# Impact of Neighborhood Similarity on Amazon HQ2 selection

--***Will good neighborhood attract new business office for Unicorn companies***

Jun Sun

May 05, 2019

## 1. Introduction

### 1.1 Background

Amazon announced the process to look for the 2nd headquarter in 2017, it led to a hot wave of biding on this proposal from 54 states, provinces, districts, and territories due to the huge financial and job potentials. 200 cities entered the final list. Among the 20 cities in the final lists, Long Island, New York and Crystal city, Virginia were selected as the locations for 2nd headquarters in November, 2018. Due to objection from various political parties, Amazon has canceled the selection Long Island NY in early 2019 while the development at Crystal city VA is still undergoing.

### 1.2 Problem

While Amazon has laid down the requirement for the 2HQ selection, such as Metropolitan areas with certain populations, close to popular center and highway/airport, availability of talents, financial incentives, etc., it will be interesting to check if the neighborhood of candidate locations/cities is an important criterion. For example, does the 2HQ have similar neighborhood as that in current HQ in Seattle, WA?

### 1.3 Interest

If the neighborhood similarity plays significant role in 2HQ selection, it will provide enough information for cities/territories authorities to set a strategic approach to attract new businesses in future.

## 2. Data acquisition and cleaning

### 2.1 Data sources

The current Amazon HQ is at South Lake Union at Seattle. The Long Island City in New York city belongs to Queens borough but Manhattan borough is also included in this study due to the close location between Long Island City and Manhattan. The Crystal City in VA belongs to Arlington borough but Washington

DC is also included due to the similar reason of close distance between Crystal City and Washington DC.

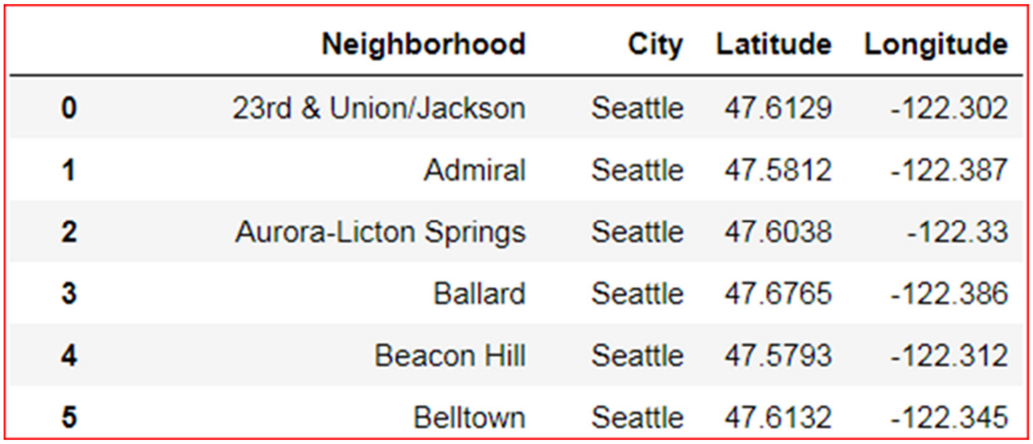
The list of neighborhoods for Seattle is obtained from Seattle gov website (4) and latitude/longitude data for each neighborhood is obtained using the Nominatim package from geopy library. Same approach was applied for Arlington neighborhood with the neighborhood data from Wiki webpage.

For Washington DC and New York, it is relatively easy to get the neighborhood name and geospatial data from government open data sources (2, 3).

## 2.2 Data cleaning/preparation

After getting the data from the different dada sources, the data is consolidated into one data table containing columns as “Neighborhood”, “City”, “Latitude” and “Longitude”. Some neighborhood names of Seattle city need to parse to get the right name when multiple names are stacked together. The neighborhoods with same latitude/longitude are combined. The neighborhoods with no geospatial data are dropped.

The data for Queens/Manhattan borough is combined and categorized under “New York” city column. Same strategy applied for Arlington and Washington DC data and are categorized under “Arlington/DC” city column. The example of final dataset is shown in Figure 1. Totally 341 unique neighborhood data is collected.

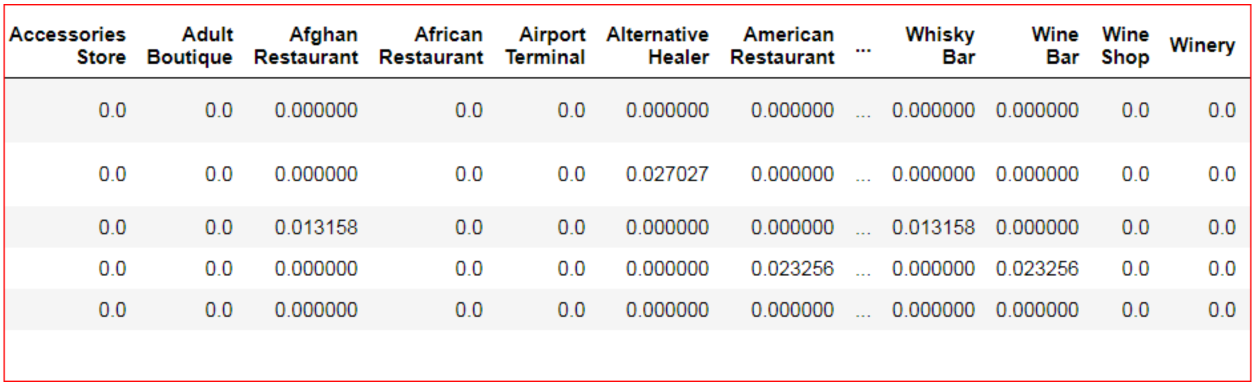


| | Neighborhood | City | Latitude | Longitude |
|---|---|---|---|---|
| **0** | 23rd & Union/Jackson | Seattle | 47.6129 | -122.302 |
| **1** | Admiral | Seattle | 47.5812 | -122.387 |
| **2** | Aurora-Licton Springs | Seattle | 47.6038 | -122.33 |
| **3** | Ballard | Seattle | 47.6765 | -122.386 |
| **4** | Beacon Hill | Seattle | 47.5793 | -122.312 |
| **5** | Belltown | Seattle | 47.6132 | -122.345 |

*Figure 1. Example of cleaned dataset for neighborhood*

## 2.3 Feature selection

The next step is to obtain top 100 venues within 500m of radius for each neighborhood by calling the Foursquare API with the geospatial data for all the neighborhoods as prepared previously. 10741 venues data points are collected and converted to category dataset with 455 categories and mean for each category is calculated for clustering. The final feature dataset is as in Figure 2.



| Accessories Store | Adult Boutique | Afghan Restaurant | African Restaurant | Airport Terminal | Alternative Healer | American Restaurant | ... | Whisky Bar | Wine Bar | Wine Shop | Winery |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.027027 | 0.000000 | ... | 0.000000 | 0.000000 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.013158 | 0.0 | 0.0 | 0.000000 | 0.000000 | ... | 0.013158 | 0.000000 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.023256 | ... | 0.000000 | 0.023256 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.0 | 0.0 |

*Figure 2 Example of feature dataset for clustering*

3. Data analysis with K-means cluster
   K-means clustering model is chosen to cluster the neighborhoods for all 3 cities.

3.1 Determine the optimal cluster for K-Means clustering

Elbow method and Silhouette score method are used to calculate the optimal k cluster. However, as shown in Figure 3, k cluster number cannot be identified by sum of squared distance, but the Silhouette score indicates k=15 is reasonable cluster number for further clustering.

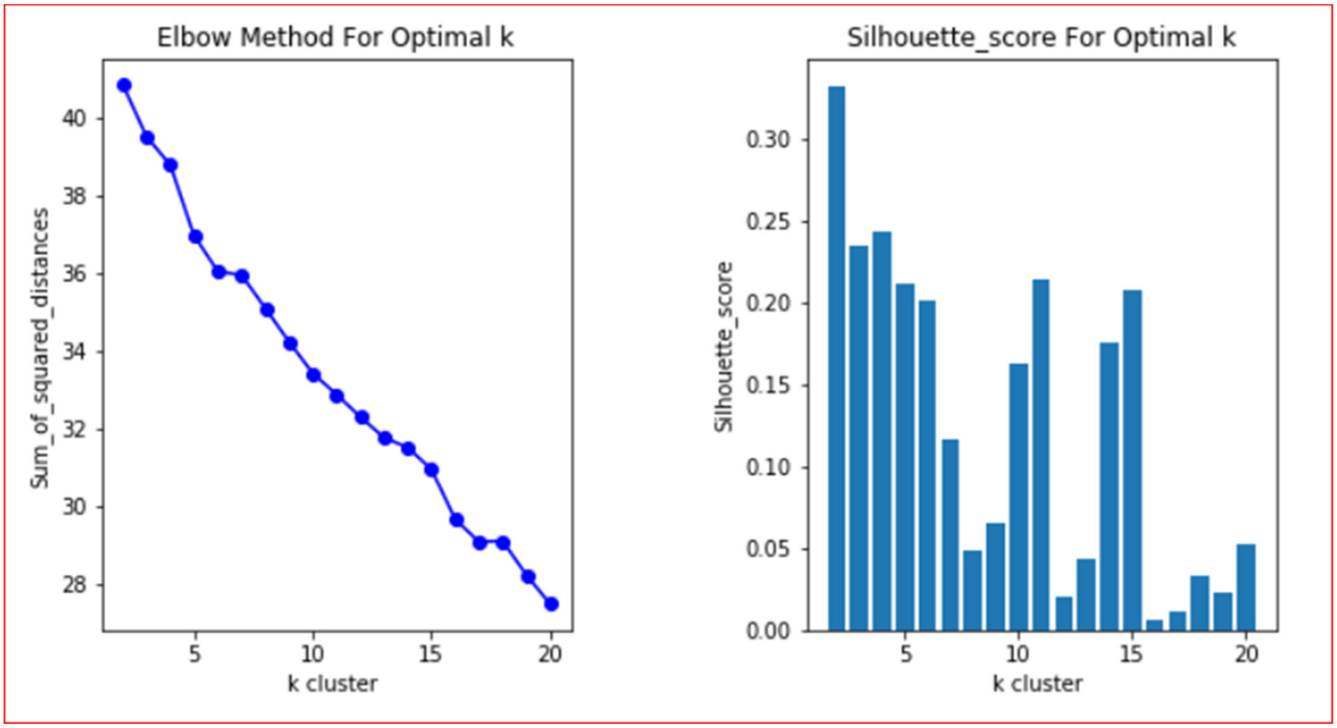
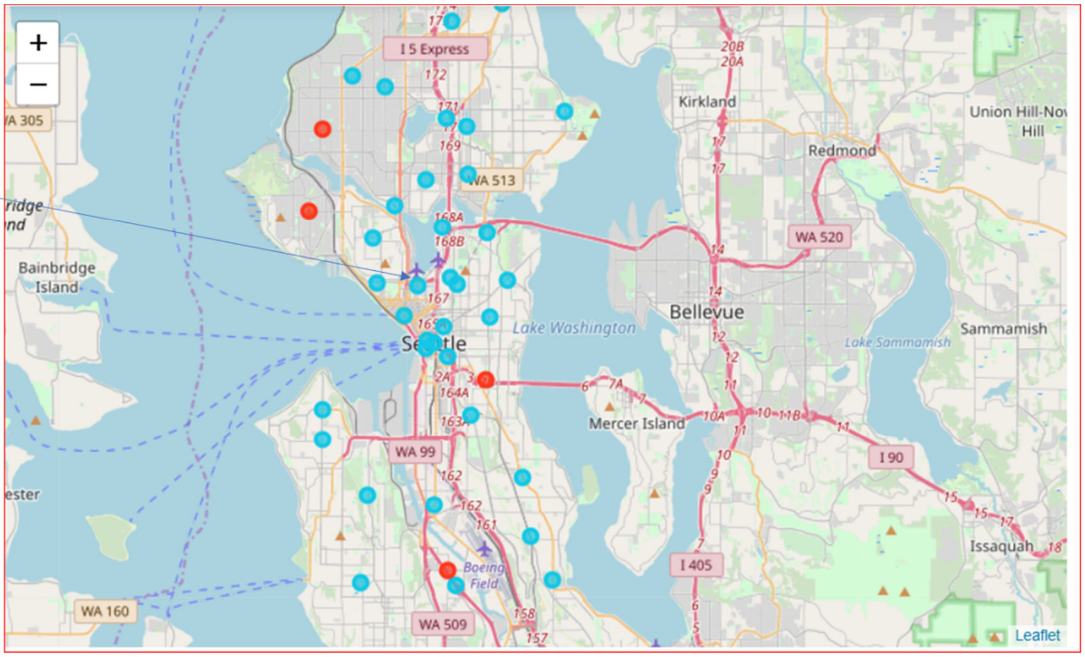


*Figure 3 Elbow and Silhouette score methods to determine optimal k cluster*

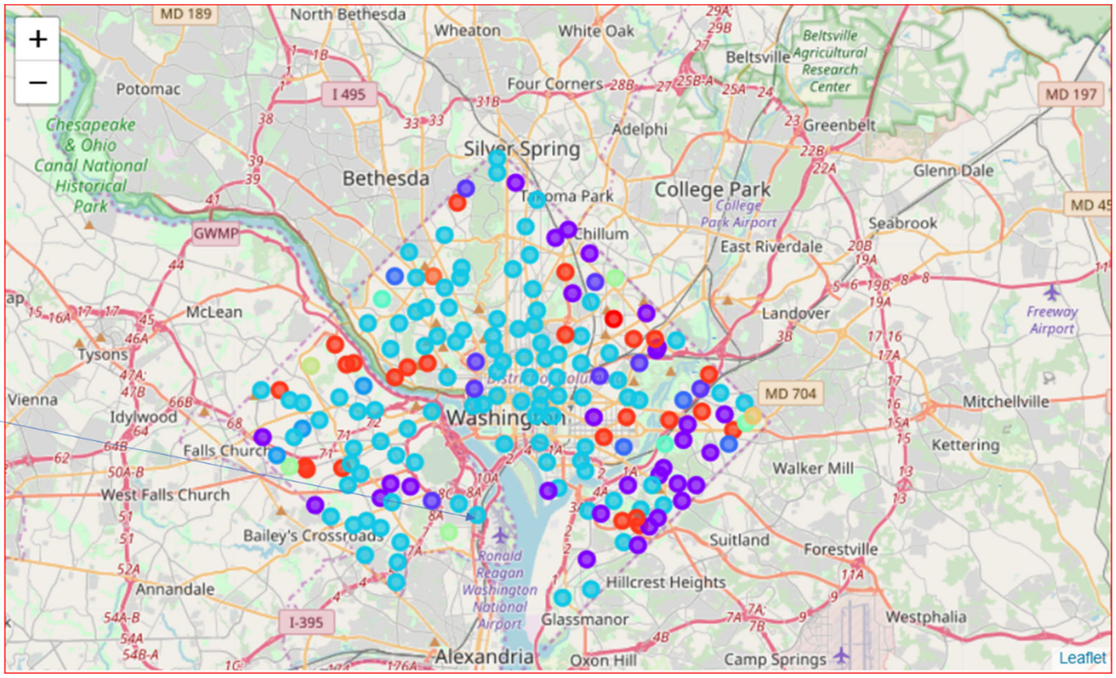
3.2 K-Means clustering on neighborhoods for 3 Amazon HQs cities

All the neighborhoods are clustered into 15 clusters with K-means algorithm and the cluster labels are merged with neighborhood geospatial dataset. Then clustered neighborhoods are plotted into city maps as shown in Figure 4, Figure 5 and Figure 6.



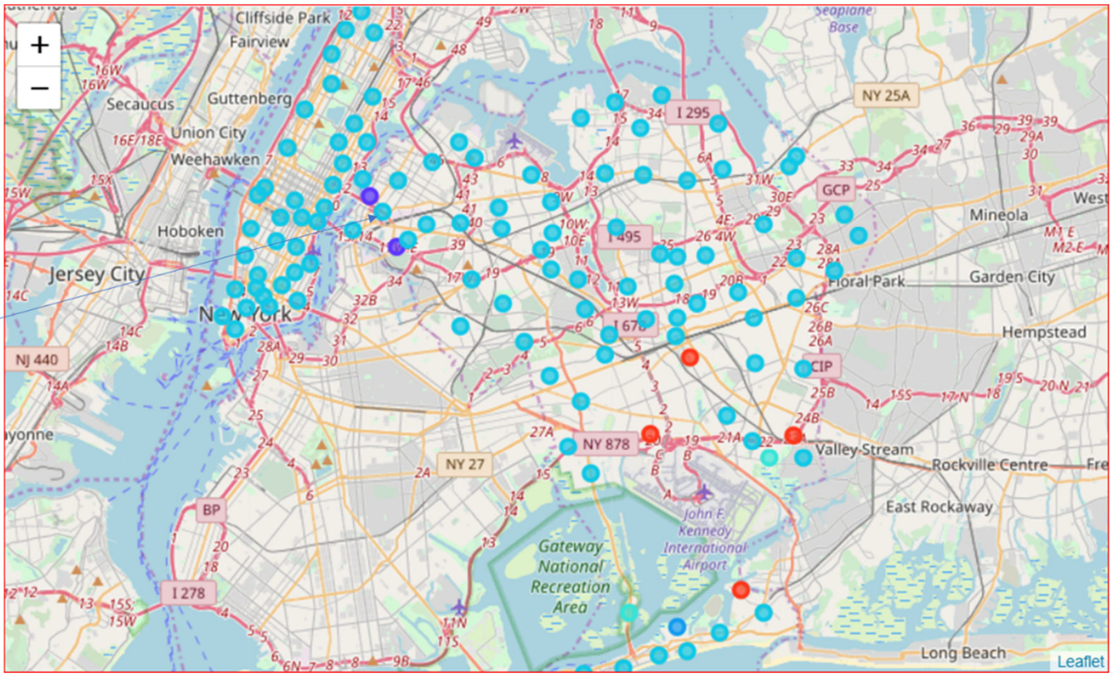
South Lake Union Amazon HQ

*Figure 4 Clustering results for Seattle neighborhoods*



Crystal City Amazon 2ndHQ

*Figure 5 Clustering results for Arlington/Washington DC neighborhoods*



Long Island City
Amazon 2ndHQ
(Cancelled)

*Figure 6 Clustering results for New York neighborhoods*

From the map data, Arlington/Washington DC showed more diversity in the neighborhood comparing to Seattle and New York. However, 56% neighborhoods in Arlington/DC, 92% neighborhoods in New York show similarity with 90% neighborhoods in Seattle as in cluster #5 as shown in Figure 7.

In addition, the exact 3 Amazon headquarter locations (South lake union at Seattle, Crystal city at Arlington and Long island city at New York) are in same cluster as shown in Figure 8.

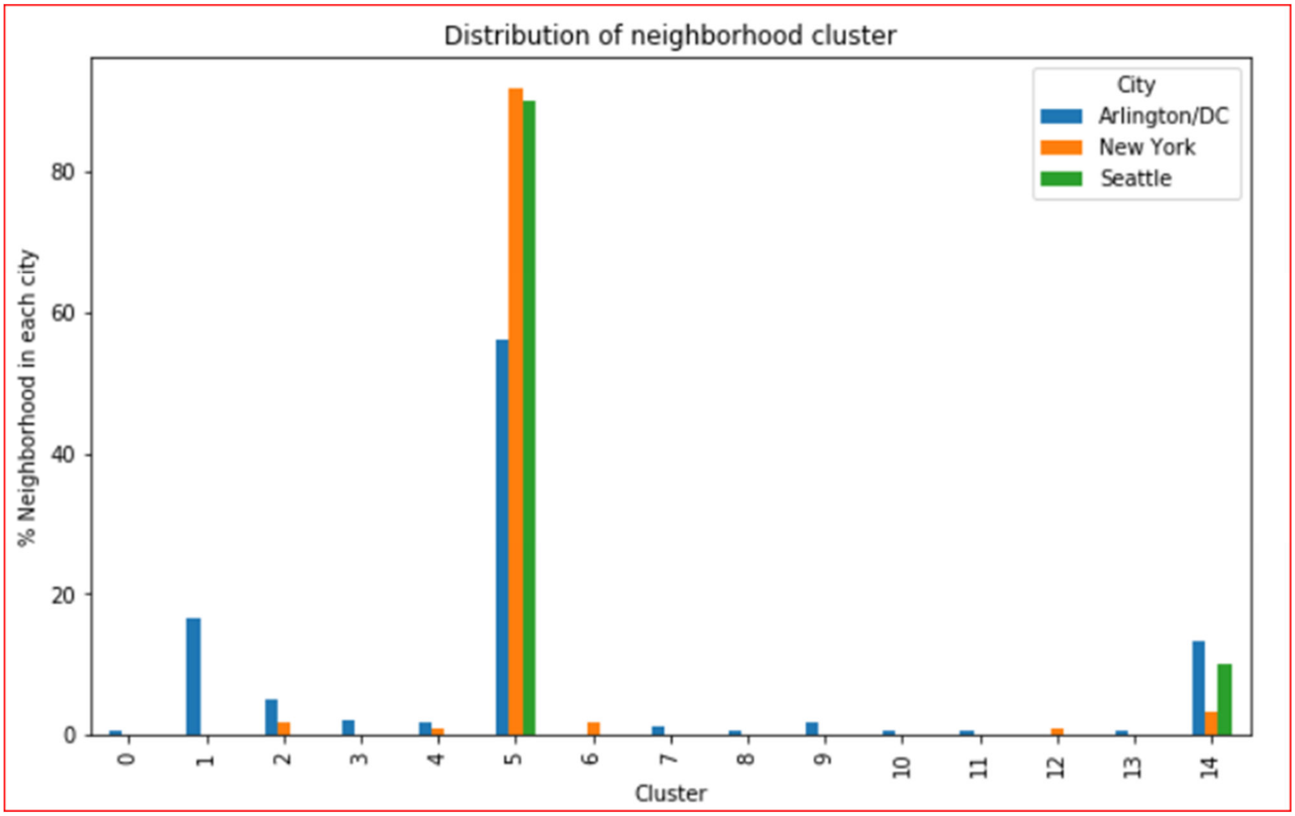
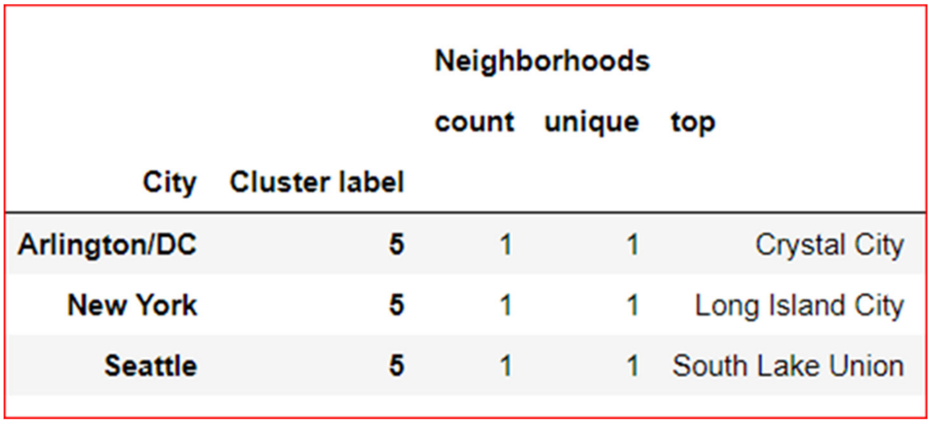


*Figure 7 Distribution of neighborhood clusters*



| | | Neighborhoods | | |
|---|---|---|---|---|
| | | count | unique | top |
| City | Cluster label | | | |
| Arlington/DC | 5 | 1 | 1 | Crystal City |
| New York | 5 | 1 | 1 | Long Island City |
| Seattle | 5 | 1 | 1 | South Lake Union |

*Figure 8 Cluster of 3 Amazon HQs neighborhoods*

## 4. Conclusions

The clustering data show that all neighborhoods in three Amazon HQ cities showed high similarity by clustering. The 3 amazon headquarters locations (South lake union at Seattle, Crystal city at Arlington and Long island city at New York) are in same cluster. From this data analysis, the neighborhood similarity might play an important role during the selection of 2nd HQ for Amazon.

To attract new business operation for unicorn companies for a city/territory, the similarity of neighborhoods between the proposed location and current company location is worthwhile to consider beside the financial/tax incentive, availability of talents and other political reasons. This provide a check point for the candidate location to bid for potential business operation.

5. References

1. https://en.wikipedia.org/wiki/Amazon_HQ2
2. http://opendata.dc.gov/datasets/neighborhood-labels/data
3. https://cocl.us/new_york_dataset
4. https://www.seattle.gov/neighborhoods/neighborhoods-and-districts