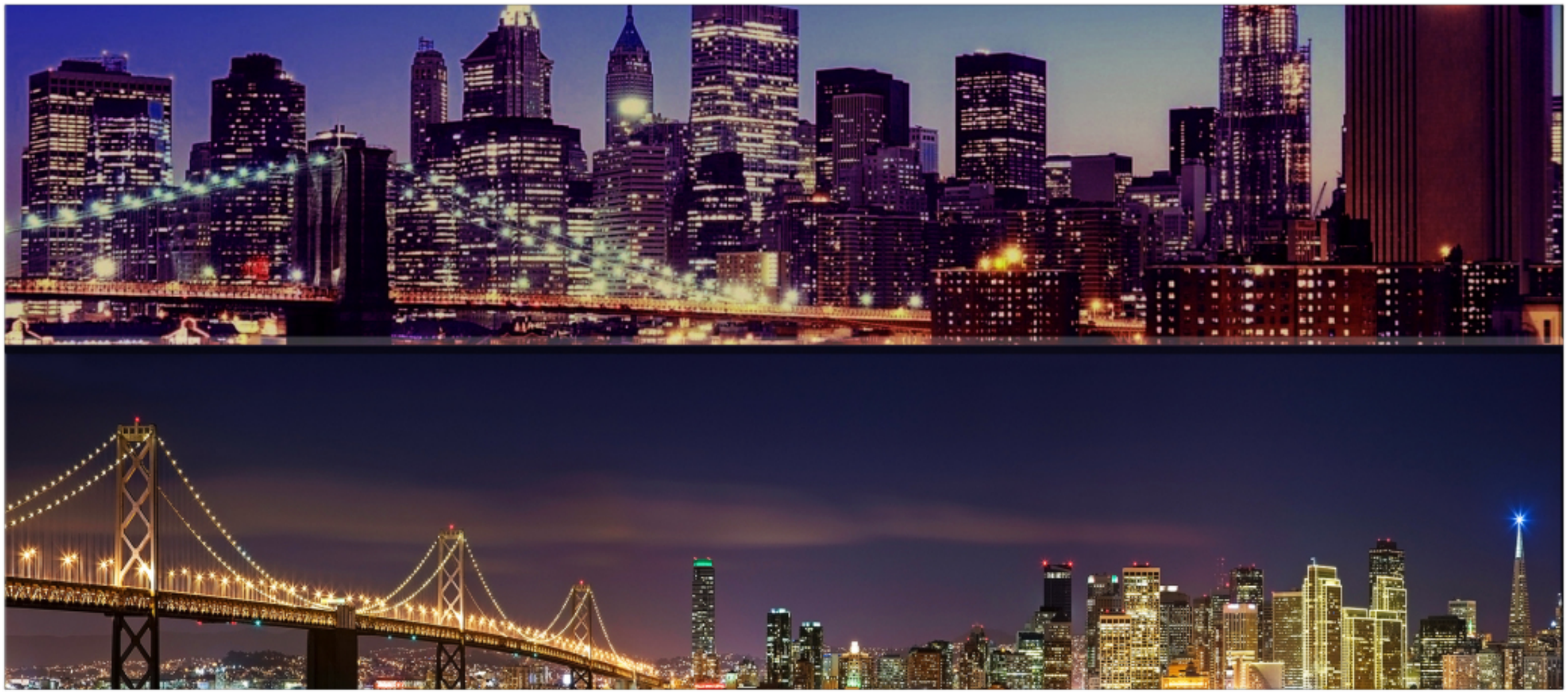

"The Battle of the Neighborhoods"

A Comparison of Manhattan and San Francisco Neighborhoods Based on Their Venues

Applied Data Science
Capstone Project

Introduction

- My company specializes in helping its clients relocate from one U.S. city to another U.S. city (it has not yet branched out to international relocations).
- Our latest client is relocating from Manhattan, a borough of New York City, to San Francisco, a city in northern California. My assignment was to research venues in the neighborhoods of each city, and similar (or dissimilar) neighborhoods, based on their venues.
- My company will use this analysis as part of their efforts to help their client find a San Francisco neighborhood they might like to live in, based on the neighborhoods of Manhattan that the client is already familiar with. In other words, my company wants to be able to tell our client, "if you like the venues in Manhattan neighborhood *A*, you will probably like San Francisco neighborhood *B*." Of course, this could also work the other way, for someone relocating from San Francisco to Manhattan.
- The scope of this study was to find similarity based on the available venues in each neighborhood. A comprehensive comparison of neighborhoods should include other factors, such as population demographics, cost of living, neighborhood "character" (urban or suburban, residential or mixed-use), and so on.

Methodology and Data

- Identify the neighborhoods in each city* and their geographical coordinates.
  - Geographical coordinates for Manhattan were taken from the GeoJSON file obtained from https://geo.nyu.edu/catalog/nyu_2451_34572. The data is published by the New York City Department of City Planning.
  - Geographical coordinates for San Francisco were derived from the GeoJSON file obtained from https://data.sfgov.org/Geographic-Locations-and-Boundaries/Planning-Neighborhood-Groups-Map/iacs-ws63. The data is published by the San Francisco Department of City Planning. This data is in the form of multipolygons, from which centroids for each neighborhood were computed. For details, see the report in the accompanying Jupyter notebook.
- Identify the various venues (restaurants, gyms, parks, and so on) of each neighborhood.
  - Foursquare (https://foursquare.com) is a social networking service that provides recommendations on places to go based on the user's current location. It has a large inventory of venue information, and an API that third-party developers can use to access that information. The Foursquare API is used in this study to obtain information about the venues available in each neighborhood.

* *Even though Manhattan is not a city proper, for convenience it is referred to as a "city" in this study.*

Methodology and Data (*continued*)

- Partition neighborhoods into "clusters" (groups) based on the commonality of venue types.
  - *k*-means clustering (https://en.wikipedia.org/wiki/K-means_clustering) is the method used to partition the neighborhoods into clusters, based on each neighborhood's venue types and frequencies.
- The neighborhoods in each cluster can be considered sufficiently alike, in terms of available venues, so that if you like one neighborhood in a given cluster, you might like another neighborhood in the same cluster.

Data Summary and Observations

- The table below shows a high-level summary of information gathered as part of this study
- City area and population statistics were obtained from Wikipedia
  - https://en.wikipedia.org/wiki/Manhattan
  - https://en.wikipedia.org/wiki/San_Francisco
- Observation: Manhattan's land area is about half the size of San Francisco's, but it has a population twice as large, and almost twice as many venues

	Manhattan	**San Francisco**
Land area	59.1 km²	121.46 km²
Population (as of 2017)	1,664,727	884,363
Population density	28,154 km²	7,282 km²
Neighborhood count	40	37
Venue count	2,930	1,559
Venue category count	326	299

Data Summary and Observations (*continued*)

- The *k*-means method requires you to specify the number of clusters ("n_clusters") into which the neighborhoods are partitioned.
- I had to experiment a bit to find a reasonable numbers. With smaller n_clusters values (12 or less) the tendency was to create one very large cluster of 50+ neighborhoods. Larger values led to smaller clusters, but also resulted in many one- or two-neighborhood clusters.
- I finally settled on n_clusters=14, which generated three medium-sized clusters and minimized the number of one-neighborhood clusters.

Data Summary and Observations (*continued*)

- The three clusters appear as 0, 5, and 7 in the report notebook
  - Cluster 0 has 31 neighborhoods (14 from Manhattan, 17 from San Francisco)
  - Cluster 5 has 16 neighborhoods (8 from Manhattan, 8 from San Francisco)
  - Cluster 7 has 14 neighborhoods (9 from Manhattan, 5 from San Francisco)
  - These represent 61 of the 77 total neighborhoods, which is 79%.
- From this data, we can infer that any neighborhood in a given cluster is similar to another neighborhood in the same cluster.
- The next three slides show the neighborhoods in their respective cities and clusters.

Data Summary and Observations (*continued*)

- Cluster 0

Manhattan	San Francisco
Marble Hill	Haight Ashbury
Manhattanville	Outer Mission
Central Harlem	Inner Sunset
East Harlem	Downtown/Civic Center
Lenox Hill	Lakeshore
Lincoln Square	Russian Hill
Midtown	Ocean View
Chelsea	Mission
Greenwich Village	Inner Richmond
Little Italy	Marina
Soho	Bayview
Midtown South	Glen Park
Tudor City	Castro/Upper Market
Hudson Yards	Bernal Heights
	Presidio
	North Beach
	Western Addition

Data Summary and Observations (*continued*)

- Cluster 5

Manhattan	San Francisco
Chinatown	Seacliff
Inwood	Diamond Heights
Upper East Side	Excelsior
Yorkville	Financial District
East Village	Visitacion Valley
Noho	Presidio Heights
Civic Center	South of Market
Flatiron	Golden Gate Park

Data Summary and Observations (*continued*)

- Cluster 7

Manhattan	San Francisco
Washington Heights	Treasure Island / YBI
Roosevelt Island	Crocker Amazon
Clinton	Parkside
Manhattan Valley	Nob Hill
Gramercy	Chinatown
Financial District	
Sutton Place	
Turtle Bay	
Stuyvesant Town	

Data Summary and Observations (*continued*)

- The remaining clusters were very small, consisting of only one neighborhood, or a small number of neighborhoods all in the same city:
  - Manhattan
    - Battery Park City
    - Carnegie Hill
    - Hamilton Heights
    - Lower East Side
    - Morningside Heights
    - Murray Hill
    - Tribeca
    - Upper West Side
    - West Village
  - San Francisco
    - Noe Valley
    - Outer Sunset
    - Pacific Heights
    - Potrero Hill
    - Outer Richmond
    - Twin Peaks
    - West of Twin Peaks
- We can infer that none of these neighborhoods from one city is sufficiently similar to another neighborhood in the other city.

Data Summary and Observations (*continued*)

- The results suggest that overall, Manhattan and San Francisco have much in common in terms of the venues available in each, even though a relatively small percentage of the neighborhoods remain unique (do not share sufficiently similar types of venues).
- Even in the three major clusters, there is much in common. In each of those clusters there are three or more of these kinds of venues:
  - Bar
  - Café
  - Coffee Shop
  - Mexican Restaurant
  - Park

Data Summary and Observations (*continued*)

- For the 16 neighborhoods that do not seem sufficiently similar to other neighborhoods, a "compromise" might be reached: Re-run the *k*-means analysis with an n_clusters value of 2. The analysis yields one cluster with 67 neighborhoods and a another cluster with 10 neighborhoods. While the similarities between neighborhoods are not as granular, it does provide a small degree of similarity that can be used for the 16 neighborhoods that did not otherwise fit into the primary analysis in this notebook.
- Another approach could be to run the *k*-means analysis on just the 16 neighborhoods, to see what similarities might exist.

Data Summary and Observations (*continued*)

- This study is focused only on venues reported by Foursquare. My experience is that the number of venues reported by the Foursquare API varied throughout the day. It is possible that there are other data sources that provide more comprehensive and consistent information about venues in a neighborhood, but this assignment required the use Foursquare. Nonetheless, despite some of the differences in the data returned by Foursquare with each call to the API, the overall results are rather consistent.
- Note: This study used Foursquare's "explore" endpoint. I also tested with the "search" endpoint, and that produced very similar results.

Conclusion

- This study shows that we can find many venue-wise similarities between the neighborhoods of Manhattan and the neighborhoods of San Francisco, so we can say "If you like neighborhood *A*, you might like neighborhood *B*."

- The scope of this study was to find similarity based on the available venues in each neighborhood. A comprehensive comparison of neighborhoods should include other factors, such as population demographics, cost of living, neighborhood "character" (urban or suburban, residential or mixed-use), and so on.