Applied Data Science Capstone

April 8, 2020

Data

This section provides a description of the data and how it will be used to solve the problem.

Data Sources

The Foursquare API will be used to obtain location data for acute care hospitals, using search_query = 'Hospital', that are within 10 miles (16.09 kilometers) from the Space Needle. One of the anticipated challenges with this search is that Foursquare will likely return a variety of facilities where the facility name includes 'Hospital', but the facility does not treat humans, e.g. a veterinarian hospital facility.

The Nominatim package from GeoPy will be used to obtain the latitude and longitude for the Space Needle.

In addition, hospital bed information will be obtained from two sources:
https://en.wikipedia.org/wiki/List_of_hospitals_in_Washington_(state)

https://www.wsha.org/our-members/member-listing/

Since the number of beds varies across hospitals, an attempt will be made to see the extent to which the hospitals returned by the API search will cluster based on number of beds. For some people, deciding which nearby hospital to use may depend on the size of the hospital, where larger hospitals tend to provide more advanced services.

Data Preparation

The Foursquare venue search will focus on three of the four response fields (name, location, and categories):

Field	Relevant Content
id	n/a
name	Hospital name
location	Address, latitude, longitude
categories	Type of facility

The results returned from the Foursquare search will be transformed into Pandas data frame and filtered for relevant columns. It is anticipated that some data cleanup will be required to address issues such as missing values, duplicates, etc. Furthermore, the 'categories' response field will be parsed to extract the value for 'name', which indicates the type of facility, e.g., 'Hospital', 'Veterinarian', 'Doctor's Office', etc. This should not be confused with the 'name' response field in the above table.

The hospital bed data will be merged with the Foursquare data frame, after which data analysis can proceed.

These data will be used to:

- Calculate the straight-line distance from the Space Needle to each hospital using the geodesic function.
- Perform a cluster analysis using the k-means unsupervised machine learning algorithm to see if the selected hospitals cluster based on number of beds.