

Machine Learning Report

Linear Regression



House Rental Price Model
June/2023
TURING CONSULTING

Table of Contents

1. Executive Summary	3
2. Data audit, cleaning and interpretation	3
3. Technical approach	4
4. Model evaluation	5
5. Conclusion and Recommendations	6
6. Technical Annex	8

1. Executive Summary

The Turing Consulting has conducted a linear approach for modeling the relationship between a scalar response, the house rental price in Madrid, and one or more explanatory variables. The goal of the model is to help estimate the rental price and find good opportunities in the market (flats that may be under their theoretical estimated price).

Initially, the dataset was audit and clean to assure the model defined would be reliable. As such, Turing consulting proceeded to eliminate null values, create new variables (e.g. Binary variables for each district), use additional artifacts to improve the analysis, and remove outliers (rents above 10,000€). A bivariate analysis was performed to investigate the relationship between the rental price and the possible explanatory variables. A strong positive relationship was found between the *rent* and the *Sq.Mt* (around 0.82), and weak relation between *rent* and *floor*. To evaluate the model two data sets were created, one for training (80%) and another for validation (20%). This was later used to confirm the robustness of the model. The criteria for including variables was avoiding multicollinearity and being statistically significant for a 95% level of confidence (p-value lower than 0.05). After running the model several times, it was concluded that the rental price in madrid could be described using the below equation:

$$\begin{aligned} \text{Predicted Rent} = & -100.9 + 10.3 * \text{Sq.Mt} + 45.1 * \text{Floor} + 211.4 * \text{Outer} \\ & + 608.8 * D_Centro + 518.1 * D_chamartin + 550 * D_chamberi - 114.2 * D_fuencarral \\ & + 781.4 * D_retiro + 965.7 * D_salamanca \end{aligned}$$

Considering the model, a pattern emerges that can reveal potential opportunities for the top 25% of highest price differences between theoretical prices and real rent prices. Similarly, properties located on higher floors also appear to be undervalued, presenting potential opportunities for investors or renters looking for better deals. Nearly all of these top price-difference properties are outward-facing, suggesting a potential bias in the market pricing these types of properties lower. As such, the main recommendation is to **focus on larger, outward-facing properties on higher floors, regardless of the district, which could uncover undervalued opportunities in the rental market.**

The consistent model performance in both training and testing phases instills confidence in the predictive ability of this model. However, to ensure the model stays robust and relevant over time, it's important to continue monitoring its performance, retrain it with fresh data as necessary, and consider incorporating additional relevant features if they become available.

2. Data audit, cleaning and interpretation

Data audit and cleaning

Before starting analyzing the relationship between the variables, a thorough inspection of the dataset was made. The dataset included data on the following variables: *District, Address, Area, Number, Rent, Bedrooms, Sq.Mt, Floor, Outer, Elevator, Penthouse, Cottage, Duplex* and *Semidetached*. Our approach was to remove the null values in the dataset, when it was impossible to infer any value (for more details see Annex A1).

Data interpretation

Once finalized the data cleaning, a bivariate analysis was done to investigate the relationship between the rental price and the candidates to explanatory variables. Some of these relationships have been visualized in a scatter plot (for more detail see Annex A2 and A3), where it was observed the positive/negative and low/high correlation between the different variables of the original dataset with the dependent variable *Rent*. The correlation matrix (Annex A4) supports the understanding of the scatter plots. It was observed a strong positive correlation between *Sq.Mt* and the dependent variable, and a weak relationship between rent and floor number.

In order to better understand the effect of different districts on the rent price, another graph was analyzed (Annex A5). The variable *District* can be an important variable to consider, as there is a considerable difference between the rental prices for the different districts. As such, in order to be incorporated into the analysis, the district was converted into more variables using the “dummyfication” technique.

In order to improve correlation between the explanatory variables and rent, two approaches were tested: logarithms transformation and the creation of new variables. Logarithm transformations help to normalize skewed data and sometimes improve correlations, with that aim following log-transformation to the continuous variables were calculated: $\ln Rent = \log(Rent)$, $\ln Sq.Mt = \log(Sq.Mt)$. The variables created were *Bedrooms/Sq.Mt*, which is the division of both variables with the aim of identifying different types of houses, and the multiplication of the following: *Elevator x Penthouse*, *Bedrooms x Duplex*, *Outer x Penthouse* and *Outer x Duplex*. Nevertheless, none of these approaches improved these relationships (Annex A6).

It is also relevant to consider the strong correlations between variables like *Sq.Mt.*, *Bedrooms*, *Elevator* and *Cottage* to avoid multicollinearity in the upcoming model fitting.

Outliers

It was observed in the scatter plot that some records do not follow the general trend, specifically for a high range of values for rent prices. Therefore, the further analysis will not consider values for *Rent* equal or above 10.000€.

3. Technical approach

The dataset was split into two sets: training and validation. This splitting was performed randomly using the `train_test_split` function of the `sklearn.model_selection` Python library with `test_size = 0.2` and `random_state = 101`.

To initial model was run considering several variables (Annex A7): *Bedrooms, Sq.Mt, Floor, Outer, Elevator, Penthouse, Cottage, Duplex, Semidetached, Bedrooms / Sq.Mt, Floor x Elevator, Elevator x Penthouse, Bedrooms x Duplex, Outer x Penthouse, Outer x Duplex*, and the binary variables of each district. Notice that the last variables are the ones created to include the district in the model (*District_argazuela* to avoid the dummy's trap).

The procedure followed implies an iterative process of elimination of the insignificant variables from the model, P-value higher than 0,05 (95% confidence level), until significant variables are obtained. After that it is important to check the correlation and eliminate the less significant one(s) to avoid multicollinearity. Several models were run (using the OLS function from the statsmodels.api Python library) and the following results were obtained:

The first model results (Annex A8):

- $R^2 = 0.783$ and adjusted $R^2 = 0.779$
- Significant variables: Bedrooms, Sq.Mt, Floor, Outer, Bedrooms / Sq.Mt, District_carabanchel, District_centro, District_chamartin, District_chamberi, District_fuencarral, District_retiro, District_salamanca and District_vicalvaro.

As per the results, for the second regression only the significant variables described above were included, *Bedrooms* was eliminated to avoid multicollinearity and also the non-significant variables were removed.

The second model results (Annex A9):

- $R^2 = 0.768$ and adjusted $R^2 = 0.768$
- Significant variables: Sq.Mt, Floor, Outer, District_centro, District_chamartin, District_chamberi, District_fuencarral, District_retiro, and District_salamanca.

Taking the conclusions from the 2nd model, only the variables described above were included in the final model (Annex A10). The result was an R^2 of 0.767 and an adjusted R^2 of 0.765. Correlation was checked one final time, and no multicollinearity issues were found (Annex A11). Even though the floor variable had previously demonstrated low correlation with rent, the low p-value means it is statistically significant, therefore it was included in the model.

4. Model evaluation

One of the main criteria to assure a good model is a high R^2 , the 76% obtained with the final model assures good predictions. On the other hand the independent variables are not highly correlated so good interpretations can be made of the parameters.

In this scatter plot (annex A12) it is observable that by applying the model to the test set the linear relationship between the test values and the predictions remains, actually, the correlation between these two variables is 0.88 which corresponds to R^2 of 0.764. This means that the model is not overfitted.

Additionally, by looking at the histogram of the errors (annex 13) it can be observed that the error is centered around zero and maintains symmetric distribution, another indicator that the model has no biases and that the errors are small.

The following metrics were extracted:

- **Mean Absolute Error (MAE):** 417.42. This metric indicates that, on average, the predictions are around 417.42 away from the actual prices, regardless of the direction. This is a good result, as it is a lower value compared to the minimum value in the dataset.
- **Mean Squared Error (MSE):** 385521.67. This indicator is more sensitive to large errors than MAE, because it squares the differences before averaging them. The relatively low value of MSE in the model suggests that there are not many outliers or large errors in the predictions.

- **Root Mean Squared Error (RMSE):** 620.90. This can be interpreted similarly to the standard deviation. This result suggests that the prediction errors typically have a magnitude of about 620.9 units. Since RMSE gives a relatively high weight to large errors, the model's low RMSE is a good indicator of quality of the model.

5. Conclusion and Recommendations

In conclusion, the regression model for predicting house rental prices has performed well in both training and testing stages. It exhibits an R-squared value of 0.76, indicating that it accounts for approximately 76% of the variability in the rental prices.

This strong performance holds true in the training phase (utilizing 80% of the data) as well as in the validation or test phase. The consistency in the R2 value in both training and testing phases implies that the model generalizes well and is not overfitting the training data.

$$\begin{aligned} \text{Predicted Rent} = & -100.9 + 10.3 * \text{Sq.Mt} + 45.1 * \text{Floor} + 211.4 * \text{Outer} \\ & + 608.8 * D_{\text{Centro}} + 518.1 * D_{\text{chamartin}} + 550 * D_{\text{chamberi}} - 114.2 * D_{\text{fuencarral}} \\ & + 781.4 * D_{\text{retiro}} + 965.7 * D_{\text{salamanca}} \end{aligned}$$

Examining the individual feature contributions:

- The most relevant variable in explaining the dependent is Sq.Mt
- As the dependent variable Rent is in level, the explanation of the beta coefficient for Sq.Mt: 1 additional square meter, has a positive effect on the rental price by, on average, 10.33 units.
- A house in the Salamanca district means that, on average, the rent will cost 965.7 units more.
- The average rent decreases in the District of Fuencarral
- Higher floor numbers are expected to increase the rental price.
- Outer properties cost 211.43 units more, on average, in rent compared to non-outer properties.

The district of the property significantly affects the rental price, with properties in districts Centro, Chamartin, Chamberi, Retiro, and Salamanca costing more compared to the reference district, while properties in district Fuencarral cost less.

The consistent model performance in both training and testing phases instills confidence in the predictive ability of this model. It implies that it can be expected to perform similarly when this model is applied to new data in future predictions.

Upon reviewing the properties that fall into the top 25% of highest price differences (for more detail see Annex A14), a pattern emerges that suggests where potential opportunities may lie. **Larger properties**, as indicated by **greater square meters (Sq.Mt)**, **tend to exhibit more significant price** differences, implying that these larger areas might be undervalued and **represent lucrative opportunities**. Similarly, properties located on higher floors also appear to be undervalued, presenting potential opportunities for investors or renters looking for better deals.

Nearly all of these top price-difference properties are outward-facing, suggesting a potential bias in the market pricing these types of properties lower.

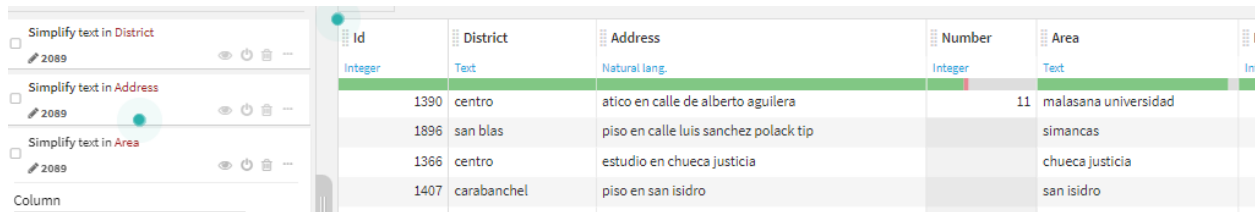
Interestingly, the district location does not seem to play a significant role in these high price-difference properties, suggesting that good investment opportunities are not confined to any specific districts and could be found throughout the market.

Thus, focusing on larger, outward-facing properties on higher floors and considering options across various districts could be a promising strategy for uncovering undervalued opportunities in the rental market.

6. Technical Annex

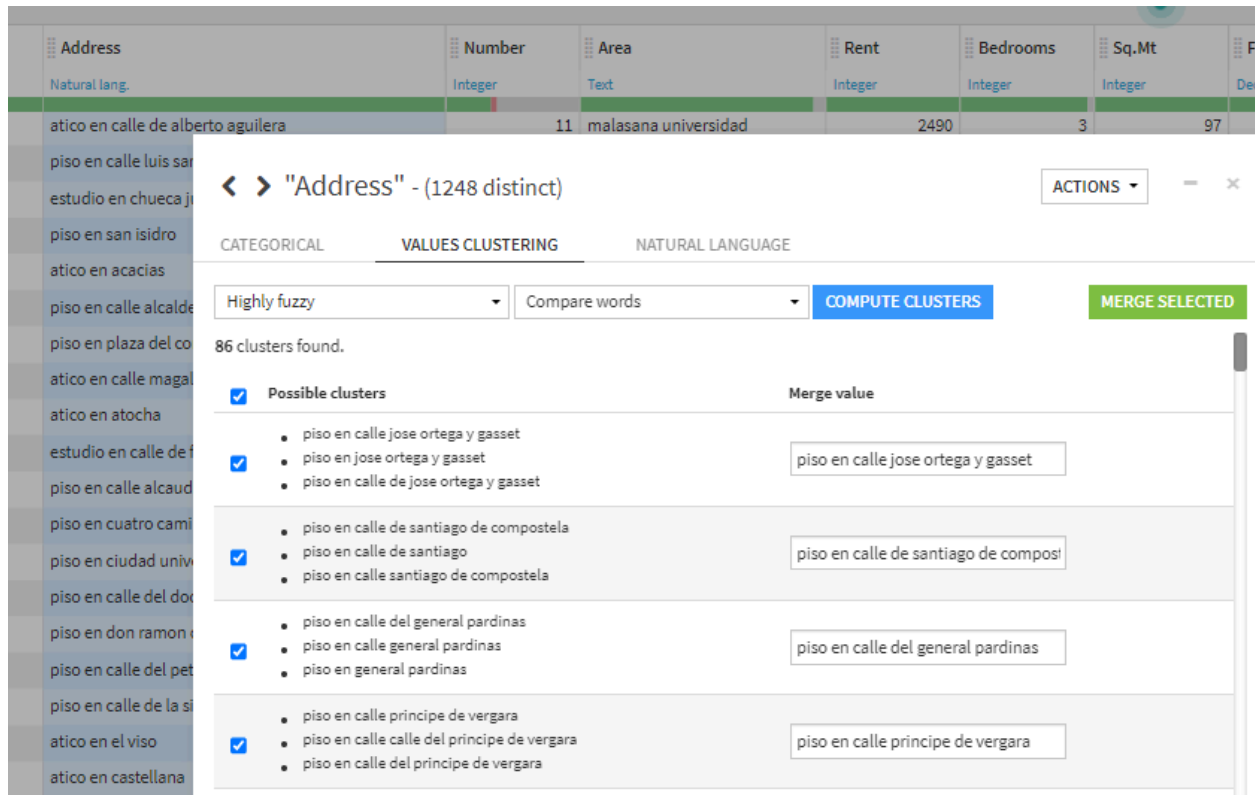
A1 : Regarding District, Address and Area special characters had to be eliminated and for Address and Area the values were harmonized by equal description. For the four blanks observations in Area, the correct value was extrapolated using the Address field of each record. As per analysis criteria we will consider that except for a “Piso” & “Ático” that could be interior, the rest will be treated as Outer.

1. Simplify text in District, Address and Area to eliminate special characters.



Id	District	Address	Number	Area
1390	centro	atico en calle de alberto aguiera	11	malasana universidad
1896	san blas	piso en calle luis sanchez polack tip		simancas
1366	centro	estudio en chueca justicia		chueca justicia
1407	carabanchel	piso en san isidro		san isidro

2. Address & Area harmonization by equal description:



Address	Number	Area	Rent	Bedrooms	Sq.Mt
atico en calle de alberto aguiera	11	malasana universidad	2490	3	97

< > "Address" - (1248 distinct)

CATEGORICAL **VALUES CLUSTERING** NATURAL LANGUAGE

Highly fuzzy Compare words **COMPUTE CLUSTERS** **MERGE SELECTED**

86 clusters found.

Possible clusters	Merge value
<input checked="" type="checkbox"/> <ul style="list-style-type: none"> piso en calle jose ortega y gasset piso en jose ortega y gasset piso en calle de jose ortega y gasset 	piso en calle jose ortega y gasset
<input checked="" type="checkbox"/> <ul style="list-style-type: none"> piso en calle de santiago de compostela piso en calle de santiago piso en calle santiago de compostela 	piso en calle de santiago de compostela
<input checked="" type="checkbox"/> <ul style="list-style-type: none"> piso en calle del general pardinas piso en calle general pardinas piso en general pardinas 	piso en calle del general pardinas
<input checked="" type="checkbox"/> <ul style="list-style-type: none"> piso en calle principe de vergara piso en calle calle del principe de vergara piso en calle del principe de vergara 	piso en calle principe de vergara

< > "Area" - (140 distinct)

ACTIONS

CATEGORICAL

VALUES CLUSTERING

Highly fuzzy

Compare words

COMPUTE CLUSTERS

MERGE SELECTED

1 clusters found.

☐

Possible clusters

Merge value

☐

- nuevos ministerios rios rosas
- en nuevos ministerios rios rosas

nuevos ministerios rios rosas

3. As per the 4 blank observations in Area we can find the right fields looking by Address.

Formula

```
if(Address='piso en ventilla almenara', 'ventanilla-almenara', if (Address='piso en cuatro caminos', 'cuatro caminos', if (Address= 'Piso en bella vistas', "bellas vistas", if (Address='piso en cuzco castillejos', 'Cuzco-castillejos',Area))))
```

Formula is valid

Preview

Examples

Documentation

Sample output	Address	Area
12 de octubre orcasur	piso en calle del petroleo	12 de octubre orcasur
12 de octubre orcasur	piso en calle de eduardo barreiros	12 de octubre orcasur
a guilas	piso en atguilas	a guilas
a guilas	piso en calle rafael finat	a guilas
abrantes	piso en avenida de abrantes	abrantes
abrantes	piso en abrantes	abrantes
abrantes	piso en calle del chimbo	abrantes
abrantes	piso en callejon juan ramon	abrantes
abrantes	piso en abrantes	abrantes

4. Number Variable for the empty fields we add a "fake Value" that we know it will never happens

Replacements

s/n	→	-99	
Trafalgar	→	-99	

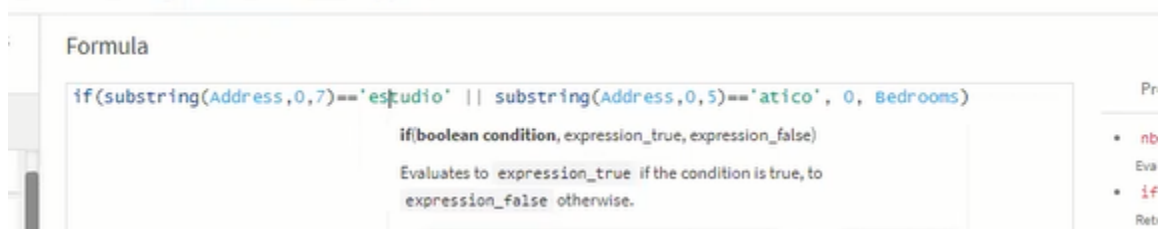
ADD REPLACEMENT

5. Bedroom Variables

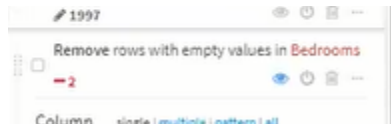
If we analyze the null observations :

Bedroom Analysis		Bedroom										Grand Total
House Type		0	1	2	3	4	5	6	7	8	(blank)	
Ático		4	32	65	37	21	4	1			5	169
Casa o					4	3	9	15	4	1		36
Caserón						1						1
Chalet					4	21	11	11	5			52
Dúplex			15	17	13	8	9				2	64
Estudio		10									82	92
Piso			420	602	371	204	64	8	2	4		1675
(blank)												
Grand Total		14	467	684	429	258	97	35	11	5	89	2089

As per analysis criteria we will add at least 0 to Atic & Estudio



The rest of null observations as per non possible definition removed: A total of 2 blanks .



1. Floor Variable

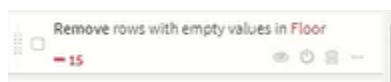
If we analyze the null observations

Floor Analysis	Flo																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																		
----------------	-----	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

As per analysis criteria we will consider that Casa, Caserón & Chalet Caseron as per Cottage equal to 1 and Semidetached equal to 1 the floor should be 0.



The rest of null observations as per non possible definition removed: A total of 15 blanks



2. Outer Variable

Outer Analysis	Out			
House Type	0	1 (blank)	Grand Total	
Ático	7	157	5	169
Casa o	1		35	36
Caserón			1	1
Chalet	3		49	52
Dúplex	6	58		64
Estudio	28	62	2	92
Piso	211	1394	70	1675
(blank)				
Grand Total	256	1671	162	2089

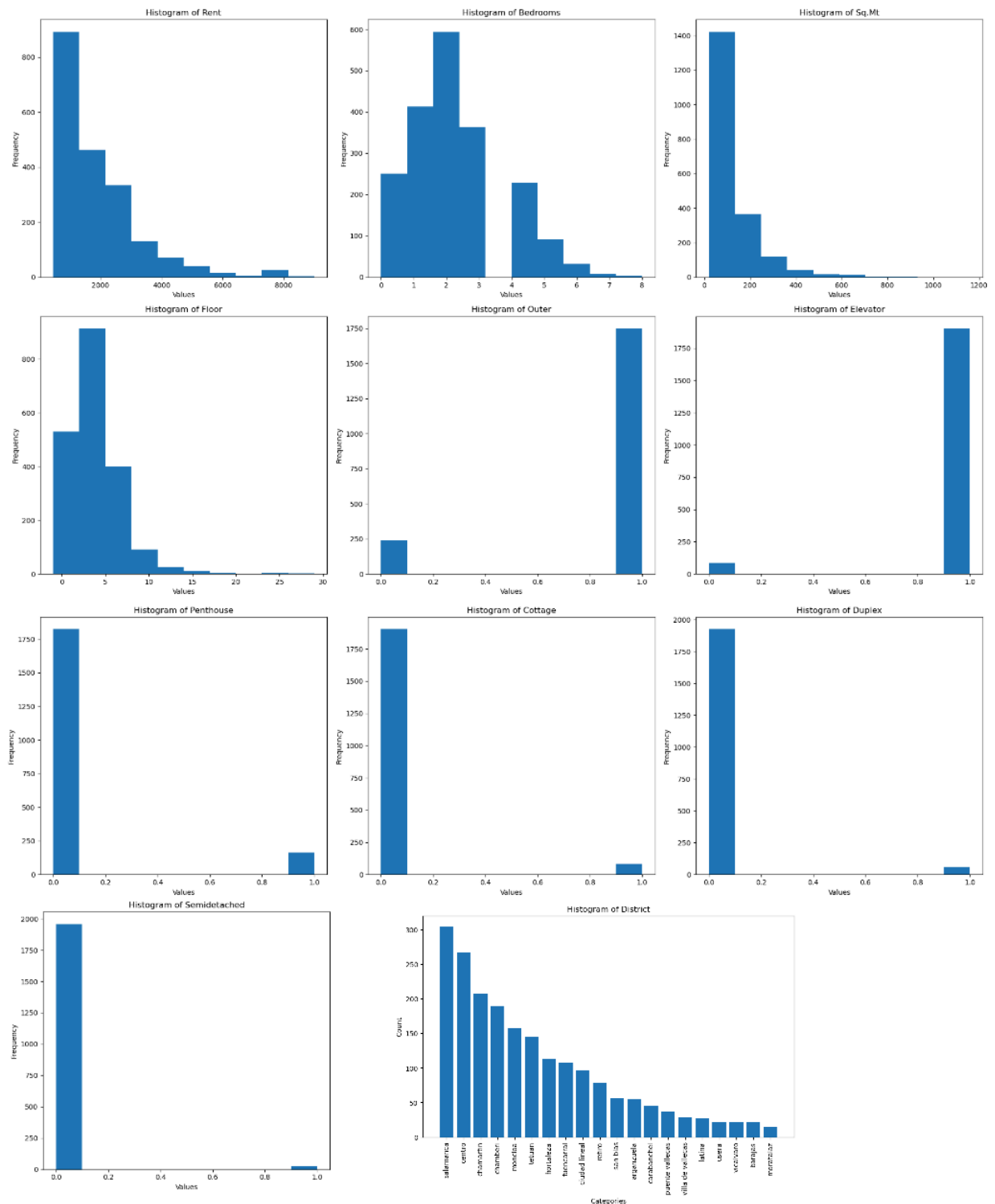
As per analysis criteria we will consider that except for a “Piso” & “Ático” that could be interior, the rest will be treated as Outer.

```
Formula
if(substring(Address,0,4)=='casa' ||
  substring(Address,0,6)=='chalet' ||
  Semidetached==1 ||
  Cottage ==1,
  1, Outer)
```

The rest of null observations as per non possible definition removed: A total of 77 blanks



A2: Variable distributions



"Rent" and "Sq.Mt" both appear to be positively skewed, with their means being larger than their medians. This suggests there might be some high-value outliers in the dataset.

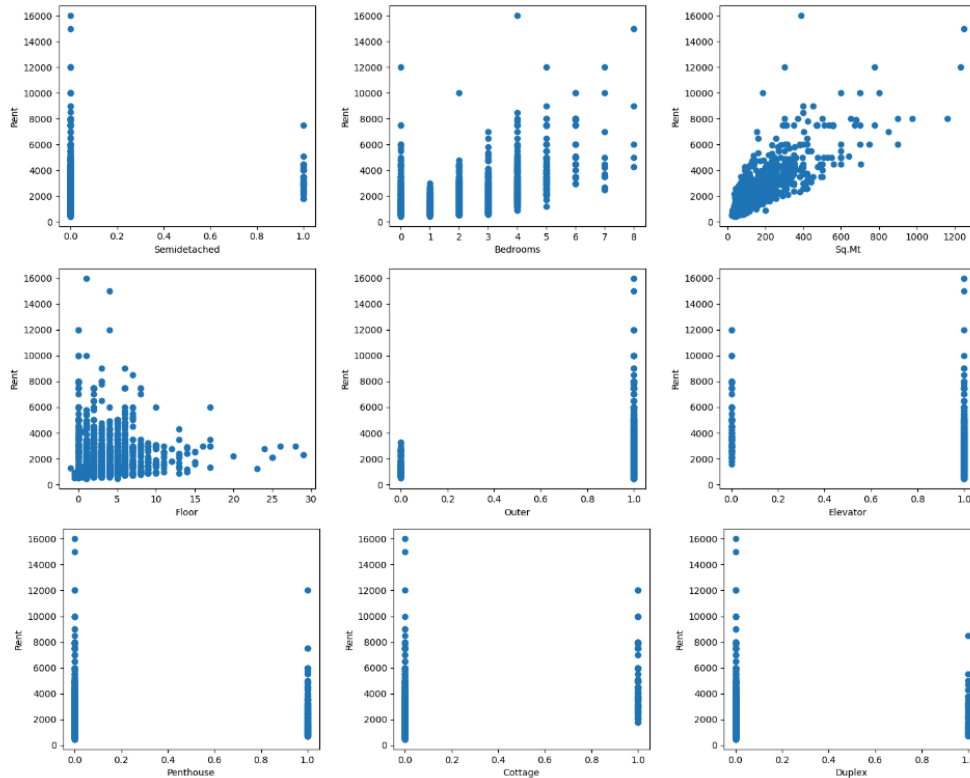
"Floor" seems to be relatively evenly distributed, but given its maximum value, there might be a few properties located in higher-floor buildings.

"Outer" and "Elevator" are mostly 1 (yes), suggesting that a majority of properties are outward-facing and equipped with an elevator.

"Penthouse", "Cottage", "Duplex", and "Semidetached" variables, which are all binary, have mean values significantly less than 0.5. This indicates that these types of properties are less common in the dataset.

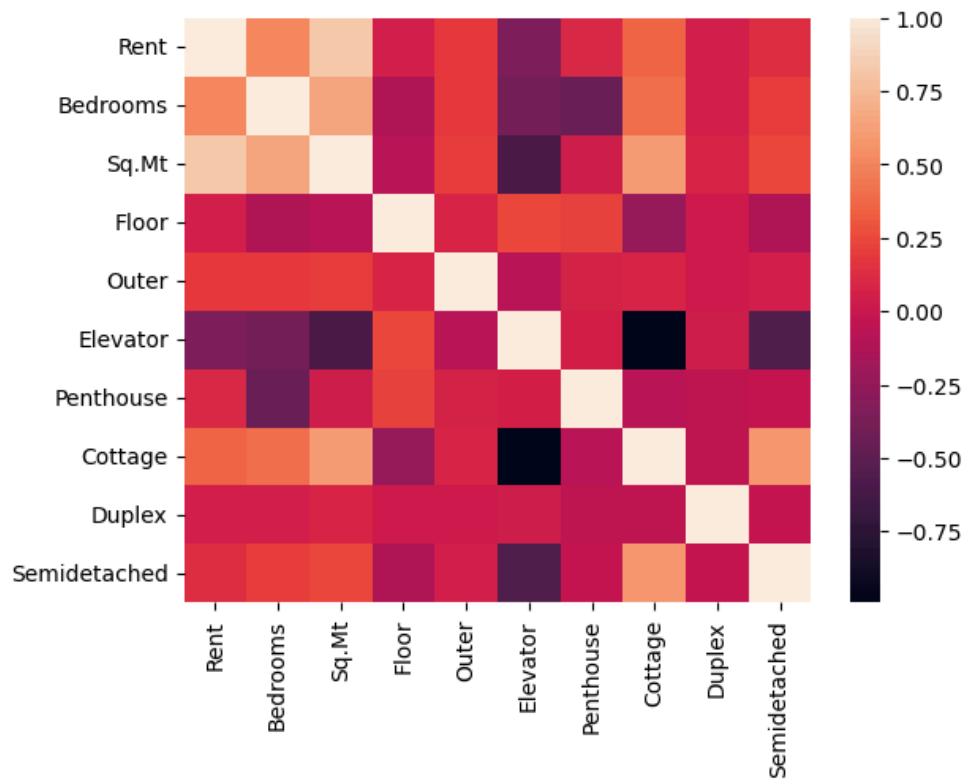
This high-level analysis offers some insight into the nature of the properties in the dataset. It also reveals areas that might warrant further investigation, such as the presence of high-value outliers and the distribution of different property types.

A3: Scatter plots

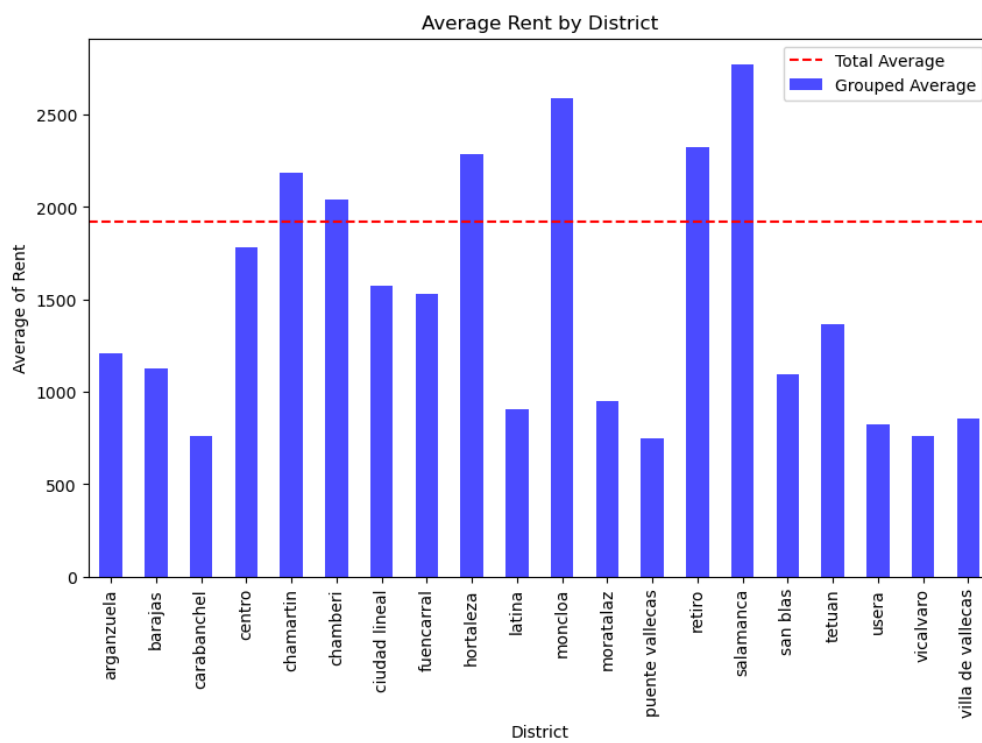


A4: Correlation matrix original variables

	Rent	Bedrooms	Sq.Mt	Floor	Outer	Elevator	Penthouse	Cottage	Duplex	Semidetached
Rent	1.000000	0.517803	0.822577	0.052758	0.185068	-0.360172	0.094679	0.362837	0.049974	0.133474
Bedrooms	0.517803	1.000000	0.647785	-0.136440	0.185439	-0.400245	-0.440413	0.401162	0.057257	0.207484
Sq.Mt	0.822577	0.647785	1.000000	-0.084319	0.206504	-0.599432	0.032500	0.600754	0.073880	0.235394
Floor	0.052758	-0.136440	-0.084319	1.000000	0.079312	0.238032	0.222508	-0.236549	0.021441	-0.135448
Outer	0.185068	0.185439	0.206504	0.079312	1.000000	-0.077545	0.070797	0.077061	0.021664	0.044126
Elevator	-0.360172	-0.400245	-0.599432	0.238032	-0.077545	1.000000	0.062840	-0.993769	0.037725	-0.569034
Penthouse	0.094679	-0.440413	0.032500	0.222508	0.070797	0.062840	1.000000	-0.062448	-0.053678	-0.035758
Cottage	0.362837	0.401162	0.600754	-0.236549	0.077061	-0.993769	-0.062448	1.000000	-0.037490	0.572602
Duplex	0.049974	0.057257	0.073880	0.021441	0.021664	0.037725	-0.053678	-0.037490	1.000000	-0.021467
Semidetached	0.133474	0.207484	0.235394	-0.135448	0.044126	-0.569034	-0.035758	0.572602	-0.021467	1.000000



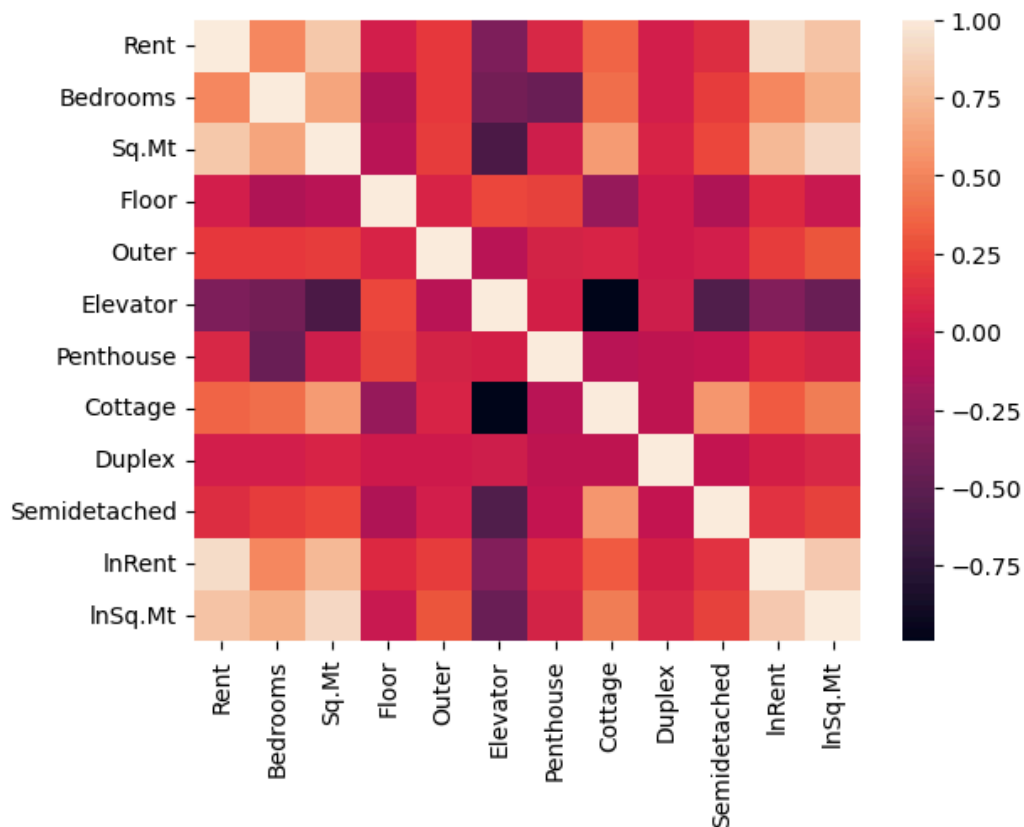
A5: Rent average by District



A6: Correlations with log-transformations

	Rent	Bedrooms	Sq.Mt	Floor	Outer	Elevator	Penthouse	Cottage	Duplex	Semidetached	InRent	InSq.Mt
Rent	1.000000	0.517803	0.822577	0.052758	0.185068	-0.360172	0.094679	0.362837	0.049974	0.133474	0.935880	0.808392
InRent	0.935880	0.515098	0.746098	0.118041	0.204231	-0.318347	0.118728	0.320138	0.061335	0.144646	1.000000	0.829507

	Rent	Bedrooms	Sq.Mt	Floor	Outer	Elevator	Penthouse	Cottage	Duplex	Semidetached	InRent	InSq.Mt
Rent	1.000000	0.517803	0.822577	0.052758	0.185068	-0.360172	0.094679	0.362837	0.049974	0.133474	0.935880	0.808392
Bedrooms	0.517803	1.000000	0.647785	-0.136440	0.185439	-0.400245	-0.440413	0.401162	0.057257	0.207484	0.515098	0.694860
Sq.Mt	0.822577	0.647785	1.000000	-0.084319	0.206504	-0.599432	0.032500	0.600754	0.073880	0.235394	0.746098	0.902279
Floor	0.052758	-0.136440	-0.084319	1.000000	0.079312	0.238032	0.222508	-0.236549	0.021441	-0.135448	0.118041	-0.002988
Outer	0.185068	0.185439	0.206504	0.079312	1.000000	-0.077545	0.070797	0.077061	0.021664	0.044126	0.204231	0.294943
Elevator	-0.360172	-0.400245	-0.599432	0.238032	-0.077545	1.000000	0.062840	-0.993769	0.037725	-0.569034	-0.318347	-0.455666
Penthouse	0.094679	-0.440413	0.032500	0.222508	0.070797	0.062840	1.000000	-0.062448	-0.053678	-0.035758	0.118728	0.071042
Cottage	0.362837	0.401162	0.600754	-0.236549	0.077061	-0.993769	-0.062448	1.000000	-0.037490	0.572602	0.320138	0.455205
Duplex	0.049974	0.057257	0.073880	0.021441	0.021664	0.037725	-0.053678	-0.037490	1.000000	-0.021467	0.061335	0.098578
Semidetached	0.133474	0.207484	0.235394	-0.135448	0.044126	-0.569034	-0.035758	0.572602	-0.021467	1.000000	0.144646	0.221912
InRent	0.935880	0.515098	0.746098	0.118041	0.204231	-0.318347	0.118728	0.320138	0.061335	0.144646	1.000000	0.829507
InSq.Mt	0.808392	0.694860	0.902279	-0.002988	0.294943	-0.455666	0.071042	0.455205	0.098578	0.221912	0.829507	1.000000



A7: Correlation matrix between the expanded dataset, native variables and new ones (without considering the dummy districts).

	Rent	Bedrooms	Sq.Mt	Floor	Outer	Elevator	Penthouse	Cottage	Duplex	Semidetached	Bedrooms / Sq.Mt	Elevator x Penthouse	Bedrooms x Duplex	Outer x Penthouse	Outer x Duplex
Rent	1.000000	0.517803	0.822577	0.052758	0.185068	-0.360172	0.094679	0.362837	0.049974	0.133474	-0.291909	0.094679	0.093904	0.102817	0.055519
Bedrooms	0.517803	1.000000	0.647785	-0.136440	0.185439	-0.400245	-0.440413	0.401162	0.057257	0.207484	0.412538	-0.440413	0.125461	-0.429864	0.063633
Sq.Mt	0.822577	0.647785	1.000000	-0.084319	0.206504	-0.599432	0.032500	0.600754	0.073880	0.235394	-0.261816	0.032500	0.125565	0.037840	0.082707
Floor	0.052758	-0.136440	-0.084319	1.000000	0.079312	0.238032	0.222508	-0.236549	0.021441	-0.135448	-0.112985	0.222508	0.027604	0.220917	0.021555
Outer	0.185068	0.185439	0.206504	0.079312	1.000000	-0.077545	0.070797	0.077061	0.021664	0.044126	-0.073465	0.070797	0.032053	0.107734	0.063429
Elevator	-0.360172	-0.400245	-0.599432	0.238032	-0.077545	1.000000	0.062840	-0.993769	0.037725	-0.569034	0.124909	0.062840	0.033506	0.061358	0.036125
Penthouse	0.094679	-0.440413	0.032500	0.222508	0.070797	0.062840	1.000000	-0.062448	-0.053678	-0.035758	-0.541316	1.000000	-0.047674	0.976419	-0.051401
Cottage	0.362837	0.401162	0.600754	-0.236549	0.077061	-0.993769	-0.062448	1.000000	-0.037490	0.572602	-0.124259	-0.062448	-0.033297	-0.060976	-0.035900
Duplex	0.049974	0.057257	0.073880	0.021441	0.021664	0.037725	-0.053678	-0.037490	1.000000	-0.021467	-0.036995	-0.053678	0.888156	-0.052412	0.957586
Semidetached	0.133474	0.207484	0.235394	-0.135448	0.044126	-0.569034	-0.035758	0.572602	-0.021467	1.000000	-0.051966	-0.035758	-0.019066	-0.034915	-0.020556
Bedrooms / Sq.Mt	-0.291909	0.412538	-0.261816	-0.112985	-0.073465	0.124909	-0.541316	-0.124259	-0.036995	-0.051966	1.000000	-0.541316	-0.023303	-0.528447	-0.041743
Elevator x Penthouse	0.094679	-0.440413	0.032500	0.222508	0.070797	0.062840	1.000000	-0.062448	-0.053678	-0.035758	-0.541316	1.000000	-0.047674	0.976419	-0.051401
Bedrooms x Duplex	0.093904	0.125461	0.125565	0.027604	0.032053	0.033506	-0.047674	-0.033297	0.888156	-0.019066	-0.023303	-0.047674	1.000000	-0.046550	0.875409
Outer x Penthouse	0.102817	-0.429864	0.037840	0.220917	0.107734	0.061358	0.976419	-0.060976	-0.052412	-0.034915	-0.528447	0.976419	-0.046550	1.000000	-0.050189
Outer x Duplex	0.055519	0.063633	0.082707	0.021555	0.063429	0.036125	-0.051401	-0.035900	0.957586	-0.020556	-0.041743	-0.051401	0.875409	-0.050189	1.000000

A8: Parameters first regression model.

	coef	std err	t	P> t	[0.025	0.975]
const	-734.2845	655.689	-1.120	0.263	-2020.412	551.843
Bedrooms	98.3062	26.563	3.701	0.000	46.203	150.410
Sq.Mt	10.0958	0.358	28.219	0.000	9.394	10.798
Floor	28.5860	5.684	5.029	0.000	17.436	39.736
Outer	145.7517	55.277	2.637	0.008	37.327	254.176
Elevator	822.7985	645.897	1.274	0.203	-444.123	2089.720
Penthouse	-119.3310	134.520	-0.887	0.375	-383.191	144.529
Cottage	-98.6895	655.605	-0.151	0.880	-1384.652	1187.273
Duplex	192.9215	313.189	0.616	0.538	-421.396	807.239
Semidetached	310.7710	169.801	1.830	0.067	-22.292	643.834
Bedrooms / Sq.Mt	-6625.7410	2744.373	-2.414	0.016	-1.2e+04	-1242.678
Elevator x Penthouse	-119.3310	134.520	-0.887	0.375	-383.191	144.529
Bedrooms x Duplex	-64.1291	66.749	-0.961	0.337	-195.056	66.798
Outer x Penthouse	503.9137	272.883	1.847	0.065	-31.344	1039.171
Outer x Duplex	-61.6173	313.805	-0.196	0.844	-677.143	553.908
District_barajas	-310.6958	179.107	-1.735	0.083	-662.012	40.620
District_carabanchel	-292.5942	144.014	-2.032	0.042	-575.077	-10.111
District_centro	487.6075	104.783	4.653	0.000	282.077	693.139
District_chamartin	393.7177	108.625	3.625	0.000	180.650	606.786
District_chamberi	402.4715	108.768	3.700	0.000	189.124	615.819
District_ciudad lineal	-130.0692	123.152	-1.056	0.291	-371.630	111.492
District_fuencarral	-264.3161	117.086	-2.257	0.024	-493.980	-34.653
District_hortaleza	-108.3331	119.130	-0.909	0.363	-342.005	125.339
District_latina	-299.3898	164.914	-1.815	0.070	-622.868	24.088
District_moncloa	86.3508	113.720	0.759	0.448	-136.710	309.412
District_moratalaz	-404.4820	217.212	-1.862	0.063	-830.541	21.577
District_puente vallecas	-131.2949	156.206	-0.841	0.401	-437.692	175.102
District_retiro	626.5219	125.777	4.981	0.000	379.811	873.233
District_salamanca	781.0716	104.660	7.463	0.000	575.783	986.361
District_san blas	-237.1152	137.569	-1.724	0.085	-506.955	32.725
District_tetuan	113.6791	114.349	0.994	0.320	-110.615	337.973
District_usera	-222.3511	182.165	-1.221	0.222	-579.665	134.963
District_vicalvaro	-373.5357	182.211	-2.050	0.041	-730.942	-16.130
District_villa de vallecas	-262.7102	164.162	-1.600	0.110	-584.713	59.292

A9: Parameters second regression model.

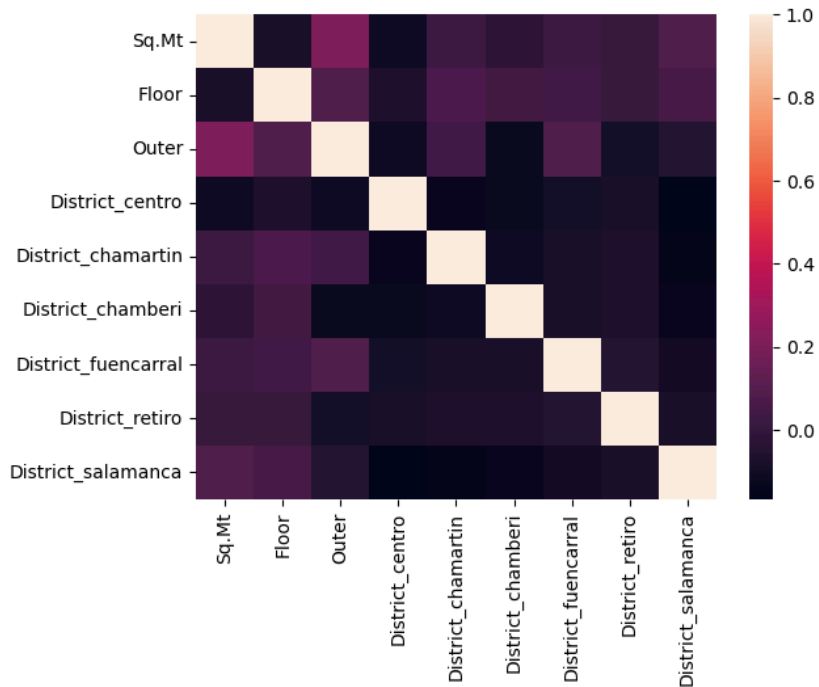
	coef	std err	t	P> t	[0.025	0.975]
const	-14.5697	74.106	-0.197	0.844	-159.926	130.786
Sq.Mt	10.2169	0.167	61.131	0.000	9.889	10.545
Floor	43.5095	5.550	7.839	0.000	32.623	54.396
Outer	212.4190	53.713	3.955	0.000	107.062	317.776
Bedrooms / Sq.Mt	-2314.5548	1636.864	-1.414	0.158	-5525.216	896.106
District_carabanchel	-208.9548	113.357	-1.843	0.065	-431.301	13.391
District_centro	576.5069	53.623	10.751	0.000	471.326	681.688
District_chamartin	496.0556	58.099	8.538	0.000	382.096	610.015
District_chamberi	521.2768	60.643	8.596	0.000	402.327	640.227
District_fuencarral	-164.1144	73.128	-2.244	0.025	-307.552	-20.677
District_retiro	759.4299	87.885	8.641	0.000	587.045	931.814
District_salamanca	941.8562	50.115	18.794	0.000	843.556	1040.156
District_vicalvaro	-301.5280	161.565	-1.866	0.062	-618.433	15.377

A10: Parameters final regression model

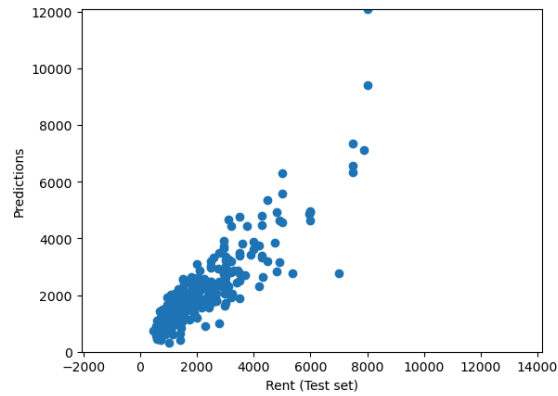
	coef	std err	t	P> t	[0.025	0.975]
const	-100.9489	58.457	-1.727	0.084	-215.611	13.713
Sq.Mt	10.3228	0.161	64.202	0.000	10.007	10.638
Floor	45.1474	5.498	8.211	0.000	34.362	55.932
Outer	211.4365	53.705	3.937	0.000	106.097	316.777
District_centro	608.8422	52.537	11.589	0.000	505.792	711.892
District_chamartin	518.0511	57.724	8.975	0.000	404.827	631.275
District_chamberi	550.0279	59.959	9.173	0.000	432.420	667.636
District_fuencarral	-144.2123	72.878	-1.979	0.048	-287.161	-1.264
District_retiro	781.3717	87.745	8.905	0.000	609.263	953.480
District_salamanca	965.7427	49.581	19.478	0.000	868.492	1062.993

A11: Correlation matrix for independent variables in final model.

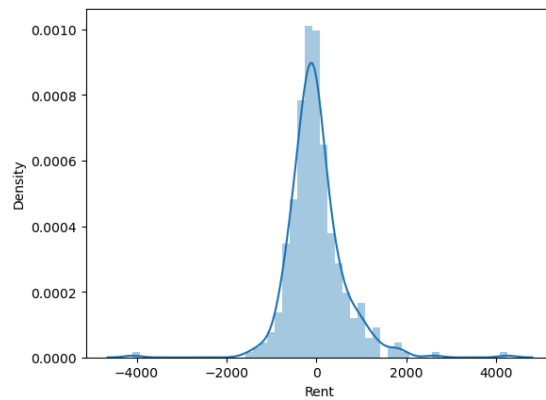
	Sq.Mt	Floor	Outer	District_centro	District_chamartin	District_chamberi	District_fuencarral	District_retiro	District_salamanca
Sq.Mt	1.000000	-0.084319	0.206504	-0.120499	0.018983	-0.020899	0.015606	0.010234	0.084404
Floor	-0.084319	1.000000	0.079312	-0.064646	0.072758	0.036294	0.038797	0.002098	0.050243
Outer	0.206504	0.079312	1.000000	-0.109810	0.039611	-0.124298	0.081651	-0.091488	-0.049816
District_centro	-0.120499	-0.064646	-0.109810	1.000000	-0.134145	-0.127163	-0.094306	-0.080041	-0.167186
District_chamartin	0.018983	0.072758	0.039611	-0.134145	1.000000	-0.110301	-0.081802	-0.069428	-0.145018
District_chamberi	-0.020899	0.036294	-0.124298	-0.127163	-0.110301	1.000000	-0.077544	-0.065815	-0.137470
District_fuencarral	0.015606	0.038797	0.081651	-0.094306	-0.081802	-0.077544	1.000000	-0.048809	-0.101950
District_retiro	0.010234	0.002098	-0.091488	-0.080041	-0.069428	-0.065815	-0.048809	1.000000	-0.086529
District_salamanca	0.084404	0.050243	-0.049816	-0.167186	-0.145018	-0.137470	-0.101950	-0.086529	1.000000



A12: Scatter plot linear relationship between the test values and the predictions



A13: Histogram of errors



A14: Good opportunities

Descriptive matrix for independent variables for those houses which have bigger differences between theoretical prices and real rent prices. This dataset corresponds to the 25% with higher differences (bigger than \$328).

	Sq.Mt	Floor	Outer	District_centro	District_chamartin	District_chamberi	District_fuencarral	District_retiro	District_salamanca
count	498.000000	498.000000	498.000000	498.000000	498.000000	498.000000	498.000000	498.000000	498.000000
mean	157.791165	3.928715	0.901606	0.126506	0.150602	0.098394	0.020080	0.084337	0.251004
std	151.008856	3.171877	0.298146	0.332753	0.358021	0.298146	0.140416	0.278173	0.434027
min	20.000000	-0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	70.500000	2.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	105.000000	3.500000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	170.750000	6.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.750000
max	1160.000000	25.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000