Clustering Of Warsaw Districts

Maciej Pawełczyk

January 19th 2020

1. Introduction

1.1 Background

Warsaw is a 517.24 km2 capital of Poland. While it has been founded in IXth century, its current urban form is a result of severe destruction during World War II (including almost complete destruction of all infrastructure on the left bank of Vistula river). As a capital city of booming economy it grows rapidly in all direction, including suburbs with new commercial and apartment building being developed on daily basis. This poses a challenge for its 1.8 mil population (and additional people migrating into Warsaw) of how to determine the best place to live.

1.2 Problem

The 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?

1.3 Interest

The target audience is any person already living in Warsaw, who consider relocating as well as any person considering migrating into Warsaw.

2. Data acquisition and cleaning

2.1 Data sources

Naturally, the project will utilize FourSquare API in order to acquire as much of the venue data as it is possible. Rather than acquiring the data based on neighborhood's (called 'Districts' in this notebook) central position and then proceeding on based on venue's location, the notebook utilize the location based on other venues acquired from different data sources.

Those data sources consist of Warsaw's infrastructure data, specifically schools, educational facilities, kindergartens etc. along with places of culture (in this instance - theaters). If the data does not come with the geographical position, a separate code is used to obtain that data.

The end result is a detailed, comprehensive summary of important Warsaw information, clearly presented using Folium maps.

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

2.2 Data cleaning

Available data has been processed to ensure that every single venue, both FourSquare API and other has geographical position available (latitude, longitude). In order to do so Nominatim geocoder has been utilized to translated address based information on the educational facilities into latitude/longitude based information.

Furthermore some of the datasets had no intrinsic district information. This information was crucial for further analysis. This is why additional Warsaw districts .geojson file has been obtained and used. Each location has been looped over to determine if specified location lies within district geographical polygon.

Finally 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

2.3 Feature selection

The dataset consisted of 372 unique venues, which have been used for the analysis. Considering that clustering Data Science technique has been used, which is an example of unsupervised learning method, none of the features have been removed.

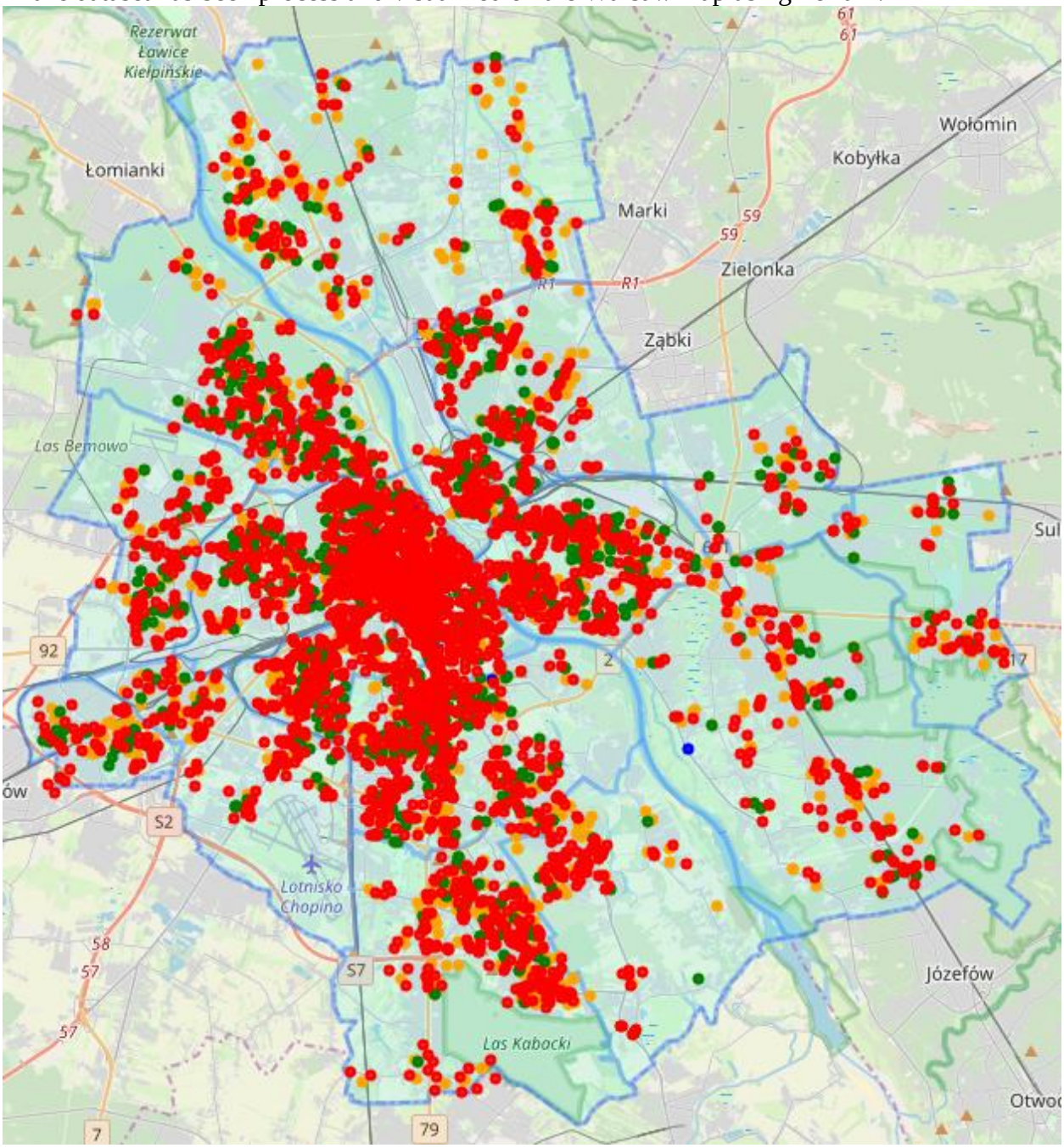
Furthermore as the number of available venues does indeed reflect the nature of the district it is a predictive feature and as such no normalization have been utilized for the number of features.

The features have been however normalized for their corresponding frequencies (summing up to 1).

3. Exploratory Data Analysis

3.1 Data visualization

Entire dataset has been process and visualized on the Warsaw map using Folium:



Each of the district can be zoomed in to determine availability of different venues.

3.2 Dataset composition

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	
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

4. Clustering

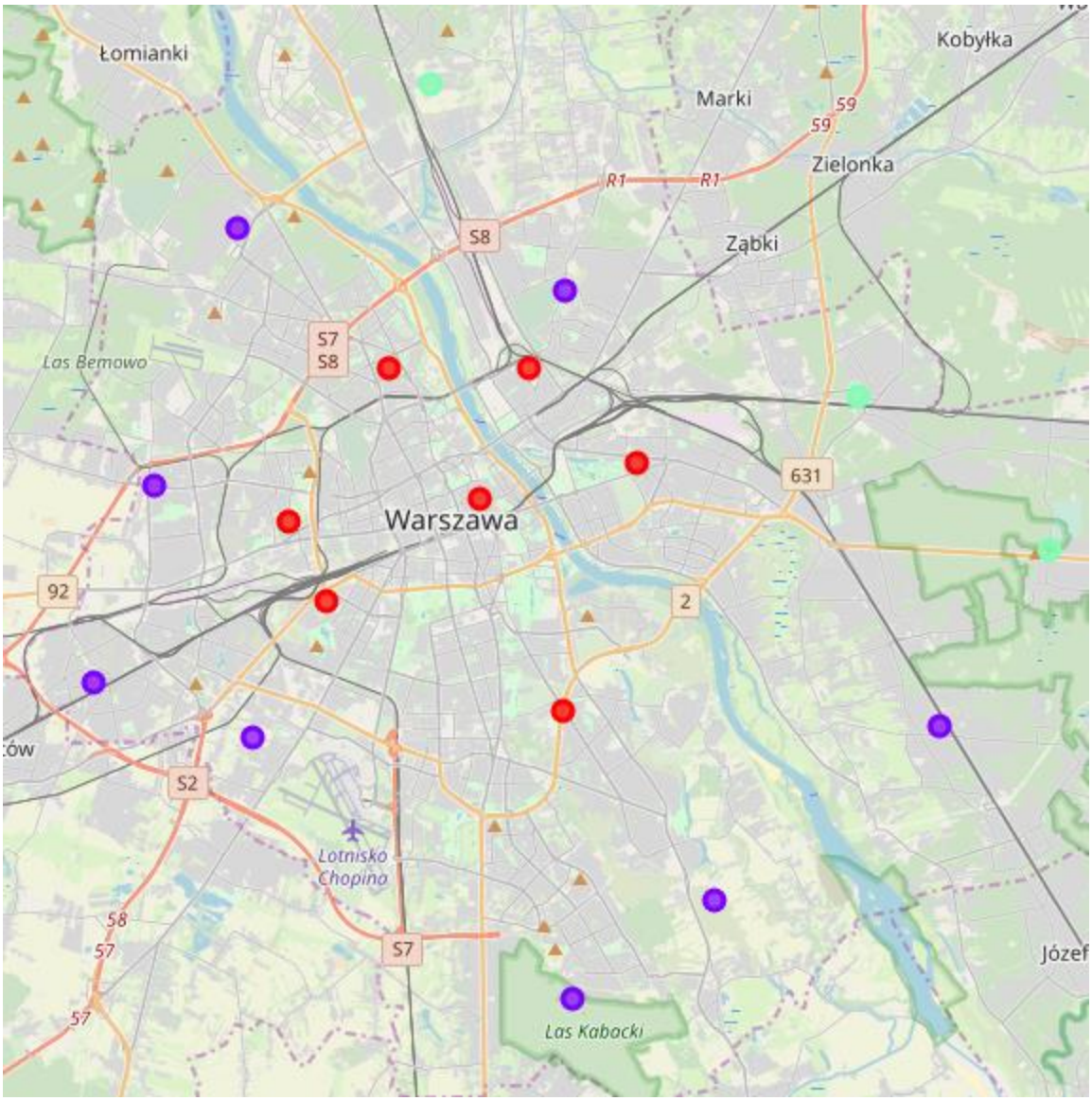
Warsaw neighborhoods have been clustered based on the most often recurring venues.

Obtained analysis has shown that the best cluster set is a set of 3 items, as it is the only one, which allows clustering of multiple districts in one cluster.

This kind of grouping is also based in reality - naturally, the inner city districts have better infrastructure, mainly due to higher density of population and longer historical background. That being said, the best fit of 3 clusters is unexpected and is mainly attributed to the data distribution.

	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

Result visualization:



5. Conclusions

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.

6. Future directions

Clustering into 3 groups has been made based on acquired data. The notebook has utilized lots of infrastructure data, but more could be obtained in the future. This includes shop, mall, hospital, cinema, park, grocery story information and multiple other sources that could be used to make the map more comprehensive.

Additional information, such as traffic information, real estate prices, heatmap of available commercial space and multiple more could be used to create a good estimate on the Warsaw's real estate price. This in turn could be used to determine real estate development opportunities as well as would allow user to find a price outlier for apartment price, eventually turning this notebook into valuable business opportunity.