Warsaw District Clustering

Background

- Finding a place to live is a daunting task as it requires user to go through multiple variables for every potential apartment.
- While one of the most important aspects of choosing the right apartment to live is finding a location with smallest distance to workplace, the second most important aspect is availability of local infrastructure, which consists of local venues as well as availability of broad range of educational infrastructure for one's children.
- Usually a person would pick a specific district to do analyze and proceed to look for available venues in the vicinity. But how to pick the right district? Does any of its 18 districts (boroughs) cluster - meaning that even though you like south you might consider living in the north?

DISTRICT CLUSTERING IS A FIRST STEP ON THE WAY TO FINDING THE BEST PLACE TO LIVE IN WARSAW

Data Acquisition and Cleaning:

Data Sources:

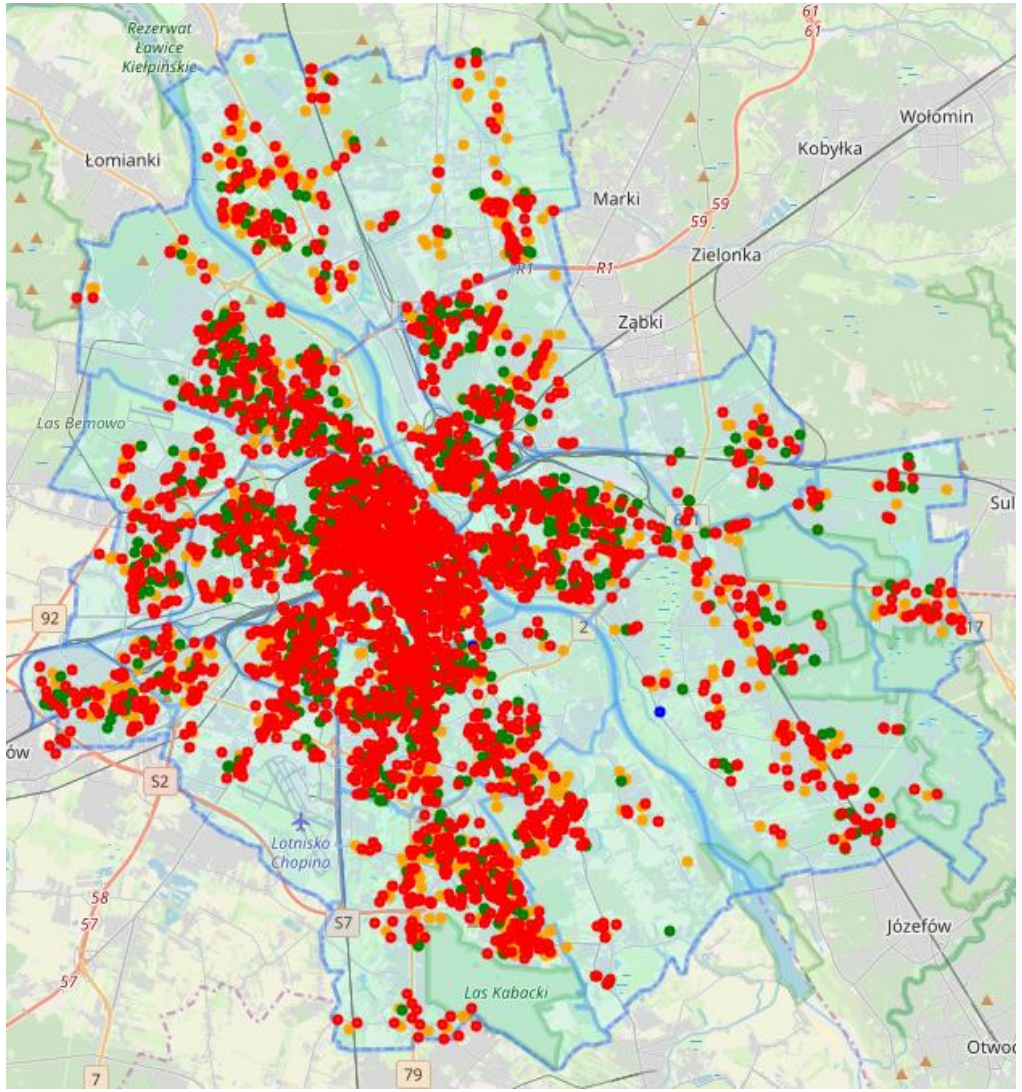
- Data API currently developed by Warsaw City: https://api.um.warszawa.pl/#
- Data API currently developed by Polish Government: https://www.dane.gov.pl/institution/65,miasto-stoleczne-warszawa?page=1&per_page=5&q=&sort=-verified
- Dataset provided by the City of Warsaw: https://edukacja.warszawa.pl/placowki/przedszkola
- GeoJson representing position of Warsaw districts: https://github.com/andilabs/warszawa-dzielnice-geojson

After processing NaNs, determining geographical position and making sure that each venue data row has a district assigned (based on its geographical position) each and every available dataset has been processed into template form, with template consisting of the following categories:

- District
- Venue
- Neighborhood Latitude
- Neighborhood Longitude
- Venue Latitude
- Venue Longitude
- Venue Category

The dataset consisted of 372 unique venues, which have been used for the analysis.

Data Visualization

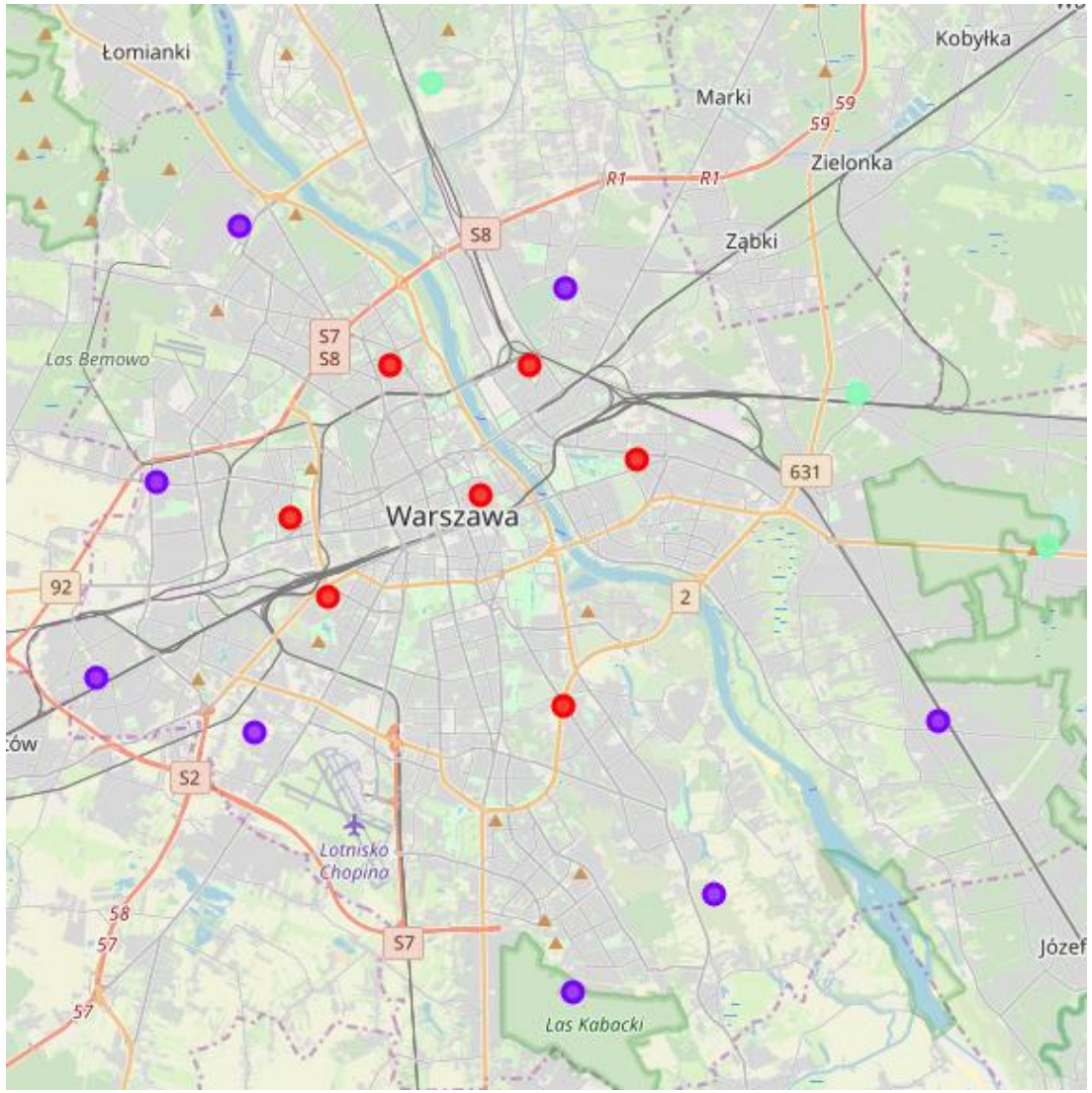


District	
Bemowo	206
Białołęka	312
Bielany	303
Mokotów	762
Ochota	309
Praga Południe	506
Praga Północ	250
Rembertów	51
Targówek	182
Ursus	128
Ursynów	418
Wawer	262
Wesoła	94
Wilanów	126
Wola	457
Włochy	175
Śródmieście	1358
Żoliborz	146

Resultant dataset consists of 6045 rows, ranging from 51 to 1358 features for the district. This categorical unbalance is a natural result of a diverse population density across Warsaw districts.

District Clustering

	District	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude
0	Bemowo	1	52.239606	20.899061
1	Białołęka	2	52.328444	20.998763
2	Bielany	1	52.296865	20.929070
3	Mokotów	0	52.189544	21.047263
4	Ochota	0	52.214050	20.961090
5	Praga Południe	0	52.244749	21.074174
6	Praga Północ	0	52.265506	21.034593
7	Rembertów	2	52.259174	21.154148
8	Targówek	1	52.283011	21.047594
9	Ursus	1	52.195898	20.876551
10	Ursynów	1	52.125437	21.050433
11	Wawer	1	52.186094	21.183994
12	Wesoła	2	52.225167	21.223771
13	Wilanów	1	52.147384	21.102316
14	Wola	0	52.231552	20.947744
15	Włochy	1	52.183478	20.934407
16	Śródmieście	0	52.236558	21.016590
17	Żoliborz	0	52.265548	20.983701



Conclusion

It seems that district clustering approach has provided a good, comprehensive, based in reality clustering division between Warsaw's districts. Inner districts are indeed the most expensive ones, followed by Vistula's (Warsaw main river) left bank outer districts and finally - outer east Vistula's bank districts.

The inner city was specified as cluster 1 and combines districts of: Mokotów, Ochota, Praga Południe, Praga Północ, Wola, Śródmieście and Żoliborz. Outer regions were clustered into cluster 2 and consist of districts of: Bemowo, Bielany, Targówek, Ursus, Ursynów, Wawer, Wilanów & Włochy. Finally the remaining neighborhoods were clustered into set 3 and consist of districts: Białołęka, Rembertów & Wesoła.