Introduction

Moving can be a stressful and tedious task. Packing & unpacking, arranging a moving company, finding the right time, cost etc. are all part of this challenging task. However, these hurdles to overcome do not have long lasting effects. For many people, the decision of "where" to move is the hardest one among all as it will have an enormous effect on their well-being during their stay in the place they moved in. One way to ease their pain on making this decision would be to cluster neighborhoods based on their similarity to each other and select from the ones that are similar to your place or to a place that you consider to be a decent place. This way, a very long list of places would significantly be narrowed down.

Business Problem & Target Audience

Say that you are a freelancer that helps customers to make a decision on where to move. In this particular case, you have a customer who currently lives in a neighborhood in Manhattan, New York and wants to move to a neighborhood in Toronto, Canada as he found a job there. The customer is very happy about his current place and wants to find a neighborhood in Toronto that is similar to his current place in terms of activities he can enjoy such as restaurants, cafes, museums etc. He asks our help to narrow down a list of potential places to move in. We will make clusters of neighborhoods including his current neighborhood and neighborhoods in Toronto. And he will select from places that are in the same cluster as his current place.

Data & Source

The following data will be required for this problem:

- List of neighborhoods in Toronto
- Coordinates of those neighborhoods as well as coordinates of his current neighborhood
- Top venue data of each neighborhood for clustering

The list of neighborhoods in Toronto will be extracted from the following Wikipedia page:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The data will be transformed into pandas data frame and will be processed into a proper version (data cleaning, wrangling etc.)

In order to get the venue data, we will use Foursquare API based on the coordinates of the neighborhoods, select the top venues, and make a few transformations on the data to be used for clustering.

Methodology

Since the customer currently lives in Chinatown, Manhattan and wants to move to Toronto, Canada. We first need some data related to Toronto. The data is extracted from the following link:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Since the data is in a raw format, it needs to be transferred into panda's data frame and be cleaned to be analyzed and compared with the customer's current address. The following steps were followed to that end (They can be clearly seen on the notebook as well):

- Transform the wikipedia table into panda dataframe
- Get rid of 'Not assigned' Boroughs
- Merge Neighbourhoods with same postcodes and seperate them with ","
- If a neighbourhood has "Not assigned" value, assign the borough name
- Load postal codes, turn it into panda dataframe, change the column name for the next step
- Add latitude and longitude to the df dataframe
- We will examine boroughs that has Toronto in them, so let's create a dataframe for that
- Now let's add the customer's neighbourhood in Newyork to this dataframe
- Define Foursquare Credentials and Version
- Let's explore the first neighborhood in our dataframe
- Now, let's get the top 100 venues that are in The Beaches within a radius of 500 meters
- Let's create GET request url and send the GET request and examine the results
- Let's create a function to repeat the same process to all the neighborhoods in Toronto and our customer's place in Newyork
- We'll check how many venues have been returned for each neighbourhood
- Analyze each neighbourhood
- Next, let's group rows by neighbourhood and by taking the mean of the frequency of occurrence of each category

- Now let's create the new dataframe and display the top 10 venues for each neighborhood
- Neighbourhood clustering
- Let's create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.
- Examine Clusters

Since this is a clustering problem **K-Means Clustering** algorithm has been used

Results & Discussion

After the cleaning of the data, Toronto was divided into 39 neighborhoods. And these neighborhoods were compared to each other and the customer's neighborhood in Manhattan. Since this is a clustering issue, K-Means Clustering algorithm was used setting the cluster number to 5. Therefore, 40 neighborhoods were separated into 5 clusters. When examining those clusters, it was found that customer's place falls into cluster number 3 along with other 34 neighborhoods out of a total of 40 neighborhoods. This means that many places in Toronto are very similar to his taste. This is very good news for the customer because he can go for the cheapest option since almost all neighborhoods in Toronto is very similar to his taste.

Conclusion

In this analysis, the customer's current place is compared to neighborhoods in Toronto using a clustering method. It was found that many neighborhoods in Toronto is similar to his current neighborhood, meaning that customer is at advantage because he can go for the cheaper houses as he does not need to worry too much about whether the neighborhood will be appropriate for him or not because almost all of them will be in a neighborhood that will be similar to his taste based on the clustering analysis.