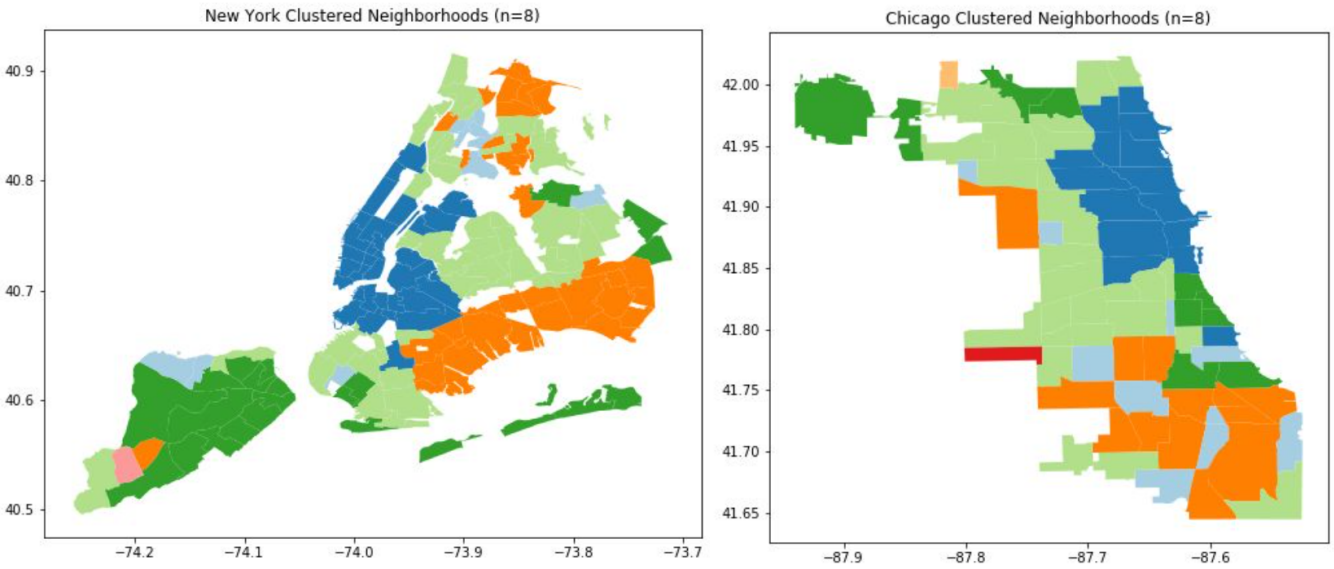
Defining the "Essence" of City Neighborhoods

Erica Zhou
ezhou@mit.edu

Lilia Staszel
lstaszel@mit.edu



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 a subjective "feel" of an area, and can we capture this neighborhood "essence" quantitatively? This project aims to build a dataset of important distinguishing 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, and more traditional demographics, we create interesting, innovative metrics and eventually categorize different types of neighborhoods.

1. INTRODUCTION

The project analyzes neighborhoods by combining multiple datasets and defining metrics to help with classification. We begin by creating a dataset that can help 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 and studying urban development. It also gives 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 neighborhood boundaries are not standardized within a city and the metrics are not comparable across cities. With our new dataset, these cross-city comparisons will be possible.

We synthesize data from unique sources such as Airbnb and Foursquare as well as traditional sources like census demographics and government reports. Using Airbnb neighborhood descriptions, we classify city areas based on common words that are used in order to capture a metric of the qualitative features of a neighborhood. Additionally, we collect a large number of quantitative features to help create a bigger picture of each area. We then combine all the metrics into one dataset, with well defined neighborhoods and boundaries between them. For this report, we start with two major US cities, New York City, NY and Chicago, IL and collect comparable metrics for the two.

Our hypothesis is that different cities have similar neighborhoods that fall into distinct categories. To test this hypothesis, we perform principal component analysis (PCA) and clustering of the neighborhoods within and across the cities and look for interesting insights.

2. RELATED WORK

This work builds upon some past work on clustering of city neighborhoods. For example, many people have used the Foursquare API to build neighborhood clusters based on venues located around the neighborhood. For example, people have published open-source work with data about various cities like London, Toronto, and Houston [5]. From the research setting, Preotiuc-Pietro et al clustered Foursquare information across 17 cities with a similar goal [4]. Similarly, Cao et al built a spatial modeling of cities to separate out regions within cities that serve different purposes [2]. From this research, an interesting question that arises is whether cities are best clustered by spatial methods or by venues or other features, and how similar the clusters generated by these methods are.

3. DATA COLLECTION

Before collecting data for each neighborhood, we had to standardize definitions for each neighborhood since the term is societally and sometimes differently defined between various sources. The final dataset we analyzed is a combination of three different datasets representing various qualitative and quantitative statistics about different neighborhoods in Chicago and New York.

3.1 Neighborhoods

We began by defining standardized neighborhoods for each city. For any city, there are many ways to define a neighborhood from census tracts to zip codes to historical "boundaries" that only locals are aware of. For New York City, we decided to go with a variant of the 195 Neighborhood Tabulation Areas (NTAs), that were defined by the city several years ago to "offer a good compromise between the very detailed data for census tracts (2,168) and the broad strokes provided by community districts (59)"[1]. Due to a few differences from the Airbnb data set, we ended up defining a new set of neighborhoods that combined the two, which will be discussed in detail in the following section. For Chicago, we used the 77 community areas defined on the city's data portal, which correspond with the Airbnb data.

3.2 Quantitative Data

We included some traditional quantitative features to supplement the qualitative data for each of the two cities. For the city of New York, this data was collected and aggregated from several different official datasets released by the city government. We collected a large number of factors from the NYC Neighborhood Health Atlas, a dataset that contains demographics (such as race, age, country of origin), social and economic conditions (education, poverty, disabilities), health outcomes (hospitalizations, premature mortality), health care (health insurance status, Medicaid enrollment), housing (density, rent burden), and neighborhood conditions (air quality, number of alcohol retailers, crime complaints) [3]. In addition, we used the NYC street pavement rating and pedestrian walk counts.

For Chicago, our quantitative features came from two city datasets. The first is from the Chicago Community Snapshot dataset, which includes information collected from several sources including the U.S. Census Bureau's 2013-2017 American Community Survey (ACS), the Illinois Department of Employment Security and the Illinois Department of Revenue. The second is a dataset of public health outcomes defined by the City of Chicago.

It is important to note that we only took a selection of data available for our datasets, and using the methodology and neighborhood boundaries we developed, more features for each neighborhood can be added to create an even more complete dataset for analysis.

3.3 Qualitative & Venue Data

We collected more non-traditional neighborhood features from two different sources: Airbnb and Foursquare. Our goal in utilizing these features was to capture more subjective aspects of a neighborhood or aspects not covered by purely demographic statistics. Airbnb users can supplement their listings with a "neighborhood description," and, to get a quantitative sense of these descriptions, we calculated the proportion of user listings that included certain words, as seen in figure 1.

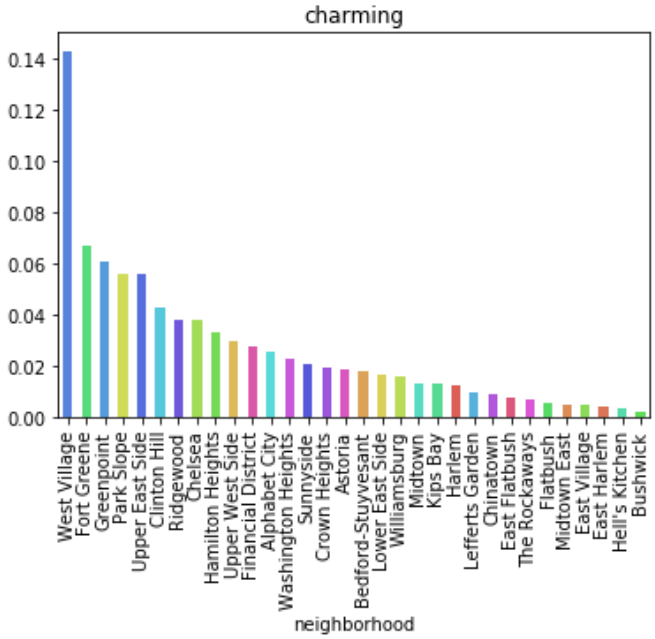


Figure 1: Proportion of Airbnb neighborhood descriptions utilizing the word "charming."

Foursquare provides an API that gives top (as defined by their platform) "venues" around a given location and defines "categories" for them (see figure 2). The final dataset includes a proportion of venues of each category in each neighborhood.

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

Figure 2: Top venue categories for a few neighborhoods in Chicago.

4. DATA PROCESSING

A few processing and cleaning steps were necessary to generate the final dataset. After defining neighborhoods, features were calculated for each neighborhood and combined to generate a consistent dataset for both cities. In sum, we collected 685 features for 146 New York City neighborhoods, 722 features for 77 Chicago neighborhoods, and 442 generalized features in a combined New York City/Chicago dataset. Jupyter notebooks containing all of the data processing steps can be found in the Github repo.

4.1 Neighborhood Definition

Due to the lack of a universal system for collecting, storing, and measuring neighborhood data within and across cities, a large amount of work went into cleaning and processing each of the datasets and synthesizing them to make fair comparisons between neighborhoods. One major difficulty was resolving differences between boundaries for Airbnb-defined neighborhoods and NTA neighborhood boundaries defined by the City of New York. We observed the differences and manually created a mapping between the two, with a few neighborhoods that were combinations of Airbnb or NTA neighborhoods. For Chicago, the Community Areas and Airbnb neighborhoods matched perfectly, so no new neighborhood definitions were needed.

4.2 Quantitative Data

Since we needed to combine some of the NTAs in New York, careful processing was necessary for the variables in the demographic datasets to combine them fairly. The features had a variety of units, such as "rate per 1000," percent, and "houses per square mile," so when combining, we made sure to multiply and add by correct proportions so no bias was introduced into our new dataset. Most of this was done manually by reading and understanding the definition of each variable and deciding how to combine it properly. In the final NYC dataset, we standardized the units so each of the variables were on similar scales and comparable across cities. For pavement quality, walking data and other location based data, we iterated through each NTA to check if the coordinates were located within the geographic boundaries for that neighborhood, and added that data to aggregated columns.

For the quantitative data for Chicago, no neighborhood combination was necessary, so most of the work involved making sure all the data was in comparable units, converting columns to proper data types and determining which columns were the most important to keep for our analysis.

4.3 Airbnb Neighborhood Descriptions

The Airbnb dataset included an entry for each Airbnb listing along with its neighborhood location as well as an optional "neighborhood description" field. For New York, neighborhoods were remapped to a new "Unique Identifier," but Chicago neighborhoods were left as is. For the text processing, we converted these descriptions to a dataset of common words and their frequencies in reviews. To begin, we tokenized descriptions into words and utilized Python's `nltk` to get generic English stopwords and manually curated of city-specific stopwords (ex. empire, state, building for New York since this landmark is location-specific). We removed stopwords, along with cleaning punctuation and capitalization, to get a final list of words. We calculated the frequency of each word in each neighborhood as:

$$freq = \frac{\#\ appearances\ in\ neighborhood\ entries}{\#\ entries\ for\ neighborhood}$$

This counts utilizing a word twice in the same review as two appearances, which may give an inflated frequency. However, this metric also gives a stronger signal for commonly-used and potentially important words. Final features were defined as words appearing in at least 1% of entries in the entire dataset to weed out location-specific terms and their frequencies per neighborhood. Additional features were also added for number of Airbnb listings and average price per person per night to generate the final dataset.

We also ran a term frequency-inverse document frequency (tf-idf) algorithm to try to generate features (results can be found at `NYC/tfidf.ipynb` in Github repo). However, the reviews include a large variety of words, which led to an extremely large dataset, and the results did not appear significantly better than the hand-pruned features, so we did not further pursue this route.

4.4 Foursquare Top Venues

The Foursquare top venues dataset closely follows the procedure detailed in [5]. Using a latitude/longitude location of each neighborhood (obtained from the geojson), we queried for the top 100 (limited by the free API) venues in each neighborhood, which returned names, locations, as well as a Foursquare-defined "category" for each venue. From this, one-hot encoding made it possible to get counts of venue categories, which were normalized to give a percentage of venues per neighborhood that fell within each category.

4.5 Merged Dataset

Once complete datasets for each city were created, we looked at which features between the two could be comparable for use in our cross-city analysis. This narrowed down the features a lot, since most outcomes were measured in different ways. We created three main datasets: one with New York neighborhoods and their full set of features, one with Chicago neighborhoods and their features, and one combined dataset with the comparable features for the two cities. Each dataset contained a combination of features from quantitative data, Airbnb, and Foursquare.

5. DATA VISUALIZATION

These datasets provide a new and standardized way at looking at features between different cities. However, on their own, they provide a lot of information without a lot of structure or clear organization. Thus, visualization and distillation of information also proved to be a large challenge for us. To explore and analyze these datasets, we built a few different visualization and clustering tools.

5.1 Folium Visualization

As a proof of concept, we created an interactive map for quick and simple visualization of different variables across neighborhoods. Because each variable is standardized (across population or area in a neighborhood) they can be easily compared using a choropleth map. We used Folium, a library that generates Leaflet maps in Python, to make a simple map with multiple layers that can be toggled on and off. For demonstration purposes, we chose ten variables from the dataset to display on the maps, but any of the hundreds of Airbnb, Foursquare, and traditional features can be added to the with just a single additional line of code. The maps are saved in html format, which can be opened in any browser. A sample of the map for New York City can be seen in Figure 3. Note that the blacked out neighborhoods are areas where there was not sufficient Airbnb description data to calculate a word frequency.

This style of mapping could be useful for someone who is trying to get a feel for the different areas of a city, and knows which features they care about. It is easy to make comparisons and spot patterns in the data. Two side by side

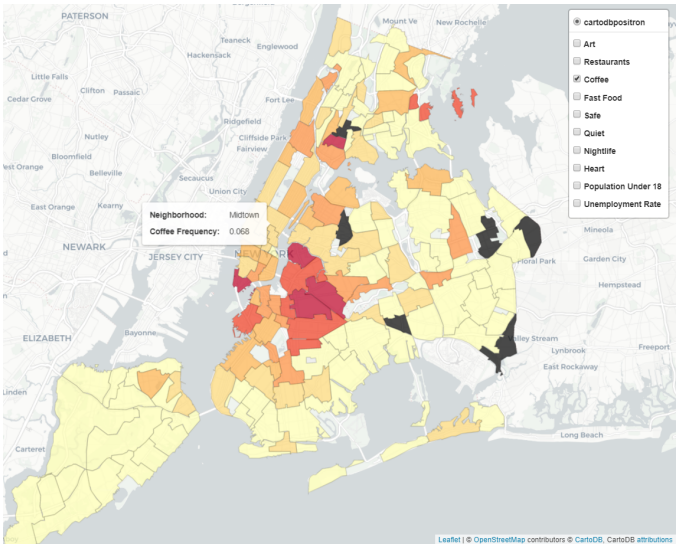


Figure 3: Interactive Folium map of New York City

maps can also be used to spot spatial correlations between certain features and gain insights from those.

5.2 K-Means Clustering

To determine similar neighborhoods within and among cities, we ran an unsupervised k-means clustering algorithm on different subsets of the data. For the number of clusters, we wanted enough to distinguish different areas of cities but few enough to pick up on similarities between neighborhoods. We also graphed the inertia (within-cluster sum of squares) for different numbers of clusters to look for a leveling-off, as seen in figure 4. Using this information, we decided upon 8 clusters.

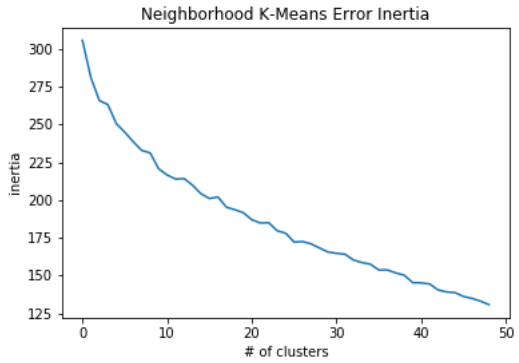


Figure 4: K-means error for different numbers of clusters (full NYC/Chicago dataset)

The subsets of data we analyzed included:

- full dataset for one city
- quantitative data only for one city
- qualitative (Airbnb/Foursquare) data only for one city
- full dataset with common features in both cities
- qualitative data only for common features in both cities

each of which yielded differing but interesting results.

One important processing step for performing k-means clustering and PCA was to normalize data values so that the importance of variables was not incorrectly determined just by their magnitudes. To do this, we normalized all values that were not already percentages to a range of 0-1. We did not further adjust the variances of the features to prevent artificial inflating of features that did not have that much variance or importance to begin with.

5.3 PCA

In addition to creating clusters of similar neighborhoods, we wanted a way to visualize clusters and the features that were responsible for clustering them together. Accordingly, we implemented principal component analysis (PCA) as a way to reduce the dimensionality of the data and quickly distill out features that were most revelatory towards defining neighborhoods. Specifically, we looked at the features and coefficients associated with the first few principal components, which accounted for the greatest amount of variance in the datasets.

To visualize the results of PCA, we plotted the first and second principal components for each neighborhood, colored by its k-means cluster. For example, the qualitative features for the cross-city dataset yielded the following principal components and clustering:

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

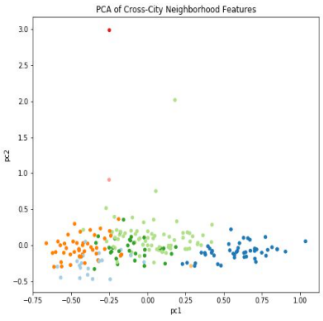


Figure 5: PCA graph and largest magnitude principal component features. Green features contribute most positively, and red features contribute most negatively.

As seen above, the first two principal components do a decent job of separating neighborhood clusters, and there is interestingly one high outlier in *pc2*. In terms of explainability, dark blue neighborhoods are highest in *pc1*, suggesting that these neighborhoods have the highest concentration of restaurants, bars, and shops.

5.4 Map Visualization

Since the data focuses on geographic areas, it makes sense to plot neighborhoods on a geographic map. Additionally, a map reveals whether there is a spatial aspect (without any explicit spatial features) to the clusters. An example of the neighborhoods corresponding to the PCA in figure 5 can be seen on page 1. The coloring on these maps corresponds to the same coloring as in the PCA graph (the k-means clusters with 8 clusters).

6. RESULTS & DISCUSSION

We tried several different clustering approaches on various subsets of features. Clustered city maps from the various datasets can be found in the `images` folders in the project Github repository, and full code can be found in `NYC/4-clustering.ipynb, Chicago/3-clustering.ipynb,` and `combined-cluster.ipynb`.

As one specific example, we will consider the combined data spanning both the New York and Chicago datasets. For this analysis, we focused on the dataset of common qualitative traits, which comprised 432 features. Clustering these features resulted in the PCA seen in figure 5 and the city maps pictured on page 1 and reveals a few interesting insights:

First, the algorithm successfully clustered the downtown/ most central areas of Chicago and New York, which are the darker blue clusters in each map. It is interesting to note that the vast majority of dark blue areas are also in the same area, as opposed to spread around. This suggests that although no individual feature contains any spatial information, the combination of features does create a sort of spatial awareness, and, in general, similar types of neighborhoods are spatially closer to each other, as opposed to there being neighborhoods of every type interspersed throughout the city. This is reasonable and expected to some extent, considering that oftentimes, neighborhoods do flow fluidly into each other and blend due to their spatial proximity.

From a mathematical perspective, the traits determined by PCA also provide a few more insights. Based on figure 5, it is clear that the blue areas are highest in *pc1*, which correlates most strongly with the presence of restaurants, bars, coffee shops, etc. Unsurprisingly, this corresponds to the downtown areas of both cities, the most concentrated areas of people as well as these types of venues. Further, as we move spatially further from downtown, the neighborhood colors generally move from dark blue to light green and then orange, which is exactly the same order as the decrease in *pc1* in the PCA graph. Again, this suggests a spatial pattern to neighborhood clustering: as we move further away from downtown into more suburban areas, *pc1*, representative of the amount of restaurants, bars, and similar venues, similarly decreases.

Although these preliminary insights provide some interesting ideas about neighborhood clustering, they also face a few limitations. In general, the variance explained by the principal components is still fairly low (upon preliminary clustering, 0.11 for *pc1* and 0.07 for *pc2*. Although enough variance was accounted for to generally draw out the downtown vs. more suburban areas, in order to get a more nuanced picture and more defined clusters, we need to further refine and narrow down our selected features to the ones that are most distinguishing within clusters and most discriminatory between different clusters.

Notably, adding back in quantitative features did significantly increase the explained variance ratio for the first few principal components, as well as further separate the clusters, as visible in figure **??**. However, in these datasets, the quantitative features predominantly focused on features related to socioeconomic status and race, which could lead to ethically questionable clustering metrics. For this reason, we focused more on the generally more positive Airbnb data, as well as venue data, for the majority of this analysis. In the future, though, it would be interesting to incorporate more

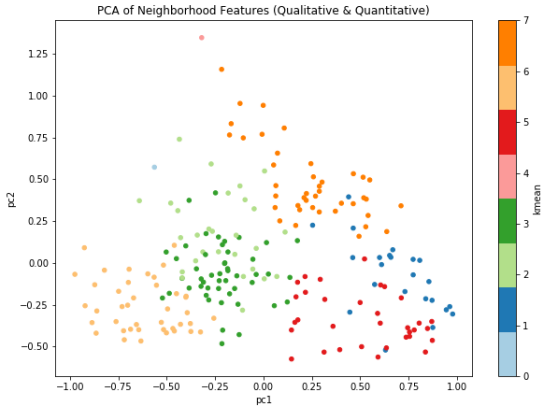


Figure 6: PCA for NYC/Chicago combined dataset, including quantitative features.

quantitative data since it does seem fairly promising for our clustering purposes.

Finally, looking at the map, one notable result is the presence of a few clusters of one neighborhood. For example, New York has one light pink neighborhood that Chicago does not have, and Chicago has a light orange and a red neighborhood that New York does not have. These singular clusters suggest that the corresponding neighborhoods are somehow different enough from any other neighborhood in either city to merit their own cluster. Although it is possible that this is true, the more likely scenario is that the features have not been refined enough to create "generalized" clusters among the two cities. Ways of rectifying this may include more selectively pruning for features or incorporating data from more cities to add more general features, which would limit the impact of individual outliers.

7. FUTURE WORK

Our full datasets provide the opportunity to explore various 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. There is also a lot of room for exploring different ways of clustering the data, or looking at clusters of an subset of the many features.

We also can expand our visualization tools to make it possible to query for specific features or neighborhoods that share similar qualities. This could be useful for someone who really likes the "vibe" of a neighborhood in New York City and wants to find something similar in a different city. Additionally if someone is searching for a place to live or stay in a city, they could use the data set to find neighborhoods that share the qualities they are looking for. Creating an easy to use tool for someone to query for similar neighborhoods could be an useful next step.

Also, adding more cities with the same set of standardized variables would give a more holistic view of features that really help to distinguish the "essence" of a neighborhood and could extend the analysis of interesting patterns in urban settings across the globe. It would also be interesting to

add some international cities and see if there are any notable differences in the features and clustering in countries outside of the United States.

Lastly, while we have hundreds of features for each city, there are still many other sources of data that we could use in the future to develop a fuller picture. One possible idea to use social media interactions from tweets or photos that are geo-tagged in specific neighborhoods. This could be another way to capture the kinds of qualitative data that add so much value to our clustering analysis.

8. CONCLUSION

While the combination of data sets and neighborhood definition was often tedious and frustrating, the final product is clean, easy to use and exciting. With the data sets we have created, this and so much is possible. Our clustering and PCA showed that cities have similar areas and demonstrated which features go into defining different areas of a city. We have only scratched the surface for analysis and insights possible with our data set. From travel, to housing purchases, to government policy and urban planning, having a unique and robust data set standardized across neighborhoods and cities allows for endless opportunities for learning more about cities. Through our careful and creative data collection, processing, and cleaning, we have made something useful for other projects in the future.

9. ACKNOWLEDGMENTS

We would like to acknowledge Professors Madden & Kraska, as well as our TAs Matt & Joana, for all of their help, feedback, and ideas throughout the development of this project.

10. REFERENCES

[1] Nyc open data. `http://www.opendata.cityofnewyork.us`. Accessed: 2019-09-30.

[2] Z. Cao, S. Wang, G. Forestier, A. Puissant, and C. F. Eick. Analyzing the composition of cities using spatial clustering. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, pages 14:1–14:8, New York, NY, USA, 2013. ACM.

[3] N. Y. C. D. of Health and M. Hygiene. New york city neighborhood health atlas. `https://www1.nyc.gov/site/doh/health/neighborhood-health/nyc-neighborhood-health-atlas.page`. Accessed: 2019-09-30.

[4] D. Preoţiuc-Pietro, J. Cranshaw, and T. Yano. Exploring venue-based city-to-city similarity measures. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, pages 16:1–16:4, New York, NY, USA, 2013. ACM.

[5] M. Zia. Segmenting clustering neighborhoods in london city. `https://medium.com/@shaikzia/segmenting-clustering-neighborhoods-in-london-city-faeac0715d99`. Accessed: 2019-12-01.

APPENDIX

The full code and instructions to run this project can be found at: `https://github.com/ericazhou7/neighborhood-similarity`.