Predicción del precio de la vivienda mediante el uso de métodos estadísticos espaciales

Andrés Delgado Mosquera, Iván Carlos Barrio Herreros, David J. K. Tofan

Tutor: Javier Poole Pérez-Palencia
Cotutor: Miguel de la Llave Montiel

Trabajo Fin de Máster presentado en la Escuela IBM & Universidad Europea, para la obtención del Título de Máster en Business Analytics.

Curso 2019-2020

Madrid, 16 de julio de 2020

Resumen

El mercado de la vivienda ofrece grandes oportunidades a quienes saben aprovecharlas. En este documento proponemos un estudio estadístico espacial del precio de la vivienda orientado tanto a obtener una predicción precisa como a que dicha predicción se sustente en un modelo estadístico sólido y estable a lo largo del espacio, garantizando unos resultados aceptables incluso cuando se trate de predecir en base a los datos de una nueva vivienda, ajena totalmente a los utilizados en la fase de modelizado.

Gracias a estas técnicas, pretendemos ofrecer un punto de vista diferente, basando nuestro estudio en la influencia que sufre el precio de una vivienda a causa del valor de sus viviendas vecinas, y planteando así un nuevo marco de trabajo e investigación poco explorado en comparación con otros modelos estadísticos más tradicionales.

Abstract

The housing market offers great opportunities to those who know how to take advantage of them. In this document we propose a spatial statistical study of the price of housing aimed both at obtaining an accurate prediction and at ensuring that said prediction is based on a solid and stable statistical model throughout the space; guaranteeing acceptable results even when trying to predict based on the data of a new house, totally unrelated to those used in the modeling phase.

Thanks to these techniques, we intend to offer a different point of view, basing our study on the influence that the price of a house suffers due to the value of its neighboring homes, and thus proposing a new framework of work and research little explored compared to other more traditional statistical models used for this topic.

Índice

1. Introducción.

¡Nuestra vida ya cambió y ahora debemos acostumbrarnos a ella!

– Anónimo

Queremos comenzar la introducción de este trabajo de fin de máster haciendo alusión a la cita mencionada, ya que debido a la pandemia generada por el Covid-19 el comportamiento económico de todas las industrias ha cambiado, y como en la guerra, hay industrias que se están beneficiando y otras que han tenido contracciones catastróficas. Pero ¿qué pasará con estas industrias una vez que se logre controlar la curva de contagios? ¿Y cómo se comportarán las industrias después de la pandemia?

La industria inmobiliaria aporta un 5,6% del PIB (2017) y emplea a más del 0,6% de la población económicamente activa[1] [2], es por lo tanto un sector clave y es de suma importancia conocer cómo predecir los precios de forma acertada.

La asociación española de Personal Shopper Inmobiliario (AEPSI) realizó un análisis del impacto que tendrá la Covid-19 en el mercado inmobiliario español. Prevén que la industria se contraiga en un 10% a nivel nacional y en ciudades como Madrid entre 5% y 7%[3].

Iñaki Unsain, presidente de AEPSI, afirma que *"la variación del precio de la vivienda depende, principalmente de la evolución del empleo, el PIB y los tipos de interés. Cuando estas variables se mueven a la baja, el precio de la vivienda desciende. Esta variación final del precio será mayor o menor en función del plazo de recuperación. Cuanto más se alargue el paro de la economía, mayor será la caída del precio"*.[4]

AEPSI también dice que, debido a la crisis, el mercado inmobiliario ha pasado a estar dominado por aquellas personas con liquidez ya sean nacionales o extranjeros. Son inversores ambiciosos que buscan oportunidades en este contexto.

[1] Statista Research Department. (2020, Enero 10). Industria de la construcción: aportación al PIB 2005-2017. Recuperado Mayo 9, 2020, de https://es.statista.com/estadisticas/549605/aportacion-del-sector-de-la-construccion-al-pib-en-espana/
[2] Statista Research Department. (2020, Enero 10). Actividades inmobiliarias: número de trabajadores 2005-2016. Recuperado Mayo 9, 2020, de https://es.statista.com/estadisticas/526311/numero-de-empleados-del-sector-inmobiliarias-en-espana/
[3] Inmodiario. (2020, Marzo 31). El precio de la vivienda caerá un 10% en España por la crisis del Covid-19. Recuperado Mayo 9, 2020, de https://www.inmodiario.com/96/28847/precio-vivienda-caera-espana-crisis-covid.html
[4] Inmodiario. (2020, Marzo 31). El precio de la vivienda caerá un 10% en España por la crisis del Covid-19. Recuperado Mayo 9, 2020, de https://www.inmodiario.com/96/28847/precio-vivienda-caera-espana-crisis-covid.html

De aquí, nace la necesidad de proporcionar al mercado un modelo estadístico que englobe aquellas variables que influyan directamente en el precio del metro cuadrado, con el cual puedan evaluar si el bien inmueble es una oportunidad de compra. Centraremos la versión piloto de nuestro análisis en la ciudad de Madrid.

Para lo cual, nos planteamos la siguiente duda:

- ¿Cuáles son aquellas variables que influyen directamente en el precio de los inmuebles de Madrid? ¿Nos ayudan a predecir un valor cercano a la realidad? ¿Influye la situación geográfica?

2. Objetivos del TFM.

Conforme a la duda planteada nos hemos fijado los siguientes objetivos en este trabajo de fin de máster:

1) Encontrar aquellas variables que tengan mayor influencia sobre el precio de la vivienda en la ciudad de Madrid.
2) Encontrar un modelo estadístico que nos ayude a predecir el precio de la vivienda de la ciudad de Madrid.
3) Definir si el precio de la vivienda de Madrid presenta dependencia espacial con los vecinos.

3. Motivación del TFM, objetivos y alcance

Contexto.

La situación en España.

España se ha beneficiado de una recuperación económica relativamente equilibrada en los últimos años, con un crecimiento medio del PIB del 2,8% desde 2014.[5] Pero ese crecimiento ha desacelerado en 2019 a medida que el crecimiento del consumo privado ha desacelerado.[6]

Con las turbulencias políticas que España ha vivido en los últimos años el gobierno ha sufrido una fuerte división;[7] por otro lado, varios países europeos experimentan con diferentes

[5] Expansión. (2020, Enero 31). PIB de España - Producto Interior Bruto 2019. Recuperado Marzo 27, 2020, de https://datosmacro.expansion.com/pib/espana

[6] Libre Mercado. (2019, Septiembre 30). España se desacelera: el PIB crece a su menor ritmo en tres años. Recuperado Marzo 28, 2020, de https://www.libremercado.com/2019-09-30/espana-se-desacelera-el-pib-crece-a-su-menor-ritmo-en-tres-anos-1276645488/

[7] EL PAÍS, (2019, Noviembre). Resultados Electorales en Total España: Elecciones Generales 2019. Recuperado Marzo 27, 2020, de https://resultados.elpais.com/elecciones/2019/generales/congreso/

movimientos políticos que provocan brechas y enfrentamientos entre diferentes grupos de la sociedad, con el consiguiente efecto económico.[8]

Y con la reciente aparición de la pandemia de enfermedad por COVID 19 (coronavirus), la incertidumbre sobre la situación económica en España y el futuro del país crece más. Ya que el país tiene una alta dependencia de los sectores de servicios, como por ejemplo el turismo, comercios, hostelería y el transporte.[9]

Ciudad de Madrid en cifras.

Madrid es la capital de España, así como su ciudad más poblada. El municipio de Madrid tiene una población de 3.223.334 de personas. Madrid es la comunidad autónoma con más crecimiento de población en España. Un crecimiento de 1,25% del año 2017 al 2018.[10] Ese crecimiento notable de población debería de nutrir a la demanda de la compra de viviendas en el municipio, relacionado con el fenómeno de "España vacía", ya que existe una alta demanda en ciudades como Madrid y una pérdida de población en ciudades más pequeñas.[11] En varios análisis de la compra-venta de viviendas se ha identificado que los precios de demanda de la vivienda han tocado máximos históricos en los últimos dos años.[12] *"Madrid [...] gana población y colapsan el mercado de la vivienda"*. [13]

Según la Encuesta Continua de Hogares del Ayuntamiento de Madrid, actualmente (2018) existen en total 1.347.909 hogares documentados en la ciudad de Madrid. De esa cantidad total, 48.387 tienen menos de 3 habitaciones, 1.216.666 tienen entre 3 y 6 habitaciones, y 82.855 tienen 7 o más habitaciones.[14]

---

[8] Bonet, E. (2019, Septiembre 6). "Si la Unión Europea no cambia de rumbo, la ultraderecha seguirá creciendo". Recuperado Marzo 30, 2020, de https://www.publico.es/politica/auge-ultraderecha-union-europea-no-cambia-rumbo-ultraderecha-seguira-creciendo.html

[9] ESADE. (2020, Marzo 13). Política económica contra el coronavirus: impacto y respuestas para España. Recuperado Marzo 27, 2020, de https://dobetter.esade.edu/es/coronavirus-politica-economica

[10] Rodrigo, N. (2019, Enero 2). Madrid, el municipio que más creció en 2018, con 40.000 habitantes más. Recuperado Marzo 27, 2020, de https://cincodias.elpais.com/cincodias/2019/01/02/economia/1546432956_524849.html

[11] El Economista. (2019, Octubre 4). Así es la España vacía: 12 gráficos para entender el problema de la despoblación en nuestro país. Recuperado Abril 11, 2020, de https://www.eleconomista.es/economia/noticias/10120949/10/19/Asi-es-la-Espana-vacia-12-graficos-para-entender-el-problema-de-la-despoblacion-en-nuestro-pais.html

[12] Sanz, E. (2020, Febrero 23). La vivienda se enfría: los vendedores recortan hasta un 36% sus expectativas de precio. Recuperado Marzo 27, 2020, de https://www.elconfidencial.com/vivienda/2020-02-23/vivienda-precios-mercado-residencial-maximos_2461336/

[13] Salvador, R. (2019, Diciembre 16). Madrid y Barcelona ganan población y colapsan el mercado de la vivienda. Recuperado Abril 10, 2020, de https://www.lavanguardia.com/economia/20191215/472233133386/precios-vivienda-barcelona-madrid-crecimiento-poblacion-venta-alquiler.html

[14] Ayuntamiento de Madrid. (2019, Junio 27). Encuesta Continua de Hogares – Hogares.xlsx (1.3). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Demografia-y-poblacion/Cifras-de-poblacion/Encuesta-Continua-de-Hogares/?vgnextfmt=default&vgnextoid=0ccf7bfbd989b610VgnVCM2000001f4a900aRCRD&vgnextchannel=a4eba53620e1a210VgnVCM1000000b205a0aRCRD

El régimen de tenencia de 625.291 viviendas es por compra propia, totalmente pagada, heredada o donada, 320.010 son propias con pagos pendientes, 353.730 son alquiladas, y 48.877 son cedidas gratis o bajo precio por otro hogar, o por empresa.[15]

120.470 de los hogares tienen menos de 46 m2 de superficie útil de la vivienda, 597.194 tienen entre 46 y 75 m2, 428.375 tienen entre 76 y 105 m2, 145.167 tienen entre 106 y 150 m2, y 56.701 tienen más de 150 m2.[16]

El número de operaciones de compra-venta de vivienda en el año 2018 en Madrid ha sido 38.330. De las cuales la mayoría (9.718) eran de tamaños entre 40 y 60 m2. Muchas de ellas (9.633) tenían entre 60 y 80 m2, y 9.286 eran de tamaño superior a 100 m2.[17]

Comportamiento del precio del metro cuadrado de la vivienda en Madrid en el Último Año

En el gráfico 1 vemos como ha sido el comportamiento del precio del metro cuadrado en la ciudad de Madrid de acuerdo con datos obtenidos de Idealista. Tiene una pendiente creciente con un rango de precios entre 3.212 euros a 3.229 euros.

[15] Ayuntamiento de Madrid. (2019, Junio 27). Encuesta Continua de Hogares – Hogares.xlsx (1.9). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Demografia-y-poblacion/Cifras-de-poblacion/Encuesta-Continua-de-Hogares/?vgnextfmt=default&vgnextoid=0ccf7bfbd989b610VgnVCM2000001f4a900aRCRD&vgnextchannel=a4eba53620e1a210VgnVCM1000000b205a0aRCRD

[16] Ayuntamiento de Madrid. (2019, Junio 27). Encuesta Continua de Hogares – Hogares.xlsx (1.4). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Demografia-y-poblacion/Cifras-de-poblacion/Encuesta-Continua-de-Hogares/?vgnextfmt=default&vgnextoid=0ccf7bfbd989b610VgnVCM2000001f4a900aRCRD&vgnextchannel=a4eba53620e1a210VgnVCM1000000b205a0aRCRD

[17] Ayuntamiento de Madrid. (2018, Abril 23). Distritos en cifras (Información de Distritos). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Distritos-en-cifras/Distritos-en-cifras-Informacion-de-Distritos-/?vgnextfmt=default&vgnextoid=74b33ece5284c310VgnVCM1000000b205a0aRCRD&vgnextchannel=27002d05cb71b310VgnVCM1000000b205a0aRCRD

Gráfico 1: Comportamiento del precio del metro cuadrado en Madrid.

Precio Promedio del M2 por Distrito de Madrid

En el gráfico 2 se observa el precio promedio por barrio del metro cuadrado en cada distrito en el año 2020. (Datos Idealista)

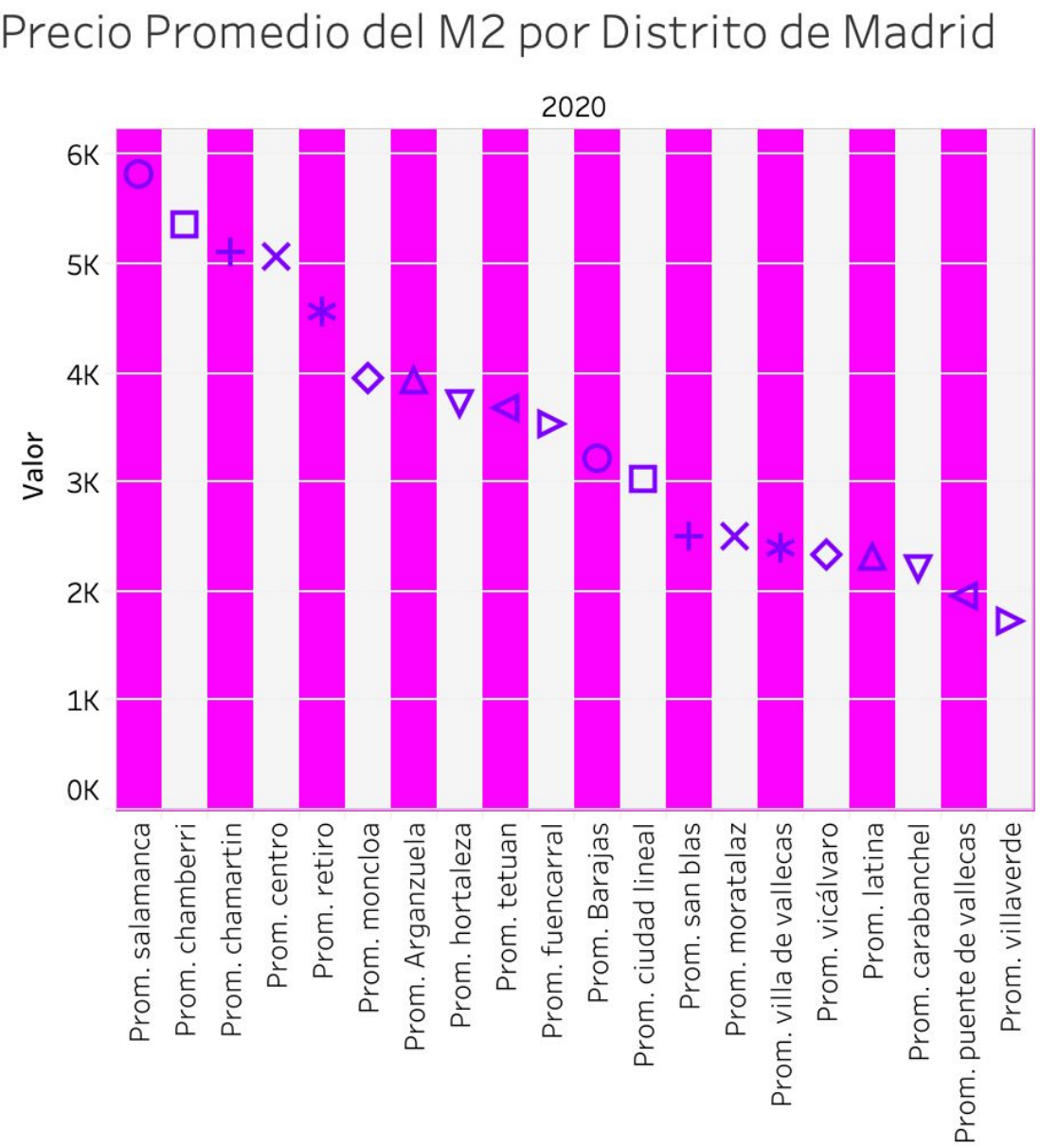


Gráfico 2: Precio Promedio por barrio del metro cuadrado en cada distrito

Existe una diferencia de 4.070 euros entre el distrito de Salamanca cuyo precio de m2 es más costoso frente al distrito de Villaverde. Entre un distrito y otro existe una diferencia superior al 300%

Porcentaje de Crecimiento del Valor M2 de Madrid Dic 2019-2021

En la tabla 1 vemos la previsión para diciembre de 2021 del precio por metro cuadrado en cada distrito. La previsión se realiza mediante la herramienta "Estimación" de Tableau, la cual maneja 8 diferentes modelos estadísticos y dependiendo del tipo de datos escoge los modelos a usarse. Luego el modelo que posea un mejor accuracy es el escogido para proveer la proyección estimada de datos. Los datos ingestados son del periodo 01/01/2010 - 01/03/2020. Los datos base se obtuvieron a través de Idealista y su previsión de Tableau.[18]

Además, se evalúa el porcentaje de crecimiento que tendrá el precio del metro cuadrado por distrito entre diciembre del 2019 y diciembre del 2021. La fluctuación del crecimiento irá desde un -24% en Arganzuela hasta un 23% en Vicálvaro.

Distrito	Precio m2 Dic 2019	Proyección m2 Dic. 2021	% de Crecimiento de Valor m2 Dic.2019-2021
Arganzuela	4.019	3.069	-24%
Barajas	3.229	3.485	8%
Carabanchel	2.215	2.262	2%
Centro	5.059	5.567	10%
Chamartín	5.070	5.841	15%
Chamberí	5.301	5.608	6%
Ciudad Lineal	3.035	2.802	-8%
Fuencarral-El Pardo	3.554	3.801	7%
Hortaleza	3.615	3.986	10%
Latina	2.330	2.356	1%
Moncloa-Aravaca	3.963	4.043	2%

[18] Idealista. (2020, Marzo). Evolución del precio de la vivienda en venta en Villaverde. Recuperado Abril 11, 2020, de https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/madrid-comunidad/madrid-provincia/madrid/villaverde/

Moratalaz	2.515	2.497	-1%
Puente de Vallecas	1.945	1.970	1%
Retiro	4.539	4.382	-3%
Salamanca	5.786	5.590	-3%
San Blas	2.505	2.326	-7%
Tetuán	3.656	3.435	-6%
Usera	2.443	2.141	-12%
Vicálvaro	2.318	2.842	23%
Villa de Vallecas	2.434	2.427	0%
Villaverde	1.719	1.865	8%

Tabla 1: Precio del metro cuadrado por distrito en diciembre 2021.

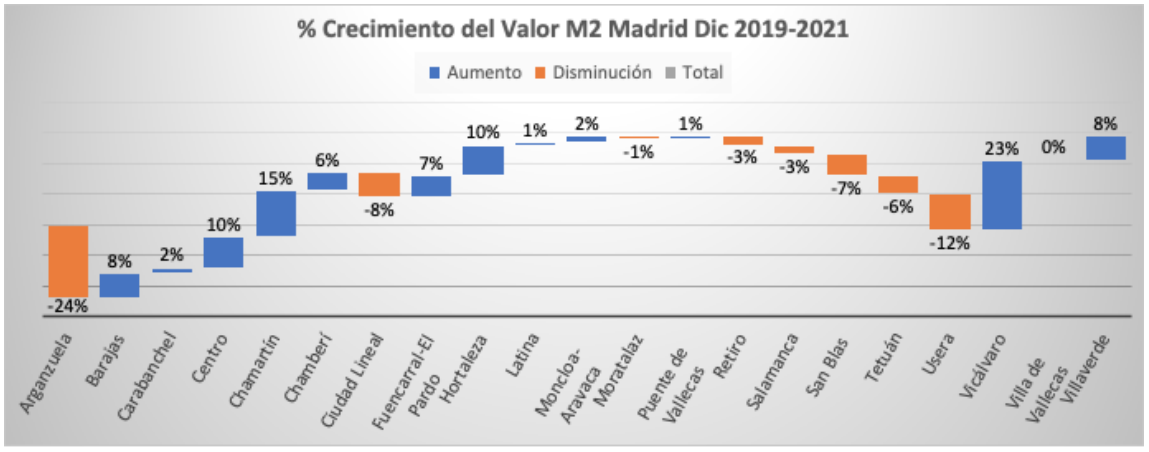
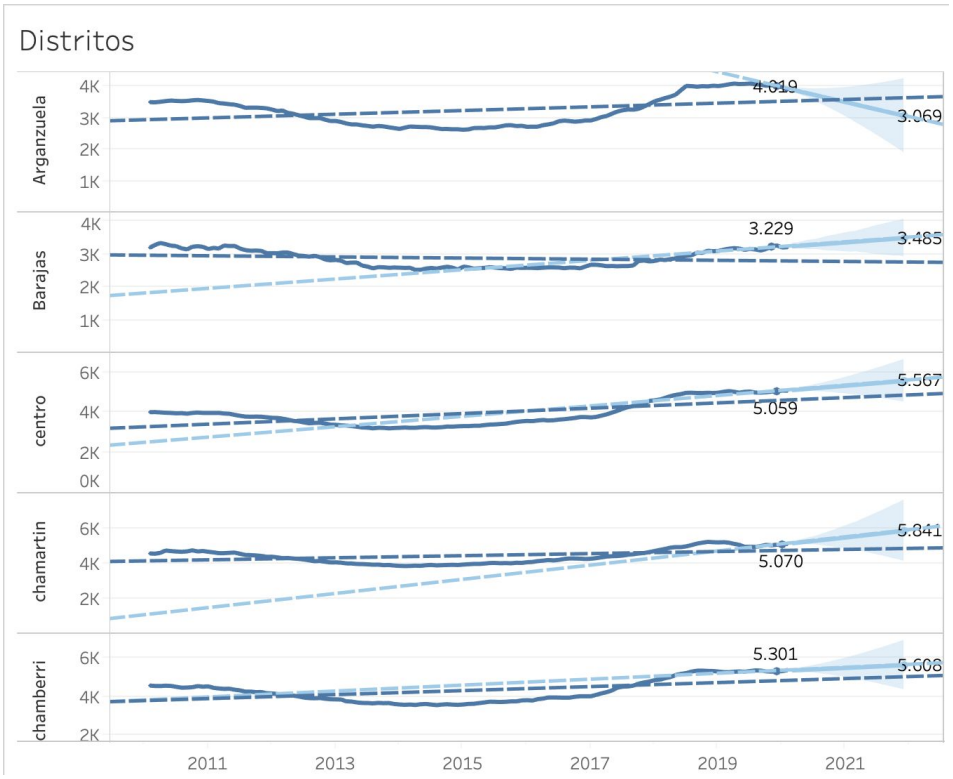
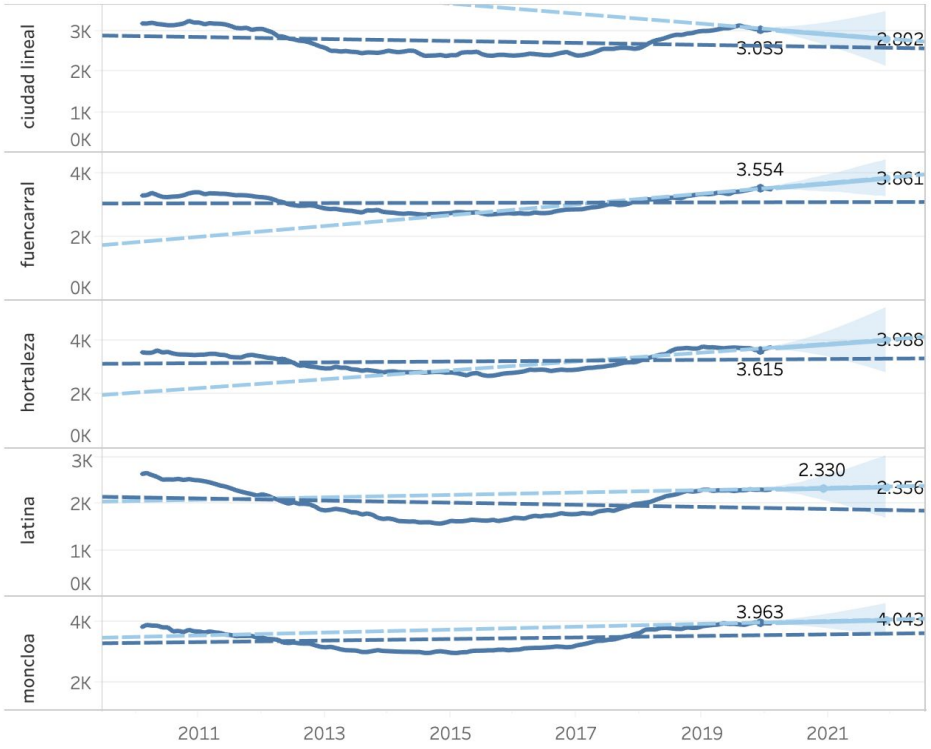


Gráfico 2.5: Crecimiento del Valor del m2 en Madrid a diciembre 2019-2021

En el gráfico 3 se observa el comportamiento que han tenido los 21 distritos de Madrid con sus respectivas proyecciones a diciembre del 2021 del precio del m2.


Distritos
Arganzuela
4K
3K
2K
1K
4.019
3.069
Barajas
4K
3K
2K
1K
3.229
3.485
centro
6K
4K
2K
0K
5.567
5.059
chamartin
6K
4K
2K
5.841
5.070
chamberri
6K
4K
2K
5.301
5.608
2011
2013
2015
2017
2019
2021


ciudad lineal
3K
2K
1K
0K
3.035
2.802
fuencarral
4K
2K
0K
3.554
3.861
hortaleza
4K
2K
0K
3.615
3.968
latina
3K
2K
1K
0K
2.330
2.356
moncloa
4K
2K
3.963
4.043
2011
2013
2015
2017
2019
2021

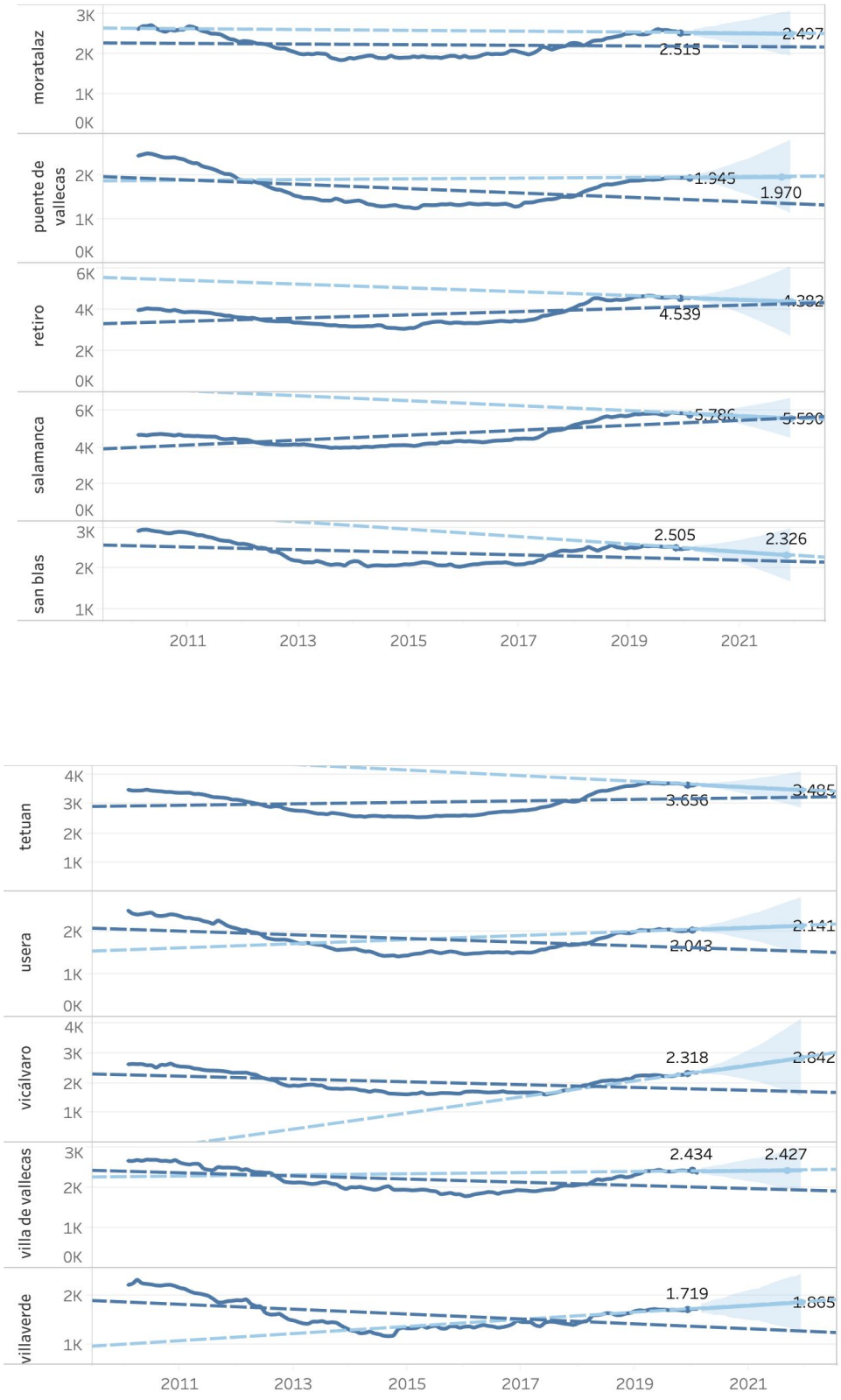


Gráfico 3: Análisis propio con datos de Idealista. Comportamiento del precio del metro cuadrado de los 21 distritos de Madrid.

Compra y Venta de Bienes Inmuebles Madrid 2019

En el gráfico 4 se evidencia el comportamiento de compra y venta de bienes inmuebles en Madrid durante el último año, en el cual se evidencia una variación del -7,64% entre enero 2019 y 2020: [19]

Gráfico 4: Estadística de Transmisiones de Derechos de la Propiedad en 2019. Del INE. Estadística de Transmisiones de Derechos de la Propiedad.

Venta de Pisos por Barrio de Madrid 2019

Por otro lado, en el gráfico 5 observamos el análisis de venta por distrito de Madrid en el año 2019. Vemos que el Centro es el distrito con mayor número de transacciones de bienes inmuebles llegando a las 4.428 viviendas vendidas frente a Barajas, que tan solo ha registrado 971 transacciones.

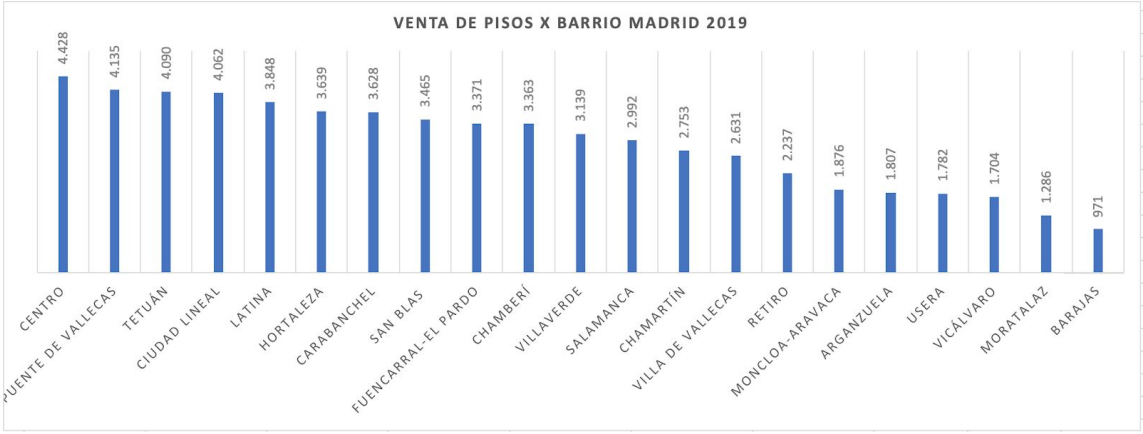


Gráfico 5: Venta de pisos por barrio en Madrid 2019.

[19] Instituto Nacional de Estadística (INE). (2020, Marzo 16). Estadística de Transmisiones de Derechos de la Propiedad. Recuperado Abril 10, 2020, de https://www.ine.es/jaxiT3/Datos.htm?t=6150#!tabs-grafico

Venta y Precio del metro cuadrado vs. Población 2019

En el gráfico 6 vemos la relación entre la venta de viviendas, el valor del metro cuadrado y el total de la población por distrito de Madrid. A grandes rasgos observamos que cuanto menor es el precio del metro cuadrado, mayor es el número de ventas, y que las zonas con menor número de habitantes tienen un precio por metro cuadrado más elevado.[20]

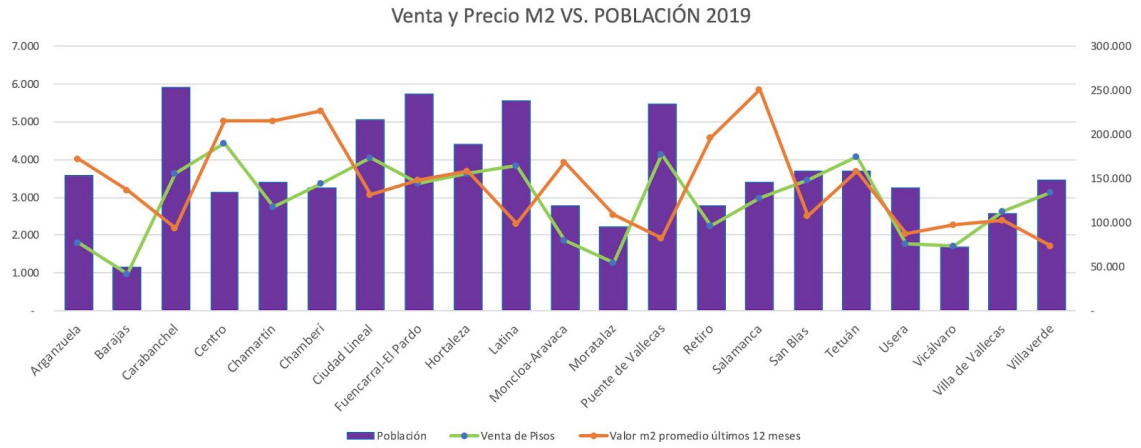
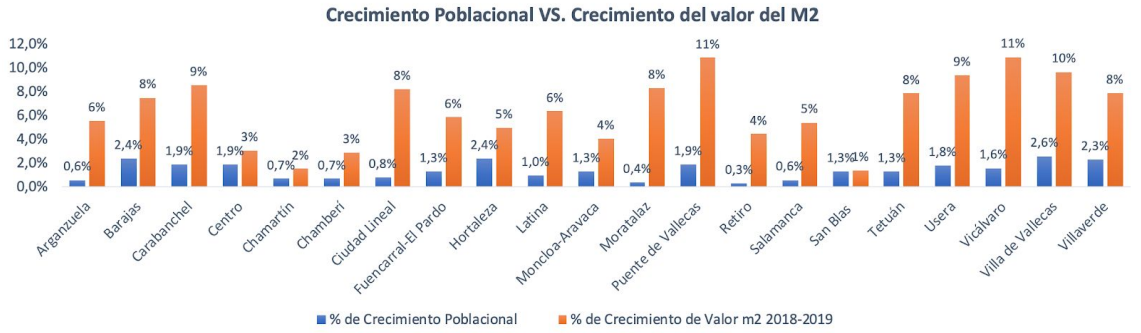


Gráfico 6: Venta y precio m2 vs. población en 2019

Crecimiento Poblacional vs. Crecimiento del valor del M2 de Madrid

En el gráfico 7 se visualiza el porcentaje de crecimiento que ha tenido la población[21] y el valor del metro cuadrado por distrito de Madrid[22] en el periodo 2018-2019 y a simple vista vemos que no existe una correlación entre estas dos variables, sin embargo, es una premisa que comprobaremos al momento de correr los modelos estadísticos.



[20] Ayuntamiento de Madrid. (n.d.). Inicio. Recuperado Abril 10, 2020, de https://www.madrid.es/portal/site/munimadrid

[21] Portal Estadístico. (2020, Marzo). Explotación del Padrón; Resto de municipios: INE, Revisión del Padrón. Recuperado Abril 10, 2020, de http://portalestadistico.com/municipioencifras/default.aspx?pn=madrid&pc=ZTV21&idp=35&idpl=1329&idioma=

[22] Idealista. (2020, Marzo). Evolución del precio de la vivienda en venta en Villaverde. Recuperado Abril 10, 2020, de https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/madrid-comunidad/madrid-provincia/madrid/villaverde/

Gráfico 7: Crecimiento poblacional vs. crecimiento del valor del metro cuadrado.

PIB vs. Crecimiento M2 vs. Crecimiento Mercado Inmobiliario Madrid

En la Tabla 2 se presenta una comparativa de la variación anual que ha tenido el PIB de España con respecto a la variación del valor del metro cuadrado que ha tenido Madrid en los últimos 10 años. Más adelante se determinará si existe o no una correlación entre estas dos variables.

AÑO	Variación Anual PIB	Crecimiento m2
2011	- 0,81%	-11%
2012	- 2,96%	-14%
2013	- 1,43%	-16%
2014	1,38%	-14%
2015	3,83%	9%
2016	3,04%	2%
2017	2,89%	4%
2018	2,35%	10%
2019	1,98%	8%

Tabla 2: Variación anual del PIB de España respecto al crecimiento anual del metro cuadrado en Madrid.

En el gráfico 8 se observa que puede existir una correlación entre la variación anual del PIB con el crecimiento que ha tenido el valor del metro cuadrado en los últimos 10 años, vemos que existe una menor correlación con la variación anual de la venta del sector inmobiliario en estos distritos, de igual forma son afirmaciones que se evaluarán en el análisis estadístico.[23]

[23] Instituto Nacional de Estadística (INE). (2019). Total Nacional. Datos ajustados de estacionalidad y calendario. Producto interior bruto a precios de mercado. Variación anual. Índices de volumen encadenados. Recuperado Abril 11, 2020, de https://www.ine.es/consul/serie.do?d=true&nocab&s=CNTR4892&nult=50

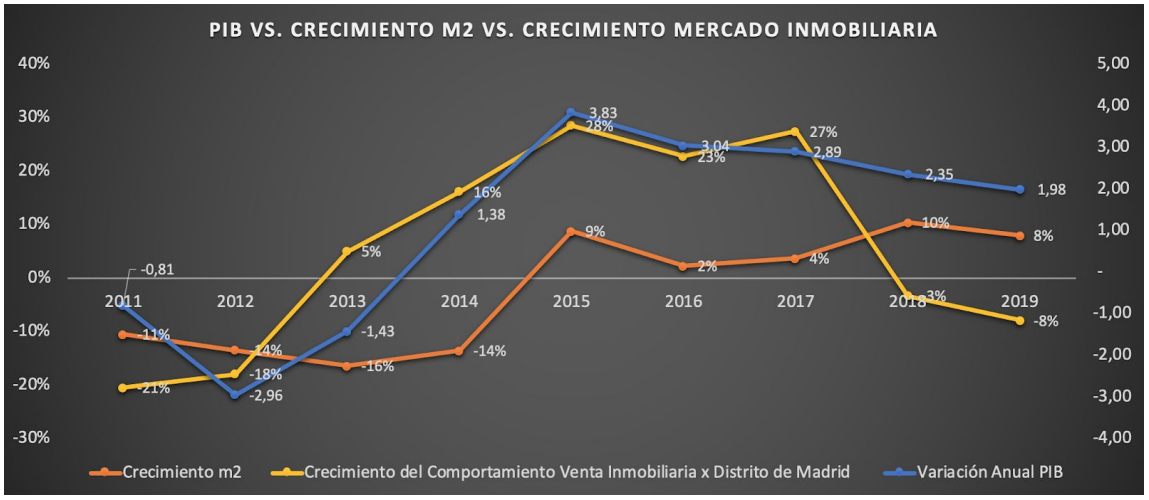


Gráfico 8: Variación anual del PIB vs. crecimiento anual del metro cuadrado vs. crecimiento del número de viviendas vendidas.

Métodos utilizados para predecir el precio del metro cuadrado.

Método PER

Las siglas PER significan "Price Earnings Ratio" y es un método utilizado para saber si es mejor comprar o alquilar una vivienda, además de identificar si el precio de un piso es barato o caro.

Las variables a utilizar para el cálculo de este método son:

- Precio de venta
- Precio de alquiler
- Rentabilidad bruta del inmueble

La manera de calcular el PER es a través de la división del valor de la vivienda para el valor que percibirá si arrienda el piso por un año.

Por ejemplo, si analizamos una vivienda a la venta por 400.000 euros y la misma en alquiler por 1.000 euros/mes (12.000 euros/año), ¿es preferible comprar, o alquilar? Lo primero que propone este método es calcular el PER de la vivienda dividiendo ambos datos (400.000/12.000) y ver que es 33 veces.

Este dato es alto, ya que es muy superior a la media española actual de 26,3 en la industria inmobiliaria. Es decir, con los números en la mano, esta casa es cara para comprar y sería mejor vivir en ella de alquiler.[24]

[24] Bankinter. (2020, Febrero 7). La tabla definitiva para saber si un piso está caro o barato. Recuperado Abril 11, 2020, de https://blog.bankinter.com/economia/-/noticia/2015/12/10/valoracion-inmuebles-pisos-baratos-caros

Método de Comparación

El método de comparación consiste en: como su propio nombre indica, tomar una muestra de inmuebles de similares características, compararlo con el bien inmueble a comercializar y fijar un precio que estará relacionado al valor de los inmuebles muestrales.

Las variables a utilizar dependen del modelo de negocio que quiera llevar cada empresa. Por ejemplo, la empresa Tercero B tiene un método que lo llaman 3BValue en el cual seleccionan 25 inmuebles de una similar tipología que cumplan con las siguientes características:

- Ubicados a menos de 700 metros del inmueble
- Superficie +/- 30% del inmueble que estamos analizando

Esto da como resultado el precio máximo al cual se recomienda vender el inmueble.

Ahora, para calcular el precio mínimo de comercialización de este inmueble, aplican el descuento medio de la zona dado por Idealista:

valor mínimo = valor máximo – descuento medio en la zona.

En el gráfico 9 se observa el promedio del porcentaje de descuento que se obtiene al negociar una propiedad por distrito.[25]

El promedio de descuento ofrecido en Madrid es del 18,9%.

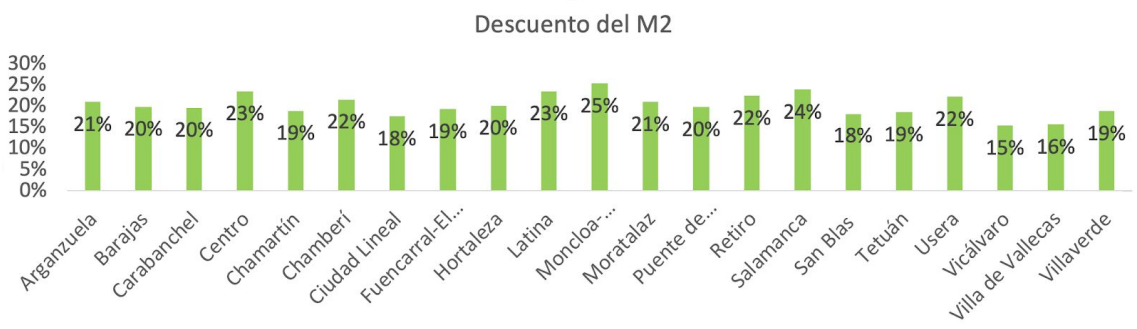


Gráfico 9: Descuento del m2.

Además, asignan una fiabilidad al precio propuesto en función de la calidad de las muestras: A+ A A- B+ B B-

[25] Idealista. (2016, Julio 26). Madrid: descuentos por pedidos por los compradores de vivienda. Recuperado Abril 10, 2020, de https://www.idealista.com/news/estadisticas/descuentos-vivienda/venta-viviendas/madrid--comunidad-de/madrid

Con este método han calculado que el error medio de su modelo será del 20% y que si el inmueble tiene plaza de aparcamiento o trastero el precio se incrementará en 27.000 euros y 3.000 euros respectivamente.[26]

Entre las empresas que utilizan este modelo está BBVA, que también nos dice si existe capacidad de negociación o no en función de la cantidad de inmuebles disponibles para su venta en la zona respecto del número total de viviendas, no recomendando negociar el precio si dicho ratio fuera inferior a la media del municipio.[27]

Bankia es otra empresa que también utiliza este método y como información adicional nos proporciona el número de zonas verdes, comercios, centros educativos, medios de transporte y vida nocturna del sector (Imagen 1: referencia de información de Bankia).[28]

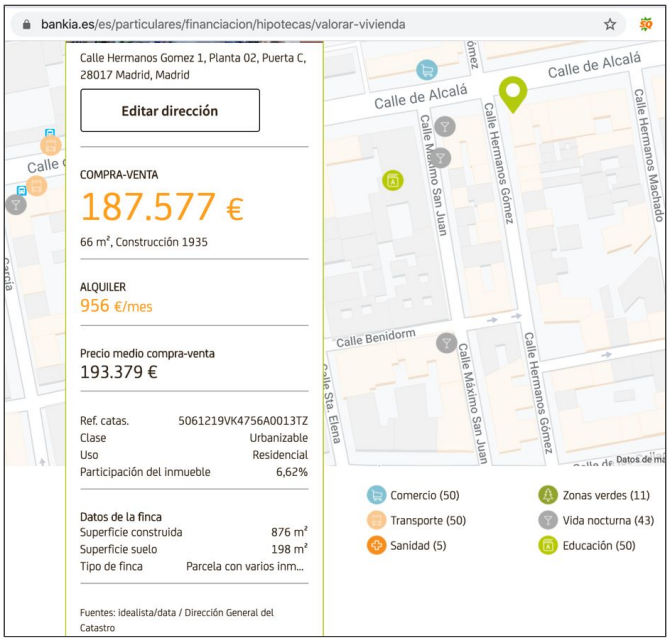


Imagen 1: Referencia de información de Bankia.

En el caso de Better Place[29], utiliza 40 propiedades para realizar la comparación y como estos, existen un sinfín de empresas (Idealista, Red Piso, Tengo tu Casa, Sociedad de Tasación, etc) que a través de un modelo u otro se encargan de predecir el valor del metro cuadrado de Madrid.

[26] TerceroB. (2014, Mayo 25). 3BValue. Recuperado Abril 10, 2020, de https://www.tercerob.com/3BValue
[27] BBVA. (2020, Abril 8). Descubre el precio medio de una zona con BBVA Valora. Recuperado Abril 11, 2020, de https://www.bbva.es/personas/experiencias/bbva-valora/analizar-barrio.html
[28] Bankia. (2017, Enero 12). Calcula el valor de tu vivienda. Recuperado Abril 11, 2020, de https://www.bankia.es/es/particulares/financiacion/hipotecas/valorar-vivienda
[29] Betterplace. (2020, Febrero 24). Madrid: informe inmobiliario del municipio: estado de compraventa y alquiler. Recuperado Mayo 10, 2020, de https://www.betterplaceweb.com/informe-inmobiliario-madrid-enero-2019/

Existen también varios estudios académicos relacionados con esta temática, de los cuales seleccionamos varios y analizamos pros y contras.

Estimación del precio del m2 a través de Redes Neuronales

Realizado por el Dpto. Estadística, Econometría, I.O. y Org. Empresas. Universidad de Córdoba en el año 2009 y focalizado en predecir el precio del metro cuadrado de las viviendas de la ciudad de Córdoba. Con dicho objetivo, utilizaron las variables referidas en la imagen 2 aunque descartaron varias argumentando lo siguiente: "En lo que respecta a las variables de entrada -inputs- a considerar, a priori es deseable incluir un número no demasiado elevado de este tipo de variables fundamentalmente porque la inclusión de un número elevado de variables no origina importantes mejoras en el poder explicativo de la red, de manera que si no se inclu- yen aquellas variables que poco puedan aportar esto tiene un reducido efecto en la significación global del modelo. De modo que ante varias redes con un poder explicativo similar se elegirá la más simple":

<table>
<tr><th colspan="3">INTERNAS DE LA VIVIENDA</th><th colspan="2">EXTERNAS DEL EDIFICIO</th></tr>
<tr><td>BÁSICAS</td><td colspan="2">Superficie const.
Dormitorios
Baños
Aseos
Terraza (*)
Teléfono (*)
Armarios empotrados(*)
Garaje(*)
Trastero(*)
Climatización</td><td>GENERALES</td><td>Año edificación
Ascensor(*)
Tendedero(*)</td></tr>
<tr><td rowspan="2">GENERALES</td><td>CALIDAD</td><td>Solados
Carpintería ext.
Carpintería int.
Amueblado(*)
Cocina amue.(*)</td><td rowspan="2">EXTRAS</td><td rowspan="2">Piscina(*)
Tenis(*)
Jardines(*)</td></tr>
<tr><td>REFORMA</td><td>Reformado(*)</td></tr>
<tr><td>ORIENTACIÓN</td><td colspan="2">Exterior(*)</td><td rowspan="2">LOCALIZACIÓN</td><td rowspan="2">Zona ubicación</td></tr>
<tr><td>ECONÓMICAS</td><td colspan="2">Gastos de comunidad
Precio de mercado</td></tr>
</table>

Imagen 2: Tabla sobre variables internas y externas de propiedades.

Variables utilizadas en el modelo final:

- Superficie: mide las dimensiones de la vivienda y está expresada en metros cuadrados construidos.
- Antigüedad: recoge el número de años que tiene el edificio en el que se ubica la vivienda.

- Ubicación: índice de ubicación del inmueble, que pondera la situación geográfica del inmueble junto con el nivel de renta de la zona.
- Anejos: índice de anejos, que recoge la existencia de garaje y trastero.
- Común: gastos de comunidad mensuales expresados en euros.
- Sol^Car: es la interacción entre la calidad de la solería y la calidad de la carpintería exterior de la vivienda.

En la Imagen 3 se observa la relación de cada una de las variables con el precio. Vemos como la influencia de estas variables fluctúa año tras año siendo la Superficie (construida) y la Ubicación (distancia al centro de la ciudad) las que tienen mayor correlación con el precio del metro cuadrado.[30]

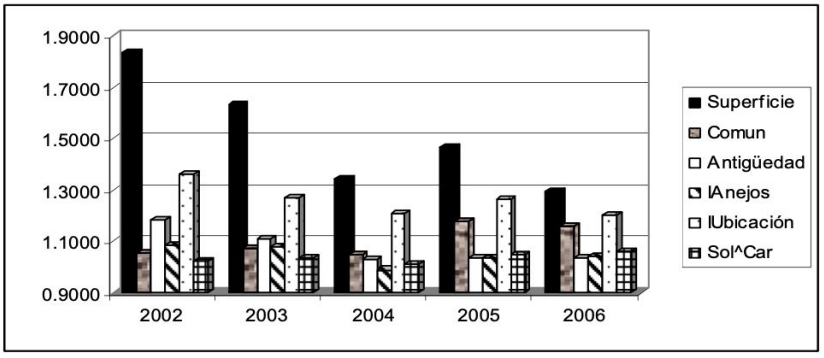


Imagen 3: Relación de las variables con el precio.

Modelo de Regresión lineal Regularizado

Este análisis fue realizado por Alexandru Papiu (Kaggle) en el cual se pretendía encontrar qué variables están relacionadas con el valor del m2 en Ames, Iowa. Tras de una limpieza de datos se realizó la regresión lineal tomando en cuenta las siguientes variables:

- Funtional: calificación de funcionalidad del hogar
- LotArea: Tamaño del lote en pies cuadrados
- Exterior1: cubierta exterior en la casa
- KitchenQual: calidad de cocina
- OverallQual: material general y calidad de acabado
- SaleType: tipo de venta

[30] Tabales, J. N., Caridad, J. M., Villamondos, N. C., & Jiménez, A. M. F. (2009, Febrero). Estimación del precio de la vivienda mediante redes neuronales artificiales (RNA) en diferentes marcos temporales. Recuperado Marzo 27, 2020, de http://casus.usal.es/pkp/index.php/MdE/article/view/994

- LandContour: llanura de la propiedad
- GarageCond: condición del garaje
- CentralAir: aire acondicionado central
- MSZoning: la clasificación general de zonificación
- SaleCondition: condición de venta
- Condición2: Proximidad a la carretera principal o al ferrocarril (si hay un segundo)
- RoofMatl: material del techo

En la imagen 4 se aprecia el grado de relación de cada una de las variables con el precio:

Imagen 4: Coefficients in the Lasso Model.

Por otro lado, Padro Marcelino (Kaggle) realizó su análisis al mismo dataset a través de la librería Kernel en Python y las variables que presentaron mayor relación con el precio del m2 fueron:[31]

- OverallQual: material general y calidad de acabado.
- GrLivArea: Superficie habitable por encima del nivel del suelo (pies cuadrados).
- GarageCars: tamaño del garaje en la capacidad del automóvil.
- GarageArea: Tamaño del garaje en pies cuadrados.
- TotalBsmtSF: pies cuadrados totales del área del sótano.
- 1stFlrSF: pies cuadrados del primer piso.

[31] Apapiu. (2017, Abril 21). Regularized Linear Models. Recuperado Abril 10, 2020, de https://www.kaggle.com/apapiu/regularized-linear-models

- FullBath: baños completos sobre rasante.
- TotRmsAbvGrd: Total de habitaciones por encima del grado (no incluye baños).
- YearBuilt: fecha de construcción original

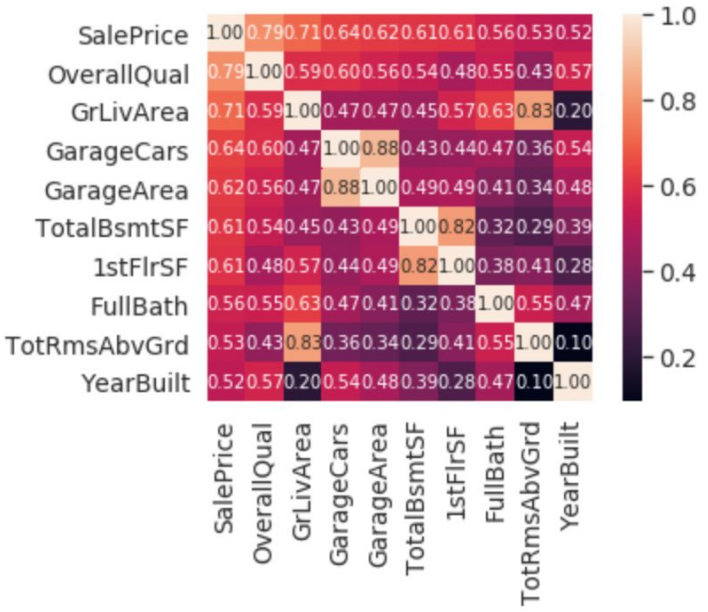


Imagen 5: 'SalePrice' correlation matrix.[32]

Vemos que la calidad del acabado y el tamaño de la superficie son las variables que más se correlacionan con el precio del m2 de las viviendas.

Insights obtenidos del estado del Arte y de las informaciones recopiladas

- El precio del m2 en Madrid se ha incrementado en 0,3% el último año.
- Existe una diferencia de hasta un 300% en el valor del m2 de un distrito a otro (Salamanca vs. Villaverde).
- La fluctuación del crecimiento del m2 por distrito irá desde un -24% en Arganzuela hasta un 23% en Vicálvaro para el año 2021 con base a una proyección obtenida con la herramienta “Estimación” de Tableau..
- La venta de viviendas en Madrid ha disminuido en un 7,64% de enero del 2019 a enero del 2020.
- Existe una diferencia de hasta el 450% en la transaccionalidad de viviendas de un distrito a otro (Centro vs. Barajas).
- Aparentemente existe una relación en la cantidad de población de un distrito vs. el valor del m2. A mayor población menor valor del m2 y viceversa.

[32] Pmarcelino. (2019, Agosto 23). Comprehensive data exploration with Python. Recuperado Abril 11, 2020, de https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

- Aparentemente existe una correlación entre la variación del PIB con el crecimiento anual del valor del m2 en Madrid de los 10 últimos años.
- Entre las principales metodologías utilizadas en Madrid para predecir el precio del m2 son el PER y el método de comparación.
- El promedio de descuento ofrecido al comercializar una vivienda en Madrid es del 18,9%.
- El método de comparación es el más utilizado por los principales players del mercado, tropicalizando así las variables a utilizar.
- Vemos que las variables que tienen una mayor correlación sobre el valor del m2 varían de una población a otra sin embargo la que se mantiene es el tamaño de su superficie

4. Metodología de trabajo-investigación del TFM:

En base a lo analizado en el estado del arte, se decidió estructurar nuestra metodología de trabajo basada en tres ejes; un análisis de las variables extrínsecas al valor del m2, un estudio de las variables intrínsecas de las viviendas y un análisis geoespacial de variables que puedan influir en el valor del m2 de las viviendas de Madrid.

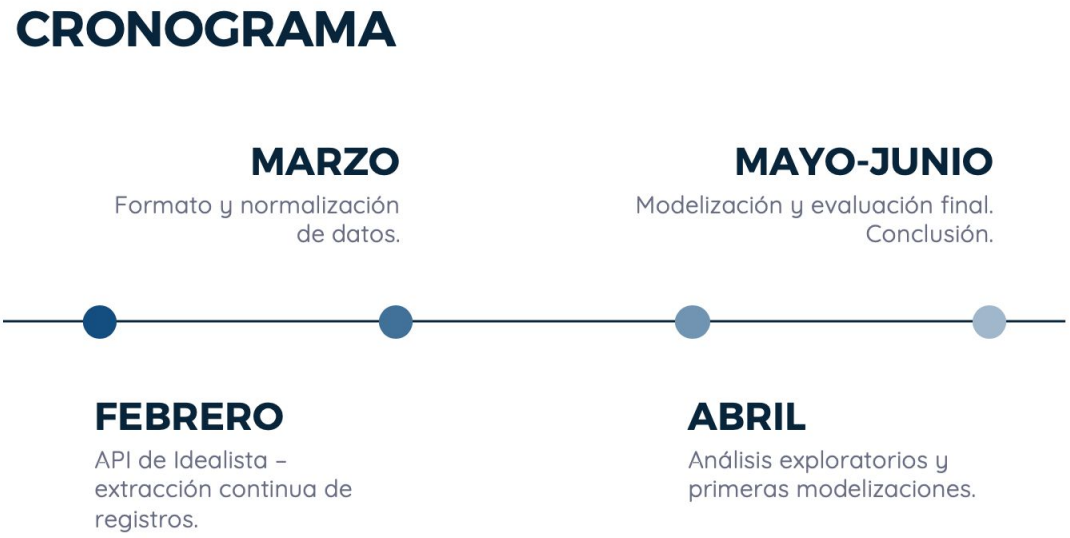
Variables Extrínsecas: Entiéndase por todas aquellas variables que no forman parte de la vivienda pero que pueden llegar a influir en el precio del m2, estas variables pueden ser: PIB, número de hipotecas, población de extranjeros viviendo en cada distrito, tasa de desempleo en el género femenino, entre otros.

Variables Intrínsecas: Al contrario de las extrínsecas son todas aquellas variables que forman parte de la vivienda tales como: número de habitaciones, cantidad de baños, si posee o no ascensor, en qué número de piso se encuentra la vivienda, entre otros.

Variables Geoespaciales: Cómo influyen en el valor del m2, aquellos servicios de primera necesidad como el metro, hospitales o escuelas que se encuentran geográficamente cercanas a las viviendas. Algunas de las variables extrínsecas cuentan, además, con determinado componente espacial, dado que pueden ser recopiladas por el organismo que corresponda (INE, Ayuntamiento de Madrid...) en base a distrito o zona.

Debido a que no contamos con una línea de temporalidad o estacionalidad en los datos intrínsecos y Geoespaciales, no utilizaremos modelos de series temporales. Utilizaremos modelos tipo GLM y SAR, que permitan añadir nuevos datos hasta obtener una precisión aceptable y unas betas robustas.

Cronograma de trabajo



5. Fuentes de datos, proceso de limpieza y tratamiento de los datos.

Fuentes de datos.

El proceso de ingesta se ha llevado a cabo mediante varias técnicas, tanto basadas en el scrapping del sitio web, como en el scrapping de los emails que, mediante suscripción, Idealista envía a nuestros correos.

Para ello hemos utilizado tanto Python como Java, en diferentes intentos de sortear las problemáticas que iban surgiendo tanto en la captura de los datos como en su posterior incorporación a la base de datos MySql. Durante este proceso utilizamos librerías del tipo BeautifulSoup, Request e imaplib para acceder a la web o al email en el caso de Python, explorando los nodos HTML y rescatando de ellos la información necesaria y google.cloud.sql.connector para incluirla en la BBDD creada a los efectos en la nube de Google.

El lector de emails equivalente en Java, desarrollado tras una serie de inconvenientes con la escritura Sql en Google Cloud a través de Python utilizaba la librería javax.mail y java.sql.

En cuanto a la imputación de coordenadas, merecen una mención especial dado el alto componente espacial del proyecto y la falta de precisión que ofrece Idealista, habitualmente de forma deliberada para ocultar la dirección exacta de la vivienda a los futuros compradores y proteger así a las agencias. En este caso hemos reconocido direcciones en tres formatos diferentes, por una parte, a nivel de barrio, a nivel de calle, y a nivel de calle con número.

Para "traducir" estos datos a coordenadas decidimos usar el API de Google Maps, en su versión de pago por consulta, llevando a cabo una doble comprobación que en caso de detectar una ubicación fuera del barrio al que pertenece el edificio, imputaba el centroide de dicho barrio, minimizando así el error.

No obstante, y dado que el ritmo de ingesta no era suficiente, decidimos finalmente utilizar el API que Idealista provee para tales efectos, consiguiendo un dataset de un total de 6571 registros. El límite de la API está fijado en 100 solicitudes al mes, y cada solicitud devuelve 50 registros de propiedades.[33]

La extracción se hizo con RStudio utilizando la URL de la API, filtrando con un máximo de precio de 500.000 euros, operación "sale" de venta, distancia al centro de 30km, y centrándonos en las coordenadas de Madrid (40.4167, -3.70325). A la solicitud de la API le añadimos un *consumer_key* y un *consumer_secret* que hemos obtenido de Idealista para ser autorizados a utilizar la API. Luego con un *httr::POST* hemos hecho la solicitud como texto guardándolo con *write_json* como archivo JSON y añadiendo un timestamp a cada solicitud, y finalmente guardando todos los archivos en Google Drive.

Los archivos JSONs incluyen 40 variables que son las siguientes: *'propertyCode', 'thumbnail', 'externalReference', 'numPhotos', 'price', 'propertyType', 'operation', 'size', 'exterior', 'rooms', 'bathrooms', 'address', 'province', 'municipality', 'district', 'country', 'latitude', 'longitude', 'showAddress', 'url', 'distance', 'hasVideo', 'status', 'newDevelopment', 'newDevelopmentFinished', 'priceByArea', 'hasPlan', 'has3DTour', 'has360', 'parkingSpace.hasParkingSpace', 'parkingSpace.isParkingSpaceIncludedInPrice', 'detailedType.typology', 'detailedType.subTypology', 'suggestedTexts.subtitle', 'suggestedTexts.title', 'neighborhood', 'floor', 'hasLift', 'parkingSpace.parkingSpacePrice', 'topNewDevelopment'*.

Proceso de limpieza y tratamiento de datos.

El proceso de limpieza se hizo en Google Colab, utilizando Python, hemos unificado todos los archivos JSONs obtenidos de la API para obtener un DataFrame que unificara todos los registros.

[33] I. (Ed.). (2009, Octubre 4). Idealista Labs. Recuperado Mayo 20, 2020, de https://www.idealista.com/labs/

Hemos transformado, limpiado y filtrado determinadas variables para un mejor manejo de los datos.

El proceso de limpieza ha sido el siguiente:

- Descartamos variables que podrían reflejar más el trabajo de la agencia que el valor propio de la vivienda, tales como fotografías profesionales, vídeos 360...
- Retiramos variables contenidas en otras pero que no aportan tanta información, como las contenedoras (caso de parkingSpace.hasParkingSpace, parkingSpace.isParkingSpaceIncludedInPrice y parkingSpace.parkingSpacePrice, donde de ser afirmativa cualquiera de las dos últimas podríamos inferir la primera)
- Finalmente filtramos precio (*price*) a todo menor o igual a 500.000 euros, tamaño (*size*) menor de 500m2, y municipio (*municipality*) de Madrid, deshaciendonos también de varios outliers.

Variables que no hemos utilizado dado que aparentemente no proporcionan información útil:

- *thumbnail*: imagen de la propiedad.
- *numPhotos*: número de imágenes que hay de la propiedad.
- *showAddress*: si enseña la dirección o no.
- *url*: página web del perfil de la propiedad.
- *externalReference*: referencia externa.
- *province*: provincia de la propiedad, que siempre va a ser Madrid.
- *operation*: operación de venta, todas son "sales".
- *country*: país de la propiedad, que siempre va a ser España.
- *detailedType.typology*: tipo de propiedad o vivienda, como por ejemplo piso, ático, dúplex, chalet, etc. Misma información que la variable *"propertyType"*.
- *suggestedTexts.subtitle*: texto metadato sugerido por Idealista como subtítulo del perfil de la propiedad.
- *suggestedTexts.title*: texto metadato sugerido por Idealista como título del perfil de la propiedad.
- *detailedType.subTypology*: subtipo de la propiedad, como por ejemplo piso, ático, dúplex, chalet, etc. Misma información que la variable *"propertyType"*.

Variables que hemos utilizado y su data type:

- *propertyCode* (variable numérica): número único de identificación para identificar cada propiedad.
- *price* (variable numérica): precio de venta de la vivienda.

- *propertyType* (variable carácter): tipo de vivienda, como por ejemplo piso, chalet, ático, dúplex, o estudio.
- *size* (variable numérica): metros cuadrados de la vivienda.
- *exterior* (variable lógica): si tiene exterior, cómo por ejemplo jardín o terraza.
- *rooms* (variable numérica): número de habitaciones que tiene la vivienda.
- *bathrooms* (variable numérica): número de baños.
- *address* (variable carácter): dirección de dónde se encuentra la propiedad.
- *municipality* (variable carácter): municipalidad de dónde se encuentra la propiedad.
- *latitude* (variable espacial): coordenada de latitud, imputada a zona si no hubiera dirección, calle y número.
- *longitude* (variable espacial): coordenada de longitud, imputada a zona si no hubiera dirección, calle y número.
- *distance* (variable numérica): distancia entre la propiedad y el centro de Madrid (Puerta del Sol).
- *status* (variable carácter): el estado de la propiedad.
    - Opciones: good, renew, newdevelopment, NaN.
- *newDevelopment* (variable lógica): si la propiedad es construcción nueva.
- *newDevelopmentFinished* (variable lógica): si la propiedad es construcción nueva y se ha terminado de construir.
- *priceByArea* (variable numérica): precio por metro cuadrado.
- *parkingSpace.hasParkingSpace* (variable lógica): si la propiedad tiene un parking o garaje.
- *parkingSpace.isParkingSpaceIncludedInPrice* (variable lógica): si el parking o garaje está incluido en el precio total.
- *parkingSpace.parkingSpacePrice* (variable numérica): precio adicional del parking o garaje que viene con la propiedad.
- *neighborhood* (variable carácter): el barrio de dónde se encuentra la propiedad.
- *floor* (variable carácter): número de planta de la propiedad donde está situada la vivienda.
- *hasLift* (variable lógica): si la propiedad o el edificio en el que se encuentra tiene ascensor.

6. Análisis EDA (Análisis exploratorio de los datos).

Conforme a lo analizado en el estado del arte vemos que existen un conjunto de variables extrínsecas e intrínsecas que influyen en el valor del metro cuadrado de una vivienda. De ahí que el análisis exploratorio de los datos lo hemos estructurado en las siguientes 3 fases:

1) Relación de las variables extrínsecas con el valor del m2 de Madrid cuyos datos abarcan el periodo del 2013 al 2019 y en el cual contrastaremos las dudas planteadas en el estado del arte con relación al efecto de las variables externas.
2) Relación de las variables intrínsecas de las viviendas con los datos obtenidos de 8 distritos de Madrid.
3) Análisis de correlación y multicolinealidad de las variables extrínsecas e intrínsecas con el valor del m2 de la vivienda en Madrid. El análisis se realizará con aquellas variables que en los análisis previos se han determinado que guardan una mayor influencia en el precio.

La estructura del análisis EDA será:

1) Detección de Outliers
2) Correlación
3) Detección de Multicolinealidad

Análisis de Variables Extrínsecas

Las variables que se contemplaron para el análisis de variables Extrínsecas con:

- *Ventas*: Ventas de viviendas por año en Madrid.
- *PIB:* PIB anual.
- *PIB per Cápita*: Producto interno bruto percibido por habitante.
- *Población Nacionales:* Total de habitantes españoles en la ciudad de Madrid.
- *Población Extranjeros:* Total de habitantes extranjeros en la ciudad de Madrid.
- *Desempleo Hombres*: Total hombres en paro.
- *Desempleo Mujeres:* Total mujeres en paro.
- *Tasa de Desempleo:* Porcentaje de desempleados vs. el total de la población.
- *Promedio de tasas de interés de hipotecas:* El promedio de las tasas de interés de hipotecas anuales.
- *Hipotecas:* Número de hipotecas otorgadas en Madrid.
- *Salario Mínimo:* Salario mínimo anual.

Detección de Outliers

Las variables que presentaron Outliers son:

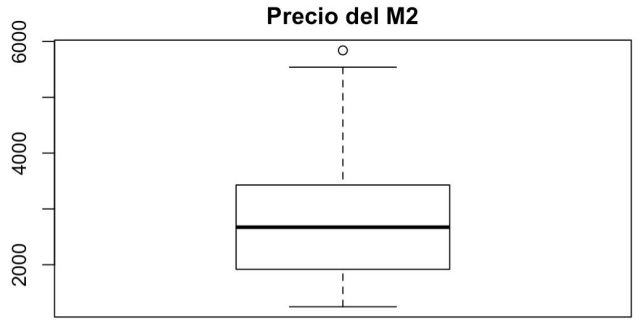


Gráfico 10: Precio del m2; datos extrínsecos

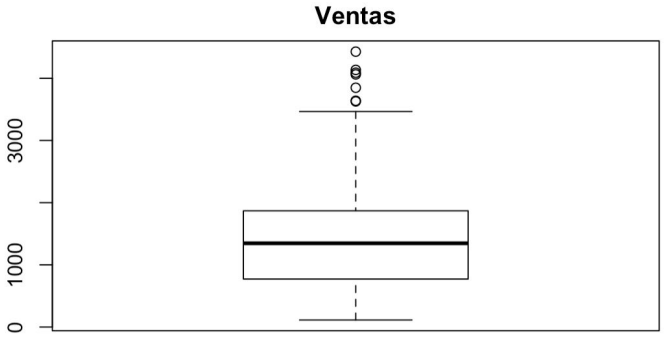


Gráfico 11: Ventas de viviendas; datos extrínsecos

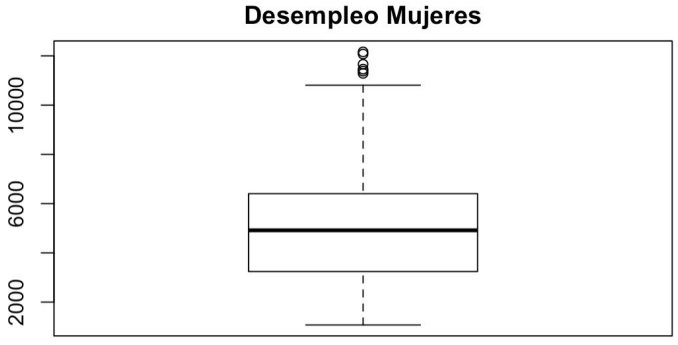


Gráfico 12: Desempleo Mujeres; datos extrínsecos

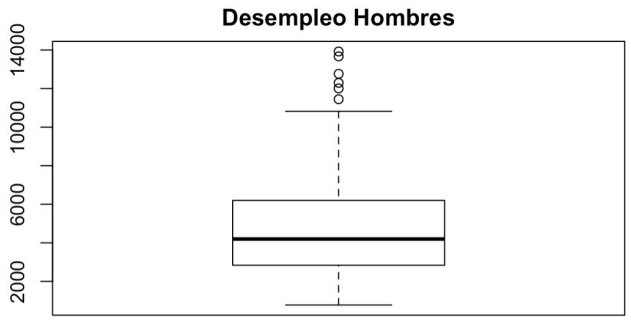


Gráfico 13: Desempleo Hombres; datos extrínsecos

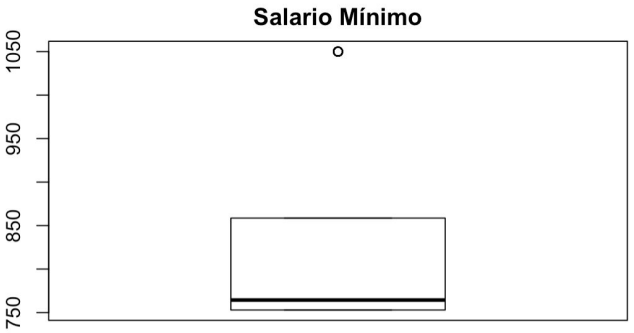


Gráfico 14: Salario Mínimo; datos extrínsecos

Se detectaron 46 datos Outliers, por lo cual se procedió a extraerlos.

Con la base de datos libres de Outliers procedimos a realizar el análisis de correlación.

Correlación

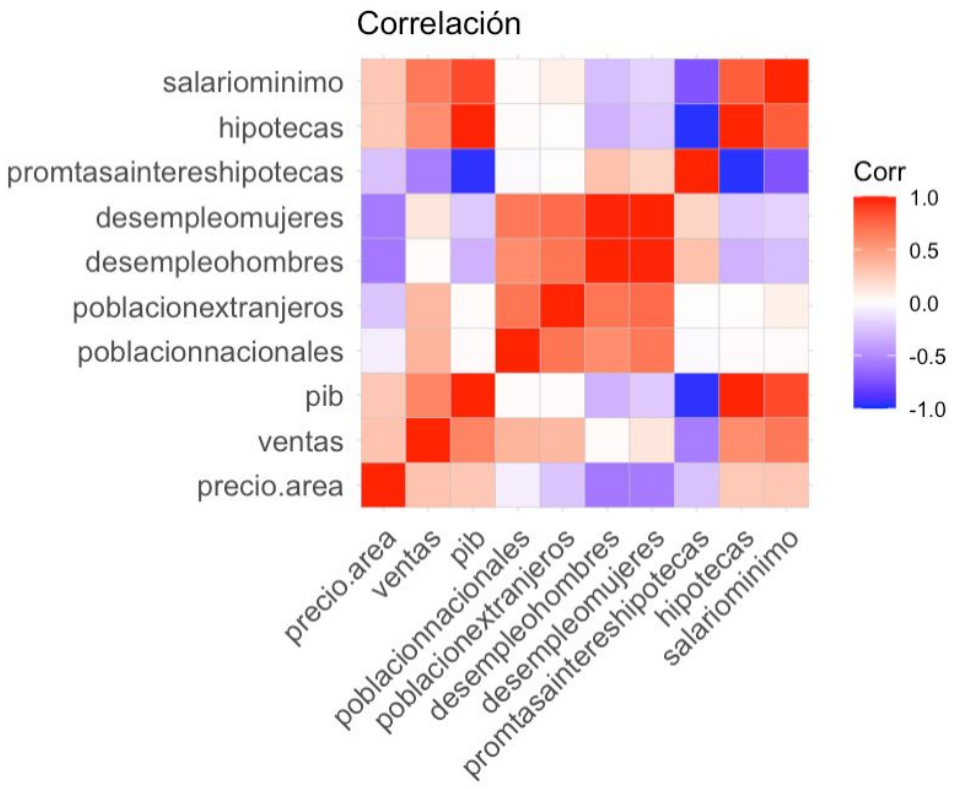


Gráfico 15: Correlación; datos extrínsecos

En el análisis de correlación vemos dos grupos de variables claramente identificados, aquellos que están correlacionados y aquellos que están inversamente correlacionados.

Variables con correlación respecto al precio de la vivienda del m2 en Madrid:

- Ventas
- PIB
- Hipotecas
- Salario Mínimo

Variables inversamente correlacionadas al precio de la vivienda del m2 en Madrid:

- Desempleo Mujeres
- Desempleo Hombres
- Población Extranjeros
- Promedio de tasas de interés de hipotecas

Análisis de Variables Intrínsecas

Realizamos un Boxplot (Diagrama de Cajas) y un Histograma con la variable principal el precio (*price*) para identificar outliers y la distribución de los datos.

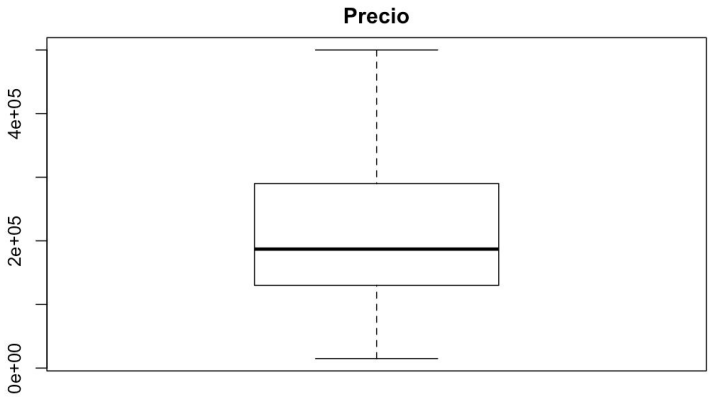


Gráfico 16: Precio del m2; datos intrínsecos

Además, al ser el precio nuestra variable a predecir es de nuestro interés saber el tipo de distribución que tienen nuestros datos, el gráfico 16 nos muestra que nuestros precios tienden a tener una distribución normal.

Curva de la distribución del precio

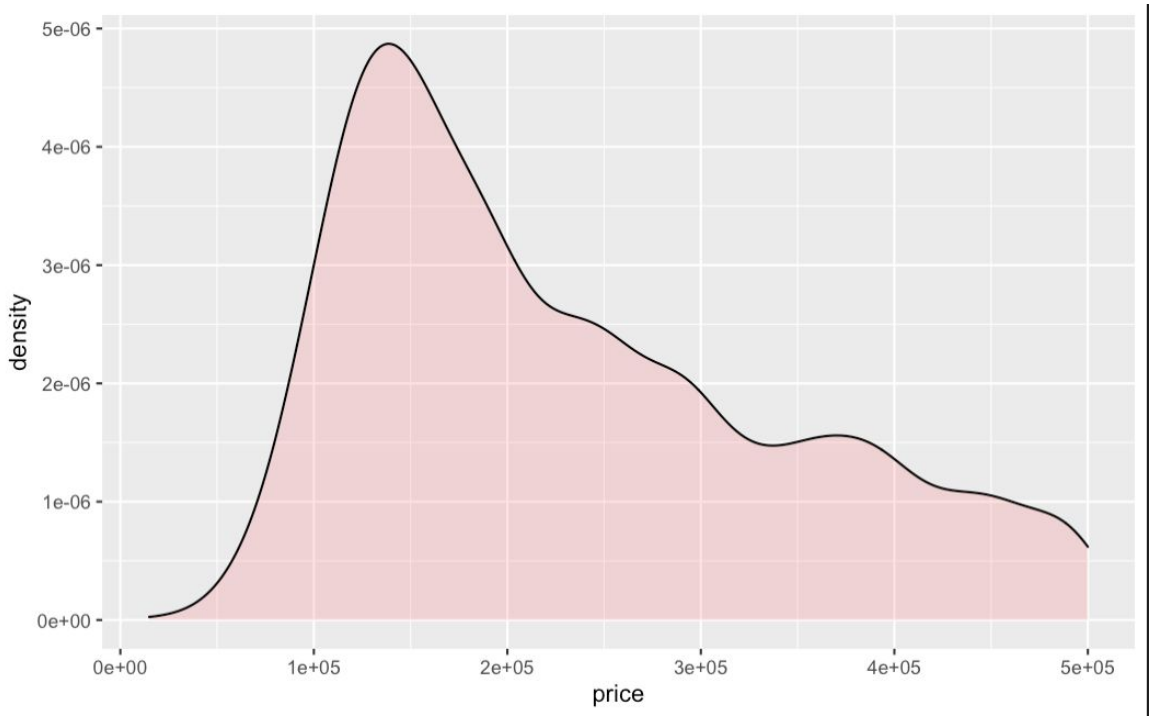


Gráfico 17: Distribución del precio

Además, para validar el tipo de distribución de nuestra variable precio realizamos un qqplot en el cual se determina que la distribución de la variable precios se aproxima a una normal.

Curva QQPlot

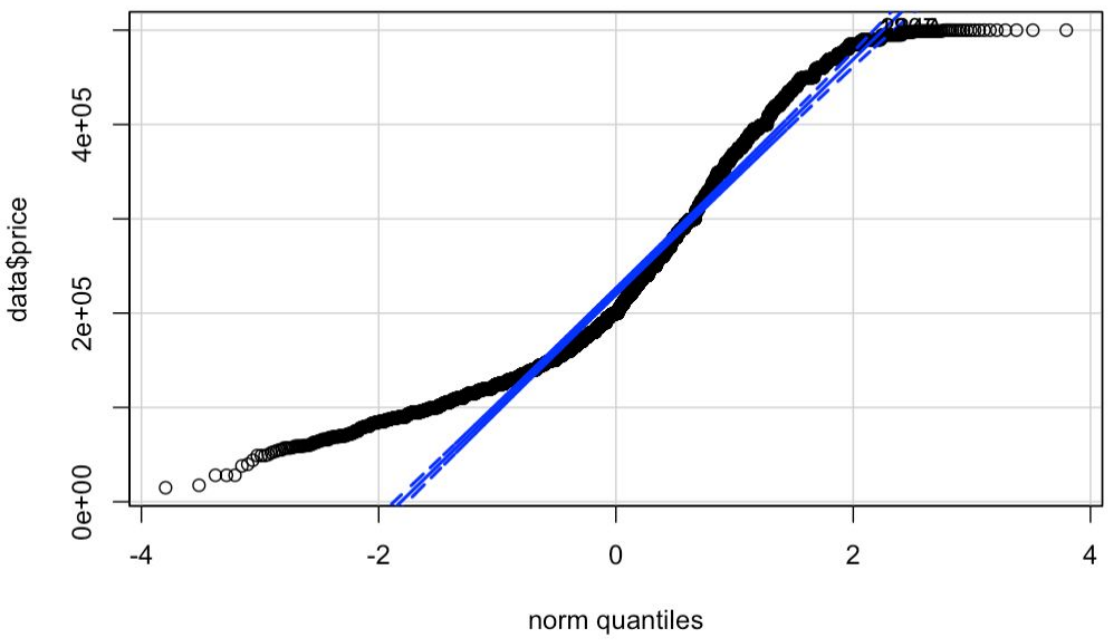


Gráfico 18: qqplot de la variable Precio

Seguimos las recomendaciones de la Escalera de Transformaciones de Tukey y probamos diferentes opciones de transformación de dicha variable para intentar que tuviera un mayor parecido con una distribución normal, pero dado que los resultados no eran significativamente mejores mantuvimos la variable precio sin modificaciones.

Para la variable del tamaño de las viviendas (*size*) también utilizamos Boxplot y Histograma para determinar más outliers y la distribución de los datos. Observamos una media y una mediana relativamente cercanas en los tamaños. Una media superior indicaría una asimetría positiva.

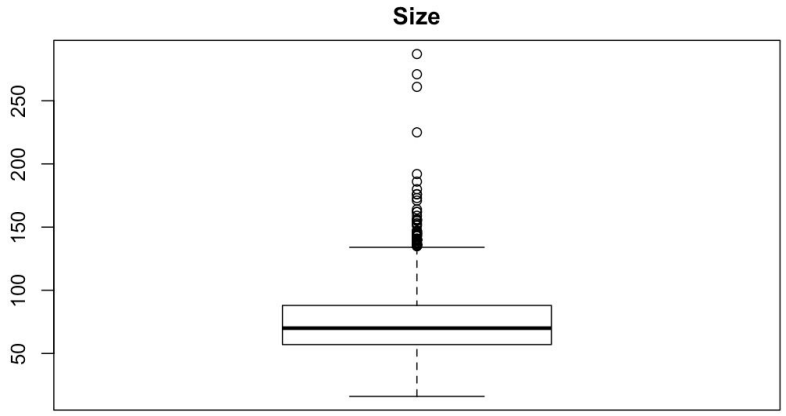


Gráfico 19: Size; datos intrínsecos

La variable de altura del piso o planta (*floor*) la agrupamos en “primera” (1), “mitad” (2,3,4,5), “alto” (6 ... 17) y “otro” (st, ss, en, -1, bj). En caso de datos perdidos (NA o NaN) le imputamos el valor “otro” por ser el más desfavorable.

La variable de tipo de propiedad (*propertyType*) la analizamos con una tabla de frecuencias de valores, y observamos que el valor “countryHouse” tiene 0 registros, eso podría deberse a algún espacio en blanco o carácter raro. Por lo que eliminamos esos registros. Además, observamos que el valor “flat” tiene una clara mayoría ante los otros valores dentro de la variable *propertyType*. Finalmente trabajaremos solo con el valor “flat”.

Gráfico de cantidad de viviendas por “propertyType”

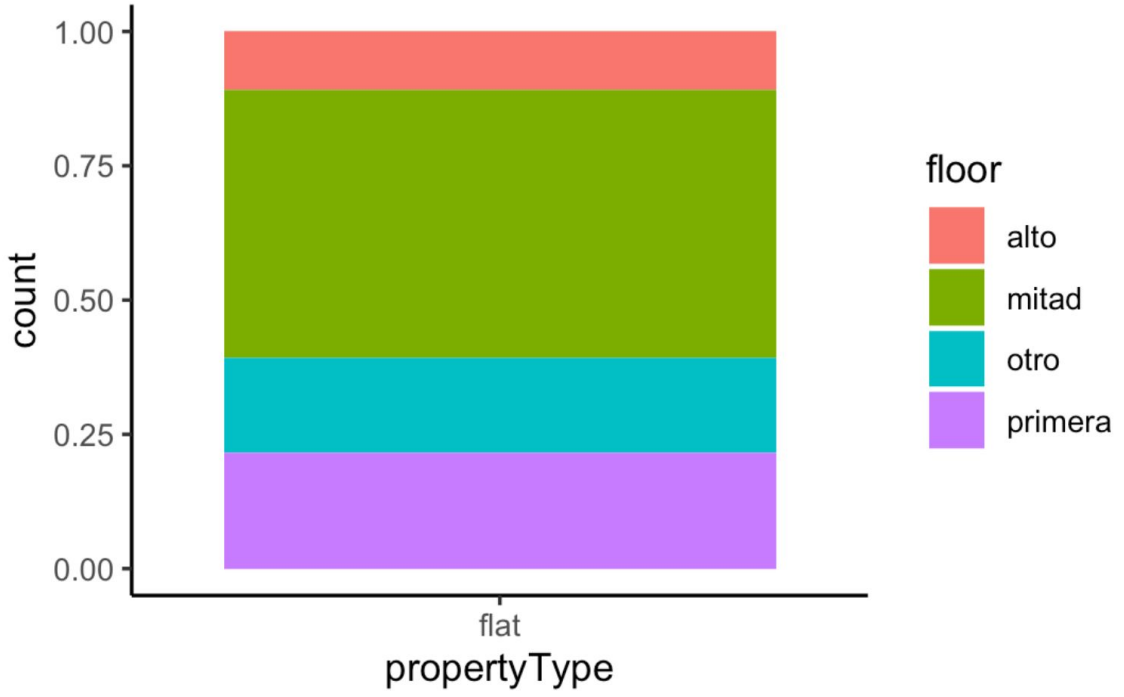


Gráfico 20: PropertyType; datos intrínsecos

Tomando en cuenta la variable del barrio (*neighborhood*), se trata de una buena forma de geoposicionar las viviendas dado que Idealista establece barrios relativamente pequeños y bien diferenciados. Las coordenadas (Latitud y Longitud) son sin duda mejores, pero debemos contar con la dificultad a la hora de capturarlas con determinados métodos fuera del API. Con una tabla de frecuencia de valores observamos los registros que aparecen con 0 registros, y las eliminamos.

Viendo otra vez las variables *latitude* & *longitude* todo está, aparentemente, en territorio “Madrid” y no hay datos perdidos (NA).

Gráfico de los centroides de viviendas por obtenido por las coordenadas (Latitud y Longitud)

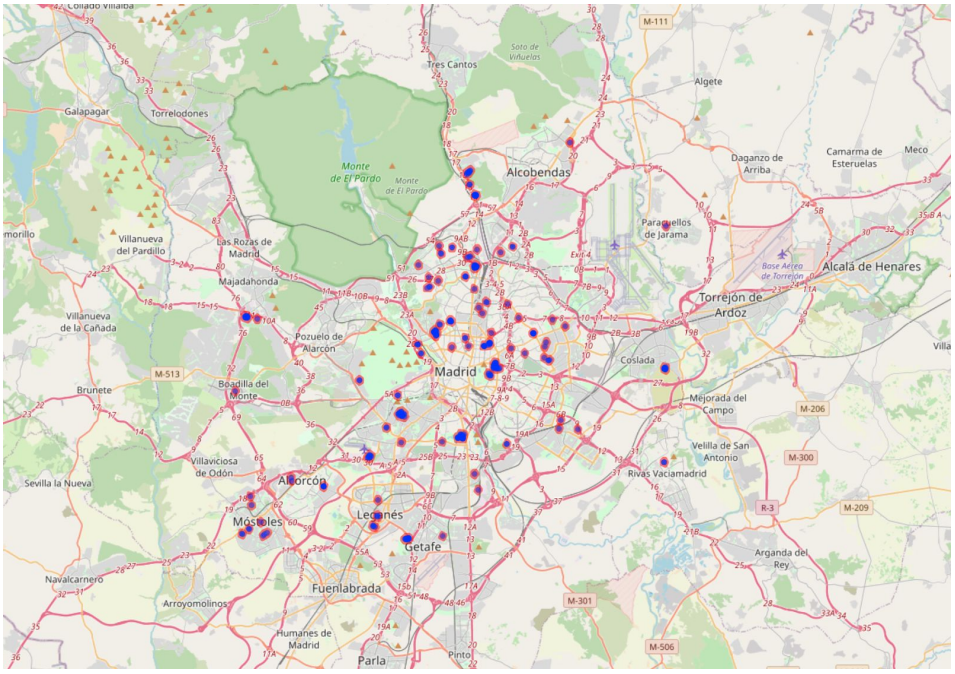


Gráfico 21: Centroides de viviendas; datos intrínsecos

En la variable *status* existen muchos datos perdidos (NA), y dado que "renew" es el factor más desfavorable se lo imputamos a todos ellos. La única razón más allá de la incidencia informática que tendría alguien para ocultar que ha reformado su vivienda es, efectivamente, que no lo ha hecho.

La variable *newDevelopment* duplica los datos de *status*, pero con menor información. Contrastamos si los registros a los que hace referencia *newDevelopment* como TRUE lo son en la variable *status* y de ser así la eliminamos.

Hacemos lo propio con la variable *newDevelopmentFinished*, dado que únicamente contiene 5 registros en el dataset y también son redundantes. Lo mismo con *topNewDevelopment*, aunque desconocemos si ha sido mal importada pero únicamente tiene datos perdidos (NA) y FALSE.

Para las variables *hasParkingSpace*, *isParkingSpaceIncludedInPrice* y *parkingSpacePrice*, si el precio está informado en parking, sumamos este precio al de la vivienda. Será *hasParkingSpace* quien refleje si tiene parking o no, y en cualquier caso estará incluido en el precio. Eliminamos isParkingSpaceIncludedInPrice y parkingSpacePrice por carecer ahora de sentido.

Gráfico de hasLift vs. floor

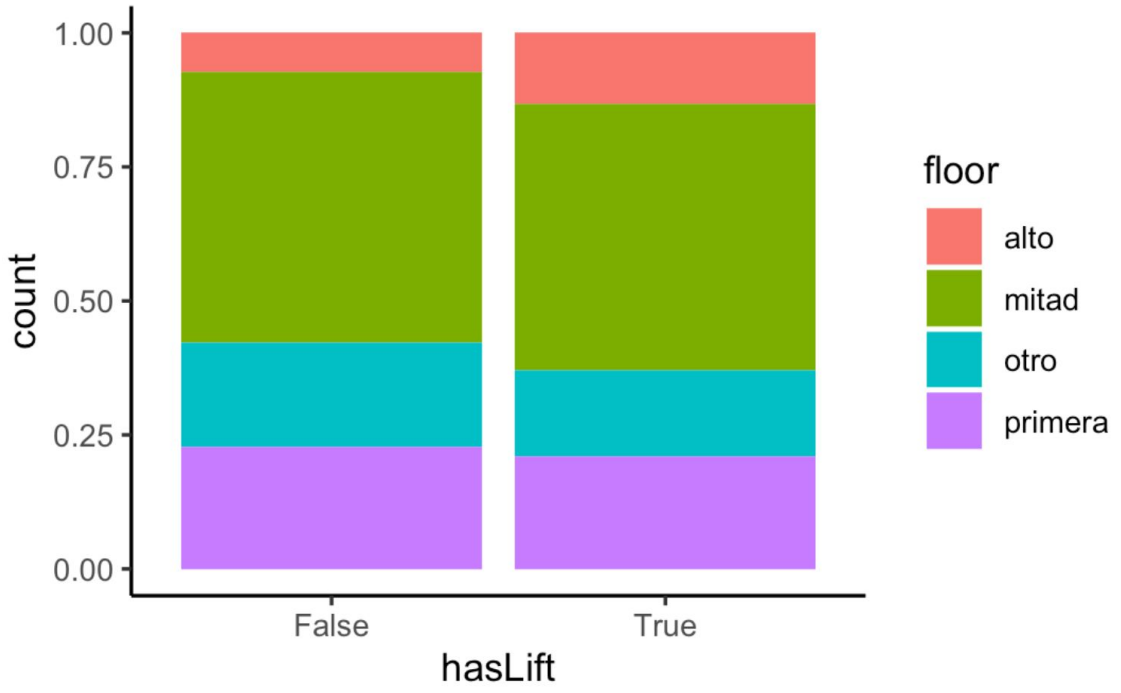


Gráfico 22: hasLift vs. floor; datos intrínsecos

En el gráfico superior se valida que existe una distribución similar entre los pisos que tienen o no tienen ascensor con relación a la altura en la que se encuentre el piso.

Gráfico de district respecto a status

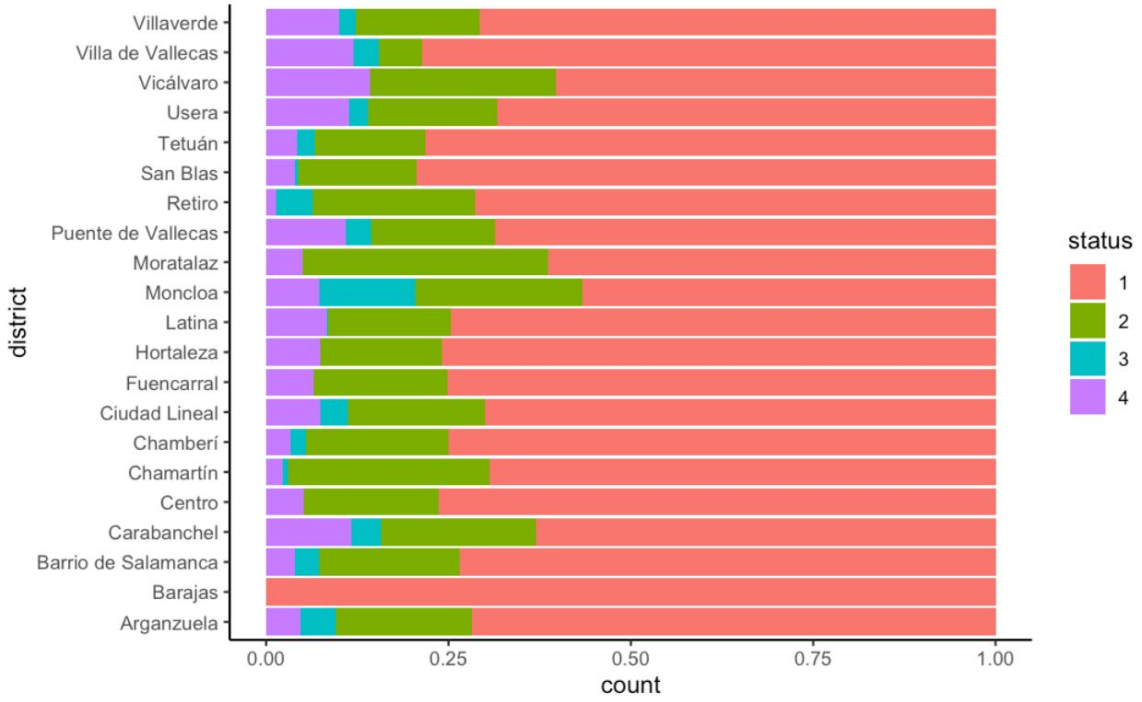


Gráfico 23: district respecto a status; datos intrínsecos

En el gráfico superior observamos como la mayoría de los pisos independientemente del distrito en el que se encuentren, poseen en su mayoría pisos en buen estado (1) y muy pocos pisos nuevos (3).

Finalmente, analizamos el resto de las variables en busca de NAs.

Correlación

Gráfico de Correlación de Variables Intrínsecas

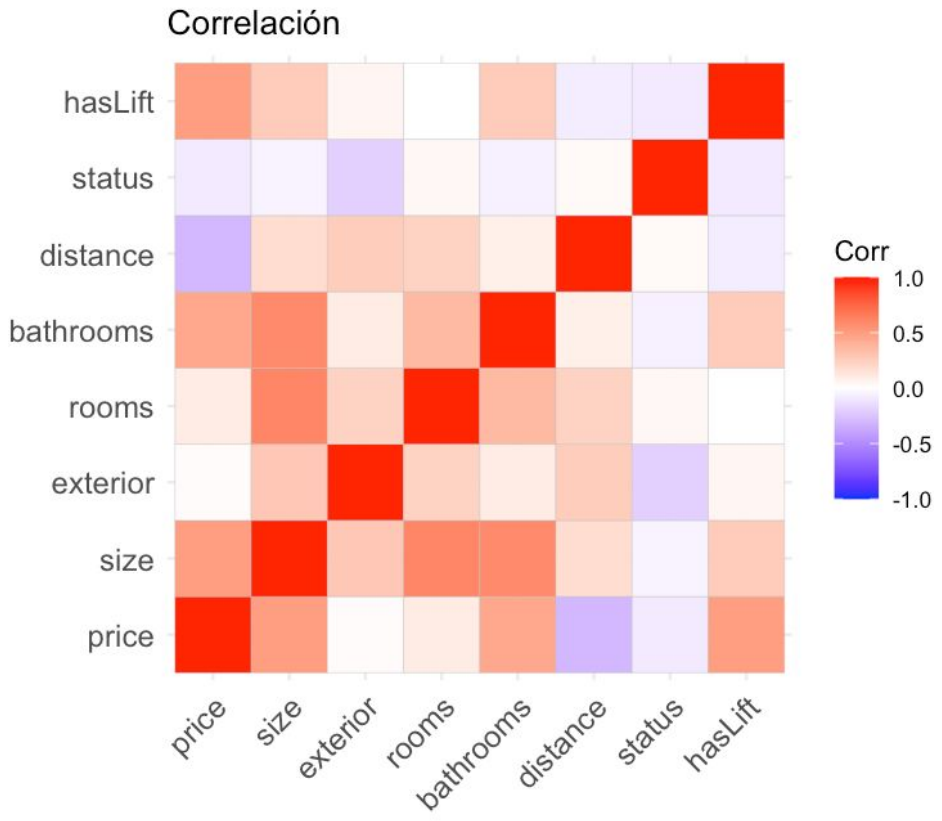


Gráfico 24: Correlación de variables; datos intrínsecos

En el gráfico de correlación se observa como las variables size, bathrooms y hasLift son las que presentan mayor correlación con precio y por otro lado, distancia y el status son las únicas variables que presentan una correlación negativa

Finalmente, y en base a la correlación que presentan las diferentes variables extrínsecas e intrínsecas, seleccionamos con cuales trabajaremos al momento de realizar los modelos estadísticos. El análisis de correlación de las variables escogidas es el que se aprecia en el gráfico 25.

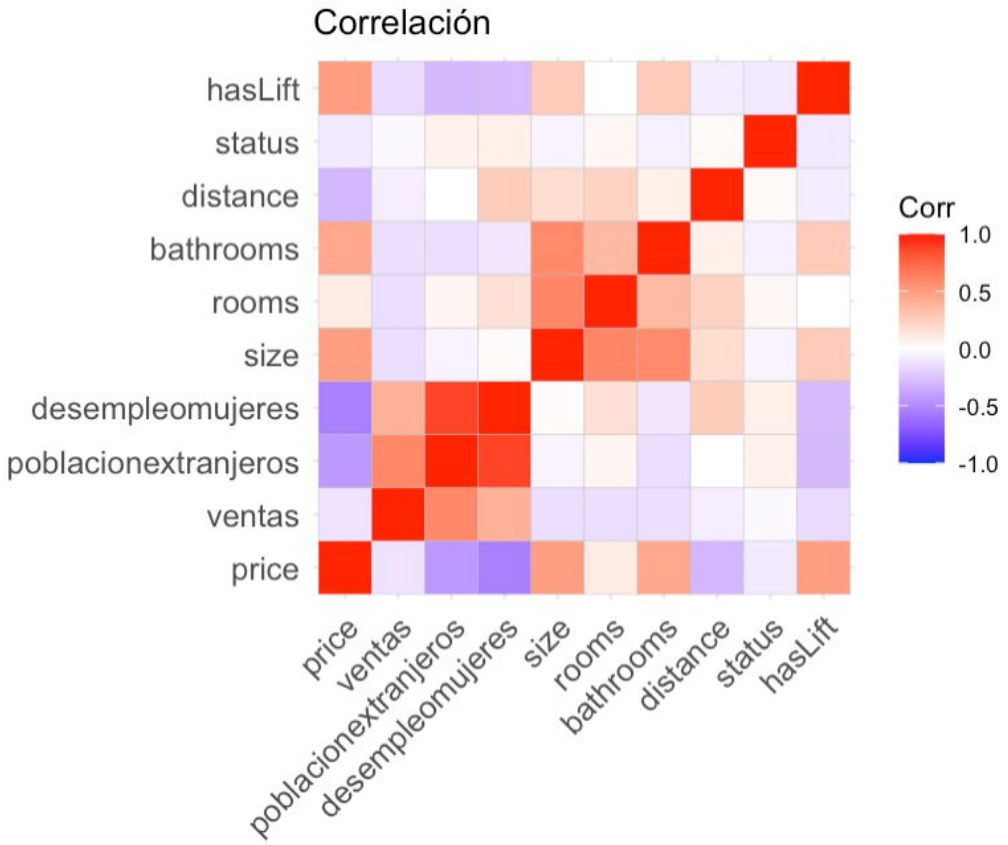


Gráfico 25: correlación de variables; datos intrínsecos y extrínsecos

Multicolinealidad

Para verificar si existe multicolinealidad en las variables extrínsecas e intrínsecas, utilizamos el análisis VIF con el cual determinamos que las únicas variables que presentan multicolinealidad son:

- *desempleohombres*
- *desempleomujeres*

Esto no significa que no debamos tenerlas en cuenta para el desarrollo del modelo. Si se determina que son importantes podrán ser utilizadas, pero debemos tener presente este factor.

	GVIF	Df	GVIF^(1/(2*Df))
distance	1.510695	1	1.229103
size	2.346232	1	1.531742
rooms	1.811106	1	1.345773
bathrooms	1.594697	1	1.262813
status	1.080425	1	1.039435
hasLift	1.279545	1	1.131170
ventas	1.891695	1	1.375389
poblacionextranjeros	8.898577	1	2.983048
desempleomujeres	7.387531	1	2.718001
floor	1.127651	3	1.020225

Imagen 6: Análisis de multicolinealidad

Heterocedasticidad

Al realizar un geom_point entre el precio de las viviendas y su tamaño, verificamos que existe heterocedasticidad en la mayoría los datos, misma que será analizada al momento de realizar los modelos estadísticos.

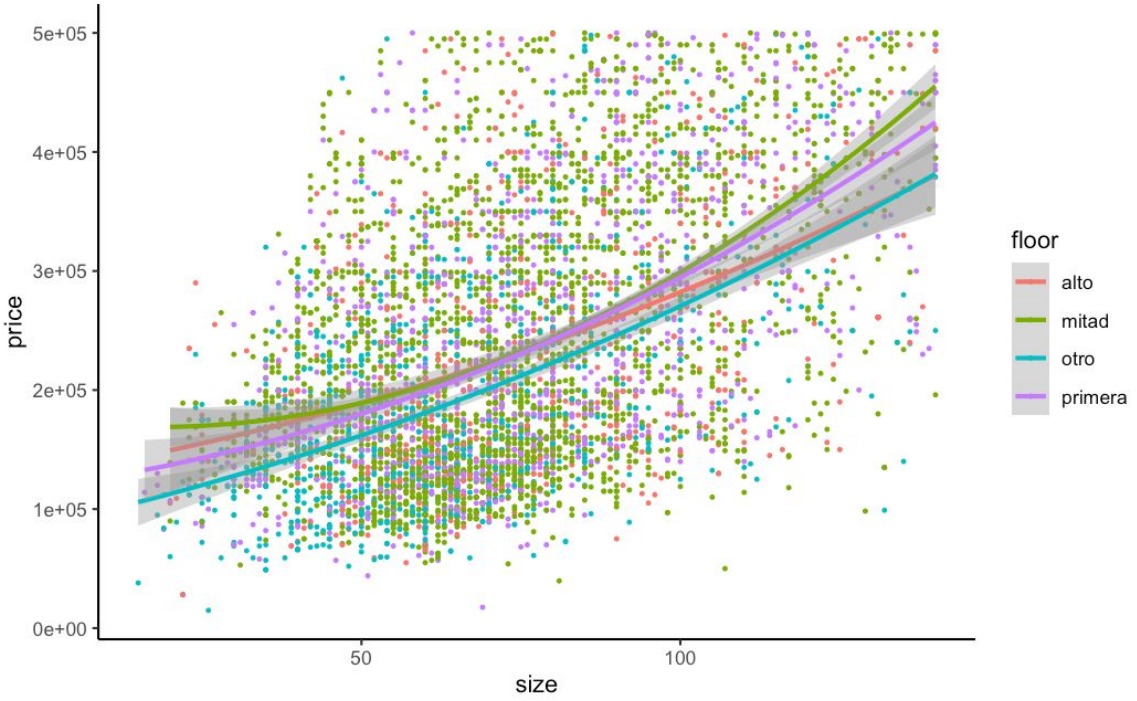


Gráfico 26: correlación de variables; datos intrínsecos y extrínsecos

7. Selección de técnicas y modelización del proyecto.

En cuanto a la modelización, tras probar varios métodos nos quedamos con tres que explican el proceso que ha seguido este TFM:

1.- GLM sin presencia de datos con componente espacial.

En este modelo tenemos únicamente en cuenta las variables sin componente espacial y que se relacionan con las características propias del piso. La altura (variable floor), se mostró no significativa en otras pruebas de GLM para todas las categorías menos para "floor_otro", por lo que únicamente usamos este estado de la variable. Además, añadimos dos variables nuevas elevando al cuadrado y al cubo el valor del tamaño de la vivienda, de manera que permitimos al modelo llevar a cabo algún tipo de spline llegado un punto de inflexión.

Además, tras varias pruebas en las que la precisión era bastante mejorable, decidimos normalizar todas las variables susceptibles de serlo salvo el propio precio, decisión que mantenemos también en los siguientes modelos.

La fórmula aplicada es la siguiente:

Price ~ size + size2 + size3 + rooms + bathrooms + status + hasLift + floor_otro

y los coeficientes devueltos:

```
## Call:
## glm(formula = formula, family = "gaussian", data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -233353   -44866    -3223    38344   260058
##
## Coefficients:
##                Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept)    237764.8   1073.4       221.498   < 2e-16 ***
## size           40267.1    1617.6       24.894    < 2e-16 ***
## size2          -3935.0    768.0        -5.124    3.08e-07 ***
## size3          1269.0     466.1        2.723     0.00649 **
## bathrooms      16560.3    1024.2       16.169    < 2e-16 ***
## status         400.2      793.3        0.504     0.61394
## hasLift        25832.1    864.9        29.867    < 2e-16 ***
## floor_otro     -25615.0   2095.6       -12.223   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 4077700819)
##
```

```
##     Null deviance: 8.1577e+13  on 6640  degrees of freedom
## Residual deviance: 2.7035e+13  on 6630  degrees of freedom
## AIC: 165817
##
## Number of Fisher Scoring iterations: 2
```

Observamos que todas las variables tienen un pvalor infinitesimal o muy bajo, por lo que podríamos dar veracidad a sus betas. El AIC de este modelo es de 165817 y el RMSE de 63803.97

```
datos$glm2predict <- predict(glm2,datos,type="response")
datos$e2 <- ((datos$price)-(datos$glm2predict))

RMSE(datos$glm2predict,datos$price)
#63803.97
```

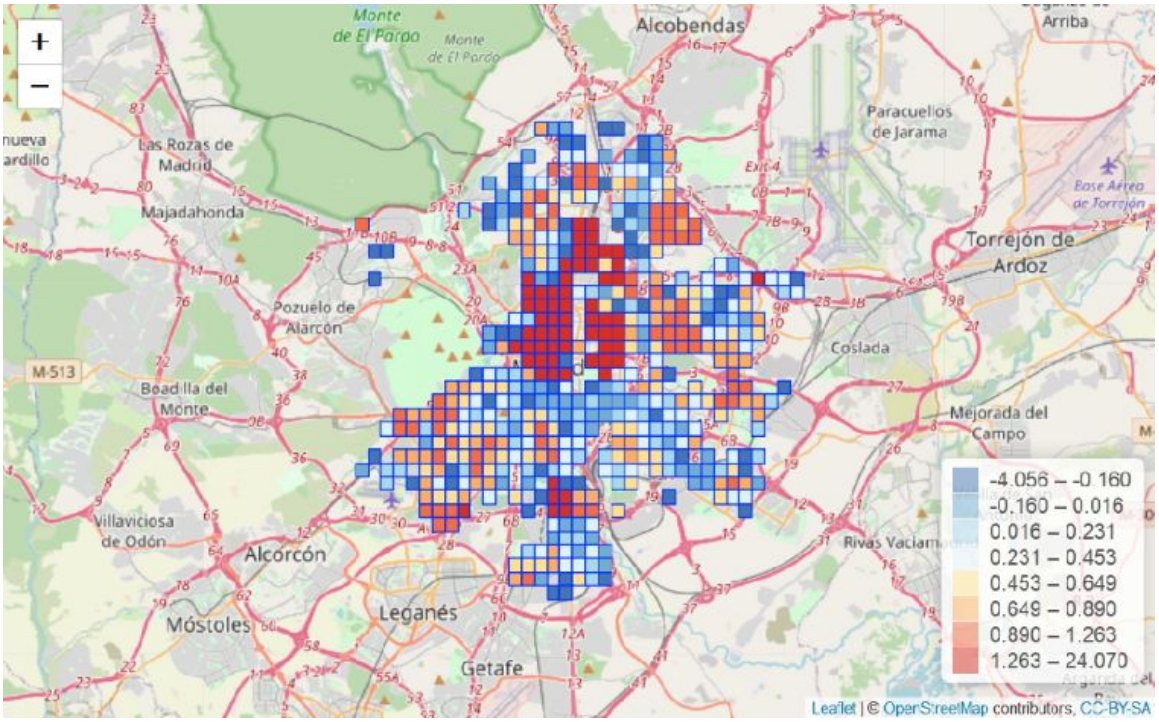


Gráfico 27: Raster 1

Si llevamos a cabo un test de Kulldorf sobre los residuos obtenemos un p-valor del 1% y observamos un clúster central donde nuestro modelo infravaloraría las predicciones. En cuanto a iMoran, el estadístico que obtenemos también es significativo y tiene un valor de 16, dada la alta dependencia espacial.

Moran I test under randomisation

data: xx$response
weights: nb2listw(knn2nb(knn))
Moran I statistic standard deviate = 16.038, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:

Moran I statistic	Expectation	Variance
0.4314091001	-0.0020202020	0.0007304031

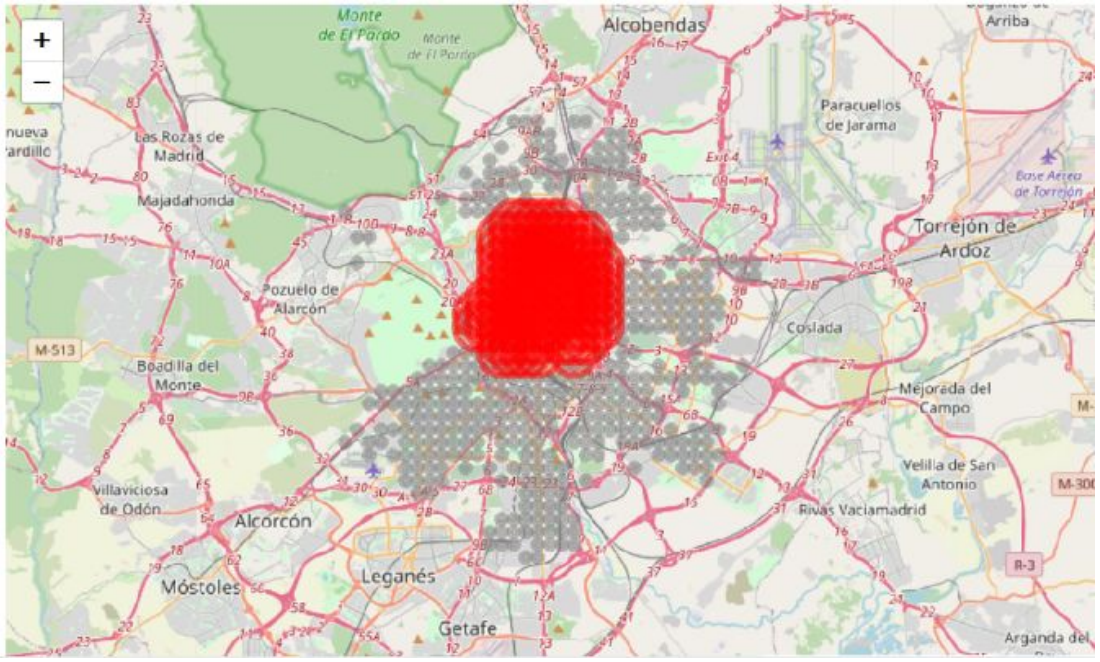


Gráfico 28: Kulldorf *1*

2.- GLM teniendo en cuenta variables con componente espacial

El segundo modelo por el que nos decantamos es una vez más GLM, aunque en esta ocasión añadimos varias variables con componente espacial. Estas son:

distancia_km_metro, distancia_km_hospital, distance, ventas, poblacionextranjeros, desempleomujeres, distancia y norte

La primera - distancia_km_metro - refleja la distancia de cada vivienda a las bocas de metro de la ciudad, habiéndose obtenido dichas ubicaciones mediante el uso de la sección Open Data de la página web del Consorcio Regional de Transportes de Madrid (fichero SHP) y la función

"distm" mediante el método “law of cosines”, que asume que la tierra es esférica para sus cálculos.

La segunda, denominada "distancia_km_hospital", guarda la distancia a hospitales de cada una de las viviendas. En este caso las ubicaciones se obtuvieron mediante el API de OSM, dado que ofrecían la calidad suficiente y no como en el caso de los Metros donde se optó por una alternativa más adecuada.

En el primer análisis Kulldorf detectamos un clúster adyacente a la Avenida Castellana, por lo que creamos un dataframe con puntos a lo largo de la misma y buscamos la distancia mínima a esta siguiendo la misma metodología que con hospitales y metros. Imputamos TRUE a cada uno de los registros que estaba a menos de 400 metros de alguno de los puntos de la avenida, obteniendo un total de 256 positivos.

Las tres variables que implican cálculo de distancia han pasado además una doble comprobación, basada en calcular la diferencia entre la distancia obtenida con "distm" y "law of cosines" y un cálculo manual.

En cuanto al resto de variables con componente espacial, observamos Distance, se trata de la distancia entre la vivienda y Puerta del Sol (forma parte del dataset de Idealista), "ventas", que informa sobre el número de ventas por distrito en el ejercicio 2019, "poblacionextranjeros" que hace lo propio con el número de extranjeros por distrito y "desempleomujeres", que refleja la tasa de desempleo en mujeres en cada distrito. En un principio se barajaron otras variables como "poblacionnacionales", aunque se descartaron por estar parcialmente incluidas en otras (poblacionextranjeros en este caso) y por lo tanto añadir ruido al modelo, y por tener un p-valor no aceptable.

La variable "norte" imputa TRUE a todos los valores de "latitude" superiores a 40.403030, dado que una vez más según Kulldorf detectamos una distribución irregular de residuos entre norte y sur.

El conjunto de variables tenidas en cuenta para este segundo modelo es, por lo tanto:

Price ~ distancia_km_metro + distancia_km_hospital +size2+size3+ Distance + size + rooms + bathrooms + status + hasLift + ventas + poblacionextranjeros + desempleomujeres + floorotro + distancia + norte

y los coeficientes obtenidos los siguientes:

```
## Call:
## glm(formula = formula, family = "gaussian", data = datos)
##
## Deviance Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -232318  -37490   -3324   32396  282533
##
## Coefficients:
##                            Estimate   Std. Error  t value  Pr(>|t|)
## (Intercept)                231167.4    1853.1     124.744  < 2e-16 ***
## distancia_km_metro         2617.6      780.0      3.356    0.000795 ***
## distancia_km_hospital      -5360.9     795.7      -6.738   1.74e-11 ***
## size2                      -7229.4     691.9      -10.449  < 2e-16 ***
## size3                      1971.4      416.6      4.732    2.26e-06 ***
## distance                   -33024.8    956.1      -34.542  < 2e-16 ***
## size                       52344.7     1563.1     33.488   < 2e-16 ***
## rooms                      -14774.7    944.1      -15.650  < 2e-16 ***
## bathrooms                  18450.0     915.2      20.159   < 2e-16 ***
## status                     1403.4      708.0      1.982    0.047489 *
## hasLift                    20158.6     789.2      25.544   < 2e-16 ***
## ventas                     15671.1     1286.2     12.184   < 2e-16 ***
## poblacionextranjeros       -16060.0    2191.0     -7.330   2.57e-13 ***
## desempleomujeres           -30722.2    2067.6     -14.859  < 2e-16 ***
## floorotro                  -30030.2    1876.2     -16.006  < 2e-16 ***
## distanciaTRUE              25984.4     3798.6     6.840    8.60e-12 ***
## norteTRUE                  15246.8     2557.8     5.961    2.64e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3237284133)
##
##     Null deviance: 8.1577e+13  on 6640  degrees of freedom
## Residual deviance: 2.1444e+13  on 6624  degrees of freedom
## AIC: 164290
##
## Number of Fisher Scoring iterations: 2
```

donde observamos que todos los p-valores son muy pequeños y por lo tanto las variables significativas.

En este caso obtenemos un AIC de 164290 y un RMSE de 56824.27

Realizada la prueba de iMoran comprobamos que el estadístico es de 12.015 y significativo.

```
Moran I test under randomisation

data:  xx$response
weights: nb2listw(knn2nb(knn))
Moran I statistic standard deviate = 12.015, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic             Expectation          Variance
```

0.322193735 -0.002020202 0.000728190

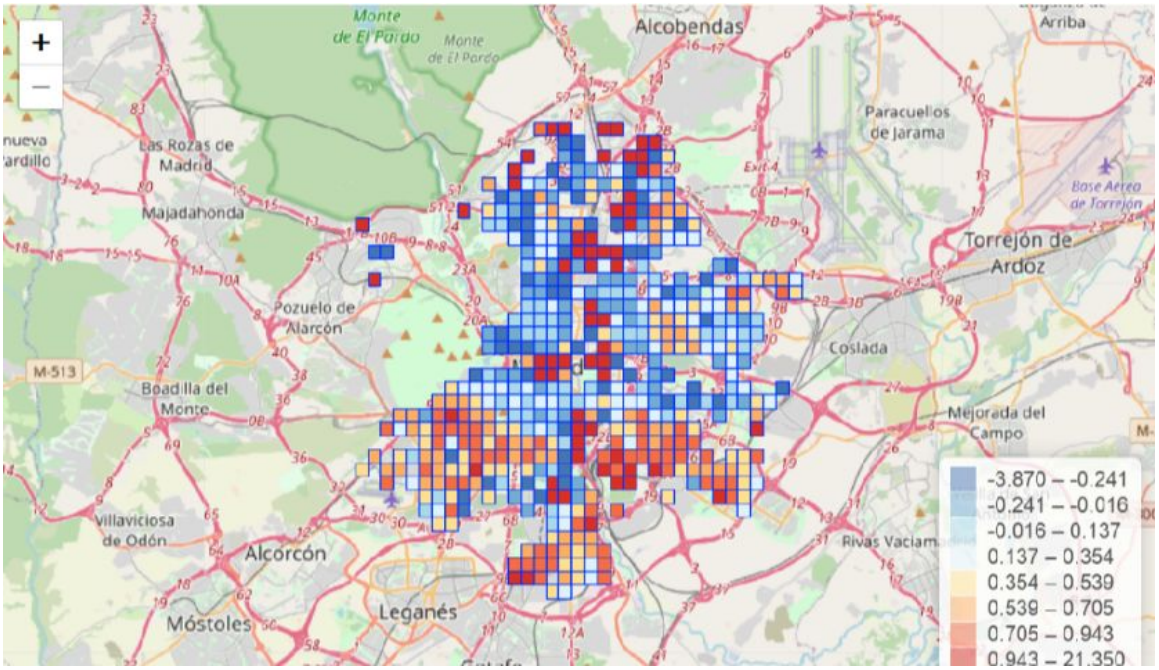


Gráfico 29: Raster 2

y un análisis Kulldorf nos indica que los clústeres no desaparecen, pero se han desplazado hacia la zona norte.

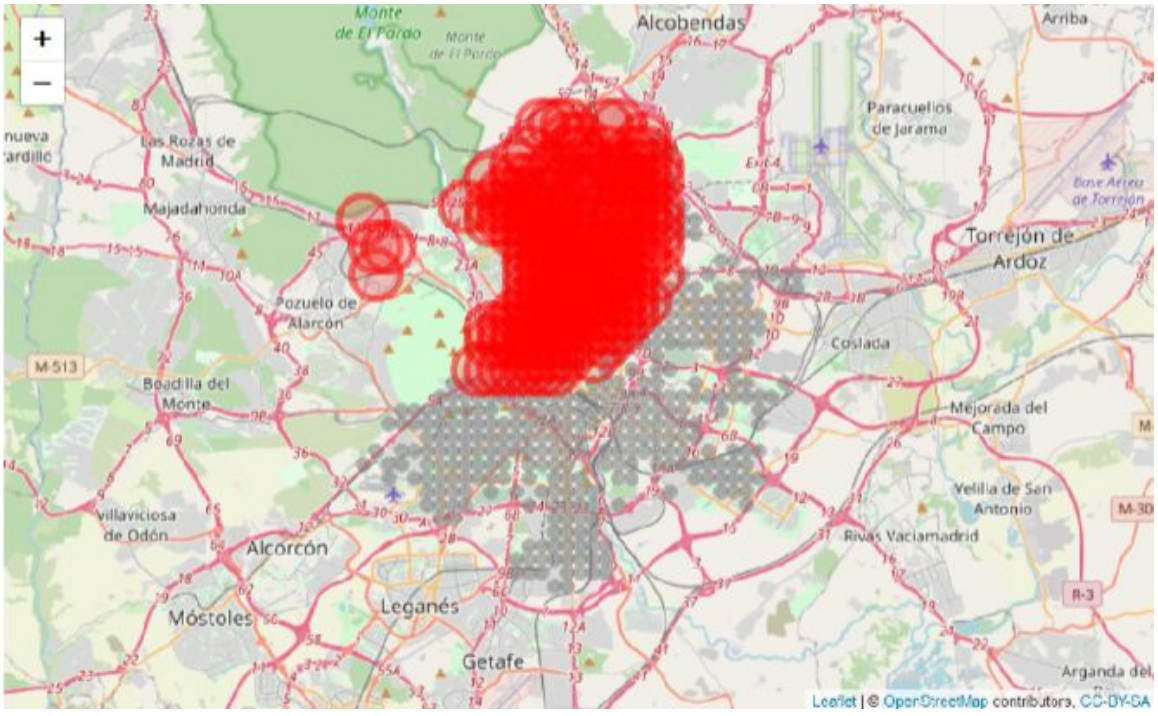


Gráfico 28: Kulldorf *2*

3.- SAR con las variables significativas en modelos anteriores, incluidas espaciales

En este caso incluimos las siguientes variables, sobre las que ya se ha hablado previamente:

nb <- knn2nb(knearneigh(cbind(datos$longitude, datos$latitude), k=5))

Price ~ distancia_km_metro +size2+size3+ distancia_km_hospital + Distance + size + rooms + bathrooms + status + hasLift + ventas + poblacionextranjeros + desempleomujeres + floorotro + distancia + norte

y obtenemos los siguientes coeficientes:

```
##Call:spatialreg::lagsarlm(formula = formula, data = data, listw = listw,
##     na.action = na.action, Durbin = Durbin, type = type, method = method,
##     quiet = quiet, zero.policy = zero.policy, interval = interval,
##     tol.solve = tol.solve, trs = trs, control = control)

##Residuals:
##       Min     1Q     Median         3Q      Max
##-225648.5 -31502.7  -2234.1   28687.1  269984.0

##Type: lag
##Coefficients: (numerical Hessian approximate standard errors)
##                          Estimate  Std. Error  z value  Pr(>|z|)
##(Intercept)               147482.54    2608.80  56.5327  < 2.2e-16
##distancia_km_metro           423.60     433.09   0.9781   0.328024
##size2                      -7243.36     612.60 -11.8239  < 2.2e-16
##size3                       1854.91     366.72   5.0581  4.235e-07
##distancia_km_hospital      -3280.02     669.47  -4.8994  9.611e-07
##distance                  -22870.67     853.14 -26.8077  < 2.2e-16
##size                       46857.10    1387.99  33.7589  < 2.2e-16
##rooms                      -9337.86     848.90 -10.9999  < 2.2e-16
##bathrooms                  15447.35     815.05  18.9527  < 2.2e-16
##status                       408.55     407.42   1.0028  0.315963
##hasLift                    14382.98     715.40  20.1049  < 2.2e-16
##ventas                     13076.08    1106.79  11.8144  < 2.2e-16
##poblacionextranjeros      -10761.77    1957.11  -5.4988  3.824e-08
##desempleomujeres          -19547.22    1869.89 -10.4537  < 2.2e-16
##floorotro                 -28910.75    1668.73 -17.3250  < 2.2e-16
##distanciaTRUE              17541.93    3375.42   5.1970  2.026e-07
```

##norteTRUE 5625.25 2139.08 2.6298
0.008545

##Rho: 0.3939, LR test value: 1358.3, p-value: < 2.22e-16
##Approximate (numerical Hessian) standard error: 0.0098148
z-value: 40.133, p-value: < 2.22e-16
##Wald statistic: 1610.7, p-value: < 2.22e-16

##Log likelihood: -81447.84 for lag model
##ML residual variance (sigma squared): 2559300000, (sigma: 50590)
##Number of observations: 6641
##Number of parameters estimated: 19
##AIC: 162930, (AIC for lm: 164290)

observando un pvalor demasiado alto en la variable “status” y distancia_km_metro, lo que nos podría llevar a descartarlas para siguientes modelos dado que únicamente aportan ruido dada su baja fiabilidad, y obteniendo un RSME de 58319.8. El valor de Rho, indicador de la dependencia espacial, es de 0.3939.

En este tercer caso el valor de iMoran tiene un valor significativo de 22.64

Moran I test under randomisation

data: xx$response
weights: nb2listw(knn2nb(knn))
Moran I statistic standard deviate = 22.648, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic Expectation Variance
0.5833773223 -0.0018214936 0.0006676259

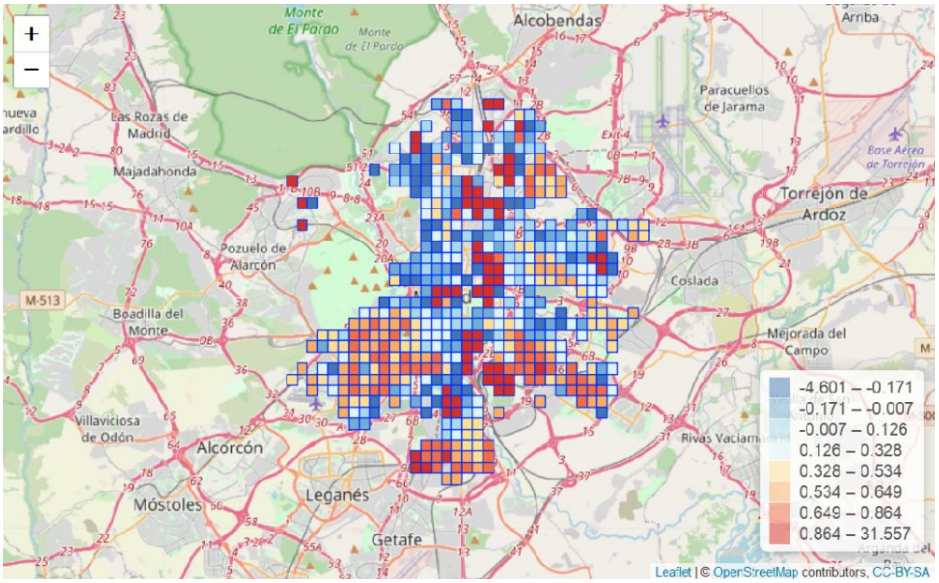


Gráfico 28: Raster 3

y al probar Kulldorf sobre estos nuevos residuos comprobamos que todos los clústeres espaciales han desaparecido, salvo uno ínfimo al norte de Pozuelo de Alarcón, y el p-valor pasa a ser 1, por lo que no significativo.

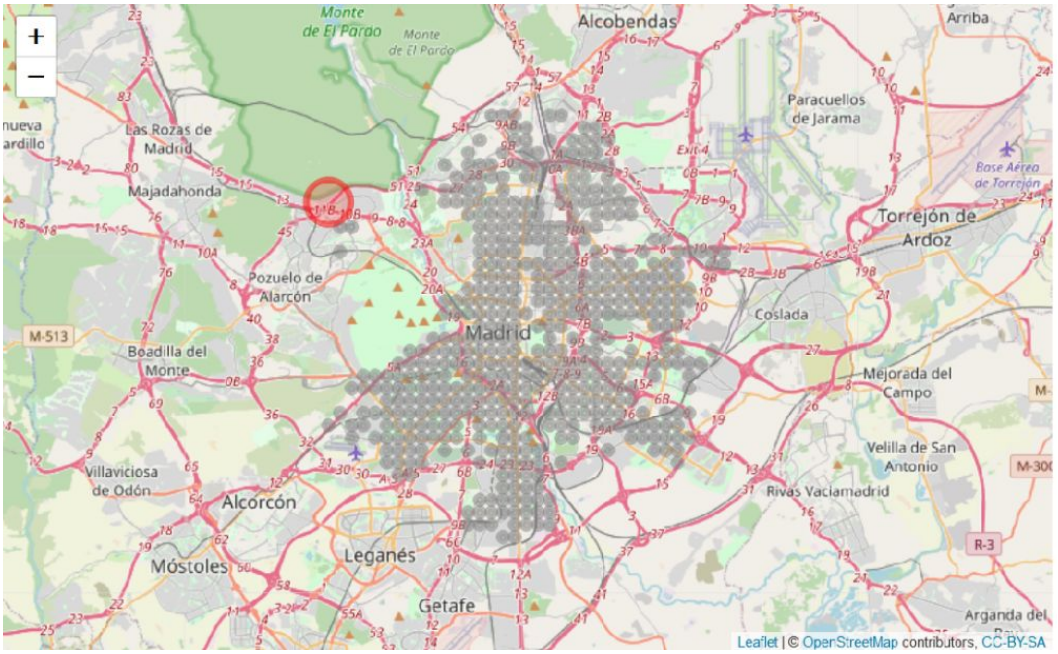


Gráfico 28: Kulldorf *3*

Habremos conseguido por lo tanto eliminar la heterocedasticidad espacial, dada la ausencia de clústeres, aunque como hemos podido comprobar en iMoran la dependencia espacial sigue

existiendo y no es significativamente inferior a la que conseguíamos con GLM añadiendo valores espaciales.

8. Resultados obtenidos y restricciones del modelo/modelos presentados.

GLM1	glm, incluye únicamente variables sin componente espacial
Resultados	AIC: 165817 RMSE: 63803.97 Porcentaje de error: 27% iMoran: 16.038 con p-valor == ~0 Kulldorf: Se detecta clúster espacial en el centro de Madrid, lo que implicaría que estaríamos infravalorando toda esta zona en pro del resto de zonas del municipio. p-valor: 0.01
Limitaciones	Este modelo no tiene en cuenta ningún factor espacial, razón por la cual existe un cluster espacial muy grande en el centro de Madrid, donde estaríamos infraestimando el precio de la vivienda.

GLM2	glm, incluye tanto variables con componente espacial como sin él
Resultados	AIC: 164290 RMSE: 56824.27 Porcentaje de error: 24% iMoran: 12.015 con p-valor == ~0 Kulldorf: Sigue existiendo clúster significativo (p-valor 0.01) pero esta vez más al norte.
Limitaciones	En este caso el clúster no desaparece y sigue siendo significativo, pero se encuentra más al norte. La dependencia espacial se reduce y sigue siendo significativa.

SAR	lagsarlm, incluye tanto variables con componente espacial como sin él
Resultados	AIC: 162930 RMSE: 58319.8 iMoran: 12.02 con p-valor: ~0 Rho: 0.3939 Porcentaje de error: 25% Kulldorf: Clúster minúsculo y no significativo (p-valor == 1) en el norte de Pozuelo de Alarcón

Limitaciones	Los valores espaciales se introducen tanto mediante la matriz de adyacencia listw como intrínsecos en las variables “de distrito” y las que hacen referencia directa a distancias a puntos. Gracias a esto conseguimos eliminar el número de clústeres y los errores del modelo pasarían a ser un “ruido blanco” y por lo tanto estarían distribuidos de forma aleatoria. La dependencia espacial no desaparece, si bien queda abierta la posibilidad de incluir nuevas variables que tengan en cuenta este aspecto de cara a mejorarlo en futuros modelos. Aun así, gracias a haber recalculado los coeficientes con lagsarlm tenemos unas betas más robustas.

Resumen de métodos utilizados:

Podríamos clasificar los métodos utilizados entre los de predicción (GLM, SAR), y los de evaluación de modelo (Kulldorf, iMoran, RMSE)

En este caso, las **funciones GLM** aplicadas se llevaron a cabo tras varias pruebas con algoritmos tipo **regresión lineal, SVM (regresión), random forest (regresión) y gradient boosting sin incluir en ellos componentes espaciales**. Estos cuatro métodos ofrecían una precisión demasiado baja y fomentaron la búsqueda de una nueva herramienta, fácilmente explicable y más precisa que las anteriores.

GLM es un modelo lineal que estima las betas mediante máxima verosimilitud, no obstante, aunque sea lineal, en nuestro caso permitimos que sea sensible a puntos de inflexión gracias al uso de splines. La especial atención a la significatividad de cada variable y la fácil interpretabilidad de esta metodología la convierten en una buena base tanto como función predictora como método para escoger variables para ser utilizados en otros modelos.

Por otra parte, y ante las carencias de GLM a la hora de eliminar los clústeres detectados mediante Kulldorf, proponemos un **modelo SAR** permite realizar la misma predicción, pero en este caso teniendo en cuenta una matriz de adyacencia según la cual ajustamos las variables explicativas dependiendo de la relación de este punto con las observaciones vecinas. Se trata de un modelo puramente espacial.

Siguiendo la siguiente fórmula:

$$Y^* = \rho W Y^* + X\beta + \varepsilon$$

$$\varepsilon \sim N(0, I_N)$$

SAR parte de la premisa de que el precio de los n vecinos más cercanos, planteados en W (matriz de adyacencia), afectan de algún modo al precio de la observación estudiada, siendo este un método capaz de modelar la interdependencia entre las medidas cercanas.

Devuelve una serie de coeficientes acompañados de sus p-valores, de forma que podemos comprender cómo afecta cada variable a nuestra variable dependiente, si bien tiene en cuenta para reestimar las betas la adyacencia de nuestra vivienda al resto de viviendas, partiendo de la premisa de que el precio de unas se ve afectado por el precio de sus vecinas. Además de otros estadísticos, nos indica el parámetro Rho, que indica la existencia de dependencia espacial, aunque en este caso llegamos a obtener un pvalor en Kulldorf de 1, por lo que los clústeres espaciales dejan de ser significativos.

Para la evaluación de nuestro modelo hemos utilizado el **indicador RMSE**, root-mean-square error, que calcula la media de las diferencias (real menos predicho) al cuadrado. Se trata de una forma de evaluar la precisión del modelo, y tiene una relación directa con las unidades en las que se expresa la unidad predicha (euros, en este caso), por lo que es más fácil de interpretar que un Criterio Akaike (AIC), por ejemplo.

En la evaluación también hemos medido, dado que lo ofrecen de forma automática los modelos, dicho **AIC**, pero en este caso únicamente nos da una indicación sobre si un modelo es mejor que otro (menor AIC implica mejor modelo).

El **índice de I-Moran**, por su parte, nos ofrece una medida de la autocorrelación espacial a lo largo del mapa. En nuestro caso lo evaluamos sobre los resíduos, para conocer si estos se encuentran agrupados, dispersos o son aleatorios. En nuestro caso, siempre obtenemos un iMoran positivo y con p-valor infinitesimal, por lo que dichos residuos están más agrupados de lo que cabría esperar de existir una aleatoriedad total, aceptamos por lo tanto que existe una dependencia espacial.

Dicho índice se calcula siguiendo una metodología muy similar a la utilizada por el modelo SAR, aunque esta vez con un objetivo de observación y control de resultados en lugar de como algoritmo de modelizado. En este caso seguimos la ecuación:

$$Y^* = \rho W Y^* + X\beta + WX\delta + \varepsilon$$

$$\varepsilon \equiv N(0, I_n)$$

para determinar si existe o no dependencia espacial teniendo en cuenta una vez más un número determinado de vecinos con los que conformamos una matriz de adyacencia.

Por otra parte, el **método Kulldorf** trata de clusterizar las diferentes observaciones. En nuestro caso, por ejemplo, los dos métodos Kulldorf que evalúan los residuos de los GLM1 y GLM2, muestran dos grandes clústeres que pasan de todo el centro de Madrid a ligeramente al norte cuando comenzamos a imputarle nuevos datos. Esto implicaría que nuestro modelo infravalora a la hora de predecir en estas zonas. En el caso del modelo SAR, Kulldorf desaparece (únicamente aparece en un pequeño punto en Pozuelo de Alarcón) y demás tiene un p-valor que descarta su validez y por lo tanto desaparece la heterocedasticidad espacial en el modelo.

9. Cuadro de mando (Data-Driven-Decision-Making).

El cuadro de mando – *dashboard* en inglés – le proporciona al usuario una visión general y visualmente más atractiva sobre los datos recolectados, así como la posibilidad de examinar un análisis exploratorio simple sobre esos. Es una herramienta, que, en este caso, también le admite al usuario introducir sus propios parámetros de una vivienda en Madrid para poder ejecutar un análisis predictivo a partir de nuestro modelo y así determinar un valor de precio sobre esa vivienda. El resultado es el precio en euros que el modelo estadístico sugiere debería tener esa vivienda con esos parámetros y características.

La estructura del cuadro de mando es la siguiente: tiene un menú principal a mano izquierda con tres tabs principales: *Parámetros*, *Análisis*, *Idealista*. También hay una sección del menú llamado *About* con un tab de *About Us* y un link que lleva al repositorio GitHub del proyecto.

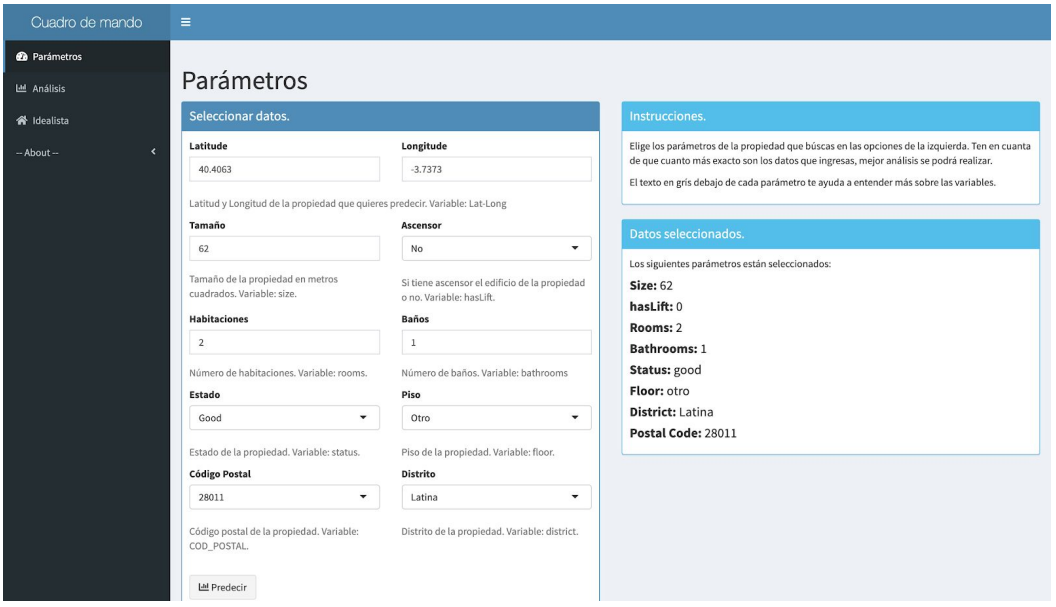


Imagen Dashboard 1. Primer tab Parámetros.

En el primer tab *Parámetros* se encuentra un pequeño formulario donde el usuario puede introducir los datos sobre la propiedad que desea analizar. Al hacer clic en el botón de *Predecir* el usuario visualizará el tab de *Análisis*.

El segundo tab *Análisis* está dividido en tres secciones. En la primera podrá observar el resultado del análisis predictivo, así como el RMSE. En la segunda visualizará un mapa interactivo de la posición de su propiedad buscada, así como otros puntos interesantes cercanos, como por ejemplo entradas a estaciones de metro y hospitales. En la tercera sección tiene a disposición el usuario una tabla de registro sobre todas las búsquedas que ha realizado durante su sesión.

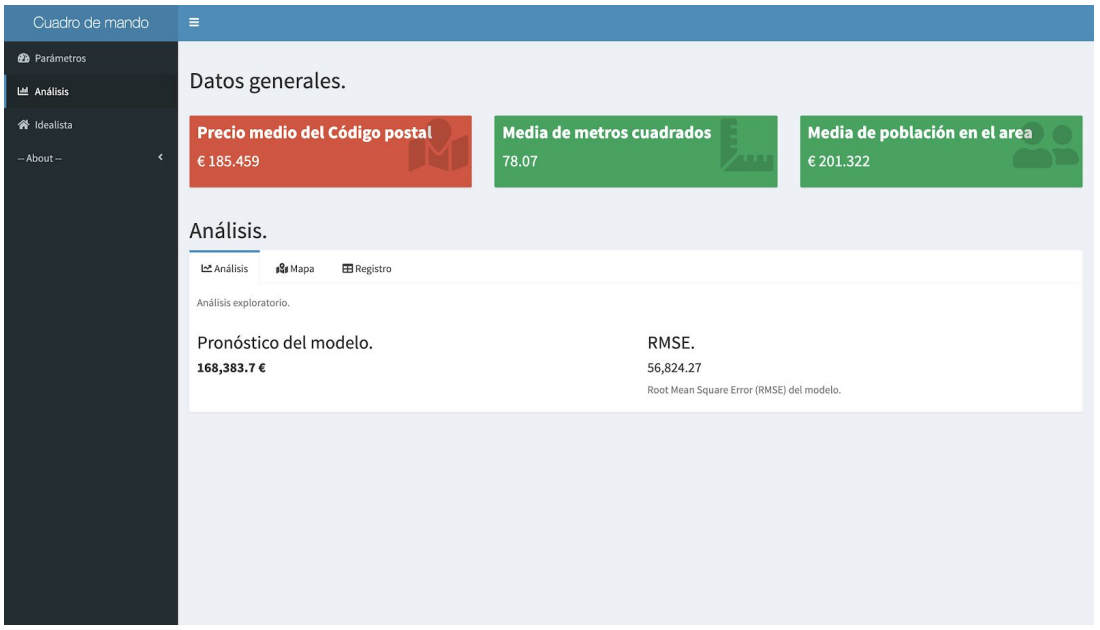


Imagen Dashboard 2. Segundo tab Análisis.

En el tercer tab *Idealista*, el usuario podrá examinar un breve análisis exploratorio sobre los datos utilizados para crear el modelo, una pequeña gráfica ilustrando la correlación entre las variables (parámetros) así como una tabla interactiva con todos los registros reales de propiedades.

Imagen Dashboard 3. Tercer tab Idealista.

10. Conclusiones y propuesta de trabajo futuro.

La conclusión principal de este trabajo gira en torno a la importancia de analizar el componente espacial a la hora de tratar de predecir el precio de la vivienda tanto en la ciudad de Madrid como, presumiblemente, en el resto de las ciudades del mundo.

Si bien la dependencia espacial sigue existiendo incluso tras el paso último de nuestra investigación, hemos conseguido minimizar las ocasiones en las que el modelo infraestima el valor de un determinado clúster de viviendas a juzgar por el último análisis Kulldorf.

Esto deja varias vías de trabajo abiertas a nivel de investigación:

Por una parte, el modelo propuesto todavía puede ser mejorado añadiendo más variables significativas basadas en distancias a puntos de la ciudad o demográficas, en busca de mejorar los estadísticos espaciales como la precisión en sí.

Por otra parte, una nueva vía de trabajo podría consistir en mejorar el proceso de ingesta. Más allá de decidir entre ingestar mediante API o scrapping, lo más interesante puede ser explorar nuevas fuentes de datos estructuradas de forma menos homogéneas, como Milanuncios o Facebook, en busca de oportunidades que escapen a las ingestas de otros proyectos similares.

Un primer paso podría ser, por ejemplo, construir un web scrapper que recorra una página como Milanuncios en busca de propiedades en Madrid, y mediante una inteligencia artificial basada en redes neuronales, detecte los diferentes datos interesantes para el modelado (número de habitaciones, metros, precio...) teniendo en cuenta tanto los campos que Milanuncios ha previsto para estos datos, como la propia descripción que habrá sido escrita por diversos vendedores y, por lo tanto, con diferentes estilos y criterios.

11. Bibliografía utilizada para realización del TFM.

El estilo de citación y referencias utilizada en este trabajo es de *American Psychological Association* (APA).

A continuación, ilustramos las referencias bibliográficas y páginas webs consultadas:

Bibliografía

Statista Research Department. (2020, Enero 10). Industria de la construcción: aportación al PIB 2005-2017. Recuperado Mayo 9, 2020, de https://es.statista.com/estadisticas/549605/aportacion-del-sector-de-la-construccion-al-pib-en-espana/

Statista Research Department. (2020, Enero 10). Actividades inmobiliarias: número de trabajadores 2005-2016. Recuperado Mayo 9, 2020, de https://es.statista.com/estadisticas/526311/numero-de-empleados-del-sector-inmobiliarias-en-espana/

Inmodiario. (2020, Marzo 31). El precio de la vivienda caerá un 10% en España por la crisis del Covid-19. Recuperado Mayo 9, 2020, de https://www.inmodiario.com/96/28847/precio-vivienda-caera-espana-crisis-covid.html

Inmodiario. (2020, Marzo 31). El precio de la vivienda caerá un 10% en España por la crisis del Covid-19. Recuperado Mayo 9, 2020, de https://www.inmodiario.com/96/28847/precio-vivienda-caera-espana-crisis-covid.html

Expansión. (2020, Enero 31). PIB de España - Producto Interior Bruto 2019. Recuperado Marzo 27, 2020, de https://datosmacro.expansion.com/pib/espana

Libre Mercado. (2019, Septiembre 30). España se desacelera: el PIB crece a su menor ritmo en tres años. Recuperado Marzo 28, 2020, de https://www.libremercado.com/2019-09-30/espana-se-desacelera-el-pib-crece-a-su-menor-ritmo-en-tres-anos-1276645488/

EL PAÍS, (2019, Noviembre). Resultados Electorales en Total España: Elecciones Generales 2019. Recuperado Marzo 27, 2020, de https://resultados.elpais.com/elecciones/2019/generales/congreso/

Bonet, E. (2019, Septiembre 6). "Si la Unión Europea no cambia de rumbo, la ultraderecha seguirá creciendo". Recuperado Marzo 30, 2020, de https://www.publico.es/politica/auge-ultraderecha-union-europea-no-cambia-rumbo-ultraderecha-seguira-creciendo.html

ESADE. (2020, Marzo 13). Política económica contra el coronavirus: impacto y respuestas para España. Recuperado Marzo 27, 2020, de https://dobetter.esade.edu/es/coronavirus-politica-economica

Rodrigo, N. (2019, Enero 2). Madrid, el municipio que más creció en 2018, con 40.000 habitantes más. Recuperado Marzo 27, 2020, de https://cincodias.elpais.com/cincodias/2019/01/02/economia/1546432956_524849.html

El Economista. (2019, Octubre 4). Así es la España vacía: 12 gráficos para entender el problema de la despoblación en nuestro país. Recuperado Abril 11, 2020, de https://www.eleconomista.es/economia/noticias/10120949/10/19/Asi-es-la-Espana-vacia-12-graficos-para-entender-el-problema-de-la-despoblacion-en-nuestro-pais.html

Sanz, E. (2020, Febrero 23). La vivienda se enfría: los vendedores recortan hasta un 36% sus expectativas de precio. Recuperado Marzo 27, 2020, de https://www.elconfidencial.com/vivienda/2020-02-23/vivienda-precios-mercado-residencial-maximos_2461336/

Salvador, R. (2019, Diciembre 16). Madrid y Barcelona ganan población y colapsan el mercado de la vivienda. Recuperado Abril 10, 2020, de https://www.lavanguardia.com/economia/20191215/472233133386/precios-vivienda-barcelona-madrid-crecimiento-poblacion-venta-alquiler.html

Ayuntamiento de Madrid. (2019, Junio 27). Encuesta Continua de Hogares – Hogares.xlsx (1.3). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-

de-informacion-estadistica/Demografia-y-poblacion/Cifras-de-poblacion/Encuesta-Continua-de-Hogares/?vgnextfmt=default&vgnextoid=0ccf7bfbd989b610VgnVCM2000001f4a900aRCRD&vgnextchannel=a4eba53620e1a210VgnVCM1000000b205a0aRCRD

Ayuntamiento de Madrid. (2019, Junio 27). Encuesta Continua de Hogares – Hogares.xlsx (1.9). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Demografia-y-poblacion/Cifras-de-poblacion/Encuesta-Continua-de-Hogares/?vgnextfmt=default&vgnextoid=0ccf7bfbd989b610VgnVCM2000001f4a900aRCRD&vgnextchannel=a4eba53620e1a210VgnVCM1000000b205a0aRCRD

Ayuntamiento de Madrid. (2019, Junio 27). Encuesta Continua de Hogares – Hogares.xlsx (1.4). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Demografia-y-poblacion/Cifras-de-poblacion/Encuesta-Continua-de-Hogares/?vgnextfmt=default&vgnextoid=0ccf7bfbd989b610VgnVCM2000001f4a900aRCRD&vgnextchannel=a4eba53620e1a210VgnVCM1000000b205a0aRCRD

Ayuntamiento de Madrid. (2018, Abril 23). Distritos en cifras (Información de Distritos). Recuperado Marzo 27, 2020, de https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Distritos-en-cifras/Distritos-en-cifras-Informacion-de-Distritos-/?vgnextfmt=default&vgnextoid=74b33ece5284c310VgnVCM1000000b205a0aRCRD&vgnextchannel=27002d05cb71b310VgnVCM1000000b205a0aRCRD

Idealista. (2020, Marzo). Evolución del precio de la vivienda en venta en Villaverde. Recuperado Abril 11, 2020, de https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/madrid-comunidad/madrid-provincia/madrid/villaverde/

Instituto Nacional de Estadística (INE). (2020, Marzo 16). Estadística de Transmisiones de Derechos de la Propiedad. Recuperado Abril 10, 2020, de https://www.ine.es/jaxiT3/Datos.htm?t=6150#!tabs-grafico

Ayuntamiento de Madrid. (n.d.). Inicio. Recuperado Abril 10, 2020, de https://www.madrid.es/portal/site/munimadrid

Portal Estadístico. (2020, Marzo). Explotación del Padrón; Resto de municipios: INE, Revisión del Padrón. Recuperado Abril 10, 2020, de

http://portalestadistico.com/municipioencifras/default.aspx?pn=madrid&pc=ZTV21&idp=35&idpl=1329&idioma=

Idealista. (2020, Marzo). Evolución del precio de la vivienda en venta en Villaverde. Recuperado Abril 10, 2020, de https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/madrid-comunidad/madrid-provincia/madrid/villaverde/

Instituto Nacional de Estadística (INE). (2019). Total Nacional. Datos ajustados de estacionalidad y calendario. Producto interior bruto a precios de mercado. Variación anual. Índices de volumen encadenados. Recuperado Abril 11, 2020, de https://www.ine.es/consul/serie.do?d=true&nocab&s=CNTR4892&nult=50

Bankinter. (2020, Febrero 7). La tabla definitiva para saber si un piso está caro o barato. Recuperado Abril 11, 2020, de https://blog.bankinter.com/economia/-/noticia/2015/12/10/valoracion-inmuebles-pisos-baratos-caros

Idealista. (2016, Julio 26). Madrid: descuentos por pedidos por los compradores de vivienda. Recuperado Abril 10, 2020, de https://www.idealista.com/news/estadisticas/descuentos-vivienda/venta-viviendas/madrid--comunidad-de/madrid

TerceroB. (2014, Mayo 25). 3BValue. Recuperado Abril 10, 2020, de https://www.tercerob.com/3BValue

BBVA. (2020, Abril 8). Descubre el precio medio de una zona con BBVA Valora. Recuperado Abril 11, 2020, de https://www.bbva.es/personas/experiencias/bbva-valora/analizar-barrio.html

Bankia. (2017, Enero 12). Calcula el valor de tu vivienda. Recuperado Abril 11, 2020, de https://www.bankia.es/es/particulares/financiacion/hipotecas/valorar-vivienda

Betterplace. (2020, Febrero 24). Madrid: informe inmobiliario del municipio: estado de compraventa y alquiler. Recuperado Mayo 10, 2020, de https://www.betterplaceweb.com/informe-inmobiliario-madrid-enero-2019/

Tabales, J. N., Caridad, J. M., Villamondos, N. C., & Jiménez, A. M. F. (2009, Febrero). Estimación del precio de la vivienda mediante redes neuronales artificiales (RNA) en diferentes marcos temporales. Recuperado Marzo 27, 2020, de http://casus.usal.es/pkp/index.php/MdE/article/view/994

Apapiu. (2017, Abril 21). Regularized Linear Models. Recuperado Abril 10, 2020, de https://www.kaggle.com/apapiu/regularized-linear-models

Pmarcelino. (2019, Agosto 23). Comprehensive data exploration with Python. Recuperado Abril 11, 2020, de https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

I. (Ed.). (2009, Octubre 4). Idealista Labs. Recuperado Mayo 20, 2020, de https://www.idealista.com/labs/

Idealista. (2019). Descuentos por distritos pedidos por los compradores de vivienda. Recuperado Abril 11, 2020, de https://www.idealista.com/news/estadisticas/descuentos-vivienda/venta-viviendas/distritos

Portal Estadístico. (2020, Marzo). Explotación del Padrón; Resto de municipios: INE, Revisión del Padrón. Recuperado Abril 10, 2020, de http://portalestadistico.com/municipioencifras/default.aspx?pn=madrid&pc=ZTV21&idp=35&idpl=1329&idioma=

Anexos.

Anexo al estado del Arte.

Distrito	Venta de Pisos	Valor m2	Población	% de Crecimiento Poblacional 2018-2019	% de Crecimiento de Valor m2 2018-2019	Descuento del m2
Arganzuela	1.807	4.031	153.830	0,6%	6%	20,90%
Barajas	971	3.181	48.972	2,4%	8%	19,80%
Carabanchel	3.628	2.189	253.040	1,9%	9%	19,50%
Centro	4.428	5.037	134.881	1,9%	3%	23,40%
Chamartín	2.753	5.041	145.865	0,7%	2%	18,80%
Chamberí	3.363	5.308	139.448	0,7%	3%	21,50%
Ciudad Lineal	4.062	3.064	216.270	0,8%	8%	17,50%
Fuencarral-El Pardo	3.371	3.472	246.021	1,3%	6%	19,20%
Hortaleza	3.639	3.702	188.267	2,4%	5%	20,10%
Latina	3.848	2.299	238.154	1,0%	6%	23,40%

Moncloa-Aravaca	1.876	3.933	119.423	1,3%	4%	25,30%
Moratalaz	1.286	2.533	94.609	0,4%	8%	20,90%
Puente de Vallecas	4.135	1.933	234.770	1,9%	11%	19,70%
Retiro	2.237	4.597	119.379	0,3%	4%	22,40%
Salamanca	2.992	5.844	146.148	0,6%	5%	23,80%
San Blas	3.465	2.519	158.166	1,3%	1%	18,20%
Tetuán	4.090	3.704	157.937	1,3%	8%	18,50%
Usera	1.782	2.044	139.501	1,8%	9%	22,10%
Vicálvaro	1.704	2.278	72.126	1,6%	11%	15,30%
Villa de Vallecas	2.631	2.411	110.436	2,6%	10%	15,60%
Villaverde	3.139	1.717	148.883	2,3%	8%	18,70%

Tabla 3: Descuentos por distritos pedidos por los compradores de vivienda.[34] [35]

[34] Idealista. (2019). Descuentos por distritos pedidos por los compradores de vivienda. Recuperado Abril 11, 2020, de https://www.idealista.com/news/estadisticas/descuentos-vivienda/venta-viviendas/distritos
[35] Portal Estadístico. (2020, Marzo). Explotación del Padrón; Resto de municipios: INE, Revisión del Padrón. Recuperado Abril 10, 2020, de http://portalestadistico.com/municipioencifras/default.aspx?pn=madrid&pc=ZTV21&idp=35&idpl=1329&idioma=