# **Capstone Project Week 5**

## **Introduction/Business Problem**

In today's world, it became incredibly common to see new bloggers who make a living by posting different pictures of places, food, art, etc., so as the demand for new things increase tremendously, it gets more and more difficult to find novelty around for the new are youth. The most popular places have started to lose their magic as now popular is the new common. Foursquare data will be used to help this audience to provide them with the list least known, least popular places which have a good chances for improvement.

## **Data to be used**

I will be inspiring from the previous labs where we used FourSquare data and clustered neighborhoods. Execute my idea I'll need some back up with Foursquare data. I will use Geopy library so that I can find latitude and longitude of the areas. Once the locations are marked out on the map, it'll be easier to use clustering method. By clustering the restaurant and least popular locations, I will be able to provide a good enough understanding to bloggers who are looking for sharing their new experiences.

## **Methodology**

First, I’ve started scraping the table of neighbourhoods, postal codes from a website. I’ve used BeautifulSoup for that which made it easier to scrape the data. Data needed some clean up so I’ve used drop function to get rid of meaningless data.

Secondly, I’ve pulled in the latitude and longitude data, and matched it with postal codes.

The next step was to plot the map and mark the location for improved visibility.

In order to get to a place’s popularity data, I’ve connected Foursquare’s API by entering my user\_id and secret. It also helped me getting nearby venues information.

The bottom venues is defined by sorting the values in an ascending order. Later, it came to clustering the results so that more accurate estimation could be done.

## **Results**

The outcome clearly stated the least popular venues around the area as on the below table for each neighbourhood.



Once I’m done with this step, it came to explore and cluster the neighborhoods in Toronto. For this step, I’ve used k-means clustering method. K-Means can group data only unsupervised based on the similarity of area attributes to each other. K-Means is a type of partitioning clustering, that is, it divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar.

After deciding the number of cluster set (5 in this case), cluster labels generated for each row in the data frame is investigated and k means labels are printed.

## **Discussion / Conclusion**

In this study, I analyzed the relationship between the locations and their popularity among others. I identified the locations that are least commonly prefered, and offered the audience take these into consideration.I built classification models to predict how these areas would be clustered. These models can be very useful in helping bloggers in in a number of ways. For example, it could help them find out new places that not many people have been yet and promote these and increase their revenue.