Volkswagen
Used Car Price Prediction

By Thai Linh Bui

Table of contents

1. Dataset Description
2. Data cleaning
3. Exploratory Data Analysis
4. Modeling - Car Price Prediction
5. Difficulties & Improvements

Dataset Description

Source: Kaggle

Description:

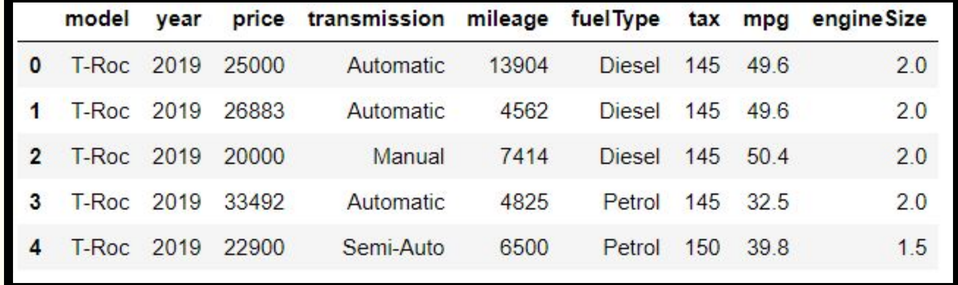
- Used cars in UK (in July 2020)
- 15 157 rows & 9 columns
- No missing values
- 3 columns out of 9 have the object data type: 'model', 'transmission' and 'fuelType'

- The oldest model: 2000
- The newest model: 2020.

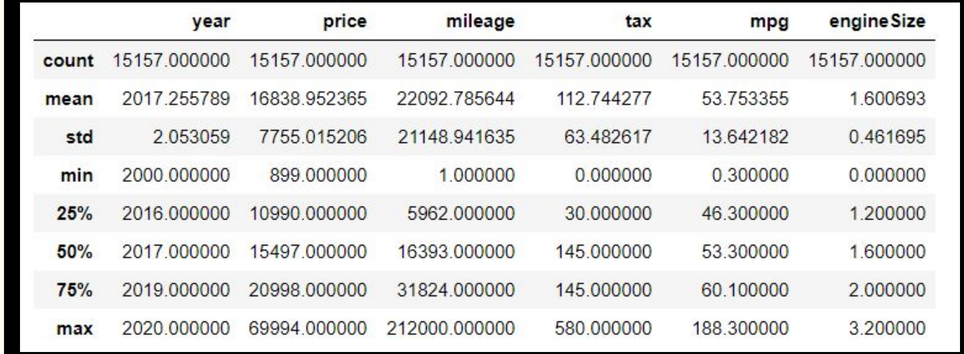
- The price range: 899£ to 69 994£
  The mean : 16K£ AND the 75th quantile: 21K£
  => there might be some outliers.

- Mileage: there might have outliers

- Tax: road tax that owners need to pay every year



	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
0	T-Roc	2019	25000	Automatic	13904	Diesel	145	49.6	2.0
1	T-Roc	2019	26883	Automatic	4562	Diesel	145	49.6	2.0
2	T-Roc	2019	20000	Manual	7414	Diesel	145	50.4	2.0
3	T-Roc	2019	33492	Automatic	4825	Petrol	145	32.5	2.0
4	T-Roc	2019	22900	Semi-Auto	6500	Petrol	150	39.8	1.5

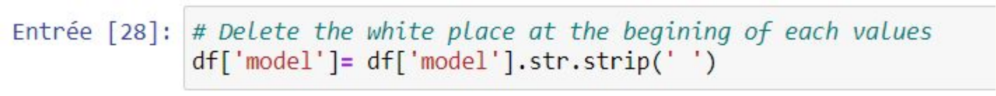


	year	price	mileage	tax	mpg	engineSize
count	15157.000000	15157.000000	15157.000000	15157.000000	15157.000000	15157.000000
mean	2017.255789	16838.952365	22092.785644	112.744277	53.753355	1.600693
std	2.053059	7755.015206	21148.941635	63.482617	13.642182	0.461695
min	2000.000000	899.000000	1.000000	0.000000	0.300000	0.000000
25%	2016.000000	10990.000000	5962.000000	30.000000	46.300000	1.200000
50%	2017.000000	15497.000000	16393.000000	145.000000	53.300000	1.600000
75%	2019.000000	20998.000000	31824.000000	145.000000	60.100000	2.000000
max	2020.000000	69994.000000	212000.000000	580.000000	188.300000	3.200000

Data cleaning

Data cleaning

- In the column 'model', there is a white space preceding each value that needsto be removed



Entrée [28]:

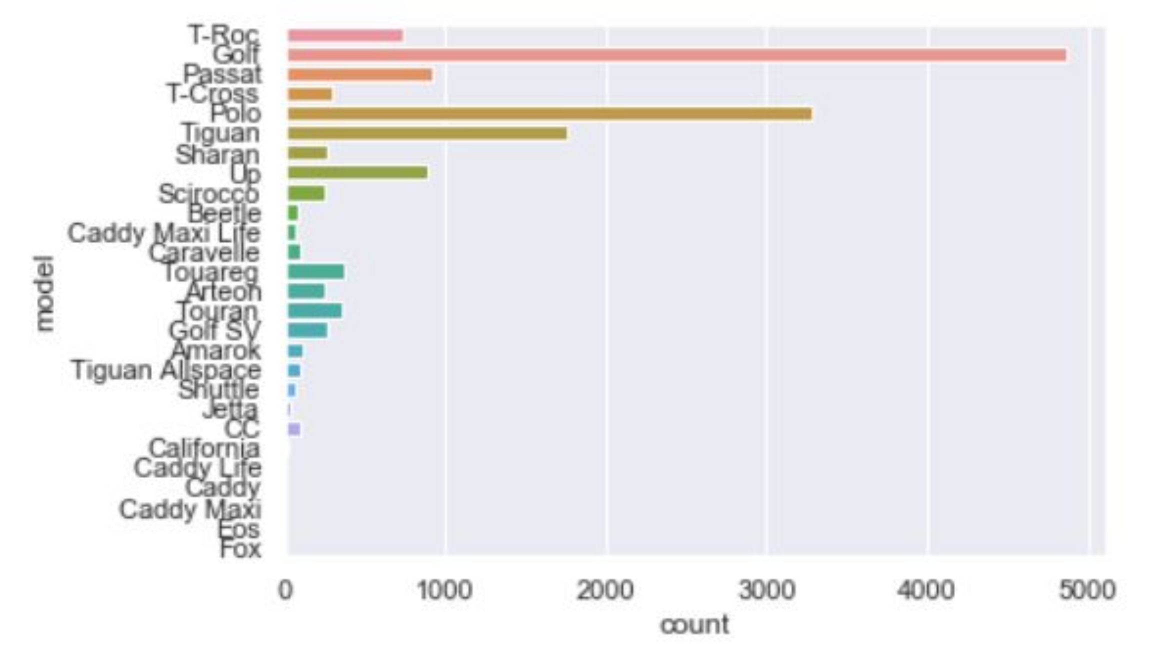
```
# Delete the white place at the begining of each values
df['model']= df['model'].str.strip(' ')
```

- There are several columns to rename

Exploratory Data Analysis

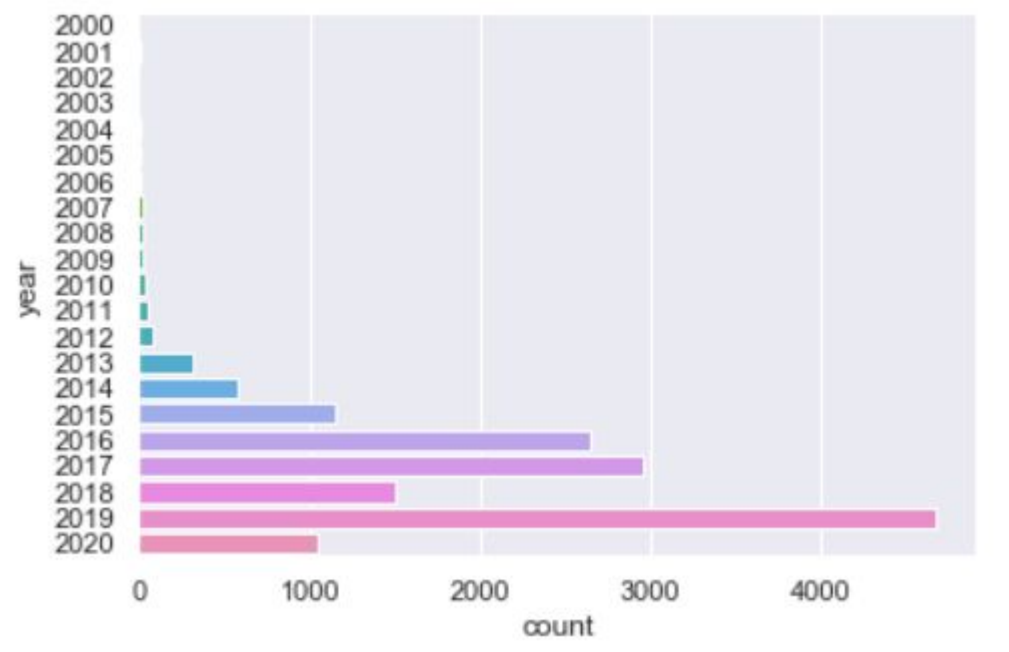
Categorical values

Distribution of categorical values



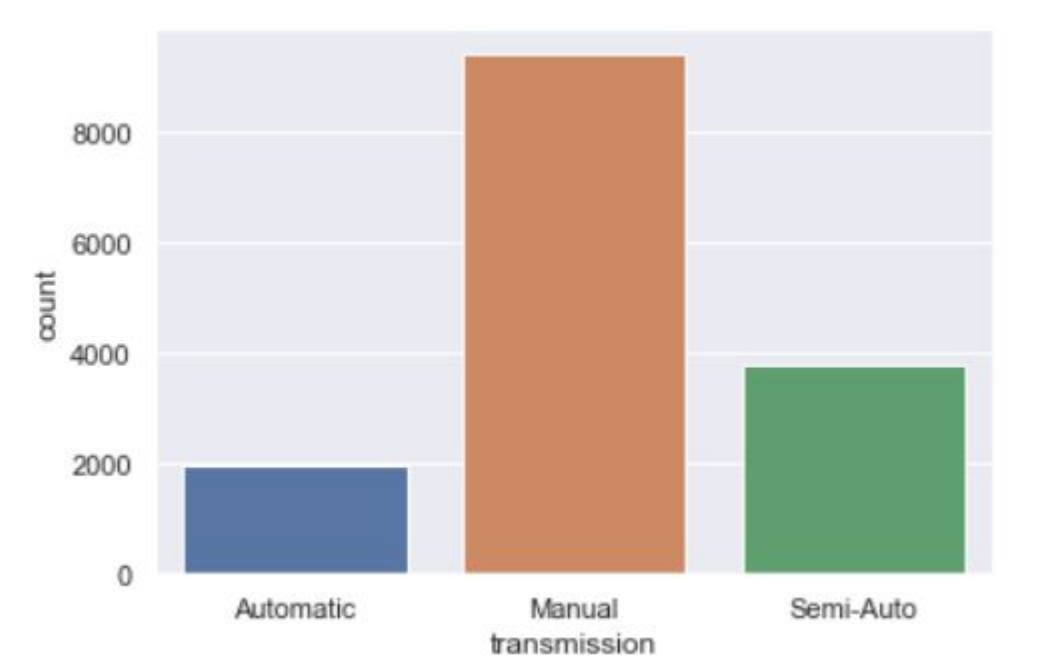
Distribution of car by model

Distribution of categorical values



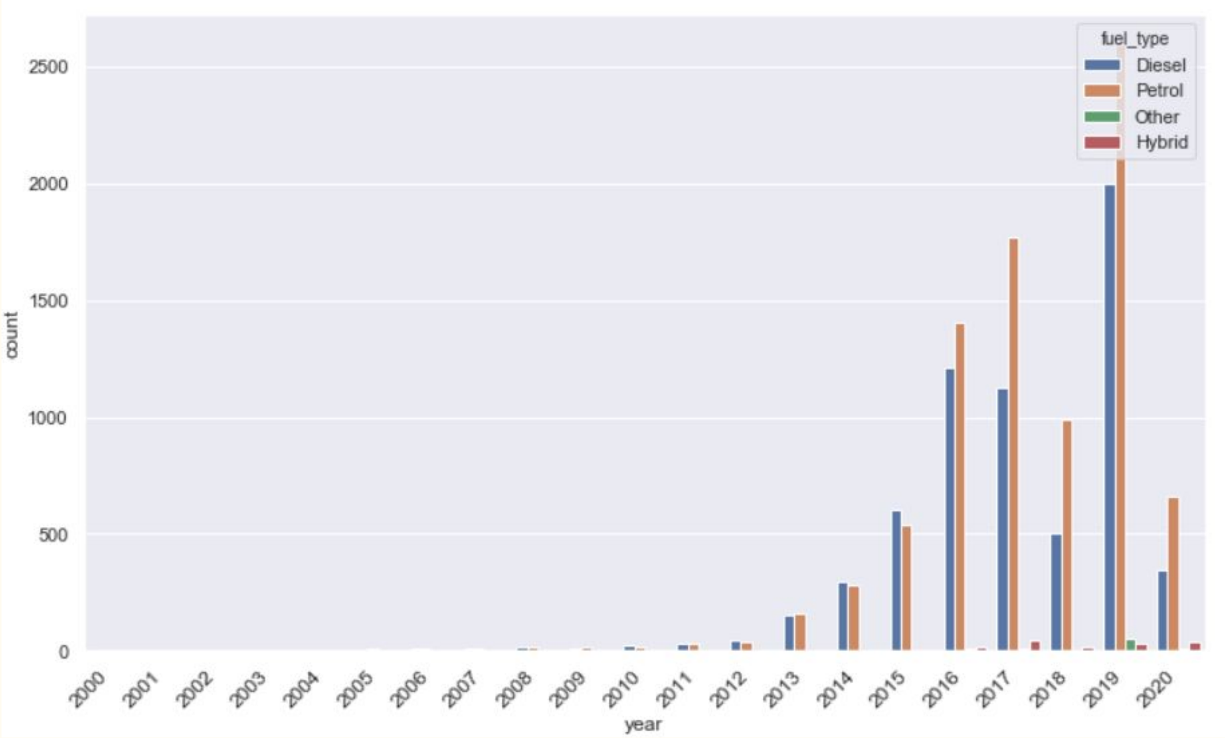
Distribution of car by year

Distribution of categorical values



Distribution of cars by types of transmission

Distribution of categorical values

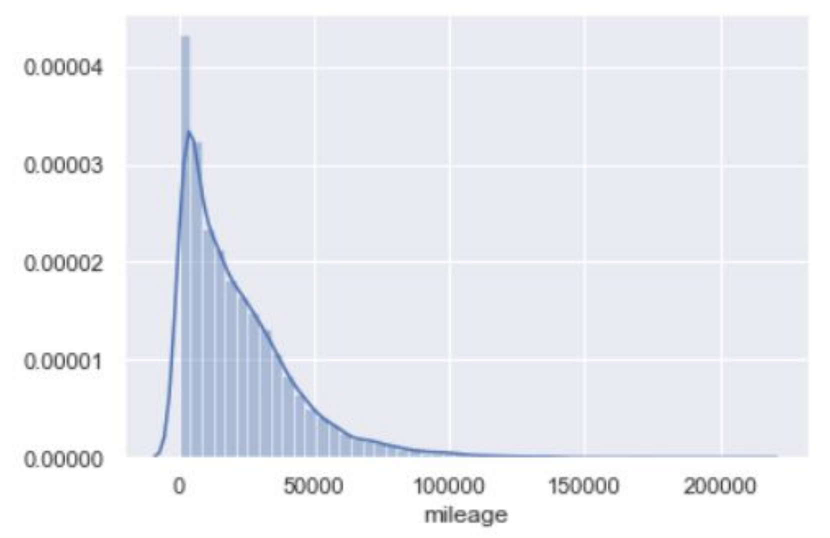


Distribution of cars by types of fuel

The petrol is used the most for the commercialized cars since 2016, while diesel was the most common before 2016.

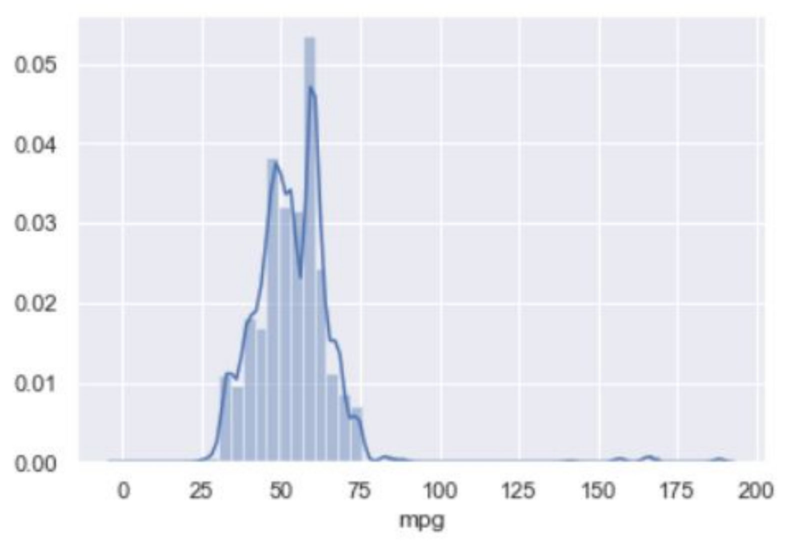
Numerical values

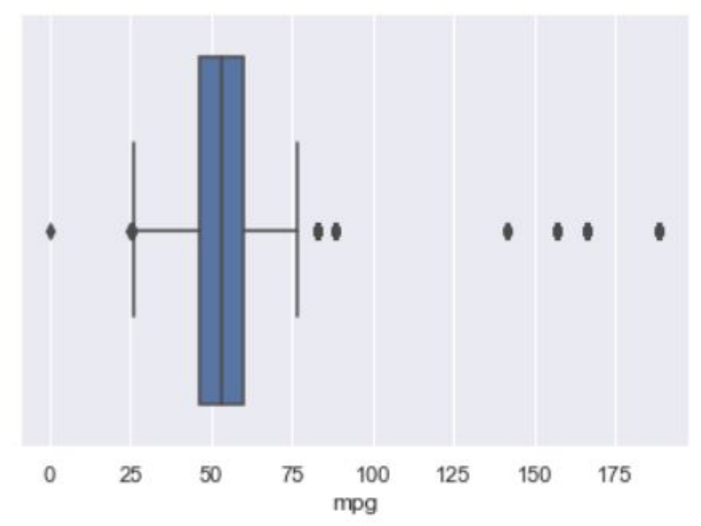
Distribution of numerical values - Mileage



Most of values are less than 50K km
=> might have outliers
=> less accurate while predicting for the cars that have the mileage value > 50K km

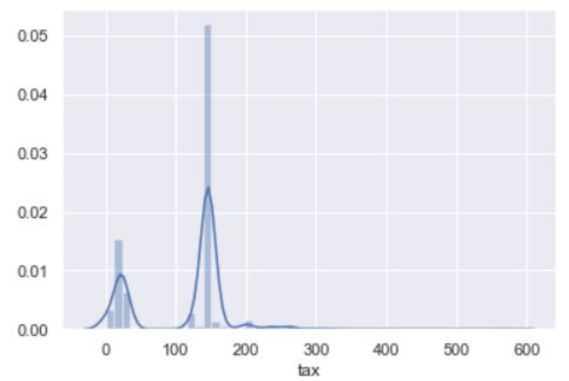
Distribution of numerical values - Mile per gallon

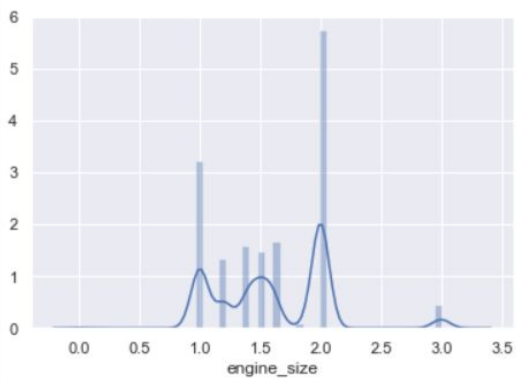


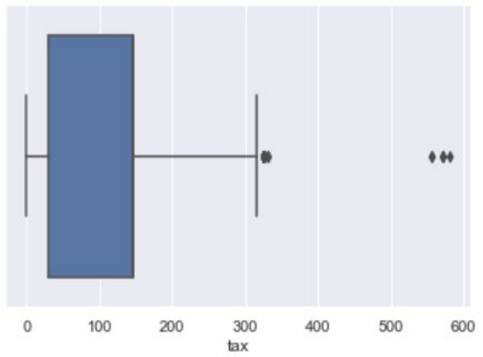


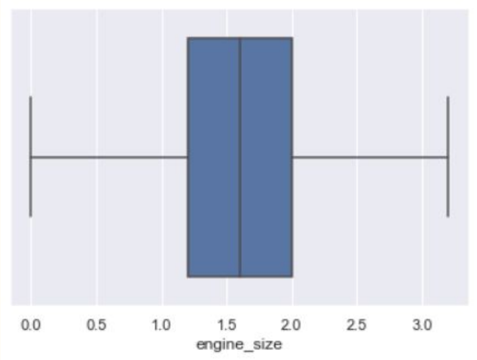
The graph doesn’t really follow the normal distribution.
=> fat tail on the right

Distribution of numerical values - Tax & Engine size



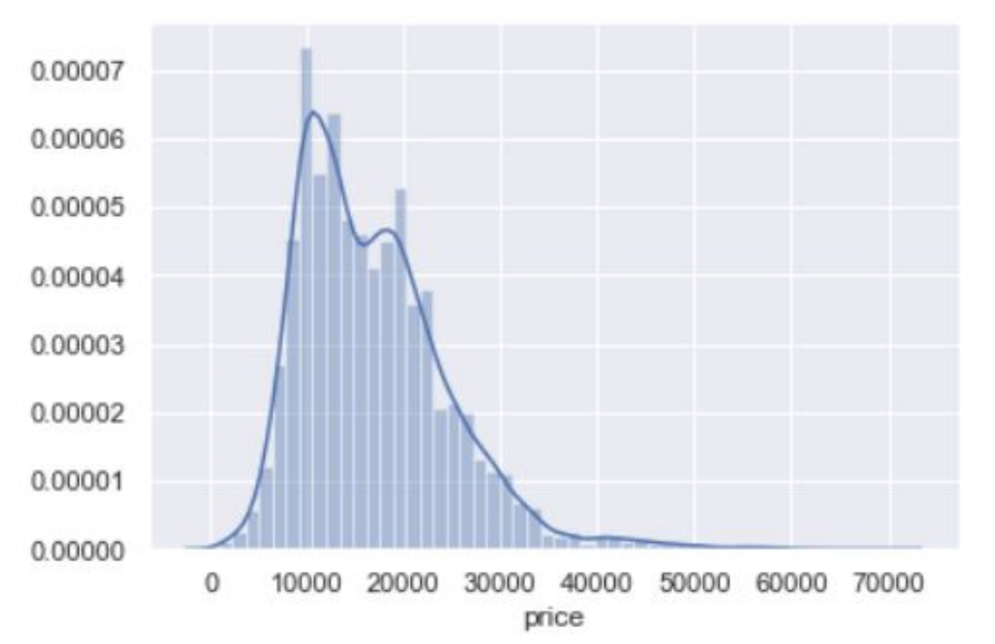






Focus on Price

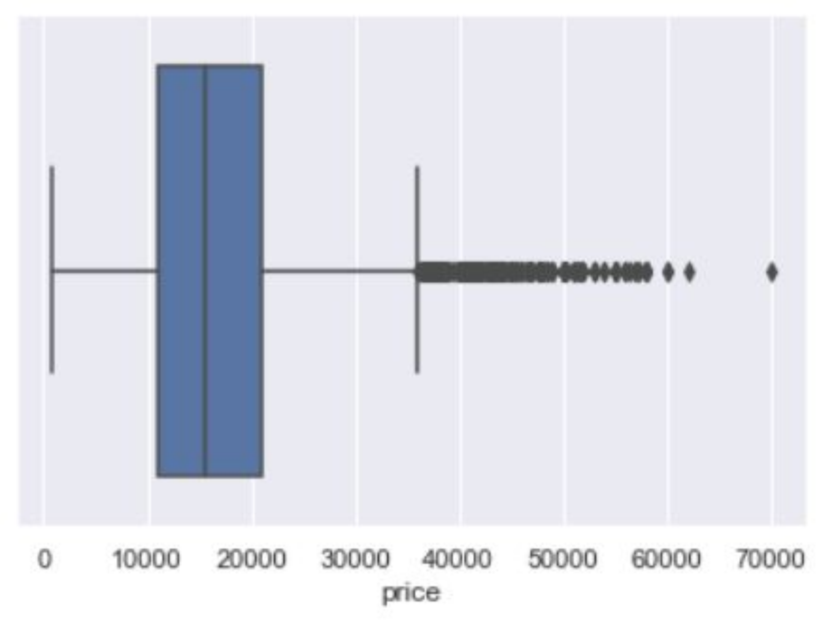
1. Distribution



Right skewed distribution

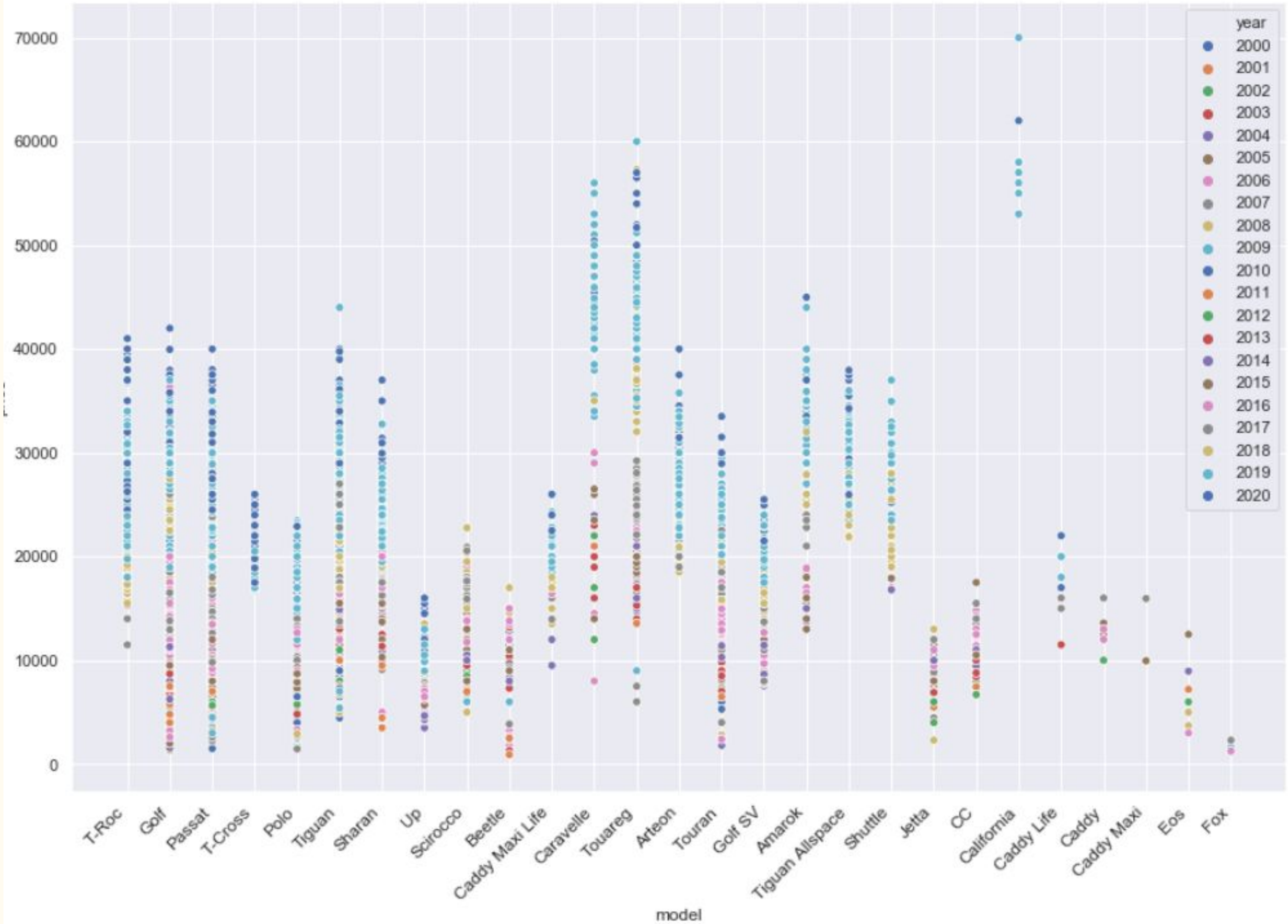
Focus on Price

2. Outliers



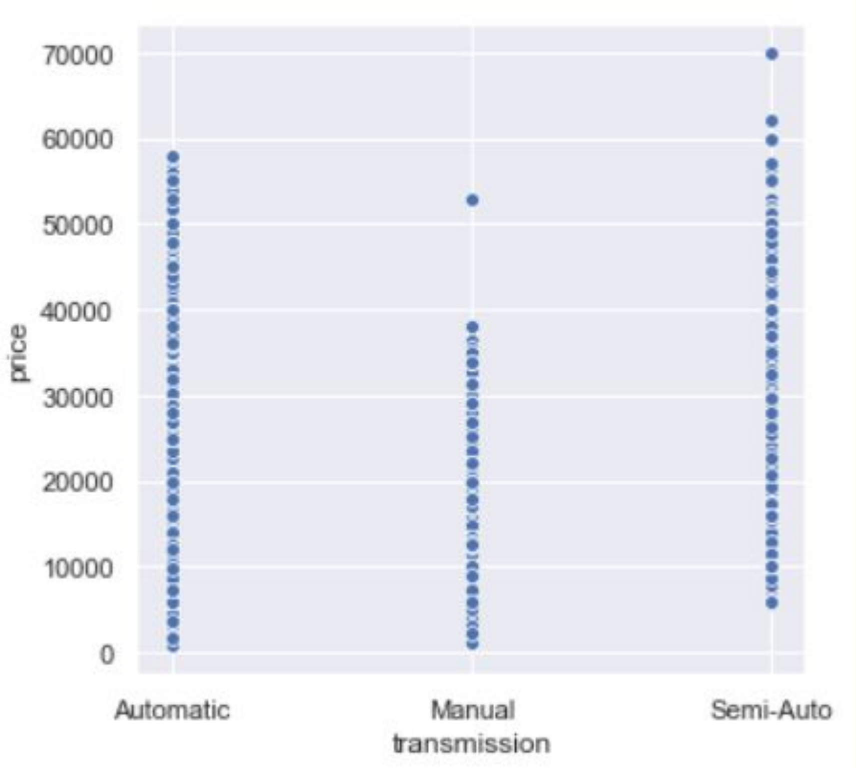
- Justify the form of the normal distribution
- Outliers from the price of 36K£

Question: What models are related the most to the outliers?


70000
60000
50000
40000
30000
20000
10000
0
year
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
T-Roc
Golf
Passat
T-Cross
Polo
Tiguan
Sharan
Up
Scirocco
Beetle
Caddy Maxi Life
Caravelle
Touareg
Arteon
Touran
Golf SV
Amarok
Tiguan Allspace
Shuttle
Jetta
CC
California
Caddy Life
Caddy
Caddy Maxi
Eos
Fox
model

Focus on Price

2. Relationship between Price and Other Variables

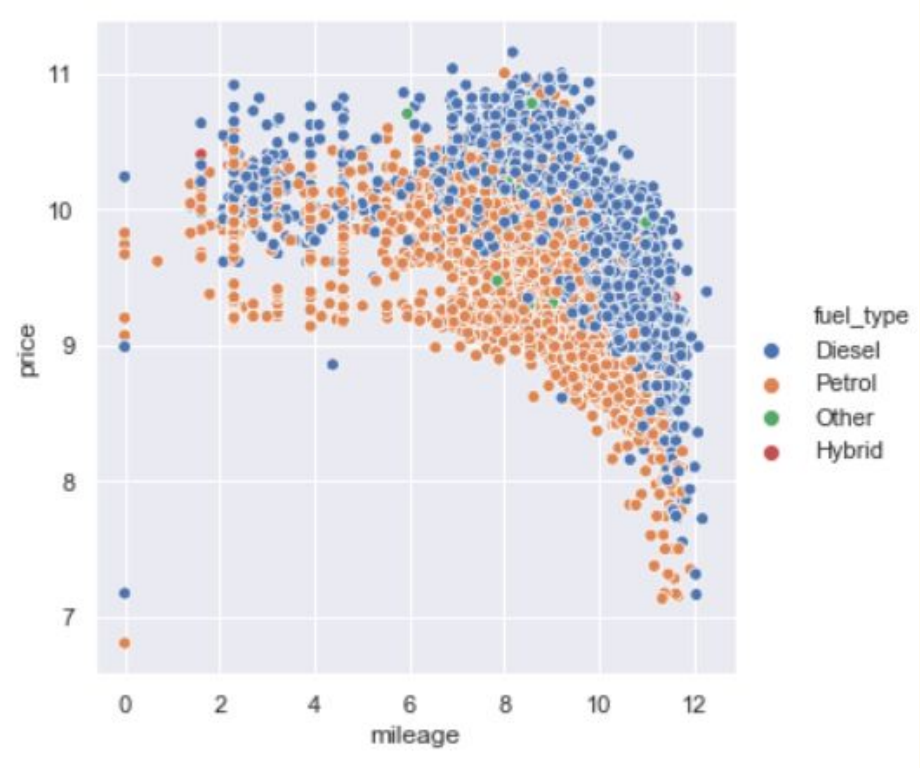


Price & Transmission

The semi-auto and automatic cars tend to be more expensive than the manual ones.

Focus on Price

2. Relationship between Price and other variables

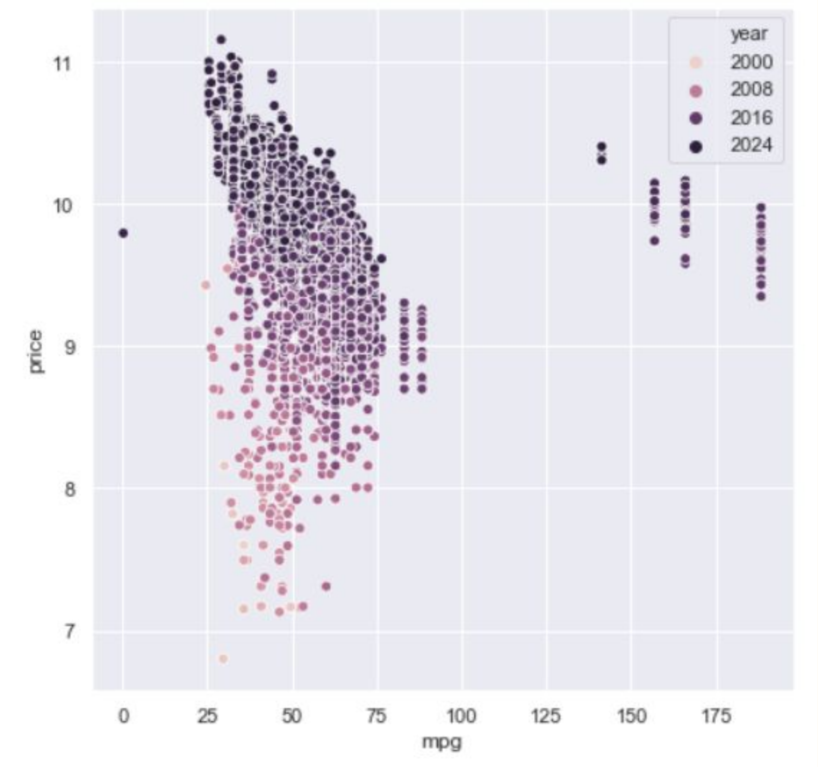


Price & Mileage

The older the car is, the less expensive it is.

Focus on Price

2. Relationship between Price and other variables



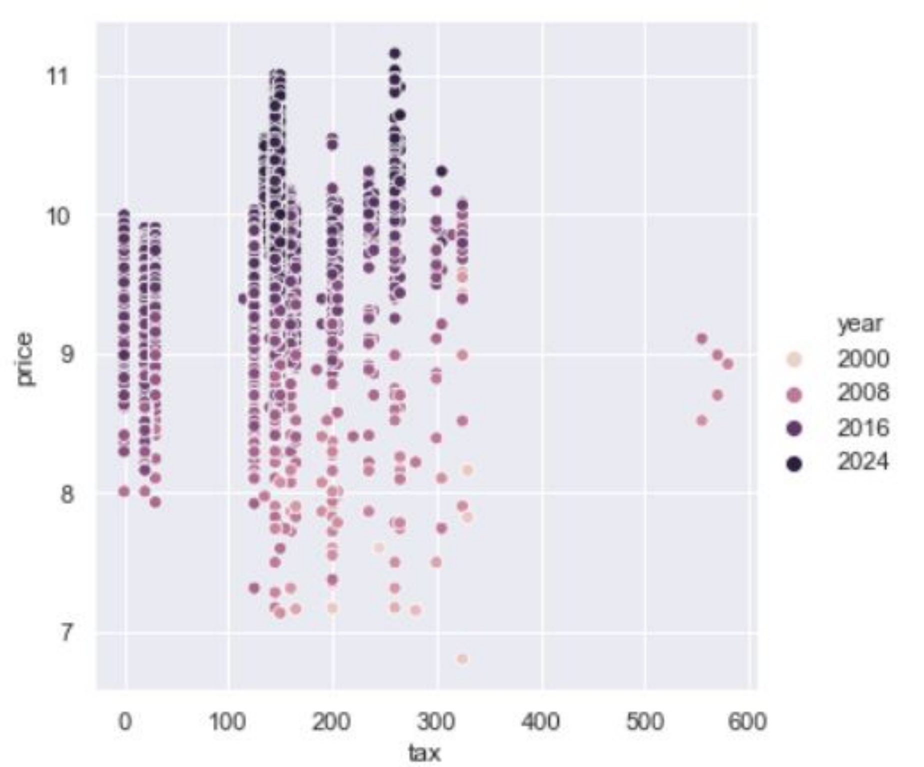
Price & Mile per gallon

The less fuel the car consumes, the more expensive it is.

The older the car is, the less expensive it is even though it consumes not much fuel.

Focus on Price

2. Relationship between Price and other variables

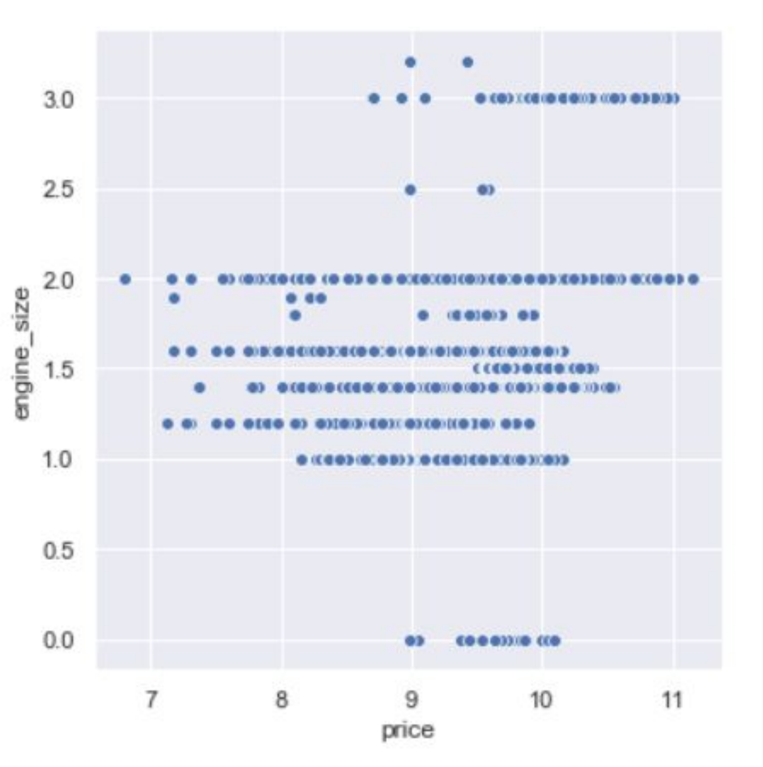


Price & Road tax

There is no evident correlation between the amount of tax road and the car price.

Focus on Price

2. Relationship between Price and other variables



Price & Engine Size

There is seemingly a sign of correlation between the price and the size of the engine.

Correlation between all variables

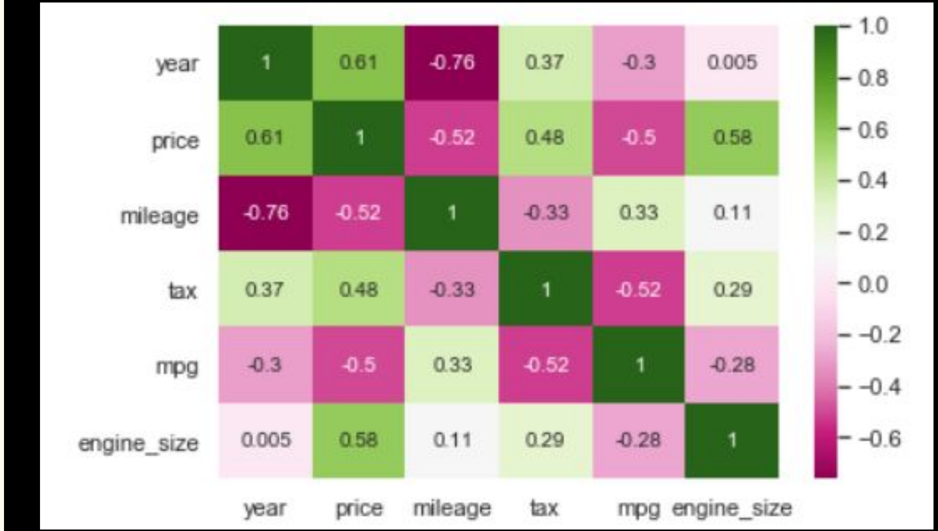
1. Moderate correlation

- Price vs Year: 0.61
- Price vs Engine_size: 0.58

- Price vs Mileage: -0.52
- Price vs Mpg: -0.5
- Road tax vs Mpg: -0.52

2. High correlation

- Mileage vs Year: -0.76

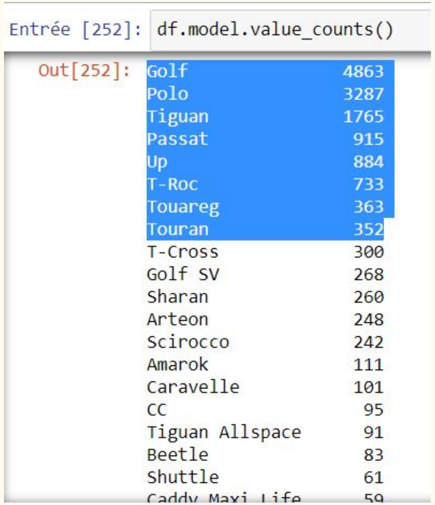


	year	price	mileage	tax	mpg	engine_size
year	1	0.61	-0.76	0.37	-0.3	0.005
price	0.61	1	-0.52	0.48	-0.5	0.58
mileage	-0.76	-0.52	1	-0.33	0.33	0.11
tax	0.37	0.48	-0.33	1	-0.52	0.29
mpg	-0.3	-0.5	0.33	-0.52	1	-0.28
engine_size	0.005	0.58	0.11	0.29	-0.28	1

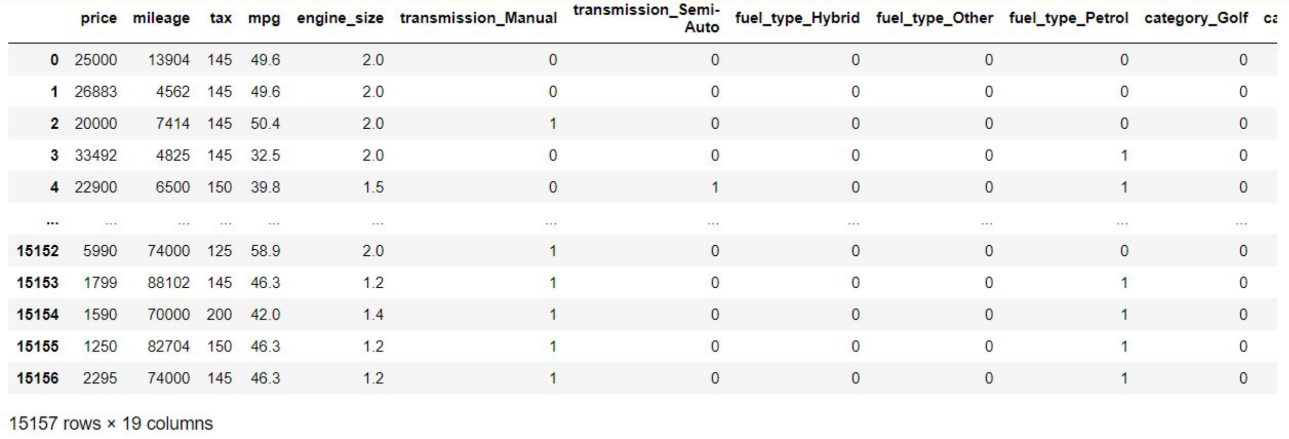
Prediction Modeling

1. Preparation

- Create dummies for categorical columns
  - Column ‘model’: keep 8 models that contribute the most data to the dataset and group other models in a category “Other”



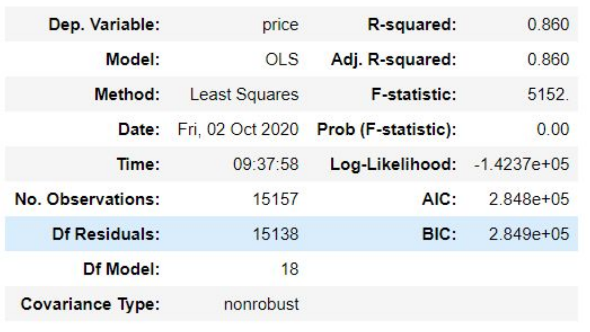
```
Entrée [252]: df.model.value_counts()

Out[252]: Golf                4863
          Polo                3287
          Tiguan              1765
          Passat               915
          Up                   884
          T-Roc                733
          Touareg              363
          Touran               352
          T-Cross              300
          Golf SV              268
          Sharan               260
          Arteon               248
          Scirocco             242
          Amarok               111
          Caravelle            101
          CC                    95
          Tiguan Allspace       91
          Beetle                83
          Shuttle               61
          Caddy Maxi Life       59
```



	price	mileage	tax	mpg	engine_size	transmission_Manual	transmission_Semi-Auto	fuel_type_Hybrid	fuel_type_Other	fuel_type_Petrol	category_Golf	ca
0	25000	13904	145	49.6	2.0	0	0	0	0	0	0	
1	26883	4562	145	49.6	2.0	0	0	0	0	0	0	
2	20000	7414	145	50.4	2.0	1	0	0	0	0	0	
3	33492	4825	145	32.5	2.0	0	0	0	0	1	0	
4	22900	6500	150	39.8	1.5	0	1	0	0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	
15152	5990	74000	125	58.9	2.0	1	0	0	0	0	0	
15153	1799	88102	145	46.3	1.2	1	0	0	0	1	0	
15154	1590	70000	200	42.0	1.4	1	0	0	0	1	0	
15155	1250	82704	150	46.3	1.2	1	0	0	0	1	0	
15156	2295	74000	145	46.3	1.2	1	0	0	0	1	0	

15157 rows × 19 columns

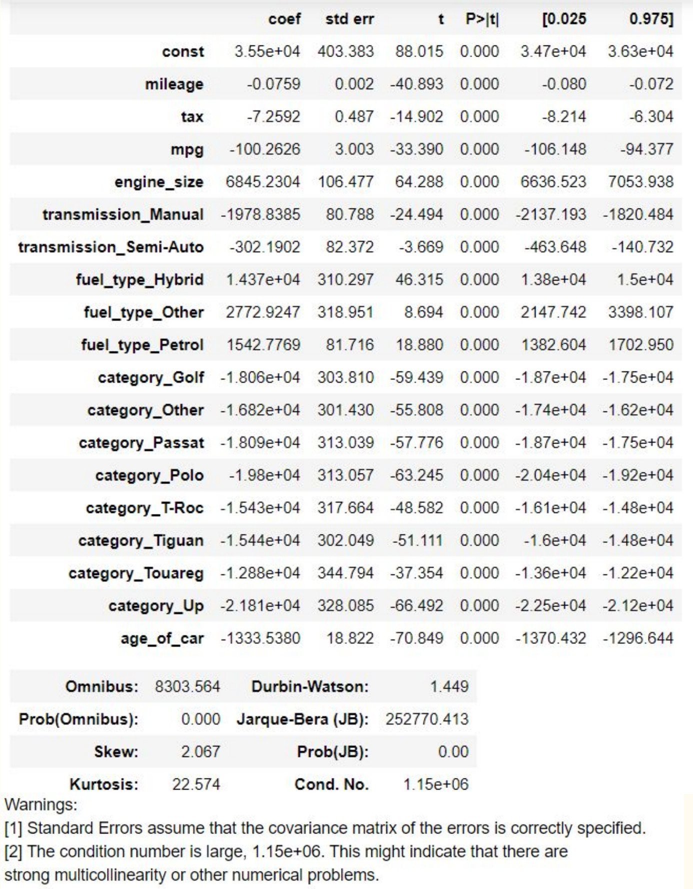
2. Modeling - First Run



Dep. Variable:	price	**R-squared:**	0.860
Model:	OLS	**Adj. R-squared:**	0.860
Method:	Least Squares	**F-statistic:**	5152.
Date:	Fri, 02 Oct 2020	**Prob (F-statistic):**	0.00
Time:	09:37:58	**Log-Likelihood:**	-1.4237e+05
No. Observations:	15157	**AIC:**	2.848e+05
Df Residuals:	15138	**BIC:**	2.849e+05
Df Model:	18		
Covariance Type:	nonrobust		

Passed indicators:
R^2 & Adj R^2: 86% - Good
Prob (F-statistic): 0 - Good
P>|t| - Pvalue: 0 - Good
Durbin-watson: 1.4 - Good

Failed indicators:
Prob(Omnibus): 0 - Not Good
Warning messages



	coef	std err	t	P>|t|	[0.025	0.975]
const	3.55e+04	403.383	88.015	0.000	3.47e+04	3.63e+04
mileage	-0.0759	0.002	-40.893	0.000	-0.080	-0.072
tax	-7.2592	0.487	-14.902	0.000	-8.214	-6.304
mpg	-100.2626	3.003	-33.390	0.000	-106.148	-94.377
engine_size	6845.2304	106.477	64.288	0.000	6636.523	7053.938
transmission_Manual	-1978.8385	80.788	-24.494	0.000	-2137.193	-1820.484
transmission_Semi-Auto	-302.1902	82.372	-3.669	0.000	-463.648	-140.732
fuel_type_Hybrid	1.437e+04	310.297	46.315	0.000	1.38e+04	1.5e+04
fuel_type_Other	2772.9247	318.951	8.694	0.000	2147.742	3398.107
fuel_type_Petrol	1542.7769	81.716	18.880	0.000	1382.604	1702.950
category_Golf	-1.806e+04	303.810	-59.439	0.000	-1.87e+04	-1.75e+04
category_Other	-1.682e+04	301.430	-55.808	0.000	-1.74e+04	-1.62e+04
category_Passat	-1.809e+04	313.039	-57.776	0.000	-1.87e+04	-1.75e+04
category_Polo	-1.98e+04	313.057	-63.245	0.000	-2.04e+04	-1.92e+04
category_T-Roc	-1.543e+04	317.664	-48.582	0.000	-1.61e+04	-1.48e+04
category_Tiguan	-1.544e+04	302.049	-51.111	0.000	-1.6e+04	-1.48e+04
category_Touareg	-1.288e+04	344.794	-37.354	0.000	-1.36e+04	-1.22e+04
category_Up	-2.181e+04	328.085	-66.492	0.000	-2.25e+04	-2.12e+04
age_of_car	-1333.5380	18.822	-70.849	0.000	-1370.432	-1296.644

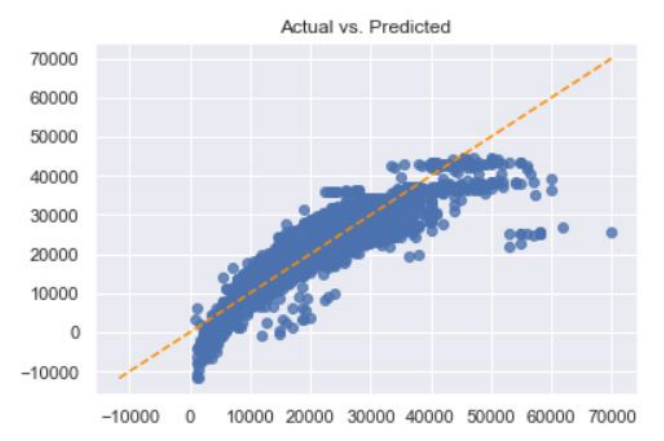
Omnibus:	8303.564	**Durbin-Watson:**	1.449
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	252770.413
Skew:	2.067	**Prob(JB):**	0.00
Kurtosis:	22.574	**Cond. No.**	1.15e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.15e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

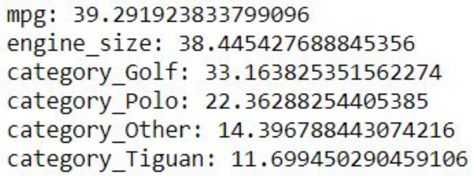
2. Modeling - First Run

Assumption check

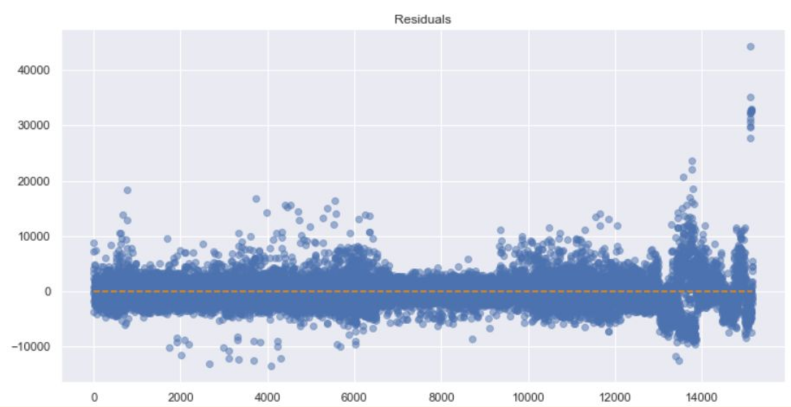
- Satisfied: Linearity



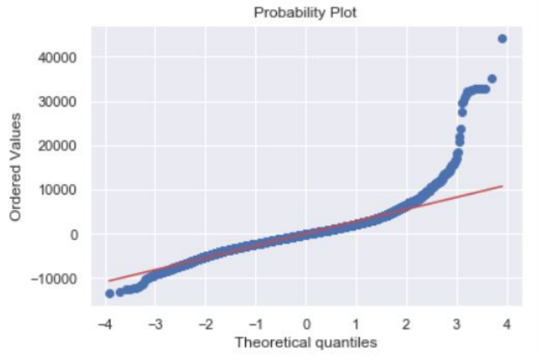
- Potentially violated:
  - Multicollinearity: 6 variables whose VIF $>$ 10
  - Heteroskedasticity:



```
mpg: 39.291923833799096
engine_size: 38.445427688845356
category_Golf: 33.163825351562274
category_Polo: 22.36288254405385
category_Other: 14.396788443074216
category_Tiguan: 11.699450290459106
```



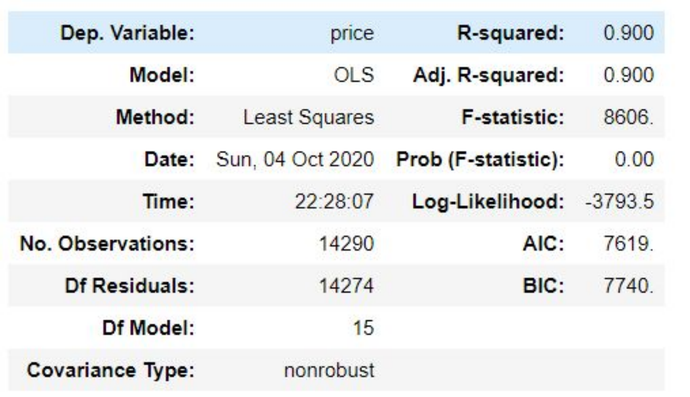
- Violated:
  - Autocorrelation: d $=$ 1.44 $<$ 1.5
  - Normality: Residuals are not normally distributed



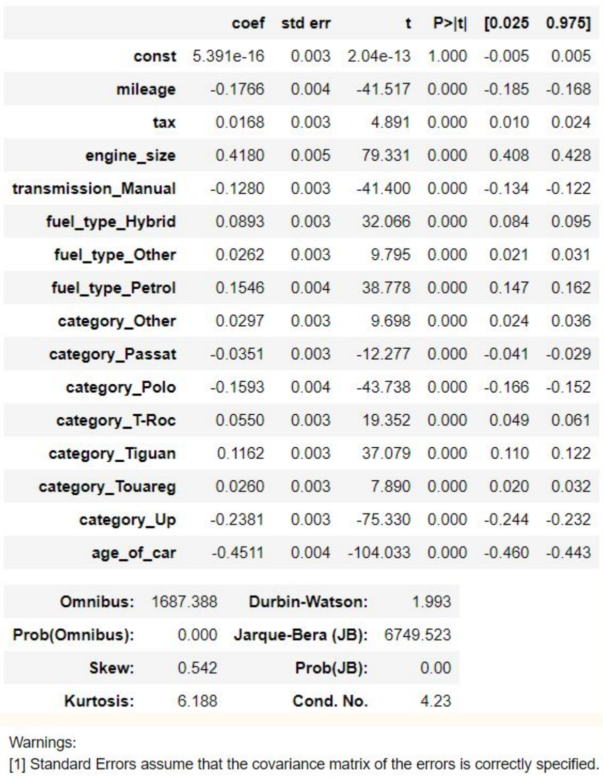
3. Modeling - Second Run

- Correction made:
    - Apply natural logarithm to the values of Price and Mileage -> Assumption - Linearity: line will be more linear
    - Drop column ‘mpg’ and ‘category_Tiguan’ -> Assumption - Multicollinearity: reduce number of columns whose VIF > 10
    - Eliminate outliers for all columns except dummies -> Assumption - Homoskedasticity & Normality: less outliers and residual line will have less fat tails.
    - Normalize the dataset
    - Shuffle the dataset -> Assumption - Autocorrelation: reduce the risk that there is a specific pattern between values formed by a random order which can impact this assumption

3. Modeling - Second Run



Dep. Variable:	price	**R-squared:**	0.900
Model:	OLS	**Adj. R-squared:**	0.900
Method:	Least Squares	**F-statistic:**	8606.
Date:	Sun, 04 Oct 2020	**Prob (F-statistic):**	0.00
Time:	22:28:07	**Log-Likelihood:**	-3793.5
No. Observations:	14290	**AIC:**	7619.
Df Residuals:	14274	**BIC:**	7740.
Df Model:	15		
Covariance Type:	nonrobust		



	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	5.391e-16	0.003	2.04e-13	1.000	-0.005	0.005
mileage	-0.1766	0.004	-41.517	0.000	-0.185	-0.168
tax	0.0168	0.003	4.891	0.000	0.010	0.024
engine_size	0.4180	0.005	79.331	0.000	0.408	0.428
transmission_Manual	-0.1280	0.003	-41.400	0.000	-0.134	-0.122
fuel_type_Hybrid	0.0893	0.003	32.066	0.000	0.084	0.095
fuel_type_Other	0.0262	0.003	9.795	0.000	0.021	0.031
fuel_type_Petrol	0.1546	0.004	38.778	0.000	0.147	0.162
category_Other	0.0297	0.003	9.698	0.000	0.024	0.036
category_Passat	-0.0351	0.003	-12.277	0.000	-0.041	-0.029
category_Polo	-0.1593	0.004	-43.738	0.000	-0.166	-0.152
category_T-Roc	0.0550	0.003	19.352	0.000	0.049	0.061
category_Tiguan	0.1162	0.003	37.079	0.000	0.110	0.122
category_Touareg	0.0260	0.003	7.890	0.000	0.020	0.032
category_Up	-0.2381	0.003	-75.330	0.000	-0.244	-0.232
age_of_car	-0.4511	0.004	-104.033	0.000	-0.460	-0.443

Omnibus:	1687.388	**Durbin-Watson:**	1.993
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6749.523
Skew:	0.542	**Prob(JB):**	0.00
Kurtosis:	6.188	**Cond. No.**	4.23

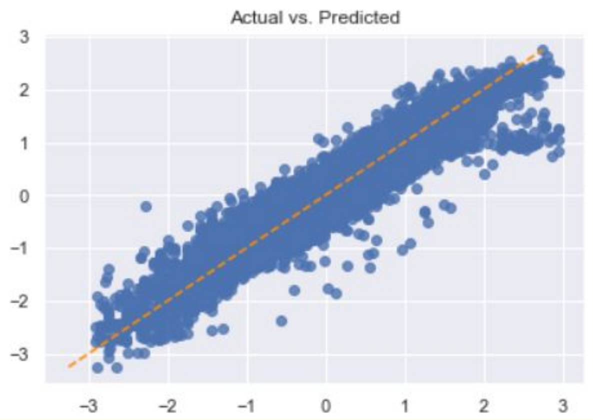
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

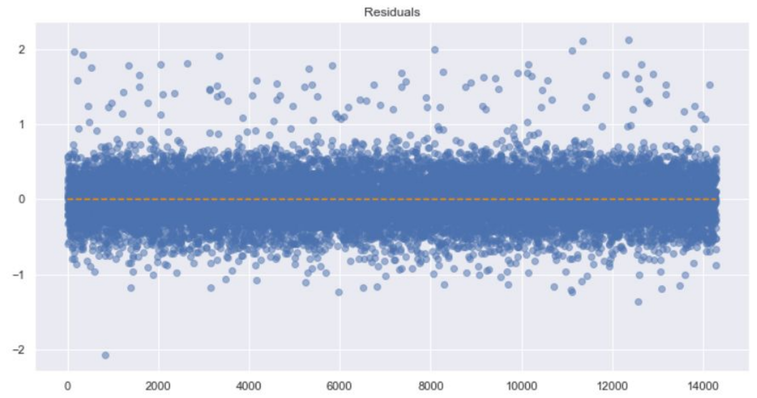
- Improved most of the key indicators
- **Question:** why P value of ‘const’ = 1?

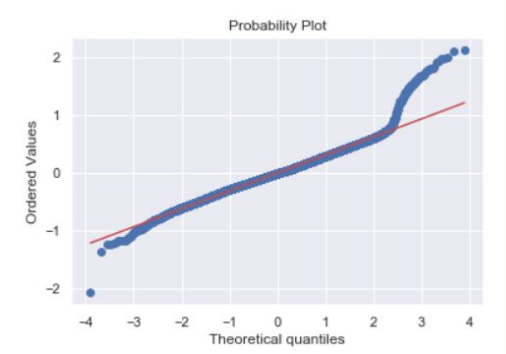
3. Modeling - Second Run

Assumption check

- Satisfied:
  - Linearity
  - Multicollinearity: no variable having VIF > 10
  - Autocorrelation: d = 1.99 > 1.5
- Potentially violated:
  - Heteroskedasticity:
- Violated:
  - Normality: Residuals are not normally distributed







Conclusion

Prediction Model:

- R-squared: 90%
- 1 violated assumption need to be remediated

Difficulties & Improvements

Difficulties

- Understanding of the reasons for errors detected in the Assumption checks
- Solutions for failed assumptions, specially the assumption ‘Normality’

Improvements

- Improve the model so that all assumptions can be validated

Github

https://github.com/EricBui0201/UK-Car-Price-Prediction

Thank you