

Análisis de datos idealista18

David Recio Pastor

Preprocesado de datos y anomalías

El set elegido para esta prueba ha sido el de la ciudad de Valencia. Importamos los datos descargados del repositorio de Github y empezamos el proceso de los datos.

Los datos tienen una dimensiones de 33622 filas y 46 columnas. Las descripciones de las columnas las encontramos en el paper “A geo-referenced micro-data set of real estate listings for the three largest Spanish cities from the Idealista website”.

Tras una primera observación de los datos, observamos que hay 4 períodos, los cuatro trimestres de 2018, que convertimos de numero entero a formato fecha por conveniencia. La primera acción que llevamos a cabo es agrupar por “ASSETID”, para ver si tenemos varias observaciones de un mismo ASSETID en diferentes períodos temporales. No es el caso, tenemos 4476 “ASSETID” con más de una observación pero todas en el mismo período. El número de observaciones por ASSETID tiene la siguiente distribución:

1	22915
2	3144
3	1029
4	207
5	79
6	14
8	1
7	1
10	1

Dejamos a un lado lo que tienen 1 observación, y tratamos el resto. Para cada “ASSETID” con más de una observación, alistamos el resto de características, buscando si valores como precio, número de habitaciones u otros son iguales o difieren para un teóricamente mismo inmueble. Ponemos un ejemplo:

ASSETID	{A9651535568269959084}
PERIOD	{201812}
PRICE	{111 000, 116 000, 119 000}
UNITPRICE	{1480, 1348.84, 1383.72}
ADTYPOLOGYID	{HOME}
ADOPERATIONID	{SALE}
CONSTRUCTEDAREA	{75, 86}
ROOMNUMBER	{2}
BATHNUMBER	{1}
HASTERRACE	{1}
HASLIFT	{1}
HASAIRCONDITIONING	{1}
AMENITYID	{2}
HASPARKINGSPACE	{1}
ISPARKINGSPACEINCLUDEDINPRICE	{1}
PARKINGSPACEPRICE	{1}
HASNORTHORIENTATION	{0}
HASSOUTHORIENTATION	{1}
HASEASTORIENTATION	{1}
HASWESTORIENTATION	{0}

Como se observa, hay 3 valores de precio y dos de superficie. Si tuviésemos referencias temporales diarias, podríamos elegir la última por entender que el anuncio se puso incorrectamente y luego se corrigió pero, como no es el caso, la elección es descartar esta observación. La regla de decisión para descar-

tar observaciones dudosas es la siguiente:

- Para todas las características inmutables (superficie, terraza, ser duplex, orientacion, planta...) directamente descartamos las observaciones, ante la imposibilidad de conocer el valor real. Descartamos 826 ASSETID.
- Cuando el único valor con más de una observación es el precio, tomamos el mínimo, entendiendo que es el último que se ha puesto y supone una rebaja.
- Con respecto a la situación geográfica, para longitud o latitud hay 3650 observaciones con más de un valor. Para resolverlo, nos quedamos con la media de los valores que existan, al igual que para el resto de valores de distancias.

Hechas estas primeras correcciones, tenemos 26.555 filas. Hemos descartado 7067, un 21,02%.

Estos serían los descartes de filas, por el lado de las columnas o características, decidimos eliminar:

- ADTYPOLOGYID, ADOPERATIONID, ADTYPOLOGY, ADOPERATION y CITYNAME porque no aportan nada al análisis.
- PARKINGSPACEPRICE porque no tiene sentido, debería ser un rango de precios pero no son correctos, en la mayoría de casos el valor es 1.
- De las variables HASPARKINGSPACE Y ISPARKINGSPACEINCLUDEDINPRICE nos deberíamos quedar con sólo una, ya que los valores son idénticos en ambas.
- CONSTRUCTIONYEAR, el procedente del anunciante tiene 10533 observaciones sin año, por lo que usamos el procedente del catastro.

El resto de variables pueden tener interés, aunque no usaremos todas en nuestro análisis.

Estadística descriptiva

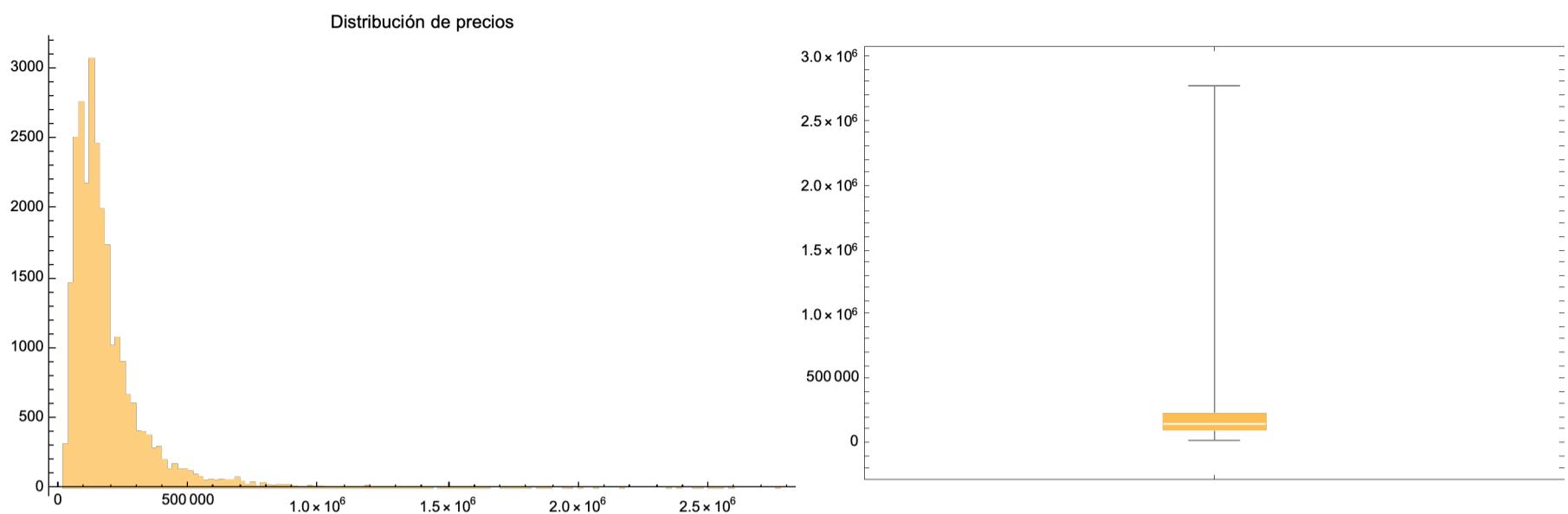
Tras la primera criba, empezamos con un análisis un poco más exhaustivo de nuestro dataset. Para ello, emplearemos diferentes métodos según características:

- Para variables continuas, crearemos una tabla de medidas descriptivas de la muestra, además de un histograma y diagrama de caja para tener una idea visual de la distribución.
- Para variables binarias, haremos un conteo y los pondremos de forma porcentual.
- Para variable ordinales, contaremos la cantidad de sucesos de cada valor y descartaremos para nuestro análisis global los menos relevantes (a nivel de generalizar).
- Las coordenadas geográficas las usaremos para situar a cada inmueble en su barrio, incluyéndolo en un cluster de la ciudad para su posterior análisis gráfico.

Variables continuas

PRICE - Precio absoluto

Mínimo	20 000
1er Cuartil	97 000
Mediana	148 000
Media	193 858.
3er Cuantil	227 000
Top 5%	491 000
Top 1%	889 000
Máximo	2 772 000
Std. Dev.	172 958.
Test de asimetría	4.18617
Kurtosis	33.7968

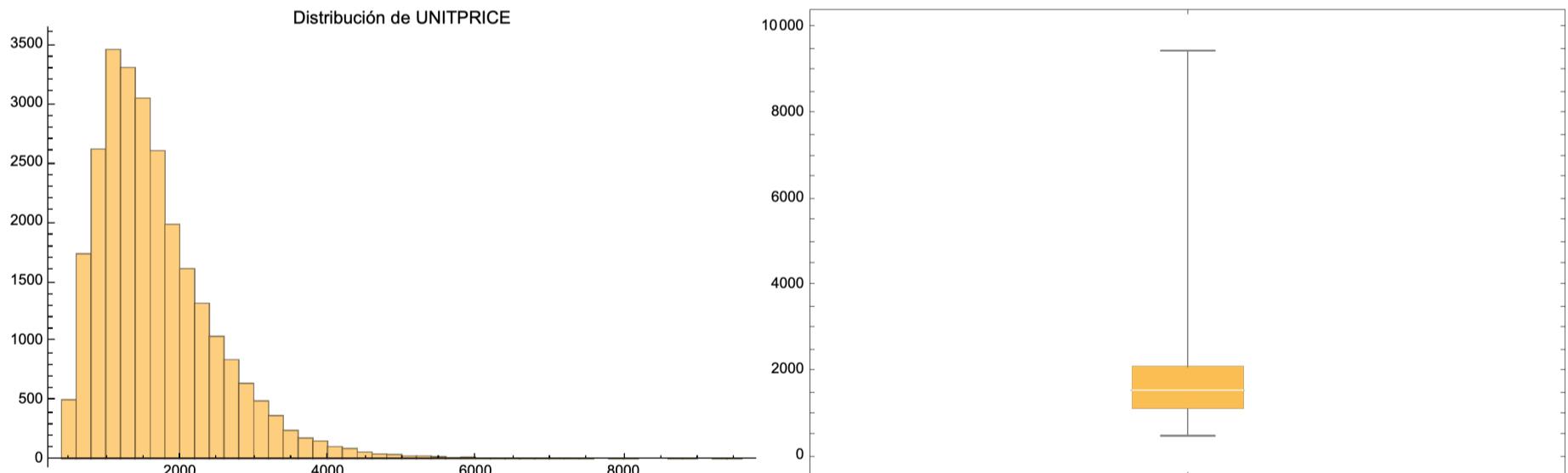


Vemos que hay valores atípicos que degeneran la distribución, produciendo una asimetría exagerada, podemos verlo en la diferencia entre media y mediana o, de forma aún más evidente, en los gráficos. Para solucionar este problema, podríamos descartar valores por encima del top 1% o 5%.

Para este caso concreto, lo normal es usar la variable normalizada por superficie, "UNITPRICE"

UNITPRICE - Precio por metro

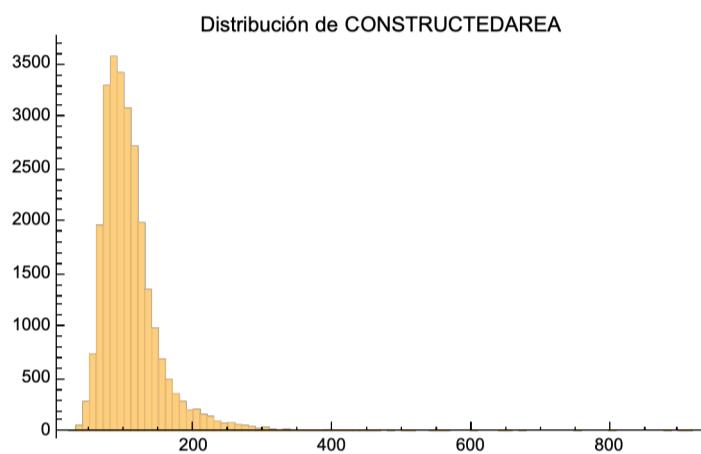
Mínimo	480.687
1er Cuartil	1098.21
Mediana	1506.41
Media	1687.73
3er Cuantil	2078.43
Top 5%	3241.67
Top 1%	4458.94
Máximo	9421.82
Std. Dev.	838.273
Test de asimetría	1.73528
Kurtosis	8.99827



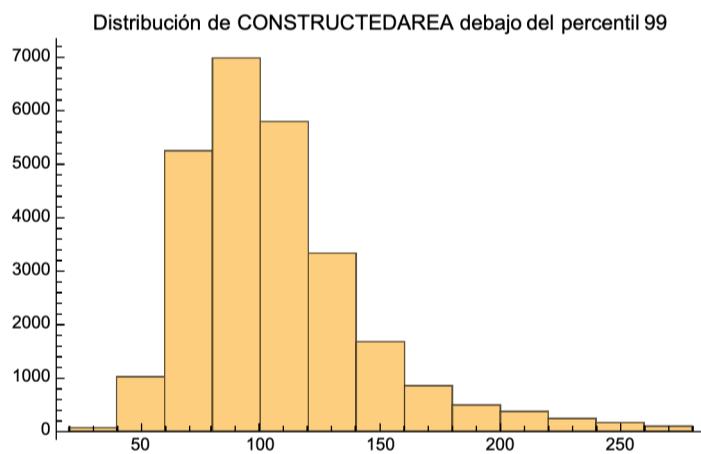
Como vemos, es mucho menos asimétrica, pero sigue teniendo algún valor atípico, ya que el máximo duplica su valor al top 1%, por lo que descartaremos las observaciones por encima de este top 1%.

CONSTRUCTED AREA

Mínimo	24
1er Cuartil	80
Mediana	99
Media	107.704
3er Cuantil	121
Top 5%	188
Top 1%	274
Máximo	912
Std. Dev.	46.6645
Test de asimetría	3.83181
Kurtosis	39.0155



Como vemos, tiene una distribución similar a la de precio absoluto, pero con mayor kurtosis, es decir, más concentración. Si eliminamos el último tanto por ciento de la muestra, nos quedaría una distribución mucho más simétrica, simplemente quitando un 1% de valores anómalos, que son 264 observaciones.



Variables discretas y categóricas

Variables binarias

Se presenta la distribución de cada tipo como proporción sobre el total, por ejemplo, hay un 75,11% de viviendas sin terraza.

0	0.751063
1	0.248937

1	0.788331
0	0.211669

1	0.465199
0	0.534801

1	0.164126
0	0.835874

0	0.876228
1	0.123772

1	0.530849
0	0.469151

0	0.932166
1	0.0678336

0	0.951289
1	0.0487107

0	0.944551
1	0.0554489

0	0.982759
1	0.0172407

0	0.994316
1	0.00568417

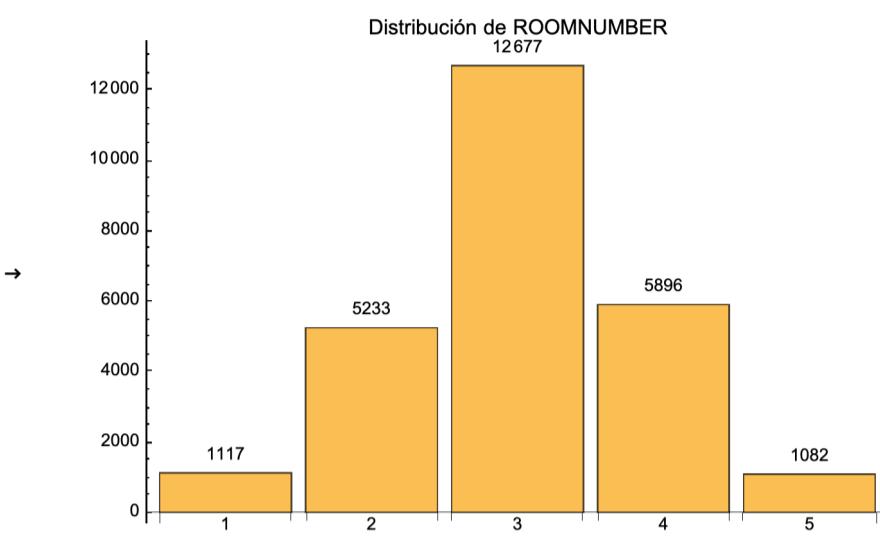
0	0.988142
1	0.0118577

Variables ordinales y categóricas

ROOMNUMBER

Agrupando las viviendas por número de habitaciones observamos que hay varios valores atípicos, un caso de una casa con 81 habitaciones que descartamos y varios valores que no van a tener ninguna relevancia a nivel de agrupación estadística. Por tanto, para nuestro estudio, vamos a quedarnos con las viviendas de 1 a 5 habitaciones.

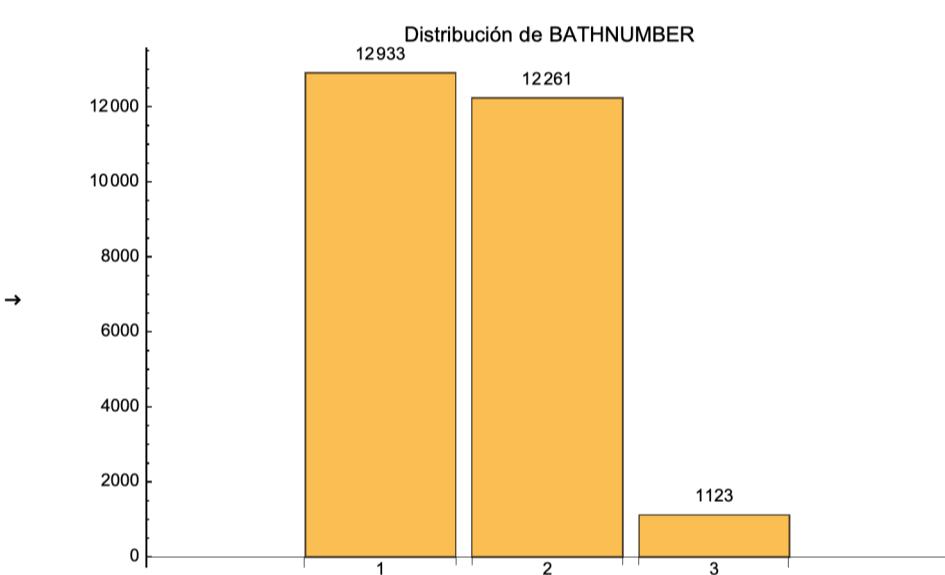
0	190
1	1117
2	5233
3	12677
4	5896
5	1082
6	272
7	69
8	10
9	5
10	8
11	3
15	2
81	1



BATHNUMBER

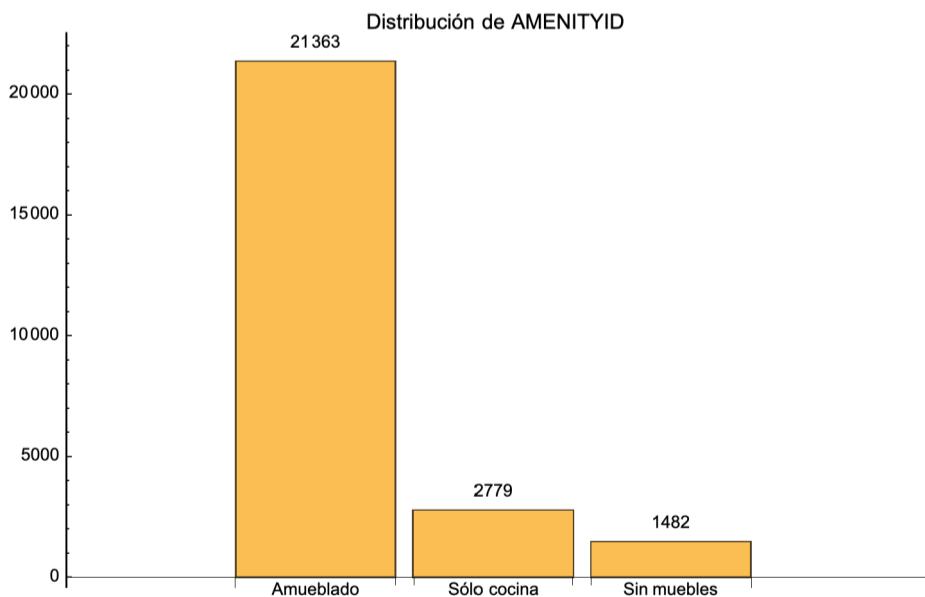
Usamos la misma metología y descartamos las viviendas con más de 3 baños y sin baños.

0	39
1	12933
2	12261
3	1123
4	170
5	25
6	6
7	2
8	2
10	3
12	1



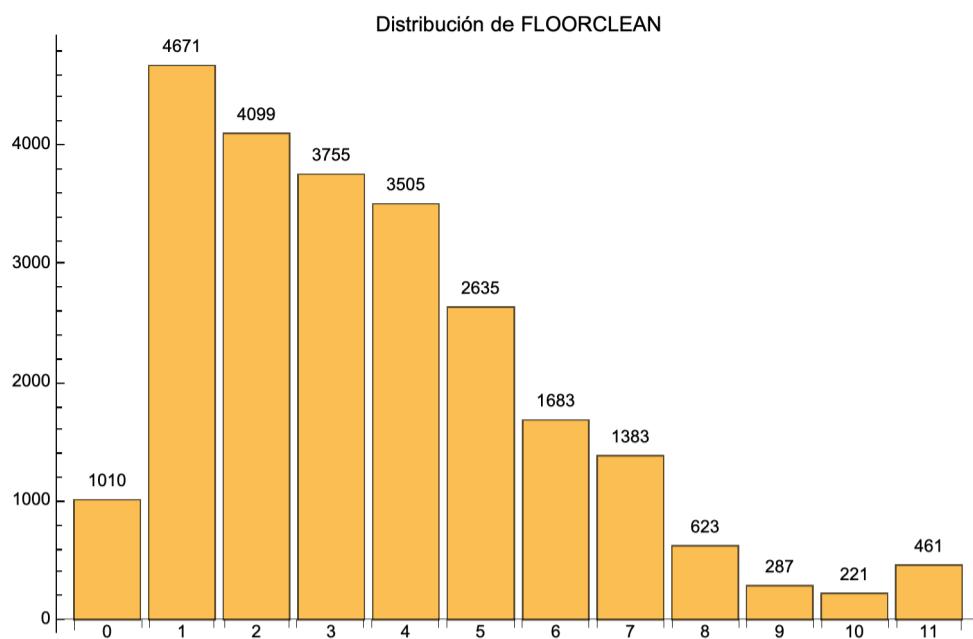
AMENITYID

Contamos el número de viviendas de cada tipo, observando que la gran mayoría de los inmuebles se venden amueblados por completo.



FLOORCLEAN

Esta variable indica la planta en la que se encuentra la vivienda, cabe mencionar que hay 1286 observaciones para las que no se tienen dato de planta, pero que no descartamos porque a priori no será una variable de diferenciación en nuestro modelo, pero si lo fuese deberían descartarse o aproximarse en función de la variable CADMAXBUILDINGFLOOR.



CADCONSTRUCTIONYEAR

Observar el año de construcción según el catastro es algo más curioso. Si hacemos una serie temporal agrupando número de viviendas por año construido vemos unas tendencias que podrían ser objeto de un estudio socioeconómico más profundo. Vemos que tenemos dos máximos locales en 1970 y en 2002 con 1213 y 416 viviendas respectivamente.

Mínimo	1591
1er Cuartil	1961
Mediana	1970
Media	1970.58
3er Cuantil	1982
Top 5%	2006
Top 1%	2012
Máximo	2018
Std. Dev.	23.5724
Test de asimetría	-1.51166
Kurtosis	13.1501



Habrá más variables para analizar en este apartado, pero he decidido pasar al siguiente por la falta de tiempo por mi parte y porque entiendo que no es tan valioso.

VARIABLES GEGRÁFICAS O ESPACIALES

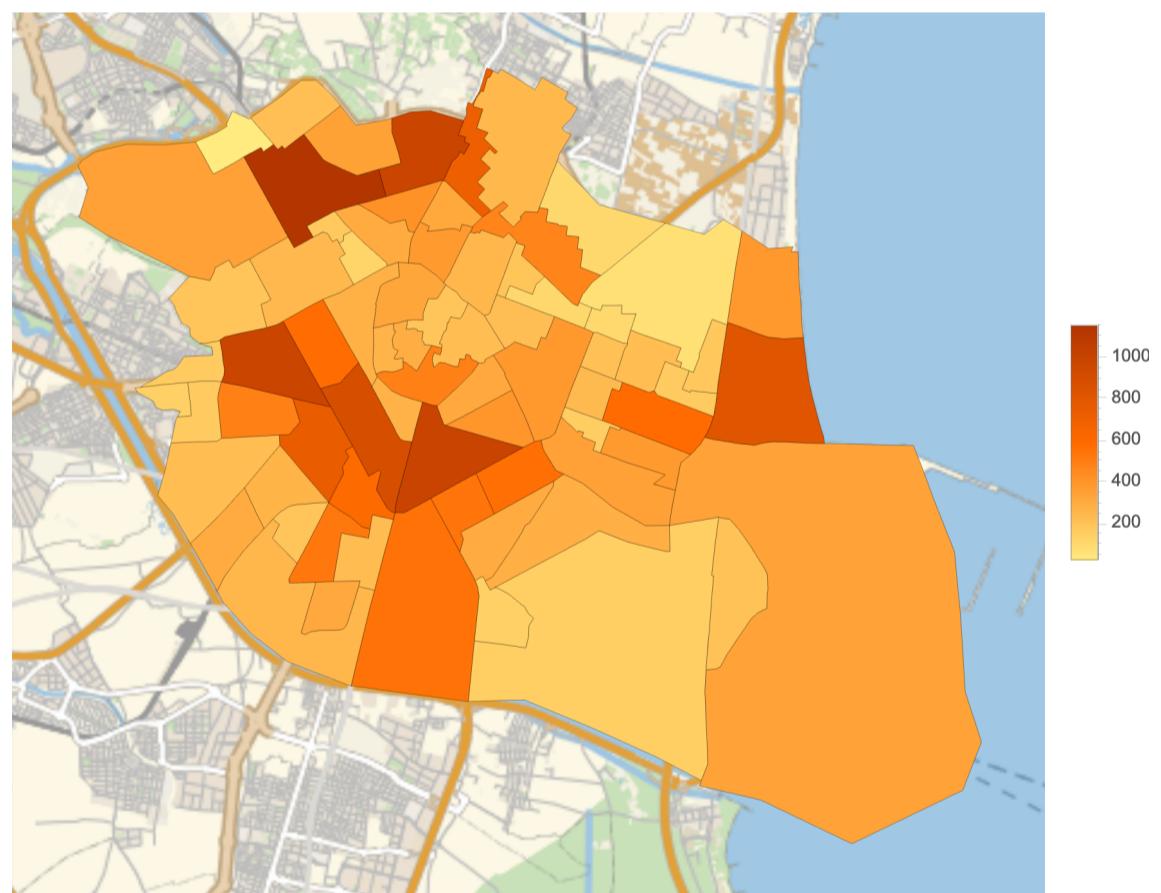
Con los datos de coordenadas que tenemos, y usando los datos del archivo “Valencia_polygons.csv” del repositorio de Github he hecho lo siguiente:

- Para cada barrio, tengo su representación espacial dada por los vértices del polígono en el archivo mencionado.
- Con las coordenadas de inmueble, calculo si un inmueble está dentro de un barrio o no.
- Genero una nueva columna de mi dataset donde a cada inmueble le asigno el barrio al que pertenece.
- Hay 28 inmuebles que no están en ninguno de los barrios, estarán fuera de ciudad de acuerdo con nuestro polígonos.

Barrios con más anuncios

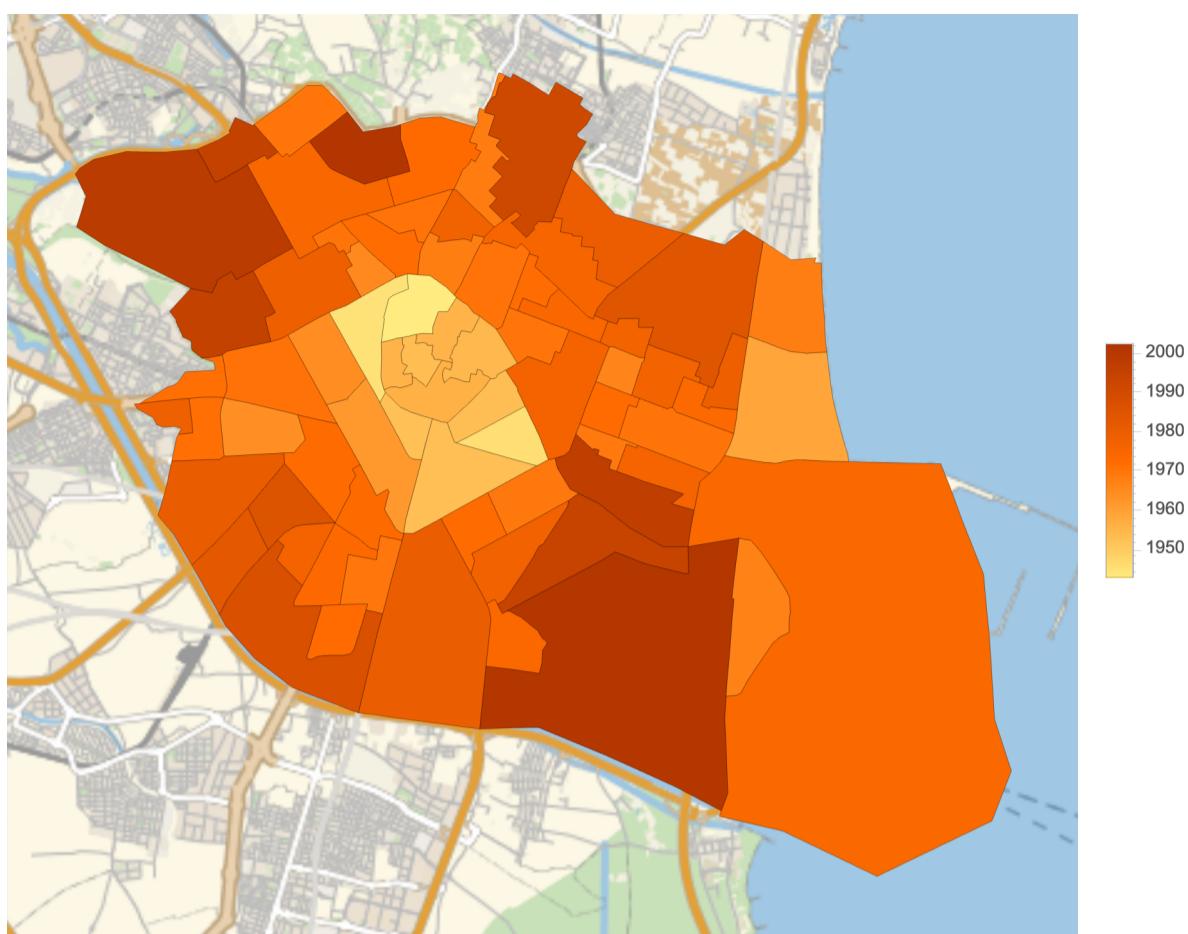
Como uno ejemplo de uso de los datos espaciales, hemos tomado los barrios que más anuncios tienen:

Benicalap	1150
Russafa	994
Torrefiel	984
Nou Moles	983
Arrancapins	885
El Cabanyal–El Canyamelar	821
Patraix	746
Els Orriols	714
La Raiosa	595
Aiora	582
La Petxina	579
Mont–Olivet	565
Malilla	552
En Corts	537
L'Hort de Senabre	526
Sant Francesc	485
Tres Forques	484
Benimaclet	472
Tormos	411
Gran VÃa	391
↙ ↘ rows 1–20 of 73 ↙ ↘	



Distribución espacial de las viviendas según año de construcción

Podemos observar que en la zona sur y noroeste es donde las viviendas son más nuevas y, evidentemente, tenemos la zona del centro histórico como la zona donde las viviendas son más antiguas.



Precios por segmentos

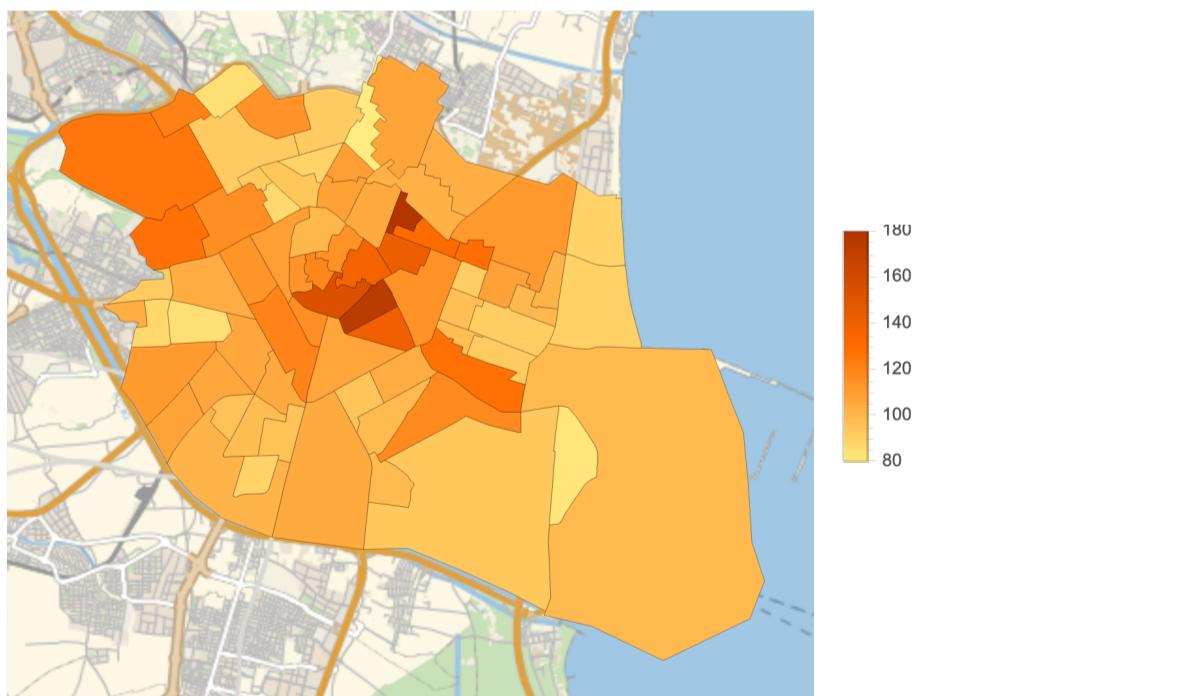
Como segmentación, vamos a tomar el número de habitaciones, aunque siempre lo vamos a filtrar por barrio a la hora de visualizar, para tener esos dos ejes de diferenciación. Podríamos, en caso de tener más tiempo, hacerlo por orientación, podríamos tomar los puntos de interés y calcular precio de las viviendas en un radio de 100 metros/200 metros/300 metros... hay posibilidades casi ilimitadas.

Número de habitaciones

Para generar los siguientes gráficos, el procedimiento que se ha seguido es el siguiente:

- Filtrar el dataset por número de vivienda, obteniendo 5 subsets.
- En cada subset, agrupar por barrio.
- Calcular por el precio medio absoluto y asignarle ese valor a cada barrio.
- Calcular el precio medio por metro cuadrado y asignarle ese valor a cada barrio.

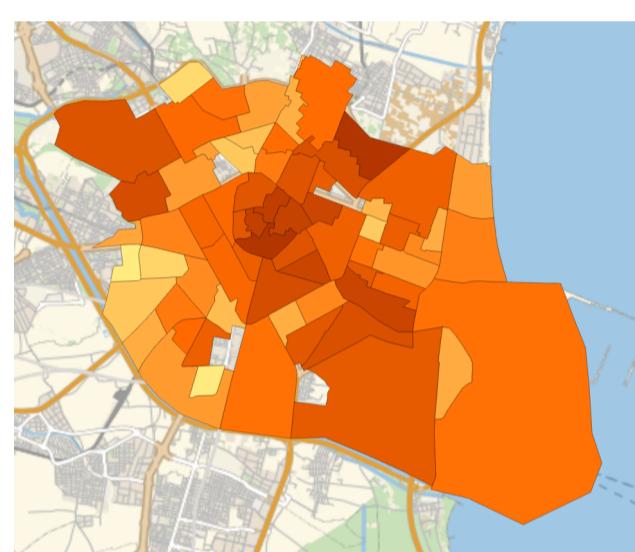
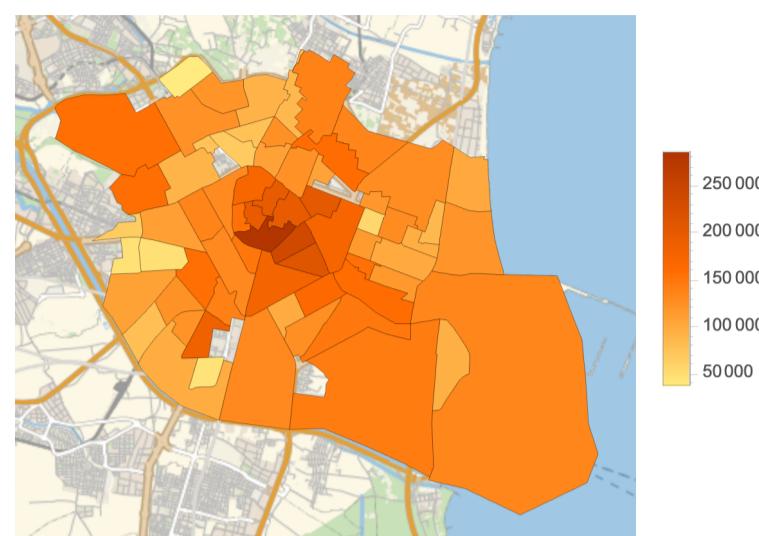
Como enlace entre los dos, he creado este mapa de la media de metros por vivienda por barrio:



El resultado muestral es el siguiente:

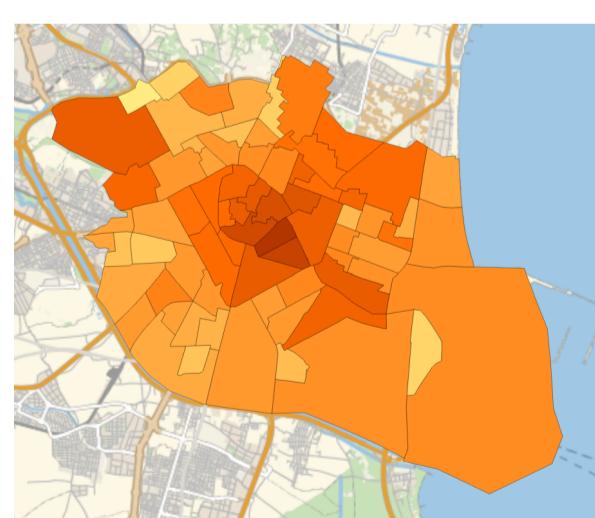
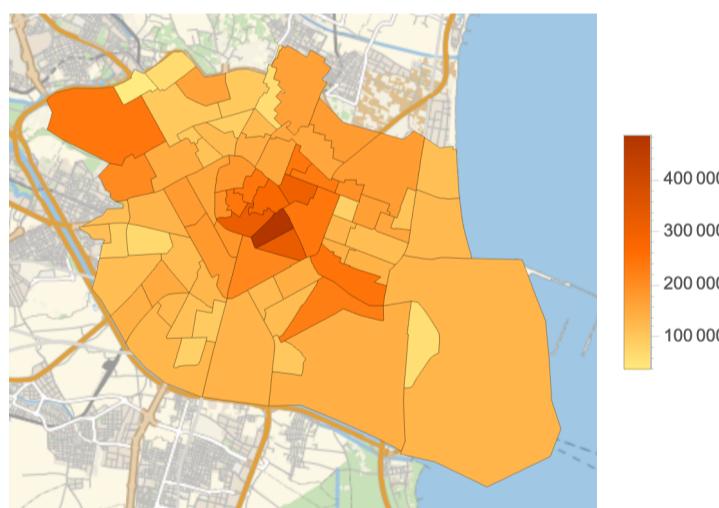
Viviendas de 1 habitación

Mínimo	25 000
1er Cuartil	91 000
Mediana	131 000
Media	142 970.
3er Cuantil	171 000
Top 5%	284 000
Top 1%	363 000
Máximo	981 000
Std. Dev.	77 899.6
Test de asimetría	2.36548
Kurtosis	17.7542



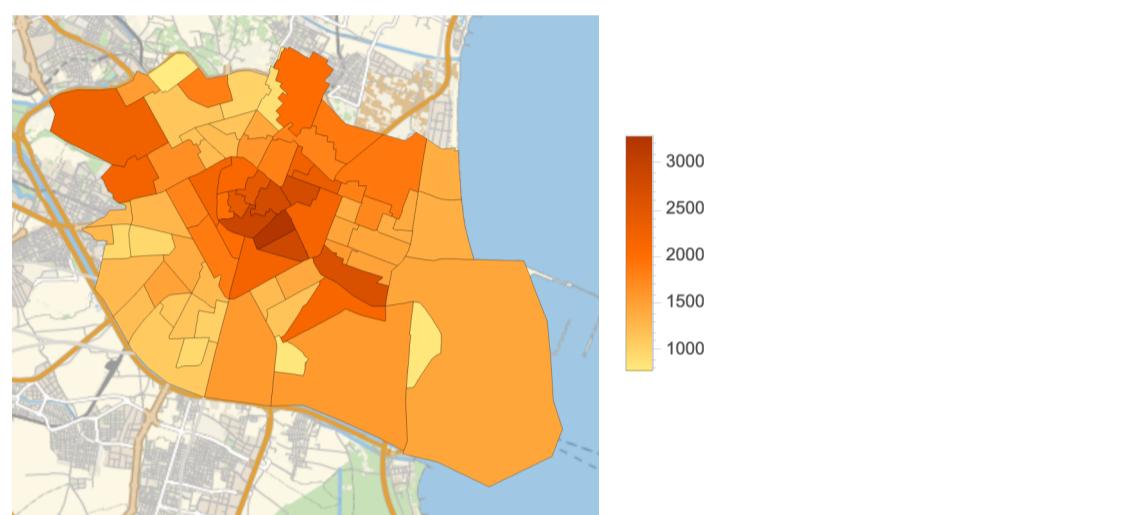
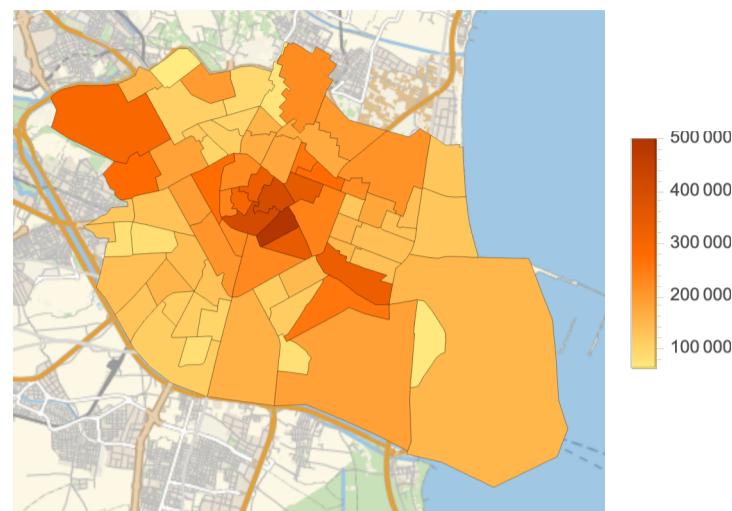
Viviendas de 2 habitaciones

Mínimo	24 000
1er Cuartil	86 000
Mediana	131 000
Media	146 641.
3er Cuantil	180 000
Top 5%	294 000
Top 1%	462 000
Máximo	863 000
Std. Dev.	86 235.1
Test de asimetría	2.01052
Kurtosis	10.7207



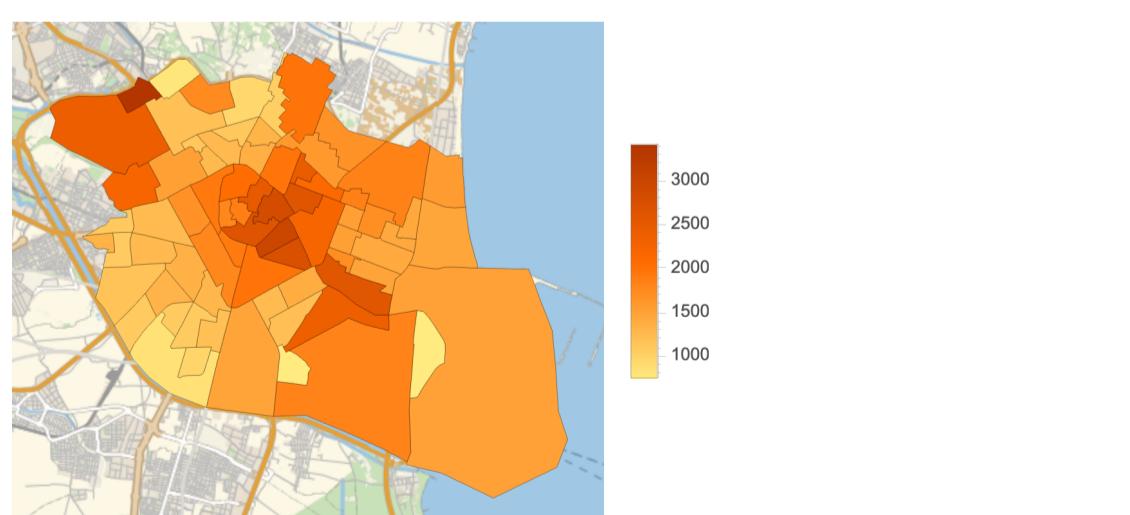
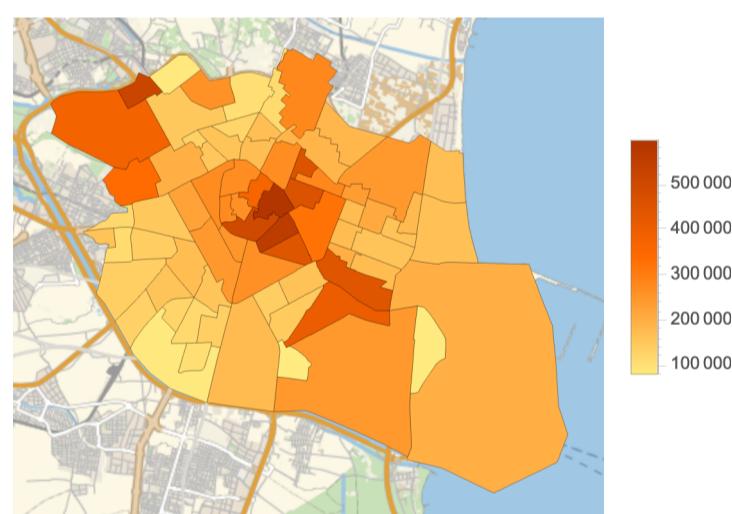
Viviendas de 3 habitaciones

Mínimo	25 000
1er Cuartil	87 000
Mediana	134 000
Media	161 798.
3er Cuantil	197 000
Top 5%	364 000
Top 1%	550 000
Máximo	2 441 000
Std. Dev.	111 936.
Test de asimetría	3.21546
Kurtosis	31.2924



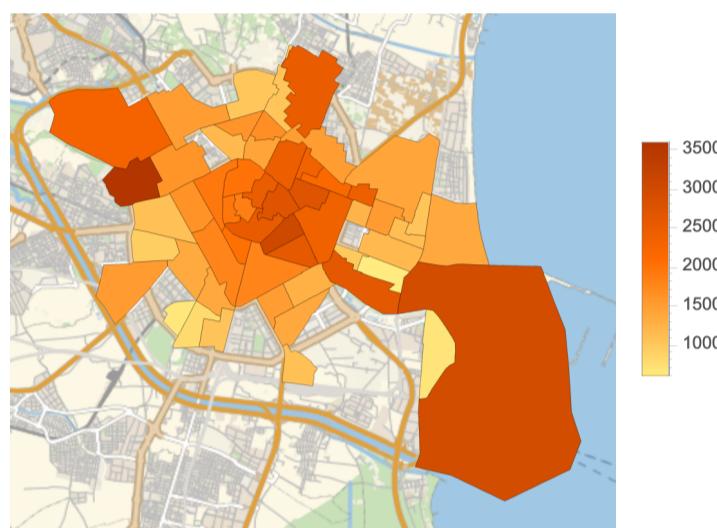
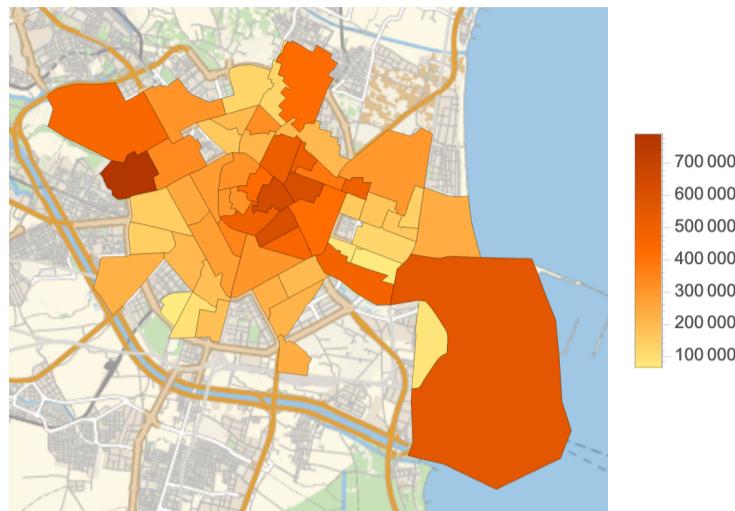
Viviendas de 4 habitaciones

Mínimo	25 000
1er Cuartil	125 000
Mediana	173 000
Media	217 631.
3er Cuantil	260 000
Top 5%	526 000
Top 1%	779 000
Máximo	1 217 000
Std. Dev.	148 833.
Test de asimetría	2.20189
Kurtosis	9.47577



Viviendas de 5 habitaciones

Mínimo	46 000
1er Cuartil	223 000
Mediana	344 500
Media	381 463.
3er Cuantil	496 000
Top 5%	760 000
Top 1%	1 004 000
Máximo	1 707 000
Std. Dev.	209 339.
Test de asimetría	1.21274
Kurtosis	5.56567



Sin conocer Valencia en profundidad, viendo estos datos podemos ver claramente que la zona centro es la más cara y a zona suroeste la más humilde para viviendas de todos los tamaños. Tenemos huecos vacíos en zonas porque no tenemos datos para todas las zonas y todos los tipos de vivienda.

Productos de datos y limitaciones

En mi opinión, la principal limitación de estos datos es la temporal. El mercado inmobiliario tiene como principal característica tener una reacción lenta ante los shocks externos. Sin duda, hay una parte del mercado que son inversiones a corto plazo (reforma y venta) pero en la mayoría de los casos este sector de la economía tiene un carácter largoplazista que hace que sus fluctuaciones en los precios sean mucho más lentas, aparte de verse habitualmente limitadas por medidas políticas (en el contexto actual, modificaciones hipotecarias a préstamos variables, moratorias en otros casos...). En un mercado corriente, un aumento del precio real de la vivienda (por la subida de los precios de la financiación), debería producir una rebaja en el precio nominal, pero esto no es algo que ocurra a corto plazo, por lo que no se puede analizar con datos anuales.

Como conclusión, una lista de ejemplo de productos y servicios que se podrían vender a partir de estos datos serían:

- Pasar de datos de barrios a nivel de calles y cruzarlos con datos de renta per cápita por calles disponibles para ver la relación entre precio de vivienda y renta.
- Evolución de las zonas en las que ha habido más anuncios de viviendas (y, como consecuencia, más actividad) a lo largo del tiempo para intentar anticipar tendencias en zonas “de moda” para que inmobiliarias puedan centrar su actividad.
- No sé cómo se hace en idealista, pero creo que es muy importante tener información veraz sobre fecha de inicio de la venta, fecha en la que se realiza la operación y precio real. Se debe incentivar a los anunciantes a proporcionar esta información para ofrecer servicios más precisos y atractivos a los clientes de la plataforma.
- Estimar el porcentaje real de operaciones que se realizan gracias/a través de idealista sobre el total de transmisiones reales (dato público) para tener una medida objetiva sobre el impacto de la plataforma sobre la que ejercer acciones de negocio.
- Hacer un filtrado de viviendas por calidad de materiales/construcción basado en el análisis de las fotografías aportadas (con métodos de image identification entrenados para reconocer materiales, defectos estructurales... y también por zonas para poder conocer los edificios) para poder ordenar viviendas de más necesitadas de reformas a nuevas y poder vender este producto a empresas de reformas que pudieran ofrecer sus servicios a anunciantes a través de idealista.