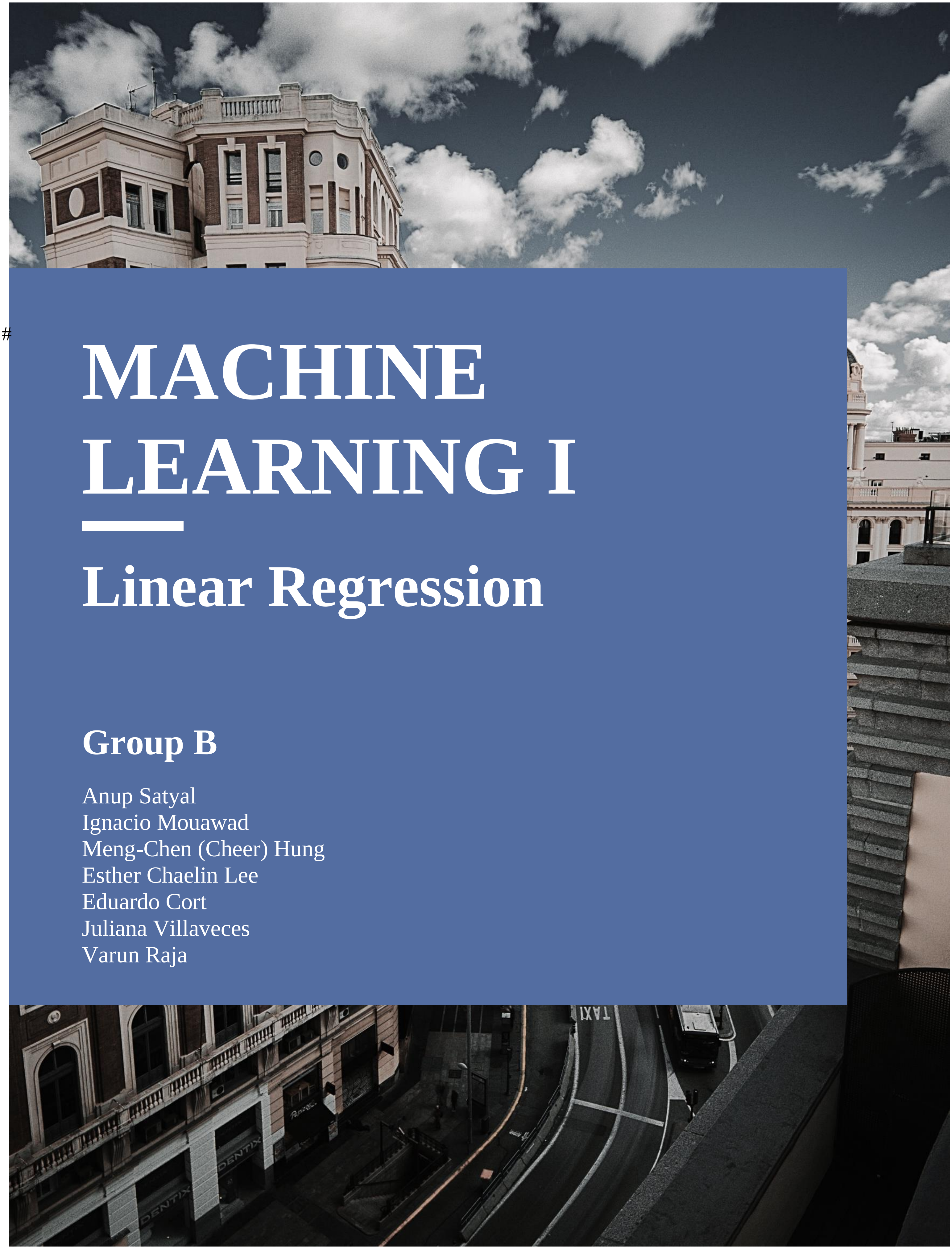

MACHINE LEARNING I

Linear Regression

Group B

Anup Satyal
Ignacio Mouawad
Meng-Chen (Cheer) Hung
Esther Chaelin Lee
Eduardo Cort
Juliana Villaveces
Varun Raja

Executive Summary

The dataset provided to us contained information on 2000 rental properties for all 21 districts in Madrid from Idealista. Additional variables included the area (*barrio*), rental price, number of bedrooms, square meters, floor, outer, elevator, penthouse, duplex, semi-detached and cottage.

A clustering model was first carried out to classify the different Areas in the dataset based on similarities such as average rent per square meters, average number of beds, proportion of properties with elevators etc (the metrics used and areas in each cluster are provided in the technical annex). A step-wise method was also carried out to identify which of the numeric variables were statistically significant in explaining the rental prices. Finally, a correlation analysis was carried out to remove information redundancy. Using the clusters and statistically significant non-correlated variables, a linear regression model was used to estimate the rental prices.

The resulting models can potentially be used by the agency to estimate theoretical rental prices for properties around Madrid. The agency, or its clients, may also find apartments that are underpriced relative to a theoretical price estimated by the model, and exploit those opportunities.

The results of the analysis suggest that such a model would can only be used to predict the rental prices in 4 out of the 5 clusters. Without additional data, we are unable to confirm its applicability in Areas that form a part of the 5th cluster (Cluster_5).

Results

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Cluster_0	- 307.4	81.7	- 3.76330	< 1e-4	★★★
Cluster_1	- 173.0	92.4	- 1.87170	0.03070	★☆☆
Cluster_2	- 176.7	82.3	- 2.14700	0.01600	★☆☆
Cluster_3	494.6	67.3	7.35260	< 1e-4	★★★
Penthouse = 0	- 204.0	75.2	- 2.71360	0.00340	★★☆
Outer = 1	370.3	54.7	6.77310	< 1e-4	★★★
Floor	17.0	7.6	2.24350	0.01250	★☆☆
Square_Meters	10.8	0.2	62.80230	< 1e-4	★★★
Intercept	113.3	113.3	1.00000	0.15880	☆☆☆
R2	0.82				

The results indicate that a mathematical function that could be used to estimate rental prices is as follows:

Rental Price = 113.3 + Coefficient (Cluster_x) – 204 * (Penthouse = 0) + 370 * (Outer = 1) + * 17.0 * (Number of Floors) + 10.8 * (Square Meters)

The theoretical interpretation for the intercept (113.3) is that for a interior property that is not a penthouse, with zero square meters, zero floors, and in Cluster_4, the rental price would be €113.3. Naturally there is no economic interpretation for this number. In addition, it should be noted that the intercept is not *statistically significant* at a 95% confidence level and therefore cannot be used to estimate prices. In the table above, this is indicated by the stars on the right-hand side. Essentially, if a variable is not statistically significant (i.e the stars are not filled in), which is the case for the intercept for example, no inference (conclusions) can be extracted from it. This is because if they are used, the actual price is likely to be more than ±5% of estimated values, which is probably higher than the risk that the agency should take. Since in practical terms, a property cannot have zero values for any of the variables above but can be located in the Areas that form a part of Cluster_4, we can assume that Cluster_4 represents the intercept and since the intercept is not significant, Cluster_4 can be said to be insignificant as well.

The modelling results show that all else being equal (c*etris paribus),* apart from first 4 clusters, the significance of which is detailed below, the number of floors, the size of the property (in square meters) and a street facing property has a positive influence on rental prices. In other words, holding everything else constant, increasing the number of floors where the apartment is located in by 1 increases the rental price by, on average, €17.0. Similarly, *cetris paribus*, increasing the size of the property by 1 square meter increases the rental price by, on average, €10.8. With regards to the street facing property, the conclusion is that if it is outer facing, the rent would increase by, on average, €370.3. By contrast, if the property were not outer facing, that is to say inner facing, the rent would decrease by an equivalent amount. A similar conclusion can be drawn if the property is a penthouse. If the property is a penthouse, holding everything else constant, the rent can be expected to increase by €204. Lastly, coefficients of the cluster indicate that an apartment located in the first three clusters will have a rent that is, on average, €307, €173 and €176.7 lower respectively, all else held equal. For the 4^{th} Cluster (Cluster_3), the rental price should be, on average, €494.6 higher.

Based on the results provided by the model, the theoretical price for a 5^{th} floor, 50 square meter penthouse property facing the street in Jeronimos (located in Cluster_3) can be estimated as follows:

Rental Price = $496.4 + 204 + 370 + 17.0 * 5 + 10.8 * 50 = 1697.9$

The final metric that is important to highlight is the R^2 score. It shows how much of the variance of Rental Prices can be explained using the variables in the model, which is its “power”.

The R^2 indicates that knowing the size of a property on a certain floor, whether it is an outer facing penthouse in one of the statistically significant clusters, the model can explain 82% of the variability in rental prices. That is to say that other factors only affect 18% of the variability in rental prices.

Data Preparation and Analysis

Prior to the modelling phase, some data preparation processes were carried out on the provided dataset. The steps followed were:

- To ensure consistency, all text variables were converted to lower case. To improve functionality during the modelling process, some special characters from the Spanish language were also converted into common English.
- Assuming that floor refers to the level where rental properties are located, elevator refers to whether the building has an elevator and outer refers to whether the property is street facing, the missing values for these fields for cottages were filled with a zero.
- The elevator column contained some inconsistent values (-0.5 and 0.5) which were substituted by 0 and 1 respectively.
- Values in the district and areas columns were incorrectly inputted. For example, even though Pau de Carabanchel is not an area (barrio) in Madrid, it was inputted as a value. To guarantee exactness, a dataset from the Madrid Mayoral Office was used to correct all the incorrect data in these columns.
- All the rows that still had missing values were removed from the dataset. In total, a 190 (<10% of the provided dataset) were deleted.
- Some outliers were observed in the Rental Prices column. To verify if they were indeed outliers, a Rent per Square Meter feature was created and all values outside 1.5x IQR of this new variable were considered to be outliers and thus removed.
- Lastly, a correlation analysis was carried out after which Cottage and Bedrooms were removed as variables due to their high correlation with other variables.

Once the data-preparation phase was completed, an explanatory analysis indicated that the >30% dataset was formed by only 3 districts (Salamanca – 13.8%, Centro – 12.1% and Chamartin 10.8%), although the Areas were largely well distributed (max 4% per area). To avoid the introduction of biases and to allow for more precise and usable estimates of rental prices, it was decided that areas rather than districts should be used as a model input.

After the features were selected, the properties in the provided dataset were aggregated based on several characterizes, the definitions for which are provided in the technical annex. The table below shows a sample of the inputs that were used for the clustering algorithm.

Area	avg_bedrooms	avg_rent_sqm	avg_floor	propor_outer	propor_elevator	propor_penthouse	propor_cottage	propr_duplex	propor_semi-detached
abrantes	3	9	1	0%	0%	0%	0%	0%	0%
acacias	2	16	4	1%	1%	2%	0%	0%	0%
adelfas	3	13	4	0%	0%	0%	0%	0%	0%
almagro	3	17	3	3%	3%	6%	0%	0%	0%
almenara	2	14	5	1%	1%	0%	1%	0%	3%

The results of the clustering algorithm were added to the statistically significant variables that resulted from a step-wise carried out on the numeric variables (table below). These provided the final inputs to be used for the linear modelling.

Variable	Pr(>|t|)	Significance
Square_Meters	-	★★★
Elevator	0.00000	★★★
Floor	0.00000	★★★
Outer	0.00421	★★★
Penthouse	0.03485	★☆☆

Finally, to avoid overfitting problems, multiple iterations of the linear model with a 70-30% train/test split of the dataset and 10 k-fold cross validations were carried out to derive the results presented above.

Conclusions

The results of our analysis indicate that the model can be used to predict the rental prices for properties located in 4 out of 5 clusters that were generated from the clustering algorithm. Such a model can explain 82% of variability in rental prices with a 5% risk that the actual price is outside the predicted figure. The only constraint with the model is that the scope of the predictions must be limited to the statistically significant clusters. The models simply cannot be used to predict prices for any other areas or using any other variables not a part of the model equations mentioned above.

Recommendations

While the scope of both models is limited to the statistically significant clusters and variables, it is not to say that devising similar models for a different dataset that yields entirely different results is not possible. In fact, 62.5% of the properties in the provided dataset are in only 6 (out of 21) districts. There is clearly a bias in the dataset and a different dataset with more properties in the other areas may yield a model that contains the areas in Cluster_4. Such an analysis is recommended such that the agency may predict the rental prices of properties in the areas not included in the model from this analysis.

In addition, while an R^2 of 82% can be considered acceptable, there might still be other statistically significant variables that are able to explain the remaining 18% of the variability in rental prices. Therefore, it is also recommended that the agency build a dataset with additional variables that may help to explain the remaining variability. Such an analysis will most likely result in a better model than the one that is possible with the provided variables.

Appendix

EXHIBIT 1 – CLUSTERING INPUT

Area	avg_bedrooms	avg_rent_sqm	avg_floor	propor_outer	propor_elevator	propor_penthouse	propor_cottage	propr_duplex	propor_semi-detached
abrantes	3	9	1	0%	0%	0%	0%	0%	0%
acacias	2	16	4	1%	1%	2%	0%	0%	0%
adelfas	3	13	4	0%	0%	0%	0%	0%	0%
almagro	3	17	3	3%	3%	6%	0%	0%	0%
almenara	2	14	5	1%	1%	0%	1%	0%	3%
almendrales	2	12	4	0%	0%	0%	0%	0%	0%
aluche	3	10	4	1%	1%	0%	0%	0%	0%
apostol santiago	3	12	5	1%	0%	1%	0%	0%	0%
arapiles	2	17	4	1%	2%	2%	0%	4%	0%
aravaca	4	12	1	1%	1%	2%	11%	0%	9%
arcos	3	9	3	0%	0%	0%	0%	0%	0%
argüelles	2	18	6	2%	3%	2%	0%	2%	0%
atalaya	4	14	2	0%	0%	0%	0%	0%	0%
bellas vistas	1	15	2	1%	1%	0%	0%	2%	0%
berruguete	2	14	2	1%	1%	1%	0%	0%	0%
buenavista	2	10	4	0%	0%	2%	0%	2%	0%
butarque	1	11	3	0%	0%	0%	0%	0%	0%
campamento	3	12	3	0%	0%	0%	0%	2%	0%
canillas	3	11	1	0%	0%	1%	5%	0%	3%
canillejas	3	9	2	0%	0%	0%	1%	0%	3%
casco historico de vallecas	2	11	2	0%	0%	0%	0%	0%	0%
casco historico de vicalvaro	2	12	2	0%	0%	0%	0%	0%	0%
castellana	3	18	4	4%	4%	8%	0%	2%	0%
castilla	2	14	6	2%	2%	1%	0%	0%	0%
castillejos	2	16	6	2%	2%	0%	0%	0%	0%

Area	avg_bedrooms	avg_rent_sqm	avg_floor	propor_outer	propor_elevator	propor_penthouse	propor_cottage	propr_duplex	propor_semi-detached
chopera	3	15	5	0%	0%	1%	0%	0%	0%
ciudad jardin	2	17	2	1%	1%	0%	1%	0%	3%
ciudad universitaria	4	12	2	1%	1%	0%	8%	0%	0%
colina	3	13	2	1%	0%	1%	0%	2%	0%
comillas	2	10	4	0%	0%	0%	0%	0%	0%
concepcion	3	12	2	1%	1%	2%	1%	2%	0%
cortes	2	18	3	1%	1%	1%	0%	2%	0%
costillares	3	13	4	1%	1%	1%	1%	2%	3%
cuatro caminos	2	16	4	3%	3%	1%	0%	4%	0%
delicias	2	13	3	1%	1%	0%	0%	0%	0%
el goloso	3	14	3	1%	1%	3%	0%	2%	0%
el plantio	5	11	0	0%	0%	0%	5%	0%	0%
el viso	3	18	4	2%	2%	5%	4%	4%	12%
embajadores	2	18	2	2%	2%	1%	0%	2%	0%
ensanche de vallecas	2	11	4	1%	1%	2%	0%	0%	0%
entrevias	3	9	3	0%	0%	0%	0%	0%	0%
estrella	3	15	6	0%	0%	0%	0%	0%	0%
fontarron	3	9	4	0%	0%	0%	0%	0%	0%
fuente del berro	3	15	3	0%	1%	0%	0%	0%	0%
fuentelarreina	4	10	2	0%	0%	0%	1%	0%	0%
gaztambide	2	18	5	0%	1%	0%	0%	0%	0%
goya	3	18	4	3%	4%	4%	0%	2%	0%
guindalera	3	16	3	1%	1%	0%	1%	2%	3%
hellin	3	12	3	0%	0%	0%	0%	0%	0%

Area	avg_bedrooms	avg_rent_sqm	avg_floor	propor_outer	propor_elevator	propor_penthouse	propor_cottage	propr_duplex	propor_semi-detached
hispanoamerica	2	17	4	3%	3%	1%	0%	4%	0%
horcajo	4	11	4	0%	0%	0%	0%	0%	0%
ibiza	3	16	3	1%	2%	1%	0%	0%	0%
imperial	2	14	3	0%	0%	0%	0%	0%	0%
jeronimos	3	20	5	1%	1%	2%	0%	0%	0%
justicia	2	19	2	3%	4%	5%	0%	2%	0%
la paz	3	12	7	1%	1%	0%	1%	0%	0%
las aguilas	3	10	6	0%	0%	0%	0%	0%	0%
legazpi	2	14	2	0%	0%	1%	0%	0%	0%
lista	2	17	3	2%	2%	2%	0%	2%	0%
los angeles	3	8	6	0%	0%	0%	0%	0%	0%
los carmenes	2	13	2	0%	0%	0%	0%	0%	0%
los rosales	3	13	3	0%	0%	0%	0%	0%	0%
lucero	1	13	2	0%	0%	1%	0%	0%	0%
marroquina	3	11	4	0%	0%	0%	0%	0%	0%
media legua	4	10	4	0%	0%	0%	0%	0%	0%
mirasierra	4	11	3	1%	1%	0%	2%	0%	0%
moscardo	2	12	2	0%	0%	0%	0%	0%	0%
niño jesus	3	17	5	0%	1%	0%	1%	0%	0%
nueva españa	3	16	4	2%	3%	1%	4%	2%	3%
numancia	2	11	2	0%	0%	0%	0%	0%	0%
opañel	2	14	3	0%	0%	0%	0%	0%	0%
orcasur	2	12	2	0%	0%	0%	0%	0%	0%
pacifico	2	16	3	1%	1%	0%	0%	0%	0%
palacio	2	19	3	2%	2%	2%	0%	0%	0%

Area	avg_bedrooms	avg_rent_sqm	avg_floor	propor_outer	propor_elevator	propor_penthouse	propor_cottage	propr_duplex	propor_semi-detached
palomas	4	12	1	0%	0%	0%	6%	0%	15%
palomeras bajas	2	13	2	0%	0%	0%	0%	0%	0%
palomeras sureste	2	11	4	0%	0%	0%	0%	0%	0%
palos de moguer	1	16	2	0%	0%	1%	0%	2%	0%
pavones	3	14	3	0%	0%	0%	0%	0%	0%
peñagrande	3	12	3	1%	1%	0%	2%	0%	3%
pilar	3	12	7	1%	1%	0%	0%	0%	0%
pinar del rey	4	11	2	0%	0%	0%	3%	0%	6%
piovera	4	12	1	1%	1%	1%	22%	4%	24%
portazgo	3	10	2	0%	0%	0%	0%	0%	0%
pradolongo	2	12	2	0%	0%	1%	0%	0%	0%
prosperidad	2	14	2	1%	1%	2%	0%	2%	0%
pueblo nuevo	3	11	3	1%	0%	1%	0%	2%	0%
puerta bonita	2	11	2	0%	0%	1%	0%	0%	0%
puerta del angel	2	14	3	0%	0%	2%	0%	0%	0%
quintana	2	14	3	0%	0%	0%	0%	2%	0%
recoletos	3	19	4	4%	5%	3%	0%	7%	0%
rejas	2	12	2	2%	2%	0%	0%	2%	0%
rios rosas	2	16	4	1%	2%	3%	0%	0%	0%
rosas	3	12	5	0%	0%	1%	0%	0%	0%
salvador	2	12	2	0%	0%	2%	1%	0%	0%
san diego	2	13	1	1%	0%	0%	0%	0%	0%
san fermin	2	10	2	0%	0%	0%	0%	0%	0%
san isidro	2	12	2	1%	0%	0%	0%	0%	0%
san juan bautista	3	14	5	1%	1%	2%	1%	11%	3%

Area	avg_bedrooms	avg_rent_sqm	avg_floor	propor_outer	propor_elevator	propor_penthouse	propor_cottage	propr_duplex	propor_semi-detached
san pascual	3	15	3	1%	1%	0%	0%	0%	0%
simancas	2	11	2	1%	1%	0%	0%	0%	0%
sol	2	18	3	1%	2%	2%	0%	2%	0%
trafalgar	2	17	4	1%	1%	2%	0%	0%	0%
universidad	2	20	3	2%	2%	3%	0%	0%	0%
valdeacederas	2	13	2	1%	1%	0%	0%	7%	0%
valdebernardo	2	11	4	0%	0%	0%	0%	0%	0%
valdefuentes	3	13	3	3%	3%	5%	5%	7%	0%
valdemarin	4	13	1	1%	1%	4%	6%	6%	0%
valdezarza	3	13	2	1%	1%	0%	2%	0%	3%
vallehermoso	3	17	4	1%	1%	0%	0%	0%	0%
valverde	2	13	3	3%	3%	5%	1%	0%	3%
ventas	2	13	3	0%	0%	0%	0%	2%	0%
vinateros	3	10	4	0%	0%	0%	0%	0%	0%
vista alegre	3	10	3	0%	0%	0%	0%	0%	0%
zofio	2	11	4	0%	0%	0%	0%	0%	0%

EXHIBIT 2 – CLUSTER AREAS

Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4
costillares	atalaya	quintana	san pascual	san juan bautista
concepcion	mirasierra	colina	pacifico	la paz
pueblo nuevo	fuentelarreina	ventas	niño jesus	pilar
valverde	pinar del rey	puerta del angel	jeronimos	apostol santiago
el goloso	canillas	lucero	ibiza	las aguilas
peñagrande	piovera	los carmenes	goya	arguelles
valdefuentes	palomas	palomeras bajas	castellana	estrella
aluche	aravaca	san diego	guindalera	chopera
campamento	ciudad universitaria	numancia	recoletos	castilla
vinateros	valdemarin	legazpi	lista	rosas
media legua	valdezarza	palos de moguer	fuente del berro	castillejos
marroquina	el plantio	imperial	acacias	almenara
pavones	horcajo	delicias	justicia	los angeles
fontarron	abrantes	san isidro	embajadores	
palomeras sureste		puerta bonita	universidad	
entrevias		prosperidad	sol	
portazgo		salvador	palacio	
adelfas		simancas	cortes	
buenavista		rejas	opañel	
vista alegre		bellas vistas	nueva españa	
comillas		berruguete	hispanoamerica	
arcos		valdeacederas	ciudad jardin	
canillejas		pradolongo	el viso	
hellin		moscardo	almagro	
zofio		orcasur	arapiles	
valdebernardo		almendrales	gaztambide	
ensanche de vallecas		san fermin	rios rosas	
los rosales		casco historico de vicalvaro	vallehermoso	
		casco historico de vallecas	trafalgar	
		butarque	cuatro caminos	

EXHIBIT 3 – DISTRUBITION OF ERRORS IN DATAIKU

Minimum	25th perc.	Median	75th perc.	90th perc.	Maximum
-963.46	-278.16	-41.252	218.33	644.39	1562.3
	Average	25.213		Standard deviation	518.87

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).

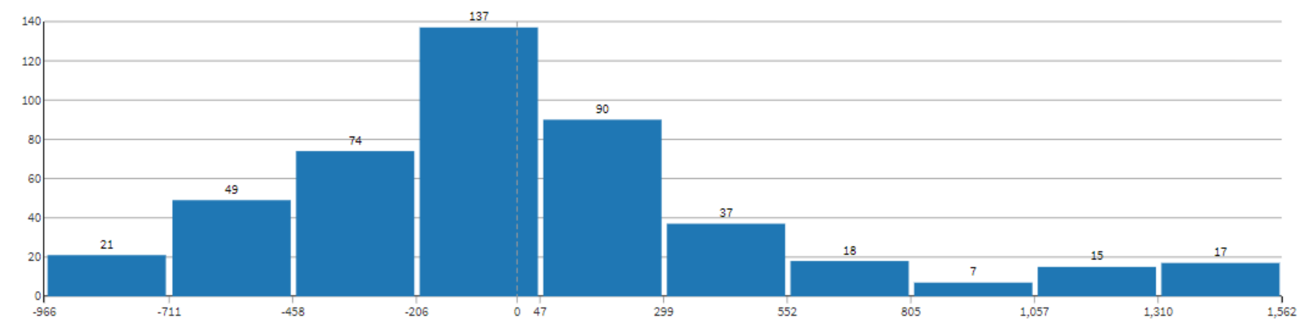

EXHIBIT 5 – VARIABLE CORRELATIONS

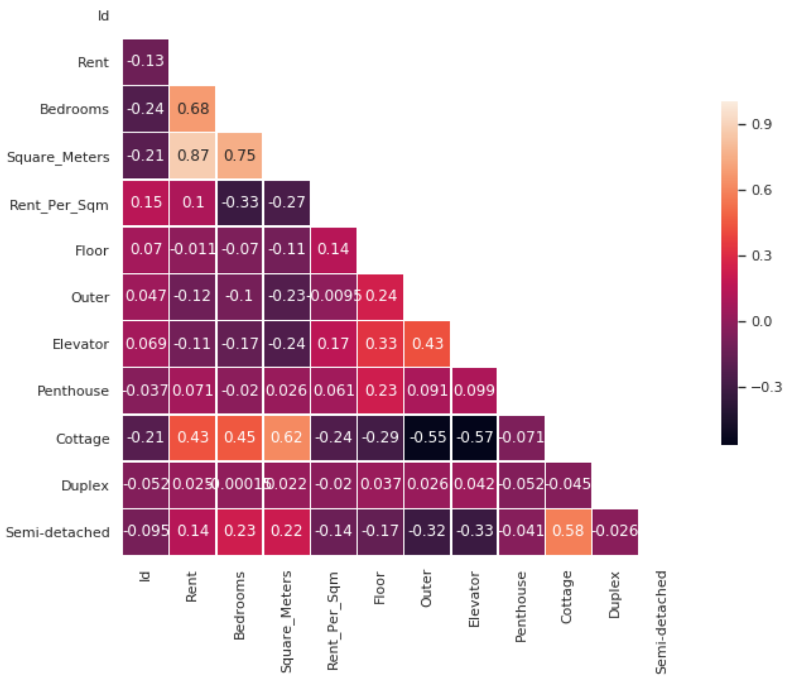

EXHIBIT 6 – MODEL RESULTS

Explained Variance Score Best possible score is 1.0, lower values are worse	0.82145
Mean Absolute Error (MAE) Average of the absolute value of the regression error	413
Mean Average Percentage Error Average of the absolute value of the regression error	21.2%
Mean Squared Error (MSE) Average of the squares of the errors	4.6456e+5
Root Mean Squared Error (RMSE) Root of the above mesure	682
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	0.90692
R2 Score (Coefficient of determination) regression score function	0.82075