## 

**ADEI: Spain Real Estate Analysis**

REPORT 1

2024-2025

Joan Marc Coll, Guillem Gaya, Dimas Noguera y Ada Peña

## 

[**1. Data Source 2**](#_f9fi31f3bvuw)

[**2. Data Description 2**](#_25djlq2nkde3)

[2.1 Description of Variables 2](#_3xsclg2qf7wo)

[2.2 Descriptive Statistics of Variables 4](#_gs99yx8c0uls)

[2.3 Frequency of Variables 9](#_6giq3v321ovz)

[**3. Data Preprocessing 11**](#_4x21r0gnlb1v)

[3.1 Data Renaming 11](#_s5nxw59ld6ip)

[3.2 Handling Missing Data 12](#_1fiidx3u73vj)

[3.3 Changing Variable Type 13](#_axi8s02f5rbh)

[**4. Description of the preprocessed dataset 14**](#_st89rqlof6a)

[4.2 Descriptive Statistics of Preprocessed Variables 15](#_vf5ngui6rhn1)

[4.3 Frequency of Preprocessed Variables 17](#_dkhpux12pda)

[**5. Plot Comparison 19**](#_sch05gz0omd)

## 

# 1. Data Source

The data for our study is available at the following link provided by Data Market:  
<https://datamarket.es/media/samples/inmuebles-sample.csv>

The primary objective is to predict the type of operation ("operation") of the properties.

# 2. Data Description

The selected dataset contains 1000 instances and 24 variables that provide information about different types of properties.

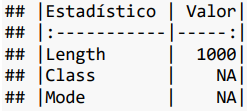
## 2.1 Description of Variables

* **website**Variable type: "character".  
  Description: Website where the property is advertised.
* **url**Variable type: "character".  
  Description: Web link to the property advertisement.  
  Additional information: Initially, this variable does not provide useful information as each instance will be unique, and it does not provide any information about the property.
* **reference**Variable type: "character".  
  Description: Advertisement identifier.
* **country**Variable type: "character".  
  Description: Country code of the property.  
  Additional information: 100% of the instances have the same value "ES," meaning all the properties are from Spain. Therefore, this variable does not provide meaningful information for the current dataset unless properties from other countries are added.
* **province**Variable type: "character".  
  Description: Province where the property is located.
* **location**Variable type: "character".  
  Description: Location of the property (street, avenue, etc.).
* **title**Variable type: "character".  
  Description: Title of the property advertisement.
* **description**Variable type: "character".  
  Description: Description of the property advertisement.
* **price**Variable type: "numeric".  
  Description: Property price in euros.
* **operation**Variable type: "character".  
  Description: Type of operation on the property (sale or rent).  
  Additional information: This is the target variable of our dataset.
* **property\_type**Variable type: "character".  
  Description: Type of property (shops, land, etc.).
* **rooms**Variable type: "integer".  
  Description: Number of bedrooms in the property.
* **baths**Variable type: "integer".  
  Description: Number of bathrooms in the property.
* **area**Variable type: "integer".  
  Description: Property area in square meters.
* **floor**Variable type: "integer".  
  Description: Number of floors in the property.
* **elevator**Variable type: "logical".  
  Description: Whether the property has an elevator or not.
* **outside**Variable type: "logical".  
  Description: Whether the property has an outdoor space or not.
* **floor.1**Variable type: "integer".  
  Description: Number of floors in the property.  
  Additional information: This variable is a duplicate of the "floor" variable.
* **images**Variable type: "integer".  
  Description: Number of images of the property on the advertiser's website.
* **latitude**Variable type: "numeric".  
  Description: Latitude of the property.
* **longitude**Variable type: "numeric".  
  Description: Longitude of the property.
* **dealer**Variable type: "character".  
  Description: Name of the dealer (companies).
* **dealer\_url**Variable type: "character".  
  Description: Web link to the dealer.
* **dealer\_is\_professional**Variable type: "character".  
  Description: Indicates whether the dealer is a professional or not.

## 2.2 Descriptive Statistics of Variables

We do not show descriptive statistics for all non-numeric variables, as they are all similar to the example shown below for “website.”

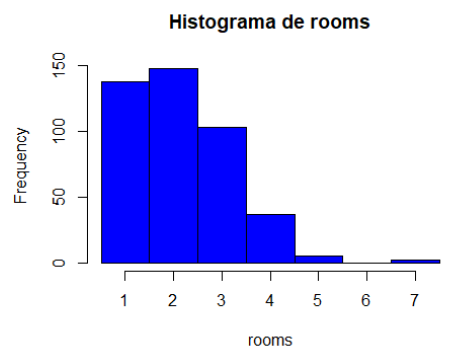
* **website**

****

* **price**

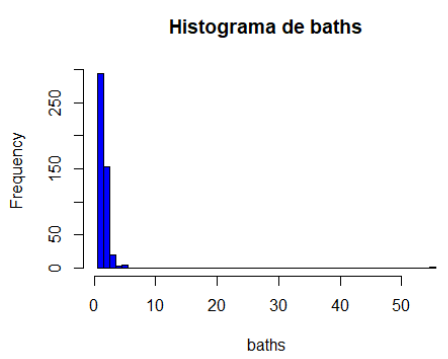
| **Statistical** | **Value** |
| --- | --- |
| Min. | 374.0 |
| 1st Qu. | 1200.0 |
| Median | 36500.0 |
| Mean | 156967.6 |
| 3rd Qu. | 196890.0 |
| Max. | 9875000.0 |
| NA's | 6.0 |

The price variable displays a broad range, from a minimum of €374 to a maximum of €9,875,000. The high mean (€156,967.6) and significant disparity from the median (€36,500) highlight the presence of high-value outliers, likely reflecting luxury properties or extensive commercial lots. The data reflects both rental and sale transactions, contributing to the diversity in price distribution.

* **rooms**

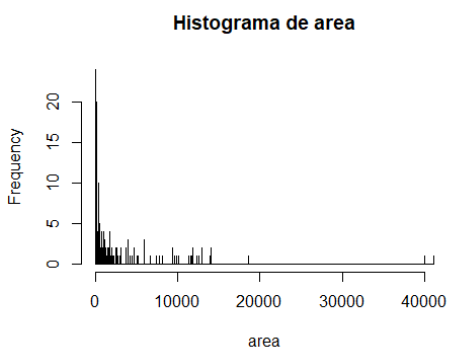
| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 1,000000 |
| Median | 2,000000 |
| Mean | 2,147806 |
| 3rd Qu. | 3,000000 |
| Max. | 7,000000 |
| NA's | 567,000000 |

Properties show a median of 2 rooms, with a maximum of 7 rooms. Interestingly, 56.7% of the data is missing, as not all property types (e.g., land) include room counts. The structured missingness suggests inherent differences in property characteristics.

* **baths**

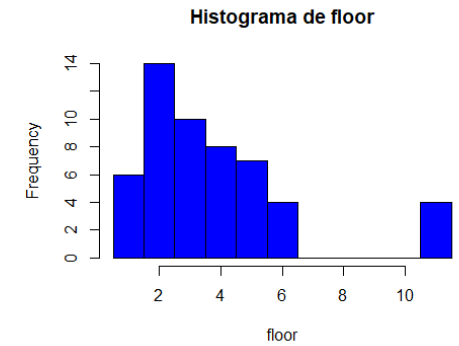
| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 1,000000 |
| Median | 1,000000 |
| Mean | 1,572632 |
| 3rd Qu. | 2,000000 |
| Max. | 55,000000 |
| NA's | 525,000000 |

The median remains 1 bathroom, but the presence of outliers, such as properties with 55 bathrooms, likely represents unique or large-scale properties. Missing values (52.5%) correspond to property types without bathrooms.

* **area**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 6,0000 |
| 1st Qu. | 70,0000 |
| Median | 120,0000 |
| Mean | 764,6817 |
| 3rd Qu. | 511,2500 |
| Max. | 41057,0000 |
| NA's | 48,0000 |

The average property size is 764.68 m², with most properties clustering below 500 m². However, outliers extend up to 41,057 m², likely large commercial or industrial lots. Missing values are minimal (4.8%), indicating robust coverage of this variable.

* **floor**

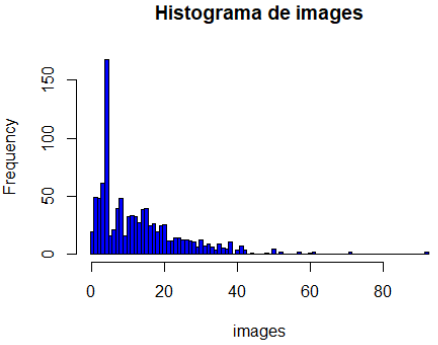
| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 2,000000 |
| Median | 3,000000 |
| Mean | 3,754717 |
| 3rd Qu. | 5,000000 |
| Max. | 11,000000 |
| NA's | 947,000000 |

94.7% of instances have missing values for the "floor" variable. The primary reason is that not all properties have floors, and some are not on any floor. Therefore, some values should be 0 instead of N/A (structured missing values). Although we could address this issue, there are too many missing values, and a significant part of them are not structured, so it is simpler to eliminate this variable.

**- floor.1**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 1,000000 |
| 1st Qu. | 2,000000 |
| Median | 3,000000 |
| Mean | 3,754717 |
| 3rd Qu. | 5,000000 |
| Max. | 11,000000 |
| NA's | 947,000000 |

This variable is an exact replica of the “floor” variable, so it makes no sense to keep it.

* **images**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 0,000 |
| 1st Qu. | 4,000 |
| Median | 10,000 |
| Mean | 12,899 |
| 3rd Qu. | 18,000 |
| Max. | 92,000 |
| NA's | - |

There may be some variation in the number of images for different types of properties due to their size and type (e.g., more images may be needed to show an industrial warehouse than a small apartment).

* **latitude**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 28,00509 |
| 1st Qu. | 28,45991 |
| Median | 39,57950 |
| Mean | 37,05438 |
| 3rd Qu. | 41,59779 |
| Max. | 43,42275 |
| NA's | - |

* **longitude**

| **Statistical** | **Value** |
| --- | --- |
| Min. | -17,767417 |
| 1st Qu. | -16,309590 |
| Median | -1,645366 |
| Mean | -5,852112 |
| 3rd Qu. | 1,661144 |
| Max. | 2,763666 |
| NA's | - |

## 

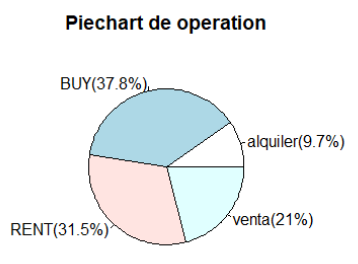
## 2.3 Frequency of Variables

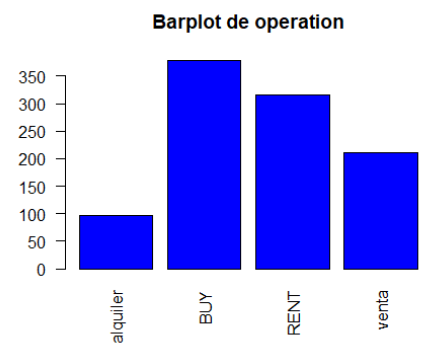
We will visualize the categorical variables to examine their content more closely and check if any modifications are necessary.

* **province**

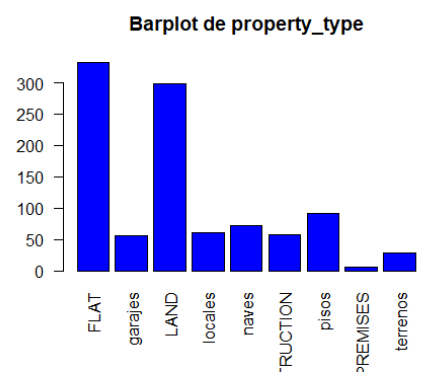
| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| alava | 18 | 0,02 |
| albacete | 26 | 0,03 |
| barcelona | 299 | 0,3 |
| caceres | 4 | 0 |
| lugo | 6 | 0,01 |
| malaga | 18 | 0,02 |
| navarra | 57 | 0,06 |
| navarra\_nafarroa | 4 | 0 |
| pontevedra | 98 | 0,1 |
| santa-cruz-de-tenerife | 314 | 0,31 |
| tarragona | 1 | 0 |
| valencia | 155 | 0,16 |

Barcelona (“B”) and Santa Cruz de Tenerife (“TF”) have the highest representation, accounting for 30% and 31% of properties, respectively. Provinces like Tarragona (“T”) have minimal representation, suggesting potential regional concentration in the dataset.

* **operation**

****

The names of the values need to be modified since “BUY” is equivalent to “venta,” and “RENT” means “alquiler.” This means there are values indicating the same thing but with different names.

* **property\_type**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| FLAT | 332 | 0,33 |
| garajes | 56 | 0,06 |
| LAND | 299 | 0,3 |
| locales | 60 | 0,06 |
| naves | 72 | 0,07 |
| NEW\_CONSTRUCTION | 57 | 0,06 |
| pisos | 91 | 0,09 |
| PREMISES | 5 | 0 |
| terrenos | 28 | 0,03 |

There is a similar issue, with values indicating the same thing but with different names, which must be fixed. Some names are also too long to be displayed correctly. Flats (“FLAT”) and land (“LAND”) are the most common property types, comprising 38% and 33%, respectively. Other categories like garages (“GARAGE”) and premises (“PREMISES”) show lesser frequencies, emphasizing a residential focus.

* **dealer\_is\_profesional**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| TRUE | 1000 | 1 |

This variable has the same value for all instances.

# 

# 3. Data Preprocessing

## 3.1 Data Renaming

Some data were in Spanish and others in English, so we unified everything into English. The main reason for this was that the English names were shorter, allowing for better visualization in the graphs.

Changes to the "operation" column:

* “RENT” instead of "alquiler"
* “BUY” instead of "venta"

Changes to the "property\_type" column:

* “FLAT” instead of "pisos"
* “LAND” instead of "terrenos"
* “N.C.” instead of "NEW\_CONSTRUCTION"
* “GARAGE” instead of "garajes"
* “I.U.” instead of "naves"
* “PREMISES” instead of "locales"

We shortened the names of provinces to make them easier to represent.

Changes to the "province" column:

* “VI” instead of "alava"
* “AB” instead of "albacete"
* “B” instead of "barcelona"
* “CC” instead of "caceres"
* “LU” instead of "lugo"
* “MA” instead of "malaga"
* “NA” instead of "navarra"
* “NA” instead of "navarra\_nafarroa"
* “PO” instead of "pontevedra"
* “TF” instead of "santa-cruz-de-tenerife"
* “T” instead of "tarragona"
* “V” instead of "valencia"

## 3.2 Handling Missing Data

In our initial dataset, some columns had either all or most of the data as N/A, so we eliminated those columns that did not contribute to the study.

Additionally, the "floor.1" column is an exact copy of the "floor" column, so regardless of the amount of N/A values in "floor.1," it was eliminated.

Columns eliminated:

* **elevator** (100% N/A values)
* **outside** (100% N/A values)
* **floor** (94.7% N/A values)
* **floor.1** (94.7% N/A values)

For the “rooms” variable, we found that some data had N/A values. When we checked the “property\_type” column and saw that the property was a “FLAT,” we assumed it was a studio, and for this, we assigned the value **STUDIO** to those instances.

**Non-random Missing Data:**After making these corrections, we addressed each remaining missing value to handle the data appropriately.  
For the “rooms” and “baths” columns, we found many structural missing values since some properties do not have rooms or bathrooms, so we replaced these values with 0.

**Random Missing Data:**In the “area” column, we encountered more missing values, this time random ones. To correctly impute the data, we applied a **KNN algorithm** for each type of property, ensuring the values resemble each other as closely as possible.  
We applied the same method to the missing values in the “prices” column, but also distinguished between “BUY” and “RENT” operations.

Notably, we encountered an instance of type “N.C.” (new construction) that we could not process due to time constraints. To simplify the study, we applied the algorithm in a general manner.

## 

## 

## 3.3 Changing Variable Type

Some variables were declared as a data type that was not appropriate for the study; since all categorical variables were classified as "character," we converted all the "character" columns to **factor** type, except for variables that are not categorical.

# 

# 4. Description of the preprocessed dataset

After processing all the information in our database and discarding unnecessary data for our study, we are left with a database that still contains 1000 instances but has been reduced to 21 variables, with each one providing the same information as before preprocessing.

The variables we will be working with are as follows:

* **website**

Variable type: character

* **url**

Variable type: character

* **reference**

Variable type: character

* **country**

Variable type: character

* **location**

Variable type: character

* **province**

Variable type: factor

* **title**

Variable type: character

* **description**

Variable type: character

* **price**

Variable type: integer

* **operation**

Variable type: factor

* **property\_type**

Variable type: factor

* **rooms**

Variable type: integer

* **baths**

Variable type: integer

* **area**

Variable type: integer

* **images**

Variable type: integer

* **latitud**

Variable type: numeric

* **longitud**

Variable type: numeric

* **dealer**

Variable type: character

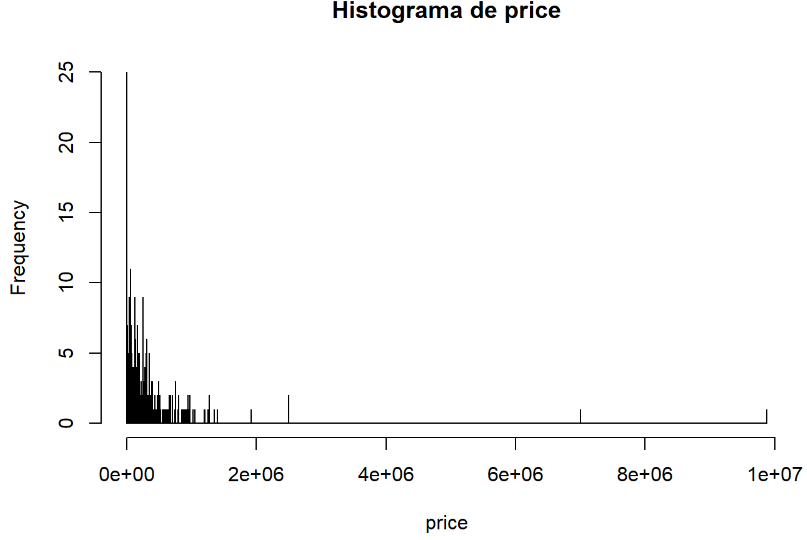
* **dealer\_url**

Variable type: character

* **dealer\_is\_profesional**

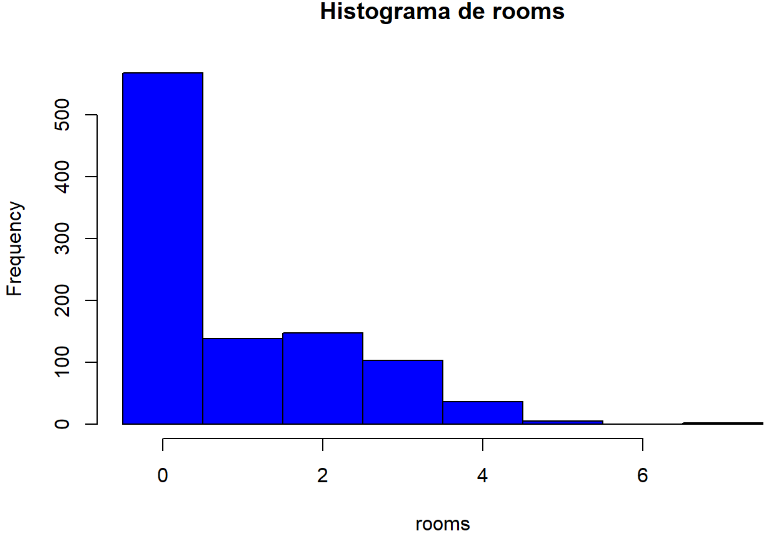
Variable type: factor

## 4.2 Descriptive Statistics of Preprocessed Variables

* **price**

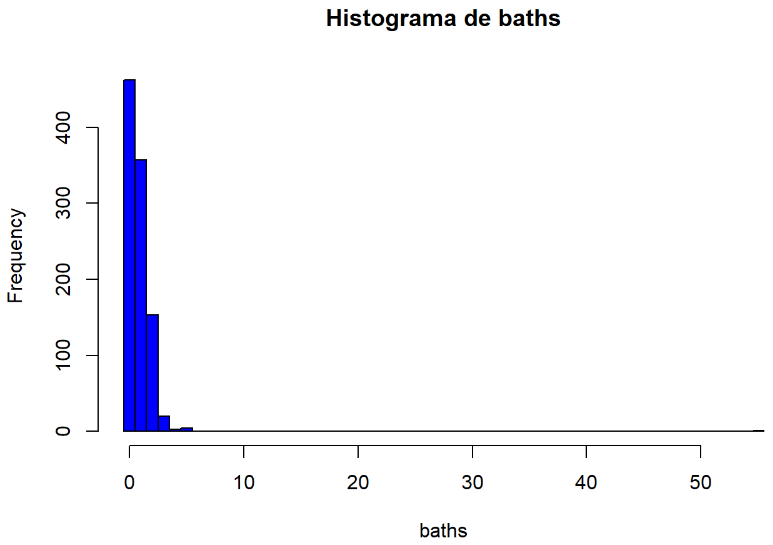
| **Statistical** | **Value** |
| --- | --- |
| Min. | 374,0 |
| 1st Qu. | 1200,0 |
| Median | 35320,0 |
| Mean | 156083,2 |
| 3rd Qu. | 195390,0 |
| Max. | 9875000,0 |
| NA's | - |

Post-preprocessing, imputation using KNN has maintained the variable’s original distribution. Key statistics, such as the mean (€156,083.2) and median (€35,320), remain consistent, ensuring no significant distortion from imputation.

* **rooms**

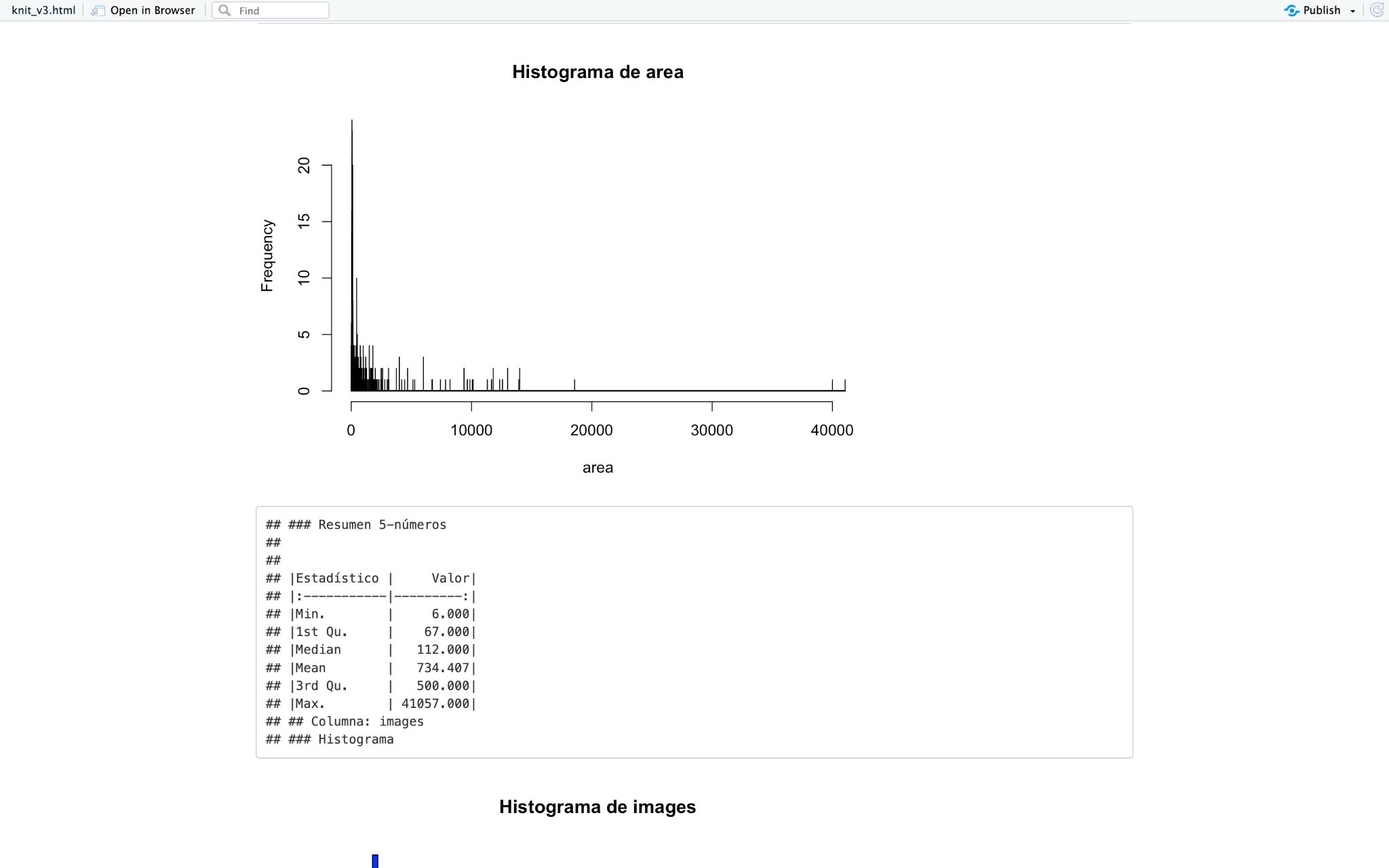
| **Statistical** | **Value** |
| --- | --- |
| Min. | 0,00 |
| 1st Qu. | 0,00 |
| Median | 0,00 |
| Mean | 0,93 |
| 3rd Qu. | 2,00 |
| Max. | 7,00 |
| NA's | - |

Structural missing values were treated, leading to a concentration at 0 for properties without rooms (e.g., land). The mean dropped to 0.93, emphasizing the impact of imputations.

* **baths**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 0,00 |
| 1st Qu. | 0,00 |
| Median | 1,00 |
| Mean | 0,81 |
| 3rd Qu. | 1,00 |
| Max. | 55,00 |
| NA's | - |

Similar to rooms, preprocessing has created a cluster at 0 for properties without bathrooms. The mean adjusted to 0.81 reflects this shift, while outliers like 55 bathrooms remain unchanged.

* **area**

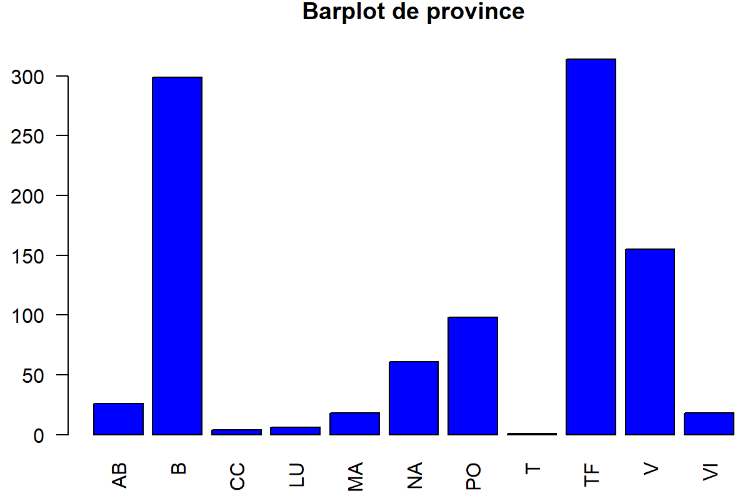
| **Statistical** | **Value** |
| --- | --- |
| Min. | 6,000 |
| 1st Qu. | 67,000 |
| Median | 112,000 |
| Mean | 734,407 |
| 3rd Qu. | 500,000 |
| Max. | 41057,000 |
| NA's | - |

Minimal changes occurred during preprocessing, as missing values were few and imputation effectively retained the original distribution. The mean size (∼734 m²) still reflects the large variation in property types.

## 

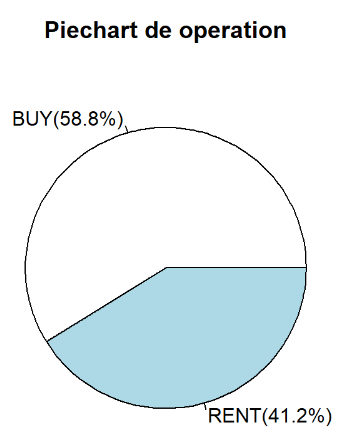
## 4.3 Frequency of Preprocessed Variables

We will visualize the categorical variables after the preprocessing to examine that their content has been modified as intended and to see if the modified dataset doesn't have any other problem we didn’t see.

**- province**

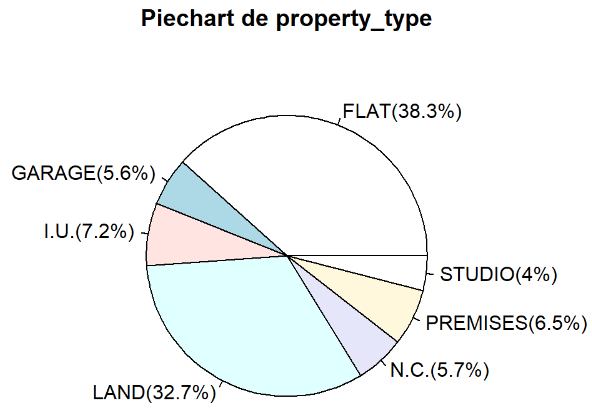
| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| AB | 26 | 0,03 |
| B | 299 | 0,3 |
| CC | 4 | 0 |
| LU | 6 | 0,01 |
| MA | 18 | 0,02 |
| NA | 61 | 0,06 |
| PO | 98 | 0,1 |
| T | 1 | 0 |
| TF | 314 | 0,31 |
| V | 155 | 0,16 |
| VI | 18 | 0,02 |

Renaming of province codes was successful, simplifying visualization and maintaining clarity. High representation in Barcelona (“B”) and Santa Cruz de Tenerife (“TF”) persists.

**- operation**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| BUY | 588 | 0,59 |
| RENT | 412 | 0,41 |

Post-renaming, the unified categories (“BUY” and “RENT”) ensure consistent analysis. The 59%-41% split underscores the dataset’s slight emphasis on sales transactions.

**- property\_type**

| **Class** | **Abs. Freq.** | **Rel. Freq.** |
| --- | --- | --- |
| FLAT | 383 | 0,38 |
| GARAGE | 56 | 0,06 |
| I.U. | 72 | 0,07 |
| LAND | 327 | 0,33 |
| N.C. | 57 | 0,06 |
| PREMISES | 65 | 0,06 |
| STUDIO | 40 | 0,04 |

Preprocessing improved category clarity by unifying naming conventions (e.g., “FLAT” for all flats). Dominance of flats and land types remains apparent, reinforcing the dataset’s residential nature.

# 5. Plot Comparison

* **rooms**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **baths**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **area**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **operation**

| **Original** | **Processed** |
| --- | --- |
|  |  |

* **property\_type**

| **Original** | **Processed** |
| --- | --- |
|  |  |