Analyzing Cities: A Case Study of Singapore and New York

Capstone Project
IBM Professional Certificate in Data Science

Michael W.D. Kurz

This Version: April 19, 2020

Table of Contents

1. Introduction

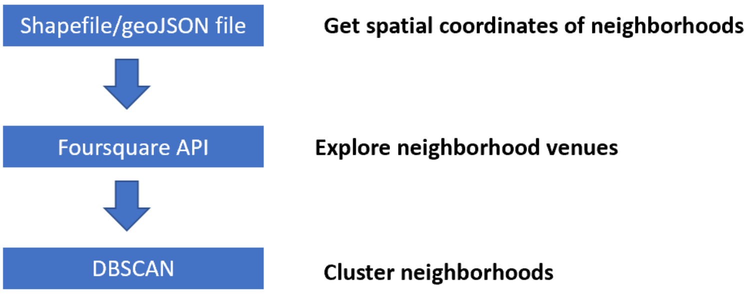
As cities' around the world grow larger it becomes increasingly important to understand their internal structure as well as the similarities and differences between cities. The analysis presented in this report focuses on Singapore and New York City. These cities present an interesting case study for a preliminary analysis as they both are major financial, commercial, and cultural hubs with large, diverse populations. However, the two cities are also situated in culturally significantly different regional environments.

For Singapore and New York City, I am using DBSCAN to find clusters of similar neighborhoods in Singapore. For this analysis, I am defining similarity in terms of the types and number of amenities in the neighborhood. Thus, I describe each neighborhood as a vector of the amenities within them. The analysis in this report is similar to the analysis of U.S. cities presented in (Preotiuc-Pietro, Cranshaw, & Yano, 2013).

I augment this by demographic data on the residents of each neighborhood to gain a better understanding of which demographic groups reside most commonly in which kind of neighborhood. I will investigate whether similarity in terms of amenities correlates with similarity in terms of the demographic characteristics of residents.

For each of the two cities, Singapore and New York, I obtain the spatial location of neighborhoods form free database available online, use the location data to download information for venues located in each neighborhood through the Foursquare API, and finally use the so-called DBSCAN algorithm to find clusters of similar neighborhoods with each of the cities. The workflow for each city is outlined in Figure 1. After having obtained clusters for each city I use agglomerative clustering applied to the city neighborhood clusters to create a hierarchy of cluster similarities that allows investigating similarities of clusters across the two cities.

Figure 1: Workflow for each city


Shapefile/geoJSON file
Get spatial coordinates of neighborhoods
Foursquare API
Explore neighborhood venues
DBSCAN
Cluster neighborhoods

A Jupyter notebook containing the Python codes used to create the results summarized in this report can be accessed on GitHub under the link shown in the Appendix.

Identifying neighborhoods that are like each other in terms of amenities and which demographic groups typically reside in them can help urban planners/city planners to make more informed decisions. It may also help the proprietors of restaurants, bars, etc. to make informed decisions about the location of new venues. Additionally, the results can help residents that need to relocate from one part of the city to another to find neighborhoods that are like the one they currently live in or to select a neighborhood with similar amenities.

2. Data

The Urban Authority of Singapore distinguishes five broader regions for planning purposes: The Central region, North region, North-East region, East region, and West region. Each of the regions is further split into planning areas and each planning area is again divided into sub-zones. For my analysis, I will focus on the latter. I obtain spatial coordinates for Singapore's subzones through a shapefile made freely available through the Government of Singapore's data portal.[1] These sub-zones are divisions of a planning area which are usually centered around a focal point such as a neighborhood center or activity node. There are in total of 323 subzones across the five regions of Singapore. Most subzones (139) are in the central region. In Figure 2 I show a map of Singapore, where each dot on the map represents a subzone. The different colors indicate the various regions: Central region (blue), East region (yellow), North region (red), North-East region (green), and West region (aqua).

For each of the neighborhoods, I use the Foursquare API to download information for the top 100 venues located in the neighborhood. This results in 5.924 unique venues in Singapore belonging to 359 unique venue categories. Many of these venues' categories include generic categories of limited usefulness in characterizing a neighborhood. Examples of such categories are, "Bridge", "Building", "Crossing", etc. I drop these venues from the dataset. Moreover, I consolidate various venue categories into a single category. For example, I am merging Airport and Airport Terminal into the same category, equally, I am merging restaurant categories indicating

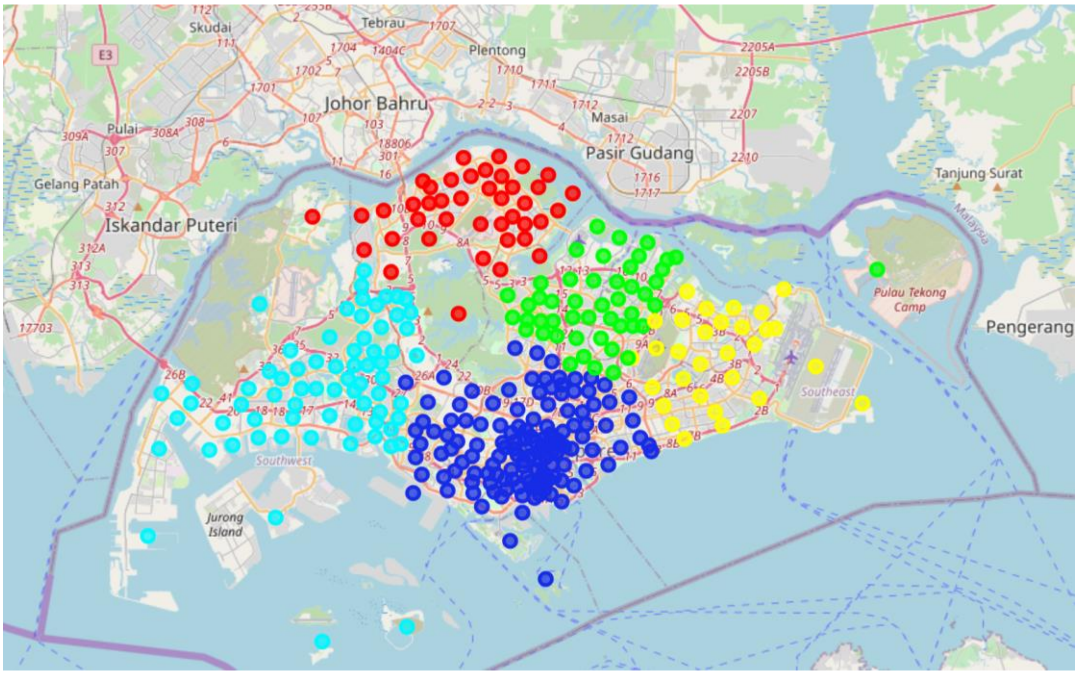
[1] The shapefile can be downloaded via https://geo.data.gov.sg/mp14-subzone-web-pl/2014/12/05/shp/mp14-subzone-web-pl.zip
The spatial coordinates of subzones in the shapefile are not latitude and longitude coordinates. It uses a transverse mercator projection centered on the coordinates 1.367°E 103.83°N. Therefore, it necessary to transform the projection to latitude and longitude coordinates.

different flavors of regional food (such as different flavors of Chinese or Italian food) into single categories. After this consolidation, I retain 152 distinct venue categories for Singapore.[2]

For Singapore I additionally obtain some socio-demographic data for each subzone which is also conveniently provided by the Singaporean government. The socio-demographic data is provided via https://data.gov.sg/dataset/resident-population-by-planning-area-subzone-age-group-and-sex-2015?resource_id=68775b41-3025-4763-970d-f479652e8b05 and https://data.gov.sg/dataset/resident-population-by-planning-area-subzone-and-type-of-dwelling-2015?resource_id=2719aca1-6b37-4f8b-ac3f-a866d32df7c6

The socio-demographic datasets represent the results of the 2015 wave of the Singapore household survey. For each subzone the socio-demographic datasets include the total resident population and breakdowns of the resident population by age group, type of dwelling, and ethnicity. This data allows us to analyze the similarity between neighborhoods of Singapore in terms of the demographic characteristics of its residents, enabling the comparison between neighborhood similarity in terms of amenities and terms of demographics.

Figure 2: Singapore's neighborhoods color-coded by region

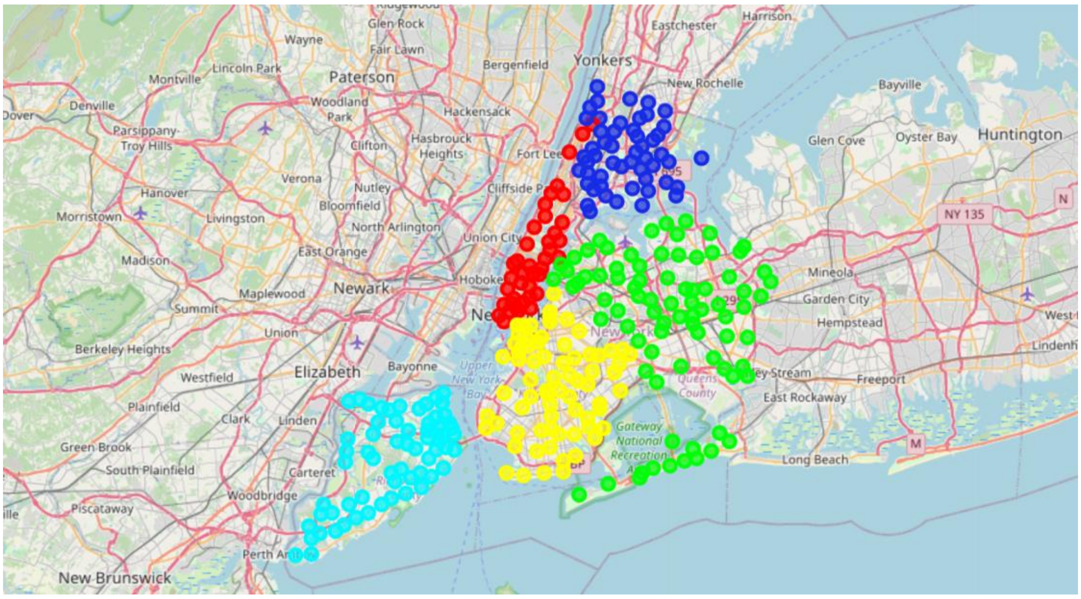


Spatial locations for New York's neighborhoods are obtained from a geoJSON file that accessible via the following links: https://cocl.us/new_york_dataset (alternatively via https://geo.nyu.edu/catalog/nyu_2451_34572). Like Singapore, I plot the locations of New York's neighborhoods on a map as shown in Figure 3. In Figure 3 blue markers represent Bronx's neighborhoods, red indicates Manhattan, aqua represents Staten Island, yellow is used for Brooklyn, and green for Queens.

Venues data for New York's neighborhoods are also obtained through the Foursquare API. I apply the identical consolidation procedure that I used for Singapore. In this way I obtain 5.924 unique venues belonging to 193 distinct venue categories for New York City.

[2] The exact consolidation procedure is shown in the Jupyter notebook accompanying this report.

Figure 3: New York City's neighborhoods color-coded by borough



3. Methodology

In this section, I outline the specific methodology that I used for clustering.

3.1 Clustering neighborhoods within a city

I use the DBSCAN[3] algorithm originally proposed by (Ester, Kriegel, Sander, & Xu, 1996) to find clusters of similar neighborhoods within the cities of Singapore and New York City. DBSCAN finds clusters by searching for areas of high density that separated by areas of low density. Using DSCAN presents some advantages over other clustering algorithms. Specifically, DBSCAN can identify arbitrarily shaped clusters, does not require the user to specify the number of clusters before-hand and it is robust to noise in the input data. The DBSCAN algorithm is implemented in the Python sklearn library which I will employ in this analysis.[4]

The computation of dense areas of the data requires a choice of a distance metric as a measure dissimilarity between neighborhoods. Recalling that I characterize neighborhoods as vectors of the frequency of occurrence of venue categories within the neighborhood, distance reflects not spatial distance but dissimilarity in terms of venue types that are available in a neighborhood. The most standard distance metric is the Euclidean distance which measures the straight-line distance between two points on a (multi-dimensional) plane.

Euclidean distance arises as a special case of the Minkowski distance which is given by

$$Minkowski = \left(\sum_{i}^{n} |x_i - y_i|^p\right)^{1/p}$$

The Euclidean distance results from setting $p = 2$. (Aggarwal, Hinneburg, & Keim, 2001) investigate the properties of various distance metrics in a high-dimensional dataset and argue that the Euclidean distance does

[3] **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise

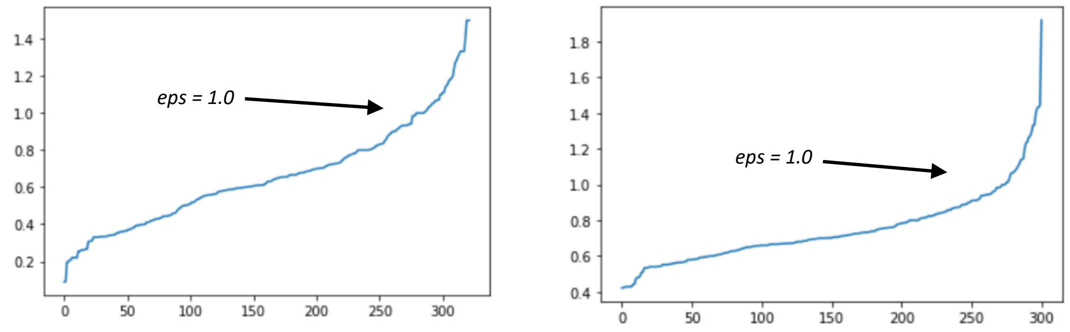
[4] For a more detailed outline of the DBSCAN algorithm see also the sklearn user guide https://scikit-learn.org/stable/modules/clustering.html#dbscan or this blog-post: https://towardsdatascience.com/dbscan-algorithm-complete-guide-and-application-with-python-scikit-learn-d690cbae4c5d

not perform well in high-dimensions. They propose that lower values for p should be chosen in high-dimensional datasets.

Even after consolidating venue categories, the Foursquare dataset for Singapore contains 152 different venue categories. Therefore, the clustering problem in this analysis is high-dimensional. Following (Aggarwal, Hinneburg, & Keim, 2001), I use the Manhattan distance (sometimes also called City-block distance) which also arises as a special case of the Minkowski distance for $p = 1$.[5]

Besides the choice of the distance metrics, DBSCAN requires specifying the maximum distance, eps, between data points such that they can be considered members of the same cluster and the minimum possible size of a cluster. I set the minimum possible size of a cluster to 2, i.e., two neighborhoods would already be enough to form a cluster. Based on this choice I determine eps following the procedure proposed in (Rahmah & Sitanggang, 2016). The procedure is based on the idea of obtaining an approximation of distances in dense areas of the data. Following (Rahmah & Sitanggang, 2016) I find, for each neighborhood the distance to its two nearest neighbors in terms of the Manhattan distance measure and plot the resulting distances for the Singapore neighborhood data and the New York City neighborhood data in Figure 4 below. Using the plot in Figure 4 I determine eps as the point of the greatest change in curvature. For both cities I select $eps = 1.0$.

Figure 4:Nearest Neighbor distance in Singapore (left) and New York City (right)



3.2 Cluster similarities between Singapore and New York City

After clusters of similar neighborhoods within Singapore and within New York City are found, I am interested in comparing similarities across the two cities. For this purpose, I am using agglomerative clustering applied to the neighborhood clusters found in Singapore and New York City to create a dendrogram linking the clusters. This approach has the primary advantage that it allows for easy visualization of the complete range of nested groups of the individual cities' neighborhood clusters.

Agglomerative clustering is based on the idea of interpreting each data point as its cluster and joining clusters pairwise with its nearest neighbor until all clusters are joint to create a hierarchical structure of clusters. Thus, at first, each cluster found in Singapore and New York is interpreted by the algorithm as its cluster. The algorithm proceeds by grouping the clusters across the cities based on their distance/similarity. As for the individual cities, clusters are characterized in terms of the frequency of venue categories that occur within each cluster. Distance

[5] Note that the Euclidean distance is the L2-norm of the data, while the Manhattan distance, accordingly, is the L1-distance of the data.

between clusters is again calculated using the Manhattan distance metric. Note that since some venue categories only occur in Singapore and some that occur only in New York, the dataset of all clusters contains 207 distinct venue categories. Since clusters are merged during the agglomerative clustering algorithm it is necessary to define how the characteristics of data points are aggregated (or in other words linked). I use "complete" linkage, i.e. algorithm computes and merges clusters to minimize the maximum distance between the clusters (i.e., the distance of the farthest elements).

4. Results

The following sub-sections of the report present the results of the clustering of neighborhoods within Singapore and New York City and an analysis of the similarities of the clusters across the two cities.

4.1 Singapore's neighborhood clusters

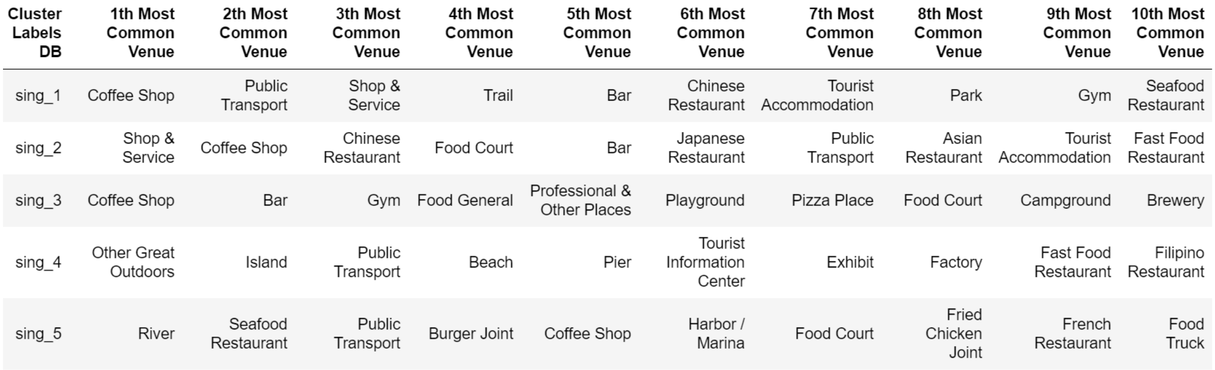
DBSCAN identifies 5 clusters of similar neighborhoods in Singapore. I label the neighborhoods generically "sing_1" to "sing_5". While the number of clusters is the same as the number regions, the size of the clusters in terms of the number of included subzones differs greatly from the regions. In Table 1 I show the sizes of clusters and regions.

Table 1: Number of neighborhoods per cluster vs. Number of Neighborhoods per region

Cluster	Sing_1	Sing_2	Sing_3	Sing_4	Sing_5	Total
# of neighborhoods	31	283	5	2	2	323
Regions	Central	North	North-East	East	West	Total
# of neighborhoods	134	41	48	30	70	323

Here it is worth mentioning that the neighborhoods in cluster "Sing_1" may not constitute a cluster in the classical sense. "Sing_1" summarizes regions of the data with low density.[6] In Table 2 I show the top 10 most common venue categories in each of the clusters.

Table 2: Most common venues per cluster in Singapore



Cluster Labels DB	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
sing_1	Coffee Shop	Public Transport	Shop & Service	Trail	Bar	Chinese Restaurant	Tourist Accommodation	Park	Gym	Seafood Restaurant
sing_2	Shop & Service	Coffee Shop	Chinese Restaurant	Food Court	Bar	Japanese Restaurant	Public Transport	Asian Restaurant	Tourist Accommodation	Fast Food Restaurant
sing_3	Coffee Shop	Bar	Gym	Food General	Professional & Other Places	Playground	Pizza Place	Food Court	Campground	Brewery
sing_4	Other Great Outdoors	Island	Public Transport	Beach	Pier	Tourist Information Center	Exhibit	Factory	Fast Food Restaurant	Filipino Restaurant
sing_5	River	Seafood Restaurant	Public Transport	Burger Joint	Coffee Shop	Harbor / Marina	Food Court	Fried Chicken Joint	French Restaurant	Food Truck

Looking at the socio-demographic characteristics of the clusters may provide further insights. Recall that these socio-demographic characteristics were not used as input for the clustering algorithm. Therefore, the arising patterns provide insights into potential applications of the clustering for understanding the socio-demographic

[6] DBSCAN labels these areas as cluster "-1", while regular clusters labeled with numbers starting from 0.

structure of Singapore. In Figure 5 to Figure 7 I show distributions of resident's age groups, dwelling types, and ethnic groups for all clusters. Notably, clusters "sing_4" and "sing_5" do not have any residents. A look at Table 2 suggests that this may not be surprising as these clusters seem to comprise primarily venues related to nature and outdoor activities.

The dwelling types in Figure 6 show the most interesting pattern. For cluster "sing_1" the most common dwelling type is from category "other dwelling" which comprises dwellings such as single-family homes etc. On the other hand, for cluster "sing_2" the most common dwelling type are 1 or 2 room apartments in the city's housing development buildings. The analysis of socio-demographic data in each of the clusters, however, should be taken with a grain of salt. The socio-demographic data is obtained from a several years old census survey. Since the Foursquare venues data is based rather on recent information submitted from Foursquare users. This timing mismatch would affect conclusions if the composition of residents in the neighborhoods changed significantly.

Figure 5: Residents' age distribution

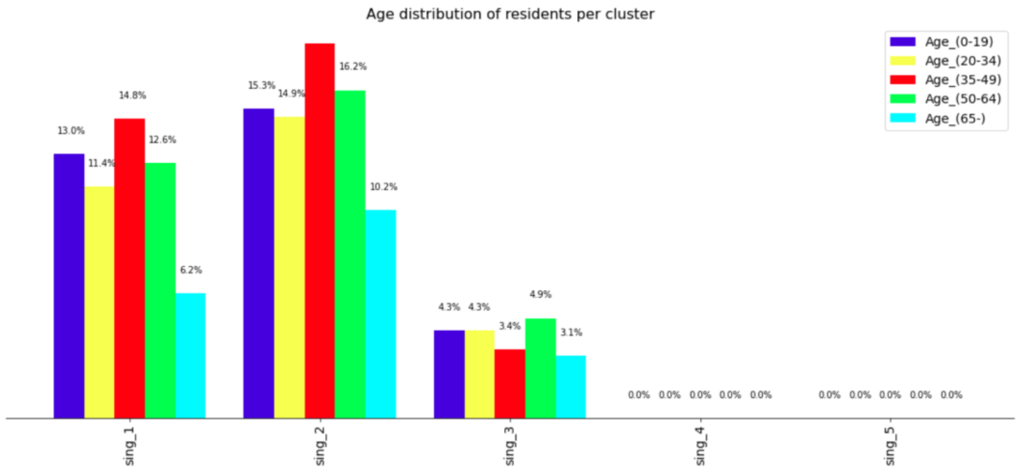


Figure 6: Residents' dwelling type distribution

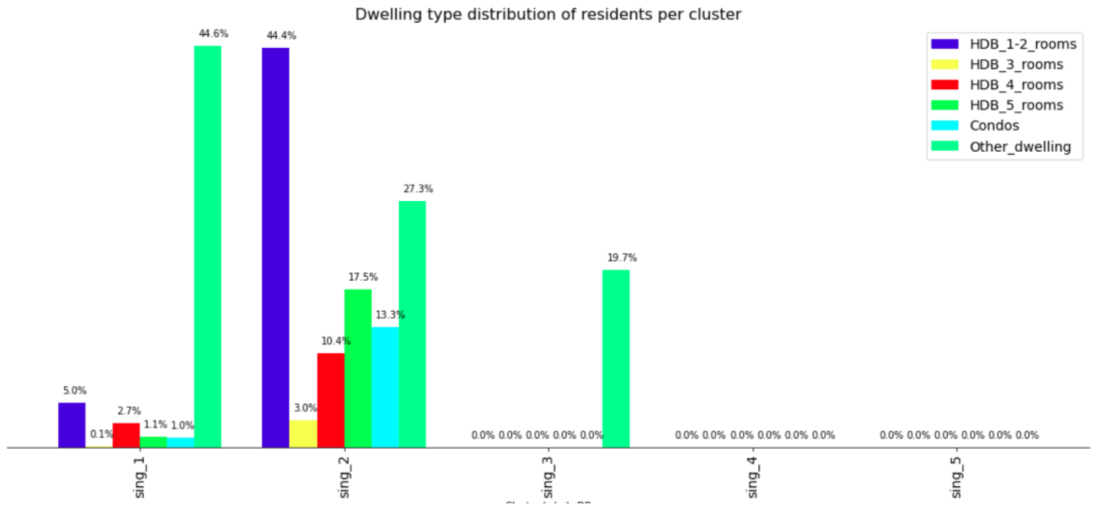
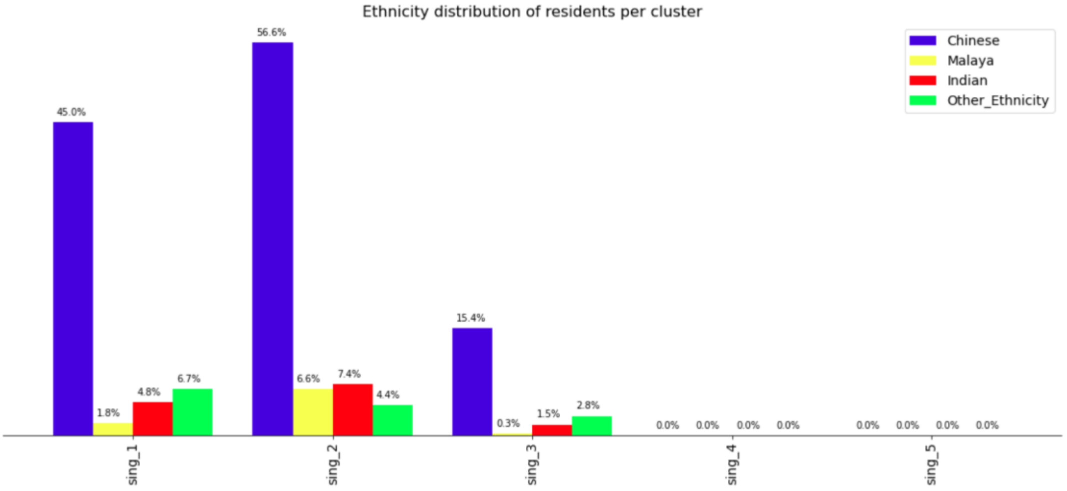


Figure 7: Distribution of residents' ethnic groups



4.2 New York City's neighborhood clusters

DBSCAN identifies 4 clusters of similar neighborhoods in New York City. In line with the analysis of Singapore, I label the neighborhood clusters generically "nyc_1" to "nyc_4". Like the clustering results for Singapore, DBSCAN finds a single cluster that encompasses most of the city's neighborhoods and the cluster "nyc_1" summarizes low-density regions of the data. The sizes of clusters and boroughs in terms of their neighborhoods are shown in Table 3. In Table 4 I show the top 10 most common venue categories found in each of the clusters. Unfortunately, I do not have any socio-demographic data comparable to the data for Singapore available.

Table 3: Number of New York neighborhoods in Clusters vs. Boroughs

Cluster	Nyc_1	Nyc_2	Nyc_3	Nyc_4	-	Total
# of neighborhoods	27	268	3	2	-	306
Borough	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Total
# of neighborhoods	52	70	40	81	63	306

Table 4: Most common venues per cluster in New York City

Cluster Labels DB	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
nyc_1	Beach	Shop & Service	Bar	Tourist Accommodation	Park	Dessert Shop	Thai Restaurant	Pizza Place	Gym	Public Transport
nyc_2	Shop & Service	Bar	Gym	Coffee Shop	Pizza Place	Dessert Shop	Italian Restaurant	Deli / Bodega	Chinese Restaurant	Bakery
nyc_3	Athletics & Sports	Dessert Shop	Playground	Park	Gym	Harbor / Marina	Mexican Restaurant	Deli / Bodega	Pizza Place	Chinese Restaurant
nyc_4	Caribbean Restaurant	Shop & Service	Gym	Music Venue	Pizza Place	Bakery	Thai Restaurant	Dessert Shop	Playground	Chinese Restaurant

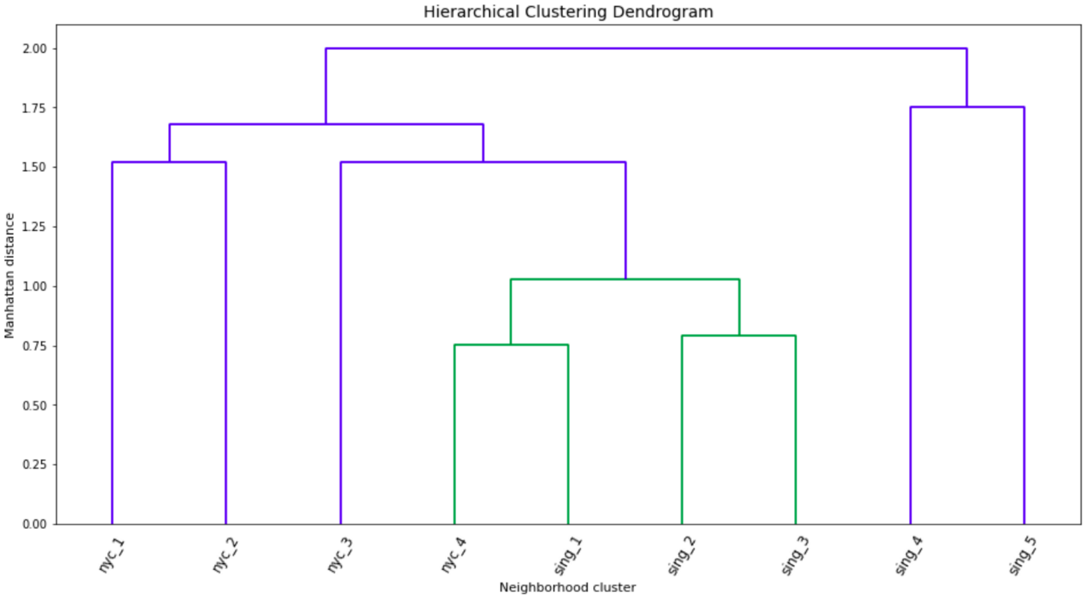
4.3 Similarities across Singapore and New York City

Finally, I compare similarities of neighborhood clusters in Singapore and New York City. The dendrogram resulting from the agglomerative clustering algorithm is shown in Figure 8 below.

There seem to two major groups of neighborhood clusters. One comprising New York's neighborhood clusters nyc_1 to nyc_4 and Singapore's clusters sing_1 to sing_3 (left group in the dendrogram), the other comprising

Singapore’s neighborhood clusters sing_4 and sing_5 (right group in the dendrogram). Within the left group in the dendrogram clusters, nyc_1 and nyc_2 seem to share fewer similarities with neighborhoods in Singapore than nyc_3 or nyc_4. Especially, New York’s cluster nyc_4 seems to comprise neighborhoods offering amenities that are comparable to neighborhoods in Singapore’s cluster sing_1. Note that the two largest clusters “sing_2” and “nyc_2” only joint at the second-highest hierarchy level, implying a significant distinctiveness of the clusters.

Figure 8: Cluster Similarities between Singapore and New York City



5. Discussion

The results of the clustering reveal a striking similarity of neighborhoods within both cities Singapore and New York. For both cities, most neighborhoods are grouped into a single cluster. This implies that for both cities most neighborhoods offer rather similar types of amenities. Also, for both cities, the second largest clusters are the low-density clusters, i.e. clusters with too little similarity to each other or to other neighborhoods to be grouped into any of the other clusters. The similarity in results is remarkable given Singapore and New York are situated in very different cultural environments and look back at a very different history. On the other hand, both cities are financial and cultural centers, focal points of global trade and commerce, and multi-cultural melting pots. Similar patterns in terms of the cluster results seem to suggest latter influence outweigh history and regional culture. This provides potentially interesting insights for urban planners but also expatriates relocating from one city to the other. In many aspects, Singapore and New York seem rather similar and especially similarities of neighborhoods within each city are high. For Urban planners this implies branding of neighborhoods in terms of amenities seems less meaningful for most neighborhoods. The unique profile of neighborhoods could be sharped by encouraging greater diversity in venues. Expatriates can expect significant differences across cities but location choices within cities seem less important as far as the availability and diversity of amenities are concerned.

However, the results presented in this report should be only preliminary. Future research should be conducted to solidify the evidence presented here. More detailed data on venues such as price ranges, visitor ratings, and frequency of visitor check-ins should be considered. I left these additional data items out of scope for the current analysis since they subject to a stricter download volume limit for free accounts in Foursquare API. Therefore, downloading them within a reasonable timeframe for all venues in two major cities was not feasible.

6. Conclusion

The analysis presented in this report is very preliminary but provides some ideas and insights for the comparison of cities and neighborhoods within cities. Further research should be conducted to refine the results presented here. The similarity in patterns of the cluster exercises for Singapore and New York City raises intriguing questions about the importance of history and regional culture as compared to economic forces in the development and planning of cities.

References

Aggarwal, C., Hinneburg, A., & Keim, D. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In V. V. Van den Bussche J., *Database Theory — ICDT 2001. ICDT 2001. Lecture Notes in Computer Science.* Berlin, Heidelberg: Springer.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).*

Preotiuc-Pietro, D., Cranshaw, J., & Yano, T. (2013). Exploring venue-based city-to-city similarity measures. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing.*

Rahmah, N., & Sitanggang, I. S. (2016). Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra. *IOP Conf. Series: Earth and Environmental Science 31.*

Appendix

The Jupyter notebook underlying the analysis presented above can be retrieved from GitHub using the following link:

https://github.com/MichaelKurz1988/CityAnalysis/blob/master/Singapore_nyc_cluster.ipynb

The notebook requires downloading shapefile and JSON file containing neighborhood coordinates for Singapore and New York City separately. Moreover, personal login credentials for the Foursquare API is required to execute the notebook.