MACHINE LEARNING I

Assignment II – Regression Analysis

Group Assignment

Mariana Narvaez, Jonas Hellevang, Daniel Bilitewski, Guillermo Chacón, Nisrine Ferahi, Rabiga Shangereyeva

25th December 2019

Introduction

The following report summarizes the real state analysis done to properties located in Madrid. In this report we will explain the method used to obtain the real state data, a description of the characteristics of the data obtained and what they mean, the cleaning process of it, the applied regression model and finally some conclusions from the data.

Market in Spain

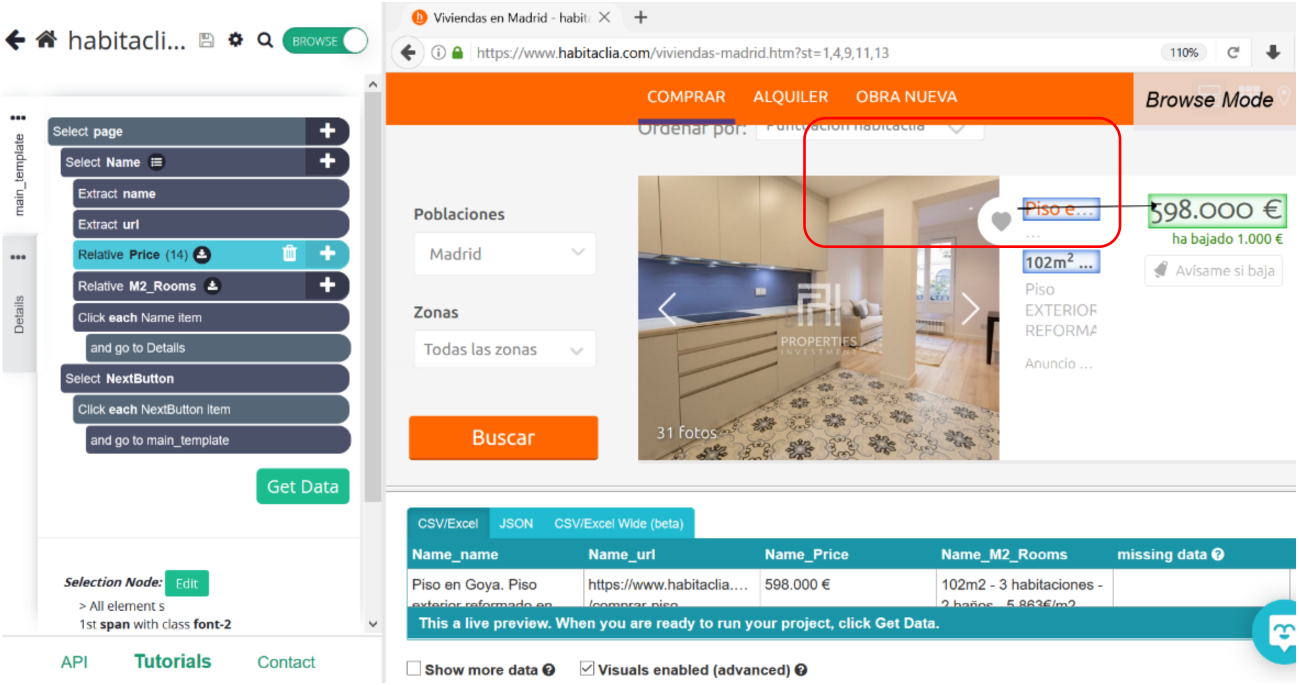
The real estate market of Madrid has the last years experienced an increase in prices, which experts have speculated in being a bubble. However, a report done by Moving2Madrid proves otherwise, showing us that the market is undergoing a significant structural change. Madrid is the fastest growing city in Spain where no city comes even close to what Madrid is experiencing, where Granada is the closest one. Why not Barcelona? Due to the Cataluña Independence movement, Barcelona has experienced capital flight, leaving less jobs. Barcelona has also 4% higher taxes on secondhand properties, and stricter regulations for Airbnb. What makes this more interesting is that several other cities in Europe like Oslo, the capital of Norway, is experiencing the same increase in prices over the last few years.

Dataset

Web Scrapping:

To obtain the data, we used **Parsehub** and create a project for extracting the data from **Habitaclia** https://www.habitaclia.com/madrid. The project was designed to get the data only for flats in Madrid, not houses. We used a main template to go over all the flats of the first page. The template extracted the description, URL and price of each flat of the first page and then we clicked the next button to go to the next page. The next button step was only executed after the **details template** was executed.

Image 1: First page web Scrapping



- Select page
- Select Name
- Extract name
- Extract url
- Relative Price (14)
- Relative M2_Rooms
- Click each Name item
- and go to Details
- Select NextButton
- Click each NextButton item
- and go to main_template

Selection Node: Edit
> All element s
1st span with class font-2

https://www.habitaclia.com/viviendas-madrid.htm?st=1,4,9,11,13

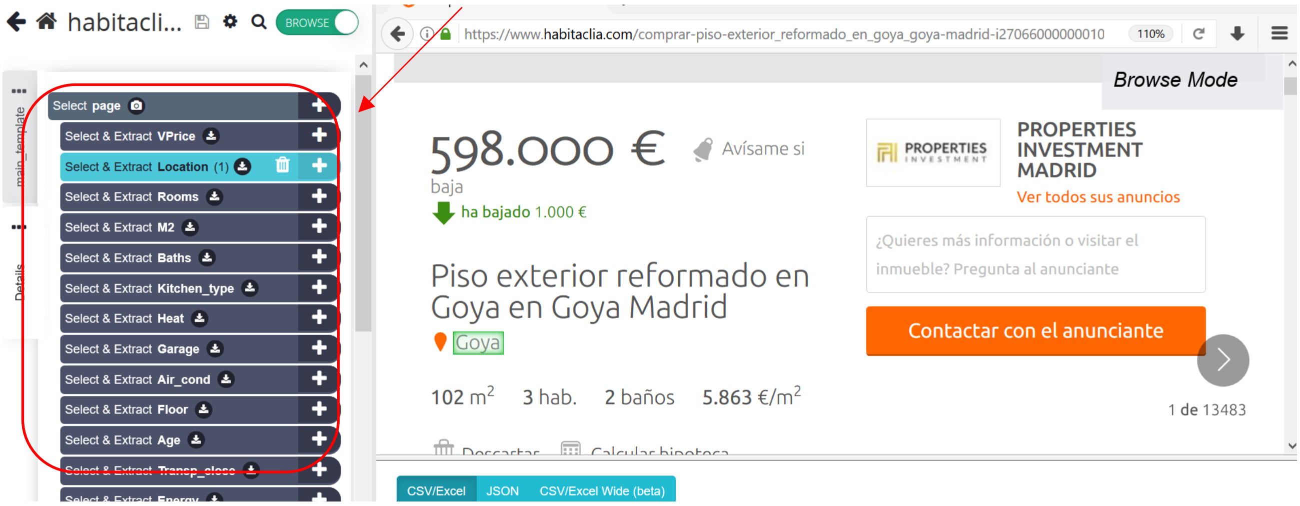
Name_name	Name_url	Name_Price	Name_M2_Rooms	missing data
Piso en Goya. Piso exterior reformado en	https://www.habitaclia.... /comprar-piso	598.000 €	102m2 - 3 habitaciones - 2 baños - 5.863€/m2	

This a live preview. When you are ready to run your project, click Get Data.

□ Show more data ☑ Visuals enabled (advanced)

The **details template** was used to go into each flat of each page and extract all the additional variables we thought was relevant for the analysis.

Image 2: First page web Scrapping



- Select page
- Select & Extract VPrice
- Select & Extract Location (1)
- Select & Extract Rooms
- Select & Extract M2
- Select & Extract Baths
- Select & Extract Kitchen_type
- Select & Extract Heat
- Select & Extract Garage
- Select & Extract Air_cond
- Select & Extract Floor
- Select & Extract Age
- Select & Extract Transp_close
- Select & Extract Energy

https://www.habitaclia.com/comprar-piso-exterior_reformado_en_goya_goya-madrid-i27066000000010

habitacli...
BROWSE
main_template
Details
Browse Mode
598.000 €
Avísame si baja
ha bajado 1.000 €
Piso exterior reformado en Goya en Goya Madrid
Goya
102 m2 3 hab. 2 baños 5.863 €/m2
Descartar
Calcular hipoteca
PROPERTIES INVESTMENT MADRID
Ver todos sus anuncios
¿Quieres más información o visitar el inmueble? Pregunta al anunciante
Contactar con el anunciante
1 de 13483
CSV/Excel
JSON
CSV/Excel Wide (beta)

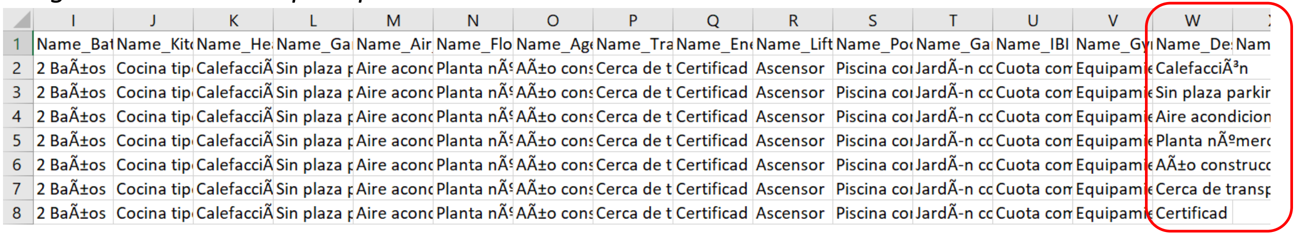
Data Preparation:

After we finally managed to get the data, we proceeded with the cleaning so we could use the information obtained in Dataiku.

Data manipulation:

Because for each flat, some of the extracted features was displayed as a list (highlighted in red) at the end of the columns containing the information of the flat, we got several rows per observation.

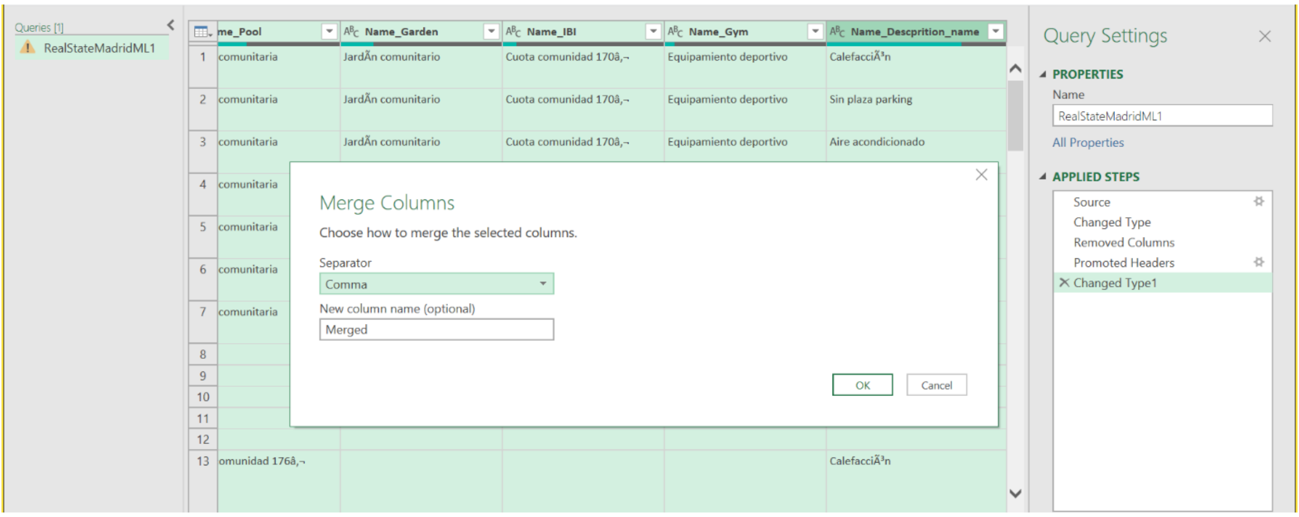
Image 3: Several rows per apartment



	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	Name_Bat	Name_Kit	Name_He	Name_Ga	Name_Air	Name_Flo	Name_Age	Name_Tra	Name_Ene	Name_Lift	Name_Poo	Name_Ga	Name_IBI	Name_Gy	Name_De	Nam
2	2 BaÃ±os	Cocina tip	CalefacciÃ	Sin plaza p	Aire acond	Planta nÃº	AÃ±o cons	Cerca de t	Certificad	Ascensor	Piscina co	JardÃ-n cc	Cuota com	Equipamie	CalefacciÃ³n	
3	2 BaÃ±os	Cocina tip	CalefacciÃ	Sin plaza p	Aire acond	Planta nÃº	AÃ±o cons	Cerca de t	Certificad	Ascensor	Piscina co	JardÃ-n cc	Cuota com	Equipamie	Sin plaza parkir	
4	2 BaÃ±os	Cocina tip	CalefacciÃ	Sin plaza p	Aire acond	Planta nÃº	AÃ±o cons	Cerca de t	Certificad	Ascensor	Piscina co	JardÃ-n cc	Cuota com	Equipamie	Aire acondicion	
5	2 BaÃ±os	Cocina tip	CalefacciÃ	Sin plaza p	Aire acond	Planta nÃº	AÃ±o cons	Cerca de t	Certificad	Ascensor	Piscina co	JardÃ-n cc	Cuota com	Equipamie	Planta nÃºmerc	
6	2 BaÃ±os	Cocina tip	CalefacciÃ	Sin plaza p	Aire acond	Planta nÃº	AÃ±o cons	Cerca de t	Certificad	Ascensor	Piscina co	JardÃ-n cc	Cuota com	Equipamie	AÃ±o construcc	
7	2 BaÃ±os	Cocina tip	CalefacciÃ	Sin plaza p	Aire acond	Planta nÃº	AÃ±o cons	Cerca de t	Certificad	Ascensor	Piscina co	JardÃ-n cc	Cuota com	Equipamie	Cerca de transp	
8	2 BaÃ±os	Cocina tip	CalefacciÃ	Sin plaza p	Aire acond	Planta nÃº	AÃ±o cons	Cerca de t	Certificad	Ascensor	Piscina co	JardÃ-n cc	Cuota com	Equipamie	Certificad	

In order to not lose any important information, we decide to keep all the rows per observation and store all the information from the different columns into one (apart from price, bedrooms, bathrooms and neighborhood). The values of the columns were separated with a comma in the new columns called **Merged.**

Image 4: Merging process



	me_Pool	Name_Garden	Name_IBI	Name_Gym	Name_Descprition_name
1	comunitaria	JardÃn comunitario	Cuota comunidad 170â‚¬	Equipamiento deportivo	CalefacciÃ³n
2	comunitaria	JardÃn comunitario	Cuota comunidad 170â‚¬	Equipamiento deportivo	Sin plaza parking
3	comunitaria	JardÃn comunitario	Cuota comunidad 170â‚¬	Equipamiento deportivo	Aire acondicionado
4	comunitaria				
5	comunitaria				
6	comunitaria				
7	comunitaria				
8					
9					
10					
11					
12					
13	omunidad 176â‚¬				CalefacciÃ³n

Merge Columns

Choose how to merge the selected columns.

Separator: Comma

New column name (optional): Merged

Query Settings

PROPERTIES

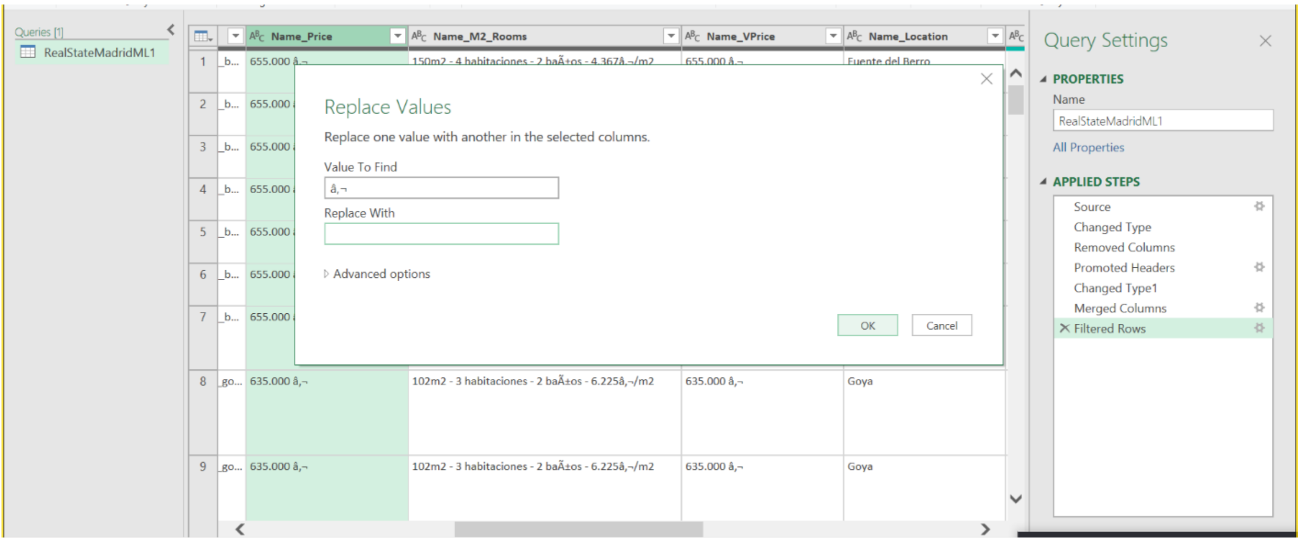
Name: RealStateMadridML1

APPLIED STEPS

- Source
- Changed Type
- Removed Columns
- Promoted Headers
- Changed Type1

Continuing with the cleaning, we extracted the important values from each separated column. E.g. For the price we only preserved the numeric value of the column.

Image 5: Extract important values



Queries [1]
RealStateMadridML1

		Name_Price	Name_M2_Rooms	Name_VPrice	Name_Location
1	_b...	655.000 â‚¬	150m2 - 4 habitaciones - 2 baÃ±os - 4.367â‚¬/m2	655.000 â‚¬	Fuente del Berro
2	_b...	655.000			
3	_b...	655.000			
4	_b...	655.000			
5	_b...	655.000			
6	_b...	655.000			
7	_b...	655.000			
8	_go...	635.000 â‚¬	102m2 - 3 habitaciones - 2 baÃ±os - 6.225â‚¬/m2	635.000 â‚¬	Goya
9	_go...	635.000 â‚¬	102m2 - 3 habitaciones - 2 baÃ±os - 6.225â‚¬/m2	635.000 â‚¬	Goya

Replace Values

Replace one value with another in the selected columns.

Value To Find: â‚¬

Replace With:

Advanced options

OK Cancel

Query Settings

PROPERTIES

Name: RealStateMadridML1

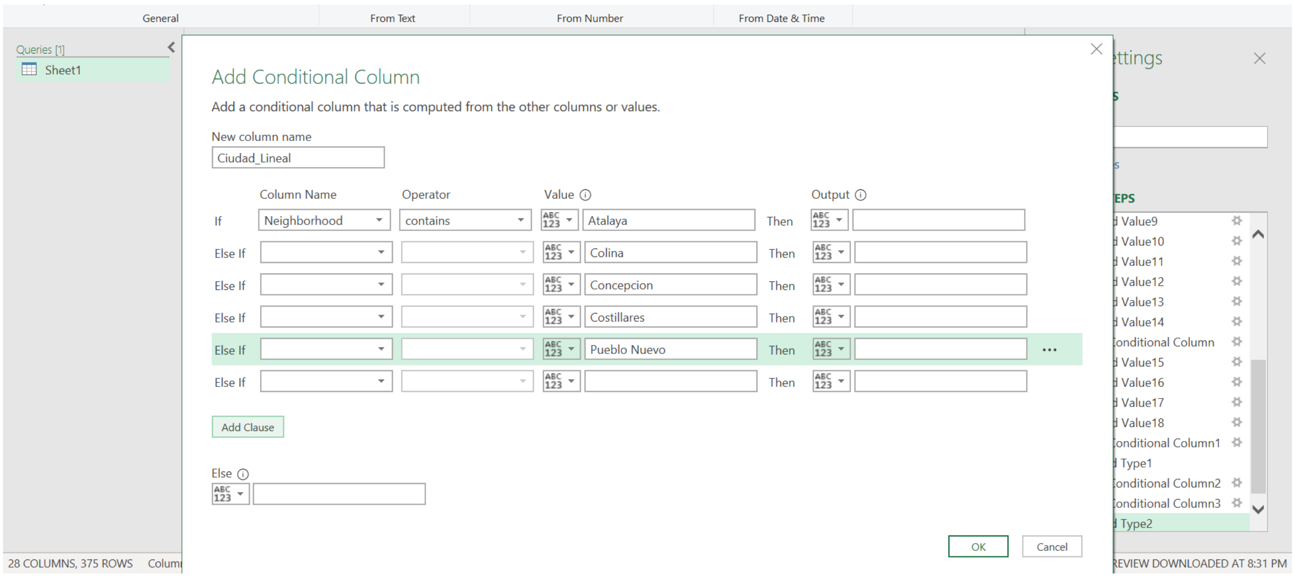
All Properties

APPLIED STEPS

- Source
- Changed Type
- Removed Columns
- Promoted Headers
- Changed Type1
- Merged Columns
- Filtered Rows

For neighborhoods we created a conditional column for each district including the corresponding neighborhoods. E.g. Ciudad Lineal.

Image 6: Conditional for each district Column



General | From Text | From Number | From Date & Time

Queries [1]
Sheet1

Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name: Ciudad_Lineal

	Column Name	Operator	Value		Output
If	Neighborhood	contains	Atalaya	Then	
Else If			Colina	Then	
Else If			Concepcion	Then	
Else If			Costillares	Then	
Else If			Pueblo Nuevo	Then	
Else If				Then	

Add Clause

Else:

OK Cancel

ttings

EPS

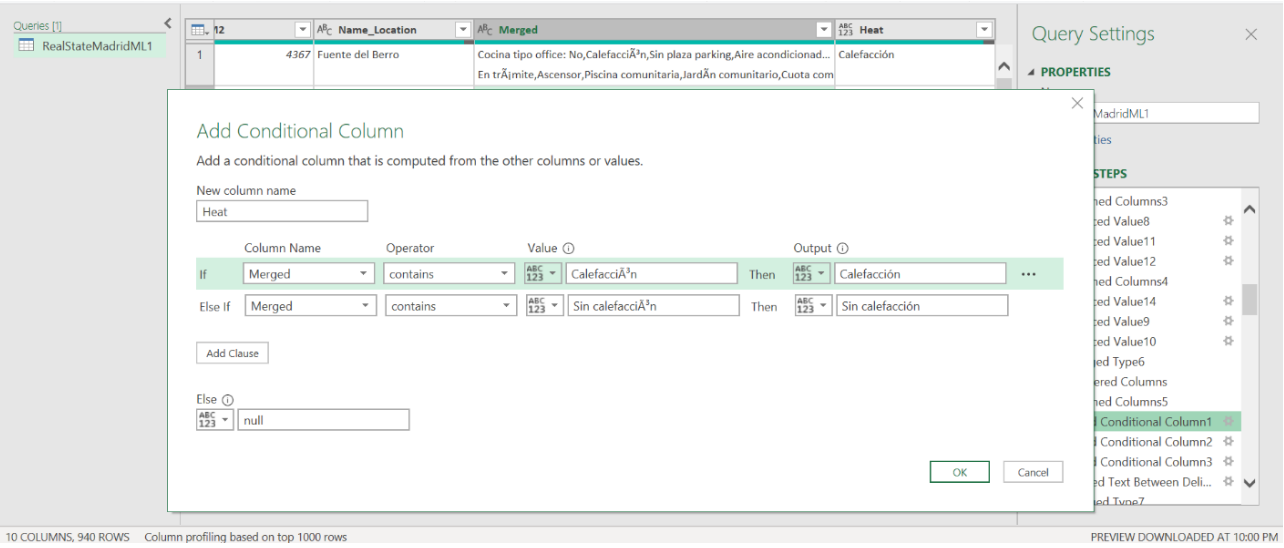
- Value9
- Value10
- Value11
- Value12
- Value13
- Value14
- onditional Column
- Value15
- Value16
- Value17
- Value18
- onditional Column1
- Type1
- onditional Column2
- onditional Column3
- Type2

28 COLUMNS, 375 ROWS Colum

REVIEW DOWNLOADED AT 8:31 PM

The next step was to create conditional columns for the rest of the features of the flats. Basically, we looked into the Merged column for the feature we were interested in and created a new column for that feature.

Image 7: Conditional for other features



12	Name_Location	Merged	Heat
1 4367	Fuente del Berro	Cocina tipo office: No,CalefacciÃ³n,Sin plaza parking,Aire acondicionad... En trÃ¡mite,Ascensor,Piscina comunitaria,JardÃn comunitario,Cuota com	Calefacción

Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name: Heat

	Column Name	Operator	Value		Output
If	Merged	contains	CalefacciÃ³n	Then	Calefacción
Else If	Merged	contains	Sin calefacciÃ³n	Then	Sin calefacción

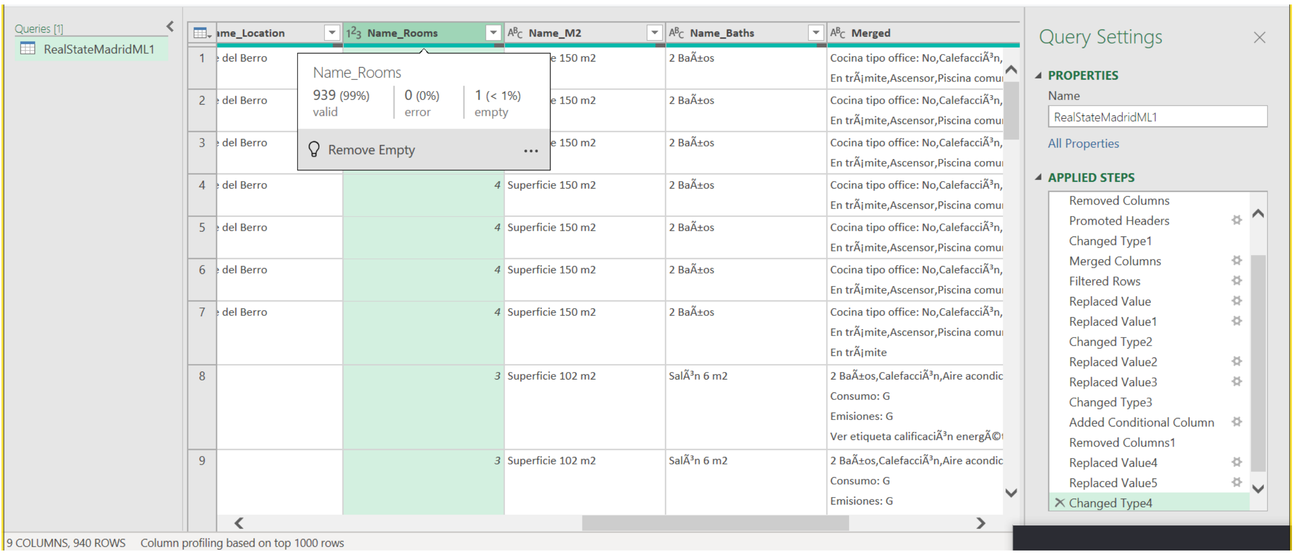
Else: null

10 COLUMNS, 940 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 10:00 PM

After we had all the columns properly separated, we started to look for missing values and to replace them with the best option considering the type of data for each column. E.g. For bedrooms we decided to impute the missing values by using the average.

Image 8: Checking for missing values



Name_Rooms

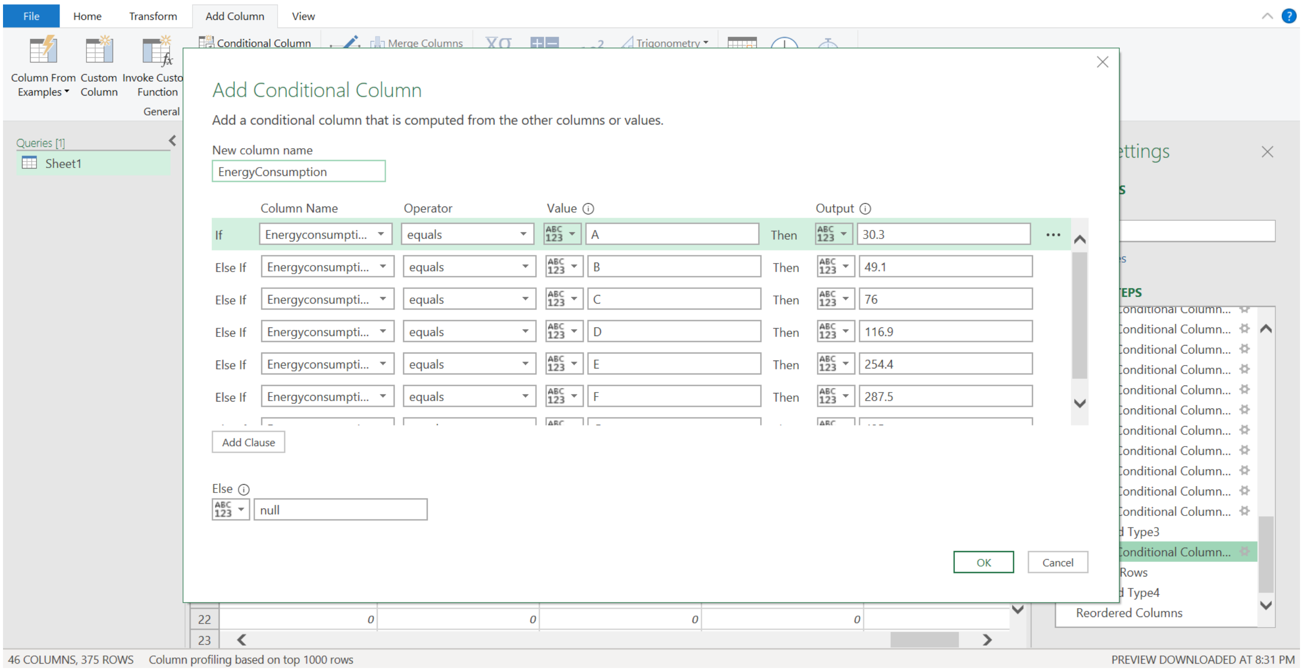
939 (99%) valid | 0 (0%) error | 1 (< 1%) empty

	ame_Location	Name_Rooms	Name_M2	Name_Baths	Merged
1	del Berro		e 150 m2	2 BaÃ±os	Cocina tipo office: No,CalefacciÃ³n, En trÃ¡mite,Ascensor,Piscina comu
2	del Berro		e 150 m2	2 BaÃ±os	Cocina tipo office: No,CalefacciÃ³n, En trÃ¡mite,Ascensor,Piscina comu
3	del Berro		e 150 m2	2 BaÃ±os	Cocina tipo office: No,CalefacciÃ³n, En trÃ¡mite,Ascensor,Piscina comu
4	del Berro	4	Superficie 150 m2	2 BaÃ±os	Cocina tipo office: No,CalefacciÃ³n, En trÃ¡mite,Ascensor,Piscina comu
5	del Berro	4	Superficie 150 m2	2 BaÃ±os	Cocina tipo office: No,CalefacciÃ³n, En trÃ¡mite,Ascensor,Piscina comu
6	del Berro	4	Superficie 150 m2	2 BaÃ±os	Cocina tipo office: No,CalefacciÃ³n, En trÃ¡mite,Ascensor,Piscina comu
7	del Berro	4	Superficie 150 m2	2 BaÃ±os	Cocina tipo office: No,CalefacciÃ³n, En trÃ¡mite,Ascensor,Piscina comu En trÃ¡mite
8		3	Superficie 102 m2	SalÃ³n 6 m2	2 BaÃ±os,CalefacciÃ³n,Aire acondic Consumo: G Emisiones: G Ver etiqueta calificaciÃ³n energÃ©
9		3	Superficie 102 m2	SalÃ³n 6 m2	2 BaÃ±os,CalefacciÃ³n,Aire acondic Consumo: G Emisiones: G

9 COLUMNS, 940 ROWS Column profiling based on top 1000 rows

For the **Energy Consumption** variable, we decided to assign the maximum value of the range corresponding to each letter.

Image 9: Changing Energy Consumption variable



Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name: EnergyConsumption

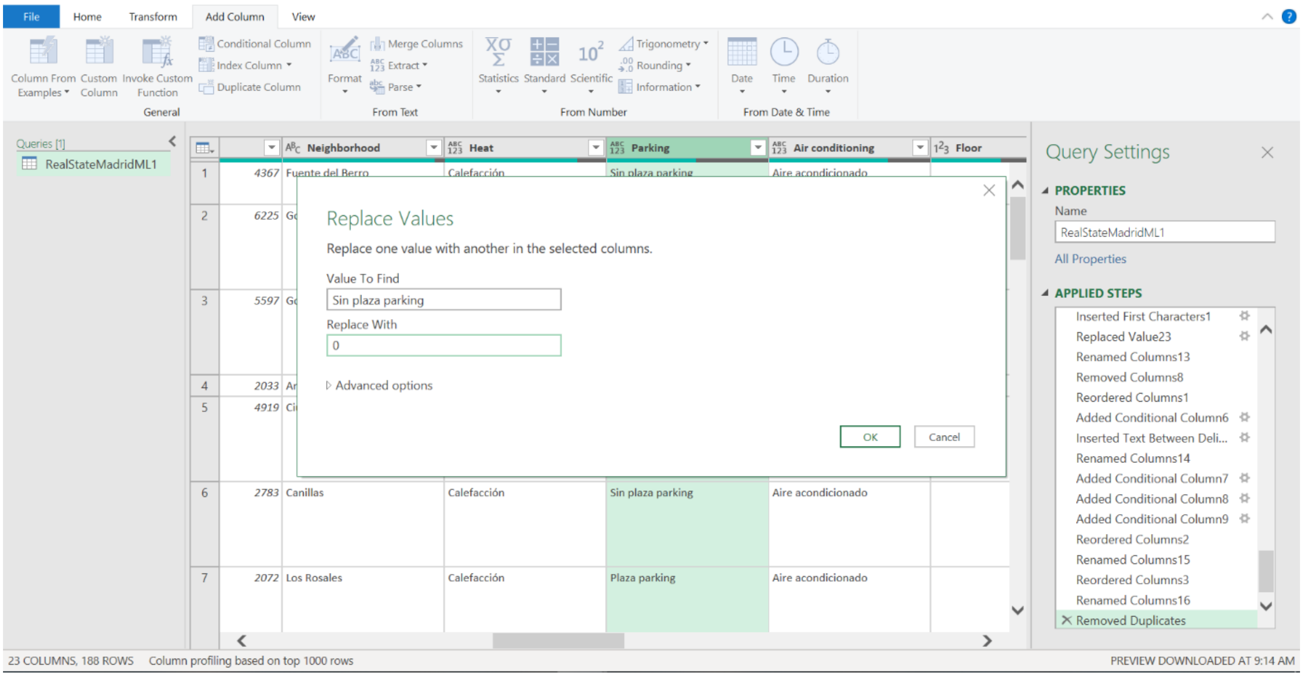
	Column Name	Operator	Value		Output
If	Energyconsumpti...	equals	A	Then	30.3
Else If	Energyconsumpti...	equals	B	Then	49.1
Else If	Energyconsumpti...	equals	C	Then	76
Else If	Energyconsumpti...	equals	D	Then	116.9
Else If	Energyconsumpti...	equals	E	Then	254.4
Else If	Energyconsumpti...	equals	F	Then	287.5

Else: null

46 COLUMNS, 375 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 8:31 PM

For all the categorical variables, we decided to replace the value by 1 and 0 depending on if the flat had or did not have the feature.

Image 10: Adjusting Categorical variables



		Neighborhood	Heat	Parking	Air conditioning	Floor
1	4367	Fuente del Berro	Calefacción	Sin plaza parking	Aire acondicionado	
2	6225	Gc				
3	5597	Gc				
4	2033	Ar				
5	4919	Ci				
6	2783	Canillas	Calefacción	Sin plaza parking	Aire acondicionado	
7	2072	Los Rosales	Calefacción	Plaza parking	Aire acondicionado	

Replace Values

Replace one value with another in the selected columns.

Value To Find: Sin plaza parking

Replace With: 0

Query Settings

PROPERTIES

Name: RealStateMadridML1

All Properties

APPLIED STEPS

- Inserted First Characters1
- Replaced Value23
- Renamed Columns13
- Removed Columns8
- Reordered Columns1
- Added Conditional Column6
- Inserted Text Between Deli...
- Renamed Columns14
- Added Conditional Column7
- Added Conditional Column8
- Added Conditional Column9
- Reordered Columns2
- Renamed Columns15
- Reordered Columns3
- Renamed Columns16
- Removed Duplicates

23 COLUMNS, 188 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 9:14 AM

Finally, we eliminated the duplicated values for each flat, using the URL that was the only value that was unique for each observation.

Descriptive analysis of the data

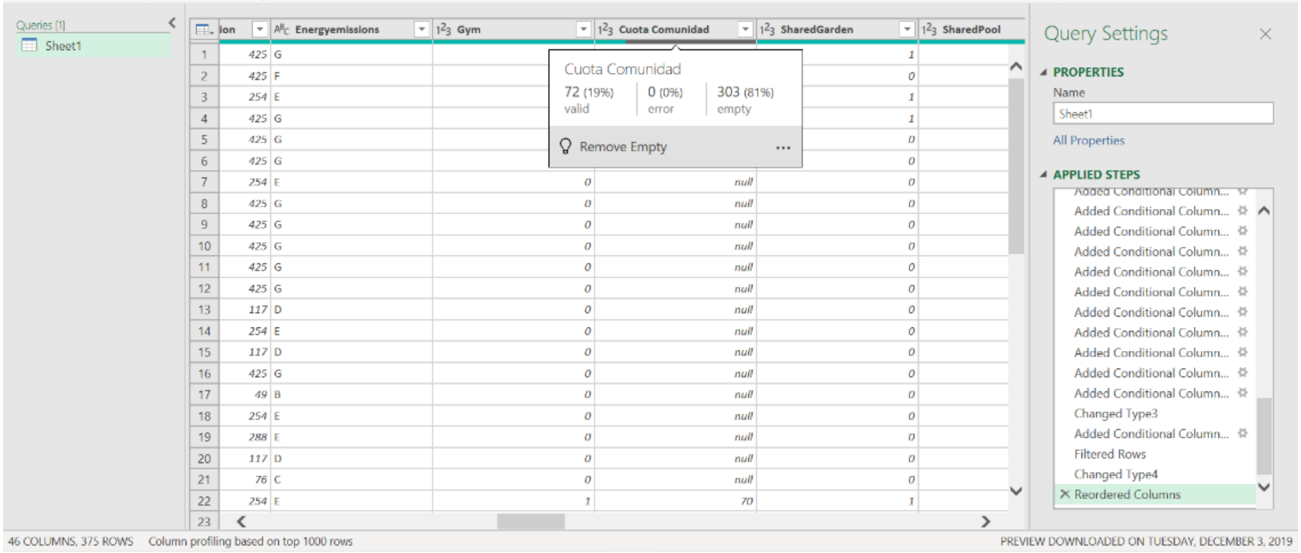
The below table show a simple description of the variables of the dataset.

Table 1: Variables Description

Variable Name	Description
Name_name	Description of the flat
Name_url	URL
Price	Price in euros of the flat
M2	Squared meter of the flat
Rooms	Number of bedrooms
Bathrooms	Number of bathrooms
Price/M2	Price (euros) per Square meter
Neighborhood	Name of the neighborhood
Heat	If it has heat or not
Parking	If it has parking or not
Air conditioning	If it has air-conditioning or not
Floor	In which floor of the building the flat is located
Building Year	The year of construction of the flat
Energyconsumption	Ranges from A to G indicating the level of energy consumption measured in kw/S^2
Energyconsumption2	Maximum value of consumption kw/S^2 for each range
Gym	If it has a gym or not
Cuota Comunidad	Administrative Fee
SharedGarden	If it has a shared garden or not
SharedPool	If it has a shared pool or not
Lift	If it has a lift or not
Surveillance	If it has surveillance or not

From this table, the only variable that we did not consider to include for the regression analysis was "Cuota comunidad" that has over 80% of missing values.

Image 11: Missing values Cuota Communidad



	ion	Energyemissions	Gym	Cuota Comunidad	SharedGarden	SharedPool
1	425	G			1	
2	425	F			0	
3	254	E			1	
4	425	G			1	
5	425	G			0	
6	425	G			0	
7	254	E	0	null	0	
8	425	G	0	null	0	
9	425	G	0	null	0	
10	425	G	0	null	0	
11	425	G	0	null	0	
12	425	G	0	null	0	
13	117	D	0	null	0	
14	254	E	0	null	0	
15	117	D	0	null	0	
16	425	G	0	null	0	
17	49	B	0	null	0	
18	254	E	0	null	0	
19	288	E	0	null	0	
20	117	D	0	null	0	
21	76	C	0	null	0	
22	254	E	1	70	1	
23						

Cuota Comunidad
72 (19%) valid | 0 (0%) error | 303 (81%) empty

Name: Sheet1

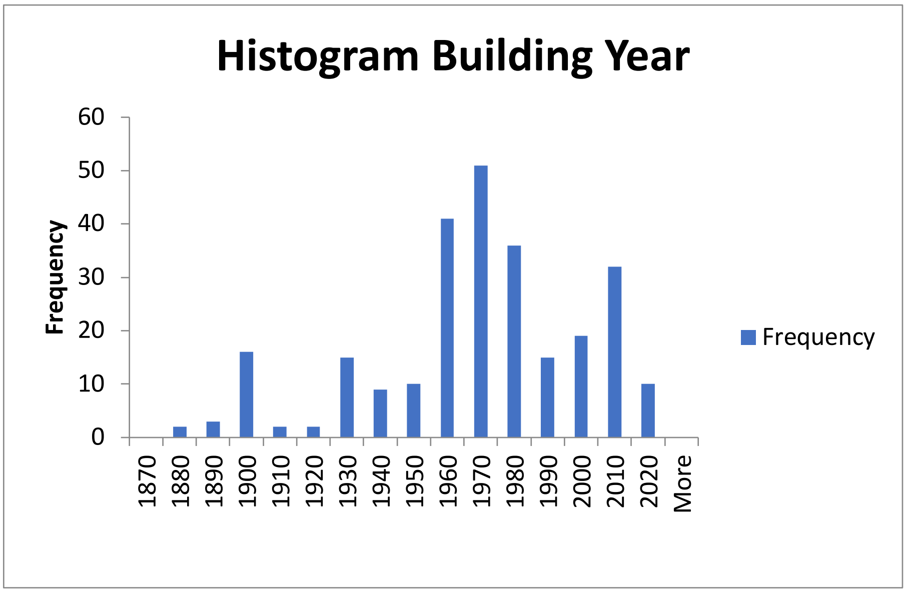
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Added Conditional Column...
Changed Type3
Added Conditional Column...
Filtered Rows
Changed Type4
Reordered Columns

46 COLUMNS, 375 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED ON TUESDAY, DECEMBER 3, 2019

For "building year", 30 % of the data had missing values that we decided to replace with the median of the variable. We used the median because of the distribution of the variable.

Graph 1: Histogram Building Year



Additionally, we looked at the correlation with the other variables, and we decided to keep building year for the analysis because even when we cannot see a higher correlation among the other variables, this variable provides additional understanding for the price.

Table 2: Correlations to Building Year

	Price	*M2*	*Rooms*	*Bathrooms*	*Total Rooms*	*BuildingYear*
BuildingYear	-0.0238058	0.14118849	0.14706701	0.1835281	0.18455623	1

We also made a correlation table for the rest of the numerical variables.

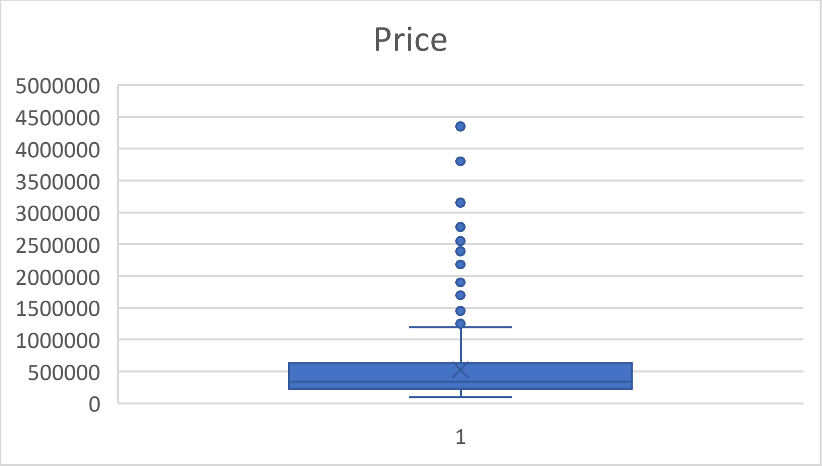
Table 3: Correlations matrix

	Price	*M2*	*Rooms*	*Bathrooms*	*Total Rooms*
Price	1				
M2	0.85367411	1			
Rooms	0.50883343	0.69029603	1		
Bathrooms	0.75897428	0.80699007	0.60197380	1	
Total Rooms	0.70165454	0.83326353	0.90500125	0.88448279	1

From this table we can observe that the square meters variable is the one that has the highest correlation with price, followed by bathrooms and bedrooms. Also, the bathrooms variable has surprisingly more correlation with square meters than the bedroom variable has. We assumed that this is because the number of bathrooms can say more about the size of a flat than the number of bedrooms that not necessarily increase when the square meters does.

Finally, during the exploration of the data, we observed that the variable Price is right skewed indicating that we have outliers corresponding to the more expensive apartments. By investigating this further, we can see that the mean price is 528,481 € while the median is 340,000 €.

Graph 2: Boxplot Price



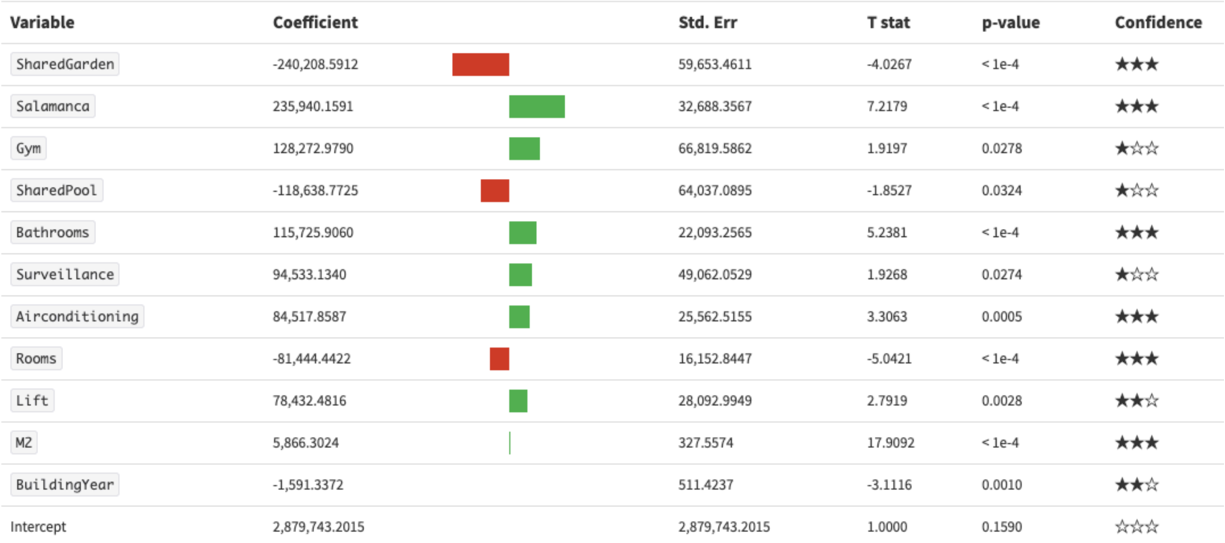
The outliers can be found in the area of Salamanca. They all have an elevator, neither has a pool or shared garden, and almost every apartment has a high energy consumption. We will see if the model addresses these variables with a positive correlation with price, explaining the outlier's nature.

TECHNICAL REPORT

Interpretation of parameter and significance test

After running several simulations on different parameters using the addition method, starting with square meter and adding the others as we saw fit and removing the ones that were not significant, we ran into our best model as shown below:

Image 12: Dataiku Template best Model



Variable	Coefficient	Std. Err	T stat	p-value	Confidence
SharedGarden	-240,208.5912	59,653.4611	-4.0267	< 1e-4	★★★
Salamanca	235,940.1591	32,688.3567	7.2179	< 1e-4	★★★
Gym	128,272.9790	66,819.5862	1.9197	0.0278	★☆☆
SharedPool	-118,638.7725	64,037.0895	-1.8527	0.0324	★☆☆
Bathrooms	115,725.9060	22,093.2565	5.2381	< 1e-4	★★★
Surveillance	94,533.1340	49,062.0529	1.9268	0.0274	★☆☆
Airconditioning	84,517.8587	25,562.5155	3.3063	0.0005	★★★
Rooms	-81,444.4422	16,152.8447	-5.0421	< 1e-4	★★★
Lift	78,432.4816	28,092.9949	2.7919	0.0028	★★☆
M2	5,866.3024	327.5574	17.9092	< 1e-4	★★★
BuildingYear	-1,591.3372	511.4237	-3.1116	0.0010	★★☆
Intercept	2,879,743.2015	2,879,743.2015	1.0000	0.1590	☆☆☆

This model and the way we constructed it helped us reach some interesting conclusions regarding the included and the excluded variables:

- There seems to be no pattern of significance on listing prices for area except for Salamanca. Salamanca stands out because the area has all the fifteen listings that are outliers. Before we filled in the median for missing values in BuildingYear, Retiro and Chamberi also showed significance as an area. Why they changed to insignificant could be explained by the fact that these buildings were of the type older, and when we changed the missing values, these two areas did no longer influence our regression with a p-value of 0.0594 and 0.0663 respectively.
- For size (m2) we ran a regression on rooms and bathrooms to see if these would create a multicollinearity problem on our regression. Our finding was, as suspected, that they did indeed have a positive effect on the size of the apartment. However, by including them in our regression together with size, we found that bathroom had a positive effect on our regression, while rooms have a negative effect. This could be because the more bathrooms you have in an apartment, the more people can live comfortably in the same apartment. You would also want more space in an apartment, and by having more rooms instead of open spaces, an apartment would seem smaller and again you would probably want to pay less.
- For the variable floor we tried multiple things to get use of this variable in our regression. Although our intuition said that price would be lower for ground floor, by testing that with the floor 0 as 1 and all other floors as 0, we did not find any significance. This can be reasoned by many of the highest priced apartments listed are on the lower floors in Salamanca.
- BuildingYear showed a negative correlation with price, decreasing the price by 1,591 € for every year closer to today's date. Why would a newer apartment be less valued in our regression? This can potentially be because of many of the buildings in Madrid are colonial, and building year represent this as a new building can't be colonial.
- Energy consumption turned out to be insignificant, this could be explained due to its many values in the same category. As mentioned above, many of the older apartments is valued higher than the newer ones, having worse energy consumption- and emission, these values shows signs of multicollinearity with BuildingYear.
- Listings that has air-conditioning, lift, surveillance, and gym tend to have higher prices than those who does not. These has the p-values 0.0005, 0.0028, 0.0274 and 0.0278 respectively. These does all intuitively make sense, e.g. Madrid gets very hot during the summer, and air-conditioning can be expensive. When a building has a lift, it will be much easier for you to reach your floor than having to take the stairs every time. When you have surveillance, you feel more secure and feeling safe in your home is important. Lastly, having a gym in your building takes the hassle out of traveling to a gym, removes membership prices and can simply be looked at as a luxury you would have to pay more for.
- Having a shared garden or shared pool had a negative effect on the price with p-values <1e-4 and 0.0324 respectively. This does intuitively not make sense but can be explained with those buildings that have this luxury tend to be further out of the city center and none of the more expensive apartments in our sample had a shared garden or shared pool.
- We also checked if the regression model could have use of heat and parking, but these had insignificant results with our limit of confidence at 95%. Heat can possible be explained by if you have air-conditioning, you also have heat.

- We decided to not use cuota comunidad (administrative fees) at all in our regression due to the high amount of missing data that did not make sense to include in the model.

Global results (R2 and measurements of errors)

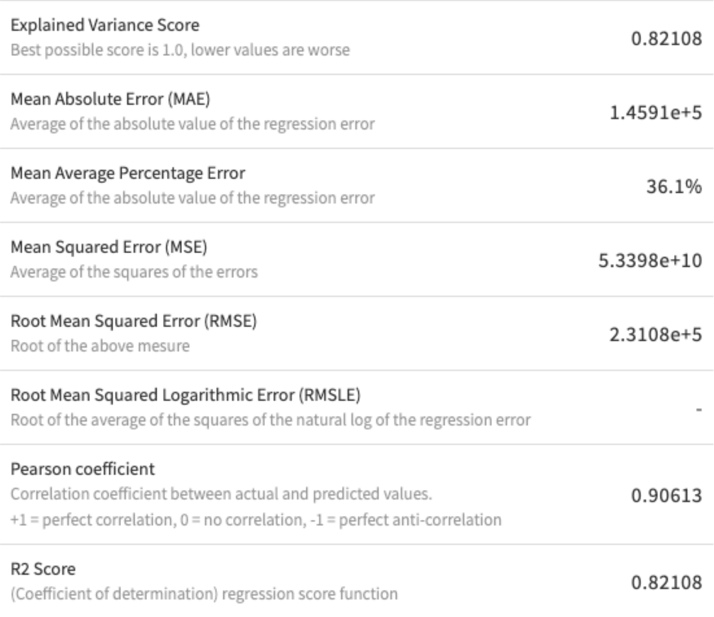
The final model was created by adding one variable at a time starting with the m2, followed by all the others. As we saw some variables had no significance, these were again taken out of the model replaced by new variables to be tested. This method led us to the highest possible R2 while still maintaining only variables with some sort of significance.

Our final model is composed of 11 independent explanatory variables:

$$\begin{aligned} = \ & 2.879.743 + 5.866 \times m^2 - 240.209 \times SharedGarden + 235.940 \times Salamanca \\ & + 128.273 \times Gym - \ 118.639 \times SharedPool + 115.726 \times Bathrooms \\ & + 94.533 \times Surveillance + 84.518 \times Airconditioning - 84.444 \times Rooms \\ & + 78.432 \times Lift - 1.591 \times BuildingYear \end{aligned}$$

Explanations of the R2 and measurements of errors:

Image 13: Model output variables description



Explained Variance Score Best possible score is 1.0, lower values are worse	0.82108
Mean Absolute Error (MAE) Average of the absolute value of the regression error	1.4591e+5
Mean Average Percentage Error Average of the absolute value of the regression error	36.1%
Mean Squared Error (MSE) Average of the squares of the errors	5.3398e+10
Root Mean Squared Error (RMSE) Root of the above mesure	2.3108e+5
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	-
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.90613
R2 Score (Coefficient of determination) regression score function	0.82108

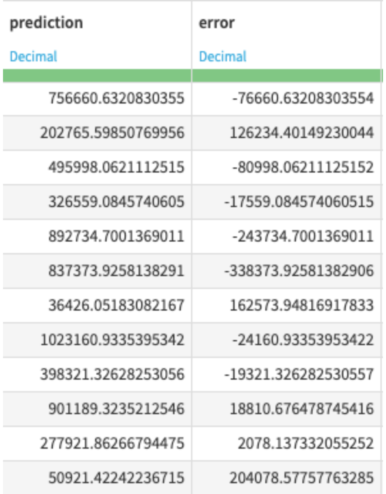
- R2 tells us the percentage of the variation of price explained by our regression. In our case, we have a value of 0.82, which again can be interpreted as our variables can explain 82% of the price of apartments.
- MAE (Mean Absolute error) is the average magnitude of errors in our prediction that ranges from zero to infinity. MAE is an appropriate measure in this case because it is three times as bad to predict 3000 € wrong than 1000 € wrong. Also, RMSE (Root of Mean Squared Error) is more sensitive to outliers because you square the MSE, and as we have a couple outliers in our regression, MAE seems to be the right choice. What does the MAE tell us here? With a value of 145,910, large errors are somewhat likely to have occurred.

- MAPE (Mean Average Percentage Error) gives us the average of our regression in absolute value. With a value of 36.1 % we can say that our MAPE score is reasonable. The percentage of average errors in our regression is 36.1 %. If we had a value below 20 % we would have had a good model, and if it was below 10 % our model would be highly accurate.
- MSE (Mean Squared Error) squares our errors while the RMSE takes the root of these. In our case, MSE has a value of 53,398,000,000 vs. 231,080 for RMSE. What does this tell us? Since values closer to zero are better, MAE and RMSE is confirming our above conclusions that the model created on our dataset shows more errors than we would have liked.

User guide and recommendations

As you can see in the predicted price in the table below, the positive error can be used to your advantage as an individual if you are looking for a new home or an investment for secondhand housing. You can even use the model to check what kind of apartment you should steer away from and not buy. The model can also be used by investment companies, people who want to rent out their apartment, or real-estate agencies. They can use the model to analyze whether to invest or not, and if they should be renting it out or fixing them to re-sell with an even bigger profit. With an R2 of 0.82 the variables give a good explanation of the price, and although there are some errors of importance, the model can be a good indicator if the apartment you are looking at is a good buy or not. How can you do this in real life? By simply filling in the values of our regression, you will get an output that is our predicted price. This will show you some sort of error, and if the result shows you that the apartment is undervalued, you should consider buying it. If the apartment is overvalued, you should not.

Table 4: Model output variables description



prediction	error
Decimal	Decimal
756660.6320830355	-76660.63208303554
202765.59850769956	126234.40149230044
495998.0621112515	-80998.06211125152
326559.0845740605	-17559.084574060515
892734.7001369011	-243734.7001369011
837373.9258138291	-338373.92581382906
36426.05183082167	162573.94816917833
1023160.9335395342	-24160.93353953422
398321.32628253056	-19321.326282530557
901189.3235212546	18810.676478745416
277921.86266794475	2078.137332055252
50921.42242236715	204078.57757763285

As you can see from the table on the next page, the top 10 best buys have many things in common. First, none of the flats have a shared pool or a gym. Second, six out of ten have a lift. Third, six out of ten have air-conditioning. Fourth, they all have from three to five bedrooms. Fifth, only one of the apartments have surveillance, and only one of the apartments have a shared garden. Sixth, the newest apartment is built in 1981. What can we infer from this? These apartments are undervalued, and you can make big bucks by buying all of these apartments.

Table 5: Top 10 best buys

Price	Prediction	Error	M2	Rooms	Bathrooms	Neighborhood	Air-conditioning	Building Year	Gym	Shared Garden	Shared Pool	Lift	Surveillance
970000	1799208	829208	391	5	3	Ciudad Lineal	0	1981	0	1	0	1	0
1950000	2714045	764045	439	5	6	Chamberi	1	1956	0	0	0	0	0
1900000	2598320	698320	439	5	5	Chamberi	1	1956	0	0	0	0	0
1250000	1824171	574171	282	4	3	Salamanca	1	**1967**	0	0	0	1	0
230000	742505	512505	132	3	4	Hortalez	0	**1967**	0	0	0	0	0
695000	1171980	476980	196	3	2	Centro	1	1890	0	0	0	1	0
795000	1237729	442729	250	5	3	Tetuan	0	1965	0	0	0	1	0
649000	1076507	427507	149	3	2	Salamanca	1	1925	0	0	0	1	0
199000	626040	427040	136	4	2	Chamberi	0	**1967**	0	0	0	1	1
670000	1088712	418712	170	4	3	Salamanca	1	1967	0	0	0	0	0

Even though our model has a R2 of 0.82 there are things that could be improved. First, increasing the number of observations can decrease the error measurements making your prediction more accurate. Second, by getting more variables that we were not able to obtain, the model could turn out very different and again reducing the margin of errors. If both of these improvements happened at the same time, the model would be more accurate, and there would be a greater chance of when you try to use the model for yourself, that your apartment purchase would be a good one.