clustering_resumido

Ana Valiente

2025-06-08

Con el objetivo de identificar patrones relevantes dentro de la base de datos, se han planteado dos análisis de clustering independientes. Esta técnica de aprendizaje no supervisado permite agrupar inmuebles en función de sus similitudes, sin necesidad de una variable respuesta. Para comenzar la exploración, seleccionamos dos grupos de variables para construir nuestros nuevos conjuntos: Clustering de Confort (dimensión interna), enfocado en las condiciones de comodidad y habitabilidad que ofrece la propia vivienda. Clustering de Servicios (dimensión externa), refleja la calidad, accesibilidad y otras características del entorno urbano del inmueble.

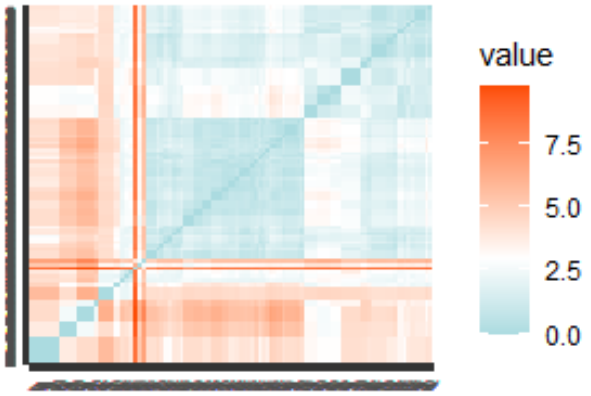
Variable	Confort	Servicios
bathrooms	1	0
rooms	1	0
surface	1	0
tieneAscensor	1	0
tieneTrastero	1	0
tieneCalefaccion	1	0
tieneAireAcondicionado	1	0
supermarket_count	0	1
pharmacy_count	0	1
hospital_count	0	1
university_count	0	1
college_count	0	1
public_transport_count	0	1

Sin embargo, debido a limitaciones de espacio, en la memoria principal se ha optado por incluir únicamente el clustering basado en los servicios de la ubicación. Hemos priorizado esta división antes que la de confort y comodidad porque consideramos que sus resultados muestran una estructura de agrupamiento más clara y relevante. Además, la clasificación solventa uno de los problemas de la base de datos original: el exceso de barrios y municipios que no permitía un análisis correcto del entorno urbano. Así pues, dado que las conclusiones de esta fase del proyecto serán incorporados en estudios posteriores; creemos que su inclusión en la memoria final es imprescindible. No obstante, la clasificación en función del confort está disponible en el anexo de CLUSTERING; en caso de que se quiera consultar.

CLUSTERING DE SERVICIOS DEL ENTORNO

El paso previo a la clasificación fue comprobar si existían tendencias de agrupación en los pisos basándonos en las variables del cluster. Para ello, elaboramos un *heatmap*. Dado que pretendíamos identificar similitudes entre las infraestructuras ofrecidas por cada ubicación, consideramos adecuado aplicar una medida de cercanía. Probamos todas las distancias para variables numéricas disponibles y, finalmente, obtuvimos una tendencia de agrupación más compacta con la euclídea. En el mapa resultante, se observan ciertos subgrupos alrededor de la diagonal principal; aunque sería necesario efectuar el clustering para identificar estructuras más claras.

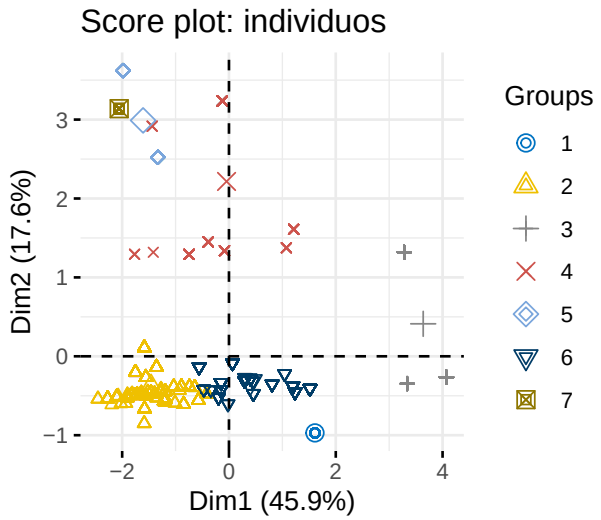
.

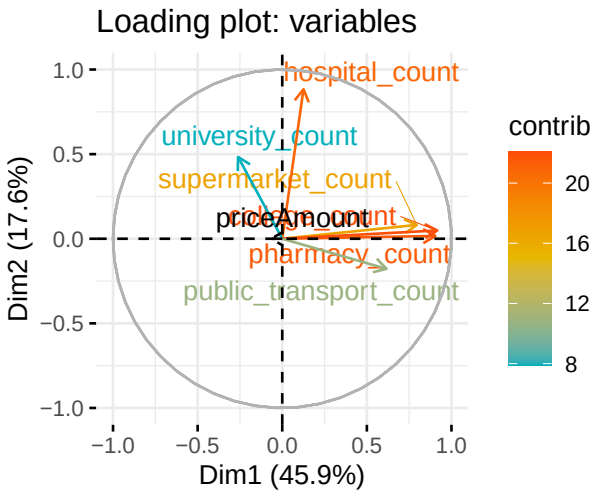


Como punto de partida, iniciamos el análisis probando varias técnicas de agrupamiento: el método jerárquico de Ward y los de partición k-medias y k-medioides. En los tres casos, analizamos simultáneamente la suma de cuadrados residual y el coeficiente de Silhouette para determinar el número óptimo de subgrupos. Así, seleccionamos k=6 con Ward; k=8 mediante PAM y k=7 aplicando k-medias. Una vez divididos los datos, comparamos el coeficiente medio de Silhouette y vimos que, en los grupos obtenidos a partir de las medias, la mayoría de los grupos estaban mejor definidos. Por ello, nos decantamos por los 7 clusters. Los gráficos relativos a la selección del método de agrupación no se han incluido en la memoria final, pero se pueden consultar en el anexo de CLUSTERING.

A continuación, estudiamos qué variables habían tenido más peso en la nueva clasificación a través de un Análisis de Componentes Principales con 2 dimensiones. Asimismo, utilizamos las columnas cluster y priceAmount como auxiliares para entender su relación con el resto de variables consideradas.

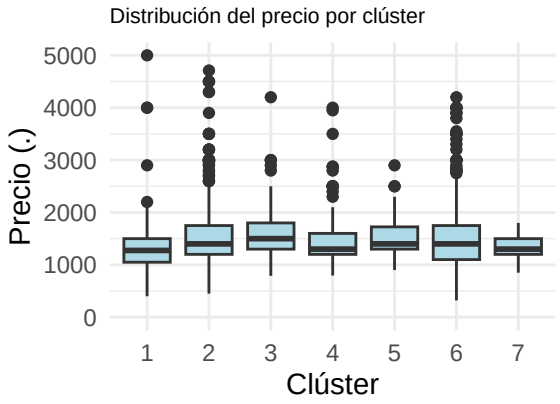
Al estudiar el score plot, identificamos siete grupos claramente diferenciados. Destacan especialmente los clusters 7 y 1, tan compactos que apenas se distinguen uno o dos individuos. Incluso en las divisiones más dispersas (3 y 4) sus puntos se encuentran bastante próximos. Por otra parte, en el loading plot se aprecia que la primera componente está positivamente relacionada con la cantidad de farmacias, colegios, supermercados y conexiones de transporte. En cambio, la segunda depende de los hospitales y universidades. No parece que el precio del piso tenga una gran importancia en esta clasificación. Paralelamente, si suporponemos ambos gráficos, parece que el cluster 3 sobresale por estar notablemente mejor comunicado que el resto de grupos. Por el contrario, el cluster 2 es aparentemente el peor ubicado.



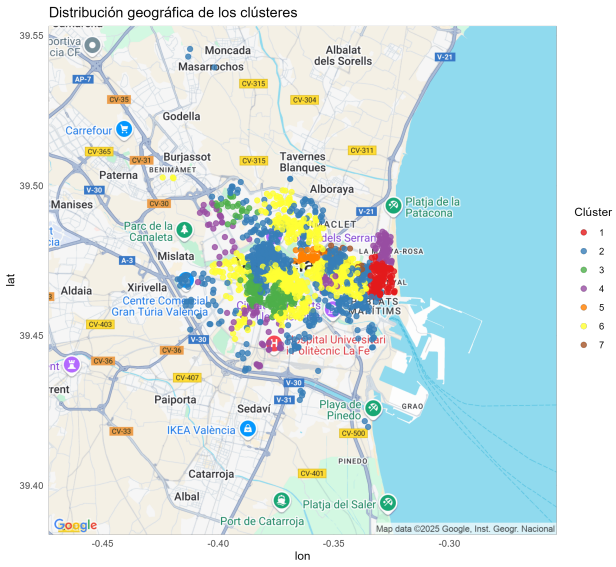


Para comparar los precios de cada cluster, calculamos sus medias y representamos las 7 distribuciones mediante sus respectivos gráficos de caja y bigotes. Tal como habíamos supuesto al examinar el PCA, no encontramos una diferencia muy acusada entre los precios medios. No obstante, resulta sorprendente que el cluster 2, que habiamos resaltado como el peor ubicado, es el más caro en promedio. Este importe elevado seguramente se deba a los pisos atípicos detectados en su *box&whiskers*, los cuales podrían estar mal clasificados o reflejar una ubicación segmentada internamente entre viviendas básicas y otras más exclusivas. El tercer cluster tiene un precio muy similar, aunque con menos pisos atípicos, por lo que seguramente sus viviendas presenten buena comunicación general.

Clúster	Precio medio (.)
1	1360.402
2	1595.870
3	1595.106
4	1451.182
5	1564.750
6	1562.079
7	1331.579



Finalmente, para visualizar geográficamente los resultados obtenidos, hemos los 7 clusters en un mapa de la ciudad de Valencia, utilizando las variables latitud y longitud. Así, fue posible determinar a qué zona corresponde cada uno de los grupos. La representación del mapa la conseguimos a partir de una API KEY de google cloud y, por motivos de seguridad, en esta memoria se incluirá en formato png. En la imagen se aprecia como, mientras algunos clusters están muy centrados en una zona concreta, otros se extienden por varios barrios valencianos.



A partir de la información obtenida en el mapa, hemos diseñado la siguiente clasificación:

Clúster 1	Poblats Marítims
Clúster 2	Centro histórico, pedanías y zona universitaria
Clúster 3	Ruzafa, Jesús y Benicalap
Clúster 4	Periferia
Clúster 5	Aragón, Blasco Ibáñez
Clúster 6	Gran Vía, Colón y zona universitaria
Clúster 7	Tarongers

En este punto, podemos aplicar las conclusiones extraídas del análisis a zonas reales y conocidas de nuestra ciudad. Por ejemplo, las zonas mejor ubicadas y con mejores prestaciones serían Ruzafa, Jesús y Benicalap. Paralelamente, entre el Centro histórico, pedanías y parte zona universitaria encontramos una ecléctica gama de viviendas, unas pocas más exclusivas y la mayoría restante con comunicaciones bastante mejorables (posiblemente las ubicadas en las afueras de la ciudad). Los clusters 5 y 7, que en el PCA aparecían altamente relacionados con la variable *university_count* se corresponden con los sectores de Aragón, Blasco y Tarongers. En conjunto, los resultados del clustering muestran una segmentación coherente con la estructura urbana y social de la ciudad, reflejando patrones territoriales que se corresponden con zonas reales y reconocibles. Esta clasificación no solo valida la calidad del análisis, sino que también permite incorporar de forma explícita la dimensión del entorno como variable explicativa en análisis posteriores; mejorando así nuestro entendimiento del mercado inmobiliario valenciano.