**IBM Data Science Capstone Project**

**Background**

In these days, it is more and more common to have to leave your country due to a job offer abroad. One of the most difficult issues in a situation like this is to find a good place to live, and good can be interpreted as a place located in an environment as similar as possible to the current place.

Let me illustrate with an example why finding similar environments it is relevant in situations like this. Imagine a person who cannot live without coffee, as many of us do, so every day before going to work, she walks to her favourite coffee shop that is just around the corner. Another habit is to have a spicy tuna roll on Wednesday’s evening in the Japanese restaurant that is two blocks from home, lastly, jogging on Saturday’s morning in the neighbourhood’s park is a most.

Although it is possible to do some research on internet, most of the blogs and articles could consider some aspects in their analysis that are not necessarily aligned with our preferences. They can prioritize to have access to subway stations or neighbourhoods that have a vivid night life, whereas what our person used as example wants is coffee shops and Japanese restaurant options nearby, notwithstanding a rent in a range of her possibilities.

**Problem Description**

Based on what it has been discussed on above, the problem identified in the scenario when a person is moving to another country, is that it is no easy to find recommendation of neighbourhoods to live that take into account our own preferences and priorities.

**Data description**

In order to build a neighbourhood recommendation model, the following data classified in two blocks is required:

* Data gathered from the user – (The user is the person who is looking for a neighbourhood to live)
  + Rent willing to pay
  + % of variation of the rent to pay, i.e. 20% variation
  + List of top 5 venue categories ranked, i.e. 1) Coffee Shop, 2) Japanese Restaurant, etc.
  + Office where the user is being relocated. It could be the office description, address, or location
* Average amount of rent per neighbourhood
* Venues data – Gathered from Foursquare
  + List of venues and their category per neighbourhood in a radius to be specified
* Neighbourhoods geolocation data
  + Latitude and Longitude of all neighbourhoods considered
* List of the top five boroughs to live in Mexico City based on statistical information or research studies
  + Used to narrow down the set of neighbourhoods and data required associated with them

**Data usefulness**

Data gathered from the user will be used to get her preferences which are the key part and differentiator from other recommender solutions. In addition to this, the amount of rent and allowed variation will be used as a filter to only consider neighbourhoods that are within the budget range.

It is essential to have a list of the average rent per neighbourhood in order to identify which neighbourhoods are the range specified by the user, and discard neighbourhoods that are out of the budget range.

The complement of having the user’s venue categories preferences, is to have the list of venue categories per neighbourhood. With all this information, a recommender system can be developed to search for neighbourhoods that are closer to the user’s preferences.

Mexico City is a place known for its dense traffic, due to this fact the distance from home to the workplace is critical. Here is where the geolocation of each neighbourhoods comes into play, in order to calculate the distance to the workplace by using the latitude and longitude.

Last but not least, Mexico City it is also well-known for its size, this means that there are a lot of neighbourhoods dispersed in a bit less than 1,500 . Considering this, it is important to narrow down the list neighbourhoods. Hence, the data to handle. One way to do this by reviewing studies or research papers to limit the options to the safer zones or boroughs that provide a better quality of life.