Predictive Analysis – Electrical car

Yousof Rahimian

Master of Data Science

Bellevue University

DSC 630 T302

(Individual Course Project)

Fall,2022

Project Overview

In this project, I am going to build a machine learning model to make this task automated with the best accuracy possible using python. I am planning to build a model that can predict the following question.

- How does price of cars relate to variables and find out the more important of item on price?

What is an Electric Car?

Electric cars have actually been around for longer than gasoline-powered cars. The first electric vehicle was a motorized carriage created by Scottish inventor Robert Anderson in the early 1830s. Unfortunately, the battery couldn't be recharged, so it was a bit of a novelty.

Rechargeable batteries appeared in 1859 and in 1884, a man named Thomas Parker built a prototype electric vehicle. A few years later in 1887, William Morrison patented his electric car in Des Moines, Iowa, and the electric race was on.

While many companies tried their hand at putting EVs on the market, Henry Ford won the battle with his cheap-to-produce Model T and the world went gasoline car crazy... until now.

EVs first came into existence in the mid-19th century, when electricity was among the preferred methods for motor vehicle propulsion, providing a level of comfort and ease of operation that could not be achieved by the gasoline cars of the time. Internal combustion engines were the dominant propulsion method for cars and trucks for about 100 years, but electric power remained commonplace in other vehicle types, such as trains and smaller vehicles of all types.

In the 21st century, EVs have seen a resurgence due to technological developments, and an increased focus on renewable energy and the potential reduction of transportation's impact on climate change and other environmental issues. Project Drawdown describes electric vehicles as one of the 100 best contemporary solutions for addressing climate change.

Data Source

The data has been source from Kaggle and have consist of the following(I just list a few notable columns):

Brand of the vehicle: This column contains brand of cars and their models.

Top speed: This column contains top speed of cars

Range/Km: This column contains range of cars based on kilometer per hours

Efficiency: This column contains efficiency of cars

FastCharge_KmH: This column contains fast charge of batterie of cars

Price: This column contains price Euro of cars

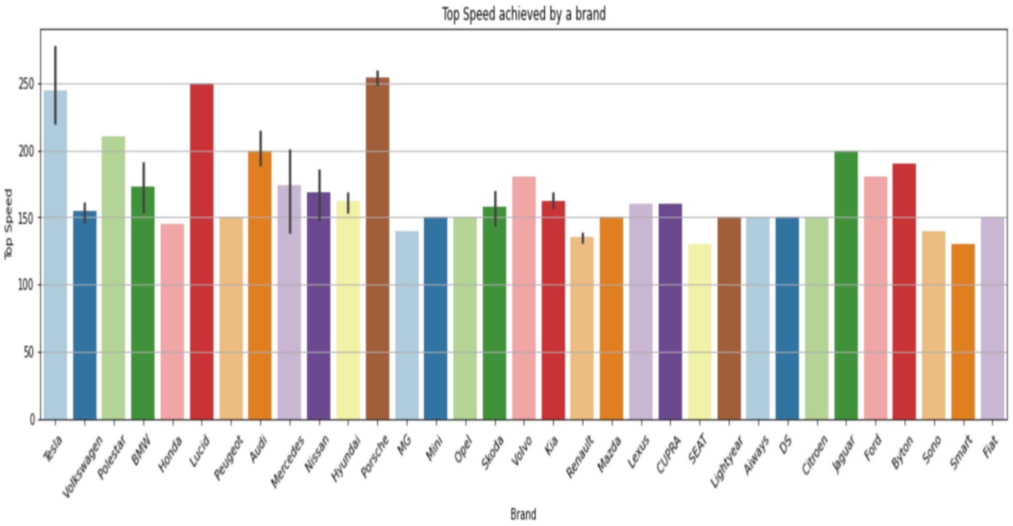
The models will be analyzed by selecting metrics that take into R-square or adjust R-square metrics. For this project I believe it will be most impactful to divide the data into a training set and a test set. Once I use a training set to create a model, I will then test the model on the test set of data. I would evaluate the results base on the 5 questions I've mentioned above, which will evaluate all electrical car company based on the available datasets. Mean square error or mean absolute error are two metrics to evaluate a regression model performance during these analyses.

The car industry is undergoing a radical transformation, with most carmakers agreeing the next 10 years will bring more change than the two previous decades. I hope to learn more about the liner regression model during this analysis to realized which variables will have the greatest impact on the price to get the best decision for buying the reliable and convenient EVs care in future. I believe almost all the variables will affect the price, but I will use the model to find out the most importantly effective variables to rich out the best price.

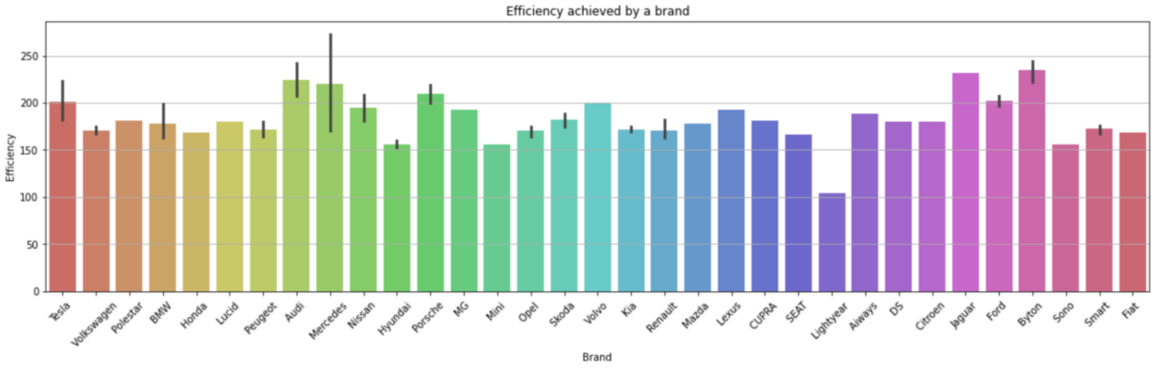
Pairplot, Heatmap, Pie Chart, and Barplot are four data visualization techniques which I've used in this graphical section.

I used Seaborn Pairplot to get the relation between each and every variable present in the data frame. Heatmap technique representing the correlation coefficient in variables and find out the highly coefficient correlated variables to focus on valuable variables. Pichart is dividing the cars and body in the circular statistical graphic to illustrate the variety of available body style in this data set. Barplot techniques show the relationship between a numeric and a categorical variable and compare the fastest versus slowest car speed, efficiency, reliability, and price of vehicles which are the most items I will work on during the analysis.

Porsche, Lucid, and Tesla produce the fastest cars and Smart the lowest one.



Byton, Jaguar, and Audi are the most efficient and Lightyear the least which show the highest efficiency.



I've chosen the Linear Regression model to describe the relation between top speed and efficiency which are the two main variables related to price in this analysis.

I've explored and cleaned the data, finding out the number of null values, visualize the data, trained, and find out the accuracy of model. Below is the list of things that have done:

Import and clean the data

Pairplot of all the columns based on Brand presence

Heatmap to show the correlation of the data

Build and evaluate the models

Use Linear regression Model

Regression Coefficients

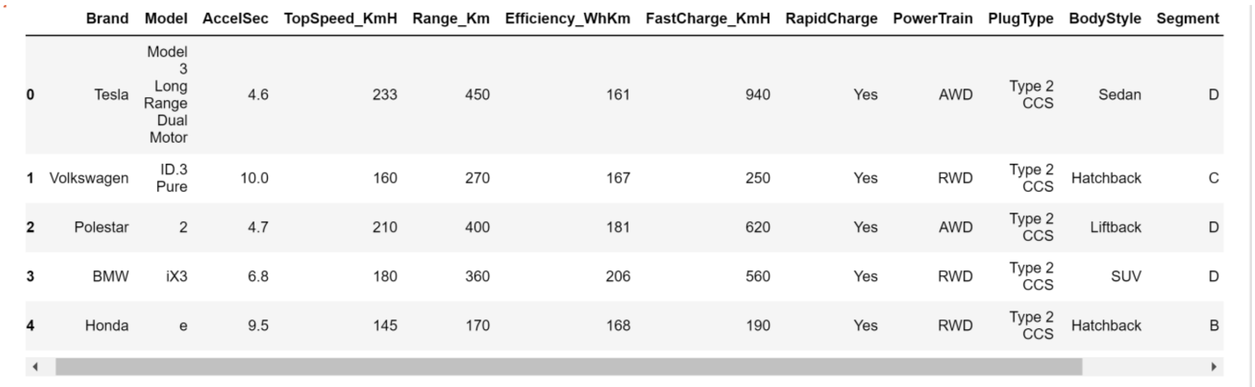
Logistic Regression

Confusion Matrix of the regression

Finding out the accuracy score

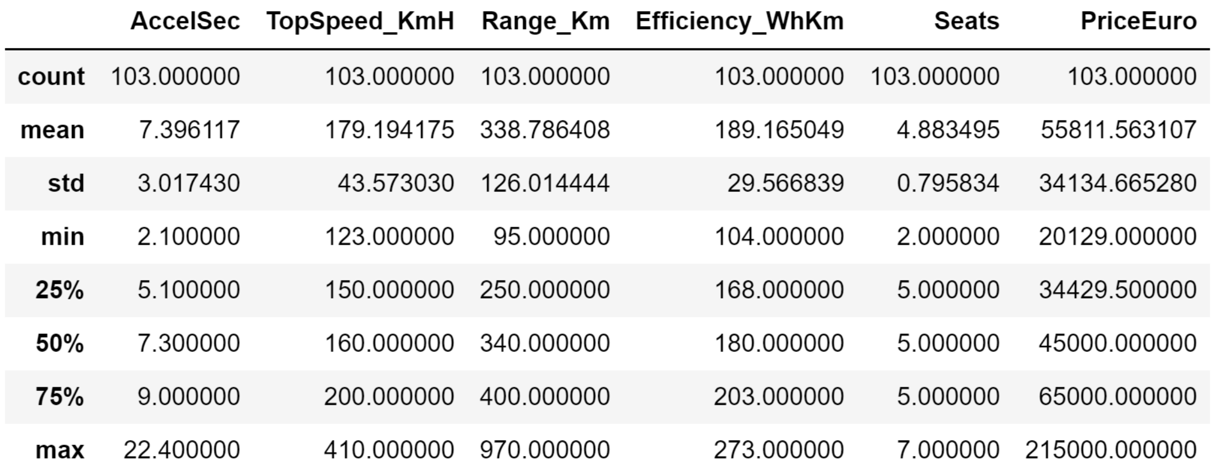
I will illustrate the process of the work in the following with screenshot of each section.

Data



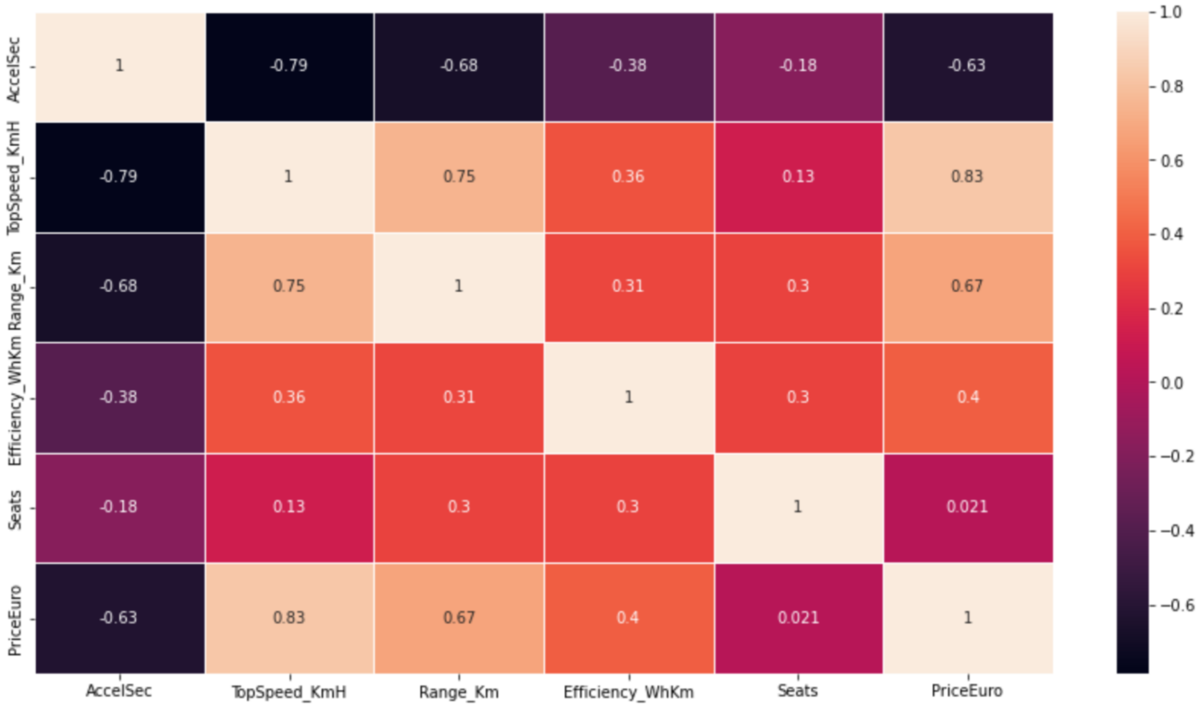
	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment
0	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	Yes	AWD	Type 2 CCS	Sedan	D
1	Volkswagen	ID.3 Pure	10.0	160	270	167	250	Yes	RWD	Type 2 CCS	Hatchback	C
2	Polestar	2	4.7	210	400	181	620	Yes	AWD	Type 2 CCS	Liftback	D
3	BMW	iX3	6.8	180	360	206	560	Yes	RWD	Type 2 CCS	SUV	D
4	Honda	e	9.5	145	170	168	190	Yes	RWD	Type 2 CCS	Hatchback	B

Descriptive Statistics of the dataset



	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	Seats	PriceEuro
count	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000
mean	7.396117	179.194175	338.786408	189.165049	4.883495	55811.563107
std	3.017430	43.573030	126.014444	29.566839	0.795834	34134.665280
min	2.100000	123.000000	95.000000	104.000000	2.000000	20129.000000
25%	5.100000	150.000000	250.000000	168.000000	5.000000	34429.500000
50%	7.300000	160.000000	340.000000	180.000000	5.000000	45000.000000
75%	9.000000	200.000000	400.000000	203.000000	5.000000	65000.000000
max	22.400000	410.000000	970.000000	273.000000	7.000000	215000.000000

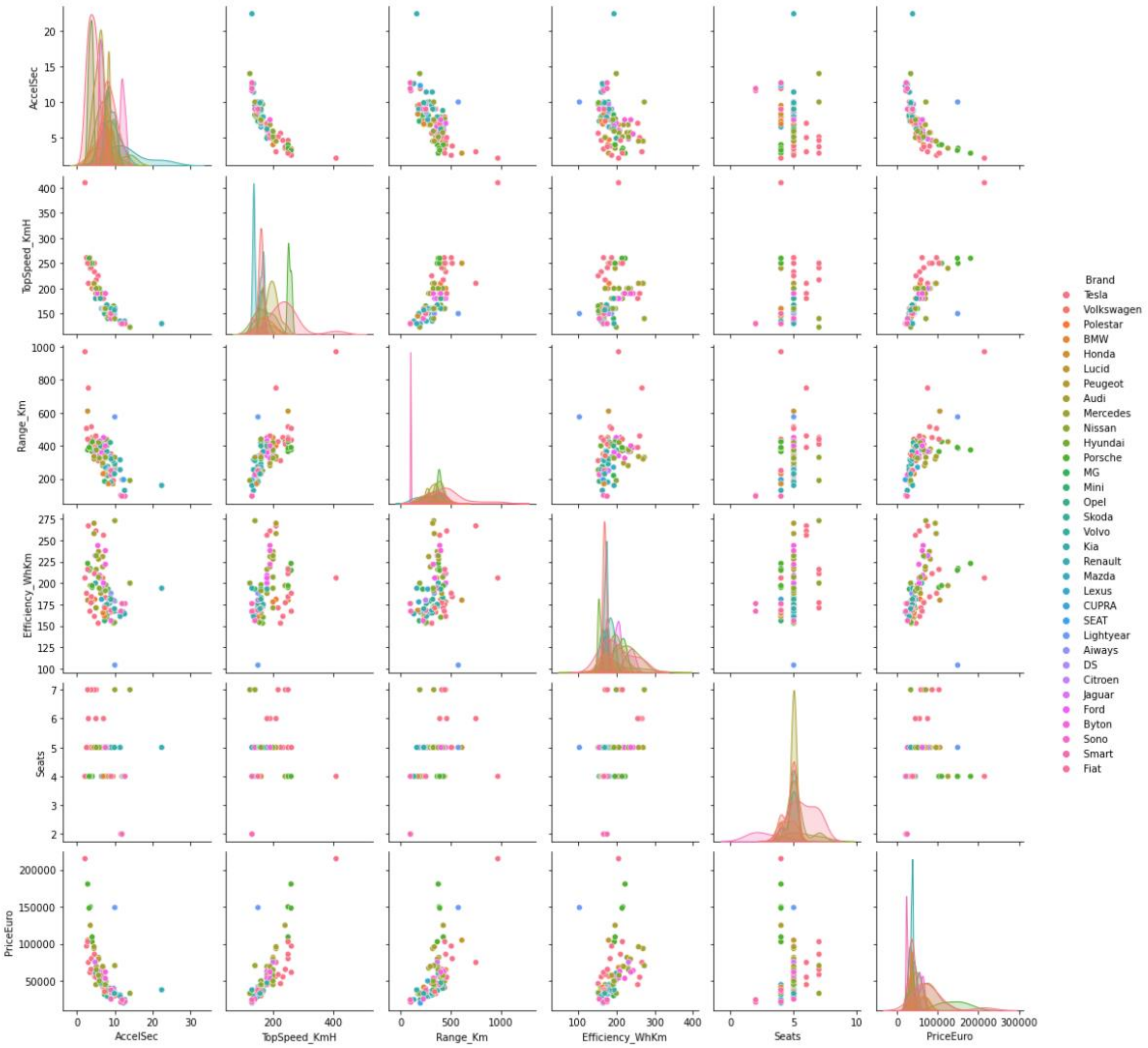
Heatmap to show the correlation of the dat

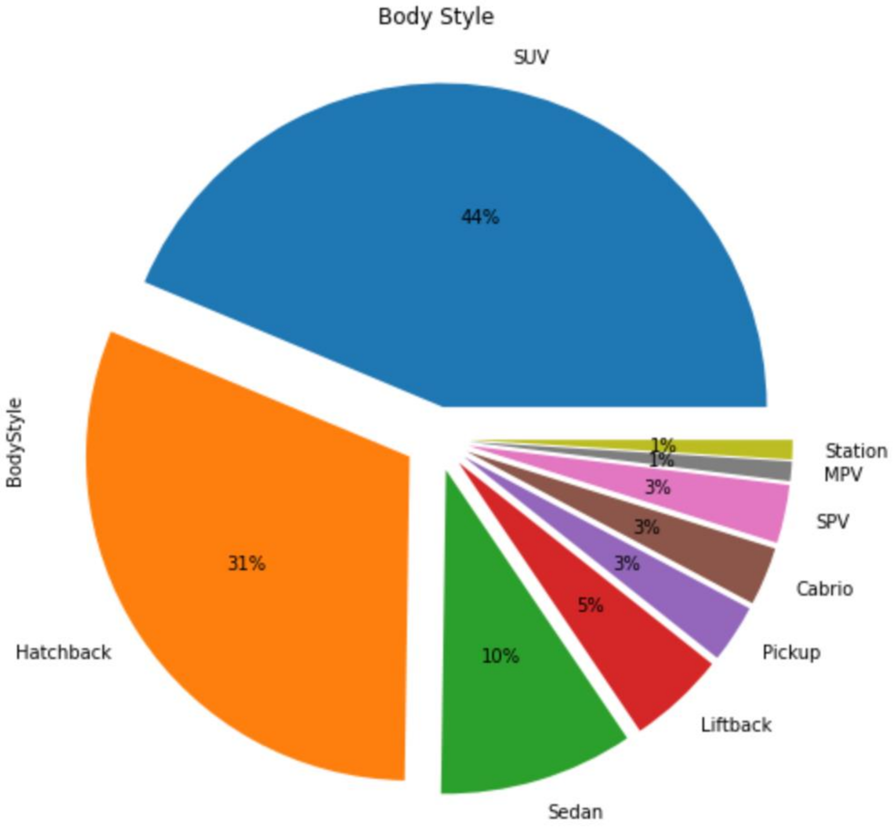


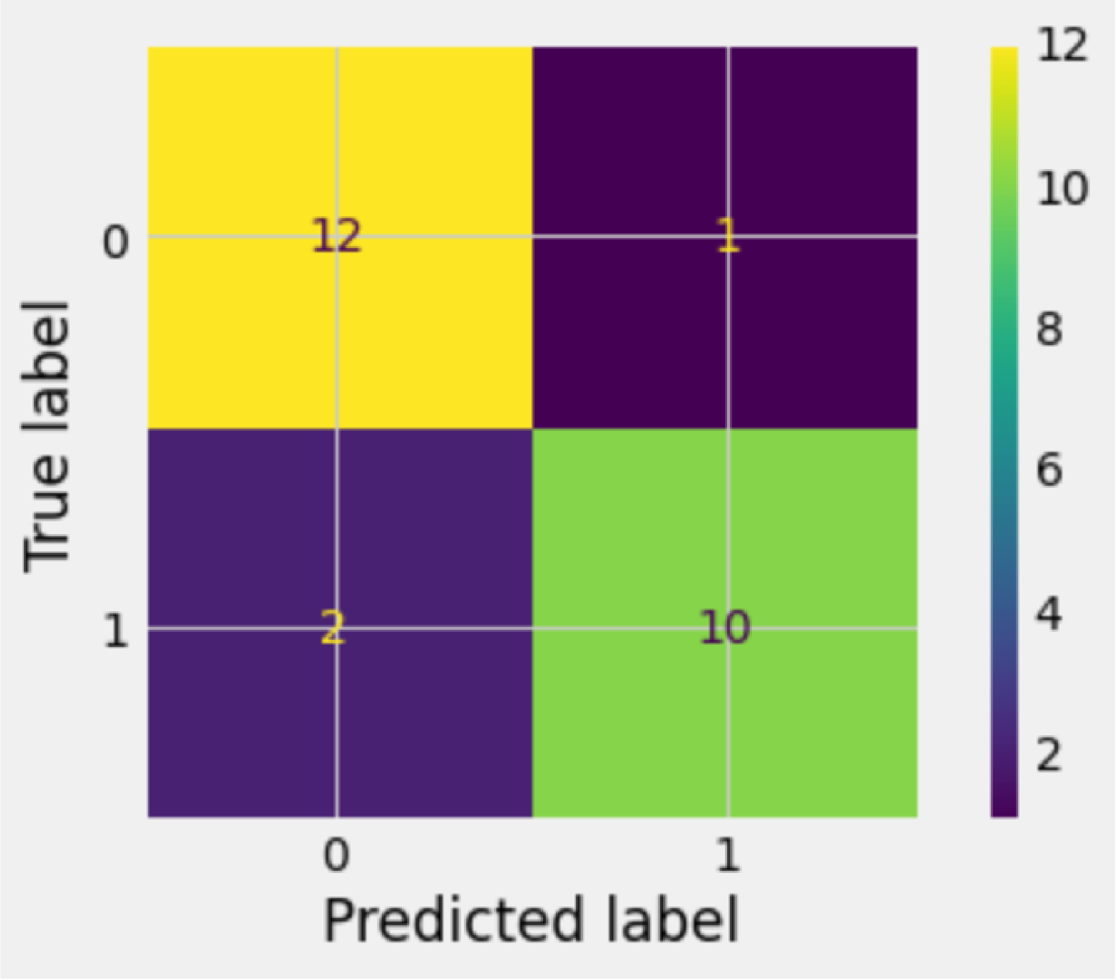
	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	Seats	PriceEuro
AccelSec	1	-0.79	-0.68	-0.38	-0.18	-0.63
TopSpeed_KmH	-0.79	1	0.75	0.36	0.13	0.83
Range_Km	-0.68	0.75	1	0.31	0.3	0.67
Efficiency_WhKm	-0.38	0.36	0.31	1	0.3	0.4
Seats	-0.18	0.13	0.3	0.3	1	0.021
PriceEuro	-0.63	0.83	0.67	0.4	0.021	1

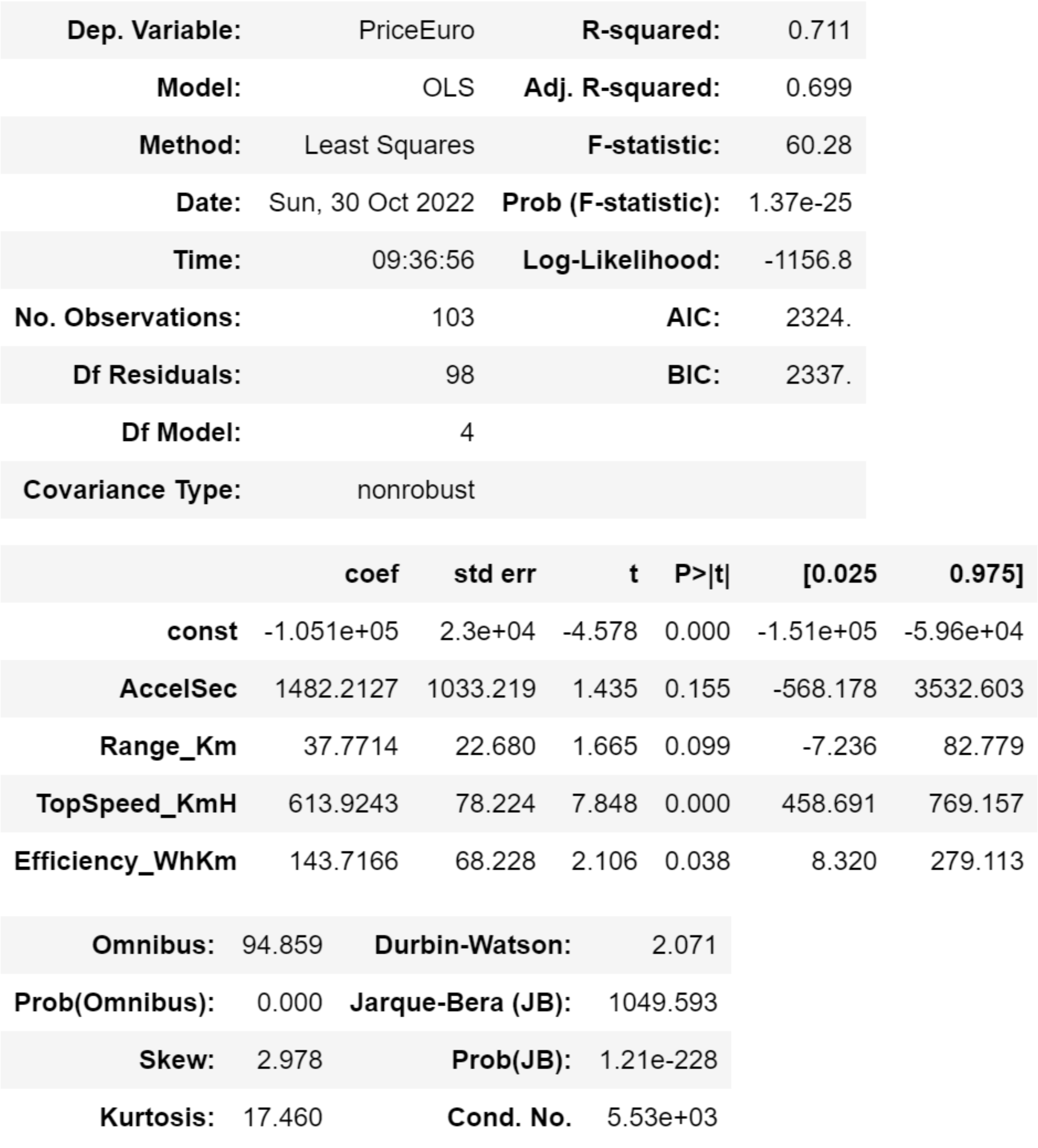
Pairplot of all the columns based on Brand presence

From the picture below, we can observe the variations in each plot. The plots are in matrix format where the row name represents x axis and column name represents the y axis. The main-diagonal subplots are the univariate histograms (distributions) for each attribute. So, in this pareplot TopSpeed_KmH is highly correlation with PriceEuro and AccelSec.




Body Style
SUV
44%
BodyStyle
31%
Hatchback
10%
Sedan
5%
Liftback
3%
Pickup
3%
Cabrio
3%
SPV
1%
MPV
1%
Station


12
1
2
10
0
1
True label
0
1
Predicted label
12
10
8
6
4
2



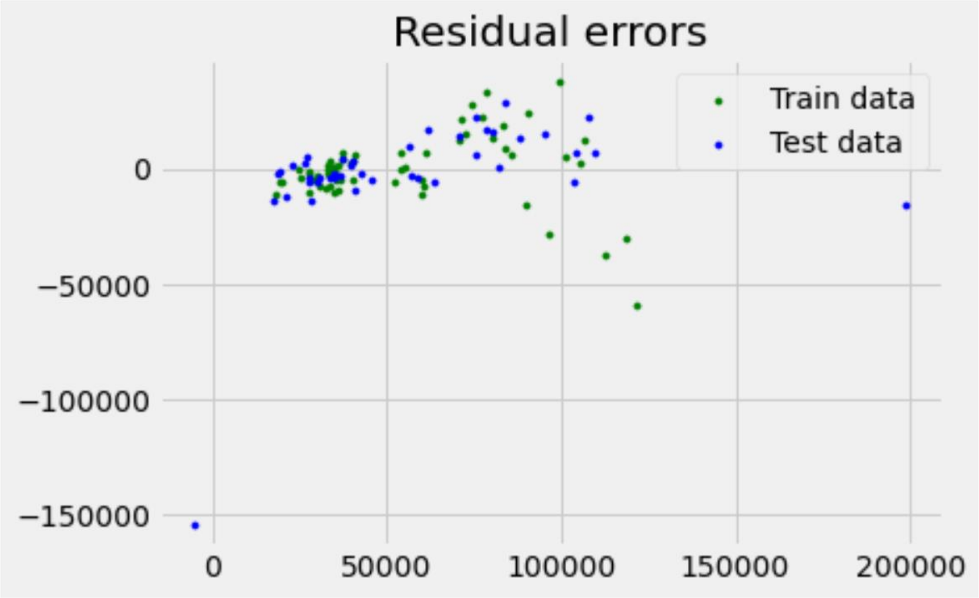
Dep. Variable:	PriceEuro	**R-squared:**	0.711
Model:	OLS	**Adj. R-squared:**	0.699
Method:	Least Squares	**F-statistic:**	60.28
Date:	Sun, 30 Oct 2022	**Prob (F-statistic):**	1.37e-25
Time:	09:36:56	**Log-Likelihood:**	-1156.8
No. Observations:	103	**AIC:**	2324.
Df Residuals:	98	**BIC:**	2337.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-1.051e+05	2.3e+04	-4.578	0.000	-1.51e+05	-5.96e+04
AccelSec	1482.2127	1033.219	1.435	0.155	-568.178	3532.603
Range_Km	37.7714	22.680	1.665	0.099	-7.236	82.779
TopSpeed_KmH	613.9243	78.224	7.848	0.000	458.691	769.157
Efficiency_WhKm	143.7166	68.228	2.106	0.038	8.320	279.113

Omnibus:	94.859	**Durbin-Watson:**	2.071
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1049.593
Skew:	2.978	**Prob(JB):**	1.21e-228
Kurtosis:	17.460	**Cond. No.**	5.53e+03

Only Top Speed and Efficieny are the two variables related to price

Linear regression model:



- In the above plot, I determine the accuracy score using Explained Variance Score.
- Variance score is around .5
- The best possible score is 1.0, lower values are worse

Summary:

We used several regression models to fit our data and it seems that they all succeeded to fit the data well and this indicated that the data preprocessing stage was also a success but we're still facing the problem of overfitting so I see that all models are truly promising and ready for the next stage of improvement to reduce overfitting.

We also noticed that the PCA could actually preserve the varience in data we reduced features number from 73 to 29 and still could manage a fair performance on our models with just a slight difference from the original dataset disregarding The huge overfitting with Linear and Lasso regression models on the reduced data.

Porsche, Lucid and Tesla produce the fastest cars and smart the lowest

Lightyear, Porsche and Lucid are the most expensive and SEAT and Smart the least

Byton, Jaguar and Audi are the most efficient and Lightyear the least

Around 78% of the dependent variable has been explained by the independent variables

Variance score is around .5, which the best possible score is 1.0, lower values are worse, and data is accurate up to 95%

References:

EVs - One Electric Vehicle Dataset - Smaller | Kaggle

EVs - One Electric Vehicle Dataset - Smaller | Kaggle

Electronics | Free Full-Text | Electric Vehicles: A Data Science Perspective Review | HTML (mdpi.com)

Data Analysis of Electric Vehicles: For Convenient and Smart Life – Hyundai Motor Group TECH

Using Data Science to Predict the Energy Consumption of Electric Vehicles | by Martin Smuts | Medium