Trabajo práctico N°1

Análisis exploratorio de un dataset de precios de propiedades Properati

Grupo N° 4

Eckerman, Luján
González, Gilda
Grao, Brenda
Guerrero, Matías
Magariño, Nicolás

Contenido

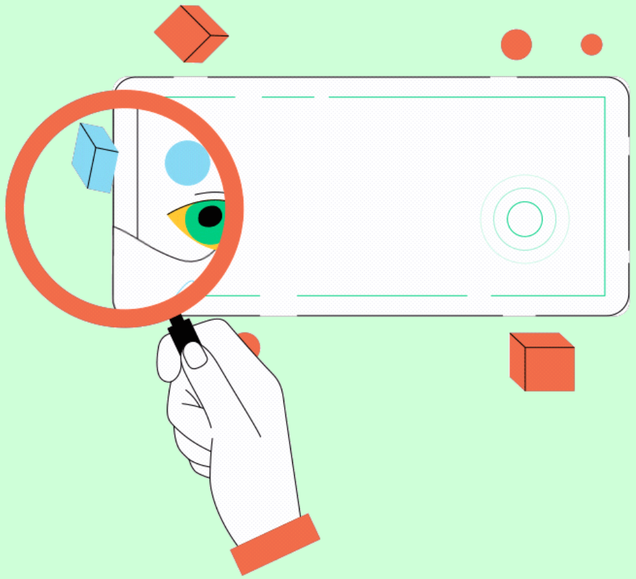
¿Qué hicimos?

1. Análisis exploratorio
2. Columnas de interés
3. Limpieza
4. Imputación
5. Limpieza final
6. Resultados

1) Análisis exploratorio

Primera aproximación al dataset

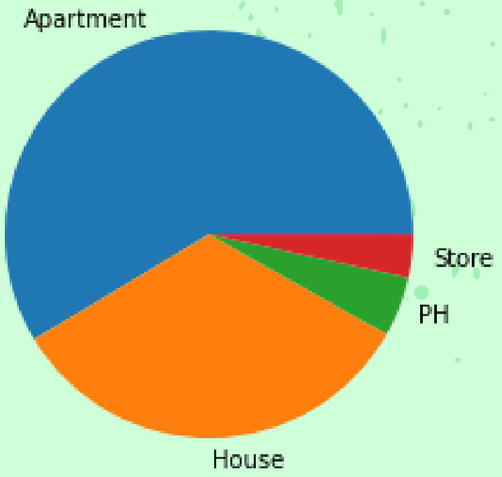
- Entender su contenido
- Detectar variables relevantes
- Analizar patrones
- Determinar estrategias a llevar

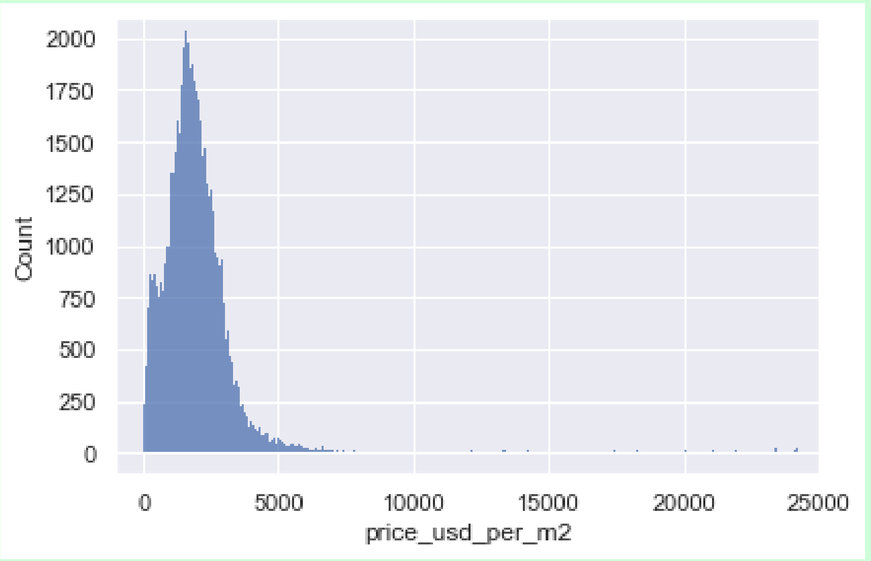


2) Evaluación de datos de interés

Columnas importantes

- Operation
- Property_type
- Place_name
- Place_with_parent_names
- State_name
- Price
- Surface_total_in_m2
- Surface_covered_in_m2
- Currency
- Lat-Lon

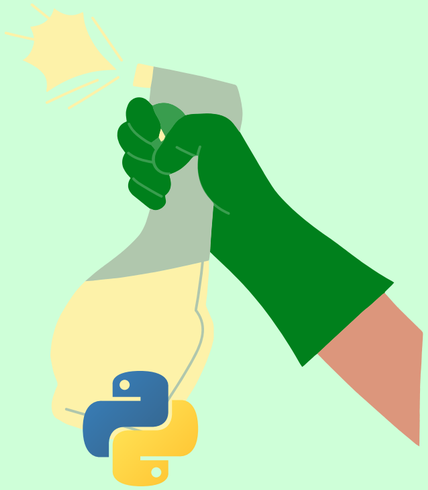




3) Limpieza

Segmentos principales

- Limpieza gruesa
- Extracción de datos de descripción y título
- Extracción de ubicación
- Quitar publicaciones duplicadas



3) Limpieza

Limpieza gruesa

- Quitar columnas innecesarias (links, operation, country, etc)
- Quitar publicaciones en otras monedas (UYU y PEN)
- Quitar publicaciones que no sean Apartment y House
- Utilizar solo publicaciones que tengan >1000 registros en esa ciudad.

3) Limpieza

Extracción de datos de descripción y título

Usando Regex en la columna títulos y descripción, se obtuvo:

- Cantidad de ambientes y cochera (dato creado)
- Superficie (complementó a la existente)
- Cantidad de habitaciones (complementó a la existente)
- Variable dummy para eliminar publicaciones falsas, proyectos, fideicomisos, etc
- Detección de amenities
- Creación de variables dummies que podrían aportar a la capacidad predictiva del modelo (tiene pileta, barrio vip, etc.)

3) Limpieza

Extracción de ubicación

Sobre la columna place_with_parent_names separada por pipes, tomamos la tercera como parámetro principal de ubicación.

En caso de no existir, se completó con place_name

Argentina |Capital Federal | Chacarita
Argentina |Santa Fé | Rosario

3) Limpieza

Quitar publicaciones duplicadas

Criterio de eliminación:

Filas que tengan exactamente lo mismo en:

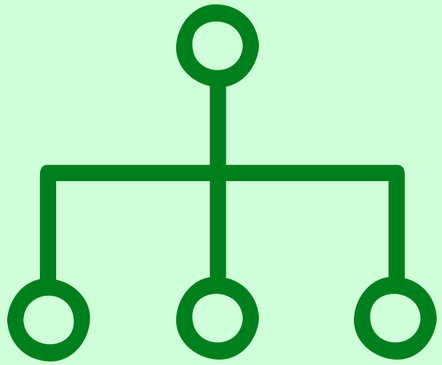
- property_type
- place_name
- place_with_parent_names
- lat-lon
- price
- currency
- surface_total_in_m2
- surface_covered_in_m2
- title

4) Imputación de precio

Jerarquía de imputación

- Dato directo de precio/m2 en dolares
- Calcular precio/m2 tomando precio / superficie
  - Análisis y limpieza de columna de superficie
  - Eliminación de outliers
- Imputación por media según zona y tipo de propiedad

Precisión



4) Imputación de precio

Dato directo precio/m2 en dólares

- 56.61% del dataset original tenía este campo completo

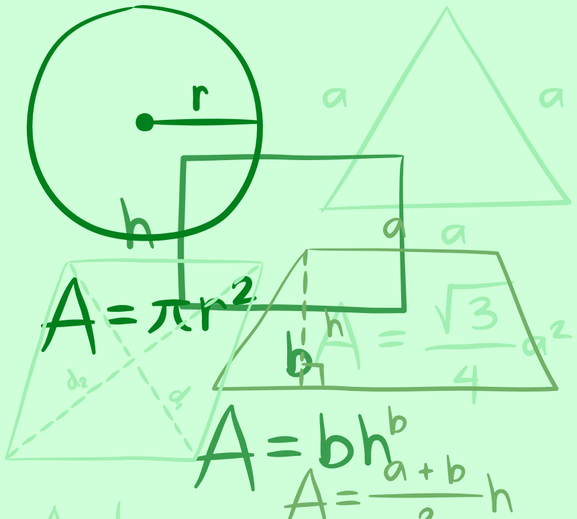
Dato directo precio/m2 en pesos

- Descartado por no ser confiable, tipo de cambio irreal

4) Imputación de precio

Análisis y limpieza de columna de superficie

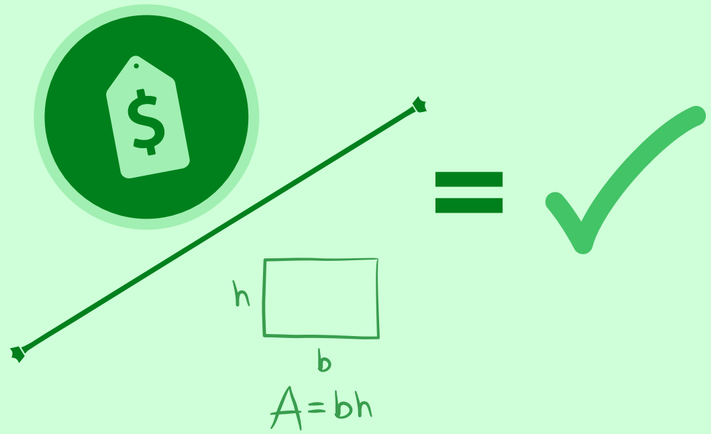
- Unificación de superficies por columnas total y cubierta
- Criterios:
  - Si una no tiene dato, se toma la que tiene.
  - Si cubierta>total, se toma cubierta.
  - Si total>cubierta, se toma cubierta.
  - Si el dato es igual, se toma cualquiera.



4) Imputación de precio

Calcular precio/m2 tomando precio/superficie

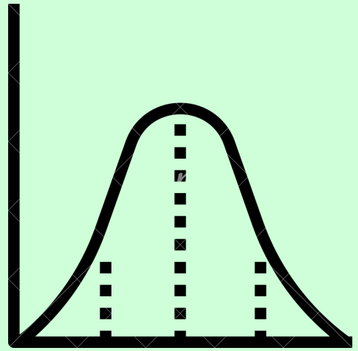
- Tomando el precio de publicación se divide por la superficie, obteniendo el precio del m2 en dólares.



4) Imputación de precio

Eliminacion de outliers

- Se decidió eliminar aquellos precios /m2 que pertenecían al:
  - Cuantilo 1% inferior (publicaciones falsas para obtener mejor ranking en web)
  - Cuantilo 4% superior (publicaciones exageradamente altas)



4) Imputación de precio

Imputación por media según zona y tipo

Imputación de precios tomando los criterios:

- Zona
- Tipo de propiedad

Se obtuvo la media en cada una de ellas y luego se imputó en las filas sin precio/m2

Ej: El promedio de m2 de un departamento en Abasto
Ej: El promedio de m2 de una casa en La Plata

5) Limpieza final

Limpieza de datos no completables

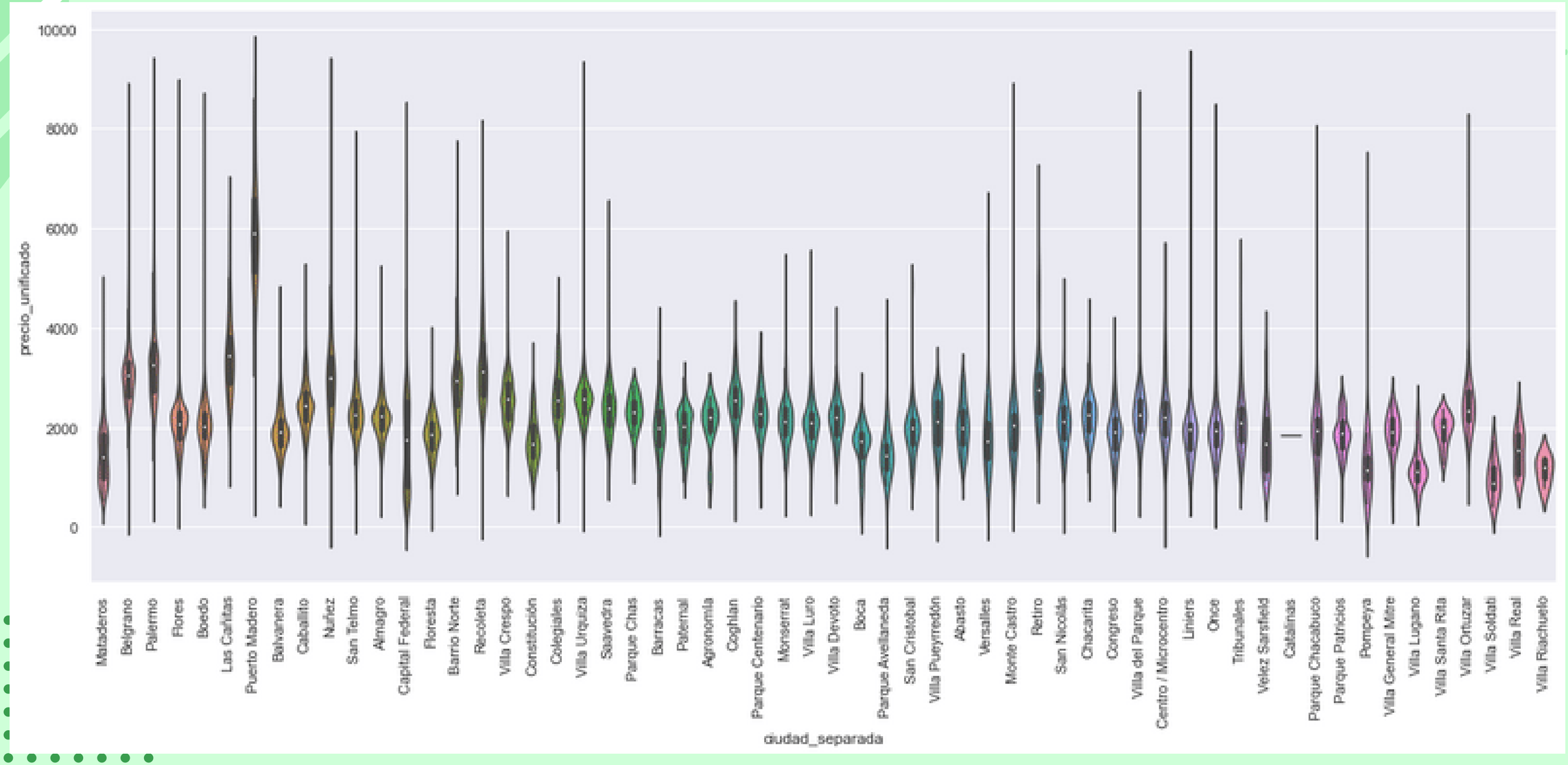
- Eliminación de filas sin precio/m2
- Eliminación de columnas que no aportan valor al caso de negocio.

6) Resultados

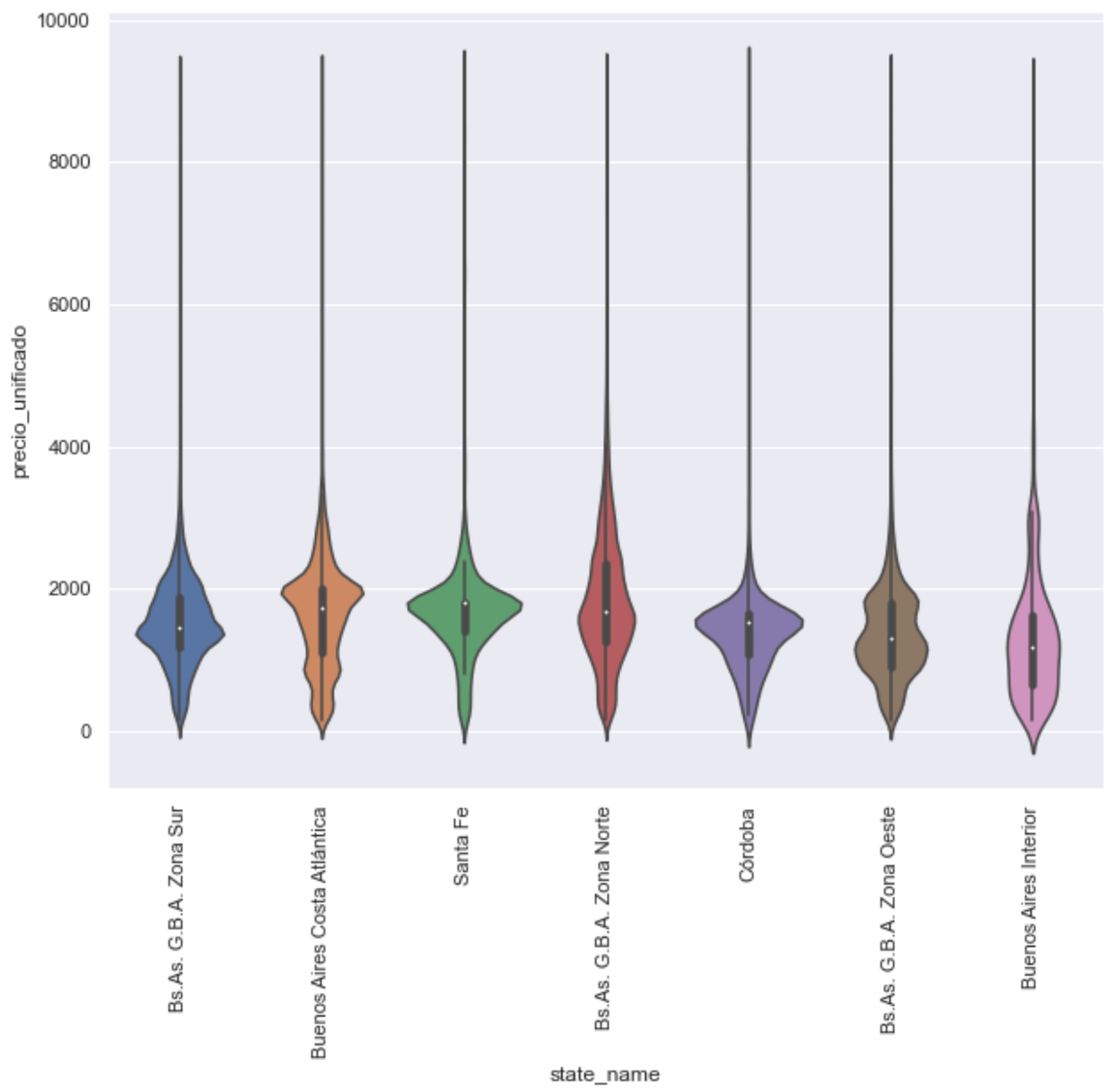
Indicadores principales de limpieza

	Dataset original	Dataset final	Dif %
Cantidad de filas	121220	86079	- 28.99 %
No nulos precio/m2	68617	86079	+ 25.45 %
No nulos superficie	81892	80328	- 1.91 %
Cant. datos en habitación	47390	60210	+ 27.05 %
Cant. datos con ambientes	0	31391	-
Cant. datos con cochera	0	37886	-

6) Resultados



6) Resultados



¡Muchas gracias!