A Venue Localizer

Michael Korbman
Final Assignment for the IBM Data Science Professional Certificate

Keeping up with a faster reality

Business background

- The dynamic of opening and closing shops is becoming faster
- Small successful shops often look for possibilities to invest their surplus of cash and to expand their business
- Identifying the right area is the main factor in the future success of the new venue
- The scale of the question is often huge

- Several factors play a role in identifying the optimal area
- Many of these factors can be more efficiently analyzed automatically
- The automatic analysis can support the selection process and provide hints for points that require a further investigation

We present a POC for a tool performing an automatic estimation of the most promising neighborhoods

Obtaining the data for the analysis

- The POC will focus on Manhattan, NY
- The data of current venues are retrieved through the API of Foursquare

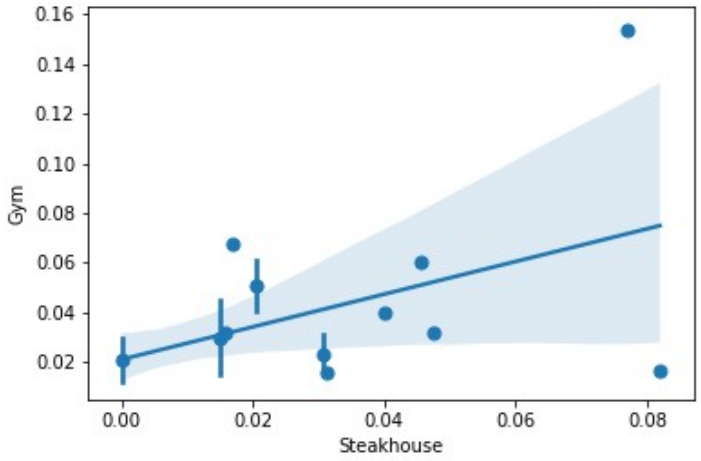
Pool of data

- 46 types of venues (Restaurants, Cafes, Theaters, ...)
- 40 neighborhoods of Manhattan (Tribeca, Little Italy, Harlem, …)
- In each neighborhood we evaluate the percentage of venues per category:
  - In Tribeca 9% of venues are American Restaurants, 6% are Italian Restaurant, and so on
  - In Little Italy 10% of venues are Bakeries, 10% are Cafes, 5% are Cocktail Bars, and so on
- This is the starting point for the predictive analysis

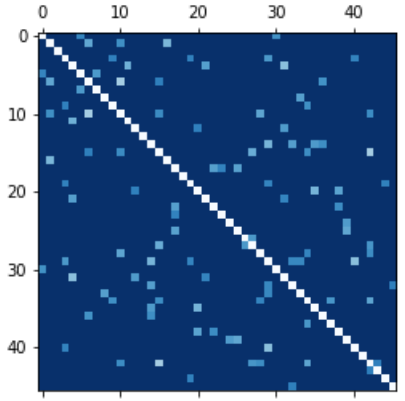
1st Approach: Cluster Analysis

Goal: Identifying groups of venues belonging together

- Some categories have often similar densities in same neighborhoods



- The Pearson correlation can be interpreted as a “distance” between categories
- A group of points close to each other (at least two points with Pearson correlation > 0.4) form a cluster
- We can thus obtain a distance matrix to perform the cluster analysis



Distance matrix for the clustering of the categories

Cluster Analysis: Results

Following clusters were identified:

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Art Gallery	Bakery	Park	Bookstore	Café	Salon / Barbershop	Italian Restaurant
Boutique	Steakhouse	Gym / Fitness Center	Burger Joint	Restaurant	Dessert Shop	Indian Restaurant
Clothing Store	Sandwich Place	Bar			Korean Restaurant	Hotel
Mediterranean Restaurant	Pizza Place	Coffee Shop			Japanese Restaurant	Sushi Restaurant
French Restaurant	Mexican Restaurant	Cocktail Bar			Vegetarian / Vegan Restaurant	Gourmet Shop
	Deli / Bodega					
	Thai Restaurant					
	Gym					

- Clusters of venue categories help to identify promising neighborhood for a new business
- Venues belonging to the same cluster become hints of a suitable area

Limitations of this approach:

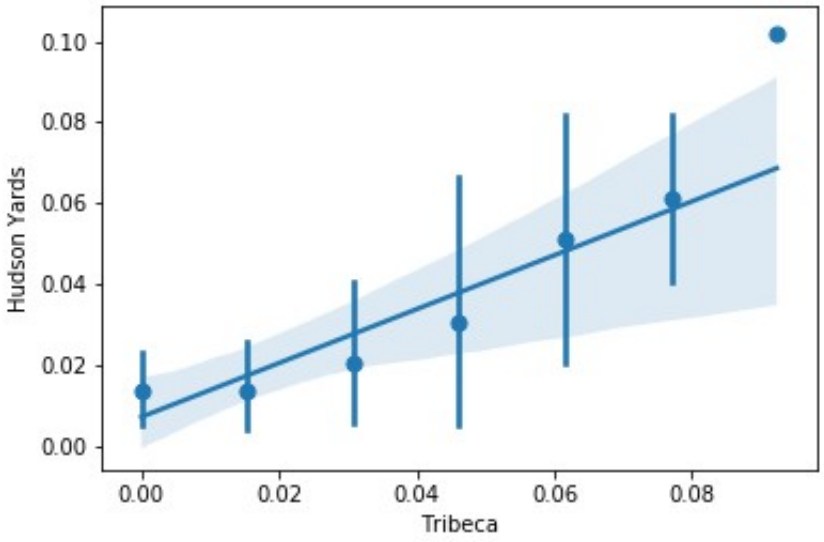
1. It does not provide quantitative estimations
2. Some categories may be left out of the clustering

2nd Approach: Recommender System

Goal: Predicting the density of a venue in a certain neighborhood

- Given a certain venue category and a target neighborhood, we can obtain an estimate of the likely density through a linear regression with another neighborhood
- Each other neighborhood provides such an estimation
- The different estimations are then weighted with the correlation coefficient

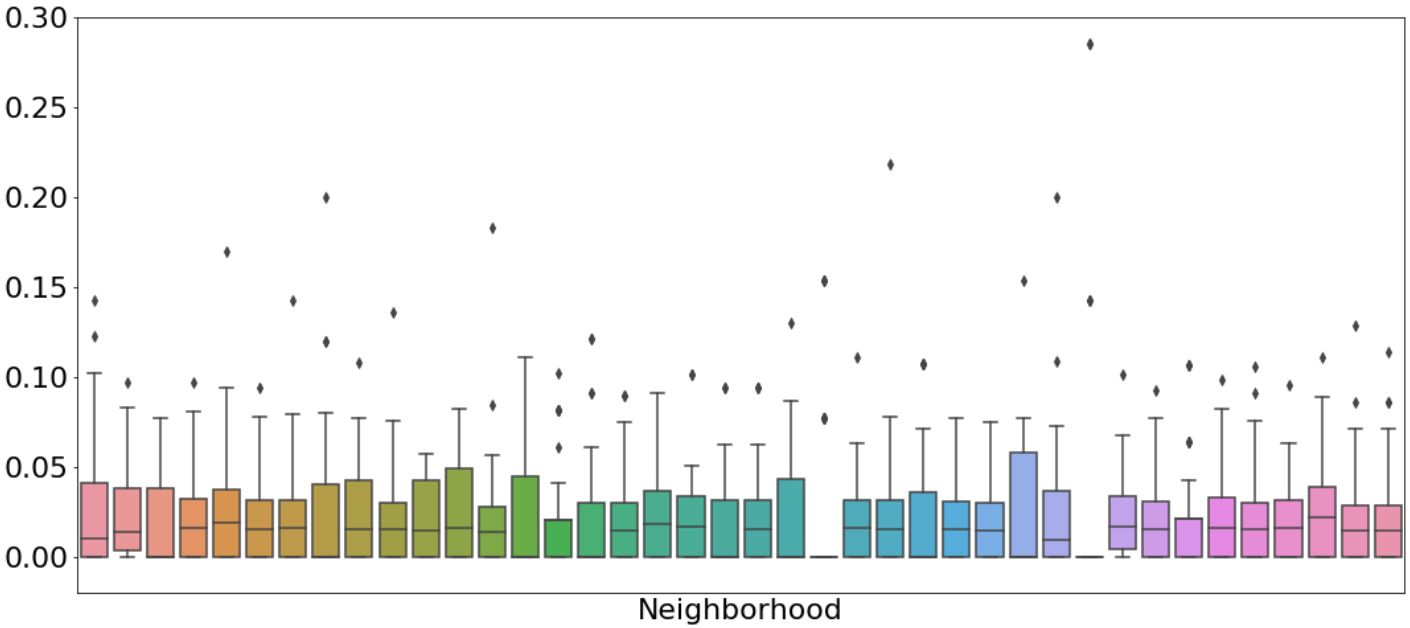
- The plot shows two neighborhoods with a relatively large correlation
- Comparing the expected density with the actual one, we obtain the confidence level of the model



Recommender System: Results

The described analysis was run for each category and neighborhood:
We obtain a deviation (σ) of 2% in the density

Density distributions in the neighborhoods



The densities range from 0 to ~20%.
A model with an average error of 2% in the relative density is therefore applicable.

Recommender System: Results

- Results are reported on an interactive map, densities with large differences from the actual values are shown
- Blue densities represents “under-represented” venues, red densities represent “over-represented“ venues

Focusing on East Village for example:

- Cafes have a density 2σ lower than expected
- Bars have a density 3σ higher than expected

It is than worth to investigate, whether it could make sense to turn some bars into cafes.

Outlook

We have shown two possible approaches for the POC

The clustering analysis

- It detects groups of venues belonging together
- It offers hints for a suitable neighborhood but no quantitative measure

The Recommender System

- It takes advantage of correlations between neighborhoods to recommend suitable target areas
- The suggestion is related to an confidence level

The POC shows some directions to improve the method and the accuracy, for example:

- Increasing the data pool investigating the effect of other cities / countries on the result
- Taking advantage of time-dependent data to improve the predictive power of the model