XYZ Ice Cream Company - Expansion Plans

New Site Evaluations

Fred Ariemma - June, 7 2019

1.0 Introduction.

1.1 Background.

This Project is the capstone to Applied Data Science on Coursera completing certification in Date Science from IBM. In the project, we have been by contracted by XYZ Ice Cream Co. to examine available retail locations in Manhattan and produce a report indicating the 10 best locations to open a new XYZ Ice Cream shop.

1.2 Business Problem.

XYZ Ice Cream Shop wants to expand and open a new location somewhere in Manhattan. Finding a good location is key for retail businesses. Choosing a poor location will waste extensive funds required during re-modeling and start-up of the new business. XYZ wants us to examine locations across Manhattan to make sure that it's a good location capable of supporting an ice cream shop is chosen before construction begins.

1.3 Assumptions.

We will use data from locations surrounding existing ice cream shops to determine the correct mix of foot traffic to support an ice cream shop. We are assuming that if an ice cream shops exists at a location then it must be a good location with a good mix of venues that generate foot traffic to support an ice cream shop.

2.0 Data Gathering.

Several resources will be used solve XYZ's problem.

2.1 Find Existing Ice Cream Shops.

First, FourSquare’s search API will be used to find all the Ice Cream shops currently operating in Manhattan. FourSquare will return the name, the address, the category, the geolocation, along with other details about the location. This data is then cleaned, as FourSquare, may return more than just ice cream shops.

2.2 Find Venues that Surround Existing Ice Cream Shops.

The location data of the ice cream shops from 2.1 is used to call FourSquare’s Explore API to find all the types of venues that surround existing ice cream shops. The data is cleaned to extract a category for each venue.

2.3 Feature Selection.

The surrounding venues data is then cleaned to create feature attributes. One hot encoding is used to create the features that are required by machine learning to evaluate new locations. The features are then summed by each location to create the training data for the regression model.

2.4 Find Available Retail Leases.

Available retail space is found by using a screen scrapper against the real estate web site Showcase.com. Showcase will return a list of available leases, their address, rent, and square footage.

2.5 Find Venues that Surround Available Locations.

The surrounding venues of the available locations are found by using FourSquare Explore API again. The data is cleaned to extract a category for each venue.

2.6 Evaluate Available Locations.

Clean the available locations venue data by creating a feature set using one hot encoding, sum the features for each available location, then map the feature columns form the available locations to match the feature set created from the existing ice cream shop locations. The columns need to match as required by the regression model. Lastly, the regression model's predict function is used to score each location.

3.0 Methodology.

3.1 Foot traffic.

Our basic premise is that foot traffic provided by nearby venues provides a good environment to support an ice cream shop. First, we need to find existing ice cream shop location so that we can examine the surrounding venues. We need to do data analysis on various levels of category codes to insure that we are only extracting ice cream shops. Figure 1, shows existing Ice Cream Shops below 90th street in Manhattan.

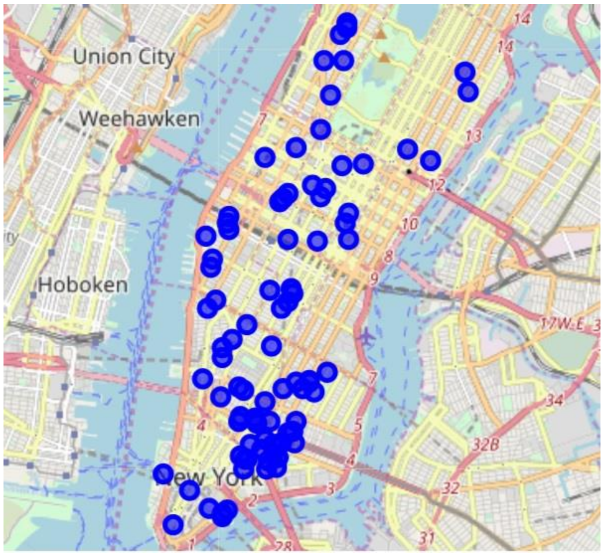

Figure 1 - Existing Ice Cream Shops in Manhattan

3.2 Venue Classification.

Next, we need to extract the surrounding venues and classify them. We will limit the range of the venue to 300 meters from the ice cream shop. After extracting all the venues for all the ice cream shops there are 335 unique venue classification categories, with a wide range of options from "Ethiopian Restaurant" to "Indie Movie Theater". These classifications will be used by machine learning to "learn" the attribute features that create a "good" environment to open an ice cream shop.

3.3 Machine Learning.

Machine Learning is used to score available retail rental locations. To do this we use SciKit Learn's Multiple Linear Regression module. The first step in running regression is to train the model, but before we can do this we must convert the category classification code into an attribute feature set using "one-hot encoding". After the model is trained it is ready to score any new locations.

3.4 Available Locations.

Available locations are found using the screen scrapper, "Beautiful Soup", against the web site "42floors.com". Three screens are scrapped providing 206 available retail rental locations. Figure 2 shows a map of available retail locations. Each location's surrounding venues are then found, extracted then turned into a feature set to be feed into the linear regression model. The regression model scores the locations based on the surrounding venues, the higher the score the better the location.

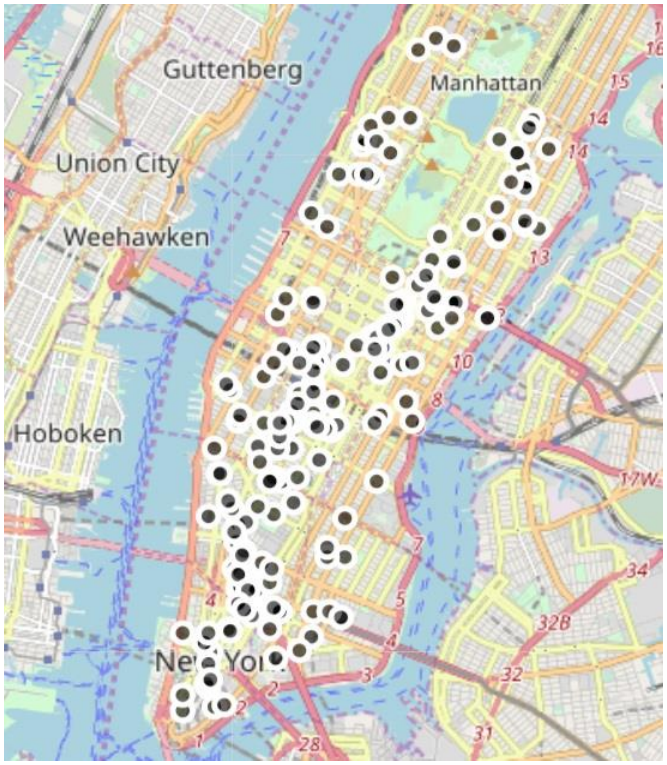

Figure 2 - Available Retail Rental Locations in Manhattan

4.0 Results.

4.1 Top Ten Available Addresses:

Scores are calculated by the Linear Regression model rating the surrounding neighbors of the available rental location. Top ten locations only take surrounding venues into account. A team should assess all the qualifications of each location such as square footage, frontage, cost etc. before making a final determination. The top ten locations are:

1. **Score: 4.18 / 666 Greenwich St , NY, NY.**
2. **Score: 3.56 / 365 7th Ave , NY, NY.**
3. **Score: 3.54 / 714 2nd Ave , NY, NY.**
4. **Score: 3.38 / 133 W 25th St , NY, NY.**
5. **Score: 3.38 / 133 W 25th St , NY, NY. Two sites available.**
6. **Score: 2.9 / 200 West End Ave , NY, NY.**
7. **Score: 2.88 / 1311 Lexington Ave , NY, NY.**
8. **Score: 2.88 / 1311 Lexington Ave , NY, NY. Two sites available.**
9. **Score: 2.82 / 328 330 E 14th St , NY, NY.**
10. **Score: 2.75 / 45 John St , NY, NY.**

4.2 Map of Top Available Locations:

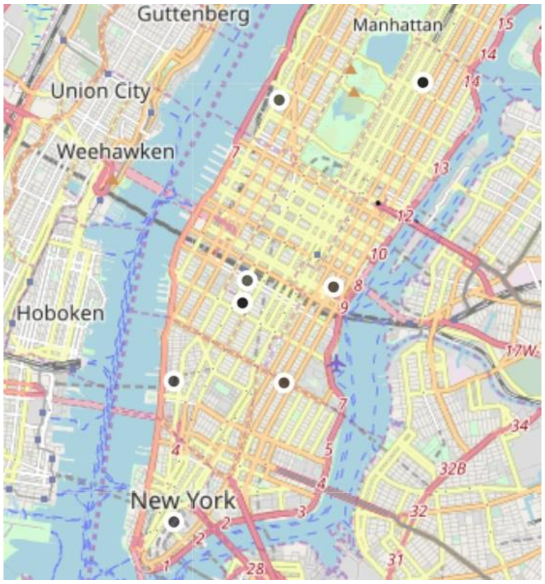

Figure 3 - Top Available Retail Locations

5.0 Discussions.

5.1 FourSquare and the Dataset.

The evaluation of each location is highly dependent on the results from FourSquare. Google Maps is a larger more widely used API providing similar functionality. I would recommend that future analyst is changed to Google Maps. FourSquare is used here mainly because of its no charge pricing, but even within FourSquare, the no charge tier has its limitations. With more detailed information about each venue, the analyst would be better able to judge the size of the venue (i.e.) is the venue a small local bookstore or a large Barnes Nobles bookstore. Using detailed data like number of check-ins would helpful in sizing the venue, plus know the hours of operations would be very useful.

5.2 Increase the Size of the Training Set.

Machine Learning algorithms like Linear Regression have improved accuracy when larger datasets are used to train the model. An idea might be to include other cities details about ice cream shops to increase the size of the data set. But including other cities data might also include other factors. More research needs should be done before expanding the dataset. Another advantage that may be able to be achieved with a larger data set is clustering the neighbors before running the regression models. This would allow us to define what makes a successful ice cream shop in a commercial area vs residential neighborhood. Having multiple regression models for each cluster may help fine tune the analysis.

6.0 Conclusion.

6.1 Top Choice - 666 Greenwich St, West Village, New York City, NY.

A ground team should be sent out to review the space in person. Our evaluation only considers the surrounding neighbor’s ability to support an ice cream shop, and doesn’t take other factors such as cost, frontage, layout, remodeling required etc. into account. 666 Greenwich is 5500 square feet, but the costs haven’t been included. Here are some of the photos of the space: