Checkpoint 1 - Grupo 02

Análisis Exploratorio

Inicialmente el dataset tiene un total de 20 columnas y 460154 registros, los cuales representan propiedades en venta o en alquiler. Dichas propiedades tienen todos sus datos relevantes como ubicación y tipo de propiedad, precio, entre otros.
Para este análisis se tomaron solamente las propiedades que se ubican en CABA, son de tipo casa, ph o departamento y están a la venta en USD.

Una vez que el dataset fue filtrado con los criterios mencionados este pasó a tener un total de 94249 registros (disminuyó casi por un factor de 5).

Luego del filtrado de datos de interés, se crearon los dataset de training y testeo, para esto se dividió el dataset filtrado en 80% de datos para entrenamiento y el 20% restante para testeo. A partir de esta división y a lo largo de la siguiente etapa (Preprocesamiento de Datos) se usaron estos mismos en lugar del dataset filtrado completo.

`Exploracion Inicial`

En esta etapa del análisis se hizo un vistazo de los datos para encontrar correlaciones entre atributos, ver la forma de los mismos con medidas de resumen, y una primera visualización de estos (utilizando el training dataset).

Primeramente se observaron distintas medidas a modo de resumen de las variables cuantitativas del dataset, donde se puede ver por ejemplo la desviación estándar de la latitud y longitud la cual es mínima ya que todas las propiedades están en CABA, así como también se pueden identificar outliers (aunque estos serán analizados más profundamente en la sección de datos atípicos) en los atributos de “rooms” y “bedrooms” ya que la moda es 3 y 1 respectivamente pero tienen valores máximos de 36 y 32, entre otras observaciones.

Luego se indagaron los valores posibles y la frecuencia de los mismos para cada variable cualitativa (se utilizaron visualizaciones de barras para mostrar estos datos). Es en esta parte que se descubrió que las variables “place_l5” y “place_l6” tienen 0 valores posibles es decir que son columnas nulas, que hacer con estas se determinará más adelante.

También en esta etapa se hizo un análisis de cuáles columnas son irrelevantes para el propósito de este trabajo y cuáles no. En base a este análisis se tomó la decisión de dejar de contar con las siguientes columnas:

- Place_l2 (ya que siempre será CABA)
- Place_l5/l6 (100% nulas)
- Operation (todos son Venta)
- Property_currency (todos son USD)
- Property_title (no tiene importancia)

Se exploró la distribución de las variables más relevantes del dataset a través de visualizaciones univariadas. Estas son, un histograma de distribución del precio, un gráfico de barras que muestra cuales son los puntos del año con más ventas y dos boxplots para las variables “room” y “bedroom” (donde se pueden ver los outliers previamente identificados).

Finalmente se le dio cierre a la sección introductoria de exploración inicial haciendo un análisis de la correlación entre diferentes variables. Para empezar este análisis se usó un pairplot entre algunas variables seleccionadas, este muestra algunas relaciones entre variables como también muestra la completa independencia entre algunas otras variables. Se analizan más profundamente las relaciones entre superficie total y superficie cubierta, precio y cantidad de ambientes/habitaciones, y entre “rooms” y “bedrooms”, ya que son las que a simple vista parecen tener relación; Para este análisis se calculó la covarianza y correlación, y se obtuvo que la correlación más alta está entre “rooms” y “bedrooms”, y las más bajas entre precio y superficie cubierta/total.

Preprocesamiento de Datos

Esta parte está comprendida por dos pilares, el análisis de los datos nulos/faltantes y el análisis de los datos atípicos/outliers. A lo largo de este análisis se irá modificando el dataset con el que se está trabajando de manera tal que sea más útil a la hora de entrenar un modelo predictivo del precio de las propiedades. Estas modificaciones serán, imputación de datos, modificación de registros por datos mal ingresados, y de ser necesario, eliminación de algunos registros.

`Datos faltantes`

En esta sección, se llevó a cabo un análisis exhaustivo de los datos faltantes en el conjunto de datos. Inicialmente, se identificaron las variables (columnas) que contenían valores nulos.

Se observó que las variables place_l6 y place_l5 no contienen datos en absoluto, lo que las hace irrelevantes para el conjunto de datos. Asimismo, la variable place_l4 presenta un 96% de datos nulos, lo que la sitúa en la misma categoría que las variables anteriores. Por lo tanto, se tomó la decisión de eliminarlas tanto del conjunto de entrenamiento como del de prueba.

Posteriormente, se analizaron las demás variables que presentaban valores nulos, como property_bedrooms, property_surface_total, latitud, longitud, property_surface_covered, property_rooms y place_l3. Se hicieron visualizaciones de la cantidad de valores nulos y no nulos en cada una de ellas. Además, se investigó si el precio de las propiedades también contenía datos nulos.

Se abordaron los datos faltantes o mal ingresados en el dataset comenzando con la imputación Cold Deck para la variable place_l3, que contiene información sobre los barrios de las propiedades. Para completar esta información fue de utilidad el archivo barrios.csv proporcionado por el Gobierno de la Ciudad de Buenos Aires.

Para las demás variables, se realizó un análisis cuantitativo de los datos. Dado que la presencia de valores negativos o cero en estas variables carece de sentido, fueron asignados como valores NaN.

Además, se verificó la presencia de datos duplicados en el dataset y se confirmó que no había ninguno.

Luego, se aplicó la imputación MICE (Multiple Imputation by Chained Equations) a las variables Property_surface_covered, Property_rooms, Property_bedrooms y Property_surface_total. Esta técnica permitió completar los datos faltantes en el conjunto de entrenamiento.

Finalmente, se compararon las distribuciones del dataset antes y después de la imputación. Se identificaron los valores que presentaban la máxima cantidad de datos para cada variable que fue completada y se analizaron los gráficos de distribución resultantes.

Valores atipicos

Para los valores atípicos, se realizó un análisis exhaustivo de los outliers siguiendo ciertos lineamientos: los triviales que podían verse a simple vista (datos incoherentes) y aquellos que se arrojaron luego de análisis LOG/isolation tree.
Para el análisis UNIVARIADO primariamente se hizo el análisis siguiendo los rangos intercuartiles y utilizando gráficos de tipo boxplot. En muchos casos, en base a los resultados obtenidos, se comparó con alguna de las otras variables para ver si existía algún error de tipo escala, decimal o tipeo. Esto se hacía muy obvio en ciertos registros. Por dar un ejemplo, se vieron casos donde estaban mal expresados los metros cuadrados o el precio tenía ceros extra. En el caso específico de los precios que tenían estas características, se buscó referencias externas de propiedades similares en barrios similares para tener en cuenta antes de realizar cualquier modificación. Una vez realizado estos cambios (a mano), y eliminado ciertos registros que no tenían sentido en ninguna de sus variables y era imposible darles un valor de veracidad, se procedió a analizar los límites inferior y superior del IQR. Se tuvo en cuenta para este análisis no sólo el límite general, sino también el límite según barrio. Si ponemos estos datos en contexto, todos sabemos que no sería el mismo el resultado del análisis si, por ejemplo, solo nos limitamos a analizar la zona sur de la Ciudad de Buenos Aires. Desde las medianas en tamaños de superficie cubierta, a precios como tiempo de publicación, cada barrio tiene tendencias específicas, que si bien colaboran en el resultado final, influencian muchísimo en el precio. Teniendo en cuenta que el objetivo final de éste problema es poder predecir el precio de una propiedad que será lanzada en venta en ésta ciudad, no se quiso eliminar outliers que realmente no lo eran si se tenía en cuenta el barrio en el que se encontraba. Por lo tanto, puede verse que para cada variable se estableció un límite levemente superior a los bigotes/ límite IQR general, y se promedió según lo que se observó por el barrio. A estos outliers extremos (que superan incluso estos límites), se los

normalizó utilizando una corrección de tipo logarítmico. Este proceso se realizó para las siguientes variables:

- property_bedrooms
- property_rooms
- property_price
- property_surface_covered
- property_surface_total

Por último, se realizó un análisis de Z Score normal y modificado. Utilizando como umbral 3.5 se observó la permanencia de muchos valores atípicos, pero observando los registros no parecían ser valores particularmente anormales, por lo que se optó por mantenerlos como estaban (al menos hasta realizar un análisis más extenso multivariado).

Es cierto que en lo que llamamos “análisis univariado” muchas veces tomamos en cuenta otras variables antes de realizar una modificación, pero siempre fue con las herramientas de análisis univariado. Dejamos el análisis LOF / isolation forest para lo que no es tan visible a simple vista.

LOF muestra algunos problemas para identificar outliers en análisis multivariados si no se filtran los datos según tipo o barrio. Igualmente, algunas anomalías detectadas con estos valores se solucionaron normalizando con MinMax.

Se establecio n_neighbors = 30 para el caso general y n=10 para los casos por tipo de propiedad. Para ciertas variables, había un tipo de propiedad que resaltaba en cantidad de outliers con respecto a la media general (por ejemplo, las casas tenían precios demasiado altos, los PHs con respecto a la superficie total, etc). Igualmente, se concluyó que eran casos razonables por la naturaleza del problema.

Visualizaciones

La visualización de datos es el proceso de representar información de manera gráfica y visual para facilitar su comprensión y análisis. Se trata de transformar datos numéricos, textuales o cualitativos en elementos visuales, como gráficos, diagramas o mapas, que puedan ser interpretados fácilmente por los usuarios.

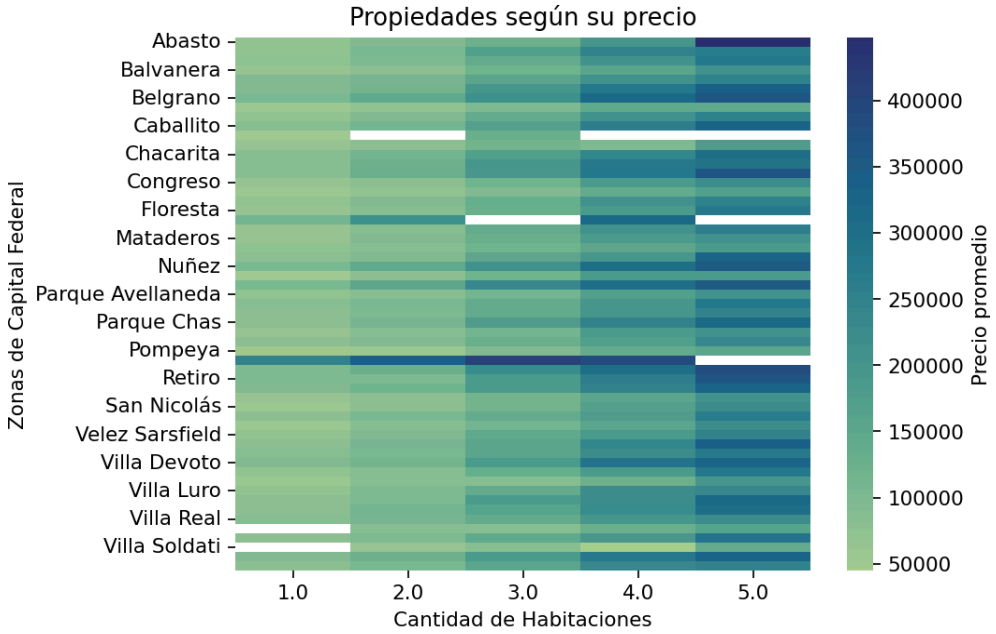
Se hicieron visualizaciones por varias razones:

Comprensión de los datos: Las visualizaciones nos permiten comprender rápidamente grandes volúmenes de datos y detectar patrones, tendencias o relaciones que podrían pasar desapercibidos en una tabla de números.

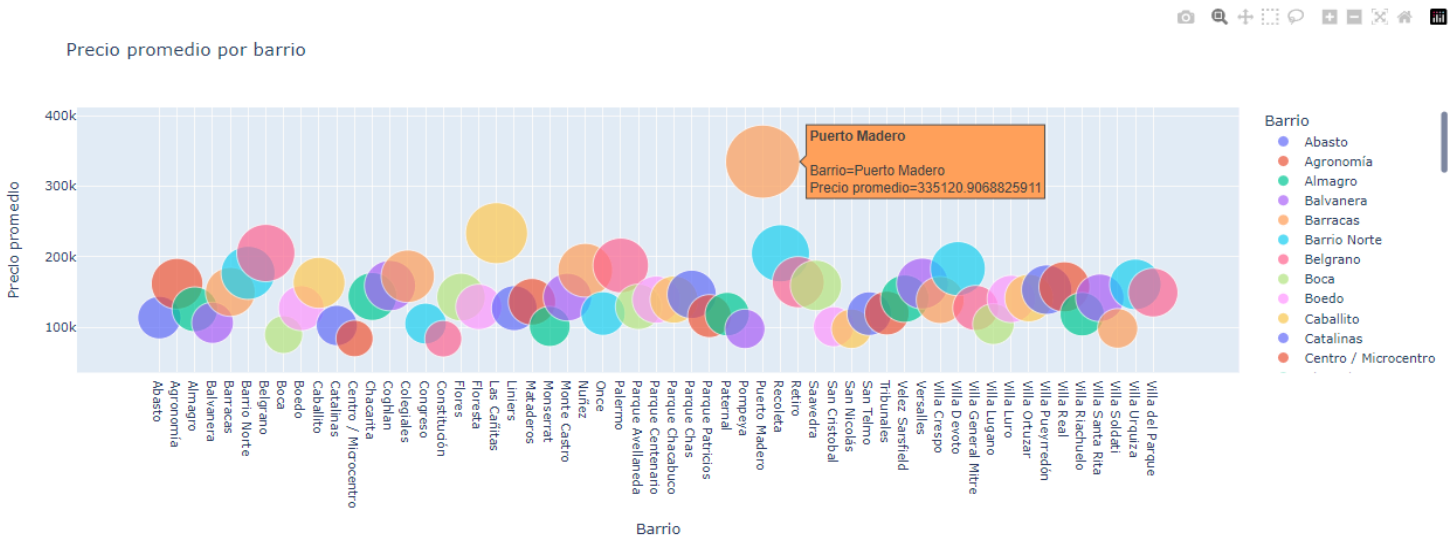
Comunicación efectiva: Las visualizaciones son una herramienta poderosa para comunicar información de manera efectiva a una audiencia. Son más intuitivas y fáciles de entender que los datos en bruto, lo que las hace ideales para presentaciones, informes o publicaciones.

Toma de decisiones: Las visualizaciones ayudan a respaldar la toma de decisiones informadas al proporcionar una representación clara y concisa de la información relevante. Permiten identificar oportunidades, riesgos o áreas de mejora.

Exploración de datos: Las visualizaciones nos permiten explorar datos desde diferentes perspectivas y niveles de detalle. Podemos interactuar con gráficos interactivos para profundizar en los detalles o ampliar nuestra comprensión.



La visualización presentada es un mapa de calor (heatmap) que muestra el precio promedio de las propiedades en función de la cantidad de habitaciones y su ubicación en diferentes zonas de la Ciudad Autónoma de Buenos Aires (CABA). Al observar el heatmap, es posible identificar patrones en los precios según la ubicación y el tamaño de la propiedad. Por ejemplo, puede haber áreas donde las propiedades con más habitaciones tienden a tener precios más altos, mientras que en otras zonas los precios son más uniformes independientemente del tamaño de la propiedad.La visualización proporciona una manera efectiva de entender cómo se relacionan el precio promedio de las propiedades, la cantidad de habitaciones y la ubicación en diferentes áreas de la Ciudad Autónoma de Buenos Aires. Ayuda a los interesados a tomar decisiones informadas sobre la compra o inversión en propiedades en la ciudad. Los valores nulos (“los espacios en blanco en la visualización”) aún persisten en el conjunto de datos ya que esta sección es previa a la sección de datos faltantes.



Esta visualización muestra un gráfico de burbujas agrupadas que representa el precio promedio de las propiedades en cada barrio.El gráfico de burbujas agrupadas proporciona una representación visual del precio promedio de las propiedades en cada barrio de interés. Cada burbuja representa un barrio, y el tamaño de la burbuja indica el precio promedio de las propiedades en ese barrio. Los colores de las burbujas pueden indicar diferentes barrios para facilitar la identificación visual, las burbujas más grandes representan barrios con precios promedio más altos, mientras que las burbujas más pequeñas representan barrios con precios promedio más bajos. El gráfico de burbujas agrupadas es interactivo, lo que permite a los usuarios explorar los datos con mayor detalle al pasar el cursor sobre las burbujas para ver información adicional, como el nombre del barrio y el precio promedio.

Clustering

Análisis de la tendencia al clustering del dataset

Primeramente se eliminaron variables que no sirven para el clustering como id, variables de fechas y la variable place_l3 ya que esta última se puede obtener conociendo latitud y longitud y luego se observó la agrupación natural de los datos mediante gráficos.

Estimación de la cantidad apropiada de grupos

Luego de lo mencionado anteriormente se procedió a hacer un minmax de las variables para que estén todas a la misma escala entre 0 y 1, y se utilizó K-means varias veces con diferentes cantidades de grupos guardando el SSD (sum of squared distances) de cada uno. Este ssd fue utilizado para analizar la cantidad óptima de grupos mediante el método del codo (Elbow method).

Evaluación de la calidad de los grupos

Dicho análisis sirvió para saber que la cantidad óptima de grupos está entre 2 y 5, pero para saber el mejor entre estas posibilidades se hizo un análisis más profundo de la calidad de los grupos con la técnica Silhouette. En este análisis se obtuvo que el silhouette score más alto se obtiene cuando hay un total de 3 grupos y es 0.82 es decir cercano a 1 (sabiendo que el silhouette score está en un rango entre -1 a 1 donde 1 significa que los grupos están bien asignados). Sin embargo si se verifica la distribución de los datos en estos grupos se ve que no están muy balanceados y que si se usan 4 o 5 grupos, estos están más balanceados pero tienen un silhouette score de ~0.2 más bajo que cuando se usan 3.

Análisis de cada grupo (incompleto)

Se examinaron las características de cada grupo para entender por qué fueron formados utilizando técnicas de visualización y estadísticas descriptivas para analizar las características distintivas de cada cluster.

Estado de Avance

1. Análisis Exploratorio y Preprocesamiento de Datos

Porcentaje de Avance: 100%/100%

2. Agrupamiento

Porcentaje de Avance: 20%/100%

Tareas en curso: puntos (d) y (e). Análisis de grupos (porque se formaron) y gráfico de grupos por colores en el mapa de CABA.
Tareas planificadas: punto (f). Análisis con 3 grupos.

Tiempo dedicado

Integrante	Tarea	Prom. Hs Semana
Aramayo Carolina	Preparación de datos para Análisis Exploratorio y procesamiento de datos Análisis de datos faltantes Reporte datos faltantes Agrupamiento	3hs 12hs 1hs 1.30hs
Utrera Maximo Damian	Exploracion Inicial Agrupamiento Análisis de grupos Reporte grupal	~11hs
Villalba Ana Daniela	Valores atípicos Agrupamiento Reporte valores atípicos	~11hs
Fiorilo Roy	Visualización de los datos Reporte Visualización	12hs