CONTEXTUALIZING AN UNFAMILIAR CITY

Clustering LA then classifying NYC neighborhoods based on nearby venues

Background

- Moving to a new city is a difficult undertaking and can result in a lot of stress
- Especially when you like the facilities and venues in your current accommodation
- To make moving home easier, you need a reference point, where potential neighborhoods in the new city are compared to your current location
- Imagine a real-estate / rental-listing website (such as zillow or realtor) being able to contextualize the potential neighborhoods based on their similarity of surrounding facilities to your current accommodation

Procedure

- Los Angeles (USA) is selected as the current city and New York City (USA) is selected as the moving location. Then a three-step process is applied with the aim to make unfamiliar New York understandable by relating the neighborhoods to Los Angeles:

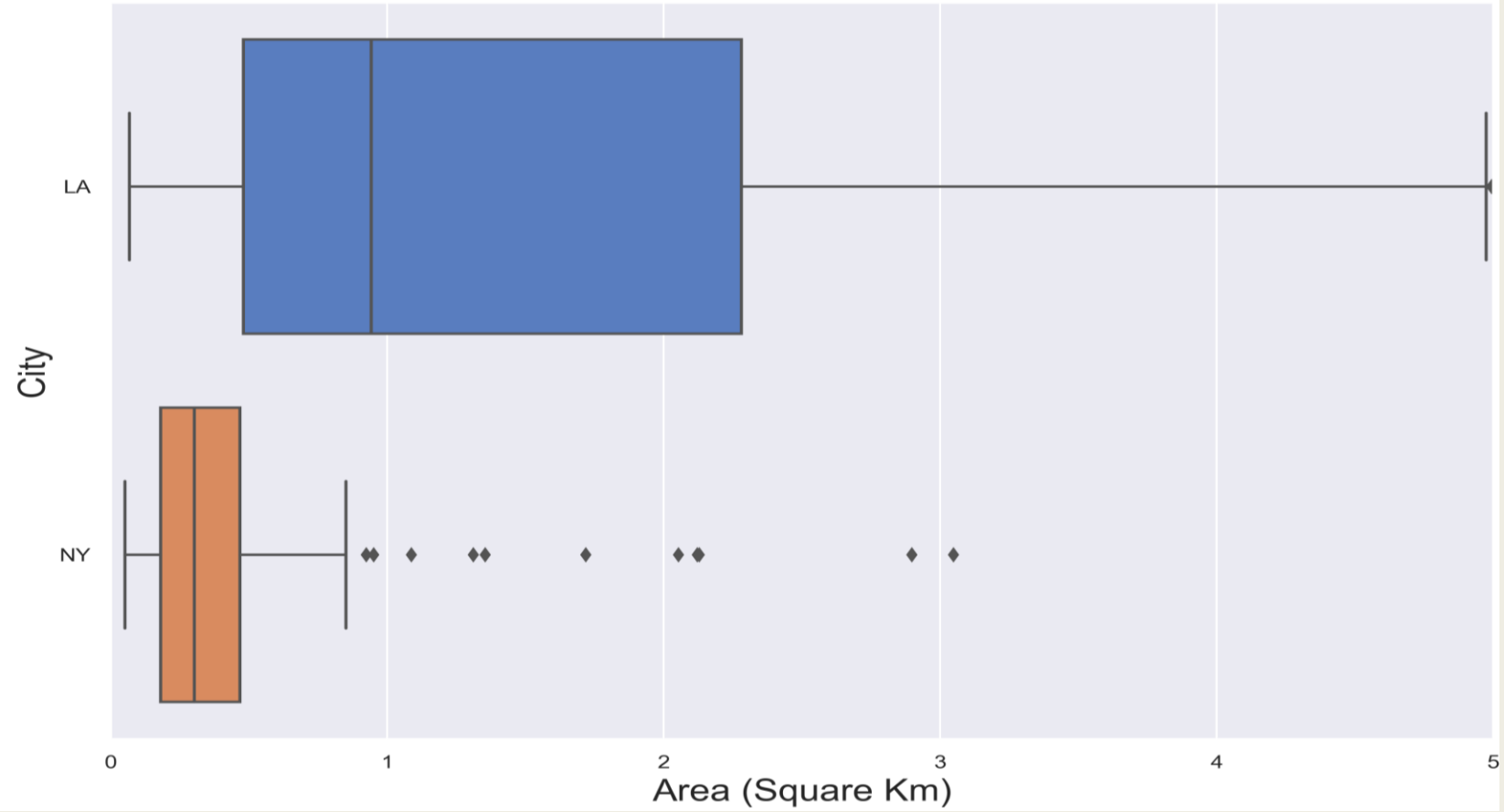
1. The neighborhoods of Los Angeles will be clustered based on their surrounding facilities.
2. These clusters will be used as labels to train a classifier.
3. The classifier will then be applied to New York City neighborhoods, providing context for the new city via analogy to the current residence.

Data

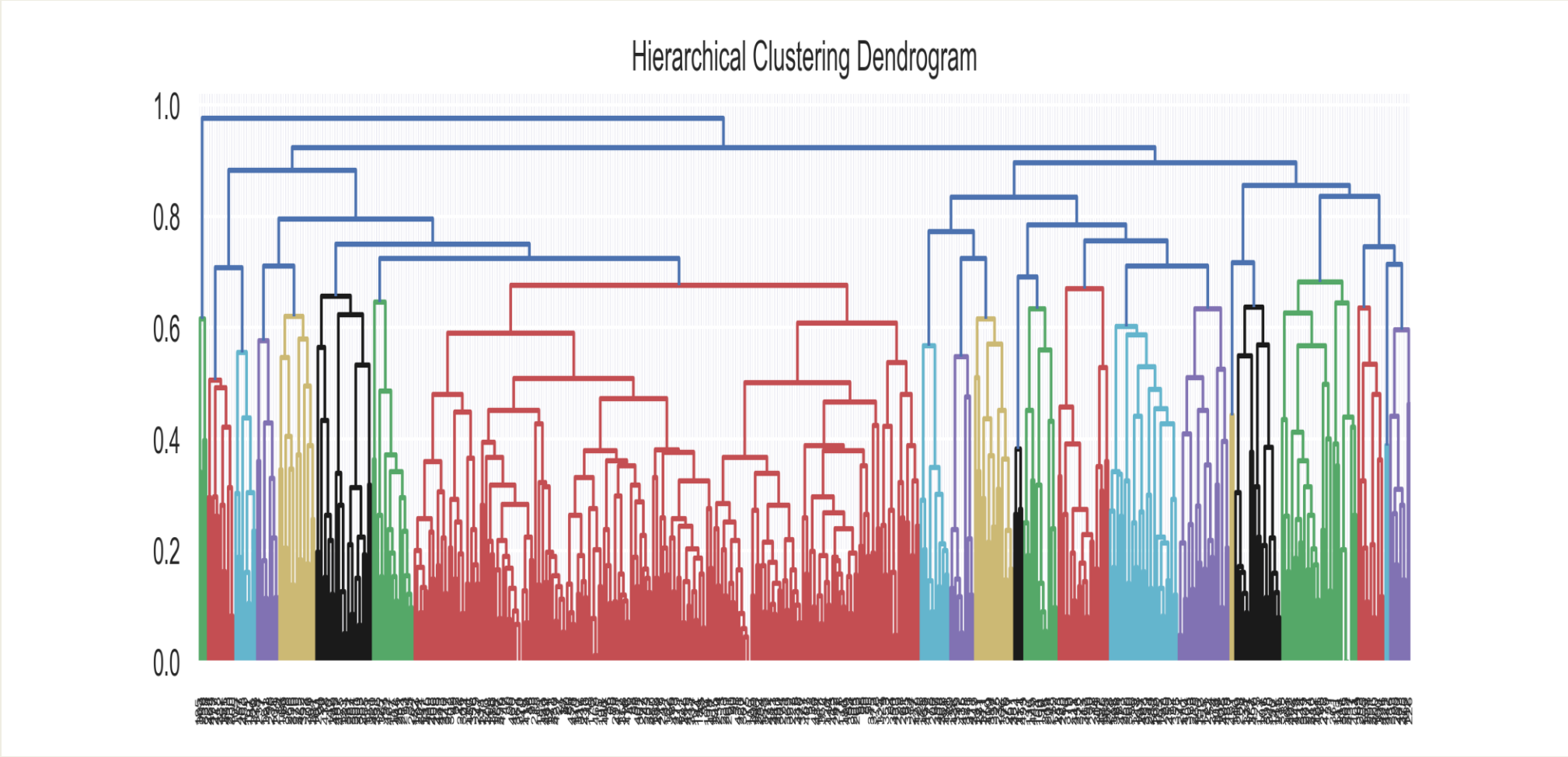
- Neighborhood names and boundaries, as defined by the latitude, longitude coordinates of bounding polygons of the will be sourced from Neighborhood Data for Social Change for Los Angeles and New York Open Data for New York City.

- The types of venues (categories) of the surrounding businesses and facilities in each neighborhood will be sourced from the Foursquare Places API.

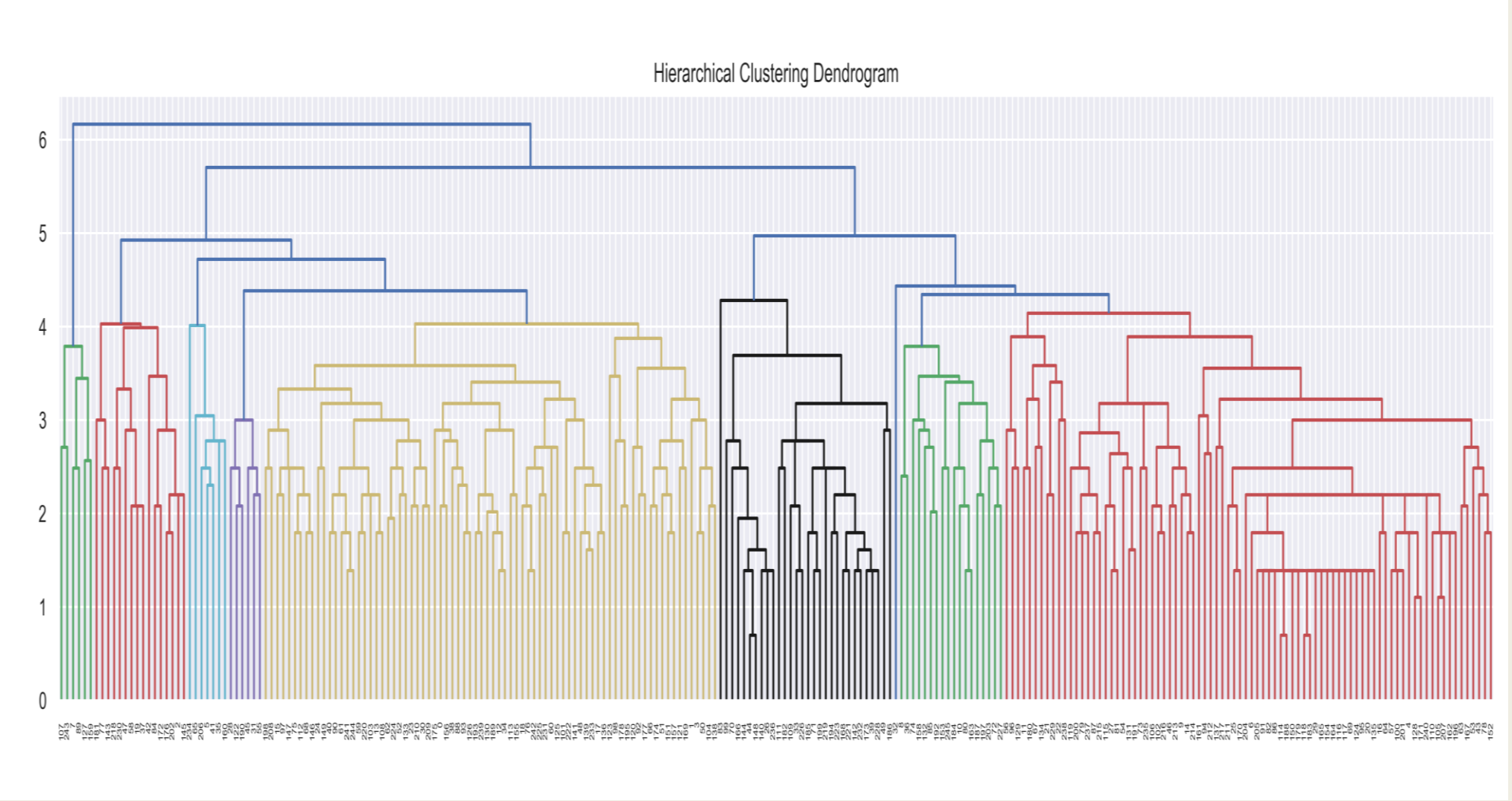
Exploration: Outliers in LA Neighborhood Area



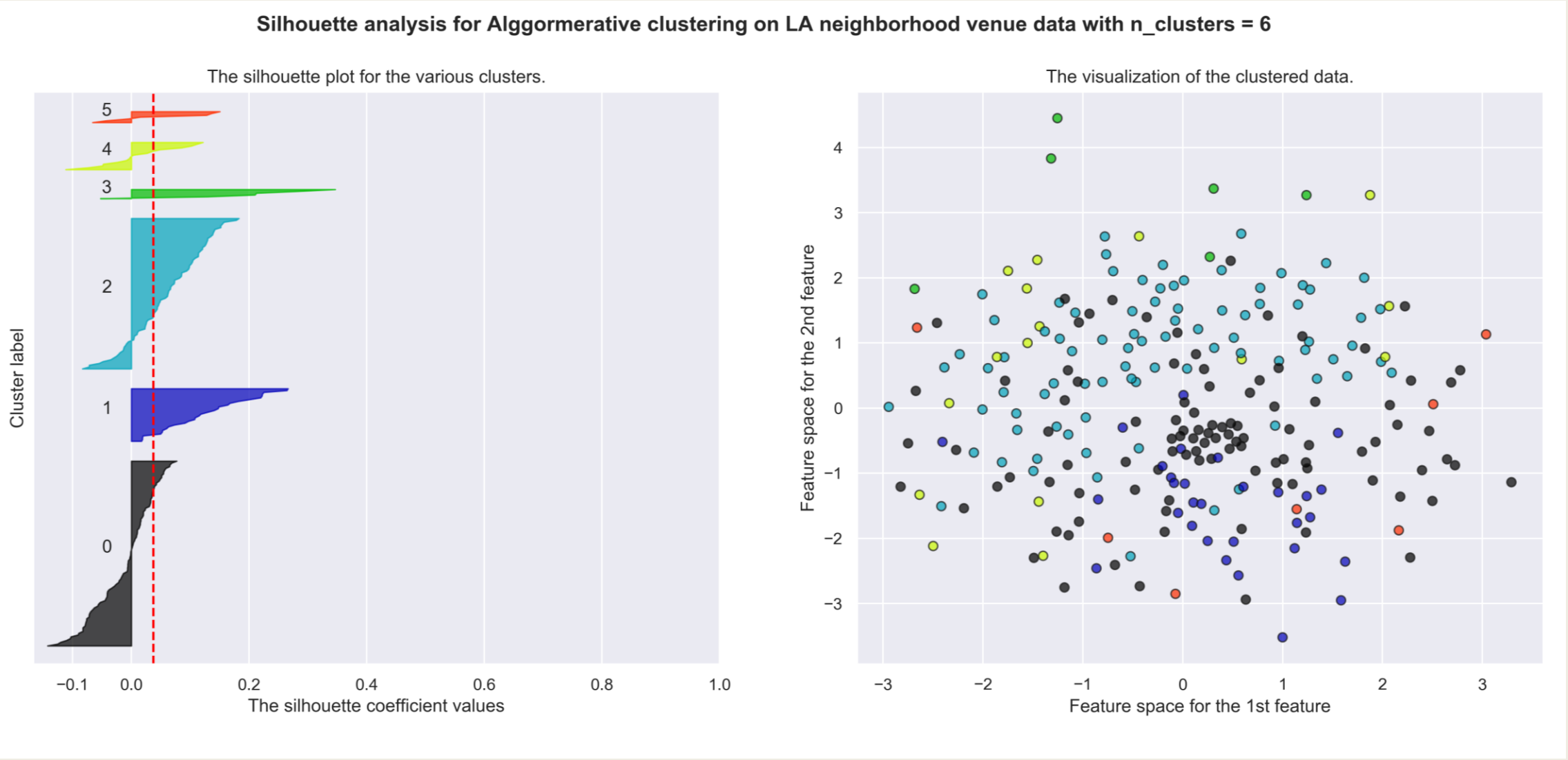
Venue Categories Clustered for Semantic meaning (semi-automated)

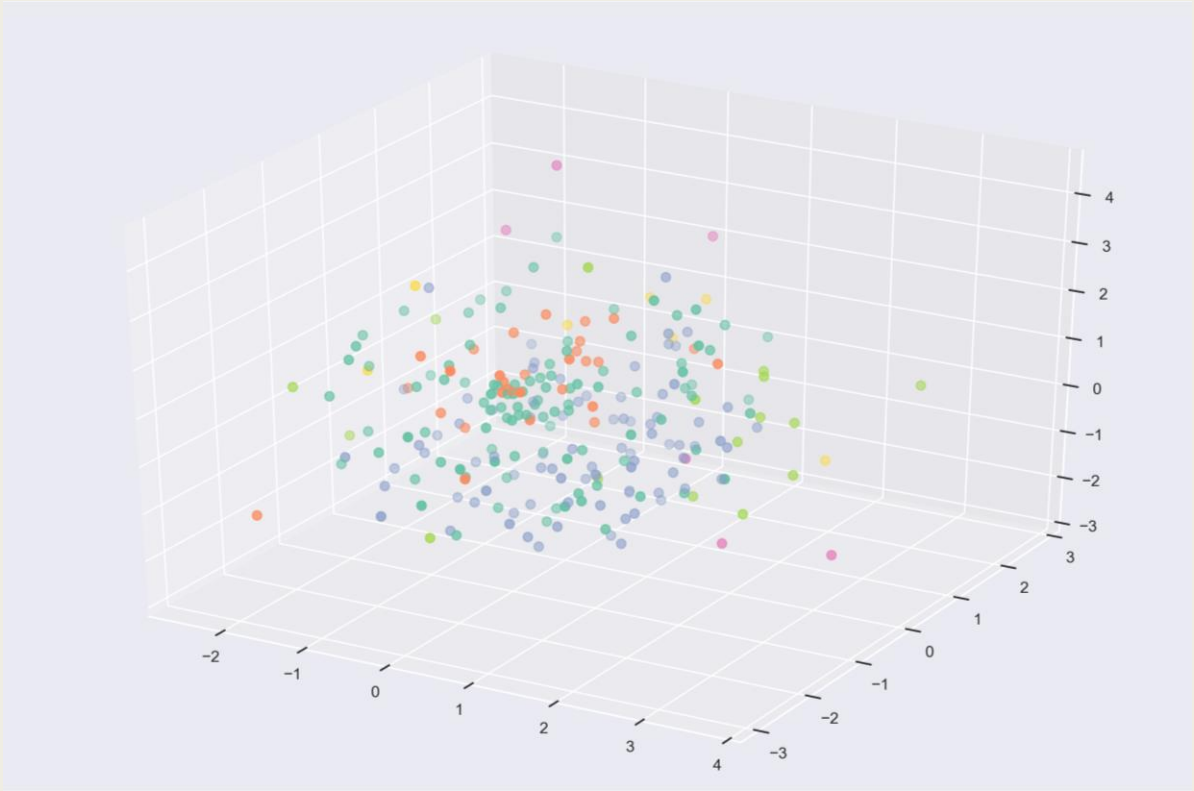


Clustering Los Angeles Neighborhoods

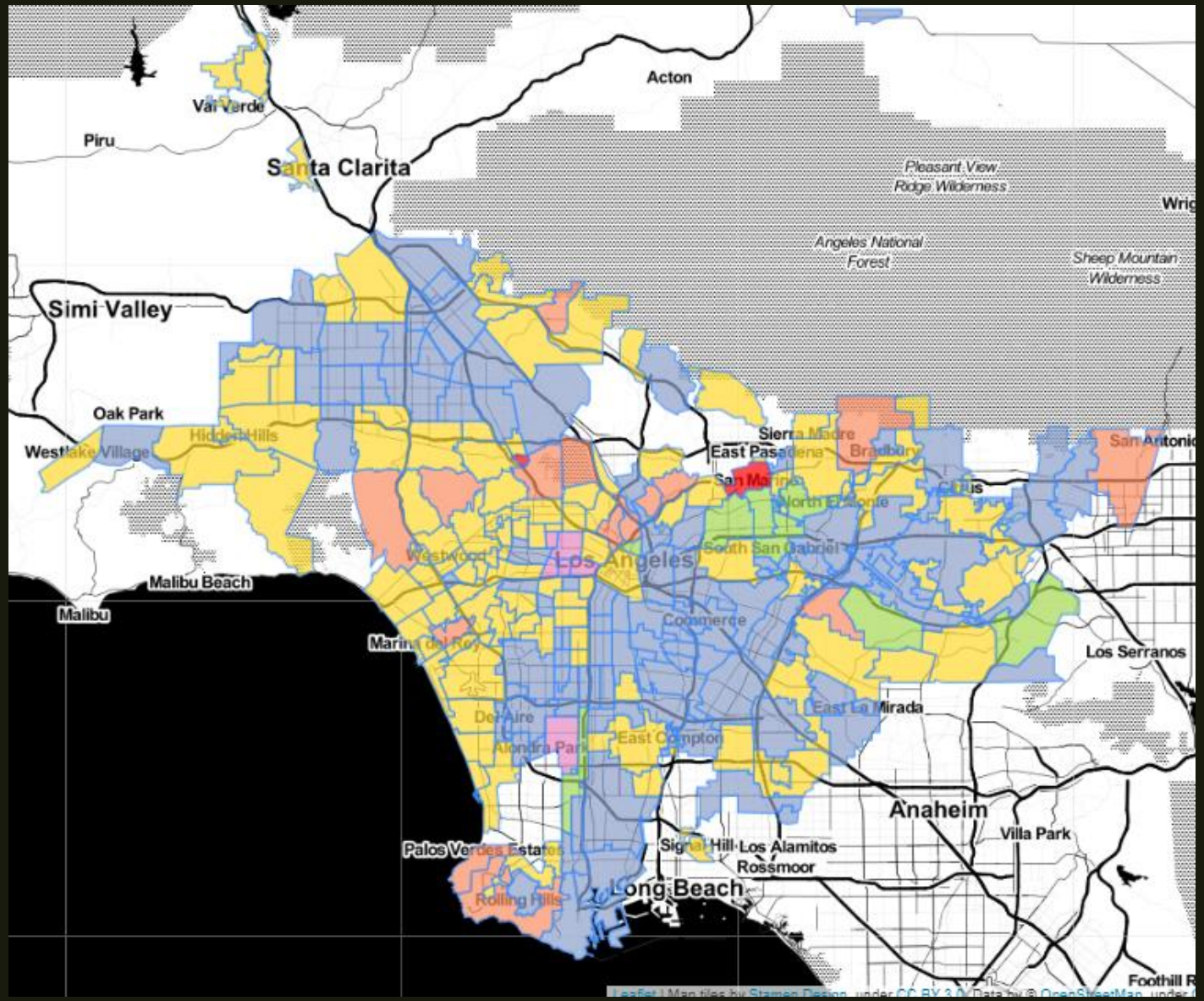


Best LA Cluster Number was Six

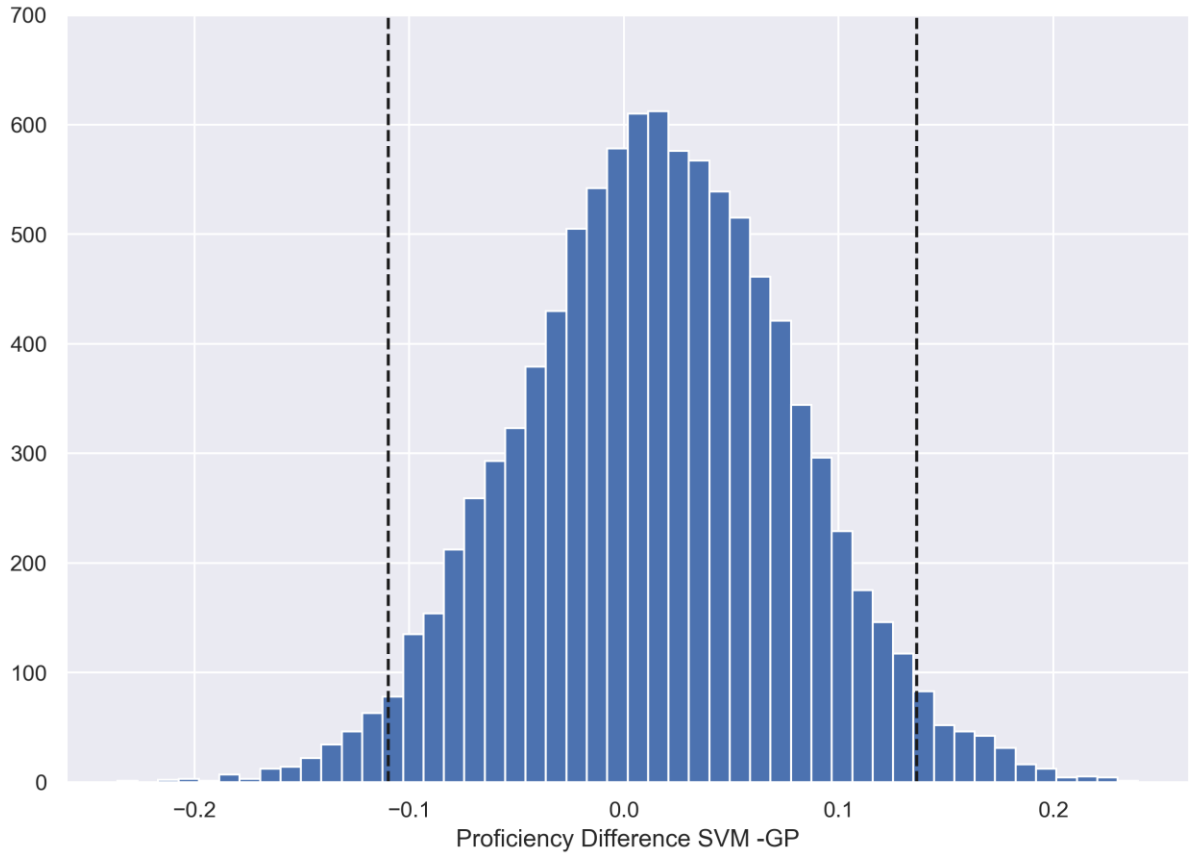




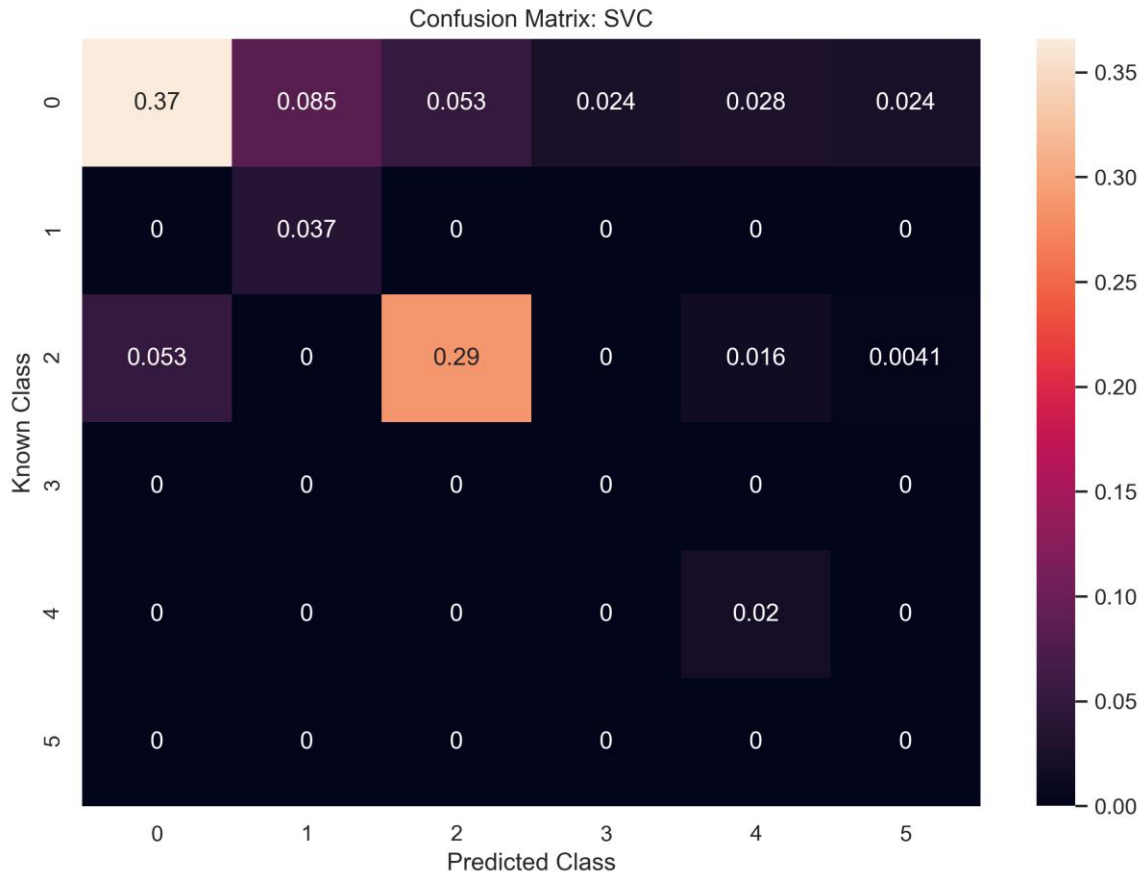
LA CLUSTERING 3D EMBEDDING (MULTI-DIMESIONAL SCALING)



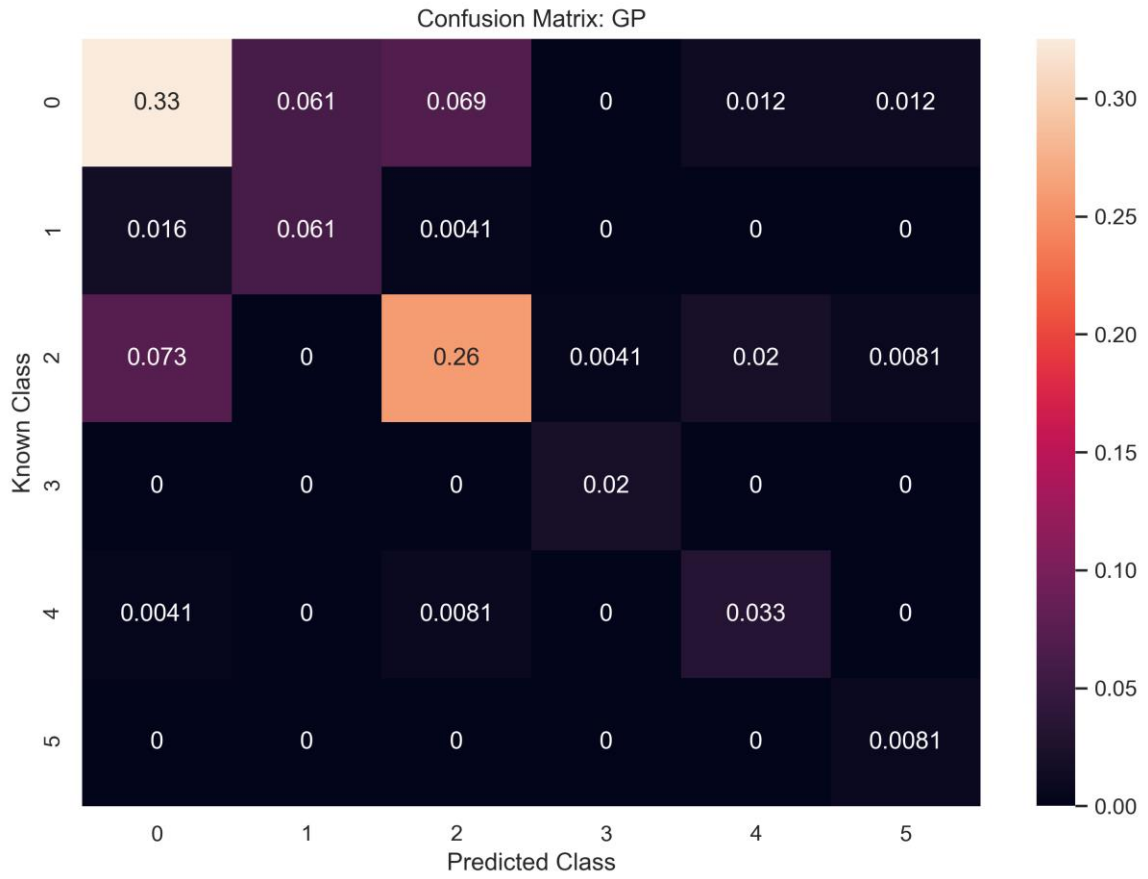
LA NEIGHBORHOOD CLUSTERING MAP (mix of venue types)



TRAINING CLASSIFIERS ON LA CLUSTERS (GAUSSIAN PROCESS AND SUPPORT VECTOR MACHINE TIED FOR BEST)

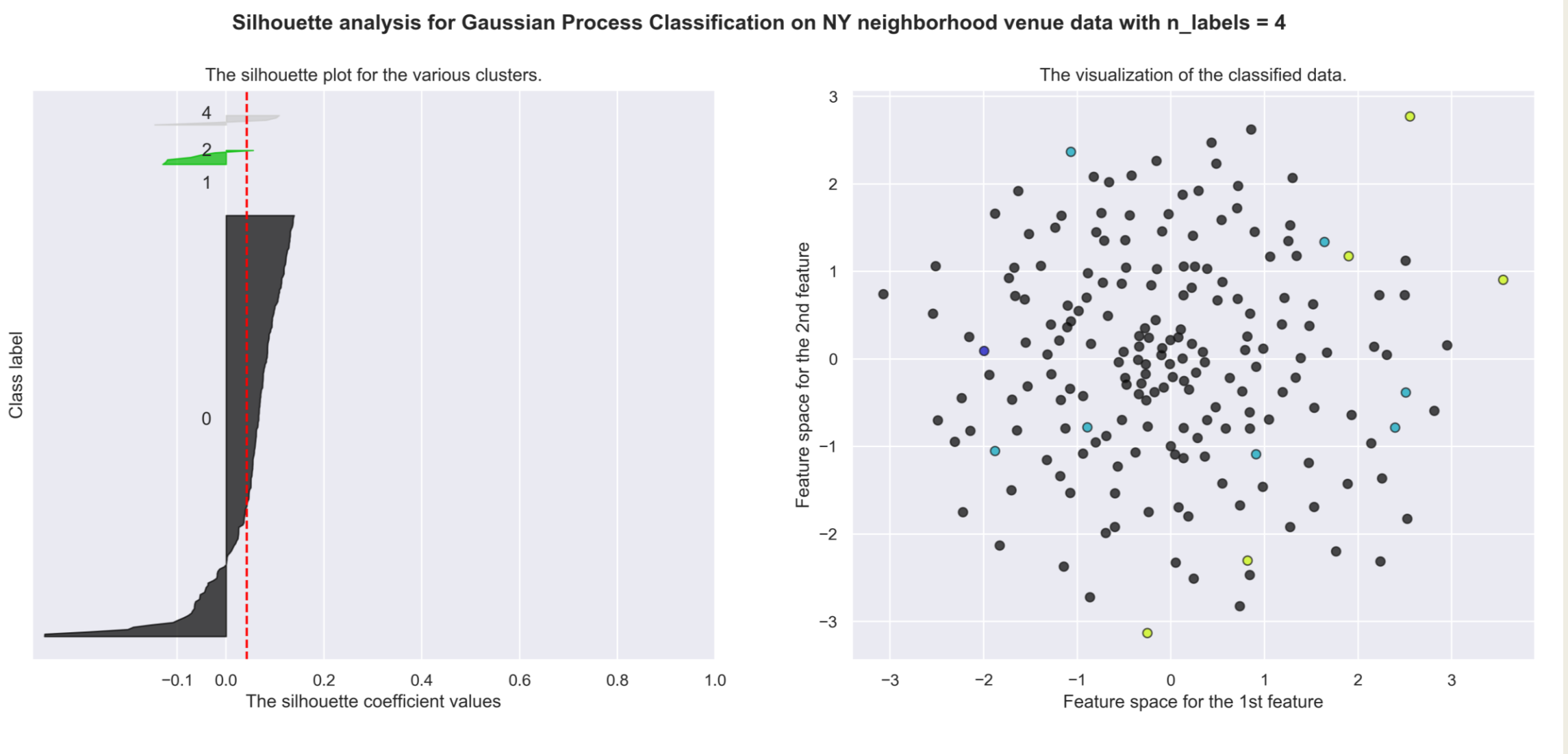


BUT SVM CONFUSION MAT SHOWS IT MISSED 2 CLASSES

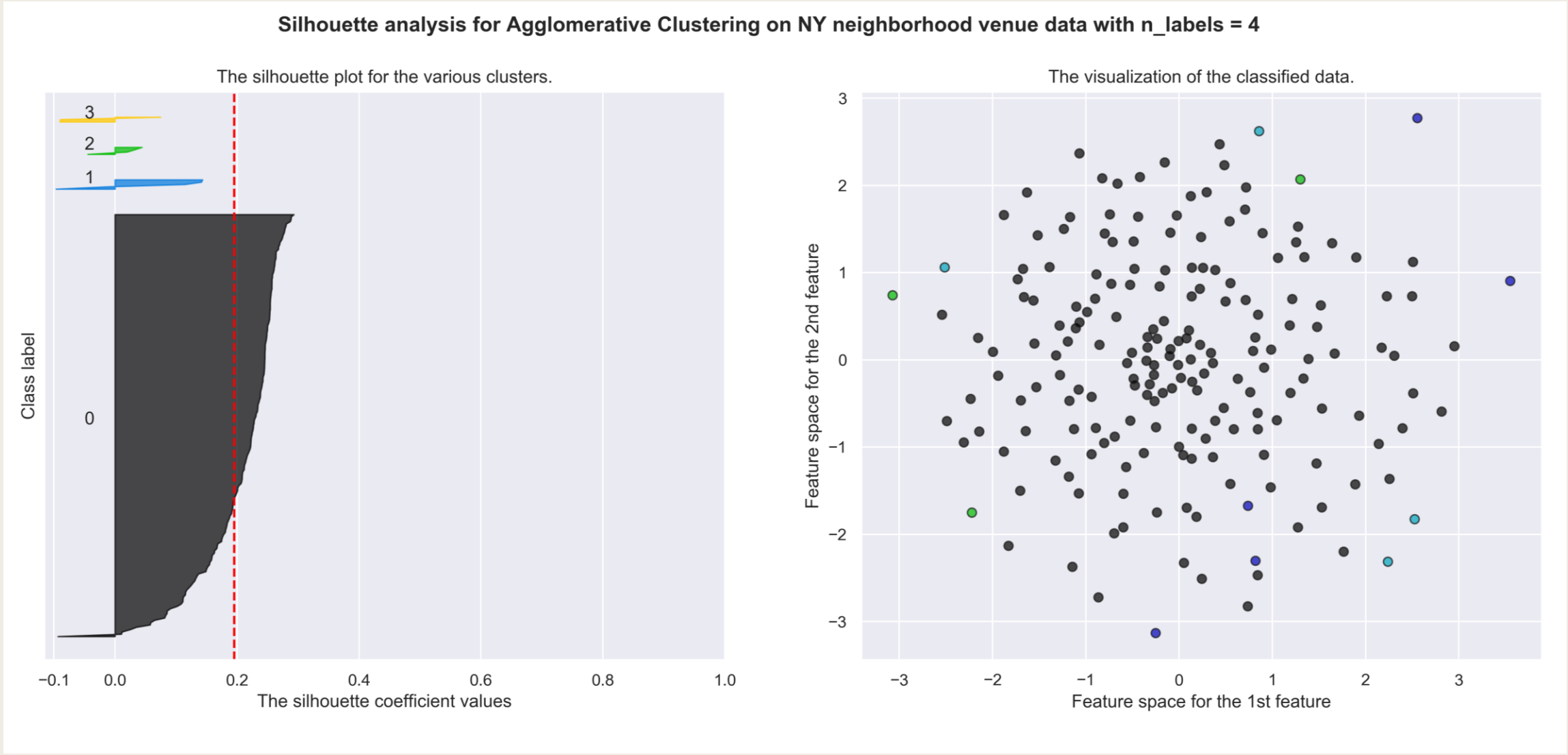


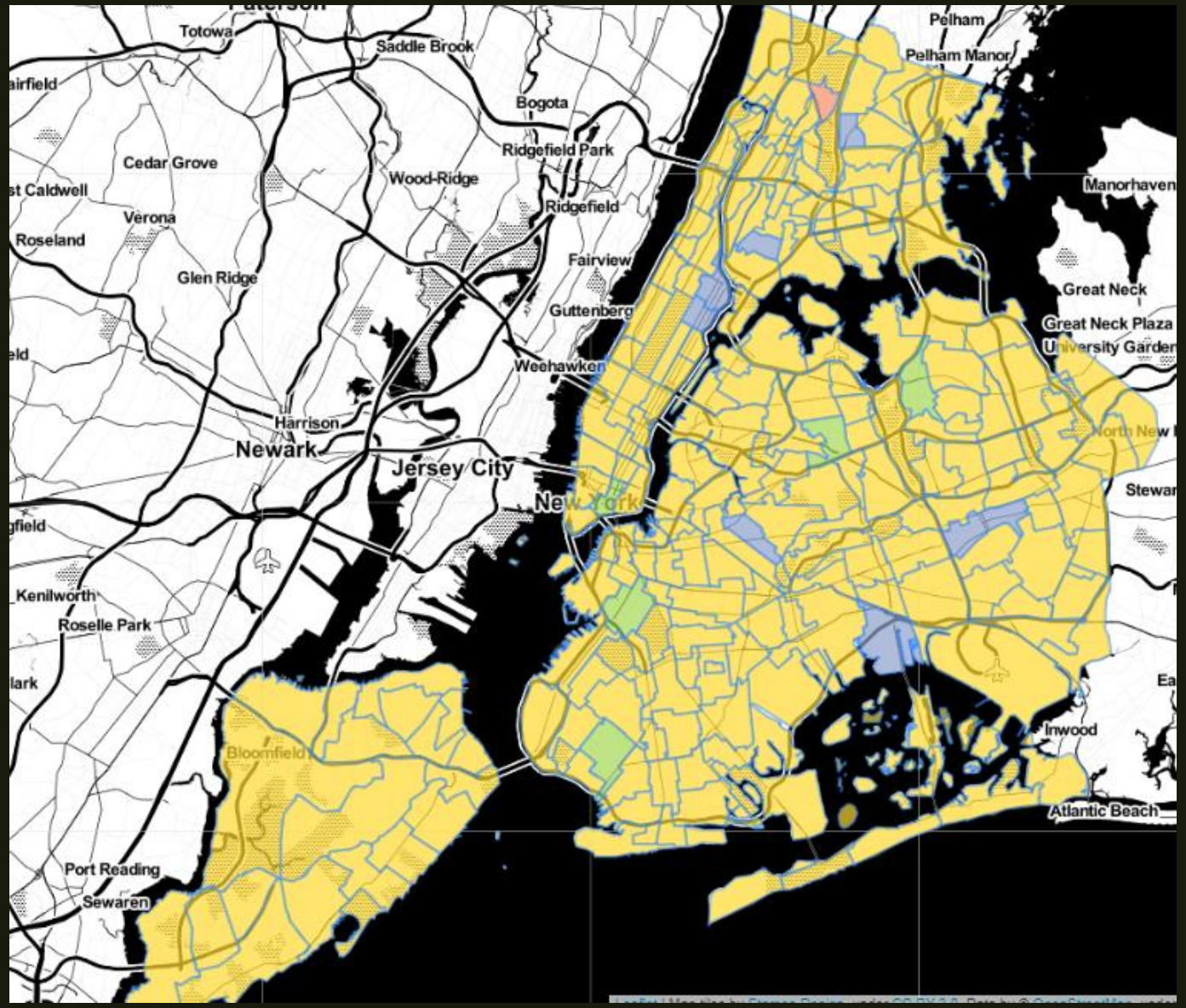
GP CONFUSION MAT CONFIRMS BETTER CLASSIFICATION FOR 2 CLASSES

NYC Neighborhood Classification



Similar Results to NYC Clustering





NYC CLASSIFIED NEIGHBORHOOD MAP (dominated by pizza places)

Conclusions and Further Directions

The approach was demonstrated to be successful with estimated classification accuracy of 70% for 6 classes and comparable results to direct clustering of the New York neighborhoods.

New York City neighborhoods were found to be much more homogeneous (Pizza Places everywhere)

Look at integrating other types of data for the clustering features, to give a more holistic comparison of neighborhoods between cities.

Use a different source for venue data, such as Google. Google venue information may be less biased to food-based venues than the foursquare social media data.

Local residents from each city could be recruited to assess the veracity of the clustering and classification.