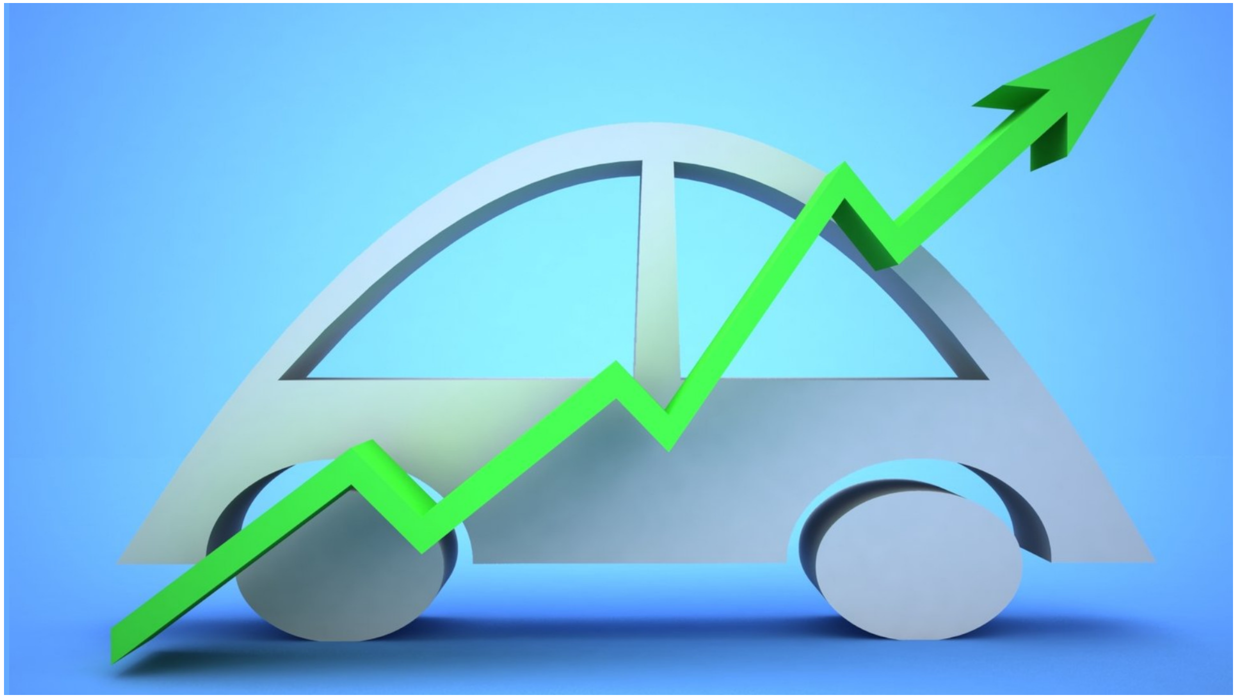


Kagan ILTER

Predicting the Price of a Used Vehicle

Springboard Data Science Career Track

Capstone Project

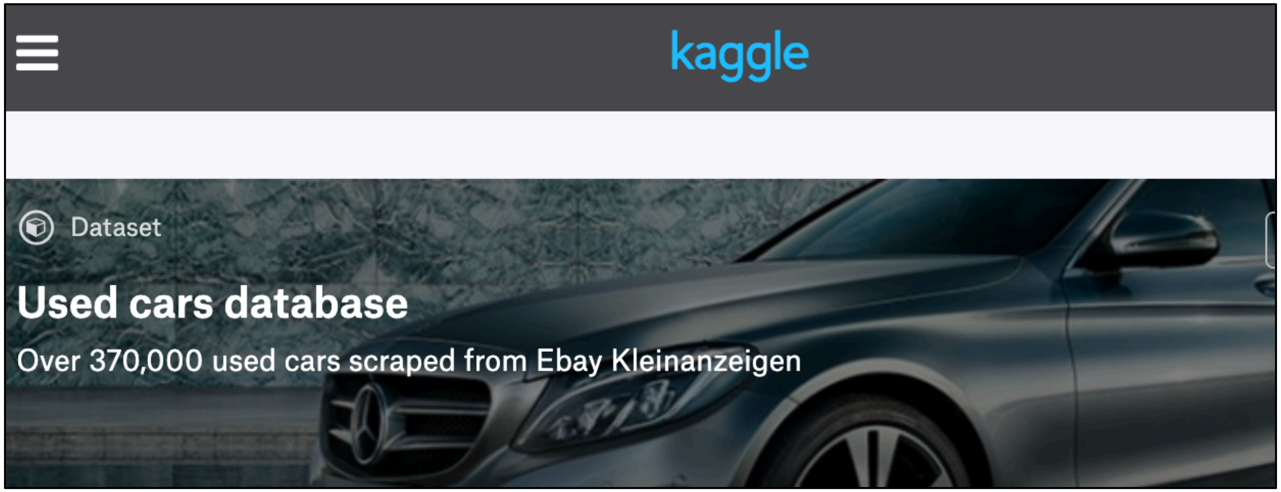
Problem Definition

The scope of this project is to predict the price of a used vehicle based on its features.

INDIVIDUAL BUYERS	DEALERS	INDIVIDUAL SELLERS	WEBSITES / APPLICATIONS

Data Information

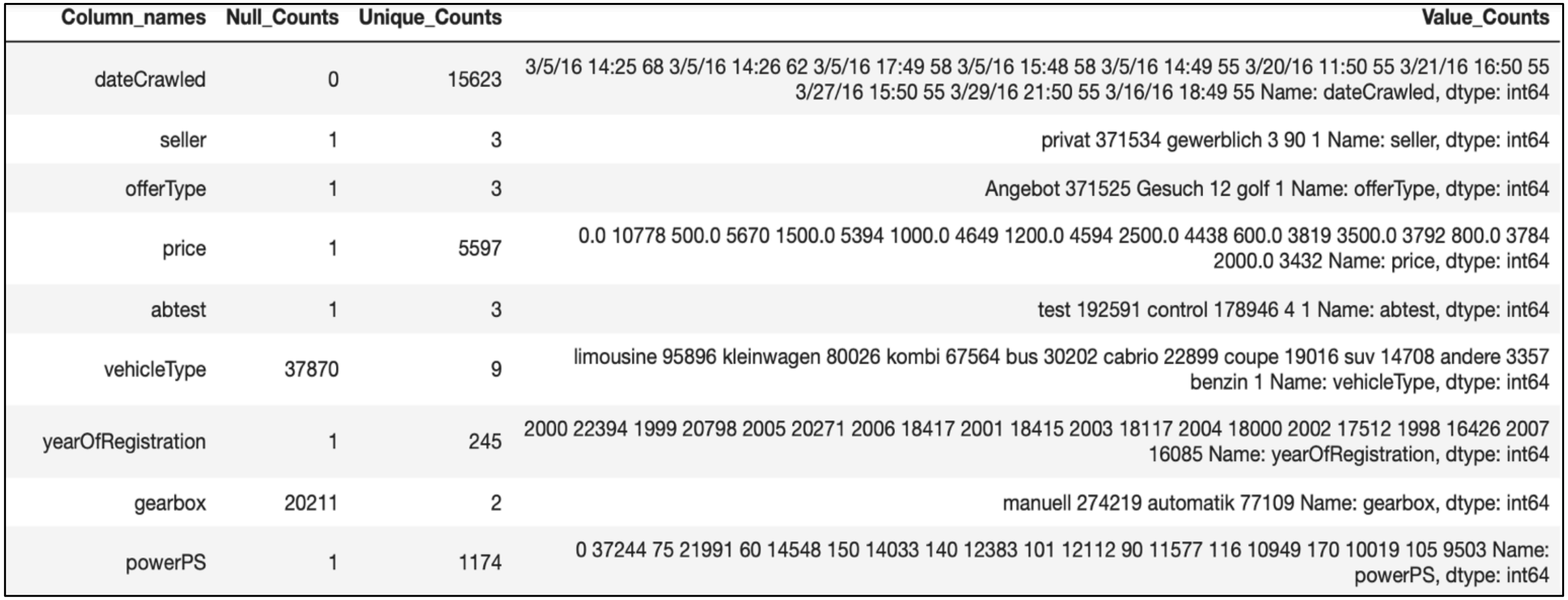
- The data was scraped with from the Ebay. Dataset is acquired from Kaggle
- 19 features and 371528 data points/ observations.
- Data points/observations from 2016.



Used cars database

Over 370,000 used cars scraped from Ebay Kleinanzeigen

Data Exploration



Column_names	Null_Counts	Unique_Counts	Value_Counts
dateCrawled	0	15623	3/5/16 14:25 68 3/5/16 14:26 62 3/5/16 17:49 58 3/5/16 15:48 58 3/5/16 14:49 55 3/20/16 11:50 55 3/21/16 16:50 55 3/27/16 15:50 55 3/29/16 21:50 55 3/16/16 18:49 55 Name: dateCrawled, dtype: int64
seller	1	3	privat 371534 gewerblich 3 90 1 Name: seller, dtype: int64
offerType	1	3	Angebot 371525 Gesuch 12 golf 1 Name: offerType, dtype: int64
price	1	5597	0.0 10778 500.0 5670 1500.0 5394 1000.0 4649 1200.0 4594 2500.0 4438 600.0 3819 3500.0 3792 800.0 3784 2000.0 3432 Name: price, dtype: int64
abtest	1	3	test 192591 control 178946 4 1 Name: abtest, dtype: int64
vehicleType	37870	9	limousine 95896 kleinwagen 80026 kombi 67564 bus 30202 cabrio 22899 coupe 19016 suv 14708 andere 3357 benzin 1 Name: vehicleType, dtype: int64
yearOfRegistration	1	245	2000 22394 1999 20798 2005 20271 2006 18417 2001 18415 2003 18117 2004 18000 2002 17512 1998 16426 2007 16085 Name: yearOfRegistration, dtype: int64
gearbox	20211	2	manuell 274219 automatik 77109 Name: gearbox, dtype: int64
powerPS	1	1174	0 37244 75 21991 60 14548 150 14033 140 12383 101 12112 90 11577 116 10949 170 10019 105 9503 Name: powerPS, dtype: int64

- There are null values in the features: Vehicletype, model, gearbox, fueltype and notrepaireddamage

- There are also some false entries such as year of registration: 9999, powerps: 10.000 or price:0

Data Exploration

- Some features should be dropped:
  - seller : only 3 observations are dealer,
  - offerType : Only 12 of all offers are gesuch (request),
  - nrOfPictures : None of the entries have pictures,
  - monthOfRegistration : Not important

- powerPS feature have 37244 of 0 values (which is a wrong entry)

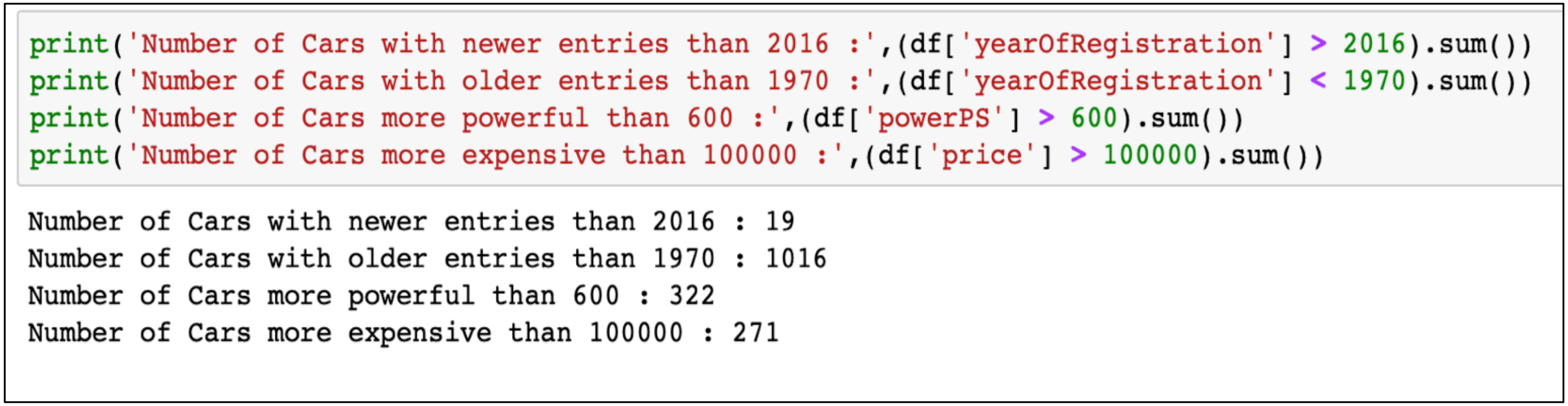
- 7 features are discrete numbers, whereas 12 features are object (string, datetime....)

- vehicleType, gearbox, model, fuelType, brand, notRepairedDamage features have missing values!!!

- Age feature is added in order to better understand the data set.

Data Wrangling

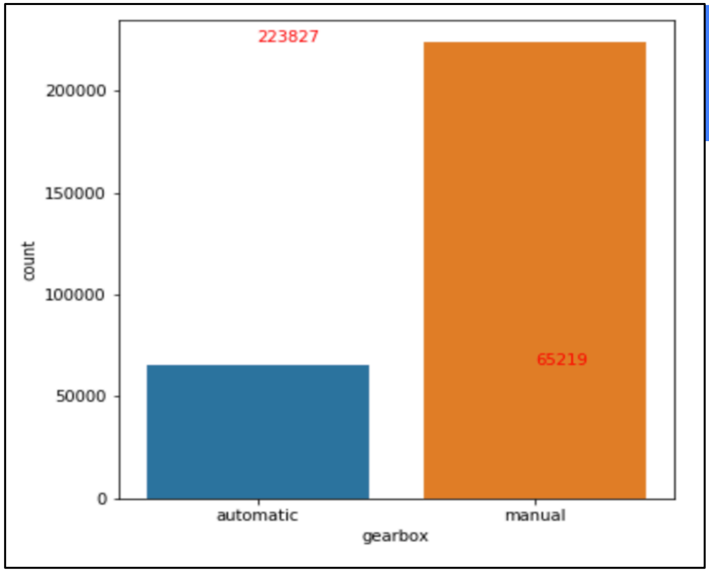
- Missing values of Fueltype, gearbox and NotRepairedDmage replaced with the most used values. (75% of the vehicles are manuel, 60% of the vehicles are gasoline, 71% of the vehicles are not damaged)
- Missing values of vehicle type and model dropped because they have balanced number of unique values.
- Nonsense data and outliers such as yearofRegistrain: 1000 or 9999, price: 0, 100000 are filtered
- The shape of the data after data wrangling is: 289046 x 14



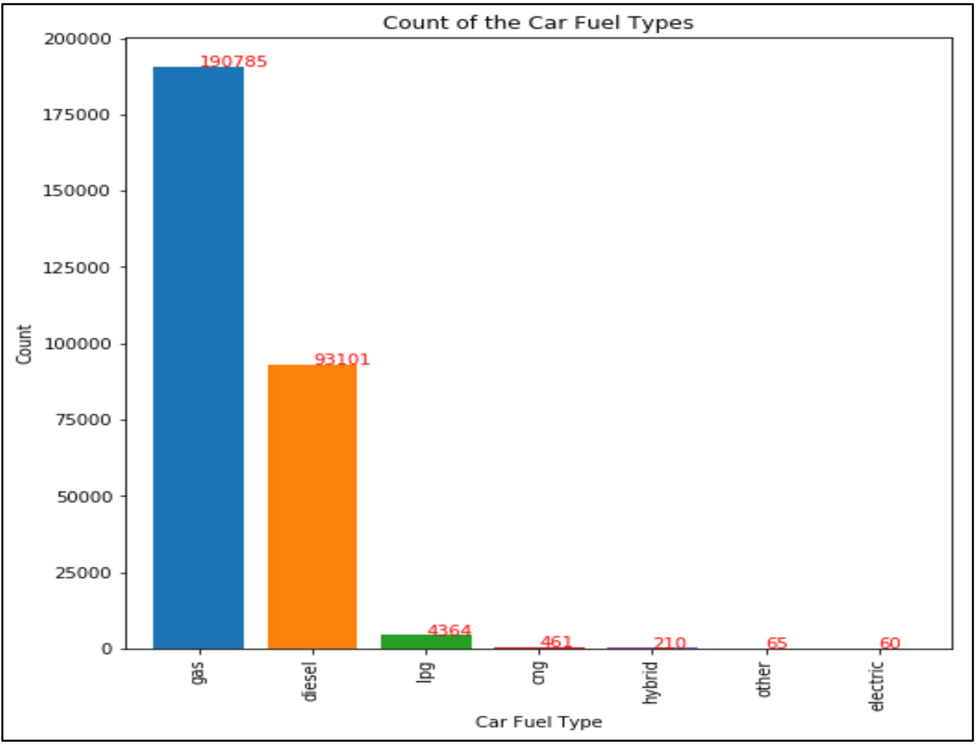
```python
print('Number of Cars with newer entries than 2016 :',(df['yearOfRegistration'] > 2016).sum())
print('Number of Cars with older entries than 1970 :',(df['yearOfRegistration'] < 1970).sum())
print('Number of Cars more powerful than 600 :',(df['powerPS'] > 600).sum())
print('Number of Cars more expensive than 100000 :',(df['price'] > 100000).sum())
```

```
Number of Cars with newer entries than 2016 : 19
Number of Cars with older entries than 1970 : 1016
Number of Cars more powerful than 600 : 322
Number of Cars more expensive than 100000 : 271
```

Data Visualization

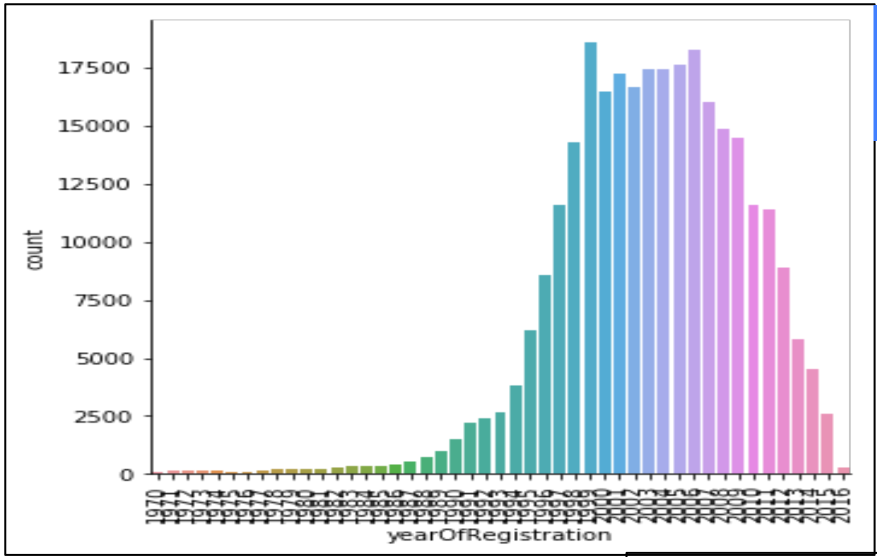


Regarding the gearbox, manual cars are more widespread

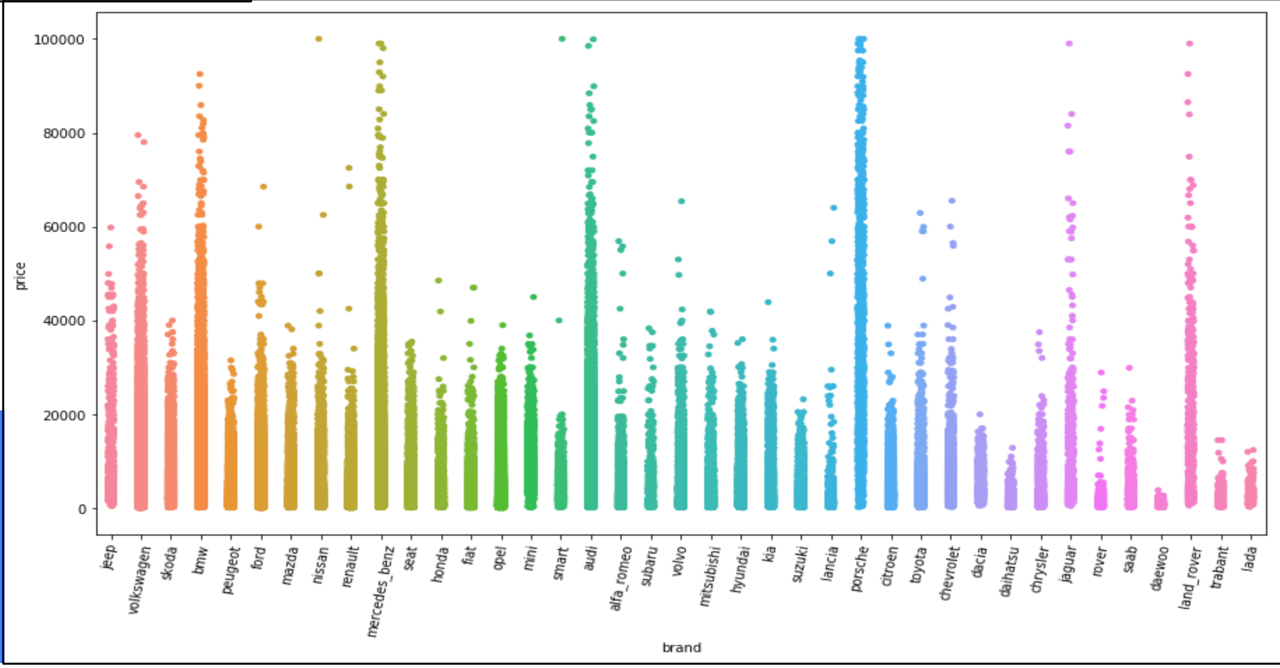


Since the gas prices in Europe are higher than the US, diesel, lpg and cng cars can be seen. However, regular gas using cars are widespread.

Data Visualization

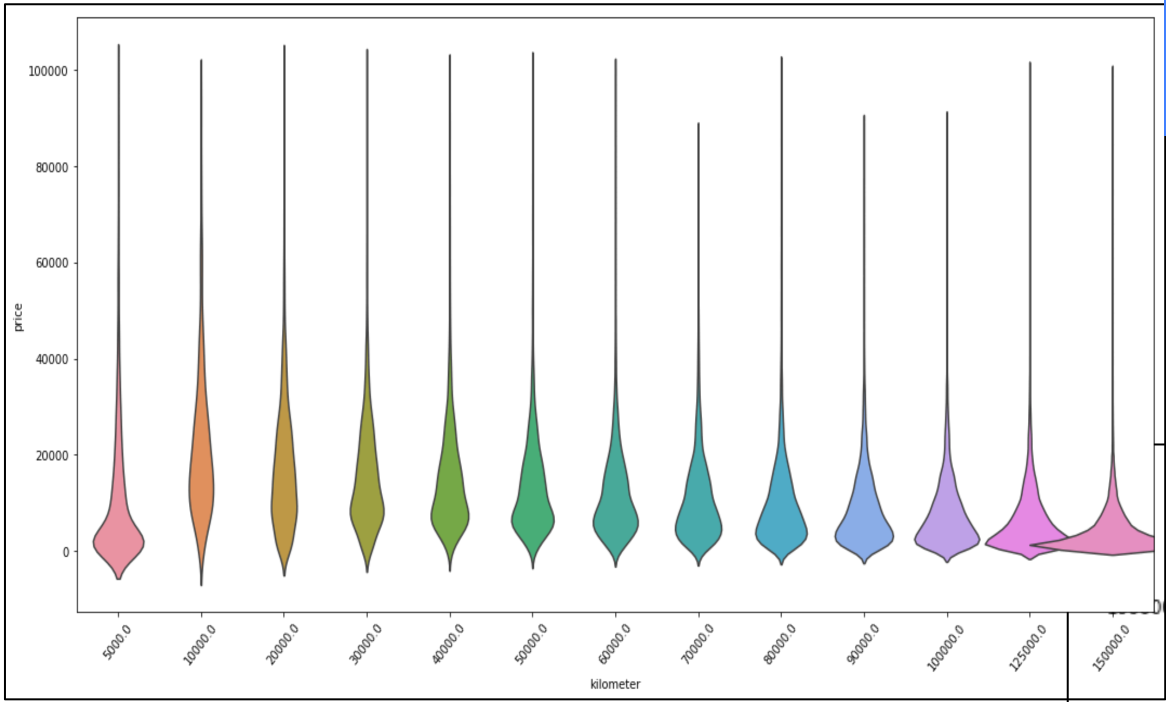


Most of the used cars are registered between 2000 and 2010

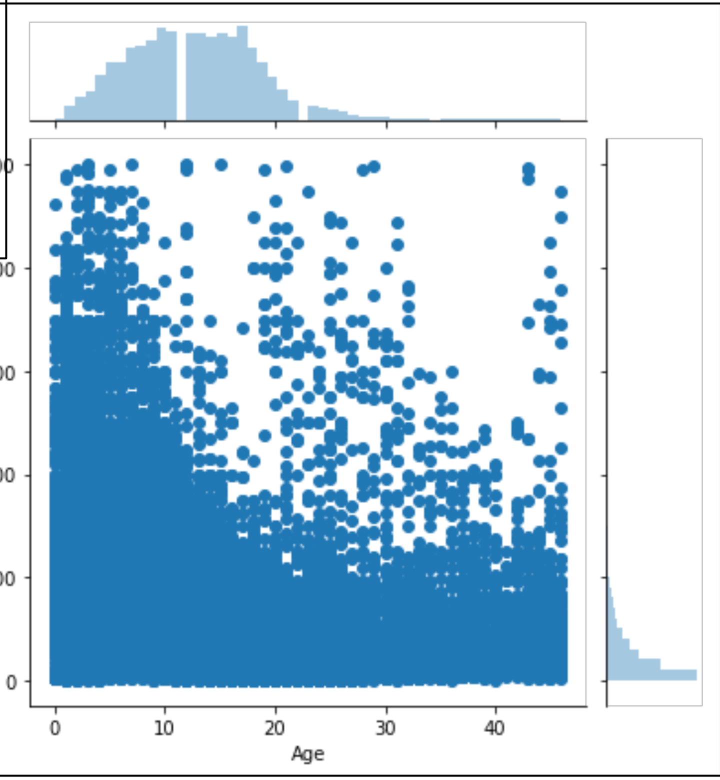


Mercedes_benz, Audi, Porsche, BMW, Land Rover are the most expensive vehicles.

Data Visualization

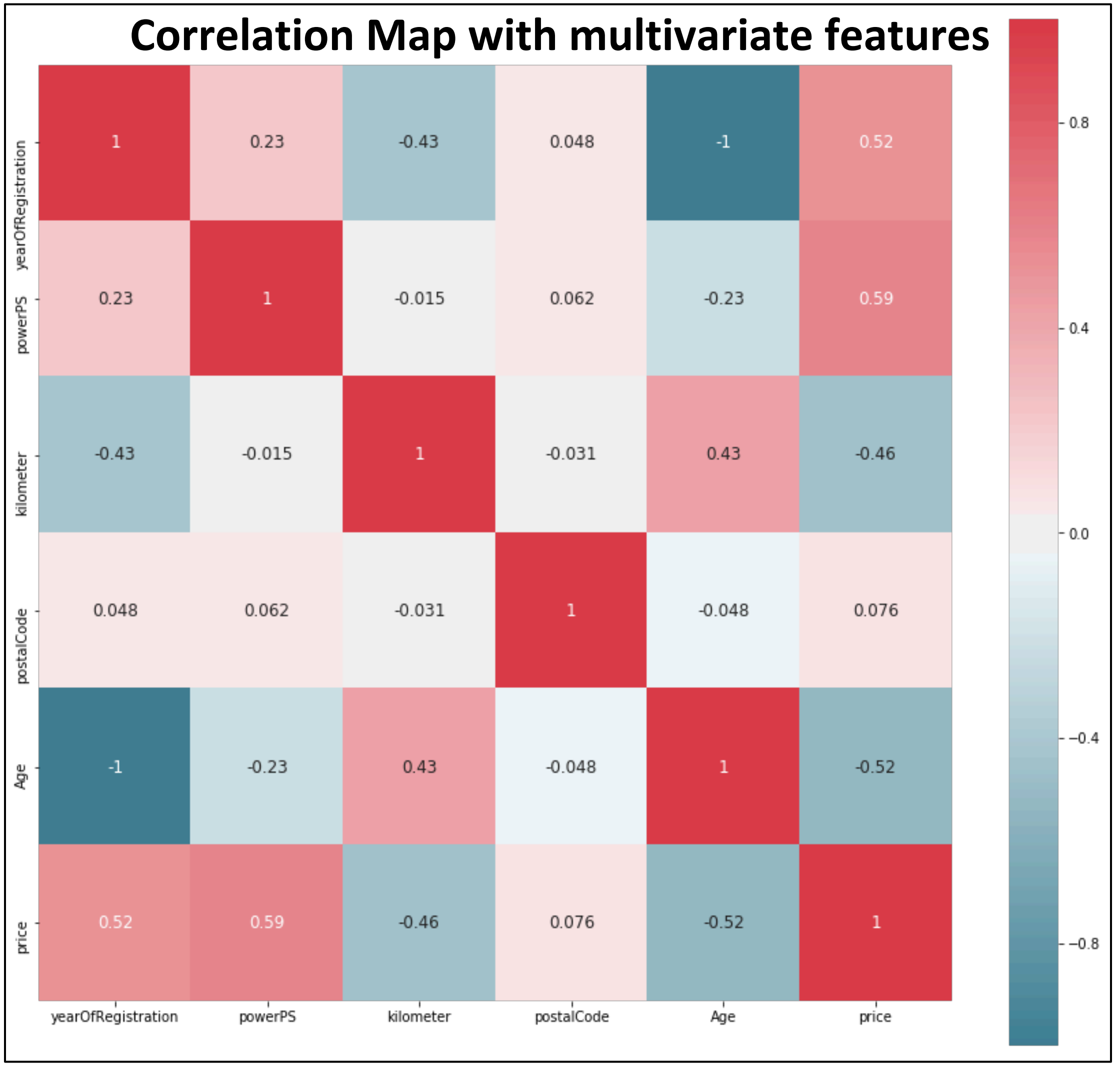


High milage cars have lower prices than the low milage cars



Age is an important factor regarding the price of a car. In this joint plot, we can see that young cars have higher prices than old cars,

Data Visualization



Correlation Map with multivariate features

	yearOfRegistration	powerPS	kilometer	postalCode	Age	price
yearOfRegistration	1	0.23	-0.43	0.048	-1	0.52
powerPS	0.23	1	-0.015	0.062	-0.23	0.59
kilometer	-0.43	-0.015	1	-0.031	0.43	-0.46
postalCode	0.048	0.062	-0.031	1	-0.048	0.076
Age	-1	-0.23	0.43	-0.048	1	-0.52
price	0.52	0.59	-0.46	0.076	-0.52	1

Target feature: price

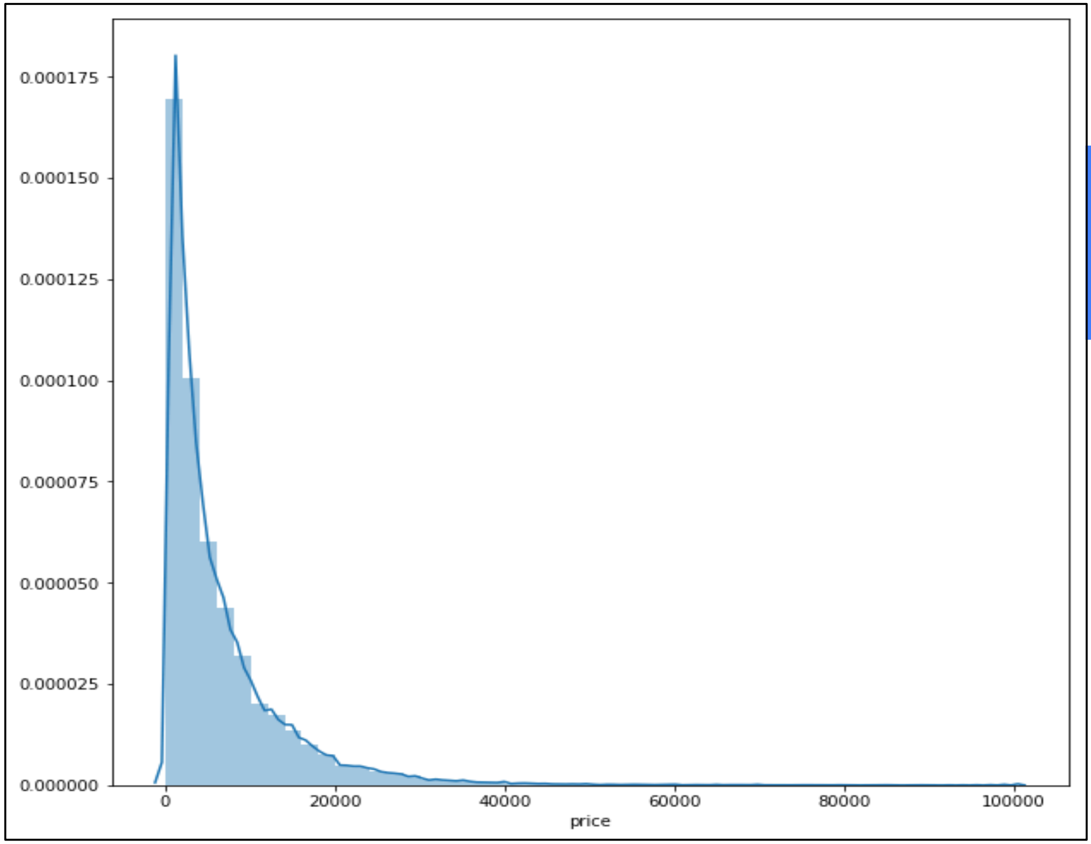
- Age has a negative relationship (-0.52),

- Kilometer has a negative relationship (-0.46)

- PowerPS has a positive relationship (0.52)

Inferential Statistics

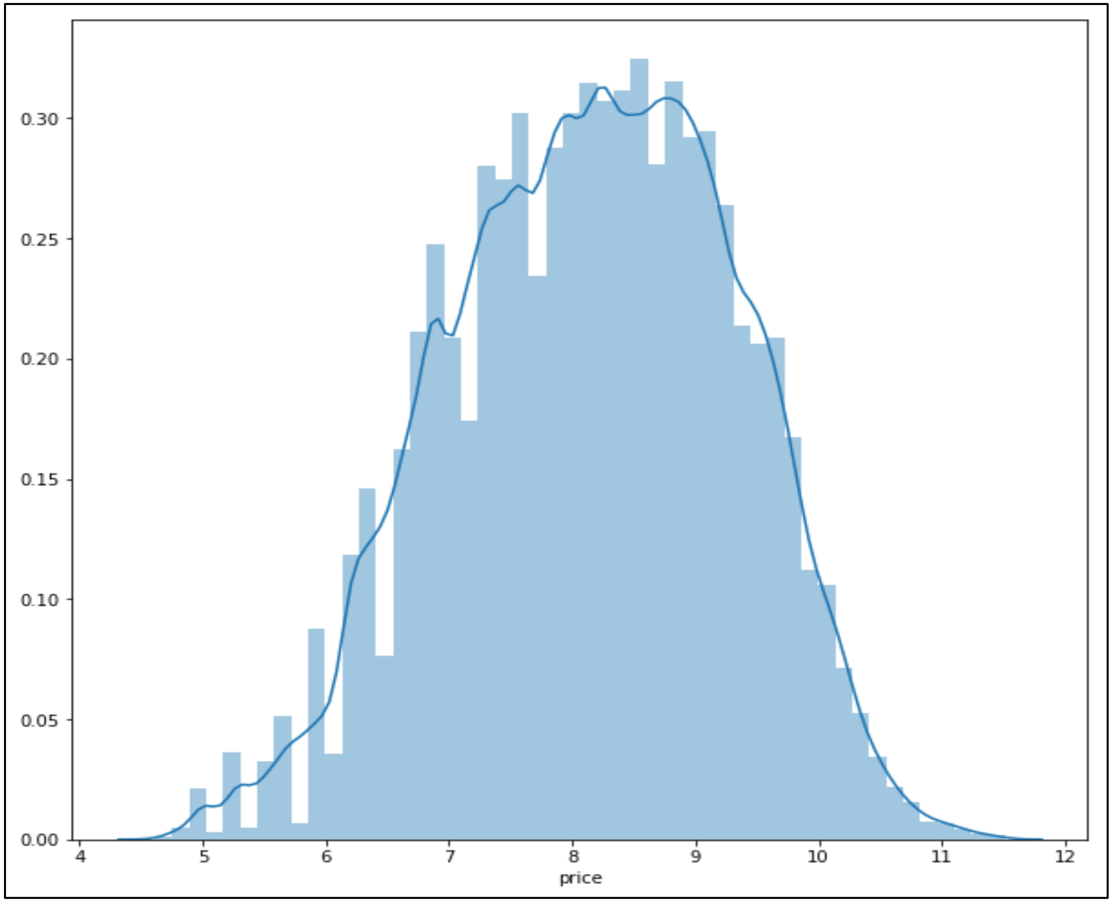
Hypothesis Testing: Distribution of value (price) of the cars normal or not



Shape of the plot looks like distribution is right skewed, we can try log value of the price

Inferential Statistics

Hypothesis Testing: Distribution of value (price) of the cars normal or not

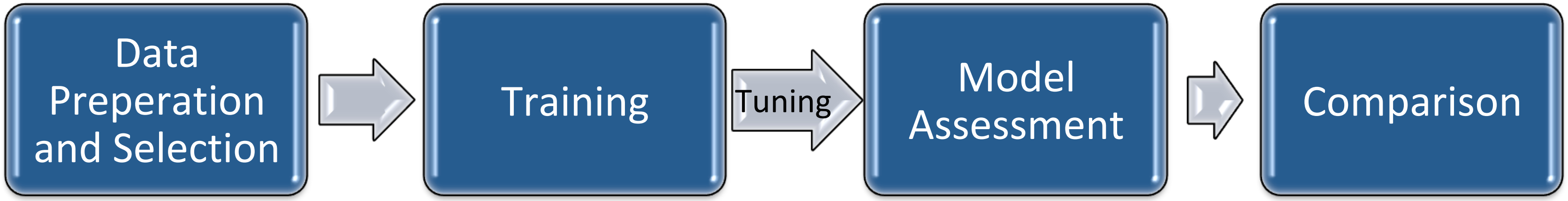


The log-price distribution now looks normally distributed. So our target feature price is log-normally distributed. Null hypothesis can not be rejected.

Predictive Modeling

Machine Learning Models

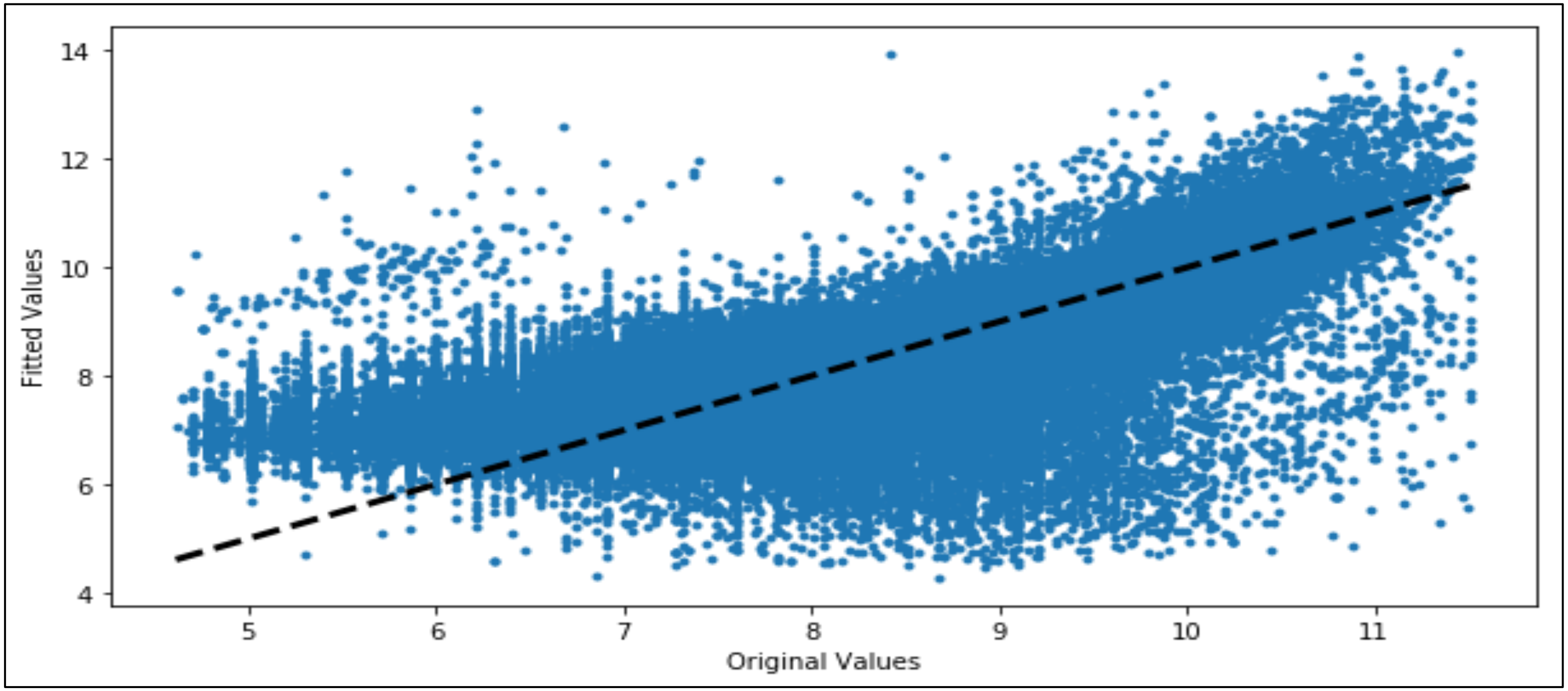
- Supervised learning regression problem
- Used the tools of Scikit Learn, Stat-models, Scipy



- Subset of whole data: 30%
- 5-fold cross validation
- Get dummies of categorical variables

Machine Learning Models

- First, applied a linear regression model to the numeric features : Age , kilometer ,powerPS and yearOfRegistration
- Low R squared (0.65), not very accurate (Linear Regression with Non-numeric Features), not a perfect positive linear correlation



Machine Learning Models

Model Comparison (Numeric Features)

Model	R-Squared	MSE
Linear Regression	0.643	0.4755
Ridge	0.640	0.4786
Lasso	0.642	0.4759
Random Forest	0.793	0.2740

- Highest score is 0.79 with random forest

Machine Learning Models

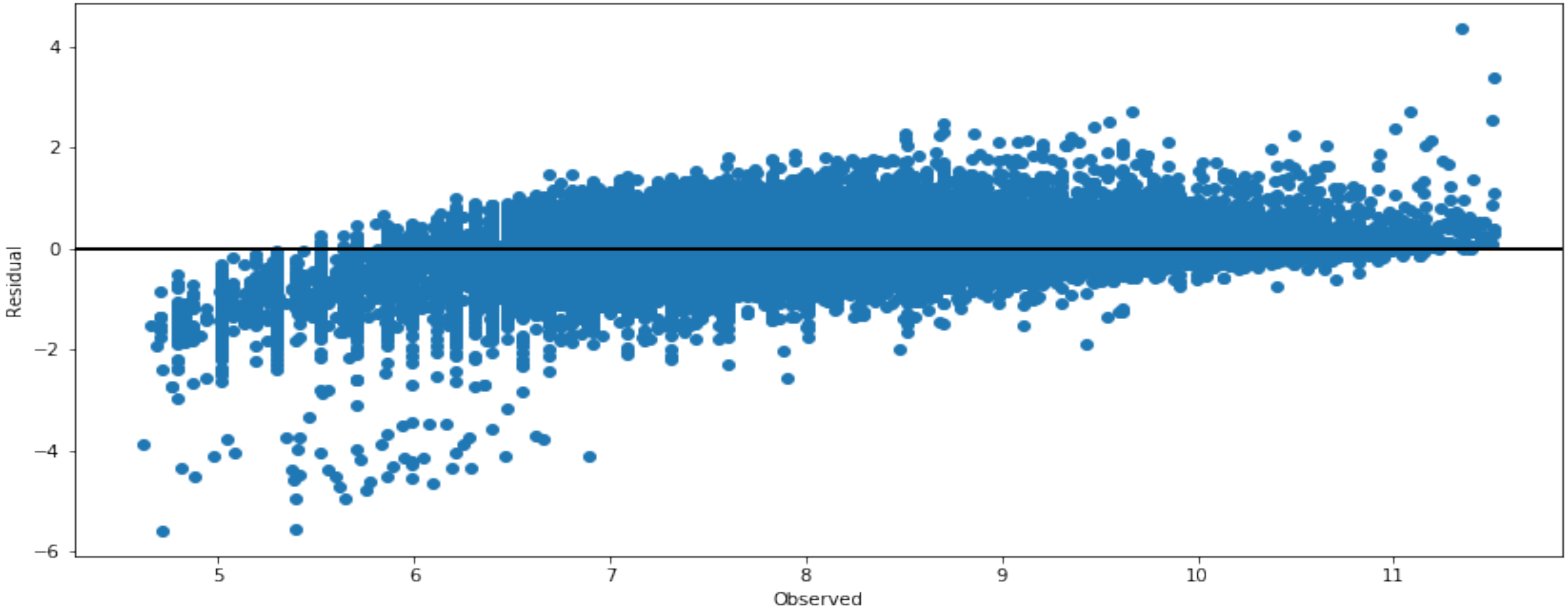
Model Comparison (All Features after dummies)

Model	R-Squared	MSE
Linear Regression	0.776	0.2966
Ridge	0.772	0.3026
Lasso	0.642	0.4759
Decision Tree	0.764	0.3138
Random Forest	0.873	0.1683

- Highest score is 0.87 with **Random Forest** (with the lowest MSE)

Machine Learning Models

Residual Plot



- We can see the small difference between real and predicted price which means that our model works fine

Conclusion

- The aim of this project was to predict the price of a used vehicle based on its features.

- After data wrangling, applying only numeric values resulted a low score prediction 64%)

- Linear regression also gave a low score (64%)

- With feature selection after adding all categorical data, out of four regression models, **Random Forest** gave the best score (**88%**)

Future Works

- Find more data for training (2017, 2018 data)
- After adding all categorical features, hyper tuning and feature selection to drop less effective features
- A new model that predicts ‘how long a vehicle would stay active in the webpage before it is sold’

Thank You