**PROYECTO PARA CLUSTERIZAR LOS BARRIOS DE TORONTO Y NUEVA YORK**

1. **Introducción.**

Los métodos de clusterización o agrupamiento nos sirven de mucho ya que nos permiten agrupar ya sea observaciones o individuos en clases, de modo que los objetos que se encuentren en una misma clase van a tener características similares o siguen un mismo patrón a diferencia de otros objetos que pertenecen a otra clase.

Este proceso o algoritmo de clusterización nos pude ayudar para agrupar a nuestro cliente, productos entre otras cosas.

1. **Público objetivo:**

El presente trabajo busca ser una guía de ayuda para aquellas personas que estén interesadas en querer abrir una franquicia de una pizzería y busca ayudarlos a escoger el mejor lugar.

1. **Problema**

Se pretende identificar que vecindarios de Toronto son similares o comparten características con vecindarios de Nueva York para así tomar una decisión y ver en que vecindarios es más factible poner una franquicia de pizzerías.

1. **Datos a usar**

Los datos que vamos a usar son datos de vecindarios de NY y Toronto, estos datos están en un dataframe y este tiene las siguientes características

Cantidad de filas y columnas: 409 filas y 5 columnas.

Nombre de columnas:

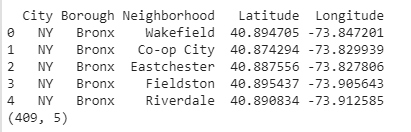
**City:** Nombre de las ciudades (Nueva York y Toronto)

**Borough:** Nombre de los distritos

**Neighborhood:** Nombre de los vecindarios

**Latitude:** Latitud del vecindario

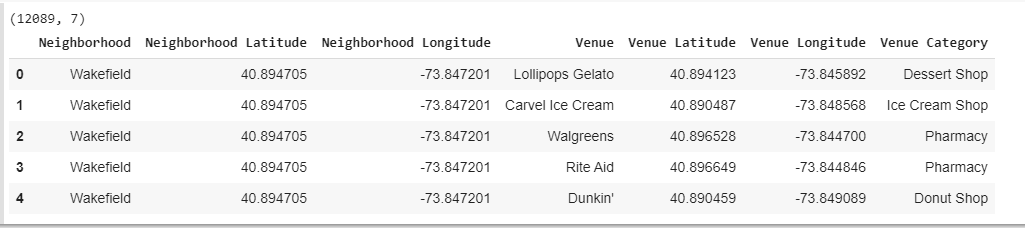
**Longitude:** Longitude del vecindario



1. También haremos uso del API de Forsquare para obtener datos de los lugares más populares que contiene cada vecindario.

Para este caso vamos a llamar a los 100 mejores lugares ubicados en un radio de 500 metros de los vecindarios, al realizar esta operación nos resulta un Data Frame con las siguientes características:

* 12089 filas y 7 columnas
* **Neighborhood**: Nombre del vecindario
* **Neighborhood Latitude**: Latitude del vecindario
* **Neighborhood Longitude**: Longitud del vecindario
* **Venue:** Lugar popular en el vecindario
* **Venue Latitude**: Latitude del lugar
* **Venue Longitude**: longitud del lugar
* **Venue Category**: Categoría del lugar



Hacemos un pequeño análisis de la columna Category

Al hacer un análisis vemos lo siguiente:

* Hay un total de 12089 filas
* 463 valores únicos de categorías
* Pizza Place es el valor que se mas repite con un total de 510

Si hacemos un top 20 de los valores más comunes en las categorías nos da lo siguiente:

|  |  |
| --- | --- |
| **Venue Category** | **Cantidad** |
| Pizza Place | 510 |
| Coffee Shop | 486 |
| Italian Restaurant | 340 |
| Deli / Bodega | 275 |
| Café | 272 |
| Bakery | 261 |
| Bar | 253 |
| Sandwich Place | 236 |
| Park | 226 |
| Chinese Restaurant | 221 |
| Grocery Store | 212 |
| Pharmacy | 196 |
| Mexican Restaurant | 187 |
| American Restaurant | 185 |
| Bank | 169 |
| Donut Shop | 169 |
| Ice Cream Shop | 158 |
| Restaurant | 151 |
| Gym | 145 |
| Sushi Restaurant | 142 |