1. PROBLEM

Regression Analysis on Predicting the Value of a Used Vehicle

a. Client:

(1) Individual Buyers:

Whoever interested in buying a new car wonders about the actual value of the specific car with the one which was asked by the seller. So it is important for all to come to a better understanding of the values of the cars.

(2) Vehicle Dealers:

Most dealers would like to learn the value of that individual car, and determine its value later on.

(3) Individual Sellers:

Most private sellers would need the value of their car since the value is not constant, it changes considering the depreciation, the repairs etc.

(4) Websites:

(Craigslist, autoscout24, mobile.de) or Applications created to help private parties or dealers sell their vehicles.

b. Data set:

(1) Link of Data from Kaggle (https://www.kaggle.com/orgesleka/used-cars-database/data)

(2) Data Set includes 19 features and 371528 data points/observations.

(3) The data seems raw. Has too many missing values, outliers. Also includes nonsense/ wrong entries like 9999 as year of Registration.

(4) This data includes the data points/observations from 2016.

(5) Target Feature is 'price'. Since it is the price of an individual car, it is continuous.

2. DATA WRANGLING

- *Examining the Features and Samples*
  - ✓ *Features*

i. *dateCrawled:*

ii. *name: Name of the Vehicle such as Jeep_Grand_Cherokee_"Overland".*

iii. *seller: Seller Type such as private or dealer.*

iv. *offerType: Fixed Price or Open to any Offer.*

v. *price: Value of the Vehicle*

vi. *abtest : abtesting*

vii. *vehicleType : Type of the Vehicle such as sedan, station, small car, bus, cabrio, coupe, SUV….*

viii. *yearOfRegistration: Year of the Registration of the Vehicle*

ix. *gearbox: Transmission type such as automatic or manual.*

x. *powerPS: the horse power of the vehicle*

xi. *model: model of the vehicle such as golf, polo, Passat….these are all VW.*

xii. *Kilometer: the Mileage of the Vehicle.*

xiii. *monthOfRegistration : Month of Registration.*

xiv. *fuelType : Type of the Fuel used by the vehicle such as gas, diesel, lpg, hybrid, electic*

xv. *brand: the Brand of the Vehicle such as VW, BMW, Mercedes, Honda, Toyota*

xvi. *notRepairedDamage: Condition of the Vehicle whether it has any repair history or damage*

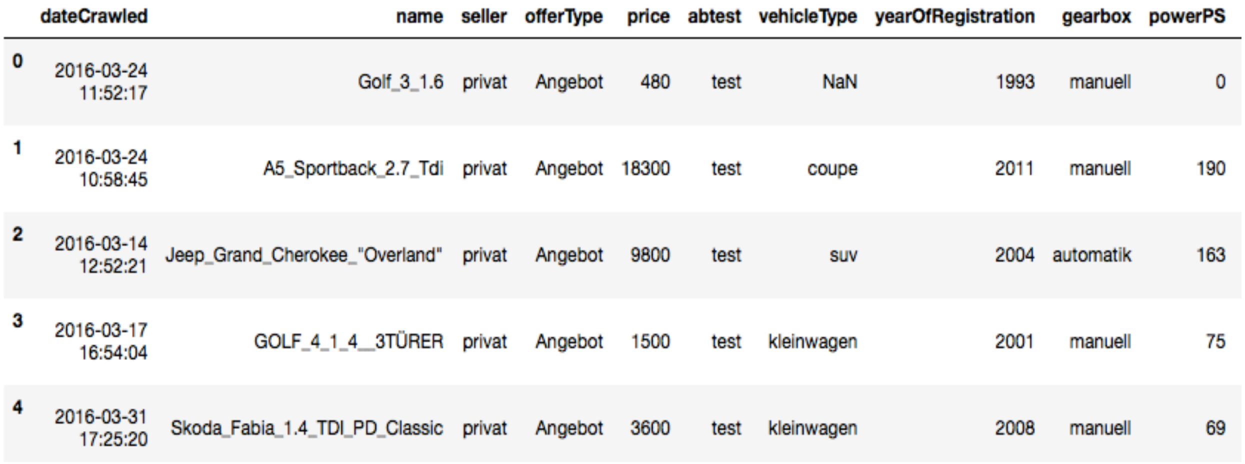
xvii. *dateCreated : date of the posting created.*

xviii. *nrOfPictures : Number of Pictures of the Posting.*

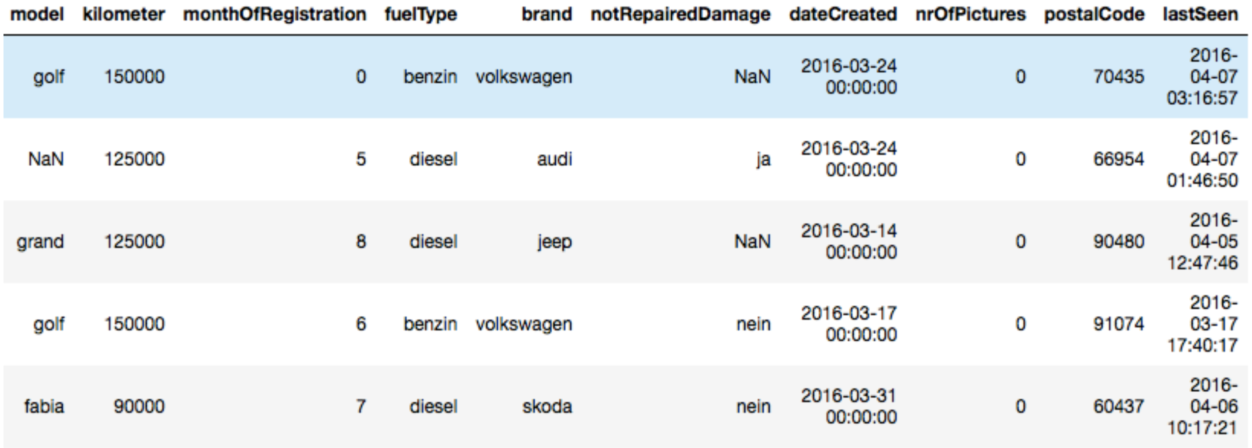
xix. *postalCode : Zipcode of the Area where the vehicle is on sale.*

xx. *lastSeen : the date of the click when it was seen by a potential customer.*

xxi. *Age: the Age of the Vehicle*

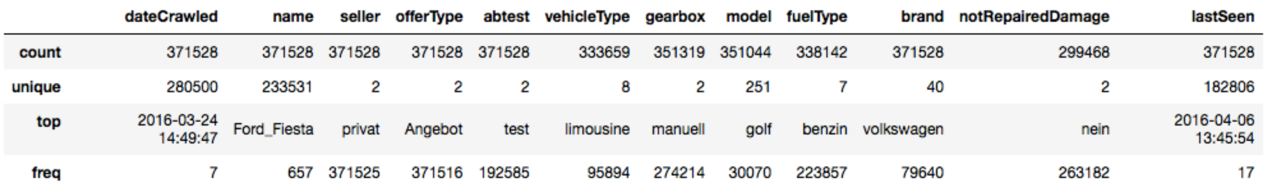


	dateCrawled	name	seller	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS
0	2016-03-24 11:52:17	Golf_3_1.6	privat	Angebot	480	test	NaN	1993	manuell	0
1	2016-03-24 10:58:45	A5_Sportback_2.7_Tdi	privat	Angebot	18300	test	coupe	2011	manuell	190
2	2016-03-14 12:52:21	Jeep_Grand_Cherokee_"Overland"	privat	Angebot	9800	test	suv	2004	automatik	163
3	2016-03-17 16:54:04	GOLF_4_1_4__3TÜRER	privat	Angebot	1500	test	kleinwagen	2001	manuell	75
4	2016-03-31 17:25:20	Skoda_Fabia_1.4_TDI_PD_Classic	privat	Angebot	3600	test	kleinwagen	2008	manuell	69



model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage	dateCreated	nrOfPictures	postalCode	lastSeen
golf	150000	0	benzin	volkswagen	NaN	2016-03-24 00:00:00	0	70435	2016-04-07 03:16:57
NaN	125000	5	diesel	audi	ja	2016-03-24 00:00:00	0	66954	2016-04-07 01:46:50
grand	125000	8	diesel	jeep	NaN	2016-03-14 00:00:00	0	90480	2016-04-05 12:47:46
golf	150000	6	benzin	volkswagen	nein	2016-03-17 00:00:00	0	91074	2016-03-17 17:40:17
fabia	90000	7	diesel	skoda	nein	2016-03-31 00:00:00	0	60437	2016-04-06 10:17:21

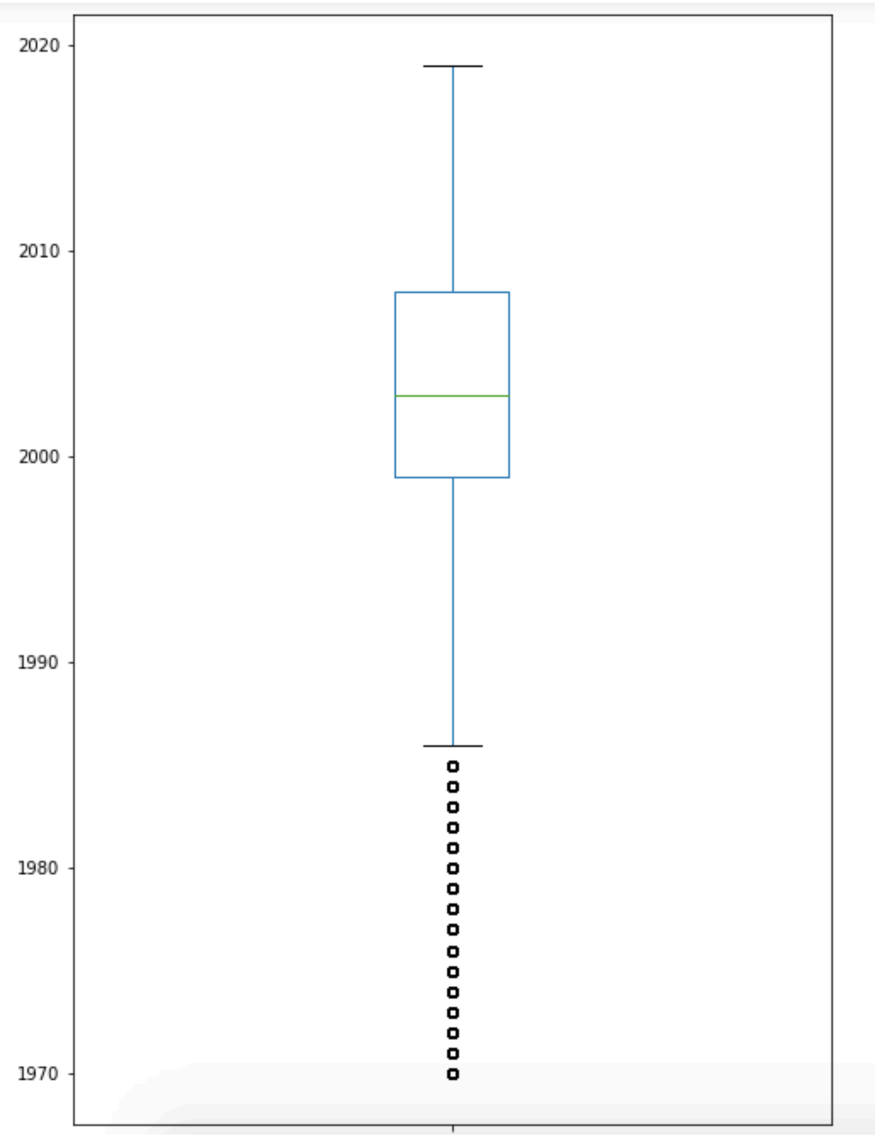
- *Determining the Features to be Dropped*



	dateCrawled	name	seller	offerType	abtest	vehicleType	gearbox	model	fuelType	brand	notRepairedDamage	lastSeen
count	371528	371528	371528	371528	371528	333659	351319	351044	338142	371528	299468	371528
unique	280500	233531	2	2	2	8	2	251	7	40	2	182806
top	2016-03-24 14:49:47	Ford_Fiesta	privat	Angebot	test	limousine	manuell	golf	benzin	volkswagen	nein	2016-04-06 13:45:54
freq	7	657	371525	371516	192585	95894	274214	30070	223857	79640	263182	17

- ✓ Seller feature: 3 out of 371528 observations are dealer. So this feature can be dropped.
- ✓ Offer Type feature : 12 out of 371528 observations are Gesuch. So this feature can be dropped.
- ✓ Number of Pictures feature has all 0 values. So this feature can also be dropped.
- ✓ PowerPS Feature : 40820 observations are 0. So this needs to be dealt with.
- ✓ vehicleType has 37869, gearbox has 20209, model has 20484, fuelType has 33386, notRepairedDamage has 72060 NULL values.
- ✓ 13 unique numbers for the 'kilometer' (mileage) exist. So we can keep them all since there is no outlier.
- ✓ 7 features are discrete, numbers, whereas
- ✓ 12 features are object (string, datetime....)
- ✓ Shape of the data is 371528x20 (‘Age’ feature has been created to better examine the data set.).
- ✓ "vehicleType, gearbox, model, fuelType, brand, notRepairedDamage" Features have missing values!!!

- *Determining the Filtering Parameters*
  - ✓ Duplicated samples should be deleted. There are only 4 rows duplicated.
  - ✓ German words like ‘nein’, ‘ja’ have to be translated with ‘replace’ function.
  - ✓ After examining the samples with horsepower higher than 600 and lower than 5, it is more likely to get rid of these as outliers.
  - ✓ Since the data was scraped in 2016, samples with a registration date of 2017 and above should be deleted.

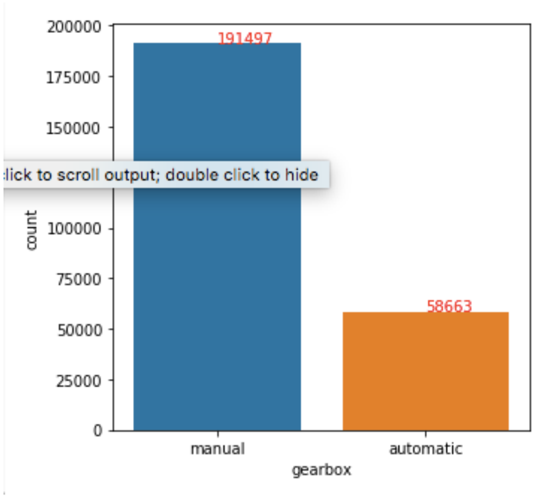


  - ✓ The number of cars with a value of above 100.000 is 403. So I am assuming these as outliers. However, despite the fact that the number of cars with a ‘0’ value is 10778, I will keep them since these vehicles could be considered to be given away as a present.
  - ✓ The percentage of samples/observations lost after filtering is % 14.5.

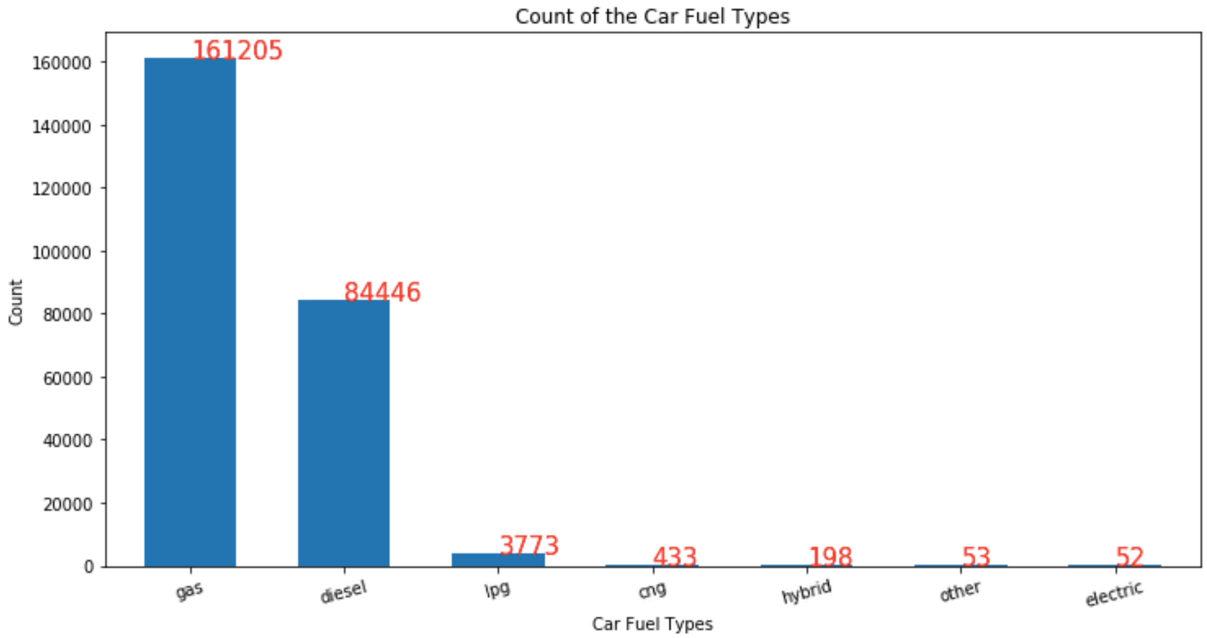
- *Determining how to handle the Missing Values*
  - ✓ It may mislead to keep the missing values by replacing with an interpolation for this data set. So considering the richness of the number of samples, I am opting for getting rid of the missing of values.
  - ✓ The percentage of samples/observations lost after deleting missing values is % 30.
  - ✓ The shape of the data set before the data wrangling is 371528x19, and its shape became 250160x16 after.

3. DATA VISUALIZATION (EDA)

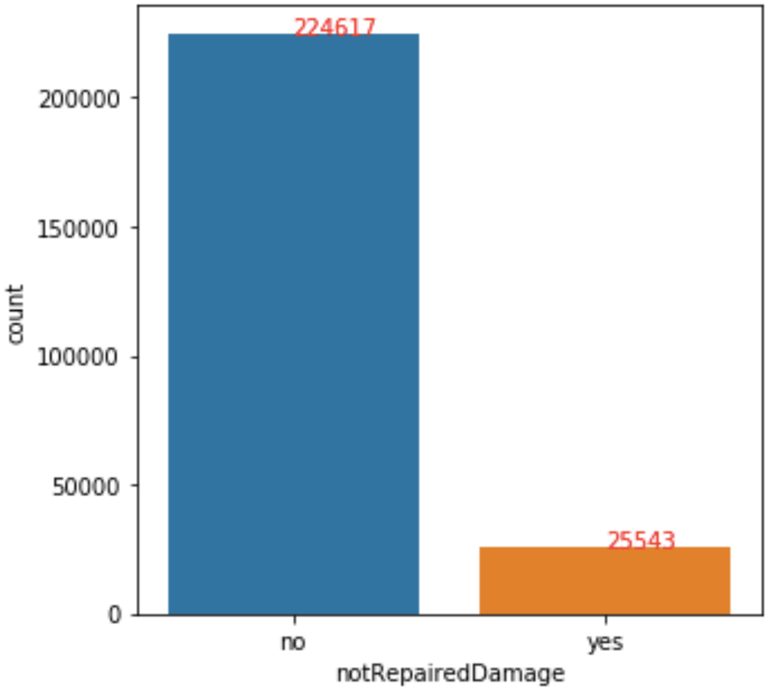
- Univariate Visualizations



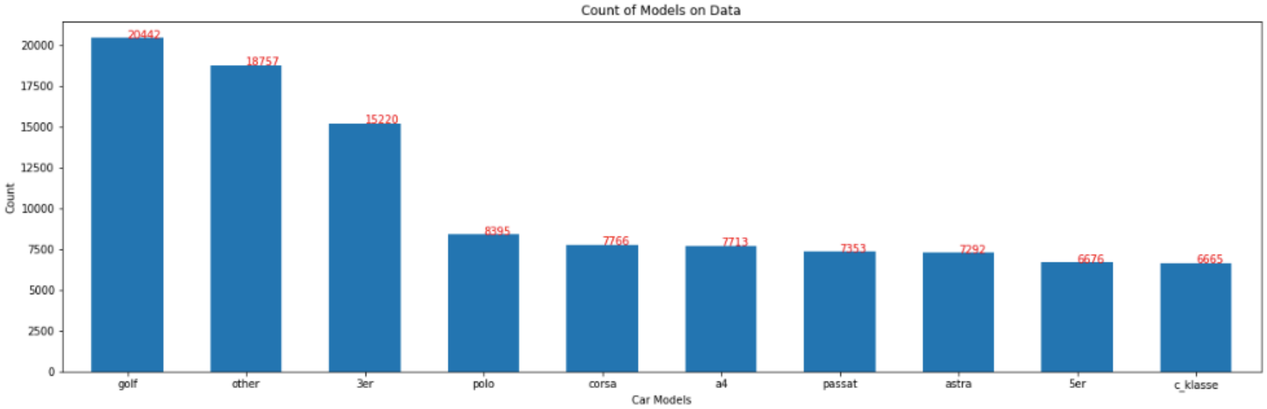
Manual vehicles are more common in Europe. (Knowing that the data set comes from a European website)



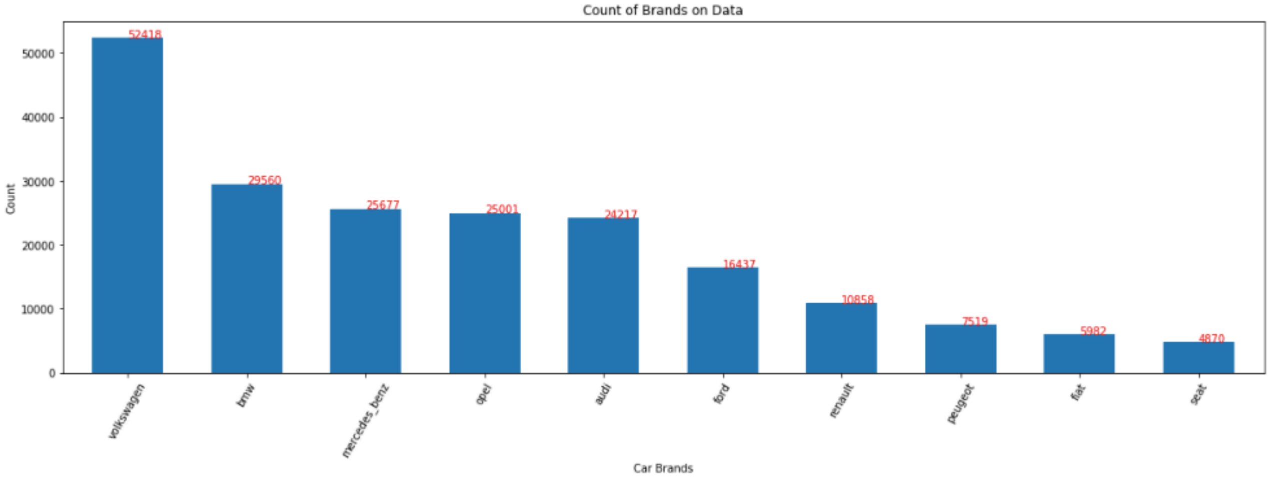
Number of vehicles operated by gas is 161.205, whereas the number of vehicles operated by diesel fueltype is 84.446. Diesel operated vehicles are very popular in Europe. In addition, liquid petroleum gas (LPG), compressed natural gas(CNG) operated vehicles exist in spite of the of the fact that their number is low. Electric and Hybrid vehicles should be instigated in Europe because their number is really low. (More could be found in the Capstone Project Report)



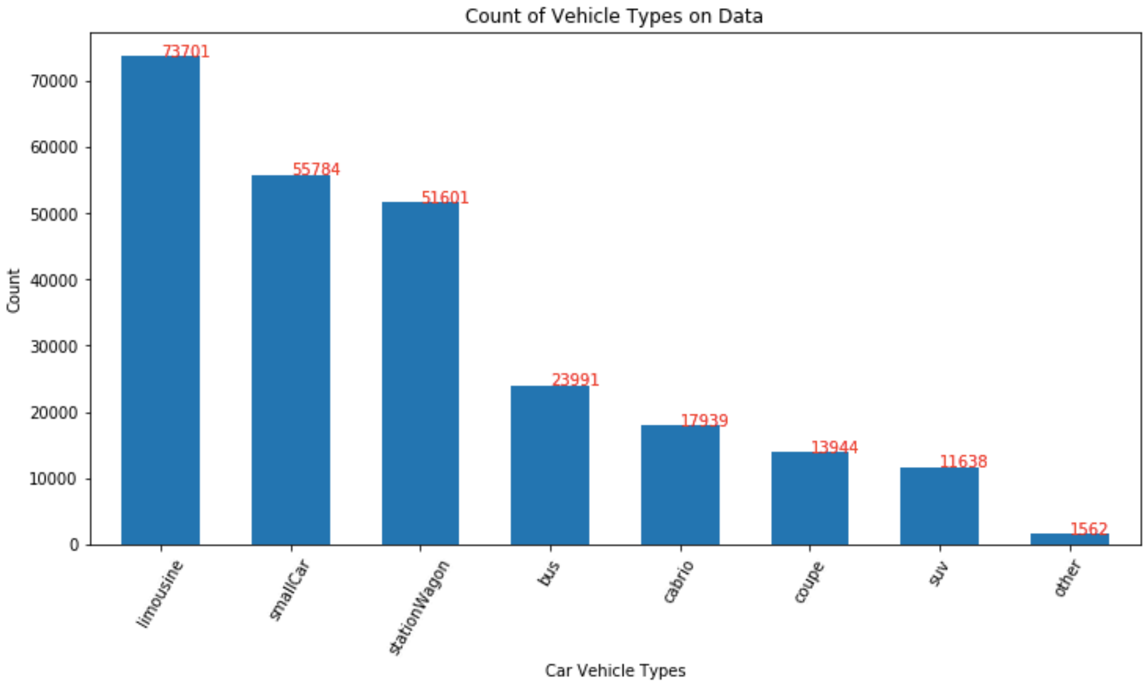
Most Vehicles do not have any repair or damage history. This may be misleading, because most pre-owned cars don't tend to speak out this information once the potential customer engages and indicates interest in the vehicle.



Golf, Polo, Corsa, Passat are the most common models in Europe.

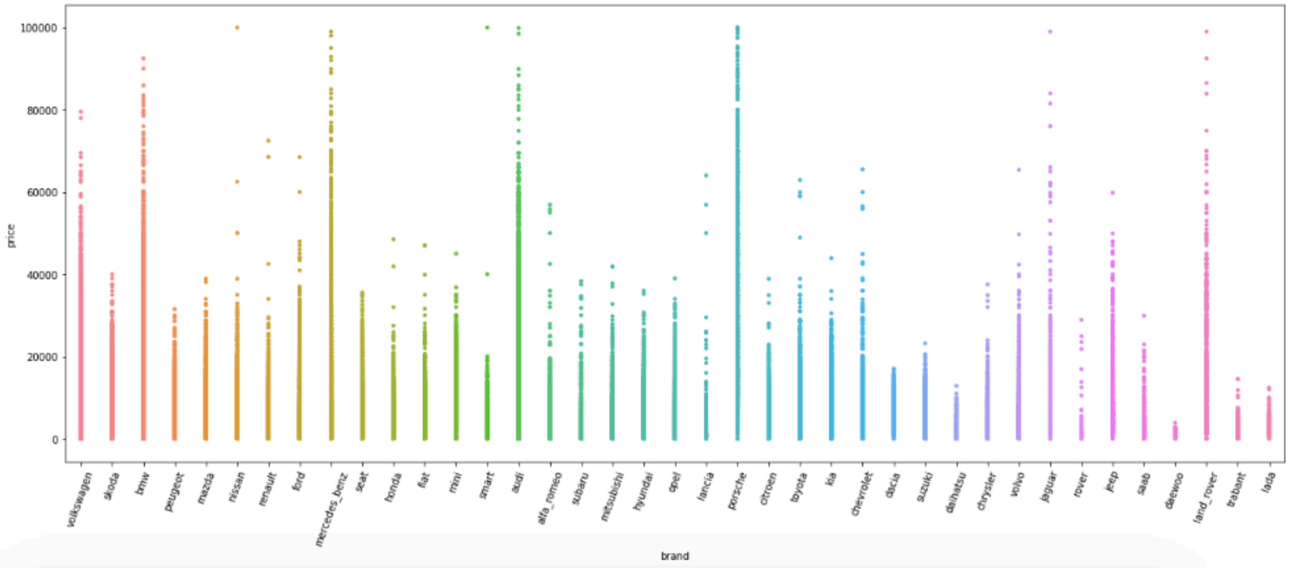


VW, BMW, Mercedes Benz, Opel, Audi, Ford keep the leading positions compared to other brands in Europe.

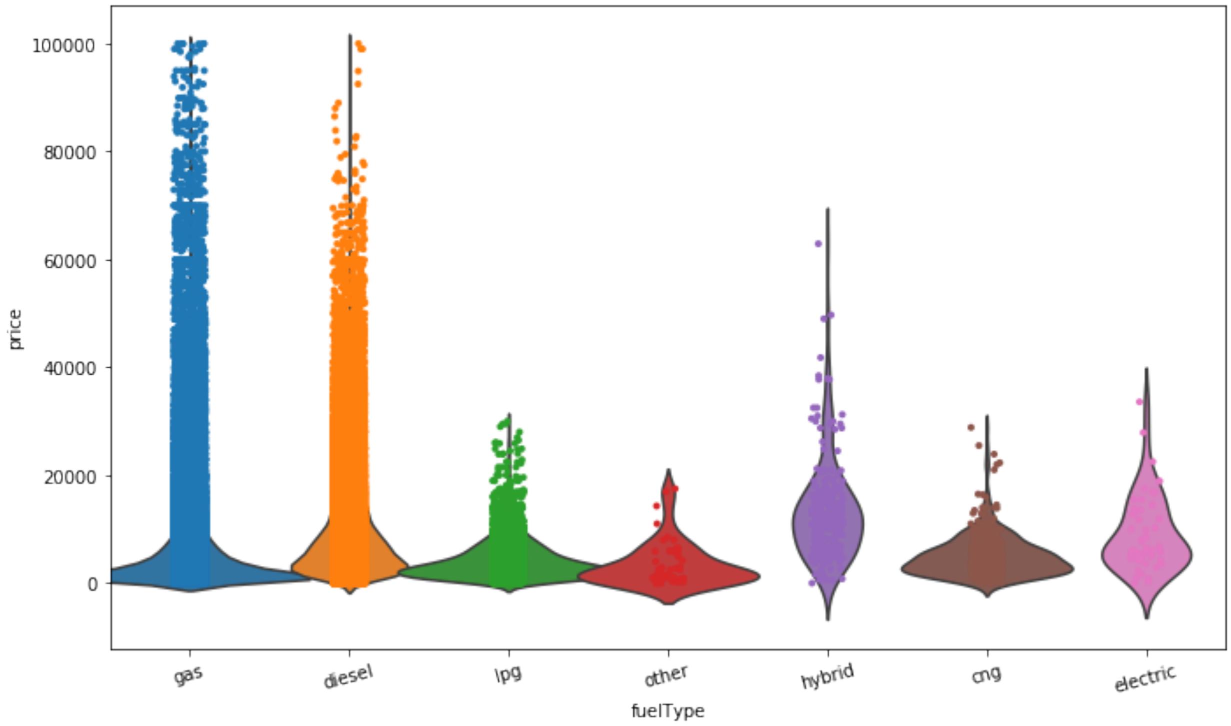


Number of Limousine (Sedan) vehicles is 73701, wheres that of SUVs is 11638.

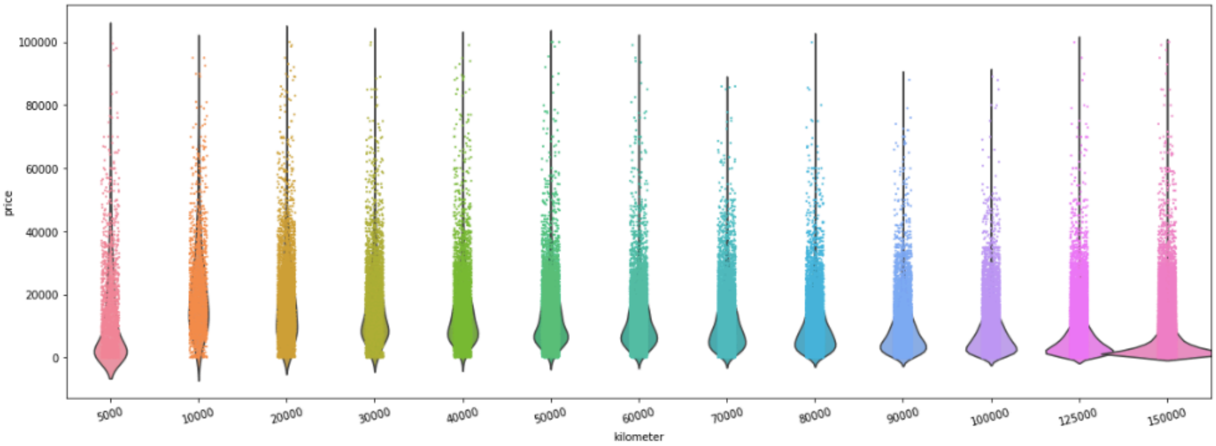
- Bivariate Visualizations



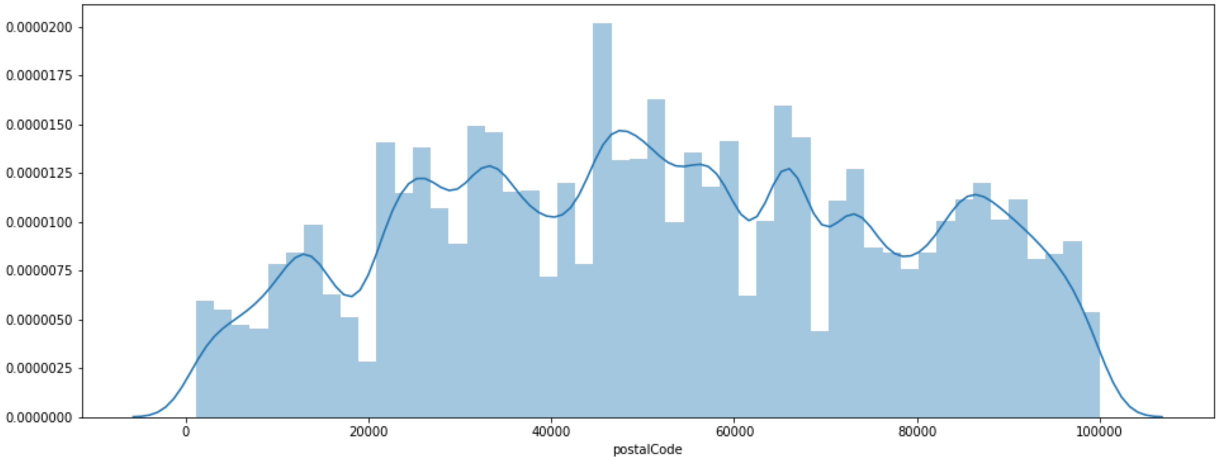
This chart is a visualization (strip plot) of brand and price features. It is easy to examine that Mercedes_benz, Audi, Porsche, BMW, Land Rover are the most expensive vehicles.



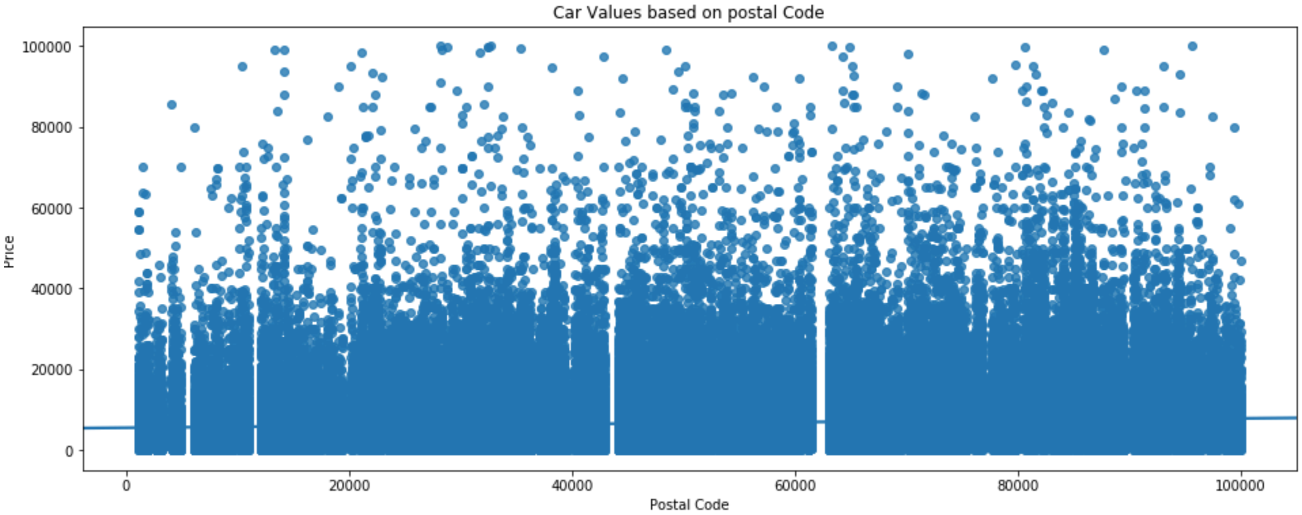
The most expensive vehicles are usually of the fueltype , gas however, there are some number of diesel operated vehicles with a high value. This chart indicates that most hybrid and electric vehicles are not of high values although they are the new generation vehicles.



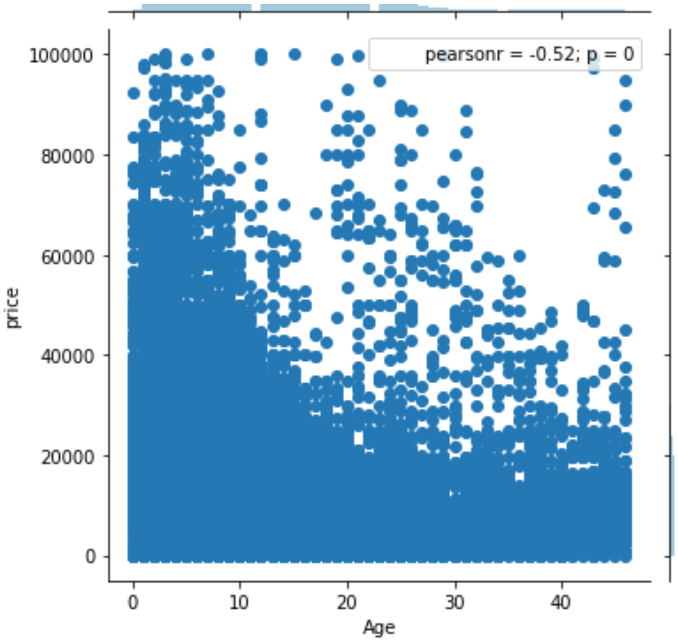
The vehicles with a low mileage tend to be the most expensive vehicles as shown on the chart. Some old vehicles with a high value exist. This is most likely due to the condition of the vehicle (classic).



Postal code of '45000' is the town that the highest number of cars are loaded into the website for sale.

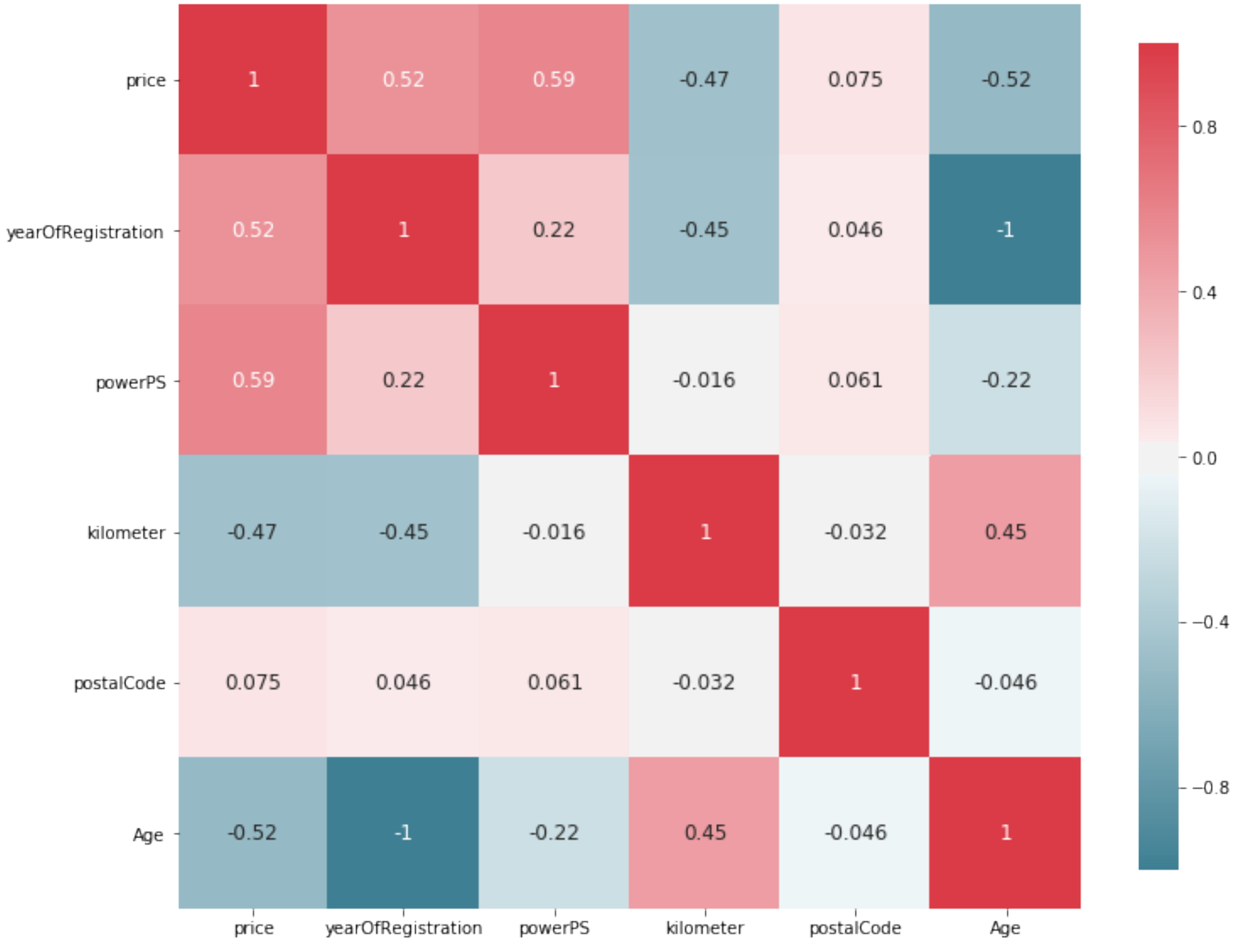


A rise on the values of the vehicles loaded from the towns with postel Codes of 80.000 and 85.000 is obvious on the scatter.



It is clear that the vehicles until the age of 10 have the highest values although some old vehicles with a high value exist. It is easy to interpret these vehicles are classic vehicles.

- Multivariate Visualizations

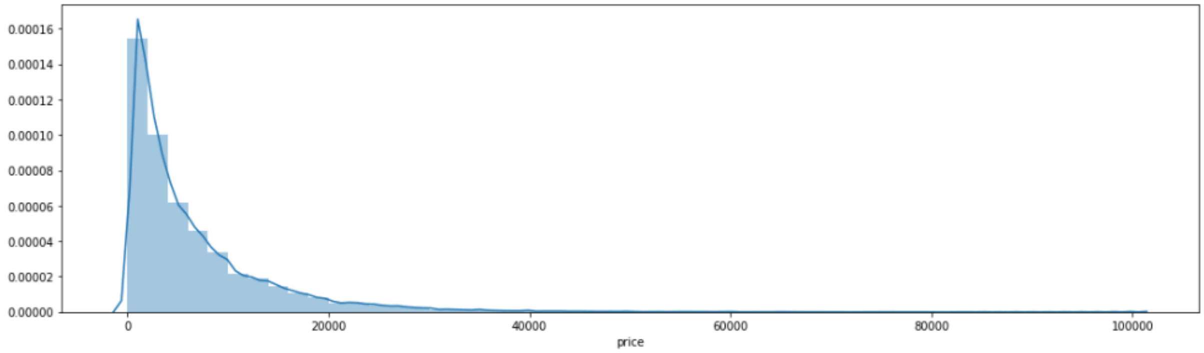


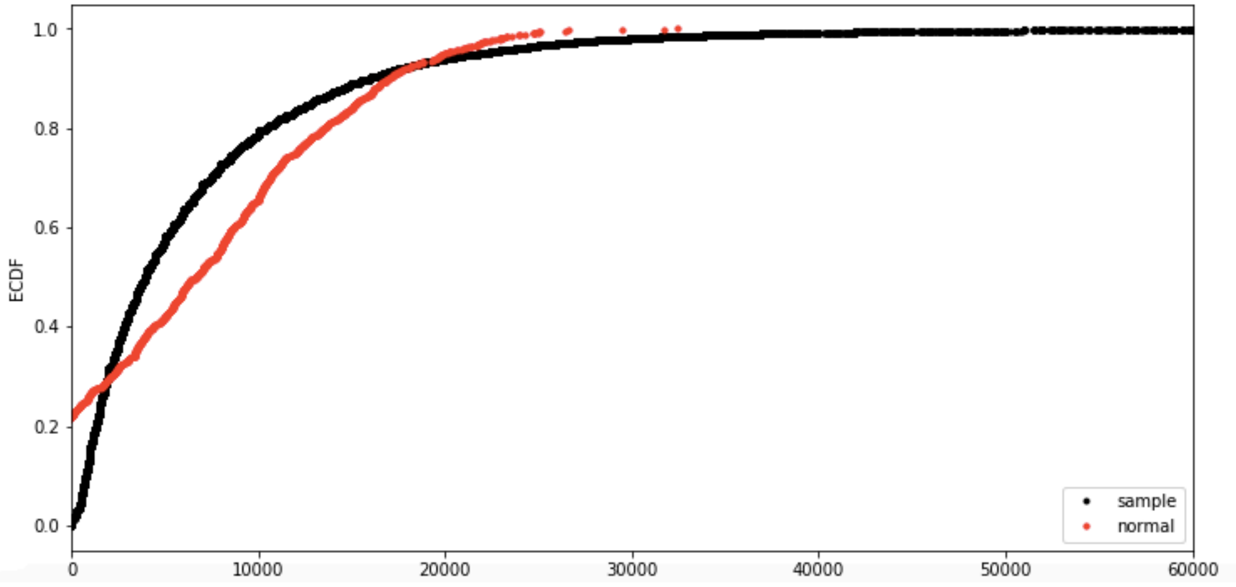
	price	yearOfRegistration	powerPS	kilometer	postalCode	Age
price	1	0.52	0.59	-0.47	0.075	-0.52
yearOfRegistration	0.52	1	0.22	-0.45	0.046	-1
powerPS	0.59	0.22	1	-0.016	0.061	-0.22
kilometer	-0.47	-0.45	-0.016	1	-0.032	0.45
postalCode	0.075	0.046	0.061	-0.032	1	-0.046
Age	-0.52	-1	-0.22	0.45	-0.046	1

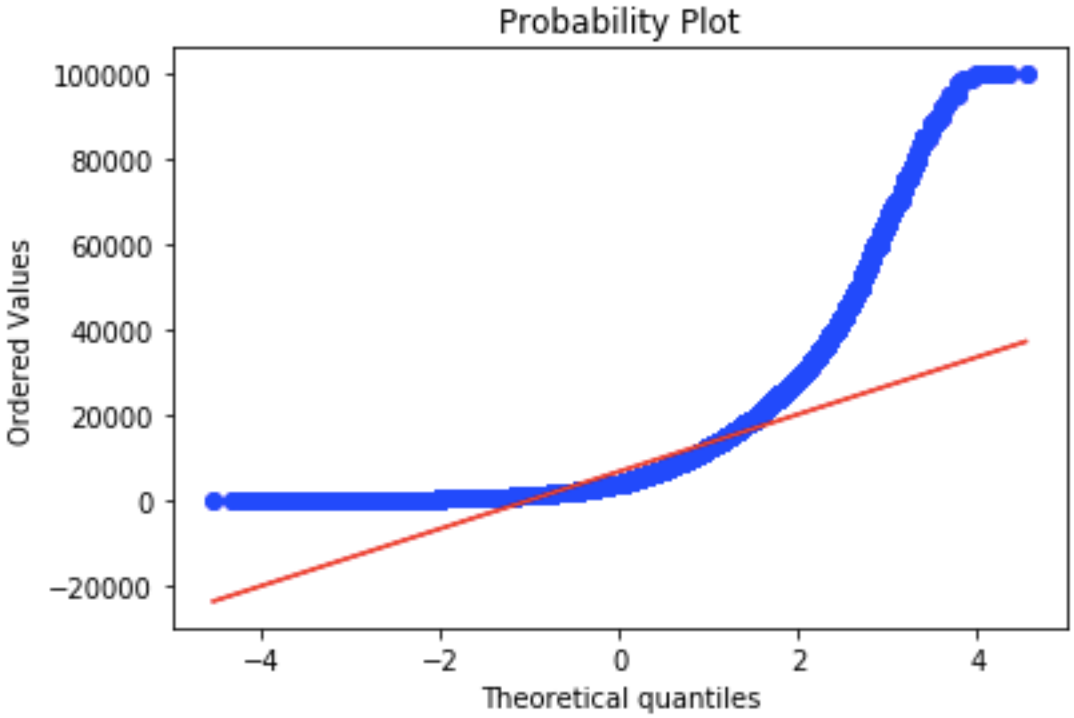
From the chart above, we can interpret that 'yearOfRegistration/Age (same indicator)' features have a positive correlation of .52 with 'price' feature. 'powerPS' feature which indicates the horsepower of the vehicles have a positive correlation of .59 with 'price' feature. 'PostalCode' has a slight correlation with 'price' feature.

- *Inferential Statistics*:

The values (price) of used vehicles data do not seem to come from a normal distribution. According to the linear regression theorem, it doesn't have to come from normal distribution taking into consideration that value feature will be the target (y) as long as the residuals (i.i.d) are normally distributed.







4. Regression Analysis

I first trained a linear regression model with only numerical features to establish a baseline of how well it predicted the target feature. However, after analyzing the results of the model, I discovered the outliers and multicollinearity should have been taken into consideration. Therefore, I also trained a Ridge regression and a LASSO regression to see if it would predict the target feature with more accuracy. Then I repeated the same procedure with including non-numeric features to the model by using get dummies function. This additional procedure helped me get better results.

To measure the accuracy of the models, I utilized the coefficient of determination and the mean squared error.

Coefficient of Determination (R-Squared):

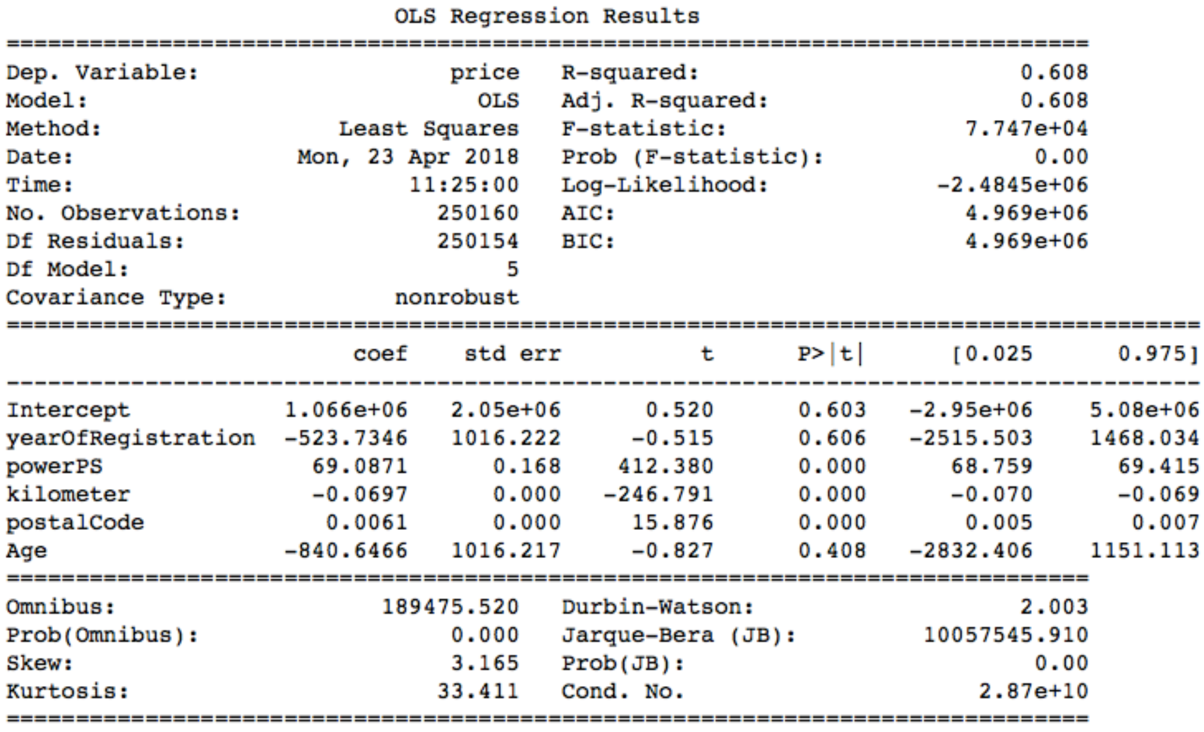
The coefficient of determination indicates how much of the variance in the target variable is explained by the model. In other words, it measures the goodness-of-fit of the model. It is given as a percentage with a high value indicating that the model explains all the variability well while a low value indicating that the model doesn't explain the variability well.

However, R-squared does not indicate whether a regression model is adequate. For example, while a model with high R-squared might explain the variance in the target variable well, it might generalize well to future data due to overfitting. As a result, we will also be another metric to assess the accuracy of the model.

Mean Squared Error (MSE):

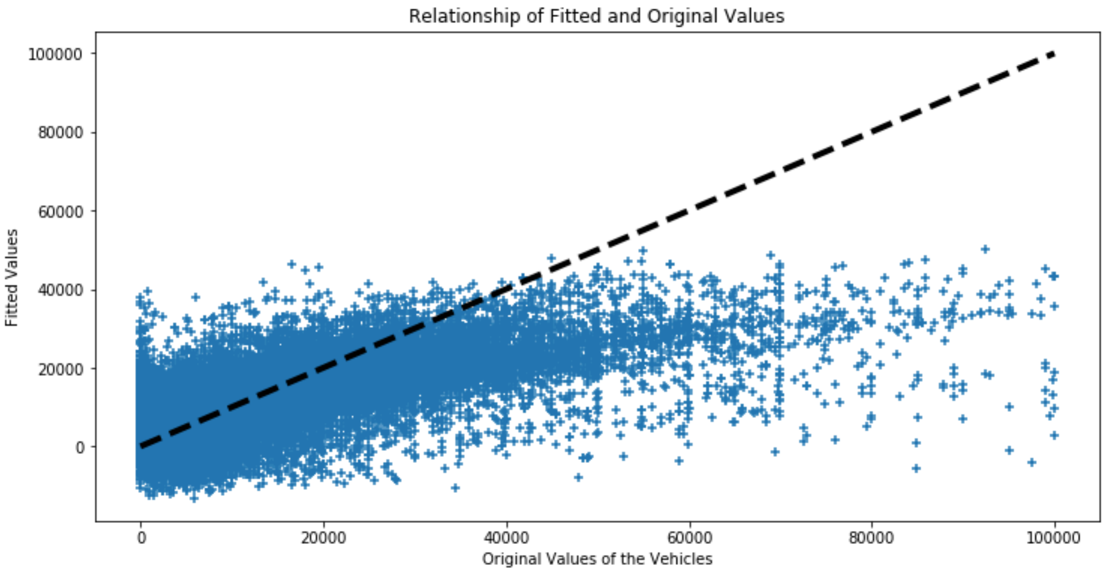
The mean squared error is a measure of the averages of the squares of the errors or deviations. In other words, the difference between the estimator and what is estimated. It is always non-negative, and values closer to zero are better. The MSE is useful for comparing different regression models or for tuning their parameters via grid search and cross-validation.

Let’s examine the output of the linear regression model from statsmodels:



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.608
Model:                            OLS   Adj. R-squared:                  0.608
Method:                 Least Squares   F-statistic:                 7.747e+04
Date:                Mon, 23 Apr 2018   Prob (F-statistic):               0.00
Time:                        11:25:00   Log-Likelihood:            -2.4845e+06
No. Observations:              250160   AIC:                         4.969e+06
Df Residuals:                  250154   BIC:                         4.969e+06
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept           1.066e+06   2.05e+06      0.520      0.603   -2.95e+06    5.08e+06
yearOfRegistration  -523.7346   1016.222     -0.515      0.606   -2515.503    1468.034
powerPS               69.0871      0.168    412.380      0.000      68.759      69.415
kilometer             -0.0697      0.000   -246.791      0.000      -0.070      -0.069
postalCode             0.0061      0.000     15.876      0.000       0.005       0.007
Age                 -840.6466   1016.217     -0.827      0.408   -2832.406    1151.113
==============================================================================
Omnibus:                   189475.520   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         10057545.910
Skew:                           3.165   Prob(JB):                         0.00
Kurtosis:                      33.411   Cond. No.                     2.87e+10
==============================================================================
```

The results show a low R-squared of 0.608 which indicates a weak goodness-of-fit. This model also produces a MSE of 25410123.9382. The MSE by itself doesn't tell how accurate the model is. We will need to calculate the MSE of each of the models to compare which is better.



It seems that there is no positive linear correlation between fitted values and original prices. Besides this, at most, the model tends to fit the original price poorly. So I will try other models in addition Linear Regression.

Let's compare the R-squared and MSE of each of the models:

Model	**R-squared**	**MSE**
Linear	0.604	25410123.9382
Ridge	0.600	25682187.4815
Lasso	0.604	25417144.6858

As we can see, Linear and Lasso models produce the same R-squared, however, the MSE for Linear regression is the lowest.

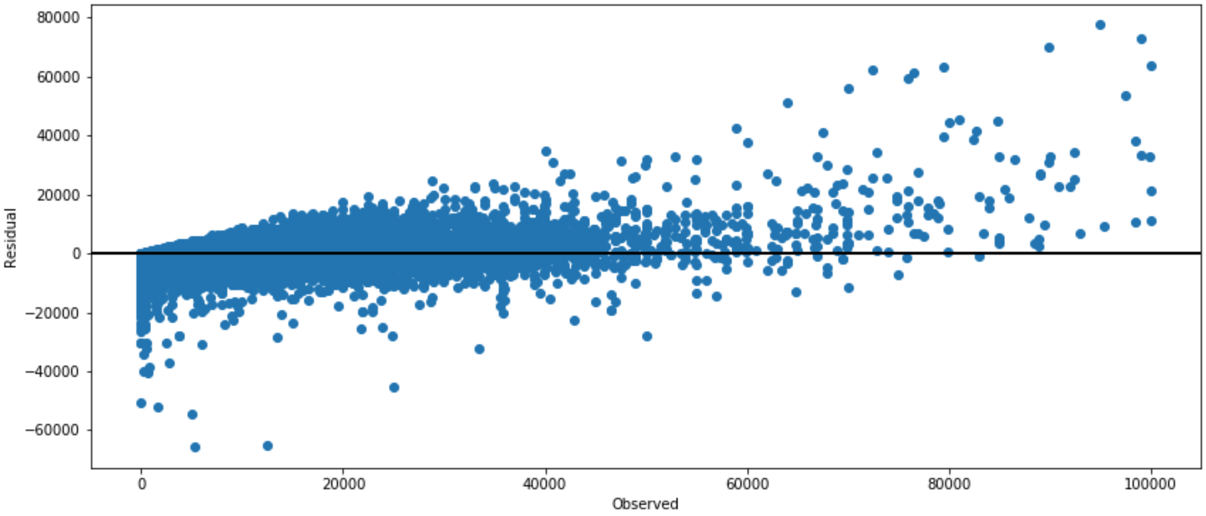
Model	**R-squared**	**MSE**
Linear	0.736	16922125.1335
Ridge	0.732	17242387.8353
Lasso	0.73	17427420.7108

Encoding the non-numerical values with categorical ones, adding to the model provided us to get better results.

5. RANDOM FOREST REGRESSION ANALYSIS

Random Forest Regression algorithm worked great on this data set. Without the categorical features, it scored 0.80 of R-squared, and after including the categorical features, the score rose up to 0.90. This score looked like one of the best scores.

Here the residual plot is shown below.



6. CONCLUSION

The analysis sought to improve the accuracy of the predicting of the values of Used Vehicles in order to better make estimations in terms for both the seller and the buyers including the dealers. We achieved this by building a statistical model utilizing several features that are identified as being correlated with the value of the vehicle.

I first trained three Linear Regression models (Linear Regression, Ridge and Lasso) with Numeric Features and then applied the same models by transforming the non-numeric features to categorical values, and compared their performance using R-squared and Mean Squared Error. I first got R-squared score of 0.608 with numeric features, and that of 0.73 with adding categorical values. All three models achieved almost similar R-squared, and MSE. Finally, I applied cross-validation with the same models, but this didn't improve the model performance.

And then, I considered how to get a better score by exploring other models like Decision Tree, and Random Forest. Random Forest with only numeric features gave me a score of 0.80, 0.899 with including non-numeric features. This achieved the best score ever on this data set.

7. Future Work

As a future scope, the analysis could include data from all around the year. In this way, we could see the ups and downs of seasonality and impact of the other periods of the year. We could also add a variable which would indicate the number of clicks to that individual postings. This would imply to the other potential buyers that specific vehicle had eyes on it, so it is valuable compared to others. Besides, we could add variables which would show the accessories the vehicle posted.

In terms of better performance of the model, I could better hyper tune the model, and then delete some features to get a better result. Last but not least, I could explore how to build this model with deep learning algorithms.