# **Checkpoint 1 - Grupo 10**

## **Análisis Exploratorio**

El conjunto de datos consta inicialmente de 20 columnas y 460154 entradas. Tras filtrar los datos que nos interesan para este TP, nos quedan 94249 entradas de interés.

Los datos que

Hay varios tipos de datos que nos interesan:

- tenemos datos cuantitativos, como las coordenadas o las propiedades, que permiten especificar cada dato y disponer de datos variados y de un conjunto de datos más completo,

- tenemos también datos cualitativos que permiten categorizar y especificar cada propiedad con diversas etiquetas.

**Preprocesamiento de Datos**

1. Eliminación de columnas

Decidimos eliminar las columnas "place\_l2", "operation" y "property\_currency", que no son útiles porque eran nuestros filtros iniciales y, por lo tanto, sólo tienen un valor en nuestro conjunto de datos. También eliminamos "place\_l5" y "place\_l6" porque estaban vacías.

Por último, decidimos eliminar "id", "property\_title", "start\_date", "end\_date" y "created\_on" porque sólo identifican una única venta y no un grupo de propiedades. No obstante, no estamos cerrados a reintegrar las columnas "start\_date", "end\_date" y/o "created\_on", que podrían ayudar a identificar tendencias mensuales en las ventas de propiedades.

1. Correlaciones

Hemos podido detectar las siguientes correlaciones :

- "property\_rooms" y "property\_bedrooms" : 0.86

- "property\_price" y "property\_rooms" : 0.48

- "property\_price" y "property\_bedrooms" : 0.42

- "property\_surface\_total" y "property\_surface\_covered" : 0.34

- "latitud" y "longitud" : -0.28

1. Nuevos features

Aún no hemos generado ninguna feature, pero está previsto crear una nueva variable, el precio del metro cuadrado por venta, que permitiría realizar nuevos análisis, sobre todo en términos de localización.

1. Valores atípicos

Por el momento, mediante un análisis unidimensional, hemos identificado algunos valores atípicos (superiores a 3 veces el rango intercuartílico) para las variables "property\_surface\_total", "property\_surface\_covered", "property\_rooms", "property\_bedrooms" y "property\_price". En vista de su escaso número, hemos decidido simplemente eliminarlos, pero un análisis multidimensional podría cambiar nuestra forma de proceder.

1. Datos faltantes

Utilizamos tres técnicas diferentes para completar los datos que faltaban.

En primer lugar, para las variables "latitud", "longitud" y "place\_l3" (respectivamente 4,21%, 4,21% y 0,48% de datos faltantes), utilizamos un CSV que delimita geográficamente todos los barrios de Buenos Aires (<https://data.buenosaires.gob.ar/dataset/barrios>) para completar la variable "place\_l3" a partir de las coordenadas y las coordenadas a partir del centro del barrio según el valor de "place\_l3". Este método no es perfecto ni preciso, pero nos permite aumentar nuestros datos de manera coherente. Las entradas que se quedaron sin una de estas variables se suprimieron porque su número era reducido (en torno al 0,6%).

En segundo lugar, suprimimos la columna "place\_l4" (95,97% de datos ausentes) y sustituimos los de "place\_l3" por los de "place\_l4" si existían, ya que sólo contenían divisiones dentro de Palermo.

Por último, hemos utilizado el principio MICE para completar los datos que faltaban en las columnas "property\_rooms" (1,14%), "property\_bedrooms" (11,74%), "property\_surface\_total" (5,54%) y "property\_surface\_covered" (3,8%) a partir de las demás columnas utilizando el principio de regresión e intentando acercarnos lo más posible a la realidad.

**Visualizaciones**

**Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement**

Este gráfico de barras representa la diversidad de los barrios presentes en el conjunto de datos. Sin embargo, podemos ver que algunos barrios, como Palermo y Belgrano, están sobrerrepresentados con relación a los demás, lo que complicará la predicción para algunos barrios menos representados, como Liniers.

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Este diagrama de dispersión pone de manifiesto la correlación positiva entre el número de habitaciones y el número de dormitorios. También podemos ver que la distribución está bastante centrada en valores inferiores a 20, lo que parece coherente dadas las propiedades de Buenos Aires.

Puede que no necesitemos transformar los datos, pero sí eliminar aquellos que parezcan incoherentes (¡como los que tienen 70 dormitorios y sólo 4 habitaciones!). Por tanto, un criterio podría ser que el número de habitaciones sea siempre mayor o igual que el número de dormitorios.

**Clustering**

No hemos tenido tiempo de empezar a implementar la agrupación en clústeres, pero esa es nuestra próxima tarea.

**Estado de Avance**

1. **Análisis Exploratorio y Preprocesamiento de Datos**

**Porcentaje de Avance:** 80%/100%

**Tareas en curso:** Análisis multidimensional de los valores atípicos y transformación de esos datos

**Tareas planificadas:** Creación de nuevas variables

**Impedimentos:** /

1. Exploración Inicial: /
2. Visualización de los datos: /
3. Datos Faltantes: /
4. Valores atípicos: Análisis multidimensional de los valores atípicos y transformación de esos datos
5. Opcional: Creación de nuevas variables

**2. Agrupamiento**

**Porcentaje de Avance:** 0%/100%

**Tareas en curso:** /

**Tareas planificadas:** detallar con qué tareas tienen pensado continuar.

**Impedimentos:** /

## **Tiempo dedicado**

| Integrante | Tarea | Prom. Hs Semana |
| --- | --- | --- |
| Alan Richmond | Filtrado de Datos Armado de Reporte  División de los conjuntos | 5 |
| Alan Mejia | Revisión y corrección de la parte 1 Análisis de outliers | 2 |
| Flavian Ferré | Exploración Inicial  Visualización de Datos  Datos Faltantes  Armado de reporte | 10 |
|  |  |  |