Capstone Project

The Battle of Neighborhoods

Investigation of Moscow regions

Moscow is one of the fastest growing European cities and the largest Russian city. Currently, the population of Moscow is more than 12 million people, and the population growth exceeds 100,000 people per year. Residents of the city and official authorities are constantly faced with problems associated with urbanization. These are environmental problems, transport, overpopulation, uneven population density, tax regulation, financing, problems of recreation and entertainment.

Introduction

One of the areas of regulation is to eliminate the asymmetry in the location of urban infrastructure centers and to promote the development of venues in certain areas where their absence is obvious. And vice versa, closing and moving venues, the number of which exceeds the necessary indicators. Significant help in managing these areas can be provided by Data Science and special methods for detecting hidden patterns. Thus, the problem is to find non-obvious patterns in the distribution of urban infrastructure centers in the context of each individual district of the city.

Business Problem

Using the Foursquare service allows you to get a list of objects for each of 146 districts of the city of Moscow. A subsequent analysis of the distribution of objects by district will allow us to find hidden dependencies. The predominance of objects belonging to a certain category over objects of other categories will allow us to build a certain rating of districts and combine them into groups. At the same time, segmentation and clustering (K-means) will allow us to quickly identify unknown patterns. A certain difficulty with this approach will lie in the process of verifying that a particular venue belongs to a specific area of the city. To solve this problem, we will need to describe the boundaries of each region in the form of a polygon. In the future, having calculated the location of the central point for each region, as well as the radius of the surrounding circle, it will be possible to filter out places that are included exclusively in a specific area. After additional transformations, an array of sites can be processed using the K-means algorithm. Then we can analyze the formed clusters.

Data descriptiion

Fortunately, we can get geo-points for the polygons of each district of the city on the website http://gis-lab.info. File format is CSV and every row is geo data for a district.

WKT,NAME,OKATO,OKTMO,NAME_AO,OKATO_AO,ABBREV_AO,TYPE_MO
"MULTIPOLYGON (((36.8031012 55.4408329,36.8031903 55.4416007,36.8035692 55.4516224,36.812528 55.4513994,36.8274471 55.4513398,36.8333688 55.4513764,36.8338034 55.4516439,36.8345763 55.4512558,36.8348594 55.4514247,36.8349932 55.4514931,36.8358013 55.4511173,36.8360591 55.4511632,36.8461554 55.4510412,36.8602864 55.4508946,36.8649423 55.4506415,36.8608407 55.4492656,36.8582649 55.4478456,36.8582898 55.447659,36.8600008 55.4466656,36.8611076 55.4473042,36.8622805 55.4467128,36.8638768 55.4472018,36.8694408 55.4489425,36.8724625 55.4502245,36.8749845 55.4513839,36.8773319 55.453135,36.8804877 55.4548177,36.8822676 55.455771,36.8833225 55.4551478,36.8837761 55.4554817,36.8846681 55.4551548,36.8855977 55.4548773,36.8863281 55.4552539,36.8952465 55.450465,36.8875942 55.4460139,36.8818703 55.4426547,36.8912361 55.437452,36.8915818 55.4377769,36.893413 55.4368788,36.8948019 55.4377928,36.8963369 55.4389852,36.8968637 55.4392912,36.8968237 55.4387451,36.8964101 55.4380589,36.8959156 55.4369562,36.8931806 55.4285387,36.8930385 55.4281378,36.892214 55.4255804,36.8920661 55.4251213,36.8921556 55.4250135,36.8929218 55.4250569,36.8939156 55.4236071,36.8938567 55.4229336,36.8931251 55.4224349,36.8916934 55.4211121,36.8919249 55.4163998,36.892068 55.4106145,36.892487 55.4102303,36.8953439 55.4078434,36.8974602 55.4060326,36.897289 55.4057843,36.8978411 55.4053155,36.9014402 55.402228,36.9016975 55.4019726,36.9016599 55.4018918,36.9037445 55.4002463,36.9042105 55.3993939,36.9042939 55.3991925,36.9043756 55.3989367,36.9046011 55.3984985,36.9049361 55.3978212,36.905001 55.3975771,36.9050214 55.3974678,36.9051339 55.3973911,36.9051533 55.3973512,36.9052789 55.3973497,36.9052676 55.3971375,36.9056722 55.3966414,36.9058245 55.392973,36.9057008 55.3921492,36.9057503 55.3910598,36.9057594 55.3908605,36.9057755 55.3906733,36.9058681 55.3896065,36.9056813 55.3894557,36.9059927 55.3892795,36.9060527 55.3891022,36.9060885 55.3889279,36.9064487 55.3884621,36.9070191 55.3877243,36.9072097 55.3875265,36.908038 55.3866655,36.9085849 55.3860978,36.9086165 55.3858614,36.9099096 55.3858468,36.9123641 55.385847,36.913446 55.3846883,36.9145318 55.3849421,36.9148241 55.3852584,36.9154068 55.385838,3...................

Sample data

In the next steps we'll load necessary data

- make some cleaning of the data
- build additional lists of geo points
- visualize preliminary results
- choose the best parameters for K-means method
- run K-means
- visualize results
- recap

Methodology

	WKT	NAME	OKATO	OKTMO	NAME_AO	OKATO_AO	ABBREV_AO	TYPE_MO	index
0	MULTIPOLYGON (((36.8031012 55.4408329,36.80319...	Киевский	45298555	45945000	Троицкий	45298000	Троицкий	Поселение	0
1	POLYGON ((37.4276499 55.7482092,37.4284863 55....	Филёвский Парк	45268595	45328000	Западный	45268000	ЗАО	Муниципальный округ	1
2	POLYGON ((36.8035692 55.4516224,36.8045117 55....	Новофёдоровское	45298567	45954000	Троицкий	45298000	Троицкий	Поселение	2
3	POLYGON ((36.9372397 55.2413907,36.9372604 55....	Роговское	45298575	45956000	Троицкий	45298000	Троицкий	Поселение	3
4	POLYGON ((37.4395575 55.6273129,37.4401803 55....	"Мосрентген"	45297568	45953000	Новомосковский	45297000	Новомосковский	Поселение	4
...	...	...	...	...	...	...	...	...	...
141	POLYGON ((37.7998089 55.7623198,37.7998143 55....	Ивановское	45263567	45306000	Восточный	45263000	ВАО	Муниципальный округ	141
142	POLYGON ((37.8360239 55.709776,37.8361995 55.7...	Косино-Ухтомский	45263573	45308000	Восточный	45263000	ВАО	Муниципальный округ	142
143	POLYGON ((37.8404157 55.7304867,37.8406349 55....	Новокосино	45263579	45310000	Восточный	45263000	ВАО	Муниципальный округ	143
144	POLYGON ((37.9061276 55.7062585,37.9070118 55....	Некрасовка	45290574	45391000	Юго-Восточный	45290000	ЮВАО	Муниципальный округ	144
145	MULTIPOLYGON (((37.290502 55.8019897,37.295422...	Кунцево	45268562	45320000	Западный	45268000	ЗАО	Муниципальный округ	145

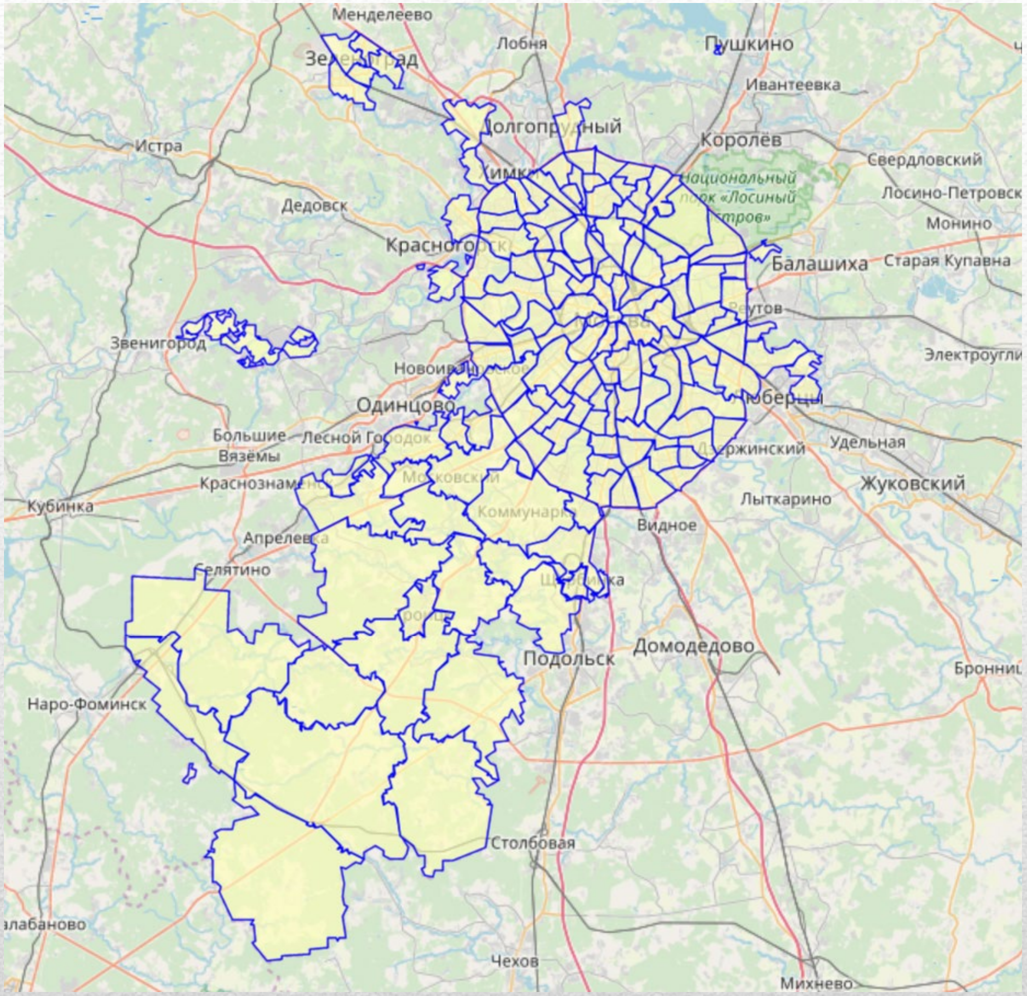
146 rows × 9 columns

Collected data

Now we have to prepare our data. For further correct calculations we have to extract several lists from the gathered table.

- List of polygonal segments which constitutes city region's borders extracted from moscow_regions.
- List of coordinates of a center for every segment. Center point is actually a geometric center of a plane figure and calculated as the arithmetic mean position of all the points. I used spherical coordinates here, and this is not entirely correct, but acceptable for our case.
- List of polygones' radiuses. It is calculated as the farthest point from the center, belonging to the polygon.

Prepare data

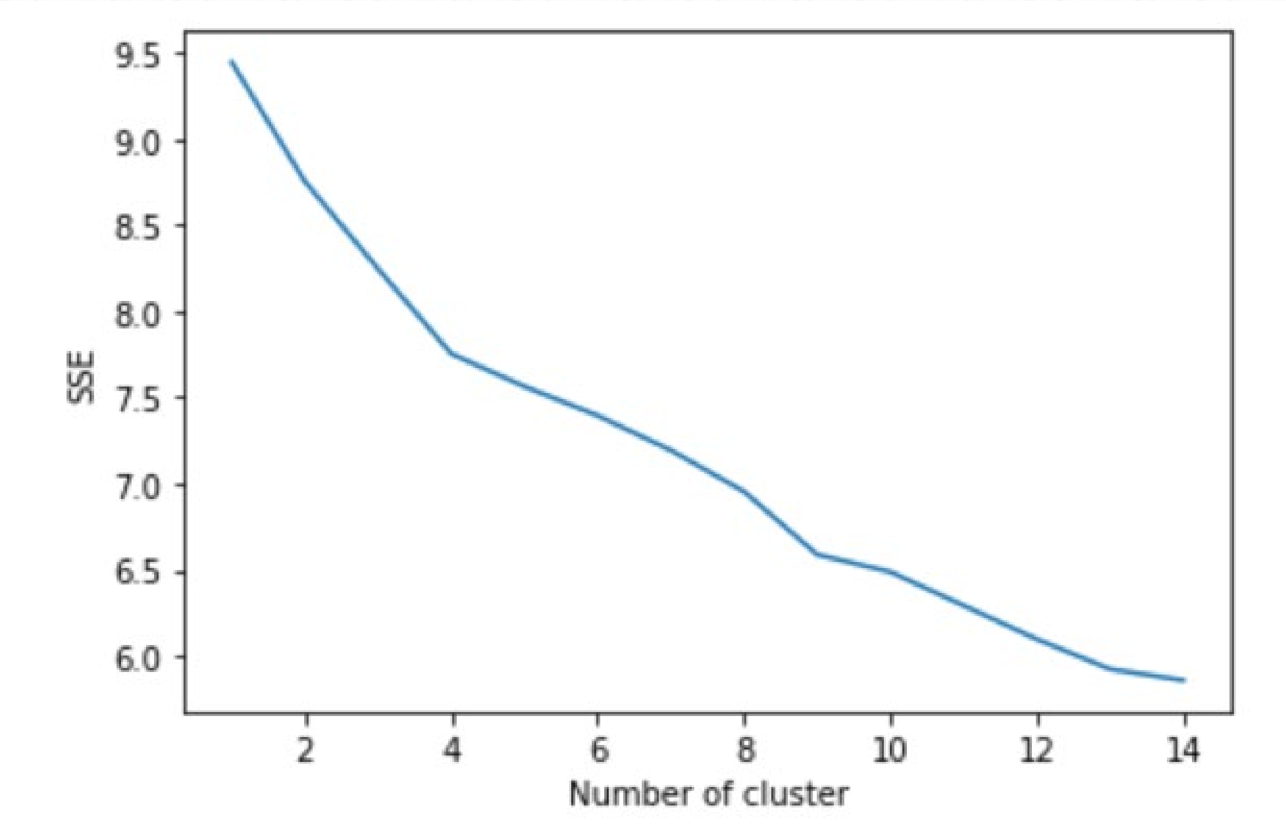


Moscow’s regions

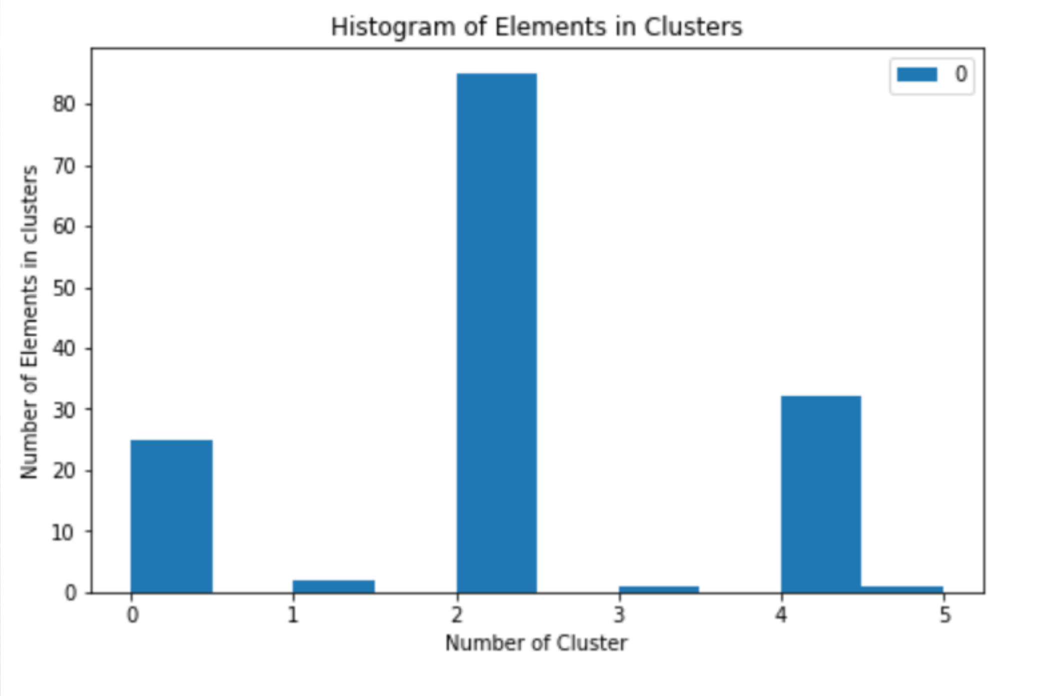
	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Region index	Venue	Venue Category	Venue Latitude	Venue Longitude
5	Kievskij	55.391999	36.907018	0.0	J/d stanzia Bekasovo-1	Train Station	55.429574	36.840377
12	Kievskij	55.320082	36.911095	0.0	Machixino Style	Castle	55.322899	36.912227
13	Filevskij Park	55.748695	37.473326	1.0	PandaPark Fili	Athletics & Sports	55.750919	37.478342
14	Filevskij Park	55.748695	37.473326	1.0	PKiO «Fili»	Park	55.747953	37.484884
16	Filevskij Park	55.748695	37.473326	1.0	Filevskaa naberejnaa	Pedestrian Plaza	55.740289	37.456460
...	...	...	...	...	...	...	...	...
13471	Kunzevo	55.785919	37.330884	145.0	Ostanovka «Rublevo»	Bus Stop	55.786702	37.355136
13473	Kunzevo	55.785919	37.330884	145.0	Asna	Pharmacy	55.782227	37.358974
13482	Kunzevo	55.808512	37.376903	145.0	Biznes poselok Rublevo-Makininskij	Hostel	55.808627	37.378484
13485	Kunzevo	55.808512	37.376903	145.0	MINI Garage	Auto Workshop	55.804565	37.375400
13486	Kunzevo	55.808512	37.376903	145.0	Appart-otel_ Rublevo-Makinino	Hotel	55.810122	37.378312

4487 rows × 8 columns

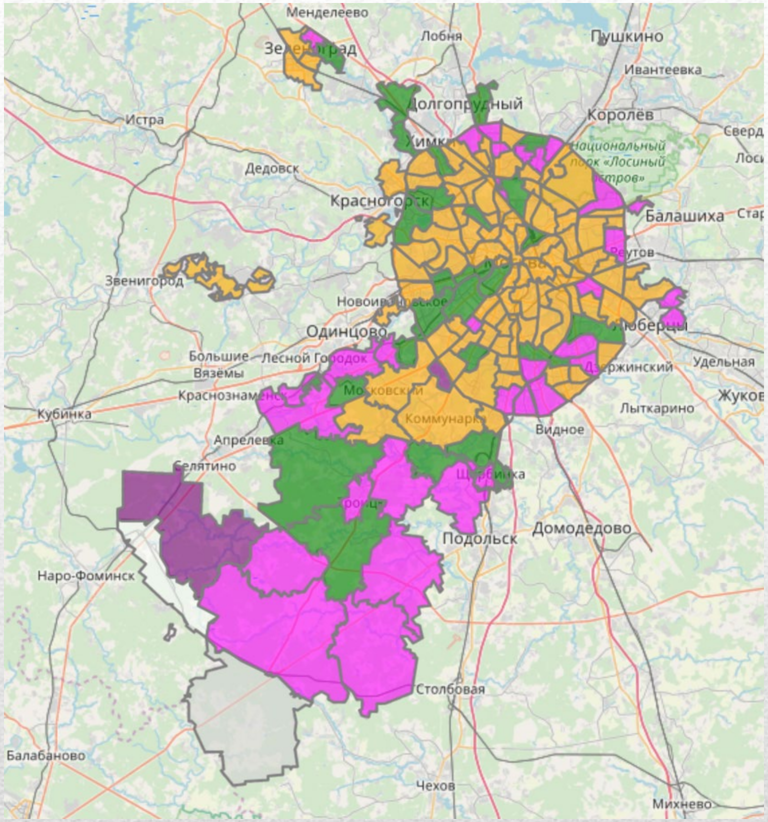
Play with Foursquare API



Finding optimal number of clusters



Clustering on six groups



Visualization of clusters

Clusters 1,3 and 5 are clear exceptions. The properties of their elements differ sharply from others, since the size of these clusters is extremely small. It makes sense to exclude them from consideration.

Clusters 0,2,4 have comparable sizes. Given the features of filling objects with clusters, the following typical names can be given.

- **Cluster 0. *Green area of parks and cafes*. This cluster is dominated by park areas, cafes and coffee shops. Other objects to a lesser extent characterize this cluster.**
- **Cluster 2. *Shops and cafes*. Park zones also prevail in this cluster, but there is a significant preponderance regarding various types of stores. Other objects to a lesser extent characterize this cluster.**
- **Cluster 4. *Shops and supermarkets*. As for cluster 2, the predominance of shopping facilities is becoming overwhelming. Other objects to a lesser extent characterize this cluster.**

Discussion

- We were able to get the division of the regions of the city of Moscow into several clusters.
- The accuracy of this method is somewhat degraded due to the lack of the number of objects returned by the Foursquare API. But this does not prevent a general idea of the properties of the obtained groupings.
- Also, some simplifications of mathematical methods introduce an error when calculating coordinates on a spherical surface, but this does not violate general principles.

Conclusion

Thanks for your time!