Integración de los datos de ofertas de vivienda en Cali.

Cesar Augusto Correa Lozano

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2022

Integración de los datos de ofertas de vivienda en Cali

Cesar Augusto Correa Lozano

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Director:
Mg. David Arango Londoño
Codirector:
Ph D. Jaime Mosquera Restrepo

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2022

Dedicatoria

A mis padres, quienes se esforzaron cada día para que yo tuviera los medios y herramientas para estudiar y quienes además, a lo largo de mi carrera, me acompañaron, ayudaron y guiaron para dar lo mejor de mi.

A mi hermano, a quien con mi ejemplo quiero servir de modelo a seguir para su formación como persona y profesional.

A mis profesores, aquí va para todos los profesores que he tenido desde el inicio de mi educación, quienes con su dedicación, paciencia y ejemplo contribuyeron en mi formación como persona y profesional.

Agradecimientos

Agradezco primeramente a Dios, al Señor Jesucristo, al Espíritu Santo, a la Santa Cruz, al Señor de los Milagros, a la Virgen María y a todos los Santos y Santas, quienes a lo largo de mi carrera me ayudaron, me dieron los medios para hacerlas, me iluminaron y acompañaron.

Agradezco a mis padres y hermano, quienes con su paciencia, dedicación y amor estuvieron para mi en el día a día de mi proceso como estudiante, ayudándo en todo lo que más podían y brindándome su apoyo y compañía.

Agradezco también a todos los profesores que con su granito de arena contribuyeron para mi formación como persona y profesional, con especial mención a los profesores José Rafael Tovar, Andrés Felipe Ochoa Muñoz y Jaime Mosquera Restrepo, quienes con ayuda me guiaron para encontrar y definir la metodología utilizada para el desarrollo de este trabajo.

Resumen

En Colombia, el mercado inmobiliario ha sido una actividad económica con grandes proyecciones y con gran participación en el desarrollo socioeconómico del país en los últimos años (OIKOS 2020). A pesar de esto, la información publicada sobre este mercado es limitada. A la vez, es frecuente encontrar deficiencias en la disponibilidad y el análisis de los precios de alquiler o venta de las viviendas, lo que dificulta la predicción de los retornos de las inversiones en el sector, así como los posibles efectos en otros sectores de la economía (Cubeddu et al. 2012, Cárdenas Rubio et al. 2019). En la búsqueda de una solución a este problema, el presente trabajo plantea una metodología replicable para la conformación (en el instante que sea requerido) de un registro amplio de la oferta del mercado inmobiliario de la ciudad de Cali (Valle del Cauca), que permitirá a quien lo consulte tener una referencia y visión amplia sobre dicho tema. La construcción de este registro comienza desde la consulta a varias paginas web, dedicadas a publicar ofertas de vivienda en Cali, a traves de un algoritmo de web scraping encargado de descargar en una hoja de datos la información publicada sobre la vivienda, para posteriormente, aplicar un proceso de limpieza de datos y eliminación de duplicados que consolide el registro. Como resultado, se entrega un aplicativo Shiny con dos tableros que resumen las características mas importantes de la oferta de viviendas en Cali y un breve análisis del precio.

Palabras clave: Web Scraping, Limpieza de Datos, Detección de duplicados, Comparación de Imágenes.

Abstract

In Colombia, the housing market has been an economic activity with good projections and with great participation in the socioeconomic development of country in recent years OIKOS (2020). Despite this, the information published on this market is limited, at the same time, deficiencies are often found in the availability and analysis of rental or sale prices of housing, making it difficult to predict returns on investments in the sector, as well as their possible effects on other sectors of the economy Cubeddu et al. (2012), Cárdenas Rubio et al. (2019). In the search for a solution to this problem, this work proposes a replicable methodology for the formation (at the time required) of a comprehensive record of the housing market supply of the Cali city (Valle del Cauca), which will allow those who consult it to have a reference and broad vision on said subject. The construction of this registry begins with the consultation of several web pages, dedicated to publishing housing offers in Cali, through a web scraping algorithm in charge of downloading published housing information into a datasheet; and then apply a process of data cleaning and duplicate detection (record linkage) that consolidates the registry. As an additional, a Shiny application is delivered with two dashboards that that summarize the most important characteristics of the housing offer in Cali and a brief analysis of its price.

Keywords: Web Scraping, Data Cleaning, duplicate detection, Image Comparison.

Contenido

Lista de Figuras

Lista de Tablas

Declaración

Me permito afirmar que he realizado el presente Trabajo de Grado de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el propio trabajo. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, se han reconocido según autor. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de Tesis o Trabajo de Grado.

Igualmente declaro que los datos utilizados en este trabajo están protegidos por las correspondientes cláusulas de confidencialidad.

Santiago de Cali, 25 de Octubre de 2022.

Cesar Augusto Correa Lozano

1. Introducción

De acuerdo con OIKOS (2020), en los últimos años el mercado inmobiliario colombiano ha presentado buenas proyecciones económicas y una gran participación en el desarrollo social del país; ya que, se han implementado diversos programas de financiamiento de vivienda, que han surgido por parte del gobierno nacional y de entidades de diversa índole. Como resultado, para el 2019, el sector inmobiliario superaba la quinta parte de la economía colombiana y generaba 1.8 millones de puestos de trabajo aproximadamente (Mutis 2019).

A pesar de ser un mercado tan prometedor, cuenta con una problemática, debido a que la información publicada es limitada, poco disponible, dificultando poder realizar análisis sobre el estado actual y real del impacto en la economía colombiana (Cubeddu et al. 2012, Cárdenas Rubio et al. 2019). En este sentido, el presente trabajo propone una metodología estadística para consolidar una hoja de datos de la oferta de vivienda para la ciudad de Cali (Valle del Cauca), disponible para ser analizada y obtener una vista general o un diagnóstico de la oferta inmobiliaria en la ciudad.

De acuerdo a lo mencionado, la metodología se compone de tres partes: la primera dedicada a la creación de algoritmos de web scraping para la extracción de los datos partiendo desde las páginas web donde se anuncia la oferta de vivienda para Cali, los cuales fueron creados y ejecutados a través del aplicativo de Google Web Scraper, en la versión 0.6.4, dando como resultado una hoja de datos por página consultada en formato de valores separados por comas CSV y permite obtener el link de las imágenes publicadas en el anuncio.

La segunda parte, consiste en la aplicación de varios procesos de limpieza de datos para unificar y eliminar inconsistencias de las hojas de datos obtenidas, mediante técnicas como la imputación de datos multivariados, la implementación de restricciones de integridad, detección de datos atípicos, entre otras. Todo esto, mediante las diferentes herramientas del software Microsoft Excel para Microsoft 365 MSO (versión 2202) y del software estadístico R versión 4.0.2, en el entorno de RStudio versión 1.3.1073.

La tercera y última parte, se enfoca a la detección y eliminación de registros duplicados a partir de las imágenes y variables asociadas a cada anuncio disponible en las páginas web consultadas de oferta de vivienda.

Como producto adicional, se realizó la creción de un aplicativo Shiny para mostrar dos tableros basados en la hoja de datos resultante de la última parte. El primer tablero visualiza la oferta de vivienda en Cali y el segundo se enfoca en un breve análisis de precios de las viviendas ofertadas. Estos tableros son creados con el objetivo de ilustrar algunos de los análisis que pueden surgir como resultado del procesamiento de la hoja de datos consultada.

1.1. Definición del Problema.

En los últimos años, el mercado inmobiliario ha sido una actividad económica con agradable proyecciones en el país, esto debido en parte a los diversos programas de financiamiento de vivienda que han surgido por parte del gobierno nacional y de entidades de diversa índole (OIKOS 2020). Dado que según Mutis (2019), el sector inmobiliario integral superaba la quinta parte de la economía colombiana y generaba 1.8 millones de puestos de trabajo aproximadamente. A la vez, el sector de la construcción (con el cual esta muy correlacionado) generaba 1.4 millones de empleos directos. En este orden de ideas, la presidenta de la Cámara Colombiana de la Construcción afirma que estos sectores tienen una gran participación en el desarrollo social y económico del país (OIKOS 2020).

A pesar de la importancia del sector inmobiliario para Colombia, de acuerdo a Cubeddu et al. (2012) y a Cárdenas Rubio et al. (2019) la información disponible sobre el mercado inmobiliario es limitada. Es frecuente encontrar deficiencias en la disponibilidad y el análisis de los precios de alquiler o venta de las viviendas, lo que dificulta la predicción de los retornos de las inversiones en el sector, así como los posibles efectos en otros sectores de la economía.

Por otra parte, según Chiquiza (2020), actualmente para el sector inmobiliario se ha habilitado el avalúo en línea, que consiste en páginas web que proporcionan un estimado del precio de una casa o apartamento a partir de características propias como la ubicación, el metraje o los espacios con los que cuenta. Chiquiza (2020) también indica que el gerente de la constructora OIKOS afirma que "*desde el punto de vista numérico, el avalúo por internet es muchísimo más exacto que el normal, pues el número de inmuebles que se verifican al momento de hacer una consulta es de cerca de 400, de características similares*".

De lo anterior, se deduce una creciente necesidad de las personas o entidades que desean comercializar un bien inmobiliario, de tener un referente informativo que le oriente dentro de este sector, de tal forma que puedan conocer el estado actual del mercado. Aunque en la actualidad existen páginas web que servirían como referente informativo, estas se encuentran

limitadas a las ofertas publicadas en ellas mismas, por lo que se tendría una visión parcial de este mercado. En este sentido, se propone en este trabajo reunir la información de varias páginas web que en la actualidad recogen las ofertas de vivienda en Cali, de tal forma que se obtenga un referente informativo unificado que proporcione una visión amplia y global sobre el mercado inmobiliario caleño.

1.2. Justificación

Al ser la vivienda el bien duradero más importante para los hogares, hace que el comportamiento del mercado tenga un importante impacto sobre la economía y política nacional (Holmes & Panagiotidis 2011). Para el caso de Colombia, de acuerdo con Mutis (2019) el sector inmobiliario integral supera la quinta parte de la economía colombiana para el 2019, con más de un 1.8 millones empleos generados. Aun así, la información disponible sobre el mercado inmobiliario es limitada y débil, lo cual dificultaba el análisis.

Por lo anterior, se hace importante tener una metodología que entregue, en el instante que sea requerida, una hoja de datos con la información actual y confiable del mercado inmobiliario que permita hacer los estudios deseados. Ejemplos de estos serían, el avalúo de un inmueble, estudio para compra o venta de una vivienda, estudio para la elaboración de planes de construcción de viviendas, estudio para la planeación del ordenamiento territorial, entrenamiento de modelos Machine Learn para la predicción de precios, estudio de la fluctuación de los precios de vivienda, entre otros.

Cabe mencionar que, si bien la metodología desarrollada en este trabajo está enfocada en el mercado inmobiliario de Cali, se le pueden hacer pequeños ajustes para que esta sea aplicable tanto a otras ciudades como a otros mercados de los cuales se pueda obtener información de páginas de internet. Esto hace que el aporte de este trabajo no solo sea para este caso en particular, sino que también pueda ser utilizado como referente para obtención y estructuración de datos de internet.

2. Objetivos

2.1. Objetivo General

Conformar una hoja de datos de referencia unificada para la consulta y análisis en tiempo real de la oferta inmobiliaria en la ciudad de Cali.

2.2. Objetivos Específicos

- Diseñar e implementar una estrategia de consulta automática para conformar un registro único, en tiempo real, de la oferta inmobiliaria en la ciudad de Cali.
- Identificar y eliminar los registros duplicados en la oferta de viviendas a través de técnicas basadas en la similitud de los campos numéricos, campos de texto e imágenes disponibles.
- Diseñar un tablero de visualización (Dashboard) para el análisis primario de la información contenida en el registro único de la oferta inmobiliaria en la ciudad de Cali.

3. Antecedentes

Se han encontrado diversos trabajos desarrollados con relación a la integración de bases de datos con ofertas de vivienda para el análisis del estado del mercado inmobiliario, esto lo hacen mediante la aplicación de una metodología de recolección de datos (manual o automatizada) y procesos de limpieza. En este orden de ideas, se realizó una revisión bibliográfica en búsqueda de algunos trabajos que sirvieran como base metodológica y temática para el presente trabajo.

Cárdenas Rubio et al. (2019) presentan una base de datos y análisis sobre la información de precios y características de venta o arriendo de casas o apartamentos en las principales ciudades de Colombia (Bogotá, Medellín, Cali, Barranquilla y Bucaramanga), sin embargo en el documento solo se presentan análisis para la ciudad de Bogotá. La metodología que utilizaron se resume a la extracción, compilación y limpieza en tiempo real de anuncios publicados en páginas de Internet especializadas en el mercado inmobiliario colombiano, mediante web scraping, hacen un análisis espacial de los precios de vivienda para mostrar el potencial de la base de datos que construyeron.

Por otra parte, los autores encontraron una limitación, manifestando que aunque se encuentra información de diferentes zonas de las ciudades, la mayoría de viviendas ofrecidas para la venta o arriendo tienden a concentrarse en las zonas de estratos altos. También, se percibe una limitación en la predicción de precios debido a que existe una diferencia entre el precio anunciado para arrendar o vender y el precio pagado cuando se realiza el negocio. En las conclusiones, los autores plantean que las plataformas web facilitan al público consultar el precio y las características de las viviendas ofrecidas para la venta o arriendo en diferentes ciudades del país, reduciendo las deficiencias de información en el análisis del mercado inmobiliario. Además, resaltan la posibilidad de consolidar una base de datos confiable mediante técnicas de recolección automatizada y una depuración adecuada de los datos.

Cuervo & Jaramillo (2014) mencionan una actualización sobre una bsase de datos de precios inmobiliarios de las viviendas en Bogota, la cual se compone de varias series con precios de alquile de vivienda, compraventa de vivienda usada y de tierra destinada a vivienda entre 1970 y 2013, diferenciadas para tres estratos de ingresos y elaboradas a partir de las ofertas publicadas en los periódicos y los precios del suelo publicados por la Lonja de Propiedad Raíz.

La metodologia implementada por los autores consiste en la revisión de las publicaciones en los periódicos y de la lonja para extraer de forma minuciosa los datos sobre las viviendas ofertadas junto con la geoposición. Lo anterior, teniendo en cuenta los principios fundamentales: el primero, consiste en recopilación de la evolución del precio de las viviendas y garantía de que los indicadores que ellos construyen se refieran a los mismos inmuebles a lo largo del tiempo; el segundo hace referencia a las series de precios que son compatibles entre sí y haciendo posible la articulación analítica. Este trabajo, presenta las limitaciones propias de utilizar datos recolectados de forma manual, que se refleja en posibles sesgos de selección, limitación en términos del volumen de datos obtenidos, de la frecuencia de recolección, así como de posibles errores en la transcripción de la información.

Tamilselvi & Saravanan (2009) plantean una metodología para la detección de datos duplicados en una data warehouse (base de grandes volúmenes de datos con duplicados no intencionales) buscando una óptima calidad de datos e incrementar la velocidad en la que se realiza el proceso de limpieza de datos. De manera ilustrativa aplican el proceso sobre una base datos de estudiantes.

La metodología que imparten consta de seis pasos donde, a excepción del paso final, cada uno cuenta con un algoritmo; los cuales son:

- **Selección de variables** tiene por función buscar cuales son las variables más relevantes o criticas para la identificación de un registro duplicado, formando una base datos a parte con estas variables.

- **Formación de llaves** aquí a partir de las variables seleccionadas se crea una llave primaria para cada registro el cual servirá para la creación de bloques.

- **Agrupación o separación en bloques de los registros** a partir de la llave primaria o de las variables seleccionadas se busca dividir el conjunto de registros en grupos más pequeños para facilitar la detección.

- **Computación de la similaridad** en cada grupo o bloque se aplica una métrica de similaridad definida (los autores proponen usar la función de similaridad Jaccard, que sirve para ver la similitud entre cadenas de textos).

- **Detección y eliminación de duplicados** aquí con base en un umbral de distancia se determina si una pareja de registros es o no un duplicado, para en caso tal se deje solo a uno de los dos.

- **Consolidación** se organiza la base de datos original sin los registros que fueron detectados ya como resultado o se puede proceder con un proceso de limpieza de datos.

Briñez et al. (2013) plantearon una técnica basada en métricas de similitud para el análisis de las imágenes obtenidas en estudios de fotoelasticidad de películas plásticas sometidas a deformación por esfuerzos mecánicos. En este trabajo los autores fotografiaron secuencialmente la deformación de películas plásticas termodeformables con tracción mecánica; posteriormente, hicieron una descomposición de las imágenes en los componentes R,G,B (formato de color basado en los colores Rojo, Verde y Azul – Red Green Blue) y describieron los cambios que presentan las películas en las imágenes. comparando de forma ordenada cada imagen generada en el proceso de descomposición con una imagen de referencia que viene siendo la primera de la secuencia.

Para la descripción, los autores compararon de forma ordenada la distancia euclidiana y la distancia de Bhattacharyya del histograma de cada imagen generada en la descomposición con el de la primera imagen de la secuencia (imagen de referencia), ayudándose también con el coeficiente de correlación de Pearson simple. De este trabajo se resalta que los autores en las conclusiones indican que según los resultados que obtuvieron la distancia euclidiana para histogramas sería la mejor métrica para este tipo de aplicaciones.

Allen et al. (2016) plantean aplicar técnicas de detección de duplicados en 18 conjuntos de datos de tesis de estudiantes de psicología que fueron entregadas para la evaluación, con tres objetivos; primero identificar aquellos trabajados donde se fabricaron, falsificaron o plagiaron resultados en algún momento de la investigación o la revisión de esta; segundo, ver si hay evidencia de conjuntos de datos de investigación duplicados; tercero, ver en caso de que haya duplicados que tan probable es que se presenten.

Los autores encontraron conjuntos de datos compuestos de al menos 40 o 50 variables, siendo normalmente del tipo likert o dicotómicas; por lo que optaron por tomar de 6 a 15 variables por conjunto y calcular el número máximo teórico de permutaciones de respuestas para el proceso de detección de duplicados.

A partir de lo anterior, aplicaron un Análisis de Componentes Principales (ACP), generando los scores del primer componente; los cuales fueron graficados y comparados para visualizar si había o no casos con duplicados. Los autores tuvieron en cuenta dos conceptos de duplicados: el completamente duplicado y el parcialmente duplicado. A la vez consideraron si dos o mas casos tienen el mismo score, serán clasificados como Casos Potencialmente Problematicos-PPCs y a cada uno se le calcula un indice de similaridad (número de variables idénticas dividido por numero total de variables del conjunto de datos en términos de porcentaje). Aquellos casos que fueron identificados como PPC, con un índice de similaridad alto, fueron analizados para ver los patrones de respuestas y la naturaleza de los mismos, además fue interrogado el autor para identificar si era consciente o no del problema y las razones que lo llevaron a eso en caso tal.

Boeing & Waddell (2016) exponen principalmente como se llevó acabo la recolección, limpieza, análisis, mapeo y visualización de la información encontrada en 11 millones de avisos de viviendas en arriendo del sitio de internet Craigslist US (United States-Estados Unidos); lo anterior con dos objetivos: primero presentar varias tendencias en este conjunto de datos poco explorados y las implicaciones para el mercado de la vivienda y segundo compartir con los especialistas y profesionales de la vivienda una poderosa metodología emergente de la ciencia de datos para recopilar e investigar datos urbanos.

Tambien, tuvieron en cuenta la descripción del mercado de viviendas en arrendamiento en Estados Unidos de Craigslist e indicar la necesidad de las técnicas de Big Data en dicho mercado, los autores pasan a la metodología donde indican que fue a través de web scraping que extrajeron la información de las viviendas del sitio de internet Craigslist US que corresponde a los anuncios existentes entre Mayo y Julio del 2014. El Web Scraper construido por los autores para dicho fin, fue creado usando Python y *scrapy* (un sistema de Web Scraping). A la base de datos obtenida mediante web scraping (la cual como característica especial contenía información sobre la área metropolitana a la que pertenecía cada vivienda) se le aplicó un proceso de limpieza de datos para eliminar duplicados y datos incoherentes.

Después de dicho proceso, los autores procedieron a hacer el análisis espacial sobre los datos enfocado a evaluar las características del mercado inmobiliario; a la vez, analizaron el nivel de accesibilidad a una vivienda y las proporciones de viviendas que se encontraban a un precio de alquiler por debajo del mercado de arriendo justo del Department of Housing and Urban Development-HUD US para 58 áreas metropolitanas, ayudándose de la base de datos del 2014 que ofrece dicho departamento de este mercado. Lo más importante de los hallazgos es la equivalencia entre los datos obtenidos y los datos del departamento, también, el hecho de considerar el 37 % de los listados en estas regiones están por debajo del Fair Market Rents (FMR-Mercado de alquiler justo) correspondiente, con algunas excepciones donde, según los análisis, los ciudadanos enfrentan grandes desafíos para acceder a una vivienda.

Finanmente, en los anteriores trabajos, se evidencia la existencia de diferentes metodologías implementadas para la obtención, organización y limpieza de datos del mercado inmobiliario, resaltando el hecho de lo útil y práctico que es la implementación de web scraping para la recolección de datos de la internet. También se encuentra un precedente de como llevar a cabo la comparación de imágenes, alrededor de evaluar la similaridad de los registros. Por otro lado, se encuentra que el proceso de detección de duplicados se realiza principalmente solo considerando los datos de los anuncios de oferta de vivienda, por tanto, resulta interesante añadir a este proceso la comparación de las imágenes de tal forma que permita enriquecerlo y hacerlo más preciso en situaciones donde la similitud no es definitiva.

4. Marco Teórico

A continuación se presentan los principales conceptos que fueron utilizados para la implementación de los métodos y procedimientos estadísticos aquí desarrollados; entre ellos se describe el proceso de Limpieza de Datos, Detección de duplicados, Análisis Factorial Múltiple, Imputación de Datos, Análisis de Clusters, las Medidas de Distancia empleadas que son las técnicas estadísticas sobre las cuales se soporta la metodología propuesta.

4.1. Marco Conceptual

Los conceptos y definiciones fundamentales para el desarrollo del trabajo se plantean aquí con el fin de brindar claridad respecto a lo que hacen referencia y en que sentido se van a usar. Para tal fin, se definen los términos: vivienda, Web Scraping, imagen, dato duplicado, dato faltante y mercado inmobiliario.

4.1.1. Vivienda

Pérez & Gardey (2010) definen como vivienda al lugar cerrado y cubierto que se construye para que sea habitado por personas. Este tipo de edificación ofrece refugio a los seres humanos y les protege de las condiciones climáticas adversas, además de proporcionarles intimidad y espacio para guardar las pertenencias y desarrollar las actividades cotidianas.

Casa, departamento, apartamento, residencia, piso, hogar, domicilio y estancia son algunos de los términos que se usan como sinónimo de vivienda. La utilización de cada concepto depende de ciertas características, generalmente vinculadas al tipo de construcción. De esta forma, las viviendas colectivas reciben nombres como apartamento o departamento, mientras que las viviendas individuales se conocen como casa, chalet, etc.

4.1.2. Mercado Inmobiliario

De acuerdo con Núñez Tabales et al.(2008,pp 9-10):

"*En sentido estricto, el mercado inmobiliario incluye el mercado o mercados de la tierra y de otros activos o recursos territoriales que permanecen inmóviles en su ubicación original y ello con mayor propiedad que las construcciones, ya que algunas de éstas se pueden trasladar o desaparecer,*

mientras que la tierra, como espacio territorial, permanece indefinidamente sin ningún tipo de movimiento o variación en el espacio y en el tiempo.(...)Aunque desde el punto de vista práctico, al hablar de mercado inmobiliario normalmente se hace referencia al mercado de inmuebles construidos y, más concretamente, al mercado de la vivienda, tanto nueva como usada entre otras razones, por la mayor magnitud de negocio de la construcción respecto a la agricultura y por la importancia de la construcción residencial frente al total, así como por la vinculación y dependencia del mercado de inmuebles no residenciales al mercado de la vivienda y, finalmente, por la importancia social de la vivienda" .

4.1.3. Web Scraping

Broucke & Baesens (2018) definen al Web Scraping como "*la construcción de un agente para descargar, analizar y organizar datos de la web de manera automatizada*". En otras palabras, lo que se hace es simular mediante un aplicativo la actividad humana de extraer los datos de la web para pegarlos en una hoja de datos (por ejemplo) de una forma más rápida y correcta de lo que puede hacerlo el humano; por lo tanto, el Web Scraping se centra en la transformación de datos no estructurados en la web en datos estructurados que pueden ser almacenados en bases de datos locales centrales o en una hoja de cálculo (Vargiu & Urru 2013).

Las técnicas de "Web Scraping" al principio fueron conocidas como "screen scraping" y se usaban para extraer datos con el fin de generar una representación visual; por lo cual, en los primeros años de la computación a menudo era reducido a un simple código base de terminales; el cual, en esos días, era usado para obtener grandes cantidades de texto de las terminales para el almacenamiento y uso posterior (Broucke & Baesens 2018). Y es que, Broucke & Baesens (2018) mencionan que "*las técnicas para la recolección de datos automática es probablemente tan vieja como el intenet mismo*".

Web Scraping juega un papel importante en varios campos del conocimiento, siendo una forma económica de obtener datos. El mercadeo es uno de los campos que más se beneficia de esta técnica, permitiendo recolectar información en tiempo real de las redes sociales sobre el pensar de la gente y las necesidades de los mismos (Análisis de Sentimientos). Otras aplicaciones del web scraping están en los productos de Google, por ejemplo Google Translate utiliza textos almacenados en la web para entrenarse y mejorarse a sí mismo. (Broucke & Baesens, 2018, pp.7-8)

Existen dos clases de formas de hacer Web Scraping. Por un lado, están las formas que usan aplicaciones (APIs-Application Programming Interface) que mediante plataformas permiten dialogar con los servidores de una página web y así obtener los datos; por otro lado, están las formas que usan algún lenguaje de programación para establecer un algoritmo que interactúe con el código HTML de la página web para así obtener la información.

4.1.4. Dato Duplicado

Uribe & Jiménez (2010), Hadzic & Sarajlic (2020) definen como dato duplicado aquella entidad que se almacena más de una vez a través de una o varias bases de datos, en tuplas (registros o individuos) con igual estructura, pero sin ser necesariamente copias/replicas exactas, pudiendo así ser un duplicado con diferencias leves o significativas en algunas variables de forma individual.

La existencia de duplicados dentro de una base de datos representa un problema, que de no ser tratado, puede llevar a falsas inferencias o a decisiones equivocadas y éstas a la vez pueden ocasionar pérdidas de tiempo, dinero y credibilidad (Uribe & Jiménez 2010). Entre las causas mas comunes para ocasionar duplicados, se encuentran: restricciones de formato, de longitud y/o en el conjunto de caracteres permitidos, errores humanos al capturar los datos, errores que surgen integrando bases de datos diferentes o haciendo migración entre sistemas, modelos de datos mal diseñados, entre otras.

4.1.5. Datos Faltantes

Un dato faltante es aquella medición de una variable que no se encuentra registrada en la matriz de datos, normalmente, señalados con un **NA**. Esta clase de dato, debido a su naturaleza, no solo reduce la precisión de la estimación y disminuye la potencia de las pruebas estadísticas usadas, sino también, bajo ciertas circunstancia, pueden llevar al sesgo de la información y a conclusiones erróneas (Ochoa 2018).

Algunos posibles mecanismos de generación de datos faltantes fueron propuestos por Rubin (2004) y tienen el objetivo de caracterizar las razones por las cuales hay observaciones faltantes. Se define R_{ij} como la variable aleatoria que indica la presencia de datos faltantes, entonces $R_{ij} = 1$ si el dato x_{ij} es faltante, $R_{ij} = 0$ si el dato x_{ij} es observado. De esta manera, cada dato x_{ij} tiene su dato R_{ij} y se hace una partición de x_{ij} , tal que x^0 representa a los datos observados y x^m a los datos faltantes. Estos datos pueden ser (Ochoa 2018, Rubin 2004):

- Faltantes Completamente Aleatorios (Missing Completely at Random, MCAR). Los datos son MCAR cuando la probabilidad de que la respuesta sea faltante no está relacionada con ningún valor de las variables que se planearon observar o de la misma variable de respuesta.

$$P(R_i|x^0, x^m, X_i) = P(R_i)$$

  Por lo tanto, los datos x_{ij} son MCAR cuando R_i es independiente de los componentes observados y no observados.

- Faltantes Aleatorios (Missing at Random, MAR). Se dice que los datos son MAR cuando la probabilidad de que una respuesta sea faltante depende de los valores observados de las variables, pero no depende que los valores no observados.

$$P(R_i|x^0, x^m, X_i) = P(R_{ij}|x^0, X_i)$$

  Esto significa que la distribución condicionada de R_i dado x_i o es independiente de x^m.

- Faltantes no Aleatorios (Missing Not at Random, MNAR). Se dice que los datos son MNAR cuando la probabilidad de que la respuesta sea faltante está relacionada no solamente con los valores observados, sino que también de los valores no observados de la variable respuesta.

$$P(R_i|x^0, x^m, X_i)$$

  Lo que indica que la distribución condicionada de R_i dado x^o, depende de al menos un componente de x^m.

4.1.6. Imagen

Morales & Azuela (2012) definen que "*Una imagen es la representación óptica de uno o más objetos iluminados por una o varias fuentes de radiación*". Por lo que, la generación de una, nace como el reflejo de los rayos de la fuente de radiación sobre una superficie, como se muestra en la Figura **4-1**.

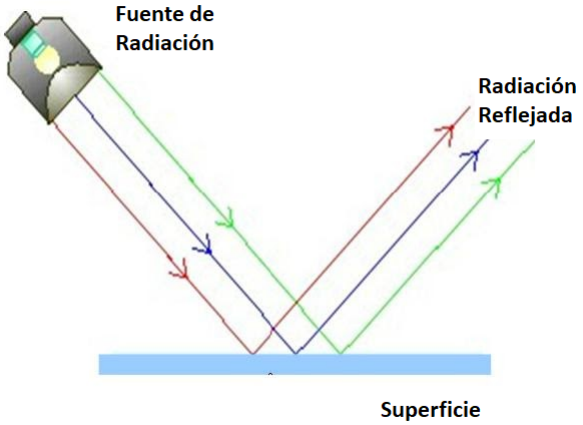


Figura 4-1.: Reflexión de la luz sobre la superficie de un objeto
Fuente: Hernández (2011)

De acuerdo con Morales & Azuela (2012), cuando la radiación reflejada es percibida por el sistema óptico se crea una imagen óptica. Esta imagen pasa a ser una imagen digital cuando es capturada por un captador o digitalizador, que la transforma en una imagen digital a través de un proceso de

cuantificación y muestreo, siendo el muestreo la digitalización de las coordenadas de la imagen y la amplitud de esta la cuantificación. A partir de aquí, se tiene una función que va a transformar cada coordenada de la imagen en un número del 0 a L-1 ($L = 2^B$, donde B es la cantidad de bits que se consideran, comúnmente B=8bits lo que genera un rango de valores enteros de [0,255]), al que se le denomina píxel; entonces una imagen se podrá representar como una matriz de píxeles con M filas y N columnas $g(i, j)$.

$$g(i,j) = \begin{pmatrix} g(1,1) & g(1,2) & \cdots & g(1,N) \\ g(2,1) & g(2,2) & \cdots & g(2,N) \\ \vdots & \vdots & \ddots & \vdots \\ g(M,1) & g(M,2) & \cdots & g(M,N) \end{pmatrix} \tag{4-1}$$

4.2. Marco Estadístico

Con el fin de hacer un apartado las técnicas estadísticas consideradas se presenta en esta sección los conceptos estadisticos claves sobre la limpieza de datos, detección de duplicados, detección de datos atípicos, métricas de distancia, análisis de clusters, Análisis Factorial Múltiple (AFM) e imputación de datos con AFM.

4.2.1. Limpieza de Datos

Müller & Freytag (2005) definen la limpieza de datos como el conjunto de procesos realizados sobre los datos para remover las anormalidades presentes y obtener una colección de datos precisa y única. Dichos procesos pueden ser automáticos o semi-automáticos.

Según Maletic & Marcus (2000) y Müller & Freytag (2005), dichos procesos se comprenden de 4 fases: La primera, consiste en identificar los tipos de anormalidades que están reduciendo la calidad de los datos (Auditoría de Datos); la segunda, consiste en la elección de un método adecuado para automáticamente detectarlos y eliminarlos (Especificación del proceso); la tercera, es la ejecución del método elegido (Ejecución del proceso) y la última, la cual es propuesta por Müller & Freytag, es la evaluación de los resultados obtenidos tras la aplicación del método elegido para ver si es necesario aplicar otro proceso (Post-Proceso). En este sentido, el proceso de limpieza de datos se define como un ciclo cuya regla de parada depende del contexto en el que se estén dando los datos.

Fases del Proceso de limpieza de datos

Si bien anteriormente se mencionaron las fases que tiene un proceso de limpieza de datos, en este apartado se dará una explicación detallada de cada fase según lo encontrado en Müller & Freytag (2005).

- **Auditoría de Datos**: Consiste en hacer una auditoría a los datos analizando a través las estadísticas descriptivas y la sintaxis. Esto con el fin de soportar la especificación de las restricciones de integridad y formatos de dominio de la colección de datos; a la vez se buscan características de los datos que pueden después servir para corregir las anomalías. Por lo tanto, el resultado final de esta fase es una indicación para cada una de las posibles anomalías sobre si ocurre dentro de la recopilación de datos y con qué tipo de característica.

- **Especificación del proceso**: Con base a lo encontrado en la auditoría, se selecciona un proceso de limpieza que permita, una vez aplicado, eliminar todas la anomalías de los datos automáticamente. A la vez en este paso se debe evaluar y probar la exactitud y eficacia del método seleccionado.

- **Ejecución del proceso**: Para esta fase se tiene el método de limpieza correctamente seleccionado, lo que procede es la aplicación sobre la colección de datos, la cual debe garantizar un rendimiento eficiente incluso en grandes volúmenes de datos. Normalmente esto demanda un gran gasto computacional y más si se quiere eliminar la totalidad de las anomalías. Por lo tanto, se hace necesario la definición de la mejor precisión de detección sin dejar de tener una velocidad de ejecución aceptable.

- **Post-proceso y Control**: Después de implementar el proceso de limpieza, los datos son inspeccionados de nuevo para verificar la exactitud de la especificación de las operaciones y aquellos que en un primer momento no pudieron ser corregidos son inspeccionados para una corrección manual. Dependiendo de los resultados se decide si se vuelve aplicar todo el proceso de limpieza de datos.

Métodos de Limpieza de Datos

Müller & Freytag (2005) plantean que los métodos más populares usados dentro de los procesos de limpieza de datos son:

- **Análisis de la Sintaxis.** Se realiza como un proceso de detección de errores de sintaxis de los datos. Para tal fin, lo que se hace es definir una gramática o dominio G y se decide si una cadena de texto dada cumple con ella o no. Estas cadenas son las tuplas completas de una instancia relacional o valores de atributo de un dominio. Cabe mencionar que, los errores de sintaxis normalmente se deben al formato en el que se guardan los datos, por lo que se debe tener cuidado con esto.

- **Transformación de Datos.** Es un método que tiene la intención asignar un formato a los datos de tal forma que se ajusten a un formato esperado para el análisis de los datos. En este método lo que se hace es afectar el esquema de las tuplas, así como los dominios de los valores para que se ajusten a esquema común que se adapte mejor a las necesidades previstas para el análisis. Aquí, la corrección de valores debe realizarse solo en los casos

en que los datos de entrada no se ajustan al esquema y provocan fallas en el proceso de transformación. Por otro lado, se tiene que la estandarización y normalización son los procesos de transformación más usados ya que eliminan las irregularidades de los datos, debida, por ejemplo, a la escala de medición o a las diferentes formas de escribirse.

- **Restricciones de Integridad.** En general, consiste en la eliminación de tuplas que no cumplen con un conjunto de reglas (restricciones de integridad) buscando mantener la calidad de la información, aún después de un proceso de inserción, eliminación o actualización de tuplas. De estas reglas existen dos enfoques diferentes, comprobación y mantenimiento de restricciones de integridad. El primero se basa en rechazar cualquier modificación a la base de datos (transacción) que de aplicarse violaría las restricciones de integridad, el segundo se refiere a modificar la transacción original para que esta se pueda aplicar sin violar las restricciones de integridad.

  La aplicación de este método en los procesos de limpieza de datos es limitada y la idea básica es identificar automáticamente desde el conjunto de restricciones de integridad, una serie de modificaciones y aplicarlas en la recopilación de datos para que luego la recopilación no contenga más violaciones de estas. Este método entonces, se limita a servir de apoyo a la limpieza de datos, dado que el control del proceso debe permanecer con el usuario todo el tiempo.

- **Detección de Duplicados** También conocido como Record Linkage ofrece varios enfoques dentro del mundo de limpieza de datos, pero cada uno requiere un algoritmo para determinar si dos o más tuplas son representaciones duplicadas de la misma entidad y para que sea eficiente se debe comparar cada una de las tuplas con las demás.

- **Análisis Estadístico** El método tiene como objetivo detectar valores faltantes o atípicos, los cuales indicarían posibles tuplas no validas, mediante un análisis descriptivo y/o multivariado; a la vez dentro de este método, se considera la metodología estadística para tratar esa clase de valores.

4.2.2. Detección de Duplicados

Uno de los más interesantes problemas de la Calidad de los Datos es la multiplicidad, los datos duplicados, que consiste en la representación o registro de diferentes formas de un objeto o individuo en un conjunto de datos (Hadzic & Sarajlic 2020). Por tanto, aplicar un proceso mediante el cual, estos registros sean detectados y diferenciados, es importante para trabajar con una base de datos. Según Amón & Jiménez (2010) "*el proceso que detecta este conflicto de duplicados se conoce por diferentes nombres:* ***Record Linkage*** *o* ***Record matching*** *entre la comunidad estadística;* ***database hardening*** *en el mundo de la Inteligencia Artificial;* ***mergepurge,*** ***data deduplication*** *o* ***instance identification*** *en el mundo de las bases de datos; otros nombres como* ***coreference resolution*** *y* ***duplicate record detection*** *también son usados*".

Uribe & Jiménez (2010) explican el proceso, comenzando por definir un umbral real $\theta \in [0, 1]$, luego se le asigna un valor numérico a la similitud entre cada una de las tuplas mediante una función. Entonces, si la similitud entre una pareja de tuplas es mayor o igual a θ, se asumen como duplicados, es decir, se consideran representaciones de una misma entidad real.

En la actualidad hay varios métodos para aplicar este proceso en una base de datos, todos buscando que los duplicados sean identificados correctamente y rápidamente en la totalidad de los casos (eficiencia y efectividad), por lo tanto *reducir la cantidad de comparaciones* a realizar es el principal objetivo, Hadzic & Sarajlic (2020) describen tres de ellos blocking methods, windowing methods y semantic methods.

Blocking methods

Los Blocking Methods consisten en dividir la base datos en subconjuntos de datos a partir de la similaridad que presenta en la blocking key de los registros, para posteriormente en cada uno aplicar el proceso de record linkage o detección de duplicados (Hadzic & Sarajlic 2020; Christen 2007). De acuerdo a Tamilselvi & Saravanan (2009) la blocking key es una variable, la concatenación de variables significativas, un array o un token que se usará como representante de los de datos de un registro, y a partir de las clases que se formen en ella crear los subconjuntos y hacer la detección de duplicados.

Windowing Methods

Hadzic & Sarajlic (2020) explican que estos métodos consisten en ir clasificando los datos de acuerdo a una llave y usando una ventana de datos corrediza, y solo se comparan los datos que estén dentro de la misma ventana; el tamaño de la ventana es ajustado para cada base de datos, buscando uno que no sea muy pequeño ni muy grande, para evitar que se hagan muchas comparaciones y que algunos duplicados queden fuera. El windowing method mas conocido es el Sorted Neighborhood, este consiste en hacer una comparación de registros a través de una llave-regla, que puede estar compuesta de una o más variables, usada para la clasificación de los registros.

Semantic Methods

Buscan aprovechar el significado y contexto del contenido o las relaciones entre los datos para así llevar a cabo una detección de duplicados. Estas técnicas comúnmente combinan similitudes semánticas y sintácticas en un valor de similitud general que se utiliza para detectar entidades similares (Hadzic & Sarajlic 2020).

4.2.3. Detección de datos atípicos y puntos influyentes

Los datos influyentes son tratados desde un contexto de regresión lineal, ya que, es donde mayores problemas genera, dado que puede cambiar los signos de las estimaciones de los parámetros de regresión, afectar la significancia y llevar a malas interpretaciones del modelo. De acuerdo con Mendoza et al. (2002), un dato atípico es una observación extrema cuyo valor no está de acuerdo con el conjunto de datos; esta clase de dato puede ser de **balanceo**, si se determina que tiene una influencia leve para un modelo de regresión, o puede ser de **influencia** si el efecto en un modelo de regresión se determina que es moderado o grave.

Mendoza et al. (2002) plantean que una forma preliminar de detectarlos es calculando los residuos del modelo, donde se espera que si todos los puntos están ubicados muy cerca de la recta, los residuos sean pequeños y si un residuo es grande, sea señal de un punto atípico. De acuerdo a los autores, existen varios métodos para la detección de datos atípicos, entre ellos se encuentran la matriz H, distancia de Cook, DFFITS, DFBETAS y el COVRATIO.

Matriz H

Hoaglin & Welsch (1978) presentan en el articulo como hacer la detección de datos atípicos a través de la matriz Hat $H = X(X^TX)^{-1}X^T$; la cual consiste, en que dado que la matriz Hat es una matriz ponderada, le da un peso a cada dato en consideración a todos los demás, en la diagonal $h_{ii} \in (0, 1)$; donde si un dato esta muy cerca al baricentro tendrá un peso h_{ii} pequeño, mientras que, si esta muy alejado va a ser más grande. Entonces, para un modelo regresión simple, se puede demostrar que los elementos de la diagonal de la matriz Hat se puede calcular como:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum(x_i - \bar{X})^2} \tag{4-2}$$

A la vez, se tiene como regla que si $h_{ii} > 2\bar{h}_{ii}$ se supone que el dato está muy alejado del resto para ser de balanceo.

4.2.4. Métricas de Distancia

De acuerdo con Martinez & Martinez (1999) "*Se da, en general, el nombre de distancia o disimilaridad entre dos individuos i y j a una medida, indicada por d(i,j) , que mide el grado de semejanza o desemejanza, entre ambos objetos o individuos, en relación a un cierto número de características cuantitativas y/o cualitativas. El valor de d(i,j) es siempre un valor no negativo, y cuanto mayor sea este valor mayor será la diferencia entre los individuos i y j*", y debe cumplir como mínimo con estas tres propiedades:

- **P1:** $d(i,j) > 0$ (no negatividad)
- **P2:** $d(i,j) = 0$ cuando $i = j$
- **P3:** $d(i,j) = d(j,i)$ simetría

La lógica, bajo la cual funciona esta medida es la de que entre mayor sea el valor de $d(i,j)$, mayor será la distancia (la diferencia) entre los individuos u objetos (Martinez & Martinez 1999). Si bien, existen varias medidas de distancia, las más conocidas son la distancia Euclidiana, la distancia de Mahalanobis, la distancia de Manhattan, la distancia de Chebyshev y la distancia de Minkowski.

Distancia Euclidiana

Es la medida de distancia más conocida y sencilla de comprender, debido a que su definición coincide con el concepto más común de distancia (Martinez & Martinez 1999). Se entiende entonces, como distancia euclidiana a la longitud del segmento o recta que une 2 puntos en un espacio vectorial.
Para un espacio vectorial R^n, se tiene que la distancia euclidiana entre 2 puntos vectores $A = \langle a_1, a_2, a_3, ..., a_n \rangle$ y $B = \langle b_1, b_2, b_3, ..., b_n \rangle$ de este espacio es:

$$d(A,B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + ... + (a_n - b_n)^2} \tag{4-3}$$

Un dato interesante sobre estas distancias es que en ellas según Martinez & Martinez (1999) "si existe un espacio vectorial R^m , con $m < n$ (siendo n el número de variables consideradas para representar a los individuos ij) y 2 puntos de ese espacio, P_i y P_j de coordenadas : $P_i = \langle P_{i1}, P_{i2}, ...P_{im} \rangle$ y $P_j = \langle P_{j1}, P_{j2}, ..., P_{jm} \rangle$ verificándose que la distancia que estamos considerando entre los individuos i y j es igual a la distancia euclídea entre los puntos P_i y P_j en R^m ; esto es: Si $d(i,j) = (P_i - P_j)$, diremos que la distancia $d(i,j)$ es euclidiana". Además los autores insisten en que se debe cumplir con la desigualdad triangular:

$$d(i,j) < d(i,t) + d(t,j)$$

De acuerdo con Martinez & Martinez (1999) esta distancia presenta 2 inconvenientes, el primero es que es una distancia sensible a las unidades de medida de las variables y el segundo es que si se presenta correlación entre las variables utilizadas, se generará una redundancia de información que sesgaría la medida.

Distancia de Mahalanobis

Resolviendo los inconvenientes de la distancia euclidiana nace esta distancia, en un ambiente más enfocado a la estadística y en busca de medir la distancia entre individuos-registros de una matriz de datos, no sólo considerando los diferentes niveles de medición de las variables, sino los diversos tipos de asociaciones entre las variables (dichas variables deben ser numéricas).

Sean A y B vectores fila que representan las mediciones de 2 individuos que pertenecen a una matriz de datos X con m filas-individuos y n columnas-variables , entonces la distancia de mahalanobis entre A y B seria:

$$A = \left\langle \begin{array}{c} x_1^A \\ x_2^A \\ \vdots \\ x_n^A \end{array} \right\rangle \qquad\qquad B = \left\langle \begin{array}{c} x_1^B \\ x_2^B \\ \vdots \\ x_n^B \end{array} \right\rangle$$

$$d(A,B) = \sqrt{(A-B)^T \Sigma^{-1} (A-B)} \tag{4-4}$$

Donde Σ seria la matriz de covarianzas de X:

$$\sum = \begin{pmatrix} \sigma_{x_1} & \sigma_{x_1 x_2} & \cdots & \sigma_{x_1 x_n} \\ \sigma_{x_2 x_1} & \sigma_{x_2} & \cdots & \sigma_{x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_n x_1} & \sigma_{x_n x_2} & \cdots & \sigma_{x_n} \end{pmatrix}$$

Esta distancia resulta ser mejor que la euclidiana, dado que tiene una mayor sensibilidad, pues tiene relación con la distribución de no solo cada individuo sino también con la de cada variable. De acuerdo a Salas Plata & Portillo (2008) la distancia de Malahanobis resulta ser muy poderosa para saber si un determinado conjunto de condiciones similares es en realidad un conjunto de condiciones ideales, y es muy útil para identificar qué partes de un escenario son las más parecidas a las de un escenario "ideal"; razón por la cual comúnmente se usa para comparar los individuos contra la media.

Distancia entre Matrices

Golub et al. (1996) mencionan que la distancia euclidiana de un punto A al origen se considera en el álgebra lineal como la norma del "vector" $\overrightarrow{OA}$, por lo que, bajo este concepto se tiene que en un espacio matricial M^{mxn}, entonces se tiene que la norma de una matriz A, también conocida como norma de Frobenius, seria:

$$\|A\| \;=\; \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \quad ; a_{ij}\ es\ un\ elemento\ de\ A \tag{4-5}$$

La cual, además de cumplir con las propiedades de una distancia euclidiana debe cumplir también con:

$$\|AB\| \;\leq\; \|A\| * \|B\|$$

A partir de lo anterior, entonces se puede definir como la distancia euclidiana entre dos matrices que pertenecen a M^{mxn} seria:

$$d(A,B) \;=\; \|A-B\| \;=\; \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}-b_{ij}|^2} \tag{4-6}$$

4.2.5. Análisis de Clusters

El Análisis de Cluster, también conocido como Análisis de Conglomerados, es una técnica estadística multivariada que intenta, mediante criterios de distancia reorganizar un conjunto de individuos en grupos descritos por n variables llamados conglomerados, donde los individuos de un conglomerado son relativamente homogéneos entre si pero muy diferentes a los que conforman otros conglomerados (De la Fuente 2011, Lebart et al. 1995).

Al ser el análisis de cluster una técnica de agrupación que recurre a procedimientos algorítmicos, se le considera un análisis exploratorio que carece de bases estadísticas y cálculos formales por lo cual hacer inferencias para una población a partir de una muestra es imposible (Lebart et al. 1995).

En general, los algoritmos de esta metodología agrupan a los individuos en conglomerados de acuerdo con una medida de semejanza o desemejanza; buscando que los individuos de un conglomerado sean similares entre si, mientras que, son diferentes a los de otro conglomerado. En este orden de ideas, los algoritmos de formación de conglomerados se agrupan en dos categorías:

- **Algoritmos No Jerárquicos o de particiones**: Método de dividir el conjunto de observaciones en k conglomerados, definidos inicialmente por el usuario o se determina mediante un criterio de convergencia.

  Algunos ejemplos de estos son K-means, Nubes Dinámicas, Análisis Modal, Método de Wolf, Block-Clustering.

- **Algoritmos Jerárquicos** Método que entrega una jerarquía de divisiones del conjunto de elementos en conglomerados. Estos a la vez se dividen en *algomerativos* que parte de que cada individuo es un conglomerado y en pasos sucesivos los va uniendo, hasta que finalmente todas las observaciones están en un único conglomerado; y en *disociativos* parte de un gran conglomerado y en pasos sucesivos se va dividiendo hasta que cada observación queda en un conglomerado distinto.

  Entre ellos están Método del Centroide, Método de Ward, Salto minimo-Vecino más cercano, Salto maximo-Vecino más lejano.

Estas técnicas presentan ventajas diferentes y pueden utilizarse conjuntamente; lo que da lugar a una estrategia de clasificación basada en un algoritmo mixto, el cual se adaptada al particionamiento de conjuntos de datos formados por millares de individuos a clasificar (Lebart et al. 1995).

En la práctica, Lebart et al. (1995) propone a los usuarios emplear un algoritmo de clasificación sobre los factores o componentes principales resultantes de aplicar algún método factorial[1] sobre los datos, para así aprovechar la ventaja de los conglomerados de ser más fáciles de interpretar que los factores.

Método de Ward

También conocido como el método de agregación según la varianza, se rige bajo el principio de buscar en cada etapa una partición tal que la varianza interna de cada grupo sea mínima y por consecuencia la varianza entre los grupos sea máxima (Lebart et al. 1995).

En este orden de ideas, Lebart et al. (1995) indica que en una etapa inicial de la aplicación de este método la varianza intra-grupos es nula y la varianza inter-grupos es igual a la varianza total de la nube, puesto que cada elemento terminal constituye a su nivel un grupo. En la etapa final, es la varianza inter-grupos la que es nula y la varianza intra-grupos es equivalente a la varianza total, dado que a ese nivel se dispone de una partición en un solo grupo. En consecuencia, a medida que se efectúan los reagrupamientos, la varianza intra-grupos aumenta y la varianza inter-grupos disminuye.

La estrategia a seguir para ejecutar este método se basa según Lebart et al. (1995) en lo que se denomina como el Criterio de Ward Generalizado, el cual sugiere que en vez de buscar individuos próximos para formar una agrupación, se debe buscar individuos x_i y x_i' que al agruparse generen un índice de nivel $\Delta I_{ii'}$ mínimo, que es la cantidad en la que se vería aumentada la varianza intra-grupos y disminuida la varianza inter-grupos.

[1] Estos métodos serian el Análisis de Componentes Principales, Análisis de Correspondencias Múltiples, Análisis Factorial Múltiple, entre otros.

El índice de nivel $\Delta I_{ii'}$, se define como la perdida de varianza inter-grupos, debido al paso de la partición en k conglomerados a la partición en $k-1$ conglomerados, lo que se expresa como:

$$\Delta_k = \Delta I_{ii'} = I_{inter(P_k)} - I_{inter(P_{k-1})} = \frac{m_i m_{i'}}{m_i + m_i'} \parallel x_i - x_{i'} \parallel = \frac{m_i m_{i'}}{m_i + m_i'} d^2(x_i, x_{i'}) \quad (4\text{-}7)$$

Donde, m_i seria el peso del individuo x_i, $I_{inter(P_k)}$ la varianza inter-grupos y $d(x_i, x_{i'})$ es la distancia entre los individuos x_i y $x_{i'}$.

4.2.6. Análisis Factorial Múltiple

Dentro de los métodos de análisis multivariados exploratorios, se encuentra el Análisis Factorial Múltiple (AFM), el cual según Escofier & Pagès (1992) es "*un método factorial adaptado para el tratamiento de tablas en las que un grupo de individuos se describe mediante varios grupos de variables*". Por lo que, como indica Husson & Josse (2013) estudia la similaridad entre los individuos en términos de todas las variables, mientras estudia los vínculos de las variables y relaciona ambos; también estudia los vínculos entre grupos de variables y compara la información aportada por cada uno.

La naturaleza de este análisis plantea que a cada grupo de variables le corresponde una tabla, por lo que como están definidos sobre los mismos individuos se pueden yuxtaponer y formar así una única tabla que cruza individuos y variables (Escofier & Pagès 1992), creando así la denominada matriz yuxtapuesta que se presenta en la Figura **4-2** (tomada de Garcia (2021)).

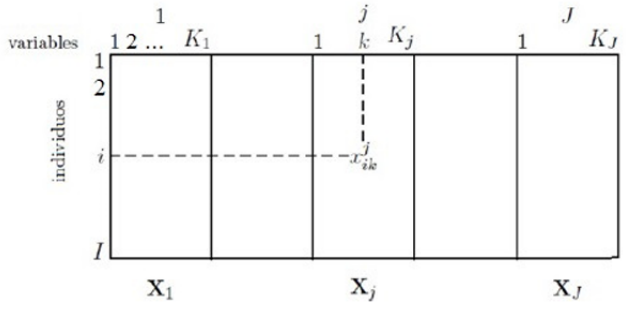


Figura 4-2.: Representación general de la estructura de la tabla de datos en AFM
Fuente: Garcia (2021)

El AFM se presenta en tres espacios, en el espacio de la nube de individuos R^K, en el espacio de las variables R^I y en el espacio de la nube de grupos de variables R^{I^2}. Por lo cual, maneja la siguiente notación para cada una de las partes involucradas:

- X : El arreglo o la matriz yuxtapuesta.
- $X_1, X_2, ..., X_J$: Las matrices que conforman los grupos de variables.
- J : Número de grupos.
- I : Número de individuos.
- K : Número total de variables.
- K_j : Número de variables en el j-esimo grupo.

Husson & Josse (2013) plantean los siguientes pasos para aplicar el AFM sobre **J** grupos de variables de **I** individuos:

1. Se crea **X** la matriz yuxtapuesta donde $\mathbf{X} = [X_1, X_2, ..., X_J]$, siendo X_j una matriz de variables continuas o de variables dummy si el j-esimo grupo es de variables categóricas.
2. Se define la matriz D_Σ que es una matriz diagonal del tamaño de $K\mathrm{X}K$ donde cada elemento es igual a uno para variables continuas o a la raíz cuadrada de la proporción de las categorías representadas como variables dummy.
3. Se define entonces la matriz **Z** como $\mathbf{X}D_\Sigma^{-1/2}$, que de forma centrada seria $\mathbf{Z} = \mathbf{X}D_\Sigma^{-1/2} - M$, donde M es una matriz de tamaño $I\mathrm{X}K$ con $M = 1m'$ siendo m el baricentro de **X**.
4. En este punto, se procede a "equilibrar" la influencia de cada grupo, mediante la multiplicación de cada uno por la raíz cuadrada de la inversa multiplicativa de la primera componente λ_1 asociada al ACP del grupo en el caso de las continuas o la asociada al ACM del grupo en el caso de las categóricas. Técnicamente, lo anterior se puede escribir como $\mathbf{Z}\Lambda^{-1/2} = \left[Z_1/\sqrt{\lambda_1^1}, Z_2/\sqrt{\lambda_1^2}, ..., Z_J/\sqrt{\lambda_1^J}\right]$
5. Finalmente, se aplica un ACP sobre la matriz $\mathbf{Z}\Lambda^{-1/2}$, para así obtener los componentes principales del AFM, que pasan a denominarse factores comunes, junto a los vectores y valores propios correspondientes.

Representaciones Gráficas

Escofier & Pagès (1994) indica que en el Análisis Factorial Múltiple se plantean cuatro representaciones, la de los individuos, la de las variables, la global de los grupos de variables y la superpuesta; estás son descritas por Garcia (2021) de la siguiente manera:

- **Representación de los individuos** : Partiendo de la definición del vector u_s como el eje de inercia de orden s de la nube de individuos N_I asociada a la tabla X en R^K. Este se deduce de la componente principal F, gracias a la relación de transición: $u_s = (1/\sqrt{\lambda_s})X^T E F_s$,

en la que λ es el valor propio de GE asociado a F_s, donde $G = \sum_j G_j = \sum_j X_j M_\lambda X_j^T$, E es la matriz diagonal de los pesos de los individuos y M_λ es una matriz cuya diagonal está compuesta por $1/\lambda_j^1$ con $j = 1, 2, ..., J$ .La proyección de N_I^j sobre u_s se escribe:

$$F_s = \widetilde{X}_j M_\lambda u_s = (1/\sqrt{\lambda_s})\widetilde{X}_j M_\lambda X_j^T E F_s = (1/\sqrt{\lambda_s}) G_j E F_s; \quad s = 1, 2, \ldots, K \tag{4-8}$$

Las J nubes de individuos, denominadas nubes parciales, se proyectan como ilustrativos puntos sobre el primer plano factorial, como se ilustra en la representación ficticia de la gráfica de la Figura **4-3**.

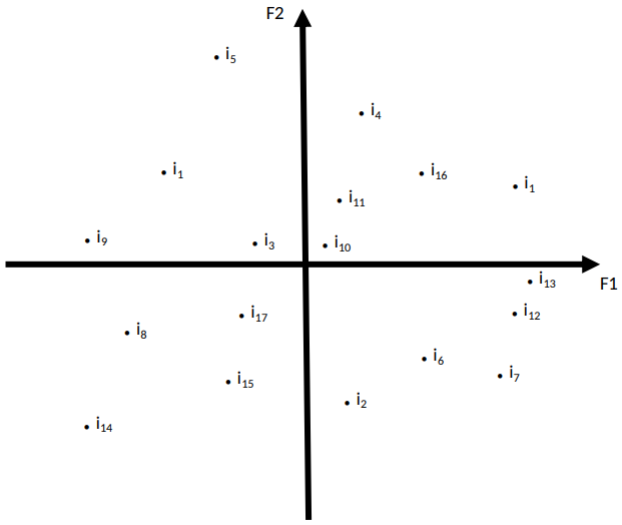


Figura 4-3.: Ejemplo de representación de los individuos en el AFM

Fuente: Garcia (2021)

- **Representación de las variables** : Se obtiene directamente del ACP de la matriz $\mathbf{Z}\Lambda^{-1/2}$, tomando las proyecciones de las variables φ que a través de las relaciones de transición se define así:

$$\varphi_s = \sqrt{\lambda_s} u_s; \quad s = 1, 2, \ldots, K \tag{4-9}$$

Cabe mencionar que esta representación ayuda a la interpretación de la nube de individuos y puede considerarse como una representación óptima de las correlaciones entre variables. Como ejemplo de una representación de este tipo está la Figura **4-4** mostrando las proyecciones de variables en el primer plano factorial.

También, se consideran los factores comunes como una forma de maximizar la inercia de las proyecciones de todas las variables, la inercia proyectada de cada nube N_K^j puede ser interpretada como la contribución de un grupo. La ponderación de los grupos (por $1/\lambda_j^1$) equilibra la influencia, con relación a que la contribución de un grupo en la construcción de un eje está limitada por el valor 1, surge de nuevo, la idea según la cual ningún grupo puede, por sí sólo, determinar el primer eje (salvo una situación de simetría excepcional) y un grupo influye e sobre tantos más ejes cuanto mayor sea la dimensión K_j.

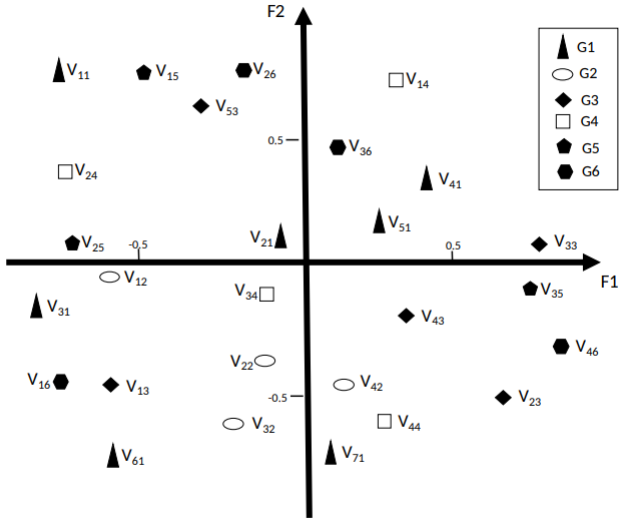


Figura 4-4.: Ejemplo de representación de las variables en el AFM

Fuente: Garcia (2021)

- **Representación de los grupos de variables** : Como el AFM es un análisis que permite estudiar los vínculos que tienen los grupos de variables, que es lo que se busca visualizar con esta representación. Escofier & Pagès (1992) plantean que "*la coordenada de un grupo sobre un eje es la inercia acumulada de las variables del grupo sobre el eje correspondiente del AFM*", esta inercia también se entiende como la correlación lineal entre las proyecciones de la nube global y las de la nube de cada grupo.

  El valor de la coordenada debido a la ponderación que se efectúa sobre las variables va a variar de 0 a 1. Por tanto, se tiene la Figura **4-5** como ejemplo ficticio de esta representación, de la cual (Garcia 2021) indica que los valores altos de las coordenadas indican que el factor está asociado a las variables de los grupos; por el contrario, valores bajos de las coordenadas indican poca relación del factor con las variables de los grupos.

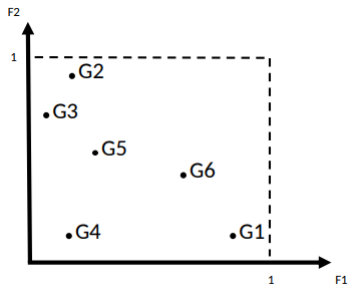


Figura 4-5.: Ejemplo de representación de los grupos de variables en el AFM
Fuente: Garcia (2021)

- **Representación superpuesta de las J nubes definidas por cada grupo de variables** : Esta representación permite ver que tan bien se encuentran representados-caracterizados los individuos en los diferentes grupos de variables.

  Gracias a que las nubes N_I^j están situadas en R^K, es posible una representación simultánea por proyección sobre un mismo subespacio, por lo tanto Garcia (2021) indica que para este fin hay que considerar lo siguiente:

  - La nube $N_I^j \in R^{K_j}$, y R^{K_j} es subespecio de R^K.
  - $R = \oplus_j R^{K_j}$ (suma directa de los subespacios ortogonales dos a dos).
  - M_λ^j es la métrica de R^{K_j}, submatriz con K_j filas y columnas de M_λ.
  - Las coordenadas de los puntos de N_I^j son las filas X_j, entonces las coordenadas de estos puntos en R^K son las filas de la matriz $\widetilde{X}_j$ definida como:

  $$\widetilde{X}_j = [0 \ \ldots 0 \ \widetilde{X}_j \ 0 \ \ldots \ 0]$$

  - $X = \sum_j \widetilde{X}_j$, entonces, con el fin de hacer coincidir $F_s(i)$ con el punto medio de los J puntos parciales $F_s^j(s)$, las filas de $\widetilde{X}_j$ se proyectan como puntos suplementarios amplificados por J.
  - La unión de las J nubes parciales forma la nube N_I^J.
  - La inercia de la nube N_I^J se puede expresar como $InerciaIntra + InerciaEntre$ de las subnubes de $N : i^J$. La inercia de la nube media N_I del AFM corresponde a la $InerciaEntre$ cuando las coordenadas parciales se amplifican por J (Escofier & Pagès 1992).

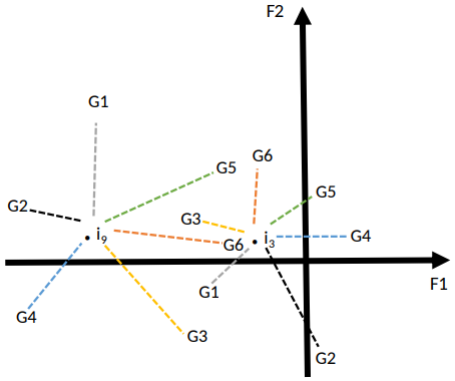


Figura 4-6.: Ejemplo de representación superpuesta en el AFM
Fuente: Garcia (2021)

4.2.7. Imputación de Datos con AFM

La presencia de datos faltantes (missing data) se puede dar en cualquier tipo de dato o situación por motivos indiscrimiados, corriendo mayor riesgo de que estén entre más grande sea el conjunto de datos o el número de recursos de donde se toma la información. Esta clase de datos representa un peligro para los análisis como se explicó anteriormente, por lo que a lo largo de la historia se han creado procesos (metodologías) para de alguna forma lidiar con ellos o estimarlos, a los cuales se les denomina "Imputación de Datos". Husson & Josse (2013) plantean tres métodos de imputación multivariada basados en análisis factoriales en el caso de presentar valores faltantes con patrón aleatorio, dispersos o estructurados. A continuación se describirá solo el algoritmo que utiliza el AFM para inputar datos faltantes, el cual según los autores dado que el núcleo del AFM es una Análisis de Componentes Principales (ACP) ponderado, consiste en aplicar el algoritmo iterativo regularizado para el ACP considerando la estructura del grupo y la ponderación específica.

1. Se inicia con $l = 0 : X^0$. Los valores faltantes son reemplazados por un valor inicial que puede ser la media para el caso de las variables continuas o la proporción de la categoría para el caso de las variables dummy, estos calculados sin considerar los datos faltantes y considerando que las entradas pueden no ser enteras para las variables dummy, pero la suma de las entradas correspondientes a un individuo y una variable debe ser igual a uno. Las matrices D_Σ^0 y $\hat{\mathbf{M}}^0$ son calculadas junto con el valor propio de cada grupo de variables para componer $\Lambda^0 = diag(\lambda_1^1, ..., \lambda_1^J)^0$.

2. Se define la tabla global $Z^{l-1}(\Lambda^{-1/2})^{(l-1)} = \left(\mathbf{X}^{(l-1)} \left(D_\Sigma^{-1/2}\right)^{(l-1)} - \hat{\mathbf{M}}^{(l-1)}\right) \left(\Lambda^{-1/2}\right)^{(l-1)}$

3. Se aplica un ACP sobre la tabla global para estimar las dimensiones ($\hat{\mathbf{F}}^l$,$\hat{\mathbf{U}}^l$); manteniendo S dimensiones. La matriz **F** es la matriz de componentes principales para los individuos mientras que la matriz **U** es la matriz de vectores propios.

4. Los valores ajustados son estimados como $\mathbf{Z}^l = \hat{\mathbf{F}}^l(\hat{\mathbf{U}}^l)^T\left(\Lambda^{-1/2}\right)^{(l-1)}$ y entonces los valores faltantes son imputados con $\hat{\mathbf{X}}^l = \left(\hat{\mathbf{M}}^{(l-1)} + \hat{\mathbf{F}}^l(\hat{\mathbf{U}}^l)^T\left(\Lambda^{-1/2}\right)^{(l-1)}\right)\left(D_\Sigma^{1/2}\right)^{(l-1)}$. Donde el nuevo conjunto de datos imputados es $\mathbf{X}^l = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}$.

5. Ya con $\mathbf{X}^l$ se recalculan las matrices $D_\Sigma^l, \hat{\mathbf{M}}^l$ y Λ^l.

6. Se repiten los pasos del 2 al 5 hasta convergencia.

Los autores plantean que el algoritmo anterior, presenta problemas de sobre ajuste, los cuales se pueden evitar teniendo en cuenta un procedimiento de regularización que consiste en umbralizar los valores propios obtenidos al realizar el ACP en el paso 3 del algoritmo. Donde cada valor propio $\sqrt{\lambda_S}$ es remplazado por $\left(\sqrt{\lambda_S} - \frac{\sigma^2}{\sqrt{\lambda_S}}\right)$ con σ^2 igual al promedio de los últimos valores propios (los autores invitan a leer "Missing values in exploratory multivariate data analysis methods" de Josse y Husson 2012 para más detalles sobre la solución del sobre ajuste).

También, los autores mencionan que el valor de S es seleccionado a priori, sugiriendo que una forma de elegir este valor para el caso de datos incompletos es aplicar validación cruzada en el ACP, bajo la lógica de que se remueve una celda de la matriz de datos y se hace la predicción con las fórmulas de reconstrucción sobre los datos sin esta celda. El procedimiento se repite para cada celda de los datos y para diferentes números de dimensiones. El número que minimiza el error cuadrático medio de la predicción se mantiene.

5. Metodología

Para dar cumplimiento a los objetivos planteados en este trabajo, se describen los pasos realizados para la generación de resultados, los cuales se encuentran ilustrados en el diagrama de la Figura **5-1** y son mencionados a continuación:

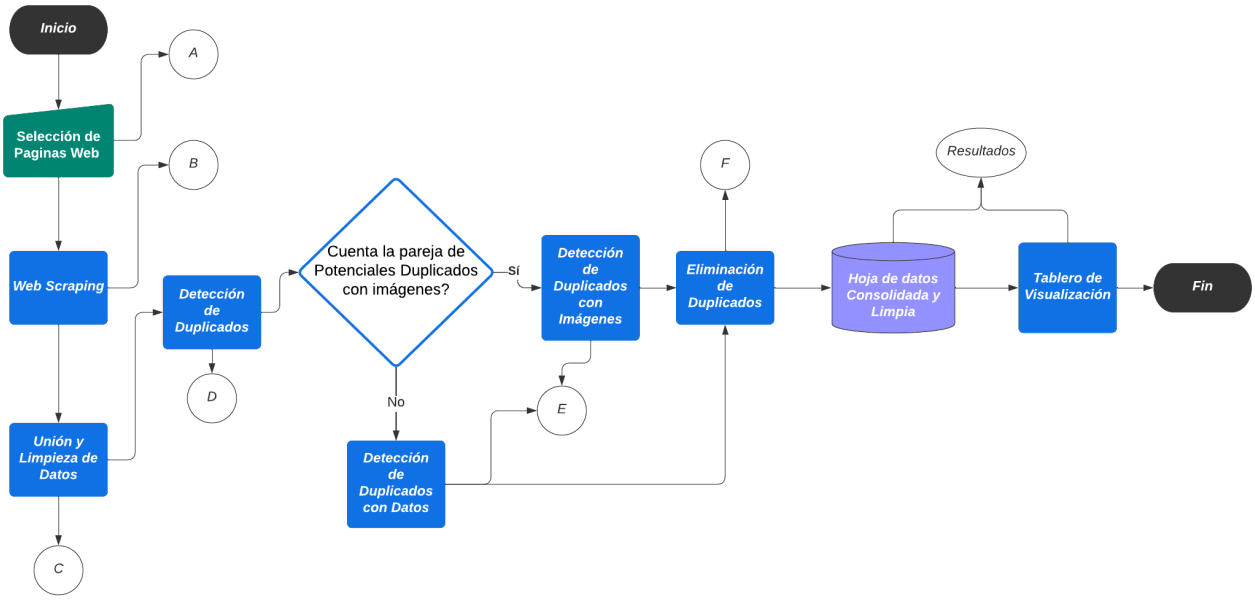


Figura 5-1.: Diagrama de flujo de actividades
Fuente: Elaboración propia

A. **Selección de Páginas web**: Se realizó una evaluación sobre algunas páginas web de donde se puede extraer la información sobre la oferta de vivienda en Cali, para de acuerdo a las característica presentadas, seleccionar las más adecuadas para hacer la captación de datos.

B. **Web Scraping**: Para cada página web selecta se creó un algoritmo para la obtención de los datos e imágenes de los anuncios de venta de vivienda de forma estructurada en hojas de datos mediante la técnica de Web Scraping.

C. **Unión y Limpieza de Datos**: A las hojas de datos obtenidas se les aplicó un proceso de organización y reestructuración, para a partir de las variables que tienen en común unirlas y formar una sola hoja de datos. Con el fin de eliminar las inconsistencias de la nueva hoja de datos se aplicaron sobre esta distintos procesos de limpieza de datos.

D. **Detección de duplicados**: Se realizan transformaciones a los datos de la hoja de datos unificada para facilitar la comparación de registros considerando todas las variables a partir de una métrica adecuada. Tras la comparación de los registros con la métrica, se hace la detección de "los Potenciales duplicados" considerando un umbral.

E. **Detección de Duplicados con Datos-Detección de Duplicados con Imágenes** : Dado que no todos los registros contaban con imágenes para hacer una comparación directa y definir una regla para determinar si son o no duplicados, se partió la detección de duplicados en 2 grupos: los que si tenia imágenes disponibles para la detección y los que no. Para cada uno se aplicó un proceso diferente para determinar si un par de registros eran o no duplicados (Comparación de Imágenes o Similitud Aproximada).

F. **Eliminación de Duplicados**: Con los registros duplicados ya detectados, se procede a realizar su correcta eliminación de la hoja de datos principal, para así consolidar una hoja de datos sobre la venta de viviendas en la ciudad de Cali limpia y estructurada.

A continuación, se describe a detalle cada de una de las etapas comentadas anteriormente.

5.1. Selección de Páginas Web

En este trabajo se evaluó la estructura de las páginas web de oferta de vivienda disponibles en internet: OLX, Fincaraiz, Mercado Libre y Metrocuadrado. Para la evaluación y selección de las páginas se tomó en cuenta la disposición de los datos por anuncio, la facilidad para extracción de las imágenes y la facilidad para crear un algoritmo de Web Scraping.

- **OLX**: Se encuentra una consistencia entre los anuncios en la cantidad de datos (variables) disponibles en esta página de consulta. A la vez, cuenta con una facilidad para la extracción de la información, dado que los campos de variables siempre se encuentran en el mismo lugar dentro del anuncio (campos fijos). También, se encontró la presencia de un espacio apartado para las imágenes de la vivienda con botones para ir cambiando entre ellas, lo cual facilita la captación. Por último, lo más atractivo de esta página web es que cuenta con sistema de cambio listado de anuncios[1] muy sencillo, que consiste en oprimir un botón para que se cargue de forma continua el siguiente listado, lo que facilita la programación del algoritmo de Web Scraping.

- **Fincaraiz**: Esta página de consulta cuenta con una amplia gama de campos de datos que cambian de anuncio a anuncio (campo variable), lo que dificulta la extracción y estructuración de los datos. Por otro lado, al igual que OLX tiene un apartado para las

[1]Entiéndase por sistema de cambio de listado de anuncios aquella parte al final de una página web donde se encuentra uno o varios botones que permiten cargar más anuncios visibles en la página cambiando de página(listado)

imágenes, lo que significa, una extracción más sencilla. Si bien tiene campos variables, estos siempre están en el mismo apartado dentro de un anuncio y también tiene campos fijos, por lo que la extracción de datos se hace más sencilla. A la vez, esta página web, debido a la dirección URL (link), no es necesario programarle un sistema de cambio de listado lo que facilita y simplifica el algoritmo de Web Scraping.

- **Mercado Libre**: Esta página es similar a Fincaraiz en cuanto a la disposición de campos de datos. La diferencia entre ellas radica en que comparten pocos campos en común entre ellas y con OLX, lo que la hace poca atractiva para este trabajo, puesto que representa problemas de compatibilidad para la unión de las hojas de datos. En cuanto a la extracción de imágenes, es igual a las anteriores páginas mencionadas y para la cuestión de cambio de listado de anuncios requiere de una programación propia dentro del algoritmo de Web Scraping, pues se debe hacer un apartado para que el algoritmo entienda como debe ir cambiando de listado y cuando hacerlo.

- **Metrocuadrado**: Presenta una página web con la peculiaridad de que cuando se ingresa a un anuncio se abre una nueva ventana o pestaña con el propio link, mecanismo sobre el cual es difícil implementar el algoritmo de Web Scraping. Por otro lado, esta página web presenta en los anuncios tanto campos variables como fijos, el mismo sistema de cambio de imagen de las anteriores páginas y el mismo sistema de cambio de listado que Mercado Libre.

Considerando las anteriores características de cada página, se descarta Metrocuadrado por la dificultad para programar un mecanismo que ingresara a los anuncios para extraer la información. Aunque, se pensaba trabajar con Mercado Libre, esta página también se descartó debido a la falta de campos similares con OLX y Fincaraiz. Cuando se analizó esta página se encontró que apenas comparte 4 campos en común, a la vez que se encontró que la mayoría de anuncios contaba con poca información sobre la vivienda, lo que representaba una hoja de datos con muchos datos faltantes.

OLX, de acuerdo con el estudio, fue considerada la página más óptima y la primera en ser selecta, dado que, los aspectos de consulta y extracción se facilitan mucho por la estructura y consistencia en la información que tiene. Fincaraiz, fue seleccionada por la compatibilidad de información disponible con OLX y la facilidad que presenta para el cambio de listado frente a las otras dos opciones.

5.2. Web Scraping

Se implementa el Web Scraping como técnica para la obtención automática de datos mediante el software Web Scraper de Google, el cual se puede descargar desde la página oficial `https://webscraper.io/` y queda instalado como una extensión del navegador Chrome

que aparece en el apartado de “Herramientas para Desarrolladores” como una pestaña.

Para cada página seleccionada se crea y diseña un algoritmo de descarga, considerando la estructura y cantidad de variables que se puede extraer de cada anuncio. En el algoritmo se indica la dirección URL del resultado de la búsqueda de “venta de vivienda en Cali” en la página web (OLX o Ficaraiz) donde se encuentran los anuncios de interés, las variables que se van a captar de cada anuncio, el proceso para la captación de las imágenes y por último se específica el mecanismo de cambios de listado de anuncios. Por otro lado, como Web Scraper entrega un link para cada imagen, se crea un código de descarga para estas imágenes en lenguaje Python. El código ingresa a cada uno de links, descarga la imagen y la guarda bajo un nombre determinado en una ubicación especifica que varía entre las páginas de origen.

La construcción de los dos algoritmos de descarga para Web Scraper inicia con la creación en el aplicativo de un caso (un Sitemap), donde se van a definir cada uno de los pasos a seguir para la extracción de datos. Para este fin, se consideran los siguientes pasos:

1. Entrar en la página de anuncios respectiva en el navegador Google Chrome, donde se busca “casas y apartamentos en venta para Cali”.
2. Presionar F12 para abrir las “Herramientas de Desarrolladores” e ir a la pestaña de "Web Scraper".
3. Dar clic a la opción “Create new sitemap” y después a “Create Sitemap”, como se ilustra en la Figura **5-2**.

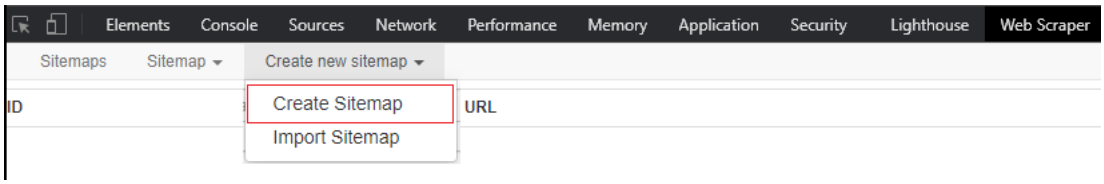


Figura 5-2.: Plataforma de Web Scraper
Fuente: Elaboración propia

4. Se abre la interfaz, tal y como se muestra en la Figura **5-3**. Aquí se asigna al algoritmo de descarga un nombre en la casilla “Sitemap name” y en la casilla “Star URL” se pone el link la página web. Se finaliza el proceso con un clic en “Create Sitemap”.

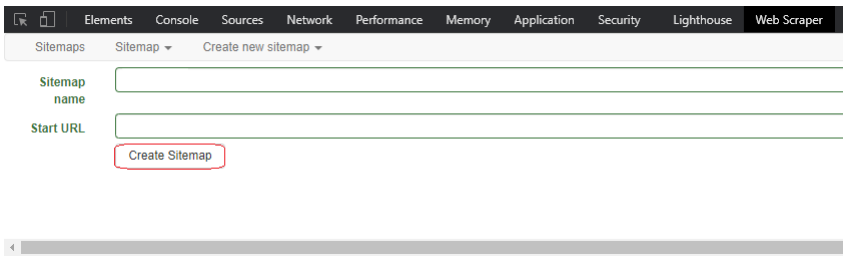


Figura 5-3.: Creación de un Sitemap (Algoritmo de Descarga)
Fuente: Elaboración propia

5.2.1. Algoritmo de Web scraping OLX

Para la página de OLX se consideró el siguiente algoritmo, en cuál se tiene en cuenta el sistema de cambio de listado de anuncios, el sistema de cambio de imágenes y el tipo de campo de dato.

1. Se ingresa al link `https://www.olx.com.co/cali_g4069078/apartamentos-casas-venta_c367` que corresponde a la página de OLX donde se anuncian las casas y apartamentos en venta para la ciudad de Cali.

2. Dentro de la página se presiona F12 para abrir las “Herramientas de Desarrolladores” y se da clic a la pestaña de "Web Scraper". En ella se procede a crear el Sitemap para OLX considerando el link del paso 1 y con el nombre de “olx”.

3. Se pasa a la interfaz, como se muestra en la Figura **5-4**, donde se procede a crear el primer y principal “selector”[2] dando clic en “Add new selector”.

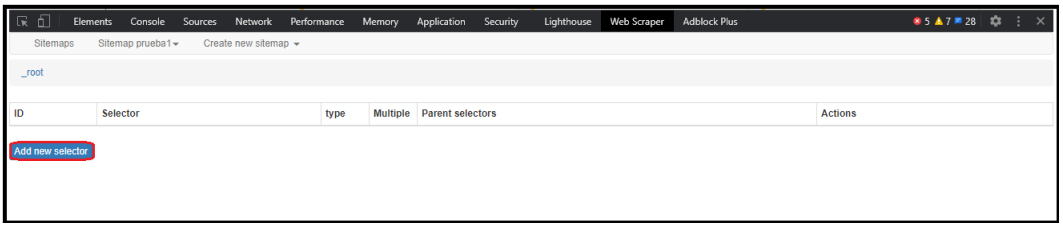


Figura 5-4.: Interfaz del Sitemap
Fuente: Elaboración propia

4. Se despliega la interfaz que se muestra en la Figura **5-5**, llenando las casillas de la siguiente manera:

   - En la primera casilla, “ID”, ira el nombre del selector. Como es el primero se llamará “anuncio”.

[2]Selector es la acción que va hacer el software según las especificaciones programadas en este.

- En la casilla “Type” se elegirá “Element click”. En este caso, dada la naturaleza de la página de OLX, automáticamente se añadirán 2 nuevas casillas “Click selector” y “Click type”.
- En la casilla “Selector”, se da clic en “Select” y se señala cada uno de los anuncios disponibles hasta el botón “Cargar más”, luego se da clic en “Done selecting”.
- En la casilla de “Click selector” se da clic en “Select” y se señala todo el botón de “Cargar más”. Seguidamente se da clic en “Done selecting”.
- Para la casilla “Click type” escogemos la opción “Click more (...)”.
- Las casilla “Click element uniqueness” y “Discard initial elements” se dejan intactas.
- Se da clic en la casilla “Multiple”, esto para indicarle al aplicativo que entre en cada uno de los anuncios disponibles en el sitio web.
- En la casilla Delay se configura el tiempo de chequeo por anuncio que se mide en milisegundos(ms), en este caso se consideran 1000 ms = 1 seg.

Terminado el proceso de llenado de las casillas se da clic en “Save selector” y se retorna a la interfaz del Sitemap, en la Figura **5-5** se aprecia como debe de quedar diligenciado.

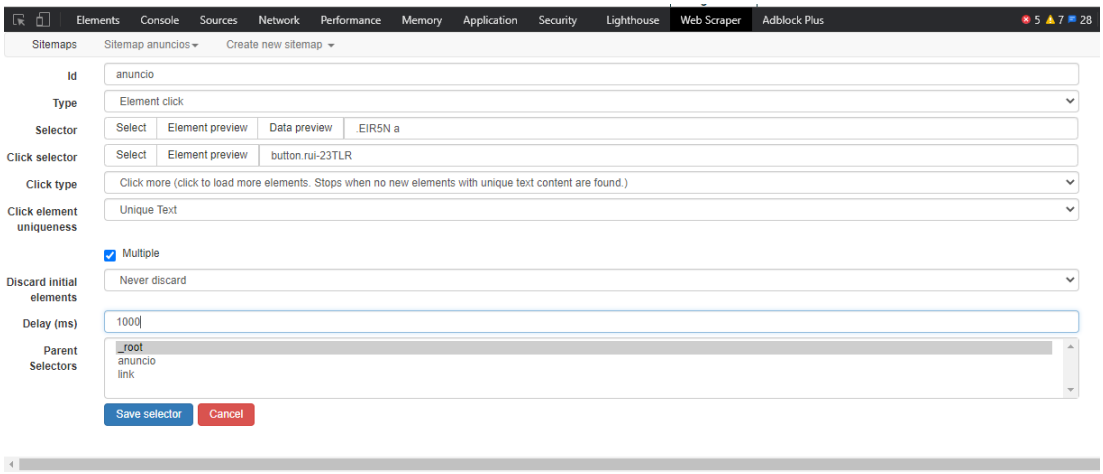


Figura 5-5.: Selector “anuncio”

Fuente: Elaboración propia

5. Se ingresa en el selector que se acaba de crear, dando clic sobre él. De nuevo aparecerá una interfaz como la mostrada en la Figura **5-4**. Seguidamente se crea otro selector dando clic en “Add new selector”, al cual se le pondrá por “ID” “link”, su “Type” sera “Link”, para “Selector” se seleccionará el primer anuncio del sitio web y por último se da clic en “Save selector”. En la Figura **5-6** se aprecia como debe quedar diligenciado. Este selector se crea con el objetivo de obtener el link de cada anuncio para que así el aplicativo visite cada anuncio.

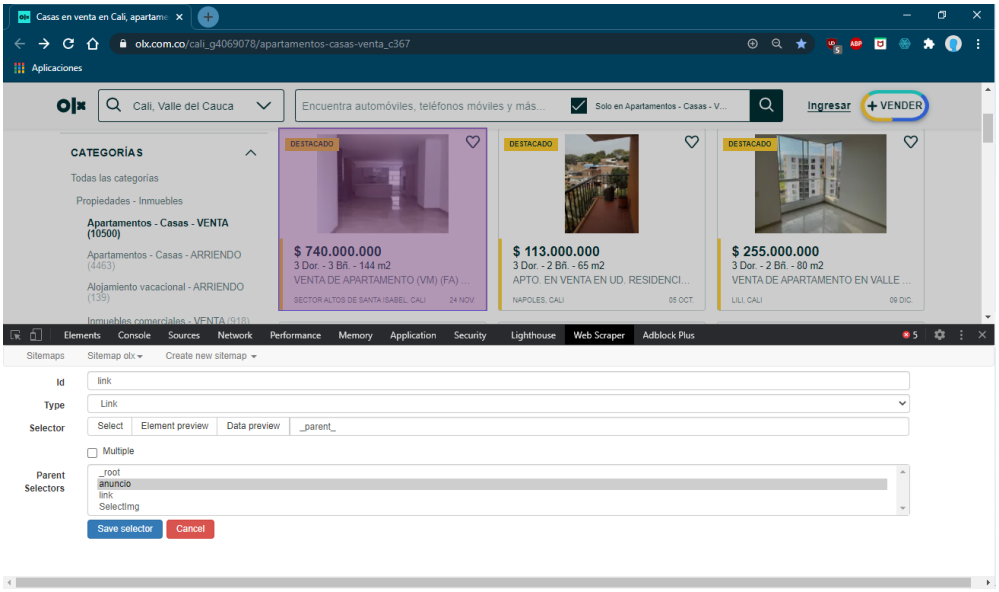


Figura 5-6.: Selector “link”

Fuente: Elaboración propia

6. Se entra en el selector “link” y en el primer anuncio, luego dentro del selector, se deberá crear cada uno de los selectores, que tomaran los datos de interes de los respectivos campos dentro del anuncio. Para efectos prácticos, solo se explicara el proceso para tomar el dato del campo de precio del anuncio de la vivienda. Como se ilustra en Figura **5-7** después de dar clic en “Add new selector” se le pone en “ID” el nombre del dato a extraer en este caso “precio”, en “Type” se deja tipo “Text”, en “Selector” señalamos el campo donde se encuentra y por último se da clic en “Save selector” para guardarlo. El proceso anterior se repetiría para cada uno de los campos disponibles en el anuncio.

7. Para la extracción de las imágenes de cada anuncio, se procede a crear un “Selector” dentro de “link”; la “ID” sería SelectImg, el “Type” será “Element Click”, en “Selector” se señalará el recuadro donde están las imágenes dentro del anuncio como se muestra en la Figura **5-8**; luego en “Click selector” se selecciona el botón de cambio de imagen del lado derecho que se encuentra dentro del recuadro anteriormente mencionado (en OLX este botón tiene la forma de una flecha), en “Click type” se elije la opción “Click more (...)” y por ultimo se guarda el selector dando clic en “Save selector”.

   Se entra al selector recién creado y se crea uno nuevo que es el que toma la imagen. Dicho selector tendrá por “ID” imagen, su “Type” será “Imagen”, para su “Selector” se seleccionaría el recuadro negro que se encuentra dentro del recuadro de imagenes del anuncio como se aprecia en la Figura **5-9**, y se da clic en el recuadro de “Multiple”, para finalmente guardarlo. Cabe mencionar que Web Scraper va a guardar como imagen un link, que posteriormente se usara para la descarga como tal de imagen en Python.

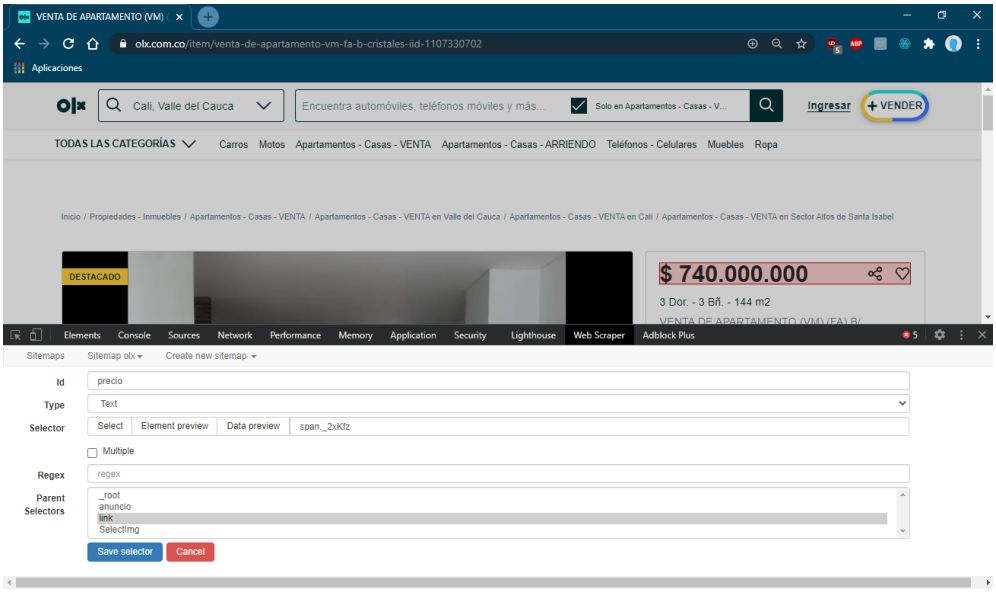


Figura 5-7.: Selección de variables

Fuente: Elaboración propia

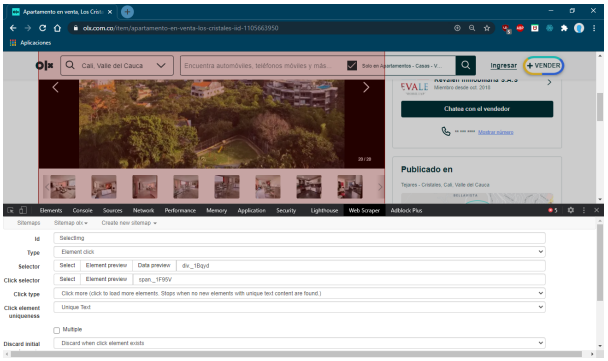


Figura 5-8.: Seleccionador de Imagen

Fuente: Elaboración propia

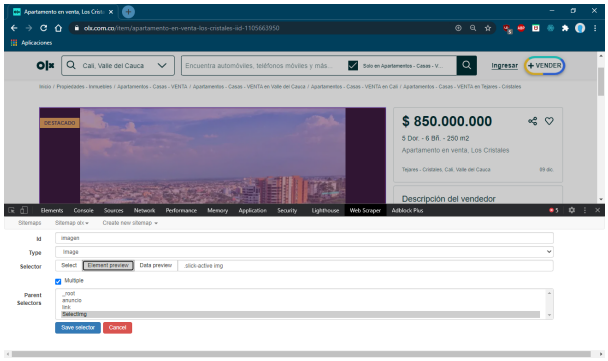


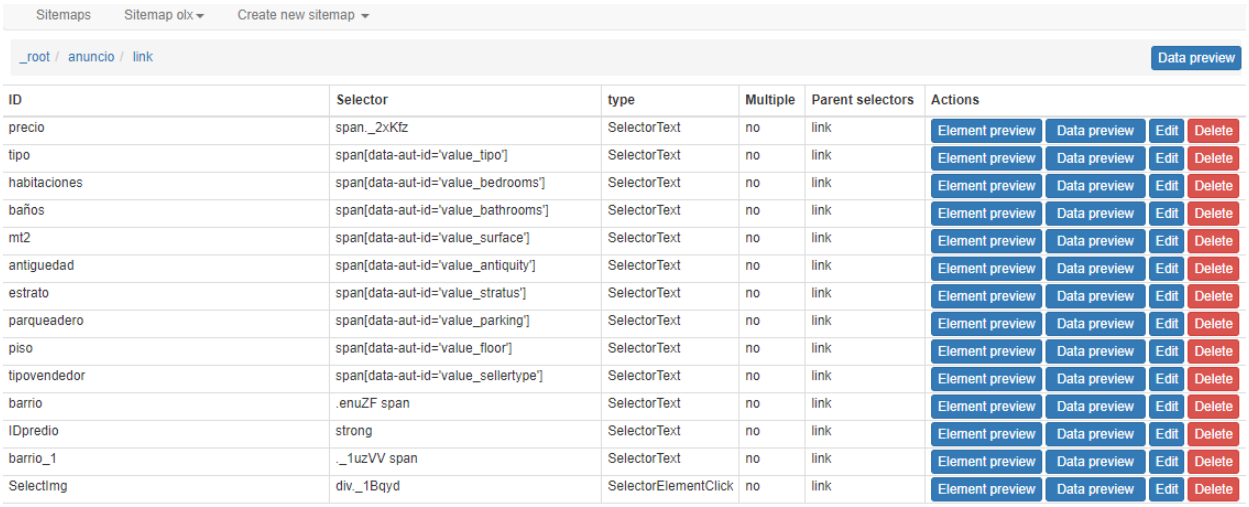
Figura 5-9.: Captador de la Imagen

Fuente: Elaboración propia

8. Ya con todos los selectores necesarios creados se presiona clic en “Sitemap olx”, desplegando así una lista de acciones a realizar, entre las cuales se selecciona “Scraper” para que inicie el proceso de captación de información del algoritmo y una vez que termina se elije “Export data” para que descargar un archivo delimitado por comas (csv) que contiene una de hoja de datos con la información captada de cada anuncio.

Este algoritmo fue ejecutado el 1 de diciembre del 2020 y con él se logró obtener información de 798 anuncios de ofertas de viviendas en Cali considerando en general 18 variables. Por un lado, 14 variables corresponden a los campos de datos que se muestran en la Figura **5-10**, de aquí 12 de ellas corresponden a las variables de interés que se describen en la Tabla **5-1**, una al dato auxiliar barrio_1 y la otra corresponde al apartado para la captación de imágenes. Las otras 4 variables las

entrega Web Scraper por defecto y hacen referencia al link del anuncio, titulo del anuncio, link principal (aquel que se usó en el paso 1 del algoritmo) y código de descarga[3].



ID	Selector	type	Multiple	Parent selectors	Actions
precio	span._2xKfz	SelectorText	no	link	Element preview Data preview Edit Delete
tipo	span[data-aut-id='value_tipo']	SelectorText	no	link	Element preview Data preview Edit Delete
habitaciones	span[data-aut-id='value_bedrooms']	SelectorText	no	link	Element preview Data preview Edit Delete
baños	span[data-aut-id='value_bathrooms']	SelectorText	no	link	Element preview Data preview Edit Delete
mt2	span[data-aut-id='value_surface']	SelectorText	no	link	Element preview Data preview Edit Delete
antiguedad	span[data-aut-id='value_antiquity']	SelectorText	no	link	Element preview Data preview Edit Delete
estrato	span[data-aut-id='value_stratus']	SelectorText	no	link	Element preview Data preview Edit Delete
parqueadero	span[data-aut-id='value_parking']	SelectorText	no	link	Element preview Data preview Edit Delete
piso	span[data-aut-id='value_floor']	SelectorText	no	link	Element preview Data preview Edit Delete
tipovendedor	span[data-aut-id='value_sellertype']	SelectorText	no	link	Element preview Data preview Edit Delete
barrio	.enuZF span	SelectorText	no	link	Element preview Data preview Edit Delete
IDpredio	strong	SelectorText	no	link	Element preview Data preview Edit Delete
barrio_1	._1uzVV span	SelectorText	no	link	Element preview Data preview Edit Delete
SelectImg	div._1Bqyd	SelectorElementClick	no	link	Element preview Data preview Edit Delete

Figura 5-10.: Campos considerados para el web scraping en OLX

Fuente: Elaboración propia

Variables	Tipo-Escala	Descripción	Valores
Precio	Cuantitativa-Razón	Precio al que se oferta la vivienda	[0,∞)
Tipo	Cualitativa-Nominal	Variable binaria indicadora de si la vivienda es casa o apartamento	Casa, Apartamento
Habitaciones	Cuantitativa-Razón	Cantidad de habitaciones que posee la vivienda	[0,∞)
Baños	Cuantitativa-Razón	Cantidad de baños que posee la vivienda	[0,∞)
Metros Cuadrados	Cuantitativa-Razón	Area en metros cuadrados de la vivienda	[0,∞)
Antiguedad	Cualitativa-Orden	Categoria de la antigüedad de la vivienda	A Estrenar En Construcción Hasta 5 años Entre 5 y 10 años Entre 10 y 20 años Entre 20 y 50 años, Más de 50 años
Estrato	Cualitativa-Nominal	Estrato socioeconómico de la zona donde se ubica la vivienda	1,2,3,4,5,6
Parqueadero	Cualitativa-Nominal	Variable binaria indicadora de si la vivienda tiene parqueadero	Si, No
Piso	Cuantitativa-Razón	Piso en el que se encuentra la vivienda o numero de pisos de la vivienda	[0,∞)
Tipo Vendedor	Cualitativa-Nominal	Tipo de persona que ofertaba la vivienda	Dueño Directo, Inmobiliaria
Barrio	Cualitativa-Nominal	Barrio de la ciudad de Cali en el que está ubicada la vivienda	Barrios de la ciudad de Cali
IDpredio	Cualitativa-Nominal	Código de identificación del anuncio en OLX	[0,∞)

Tabla 5-1.: Tabla de variables de interés de OLX

[3]Al descargar un anuncio Web Scraper le asigna un código único

5.2.2. Algoritmo de Web scraping Fincaraiz

En el caso de Fincaraiz, que cuenta con campos variables, se considera una estrategia propia para extraer la información de forma estructurada. Para las imágenes se usa el mismo proceso que en OLX y para lo referente al sistema de cambio de listado de anuncios se hace un proceso que se describe en el paso 7 desde el link principal.

1. Se ingresa al link `https://www.fincaraiz.com.co/apartamento-casa/venta/cali/?ad=30|[1]||||1||8,9|||82|8200006|||||||||||||||1|||1||griddate%20desc||||||` que corresponde a la página de Fincaraiz donde se anuncian las casas y apartamentos en venta para la ciudad de Cali y crear el Sitemap para Fincaraiz, bajo el nombre de "fincaraiz" con el link del paso 1.

2. Se crea el selector anuncio, que como se aprecia en la Figura **5-11** es "Type" "Element", donde en "Selector" se seleccionan todos los anuncios disponibles en la primera página y se señala la opción de "Múltiple", para después guardarlo dando clic en "Save selector".

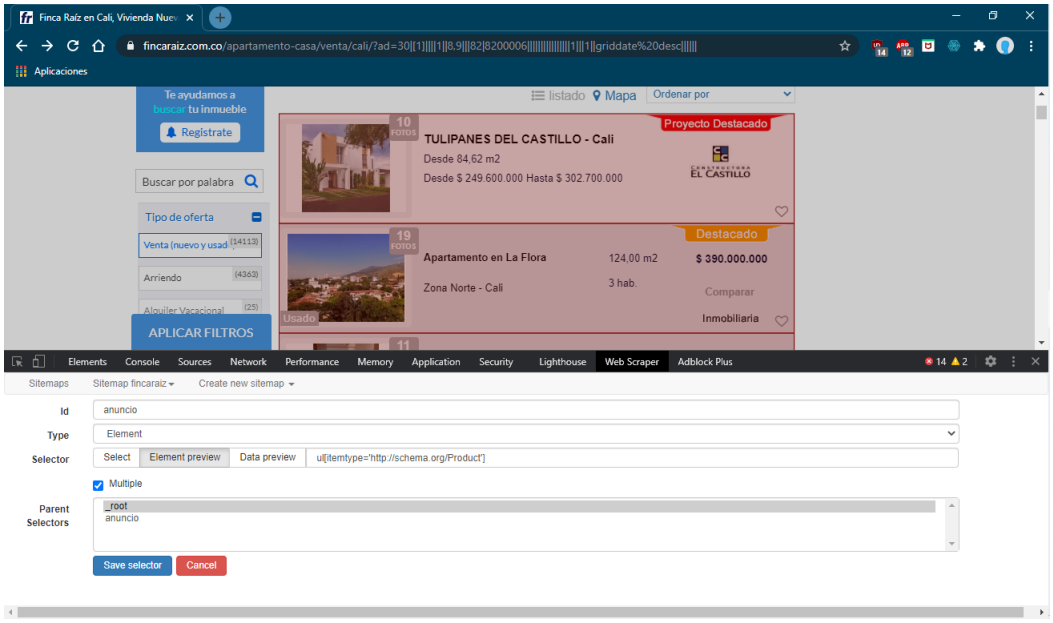


Figura 5-11.: Selector anuncio de Fincaraiz

Fuente: Elaboración propia

3. Dentro del selector "anuncio" se crea el selector "Link", que es de "Type" "Link" y en selector solo se toma el nombre del primer anuncio en la página como se muestra en la Figura **5-12**.

4. Con "Link" creado, dada la naturaleza del sitio web, los campos fijos que se podían obtener de cada anuncio son precio, baños, habitaciones, área y número parqueaderos como se muestra en la Figura **5-13**, donde se creo un selector para cada una, siguiendo el mismo proceso que se explicó para los selectores de los datos de las variables en el caso de OLX.

5. Para el caso de los campos variables, estos son representados en un recuadro que se compone de una cantidad cambiante de campos de datos, como se muestra la Figura **5-14**. Entre los campos se podría encontrar el área privada, el área construida, el estrato, ubicación/sector/barrio , antigüedad, precio por metro cuadrado, el numero del piso donde se encontraba, entre otros. Tras revisar varios anuncios se encontró que la cantidad de campos que aparecía en el recuadro por anuncio era a lo sumo 10; por lo que para la creación de los selectores de estos campos se toma como base un anuncio con los 10 campos y se sigue el proceso de creación de selectores de datos de OLX, denominándolos "c_i" ($i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$).

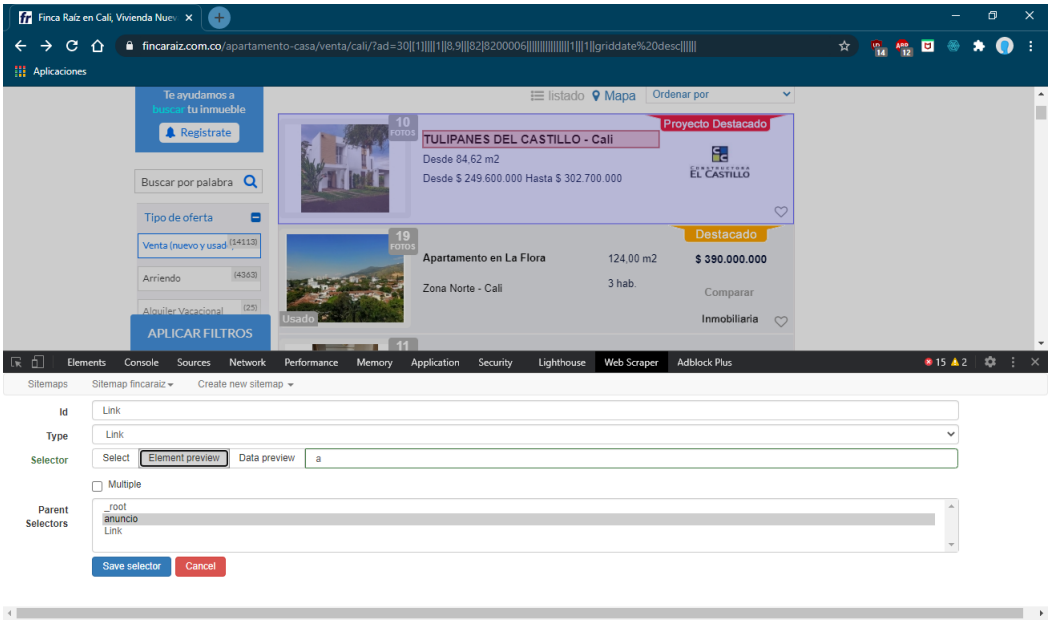


Figura 5-12.: Selector Link de Fincaraiz

Fuente: Elaboración propia

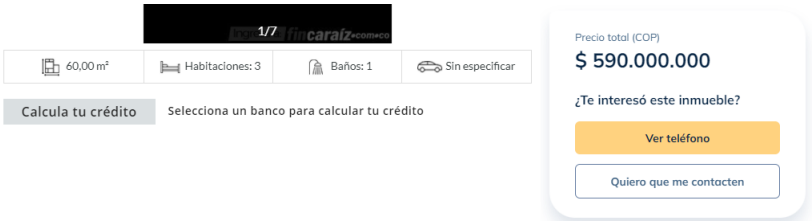


Figura 5-13.: Campos fijos disponibles en Fincaraiz

Fuente: Elaboración propia

6. Para las imágenes y la descarga de la hoja de datos se realizó el mismo proceso que en el algoritmo de OLX.

Figura 5-14.: Campos variables disponibles en Fincaraiz

Fuente: Elaboración propia

7. Para lo referente al sistema de cambio de listado, se percibió en, el caso de Fincaraiz, que al ir cambiando de listado (página) manualmente en el link cambiaba de forma secuencial solo un número en especifico; por lo que, para este caso lo que se hace, es que al momento de crear el algoritmo en Web Scraper en la casilla "Strat URL" se paga el link modificado con el rango de listados disponibles, que en este caso es de 1-492 y el link quedaria asi `https://www.fincaraiz.com.co/apartamento-casa/venta/cali/?ad=30|[1-492]||||1||8,9|||82|8200006|||||||||||||||1|||1||griddate%20desc||||||`

En el caso de Fincaraiz, el algoritmo fue ejecutado el 16 de diciembre del 2020 y gracias a él, se obtuvo información de 8438 anuncios sobre la oferta de vivienda en Cali, considerando 21 campos por anuncio, como se muestra en la Figura **5-15**, entre los cuales se encuentran el link del anuncio, titulo del anuncio, código de descarga y el link principal. También están los 18 campos de datos, de las cuales 6 son variables propiamente que se describieren en la Tabla**5-2**, 1 representa el link de la imagen y los otros 10 hacen referencia a los campos variables.

Variables	Tipo-Escala	Descripción	Valores
Habitaciones	Cuantitativa-Razón	Cantidad de habitaciones que posee la vivienda	[0,∞)
Baños	Cuantitativa-Razón	Cantidad de baños que posee la vivienda	[0,∞)
Parqueaderos	Cuantitativa-Razón	Cantidad de parqueaderos de la vivienda	[0,∞)
Precio	Cuantitativa-Razón	Precio al que se oferta la vivienda	[0,∞)
Vendedor	Cualitativa-Nominal	Tipo de persona que ofertaba la vivienda	Nombre persona natural Nombre persona juridica
Área	Cuantitativa-Razón	Área en metros cuadros de la vivienda	[0,∞)
c_1,c_2,c_3,c_4,c_5, c_6,c_7,c_8,c_9,c_10		Datos no definidos sobre la vivienda	

Tabla 5-2.: Tabla de variables disponibles de Fincaraiz

Sitemaps Sitemap olx Create new sitemap

_root / anuncio / link Data preview

ID	Selector	type	Multiple	Parent selectors	Actions
habitaciones	span.advertRooms	SelectorText	no	Link	Element preview Data preview Edit Delete
baños	span.advertBaths	SelectorText	no	Link	Element preview Data preview Edit Delete
parqueaderos	span.advertGarages	SelectorText	no	Link	Element preview Data preview Edit Delete
selectimg	img.DetailImage	SelectorElementClick	yes	Link	Element preview Data preview Edit Delete
precio	.price h2	SelectorText	no	Link	Element preview Data preview Edit Delete
titulo	h1 span	SelectorText	no	Link	Element preview Data preview Edit Delete
vendedor	p:nth-of-type(3)	SelectorText	no	Link	Element preview Data preview Edit Delete
area	span.advertSurface	SelectorText	no	Link	Element preview Data preview Edit Delete
c_1	ul.boxcube li:nth-of-type(1)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_2	ul.boxcube li:nth-of-type(2)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_3	ul.boxcube li:nth-of-type(3)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_4	ul.boxcube li:nth-of-type(4)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_5	ul.boxcube li:nth-of-type(5)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_6	ul.boxcube li:nth-of-type(6)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_7	ul.boxcube li:nth-of-type(7)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_8	ul.boxcube li:nth-of-type(8)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_9	ul.boxcube li:nth-of-type(9)	SelectorText	no	Link	Element preview Data preview Edit Delete
c_10	ul.boxcube li:nth-of-type(10)	SelectorText	no	Link	Element preview Data preview Edit Delete

Figura 5-15.: Campos considerados para Fincaraiz

Fuente: Elaboración propia

5.2.3. Descarga de Imágenes

Las imágenes en las hojas de datos están representadas por un link; por lo que, se necesitó de un código de descarga para lograr obtener las imágenes de cada anuncio. Dicho código fue generado en lenguaje Python y ejecutado en el IDLE Shell de Python 3.9.1 . La ejecución de este se realizó durante todo el mes de febrero del 2021. Cabe mencionar, que debido a la masiva cantidad de imágenes que se tiene para cada pagina web, y para que estas no detectaran un ataque de denegación de servicio (DoS), cada hoja de datos con los links de la imágenes se dividió de tal forma que se crean hojas con 15000 o 17000 filas (imágenes), que serán las que se utilicen finalmente para la descarga.

El código de descarga utilizado para captar las imágenes considera las librerías *pandas*, *numpy* [4], *requests* [5], *os* [6] , *time* [7]. En esencia, el código llama a la hoja de datos con los links de la imágenes mediante la función pd.read_csv. Luego crea un objeto tipo list con la columna donde se encuentran los códigos generados por Web Scraper, para luego mediante un contador

[4]Para el manejo de las hojas de datos en python
[5]Para acceder a los links y la descarga de la imagen
[6]Para configuración del directorio de trabajo
[7]para la generación de un retraso en el acceso al link, con le objetivo de prevenir un ataque DoS

y un ciclo saber cuantas imágenes hay por anuncio, esto con el objetivo de crear un arreglo donde se guarden los nombres de cada imagen bajo el formato " inicial de la pagina web junto con el codigo de Web Scraper_número de imagen del anuncio que le corresponde ", por ejemplo "F1606408239_6"; los cuales son generados mediante un ciclo y un objeto tipo texto iterativo. Con los nombres listos y los links en una matriz, se configura el directorio de trabajo para que las imágenes queden guardadas por pagina web de donde vienen y se ejecuta un ciclo que recorre la matriz de links, accede a ellos con requests.get y con *nombreimg=nombreA[i]; with open(nombreimg,'wb') as imagen: imagen.write(img.content); time.sleep(2)* hacer como tal la captura/descarga de la imagen.

Cabe mencionar que hubo casos donde no se pudieron captar las imágenes del anuncio, casos donde en vez de salir un archivo de imagen, se obtuvo un archivo error sin peso. La razón de esto se cree que es por vencimiento de los links pero no se pudo verificar si era por esto o por algún otro motivo, la base de esta creencia está en que la descarga de las imágenes se realizó un mes después de la obtención de los links.

5.3. Unión y Limpieza de las hojas de datos

Realizada la descarga de la información de las páginas web, se tienen 2 hojas de datos con los registros de los anuncios junto con los links de las imágenes (una correspondiente a OLX y otra a Fincaraiz). Las hojas de datos tienen por columnas las variables extraídas en cada caso (ver Tabla **5-1** para las variables de OLX y Tabla **5-2** para las de Fincaraiz), mientras que en las filas estarían los datos de los anuncios, con la particularidad que cada anuncio tiene asignadas tantas filas como imágenes tenga, debido a que se capta por imagen un link.

A cada hoja de datos se le aplicó un proceso de organización, de acuerdo a las necesidades. Se unieron en una sola hoja de datos, a la cual se le aplicó un proceso de limpieza. Como se puede apreciar en el diagrama de flujo de la Figura **5-16**, todo el proceso de unión y limpieza de las hojas de datos se constituye de 6 sub-procesos que se explican a detalle en la sección de resultados y son:

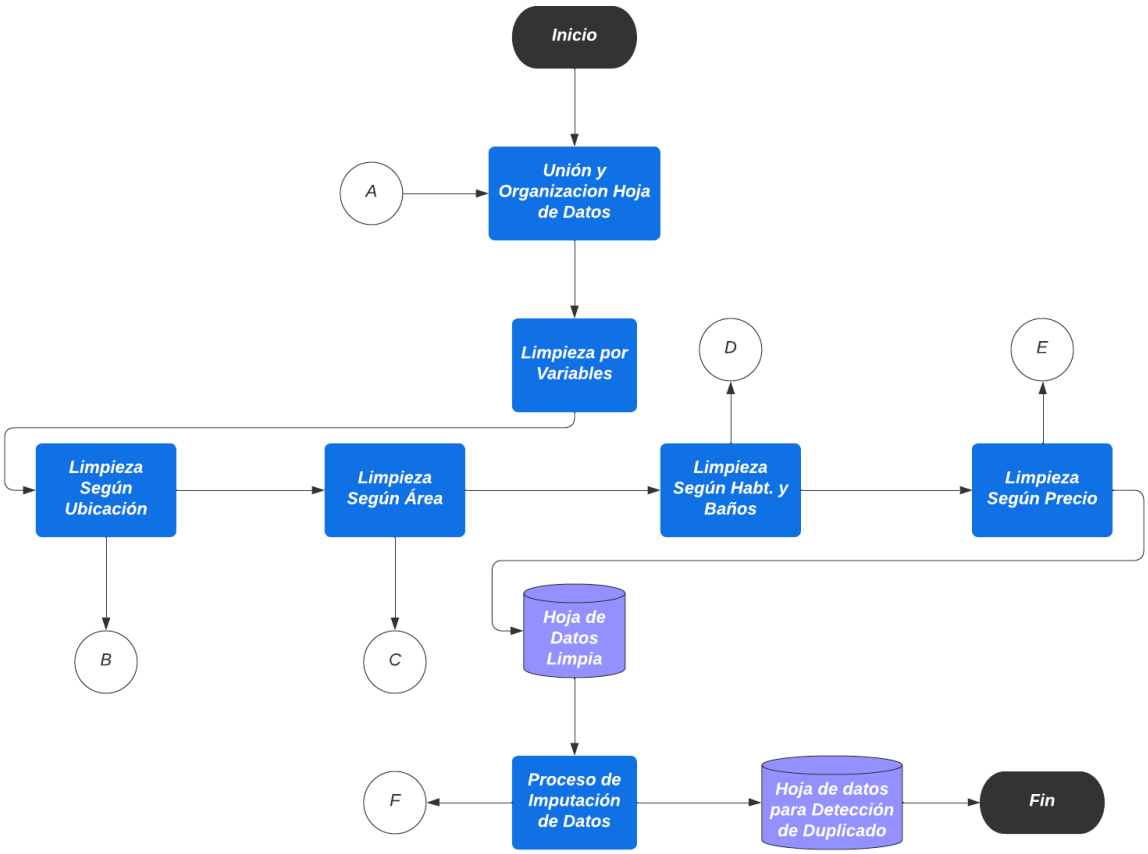


Figura 5-16.: Diagrama de flujo del proceso de limpieza de datos

Fuente: Elaboración propia

A. **Unión y Organización**: Se organizan las hojas de datos, teniendo en cuenta solo las variables que tienen en común, sin inconsistencias en la sintaxis para las cualitativas y sin las unidades de medida para las cuantitativas. Seguidamente, se juntan por bloques fila en una sola hoja de datos, sobre la cual se aplicará la limpieza de datos. Como parte del proceso, se creó la variable conteo de imágenes para así dejar una sola fila por anuncio.

B. **Limpieza Según Ubicación**: Sobre la hoja de datos unificada se aplica primero una regla de validación sobre la variable barrio, para eliminar aquellos anuncios que no corresponden a viviendas ubicadas en Cali. Debido principalmente a que la variable barrio tiene demasiadas categorías lo que la hace compleja de procesar y trabajar, se procede a crea una nueva variable que haga referencia a la ubicación llamada *zona*, a partir de la variable barrio ya depurada y de un reporte del DAGMA (2019).

C. **Limpieza Según Área**: A partir de la nueva hoja de datos unificada, se aplica una regla de restricción que busca eliminar todos los anuncios de viviendas que tenían una área que se salga de lo estipulado por el Decreto 1077 de 2015 del Ministerio de Vivienda, Ciudad y Territorio.

D. **Limpieza Según Habitaciones y Baños**: Tras la aplicación de los 2 procesos anteriores, se encontraron datos inconsistentes en las variables habitaciones (Habt.) y baños, por lo que se eliminaron los registros que los contenían de la hoja de datos.

E. **Limpieza Según Precio**: Al analizar la distribución de las variables, se encontró principalmente que la variable precio presentaba valores atípicos. Por lo que se procedió a hacer una detección de datos atípicos y puntos influyentes para esta variable, mediante el método de la matriz HAT y la consideración de los residuales.

F. **Proceso de Imputación de Datos**: Con una hoja de datos unificada y limpia se procedió a la revisión de datos faltantes de forma gráfica; aquellos que fueron detectados se estimaron mediante la imputación de datos con AFM dada la necesidad de una imputación multivariada por la concordancia y coherencia que deben haber entre los datos de cada registro.

Con los procesos de organización, unión, limpieza e imputación aplicados queda lista una hoja de datos para proceder con la detección de duplicados, con un listado de variables disponibles que se puede apreciar en la Figura **5-17**.

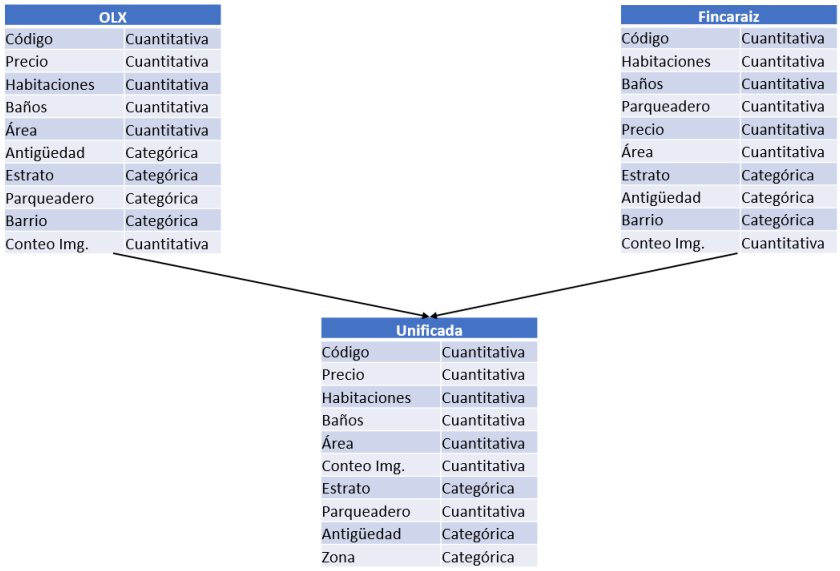


Figura 5-17.: Diagrama de disposición de variables de las hojas de datos

Fuente: Elaboración propia

5.4. Detección y Eliminación de Duplicados

Una secuencia de procesos fueron ejecutados y desarrollados para detectar los registros que se clasificarán como "Potenciales duplicados" mediante técnicas estadísticas. Debido a que solo unos cuantos de estos pares de registros tenían imágenes disponibles, se creó un grupo donde se encontraban aquellos a los que se les podía examinar las imágenes para definir si realmente eran duplicados o no (Comparación de Imágenes), y otro grupo donde se encontraban los potenciales duplicados a los cuales solo se les podía examinar los datos para determinar si eran o no (Similitud aproximada).

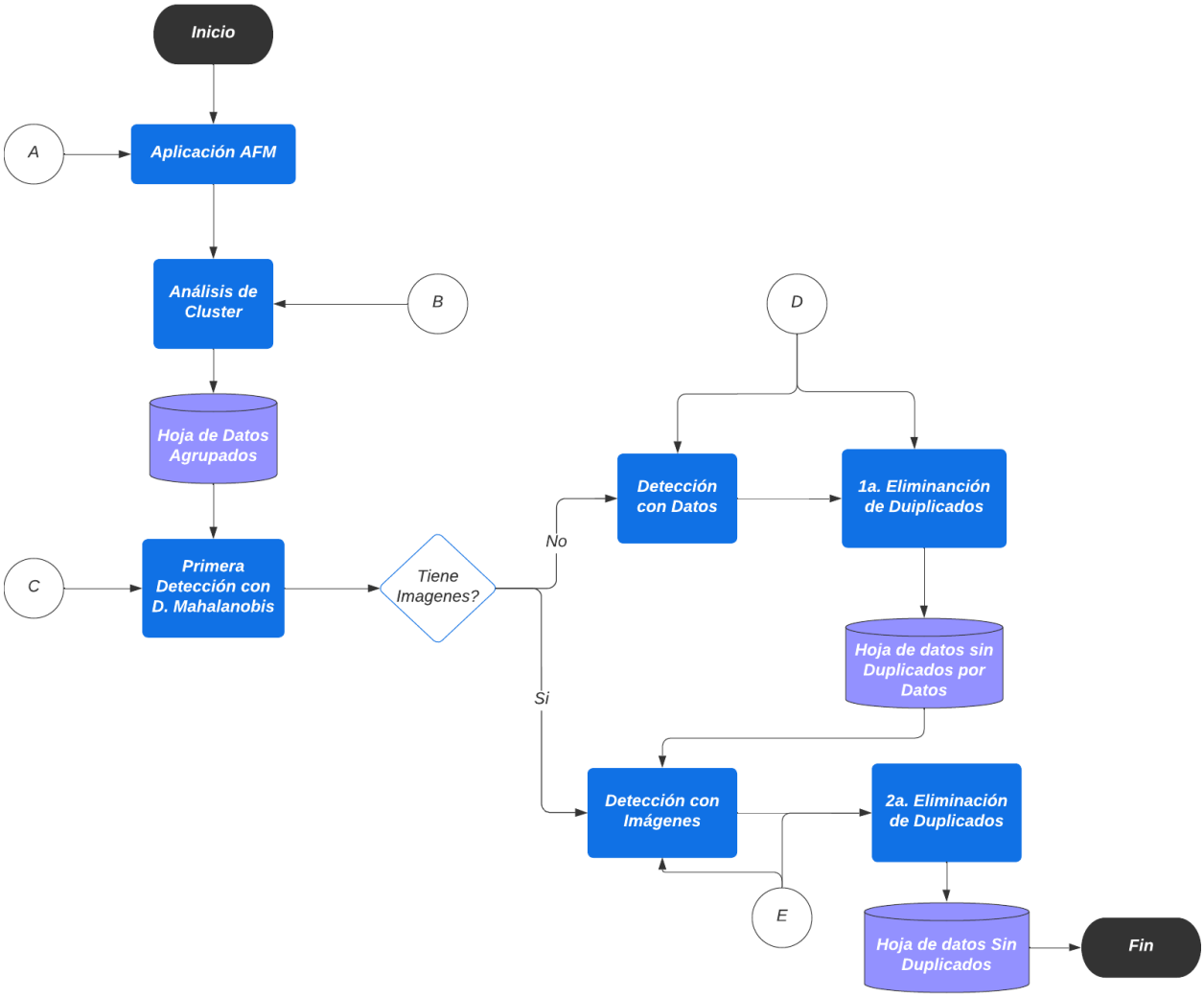


Figura 5-18.: Diagrama de flujo del proceso de Detección de Duplicados

Fuente: Elaboración propia

El diagrama de la Figura **5-18** muestra la secuencia de los procesos principales realizados para la detección de duplicados. A continuación se explican brevemente:

A. **Aplicación AFM**: En primer lugar, se procede a aplicar un AFM sobre la hoja de datos resultante del proceso de unión y limpieza de las hojas de datos, con el objetivo de trabajar con los factores comunes que representen la totalidad de varianza sin considerarlos a todos y así tener un solo tipo de variable, para facilitar el proceso de detección.

B. **Análisis de Clusters**: Tomando como referencia la metodología de blocking, se procede a hacer las agrupaciones de los registros sobre los factores comunes considerados del AFM mediante el método de Ward, para luego en el siguiente proceso mediante la distancia de Mahalanobis comenzar la detección de duplicados.

C. **Primera Detección con D. Mahalanobis**: A partir de las agrupaciones, se procede con la detección de duplicados en cada grupo calculando la distancia de Mahalanobis entre los registros a partir de los factores comunes considerados. Para luego, considerar: "Si un par de registros tiene una distancia de Mahalanobis inferior a un umbral [8] se consideran Potenciales Duplicados".

   Tras la identificación de los Potenciales duplicados, se realizaron las respectivas comparaciones entre las imágenes para clasificar definitivamente los registros como duplicados o no. Dado que, no todos los registros tienen imágenes disponibles, se dividió el conjunto de Potenciales duplicados en 2, el grupo donde se puede comparación con imágenes y el grupo donde se examinan los datos para poder clasificarlos.

D. **Detección con Datos y 1a Eliminación de Duplicados**: Se considero primero eliminar las parejas de registros que tenían una distancia de Mahalanobis de 0, para luego al grupo de Potenciales duplicados a los que se les va a examinar de los datos, aplicarles el concepto de "similitud aproximada" donde a partir de las variables en las que se diferencian se calcula una razón de la Ecuación 6-1 y se define que si una pareja tiene un valor inferior al 5 % en ella, se clasifica la pareja como duplicados y se elimina uno de los registros.

   La eliminación de los duplicados se da en la hoja de datos unificada y se hace un remplazo de registros en el grupo de los donde se puede hacer la comparación de imágenes para el caso en el que uno de los miembros de las parejas de duplicados se encuentra en este grupo.

E. **Detección con Imágenes y 2a Eliminación de Duplicados**: Al grupo de Potenciales duplicados a los que se le puede comprar las imágenes, se aplicó un proceso donde para clasificar una pareja como duplicados, debe suceder que la primera imagen de uno de

[8] El valor del umbral u se definió a partir del cuantil de orden menor que represente una distancia aproximada de 0.1

los dos registros se encuentre entre las imágenes del otro. Para verificar lo anterior, se consideran y comparan las matrices de pixeles mediante la distancia euclidiana para matrices definida en la sección 4.2.4, donde si resulta que un par de matrices tienen una distancia de 0, las imágenes a las que hacen referencia son la misma.

Tras la detección de duplicados con las imágenes, se eliminaron los registros respectivos de la hoja de datos unificada. Por lo cual, finalmente se tiene una hoja de datos descargados de internet sobre la oferta de vivienda en Cali totalmente limpia y sin duplicados, lista para usarse en diferentes análisis y proyectos.

5.5. Tableros de Visualización

Un tablero de visualización de datos o Dashboard es un medio gráfico que permite la visualización ordenada de datos con base a unos parámetros, diseñado principalmente para conocer de forma puntual o gráfica indicadores de interés para la toma de decisiones. En este orden de ideas, se crearon 2 tableros interactivos para demostrar la utilidad de la hoja de datos obtenida en casos prácticos y toma de decisiones.

El primer tablero, mostrado en la Figura **5-19**, permite a quien lo consulte obtener información sobre algunas características de la oferta de vivienda en Cali; este ofrece un par de gráficos que muestran como se distribuye la oferta de vivienda en la ciudad de Cali por zonas y por estratos de forma gráfica; además, da a conocer medidas de tendencia central o de variación para el precio, el área, la cantidad de habitaciones y la cantidad de baños de las viviendas dependiendo de los parámetros interactivos de zona, antigüedad, el estrato y parqueadero que el usuario seleccione.

El segundo tablero, , mostrado en la Figura **5-20**,permite conocer como se distribuye e interactúa el precio con una de las demás variables, mostrando a partir de la variable seleccionada, gráficos diferentes; por ejemplo, al seleccionar la variable estrato mostrará un gráfico de cajas para ver la distribución del precio por estrato, un gráfico de barras mostrando por estrato como es el precio promedio por área, un gráfico de la dispersión del precio y otro del precio vs el área del estrato seleccionado. A la vez, el tablero muestra el precio promedio, el área promedio, la cantidad y porcentaje de registros que hay dependiendo de los filtros aplicados sobre la variable selecta en todo momento.

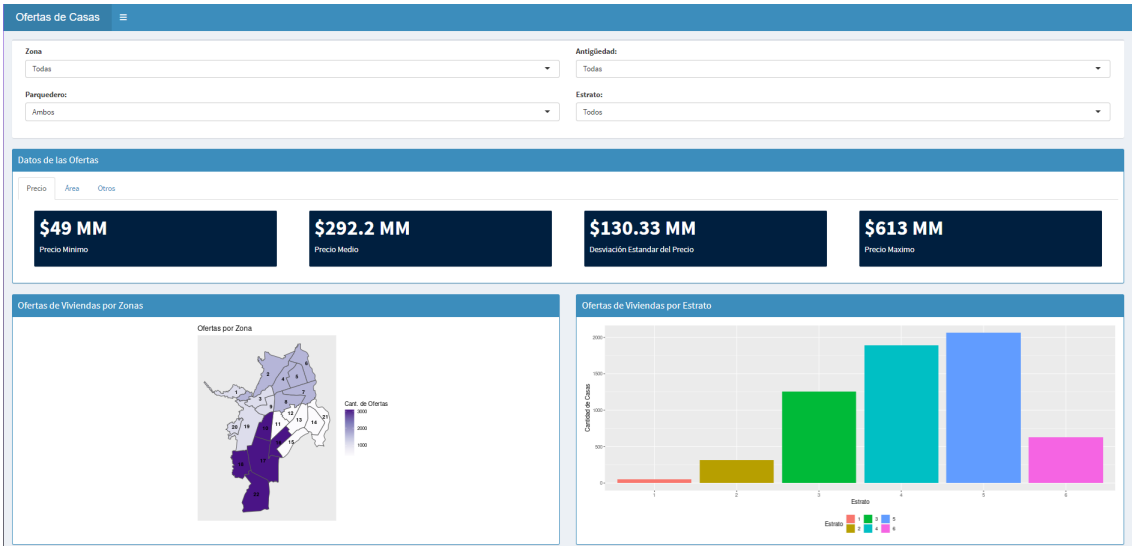


Figura 5-19.: Tablero 1: Análisis de Oferta

Fuente: Elaboración propia

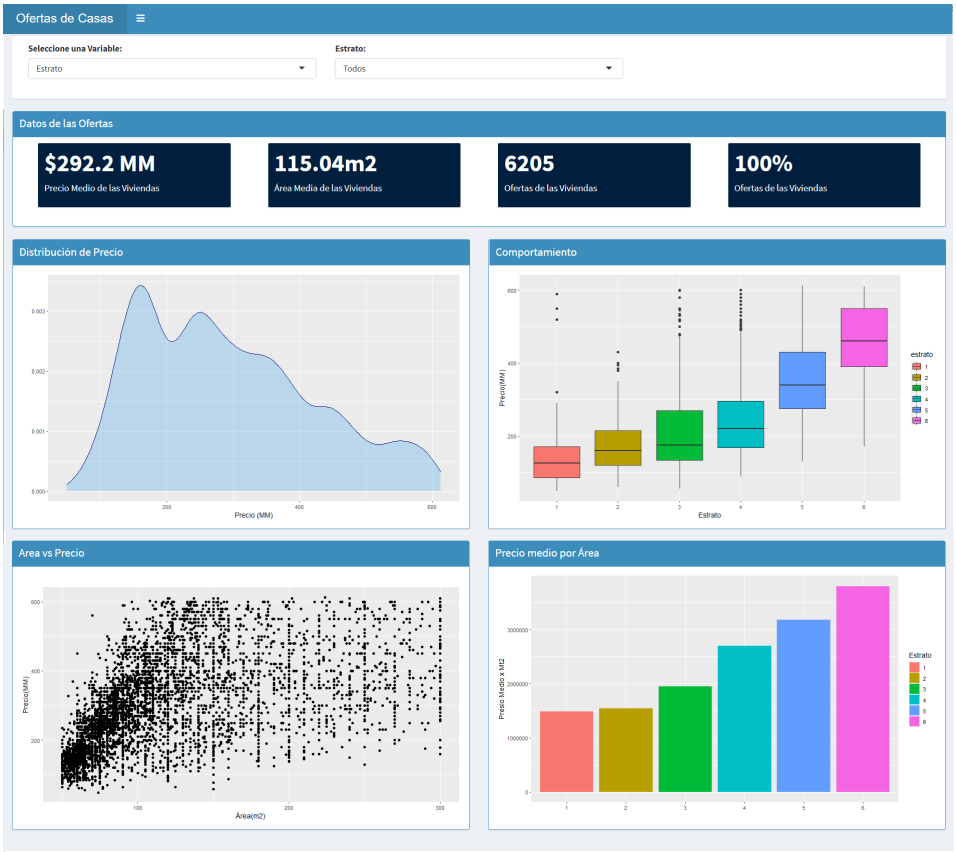


Figura 5-20.: Tablero 2: Visualización del Precio

Fuente: Elaboración propia

6. Resultados

En esta sección se describen los resultados obtenidos en la aplicación y/o desarrollo de las metodologías para la limpieza de datos, la detección de registros duplicados y la creación de un par de tableros de visualización de los datos de oferta inmobiliaria. Se inicia con la descripción del proceso de organización, unión y limpieza realizada a las hojas de datos obtenidas del Web Scraping, con el objetivo de dar mayor claridad acerca de como fueron abordas las particularidades de hojas de datos. Luego, se describe como se hizo la detección de duplicados, mencionando los aspectos considerados para lograr detectar la mayor cantidad de verdaderos duplicados posibles. Finalmente, se hace una breve interpretación de algunos resultados que se pueden visualizar en los tableros.

6.1. Unión y Limpieza de las hojas de datos

Sobre las 2 hojas obtenidas del proceso de web scraping, se aplica el proceso de unión y limpieza de los datos con la intención de estructurar las hojas de datos para conservar la mayor cantidad de variables posibles en la unión. Posteriormente, se aplica el proceso de limpieza de datos para dejar una hoja de datos con las ofertas de vivienda en Cali sin inconsistencias y bien estructurada para hacer una detección de duplicados eficiente.

La Figura **6-1** muestra como inicia y termina el proceso de limpieza, ilustrando paso a paso, así como la cantidad de datos se va reduciendo. La letra O hace referencia a que la cantidad mostrada es de la hoja de datos de OLX, la letra F al caso de la hoja de datos de Fincaraiz y la letra U representa la cantidad de datos de la Unión.

6.1.1. Organización de las hojas de datos

Dado que cada página presentó una estructura y cantidad diferente de variables, se hizo necesario organizar las hojas de datos descargadas para que ambas tenga una estructura de datos bien definida y en el caso de Fincaraiz organizar la mayor cantidad de variables posibles.

- **Organización de la hoja de datos de OLX**: Desde el software Excel se eliminaron las filas duplicadas, generadas por un error en el aplicativo Web Scraper al captar las imágenes. Seguido a esto, se creó una copia de la hoja de datos, en cual solo estarán los links de

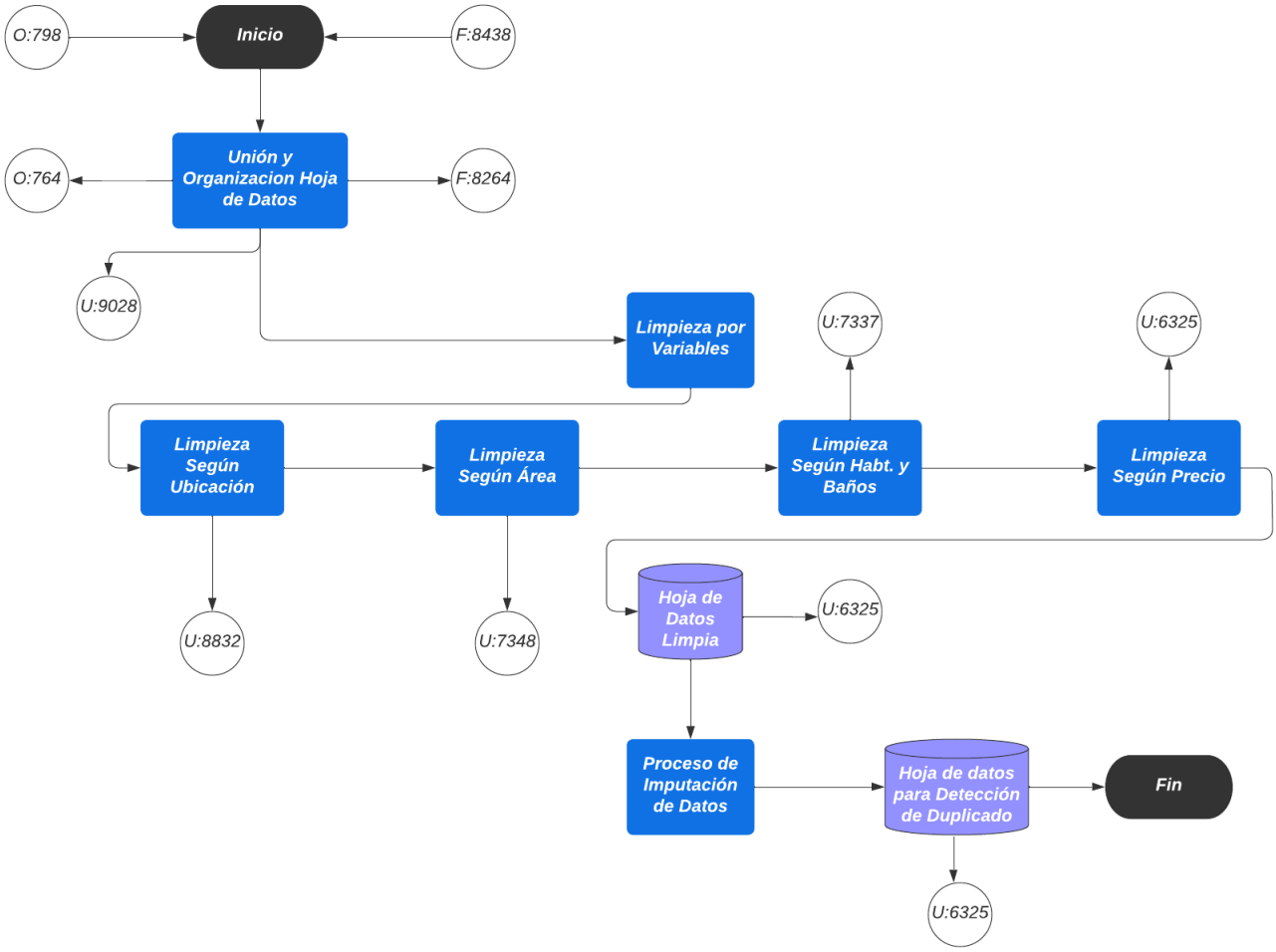


Figura 6-1.: Diagrama de flujo con el impacto del proceso de Limpieza de Datos

Fuente: Elaboración propia

imágenes y el código de descarga (el cual es único para cada anuncio y se usará para identificar a que anuncio pertenece cada link). A la hoja de datos original se le elimina la columna donde esta los links de la imágenes y se le deja una sola fila por anuncio, lo que da como resultado una hoja de datos con 798 registros (anuncios); después, se eliminaron las columnas generadas por defecto por el aplicativo para llevar un conteo y control de los anuncios a los cuales se entró a extraer información. Finalmente, se creó la variable conteo de imágenes, tal y como el nombre lo indica, es el número de imágenes que posee un anuncio. Cabe mencionar, que a las variables de interés numéricas se les quitó la unidad de medida, mientras que las categóricas se estandarizaron[1].

- **Organización de la hoja de datos de Fincaraiz**: Con esta hoja de datos, primero se

[1]Con estandarizar una variable tipo texto o categórica se hace referencia a poner toda la cadena de texto en minúscula

procedió a subir los datos al software R para eliminar los saltos de página ($\backslash n$) que Web Scraper generó de forma automática entre los datos de cada fila, esto se hizo mediante la librería *NPL* y *tm* con la función "*stripWhitespace*", seguido a eso se exportaron los datos sin salto de linea en las filas con la función "*write.csv*".

Con los datos sin saltos de línea se procedió a separar en 2 hojas de datos la información de los anuncios, una hoja contiene los datos de los anuncios y la otra los links de las imágenes. Posteriormente, se organizaron mediante filtros los datos de las campos cambiantes de la hoja de datos con la información de los anuncios para consolidar una hoja de datos estructural con 14 variables de interés, que se describen en la Tabla **6-1**. Los datos de cada uno de los anuncios dentro de esta hoja se encuentran en el mismo orden y aquellos que no cuenten tienen por valor "null". Finalmente, a la hoja de datos final de Fincaraiz también se le añadió la variables conteo de imágenes y a las variables de interés se les estandarizó y removió las unidades de medida de acuerdo a su tipo. Razón

Variables	Tipo-Escala	Descripción	Valores
Habitaciones	Cuantitativa-Razón	Cantidad de habitaciones que posee la vivienda	[0,∞)
Baños	Cuantitativa-Razón	Cantidad de baños que posee la vivienda	[0,∞)
Parqueaderos	Cuantitativa-Razón	Cantidad de parqueaderos de la vivienda	[0,∞)
Precio	Cuantitativa-Razón	Precio al que se oferta la vivienda	[0,∞)
Vendedor	Cualitativa-Nominal	Tipo de persona que ofertaba la vivienda	Nombre la persona natural Nombre persona jurídica
Área	Cuantitativa-Razón	Área en metros cuadrados de la vivienda	[0,∞)
Área privada	Cuantitativa-Razón	Área en metros cuadrados sin contracción del lote donde estaba la vivienda	[0,∞)
Área construida	Cuantitativa-Razón	Área en metros cuadrados construida de la vivienda	[0,∞)
Precio x mt2	Cuantitativa-Razón	Precio por metro cuadrado de la vivienda	[0,∞)
Estrato	Cualitativa-Nominal	Estrato socioeconomico de la zona donde se ubica la vivienda	1,2,3,4,5,6
Antigüedad	Cualitativa-Orden	Categoría de la antigüedad de la vivienda	Menos de 1 año, 1 a 8 años, 9 a 15 años, 16 a 30 años, Más de 30 años
Sector	Cualitativa-Nominal	Hace referencia al barrio o zona de la ciudad de Cali en el que está ubicada la vivienda	Barrios y zonas de la ciudad de Cali
Titulo		Titulo del anuncio de la vivienda	
Piso	Cuantitativa-Intervalo	Piso en donde se encontraba la vivienda	[1,∞)

Tabla 6-1.: Tabla de variables consideradas de Fincaraiz

6.1.2. Unión de las hojas de datos

Con las hojas de datos organizadas y estructuradas, se procede a ver las variables de interés que tienen en común para ajustarlas mediante un proceso de limpieza con el fin de hacer que ambas hojas sean más compatibles. Dicho proceso, como explica en el Anexo A, consistió en la eliminación de anuncios que no hacían referencia propiamente a viviendas de Cali en venta, como es el caso de proyectos de construcción o fincas; en la hoja de datos de Fincaraiz, se transforma en variable indicadora la variable parqueaderos, se dejó como única variable que

hace referencia al área de la vivienda a las variables área construida y se creó la variable barrio para esta hoja de datos mediante las variables título y sector; cabe mencionar, que para el caso de variable antigüedad no se modificaron las categorías de rango de años de cada hoja para la unión dada la poca concordancia había entre ellas.

En este punto, se unen las hojas de datos considerando las variables habitaciones, baños, precio, área, estrato, parqueadero, antigüedad, barrio y conteo de imágenes, para crear la hoja de datos "***Unificada***" que cuenta con 9028 registros y 10 columnas (aquí se cuenta el código de descarga que servirá de ID para cada anuncio) .

6.1.3. Limpieza de registros por Variables

En el proceso de organización se aplicaron técnicas de limpieza de datos para lograr estructurar los datos, pero la verdadera limpieza de los datos se realizó de la siguiente manera considerando las irregularidades encontradas en ciertas variables.

Limpieza Ubicación

Partiendo de la primera versión de la hoja Unificada, se procedió a efectuar una regla de restricción de integridad donde se establece, un criterio de exclusión, teniendo en cuenta los barrios que se encuentran en la lista de barrios y comunas de Cali [2] es eliminado. Tras esto se eliminaron 186 registros, entre ellos se encontraban anuncios de venta de vivienda en Ciudad del Campo y en varios barrios de Jamundí (como el El castillo, Bonanza o Alfaguara). Quedando así en la hoja Unificada con 8842 registros.

De acuerdo al listado de barrios y comunas de alcaldía de Cali, la variable barrios tendría 335 clases, de las cuales no todas están representadas o tienen una cantidad considerable de registros. Lo anterior, se entiende como una problemática tanto metodológica como computacional dado que a la hora de aplicar los análisis multivariados que se tienen contemplados estos presentarían limitaciones e inconsistencias por la cantidad de cálculos que se tendrían que realizar.

Para evitar este problema y como a la vez se cuenta con 290 registros que tienen por barrio una zona de Cali (Centro, Norte, Sur, Este u Oeste) se plantea crear la variable "comuna" con ayuda del listado de la alcaldía y la cual tendría 22 clases, para posteriormente crear la variable "Zona" a partir del reporte de DAGMA (2019) con 4 clases y la siguiente distribución por comuna e ilustrada por la Figura **6-2**.

- **Zona Norte** Comunas 2, 4, 5, 6, 7 y 8.

[2] Se descargó de la página de la alcaldía de Cali `https://www.cali.gov.co/planeacion/publicaciones/3560/idesc/` los datos de geoposición de las comunas y barrios de Cali en formato shapefile que luego fueron leidos en R mediante la función *st_read* de la libreria *sf* para crear una matriz con nombres de los barrios y la comuna a la que pertenecían

- **Zona Sur** Comunas 10, 16, 17, 18 y 22.
- **Zona Oriente (Este)** Comunas 11, 12, 13, 14, 15 y 21.
- **Zona Occidente (Oeste)** Comunas 1, 3, 9, 19 y 20.

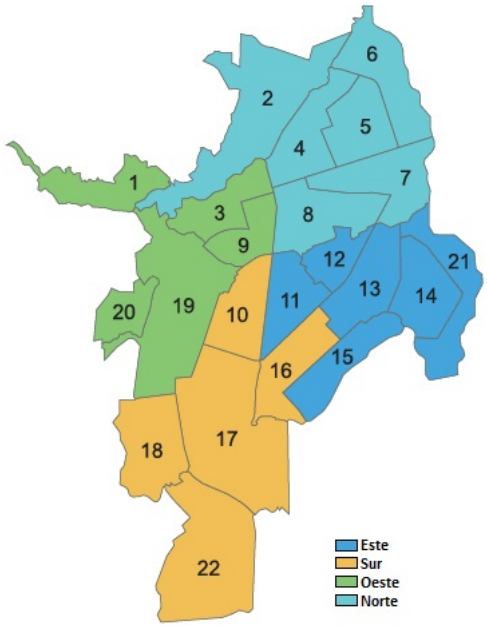


Figura 6-2.: Distribución por Zonas de la Ciudad de Cali
Fuente: DAGMA (2019)

Finalmente, se eliminaron 10 registros que tenían por valor "zona centro" en la variable barrio, dado que no se pudo asignar una zona o comuna para ellos de acuerdo con la información oficial. Lo que da como resultado que después de este proceso de limpieza queden 8832 registros en la hoja Unificada y 12 columnas variables.

Limpieza Área

Al revisar la variable área, se encontraron registros con valores de más de 1000 m^2 en esta variable como se puede observar en la Figura **6-3**, los cuales fueron percibidos como atípicos y para tratarlos se optó por aplicar una restricción de integridad. Esta restricción indica que el área mínima de una vivienda a considerar es de 50 m^2 y el área máxima es de 300 m^2, de acuerdo a lo encontrado en el Decreto 1077 de 2015 del Ministerio de Vivienda, Ciudad y Territorio. El resultado de esta restricción fue la eliminación de 1484 registros que no la cumplían, por lo que ahora Unificada tiene 7348 registros.

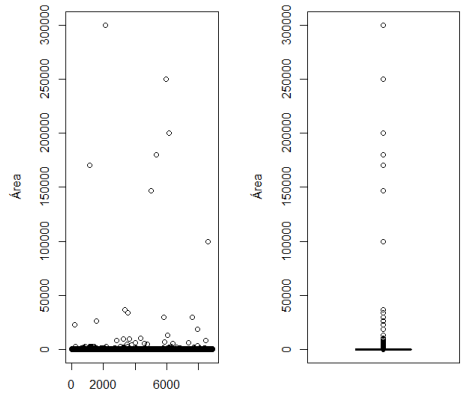


Figura 6-3.: Nube de puntos y Gráfico de cajas de variable Área con datos atípicos

Fuente: Elaboración propia

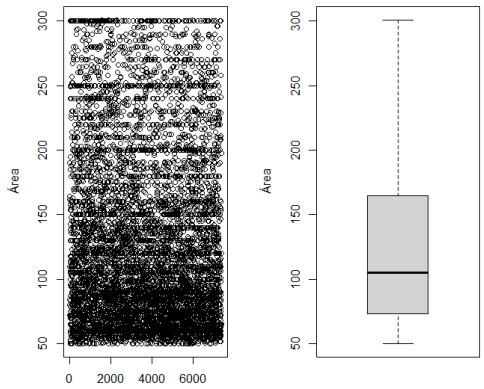


Figura 6-4.: Nube de puntos y Gráfico de cajas de variable Área sin datos atípicos

Fuente: Elaboración propia

Limpieza cantidad de Habitaciones y Baños

Dentro de la variable habitaciones se encontraron 7 registros que resultaron inconsistentes en su valor, 5 por tener como valor "Estudio" y 2 por inconsistencias. Para el caso baños se encontraron 4 registros que tenia por valor en esta variable 0, por lo que se clasificaron como inconsistentes y fueron eliminados junto a los del caso de habitaciones dejando así a Unificada con 7337 registros.

Limpieza Precio

Por último, se revisaron los valores de la variable precio y en ella, como se puede apreciar en la Figura **6-5**, se presentan en un inicio 4 datos muy atípicos; por lo que se decide realizar un proceso de detección de atípicos ya que a diferencia del caso de la variable área no existe una normativa que indique máximos y mínimos para el precio de una vivienda.

Para este proceso, se aplicó 3 veces el método de la matriz HAT propuesto por Hoaglin & Welsch (1978) a través de R, dado que fue el punto, donde se consiguió un rango de valores para la variable sin puntos por fuera del grafico de caja, como se puede apreciar en Figura **6-6**. En la primera vez, se clasificaron como puntos atípicos los 4 valores que como pueden observar en la Figura **6-5** están muy alejados de los demás datos y que el método de la matriz HAT no considero como puntos de balanceo. En la segunda vez, tras verificar gráficamente que los datos presentan aún datos atípicos se volvió a aplicar el método de la matriz HAT, después del cual se eliminaron aquellos registros que no fueron considerados como de balanceo; también, se eliminaron otros 4 registros que presentaban valores poco coherentes para el precio de una vivienda en Cali ($1.100.000, $10.000.000, $3.500.000 y $29.000.000) y que el método no estaba detectando. Dado

que al revisar los gráficos por tercera vez, aun se presentaban precios fuera del gráfico de caja se volvió aplicar el método y ya con esto se logro tener un mejor rango de precios como se puede apreciar en la Figura **6-6**. Como resultado se eliminaron 1010 registros por tener un valor atípico en el precio, por tanto, se tiene que ahora la tabla de datos Unificada cuenta con 6325 registros y 12 columnas-variables.

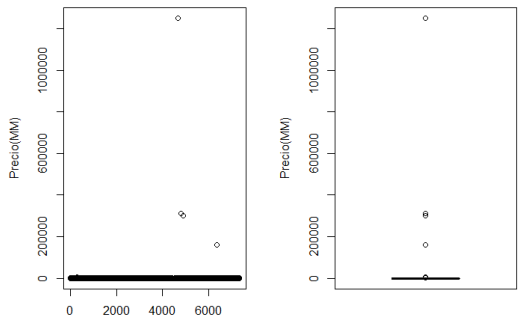


Figura 6-5.: Nube de puntos y Gráfico de cajas de la variable Precio Inicial

Fuente: Elaboración propia

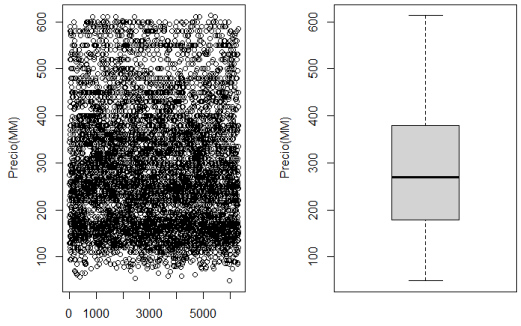


Figura 6-6.: Nube de puntos y Gráfico de cajas de la variable Precio sin datos atípicos

Fuente: Elaboración propia

6.1.4. Imputación de datos faltantes

Dada la forma en la que se obtuvieron los datos, y al hecho que en el caso de Fincaraiz, se organizó la información obtenida para definir las variables, considerando la presencia de datos faltantes, quienes de acuerdo a la teoría de Rubin (2004) califican como MCAR- faltantes completamente aleatorios dado a que se deben a la información que no suministró el usuario que publicó el anuncio en la página web y no por la observación de algún evento. Como se mencionó antes, esta clase de datos es muy problemática a la hora de hacer análisis; por lo que, se decide hacer una imputación de datos faltantes que para este caso seria multivarida a través el método planteado por Husson & Josse (2013) "Algoritmo iterativo AFM", dada la coherencia que debe existir entre las variables de un registro y a que con este método se trabaja con variables numéricas y categóricas a la vez.

En la hoja de datos Unificada, que en este punto cuenta con 6325 registros, se encuentra un 2.5 % aproximadamente de datos faltantes como se muestra en la Figura **6-7**, donde la mayoría de datos faltantes se presentan en las variables antigüedad y parqueadero (con un 1.638 % y 0.519 % respectivamente). Para la imputación se utilizo la función "*imputeMFA*"[3] de la librería

[3]Esta función se creo a partir del trabajo hecho por Husson & Josse (2013)

missMDA de R considerando como variables numéricas el precio, habitaciones, baños, el conteo de imagines y el área (estas variables fueron estandarizadas antes de aplicar la función pues es un supuesto ella considera), mientras que como variables categóricas a el estrato, la antigüedad, parqueadero, zona y barrio (se incluye la variable barrio para ayudar al algoritmo a imputar mejor los datos faltantes de zona). Ya con este proceso aplicado se obtiene la versión de Unificada que se usara para los procesos de detección de duplicados.

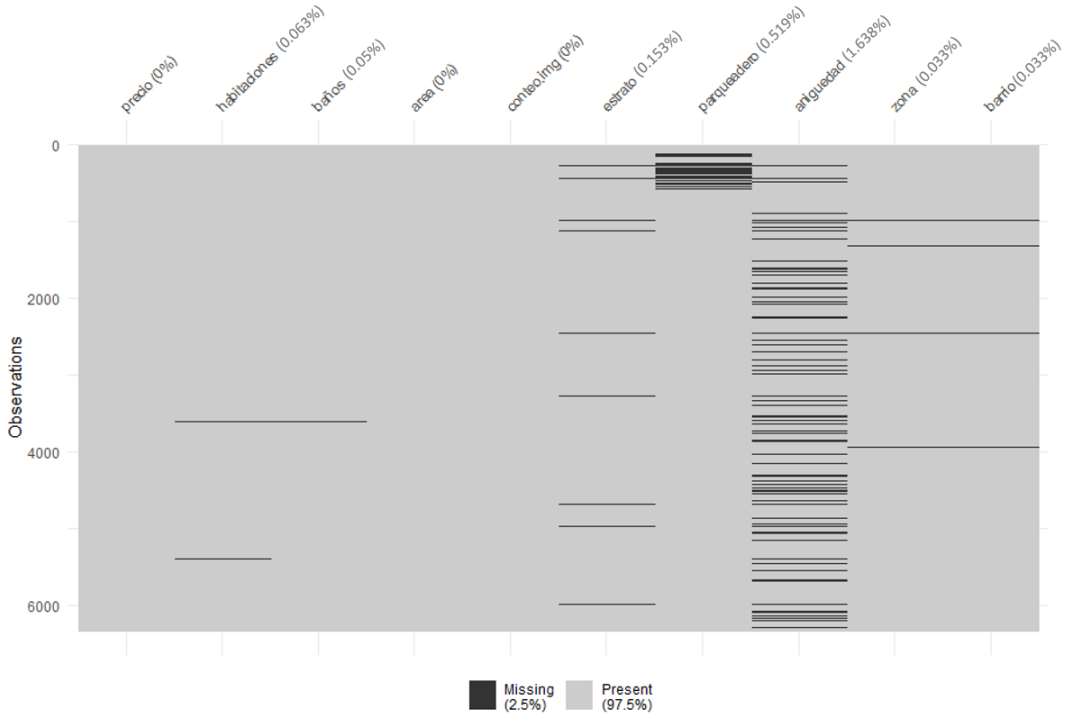


Figura 6-7.: Gráfica de los datos faltantes dentro de la hoja de datos
Fuente: Elaboración propia

6.2. Detección de Registros Duplicados

El proceso de detección de duplicados inicia con la aplicación de un Análisis Factorial Múltiple - AFM en R a través la función *MFA* de la librería *FactoMineR*, sobre una versión de la hoja Unificada donde se descartan las variables-columnas barrio y comuna debido a que contienen demasiadas clases (lo que implicaría un problema para el AFM) y código de descarga que solo se usa como ID de los anuncios; como resultado se tiene una nueva hoja de datos con los mismos 6325 registros pero con 25 columnas.[4]. La razón del análisis factorial está en la posibilidad de tener una nueva

[4] Al revisar la varianza acumulada en los factores se encontró que bastaba con considerar 25 factores pues en este punto ya se recolectaba el 100 % de varianza explicada, en total habían 29 factores comunes debido a que por cada

hoja de datos, donde todas las variables sean numéricas, para así, implementar una medida disimilitud o similitud clásica que permita hacer la búsqueda de duplicados lo más sencilla posible.

De forma consecutiva, se realiza un análisis de cluster sobre la nueva hoja de datos para la creación de los conglomerados en los cuales se van a buscar los pares de duplicados (esencia del método blocking). Dicho análisis es implementado mediante la aplicación del método de Ward y se considera una partición en 100 bloques o grupos (C=100). Teniendo en cuenta, el número de particiones o conglomerados, se eligió a partir de considerar un valor de 0.01 en la perdida inercia absoluta (la Figura **6-8** muestra la gráfica de pérdida de inercia absoluta para cada partición) dado que se espera que los duplicados estén agrupados y se encuentren muy cerca el uno del otro[5].

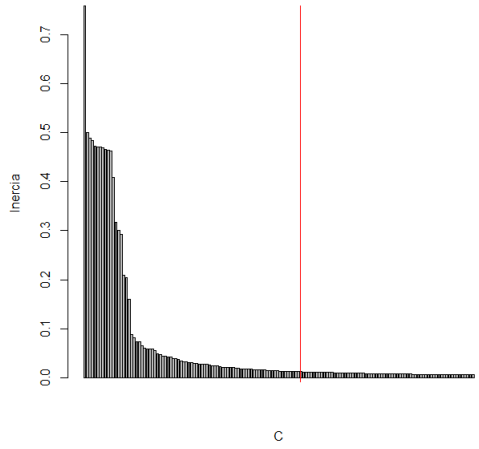


Figura 6-8.: Gráfico de Pérdida Inercia Absoluta
Fuente: Elaboración propia

Tras la división en grupos, se presentan dos consideración importantes: la primera es que el número de la fila de un registro dentro de la hoja de datos del AFM es ahora la ID del registro, ya que, la matriz de datos de cada grupo tiene las filas nombradas con dicho número; la segunda, considerando una matriz cuadrada simétrica para cada grupo y un vector general con las distancias mahalanobis entre los registros de cada grupo considerando como Σ a la matriz covarianzas de la hoja de datos, con el objetivo de usar los cuantiles de orden menor de este vector de distancias para definir el umbral que permitirá clasificar como duplicados a un par de registros. Se consideran entonces los cuantiles inferiores al 5 % de 0.01 % en 0.01 %, aquí se busco el cuantil cerrado que dé la distancia más cercana a 0.1, el cual es el cuantil 0.1 % [6] y por el cual el umbral de

clase de las variables categóricas se crea una componente

[5]En definitiva la partición con la pérdida de inicia absoluta más cercana a 0.01 es 97 grupos pero por efectos prácticos se considera una partición en 100 grupos.

[6]El cuantil exacto que más cercano encontrado fue 0.07 % con un valor de 0.1001 aproximadamente

distancia toma el valor de 0.13448 aproximadamente, entonces, “Si la distancia de mahalanobis entre un par de registros es inferior o igual 0.13448 se considera como pareja Potencial duplicado”.

En este punto, se plantea para cada grupo una búsqueda de aquellas distancias que no cumplen con el umbral, para ello se transforma en una matriz triangular inferior sin diagonal a la matriz de distancias correspondiente con el objetivo de minimizar las referencias múltiples y que los elementos de la diagonal no generen falsos positivos, ya entonces se recorre elemento por elemento de esta matriz y donde se encuentre una distancia que sea Potencial duplicado se guarda en un vector indicador, que se denominara *duplicados*, el nombre de la fila y el de la columna del elemento donde se encontró junto con el número del grupo al que pertenece la matriz de distancia. Una vez identificadas todas las parejas Potencial duplicado, se procede cambiar los nombres de las filas y columnas guardados en duplicados por el código de descarga correspondiente.

La cantidad de Potencial duplicados encontrados corresponde a 327 parejas,encontrando que 34 de ellas corresponden al caso donde bastó con el primer proceso para ser clasificadas como par de duplicados, mientras que, 293 corresponden al caso donde hay que revisar los datos y/o las imágenes para poder clasificarlas.

Con respecto al caso de las 34 parejas, se eliminó uno de los de registros de hoja Unificada de cada pareja y los códigos de descarga de los eliminados fueron buscados dentro de las 293 parejas correspondientes al otro caso por si resultaba que un registro eliminado figuraba como Potencial duplicado en otra pareja y en caso de que si, se reemplazó por el código de descarga del registro que se conservó. Así entonces, se eliminaron los primeros registros que la metodología desarrollada en este trabajo detecto como duplicados y quedando en la hoja Unificada 6291.

Para el caso de las 293 parejas a las que se les debería revisar los datos y/o imágenes para ser clasificadas como par duplicados, al revisar salieron 93 parejas a las que se les puede aplicar la revisión de imagenes y 200 parejas a las cuales por no contar imágenes uno o ambos registros de la pareja no se les puede hacer revisión de imágenes, como solución a esto se plantea hacer la clasificación de estas parejas bajo el concepto de “Similitud Aproximada”.

La revisión de imágenes a las 93 parejas que se les puede aplicar se realizó de la siguiente manera:

1. Se organizó un archivo csv a partir del archivo que contiene los links de las imágenes para el conteo de imágenes por registro y ubicar si es de Fincaraiz o OLX.

2. Se reecreó el generador de nombres usado para la descarga de la imágenes, para tener los nombres de las imágenes a consultar.

3. Para cada pareja, se procede a cargar la primera imagen de uno de los registros mediante la función *readJPEG* de la librería *jpeg* que se denomina imgA; luego se carga y se compara con imgA cada una de la imágenes, imgB, del otro registro, donde si resulta que imgA es igual a una imgB se define que ambos registros tienen las mismas imágenes y por ende son duplicados. La comparación se realiza sobre las matrices de píxeles en escala de grises de las imágenes, examinando primeramente si son de las mismas dimensiones y posteriormente en caso de que si lo sean se mide la distancia euclidiana para matrices entre ellas donde si está distancia es cero se considera que las imágenes comparadas son iguales y en caso contrario se consideran como diferentes.

Como resultado, se encontró que 5 parejas de las 93 eran de pares de duplicados, por lo que se eliminaron de la hoja Unificada uno de los registros de cada uno de los pares. Posteriormente, se realizó el proceso de búsqueda y reemplazo de códigos en caso de que uno de los eliminados se encuentre dentro de las 200 parejas que faltan por revisar.

Para la clasificación, considerando la Similaridad Aproximada se estableció la siguiente regla para una pareja Potencial duplicado, se cataloga como no duplicado primeramente si el barrio del par de registros considerados es diferente, en el caso de que el barrio sea igual en ambos se calcula una razón de diferencia que considera el precio y el área [7] de los registros como se muestra en la Ecuación 6-1 y donde si esta razón es menor a 5 % se clasifica la pareja como par de duplicados.

$$Rdif = \left(\frac{max\{PrecioReg1, PrecioReg2\}}{min\{PrecioReg1, PrecioReg2\}} - 1\right) + \left(\frac{max\{AreaReg1, AreaReg2\}}{min\{AreaReg1, AreaReg2\}} - 1\right) \tag{6-1}$$

A partir de la regla se encontraron 81 parejas clasificadas como par de duplicados, por lo que se procede a la eliminación de uno los registros de cada pareja en la hoja Unificada y con esto final se obtiene una hoja de datos totalmente limpia, sin inconsistencias y sin duplicados, con 6205 registros y considerando para la versión final las siguientes variables: precio, área, habitaciones, baños, estrato, zona, parqueadero, antigüedad y el código de descarga.

6.3. Tablero de Visualización

Con el objetivo de ilustrar lo que se puede hacer con la hoja de datos ya limpia y libre de duplicados, se diseño un aplicativo Shiny [8] que se encuentra disponible en `https://proyectodashboard.shinyapps.io/TableroOfertadeViviendaCali/`. El

[7] Área y precio fueron las variables que a la hora de revisar puntualmente las 200 parejas en las que se presentaban principalmente las diferencias

[8] Shiny es una extensión de Rstudio que facilita la creación de aplicaciones y/o tableros interactivos

aplicativo contiene 2 tableros iterativos creados sobre los datos recolectados; uno fue diseñado para un análisis sencillo del precio y el otro se diseño para un análisis de la oferta. Como el aplicativo y la hoja de datos se prestan para muchas interpretaciones y análisis, a continuación se muestran los tableros y se plantean interpretaciones de algunos de posibles resultados que se pueden generar.

6.3.1. Tablero 1: Análisis de Ofertas

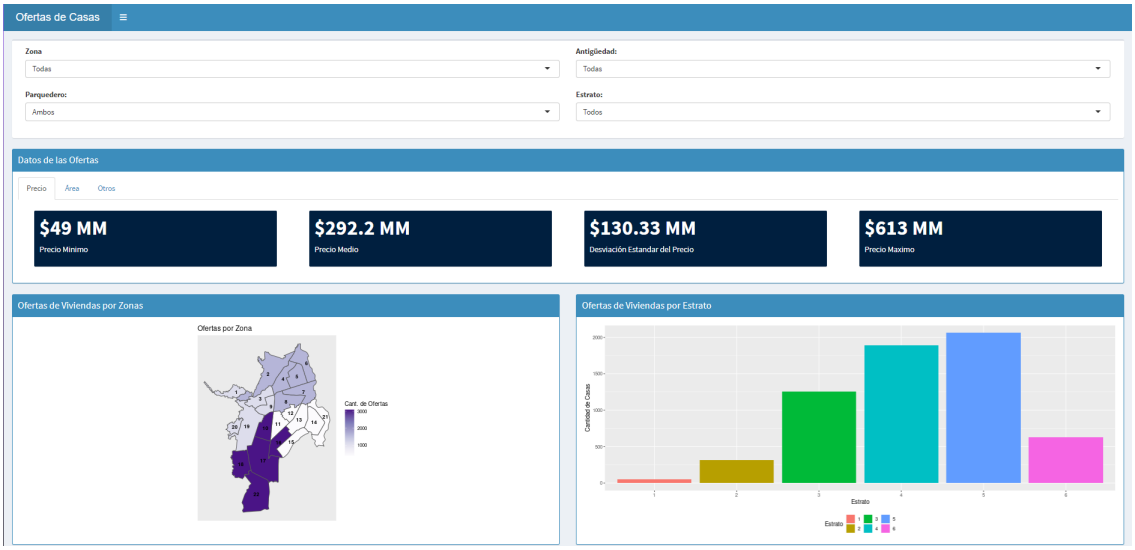


Figura 6-9.: Tablero 1: Análisis de Ofertas

Fuente: Elaboración propia

En este tablero se encuentran 2 gráficos, el de ofertas de viviendas por Zona y el de ofertas de viviendas por Estrato. Del primer gráfico se puede interpretar que las zonas Norte y Sur son en las que más viviendas se ofertan, mientras en la zona Este es en la que menos se hacen ofertas de viviendas. El segundo gráfico refleja la oferta de viviendas por estrato de acuerdo a los valores que se seleccionan en zona, antigüedad y parqueadero, la interpretación de este gráfico va de la mano de los datos que se muestran en la caja Datos de las Ofertas que muestra medidas de tendencia central o de variación del precio, área, cant. de habitaciones y cant. de baños; entonces, con este par se puede dar respuesta a preguntas como: las casas de la zona norte en promedio que precio y área tienen, una casa de estrato 4 de la zona norte normalmente cuantas habitaciones y/o baños tiene, en promedio cual es su área y cuanto vale, otro ejemplo, seria mirar como está la oferta de viviendas en determinada zona y en que estratos se encuentran estas viviendas. Ya ahí queda es jugar con los valores de configuración para saber como está la oferta de vivienda y como esta se distribuye.

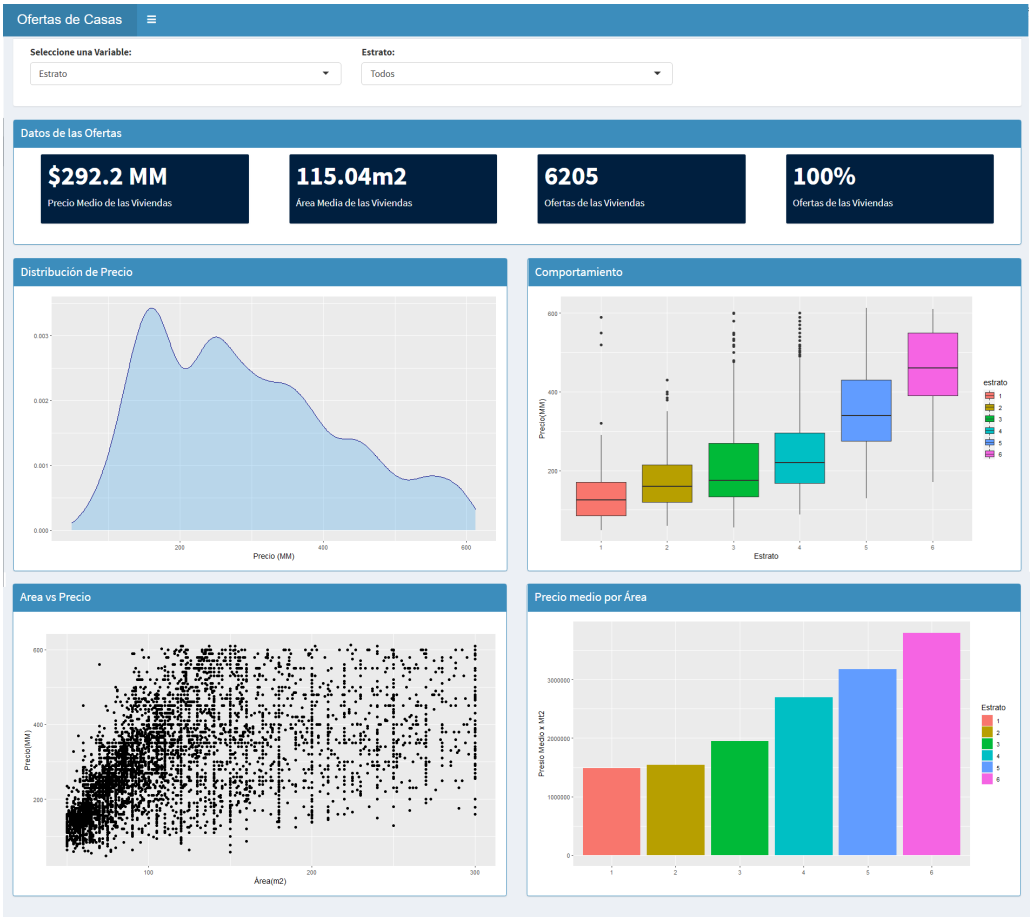


Figura 6-10.: Tablero 2: Visualización del Precio

Fuente: Elaboración propia

6.3.2. Tablero 2: Visualización del Precio

El tablero de análisis de precio mostrado en la Figura **6-10** varia sus resultados dependiendo de la variable seleccionada y dado que hay tanto variables numéricas(cantidad de habitaciones y cantidad de baños) como categóricas se presentan dos casos posibles a considerar en la selección.

El primer caso, se define para cuando se selecciona una variable categórica, en este en la caja de Comportamiento resulta un gráfico de caja o cajas dependiendo si se consideran todas o alguna clase de la variable selecta, a la vez, en la caja de Precio medio por Área se presenta un gráfico de barras del precio medio por metro cuadrado de cada clase o de la clase seleccionada de la variable selecta.

El segundo caso, se define para la selección de una de las variables cuantitativas, en este caso se abren dos nuevas listas para filtrar los resultados de acuerdo a la zona y al estrato. En este caso en la caja de Comportamiento, se aprecia un gráfico de burbujas de Área vs Precio

considerando la variable selecta para el tamaño de las burbujas; en la caja de Precio medio por Área, se encuentra el precio promedio por metro cuadrado de las viviendas que se encuentren según el filtro.

Finalmente, a partir de una variable y filtro seleccionado, se tiene la caja de Distribución de Precio que muestra un gráfico de densidad de los precios de las viviendas, la caja de Área vs Precio donde se muestra la nube de puntos correspondiente de cruzar las áreas y precios de la viviendas y por ultimo la caja de Datos de las Ofertas donde se muestra el precio y área media de las viviendas además de cuantas viviendas se encuentran ofertas para el selecto filtro junto con el porcentaje del total que representa esta cantidad.

Es importante considerar para culminar, posibles interpretaciones encontradas en el tablero, donde se tiene por ejemplo: 1. Reconocer en estrato determinado con relación al precio y área; 2. Distribución de los precios de viviendas ofertas y; 3. Observación de precios rango en la mayor cantidad de viviendas del estrato seleccionado. Adicional, es importante reconocer el tablero como una idea de cotización de metro cuadrado, teniendo en cuenta tanto estrato, zona o antigüedad, permitiendo visualizar un gráfico de burbujas entre precios y cantidad de habitaciones de vivienda.

7. Conclusiones y recomendaciones

7.1. Conclusiones

La metodología de web scraping aplicada a través de la app Web Scraper, permitió la recopilación de información de manera sencilla sobre 9236 anuncios de viviendas ofertadas en Cali. Se resalta hecho que no se requiere de programación compleja en un determinado lenguaje para descargar la información desde el aplicativo, sino de establecer de forma jerárquica un paso a paso de consulta sobre la pagina web en cuestión para extraer la información de forma automática.

Se espera que los algoritmos en Web Scraper y la organización de los datos solo se cambiarían o ajustarían en el caso de que las compañías dueñas de las páginas web modifiquen la estructura o presentación de estas, para brindar mejor experiencia o comodidad a los usuarios de estas. Es necesario resaltar que una vez se cree un algoritmo de descarga para una página web en Web Scraper siempre y cuando esta no cambie se puede ejecutar cuantas veces se desee el algoritmo desde las herramientas de desarrollador de Chrome.

Con la limpieza de datos se logró organizar de forma estructurada las hojas de datos obtenidas del proceso de web scraping. También, permitió a través de transformaciones de las variables, unificar en una sola hoja de datos la información obtenida. Con ella se eliminaron de la hoja de datos Unificada todos los datos inconsistentes según las reglas de restricción consideradas o que podían perturbar los análisis futuros. Cabe mencionar, que la limpieza de datos y la detección de duplicados se debe supervisar para verificar que las inconsistencias presentadas sean las misma a las aquí consideradas, debido a que si las páginas web cambian se pueden presentar variaciones en las hojas de datos obtenidas.

La implementación de un Análisis Factorial Múltiple (AFM) para transformar una hoja de datos con diferentes tipos de variables en una hoja de datos con solo variables numéricas conservando el nivel de variabilidad, resulto ser de gran conveniencia en la detección de duplicados, pues con esta transformación se pudo optar por la distancia de mahalanobis para medir la disimilaridad entre los registros. A la vez, el Análisis de Clusters permitió simplificar la aplicación del método de blocking (método guía seleccionado para la detección de duplicados), al crear los grupos de registros sin crear la necesidad de hacer un proceso de selección de variables para la creación de una blockingkey a través de la cual se formarían los grupos.

7.2. Recomendaciones

Se recomienda hacer la descarga de las imágenes, inmediatamente después de la descarga de datos, para así evitar que por temas de vencimiento de los links, no se puedan obtener.

A la hora de seleccionar la pagina web a considerar para la descarga de datos, es importante tener en cuenta la información a captar, revisando que esta, se encuentre en lo posible, en la misma parte y con el mismo nombre dentro de cada anuncio, con el fin de evitar las llamadas: variables cambiantes, implicando asi, un esfuerzo adicional para sacar información y organizar cada una. A la vez, también revisar la forma de paginación de la página pues dependiente de como sea se deben considerar diferentes pasos en el algoritmo de descarga.

Para la comparación de imágenes se recomienda usar los histogramas de las imágenes, ya que con estos puedo medir la semejanza entre cualquier par de imágenes, mientras que al considerar la matriz de píxeles se tiene la limitación que para poder compararlas deben ser del mismo tamaño, lo cual impide detectar imágenes iguales con diferentes tamaños. Briñez et al. (2013) hablan sobre como hacer estas comparaciones con los histogramas de la imágenes. Cabe mencionar, que no se aplicó esta forma de comparación en este trabajo porque fue descubierto en la ahora de escribir el presente documento.

Bibliografía

Allen, P. J., Lourenco, A. & Roberts, L. D. (2016), 'Detecting duplication in students' research data: a method and illustration', *Ethics & Behavior* **26**(4), 300–311.

Amón, I. & Jiménez, C. (2010), 'Funciones de similitud sobre cadenas de texto: una comparación basada en la naturaleza de los datos', **58**.

Boeing, G. & Waddell, P. (2016), 'New insights into rental housing markets across the united states: Web scraping and analyzing craigslist rental listings', *Journal of Planning Education and Research* pp. 457–476.

Briñez, J. C., Martinez, A. R. & Giraldo, F. E. L. (2013), 'Métricas de similitud aplicadas para análisis de imágenes de fotoelasticidad', *Dyna* **80**(179), 42–50.

Broucke, S. v. & Baesens, B. (2018), *Practical Web Scraping for Data Science : Best Practices and Examples with Python.*, Apress.
URL: *http://search.ebscohost.com/login.aspx?direct=true&db=edsebk&AN=1795255&lang=es&site=eds-live*

Chiquiza, J. (2020), 'Aún no basta con internet para un avalúo', *el Colombiano* .
URL: *https://www.elcolombiano.com/negocios/finanzas/aun-no-basta-con-internet-para-realizar-un-avaluo-GF12424424*

Christen, P. (2007), Towards parameter-free blocking for scalable record linkage, Technical report, Canberra, ACT: Dept. of Computer Science, Faculty of Engineering.

Cubeddu, L., Tovar, C. & Tsounta, E. (2012), 'Latin america: Vulnerabilities under construction?', *IMF Working Papers* **12**.

Cuervo, N. & Jaramillo, S. (2014), 'Precios inmobiliarios de vivienda en bogotá 1970-2013', *Documentos CEDE* .

Cárdenas Rubio, J. A., Chaux Guzmán, F. J. & Otero, J. (2019), 'Una base de datos de precios y características de vivienda en colombia con información de internet', *Revista de Economía del Rosario* **22**(1), 25.

DAGMA (2019), 'Consulta de árboles reportados, ya programados para poda'. Recuperado de: https://www.cali.gov.co/dagma/publicaciones/146160/a-traves-de-codigos-qr-la-ciudadania-podra-saber-la-programacion-de-podas-en-su-comuna/.

De la Fuente, S. (2011), Análisis conglomerados, Technical report, Fac. Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid.

Escofier, B. & Pagès, J. (1992), *Análisis Factoriales Simples y Múltiples:Objetivos, métodos e interpretaciones*, Dunod Paris.

Escofier, B. & Pagès, J. (1994), 'Multiple factor analysis (afmult package)', *Computational Statistics & Data Analysis* **18**(1), 121–140.

Garcia, C. C. (2021), 'Análisis factorial múltiple para describir las condiciones de salud sentidas de la población priorizada de la ciudad de cali en el año 2018.'.

Golub, G., Van Loan, C., Van Loan, P. & Van Loan, C. (1996), *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press.
URL: *https://books.google.com.co/books?id=mlOa7wPX6OYC*

Hadzic, D. & Sarajlic, N. (2020), 'Methodology for fuzzy duplicate record identification based on the semantic-syntactic information of similarity', *Journal of King Saud University - Computer and Information Sciences* **32**(1), 126–136.
URL: *https://www.sciencedirect.com/science/article/pii/S1319157817304512*

Hernández, P. (2011), 'Reflexión de la luz (blog física para bachillerato)'.
URL: *http://curso2012fisica.blogspot.com/p/4-ano-reflexion-y-refraccion-de-la-luz.html*

Hoaglin, D. C. & Welsch, R. E. (1978), 'The hat matrix in regression and anova', *The American Statistician* **32**(1), 17–22.
URL: *https://www.tandfonline.com/doi/abs/10.1080/00031305.1978.10479237*

Holmes, Mark J.and Otero, J. & Panagiotidis, T. (2011), 'Investigating regional house price convergence in the united states: Evidence from a pair-wise approach', *Economic Modelling* **28**(6), 2369–2376.
URL: *https://www.sciencedirect.com/science/article/pii/S0264999311001477*

Husson, F. & Josse, J. (2013), 'Handling missing values in multiple factor analysis', *Food Quality and Preference* **30**(2), 77–85.
URL: *https://www.sciencedirect.com/science/article/pii/S0950329313000700*

Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Vol. 3, Dunod Paris.

Maletic, J. I. & Marcus, A. (2000), 'Data cleansing: Beyond integrity analysis.', *IQ* pp. 200–209.

Martinez, de Lejarza E., J. & Martinez, de Lejarza E., I. (1999), *HipEstat. Hipertexto de Estadística Económica y Empresarial*, Universidad de Valencia.
URL: *https://www.uv.es/ceaces/multivari/cluster/CLUSTER.htm*

Mendoza, H., Vargas, J., Lopez, L. & Bautista, G. (2002), 'Métodos de regresión'. Licencia: Creative Commons BY-NC-ND.
URL: *http://red.unal.edu.co/cursos/ciencias/2007315/index.html*

Morales, R. R. & Azuela, J. H. S. (2012), *Procesamiento y análisis digital de imágenes*, Alfaomega.

Müller, H. & Freytag, J.-C. (2005), *Problems, methods, and challenges in comprehensive data cleansing*, Professoren des Inst. Für Informatik.

Mutis, S. C. (2019), 'Peso del sector inmobiliario', *La República* .
URL: *https://www.larepublica.co/analisis/sergio-mutis-caballero-500033/el-peso-del-sector-inmobiliario-2836824*

Núñez Tabales, J. M. et al. (2008), *Mercados inmobiliarios: modelización de los precios*, Universidad de Córdoba, Servicio de Publicaciones.

Ochoa, A. F. (2018), Análisis de correspondencias multiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo nipals(acmpdd), Master's thesis, Universidad del Valle.
URL: *https://bibliotecadigital.univalle.edu.co/handle/10893/15737*

OIKOS (2020), '¿cómo va el sector inmobiliario?', *OIKOS Noticias* .
URL: *https://www.oikos.com.co/inmobiliaria/noticias-inmobiliaria/como-va-el-sector-inmobiliario*

Pérez, J. & Gardey, A. (2010), 'Definición de vivienda'.
URL: *https://definicion.de/vivienda/*

Rubin, D. B. (2004), *Multiple imputation for nonresponse in surveys*, Vol. 81, John Wiley & Sons.

Salas Plata, J. & Portillo, M. (2008), 'P. ch. mahalanobis y las aplicaciones de su distancia estadística', *CULCyT: Cultura Científica y Tecnológica, ISSN 2007-0411, Nº. 27, 2008, pags. 13-20* **5**.

Tamilselvi, J. J. & Saravanan, V. (2009), 'Detection and elimination of duplicate data using token-based method for a data warehouse: a clustering based approach', *International Journal of Computational Intelligence Research* **5**, 191+.
URL: *https://link.gale.com/apps/doc/A234310771/CDB?u=univalle&sid=CDB&xid=79927643*

Uribe, I. A. & Jiménez, C. (2010), 'Detección de duplicados: Una guía metodológica', *Revista colombiana de computación* **11**(2), 7–23.

Vargiu, E. & Urru, M. (2013), 'Exploiting web scraping in a collaborative filtering-based approach to web advertising.', *Artif. Intell. Research* **2**(1), 44–54.

A. Anexo: Detalles de la unión de las hojas de datos

La unión de las hojas de datos se realizó en Excel juntándose por bloques fila y considerando solo las variables en común, antes de esto, se realizaron los siguientes procesos con el fin de hacerlas más compatibles.

1. Se aplica una restricción de integridad para eliminar de ambas hojas de datos los anuncios que hacían referencia a proyectos de construcción. Esta regla consiste en verificar que los anuncios en las variables habitaciones o baños no presentaran rangos de valores; a la vez, se considero que el precio de la vivienda fuera una cantidad de dinero determinada y no un rango de precios. En este aspecto en OLX se elimino un registro y en Fincaraiz se eliminaron 168 registros.

2. En ambas hojas de datos es descartada la variable piso debido a que la mayoría de los datos que la conforman son datos faltantes.

3. En el caso de Fincariaz, la variable parqueaderos es transformada en una variable indicadora, asignando un *No* a los casos donde esta variable toma el valor de "*Sin Especificar*" y un *Si* a los casos contrarios diferentes de "*null*".

4. Se descartan las variables tipo de vendedor para el caso de OLX y vendedor para el caso de Fincaraiz, debido al nivel de complejidad al definir el tipo de vendedor para Fincariaz a partir del nombre del vendedor.

5. De las tres variables que hacen alusión al área de la hoja de datos de Fincaraiz, se toma la variable área construida como la principal y las otras dos son descartadas; esto debido a que se consideró la mas indicada para ser la verdadera área de las viviendas, excluyendo la variable de menor claridad.

6. Se construye para el caso de Fincaraiz la variable barrio a partir de las variables titulo y sector, dado que dentro de una o ambas variables se encontraba el barrio o la zona donde se ubica la vivienda anunciada. Seguidamente, sobre esta nueva variable se lleva a cabo un análisis de la sintaxis para verificar que los nombres de los barrios estén bien escritos y no se presenten casos de multireferencia al mismo barrio, este se realizó con ayuda del listado de los barrios y comunas de la ciudad de Cali.

7. De la hoja de OLX se eliminan 33 registros cuyo valor en la variable tipo era “Finca”, dado que este tipo inmobiliario no se considera en este trabajo. En Fincaraiz paso algo similar pero en este caso aparecieron 3 registros con estrato “Campestre”, por lo que se interpreto como inmuebles tipo finca y se eliminaron.

8. Se encuentran en Fincaraiz tres registros catalogados como inconsistentes debido a que presentaban varios datos faltantes o valores extraños en alguna(s) de las variables.

9. En la unión resultó imposible organizar correctamente la variable antigüedad, partiendo de que en ambos casos esta es una variable categórica con rangos de años por clases, al revisar si los rangos de uno contenían a los del otro, resulto que estos apenas se interceptaban, por tanto, se decidió que cada registro conservara en la unión la clase de antigüedad que tenia en la hoja de datos origen.

B. Anexo: Creación de los Tableros en Shiny

Los aplicativos Shiny se dividen en tres partes: una denominada *UI* que se destinará para la programación de la interfaz gráfica, las entradas (input) y el espacio de las salidas (output) de la app. Otra llamada *server*, donde se programa y genera cada salida que se desee, se configuran los gráficos, datos, tablas, entre otros resultados que se mostraran en la app. La última parte es la *global* que contiene principalmente lo que es referente a la carga de las librerías y la hoja de datos, código de funciones generales que se usaran en el *server*, entre otras cosas que sirvan de insumo o sean requeridas en la generación de varias salidas.

Para este caso la *UI* en general se configuró en un ambiente de dashboardPage (shinydashboard), que es más amigable con el usuario que clásico shiny y se estructura de la siguiente manera:

- Se tienen 2 ventanas para acceder: una para el tablero de análisis de ofertas y otra para el del análisis de precio.
- Dentro de la ventana del tablero de análisis de ofertas, se configuraron 3 cajas para salidas y una para entradas. La de entradas consiste en 4 listas desplegables configuradas como filtros con las clases de las variables categóricas de la hoja de datos final. Mientras las de salida, una fue destinada a mostrar mediante un mapa de la ciudad la distribución de la oferta por zonas; otra fue para un gráfico de barras del estrato según el filtro de entrada y la ultima caja fue para mostrar las medidas de tendencia central y de variación del precio y el área, la moda de la cant. habitaciones y la moda de la cant.baños de los registros que cumplen con los filtros.
- Dentro de la ventana del tablero de análisis de precios, se tiene 6 cajas en total: una destinada contener las entradas tipo filtro, otra en la que se muestra el comportamiento del área vs el precio, otra donde de acuerdo a la variable seleccionada en el filtro salen gráficos de cajas o de burbujas, otra mostrando la densidad de los precios de los registros que cumplen con el filtro, otra mostrando una gráfica de barras de los estratos con el precio medio por área y por ultimo una caja destinada a mostrar la media del precio, la media del área,la cantidad y porcentaje de los registros que cumplen con el filtro.

En el *server* del aplicativo se programaron 15 salidas, es decir, se generaron 15 resultados a mostrar en la interfaz configurada en la *UI*. Las salidas constan de 8 cantidades y 7 gráficos,

configurados para que de acuerdo a las entradas que seleccione el usuario, vayan cambiando los resultados mostrados. Cabe mencionar, que dicha configuración no aplica para el mapa de la distribución de la oferta por zonas, debido a la decisión de utilizar un gráfico fijo.

En *global*, se dejo aquí la carga de las librerías consideradas, de la hoja de datos unifica final y de los archivos de las coordenadas del mapa de Cali por comunas, para poder hacer que R ubique las cantidades de viviendas ofertadas por zona y dibuje el mapa de la ciudad.