4. Análisis de conglomerados

Para hacer clustering con k-means la primera cuestión que surge es el número de grupos a considerar en los datos. En la tabla 8 se presentan los clusters sugeridos por distintos tipos de índices.

Cuadro 8: Indices para selección de K

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW
Number_clusters	**2.0000**	**2.000**	**3.000**	**2.0000**	**3.0000**	**4.00000e+00**	**3**	**3.0**
Value_Index	3.8287	544.615	96.955	62.6412	736.6544	4.75383e+51	7389212930	226190.9

	Friedman	Rubin	Cindex	DB	Silhouette	Duda	PseudoT2	Beale	Ratkowsky
Number_clusters	**6.0000**	**3.000**	**2.0000**	**2.0000**	**2.0000**	**2.0000**	**2.000**	**2.0000**	**2.000**
Value_Index	19.4276	-2.261	0.2354	1.3099	0.3546	1.5634	-298.024	-2.4194	0.428

El valor sugerido en la mayoría de estos índices es $k = 2$ (entre ellos algunos de los populares como *Silhouette* y *Duda*), pero también hay algunos como *Hartigan* y *TrCovW* que sustentan usar $k = 3$.

Después de explorar los casos de $k = 2$ y $k = 3$ en los datos originales, no se obtuvieron resultados con interpretación significativa (chunk: $pca - result - orig$), ya que sólo se separaban los grupos en individuos que le dan importancia a todas las variables o a ninguna. Dado esto, se optó por hacer alguna transformación [1] útil para obtener una mejor distribución de los grupos. En la figura 13 se presentan los clusters obtenidos con $k = 3$ en los datos transformados, con los cuales se obtuvieron los mejores resultados.

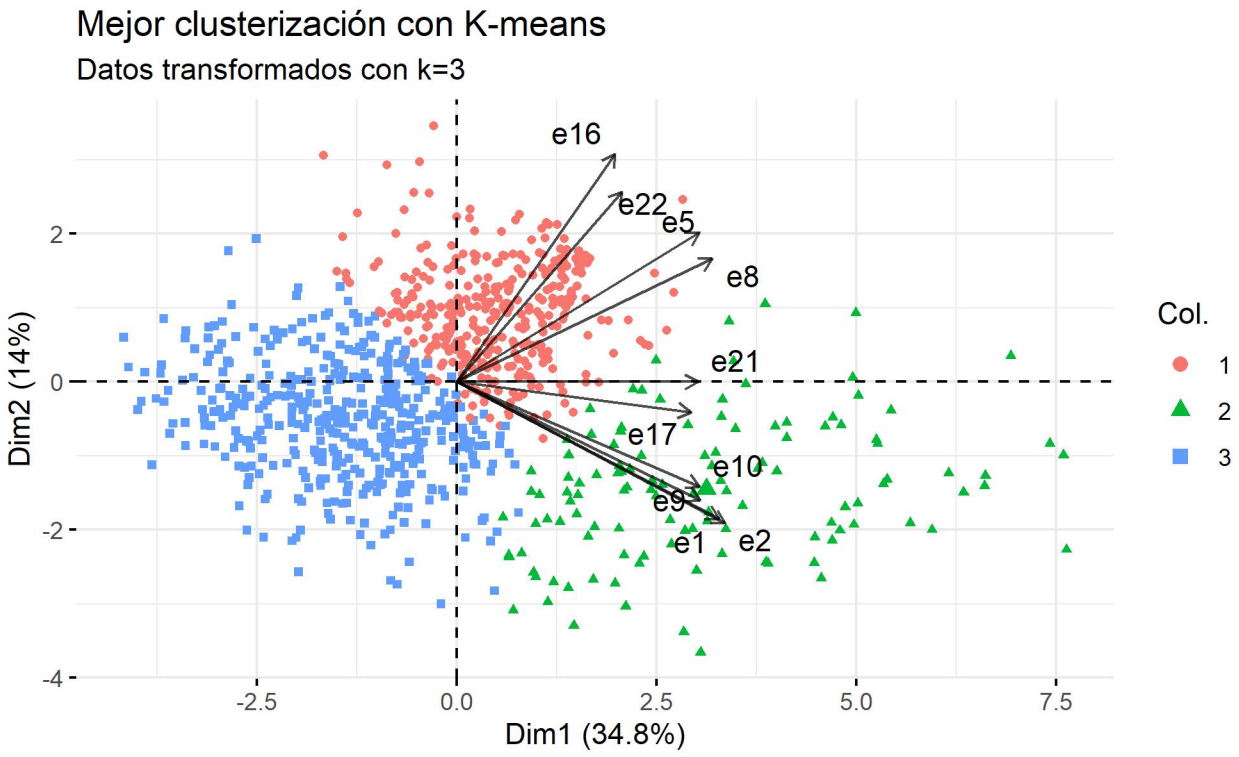

Figura 13: Clusters obtenidos con k-means.

Una alternativa a k-means es aplicar el método de conglomerados jerárquico aglomerativo. Para probar una amplia gamma de opciones, en este se consideraron 3 distancias: euclideana, manhattan y canberra, para medir la disimilaridad entre las observaciones. Para la disimilaridad entre clusters se consideraron las ligas: average, complete, Ward, centroid y median. De esta

[1] Para cada columna X, se realizó $T(X) = |X - \mu_X| \left(\frac{X}{\sigma_X}\right)^{1/3}$

forma, se obtuvieron 15 pares de distancias y disimilaridades a probar en cada dataset (datos originales y datos transformados), por lo tanto, se evaluaron 30 modelos de aglomeración distintos.

Después de revisar los clusters obtenidos por cada método (chunk: *dendogramas*), se observa que en general los resultados convenientes para la interpretación son los obtenidos con el enlace Ward. Esta clusterización se presenta en el espacio de los componentes principales de los datos transformados, lo cual se encuentra en la figura 14.

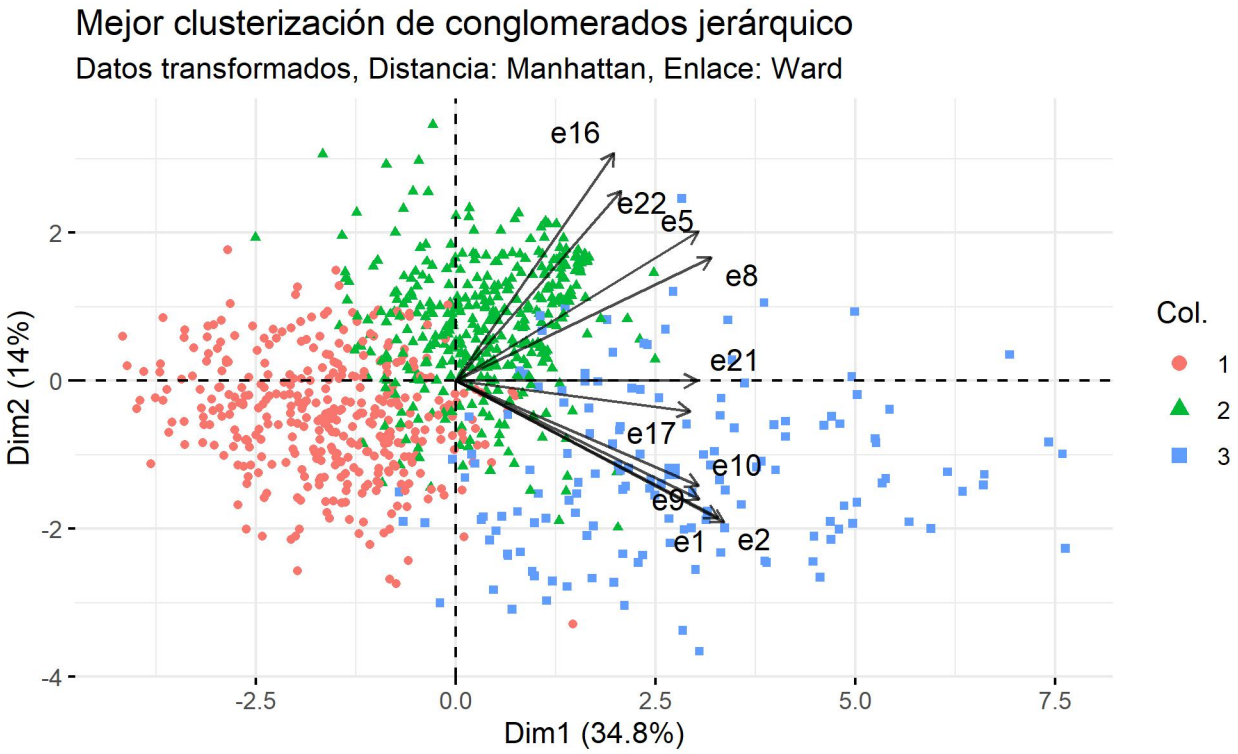

Figura 14: Clusters obtenidos con el método de conglomerados jerárquico.

Por otro lado, también se exploró la posibilidad de aplicar tanto k-means como el método de aglomeración jerárquica con los componentes principales de los datos (y no con los datos per se). Para esto se seleccionaron suficientes componentes para recuperar almenos el 80 % de la varianza en cada conjunto de datos.

En la figura 15, se presentan los clusters resultantes de ambos métodos tomando los componentes principales, en esta se observa que se obtuvieron resultados muy similares a los primeros dos.

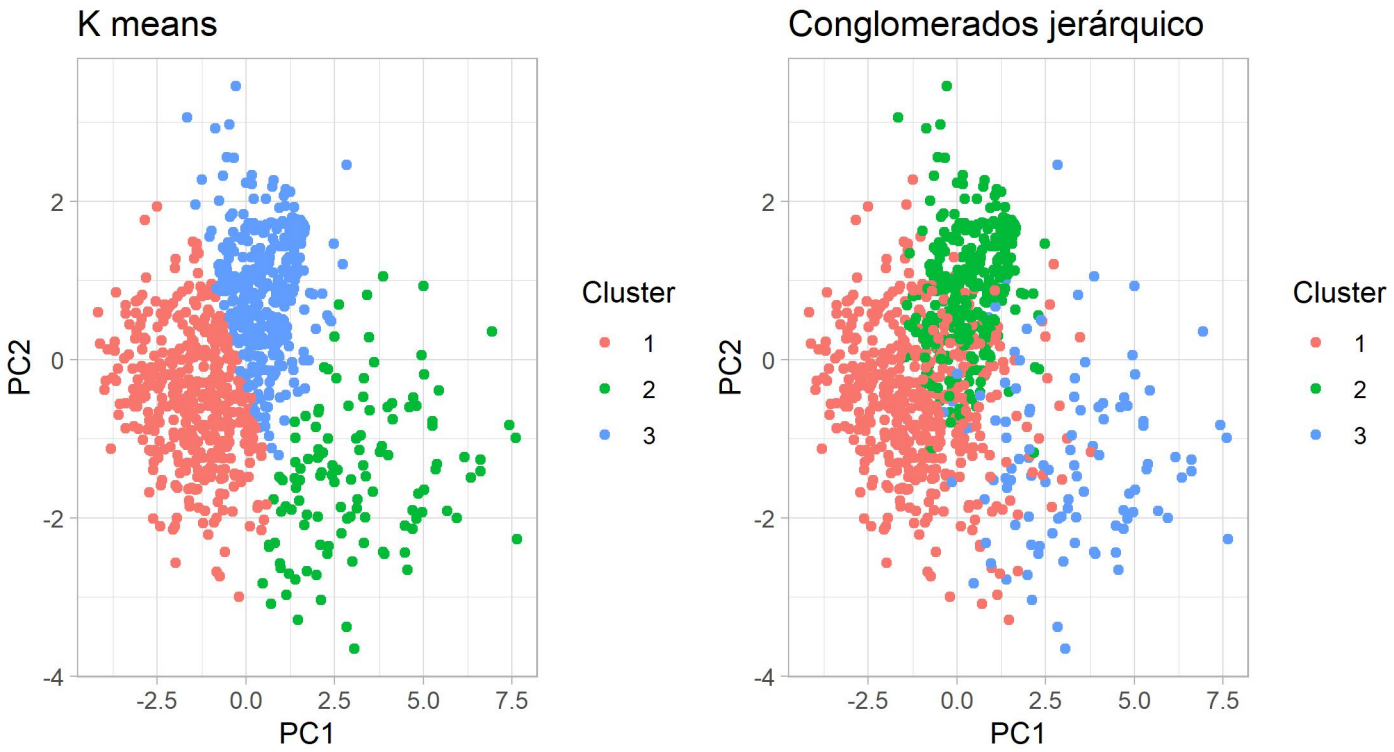

Figura 15: Clusters obtenidos al aplicar kmeans (izquierda) y método jerárquico (derecha) sobre los componentes principales.

Con todo lo anterior, es más claro en la figura 13 analizar los clusters para realizar una descripción de los clientes, ya que en este vemos como el primer cluster (rosa) se orienta principalmente en el sentido de las variables $e5, e8, e16$ y $e22$; las cuales en conjunto describen la variedad de alimentos y servicios de entretenimiento, además de la comodidad y satisfacción con el viaje. Por lo tanto en este primer cluster se encuentran las personas que le dan más importancia a estos aspectos. Por otro lado, los individuos del cluster 3 (azul) se encuentran en el sentido de las variables $e1, e2, e9$ y $e10$; las cuales en general se pueden interpretar como un grupo al que le importa más la seguridad y la llegada puntual en los vuelos (sin retrasos ni imprevistos). Estos dos clusters, tienen en común también una importancia positiva en las variables $e17$ y $e21$, que indican la calidad del servicio (hospitalario y sin complicaciones). Por último, esta el cluster 2 (verde) que se encuentra directamente opuesto a las variables $e17$ y $e21$, por lo que a estos individuos no les parece tan importante estos aspectos y además en los demás se encuentran con una orientación media, por lo que se podrían considerar como un grupo neutro en ese sentido.