

1) Data wrangling

August 8, 2023

1 1) Preparación previa

1.0.1 Carga de librerías

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

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")
data.head()
```

```
[2]: Unnamed: 0 operation property_type place_name \
0      0      sell      PH Mataderos
1      1      sell  apartment La Plata
2      2      sell  apartment Mataderos
3      3      sell      PH Liniers
4      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 ... surface_covered_in_m2 price_usd_per_m2 price_per_m2 \
```

	floor	rooms	expenses	properati_url	\
0	-34.661824	...	40.0	1127.272727	1550.000000
1	-34.903883	...	NaN	NaN	NaN
2	-34.652262	...	55.0	1309.090909	1309.090909
3	-34.647797	...	NaN	NaN	NaN
4	-38.002626	...	35.0	1828.571429	1828.571429

	floor	rooms	expenses	properati_url	\
0	NaN	NaN	NaN	http://www.properati.com.ar/15bo8_venta_ph_mat...	
1	NaN	NaN	NaN	http://www.properati.com.ar/15bob_venta_depart...	
2	NaN	NaN	NaN	http://www.properati.com.ar/15bod_venta_depart...	
3	NaN	NaN	NaN	http://www.properati.com.ar/15boh_venta_ph_lin...	
4	NaN	NaN	NaN	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 26 columns]

1.0.3 Separación de columna con muchas ubicaciones

```
[3]: # La columna "place_with_parent_names" tiene información separada con '|'. Se
      ↪separa para obtener info adicional
separado = data["place_with_parent_names"].str.split('|', expand = True)
separado.head()
```

[3]:	0	1	2	3	4	5	\
0	Argentina		Capital Federal	Mataderos		None	
1	Argentina		Bs.As. G.B.A. Zona Sur	La Plata		None	
2	Argentina		Capital Federal	Mataderos		None	

3	Argentina		Capital Federal		Liniers	None
4	Argentina	Buenos Aires	Costa Atlántica		Mar del Plata	Centro

```

6
0 None
1 None
2 None
3 None
4 None

```

1.0.4 Agregado de la nueva información en nuevas columnas

```

[4]: separado.columns = ['??', 'Pais', 'Zona', 'Partido', 'Barrios', 'Country',
    ↪ 'Otra']
data_concat = pd.concat([data, separado], axis=1)

```

1.0.5 Cálculo del tipo de cambio para corroborar

```

[5]: data_concat['TC'] = data_concat['price_aprox_local_currency'] /
    ↪ data_concat['price_aprox_usd']
    # Calculando el promedio, decidimos tomar la columna 'price_aprox_usd' como el
    ↪ $ de las propiedades
data_concat['TC'].mean().round(2)

```

```

[5]: 17.64

```

1.0.6 Revisión de registros nulos según columnas

```

[6]: data_concat.isnull().sum()

```

```

[6]: Unnamed: 0          0
operation              0
property_type         0
place_name            23
place_with_parent_names 0
country_name          0
state_name            0
geonames_id          18717
lat-lon              51550
lat                  51550
lon                  51550
price                20410
currency             20411
price_aprox_local_currency 20410
price_aprox_usd       20410
surface_total_in_m2    39328
surface_covered_in_m2  19907

```

price_usd_per_m2	52603
price_per_m2	33562
floor	113321
rooms	73830
expenses	106958
properati_url	0
description	2
title	0
image_thumbnail	3112
??	0
Pais	0
Zona	0
Partido	0
Barrios	4780
Country	80780
Otra	120672
TC	20411
dtype:	int64

1.0.7 Creación de función para limpiar superficie

```
[7]: # La siguiente función nos permite limpiar la superficie según las
      ↪ inconsistencias entre la total y la cubierta.
      # Se eligió dicha columna con respecto a 'surface_total_in_m2' ya que tenía
      ↪ menor cantidad de registros nulos
      def limpieza_superficie(sup_total, sup_cubierta):
          if sup_total is not None and sup_cubierta is not None:
              if sup_total >= sup_cubierta:
                  superficie = sup_total
              else:
                  superficie = np.NaN
          elif sup_total is not None:
              superficie = sup_total
          elif sup_cubierta is not None:
              superficie = sup_cubierta
          else: superficie = np.NaN

          return superficie
```

1.0.8 Aplicación de la susodicha función

```
[8]: data_concat["superficie"] =
      ↪ data_concat[['surface_total_in_m2', 'surface_covered_in_m2']].apply(lambda
      ↪ data_concat:
      ↪ limpieza_superficie(data_concat['surface_total_in_m2'], data_concat['surface_covered_in_m2'])
      data_concat.head(10)
```

```

[8]: Unnamed: 0 operation property_type place_name \
0      0      sell      PH      Mataderos
1      1      sell      apartment      La Plata
2      2      sell      apartment      Mataderos
3      3      sell      PH      Liniers
4      4      sell      apartment      Centro
5      5      sell      house      Gualeguaychú
6      6      sell      PH      Munro
7      7      sell      apartment      Belgrano
8      8      sell      apartment      Belgrano
9      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      ...      image_thumbnail      ?? \
0      -34.661824      ...      https://thumbs4.properati.com/8/BluUYiHJLhgIIK...
1      -34.903883      ...      https://thumbs4.properati.com/7/ikpVBu2ztHA7jv...
2      -34.652262      ...      https://thumbs4.properati.com/5/SXKr34F_IwG3W_...
3      -34.647797      ...      https://thumbs4.properati.com/3/DgIfX-85Mog5SP...
4      -38.002626      ...      https://thumbs4.properati.com/5/xrRqlNcSI_vs-f...
5      -33.014071      ...      https://thumbs4.properati.com/6/q-w68gvaUEQVXI...
6      -34.532957      ...      https://thumbs4.properati.com/5/6G0XsHCyDu1aGx...
7      -34.559873      ...      https://thumbs4.properati.com/1/IHxARynlr8sPEW...
8      -34.559873      ...      https://thumbs4.properati.com/2/J3z0jgaFhrkvnv...
9      -32.942031      ...      https://thumbs4.properati.com/8/RCf1YEWdF4rv98...

```

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

	Otra	TC	superficie
0	None	17.6445	55.0
1	None	17.6445	NaN
2	None	17.6445	55.0
3	None	17.6445	NaN
4	None	17.6445	35.0
5	None	NaN	NaN
6	None	17.6445	106.0
7	None	17.6445	45.0
8	None	17.6445	65.0
9	None	17.6445	NaN

[10 rows x 35 columns]

1.0.9 Creación de la columna de precios por metros cuadrados con la nueva superficie

```
[9]: data_concat['Precio_USD_por_M2']=data_concat.price_aprox_usd/data_concat.
      ↳superficie
data_concat.head()
```

```
[9]: Unnamed: 0 operation property_type place_name \
0          0      sell          PH  Mataderos
1          1      sell    apartment  La Plata
2          2      sell    apartment  Mataderos
3          3      sell          PH    Liniers
4          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	...	??	Pais	Zona	Partido \
0	-34.661824	...		Argentina	Capital Federal	Mataderos
1	-34.903883	...		Argentina	Bs.As. G.B.A. Zona Sur	La Plata
2	-34.652262	...		Argentina	Capital Federal	Mataderos
3	-34.647797	...		Argentina	Capital Federal	Liniers
4	-38.002626	...		Argentina	Buenos Aires Costa Atlántica	Mar del Plata

	Barrios	Country	Otra	TC	superficie	Precio_USD_por_M2
0		None	None	17.6445	55.0	1127.272727
1		None	None	17.6445	NaN	NaN
2		None	None	17.6445	55.0	1309.090909
3		None	None	17.6445	NaN	NaN
4	Centro		None	17.6445	35.0	1828.571429

[5 rows x 36 columns]

1.0.10 Eliminación de columnas

```
[10]: # Algunas se eliminan por no corresponder con nuestro analisis y otras por
      ↪ estar repetidas
data_eliminacion = data_concat.drop(['Unnamed: 0', 'operation',
      ↪ 'place_with_parent_names', 'place_name', 'country_name', 'state_name',
      ↪ 'description', 'geonames_id', 'lat-lon',
      ↪ 'floor', 'rooms', 'expenses', 'properati_url',
      ↪ 'image_thumbnail', 'title', '??'], axis =
      ↪ 1)
data_eliminacion.head()
```

	property_type	lat	lon	price	currency \
0	PH	-34.661824	-58.508839	62000.0	USD
1	apartment	-34.903883	-57.964330	150000.0	USD
2	apartment	-34.652262	-58.522982	72000.0	USD
3	PH	-34.647797	-58.516424	95000.0	USD
4	apartment	-38.002626	-57.549447	64000.0	USD

	price_aprox_local_currency	price_aprox_usd	surface_total_in_m2 \
0	1093959.0	62000.0	55.0
1	2646675.0	150000.0	NaN
2	1270404.0	72000.0	55.0
3	1676227.5	95000.0	NaN

```
4          1129248.0          64000.0          35.0
```

```

    surface_covered_in_m2  price_usd_per_m2  price_per_m2    Pais \
0          40.0          1127.272727    1550.000000  Argentina
1           NaN           NaN           NaN    Argentina
2          55.0          1309.090909    1309.090909  Argentina
3           NaN           NaN           NaN    Argentina
4          35.0          1828.571429    1828.571429  Argentina

```

```

              Zona      Partido Barrios Country  Otra      TC \
0      Capital Federal      Mataderos      None  None  17.6445
1      Bs.As. G.B.A. Zona Sur      La Plata      None  None  17.6445
2      Capital Federal      Mataderos      None  None  17.6445
3      Capital Federal      Liniers      None  None  17.6445
4  Buenos Aires Costa Atlántica  Mar del Plata  Centro      None  17.6445

```

```

    superficie  Precio_USD_por_M2
0      55.0          1127.272727
1       NaN           NaN
2      55.0          1309.090909
3       NaN           NaN
4      35.0          1828.571429

```

```
[11]: # Corroboramos el tamaño del resultado
data_eliminacion.shape
```

```
[11]: (121220, 20)
```

1.0.11 Eliminamos los valores nulos del precio nuevo

```
[12]: df_final = data_eliminacion[data_eliminacion['Precio_USD_por_M2'].notna()]
df_final
```

```

[12]:      property_type      lat      lon      price currency \
0          PH -34.661824 -58.508839    62000.0      USD
2      apartment -34.652262 -58.522982    72000.0      USD
4      apartment -38.002626 -57.549447    64000.0      USD
6          PH -34.532957 -58.521782   130000.0      USD
7      apartment -34.559873 -58.443362   138000.0      USD
...      ...      ...      ...      ...      ...
121215  apartment      NaN      NaN    870000.0      USD
121216    house      NaN      NaN    498000.0      USD
121217  apartment -34.570639 -58.475596   131500.0      USD
121218  apartment      NaN      NaN     95900.0      USD
121219  apartment      NaN      NaN   129000.0      USD

      price_aprox_local_currency  price_aprox_usd  surface_total_in_m2 \

```


0	1093959.00	62000.0	55.0
2	1270404.00	72000.0	55.0
4	1129248.00	64000.0	35.0
6	2293785.00	130000.0	106.0
7	2434941.00	138000.0	45.0
...
121215	15350715.00	870000.0	113.0
121216	8786961.00	498000.0	360.0
121217	2320251.75	131500.0	46.0
121218	1692107.55	95900.0	48.0
121219	2276140.50	129000.0	77.0

	surface_covered_in_m2	price_usd_per_m2	price_per_m2	Pais \
0	40.0	1127.272727	1550.000000	Argentina
2	55.0	1309.090909	1309.090909	Argentina
4	35.0	1828.571429	1828.571429	Argentina
6	78.0	1226.415094	1666.666667	Argentina
7	40.0	3066.666667	3450.000000	Argentina
...
121215	93.0	7699.115044	9354.838710	Argentina
121216	360.0	1383.333333	1383.333333	Argentina
121217	39.0	2858.695652	3371.794872	Argentina
121218	48.0	1997.916667	1997.916667	Argentina
121219	77.0	1675.324675	1675.324675	Argentina

	Zona	Partido	Barrios	Country \
0	Capital Federal	Mataderos		None
2	Capital Federal	Mataderos		None
4	Buenos Aires Costa Atlántica	Mar del Plata	Centro	
6	Bs.As. G.B.A. Zona Norte	Vicente López	Munro	
7	Capital Federal	Belgrano		None
...
121215	Capital Federal	Belgrano		None
121216	Bs.As. G.B.A. Zona Norte	San Isidro	Beccar	
121217	Capital Federal	Villa Urquiza		None
121218	Buenos Aires Costa Atlántica	Mar del Plata	Plaza Colón	
121219	Capital Federal		None	None

	Otra	TC	superficie	Precio_USD_por_M2
0	None	17.6445	55.0	1127.272727
2	None	17.6445	55.0	1309.090909
4	None	17.6445	35.0	1828.571429
6	None	17.6445	106.0	1226.415094
7	None	17.6445	45.0	3066.666667
...
121215	None	17.6445	113.0	7699.115044
121216	None	17.6445	360.0	1383.333333

121217	None	17.6445	46.0	2858.695652
121218	None	17.6445	48.0	1997.916667
121219	None	17.6445	77.0	1675.324675

[62375 rows x 20 columns]

2 2) Análisis estadístico breve del resultado

2.0.1 Cálculo de % de registros según la provincia/región

```
[13]: (df_final.Zona.value_counts() / df_final.Zona.size) * 100
```

```
[13]: Capital Federal          35.631263
Bs.As. G.B.A. Zona Norte    26.295792
Bs.As. G.B.A. Zona Sur      8.883367
Buenos Aires Costa Atlántica 8.219639
Bs.As. G.B.A. Zona Oeste    6.613226
Santa Fe                    5.539078
Córdoba                     5.014830
Buenos Aires Interior       1.426854
Corrientes                   0.472946
Mendoza                      0.463327
Neuquén                      0.343086
Misiones                     0.213226
San Luis                     0.189178
Río Negro                    0.160321
Tucumán                      0.153908
Entre Ríos                   0.120240
Salta                        0.078557
Chubut                       0.054509
Tierra Del Fuego            0.041683
Chaco                        0.032064
La Pampa                     0.016032
Santa Cruz                   0.012826
Catamarca                    0.008016
Jujuy                        0.006413
Santiago Del Estero          0.004810
San Juan                     0.003206
La Rioja                     0.001603
Name: Zona, dtype: float64
```

2.0.2 Agrupamos por provincia/región

```
[14]: data_agrupada_prov = df_final.groupby('Zona')
data_agrupada_prov
```

```
[14]: <pandas.core.groupby.generic.DataFrameGroupBy object at 0x00000251A4EB1710>
```

2.0.3 Calculamos las métricas según provincia/región

```
[15]: medidas = data_agrupada_prov[["Precio_USD_por_M2"]].describe().round(2)
```

```
[16]: # Ordenamos según count
medidas.sort_values(by=[('Precio_USD_por_M2', 'count')], ascending = False)
```

```
[16]:
```

	Precio_USD_por_M2				
	count	mean	std	min	\
Zona					
Capital Federal	22225.0	3123.37	3720.31	24.21	
Bs.As. G.B.A. Zona Norte	16402.0	1833.34	1789.96	4.17	
Bs.As. G.B.A. Zona Sur	5541.0	1521.41	1108.30	19.31	
Buenos Aires Costa Atlántica	5127.0	1522.44	807.09	3.00	
Bs.As. G.B.A. Zona Oeste	4125.0	1289.72	1582.14	7.00	
Santa Fe	3455.0	2251.97	4445.15	0.60	
Córdoba	3128.0	1186.38	860.71	1.18	
Buenos Aires Interior	890.0	934.20	688.90	2.07	
Corrientes	295.0	1499.20	627.57	83.68	
Mendoza	289.0	1683.17	2602.58	16.47	
Neuquén	214.0	1815.42	1460.07	72.12	
Misiones	133.0	730.31	650.03	9.88	
San Luis	118.0	1565.47	3550.66	11.22	
Río Negro	100.0	2301.48	1295.92	82.04	
Tucumán	96.0	1823.39	3720.90	62.65	
Entre Ríos	75.0	1260.00	2754.70	5.61	
Salta	49.0	2624.66	5055.14	51.75	
Chubut	34.0	2099.60	2329.06	13.50	
Tierra Del Fuego	26.0	1011.24	675.48	362.21	
Chaco	20.0	962.12	581.70	163.03	
La Pampa	10.0	899.38	335.63	392.59	
Santa Cruz	8.0	778.64	603.60	315.97	
Catamarca	5.0	611.17	610.88	145.45	
Jujuy	4.0	617.08	711.25	210.31	
Santiago Del Estero	3.0	269.77	224.19	72.00	
San Juan	2.0	1648.24	2182.48	105.00	
La Rioja	1.0	243.03	NaN	243.03	

	25%	50%	75%	max
Zona				
Capital Federal	1969.70	2469.44	3068.18	206333.33
Bs.As. G.B.A. Zona Norte	1128.73	1666.67	2327.11	48296.22
Bs.As. G.B.A. Zona Sur	1000.00	1500.00	1937.50	23140.76
Buenos Aires Costa Atlántica	909.09	1579.17	2045.44	6422.22
Bs.As. G.B.A. Zona Oeste	714.29	1225.00	1666.67	80000.00
Santa Fe	1266.79	1585.44	1889.57	57382.08

Córdoba	801.96	1192.60	1542.53	30000.00
Buenos Aires Interior	314.92	833.33	1476.13	4166.67
Corrientes	1078.25	1461.54	1865.32	3500.00
Mendoza	992.08	1428.57	1684.21	31578.95
Neuquén	962.86	1854.00	2444.22	17073.17
Misiones	224.33	467.36	1215.69	3115.38
San Luis	252.76	759.51	1283.17	33333.33
Río Negro	1309.28	2305.84	3304.38	5000.00
Tucumán	785.17	785.17	1304.54	21666.67
Entre Ríos	385.58	800.00	1518.93	24038.46
Salta	690.26	1105.26	1633.93	23750.00
Chubut	1275.88	1765.69	2271.88	14540.55
Tierra Del Fuego	543.17	869.30	1172.63	3333.33
Chaco	542.14	871.87	1145.54	2386.36
La Pampa	632.69	855.94	1186.39	1402.09
Santa Cruz	434.70	565.74	793.93	2166.67
Catamarca	243.03	471.86	528.85	1666.67
Jujuy	262.78	287.74	642.04	1682.51
Santiago Del Estero	148.00	224.00	368.65	513.31
San Juan	876.62	1648.24	2419.87	3191.49
La Rioja	243.03	243.03	243.03	243.03

2.0.4 Métricas generales del data set

```
[17]: medidas_finales = df_final[["Precio_USD_por_M2"]].describe().round(2)
medidas_finales
```

```
[17]:      Precio_USD_por_M2
count      62375.00
mean        2176.12
std         2814.35
min           0.60
25%        1222.22
50%        1818.18
75%        2500.00
max       206333.33
```

2.0.5 Exportación del DF final

```
[18]: df_final.to_csv('DF_Final.csv', index = False)
```