Applied Data Science Capstone - Coursera course

Introduction - final project

Introduction / Problem

The idea of the project is to determine the optimal neighborhood to pick a hotel in, or similar, for a family trip to one of Canada's 5 largest cities. The cities and their corresponding neighborhoods are Toronto, Montreal, Vancouver, Calgary and Edmonton. The important underlying assumption are, that the family has planned trips to site seeing spots during the day, but in the evenings and mornings, everyone wants to pursue their individual interests. Further, the family is indifferent between the cities.

Therefore, the goal is to find the neighborhood that best meets the interests and demands of the made-up, stereotypical family, consisting of the parents, a son and a daughter. The evaluation of the neighborhoods is based on a scoring system, in which, to allow the parents an as much as possible stress-free trip, the kids' priorities have higher weights.

It is assumed that each family member has a list of 5 venues that they would like to visit with the least time spent travelling from one neighborhood to another, in order to better be able to visit more venues. Thus, the ideal neighborhood includes as many of the preferred venues as possible. The venues have been chosen from the list of available ones in the neighborhoods of the five cities.

Parents	**Mom**	**Dad**
40 pts	Yoga Studio	Golf Course
30	Café	Bagel Shop
20	Museum	Karaoke Bar
10	Jewelry Store	Wine Bar
5	Shoe Store	Steakhouse

Kids	**Son**	**Daughter**
50 pts	Hockey Arena	Recreation Center
40	Nightclub	Accessories Store
30	Bar	Movie Theater
20	Comedy Club	Shopping Mall
10	Pizza Place	Gym

The code resulting from this project will produce two major results. First, it will recommend the neighborhood and city to which a trip will lead to the most overall points, thus the greatest common benefit. Second, it will recommend a second option that maximizes the outcome of the most dissatisfied family member. Hence, the second option maximizes the minimum points of the individual family members, whereas the first option maximizes the overall sum.

Data used for the project

The data for the project is obtained from three sources. First, the postal codes and names of the neighborhoods in the five cities are scraped from Wikipedia. These codes are then matched, by postal code, with the latitude and longitude geospatial data obtained from geonames.org. Lastly, the venue data is accessed through the foursquare API.

Combining the information included in the data, it is possible to rank the neighborhoods according to the order in which they meet the preferences outlined above.