

# Data Science 2022

## TRABAJO PRÁCTICO I

# Análisis exploratorio de un dataset de precios de propiedades

### **Grupo 4**

Atienza Rela, Guadalupe

Atienza Rela, Macarena

Carrero, Luis

Federico, Santiago Raul

Roa Herrera, Juan

## ÍNDICE

1

Exploración del dataset

2

Limpieza y transformación de datos

3

Análisis y conclusiones finales

1

Exploración del dataset

2

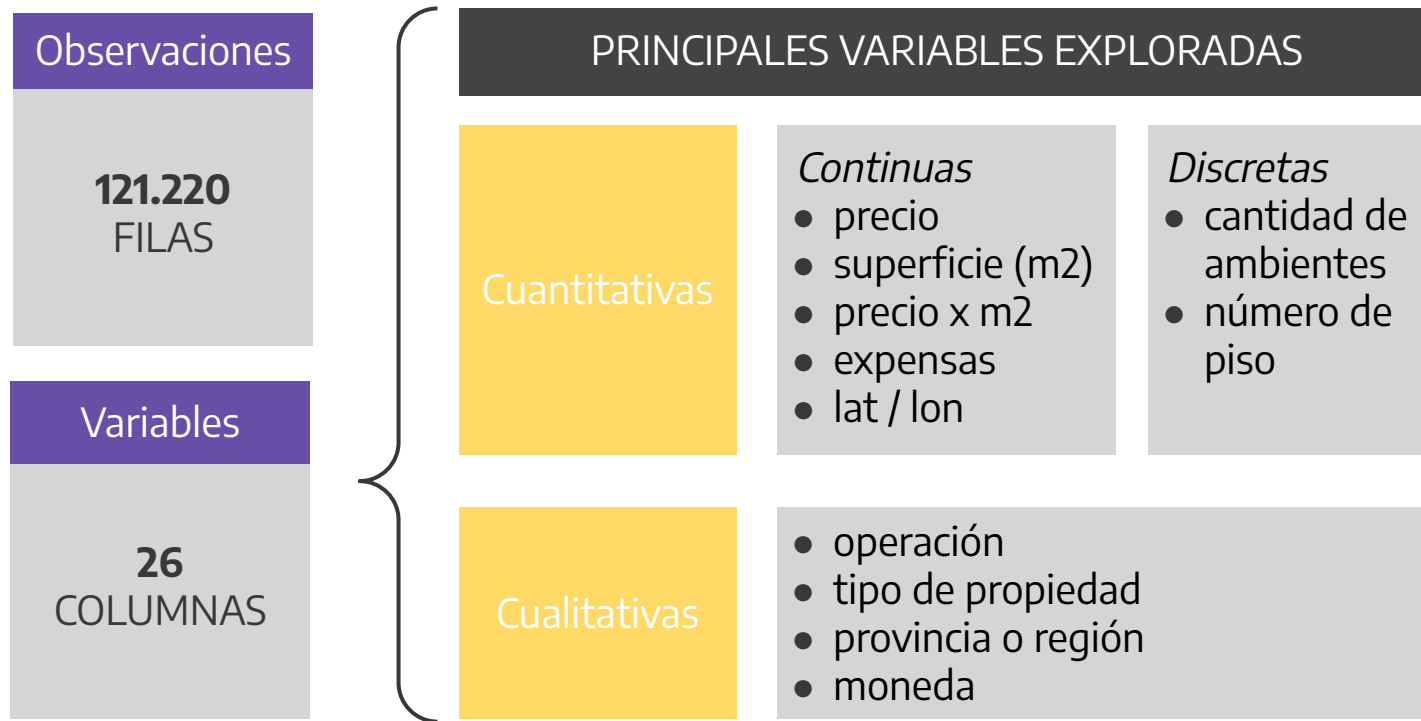
Limpieza y transformación de datos

3

Análisis y conclusiones finales

# Exploración del dataset

# Exploración del dataset



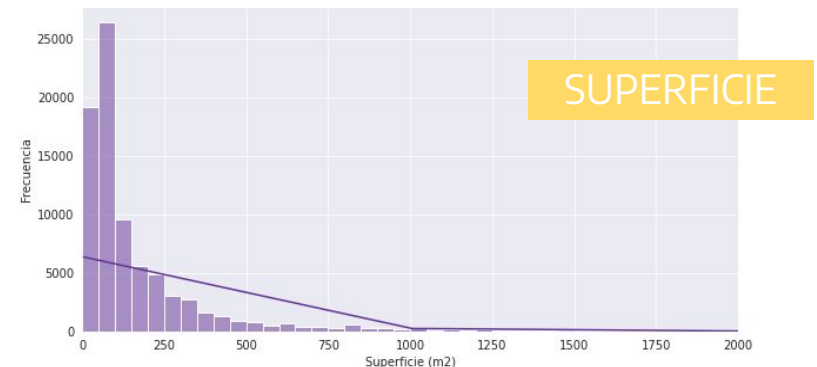
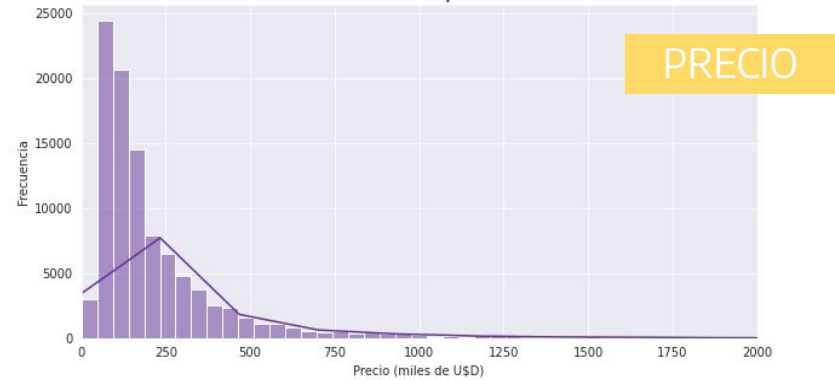
# Exploración del dataset: variables cuantitativas

Medidas de dispersión y de tendencia central

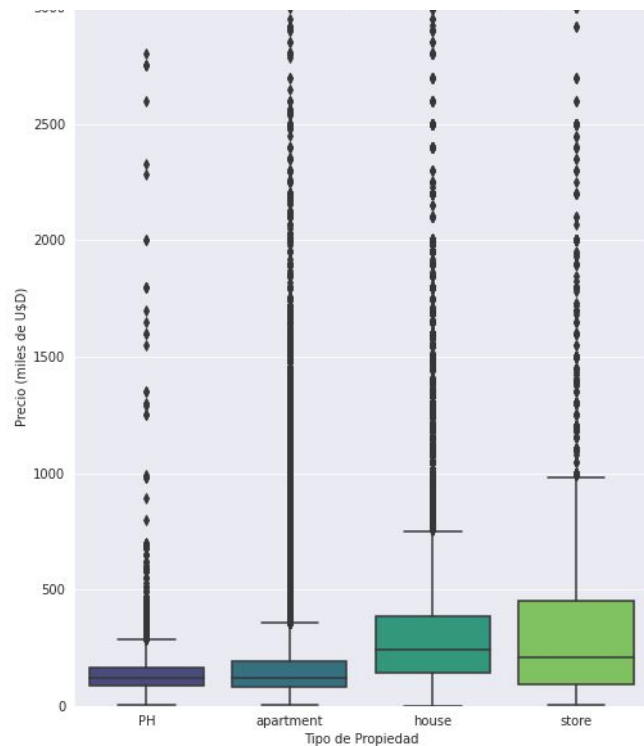
	Precio (U\$D)	Sup. T (m2)	Ambientes	Sup. C (m2)
count	100.810	81.892	47.390	68.617
promedio	239.701	234	3	2.160
desvío estándar	391.324	1782	2	2.759
min	0	0	1	1
1Q 25%	89.734	50	2	1.218
2Q 50% mediana	145.000	84	3	1.800
3Q 75%	265.000	200	4	2.486
max	46.545.445	200.000	32	206.333

# Exploración del dataset: variables cuantitativas

- la distribución de las variables presenta una gran dispersión, con **valores atípicos muy extremos**
- la distribución es **asimétrica sesgada a la derecha** (los valores atípicos más extremos están ubicados a la derecha de la curva, quedando la moda y la mediana a la izquierda de la media)
- la dispersión en la distribución varía según el **tipo** de inmueble y su **localización**



# Exploración del dataset: distribución variables cuantitativas

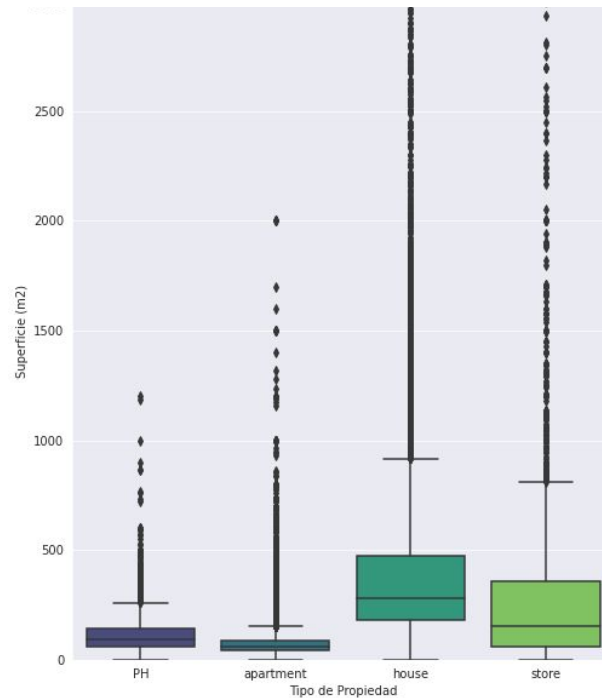
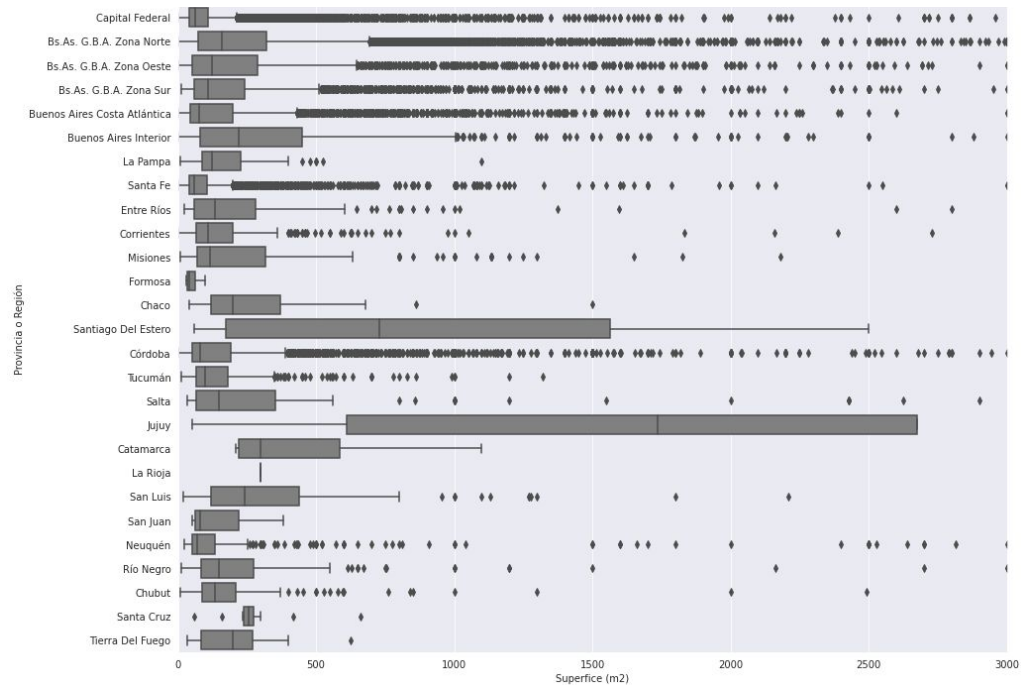


PRECIO



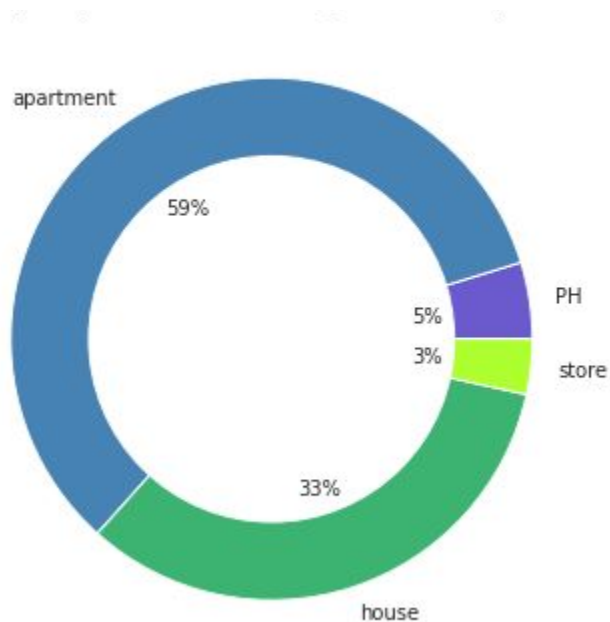
# Exploración del dataset: distribución variables cuantitativas

## SUPERFICIE

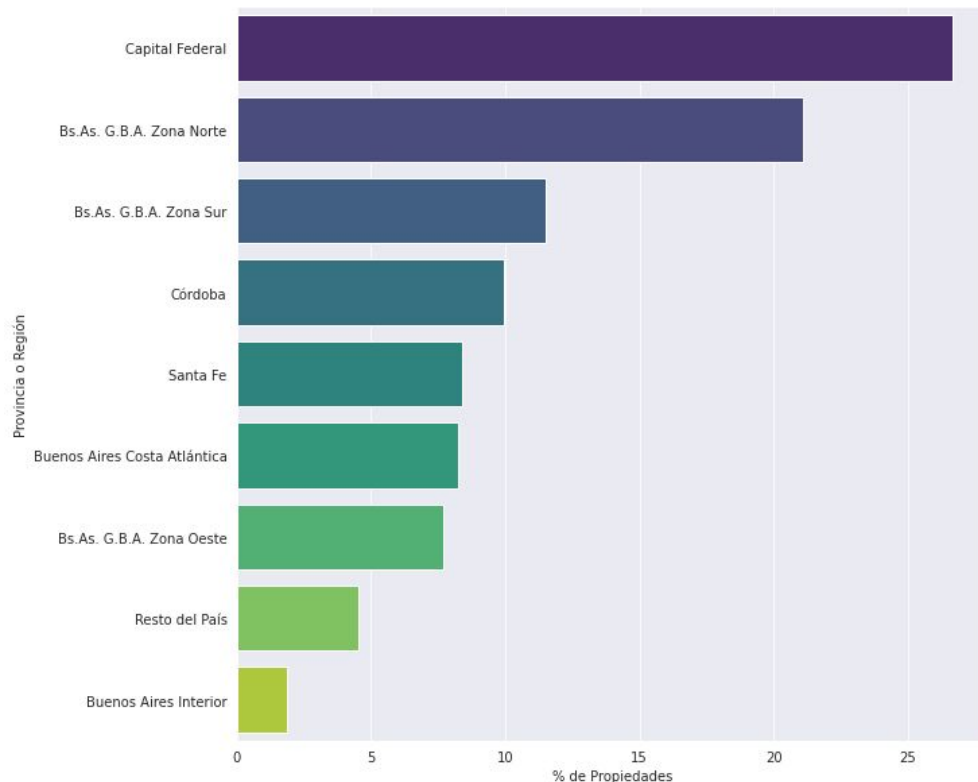


# Exploración del dataset: distribución variables cualitativas

## Propiedades según tipo

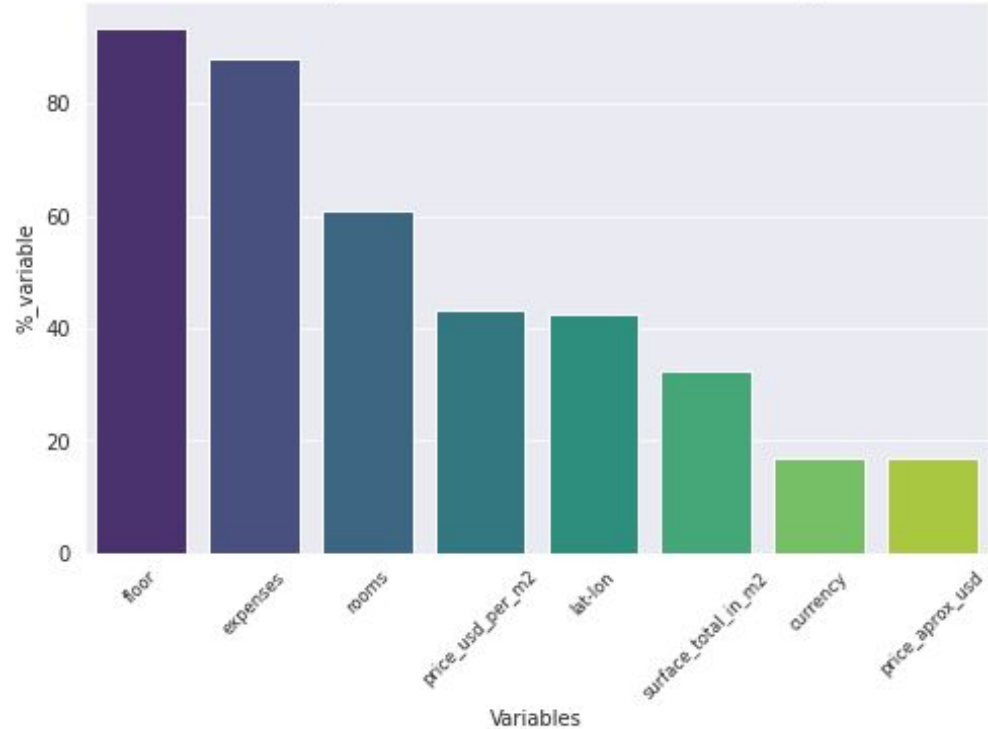


## Propiedades por provincia / región

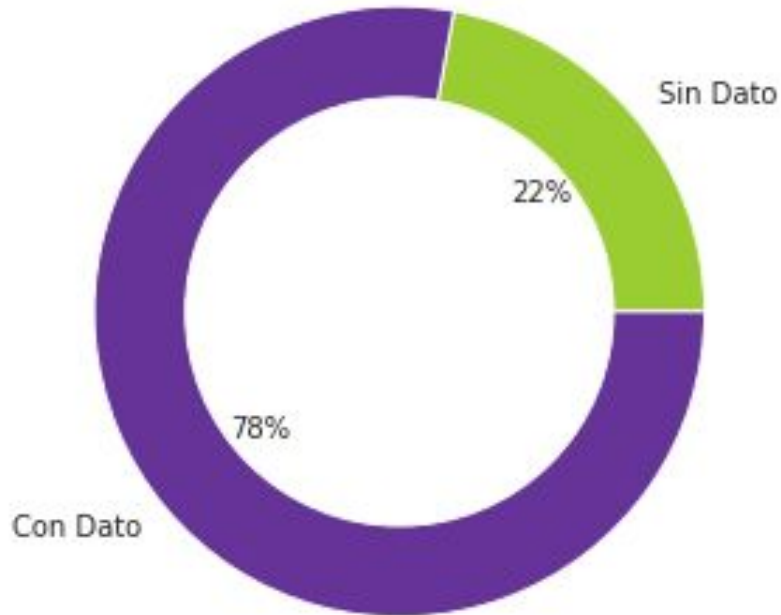


# Exploración del dataset: *missing*

	cantidad	%_missing	%_variable
floor	113321	16.24	93.48
expenses	106958	15.33	88.23
rooms	73830	10.58	60.91
price_usd_per_m2	52603	7.54	43.39
lat-lon	51550	7.39	42.53
surface_total_in_m2	39328	5.64	32.44
currency	20411	2.93	16.84
price_aprox_usd	20410	2.93	16.84



# Exploración del dataset: *missing*



- La mayor parte de los missing se concentra en las variables número de piso, monto de expensas y cantidad de ambientes.
- Si bien la localización de los inmuebles se encuentra completa en cuanto a la localidad, provincia y país, el dataset no cuenta con datos completos en relación a la georreferenciación de su ubicación.
- En cuanto a los precios, la mayor proporción de datos faltantes corresponde a los precios por metro cuadrado, por la falta de datos en al menos alguna de las variables a partir de las cuales se calcula (precio / superficie).
- El resto de las variables de precio tienen la misma proporción de datos faltantes que la de moneda, ya que las expresiones de los precios en moneda local y en dólares están calculadas a partir de ésta.

## Exploración del dataset: objetivos del proceso de *data wrangling*

---

- Imputación de datos en los campos faltantes en nuestras variables de interés cuando ésto sea posible
- Corrección o reimputación de datos en los campos en que se encuentren incongruencias
- Estandarización de los tipos de datos para facilitar su análisis, visualización y posterior modelización

1

Exploración del dataset

2

Limpieza y transformación de datos

3

Análisis y conclusiones finales

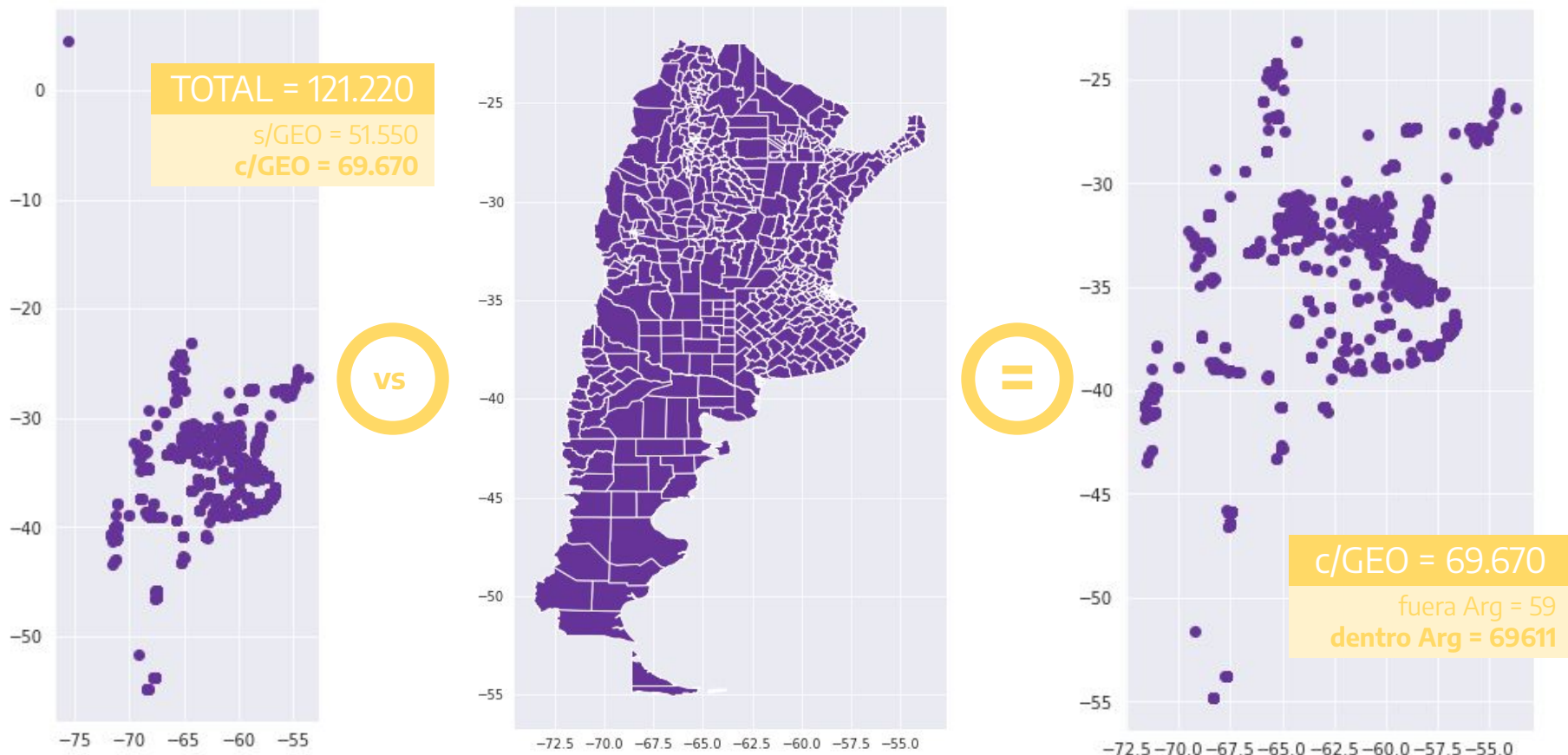
# Limpieza y transformación de datos

# Limpieza y transformación de datos. Localización

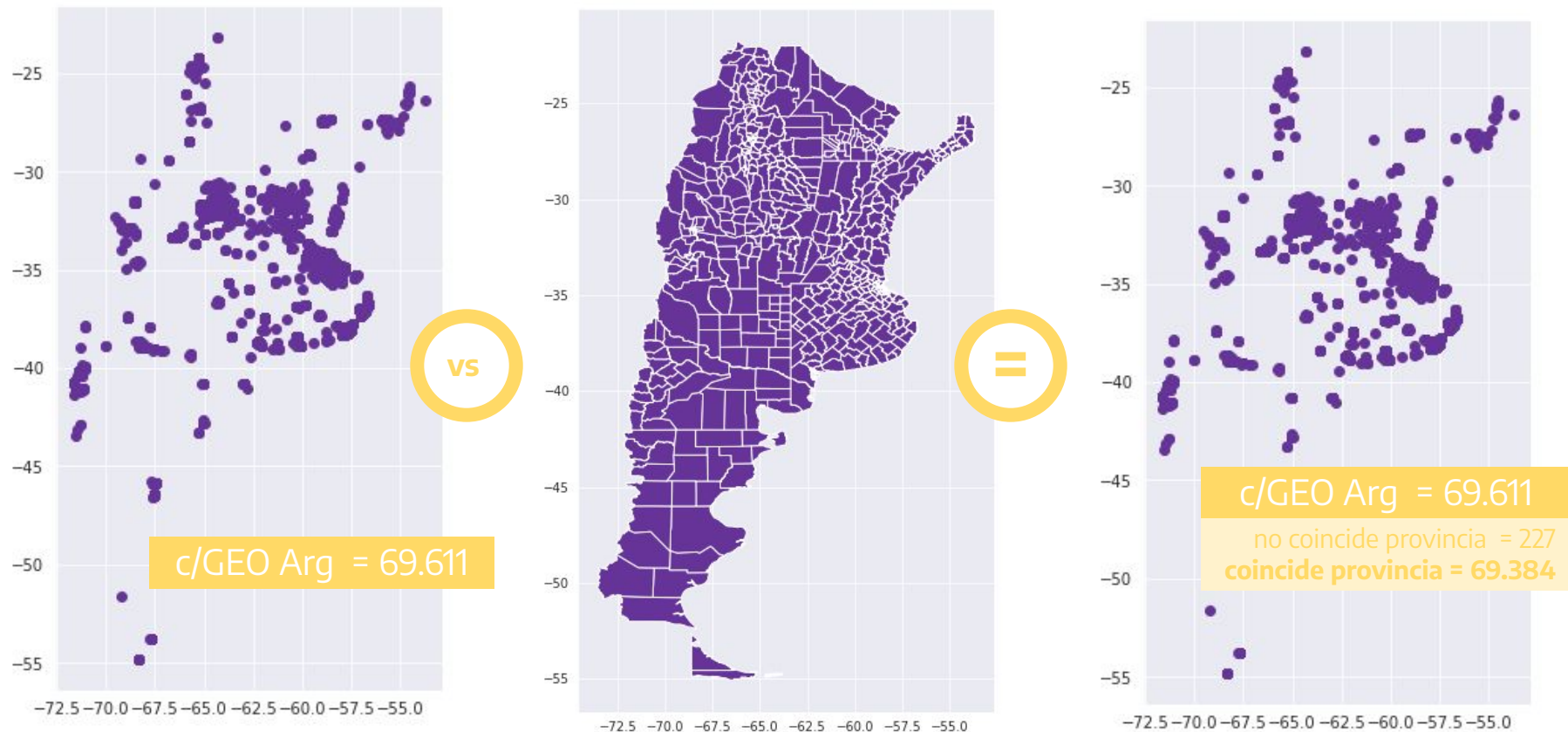
	QUÉ HICIMOS	Variables utilizadas	<i>missing</i> antes	<i>missing</i> después
<b>Ambientes</b>	buscamos los datos faltantes con patrones para identificar ambientes, dormitorios y monoambientes	• Título • Descripción	73.830	29.184
<b>Superficie Total</b>	buscamos los datos faltantes con patrones para identificar m2 cubiertos y totales	• Título • Descripción	39.328	38.021
<b>Superficie Cubierta</b>			19.907	19.836
<b>Nro de Piso</b>	corregimos inconsistencias y buscamos los datos faltantes con patrones para identificar n° de piso	• Título • Descripción	113.331	96.820
<b>Precio</b>	redujimos la cantidad de monedas en que se expresan los precios a dos	Precio • Moneda • Pr. en moneda local • Pr. en U\$D	20.410	20.410
<b>Precio x m2</b>	calculamos los datos faltantes para las observaciones en las cuales disponíamos de información para hacerlo	• Precio en U\$D • Superficie total	52.603	51.329
<b>Expensas</b>	buscamos los datos faltantes con patrones para identificar expensas discriminando por tipo de propiedad	• Título • Descripción	106.958	68.734



# Limpieza y transformación de datos. Localización: cruce 1



## Limpieza y transformación de datos. Localización: cruce 2



# Limpieza y transformación de datos. Localización

	CANTIDAD	REF
<b>Cantidad de observaciones sin coordenadas geográficas</b>	<b>51.550</b>	1
<b>Cantidad de observaciones con coordenadas geográficas</b>	69.670	2
• <b><i>fuera</i></b> del polígono del país (2.1)	59	2.1
• <b><i>dentro</i></b> del polígono del país (2.2)	69.611	2.2
- en provincias diferentes a las declaradas (2.2.i)	227	2.2.1
- en provincias que coinciden con las declaradas (2.2.ii)	69.384	2.2.2
<b>Cantidad de observaciones sin coordenadas geográficas o con coordenadas fuera del polígono</b>	<b>51.609</b>	1 + 2.1
<b>Cantidad de observaciones sin coordenadas geográficas, con coordenadas fuera del polígono o diferentes a la localización declarada</b>	<b>51.836</b>	1 + 2.1 + 2.2.1

1

Exploración del dataset

2

Limpieza y transformación de datos

3

Análisis y conclusiones finales

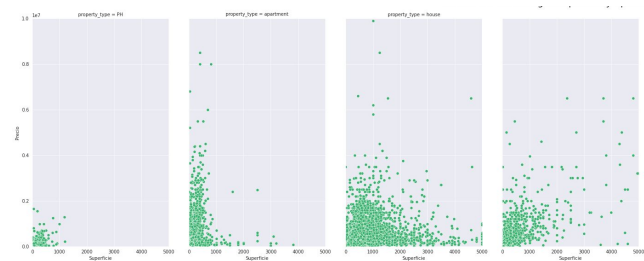
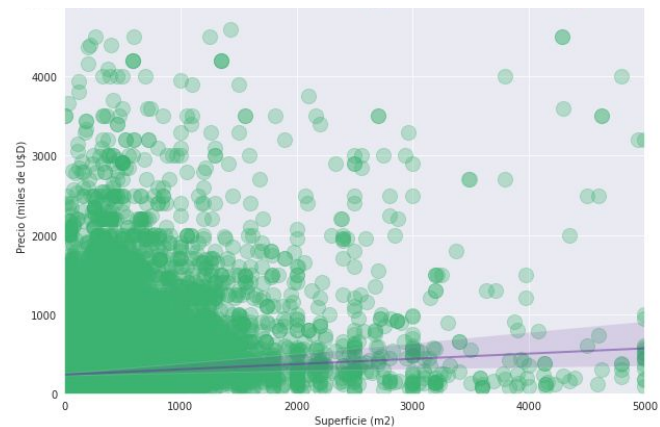


# Análisis y conclusiones finales

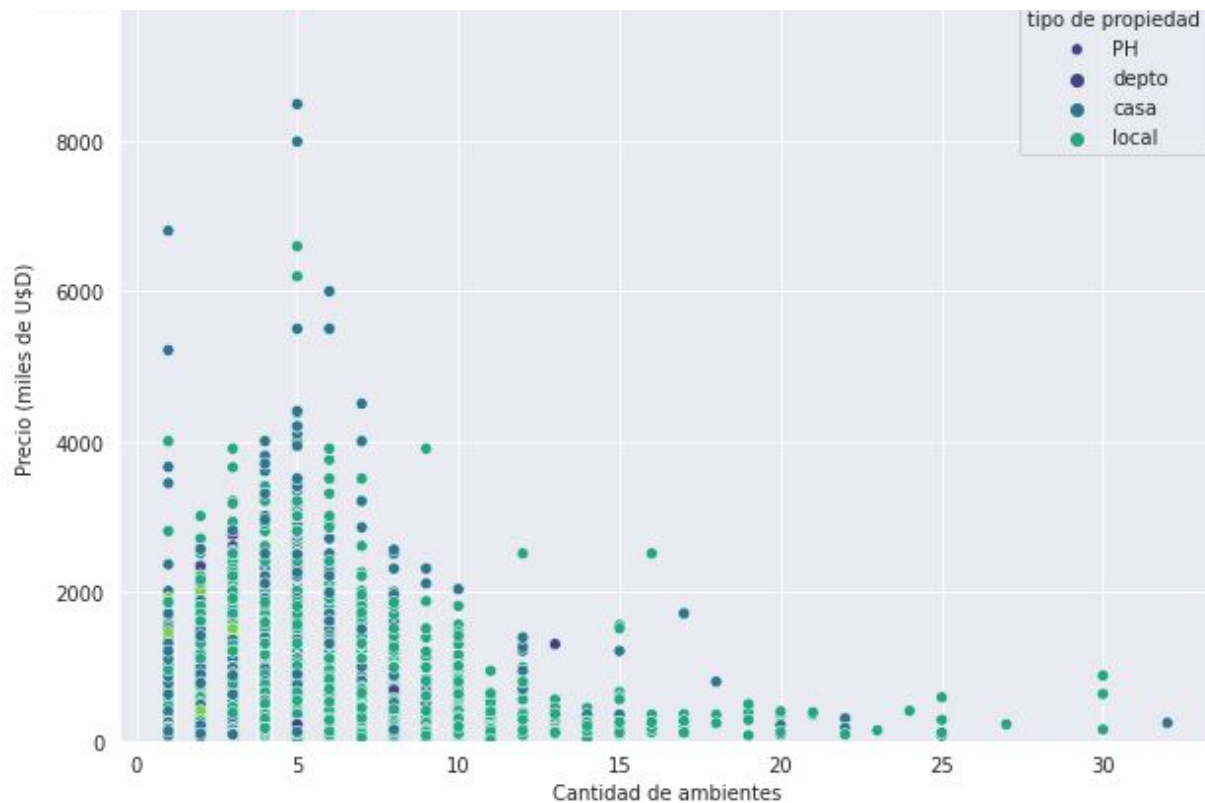
---

	<i>missing antes</i>	<i>missing después</i>
<b>Ambientes</b>	73.830	29.184
<b>Superficie Total</b>	39.328	38.021
<b>Superficie Cubierta</b>	19.907	19.836
<b>Nro de Piso</b>	113.321	96.820
<b>Precio</b>	20.410	20.410
<b>Precio x m2</b>	52.603	51.329
<b>Expensas</b>	106.958	68.734
<b>Coord. Geográficas</b>	51.550	51.836

# Análisis y conclusiones finales: precio según superficie



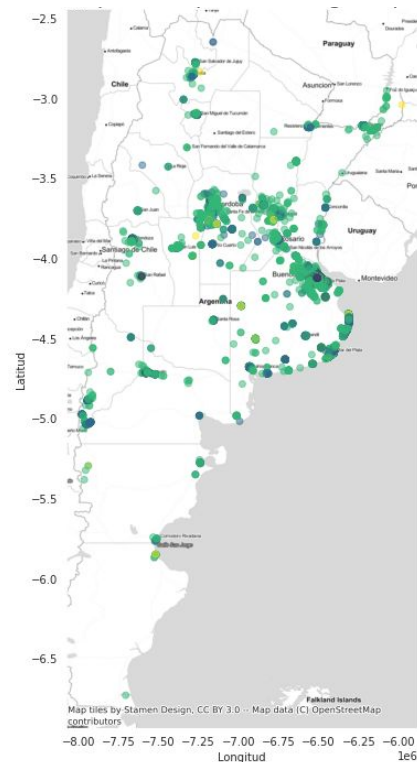
# Análisis y conclusiones finales: precio según cantidad de ambientes





# Análisis y conclusiones finales

	PRECIO (U\$D)				SUPERFICIE TOTAL (m2)			
	ph	dpto	casa	local	ph	dpto	casa	local
<b>CABA</b>	189.433	245.347	501268	514980	129	88	461	299
<b>GBA Norte</b>	136.045	201.113	426327	550361	114	87	493	734
<b>GBA Oeste</b>	110.475	114.466	218530	404753	114	65	390	636
<b>GBA Sur</b>	110.617	124.159	250338	392463	118	79	340	870
<b>PBA Costa</b>	89.489	117.381	213872	261157	101	85	502	410
<b>PBA Interior</b>	102.303	148.960	209488	245354	148	89	1336	432
<b>Córdoba</b>	351.598	139.767	292671	289199	88	225	816	632
<b>Santa Fe</b>	81.173	127.895	187405	251078	85	109	372	197



When it's been 7 hours and you still  
can't understand your own code



¡Muchas Gracias!