La base de datos original, se obtuvo de la página para finca raíz “properaty y se presenta como una fuente valiosa para abordar el problema planteado en este taller, que consiste en predecir los precios de las viviendas en Bogotá. La principal ventaja de esta base de datos radica en su amplitud, ya que contiene precios recientes de propiedades en toda la ciudad, lo que proporciona una visión completa del mercado inmobiliario. Además, esta base de datos incluye descripciones detalladas de las propiedades, lo que permite extraer información relevante para la solución del problema.

Sin embargo, es importante señalar que la base de datos presenta desafíos significativos. Se observan muchos valores faltantes, lo que introduce incertidumbre en el análisis y modelado de datos. Además, se han identificado algunos errores de digitación o extracción durante el proceso de web scraping, lo que ha resultado en valores astronómicamente altos en algunas propiedades. Estos desafíos deben abordarse cuidadosamente en el proceso de limpieza y preparación de datos para garantizar que los modelos de predicción se basen en información precisa y confiable.

Variables:

Originalmente se tenía:

|  |  |
| --- | --- |
| **Variable** | **Descripción** |
| property\_id | ID de propiedad |
| city | Ciudad de la propiedad |
| price | Precio de la propiedad |
| month | Mes del anuncio |
| year | Año del anuncio |
| surface\_total | Superficie de la propiedad |
| surface\_covered | Superficie de la propiedad |
| rooms | Número de Habitaciones |
| bedrooms | Número de Alcobas |
| bathrooms | Número de Baños |
| property\_type | Si es casa o apartamento |
| operation\_type | Tipo de operación (Venta en este caso) |
| lat | Latitud |
| lon | Longitud |
| title | Título de la propiedad |
| description | Párrafo descriptivo de la propiedad |

Missings originales:

|  |  |  |
| --- | --- | --- |
| **Missing values Analysis** | | |
|  | | |
| Variables | Missing |  |
|  | | |
| property\_id | 0 |
| city | 0 |
| price | 0 |
| month | 0 |
| year | 0 |
| surface\_total | 30,790 |
| surface\_covered | 30,079 |
| rooms | 18,260 |
| bedrooms | 0 |
| bathrooms | 10,071 |
| property\_type | 0 |
| operation\_type | 0 |
| lat | 0 |
| lon | 0 |
| title | 22 |
| description | 9 |

Como se mencionó en las desventajas, dentro la base de datos original se han identificado valores faltantes en varias variables que desempeñan un papel fundamental en la predicción de precios de viviendas. Específicamente, las variables "surface\_total" y "surface\_covered" muestran un alto número de valores faltantes, con 30,790 y 30,079 registros incompletos respectivamente. La superficie de una propiedad es un factor crucial en la determinación de su precio, por lo que la presencia de estos valores faltantes representa un desafío significativo. Además, la variable "rooms" exhibe 18,260 valores faltantes, mientras que "bathrooms" tiene 10,071 valores faltantes. Ambas variables, que describen las características de las viviendas, son relevantes para la predicción de precios. Además, la variable "title" presenta 22 valores faltantes, y la variable "description" tiene 9 valores faltantes. Ambas variables se utilizan para describir las propiedades y pueden contener información valiosa para el análisis.

Con esto en mente, la construcción de la base de datos implicó la incorporación de variables tanto a partir del texto descriptivo como de fuentes externas de datos espaciales. En cuanto a las variables de texto, se incluyeron las siguientes:

1. **Dummy Parqueadero**: Esta variable se creó para medir si en la descripción de la propiedad se mencionaba la disponibilidad de un parqueadero. Tomaba el valor de 1 si se mencionaba un parqueadero y 0 en caso contrario.
2. **En qué piso es el apto**: A través del análisis del texto descriptivo, se extrajo el número de piso en el que se encuentra el apartamento, lo que puede ser un factor relevante en la valoración de la propiedad.
3. **Dummy Penthouse**: Esta variable se utilizó para identificar si la vivienda era un penthouse, lo que generalmente implica un estatus premium. Tomaba el valor de 1 si se mencionaba que la propiedad era un penthouse y 0 en caso contrario.

En cuanto a las variables externas, se incorporaron datos provenientes de fuentes abiertas:

1. **Ciclovías**: La distancia a la ciclovía más cercana se obtuvo a partir de los datos disponibles en el portal de datos abiertos de Bogotá, lo que puede ser relevante para aquellos interesados en la accesibilidad en bicicleta.
2. **Transmilenio**: La distancia a la estación de Transmilenio más cercana se extrajo de OpenStreetMap, lo que proporciona información crucial sobre la accesibilidad al transporte público en la ciudad.
3. **UPL (Unidad Administrativa)**: Se incluyó la unidad administrativa a la que pertenece la propiedad, utilizando datos disponibles en el portal de datos abiertos de Bogotá.
4. **Distancia al Centro Comercial más cercano**: La distancia a los centros comerciales más cercanos se obtuvo de OpenStreetMap, lo que puede ser relevante para aquellos que valoran la proximidad a centros comerciales.
5. **Estrato**: El estrato de la vivienda, información también obtenida del portal de datos abiertos de Bogotá, desempeña un papel importante en la valoración del precio por metro cuadrado de la propiedad.
6. **Distancia al Parque más cercano**: La distancia al parque más cercano se extrajo de OpenStreetMap, lo que proporciona información sobre la accesibilidad a áreas verdes.
7. **Distancia al CAI más cercano**: La distancia al CAI (Comando de Atención Inmediata), que puede ser indicativa de la seguridad en la zona, se obtuvo de OpenStreetMap.

Imputaciones:

La corrección de los valores faltantes en la base de datos se llevó a cabo de manera exhaustiva. Inicialmente, se intentó imputar los valores faltantes a partir del texto descriptivo utilizando expresiones regulares para extraer información sobre la superficie de la propiedad, el número de alcobas y baños. Sin embargo, persistieron numerosos valores faltantes.

Para abordar los valores faltantes en la variable "superficie" (mts2), se implementó un enfoque en el conjunto de entrenamiento (train). Se calculó el precio por metro cuadrado de la propiedad y se buscó el valor más cercano sin faltantes. Luego, se utilizó este valor para estimar la superficie de las propiedades con valores faltantes.

En el caso de las variables de baños y habitaciones, se realizaron divisiones de las propiedades en bandas de metros cuadrados: "<100", "100-200", "200-300", "300+" y, a partir de estas categorías, se imputaron los valores faltantes utilizando la moda en cada banda.

La variable de "estrato" también presentaba un alto número de valores faltantes. Se imputaron estos valores utilizando la propiedad más cercana a 700 metros en el conjunto de entrenamiento, mientras que en el conjunto de prueba (test) se imputaron los valores faltantes con "estrato 5", ya que es común en propiedades en la zona de Chapinero.

Finalmente, se realizó una limpieza de la base de datos de entrenamiento (train) para eliminar observaciones atípicas. Se excluyeron propiedades con precios por metro cuadrado superiores a 12 millones y propiedades con más de 500 metros cuadrados. Como resultado, el conjunto de datos de entrenamiento pasó de 38,644 observaciones a 36,629, lo que contribuye a la calidad y confiabilidad de los análisis y modelos subsiguientes.

Análisis descriptivo.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Numeric variables analysis** | | | | | |
|  | | | | | |
| Statistic | N | Mean | St. Dev. | Min | Max |
|  | | | | | |
| price | 36,621 | 648,639,006.0 | 307,081,200.0 | 300,000,000 | 1,650,000,000 |
| distancia\_bus | 36,621 | 948.7 | 681.1 | 3.6 | 5,856.8 |
| distancia\_parque | 36,621 | 160.4 | 99.0 | 1.0 | 1,102.8 |
| distancia\_cc | 36,621 | 658.0 | 382.2 | 0.6 | 4,004.8 |
| distancia\_police | 36,621 | 1,012.4 | 507.6 | 2.4 | 2,236.9 |
| surface\_total | 36,621 | 136.7 | 79.7 | 28.0 | 500.0 |

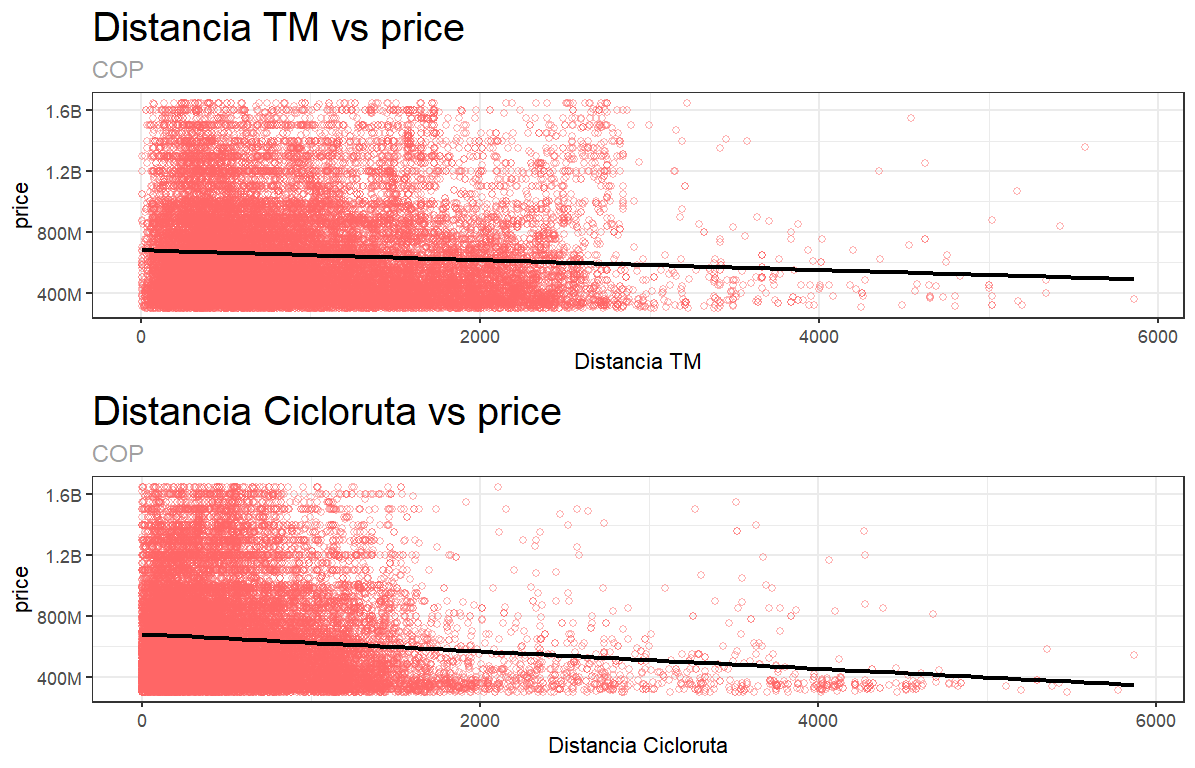
Descriptivas numéricas: Los datos facilitados ofrecen información valiosa sobre seis variables numéricas con 36.621 observaciones cada una. La variable "Precio" muestra una amplia gama de precios inmobiliarios, con una media de 648.639.006,0 y una importante desviación típica de 307.081.200,0, que oscila entre 300.000.000 y 1.650.000.000. Esta información es esencial para comprender los valores inmobiliarios de la zona.

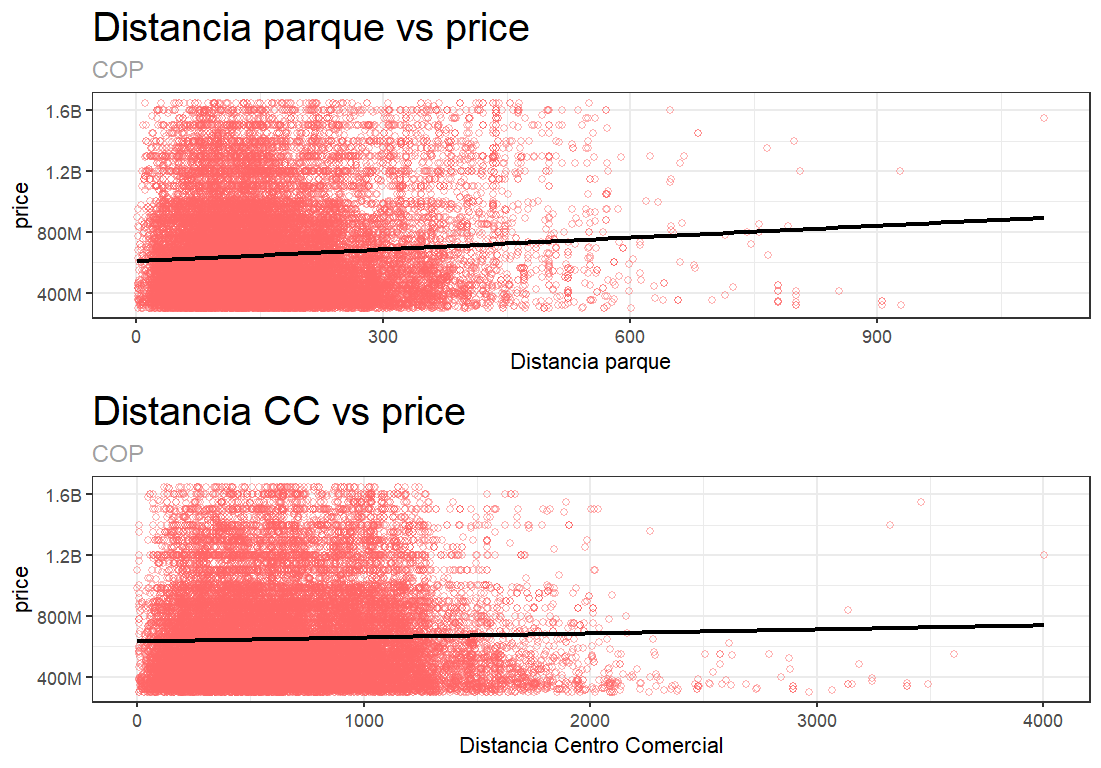
Las variables de distancia ("Parada de autobús", "Parque", "Centro comercial" y "CAI") proporcionan información sobre la accesibilidad a los servicios. Por ejemplo, la "Distancia al parque" tiene una media de 160,4 y una desviación típica moderada de 99,0, lo que indica que los parques son generalmente accesibles, con cierta variabilidad en su proximidad. Estas estadísticas ayudan a urbanistas, residentes y responsables políticos a comprender la dinámica espacial.

La variable "Superficie total", con una media de 136,7 y una desviación típica de 79,7, muestra la distribución del tamaño de las propiedades. Este dato es crucial para que compradores y promotores evalúen el tamaño de los inmuebles disponibles. En resumen, estas estadísticas ofrecen una visión global del precio, la accesibilidad y el tamaño de los inmuebles, lo que ayuda a tomar decisiones informadas en el mercado inmobiliario.

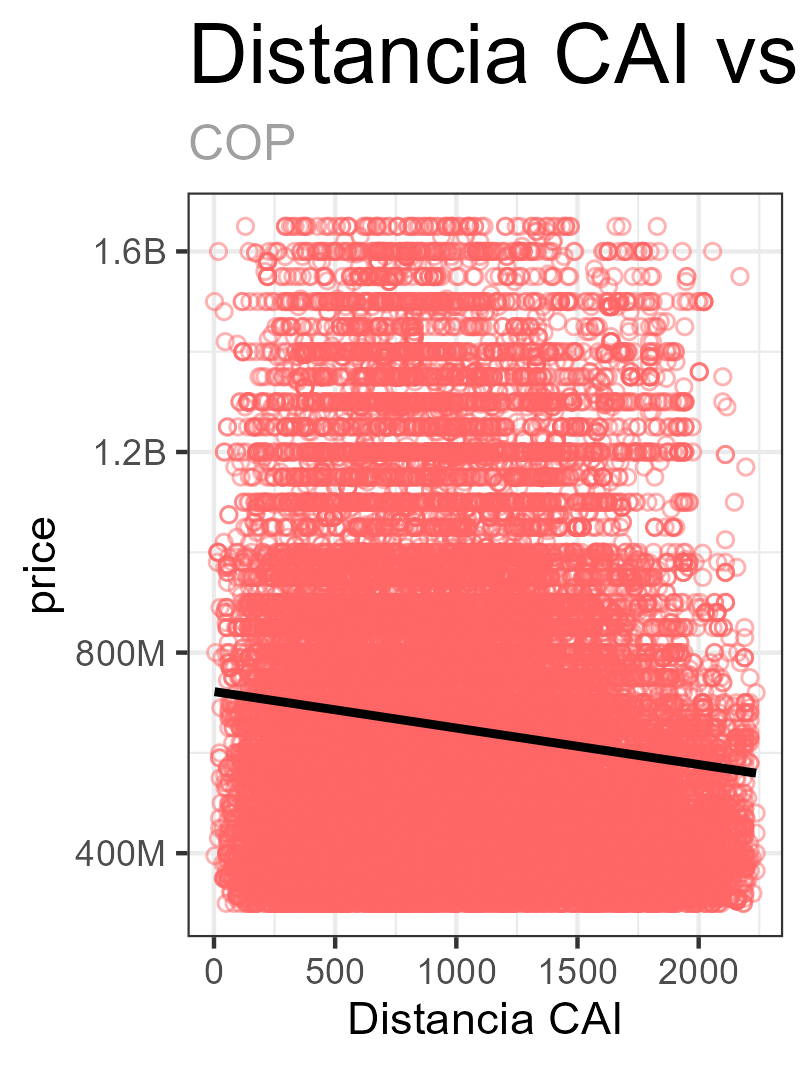
Las gráficas y mapas presentados ofrecen una visión detallada de las relaciones entre diversas variables y los precios de las propiedades en Bogotá, brindando valiosas perspectivas para los analistas y posibles compradores de viviendas.

\*\*Gráfica 1 - Distancias a Transmilenio y Ciclorutas:\*\* Esta gráfica revela una tendencia decreciente en los precios de las propiedades a medida que aumenta la distancia a las estaciones de Transmilenio y ciclorutas. Sin embargo, la disminución de valor se hace más pronunciada a partir de los 2 kilómetros. Esto sugiere que los compradores valoran en gran medida la conectividad y la facilidad de desplazamiento en la ciudad.



\*\*Gráfica 2 - Distancia a Parques y Centros Comerciales:\*\* En el caso de la distancia a los parques, se observa una relación ligeramente creciente, indicando que propiedades ubicadas más lejos de los parques tienden a tener un mayor valor. No obstante, esta relación puede estar relacionada con la seguridad, ya que en áreas menos seguras, la proximidad a parques podría afectar negativamente el precio. En cuanto a los centros comerciales, no se aprecia una relación clara.

\*\*Gráfica 3 - Distancia al CAI más cercano:\*\* A medida que la distancia al Comando de Atención Inmediata (CAI) más cercano aumenta, los precios de las propiedades tienden a disminuir. Esto sugiere que la proximidad a un CAI es un activo importante para los compradores, ya que contribuye a una mayor sensación de seguridad.



\*\*Gráfico 4 - Superficie de la Propiedad:\*\* La gráfica resalta una relación sólidamente creciente entre la superficie de la propiedad y el precio. En términos generales, las propiedades más grandes tienen precios más elevados.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

\*\*Gráfica 5 - Número de Baños:\*\* Se observa un aumento en el precio a medida que se incrementa el número de baños. Esto se alinea con la tendencia en Bogotá, donde propiedades lujosas suelen ofrecer baños privados para cada habitación.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

\*\*Gráfico 6 - Número de Cuartos:\*\* A diferencia de los baños, el número de cuartos no muestra una relación clara y creciente con el precio. Esto podría deberse a que, en cierto punto, agregar más habitaciones podría resultar en cuartos más pequeños y menos cómodos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

\*\*Gráfico 7 - Parqueadero:\*\* La presencia de un parqueadero no parece influir significativamente en el valor de una propiedad. Esto sugiere que la cantidad de parqueaderos es un factor relevante en lugar de la simple existencia de uno.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

\*\*Gráfica 8 - Comparación entre Casas y Apartamentos:\*\* Las casas tienden a tener precios más elevados que los apartamentos, lo que puede deberse a diferencias en tamaño, ubicación y características.

Gráfico, Gráfico de cajas y bigotes

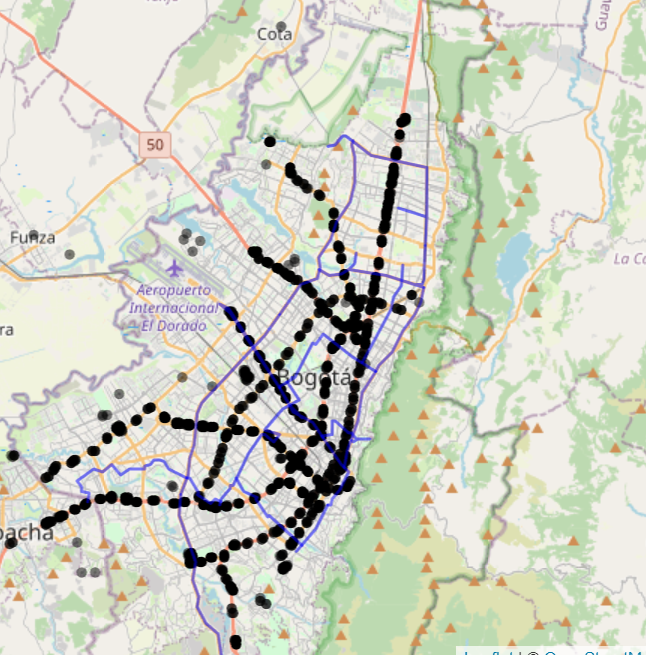
Descripción generada automáticamente

\*\*Gráfica 9 - Penthouse:\*\* Los penthouses claramente poseen un premium sobre otras propiedades, lo que indica que se valora la exclusividad y las características especiales que ofrecen.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

\*\*Mapa 1 - Conectividad de Transmilenio y Ciclorutas:\*\* Este mapa ilustra la intersección de las rutas de Transmilenio y las ciclovías en avenidas principales. Es importante destacar que, aunque las ciclovías tienen una mayor cobertura, no reemplazan a Transmilenio, sino que son complementos, y reflejan las opciones de movilidad de las personas según la proximidad de sus propiedades a estas infraestructuras. La mayoría sigue de todos modos concentrada en el oriente y norte de la ciudad.



Mapa 2