1) Data wrangling

August 8, 2023

1 1) Preparación previa

1.0.1 Carga de librerías

[1]:
```
import pandas as pd
import numpy as np
import re
```

1.0.2 Lectura del dataset original de Properati

[2]:
```
data = pd.read_csv("https://media.githubusercontent.com/media/Agustin-Bulzomi/
↪Projects/main/Programming/Digital%20House%20(Python)/Support%20Files/
↪Project%201/Properati.csv", index_col=0)
data.head(5)
```

[2]:
```
  operation property_type place_name  \
0      sell            PH  Mataderos
1      sell     apartment   La Plata
2      sell     apartment  Mataderos
3      sell            PH    Liniers
4      sell     apartment     Centro

                           place_with_parent_names country_name  \
0             |Argentina|Capital Federal|Mataderos|    Argentina
1       |Argentina|Bs.As. G.B.A. Zona Sur|La Plata|    Argentina
2             |Argentina|Capital Federal|Mataderos|    Argentina
3               |Argentina|Capital Federal|Liniers|    Argentina
4  |Argentina|Buenos Aires Costa Atlántica|Mar de…    Argentina

                    state_name  geonames_id                  lat-lon  \
0              Capital Federal    3430787.0  -34.6618237,-58.5088387
1       Bs.As. G.B.A. Zona Sur    3432039.0  -34.9038831,-57.9643295
2              Capital Federal    3430787.0  -34.6522615,-58.5229825
3              Capital Federal    3431333.0  -34.6477969,-58.5164244
4  Buenos Aires Costa Atlántica    3435548.0  -38.0026256,-57.5494468

         lat        lon  …  surface_covered_in_m2 price_usd_per_m2  \
```

```
0 -34.661824 -58.508839  …                   40.0      1127.272727
1 -34.903883 -57.964330  …                    NaN              NaN
2 -34.652262 -58.522982  …                   55.0      1309.090909
3 -34.647797 -58.516424  …                    NaN              NaN
4 -38.002626 -57.549447  …                   35.0      1828.571429

   price_per_m2  floor  rooms  expenses  \
0   1550.000000    NaN    NaN       NaN
1           NaN    NaN    NaN       NaN
2   1309.090909    NaN    NaN       NaN
3           NaN    NaN    NaN       NaN
4   1828.571429    NaN    NaN       NaN

                                     properati_url  \
0  http://www.properati.com.ar/15bo8_venta_ph_mat…
1  http://www.properati.com.ar/15bob_venta_depart…
2  http://www.properati.com.ar/15bod_venta_depart…
3  http://www.properati.com.ar/15boh_venta_ph_lin…
4  http://www.properati.com.ar/15bok_venta_depart…

                                       description  \
0  2 AMBIENTES TIPO CASA PLANTA BAJA POR PASILLO,…
1  Venta de departamento en décimo piso al frente…
2  2 AMBIENTES 3ER PISO LATERAL LIVING COMEDOR AM…
3  PH 3 ambientes con patio. Hay 3 deptos en lote…
4  DEPARTAMENTO CON FANTÁSTICA ILUMINACIÓN NATURA…

                                              title  \
0                 2 AMB TIPO CASA SIN EXPENSAS EN PB
1  VENTA Depto 2 dorm. a estrenar 7 e/ 36 y 37   …
2           2 AMB 3ER PISO CON ASCENSOR APTO CREDITO
3                       PH 3 amb. cfte. reciclado
4  DEPTO 2 AMB AL CONTRAFRENTE ZONA CENTRO/PLAZA …

                                    image_thumbnail
0  https://thumbs4.properati.com/8/BluUYiHJLhgIIK…
1  https://thumbs4.properati.com/7/ikpVBu2ztHA7jv…
2  https://thumbs4.properati.com/5/SXKr34F_IwG3W_…
3  https://thumbs4.properati.com/3/DgIfX-85Mog5SP…
4  https://thumbs4.properati.com/5/xrRqlNcSI_vs-f…

[5 rows x 25 columns]
```

1.0.3 Análisis de nulos según columnas

[3]:
```
nulos = data.isnull().sum()
```

[4]:
```
nulos_porcentaje = nulos / data.shape[0] * 100
nulos_porcentaje
```

[4]:
```
operation                     0.000000
property_type                 0.000000
place_name                    0.018974
place_with_parent_names       0.000000
country_name                  0.000000
state_name                    0.000000
geonames_id                  15.440521
lat-lon                      42.525986
lat                          42.525986
lon                          42.525986
price                        16.837156
currency                     16.837981
price_aprox_local_currency   16.837156
price_aprox_usd              16.837156
surface_total_in_m2          32.443491
surface_covered_in_m2        16.422208
price_usd_per_m2             43.394654
price_per_m2                 27.686850
floor                        93.483749
rooms                        60.905791
expenses                     88.234615
properati_url                 0.000000
description                   0.001650
title                         0.000000
image_thumbnail               2.567233
dtype: float64
```

En base a los resultados se llegó a las siguientes conclusiones:

1) Se tomará la superficie cubierta en vez de la total. En el anterior desafío se creó una función para tomar una mezcla de ambas ignorando las inconsistencias pero daba más nulos aún.

2) Se tomará el precio aprox en dólares en vez del precio per m2 en dólares. Al mismo se dividirá por la superficie cubierta para tener el valor por m2

3) "rooms" tiene pocos valores, se tendrá que imputar los datos faltantes

4) "description" y "title" servirán para obtener información extra

1.0.4 Separación de columna con muchas ubicaciones

[5]:
```
# La columna "place_with_parent_names" tiene información separada con '|'. Se␣
 ↪separa para obtener info adicional
separar_zona = data["place_with_parent_names"].str.split('|', expand = True)
separar_zona.columns = ['??', 'Pais', 'Zona', 'Partido', 'Barrios', 'Country',␣
 ↪'Otra']
```

1.0.5 Agregado de la nueva información en nuevas columnas

[6]:
```
data_concat = pd.concat([data, separar_zona], axis=1)
data_concat.head(10)
```

[6]:

	operation	property_type	place_name	\
0	sell	PH	Mataderos	
1	sell	apartment	La Plata	
2	sell	apartment	Mataderos	
3	sell	PH	Liniers	
4	sell	apartment	Centro	
5	sell	house	Gualeguaychú	
6	sell	PH	Munro	
7	sell	apartment	Belgrano	
8	sell	apartment	Belgrano	
9	sell	house	Rosario	

	place_with_parent_names	country_name	\
0	|Argentina|Capital Federal|Mataderos|	Argentina	
1	|Argentina|Bs.As. G.B.A. Zona Sur|La Plata|	Argentina	
2	|Argentina|Capital Federal|Mataderos|	Argentina	
3	|Argentina|Capital Federal|Liniers|	Argentina	
4	|Argentina|Buenos Aires Costa Atlántica|Mar de…	Argentina	
5	|Argentina|Entre Ríos|Gualeguaychú|	Argentina	
6	|Argentina|Bs.As. G.B.A. Zona Norte|Vicente Ló…	Argentina	
7	|Argentina|Capital Federal|Belgrano|	Argentina	
8	|Argentina|Capital Federal|Belgrano|	Argentina	
9	|Argentina|Santa Fe|Rosario|	Argentina	

	state_name	geonames_id	lat-lon	\
0	Capital Federal	3430787.0	-34.6618237,-58.5088387	
1	Bs.As. G.B.A. Zona Sur	3432039.0	-34.9038831,-57.9643295	
2	Capital Federal	3430787.0	-34.6522615,-58.5229825	
3	Capital Federal	3431333.0	-34.6477969,-58.5164244	
4	Buenos Aires Costa Atlántica	3435548.0	-38.0026256,-57.5494468	
5	Entre Ríos	3433657.0	-33.0140714,-58.519828	
6	Bs.As. G.B.A. Zona Norte	3430511.0	-34.5329567,-58.5217825	
7	Capital Federal	3436077.0	-34.5598729,-58.443362	
8	Capital Federal	3436077.0	-34.5598729,-58.443362	

```
9                     Santa Fe    3838574.0   -32.942031,-60.7259192

         lat        lon  …  \
0 -34.661824 -58.508839  …
1 -34.903883 -57.964330  …
2 -34.652262 -58.522982  …
3 -34.647797 -58.516424  …
4 -38.002626 -57.549447  …
5 -33.014071 -58.519828  …
6 -34.532957 -58.521782  …
7 -34.559873 -58.443362  …
8 -34.559873 -58.443362  …
9 -32.942031 -60.725919  …

                                          description  \
0  2 AMBIENTES TIPO CASA PLANTA BAJA POR PASILLO,…
1  Venta de departamento en décimo piso al frente…
2  2 AMBIENTES 3ER PISO LATERAL LIVING COMEDOR AM…
3  PH 3 ambientes con patio. Hay 3 deptos en lote…
4  DEPARTAMENTO CON FANTÁSTICA ILUMINACIÓN NATURA…
5  Casa en el perímetro del barrio 338, ubicada e…
6  MUY BUEN PH AL FRENTE CON ENTRADA INDEPENDIENT…
7  EXCELENTE MONOAMBIENTE A ESTRENAR AMPLIO SUPER…
8  EXCELENTE DOS AMBIENTES  ESTRENAR AMPLIO SUPER…
9  MEDNOZA AL 7600A UNA CUADRA DE CALLE MENDOZAWH…

                                              title  \
0                  2 AMB TIPO CASA SIN EXPENSAS EN PB
1  VENTA Depto 2 dorm. a estrenar 7 e/ 36 y 37   …
2           2 AMB 3ER PISO CON ASCENSOR APTO CREDITO
3                       PH 3 amb. cfte. reciclado
4  DEPTO 2 AMB AL CONTRAFRENTE ZONA CENTRO/PLAZA …
5  Casa Barrio 338. Sobre calle 3 de caballería, …
6  MUY BUEN PH AL FRENTE DOS DORMITORIOS , PATIO,…
7  JOSE HERNANDEZ 1400 MONOAMBIENTE  ESTRENAR CAT…
8  JOSE HERNANDEZ   1400 DOS AMBIENTES ESTRENAR ,…
9         WHITE 7637 - 2 DORMITORIOS CON PATIO

                                    image_thumbnail  ??       Pais  \
0  https://thumbs4.properati.com/8/BluUYiHJLhgIIK…      Argentina
1  https://thumbs4.properati.com/7/ikpVBu2ztHA7jv…      Argentina
2  https://thumbs4.properati.com/5/SXKr34F_IwG3W_…      Argentina
3  https://thumbs4.properati.com/3/DgIfX-85Mog5SP…      Argentina
4  https://thumbs4.properati.com/5/xrRqlNcSI_vs-f…      Argentina
5  https://thumbs4.properati.com/6/q-w68gvaUEQVXI…      Argentina
6  https://thumbs4.properati.com/5/6GOXsHCyDu1aGx…      Argentina
7  https://thumbs4.properati.com/1/IHxARynlr8sPEW…      Argentina
```

```
8  https://thumbs4.properati.com/2/J3zOjgaFHrkvnv…      Argentina
9  https://thumbs4.properati.com/8/RCf1YEWdF4rv98…      Argentina

                          Zona        Partido  Barrios  Country  Otra
0               Capital Federal      Mataderos              None  None
1        Bs.As. G.B.A. Zona Sur       La Plata              None  None
2               Capital Federal      Mataderos              None  None
3               Capital Federal        Liniers              None  None
4  Buenos Aires Costa Atlántica  Mar del Plata   Centro           None
5                    Entre Ríos   Gualeguaychú              None  None
6      Bs.As. G.B.A. Zona Norte  Vicente López    Munro           None
7               Capital Federal       Belgrano              None  None
8               Capital Federal       Belgrano              None  None
9                     Santa Fe        Rosario              None  None

[10 rows x 32 columns]
```

1.0.6 Creación de la columna de precios por metros cuadrados usando la superficie cubierta

[7]:
```
data_concat['precio_usd_por_m2'] = data_concat.price_aprox_usd/data_concat.
 ↪surface_covered_in_m2
```

2 2) Imputación

Debido a la insuficiente cantidad de datos de ambientes, vamos a intentar obtener más

2.0.1 1) Imputación en base a título y descripción

Descripción

[8]:
```
patron_amb = "(?P<ambiente>\d\s)((A|a)(M|m)(B|b))"
regex_amb = re.compile(patron_amb)

data_amb_serie = data_concat["description"]
data_amb_match = data_amb_serie.apply(lambda x: x if x is np.NaN else regex_amb.
 ↪search(x))

mask_amb_notnull = data_amb_match.notnull()

data_ambientes = data_amb_match[mask_amb_notnull].apply(lambda x: x.
 ↪group("ambiente"))

data_concat.loc[mask_amb_notnull, 'ambientes_desc'] = \
    data_amb_match[mask_amb_notnull].apply(lambda x: x.group('ambiente'))
```

[9]:
```
data_concat.loc[mask_amb_notnull, ["description", "ambientes_desc"]]
```

[9]:

	description	ambientes_desc
0	2 AMBIENTES TIPO CASA PLANTA BAJA POR PASILLO,…	2
2	2 AMBIENTES 3ER PISO LATERAL LIVING COMEDOR AM…	2
3	PH 3 ambientes con patio. Hay 3 deptos en lote…	3
11	Entrada de Coche, Jardin, Living en desnivel, …	1
12	EXCELENTE DEPARTAMENTO 2 AMBIENTES CONTRAFRENT…	2
…	…	…
121127	Amplio departamento de 2 Amb. con vestidor y t…	2
121131	Excelente semipiso al frente de 4 ambientes. E…	4
121138	VENTA CON RENTA DEPARTAMENTO DE 2 AMBIENTES A …	2
121139	Condominio Royal Residence, espectacular empre…	3
121218	2 Amb al contrafrente, luminoso. El departame…	2

[37540 rows x 2 columns]

Título

[10]:
```
patron_amb2 = "(?P<ambiente_title>\d\s)((A|a)(M|m)(B|b))"
regex_amb2 = re.compile(patron_amb2)

data_amb_serie2 = data_concat["title"]
data_amb_match2 = data_amb_serie2.apply(lambda x: x if x is np.NaN else␣
 ↪regex_amb2.search(x))

mask_amb_notnull2 = data_amb_match2.notnull()

data_ambientes2 = data_amb_match2[mask_amb_notnull2].apply(lambda x: x.
 ↪group("ambiente_title"))

data_concat.loc[mask_amb_notnull2, 'ambientes_t'] = \
    data_amb_match2[mask_amb_notnull2].apply(lambda x: x.
 ↪group('ambiente_title'))
```

[11]:
```
data_concat.loc[mask_amb_notnull2, ["ambientes_desc", "ambientes_t"]]
```

[11]:

	ambientes_desc	ambientes_t
0	2	2
2	2	2
3	3	3
4	NaN	2
12	2	2
…	…	…
121131	4	3
121138	2	2
121153	NaN	3
121215	NaN	3
121218	2	2

```
[21092 rows x 2 columns]
```

Unificación de la nueva información Se crea una función para resumir ambas columnas en una nueva

[12]:
```
def limpieza_amb(ambientes_desc, ambientes_t):
    if pd.isnull(ambientes_desc) and pd.isnull(ambientes_t):
        ambientes = np.NaN
    elif pd.isnull(ambientes_desc):
        ambientes = ambientes_t
    else:
        ambientes = ambientes_desc
    return ambientes
```

[13]:
```
# Se aplica la función
data_concat["ambientes"] = data_concat.apply(lambda data_concat:␣
 ↪limpieza_amb(data_concat['ambientes_desc'],data_concat['ambientes_t']),axis=1)
data_concat.head(15)
```

[13]:

	operation	property_type	place_name	\
0	sell	PH	Mataderos	
1	sell	apartment	La Plata	
2	sell	apartment	Mataderos	
3	sell	PH	Liniers	
4	sell	apartment	Centro	
5	sell	house	Gualeguaychú	
6	sell	PH	Munro	
7	sell	apartment	Belgrano	
8	sell	apartment	Belgrano	
9	sell	house	Rosario	
10	sell	house	Córdoba	
11	sell	house	San Miguel	
12	sell	apartment	Martínez	
13	sell	apartment	Palermo Soho	
14	sell	apartment	Palermo Soho	

	place_with_parent_names	country_name	\
0	|Argentina|Capital Federal|Mataderos|	Argentina	
1	|Argentina|Bs.As. G.B.A. Zona Sur|La Plata|	Argentina	
2	|Argentina|Capital Federal|Mataderos|	Argentina	
3	|Argentina|Capital Federal|Liniers|	Argentina	
4	|Argentina|Buenos Aires Costa Atlántica|Mar de…	Argentina	
5	|Argentina|Entre Ríos|Gualeguaychú|	Argentina	
6	|Argentina|Bs.As. G.B.A. Zona Norte|Vicente Ló…	Argentina	
7	|Argentina|Capital Federal|Belgrano|	Argentina	
8	|Argentina|Capital Federal|Belgrano|	Argentina	
9	|Argentina|Santa Fe|Rosario|	Argentina	

```
10                        |Argentina|Córdoba|Córdoba|    Argentina
11    |Argentina|Bs.As. G.B.A. Zona Norte|San Miguel|    Argentina
12  |Argentina|Bs.As. G.B.A. Zona Norte|San Isidro…    Argentina
13   |Argentina|Capital Federal|Palermo|Palermo Soho|    Argentina
14   |Argentina|Capital Federal|Palermo|Palermo Soho|    Argentina

                     state_name  geonames_id                  lat-lon  \
0                Capital Federal    3430787.0  -34.6618237,-58.5088387
1         Bs.As. G.B.A. Zona Sur    3432039.0  -34.9038831,-57.9643295
2                Capital Federal    3430787.0  -34.6522615,-58.5229825
3                Capital Federal    3431333.0  -34.6477969,-58.5164244
4   Buenos Aires Costa Atlántica    3435548.0  -38.0026256,-57.5494468
5                     Entre Ríos    3433657.0   -33.0140714,-58.519828
6       Bs.As. G.B.A. Zona Norte    3430511.0  -34.5329567,-58.5217825
7                Capital Federal    3436077.0   -34.5598729,-58.443362
8                Capital Federal    3436077.0   -34.5598729,-58.443362
9                      Santa Fe    3838574.0   -32.942031,-60.7259192
10                      Córdoba    3860259.0  -31.4200833,-64.1887761
11      Bs.As. G.B.A. Zona Norte          NaN  -34.5390571,-58.7196093
12      Bs.As. G.B.A. Zona Norte    3430813.0  -34.4860195,-58.5038139
13               Capital Federal    3430234.0                      NaN
14               Capital Federal    3430234.0                      NaN

          lat        lon  …       Pais                         Zona  \
0  -34.661824 -58.508839  …  Argentina               Capital Federal
1  -34.903883 -57.964330  …  Argentina        Bs.As. G.B.A. Zona Sur
2  -34.652262 -58.522982  …  Argentina               Capital Federal
3  -34.647797 -58.516424  …  Argentina               Capital Federal
4  -38.002626 -57.549447  …  Argentina  Buenos Aires Costa Atlántica
5  -33.014071 -58.519828  …  Argentina                     Entre Ríos
6  -34.532957 -58.521782  …  Argentina      Bs.As. G.B.A. Zona Norte
7  -34.559873 -58.443362  …  Argentina               Capital Federal
8  -34.559873 -58.443362  …  Argentina               Capital Federal
9  -32.942031 -60.725919  …  Argentina                      Santa Fe
10 -31.420083 -64.188776  …  Argentina                       Córdoba
11 -34.539057 -58.719609  …  Argentina      Bs.As. G.B.A. Zona Norte
12 -34.486019 -58.503814  …  Argentina      Bs.As. G.B.A. Zona Norte
13        NaN        NaN  …  Argentina               Capital Federal
14        NaN        NaN  …  Argentina               Capital Federal

          Partido       Barrios  Country  Otra  precio_usd_por_m2  \
0       Mataderos                  None  None        1550.000000
1        La Plata                  None  None                NaN
2       Mataderos                  None  None        1309.090909
3         Liniers                  None  None                NaN
4   Mar del Plata        Centro          None        1828.571429
5    Gualeguaychú                  None  None                NaN
```

```
6   Vicente López         Munro           None        1666.666667
7        Belgrano                  None  None        3450.000000
8        Belgrano                  None  None        3250.000000
9         Rosario                  None  None                NaN
10        Córdoba                  None  None                NaN
11     San Miguel                  None  None                NaN
12     San Isidro      Martínez           None        3194.444444
13        Palermo  Palermo Soho           None        3723.333333
14        Palermo  Palermo Soho           None        4770.967742

    ambientes_desc  ambientes_t  ambientes
0                2            2          2
1              NaN          NaN        NaN
2                2            2          2
3                3            3          3
4              NaN            2          2
5              NaN          NaN        NaN
6              NaN          NaN        NaN
7              NaN          NaN        NaN
8              NaN          NaN        NaN
9              NaN          NaN        NaN
10             NaN          NaN        NaN
11               1          NaN          1
12               2            2          2
13               3          NaN          3
14             NaN          NaN        NaN

[15 rows x 36 columns]
```

[14]:
```
data_concat.shape
```

[14]:
```
(121220, 36)
```

[15]:
```
data_concat.ambientes.notnull().sum() / data_concat.shape[0] * 100
```

[15]:
```
36.636693614915025
```

Se crea una función para resumir la información entre la nueva columna y rooms.

En la enorme mayoría de los casos en donde se tenía el dato de rooms original, la cantidad de ambientes obtenida por imputación concordaba con el valor de rooms original. Esto indica que ambos términos son intercambiables al menos en este dataset.

[16]:
```
def limpieza_amb2(rooms, ambientes):
    if pd.isnull(rooms) and pd.isnull(ambientes):
        ambientes_train = 0
    elif pd.isnull(rooms):
        ambientes_train = ambientes
```

```
    else:
        ambientes_train = int(rooms)
    return int(ambientes_train)
```

[17]:
```
# Se aplica la función. Se llama a la nueva variable "train" pues es la que␣
 ↪será usada para entrenar al modelo
data_concat["ambientes_train"] = data_concat.apply(lambda x:␣
 ↪limpieza_amb2(x['rooms'],x['ambientes']),axis=1)
data_concat.ambientes_train.value_counts()
```

[17]:
```
0     52382
3     19603
2     19404
4     12235
1      8583
5      5110
6      1813
7       921
8       441
10      226
9       200
11       72
12       65
13       33
14       27
15       26
17       16
16       11
20       11
22        8
18        6
21        5
19        4
30        4
25        4
32        3
23        2
24        1
31        1
29        1
27        1
28        1
Name: ambientes_train, dtype: int64
```

2.0.2 2) Imputación en base a la superficie

Tomando la mediana de las superficies agrupadas según ambientes definir un punto medio entre cada mediana.

La misma nos permitiría definir un divisor que delimite cuándo una superficie es más probable que pertenezca a una cantidad de ambientes. Al ser una imputación no tan certera, se dejará afuera de la serie "train".

[18]:
```
# Se calcula cuánta información nueva podría obtenerse
superficie_not_null = data_concat['surface_covered_in_m2'].notnull()
ambientes_zero = data_concat['ambientes_train'] == 0
filtro = superficie_not_null & ambientes_zero
print(filtro.sum())
```

```
41269
```

2.0.3 Divisores de ambientes

[19]:
```
amb_1 = data_concat['ambientes_train'] == 1
amb_2 = data_concat['ambientes_train'] == 2
amb_3 = data_concat['ambientes_train'] == 3
amb_4 = data_concat['ambientes_train'] == 4
amb_5 = data_concat['ambientes_train'] == 5
amb_6 = data_concat['ambientes_train'] == 6
amb_7 = data_concat['ambientes_train'] == 7

divisor1 = (data_concat[amb_1].surface_covered_in_m2.median() +␣
 ↪data_concat[amb_2].surface_covered_in_m2.median())/2
divisor2 = (data_concat[amb_2].surface_covered_in_m2.median() +␣
 ↪data_concat[amb_3].surface_covered_in_m2.median())/2
divisor3 = (data_concat[amb_3].surface_covered_in_m2.median() +␣
 ↪data_concat[amb_4].surface_covered_in_m2.median())/2
divisor4 = (data_concat[amb_4].surface_covered_in_m2.median() +␣
 ↪data_concat[amb_5].surface_covered_in_m2.median())/2
divisor5 = (data_concat[amb_5].surface_covered_in_m2.median() +␣
 ↪data_concat[amb_6].surface_covered_in_m2.median())/2
divisor6 = (data_concat[amb_6].surface_covered_in_m2.median() +␣
 ↪data_concat[amb_7].surface_covered_in_m2.median())/2
divisor7 = (data_concat[amb_7].surface_covered_in_m2.median() +␣
 ↪data_concat[amb_7].surface_covered_in_m2.max())/2

# Como no hay de 8 ambientes, se utiliza el valor máximo de 7 ambientes como␣
 ↪tope para calcular el divisor 7
```

[20]:
```
# Se crea una función para asignar ambientes según los divisores

def asignar_ambientes_segun_superficie(surface_covered_in_m2):
    #if superficie.isnull():
    #    return 0
    #elif 0 < superficie <= divisor1:
    if 0 < surface_covered_in_m2 <= divisor1:
        return 1
```

```
    elif divisor1 < surface_covered_in_m2 <= divisor2:
        return 2
    elif divisor2 < surface_covered_in_m2 <= divisor3:
        return 3
    elif divisor3 < surface_covered_in_m2 <= divisor4:
        return 4
    elif divisor4 < surface_covered_in_m2 <= divisor5:
        return 5
    elif divisor5 < surface_covered_in_m2 <= divisor6:
        return 6
    elif divisor6 < surface_covered_in_m2 <= divisor7:
        return 7
    else:
        return np.NaN
```

[21]:
```
# Se aplica la función para crear una columna de ambientes imputados con␣
 ↪valores en las filas que no tienen valores de ambientes_train

data_concat["ambientes_imputados"] = data_concat.apply(lambda x:␣
 ↪asignar_ambientes_segun_superficie(x['surface_covered_in_m2']) if␣
 ↪int(x['ambientes_train']) == 0 else 0, axis=1)
```

[22]:
```
data_concat.ambientes_imputados.value_counts()
```

[22]:
```
0.0    68838
7.0     8169
2.0     7263
1.0     6462
4.0     6068
3.0     5920
5.0     5206
6.0     2005
Name: ambientes_imputados, dtype: int64
```

[23]:
```
# Se suman ambas columnas al ser excluyentes: ambientes_final no tiene 0, cada␣
 ↪fila tiene un valor original o imputado

data_concat["ambientes_final"] = data_concat["ambientes_train"] +␣
 ↪data_concat["ambientes_imputados"]
```

[24]:
```
data_concat
```

[24]:
```
       operation property_type       place_name  \
0           sell            PH        Mataderos
1           sell     apartment         La Plata
2           sell     apartment        Mataderos
3           sell            PH          Liniers
```

```
4           sell     apartment           Centro
…           …          …               …
121215      sell     apartment         Belgrano
121216      sell         house           Beccar
121217      sell     apartment    Villa Urquiza
121218      sell     apartment      Plaza Colón
121219      sell     apartment  Capital Federal

                                 place_with_parent_names country_name  \
0                    |Argentina|Capital Federal|Mataderos|    Argentina
1              |Argentina|Bs.As. G.B.A. Zona Sur|La Plata|    Argentina
2                    |Argentina|Capital Federal|Mataderos|    Argentina
3                      |Argentina|Capital Federal|Liniers|    Argentina
4       |Argentina|Buenos Aires Costa Atlántica|Mar de…    Argentina
…                                                    …          …
121215               |Argentina|Capital Federal|Belgrano|    Argentina
121216  |Argentina|Bs.As. G.B.A. Zona Norte|San Isidro…    Argentina
121217          |Argentina|Capital Federal|Villa Urquiza|    Argentina
121218  |Argentina|Buenos Aires Costa Atlántica|Mar de…    Argentina
121219                        |Argentina|Capital Federal|    Argentina

                         state_name  geonames_id  \
0                    Capital Federal    3430787.0
1              Bs.As. G.B.A. Zona Sur    3432039.0
2                    Capital Federal    3430787.0
3                    Capital Federal    3431333.0
4       Buenos Aires Costa Atlántica    3435548.0
…                               …          …
121215               Capital Federal    3436077.0
121216     Bs.As. G.B.A. Zona Norte    3436080.0
121217               Capital Federal    3433775.0
121218  Buenos Aires Costa Atlántica          NaN
121219               Capital Federal    3433955.0

                              lat-lon        lat        lon  …      Barrios  \
0             -34.6618237,-58.5088387 -34.661824 -58.508839  …
1             -34.9038831,-57.9643295 -34.903883 -57.964330  …
2             -34.6522615,-58.5229825 -34.652262 -58.522982  …
3             -34.6477969,-58.5164244 -34.647797 -58.516424  …
4             -38.0026256,-57.5494468 -38.002626 -57.549447  …       Centro
…                                …        …        …  …          …
121215                          NaN        NaN        NaN  …
121216                          NaN        NaN        NaN  …       Beccar
121217  -34.5706388726,-58.4755963355 -34.570639 -58.475596  …
121218                          NaN        NaN        NaN  …  Plaza Colón
121219                          NaN        NaN        NaN  …         None
```

```
        Country  Otra  precio_usd_por_m2  ambientes_desc  ambientes_t  \
0          None  None        1550.000000               2            2
1          None  None                NaN             NaN          NaN
2          None  None        1309.090909               2            2
3          None  None                NaN               3            3
4                None        1828.571429             NaN            2
…         …    …                     …               …            …
121215     None  None        9354.838710             NaN            3
121216           None        1383.333333             NaN          NaN
121217     None  None        3371.794872             NaN          NaN
121218           None        1997.916667               2            2
121219     None  None        1675.324675             NaN          NaN

         ambientes  ambientes_train  ambientes_imputados  ambientes_final
0                2                2                  0.0              2.0
1              NaN                0                  NaN              NaN
2                2                2                  0.0              2.0
3                3                3                  0.0              3.0
4                2                2                  0.0              2.0
…              …                …                    …              …
121215           3                3                  0.0              3.0
121216         NaN                0                  7.0              7.0
121217         NaN                0                  1.0              1.0
121218           2                2                  0.0              2.0
121219         NaN                0                  3.0              3.0

[121220 rows x 39 columns]
```

3 3) Búsqueda de amenities

Se analiza la descripción de cada fila para encontrar palabras clave que indiquen amenities con valor agregado

[25]:
```
patron_balcon = "(?P<balcon>(B|b)(A|a)(L|l)(C|c)(O|n)(N|n))"
regex_balcon = re.compile(patron_balcon)
data_balcon = data_concat["description"]
data_match_balcon = data_balcon.apply(lambda x: x if x is np.NaN else␣
 ↪regex_balcon.search(x))
mask_notnull_balcon = data_match_balcon.notnull()
data_balcon = data_match_balcon[mask_notnull_balcon].apply(lambda x: x.
 ↪group("balcon"))
data_concat.loc[mask_notnull_balcon, 'balcon'] = \
data_match_balcon[mask_notnull_balcon].apply(lambda x: x.group('balcon').
 ↪lower())

patron_parrilla = "(?P<parrilla>(P|p)(A|a)(R|r)(R|r)(I|i)(L|l)(L|l)(A|a))"
regex_parrilla = re.compile(patron_parrilla)
```

```
data_parrilla = data_concat["description"]
data_match_parrilla = data_parrilla.apply(lambda x: x if x is np.NaN else␣
 ↪regex_parrilla.search(x))
mask_notnull_parrilla = data_match_parrilla.notnull()
data_parrilla = data_match_parrilla[mask_notnull_parrilla].apply(lambda x: x.
 ↪group("parrilla"))
data_concat.loc[mask_notnull_parrilla, 'parrilla'] = \
data_match_parrilla[mask_notnull_parrilla].apply(lambda x: x.group('parrilla').
 ↪lower())

patron_pileta = "(?P<pileta>(P|p)(I|i)(L|l)(E|e)(T|t)(A|a))"
regex_pileta = re.compile(patron_pileta)
data_pileta = data_concat["description"]
data_match_pileta = data_pileta.apply(lambda x: x if x is np.NaN else␣
 ↪regex_pileta.search(x))
mask_notnull_pileta = data_match_pileta.notnull()
data_pileta = data_match_pileta[mask_notnull_pileta].apply(lambda x: x.
 ↪group("pileta"))
data_concat.loc[mask_notnull_pileta, 'pileta'] = \
data_match_pileta[mask_notnull_pileta].apply(lambda x: x.group('pileta').
 ↪lower())

patron_patio = "(?P<patio>(P|p)(A|a)(T|t)(I|i)(O|o))"
regex_patio = re.compile(patron_patio)
data_patio = data_concat["description"]
data_match_patio = data_patio.apply(lambda x: x if x is np.NaN else regex_patio.
 ↪search(x))
mask_notnull_patio = data_match_patio.notnull()
data_patio = data_match_patio[mask_notnull_patio].apply(lambda x: x.
 ↪group("patio"))
data_concat.loc[mask_notnull_patio, 'patio'] = \
data_match_patio[mask_notnull_patio].apply(lambda x: x.group('patio').lower())

patron_quincho = "(?P<quincho>(Q|q)(U|u)(I|i)(N|n)(C|c)(H|h)(O|o))"
regex_quincho = re.compile(patron_quincho)
data_quincho = data_concat["description"]
data_match_quincho = data_quincho.apply(lambda x: x if x is np.NaN else␣
 ↪regex_quincho.search(x))
mask_notnull_quincho = data_match_quincho.notnull()
data_quincho = data_match_quincho[mask_notnull_quincho].apply(lambda x: x.
 ↪group("quincho"))
data_concat.loc[mask_notnull_quincho, 'quincho'] = \
data_match_quincho[mask_notnull_quincho].apply(lambda x: x.group('quincho').
 ↪lower())

patron_gimnasio = "(?P<gimnasio>(G|g)(I|i)(M|m)(N|n)(A|a)(C|c|S|s)(I|i)(O|o))"
```

```
regex_gimnasio = re.compile(patron_gimnasio)
data_gimnasio = data_concat["description"]
data_match_gimnasio = data_gimnasio.apply(lambda x: x if x is np.NaN else␣
 ↪regex_gimnasio.search(x))
mask_notnull_gimnasio = data_match_gimnasio.notnull()
data_gimnasio = data_match_gimnasio[mask_notnull_gimnasio].apply(lambda x: x.
 ↪group("gimnasio"))
data_concat.loc[mask_notnull_gimnasio, 'gimnasio'] = \
data_match_gimnasio[mask_notnull_gimnasio].apply(lambda x: x.group('gimnasio').
 ↪lower().replace("gimnacio", "gimnasio"))

patron_sum = "(?P<sum>(S|s)(U|u)(M|m))"
regex_sum = re.compile(patron_sum)
data_sum = data_concat["description"]
data_match_sum = data_sum.apply(lambda x: x if x is np.NaN else regex_sum.
 ↪search(x))
mask_notnull_sum = data_match_sum.notnull()
data_sum = data_match_sum[mask_notnull_sum].apply(lambda x: x.group("sum"))
data_concat.loc[mask_notnull_sum, 'sala_usos_multiples'] = \
data_match_sum[mask_notnull_sum].apply(lambda x: x.group('sum').lower())

patron_cochera = "(?
 ↪P<cochera>(C|c)(O|o)(C|c)(H|h)(E|e)(R|r)(A|a)|(E|e)(S|s)(T|t)(A|a)(C|c)(I|i)(O|o)(N|n)(A|a)
regex_cochera = re.compile(patron_cochera)
data_cochera = data_concat["description"]
data_match_cochera = data_cochera.apply(lambda x: x if x is np.NaN else␣
 ↪regex_cochera.search(x))
mask_notnull_cochera = data_match_cochera.notnull()
data_cochera = data_match_cochera[mask_notnull_cochera].apply(lambda x: x.
 ↪group("cochera"))
data_concat.loc[mask_notnull_cochera, 'cochera'] = \
data_match_cochera[mask_notnull_cochera].apply(lambda x: x.group('cochera').
 ↪lower().replace("estacionamiento", "cochera"))

patron_seguridad = "(?
 ↪P<seguridad>(S|s)(E|e)(G|g)(U|u)(R|r)(I|i)(D|d)(A|a)(D|d)|(P|p)(O|o)(R|r)(T|t)(E|e)(R|r)(O|
regex_seguridad = re.compile(patron_seguridad)
data_seguridad = data_concat["description"]
data_match_seguridad = data_seguridad.apply(lambda x: x if x is np.NaN else␣
 ↪regex_seguridad.search(x))
mask_notnull_seguridad = data_match_seguridad.notnull()
data_seguridad = data_match_seguridad[mask_notnull_seguridad].apply(lambda x: x.
 ↪group("seguridad"))
data_concat.loc[mask_notnull_seguridad, 'seguridad'] = \
data_match_seguridad[mask_notnull_seguridad].apply(lambda x: x.
 ↪group('seguridad').lower().replace("portero", "seguridad"))
```

```
patron_jardin = "(?P<jardin>(J|j)(A|a)(R|r)(D|d)(I|i)(N|n))"
regex_jardin = re.compile(patron_jardin)
data_jardin = data_concat["description"]
data_match_jardin = data_jardin.apply(lambda x: x if x is np.NaN else␣
 ↪regex_jardin.search(x))
mask_notnull_jardin = data_match_jardin.notnull()
data_jardin = data_match_jardin[mask_notnull_jardin].apply(lambda x: x.
 ↪group("jardin"))
data_concat.loc[mask_notnull_jardin, 'jardin'] = \
data_match_jardin[mask_notnull_jardin].apply(lambda x: x.group('jardin').
 ↪lower())

patron_frente = "(?P<frente>(F|f)(R|r)(E|e)(N|n)(T|t)(E|e))"
regex_frente = re.compile(patron_frente)
data_frente = data_concat["description"]
data_match_frente = data_frente.apply(lambda x: x if x is np.NaN else␣
 ↪regex_frente.search(x))
mask_notnull_frente = data_match_frente.notnull()
data_frente = data_match_frente[mask_notnull_frente].apply(lambda x: x.
 ↪group("frente"))
data_concat.loc[mask_notnull_frente, 'frente'] = \
data_match_frente[mask_notnull_frente].apply(lambda x: x.group('frente').
 ↪lower())

data_concat
```

[25]:

	operation	property_type	place_name	place_with_parent_names	country_name
0	sell	PH	Mataderos	|Argentina|Capital Federal|Mataderos|	Argentina
1	sell	apartment	La Plata	|Argentina|Bs.As. G.B.A. Zona Sur|La Plata|	Argentina
2	sell	apartment	Mataderos	|Argentina|Capital Federal|Mataderos|	Argentina
3	sell	PH	Liniers	|Argentina|Capital Federal|Liniers|	Argentina
4	sell	apartment	Centro	|Argentina|Buenos Aires Costa Atlántica|Mar de…	Argentina
…	…	…	…	…	…
121215	sell	apartment	Belgrano		
121216	sell	house	Beccar		
121217	sell	apartment	Villa Urquiza		
121218	sell	apartment	Plaza Colón		
121219	sell	apartment	Capital Federal		

```
121215              |Argentina|Capital Federal|Belgrano|    Argentina
121216  |Argentina|Bs.As. G.B.A. Zona Norte|San Isidro…    Argentina
121217         |Argentina|Capital Federal|Villa Urquiza|    Argentina
121218  |Argentina|Buenos Aires Costa Atlántica|Mar de…    Argentina
121219                       |Argentina|Capital Federal|    Argentina

                        state_name  geonames_id  \
0                   Capital Federal    3430787.0
1             Bs.As. G.B.A. Zona Sur    3432039.0
2                   Capital Federal    3430787.0
3                   Capital Federal    3431333.0
4       Buenos Aires Costa Atlántica    3435548.0
…                               …          …
121215              Capital Federal    3436077.0
121216     Bs.As. G.B.A. Zona Norte    3436080.0
121217              Capital Federal    3433775.0
121218  Buenos Aires Costa Atlántica          NaN
121219              Capital Federal    3433955.0

                             lat-lon        lat        lon  …  parrilla  \
0             -34.6618237,-58.5088387 -34.661824 -58.508839  …       NaN
1             -34.9038831,-57.9643295 -34.903883 -57.964330  …       NaN
2             -34.6522615,-58.5229825 -34.652262 -58.522982  …       NaN
3             -34.6477969,-58.5164244 -34.647797 -58.516424  …       NaN
4             -38.0026256,-57.5494468 -38.002626 -57.549447  …       NaN
…                                 …        …         …  …        …
121215                          NaN        NaN        NaN  …  parrilla
121216                          NaN        NaN        NaN  …  parrilla
121217  -34.5706388726,-58.4755963355 -34.570639 -58.475596  …  parrilla
121218                          NaN        NaN        NaN  …       NaN
121219                          NaN        NaN        NaN  …       NaN

        pileta  patio  quincho  gimnasio  sala_usos_multiples  cochera  \
0          NaN  patio      NaN       NaN                  NaN      NaN
1          NaN    NaN      NaN       NaN                  NaN  cochera
2          NaN    NaN      NaN       NaN                  NaN      NaN
3          NaN  patio      NaN       NaN                  NaN      NaN
4          NaN    NaN      NaN       NaN                  NaN      NaN
…           …    …       …         …                    …       …
121215  pileta    NaN  quincho       NaN                  NaN  cochera
121216  pileta    NaN  quincho       NaN                  NaN      NaN
121217     NaN    NaN      NaN       NaN                  NaN  cochera
121218     NaN    NaN      NaN       NaN                  NaN      NaN
121219     NaN    NaN      NaN       NaN                  NaN      NaN

        seguridad  jardin  frente
0             NaN     NaN     NaN
```

```
1              NaN     NaN  frente
2              NaN     NaN     NaN
3              NaN     NaN     NaN
4              NaN     NaN  frente
…              …       …       …
121215  seguridad      NaN     NaN
121216         NaN  jardin  frente
121217         NaN     NaN     NaN
121218         NaN     NaN  frente
121219         NaN     NaN  frente

[121220 rows x 50 columns]
```

4 4) Eliminación de nulos, ceros, outliers e información innecesaria

4.1 Datos innecesarios

[26]:
```
# No es de interés para el análisis actual la información inmobiliaria de␣
↪tiendas

mask_not_store = data_concat['property_type'] != 'store'
```

[27]:
```
data_concat = data_concat[mask_not_store]
```

[28]:
```
data_concat.shape
```

[28]:
```
(117084, 50)
```

4.2 Columnas innecesarias

[29]:
```
data_sin_columnas = data_concat.drop(['operation', 'place_with_parent_names',␣
↪'place_name', 'country_name', 'state_name',
                                      'geonames_id', 'lat-lon', 'floor',␣
↪'expenses', 'properati_url', 'image_thumbnail', '??', 'price_usd_per_m2',
                                      'place_name', 'currency',␣
↪'price_aprox_local_currency', 'surface_total_in_m2', 'price_per_m2',
                                      'price_aprox_usd', "lat", "lon", ␣
↪"Country", "Otra", "Barrios", "Pais", "Zona", "ambientes_desc",
                                       "ambientes_t", "ambientes", "rooms",␣
↪"title", "description"], axis = 1)
data_sin_columnas.head()
```

[29]:
```
  property_type     price  surface_covered_in_m2        Partido  \
0            PH   62000.0                   40.0      Mataderos
1     apartment  150000.0                    NaN       La Plata
2     apartment   72000.0                   55.0      Mataderos
3            PH   95000.0                    NaN        Liniers
```

```
4     apartment   64000.0                   35.0  Mar del Plata

   precio_usd_por_m2  ambientes_train  ambientes_imputados  ambientes_final  \
0        1550.000000                2                  0.0              2.0
1                NaN                0                  NaN              NaN
2        1309.090909                2                  0.0              2.0
3                NaN                3                  0.0              3.0
4        1828.571429                2                  0.0              2.0

  balcon parrilla pileta  patio quincho gimnasio sala_usos_multiples  cochera  \
0    NaN      NaN    NaN  patio     NaN      NaN                  NaN      NaN
1    NaN      NaN    NaN    NaN     NaN      NaN                  NaN  cochera
2    NaN      NaN    NaN    NaN     NaN      NaN                  NaN      NaN
3    NaN      NaN    NaN  patio     NaN      NaN                  NaN      NaN
4    NaN      NaN    NaN    NaN     NaN      NaN                  NaN      NaN

  seguridad jardin  frente
0       NaN    NaN     NaN
1       NaN    NaN  frente
2       NaN    NaN     NaN
3       NaN    NaN     NaN
4       NaN    NaN  frente
```

4.3 Nulos y Ceros

[30]:
```
# Se procede a eliminar en cada renglón las filas con nulos o ceros

data_partido_not_null = data_sin_columnas.dropna(subset = ["Partido"], how =␣
↪"any")
data_partido_not_empty = data_partido_not_null[data_partido_not_null.Partido !=␣
↪""]
data_ambientes_not_zero = data_partido_not_empty[(data_partido_not_empty.
↪ambientes_train > 0) | (data_partido_not_empty.ambientes_imputados != 0)]
data_surface_not_zero = data_ambientes_not_zero[data_ambientes_not_zero.
↪surface_covered_in_m2 > 0]
data_surface_not_null = data_surface_not_zero.dropna(subset =␣
↪["surface_covered_in_m2"], how = "any")
data_price_not_zero = data_surface_not_null[data_surface_not_null.
↪precio_usd_por_m2 > 0]
data_price_not_null = data_price_not_zero.dropna(subset =␣
↪["precio_usd_por_m2"], how = "any")
data_price_not_null
```

[30]:
```
      property_type     price  surface_covered_in_m2        Partido  \
0                PH   62000.0                   40.0      Mataderos
2         apartment   72000.0                   55.0      Mataderos
```

```
4          apartment   64000.0                  35.0  Mar del Plata
6                 PH  130000.0                  78.0  Vicente López
7          apartment  138000.0                  40.0       Belgrano
…                …        …                     …             …
121158     apartment  165000.0                  39.0       Recoleta
121215     apartment  870000.0                  93.0       Belgrano
121216         house  498000.0                 360.0     San Isidro
121217     apartment  131500.0                  39.0  Villa Urquiza
121218     apartment   95900.0                  48.0  Mar del Plata

        precio_usd_por_m2  ambientes_train  ambientes_imputados  \
0             1550.000000                2                  0.0
2             1309.090909                2                  0.0
4             1828.571429                2                  0.0
6             1666.666667                0                  3.0
7             3450.000000                0                  1.0
…                      …                …                    …
121158        4230.769231                0                  1.0
121215        9354.838710                3                  0.0
121216        1383.333333                0                  7.0
121217        3371.794872                0                  1.0
121218        1997.916667                2                  0.0

        ambientes_final  balcon  parrilla  pileta  patio  quincho gimnasio  \
0                   2.0     NaN       NaN     NaN  patio      NaN      NaN
2                   2.0     NaN       NaN     NaN    NaN      NaN      NaN
4                   2.0     NaN       NaN     NaN    NaN      NaN      NaN
6                   3.0     NaN       NaN     NaN  patio      NaN      NaN
7                   1.0     NaN       NaN  pileta    NaN      NaN      NaN
…                     …       …         …       …      …        …        …
121158              1.0     NaN  parrilla     NaN    NaN      NaN      NaN
121215              3.0     NaN  parrilla  pileta    NaN  quincho      NaN
121216              7.0     NaN  parrilla  pileta    NaN  quincho      NaN
121217              1.0  balcon  parrilla     NaN    NaN      NaN      NaN
121218              2.0     NaN       NaN     NaN    NaN      NaN      NaN

       sala_usos_multiples  cochera  seguridad  jardin  frente
0                      NaN      NaN        NaN     NaN     NaN
2                      NaN      NaN        NaN     NaN     NaN
4                      NaN      NaN        NaN     NaN  frente
6                      NaN      NaN        NaN     NaN  frente
7                      NaN      NaN        NaN     NaN     NaN
…                        …        …          …       …       …
121158                 NaN      NaN        NaN     NaN  frente
121215                 NaN  cochera  seguridad     NaN     NaN
121216                 NaN      NaN        NaN  jardin  frente
121217                 NaN  cochera        NaN     NaN     NaN
```

```
121218                NaN       NaN         NaN      NaN  frente

[82339 rows x 19 columns]
```

4.4 Outliers

Superficie

[31]:
```
q1_surface = data_price_not_null.surface_covered_in_m2.quantile(0.25)
q2_surface = data_price_not_null.surface_covered_in_m2.quantile(0.5)
q3_surface = data_price_not_null.surface_covered_in_m2.quantile(0.75)

higher_bound_surface = q3_surface + 1.5 * (q3_surface - q1_surface)
lower_bound_surface = q1_surface - 1.5 * (q3_surface - q1_surface)

print("El límite inferior es ", lower_bound_surface, " y el superior es ",␣
 ↪higher_bound_surface)

# Considerando que el límite inferior da negativo, se usará un estadístico␣
 ↪propio para el límite inferior

lower_bound_surface_nuevo = q1_surface.mean() * 0.25
print("El nuevo límite inferior es", lower_bound_surface_nuevo)

outlier_mask_up = data_price_not_null.surface_covered_in_m2 <␣
 ↪higher_bound_surface
outlier_mask_down = data_price_not_null.surface_covered_in_m2 >␣
 ↪lower_bound_surface_nuevo
outlier_mask = np.logical_and(outlier_mask_up, outlier_mask_down)
data_sin_outliers_superficie = data_price_not_null[outlier_mask]
data_sin_outliers_superficie
```

```
El límite inferior es  -108.0  y el superior es  300.0
El nuevo límite inferior es 11.25
```

[31]:

	property_type	price	surface_covered_in_m2	Partido	\
0	PH	62000.0	40.0	Mataderos	
2	apartment	72000.0	55.0	Mataderos	
4	apartment	64000.0	35.0	Mar del Plata	
6	PH	130000.0	78.0	Vicente López	
7	apartment	138000.0	40.0	Belgrano	
…	…	…	…	…	
121157	apartment	128000.0	35.0	Belgrano	
121158	apartment	165000.0	39.0	Recoleta	
121215	apartment	870000.0	93.0	Belgrano	
121217	apartment	131500.0	39.0	Villa Urquiza	
121218	apartment	95900.0	48.0	Mar del Plata	

```
        precio_usd_por_m2  ambientes_train  ambientes_imputados  \
0             1550.000000                2                  0.0
2             1309.090909                2                  0.0
4             1828.571429                2                  0.0
6             1666.666667                0                  3.0
7             3450.000000                0                  1.0
…                     …                …                    …
121157        3657.142857                0                  1.0
121158        4230.769231                0                  1.0
121215        9354.838710                3                  0.0
121217        3371.794872                0                  1.0
121218        1997.916667                2                  0.0

        ambientes_final  balcon  parrilla  pileta  patio  quincho  gimnasio  \
0                   2.0     NaN       NaN     NaN  patio      NaN       NaN
2                   2.0     NaN       NaN     NaN    NaN      NaN       NaN
4                   2.0     NaN       NaN     NaN    NaN      NaN       NaN
6                   3.0     NaN       NaN     NaN  patio      NaN       NaN
7                   1.0     NaN       NaN  pileta    NaN      NaN       NaN
…                   …       …         …       …      …        …         …
121157              1.0     NaN  parrilla     NaN    NaN      NaN  gimnasio
121158              1.0     NaN  parrilla     NaN    NaN      NaN       NaN
121215              3.0     NaN  parrilla  pileta    NaN  quincho       NaN
121217              1.0  balcon  parrilla     NaN    NaN      NaN       NaN
121218              2.0     NaN       NaN     NaN    NaN      NaN       NaN

       sala_usos_multiples  cochera  seguridad jardin  frente
0                      NaN      NaN        NaN    NaN     NaN
2                      NaN      NaN        NaN    NaN     NaN
4                      NaN      NaN        NaN    NaN  frente
6                      NaN      NaN        NaN    NaN  frente
7                      NaN      NaN        NaN    NaN     NaN
…                      …        …          …      …       …
121157                 sum  cochera        NaN    NaN  frente
121158                 NaN      NaN        NaN    NaN  frente
121215                 NaN  cochera  seguridad    NaN     NaN
121217                 NaN  cochera        NaN    NaN     NaN
121218                 NaN      NaN        NaN    NaN  frente

[76644 rows x 19 columns]
```

Precio

[32]:
```
q1_price = data_sin_outliers_superficie.precio_usd_por_m2.quantile(0.25)
q2_price = data_sin_outliers_superficie.precio_usd_por_m2.quantile(0.5)
q3_price = data_sin_outliers_superficie.precio_usd_por_m2.quantile(0.75)

higher_bound_price = q3_price + 1.5 * (q3_price - q1_price)
```

```
lower_bound_price = q1_price - 1.5 * (q3_price - q1_price)

print("El límite inferior es ", lower_bound_price, " y el superior es ",␣
 ↪higher_bound_price)

# Considerando que el número da negativo, se usará un estadístico propio para␣
 ↪el límite inferior

lower_bound_price_nuevo = q1_price.mean() * 0.25
print("El nuevo límite inferior es", lower_bound_price_nuevo)

outlier_mask_up = data_sin_outliers_superficie.precio_usd_por_m2 <␣
 ↪higher_bound_price
outlier_mask_down = data_sin_outliers_superficie.precio_usd_por_m2 >␣
 ↪lower_bound_price_nuevo
outlier_mask = np.logical_and(outlier_mask_up, outlier_mask_down)
data_sin_outliers_price = data_sin_outliers_superficie[outlier_mask]
data_sin_outliers_price
```

```
El límite inferior es  -480.34591194968584  y el superior es  4756.132075471698
El nuevo límite inferior es 370.8333333333333
```

[32]:

	property_type	price	surface_covered_in_m2	Partido	precio_usd_por_m2	ambientes_train	ambientes_imputados
0	PH	62000.0	40.0	Mataderos	1550.000000	2	0.0
2	apartment	72000.0	55.0	Mataderos	1309.090909	2	0.0
4	apartment	64000.0	35.0	Mar del Plata	1828.571429	2	0.0
6	PH	130000.0	78.0	Vicente López	1666.666667	0	3.0
7	apartment	138000.0	40.0	Belgrano	3450.000000	0	1.0
…	…	…	…	…	…	…	…
121156	house	170000.0	130.0	Pilar	1307.692308	0	4.0
121157	apartment	128000.0	35.0	Belgrano	3657.142857	0	1.0
121158	apartment	165000.0	39.0	Recoleta	4230.769231	0	1.0
121217	apartment	131500.0	39.0	Villa Urquiza	3371.794872	0	1.0
121218	apartment	95900.0	48.0	Mar del Plata	1997.916667	2	0.0

```
        ambientes_final  balcon  parrilla  pileta  patio quincho  gimnasio  \
0                   2.0     NaN       NaN     NaN  patio     NaN       NaN
2                   2.0     NaN       NaN     NaN    NaN     NaN       NaN
4                   2.0     NaN       NaN     NaN    NaN     NaN       NaN
6                   3.0     NaN       NaN     NaN  patio     NaN       NaN
7                   1.0     NaN       NaN  pileta    NaN     NaN       NaN
…                   …       …         …    …      …       …         …
121156              4.0     NaN  parrilla  pileta    NaN     NaN       NaN
121157              1.0     NaN  parrilla     NaN    NaN     NaN  gimnasio
121158              1.0     NaN  parrilla     NaN    NaN     NaN       NaN
121217              1.0  balcon  parrilla     NaN    NaN     NaN       NaN
121218              2.0     NaN       NaN     NaN    NaN     NaN       NaN

       sala_usos_multiples  cochera seguridad jardin  frente
0                      NaN      NaN       NaN    NaN     NaN
2                      NaN      NaN       NaN    NaN     NaN
4                      NaN      NaN       NaN    NaN  frente
6                      NaN      NaN       NaN    NaN  frente
7                      NaN      NaN       NaN    NaN     NaN
…                      …        …         …      …       …
121156                 NaN      NaN       NaN    NaN     NaN
121157                 sum  cochera       NaN    NaN  frente
121158                 NaN      NaN       NaN    NaN  frente
121217                 NaN  cochera       NaN    NaN     NaN
121218                 NaN      NaN       NaN    NaN  frente

[72172 rows x 19 columns]
```

Ambientes

[33]:
```
q1_ambientes = data_sin_outliers_price.ambientes_train.quantile(0.25)
q2_ambientes = data_sin_outliers_price.ambientes_train.quantile(0.5)
q3_ambientes = data_sin_outliers_price.ambientes_train.quantile(0.75)

higher_bound_ambientes = q3_ambientes + 1.5 * (q3_ambientes - q1_ambientes)
lower_bound_ambientes = q1_ambientes - 1.5 * (q3_ambientes - q1_ambientes)
print("El límite inferior es ", lower_bound_ambientes, " y el superior es ",␣
 ↪higher_bound_ambientes)
```

```
El límite inferior es  -4.5  y el superior es  7.5
```

[34]:
```
# Considerando que el límite inferior da negativo, no se usará pues solo tiene␣
 ↪lógica que un departamento tenga al menos 1 ambiente.

mask = data_sin_outliers_price.ambientes_train > 7
mask2 = data_sin_outliers_price.loc[mask, :]
```

```
data_sin_outliers_ambientes = data_sin_outliers_price.drop(mask2.index, axis =␣
↪0)
data_sin_outliers_ambientes
```

[34]:

	property_type	price	surface_covered_in_m2	Partido	\
0	PH	62000.0	40.0	Mataderos	
2	apartment	72000.0	55.0	Mataderos	
4	apartment	64000.0	35.0	Mar del Plata	
6	PH	130000.0	78.0	Vicente López	
7	apartment	138000.0	40.0	Belgrano	
…	…	…	…	…	
121156	house	170000.0	130.0	Pilar	
121157	apartment	128000.0	35.0	Belgrano	
121158	apartment	165000.0	39.0	Recoleta	
121217	apartment	131500.0	39.0	Villa Urquiza	
121218	apartment	95900.0	48.0	Mar del Plata	

	precio_usd_por_m2	ambientes_train	ambientes_imputados	\
0	1550.000000	2	0.0	
2	1309.090909	2	0.0	
4	1828.571429	2	0.0	
6	1666.666667	0	3.0	
7	3450.000000	0	1.0	
…	…	…	…	
121156	1307.692308	0	4.0	
121157	3657.142857	0	1.0	
121158	4230.769231	0	1.0	
121217	3371.794872	0	1.0	
121218	1997.916667	2	0.0	

	ambientes_final	balcon	parrilla	pileta	patio	quincho	gimnasio	\
0	2.0	NaN	NaN	NaN	patio	NaN	NaN	
2	2.0	NaN	NaN	NaN	NaN	NaN	NaN	
4	2.0	NaN	NaN	NaN	NaN	NaN	NaN	
6	3.0	NaN	NaN	NaN	patio	NaN	NaN	
7	1.0	NaN	NaN	pileta	NaN	NaN	NaN	
…	…	…	…	…	…	…	…	
121156	4.0	NaN	parrilla	pileta	NaN	NaN	NaN	
121157	1.0	NaN	parrilla	NaN	NaN	NaN	gimnasio	
121158	1.0	NaN	parrilla	NaN	NaN	NaN	NaN	
121217	1.0	balcon	parrilla	NaN	NaN	NaN	NaN	
121218	2.0	NaN	NaN	NaN	NaN	NaN	NaN	

	sala_usos_multiples	cochera	seguridad	jardin	frente
0	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	frente

```
6                      NaN      NaN       NaN    NaN  frente
7                      NaN      NaN       NaN    NaN     NaN
…                      …        …         …      …       …
121156                 NaN      NaN       NaN    NaN     NaN
121157                 sum   cochera      NaN    NaN  frente
121158                 NaN      NaN       NaN    NaN  frente
121217                 NaN   cochera      NaN    NaN     NaN
121218                 NaN      NaN       NaN    NaN  frente

[71766 rows x 19 columns]
```

5 5) Exportación del dataset final

[35]:
```
data_final = data_sin_outliers_ambientes.copy()
data_final.to_csv('data_final.csv', index = False, sep=';')
```