Coursera Capstone Project

IBM Applied Data Science Capstone

Part of IBM Data Science Processional Certificate

Property opportunities in Olsztyn, Poland

photo copyright: https://bit.ly/2Swn1vt

Problem Statement

I would like this project to help me in my personal dilemma. Therefore, I will neither be focusing on a business plan as such, nor will I try to sell this idea to a client (other than myself). Because of that I believe I should start with a short introduction. Hi, I'm Piotr. I live and work in UK, but I was born and raised in Poland. Olsztyn is a town located in North East Poland. Its surrounded by forest, 1000+ lakes (so they say) and It's a place where I was born couple of decades ago. My family is still here, and I do visit them as often as I can.

Some time ago I started to explore the idea of purchasing a house in Olsztyn. The problem is that the town changed so much since I left, that I wouldn't have the slightest idea which areas should I consider. Although my family is still here, they have no interest in property market and I would need to rely on my own research.

This is where this assignment comes into play: I plan to leverage Foursquare location data to help me make aforementioned decision and re-discover the area. There are multiple criteria that can be taken into consideration here: access to medical services, proximity to shopping centers, entertainment venues, roads, schools, lakes, forests, prices of land, building materials etc. In order to simplify this task, I will focus on the selected data available through Foursquare and will list it in the next paragraph. The goal is to find a location that's in a proximity to desired services and can potentially increase in value over time. Naturally, this analysis will be also beneficial for those who don't plan to settle in Olsztyn but instead are looking for investment opportunity.

Data

This paragraph will include a list of necessary data sources and the ways I plan to utilize them in order to solve my problem.

1. List of neighborhoods in Olsztyn; taken from Wikipedia.
2. Coordinates (Latitude and Longitude) of those neighborhoods that will be used to plot a map of the area; using geocoding library
3. List of POI's and venues, particularly related to (sourced using Foursquare API):
    - arts & entertainment
    - college & university
    - medical services

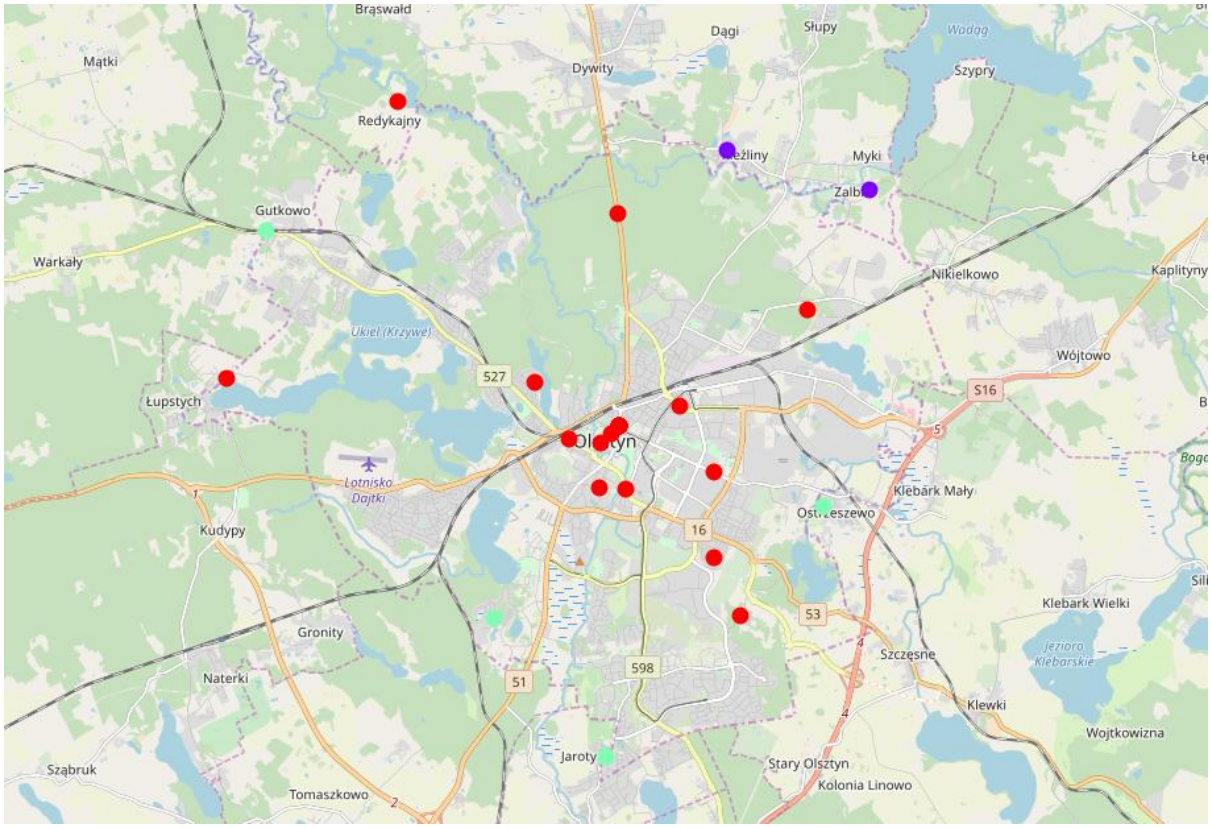
In order to get the list of neighborhoods in the area, I will use BeautifulSoup library. As you may know, it's a Python package that is used for extracting data from HTML documents by scraping them and creating a parse tree. After obtaining that list I will use geocoder library to allocate latitude and longitude coordinates accordingly to aforementioned areas and later visualize a map using folium. The last step would be to utilize Foursquare API and fetch data related to entertainment, schools and medical services (in order to simplify task at hand).

Methodology

Firstly, I need to get the list of areas in Olsztyn, Poland. Fortunately, the list is available in the Wikipedia page. I will do web scraping using beautifulsoup to extract the list of neighborhoods data - names. After that I need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. Here, Geocoder package will come in handy. After gathering the data, I will populate the data into a pandas DataFrame and then visualize the neighborhoods on a map using Folium package. This helps to perform a sanity check and visualize the data. Using Foursquare API I would get top 100 venues that are within a radius of 2000 meters. After registering with a developer account, we push Foursquare ID and Foursquare secret key when making and API call. Foursquare will return the venue data in JSON. I will extract the venue name, venue category, venue latitude and longitude. With the data, I can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, I will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, I am also preparing the data for use in clustering. I then will filter for data categories mentioned in the Data paragraph. Lastly, clustering will be performed on the data by using k-means. This 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. I will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for data categories mentioned in the data paragraph. This will then help me to answer the question as to which area is most suitable to buys a new house.

Results, conclusions & suggestions for the future

I much say, I was disappointed with the quantity of venues available for Olsztyn and the number of categories available. I should have checked it before starting this project, but it’s certainly a lesson I will remember. I had to choose at some point categories that were in a quantity significant enough to make a difference and relevant to the project. I selected categories with the highest proxy to sports. It’s seems that Clusters 1 and 2 contain neighborhoods with low number of sport activities. In contrast, most sport activities are located near those neighborhoods assigned to Cluster 0. This would indicate where one would have to look to buy a property. The results are shown on the map below:



References

Wikipedia, Category: Neighborhoods in Olsztyn:

https://pl.wikipedia.org/wiki/Kategoria:Dzielnice_i_osiedla_Olsztyna

Global Property Guide, Buying costs are low in Poland, May 2019:

https://www.globalpropertyguide.com/Europe/Poland/Buying-Guide

Business Insider, Most important things (…), 2015:

https://www.businessinsider.com/most-important-things-britons-want-when-buying-a-house-2015-6?r=US&IR=T

Narodowy Bank Polski (National Polish Bank), Property Market Report, 2016:

https://www.nbp.pl/publikacje/rynek_nieruchomosci/raport_2016.pdf