Identifying the Optimal Location for a New Business
Midterm Report
October 30, 2017

Wei Dai (wd2281)
Amla Srivastava (as5196)
Richie Castellanes (rbc2126)
Daniel First (df2450)

Problem Statement

The objective of this project is to identify the optimal location to open up a new business. We will for the sake of simplicity assume we are opening a Chinese restaurant in Manhattan, at least for the first iteration. To simplify the model, we focus on Chinese restaurant as a general concept and not differentiate between subtypes. Once we develop an initial prototype, we can expand the project to cover to other businesses such as laundromats, groceries, coffee shops, etc.

We also assume that the person opening the business wants to optimize profit for this particular restaurant – rather than, for example, raise awareness of a market brand name.
Because **Profit = Revenue – Cost**, we can begin to estimate Revenue and Cost for each location, as explained in the next section.

Methodology

Estimating Revenue:

We assume the only source of revenue for a Chinese restaurant is orders. The revenue, then, would be the **average price paid per customer** * **the number of customers**.

To estimate the **price paid per customer**, we interviewed key Chinese restaurant managers, who gave us a price range between $10 to $30, depending on the type of restaurant and/or inclusion of alcohol. That said, we can limit this to a $15 price paid per customer. From the Yelp dataset, we have an estimate of price ($, $$, $$$, $$$$).

To estimate the **number of customers**, we plan to use a multi-stage approach

(1) *Baseline followed by adjustment*:
    a. To acquire a baseline, we will speak to the owner of a Chinese restaurant, for whom we have Yelp and Foursquare data
    b. For each other restaurant, we will use the relative data as found on Foursquare (check-ins, tips, etc.) and Yelp (ratings, reviews) to estimate the number of customers.

    This, of course, is an imperfect measure, but it is our closest proxy.

(2) We will also take into consideration the surrounding **competition**, the number of Chinese restaurants in the area from Yelp and/or Foursquare, to temper our estimation of the # of customers.

Estimating Costs:

The costs for a Chinese restaurant can be broken down into *fixed* and *variable* costs.

Fixed costs: Include overhead, electricity, designing the restaurant. Based on our discussion with a Chinese restaurant owner, these fixed costs should be the same, assuming a constant size of restaurants. There are also other costs associated with the restaurant that seem variable but can be assumed to be fixed – namely, the cost of the food and the preparation per customer, as well as the number of workers to hire.

Variable costs: Rent is the biggest variable in this problem as rent across retail locations differs vastly (e.g. retail stores on 5th Avenue vs. retail stores near Columbia). We will estimate this with available rent data for retail locations.

Combining Revenue and Costs:

We are trying to estimate profit (or a profit score) for a Chinese restaurant that will be opened in a given location. We will define a metric that will allow us to estimate it for a given location. If we are searching for the best place to open a Chinese restaurant, our algorithm would calculate the projected profit for each of the potential locations. It would then return the location with the highest expected profit.

Modelling approach:

The next task is to learn a model that will be able to predict an estimated profit for any location based on several location-specific features. These includes features including but not restricted to:

(1) Distance from nearby subways, bus stops, etc. to look into accessibility of the location and in turn of a prospective business
(2) Crime in the area which can be an indicator of popularity of a location
(3) In addition, populations by zip code tied to age and gender as well as overall population density, will be used to determine number of customers who are able to access the business accordingly
(4) Another factor that might influence the profit for a particular establishment is the no. of supplementary businesses in the area. We can look into other businesses such as cafes, dessert shops that might draw more customers.

After building a baseline model, we will experiment with different machine learning techniques to come up with a model with highest accuracy.

Data Collection and Preparation

(1) **Yelp:** The Yelp Fusion API was used to extract information for ~1,000 Chinese restaurants in Manhattan. The data pulled for each restaurant included the name, price, rating, zipcode, latitude, longitude, count of reviews and category information.
(2) **Foursquare:** The Foursquare Places API has check-in data detailed enough that we can infer relative check-ins between restaurants that we can combine with Yelp data accordingly.

(3) **Google Places:** Google Places API has a feature where it provides on a daily basis the most popular times for a business location. We considered using this information, however, the Google Places API currently does not allow this.

(4) **Zillow:** We used the Zillow GetRegion Children API data to get a sense of residential real estate prices that may impact income / spend levels of residents accordingly.

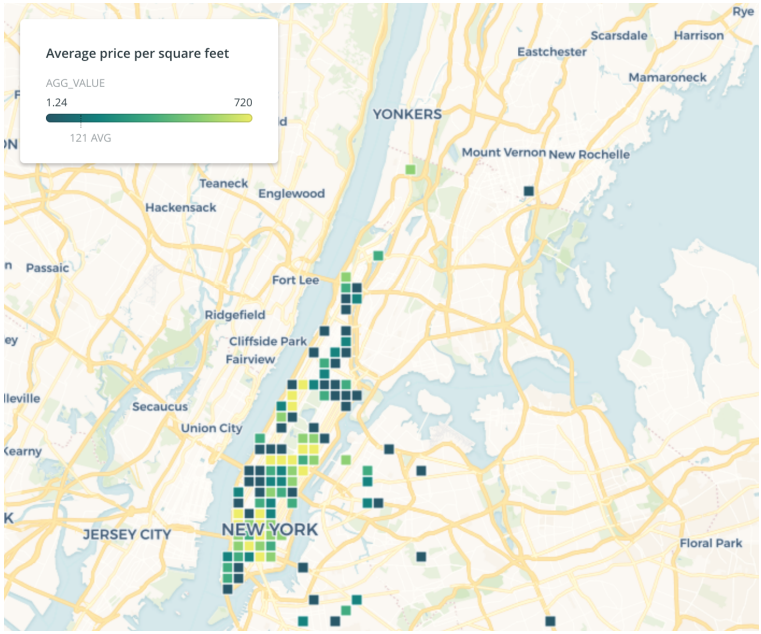
(5) **NYC Open Data:** We gathered population data by zip code cut by age and gender. This will be critical in determining targets that will ultimately help us in determine number of customers. We also pulled NYPD crime statistics from which we will estimate total no. of incidents (potentially broken down by type e.g. fraud, assault, burglary) in a region.

(6) **Loopnet:** We scraped retail rental data from Loopnet, which did not have an API, using BeautifulSoup. We were able to identify properties with retail rental locations available for rent, and their respective price per square foot per yard. We will use this to project our rental costs, which is the most important variable in deriving the cost piece of profit.

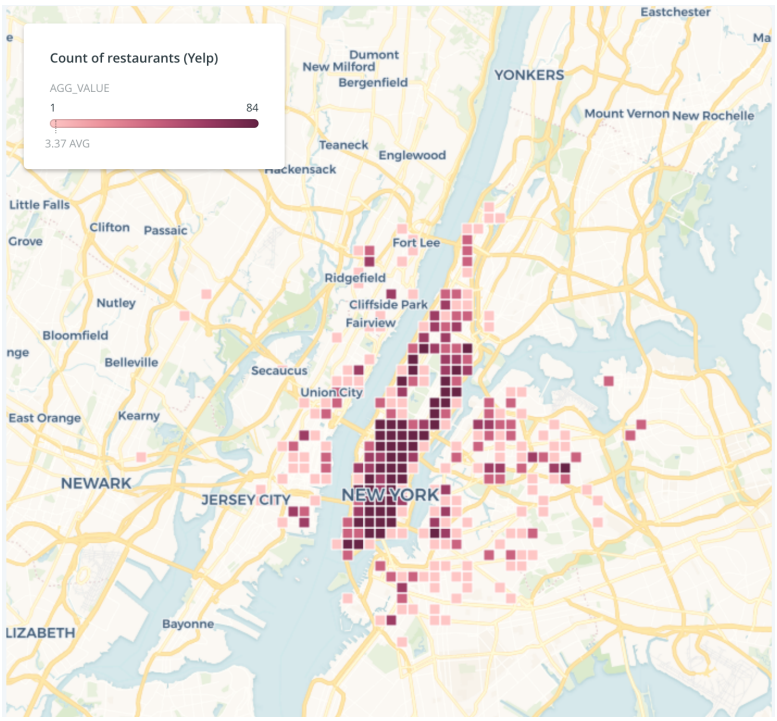
Initial Data Exploration

The datasets pulled from the various sources were cleaned and transformed to a certain degree. We then did an initial data exploration to get an idea of what kind of information is available. As of now, the data can be cut up to zipcode level. For the next steps, we hope to define more granular regions for analysis. Some of our findings are explained below.

First, we looked at price per square foot per year to get an estimate for rent. Most of the high rent areas seem to be in Midtown and the Upper East Side.



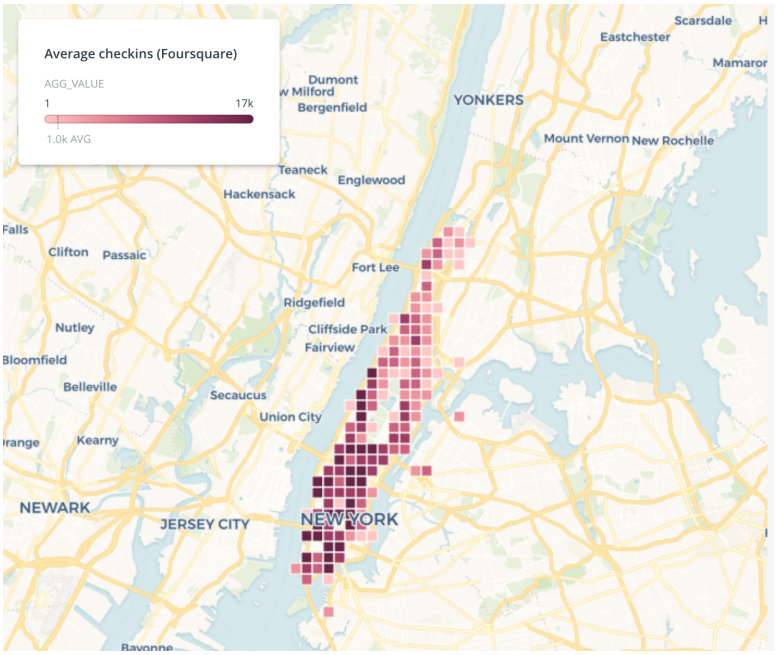
[Figure 1: Average price of square foot per year by retail location (Loopnet as of Oct 2017)]



Next, we were able to get an understanding of volume of Chinese restaurants accordingly, to get an idea of competition. Understandably, there is a high volume of restaurants tied to locations near or around Chinatown as can be seen from the figure below. The bulk of the Chinese restuarants seems to be in lower Manhattan according to Yelp data.

[Figure 2: Count of Chinese restaurants per restaurant location in NYC (Yelp)]

Using Foursquare data, we looked at number of checkins. It seems that Upper Manhattan has significantly lower checkins than locations in Midtown and Lower Manhattan.



[Figure 3: Average of checkins per location in NYC (Foursquare)]

Statement of Goals

(1) *Define location areas.* We have displayed the data at a location level. Based on this, we will need to group these individual locations into more meaningful areas that we will use for our recommendations. This will likely be driven by price per square foot per year among the features, given that there are clear variabilities at certain areas and this will likely be critical if we are to expand the scope of our problem to other types of businesses.

(2) *Model target score based on defined area.* The biggest challenge we currently have in making a recommendation is modeling out profits. While we have laid out features earlier, we will need to ensure that our model outputs scores that make meaningful recommendations. The goal is to fine tune a target score that will provide an accurate recommendation that will likely be extended beyond the current scope of the project.

(3) *Create web app that will provide recommendations based on target.* Once we have the target scores and areas defined, we will create a web app that makes a recommendation on a new business based on some user input on parameters of a business (e.g. price level, target customers by age and sex, and general neighborhoods). This will be our end product that will capture our efforts in making a recommender system for a new business.

Literature Review

The literature review focused on looking into existing applications as well as machine learning algorithms and approaches used in recommendation systems relevant to our problem statement.

Recommendation systems aim to profile user preferences over products and model these relationships which are then used to recommend products that would fit the user's tastes [2]. The systems are usually of three types - collaborative filtering, content filtering and hybrid filtering [1]. Collaborative filtering predicts preferences by calculating the similarity between one user's preferences and the preferences of other individuals. It ignores any a priori information about both users and items. On the other hand, content-based filtering uses information about products to make a recommendation. Collaborative filtering approaches can be applied to recommender systems independently of the domain and are usually more feasible[2]. Nowadays, hybrid techniques of both collaborative and content-based filtering are being increasingly accepted to improve rating prediction. However, those approaches ignore contextual information such as time, weather, location etc. In context-aware recommendation systems (including location aware systems) , user preferences for items depend not only on the items themselves, but also on the context in which items are being considered. [3]

Recommendation systems are used to suggest new products or services to users based on existing information about both users and items [1]. Most of the preceding works focus on location recommendation for location-based social networks using unsupervised methods [4, 5,

6]. They use clustering (Co-Clustering [5]) or matrix factorization (user-item rating matrix) methods. The broad idea of matrix factorization is to construct a user-item preference matrix first, using Collaborative Filtering or probabilistic matrix factorization [6] to fill the matrix, and recommend locations using the similarities between items. The clustering method also, uses user-item data as input, and choose top N nearest clusters as the recommendations for each user [5]. But this scenario is different from our project. We are not looking for recommendation for a business type rather than some specific users. But we can still get what we want from these algorithms, by finding locations shared by most of these users as our output [6]. But this raises a high requirement for our dataset. We will have to own both the user preference on business and the location information of the users.

References

[1] Portugal, I., Alencar, P., & Cowan, D. (2015). The use of machine learning algorithms in recommender systems: a systematic review. *arXiv preprint arXiv:1511.05263*.

[2] Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2009). Scalable collaborative filtering approaches for large recommender systems. *Journal of machine learning research*, *10*(Mar), 623-656.

[3] del Carmen, M., Ilarri, S., Trillo-Lado, R., & Hermoso, R. (2015). Location-Aware Recommendation Systems: Where We Are and Where We Recommend to Go. In *Proceedings of the Workshop on Location-Aware Recommendations* (pp. 1-8).

[4] Ye, M., & Yin, P .(2010). Location Recommendation for Location-based Social Networks.

[5] Leung, K. & Lee, D. (2011). CLR: A Collaborative Location Recommendation Framework based on Co-Clustering

[6] Yang, D. & Zhang, D.(2013). A Sentiment-Enhanced Personalized Location Recommendation System.