CAPSTONE PROJECT

THE BATTLE OF NEIGHBORHOODS

WARSAW

TABLE OF CONTENTS

INTRODUCTION: BACKGROUND

Prices of flats in Poland go up faster than inflation, according to the report of money.pl website. Rising apartment prices on the market effectively obscure another problem - the increase in rental prices. This trend affects students, young workers without their own flats or economic immigrants. According to the analysis of experts at Rynekpotny.pl, the increases reached even 23%. Despite everything, life in Warsaw tempts many young people.

INTRODUCTION: BUSINESS PROBLEM

The capital is mainly attracting to itself those who are focused on making dizzying careers or artists and creative people. The heart of the city is the City Centre, which is vibrant with life at any time of day or night. It is one of eighteen districts, but each of them has different advantages. In this scenario, machine learning tools should be used to assist people coming to Warsaw to make wise and effective decisions. As a result, the business problem is:

- **How can we help people moving to the capital to choose the right location to rent an flat in Warsaw?**

In order to solve this business problem, we intend to merge Warsaw districts into a cluster in order to recommend facilities. We will recommend facilities according to the amenities and necessary equipment of the surrounding facilities such as: **Café**, **Bus Station**, **Pizza Place**.

DATA

To consider the objective stated above, we can list the below data sources used for the analysis.

- **Districts of Warsaw** Wikipedia page was scraped to pull out the necessary information;
- **Coordinate data** for each Districts of Warsaw obtained through Nominatim search engine for OpenStreetMap data;

In order to investigate and target recommended locations in different locations depending on the presence of facilities and necessary objects, we will access the data through the **FourSquare API** and arrange it as a data frame for visualization. By combining data about districts in Warsaw and data about amenities and essential facilities surrounding such properties from the FourSquare API, we will be able to recommend an appropriate location.

METHODOLOGY

The Methodology section will describe the main elements of the analysis and prediction system. The methodological part consists of four stages:

1. Data Preparation
2. Visualization and Data Exploration
3. Data preparation and Preprocessing
4. Modeling

DATA PREPARATION

Scrape the Wikipedia page and gathering data into a Pandas dataframe

To start with our analysis, we used the BeautifulSoup package to transform the data in the table on the Wikipedia page into the below pandas dataframe. Subsequently, we transform the data into a pandas dataframe.

	District	Neighborhood
0	Bemowo	Bemowo Lotnisko
1	Bemowo	Boernerowo
2	Bemowo	Chrzanów
3	Bemowo	Fort Bema
4	Bemowo	Fort Radiowo

DATA PREPARATION

	District	Neighborhood	Latitude	Longitude
0	Bemowo	Bemowo Lotnisko	52.261261	20.910737
1	Bemowo	Boernerowo	52.262390	20.901451
2	Bemowo	Chrzanów	52.216759	20.882969
3	Bemowo	Fort Bema	52.256562	20.938620
4	Bemowo	Fort Radiowo	52.257211	20.891900

Use geopy library to get the latitude and longitude values of Warsaw Localities

After we have built a dataframe of Warsaw localities along with the district name and neighborhood name, in order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighborhood. It possible to export data to a csv file for easier loading later.

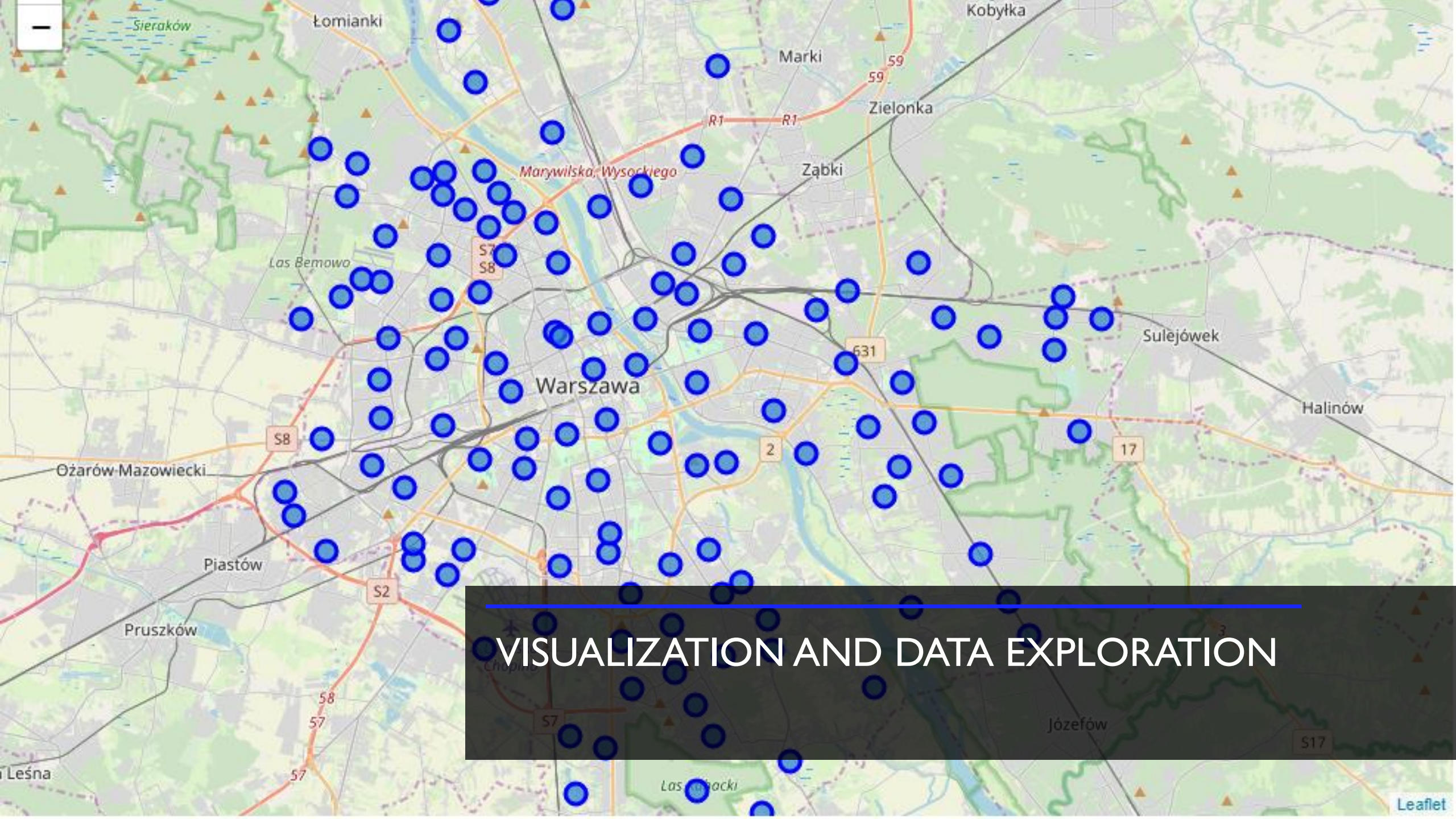
DATA PREPARATION

Utilizing Foursquare API to explore the neighborhoods

Foursquare is the most trusted, independent location data platform for understanding how people move through the real world. We have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots for each neighborhoods of Warsaw. The recommended location needs to have many eating and shopping venues nearby. Convenient public transport is also required.

```
Total 1427 of venues are found
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bemowo Lotnisko	52.261261	20.910737	Goldwings	52.260579	20.910778	Flight School
1	Bemowo Lotnisko	52.261261	20.910737	Hostel Kingroom	52.264355	20.912432	Hostel
2	Bemowo Lotnisko	52.261261	20.910737	Dach Nacipanej Vistuli	52.259773	20.915648	Airport Service
3	Bemowo Lotnisko	52.261261	20.910737	Garaże	52.258142	20.914198	Beer Garden
4	Bemowo Lotnisko	52.261261	20.910737	Place4Us	52.263905	20.915367	Hotel



VISUALIZATION AND DATA EXPLORATION

VISUALIZATION AND DATA EXPLORATION

- There are 234 uniques categories.

- Examples of Neighborhood meeting the Venue Category: **Café**

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
16	Fort Bema	52.256562	20.938620	Cafe Jurta Forty Bema	52.257985	20.935512	Café
25	Górce	52.245431	20.913714	CieKawa	52.242059	20.913374	Café
63	Kobiałka	52.354573	21.043018	Cafe Karolinka	52.355282	21.038461	Café
80	Tarchomin	52.318028	20.954304	Carmelia	52.321284	20.955756	Café
138	Słodowiec	52.276825	20.960235	COSTA Stare Bielany	52.275127	20.961906	Café

DATA PREPARATION AND PREPROCESSING

- In order to limit the number of neighborhood we limit categories to places that meet business requirements.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue
8	Grochów	Café	Restaurant
25	Natolin	Restaurant	Park
31	Powiśle	Café	Restaurant
42	Stara Praga	Restaurant	Park
46	Stary Mokotów	Café	Restaurant
47	Stary Żoliborz	Café	Restaurant
62	Śródmieście Południowe	Café	Restaurant
63	Śródmieście Północne	Café	Restaurant

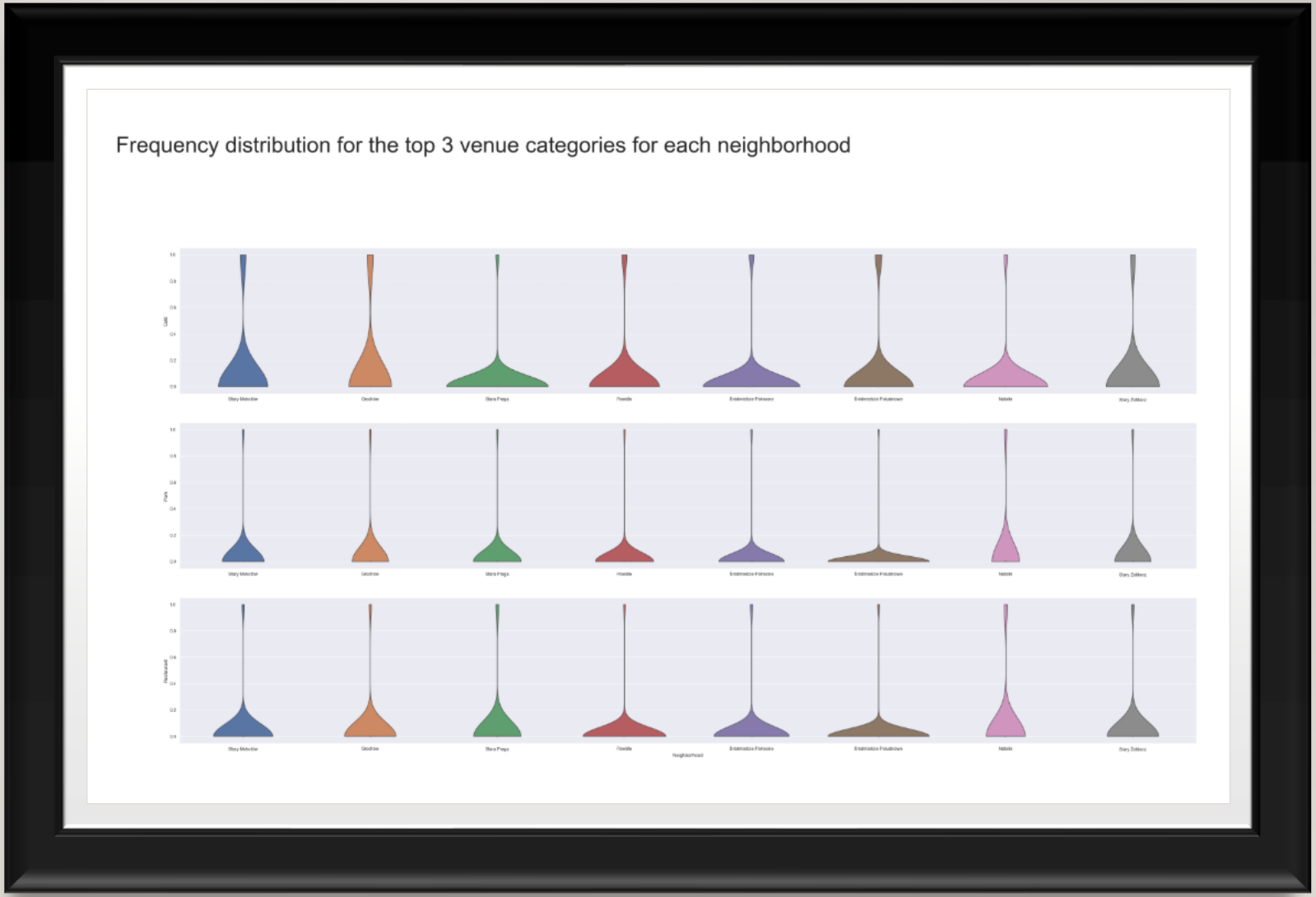
DATA PREPARATION AND PREPROCESSING

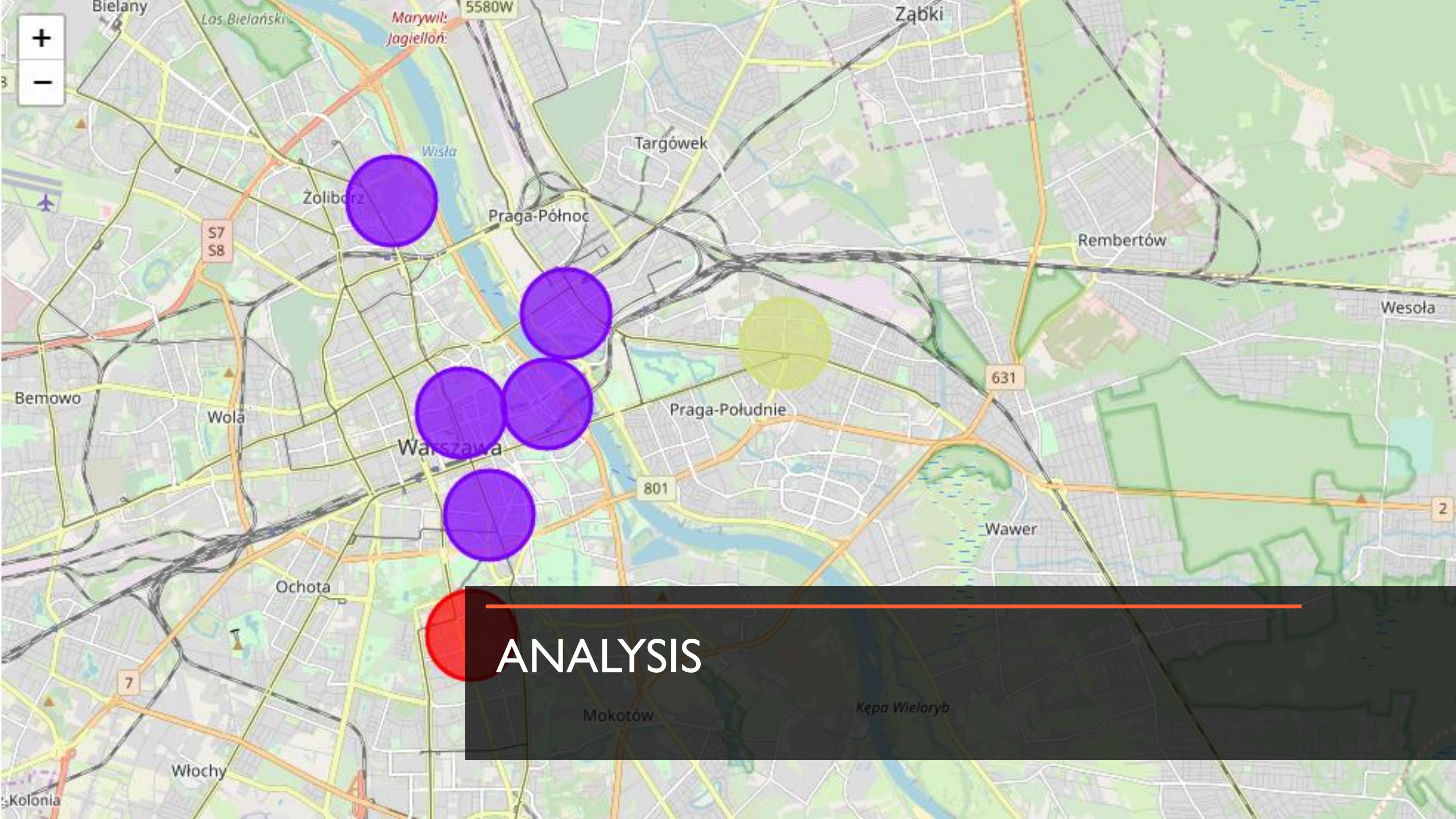
	District	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
43	Mokotów	Stary Mokotów	52.205272	21.011551	0	Bakery	Café	Ice Cream Shop	Italian Restaurant	Convenience Store	Coffee Shop	Dessert Shop	Pizza Place
53	Praga-Południe	Grochów	52.246707	21.084637	3	Café	Dessert Shop	Bus Station	Supermarket	Pizza Place	Fast Food Restaurant	Flea Market	Restaurant
59	Praga-Północ	Stara Praga	52.250981	21.033605	1	Diner	Restaurant	Hotel	Coffee Shop	Middle Eastern Restaurant	Road	Public Art	Plaza
66	Śródmieście	Powiśle	52.238055	21.029351	1	Pizza Place	Café	Eastern European Restaurant	Asian Restaurant	Pub	Polish Restaurant	Bar	Italian Restaurant
69	Śródmieście	Śródmieście Północne	52.236806	21.009433	1	Nightclub	Coffee Shop	Cocktail Bar	Café	Hotel	Italian Restaurant	Restaurant	Beer Bar
70	Śródmieście	Śródmieście Południowe	52.222253	21.015700	1	Café	Vegetarian / Vegan Restaurant	Coffee Shop	Cocktail Bar	Italian Restaurant	Sushi Restaurant	Hostel	Bistro
89	Ursynów	Natolin	52.141101	21.056435	2	Sushi Restaurant	Restaurant	Park	Coffee Shop	Indian Restaurant	Italian Restaurant	Sandwich Place	Café
141	Żoliborz	Stary Żoliborz	52.266810	20.992990	1	Café	Thai Restaurant	Polish Restaurant	Coffee Shop	Plaza	Burger Joint	Restaurant	Public Art

MODELING

```
kclusters = 4


warsaw_grouped_clustering = warsaw_grouped.drop('Neighborhood', 1)


kmeans = KMeans(n_clusters=kclusters,
random_state=0).fit(warsaw_grouped_clustering)

kmeans.labels_[0:10]
```

ANALYSIS


Frequency distribution for the top 3 venue categories for each neighborhood


Bielany
Las Bielański
Marywil
Jagielloń
5580W
Ząbki
Targówek
Wisła
Żoliborz
Praga-Północ
S7
S8
Rembertów
Wesoła
631
Bemowo
Wola
Praga-Południe
Warszawa
801
2
Wawer
Ochota
7
Mokotów
Kępa Wieloryb
Włochy
Kolonia

ANALYSIS

Examine Cluster 0

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 0, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	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
43	Stary Mokotów	Bakery	Café	Ice Cream Shop	Italian Restaurant	Convenience Store	Coffee Shop	Dessert Shop	Pizza Place	Movie Theater	Eastern European Restaurant

Examine Cluster 1

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 1, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	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
59	Stara Praga	Diner	Restaurant	Hotel	Coffee Shop	Middle Eastern Restaurant	Road	Public Art	Plaza	Park	Movie Theater
66	Powiśle	Pizza Place	Café	Eastern European Restaurant	Asian Restaurant	Pub	Polish Restaurant	Bar	Italian Restaurant	Science Museum	Restaurant
69	Śródmieście Północne	Nightclub	Coffee Shop	Cocktail Bar	Café	Hotel	Italian Restaurant	Restaurant	Beer Bar	Polish Restaurant	Greek Restaurant
70	Śródmieście Południowe	Café	Vegetarian / Vegan Restaurant	Coffee Shop	Cocktail Bar	Italian Restaurant	Sushi Restaurant	Hostel	Bistro	Plaza	Hotel
141	Stary Żoliborz	Café	Thai Restaurant	Polish Restaurant	Coffee Shop	Plaza	Burger Joint	Restaurant	Public Art	Playground	Breakfast Spot

ANALYSIS

Examine Cluster 2

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 2, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	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
89	Natolin	Sushi Restaurant	Restaurant	Park	Coffee Shop	Indian Restaurant	Italian Restaurant	Sandwich Place	Café	Convenience Store	General Entertainment

Examine Cluster 3

```
warsaw_merged.loc[warsaw_merged['Cluster Labels'] == 3, warsaw_merged.columns[[1] + list(range(5, warsaw_merged.shape[1]))]]
```

	Neighborhood	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
53	Grochów	Café	Dessert Shop	Bus Station	Supermarket	Pizza Place	Fast Food Restaurant	Flea Market	Restaurant	Coffee Shop	Mexican Restaurant

ANALYSIS

RESULTS AND DISCUSSION

I think it is no surprise that all these districts are very centrally located in the circular layout of Warsaw. Locations meeting the criteria of popular places would usually be in central locations in many cities around the world. From this visualization it is clear that on a practical level, without data on the basis of which decisions could be made, the circle of 103 locations is very large. We have significantly narrowed the search area from 8 potential districts to 5, which should respond to the business problem.

Moreover, FourSquare is not popular in Warsaw, the data maybe out-dated or unreliable, the report should gather more data from other location data source such as Google Place API.

CONCLUSION

Different applications of this analysis are available based on a different methodology and possibly different data sources. The stakeholder problem has been resolved. The stakeholder wants to find the best place to live in Warsaw, and the "best location" factors are based on the number of places in the food, cafe and park category around the location. Machine learning technique based on content filtering is the most appropriate method to solve the problem. Eight destination locations may not be a good choice, but I can quickly choose other locations and issue a recommendation again.