Opening a French Restaurant in New York

# Introduction

In this scenario, an investor is seeking to decide whether to support a chef who wants to open a new French restaurant. The restaurant scene in New York is deeply competitive, and margins are slim, so the choice of location is an essential factor in developing confidence that the restaurant can be successful. The investor wants to start looking for properties to lease, but wants to identify characteristics of a good location to search. The presence of competing restaurants will help determine in which neighborhood to search for available properties; a neighborhood with a competitor already open is a bad choice, but a neighborhood too unlike competitors’ does not inspire confidence that the other local venues are indicators that likely guests will be desirous of a French restaurant in the neighborhood.

# Data

Fortunately, Foursquare provides collected data on venues, conveniently segmented by neighborhood in the city. This data provides location, venue type, and neighborhood of the venue, and will be used in aggregate to generate a characteristic representation of each neighborhood, based on the venues within 500 meters of the geographic center of the neighborhood.

The Foursquare data is available through the Foursquare API, and can be queried throughout the development process, including “hot now” queries that can be run during the days and times the restaurant plans to be open to refine the characteristics of target neighborhoods.

# Methodology

I obtained a latitude/longitude marker for each of the 306 neighborhoods in New York and queried Foursquare for the 100 top venues within 500 meters of the GPS coordinates. This may result in certain venues being captured for multiple neighborhoods with nearby centers, but I decided not to de-duplicate based on the closest neighborhood center, as diners will likely be willing to cross neighborhood boundaries, or not even realize they are doing so, to move between multiple venues they wish to visit.

Most neighborhoods had fewer than 100 venues, so we may need to dig more deeply into neighborhoods with very few venues to determine whether their exclusion will yield more descriptive results. I identified 433 distinct categories of venues, which represent plenty of features to encode distinctions between neighborhoods. I converted the data to a one-hot list of venues by neighborhood, considering that for an initial investigation, the presence of a type of venue is more important than the quantity of a particular type of venue.

Next, I segmented the data into two dataframes: one containing only the neighborhoods in which French restaurants are currently open and one containing the neighborhoods without a French restaurant. There are forty neighborhoods with French restaurants and 262 without. I performed *k*-means clustering on the neighborhoods with French restaurants, segmenting them into five classes. Similar clustering examination, with various values of *k*, consistently lumped nearly all restaurants without French restaurants into a single class, so I decided not to progress with clustering the target neighborhoods, but rather to do a simple ranking based on proximity to the centers of the interesting classes of neighborhoods with competing restaurants.

Two of the five classes of neighborhoods with French restaurants had only one member, so I exclude those classes as not containing sufficient information to interpret characteristics of a desirable neighborhood. Similarly, four of the five classes of potential target neighborhoods contained fewer than five neighborhoods, so I discarded those neighborhoods from consideration on the grounds that they are sufficiently distinct to be unreliable for predicting a good return on investment.

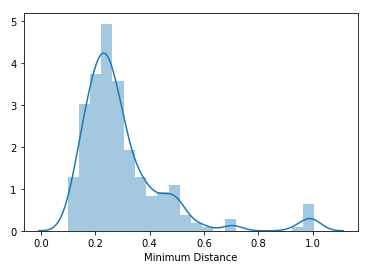
The next task was to assign some priority to the potential target neighborhoods. I decided that a simple feature proximity would be a good starting point, and if the investor is interested, I could augment this with physical proximity to the nearest competing restaurant. I left out this metric in this initial examination because I am already only targeting neighborhoods without a French restaurant, enforcing some minimum distance natively. Additionally, proximity to competitors is likely to be more important once candidate sites are identified, which requires a real estate search that is based on the outcome of this project. For a final priority list, I sorted candidate neighborhoods by feature similarity to the means of the three classes of neighborhoods containing French restaurants.

# Results

After sorting candidate neighborhoods by similarity to the centers of clusters of neighborhoods with French restaurants, I identified the following ten neighborhoods (in order) as initial candidates for a real estate search:

* South Side
* Brooklyn Heights
* East Village
* Flatiron
* Gowanus
* Gramercy
* Washington Heights
* Hunters Point
* Prospect Heights
* Astoria

I wanted to make sure this was not arbitrary, so I produced a histogram of feature vector distances:



The distribution indicated that the feature distances were not just clustered at low values, so this helped assess the identified neighborhoods as valid candidates. I checked the top ten most common venues for each of these neighborhoods, and they consistently included bars, cafés, Italian and pizza restaurants, among other common venue types, at the top of their lists.

# Discussion

My results indicated that neighborhoods with French restaurants can be classified based on the other venues in the neighborhood. My recommendations for neighborhoods to search for possible locations were based on identification of features of candidate neighborhoods similar to those with a French restaurant.

It may be easy to narrow the search further based on property prices, and to further investigate the physical proximity of candidate locations to other French restaurants. I would recommend seeking historical information about where French restaurants have recently closed to avoid trying in a location that has already failed, and potentially exclude recently opened restaurants as we should avoid relying on them due to the inherent uncertainty of their success. Parking availability or walkability will be important factors based on the target clientele.

Based solely on the feature set of venues in the neighborhood, this data was sufficiently descriptive to disambiguate neighborhoods and make ranked recommendations for an initial location search.