**Choosing the Best Districts for us to Live in Warsaw**

Jun Jie, Wong

2 Aug 2020

1. **Introduction**
   1. **Background**

My wife and I reside in Warsaw, Poland and we have been renting out an apartment ever since we got married. As part of our plan to start a new family, this Fall, we are planning to get a place of our own.

* 1. **Problem**

Different families have different needs and my wife and I are no exception. With 18 districts to choose from, we believe there are some that are more suitable for us than others. That is, some districts have venue categories that are more important to us than others. However, there is no such analysis tailored to us anywhere. Thus, once I was presented with the skills and opportunity to apply my data science skills on a project, I set forth to discover which districts are the most suitable for us to live in Warsaw, Poland.

**2. Data**

The following are the data that I needed:

1. List of boroughs in Warsaw
2. Boroughs’ respective latitudes and longitudes
3. Warsaw map overview of all boroughs
4. Relationship between boroughs and radius around borough
5. Venue categories that are within the aforementioned radius
6. Filtered venue categories
7. Merged venue categories

Let me walk through each point in detail.

1. List of boroughs in Warsaw:

The list of boroughs was manually obtained from Wikipedia [https://en.wikipedia.org/wiki/Category:Districts\_of\_Warsaw]. It was done manually because the data was not in table format and so I could not use pandas to web scrape the data from the website.

2b) Boroughs’ respective latitudes and longitudes:

This data was needed to feed into the Foursquare API later to find venue categories of each borough. Unfortunately, the latitudes and longitudes were not features in the dataset, so I used Geopy’s Nominatim to obtain that data based on the names of the boroughs in 2a.

*2a dataframe 2b dataframe*

2c) Warsaw map overview of all boroughs:

An overview of the location of the boroughs were mapped to ensure location accuracy.

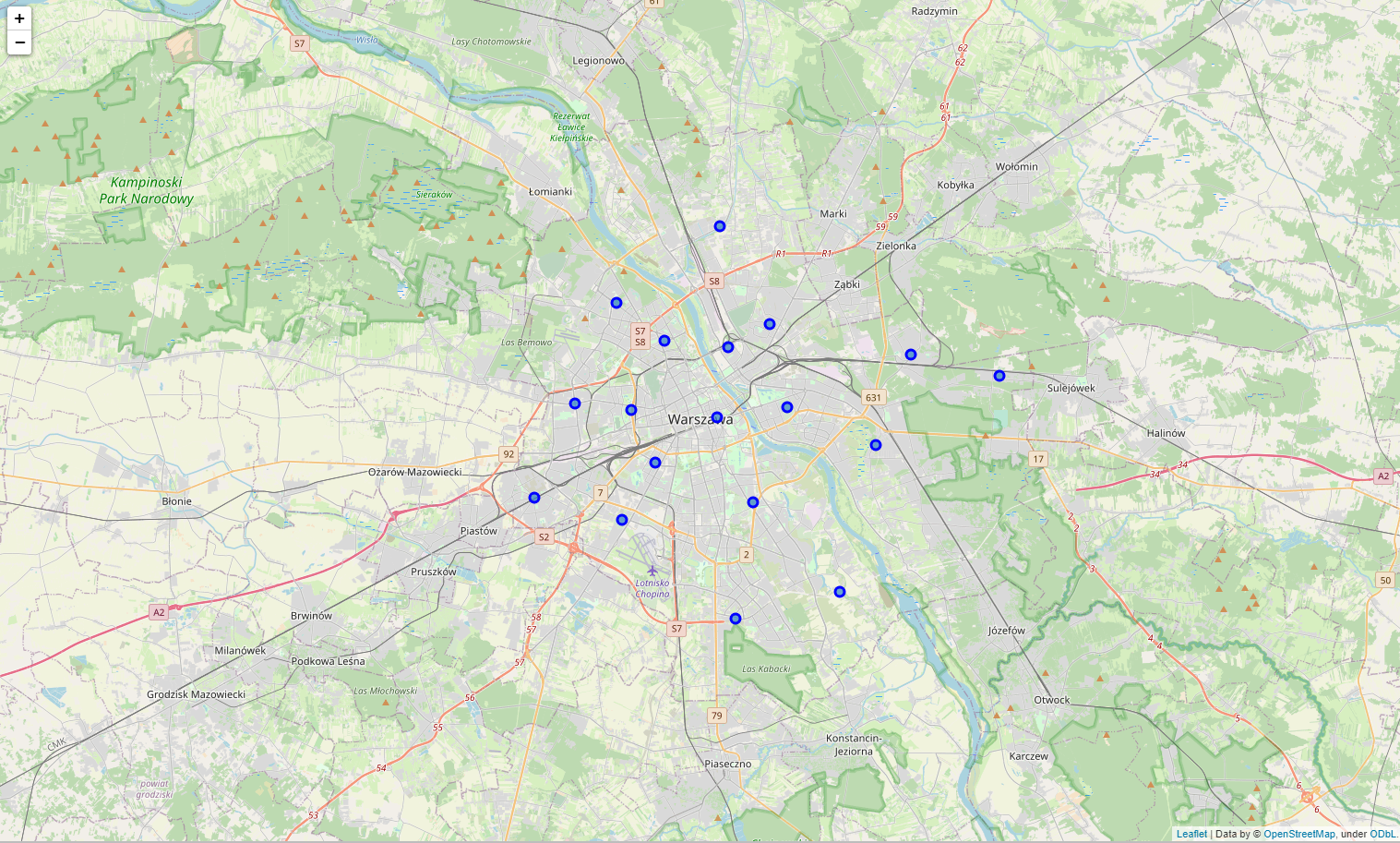


Figure 1: Location of boroughs in Warsaw, Poland

1. Relationship between boroughs and radius around borough:

The boroughs were mapped with different radius values around each borough in order to determine a sensible radius value that will feed into the Foursquare API. The API will return all the venue categories that are within the radius from the center of each district. A too large radius value will return too many overlapping venue categories while a too small radius value will return too little results. The final radius value was determined to be 2km.

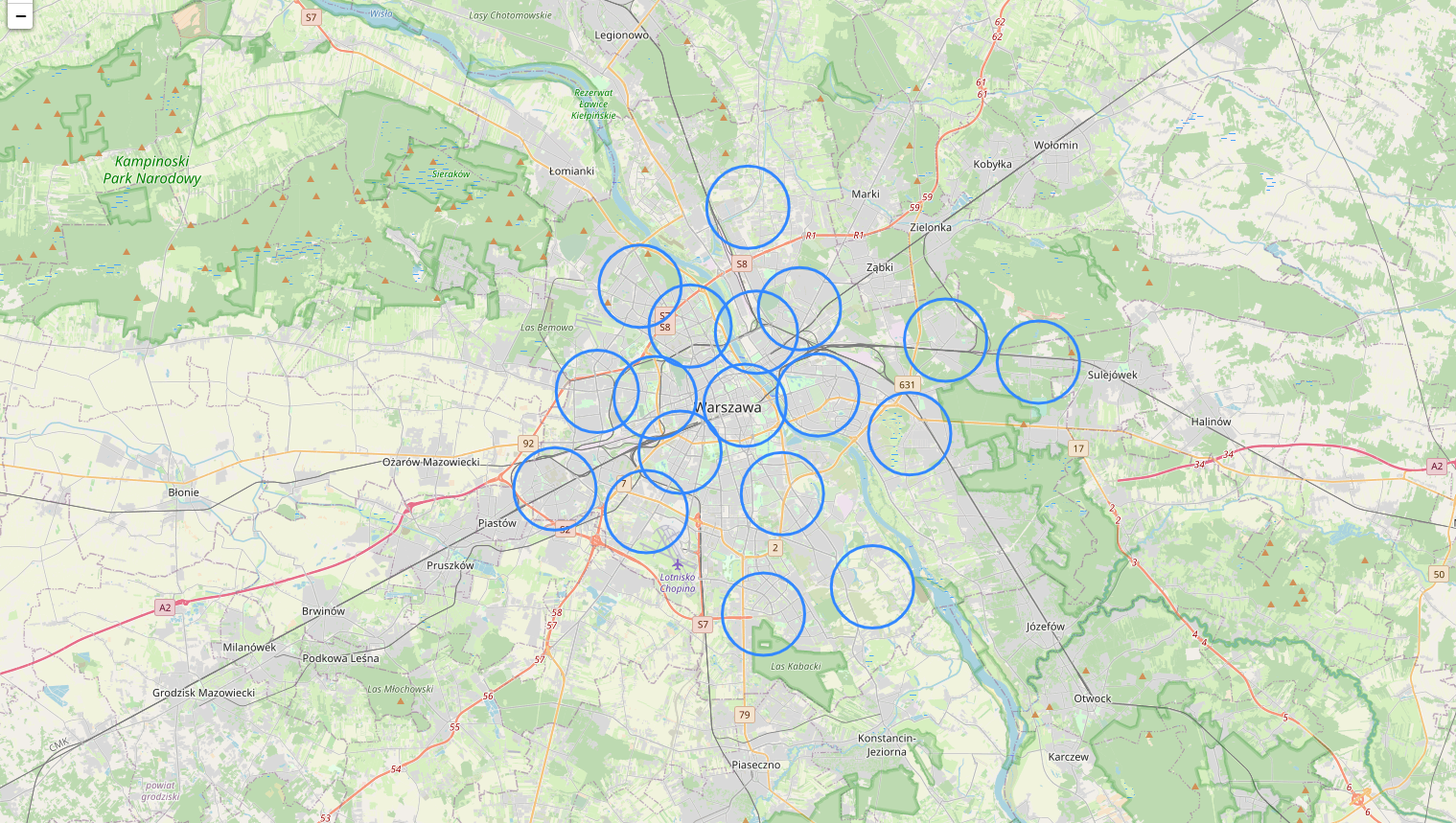


Figure 2: Map of area covered from center of each borough with radius of 2km

1. Venue categories that are within the aforementioned radius:

This data was gotten using Foursquare API by using the latitudes, longitudes, and radius as parameters. The output was all the venue categories that are within a 2km radius for each district.

1. Filtered venue categories:

Many of the returned venue categories were not applicable towards making housing decision for my wife and I. Hence, we filtered only venue categories that were relevant (and thus suitable) to us. See appendix 7a for full list of filtered venue categories.

1. Merged venue categories:

Some of the venue categories were very similar to each other. For example, ‘Outlet Mall’, ‘Outlet Store’, ‘Shopping Mall’, ‘Department Store’ could all be grouped under ‘Shopping’, and ‘Drugstore’, ‘Pharmacy’ could be grouped under ‘Pharmacy Amenities’. See Appendix 7b for full list of grouped venue categories. This dataframe can finally be fed into a machine learning algorithm to cluster the districts based on the types of venue categories. The output will also allow us to find out how each cluster was made to narrow down which was the most suitable for us. The districts in that particular cluster would be the best for my wife and I to choose to buy our new home.

**3. Methodology**

* 1. GeoPy’s Nominatim was used to obtain each borough’s latitudes and longitudes using only the name of the borough and its city (Warsaw) as input parameters. The resulting data frame is the 2b dataframe as shown above.
  2. Folium was used on three occasions. The first occasion to visualize the geographic accuracy of the boroughs in Warsaw, Poland. I created a map of Warsaw with boroughs superimposed on top. The second occasion was to visualize the radius value for each borough in order to obtain sensible data as mentioned in point 2d above. The third occasion is to visualize the clusters by giving each cluster its own color code as seen in Fig. 3 below.
  3. Foursquare API was used to explore the boroughs and cluster them. I designed the limit as 200 venue and the radius of 2000 meter for each borough from their given latitude and longitude information.
  4. K-means algorithm, with a 5 clusters input was used to segment the boroughs based on their venue categories. The results are as shown in ‘Results’ below.

**4. Results**

Four clusters were obtained.

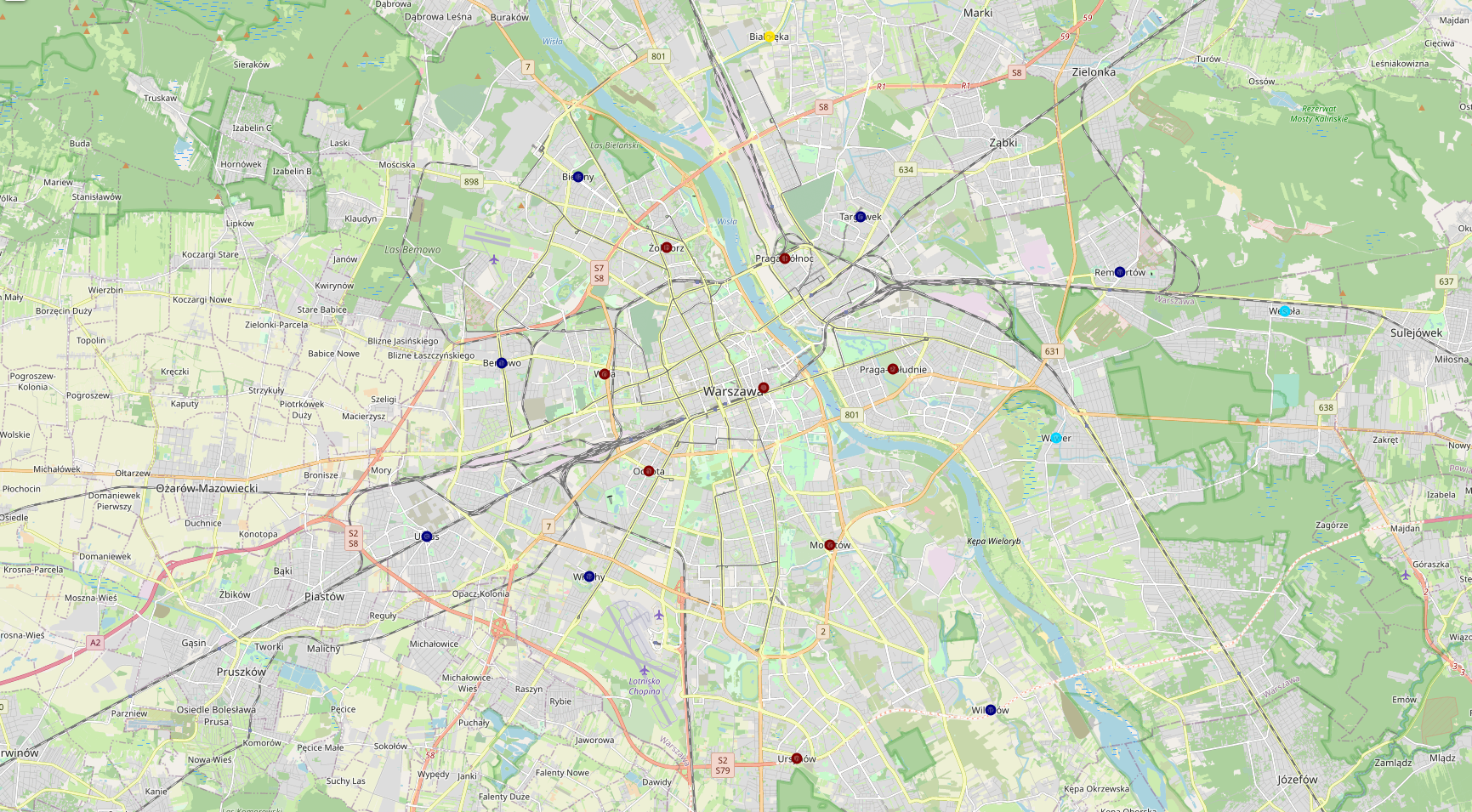
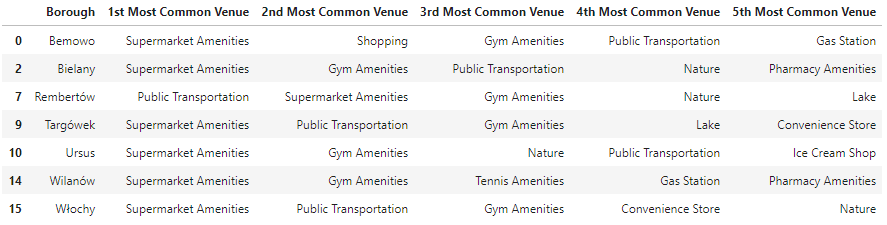


Figure 3: 4 clusters of boroughs differentiated by color

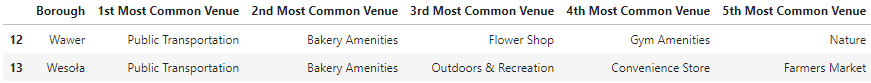
Cluster 1 (Dark red):



Cluster 2 (Dark blue):



Cluster 3 (Light blue):



Cluster 4 (Yellow):



**5. Discussion**

Each cluster of borough(s) has the following characteristics:

|  |  |
| --- | --- |
| Cluster | Characteristics |
| 1 | These boroughs have Nature and Gym Amenities as the most common type of venue. They are most suitable for families who would enjoy nature or the occasional walk in the park and/or working out. |
| 2 | These boroughs have supermarkets as the most common type of venue. They are most suitable for families who purchase groceries frequently and with specific items/brands to purchase. These families would probably want to be practical and live close to different supermarkets since different supermarkets carry different groceries at any one time - being in a borough with many supermarkets will almost guarantee the availability of items the families desire. |
| 3 | These boroughs have public transportation as the most common type of venue. They are most suitable for families who probably do not own a car and would want to be leverage on public transportation such as the bus and metro stations the most. |
| 4 | With cinema as most common type of venue, this borough is best for families who are frequent movie-goers. |

**6. Conclusion**

Since my wife and I enjoy nature more than other housing decision factors, we would prefer to purchase our new place in one of the boroughs as listed in Cluster 1. The results also tell us to avoid boroughs from Cluster 3 since we already own a vehicle and public transportation is not a priority for us.

**7. Appendix**

7a) Full list of venue categories:

Bakery, Bus Line, Bus Station, Bus Stop, Beach, Convenience Store, Deli / Bodega, Department Store, Drugstore, Farmers Market, Flower Shop, Gas Station, Garden, Grocery Store, Gym, Gym / Fitness Center, Gym Pool, Ice Cream Shop, Lake, Market, Metro Station, Movie Theater, Multiplex, Outlet Mall, Outlet Store, Park, Outdoors & Recreation, Pharmacy, Scenic Lookout, Shopping Mall, Supermarket, Tennis Court, Tennis Stadium, Train Station, Tram Station

7b) Full list of grouped venue categories

Table 1: Grouping similar venue categories

|  |  |
| --- | --- |
| **Grouping** | **Venue categories** |
| Bakery Amenities | Bakery, Deli / Bodega |
| Cinema | Movie Theater, Multiplex |
| Gym Amenities | Gym, Gym / Fitness Center, Gym Pool |
| Nature | Park, Garden, Lake, Outdoors & Recreation, Scenic Lookout |
| Public Transportation | Bus Line, Bus Stop, Bus Station, Metro Station, Train Station, Tram Station |
| Pharmacy Amenities | Drugstore, Pharmacy |
| Shopping | Outlet Mall, Outlet Store, Shopping Mall, Department Store |
| Supermarket Amenities | Grocery Store, Supermarket, Market |
| Tennis Amenities | Tennis Court, Tennis Stadium |