Lilia Staszel lstaszel@mit.edu
Erica Zhou ezhou@mit.edu
Lindsey McAllister lindseym@mit.edu

Neighborhood Metrics

Abstract
People commonly make comparisons between different cities or even neighborhoods within cities, but how much of these comparisons is grounded in fact versus just “feel” of an area, and can we capture this “feel” quantitatively? In our project, we plan to determine important features of a neighborhood and cluster similar ones in different cities by developing standardized metrics from various datasets. Using data from many sources such as retail, housing/ Airbnb, restaurants, social media, we want to create interesting, innovative metrics and eventually categorize different types of neighborhoods.

Intro
Our project aims to analyze neighborhoods by combining multiple datasets and defining metrics to help with classification. We eventually hope to determine similar neighborhoods across large cities. This could be useful for tourists who are planning a trip, someone who is looking to move, or local governments developing policies. It also can give insights about how neighborhoods compare within a city and overall trends in specific factors. There are existing reports of neighborhoods in cities, but often the metrics are not standardized or comparable across cities. With our project, these cross-city comparisons will be possible. We will also pull data from interesting/ unique sources such as Airbnb and social media that these government reports do not typically have. Our hypothesis is that different cities have similar neighborhoods that fall into distinct categories.

Methodology
To begin, we plan to think about common “defining” traits of neighborhoods that we can quantify, as well as a few cities that seem to have adequate data. In order to get a quantitative measure of how neighborhoods fall on different scales (ex. bustling, “homey”, safe), we plan to combine data from a few different datasets. For these measures, we can get some quantitative measures (ex. restaurants per square foot, population), but we can also use more subjective features (ex. restaurant or airbnb reviews, tweets) and combine these into some kind of quantitative score. Once we have a listing of neighborhoods with scores on various topics, we can build a machine learning algorithm to try to cluster different neighborhoods based on similar scores, and it would be interesting as well to “backtrack” and figure out if there are common features or strong deciding factors for placing different neighborhoods in their clusters. If we have time, we could also build a visualization to convey this information in a more user-friendly manner.

Evaluation
To measure our success, we will compare our model’s results and its most important features with personal experience and other researchable information. If necessary, we will do manual research into specific neighborhoods, and ideally, we will be able to pull out defining features of specific neighborhoods and neighborhoods in general.

Data

We are planning to use publicly available data from a variety of sources: Airbnb, Yelp, and local governments. All of these entities make large volumes of location-based publicly available. A large portion of the work for this project will be collecting, cleaning, and integrating these datasets together so that they are standardized and comparable across cities and neighborhoods. Lat/long information will be useful for joining datasets, although we will need to determine the “neighborhood” that each business, business, etc. is in. Additionally, we need to figure out how to quantify some of the features, as well as how to integrate different features to build a model.

Task List & Timeline

The plan for dividing up work in the beginning of the project is to each take charge in investigating a specific dataset and set of features. We will meet weekly to discuss and take charge of more individual tasks.

Time	**Goal**
Week of Oct 14	Pick feasible cities & features (based on dataset quality, outside knowledge)
Week of Oct 21	Choose specific datasets & features: 1 feature/person for 1 city
Week of Oct 28	Have preliminary feature measures (start with more concrete 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)
Week of Nov 11	Analyze further features & begin looking at correlations between neighborhoods
Week of Nov 18	Build clustering model using best features, think about how to visualize
Week of Nov 25	Refine model, build visualization
Week of Dec 2 *(project report)*	Have final visualization/model
Week of Dec 9 *(poster presentation)*	Refine visualization, make poster

Deliverables

Our deliverable will be a representation, either a pictorial visualization or results from a clustering model, for each neighborhood in a city. A second deliverable would be a report of clusters of similar neighborhoods among cities.