The Battle of Neighborhoods

Brendan Hargrove

Introduction & Problem Statement

Many people are forced to relocate for their careers and/or choose to move to a new city for better opportunities. However, most still have a home neighborhood they are fond of, or a favorite neighborhood from their hometown. This notebook will attempt to characterize the venues of a chosen neighborhood from one city and compare it to all neighborhoods of a different city, returning the neighborhoods with similar businesses, restaurants, etc. In this way, a user could implement this notebook to match a known neighborhood in one city with similar neighborhoods in a different city, thereby creating a “short list” of potential neighborhoods to investigate for housing, demographics, etc. However, the housing and demographic data are not within the intended scope of this notebook.

Beyond the personal use of this notebook, finding a neighborhood in a new city that is most similar to a favorite neighborhood back home, this workflow has business applications as well.

Say a company is looking to expand into a new city. By characterizing the area around an already high-performing location, and comparing that characterization against all neighborhoods in the new city, the business might be able to predict which new locations would be most likely to perform well. Similarly, by characterizing very low-performing locations and utilizing the same workflow, it might be possible to predict which neighborhoods should be avoided.

Description of Data

The data required for this project includes: neighborhood names for Houston, TX and Seattle, WA, location information (latitude-longitude) for each neighborhood/city center, and a list of venues and venue categories. All of these data were available through public websites and/or online tools/repositories and required no previous cleaning, organizing, or processing.

Neighborhood names for Houston were obtained from https://en.wikipedia.org/wiki/List_of_Houston_neighborhoods.

Neighborhood names for Seattle were obtained from https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle.

Location data for each neighborhood was gathered by iteratively feeding each neighborhood name into GeoPy Geocoder (https://geopy.readthedocs.io/en/stable/).

Once neighborhood locations were defined, each lat-long point was queried in Foursquare (https://developer.foursquare.com/) for venue information.

In addition, the following python libraries were used in data import, processing, analysis, and display: Pandas, Numpy, urllib, BeautifulSoup, GeoPy, json, MatPlotLib, SKLearn, Folium.

Methodology

Data Acquisition & Processing

For this project, I will select a neighborhood from Houston, TX and a list of all neighborhoods from Seattle, WA. These will be merged into a single DataFrame and analyzed using venue data from Foursquare. I will then run a clustering algorithm on the DataFrame and return the cluster that the Houston neighborhood belongs to. Finally, I will produce a map that depicts all neighborhoods in Seattle within the same cluster as the input neighborhood from Houston.

First, the neighborhood data for Seattle was pulled from wikipedia and parsed into html using BeautifulSoup. A “for loop” was then employed to iterate through the html and parse the information into a dataframe. All unneeded columns from the html were dropped to create a clean dataframe containing neighborhood and district names for Seattle, WA.

Next, a new dataframe was created using only neighborhood names, and appending city and state information to them to create geographically-complete location names (for example, “Bitter Lake” became “Bitter Lake, Seattle, WA”). Location information for each neighborhood in our dataframe was gathered using GeoPy Geocoder’s “Nominatim” function by feeding the names iteratively through Nominatim, acquiring the latitude-longitude coordinates, and appending them to the dataframe. Any neighborhoods that Nominatim could not locate or recognize were noted as an additional output of the loop.

```
Location information for North Beach / Blue Ridge not unavailable
Location information for North College Park
(Licton Springs) not unavailable
Location information for Portage Bay / Roanoke not unavailable
Location information for Pike-Pine Corridor / Pike/Pine not unavailable
Location information for International District ("ID") not unavailable
Location information for Central Area / Central District ("CD") not unavailable
Location information for Cherry Hill & Squire Park not unavailable
Location information for South End not unavailable
Location information for Dunlap / Othello not unavailable
Location information for Rainier Beach / Atlantic City Beach not unavailable
Location information for Mid Beacon Hill (Maplewood) not unavailable
Location information for South Beacon Hill / Van Asselt not unavailable
Location information for Industrial District not unavailable
Location information for North Admiral / Admiral District not unavailable
Location information for Junction / West Seattle Junction / Alaska Junction not unavailable
Location information for Seaview / Mee-Kwa-Mooks not unavailable
```

	Neighborhood	Latitude	Longitude
0	North Seattle	47.660773	-122.291497
1	Broadview	47.722320	-122.360407
2	Bitter Lake	47.726236	-122.348764
3	Crown Hill	47.694715	-122.371459
4	Greenwood	47.690981	-122.354877
5	Northgate	47.713153	-122.321231
6	Haller Lake	47.719748	-122.333751
7	Pinehurst	47.603832	-122.330062
8	Maple Leaf	47.693987	-122.322905
9	Lake City	47.719162	-122.295494

A brief comparison of the dataframe length before and after running Nominatim shows that out of 127 initial neighborhoods, 111 were successfully geolocated. The remaining 16 neighborhoods were likely not recognized by Nominatim due to text formatting (multiple neighborhoods combined together, or additional notes/text included in the neighborhood name). However, 111 neighborhoods was deemed to be a significant enough number that further cleaning and formatting was not necessary.

I then selected my “input neighborhood” from a different city, to be compared and clustered against the neighborhoods in Seattle. For this example, I used Houston Heights, a trendy neighborhood in Houston, TX known for its parks, coffee shops, restaurants, and general walkability. Since my input neighborhood was a single object, iterating a list through Nominatim was not necessary. I created a new dataframe for Houston Heights, fed the address into Nominatim, and appended the location information to the dataframe. Finally, I used the Pandas concatenate function to merge the Houston Heights dataframe with the Seattle dataframe. The resulting dataframe

includes the input neighborhood (Houston Heights) and all neighborhoods in Seattle, along with location information for each.

	Neighborhood	Latitude	Longitude
0	Houston Heights, Houston, TX	29.797687	-95.398446
1	North Seattle	47.660773	-122.291497
2	Broadview	47.722320	-122.360407
3	Bitter Lake	47.726236	-122.348764
4	Crown Hill	47.694715	-122.371459

Venue Acquisition Using FourSquare

Each neighborhood location was fed iteratively into Foursquare using the Venues/Explore endpoint and a “Get” request was made for the “groups” and “items” response fields, limited to 100 entries per neighborhood within a 1000 meter radius of the location. The results were placed into a new dataframe and reordered such that our input neighborhood, Houston Heights, was placed at the top.

Venue data for each neighborhood were one-hot encoded by category using Pandas “get_dummies” tool, grouped by neighborhood, and averaged to obtain a weighted value of the occurrence of each venue category within each neighborhood radius. Finally, I defined a function that gathered the 20 most frequently-occurring venues for each neighborhood and ranked them by prevalence. These “most common venues” were the basis of neighborhood clustering. In preparation for clustering, neighborhood names were removed.

K-Means Clustering

I chose to utilize k-means clustering for this project, as it was the most familiar. Each neighborhood’s 20 most frequent venue types were used to characterize clusters of “like neighborhoods” based on order of occurrence. Initially, the algorithm was run using a very small cluster count (6 clusters), but the resulting cluster groups were too large to be useful. I reassigned the number of clusters to 25 in an attempt to drastically limit the number of neighborhoods within each cluster. Finally, I defined a new dataframe limited to just those neighborhoods within the same cluster as our input neighborhood, Houston Heights.

Mapping

The final step in this analysis was to generate a map of Seattle, WA highlighting all neighborhoods within the same cluster as our input neighborhood. First, I used Nominatim to acquire the location of Seattle. This served as the center coordinates of my map. I generated a map using Folium, centered on Seattle, WA, and added markers at the coordinates of each neighborhood within my final cluster. I then added labels for each neighborhood name.

Results

The initial list of neighborhoods in Seattle consisted of 127 objects. Many of these are also larger district names, with smaller neighborhoods differentiated within them. However, the centers of these districts do not correspond to the centers of smaller neighborhoods, and I therefore deemed to leave all entries in the dataframe. After searching for location data using Nominatim, not all locations were recognized by GeoPy. The remaining neighborhoods, those with geolocation data, were 111. After adding the input neighborhood, the final dataframe contained 112 entries.

After running the neighborhoods dataframe through Foursquare and gathering the associated venues, the top 20 most frequently-occurring venues were ranked for each neighborhood. The top 20 venues for the input neighborhood, Houston Heights, are noted below.

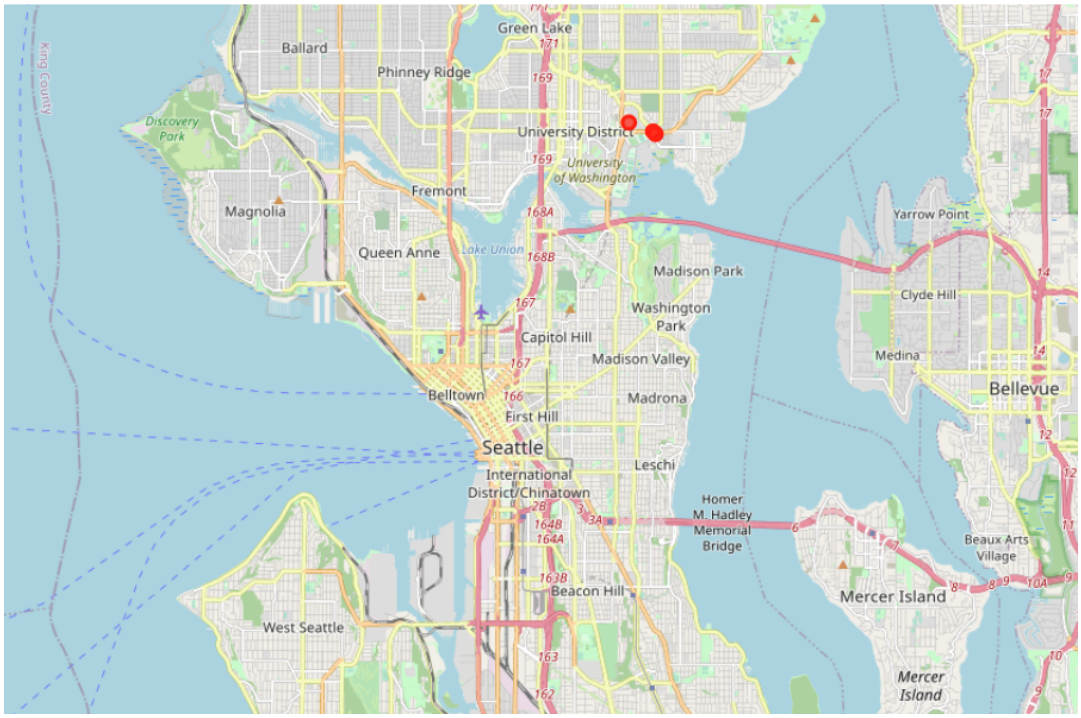
Neighborhood	Houston Heights, Houston, TX
1st Most Common Venue	Coffee Shop
2nd Most Common Venue	Burger Joint
3rd Most Common Venue	Park
4th Most Common Venue	Pharmacy
5th Most Common Venue	Flower Shop
6th Most Common Venue	Furniture / Home Store
7th Most Common Venue	New American Restaurant
8th Most Common Venue	Thrift / Vintage Store
9th Most Common Venue	Mexican Restaurant
10th Most Common Venue	Trail
11th Most Common Venue	Diner
12th Most Common Venue	Pet Store
13th Most Common Venue	Sandwich Place
14th Most Common Venue	Gift Shop
15th Most Common Venue	Cosmetics Shop
16th Most Common Venue	Spa
17th Most Common Venue	Pizza Place
18th Most Common Venue	Italian Restaurant
19th Most Common Venue	Movie Theater
20th Most Common Venue	Indian Chinese Restaurant

After analyzing the data using k-means clustering and 25 defined clusters, the input neighborhood fell into cluster 24, with three neighborhoods/districts from Seattle. Cluster 24, including the four neighborhoods and their associated characteristic venue categories, are included below.

	Neighborhood	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	...
0	Houston Heights, Houston, TX	29.797687	-95.398446	24	Coffee Shop	Burger Joint	Park	Pharmacy	Flower Shop	Furniture / Home Store	...
1	North Seattle	47.660773	-122.291497	24	Coffee Shop	Furniture / Home Store	Clothing Store	Women's Store	Arts & Crafts Store	Pizza Place	...
25	University District (U District)	47.661191	-122.292083	24	Coffee Shop	Furniture / Home Store	Clothing Store	Arts & Crafts Store	Pizza Place	Italian Restaurant	...
26	University Village	47.662740	-122.298925	24	Pizza Place	Arts & Crafts Store	Burger Joint	Thai Restaurant	Coffee Shop	Italian Restaurant	...

Discussion

The neighborhoods that fall within Cluster 24, the same cluster as Houston Heights, all occur within a small area of Seattle centered around the University of Washington. Although Houston Heights is not near a university itself, it shares many qualities typical of neighborhoods near college campuses. These qualities are well illustrated in the frequency occurrence of venues within Cluster 24. Coffee shops, parks, casual shopping, and American dining venues all feature prominently within these neighborhoods.



Conclusion

In conclusion, I have quantitatively demonstrated that the nieghborhoods in Seattle most similar to Houston Heights are those located immediately around the University of Washington, within University VIllage and University District. A former resident of Houston Heights, if wishing to find neighborhoods with familiar amenities, would want to focus their search for housing within this area. Similarly, if a business with a successful location within Houston Heights was looking to expand into a new market, I might suggest these neighborhoods as likely contenders.

There is quite a lot of additional work that could be done to expand this project. For example, the number of clusters could be decreased to group additional areas into the cluster with our input neighborhood. Additional analysis could be done to statistically determine the ideal cluster count to minimize inter-cluster error. Furthermore, additional data cleaning could be done to exclude any neighborhoods that also occur as recognized districts, resulting in a more consistent clustering match. Finally, additional cities could be brought into the analysis to look for better matches in other cities, or cities near Seattle within commuting distance. All of these workflows are outside the scope of this project, but will likely be pursued separately. Thank you for reading my report.