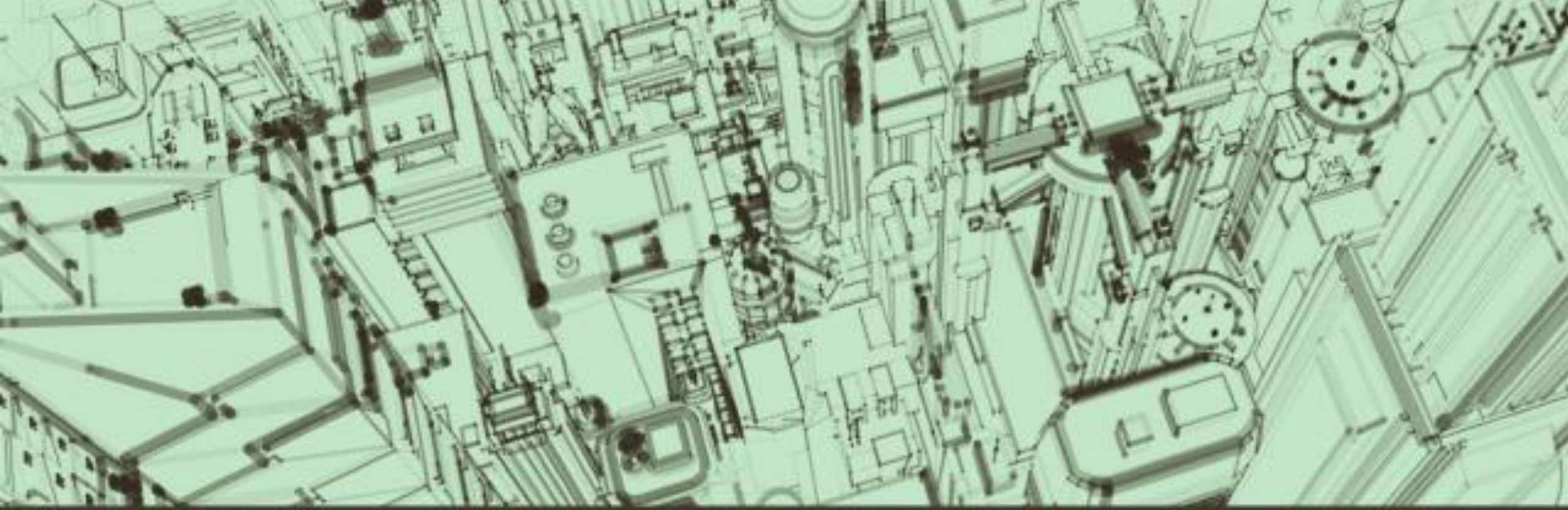


Análisis Exploratorio de un Dataset de Precios de Propiedades

Grupo 3: Gastón, Juan, Daniel

Etapas


Instalación de Geopandas y GDrive
Exploración de datos
Análisis Descriptivo
Limpieza de Datos
Análisis Posterior

Exploración de Datos

Se obtiene información general del dataset, tal como:

- Cantidad de entradas
- Cantidad de columnas
- Tipo de dato de cada campo
- Celdas no nulas por campo
- Resumen de tipos de datos
- Memoria ocupada
- Gráficos básicos

Datos Completos

Datos Parcialmente Completos

Datos Complementarios

```
[ ]  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121220 entries, 0 to 121219
Data columns (total 26 columns):
Unnamed: 0                  121220 non-null int64
operation                   121220 non-null object
property_type               121220 non-null object
place_name                  121197 non-null object
place_with_parent_names     121220 non-null object
country_name                121220 non-null object
state_name                  121220 non-null object
geonames_id                 102503 non-null float64
lat-lon                     69670 non-null object
lat                         69670 non-null float64
lon                         69670 non-null float64
price                       100810 non-null float64
currency                    100809 non-null object
price_aprox_local_currency  100810 non-null float64
price_aprox_usd             100810 non-null float64
surface_total_in_m2         81892 non-null float64
surface_covered_in_m2       101313 non-null float64
price_usd_per_m2            68617 non-null float64
price_per_m2                87658 non-null float64
floor                       7899 non-null float64
rooms                       47390 non-null float64
expenses                    14262 non-null float64
properati_url               121220 non-null object
description                 121218 non-null object
title                       121220 non-null object
image_thumbnail             118108 non-null object
dtypes: float64(13), int64(1), object(12)
memory usage: 24.0+ MB
```

Exploración de Datos

Columnas	Estado Inicial
operation, property_type	Completos y sin error
place_name, place_with_parent_names country_name, state_name	Prácticamente completos. Solo 23 registros de Tigre sin place_name
geonames_id, lat-lon, lat,lon	Parcialmente completos y con errores
price, currency, price_aprox_local_currency, price_aprox_usd, surface_total_in_m2, surface_covered_in_m2, price_usd_per_m2, price_per_m2,	Información incompleta, Distintas monedas, valores erróneos, precios por metro cuadrado mal calculados
floor, rooms, expenses,	Información incompleta, datos errónos
properati_url, description, title, image_thumbnail	Campos con información que podría servir para completar datos sobre las propiedades. Fuentes alternativas de información

Análisis Descriptivo de la Información

Se calculan los parámetros estadísticos (media, desvío, cuartiles, máximos y mínimos) para los campos correspondientes a variables cuantitativas:

- Precio
- Precio aproximado en moneda local
- Precio aproximado en dólares
- Superficie total en m^2
- Superficie cubierta en m^2
- Precio por m^2 (en dólares)
- Precio por m^2 (en pesos)
- Nº de piso
- Cantidad de habitaciones
- Valor de expensas

Relleno de campo lat-long

- **Campo geonames_id**
- **Cruce con base de datos de codificación geográfica**
- **Resultado: 95% de los datos georreferenciados**

Limpieza de Datos

- **Operaciones iniciales (indicadores de missing values y limpieza de caracteres)**
- **Columnas indicadoras de nulos iniciales**
- **Definición de funciones para extraer expresiones regulares**
- **Creación de columnas descriptivas con los caracteres limpios**
- **Análisis de las relaciones de las columnas de precios, superficies y precios por m^2**
- **Detección y exclusión de valores fuera de serie (outliers)**
- **Corrección de precios (imputación)**
- **Búsqueda de expresiones regulares**

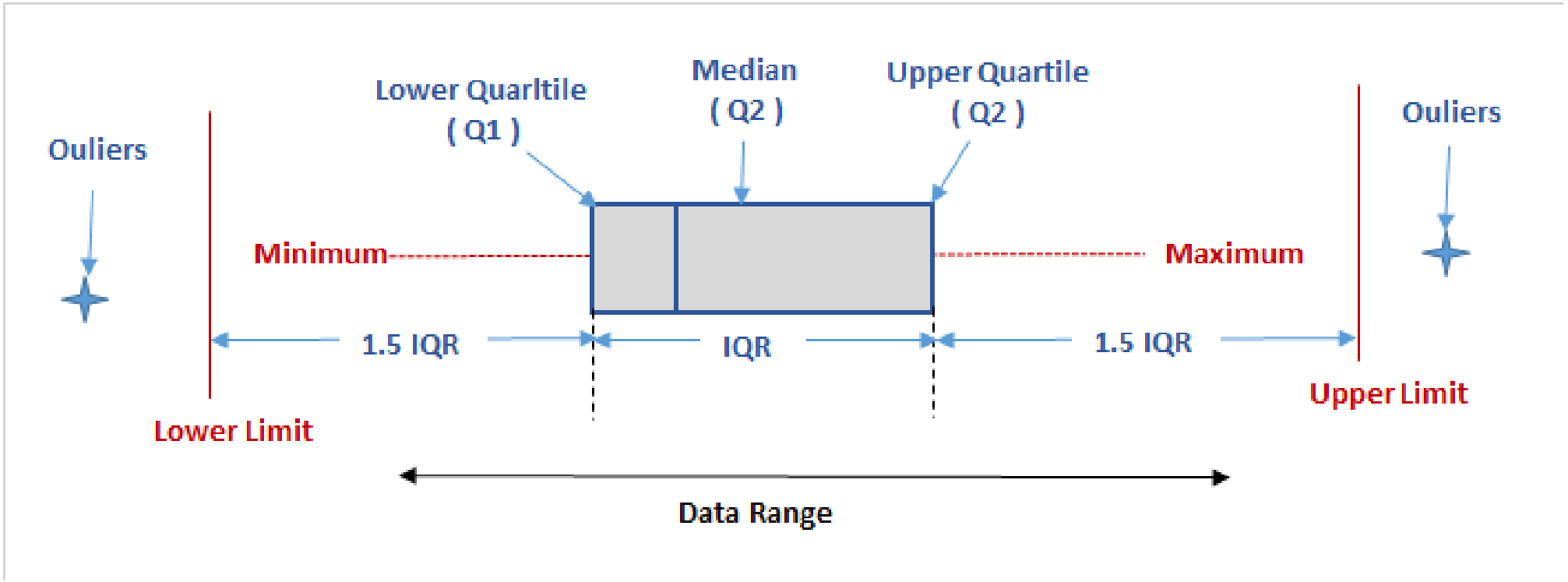
Campos de precio

- price . currency = price_aprox_local_currency
- price . currency = price_aprox_usd
- price_usd_per_m2 = price_aprox_usd / surface_total_in_m2
- price_per_m2 = price / surface_covered_in_m2

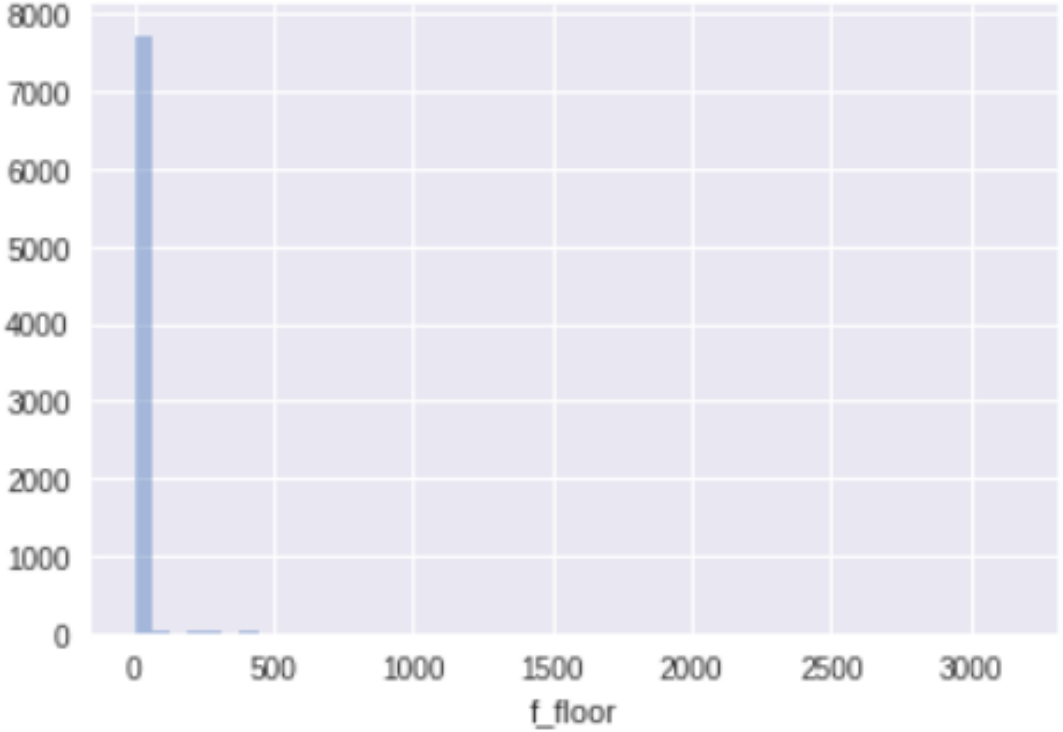
Limpieza de caracteres

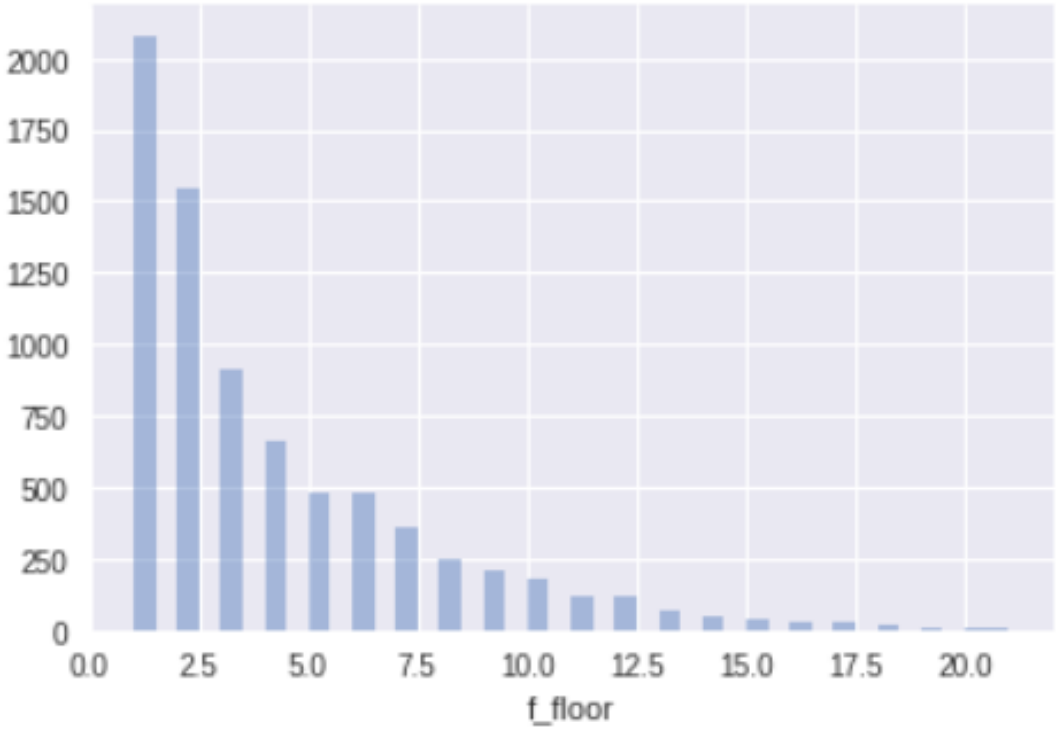
- **Todas a minúsculas**
- **Separador por símbolos**
- **Separador por caracteres**

Outliers

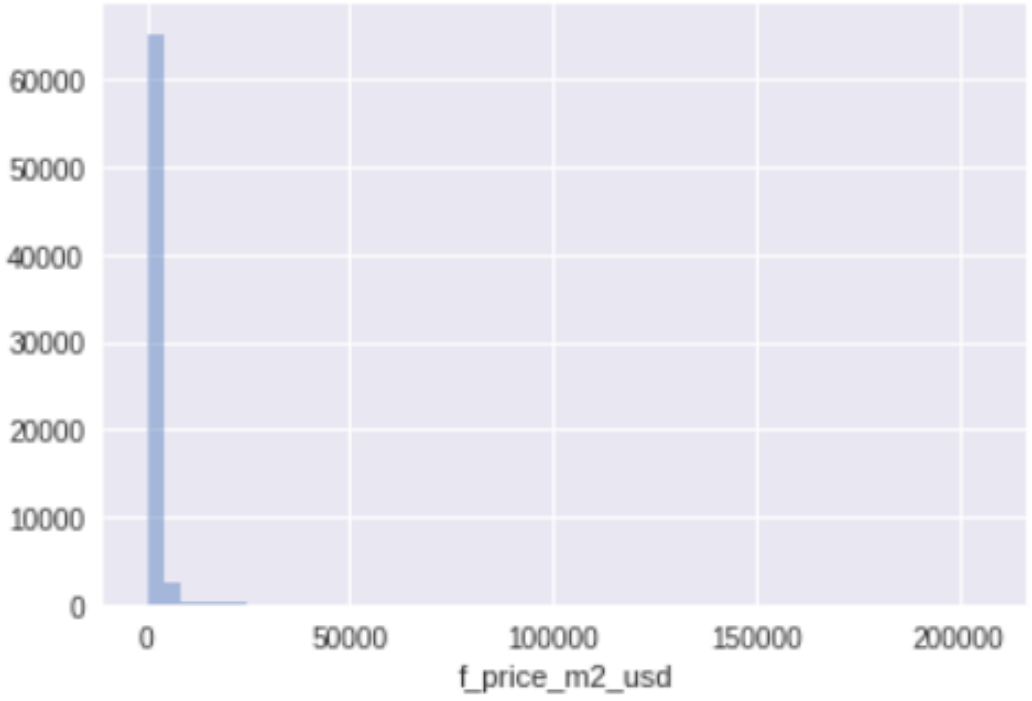

Ouliers
Lower Quarltile
(Q1)
Median
(Q2)
Upper Quartile
(Q2)
Ouliers
Minimum
Maximum
1.5 IQR
IQR
1.5 IQR
Lower Limit
Upper Limit
Data Range

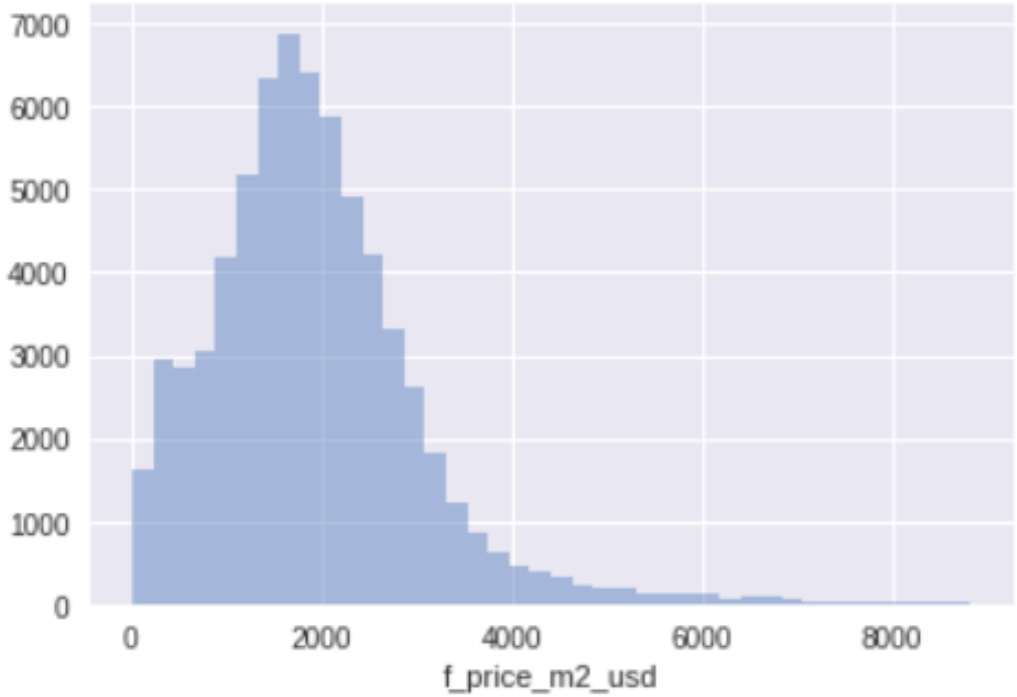
Outliers


8000
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
3000
f_floor

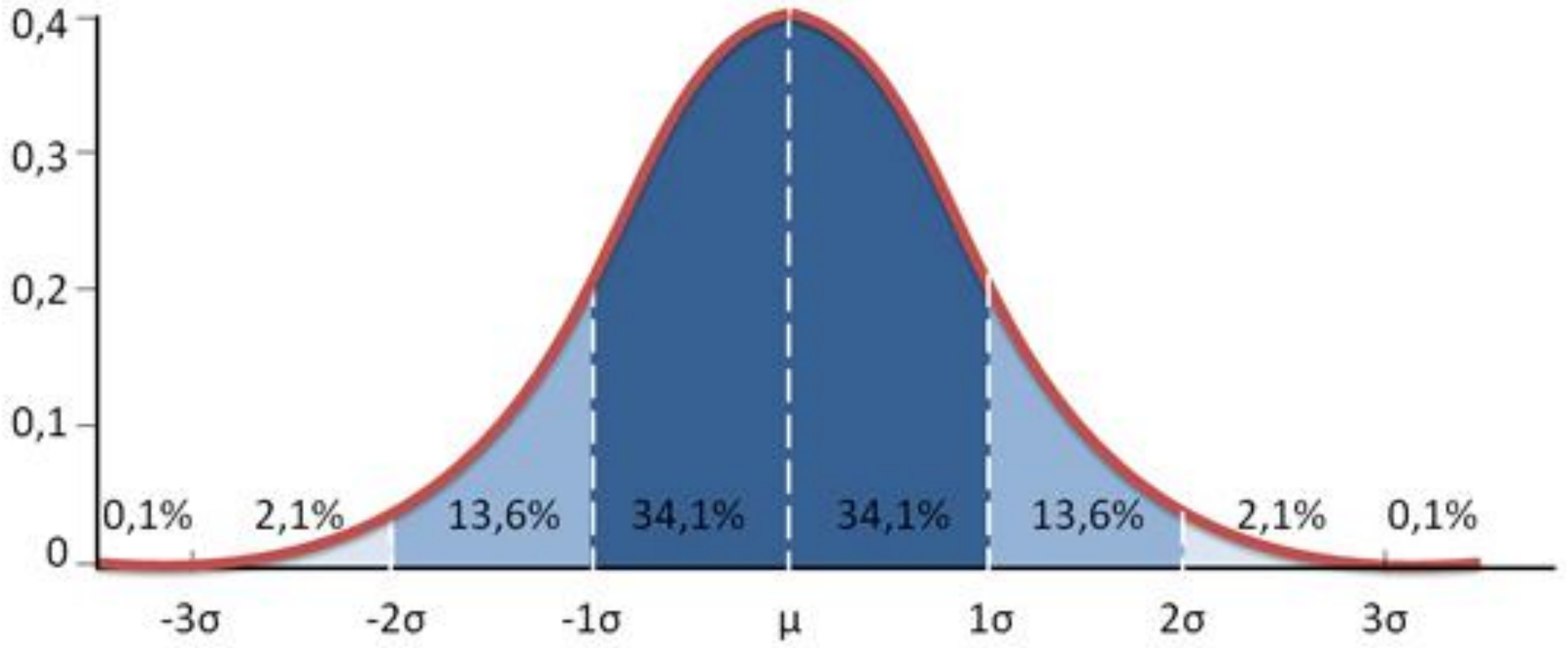

2000
1750
1500
1250
1000
750
500
250
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
f_floor

Outliers


60000
50000
40000
30000
20000
10000
0
0
50000
100000
150000
200000
f_price_m2_usd


7000
6000
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
f_price_m2_usd

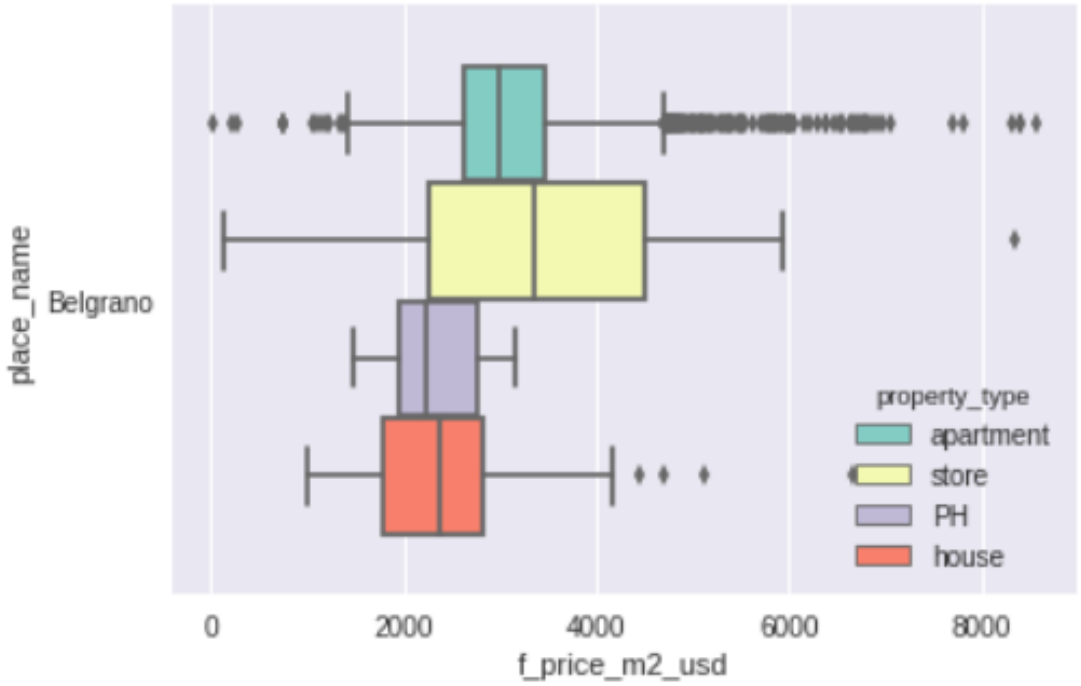
Imputación de valores

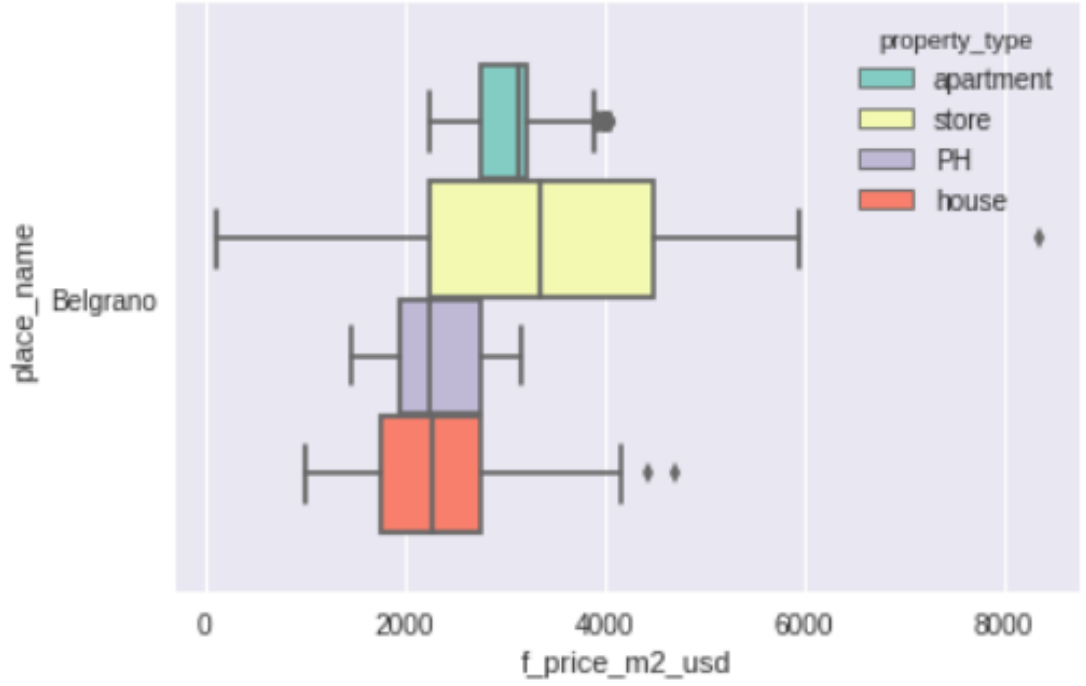


Imputación de valores (precio/m^2)

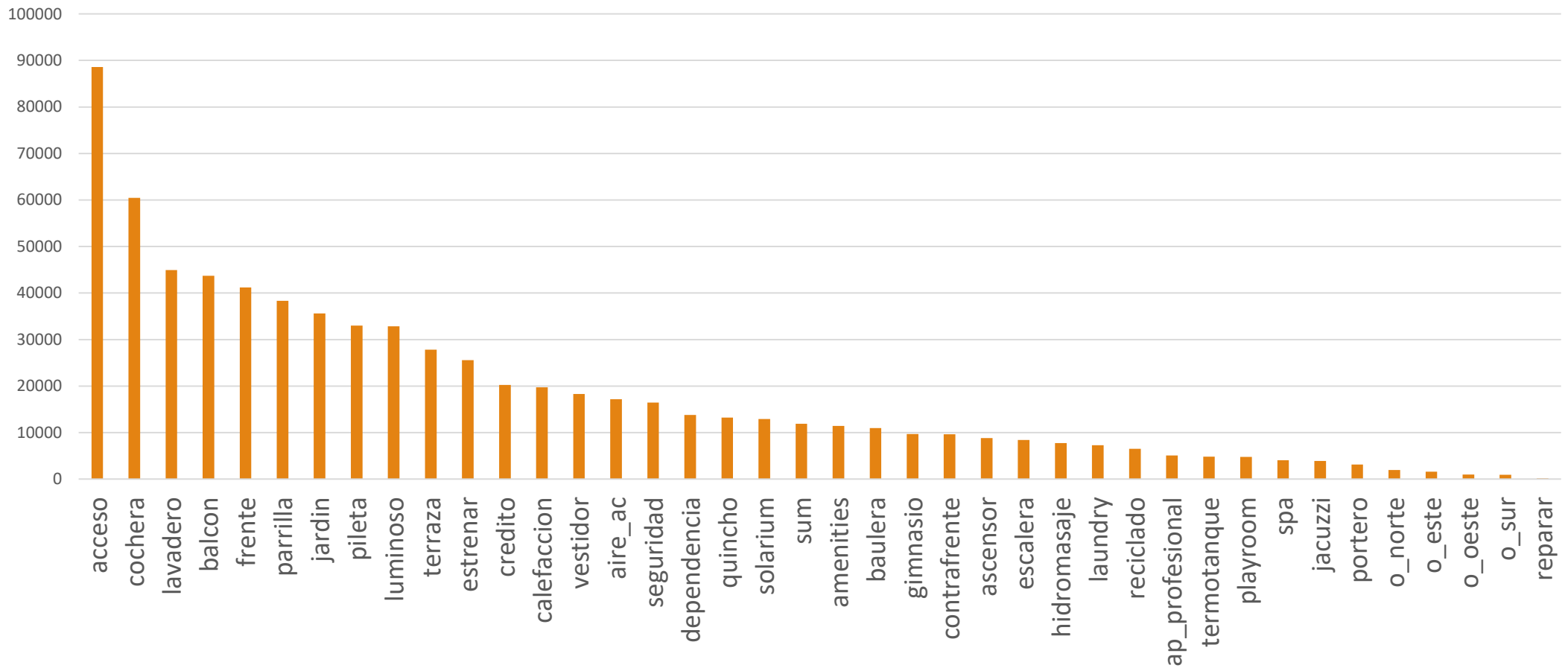
- **Por localidades/barrios**
  - **N limite**
  - **Z limite**
  - **Media del barrio**
- **Por prov/distritos**
  - **N limite**
  - **Z limite**
  - **Media de la provincia**

Imputación de valores





Dummys

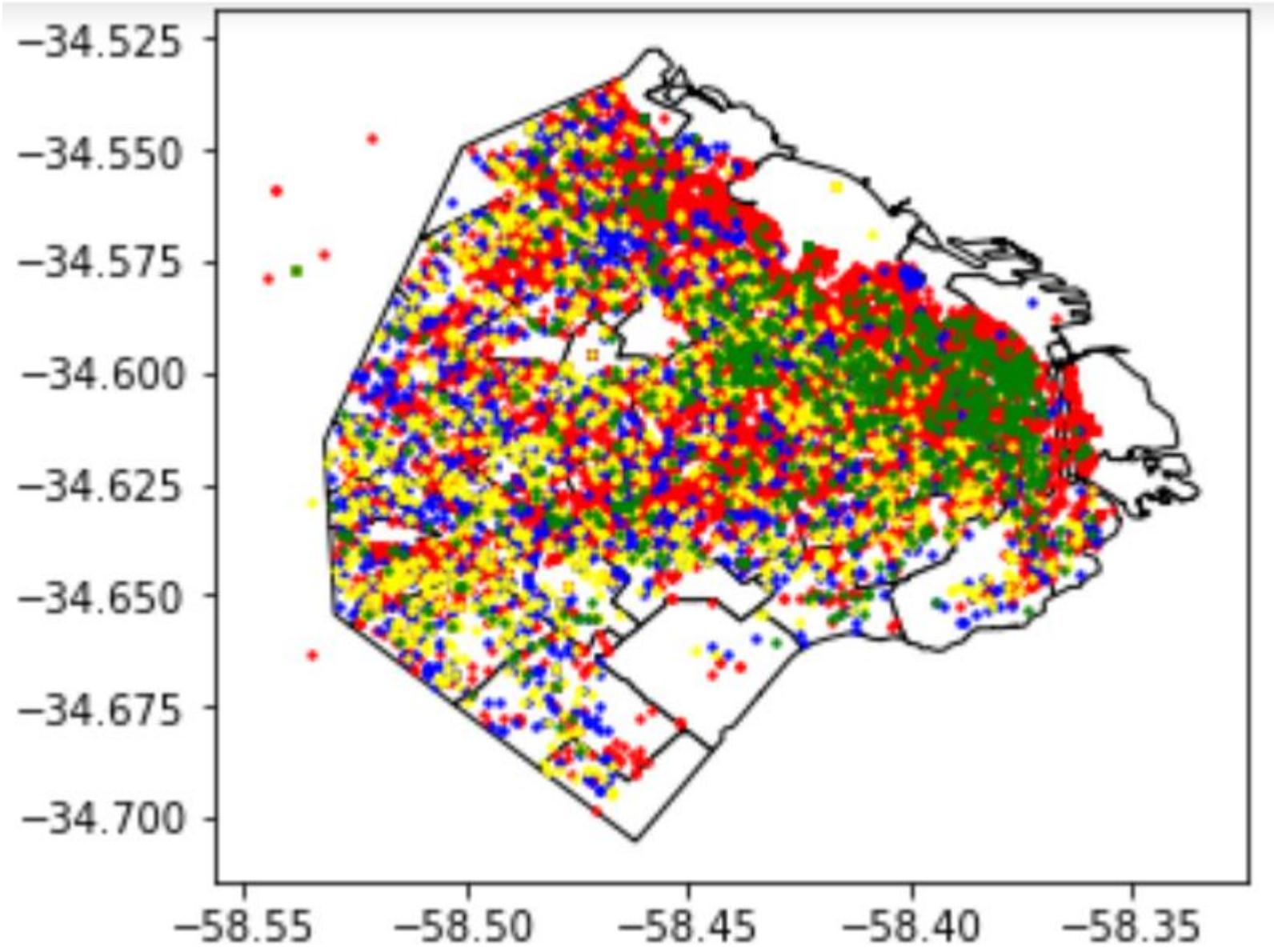

100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
acceso
cochera
lavadero
balcon
frente
parrilla
jardin
pileta
luminoso
terraza
estrenar
credito
calefaccion
vestidor
aire_ac
seguridad
dependencia
quincho
solarium
sum
amenities
baulera
gimnasio
contrafrente
ascensor
escalera
hidromasaje
laundry
reciclado
ap_profesional
termotanque
playroom
spa
jacuzzi
portero
o_norte
o_este
o_oeste
o_sur
reparar

Expresiones regulares

- **Campos descriptivos**
- **Cantidad de ambientes**
- **Diccionario de reemplazo de valores**

Resultados

- **Campos filtrados sin outliers**
- **Imputación de valores**
- **Relleno de faltantes según descripciones**
- **Recalculo de campo precio/m^2**
- **Ubicación de datos georreferenciados**


−34.525
−34.550
−34.575
−34.600
−34.625
−34.650
−34.675
−34.700
−58.55
−58.50
−58.45
−58.40
−58.35

MUCHAS GRACIAS