Capstone Coursera Applied Data Science Final Project - Problem Submission

Background

Given that we now live in a global village, people are highly mobile and relocate extremely easily given the potential for career advancement, better prospects for themselves or family, crises in their country, war, genocide etc.

Large multi-national companies also need talented individuals to be mobile in order to satisfy resource requirements for various projects or work situations.

The overwhelming question that comes to mind of the person that is relocating is, in which area of the major location they are moving to, do they settle.

Problem Statement

This question being posed on numerous occasions to the Human Resources division of a major global company prompted them to embark on trying to solve it by examining the feasibility of building an application to model the following:

- What are the crime profiles of the various neighbourhoods surrounding their company head offices in the different locations globally within a given acceptable travel radius?
- Where acceptable levels of crime are found in the said neighbourhoods, what facilities are available to their prospective employees in a given radius?

The scenario posed was that a person, either an existing employee or a new recruit, would need to relocate to one of the company's offices in another city. The system should profile the crime levels in the local neighbourhoods of the target city within the given travel radius of the office and present the neighbourhoods with the lowest crime levels.

Once the potential neighbourhood for relocation is selected the system should then model the areas facilities, venues and amenities and present the most acceptable areas within the profile.

This would be an iterative process based on user choice, until the acceptable neighbourhood is shortlisted. The user could then examine the suitable housing choices available for relocation.

Data Requirements

Given that Crime is a major factor in neighbourhood choice, data that indicates crime levels per areas chosen will be of paramount importance. The crime statistics need to be representative of the areas and dated as current as possible so as to represent the situation at the time reported.

The crime will need to have location data associated with it and be reported in such a way as to be at a granular level yet able to be rolled up into crime categories.

The facilities and amenities that are in the various locations need to be representative of the categories such as Schools, College & University, Shops, Restaurants, Malls, Health and Fitness facilities, Sports Stadiums etc.

Postal or Zip code data will be needed to map the neighbourhoods to their location coordinates.

Data Sources

The immediate sources envisaged for the data would be the internet as that data is digital enabling easy manipulation. Crime data should be available from the various government law enforcement agencies' sites.

The postal services and other similar web sites would provide the Zip / Postal code and area mappings. Various other sources are also available to provide coordinate mappings.

Venue and facilities data would be gleaned from sources such as Foursquare, Google etc.

These sources would need to be examine for suitability and type of data source. A mixture of web scaping, RESTful API calls and source files from websites (Excel or CSV) would be used as available.

Some manual data population may be required as well as required.

Process to be followed

A suitable Head Office location of the company will be chosen as a starting point for the model.

The first phase of the model would be to profile crime levels in the various location areas / neighbourhoods and once a suitable one is selected then profile the associated amenities, venues and facilities for that area.

The model will be built, tested and evaluated. Once management are satisfied that the results are providing the correct information the model will be updated to include the next sets of office locations based upon priority in order to leverage maximum results.

Additional functionality could also be added in future releases e.g. Housing data.