

Análisis del mercado inmobiliario en Valencia. Grupo A1-12

Ana Vílchez, Ana Valiente, Fátima Taberner, Miguel Gil y Oscar Antonino

2025-06-08

Contents

1 Introducción.	2
2 Análisis por componentes principales: PCA.	2
2.1 Análisis de atípicos.	3
2.2 Interpretación.	3
3 Clustering.	5
3.1 Clustering de servicios del entorno.	6
3.1.1 Representación de los clusters en la ciudad de Valencia.	8
4 Uso del Clustering para completar el PCA.	8
5 AFC simple: Subtipo y Precio	10
6 Reglas de Asociación.	10
7 PLS-DA.	11
8 PLS	14
9 ANEXO.	17
9.1 PCA.	17
9.1.1 Atípicos.	17
9.1.2 Gráficos de las contribuciones de las variables a las dimensiones.	22
9.2 Clustering.	22
9.2.1 Clustering de Servicios	22
9.2.2 Clustering de Confort	24
9.3 AFC.	28
9.3.1 AFC simple: Municipio vs. Precio	28
9.3.2 Justificación AFC simple: Subtipo vs. Precio	30

9.4	Reglas de Asociación.	31
9.5	Análisis Discriminante de Fisher.	33
9.5.1	Correlación entre variables.	33
9.5.2	Entrenamiento y test.	34
9.5.3	Normalidad y homocedasticidad.	35
9.6	PLS.	36

1 Introducción.

Nuestro conjunto de datos está compuesto por 1867 observaciones; representando cada una de ellas un inmueble en alquiler en la ciudad de Valencia, con sus rasgos característicos. En total, cuenta con 44 variables como el precio del alquiler de la vivienda, sus habitaciones, donde esta localizada, servicios de la zona...

La fuente principal del estudio ha sido el web scraping a la plataforma Fotocasa, filtrando únicamente las ofertas localizadas en el municipio de Valencia.

Y el objetivo principal del estudio es analizar y predecir el precio del alquiler en Valencia en función de sus características y ubicación, además de estudiar qué relación existe entre las diferentes variables y agrupar las viviendas en bloques según sus características.

Para ello, emplearemos distintas técnicas de análisis descriptivo y predictivo.

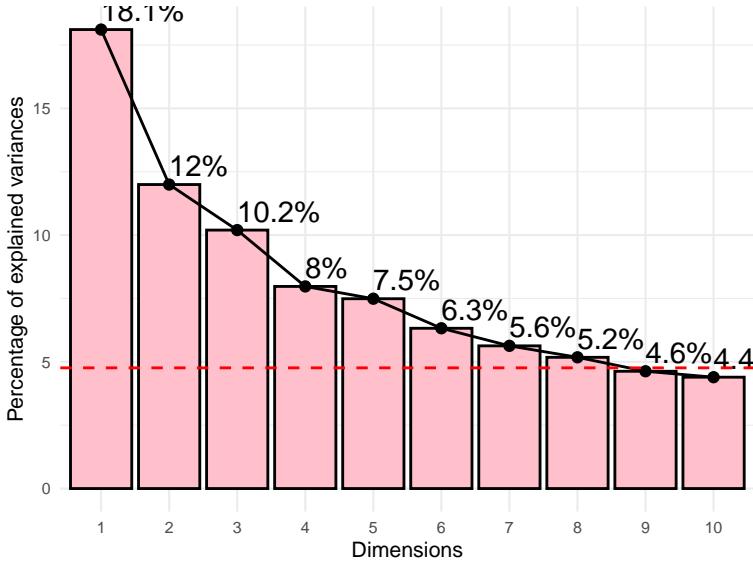
2 Análisis por componentes principales: PCA.

Como primer paso para analizar los datos que obtuvimos a través de la web de Fotocasa, realizamos un análisis de componentes principales (PCA). El objetivo era reducir la dimensionalidad del conjunto de datos de viviendas (manteniendo la mayor varianza posible en los datos) e identificar patrones o tendencias comunes en las características de los inmuebles ofertados.

Para la realización del PCA seleccionamos de nuestra base de datos 22 variables: 2 de ellas categóricas ordinales, 4 binarias y el resto numéricas. Nuestra variable a predecir es “priceAmount” por lo que la tratamos como variable suplementaria para el PCA, es decir, no interviene en la creación de las componentes principales, pero sí se proyecta sobre ellas para ver que variables están relacionadas con ella positivamente, cuales lo están negativamente y cuales no tienen relación.

Una vez seleccionadas las variables, procedimos a centrarlas para que el algoritmo identificase correctamente las direcciones de mayor variabilidad partiendo de un centro común (el origen). También las estandarizamos, debido a que estaban medidas en distintas unidades (como metros cuadrados, unidades, etc.). Con esta estandarización nos aseguramos que todas las variables tuvieran el mismo peso en el análisis, y evitar así que aquellas con magnitudes numéricas mayores influyeran de forma desproporcionada en la construcción de las componentes principales.

Seguidamente analizamos la varianza explicada por cada una de las componentes principales para seleccionar el número de componentes más adecuado.



Como resultado vimos que empleando este método (varianza explicada acumulada), con 8 componentes conseguíamos explicar el 72% de la variabilidad total de nuestros datos, sin embargo, seleccionar 8 componentes complicaría excesivamente la interpretación de los resultados, por lo que decidimos estudiar otros criterios para seleccionar el número de componentes adecuado. Empleamos el método del codo, y obtuvimos que podríamos seleccionar 4 componentes principales, por tanto nos quedamos con esta última cifra para posteriormente facilitar el análisis de los resultados.

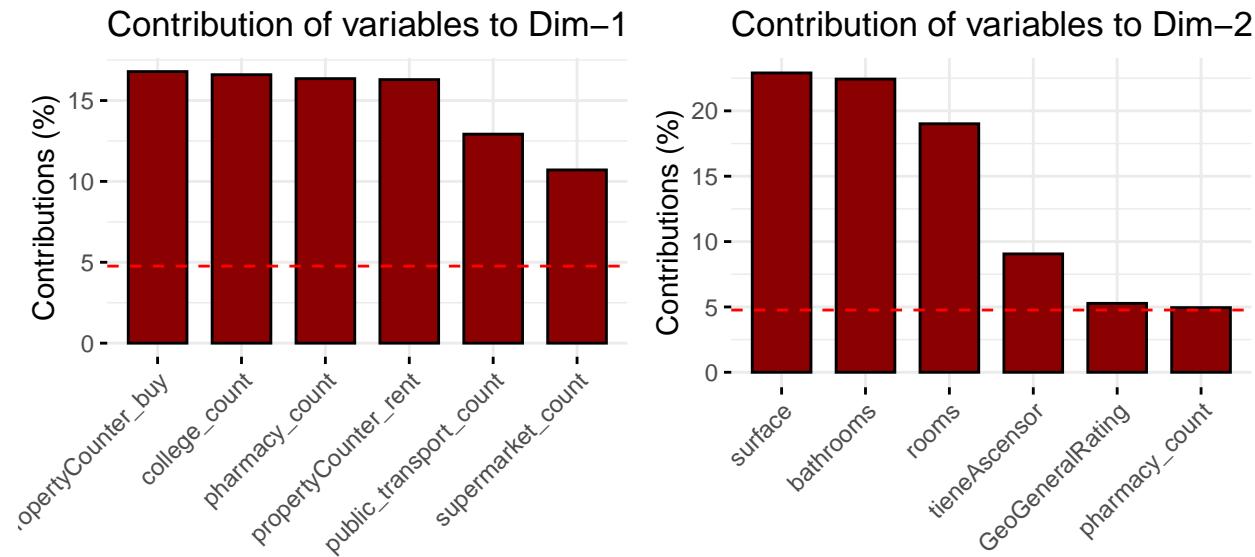
2.1 Análisis de atípicos.

Tras la selección de componentes, procedimos a evaluar primeramente la presencia de observaciones atípicas extremas dentro del espacio generado por las componentes principales. Para ello, utilizamos el estadístico T^2 de Hotelling, que nos permitió medir la distancia de cada observación al centro de la nube de puntos proyectada sobre las primeras 4 componentes. Seguidamente realizamos también un gráfico de distancia al modelo, basado en la Suma de Cuadrados de los Residuos (SCR). El objetivo en este caso era identificar atípicos moderados, es decir, aquellos inmuebles cuyas características no estaban siendo capturadas adecuadamente por las componentes principales seleccionadas. Tras detectar, en cada caso, las observaciones que tenían una T^2 de Hotelling o una SCR superior al límite establecido, seleccionamos la observación con la T^2 de Hotelling más alta, y una observación detectada como atípico moderado. Analizamos las contribuciones individuales de cada variable a sus altos valores y obtuvimos que por un lado, en el caso del atípico extremo, se debía a que era un ático de lujo en Valencia, por lo que tenía un número elevado de superficie, baños y habitaciones. Y por otro lado, en el caso del atípico moderado, se debía a que el inmueble estaba localizado en una zona con un número elevado de universidades cerca. Por tanto, se asumió que el resto de observaciones con distancias elevadas no se correspondían a errores de codificación ni valores anómalos injustificados, sino que respondían a casos reales y válidos. Por ello, no se consideró necesario excluirlas del análisis, ya que la técnica de PCA captura las tendencias generales del conjunto, pero es esperable que algunas viviendas reales, aunque menos frecuentes, se representen peor o no tengan valores similares al resto. Manteniendo estas observaciones podemos reflejar con mayor fidelidad la diversidad y complejidad del mercado inmobiliario analizado. Todos los gráficos respectivos al análisis de atípicos están en el 9.1.1.

2.2 Interpretación.

Como primer paso en la interpretación del PCA, analizamos la contribución de cada variable a las cuatro primeras componentes principales. Para ello, generamos gráficos de barras que muestran el porcentaje de

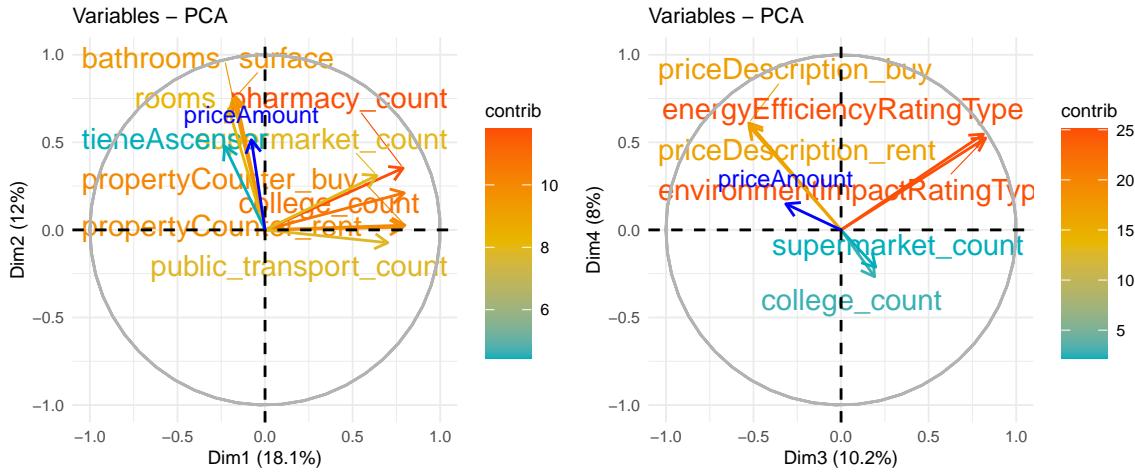
contribución de cada variable a las dimensiones 1, 2, 3 y 4, eliminando variables que no tienen gran peso en la componente, para así visualizar mejor la información que nos interesa (Gráficos Dim3 y Dim4 en 9.1.2).



A partir de los resultados obtenidos extrajimos las siguientes conclusiones:

- **Dimensión 1:** dominada por variables relacionadas con el entorno y la oferta de servicios (como propertyCounter_buy, college_count, pharmacy_count, propertyCounter_rent, etc.), lo cual sugiere que esta componente resume principalmente la accesibilidad y densidad de servicios en la zona.
- **Dimensión 2:** captura características físicas del inmueble, como surface, bathrooms, rooms, y también la presencia de ascensor (tieneAscensor). Esta dimensión parece asociada al tamaño y equipamiento de las viviendas.
- **Dimensión 3:** marcada por environmentImpactRatingType, energyEfficiencyRatingType, así como priceDescription_buy y priceDescription_rent. Aquí vemos que esta dimensión recoge información relativa a aspectos energéticos y al precio de las viviendas.
- **Dimensión 4:** vuelve a resaltar variables como priceDescription_buy, priceDescription_rent, junto con tieneAireAcondicionado o college_count. Esta dimensión podría reflejar un eje entre características de confort del inmueble y variables de contexto económico/educativo.

Para reforzar la información obtenida del gráfico de contribuciones a cada componente principal generamos varios gráficos donde se muestran las variables proyectadas sobre las diferentes componentes principales. Concretamente, representamos los planos formados por las combinaciones de componentes (1,2) y (3,4). Para una mejor visualización hemos proyectado solo las 10 y 6 variables que más contribuían a la formación de los respectivos ejes.



En estos gráficos, cada flecha representa una variable, y su dirección e intensidad nos indican cómo contribuye a la componente correspondiente. Cuanto más larga y más cálido es el color (como naranja o rojo), más relevante es esa variable para definir el eje.

Por ejemplo, en el plano (1,2), vemos claramente que variables como “surface”, “rooms” o “bathrooms” tienen mucho peso sobre la Dim2 y están correlacionadas entre si, lo que tiene sentido porque están muy relacionadas con las características físicas del inmueble, además también vemos como variables como “college_count”, “supermarket_count” y “pharmacy_count” también están correlacionadas entre si y tienen un gran peso sobre la Dim1, lo que es coherente con lo dicho anteriormente (que la dimensión 1 resumía principalmente la accesibilidad y densidad de servicios en la zona), es decir, si en una zona hay muchos supermercados, habrá muchas farmacias, colegios, etc. Por otro lado, vemos como en los planos (3,4), en el eje de la Dim3, “environmentImpactRatingType” y “energyEfficiencyRatingType” están correlacionadas positivamente, al igual que “priceDescription_buy” y “priceDescription_rent”, y a su vez están correlacionadas negativamente entre ellas. Esto nos indica que pisos con menor tasa de contaminación y consumo, tienen mayores precios. Por otro lado, en la Dim4, vemos como pisos en zonas con servicios como colegios o supermercados, tienden ligeramente a tener un menor precio y menor consumo. Además cabe mencionar de nuevo que la variable “priceAmount” la proyectamos como auxiliar, de esta forma se ve como esta relacionada con el resto de variables sin influir en la creación de las componentes. En este caso que comentábamos se ve claramente que a más superficie, baños y habitaciones claramente más precio, y en el segundo gráfico se comporta como las variables “priceDescription_buy” y “priceDescription_rent”, lo que confirma que a menor consumo y contaminación mayor precio.

Hasta ahora no habíamos representado los scores (cada inmueble) en el espacio de componentes principales, ya que al tener tantas observaciones no podíamos interpretarlo fácilmente. Para tratar de distinguir tipos de inmuebles por sus características barajamos la opción de resaltar los scores según al barrio de Valencia al que pertenezcan, pero al ser mas de 10 no se distinguían fácilmente. Por todo esto, tras la explicación del clustering realizado a nuestros datos, completaremos el PCA coloreando los inmuebles en el espacio generado por las componentes principales según al cluster al que pertenezcan.

3 Clustering.

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

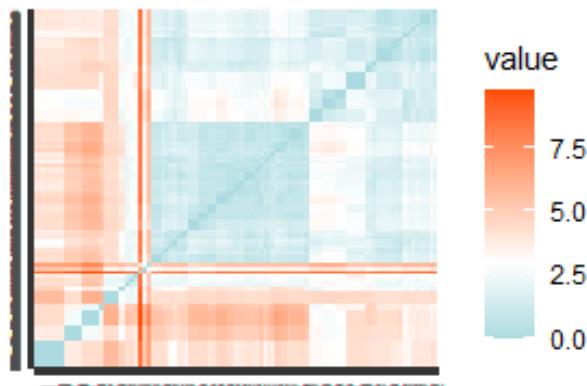
grupos de variables para construir nuestros nuevos conjuntos: Clustering de Confort (dimensión interna), enfocado en las condiciones de comodidad y habitabilidad que ofrece la propia vivienda. Clustering de Servicios (dimensión externa), refleja la calidad, accesibilidad y otras características del entorno 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 9.2.2; en caso de que se quiera consultar.

3.1 Clustering de servicios del entorno.

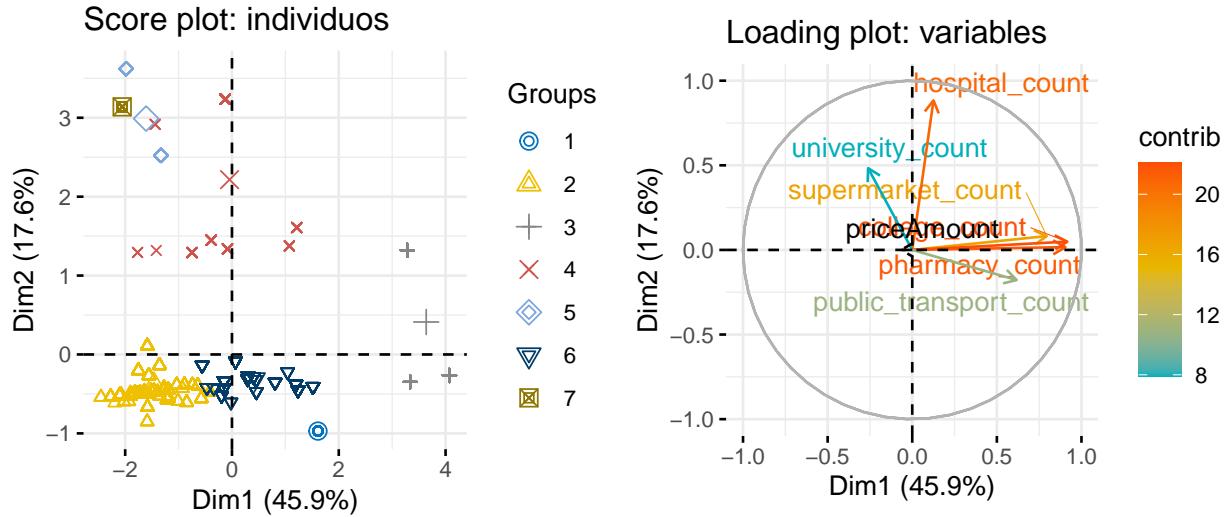
El paso previo a la clasificación fue comprobar si existían tendencias de agrupación en los pisos basándose en las variables del cluster. Para ello, elaboramos un *heatmap*. Dado que pretendíamos identificar similitudes entre las infraestructuras ofrecidas por cada ubicación, consideramos adecuado aplicar 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-medoides. 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 9.2.1.

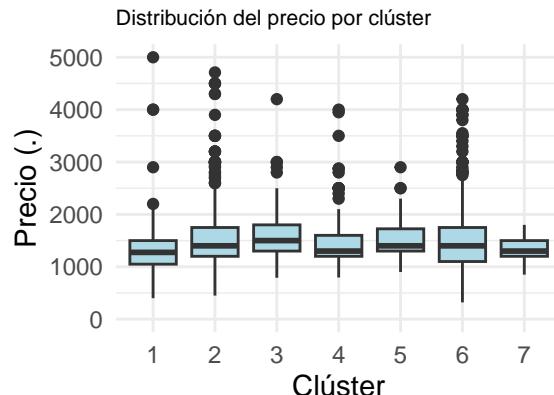
A continuación, estudiamos qué variables habían tenido más peso en la nueva clasificación a través de un Análisis de Componentes Principales con 2 dimensiones. Asimismo, utilizamos las columnas cluster y priceAmount como auxiliares para entender su relación con el resto de variables consideradas.

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 superponemos 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 habíamos 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



3.1.1 Representación de los clusters en la ciudad de Valencia.

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.

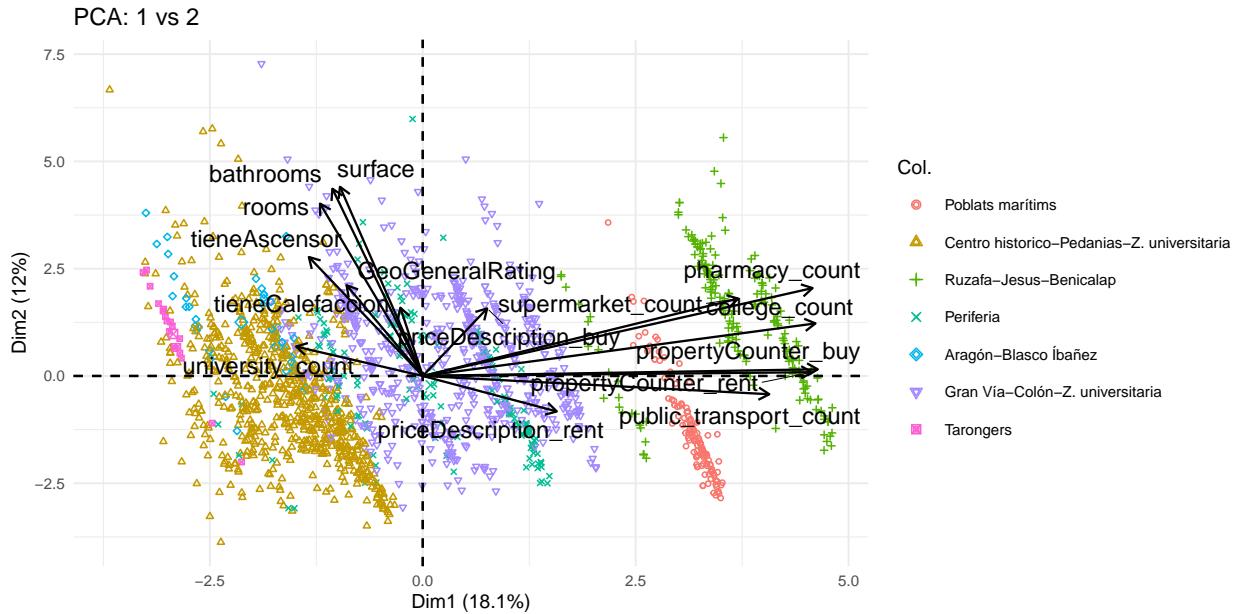


A partir de la información obtenida en el mapa, hemos diseñado la siguiente clasificación:

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

4 Uso del Clustering para completar el PCA.

En este punto, podemos aplicar las conclusiones extraídas del análisis a zonas reales y conocidas de nuestra ciudad. Utilizaremos estas agrupaciones para colorear en el PCA realizado anteriormente.



Tras comparar los anteriores gráficos con el mapa de Valencia, y ver a que zona pertenece cada grupo, extrajimos las siguientes conclusiones:

-> **Poblets marítims:** Es una zona bien comunicada, con buena oferta de propiedades, alta presencia de supermercados, transporte público, y cercanía a colegios. Parece un entorno atractivo para jóvenes y estudiantes, donde la movilidad y el acceso a servicios básicos son clave, o también para estancias vacacionales, porque esta cerca del puerto y la playa. Puede tener precios de alquiler más reducidos, pero no destacan por la amplitud o el número de baños.

-> **Centro histórico-pedanías.Z. universitaria:** representa una zona heterogénea: comparte una escasa accesibilidad a servicios urbanos, pero muestra gran variabilidad en cuanto al tamaño, equipamiento y precio de las viviendas. Algunas áreas pueden ofrecer viviendas espaciosas con ascensor, calefacción y a un precio mayor, mientras que otras tienen una oferta más básica. Esta diversidad podría explicarse por la mezcla entre áreas más históricas, otras más funcionales (posiblemente vinculadas al entorno universitario) y otras alejadas de la ciudad.

-> **Ruzafa-Jesús-Benicalap:** Zonas urbanas bien conectadas y con abundantes servicios (supermercados, transporte, farmacias). Aunque en general las viviendas no destacan por su tamaño, en algunas subzonas del grupo sí hay inmuebles amplios y bien equipados. Son áreas atractivas para quienes valoran el entorno y la accesibilidad, sin descuidar del todo el confort del hogar.

-> **Periferia:** Es una zona heterogénea. En áreas cercanas a universidades se encuentran viviendas más amplias y equipadas, mientras que en otras zonas, con mejor acceso a servicios y transporte, predominan inmuebles más pequeños.

-> **Aragón-Blasco Ibáñez:** Zonas funcionales con viviendas medianas o amplias y bien equipadas. Aunque no destacan por sus servicios, su cercanía a áreas universitarias y buena eficiencia energética las hace atractivas para jóvenes y familias que valoran el espacio y la ubicación, incluso a un coste algo mayor.

-> **Gran Vía-Colón-Z.universitaria:** representa zonas variadas: algunas con buenos servicios y precios altos, pero viviendas pequeñas; otras, cercanas a universidades, con más espacio y mejor equipamiento, aunque con menos servicios. En general, combina ubicaciones atractivas con viviendas heterogéneas, ideales para quienes priorizan la localización.

-> **Tarongers:** vinculado al entorno universitario. Zonas donde las viviendas ofrecen un adecuado tamaño y confort, y con excelente acceso a centros educativos aunque no a tantos servicios. Son áreas preferidas por estudiantes o jóvenes que buscan funcionalidad a buen precio.

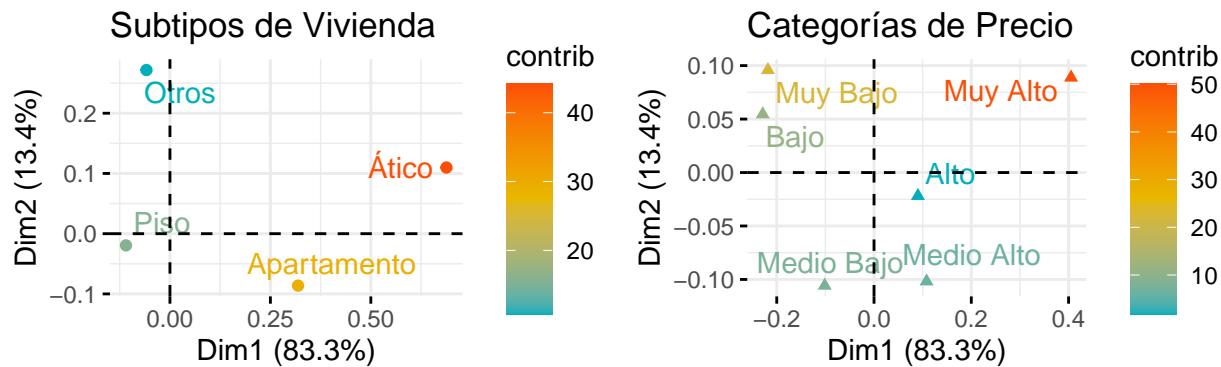
En conjunto, los resultados del clustering muestran una segmentación coherente con la estructura urbana y social de la ciudad, reflejando patrones territoriales que se corresponden con zonas reales y reconocibles. Esta clasificación no solo valida la calidad del análisis, sino que también permite incorporar de forma explícita la dimensión del entorno como variable explicativa en análisis posteriores; mejorando así nuestro entendimiento del mercado inmobiliario valenciano.

5 AFC simple: Subtipo y Precio

Además de este AFC simple se ha realizado también otro AFC con las variables Municipio y Precio, sin embargo como el clustering aporta más información que el AFC sobre dichas variables se ha decidido dejarlo en Anexos (9.3.1) como consulta en lugar de presentarlo en el proyecto.

A continuación, se analizan los resultados del Análisis de Correspondencias (AFC) realizado entre el tipo de vivienda y el rango de precios. Este análisis permite identificar patrones entre distintas categorías de vivienda (como áticos o pisos) y los niveles de precio en los que suelen encontrarse. Por esta razón, se ha optado por aplicar un AFC a estas variables.

La justificación detallada del análisis, así como su desarrollo completo, se encuentra en el apartado de Anexos (véase 9.3.2). En los gráficos siguientes, la interpretación se centrará exclusivamente en el eje horizontal, dado que este representa el mayor porcentaje de variabilidad explicada (83,3%). En consecuencia, la posición en el eje vertical (Dim2) no será tenida en cuenta en esta lectura.



Al interpretar ambos gráficos conjuntamente, y centrándonos exclusivamente en la Dimensión 1, se observa un patrón claro y coherente que vincula determinados subtipos de vivienda con categorías específicas de precios. Los subtipos más exclusivos, como el “Ático”, se alinean con las categorías más altas de precio, mientras que las categorías de precio más bajo se asocian con subtipos menos definidos o de menor valor. Esta estructura evidencia la existencia de una dimensión principal que organiza las relaciones entre tipo de vivienda y nivel de precios, permitiendo una interpretación robusta del mercado residencial en función de estas variables. La correspondencia entre ambos conjuntos de categorías es clara, especialmente porque la proyección de filas y columnas sobre el eje principal se refleja de manera simétrica.

6 Reglas de Asociación.

Con el objetivo de facilitar el análisis de reglas de asociación, se realizó una transformación de las variables de la base de datos. Las variables numéricas `precio` y `superficie` se agruparon en cuartiles, generando variables categóricas como “`precio_bajo`” o “`superficie_alta`”. Asimismo, se transformaron variables booleanas en etiquetas comprensibles (por ejemplo, `tieneAscensor = 1` se convierte en “`con_ascensor`”).

Una vez procesados los datos, se aplicó el algoritmo **Apriori** con un umbral mínimo de soporte del 1% y confianza del 50% obteniendo un total de 12745 reglas. Posteriormente se han eliminado reglas redundantes, esto permite conservar las reglas más significativas. Las reglas maximales no se han priorizado, ya que suelen perder especificidad, dificultando su aplicación práctica. Tras eliminar las redundantes nos quedamos con un total de 2408 reglas. Por último, filtramos las reglas con los siguientes umbrales quedándonos con las 21 reglas más relevantes: Soporte > 0.015, Confianza > 0.7 y Lift > 3.5

Estas condiciones permiten identificar patrones **robustos y estadísticamente relevantes** que relacionan ciertas configuraciones de un inmueble con su probabilidad de pertenecer a un rango de precio específico. A continuación se muestran las **cinco reglas más destacadas**, ordenadas por su *lift*. Es importante destacar que las 21 reglas comparten el consecuente de **Precio Alto**.

Table 1: Reglas de asociación más relevantes (ordenadas por lift)

	rules	support	confidence	coverage	lift	count
1001	{rooms: 4+, habitaciones: tieneAscensor con ascensor, bathrooms: 3+ banios} => {priceAmount: precio alto}	0.017	0.970	0.018	3.879	32
8593	{surface: surface alto, tieneAscensor con ascensor, tieneAireAcondicionado con aire, bathrooms: 3+ banios, tieneCalefaccion con calefaccion} => {priceAmount: precio alto}	0.017	0.970	0.018	3.879	32
4075	{surface: surface alto, tieneAscensor con ascensor, tieneAireAcondicionado con aire, bathrooms: 3+ banios} => {priceAmount: precio alto}	0.022	0.952	0.023	3.810	40
4063	{surface: surface alto, tieneAscensor con ascensor, bathrooms: 3+ banios, tieneCalefaccion con calefaccion} => {priceAmount: precio alto}	0.018	0.943	0.019	3.771	33
4087	{tieneAscensor con ascensor, tieneAireAcondicionado con aire, bathrooms: 3+ banios, tieneCalefaccion con calefaccion} => {priceAmount: precio alto}	0.018	0.943	0.019	3.771	33

Dado que dichas reglas están asociadas a precios altos, exploramos ahora aquellas cuya consecuencia es `priceAmount = precio_bajo`. Este tipo de reglas resulta útil para detectar inmuebles con condiciones objetivas que los hacen significativamente más asequibles. Para ello, filtramos las reglas con: Soporte > 0.0107, Confianza > 0.8 y Lift > 2.7

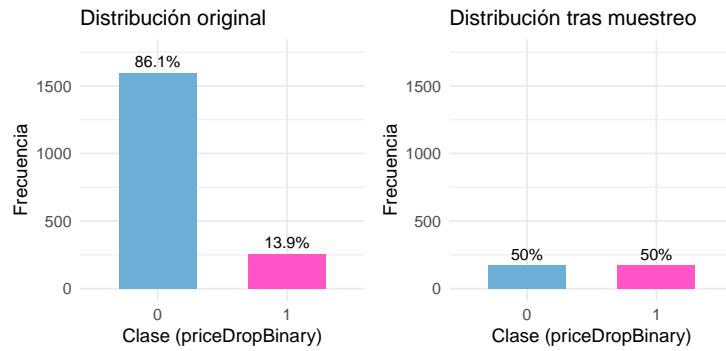
Table 2: Reglas de asociación más relevantes asociadas a Precios Bajos (ordenadas por lift)

	rules	support	confidence	coverage	lift	count
4271	{tieneAscensor sin ascensor, rooms: 1 habitacion, surface: surface_bajo, tieneAireAcondicionado sin aire} => {priceAmount: precio_bajo}	0.011	0.840	0.013	3.360	21
8705	{tieneTrastero sin trastero, tieneCalefaccion sin calefaccion, rooms: 1 habitacion, surface: surface_bajo, tieneAireAcondicionado sin aire} => {priceAmount: precio_bajo}	0.012	0.821	0.015	3.286	23
4284	{tieneCalefaccion sin calefaccion, rooms: 1 habitacion, surface: surface_bajo, tieneAireAcondicionado sin aire} => {priceAmount: precio_bajo}	0.013	0.806	0.017	3.226	25

Este análisis ha revelado patrones sólidos entre las características de los inmuebles y su rango de precio. En particular, propiedades con gran superficie, múltiples baños y habitaciones, ascensor, calefacción y aire acondicionado muestran una alta probabilidad de pertenecer al segmento de precio alto. En contraste, los inmuebles con superficie reducida, una sola habitación y ausencia de comodidades tienden a asociarse con precios bajos, lo que puede indicar oportunidades de inversión o infravaloración. Además se han creado una serie de gráficos que refuerzan este análisis (consultar en el anexo 9.4).

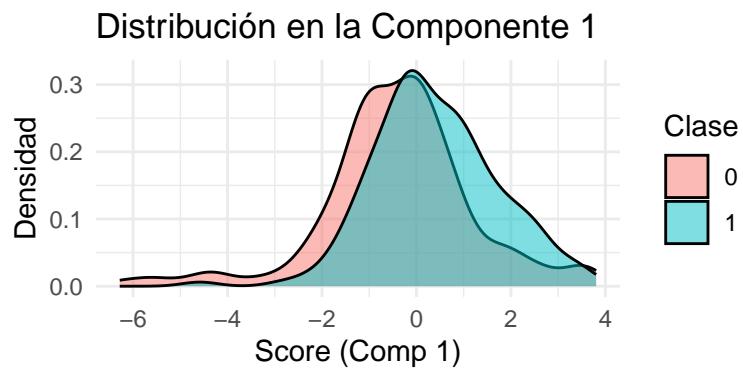
7 PLS-DA.

Antes de realizar este análisis, comprobamos que nuestra base de datos no cumplía las condiciones de homocedasticidad ni normalidad (ver en el 9.5). Esto quiere decir que no podemos aplicar un Análisis Discriminante de Fisher para predecir qué inmuebles serán rebajados. En su lugar, emplearemos un PLS discriminante; en el que consideraremos como POSITIVOS las viviendas cuyo precio sea menor al precio original. Para ello, hemos convertido la variable primitiva `priceAmountDrop` en `priceDropBinary` (0 si el precio baja, 1 si se mantiene estable). En la primera fase del análisis, seleccionamos la mayoría de variables disponibles (incluyendo el cluster por servicios diseñado en análisis previos); descartando aquellas cuya información resultaba inservible o redundante. Para evitar el diseño de un modelo sesgado, escalamos las variables numéricas y dividimos los datos en los conjuntos de entrenamiento (70%) y test (30%). Además, realizamos un undersampling en la base de entrenamiento eliminando aleatoriamente parte de los pisos sin bajada de precio; puesto que había un fuerte desbalanceo que podía favorecer sistemáticamente las predicciones de la clase predominante.

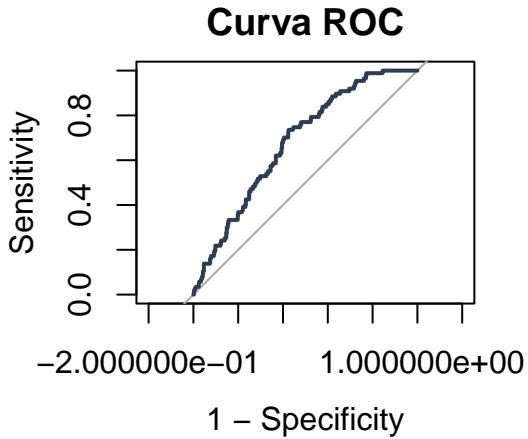


Seguidamente, propusimos el modelo inicial. Aplicamos la validación cruzada con 5 grupos o folds y 10 repeticiones hasta quedarnos con 1 única componente principal, minimizando así el error de predicción. Una vez aplicado el modelo al conjunto de test, maximizamos su exactitud balanceada de 0.59 a 0.632 variando el umbral óptimo. Seleccionamos esta métrica como medida de calidad porque tiene en cuenta los aciertos en ambas clases de pisos (0, 1), sin proporcionar resultados engañosos ni dejarse condicionar por el desbalanceo. Esta primera versión mostraba una capacidad predictiva aceptable, generando mejores resultados que los que se obtendrían por azar. Cabe recalcar que la bajada de precio de las viviendas está muy condicionada por el factor humano y las condiciones individuales de cada propietario. Por lo tanto, podríamos considerarla una variable de naturaleza impredecible o, al menos, con un comportamiento difícil de modelar. A continuación, nos propusimos incrementar la eficacia mediante un test de independencia. El objetivo era proponer un nuevo modelo formado por las variables que presentaran diferencias significativas entre clases. Así, descartamos la inclusión de variables sin capacidad predictiva, que únicamente generan ruido y disminuyen la fiabilidad de las predicciones. A las variables numéricas les aplicamos el test t y a las categóricas el test chi-cuadrada o, en caso de muestras pequeñas, la prueba de Fisher. Una vez realizada la selección, volvimos a dividir, escalar y reducir la nueva base de datos y generamos el segundo modelo; compuesto nuevamente por una sola componente principal.

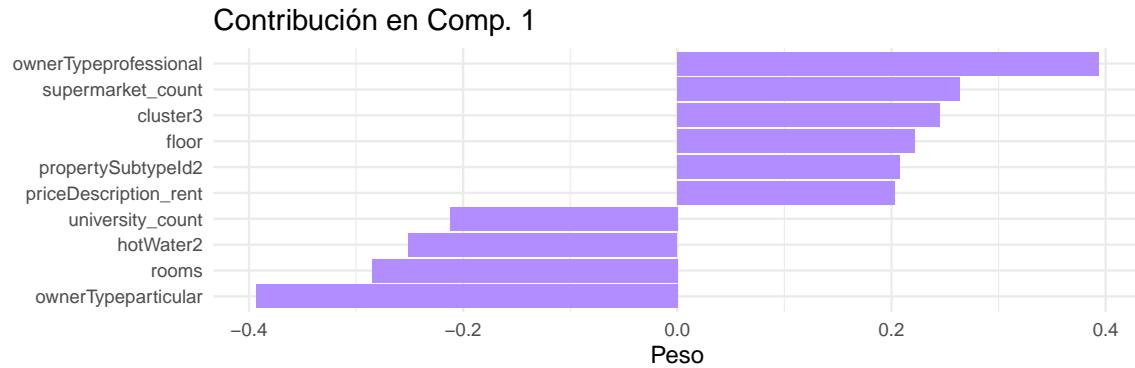
```
## Componentes óptimos: 1
```



Aunque la distribución obtenida era relativamente uniforme, maximizamos una vez más la exactitud balanceada a través de la curva ROC:



Finalmente, conseguimos un modelo que predice las bajadas de precio de los inmuebles con un 65% de exactitud balanceada. Esto es, cada 100 predicciones se consigue una media razonable de 65 recalls. Tal como hemos indicado en el inicio del análisis, esta variable está fuertemente condicionada por factores humanos: decisiones personales de propietarios, estrategias de agencias inmobiliarias, urgencia de venta o negociación, entre otros elementos difíciles de cuantificar. Por tanto, ningún modelo podrá predecirla con exactitud absoluta, y siempre quedarán fuera variables imposibles de medir que condicionarán los resultados. Aun así, cualquier herramienta que logre predecir las fluctuaciones en los precios, aunque sea parcialmente, es altamente valiosa; ya que permite orientar decisiones y ofrecer una ventaja estratégica en el análisis de mercado inmobiliario. Una vez elaborado el modelo definitivo, estudiaremos qué variables tienen más peso en la componente principal para entender qué características hacen a un piso más propenso a ser rebajado.

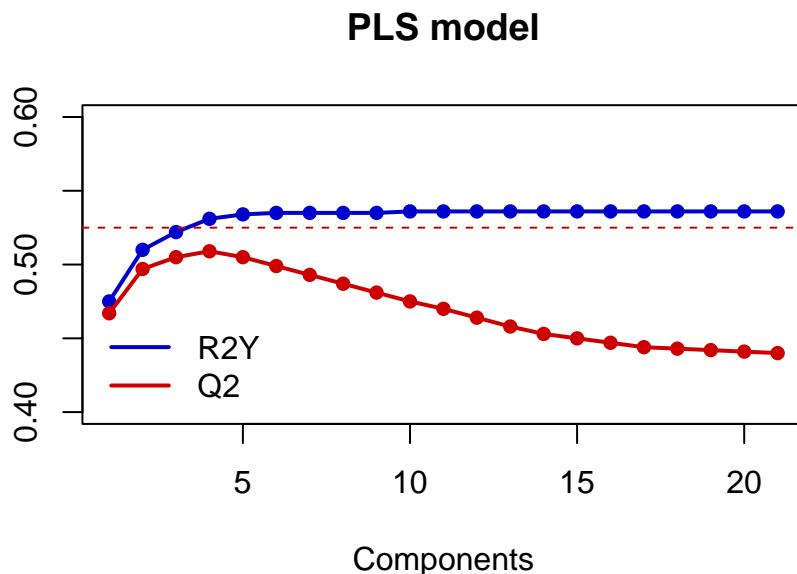


Resalta claramente que los propietarios profesionales (inmobiliarias, plataformas online, entidades bancarias...) son más propensos a reducir el precio que los propietarios particulares. Por otra parte, si nos centramos en la ubicación, en las zonas de Ruzafa, Benimaclet y Jesús (cluster3) las viviendas son más propclives a las rebajas. Asimismo, cuanto mayor sea el precio medio del alquiler en esta ubicación, más probable será que estos precios bajen. La cantidad de supermercados próximos también contribuye positivamente a dicha probabilidad. Por el contrario, precios de pisos en zonas universitarias son más estables. Respecto a las características del piso per se, las plantas altas tienen más probabilidad de ser rebajadas. En cambio, si tiene muchas habitaciones y/o calentador eléctrico (hotWater2) tiende a experimentar menos variaciones; especialmente en el caso de los apartamentos (porertySubtype2).

8 PLS

Escalamos tanto la matriz Y como la X. Estimaremos el número de componentes óptimo mediante validación cruzada. En este caso, al tener un número tan alto de observaciones, optaremos por el procedimiento “k-fold”, en nuestro caso generaremos 10 folds.

De acuerdo con el criterio de la función *opls*, el número óptimo de componentes sería 3. No obstante, vamos a generar nuestro propio gráfico para estimar mejor el número óptimo de componentes del modelo:



En el gráfico anterior podemos observar que con hasta 4 componentes el valor de Q^2 aumenta a la vez que el valor de R^2 . Sin embargo a partir de 5 componentes Q^2 empieza a disminuir ligeramente y R^2 se mantiene prácticamente constante.

Así pues, parece más adecuado seleccionar 4 componentes. Generamos a continuación el modelo con 4 componentes.

```
## PLS
## 1856 samples x 21 variables and 1 response
## standard scaling of predictors and response(s)
##      R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total     0.421     0.531    0.509    474    4   0 0.05 0.05
```

En el resumen se muestra que el modelo cuenta con una buena capacidad predictiva, con un Q^2 mayor a 0.5, es decir, explica más de un 50% de la variabilidad de la variable estudiada, PriceAmount, así como un error aceptable. Para poder entender mejor el modelo, vamos a clasificar las variables en función a la componente a la que pertenecen.

```
##                                         Variable      Peso
## surface                               surface 0.5441573
## bathrooms                            bathrooms 0.5124531
## rooms                                 rooms 0.3888462
## priceDescription_buy priceDescription_buy 0.3256391
## tieneAscensor                         tieneAscensor 0.1845652
```

Se han calculado las variables que más peso en cada componente. En este caso, se muestra solo la de la primera componente, puesto que es la que mayor variabilidad explica del modelo. Las variables superficie, número de baños, número de habitaciones, el precio de compra o si tiene ascensor son las más importantes de la primera componente.

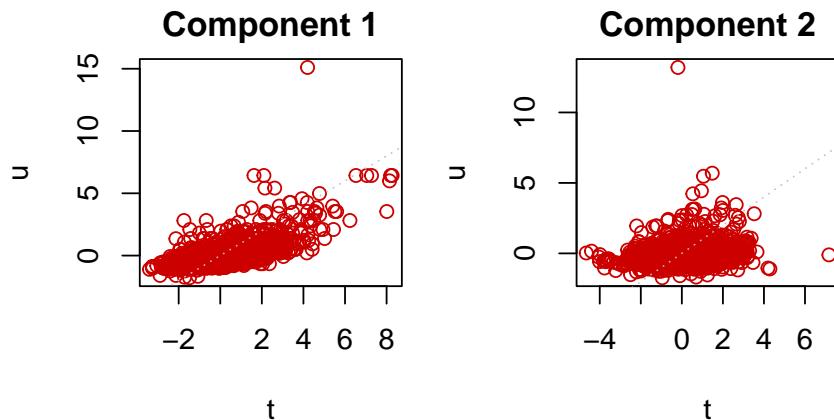
Al analizar el gráfico de scores con la elipse de Hotelling T2 (incluida en el anexo), se observa que la mayoría de las observaciones se encuentran dentro del límite esperado, lo que indica un comportamiento multivariado normal. Sin embargo, algunas observaciones se sitúan fuera de la elipse, lo que sugiere la posible presencia de valores atípicos.

Se realizarán análisis más específicos para identificarlos con mayor evidencia estadística, incluidos en el anexo.

Los resultados muestran que la mayoría de las observaciones están bien representadas y no presentan valores extremos. Sin embargo, se han identificado algunas que superan claramente los límites de confianza, tanto al 95% como al 99%. En particular, la observación n.^o 355 destaca como un outlier evidente según ambos criterios, mientras que otras, como las observaciones 136, 445 o 372, muestran discrepancias más moderadas pero persistentes en su representación por el modelo.

Tras revisar manualmente sus características y precios, comprobamos que se trata de viviendas reales, con valores justificados por su ubicación, dimensiones o nivel de equipamiento. Por tanto, al no presentar errores evidentes y contribuir a explicar parte de la variabilidad del mercado, decidimos mantener estas observaciones en el análisis.

A continuación, debemos evaluar uno de los supuestos del PLS, que es el supuesto de linealidad entre los scores de las variables predictoras y los scores de la variable respuesta. Buscamos representar dicha relación en un gráfico para analizar la relación.



En lo que respecta al gráfico de la componente 1, podemos observar una relación positiva entre t y u , aunque para valores más altos de u se observa cierta dispersión. Sin embargo, como la mayoría de puntos sigue la relación buscada, podemos afirmar que la primera componente captura una relación lineal significativa entre X e Y. En cuanto al gráfico de la componente 2, la relación es mucho menos evidente, puesto que no se observa una relación clara y los puntos están considerablemente más dispersos. Esto sugiere que dicha componente no contribuye en gran medida a la relación lineal entre las matrices, ya sea por estar capturando ruido o por tener una estructura de menor relevancia. Con el fin de obtener una mayor evidencia en las conclusiones, se va a calcular cada una de las matrices de correlaciones de los componentes.

```
##          p1          p2          p3          p4
## 0.6890911 0.2586998 0.1534190 0.1431952
```

Podemos observar que la primera componente presenta una correlación moderadamente alta, lo que indica una relación lineal apreciable entre las variables latentes de los conjuntos predictivo y de respuesta. En contraste, las componentes restantes muestran correlaciones débiles, lo que refuerza las conclusiones obtenidas a partir de los gráficos anteriores.

Siguiendo con el análisis, se debe de evaluar la capacidad del modelo creado para reconstruir las variables predictoras originales. Esto se consigue mediante el coeficiente de determinación R², calculado en la siguiente tabla.

Table 3: Top 10 variables según R²

	Variable	R2
surface	surface	0.78
pharmacy_count	pharmacy_count	0.77
college_count	college_count	0.74
propertyCounter_buy	propertyCounter_buy	0.70
propertyCounter_rent	propertyCounter_rent	0.65
bathrooms	bathrooms	0.61
supermarket_count	supermarket_count	0.59
public_transport_count	public_transport_count	0.48
rooms	rooms	0.47
tieneCalefaccion	tieneCalefaccion	0.47

El R² global indica que el modelo explica aproximadamente el 42% de la varianza total de las variables X, lo cual sugiere una capacidad explicativa moderada. A nivel individual, algunas variables como surface, college_count y pharmacy_count presentan valores de R² notablemente altos, evidenciando que el modelo captura bien su estructura. En contraste, variables como tieneTrastero o hospital_count muestran una explicación muy limitada, lo que podría deberse a ruido, no linealidad, o falta de relación con las componentes extraídas.

Tras evaluar la calidad de la reconstrucción del espacio de las variables independientes X, procedemos a calcular el coeficiente de determinación R² en el espacio de las variables dependientes Y. Este valor nos permite estimar qué proporción de la varianza de la(s) variable(s) objetivo está siendo explicada por el modelo PLS construido.

```
## $R2_kcum
## [1] 0.5313373
##
## $R2cum
## [1] 0.5313373
```

El valor global de R² obtenido para Y ha sido de 0.5313, lo que indica que el modelo explica aproximadamente el 53.13% de la varianza de la(s) variable(s) dependiente(s). Este resultado sugiere que el modelo tiene una capacidad predictiva moderadamente buena sobre las variables objetivo, y en este caso, mejor que la reconstrucción obtenida en el espacio de X (que fue de 42%). Por tanto, se puede considerar que el modelo PLS ha conseguido capturar relaciones relevantes entre los componentes latentes y la variable de respuesta.

Con el fin de comprobar si se puede conseguir un modelo más simple con mejores resultados, se va a crear un nuevo modelo con las variables más significativas. Todos los cálculos y resultados se adjuntan en el anexo.

Ambos modelos muestran una tendencia positiva coherente respecto a la diagonal (línea de identidad), lo cual indica una capacidad predictiva aceptable. Luego, aunque el modelo reducido tiene una ligera pérdida de precisión (con un R² algo más bajo y un RMSE más alto que el modelo completo), gana en simplicidad y es mucho más fácil de interpretar. Usar solo las variables más relevantes puede ser útil, sobre todo si queremos entender mejor el modelo o si no siempre es posible recoger toda la información disponible. Por eso, aunque el modelo completo tiene mejor rendimiento, el modelo reducido con las variables seleccionadas por VIP también es una buena opción dependiendo del objetivo que tengamos.

En este caso, el objetivo es crear un modelo de predicción de precios de vivienda lo más preciso posible, luego optamos por el modelo con todas las variables ya que a pesar de ser más complejo, ofrece una capacidad predictiva mayor que el modelo simplificado.

Para finalizar, realizaremos un análisis para tratar de detectar posibles errores de predicción del modelo.

```
##      R2      RMSE      MAE     MAPE
## 1 0.5356 471.2649 292.1782 18.868
```

Una vez ajustado el modelo PLS con todas las variables, se calcularon diversas métricas de error para evaluar su rendimiento. En concreto, se obtuvieron los siguientes valores: $R^2 = 0.5356$, $RMSE = 471.2649$, $MAE = 292.1782$ y $MAPE = 18.868\%$. Estas métricas permiten valorar la calidad del ajuste del modelo, proporcionando información sobre su capacidad explicativa y el tamaño medio del error.

El valor de R^2 indica que el modelo es capaz de explicar aproximadamente el 53% de la variabilidad presente en los datos de salida. Aunque no se trata de un valor extremadamente alto, sí sugiere que el modelo capta cierta estructura subyacente en los datos. Por otro lado, el $RMSE$ y el MAE muestran el tamaño medio del error de predicción en las mismas unidades que la variable respuesta, lo que permite dimensionar cuánto puede desviarse una predicción típica respecto al valor real. En este caso, los errores medios rondan entre 296 y 473 unidades, dependiendo de cómo se midan (absoluto o cuadrático). Finalmente, el $MAPE$, con un 18.868%, señala que el modelo comete, de media, un error del 19% respecto al valor real, lo cual puede considerarse aceptable o no dependiendo del contexto de la aplicación.

Es importante tener en cuenta que estos resultados se han obtenido utilizando los mismos datos con los que se entrenó el modelo. Por tanto, no reflejan necesariamente cómo se comportaría el modelo ante nuevos datos no observados, sino que indican su nivel de ajuste dentro del propio conjunto utilizado para construirlo.

Por tanto, se concluye que el modelo presenta un equilibrio adecuado entre ajuste y capacidad predictiva, lo que justifica su utilidad para el análisis. Con 4 componentes, explica un 42.1% de la variabilidad de las variables explicativas y un 53.1% de la variabilidad de la variable respuesta, con una capacidad predictiva interna (Q_2) del 50.9%. Además, se detectaron tres observaciones potencialmente anómalas, lo que resulta relevante al interpretar la estabilidad del modelo. En conjunto, estos valores respaldan la calidad del ajuste realizado y proporcionan una base sólida para el análisis posterior.

9 ANEXO.

9.1 PCA.

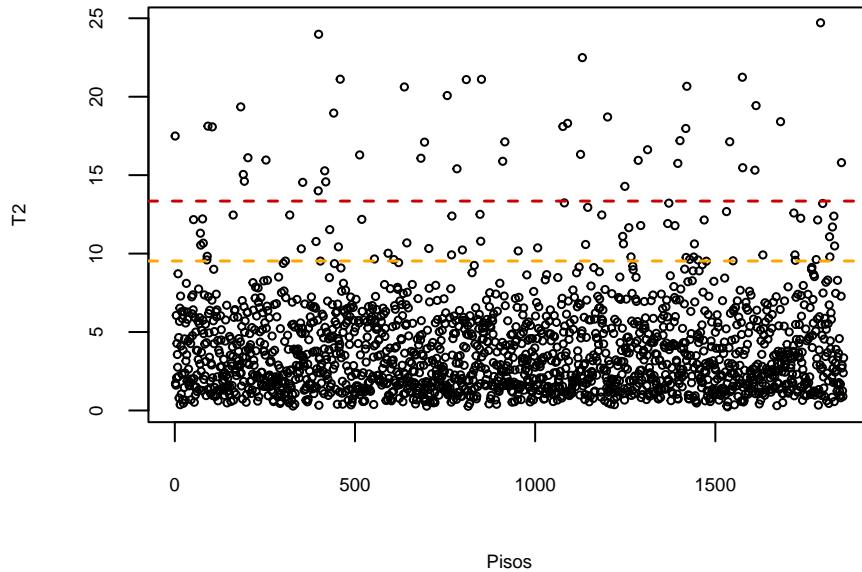
9.1.1 Atípicos.

Tras la selección de componentes, procedimos a evaluar primeramente la presencia de observaciones atípicas extremas dentro del espacio generado por las componentes principales. Para ello, utilizamos el estadístico T^2 de Hotelling, que nos permitió medir la distancia de cada observación al centro de la nube de puntos proyectada sobre las primeras 4 componentes.

Este análisis lo llevamos a cabo utilizando las coordenadas (o scores) de cada piso sobre las componentes principales más relevantes, y establecimos umbrales de referencia basados en la distribución F para los niveles de confianza del 95% y 99%. De esta forma, pudimos identificar qué observaciones se alejaban significativamente del comportamiento general del conjunto de datos.

Visualizamos estas distancias T^2 en un gráfico, donde marcamos las observaciones que superaban los valores críticos. Estas fueron clasificadas como potenciales outliers o casos extremos.

Con este procedimiento, buscábamos validar la estructura de los datos en el nuevo espacio reducido y garantizar que las observaciones atípicas no afectarán negativamente al análisis.

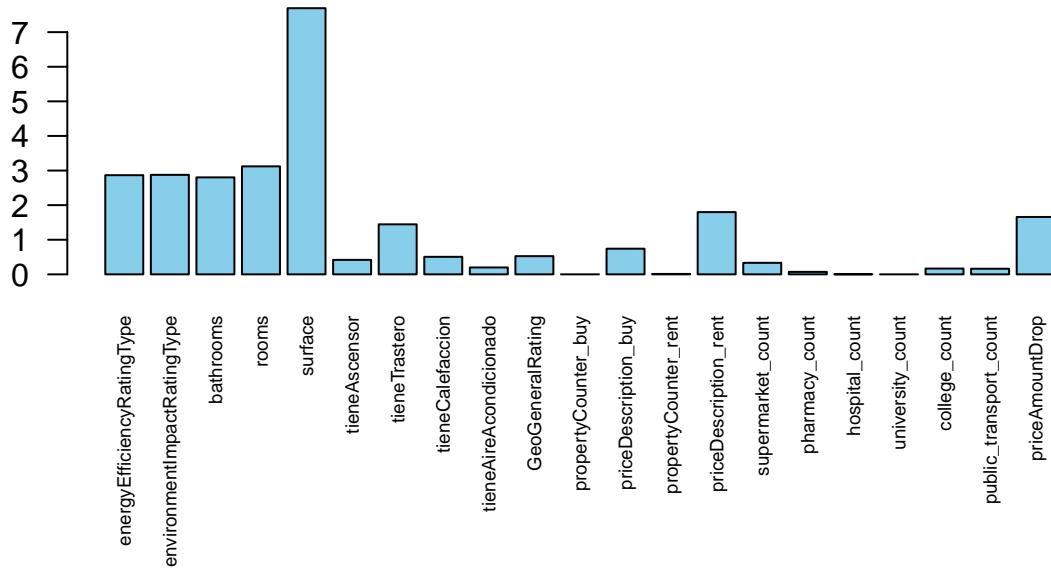


De los pisos detectados como posibles outliers según la distancia T^2 de Hotelling, se analizó en detalle aquel con el valor más extremo, utilizando una función propia diseñada para este propósito. Esta función calcula las contribuciones individuales de cada variable a la distancia T^2 de una observación, lo que permite entender con mayor precisión por qué un piso se aleja del centro del espacio PCA. Para ello, se usaron los autovectores desescalados, recuperando la importancia real de cada variable en los componentes principales, y se identificaron aquellas que más aportaban al valor elevado de T^2 .

```
## [1] "https://www.fotocasa.es/es/alquiler/vivienda/valencia-capital/aire-acondicionado-calefaccion-te"
##           Dim.1      Dim.2      Dim.3      Dim.4
## -2.5791645  5.6962296 -4.5660363 -0.7612643
```

Antes de realizar el gráfico vimos que las características del piso analizado eran coherentes. El piso seleccionado es un ático de lujo en el centro de Valencia, por lo que es normal que tenga características fuera de lo normal como un número elevado de baños o mucha superficie, y no quiere decir que sean errores de codificación ni valores anómalos injustificados. Por tanto, se asumió que el resto de observaciones con distancias elevadas también respondían a casos reales y válidos, y no se consideró necesario excluirlas del análisis.

Ático Valencia

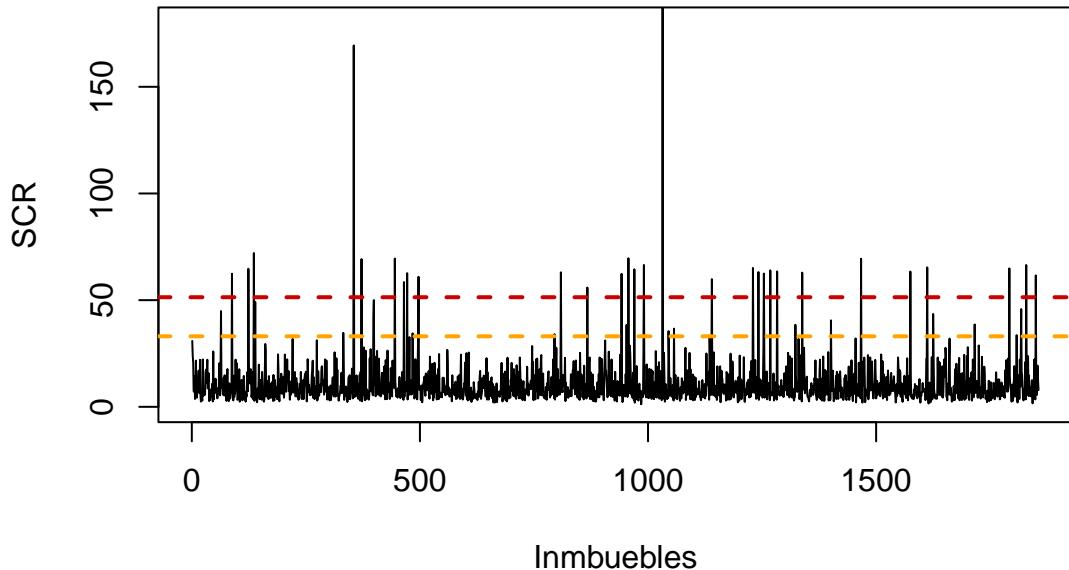


Como vemos efectivamente, las variables que más contribuyen para que el modelo detecte al inmueble como atípico son variables como “surface”, “bathrooms” o “rooms”.

Tras analizar la distancia de Hotelling T^2 , el siguiente paso fue realizar un gráfico de distancia al modelo, basado en la Suma de Cuadrados de los Residuos (SCR). El objetivo en este caso era identificar atípicos moderados, es decir, aquellos inmuebles cuyas características no estaban siendo capturadas adecuadamente por las componentes principales seleccionadas. Representamos gráficamente los SCR incorporando límites de referencia basados en una distribución chi-cuadrado (al 95% y 99% de confianza), con el objetivo de detectar qué pisos presentan información que no es bien explicada por los componentes principales seleccionados.

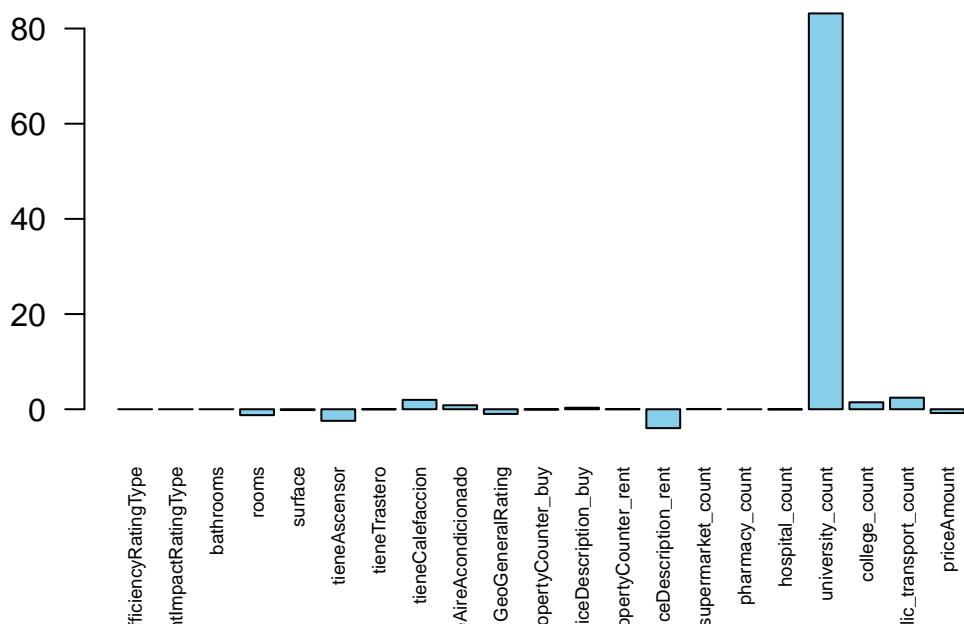
Con este gráfico detectamos observaciones que aunque no se encontraban particularmente alejadas del centro del modelo (baja T^2), no se ajustaban bien a la estructura global de los datos.

Distancia al modelo



Tras identificar los pisos que no se ajustaban bien al modelo, es decir, con un SCR superior a los umbrales establecidos, procedimos a analizar sus características particulares. Analizamos en detalle uno de ellos mediante el cálculo de las contribuciones individuales de cada variable a su alto SCR. Así podríamos ver que variables contribuían a que el inmueble no fuera bien representado por el modelo.

Contribuciones a SCR para un piso



Observamos que este inmueble no presentaba errores evidentes ni valores imposibles, sino que simplemente se trataba de una vivienda con una configuración no tan común, ya que tiene muchas universidades cerca.

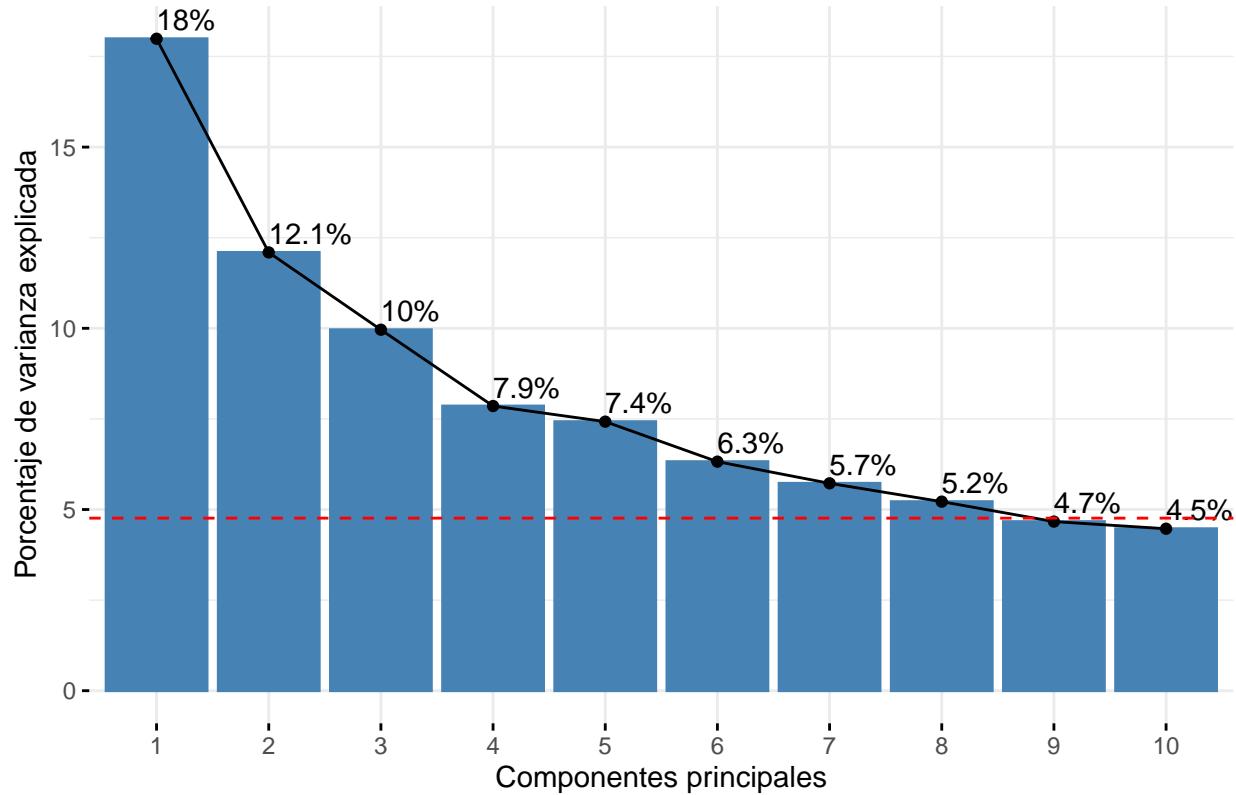
Dado que el comportamiento atípico del inmueble se explicaba por una combinación no tan común, pero coherente de variables, consideramos que las demás observaciones con alta SCR probablemente reflejan también perfiles particulares dentro del mercado, y no errores en los datos.

Por tanto, decidimos no excluirlas del análisis, ya que la técnica de PCA captura las tendencias generales del conjunto, pero es esperable que algunas viviendas reales, aunque menos frecuentes, se representen peor.

Manteniendo estas observaciones podemos reflejar con mayor fidelidad la diversidad y complejidad del mercado inmobiliario analizado.

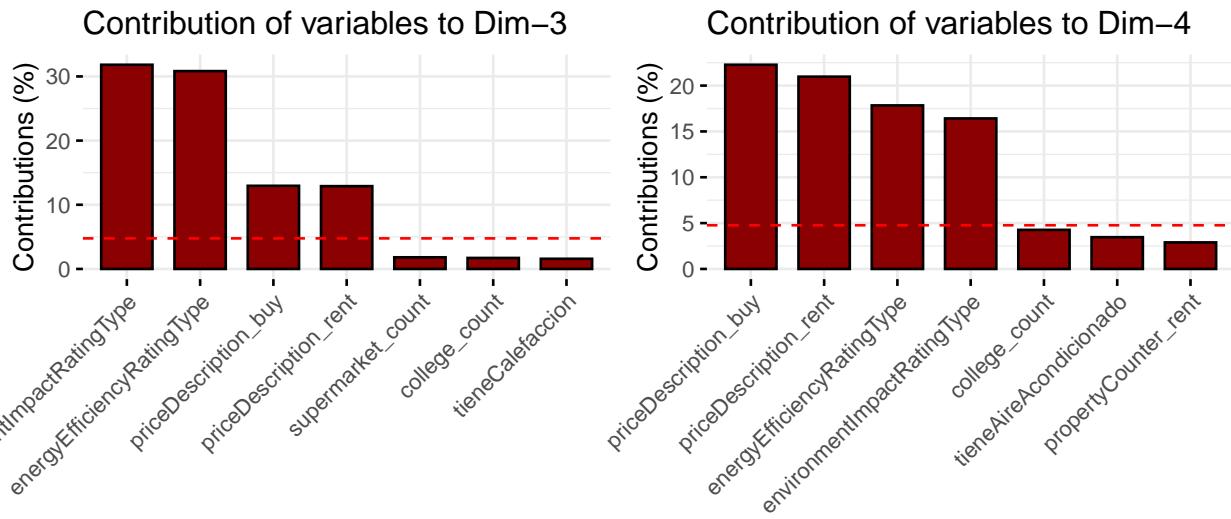
Cabe mencionar que a pesar de ya haber analizado los datos atípicos y concluir que no eran datos erróneos, realizamos un nuevo modelo PCA sin los atípicos extremos (T_2 alta), para ver si estaban afectando mucho a la hora de crear las componentes principales. Pero, tras filtrar los pisos anómalos extremos y excluirlos del modelo, el resultado del PCA apenas variaba. Por lo tanto, estos inmuebles no estaban generando componentes innecesarias ni influyendo en los resultados del análisis. Teniendo esto en cuenta, continuamos el análisis manteniendo los pisos atípicos en el modelo.

Varianza explicada por cada componente (PCA)



Cabe mencionar que a pesar de ya haber analizado los datos atípicos y concluir que no eran datos erróneos, realizamos un nuevo modelo PCA sin los atípicos extremos (T_2 alta), para ver si estaban afectando mucho a la hora de crear las componentes principales. Pero, tras filtrar los pisos anómalos extremos y excluirlos del modelo, el resultado del PCA apenas variaba. Por lo tanto, estos inmuebles no estaban generando componentes innecesarias ni influyendo en los resultados del análisis. Teniendo esto en cuenta, continuamos el análisis manteniendo los pisos atípicos en el modelo.

9.1.2 Gráficos de las contribuciones de las variables a las dimensiones.



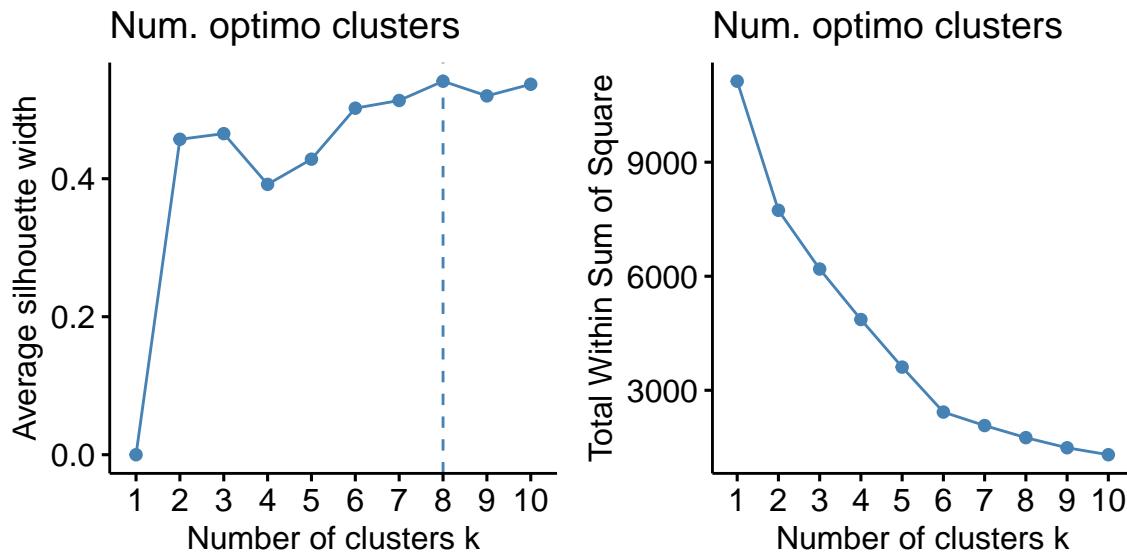
9.2 Clustering.

9.2.1 Clustering de Servicios

9.2.1.1 Prueba de distancias A continuación, se muestran los resultados obtenidos con distintas métricas de distancias que fueron evaluadas pero finalmente descartadas en favor de la distancia euclídea, por ofrecer menor cohesión entre grupos o menor interpretabilidad.

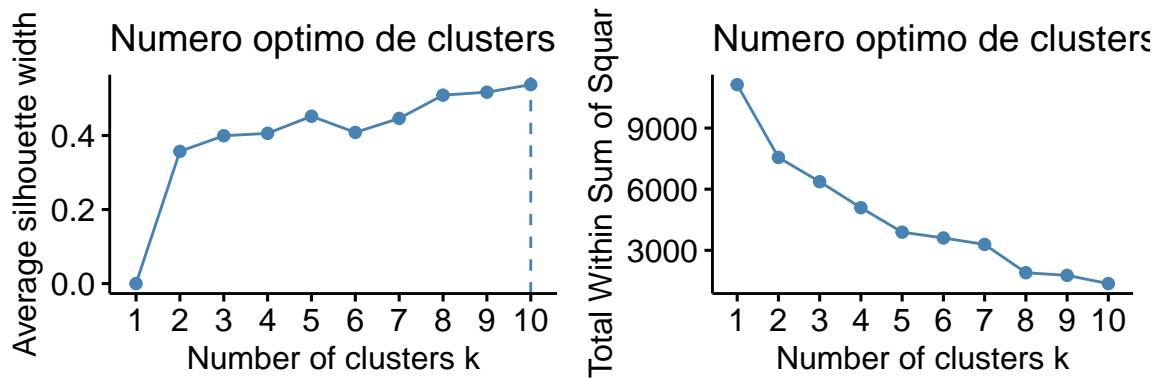


9.2.1.2 Selección de la técnica de agrupación y del número de clusters -Método jerárquico de WARD En este caso, seleccionamos el resultado con **8 clusters**. Se trataba de la organización con mayor coeficiente de Silhouette, además de una suma de cuadrados residual muy reducida.



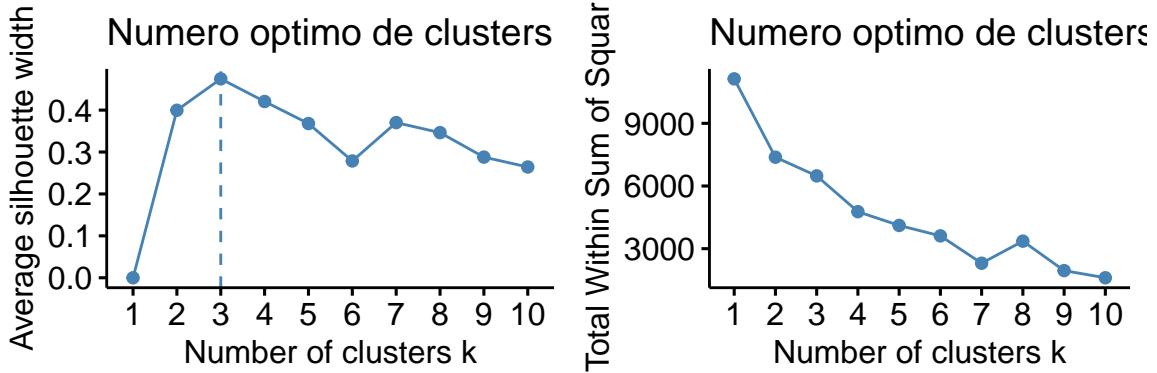
-Método de partición PAM

A través de la técnica PAM (k-medoides), obteníamos el máximo coeficiente y el mínimo SCR con 10 clusters. Sin embargo, la partición en **8 grupos** reducimos la complejidad del análisis con índices de calidad muy similares.

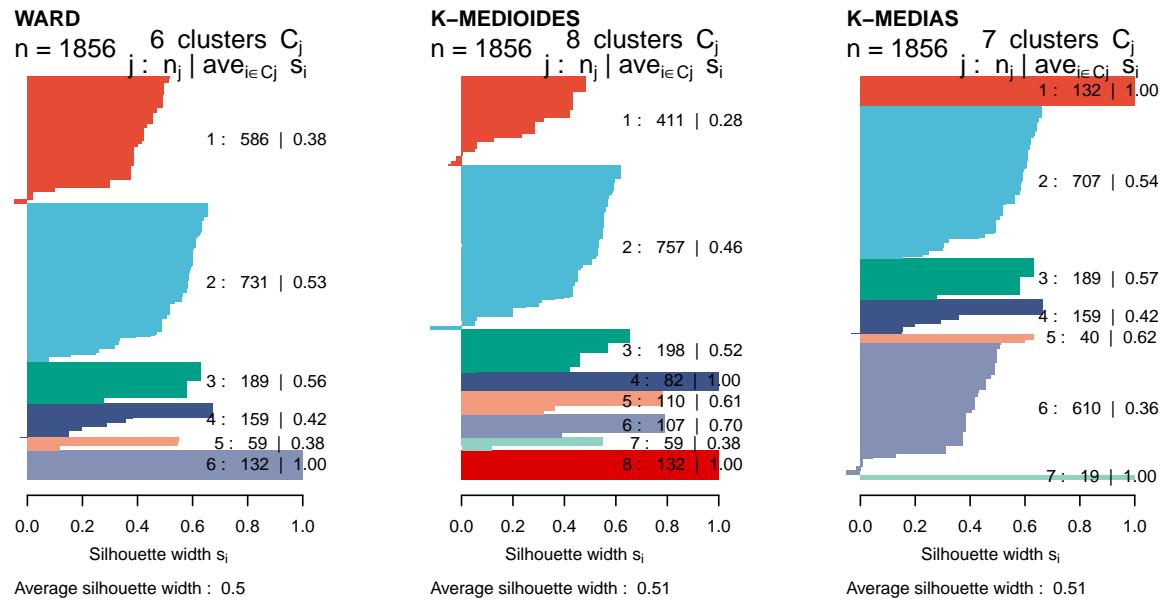


-Método de partición k-medias

Finalmente, con k-medias se maximizaba el coeficiente de Silhouette dividiendo los pisos en 3 grupos. No obstante, la SCR era todavía excesivamente elevada. La opción que mejor combinaba ambos parámetros era **k=7**.

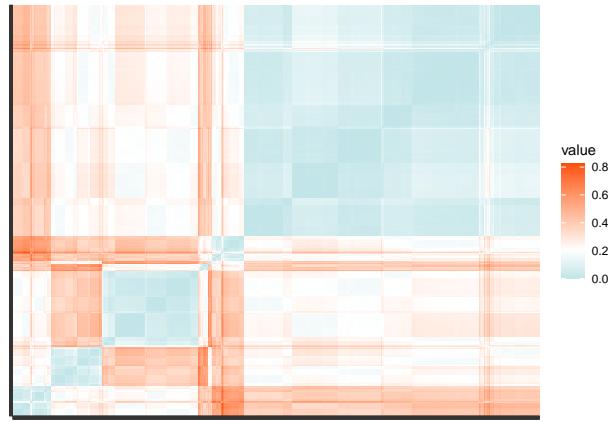


Una vez decidido el número de divisiones en cada método, comparamos los clusters resultantes. Vimos que, aunque los resultados eran similares, las **particiones de k-medias** contenían generalmente menos inmuebles mal clasificados.



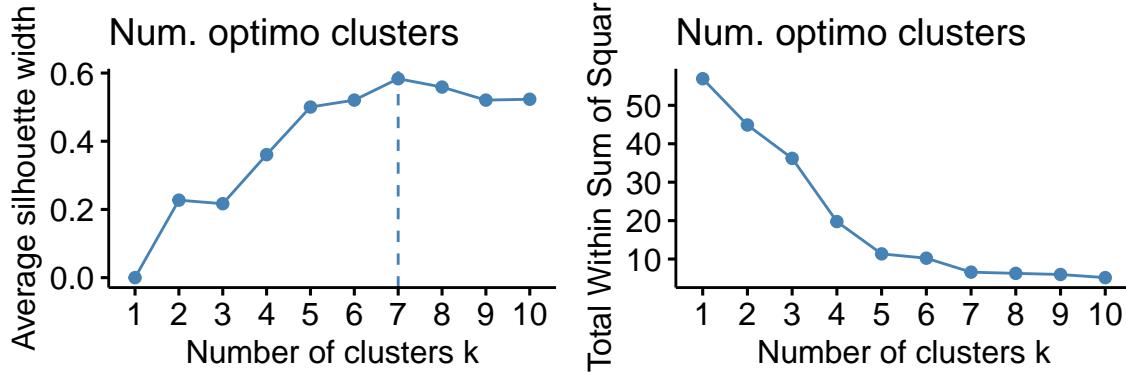
9.2.2 Clustering de Confort

9.2.2.1 Distancias y selección del método de agrupación Dado que el conjunto que describe las comodidades ofrecidas por cada piso contiene tanto variables categóricas como numéricas, únicamente podíamos consultar las tendencias de agrupamiento con la distancia de Gower.

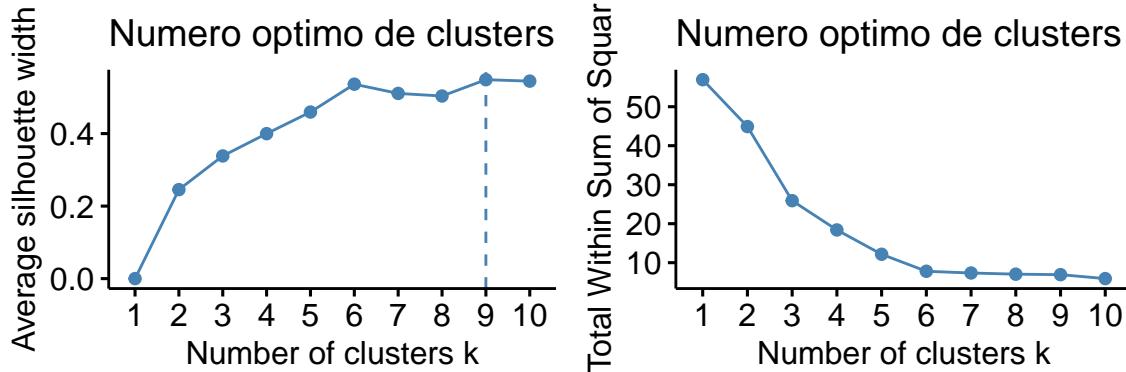


En el mapa de calor encontramos evidencias de la existencia de subgrupos en el conjunto de datos. Con el objetivo de detectar dichos clusters e interpretarlos, comenzamos el análisis probando las técnicas de Ward y k-medoides para determinar cuál generaba resultados más definidos. No aplicamos k-medias porque no habíamos utilizado la distancia euclídea.

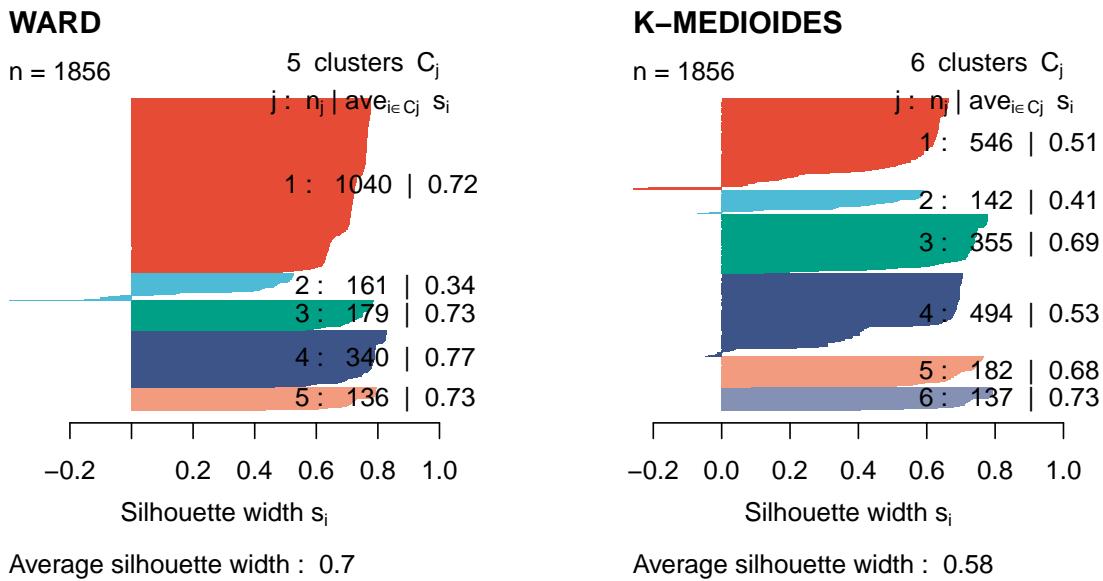
- Método de WARD ($k=5$)



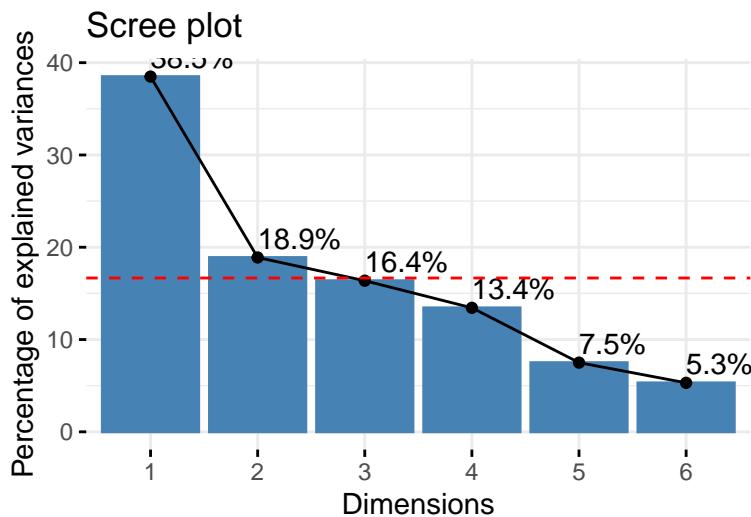
-Método PAM ($k=6$)



Finalmente, nos decantamos por el método *Ward* porque su coeficiente medio de Silhouette era notablemente superior.

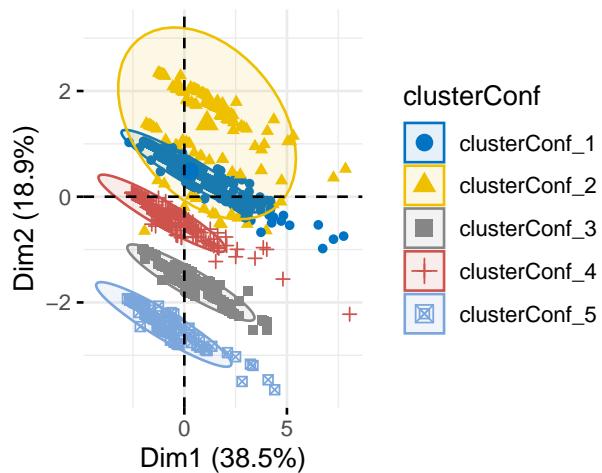


9.2.2.2 Análisis de Componentes Principales Se realizó un PCA de 2 componentes (casi el 60% de la variabilidad) para comprender mejor las características de cada Cluster.

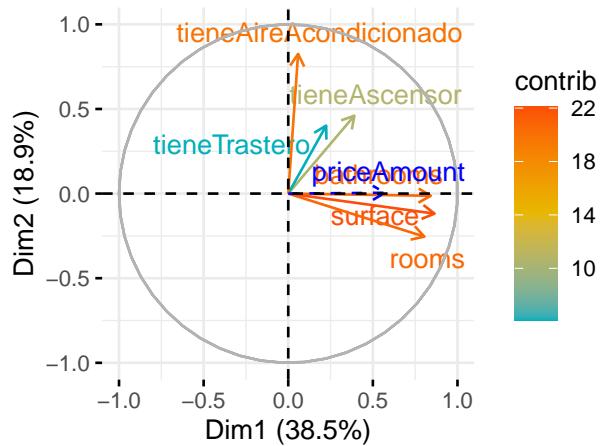


- Primera componente (*Tamaño útil*): mayor superficie, más habitaciones y baños (tamaño del piso), positivamente relacionado con el precio. Los clusters se reparten de forma relativamente uniforme, aunque el grupo 1 tiene más viviendas con valores altos en esta variable latente.
 - Segunda componente (*Equipamiento interior*): más pisos con aire acondicionado, calefacción, ascensor... Entre otras infraestructuras que aportan valor añadido a la vivienda. En esta dimensión hay una segmentación más clara de los clusters. El cluster 2 destaca positivamente, seguido del 1. Los clusters 4 y 5 son los que peores condiciones ofrecen.

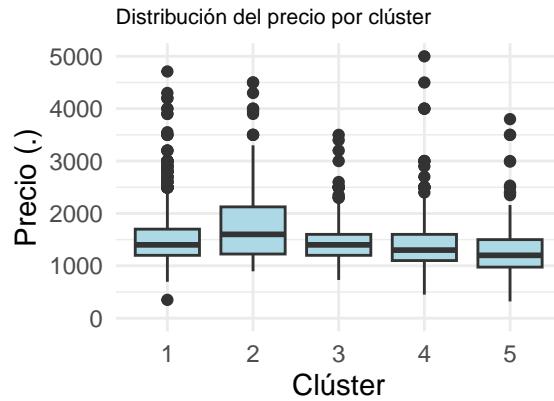
Individuos por grupo (con elipses)



Loading plot: variables



Clúster	Precio medio (.)
1	1577.748
2	1852.099
3	1476.313
4	1454.141
5	1346.397

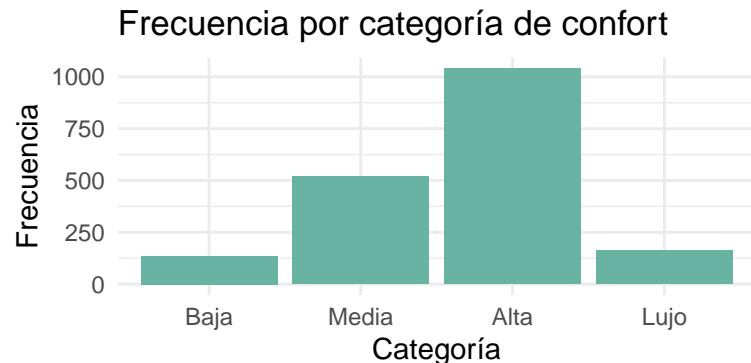


Si comparamos el precio medio de cada cluster, vemos el orden es casi idéntico a la estructura que siguen a lo largo de la segunda componente; únicamente se intercambian los clusters 3 y 4, que a su vez son los que tienen un promedio más parecido. Por lo tanto, podríamos considerar el cluster como una variable ordinal que nos indica el la categoría de cada vivienda (muy relacionada con su precio). Esta división no coincide estrictamente con el tamaño de la vivienda, sino que está más basada en sus condiciones de habitabilidad. Los niveles serían los siguientes:

Table 4: Correspondencia entre nivel de calidad y clústers

Nivel	Calidad	Clusters_originales
1	Baja	5
2	Media	3-4
3	Alta	1
4	Lujoso	2

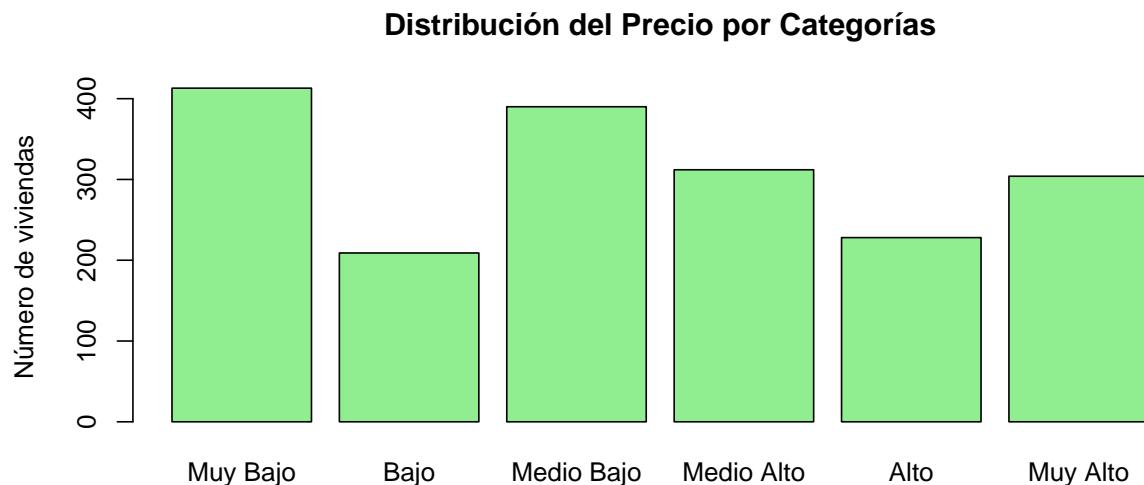
A continuación, se muestran las frecuencias de pisos en cada una de las nuevas categorías.



9.3 AFC.

9.3.1 AFC simple: Municipio vs. Precio

9.3.1.1 Transformación: precios a categorías Con el fin de realizar un Análisis Factorial de Correspondencias sobre las variables “Municipio” y “Precio” se ha realizado una transformación previa sobre la variable precio para poder realizar el AFC. “Precio” se ha transformado a 6 categorías ordinales clasificándolo de precio muy bajo a precio muy alto.



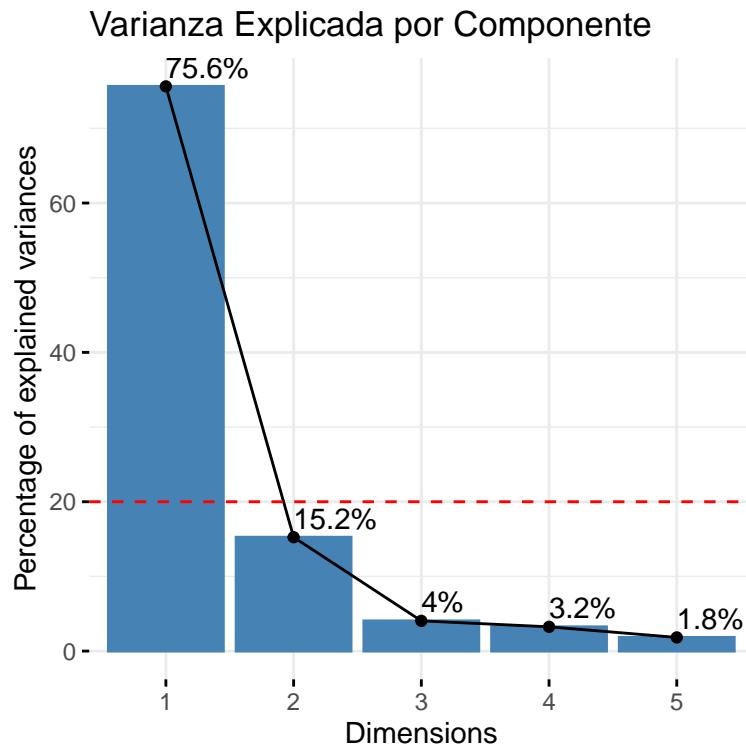
Este gráfico de barras muestra la distribución de las viviendas según su categoría de precio, creada a partir de cuantiles. El eje Y indica el número de viviendas que pertenecen a cada categoría (por ejemplo: “Muy Bajo”, “Bajo”, “Alto”, etc.).

9.3.1.2 Tabla de contingencia y Chi-cuadrado Con el fin de contrastar la hipótesis nula de independencia entre ambas variables, se aplicará un test de independencia χ^2 . Para ello se genera la tabla de contingencia de Municipio y Precio, y se aplica el test.

Como resultado se obtiene un p-valor de 0,0005, por lo que se rechaza la hipótesis nula y se concluye que existe una dependencia estadísticamente significativa entre el municipio y el precio. Así pues, tiene sentido que estudiemos la naturaleza y causas de dicha dependencia con un AFC simple.

A continuación se realiza el AFC comenzando por la selección del número de componentes. Según el siguiente gráfico se puede ver que la primera componente ya explica un 75,6% de la inercia total de la tabla de contingencia, mientras que la segunda apenas explica un 15,2%, por debajo de la inercia media teórica (20%). Por tanto se elegiría solo la primera componente, no obstante, se elgiran las dos primeras para poder generar los gráficos de filas y columnas para poder interpretar el AFC no dándole importancia a la segunda componente.

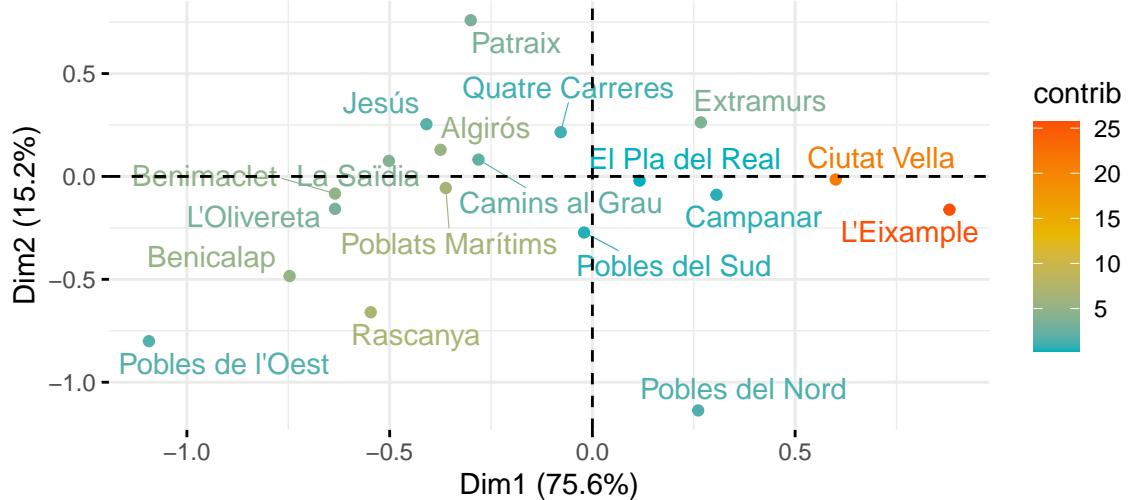
9.3.1.3 Selección del número de dimensiones



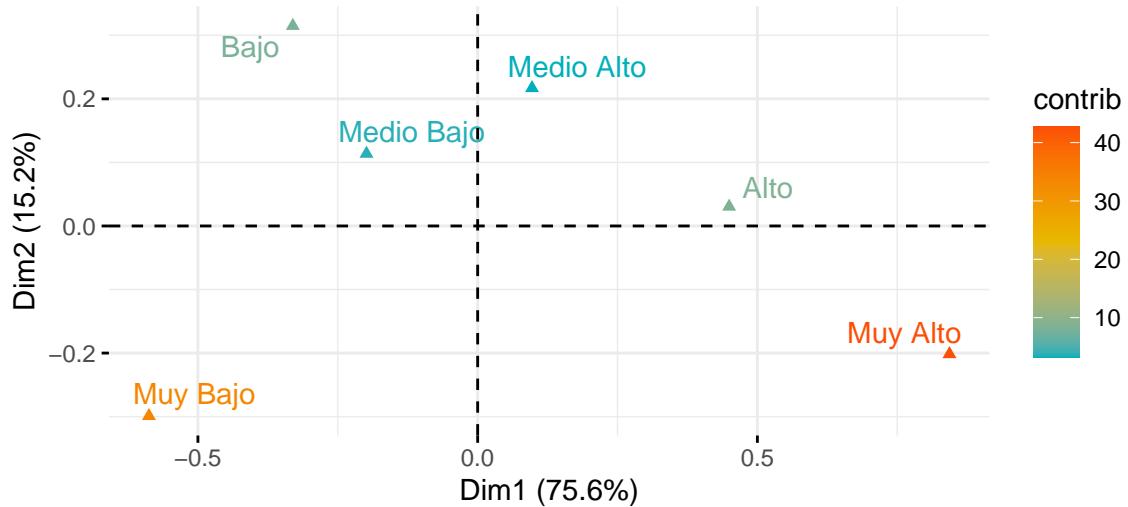
9.3.1.4 Interpretación en el plano factorial El primer eje (Dim1), que explica la mayor parte de la inercia (75,6%), separa claramente los municipios según el nivel de precios: a la izquierda se agrupan distritos como Benicalap o Pobles de l'Oest, asociados a precios bajos, mientras que a la derecha destacan L'Eixample y Ciutat Vella, vinculados a precios muy altos. El segundo eje (Dim2), con menor peso (15,2%), añade variabilidad secundaria sin una interpretación tan clara.

En conjunto, el AFC refleja una estructura clara de dependencia entre municipio y nivel de precio, coherente con la geografía socioeconómica de la ciudad.

Municipios según su perfil de precios



Categorías de Precio según municipios



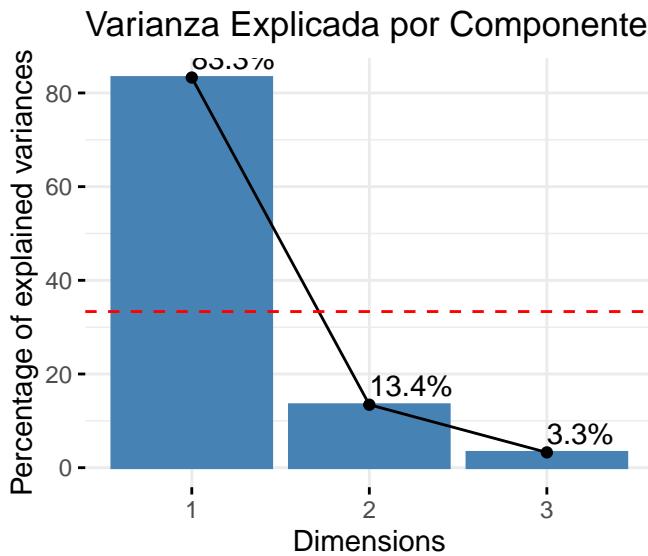
9.3.2 Justificación AFC simple: Subtipo vs. Precio

Con el fin de realizar un Análisis Factorial de Correspondencias sobre las variables “Subtipo” y “Precio” se han realizado transformaciones previas sobre dichas variables para poder realizar el AFC. La variable “Subtipo” se ha mapeado para traducir los códigos de propertySubtypeId a nombres descriptivos. Por otro lado, el Precio se ha transformado a 6 categorías ordinales clasificándolo de precio muy bajo a precio muy alto.

9.3.2.1 Dependencia entre Subtipo y Precio Con el fin de contrastar la hipótesis nula de independencia entre ambas variables, se aplicará un test de independencia χ^2 . Para ello se genera la tabla de contingencia de Subtipo y Precio, y se aplica el test.

Como resultado se obtiene un p-valor de 0,0005, por lo que se rechaza la hipótesis nula y se concluye que existe una dependencia estadísticamente significativa entre el subtipo de vivienda y el precio. Así pues, tiene sentido que estudiemos la naturaleza y causas de dicha dependencia con un AFC simple.

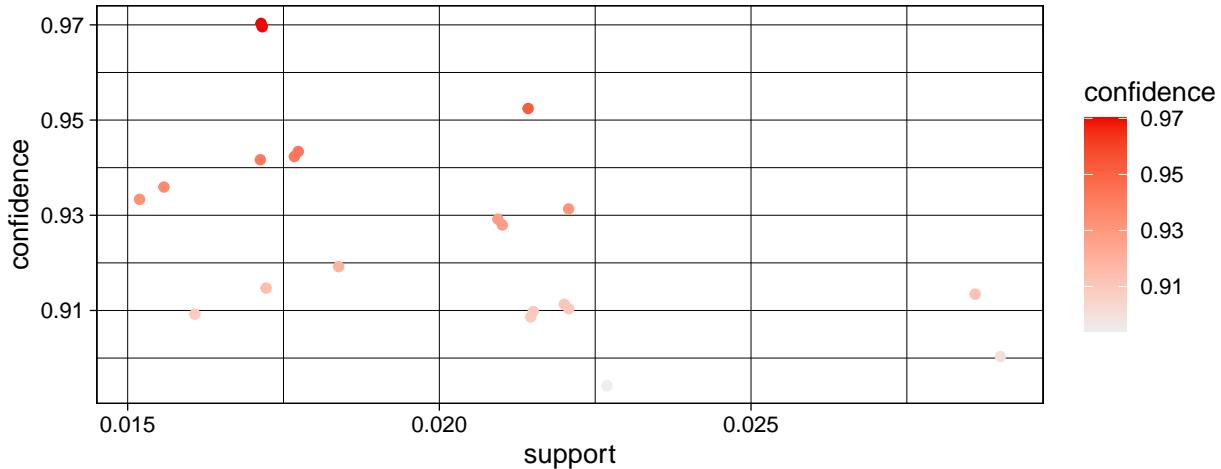
9.3.2.2 Análisis Factorial de Correspondencias A continuación se realiza el AFC comenzando por la selección del número de componentes. Según el siguiente gráfico se puede ver que la primera componente ya explica un 83,3% de la inercia total de la tabla de contingencia, mientras que la segunda solo explica un 13,4%, claramente por debajo de la inercia media teórica (aproximadamente el 32%). Por tanto se elegiría solo la primera componente, no obstante, se elgiran las dos primeras para poder generar los gráficos de filas y columnas para poder interpretar el AFC no dándole importancia a la segunda componente.



9.4 Reglas de Asociación.

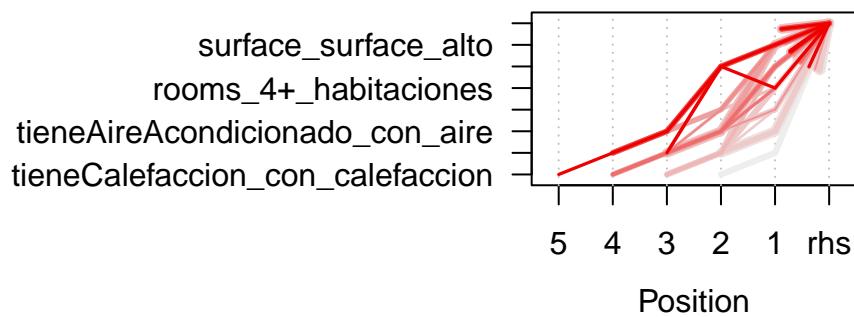
A continuación, se han empleado gráficos para explorar visualmente las reglas de asociación: un **diagrama de dispersión (scatter plot)** para analizar la relación entre soporte y confianza, y un **gráfico de coordenadas paralelas** para examinar la estructura interna de las reglas en términos de los atributos que las componen. No se utilizaron representaciones como el grouped matrix ni la matriz de calor, ya que estas técnicas son más adecuadas con combinaciones simples entre pocos ítems. Al estar las reglas formadas por múltiples atributos simultáneos, genera una elevada dimensionalidad y dificulta la interpretación gráfica en formatos matriciales.

Scatter plot for 21 rules



El gráfico de dispersión muestra la relación entre el soporte y la confianza de las 21 reglas de asociación seleccionadas, con una escala de color que representa el nivel de confianza. Todas las reglas tienen un soporte comprendido entre 0.015 y 0.025, lo cual indica que aunque no son extremadamente frecuentes. En términos de confianza, todas superan el umbral mínimo de 0.7, alcanzando valores próximos al 0.97. Esto indica que cuando se cumplen las condiciones del antecedente, la probabilidad de que el precio sea alto es muy elevada. Las reglas con mayor confianza tienden a tener un soporte ligeramente menor, lo que sugiere que, si bien son altamente fiables, aplican a un subconjunto más específico del mercado inmobiliario.

Parallel coordinates plot for 21 rules



El gráfico de coordenadas paralelas representa visualmente la estructura interna de las reglas más representativas. Cada línea corresponde a una regla individual, y su recorrido conecta los distintos atributos que forman el antecedente, finalizando en el consecuente (priceAmount_precio_alto).

Se aprecia una clara convergencia de reglas hacia ciertos atributos comunes, como: Superficie elevada del inmueble (surface_surface_alto), Existencia de ascensor (tieneAscensor_con_ascensor), Más de 3 baños (bathrooms_3+_banios) y Existencia de calefacción (tieneCalefaccion_con_calefaccion). Estas características aparecen reiteradamente como predictores del precio alto, lo que refuerza su relevancia dentro del conjunto de datos analizado.

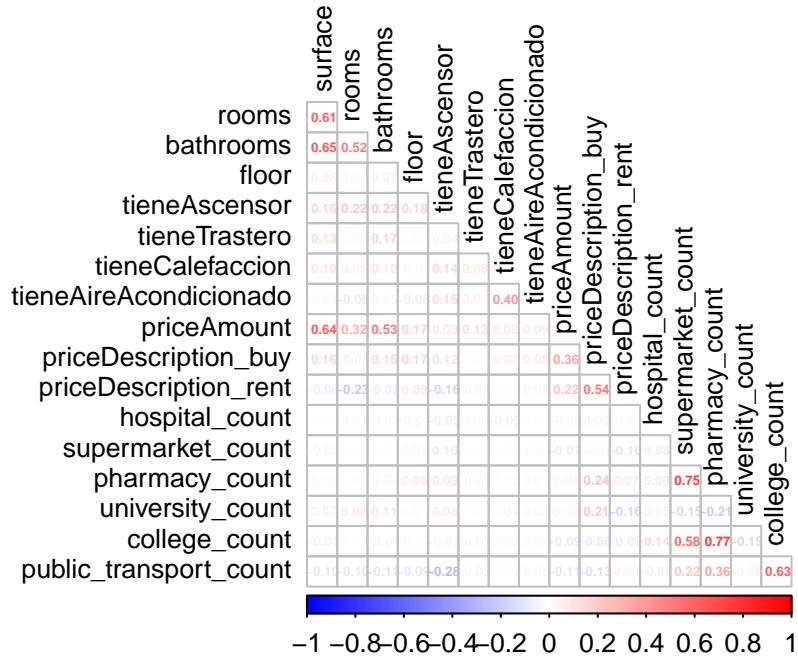
9.5 Análisis Discriminante de Fisher.

En este análisis, intentaremos aplicar la función lineal de Fisher para predecir, en función de sus características, qué inmuebles bajarán de precio. Lo haremos creando binaryDrop a partir de la variable numérica priceAmountDrop. Los valores de esta nueva variable serán 0 si priceAmountDrop=0 y 1 en el caso contrario (el precio ha bajado recientemente). Consideraremos positivo binaryDrop=1. Puesto que las clases se encuentran severamente desbalanceadas (únicamente un 14% de la muestra han bajado de precio), aplicaremos un undersampling para conseguir una proporción 2:1, quedándonos con todos los positivos y parte de los negativos.

9.5.1 Correlación entre variables.

En el análisis discriminante de Fisher, es fundamental que se cumplan ciertos supuestos para garantizar la validez de los resultados. Entre ellos destacan: la normalidad multivariante de las variables predictoras dentro de cada grupo, la homogeneidad de las matrices de covarianzas (igualdad de varianzas y covarianzas entre grupos) y la independencia de las observaciones. Un supuesto especialmente importante es la ausencia de colinealidad entre las variables independientes, ya que la colinealidad puede distorsionar los coeficientes discriminantes y afectar la estabilidad del modelo. Cuando las variables están altamente correlacionadas entre sí, se pierde información útil y se dificulta la identificación de las verdaderas diferencias entre los grupos, reduciendo así la capacidad discriminativa del análisis.

Por ello para empezar identificaremos qué variables están correlacionadas entre sí, y tras crear nuestro modelo comprobaremos si cumple los supuestos para dar por válido el modelo o por el contrario no podemos aplicar el FDA a nuestros datos.



Como vemos hay variables que presentan correlación por lo que debemos excluirlas de nuestro análisis.

```
## Variables con colinealidad (corr > 0.7 ):
```

```

## [1] "pharmacy_count"

##
## Variables finales sin colinealidad:

## [1] "binaryDrop"           "surface"          "rooms"
## [4] "bathrooms"            "floor"             "tieneAscensor"
## [7] "tieneTrastero"        "tieneCalefaccion" "tieneAireAcondicionado"
## [10] "priceAmount"          "priceDescription_buy" "priceDescription_rent"
## [13] "hospital_count"       "supermarket_count"   "university_count"
## [16] "college_count"        "public_transport_count"

```

Var1	Freq
0	516
1	258

9.5.2 Entrenamiento y test.

Una vez seleccionada la base de datos sobre la que aplicaremos el análisis, dividimos las observaciones en un 80% de datos de entrenamiento y 20% de datos de test.

```

##      Resample1
## [1,]      2
## [2,]      3
## [3,]      4
## [4,]      5
## [5,]      6
## [6,]      7

```

	num	perc
0	413	66.61
1	207	33.39

Para entrenar el modelo, realizaremos la validación cruzada mediante el método k-folds repetido. Las observaciones se dividirán en k=10 grupos y todo el proceso se repetirá r=30 veces.

```

## Linear Discriminant Analysis
##
## 620 samples
## 16 predictor
## 2 classes: '0', '1'
##
## Pre-processing: (None)
## Resampling: Cross-Validated (10 fold, repeated 30 times)
## Summary of sample sizes: 558, 558, 557, 558, 559, 559, ...
## Resampling results:
##
##    Accuracy   Kappa
##    0.6723505 0.1481655

```

El índice de Kappa del modelo planteado es relativamente bajo; señal de un modelo pobre con poca capacidad predictiva. La variable binaryDrop está muy condicionada por el factor humano y las necesidades y/o intereses individuales del propietario. Puede ser que estas características la vuelvan una variable de naturaleza impredecible.

No obstante, tampoco hemos comprobado si los datos con los que se ha realizado el modelo cumplen con los requisitos del FDA, por lo que también podría ser esta la causa de estos resultados.

A continuación comprobaremos los supuestos de normalidad multivariante y homocedasticidad.

```
## Grupo 0 (BajadaPrecio = 0) :  
## • p-valor Henze-Zirkler: <0.001 → Se rechaza normalidad multivariante  
## • Variables no normales univariantes (p < 0.05): 16 de 16  
  
## Grupo 1 (BajadaPrecio = 1) :  
## • p-valor Henze-Zirkler: <0.001 → Se rechaza normalidad multivariante  
## • Variables no normales univariantes (p < 0.05): 16 de 16
```

9.5.3 Normalidad y homocedasticidad.

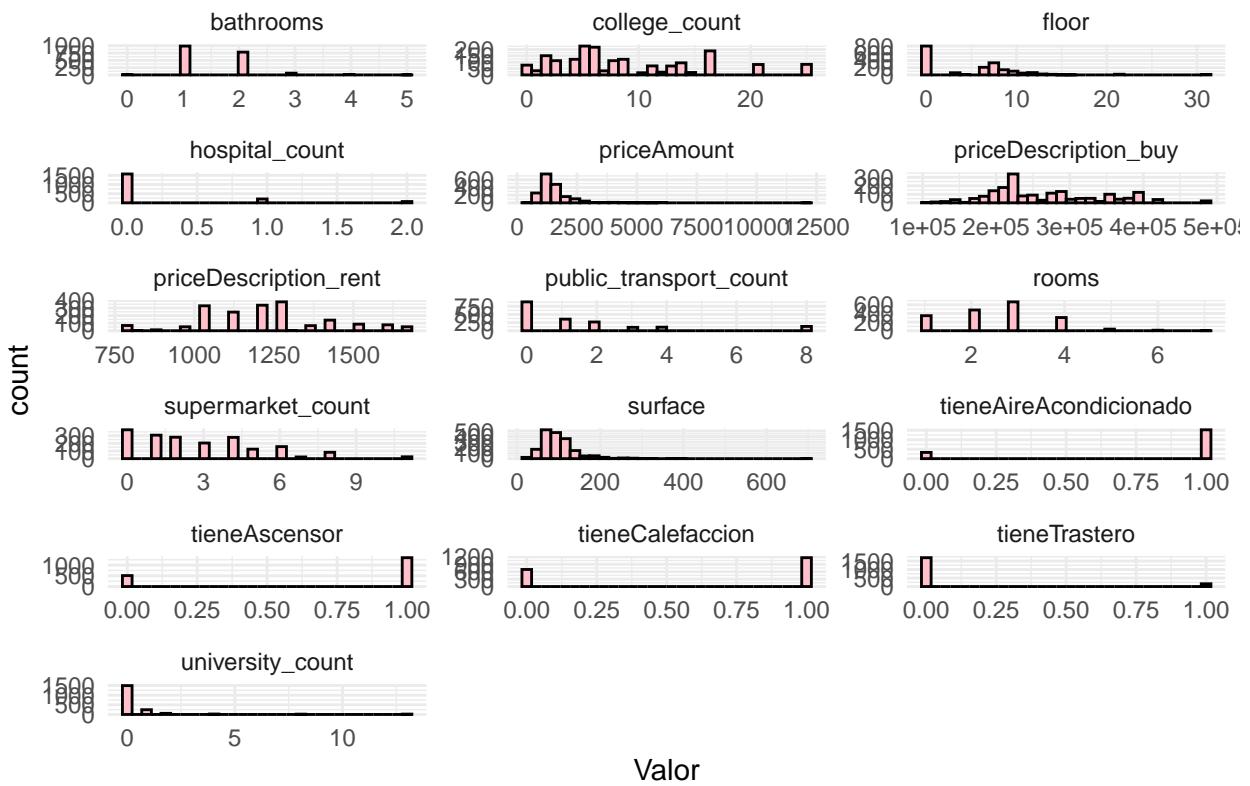
Se evaluó el supuesto de normalidad multivariante mediante la prueba de Henze-Zirkler en ambos grupos de la variable binaryDrop (0 = sin bajada de precio, 1 = con bajada).

En ambos casos, el p-valor fue < 0.001 , por lo que se rechaza la hipótesis nula de normalidad multivariante.

Además, las pruebas univariantes (Anderson-Darling) para cada variable predictora también resultaron significativas ($p < 0.001$), lo que refuerza la evidencia de que los datos no siguen una distribución normal en ninguna dimensión.

En consecuencia, no se cumplen los supuestos necesarios para aplicar el análisis discriminante de Fisher (FDA) de forma válida.

Distribución de variables



Aquí se puede apreciar claramente como las variables univariantes no siguen una distribución normal como nos confirmaban las pruebas univariantes (Anderson-Darling) para cada variable predictora.

```
##  
## Box's M-test for Homogeneity of Covariance Matrices  
##  
## data: fotocasa_filtrado[, -1]  
## Chi-Sq (approx.) = 476.58, df = 136, p-value < 2.2e-16
```

Tras realizar un contraste de hipótesis para comprobar si podíamos asumir o no homocedasticidad en nuestros datos vemos que el p-value obtenido es muchísimo menor que 0.05, y por tanto se rechaza la hipótesis nula de igualdad de matrices de covarianza.

No se cumple la homocedasticidad, ya que las matrices de covarianzas no son iguales entre los grupos.

Con todo esto vemos que nuestros datos son incompatibles con el FDA y por tanto no podemos validar nuestro modelo. Lo descartamos y trataremos de obtener resultados aplicando otros métodos como el PLS-DA.

9.6 PLS.