Introducción y motivación

Los Pollos Hermanos es una cadena de comida rápida extranjera que como parte de su expansión a Europa ha abierto el año pasado un restaurante piloto en Madrid.

Los resultados obtenidos por el restaurante han superado las expectativas de la empresa por lo que han decidido continuar su expansión. Para la primera fase han decidido abrir cinco restaurantes más en la ciudad de Madrid.

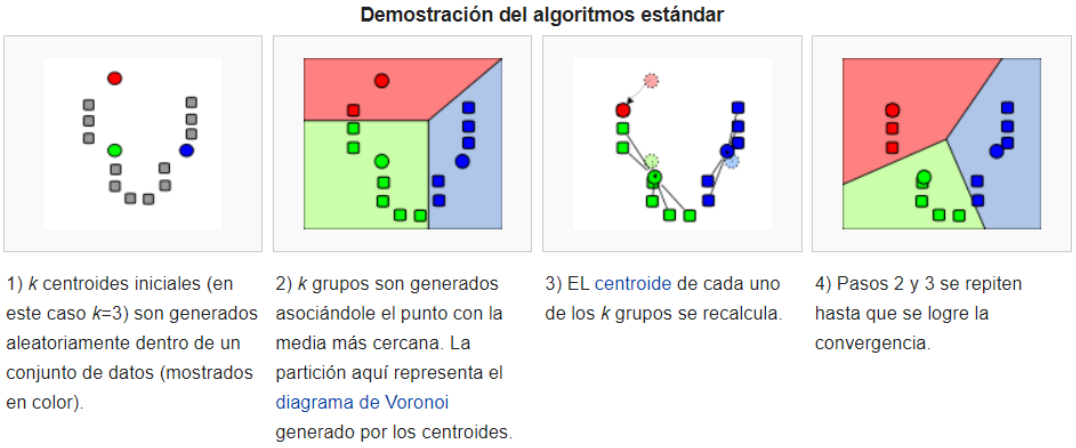
La dirección de la empresa no está familiarizada con la ciudad por lo que le ha encargado a una consultora, la misma que eligió la ubicación del primer restaurante, que busque cuales son los barrios más parecidos al actual. Ya que es donde la dirección de la empresa espera que se encuentre en mayor proporción su público objetivo.

El propósito del estudio es **determinar cuáles, de los 131 barrios de la ciudad de Madrid, son más parecidos al barrio en el que se encuentra el restaurante piloto**.

Material y Métodos

K-means

Para ello la empresa consultora ha decidido agrupar los barrios en clústeres utilizando K-means, una técnica de aprendizaje no supervisado. Esta técnica utiliza las características de los barrios, en este caso los tipos de negocios y lugares de interés que hay en cada barrio, para agruparlos en K grupos. Este algoritmo busca maximizar la distancia entre grupos y minimizar la distancia intra-grupo.



[Imagen de Wikipedia: https://es.wikipedia.org/wiki/K-medias]

Gracias a un primer análisis realizado empleando información demográfica se ha determinado que en la ciudad de Madrid hay 7 tipos de barrios. Por lo que en este caso se agruparan los barrios en 7 clústeres.

Datos

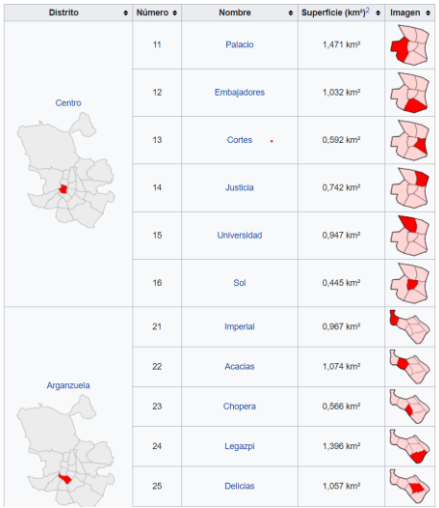
Los datos para realizar el análisis se obtendrán de diversas fuentes:

WIKIPEDIA
La enciclopedia libre

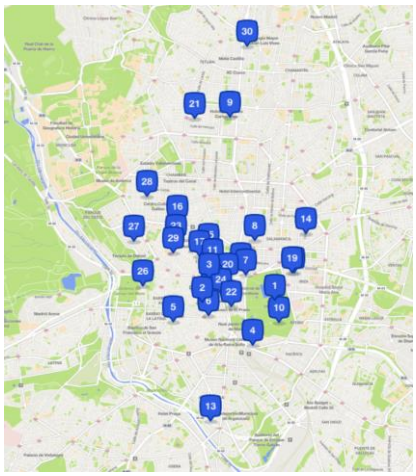
Se utilizará Wikipedia para obtener la lista de barrios y distritos de la Ciudad de Madrid.

https://es.wikipedia.org/wiki/Anexo:Distritos_de_Madrid

https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid



Se utilizará Foursquare para conseguir información sobre los barrios y distritos, concretamente sobre el tipo de negocios y lugares de interés que hay en cada zona. Para ello accederemos a los datos de Foursquare mediante su REST API. Esta información se utilizará para agrupar los barrios de Madrid en clústeres en función del tipo de negocios y lugares de interés que hay en cada barrio.



Nominatim se utilizará para obtener las coordenadas de cada barrio de Madrid. Nominatim utiliza los datos de OpenStreetMap para encontrar las coordenadas de una dirección (geocoding). La API de Nominatim ha dado error en 2 de los 131 barrios de Madrid:

```
Madrid_df.iloc[128,:]
```

```
Distrito                      Barajas
Código                            213
Barrio      Casco Histórico de Barajas
Tamaño                          0.609
Name: 128, dtype: object
```

```
Madrid_df.iloc[116,:]
```

```
Distrito      Vicálvaro
Código              193
Barrio      Valderrivas
Tamaño          4.72291
Name: 116, dtype: object
```

Al tratarse únicamente de dos errores las coordenadas de estos dos barrios se han obtenido manualmente buscándolas en geohack:

https://geohack.toolforge.org/geohack.php?language=es&pagename=Casco_Hist%C3%B3rico_de_Barajas&params=40.47361111_N_-3.57722222_E_type:city

https://geohack.toolforge.org/geohack.php?language=es&pagename=Valderrivas&params=40.401583333333_N_-3.5990944444444_E_type:city

Etapas de la preparación de los datos:

- Descargar la lista de barrios de Madrid de Wikipedia y darle el formato adecuado para las etapas posteriores.

	Distrito	Número	Nombre	Superficie (km²)[2]	Imagen
0	Centro	11	Palacio	1,471 km²	NaN
1	Centro	12	Embajadores	1,032 km²	NaN
2	Centro	13	Cortes	0,592 km²	NaN
3	Centro	14	Justicia	0,742 km²	NaN
4	Centro	15	Universidad	0,947 km²	NaN

	Distrito	Código	Barrio	Tamaño
0	Centro	11	Palacio	1.471
1	Centro	12	Embajadores	1.032
2	Centro	13	Cortes	0.592
3	Centro	14	Justicia	0.742
4	Centro	15	Universidad	0.947

- El siguiente paso es añadir al dataframe las coordenadas de cada barrio de Madrid que obtenemos con Nominatim.

	Distrito	Código	Barrio	Tamaño	Latitud	Longitud
0	Centro	11	Palacio	1.471	40.415129	-3.715618
1	Centro	12	Embajadores	1.032	40.409681	-3.701644
2	Centro	13	Cortes	0.592	40.414348	-3.698525
3	Centro	14	Justicia	0.742	40.423957	-3.695747
4	Centro	15	Universidad	0.947	40.425310	-3.706630

- El ultimo paso de la obtención de datos es la obtención de los lugares de interés en cada distrito utilizando la Rest API de Foursquare. Los datos se han estructurado en el siguiente dataframe:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Palacio	40.415129	-3.715618	Santa Iglesia Catedral de Santa María la Real ...	40.415767	-3.714516	Church
1	Palacio	40.415129	-3.715618	Plaza de La Almudena	40.416320	-3.713777	Plaza
2	Palacio	40.415129	-3.715618	Taberna Rayuela	40.413179	-3.713496	Tapas Restaurant
3	Palacio	40.415129	-3.715618	Corral de la Morería	40.412619	-3.714249	Performing Arts Venue
4	Palacio	40.415129	-3.715618	Palacio Real de Madrid	40.417940	-3.714259	Palace

Aquí concluye la preparación de los datos.