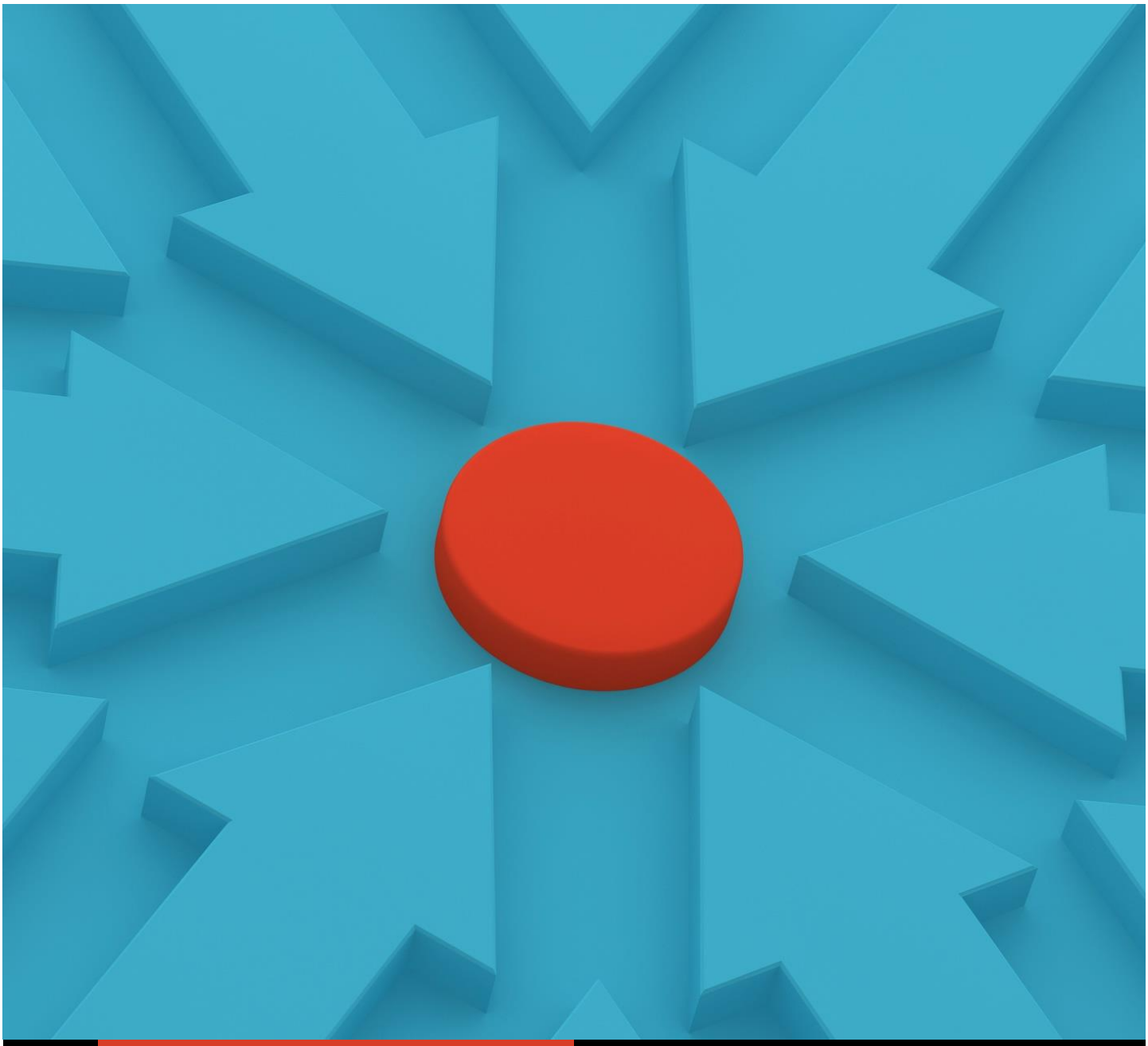
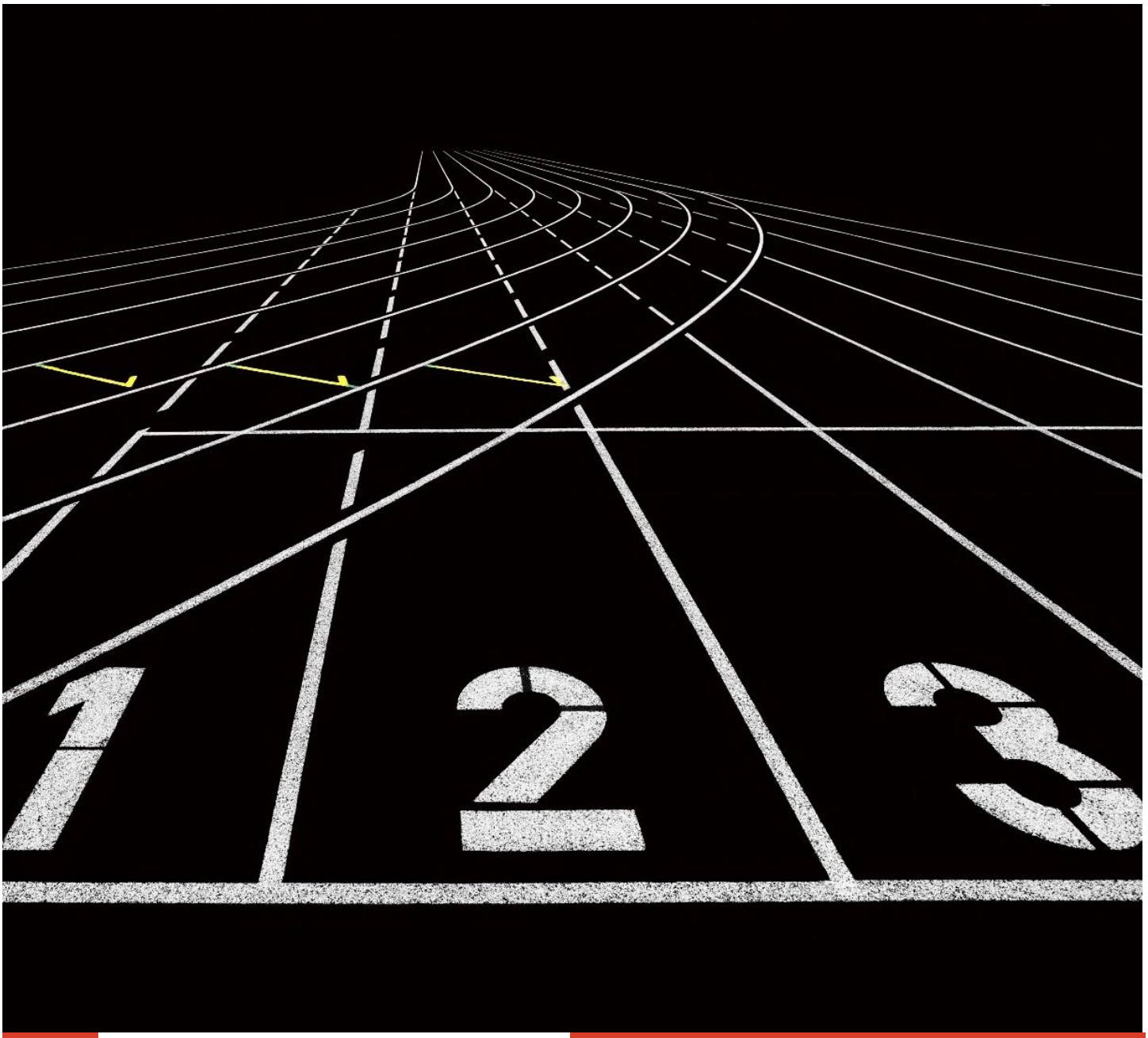


BARCELONA

DISTRITO A DISTRITO

JAVIER VIQUEIRA

IBM Data Science
Capstone Final Proyect



HOLA, BARCELONA

Cada año miles y miles de turistas viajan a Barcelona, algunos van buscando ocio y emociones, la playa o los parques de aventuras, otros buscan el descanso, la arquitectura de la ciudad o la gastronomía, unos viajan por trabajo y otros por placer, con amigos, solos o en pareja.

Barcelona es solo un ejemplo, hoy en día el mundo está en continuo movimiento y necesitamos comprender rápidamente cómo funciona la ciudad a la que recién llegamos, necesitamos un plano, sencillo, rápido que nos permita responder con agilidad a 2 preguntas, ¿qué y dónde?

INTRODUCCIÓN

Las ciudades generalmente construyen su oferta de servicios por zonas, para los residentes identificarlas suele ser una tarea fácil pero no es así para quienes llegan a la ciudad por primera vez, turistas sí, pero también empresas grandes y pequeñas, de todos los sectores. Empresas que ya están o empresas que piensan en instalares o crear nuevos proyectos en la ciudad.

Utilidad

Poder identificar las principales zonas de la ciudad de un solo vistazo es lo que vamos a conseguir en este proyecto mediante la aplicación del algoritmo de Machine Learning K-Means.

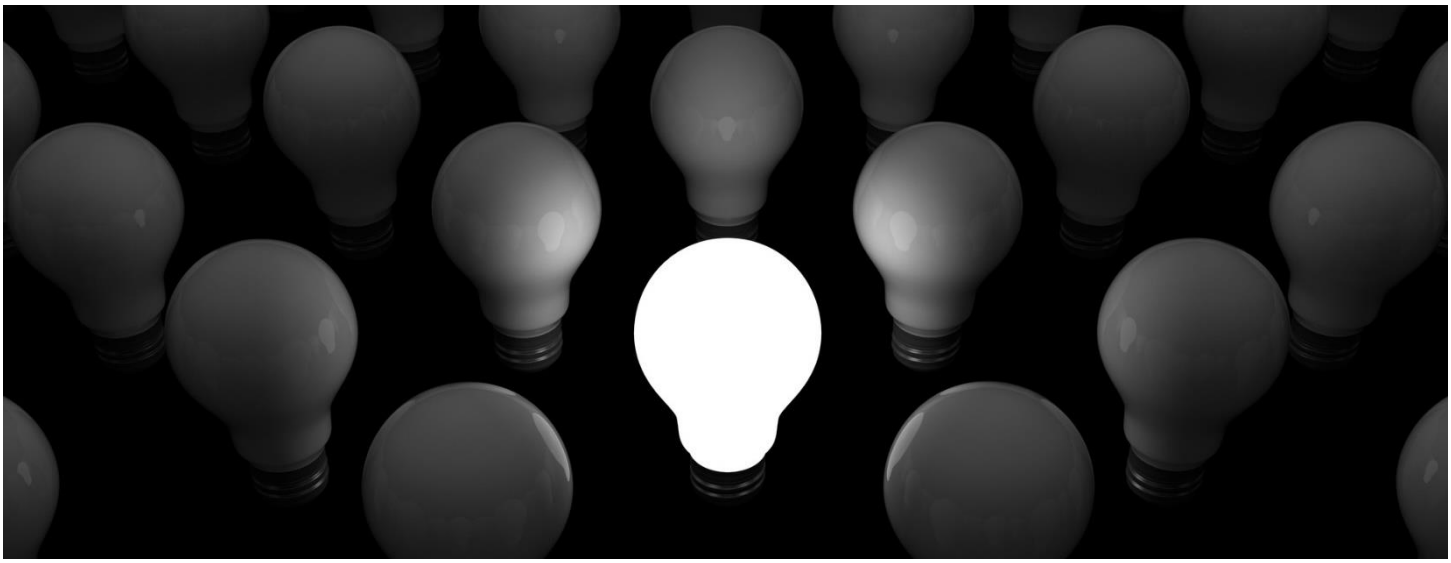


Ilustración 1

k-means

K-Means es un algoritmo no supervisado de Clustering. Se utiliza cuando tenemos un montón de datos sin etiquetar. El objetivo de este algoritmo es el de encontrar “K” grupos (clusters) entre los datos crudos.

DATOS

Utilizaremos varias fuentes de datos para poder llevar a cabo este proceso.

Wikipedia
Obtendremos los nombres de los diferentes distritos de Barcelona de la siguiente tabla de la wikipedia. https://es.wikipedia.org/wiki/Distritos_de_Barcelona

Foursqare
Mediante la API de Foursquare conseguiremos los datos de los sitios y negocios alrededor de nuestros lugares de interés.

METODOLOGÍA

Utilizaremos varias fuentes de datos para poder llevar a cabo este proceso.

PARTE 1: DATOS
Recopilación y procesado de datos, limpiaremos y filtraremos los datos obtenidos para nuestros propósitos mediante el uso de las librerías pandas y numpy principalmente. Una vez realizado este proceso añadiremos las coordenadas GPS a los distritos mediante geopy.

PARTE 2: EXPLORAR DATOS
Importaremos todas las librerías necesarias para el proceso de exploración, es de vital importancia conocer y comprender los datos antes de tratar de realizar un modelo con ellos.

Ubicación

Primero ubicaremos los diferentes distritos sobre el mapa de Barcelona

Descripción

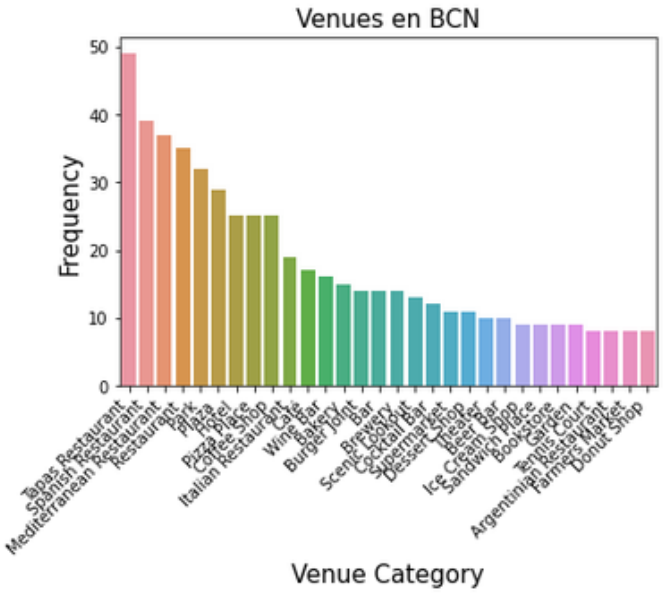
Agruparemos los datos obtenidos por categoría de negocio y por barrio para establecer relaciones entre los datos

Análisis

Analizaremos los rasgos más característicos de cada distrito de la ciudad

Visualización de datos

La visualización de datos mediante diferentes tipos de gráficos es una poderosa arma para comprender qué nos dicen los datos. A continuación un pequeño ejemplo extraído del notebook



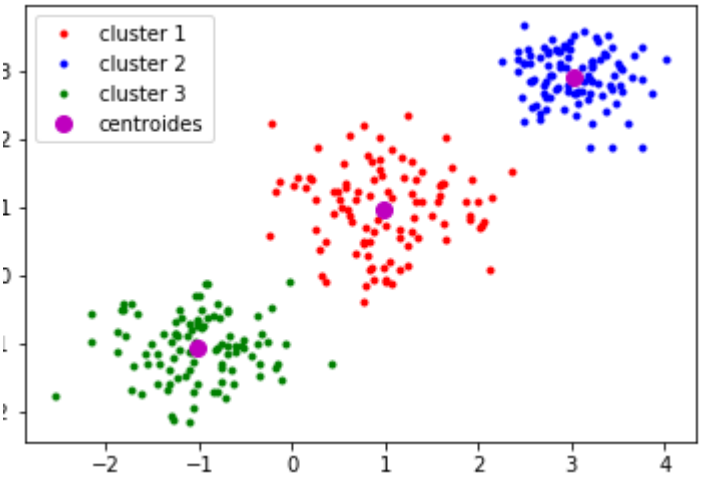
Agrupación y pivoteado

Por último, agruparemos y pivotearemos los datos de diferentes formas para comprender las relaciones entre ellos. Pueden crearse heatmaps o tablas que estudien su correlación. En nuestro caso agruparemos las diferentes categorías de negocio en cada uno de los diferentes distritos.

PARTE 3: Modelo

El algoritmo k-means

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática.



El algoritmo consta de tres pasos:

Inicialización: una vez escogido el número de grupos, k, se establecen k centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.

Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano.

Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

El algoritmo k-means resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su clúster.

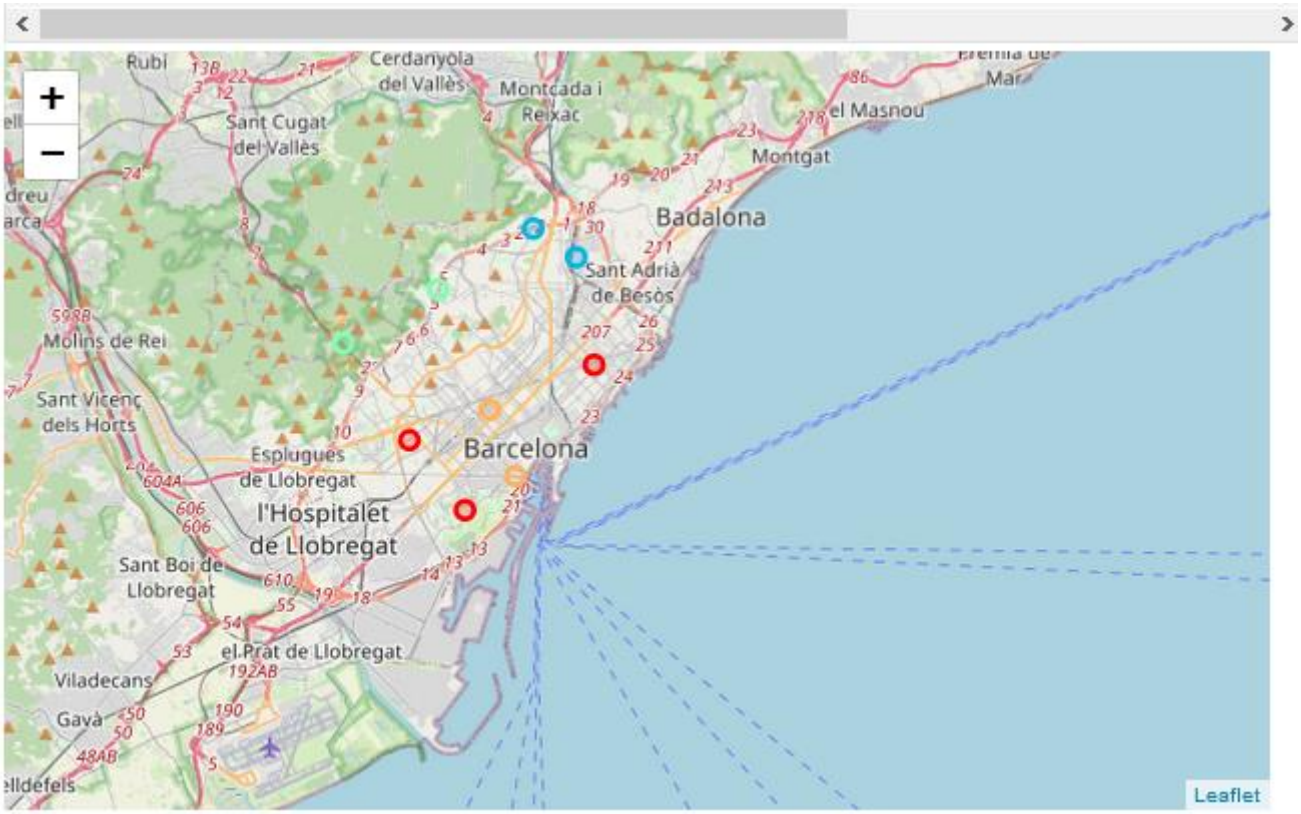
Tras procesar los datos y agruparlos como explicamos anteriormente, aplicando este algoritmo construiremos 5 clusters que recogerán las diferentes partes de la ciudad según sus características.

PARTE 4: Resultado y Análisis

En esta sección analizaremos los resultados y daremos un nombre descriptivo a cada uno de los clústeres. Así mismo mostraremos de nuevo el mapa de Barcelona, pero esta vez con el resultado final aplicado en forma de color a los diferentes distritos de la ciudad:

RESULTADOS

A continuación, se puede observar el mapa final de Barcelona y los diferentes distritos con sus nuevos colores identificativos.



En la siguiente página pueden observarse los diferentes clústeres asignados por el algoritmo, así como los principales sitios que podremos encontrar en cada uno de ellos.

Debajo hemos asignado un nombre representativo a cada uno de los clusters.

Parte 4: EVALUACIÓN CLUSTERS

1.- Examinamos los clusters.

Examinamos y damos nombre a cada cluster basandonos en sus características.

Cluster 1

```
BCN_merged.loc[BCN_merged['Cluster Labels'] == 0, BCN_merged.columns[[1] + list(range(5, BCN_merged.shape[1]))]]
```

	Superficie km^[1]	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	2262	0	Tapas Restaurant	Park	Wine Bar	Pizza Place	Italian Restaurant	Scenic Lookout	Café	Cocktail Bar	Coffee Shop	Beer Bar
3	602	0	Mediterranean Restaurant	Restaurant	Pizza Place	Japanese Restaurant	Spanish Restaurant	Café	Wine Bar	Cocktail Bar	Coffee Shop	Gastropub
9	1029	0	Park	Italian Restaurant	Mediterranean Restaurant	Restaurant	Beach	Bakery	Brewery	Historic Site	Spanish Restaurant	Supermarket

RESTAURANTES TEMÁTICOS, BARES

Cluster 2

```
BCN_merged.loc[BCN_merged['Cluster Labels'] == 1, BCN_merged.columns[[1] + list(range(5, BCN_merged.shape[1]))]]
```

	Superficie km^[1]	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	419	1	Plaza	Mountain	Pedestrian Plaza	Ethiopian Restaurant	Furniture / Home Store	Fountain	Food & Drink Shop	Food	Flea Market	Fish & Chips Shop

PLAZAS, MONTAÑAS

Cluster 3

```
BCN_merged.loc[BCN_merged['Cluster Labels'] == 2, BCN_merged.columns[[1] + list(range(5, BCN_merged.shape[1]))]]
```

	Superficie km^[1]	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	805	2	Tapas Restaurant	Spanish Restaurant	Plaza	Restaurant	Supermarket	Pizza Place	Burger Joint	Brewery	Performing Arts Venue	Mediterranean Restaurant
8	659	2	Tapas Restaurant	Restaurant	Plaza	Spanish Restaurant	Park	Brewery	Mediterranean Restaurant	Pizza Place	Farmers Market	Italian Restaurant

TAPAS

Cluster 4

```
BCN_merged.loc[BCN_merged['Cluster Labels'] == 3, BCN_merged.columns[[1] + list(range(5, BCN_merged.shape[1]))]]
```

	Superficie km^[1]	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	1991	3	Scenic Lookout	Theme Park	Mediterranean Restaurant	Spanish Restaurant	Park	Tapas Restaurant	Restaurant	Train Station	Trail	Theme Park Ride / Attraction
6	1196	3	Spanish Restaurant	Plaza	Restaurant	Park	Tapas Restaurant	Mediterranean Restaurant	Scenic Lookout	Sandwich Place	Mountain	Soccer Field

OCIO, RESTAURANTE ESPAÑOL

Cluster 5

```
BCN_merged.loc[BCN_merged['Cluster Labels'] == 4, BCN_merged.columns[[1] + list(range(5, BCN_merged.shape[1]))]]
```

	Superficie km^[1]	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	411	4	Hotel	Tapas Restaurant	Pizza Place	Coffee Shop	Plaza	Italian Restaurant	Beer Bar	Donut Shop	Spanish Restaurant	Bookstore
1	765	4	Coffee Shop	Hotel	Mediterranean Restaurant	Tapas Restaurant	Bookstore	Wine Bar	Dessert Shop	Donut Shop	Bakery	Burger Joint

HOTELES, CAFFES, TIENDAS

DISCUSION Y CONCLUSIONES

Siendo un modelo sencillo y rápido de ejecutar arroja unos resultados prometedores. Sería buena idea aplicar el modelo por calles para obtener mayor precisión, así mismo podrían optimizarse también el número de clusters que creamos con el algoritmo k means como también el número de sitios de cada una de las zonas que empleamos en los cálculos. En nuestro caso, al emplear zonas grandes como son los diferentes distritos no hemos obtenido demasiada sensibilidad al cambio en estos parámetros, seguramente a nivel de calles si sea mucho más sensible y deberemos evaluarlo procurando utilizar todas las directrices para evitar el overfit.

En cuanto al resultado estamos satisfechos pues coincide con nuestro conocimiento de la ciudad y pese a haber sido realizado con objetivos didácticos cumple su función y podría ser base de nuevos desarrollos.

Proyecto Capstone - La Batalla de los Vecindarios

Javier Viqueira

IBM Data Science Professional Certificate | Coursera 2021