Checkpoint 1 - Grupo 08

Análisis Exploratorio

Análisis exploratorio:
460154 registros 20 columnas

```
RangeIndex: 460154 entries, 0 to 460153
Data columns (total 20 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   id                        460154 non-null  object
 1   start_date                460154 non-null  object
 2   end_date                  460154 non-null  object
 3   created_on                460154 non-null  object
 4   latitud                   419740 non-null  float64
 5   longitud                  419740 non-null  float64
 6   place_l2                  460154 non-null  object
 7   place_l3                  437665 non-null  object
 8   place_l4                  139020 non-null  object
 9   place_l5                  2430   non-null  object
 10  place_l6                  0      non-null   float64
 11  operation                 460154 non-null  object
 12  property_type             460154 non-null  object
 13  property_rooms            368498 non-null  float64
 14  property_bedrooms         344113 non-null  float64
 15  property_surface_total    397813 non-null  float64
 16  property_surface_covered  427916 non-null  float64
 17  property_price            442153 non-null  float64
 18  property_currency         441590 non-null  object
 19  property_title            460154 non-null  object
```

Hipotesis:

L2 = Provincia (Capital Federal, GBA Zona Norte, GBA Zona Sur, Entre Ríos, Neuquen, etc)

L3 = Ciudad / Barrio (Nordelta, Tigre, Rosario, Pilar). Tambien ambiguo, pero filtrando por Capital Federal son los barrios de CABA

L4 = Localidad (Olivos, Nordelta, Jose C. Paz, etc). Un poco ambiguo, los resultados en L2 Capital Federal son Palermo Chico, Hollywood, Soho y Viejo

L5 = Barrio privado (BarrioPortezuelo, Barrio La Alameda, Barrio El Yacht, etc)

Nota: Todos se encuentran en Tigre, no es relevante para nosotros

L6 = Nada. Se descartará

Datos estadisticos:

index	latitud	longitud	Ambientes	Habitaciones	Superficie total	superficie cubierta	Precio
promedio	-34,2	-59,6	3,3	2,3	420,6	9605,0	306327,3
std	3,3	2,9	1,9	1,7	4026,4	3440367,0	4899613,1
min	-54,8	-103,2	1,0	-3,0	-1,0	-3,0	0,0
25,00 %	-34,6	-58,9	2,0	1,0	50,0	45,0	44900,0
50,00 %	-34,6	-58,5	3,0	2,0	90,0	78,0	98000,0
75,00 %	-34,4	-58,4	4,0	3,0	210,0	170,0	199000,0
max	42,6	-35,0	40,0	390,0	200000,0	2147483647,0	1500000000,0

Como primer punto, podemos observar como la latitud y la longitud se encuentran muy distribuidas respecto del promedio, con una desviación de aproximadamente tres grados en ambos casos. Esto nos indica luego de un primer vistazo que la información debe tener errores ya que tres grados en corresponden a un arco de trescientos kilometros, comparable con la distancia de CABA a Rosario. Luego, las habitaciones, cuartos parecen ser correctas y con una distribución similar, pero con valores extraños para minimo y maximo (390 habitaciones esta lejos de ser una casa).

Preprocesamiento de Datos

Detallar las tareas más importantes que realizaron sobre el dataset, les dejamos algunas preguntas cómo guía:

1. **¿Se eliminaron columnas (Nombre de la columna y motivo de eliminación?**
2. **¿Detectaron correlaciones interesantes (entre qué variables y qué coeficiente)?**
3. **¿Generaron nuevos features?**

4. ¿Encontraron valores atípicos?¿Cuáles?¿Qué técnicas utilizaron y qué decisiones tomaron?
5. ¿Qué columnas tenían datos faltantes?

   ¿En qué proporción? ¿Qué se hizo con estos registros?

Visualizaciones

Mostrar dos gráficos realizados: dispersión entre variables, histogramas, heatmaps, etc. que sean descriptivos del problema. Seleccionar aquellos que permitan entender cómo se distribuyen los datos, cómo se relacionan la variables etc.
Comentar brevemente qué se está visualizando en cada caso y por qué los eligieron.

Clustering

Mencionar -si se analizó- la tendencia al clustering y la cantidad apropiada de grupos que se deben formar. Mostrar cómo llegaron a esa conclusión..

Estado de Avance

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

   Porcentaje de Avance: 15%/100%

   Tareas en curso: detallar en qué puntos están trabajando.

   Tareas planificadas: detallar con qué tareas tienen pensado continuar.

   Impedimentos: mencionar temas que sean bloqueantes para ustedes.

   a) Exploración Inicial: detallar tareas pendientes si las hubiera.
   b) Visualización de los datos: detallar tareas pendientes si las hubiera.
   c) Datos Faltantes: detallar tareas pendientes si las hubiera.
   d) Valores atípicos: detallar tareas pendientes si las hubiera.

e) Opcional: si están trabajando en alguna tarea adicional de preprocesamiento detallar aquí

2. Agrupamiento

Porcentaje de Avance: 00%/100%

Tareas en curso: detallar en qué puntos están trabajando.

Tareas planificadas: detallar con qué tareas tienen pensado continuar.

Impedimentos: mencionar temas que sean bloqueantes para ustedes.

Nota: si avanzaron con los puntos 3 y 4 detallar aquí también.

Tiempo dedicado

Indicar brevemente en qué tarea trabajó cada integrante del equipo durante estas semanas.Si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte). Deben indicar el promedio de horas semanales que dedicaron al trabajo práctico.

Integrante	Tarea	Prom. Hs Semana
Testa, Santiago Tomas	Analisis exploratorio	5hs
Ramirez, Jose Israel	Analisis exploratorio	5hs
Pratto, Federico Nicolas		
Torres, Santiago/Danny		