Defining the “Essence” of City Neighborhoods

Erica Zhou, Lilia Staszel

Software Systems for Data Science (6.S080) - Fall 2019

Motivation

- Plenty of data is available about areas in cities, but boundaries and metrics are **not standardized** or comparable across cities.
- We analyzed neighborhoods by combining multiple datasets and defining metrics to **quantify the “essence”** of neighborhoods.
- Based on the collected data, we built **neighborhood clusters** based on features explorable via PCA.
- This dataset can be used to study **neighborhood patterns** through clustering and visualization, as well as to find similar areas **within and between cities**.

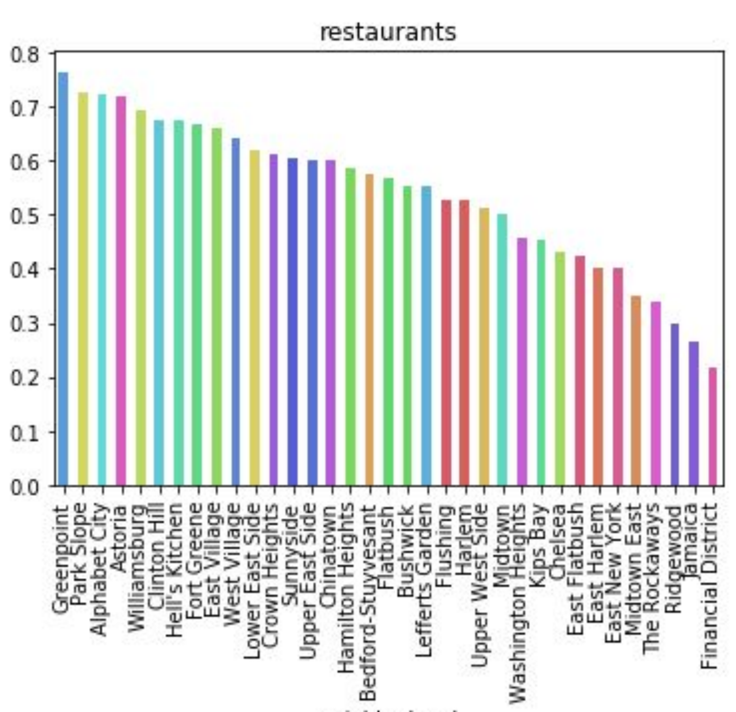
Data Sets

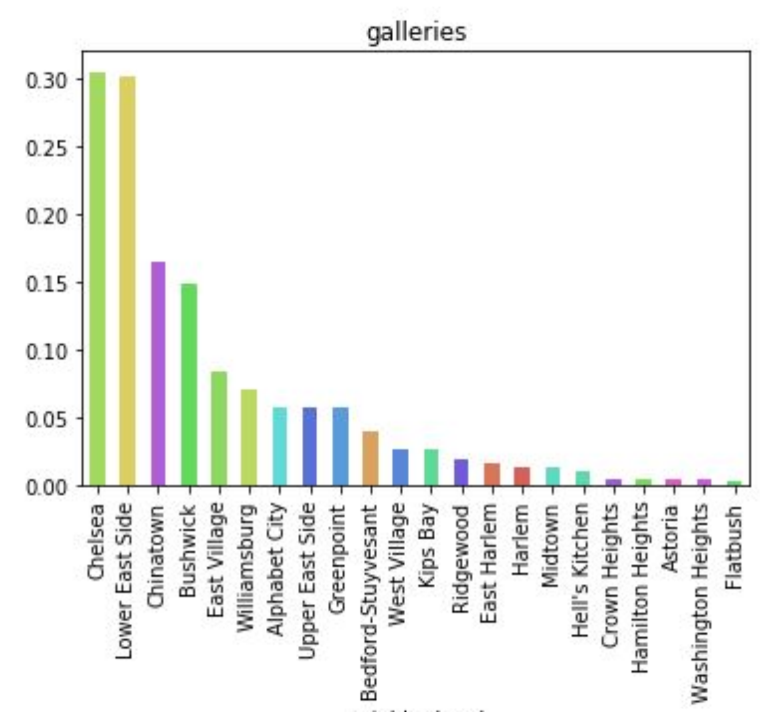
We utilized many datasets to create neighborhood profiles for New York City and Chicago.

Traditional Quantitative Features:

- Demographics
- Social and economic conditions
- Health outcomes
- Housing and neighborhood conditions

Airbnb Neighborhood Reviews:





Foursquare Top Venues:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
ALBANY PARK	Park	Coffee Shop	Bar	Convenience Store	Pizza Place	Middle Eastern Restaurant	Grocery Store
ARCHER HEIGHTS	Mexican Restaurant	Donut Shop	Taco Place	Fast Food Restaurant	Grocery Store	Pharmacy	Pizza Place
ARMOUR SQUARE	Chinese Restaurant	Bar	Pizza Place	Mexican Restaurant	Park	Grocery Store	Coffee Shop
ASHBURN	Discount Store	Grocery Store	Pharmacy	Fast Food Restaurant	Park	Pizza Place	Bank
AUBURN GRESHAM	Discount Store	Fast Food Restaurant	Grocery Store	Park	Sandwich Place	Seafood Restaurant	Pharmacy

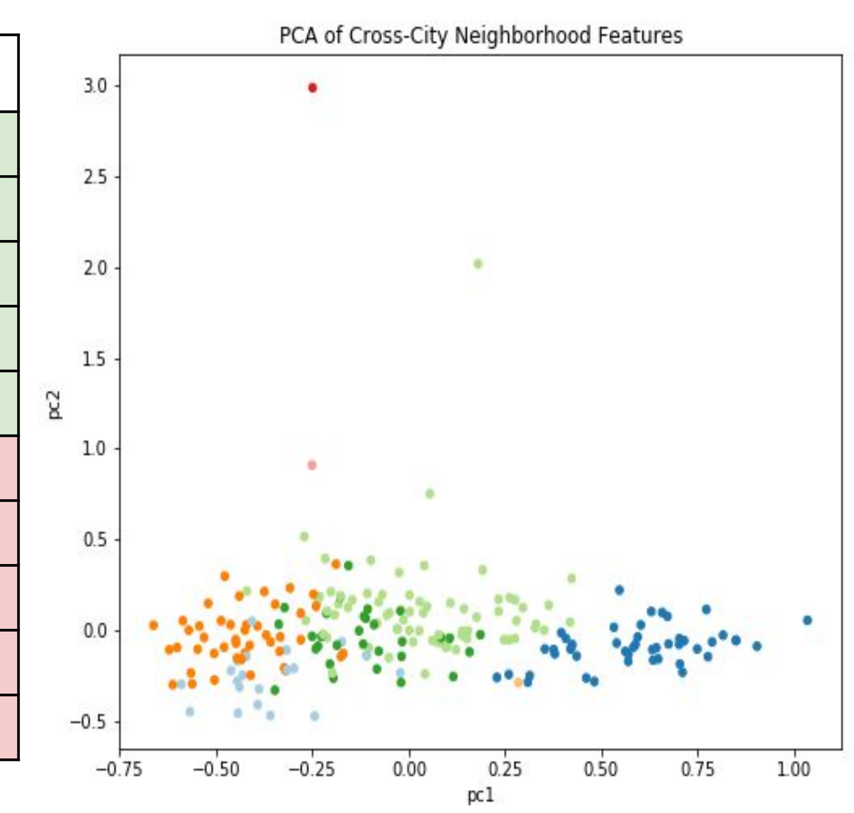
Data Processing

1. Define **neighborhood boundaries** for each city.
2. Collect **“traditional” metrics** from datasets and standardize values across neighborhoods, determining which are comparable between cities.
3. Generate **qualitative features** by calculating relative frequencies of common words in neighborhood descriptions on Airbnb listings.
4. Get common **venue types** by neighborhood boundary from Foursquare API.
5. Combine all features for 3 **finalized datasets**:
   i. 146 New York neighborhoods and 685 features
   ii. 77 Chicago neighborhoods and 722 features
   iii. Neighborhoods from both cities and comparable features between them

PCA & Clustering

We utilized **PCA** and **k-means clustering** to determine similar neighborhoods and features that drew them together.

PC1 (0.1093)	**PC2** (0.06820)
restaurants	walking
bars	eat
Coffee Shop	pizza
shops	local
Café	restaurant
quiet	bars
Discount Store	shopping
Fast Food Restaurant	Park
Pharmacy	Discount Store
Donut Shop	museum



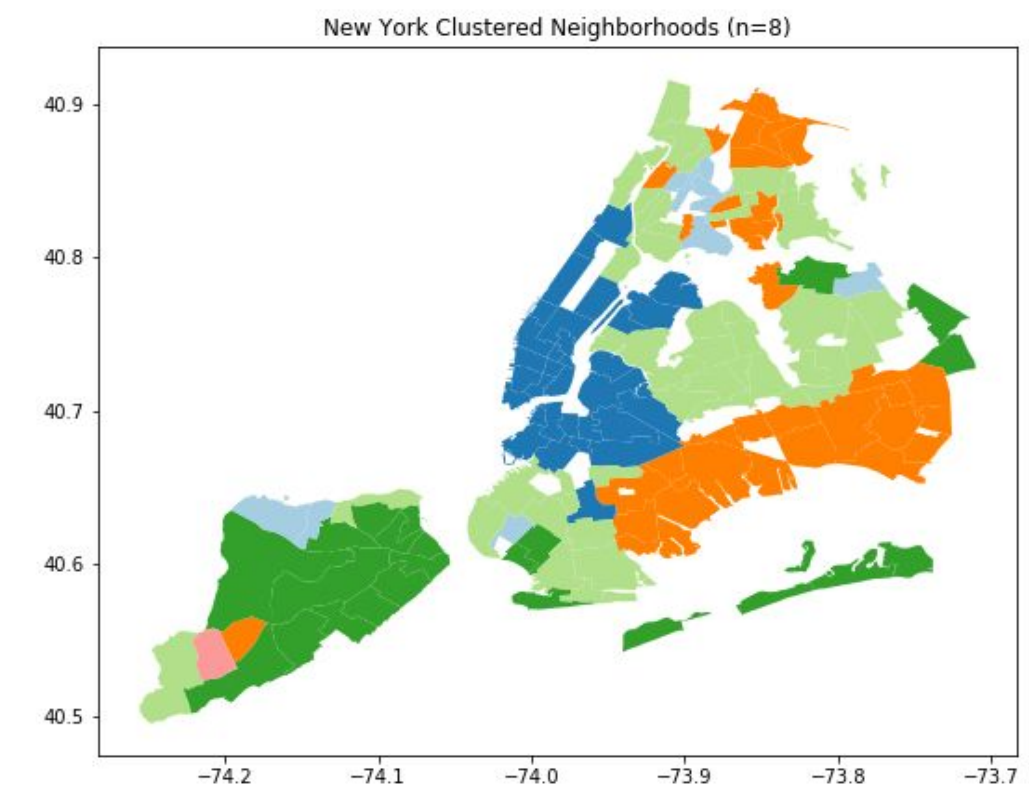
The PCA coloring corresponds to the k-means neighborhood clusters depicted to the right.

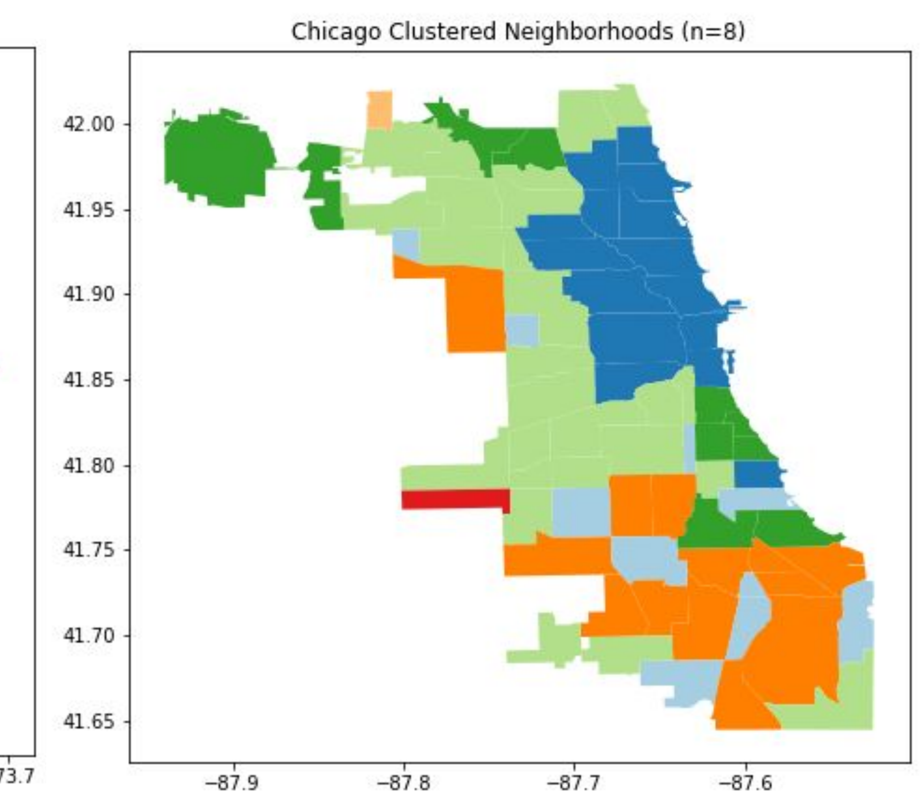
Challenges

- Neighborhood **boundaries** are not standardized
  - Combining/aggregating data
- Conversion of **qualitative data** to quantitative features
  - Attempted tf-idf
  - Hand-pruned stopwords & selected word features
- Effective **feature selection**
  - Lots of features: some strongly correlated
  - Data transformation/PCA
- **Visualization/presentation** of all forms of data collected
  - Balance of information & explainability

Results

As a final step, we built a dataset of **common qualitative traits** (from Airbnb reviews & Foursquare venues) for New York and Chicago and ran a **combined clustering algorithm** to look for revelatory features that spanned both datasets.





- The algorithm successfully clustered the **downtown areas** of Chicago and New York (dark blue), and these areas are most highly correlated with presence of restaurants, bars, and shops.
- As we would expect, ***pc1* also decreases** as we move further away from downtown to more suburban areas.
- However, the **variance explained** by the principal components is still fairly low, suggesting that we need to further refine our selected features to improve the overall clustering.

Conclusion & Future Work

- Full datasets provide the opportunity to explore **variegated aspects of urban life** by adding features not immediately quantifiable by demographic information.
- Fine tuning of features/**dimensionality reduction** would allow for better clustering of neighborhoods.
- Refined **visualization tools** would make it possible to query for specific features or neighborhoods that share similar qualities.
- Adding **more cities** would give a more holistic view of features that really help to distinguish the “essence” of a neighborhood.

References & Data Sources

Data Sources:

- NYC Open Data (data.cityofnewyork.us)
- Chicago Data Portal (data.cityofchicago.org)
- Inside Airbnb (insideairbnb.com)
- Foursquare Places API (developer.foursquare.com/places)

References:

- https://medium.com/@shaikzia/segmenting-clustering-neighborhoods-in-london-city-faeac0715d99