Capstone Project- Abstract

1 Introduction

1.1 Scenario and Introduction:

I will be graduating during the week of May 21 from New York University. I am inviting my family to spend a week in New York City to attend the graduation ceremony and spend some time travelling and exploring the city. My family likes to explore cultural sites within the city. I intend to use location-based data and data science techniques to select the best possible hotel for my duration that addresses their travelling and comfort needs.

1.2 Problem Definition:

The hotel selected should have the following characteristics

- The total cost of stay for a week should be less than the budget of USD 1000
- The hotel should at least have a rating of 3 stars
- The hotel should be in a neighborhood located close to a cluster of culturally significant tourist attractions like museums, art galleries, auditoriums etc.
- The hotel should be located within walking distance to a subway station
- The hotel should not be more than 1 hour away from the closest airport
- The hotel should be in a neighborhood with a high proportion of Indian restaurants and halal food options

2 Data

2.1 Data Requirements and Sources:

There are numerous types of data that would be required for this project:

- List of neighborhoods and their geographical coordinates
    - Potential Sources: New York University (Catalog of NYC Neighborhoods)
- Location data of Subway stations in New York City
    - Potential Sources: MTA, Wikipedia etc.
- Information on venues in Neighborhoods and category information
    - Potential Sources: Foursquare API
- Hotel information including price, amenities, and location information
    - Potential Source: Hotel Aggregating sites like Trivago, Google etc.

2.2 Data Pre-Processing:

- List of neighborhoods was obtained from the NYU catalog and the same file used in the course lab with the requisite latitude and longitude information was used
- The list of subway stations was obtained from developer data from MTA
- The hotels data was generated by going to Hotels.com, entering the criteria for the dates, the amenities, and the start-rating of hotels under consideration. The BeautifulSoup package was then used in conjunction with the selenium and chromedriver packages to scrape the data as the page was had an infinite scrolling layout which was difficult to scrape from the IBM Watson studio environment due to difficulty installing the local packages. The code was then run on my local machine and the output csv file is used as the starting point in this code base.
- Location data for hotels and neighborhoods was obtained using the nominatim package