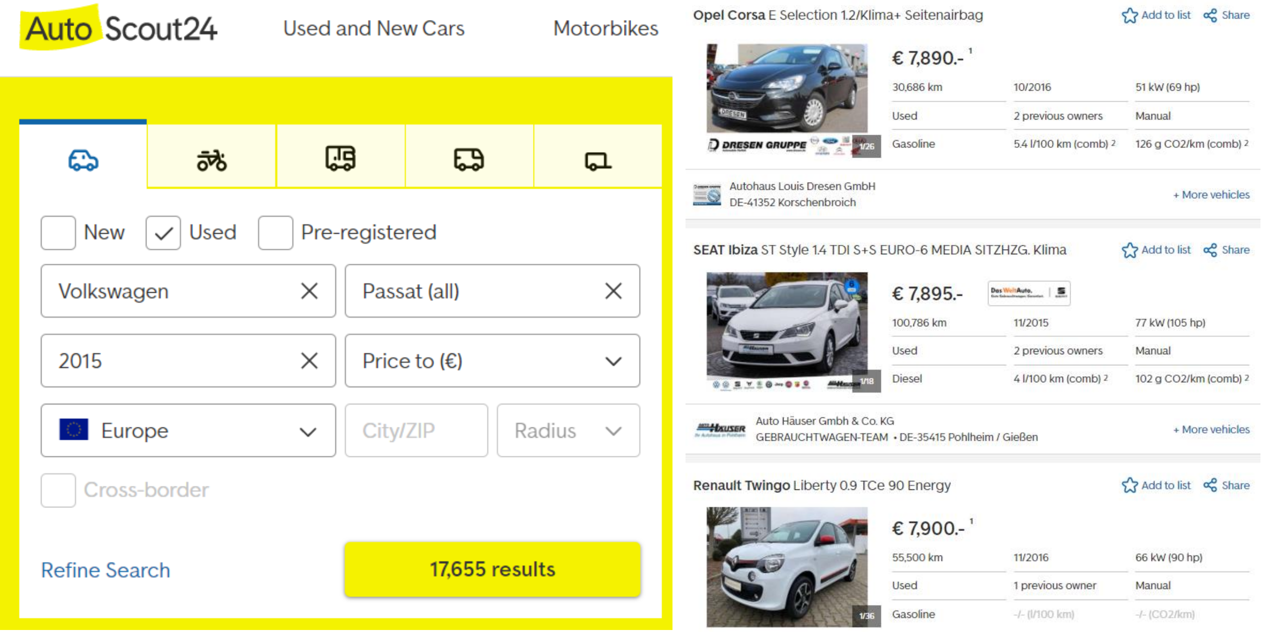


Project-02:

Web Scraping & Machine Learning

Predicting *Used Car Prices* in EU

Muharrem SELBİÇER

ISTANBUL DATA SCIENCE ACADEMY

Introduction

Target:

Predicting **prices** of used cars in Europe by;

① *collecting* data → ads(*www.autoscout24.com*)

② *determining* **the best regression model,**

③ *getting* ***a*** **successful score.**

Methodology

- **Dataset: *Web Scraping with Beautiful Soup***

Brand	Body_Type	Year	Power	Gear_Type	Fuel_Type	Consumption	CO2_Emission	Kilometer	Price_EUR
Audi A1	1	\n11/2017\n	\n70 kW (95 hp)\n	\nAutomatic\n	\nGasoline\n	\n-/- (l/100 km)\n	\n-/- (CO2/km)\n	\n13,000 km\n	\n€ 17,600.- \n
Abarth 595 Pista	1	\n01/2019\n	\n118 kW (160 hp)\n	\nManual\n	\nGasoline\n	\n6.9 l/100 km (comb)\nYou can obtain more inf...	\n158 g CO2/km (comb)\nYou can obtain more inf...	\n13,500 km\n	\n€ 16,499.- \n
Hyundai i10	1	\n06/2018\n	\n49 kW (67 hp)\n	\nManual\n	\nGasoline\n	\n4.7 l/100 km (comb)\nYou can obtain more inf...	\n-/- (CO2/km)\n	\n13,600 km\n	\n€ 8,990.- \n
Trabant P601	1	\n04/1964\n	\n19 kW (26 hp)\n	\nManual\n	\nGasoline\n	\n-/- (l/100 km)\n	\n-/- (CO2/km)\n	\n14,300 km\n	\n€ 1,200.- \n
Peugeot 206	1	\n07/2011\n	\n55 kW (75 hp)\n	\nManual\n	\nGasoline\n	\n6 l/100 km (comb)\nYou can obtain more infor...	\n-/- (CO2/km)\n	\n14,600 km\n	\n€ 2,700.- \n

Brands	Models	Body_Type	Age	Hp	Gear	Fuel	Cons	CO2_E	Km	Price
Abarth	595 Pista	Compact	1	160	Manual	Gasoline	6.9	158	13500	16499
Opel	Corsa	Compact	2	90	Manual	Gasoline	6.4	128	16000	10450
Toyota	Yaris	Compact	1	111	Manual	Gasoline	5.1	116	17000	14350
Suzuki	Swift	Compact	1	111	Manual	Electric/Gasoline	4.3	123	17000	11200
Toyota	Aygo	Compact	12	77	Automatic	Gasoline	4.6	108	17500	5999

Methodology

- **EDA**

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7910 entries, 0 to 7909
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Brands     7910 non-null   object
 1   Models     7910 non-null   object
 2   Body_Type  7910 non-null   object
 3   Age        7910 non-null   int64
 4   Hp         7910 non-null   int64
 5   Gear       7910 non-null   object
 6   Fuel       7910 non-null   object
 7   Cons       7910 non-null   float64
 8   CO2_E      7910 non-null   float64
 9   Km         7910 non-null   int64
 10  Price      7910 non-null   int64
dtypes: float64(2), int64(4), object(5)
memory usage: 741.6+ KB
```

	Age	Hp	Cons	CO2_E	Km	Price
count	7910.000000	7910.000000	7910.000000	7910.000000	7910.000000	7910.000000
mean	9.960556	152.088496	6.123724	150.569512	131916.510240	10786.250822
std	5.795627	72.415660	1.700440	39.910772	67795.149498	9660.568046
min	0.000000	15.000000	0.600000	7.000000	10022.000000	500.000000
25%	5.000000	105.000000	4.900000	122.000000	78000.000000	3850.000000
50%	9.000000	140.000000	5.800000	143.000000	129000.000000	7999.000000
75%	14.000000	184.000000	7.000000	170.000000	180000.000000	14950.000000
max	44.000000	751.000000	18.000000	462.000000	300000.000000	98000.000000

python™

jupyter

pandas

NumPy

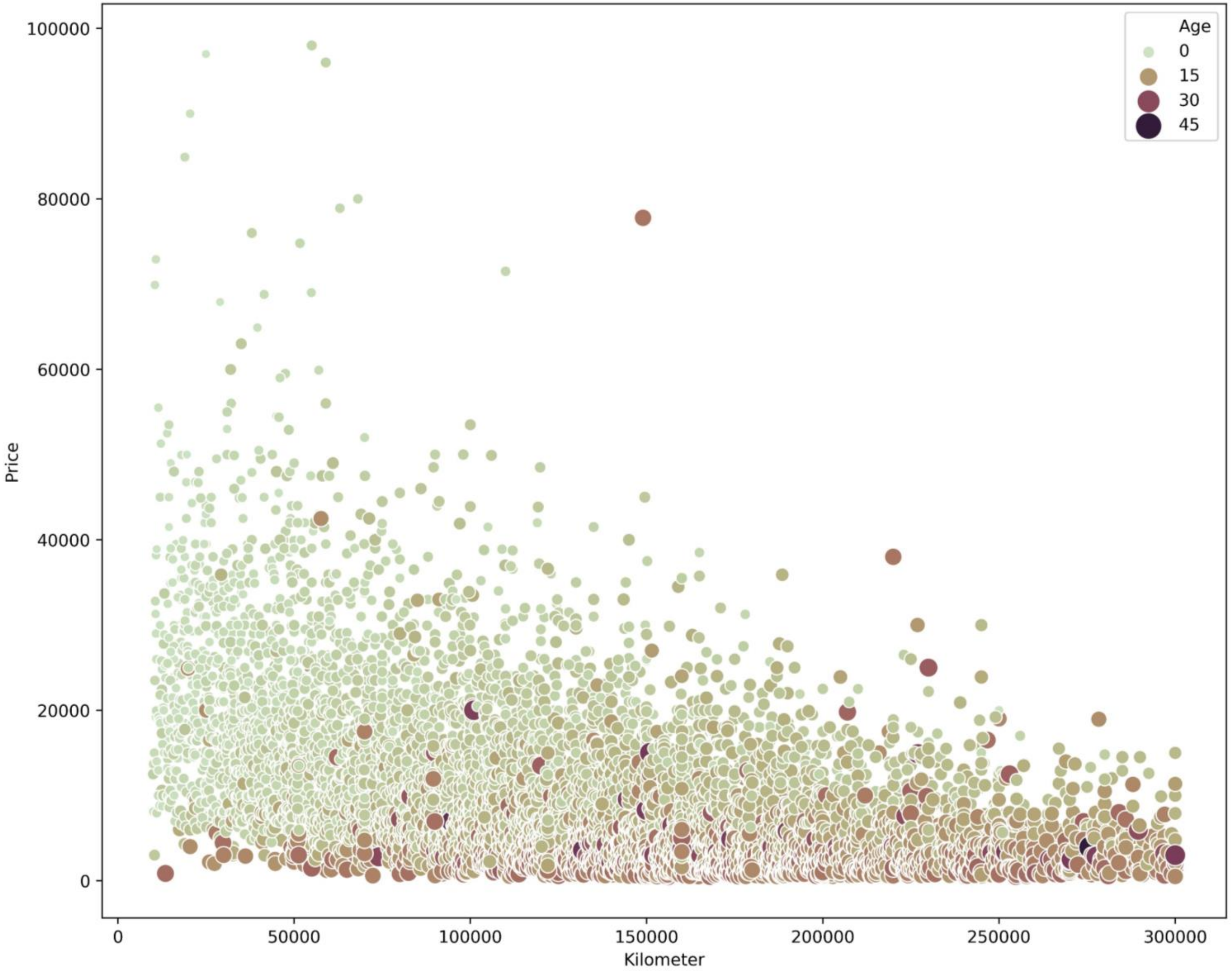
matplotlib

seaborn

Methodology

- **EDA**

```
<class '
Int64Ind
Data col
 #   Col
---  ---
 0   Bra
 1   Mod
 2   Bod
 3   Age
 4   Hp
 5   Gea
 6   Fue
 7   Con
 8   CO2
 9   Km
 10  Pri
dtypes:
memory u
```



Price
0.000000
6.250822
0.568046
0.000000
0.000000
9.000000
0.000000
0.000000

python™

jupyter

pandas

NumPy

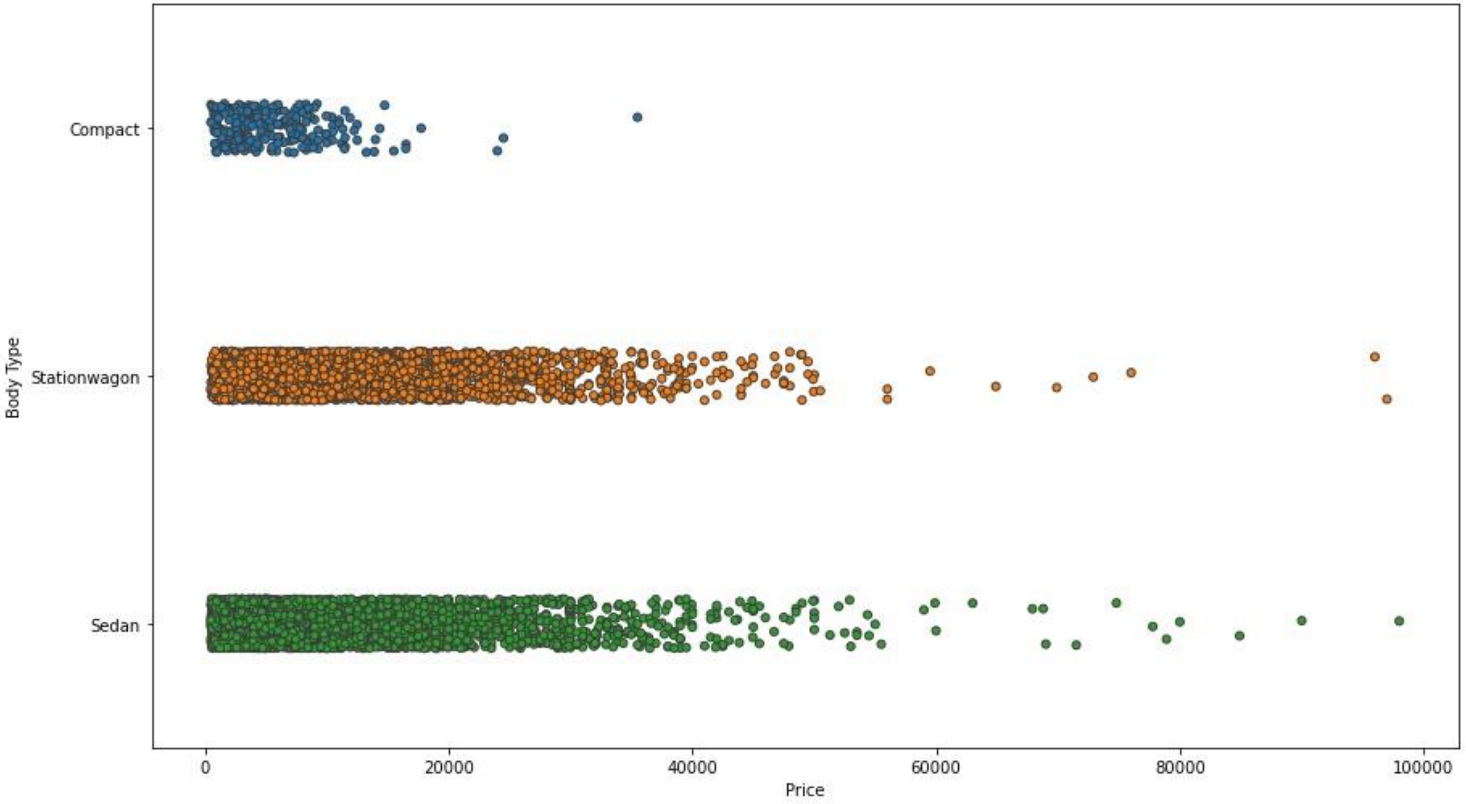
matplotlib

seaborn

Methodology

- **EDA**

```
<clas
Int64
Data
 #
---
 0
 1
 2
 3
 4
 5
 6
 7
 8
 9
 10
dtype
memor
```



Price
7910.000000
10786.250822
9660.568046
500.000000
3850.000000
7999.000000
14950.000000
98000.000000

python

jupyter

pandas

NumPy

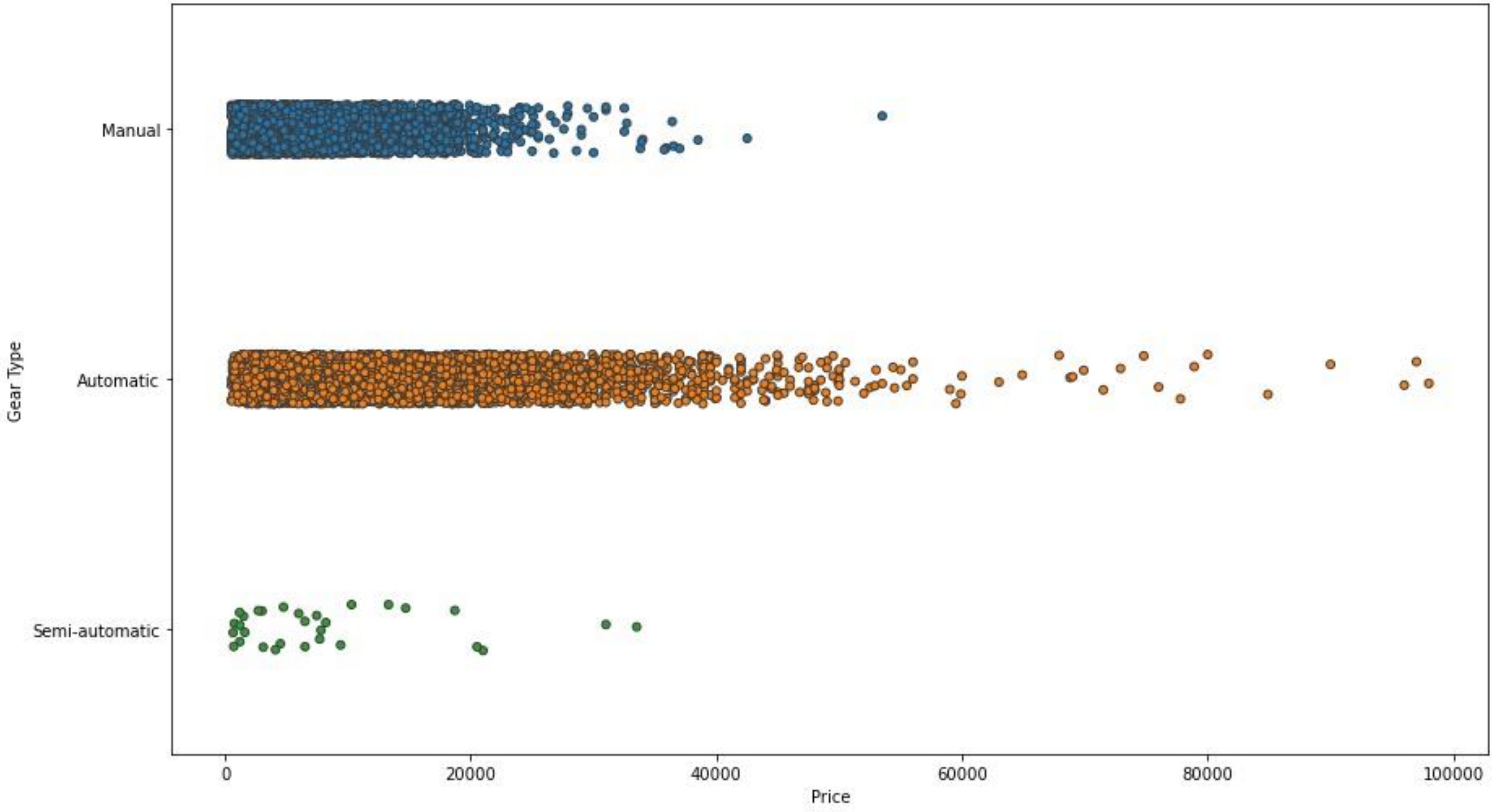
matplotlib

seaborn

Methodology

- **EDA**

```
<cla
Int6
Data
 #
---
 0
 1
 2
 3
 4
 5
 6
 7
 8
 9
 10
dtyp
memo
```

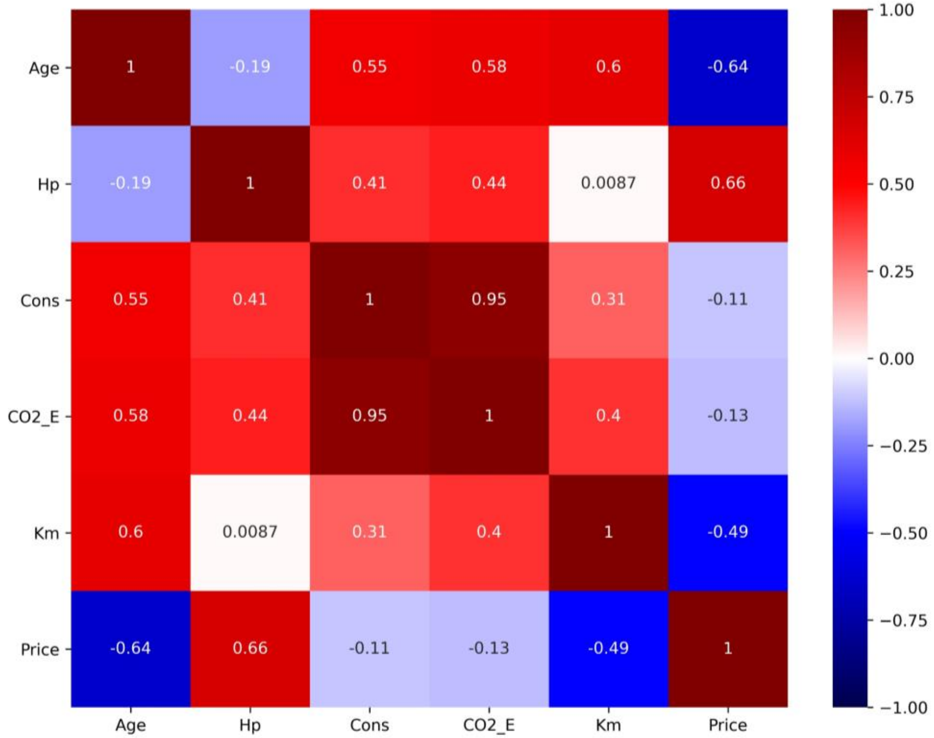


Price
00000
50822
68046
00000
00000
00000
00000
00000

Methodology

- **Regression Modelling (Correlation)**

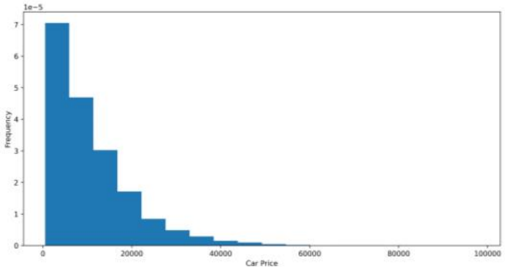
	Age	Hp	Cons	CO2_E	Km	Price
Age	1.000000	-0.194873	0.548982	0.575009	0.595268	-0.640110
Hp	-0.194873	1.000000	0.411986	0.442771	0.008664	0.661239
Cons	0.548982	0.411986	1.000000	0.949418	0.306613	-0.114067
CO2_E	0.575009	0.442771	0.949418	1.000000	0.399440	-0.125840
Km	0.595268	0.008664	0.306613	0.399440	1.000000	-0.494650
Price	-0.640110	0.661239	-0.114067	-0.125840	-0.494650	1.000000

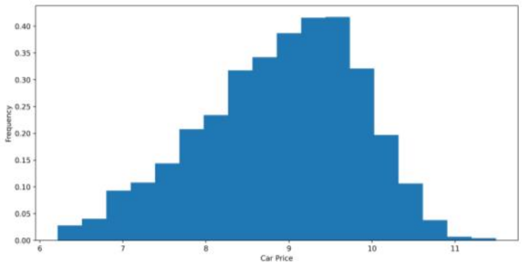


Methodology

- **Regression Modelling (Feature Engineering)**

Price → log Price





Dummy → Brand (+15 Col.) , Body Type (+3 Col.) ,
Gear (+3 Col.)

Label Encoding → Fuel Type

Drop Columns → Low Corr. (Cons., CO2 Em.)

Results

- **Train Val Test Analysis** *(Ridge alpha=0,5)*

Numeric Col.

```
Linear Regression train R^2: 0.773
Linear Regression val R^2: 0.779
Linear Regression test R^2: 0.789
Degree 2 polynomial regression train R^2: 0.852
Degree 2 polynomial regression val R^2: 0.860
Degree 2 polynomial regression test R^2: 0.855
Degree 2 Ridge polynomial regression train R^2: 0.852
Degree 2 Ridge polynomial regression val R^2: 0.860
Degree 2 Ridge polynomial regression test R^2: 0.855
Degree 3 polynomial regression train R^2: 0.734
Degree 3 polynomial regression val R^2: 0.716
Degree 3 polynomial regression test R^2: -5.904
Degree 3 Ridge polynomial regression train R^2: 0.863
Degree 3 Ridge polynomial regression val R^2: 0.867
Degree 3 Ridge polynomial regression test R^2: 0.749
Degree 4 polynomial regression train R^2: -1.766
Degree 4 polynomial regression val R^2: -2.313
Degree 4 polynomial regression test R^2: -66.575
Degree 4 Ridge polynomial regression train R^2: 0.873
Degree 4 Ridge polynomial regression val R^2: 0.849
Degree 4 Ridge polynomial regression test R^2: -12.818
Degree 5 polynomial regression train R^2: -0.194
Degree 5 polynomial regression val R^2: -0.404
Degree 5 polynomial regression test R^2: -0.150
Degree 5 Ridge polynomial regression train R^2: 0.888
Degree 5 Ridge polynomial regression val R^2: -135.101
Degree 5 Ridge polynomial regression test R^2: -5997.215
Ridge Regression train R^2: 0.773
Ridge Regression val R^2: 0.779
Ridge Regression test R^2: 0.789
Lasso Regression train R^2: 0.773
Lasso Regression val R^2: 0.779
Lasso Regression test R^2: 0.789
```

Results

- **Train Val Test Analysis** *(Ridge alpha=0,5)*

Numeric Col.

```
Linear Regression train R^2: 0.773
Linear Regression val R^2: 0.779
Linear Regression test R^2: 0.789
Degree 2 polynomial regression train R^2: 0.852
Degree 2 polynomial regression val R^2: 0.860
Degree 2 polynomial regression test R^2: 0.855
Degree 2 Ridge polynomial regression train R^2: 0.852
Degree 2 Ridge polynomial regression val R^2: 0.860
Degree 2 Ridge polynomial regression test R^2: 0.855
Degree 3 polynomial regression train R^2: 0.734
Degree 3 polynomial regression val R^2: 0.716
Degree 3 polynomial regression test R^2: -5.904
Degree 3 Ridge polynomial regression train R^2: 0.863
Degree 3 Ridge polynomial regression val R^2: 0.867
Degree 3 Ridge polynomial regression test R^2: 0.749
Degree 4 polynomial regression train R^2: -1.766
Degree 4 polynomial regression val R^2: -2.313
Degree 4 polynomial regression test R^2: -66.575
Degree 4 Ridge polynomial regression train R^2: 0.873
Degree 4 Ridge polynomial regression val R^2: 0.849
Degree 4 Ridge polynomial regression test R^2: -12.818
Degree 5 polynomial regression train R^2: -0.194
Degree 5 polynomial regression val R^2: -0.404
Degree 5 polynomial regression test R^2: -0.150
Degree 5 Ridge polynomial regression train R^2: 0.888
Degree 5 Ridge polynomial regression val R^2: -135.101
Degree 5 Ridge polynomial regression test R^2: -5997.215
Ridge Regression train R^2: 0.773
Ridge Regression val R^2: 0.779
Ridge Regression test R^2: 0.789
Lasso Regression train R^2: 0.773
Lasso Regression val R^2: 0.779
Lasso Regression test R^2: 0.789
```

After Feature Eng.

```
Linear Regression train R^2: 0.869
Linear Regression val R^2: 0.870
Linear Regression test R^2: 0.866
Degree 2 polynomial regression train R^2: 0.907
Degree 2 polynomial regression val R^2: 0.899
Degree 2 polynomial regression test R^2: 0.901
Degree 2 Ridge polynomial regression train R^2: 0.907
Degree 2 Ridge polynomial regression val R^2: 0.899
Degree 2 Ridge polynomial regression test R^2: 0.902
Degree 3 polynomial regression train R^2: 0.922
Degree 3 polynomial regression val R^2: 0.871
Degree 3 polynomial regression test R^2: 0.895
Degree 3 Ridge polynomial regression train R^2: 0.927
Degree 3 Ridge polynomial regression val R^2: -273.843
Degree 3 Ridge polynomial regression test R^2: -105.475
Degree 4 polynomial regression train R^2: 0.925
Degree 4 polynomial regression val R^2: 0.751
Degree 4 polynomial regression test R^2: 0.878
Degree 4 Ridge polynomial regression train R^2: -13.760
Degree 4 Ridge polynomial regression val R^2: -13.473
Degree 4 Ridge polynomial regression test R^2: -7.705
Degree 5 polynomial regression train R^2: 0.901
Degree 5 polynomial regression val R^2: 0.416
Degree 5 polynomial regression test R^2: 0.733
Degree 5 Ridge polynomial regression train R^2: -12.936
Degree 5 Ridge polynomial regression val R^2: -12.839
Degree 5 Ridge polynomial regression test R^2: -9.920
Ridge Regression train R^2: 0.869
Ridge Regression val R^2: 0.870
Ridge Regression test R^2: 0.866
Lasso Regression train R^2: 0.803
Lasso Regression val R^2: 0.799
Lasso Regression test R^2: 0.798
```

Results

- **Train Val Test Analysis** *(Ridge alpha=0,5)*

Numeric Col.

```
Linear Regression train R^2: 0.773
Linear Regression val R^2: 0.779
Linear Regression test R^2: 0.789
Degree 2 polynomial regression train R^2: 0.852
Degree 2 polynomial regression val R^2: 0.860
Degree 2 polynomial regression test R^2: 0.855
Degree 2 Ridge polynomial regression train R^2: 0.852
Degree 2 Ridge polynomial regression val R^2: 0.860
Degree 2 Ridge polynomial regression test R^2: 0.855
Degree 3 polynomial regression train R^2: 0.734
Degree 3 polynomial regression val R^2: 0.716
Degree 3 polynomial regression test R^2: -5.904
Degree 3 Ridge polynomial regression train R^2: 0.863
Degree 3 Ridge polynomial regression val R^2: 0.867
Degree 3 Ridge polynomial regression test R^2: 0.749
Degree 4 polynomial regression train R^2: -1.766
Degree 4 polynomial regression val R^2: -2.313
Degree 4 polynomial regression test R^2: -66.575
Degree 4 Ridge polynomial regression train R^2: 0.873
Degree 4 Ridge polynomial regression val R^2: 0.849
Degree 4 Ridge polynomial regression test R^2: -12.818
Degree 5 polynomial regression train R^2: -0.194
Degree 5 polynomial regression val R^2: -0.404
Degree 5 polynomial regression test R^2: -0.150
Degree 5 Ridge polynomial regression train R^2: 0.888
Degree 5 Ridge polynomial regression val R^2: -135.101
Degree 5 Ridge polynomial regression test R^2: -5997.215
Ridge Regression train R^2: 0.773
Ridge Regression val R^2: 0.779
Ridge Regression test R^2: 0.789
Lasso Regression train R^2: 0.773
Lasso Regression val R^2: 0.779
Lasso Regression test R^2: 0.789
```

After Feature Eng.

```
Linear Regression train R^2: 0.869
Linear Regression val R^2: 0.870
Linear Regression test R^2: 0.866
Degree 2 polynomial regression train R^2: 0.907
Degree 2 polynomial regression val R^2: 0.899
Degree 2 polynomial regression test R^2: 0.901
Degree 2 Ridge polynomial regression train R^2: 0.907
Degree 2 Ridge polynomial regression val R^2: 0.899
Degree 2 Ridge polynomial regression test R^2: 0.902
Degree 3 polynomial regression train R^2: 0.922
Degree 3 polynomial regression val R^2: 0.871
Degree 3 polynomial regression test R^2: 0.895
Degree 3 Ridge polynomial regression train R^2: 0.927
Degree 3 Ridge polynomial regression val R^2: -273.843
Degree 3 Ridge polynomial regression test R^2: -105.475
Degree 4 polynomial regression train R^2: 0.925
Degree 4 polynomial regression val R^2: 0.751
Degree 4 polynomial regression test R^2: 0.878
Degree 4 Ridge polynomial regression train R^2: -13.760
Degree 4 Ridge polynomial regression val R^2: -13.473
Degree 4 Ridge polynomial regression test R^2: -7.705
Degree 5 polynomial regression train R^2: 0.901
Degree 5 polynomial regression val R^2: 0.416
Degree 5 polynomial regression test R^2: 0.733
Degree 5 Ridge polynomial regression train R^2: -12.936
Degree 5 Ridge polynomial regression val R^2: -12.839
Degree 5 Ridge polynomial regression test R^2: -9.920
Ridge Regression train R^2: 0.869
Ridge Regression val R^2: 0.870
Ridge Regression test R^2: 0.866
Lasso Regression train R^2: 0.803
Lasso Regression val R^2: 0.799
Lasso Regression test R^2: 0.798
```

- **CV Analysis** *(Ridge alpha=0,5)*

After Feature Eng.

```
Linear Regression CV train : 0.8673945645622518
Degree 2 Polynomial CV train : 0.895507616259426
Degree 2 Polynomial Ridge CV train : 0.8963240770819392
Degree 3 Polynomial CV train : 0.8284630141130668
Degree 3 Polynomial Ridge CV train : -5376.268427676414
Degree 4 Polynomial CV train : 0.24871446919684445
Degree 4 Polynomial Ridge CV train : -0.054064130521525984
Degree 5 Polynomial CV train : -0.6731759091070364
Degree 5 Polynomial Ridge CV train : -0.6300491689780677
Ridge CV train : 0.867400478768637
Lasso CV train : 0.8034767893908513
```

- **MAE Analysis** *(Ridge alpha=0,5)*

After Feature Eng.

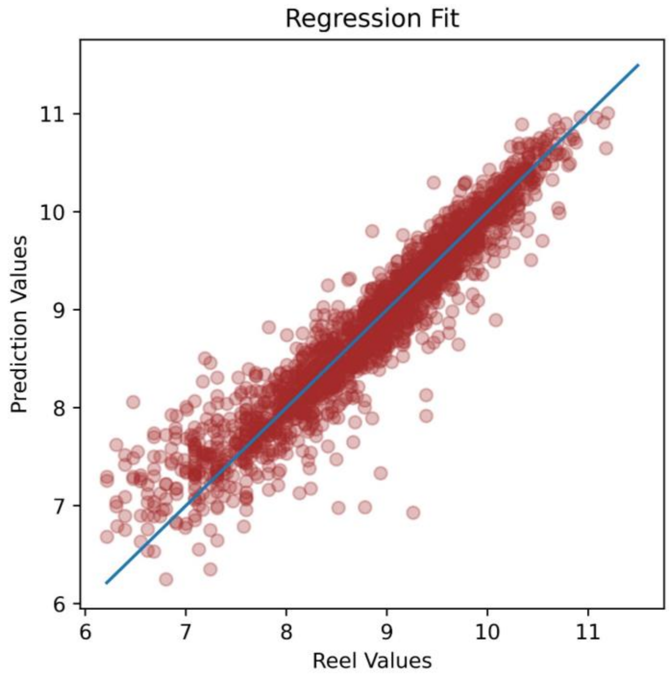
```
Linear Regression MAE : 0.23522432179014613
Degree 2 Polynomial MAE : 0.21659716304903634
Degree 2 Poly Ridge MAE : 0.2144138926834761
Degree 3 Polynomial MAE : 0.25914612388703795
Degree 3 Poly Ridge MAE : 1.6651563929495823
Ridge MAE : 0.23521255436686897
Lasso MAE : 0.32350936802981106
```

Conclusions

- *We choose* ***2nd degree polynomial ridge regression***

with ***R^2 Score : 0.8963*** *of* ***CV*** *&* ***0.2144*** *of* ***MAE.***

Reg fit plot:

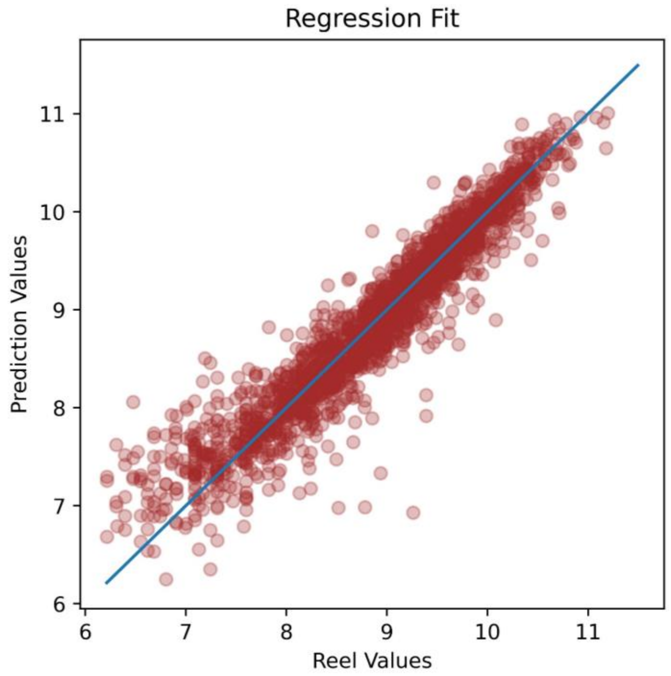


Conclusions

- *We choose* ***2nd degree polynomial ridge regression***

with ***R^2 Score : 0.8963*** *of* ***CV*** *&* ***0.2144*** *of* ***MAE.***

Reg fit plot:



Future Works

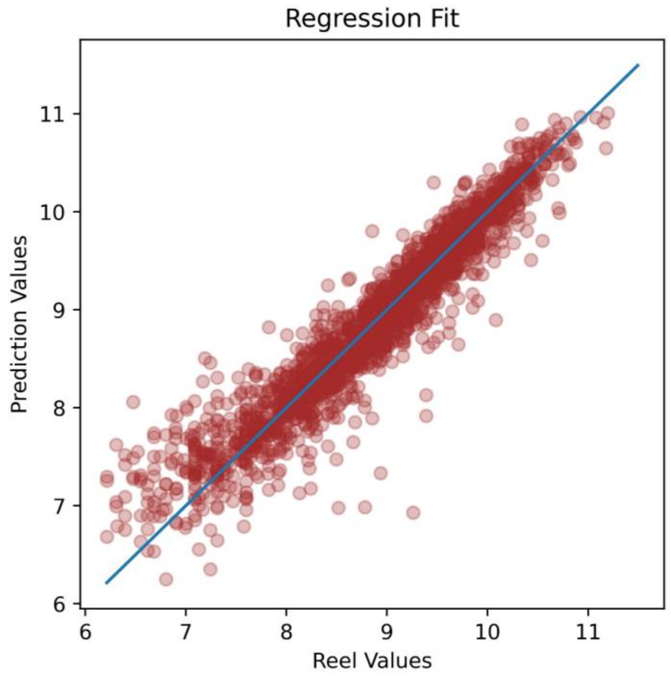
- ***Improving feature engineering,***
- ***Getting higher score,***
- ***Collecting more data.***

Conclusions

- *We choose* ***2nd degree polynomial ridge regression***

with ***R^2 Score : 0.8963*** *of* ***CV*** *&* ***0.2144*** *of* ***MAE.***

Reg fit plot:



Future Works

- ***Improving feature engineering,***
- ***Getting higher score,***
- ***Collecting more data,***
- ***Productising.***

Thank You!