Identifying demand in New York to start 'Gym/Fitness' centers

Shripad Ajgaonkar

28 September, 2019

1. Introduction

1.1 Background

With the rise in obesity rate and various diseases related to lack of body fitness, every American has now understood the importance of Gym and fitness. Awareness about body fitness and exercise is widely spread among all age groups similarly. People are finding some time from their busy schedule and are visiting the nearest 'Gym/fitness' center to maintain their body fitness. Hence the demand for such types of fitness centers is rapidly growing and there is huge potential to enter in Health and Fitness business.

1.2 Problem

One popular fitness brand wants to start its fitness center in New York. But the real problem is, to know, in which neighbourhood they should open the center. In order to bring higher 'Return on Investment' (ROI), there should be sufficient number of people who want to join the center. How can they know in advance, which neighbourhoods are best to start their business in order to gain more customers and earn higher revenues?

1.3 Interest

To acquire knowledge about potential demand, fitness brands are dependent on geolocation data analysis and conclusion obtained from it. Hence this analysis serves vital importance to find out neighbourhoods with potential demand and type of "demand segments" they are in.

2. Data acquisition

2.1 Data source

The New York City neighbourhood data are available and can be downloaded from the link https://geo.nyu.edu/catalog/nyu_2451_34572. This data contains data of 5 boroughs and 306 neighbourhoods with latitudes and longitudes. This data forms the foundation of geographical analysis of New York City, since entire New York City can be perceived as formed of these neighbourhoods. This database answers the part of the question, "Which place in New York City".

Secondly, venue data are obtained from Foursquare API. This data helps us to explore the neighbourhoods further. By providing Foursquare credentials and version, venue data can be requested for every neighbourhood in New York City. For the purpose of analysis, I am obtaining venue data for 100 top venues that are in every neighbourhood given in a neighbourhood database within a radius of 500 meters. This data contains venue name, venue category, venue latitude and venue longitude. Venue category comprises of various categories like 'Accessories store', 'Restaurant', 'Gym/Fitness Center'.

2.2 Use of data to serve the purpose of analysis

By using venue data, I will find out most commonly populated categories at every neighbourhood in New York City. This information can be used as a proxy to assess the demand of interested category ('Gym/Fitness' Center) in the selected neighbourhood. The underlined principle is, the obvious demand or need for the set of specific venue categories from the given neighbourhood, encourages commencement and establishment of those categories in that particular neighbourhood.

Based on this information about commonly populated categories, neighbourhoods will be segmented in various 'demand segments' which in turn, will be useful to find out what other supportive categories to 'Gym/Fitness' center are demanded in the same neighbourhood. For example, the more the demand for health and fitness related supportive categories like Health Food, Yoga/Meditation centers, more the scope for growth in the demand of 'Gym/Fitness' category.

Similarly, categories like 'Recreation' would discourage the establishment of 'Gym/Fitness' center in the area.

Apart from this, extent of penetration of 'Gym/Fitness' Center and other venue categories in the neighbourhoods will be found out to assess the association and identify peculiar patterns. This information will be aggregated at segment level to determine the potential for 'Gym/Fitness' center.

2.3 Features extracted from data

Following features are created by making some assumptions,

- **Boutique** – Existence of boutique emphasizes the more localization of the neighbourhood. Surrounding of good amount of boutiques symbolizes neighbourhood is residential type. Since the people have a tendency to visit nearby 'Gym/Fitness' center, highly residential neighbourhood has greater potential.
- **Arcade** – Arcades are the constructions where people meet or gather for completely different purposes other than for 'Gym/Fitness' centers. They can travel far away to visit these places. This place represents non-native type of the neighbourhood. These places are concentrated in much specialized area of neighbourhood, for example, one can resemble, commercialized area.
- **Cafeteria** – It is the prominent meeting, gathering or discussion area for small group of family as well as business people. Its existence is scattered across native-residential as well as diverse-commercial zones. Hence the area surrounded by Cafeteria keeps high potential for the growth of 'Gym/Fitness' centers.
- **Food & Wine** – Its existence also is scattered across native-residential as well as diverse-commercial area. Although, it brings a little bit more residential flavour due to family dine out or local street food requirements. Hence anticipates little more potential for a 'Gym/Fitness' center.
- **Super market & traditional Grocery** – In recent time span, supermarket owners are trying to open 'Gym/Fitness' center on market premises. It has been proven a good experimentation to keep people attracted to the market by increasing their engagement levels. While the penetration of traditional grocery stores indicate localization of neighbourhood and necessitates the demand for fitness in the proximity of residential area with the increasing

potential, business potential for a ‘Gym/Fitness’ center has been swamped in the area of Super market.

- **Health** – In the residential area where people are more conscious towards health and particularly following their health habits or diets, these areas are keeping more potential for ‘Gym/Fitness’ centers. Hence, the area where services like, 'Gluten-free Restaurant', 'Health & Beauty Service', 'Health Food Store', 'Massage Studio', 'Organic Grocery', 'Physical Therapist', 'Salad Place', 'Spa', 'Spiritual Center', 'Weight Loss Center', 'Yoga Studio' are found, these areas are good prospects for ‘Gym/Fitness’ center.
- **Recreation** – These are the places, where generally people come for enjoyment, fun, rest or entertainment purpose. While they are here, they are in the complete resting mood. The purpose they seek is completely different than ‘Gym/Fitness’ centers can offer. Hence the area where recreational places are situated are not suitable for ‘Gym/Fitness’ centers.
- **Sports** – These are the places where sports activities are conducted. This serves the substitute purpose of the ‘Gym/Fitness’ center. All the sports places offer a substitute service of fitness in addition. Hence the area where sports activities are conducted are not suitable places for ‘Gym/Fitness’ centers.
- **Public transport** – The purpose served by the ‘Gym/Fitness’ center has very little demand in the area where public transport like bus, train or metro station are located. This is because, ‘Gym/Fitness’ centers are sought by people in the nearby residential area.
- **Feminine** – It is said that women are more conscious about their health and fitness. Hence now days there is increasing demand from women. As their engaging nature, they prefer their training centers in the area where services like 'Baby Store', 'Cooking School', 'Cosmetics Shop', 'Lingerie Store', 'Nail Salon', 'Women's Store' are being provided. Hence these types of areas have more potential for woman ‘Gym/Fitness’ centers.

3. Methodology

3.1 Neighbourhood data download and data frame creation

As described above in data acquisition session, I downloaded neighbourhood data of New York City from the link https://geo.nyu.edu/catalog/nyu_2451_34572. This data contains data of 5 boroughs and 306 neighbourhoods with latitudes and longitudes. The data is in JSON format. This data is then loaded in python data frame.

3.2 Venue data acquisition from Foursquare API

After providing ‘Foursquare’ credentials, venue data is obtained from ‘Foursquare’ API for every neighbourhood listed in neighbourhood data. For each neighbourhood, top 100 venues are extracted within the radius of 500 meters along with specific venue category. This venue data is then loaded in python data frame. This data consists of total 10,381 venues for all listed neighbourhoods. Initially, there were 428 unique venue categories.

3.3 Data pre-processing and feature creation

Now these venue categories are regrouped to create new feature categories based on our business ideas or assumptions the way business works as specified in above session, 2.3 (Feature extracted from data).

Sr. No.	New Genre/Feature	Venue Categories Grouped
1.	**Boutique**	All types of Boutique shops, Clothing shops
2.	**Arcade**	Arcades, Buildings, Community centers, Landscapes, Event places, All types of markets (Farmer/Fish/Flea), Heliport, Hostel, Memorial sites, Monument/Landmark, Motels, Newsstands, Offices/Co-working spaces, Outdoors, Pier, Platforms, Post Office, Social clubs
3.	**Cafeteria**	Tea shops, Cafeteria, Coffee shops
4.	**Food & Wine**	All types of Restaurants, BBQ points, Breakfast shops, Bagel shops, Bakery, Burger shops, Burrito places, Food/Fast food shops, Gourmet shops, Hot Dog shops, Ice cream shops, Pastry/Pie shops, Sandwich/Pizza shops All types of beer bars, cocktail bars, Liquor stores, Pubs, Lounge, Night clubs, Whisky bars, Wine bars
5.	**Super market**	All types of Department stores, Discount stores, Shopping malls, Super markets
6.	**Grocery**	All types of Convenience stores, Deli/Bodega, Grocery stores
7.	**Health**	Gluten-free Restaurants, Health & Beauty Services, Health Food Stores, Massage Studios, Organic Grocery stores, Physical Therapist, Salad Places, Spa, Spiritual Centers, Weight Loss Centers, Yoga Studios
8.	**Recreation**	Art galleries/Museums, Craft stores, Entertainment centers, Auditoriums, Bath houses, Club houses, Comedy clubs, Concert halls, History museums, Movie theatres/Multiplexes, Jazz clubs, Opera houses, Piano bars, Performing arts venues, Plaza, Public arts, Parks/Theme parks, Game centers, Beaches, Lakes, Harbours, Fountains
9	**Sports**	All types of sports activity places in Athletics, Baseball, Basketball, Bowling, College sports, Golf, Climbing, Skating, Soccer, Sports goods, Sports clubs, Surfing, Tennis, Volleyball etc.
10.	**Public Transport**	Bus/Train/Metro stations
11.	**Feminine**	All types venues related to female like Baby stores, Cooking schools, Cosmetic shops, Lingerie stores, Nail salons and other Women's stores

After regrouping, there remained 143 unique new venue categories. This new venue categories are used for analysis.

3.4 Feature extraction

For all venue entries in the neighbourhoods, dummy variables are created for newly created features categories. For example, for a specific Food/Restaurant venue situated in the specific neighbourhood, '1' will be coded as new feature category - 'Food & Wine' so on. So, for 10,381 venue entries within neighbourhoods in database, 143 dummy variables for new feature categories are created.

Now, this data is grouped at the neighbourhood level and means/averages are calculated for all 143 feature categories. The data consist of means/averages of 143 feature categories for all neighbourhoods. Mean/average indicates penetration/distribution of each feature category in the given neighbourhood.

This penetration/distribution data is now ready to analyse further.

3.5 Cluster analysis – Using 'K-means' cluster algorithm

Here, our core purpose is to identify distinct neighbourhood segments based on patterns in penetration data. Every neighbourhood segment may be unique in itself and distinct from other neighbourhood segments. Every segment will be consisting of a peculiar penetration pattern of some specific features. These segments will be studied/profiled in terms of these patterns. These segment profiles will help us to decide the neighbourhood segment and hence finally the neighbourhood to plan the business.

1. Finding the appropriate cluster number

Unlike supervised learning, where we have actual event data to evaluate the model's performance, clustering analysis doesn't have a solid evaluation metric that we can use to evaluate the outcome of the model. Also, since K-means requires a number of clusters as an input and this information cannot be derived from data, there is no right answer in terms of the number of clusters. In the cluster-predict methodology, we can evaluate how well the models are performing based on different K clusters since clusters are used in the downstream modeling.

Elbow method is used to assess the right number of cluster with minimum WCSS (Within Cluster Sum of Squares). Based on Elbow method 3 is right number of clusters.

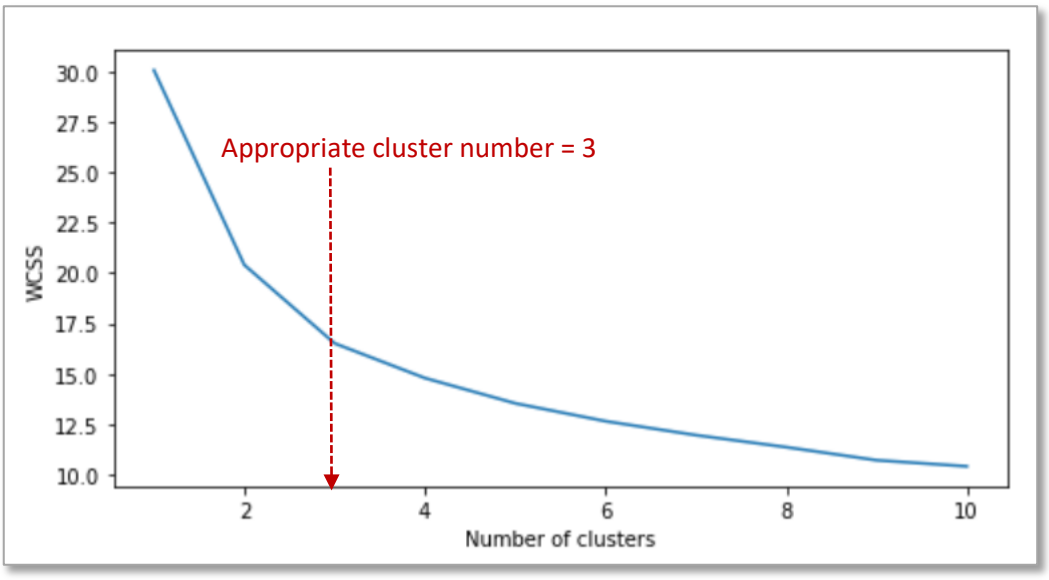

Figure 1: Using elbow method to derive appropriate number of clusters

2. Cluster formation

Based on appropriate cluster number, that is 3, I ran K-means cluster algorithm to form 3 clusters. Here, three neighbourhood clusters are formed.

3. Finding correlation of Gym/Fitness centers with other venue categories

After segmentation, penetration data is averaged at segment level (obtained from model). Pearson correlation is calculated between 'Gym' and other feature categories by using this data. It is plotted by using bar plot.

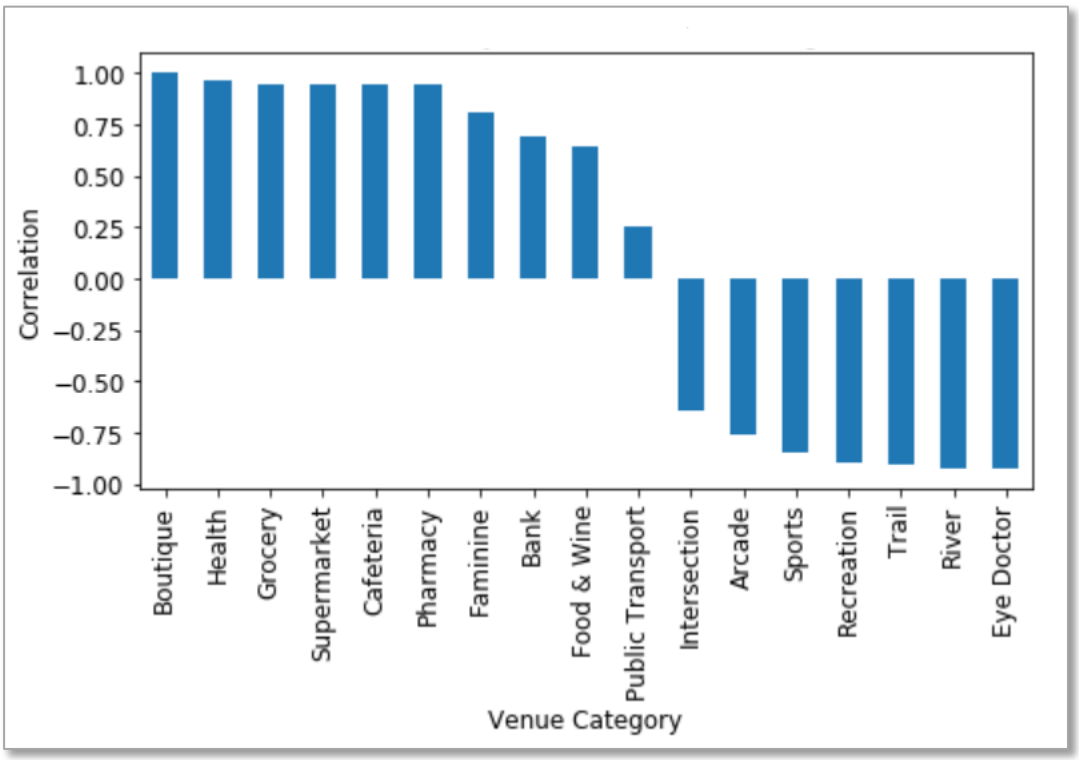

Figure 2: Associating penetration of 'Gym/Fitness' centers with other feature venue categories

4. Understanding segment profiles

Now, each segment is separately studied for penetration of all feature categories. Distinct peculiarity of each segment will be found out based on penetration pattern within the segment. Based on profile, we will come to know potential of each segment to start the business of 'Gym/Fitness' center.

I have plotted penetration of features for every segment.

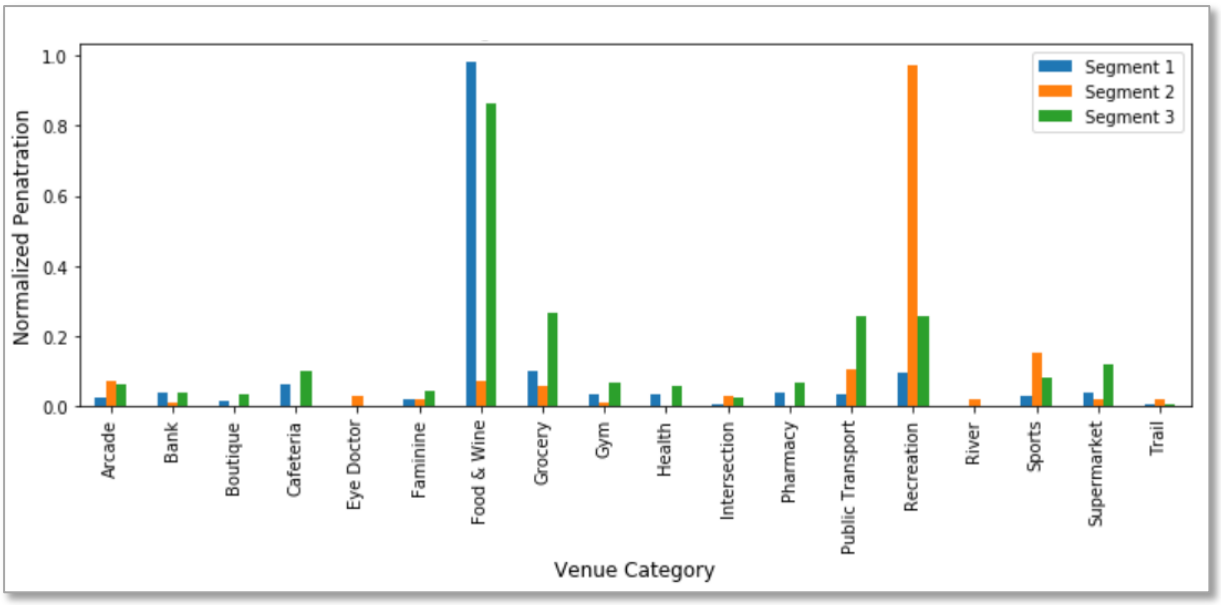

Figure 3: Analysis of penetration patterns of feature categories across segments

4. Results

Part I

Based on correlation analysis, presented in part 3 (Finding correlation of Gym/Fitness centers with other venue categories) of session 3.5, there are certain observations we can make out,

1. 'Boutique', 'Health', 'Grocery', 'Supermarket', 'Cafeteria' and 'Pharmacy' have high positive correlation with 'Gym' penetration.
2. 'Feminine', 'Bank', 'Food & Wine' have moderate positive association with 'Gym' penetration.
3. 'Public Transport' has least positive association with 'Gym' penetration.
4. 'Intersection', 'Arcade' and 'Sports' have moderate negative association with 'Gym' penetration.
5. 'Recreation', 'Trail', 'River' have high negative association with 'Gym' penetration.

Part II

Based on the profile analysis in part 4 (Understanding segment profiles) of session 3.5, we arrived at segment structure like,

1. **Segment 1 – Moderate potential for Gym**
   - **High penetration** of 'Food & Wine', 'Bank'
   - **Moderate penetration** of 'Cafeteria', 'Boutique', 'Feminine', 'Grocery, 'Super market', 'Health', 'Pharmacy'
   - **Low penetration** of 'Arcade', 'Intersection', 'Public Transport', 'Recreation', 'Sports', 'Trail'
2. **Segment 2 – Low potential for Gym**
   - **High penetration** of 'Arcade', 'Eye Doctor', 'Intersection', 'Recreation', 'River', 'Sports', 'Trail'
   - **Moderate penetration** of 'Public Transport', 'Feminine'
   - **Low penetration** of 'Bank', 'Food & Wine', 'Grocery', 'Super market'
3. **Segment 3 – High potential for Gym**
   - **High penetration** of 'Grocery', 'Health', 'Pharmacy', 'Public Transport', 'Super market', 'Boutique', 'Cafeteria', 'Feminine'

- **Moderate penetration** of ‘Arcade’, ‘Food & Wine’, ‘Bank’, ‘Intersection’, ‘Recreation’, ‘Sports’
- **Low penetration** of ‘Trail’

5. Discussion

‘Boutique’, ‘Health’, ‘Grocery’, ‘Supermarket’, ‘Cafeteria’ and ‘Pharmacy’ represent localised nature of neighbourhood through native residential needs like consumption, engagement and fitness consciousness. High penetration of these categories associates with high demand for Gym/Fitness centers and indicates high potential for the business.

These categories highly fabricate profile of Segment 3.

‘Feminine’, ‘Bank’, ‘Food & Wine’ mostly represent engagement needs of people. Penetration of these categories is associated with moderate demand for ‘Gym/Fitness’ centers and indicates moderate potential for the business.

These categories highly fabricate profile of Segment 1. ‘Feminine’ forms the structure of Segment 2 as well.

‘Public Transport’, ‘Intersection’, ‘Arcade’, ‘Sports’, ‘Recreation’, ‘Trail’, ‘River’ indicate very different purposes, needs or goals that people are interested in with respect to ‘Gym/Fitness’. Hence penetration of these categories associates very little or no demand.

These categories highly fabricate the profile of Segment 2.

6. Conclusion - How this analysis can be used…?

Since all the neighbourhoods in the New York City are segmented in 3 segments and are labelled for segment membership, every single neighbourhood draws the picture of future potential for the ‘Gym/Fitness’ center by virtue of segment profile.

For example, neighbourhood Allerton is classified in Segment 1, then it can be perceived as moderate potential neighbourhood for ‘Gym/Fitness’ business.

All the neighbourhoods with segment membership for segment 3 are prioritized first, since this segment has high potential for ‘Gym/Fitness’ business.

All the neighbourhoods with segment membership for segment 2 are avoided, since this segment has very little or no potential for the business.