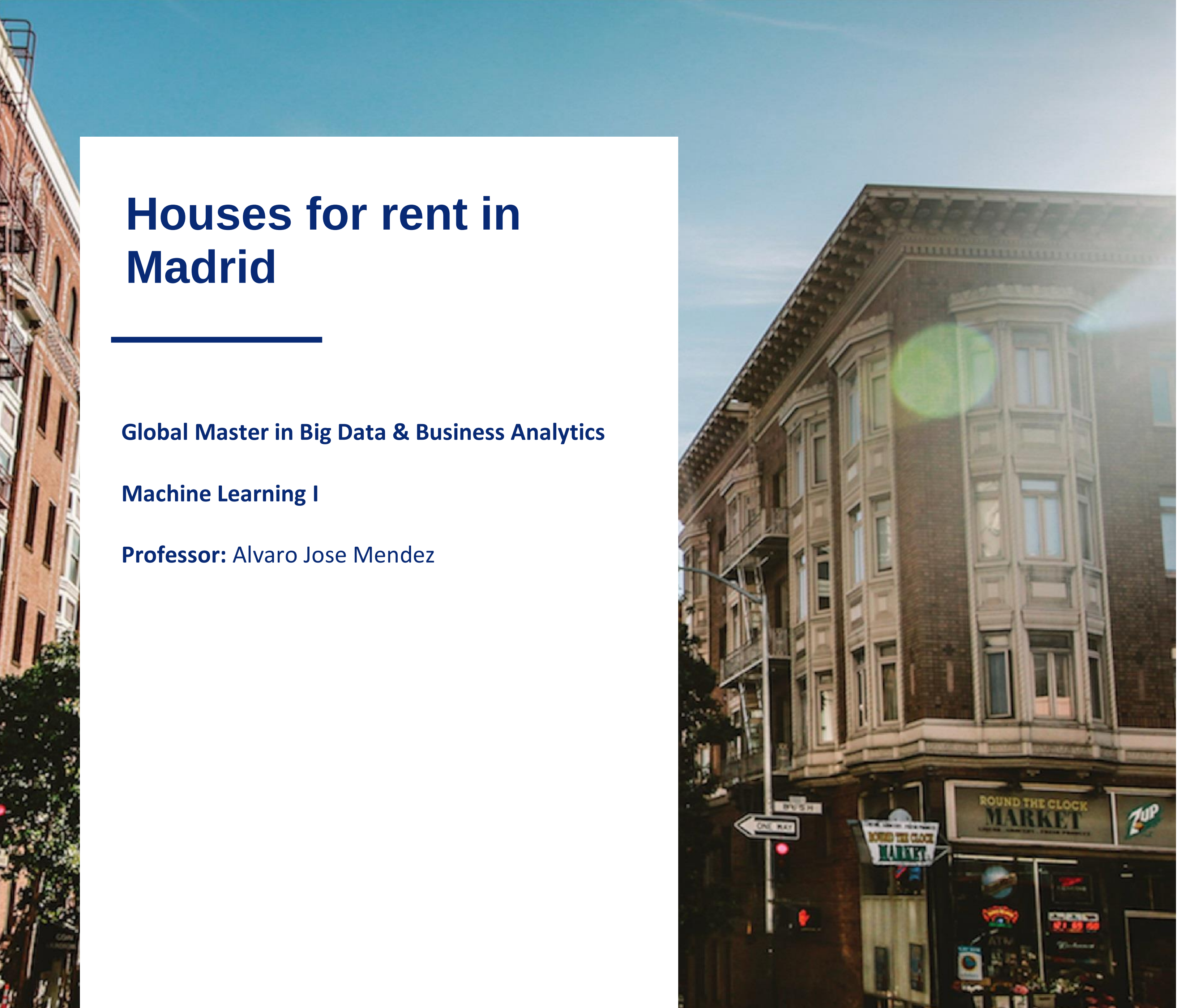


Houses for rent in Madrid

Global Master in Big Data & Business Analytics

Machine Learning I

Professor: Alvaro Jose Mendez

JULY, 2020

TEAM A
Authored by: Abdulaziz Alhagbani, Ckalib Nelson, Michael Wagdy, Sarang Zendehrooh, Nabil Massri

SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Table of Contents

Introduction

The purpose of this executive summary is to explain the model used to **estimate house rental prices** based on a set of explanatory variables related with their characteristics. The dataset used has more than 2000 houses that are available for rent in Madrid. Team A followed the following **machine learning process** to build a Linear Regression Model:

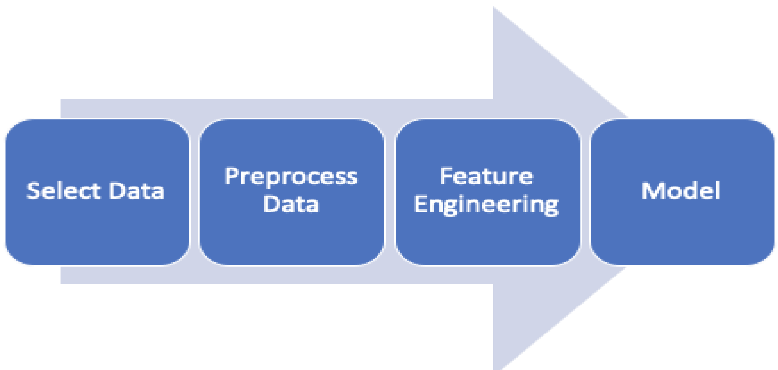


Illustration: *Process of Machine Learning*

Key Findings

- *The model allows users to estimate rental prices of properties not included in the dataset yet have similar characteristics as those houses in the dataset*

Pre-processing

After selecting the data, Team A completed data pre-processing, which includes checking for missing values, checking for categorical variables, and feature scaling. Out of 15 features, there are 6 features that have missing values: Number, Area, Bedrooms, Outer, and Elevator. For the **numerical features** that are missing values, such as Bedrooms, **missing values were imputed with the average of values**. Similar to ID and Address, since Number does not provide any added value, Team A considered it irrelevant for the analysis and thus, excluded the feature model. For the **categorical features** that are missing values, such as Area, Outer, Floor, and Elevator, **missing values were imputed with the most frequent values** and handled using **dummy encoding**. In regard to feature scaling, Team A decided to apply standard rescaling to numerical features.

Feature engineering

Although there were originally 14 explanatory variables, Team A **decided to only use 10 explanatory variables.** The variables Team A excluded are Address, Number, District, and ID (Rent is the dependent variable, meaning that it was excluded as an explanatory variable by default). **Team A also did not create additional variables.** Here is an illustration of the explanatory variables in the model:

Input features				
Floor	Input	**"A"**	Category	Dummy-encode, impute missing
Penthouse		**"A"**	Category	Dummy-encode, impute missing
Outer		**"A"**	Category	Dummy-encode, impute missing
Semidetached		**"A"**	Category	Dummy-encode, impute missing
Duplex		**"A"**	Category	Dummy-encode, impute missing
Area		**"A"**	Category	Dummy-encode, impute missing
Sq. Mt		**#**	Numeric	Avg-std rescaling
Bedrooms		**#**	Numeric	Avg-std rescaling
Elevator		**"A"**	Category	Dummy-encode, impute missing
Cottage		**"A"**	Category	Dummy-encode, impute missing
District		**"A"**	Category	Dummy-encode, impute missing
Rent	Target	**#**	Numeric	

***Table:** Input features*

In an effort to **prevent collinearity**, or a condition in which some of the independent variables are highly correlated, Team A decided **to exclude one of the geographic variables: District**. Team A decided to exclude District over Area given that the later is a more granular explanatory variable, representing a smaller geographic region in Madrid than Area.

Model

The algorithm used to build a linear regression model in Dataiku Data Science Studio (DSS) was the **Ordinary Least Squares (OLS) regression algorithm**. Given that the goal is to explain house rental prices, the target of this model is the variable Rent. The train-test ratio is 80-20, which is the default option in DSS. Roughly 1748 rows were used as the training set and 440 rows were used for the test set. The method of **optimization is based on the R2** since it tends to be the most prominent evaluator for linear regression models.

Coefficients

In linear regression, **coefficients are the values per explanatory variable that change the dependent variable**. For example, per the following illustration, if a house in Madrid does not have a

Penthouse (a binary explanatory variable), the rental price decreases by roughly 273 Euros. Adversely, a house in the Area of Jerónimos, the rental price increases by roughly 1624 Euros.

Regression coefficients	
Variable	***Coefficient***
Area is Jerónimos	1,624
Area is Recoletos	1,416
Area is El Viso	1287
Area is Nino Jesus	1020
Area is Castellana	1019
Area is El Viso	964
Area is Almagro	921

Table: *Sample of Regression Coefficients*

Evaluation Metrics

Given that Team A built a linear **regression model**, there are several evaluation metrics to consider, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2; however, the metrics Team A **decided to optimize the hyperparameters for was R2**. The following illustration details the evaluation metrics of the linear regression model:

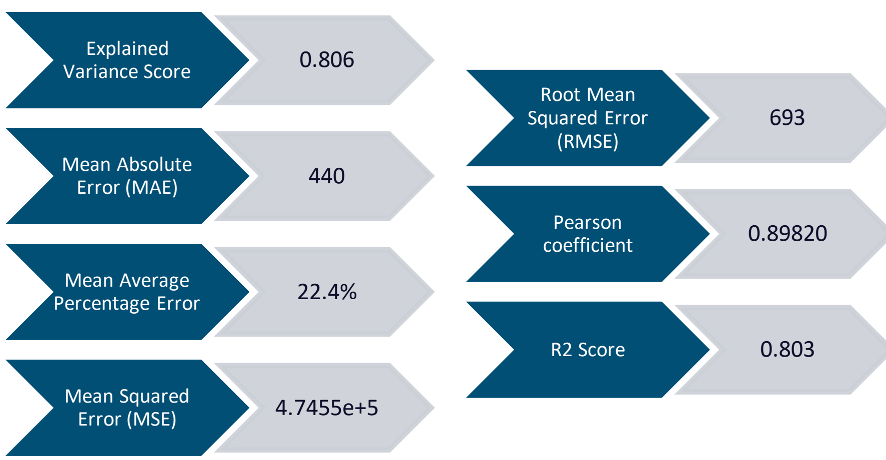


Illustration: *Detailed Metrics*

MAE, MSE, & R2

The **MAE** is the difference between the predicted values and observed values, or the "error" between the predicted values and the observed values. A MAE of 440 is reasonably good knowing that **lower MAE values are better**. Previously run models with less significant variables yielded a greater MAE than 440. The MSE is the summation of the square of distance between predicted values and observed values. An MSE of 4.7455e+5 is fairly good knowing that the **lower MSE values are better**. Previously run models with less significant variables yielded a greater MSE than 47455e+5. **The R2 explains how well the explanatory variables explain the variability** in the dependent model. An R2 of roughly 80% is satisfactorily good knowing that R2 scores closer to 1 are better.

Conclusions

Given that the explanatory variables explain roughly 80% of the variability in the dependent variable, Team A has a good OLS model. As a result, the value add of this model is twofold:

The model allows for the **comparison of rental prices of properties with similar characteristics** in different

The model helps identify the **most and the least expensive districts** in Madrid and recognize the **key features** that affects the variability of the house rentals in Madrid

For example, a house in Madrid, particularly in the Area of Colina, might have the following characteristics:

Area	*Bedrooms*	*Sq. Mt*	*Floor*	*Outer*	*Elevator*	*Penthouse*	*Cottage*	*Duplex*	*Semi Detached*
Colina	2	90-110	3	Yes	Yes	No	No	No	No

Table: *House Characteristics Example*

The model has predicted the rental price of houses with similar characteristics in the area of Colina at roughly 1,290 Euros. This means that rental properties in Colina with similar characteristics have an estimated rental price of 1,290 Euros. Unfortunately, there will be instances in which the predicted value is greater than the actual value; however, **the OLS modelling method is one of the best possible options to determine an estimated rental price**. Furthermore, the **model yields a comparison of estimated rental prices of properties in different Areas**, as shown in the Tableau Dashboard Team A created called Madrid's Houses Rentals. If a potential tenant wanted a rental property with the aforementioned characteristics but had a specific willingness to pay (WTP), this model could help them evaluate their options based on their WTP.

Technical Annex

Data Dictionary

Field Name	Data Type	Data Format	Field Size	Description	Example
ID	Integer			Identifier	3
District	Text			District in Madrid	Ciudad Lineal
Address	Text			Address in Madrid	Piso en calle de Vicente Muzas
Number	Integer			House #	4
Area	Text			Area in Madrid	Colina
Rent	Integer			Monthly rent	1300
Bedrooms	Integer			# of bedrooms	2
Sq.Mt	Integer			Square meters	100
Floor	Integer			# of floors	3
Outer	Integer			Has outer space	1
Elevator	Integer			Has elevator	1
Penthouse	Integer			Has penthouse	0
Cottage	Integer			Has cottage	0
Duplex	Integer			Is a duplex	0
Semidetached	Integer			Is semidetached	0

Regression Summary Table

The regression summary table provides a valuable insight into the model, including: **Coefficient, Standard Error and P Value (Significant level)**.

Out[19]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.801
Model:	OLS	**Adj. R-squared:**	0.788
Method:	Least Squares	**F-statistic:**	59.79
Date:	Thu, 09 Jul 2020	**Prob (F-statistic):**	0.00
Time:	08:06:05	**Log-Likelihood:**	-17378.
No. Observations:	2188	**AIC:**	3.503e+04
Df Residuals:	2049	**BIC:**	3.582e+04
Df Model:	138		
Covariance Type:	nonrobust		

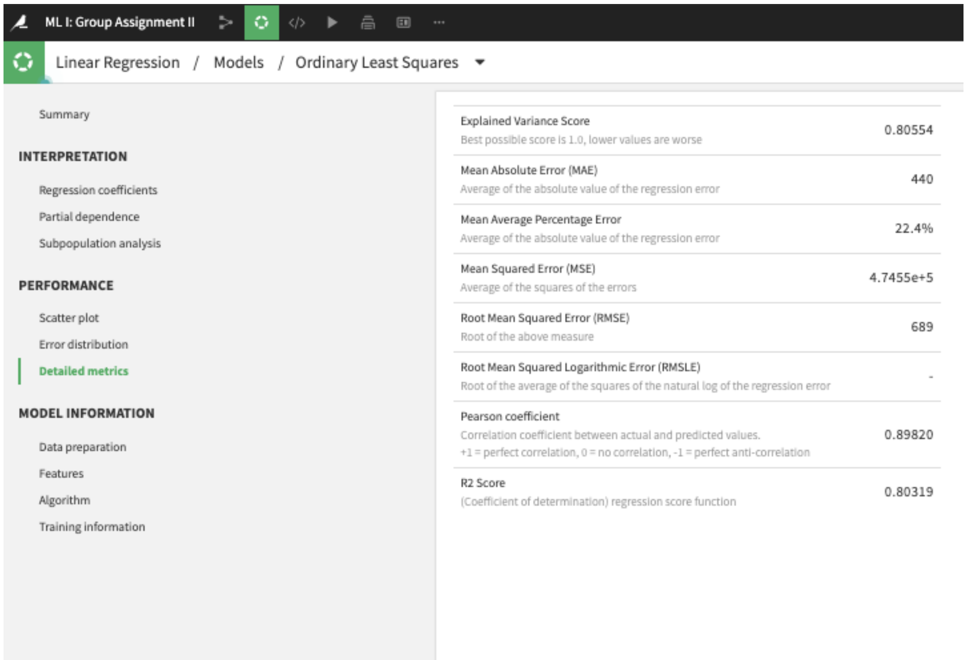
	coef	std err	t	P>|t|	[0.025	0.975]
const	415.2375	141.320	2.938	0.003	138.091	692.384
Bedrooms	80.4286	24.362	3.301	0.001	32.652	128.205
Sq.Mt	1264.2635	30.704	41.176	0.000	1204.049	1324.478
Penthouse_value_0	91.5887	75.478	1.213	0.225	-56.432	239.609
Penthouse_value_1	323.6487	79.783	4.057	0.000	167.185	480.113
Semidetached_value_0	287.9235	103.568	2.780	0.005	84.814	491.033
Semidetached_value_1	127.3140	113.397	1.123	0.262	-95.072	349.700

Table: *Regression Summary*

**We have exported our model into python to get the regression summary table, please find attached python file for full view of the summary table.*

***We acknowledge the values from the summary table are very close to what DSS UI highlights, however, they are not 100% exact.*

Detailed Metrics



Test Data with Predictions

ML I: Group Assignment II

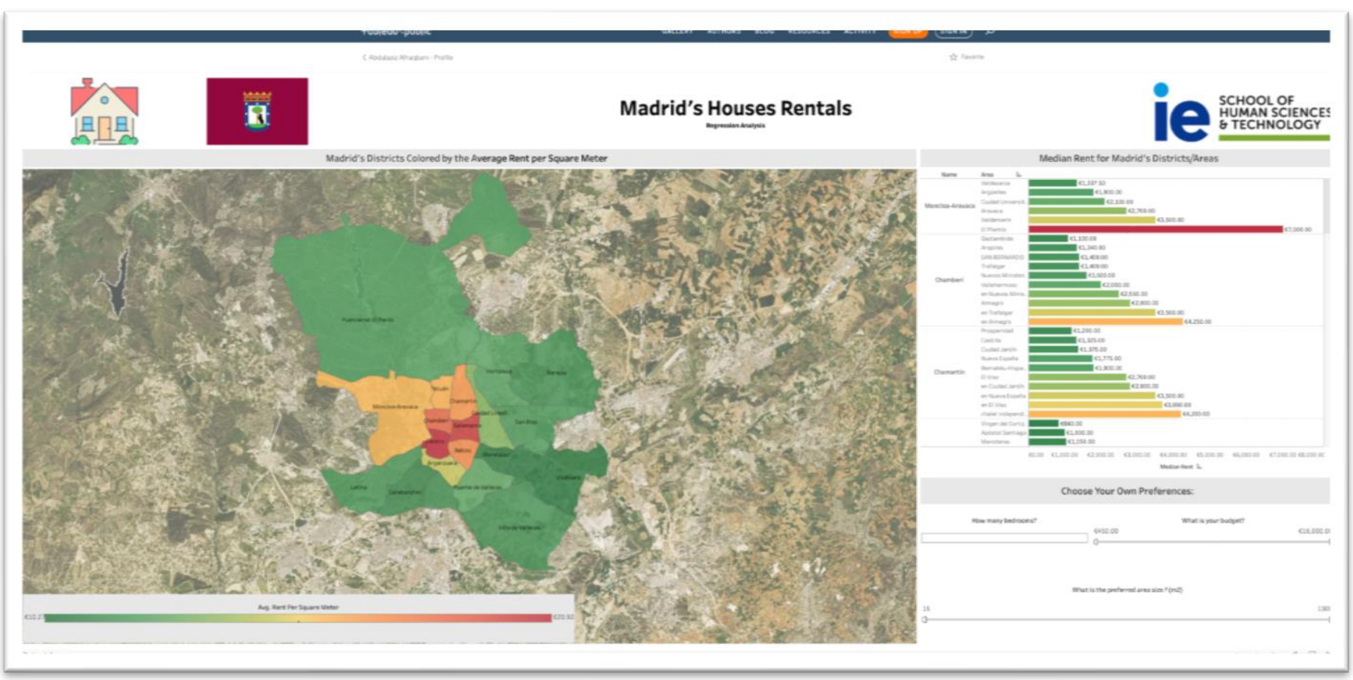
Linear Regression / Models / Ordinary Least Squares

Report | Predicted data | Charts | PUBLISH | DEPLOY

440 matching rows

Id	District	Address	Number	Area	Rent	Bedrooms	Sq.Mt	Floor	Outer	Elevator	Penthouse	Cottage	Duplex	Semidetached	prediction	error	relative_error	error_decile	abs_error_decile
Integer	Text	Natural lang.	Integer	Text	Integer	Integer	Integer	Decimal	Integer	Integer	Integer	Integer	Integer	Integer	Decimal	Decimal	Decimal	Integer	Integer
3	Ciudad Lineal	Piso en calle de Vicente Muzas	4	Colina	1300	2	100	3.0	1	1	0	0	0	0	1290.5947089118768	9.40529108812325	0.007234839298556345	3	0
4	Ciudad Lineal	Piso en calle Badajoz		San Pascual	1600	3	120	4.0	1	1	0	0	0	0	1765.5707366030642	-165.57073660306423	-0.10348171037691514	2	0
6	Ciudad Lineal	Ático en calle José Silva		San Juan Bautista	1850	3	101	5.0	1	1	1	0	0	0	1537.0990848370084	312.9009151629916	0.1691356298178333	4	1
9	Ciudad Lineal	Piso en calle Jazmín	17	Costillares	850	1	52	1.0	1	1	0	0	0	0	819.5206990020456	30.479300997954397	0.03585800117406399	3	0
15	Ciudad Lineal	Piso en calle de Arturo Soria		San Pascual	1700	3	130	1.0	1	1	0	0	0	0	2023.333307633309	-323.333307633309	-0.19019606331371117	2	1
20	Ciudad Lineal	Piso en Ventas		Ventas	830	1	60	3.0	1	1	0	0	0	0	790.741614037536	39.25838596246399	0.04729926019573975	3	0
23	Ciudad Lineal	Piso en Virgen de los reyes		Concepción	1300	3	100	1.0	1	1	0	0	0	0	1275.0794241293675	24.920575870632547	0.01916967374664042	3	0
27	Ciudad Lineal	Piso en calle de Carlota Oâ€™Neill		San Pascual	1100	1	45	4.0	1	1	0	0	0	0	897.5228329435861	202.47716705641392	0.1840701518694672	4	0
36	Ciudad Lineal	Piso en Arturo Soria		Concepción	1200	2	70	3.0	1	1	0	0	0	0	690.5575906817753	509.4424093182247	0.4245353410985206	5	2
40	Ciudad Lineal	Ático en Arturo Soria		Concepción	3000	5	400	4.0	1	1	1	0	0	0	4466.934296388081	-1108.0387373232388	-0.48897809879602694	0	5
53	Ciudad Lineal	Dúplex en San Juan Bautista		San Juan Bautista	3200	3	330	3.0	1	1	0	0	1	0	3471.6621050982867	-271.6621050982867	-0.0848944078432146	2	1
56	Ciudad Lineal	Piso en sor maría de agreda	4	Ventas	680	1	55	3.0	1	1	0	0	0	0	741.6215121392247	-61.621512139224706	-0.09061987079297751	3	0
60	Ciudad Lineal	Piso en calle caribe		San Pascual	2500	3	170	5.0	1	1	0	0	0	0	2234.3289846269968	265.67101537300323	0.10626840614920129	4	1
66	Ciudad Lineal	Estudio en Costillares		Costillares	780		40	0.0	1	1	0	0	0	0	696.7448505251234	83.25514947487659	0.10673737112163666	3	0
67	Ciudad Lineal	Piso en calle de Zigia		Quintana	750	1	49	0.0	1	0	0	0	0	0	694.394667903884	55.605332096115944	0.07414044279482127	3	0
72	Ciudad Lineal	Piso en avenida de Badajoz	23	San Pascual	1600	3	120	4.0	1	1	0	0	0	0	1765.5707366030642	-165.57073660306423	-0.10348171037691514	2	0
74	Ciudad Lineal	Piso en plaza de Viaflor	6	Pueblo Nuevo	700	1	55	1.0	1	0	0	0	0	0	537.8690060497092	162.13099395029076	0.23161570564327252	3	0
76	Ciudad Lineal	Piso en calle de Santa Genoveva		Ventas	825	3	82	1.0	1	1	0	0	0	0	1362.2944720402352	-537.2944720402352	-0.6512660267154365	1	2
77	Ciudad Lineal	Piso en LOPEZ DE HOYOS		Colina	900	2	65	1.0	1	0	0	0	0	0	1093.8610364283877	-193.8610364283877	-0.21540115158709744	2	0
86	Ciudad Lineal	Piso en plaza PLATON		San Pascual	790	2	69	2.0	1	1	0	0	0	0	1252.8258569960128	-462.8258569960128	-0.5858555151848264	1	2
88	Ciudad Lineal	Piso en Normas		San Juan Bautista	1200	3	78	3.0	1	1	0	0	0	0	996.0089694234032	203.9910305765968	0.16999252548049734	4	0
93	Ciudad Lineal	Piso en travesía Cañas	1	Colina	900	2	80	2.0	1	0	0	0	0	0	1135.6023222378267	-235.60232223782668	-0.26178035804202965	2	1
122	Fuencarral	Piso en Las Tablas		La Paz	1250	2	101	3.0	1	1	0	0	0	0	910.2986546176409	339.70134538235914	0.2717610763058873	4	1
126	Fuencarral	Piso en Las Tablas		Las Tablas	1300	2	160	5.0	1	1	0	0	0	0	1766.20914612943	-466.2091461294301	-0.35862242009956163	1	2
138	Fuencarral	Ático en Las Tablas		Las Tablas	1100	1	70	6.0	1	1	1	0	0	0	1203.762535010783	-103.76253501078304	-0.09432957728253004	3	0
144	Fuencarral	Piso en calle ginzo de limia		Pilar	900	3	80	8.0		1	0	0	0	0	999.0677561806917	-99.06775618069173	-0.11007528464521303	3	0
149	Fuencarral	Piso en calle Capiscol		Las Tablas	880	1	61	4.0	1	1	0	0	0	0	750.4507119096429	129.5492880903571	0.14721510010267852	3	0
154	Fuencarral	Casa o chalet independiente en calle de la Peña ...		Mirasierra	2900	4	200				0	1	0	0	1858.065824941381	1041.934175058619	0.3592876465719376	6	4
160	Fuencarral	Piso en avenida Monasterio de el Escorial		Montecarmelo	1350	3	103	1.0	1	1	0	0	0	0	1515.1132703572127	-165.11327035721274	-0.12230612619052796	2	0
161	Fuencarral	Piso en avenida del Santuario de Valverde	98	Montecarmelo	1500	3	135	3.0	1	1	0	0	0	0	1605.3038880410845	-105.30388804108452	-0.07020259202738968	3	0
164	Fuencarral	Piso en calle Julio Palacios	1	La Paz	2300	4	205	5.0	1	1	0	0	0	0	2105.4560455598416	194.54395444015836	0.08458432801746016	4	0
165	Fuencarral	Piso en calle de Ponferrada	19	Pilar	750	2	60	3.0	1	1	0	0	0	0	542.5840193201332	207.41598067986683	0.2765546409064891	4	0
171	Fuencarral	Piso en calle Arzobispo Morcillo	32	La Paz	1500	3	120	5.0	1	1	0	0	0	0	1204.7911256961465	295.20887430385346	0.196805916202569	4	1
174	Fuencarral	Piso en calle Castillo De Candanchú	6	Las Tablas	1200	2	80	0.0	1	1	0	0	0	0	1059.8935808601695	140.10641913983045	0.11675534928319205	3	0
175	Fuencarral	Piso en Pedro Rico	s/n	La Paz	1900	4	165	2.0	1	1	0	0	0	0	1788.8413486806592	111.15865131934083	0.05850455332596886	3	0
177	Fuencarral	Piso en camino de Santiago	33	Las Tablas	1100	3	90	3.0	1	1	0	0	0	0	1101.9380108729606	-1.9380108729606036	-0.0017618280663278215	3	0
190	Fuencarral	Piso en san juan de ortega	20	Las Tablas	945	1	90	0.0	1	1	0	0	0	0	1092.5105970643867	-147.51059706438673	-0.15609586990940394	2	0
196	Fuencarral	Piso en Fuentelarreina		Fuentelarreina	2300	4	275	5.0	1	1	0	0	0	0	2709.169089817247	-409.1690898172469	-0.17789960426836823	2	1
201	Fuencarral	Piso en calle del pico balaitus	3	Peñagrande	650	2	60	3.0	1	1	0	0	0	0	400.7532864480245	249.24671355197552	0.3834564823876546	4	1
206	Fuencarral	Piso en Montecarmelo		Montecarmelo	1350	3	116			1	0	0	0	0	1418.647500827502	-68.64750082750197	-0.05085000061296442	3	0
212	Fuencarral	Piso en calle Fresnedillas	18	Fuentelarreina	880	2	70	1.0	1	0	0	0	0	0	668.8926813338499	211.10731866615015	0.23989468030244335	4	0
214	Fuencarral	Piso en La Paz		La Paz	1900	4	170	1.0	1	1	0	0	0	0	1943.5804704644652	-43.58047046446518	-0.02293708971813957	3	0
215	Fuencarral	Piso en Peñagrande		Peñagrande	1300	3	95	1.0	1	0	0	0	0	0	957.3242281332971	342.6757718667029	0.2635967475897715	4	1
218	Fuencarral	Piso en calle Santiago de Compostela		Pilar	1600	4	160	9.0	1	1	0	0	0	0	1981.3208490572606	-381.3208490572606	-0.23832553066078788	2	1
219	Fuencarral	Piso en calle Ginzo de Limia	37	Pilar	970	3	80	8.0	1	1	0	0	0	0	999.0677561806917	-29.06775618069173	-0.029966758949166734	3	0
225	Hortaleza	Chalet pareado en Canillas		Canillas	4900	5	300				0	1	0	0	3145.4070076610456	1754.5929923389544	0.358080202518154	8	8

Madrid's House Rentals



(click on illustration to visit Tableau Dashboard)