IBM Applied Data Science Capstone

Powered by Coursera

Project report: Preferred location map for the online pharmacy warehouses in London

By Anton Gil

May 2020

Table of content

Introduction

During the pandemic caused by COVID-19 outbreak the pharmacy retail business in most of the CIS countries got spotlighted with number of inefficiencies. With the requirement of self-isolation, population and businesses had to adjust to a new model of living and operating. While some day-to-day needs could have been done in isolation such as remote working, food delivery, fitness etc. the process of buying medicine still hasn't shifted online on an appropriate level. There are numbers of inconveniences related to it such as:

- No option of contactless ordering, instead, client is forced to go to a pharmaceutical shop in person and get in contact with potentially sick people;
- Increased danger of the pharmaceutical shops' staff through;
- No online stock monitoring causing clients to visit number of shops «blindly» in order to find the requirements;
- Ability to perform over-the-counter manipulations (no receipt or prescription sales) or overcharging for the highly demanded goods (masks, antiseptics, etc.).

A possible solution could be a set up of an online pharmacy network which would meet certain core criteria and features:

- < 2 hours on-demand delivery time though either own courier fleet or other transportation service providers;
- Unified meet of industry regulation i.e. direct connection to the authorized medical centers and doctors for verification of prescriptions through reliable technology;
- Development and executions of medicine delivery schedule based on the prescription;
- Track the type and volume of medicine purchase and prescriptions for better planning and analysis;
- Maintain the patient profile history.

Business problem

The goal of this capstone project is not to develop a business plan of an online pharmacy, but rather to try to solve one of the core problems of this process which is the pharmacy warehousing optimal location detection. Therefore, the problem I am trying to solve through the usage of data science methodology, machine learning algorithm and location data is to define the optimal location of the pharmacy warehouses.

The key assumption is that the optimal location for the warehouse should be in a place with a high density of population, hence a high demand for pharmacy products. With data available I will take the supply analysis approach and track the areas with the highest density of pharmacy stores, classifying those as the preferred locations for the warehouse location.

This methodology should be applicable to number of countries, for the reason of data availability I will use the London city as a subject of this capstone. By London I surely mean the one standing on the river Thames.

Target Audience of this project

Among of the stakeholders who might be interested in this project are:

- Healthcare regulators;
- Retail pharmacy market participants (owners, suppliers, producers);
- Hospitals and health insurance companies;

— City transportation companies;
— Customers.

Data

What data do I need to solve this problem? What are the sources?

To solve the stated problem, I will require the data illustrated in the following sum-up table along with the purpose, extraction method and the source:

Data	Purpose	Extraction method	Source
List of neighborhoods in London	To define the area for the analysis	Web scraping (Beautifulsoup)	Wikipedia.org
Latitude and longitude coordinates	City map plotting with pharmacy stores locations	Python Geocoder package	-
List of pharmacy stores with locations	To perform clustering on the neighborhoods	API Requests	Foursquare

So, how can I use data to answer this question?

In order to better define the area for this project I will split the territory of London into neighborhoods. The list of these neighborhoods can be found on Wikipedia page and can be gathered with the web scraping technique and using Python and BeautifulSoup package.

It will be required to understand the exact location of the neighborhoods and pharmacy stores. For that purpose, I will be using Geocoder package – which is a simple and consistent geocoding library written in Python, that can be a reasonable alternative to Google or Bing geocoding services.

Finally, in order to get the information and locations of the actual operating pharmacy stores I will be using the Foursquare API as a requisite source for this Capstone. Nonetheless, Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers and had a wide use in London, which meets the requirements of the stated tasks.

In the next section I will discuss the methodology used in order to execute this project.

Methodology

For this capstone I followed the data science methodology based on **CRISP-DM**. It consists of 6 main steps:

1. **Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**

Business Understanding This stage is the most important because this is where the intention of the project is outlined. My goal is to find the preferred location for the pharmacy warehouse. The key assumption is that the optimal location for the warehouse should be in a place with a high density of population, hence a high demand for pharmacy products.

Data Understanding & Data Preparation Data understanding relies on business understanding. With data available I will take the supply analysis approach and track the areas with the highest density of pharmacy stores, classifying those as the preferred locations for the warehouse location. I have used data from open sources such as Wikipedia, Geocoder library and Foursquare. Let's get into the process of data collection:

First step was to get the list of London neighborhoods, these are available on Wikipedia.org and ready to for use once you know how to use the web scrapping. In order to successfully do so I have preinstalled and used the BeautifulSoup library, parsed the data from the html into a BeautifulSoup object, created a storage list and appended the data into the list. As a result, I received a data frame with a list of neighborhoods of London (pic 1). It is crucial for the data preparation process to check if all the data is correct and in place.

	Neighborhood
0	Abbey Wood
1	Alperton
2	Anerley
3	Archway
4	Barnes
...	...

Picture 1 – Neighborhood data frame

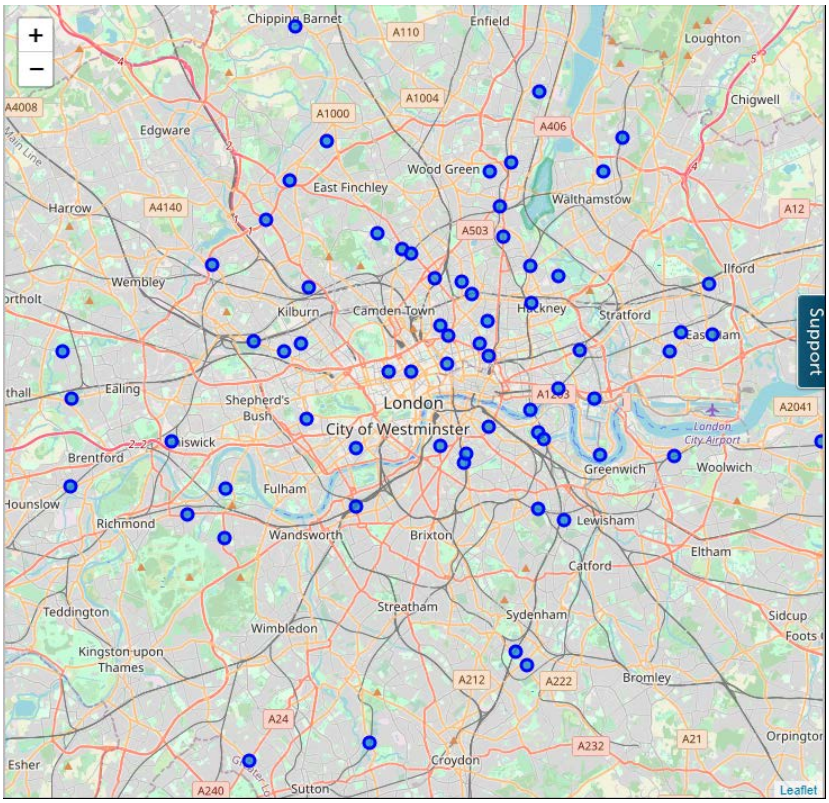
As the next step, I've collected the geographical coordinates of each neighborhood location which will be further required to use the Foursquare API. For the purpose geo data collection, a function is required which will allow to get the coordinates with the use of Geocoder on the specified address. Basically, the function used is a loop that repeats this request for each of the neighborhoods in the list. The received coordinates of latitude and longitude are merged with the list of neighborhoods (Pic 2).

	Neighborhood	Latitude	Longitude
0	Abbey Wood	51.492450	0.121270
1	Alperton	51.526871	-0.206440
2	Anerley	51.412330	-0.065390
3	Archway	51.565747	-0.134919
4	Barnes	51.474570	-0.242120
...	...	...	...
64	Walworth	51.487640	-0.095420
65	Wapping	51.504580	-0.055990
66	West Drayton	51.595020	-0.011722
67	Worcester Park	51.370997	-0.228087
68	Yiewsley	51.512630	-0.472590

69 rows × 3 columns

Picture 2 – Data frame with geodata

As a milestone and a proof check it's a reasonable action to visualize what is gathered so far and building a map is a great way to do so. To build a map the latitude and longitude values of neighborhood and London itself are required along with the folium package to bring it to live. The result is on the picture 3.



Picture 3 – Map of London neighborhoods

To follow up, I have used Foursquare API to explore the neighborhoods for the Pharmacy shops around. In order to do so it is required to have a developer account registered in order to receive personal credentials. Along with the credentials there are few arguments that must be specified prior to make an API request. The arguments used are:

- Limit – limit of number of venues returned by Foursquare API for each location;
- Radius – search radius in meters from the specified location;
- Query – specified query request, venue type in this case ("Pharmacy").

For this case it turned out to be the most challenging step of the entire process since there certain limitations of the free developer account and request format that must be considered. Foursquare returns the specified data in a json file format that should be transformed into a data frame. After few manipulations the example of the request result is illustrated on the picture 4.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	51.49245	0.12127	Superdrug	51.506883	0.106192	Pharmacy
1	Abbey Wood	51.49245	0.12127	Morrisons Pharmacy	51.507656	0.105978	Pharmacy
2	Abbey Wood	51.49245	0.12127	LloydsPharmacy	51.483370	0.147426	Pharmacy
3	Abbey Wood	51.49245	0.12127	Superdrug	51.462711	0.107610	Pharmacy
4	Abbey Wood	51.49245	0.12127	Superdrug	51.490997	0.067669	Pharmacy
...	...	...	...	...	...	...	...
3371	Yiewsley	51.51263	-0.47259	Boots	51.545489	-0.477378	Pharmacy
3372	Yiewsley	51.51263	-0.47259	Boots	51.470376	-0.458656	Pharmacy
3373	Yiewsley	51.51263	-0.47259	Adell Pharmacy	51.551479	-0.448780	Pharmacy
3374	Yiewsley	51.51263	-0.47259	Savers	51.546329	-0.480300	Pharmacy
3375	Yiewsley	51.51263	-0.47259	Boots	51.472923	-0.487690	Pharmacy

3292 rows × 7 columns

Picture 4 – Data frame of pharmacy shops around London neighborhoods

Now, as a final step of the data collection and preparation it is important to analyze the data we are about to work with on the Modeling step. In order to differentiate the data gathered and create a separate feature with the use of one hot encoding. After I group rows by neighborhood and by taking the mean of the frequency of occurrence of each category the data frame looks as follows:

	Neighborhood	Pharmacy
64	Walworth	0.959184
65	Wapping	0.956522
66	West Drayton	0.952381
67	Worcester Park	1.000000
68	Yiewsley	1.000000

Picture 5 – Data frame after one hot encoding

After this step the data set is ready for Modeling.

Modeling For the task stated I have used clustering using k-means clustering as it is reasonably suited for the matters of such sort. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. [1] I've separated the neighborhoods into three clusters based on the frequency of occurrence (density) of a Pharmacy shop in each neighborhood. The cluster with the highest density is the one to be considered for the warehouse to be placed.

Results evaluation

Based on the model built here are the sorted results of clustering:

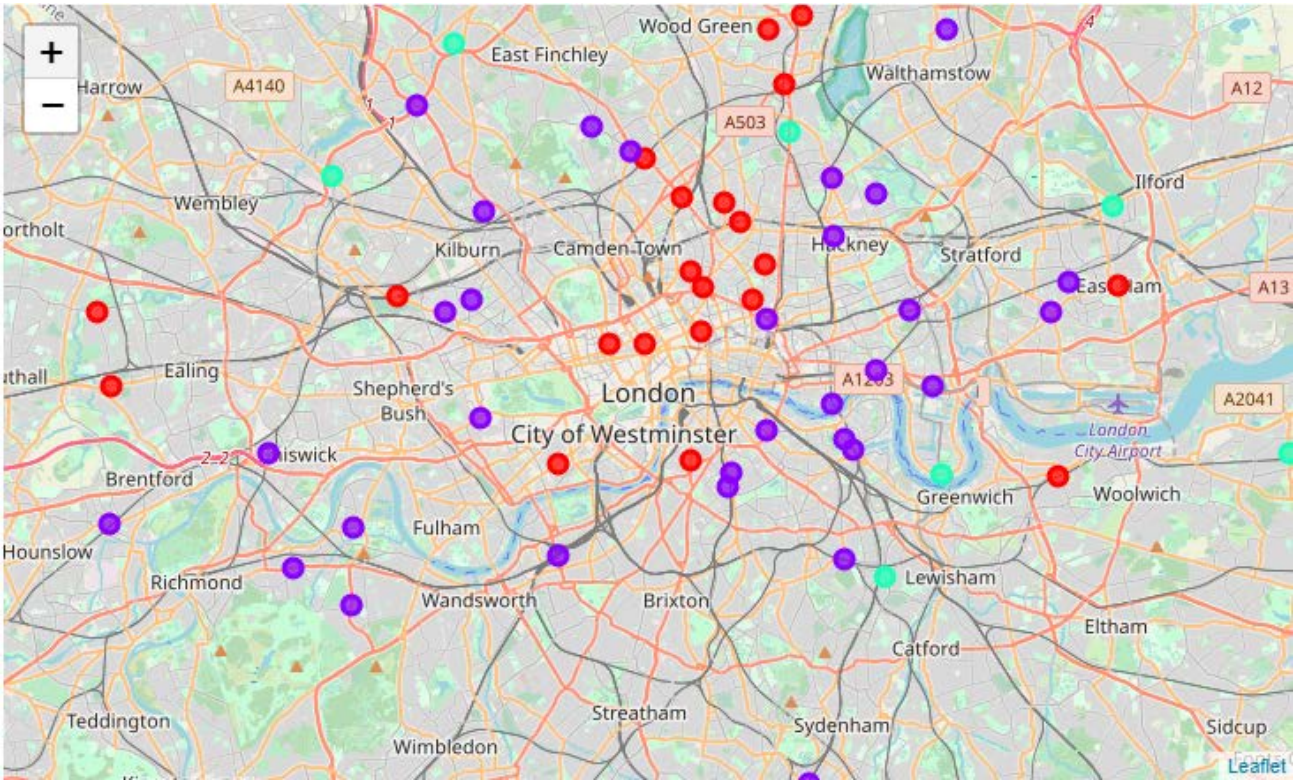
	Neighborhood	Pharmacy	Cluster Labels	Latitude	Longitude
68	Yiewsley	1.000000	0	51.512630	-0.472590
20	Elmers End	1.000000	0	51.407156	-0.058434
42	Kensal Green	0.981132	0	51.530540	-0.225480
22	Greenford	1.000000	0	51.526707	-0.342207
24	Hackbridge	1.000000	0	51.377690	-0.154170
...	...	...	...	...	...
12	Brockley	0.942857	2	51.462680	-0.035580
39	Isle of Dogs	0.928571	2	51.487210	-0.013810
47	Neasden	0.933333	2	51.559708	-0.250301
58	Stamford Hill	0.944444	2	51.570230	-0.072830
0	Abbey Wood	0.928571	2	51.492450	0.121270

69 rows × 5 columns

Picture 5 – Data frame after one hot encoding

[1] Wikipedia.org – https://en.wikipedia.org/wiki/K-medoids

Now, let's visualize the result of clustering on the pre-created London map.



Picture 6 – Map of London with clustered neighborhoods

Legend: Cluster 0 – red, Cluster 1 – purple, Cluster 2 – light blue

Cluster 0 – the cluster of neighborhoods marked red is the one we are interested in, since it has the highest density of Pharmacy shops and these locations should be considered for the warehouse placement.

Cluster 1 – the cluster of neighborhoods marked purple is the one that is a second tier with a moderate density of pharmacy shops.

Cluster 2 – the cluster of neighborhoods marked light blue is the one with the lowest density of pharmacy shops.

Here is the table of neighborhoods which got into cluster 0:

Elmers End	Harold Wood	Highams Park	Chipping Barnet
Islington	Highbury	Seven Sisters	Tottenham
Greenford	Worcester Park	Gidea Park	Kensal Green
Hackbridge	Holloway	Bedford Park	St Helier

Hanwell	Northolt	Upper Holloway	Canonbury
Hanworth	Ickenham	Barnsbury	Chelsea
Clerkenwell	Lambeth	Beckton	Charlton

Discussion

Based on the gathered results I do recommend to start the consideration of the online pharmacy warehouse deployment in the neighborhoods that got in cluster 0 as it is the one that has the highest density of current pharmacy shops. The high concentration of competitors surely might be a more complicated locating to start the online business, therefore if the business model analysis would suggest that the business should start from the places with the lowest competition possible the cluster 2 is the one to consider.

Limitations and Suggestions for the Future Research

I would like to note that the approach used is market-supply-driven. Prior to making a decision it is important to consider the demand side of the market. The population density analysis and population age groups are factors to consider in further model development. From the business side in-depth market analysis is recommended to define customer preferences and readiness to use online pharmacy services on day-to-day basis.

Deployment In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. This methodology should be tested on other locations.

Conclusion

The preferred location for an online pharmacy warehouse in London, UK has been defined and characterized by the Cluster 0. In order to do so the CRISP-DM methodology has been used along with open data sources such as Wikipedia, geocoder library and Foursquare, one-hot-encoding and k-means have been implemented to define the clusters. The findings of this work should assist stakeholders in business plan implementation.