Checkpoint 2 - Grupo 02

Indice

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 y de llevar a cabo una exploración inicial de la cual se hablará en esta sección, se crearon los datasets de training y testing, para esto se dividió el dataset filtrado en 80% de datos para entrenamiento y el 20% restante para testeo.

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 dataset filtrado).

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”, seguida de la correlación entre surface_total y surface_covered, 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

ANÁLISIS UNIVARIADO

Se buscó los valores atípicos en estas variables:

- property_bedrooms
- property_rooms
- property_price
- property_surface_covered
- property_surface_total

Se buscó valores atípicos según el índice de rango intercuartil y se graficó con boxplots con el campo en general.

De estos, se identificaron outliers extremos mediante la observación de puntos muy aislados en los boxplots de las variables analizadas. Estos puntos, al estar significativamente separados del resto de los datos en el gráfico, indicaron valores que pudieron influir de manera considerable en los resultados.

Al realizar la inspección de los datos con estas características, se identificaron registros que presentaban información incoherente o improbable en el contexto de la venta de inmuebles y/o el diseño de una propiedad habitable, sugiriendo posibles errores de ingreso de datos o inconsistencias en la recolección de información. Algunos ejemplos de estos errores y su correspondiente corrección y análisis:

Propiedades con > 40 habitaciones	Se eliminaron los registros.
Propiedades con > 15 habitaciones	En los casos donde el número de ambientes era < 10, se observó un error decimal, corregido.
Propiedades con precio > 8.000.000	Se buscó referencias en sitios de venta de propiedades (Zonaprop, Remax, Argenprop, Properati). No se observaron propiedades de esas dimensiones y/o en esos barrios con ese precio en dólares estadounidenses. Realizar una conversión (en el caso de haber un error y ser un precio expresado en ARS, lo cual sería raro en el mercado) o realizar una reducción de un décimo, no parecerían adecuar los precios. No se realizaron modificaciones.
Superficie cubierta < 10m2	Si bien según el código urbanístico de la Ciudad de Buenos Aires no se permiten departamentos tan pequeños, esto puede infringirse. Bajo estos supuestos, y viendo los registros que aparecen, se reemplazó la superficie cubierta por la superficie total.
Superficie cubierta > 15000	Se realizó corrección de unidad (división por 1000)
Superficie cubierta > 4000	Se realizó corrección de unidad (división por 100)
Superficie cubierta > 1100	En los casos donde el número de habitaciones estaba dentro de los valores normales (<= 5), se realizó una corrección logarítmica.

A su vez, considerando la naturaleza del problema, se realizaron boxplots según barrios. Al observar estos boxplots se notó la diferencia que existía al realizar esta simple consideración. Si se ponen estos datos en contexto, todos saben que no serían los mismos los resultados del análisis si, por ejemplo, solo se limitaran a analizar las zonas sur de la Ciudad de Buenos Aires. Desde las medianas en tamaños de superficie cubierta, hasta los precios y el tiempo de publicación, cada barrio tiene tendencias específicas que, si bien colaboran en el resultado final, influyen mucho en el precio. Teniendo en cuenta que el objetivo final de este problema es poder predecir el precio de propiedades que serán lanzadas en venta en esta ciudad, no se quisieron eliminar outliers que realmente no lo eran si se tenía en cuenta el barrio en el que se encontraban.

Teniendo en cuenta esto, se determinaron para cada variable límites que fueran outliers extremos para todos los barrios. Se realizó una corrección logarítmica sobre

las variables (únicamente) que superaron esos límites, los cuales se detallan a continuación:

Variable	Limite
property_bedrooms	6
property_rooms	10
property_surface_total	3000

Se procedió a analizar el Z Score, normal y modificado, utilizando como límite de z-score 3.5 y 5. Los límites no fueron lo suficientemente claros para determinar qué valores eran realmente atípicos y necesitaban ser modificados. Ante la falta de claridad en los resultados, se optó por no realizar cambios en los datos debido a la preocupación de no alterar los datos de manera innecesaria sin una comprensión clara de cuáles valores son verdaderamente atípicos. Se postergó cualquier modificación al análisis multivariado más exhaustivo que permitiera comprender mejor cómo podrían influir en el conjunto estos datos.

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

ANÁLISIS MULTIVARIADO

En primer lugar, se realizó un análisis de isolation forest. Revisando las anomalías y viendo los scatterplot que resultaron de éste modelo, en su mayoría parecerían seguir las tendencias de los registros no considerados anómalos pero con valores levemente inferiores/superiores a la mediana.. Parecen ser anomalías genuinas y decidimos mantenerlas por contexto. Sabemos que existe un riesgo inherente en mantener los datos como están pero de detectarse en el entrenamiento que los resultados no son los indicados se revisará esta sección.

LOF mostró algunos problemas para identificar outliers en análisis multivariados al no filtrarse los datos según tipo o barrio. En el caso específico de contraponer (property_surface_total, property_surface_covered) y (property_price,

property_surface_total), al inspeccionar las anomalías determinadas por LOF, se encuentran anomalías extrañas que se normalizan con MinMax para poder analizarlas nuevamente. Luego de la normalización se vuelve a analizar con LOF pero no se encuentran mejoras notables en los resultados.

Se estableció 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, concluyeron que eran casos razonables por la naturaleza del problema y no se descartaron estos registros.

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.

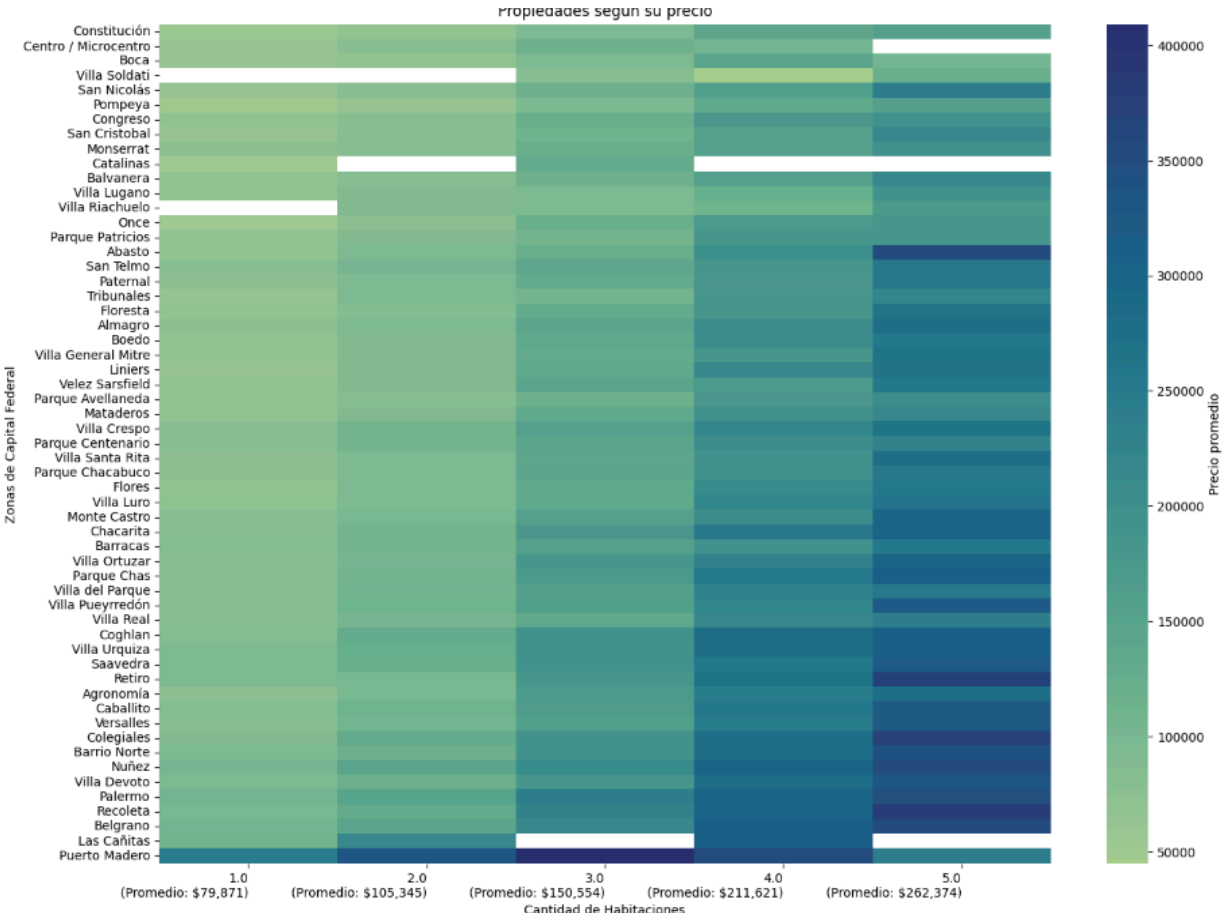
Se hicieron visualizaciones por varias razones:

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.

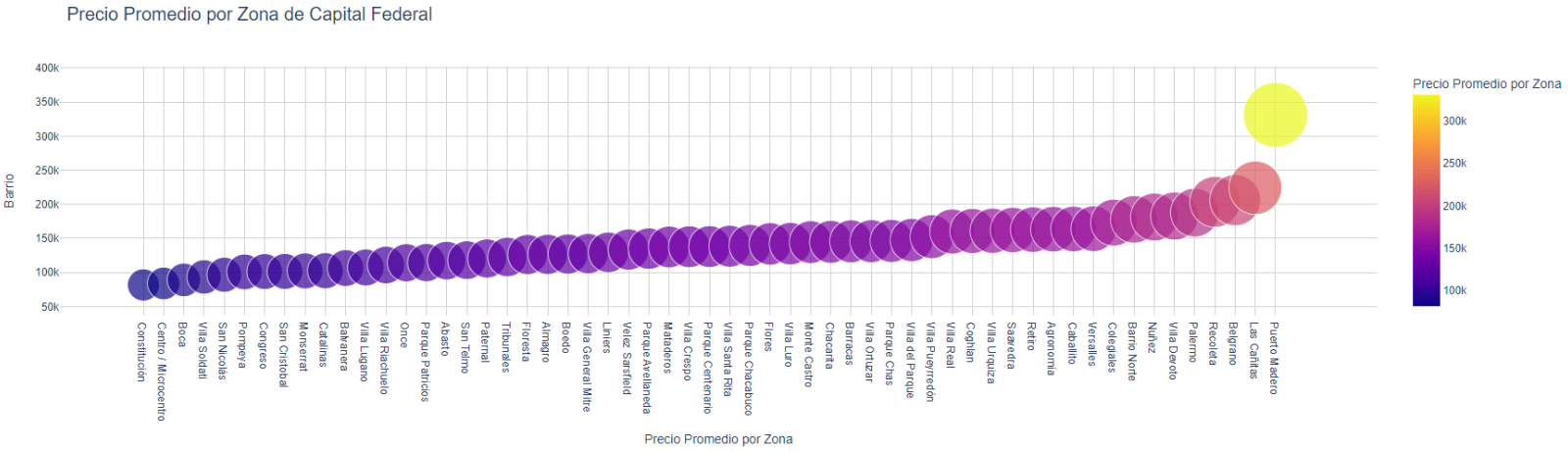
Comunicación efectiva: Las visualizaciones son una herramienta poderosa para comunicar información de manera efectiva a una audiencia. Son más intuitivas y fáciles de entender que los datos en bruto, lo que las hace ideales para presentaciones, informes o publicaciones.

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.

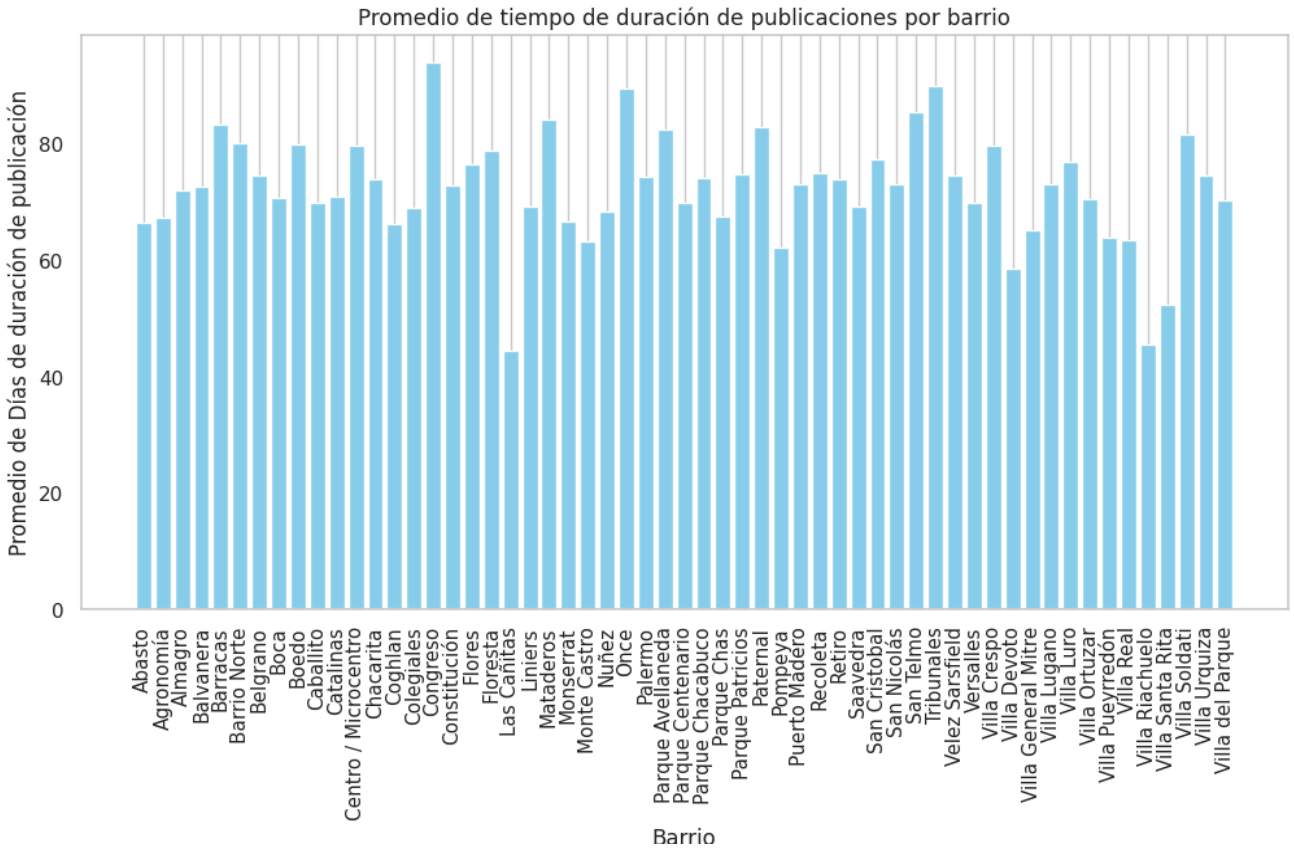
Exploración de datos: Las visualizaciones nos permiten explorar datos desde diferentes perspectivas y niveles de detalle. Podemos interactuar con gráficos interactivos para profundizar en los detalles o ampliar nuestra comprensión.



La visualización presentada fue un mapa de calor (heatmap) que mostraba 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, fue posible identificar patrones en los precios según la ubicación y el tamaño de la propiedad. Por ejemplo, había áreas donde las propiedades con más habitaciones tendían a tener precios más altos, mientras que en otras zonas los precios eran más uniformes independientemente del tamaño de la propiedad. La visualización proporcionó una manera efectiva de entender cómo se relacionaban el precio promedio de las propiedades, la cantidad de habitaciones y la ubicación en diferentes áreas de la Ciudad Autónoma de Buenos Aires. Ayudó a los interesados a tomar decisiones informadas sobre la compra o inversión en propiedades en la ciudad. Cabe destacar que los precios estaban ordenados, lo que facilitó la percepción de los degradés de colores en el heatmap. Sin embargo, es importante señalar que los valores nulos ("los espacios en blanco en la visualización") aún persistieron en el conjunto de datos, ya que esta sección fue previa a la sección de datos faltantes.



Esta visualización mostró un gráfico de burbujas agrupadas que representaba el precio promedio de las propiedades en cada barrio de la zona de Capital Federal. Cada burbuja representaba un barrio, y el tamaño de la burbuja indicaba el precio promedio de las propiedades en ese barrio. Los colores de las burbujas indican diferentes rangos de precios para facilitar la identificación visual: los tonos más oscuros representaban barrios con precios promedio más bajos, mientras que los tonos más claros representaban barrios con precios promedio más altos. El gráfico de burbujas agrupadas era interactivo, lo que permitió 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. Esta representación visual proporcionó una manera intuitiva de comprender la distribución de precios en los diferentes barrios de la zona de Capital Federal, después de haberse realizado todas las correcciones sugeridas por el corrector, llegando a la conclusión de que era una variación de un heatmap.



Esta visualización presentó un gráfico de barras que mostraba el promedio de tiempo de duración de las publicaciones por barrio en la Ciudad Autónoma de Buenos Aires (CABA). Cada barra en el gráfico representaba un barrio específico, y su altura indicaba el promedio de días que las publicaciones permanecieron activas en ese barrio. La visualización permitió comparar fácilmente la duración promedio de las publicaciones entre diferentes barrios de la ciudad. Las barras más altas indicaron que las publicaciones en esos barrios tendieron a permanecer activas por más tiempo, mientras que las barras más cortas indicaron una duración promedio más corta de las publicaciones. Los barrios con barras más altas podrían indicar áreas de la ciudad donde la demanda de propiedades era más alta o donde las propiedades tendían a permanecer en el mercado por períodos más prolongados. Esto pudo haber sido útil para comprender las dinámicas del mercado inmobiliario en diferentes áreas de la ciudad.

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 luego se observó la agrupación natural de los datos mediante gráficos y se procesaron los datos haciendo una regresión logística de la columna property_type, encodeando los barrios y normalizando las variables numéricas por medio de Min-Max.

Estimación de la cantidad apropiada de grupos

Primeramente 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). Dicho método dio como resultado que el mejor K estaría entre 2 y 5, y para elegir el mejor de estos es que se calcula el silhouette score de cada uno y se lo grafica. El K con el mayor score fue K=2 con un total de 0.66, sin embargo los datos no estaban bien balanceados y había un cluster con muchas más observaciones que el otro. Es por esto que al evaluar una tabla con la información obtenida:

- k = 2, score: 0.66, division: 57k/14k
- k = 3, score: 0.61, division: 34k/26k/11k
- k = 4, score: 0.60, division: 34k/23k/8k/5k
- k = 5, score: 0.59, division: 24k/22k/10k/8k/5k

Se decidió como mejor opción el K=4, y es con el que se trabajó.

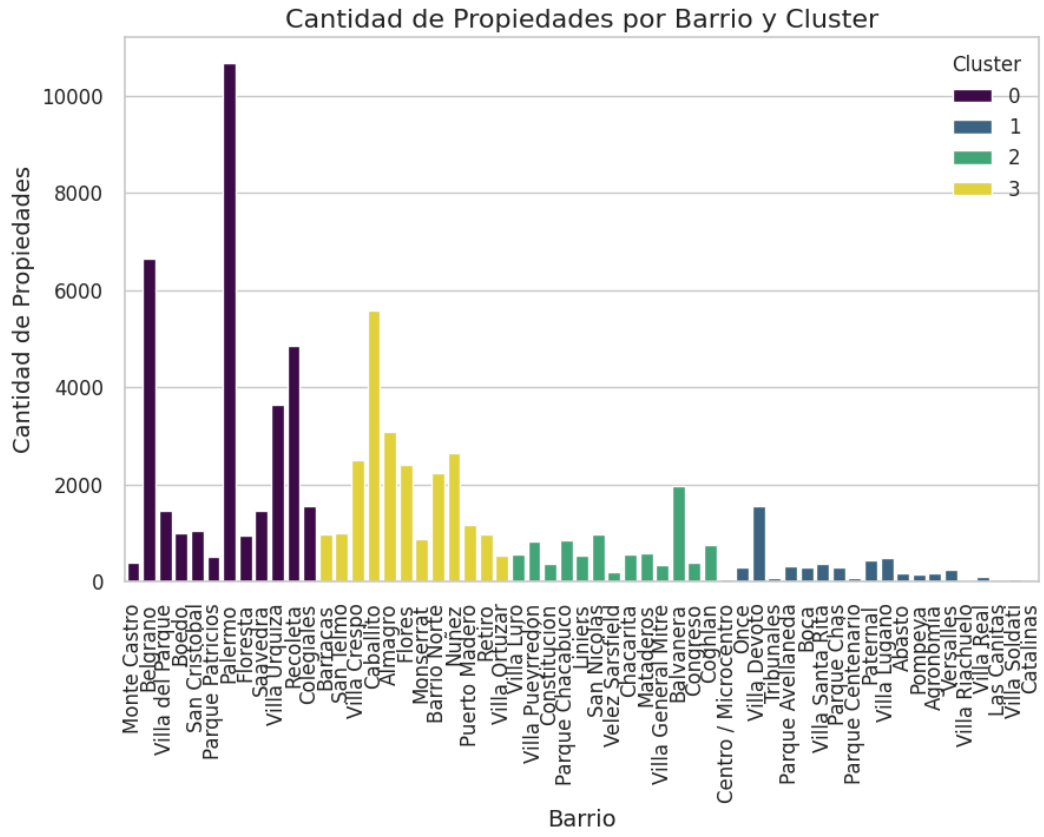
Análisis de grupos

Se examinaron a detalle todas las variables utilizando distintas visualizaciones y mostrando cómo cada una impactó a la creación de los clusters. Más específicamente se observó si la agrupación sucedió:

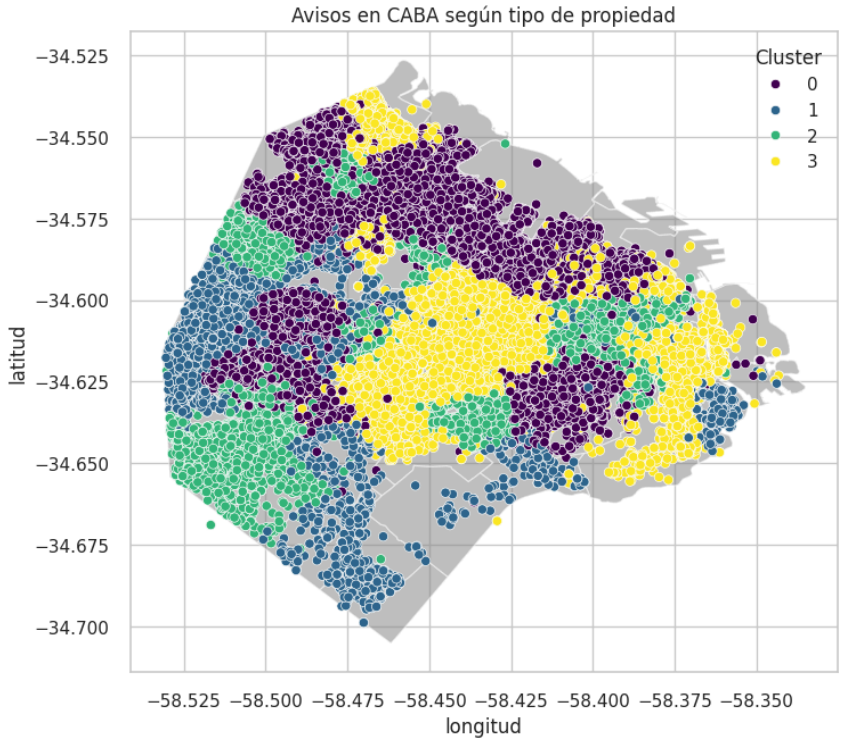
- Por cercania geografica
- Por barrio
- Por tipo de propiedad

- Por precio
- Por cantidad de habitaciones/ambientes
- Por superficie

Se concluyó que hay una clara división geográfica entre clusters, y esta se dio debido a los barrios. Cada barrio pertenece a un y solo un cluster, es decir no hay intersecciones, lo que quiere decir que es una agrupación perfecta en función de dicha variable (ver el gráfico de abajo). Adicionalmente se intentó ver si alguna otra variable también tuvo algún grado de influencia pero dicha hipótesis se descartó luego de analizar las visualizaciones de las otras variables.



Gráficos de los clusters en el mapa de CABA



Análisis con tres grupos

En esta sección se hizo otro análisis de grupos pero esta vez utilizando el algoritmo K-means con un K=3. A su vez esta vez se eliminó la columna de barrios ya que de lo contrario este análisis terminaba dando un resultado muy parecido al anterior.

Primeramente luego de clusterizar se hizo el mismo gráfico mostrado en el anterior análisis, pero esta vez se observó que ya no hay agrupación geográfica alguna. Por lo tanto se procedió a analizar variable por variable el impacto de cada una, y en base a esto se encontró que nuevamente ocurrió una agrupación perfecta pero esta vez en base al tipo de propiedades, es decir, se crearon tres clusters cada uno conteniendo el 100% de cada tipo de propiedad y sin intersecciones.

Entonces se concluye que esta clusterización no fue muy útil ya que las propiedades ya estaban ‘agrupadas’ por su tipo previamente.

Clasificación

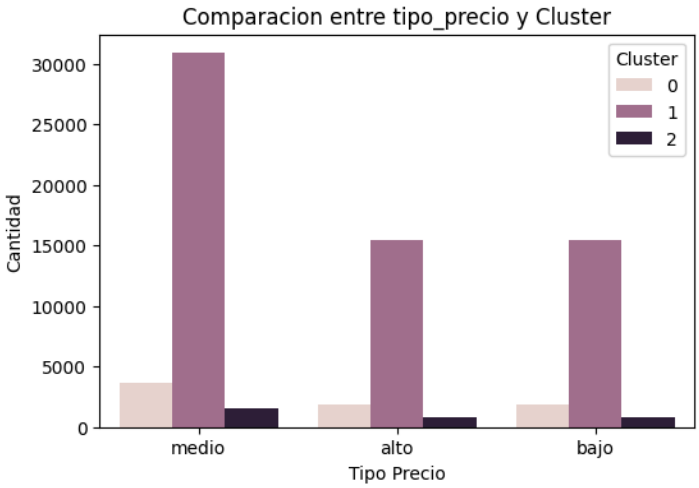
Mencionar cuál es la alternativa seleccionada para construir la variable “tipo_precio” justificando su elección. Mostrar en un mapa de CABA los avisos coloreados por tipo_precio ¿Qué diferencias o similitudes encuentran con el agrupamiento realizado por K-Means con 3 grupos?
Si llegaron a entrenar alguno de los modelos, mencionar cuáles y qué métricas obtuvieron en test y si realizaron nuevas transformaciones sobre los datos (encoding, normalización, etc) completando los ítems a y b:

Esta sección se abre haciendo una construcción un target dataset, para el cual se creó una nueva feature llamada “tipo_precio” que se distribuye en “bajo”, “medio” o “alto”. Para la asignación de una categoría de precio a cada propiedad se evaluaron tres alternativas que tienen en cuenta el precio por metro cuadrado (pxm2) de las propiedades. Estas fueron:

- División del pxm2 equitativa entre las tres categorías.
- División de pxm2 de 25% - 50% - 25% entre las tres categorías.
- División de pxm2 de 25% - 50% - 25% de cada tipo de propiedad, entre las tres categorías.

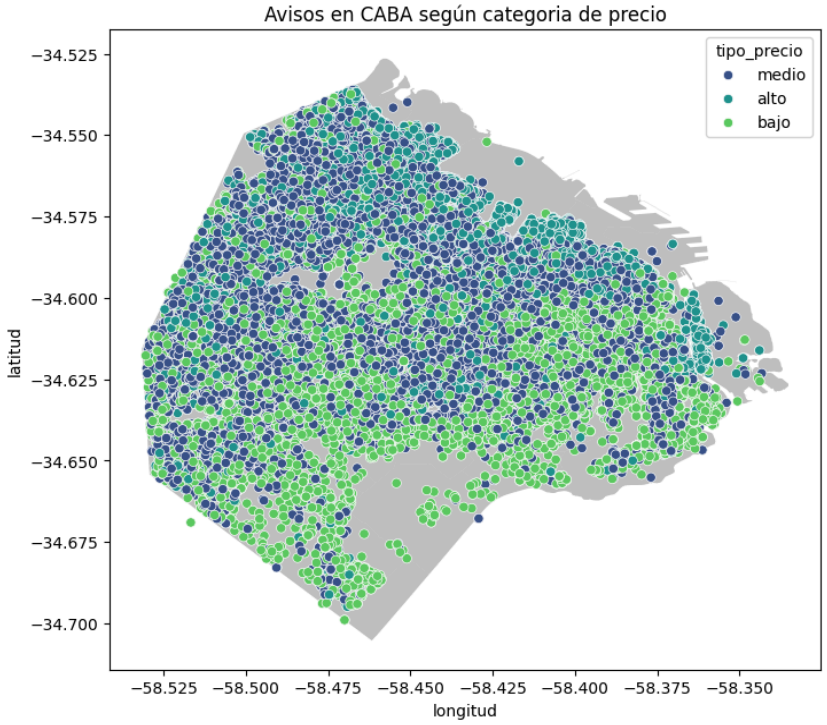
Se decidió que el mejor método fue el tercero ya que tiene una distribución más justa (no simplemente arbitrariamente equitativa) y a su vez tiene en cuenta que por ejemplo puede ocurrir que un precio X sea barato para un PH pero caro para un Departamento.

Luego se compararon las variables de tipo precio con los clusters mencionados anteriormente cuando el K=3, y se ve cómo se distribuyen los tipos de propiedad en cada categoría de precio.



- Cluster 0: PHs
- Cluster 1: Departamentos
- Cluster 2: Casas

Finalmente antes de comenzar con la sección de entrenamiento y predicción, se observó mediante un gráfico en el mapa de CABA que hay mayor densidad de precios baratos en el sur de la ciudad, y este aumenta a medida que se va más al norte, siendo que en el borde norte hay más densidad de precios altos.



Ingeniería de características

En esta sección se eliminaron las columnas que den información sobre el precio de las propiedades ya que se quiere predecir la categoría de precio a partir del resto de columnas. También se eliminaron todas las columnas consideradas irrelevantes como ID, fechas y una de las columnas con una alta correlación entre sí (por ejemplo, rooms y bedrooms). Luego se procedió a encodear las variables cualitativas, y a normalizar las variables numéricas utilizando Min-Max.

Concluye este proceso con un dataset de las 5 columnas place_l3, property_type, property_rooms, property_surface_total, tipo_precio adecuadamente procesadas.

a. Construcción del modelo (Pendiente)

b. Cuadro de Resultados (Pendiente)

Regresión

Ingeniería de características

En esta sección se aplica la ingeniería de características anteriormente utilizada para la parte de Clasificación con la diferencia de que esta vez el dataset a utilizar para entrenar los modelos está conformado por las columnas place_l3, property_type, property_rooms, property_surface_total, property_price.

a. **Construcción del modelo (Pendiente)**
b. **Cuadro de Resultados (Pendiente)**

Estado de Avance

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

   Porcentaje de Avance: 100%/100%

2. **Agrupamiento**

   Porcentaje de Avance: 100%/100%

3. **Clasificación**

   Porcentaje de Avance: 40%/100%

   Tareas en curso: Árbol de decisión

   Tareas planificadas: Random forest, Modelo a elección.

4. **Regresión**

   Porcentaje de Avance: 20%/100%

   Tareas planificadas: Knn, Xgboost, Modelo a elección.

Tiempo dedicado

Integrante	Tarea	Prom. Hs Semana
Aramayo Carolina	Agrupamiento	1
Utrera Maximo Damian	Finalización de tareas de agrupamiento Finalización de análisis de clusters Construcción del target (clasificación) Ingeniería de caracteristicas (clasificacion y regresion) Reporte grupal	~12
Villalba Ana Daniela	Agrupamiento Arbol de decisión	1
Fiorilo Roy	Finalización de análisis de tres clusters Construcción del target (clasificación) Reporte grupal	~12