**VIVIAN CABANZO FERNÁNDEZ**
**LAURA DANIELA DIAZ TORRES**
**CRISTIAN FELIPE MUÑOZ GUERRERO**
**ZENNETH OLIVERO TAPIA**

**Problem Set 3 – Making Money with ML?**

**BIG DATA Y MACHINE LEARNING PARA ECONOMIA APLICADA**

**2025-02**

data_equipo_08

**Limpieza de datos, construcción y transformación de variables, definición de bases de datos finales.**

Universidad de los Andes

# Best model overview – Random Forest CV espacial

| Elemento | Descripción |
|---|---|
| Tipo de modelo | Random Forest con validación cruzada espacial (ensamble por folds) |
| Objetivo / Métrica | Minimizar MAE (validación 5-fold) |
| Hiperparámetros | num.trees = 1000; mtry ≈ √p; min.node.size = 10; sample.fraction = 0.80; splitrule = variance; seed = 2025 |
| Esquema de validación | Regular 5-fold: MAE = 120,886,984; Espacial 5-fold: MAE = 131,308,700 |
| Estrategia de predicción | Promedio de las K predicciones por fold espacial (ensamble por folds) |
| Top variables | surface_total, bedrooms, estrato, dist_centros_comerciales, has_parqueadero_garaje, localidad |
| Kaggle (public LB) | 209,389,203 |

Usando el CV espacial se generalizaron mejor las zonas no vistas. El ensamble por folds reduce varianza y el bosque capta no linealidades e interacciones entre tamaño, ubicación, distancias y amenidades, entregando predicciones estables sin sobreajuste.

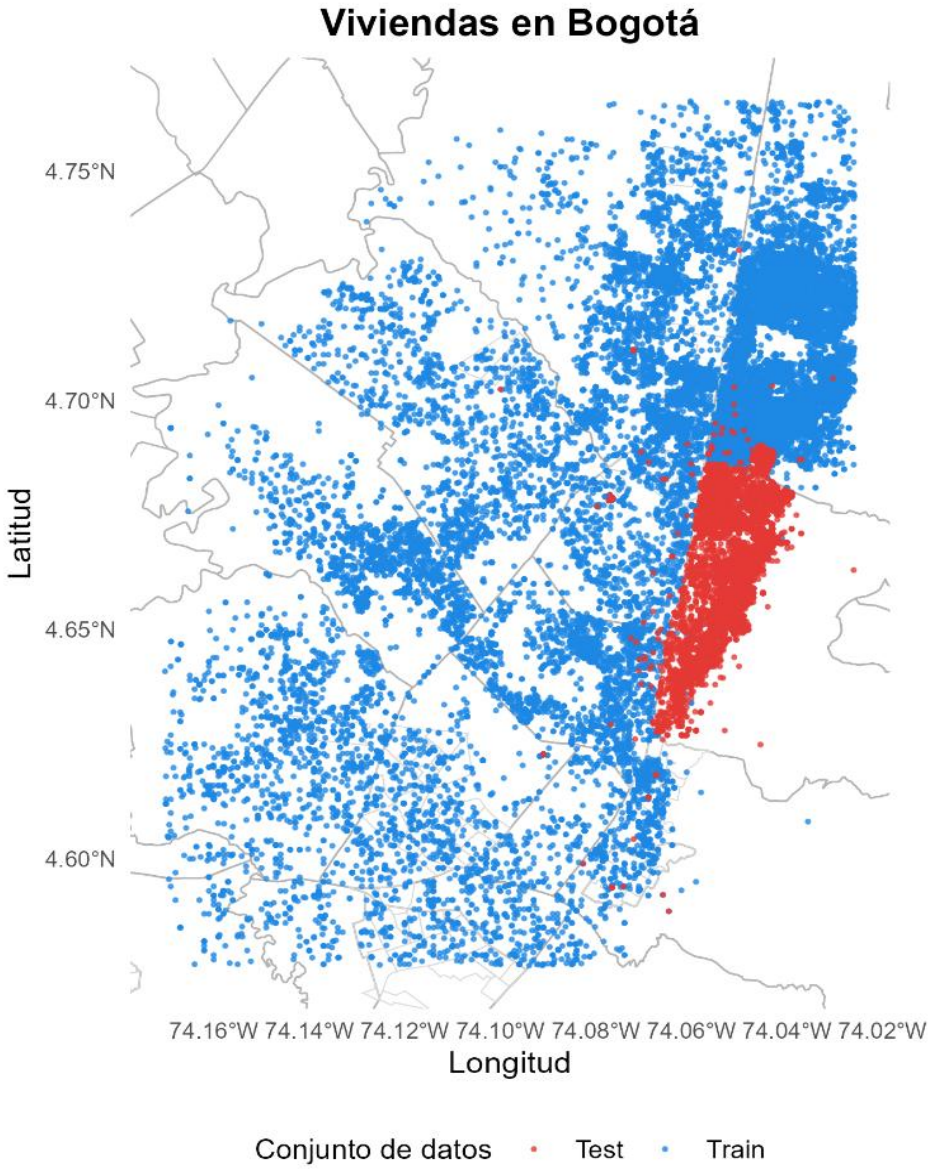
# BASES DE DATOS INICIALES

Las bases de datos iniciales de entrenamiento y de prueba, fueron construidas a partir de los datos recopilados del portal web https://www.properati.com.co

- train.csv cuenta con 38.644 observaciones y 16 variables
- test.csv cuenta con 10.286 observaciones y 16 variables

| VARIABLE | DESCRIPCIÓN | VARIABLE | DESCRIPCIÓN |
|---|---|---|---|
| **Property_id** | Identificador único asignado a cada propiedad, usado para diferenciar registros y evitar duplicados. | **Bedrooms** | Cantidad de habitaciones destinadas a dormitorio. |
| **City** | Ciudad en la que se ubica la propiedad según la publicación en el portal. | **Bathrooms** | Número total de baños disponibles en la propiedad. |
| **Price:** | Precio de venta ofertado para la propiedad, reportado directamente por el anunciante. | **Property_type** | Tipo de propiedad anunciada (casa o apartamento). |
| **Month** | Mes de publicación del anuncio en el portal. | **Operation_type** | Tipo de operación del anuncio (venta o arriendo). |
| **Year** | Año de publicación del anuncio. | **Lat** | Latitud geográfica de la propiedad. |
| **Surface_total** | Superficie total del inmueble en metros cuadrados (m2). | **Lon** | Longitud geográfica de la propiedad, utilizada para ubicar espacialmente el inmueble. |
| **Surface_covered** | Superficie construida o cubierta del inmueble en m2. Puede diferir de la superficie total. | **Title** | Titulo del anuncio publicado en el portal, usualmente con información breve y destacada. |
| **Rooms** | Número total de ambientes o espacios de la vivienda (sala, comedor, estudio). | **Description** | Descripción de texto del anuncio con detalles adicionales sobre características del inmueble. |

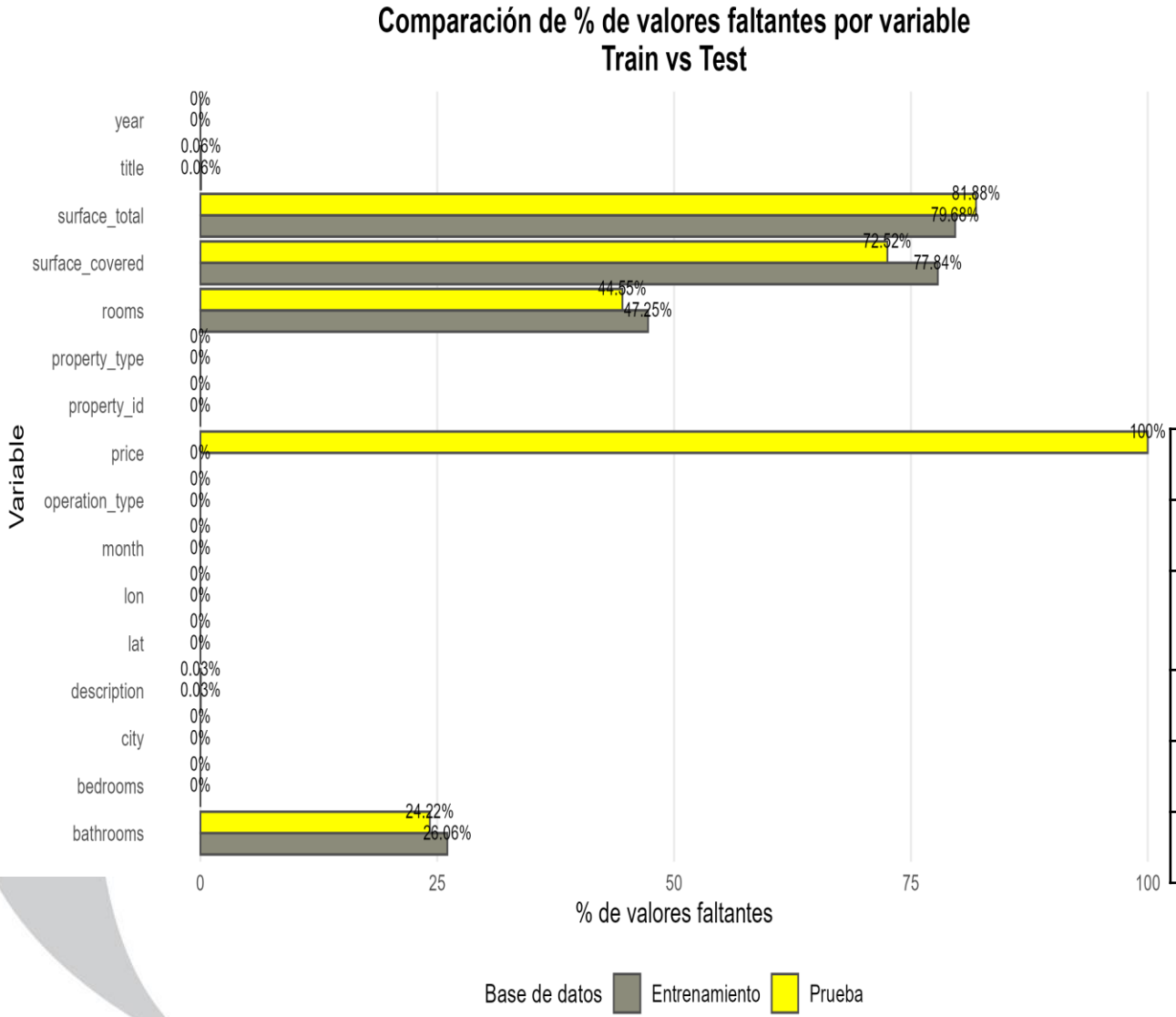
# DISTRIBUCIÓN ESPACIAL DE LAS BASES




Exploración espacial de las bases de datos, de acuerdo con las variables de Longitud y Latitud.

- Viviendas con información para el entrenamiento de modelos – Bogotá.
- Viviendas sobre las cuales se realizarán predicciones de precios de venta – Localidad de Chapinero.

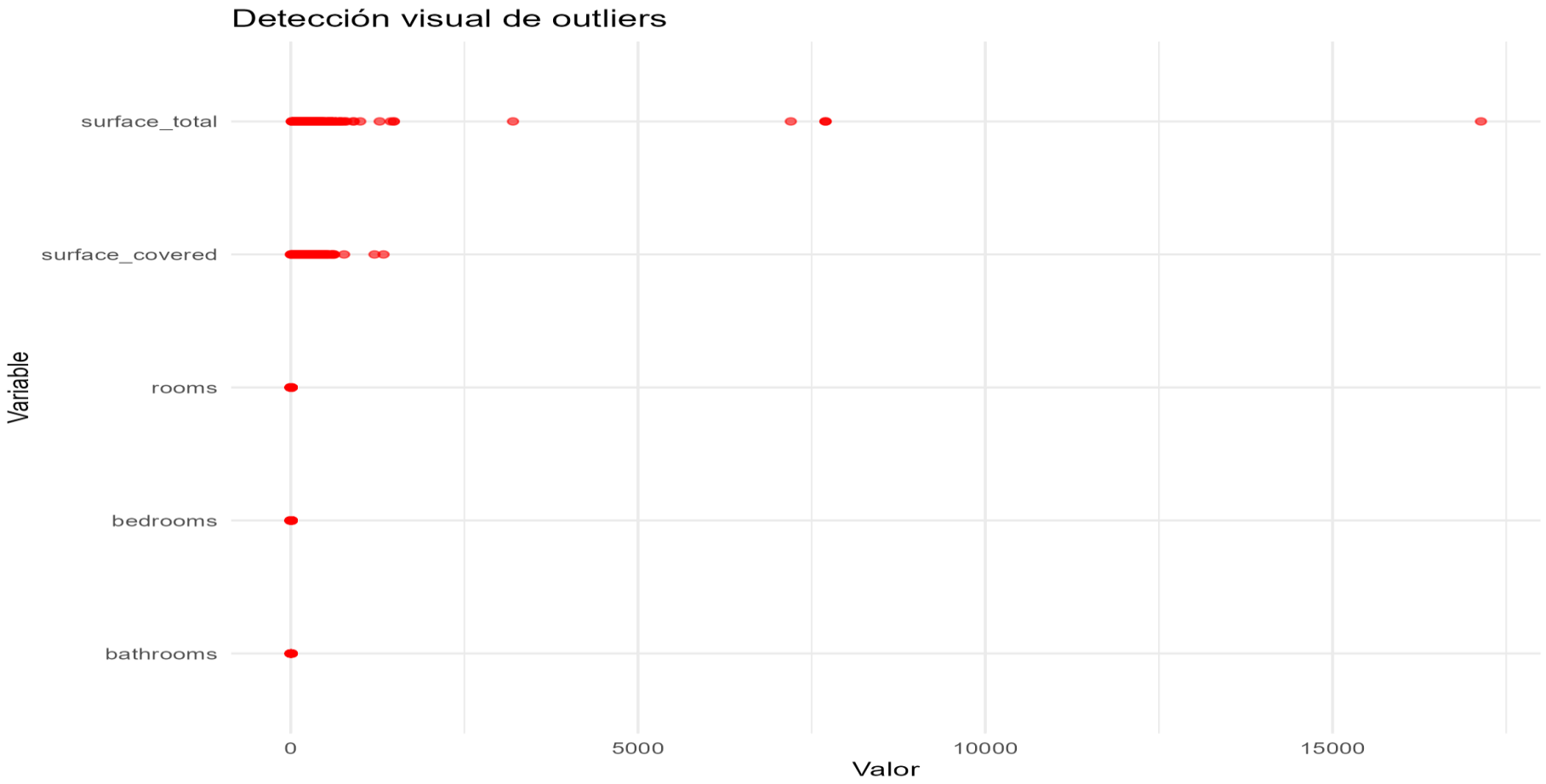
# DATOS FALTANTES POR BASE Y VARIABLE




Porcentaje de datos faltantes en las bases de entrenamiento y de prueba:

| Variable | Train | Test |
|---|---|---|
| Surface total | 79,68% | 81,88% |
| Surface covered | 77,84% | 72,52% |
| Rooms | 47,25% | 44,25% |
| Bathrooms | 26,06% | 24,22% |
| Price | 0% | 100% |

# VALORES ATÍPICOS




La variable Surface_total a pesar de tener un porcentaje alto de faltantes, también presenta valores atípicos altos, con superficies superiores a los 2.000 m2.

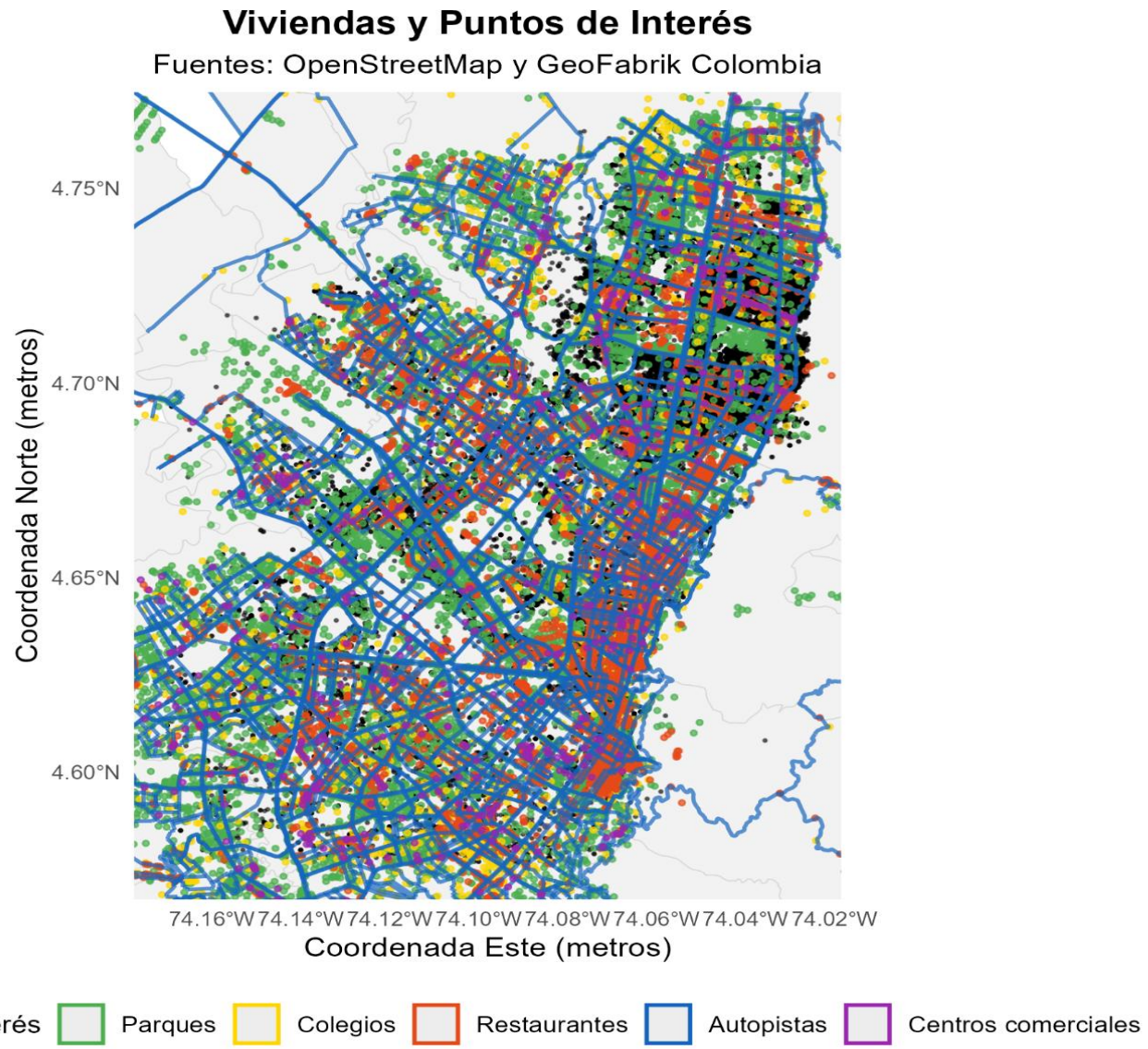
Universidad de los Andes

# ADECUACIÓN DE LOS DATOS Y CONSTRUCCIÓN DE NUEVAS VARIABLES

| PROCESO | DESCRIPCIÓN |
|---|---|
| Procesamiento de texto | - Normalización y unificación de texto<br>- Diccionario de palabras con atributos clave (parqueadero, balcón, patio, etc.)<br>- Tokenización de palabras para identificar su presencia y frecuencia<br>- Creación de 19 variables dicotómicas adicionales |
| Análisis espacial y geográfico | - Información de OpenStreetMap – OSM de Colombia<br>- Delimitación de la ciudad de Bogotá<br>- Identificación de puntos de interés urbano (parques, colegios, centros comerciales, restaurantes y autopistas), calculo de la distancia de las viviendas a dichos puntos.<br>- Creación de 7 variables adicionales |
| Inclusión del estrato socioeconómico | - Información del DANE:<br>- i). Un archivo geoespacial LOTE.gpkg<br>- ii). Una tabla estrato.csv<br>- El proceso crea 1 variable que asocia a cada lote su estrato correspondiente. |

Universidad de los Andes

# ANÁLISIS GEOESPACIAL



# TRATAMIENTO DE DATOS FALTANTES Y VALORES ATÍPICOS

| PROCESO | DESCRIPCIÓN |
|---|---|
| Eliminación de observaciones | - 414 observaciones con descripción que indica que la propiedad no se encuentra en Bogotá a pesar de que su ciudad de registro indica que si.<br>- 1.731 por el buffer o franja de 500 metros alrededor de Chapinero en la base de entrenamiento para evitar solapamiento y autocorrelación espacial.<br>- 4.426 observaciones idénticas, análisis realizado por combinación de titulo, descripción, precio, latitud y longitud. |
| Imputación de Surface total y superficie construida | - Se identificaron los datos mayores a 1.500 m2, se reemplazaron con NA para no perder la información de su registro.<br>- Se asignó el valor de superficie construida cuando esta superaba a la superficie total reportada.<br>- Si una de las dos superficies estaba disponible, se utilizó dicho valor para completar la variable faltante.<br>- Los casos restantes se imputaron empleando la mediana de superficie, estimada no por lo valores totales sino por tipo de propiedad y número de habitaciones. |
| Imputación de Baños | - Se implementó una regla de imputación basada en la cantidad de habitaciones y el tipo de propiedad, bajo el supuesto que toda vivienda debe contar con al menos un baño. |
| Imputación de Ambientes – romos | - De manera análoga, se imputó la cantidad mínima de ambientes en función del tipo de propiedad, garantizando la presencia de, por lo menos, un espacio correspondiente a sala–comedor. |

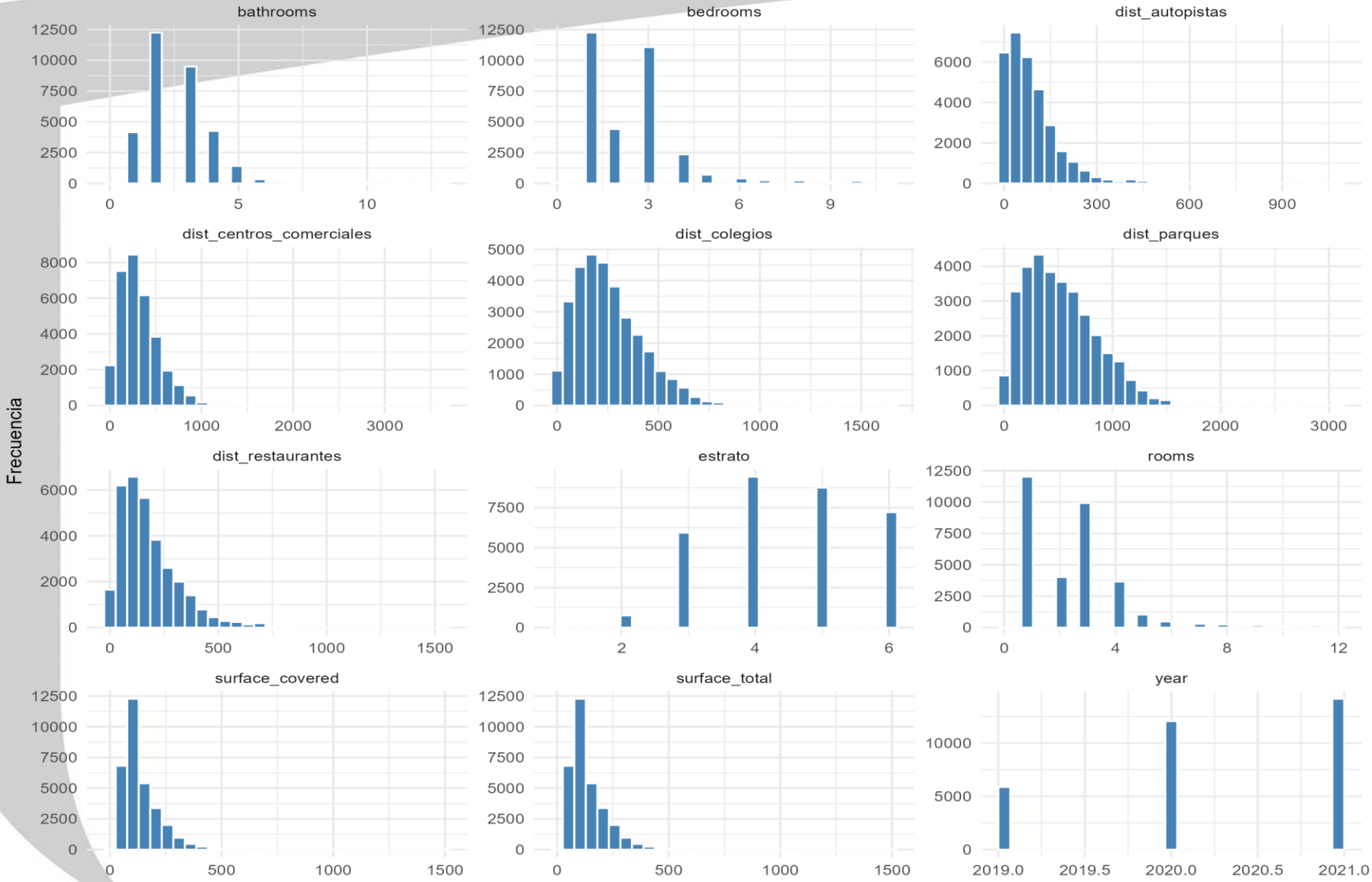
# BASES DE DATOS FINALES

En resumen: La información final se obtuvo mediante la unión y depuración de las distintas fuentes disponibles, incorporando transformaciones de variables, extracción automatizada de atributos a partir del texto, y la creación de variables relacionadas con características físicas, atributos geográficos y condiciones socioeconómicas.

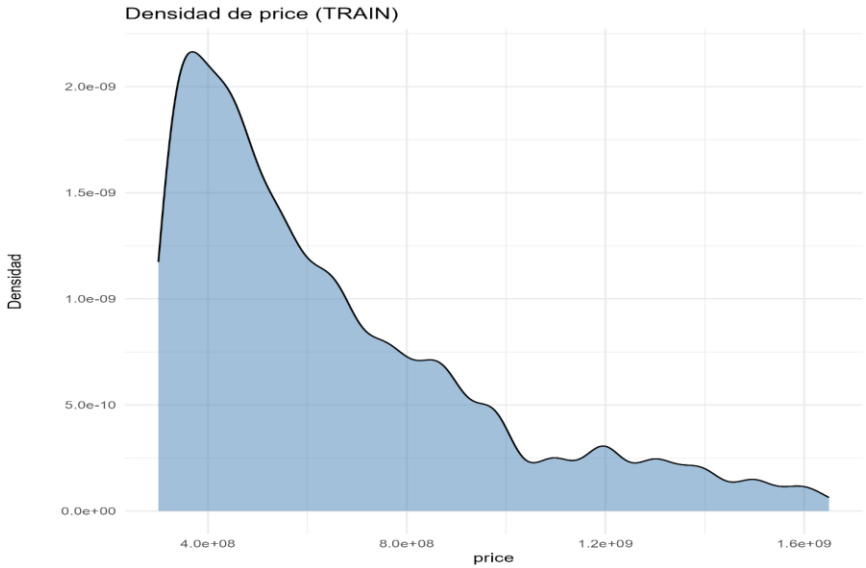
| CATEGORÍA | CANTIDAD | VARIABLES |
|---|---|---|
| Variables Iniciales | 16 | Identificación de la propiedad, ciudad, precio, mes, año, superficie total, superficie construida, ambientes, baños, habitaciones, tipo de propiedad, tipo de operación del anuncio, latitud, longitud, título y descripción. |
| Variables dicotómicas a partir de texto | 19 | Presencia de: parqueadero o garaje, seguridad, ascensor, gimnasio, piscina, BBQ, salón social, zona infantil, balcón, terraza, patio, jardín exterior, chimenea, cocina integral, depósito, estudio, remodelado, vista y pisos de vivienda. |
| Variables de distancia a puntos de interés | 5 | Distancias a parques, colegios, restaurantes, autopistas y centros comerciales. |
| Otras variables | 3 | Localidad, barrio y estrato |
| TOTAL | 43 | Variables en train_final.csv y test_final.csv |

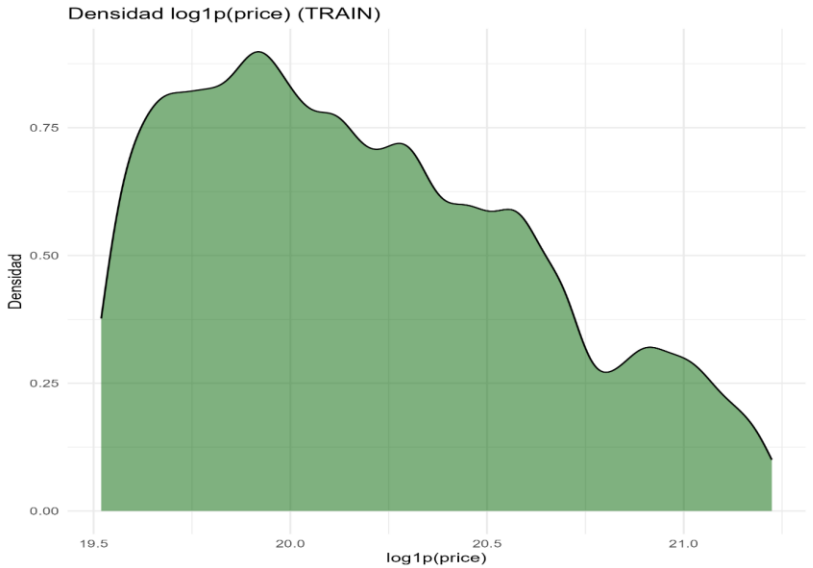
# ESTADÍSTICAS DESCRIPTIVAS

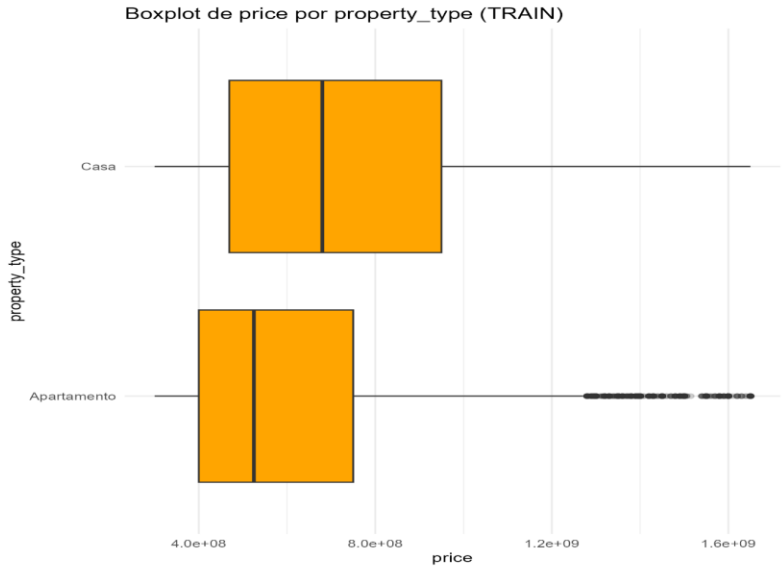
Histogramas de variables númericas de la base de entrenamiento

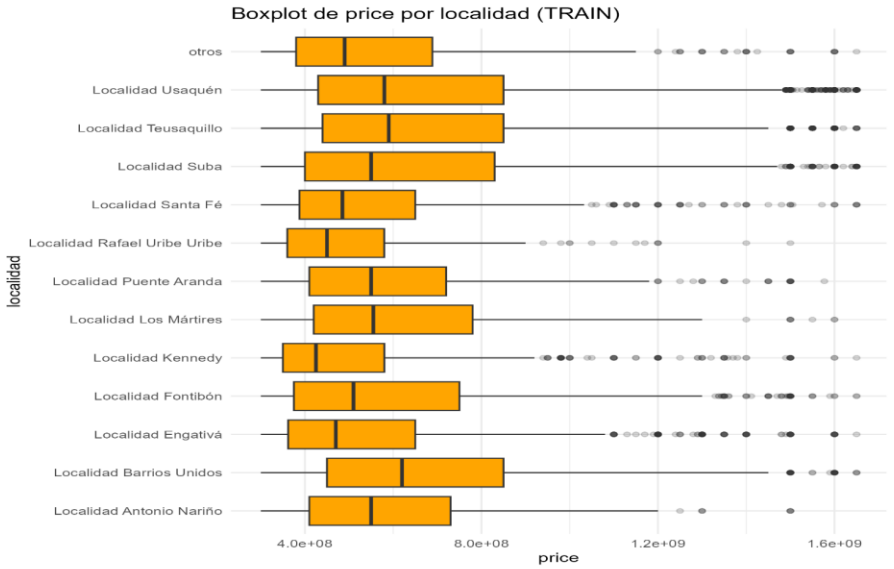


# ESTADÍSTICAS DESCRIPTIVAS



Universidad de los Andes

# RELEVANCIA DE LA INFORMACIÓN PARA LA PREDICCIÓN DE PRECIOS DE VIVIENDA EN CHAPINERO

Las bases de datos consolidadas, resultado de los procesos previamente descritos, son pertinentes para la predicción de precios de venta en la localidad de Chapinero, debido a que integran información que captura de manera adecuada los factores que explican la variación del valor de las viviendas en Bogotá.