DATA SCIENCE
GRUPO 6

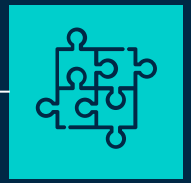
- Laura Martinez Cardona
- Juan Marin
- Luisa Correa
- Rafael Figueredo
- Alejandro Pacheco
- Ignacio Neves

GENERALIDADES

1. Librerías.
2. Lectura y descripción de datos.
3. Análisis exploratorio de datos.
4. Identificación de outliers.
5. Correlación de variables
6. Reducción de dimensionalidad del dataset.
7. Definición de nuevo dataset.
8. Visualización de dataset.
9. Conclusiones.

.

PASO A PASO



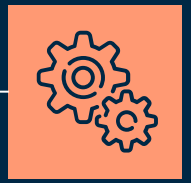
01

DESCRIPCIÓN

Importar Librerías.

Lectura y descripción de datos.

Análisis exploratorio de datos.



02

DESARROLLO

Identificación de outliers.

Correlación de variables

Reducción de dimensionalidad del dataset.

03

CONCLUSIONES

Definición de nuevo dataset.

Visualización de dataset.

Conclusiones.

1.LIBRERÍAS

Pandas - Manejo de Data frame

Matplotlib - Visualización

Seaborn - Visualización

Missingno - Visualización valores faltantes

2. LECTURA Y DESCRIPCIÓN

El primer paso que se realizó para poder acercarnos de manera más detallada al dataset y analizarlo de manera correcta, es permitir la visualización de las columnas, en el gráfico se ve la cantidad de datos nulos y el tipo de dato que se encuentra , también el nombre de estas y que columnas no aportan mucho en el desarrollo final de la limpieza

```
Data columns (total 26 columns):
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   Unnamed: 0                 121220 non-null  int64
 1   operation                  121220 non-null  object
 2   property_type              121220 non-null  object
 3   place_name                 121197 non-null  object
 4   place_with_parent_names    121220 non-null  object
 5   country_name               121220 non-null  object
 6   state_name                 121220 non-null  object
 7   geonames_id                102503 non-null  float64
 8   lat-lon                    69670 non-null   object
 9   lat                        69670 non-null   float64
 10  lon                        69670 non-null   float64
 11  price                      100810 non-null  float64
 12  currency                   100809 non-null  object
 13  price_aprox_local_currency 100810 non-null  float64
 14  price_aprox_usd            100810 non-null  float64
 15  surface_total_in_m2        81892 non-null   float64
 16  surface_covered_in_m2      101313 non-null  float64
 17  price_usd_per_m2           68617 non-null   float64
 18  price_per_m2               87658 non-null   float64
 19  floor                      7899 non-null    float64
 20  rooms                      47390 non-null   float64
 21  expenses                   14262 non-null   float64
 22  properati_url              121220 non-null  object
 23  description                121218 non-null  object
 24  title                      121220 non-null  object
 25  image_thumbnail            118108 non-null  object
dtypes: float64(13), int64(1), object(12)
```

3. ANÁLISIS EXPLORATORIO

121.220 FILAS X 26 COLUMNAS

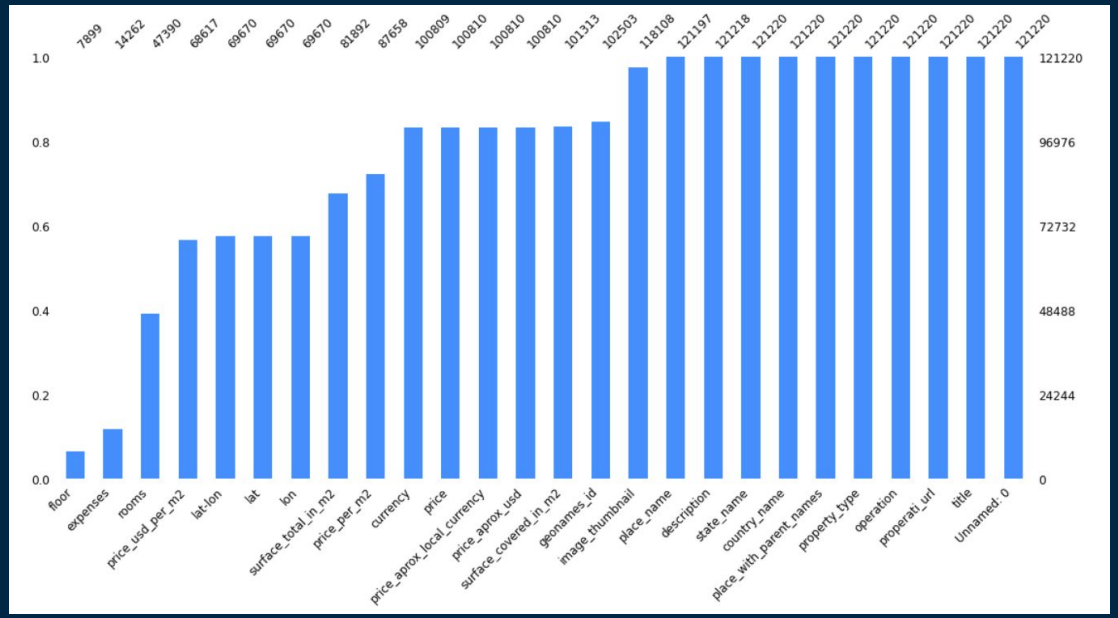
TIPOS DE DATOS

50% Datos categóricos
50% Datos numéricos

Variable: PRICE_USD_PER_M2

Count	68617	Min	0.60
Mean	2160.08	25%	1218.18
Std	2759.28	50%	1800
		75%	2486.41
		Max	206333

3. ANÁLISIS EXPLORATORIO

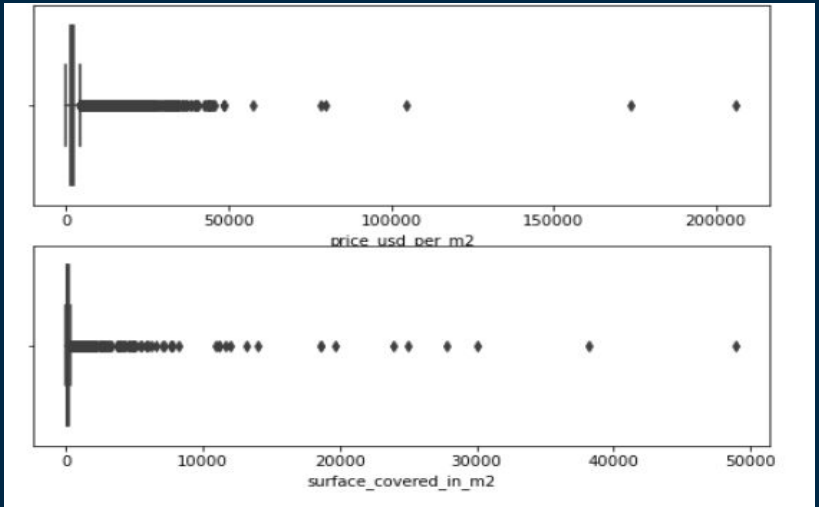


DATOS FALTANTES

Por medio de missingno visualizamos datos totales y faltantes en cada columna disponible

4. IDENTIFICACIÓN DE OUTLIER

Los Outliers que se encontraron fueron valores significativamente distintos unos de otros; en este dataset las variables están generalmente relacionadas, pudiéndose encontrar o no valores atípicos alejados entre ellos.

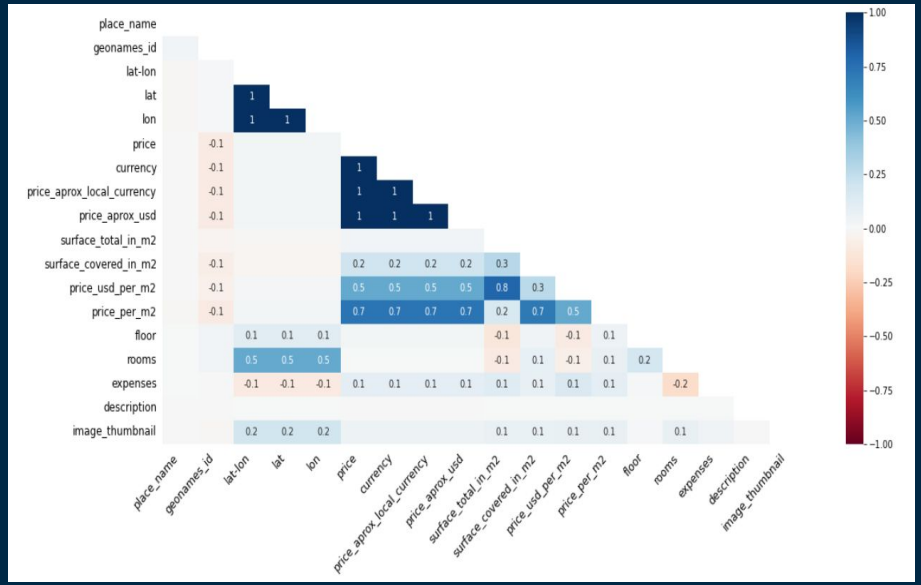


Como puedes observar, dibujamos un gráfico Boxplot donde mostramos la concentración de la media y un valor atípico fuera del rango intercuartílico.

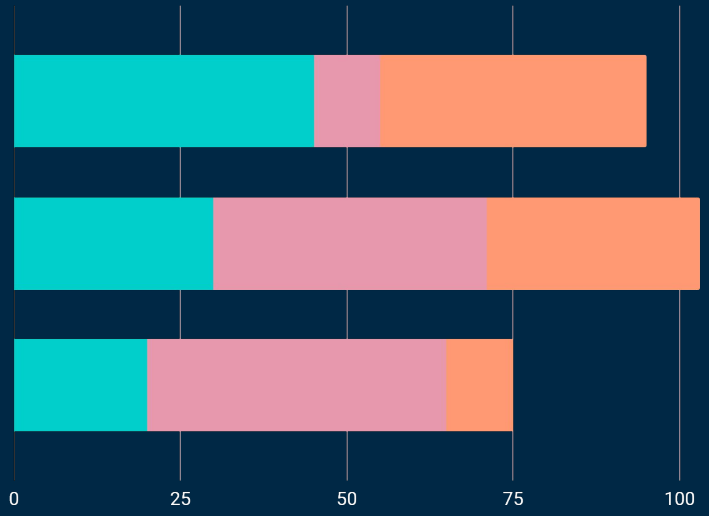
5. CORRELACIÓN DE VARIABLES

Según el mapa de calor y analizando las variables se opta por :
Identificar 5 variables que tienen la misma información, con coeficiente de correlación de "1", será seleccionada price_usd_per_m2 ya que da amyor información y estándar (USD/m2). Por lo tanto se eliminan las siguientes columnas

price
currency
price_aprox_local_currency
price_aprox_usd



6. REDUCCIÓN DE DIMENSIONALIDAD



ELIMINACIÓN DE COLUMNAS

Bajo el criterio de cantidad de datos faltantes y su importancia en el modelo

Unnames: 0, operation, place_with_parent_names, properati_url, description, title, image_thumbnail, country_name, floor, expenses...

Con base a la correlación de variables cuantitativas.

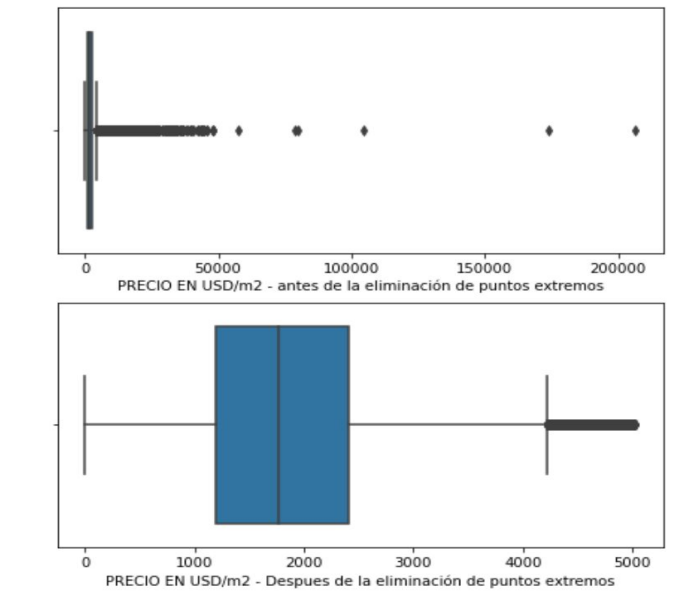
Price, currency, price_aprox_usd, price_aprox_local_currency

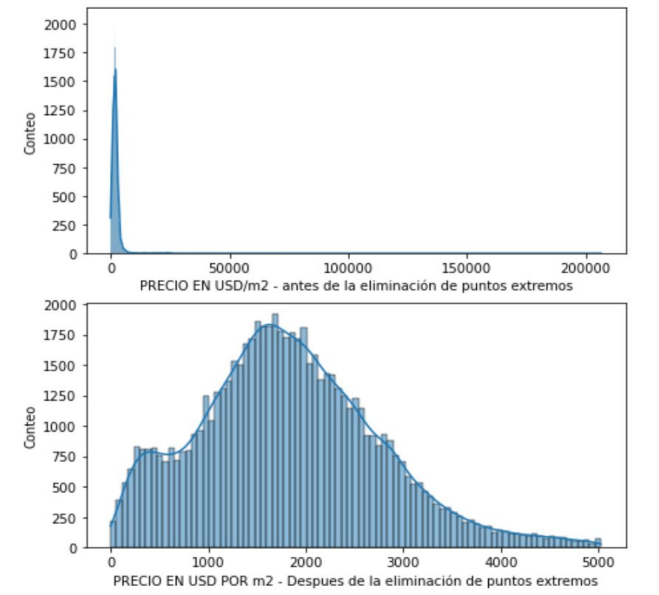
7. DEFINICIÓN DE NUEVO DATASET

Después de depurar las columnas podemos contar con un nuevo dataset, donde se visualiza una información más concreta y limpia.

	property_type	place_name	state_name	price_aprox_usd	surface_total_in_m2	surface_covered_in_m2	price_usd_per_m2	price_per_m2
0	PH	Mataderos	Capital Federal	62000.0	55.0	40.0	1127.272727	1550.000000
2	apartment	Mataderos	Capital Federal	72000.0	55.0	55.0	1309.090909	1309.090909
4	apartment	Centro	Buenos Aires Costa Atlántica	64000.0	35.0	35.0	1828.571429	1828.571429
6	PH	Munro	Bs.As. G.B.A. Zona Norte	130000.0	106.0	78.0	1226.415094	1666.666667
7	apartment	Belgrano	Capital Federal	138000.0	45.0	40.0	3066.666667	3450.000000

8. VISUALIZACIÓN DEL DATASET

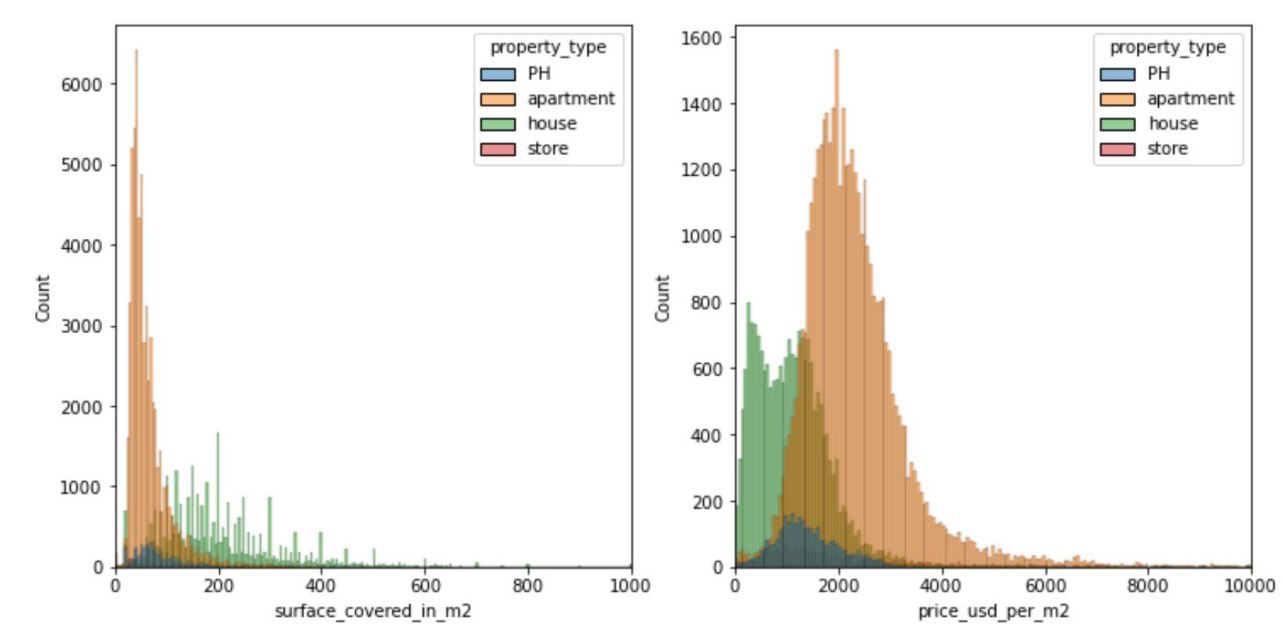




Visualizamos el comportamiento de las variables en el dataset limpio, comparamos el comportamiento de price_usd_per_m2 antes y después de la eliminación de outliers.

Observamos el comportamiento de price_usd_per_m2 por tipo de propiedad en un rango entre 0 y 10.000 usd/m2

8. VISUALIZACIÓN DEL DATASET



Histogramas para las variables “Surface_covered_in_m2” y “Price_usd_per_m2” después de la limpieza de datos

9. CONCLUSIONES

1.En la limpieza de datos, La selección de variable y búsqueda de datos extremos fueron las actividades más importantes, ya que debíamos pensar a futuro qué finalidad va tener el dataset.

2. La visualización de datos nos permite identificar variables con poca información para su eliminación, aunque inicialmente precian “interesantes” para incluir en los modelos

3. La librería Missingno ayudó a visualizar los datos nulos de cada variable gráficamente, lo que nos permitió darnos cuenta que variables nos podrían afectar en el futuro modelo.

4. La librerías que permiten visualizar los Outlier gráficamente, nos permitieron identificarlos mucho más fácilmente.

5. La limpeiza de datos es un proceso impresindible para el posterior analisis de datos y desarrollo de modelos

GRACIAS