Checkpoint 1 - Grupo 06

Análisis Exploratorio

Descripción del dataset

El dataset provisto consiste de una lista de anuncios de propiedades en venta o en alquiler, generado en el año 2021 por la empresa Properati. Las propiedades se encuentran mayormente en Argentina, pero es posible identificar algunas de ellas ubicadas en Uruguay, Brasil, e incluso Estados Unidos. Se incluyen desde avisos de casas y departamentos, hasta locales comerciales y casas de campo.

El dataset completo se compone de 460.154 registros, con información distribuida en 20 columnas.

Las columnas están comprendidas por:

- ***id*:** cadena de caracteres única para identificar cada publicación.
- ***start_date:*** fecha en la que el anuncio se abre para la venta/alquiler. Coincide con la columna *created_on*.
- ***end_date*:** fecha en la que el anuncio se cierra.
- ***created_on*:** fecha en la que se creó la publicación. Coincide con la columna *start_date*.
- ***latitud*:** coordenada de latitud de la propiedad en cuestión.
- ***longitud*:** coordenada de longitud de la propiedad en cuestión.
- ***place_l2*:** detalle de la provincia/departamento donde está la propiedad.
- ***place_l3*:** detalle del partido/barrio/pueblo donde está la propiedad.
- ***place_l4*:** detalle de localidades en un partido.
- ***place_l5*:** detalle para barrios cerrados.
- ***place_l6*:** subdivisión adicional. No es usada por ninguna propiedad.
- ***operation*:** tipo de contrato del aviso: venta, alquiler, o alquiler temporal.
- ***property_type*:** tipo de inmueble del aviso.
- ***property_rooms*:** cantidad de ambientes detallados en el aviso.
- ***property_bedrooms*:** cantidad de dormitorios detallados en el aviso.
- ***property_surface_total*:** superficie total que ocupa el terreno donde se encuentra la propiedad.

- ***property_surface_covered*:** superficie cubierta correspondiente a la propiedad.
- ***property_price*:** precio de la propiedad detallado en el aviso.
- ***property_currency*:** moneda del precio de la propiedad.
- ***property_title*:** título del aviso cargado en la página web.

La consigna del trabajo práctico pide trabajar con una muestra del dataset original, que deberá incluir únicamente ofertas de venta en dólares de PH, casas, o departamentos, ubicados en Capital Federal.

Con estos parámetros, la cantidad de registros del dataset se reduce a **92.249**.

Por último, dividimos el dataset en conjuntos de entrenamiento y de prueba, de forma que se pueda desarrollar el modelo con uno de ellos, y luego evaluar el rendimiento del modelo generalizado a un conjunto de datos independiente.

El criterio para la división fue de 80% (**75.399**) de los registros para el conjunto de entrenamiento, y el 20% (18.850) restante para el conjunto de prueba.

Preprocesamiento de Datos

Selección de variables

Un análisis inicial del contenido del dataset revela que algunas de las columnas no son relevantes para el modelo de predicción.

- La columna *id* contiene una cadena única para cada registro del dataset. Este valor no muestra ninguna relación aparente con el resto de los datos del registro, por lo que puede omitirse del análisis.
- Las columnas utilizadas para el filtro inicial de la información que contenían un solo dato posible ya sirvieron su propósito y pueden ser retiradas. Estas columnas serían *operation* ('Venta'), *property_currency* ('USD'), y *place_l2* ('Capital Federal').
- Los atributos *start_date*, *end_date*, y *created_on*, representan fechas relevantes de cada aviso en el dataset. Sin embargo, para la predicción del precio de venta de una propiedad, estas variables no resultan significativas.
- Las columnas *place_l5* y *place_l6* permanecen vacías en todos los registros del dataset filtrado, por lo que pueden ser retiradas sin complicaciones.

Luego, las columnas relevantes para el modelo de aquí en adelante son:

latitud, longitud, place_l2 (city), place_l3 (zone), place_l4, property_type, property_rooms, property_bedrooms, property_surface_total, property_surface_covered, property_price, y *property_title*.

Correlaciones entre variables

En cuanto a correlaciones existentes, se planteó un heatmap entre las variables numéricas del problema, respecto al precio de la propiedad.

- La más evidente ocurre con *property_rooms* y *property_bedrooms*. La correlación es fuertemente positiva (0,87) en estos casos, por lo que el precio de una propiedad es propenso a aumentar en base a estos dos atributos.

Para profundizar en este aspecto, se generaron algunos features, particularmente ***property_bedrooms_ratio*** y ***property_surface_ratio***, para definir si alguna de estas proporciones influye de alguna manera en el precio de la propiedad.

- El coeficiente de correlación entre habitaciones versus ambientes y el precio resultó prácticamente nulo (0,062126).
- De modo similar, el coeficiente entre superficie cubierta versus superficie total y el precio no permitió sacar mayores conclusiones (-0,001285).

Sin embargo, consideramos que es una evaluación temprana y que tras una posterior limpieza de los datos, se podrían observar correlaciones más significativas.

Datos faltantes

Para tratar los datos faltantes, inicialmente se calcularon los porcentajes de datos nulos por columna. La mayoría de estos muy probablemente se deba a faltas al momento de la carga de datos, ya que no todas las variables son requeridas a la hora de publicar un aviso.

- Las variables mencionadas anteriormente que fueron usadas como filtro, no tienen datos faltantes. A estas se suman *id, property_price,* y *property_title*.
- Variables como *latitud, longitud, zone, property_rooms, property_surface_total,* y *property_surface_covered*, tienen alrededor del 5% o menos de datos faltantes.
- Las dos variables restantes, *place_l4* y *property_bedrooms*, tienen porcentajes de datos faltantes más significativas (96,19% y 11,67% respectivamente), por lo que continuaremos con el análisis de estas dos variables en particular.

Los faltantes en ***place_l4*** se explican ya que esta columna no es más que una subdivisión zonal del barrio de Palermo. Para no mantener esta columna, reemplazamos cada instancia de Palermo en la columna *zone*, por su correspondiente subdivisión.

Para los faltantes de ***latitud***, ***longitud*** y ***zone***, se plantearon múltiples escenarios:

- Si el registro contiene un dato de *zone*, pero no *latitud* y/o *longitud*, se reemplazan los datos de latitud y longitud por la media de todas las filas de la misma zona.
- Si el registro contiene *latitud* y *longitud*, pero no *zone*, se calcula la diferencia entre la ubicación del registro y la ubicación media de cada zona. Y se asignará la zona que presente una menor diferencia respecto de la ubicación.
- Si el registro no contiene *latitud*, *longitud*, ni *zone*, ese dato tendrá que quitarse, ya que no pueden hacerse esas asignaciones basándose únicamente en precios o superficies.

Luego de este filtrado, el número de registros pasa a ser **75.267**.

Para los faltantes en ***property_surface_total***, se le asignará teniendo en cuenta el precio medio por metro cuadrado por zona, dado que el dato de *property_price* no tiene faltantes. Mediante una regla de tres simple, se puede estimar la superficie total.

Para los faltantes en ***property_surface_covered***, se supone que serán una fracción del dato de *property_surface_total*. Entonces se calculó el dato de porcentaje de superficie cubierta en una propiedad, por zona, y se usó para imputar la superficie cubierta en cada fila.

Luego, completamos los datos faltantes en ***property_rooms***, según la información de superficie y habitaciones de otras propiedades de la misma zona. La cantidad de habitaciones en cada propiedad puede estimarse como la superficie cubierta de la propiedad, dividido la superficie promedio por habitación en la zona.

De manera similar, ***property_bedrooms*** tiene que ser una fracción de *property_rooms*. Se calculó el porcentaje de dormitorios dadas las habitaciones, de las propiedades de una misma zona. Se usó este porcentaje para estimar la cantidad de dormitorios por propiedad.

Valores atípicos

A la hora de analizar los valores atípicos, se planteó primero un caso trivial.

- Registros con valores numéricos menores a 0. En un caso particular, la cantidad de habitaciones de una propiedad era negativa. Sin embargo, se consideró un error de tipeo ya que el título del aviso contemplaba este valor como positivo.

La identificación de otros valores atípicos se llevó a cabo gráficamente y mediante sucesivos análisis univariados y multivariados.

Para detectar outliers en *property_rooms* y *property_bedrooms*, se realizaron acercamientos con box plots. A simple vista se podían observar registros con mucho más de 3 IQR de distancia del resto de los datos, para dimensionar la presencia de outliers.

Luego se calcularon los correspondientes Z-score y Z-score modificado de *property_rooms* y *property_bedrooms,* para cada uno de los registros. Se consideraron como outliers aquellos registros que cumplían a la vez con las "**reglas de oro**" de ambas métricas (**Z > 3, Zmod > 3,5**).

- Del lado de *property_bedrooms*, el resultado del análisis arrojó 47 registros para descartar, cuyos valores en esta variable oscilaban entre 75 y 6.028 dormitorios. Luego de eliminarlos, si bien quedaban alrededor de 200 registros con un Z-score modificado mayor a 3,5, la descripción de estas publicaciones mayormente avalaba estos valores elevados.
- Por otro lado, el análisis gráfico de *property_rooms* también mostraba una cantidad excesiva de habitaciones en algunos registros. En base a los Z-score, 52 registros fueron descartados.

Como análisis multivariado, aplicamos el método de distancia de Mahalanobis para las variables *property_bedrooms* y *property_rooms*.

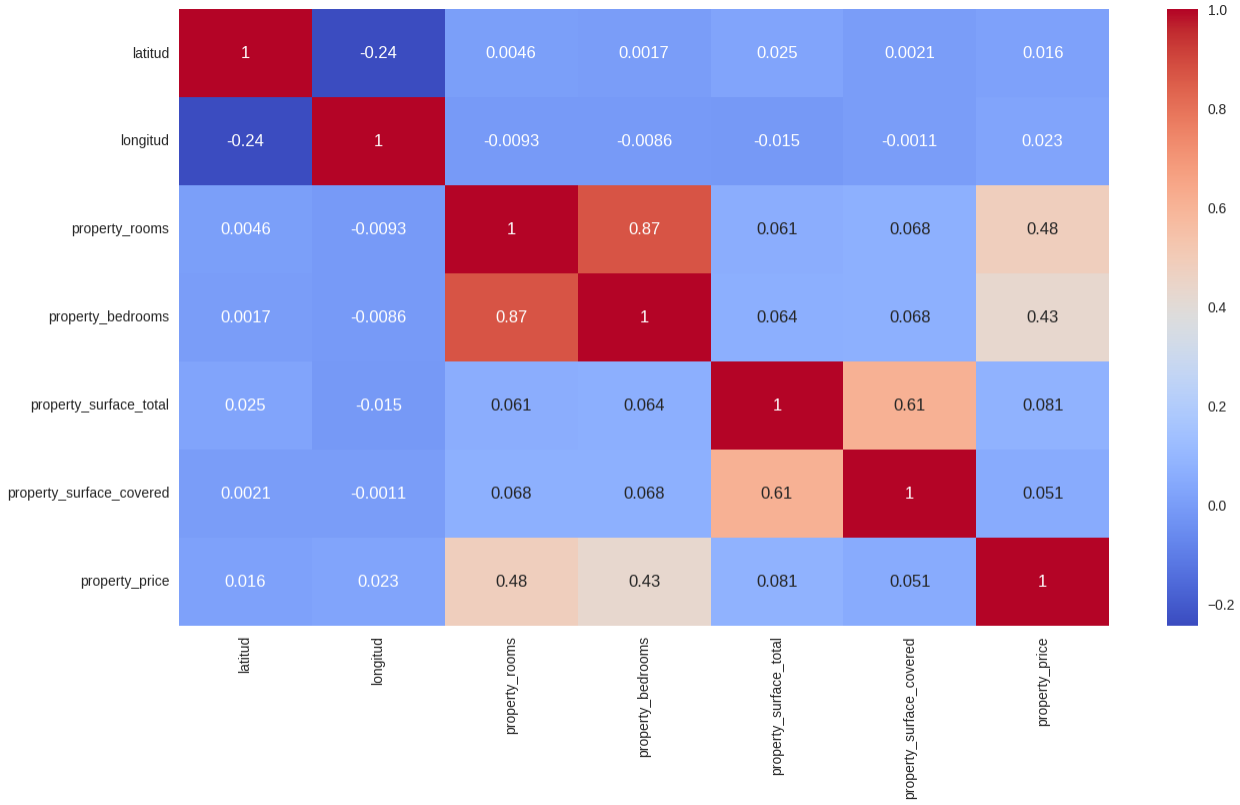
- Un scatter plot inicial revelaba la presencia de outliers, aunque no demasiadas. Un ejemplo particular es una propiedad con más de 70 dormitorios, pero pocas habitaciones.

Tras el cálculo de la distancia para cada registro, un boxplot de esta distancia revelaría una apariencia bastante similar al análisis univariado previo, con algunos outliers extremadamente alejados del rango intercuartílico. Definiendo un valor umbral de 1.100, se aislaron los 5 valores más alejados de la nube de puntos, de los cuales uno presentaba un error de tipeo.

Posteriormente, al analizar del mismo modo la relación entre las variables ***property_price*** y ***property_surface_total***, seguían existiendo valores anómalos sin detectar. Planteando un umbral de 1.000, se pudieron poner bajo observación otros casos bordes, propiedades de muy extensa superficie a precios anormalmente reducidos, y viceversa. Dado que al momento no fue posible definir si se trata de errores de tipeo, o estos valores se deben a estar en zonas particulares, conservamos estos valores en el dataset hasta encontrar otra alternativa.

Visualizaciones

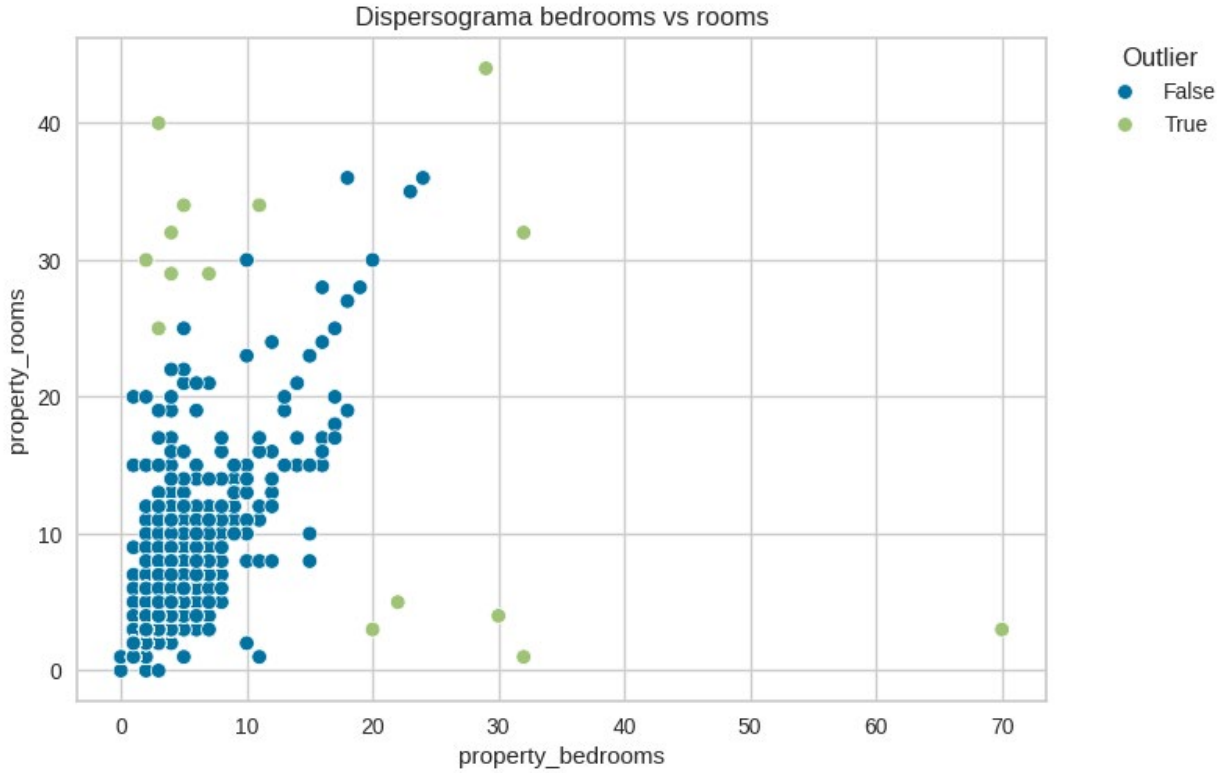
Heatmap de las variables numéricas del dataset, respecto del precio



El gráfico representa las distintas correlaciones entre pares de variables del dataset, respecto del precio de la propiedad. Principalmente se puede observar como *property_rooms* y *property_bedrooms* tienen correlaciones positivas, lo que significa que el precio de la propiedad tiene tendencia a aumentar junto con la cantidad de habitaciones y dormitorios.

La segunda correlación más significativa según el gráfico es la que vincula *property_surface_covered* y *property_surface_total*, indicando que el precio de la propiedad suele aumentar junto a la superficie de ésta, aunque resulta menos influyente que la cantidad de habitaciones.

Scatter plot entre property_rooms y property_bedrooms, distinguiendo outliers según Mahalanobis



El gráfico surgió durante la detección de valores atípicos de forma multivariada. Cada punto representa los valores que cada observación del dataset posee en las columnas *property_rooms* y *property_bedrooms*. La correlación entre estas variables es positiva, teniendo en cuenta que la mayoría del conjunto de datos aumenta las cantidades de forma pareja, con la cantidad de habitaciones superando la de dormitorios consistentemente.

El análisis realizado con las distancias de Mahalanobis permite distinguir cuáles de estos valores tienen una mayor probabilidad de ser outliers. El valor de umbral utilizado para este gráfico es de 500.

Clustering

Para el análisis inicial de clustering, se intentará utilizar solamente las columnas *property_rooms, property_bedrooms, property_surface_total, property_surface_covered,* y *property_price*, de forma tal que los potenciales grupos guarden una relación numérica clara.

Para tener una idea de la cantidad de grupos a utilizar, se planteó el *Elbow method* de forma preliminar. Según su gráfico, obtenido mediante el algoritmo *K-means*, la cantidad recomendada sería de 4 o 5 clusters.

Sin embargo, al analizar la calidad de los grupos con el *Silhouette method*, el score favorece los 8 o 9 clusters.

Esto significa que el dataset presenta algo de ruido o ambigüedad en algunos registros, por lo que antes de proseguir con el análisis de clusters debería realizarse una mayor limpieza en los datos.

Estado de Avance

1. Análisis Exploratorio y Preprocesamiento de Datos

Porcentaje de Avance: 90%/100%

Tareas en curso: -

Tareas planificadas: Análisis adicionales para detectar outliers.

Impedimentos: -

a) Exploración Inicial: -

b) Visualización de los datos: -

c) Datos Faltantes: -

d) Valores atípicos: profundizar en los análisis multivariados.

e) Opcional: estandarizar nombres de variables para su reutilización. Mejoras en el contenido de la notebook para facilitar la lectura.

2. Agrupamiento

Porcentaje de Avance: 10%/100%

Tareas en curso: calcular tendencia a la clusterización del dataset.

Tareas planificadas: reevaluar cantidad necesaria de clusters.

Impedimentos: registros con valores atípicos no identificados.

Tiempo dedicado

Integrante	**Tarea**	**Prom. Hs Semana**
Verónica Leguizamón	Tratamiento de valores faltantes	5
Henry Maldonado	Análisis explorativo Preprocesamiento de datos	5
Augusto Rivera Lofrano	Detección de valores atípicos	5
Alejandro Vargas	Análisis explorativo Armado de reporte	5