Checkpoint 1 - Grupo 11

Análisis Exploratorio

El dataset original consta de unas 460154 registros y 20 columnas. Todos estos registros resultan ser anuncios de propiedades, ya sea en venta, alquiler o alquiler temporal, no solo en la Argentina, sino incluso en Uruguay y otros países. Además, hay anuncios no solo para casas, PHs y departamentos, sino también para lotes, locales comerciales, depósitos, oficinas, casas de campo, cocheras y otros.

Dentro de los features más destacables del dataset hallamos:

- **operation:** Detalla el tipo de operación bajo la cual está listada cada propiedad Es una variable categórica cuyo valores incluyen: venta, alquiler, alquiler temporal. En este caso particular, sólo nos interesaban las propiedades en venta.
- **property_price:** Indica el precio de una propiedad. Es el valor que deseamos predecir. Es una cuantitativa variable continua.
- **property_currency:** Detalla el tipo de moneda del precio de la publicación Es una variable categórica cuyos posibles valores incluyen: ARS, USD. Solo nos interesaban las propiedades listadas en USD.
- **property_surface_total:** Indica la superficie total de la propiedad. Es una variable cuantitativa continua.
- **property_rooms:** Detalla la cantidad de habitaciones en la propiedad. Es una variable cuantitativa discreta.
- **place_l3**: Renombrada a “neighbourhood”, describe el barrio donde se encuentra la propiedad. Es la variable categórica que más llama la atención, ya que la información acerca de la ubicación suele ser de suma importancia cuando se busca comprar una propiedad.

Supuestos e hipótesis tomados:

- Los datos de superficie están en metros cuadrados.
- Los precios indicados en dólares efectivamente estaban en dólares.

Preprocesamiento de Datos

En el preprocesamiento de datos se decidió eliminar las columnas **place_l4, place_l5, place_l6**. Posterior al filtrado del dataset se decidió por eliminar las columnas **place_l2**, **operation** y **property_currency** debido a que todas estas columnas quedaban solamente con registros de un mismo valor (Capital Federal, Venta y USD respectivamente).

Además se detectó que las variables **property_rooms** y **property_bedrooms** tenían una correlación muy elevada entre ellas, de 0,87. Previo al análisis de outliers e imputación de datos faltantes, las otras variables que resultaron tener una correlación algo elevada son **property_surface_total** y **property_surface_covered**, con un valor de 0,6.

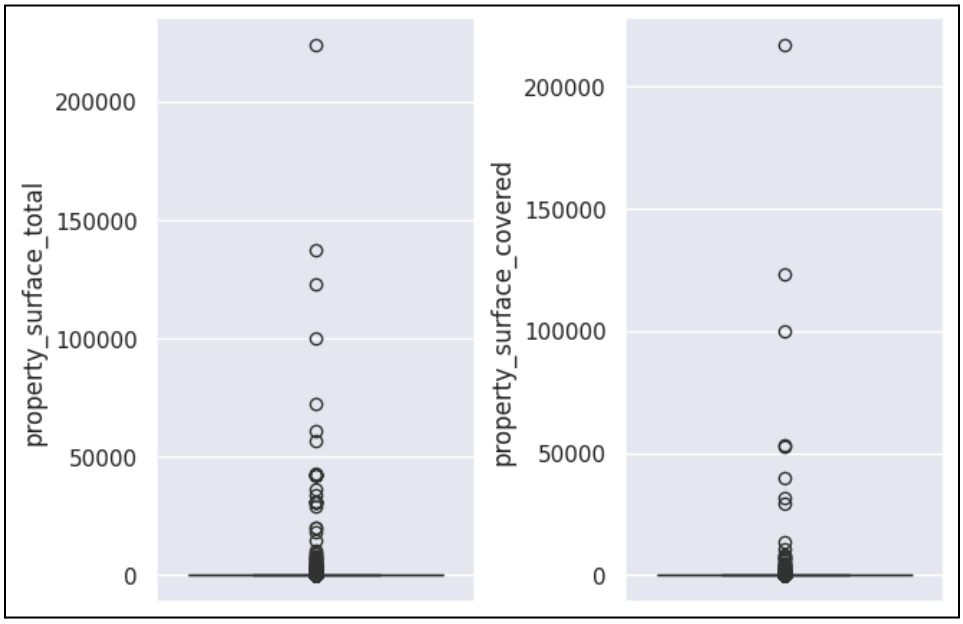
Durante el tratamiento de los análisis nulos de las variables **property_surface_total** y **property_surface_covered** se consideró utilizar un método de imputación basado en un modelo de regresión lineal que predijera el valor de esta última a partir de la superficie total y el número de cuartos. Para mejorar la correlación entre las variables se probó escalando las variables de superficie para minimizar la importancia de las diferencias entre las unidades de medida de las cantidad de cuartos y las superficies. Se terminó hallando que la mejor opción era aplicar el logaritmo a las superficies. Sin embargo, finalmente se terminó optando el método MICE por practicidad, ya que el modelo de regresión también requería de un análisis de nulos adicional para evitar logaritmos de valores nulos o ceros.

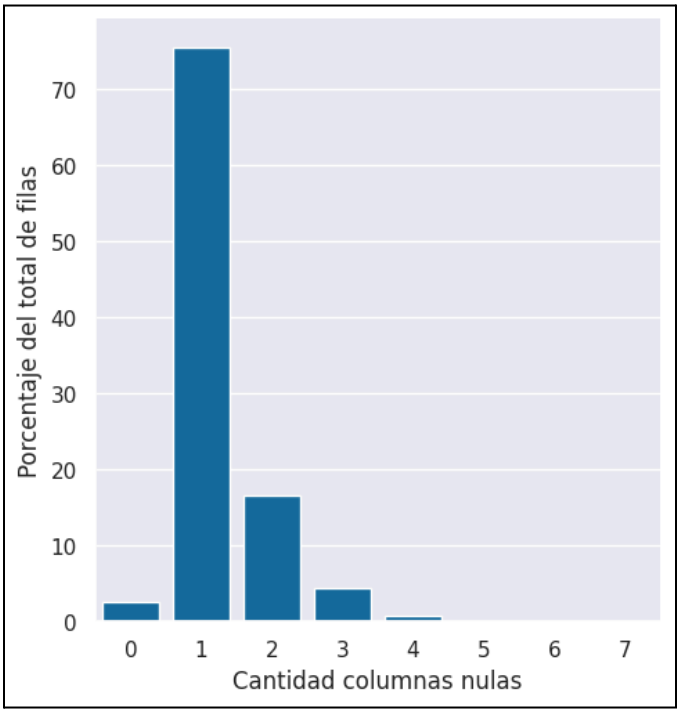
Se encontraron muchos valores atípicos sobre todo en los valores de las superficies, teniendo algunas superficies totales que eran menores que las superficies cubiertas, o valores que resultaban muy grandes y con poco sentido para la cantidad de habitaciones de algunas propiedades (ej. monoambientes con 51100m^2 de superficie cubierta). Algunos se analizaron y se pudieron salvar corroborando con la información de optó por utilizar las técnicas de boxplots, iqr y scatterplots para el análisis multivariado.

Luego de analizar los outliers y realizar imputación, se detectó que las variables **property_surface_total** y **property_surface_covered** incrementaron su correlación a 0,93.

Visualizaciones

Los gráficos que consideramos más descriptivos del problema fueron los siguientes dos:





Los boxplots muestran una gran cantidad de outliers extremos para las columnas de superficie total y superficie cubierta de los anuncios. Esta situación se dio también con la cantidad de habitaciones y dormitorios. Este tipo de gráficos nos permitieron identificar ese problema, que luego pudimos resolver modificando datos puntuales y deshaciéndonos del resto.

Por su parte, el gráfico de barras muestra otro de los problemas con el que nos enfrentamos en varios lados, que era el de los datos nulos o NaNs. Se puede ver en el gráfico que un 75% de las filas tienen al menos una columna nula, y las filas que tienen todos los datos representan un porcentaje muy pequeño del total. Este era un problema que atravesaba todos los datos, y lo resolvimos aplicando diferentes técnicas de imputación como MICE e imputación Cold-Deck. En algunos pocos casos restantes, también recurrimos a eliminarlos directamente.

Clustering

Pudimos descubrir mediante la estadística de Hopkins que el dataset presenta una fuerte tendencia al clustering, con un valor de alrededor de 0.000276 (para la librería 'pyclustertend' cuanto más cercano a 0 es más fuerte es la tendencia al clustering). La cantidad apropiada de grupos que se deben formar resultó ser 2, llegamos a esta cantidad de grupos analizando el score de Silhouette para valores desde 2 clusters hasta 10 clusters y analizando y comparando el score de cada grupo. Este score cae cuánto más se incrementa la cantidad de grupos, siendo como se mencionó antes el más elevado de 0,75 aproximadamente para 2 grupos.

Estado de Avance

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

   Porcentaje de Avance: 90%/100%

   Tareas en curso: Análisis de relación entre precio de venta y superficie.

   Tareas planificadas: Análisis de relación entre precio de venta y superficie.

   Impedimentos: -.

2. Agrupamiento

   Porcentaje de Avance: 90%/100%

   Tareas en curso: Ajuste de gráficos.

   Tareas planificadas: Ajuste de gráficos.

Impedimentos: -.

Nota: En la sección de agrupamiento, la celda en la que se calcula el coeficiente de Silhouette tarda alrededor de 15 minutos en ejecutarse.

Tiempo dedicado

Integrante	Tarea	Prom. Hs Semana
Carlos Castillo	Exploración de datos Análisis de Correlaciones Análisis de Valores Faltantes Detección y tratamiento de Outliers Armado de Reporte	10
Juan Pablo Destefanis	Análisis de Valores Faltantes Imputación de Datos Detección y tratamiento de Outliers Clustering Armado de Reporte	8
Celeste Gómez	Análisis de Valores Faltantes Imputación de Datos Detección y tratamiento de Outliers Armado de Reporte	8
Facundo Agustín Polech		