Group Assignment #2

Machine Learning I

PROFESSOR ALVARO JOSÉ MÉNDEZ LÓPEZ

Group D

Adilet Gaparov
Amanda Marques
Benjamín Chumaceiro
David Facusse
Salim Aouzai
Tingting Sun

SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Dearest Real Estate Agency manager,

After carefully reviewing the data with respect to the 2000 ads posted on Idealista, we have created a regression model to help explain and predict rent prices in Madrid more precisely.

In our analysis, we identified that the most important variable for determining the price of rent is the **square meters** of the real estate. **District and average purchasing price** per square meters follow on importance of impacting the rent price. In addition, we found out that in Madrid the floor and elevator had the smallest impact on the price. Finally, the area with higher prices per square meter are Recoletos, Justicia (Chueca) and Universidad (Malasaña) [1].

Please find a list of recommendations at the end of the document, along with a brief explanation of our model.

Data understanding

Data were stored in an Excel file with 12 different attributes. They were:

1. 4 numerical attributes: rent price, number of bedrooms, area of real estate in square meters, and floor.
2. 2 categorical attributes: district and area (*Barrio*).
3. 6 attributes, indicating the existence or presence (yes or no): outer, elevator, penthouse, cottage, duplex and semi-detached.

After a thorough revision of the dataset several inconsistencies were identified, including missing values for area (*barrio*), floor, number of bedrooms, outer, elevator, as well inconsistencies in the names of areas (*barrio*) and districts. Moreover, the dataset contained outliers in terms of square meters of area of real estate.

Data preparation

Area (*Barrio*) had approximately 2% of missing values. The number was not relevant for the total dataset, so the rows with missing values for this attribute were excluded. Considering that we wanted to analyze the prices per area (*barrio*), imputing missing values would affect the results of the analysis.

In addition, we deleted those rows where **Floor**, **Outer** and **Elevator** had no values at the same time. For other cases, where only one of these attributes were missing, we imputed the values. Specifically, Floor values were filled with the median of the remaining data, while for Outer and Elevator we took a conservative approach and filled the values with 0: meaning interior apartment and no elevator. In the Floor column, there were values in date format (which we excluded) and values referring to -0.5, which we interpreted as being what is called "*Sótano*" in Madrid. In addition, another variable called "Elevator_is_useful" was created after considering that it might only be important to have an elevator in the building if the apartment is on the 3rd floor or higher.

Missing values were also found in the number of bedrooms column (**Bedrooms**). In this case, the amount of missing values was more significant and excluding the rows or manipulating the empty fields could bias the dataset. Therefore, a Pearson correlation analysis was applied among features to check if this variable was correlated with the rest. If this was the case, there would be no need to use this variable in our model. It turns out that Bedrooms are approximately 75%

[1] *(Refer to Annex Figure 1)*

explained by area of real estate in square meters and therefore we discarded it for our regression analysis.

We understand that having a predictive model for rent price would be more useful for the agency if the **granularity of analysis is done by Area** (*barrio*) and district. Therefore, we went further on exploring and analyzing the values in the dataset for these variables and found:

1. The model would be **more reliable** if we could condense the Area(*barrio*) variable as it contains 117 categories, which is leading to high cardinality issue;

2. Misspelling of the Areas(*barrio*):

For instance, Nueva España vs. en Nueva España; El Viso vs. En El Viso; Nuevos Ministerios vs en Nuevos Ministerios; Almagro vs en Almagro; Trafalgar vs en Trafalgar; Tres Olivos Valverde vs Tres – Olivos – Valverde. These names refer to the same area and hence they were re-spelled accordingly.

3. Some inconsistencies in the names of the Areas*(barrio)* by district:

District Arganzuela contained areas which are not valid: Esperanza, Gasometro, Oriana, Toledo, Sodio. They are names of streets/metro stations and therefore all of them were merged into the nearest areas (barrios).
District Carabanchel: "Pau de Carabanchel" is part of area "Buena Vista", not an independent area.
District Centro contained areas which are not valid: Plaza de España, Escalinata. Former is a square, latter is street. We merged them to corresponding areas (barrios)
District Chamberi: not valid area is SAN BERNARDO. It is long street that covers several barrios of Chamberi. Since we couldn't relate this to one area, we decided to exclude this advertisement from our dataset.
District Fuencarral: not valid areas are Las Tablas, Montacarmelo, Pena; Montacarmelo is a residential area of the barrio called "El Goloso". Pena refers to Peñagrande, while Las Tablas is part of barrio Valverde.
District Hortaleza: not valid areas are Manoteras, Sanchinarro, Virgen del Cortijo-Manoteras. They are parts of different areas (barrios) in Hortaleza and were replaced by corresponding areas.
District Vicalvaro: some of the areas listed under district Vicalvaro actually belong to another district Villaverde and therefore we added this 21st district to the dataset.

In order to solve for the huge amount of areas, we have included data from external datasets from the *"Ayuntamiento de Madrid"* website (Open Data). We have then included:

- **The surface and population per area & per age ranges**, to get the density.
- **The percentage of foreigners** living in that Area*(barrio)*.
- **The number of restaurants**, bars, cafes and similar venues in the area.
- **The average purchasing price** per square meter of apartments in each area.

Finally, given that regression analysis is sensitive to outliers, we excluded the advertisements from our dataset where the in square meters of the apartment was beyond 482 sq. meters, that is beyond 5 times interquartile range (IQR). After having cleaned the dataset as describe earlier, the dataset contained 8 such outliers. As a result, our model is applicable for those apartments and houses where overall area is below or equal to 482 square meters.

Modeling and validation:

After exploratory data analysis, testing variables on statistical significance and different models, the following were used to build the final model:

- Square meters of the real estate in, floor, overall population of the Area (*barrio*), district, and average price per square meter to buy.

The model is built using **Ordinary Least Squares regression algorithm** and the majority of numerical variables were transformed by applying natural logarithm. The final satisfactory result was Adjusted **R-Squared value of 0.817**, which suggests that approximately 82% of the target variable is explained by the variance of the explanatory variables used.

To validate the accuracy of the model, we checked the average percentage difference between the predicted rent price and the one provided. **The average difference was around 17%**, suggesting the potential opportunities for the real estate agency to increase the price of the rent in some areas. Please find them detailed below.

The final linear regression equation can be found in Annex.

Conclusions and recommendations

After careful analysis and successfully creating and testing the model the conclusion and recommendations taken from the investigation are:

- According to the model, the Real Estate Agency could potentially charge a higher rent for properties in the following areas:
  - Plantio (on average 46% below predicted)
  - Ciudad Universitaria (on average 27% below predicted)
  - Canillas (on average 26% below predicted)
  - Portazgo (on average 24% below predicted)
  - Fontarron (on average 19% below predicted);

- Although it might seem that Salamanca is the most expensive district to rent an apartment in *(figure 7)*, when comparing rent price per square meters, Salamanca is the second expensive district, while Centro is the most expensive one *(figure 8).* [2]
- Apartments located in the areas of Recoletos, Justicia (Chueca) and Universidad (Malasaña) typically have higher rent prices per square meter. [3]
- The most important areas for predicting the rent price are:

Almagro, Arguelles, Castellana, Chueca - Justicia, Concepción, Conde Orgaz-Piovera, Cuatro Caminos, El Viso, Fuentelarreina, Goya, Huertas-Cortes, Ibiza, Jerónimos, Lavapiés-Embajadores, Lista, Malasaña-Universidad, Mirasierra, Niño Jesús, Nueva España, Palacio, Recoletos, Rejas, Sol, Trafalgar, Vallehermoso;

- To have a more precise analysis it might be useful to include other variables related to acquisition power per area;

[2] *(Refer to Annex Figure 7, Figure 8)*

[3] *(Refer to Annex Figure 1)*

- A potential way to further improve the model is to have more information about the apartment itself. From this dataset, there is no way to differentiate between two apartments with the same sq. meters in the same area. Therefore, having more detailed information regarding the interior of the apartments (decoration, lights, disposition of rooms, doorman, bathrooms) could allow to predict rent price with smaller errors.

Technical annex

The final equation for linear regression model can be stated as follows:

$$\begin{gathered}\ln Rent\ Price = 4.080718344 + 0.71157305 * \ln Sq.meters + 0.072510812 * \\ \ln(Floor + 2) + 0.192502592 * standardized\ Price\ per\ sq.meter - 0.025063408 * \\ \ln Population\ of\ \ Area + b\_district,\end{gathered}$$

where *ln* refers to natural logarithm, *b_district* coefficient depends on the district (refer to Table 1), *“standardized Price per sq.meter”* and “ln *Population of Area”* can be taken from Excel file, we attached to this letter. Since we cannot take logarithm from 0 and Floor values include 0, we are adding 2 to *Floor* before taking natural logarithm.

Table 1. Coefficient to add to the linear regression equation, depending on the district

District	*b_district*
Moncloa	0.210604406
Centro	0.209849816
Retiro	0.177003707
Chamberi	0.152015541
Tetuan	0.151148727
Chamartin	0.135350512
Puente de Vallecas	0.132710564
Ciudad Lineal	0.130465809
Salamanca	0.129271996
Arganzuela	0.08309085
Hortaleza	0.062574679
Fuencarral	0.054858566
Usera	0.044049073
Latina	0.026634997
San Blas	0.003086894
Vicalvaro (reference)	0
Carabanchel	-0.003420285
Villaverde	-0.054400431
Villa de Vallecas	-0.057756388
Barajas	-0.063567361
Moratalaz	-0.068510243

Figure 1. The most expensive areas to rent on average (predicted rent price per square meter)

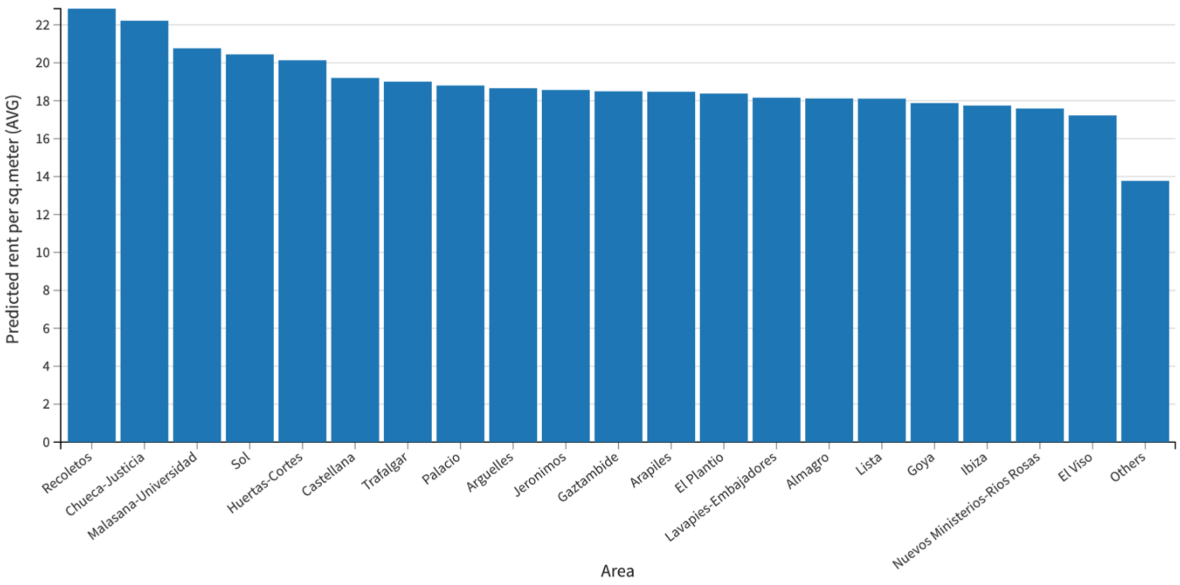
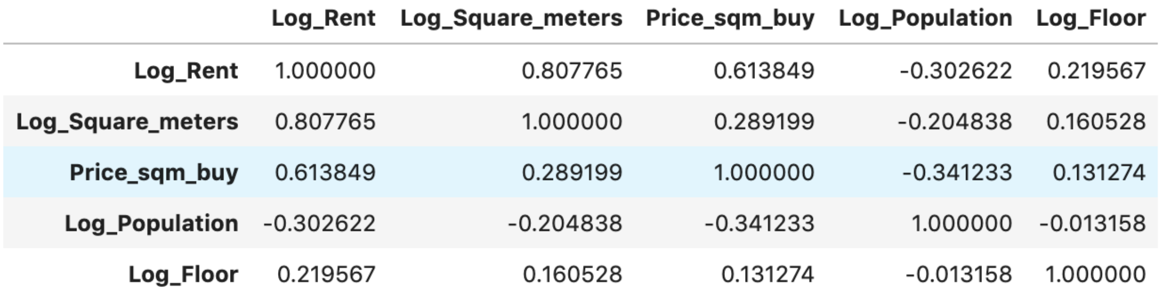


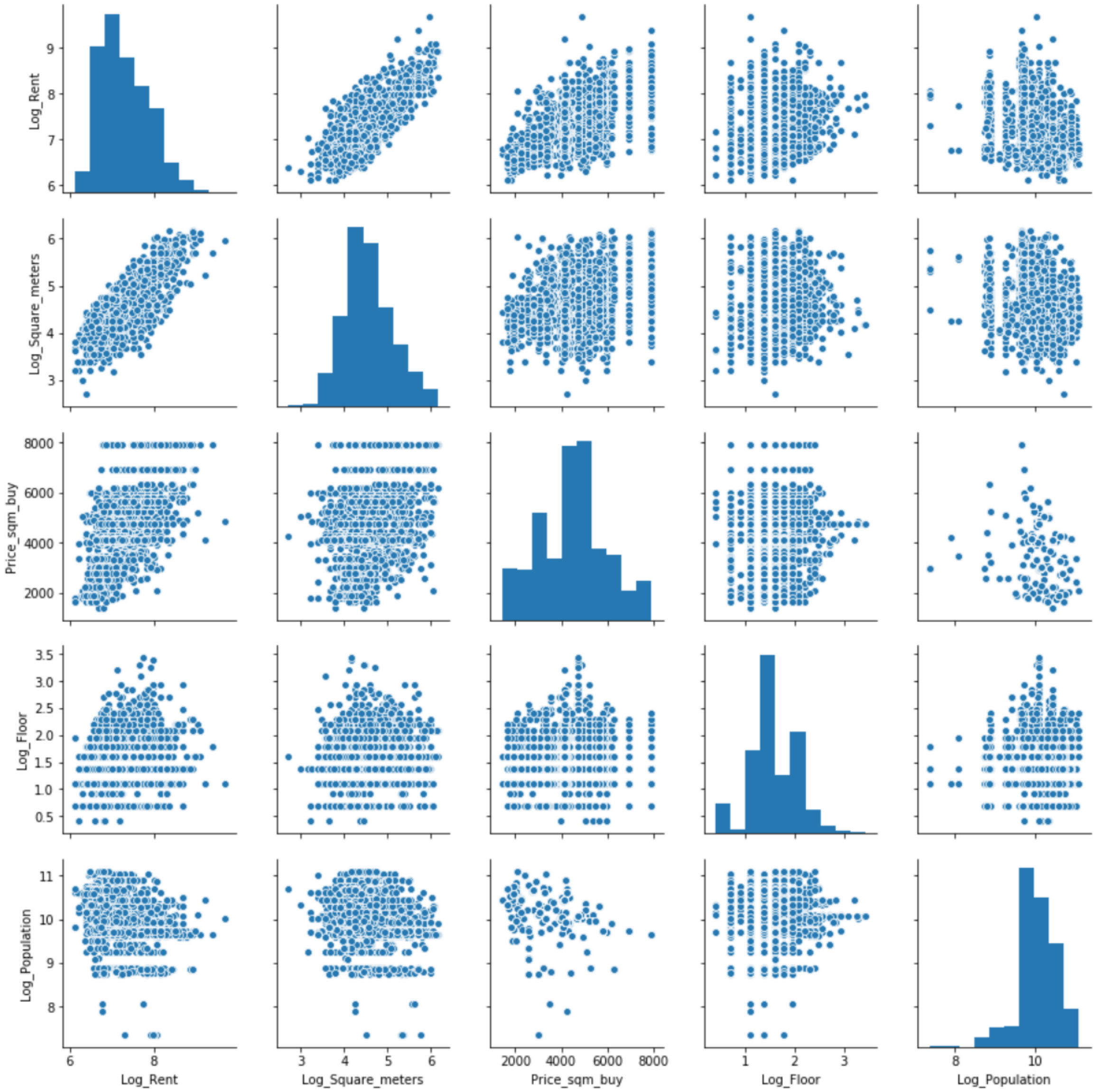
Figure 2. Pearson Correlation Analysis for final list of attributes



	Log_Rent	Log_Square_meters	Price_sqm_buy	Log_Population	Log_Floor
Log_Rent	1.000000	0.807765	0.613849	-0.302622	0.219567
Log_Square_meters	0.807765	1.000000	0.289199	-0.204838	0.160528
Price_sqm_buy	0.613849	0.289199	1.000000	-0.341233	0.131274
Log_Population	-0.302622	-0.204838	-0.341233	1.000000	-0.013158
Log_Floor	0.219567	0.160528	0.131274	-0.013158	1.000000

Price_sqm_buy refers to average price per square meters for each area.

Figure 3. Relationship between variables and their distributions



Price_sqm_buy refers to average price per square meters for each area.

Figure 4. Actual values of *ln Rent* vs predicted values of *ln Rent*

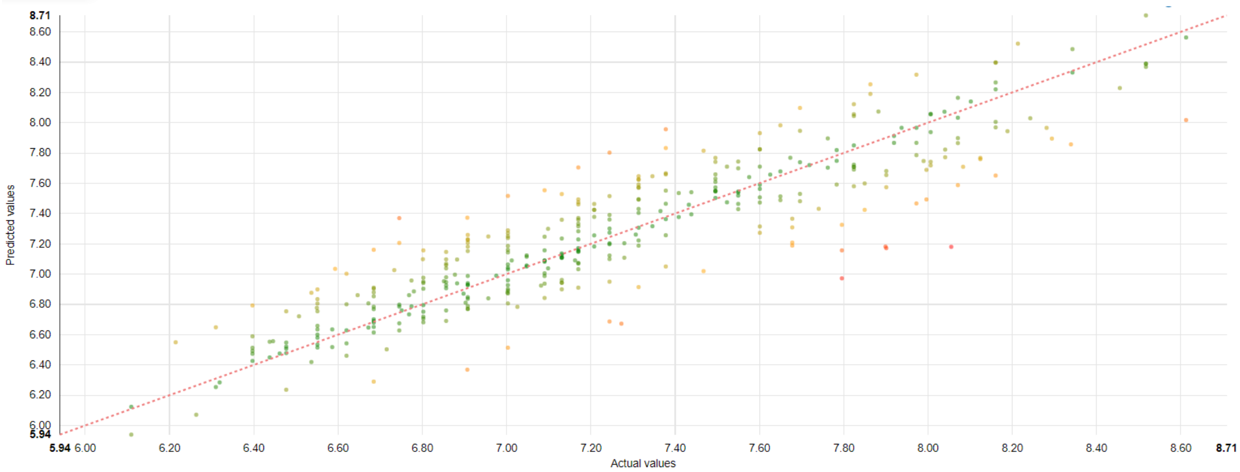
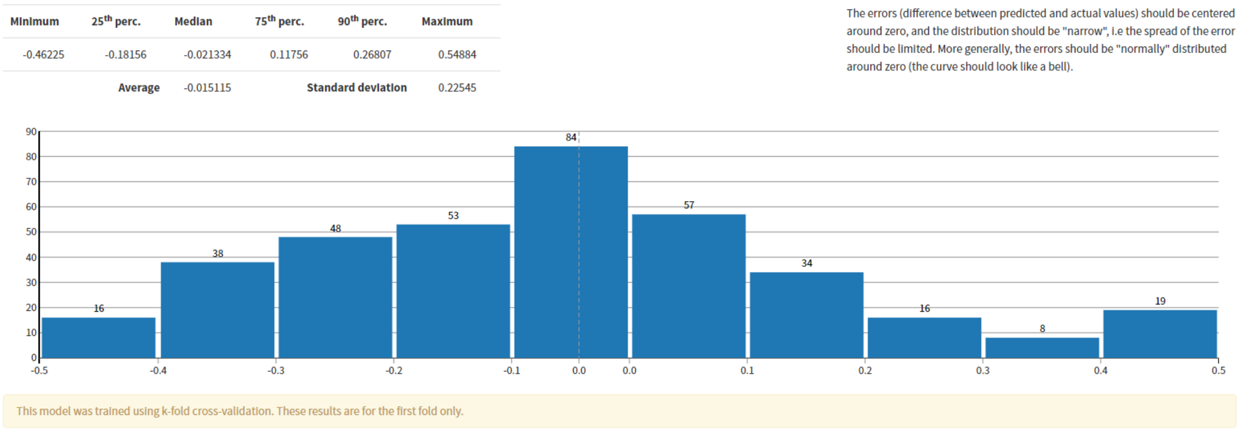


Figure 5. Error are normally distributed and centered around zero, as expected according to assumptions of linear regression.

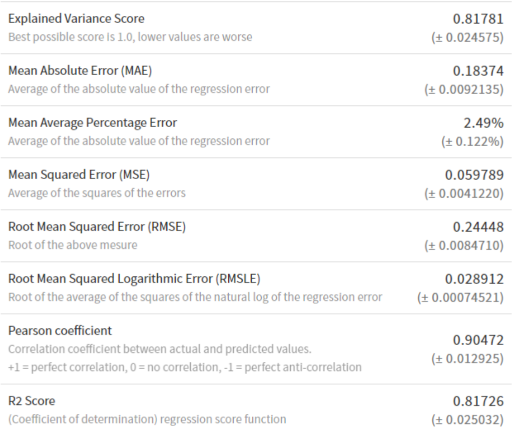


Minimum	25th perc.	Median	75th perc.	90th perc.	Maximum
-0.46225	-0.18156	-0.021334	0.11756	0.26807	0.54884
	Average	-0.015115		Standard deviation	0.22545

The errors (difference between predicted and actual values) should be centered around zero, and the distribution should be "narrow", i.e the spread of the error should be limited. More generally, the errors should be "normally" distributed around zero (the curve should look like a bell).

This model was trained using k-fold cross-validation. These results are for the first fold only.

Figure 6. Detailed review of evaluation metrics



Explained Variance Score Best possible score is 1.0, lower values are worse	0.81781 (± 0.024575)
Mean Absolute Error (MAE) Average of the absolute value of the regression error	0.18374 (± 0.0092135)
Mean Average Percentage Error Average of the absolute value of the regression error	2.49% (± 0.122%)
Mean Squared Error (MSE) Average of the squares of the errors	0.059789 (± 0.0041220)
Root Mean Squared Error (RMSE) Root of the above mesure	0.24448 (± 0.0084710)
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.028912 (± 0.00074521)
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.90472 (± 0.012925)
R2 Score (Coefficient of determination) regression score function	0.81726 (± 0.025032)

Figure 7. The most expensive districts by average rent price

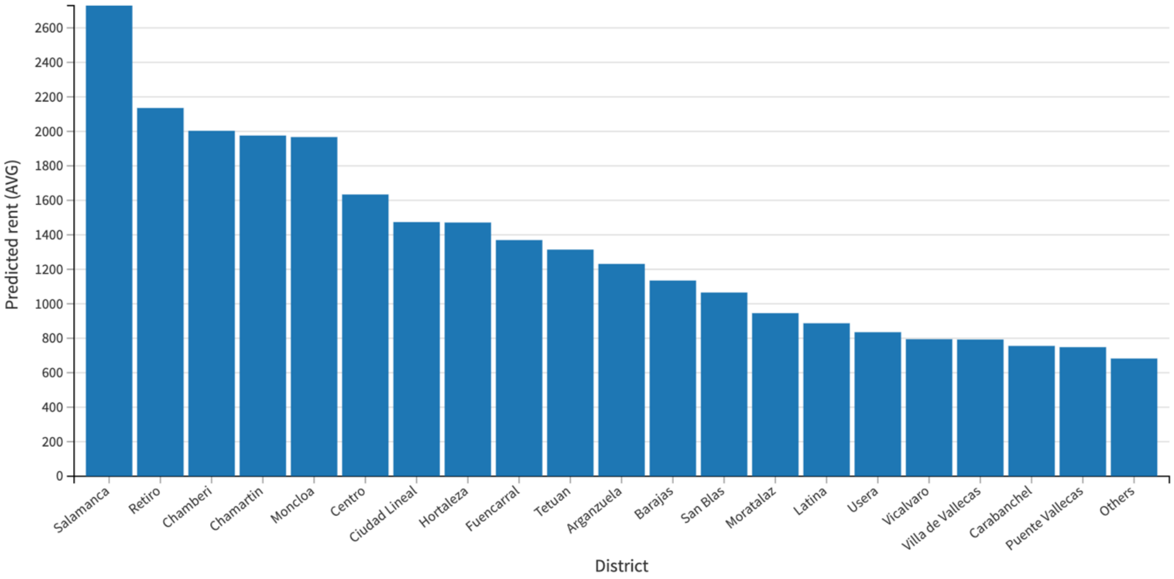
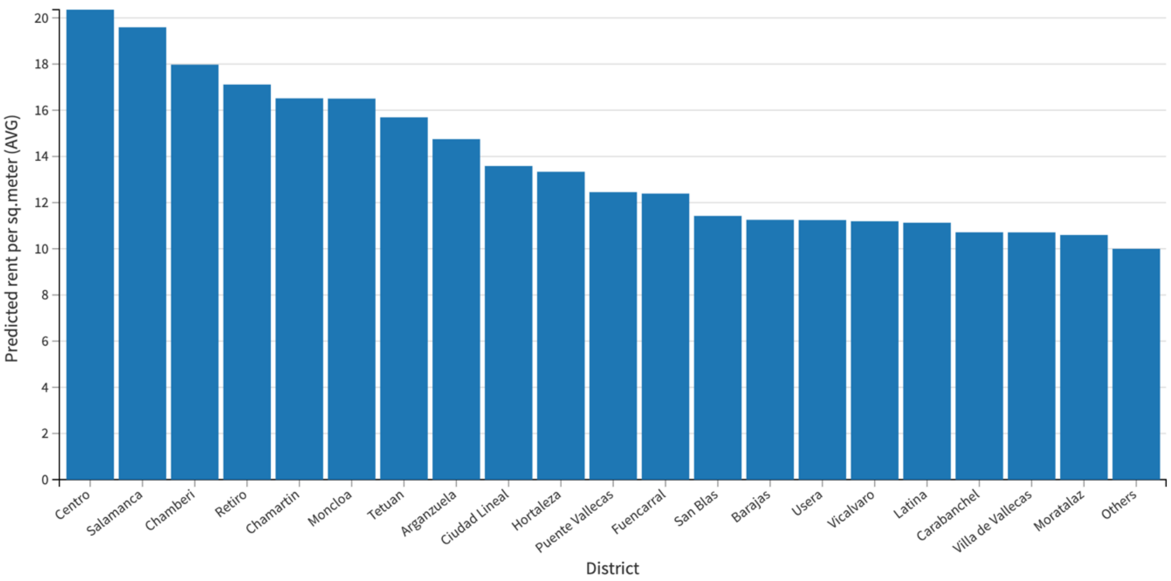


Figure 8. The most expensive districts by average rent price per square meters


Predicted rent per sq.meter (AVG)
20
18
16
14
12
10
8
6
4
2
0
Centro
Salamanca
Chamberi
Retiro
Chamartin
Moncloa
Tetuan
Arganzuela
Ciudad Lineal
Hortaleza
Puente Vallecas
Fuencarral
San Blas
Barajas
Usera
Vicalvaro
Latina
Carabanchel
Villa de Vallecas
Moratalaz
Others
District