Finding the most popular venues in districts of Moscow

Viktor Klementev, klementiev.victor@gmail.com

Introduction

Business problem

As the topic for our project should be related to "battle of neighborhoods", I've decided to compare districts of Moscow, Russia. There are 119 districts here and there should be some similarities among them as there are differences in real estate prices, the number of population as well as the number of workplaces (or district has residential status, by that I mean the district where people mostly live but go for work to another district where there are office centers or other working facilities). Foursquare API gives the opportunity to explore the list of venues for Moscow. The means of pandas, folium and some other python extensions will help to divide Moscow territory by districts and find out some similarities or differences among districts based on venues popularity analysis.

A brief description of research

I'm going to analyze the following parameters for every district: Moscow foursquare data on macro level as well as some socioeconomic indicators related to these districts (population, real estate price, feature of work/residential district). Using this information, I'm going to analyze Moscow districts from two points – district' grouping based on analysis of selected socioeconomic indicators and clustering based on analysis of foursquare data.

In overall I plan that clustering Foursquare data on Moscow districts on the one side and socioeconomic indicators groupings on the other side will give a sufficient characteristic of every Moscow district.

Who would be interested in this project

So that the results of my research will be interesting for people who study city environments to develop their business as well as people studying what Moscow district to move based on its features I've got in this research (by the way Moscow is the most dynamic Russian location right now and there is huge migration inflow out of other country regions), interested parties for this research could also be sociologists who study city environments. Nowadays Moscow is rapidly changing and there is high attention for urban studies there. I hope some insights into this area will be helpful for these urbanists too and make a little contribution into knowledge of current state of affairs in Moscow.

The information I get will be helpful to find similarities among districts, benchmark the most popular districts venues and elicit some new business opportunities based on absence/presence of venues in similar districts. Similarities of districts will be found based on socioeconomic indicators. I also hope that combination of some statistics information with Foursquare data would be unique to some extent.

I understand that the results of my work would be an overview as there will be no any advices in what neighboorhood it is fine to open a new coffee shop. However, my research could help to understand the overall situation in every Moscow district and find out what business opportunities (related to eating/entertainment business) are hidden in every district of Moscow (e.g. to find two districts with similar population and real estate prices and reveal that some venues that are very popular in one district could also be popular for another same one).

Data section

The data I'm going to analyze will be numeric, in the format of table. I'm going to extract several data columns from different sources and merge them into one dataframe. The basic column will be the list of Moscow districts. I will get this information in Russian so that i need to fond some option to translate this list to English. When my merged dataframe will be ready I will add additional dummies-like columns that will group districts into several cohorts based on price and work/residential feature.

Data that will be used and analyzed:

1. socioeconomic indicators
2. foursquare data on venues in Moscow
3. json data with spatial data for Moscow districts to build choropleth maps (but this data for use only, not for analysis).

1. For my research I choose the following socioeconomic indicators to analyze:

– population of every district (Source – Wikipedia, there is a list of Moscow districts and their population, in Russian only, *https://ru.wikipedia.org/wiki/%D0%A1%D0%BF%D0%B8%D1%81%D0%BE%D0%BA_%D1%80%D0%B0%D0%B9%D0%BE%D0%BD%D0%BE%D0%B2_%D0%B8_%D0%BF%D0%BE%D1%81%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B9_%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D1%8B*),

– real estate prices per sq m for every district (there is web-site with information in Russian only, *https://www.irn.ru/rating/moscow/*, it is the popular Russian real estate web site with classified on real estate topics and related analytics),

– the estimation of whether the district in mostly consists of office and workplaces or is residential area (there was a research made by Yandex where they estimate what percentage of people living in certain district constantly and what percentage are moving daily to work and back home out of their district, *https://cache-spb02.cdn.yandex.net/download.yandex.ru/company/figures/2016/home_work/districts_f.html*).

To get population and real estate prices data I'm going to use web scrapping, for work/home data I collect information manually as it was in folium-like map and right now I'm not able to do web-scrapping from interactive maps.

During analysis, I'm going to use the following python commands and modules:

– geopy and nominatim for getting district coordinates, requests and beautifulsoup4 for web-scrapping,

– replace, split, rename, frop column, change the order of columns, delete some columns, change the data type in column from object to integer, set/reset index, open/save as excel and csv, merge and join columns.

This data will help to sort district on business/residential and sort them based on real estate prices and its population. I hope this information will be helpful to explain why certain types of Foursquare venues are popular in some districts e.g. I expect supermarkets to be popular in the residential areas whether in business locations coffee shops or restaurants will be on the top. There will be information on 119 districts of Moscow and more then 2 600 foursquare venues to analyze.

2. As for Foursquare data, I'm going to use the following python commands and modules to extract and work with data:

– json, pandas.io.json, getnearbyVenues, onehot, commands to relate venue location to the name of Moscow district, kmeans clustering.

– from Foursquare i will get the type, latitude and longitude of the venue.

– clustering and closeness in socioeconomic features i've listed will show how districts are similar and different among each other and what business needs could be implemented there.

3. I've already bought prepared data on Moscow districts' borders from the site *https://my.nextgis.com/*, they use OpenStreetMap data as a source. I also used Google Map to get the coordinates of Moscow center, I choose Zaryadye Park that is situated near Kremlin to get these coordinates. As for center coordinates of the every district, I've extracted them automatically using 'Nominatim'.

Working plan

Here is my plan to work with the data. Analysis will be held in table format in dataframes and folium maps:

– to do web-scrapping with 'beautifulsoup' from two web-sites. Web-pages I need to scrap have a very 'dirty' code neigborhoods. So that I expect to spend quite a big amount of time to clean the data. In data cleaning I will use such code excerpts as 'table.replace(,)', 'table.split()', 'df.drop';

– as I'm working with Russian data I need to have some translation. I will try to use transliteration modules to convert Russian names of districts ('transliterate', 'google translate') but it was in vain because the wording of districts was not the same I it should be officially. The other way is to translate them using web scrapping and 'join/merge method' or manually. I chose to do it manually;

– then I need geographic coordinates for every district. I will use 'geopy.geocoders' extension and import 'Nominatim'. For the purpose of map creation I need to have latitude and longitude as two separate columns but 'Nominatim' give them as one so that I need to find a solution to separate column in dataframe;

– as I will get three separate tables with data, I need to merge them using the column with district names in English;

– I'm going to add new columns with characteristics of socioeconomic indicators to put them easily on a map and classify further in cluster analysis results. E.g. for 'residential areas' or 'areas mostly for work' I will not use dummy variables as '0' and '1' meanings come in two columns;

– for Foursquare I'm planning to analyze the overall number of venues to get a big picture of what is going on in Moscow and in what district;

– then there will be imports of 'folium' extension to build the map and k-means to do clustering. I'm going to get the data of district ('neighborhood') name not from json file but out of web-scrapping and usage of 'Nominatim' extension;

– in the final cluster analysis i'm going to match the data i've obtained during clustering analysis and comparisons of socioeconomic indicators i've described earlier. Thus making conclusions using final tables of clusters will be easier.

Methodology

My work will consist of three parts:

1. Data extraction and preparation. Mapping the data (though in code part I divided those as two different parts).

Data extraction from web-sources I've mentioned above to get statistic information for every district of Moscow, then cleaning and preparation of data. Finally for this point, I will group districts based on their similarities on population density real estate prices (e.g. groups of 'cheap' real estate less than 2.500 US dollars, medium prices up to 4.000 US dollars, expensive housing with prices above 4.000 US dollars) and work/residential feature of every district (e.g. working district could be considered as the one where 60% of population live constantly and 40% are people who come there for working hours from other parts of Moscow, however - this estimation is intuitive, i chose this proportion as there is a share of people who work and live at the same district). As a result I will divide Moscow territory on working/living areas, areas with expensive, medium and relatively cheap real estate, areas depending on size of population. I will get the number of clusters that will further use in kmean analysis.

Example of web scrapping of Wikipedia page (English translation is provided in the code). Before:

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	Академический	Академический	ЮЗАО	5	83	109387	18762.78	2467	0	22	7
1	2	Алексеевский	Алексеевский	СВАО	5	29	80534	15223.82	1607	9	20	5

After the data cleaning:

	District	Population	Population density	District square	address for Nominatim	location	point
0	Академический	109387	18762.78	5.83	Академический , ЮЗАО	(Академический, Москва, Юго-Западный администр...	(55.6897377, 37.5767712, 0.0)
1	Алексеевский	80534	15223.82	5.29	Алексеевский , СВАО	(Алексеевский, Москва, Северо-Восточный админи...	(55.8148783, 37.6506684, 0.0)

Another example of web scrapping of real estate web site page. Before:

```
table1
```


```
'\n\n,1,\n,\nОстоженка ,\nКропоткинская, \t\t\t\t\t\t\t\tПарк культуры ,395\xa0794,+1,3%\n\n,2,\n,\nЯкиманка ,\nНовокузнецкая,
\t\t\t\t\t\t\t\t\tПолянка, \t\t\t\t\t\t\t\t\tТретьяковская ,376\xa0733,+0,8%\n\n,3,\n,\nАрбат ,\nАлександровский сад, \t\t\t\t\t
\t\t\t\tАрбатская, \t\t\t\t\t\t\t\t\tБиблиотека имени Ленина, \t\t\t\t\t\t\t\t\tБоровицкая, \t\t\t\t\t\t\t\t\tСмоленская ,364\xa
0093,+0,5%\n\n,4,\n\t\t\t\t+1\t\t\t,\nЦентр Москвы ,\nКитай-город, \t\t\t\t\t\t\t\t\tКузнецкий мост, \t\t\t\t\t\t\t\t\tЛубянка,
\t\t\t\t\t\t\t\t\tОхотный ряд, \t\t\t\t\t\t\t\t\tПлощадь Революции, \t\t\t\t\t\t\t\t\tТеатральная ,355\xa0644,+1,1%\n\n,5,\n\t\t
\t\t-1\t\t\t,\nТверской ,\nМаяковская, \t\t\t\t\t\t\t\t\tПушкинская, \t\t\t\t\t\t\t\t\tТверская, \t\t\t\t\t\t\t\t\tЧеховская ,34
9\xa0883,-0,8%\n\n,6,\n,\nХамовники ,\nЛужники, \t\t\t\t\t\t\t\t\tСпортивная, \t\t\t\t\t\t\t\t\tФрунзенская ,322\xa0677,+0,8%\n
\n,7,\n\t\t\t\t+1\t\t\t,\nЗамоскворечье ,\nДобрынинская, \t\t\t\t\t\t\t\t\tОктябрьская, \t\t\t\t\t\t\t\t\tПавелецкая, \t\t\t\t\
\t\t\t\tСерпуховская ,306\xa0195,+2,1%\n\n,8,\n\t\t\t\t-1\t\t\t,\nПресненский ,\nБаррикадная, \t\t\t\t\t\t\t\t\tВыставочная, \t
```


After the data cleaning:

	District	Price per sqm
0	Остоженка	395794
1	Якиманка	376733
2	Арбат	364093
3	Центр Москвы	355644
4	Тверской	349883

2. Venues extraction from Foursquare and clustering. For Moscow there are 2.683 enues and 297 unique venues categories.

Analysis of Foursquare venues in Moscow and clustering them. Explored with machine learning tools Foursquare data will show whether my logical division of Moscow territory from point 1 is true or offer another clustering based on Moscow venues popularity.

After I get information on socioeconomic indicators in the format I need, as well as I conduct the analysis of foursquare data using clustering, I'm going to integrate Foursquare results analysis with clustering I've got on socioeconomic indicators analysis. As a result, I will show top-3 venues that are popular for every district and notify the characteristics of these districts out of socioeconomic indicators I've got.

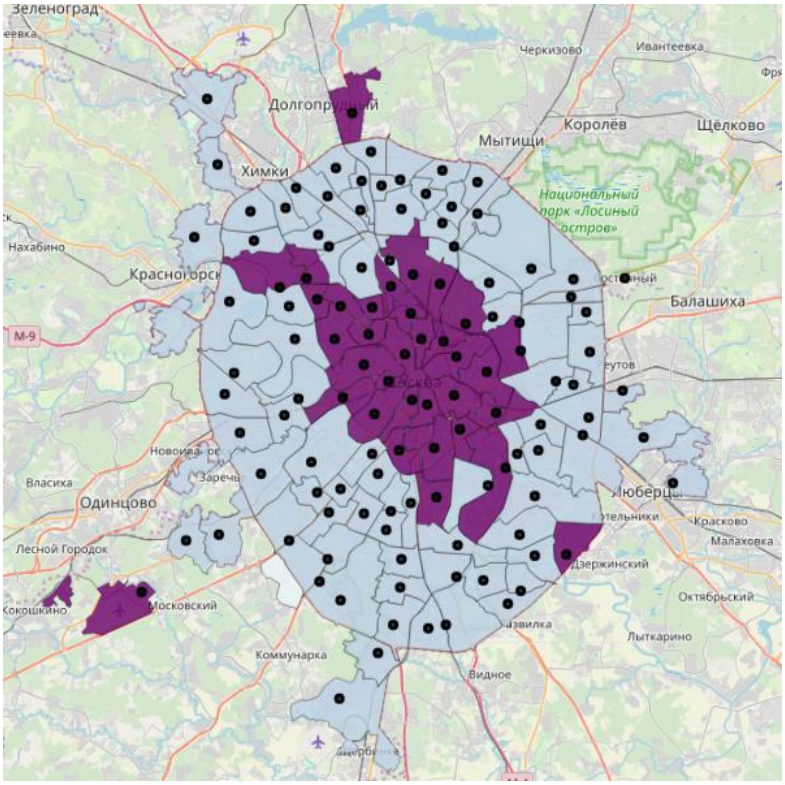
3. Visualized outcomes of district analysis for target audience in the form of maps.

Results section

After data extraction and preparation I've got the following table that I've used further for maps creation. Through the process of maps creation I want to distinguish the number of groups by which Moscow districts could be divided. Then I will use this number to set up the number of clusters in kmeans analysis. Pay attention that there are dummies like columns signifies work/residential district and price categories in numbers, I made them to put this division into maps.

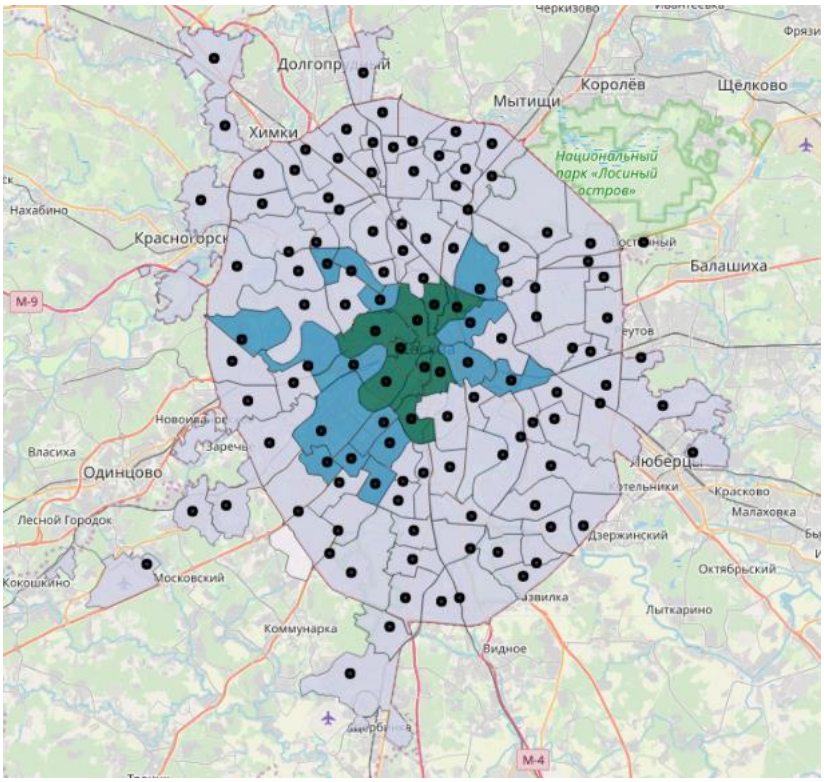
	District	District_en	Population	% of homeplaces	% of workplaces	Price per sq m	Latitude	Longitude	Work/residential district	Dummies for work/residential district	Price categories	Dummies for price
0	Академический	Akademichesky District	109387	41.5	58.5	205702	55.689738	37.576771	residential	0	average	3
1	Алексеевский	Alexeyevsky District	80534	35.4	64.6	191495	55.814878	37.650668	mostly for work	1	mass-segment	2
2	Алтуфьевский	Altufyevsky District	57596	60.5	39.5	147175	55.880255	37.581635	residential	0	mass-segment	2
3	Арбат	Arbat District	36125	12.7	87.3	362197	55.751199	37.589872	mostly for work	1	expensive	4
4	Аэропорт	Aeroport District	79486	38.5	61.5	206218	55.800402	37.533156	mostly for work	1	average	3
5	Бабушкинский	Babushkinsky District	88537	60.0	40.0	154080	55.865958	37.663894	residential	0	mass-segment	2
6	Басманный	Basmanny District	110694	17.0	83.0	224197	55.767281	37.669773	mostly for work	1	average	3
7	Беговой	Begovoy District	42781	18.8	81.2	237475	55.781917	37.566300	mostly for work	1	average	3
8	Бескудниковский	Beskudnikovsky District	79603	75.1	24.9	149875	55.863739	37.557322	residential	0	mass-segment	2

The first folium map I created includes popups that are coordinates of some point inside every Moscow district that I've got out of Nominatim. District borders I've extracted from json file I've downloaded from nextgis.com (in turn they created it using information from OpenStreetMaps). Chronolepth map here shows the distribution of districts where people are mostly working (dark color) and where they are mostly living (light color). With this map we get two groups of districts division.

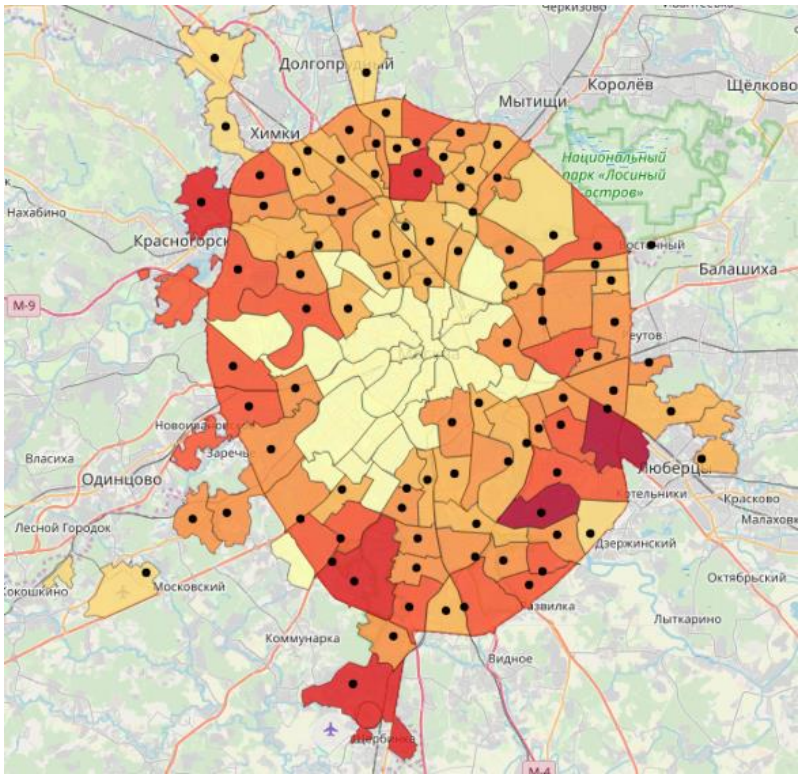


The second chronolepth map below shows the price differences for real estate in three groups of price division – expensive, average and mass-segment. The most expensive real estate in the districts for work, in the city center so that we now have three groups of districts division:

expensive districts for work, districts for work with average real estate prices, districts for living with average real estate prices. Districts with mass-segment prices will be classified further.



The third indicator I included into my research is population size of districts. I showed it for districts with mass-segment price for real estate. There are six different colors that could be separated into three groups.



So that I got 6 groups of districts and will use this number for cluster analysis.

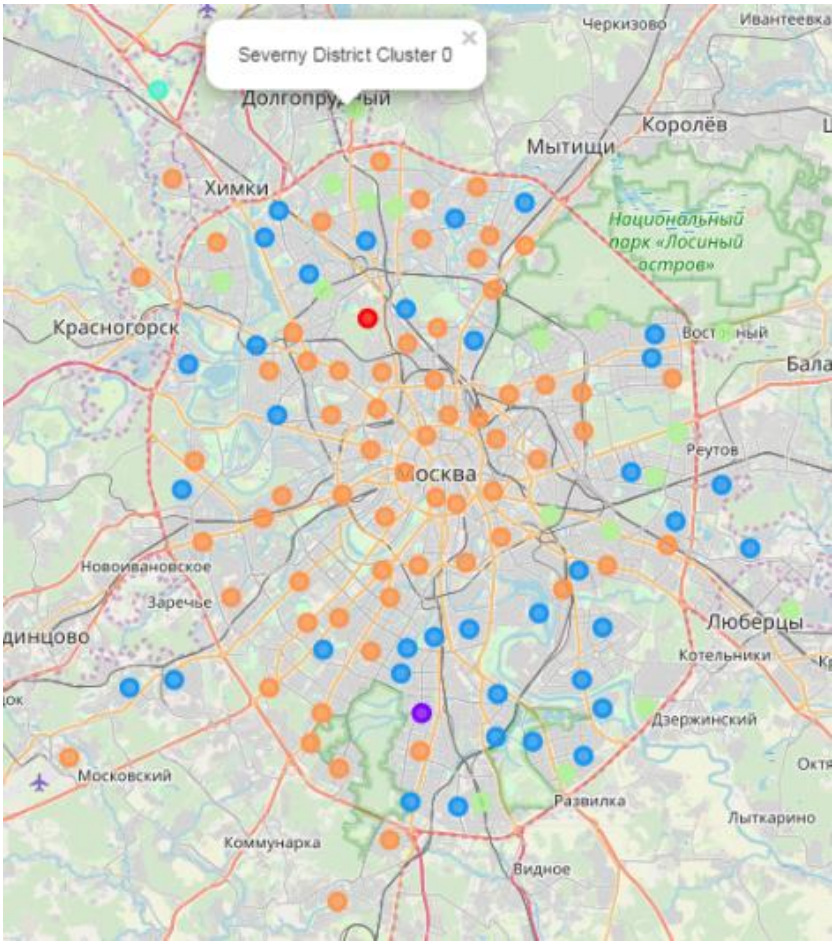
Here is the beginning of the table with cluster analysis results. Each district now has cluster label.

	Neighborhood	Work/residential district	Price categories	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
	Aeroport District	mostly for work	average	55.800402	37.533156	1	Coffee Shop	Cosmetics Shop	Café	Convenience Store	Flower Shop
	Akademichesky District	residential	average	55.689738	37.576771	1	Pharmacy	Bakery	Beer Store	Pub	Park
	Alexeyevsky District	mostly for work	mass-segment	55.814878	37.650668	4	Hotel	Auto Workshop	Supermarket	Mobile Phone Shop	Gym / Fitness Center
	Altufyevsky District	residential	mass-segment	55.880255	37.581635	0	Supermarket	Grocery Store	Convenience Store	Building	Café
	Arbat District	mostly for work	expensive	55.751199	37.589872	1	Coffee Shop	Hotel	Bar	Museum	Yoga Studio
	Babushkinsky District	residential	mass-segment	55.865958	37.663894	1	Mobile Phone Shop	Cosmetics Shop	Notary	Middle Eastern Restaurant	Café
	Basmanny District	mostly for work	average	55.767281	37.669773	1	Park	Hostel	Auto Workshop	Public Art	Recording Studio
	Begovoy District	mostly for work	average	55.781917	37.566300	1	Dance Studio	Gym / Fitness Center	Café	Coffee Shop	Furniture / Home Store
	Beskudnikovsky District	residential	mass-segment	55.863739	37.557322	4	Cosmetics Shop	Liquor Store	Moving Target	Soccer Field	Park
	Bibirevo District	residential	mass-segment	55.883894	37.603577	1	Fast Food Restaurant	Clothing Store	Cosmetics Shop	Toy / Game Store	Pizza Place

Results of cluster analysis are the following:

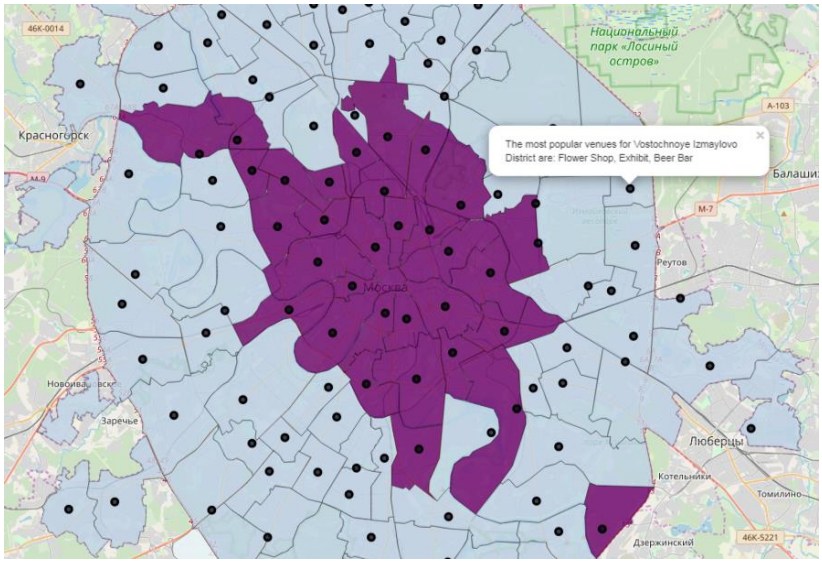
1. Actually there are 3 clusters, as 3 other consist of one district.
2. Some food and everyday supply venues are pre-dominantly concentrated in residential mass-segment districts.
3. The majority of work districts are in one cluster.
4. In cluster 1 supermarket is the most popular venue for residential mass-segment price districts.
5. Coffee shop is the most popular and common venue for mostly for work and expensive districts.

Here is the map of Moscow district clusters:



There are approximately 2.600 venues extracted with the help of Foursquare API. Clustering was made based on their location analysis. Following the goal of my research I revealed the most popular venues for every district and put top-3 of them on popups of maps with real-estate prices, work/residential feature of district and size of its population.

Example of a map with districts division on working/residential and popups with the top-3 most popular venues.



Discussion section

There are 119 districts in Moscow and they are all different – a lot of approaches can be tried in cluster and classify them. I've made two different approaches using logic based on limited number of socioeconomic indicators and elements of machine learning (kmeans clustering) but there could be some other results of clustering and classification depending on indicators and volume of data researcher uses.

Limitations for the study is that I generalize information based on district boundaries (as there could be different population density or real estate prices within one district) and Foursquare venues are limited in numbers (obviously there are more than >2600 venues in Moscow that were founded during analysis). But right now I don't have data on population and real estate prices for every home in Moscow as well as the full list of all existing Moscow venues with its coordinates. Actually the analysis of venues offered by Foursquare is not full and objective as in its venues list there are primarily venues related to food, entertainment or public transportation. There should be more places that people normally visit. And as I understood Foursquare data covers just the users of its app, that is a small percentage of population.

Conclusions

Now it's easy to identify what venues are the most popular for every district and do some urban or benchmarking studies to reveal some new business opportunities in similar by socioeconomic situation districts.

Future direction of research – to analyze data that covers more Moscow venues and places of interest, to find opportunities to connect API of some Moscow real estate web site and get more detailed information on real estate prices with precision to every house.