Find the best district to open a Mall in Warsaw

Introduction

A shopping mall is a place where most people visit to buy items, clothes, groceries, etc. almost everything that a person needs. And because a mall has almost everything some families go there to spend some time together. And some mall actually has become a place to meet people and talk about their experiences.

That's why malls are highly valuated in the city planning and having a great place to build one is hard, because the property developers needs to look to every aspect related to location. One of this aspect its the competition, because you need to find a place that there's less or no other malls in the area.

Warsaw is one of the fastest growing cities in Central Europe and have a lot of building constructions in every part of the city, that's why I'm choosing this city for this project.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Warsaw, Poland to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Warsaw, Poland, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

Data

• List of districts in Warsaw. To define the scope for this project which is confined to the city of Warsaw.

• Latitude and longitude coordinates of those Districts. This is required in order to plot the map and also to get the venue data.

• Data of the venues in the city, this data should correspond to shopping malls. This data is used to perform clustering on the districts.

Sources of data

To get the districts of Warsaw i'll use this This Wikipedia page (https://en.wikipedia.org/wiki/Districts_of_Warsaw) which contains all the districts of the city.

Use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages.

Then get the geographical coordinates of the districts using Python Geocoder package which will give the latitude and longitude coordinates of the districts.
Later with the help of Foursquare API i'll get the venue data for those districts, and then filter to only use the Shopping Mall data.

For this is project will use many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

The first thing to do is get the list of districts of the city of Warsaw. As I mentioned before the list of districts is available in the Wikipedia page.

I'll do some web scrapping using python request and Beautifulsoup packages to extract the list of districts. But this list is not enough, the important information is get the Coordinates in the form of latitude and longitude in order to be able to use the Foursquare API.

To get that coordinates I'll use the Geocoder package to later match that coordinates with the correct district in a pandas dataframe. Just to confirm that the coordinates retrieved are correct I'll plot that coordinates on a map.

After I'll use the Foursquare API to get the top venues that are within a radius of 2000 meters, using the Foursquare Developer Account with the Foursquare ID and Foursquare secret key, I make API calls to Foursquare passing the geographical coordinates of the districts in a python loop, Foursquare will return the venue data in Json format and I'll extract the venue name, venue category, venue latitude and longitude. With that data I can check how many venues were returned for each district and the quantity of categories. With that data I'll group the rows by district and take the mean of the frequency of occurrence of each venue. And just after I'll filter the Shopping malls as venue category.

The last part consists in performing a clustering by k-means. That algorithm identifies the k number of centroids and the allocates every data point to the nearest cluster. The districts will be clustered in 3 groups based on the frequency of occurrence for Shopping Mall. That will allow to identify which districts have a higher concentration of Malls and who districts have lower.

With that information it could be possible to answer the question of what districts are most suitable to open new shopping malls.

Results

The results show that the districts are clustered in 3 groups.

Cluster 0.

Districts with low number or don't have malls

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
8	Targówek	0.0	0	52.23560	21.01037
14	Wola	0.0	0	52.27697	20.94778
13	Wilanów	0.0	0	52.26579	21.16994
12	Wesoła	0.0	0	52.23560	21.01037
11	Wawer	0.0	0	52.25680	21.02976
10	Ursynów	0.0	0	52.23903	20.97123
7	Rembertów	0.0	0	52.22292	21.23074
17	Żoliborz	0.0	0	52.19141	20.95212
5	Praga Południe	0.0	0	52.15418	21.03786
2	Bielany	0.0	0	52.27726	21.06594
1	Białołęka	0.0	0	52.21314	20.97069

Cluster 1.

Districts with the more concentration of malls

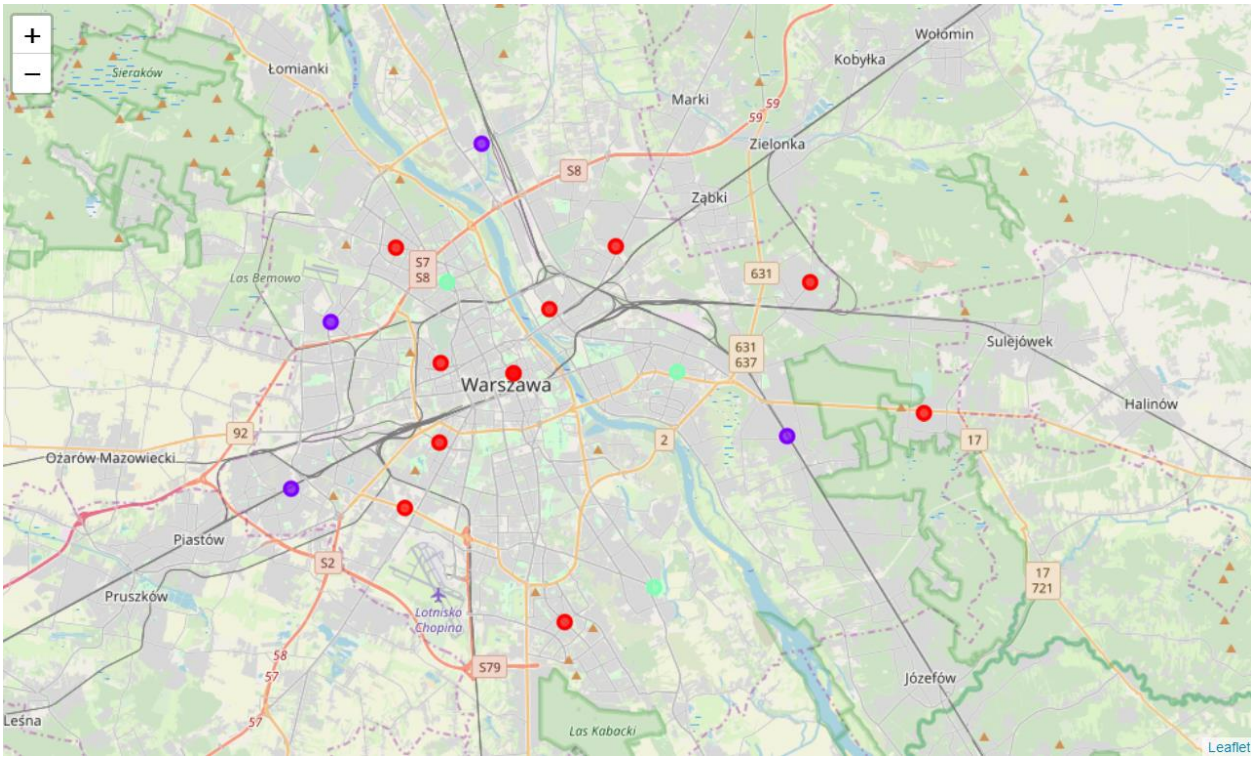
	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
16	Śródmieście	0.033898	1	52.25269	20.91244
4	Ochota	0.032258	1	52.21505	21.15758
6	Praga Północ	0.037736	1	52.19786	20.89114
0	Bemowo	0.043478	1	52.31097	20.99324

Cluster 2.

Districts with a moderate number of malls

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
9	Ursus	0.010000	2	52.26572	20.97483
3	Mokotów	0.010417	2	52.23633	21.09840
15	Włochy	0.012048	2	52.16566	21.08649

All of this result can be seen in the map below.



Discussion

Everything that I showed was using tools and knowledge of the Coursera Data Science Course and the results of this project are a great example of a real-world application. Some more experienced people maybe can use better or more advanced web scrapping techniques or other APIs instead of Foursquare which could lead in a different or more precise result.

Conclusion

With the results obtained we can see that most of the malls area located in cluster 1, while cluster 2 has less malls and cluster 0 has no shopping malls. This means that the districts located in cluster 1 have more competition than in other clusters.

This project recommends property developers to capitalize on these findings to open new shopping malls in districts in cluster 0 with little to no competition and property developers with unique selling propositions can also open new shopping malls in districts in cluster 2 with moderate competition.