

U1

Rodriguez González José Adrián

August 2024

1 Abstract

In data analysis we need to understand that every analysis problems requires its own time before to create a predictive model. This is because we need to verify if our model would it be useful. We have two problems to practice that it has been learned about data analysis. As first checkout it'll be a simple dataset that only have two features *xandy*. So at the first sight seems to be quick but before to create a predictive model, it'll be required to take a look into the data as first step. And also, we got another case about prediecte the price of cars.

2 Introduction

With the objective to checkout how does it work the linear regression and look out for troubles on the real life. It has been presented two datasets. The first one consist in sintetic data. And the second one is related wth the prices of cars, and involves several features and the main objective on these datasets is to create a predictor that may predict future data.

3 Related works

Some of the related works(by now, I'll just let some of the books and the documentation that I could to see)

Article	Year	Techniques	Data	Results
Consumers' preferences on the Swiss car market: A revealed preference approach	2019	hedonic pricing approach and linear regression	Were obtained on the auto-Schweiz website	How lastly the costumers have been prefered lighter cars than heaviear despite the increasing of the weight cars and how it is an important factor for the costumer the efficence on the fuel consumption
İkinci El Otomobil Fiyat Artışına Etki Eden Faktörlerin Yapısal Eşitlik Modeli ile Tespit Edilmesi: Van İli Örneği(Determination of the Factors Affecting the Used Car Price Increase with the Structural Equation Model: The Case of Van Province)	2023	AFA, and structured equation model	Were obtained due to surveys on the province of Van	It has been found how the several features as economics, marketing, strategies and supplying are correlated with the amount of prices, however the economics are is stronger against the other factors
Are Used Cars More Sustainable? Price Prediction Based on Linear Regression	2023	Linear regression	the data was obtained on a website	The linear regression has shown effective results and also it could help to find out that the make of a car influence the price of it, the transmission doesn't seem to have a relationship with the price car

4 Methods and materials

The material used for this analysis were the usage of python and its libraries for data analysis and machine learning:

- Numpy: For mathematical calculations
- Pandas: For handle datasets
- scipy: to evaluate statistical parameters
- Scikit-learn: to train our model and check more parameters related with the model chosen.

The methodology followed it's the mix of scientific methodology with the abstraction for a data scientist.

- Obtain the data
- Make an exploration into the data.
- Check various parameters from the dataset.(These parts involves most of the section of exploratory analysis, as also, this step gives several hypothesis to check out at the dataset)
- Now that we have our Hypothesis planted, and also, with the help of the last step that it can be plotted the data. Now it'll be cleaned the data, and for this section it inquires in several steps
 - Hypothesis proposal
 - Transformations (logarithm, square, box-cox, capping and flooring)
 - Check the metrics of skewness, R^2 , MSE , $RMSE$
 - Make the model of linear regression
- check its metrics
- Propose another hypothesis.

Also, the main models that has been studied were Linear regression and Knn regressor

4.1 Linear regression

When we try to represent something complex it is usual that it has to be create a structure that simplifies the process however, it must contain the enough data and values that can approach the behaviour of phenomenon. In science areas, it is commonly called model. A model can explain a complex phenomenon in simple terms that can be easily handled and understandable. One of the most simplest models that exist on data science is the linear regression. Knowing:

$$f : \mathbb{R} \rightarrow \mathbb{R}$$

we have:

$$f(x) = mx + b$$

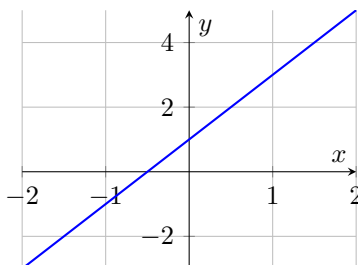


Figure 1: An example of a lineal function

This is a linear function(Figure 1), one of the simplest functions that can be viewed on calculus, sometimes a simple phenomenon can be viewed without getting in a huge amount of complexity and could be described with simple models, and when we have a problem or a situation in the reality when we need to create a machine hat can predict future data, what can we do? In those cases, as it has been portrayed, a linear function can be used as a mathematical feature that attempts to fit on the data. Despite that nowadays exist several

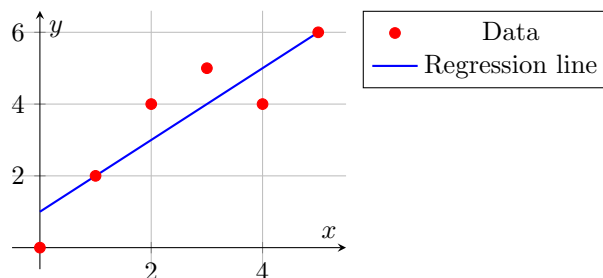


Figure 2: A simple example of a lineal regression

types of models that can predict future data, lineal regression as it is one of the simplest due to it just use a mathematical function that is easy to use, a lot of new models can be tested with a contraposition of lineal regression. So, the Lineal regression is described as

$$y = \beta_0 x + \beta_1 + e$$

β_0 it is the slope and β_1 is the intercept. And e is the error tha exist on the model.

However it'll be required to know some metrics that will help us to find out how efficient is our model and to know the error that we have in our model too

- R^2 : This metric can measure how much the model can describe the phenomenon. The $R^2 \in [0, 1]$ and if the model tends to 1 the model can describe better the phenomenon and can be more precise in the predictions.
- MSE and $RMSE$ = Are metrics that measure the distance between the value predict against the real value.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

The difference between MSE and RMSE are that MSE is susceptible to detect outliers values due to the nature of a square function, and also it helps to know the variance that we got in our model. Nevertheless, RMSE use the same scale of the values and want to measure the error as it is on the data.

4.2 Knn model

It consist in to look out on the data and predict a value according to the vecinity of the values nearest from the data that is presented. Knn presents two basci hyperparameters, the first one is the number of vecinity and the second is the type of distante that it'll be chosen. Knn model has some different parameters if we measure it as classifier. However at this case it'll be used the Knn model as a regressor, so it'll be usable the metrics that it has been described before. (figure 3) After knowing the main features that were used at this study, let's look out for the process that has been followed. As it'll have seen, the process aims to use the scientific method, however, with the feauters of a data analyst

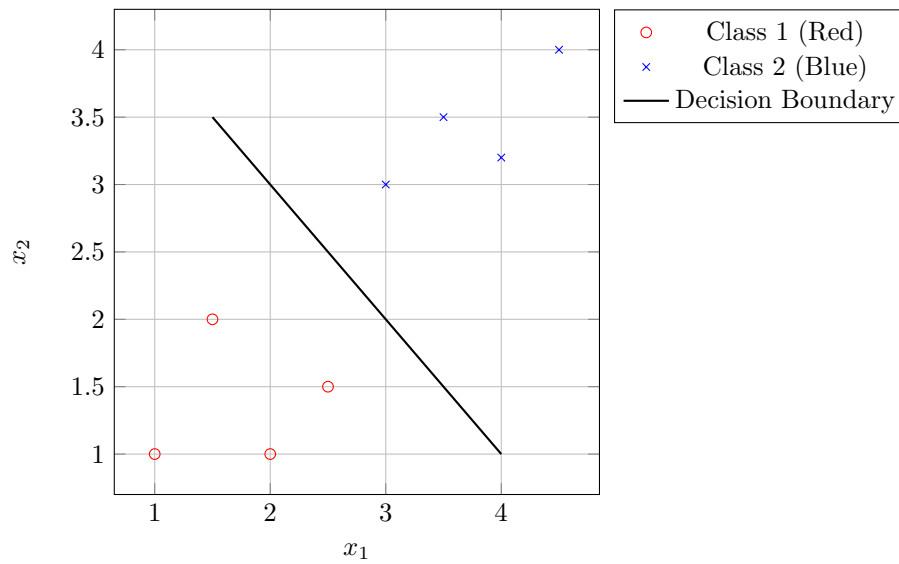


Figure 3: A simple example of knn regressor

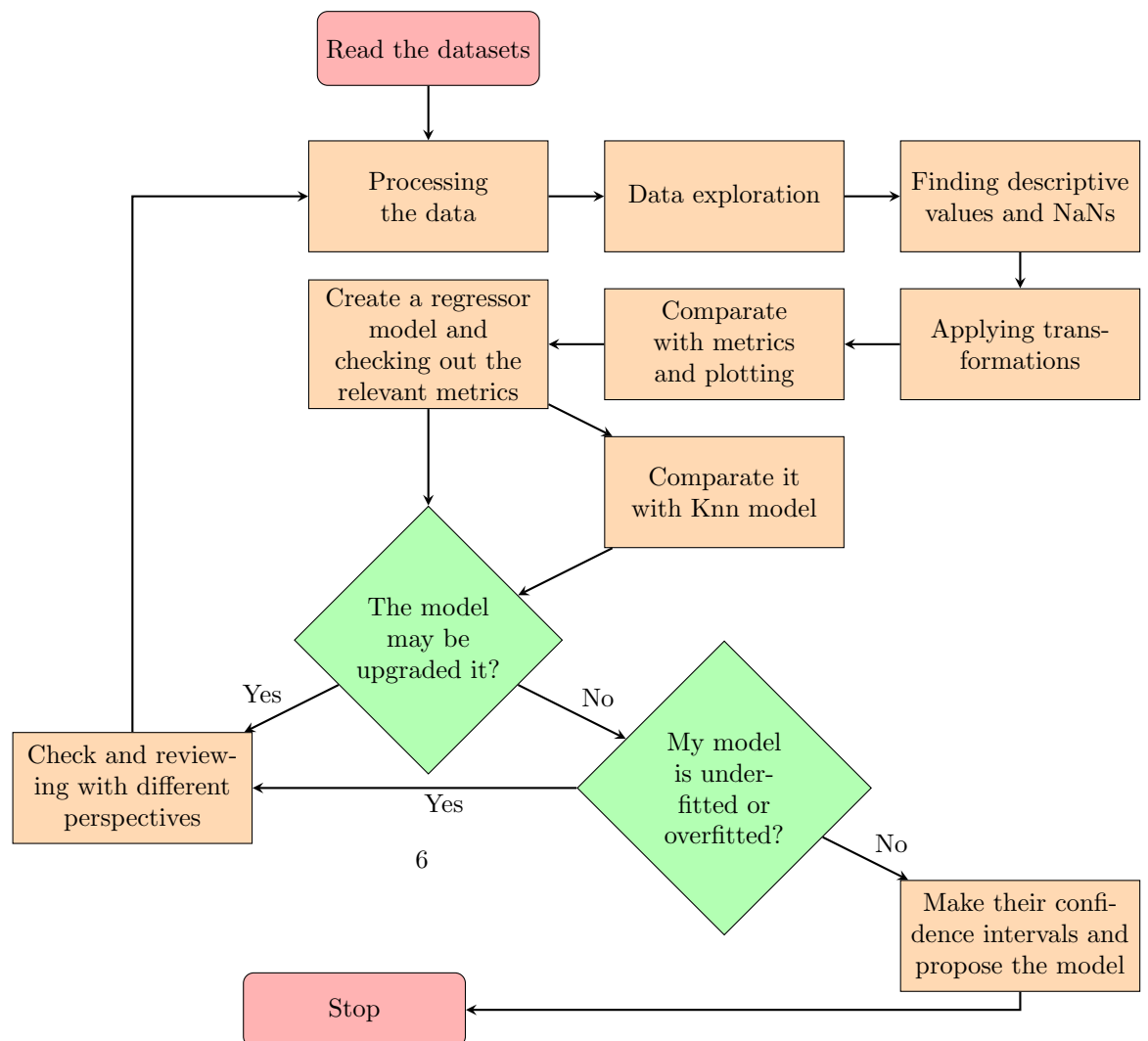


Figure 4: The procedural followed at the investigation

These were the main steps followed at the work. The part of reading some material and check information from several sources, is mainly the first step before to begin the work. However if there's something missing when the hypothesis are covered, it'll be useful to look for more sources to create other hypothesis.

5 Data analysis

5.1 first dataset

The first data analysis that has been studied was the dataset with sintetic data. As it had been mentioned, following the scientific methodology we look for the data and it started with a data exploration. It obtained the following values

parameter	y	x
number of values	506.000000	506.000000
mean	22.528854	3.613524
std	9.182176	8.601545
minimum	5.000000	0.006320
percentile 25	17.025000	0.082045
percentile 50	21.200000	0.256510
percentile 75	25.000000	3.677083
maximum	50.000000	88.976200

Now that we have assured that the data are completed we can see the value of the mean and the 50 percentile are quited different, mostly on the variable x, it requires to be cleaned the data. Nevertheless, it is useful to check the correlation with variable x to y too. The correlation matrix resulted on

	y	x
y	1.000000	-0.389582
x	-0.389582	1.000000

The correlation is quite low between x,y. That means that it'll be hard to find a relation with the x and y with linear regression. And also, something that we can notice with the plots that they are not normally distribuited. The x-axis it seems more a exponential distribution than normal. Other way to demonstrate it could making a hypothesis test and comprobatng it with a Shapiro-walk; it consist to propose a null hypothesis, that if the value of statistic is closer to 1 or great, it will reject the hypothesis, therefore the distribution doesn't follow a normal distribution. By other hand if we encounter a value smaller it will indicate is that probably the distribution follows a normal dsitrbuton. And also we have de p value that if is greater than 0.05 will indicate that is a normal distribution, in the oppostive way it is a non-normal distribution. The test

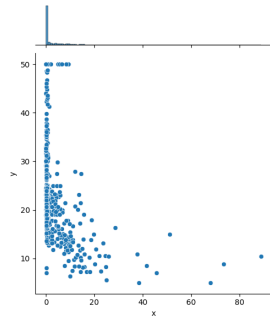


Figure 5: E

concluded with a distribution of $4.85 * 10^{-28}$ so it is not a normal distribution. At this point we can choose several ways to follow

- Trying to normalizing
- Trying with models more associated with exponential distributions or models that will no depend on the distribution.
- Infere with other features

Due to the objectives of the work, the first attempt it'll be used. At this dataset has been followed 4 approaches to look the effcence of methods at the cleaning of the data. However, it has been tested several types of transformations to check out if it'll be more useful the logarithm transformation according with the parameter of skew

type-skew	x	y
original	5.223148798243851	1.1109118502479587
log	1.2692005882725572	-0.24563979611568673
sqrt	2.024382103123676	0.4381663127860419
1/x	-0.5772191040682719	1.9393215717506038
box-cox	0.093649	...
outliers	0.40335	1.058543

It could seems some techniques are better however logarithm tranformation has a better resolution in the skew of both variables.

5.2 first method

The firsrt approach to make a model was transforming with logarithm and making a cross validation And therefore it was contrasted with the Knn model, getting the next values

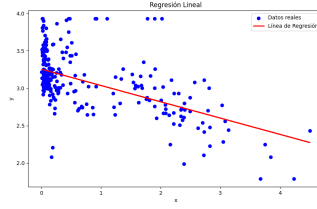


Figure 6: Lineal regression

metric	Lineal regression	Knn
R^2	35.488840060699256%	43.57
MSE	0.09611374428381783	0.11904
RMSE	0.30971201639466434	0.33166

The model has been contrasted with both models and with the same cleaned data, as also trained at the same way respectively at the model. However, the Lineal regressor has been predicting the data within the confidence. For example, it added a value x on 2.3 and the result in y is 2.7534. The confidence value on 95% interval is from (2.172214750599946, 3.334689169873629) and it has been seen the value was fitted correctly. However it has been made a gen-

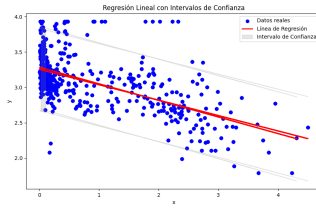


Figure 7: Lineal regresion with the intervals set

eralized model to linear regression in a set of dots (fig8). The equation for this model is

$$y = -0.21773690693332431x + 3.2542468461834333$$

5.3 Second method

The second approach it consisted to just transforming the value with logarithm without CV 5x2

metric	Lineal regression	Knn
R^2	15.0589474050694%	22.562569548739275%
MSE	0.1127456496305317	0.11904364904719832
RMSE	0.3357761897909554	0.33166

The values obtained are quite low, and the test of interval values with a random value it has been displayed. The random value tested was $x = 2.3$ $y =$

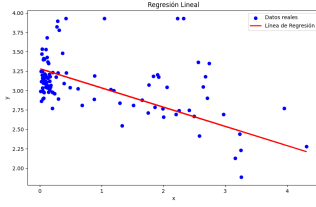


Figure 8: Lineal regresion with the confidence intervals

2.712267315 The CI of 95% for 2.3 (2.1310301053631586, 3.2935045246368415) so the equation resulting is:

$$y = 3.28236728 - 0.24786955x$$

5.4 third method

The outliers were fixed with capping and flooring methos using the median

metric	Lineal regression	Knn
R^2	32.19327627743514%	32.92025378085354%
MSE	0.002590185931743603	0.09954163106601016
RMSE	0.3097120163946644	0.3155021886

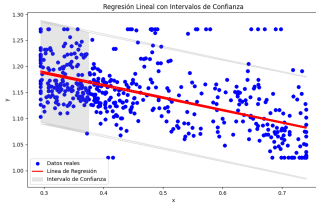


Figure 9: Lineal regresion with the confidence intervals

The random value tested was $x = 0.4$ $y = 1.1624568760450196$ The CI of 95% for 2.3 (1.0655741648818118, 1.2593395872082274) so the equation resulting is:

$$y = 1.2573964360450196 - 0.23734890073453882x$$

5.5 fourth method

The last methods were used at this model. And also the test of Knn, was a little bit different to due it has beren included the type of distance

metric	Lineal regression	Knn
R^2	37.1 %	38.88337562736399%
MSE	0.08067515724063212	0.06685108607726611
RMSE	0.2839075215367599	0.3155021886

Although, for this search of parameters, it got the Minkowski dataset, and finally the value of the intervals and equation are described as The random value tested was $x = 0.4$ $y = 3.1646101221355143$ The CI of 95% for 2.3 (2.624302151585808, 3.7049180926852205) so the equation resulting is:

$$y = 3.2466778431246275 - 0.20516930247278303x$$

The methods were contrasted by the simplest model and also the intervals were obtained, however that section it comprehends the results discussion

And one of the most important things that it has obtained is this table that compiles all the metrics and parameters that has been obtaining

Métrica	model1	model2
R^2	35.4%	15.05%
MSE	0.0961	0.1127
RMSE	0.3097	0.3358
EQUATION	$y = -0.2177x + 3.2542$	$y = -0.2479x + 3.2824$
Adjusted R^2	36.7%	39.8%
F-statistic	147.1	267.1
$P(F - statistic)$	5.97e-27	2.11e-46
Log-Likelihood	-50.079	-91.384
AIC	104.2	186.8
BIC	111.2	194.8
std error	const=0.024, x=0.018	const=0.019, x=0.015
σ	const=135.161, x=-12.128	const=170.036, x=-16.343
$P > t $	const: 0.000, x: 0.000	const: 0.000, x: 0.000
Confidence Intervals	const=[3.2068, 3.3017], x=[-0.2531, -0.1824]	const=[3.2444, 3.3203], x=[-0.2777, -0.2181]
Omnibus	41.738	37.308
Prob(Omnibus)	0.000	0.000
Jarque-Bera (JB)	64.331	64.192
Prob(JB)	1.07e-14	1.15e-14
Skew	0.960	0.581
Kurtosis	4.553	4.570
Durbin-Watson	0.983	1.902
Condition Number	2.26	2.16

Table 1: Model comparsion 1 and 2

Métrica	model3	model4
R^2	32.193%	37.1%
MSE	0.014957	0.0807
RMSE	0.12224	0.2839
EQUATION	$y = -0.30537x + 1.85673$	$y = -0.2052x + 3.2467$
Adjusted R^2	33.5%	36.8%
F-statistic	128.2	147.8
$P(F - statistic)$	2.76e-24	4.72e-27
Log-Likelihood	181.27	-32.091
AIC	-385.5	68.18
BIC	-351.5	75.25
std error	const=1.8567, x=-0.3054	const=0.022, x=0.017
σ	const=0.015, x=0.027	const=144.464, x=-12.159
$P > t $	const: 0.000, x: 0.000	const: 0.000, x: 0.000
Confidence Intervals	const=[1.8279, 1.886], x=[-0.3584, -0.2522]	const=[3.2024, 3.2909], x=[-0.2384, -0.1719]
Omnibus	40.637	35.555
Prob(Omnibus)	0.000	0.000
Jarque-Bera (JB)	56.377	46.953
Prob(JB)	5.73e-13	6.37e-11
Skew	1.021	0.938
Kurtosis	4.087	3.966
Durbin-Watson	0.966	0.958
Condition Number	4.45	2.26

Table 2: Model comparision 3 and 4

5.6 Second dataset

The next dataset that it has been studying was about cars. It is a car's magazine from 1985 that compiles prices and features from several cars. For this dataset it has been read some papers(include the cite) to understand some of the features and its relationship. As a one of the first steps was the cleaning the data, and a data exploration was made to look for missing values and

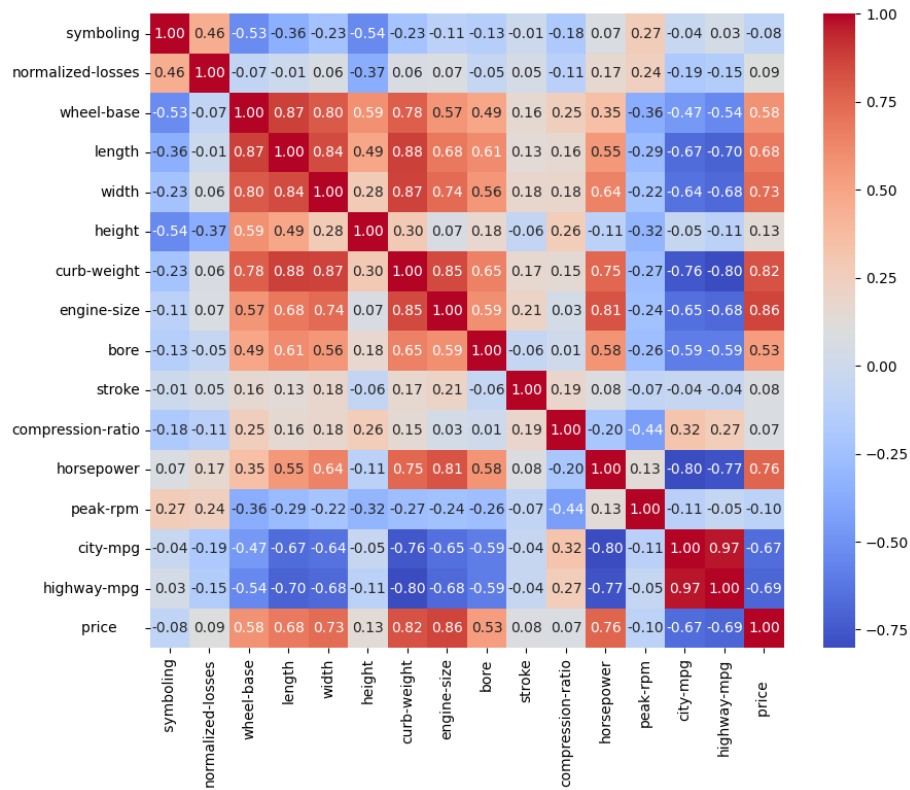


Figure 10: The heatmap of correlation

The heatmap (Figure 10) represents how is related a variable with other. As the objective of the study is to making a model that can predict the price of cars according to certain variables. However a tool that can save time it is to see the correlation with continuous variables. So those variables that have been analyzed were the Curb weight, engine size, horsepower city-mpg and highway-mpg. The procedure to clean up the data was the same as the previous dataset however some columns contained a few NaN values and they must be changed. As also, some features are measured by different magnitudes, for example, amounts of money and horse power, so it would be useful to standardize the data and also it could be applied a logarithm transformation due to some the dataset wasn't following a normal distribution, instead it was an exponential distribution (figure 11,12).

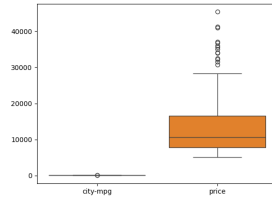


Figure 11: *
Example when the
data is not standardized

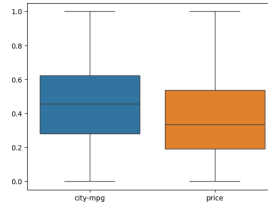


Figure 12: *
Example when data uis
standarized and
transformed

5.7 Engine size

metric	Lineal regression	Knn
R^2	73.62087367142167 %	78.14960577066459%
MSE	0.00992703573774292	0.008260694761950247
RMSE	0.09960553714701531	0.09078358735816178

Prediction for x=0.9: 0.873558685297521 Ci of 95%for x=0.9: (0.6755512178129525, 1.0715661527820897)

$$y = 1.0421234211972827x - 0.06435239378003343$$

5.8 Miles per gallon on city

metric	Lineal regression	Knn
R^2	60.45640689950125 %	63.98078808830601%
MSE	0.020279279592056144	0.01864935309235264
RMSE	0.14237431771458053	0.13622653661454523

Prediction for x=0.9: 0.001530863668857796 Ci of 95%for x=0.9: (-0.2679841644933157, 0.2710458918310313)

$$y = -0.847735376275263x + 0.7644927023165945$$

5.9 Highway miles per galon

metric	Lineal regression	Knn
R^2	60.26238063837296 %	71.4051650378093%
MSE	0.020296748345956392	0.01473467535564263
RMSE	0.14246665691899474	0.12126069885616314

Prediction for x=0.9: 0.029920806311725112 Ci of 95%for x=0.9: (-0.23756924465403179, 0.297410857277482)

$$y = -0.8749508264218194x + 0.8173765500913626$$

5.10 Crub weight

metric	Lineal regression	Knn
R^2	75.95183149109518 %	75.37829642479748%
MSE	0.012140499045124736	0.012413594738192302
RMSE	0.1262230883265539	0.11115941370057997

Prediction for x=0.9: 0.8331700416235279 Ci of 95%for x=0.9: (0.6006405680156519, 1.065699515231404)

$$y = 0.9548221722260348x - 0.026169913379903483$$

5.11 Horse power

metric	Lineal regression	Knn
R^2	R ² : 69.07940114232951 %	43.319290242601016%
MSE	0.08221699892102004	0.13686099980559793
RMSE	0.28673506747696564	0.3696553507791872

Prediction for x=0.9: 0.8188622328838921 Ci of 95%for x=0.9: (0.5575508676925711, 1.0801735980752132)

$$y = 0.8850130631724356x + 0.022350476028700144$$

5.12 Make

metric	Lineal regression	Knn
R^2	64.3410145660548 %	55.39383851984723%
MSE	0.010664898715086689	0.022665468018166633
RMSE	0.018231972174382653	0.15049592765441985

Prediction for x=0.7: 9.776875346817139 Ci of 95%for x=0.9: (9.502410532700683, 10.051340160933595)

$$y = 0.18958328x + 9.644167050817138$$

This model it has to be taken with careful due is a colineal model that has a several brands of cars. At also, the slopes are

$$slopes = 0.18958328, 0.50766223, -1.26198477, -0.99000114, -0.92894517, -0.6305545, 1.16321074, -0.682593$$

The way that it has been chosen the categorical variable of Make thna the other variables, it is because they were tested with ANOVA, and the statistical significance

Categoría	Suma Cuadrada (sum_sq)	Grados de Libertad (df)	F	Valor PR($\geq F$)
make	9.750960e+09	21.0	29.502216	1.019818e-47
fuelType	1.534125e+08	1.0	2.495859	0.115703
aspiration	3.969954e+08	1.0	6.587290	0.010991
numofdoors	2.672851e+07	1.0	0.427171	0.514127
body	1.960055e+09	4.0	9.183927	7.844576e-07
drive	5.060130e+09	2.0	67.503667	3.539271e-23
engineloc	1.383992e+09	1.0	24.979629	0.000001
type	2.538008e+09	6.0	8.298116	5.008078e-08
cylinders	7.336567e+09	6.0	45.727054	7.149270e-35
system	4.352758e+09	7.0	14.797402	1.865735e-15

Table 3: Resultados ANOVA con valor PR($>F$) para diferentes categorías.

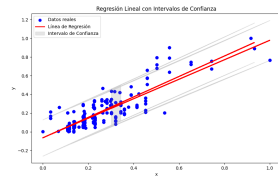


Figure 13: *
Engine size regressor

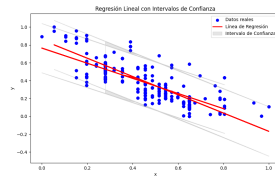


Figure 14: *
Miles per gallon city

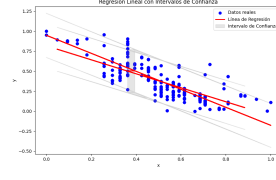


Figure 15: *
Highway miles per
gallon

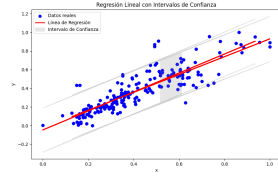


Figure 16: *
Curb weight

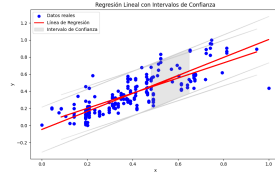


Figure 17: *
Horse power

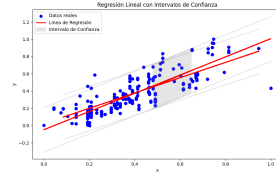


Figure 18: *
Make

Another thing to take in mind is the tendency that follows the variables

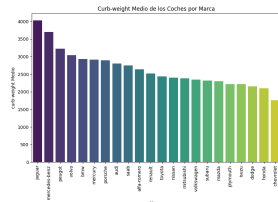


Figure 19: *
Curb weight and brand

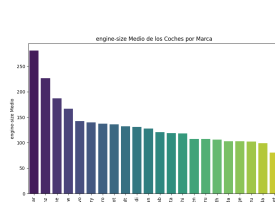


Figure 20: *
Engine size and brand

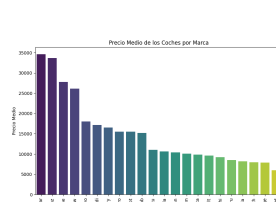


Figure 21: *
Prices and brand

6 Discussion

6.1 First data-set

As we can observe on the table 1 and table 2 it proposed 4 different models with different type of data cleaning, however we have seen different results in each one. If we see the R^2 of the models the worst is the second model, and also, as it saw in the tables from each table, the value of the Knn and The regressor is the lowest although, is close among them. The best R^2 and also, it has a one of the closest values with Knn regressor, is the fourth model. Now in terms of MSE and RMSE the third model had a lowest MSE and RMSE but the fourth model it has also a good resolution in that terms. Always in Data science it can be created different models to represent different phenomenon or different aims. In this case, we have seen that despite that the fourth model doesn't have the best MSE and RMSE from all models however it is no the worst, the percentage of describing data is the highest so if we have this model with just this two variables and type of regressor, it can choose the fourth model. However the best thing that it can be done on this dataset could it be use different models or different treatments on the distribution

6.2 Second data-set

In base on the related works, (Polat and Bulut, 2023) used a similar approach in the creation of the model and also at the prediction of price of cars. And here it found that the cars with a certain make have more influence in a price than cars with a unknown brand. And (Alhakamy et al., 2023) the way to describe different features that can lead to different prices of car in the Van regions had a similarity workflow than this work. As also they used similar metrics. And (Weber, 2019) they found how the fuel consumption has a huge weight into the prices of cars. As also the make is other important factor to see. In the last analysis, it could have seen that the main features had a strong relation were those related with the motor and the structure of the car. And also the efficiency of the computation of fuel. In the case of categorical variables, the variable that had a better resolution in predictions was the make, so here it can be refuted the past works. To indicate the best model in terms of the car's infrastructure; could be the curb weight because it had a r^2 of 75% and its Knn model has similar approach. Also its MSE and RMSE despite that is not the best of all, it is second model that has a good resolution on MSE and RMSE. These variables which are related with the structure of the car, tend to follow the mark and the prices respect to the mark has a similar approach. So it can be deduce that relevant and expensive cars would try to make their cars as better as possible, such in the efficiency of consumption of fuel or the curb weight, horse power. Meanwhile those brands that has a cheaper cost don't tend to have the better structure features.(figure19,20,21)

7 Conclusions and future work

In both datasets can be used other type of models and see its efficiency in the prediction, as also it could be useful try to use regression that are more associated with exponential distributions. In the second part it could aim to see the colinearity and compare the r^2 obtained in the last models respect to the colinear model. To make a good analysis it can be challenging because several variables could be related among them, but when it makes an analysis the best way to know if it is following the correct path is looking for related works and its results, with this as priority it could be good help to delimate the variables to use. Also, the dataset 2 shows how tendencies works on and it is important to look out at them.

References

- Alhakamy, A., Alhowaity, A., Alatawi, A. A., and Alsaadi, H. (2023). Are used cars more sustainable? price prediction based on linear regression. *SUSTAINABILITY*, 15(2).
- Polat, M. and Bulut, C. (2023). İkinci el otomobil fiyat artışına etki eden faktörlerin yapısal eşitlik modeli ile tespit edilmesi: Van İli Örneği. *Sosyoekonomi*, 31(55):347–369.
- Weber, S. (2019). Consumers’ preferences on the swiss car market: A revealed preference approach. *Transport Policy*, 75:109–118.