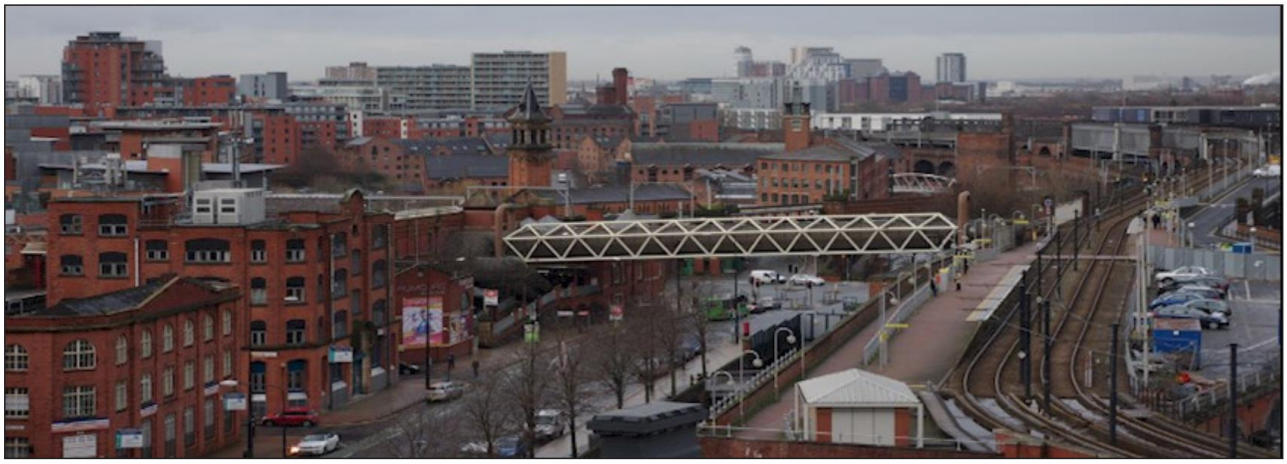
Coursera Capstone Project

IBM Applied Data Science Capstone

Location Analysis of Pharmacy Businesses in Manchester, UK.



Prepared by: Maxim Lukin

Date of the report: 12/8/19

Introduction

One of the quality life criteria in big cities is the accessible environment including the availability of such facilities as a pharmacy. A pharmacy near the house is in demand for social groups, such as pensioners, people with disabilities, young families with children. For my analyzes, the city of Manchester UK was chosen. Manchester is a major city and metropolitan borough in Greater Manchester, England, with a population of 534,982 as of 2018 (4th most populous English district). It lies within the United Kingdom's second-most populous urban area, with a population of 2.9 million, and third-most populous metropolitan area, with a population of 3.3 million.

Business Problem

The objective of the Capstone project is to analyse and select the best locations in the City of Manchester UK for opening new pharmacy business using data Science methodology and techniques such as clustering. this project aims to provide solutions to answer the question: in case new pharmacy business is looking to be open, which area will would be recommended for that purpose?

Data and Data Sources

In order to solve the problem in the following data will be used:

- list of areas of Manchester, UK
- Latitude and longitude coordinates of all areas
- Venues data from Foursquare API

Methodology

1. I got the list of areas of the city of Manchester, UK. The list is available in the Wikipedia page[1].
2. I did web scraping using Python requests and Beautiful Soup packages[2] to extract the data in the form of a list of areas.
3. I got the geographical coordinates in the form of latitude and longitude using Geocoder package[3].
4. After gathering the data, data will be transferred into a pandas DataFrame and then visualize the in a map using Folium package[4]. This helps to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted.
5. I used Foursquare API to get the top 100 venues that are within a radius of 2000 meters. All API calls have been passed to in the geographical coordinates of the area in a Python loop. Foursquare returned the venue data in JSON format and the venue name, venue category, venue latitude and longitude have been extracted. The number of venues has been checked and examined how many unique categories we have.

[1] https://en.wikipedia.org/wiki/Category:Areas_of_Manchester
[2] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[3] https://geocoder.readthedocs.io/
[4] https://python-visualization.github.io/folium/

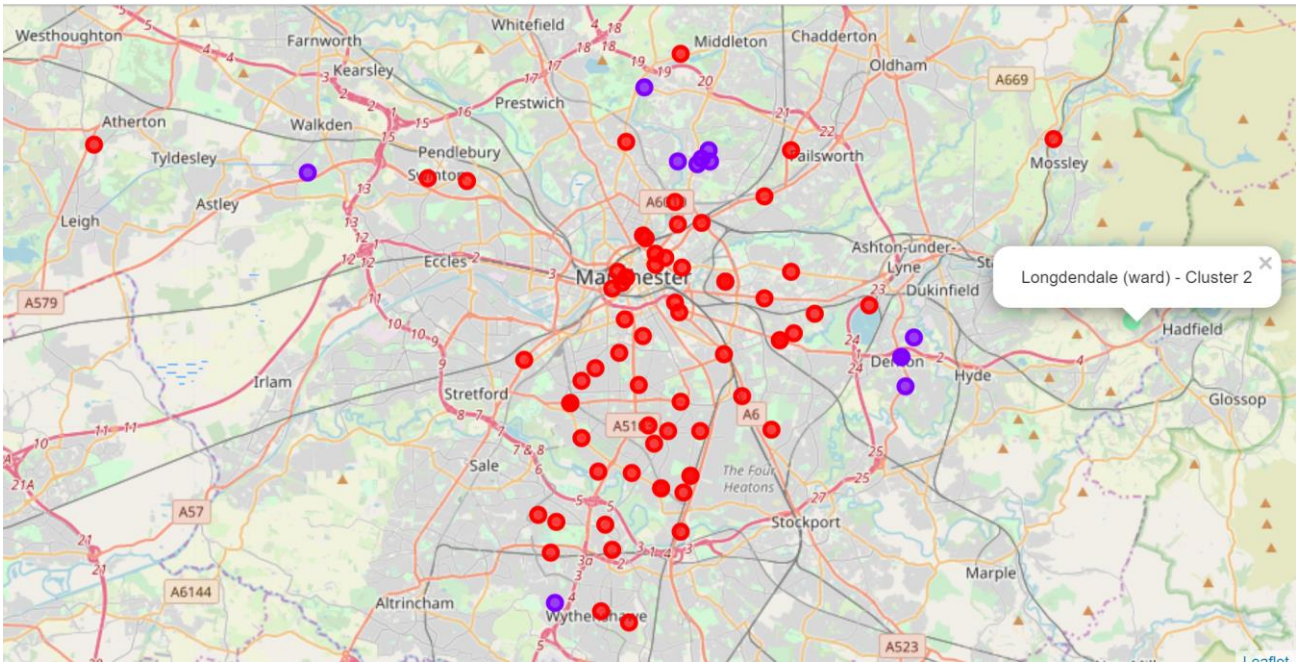
6. All data has been grouped by areas and the frequency of occurrence mean number has been calculated. The data has been prepared for further clustering.
7. Dataframe has been filtered by “Pharmacy” venue category as venue category for the research.
8. K-means clustering has been performed. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.
9. 3 clusters have been created based on their frequency of occurrence for “Pharmacy”. The results allowed to identify which areas have higher concentration of Pharmacy business while which areas have fewer number of Pharmacy businesses. Based on the occurrence of Pharmacy in different areas, it helps to answer the question as to which areas are most suitable to open new Pharmacy business.

Results

The results from the k-means clustering shows that there are 3 clusters based on the frequency of occurrence for “Pharmacy”:

- Cluster 0: Areas with low concentration of Pharmacy businesses (red)
- Cluster 1: Areas with high concentration of Pharmacy businesses (violet)
- Cluster 2: Areas with high concentration of Pharmacy businesses and far distance from Manchester city (chlorine)

The results of the clustering are visualized in the map below:



Discussion

According to obtained results there are 32 Pharmacies within Manchester’s dataframe. Based on the retrieved data and combined analysis, Cluster 0 contains the most relevant locations for new Pharmacy business.

Several weaknesses have been identified. First one lacks gathered information from several Manchester areas. Another one is related to obtained results itself: city center has relatively low concentration of pharmacies, which is contradict with modern urbanistic cities trends. I beleave more detailed analysis could be performed with consideration of population, distribution, city crime statistics by area, daily movements of city residents, the main centers of concentration of people and businesses etc.

Conclusion

In the project, we have gone through the process of searching for solutions for the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities and providing recommendations of the best locations where new pharmacy could be opened. To answer the business question that was raised in the introduction section, the areas in cluster 0 are the most accurate locations for new pharmacy businesses.