Universidad de los Andes
Big Data y Machine Learning
Maestría en Economía Aplicada
Anamaría Eleonora Rodríguez
Laura Almario
Daniel Orjuela
Daniela Rojas Benitez

Conjunto de problemas 2.

Enlace de Github: https://github.com/laura-Almario/Taller_2_Chapi

1. Introducción. La introducción expone brevemente el problema y si hay antecedentes. Describe brevemente los datos y su idoneidad para abordar la cuestión del conjunto de problemas. Contiene una vista previa de los resultados y las principales conclusiones.

El precio de los inmuebles es, sin duda, un problema que afecta a todas las personas ya sea para vivienda, inversión o fines comerciales, conseguir un precio que ajuste al mercado y, además cumpla con las expectativas de compradores y vendedores no es tarea fácil. Tanto los factores internos del inmueble tales como el área, el número de habitaciones, los baños, los parqueaderos, etc., así como los factores externos como la cercanía a parques, colegios y la facilidad para transportarse pueden afectar el precio de un inmueble. Así, dos lugares con características propias idénticas pueden tener precios muy distintos en función del lugar en el que se encuentran ubicados.

En el presente estudio, buscaremos predecir los precios de los inmuebles en la localidad de Chapinero, para esto, tomaremos los datos de la página web propertati.com.co a partir de los cuales se obtuvieron las muestras de entrenamiento y prueba; igualmente, a través de las descripciones realizadas por los vendedores y el posicionamiento geoespacial buscaremos factores adicionales que puedan afectar el precio.

Limpieza de información

En un primer momento, se realiza la limpieza de la base de datos, para los efectos de este análisis se restringe la información para la localidad de Chapinero en la ciudad de Bogotá.

Con esta información en detalle tenemos:

- La variable parking se construye de la información que ponen los oferentes de bienes en la descripción de sus inmuebles en el portal, en el programa se realizar una lectura de los textos de “garaje” y “parqueadero” y posteriormente se construye la variable teniendo en cuenta si incluye alguna de las palabras en la descripción
- La variable estudio se construye a partir de la lectura en el programa de la descripción para la palabra “estudio”

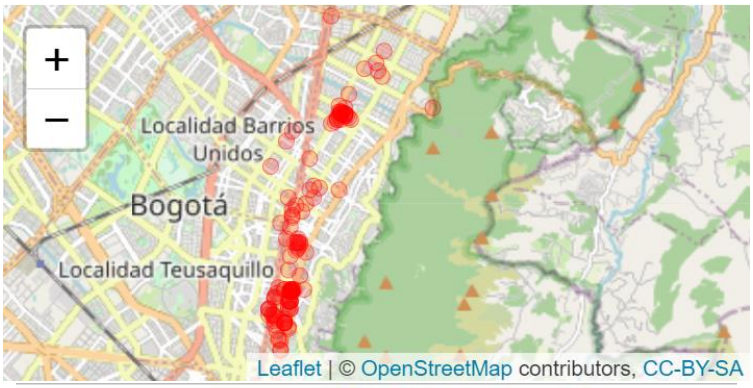
Los atributos del entorno se extraen por medio del programa de la base de datos y visualización geográfica “Open Street Map” para los atributos de paraderos de bus, restaurantes, parques y condiciones de seguridad.

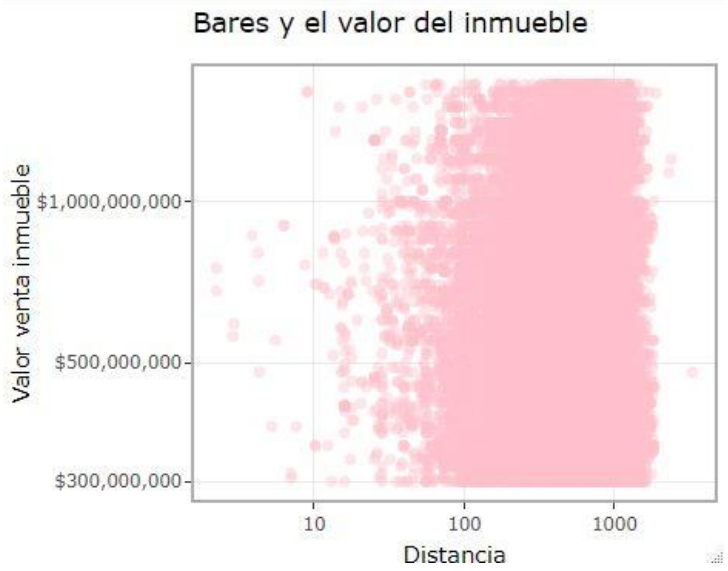
2. Include a descriptive analysis of the data. At a minimum, you should include a descriptive statistics table and two maps with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.

Análisis de las variables:

Cantidad de bares

Se puede observar en Chapinero una zona de alta concentración de bares alrededor de la avenida Caracas. Esta zona se caracteriza por estar cerca de varias universidades y sectores de oficinas, por lo cual no es muy residencial. Por otro lado, se observa una alta concentración en la zona de la 85 y la 93, donde es otro tipo de bares que son de más alto perfil y en las cuadras aledañas si se encuentran zonas residenciales.

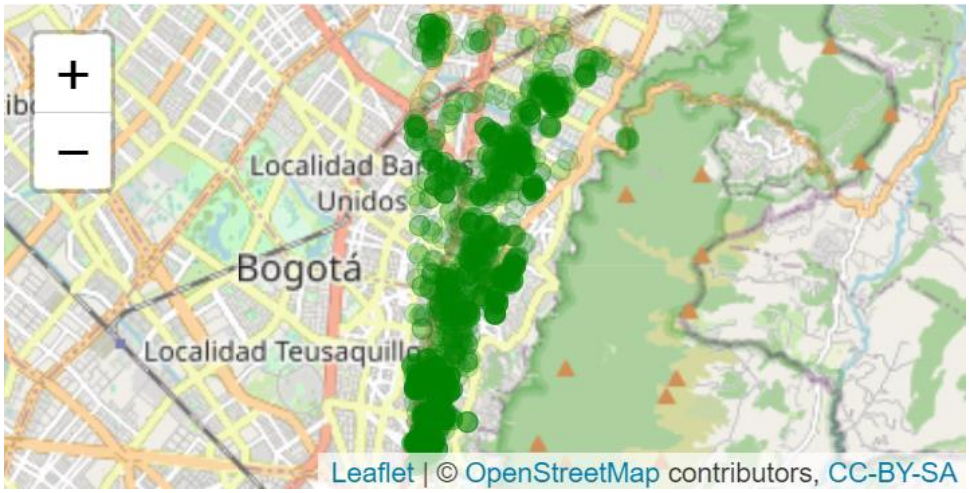


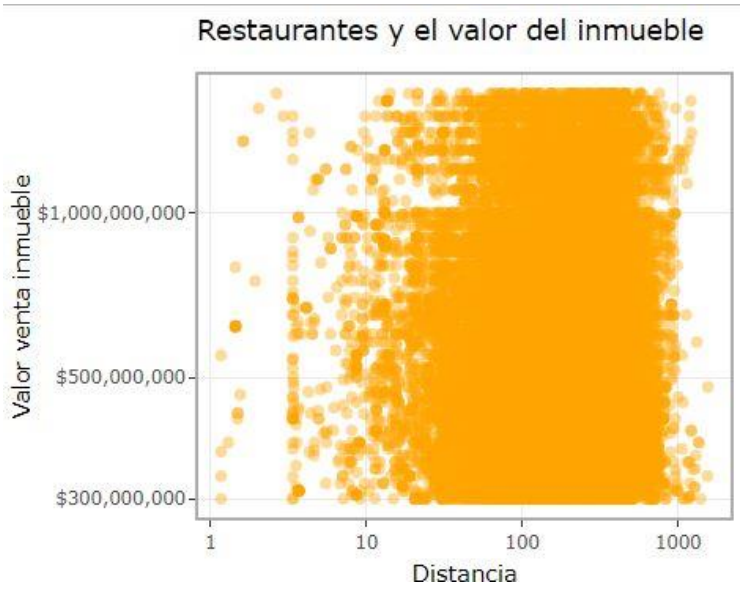


Cantidad de restaurantes

Chapinero es una reconocida en Bogotá por ser una localidad gastronómica de la ciudad. Es por esto que en la distribución del mapa se puede observar una alta concentración de restaurantes dentro de toda la localidad. Cabe resaltar que en las zonas donde se ve más concentración están los lugares más concurridos por los bogotanos, como lo puede ser el Parque de la 93, la zona aledaña a la Universidad Javeriana, la Zona G y la 85. Asimismo, como

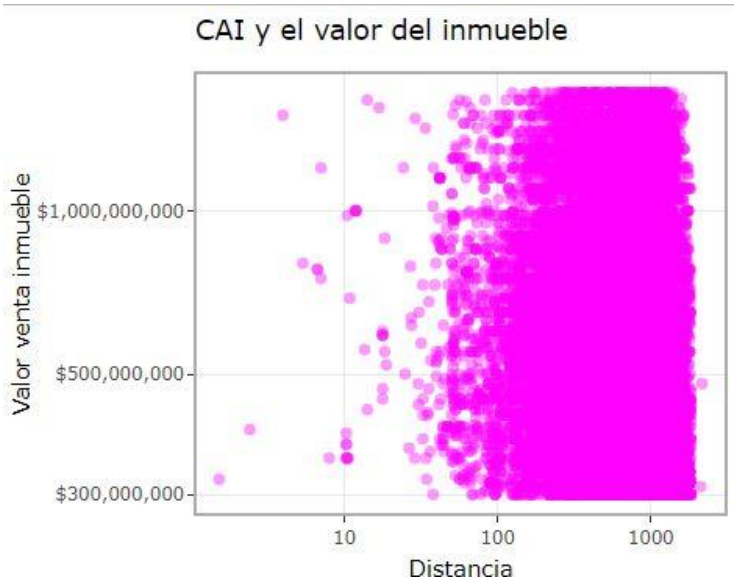
se puede observar, en comparación con el mapa de bares, estos también se encuentran en las principales zonas de los restaurantes, confirmando la importancia en los planes de los bogotanos de encontrar lugares para comer y rumbear en la misma zona.





Seguridad: Centros de Atención Inmediata CAI

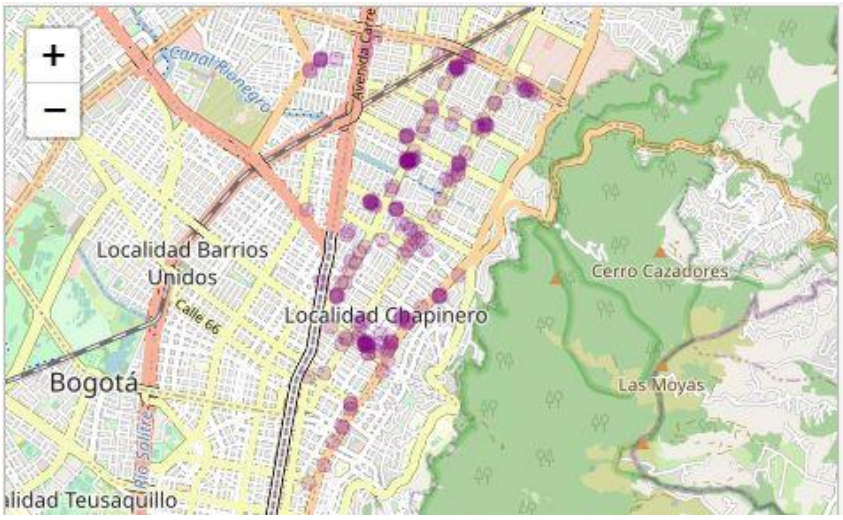
Chapinero es una de las localidades con un alto nivel de robo. De acuerdo a infobae, Chapinero fue la localidad más afectada en 2022, en donde el incremento, con respecto al 2021 fue del 65%, pasando de 1.433 hurtos a 2.366. Esto se puede analizar respecto a la cantidad de centros de atención inmediata en la localidad al igual que locales, bares y restaurantes que aumentan la cantidad de personas y oportunidades de robo. Y es que, comparado con la ubicación de estos centros de ocio en la localidad, los CAI no necesariamente se encuentran en ubicaciones cercanas a todos, alargando los tiempos de respuesta y la posibilidad de tener una atención realmente inmediata.

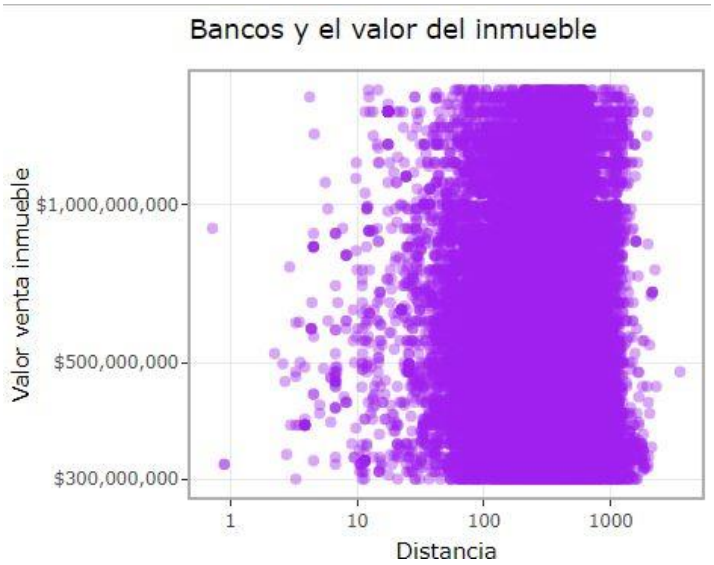


Fuente: Elaboración propia con datos de open street maps.

Bancos

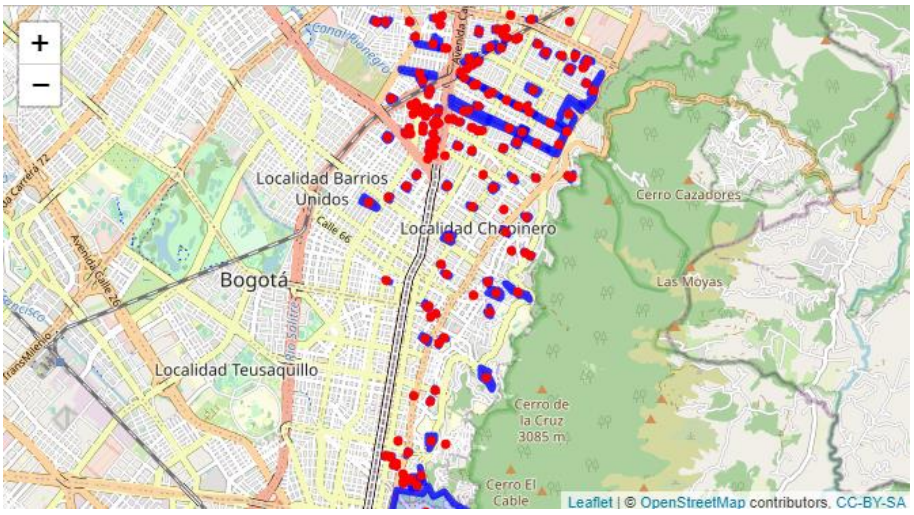
Chapinero cuenta con la zona conocida como la zona financiera de Bogotá, teniendo a la bolsa de valores y las oficinas de servicios financieros. La importancia de este sector para Bogotá es una de las razones por las cuales es importante priorizar la variable. Su relevancia para todas las industrias y la atracción de personas a la zona representan razones por las cuales se puede incrementar o bajar el precio de la zona.

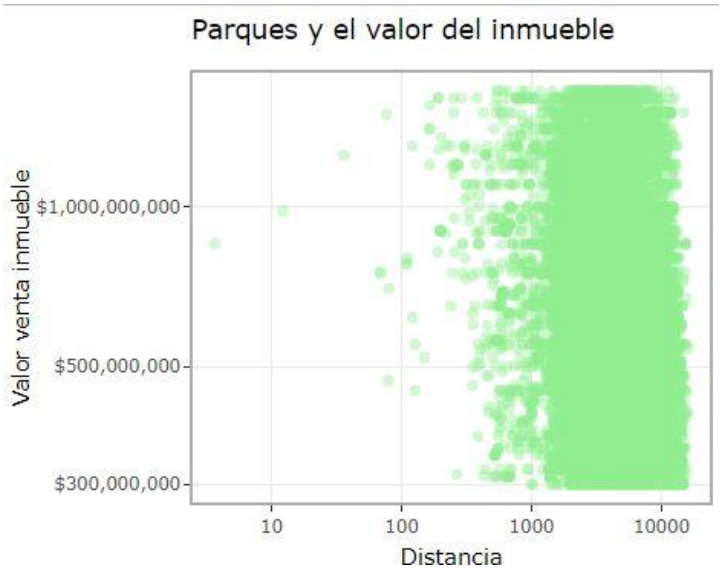




Fuente: Elaboración propia con datos de open street maps.

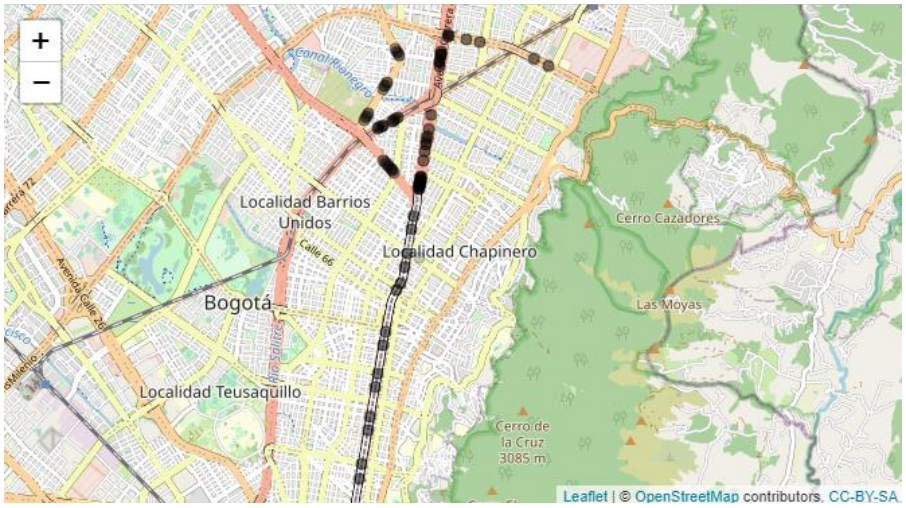
Parques Chapinero

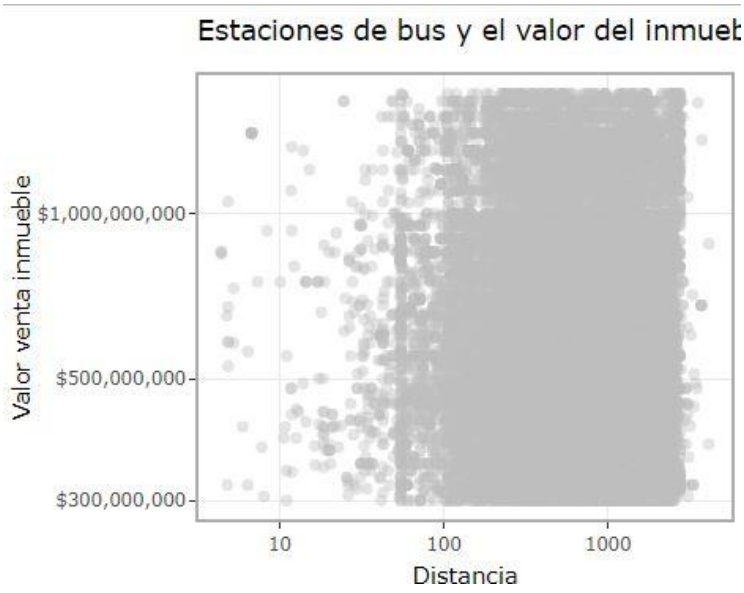




Fuente: Elaboración propia con datos de open street maps.

Estaciones de bus





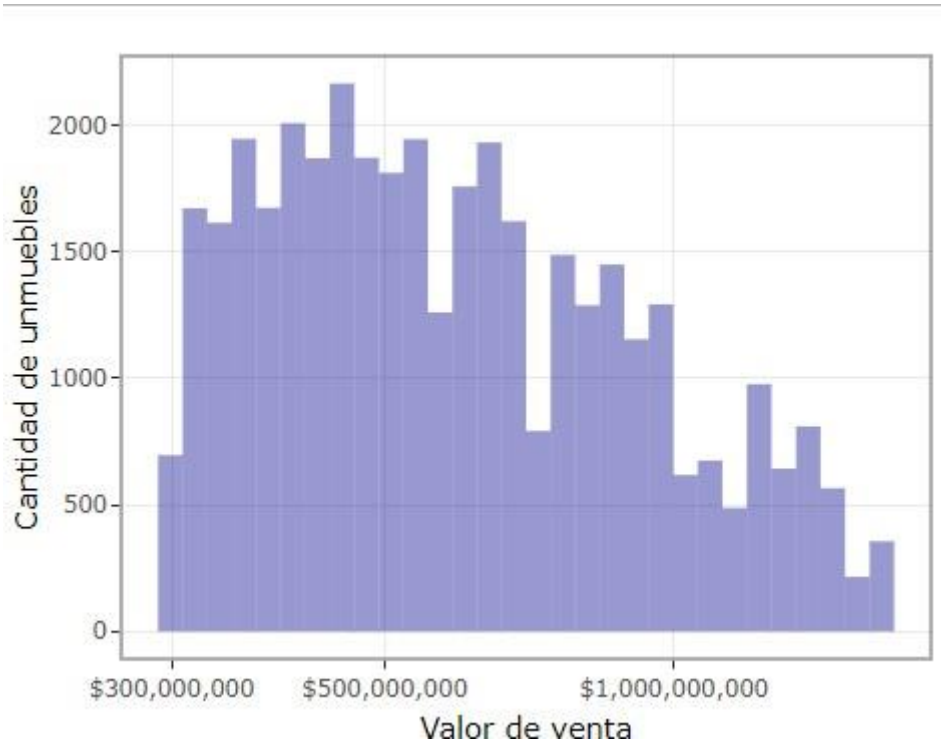
Reconocemos que, por la calidad de actividad comercial, de consumo y residencial, la comparación con otras localidades puede no ser la mejor manera de predecir, sin embargo, usaremos la totalidad de la ciudad de Bogotá para predecir los precios de la localidad de Chapinero.

Finalmente, para poder asegurar que las bases de datos estén completas, terminamos la limpieza de la base de datos identificando las variables que contienen datos NA y al ser una proporción pequeña de la muestra, realizamos un análisis exhaustivo sobre estas variables para determinar la completitud de la información.

Este ejercicio lo realizamos para las variables baño y área total. Para la primera variable, realizamos una lectura del número de datos que tienen NA y al ser menos del 5% de la base de datos, realizamos una imputación al promedio de estas observaciones. Por su parte, al revisar la información de la superficie total del inmueble, encontramos que la información faltante era significativa, por lo que utilizamos la información contenida en la descripción y realizamos la imputación teniendo en cuenta esos datos.

Estadísticas descriptivas

En este apartado, nos interesa conocer cómo se comporta la información que tenemos en la base de datos, por tal revisamos la cantidad de inmuebles para cada rango de precio.



Fuente: Elaboración propia

Del Histograma anterior extraemos que, la mayoría de los inmuebles se concentran a la izquierda del gráfico, lo que indica que, a medida que aumenta el precio, hay menor cantidad

de inmuebles para ese precio, encontrando que la mayoría de los inmuebles se sitúan cerca de los $460.000.000 COP.

Estadísticas	Precio	Superficie total	Superficie cubierta	Cuartos	Habitaciones	Baños
Cantidad	38644,00	38644,00	8565,00	20384,00	38644,00	38644,00
Min	300000000,00	2,00	2,00	1,00	0,00	1,00
Max	1650000000,00	76610,00	1336,00	11,00	11,00	13,00
Rango	1350000000,00	76608,00	1334,00	10,00	11,00	12,00
Mediana	559990000,00	523,25	108,00	3,00	3,00	3,00
Media	654534675,29	810,62	131,93	3,01	3,14	2,96
Desviación estándar	311417886,95	2253,68	76,62	1,37	1,53	0,97

Estadísticas	Distancia a bar	Distancia a restaurante	Distancia a Banco	Distancia a CAI	Distancia a trasporte	Distancia a parque
Cantidad	38644,00	38644,00	38644,00	38644,00	38644,00	38644,00
Min	2,23	1,17	0,72	1,53	4,42	3,67
Max	3318,14	1551,93	3525,55	2165,18	4114,00	15944,44
Rango	3315,91	1550,76	3524,83	2163,65	4109,58	15940,77
Mediana	513,41	189,99	308,66	686,88	730,23	4645,90
Media	565,40	223,93	366,44	725,34	888,65	5072,16
Desviación estándar	323,59	158,94	252,09	373,16	633,11	2574,69

Fuente: Elaboración propia

Modelos y Resultados

En este trabajo, se estiman 4 modelos de predicción de información con 9 variables explicativas.

Variables incluidas en el modelo

El modelo teórico de precios hedónicos sugiere la inclusión como variables independientes de variables características del inmueble y del entorno, por tanto, como parte de las variables del inmueble se incluyó:

- Superficie total del inmueble en metros cuadrados
- Número de habitaciones
- Número de baños
- Tipo de propiedad (casa o apartamento)
- Parqueadero
- Estudio

Estas variables presentan la ventaja de generar un criterio sobre las amenidades y el espacio del inmueble que se pretende vender o arrendar. Por otro lado, como variables del entorno se incluyó:

- Distancia a los paraderos de bus
- Distancia a los parques

Estas dos variables indican la cercanía del inmueble al espacio público de ocio en este caso los parques y a los puntos de acceso de las estaciones de Transmilenio.

Los modelos de predicción que se estiman incluyen las mismas variables para facilitar la comparación bajo los criterios estadísticos, estos son:

1. El primer modelo que se estima es una regresión lineal simple con las variables que se mencionaron anteriormente, con esto se genera una predicción del ajuste del modelo para los valores de precio en la base de datos de entrenamiento.
2. El segundo modelo que se estima es un modelo de Árbol, en el cual se usa un modelo de rpart y se establece un tunelength de 5, es decir, explora 5 combinaciones diferentes de hiper parámetros. Igualmente, se hace una partición de 5 para realizar el cross validation.
3. El tercer modelo que se estima es el modelo de árbol con regresión lineal, el cual se usa bajo los mismos parámetros del modelo anterior, con único cambio es que el método usado es de regresión lineal.
4. El cuarto modelo es el modelo de Árbol con el proceso de Random Forest, este fue el modelo más ajustado, es decir con mínimo error de los que se estimaron para obtener la predicción sobre el valor de los precios del inmueble. En este modelo, se determina un mtry de 3 y para el parámetro de min.node.size se pone 5, y como parámetro de partición de la muestra de entrenamiento, se establece la varianza.
5. Por último, se estimó un modelo de Árbol con el proceso de Random Forest y boosting, con los mismos parámetros del modelo anterior, con el objetivo de revisar si esto pudiese genera mejores resultados.

Para cada modelo se estimaron los estadísticos de RMSE, R2 y MAE. Sin embargo, la decisión final de elección del modelo se hace de acuerdo con el ajuste de la competencia del score de la plataforma Kaggle.

Tabla con los estadísticos de resultado

Modelo	**Score kaggle**
Regresión Lineal Simple	377305776,18619
Árbol	314529235,17245
Árbol con Regresión Lineal	288393175,96712
Árbol con proceso de Random Forest	211649199,98799

Fuente: elaboración propia.

Conclusiones y recomendaciones

A modo de conclusión, el modelo de Árbol con el proceso de Random Forest (modelo 4) se destacó como el más ajustado y preciso, superando a los otros modelos estimados en términos de error. La capacidad de capturar relaciones no lineales y considerar múltiples características en cada árbol permitió obtener una predicción más precisa sobre los precios de los inmuebles.

De lo anterior, concluimos que el mejor modelo es el Número 4, éste es Árbol con el proceso de Random Forest, el cual nos arroja el menor RMS.

RMSE	Rsquared	MAE
0.3237857	0.5500876	0.2616759

Para finalizar, la estimación de predicciones con los diferentes modelos vistos en clase tiene muchas condiciones previas para poder llegar a la estimación. El proceso de limpieza y consolidación de base de datos es fundamental para poder tener estimaciones con mayor predicción y con mejores datos, esto fue un aprendizaje importante para la elaboración de este trabajo.

Al momento de correr los modelos, será de especial importancia que los investigadores conozcan las ventajas de generar predicciones con los distintos modelos y los parámetros que se deben establecer. En conclusión, realizar múltiples testeos de prueba y error para realizar la verificación de la calidad en las predicciones es determinante en tener un resultado acorde con lo que se espera de un análisis de este tipo.

Nota: otra información relevante será archivada en el repositorio de Github.