Impact of Neighborhood Similarity on Amazon HQ2 selection

Will good neighborhood attract new business office for Unicorn companies

Jun Sun

05/05/2019

Introduction

- 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 Nov, 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.
- 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 criteria. For example, does the 2HQ have similar neighborhood as that in current HQ in Seattle, WA? 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.

Methodology

The neighborhood info or list for three cities, Seattle WA, Arlington VA and Queens, NY can be obtained from open data source or Wiki website. Then the latitude and longitude information for each neighborhood can been obtained using geopy package.

After consolidation of all datasets, the top 100 popular venues from each neighborhood can be retrieved using Four Square API.

K-cluster algorithm will be used to cluster all neighborhoods. The optimal cluster can be obtained by Sum of squared distance, or Silhouette score. The all neighborhoods in3 cities will be clustered using the optimal cluster.

Finally the similarity of neighborhood from each cities will be compared to check how much of similar neighborhood from 3 cities. This will lead to conclusion regarding the impact of neighborhood similarity on Amazon 2HQ selection.

Obtain/preparation Neighborhood geospatial data

- 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 DC.
- The list of neighborhoods for Seattle is obtained from Seattle gov website and Latitude/Longitude data for each neighborhood is obtained using the Nominatim package from geopy library.
- Same approach was applied for Arlington neighborhood. The neighborhood data for Washington DC is obtained from DC gov site. And the Arlington and DC neighborhood data is combined into one dataset using "Arlington/DC" as city label.
- For Queens and Manhattan neighborhood data, it is available from this New York geospatial json file with some data manipulation.
- All the datasets from 3 cities is consolidate as a single dataset for further analysis.

	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

Obtain/preparation Neighborhood geospatial data

- 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 DC.
- The list of neighborhoods for Seattle is obtained from Seattle gov website and Latitude/Longitude data for each neighborhood is obtained using the Nominatim package from geopy library.
- Same approach was applied for Arlington neighborhood. The neighborhood data for Washington DC is obtained from DC gov site. And the Arlington and DC neighborhood data is combined into one dataset using "Arlington/DC" as city label.
- For Queens and Manhattan neighborhood data, it is available from this New York geospatial json file with some data manipulation.
- All the datasets from 3 cities is consolidate as a single dataset for further analysis. Part of dataset is shown as below.

	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

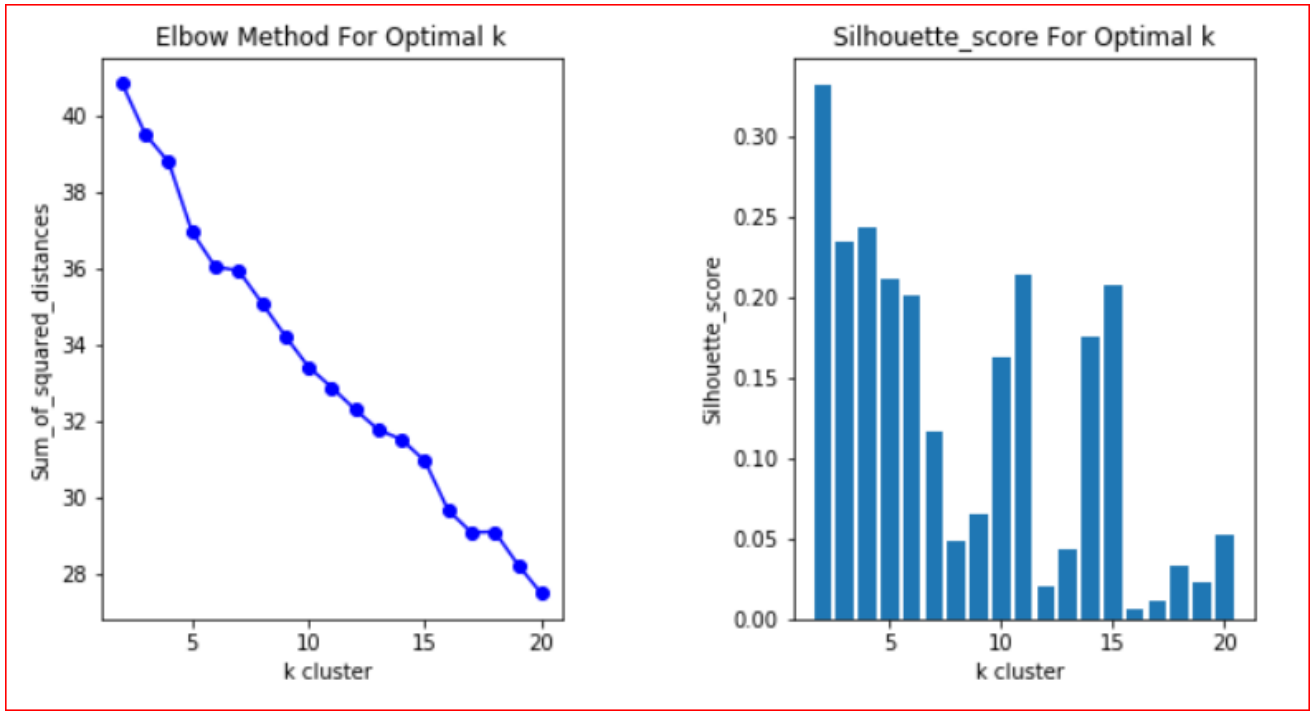
Obtain venues data for all neighborhoods for clustering

- The top 100 venues within 500m of radius for each neighborhood is obtained by calling the FourSquare API with the geospatial data for all the neighborhoods.
- 10741 venue records are obtained and are transformed into a dataframe as below for clustering.

	Neighborhoods	City	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	Alternative Healer	American Restaurant	...	Whisky Bar	Wine Bar	Wine Shop	Winery
0	23rd & Union/Jackson	Seattle	0	0	0	0	0	0	1	0	...	0	0	0	0
1	23rd & Union/Jackson	Seattle	0	0	0	0	0	0	0	0	...	0	0	0	0
2	23rd & Union/Jackson	Seattle	0	0	0	0	0	0	0	0	...	0	0	0	0
3	23rd & Union/Jackson	Seattle	0	0	0	0	0	0	0	0	...	0	0	0	0
4	23rd & Union/Jackson	Seattle	0	0	0	0	0	0	0	0	...	0	0	0	0

Determine the optimal cluster for K-Means clustering

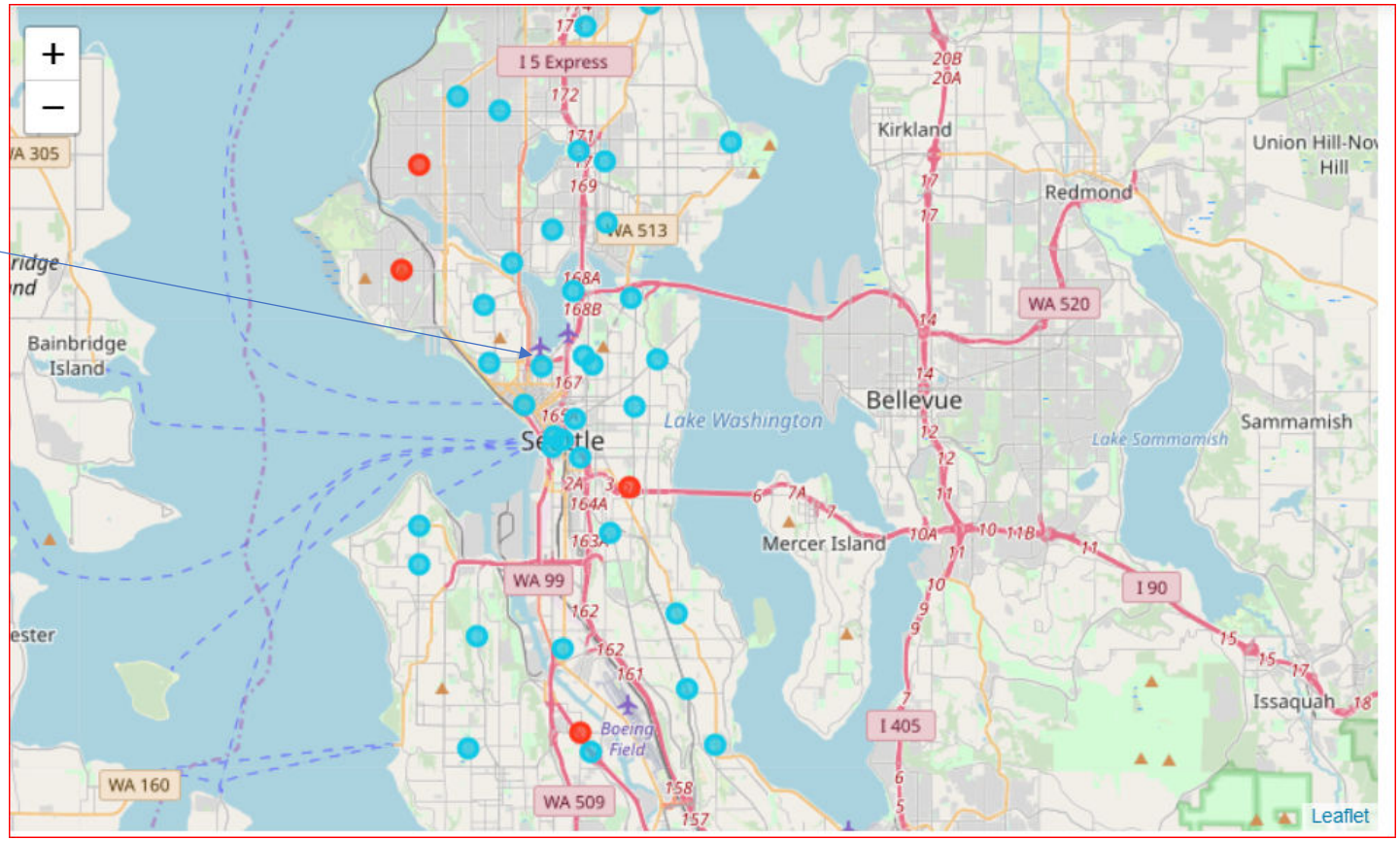
k cluster number cannot be identified by sum of squared distance, but the Silhouette_score indicates k=15 is reasonable cluster number for further clustering as the figure shown below.



K-Means clustering on neighborhoods for 3 Amazon HQs cities

Clustering results for Seattle neighborhoods

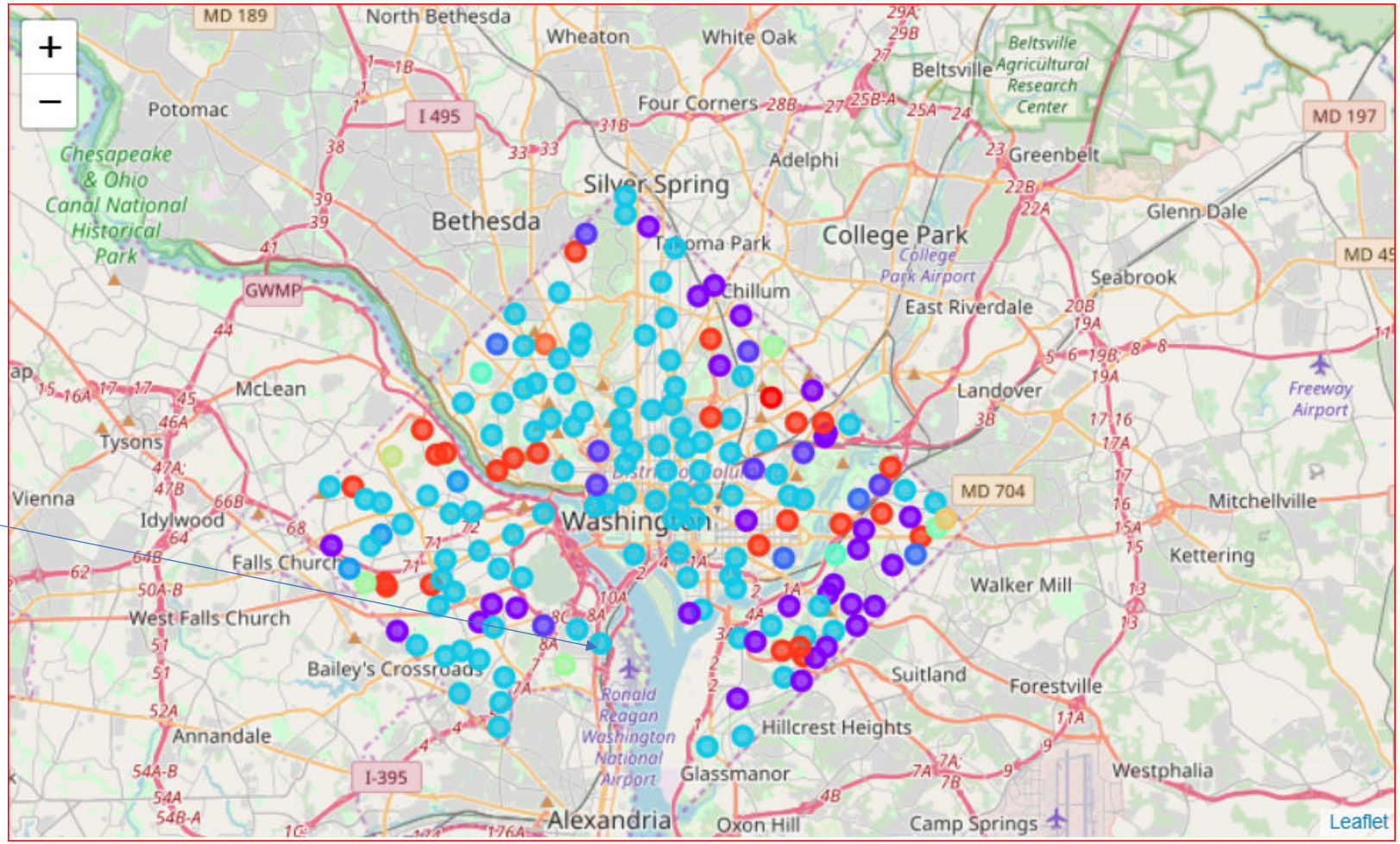
South Lake Union
Amazon HQ



K-Means clustering on neighborhoods for 3 Amazon HQs cities

Clustering results for Arlington/Washington DC neighborhoods

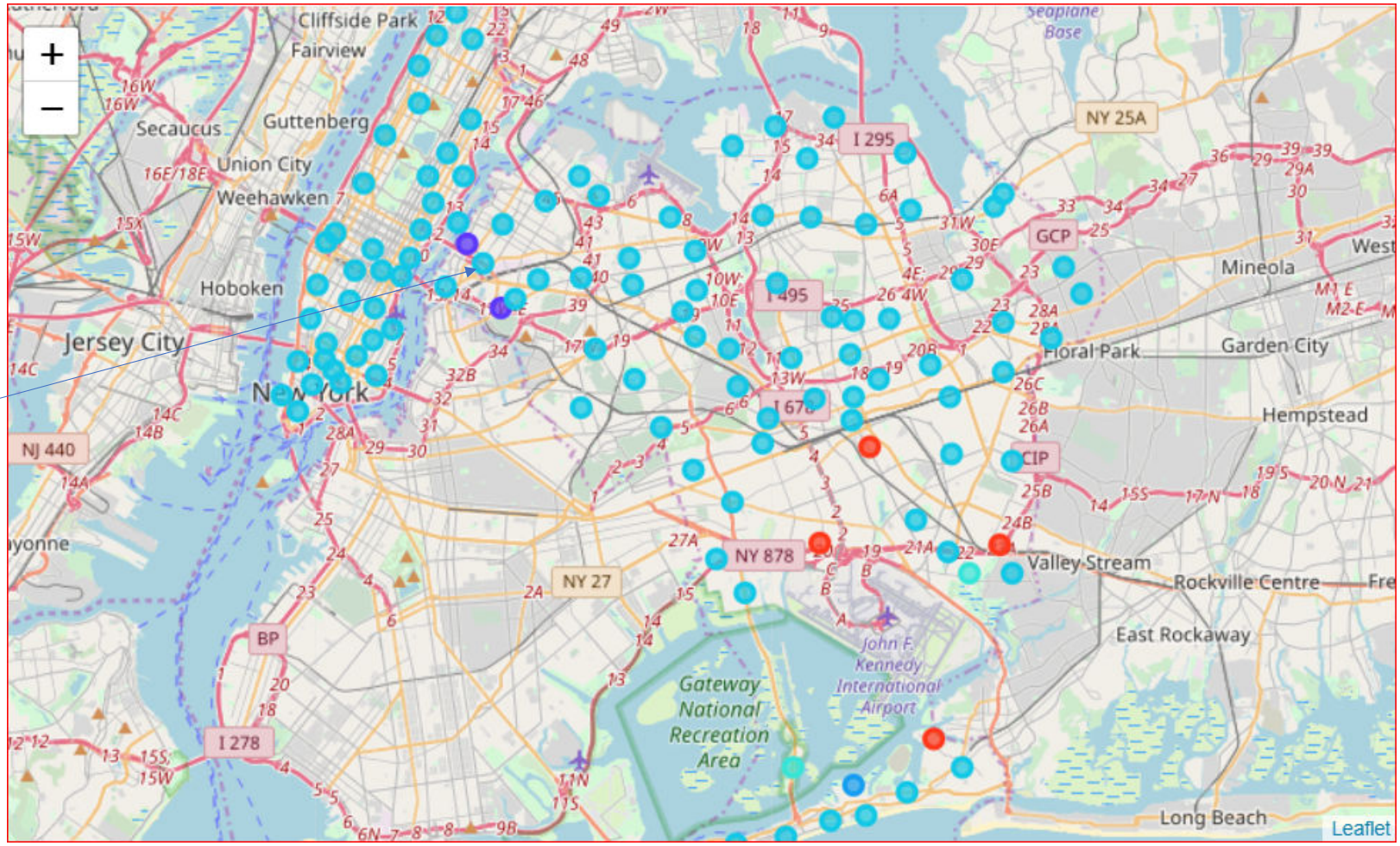
Crystal City
Amazon 2ndHQ



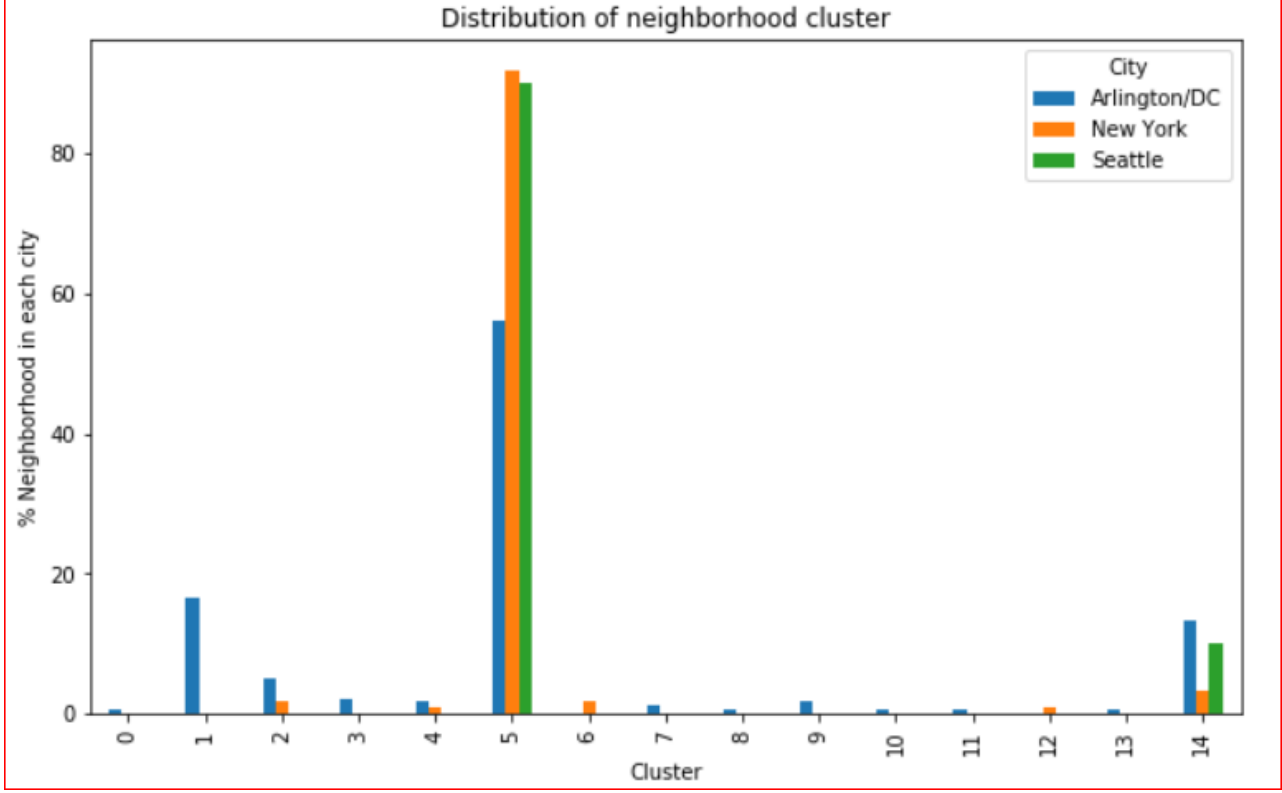
K-Means clustering on neighborhoods for 3 Amazon HQs cities

Clustering results for New York neighborhoods

Long Island City
Amazon 2ndHQ
(Cancelled)



- 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.
- The exact 3 Amazon headquarters locations (South lake union at Seattle, Crystal city at Arlington and Long island city at New York) are in same cluster.



		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

Conclusions and Discussions

- All neighborhoods in three Amazon HQ cities showed high similarity by clustering.
- The exact 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, 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.

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

The Jupyter notebook for this project is available at this link.