Anexo 1 - Procesamiento de la base de datos

A continuación, se presenta la descripción del procesamiento de datos realizado para obtener la base de datos descrita en el numeral 3 del documento:

1. Exploración de datos

Se unieron las bases ***train*** y ***test*** descargados de la página web de la competencia en Kagel[1] para luego poder hacer filtros por ubicación, en caso de que fuera necesario, y se creó la variable "*sample"* para identificar la fuente de cada una de las bases.

En una exploración inicial a los datos, se observa que hay una gran cantidad de valores omitidos (NA) en las variables *"surface_total", "surface covered", "rooms", "bathrooms"*, además de algunas en *"title"* y "*description*". Además, más del 50% de las viviendas (casas y apartamentos) tienen un precio menor a 6 millones de pesos. Por otro lado, al visualizar los precios de las viviendas respecto al área, podemos notar muchos valores atípicos, esto ya que, en términos generales, uno esperaría que los precios de las viviendas sean mayores a medida que incrementa el área. Sin embargo, esto también se puede deber a la heterogeneidad de las viviendas, por la ubicación y otras características. Otra característica importante es que el 76% de la base ***train*** corresponde a información de Apartamentos, mientras que en la base ***test*** es el 97%. Finalmente, analizamos la variable "*description*", donde encontramos información valiosa en forma de texto (chr), a partir de la cual se extrajo información a partir del procesamiento del lenguaje natural.

2. Creación de variables

Para poder obtener las características físicas de las viviendas, se optó por extraer la información de la variable *"description"*, siguiendo los siguientes pasos:

1. **Stopwords**: se eliminaron las *stopwords* de la variable *"description"*.
2. **Tokenización**: para 1, 2 y 3 palabras. Se mantuvieron las palabras repetidas en los tokens de 1.
3. **Lista de variables de interés:** se hizo una lista de variables que podrían afectar el precio de las viviendas. Para esto, nos basamos en características relevantes comunes en portales inmobiliarios y en la bibliografía de referencia (Herart & Maier, 2010). Creamos un *loop* que busca cada una de estas variables de interés en los tokens de cada observación. Si encuentra esa variable en los tokens, rompe el *loop* y crea una dummy con un valor de 1.

3. Imputación de Datos

Para reducir los valores omitidos de las variables de área y baños, se aplicaron *loops* en los tokens de 2 palabras, se capturaron los números que estuvieran en los tokens con las palabras de interés (e.g. bano, bao, baos, mts, mts, m2…). Además, se pusieron rangos máximos como filtros para evitar errores. Finalmente, en el caso de baños, para las observaciones donde no se encontró valores numéricos se aplicó un *loop* en los tokens de 1 palabra y se realizó un conteo de las veces que la palabra de interés estuviera presente (bano, banos, bao, baos).

[1] https://www.kaggle.com/competitions/uniandes-bdml-202313-ps2

Una vez se capturaron esos datos, se pudo reducir los valores faltantes de ambas variables (de 30,790 a 21,099 para área y de 10,071 a 3,826 para baños). Para el resto de los valores, se realizó una imputación de vecino más cercano, utilizando el paquete *mice.*

4. Tratamiento de valores atípicos de las variables

Con el objetivo de evaluar los valores atípicos (outliers), se utilizó una gráfica de bigotes y una prueba de outliers estandarizada para las variables *"price", "bedrooms", "banos" y "area"*. Se hizo una exploración de estos datos atípicos capturados, teniendo en cuenta otras características de las propiedades incluyendo *"property_type", "localidad" y "description"*. Se encontró que las observaciones con muchas habitaciones y baños eran aquellas que eran casas comerciales, con múltiples habitaciones y en localidades como Suba, Barrios Unidos y Engativá. Por otro lado, los *outliers* en precios se presume son debido a falta de puntos (e.g. un apartamento de 17350 m2 se esperaría que fuera en realidad 173.50m2). Sin embargo, debido a que no es posible realizar estas presunciones, se decidió eliminar todos estos valores.

5. Incorporación de datos del entorno a partir de la unión de datos espaciales

Para complementar la base de datos se realizó la identificación de variables relevantes en las bases de datos de la administración distrital y nacional, entre las que se destacan:

- Identificación de áreas de planeamiento y división espacial (Localidades, UPZ, Barrios).
- Datos catastrales (Estrato socioeconómico, Valores de referencia del suelo).
- Indicadores de turismo (Densidad de establecimientos de Gastronomía y Bar, Densidad de establecimientos de alojamiento turístico, identificación de zonas de interés Turístico).
- Indicadores de criminalidad (Numero de incidentes delictivos y delitos de alto impacto reportados en diferentes categorías).
- Indicador de espacio público efectivo.
- Datos poblacionales (Densidad poblacional, Numero de habitantes y Numero de viviendas en la manzana).
- Ubicación de servicios relevantes (Parques, Colegios, Centros Comerciales, Estaciones de transporte publico).

Una vez descargados estos datos, se realizó la unión espacial de los datos en R usando la librería *sf*, de manera tal que los puntos geográficos tomaran los valores del sector catastral, upz o manzana en la que se encuentran según la escala de agregación disponible. Para identificar la distancia de los inmuebles a los servicios relevantes se utilizó nuevamente la librería *sf* para calcular la distancia de cada inmueble con el punto de interés mas cercano en cada caso.

En la Tabla 1 se detallan las bases utilizadas, así como las variables de interés extraídas, su escala de agregación, fuentes y enlaces de descarga de los archivos utilizados los cuales también se pueden descargar en este enlace: https://drive.google.com/file/d/1pvOZqn-tUOfV-S8v01EGpBHa63mAj9E/view?usp=sharing

Tabla 1.Identificación de Fuentes externas

Nombre	Variables de interés	Escala de agregación	Fuente	Url descarga	Nombre archivo
Sectores catastrales (barrios)	Código de sector (ID) Nombre del Barrio	Sector catastral	UAECD	https://datosabiertos.bogota.gov.co/dataset/sector-catastral	SECTOR.geojson
Valor de referencia comercial m2 de terreno	valor del suelo (precio)	Manzana	UAECD	https://datosabiertos.bogota.gov.co/dataset/valor-de-referencia-por-metro-cuadrado-de-terreno	valor_ref_2023.geojson
Estratos	Estrato Socioeconómico	Manzana	UAECD	https://datosabiertos.bogota.gov.co/dataset/estratificacion-para-bogota	manzanaestratificacion.zip ManzanaEstratificacion.shp
Establecimiento de Gastronomía y Bar	Densidad de establecimientos de gastronomía y bar	UPZ	Instituto Distrital de Turismo	https://datosabiertos.bogota.gov.co/dataset/establecimiento-de-gastronomia-y-bar-bogota-d-c	establecimientos gastronomia y bar.geojson
Establecimiento de Alojamiento Turístico	Densidad de establecimientos de alojamiento turístico	UPZ	Instituto Distrital de Turismo	https://datosabiertos.bogota.gov.co/dataset/establecimiento-de-alojamiento-turistico-bogota-d-c#	establecimientos alojamiento turistico.geojson
Colegios	Ubicación de colegios	coordenadas	Secretaria Distrital de Educación	https://datosabiertos.bogota.gov.co/dataset/colegios-bogota-d-c	colegios.geojson
incidentes delictivos	Número de incidentes delictivos reportados en las categorías: - Riñas -Narcóticos -Orden Público -Maltrato	Sector catastral	Secretaría Distrital de Seguridad, Convivencia y Justicia	https://datosabiertos.bogota.gov.co/dataset/incidente-reportado-bogota-d-c	IRSCAT.geojeson
delitos de alto impacto	Número de delitos de alto impacto reportados en las categorías: - Homicidio -Hurto -Violencia sexual -Violencia intrafamiliar	Sector catastral	Secretaría Distrital de Seguridad, Convivencia y Justicia	https://datosabiertos.bogota.gov.co/dataset/delito-de-alto-impacto-bogota-d-c	DAISCAT.geojeson

Nombre	Variables de interés	Escala de agregación	Fuente	Url descarga	Nombre archivo
Zonas de interés turístico	Identificación de inmuebles ubicados en zonas de interés turístico por tipología	polígonos	Instituto Distrital de Turismo	https://datosabiertos.bogota.gov.co/dataset/zonas-interes-turistico-bogota-d-c#	zitu.geojson
parques	Ubicación de parques	polígonos	Secretaria distrital de planeación	https://datosabiertos.bogota.gov.co/dataset/parque-bogota-d-c	parques.zip Parque.shp
espacio público efectivo	indicador de espacio público efectivo (m2 de espacio público por habitante)	UPZ	DADEP	https://datosabiertos.bogota.gov.co/dataset/espacio-publico-efectivo-upz-2021	EPE_UPZ.shp
estaciones TM	ubicación estaciones TM	coordenadas	Transmilenio S.A	https://datosabiertos-transmilenio.hub.arcgis.com/datasets/estaciones-troncales-de-transmilenio/explore	Estaciones_TM.geojson
Densidad poblacional	densidad poblacional en la manzana Número de habitantes en la manzana Número de viviendas en la Manzana	Manzana	DANE	https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/visor-descarga-geovisores/#gsc.tab=0	MGN_ANM_MANZANA.geojson
localidad	localidad (ID) Nombre de la localidad	localidad	SDP	https://datosabiertos.bogota.gov.co/dataset/localidad-bogota-d-c	localidad.zip localidad.shp
centros comerciales	ubicación centros comerciales	coordenadas	SDP	https://datosabiertos.bogota.gov.co/dataset/gran-centro-comercial-bogota-d-c	Grandes_centros_comerciales.zip Grandes_centros_comerciales.shp