Halil AKSU

MS: University of Southern California, Operations Research

Predicting the Value of a Used Vehicle

May 2nd, 2018

- Personel Background:
- 16 years of experience on management and analysis
- Additional background on international relations and public affairs
- Valuable experience on logistics and personnel management
- BS degree on Industrial Engineering (graduated 3rd out of 242)(3.84)
- MS degree on Operations Research (3.79)
- MA degree on Leadership and Management (graduated 3rd place)
- Numerous presentations before various VIP audience

Predicting the Value of a Used Vehicle

Data Science Career Track Capstone Project, February 05th 2018 Cohort

Springboard

Problem Introduction

Who might care?

Individual Buyers

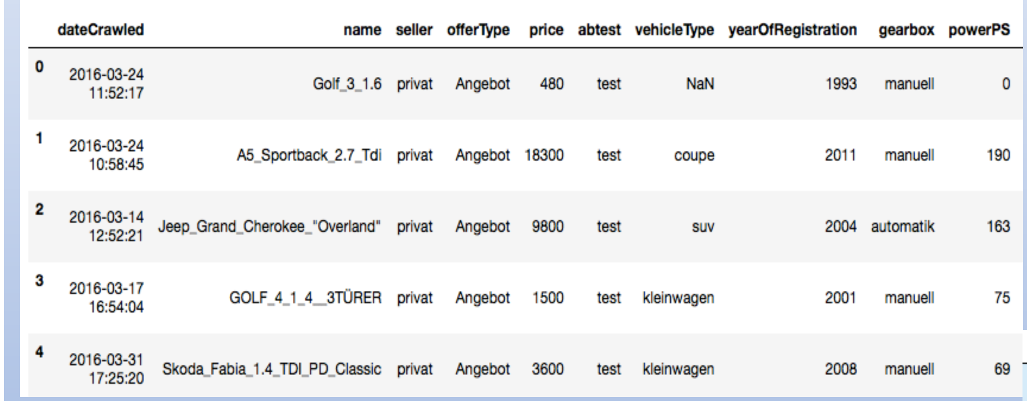
Dealers

Websites/ Mobile Apps

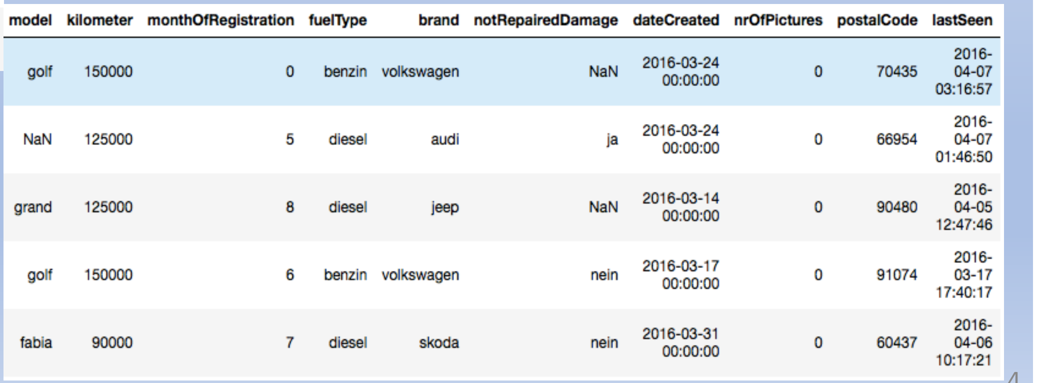
craigslist

Data Set :

- From Kaggle
- 19 features and 371528 data points/ observations
- Target Feature is ‘price’



	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



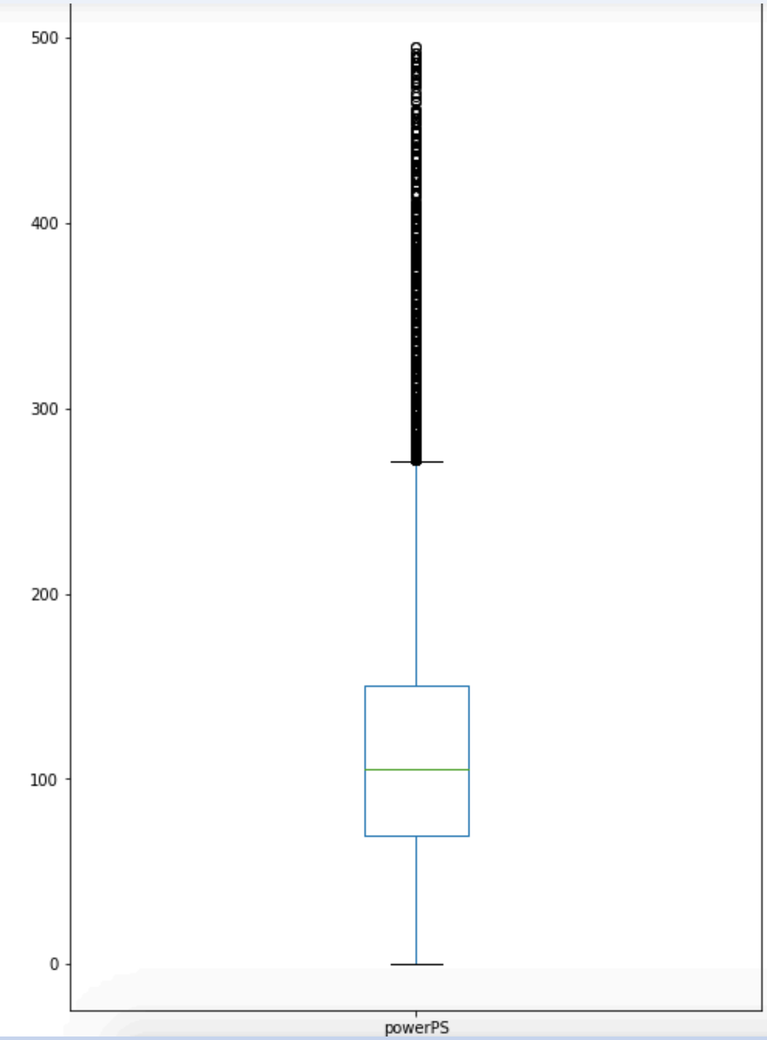
model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage	dateCreated	nrOfPictures	postalCode	lastSeen
golf	150000	0	benzin	volkswagen	NaN	2016-03-24 00:00:00	0	70435	2016-04-07 03:16:57
NaN	125000	5	diesel	audi	ja	2016-03-24 00:00:00	0	66954	2016-04-07 01:46:50
grand	125000	8	diesel	jeep	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

Data Exploration :

- ✓ 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!!!

Filtering Parameters:

After examining the samples with horsepower higher than 600 and lower than 5, it is more likely to get rid of these as outliers.

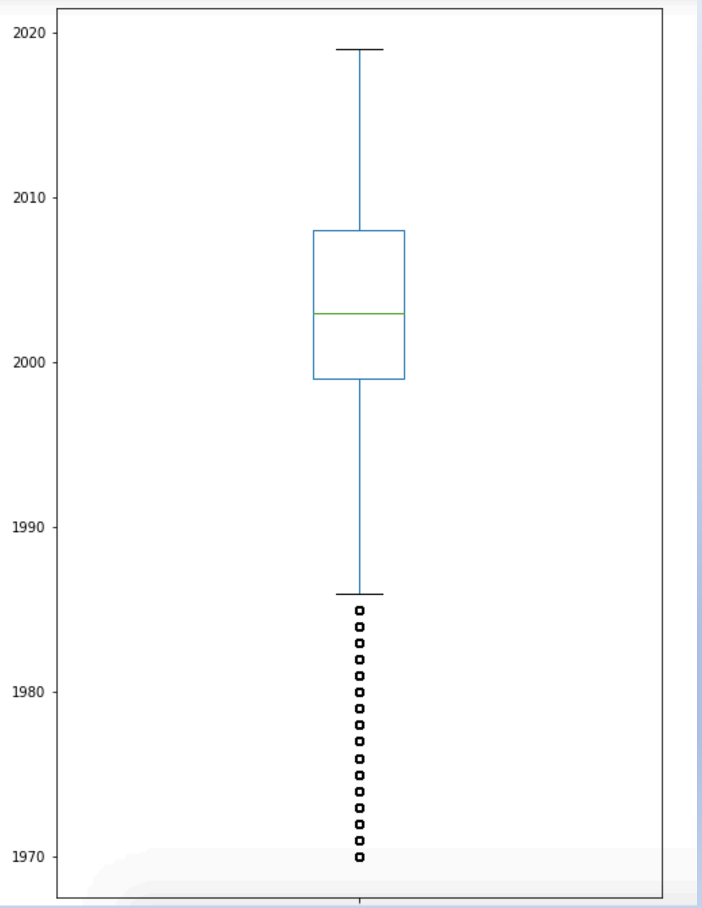


Filtering Parameters:

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.



Handling Missing Values:

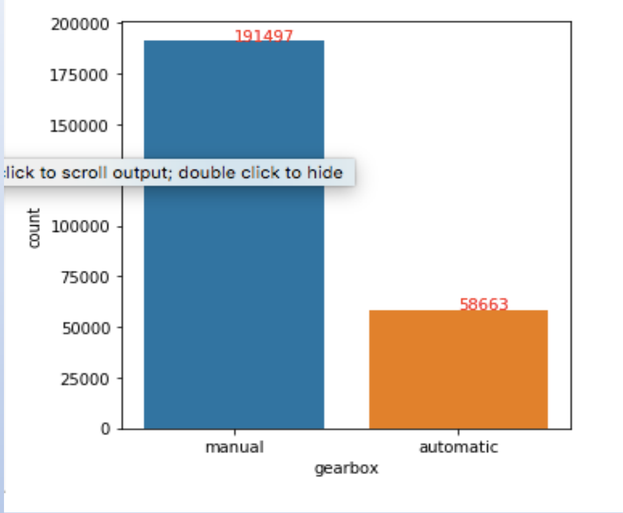
- It may mislead to keep the missing values by replacing with an interpolation for this data set. So considering the richness of the number of samples, I am opting for getting rid of the missing of values.

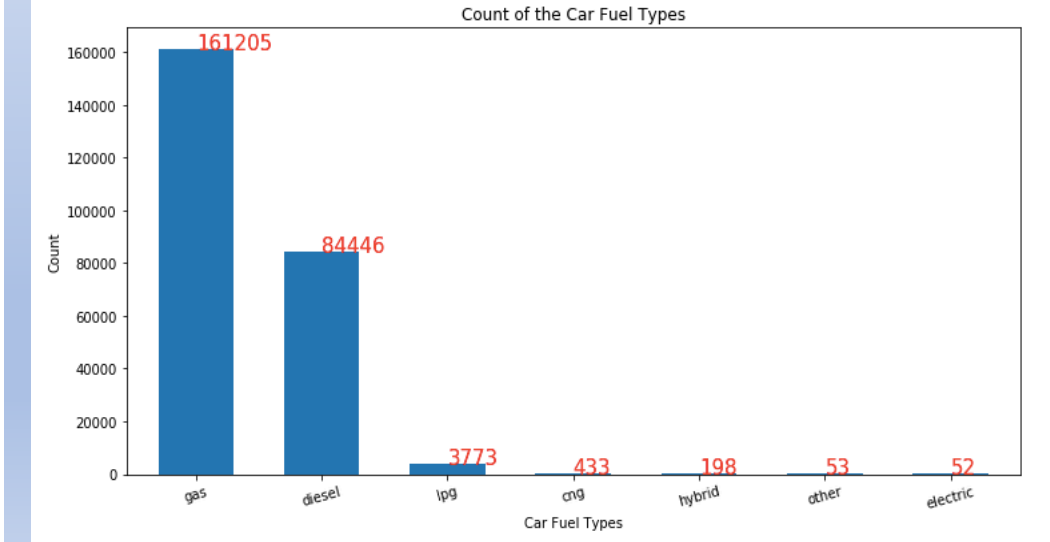
- The percentage of samples/observations lost after deleting missing values is % 30.

- The shape of the data set before the data wrangling is 371528x19, and its shape became 250160x16 after.

Data Visualization:

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

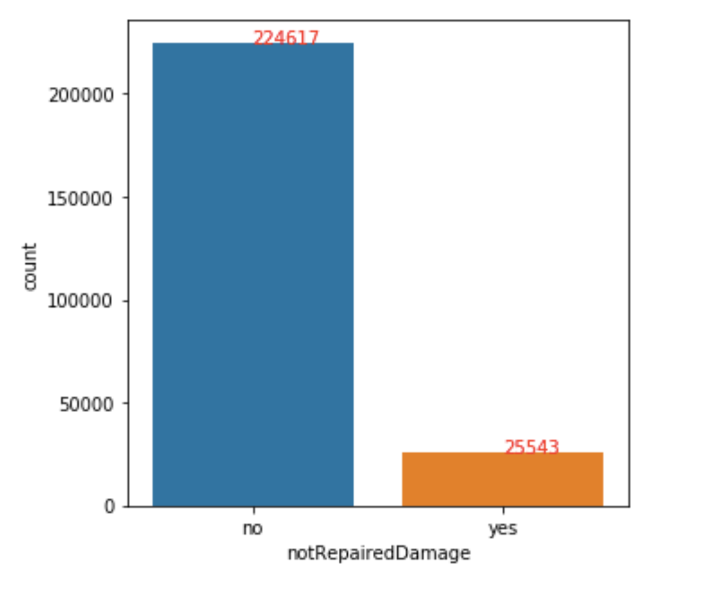


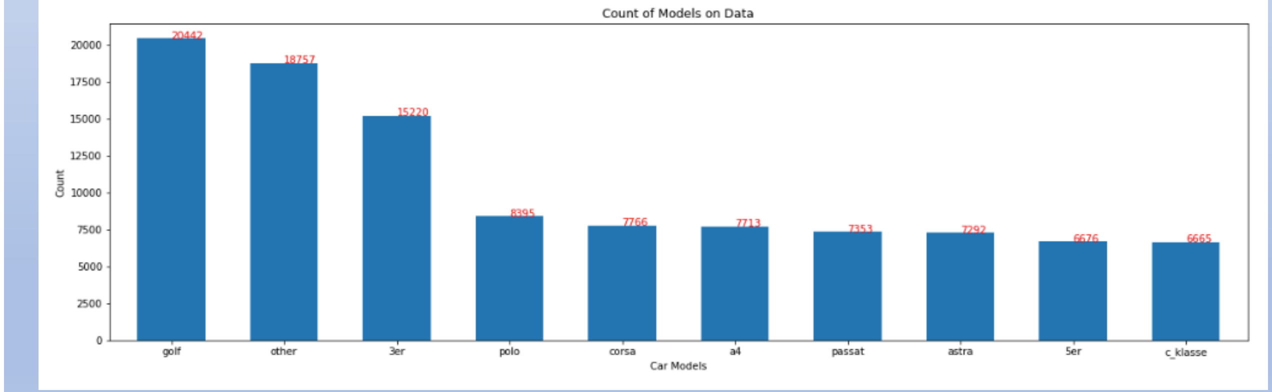


Number of vehicles operated by gas is 161.205, whereas the number of vehicles operated by diesel fueltype is 84.446. Diesel operated vehicles are very popular in Europe. In addition, liquid petroleum gas (LPG), compressed natural gas(CNG) operated vehicles exist in spite of the of the fact that their number is low.

Data Visualization:

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.

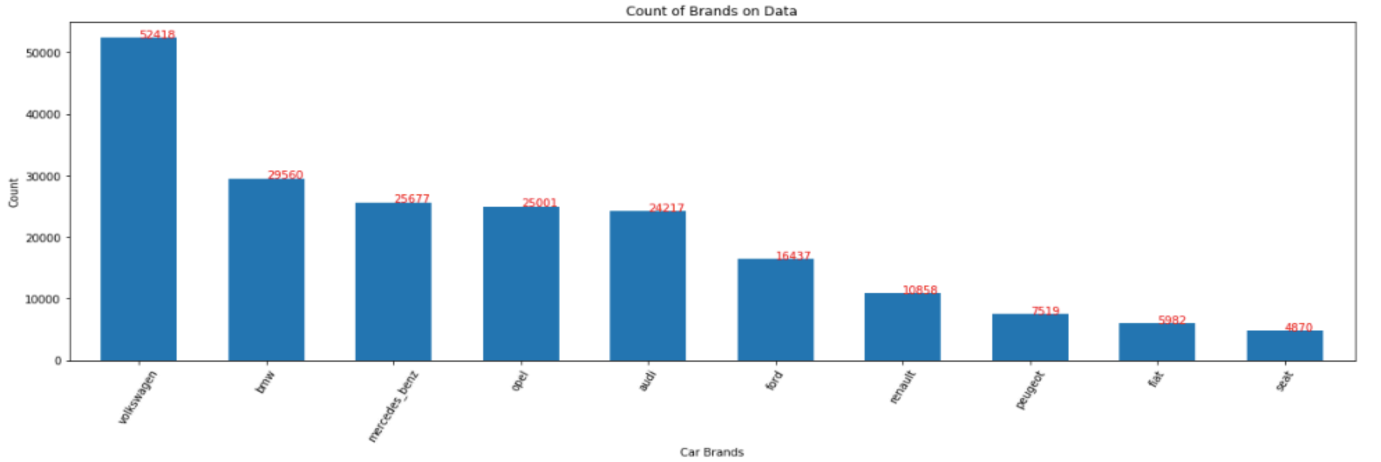


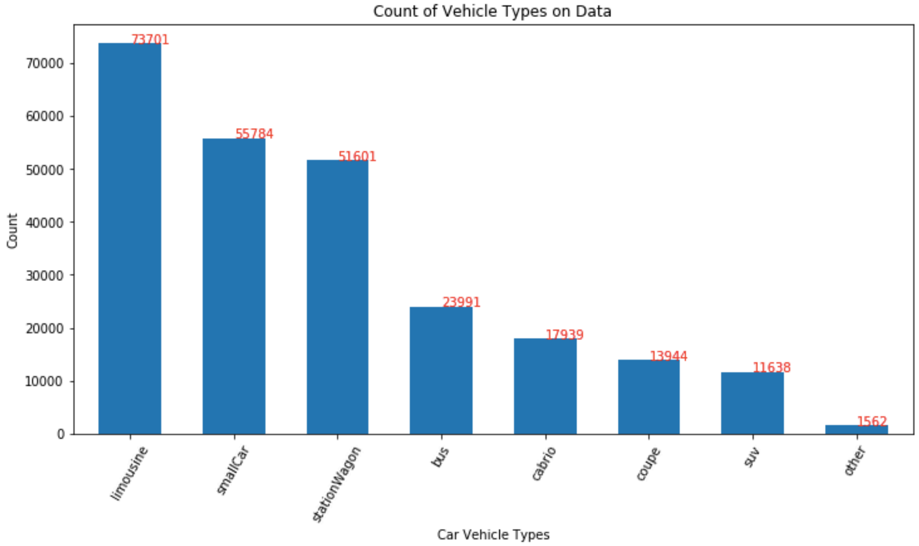


Golf, Polo, Corsa, Passat are the most common models in Europe.

Data Visualization:

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

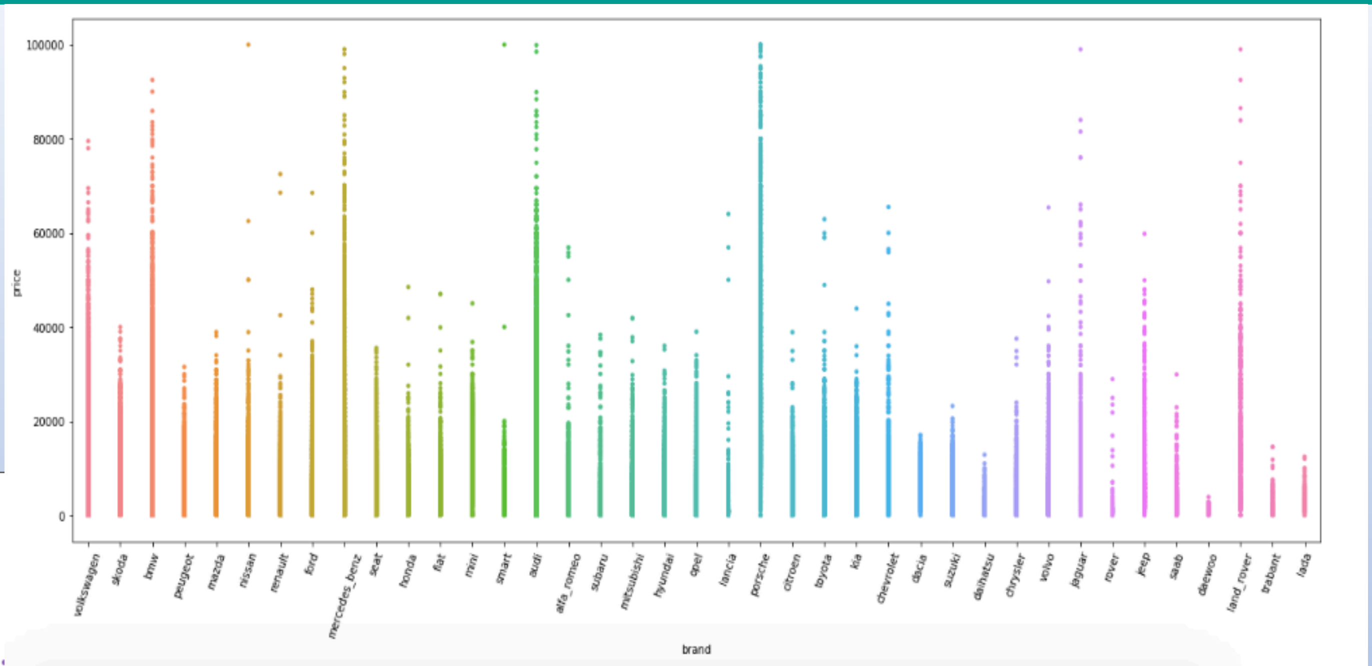


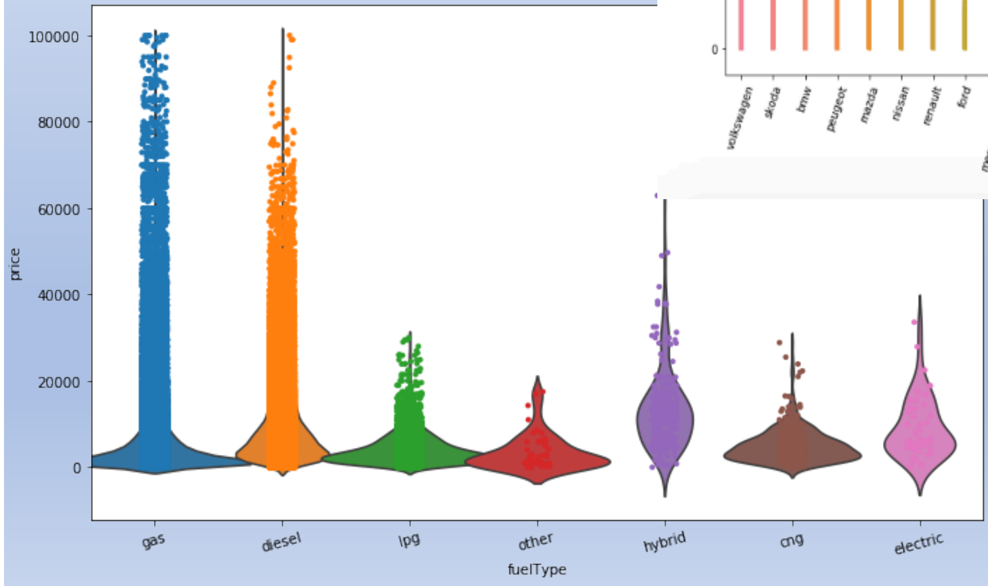


Number of Limousine (Sedan) vehicles is 73701, wheres that of SUVs is 11638.

Data Visualization:

It is easy to examine from this chart that Mercedes_benz, Audi, Porsche, BMW, Land Rover are the most expensive vehicles

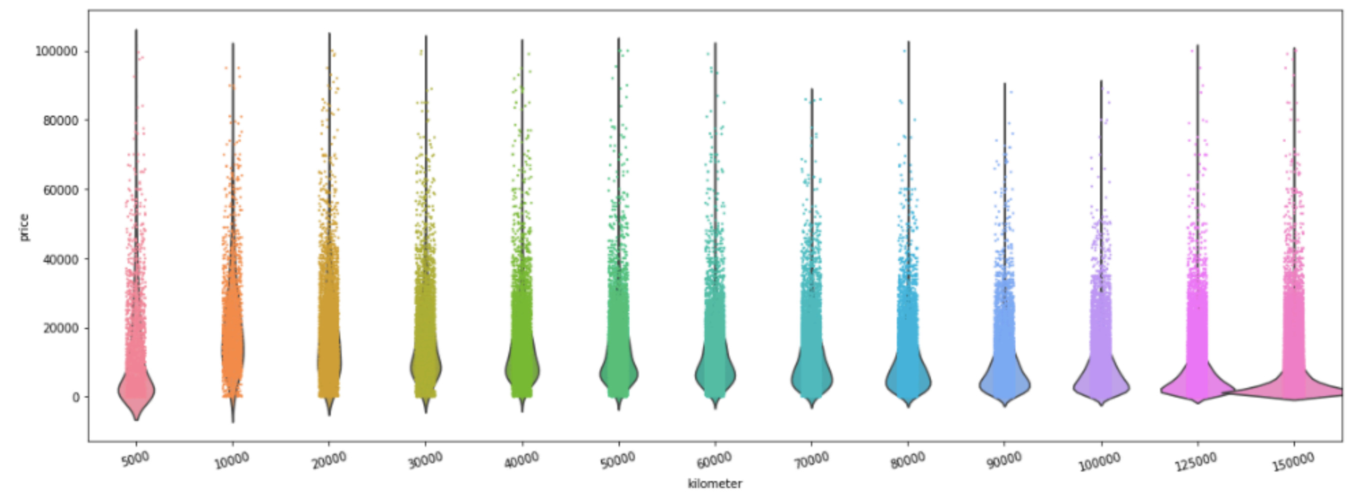


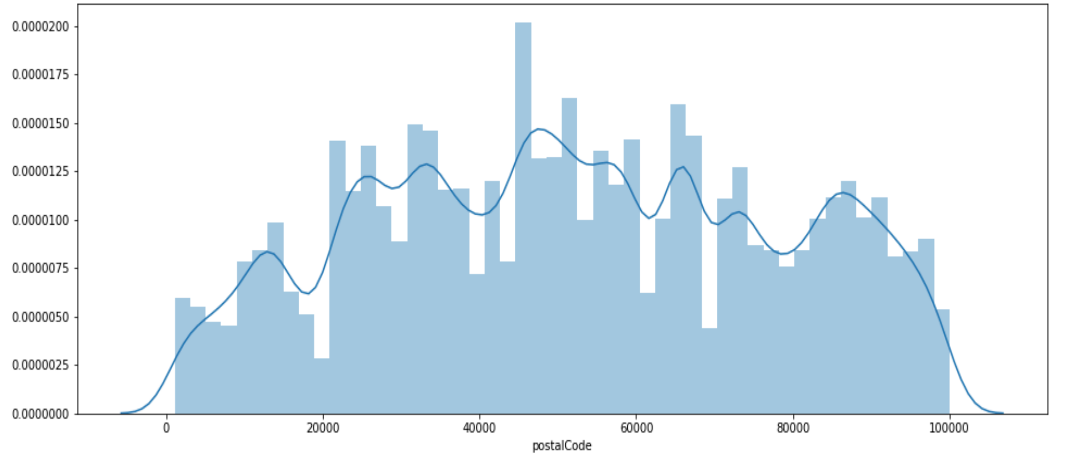


The most expensive vehicles are usually of the fueltype , gas however, there are some number of diesel operated vehicles with a high value. This chart indicates that most hybrid and electric vehicles are not of high values although they are the new generation vehicles.

Data Visualization:

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).

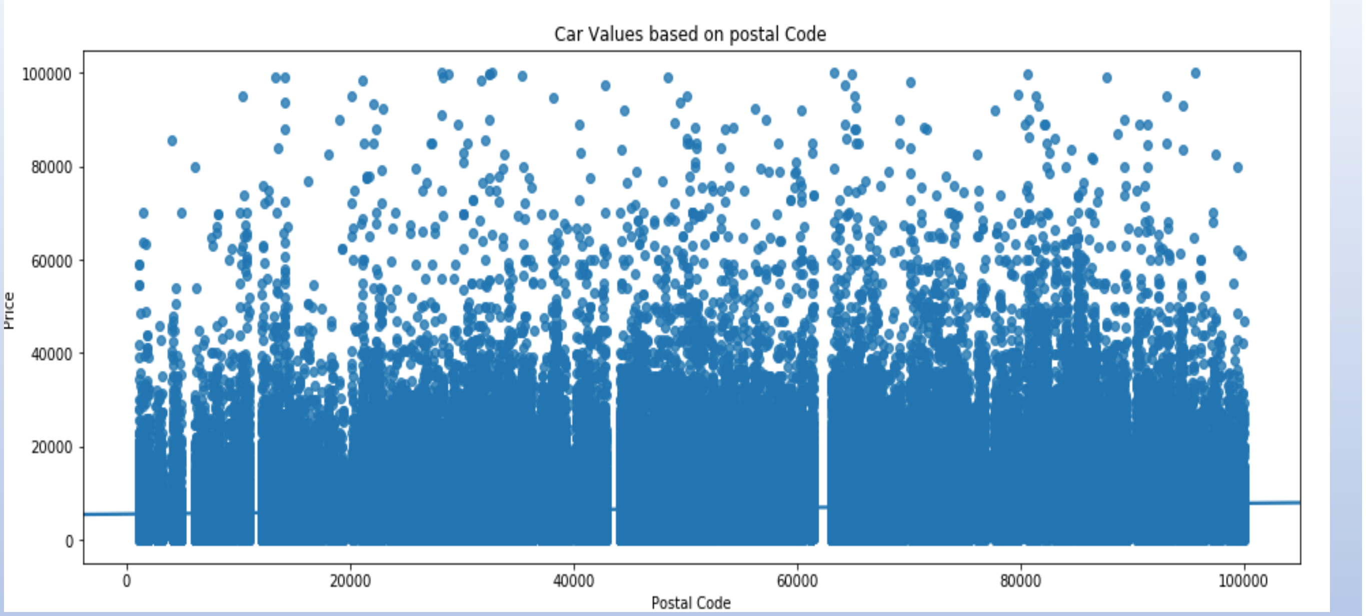


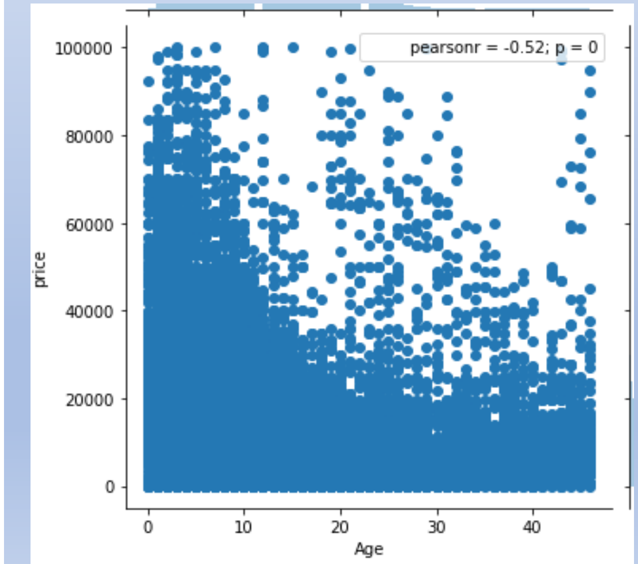


Postal code of '45000' is the town that the highest number of cars are loaded into the website for sale.

Data Visualization:

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.

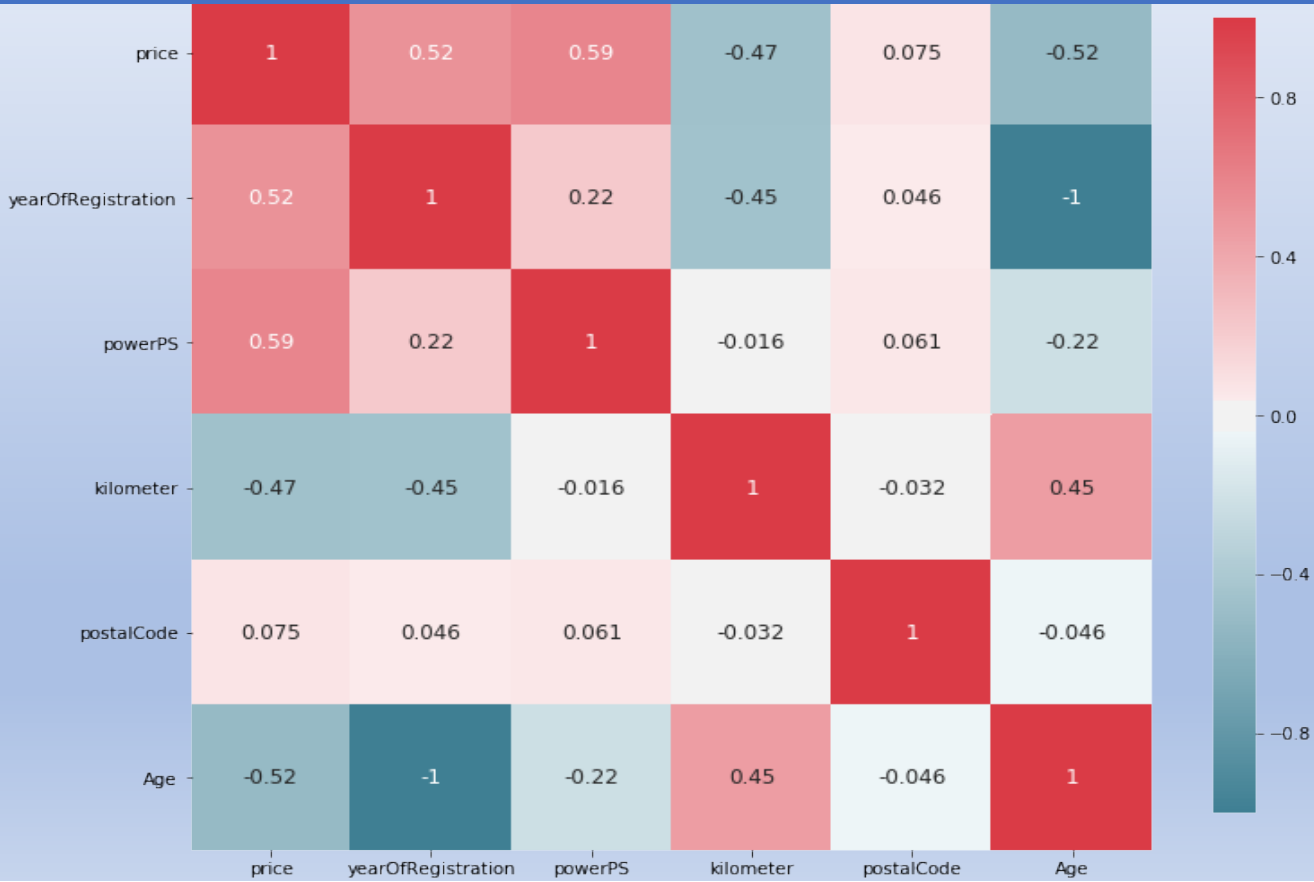




It is clear that the vehicles until the age of 10 have the highest values although some old vehicles with a high value exist. It is easy to interpret these vehicles are classic vehicles.

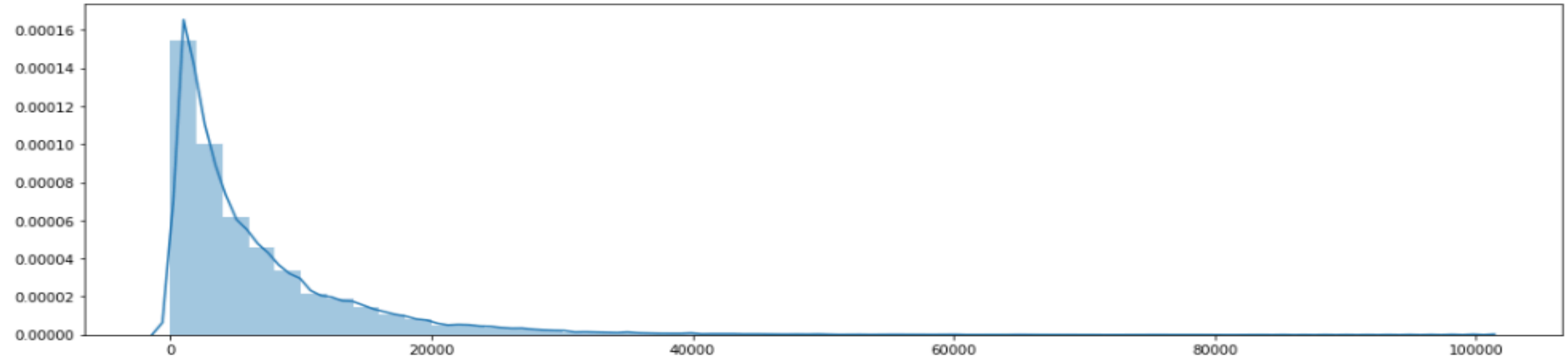
Data Visualization:

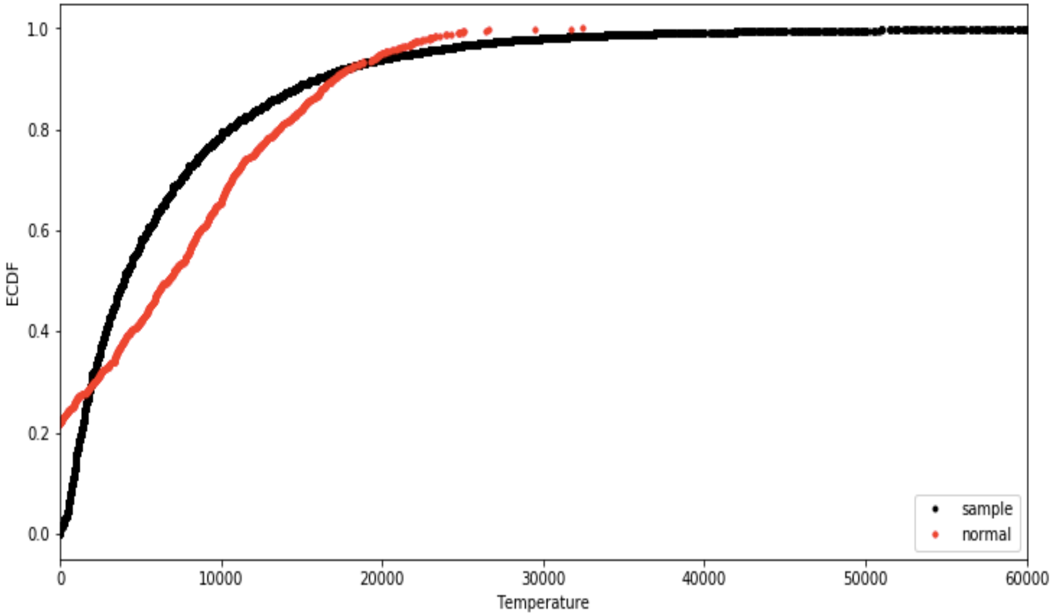
From the chart above, we can interpret that ‘yearOfRegistration/Age (same indicator)’ features have a positive correlation of 0.52 with ‘price’ feature. ‘powerPS’ feature which indicates the horsepower of the vehicles have a positive correlation of 0.59 with ‘price’ feature. ‘PostalCode’ has a slight correlation with ‘price’ feature.



	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

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.

Predictive Modeling

Modeling Overview

Type: Supervised learning

Continuous Numerical Value

Tools: Python’s scikit learn(main), stat-models, scipy

Model Assumptions, Limitations and Disclaimers

- Assume that all postings are independent

- Used the data only from 2016 (Future Work: Should be expanded)

- Utilized numerical features first, and then included non-numerical features with get_dummies function

- Tried Scaling, though it didn’t improve the model performance.

Regression:

Data preparation and selection

1. Get_dummies of all categorical variables
2. Randomly select a small subset of the whole data **(30%)**

Cross validation for model assessment and selection

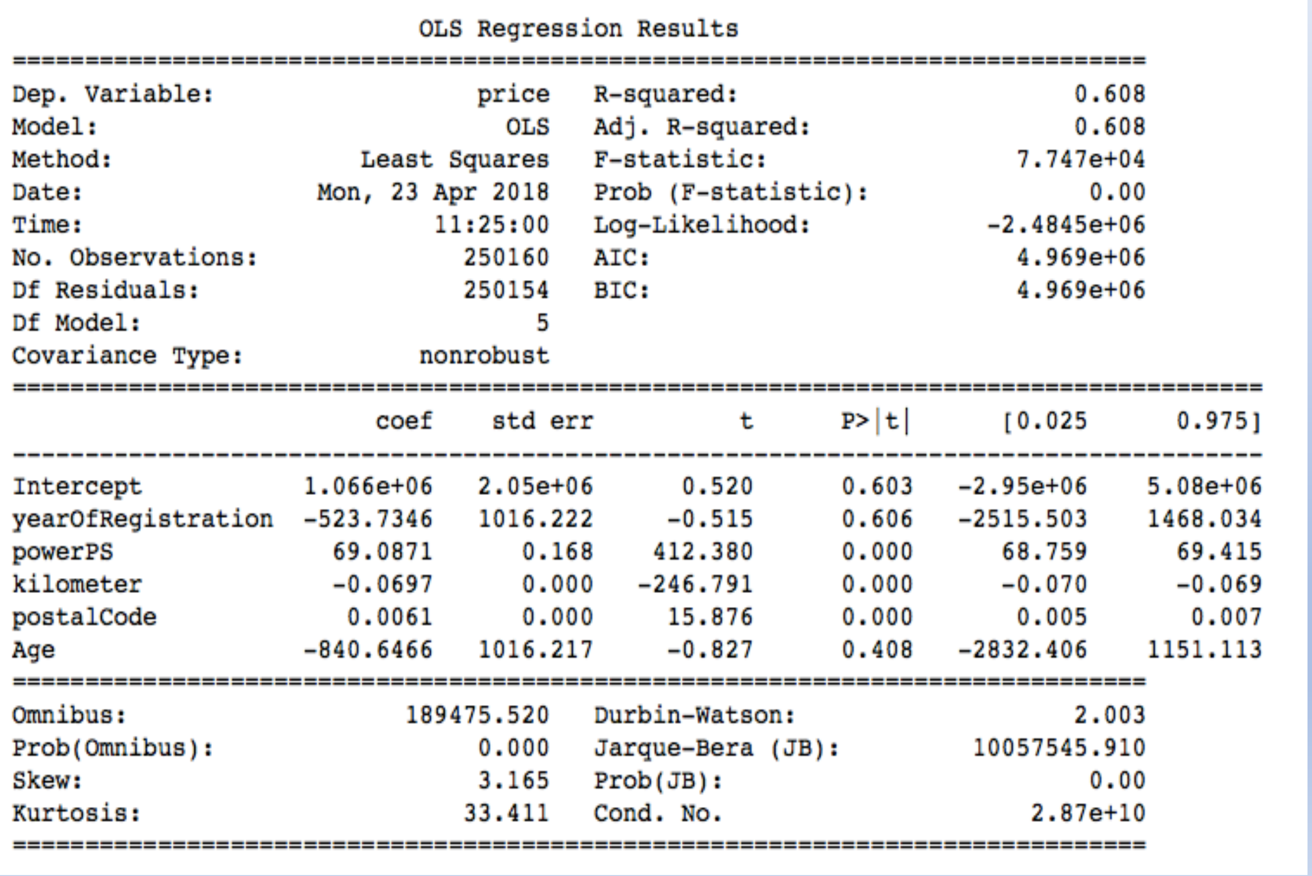
- 5-fold cross validation
- At each iteration, perform the data pre-processing and train the classifier using data contained in 4-folds
- Evaluate the model using the data contained in remaining one fold

Comparison:

1. Compared R-Squared and MeanSquared Error all scores

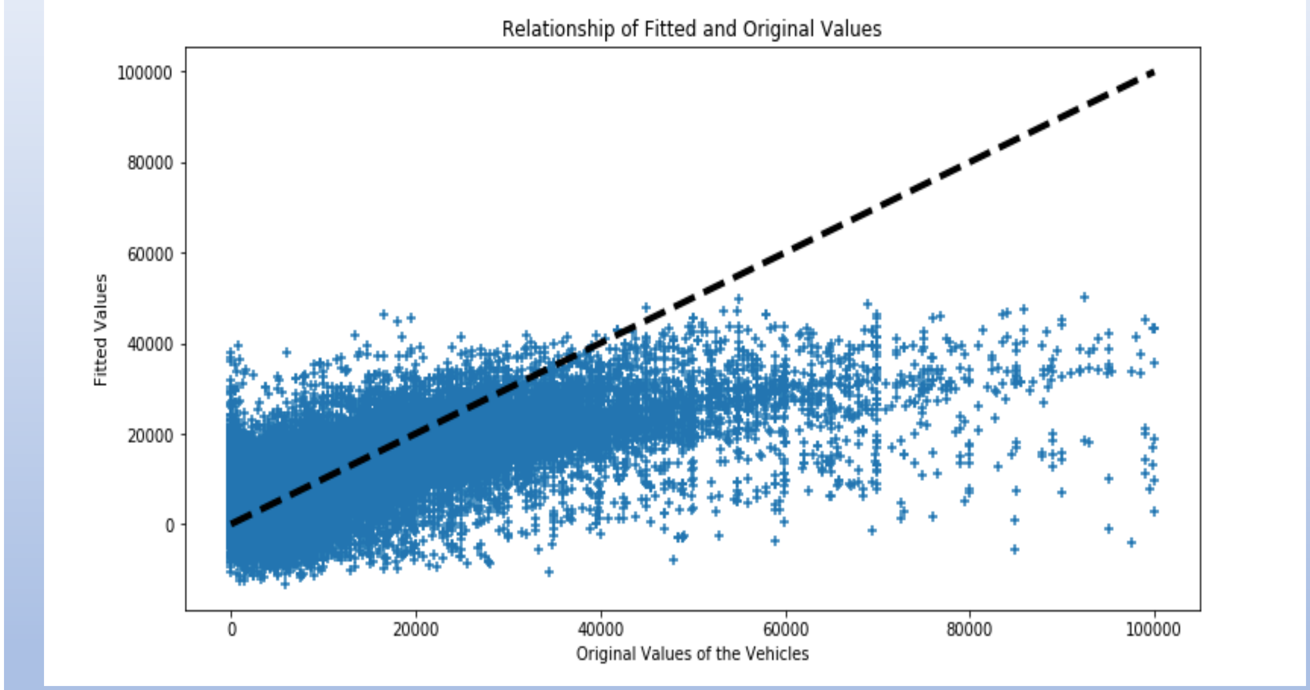
Regression:

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.



```
                            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
==============================================================================
```

Regression:



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.

Regression:

With only Numeric Features

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

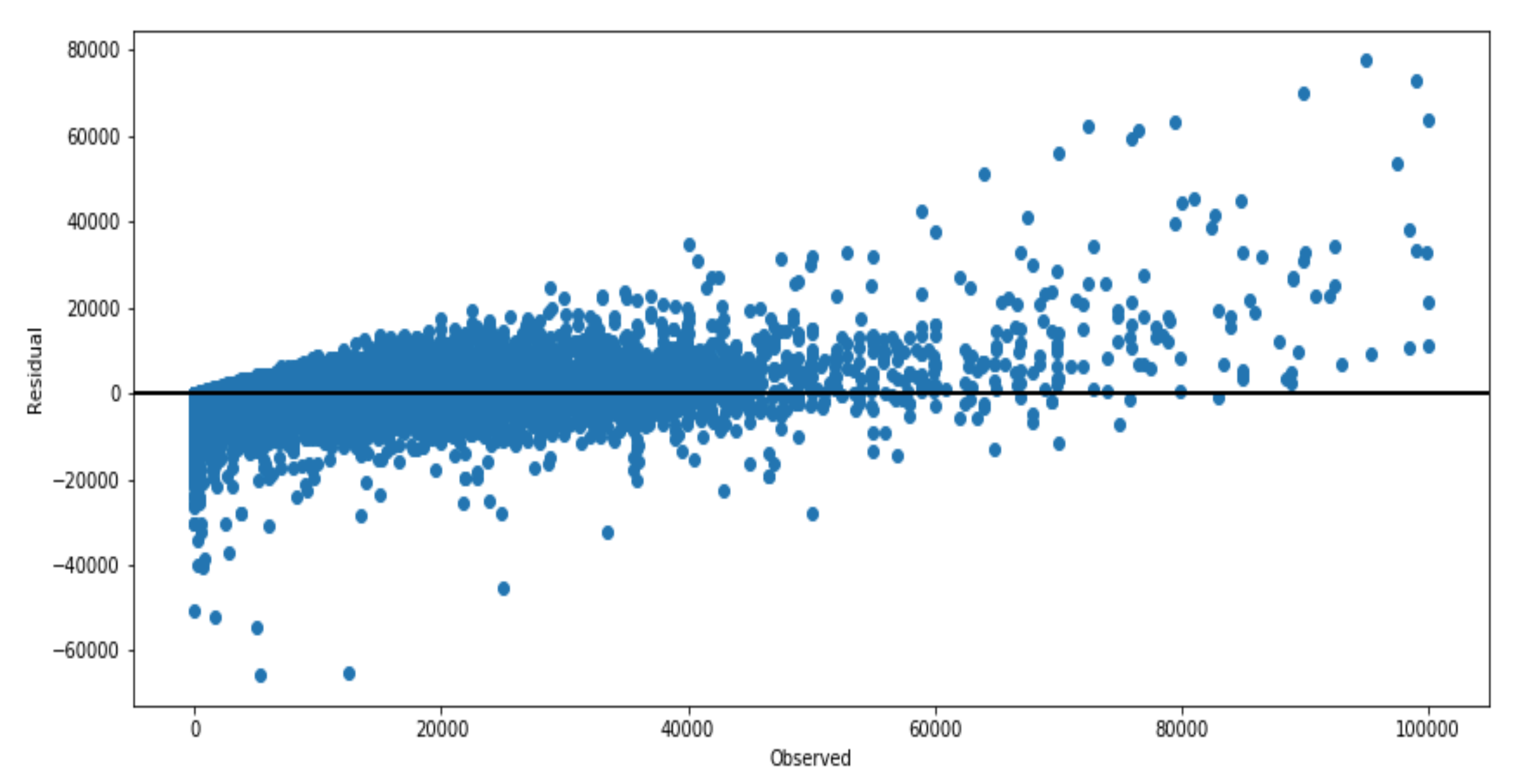
Including also Non-Numeric Features

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

Regression:

Model	R-squared	MSE
Random Forest with Numeric	0.804160991899	12592910.4069
Random Forest with all	0.899070707067	6489991.73177

Residual Plot :



More Ideas to Improve Model in Future

- Engineer more features related with Number of Clicks to each Posting, etc.
- Extract more data from all year around.
- Use other Algorithms/Models to get better results.

Conclusions

- All features of dataset contributed to the predictive power of the model
- Out of 4 supervised regression models, the Random Forest Regressor provided the best result: R-Squared=0.899 and MSE=
- With more ideas, the model can improve in the future

Thank you!

Halil Aksu
Email: halilaksu79@gmail.com
https://www.linkedin.com/in/halil-aksu-38bab6129/
https://github.com/CoderModer79