**Peer-graded Assignment Capstone Project - The Battle of Neighborhoods**

***Applied Data Science Capstone by IBM/Coursera***

**Introduction: Business Problem**

For potential homeowners who are interested in buying a property, the listings on websites often do not give great details on neighborhood’s amenities (i.e. Park, school, restaurants, bus stop, etc.). A great neighborhood should also include important amenities such as grocery stores, shops, and restaurants. Most people like to frequent places that are convenient. Schools are another important amenity. Even if you don't have kids, if you want to sell your home in the future, many buyers will be on the lookout for good schools. The quality of local schools and the distance from the house are both important factors to consider.

Access to parks, recreation, shopping, restaurants and coffee shops is another key factor to a great

neighborhood. When the potential owner try to move to a new neighborhood, these are the neighborhood amenities to look into before purchasing a home. In this project we will try to utilize Foursquare's location data to explore a neighbourhood's geographical location of multiple addresses (i.e. listing available on real estate websites), as well as using KNN clustering, return the user the best or clusters of similar neighborhood within the selected city according to a set of priorities that the user rates.

We will define the acceptable driving distance to be within 1.5 km (walking distance)

**Data:**

Based on definition of our problem, the following factors we must consider:

* user's input of selected addresses
* type and number of venues in the surrounding area of selected address

It is better to use regularly spaced grid of locations, centered around city center, to set as a starting point to scrape the amenities in the selected neighborhood.

Following data sources will be needed to extract/generate the required information:

center of selected cities will be generated algorithmically and approximate addresses of centers of those areas will be obtained using **Google Maps API reverse geocoding**

type and number of amenities and their type and location in every neighborhood will be obtained using **Foursquare API**

**Methodology**

Assuming that the user is interested in the following 7 properties listed on realtor.ca, the user would like to know, within walking distance of ~1.5 km, what are the venues nearby.

**Neighborhood Candidates**

For this project, we would need to select centroids of latitude & longitude coordinates to scan nearby areas. The method will be creating grids of cells within 1.5 km of our selected addresses, which is approximate 3 x 3 km centered around the selected address.

We will load the addresses into a dataframe:

Table

Description automatically generated

I have created a grid of area candidates, equally spaced, centered around the given property address and within ~0.2143km from the center. Each grid "bubble" surrounding the address will be defined as circular areas with a radius of 0.75 km (within 1.5 km, there for the radius is half of that), so grid bubbles' center will be 0.4286 km apart.

The very next step is to create a **hexagonal grid of cells**: we offset every other row, and adjust vertical row spacing so that **every cell center is equally distant from all it's neighbors**.

Map

Description automatically generated

**Foursquare**

Next, we are going to start utilizing the Foursquare API to explore the addresses and segment them.

Lets look at the dataframe to see how many venues was returned

Table

Description automatically generated

so, in total we have 167 venues from all grid bubbles lets find out all unique venue categories. Lets check how many venues were returned from all of these addresses:

Table

Description automatically generated

**Analyze each Property's surrounding area**

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category Lets print each property's along with the top 7 most common venues

Table

Description automatically generated

**Cluster Neighborhoods**

Run *k*-means to cluster the neighborhood into n clusters and visualize them on a map:

Map

Description automatically generated

**Examine Clusters**

Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster.

Table

Description automatically generated

**Results and Discussion**

Our analysis shows that for each address, the surrounding bubbles have different categories of venues. Each have a very unique blend of amenities and hence, the criteria for selecting the "right" property would depend directly on the personal preference of the buyer. Note that here, in this exercise, we have not included crime incidents, pricing, square footages as well as neighborhood density. To further this project, we can gather more data to support the decision of the buyer.

**Conclusion**

For myself, if I am the buyer, assuming all of these properties are similar in price, housing sizes, I would prefer to have a area closer to a park for my kids, close to restaurants, as well as having a gym nearby for exercising. It seems that Property index 301 is the ideal place.