Informe técnico análisis de datos precio de apartamento en venta en la ciudad de Bogotá

Elaborado por Caris Andrea Chia Amaya

Enlace de repositorio de Github: https://github.com/caris-chia/test-xpecta

Enlace de aplicativo Streamlit: https://test-xpecta-pcappzfdnxjqxk8ztvb7zo.streamlit.app/

Consideraciones iniciales

Durante la fase de exploración para elegir el tema de este proyecto, consideré opciones relacionadas con psicología, como un modelo de predicción de los resultados del ICFES en función de características sociodemográficas, análisis de violencia de género usando la base de datos de Medicina Legal, o el análisis de competencias de candidatos en procesos de selección. Sin embargo, esta prueba técnica requería la extracción de datos mediante Webscraping. En mi campo, es más común obtener bases de datos a través de entidades públicas o datos abiertos, lo cual limitaba las opciones a análisis bibliométricos, área en la que tengo considerable experiencia, ya que los repositorios suelen ofrecer medios confiables para descargar la información en diversos formatos. También evalué la posibilidad de realizar un análisis de sentimientos de comentarios en redes sociales, pero la calidad y disponibilidad de esos datos no me convencía para demostrar mis habilidades en esta prueba técnica.

Por ello, decidí optar por un tema relacionado con el sector inmobiliario, ya que tengo experiencia analizando este tipo de datos durante mis estudios de maestría. Además, me encuentro aprendiendo sobre análisis espaciales, un área que me interesa mucho por su aplicabilidad transversal a distintos campos. Elegir una base de datos inmobiliaria me permitía aplicar estos conocimientos y explorar nuevas perspectivas que enriquecen mi formación.

Descripción general del proceso

El análisis del costo de los apartamentos a la venta en Bogotá es de gran relevancia, ya que contribuye a comprender las dinámicas del mercado inmobiliario en una de las ciudades más importantes de Colombia. Esta información no solo es fundamental para potenciales compradores, sino también para inversores y planificadores urbanos que buscan identificar tendencias y evaluar la accesibilidad de la vivienda en la capital. En concordancia, según Garay y Rodríguez (2021), el comportamiento del mercado inmobiliario en Bogotá refleja aspectos económicos y sociales que impactan tanto en la calidad de vida de los residentes como en las estrategias de desarrollo urbano sostenible, por lo tanto, estudiar estas variables permite un entendimiento más profundo de cómo se configuran los precios y su relación con factores espaciales y socioeconómicos.

Para llevar a cabo el presente análisis del mercado inmobiliario en Bogotá, se realizó un proceso de web scraping en la página de Finca Raíz con el objetivo de extraer datos sobre

apartamentos en venta en la ciudad. Aunque inicialmente el sitio web reportaba 42.055 resultados, se identificó que a partir de la página 476 ya no se mostraban más resultados, lo cual resultó en un total aproximado de 9.996 registros. En el proceso, se utilizaron las librerías Selenium, BeautifulSoup y otras herramientas encontradas en el código fuente (leer el Readme.md de la carpeta para entender el funcionamiento de cada script), con lo cual se logró extraer los enlaces individuales de cada propiedad y posteriormente los datos detallados, como el precio, la ubicación, el estrato, la cantidad de habitaciones, espacios de parqueadero, antigüedad del apartamento, entre otros. Tras eliminar 82 registros correspondientes a proyectos de construcción, debido a su estructura inconsistente, se obtuvieron 8.480 registros finales que serán analizados. Esta base de datos permite no solo estudiar el comportamiento de los precios en el mercado inmobiliario de Bogotá, sino también explorar otros factores que influyen en la oferta de viviendas.

Respecto a la legalidad del proceso de web scraping realizado sobre el portal de Finca Raíz, es importante señalar que los términos y condiciones del sitio web (https://www.fincaraiz.com.co/informacion#terminos-y-condiciones) establecen claramente la prohibición de realizar scraping o crawling de las páginas sin la autorización explícita del portal. En particular, se indica que no está permitido extraer información para la construcción de bases de datos, fines comerciales o difusión masiva de información.

En el caso del presente análisis, es importante mencionar que su propósito es estrictamente académico, en el contexto de una prueba técnica, y no será divulgado ni utilizado con fines comerciales. El objetivo del scraping fue únicamente demostrar competencias técnicas en la extracción y análisis de datos, sin intención de afectar los derechos de propiedad intelectual o los intereses comerciales de Finca Raíz. Por lo que para usar este análisis con fines comerciales hay que solicitar autorización de Finca Raíz.

Limpieza de base de datos y Análisis Exploratorio (EDA)

Con el objetivo de realizar la limpieza e imputación de valores nulos en el conjunto de datos para preparar el modelo de predicción de precios de viviendas, se comenzó identificando las columnas con valores faltantes en el dataset. Las principales columnas con valores nulos fueron:

- social_stratum: 14 valores nulos.
- bathrooms: 54 valores nulos.
- private_area: 899 valores nulos.
- age: 617 valores nulos.
- rooms: 21 valores nulos.
- parking_spaces: 1,584 valores nulos.
- administration: 1,043 valores nulos.

- floor_number: 2,258 valores nulos (estos se marcaron inicialmente como "No reporta").

Para hacer el proceso de limpieza y transformación de datos se imputó usando varios enfoques:

- **social_stratum (estrato social):** Se imputaron los valores nulos utilizando la moda por barrio (main_location), ya que el estrato es una variable ordinal. Para los valores restantes, se imputó utilizando la moda general.
- **bathrooms y rooms (número de baños y habitaciones):** Se usó la mediana por cuartiles de área construida (constructed_area) para imputar estos valores, dado que con un análisis de correlaciones se identificó que hay una tendencia de más baños y habitaciones en apartamentos más grandes.
- **private_area (área privada):** Se imputó utilizando la columna constructed_area, ya que en muchos casos el área construida y el área privada son iguales o los usuarios la reportaban así.
- **age (antigüedad de la propiedad):** Se reemplazaron los valores nulos por “No reporta” dado que no se identifican relaciones significativas con otras variables.
- **parking_spaces (espacios de parqueo):** Se utilizó la mediana de espacios de parqueo por cuartiles de área construida.
- **administration (costo de administración):** Se imputaron los valores faltantes utilizando la mediana del estrato social (social_stratum), debido a que el costo de administración tiende a variar según el estrato.
- **floor_number (número de piso):** Se optó por imputar los valores nulos con la categoría "No reporta", ya que no se observó una relación significativa con otras variables como el precio.

Adicionalmente, algunas columnas, como bathrooms, rooms, y parking_spaces, contenían valores flotantes cuando deberían ser enteros, por lo que se realizó una conversión de tipo para asegurar la consistencia de los datos.

Análisis descriptivo

Utilizando las librerías shapely y pandas, se construyó un polígono para delimitar la base de datos de los apartamentos en venta en Bogotá, concentrándose en una zona específica que incluye las calles 26 a 72 y las carreras 30 hasta la Avenida Boyacá. Este polígono permitió filtrar los datos de la base general para obtener un conjunto más pequeño y específico sobre el cual realizar análisis descriptivos de las características de los apartamentos.

En cuanto a la base de datos que contiene todos los apartamentos de Bogotá, el plan inicial era utilizarla para crear un modelo de predicción. Sin embargo, el análisis se limitó a lo

descriptivo debido a una limitante importante, el campo “main_location”, que indica el barrio, no siempre es preciso porque se selecciona manualmente por los usuarios, lo que genera inconsistencias. Para solucionar este problema, se intentó una validación a partir de la localización geográfica usando la librería geopy. En la carpeta del código se encuentra un script llamado prueba_barrio, que realiza esta tarea con una muestra de la base de datos. Este método utiliza el API de Google Maps, pero, debido al alto número de solicitudes necesarias para la base de datos completa, el costo de las peticiones superaría los 40 USD por lo que no lo ejecuté. Otra opción considerada fue el uso del servicio Nominatim de OpenStreetMap, pero alcancé a implementar la prueba. Una vez se pueda resolver este problema de validación geográfica, será posible obtener datos confiables sobre el barrio y la localidad, lo cual permitirá avanzar con la creación del modelo predictivo y el análisis más detallado del mercado inmobiliario en Bogotá.

Posibles mejoras futuras

- Respecto a la extracción de barrios se dejo implementado el código funcional para una base de datos más pequeña, que no excede el límite de tokens gratuitos por lo que se puede ajustar para usar Nominatim de OpenStreetMap a través de geocodificación inversa como alternativa viable.
- Otra posible mejora es la extracción de imágenes de los inmuebles y del texto descriptivo que redacta el usuario que publica el anuncio. Esto permitiría realizar un análisis de estas descripciones conectándose a la API de ChatGPT para evaluar el contenido y obtener información adicional.
- En el repositorio se incluye un archivo compose.yml. Dado que los datos se encontraban en una sola base de datos, no fue necesario dividirlos para trabajar con SQL. Sin embargo, de cara a futuros desarrollos, como la implementación de un modelo de machine learning, se podrían integrar otras fuentes de datos, esto haría necesaria la implementación de una infraestructura más compleja, justificada por la utilidad del archivo compose.yml para manejar múltiples bases de datos de manera eficiente.
- Se puede mejorar la interfaz de usuario explorando la documentación de Streamlit. Esto permitiría desarrollar una aplicación interactiva y accesible para visualizar los resultados del análisis, facilitando la interpretación y presentación de los datos de una manera amigable y eficiente.