

**MAESTRÍA EN INTELIGENCIA ARTIFICIAL.**

**ERICK CHAPARRO QUINTANAR**

**JUAN MANUEL ALEGRÍA BOBADILLA**

**LILIANA MONDRAGÓN QUINTANA**

**INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO (IAA)**

**DR. GADDIEL DESIRENA LÓPEZ**

**PROYECTO FINAL**

**SEPTIEMBRE 2025**

ÍNDICE

[INTRODUCCIÓN 1](#_Toc209551220)

[1. DEFINICIÓN DEL PROBLEMA. 2](#_Toc209551221)

[2. VISUALIZACIÓN, EXPLORACIÓN Y ANÁLISIS DE DATOS 4](#_Toc209551222)

[2.1. Estadísticas Básicas: 4](#_Toc209551223)

[2.2. Insignias y gráficos (mínimo 5). 5](#_Toc209551224)

[3. CLUSTERIZACIÓN DE LOS DATOS 10](#_Toc209551225)

[3.1. Clustering jerárquico (codo y gradiente) 10](#_Toc209551226)

[3.2. Kmeans (codo y gradiente) 15](#_Toc209551227)

[CONCLUSIÓN. 17](#_Toc209551228)

# INTRODUCCIÓN

En el presente proyecto se realiza un análisis exploratorio de datos y una segmentación mediante técnicas de aprendizaje automático no supervisado (clustering), tomando como base el archivo *tripadvisor\_review.csv*. Este conjunto de datos contiene información recopilada de usuarios en el sitio web TripAdvisor.com, centrada en reseñas sobre destinos turísticos en la región del Este de Asia.

El dataset fue generado a partir del rastreo de opiniones de viajeros y agrupa las calificaciones promedio otorgadas por usuarios a distintos tipos de destinos turísticos. Las opiniones se categorizan en una escala de 0 (terrible) a 4 (excelente).

Cada registro del dataset corresponde a un usuario único, identificado mediante un ID, e incluye sus calificaciones promedio en las siguientes 10 categorías de atracciones turísticas: Galerías de arte, Clubes de baile, Bares de jugos, Restaurantes, Museos, Complejos turísticos, Parques o lugares de picnic, Playas, Teatros e Instituciones religiosas.

# DEFINICIÓN DEL PROBLEMA.

Realizar un análisis exploratorio de datos y aplicar técnicas de clustering al dataset *tripadvisor\_review.csv*. Este dataset contiene calificaciones de usuarios a diferentes atracciones turísticas. El propósito es identificar patrones y segmentaciones significativas entre los usuarios o atracciones, lo que podría facilitar estrategias de marketing, recomendaciones personalizadas, o mejoras en la experiencia del cliente.

El archivo tripadvisor\_review.csv contiene las siguientes características:

1. Filas: Cada fila representa un usuario.
2. Columnas: Cada columna representa la puntuación de ese usuario a una atracción específica.
3. Tipo de valores: Numéricos (calificaciones del 0 al 4).
   * Excelente (4)
   * Muy Buena (3)
   * Regular (2)
   * Mala (1)
   * Terrible (0)
4. Variables:

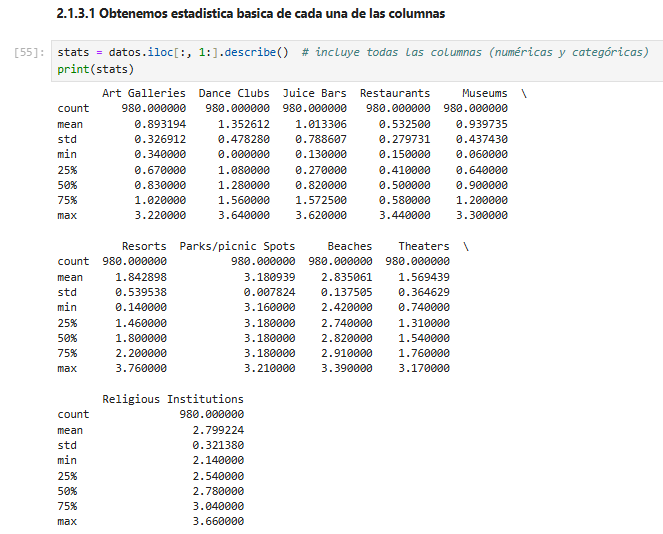
* ID de usuario único.
* Comentarios promedio de los usuarios sobre galerías de arte.
* Comentarios promedio de los usuarios sobre clubes de baile.
* Comentarios promedio de los usuarios sobre bares de jugos.
* Comentarios promedio de los usuarios sobre restaurantes.
* Comentarios promedio de los usuarios sobre museos.
* Comentarios promedio de los usuarios sobre complejos turísticos.
* Comentarios promedio de los usuarios sobre parques/lugares de picnic.
* Comentarios promedio de los usuarios sobre playas.
* Comentarios promedio de los usuarios sobre teatros.
* Comentarios promedio de los usuarios sobre instituciones religiosas.

No se encontraron valores nulos, y todas las variables fueron consideradas para el análisis.

## VISUALIZACIÓN, EXPLORACIÓN Y ANÁLISIS DE DATOS

Extracción de información estadística básica en los datos analizados

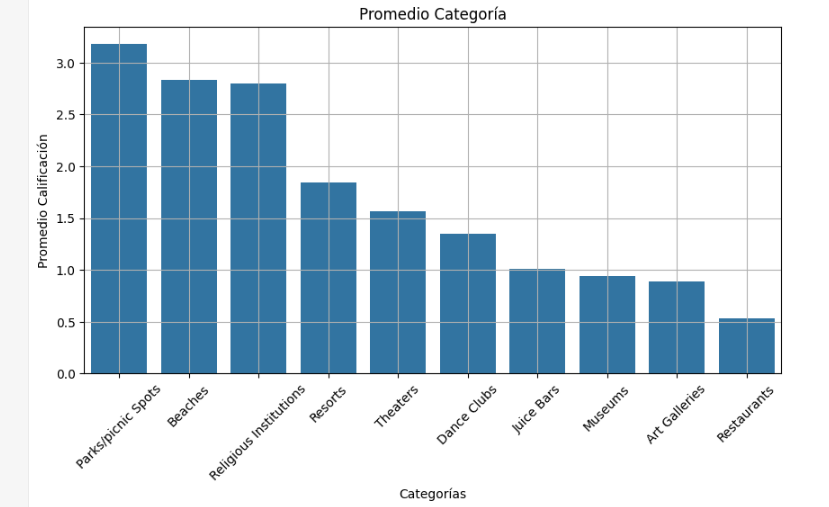
## Estadísticas Básicas:

Se extrajeron estadísticas descriptivas como valores no nulos, la media, desviación estándar, mínimo, percentiles (25%, 50%, 75%) y máximo, las cuales se muestran a continuación con base en las calificaciones otorgadas por los usuarios en las diversas atracciones.

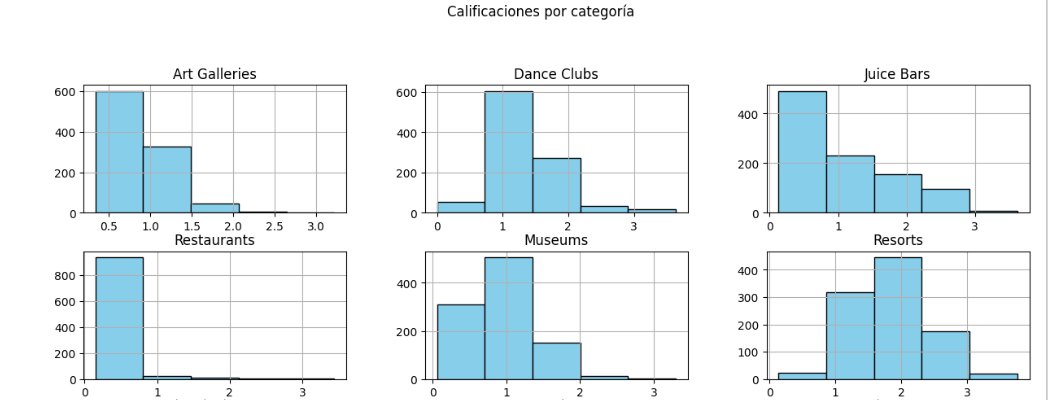
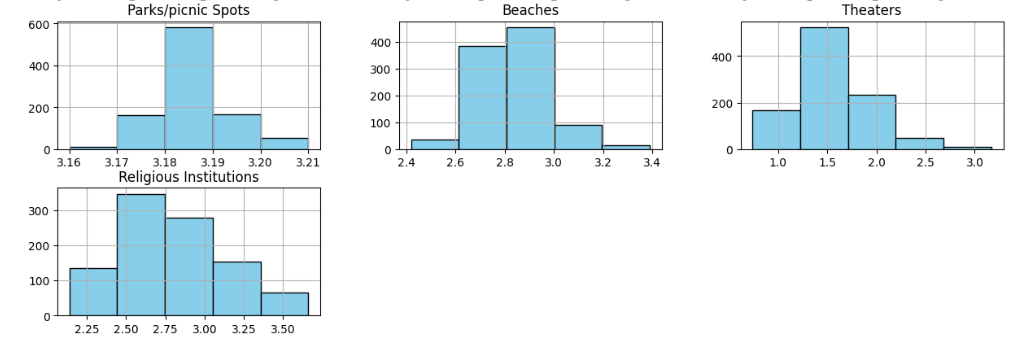
## Insignias y gráficos (mínimo 5).

A continuación, se presentan 5 insignias obtenidos a partir de visualizaciones y análisis:

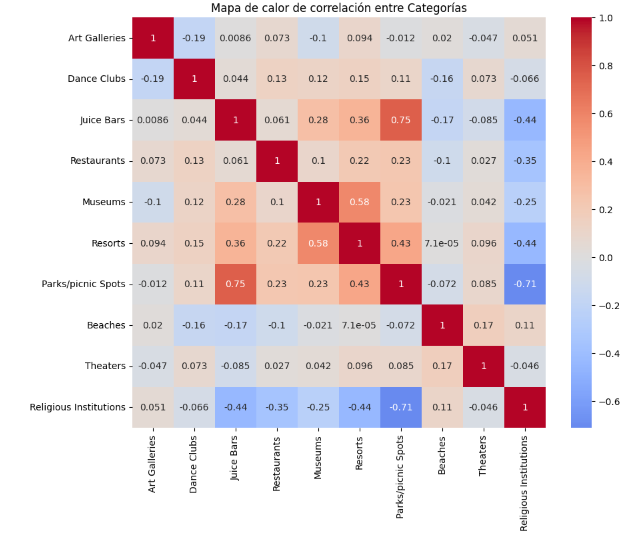
INSIGNIA 1: Promedio de calificación asignada por los usuarios con base a cada una de las atracciones evaluadas; el gráfico que a continuación se presenta refleja los resultados obtenidos de forma independiente por cada una de las atracciones con base en la evaluación de los usuarios, dicha información refleja la atracción mejor evaluada **Park/picnic Spots**, así como la pero evaluada **Restaurants**.



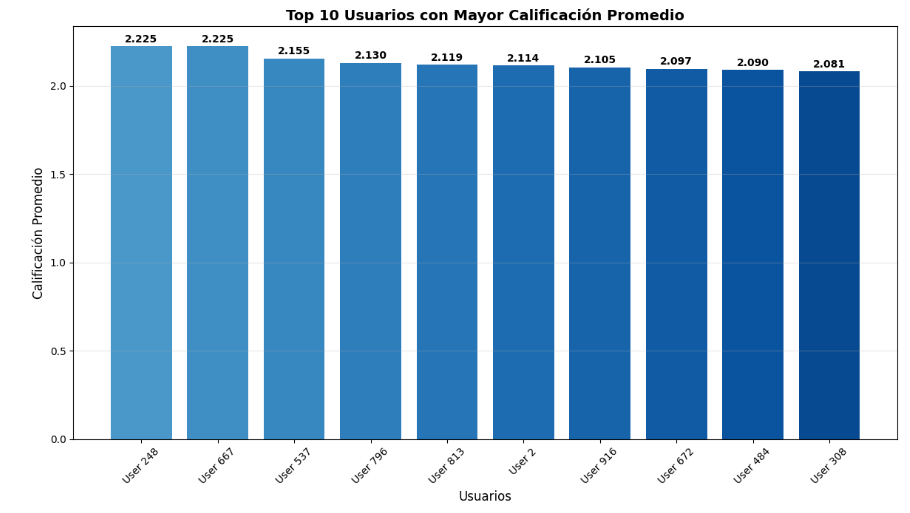
INSIGNIA 2: Calificaciones por categoría de forma distribuida, para visualizar el comportamiento de cada uno de los atractivos y su calificación asignada con base en la votación realizada por parte de los usuarios. Estos gráficos permiten realizar una valoración de las calificaciones obtenidas en cada una de las atracciones, desde la más alta hasta la más baja, así como la cantidad de votantes de las mismas para poder realizar un ajuste en las atracciones disponibles.



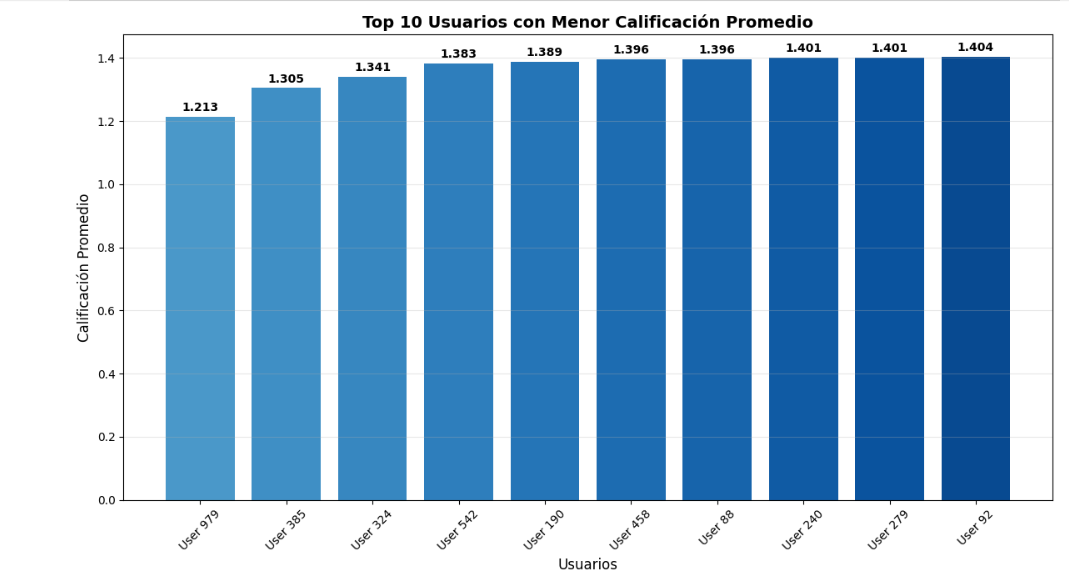
INSIGNIA 3: Identificación de similitudes entre las categorías evaluadas por parte de los usuarios de Reseñas de viajes mediante un mapa de calor, el cual refleja que a los que les gusta **Parks/picnic Spots** de igual forma les gusta **Juice Bars**.



INSIGNIA 4: El gráfico que se muestra enseguida refleja el top 10 de los usuarios con calificaciones más altas en promedio para todas las atracciones, desde la calificación más alta 2.225 por parte de los usuarios 248 y 667, así mismo con una calificación de 2.081 por parte del usuario 308. La información graficada permite realizar una valoración final mediante alguna encuesta extra para saber su opinión de dichos usuarios y poder realizar recomendaciones para aumentar la valoración de otros visitantes.



INSIGNIA 5: Por igual es importante reflejar el resultado más bajo en cuento al top 10 de los usuarios con calificaciones mas bajas en promedio para todas las atracciones, lo cual se muestra en el gráfico siguiente. Siendo la calificación mas baja 1.213 por parte del usuario 979 y la mas alta respecto a niveles bajos de 1.401 por parte de los usuarios 240, 279 y 92. Respecto a este gráfico es importante tomar acciones para elevar el nivel de las calificaciones obtenidas y lograr una mejor estrategia de marketing.



# CLUSTERIZACIÓN DE LOS DATOS

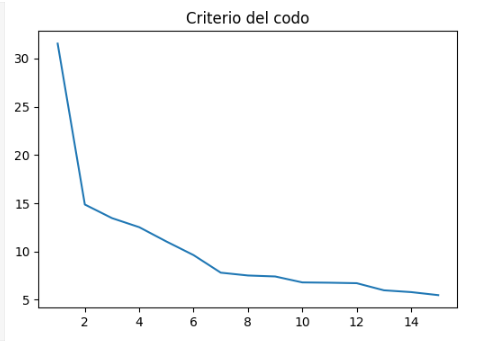
## Clustering jerárquico (codo y gradiente)

Para determinar el número óptimo de grupos al aplicar el método de clustering jerárquico, se utilizó el enfoque del codo mediante el análisis de la distancia de enlace entre los últimos pasos de agrupamiento.

En este caso, se utilizó el método de Ward para construir el dendrograma, y se extrajeron los últimos valores de fusión para visualizar cómo varía la distancia entre los grupos al ir disminuyendo el número de clusters.

En el gráfico generado, se puede observar una caída significativa en el gradiente de las distancias de enlace, lo cual indica el punto óptimo en el que detener la agrupación. Ese punto representa el número ideal de clusters antes de que las fusiones empiecen a agrupar elementos demasiado diferentes entre sí.

En base al gráfico resultante, se identificó que el número óptimo de clusters es aproximadamente 5, ya que es en ese punto donde se produce el codo en la curva de distancias, indicando una agrupación natural en los datos.

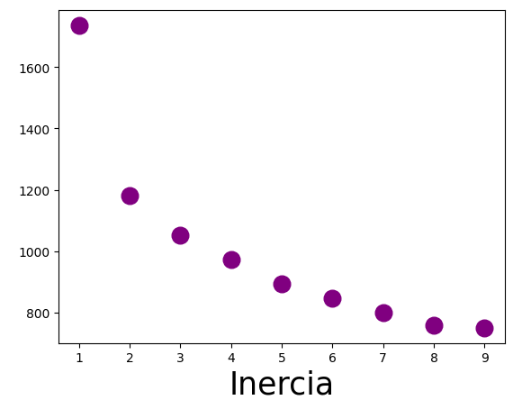


Para aplicar el algoritmo de KMeans, se requiere conocer de antemano el número óptimo de clusters (k). Para determinar este valor, se utilizó el método del codo, el cual se basa en la evaluación de la inercia en función del número de clusters.

En este análisis, se ejecutó el algoritmo KMeans para valores de k del 1 al 9. En cada caso, se registró la inercia, es decir, la suma de las distancias cuadradas de cada punto a su centroide más cercano.

Luego, se graficaron los valores de inercia frente al número de clusters. En este tipo de gráfico, el punto donde la reducción de inercia comienza a disminuir significativamente, indica el número óptimo de clusters.

En el gráfico generado, se observa que la inercia disminuye rápidamente hasta k = 5, y luego la mejora se vuelve marginal. Esto indica que 5 clusters es un valor apropiado para aplicar KMeans, ya que permite una segmentación eficiente sin sobre ajustar el modelo.

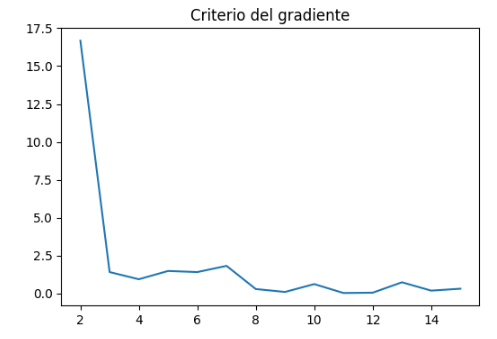


El criterio del gradiente se basa en analizar la variación de la inercia entre valores consecutivos de k, observando cómo disminuye la ganancia de información al aumentar el número de grupos.

En este análisis, se calcularon las diferencias de inercia (gradiente) y posteriormente se graficaron en orden inverso, obteniendo así una representación de la velocidad con la que cambia la inercia a medida que se incrementa el número de clusters.

En el gráfico generado se observa que el gradiente presenta un valor muy alto al inicio, lo que indica una gran mejora al pasar de un solo cluster a dos. Sin embargo, a partir de aproximadamente k = 7, las variaciones del gradiente se estabilizan y los cambios se vuelven mínimos.

Esto sugiere que 7 clusters constituye un número adecuado para la aplicación de KMeans, ya que a partir de este punto el incremento en el número de grupos no genera una mejora significativa en la calidad de la segmentación.

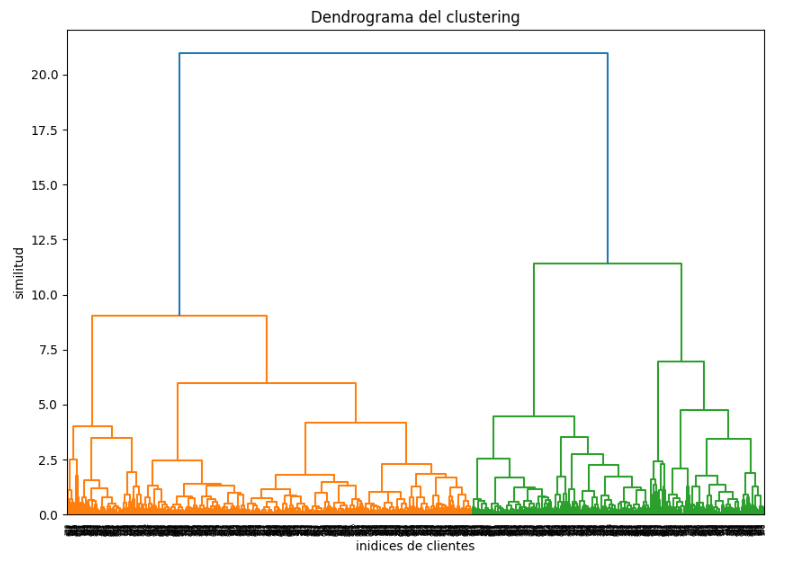


Para complementar el análisis de segmentación, también se aplicó un clustering jerárquico aglomerativo, cuya representación gráfica se obtiene a través de un dendrograma.

Este método se basa en calcular una matriz de similitud entre los clientes utilizando la similitud del coseno, y posteriormente aplicar la métrica euclidiana junto con el método de Ward, que busca minimizar la varianza dentro de cada grupo en cada paso de la agrupación.

En el dendrograma, cada cliente se representa en el eje X, mientras que el eje Y indica el nivel de similitud (o distancia) en el que se van formando las uniones entre los diferentes elementos o grupos.

En el gráfico se observa cómo, a medida que se asciende, los clientes se van agrupando en ramas cada vez más grandes hasta formar un único conjunto. Cortando el dendrograma a una determinada altura, se puede definir el número de clusters más adecuado. En este caso, se aprecia que existe una división natural en dos grandes grupos principales, aunque también podrían identificarse subgrupos más pequeños dependiendo del nivel de segmentación que se desee aplicar.



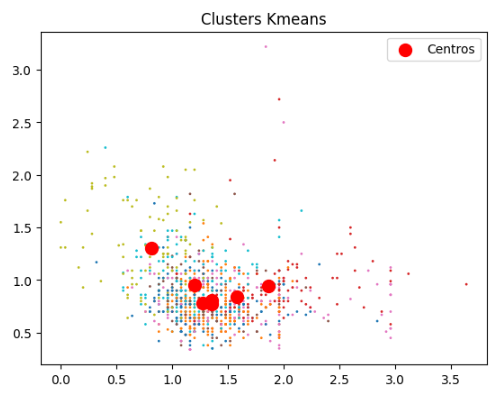
## 3.2. Kmeans (codo y gradiente)

Para finalizar el análisis, se aplicó el algoritmo KMeans con el número de clusters óptimo previamente determinado mediante los criterios del codo y del gradiente, que sugirieron la elección de 7 grupos.

En la ejecución del algoritmo, cada observación fue asignada a uno de los 7 clusters, representados en el gráfico por distintos colores. Además, los centroides de cada grupo se muestran resaltados en color rojo y con un tamaño mayor, lo que permite identificar la posición promedio de cada cluster dentro del espacio de características.

El resultado visual evidencia cómo los datos se agrupan en diferentes regiones, mostrando una segmentación clara en siete conjuntos bien diferenciados. Esta distribución confirma la validez de la elección de k = 7, ya que los clusters mantienen coherencia interna y separación respecto a los demás grupos.

En conclusión, el gráfico de KMeans permite observar de manera intuitiva la estructura de la segmentación obtenida, donde cada cluster representa un grupo de clientes con características similares y los centroides actúan como representantes de cada segmento.



# CONCLUSIÓN.

* El análisis permitió descubrir patrones claros en las preferencias de los usuarios de TripAdvisor.
* La consistencia entre el clustering jerárquico y KMeans valida la segmentación obtenida.
* Este tipo de segmentación es útil para personalizar recomendaciones, mejorar experiencias de viaje y diseñar campañas de marketing dirigidas.
* En general, los usuarios tienen una percepción positiva de los destinos turísticos del Este de Asia.