ANÁLISIS DE HOTELES DE LUJO EN EUROPA

Manuel Díaz Gil

Lucía Falcón García

José D. Sánchez Jiménez

Índice

Análisis de hoteles de lujo en Europa

1. Reproducción de scripts en R:

Para la realización de este apartado hemos usado la herramienta RStudio y el lenguaje de programación R.
El conjunto de datos escogido ha sido el siguiente: https://www.kaggle.com/sampsonsimpson/exploring-515k-european-hotel-reviews#L258

Hemos utilizado diferentes scripts, y a partir de ellos hemos llevado a cabo una serie de modificaciones en el código original para poder llegar al resultado deseado.

- **Modificación 1**: añadir tres nuevos campos al conjunto de datos original (país, ciudad del hotel y turista [Si-No]). Estos campos los hemos podido calcular a partir del campo “Hotel_Adress” y “Reviewer_Nationality”.

- **Modificación 2**: originalmente las gráficas no tenían un orden lógico, por ello, decidimos realizar una modificación para ordenarlas de forma ascendente y descendente, según el caso. Gracias a ello pudimos interpretar el resultado de una forma más clara y sencilla.

- **Modificación 3:** en las gráficas de nubes de palabras hemos quitado algunos artículos que aparecían y que eran irrelevantes para el resultado final. Uno de estos artículos es: **“The”**.

- **Modificación 4:** en algunas de las gráficas de puntuaciones hemos hecho modificaciones para que la ciudad analizada fuera Barcelona, ya que originalmente venía para otra ciudad que no era de España. Por lo tanto, nos parecía más interesante el análisis a una ciudad española que a una extranjera..

- **Modificación 5:** se ha añadido en el eje x un rango de días en la gráfica, en la cual se relacionan las puntuaciones medias que dejan los usuarios y el número de días transcurridos desde que abandonaron el hotel.

El resultado, y correspondiente análisis de cada gráfica, se puede encontrar dentro del proyecto en R llamado **“Proyecto_Kaggle.Rproj”** que se adjunta con la demás documentación. Se ha exportado el proyecto a HTML con el nombre de **“quick-visualization-of-data.html”**. En este HTML, no se muestran todas las gráficas del componente wordcloud.

2. Clustering

Antes de hacer clustering con Knime tuvimos que encontrar el número de clusters adecuado para ello. Este número pudimos conseguirlo a través de la técnica promedio Silhouette en python.

2.1 Algoritmo Silhouette

Debido al gran tamaño de nuestro conjunto de datos (515.000 filas), tuvimos que hacer pruebas con una cantidad menor de registros.

- 1ª prueba: 1.000 registros:

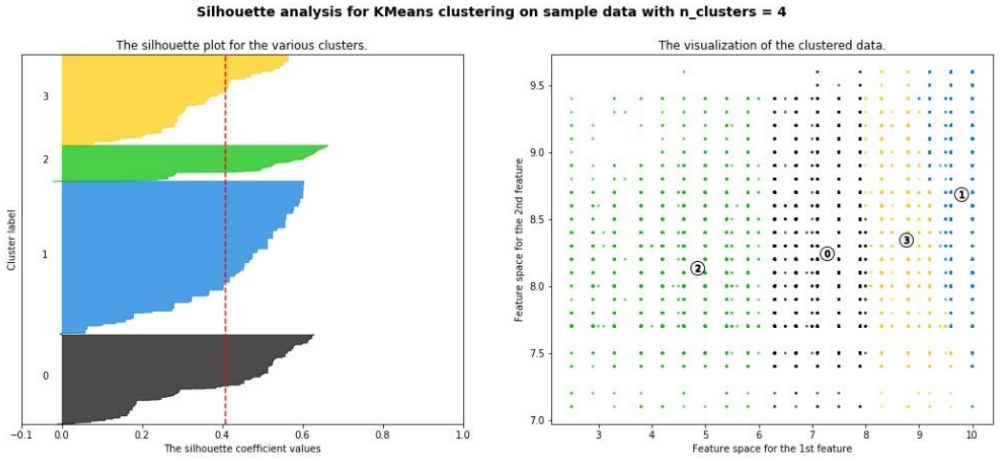
```
For n_clusters = 2 The average silhouette_score is : 0.5882203351480803
For n_clusters = 3 The average silhouette_score is : 0.5592659805680047
For n_clusters = 4 The average silhouette_score is : 0.4719227588640808
For n_clusters = 5 The average silhouette_score is : 0.48103099766997226
For n_clusters = 6 The average silhouette_score is : 0.48004829685075084
```

- 2ª prueba: 10.000 registros:

```
For n_clusters = 2 The average silhouette_score is : 0.5632929271807481
For n_clusters = 3 The average silhouette_score is : 0.5294968842103306
For n_clusters = 4 The average silhouette_score is : 0.4351399227297383
For n_clusters = 5 The average silhouette_score is : 0.41361113504279884
For n_clusters = 6 The average silhouette_score is : 0.44421797628255955
```

- 3ª prueba: 50.000 registros (última):

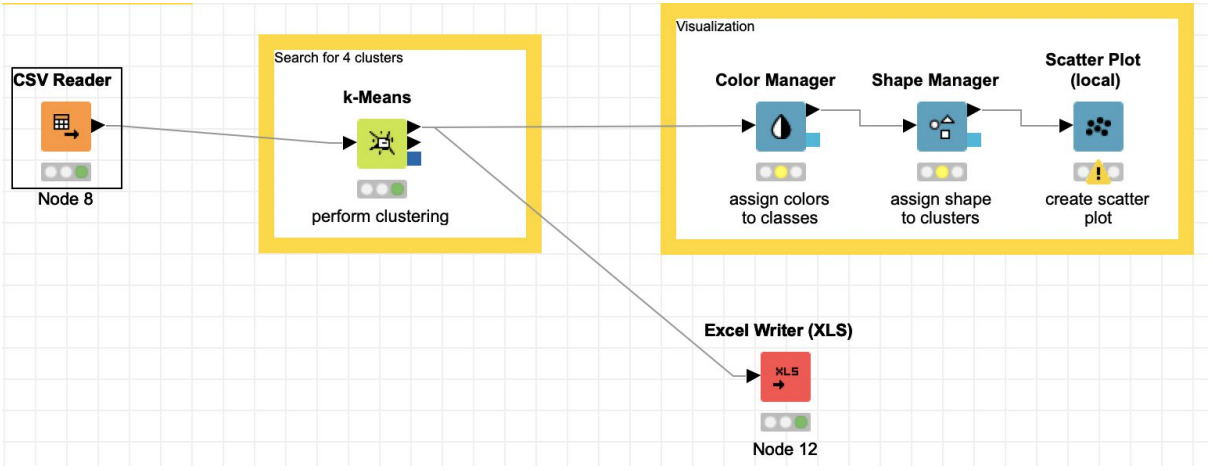
```
For n_clusters = 2 The average silhouette_score is : 0.5794782947140055
For n_clusters = 3 The average silhouette_score is : 0.5281016741744753
For n_clusters = 4 The average silhouette_score is : 0.40847991450735377
For n_clusters = 5 The average silhouette_score is : 0.4057800505723598
For n_clusters = 6 The average silhouette_score is : 0.39379789460370085
```



Para todos los casos se puede observar como el cambio más brusco en el coeficiente se produce entre el número 3 al 4, por lo que finalmente nos decantamos por elegir 4 como número óptimo de clusters, que es cuando empieza a mantenerse constante el valor del coeficiente al aumentar el número de clusters.

2.2 Análisis con Knime

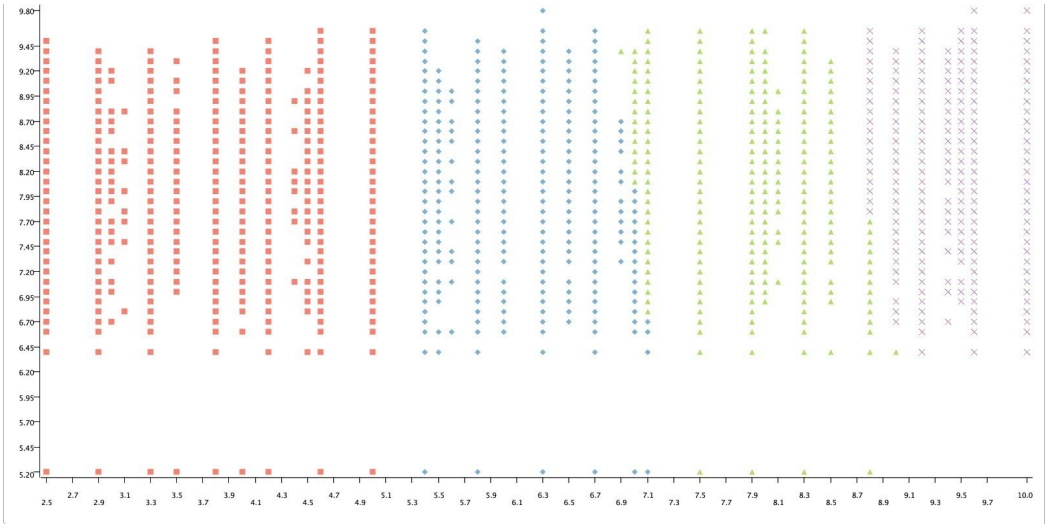
Para hacer clustering con Knime hicimos uso del algoritmo k-means, basándonos en las siguientes características: puntuación media del hotel y la puntuación que ha dejado el cliente en la review. El proceso fué el siguiente:



Una vez ejecutado, obtuvimos los siguientes centroides:

Row ID	Average_Score	Reviewer_Score
cluster_0	7.964	4.175
cluster_1	8.571	9.56
cluster_2	8.247	7.802
cluster_3	8.116	6.176

Resultado gráfico:



En ambas capturas de pantallas podemos observar como el cluster crítico es el número 0. Los hoteles que pertenecen a este cluster tienen en general una buena nota media, sin embargo, los clientes le están dando una mala puntuación (4,1 de media). Como consecuencia, esto supone que la nota media de estos hoteles podría disminuir bastante. Por ello, vamos a centrarnos en el cluster número 0, para poder analizarlo más a fondo.

2.3 Tablas de contingencia en Excel

Una vez exportado el fichero de Knime a formato Excel, con los datos asignados a cada cluster, hicimos tablas de contingencias, usando tablas

dinámicas. Las variables asociadas a la tabla son: país donde está ubicado el hotel y el cluster al que pertenece.

En la siguiente tabla obtenemos el número total de hoteles por países que hay en cada cluster.

Cuenta de Reviewer_Score	Etique				
Paises	cluster_0	cluster_1	cluster_2	cluster_3	Total general
Austria	1737	23522	10214	3466	38939
France	3704	34393	15702	6129	59928
Italy	2320	20451	10390	4046	37207
Netherlands	3087	33320	14875	5932	57214
Spain	2876	36584	15125	5564	60149
United Kingdom	16979	141232	72771	31319	262301
Total general	**30703**	**289502**	**139077**	**56456**	**515738**

Si pasamos estos datos a porcentajes, obtenemos la siguiente tabla:

Elemento cluster / nº total fila * 100				
	cluster_0	cluster_1	cluster_2	cluster_3
Austria	4,46%	60,41%	26,23%	8,90%
France	6,18%	57,39%	26,20%	10,23%
Italy	6,24%	54,97%	27,92%	10,87%
Netherlands	5,40%	58,24%	26,00%	10,37%
Spain	4,78%	60,82%	25,15%	9,25%
United Kingdom	6,47%	53,84%	27,74%	11,94%

En la tabla anterior podemos observar para cada país los porcentajes de hoteles que hay en cada clúster.

Por ejemplo, para Austria, del 100% que representa la fila, un 4,46% de hoteles estarían situados en el cluster 0. Los hoteles de este cluster como hemos visto anteriormente son los que peor puntuación de media están recibiendo.

En la siguiente tabla, el porcentaje indica que los hoteles británicos son mayoría en el cluster crítico 0, es decir, los peores puntuados con diferencia. Haciendo referencia a la tabla anterior donde se indica que el 6,47% de ese 55,30% son los que peor puntuación consiguen.

Elemento cluster / nº total columna * 100				
	cluster_0	cluster_1	cluster_2	cluster_3
Austria	5,66%	8,12%	7,34%	6,14%
France	12,06%	11,88%	11,29%	10,86%
Italy	7,56%	7,06%	7,47%	7,17%
Netherlands	10,05%	11,51%	10,70%	10,51%
Spain	9,37%	12,64%	10,88%	9,86%
United Kingdom	55,30%	48,78%	52,32%	55,48%

Hicimos otra tabla dinámica que relacionaba la nacionalidad del cliente que puntuaba en lugar del país donde se encontraba el hotel, pero decidimos eliminarlo ya que al existir unas 220 nacionalidades distintas la tabla salía demasiado grande, y con porcentajes pequeños, por lo que no llegamos a un resultado que fuera representativo.

3. Conclusiones

Hemos podido identificar el cluster crítico para analizarlo y sacar la siguiente información:

- Los hoteles británicos con diferencia son los que más puntuaciones bajas reciben por parte de los clientes.

- Analizamos la procedencia de los clientes que daban peores puntuaciones pero al haber tantas nacionalidades no llegamos a ningún resultado concreto.

4. Bibliografía

Dataset y Script R

- https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe/data

Silhouette

- https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Script K-means

- Ejemplos de Knime