



UNIVERSIDAD CARLOS III DE MADRID

ANÁLISIS MULTIVARIANTE, GRADO EN ESTADÍSTICA Y EMPRESA

Práctica II: ACP y MDS aplicada a datos de Airbnb en Madrid

Jorge Salas y Marc Pastor

Índice

1 Análisis de Componentes Principales (PCA)	3
2 Análisis de Coordenadas Principales (MDS)	9
2.1 Recategorización de las variables categóricas	9
2.2 Lectura de datos	9
3 Matriz de similaridades de Gower	10
3.1 Variables correladas con la primera coordenada principal	13
3.2 Variables correladas con la segunda coordenada principal	16

1 Análisis de Componentes Principales (PCA)

En este apartado realizamos PCA a nuestro conjunto de datos, con el objetivo de reducir la dimensión del mismo. Los **supuestos** necesarios para el PCA son: tener datos numéricos centrados, es decir, con vector de medias cero. En nuestro caso, utilizaremos el PCA basado en la matriz de correlaciones **R**, ya que tenemos variables medidas en unidades muy distintas.

A continuación simplificaremos la notación:

Variable	Descripción
X_1	host_acceptance_rate
X_2	price
X_3	review_scores_rating
X_4	review_scores_accuracy
X_5	review_scores_cleanliness
X_6	review_scores_checkin
X_7	review_scores_communication
X_8	review_scores_location
X_9	review_scores_value

En nuestro caso, utilizaremos el PCA basado en la matriz de correlaciones **R**, ya que tenemos variables medidas en unidades muy distintas.

Consideremos las transformaciones de Box-Cox de las que hablábamos en el [apartado 5.3 de la primera entrega](#). Utilizando las variables *lambda_1* y *lambda_2* que obtuvimos en la entrega anterior:

```
% Obtenemos lambda_1 y lambda_2 de las transformaciones de Box-Cox de las
% variables de X_analisis d
%[X_continuas_transformada, lambda_2, lambda_1, p_valores] =
%    transformar_boxcox(X_continuas, nombres_variables_numericas, false) %
%    para que no se impriman los histogramas
lambda_2_analisis = lambda_2(setdiff(1:size(X_continuas_transformada, 2),
%    [2,3]))'
lambda_1_analisis = lambda_1(setdiff(1:size(X_continuas_transformada, 2),
%    [2,3]))'

lambda_2_analisis =
1.0e-03 *

1.0000
0
0
0
0
0
0
0
0
0
```

```

lambda_1_analisis =
2.2036
-0.1366
8.1599
11.4881
9.6293
14.9914
14.7061
15.9814
7.6131

```

Tras realizar las transformaciones de Box-Cox, obtenemos las siguientes variables transformadas.

$$\begin{aligned}
X_1(\boldsymbol{\lambda}) &= \frac{(X_1 + \lambda_2)^{\lambda_1 - 1}}{\lambda_1} = \frac{(X_1 + 1 \cdot 10^{-3})^{2.2036 - 1}}{2.2036} \\
X_2(\boldsymbol{\lambda}) &= \frac{(X_2)^{-0.1366 - 1}}{-0.1366} \\
X_3(\boldsymbol{\lambda}) &= \frac{(X_3)^{8.1599 - 1}}{8.1599} \\
X_4(\boldsymbol{\lambda}) &= \frac{(X_4)^{11.4881 - 1}}{11.4881} \\
X_5(\boldsymbol{\lambda}) &= \frac{(X_5)^{9.6293 - 1}}{9.6293} \\
X_6(\boldsymbol{\lambda}) &= \frac{(X_6)^{14.9914 - 1}}{14.9914} \\
X_7(\boldsymbol{\lambda}) &= \frac{(X_7)^{14.7061 - 1}}{14.7061} \\
X_8(\boldsymbol{\lambda}) &= \frac{(X_8)^{15.9814 - 1}}{15.9814} \\
X_9(\boldsymbol{\lambda}) &= \frac{(X_9)^{7.6131 - 1}}{7.6131}
\end{aligned}$$

En este caso realizaremos el Análisis de Componentes Principales mediante la matriz de correlaciones \mathbf{R} , lo cuál es equivalente a trabajar con las variables estandarizadas:

$$\tilde{X}_j(\boldsymbol{\lambda}) = \frac{X_j(\boldsymbol{\lambda})}{\sqrt{\widehat{\text{Var}}[X_j]}} \quad \forall j = 1, \dots, p$$

Primero comprobamos que la matriz $\mathbf{X}(\boldsymbol{\lambda}) = (X_1(\boldsymbol{\lambda}), \dots, X_9(\boldsymbol{\lambda}))$ no está centrada (lo cuál es un supuesto del PCA) ya que su vector de medias no es $\mathbf{0}$. Entonces utilizando la matriz de centrado \mathbf{H} , definimos $\mathbf{X}(\boldsymbol{\lambda})_0 = \mathbf{H}\mathbf{X}(\boldsymbol{\lambda})$. Una vez centrados los datos, realizamos el PCA normalmente.

```

%% X_analisis
mean(X_analisis) % vemos que el vector de medias no es cero, por lo que la
                   → matriz no está centrada

```

```

p = size(X_analisis, 2);
n = size(X_analisis, 1);
H = eye(n) - 1/n * ones(n); % matriz de centrado
X_analisis_0 = H*X_analisis;
mean(X_analisis_0) % la media es cero, por lo tanto la matriz está
↪ centrada

ans =
1.0e+09 *
-0.0000    0.0000    0.0000    0.0060    0.0004    1.3287    0.8748    6.1408    0.0000

ans =
1.0e-04 *
-0.0000   -0.0000    0.0000    0.0001    0.0000   -0.0058    0.0122   -0.1595   -0.0000

```

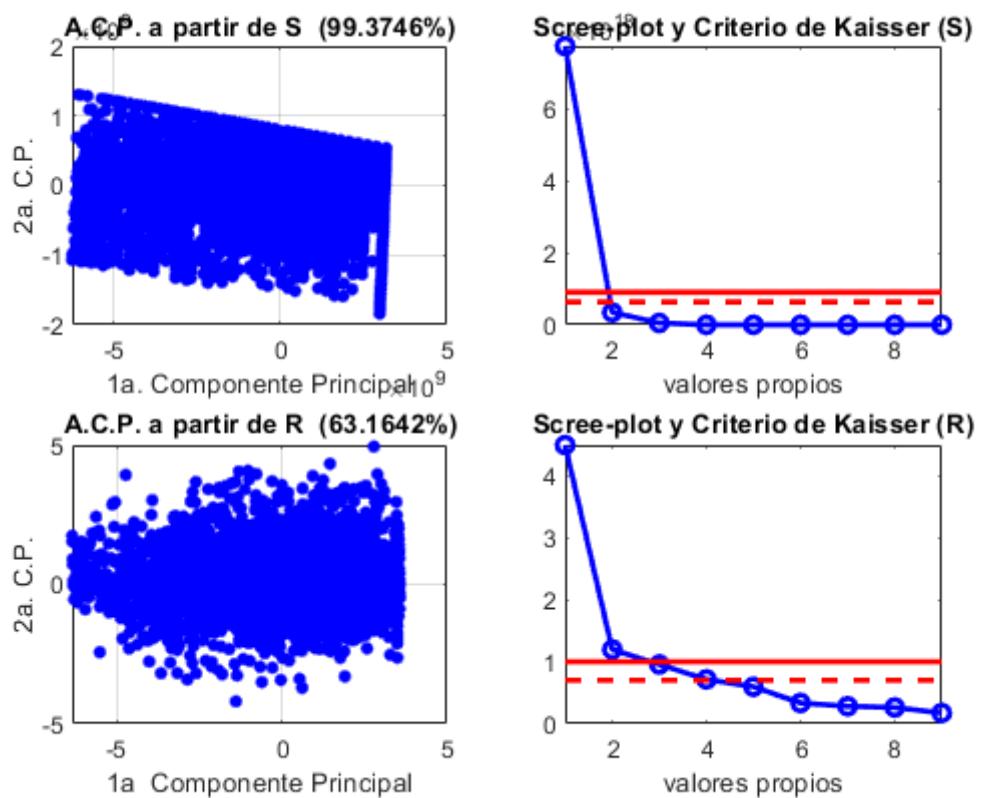
Realizamos el PCA:

```

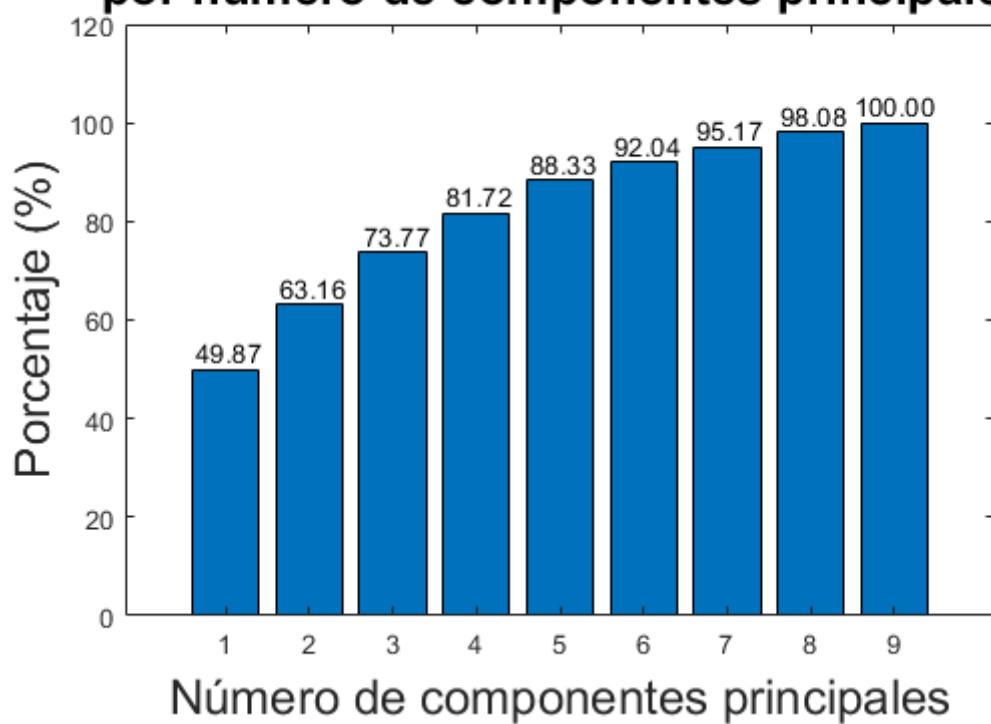
[T1,D1,Y1,acum1,T2,D2,Y2,acum2]=comp2(X_analisis_0); % supuestamente
↪ deberíamos escoger
% utilizaremos la matriz de correlaciones

% Gráfico de porcentaje de variabilidad explicada
index = 1:p;
figure
bar(index, acum2);
xlabel('Número de componentes principales', 'Fontsize', 18);
ylabel('Porcentaje (%)', 'Fontsize', 18);
title('Porcentaje de variabilidad explicada \newline por número de
↪ componentes principales', 'fontsize', 18);
text(index, acum2,num2str(acum2','%0.2f'),...
'HorizontalAlignment','center',...
'VerticalAlignment','bottom');

```



Porcentaje de variabilidad explicada por número de componentes principales



Recordemos que el criterio de Kaisser basado en la modificación de Jolliffe nos dice que excluyamos aquellas componentes cuyos autovalores son menores que $0.7\bar{\lambda} = 0.7tr(S)/p$ (si calculamos las componentes en base a la matriz de covarianzas \mathbf{S}) o que 0.7 si lo hacemos a partir de \mathbf{R} (nuestro caso). Habría que ver si el cuarto autovalor más grande es superior a 0.7 o no.

A continuación se muestra la tabla de los autovalores de la matriz \mathbf{R} , ordenados de mayor a menor:

D2

D2 =

4.4886
1.1962
0.9548
0.7148
0.5949
0.3342
0.2816
0.2623
0.1726

Podemos ver que $\lambda_4 > 0.7$, por lo que según el criterio anterior, deberíamos seleccionar las 4 primeras componentes.

Las 4 componentes principales son:

$$\begin{aligned} Y_1 &= -0.0264\tilde{X}_1(\boldsymbol{\lambda}) + 0.0131\tilde{X}_2(\boldsymbol{\lambda}) + 0.4287\tilde{X}_3(\boldsymbol{\lambda}) + 0.4127\tilde{X}_4(\boldsymbol{\lambda}) + 0.3823\tilde{X}_5(\boldsymbol{\lambda}) + 0.3741\tilde{X}_6(\boldsymbol{\lambda}) + 0.3848\tilde{X}_7(\boldsymbol{\lambda}) \\ Y_2 &= -0.4905\tilde{X}_1(\boldsymbol{\lambda}) - 0.7375\tilde{X}_2(\boldsymbol{\lambda}) + 0.0426\tilde{X}_3(\boldsymbol{\lambda}) + 0.0132\tilde{X}_4(\boldsymbol{\lambda}) - 0.0231\tilde{X}_5(\boldsymbol{\lambda}) + 0.0843\tilde{X}_6(\boldsymbol{\lambda}) + 0.0635\tilde{X}_7(\boldsymbol{\lambda}) \\ Y_3 &= 0.8426\tilde{X}_1(\boldsymbol{\lambda}) - 0.3791\tilde{X}_2(\boldsymbol{\lambda}) + 0.0635\tilde{X}_3(\boldsymbol{\lambda}) + 0.0723\tilde{X}_4(\boldsymbol{\lambda}) + 0.1202\tilde{X}_5(\boldsymbol{\lambda}) - 0.0914\tilde{X}_6(\boldsymbol{\lambda}) - 0.0626\tilde{X}_7(\boldsymbol{\lambda}) \\ Y_4 &= -0.1461\tilde{X}_1(\boldsymbol{\lambda}) + 0.5078\tilde{X}_2(\boldsymbol{\lambda}) + 0.1456\tilde{X}_3(\boldsymbol{\lambda}) + 0.1510\tilde{X}_4(\boldsymbol{\lambda}) + 0.3031\tilde{X}_5(\boldsymbol{\lambda}) - 0.1857\tilde{X}_6(\boldsymbol{\lambda}) - 0.1474\tilde{X}_7(\boldsymbol{\lambda}) \end{aligned}$$

Resumido en tabla:

Variable	Y_1	Y_2	Y_3	Y_4
$\tilde{X}_1(\boldsymbol{\lambda})$	-0.0264	-0.4905	0.8426	-0.1461
$\tilde{X}_2(\boldsymbol{\lambda})$	0.0131	-0.7375	-0.3791	0.5078
$\tilde{X}_3(\boldsymbol{\lambda})$	0.4287	0.0426	0.0635	0.1456
$\tilde{X}_4(\boldsymbol{\lambda})$	0.4127	0.0132	0.0723	0.1510
$\tilde{X}_5(\boldsymbol{\lambda})$	0.3823	-0.0231	0.1202	0.3031
$\tilde{X}_6(\boldsymbol{\lambda})$	0.3741	0.0843	-0.0914	-0.1857
$\tilde{X}_7(\boldsymbol{\lambda})$	0.3848	0.0635	-0.0626	-0.1474
$\tilde{X}_8(\boldsymbol{\lambda})$	0.2222	-0.4451	-0.3072	-0.7217
$\tilde{X}_9(\boldsymbol{\lambda})$	0.4019	0.0617	0.1263	0.0881

La interpretación de estas componentes resulta un tanto difícil. Observando la primera componente principal, podemos ver que las variables referidas a reviews del usuario, tienen un peso similar (alrededor de 0.4 en promedio, excepto $\tilde{X}_8(\boldsymbol{\lambda}) = \frac{(X_8)^{15.9814-1}}{15.9814}$, la transformación de la review de la ubicación), que contribuyen positivamente al eje. Es

decir, podemos interpretar este eje como un indicador de la calidad del piso: a mayor valor de las variables de review, mayor será la calidad del piso.

El segundo eje está dominado por los pesos negativos del ratio de aceptación del anfitrión (transformado) $\tilde{X}_1(\boldsymbol{\lambda}) = \frac{(X_1 + 1 \cdot 10^{-3})^{2.2036-1}}{2.2036}$, el precio transformado $\tilde{X}_2(\boldsymbol{\lambda}) = \frac{(X_2)^{-0.1366-1}}{-0.1366}$ y la review de la ubicación del apartamento (transformada) $X_8(\boldsymbol{\lambda}) = \frac{(X_8)^{15.9814-1}}{15.9814}$. Podríamos interpretar este eje como una medida de la accesibilidad del piso. Para valores altos de precio, ratio de aceptación y de puntuación de la ubicación, los pesos contribuirán de forma muy negativa al eje, dando lugar a valores pequeños, es decir, apartamentos poco accesibles. Es decir, si un piso está bien ubicado, es muy caro y el casero es muy selectivo a la hora de seleccionar a sus huéspedes, el piso será poco accesible. En cambio, si un apartamento está mal ubicado, tiene bajos precios y su dueño es menos selecto, el piso será más fácilmente accesible.

El tercer y cuarto eje no tienen una interpretación clara, pero los conservaremos porque como se puede ver a continuación, la variabilidad explicada resulta bastante alta, y se cumple el criterio de Kaisser (con modificación de Jollife).

Si incluyeramos solo 2 componentes, la variabilidad explicada sería del 63% pero podríamos interpretar ambos ejes. Si incluyeramos 3 componentes se explicaría un 73.77% (un 10% más, aunque sigue siendo un porcentaje medio-bajo de variabilidad explicada), pero no podríamos interpretar Y_3 . Finalmente, con 4 componentes se explica un 81.72% de la variabilidad (un 8% más), siendo un porcentaje de variabilidad bastante alto aunque perdemos la interpretabilidad de Y_3 e Y_4 . A partir del quinto la pendiente disminuye significativamente.

A continuación vemos las correlaciones entre las componentes principales y las variables originales:

Variable	Y_1	Y_2	Y_3	Y_4
\tilde{X}_1	-0.0560	-0.5364	0.8234	-0.1235
\tilde{X}_2	0.0278	-0.8066	-0.3704	0.4294
\tilde{X}_3	0.9082	0.0466	0.0620	0.1231
\tilde{X}_4	0.8743	0.0144	0.0706	0.1277
\tilde{X}_5	0.8101	-0.0253	0.1174	0.2563
\tilde{X}_6	0.7926	0.0922	-0.0893	-0.1570
\tilde{X}_7	0.8152	0.0694	-0.0611	-0.1246
\tilde{X}_8	0.4707	-0.4868	-0.3002	-0.6102
\tilde{X}_9	0.8515	0.0675	0.1234	0.0745

El primer eje principal Y_1 está altamente correlado con X_3 y X_4 , que son las variables de review con más peso. Es decir, que lo más relevante a la hora de determinar la calidad de un apartamento de Airbnb es su puntuación global (X_3) y el grado de detalle/exactitud del anuncio (X_4), que indica el grado de exactitud entre los servicios y características que se anuncian, y los que obtiene el cliente en realidad.

El segundo eje principal está altamente correlado (en valor absoluto) con el precio por noche (X_2) y la proporción de reservas que acepta el huésped (X_1). Es decir, lo que determina en mayor medida el grado de accesibilidad de un apartamento es su precio, su dueño y su ubicación. Cuanto mejor ubicado, más cara es la estancia y más selecto el huésped (ídem en el caso contrario).

Variable	Tipo	Descripción
host_acceptance_rate	Numérica continua	Porcentaje de ofertas que acepta el propietario.
host_total_listings_count	Numérica discreta	Número total de propiedades distintas ofrecidas.
latitude	Numérica continua	Latitud del apartamento.
longitude	Numérica continua	Longitud del apartamento.
accommodates	Numérica discreta	Capacidad (personas) del apartamento.
bathrooms	Numérica discreta	Número de baños.
bedrooms	Numérica discreta	Número de habitaciones.
beds	Numérica discreta	Número de camas.
price	Numérica continua	Precio por noche (en \$).
minimum_nights	Numérica discreta	Número mínimo de noches.
maximum_nights	Numérica discreta	Número máximo de noches.
availability_30	Numérica discreta	Número de días disponibles en los 30 días siguientes.
availability_90	Numérica discreta	Número de días disponibles en los 90 días siguientes.
number_of_reviews	Numérica discreta	Número de reseñas.
number_of_reviews_ltm	Numérica discreta	Número de reseñas en el último mes.
review_scores_rating	Numérica continua	Puntuación media general del apartamento.
review_scores_accuracy	Numérica continua	Puntuación media de la exactitud y detalle de la descripción.
review_scores_cleanliness	Numérica continua	Puntuación media de la limpieza del apartamento.
review_scores_checkin	Numérica continua	Puntuación media del checkin del apartamento.
review_scores_communication	Numérica continua	Puntuación media de la comunicación con el anfitrión.
review_scores_location	Numérica continua	Puntuación media de la ubicación del apartamento.
review_scores_value	Numérica continua	Puntuación media de la calidad/precio del apartamento.
reviews_per_month	Numérica discreta	Número de reseñas por mes.
neighbourhood_cleansed	Categórica multiestado	Indica el distrito en el que se encuentra el apartamento.
neighbourhood_group_cleansed	Categórica multiestado	Indica el barrio en el que se encuentra el apartamento.
property_type	Categórica multiestado	Indica si se trata de un apartamento, habitación compartida, etc.
room_type	Categórica multiestado	Indica el tipo de habitación.

2 Análisis de Coordenadas Principales (MDS)

Para llevar a cabo este análisis, hemos usado más variables que en apartado anterior, ya que la principal ventaja del MDS es que nos permite trabajar con datos de tipo mixto. Hemos usado las siguientes variables:

Como se puede ver, no hemos incluido variables binarias, ya que lo hemos probado y provocaban errores de cálculo (Nan's) en la función `correlaciones2.m`, además, la variabilidad explicada de las 2 primeras coordenadas principales era mucho menor que con la opción actual.

2.1 Recategorización de las variables categóricas

Hemos tenido que recategorizar las variables multi-estado, para que tengan niveles numéricos, para poder usar las funciones vistas en clase. [Resumen de la codificación](#)

2.2 Lectura de datos

```
datoslimpiospreMDS = readtable('datos_limpios_pre_MDS.csv');
nombres_variables_pre_MDS = datoslimpiospreMDS.Properties.VariableNames;
```

```
X_pre_mds = table2array(datoslimpiospreMDS);

num_variables_numericas = 23;
num_variables_binarias = 0;
num_variables_categoricas = 4;
```

3 Matriz de similaridades de Gower

Hemos construido una matriz de similaridades en base a la distancia de Gower, ya que tenemos variables de tipo mixto. Esto implica que no tiene sentido utilizar distancias tradicionales como la euclídea para variables categóricas, ya que a pesar de la codificación numérica de las mismas, las operaciones aritméticas habituales no tienen sentido.

```
% Matriz de similaridades de Gower
S = gower2(X_pre_mds,num_variables_numericas, num_variables_binarias,
    ↵ num_variables_categoricas);
% Matriz de distancias al cuadrado de Gower
D2 = ones(size(S))-S;
```

A continuación realizamos un análisis de coordenadas principales:

```
[Y,vaps,percent,acum] = coorp(D2);
```

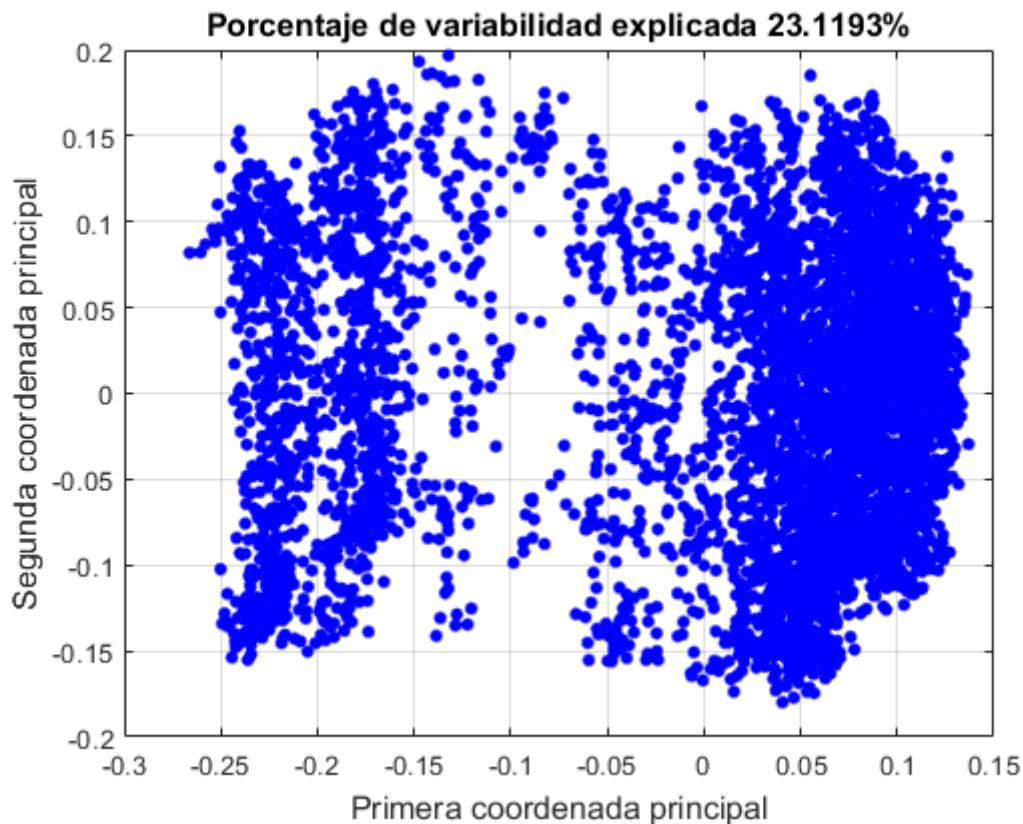


Figure 1: png

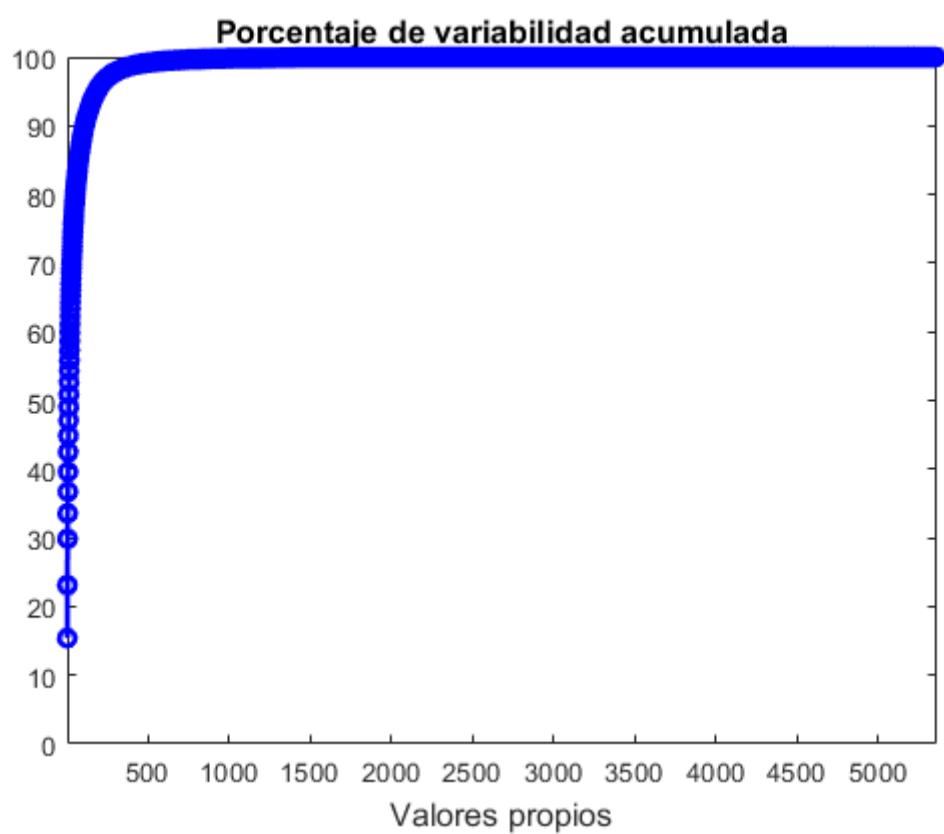


Figure 2: png

Podemos comprobar que las dos primeras componentes principales están incorreladas, y que con 13 coordenadas principales (aproximadamente igual a la mitad de variables, 27), se explica un 54% de la variabilidad, lo cuál no es demasiado.

Calculamos las correlaciones (de variables numéricas) y la V de Cramer (para variables categóricas), respecto de las coordenadas principales:

```
p_cuant = num_variables_numericas;
p_nominal = num_variables_binarias + num_variables_categoricas;
corr_table = correlaciones2(X_pre_mds, Y(:, 1:3), p_cuant, p_nominal)
```

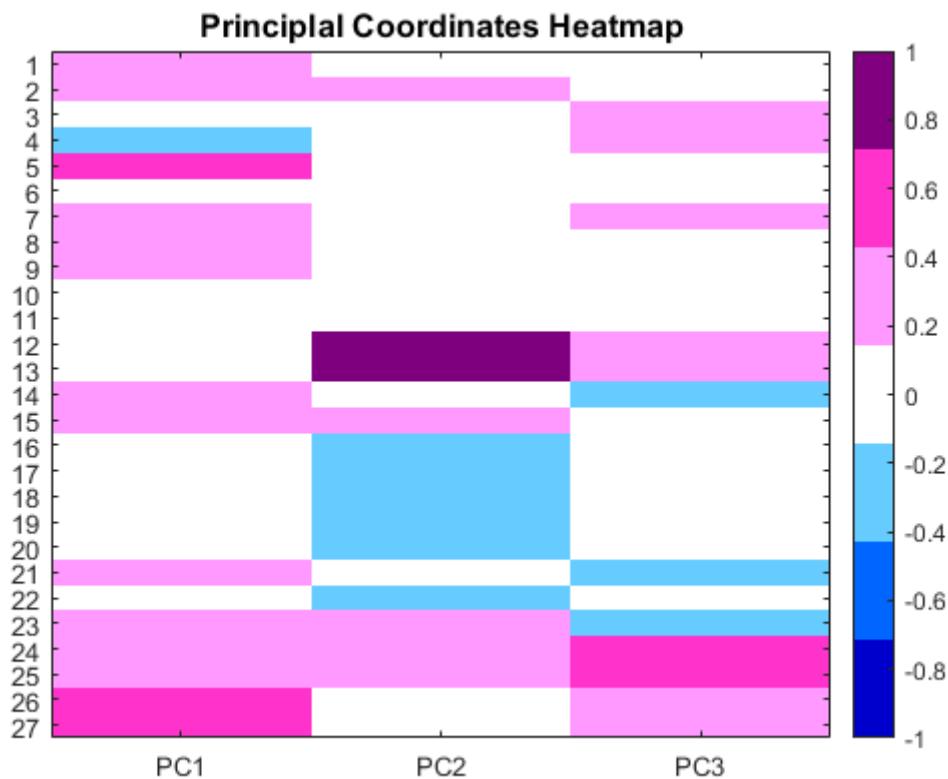


Figure 3: png

Las variables más correladas con la primera coordenada principal son X_5 (accommodates), X_{26} (property_type) y X_{27} (room_type).

Las variables más correladas con la segunda componente son X_{12} (availability_30) y X_{13} (availability_90).

La variable más correlada con la tercera componente serían X_{24} y X_{25} , relacionadas con el tipo de Neighbourhood.

3.1 Variables correladas con la primera coordenada principal

```
identif_cuantis(X_pre_mds(:, 5), Y);
```

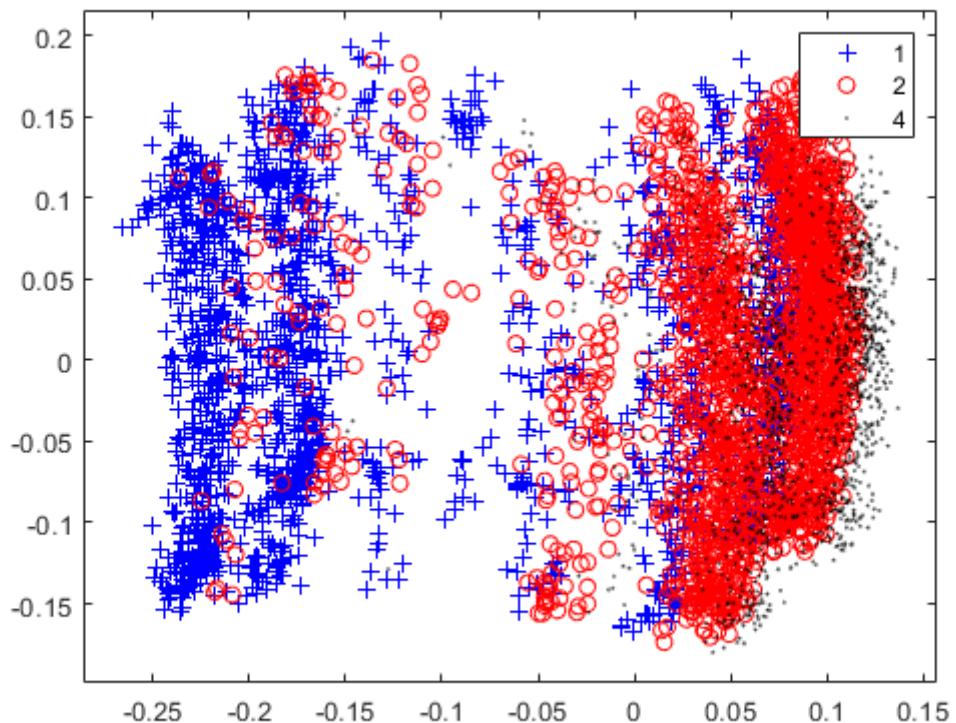


Figure 4: png

La variable accommodates tiene una correlación positiva con la primera coordenada principal, y por lo tanto, a medida, que la variable aumenta su valor, el individuo se coloca más a la derecha en la representación MDS.

```
identif_cualis(X_pre_mds(:, 26), Y);
```

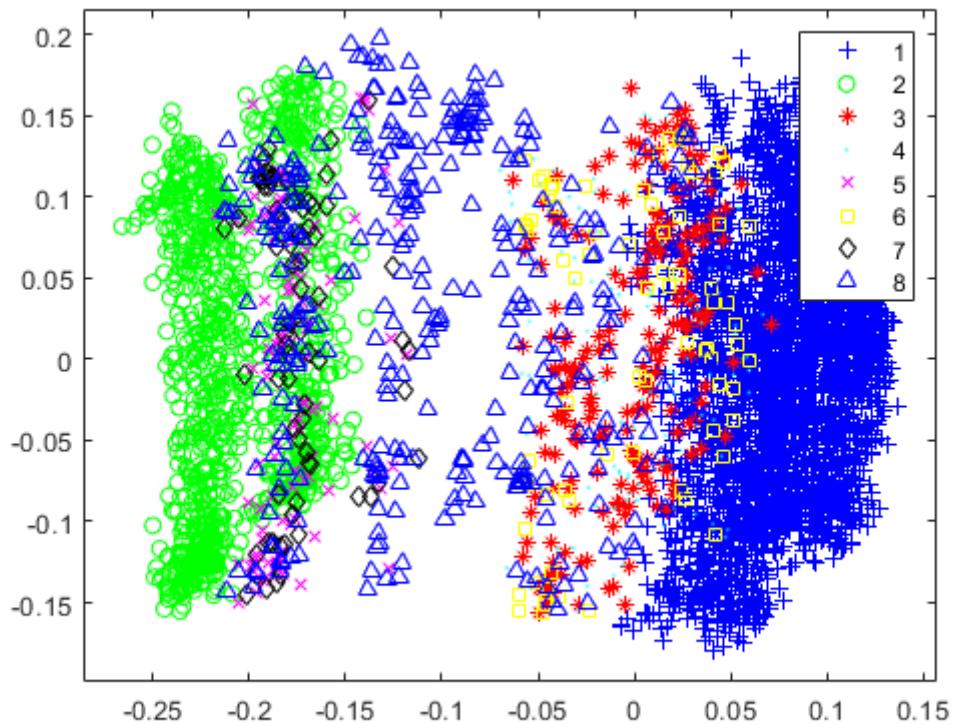


Figure 5: png

Este gráfico nos indica que podemos diferenciar entre dos perfiles de anuncios en Airbnb según la primera coordenada principal: habitaciones privadas (en verde, a la izquierda) y apartamentos (cruces azules, a la derecha), que tendrían valores contrarios en su representación MDS.

```
identif_cualis(X_pre_mds(:, 27), Y);
```

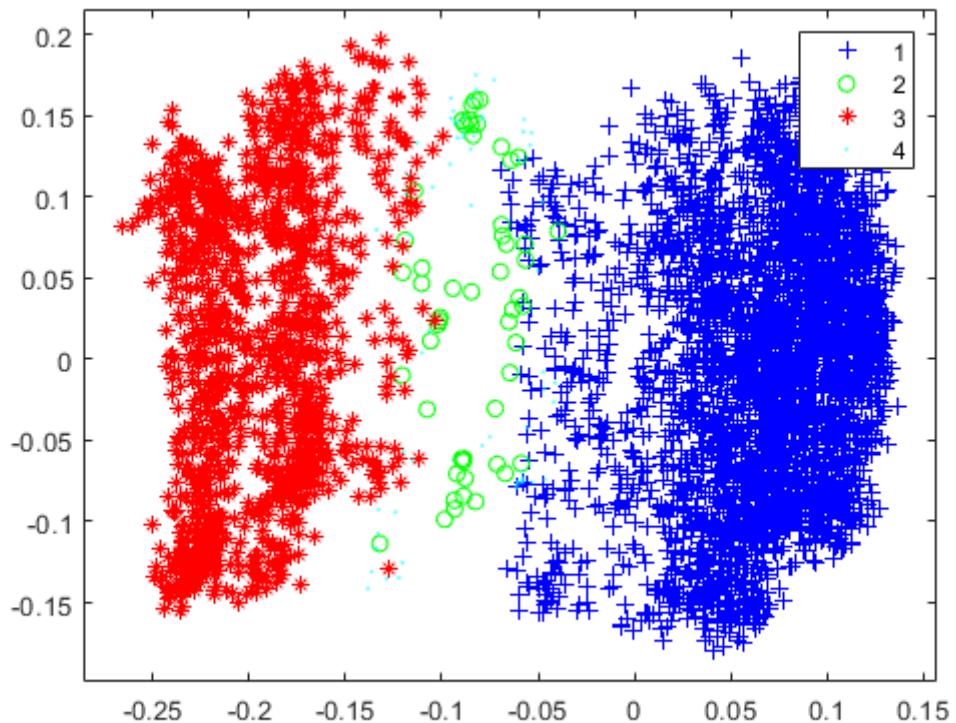


Figure 6: png

Podemos diferenciar claramente dos grupos, según la primera coordenada principal: en azul y con valores a la derecha del primer eje: los apartamentos enteros, y en rojo y a la izquierda del mismo eje: habitaciones de hotel.

3.2 Variables correladas con la segunda coordenada principal

```
identif_cuantis(X_pre_mds(:, 12), Y);  
identif_cuantis(X_pre_mds(:, 13), Y);
```

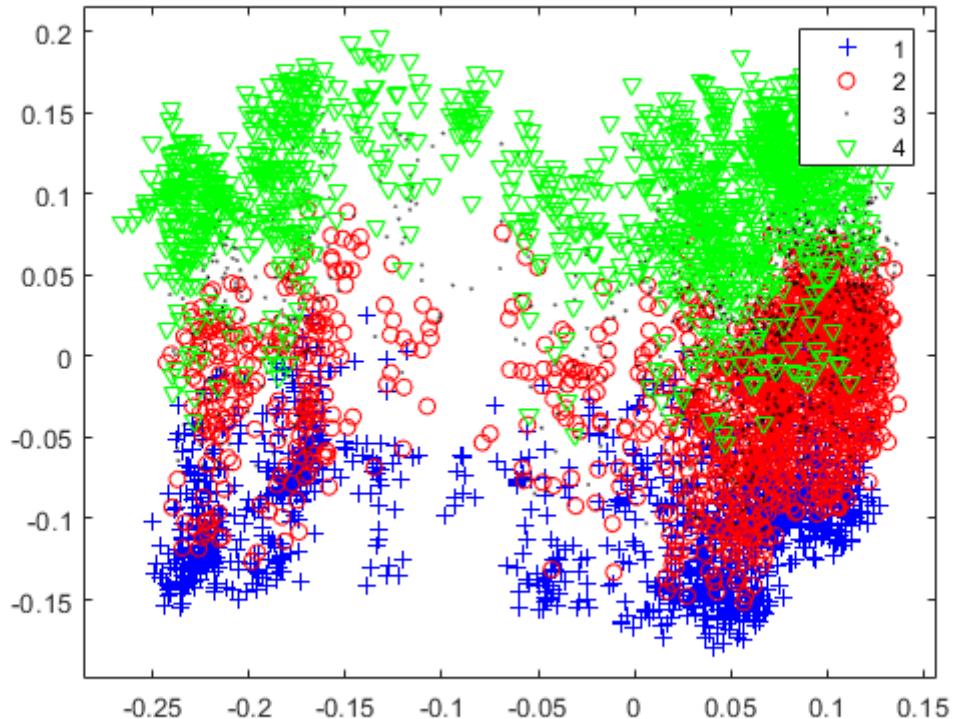


Figure 7: png

La variable crece en la segunda coordenada principal mientras mayor disponibilidad mensual o trimestral tenga el apartamento.

En base a este análisis, podemos concluir que existen 2 grupos principales en función del número de *accommodes*, y el tipo de apartamentos: las habitaciones privadas, están a la izquierda de la primera coordenada principal, al tener normalmente menos capacidad que los apartamentos, que están situados en la parte derecha del eje, y suelen tener mayor capacidad. En cuanto al segundo eje, no se observan grupos claramente diferenciados, ya que hay mucho solapamiento de puntos.