**Report – Battle of Neighbourhoods**

**Introduction – Business Problem:**

In my project, I assume to represent Relocation Specialist team who specializes in showing accommodation options to Indian families relocating to an office in New York. Requirements of each family differs and it is difficult to keep track of amenities available in each area unless you are an experienced professional. New recruits find it hard to handle the clients and show them the desired options hence losing the confidence of the client and inturn of the agency. I would like to develop a package which can show the amenities of an area given the location which helps the relocation specialists.

For example, I would take the case of an Indian family with a young kid who are looking for reasonable options. The criteria of the family is as below:

1. Reasonable accommodation costs

2. Proximity to Indian families to feel at home

Research has been done and since New York is a high cost destination New Jersey has been chosen to be a reasonable place. It also shows that most of the Indian families live in Essex County in New Jersey State.

So the solution is to find an area in Essex county which fits the criteria mentioned below by using clustering approach which the family thinks adds value to their life style choices and increases their chances of a happy and healthy life.

4. Nearby parks/green areas for the kid to play and enjoy

5. Availability of fitness centres to keep the young adults fit

6. Availability of Sport centres for the kid to learn extra curricular activities

7. Easy and nearby entertainment options like theatres etc

8. Restaurant, Bank facilities nearby

**Data:**

For our analysis we need an overall idea of New Jersey state in general and Essex county in particular.

Use 'https://www.zipcodestogo.com/New%20Jersey/' link to get the postal codes of counties and cities.

Then use "Beautiful soup" to retrieve the postal codes, county and city information from the website.

Geo co-ordinates: The geo location of the city-county should be obtained from Geocoder.

Nearby Venue locations: The venues for each city-county should be obtained from FourSquare.

**Methodology:**

Perform exploratory analysis on data to check if it suits the requirements.

Data with Null values should be removed.

Data should not contain duplicate values.

Data should be formatted correctly and have the correct domain.

Once the data is standardized the qualified data is ready to be processed.

While using Geocoder, it was found that for some of the places the latitudes and longitudes could not be retrieved. The data has been massaged to remove those city-county rows as it doesn’t add value.

While using FourSquare, it was found that for some of the places the nearby venue locations could not be retrieved. The data has been massaged to remove those city-county rows as it doesn’t add value

Then cluster the nearby venues to find out the top ten amenities available for each location and extract those features. Based on the amenities cluster the neighbourhoods which have similar characteristics. These venues can be shown to the clients to choose a location.

Advantage is that given a location, we get an overview of the facilities available and it gives us the knowledge to choose the best options.

**Results**

Based on the amenities available in each area, the city-county combinations are divided into 5 clusters.

Each cluster has a unique amenity.

For example Cluster 1 is the grouping of most happening places consisting of restaurants, malls, banks & entertainment centers

Cluster 2 is the group which contains a natural setting of a lake and some food joints

Cluster 3 & 5 is mostly about food joints and super markets

Cluster 4 is the group which has a historic site, food joints and convenience store

**Discussion**

Though the data has gone through exploratory analysis, some the issues can be found during actual run of data. For ex: Sussex County has passed through exploratory analysis but could not find a geo location and hence the need to delete the data at a later stage of the process Similar is the case with “Short Hills” area where FourSquare could not find the nearby venues

Hence I see that the Data Science methodology is a highly iterative process which needs going back and forth to tune the data as needed.

We could also find the relevant data which can be shown to the Indian family with a kid as possible accommodation options.

**Conclusion**

From the data, we find that Montclair in Essex county is the most suitable option for the family given the below facilities

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Park | Chinese Restaurant | Event Space | Athletics & Sports | Train Station | Bookstore | Theater | Café | Bank | French Restaurant |