Searching a location for the new Italian restaurant

IBM/Coursera Applied Data Science specialization course

Alexander Bagma

July 24, 2021

Table of contents

1. Introduction

1.1 The business problem

The business problem we are solving is a search of an optimal location for a restaurant. The optimal location should be defined for *the Moscow, Russia*. Another case is that stakeholders planning to open an Italian cuisine restaurant. Also location should be not so far from the city center and in the not so crowded by other restaurants area. Stakeholders are interested in locations within 6 kilometers from the city center.

1.2 Solution planning

There are a lot of restaurants at the Moscow center, so we need to define areas with a moderate competition level. This will require us to select areas with a limited number of restaurants around. Also since we are searching a place for the Italian restaurant we should avoid placing the new restaurant near any existing Italian restaurant.
Stakeholders are interested in the location which will be as close as possible to the city center and their interest is decreasing with the distance. This requirement will surely interfere with amount of restaurants in location. Because city centers are usually concentrate all kinds of business including restaurants of all sort. But in general we should fulfill first two requirements(concurrency/no other Italian restaurants nearby) before the distance from center.
Based on these requirements we will define criteria that will help us to define the most promising locations and select the most promising.

2. Data sources

2.1 Data discussion

Based on provided requirements we can define parameters that will affect our solution:

- total number of restaurants in an area
- distance to the nearest Italian restaurant located in an area
- distance of an area center to the city center

Another task is necessity to find a definition of an area or location for which we will define parameters. We defined single area as a cell in hexagonal grid which covers the city center. Grid should cover all areas up to 6 kilometers from the center.

Centers of areas and whole areas grid can be generated programmatically. With help of geocoding libraries we are be able to define coordinates of the Moscow city center. With this coordinates we can calculate coordinates of all areas in the city grid and distance the city center. Reverse geocoding functionality will allow us to find addresses for the center of any area in the grid.

Restaurants data (including location coordinates and type) can be obtained through Foursquare API, this data then can be recalculated into amount of restaurants in each area and distance to the nearest Italian restaurant.

According to initial requirements this is all information that we need. All information available through the public sources. So there is no problems with current work process.

2.2 Areas grid generation

We generated hexagonal grid which cowered the city center and all of it's surroundings in predefined radius. Each generated point in the grid was the center of a local area. All center points calculated in order to cover all ground without free spaces and satisfy total locations size.

We defined geocoding and coordinates recalculation functions. This helped us in our work progress. For geocoding we used Open Street Map services extensions. They are covered all our requirements for geolocation with appropriate accuracy. For grid generation and distance calculation we need to convert latitude/longitude coordinates (WGS84 spherical coordinate) into UTM (Cartesian) meters coordinates. Also we need to convert calculated UTM coordinates back into latitude/longitude.

We found coordinates of the city center with help geocoding services, then converted them into UTM. This coordinates are appropriate for the areas grid generation. We've searched city center by it's simple address: "*Moscow, RU".* The geocoding functions then defined this location by coordinates with **latitude** `55.7504461` and **longitude** `37.6174943`. These coordinates recalculated to UTM coordinates gave us plain location of the city center. **X** coordinate was `37079.00` meters and **Y** was `6203013.36` meters.

Next we've defined algorithm which created a hexagonal grid in Cartesian coordinates around the provided center up to the defined grid size radius. All cells of the provided grid was covered an area defined by provided cell radius.

Stakeholders are interested in locations around city center not further than 6 kilometers from it. So the grid radius should be the same 6000 meters. Size of a single search area was defined as a circle with radius of 800 meters, so single area will cover 1,6 kilometers in diameter.

After the grid generation we must recalculate its Cartesian coordinates into latitude/longitude. By finishing this we was able to use them for further work. For a newly calculated coordinates we was able to find addresses of of each area in the grid. Addresses help us to represent target locations in the human readable form. All data was combined into one data frame for simplicity of further use with an additional identification number attached to every area centers.

In the picture below you can see generated grid of neighbor areas. Cells are partially intersecting but covering all city center without spaces. White circle defines radius of 6000 meters from the city center.

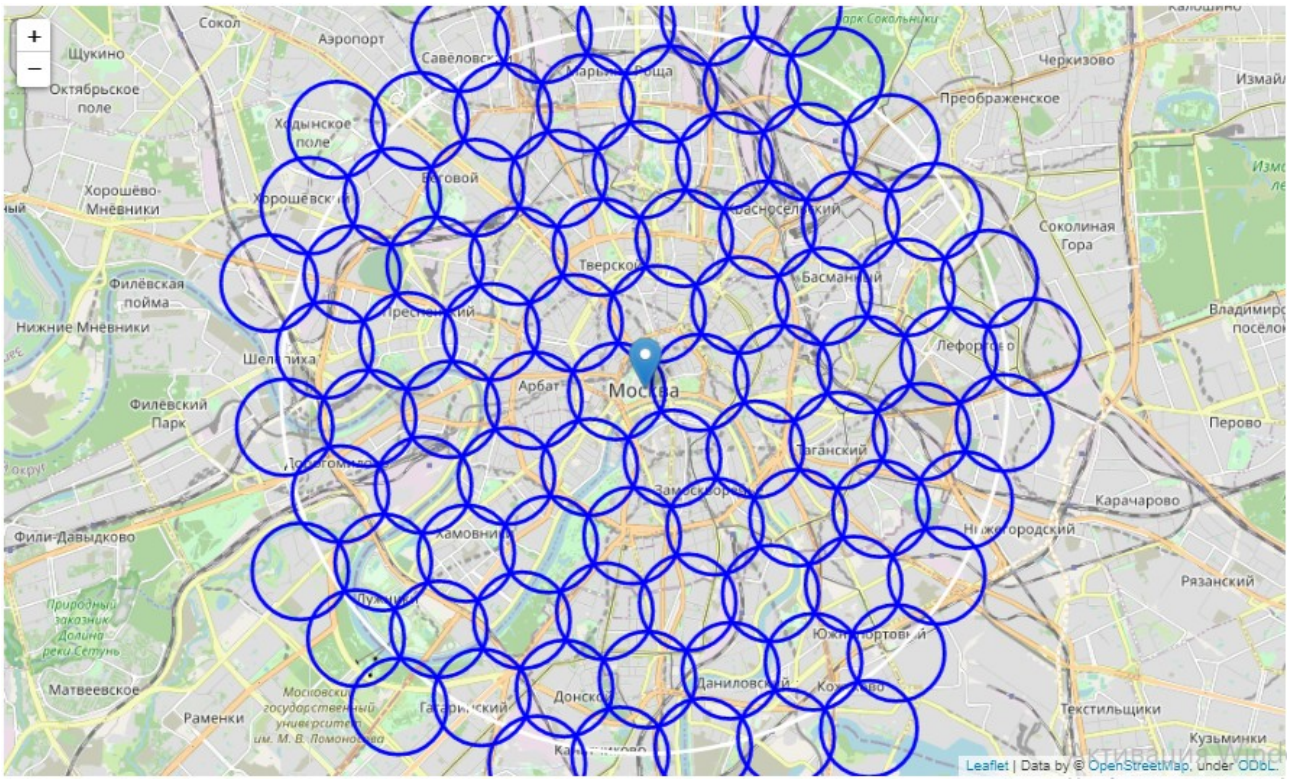

Image 1: City coverage by generated areas grid

In the table below you can see example of data generated for location grid:

id	X	Y	Distance	Latitude	Longitude	Address
0	34650.281946	6.197050e+06	6439.179620	55.694986	37.589250	Россия, Москва, улица Карьер, 2 с1
1	36035.922592	6.197050e+06	6054.120206	55.696310	37.611111	Россия, Москва, Загородное шоссе
2	37421.563238	6.197050e+06	5973.416840	55.697631	37.632974	Россия, Москва, улица Татлина
3	38807.203885	6.197050e+06	6208.948866	55.698948	37.654838	Россия, Москва, 2-й Кожуховский проезд, 29 к6
4	40192.844531	6.197050e+06	6727.583764	55.700261	37.676704	Россия, Москва, 2-й Южнопортовый проезд

It contains both Cartesian and spherical coordinates, distance from city center and address of every location. Also every location has it's own identifier which will help in further operations. All data looked normal so we proceeded in our work.

Now we can move further and retrieve an information about restaurants located in the city center.

2.3 Restaurants data

Now we need to gather an information about restaurants in the city center and theirs alignment to one of areas in the previously generated grid. We can use Foursquare API to get info on restaurants and theirs location.

We're interested in venues of 'food' category. This category contains many different types of food venues not only restaurants. So we should filter all that we can't count as a restaurant. The reason of a such behavior is that other types of food venues (like bakery or fast food) are not direct competitors.

We should include into restaurants list only venues with 'restaurant', 'diner', 'taverna' or 'steakhouse' words in category name. Also we should find all venue categories that correspond to an Italian restaurant. So we should get from 'food' categories list category with name 'italian' and all of it's subcategories.

First we should define functions to operate with categories data. We need to retrieve it, find subcategory in it, convert tree like structure of subcategories into flat list and we need a functionality to remove all unnecessary parts of received objects.

Then we should request categories list from Foursquare and find food category in it. Then we will search for all Italian restaurant categories.

The general **food category** id is **`4d4b7105d754a06374d81259`**.

Results of categories exploration you can see in the table below.

Category name	Category id
Italian Restaurant	4bf58dd8d48988d110941735
Abruzzo Restaurant	55a5a1ebe4b013909087cbb6
Agriturismo	55a5a1ebe4b013909087cb7c
Aosta Restaurant	55a5a1ebe4b013909087cba7
Basilicata Restaurant	55a5a1ebe4b013909087cba1
Calabria Restaurant	55a5a1ebe4b013909087cba4
Campanian Restaurant	55a5a1ebe4b013909087cb95
Emilia Restaurant	55a5a1ebe4b013909087cb89
Friuli Restaurant	55a5a1ebe4b013909087cb9b
Ligurian Restaurant	55a5a1ebe4b013909087cb98
Lombard Restaurant	55a5a1ebe4b013909087cbbf
Malga	55a5a1ebe4b013909087cb79

Marche Restauran	55a5a1ebe4b013909087cbb0
Molise Restaurant	55a5a1ebe4b013909087cbb3
Piadineria	55a5a1ebe4b013909087cb74
Piedmontese Restaurant	55a5a1ebe4b013909087cbaa
Puglia Restaurant	55a5a1ebe4b013909087cb83
Romagna Restaurant	55a5a1ebe4b013909087cb8c
Roman Restaurant	55a5a1ebe4b013909087cb92
Sardinian Restaurant	55a5a1ebe4b013909087cb8f
Sicilian Restaurant	55a5a1ebe4b013909087cb86
South Tyrolean Restaurant	55a5a1ebe4b013909087cbb9
Trattoria/Osteria	55a5a1ebe4b013909087cb7f
Trentino Restaurant	55a5a1ebe4b013909087cbbc
Tuscan Restaurant	55a5a1ebe4b013909087cb9e
Umbrian Restaurant	55a5a1ebe4b013909087cbc2
Veneto Restaurant	55a5a1ebe4b013909087cbad

With categories data at hands we are able to request restaurants data from Foursquare api.
We need a way to request a list venues around of the provided coordinates, we need to define if received venue is a restaurant or not. Also we need a way to define if venue category is one of required Italian categories. And we need to request restaurants list for every area in the previously generated grid.

About the grid, as we saw earlier areas are intersecting, so it's possible for one venue to be returned for two areas. To deal with this possibility we will order areas by distance from city center and will request venues for areas in that order. Also we will track all previously assigned venues and forbid them from further areas. In the end we will have no repeated venues in the list.

From requested data we created data frame for simplicity of further work with information and update it with necessary data. We defined their Cartesian coordinates for further distance calculations and we will mark Italian restaurants in the list.

You can see an example of received data in the table below.

Id	Name	Lat	Lon	Address	Categories	Area	Distance	X	Y	Italian
4d05f594dc45a0936f9cf1c6	Корчма «Тарас Бульба»	55.750644	37.610157	Моховая ул., 8, стр. 1, 119019, Москва, Россия	52e928d0bcbc57f1066b7e96	42	34	36622.236463	6.203084e+06	No
55e1cfcb498ee04c095f4c7a	Пян-се	55.752643	37.609967	ул. Воздвиженка, 4/7, Москва, Россия	4bf58dd8d48988d108941735	42	257	36634.107861	6.203308e+06	No
4c409b03d7fad13a5	Ширван	55.751243	37.607539	Староваганьковский пер.,	5293a7d53cf9994f4e043a	42	195	36465.5406	6.203169e	No

Id	Name	Lat	Lon	Address	Categories	Area	Distance	X	Y	Italian
13c06da				19, стр. 7, 111222, M...	45			21	+06	
5b3361f060255e002c1f4270	Il Pizzaiolo	55.748705	37.609088	Волхонка 6, 119019, Москва, Россия	4bf58dd8d48988d110941735	42	194	36532.278887	6.202877e+06	Yes
53b94da2498eafad5864e500	Dolmama	55.753999	37.609699	Романов пер., 2/6, стр. 13, Москва, Россия	5f2c2b7db6d05514c7044837	42	409	36633.441482	6.203460e+06	No

So we have a list of all restaurants in the city center with Cartesian and spherical coordinates, addresses and categories. Also this data contains if restaurant is Italian and number of area in the grid to which it corresponds.

This is all information we required to move further, but before let's try to look on to it and define some simple metrics like amount of Italian restaurants, average amount of restaurants in area and average amount of Italian restaurants in every area.

This information will help us to negotiate with stakeholders the definition of crowded areas and preferable distance to nearest Italian restaurant.

The total number of restaurants: **1543**
Total number of Italian restaurants: **162**
Percentage of Italian restaurants: **10.50%**
Average number of restaurants in area: **18.4**
Average number of Italian restaurants in area: **2.7**

You can see locations of all restaurants on the image below. Italian restaurants are marked with red color all other restaurants are marked with blue color.

It looks like most of the restaurants tends to align with main city roads. This can be helpful when stakeholders will try to define specific location of a restaurant.

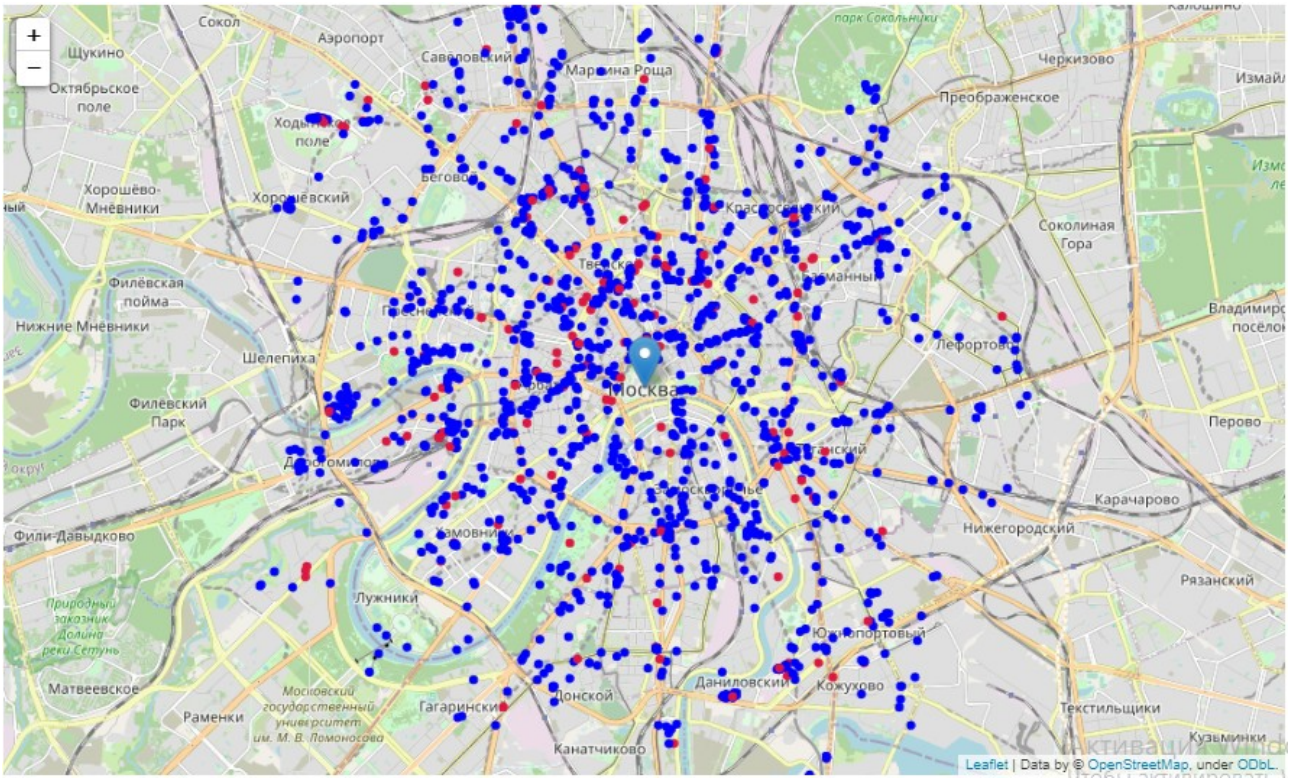

Image 2: Restaurants map

Now we have all required data, next we'll try analyze it to find required area.

3. Methodology

In this project we need to find areas of Moscow that have low restaurant density, and even less Italian restaurants. Search area are limited within 6 kilometers radius from city center.

As a first step we have collected data: location and type (category) of every restaurant within 6 kilometers distance from the city center. We have also identified Italian restaurants (according to Foursquare categorization) and defined allocation area for all of them.

Second step will be calculation and exploration of 'restaurant density' across different generated areas. For this purpose we will use heat-maps visualization to identify a few promising areas as to the center with low number of restaurants in general and focus our attention on those areas.

The third step we will create a metric of area quality. This metric will help us range areas and select the best position for a new restaurant. With such metric in hand we will be able to group areas by their quality. This will help to negate effects caused by grid cells allocation and find the target location.

4. Analysis

4.1 Restaurants allocation analysis

Let's perform analysis on the gathered data. First we can count the number of restaurants in every area. We will define new 'locations' data frame which will help up with data processing. For this we will copy cells data and enrich it with location data from restaurants. Also we will need to fill empty values for areas without restaurants. It's possible and empty values will be defined as NaN. It can ruin further calculations so let's fill them with zeros.

Also we need to calculate distance from center of every area to nearest Italian restaurant. Here you can see the result of all calculations. The table below contains only first 5 rows of calculated data but it can show you what kind of information we have.

Id	Latitude	Longitude	Address	Restaurants count	Distance to nearest Italian
0	55.694986	37.589250	Россия, Москва, улица Карьер, 2 с1	1.0	1094.168986
1	55.696310	37.611111	Россия, Москва, Загородное шоссе	1.0	921.820247
2	55.697631	37.632974	Россия, Москва, улица Татлина	6.0	477.141730
3	55.698948	37.654838	Россия, Москва, 2-й Кожуховский проезд, 29 к6	12.0	777.545625
4	55.700261	37.676704	Россия, Москва, 2-й Южнопортовый проезд	3.0	858.337066

From the obtained data we can calculated an average **distance to the nearest Italian restaurant**. And it was **650 meters**.

On the picture below you can see a heat map of all restaurants. Let's see it and try to find some additional information for our task. White circles splitting all radius into 3 parts for simplicity of navigation.

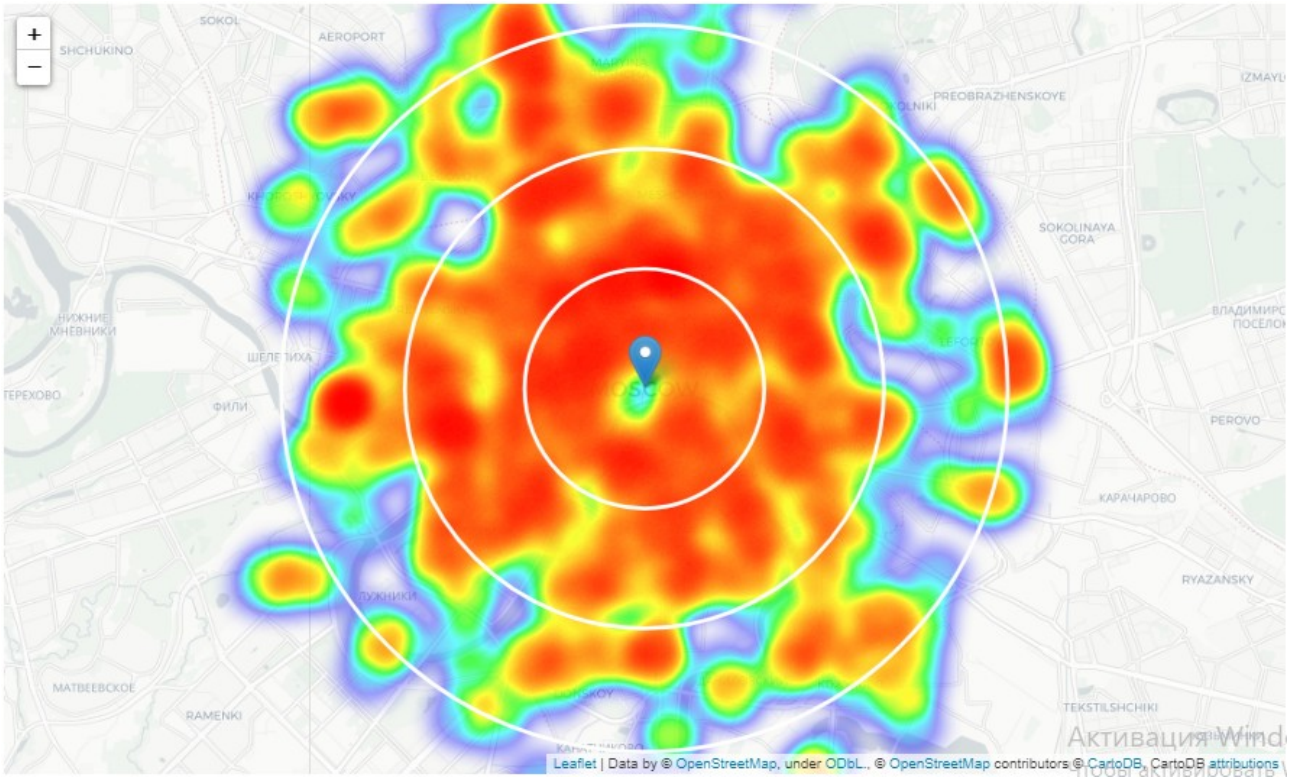

Image 3: Restaurants heat map

The same for Italian restaurants.

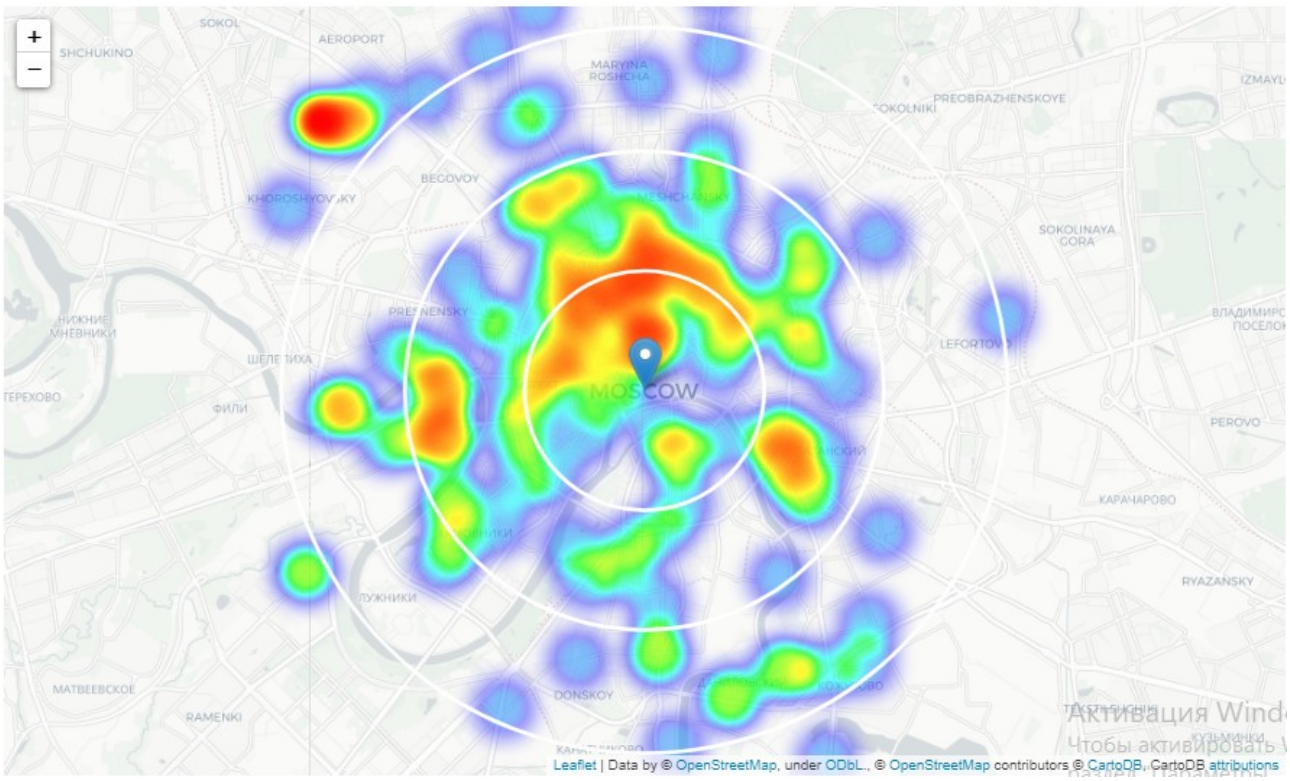

Image 4: Italian restaurants heat map

Based on what we can see center mostly crowded with restaurants and more or less free space we can find on the outer third of city center area. There are a few windows in the west and south which can be decided as an interesting regions.

Italian restaurants map shows a little amount of restaurants in third part of the radius. We have no free space in the north in the first and second parts but we can see free spaces on the west or south east.

So both maps shows an empty third part and crowded north half with some spots on west.

That's interesting and can give us a direction for further exploration. But how we defined that norther area is crowded, what was definition of 'crowded'. We saw red map parts but that's not enough. We should create some measurable way to check quality of any area. And with help of this tool we will be able to find the best location.

First we should know why this location is the best, we should be able to describe how we selected this location and we should be able to correct and calibrate this tool according the stakeholders requirements.

So we need to define a quality metric for our areas.

4.2 Quality metric

Let's define the quality measurement metric which we will apply to every area. Thus we will be able to range them and find the best one.

Our stakeholders not interested in areas further than 6 kilometers from center, so let it return zero for all areas further then this distance and let it return one at the center of the city. We can define this as a difference between our grid range and area distance to the center both divided by grid range.
We can return one for every area where amount of restaurants lower than required, for other areas we will return maximal amount or restaurants allowed in area divided by current amount of restaurant. So this metric will decrease to zero while amount of restaurants in area increasing.

The same way we can behave for the distance to the nearest Italian restaurant. We will return one for areas with distance below above required and divide current distance by maximal allowed for all others.

So we have a set from 3 separate metrics each can return a value from zero to one. We can combine them simply by multiplying and receive a new metric which will return a value from zero to one.

By applying this metric to all locations we will receive a value which will allow to range them and find the best location.

But first we need to define maximal allowed amount of restaurants in area and minimal distance to the nearest Italian restaurant. Since awerage number of restaurants in area was calculated as 18.5, stakeholder defined maximal amount of restaurants on the level of 15. And since average distance to nearest Italian restaurant was 664 meters, allowed distance was set at the 400 meters.

We calculated previously defined metric for every location, and normalized quality values for the simplicity of visualization. In the table below you can see example of calculated data.

Id	Latitude	Longitude	Address	Restaraunt count	Distance to Italian	Quality
0	55.694986	37.589250	Россия, Москва, улица Карьер, 2 с1	1.0	1094	0.000000
1	55.696310	37.611111	Россия, Москва, Загородное шоссе	1.0	922	0.000000
2	55.697631	37.632974	Россия, Москва, улица Татлина	6.0	477	0.010985
3	55.698948	37.654838	Россия, Москва, 2-й Кожуховский проезд, 29 к6	12.0	778	0.000000
4	55.700261	37.676704	Россия, Москва, 2-й Южнопортовый проезд	3.0	858	0.000000

On the images below you can see areas greed visualized according to theirs quality. The higher quality the brighter will be circle bordering the area. This helps us visually define most of the promising areas and find quality allocation patterns.

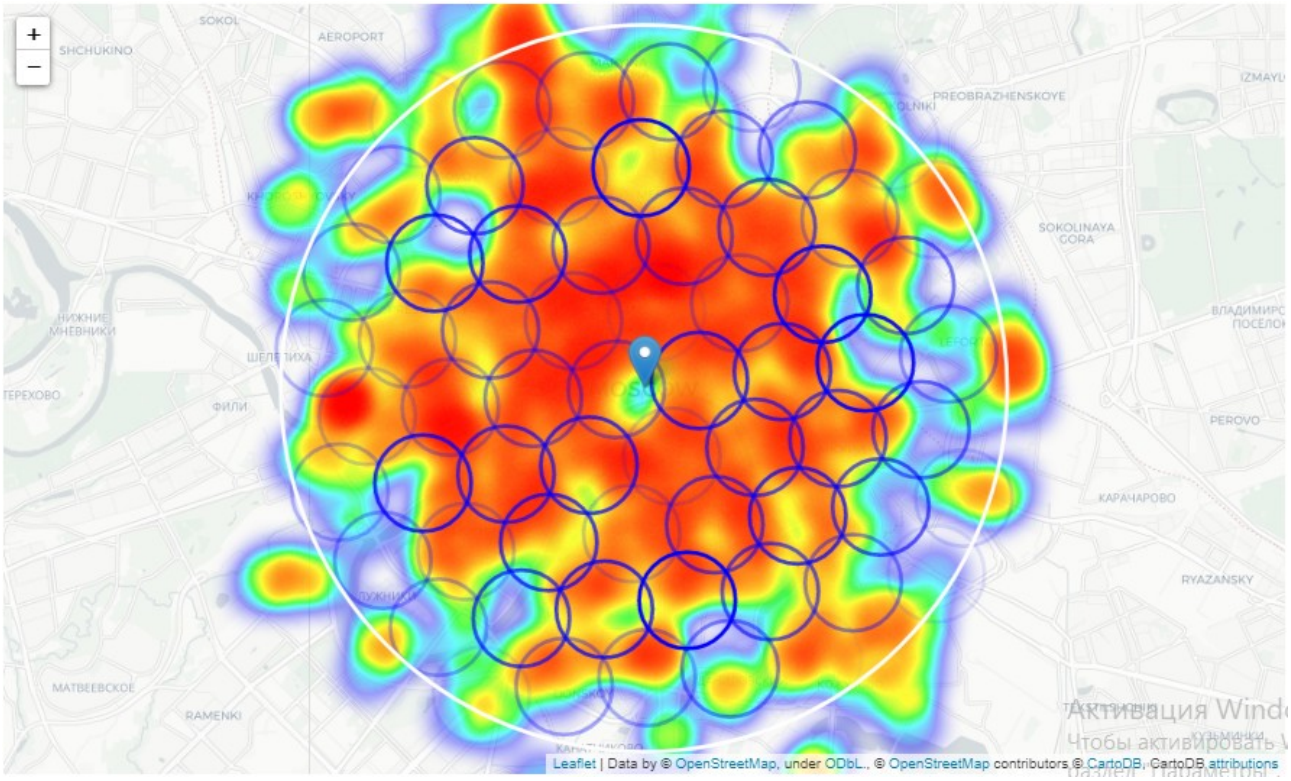

Image 5: Locations quality on restaurants map

Than we can see the same quality map for the Italian restaurants only.

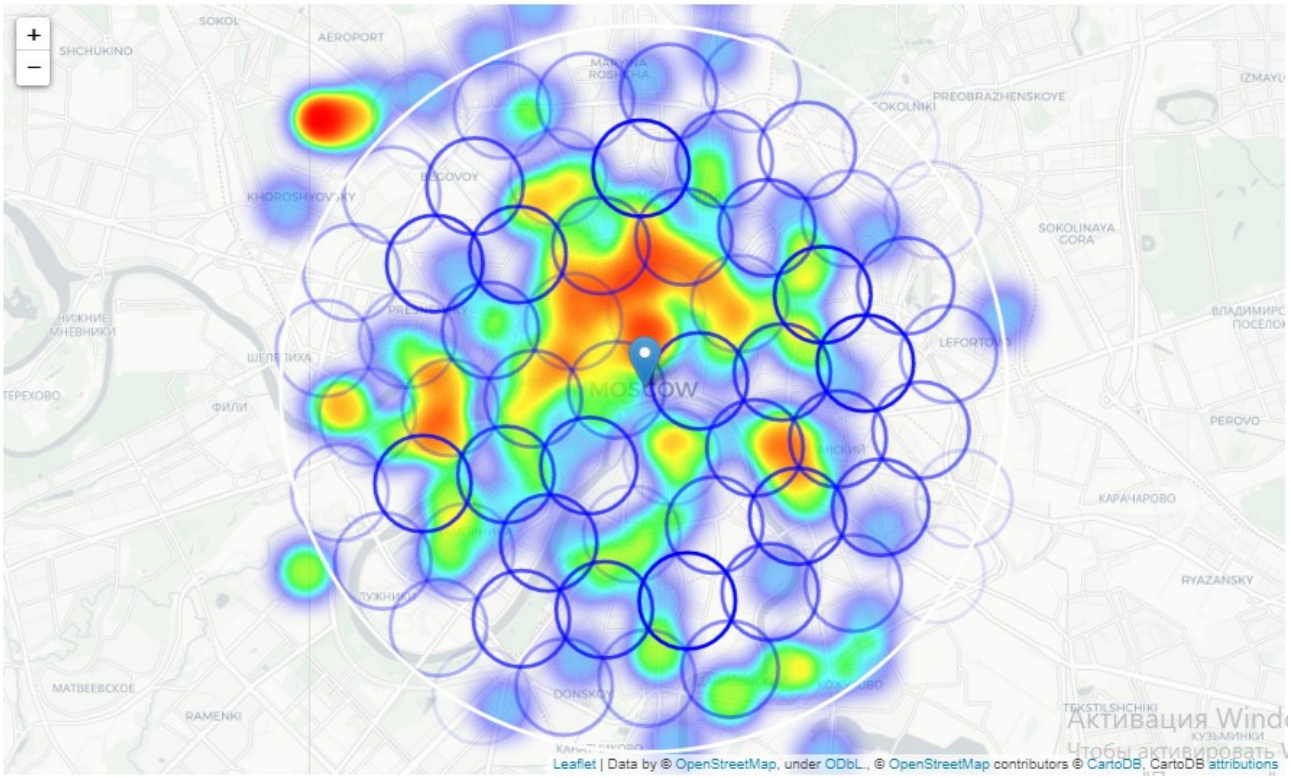

Image 6: Locations quality on Italian restaurants heat map

We are able to define some interesting locations most of them on the east but some can found on the west and even north.

Let's select **top ten best locations** and visualize them on separate map. Contrast of color will be increased so we will be able to distinct quality order with ease.

Id	Latitude	Longitude	Address	Restaurants count	Distance to nearest Italian restaurant	Quality
16	55.719018	37.628926	Россия, Москва, 3-й Павловский переулок, 2 с1	14.0	505.698918	1.000000
70	55.783176	37.616752	Россия, Москва, Большая Екатерининская улица, ...	13.0	594.933627	0.970041
45	55.754287	37.675889	Россия, Москва, Строгановский проезд	5.0	492.803866	0.950557
54	55.764427	37.664616	Россия, Москва, Малый Демидовский переулок, 3	18.0	421.446092	0.912424
30	55.736424	37.559221	Россия, Москва,	12.0	524.468106	0.831455

Id	Latitude	Longitude	Address	Restaurants count	Distance to nearest Italian restaurant	Quality
			Бережковская набережная, 16			
59	55.769062	37.562335	Россия, Москва, Ходынская улица, 3	8.0	415.079106	0.807921
43	55.751656	37.632098	Россия, Москва, улица Варварка, 6 с4	39.0	402.621934	0.805962
60	55.770393	37.584235	Россия, Москва, Тишинская площадь, 6	23.0	635.427201	0.793661
26	55.731586	37.658031	Россия, Москва, Крестьянская площадь	17.0	313.143944	0.768508
32	55.739081	37.602988	Россия, Москва, Бутиковский переулок, 3	37.0	547.396688	0.743358

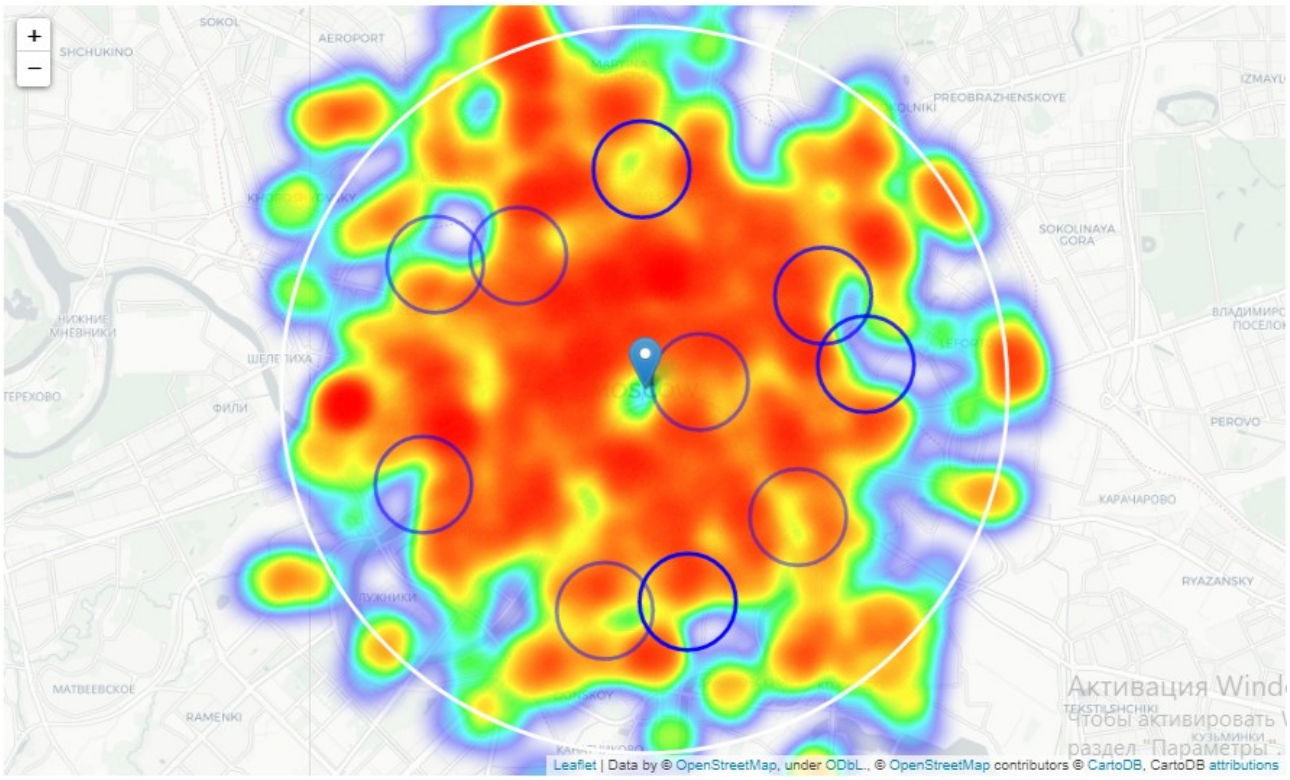

Image 7: Ten best locations

Here we can see even the promising area in the center of the city. Not the best according to color but surely promising. The best location on the south and it should the first place we will visit. Also we can see a group of good locations on the east. On the west and north we can find separate locations that are suitable for our purposes, but all of them at the second third of allowed radius.

4.3 Location groups analysis

Now we can investigate internal structure of the quality grid. We will group locations according to their coordinates and quality and try get some additional information from that.

For this task we can use K-Means clustering algorithm. It can provide us with an interesting data about the areas quality distribution. This information can help us find an additions directions for further exploration and analysis.

Amount of clusters for K-Means algorithm was 15. Results of optimization you can see in the table below. Which contains all cluster centers with their coordinates and quality. Also all centers was added to the map below with the 10 best locations.

Now we can analyze this image and check our best locations selection.

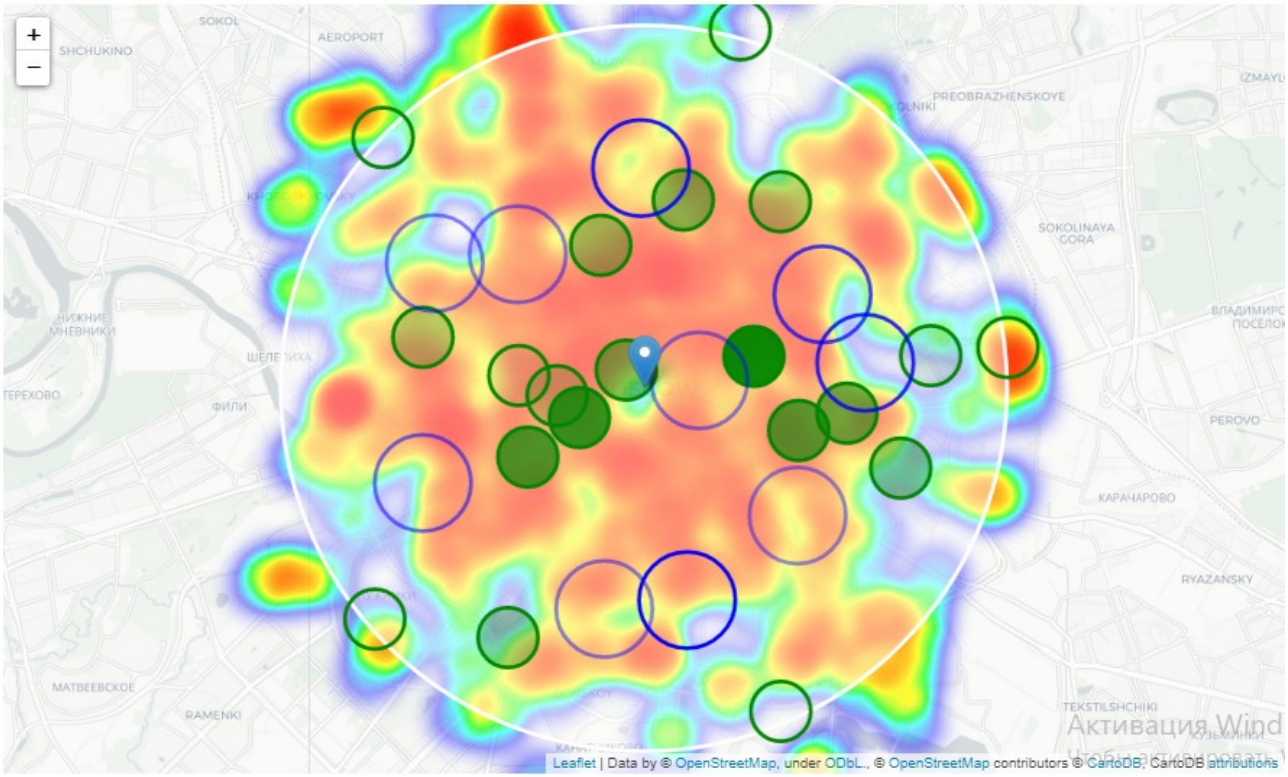

Image 8: Locations clustering

Cluster number	Latitude	Longitude	Quality
5	55.755227	37.646546	0.958256
17	55.758626	37.568597	0.811013
2	55.745405	37.648041	0.770484
19	55.740271	37.586919	0.712236
10	55.731138	37.650636	0.651728
7	55.753680	37.665790	0.568543
0	55.759955	37.590491	0.489676
15	55.777623	37.632293	0.463520
14	55.728710	37.670220	0.428157
6	55.713451	37.581682	0.342852
9	55.777621	37.632290	0.284082
13	55.757754	37.554282	0.278478
3	55.749387	37.594631	0.191436
11	55.755321	37.693155	0.126941
18	55.733351	37.597704	0.086900
12	55.788156	37.550305	0.011563
8	55.702476	37.653605	0.007605
16	55.756536	37.713515	0.000131
4	55.716182	37.546567	0.000000
1	55.803678	37.642909	0.000000

We can see that cluster with the best quality have nearly all possible value it's score 0.95. So this location is really promising it will be easy to fit a new Italian restaurant nearby. Also it's combines the group of good locations with the nearest good to the center. So east direction we can define as a main. Also according to clusters data a good location can be found at the south west direction from the center. Where as north in general have a pure quality. It will be hard to find a good place. The same could told about south east.

Several interesting locations are connected by the same cluster so it supports our choice. Now we can define addresses for the best locations of the future restaurant and transfer them to the stakeholders.

5. Results and discussion

So we finished our analysis of the locations data. Now we able to provide the best region to start ‘street’ search. In my opinion the first place that we should visit is the surroundings of best cluster center. It combines several best locations and have the perfect quality score.

Here you can see this quality cluster address:
Россия, Москва, Большой Трёхсвятительский переулок, 3

The best chances to find required location will be around there. The same is the nearest to the center good point. If stakeholders will be unable to find specific building in this location then they will be able to shift theirs attention to the east but a little further. Where they will be able to search another 2 good locations.

Another options is to rely completely on locations quality and search locations by theirs order. But even than I’d advice to look at the closest to center location as a second one. Below you can see data for the best and the closest locations.

The best location according to our research:

ID	Distance to center in meters	Latitude	Longitude	Address	Restaurants in area	Distance to nearest Italian restaurant	Quality
16	3580	55.719018	37.628926	Россия, Москва, 3-й Павловский переулок, 2 с1	14.0	506	1.0

The closest to the center good location:

ID	Distance to center in meters	Latitude	Longitude	Address	Restaurants in area	Distance to nearest Italian restaurant	Quality
43	929	55.751656	37.632098	Россия, Москва, улица Варварка, 6 с4	39 .0	402	0.81

On the picture below you can see the best and closest to the center locations combined with the best quality cluster center.

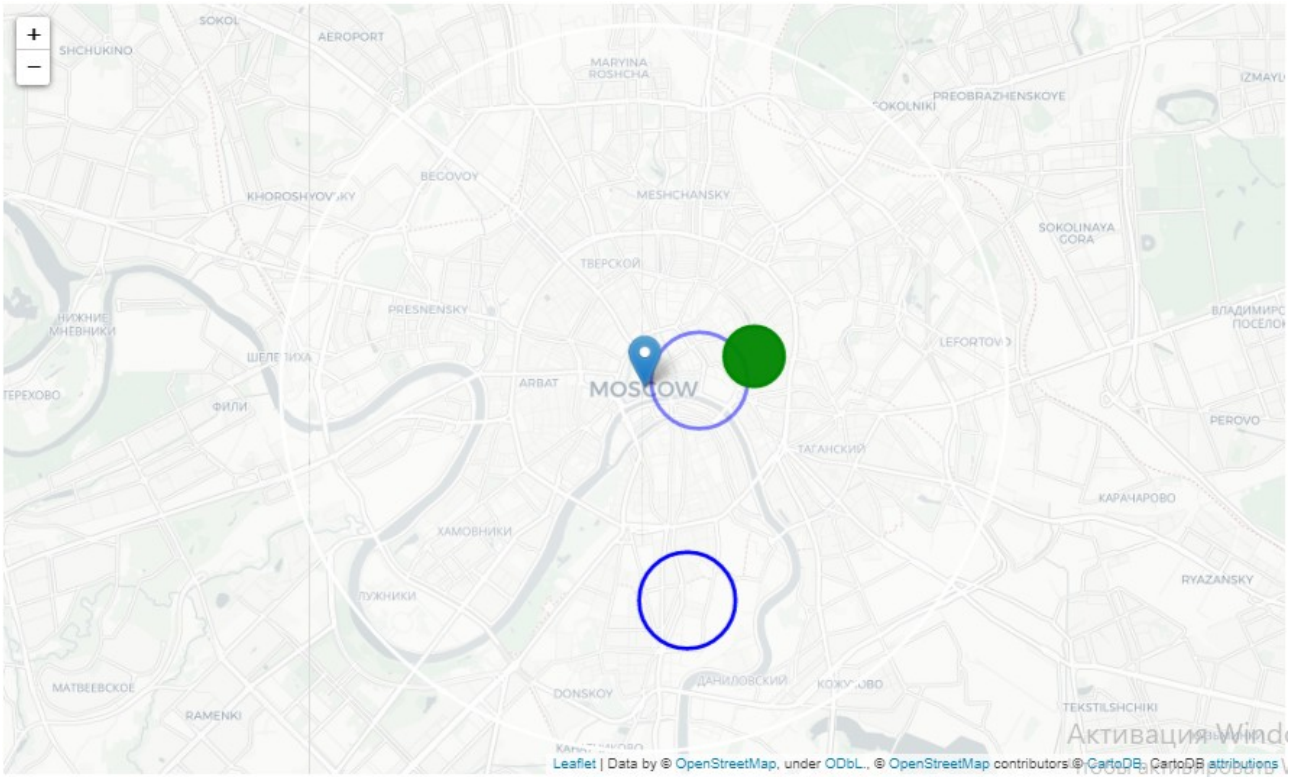

Image 9: Proposed locations map

So we can provide to our stakeholders this two most interesting locations along with list of the best ones. It will allow to start 'street' search of specific building for a new restaurant.

Our analysis shows that we are able to find a good place for an Italian restaurant even in so dense city as a Moscow, there are pockets of low restaurant density fairly close to city center. Highest concentration of restaurants was detected north and west from the city Center, and the best positions should be at the eastern part of the city. Some additional locations was identified as interesting even on the crowded northern part of the city. But we was concentrated on the most promising ones.

We have proved our choise of the best places with help of clustering which shows that both locations selected by us are connecting with a cluster with the best quality.

Result of 10 best locations with best quality are also available to stakeholders. This, of course, does not imply that those zones are actually optimal locations for a new restaurant. This analysis was to only provides info on areas close to the Moscow center but not crowded with existing restaurants (particularly Italian). It is possible that there is a good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant allocation. Also it's possible that additional factors and smaller locations grid will provide more specific results. But non the less provided locations can be considered as a starting point for more detailed analysis which could result with more detailed specifications for an analysis or allocation of a new restaurant.

6. Conclusion

Purpose of this project was to identify locations in Moscow close to center with low number of restaurants (especially Italian) in order to aid stakeholders in the search for optimal location for a new Italian restaurant. By processing Foursquare data we was able to identify areas of interest. We found some promising areas which satisfy all requirements. Clustering algorithms prove our locations selection. And we were able provide addresses of the most interesting locations to the stakeholders. Also was found additional good locations that was indistinguishable with simple eye. This was found with help of simple and configurable metric which could be discussed and improved if necessary.

Final decisions on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended area.