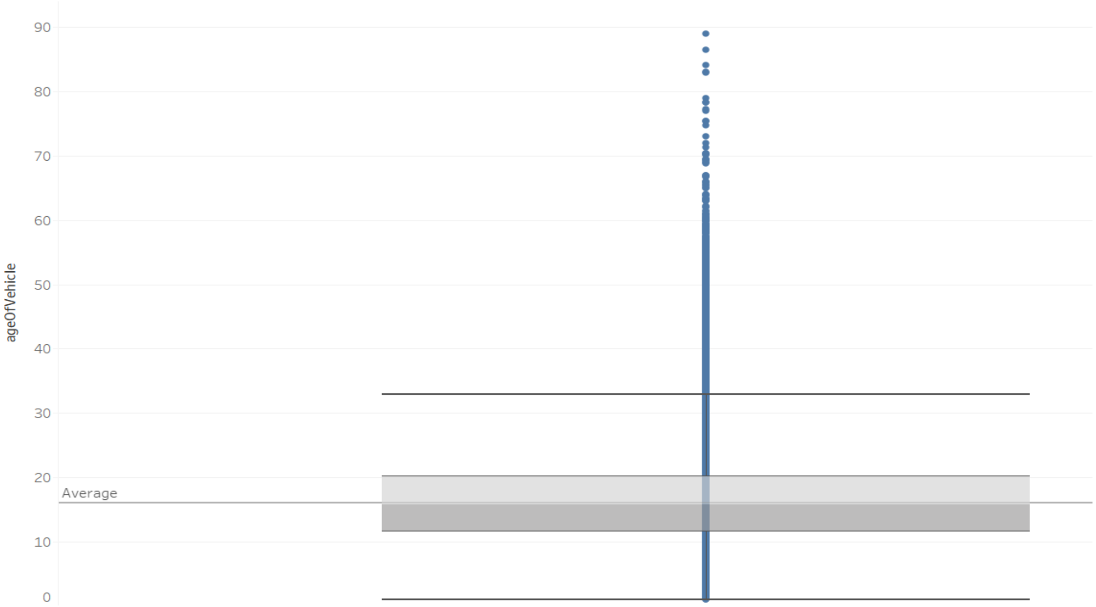


Fig. 6: Box-Whisker Plot: ageOfVehicle

iv. No_of_days_online

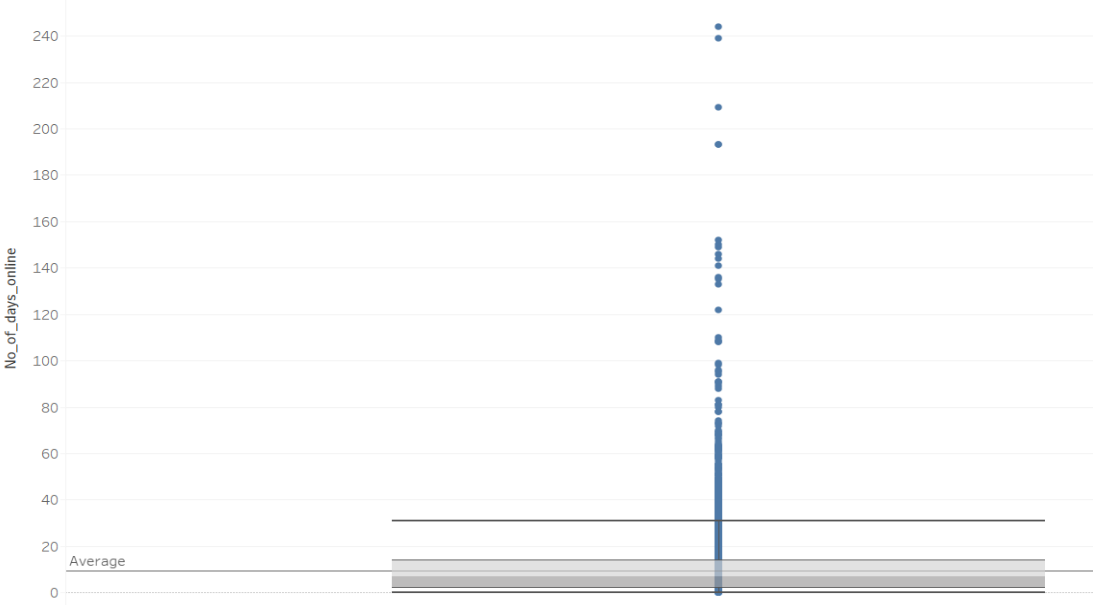


Fig. 7:Box-Whisker Plot: No_of_days_online

Statistical Information	ageOfVehicle	No_of_days_online
25% (lower quartile)	1.08	0
50% (Median – middle quartile)	15.75	7
75% (upper quartile)	33	31
Mean	16.10	9.24

Table 5: Statistical Information for ageOfVehicle and No_of_days_online

Inference: Transformation technique was applied to reduce the effect of outliers only for the target variable price.

v. Count of brand

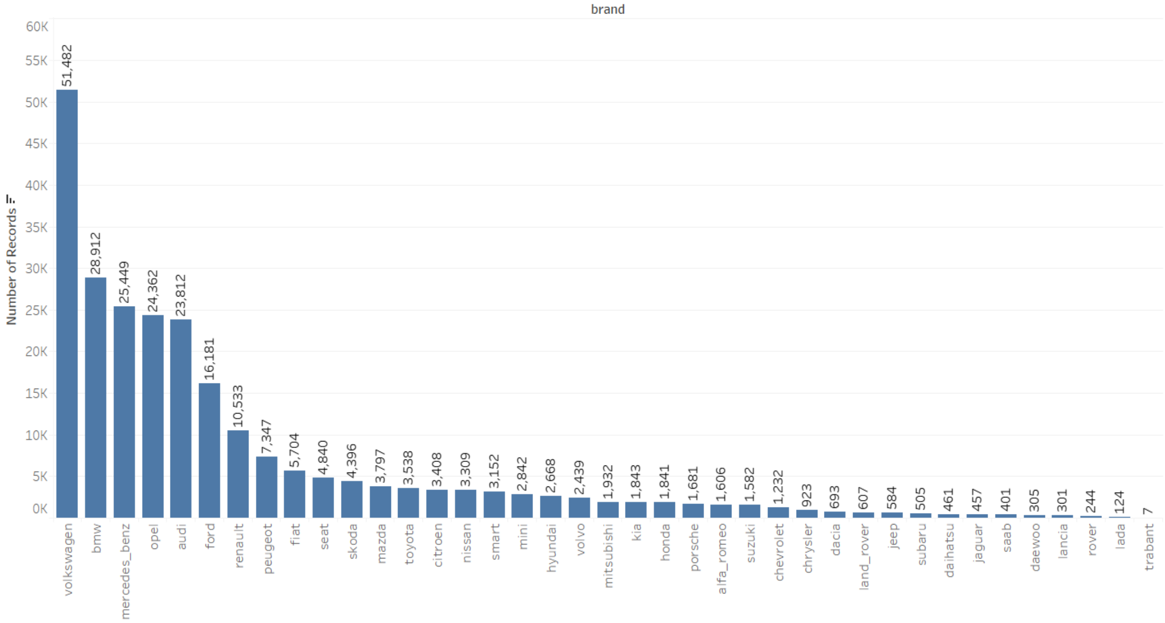


Fig. 8: Count of brand

Inference: Volkswagen, BMW, Mercedes Benz, Opel and Audi are the top most brands available in used car segment.

vi. Count of vehicleType

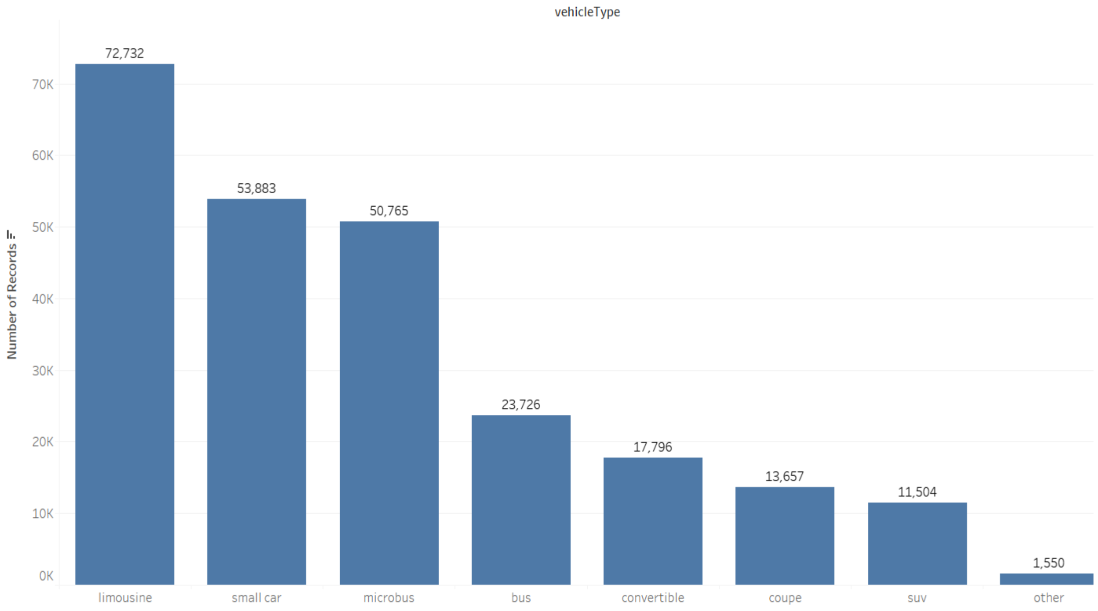


Fig. 9: Count of vehicleType

Inference: Limousine in Germany means Sedan vehicle type, from the count plot we can see that Limousine, small car and microbus are the most used vehicle type. Brands like BMW and Mercedes Benz sedan cars are used as taxi in Germany. Hence, Limousine are most in demand.

vii. Count of notRepairedDamage

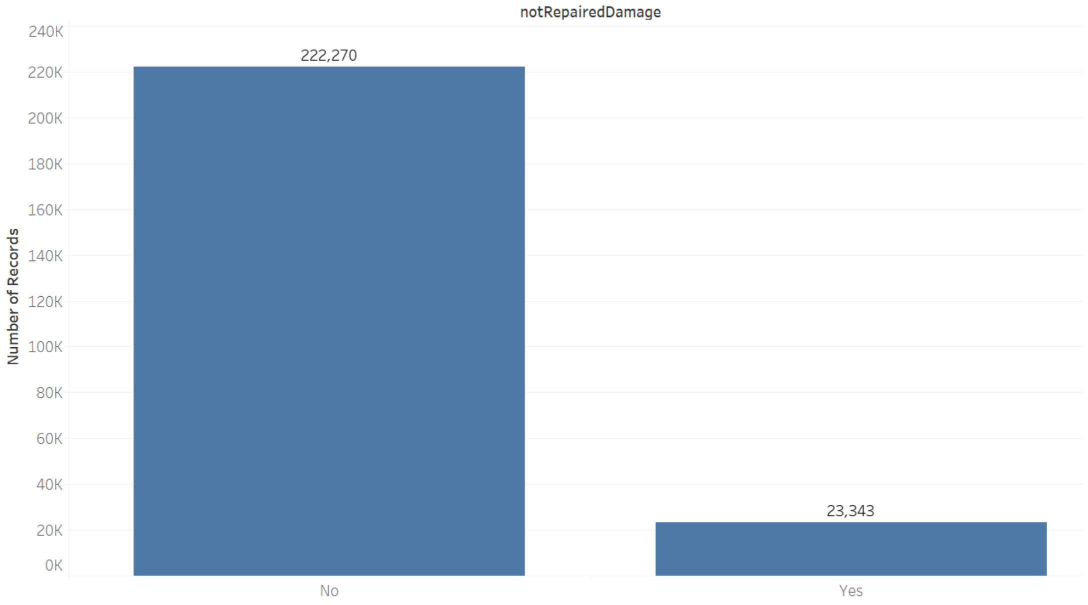


Fig. 10: Count of notRepairedDamage

viii. Count of fuelType

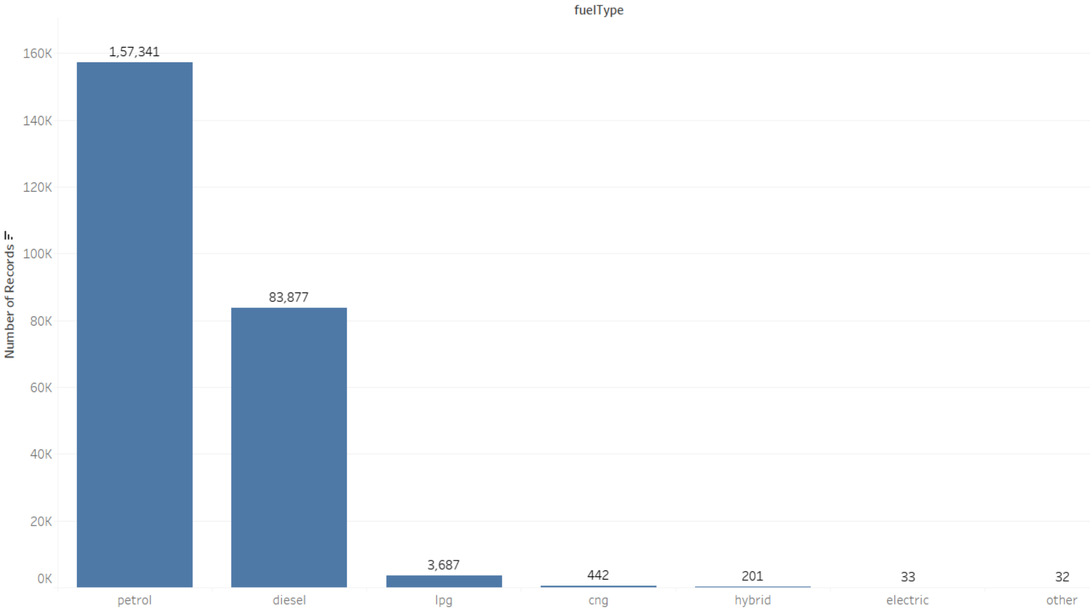


Fig. 11: Count of fuelType

Inference: Petrol and Diesel cars are most in demand and as the maintenance cost for petrol cars is less and more reliable compared to diesel cars, customers tend to buy petrol cars often.

5.2.Bivariate Analysis

i. Price v/s fuelType

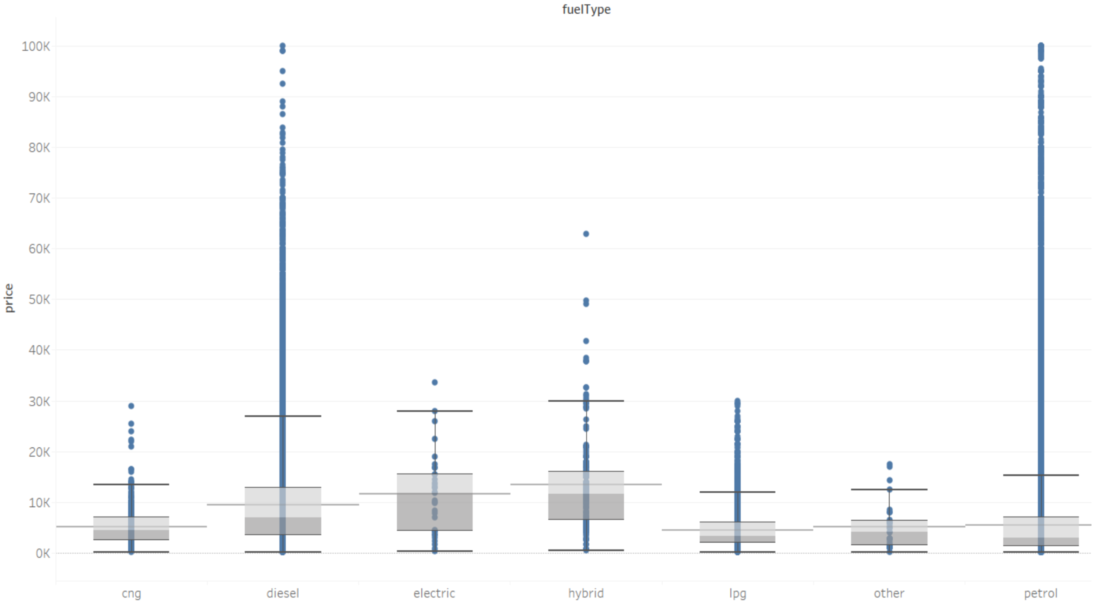


Fig. 12: Price v/s fuelType

Inference: Cars with different fuel type have different prices, from the plot we can see that cars of electric and hybrid fuel type are expensive as compared to other fuel types, Diesel and petrol fuel type cars belong to moderate range of price. CNG and LPG cars are the least expensive. The horizontal grey line shows the mean price of each fuel type cars, also p-value of fuel type is 0 when One-way ANOVA test was performed, hence fuel type of car becomes significant to predict price.

ii. Price Vs gearbox

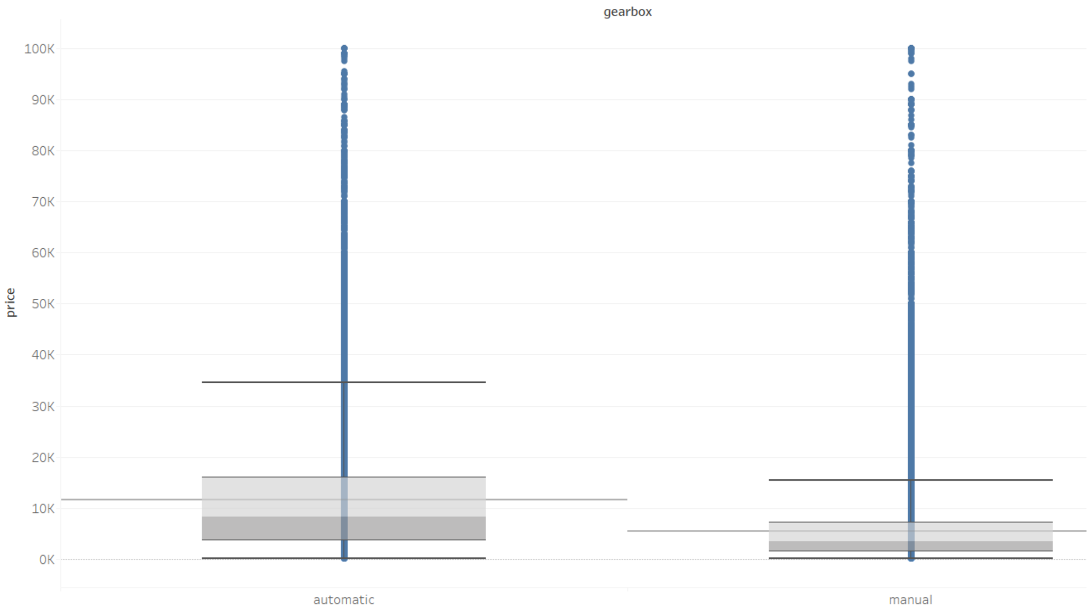


Fig. 13: Price Vs gearbox

Inference: Cars with automatic gearbox are of higher prices as compared to cars with manual gearbox. Mean price of cars with automatic and manual gearbox are different, p-value of gearbox is 0 which indicates that it is a significant feature in price prediction.

iii. Price v/s vehicleType

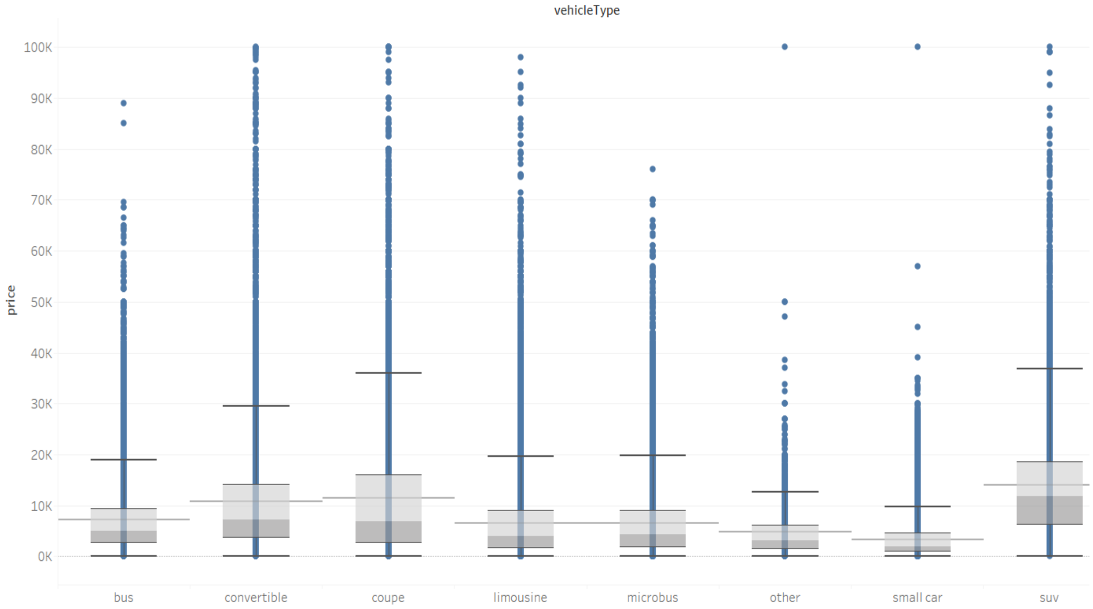


Fig. 14: Price v/s vehicleType

Inference: SUV, coupe and convertible vehicle types comes under higher price ranges. The mean price difference between convertible and coupe; limousine and microbus is less hence for price prediction one may consider only one vehicle type among them, though here we have considered each vehicle type distinctly.

iv. Price v/s notRepairedDamage

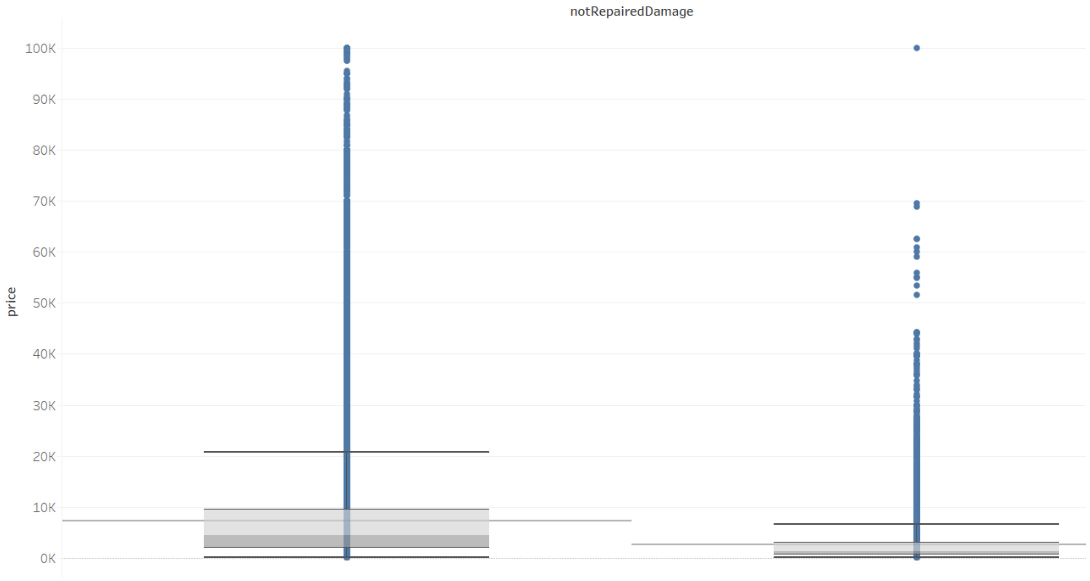


Fig. 15: Price v/s notRepairedDamage

Inference: Cars that do not have unrepaired damages have higher prices compared to the ones that have not repaired damages. Customers generally prefer cars in good condition over the ones that would bring about additional maintenance cost.

v. Average price of brands

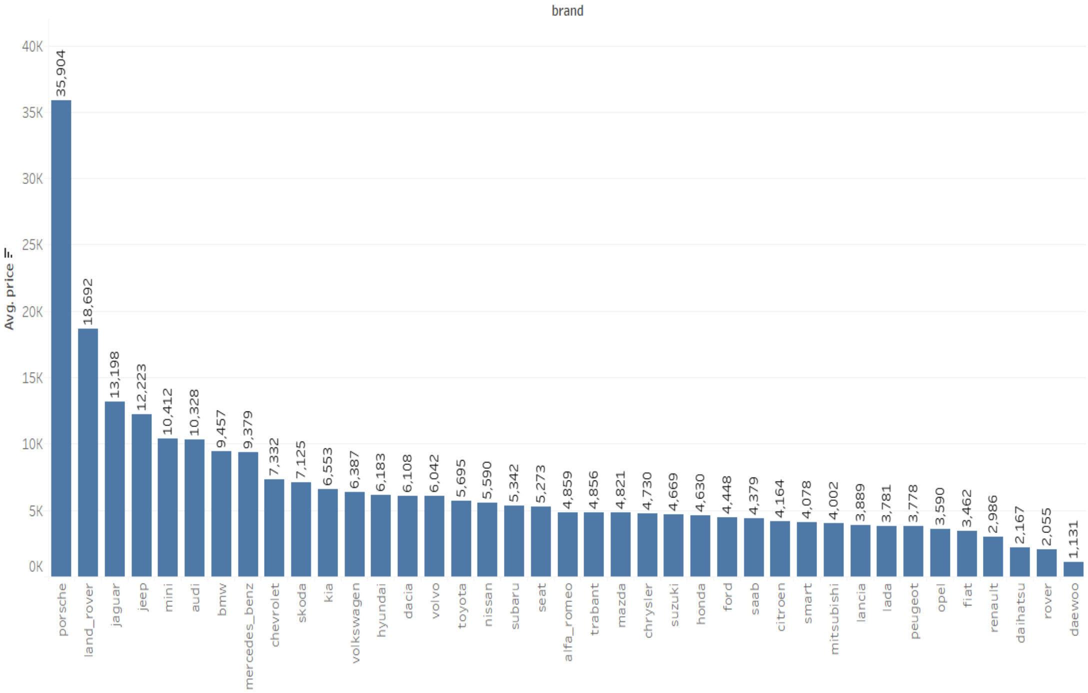


Fig. 16: Average price of brands

Inference: Porsche, land rover and jaguar top the brand list being the most expensive used cars.

vi. Average price v/s kilometer

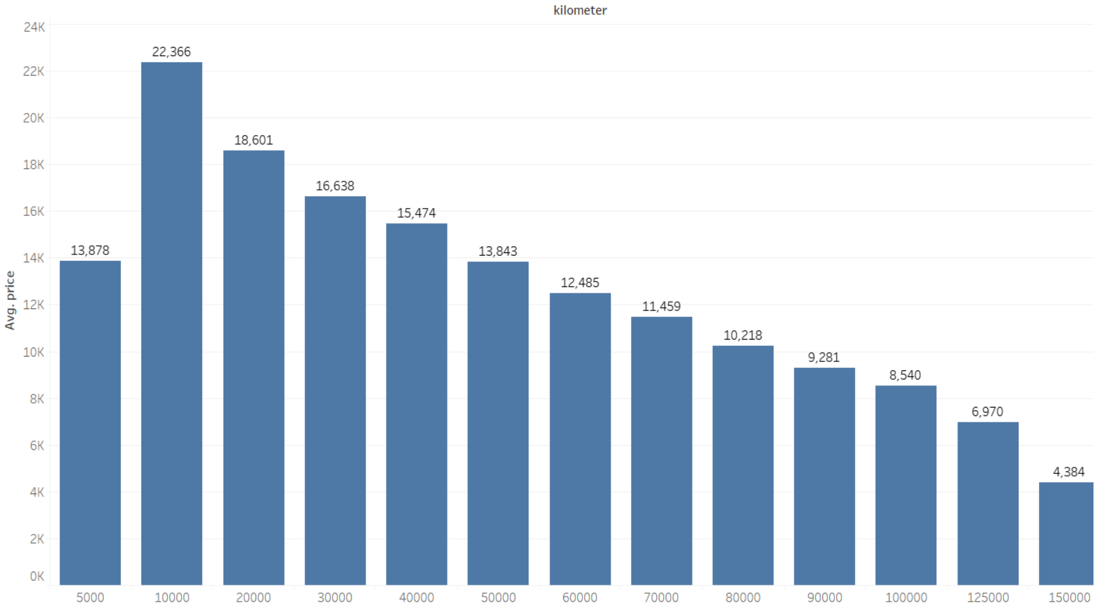


Fig. 17: Average price v/s kilometer

Inference: Here, we have considered kilometer as a categorical feature against average price. As the kilometers driven increases, average price decreases, this could probably mean that customers are keen on buying used cars that are less driven and are relatively in good condition. Also, we can see the average price for 5000 km is less as compared to 10000 km, there could be three possible reasons:

1. The data does not comprise much information about used cars that are driven 5000 kilometers.
2. The cars might have some not be repaired damages; that would result in lower prices.
3. Fuel type and vehicle type could rank higher when it comes to used car price determination.

vii. Price v/s ageOfVehicle

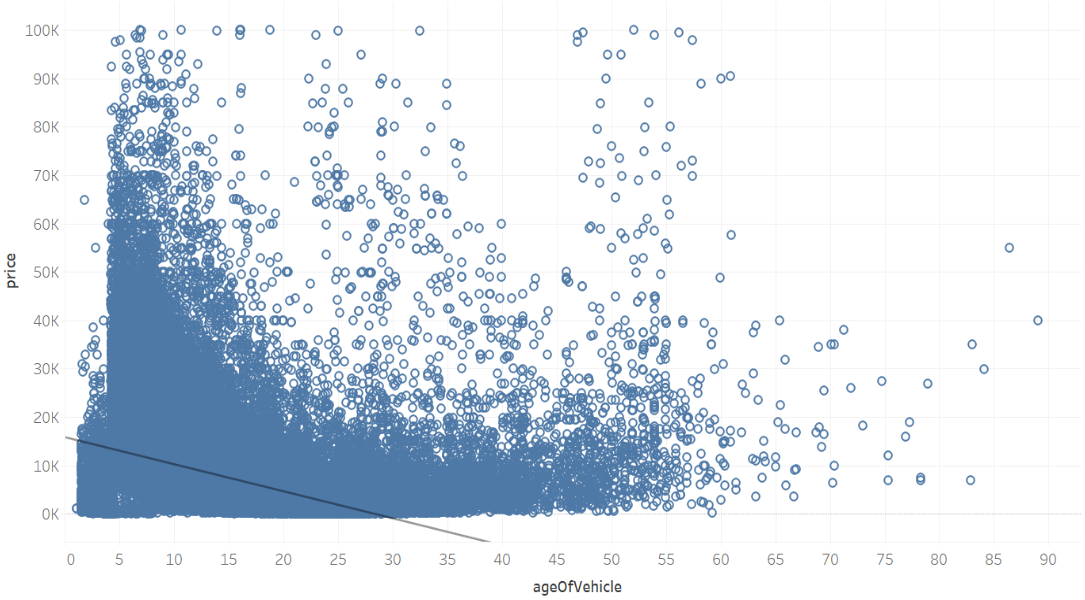


Fig. 18: Price v/s ageOfVehicle

Inference: The trend line in the above scatter plot indicates a strong negative correlation between price and age of vehicle. Exceptions include Porsche, Land Rover, Mercedes Benz and BMW that belong to luxury cars, wherein even though the age of vehicle is greater than 20 years, price would be higher shows the reliability of the German cars.

viii. Price v/s PowerPS

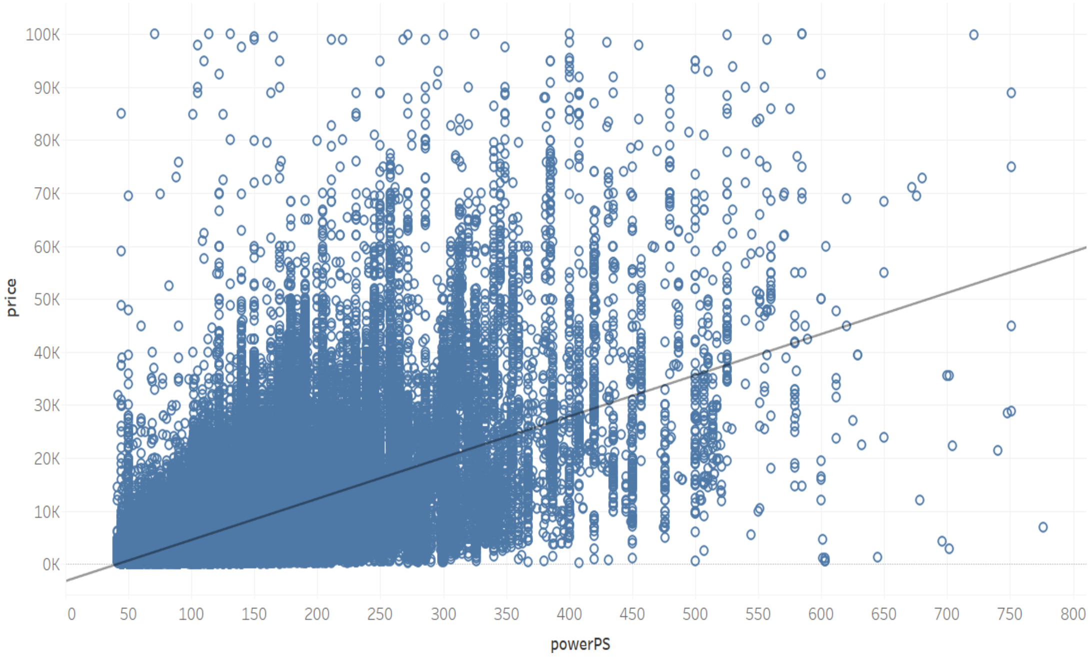


Fig. 19: Price v/s PowerPS

Inference: The above trend line shows a positive correlation between price and powerPS. Basic vehicle type like small cars and bus have less powerPS as compared to sport cars that has powerPS more than 300. Such high powerPS cars belongs to brands like Mercedes Benz, Porsche, BMW, Audi etc., that are well known for manufacturing sports cars.

ix. Average powerPS v/s brand

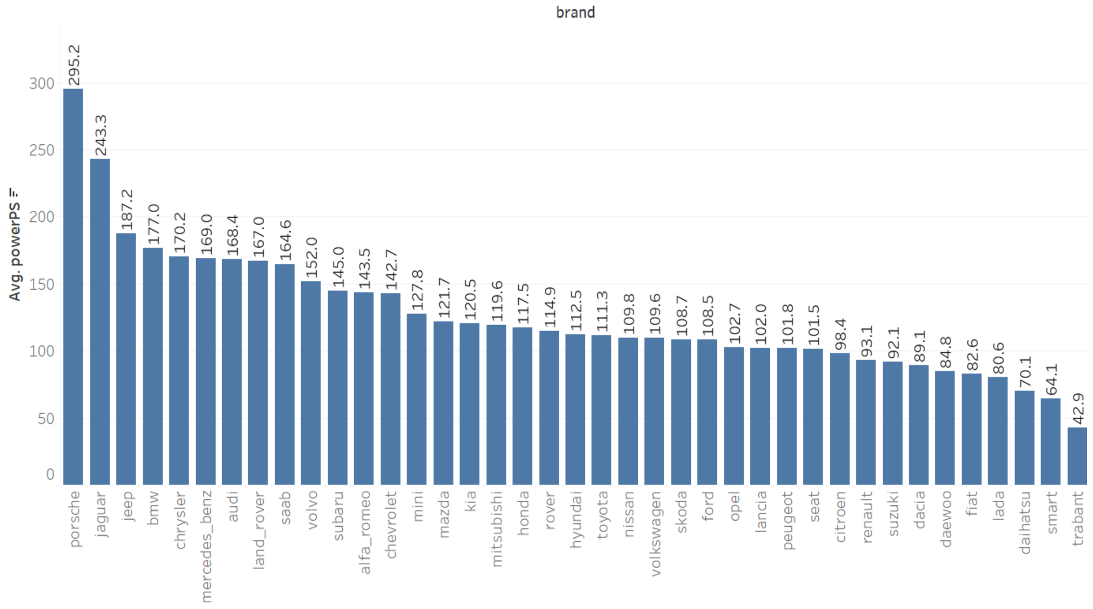


Fig. 20: Average powerPS v/s brand

Inference: As we have seen in earlier plot between average price and brand; Porsche, Land Rover, Jaguar, Jeep, BMW are one of those brands whose average price is higher as compared to other brands. The bar plot between average price and powerPS also states same brands justifying the positive correlation between price and powerPS.

x. Average powerPS v/s fuelType

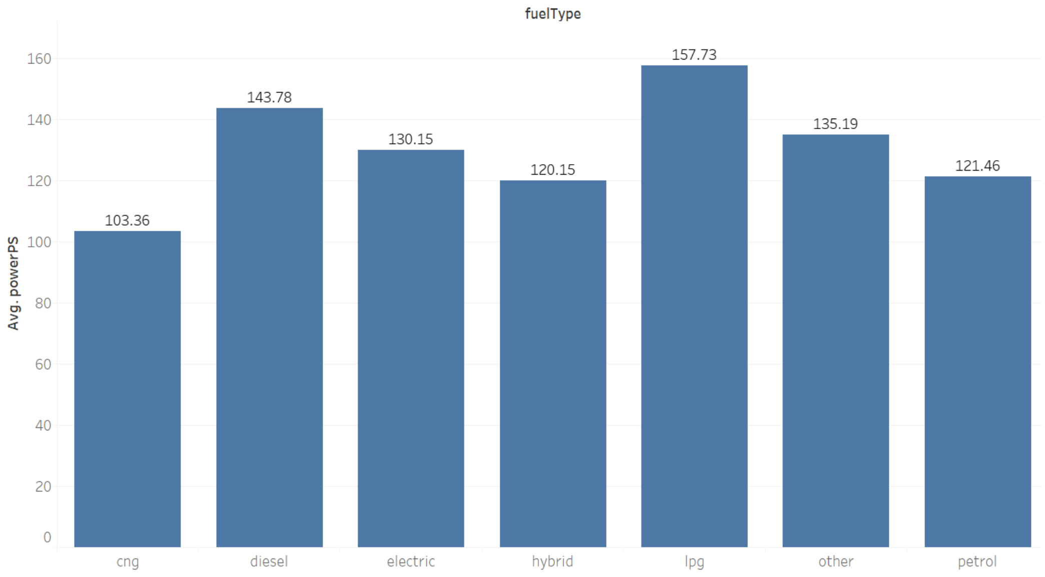


Fig. 21: Average powerPS v/s fuelType

Inference: As Germany is more likely to ban diesel and petrol cars in the country by 2030, LPG and electric cars seems to be an efficient alternative in terms of powerPS performance, also LPG cars are the cheapest cars, this could be one of the main reason as customers are selling diesel and petrol cars, hence available in huge amount in dealer's inventory. Diesel cars with automatic gearbox gives 2nd best average powerPS hugely comes within higher price ranges could be mainly sport cars.

xi. Average powerPS v/s gearbox

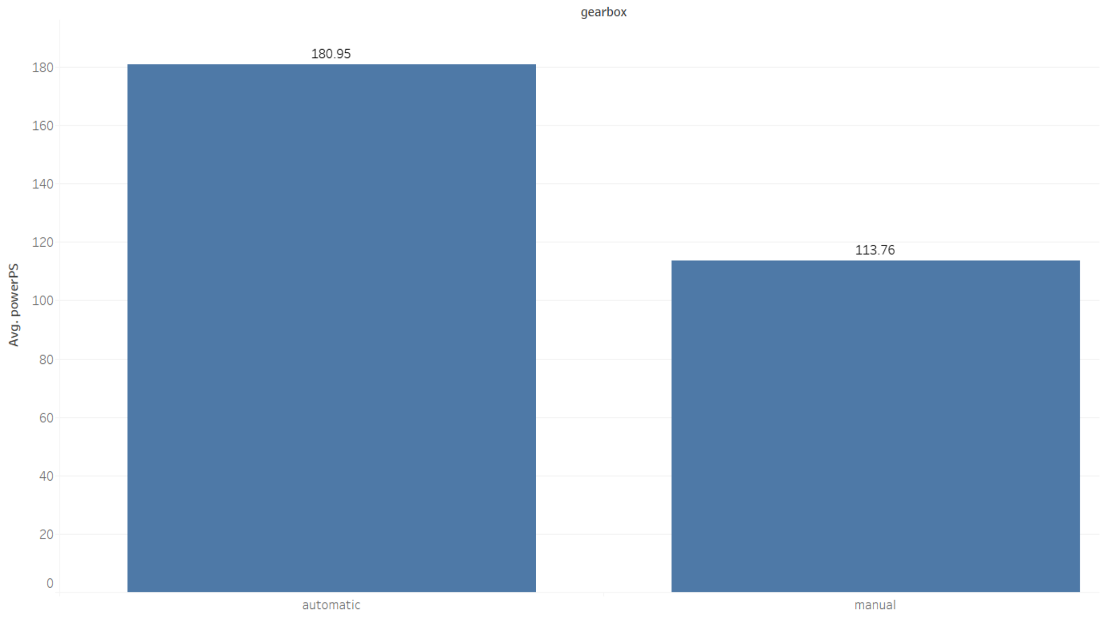


Fig. 22: Average powerPS v/s gearbox

Inference: Cars with low power generally have manual gearbox, whereas, mid-range cars will have an equal number of manual and automatic gearboxes. Nowadays, in many high range cars, manufacturers prefer automatic gearboxes because only then they can achieve their said 0-100 speeds. Additionally, launch control and gear response time will be better in cars with automatic gearboxes.

xii. Count of fuelType in terms of gearbox

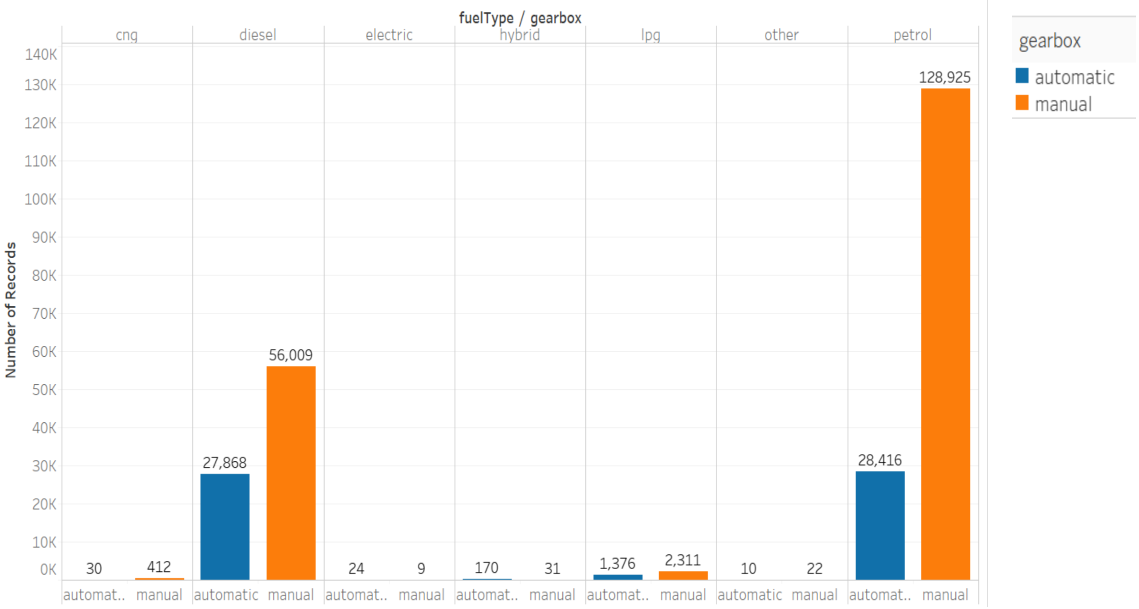


Fig. 23: Count of fuelType in terms of gearbox

xiii. Count of vehicleType in terms of fuelType

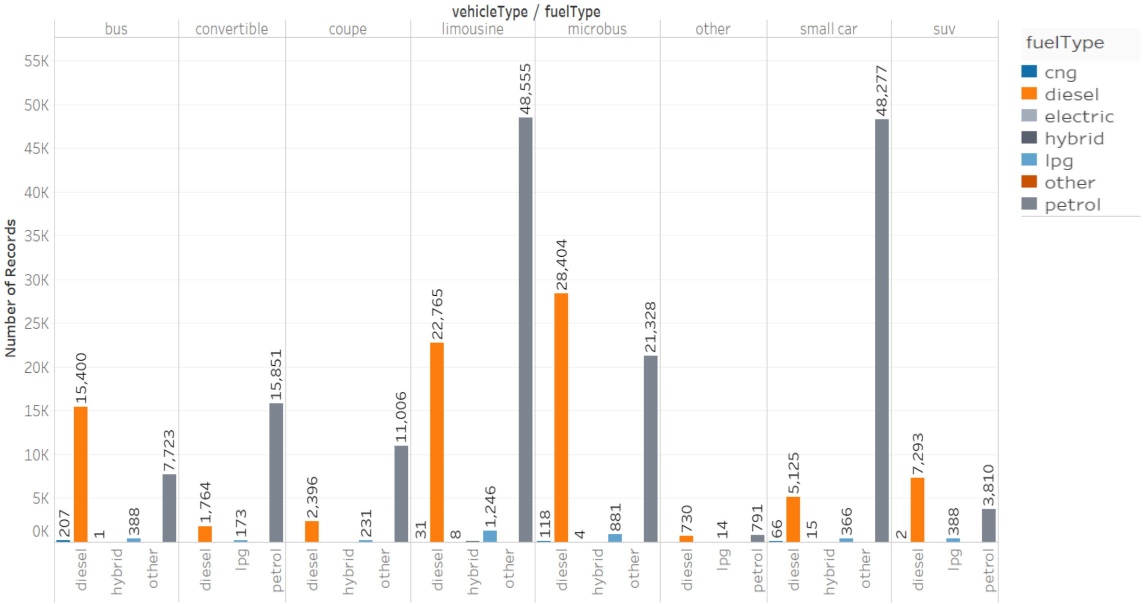


Fig. 24: Count of vehicleType in terms of fuelType

xiv. Count of vehicleType in terms of gearbox

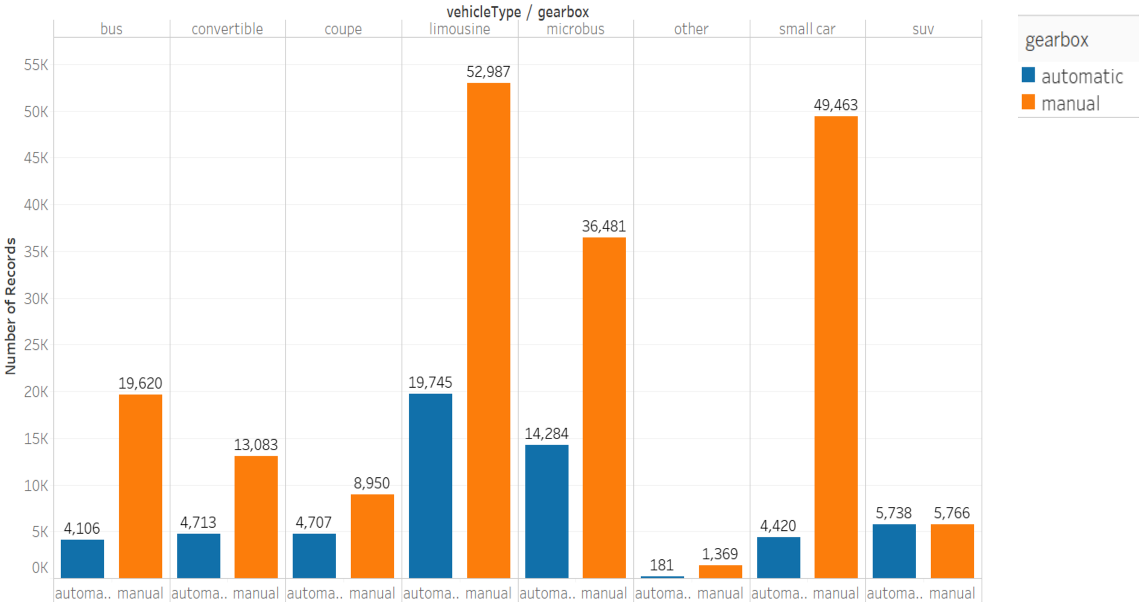


Fig. 25: Count of vehicleType in terms of gearbox

xv. Count of vehicleType in terms of notRepairedDamage

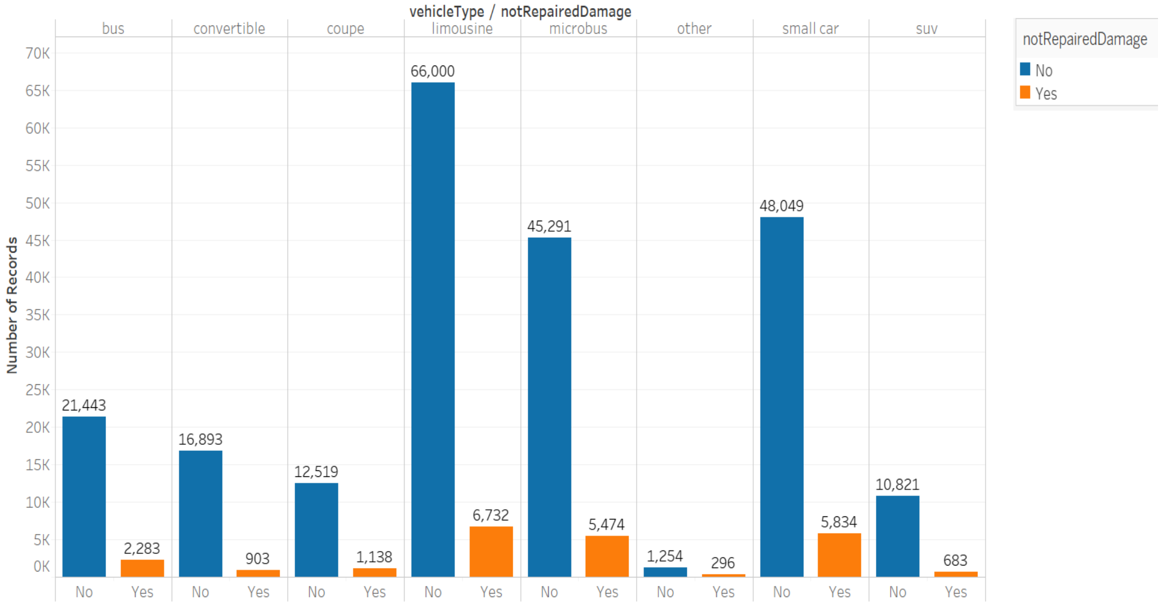


Fig. 26: Count of vehicleType in terms of notRepairedDamage

5.3.Multivariate Analysis

i. Price, vehicleType and gearbox

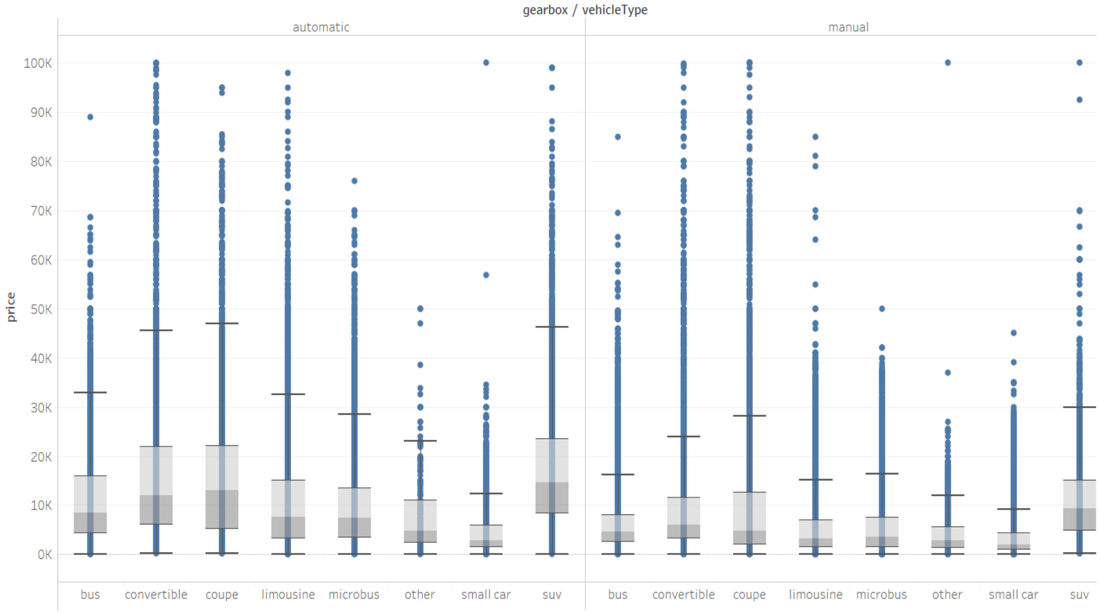


Fig. 27: Price, vehicleType and gearbox

Inference: As we have seen earlier in the plot between price and vehicle type; SUV, convertible and coupe are of higher prices as compared to other vehicle types; when compared in terms of gearbox we get the same results in both automatic and manual gearboxes. Thus in general, SUV, convertible and coupe are of higher prices irrespective of gearbox.

ii. Price, fuelType and gearbox

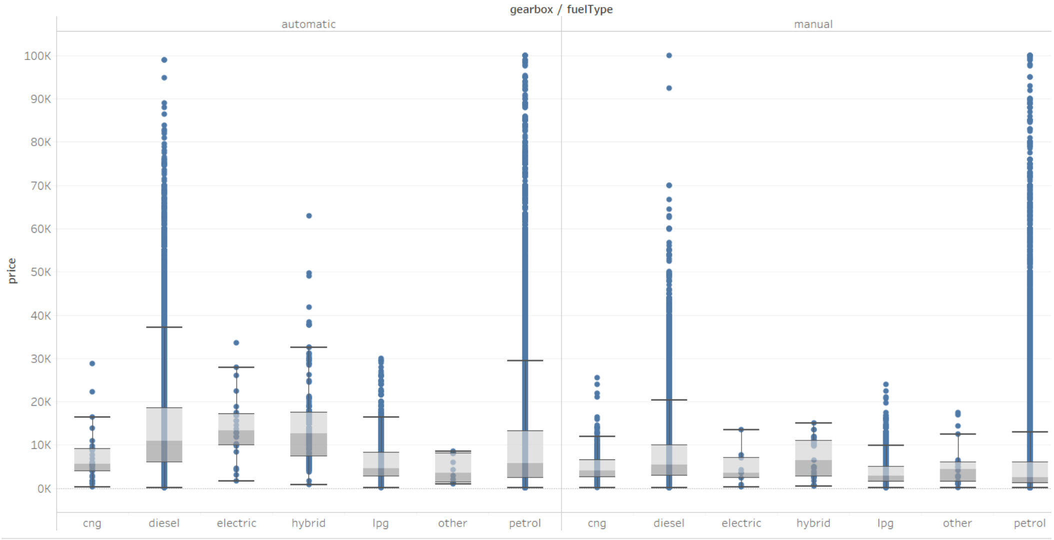
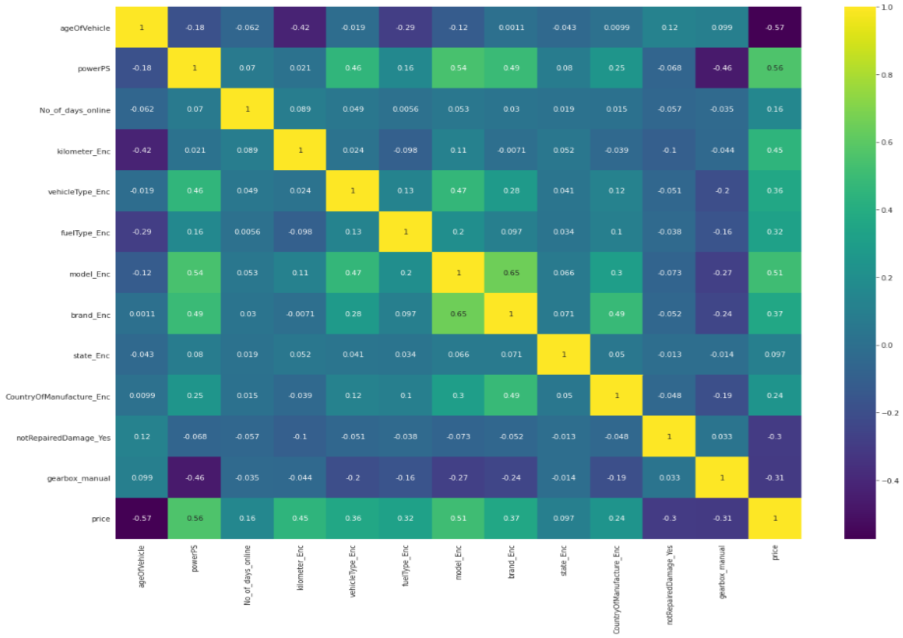


Fig. 28: Price, fuelType and gearbox

Inference: Box plot between price and fuel type stated that the electric and hybrid fuel type cars belong to higher price ranges. But when compared in terms of gearbox, diesel cars are as expensive as electric and hybrid for automatic gearbox; while for manual gearbox, electric cars are of lesser price as compared to diesel and hybrid cars. Hence, gearbox plays a major role for selecting cars in terms of fuel type.

iii. Correlation Matrix



	ageOfVehicle	powerPS	No_of_days_online	kilometer_Enc	vehicleType_Enc	fuelType_Enc	model_Enc	brand_Enc	state_Enc	CountryOfManufacture_Enc	notRepairedDamage_Yes	gearbox_manual	price
ageOfVehicle	1	-0.18	-0.062	-0.42	-0.019	-0.29	-0.12	0.0011	-0.043	0.0099	0.12	0.099	-0.57
powerPS	-0.18	1	0.07	0.021	0.46	0.16	0.54	0.49	0.08	0.25	-0.068	-0.46	0.56
No_of_days_online	-0.062	0.07	1	0.089	0.049	0.0056	0.053	0.03	0.019	0.015	-0.057	-0.035	0.16
kilometer_Enc	-0.42	0.021	0.089	1	0.024	-0.098	0.11	-0.0071	0.052	-0.039	-0.1	-0.044	0.45
vehicleType_Enc	-0.019	0.46	0.049	0.024	1	0.13	0.47	0.28	0.041	0.12	-0.051	-0.2	0.36
fuelType_Enc	-0.29	0.16	0.0056	-0.098	0.13	1	0.2	0.097	0.034	0.1	-0.038	-0.16	0.32
model_Enc	-0.12	0.54	0.053	0.11	0.47	0.2	1	0.65	0.066	0.3	-0.073	-0.27	0.51
brand_Enc	0.0011	0.49	0.03	-0.0071	0.28	0.097	0.65	1	0.071	0.49	-0.052	-0.24	0.37
state_Enc	-0.043	0.08	0.019	0.052	0.041	0.034	0.066	0.071	1	0.05	-0.013	-0.014	0.097
CountryOfManufacture_Enc	0.0099	0.25	0.015	-0.039	0.12	0.1	0.3	0.49	0.05	1	-0.048	-0.19	0.24
notRepairedDamage_Yes	0.12	-0.068	-0.057	-0.1	-0.051	-0.038	-0.073	-0.052	-0.013	-0.048	1	0.033	-0.3
gearbox_manual	0.099	-0.46	-0.035	-0.044	-0.2	-0.16	-0.27	-0.24	-0.014	-0.19	0.033	1	-0.31
price	-0.57	0.56	0.16	0.45	0.36	0.32	0.51	0.37	0.097	0.24	-0.3	-0.31	1

Fig. 29: Correlation Matrix

6. ASSUMPTIONS OF LINEAR REGRESSION

6.1.No Auto Correlation

For Durbin-Watson test we got a value of 2.004 which lies in the acceptance range of 1.5 - 2.5. Moreover, the ACF plot also clearly confirms that there is no auto-correlation between the residuals as clearly shown below.

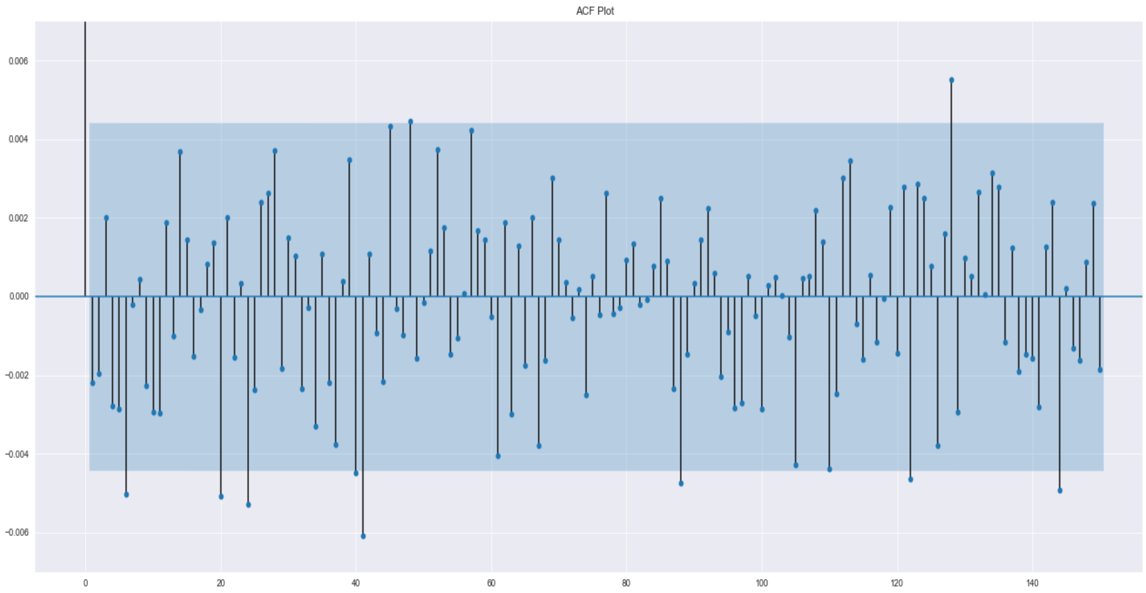


Fig. 30: ACF plot

6.2.Normality of residuals

The Jarque Bera test resulted in a p-value of 0.0 along with a test statistic value of 279629.60 which is greater than the t-critical value of 5.99. Moreover, our residuals deviated from normality towards the extreme which we can clearly see from the Q-Q - plot below. So we rejected the null hypothesis and concluded that residuals are not normal.

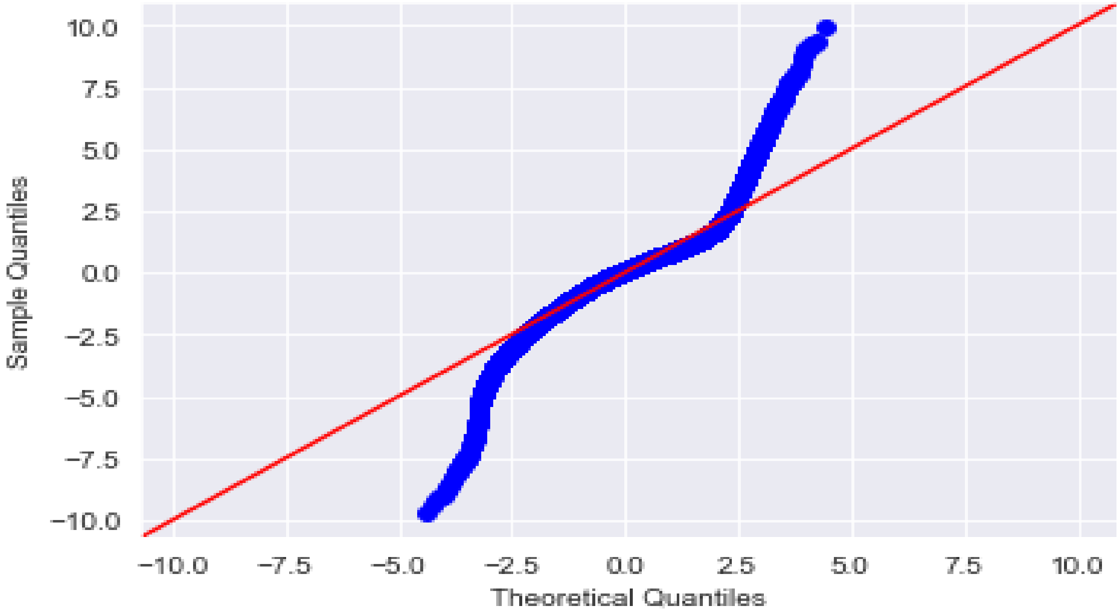


Fig. 31: Q–Q plot

6.3.Linearity of Residuals

We get a p-value of 0.399 for Linear Rainbow test which is higher than 0.05. Moreover, from the below scatter plot the residuals are symmetrically distributed in the former one and around horizontal line in the latter one. In both cases linearity is observed.

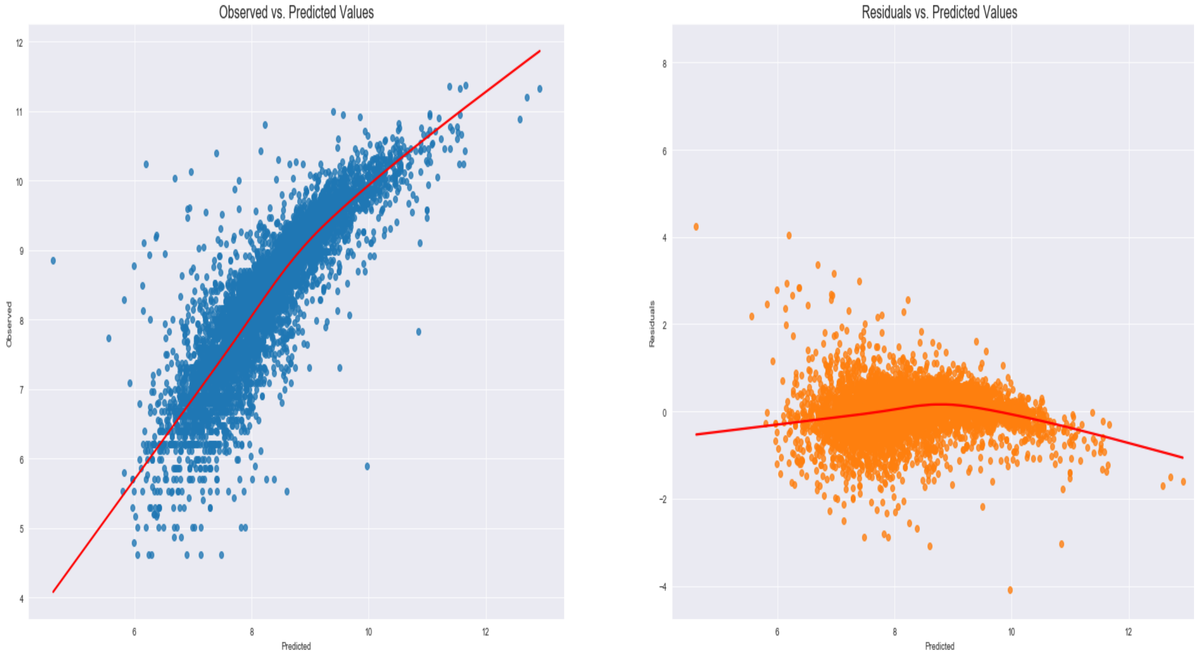


Fig. 32: Scatter Plot of Residuals

6.4.Homoscedasticity

If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic. The Goldfeld-Quandt test gives a p-value of 0.3132 which is higher than 0.05. Moreover, we can visually see that Homoscedasticity is present as shown below.

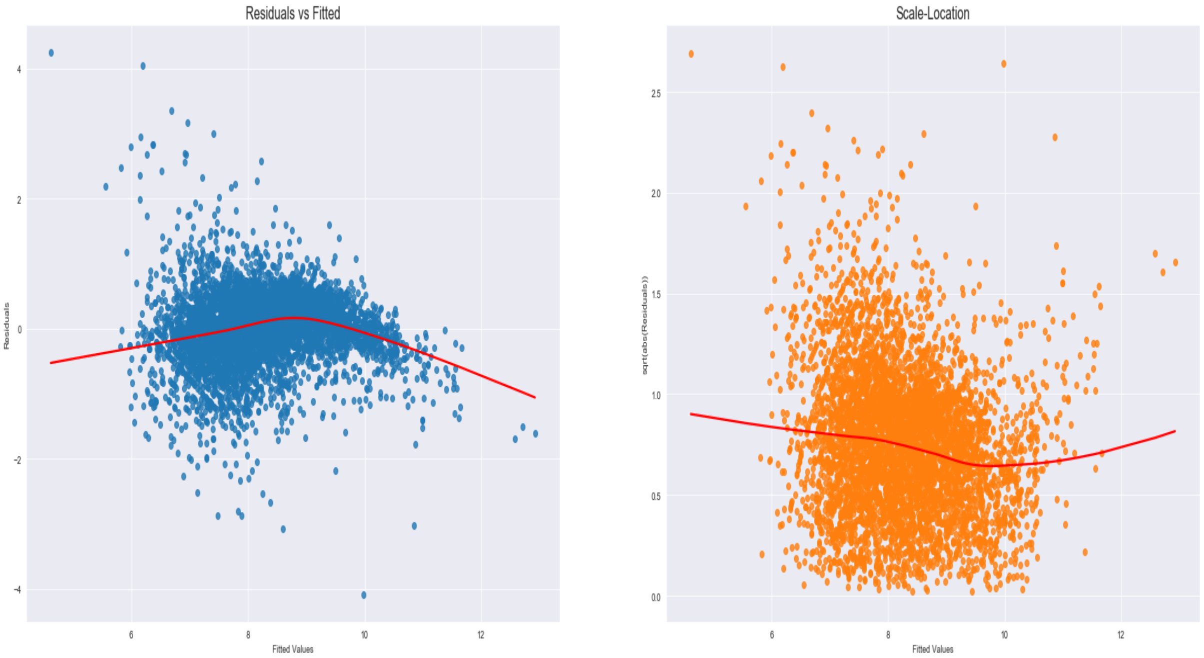


Fig. 33: Scatter plot of residuals v/s fitted

6.5.No Multicollinearity

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.

Following are the results for variance inflation factor test:

Feature	VIF
ageOfVehicle	1.482666
powerPS	1.975102
No_of_days_online	1.015263
vehicleType	1.421693
kilometer	1.356916
fuelType	1.236748
model	2.265722
brand	2.214775
state	1.012671
CountryOfManufacture	1.333841
notRepairedDamage	1.025145
gearbox	1.304639

Table 6: VIF values for the features

The VIF for all the variables are very low as we can see from the above values. So we can conclude that there is no multicollinearity present in our data.

7. SCOPE OF FURTHER WORK

This interim report work examined the dataset and performed the necessary data preparation and exploration required for model building. It could be concluded that, most of the linear regression assumptions were satisfied and hence, linear models cannot be written off completely.

In the further part of the study, both linear as well as non-linear algorithms would be considered for building the model to predict price of used cars.

8. REFERENCES

1. Amjadh Ifthikar and Kaneeka Vidanage (2018) “Valuation of Used Vehicles: A Computational Intelligence Approach”.

2. Dr. Shiva Shankar. K.C (2016) “A Study on Consumer Behavior Towards PreOwned Cars in India”, Indian Journal of Research, 5(11).

3. Edmunds Forecasts (2020) “New Vehicle Sales Drop in March to Close a Down First Quarter in 2020”. Available from: https://www.edmunds.com/industry/press/new-vehicle-sales-drop-in-march-to-close-a-down-first-quarter-in-2020-edmunds-forecasts.html [Accessed 09 June, 2020].

4. Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric (2019) “Car Price Prediction using Machine Learning Techniques”, TEM Journal, 8 (1), 113-118.

5. Evgeniya Koptyug (2019) “Leading import countries for motor vehicles from Germany based on export value 2018”. Available from: https://www.statista.com/statistics/587701/leading-import-countries-german-motor-vehicles-by-export-value/ [Accessed 09 June, 2020].

6. Evgeniya Koptyug (2020) “Revenue of the market for second-hand cars in Germany from 2000 to 2019 (in billion euros)”. Available from: https://www.statista.com/statistics/589610/revenue-used-cars-germany/ [Accessed 19 June, 2020].

7. Graham Rapier (2020) “Used car may get even cheaper than in the last recession as the coronavirus forces dealerships to offer unprecedented deals”. Available from: https://www.businessinsider.in/business/news/used-cars-may-get-even-cheaper-than-in-the-last-recession-as-the-coronavirus-forces-dealerships-to-offer-unprecedented-deals/articleshow/75072952.cms [Accessed on 10 June, 2020].

8. Greg Gardner (2018) “Auto Sales Are Down. Here's Why They'll Continue To Fall (Forbes)”. Available from: https://www.forbes.com/sites/greggardner/2018/03/12/auto-sales-are-down-heres-why-theyll-continue-to-fall/#1612ab622dcb [Accessed 09 June, 2020].

9. Harikrushna Vanpariya (2018) “Using Different Machine Learning Techniques for Predicting the Price of Used Cars”, International Journal for Scientific Research & Development, 6(10).

10. Ken Research (2020) “Germany Used Car Market is expected to reach about EUR 105 Billion in Gross Transaction Value (GTV) by 2023: Ken Research”. Available from: https://www.kenresearch.com/blog/2020/02/germany-used-car-market-value/ [Accessed 25 June, 2020].

11. Mariana Listiani (2009) “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application”, Master Thesis.

12. Monika Singh (2019) “Germany Used Car Market Outlook to 2023 - Surge in Multi-Brand Dealerships Coupled with Improved Quality and Inspection of Used Cars to boost Used Cars

Market". Available from: https://www.kenresearch.com/automotive-transportation-and-warehousing/automotive-and-automotive-components/germany-used-car-market-outlook/277739-100.html [Accessed 18 June, 2020].

13. Mordor Intelligence "Used car market– growth, trends, and forecast (2020 - 2025)". Available from: https://www.mordorintelligence.com/industry-reports/global-used-car-market-growth-trends-and-forecast-2019-2024 [Accessed 10 June, 2020].

14. Sameerchand Pudaruth (2014) "Predicting the Price of Used Cars using Machine Learning Techniques", International Journal of Information & Computation Technology, 4(7), 753-764.

15. Statista Research Department (2020) "Worldwide car sales 2010 – 2020". Available from: https://www.statista.com/statistics/200002/international-car-sales-since-1990/ [Accessed 09 June, 2020].

16. Stephen Edelstein (2020) "COVID-19 may already be causing new car sales to fall". Available from: https://www.digitaltrends.com/cars/coronavirus-set-to-torpedo-new-car-sales-for-march-2020-and-beyond-report-says/ [Accessed 09 June, 2020].