UNIVERSIDAD EAFIT

TRABAJO DE GRADO

Metodología enfocada a la identificación de oportunidades de inversión en el mercado inmobiliario

---

Autor:
Obed Ríos-Ruiz

Director:
Dr. Henry Laniado

Co-director:
Andrés Montoya López

Documento emitido en cumplimiento con los requerimientos para el título Magister en Ciencia de Datos y Analítica

en la

Maestría de Ciencia de Datos y Analítica
Departamento de Ingeniería

23 de octubre de 2021

UNIVERSIDAD EAFIT

Resumen

Maestría en Ciencia de Datos y Analítica

Departamento de Ingeniería

Magister en Ciencia de Datos y Analítica

Metodología enfocada a la identificación de oportunidades de inversión en el mercado inmobiliario

por Obed Ríos-Ruiz

Este proyecto explora la posibilidad de construir un esquema de recomendación para la inversión en el mercado inmobiliario en Colombia. Mediante el diseño y edificación de una arquitectura que logra procesar grandes volúmenes de información de la web, y con la posterior aplicación de una metodología enfocada a la captura de la dinámica del mercado por localidades, se pretende poner al alcance una descripción clara enfocada en datos y métodos robustos que sirva de referente para la toma de decisiones en cuanto a inversión en bienes raíces. Este aglomerado de componentes, primero en su tipo en el contexto colombiano, democratizaría el acceso a la información y entregaría agregaciones de los estados de procesos de oferta y demanda que a la fecha el público general desconoce.

Mediante la aplicación consecutiva de actividades de minería de datos, métodos estadísticos y de ciencia de datos, y finalmente de agregación y presentación de los resultados, se generan visuales clave con fundamento matemático y real de cómo se comporta el mercado dando entrada a la identificación de oportunidades de inversión. En principio, la condición actual del sistema se obtiene mapeando la oferta de bienes en múltiples departamentos y municipios desde avisos publicitarios en sitios web especializados. Estos contienen detalles de las propiedades, lo que incluye la oferta en sí misma, varias características superficiales e información de localidad. A continuación, la información es procesada rigurosamente para obtener una base sólida con la que se calculan variables que reflejen el comportamiento de la oferta en las localidades. En esta etapa juega un rol central la estadística, en especial los métodos robustos, en tanto se desean capturar patrones que permitan caracterizar lo que denominamos “oportunidades” o bienes que se encuentran fuera de la normalidad (por debajo) del total de la oferta. Finalmente, las propiedades identificadas se analizan con minuciosidad en pro de reforzar los hallazgos.

El éxito de este proyecto daría como fruto un nuevo método para interactuar con el mercado de bienes raíces donde cualquier interesado podría informarse y aprovechar el beneficio de los datos. El potencial de la metodología de identificación depende fuertemente de qué tan asequibles sean los resultados para el público interesado, de forma que cualquier agente pueda comprender las razones detrás de las sugerencias a partir de una historia y contexto desde los datos.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Descripción del proyecto

1.1. Planteamiento del Problema

Actualmente crear y simplificar métodos de interacción con los mercados tradicionales se ha convertido en uno de los medios para desplegar nuevas economías, democratizar el potencial en el uso de la información y desmitificar los procesos de oferta y demanda. En el caso colombiano la informalidad de los mercados ha creado barreras que impiden generar vistas claras de los procesos y como consecuencia se encuentran muy pocos actores que puedan explotar las oportunidades en el mercado mismo. Adicionalmente, como se expone en Salcedo-Perez y col. (2020), la informalidad conlleva consecuencias abismales en la economía y el gobierno nacional, y alternamente promueve organizaciones y medios perversos a los que luego la población general se debe atener.

El caso del mercado inmobiliario no es una excepción en esta declaración, en especial porque el público general no cuenta con accesos directos a la información y los datos entregados generalmente son a nivel nacional o se encuentran agregados y cuentan con un atraso de meses o años, lo que restringe el poder de tomar de decisiones estratégicas al instante que puedan generar beneficio común.

Adicionalmente, la carencia de información sobre el estado del mercado conlleva a una descentralización de la información que imposibilita crear referentes claros o comunes sobre los que se puedan comparar las ofertas, imposibilitando el poder de elección de los compradores. Los procesos de oferta y demanda se convierten luego en tratos clandestinos sin garantías en los que una de las dos partes resulta en desventaja y el otro obtiene ganancias considerablemente mayores. Indirectamente, la confianza general de los consumidores en los procesos y actores oficiales del mercado, tal como lo discute extensamente Zapata-Vega (2013) en su monografía sobre el factor de desconfianza en el sector inmobiliario, disminuye sustancialmente y aporta en la generación de un ciclo vicioso.

Una forma en la que se trata la informalidad es por medio de las llamadas “agencias inmobiliarias”, las cuales prestan servicios de mediación, asesoramiento y gestión en transacciones inmobiliarias. Dichas agencias están reguladas parcialmente bajo la ley (las normas se presentan como guías para un buen servicio más que como obligaciones en la prestación de este, ver “Unidad Sectorial de Normalización de Servicios Inmobiliarios, Fedelonjas”) y cuentan generalmente con extensos portafolios para la compraventa, alquiler, permutación o cesión de bienes inmuebles. Los propietarios optan por las agencias por diferentes razones, sea por temas de asesoría profesional, ahorro de tiempo, protección en las transacciones o exposición en el mercado. En Colombia, tal como se demuestra en el informe de demanda de inmuebles en

venta y arriendo para el 2018 por el portal fincaraiz.com.co, la oferta inmobiliaria en las principales urbes suele estar predominada por las agencias inmobiliarias, lo que dado su funcionamiento acarrea dos situaciones: primero, tiende a desplazar a los vendedores independientes y les asigna riesgos; y segundo, obliga a los compradores a dirigirse a estos tramitadores para enterarse de sus opciones. Ambas situaciones son claros síntomas de un mercado monopolizado por una sola clase de agentes que a la vez lo atomizan ya que fragmentan la oferta y ninguno puede influir por sí solo en el equilibrio del mercado. Sobre lo último, esta atomización del mercado también sesga enormemente las posibilidades del comprador ya que debe ceñirse al portafolio del agente inmobiliario. Finalmente, el atraso tecnológico de estas agencias agrava aún más la situación general en tanto, por un lado, restringe la consulta de información a un acto presencial y a registros físicos, lo que en la actualidad claramente reduce la dinámica del mercado, y por otro lado, ralentiza la implementación de nuevos esquemas de negocio y tecnología (ver Instituto Nacional de Contadora Públicos (*Proptech: el avance tecnológico para el sector inmobiliario en América Latina y el Caribe*), por INCP, 2019).

Una alternativa reciente han sido los portales web dedicados al mercado de bienes raíces. Dichos portales, de acceso público, presentan anuncios clasificados de constructoras, inmobiliarias o particulares para vivienda usada, inmuebles comerciales, proyectos de vivienda nueva y otras propiedades, de forma que el público interesado puede informarse independientemente sobre la oferta. En cuanto a información, estas plataformas generalmente se restringen a mostrar los avisos en el instante de consulta y ponen al alcance del usuario una serie de filtros de acuerdo con las características superficiales o locación de los inmuebles. Tal como lo discute Avila, 2019, en Avila (*Cómo funcionan los modelos que están transformando el negocio inmobiliario*), ha habido una transformación en los modelos de negocio de dichos portales y actualmente se centran en las pautas y exposición particular de inmuebles, así, por ejemplo, se priorizan las llamadas "oportunidades", que representan avisos pagados para mostrarse primero al usuario. El objetivo de estos portales es entonces enriquecerse y ofrecer de un banco de anuncios, carente de transformaciones estratégicas para ofrecer las mejores opciones de acuerdo con la capacidad financiera del interesado o inversor.

El potencial de consolidar la información partiendo de los portales web y anexar información clave desde dinámicas macroeconómicas y de mercado, hasta nuestro conocimiento, no se ha explorado hasta ahora. La oportunidad de dar tratamiento a la problemática aquí expuesta mediante una herramienta que no solamente procese los datos sino también que los transforme y presente como un conjunto de recomendaciones con una justificación desde los datos es entonces el objetivo de esta tesis.

1.2. Justificación

El mercado inmobiliario en Colombia es estimulado principalmente por la puesta en el mercado de nuevas viviendas y por la rotación de viviendas usadas. En cuanto a la vivienda nueva este segmento representa uno de los más relevantes en el sector de la construcción, aportando hasta 9,2 billones al año 2019 según el DANE. Por otro lado, el mercado de los usados, según datos de la Superintendencia Financiera para diciembre del mismo año, acumuló montos hasta de 45,9 billones de pesos en saldos de cartera hipotecaria, lo que es considerablemente mayor. Estos dos segmentos representaron en total 1,9 % y 4,6 %, respectivamente, del PIB nacional de dicho año, y

son producto de aproximadamente 19 millones de interacciones entre compradores y vendedores según reporte de demanda por fincaraiz.com.co. La importancia de este mercado radica, según el Fondo Monetario Internacional por medio de su informe Zhu («Los mercados inmobiliarios, la estabilidad financiera y la economía»), en primera instancia, en que los bienes raíces son un facilitador de la actividad económica al ofrecer espacios para que las compañías operen proporcionando así infraestructura empresarial; en segundo lugar, los bienes inmuebles son fuente de empleo en muchas áreas implicando diversos actores en toda la industria de la construcción; en tercer lugar, los bienes son una clase de activos relevante para inversores institucionales y privados; y finalmente, desempeña un importante papel en la provisión de infraestructura en todo el país. Este mercado es considerado, por tanto, como uno de los conductores principales de la economía colombiana, ya que arrastra una serie extensa de interacciones previas asociadas a crecimiento o empeoramiento económico nacional, y posteriormente diluye sus impactos a lo largo de otros mercados que dependen directamente de su dinámica.

En cuanto a los sistemas de información anexos a dicho mercado que se encuentran al alcance del público general, no se encuentra alguno que explote o ponga a disposición el potencial de la aplicación de analítica, para la toma de decisiones hacia la inversión estratégica en dicho mercado. Adicionalmente, las condiciones de la información y la atomización del mercado impiden que existan visuales claras y únicas del estado del mercado, impidiendo la introducción de alternativas públicas basadas en datos que se puedan consultar en pro de transformar la inversión en bienes raíces en un proceso inteligente y justo.

Una forma en la que se puede liberar el proceso de oferta y demanda bajo las condiciones actuales del mercado es por medio de lo que se denomina "arbitraje informado" (en el contexto financiero), definido como la estrategia con fines de aprovechar la diferencia de precio entre diferentes mercados sobre un mismo activo financiero para obtener beneficio económica, normalmente sin riesgo (Economipedia (*Arbitraje financiero*), en Economipedia, 2016). Esta clase de arbitraje aplicado al mercado de bienes raíces tiene sentido en cuanto a que, si un inversor cuenta con un conjunto de información suficientemente amplio y certero que sirva como aproximación de la dinámica real del mercado, puede luego participar con sus fondos bajo la expectativa (con baja incertidumbre) de retorno positivo y aportar en la transformación del mercado inmobiliario hacia uno mucho más activo y robusto. El potencial asociado a esta clase de participación se basa luego en el sesgo de información entre las partes activas del mercado, lo que habilita dos posibilidades, primero, la de explotarlo en vistas de ubicar negociaciones convenientes para el inversor, y segundo, de reformar el entorno general del mercado de modo que todos los interesados puedan interactuar de forma estratégica.

Así, la ventana de oportunidad se presenta en la construcción de herramientas modernas que conduzcan o apoyen al inversor en la toma de decisiones sobre los bienes de interés. Mediante la exposición clara del estado del mercado se reduce la incertidumbre y se generan nuevos métodos de inversión. Luego, las herramientas que se construyan deben explotar el potencial hasta ahora inexplorado de los datos y descubrir las dinámicas ocultas del mercado. Mediante la atención de esta oportunidad también se clarifican y establecen reglas bajo una fundamentación rigurosa y matemática hasta nuestro conocimiento nunca resuelta en detalle en el país. En su máxima expresión el sistema que trate la problemática expuesta podría fácilmente

llevarse a escalas mayores y convertirse en un producto de alta utilidad que pueda ser usado por toda la población interesada.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar una metodología práctica como prueba piloto para la identificación de oportunidades de inversión en el mercado inmobiliario de Colombia. Dicha metodología debería considerar efectos de la oferta por localidades y contar un componente teórico robusto para la estimación del precio de venta o arriendo según sea el caso.

1.3.2. Objetivos específicos

- Desarrollar procedimiento para la consulta y consolidación regular de información sobre el mercado inmobiliario del país
- Constituir un conjunto de medidas que reflejen las variaciones del precio producto de la oferta en las zonas aledañas a los inmuebles bajo estudio
- Diseñar una herramienta estadística para predecir las dinámicas del mercado, especialmente sobre los precios de los inmuebles
- Construir modelo analítico interpretable para generación de recomendaciones de inversión en el mercado

Capítulo 2

Generalidades

2.1. Estado del arte

Los modelos de compra y venta de inmuebles no han sufrido modificaciones sustanciales en los últimos años como lo discute Avila en Avila (*Cómo funcionan los modelos que están transformando el negocio inmobiliario*). La mayoría de los procesos se transan persona a persona por fuera de línea, y la única novedad ha sido el medio de publicación de las propiedades que hoy en día es por medio de portales web, aunque después del primer contacto en línea se retorna al proceso tradicional. Las plataformas virtuales en el contexto colombiano no cuentan con componentes analíticos o auto asistidos que le faciliten, por un lado, al comprador ubicar la propiedad que maximice sus intereses, sean estos de carácter financiero o personal, o, por otro lado, al vendedor presentar sus ofertas de tal forma que sean lanzadas directamente a los interesados. Estas características son evidencia clara del atraso de la estrategia de transformación digital en el mercado inmobiliario tradicional.

A nivel internacional las inmobiliarias han modificado sus modelos de negocio para presentarse más atractivas a un público que hoy en día exige virtualidad y transparencia en los procesos de oferta y demanda. Algunos de estos modelos son soportados por compañías de tecnología que aprovechan y explotan la madurez del mercado para ofrecer servicios especializados que apoyan al vendedor y construyen sus ofertas de forma que se acomoden al perfil de cada comprador. Mediante los nuevos modelos, como el promovido por Proptech en una de sus ramas de innovación (definición ampliada en Lecamus (*PropTech: What is it and how to address the new wave of real estate startups?*), Lecamus), el tema inmobiliario se convierte en una aplicación más de la analítica avanzada, en la cual el objetivo central es diseñar algoritmos y técnicas cuyas salidas, en este caso, sean sugerencias de propiedades a adquirir bajo un capital dado. Las compañías tradicionales en búsqueda de contrarrestar el avance de los nuevos modelos invierten capital para fondear o comprar bienes que puedan monopolizar; no obstante, esta clase de estrategias es altamente costosa e ineficiente en el tiempo.

En el contexto latinoamericano, y en particular el colombiano, como se clarifica en Cubeddu, Tovar-Mora y Tsounta («Latin America: Vulnerabilities under construction?») por Cubeddu, Tovar-Mora y Tsounta, el panorama es considerablemente plano, ya que se identifica una clara ausencia de fuentes de información en línea sobre el mercado con las cuales la oferta pueda ser consultada y los usuarios puedan usar la información para decidir sobre en qué propiedad invertir sus capitales. Aún más, sin estas bases de información se imposibilita poner al alcance del público herramientas especializadas cuyo propósito sea aproximarse a sus demandas o

identificar oportunidades en las cuales el posible retorno sea mayor a la inversión inicial.

El tema central en todos los casos es como dar tratamiento a la valorización de los inmuebles, cuya comprensión se limita, generalmente, al precio. Tanto el vendedor como el comprador emiten valoraciones sobre los bienes y éstas pueden estar en línea o en desacuerdo; sin embargo, la justificación detrás de cada propuesta suele estar atada a un criterio personal y fallan en no tener en consideración las dinámicas generales del mercado o medidas de comparación que permitan posicionar la transacción como una oportunidad de retorno positivo o una mala inversión.

Los acercamientos presentes en la literatura para resolver el tema de los precios en el mercado inmobiliario, según lo presenta Cho (1996) en su famoso trabajo «House Price Dynamics: A Survey of Theoretical and Empirical Issues», suelen estar presentes en una de dos categorías. Por un lado, hay trabajos que se centran en estudiar el problema desde una perspectiva generalista en la cual se crean indicadores que representan el mercado como un sistema y permiten hacerle seguimiento para así tomar decisiones. Por otro lado, hay una vertiente que se centra en la estimación de los precios de cada propiedad partiendo de las características superficiales, de entorno y otras variables exógenas que permiten describir la dinámica de valorización en el mercado partiendo de cada unidad de estudio. Actualmente no existe consenso sobre que metodología debería prevalecer dado que cada mercado, de acuerdo con la región, por mencionar un caso particular, debe estudiarse por separado y las aplicaciones en cada caso pueden ser completamente diferentes una de la otra; no obstante, cada uno de estos enfoques parte del interés compartido de encontrar patrones que le permitan a un público de interés fiarse o abstenerse de participar en el mercado bien sea comprando o vendiendo.

El método más común de estimar el valor de las propiedades es conocido como regresión hedónica, en la cual cada propiedad es vista como la agregación de componentes o atributos individuales (de carácter superficial) que en su totalidad representan la utilidad del bien; no obstante, este acercamiento ha sido demostrado como insuficiente en trabajos como los de Wing y Chin (2003) para capturar dinámicas de mayor nivel asociadas a la localidad y a efectos de la economía que claramente determinan la valorización de las propiedades. Este método falla entonces en proveer de mecanismos dinámicos y estables para la comprensión del mercado, lo que impide tomar decisiones con confianza. A causa de las falencias presentes en la teoría hedónica se han propuesto acercamientos que abarcan desde la inclusión de más variables en la regresión hasta la implementación de métodos basados en "deep learning" para simular la dinámica de todo el mercado, que, sin embargo, no satisfacen la necesidad de entregar información clara, comprensible y justificable para los posibles inversionistas.

Posterior al tratamiento del valor de las propiedades se procede a identificar aquellas que puedan ser objetos de inversión, es decir, de aquellas cuyo valor estimado sea superior al de oferta real en el mercado, tal como en Baldominos y col. (2018). Estas propiedades son de alto interés debido a su potencial de retorno económico; no obstante, este acercamiento parte del supuesto no comprobable que los datos en la entrada se encuentran lo suficientemente diferenciados para efectivamente, por un lado, construir un modelo robusto que pueda luego discernir entre un inmueble "normal" y una oportunidad, y por otro lado, generar luego una estimación de retorno real.

En el contexto colombiano se encuentran pocas referencias que contemplen el problema bajo una perspectiva basada en datos. En algunos casos si bien se hace aplicación de métodos y análisis estadísticos y otros basados en ciencia de datos para resolver el problema su alcance se limita por las aplicaciones reales de las herramientas construidas y en otros casos por falta de fundamentación teórica e inclusive por falta de datos. Un ejemplar del alcance de las aplicaciones se expone y trata en Rubio, Guzmán y Otero (2019), donde si bien la investigación se orienta a la utilización de la ciencia de datos para para la creación y análisis de una base con información de precios y características de venta o arriendo de casas o apartamentos en las principales ciudades de Colombia, este se limita a presentar resultados descriptivos.

El escenario planteado en este proyecto busca implementar un proceso completo que incluya una base robusta de datos con que alimentar un método inteligente y finalice con el despliegue de información relevante e informativa para los inversionistas. Lograr este cometido sentaría una base con que múltiples aplicaciones posteriores pudieran construirse y proveería al público de opciones confiables e inteligentes basadas en datos con que interactuar en el mercado.

2.2. Metodología

2.2.1. Mercado inmobiliario

Las metodologías que predominan actualmente en el estudio de la dinámica del mercado inmobiliario se congregan entre algunas aproximaciones tradicionales por medio de aplicaciones estadísticas y otras más novedosas que mediante el uso de recursos computacionales de gran capacidad y algoritmos inteligentes buscan identificar patrones en el sistema.

A continuación, se detallan varios de los métodos que conforman el estándar en las aplicaciones al problema bajo estudio.

Las aplicaciones estadísticas han sido el estatus quo en la búsqueda de estudiar el precio de los inmuebles, así, Fik, Ling y Mulligan (2003) exploran, por ejemplo, la extensión de ecuaciones regresivas mediante la adición de variables asociados a la localización de los inmuebles que permitiese capturar las variaciones en el precio en las propiedades. De dicho estudio se pudo obtener una aproximación suficientemente robusta para los estándares de la época que serviría posteriormente de base para modificaciones adicionales a esta clase de modelos y para comprender el efecto de atributos no propios de los inmuebles en su valorización. Las bases generales de estas aplicaciones se resumen en la presentación del precio, $\overline{P}$, de los inmuebles como un sistema que es afectado por características estructurales ($\overline{S}$) como el tamaño del inmueble, la cantidad de habitaciones o baños, el piso, entre otros, atributos de las localidades ($\overline{N}$) tales como calidad de las vías, accesos a servicios de transporte, etc., y algunos efectos del ambiente ($\overline{E}$) como calidad del aire y proximidad a parques. Así pues, el objetivo se central en desplegar $\overline{P}$ como $\overline{P} = f(\overline{S}, \overline{N}, \overline{E})$. Este sistema es denominado "regresión hedónica" [Kain y Quigley (1970)] y es formalmente presentado como un modelo lineal (P_n^t) log-lineal ($ln(P_n^t)$) donde el precio P_n^t de la propiedad n en el período t es una función de un número K de características medidas por cantidades z_{nK}^t forma que

$$P_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nK}^t + \epsilon_n^t, \tag{2.1}$$

para el caso lineal y la fórmula a continuación para el caso log-lineal

$$lnP_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nK}^t + \epsilon_n^t, \tag{2.2}$$

donde β_0^t y β_k^t son el intercepto y los términos de las características a estimar, y ϵ_n^t es un término de error. Este modelo ha sido explorado en Haan y Diewert (2013).

Las diferencias entre los estudios presentes en la literatura radican en la presentación de alguna de las ecuaciones anteriores bien sea agregando efectos en el tiempo, cambiando la base de variables que afectan la función objetivo, generalizando el sistema o modificando la estructura del factor de error. Esta clase de métodos, como generalidad, entrega buenos resultados según lo discuten diferentes autores como por ejemplo Goh, Costello y Schwann (2012); Shimizu y Nishimura (2010); y Wallace y Meese (1997); no obstante, y a pesar del gran volumen de información disponible en la web, por mencionar un medio, requiere de muchos datos, los cuales no siempre se encuentran disponibles o son relativamente complicados de obtener y esto puede fácilmente afectar el validez de los resultados. Adicionalmente, en muchos casos se pierde capacidad de interpretación y se presenta un alto riesgo de sobreajuste como se expone en Bourassa, Hoesli y Peng (2003) y en Heene y col. (2014) respectivamente.

Alterno al acercamiento estadístico se presenta también el método por "ventas repetidas", el cual se aplica tempranamente en Bailey, Muth y Nourse (1963) y posterior en Case y Shiller (1987). Dicho método busca medir la calidad de las propiedades midiendo su valor para cada bien en dos períodos de tiempo diferentes y tiene la ventaja de no requerir información específica sobre los inmuebles; no obstante, exige cantidades de datos transaccionales que permitan analizar la variación en el tiempo, lo que conlleva a situaciones de mayor estrés ya que deben emparejarse las ventas históricas y éstas deben haberse transado lo suficiente para obtener una aproximación del mercado. Tal como lo expone Dombrow, Knight y Sirmans (1997), este método conduce a sesgos de agregación considerablemente altos y conlleva a conclusiones erróneas de la condición del mercado.

En los dos casos previos las metodologías requieren poder identificar cada objeto de estudio, lo que no se presenta en los métodos de tendencia central del precio. La idea de estos modelos es poder eliminar el ruido en las estimaciones por medio de la agregación de grandes cantidades de datos y apoyándose en la ley de los grandes números para obtener finalmente índices de representación. Algunas críticas asociadas a este acercamiento, declaradas en Goh, Costello y Schwann (2012), se centran en que al no considerar los detalles de las propiedades se puede resultar en índices incorrectos o susceptibles a sesgos por la demanda en un período de tiempo. Versiones simplificadas como la de Maguire y col. (2016) sobre esta aproximación trabajan en presentar mejoras sobre la cantidad de información requerida y aumentar su flexibilidad.

Como inicio a la aplicación inteligente de información de índole no superficial surge en el mapa de métodos el análisis espacial como alternativa para identificar patrones espaciales en las variaciones de los precios. En particular, Li y col. (2017) exponen un esquema de análisis que no solamente explota el potencial de la dinámica espacial para comprender la evolución de los precios en el tiempo sino también de la "big data" al operar sobre una base de datos proveniente de un sitio web de alto tráfico

destinado a propósitos comerciales sobre inmobiliarios. Los esquemas resultantes de esta clase de metodologías aportan al fortalecimiento de una teoría que exige la integración de herramientas modernas de minería de datos con algoritmos eficientes e inteligentes que reflejen la realidad de un mercado con aristas de alta complejidad.

En la tarea de aplicar métodos modernos para modelar la dinámica de los precios del mercado los árboles de decisión, regresiones de soporte vectorial y árboles de regresión han entregado igualmente resultados prometedores. Park y Bae (2015) presentan un procedimiento basado fuertemente en árboles de decisión que se apoya en otros métodos como *C4.5*, RIPPER ("Repeated Incremental Pruning to Produce Error Reduction"), Naïve Bayesian y AdaBoost para identificar las características más relevantes y aprender las reglas que afectan el precio de venta. Alternativamente, las aplicaciones por redes neuronales y aprendizaje profundo en temas de predicción en los sistemas inmobiliarios, que tradicionalmente se planteaban sólo como contrapartes al esquema puramente hedónico, han adquirido suficiente fuerza para formar tópicos de investigación que a la fecha se encuentran en auge, así, Selim (2009) han expuesto por medio de sus análisis múltiples dimensiones del mercado inmobiliario que no se habían podido explorar debido a las limitaciones de las técnicas habituales.

La mayor agravante, no obstante, de los trabajos previamente mencionados, se centra en la falta de aplicación real de los resultados o de disposición de herramientas que efectivamente puedan ser consumidas por un público general. El objetivo común de dichos estudios es acumular o establecer bases formales sobre las cuales estudios posteriores se puedan justificar. Pérez Rave (2019), contrario al estándar académico, expone una forma en la que herramientas basadas en datos y analítica pueden servir como medio para interactuar y entender el mercado, particularmente el colombiano. A pesar de que la propuesta de este último autor parece tratar el objetivo de nuestra propuesta, éste no logra conducir su aplicación hacia actores que puedan tomar acciones reales y sus resultados se limitan a visuales descriptivas. El potencial de dirigir los procesos hacia temas de inversión es más frecuente en investigaciones informales como en Durrani (*Real Estate Investment: Buy to Sell or Buy to Rent?*), Przytuła (*Are you buying an apartment? How to hack competition in the real estate market*), y Kim y col. (*Which house shall we invest in?*), por Durrani, 2017, Przytuła, 2018, y Kim y col., 2019, respectivamente, donde la aplicación en inversión se presenta como resultado secular de la estimación de los precios y proyección del mercado.

Las metodologías previamente expuestas, en compañía de los extensos resultados que se han logrado a lo largo de los años sobre el tema del análisis de las dinámicas del mercado inmobiliario, se presentan entonces como precedentes para comprender, en primera instancia, la profundidad del problema mismo, y en segunda instancia, la oportunidad de implementar acercamientos alternativos o nuevos que presenten el tema con una nueva luz. En lo que concierne a este proyecto se apoyará fuertemente en algunos de ellos, especialmente los estadísticos, soportándose en herramientas modernas para obtener y transformar los datos, con un foco hacia el sentido práctico de los resultados y el carácter robusto de la estadística aplicada al análisis y construcción de métodos.

2.2.2. Estadística robusta

Distancia de Mahalanobis

Un acercamiento tradicional en la introducción de componentes robustas a la estadística es la denominada distancia de *Mahalanobis*. Entre sus múltiples aplicaciones, dicha distancia es utilizada para la detección de anomalías en conjuntos de datos multivariados, tal como en Ghorbani (2019), y expresa la cantidad de desviaciones estándar a la que se encuentra un punto $\mathbf{x} \in \mathbf{R}^p$ proveniente de una distribución p-variada $f_{\mathbf{x}}(\cdot)$ de la media $\mu = E(\mathbf{X})$ de su distribución.

La distancia de *Mahalanobis* es una medida invariante ante cambios de escala y se prefiere sobre las distancias usuales como la euclideana ya que considera, adicionalmente, las correlaciones entre variables. Considerando entonces que $f_{\mathbf{x}}(\cdot)$ tiene varianza finita y matriz de covarianzas dada por como $\Sigma = E[(\mathbf{X} - \mu)'(\mathbf{X} - \mu)]$, la métrica se define por la expresión a continuación (ver Peña, 2002, p. 88).

$$D(\mathbf{X}, \mu) = \sqrt{(\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu)}, \tag{2.3}$$

donde Σ debe ser una matriz no singular y definida positiva. Nótese que en el caso $\Sigma = \mathbf{I}$ se trata entonces con la distancia euclídea.

El potencial de esta distancia no se limita exclusivamente a problemas de clasificación o clusterización, aunque son los casos donde más claramente es necesario establecer correlación entre diferentes grupos de datos. El análisis de discriminantes y de patrones son otros campos teóricos en los que la métrica también juega un papel central.

Estimador de mínimos cuadrados recortados

Este método estadístico es un acercamiento robusto al ajuste de una función a un conjunto de datos manteniéndose estable frente a la posible presencia de "outliers". Dicho método fue propuesto en Rousseeuw (1984) como una alternativa al método tradicional de mínimos cuadrados mediante el reemplazo de la medida de dispersión, de modo que la función objetivo se replantea como en la ecuación abajo (2.4).

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{h} (r^2)_{(i)}, \tag{2.4}$$

donde $(r^2)_{(1)} < (r^2)_{(2)} < \ldots < (r^2)_{(n)} <$ son los residuales cuadrados en orden y h el parámetro asociado a las iteraciones.

En resumen, al no agregar todos los residuales cuadrados en cada iteración, el método resulta en el mejor ajuste a la mayoría de los datos al asociar los residuales más grandes a aquellos posibles "outliers". Un análisis más extenso sobre este tipo de regresión se puede encontrar en Rousseeuw y Hubert (2017).

Si bien este acercamiento es relativamente simple, tiene alto impacto en cuanto a sus aplicaciones. Su forma facilita la adecuación del método dentro de metodologías de mayor tamaño, lo que lo hace bastante versátil, de modo que, por ejemplo, en el desarrollo de esta tesis, será orientado hacia un tipo específico de regresión hedónica que permitirá identificar oportunidades de inversión en el mercado de bienes raíces bajo estudio.

2.2.3. Métricas de evaluación

Previo al desarrollo central de este proyecto, se debe elegir una función de pérdida para evaluar los resultados obtenidos. Entre las métricas usuales en aplicaciones de regresión prima el error cuadrático medio [Alpaydin (2010)]; no obstante, para la valoración de propiedades en el mercado de bienes raíces es más común acceder a la raíz cuadrada del error cuadrático medio [Afonso y col. (2019)], al error absoluto medio porcentual o *MAPE* [Pérez Rave, Correa y Echavarría (2019)] y al coeficiente de dispersión [Marjan y col. (2018)].

En la aplicación presente se hace elección del error absoluto medio porcentual (*MAPE*). Dicha métrica se define como

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{F_t} \right|, \tag{2.5}$$

donde A_t es el valor del precio real y F_t el predicho.

2.2.4. Método analítico propuesto

El contexto previamente descrito sobre cada tema se encuentra en complemento con el método analítico central de este trabajo de tesis. Con miras a exponer la interacción entre los distintos acercamientos se siguen los diagramas 2.1 y 2.2. El primer diagrama corresponde a la dinámica entre los paradigmas mencionados para la generación de variables insumo que aporten en la estimación del valor. Por su lado, el segundo diagrama explica de forma compacta como la regresión hedónica es combinada con el método de mínimos cuadrados recortados (**LTS**) generando así una "regresión hedónica robusta". El error absoluto medio porcentual se ubica como métrica de evaluación central en el rendimiento del modelo.

Apoyándose en el diagrama a continuación, el método analítico propuesto consiste en la transformación inteligente de las entradas de información como fuente para la generación de nuevas variables y su interrelación. Luego, el algoritmo central consume todos los datos disponibles y es puesto en competencia contra un referente inicial y algunas variaciones de si mismo. En cada punto el rendimiento se evalúa por medio de métricas estándar en este campo matemático. Finalmente, las salidas del método elegido se analizan en detalle y son puestas en contexto bajo el mercado inmobiliario y de construcción, de modo que se entrega como producto final un análisis completo que congrega bienes dignos de ser revisados y negociados, logrando así el objetivo de la metodología.

2.2.5. Ambiente de programación

Todos los métodos han sido implementados utilizado el lenguaje "open-source" `Python`. Las funciones particulares son de autoría propia y los resultados de tipo georreferencial han sido almacenados en formato GIS, lo que facilita su manipulación y análisis en otros programas comerciales.

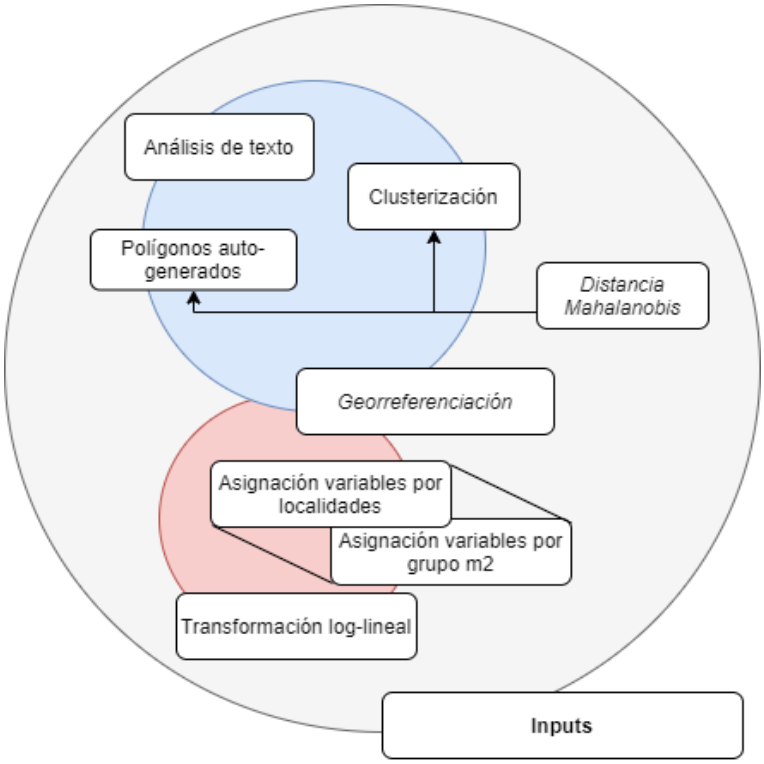


FIGURA 2.1: Métodos para la generación de inputs al modelo, elaboración propia

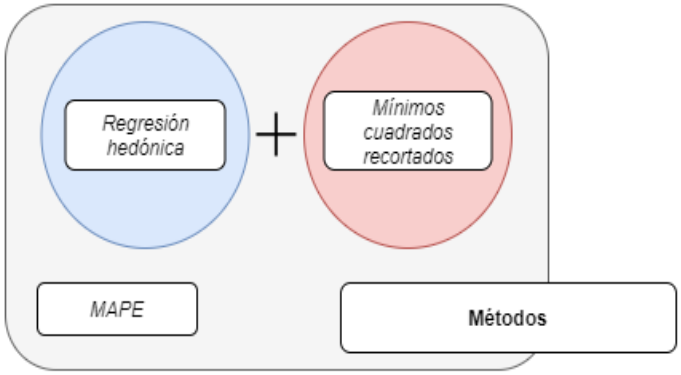


FIGURA 2.2: Algoritmos centrales en la metodología, elaboración propia

FIGURA 2.3: Flujo completo del método, elaboración propia

Capítulo 3

Datos

En este proyecto el conjunto de datos comprende una selección extensa de avisos en Internet a nivel país para arriendo y venta de inmuebles nuevos y usados. Mediante el uso de métodos automatizados que se encargan de consultar múltiples portales públicos se logra generar una base de datos local robusta con detalle de las características superficiales de los inmuebles e información de los precios en el mercado. Mediante *web scrapping* y haciendo uso intenso de métodos de minería de datos con apoyo de funciones intermedias que pre-procesan la información, se recolectan los datos necesarios para todos los departamentos del país, aunque para propósitos de este proyecto se limita a datos del departamento de Antioquia (Figura 3.1).

FIGURA 3.1: Mapa de subregiones de departamento de Antioquia, Colombia, tomado de Wikipedia, the free encyclopedia (2015)

Dicho departamento actualmente cuenta un aproximado de 6 millones de habitantes y se encuentra en nor-occidente del país. El departamento se divide principalmente por subregiones y municipios, y su capital es la ciudad de Medellín, situada en la zona central del Valle de Aburrá (Figura 3.2).

FIGURA 3.2: Medellín, capital del departamento de Antioquia, y su área metro., tomado de Wikipedia, the free encyclopedia (2017)

La dinámica comercial del departamento a la par con sus características geográficas y demográficas lo posicionan como foco de migración y produce que en su interior haya un mercado inmobiliario de gran envergadura. Así pues, la oferta y demanda de inmuebles en la región son constantes, por lo que debe programarse una ejecución semanal de extracción de datos de los portales en búsqueda de capturar dicha dinámica. Con más detalle los avisos en línea tienden a mostrarse como el ejemplo de la Figura 3.3, desplegando inclusive, en algunos casos, las referencias geográficas de los bienes, aunque es más frecuente encontrar descripciones informales de la ubicación de los mismos de acuerdo a la ciudad.

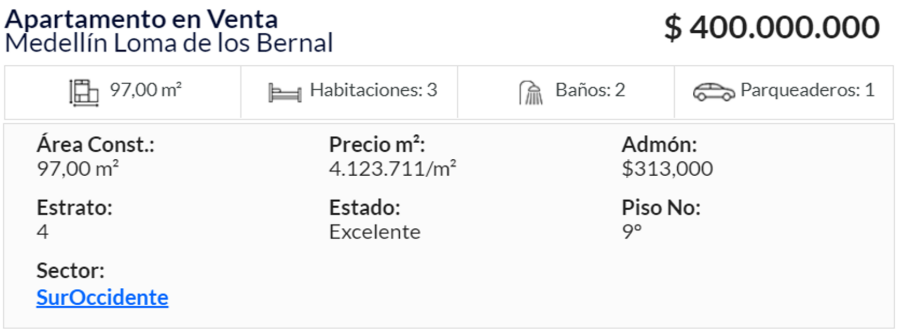


FIGURA 3.3: Ejemplo de anuncio publicitario, tomado de Fincaraiz (2020)

Debido a la alta frecuencia con que se renueva la información, la probabilidad de encontrar avisos duplicados, o bienes publicados en múltiples portales, es considerablemente alta, de modo que se hace imperante eliminar esta multiplicidad al momento de consumir la información.

La base de datos cuenta entonces con registros a partir del mes de febrero de 2020, con un promedio de 778 páginas web consultadas en cada iteración y una media de 9,700 avisos distintos por mes para arriendo y 14,200 para venta. En total se cuenta con 3,780,975 observaciones.

3.1. Descripción

Los datos, como fue previamente mencionado, son ligeramente procesados antes de ser almacenados en las respectivas bases de datos. Entre los pasos aplicados se encuentran análisis de expresiones regulares y asignación de tipos a las variables. En síntesis, la información cruda es llevada a una estructura estática en la cual se consolidan los atributos en tres categorías principales, siendo estas ubicación, características del inmueble e información superficial del aviso. Los dos últimos grupos suelen contener los mismos datos ya que la primera visual de los anuncios tiende a resumir toda la información del inmueble.

Las variables obtenidas son las siguientes:

- **Aviso** | *Period* : Fecha de consulta.
- **Aviso** | *Link* : Dirección web del aviso.
- **Aviso** | *Rooms sup* : Cantidad de habitaciones en el inmueble.
- **Aviso** | *Price sup* : Precio total en pesos colombianos según modalidad.
- **Aviso** | *Ext sup* : Extensión en *mt2* de la propiedad.
- **Aviso** | *Title sup* : Descripción corta.

- **Inmueble** | *Rooms* : Cantidad de habitaciones en el inmueble.
- **Inmueble** | *Price* : Precio total en pesos colombianos según modalidad.
- **Inmueble** | *Surface* : Superficie utilizable en *mt2*.
- **Inmueble** | *Area* : Extensión habitable en *mt2*.
- **Inmueble** | *Group m2* : Grupo de acuerdo a extensión en *mt2*, de $0 - 40$ (grupo 1), de $40 - 60$ (grupo 2), de $60 - 80$ (grupo 3), de $80 - 100$ (grupo 4), de $100 - 120$ (grupo 5), de $120 - 160$ (grupo 6), de $160 - 200$ (grupo 7) o > 200 (grupo 8).
- **Inmueble** | *Title* : Descripción completa.
- **Inmueble** | *Baths* : Cantidad de baños en el inmueble.
- **Inmueble** | *Price m2* : Precio por *mt2* según modalidad.
- **Inmueble** | *Includes administration* : Variable booleada indicando si el valor publicado incluye rubro de administración.
- **Inmueble** | *Administration price* : Rubro en pesos colombianos por concepto de administración de la propiedad.

- **Inmueble** | *Stratum* : Estrato socio-económico asociado al inmueble.
- **Inmueble** | *Floor* : Piso en el que se ubica la vivienda.
- **Inmueble** | *Ages* : Grupo de acuerdo a edad de construcción, *desconocido* (grupo 0), $<$ 1 año (grupo 1), de $1-8$ (grupo 2), de $9-15$ (grupo 3), de $16-30$ (grupo 4) o $>$ 30 años (grupo 5).
- **Inmueble** | *Type* : Modalidad, venta o arriendo.
- **Inmueble** | *Category 1* : Tipo de inmueble (apartamento, casa, oficina o local).
- **Inmueble** | *Contract type* : Tipo de contrato.
- **Inmueble** | *Origin* : Portal de origen.

- **Ubicación** | *Latitude* : Componente de latitud de la coordenada geográfica.
- **Ubicación** | *Longitude* : Componente de longitud de la coordenada geográfica.
- **Ubicación** | *Location 1* : Departamento.
- **Ubicación** | *Location 2* : Municipio.
- **Ubicación** | *Neighborhood* : Barrio.
- **Ubicación** | *Address* : Dirección local.
- **Ubicación** | *Approx loc* : Localidad estimada.

3.2. Georreferenciación

Las variables asociadas a la ubicación del inmueble son clave para la formación del componente $\overline{N}$ mencionado en la metodología (Sección 2.2.1). Dados los componentes de las coordenadas geográficas de las propiedades sería posible, en teoría, agruparlas por zonas y construir métodos basados en distancias para determinar la relación entre su valorización y su cercanía con puntos de interés en la ciudad.

Una evaluación previa sobre el conjunto de datos bajo estudio demuestra que las componentes *Latitude* y *Longitude* han sido determinados con base en la dirección local (*Address*) ingresada en la publicación de las viviendas. Así pues, es común encontrar publicaciones que no cuentan con el detalle georreferencial completo o inclusive que no concuerdan con la referencia a nivel de barrio o localidad. Un ejemplo de esta inconsistencia se puede visualizar en la imagen a continuación, donde, a pesar de tratarse de un local en la zona de *La Candelaria* en la ciudad de Medellín, las respectivas coordenadas lo ubican en la ciudad de Bogotá, Cundinamarca.

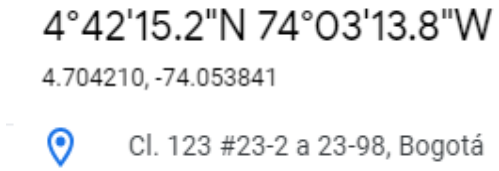


FIGURA 3.4: Ejemplo de inmueble mal ubicado

Teniendo en cuenta la relevancia de estas variables de localización en la valorización de las propiedades, en este trabajo se implementarán técnicas de análisis de texto y de agrupamiento para mitigar el impacto que produce una propiedad georeferenciada incorrectamente.

3.2.1. Clusterización

En vista de esta condición sobre los datos, se opta, inicialmente, por desarrollar un método que permita agrupar los inmuebles alrededor de sus descripciones de ubicación, más exactamente tomando como base la variable *Neighborhood*. La decisión de proceder con un método de dicha naturaleza se justifica en el supuesto de que si bien quienes publican sus propiedades, tanto para arriendo como para venta, no logran posicionarlas exactamente en el mapa, al menos se aseguran que las descripciones asociadas a su ubicación sean correctas.

El método de "clusterización" fue diseñado haciendo uso intensivo de técnicas de tratamiento de texto, en especial métodos tipo *fuzzy*, y expresiones regulares. Posterior a la agrupación se asigna como *tag* aquella referencia con mayor frecuencia en el subconjunto de datos. La tabla a continuación ejemplifica las salidas del método.

CUADRO 3.1

Salidas método de clusterización

Referencias informales	**Frecuencia**	**Tag asociado**
Altos de Niquia	20	
Niquia Edificio Gales	9	
Cabañas de Niquia	5	Altos de Niquia
Exito de Niquia	4	
Niquia	2	

Nota. Frecuencia como unidades asociadas a la referencia informal.

3.2.2. Polígonos

Como fue mencionado previamente, los datos de coordenadas geográficas pueden estar errados para los elementos individuales, lo que impide consumirlos para crear zonas o definir relaciones de cercanía. Luego, aprovechando la salida del paso previo, es posible crear objetos aglomerados robustos y poco sensibles a "outliers".

Los pasos lógicos del método de generación de objetos georreferenciados siguen el flujo desplegado en la Figura 3.5

El paso de mayor interés en este flujo es la **eliminación de outliers**, donde se ha recurrido a la eliminación de registros por percentiles, más exactamente del 90, sobre la distancia de *Mahalanobis* (Ver sección 2.2.2). Este acercamiento es de carácter robusto y facilita la identificación de coordenadas por fuera de la región de interés. Posterior a la limpieza hay dos posibles escenarios: primero, que los puntos se encuentren muy cercanos o sobre el mismo eje (viviendas situadas en una misma calle), en cuyo caso el polígono no es cerrado y se forma una línea; segundo, que los puntos se encuentren bien distribuidos y su polígono asociado sea cerrado. Ejemplos de ambas geometrías se despliegan en la Figura 3.6.

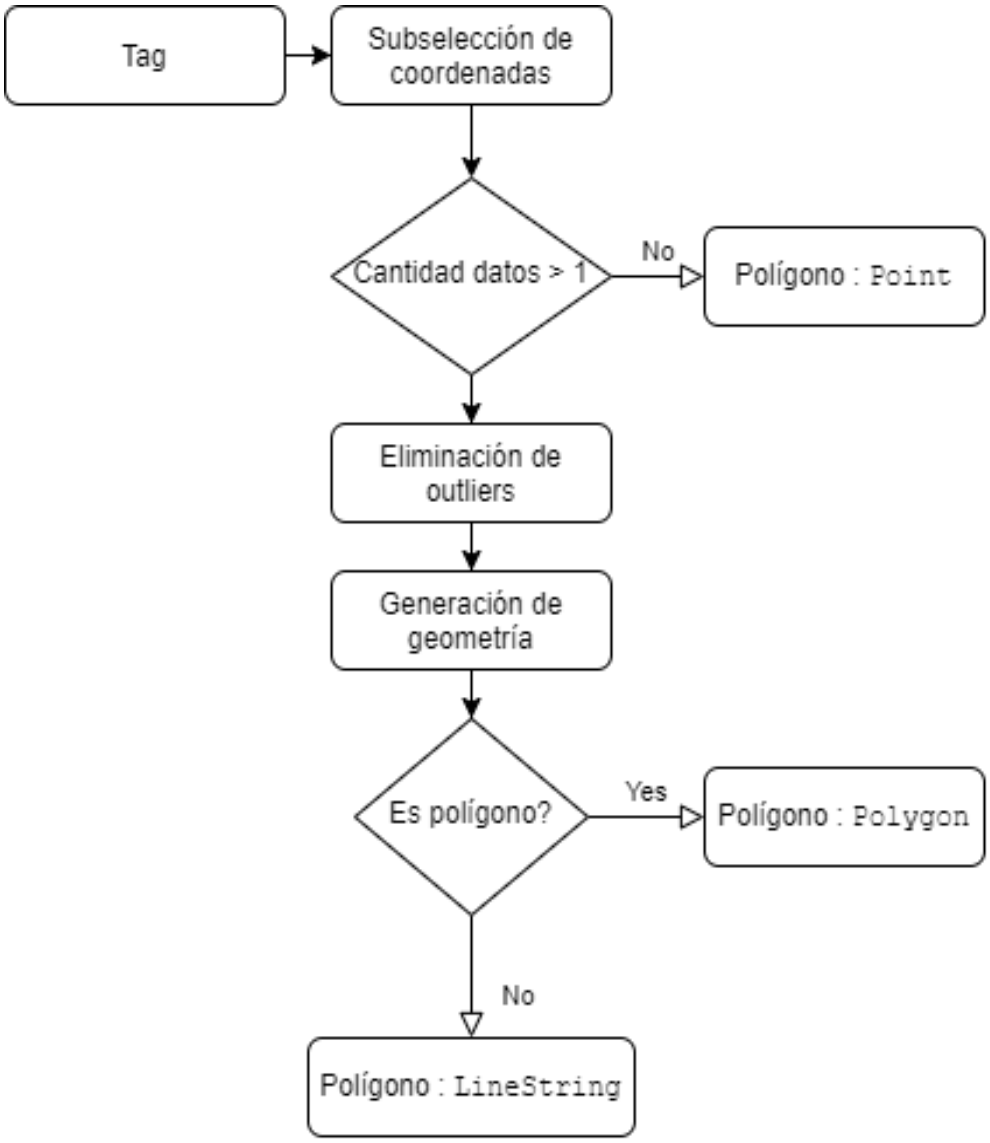
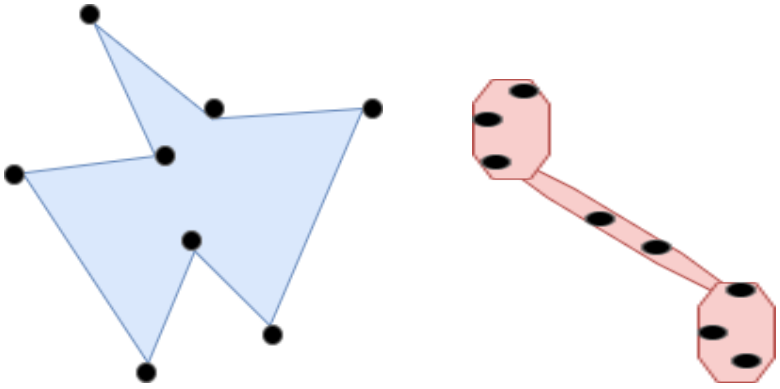


FIGURA 3.5: Pasos lógicos método de polígonos



(a) Polígono multidimensional (b) Objeto lineal

FIGURA 3.6: Polígonos generados

3.2.3. Variables por ubicación

Los resultados obtenidos en la sección anterior permiten, finalmente, calcular características asociadas a las localidades que de otro modo no hubiera sido posible estimar. Partiendo de los polígonos "auto-generados" y sus respectivos centroides, se computa ahora una matriz de distancias y se filtra para cada grupo los 10 polígonos más cercanos, de modo que se trata ahora de vecindades.

Siguiendo la referencia en el Cuadro 3.1 para el caso *Altos de Niquia* se tendrían, por ejemplo, los siguientes objetos catalogados como cercanos:

CUADRO 3.2

Localidades cercanas a polígono

Tag	Posición	Localidad
Altos de Niquia	1	Mirador
	2	Niquia
	3	Niquia Ceiba del Norte
	4	Terranova
	5	Carmen de Viboral
	6	San Jeronimo
	7	Nikia
	8	Niquia San Francisco
	9	Inversión Niquia
	10	Niquia Casa de Justicia

La dinámica producto de la oferta por vecindades, así como el efecto municipal, se capturan luego mediante un conjunto de 32 variables calculadas como el promedio del precio de venta (o arriendo) por metro cuadrado de acuerdo a la categoría. Se ha elegido la media como métrica de evaluación con la intención de introducir sensibilidad frente a inmuebles con precios anormalmente altos. Luego, las variables *Price m2* y *Grupo m2* son centrales, ya que permiten asociar promedios de precios de venta y arriendo por categoría de extensión en metros cuadrados. El diagrama en la Figura 3.7 resume la metodología del cálculo.

Al considerar ambas modalidades de negocio se espera capturar el efecto que tiene un posible exceso de oferta de propiedades en venta (o arriendo) en una localidad sobre los precios de arriendo (o venta).

3.3. Consolidación

El conjunto de datos final tiene un total de aproximadamente 216,151 filas (registros únicos) por 62 columnas. A este nivel, se han aplicado dos filtros a la información. En primer lugar, la cantidad de columnas, que corresponde a las variables, ha sido reducido. Esta supresión se debe a que existen variables procedentes del paso de extracción que no aportan a la estimación, algunas otras son redundantes y aquellas relacionadas con la ubicación no son requeridas posterior al cálculo de las relaciones por localidad. En segundo lugar, hay un paso de razonabilidad que debe tenerse en cuenta sobre la clase de propiedades a analizar en este proyecto. Más precisamente, el análisis debe enfocarse en viviendas familiares que no requieran niveles absurdamente altos de inversión, lo que se traduce luego en propiedades cuya extensión

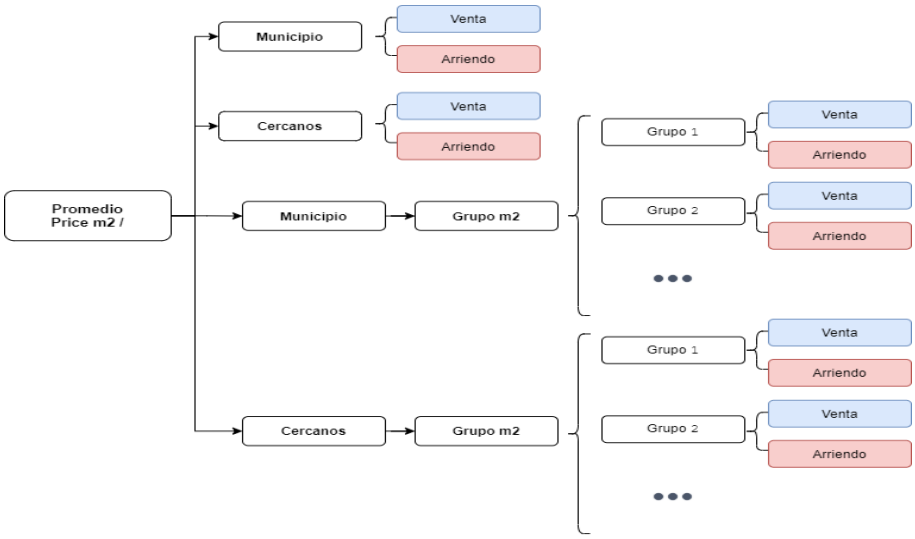


FIGURA 3.7: Lógica de cálculo de variables por localidad, elaboración propia

total no supere los 200 metros cuadrados y cuyo precio por metro cuadrado tampoco exceda los 20 millones de pesos.

Debe tenerse en cuenta la posibilidad de que algunas de las variables no tengan información disponible o no sean verídicas. Esto ocurre habitualmente en los casos en los que quienes publican las ofertas, particulares o agencias inmobiliarias, optan por ser ambiguos en sus descripciones para así hacerlas más atractivas. Desafortunadamente no existen registros oficiales de acceso público que permitan complementar o cerciorarse de la información. Asimismo, crear métodos de imputación de datos faltantes introduciría sesgos no deseados e implicaría cuestionamientos éticos sobre la pertinencia del uso, por ejemplo, de información demográfica para asociarla a las condiciones de las viviendas.

Más información sobre los datos y los rangos de valores para las variables finales está disponible en la Tabla 3.3.

CUADRO 3.3

Descripción de los datos

Variable	Tipo	Rango	Media (Desv. est.)	Valores nulos
Rooms (Rooms sup)	Numérico	$0 - 116$	2,58 (1,11)	0
Baths	Numérico	$0 - 80$	2,08 (0,94)	0
Administration price	Numérico	$0 - 8{,}78$	$3{,}69 * 10^5$ $(1{,}02 * 10^7)$	0
Area (Surface / Ext sup)	Numérico	$1 - 200$	86,38 (37,12)	0
Stratum	Numérico	$1, 2, 3, 4, 5, 6$	4,2 (1,09)	2355
Floor	Categórico	$?, 1 - 16$	–	0
Grupo m2	Categórico	$1 - 6$	–	0
Ages	Categórico	$0 - 5$	–	0
Category 1	Categórico	Apto, Casa, Local, Oficina	–	0
Contract type	Categórico	Particular, Professional	–	0
Type	Categórico	Venta, arriendo	–	0
Location 2	Categórico	–	–	0
Neighborhood	Categórico	–	-	0
Includes administration	Booleano	T (203,707) / F (12,444)	–	0

Nota. En **Rango** "T" corresponde a valor afirmativo, y "F" su contrario.

3.4. Análisis exploratorio

Previo al consumo de los datos para el modelo, se procede a explorar ligeramente como se comportan éstos en relación a la variable del precio y las demás características entre ellas mismas.

Para efectos de simplificación se hará énfasis en la ciudad de *Medellín*. A continuación, se hace necesario separar entre las modalidades, ya que de otro modo se trataría con escalas de valores completamente diferentes, tal como se demuestra en las Figuras 3.8 y 3.9, donde el eje horizontal hace referencia a la variable *Group m2* (grupo por extensión total en *m2*) y el vertical al precio metro cuadrado. Finalmente, los análisis se aplicarán sobre modalidad de negocio *Venta*, lo que en últimas reduce la cantidad de registros totales a 54,981 filas.

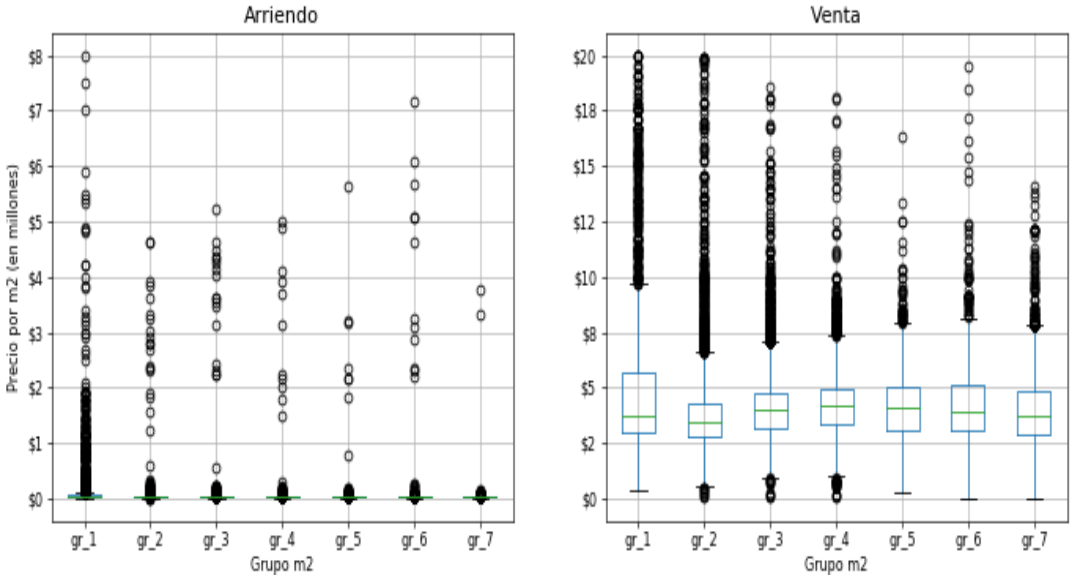


FIGURA 3.8: Boxplot de la distribución de precio por metro cuadrado basado en el grupo por extensión, elaboración propia

Nótese como en la Figura anterior en el segmento *arriendo* el valor máximo se mantiene alrededor de los 8 millones de pesos y sus promedios (Figura 3.9) entre 20,000 y 50,000 pesos por metro cuadrado aproximadamente. Aquellas propiedades bajo arriendo con extensión de área menor a 60 metros cuadrados (grupo 1 y 2) cuyo valor asociado por metro cuadrado sea superior a $1,000,000 corresponden, usualmente, a oficinas o locales situados en zonas exclusivas de la ciudad. Por su lado, los precios de *venta* son relativamente estables y se mantienen entre los 2 y 6 millones de pesos, lo que, para apartamentos y casas, es un rango típico en el mercado inmobiliario (ver Arquitectura & Concreto (*Manual de Inversión:Valor del metro cuadrado en Medellín*), de Arquitectura & Concreto, 2021). También es de notar que hay una población con densidad considerable con valores de venta superiores al promedio mencionado, que, así como con la modalidad de *arriendo*, corresponden a oficinas o locales, y donde es usual encontrar costos de metro cuadrado de esta magnitud debido a las actividades comerciales que allí usualmente se ejecutan y valorizan por tanto la propiedad; de lo contrario, si se trata de apartamentos o casas, se consideran luego como puntos anómalos bajo la muestra de mercado que se está analizando.

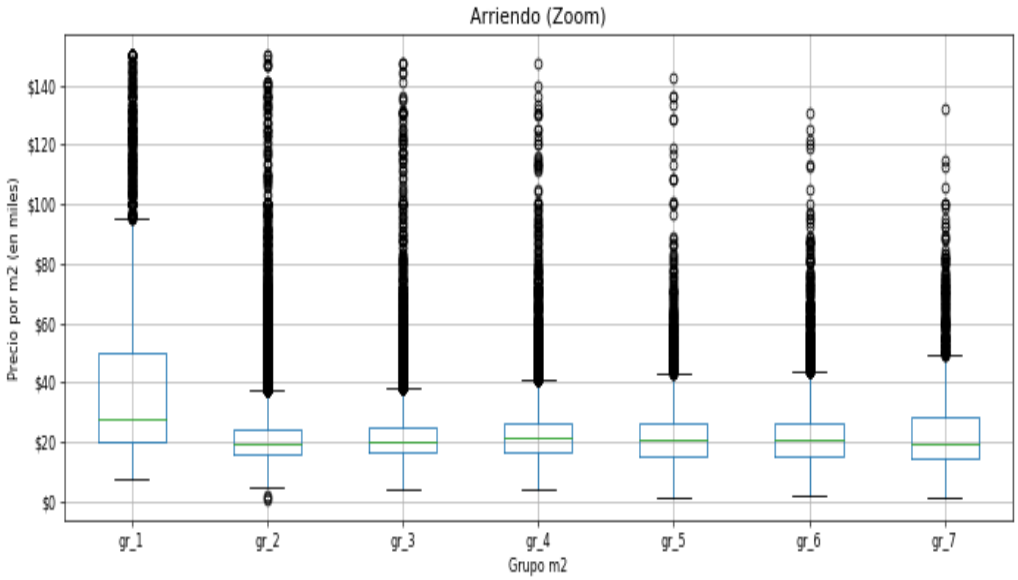


FIGURA 3.9: Boxplot (zoom) de la distribución de precio por metro cuadrado basado en el grupo por extensión, elaboración propia

La tabla a continuación resume la cantidad de registros disponibles posterior al filtro de municipio y modalidad de negocio. El mapa en 3.10 presenta su respectiva distribución espacial. Los apartamentos con superficie entre los 60 y 100 metros cuadrados abundan sobre el resto y su población total supera considerablemente la de los otros tipos de propiedad, lo que es un reflejo de varios procesos, tanto a nivel de ciudad como de sociedad entera.

CUADRO 3.4

Cantidad de registros por tipo de propiedad y grupo por extensión

Grupo m2	Tipo propiedad			
	Apartamento	**Casa**	**Local**	**Oficina**
Grupo 1 - $[0, 40)$	1,220	48	384	135
Grupo 2 - $[40, 60)$	8,062	422	179	200
Grupo 3 - $[60, 80)$	10,988	1,290	125	154
Grupo 4 - $[80, 100)$	9,478	1,307	72	108
Grupo 5 - $[100, 120)$	6,043	1,446	52	52
Grupo 6 - $[120, 160)$	8,154	2,096	83	71
Grupo 7 - $[160, 200)$	3,694	1,447	53	58
Total	47,639	8,056	948	778

Nota. Cada grupo detalla los rangos en m^2 de los inmuebles que agrupan.

Sobre lo último, según el DANE en su reporte DANE (*Censo de Población y Vivienda 2018*), la tendencia de los colombianos de vivir en apartamentos se elevó del 25 % al 32 % en tan sólo 13 años dado que las familias son cada vez más reducidas (en promedio de 3,9 a 3,1 personas), y en particular en el departamento de *Antioquia* el 46 % de la población total convive en apartamentos. Adicionalmente, como lo documenta CAMACOL, los proyectos de construcción se están centrando hoy en día en unidades residenciales de gran capacidad, lo que en últimas incrementa la oferta y reduce las tasas de interés, de modo que las familias de estratos bajo y medio pueden

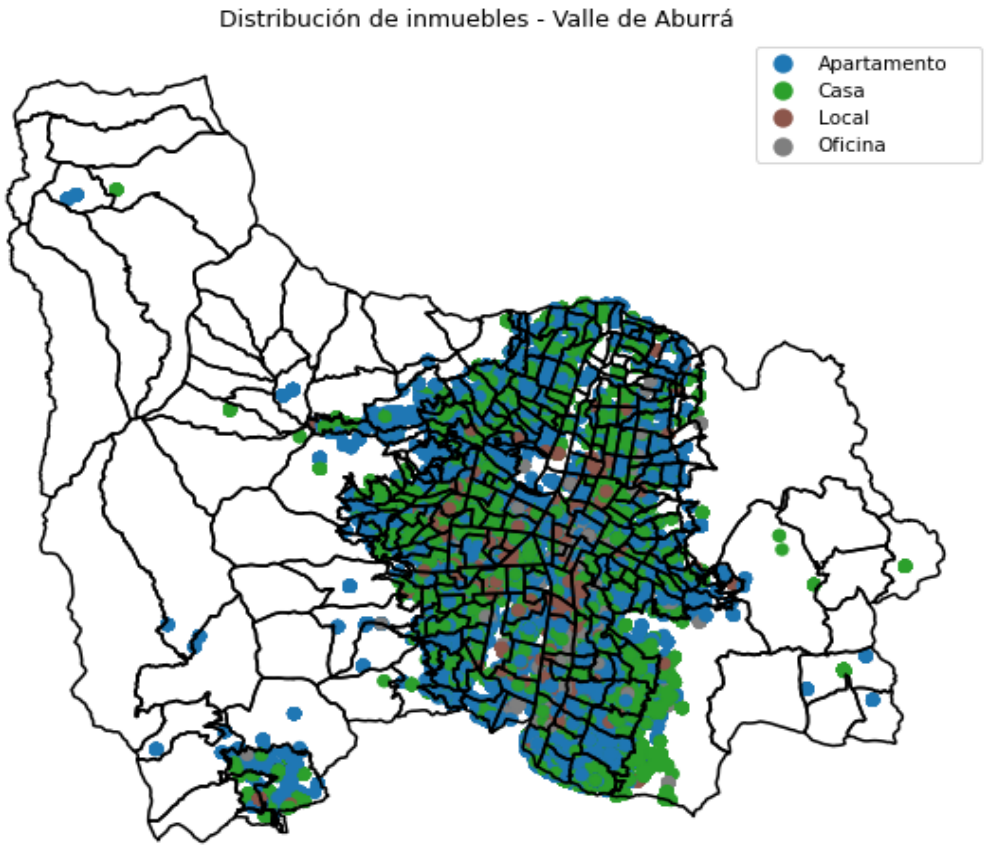


FIGURA 3.10: Mapa de inmuebles ubicados en la ciudad, elaboración propia

acceder a vivienda propia. En paralelo, el estado ha lanzado varias modalidades de acceso a la vivienda con apoyo de varias entidades. Los incentivos tienen la finalidad última de aumentar la probabilidad de aseguramiento de vivienda, y lo hacen por medio de mecanismos como subsidios para viviendas de interés social (VIS) y apoyos directos con las entidades bancarias, propiciando así demanda y oferta sana en el mercado.

En relación a la estabilidad de los precios para la modalidad de estudio, esta se extiende a las distintas categorías de propiedad en el conjunto de datos tal como se muestra en la Figura 3.11. Los rangos estándar de valores de venta tanto para apartamentos como para casas se comprueban también allí.

Debido a las actividades comerciales que se llevan a cabo en los *locales* su distribución de precios tiende a ser superior que cualquier otro tipo de propiedad. En cuanto a los *apartamentos* y *casas* se mantienen todavía sobre el rango previamente mencionado. De notar es que las *casas* fluctúan menos que los *apartamentos* por categoría de superficie total, lo que es un motivo más para interesarse en ellas. Además, desde la perspectiva de un inversor, es más rentable explotar terrenos para construir complejos habitacionales (para posterior venta o arriendo) que una sola unidad de vivienda. Sobre las *oficinas*, aunque también abundan, no son foco de interés en el desarrollo actual, aunque existe un nicho de mercado bastante dinámico especializado en esta clase de inmuebles.

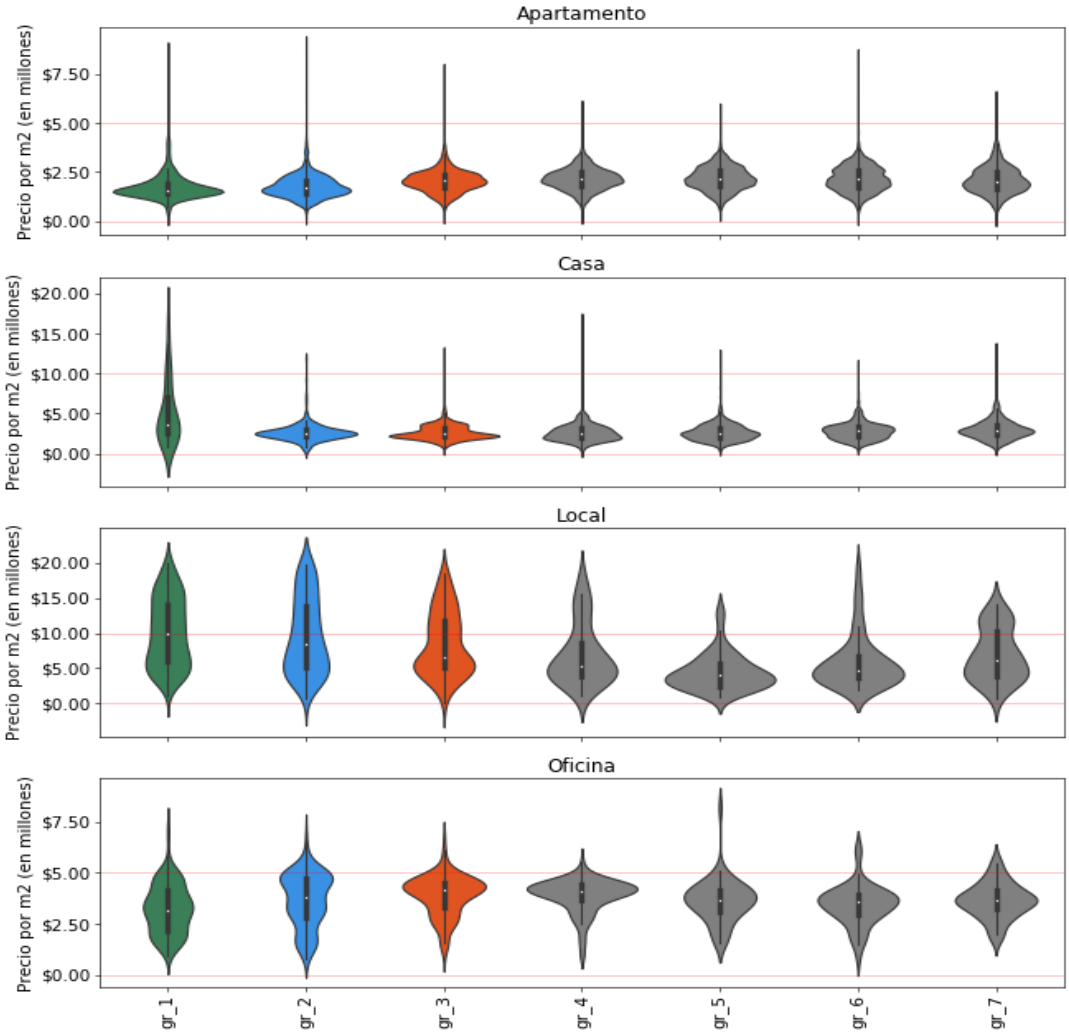


FIGURA 3.11: Violinplot de la distribución de precio por categoría de propiedad basado en el grupo por extensión, elaboración propia

En el cuadro 3.5 se resume la agregación de precios promedio más bajos por tipo de vivienda y grupo m2 para las localidades con mayor densidad en el conjunto de datos. Para el tipo *casa* con extensión total menor a $40m^2$ la cantidad de registros no es significativa (Ver Cuadro 3.4) por lo que se omite dicho grupo. El hallazgo más relevante en este punto es que los precios por metro cuadrado a medida que se habla de propiedades más extensas tiende a ser menor, para ambos tipos de propiedades. Las localidades que surgen en este análisis no se ubican en las zonas típicamente más apetecidas en la ciudad al noroccidente como lo son *El Poblado*, *Envigado* o *Laureles*, sino en sectores crecientes y populares de la ciudad.

El diferencial en precios promedio respecto las localidades con mayores precios (Tabla 3.6) es considerablemente alto, con un máximo de hasta 5 millones de pesos para los apartamentos del grupo 3 ($60 - 80\ m^2$) y 3 millones para las casas del grupo 4 ($80 - 100\ m^2$). A valores totales se estaría hablando de montos de hasta 464 millones para una vivienda usada tipo apartamento de 64 m^2 en comparación a otra por 119 millones con las mismas características.

En este punto se hace evidente la heterogeneidad en la información sobre los valores de venta, y el estado de estos precios en estas zonas es clara evidencia del apetito voraz por parte de las constructoras y las inmobiliarias en ofrecer estilos de vida por encima del promedio de la capacidad adquisitiva de la mayoría de la población colombiana, que sólo algunos pueden alcanzar.

CUADRO 3.5

Listado de localidades con menores precios por tipo de vivienda y grupo m2

Tipo de vivienda	Grupo m2	Localidad	Price m2
Apartamento	Grupo 1 - $[0, 40)$	San Antonio de Prado	$2,655,948
	Grupo 2 - $[40, 60)$	Robledo La Aurora	$2,075,769
	Grupo 3 - $[60, 80)$	Castilla	$1,872,129
	Grupo 4 - $[80, 100)$	Campo Valdes	$2,066,019
	Grupo 5 - $[100, 120)$	San Javier	$1,552,933
	Grupo 6 - $[120, 160)$	Manrique	$1,586,665
	Grupo 7 - $[160, 200)$	Prado	$1,614,275
Casa	Grupo 2 - $[40, 60)$	Enciso	$1,759,518
	Grupo 3 - $[60, 80)$	Manrique	$1,872,258
	Grupo 4 - $[80, 100)$	Belen Rincon	$1,562,762
	Grupo 5 - $[100, 120)$	San Javier	$1,696,815
	Grupo 6 - $[120, 160)$	Castilla	$1,373,757
	Grupo 7 - $[160, 200)$	Manrique	$1,378,388

Nota. Cada grupo detalla los rangos en m^2 de los inmuebles que agrupan.

CUADRO 3.6

Listado de localidades con mayores precios por tipo de vivienda y grupo m2

Tipo de vivienda	Grupo m2	Localidad	Price m2
Apartamento	Grupo 1 - $[0, 40)$	El Poblado	$6,202,170
	Grupo 2 - $[40, 60)$	Ciudad del Rio	$7,023,008
	Grupo 3 - $[60, 80)$	Milla de Oro	$7,256,233
	Grupo 4 - $[80, 100)$	Milla de Oro	$6,567,417
	Grupo 5 - $[100, 120)$	La Calera	$6,645,841
	Grupo 6 - $[120, 160)$	Ciudad del Rio	$6,111,776
	Grupo 7 - $[160, 200)$	La Calera	$5,703,487
Casa	Grupo 2 - $[40, 60)$	Guayabal	$3,832,685
	Grupo 3 - $[60, 80)$	Santa Lucia	$4,296,171
	Grupo 4 - $[80, 100)$	La Mota	$4,191,853
	Grupo 5 - $[100, 120)$	La Castellana	$4,214,490
	Grupo 6 - $[120, 160)$	San Lucas	$5,552,629
	Grupo 7 - $[160, 200)$	San Lucas	$5,281,836

Nota. Cada grupo detalla los rangos en m^2 de los inmuebles que agrupan.

La burbuja inmobiliaria presente en las zonas del "top" 5, visible inclusive bajo la distribución de los datos bajo estudio (Ver Figura 3.12), puede ser muy peligrosa para el mercado total, principalmente debido al componente especulatorio e inflacional de la oferta, que fuerza al comprador acreer que los costos están aumentando a causa de una escasa oferta, cuando en realidad el resto de la ciudad se está desarrollando. Adicionalmente, el radio de dicha burbuja puede alcanzar las zonas aledañas y poco a poco incidir en el sobre-aumento de sus costos de vida y de propiedad.

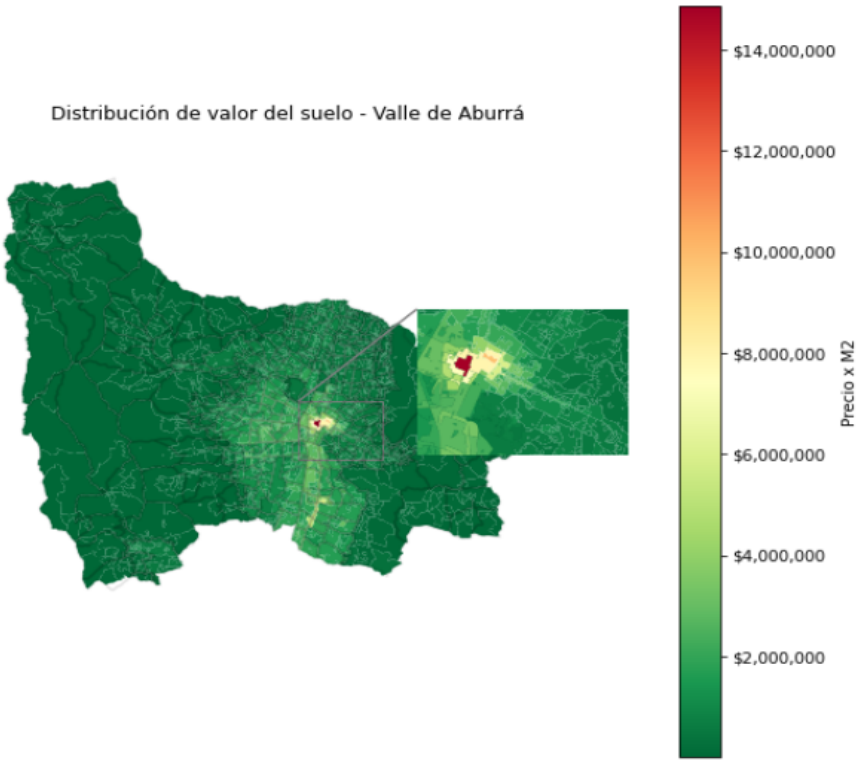


FIGURA 3.12: Mapa de distribución de precios de venta evidenciando burbuja en el sector noroccidental, elaboración propia

Esta situación, puede, sin embargo, verse también como favorable para los inversores de pequeña y mediana capacidad. La dinámica en las zonas menos "exclusivas" tiende a ser mayor, lo que incentiva a la rotación de las propiedades, es decir, a mayor transacciones y mejores ventanas de precios y retorno, además que son inmuebles que raramente se encuentran desocupados, generando un flujo de caja constante al dueño. Así pues, la oportunidad se encuentra en ubicar propiedades con características deseadas (por área total, estado o ubicación) que puedan reformarse para re-lanzarlas al mercado con precios más rentables.

3.4.1. Transformación de variable dependiente

La magnitud de la variable de salida, tal como se demuestra en la Figura 3.13, puede ser inconveniente para el diseño de un algoritmo regresivo que capture posibles relaciones lineales. Siguiendo luego la ecuación 2.2 se opta por tomar la transformación de escala bajo el logaritmo natural del precio por metro cuadrado (*Price m2*) como variable dependiente.

Es de mencionar que esta transformación es usual en la búsqueda de una ecuación lineal que describa el comportamiento de la variable de precio en cuanto a las características de los inmuebles.

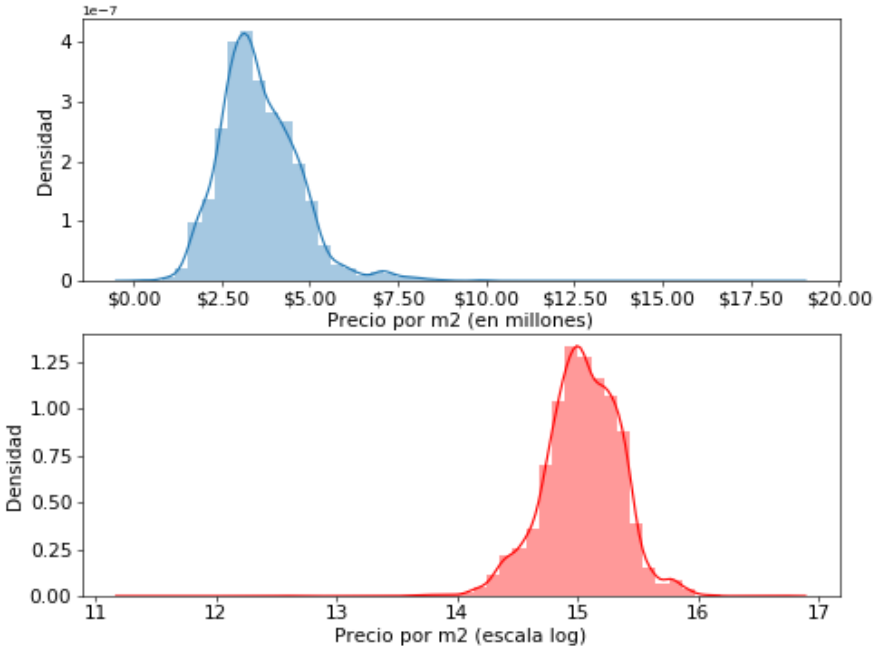


FIGURA 3.13: Distribución de variable precio m2 en escala original y transformada, elaboración propia

3.4.2. Correlaciones

El análisis de relaciones entre las variables al alcance se ejecuta tomando el coeficiente de correlación de *Pearson* para las variables continuas y el coeficiente de correlación biserial puntual para las variables binarias. Previamente se ha hecho una codificación a las variables categóricas, de modo que se extiende el número de variables por los respectivos rangos de cada variable, así, por ejemplo, la variable *Category 1* da nacimiento a dos variables *Category 1 - Apartamento* y *Category 1 - Casa*.

En la Figura 3.14 se presenta un resumen visual de los coeficientes calculados, omitiendo inicialmente las variables asociadas a la localidad por motivos de tamaño.

De esta matriz se puede observar que las variables que afectan principalmente al precio (variable transformada) son el estrato, el tipo de vivienda y la cantidad de habitaciones. Por razones naturales la variable área se encuentra altamente relacionada con los grupos por extensión total y con aquellas asociadas a las características estructurales de la propiedad. Curiosamente los variables correspondientes a los grupos de edad de construcción, todas de tipo booleano, parecen no tener efectos suficientes sobre el precio por metro cuadrado.

El Cuadro 3.7 resume las correlaciones más relevantes a nivel de variables de localidad. Los coeficientes obtenidos desafortunadamente no son significativos, aunque se espera que su efecto combinado refuerce la estimación final.

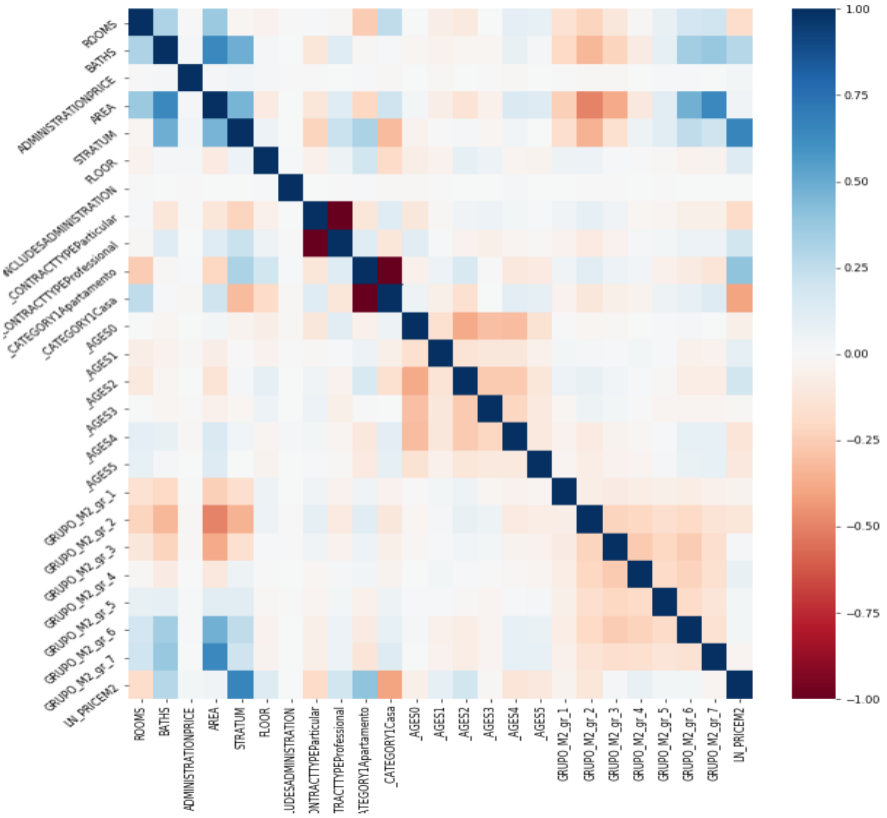


FIGURA 3.14: Matriz de correlaciones entre cada par de variables, elaboración propia

CUADRO 3.7

Correlación asociada a variables de ubicación

Variable	Correlación
Mean (cercano) Gr. 7 Venta	0,170
Mean (cercano) Gr. 5 Venta	0,157
Mean (cercano) Gr. 3 Venta	0,151
Mean (cercano) Gr. 1 Venta	0,129
Mean (cercano) Venta	0,126
Mean (cercano) Gr. 7 Arriendo	−0,081
Mean (cercano) Gr. 2 Arriendo	−0,089
Mean (cercano) Arriendo	−0,125
Mean (cercano) Gr. 1 Arriendo	−0,146
Mean (cercano) Gr. 6 Arriendo	−0,177

Nota. "Mean" hace referencia a la media aritmética.

Capítulo 4

Resultados y discusión

4.1. Aplicación

Posterior al proceso de subselección de la información y de procesamiento de las variables asociadas a las propiedades, se procede a ejecutar un modelo regresivo para estimar los precios de los inmuebles en el conjunto de datos final.

En este punto es importante mencionar que, específicamente las variables *Area* y *Administration price* han sufrido transformaciones. Por un lado, ambas variables han sido normalizadas entre 0 y 1, esto debido a lo inconveniente de su escala original. Por otro lado, el *Area* fue, adicionalmente, estandarizada, de modo que queda con media cero y varianza unitaria. Este proceder es común en las aplicaciones para el mercado de bienes raíces.

Siguiendo la metodología expuesta en Rousseeuw y Hubert (2017) y basándose en lo expuesto en la Sección 2.2.2, es pertinente aplicar métodos estadísticos robustos de eliminación de puntos anómalos en conjunto con el modelo de regresión. La idea general consiste entonces en eliminar, dentro de una cantidad finita y justa de iteraciones, las observaciones que produzcan los mayores residuales para así lograr un mejor ajuste global.

Para explorar luego el efecto de la eliminación de "outliers" se lanzan varios modelos de regresión siguiendo diferentes direcciones de selección sobre los datos. La expresión que han de seguir todos los modelos es como sigue

$$
\begin{aligned}
LN\_PRICEM2 \sim & ROOMS + BATHS + ADMINISTRATIONPRICE + STRATUM + \\
& FLOOR + INCLUDESADMINISTRATION + \\
& CONTRACTTYPE\_Particular + CONTRACTTYPE\_Professional + \\
& CATEGORY1\_Apartamento + CATEGORY1\_Casa + AREA\_sq + \\
& AGES\_0 + AGES\_1 + AGES\_2 + AGES\_3 + AGES\_4 + AGES\_5 + \\
& GRUPO\_M2\_gr\_1 + GRUPO\_M2\_gr\_2 + GRUPO\_M2\_gr\_3 + \\
& GRUPO\_M2\_gr\_4 + GRUPO\_M2\_gr\_5 + GRUPO\_M2\_gr\_6 + \\
& GRUPO\_M2\_gr\_7 + \\
& mean\_related\_res\_arre + mean\_related\_res\_vent + \\
& mean\_mpio\_res\_arre + mean\_mpio\_res\_vent + \\
& mean\_gr\_1\_res\_arre + mean\_gr\_1\_res\_vent + \\
& \dots \\
& -1
\end{aligned}
\tag{4.1}
$$

Inicialmente se genera una estimación con el método de regresión típico tomando todo el aglomerado de datos como base de comparación. Como alternativas se tienen luego:

(I) **Grupos** : Segmentando por los niveles de la variable *Grupo m2*.

(II) **LST** : Aplicando el método de eliminación de mayores residuales, con un 5 % de la población total.

(III) **Grupos** y **LST** : Eliminando el 5 % de casa subpoblación obtenida segmentando como en (**I**).

(IV) **Librería** : Aplicando una regresión robusta desde la librería `statsmodels`.

Sobre el método **IV**, la librería ha sido instanciada utilizando un estimador tipo *Huber* con un valor $c = 1{,}345$ para el parámetro `M` (norma). Para más detalle se consultar la documentación técnica del método en Statsmodels (*Documentación de modulo statsmodel en Python*) y su justificación teórica en Wang y col. (2007).

La Tabla 4.1 muestra los valores de la métrica de evaluación (MAPE) en compañía de algunas otras usuales. El conjunto total de datos ha sido particionado en dos muestras, una para entrenamiento con el 90 % de la información total (N : 49,482 observaciones).

CUADRO 4.1

Resultados de las regresiones mediante varios acercamientos para variable $ln(Price\ m2)$

	Métodos				
	Base	Grupos (**I**)	LST (**II**)	Grupos + LST (**III**)	Librería (**IV**)
Observaciones			49,482		
MAPE	1,12	1,19	1,13	1,01	1,11
RMSE	0,23	0,97	0,23	0,20	0,23
$\mathbf{R^2}$	0,59	0,66	0,68	0,74	0,59
Adjusted $\mathbf{R^2}$	0,59	0,65	0,68	0,74	0,59

La oscilación de la medida de R^2 va entre el 59 % y el 74 %. Para los casos base y **IV** se tienen evaluaciones similares y sus coeficientes no son lo suficientemente altos para asegurar estimaciones confiables. En cuanto al método I sus resultados son eclipsados por el efecto de la eliminación de observaciones con mayores residuales (**II**), que en combinación alcanzan el máximo rendimiento en todas las métricas. Así pues, la metodología **III** genera las estimaciones más certeras y tiene la ventaja de haberse creado con componentes robustos.

Con el respectivo modelo ya seleccionado es posible retornar al conjunto de datos validación y enfocarse en identificar aquellas observaciones cuyo valor real sea considerablemente menor al arrojado por la estimación. En la Figura 4.1 se ha desplegado el resultado de una simple regresión lineal entre los valores reales y estimados, de modo que el interés se centra en aquellos puntos por encima de la línea de color rojo, es decir, en aquellos cuya relación de error $estimado - real$ sea mayor.

La definición de oportunidad de inversión se complementa analizando el comportamiento del entorno (variables de localidad), de modo que sea posible discernir entre los efectos por error del modelo de regresión y las características del inmueble que lo particularizan entre los resultados.

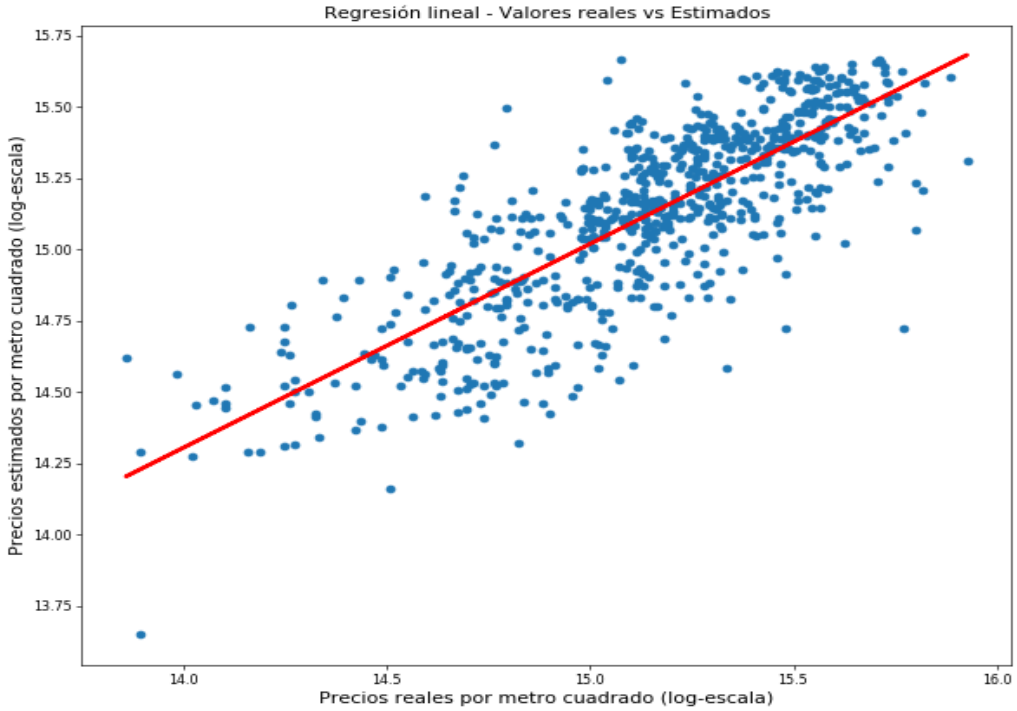


FIGURA 4.1: Regresión lineal simple entre valores reales y estimados, elaboración propia

La Tabla 4.2 resume la información más relevante de las 3 propiedades que mayor diferencia presentan, con un rango entre 2,5 y 2,8 millones de pesos colombianos. Analizando luego las características de los inmuebles se trata de apartamentos con alrededor de 112 metros cuadrados (pertenecientes al Grupo $m2$ 5) con una cantidad justa de habitaciones y baños, de buen estrato socio-económico, ubicados todos en zonas distintas de la ciudad. En general, las propiedades de todos los inmuebles son suficientemente heterogéneas, lo que es una buena señal, ya que el modelo podría estar sesgado fuertemente por alguna de las variables. Adicionalmente, éstas comparten cercanía y el sector donde se ubican es uno de los más cotizados en la ciudad.

En cuanto a las variables de localidad nótese que, en particular para aquellas asociadas a la modalidad de venta, el valor de oferta de para cada una de las propiedades se encuentran por debajo del promedio de aquellas en la localidad para el respectivo grupo por metro cuadrado. Está última condición se mantiene igualmente a nivel de municipio y de toda la localidad. Finalmente, sólo en cuanto a los precios promedio de los apartamentos con extensión de superficie similar las 3 propiedades bajo análisis se encuentran a una distancia promedio de 900 mil pesos colombianos, lo que llevado a valor total implica una diferencia media de 90 millones de pesos. En cuanto a las variables por modalidad de arriendo se evidencian también condiciones favorables para identificar las viviendas como oportunidad, principalmente debido al costo de arriendo para inmuebles entre los 0 y 40 metros cuadrados, ya que un posible comprador podría reformar los apartamentos habilitando aparta-estudios y de allí generar rentabilidad estable a largo plazo.

CUADRO 4.2

Características de inmuebles estimados superior a su valor real

Variable / Propiedad	**1**	**2**	**3**
Valor original | real	$3,527,273	$2,666,667	$3,401,961
Estimación	$6,363,421	$5,374,100	$5,925,799
Diferencia	$2,836,148	$2,707,433	$2,523,838
Area	110	120	102
Rooms	3	3	3
Baths	4	3	2
Stratum	6	5	6
Floor	14	0	0
Includes administration	0	0	0
Contract Type	Profesional	Particular	Profesional
Category 1	Apartamento	Apartamento	Apartamento
Antigüedad de construcción	Grupo 2	Grupo 3	Grupo 1
Localidad	El Poblado	Envigado	El Tesoro
mean_related_res_vent	$4,122,862	$4,388,362	$4,122,862
mean_mpio_res_vent	$4,075,773	$4,075,773	$4,075,773
mean_gr_1_res_vent	$5,205,803	$5,205,803	$5,205,803
mean_gr_2_res_vent	$3,761,085	$3,761,085	$3,761,085
mean_gr_3_res_vent	$4,061,514	$4,061,514	$4,061,514
mean_gr_4_res_vent	$4,192,055	$4,192,055	$4,192,055
mean_gr_5_res_vent	$4,091,866	$4,091,866	$4,091,866
mean_gr_6_res_vent	$4,077,669	$4,077,669	$4,077,669
mean_gr_7_res_vent	$3,959,543	$3,959,543	$3,959,543
mean_related_gr_1_res_vent	$5,769,801	$8,780,869	$5,769,801
mean_related_gr_2_res_vent	$3,629,864	$3,870,825	$3,629,864
mean_related_gr_3_res_vent	$4,181,323	$4,263,800	$4,181,323
mean_related_gr_4_res_vent	$3,696,863	$4,558,796	$3,696,863
mean_related_gr_5_res_vent	$4,352,195	$4,493,666	$4,352,195
mean_related_gr_6_res_vent	$3,793,193	$4,464,993	$3,793,193
mean_related_gr_7_res_vent	$4,307,467	$4,343,176	$4,307,467
mean_related_res_arre	$20,726	$27,657	$20,726
mean_mpio_res_arre	$32,881	$32,881	$32,881
mean_gr_1_res_arre	$85,744	$85,744	$85,744
mean_gr_2_res_arre	$26,993	$26,993	$26,993
mean_gr_3_res_arre	$26,687	$26,687	$26,687
mean_gr_4_res_arre	$27,394	$27,394	$27,394
mean_gr_5_res_arre	$25,984	$25,984	$25,984
mean_gr_6_res_arre	$30,738	$30,738	$30,738
mean_gr_7_res_arre	$25,979	$25,979	$25,979
mean_related_gr_1_res_arre	$23,052	$104,564	$23,052
mean_related_gr_2_res_arre	$17,561	$25,204	$17,561
mean_related_gr_3_res_arre	$21,672	$23,526	$21,672
mean_related_gr_4_res_arre	$24,178	$25,991	$24,178
mean_related_gr_5_res_arre	$24,190	$22,263	$24,190
mean_related_gr_6_res_arre	$19,996	$21,782	$19,996
mean_related_gr_7_res_arre	$21,655	$26,848	$21,655

En conclusión, producto de la aplicación de la metodología propuesta en esta tesis y considerando todos los resultados y análisis anteriores, es posible catalogar estos inmuebles como potenciales oportunidades de inversión.

La contribución de este trabajo es clara al llevar a la práctica un método analítico novedoso cuyas bases teóricas en conjunción con contexto de mercado permiten procesar los datos inteligentemente y obtener de ellos información de primera mano para proceder a tomar acciones que generan beneficio. La mixtura entre la regresión hedónica y el método de cuadrados recortados asumen el rol principal en el análisis producido tanto en el sentido teórico como en el aplicado, ya que, como método robusto es suficientemente estable y fortalece la regresión orientada a la estimación del precio de los inmuebles, y así mismo, por su versatilidad y simplicidad, puede usarse para casos diferentes en el mismo ambiente inmobiliario.

4.2. Conclusiones

El mercado de bienes raíces presenta buenas condiciones para la inversión. Las características de los objetos transados en dicho mercado los presenta como inversiones estables y seguras con posibilidades de retornos considerables. Adicionalmente, las dinámicas de ciudad y país en cuanto a la demanda alimentan el escenario ideal para ajustar los precios de venta y arriendo a puntos de rentabilidad óptimos.

En este documento se ha explorado un acercamiento a la identificación de inmuebles como oportunidades de inversión en el sentido que los interesados pueden tomarlos y explotarlos de mejor forma en el mercado. Desde una perspectiva de precios, las propiedades detectadas se encuentran subvalorizadas para el entorno en el que se localizan, lo que resalta la importancia de tener en cuenta el comportamiento de la oferta en zonas aledañas para así estimar razonablemente el potencial en la inversión.

Si bien los métodos en los desarrollos de este proyecto ya han sido ampliamente estudiados y aplicados para la estimación de precios de venta de inmuebles, la estrategia construida en este trabajo es novedosa y no se encuentran registros de metodologías similares. Como soporte, la utilización de algoritmos de minería de datos para extraer y procesar masivamente la información, así como el uso de métodos basados en texto para las agrupaciones georreferenciadas, demuestra su potencial para tratar problemas de "big data". En el centro, el acercamiento robusto de Rousseeuw para la regresión hedónica es verdaderamente la clave para consolidar el paso a paso sobre como localizar las denominadas "oportunidades".

Los pasos a continuación deberían enfocarse en masificar y sistematizar la metodología expuesta, de modo que la identificación alcance niveles nacionales y sea entregada por un programa o plataforma de forma automática. Un sistema a gran escala tendría la capacidad de detectar propiedades tipo oportunidad en tiempo real y reportar alertas a los interesados, optimizando así el proceso completo. A largo plazo el sistema podría habilitarse como un servicio por demanda que algunas compañías puedan anexar a sus plataformas en pro de robustecer sus productos o portafolios.

Adicionalmente, aunque las métricas de evaluación alcanzaron niveles relativamente positivos, todavía hay un inventario extenso de métodos por explorar que, posiblemente, entreguen mejores resultados, y también permitan analizar el sistema desde otras perspectivas. Los métodos expuestos en el Estado del Arte (2.1) son el mejor ejemplo de todos los posibles acercamientos, tanto teóricos como prácticos,

para tratar el sistema inmobiliario y explotar la información obtenida en pro de un beneficio común. Asimismo, los pasos aplicados en la metodología presente deberían ser estudiados con más profundidad y alterados buscando resumir de mejor forma su efecto sobre las estimaciones e identificación de anomalías.

Como trabajo futuro se podría, inclusive, anexar variables de nivel macro-económico y demográfico tales como el índice de precio al consumidor, la estimación de las tasas de interés para vivienda, la cantidad de licencias de construcción habilitadas, los planes de ordenamiento territorial, las proyecciones de crecimiento poblacional, entre otras, al análisis, permitiendo así proyectar la valorización del mercado a períodos de mayor plazo y aportando en la determinación de los momentos óptimos en que podrían liberarse las propiedades al mercado.

4.3. Plan de gestión de datos

El plan de gestión de datos asociado a este proyecto contempla los siguientes puntos:

- Los datos de insumo se concentran en la información transaccional. Estos registros digitales se encuentran consolidados en tablas bajo un esquema de base de datos `SQL`, el cual se materializa por medio de archivos fuente particionados con extensión `sqlite`.
- Sobre los datos de avisos publicitarios no se conoce en Colombia, y no es común en el mundo, normas de formato o formatos acordados. Los detalles de cada aviso son puestos en la red por particulares y las páginas web no tienden a exigir niveles de estandarización sobre los mismos.
- El acceso a los datos está restringido a la aplicación de este proyecto y la propiedad intelectual de los mismos queda a cargo de los participantes del equipo. La reutilización de los datos podrá efectuarse estrictamente para aplicaciones académicas bajo estricta aprobación de los participantes del equipo y deberán ser entregados bajo agregaciones, sin disponer de las fuentes crudas.
- Los datos, al finalizar este proyecto, serán dispuestos en una fuente física externa y en un repositorio de datos con estrictos permisos de acceso a los integrantes del equipo.

4.4. Condiciones de uso

Los datos serán utilizados exclusivamente para el cumplimiento del objetivo de este proyecto. La información recolectada ha de ser tratada de forma que ningún dato de carácter personal sea incluido o tal que algún particular pueda ser afectado. Los beneficios se limitan a lograr construir una base robusta de componentes intelectuales y prácticos que puedan exponerse a la comunidad académica e interesados con miras de explotar la información obtenida y generar nuevos medios de interacción con el mercado beneficiando equitativamente a todos. Los datos serán usados sólo por el estudiante para este proyecto y cualquier resultado liberado al público ha de contener información agregada o podrá ser consultado directamente en la web, de forma que la base original no es exclusiva ni generada a causa del proyecto.

Adicionalmente, como se trata de avisos publicitarios y de referencias que se pueden encontrar en la web, no se requiere de consentimiento de propiedad y tampoco de

anonimización de los datos ya que los valores consultados son de carácter superficial.

4.5. Aspectos éticos

La metodología expuesta en este documento tiene el propósito último de ser utilizado como medio de apoyo en la toma de decisiones sobre el mercado inmobiliario. A largo plazo, en caso de ser masificado y automatizado, podría afectar a quienes se dedican a trabajar en las casas inmobiliarias y convertirse inclusive en una herramienta de acceso privilegiado. El primer componente no puede ser mitigado a menos que los especialistas en inmuebles se aprovechen también del desarrollo y lo hagan más robusto y especializado. Sobre el segundo, es imperativo resaltar la importancia de liberar la metodología a todo público interesado, de modo que todos tengan la oportunidad de participar en el mercado basándose en los datos.

Un componente adicional a tener en cuenta es que, como herramienta, el método construido no está blindado a posibles errores, y por tanto deben haber capas de validación posteriores. El desarrollo puede dar lugar a la falsa ilusión de que no es necesario a partir del momento hacer seguimiento al proceso o sistema bajo análisis, así pues los consumidores de los resultados deben ser precavidos y evitar disparar cadenas de eventos que puedan resultar perjudiciales para el mercado o los particulares.

Bibliografía

Afonso, Bruno y col. (2019). «Housing Prices Prediction with a Deep Learning and Random Forest Ensemble». En: *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. Salvador: SBC, págs. 389-400. DOI: 10.5753/eniac.2019.9300. URL: https://sol.sbc.org.br/index.php/eniac/article/view/9300.

Alpaydin, Ethem (2010). *Introduction to Machine Learning*. 2nd. The MIT Press. ISBN: 026201243X.

Arquitectura & Concreto. *Manual de Inversión:Valor del metro cuadrado en Medellín*. URL: https://arquitecturayconcreto.com/blog/valor-del-metro-cuadrado-en-medellin/.

Avila, Roman. *Cómo funcionan los modelos que están transformando el negocio inmobiliario*. URL: https://andavip.medium.com.

Bailey, Martin J., Richard F. Muth y Hugh O. Nourse (1963). «A Regression Method for Real Estate Price Index Construction». En: *Journal of the American Statistical Association* 58.304, págs. 933-942. ISSN: 01621459. URL: http://www.jstor.org/stable/2283324.

Baldominos, Alejandro y col. (nov. de 2018). «Identifying Real Estate Opportunities Using Machine Learning». En: *Applied Sciences* 8, pág. 2321. DOI: 10.3390/app8112321.

Bourassa, Steven C., Martin Hoesli y Vincent S. Peng (2003). «Do housing submarkets really matter?» En: *Journal of Housing Economics* 12.1, págs. 12-28. ISSN: 1051-1377. DOI: https://doi.org/10.1016/S1051-1377(03)00003-2. URL: https://www.sciencedirect.com/science/article/pii/S1051137703000032.

CAMACOL. *Construcción en cifras*. URL: https://camacol.co/documentos/construccion-en-cifras.

Case, Karl E y Robert J Shiller (sep. de 1987). *Prices of Single Family Homes Since 1970: New Indexes for Four Cities*. Working Paper 2393. National Bureau of Economic Research. DOI: 10.3386/w2393. URL: http://www.nber.org/papers/w2393.

Cho, Man (mayo de 1996). «House Price Dynamics: A Survey of Theoretical and Empirical Issues». En: *Journal of Housing Research* 7.

Cubeddu, Luis, Camilo Tovar-Mora y Evridiki Tsounta. «Latin America: Vulnerabilities under construction?» En: *International Monetary Fund*. URL: https://www.imf.org/en/Publications/WP/Issues/2016/12/31/Latin-America-Vulnerabilities-Under-Construction-26130.

DANE. *Censo de Población y Vivienda 2018*. URL: http://www.dane.gov.co.

– *Producto Interno Bruto (PIB) Históricos*. URL: http://www.dane.gov.co.

Dombrow, Jonathan, J. R. Knight y C. F. Sirmans (ene. de 1997). «Aggregation Bias in Repeat-Sales Indices». En: *The Journal of Real Estate Finance and Economics* 14.1, págs. 75-88. ISSN: 1573-045X. DOI: 10.1023/A:1007720001268. URL: https://doi.org/10.1023/A:1007720001268.

Durrani, Mohammad Haseeb. *Real Estate Investment: Buy to Sell or Buy to Rent?* URL: https://nycdatascience.com/blog/student-works/zillow/.

Economipedia. *Arbitraje financiero*. URL: https://economipedia.com/definiciones/arbitraje.html.

Fik, Tim, David Ling y Gordon Mulligan (feb. de 2003). «Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach». En: *Real Estate Economics* 31, págs. 623-646. DOI: 10.1046/j.1080-8620.2003.00079.x.

Fincaraiz (2020). *Apartamento en venta*. [Online; accessed September 21, 2020]. URL: https://www.fincaraiz.com.co/apartamentos/venta/medellin.

Ghorbani, Hamid (oct. de 2019). «Mahalanobis distance and its application for detecting multivariate outliers». En: *Facta Universitatis Series Mathematics and Informatics* 34, pág. 583. DOI: 10.22190/FUMI1903583G.

Goh, Yen Min, Greg Costello y Greg Schwann (2012). «Accuracy and Robustness of House Price Index Methods». En: *Housing Studies* 27.5, págs. 643-666. DOI: 10.1080/02673037.2012.697551. eprint: https://doi.org/10.1080/02673037.2012.697551. URL: https://doi.org/10.1080/02673037.2012.697551.

Haan, Jan de y Erwin Diewert (2013). *Hedonic Regression Methods*. DOI: https://doi.org/https://doi.org/10.1787/9789264197183-7-en. URL: https://www.oecd-ilibrary.org/content/component/9789264197183-7-en.

Heene, Moritz y col. (2014). «Crisis in cognitive science? Rise of the undead theories». En: *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Society*. Cognitive Science Society, págs. 82-83. URL: http://mural.maynoothuniversity.ie/12627/.

Instituto Nacional de Contadora Públicos. *Proptech: el avance tecnológico para el sector inmobiliario en América Latina y el Caribe*. URL: https://incp.org.co/proptech-avance-tecnologico-sector-inmobiliario-america-latina-caribe/.

Kain, John F. y John M. Quigley (1970). «Measuring the Value of Housing Quality». En: *Journal of the American Statistical Association* 65.330, págs. 532-548. ISSN: 01621459. URL: http://www.jstor.org/stable/2284565.

Kim, Wonchan y col. *Which house shall we invest in?* URL: https://nycdatascience.com/blog/student-works/which-house-shall-we-invest-in/.

Lecamus, Vincent. *PropTech: What is it and how to address the new wave of real estate startups?* URL: https://medium.com.

Li, Shengwen y col. (sep. de 2017). «Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective». En: *Applied Spatial Analysis and Policy* 10.3, págs. 421-433. ISSN: 1874-4621. DOI: 10.1007/s12061-016-9185-3. URL: https://doi.org/10.1007/s12061-016-9185-3.

Maguire, Phil y col. (oct. de 2016). «A robust house price index using sparse and frugal data». En: *Journal of Property Research* 33, págs. 293-308. DOI: 10.1080/09599916.2016.1258718.

Marjan, Čeh y col. (mayo de 2018). «Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments». En: *ISPRS International Journal of Geo-Information* 7, pág. 168. DOI: 10.3390/ijgi7050168.

Park, Byeonghwa y Jae Kwon Bae (2015). «Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data». En: *Expert Systems with Applications* 42.6, págs. 2928-2934. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.11.040. URL: https://www.sciencedirect.com/science/article/pii/S0957417414007325.

Peña, Daniel (2002). *Análisis multivariante de datos*. McGraw-Hill Interamericana de España S.L. ISBN: 9788448136109. URL: https://books.google.com.co/books?id=TrVlAAAACAAJ.

Przytuła, Paweł. *Are you buying an apartment? How to hack competition in the real estate market*. URL: https://www.kdnuggets.com/2018/10/apartment-hack-competition-real-estate-market.html.

Pérez Rave, Jorge (mayo de 2019). «Statihouse: desarrollo tecnológico basado en Ciencia de Datos para explorar estadísticamente el sector inmobiliario». En: *Ingeniare* 27, págs. 113-130. DOI: 10.4067/S0718-33052019000100113.

Pérez Rave, Jorge, Juan Correa y Favián Echavarría (mar. de 2019). «A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes». En: *Journal of Property Research* 36, págs. 59-96. DOI: 10.1080/09599916.2019.1587489.

Rousseeuw, Peter J. (1984). «Least Median of Squares Regression». En: *Journal of the American Statistical Association* 79.388, págs. 871-880. DOI: 10.1080/01621459.1984.10477105.

Rousseeuw, Peter J. y Mia Hubert (2017). «Anomaly detection by robust statistics». En: *WIREs Data Mining and Knowledge Discovery* 8.2. ISSN: 1942-4795. DOI: 10.1002/widm.1236. URL: http://dx.doi.org/10.1002/widm.1236.

Rubio, Jeisson, Francisco Guzmán y Jesús Otero (abr. de 2019). «Una base de datos de precios y características de vivienda en Colombia con información de Internet». En: *Revista de Economía del Rosario* 22, pág. 25. DOI: 10.12804/revistas.urosario.edu.co/economia/a.7768.

Salcedo-Perez, Carlos y col. (2020). «Economía informal en Colombia: iniciativas y propuestas para reducir su tamaño». En: *Espacios* 4.3, pág. 22.

Selim, Hasan (2009). «Determinants of house prices in Turkey: Hedonic regression versus artificial neural network». En: *Expert Systems with Applications* 36.2, Part 2, págs. 2843-2852. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.01.044. URL: https://www.sciencedirect.com/science/article/pii/S0957417408000596.

Shimizu, Chihiro y Kiyohiko Nishimura (ago. de 2010). «Housing Prices in Tokyo: A Comparison of Hedonic and Repeat Sales Measures». En: *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)* 230, págs. 792-813. DOI: 10.1515/jbnst-2010-0612.

Statsmodels. *Documentación de modulo statsmodel en Python*. URL: https://www.statsmodels.org/stable/rlm.html.

Superintendencia Financiera. *Saldos de cartera por producto*. URL: https://www.superfinanciera.gov.co/jsp/.

Wallace, Nancy E. y Richard A. Meese (ene. de 1997). «The Construction of Residential Housing Price Indices: A Comparison of Repeat-Sales, Hedonic-Regression, and Hybrid Approaches». En: *The Journal of Real Estate Finance and Economics* 14.1, págs. 51-73. ISSN: 1573-045X. DOI: 10.1023/A:1007715917198. URL: https://doi.org/10.1023/A:1007715917198.

Wang, You-Gan y col. (jun. de 2007). «Robust Estimation Using the Huber Function With a Data-Dependent Tuning Constant». En: *Journal of Computational and Graphical Statistics* 16, 468–481. DOI: 10.1198/106186007X180156.

Wikipedia, the free encyclopedia (2015). *Subregiones de Antioquia*. [Online; accessed January 15, 2021]. URL: https://commons.wikimedia.org/wiki/File:Mapa_de_Antioquia_(subdivisiones).svg.

– (2017). *Area Metropolitana de Medellin*. [Online; accessed January 15, 2021]. URL: https://commons.wikimedia.org/wiki/File:Mapa-Area_Metropolitana_de_Medellin.png.

Wing, Chau Kwong y T. Chin (jun. de 2003). «A Critical Review of Literature on the Hedonic Price Model». En: *International Journal for Housing Science and Its Applications* 27, págs. 145-165.

Zapata-Vega, José Luis (2013). *Informalidad:Factor de desconfianza en el sector inmobiliario*. Universidad del Rosario.

Zhu, Min. «Los mercados inmobiliarios, la estabilidad financiera y la economía». En: *International Monetary Fund*. URL: https://www.imf.org/es/News/Articles/2015/09/28/04/53/sp060514.