Proyecto Capstone: Análisis de los vecindarios de Toronto y Nueva York

Ciencia de Datos Aplicada - Curso Capstone

IBM / Coursera (Agosto, 2022)

Integrante: Sebastián Grimberg Saralegui.

Contenido del informe:

1 - Introducción

1.1 - Acerca del informe

En este proyecto se propone realizar tanto un análisis descriptivo como así mismo una comparación de los distintos vecindarios de las ciudades capitales de Canadá y de Estados Unidos: Toronto y Nueva York. Primero vamos a observar como están compuestas estas ciudades y luego nos adentramos en el análisis y se buscarán similitudes y diferencias entre los diferentes vecindarios tomando en cuenta los lugares más comunes de estos y sus categorías correspondientes. El informe se encuentra dirigido a una agencia de viajes interesada en recabar información acerca de estos dos grandiosos destinos como parte de un estudio de mercado.

1.2 - Acerca de su importancia

La importancia de este análisis radica en poder conocer más en profundidad acerca de estas dos grandes ciudades, conocer las características de sus vecindarios que los hacen tanto más diferentes o similares en una y otra ciudad, información muy valiosa por ejemplo para turistas, emprendedores, profesionales, estudiantes, etc. que puedan estar pensando mudarse a una gran ciudad para llevar a cabo sus planes de vida, de negocios, estudio o simplemente sus planes turísticos.

2 - Datos

2.1 - Primer etapa de recolección

Obtenemos los datos geoespaciales correspondientes a los municipios y sus respectivos vecindarios de cada una de las ciudades. Luego aplicamos la integración de los datos de ambas ciudades para tenerlos en una sola tabla.

Observamos las primeras 5 filas de la tabla:

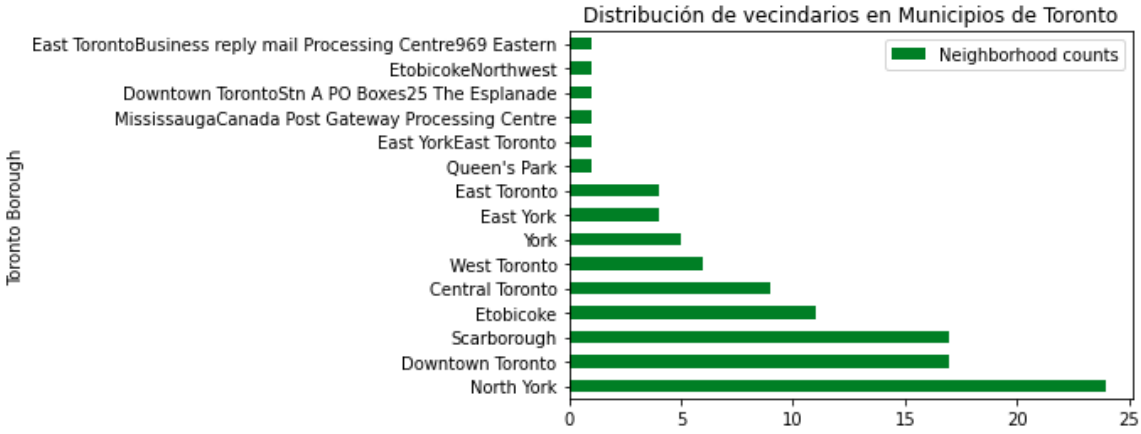
	City	Borough	Neighborhood	Latitude	Longitude
0	Toronto	North York	Parkwoods	43.753259	-79.329656
1	Toronto	North York	Victoria Village	43.725882	-79.315572
2	Toronto	Downtown Toronto	Regent Park , Harbourfront	43.654260	-79.360636
3	Toronto	North York	Lawrence Manor , Lawrence Heights	43.718518	-79.464763
4	Toronto	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

Observamos las últimas 5 filas de la tabla:

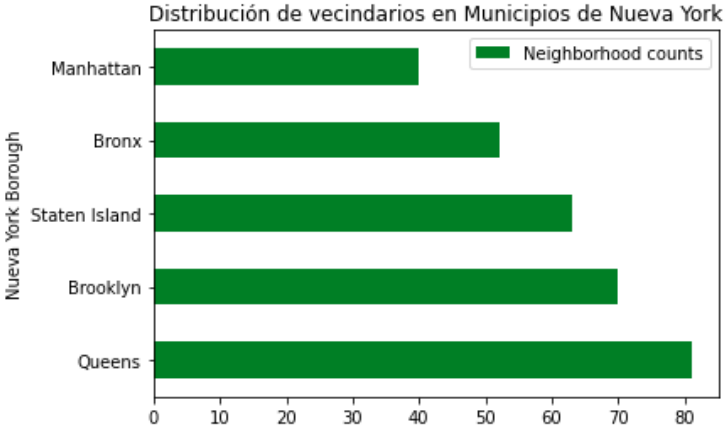
	City	Borough	Neighborhood	Latitude	Longitude
404	Nueva York	Manhattan	Hudson Yards	40.756658	-74.000111
405	Nueva York	Queens	Hammels	40.587338	-73.805530
406	Nueva York	Queens	Bayswater	40.611322	-73.765968
407	Nueva York	Queens	Queensbridge	40.756091	-73.945631
408	Nueva York	Staten Island	Fox Hills	40.617311	-74.081740

En los siguientes gráficos de barra visualizamos cómo se distribuyen los vecindarios dentro de los municipios de ambas ciudades:

En Toronto:

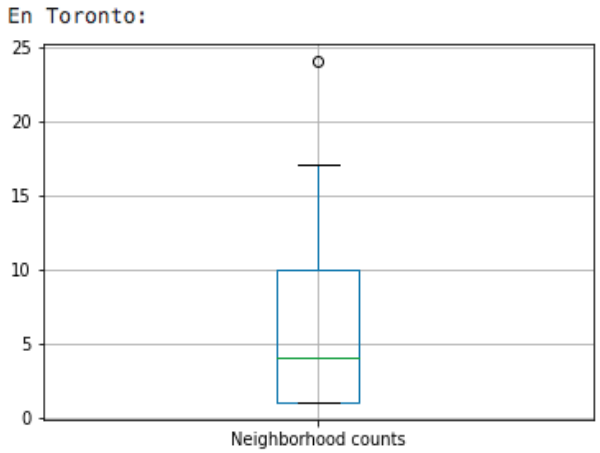


En Nueva York:

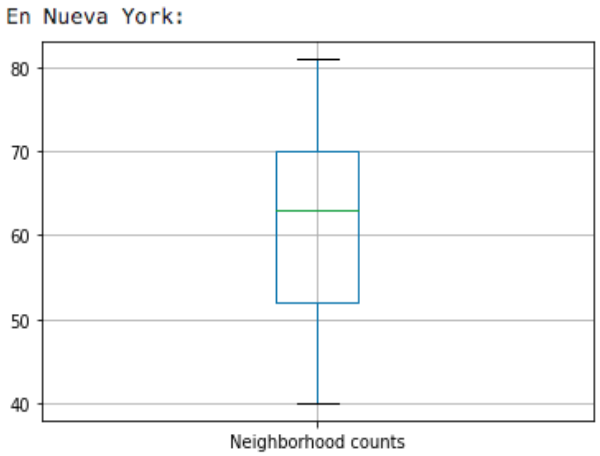


Se puede observar que en Toronto se tiene más municipios y están conformados por menos vecindarios. Observamos las distribuciones para cada ciudad en los gráficos de caja:

En Toronto:



En Nueva York:



2.2 - Segunda etapa de recolección

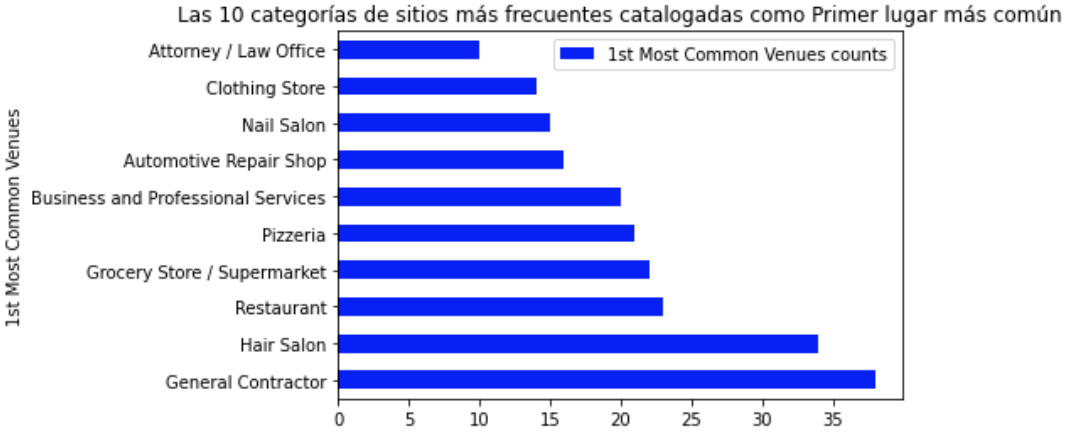
A través de la API de Foursquare se extrajeron los datos de ubicaciones de distintos lugares y sus categorías.

Observamos las primeras 5 filas de la tabla:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.754751	-79.328439	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751978	-79.333389	Food and Beverage Retail
2	Parkwoods	43.753259	-79.329656	Money in Motion	43.752947	-79.332418	Financial Service
3	Parkwoods	43.753259	-79.329656	Toronto Custom Lights	43.752947	-79.332418	Business and Professional Services
4	Parkwoods	43.753259	-79.329656	Pneutrans Systems Ltd	43.754856	-79.327753	Business and Professional Services

Ahora que se tienen los datos de distintos lugares y sus categorías por vecindarios, se pueden observar cuales son las categorías que frecuentan más. En este caso se extrajeron las 10 categorías relacionadas a los lugares más comunes y aquí tenemos el resultado en el siguiente gráfico de barras:

También se ha observado cuales fueron las 10 categorías que fueron catalogadas como "primer lugar más común" en los vecindarios, y aquí tenemos el resultado en el siguiente gráfico de barras:



Estas son las categorías más predominantes entre muchas categorías para las dos ciudades en su conjunto. Se puede observar por ejemplo un claro predominio de la categoría "Hair Salon" que es la que más frecuenta entre las 10 categorías más comunes de los vecindarios y además tiene una notoria diferencia junto a la categoría "General Contractor" siendo estas las que aparecen por gran diferencia (comparado a las demás) como primer categoría más común en los vecindarios.

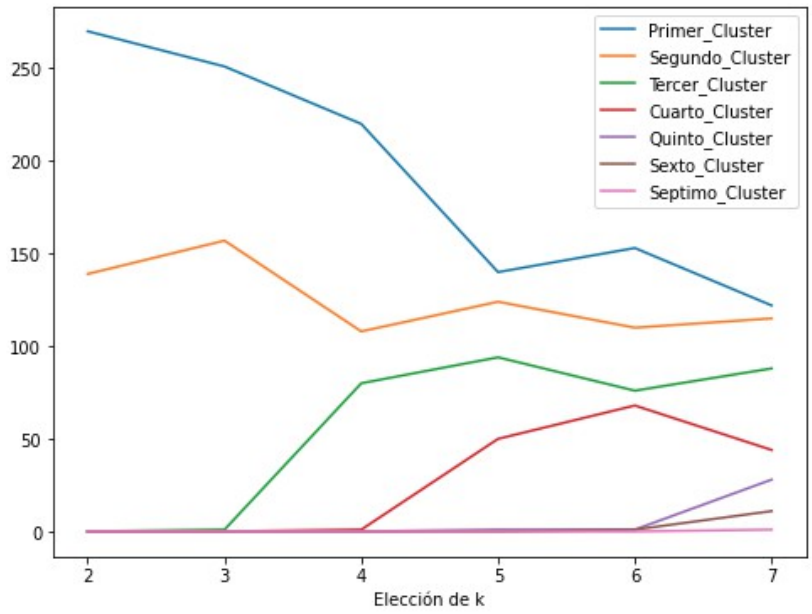
Si observamos la categoría “Pizzería” por ejemplo, observamos que si bien es la segunda más importante tomando en cuenta las 10 categorías más comunes en los vecindarios, sin embargo no es de las que frecuenta más como primer categoría más común.

3 - Metodología

Ahora que ya contamos con los datos de los distintos vecindarios y las categorías de sitios más comunes dentro de cada uno de ellos, nos enfocamos en aplicar el modelo de aprendizaje automático no supervisado: **Método de agrupamiento k-medias**. A través de esta técnica buscamos agrupar los distintos vecindarios sin tomar en cuenta si pertenecen a una u otra ciudad, sino que nos interesan similitudes y diferencias de estos para poder agruparlos, y para ello partimos de la información sobre categorías de sitios más comunes. Se aplicarán así mismo **métodos de visualización de datos** para poder observar en el mapa de ambas ciudades como quedaron conformados los grupos asignando distintos colores a las marcas de ubicación según donde han sido asignados.

3.1 - Elección del parámetro k del método k - medias

Experimentando con distintos valores de k, se ha recolectado datos sobre cuántos vecindarios conforman los distintos cluster para cada k elegido. Luego visualizamos gráficamente esto mismo para tomar una decisión:



Para **k=5** notamos una fuerte caída del número de vecindarios en el mayor cluster, y podemos ver a su vez cómo se estaría conformando un nuevo cluster bastante significativo en cantidad de vecindarios a su vez que crecen significativamente otros dos. Luego con k mayores a 5 las variaciones comienzan a ser poco significativas, por tanto en nuestro caso este número de cluster a formar será nuestra elección.

3.2 - Visualización de mapas y agrupaciones de vecindarios

Se ha aplicado el método k - medias tomando en cuenta las 10 categorías asociadas a los lugares

más comunes de los vecindarios, conformando 5 grupos de vecindarios ya que anteriormente elegimos el parámetro k=5. Generamos una tabla en la cuál agregamos la información de los cluster a la cuál pertenecen los vecindarios según el modelo aplicado.

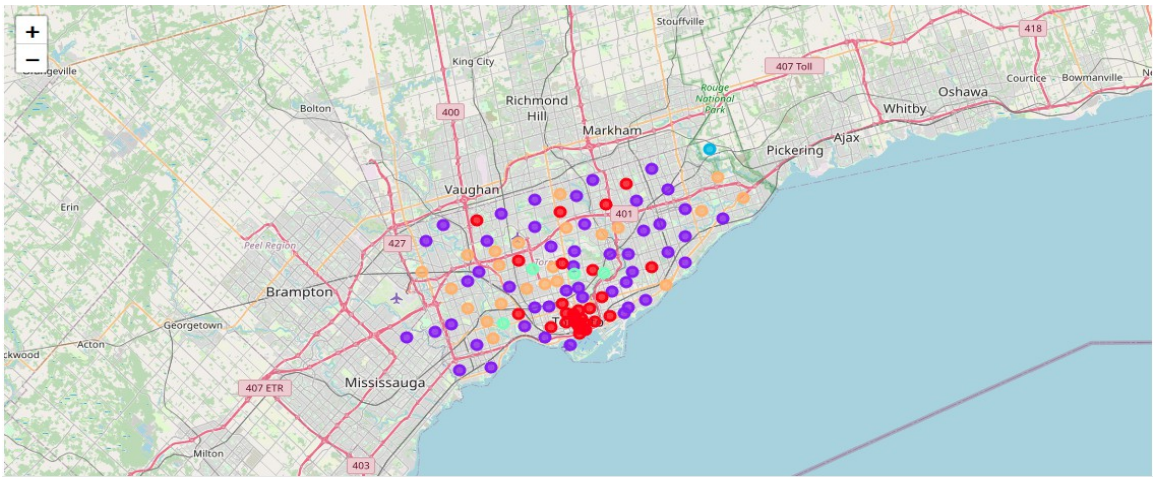
Observamos las primeras 5 filas de la tabla:

	City	Borough	Neighborhood	Latitude	Longitude	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	Toronto	North York	Parkwoods	43.753259	-79.329656	4	Business and Professional Services	Accounting and Bookkeeping Service	Community and Government	Financial Service	Engineer	Health Food Store	Park	Website Designer	Audiovisual Service	Vintage and Thrift Store
1	Toronto	North York	Victoria Village	43.725882	-79.315572	1	Car Dealership	General Contractor	Print Store	Media Agency	Organization	Automotive Repair Shop	Bookstore	Tailor	Bridal Store	Burger Joint
2	Toronto	Downtown Toronto	Regent Park , Harbourfront	43.654260	-79.360636	0	Automotive Repair Shop	Car Dealership	Park	Restaurant	Furniture and Home Store	Bakery	Music Venue	Coffee Shop	Italian Restaurant	Arts and Entertainment
3	Toronto	North York	Lawrence Manor , Lawrence Heights	43.718518	-79.464763	0	Clothing Store	Housewares Store	Carpet and Flooring Contractor	Cosmetics Store	Event Service	Loans Agency	Men's Store	Metals Supplier	Hair Salon	Gymnastics Center
4	Toronto	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	0	Café	Cafes, Coffee, and Tea Houses	Hair Salon	Bank	Fried Chicken Joint	Deli	Coffee Shop	Italian Restaurant	Diner	Organization

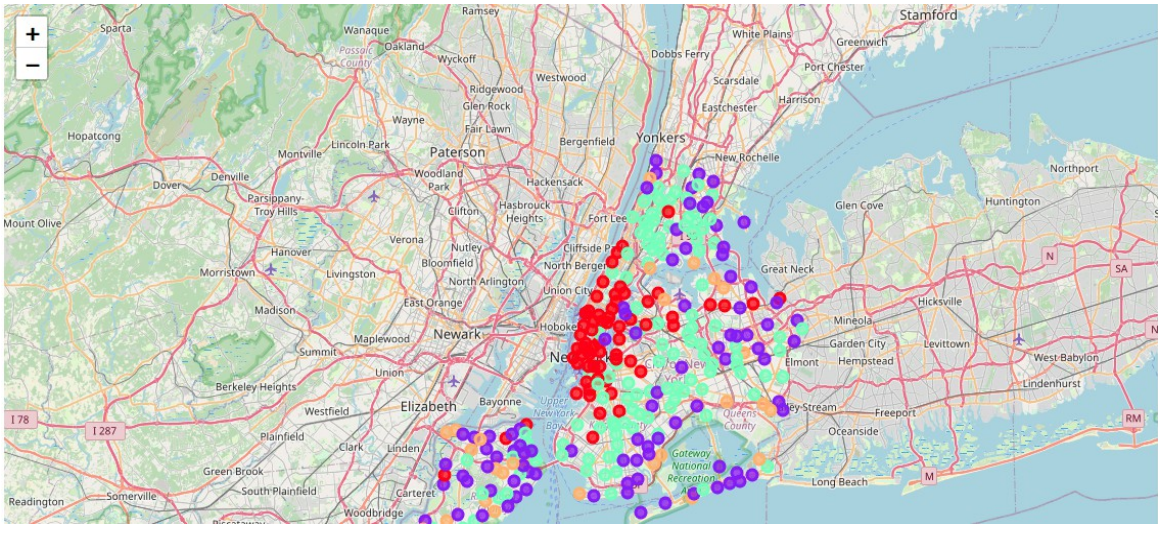
Observamos las últimas 5 filas de la tabla:

	City	Borough	Neighborhood	Latitude	Longitude	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
404	Nueva York	Manhattan	Hudson Yards	40.756658	-74.000111	0	Restaurant	American Restaurant	Coffee Shop	Cocktail Bar	Café	Grocery Store / Supermarket	Art Gallery	Furniture and Home Store	Music Venue	Pet Supplies Store
405	Nueva York	Queens	Hammels	40.587338	-73.805530	1	Beach	Playground	Used Car Dealership	Community and Government	Dog Park	Automotive Repair Shop	Real Estate Agency	Building and Land Surveyor	Surf Spot	Sports and Recreation
406	Nueva York	Queens	Bayswater	40.611322	-73.765968	4	Playground	Restaurant	General Contractor	Pest Control Service	Travel Agency	Government Department / Agency	Heating, Ventilating and Air Conditioning Cont...	Jewelry Store	Accounting and Bookkeeping Service	New American Restaurant
407	Nueva York	Queens	Queensbridge	40.756091	-73.945631	1	Deli	Print Store	Public and Social Service	Baseball Field	Playground	General Contractor	Chinese Restaurant	Grocery Store / Supermarket	Park	Barbershop
408	Nueva York	Staten Island	Fox Hills	40.617311	-74.081740	1	Grocery Store / Supermarket	ATM	Business and Professional Services	General Contractor	Organization	Electrician	Business and Strategy Consulting Office	Dining and Drinking	Drugstore	Entertainment Agency

Mapa de Toronto y los grupos de vecindarios conformados:



Mapa de Nueva York y los grupos de vecindarios conformados:



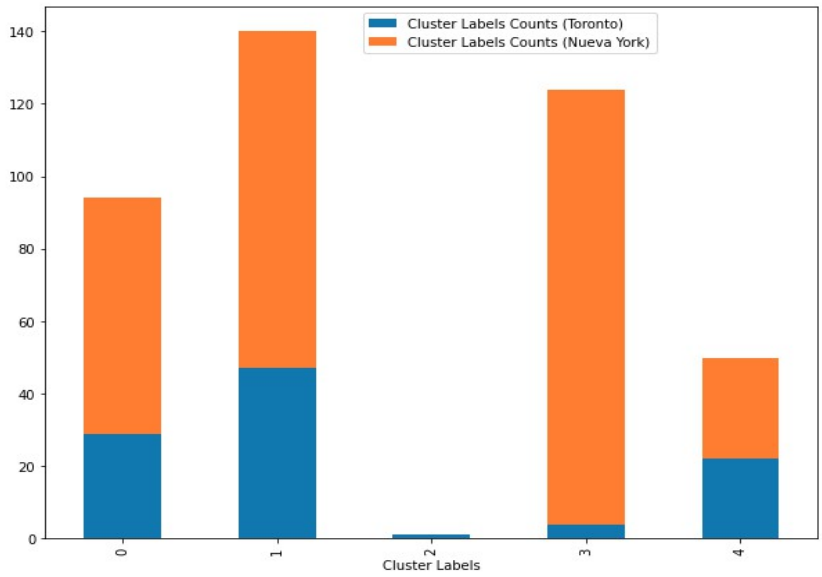
4 - Resultados

Observamos gráficamente la conformación de los distintos cluster que han quedado conformados en nuestro modelo.

Observamos la tabla para cada cluster discriminada por ciudad:

	Cluster Labels	Cluster Labels Counts (Toronto)	Cluster Labels Counts (Nueva York)	Cluster Labels Counts (Sum)
0	0	29	65	94
1	1	47	93	140
2	2	1	0	1
3	3	4	120	124
4	4	22	28	50

Visualizamos el gráfico de barras:

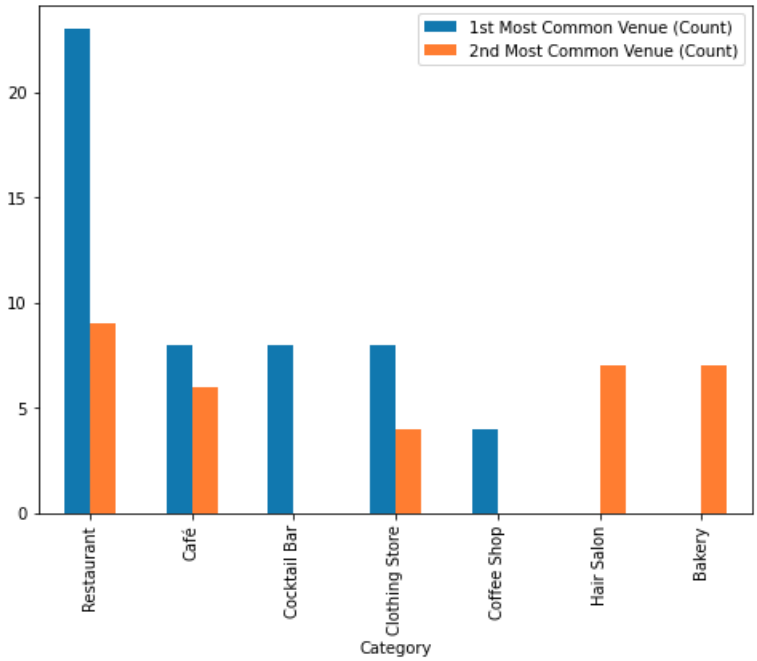


5 - Análisis

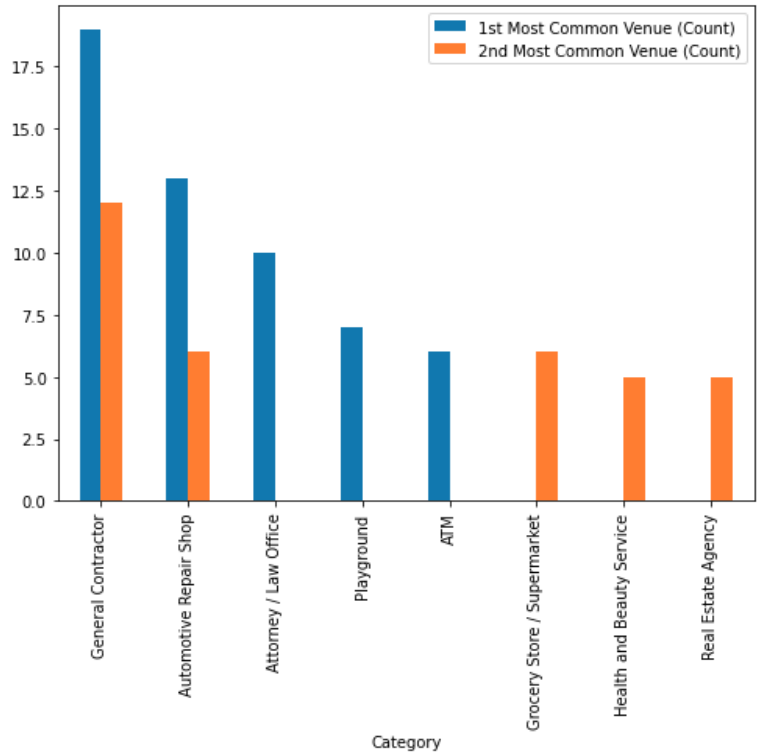
Realizamos un análisis descriptivo acerca de las características de los distintos cluster de vecindarios conformados. En este análisis nos enfocaremos en las categorías que aparecen con más frecuencias en cada cluster tomando en cuenta los dos lugares más comunes por vecindario. Así podremos observar si hay o no características que sobresalgan en cada uno de los cluster.

Visualizamos los gráficos de barra para cada cluster:

Primer agrupación (cluster 0):



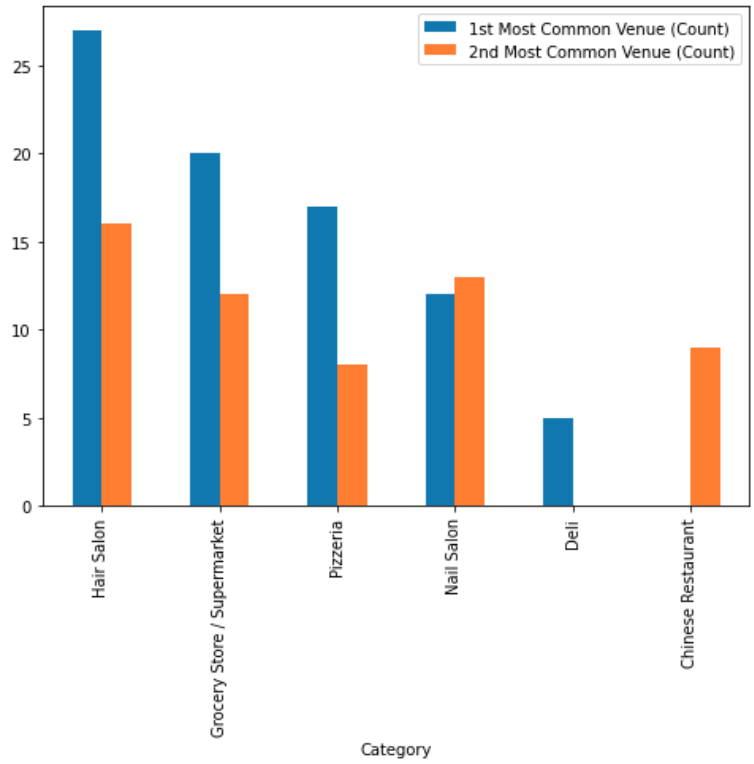
Segunda agrupación (cluster 1):



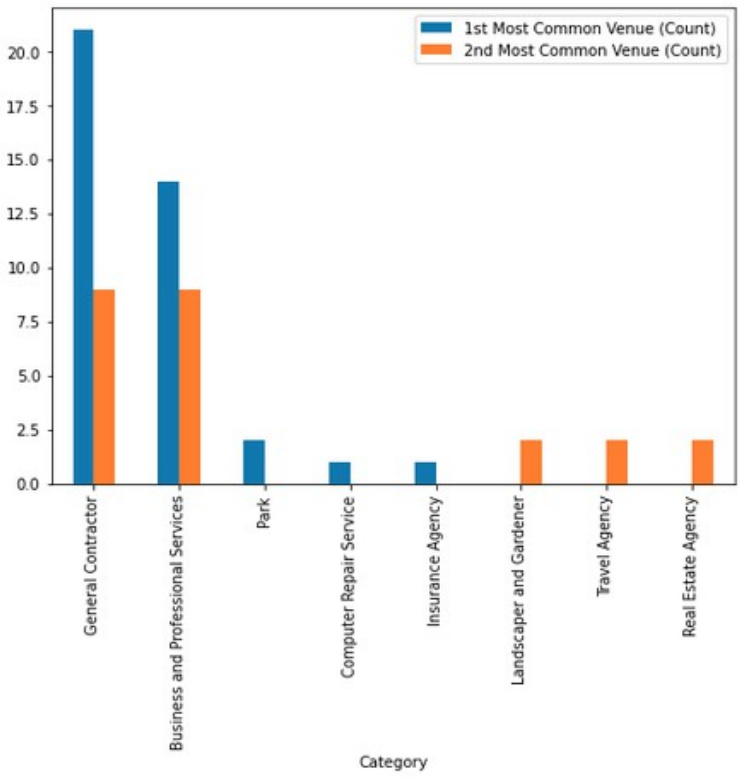
Tercer agrupación (cluster 2):

Solamente está conformado por un vecindario.

Cuarta agrupación (cluster 3):



Quinta agrupación (cluster 4):



6 - Conclusiones

Hemos analizado dos grandes ciudades capitales, las cuales han de ser ciudades multiculturales con muchas características que definen a los distintos vecindarios que las conforman, que los hacen más o menos similares o diferentes unos con los otros. Hemos podido investigar a grades rasgos qué vecindarios de una y otra ciudad tienen unas u otras características que nos permiten mediante aplicación de aprendizaje automático agruparlos de determinada forma. Como se ha observado en nuestro análisis descriptivo hay características que sobresalen del resto bastante más en unos grupos que en otros. Sin duda que este informe es apenas una puntita de toda la información que podemos recabar y que a partir de estos datos se puede seguir profundizando mucho más en el análisis.