A Venue Placer

Michael Korbman

Final assignment for the IBM Data Science Professional Certificate

Contents

1. Business Problem

Until a few years ago there was a clear distinction between family shops, operated in a single physical place in one town (for example a pizzeria), and shops owned by big franchise (for example McDonald). This difference is becoming blurrier while more and more small family-owned companies are opening more than one sample of their business in different towns or even in different neighborhoods of the same town. While the big franchise will have branches in each possible promising spot, for the smaller companies it is crucial to identify the optimal neighborhood to open the next shop of their small chain. They would also benefit from a forecast of the expected demand (quality and size) as a stronger basis for their strategic evaluation.

In this report we will describe a tool performing a statistical analysis of the neighborhoods in a city. Leveraging the data of existing venues, the tool will be able to forecast – within a certain margin of error – which neighborhood is “missing” a shop of a certain type and, on the other side, which shops are “inflated” in the same area. According to the scope of the project, the results here presented are no more than a Proof-of-Concept (POC) in a specific area (Manhattan, NY) but it will reveal what the potentials of the tool are, as well as which areas would require a more detailed development.

The human intervention is still fundamental at this stage, for example in order to identify whether an observed large deviation from the average can be explained from concurrent reasons. Given the scale of the problem (the number of cities multiplied by the number of neighborhoods in each city) however, an automated screening will be necessary in most cases.

The proposed tool can then be offered to the small chains, evaluating their specific kind of shops, as a support in the process of localizing the area to open a new branch.

2. Data

For the POC we will focus on the neighborhoods of Manhattan, New York. We want to take advantage of correlations between the neighborhoods, which we will evaluate based on the relative density of venues. To retrieve data from the venues, the Foursquare API is called providing as input parameters the coordinates of the neighborhoods and the radius. We are using a radius of 500 meters (thus retrieving venues in a 0.8 Km^2 circle around the center).

In order to increase the statistical significance of the results, we discard all venues having less than 20 shops in Manhattan. This reduces the 339 types of venues returned from the API to just 46. In each Manhattan neighborhood the absolute number of a certain type of venues is normalized over the total number of venues in that neighborhood (i.e. averaged). The resulting table (`df_nfreq` in the notebook) contains the average density of venues in the neighborhoods sorted according to their type; we will perform the analysis described in the following sections using this data.

3. Methodology

The tool should be able to provide insights about suitable neighborhoods for an input category of shop. We will pursue two different approaches to the question:

a. A cluster analysis of the venues. Once that venues “belonging together” have been identified, target neighborhoods can be pointed out according to the venues they have.
b. A recommender system for neighborhoods. Drawing from correlations between neighborhoods we can predict the expected value of a certain category of venue. Comparing this value with the actual value a “demand” or an “oversupply” of a certain type of businesses in a certain neighbor can be estimated.

We will show that the method (a) does deliver an indication of which venues to search for, in order to identify a promising neighborhood. However, since the result is difficult to interpret quantitatively and the statistic in the POC is a bit weak to obtain meaningful results, we will not push this approach much further and will focus mostly on method (b). Both approaches are explained in detail below. In both cases we start from the table containing the relative frequencies of venues in the neighborhoods.

- `df_freq` contains the venue categories in the columns
- `df_nfreq` contains the neighborhoods in the columns

3. (a) Cluster Analysis

The first step is to systematically identify which type of venues often occur together i.e. which venues have a high (low) relative frequency in the same neighborhood. This can be evaluated quantitatively calculating the Pearson correlation between venues across neighborhoods. The value of the correlations is not expected to be large, since in general the same venue category will occur in very different kinds of neighborhood. For example, even though *Steakhouse* has the highest correlation (r = 0.45) with *Gym* (Fig. 1), many points lay very far from the regression line.

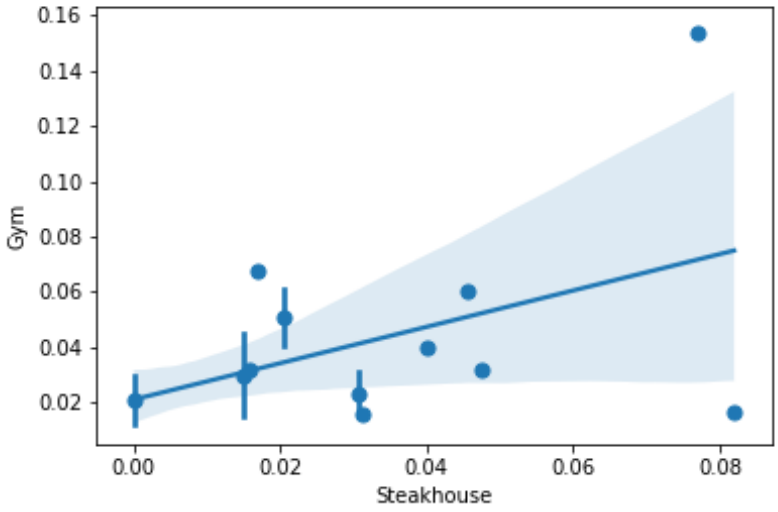

Figure 1: Scatter plot of Gym / Steakhouse. In case of multiple points for the same x-coordinate the average is taken. The bar shows the points' spread.

Calculating the correlations over the whole dataframe we obtain a symmetric matrix. The diagonal contains the correlation of each category with itself and is therefore equal to 1. We can now read the correlation as a "distance" between the categories: the higher the correlation the "closer" the points are to each other. We translate this interpretation in the matrix performing the following operations:

- Setting the negative correlations to zero
- Defining the matrix `dmatrix` (numpy array of arrays) as 1 minus the correlation value

`dmatrix` can then be taken as a distance matrix: the entry in the cell *i,j* contains a sort of normalized distance (range between 0 and 1) between the categories *i* and *j*. We set distances larger than 0.7 (i.e. correlations smaller than 0.3) to zero, to prevent these points from being clustered together. The resulting matrix is displayed in Fig. 2.

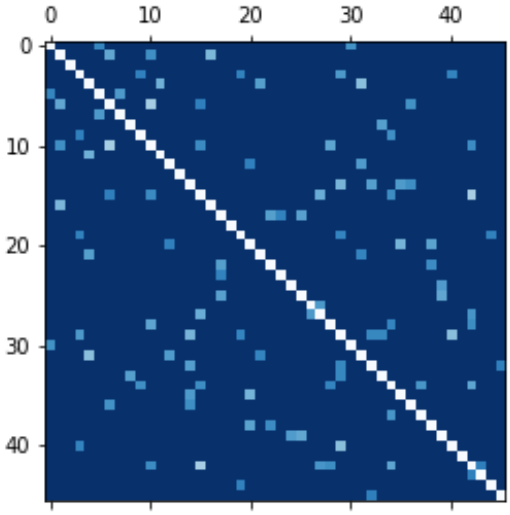

Figure 2 Distance matrix of the venue categories. White represents distance = 0, dark blue represents distance = 1

Having defined the relative distance between the categories we can cluster them. For the clustering we use the DBSCAN algorithm, defining a cluster as at least two points closer than 0.6 (i.e. with correlation larger than 0.4). The outcome of the clustering is discussed in the section 4 (a)

3. (b) The Neighborhood recommender system

Knowing the relative density of venues in many neighborhoods, we can run a statistical analysis to predict the expected density of a venue category in a certain neighborhood.

Let us assume that we want to predict the density $D_{x,y}$ of the venue category c_x in the neighborhood N_y. We will use as weights for the recommender system the correlation $K_{i,j}$ between neighborhoods N_i and N_j , evaluated using the presence of similar venues.

In general, each neighborhood N_i has at least three or four neighborhoods it is strongly ($r > 0.4$) correlated to, and that can thus provide a significant estimation. For example, Tribeca has the following largest correlations:

Hudson Yards	0.54
Civic Center	0.5
Turtle Bay	0.47
Lincoln Square	0.47
West Village	0.43

We can have a look at the linear correlation plot of Tribeca and Hudson Yards:

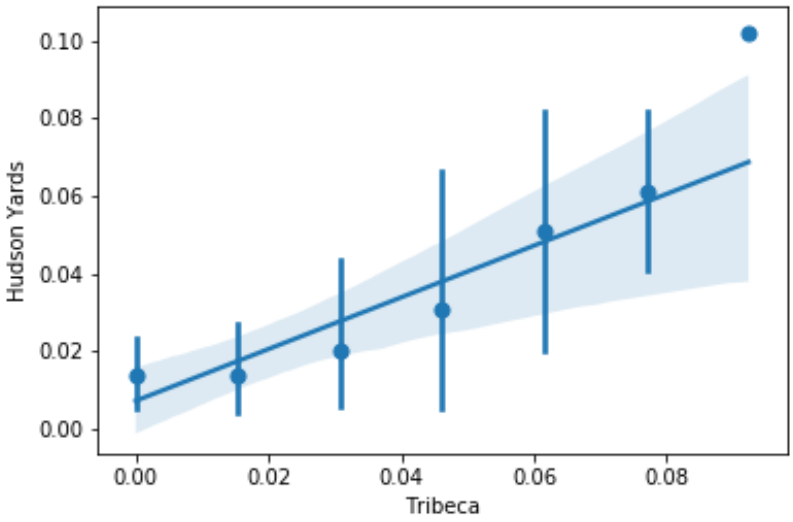

Figure 3 Correlation of two neighborhoods

From each other neighborhood N_j we will extract a predicted value which we will include in the average, weighting it with the $K_{y,j}$ coefficients. There are different ways to calculate the predicted value.

- **Direct**
  The predicted value is the value of the unknown category c_x in the neighborhood N_j.

- **Interpolated**
  Naming v the value of the category in the neighborhood the predicted value is f(v), where f is the line resulting from the linear regression.

Let us make an example to clarify the difference. Let us predict the density value for the category Café in the neighborhood Hudson Yards using Tribeca as reference point. Hudson Yard and Tribeca have a linear correlation coefficient of 5.4, the regression line is depicted in Figure 3; the relative density value of Cafés in Tribeca is 0.06.

- If we use the direct method, we predict for Hudson Yards the same value 0.06 and underestimate the actual value (0.08) by about 25%.
- If instead we use the interpolation method, we take the value of 0.06 in Tribeca and feed it to the interpolation line (Fig. 3) obtaining 0.04, thus underestimating the actual value by about 50%.

In the following we discuss the calculation of the predicted values as well as methods to evaluate the error and to decide between different methods and parameters.

The evaluation of the expected value $E_{x,y}$ takes place as follows:

- We drop the category c_x from the whole table. This category will be the “test” while all the other are the “training” set.
- We calculate the correlations $K_{i,j}$ without the “test” category. These will be the weights in the evaluations. Negative correlations are set to 0 and do not play any role in the estimation.
- We define the expected value $E_{x,y}$ as the weighted average of the densities of c_x in the other neighborhoods:

$$E_{x,y} = \sum_{j \neq y} [k_{x,j} \cdot D_{x,j}] \ \frac{1}{\sum_{j \neq y} k_{x,j}}$$

.

Running the calculation on all the categories and neighborhoods, we obtain a set of expected values which we can compare against the known values $D_{x,y}$. The deviation σ is

$$\sigma = \sqrt{\frac{1}{N} \sum_{x,y} (E_{x,y} - D_{x,y})^2}$$

This number gives us a measure of the accuracy of the model. We can use it to evaluate which set of parameters optimizes the result of the model i.e. minimizes σ. For example, comparing the direct and interpolation mentioned above as well as their average we obtain following values for σ:

- Direct: σ = 0.0220
- Interpolated: σ = 0.0236
- Average: σ = 0.0225

Therefore, we will use the direct method in the evaluation. An error of 0.02 means that the model will not have a good accuracy on low values of the density, while results on venues with larger density (e.g. >0.04) will have a more adequate confidence level.

The deviation σ can also be evaluated per neighborhood or per category, according to the required estimation.

4. Results

As mentioned previously, we present separately the results of the two methods: we discuss an overview of the clusters obtained from (a) and a more extensive analysis of the predictions from (b).

4. (a) Results of the Cluster Analysis

The analysis was run over the 46 categories selected in the exploratory phase. 30% (14) of the categories are identified as outliers, i.e. they do not belong to any cluster; in other words, the tool would not be able to perform any analysis for them. Cosmetic Shops, Yoga Studios, Spas fall into this group.

The remaining categories are clustered in the following groups:

Group 1	Group 2	Group 3
Art Gallery	Bakery	Park
Boutique	Steakhouse	Gym / Fitness Center
Clothing Store	Sandwich Place	Bar
Mediterranean Restaurant	Pizza Place	Coffee Shop
French Restaurant	Mexican Restaurant	Cocktail Bar
	Deli / Bodega	
	Thai Restaurant	
	Gym	

Group 4	Group 5	Group 6	Group 7
Bookstore	Café	Salon / Barbershop	Italian Restaurant
Burger Joint	Restaurant	Dessert Shop	Indian Restaurant
		Korean Restaurant	Hotel
		Japanese Restaurant	Sushi Restaurant
		Vegetarian / Vegan Restaurant	Gourmet Shop

The analysis says, for example, that in Manhattan Clothing Stores and French Restaurants tend to be present in the same neighborhood (Group 1). The results of the clustering make sense in several case: many of the venues falling in the same group do have some kind of relation or a similar target (e.g. Pizza and Sandwich Places). In a couple of cases however, the results do not seem completely consistent: for example, Café and Coffee Places end up in different clusters, and Restaurant does not belong to any cluster with other restaurants. This can be due to a lack of statistics - there were only 23 venues belonging to the Restaurant category, just above the threshold we set - or to a different kind of categorization - Café and Restaurant are both more generic categories than the others.

4. (b) Results of the recommender system

The expected density of venues was estimated per category and neighborhood. The difference (delta) between each expected value and the corresponding actual value was then compared with the average deviation σ:

- If the delta is positive and larger than σ, it is a hint that there could be an unsatisfied demand for the category in the neighborhood under examination
- If the delta is negative and its absolute value is larger than σ, the venue category might be over-represented in the neighborhood

The delta is conveniently measured in units of σ. All deltas in module larger than sigma are reported on a map of Manhattan, the markers are clustered around the neighborhood centers.

Figure 4 Screenshot of the interactive map, showing the venues over / under populated according to the algorithm

Blue markers correspond to positive deltas i.e. the expected value is larger than the actual one and the neighborhood could offer a good spot for the venue to thrive. Red markers correspond to negative deltas, i.e. venue categories which are more frequent than expected. The darker the color of the marker, the larger the delta (w.r.t. σ). This overview allows a quick check of promising spots for a certain venue; in addition to that, a company interested in opening a certain shop may start checking whether one of the venues belonging to the “red” categories is interested in selling his place.

Let us take one case to exemplify the result. Looking at East Village, it seems that some of the Bars could be converted to Cafés

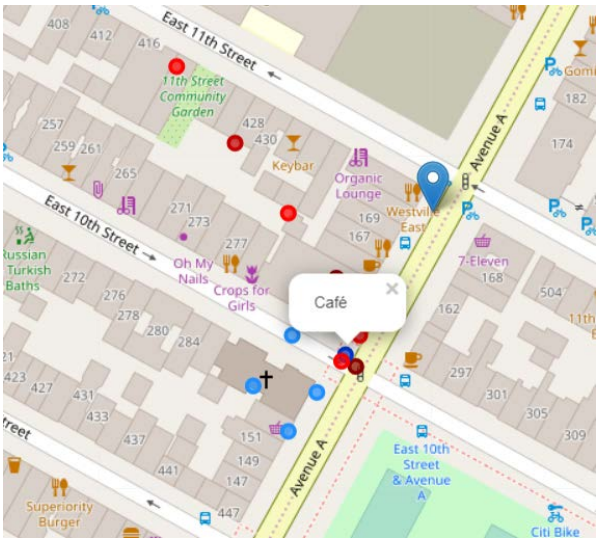

Figure 5 East Village - Extract of the map

East Village

Category	Delta / σ
Café	+1.9
Bars	-3.2

We can use this example to quickly check the algorithm procedure and try to make sense of the outcome. East Village has its strongest correlation with Gramercy: both have a relatively large amount of Bagel Shops (4% to 6%), Coffee Shops (5%), Ice Cream Shops (5% to 7%), and Pizza Places (6%). This sounds indeed as a suitable environment for a Café. However, while Gramercy does have a 3% density of Cafes, there are none in East Village. While the question will require an investigation before taking any business action, the model has correctly spotted this anomaly.

Looking at Bars, both East Village and Gramercy have a large amount of them. In East Village however, they make up to 10% of all the venues in the neighborhood. The numbers indeed point to a likely strong competition and to a strategic advantage in diversifying the offer.

5. Discussion

Below we discuss the outcome of the two methods presented.

5. (a) Cluster Analysis

The cluster analysis offers an indication of the kind of venues which would be suitable with a prospective new one. For example, if the client is a chain of fine French restaurants, the tool suggests looking for areas with boutiques and art galleries.

This kind of analysis, however, has two main limitations:

- It does not provide any quantitative measure of the reliability of the suggestion
- It leaves out all the categories which could not be clustered

The first point can be tackled going back to the correlation values and defining a proper figure of merit to evaluate the results. The second point is partially a consequence of setting low correlations to zero. Improving the degree of confidence requires a better statistic and a specific analysis; on the other hand, having to adjust the tool for each query would be a major flaw.

In order to improve the statistic, it would be interesting to study whether similar patterns are found in different areas of a city or in different cities. For example, if we found that the venues of the whole north America follow similar behaviors (within a certain margin of error), the pool of data could increase by several orders of magnitude.

5. (b) The neighborhood recommender system

The recommender system has many advantages with respect to a simple cluster analysis. The model is transparent, in terms of result calculation as well as in the confidence level of the result. Adjusting just a few input parameters, it can be optimized for a single venue category or for certain neighborhoods: the confidence level can be calculated for the required settings.

The results of the POC are encouraging, showing a certain ability of the model to support the strategic selection process of target areas for new venues. As for the cluster analysis, the model would indeed benefit from a better statistic, for example using more data. The advantage of the recommender system is that it can be trained to automatically recognize whether a new pool of data - e.g. a new city - is improving the overall accuracy of the model or not, and therefore automatically include the new data or reject them. In other terms, a table with the coordinates of all neighborhoods where Foursquare can deliver data is all it takes to enlarge the training dataset from Manhattan to potentially the whole world.

.

6. Conclusion

We have discussed two possible approaches to the problem of spotting a suitable neighborhood for a certain venue. While the clustering analysis delivers more qualitative results, the recommender system provides an expected value and the confidence interval to compare it with.

From the POC we can thus draw promising directions for a complete implementation; at the same time, it helps us to identify the points which would require some refinements. The POC is also able to deliver first meaningful results for the area it was applied to.

The main conceptual criticism which can be directed to the model, is that most of the times the tool will be evaluating an equilibrium situation, therefore the order of magnitude of the deviations identified – 2% of the relative density – would have likely already been filled if there were an actual pocket of profitability. While a more accurate analysis, and thus a smaller σ, could be obtained using a larger pool of training data, the model could also be improved recognizing trends over time, in order to be able to deal with changing situations instead of with a static environment alone.