Exploring "Drugstore Errand Runs" in Manhattan Neighborhoods

Antonio Miceli

April 29, 2020

1. Introduction

1.1. Background

Despite the growing use of delivery services in a post-COVID outbreak world, including drugstore deliveries by Amazon Pillpack, Doordash, and local mail-order pharmacies, many Americans still go on "drugstore errand runs" at least once a month.[1]

Not only do seven out of ten Americans regularly take a prescription medication, but 45% of them (119 million in 2016) also take controlled medications, many of which cannot be shipped by mail-order pharmacies because they are scheduled as "Controlled Substances" by the DEA (Drug Enforcement Administration). Moreover, according to J.D. Power's 2019 U.S. Pharmacy Study, respondents still vastly prefer their brick-and-mortar pharmacy to currently available digital options.[2] That consumers prefer their neighborhood pharmacy to delivered options may change in a post-COVID world, but perhaps a more interesting question is, what's not changing?

Unless the DEA government relaxes regulations on the delivery of controlled medications, and, until consumers readily adopt mail-order pharmacies in a post-COVID world, there will still be plenty of people regularly visiting their brick-and-mortar drugstore, with other goods and services potentially being picked up along the way.

1.2. Business Opportunity

While drugstore companies can run internal product mix and customer profile analyses on what non-prescription drug goods are being sold at their own locations, an interesting question remains: if not immediately home, where do their customers go after? Are there unexplored revenue opportunities and competitive advantages?

At the same time, the "convenience" retail landscape is experiencing a paradigm shift— cashier and cashless options such are opening in dozens of city centers. These operators are quickly iterating their offerings based on internal data from its users. Apart from being able to implement pharmacies onsite (and front-end automation of the pharmacy counter is another interesting question,) are these automated stores offering their neighborhood customers what they are looking for on their regular drugstore errand run? What are consumer foot traffic patterns on these essential "errand runs"? What are neighborhood locations are promising for automated stores?

This short study explores potential answers to these questions in Manhattan neighborhoods by mapping consumer location data and employing machine-learning algorithms and statistical methods to discover interesting relationships that could be extended to predictive models for other cities and metropolitan areas.

1.3. Interest

As initially presented in the previous questions, this exploratory study is of primary interest to existing (1) pharmacies and drugstores, as well as (2) new-entrant automated retailers who are validating and

[1] cvshealth.com/thought-leadership/by-the-numbers-how-do-consumers-interact-with-pharmacists
[2] jdpower.com/business/press-releases/2019-us-pharmacy-study

rapidly expanding their business models.[3] It can also be of interest to (3) marketing and consumer research firms, (4) existing brick-and-mortar neighborhood services looking to adapt their business models, as well as (5) other startups of various categories with innovative business models.

2. Data and Methodology

2.1. Sources

For this short study, the following sources and uses are employed:

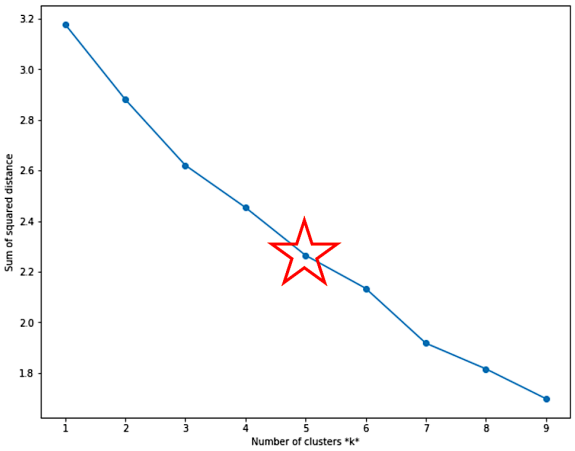
i. Venue and user location data from Foursquare's Places API[4]
ii. Neighborhood coordinate data from NYU's Spatial Data Repository[5]

2.2. Preparation, Selection, and Methods

i. Coordinate points for 40 Manhattan neighborhoods were extracted from data source (ii). These points establish the neighborhood centroids, which were used for the initial Foursquare API data call.
ii. The initial Foursquare API data call queried for the most visited pharmacies for each neighborhood centroid, limited to five venues within a 500-meter radius. 185 unique venues were returned. It is worth mentioning that a more conclusive method would identify the commercial zones across a city or specified metropolitan area, as this could better adapt the model for other cities whose neighborhoods are less dense than Manhattan (the eighth densest city in the world.)
iii. I cleaned up the retrieved data by removing locations that do not contain pharmacy counters that sell doctor-prescribed medications. From the 185 returned venues, the initial dataset ultimately yielded 169 pharmacies in 40 Manhattan neighborhoods. This cleaned dataset was used to make the subsequent API call.
iv. After the initial API call, the unique Foursquare venue identification numbers for each pharmacy were selected to make a secondary API call, which queried the most popular "NextVenues" visited after each Venue Id. By default, a NextVenues query returns up to five NextVenues. It is important to mention that the venue data in this short study only examines locations visited *after* the queried drugstores. A more comprehensive study would include venues visited before the drugstore venue as well, as we cannot reasonably assume that
v. The secondary API call data was transformed and prepared for K-Means clustering by organizing the NextVenue Categories from each pharmacy venue and calculating the mean value of the NextValue Categories by neighborhood via one-hot encoding.
vi. Before fitting the data for clustering, it's worth mentioning some of the limitations to the Foursquare dataset. For example, ideally there should have 25 next venues for each neighborhood (5 from each of the 5 pharmacies,) but not every pharmacy venue returned the full 5 NextVenues, and some did not return any NextVenues at all. So, neighborhoods that returned less than 5 results were excluded. From the initial 200 pharmacy venues, there were ultimately 105 pharmacies whose NextVenues were used for clustering.
vii. However, for the ones that did return, they are sorted by popularity. The dataset used for KMeans ultimately contained 522 NextVenues in 36 Manhattan neighborhoods.

[3] forbes.com/sites/andriacheng/2019/06/26/amazon-gos-even-bigger-rollout-is-not-a-matter-of-if-but-when/#2fd32aa16f52
[4] developer.foursquare.com/docs/places-api/
[5] geo.nyu.edu/catalog/nyu_2451_34572

viii. The optimal numbers of k-clusters were evaluated via “elbow” and “silhouette-score methods”, resulting in 5 clusters. It's worth mentioning that the silhouette method actually achieved a higher score when k=8 (.0847), but those results presented a seemingly overfitted model that did not provide an actionable level of insights compared with k=5 (with a score of 0.0674). See Plot 1 below.



Plot 1. “Elbow Method” line plot of k-clusters

ix. Apart from cluster analysis, simple linear regression, logistic regression and Support Vector Machine (using RBF kernel) modeling were conducted to test a hypothesis derived from examining relationships between features of the first and second API calls and to provide a basis for further statistical evaluation in other geographical areas. The continuous regression model attempts to predict the number of pharmacy NextVenues based on the number of medical venues in the same neighborhood. The categorical logistic and SVM models attempt to predict the presence of pharmacy NextVenues Additional data required for hypothesis testing was supplied by a third Foursquare API query. Train/test split for the linear model was set at 75/25 of the compiled dataset.

3. Results

3.1. KMeans Clustering Calculation and Relationships

Fig. 1 below displays the 169 pharmacies queried for NextVenues in 40 Manhattan Neighborhoods.

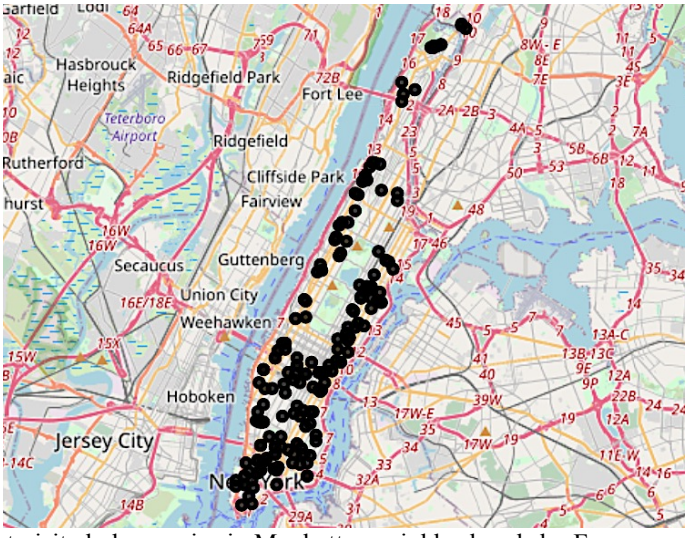


Fig 1. Most visited pharmacies in Manhattan neighborhoods by Foursquare Checkins

Fig 2 below shows the map-plotted results of machine unsupervised Clustering. Clusters 1-5 are colored orange, red, purple, light blue and light green, respectively.

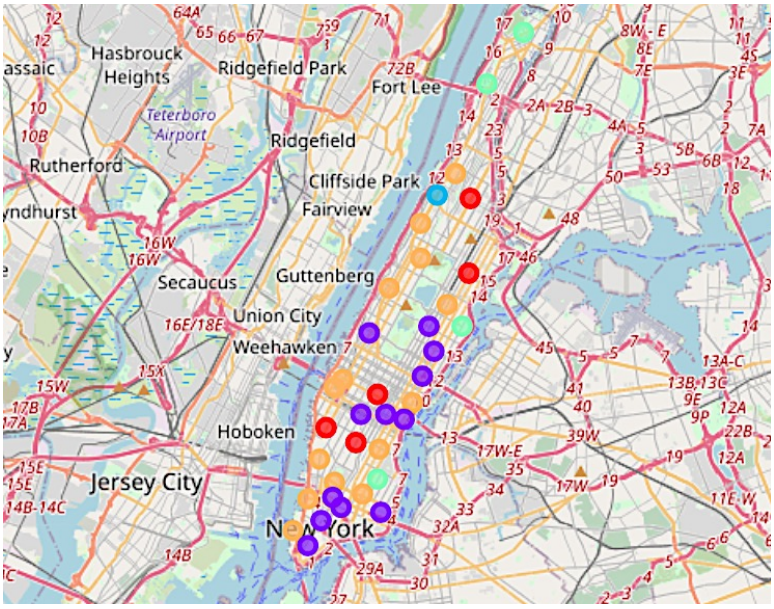


Fig 2. K-Means clustering results by neighborhood for NextVenues to pharmacies in Fig 1.

Clusters 1-5 yielded neighborhood counts of 14, 5, 12, 1, and 4, respectively. Cluster 4, consisting of only one neighborhood (Manhattanville), is characterized by containing NextVenues consisting entirely of food-oriented establishments, with all locations being restaurants except for one grocery store. This single neighborhood cluster is bordered by neighborhoods labeled Clusters 1 and 2.

Clusters 2 and 5 are both characterized by their "1st Most Common NextVenue" containing almost all of the same categories. Cluster 2 contains "Park" as a NextVenue than any other cluster, whereas Cluster 5 proportionally contains "Grocery Store" and "Supermarket" more than any other cluster.

Clusters 1 and 3 contain the most neighborhoods than the other three clusters, with 14 and 12 neighborhoods in each cluster, respectively. These two clusters share many of the same NextVenue Categories and it is harder to apparently see what could characterize their different classification.

One interesting feature is that Cluster 3 contains seven pharmacy NextVenues (more than any other Cluster), while Cluster 1 only contains two pharmacies. This result, along with the feature of pharmacies as NextVenues for a pharmacy in general is interesting and will be explored further.

3.2. Pharmacy NextVenue Calculations and Relationships

Table 1 and Fig. 3 below show the frequency of the "Top 10" NextVenue Categories. Grocery Store and Supermarket (as related venues to purchase non-restaurant food) make up more than 20% of the Categories, while Coffee Shop is the strongest independent category, at 11%. Along with Pharmacy, these five categories make up almost half (48%) of the NextVenue Categories from the most popular Manhattan Pharmacies. It is interesting that 7% of the NextVenues are another pharmacy, as this can suggest that the outcome of the initial visit was not sufficient to satisfy the purchase intention.

	NextVenue Category	Frequency
0	Grocery Store	69
1	Coffee Shop	57
2	Park	46
3	Supermarket	41
4	Pharmacy	37
5	Department Store	21
6	Plaza	17
7	Sandwich Place	14
8	Deli / Bodega	10
9	Other	209

Table 1. Frequency of top NextVenues by Category

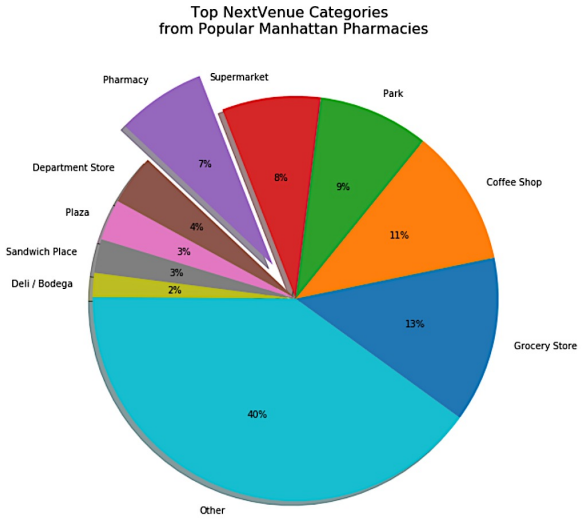


Fig 3. Pie chart of Table 1 showing percentages for top NextVenues

Although the data is limited as to item purchases, are there any interesting relationships between the initial venues and the pharmacy NextVenues? Hypotheses for potential relationships include (1) "Chain" and "Independent" pharmacies, and (2) neighborhoods with a large concentration of hospitals or medical facilities.

With respect to hypothesis (1), Fig. 4 below is a modified verision of Fig 1., which shows the same five most visited pharmacies in each neighborhood, but also distinguishes chain from independent pharmacies by color (blue markers represent chains while red markers represent independents.) When the NextVenues for were grouped for independent pharmacies, there was only one case where a NextVenue for an independent pharmacy is a chain pharmacy. Additionally, no independent pharmacies were NextVenues for chain pharmacies.

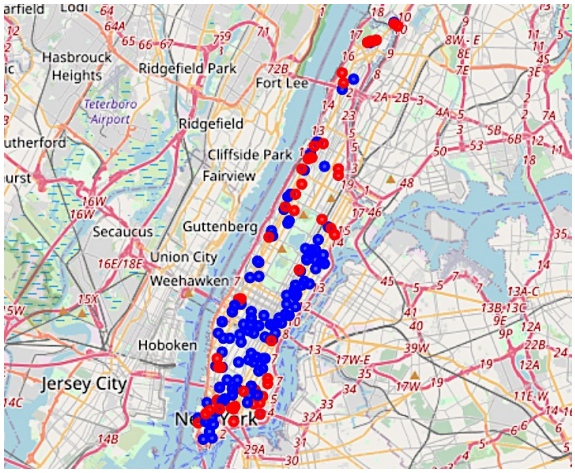


Fig 4. Most visited pharmacies in Manhattan neighborhoods by chain or independent brand

Although these observations would suggest no significant relationship between chain and independent pharmacies after a Foursquare check-in for either kind of pharmacy, it is worth mentioning that out of the 50 independent pharmacies queried for NextVenues, only 17 returned any NextVenues at all. A more determinate rejection of this hypothesis would at least include NextVenues from all independent pharmacies throughout the originally queried neighborhoods.

Notwithstanding, granular observations at the neighborhood level can provide some actionable insights for organizing feature selection in follow-up analyses with more complete datasets. Fig. 5 below

visualizes the sole observation where a chain pharmacy in Tribeca is a top NextVenue for an independent pharmacy in the Civic Center neighborhood. In this case, users will readily visit the chain pharmacy after the independent pharmacy. Also, both pharmacies share a grocery store (Whole Foods Market) as a NextVenue (where the colored circles overlap on the corner of Greenwich and Warren Streets).

Chain/Indep	Neighborhood_y	Pharmacy	NextVenue	NextVenue Latitude	NextVenue Longitude	NextVenue Category
Independent	Civic Center	Kings Pharmacy	Whole Foods Market	40.715579	-74.011368	Grocery Store
Independent	Civic Center	Kings Pharmacy	Morgan's Market	40.716298	-74.009294	Deli / Bodega
Independent	Civic Center	Kings Pharmacy	Duane Reade	40.719183	-74.010775	Pharmacy
Independent	Civic Center	Kings Pharmacy	Bed Bath & Beyond	40.715807	-74.011804	Furniture / Home Store
Independent	Civic Center	Kings Pharmacy	Bogardus Plaza	40.715962	-74.009064	Plaza
Chain	Tribeca	Duane Reade	Whole Foods Market	40.715579	-74.011368	Grocery Store
Chain	Tribeca	Duane Reade	Benvenuto Cafe Tribeca	40.719503	-74.010269	Sandwich Place
Chain	Tribeca	Duane Reade	Pier 25 - Hudson River Park	40.720193	-74.012950	Park
Chain	Tribeca	Duane Reade	Maison Kayser	40.718909	-74.010331	Bakery
Chain	Tribeca	Duane Reade	Washington Market Park	40.717046	-74.011095	Playground

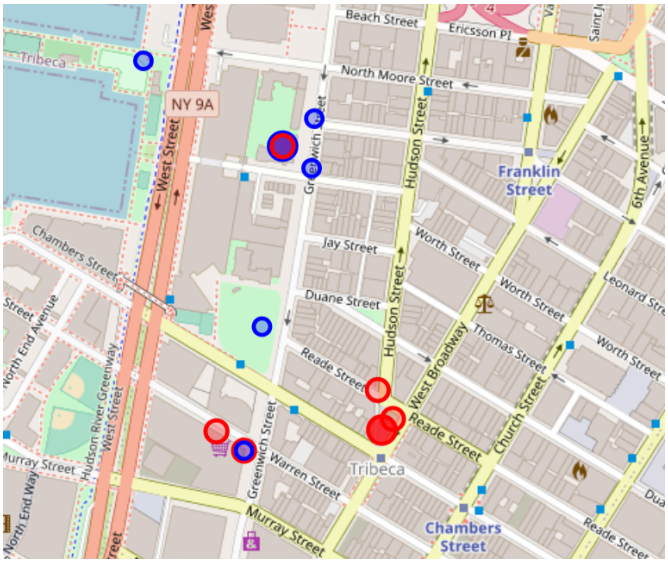


Table 2. NextVenues data for neighboring independent (red rings in Fig. 5) and chain (blue rings) pharmacies
Fig 5. Chain (solid blue circle) and independent (solid red circle) pharmacies with their respective NextVenues

While hypothesis (1) occurred from looking at the brand names of the originally returned venues, hypothesis (2) occurred only after plotting the pharmacies and their pharmacy NextVenues. Fig. 6 below shows the initial pharmacy venues represented with blue locations markers, along with their pharmacy NextVenues in purple circles. Red lines join the venues and their NextVenues.

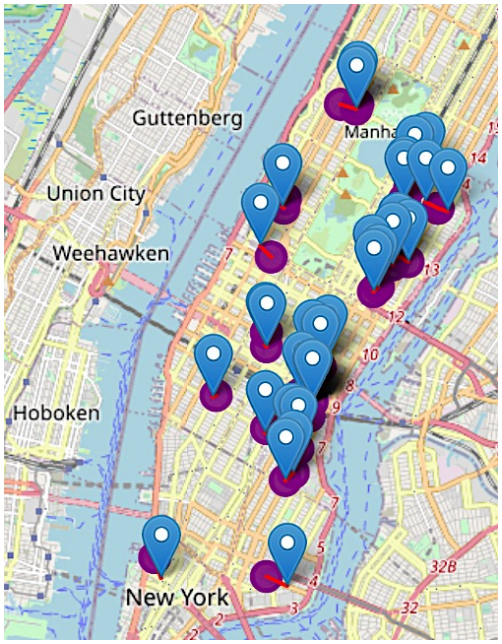
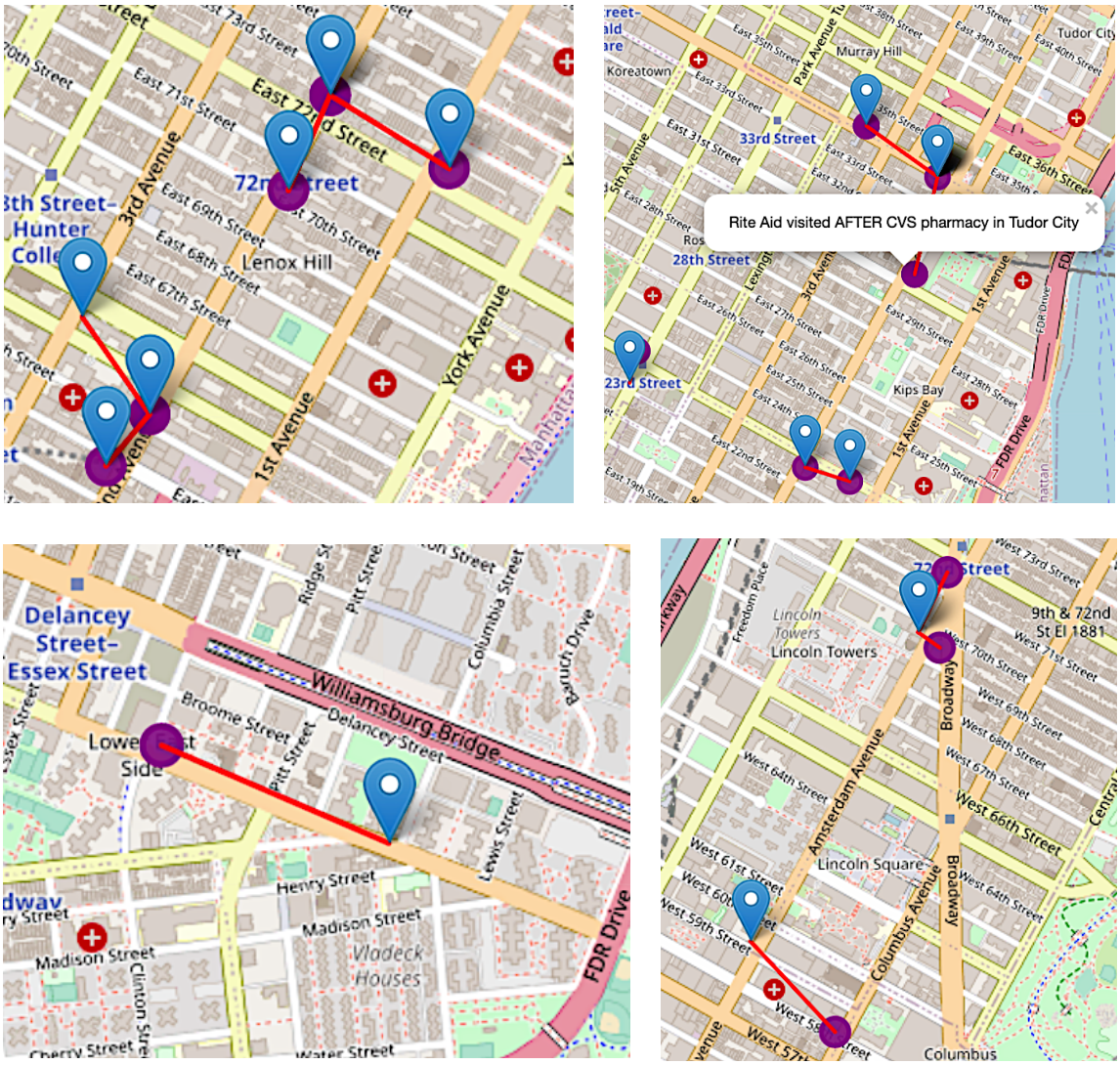


Fig 6. Manhattan pharmacy venues and their pharmacy NextVenues

This spatial map plotting provided two unexpected insights: (1) for the pharmacy NextVenues (7% of total NextVenues), these pharmacy NextVenues occurred exclusively in 17 out of the 36 researched

Manhattan neighborhoods, and (2) a supermajority of the pharmacy NextVenues are pharmacy venues themselves! Why these neighborhoods and not others?

This unequal spatial and statistical distribution led to closely examining the map in Fig. 7, where several repeated venue map features were observed. Images 1-4 below are salient examples of those observations, which show pharmacy NextVenues repeatedly occurring in areas with hospitals/medical centers, and, adjacent centrally planned complexes in different parts of Manhattan. These observations led to hypothesis (2).



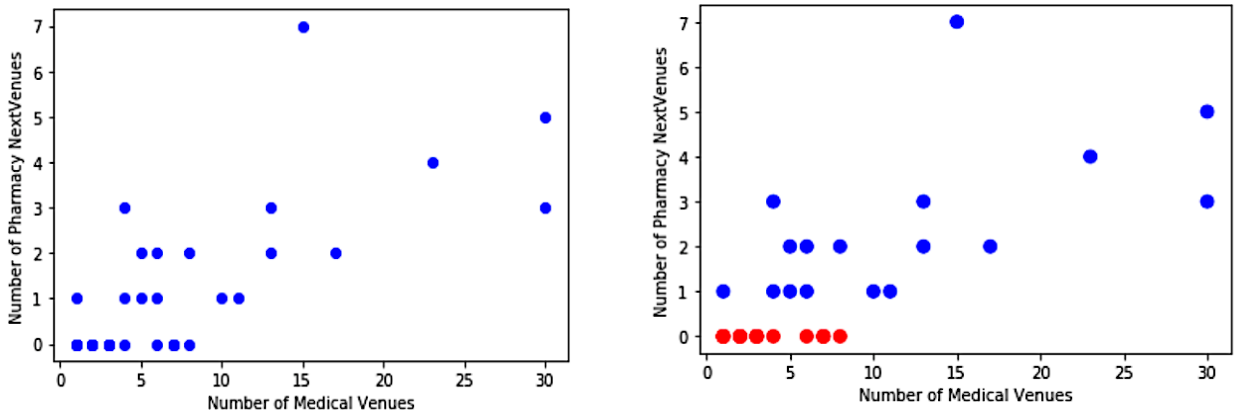
Figs 6.1=6.4 . Close-ups of Fig 6 showing Lower East Side, Murray Hill,
Upper East Side, Upper West Side neighborhoods (from top right to bottom left)

The presence of pharmacy Venue/NextVenues suggest unmet demand at individual locations, and, that there may be uncovered opportunities for automated retailers (including top NextVenue categories such as coffee shops and grocery stores) to on or nearby these Venue/NextVenue areas. This short study assumes that this is an advantageous feature for future locations. More research would need to be conducted to validate this business assumption.

Plots 2.1 and 2.2 below are scatter plot representations of Fig. 6, where each plot is a neighborhood placed by the number of medical center venues in that neighborhood (x-axis) as well as the number of pharmacy NextVenues in that neighborhood (y-axis).

As can be viewed from the below seventeen neighborhoods have at least one pharmacy NextVenue with a medical venue in that same neighborhood, compared with seven neighborhoods where no medical venue exists in that neighborhood. Pearson correlation for Plot 2.1 is 0.57, which signifies a moderately positive relationship.

Plot 2.2 is identical to 2.1, except that it plots neighborhoods that have zero pharmacy NextVenues in red. This distinction defines the target of the categorical predictive models where we lack reliable pharmacy NextVenue data; in other words, we want to be able to predict the presence of pharmacy NextVenues based on the number of medical venues in that neighborhood.

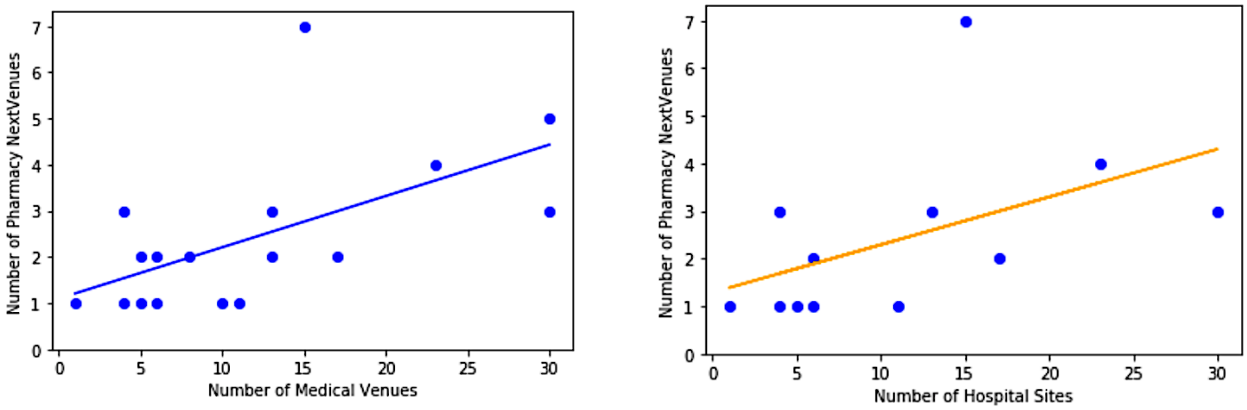


Plot 2.1-2.2. Scatter plot of Manhattan neighborhoods by number of pharmacy NextVenues and medical venues within that neighborhood

4. Predictive Modeling

One continuous and two categorical models were created to help predict the presence and quantity of pharmacy NextVenues based on the number of surrounding medical venues. Plots 3 and 4 below are training and tests sets of a simple regression model that assumes that pharmacy NextVenues exist in neighborhoods where there are medical venues.

While an r^2 score of 0.34 is not characteristically robust, it is not out of line with our eye-level map observations that identified hospital/medical venues as one of several potential explanations for the pharmacy NextVenues. We also noticed centrally planned residential complexes as a repeated observation. A suggestion for additional analysis would include additional features and the data should be normalized accordingly.



Plot 3 and 4. Training (left) and test (right) set results of simple linear regression model

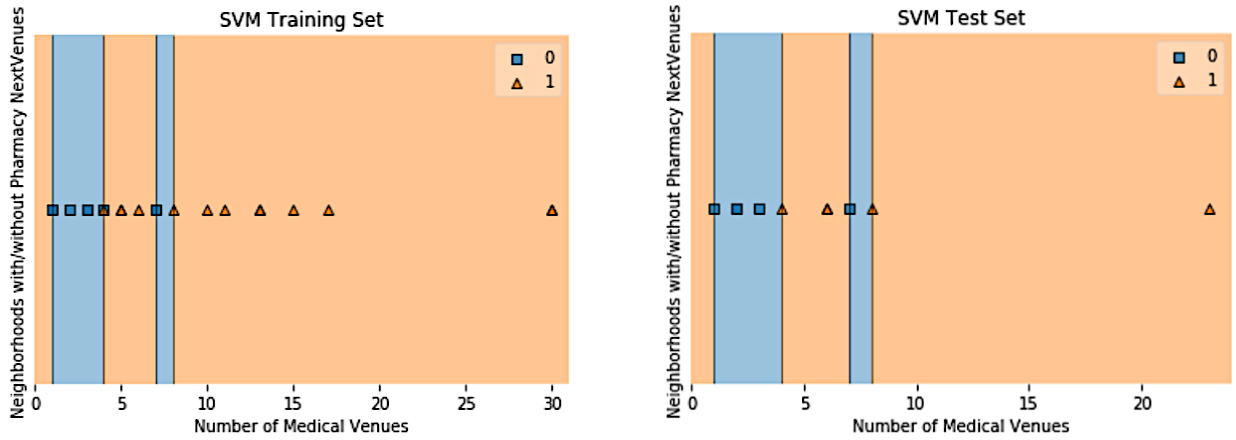
	MAE	MSE	R-Squared
Linear R	0.58	0.51	0.34

Table 4. Evaluation metrics of OLS model

While the continuous model above excludes the possibility of no pharmacy NextVenues, that is not always a likely scenario, as shown from the collected dataset in Plot 2. Predictive models that help assess the presence of pharmacy NextVenues from the quantity of hospital sites can be more helpful if the goal is to explore a future retail site within that neighborhood/area, because only a pair of pharmacies that are commonly visited together might be enough to warrant opening a nearby location.

Plots 5 and 6 below show training and tests sets for the SVM model. The region shading helps show that there should be at least four and to be more sure, closer to seven medical venues in a neighborhood to consider opening a competing retail location or complementary offering.

Table 3 shows the performance of the logistic regression and SVM models. The SVM model yielded a more accurate result upon testing, as evidenced by the higher Jaccard Similarity and F-1 Scores.



Plot 5 and 6. Training (left) and test (right) set results of Support Vector Machine Model. "0" denotes no pharmacy NextVenues, "1" denotes at least one pharmacy NextVenue

	Jaccard	F-1	Log Loss
Log R (liblinar)	0.36	0.19	0.74
SVM (RBF)	0.72	0.73	N/A

Table 4. Evaluation metrics of predictive models

5. Conclusions and Further Considerations

In this short exploratory study, we claim that pharmacy errand runs will be a stable consumer activity and retail opportunity exists while consumers are on their "drugstore errand run". We recognize that (1) not everyone visits drugstores for prescriptions, (2) it would be best to have data on venues visited before as well as after their drugstore, (3) grouping by neighborhood might not always be the best classifier for all urban areas, particularly those that are less walkable and spread apart.

We started by compiling the most popular pharmacies in each Manhattan neighborhood in order to find out the most popular spots visited after and see what opportunities based on the frequency and character of the NextVenue Categories via K-Means Clustering Analysis.

While we were not surprised by the appearance of supermarkets, grocery stores, coffee shops, and food establishments in the NextVenue Category results, we were intrigued by (a) parks and plazas as well as (b) other pharmacies themselves. As for (a), the neighborhoods in Cluster 2 with parks and plazas do not have large parks, so it seems that these consumers might be traversing the park or plaza while completing their regular household shopping. This insight can help various retail businesses and organizations with exploring future sites or for guerilla marketing/awareness campaigns.

As for (b), we first hypothesized that pharmacy NextVenues were related to chain and independent pharmacies, but it turns out that there was no apparent relationship except for one case (however, we note that there is a lack of reliable NextVenue data on independent pharmacies). However, upon mapping the pharmacy NextVenues, we realized that (1) they are unequally distributed throughout the city and concentrated in a specific number of neighborhoods, and (2) the pharmacy visited after the first is often also another top visited pharmacy in that neighborhood.

Operating on the intuition that this implies untapped demand, or at least, a steady route of foot traffic for drugstore products shopping, we attempt to identify similar features that characterize these neighborhood in order to understand (1) what is driving this phenomenon and (2) could we predict this phenomenon based on these features when there is a lack or unreliable user location data.

We observed that medical buildings such as hospitals and centrally planned residential complexes occur in almost all of the neighborhoods. While a more comprehensive study should include additional features, we selected to study the effect of hospital venues on the occurrence and frequency of pharmacy NextVenues.

The data shows there to be a modest positive relationship between medical venues and pharmacy NextVenues. As hospitals are a healthcare focused service, this relationship also helps additional research focus on the neighborhoods that aim to study consumer location data that are visiting drugstores primarily for prescription medications.

For businesses that seek retail locations where there is consistent prescription medication-driven foot traffic, we predictive SVM modeling shows that it takes at least four medical venues in a neighborhood to predict the positive likelihood of a pharmacy NextVenues, and the increase in certainty approaches 1 at when there are eight medical venues in that neighborhood. For followup research, it would be important to classify the kind of medical venues and weight their impact on foot traffic or similar desired metrics accordingly.