IBM Data Science Capstone Project

Subject: Best districts to open restaurant in Warsaw

28th September 2019

Introduction/Business Problem

The objective of this problem is to analyse restaurant locations in districts of Warsaw (the capital city of Poland). This analysis will allow to select the best places to open a new restaurant. The main audience of this project are entrepreneurs willing to launch a new restaurant or restaurants chains. Using insights from this project, stakeholders will be able to make better and more effective business decisions. Especially, the best districts The project will use analytical approach, especially clustering algorithms will be used.

Data

The following data sets will be used in project:

- List of districts in Warsaw (web scraping from Wikipedia: https://en.wikipedia.org/wiki/Districts_of_Warsaw)
- Geographical coordinates of districs (using geopy package)
- Foursquare data (extracted by Foursquare API using geographical coordinates, data will include information about restaurants in 3 km radius from the center of each district), for each district the number of restaurant of each type will be counted

Methodology

First step of the project involved downloading data from Wikipedia with information about Warsaw districts. Data was downloaded from the following page:

https://en.wikipedia.org/wiki/Districts_of_Warsaw

Python libraries ‘requests’ and ‘’BeautifulSoup’ were used to perform web scraping. Table with specific HTML tags was extracted and converted into pandas dataframe:

	District	Population	Latitude	Longitude
0	Mokotów	220,682	0	0
1	Praga Południe	178,665	0	0
2	Ursynów	145,938	0	0
3	Wola	137,519	0	0
4	Bielany	132,683	0	0
5	Targówek	123,278	0	0
6	Śródmieście	122,646	0	0
7	Bemowo	115,873	0	0
8	Białołęka	96,588	0	0
9	Ochota	84,990	0	0
10	Wawer	69,896	0	0
11	Praga Północ	69,510	0	0
12	Ursus	53,755	0	0
13	Żoliborz	48,342	0	0
14	Włochy	38,075	0	0
15	Wilanów	23,960	0	0
16	Rembertów	23,280	0	0
17	Wesoła	22,811	0	0

The second step of analysis required information about geographical coordinates of each district. The data came from python geopy package, which

found the central point in each district. Below, there is a map of Warsaw with marks in the center of districts:

Next stage of project involved Foursquare API data. For each district, data about restaurant venus in each district was collected. API calls were made by passing information about geographical coordinates from the step above. After that, data was converted into pandas dataframe for json format. API returned 1120 rows of venues data about Warsaw restaurants. Below there is a snippet from the dataset:

	District	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mokotów	52.193987	21.045781	Stary Dom	52.195544	21.024004	Polish Restaurant
1	Mokotów	52.193987	21.045781	NABO Cafe	52.189653	21.068752	Scandinavian Restaurant
2	Mokotów	52.193987	21.045781	MEZZE hummus & falafel	52.203548	21.022705	Falafel Restaurant
3	Mokotów	52.193987	21.045781	Gringo Bar Burritos Tacos & More	52.201305	21.020496	Burrito Place
4	Mokotów	52.193987	21.045781	Targ Śniadaniowy Mokotów	52.189239	21.022857	Breakfast Spot
5	Mokotów	52.193987	21.045781	Boston Port	52.197249	21.024606	Seafood Restaurant
6	Mokotów	52.193987	21.045781	Pekin Express - duck & more	52.199367	21.023589	Asian Restaurant
7	Mokotów	52.193987	21.045781	Burger Bar	52.199293	21.023540	Burger Joint
8	Mokotów	52.193987	21.045781	Restauracja Polska Różana	52.208468	21.023531	Polish Restaurant
9	Mokotów	52.193987	21.045781	Ciao a Tutti Due	52.201612	21.016697	Pizza Place

Next steps involved creating dummy variables for each row to describe each restaurant type. There were 79 unique categories of restaurants in Warsaw. After that, the frequency of each type of restaurant was calculated for each district:

	District	African Restaurant	American Restaurant	Argentinian Restaurant	Asian Restaurant	Bakery	Bistro	Breakfast Spot	Buffet	Bulgarian Restaurant	Burger Joint	Burrito Place	Cafeteria
0	Bemowo	0.000000	0.000000	0.00	0.032258	0.032258	0.016129	0.032258	0.032258	0.00	0.016129	0.00	0.000000
1	Białołęka	0.000000	0.000000	0.00	0.166667	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000
2	Bielany	0.000000	0.000000	0.00	0.093750	0.031250	0.031250	0.000000	0.000000	0.00	0.000000	0.00	0.000000
3	Mokotów	0.000000	0.000000	0.00	0.030000	0.010000	0.020000	0.030000	0.000000	0.00	0.080000	0.01	0.000000
4	Ochota	0.000000	0.050000	0.01	0.010000	0.020000	0.040000	0.010000	0.000000	0.00	0.000000	0.00	0.000000
5	Praga Południe	0.000000	0.000000	0.00	0.010000	0.010000	0.040000	0.010000	0.000000	0.01	0.010000	0.00	0.000000
6	Praga Północ	0.000000	0.000000	0.00	0.020000	0.020000	0.030000	0.020000	0.000000	0.00	0.040000	0.00	0.000000
7	Rembertów	0.000000	0.000000	0.00	0.000000	0.250000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000
8	Targówek	0.000000	0.000000	0.00	0.030769	0.000000	0.061538	0.000000	0.015385	0.00	0.046154	0.00	0.000000
9	Ursus	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000
10	Ursynów	0.000000	0.010638	0.00	0.021277	0.010638	0.000000	0.010638	0.000000	0.00	0.063830	0.00	0.000000
11	Wawer	0.022727	0.045455	0.00	0.045455	0.045455	0.068182	0.000000	0.000000	0.00	0.022727	0.00	0.000000
12	Wesoła	0.000000	0.000000	0.00	0.000000	0.166667	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.166667
13	Wilanów	0.000000	0.000000	0.00	0.034483	0.000000	0.000000	0.000000	0.000000	0.00	0.068966	0.00	0.000000
14	Wola	0.000000	0.010000	0.00	0.020000	0.030000	0.030000	0.000000	0.000000	0.00	0.020000	0.00	0.000000
15	Włochy	0.000000	0.013889	0.00	0.041667	0.013889	0.055556	0.000000	0.000000	0.00	0.027778	0.00	0.027778
16	Śródmieście	0.000000	0.000000	0.00	0.020000	0.020000	0.040000	0.020000	0.000000	0.00	0.020000	0.00	0.000000
17	Żoliborz	0.000000	0.000000	0.00	0.022989	0.057471	0.034483	0.022989	0.011494	0.00	0.057471	0.00	0.000000

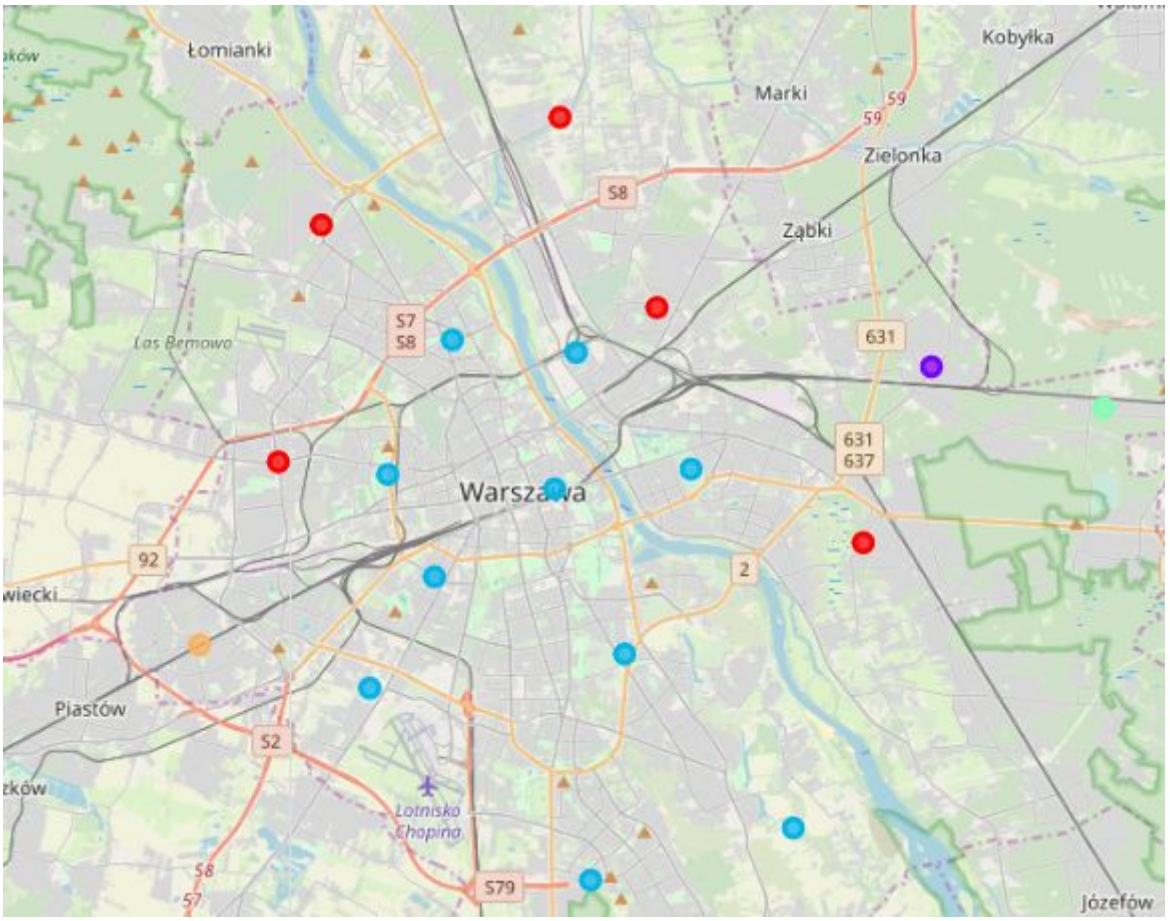
Afterwards, the ten most common restaurant types for each district were calculated:

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bemowo	Café	Pizza Place	Italian Restaurant	Fast Food Restaurant	Indian Restaurant	Sushi Restaurant	Food Court	Restaurant	Asian Restaurant	Bakery
1	Białołęka	Fast Food Restaurant	Asian Restaurant	Café	Pizza Place	Diner	Restaurant	Mediterranean Restaurant	Vietnamese Restaurant	Italian Restaurant	Deli / Bodega
2	Bielany	Café	Pizza Place	Asian Restaurant	Indian Restaurant	Polish Restaurant	Fast Food Restaurant	Italian Restaurant	Deli / Bodega	Food Court	Bakery
3	Mokotów	Café	Italian Restaurant	Burger Joint	Sushi Restaurant	Pizza Place	Restaurant	Polish Restaurant	Diner	Asian Restaurant	Breakfast Spot
4	Ochota	Café	Italian Restaurant	Sushi Restaurant	Pizza Place	American Restaurant	Vietnamese Restaurant	Korean Restaurant	Bistro	Indian Restaurant	Eastern European Restaurant

Results

The final step of project involved application of clustering algorithm. Districts were divided into five clusters based on the frequency of restaurants in each one of them. Because of the low number of rows, there were a few clusters with limited numbers of districts in them.

Later, clusters label information has been applied into map of Warsaw, which can be find below:



Discussion

Most of the restaurants in Warsaw are concentrated in blue cluster that consists of the most populous districts in the city. It may be seen that, the peripheral districts are outlier in terms of restaurants concentration. Beneath we can see data from the largest cluster:

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Café	Italian Restaurant	Burger Joint	Sushi Restaurant	Pizza Place	Restaurant	Polish Restaurant	Diner	Asian Restaurant	Breakfast Spot
2	Café	Italian Restaurant	Restaurant	Pizza Place	Polish Restaurant	Bistro	Mexican Restaurant	Sandwich Place	Sushi Restaurant	Indian Restaurant
2	Café	Pizza Place	Italian Restaurant	Chinese Restaurant	Sushi Restaurant	Burger Joint	Indian Restaurant	Eastern European Restaurant	Restaurant	Fast Food Restaurant
2	Café	Italian Restaurant	Pizza Place	Sushi Restaurant	Chinese Restaurant	Korean Restaurant	Bakery	Bistro	Indian Restaurant	Eastern European Restaurant
2	Café	Italian Restaurant	Vegetarian / Vegan Restaurant	Sushi Restaurant	Polish Restaurant	Pizza Place	Bistro	Mexican Restaurant	Modern European Restaurant	Restaurant
2	Café	Italian Restaurant	Sushi Restaurant	Pizza Place	American Restaurant	Vietnamese Restaurant	Korean Restaurant	Bistro	Indian Restaurant	Eastern European Restaurant
2	Café	Polish Restaurant	Italian Restaurant	Pizza Place	Sushi Restaurant	Restaurant	Thai Restaurant	Burger Joint	Vegetarian / Vegan Restaurant	Bistro
2	Café	Italian Restaurant	Burger Joint	Bakery	Polish Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant	Bistro	Sushi Restaurant	Diner
2	Italian Restaurant	Fast Food Restaurant	Café	Restaurant	Sushi Restaurant	Bistro	Pizza Place	Turkish Restaurant	Asian Restaurant	Chinese Restaurant
2	Italian Restaurant	Burger Joint	Restaurant	Eastern European Restaurant	Café	Japanese Restaurant	Mediterranean Restaurant	Diner	Doner Restaurant	Pizza Place

The most popular venues are cafes. It is recommended not to open the most popular restaurant category in each districts because there are too saturated. In each district, the best way is to open restaurant from the second half of the most popular venues,

because those restaurants are popular in each district but the market for them is not saturated.

Conclusion

In this report, analysis of the Warsaw districts in terms of their restaurants frequency were carried out. Districts were divided into 5 clusters, which can help in search of the best location to open a new restaurant in Warsaw. Analysis gives the overview of the restaurants saturation in Warsaw and more detailed steps can be applied to look for the best place to open a new restaurant, for example by using survey data.