Data Science Capstone Report

JMc

1. **Introduction / General Business Problem**

General quality of life is a factor in recruiting and keeping talent for technology innovation operations. For example, Amazon recently went through a process to select a location for a secondary headquarters location from it’s original headquarters in Seattle. Selection of location on the east coast would have presumably added a diversity to the Amazon geographic base and provided a desired distribution of presence. Intuitively, large cities in the east would provide the best opportunity for Amazon to recruit a large base of the best and brightest to live and work. Large cities, however, have undeniably higher costs of living. It would be ideal for companies like Amazon to be able to locate in smaller cities that have adequate stock of University academic talent but may not have the amenities of the larger and more expensive cities. The question is, “Are there neighborhoods in smaller cities that have the quality of living profile (based on amenities) of the neighborhoods in the larger cities?” For the purposes of assessment, we will compare the amenity profiles of Toronto and New York with a smaller city such as Pittsburgh. The working hypothesis is that Toronto and New York have adequate food, drink, entertainment and Recreation to support a quality of life that is attractive. Do these amenities exist in Pittsburgh so that locating there would provide an attractive life/work location at a lower cost? We will use Foursquare location data from all three cities to analyze neighborhood data to assess the amenity profiles. This report is useful to any companies checking out locations (in this case Pittsburgh) to locate in when attracting talent has an element of cost and quality of life tradeoff.

1. **Data Sources and Use**

I will be using the analysis of New York and Toronto previously performed in week 3 of the course as reference artifacts [1]. I will be performing a similar analysis of Pittsburgh. I will be gathering information about Pittsburgh from a Wikipedia site that lists the neighborhoods of Pittsburgh [2]. I will use the listed location/position coordinates listed on the pages referenced. Where geo data available about the neighborhood is unavailable or suspect on the neighborhood Wikipedia page, I will use the Google Maps Geocoder RESTful API [3] to resolve any missing or questionable data. Ultimately, I will wrangle the Pittsburgh neighborhood and Geocoding data into a .csv file that will be used as the basis for querying Foursquare [4] venue data about Pittsburgh. I will perform an neighborhood classification analysis of Pittsburgh similar to what was done in course week 3 (classification of similarity through clustering by unsupervised machine learning) for Manhattan and Toronto. Armed with three individual analyses of the three cities, I will perform an analysis of all three cities as a whole to see how Pittsburgh stacks up to the larger cities from a similarity of amenities perspective.

**References**

1. Coursera Data Science Capstone, Week 3

<https://www.coursera.org/learn/applied-data-science-capstone/home/week/3>

1. List of Pittsburgh Neighborhoods

<https://en.wikipedia.org/wiki/List_of_Pittsburgh_neighborhoods>

1. Google Geocoder API

<https://developers.google.com/maps/documentation/geocoding/start>

1. Foursquare

<https://developer.foursquare.com/>