Group 5

Analysis of Rental Market for Madrid Housing

Segmentation and Predictive Modeling for Madrid's Housing rental market

Proposed by:

Alejandro Gutierrez Werner

Apilash Balasingham

Fhadah Alromy

Lucia Pellicer Cascales

Luis Soto Guareschi

Ignacio Sahonero Vadillo

Professor:

Alvaro Jose Mendez Lopez

MBDS: Master in Business Analytics and Data Science

Machine Learning I

Madrid, Spain

www.ie.edu

October, 2025

Executive Summary

The study analyzes Madrid's rental housing market using segmentation and predictive modeling based on 2,089 listings from idealista.com.

K-means clustering identified five clear market segments (Silhouette Score = 0.72), distinguishing high-end central areas like Salamanca and Chamberí from more affordable, spacious outer districts—useful for targeted pricing and marketing.

A linear regression model explained about 70% of rent price variation ($R^2 = 0.712$), showing that **location and size** are the main determinants of rental prices. Properties in Salamanca and Retiro command higher rents, while larger size and higher floors also increase price.

Overall, the findings reveal a structured rental market and provide practical tools for agencies to improve pricing accuracy, identify mispriced properties, and tailor offerings to specific market segments.

Data Audit, Cleaning and Manipulation

The dataset contains 2,089 rental advertisements including district, address, number, area, rent price, number of bedrooms, size in square meters, floor level and some specification flags like outer, elevator, penthouse, cottage, duplex and semi detached. It has a mix of different numerical, categorical and string variables. During the data audit, we identified the presence of unique identifiers, such as Address and Number, which are not helpful to create a general model and therefore we removed these columns from the dataset. Similarly, the variable Area was also removed to avoid the creation of sparse and noisy data from excessive one-hot encoding.

The exploratory analysis also uncovered missing values in a few numerical fields. The columns Bedrooms, Floor, Outer, and Elevator were lacking some values. To ensure consistency, we imputed the bedroom variable with the mean. We filled the missing floor variables with zeros, as a further investigation of these ads described a housing on the ground floor. For binary flags like the outer and elevator variables, we imputed them with 0.5, as this is the mean of the possible values. Furthermore, outliers were observed affecting the segmentation, therefore, we decided to detect and treat those anomalies. Outlier analysis was performed using the Isolation Forest algorithm (to identify rare or extreme observations without relying on distance or distributional assumptions).

Finally, to prepare the data for the segmentation analysis and the predictive model, we standardized the numerical features using the StandardScaler. This ensures that all features contribute equally to the distance calculations in the segmentation analysis and prevents variables with larger ranges from destroying the clustering solution.

After ensuring the dataset was clean and free of missing or anomalous entries, the next stage focused on evaluating the variables for analytical relevance. We examined each numeric

variable's variance and pairwise correlations to identify redundant features for avoiding multicollinearity. We also binned the District variable, districts were grouped into three macrozones according to their geographical location; center, north and south.

Segmentation Analysis

Our segmentation followed an iterative, data-driven process aimed at balancing statistical robustness and business interpretability. We began by defining a relevant feature set, applied the K-Means algorithm, and evaluated each configuration both quantitatively and conceptually. When results lacked clarity or separation, we refined the feature selection and repeated the process. After several iterations, the analysis converged on a stable and meaningful five-cluster solution.

The final model was built on features capturing the essential dimensions of Madrid's rental market—Rent, Bedrooms, Sq.Mt, Floor, Outer, Elevator, Penthouse, Cottage, Duplex, Semidetached, and Zone. This combination allowed us to account for economic, structural, and locational characteristics, leading to clusters that are understandable from a business perspective.

To determine the optimal number of clusters, we applied an iterative combination of the Elbow Method and Silhouette analysis. Starting with a broad range of potential cluster counts, we examined how the within-cluster sum of squares (WCSS) declined as K increased. The curve showed a clear inflection around five clusters, where additional splits provided diminishing returns in explained variance. This technical evidence was supported by the Silhouette scores, which indicated consistent internal cohesion and separation at K = 5. Those five clusters are the following:

Segmentation Profiles

Segment 1 – Upper Mainstream Market (51.9% | 2,191 €)

The largest and most representative cluster, composed of spacious exterior apartments with elevators, located in high-demand districts such as Salamanca, Centro, and Chamartín. These properties offer comfort and accessibility at upper-middle market prices, providing strong rental stability and solid investment potential.

Segment 2 – Luxury Segment (7.9% | 2,467 €)

A small but premium cluster concentrated in Salamanca and Centro, defined by penthouses and cottages with outstanding views and amenities. Positioned on higher floors and outer façades, these listings command the city's highest rents and represent Madrid's luxury rental benchmark.

Segment 3 – Family Mid-Market (15.8% | 1,353 €)

Mid-range apartments mainly found in Fuencarral, Ciudad Lineal, and Hortaleza. These homes are well-connected and spacious but not luxurious, appealing to families and professionals seeking practical, comfortable living at moderate prices.

Segment 4 – Peripheral Affordable Housing (10.8% | 780 €)
Compact, budget apartments concentrated in districts like Puente Vallecas, Latina, and Vicálvaro. With smaller layouts and limited amenities, they form Madrid's most accessible rental segment**.**

Segment 5 – Mid-Market Accessible (13.5% | 1,877 €)
Centrally located apartments and cottages in areas such as Centro, Salamanca, and Chamberí. This group bridges comfort and affordability, offering mid-to-high prices that remain below the luxury threshold. It attracts professionals and small families seeking good quality of life without entering the premium range.

Linear Regression Model

The linear regression model aims to explain rental price behavior using explanatory variables derived from the segmentation analysis. It was developed iteratively with a train/test split approach, involving data preparation, feature extraction, model building, and performance evaluation.

Data integrity and feature selection followed principles from the segmentation phase, emphasizing business relevance while applying necessary modeling adjustments. Since the target variable ("Rent") showed a right-skewed distribution with outliers, a square root transformation and outlier handling were applied before feature engineering to improve normalization.

Feature selection combines stepwise selection and correlation analysis to ensure predictor effectiveness and prevent multicollinearity. The final model included **eight predictors** evaluated across **811 observations**.

Model Performance

General model fitness:

- **R-Squared** = 0.712, the model can explain 71% of the target behaviour through the selected predicting variable features.
- **Adjusted R-Squared** = 0.709 showing a useful model accounting for all 8 variables
- **F-Statistic** = 248.1 the model is significant and predictors help to explain the behavior of rent in a meaningful way.

Model Evaluation

The original variable of 'Rent' is expressed in euros (€), the final model evaluates the relationship of the predictors with the dependent variable of the square root of 'Rent' in euros. The dependent variable has been transformed for normalization purposes, error performance, and for overall model integrity. Therefore, each coefficient can be interpreted as the change in the square root of 'Rent' and should not be interpreted as a 1:1 relationship with the predictor variable scale.

Three distinct classes of predictors can be defined: District, Standardized SqMt, Standardized Floor. Top predictors by class based on t-statistic (see annex for full list):

Variable	Coefficient	Interpretation (impact on rent per month)
D-Salamanca	7.5085	Location in Salamanca will increase delta sqrt. rent between 6.037 & 8.980
SqMt (Scaled)	9.0491	1 Stdv increase in Sq.Mt will increase delta sqrt. rent between 8.592 & 9.506
Floor (Scaled)	1.4092	1 Stdv increase in Floor will increase delta sqrt. rent 0.967 & 1.851

Train vs Test Evaluation

Err. Metr.	Train	Test	Interpretation
MAE	4.730	4.926	Model predictions are minor with minor variance from train to test
MSE	39.341	39.344	Model is neither overfitting or underfitting
RMSE	6.272	6.272	No change, stable
MAPE%	10.92	11.548	0.63% minor difference and still a moderately accurate model

Model stability, proper fit, and overall performance were key to evaluating our model success.

Regression Conclusions & Business Recommendations

The linear regression model achieved strong performance (**$R^2 = 0.712$; Adjusted $R^2 = 0.709$**), explaining around **70% of rent price variation**. Error metrics (MAE = 4.9, RMSE = 6.27, MAPE ≈ 11%) confirm **stable, reliable predictions without overfitting.**

Key Drivers of Rental Prices in Madrid

- Location Matters: Premium districts like Salamanca, Retiro, and Chamberí command the highest rents due to prestige and demand.
- Size Drives Price: Larger properties consistently rent for more, showing the market's preference for space.
- Floors Cost: Higher floors increase rent, reflecting better light, views, and comfort.

Business Applications and Market Opportunities

The model supports data-driven pricing and valuation, **helping detect undervalued properties and guide renovation or investment decisions**. It highlights **growth potential in emerging districts** such as Fuencarral, Hortaleza, and Ciudad Lineal, while confirming **stability in premium areas**. Overall, **these insights enable balanced portfolio strategies**, combining steady high-yield assets with value-growth opportunities across Madrid's rental market.

Technical Annex

Here we want to include the details (screenshot, table, or results,) that support any of the claims made above, extracted from our two files.

Data Preparation:

- Missing values, counts

```
Id                0
District          0
Rent              0
Bedrooms         89
Sq.Mt             0
Floor           141
Outer           162
Elevator        133
Penthouse         0
Cottage           0
Duplex            0
Semidetached      0
dtype: int64
```

- Outlier detection

```python
from sklearn.ensemble import IsolationForest

# Detect outliers using IsolationForest
iso = IsolationForest(contamination=0.05, random_state=42)
outlier_pred = iso.fit_predict(x)

# Keep only non-outliers (outlier_pred == 1)
x = x[outlier_pred == 1].reset_index(drop=True)

# Also filter df_original to keep only non-outlier rows
df_original = df_original[outlier_pred == 1].reset_index(drop=True)
```

- New features (calculations, methods, etc.)

```python
import numpy as np
# Add new features

CENTER_TITLE = 'Center'
NORTH_TITLE = 'North'
SOUTH_TITLE = 'South'

# Create a mapping from district to zone
district_zone_map = {
    'Centro': CENTER_TITLE,
    'Arganzuela': CENTER_TITLE,
    'Retiro': CENTER_TITLE,
    'Salamanca': CENTER_TITLE,
    'Chamartín': CENTER_TITLE,
    'Chamberí': CENTER_TITLE,
    'Moncloa': CENTER_TITLE,
    'Tetuán': CENTER_TITLE,

    'Fuencarral': NORTH_TITLE,
    'Hortaleza': NORTH_TITLE,
    'Barajas': NORTH_TITLE,
    'Ciudad Lineal': NORTH_TITLE,
    'San Blás': NORTH_TITLE,

    'Latina': SOUTH_TITLE,
    'Moratalaz': SOUTH_TITLE,
    'Puente Vallecas': SOUTH_TITLE,
    'Carabanchel': SOUTH_TITLE,
    'Usera': SOUTH_TITLE,
    'Vicálvaro': SOUTH_TITLE,
    'Villa de Vallecas': SOUTH_TITLE,
}

# Add the new 'Zone' column
x['Zone'] = x['District'].map(district_zone_map)
x = x.drop(columns=['District'])
```

- Correlations results and handling: Correlation Matrix, threshold = 0.6

	Variable 1	Variable 2	Correlation
0	numerical__Rent	numerical__Sq.Mt	0.826007
1	numerical__Bedrooms	numerical__Sq.Mt	0.740397
2	numerical__Rent	numerical__Bedrooms	0.612110
3	numerical__Sq.Mt	remainder__Cottage	0.605316

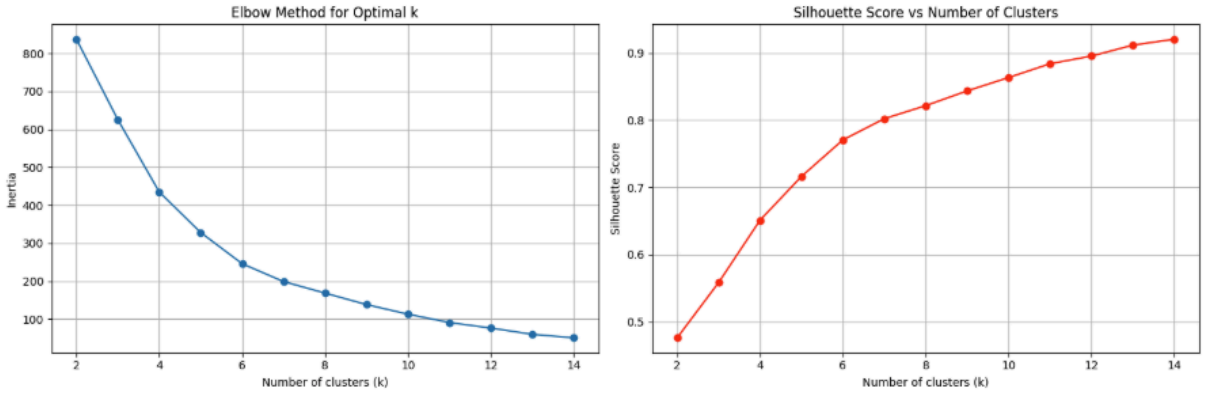
The following variables were removed as a result of the correlation analysis:
['numerical__Sq.Mt', 'numerical__Bedrooms', 'numerical__Rent']

Segmentation:

The evaluation of potential cluster counts combined the Elbow Method and Silhouette analysis to identify the optimal number of groups. As illustrated in the plots above, the Elbow curve (left) shows a marked inflection at K = 5, where the rate of inertia reduction begins to flatten, indicating diminishing gains in compactness from additional clusters.

The Silhouette curve (right) supports this result, displaying a consistent increase in score with a local plateau around the same value. This balance between inertia and Silhouette suggests that five clusters capture sufficient internal cohesion while maintaining clear separation. In

business terms, the five-cluster configuration offers a solution that is both statistically sound and easily interpretable, reflecting real distinctions within Madrid's rental market.




```
Optimal number of clusters based on silhouette score: 14
Best silhouette score: 0.921
```

Silhouette scores and inertia for the different ranges of clusters

```python
# Enhanced elbow method with silhouette analysis
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import numpy as np

RANDOM_STATE = 42

# Test different numbers of clusters
MAX_CLUSTERS = 15
K_range = range(2, MAX_CLUSTERS)
inertias = []
silhouette_scores = []

for k in K_range:
    kmeans = KMeans(n_clusters=k, random_state=RANDOM_STATE, n_init=10)
    kmeans.fit(x)
    inertias.append(kmeans.inertia_)
    silhouette_scores.append(silhouette_score(x, kmeans.labels_))
```

K =	**Silhouette Score**	**Inertias**
3	0.558	626
4	0.651	435
5	0.718	326
6	0.771	246
7	0.8	201

Final Segmentation Selection: 5 Clusters

The clustering model achieved a Silhouette Score of 0.718, indicating a strong degree of separation and internal cohesion among clusters. This means that listings within each group are highly similar to one another while remaining well-differentiated from other segments—a robust technical outcome that also aligns with intuitive, business-level distinctions.

The model's inertia value of 244.237 reflects low within-cluster variance, confirming that the chosen K-Means configuration efficiently minimizes dispersion around cluster centers. In practical terms, this indicates compact, well-defined groups that capture meaningful patterns in Madrid's rental market rather than noise or overlap.

Cluster	**Count**	**% Total**
1	1037	51.2
2	130	6.5
3	343	17.2
4	216	10.8
5	270	13.5

Analyzing the relevance of the variables in the clustering, Outer (1.07) emerged as the primary differentiator, highlighting how exterior exposure and natural light distinctly separate property groups. Penthouse (0.83) and Elevator (0.13) followed, emphasizing the premium associated with top-floor access and building comfort. In contrast, variables such as Sq.Mt, Bedrooms, and Duplex had relatively minor contributions, suggesting that rental value and market segmentation depend more on property type and amenities than on basic spatial metrics.

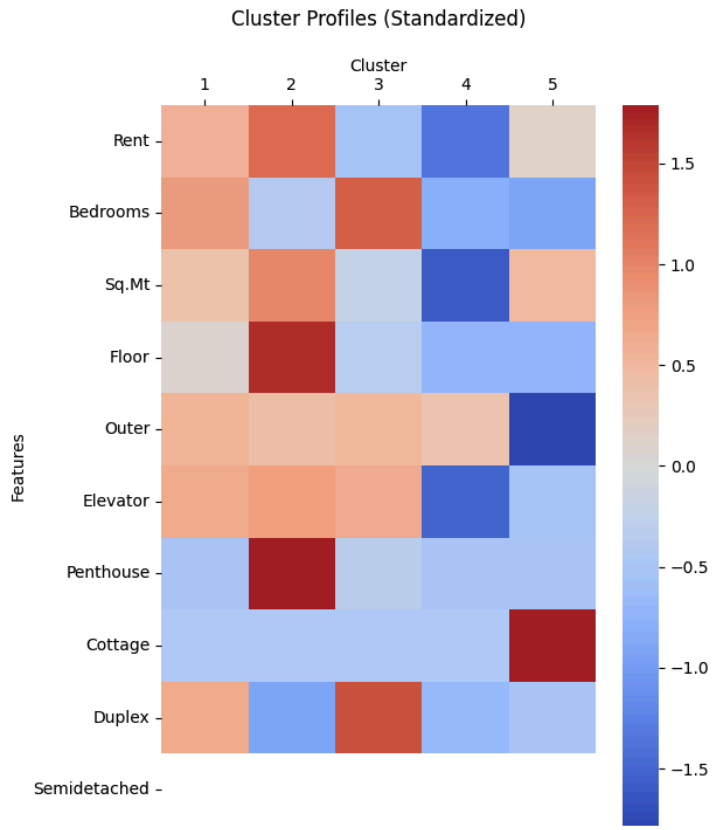
```
Numeric variable importances:

             relative importance
Outer                   1.074584
Penthouse               0.828037
Elevator                0.127308
Rent                    0.121905
Cottage                 0.100850
Floor                   0.061976
Sq.Mt                   0.032673
Bedrooms                0.009717
Duplex                  0.007556
Semidetached                 NaN
```

The combination of the summary table and visualizations provides a complete picture of how each cluster differentiates itself both numerically and spatially. The table of averages highlights the key quantitative contrasts between clusters, such as the variation in rent, size, and building features, offering a precise numerical foundation for interpretation.

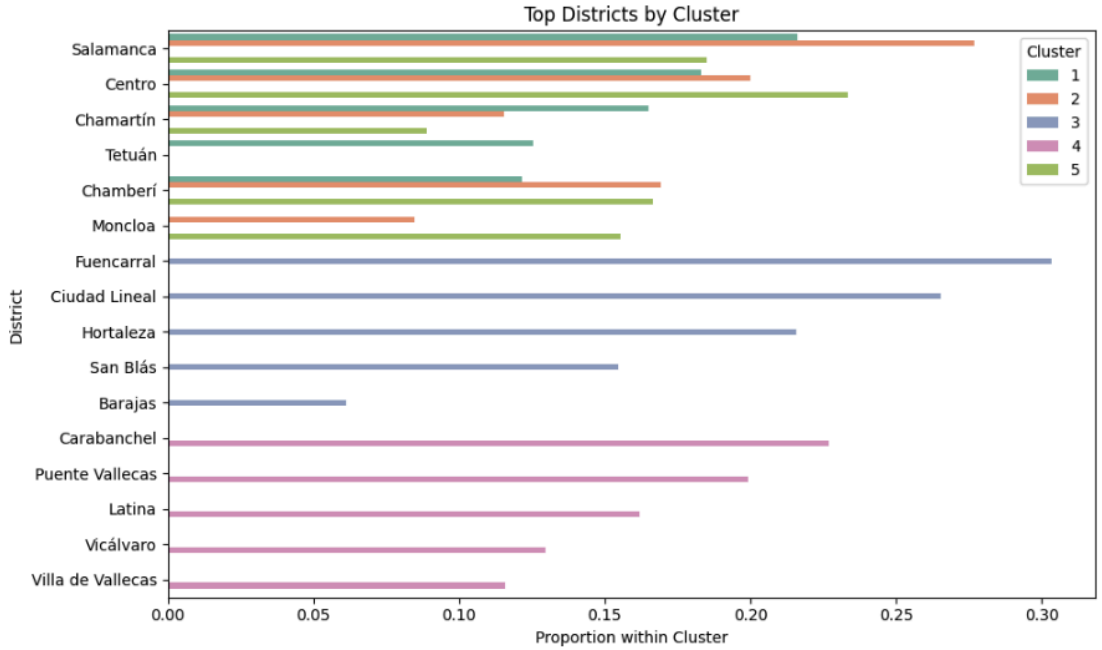
```
import pandas as pd

# df_prof will store cluster averages to calculate profile and global metrics
df_prof = df_original.drop(columns="Id", axis=1).select_dtypes(include=['number'])
df_prof = df_prof.groupby("Cluster").mean().transpose()
df_prof["Average"] = df_prof.mean(axis=1)
df_prof["Std.dev"] = df_prof.std(axis=1)
df_prof
```

Cluster	1	2	3	4	5	Average	Std.dev
Rent	2190.910318	2643.492308	1373.594752	779.773148	1876.755556	1772.905216	646.404702
Bedrooms	2.470352	2.293651	2.544643	2.232227	2.215768	2.351328	0.132231
Sq.Mt	129.423337	146.784615	111.655977	71.847222	132.200000	118.382230	25.808067
Floor	3.710756	5.676923	3.189759	2.670984	2.690583	3.587801	1.112074
Outer	1.000000	0.952756	0.981873	0.910526	0.000000	0.769031	0.385703
Elevator	0.939981	0.953846	0.934328	0.614213	0.757848	0.840043	0.133853
Penthouse	0.000000	1.000000	0.078717	0.000000	0.000000	0.215743	0.393312
Cottage	0.000000	0.000000	0.000000	0.000000	0.114815	0.022963	0.045926
Duplex	0.027001	0.000000	0.040816	0.004630	0.007407	0.015971	0.015466
Semidetached	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

The heatmap of standardized cluster profiles transforms these differences into an intuitive visual form, allowing immediate recognition of patterns—such as the dominance of luxury indicators (e.g., penthouse, elevator) in Cluster 2 or the low-rent, low-amenity structure of Cluster 4. This visualization translates statistical distinctions into a clear, interpretable gradient of property characteristics.

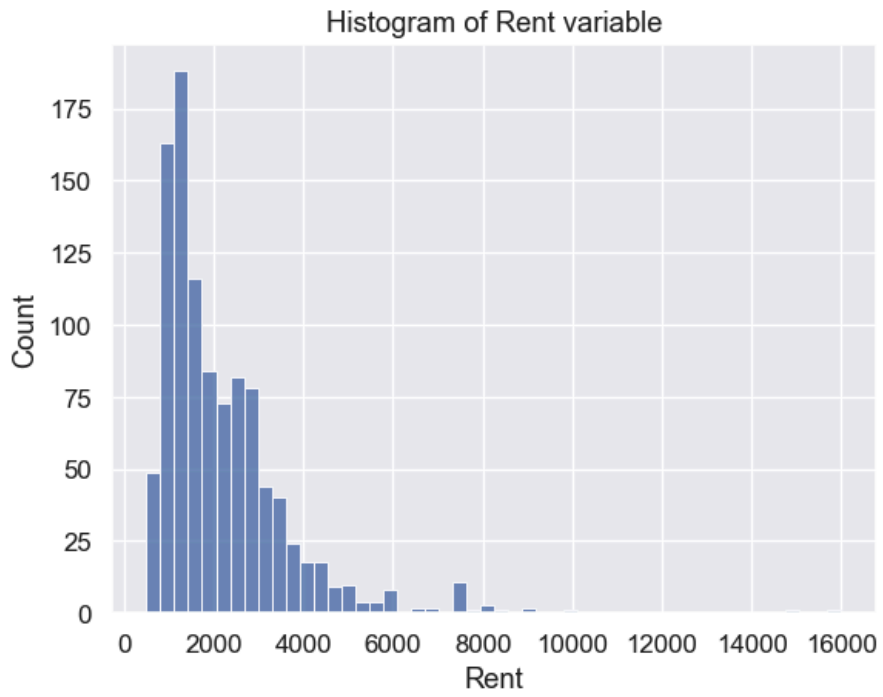


Finally, the district distribution bar chart links each cluster to its geographic footprint, revealing where each housing segment is most prevalent. By connecting data patterns to physical locations, it validates the clustering's business relevance, showing that technical segmentation aligns naturally with Madrid's known socio-spatial structure.



Regression Modeling

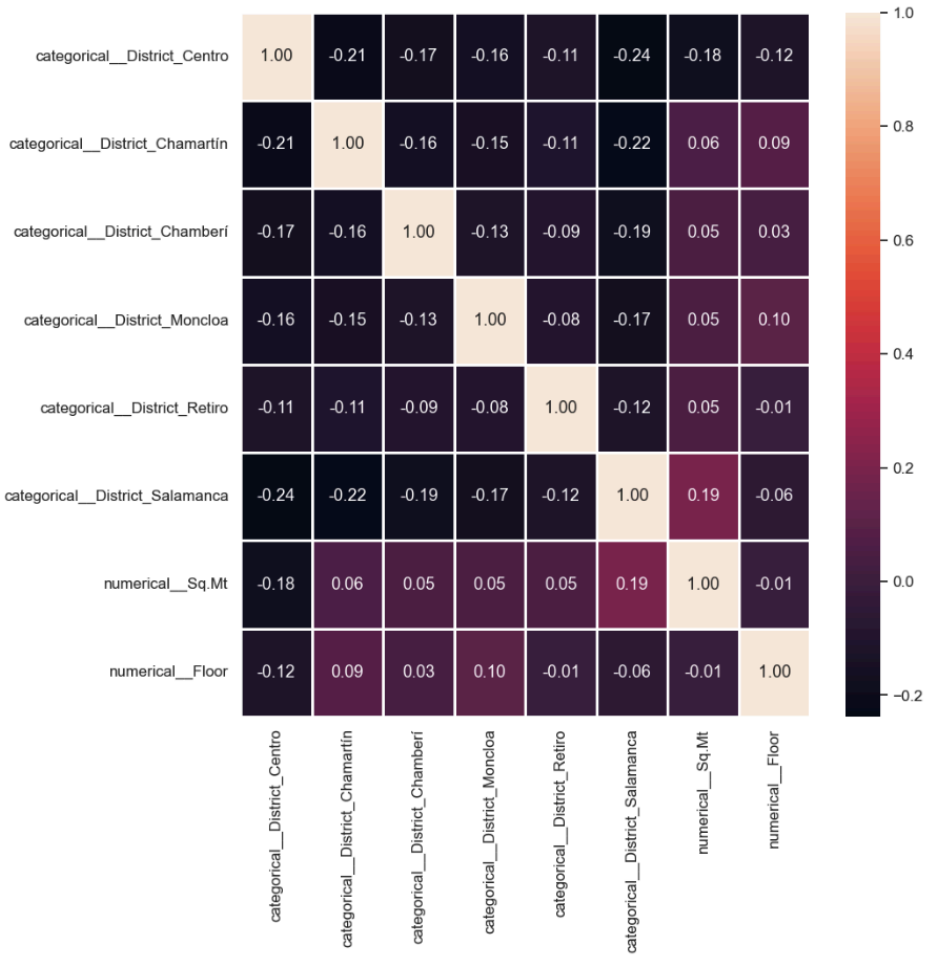
- Histogram of Rent



Analysis of 'Rent' variable to determine shape, presence, and impact of outliers to look for transformation options. We decided to attempt to apply a log transformation which was too aggressive and chose to follow a square root application. This managed to help improve the residuals and error performance, but we were still dealing with skewing due to outliers. We then applied an outlier solution looking at z-scores and excluding values outside of 3.

- Stepwise Results:

```
Add  numerical__Sq.Mt               with p-value 4.481e-190
Add  categorical__District_Tetuán   with p-value 2.215e-10
Add  categorical__District_Salamanca with p-value 6.616e-08
Add  numerical__Floor               with p-value 8.197e-09
Add  categorical__District_Centro   with p-value 0.0009036
Add  categorical__District_Chamberí with p-value 0.002642
Add  categorical__District_Retiro   with p-value 0.002246
Add  categorical__District_Chamartín with p-value 0.01687
Drop categorical__District_Tetuán   with p-value 0.1199
Add  categorical__District_Moncloa  with p-value 0.01181
resulting features:
['numerical__Sq.Mt', 'categorical__District_Salamanca', 'numerical__Floor', 'categorical__District_Centro', 'cat
egorical__District_Chamberí', 'categorical__District_Retiro', 'categorical__District_Chamartín', 'categorical__D
istrict_Moncloa']
```



Full Correlation Matrix: applied to chosen features after stepwise selection to ensure we are not going to have any multicollinearity issues.

- Regression output

OLS Regression Results

Dep. Variable:	Rent	R-squared:	0.712
Model:	OLS	Adj. R-squared:	0.709
Method:	Least Squares	F-statistic:	248.1
Date:	Thu, 09 Oct 2025	Prob (F-statistic):	5.06e-211
Time:	20:32:12	Log-Likelihood:	-2639.9
No. Observations:	811	AIC:	5298.
Df Residuals:	802	BIC:	5340.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	39.5742	0.545	72.559	0.000	38.504	40.645
categorical__District_Centro	5.6198	0.751	7.485	0.000	4.146	7.094
categorical__District_Chamartín	3.4161	0.778	4.389	0.000	1.888	4.944
categorical__District_Chamberí	5.2421	0.843	6.218	0.000	3.587	6.897
categorical__District_Moncloa	2.2344	0.885	2.524	0.012	0.496	3.972
categorical__District_Retiro	5.6353	1.084	5.201	0.000	3.508	7.762
categorical__District_Salamanca	7.5085	0.750	10.013	0.000	6.037	8.980
numerical__Sq.Mt	9.0491	0.233	38.846	0.000	8.592	9.506
numerical__Floor	1.4092	0.225	6.257	0.000	0.967	1.851

Omnibus:	117.910	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	197.455
Skew:	0.925	Prob(JB):	1.33e-43
Kurtosis:	4.557	Cond. No.	7.63

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Key points

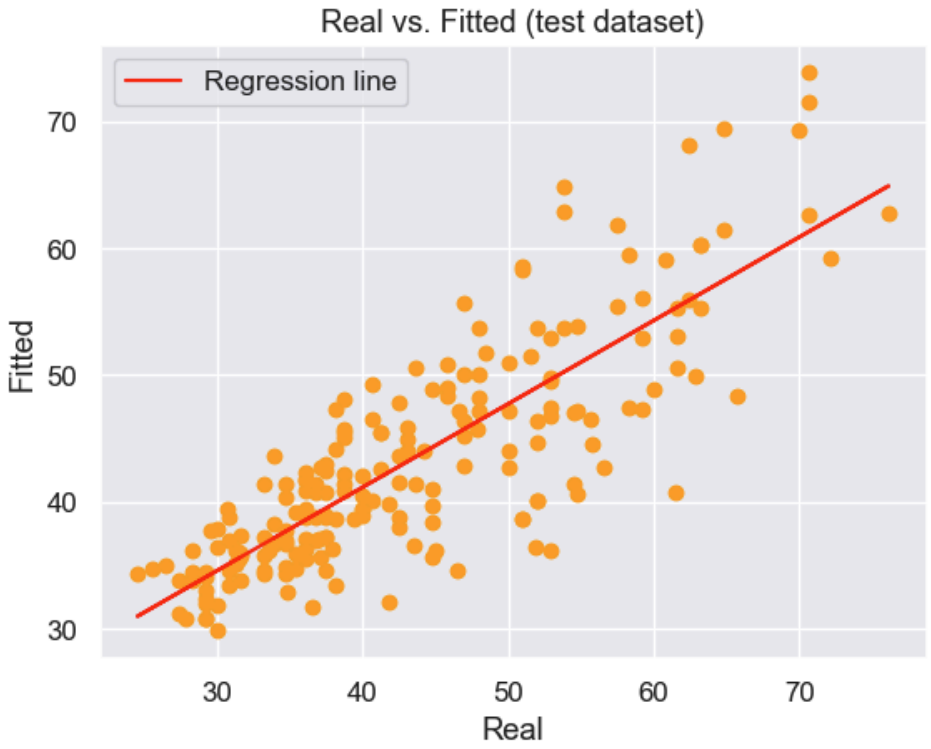
- R^2: 0.712, selected variables explain almost 70% of rent behavior as a target variable.
- Adj. R^2: 0.709, very close to the R^2 value, indicating the model is likely not at risk of overfitting.
- F-Stat= Reject null hypothesis, there are significant differences between groups and variables with rent as a target variable in this model.
- Prob (F-Stat): close to zero, variables likely have a real impact on rent as a target variable.
- All P>| t | < 0.05 (integrity check)
- Confidence Intervals do not include 1.00 (integrity check)
- 811 Observations for 8 predictors (integrity check)
- Scaled predictors: numerical_Sq.Mt coefficient interpretation:
  Sq.Mt. (t-val: 38.846) with a coefficient of +9.0491, has been scaled in the feature engineering process using StandardScaler, the impact of which results in that the output of the coefficient will increase +9.0491 square root of rent per each standard deviation increase.
- Skewness: still a little skewed but corrections improved model
- Kurtosis:4.557, A few extreme rent values remain, but model assumptions of normality are largely acceptable
- DW: no autocorrelation present in the model Cond. No. no multicollinearity (we already knew that though)

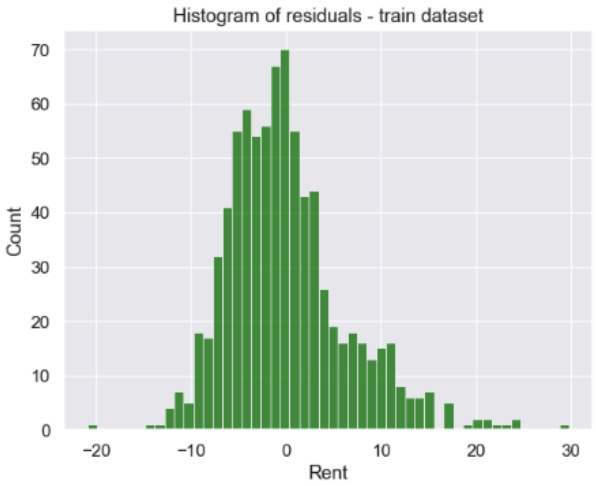
Explanation of significant variables:

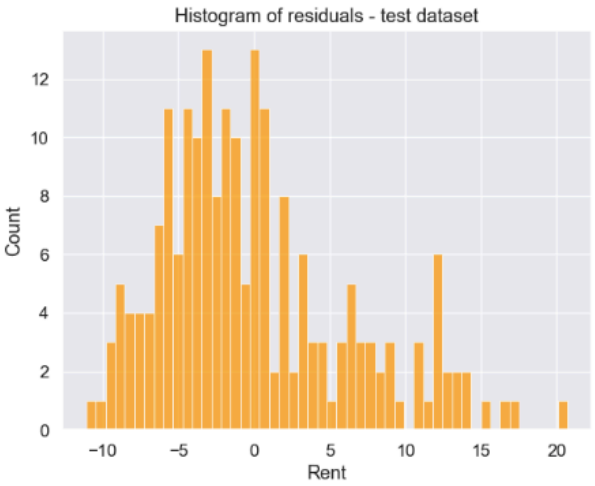
Variable	Coefficient	Interpretation (impact on rent per month)
D-Centro	5.6198	Location in Centro will increase delta sqrt. rent +5.6198
D-Chamartín	3.4161	Location in Chamartín will increase delta sqrt. rent +3.4164
D-Chamberí	5.2421	Location in Chamberí will increase delta sqrt. rent +5.2421
D-Moncloa	2.2344	Location in Moncloa will increase delta sqrt. rent +2.2344
D-Retiro	5.6353	Location in Retiro will increase delta sqrt. rent +€5.6353
D-Salamanca	7.5085	Location in Salamanca will increase rent +€687
SqMt (Scaled)	9.0491	1 Stdv increase in Sq.Mt will increase delta sqrt. rent 9.0491
Floor	1.4092	1 Stdv increase in Floor will increase delta sqrt. rent 1.4092

Model Technical Performance/Evaluation

- Heteroscedasticity
- Normality on residuals
- Actual vs predicted plots
- Train/Testing results


Real vs. Fitted (test dataset)
Regression line
Fitted
Real
70
60
50
40
30
30
40
50
60
70


Histogram of residuals - train dataset
Count
Rent
70
60
50
40
30
20
10
0
−20
−10
0
10
20
30


Histogram of residuals - test dataset
Count
Rent
12
10
8
6
4
2
0
−10
−5
0
5
10
15
20