Lilia Staszel lstaszel@mit.edu
Erica Zhou ezhou@mit.edu

Neighborhood Metrics: Mid-Term Report

Status

Our project aims to determine important features of a neighborhood and cluster similar ones in different cities by highlighting metrics from various datasets. Thus far, we have focused on New York City neighborhoods and compiled a set of qualitative and quantitative features. We hope to cluster neighborhoods based on these features, and if this goes well, build a similar dataset for another city to draw parallels and discover defining neighborhood characteristics.

Completed Tasks

We have done most of the initial data processing for New York City neighborhoods. We have identified 147 neighborhoods across the city and created a table synthesizing many datasets to capture both qualitative and quantitative aspects of each neighborhood. We have also created visualizations of the data to give a rough view of immediate insights.

To begin, we developed a unified “neighborhood” identifier by combining Airbnb’s neighborhood map with New York’s NTA (neighborhood tabulation area) map. Deciding on a list of neighborhoods was difficult because there are many different definitions of the term and boundaries thus vary between datasets. For conflicting data, we defined the neighborhood with larger area as the “neighborhood” and pieced together a similar area for the other dataset by combining multiple neighborhoods. Although not perfect, this resolved most conflicts without having to combine too many large/well-known neighborhoods.

We split the initial data collection, cleaning, and remapping of features. Erica worked on parsing the Airbnb neighborhood description data to build a word frequency mapping to capture the qualitative feel of a neighborhood. Lilia worked on collecting quantitative metrics such as demographic breakdown, crime and public health rates, and population density. We joined these two datasets with the common identifier.

Future Tasks

Moving forward, we will both work on developing clustering techniques and iterating on our dataset to improve the clustering. Initially, Erica will focus on qualitative traits and Lilia on quantitative traits, but we will work together on the combined dataset. If this goes well, we will try to extend our analyses to other cities, in which case we will each take on another city and come together to do the clustering.

Completed Deliverables

We currently have a dataset of qualitative and quantitative features of NYC neighborhoods. The qualitative features are words with corresponding frequencies in Airbnb users’ neighborhood descriptions. Currently, we have about 15 candidate words, but they can be adjusted based on which are most useful for clustering. Our 55 quantitative features include items like number of hospitals, population, and prevalence of different health conditions.

Additionally, we built a few preliminary visualizations to see the variation of features between different neighborhoods, as seen in figures (1) and (2).

Lilia Staszel lstaszel@mit.edu
Erica Zhou ezhou@mit.edu

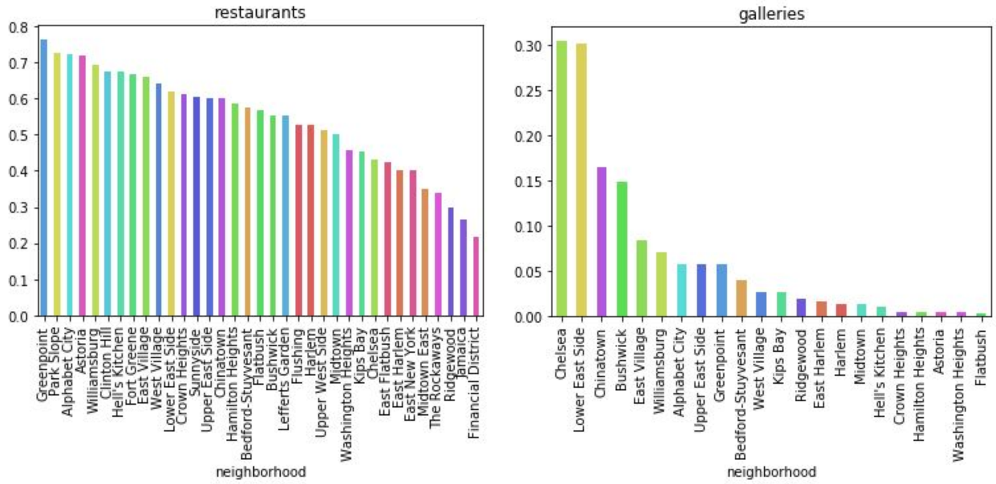

Figure 1. Normalized frequency of ‘restaurants’ and ‘galleries’ among Airbnb hosts’ neighborhood descriptions.

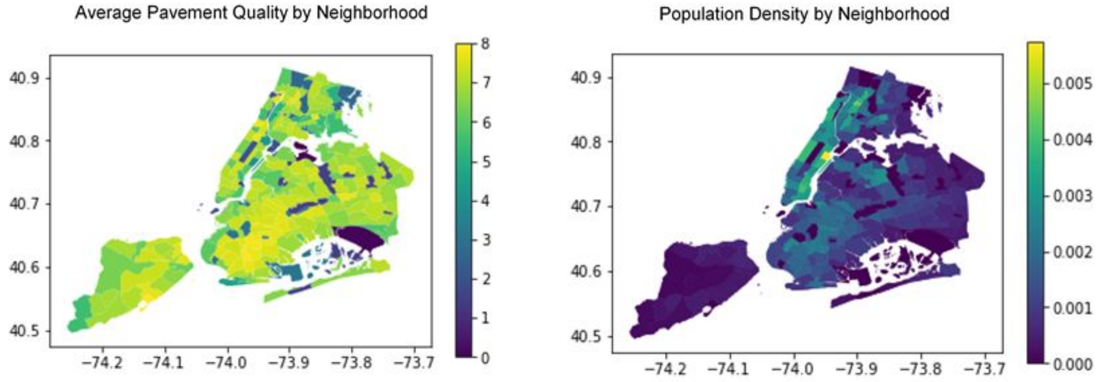

Figure 2. Heatmaps depicting quantitative features of various neighborhoods.

Potential Problems

One problem that we resolved for New York but that would reappear for another city was defining a “neighborhood” since the term is essentially arbitrarily defined.

Another difficulty we encountered was in defining features that are useful and usable. For example, some neighborhoods only have a few airbnb reviews, so we will have to decide whether to keep those in our dataset or remove them altogether. We may also have to iterate upon our original features to pull out the most notable ones.

Finally, because we did lose a team member, we will likely have to scale our project accordingly. This may mean scaling down the number of cities we use/amount of data we process or potentially changing the visualization component of the project.

Timeline

For the most part, we have stuck to our timeline so far. We changed it to begin working on clustering earlier so that we have more time to iterate upon our selected features and strategies. Additionally, we want to focus more on having an extendable clustering algorithm among multiple cities, so we have simplified the visualization component and added more time for extending the dataset to more cities.

Bold italics - accomplished/changed

Time	Goal
Week of Oct 14	Pick feasible cities & features (based on dataset quality, outside knowledge) ***Picked NYC - Airbnb & NYC gov datasets***
Week of Oct 21	Choose specific datasets & features: 1 feature/person for 1 city ***Erica: qualitative features*** ***Lilia: quantitative features***
Week of Oct 28	Have preliminary feature measures (start with more concrete features) ***Have joined dataset of quantitative and qualitative features***
Week of Nov 4 *(midterm eval)*	~~Evaluate success of metrics so far, decide on new features worth looking into (potentially more subjective ones - tweets)~~ ***Build clustering model with preliminary features, evaluate & decide if different/more data necessary***
Week of Nov 11	~~Analyze further features & begin looking at correlations between neighborhoods~~ ***Collect more data if necessary, refine clustering model for NYC***
Week of Nov 18	~~Build clustering model using best features, think about how to visualize~~ ***Try to build a similar dataset for other cities with similar available data***
Week of Nov 25	~~Refine model, build visualization~~ ***Run clustering on multiple cities and refine model given new data***
Week of Dec 2 *(project report)*	Have final ~~visualization~~/model ***Develop visualization & human-readable descriptions for neighborhood clusters***
Week of Dec 9 *(poster presentation)*	Refine visualization, make poster