An Investigation of possible locations for my future apartment based on specific requirements

Sergey Kosenko

January 22, 2021

Table of contents

1 INTRODUCTION

1.1 Background

This year a several important events await me:

1. Buy a new apartment

Although I live in a province city (Samara, Russia). It is still quite a big city for us (around 1.156 million citizens in it) so it is the problem for me as I do not know exactly in which district, I should buy a new apartment for my family. As the same time, buying apartment is a quite expensive purchase, so my choice should be conscious.

2. My child must go to school at the first grade

In Russia, a child has to go at school for 8 or 10 years depending on education. This is quite a long time, so apartments should be selected the way that it would be convenient to go from apartments to school on daily basis during 8-10 years as for my child as for me as I have to accompany him (especially on the first grades).

1.2 Problem

So, I have decided to develop a solution which can help me define areas in my city, which I have to consider when buying an apartment. Additionally, I setup some requirements which should be completed during an implementation of this solution:

1. Desired distance from apartments to the nearest underground station
It should be **no more than 1500 meters**. As it is the quickest way to go around the city (except for a personal car)

2. Desired distance from apartments in question to the nearest school:
It should be **no more than 1000 meters**. As it is easier (less time for movements) and safer (faster to go to home) to my child go to at school himself.

3. ***Interesting places (for rest, art and sport) should be nearby of possible home locations****.*
If no interesting places will be nearby of possible home locations it does not make sense to buy apartments in that place as it will be very boring life

It's quite important problem for me:

1. The good education defines the future life and success of children.
2. Apartments are expensive and it's absolutely important to make a right choice and buy the best appropriate apartments. So, I am absolutely interested in receiving good results during the project

2 DATA ACQUISITION AND CLEANING

For my analysis I will use the following datasets:

1. **A list of Samara's school published on an official government site:**
   https://samadm.ru/city_life/obrazovanie/shkoly/the-list-of-schools/

   An example of data:
   Each school in the list have the following data: id, district, Institution, Address, Phones, Site

Список школ

id	District	Institution	Address	Phones	Site
Список школ					
№ п/п	Район	Наименование образовательного учреждения	Юридический адрес	Телефоны	Адрес сайта
1	Железнодорожный	муниципальное бюджетное общеобразовательное учреждение "Лицей "Классический" городского округа Самара	443030, г. Самара, ул. Владимирская, 31а	241-35-32; 241-82-22	http://www.classic-licey-samara.ru/
2	Железнодорожный	муниципальное бюджетное общеобразовательное учреждение "Школа № 18" городского округа Самара	443017, г. Самара, ул. Структурная, 48	372-48-68	http://school18.edu.ru/

   Necessary Steps:
   1. The dataset is provided on the external site; I have to parse the page and download only the table from the page into my Jupyter notebook.
   2. The dataset provided in Russian, but the Coursera course is intended for English-learners, I will use a special Python library for translating data from the dataset in English
   3. Some of the schools have a few branches: it can be seen in the **Address** column where addresses are separated by ';' sign:

77	Ленинский	муниципальное бюджетное общеобразовательное учреждение "Самарская Вальдорфская школа" городского округа Самара (с дошкольным отделением)	443041 г. Самара, ул. Буянова, 105; 443001, ул. Пушкина, 284; 443030, ул. Спортивная, 23	333-30-97; 337-60-57; 270-45-95	http://www.waldorf-samara.ru/

   we have to separate each branch in a specific row (as each of the school should be considered in the investigation).

   4. As the list of schools do not provide us with school coordinates, I should use a geocoding library in order to receive it coordinates
   5. Some of the columns (Phones, Site) are not useful for my investigation and can be safely excluded on the step.

2. **A manually prepared list of metro (underground) stations:**
   https://www.dropbox.com/s/9in85psy0tqb266/samara_underground.csv?dl=1

   As Samara have only 10 underground stations, and no aggregated information was found about locations about the stations it was decided to create the dataset manually and upload it into cloud storage for the further uploading.

3. Data from **Foursquare API** for finding venues nearby possible home locations

School dataset

After uploading, cleaning, translating and geocoding my dataset into Jupyter notebook I have received the following dataframe:

	id	District	Institution	Address	In-depth study	Pre-school education	schoolLat	schoolLong
0	0	Railway	MBOU Lyceum Classic	443030, Samara, st. Vladimirskaya, 31a	0	0	53.19387	50.13831
1	1	Railway	MBOU School № 18	443017, Samara, st. Structural, 48	0	0	53.17531	50.17021
2	2	Railway	MBOU School № 37	443013, Samara, st. Tukhachevsky, 224	0	0	53.19948	50.16072
3	3	Railway	MBOU School number 40 named after twice Hero of the Soviet Union Marshal A.M. Vasilevsky	443030, r. Samara, st. Novo-Uritskaya, 1	0	0	53.18540	50.14665
4	4	Railway	MBOU School № 42	443030, Samara, st. Uritskogo, 1	1	0	53.19040	50.12738

As you might see from the initial data frame, two columns were also elaborated from the school naming: **In-depth study** and Pre-school education which also will be used for the further analysis, as it is good useful parameters for schools.

		id	District	Institution	Address
Pre-school education	**In-depth study**				
0	**0**	108	108	108	108
	1	21	21	21	21
1	**0**	23	23	23	23
	1	7	7	7	7

Our dataset has the following shape:

```
df_schools_samara.shape
```

```
(167, 12)
```

For the simplicity the dataset was also saved into cloud storage and will be uploaded from there that do not repeat the same steps again.

Metro / Underground dataset

The dataset includes information about Samara's metro stations with their coordinates

	id	stationName	latitude	longitude
0	0	Alabinskaya	53.209704	50.132766
1	1	Rossiyskaya	53.212176	50.149211
2	2	Moskovskaya	53.202869	50.159973

3 METHODOLOGY SECTION

Defining possible locations for buying apartments

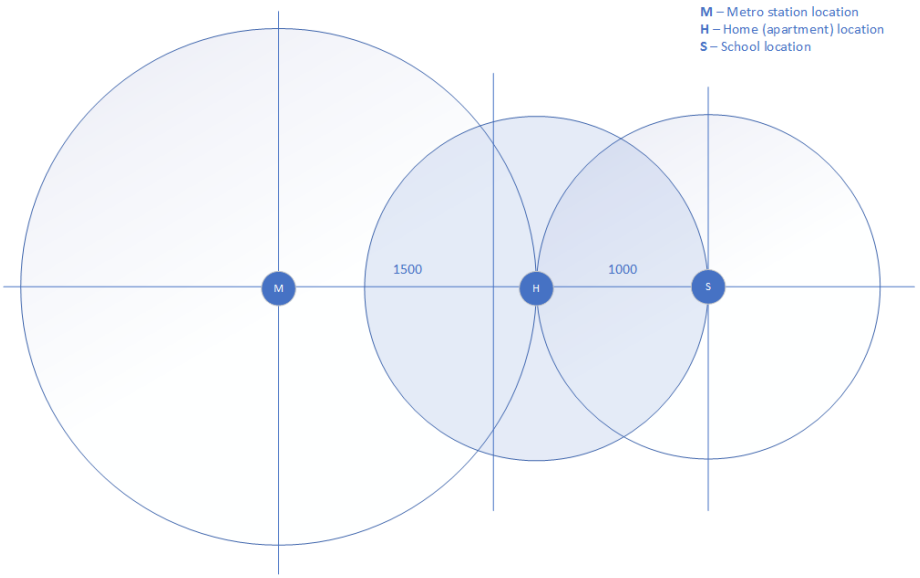
On the problem section it was mentioned that I have requirements for possible locations of new apartments regarding distances from Metro to Home and from School to Home. We will create variables describing the requirements.

```
#requirements
# desired distance from a house in question to any underground station (in meters):
house_underground_distance = 1500

# desired distance from a house in question to any school (in meters):
house_school_distance = 1000

#so I will also introduce another variable which will define the max distance between Metro and School allowed
max_available_distance = house_underground_distance + house_school_distance
```

Based on the requirements we can also define a number of coordinates for home locations (for each school) which could be a center point for area of possible home locations which I have to consider. Taken into account that the worst case will be when Home is located on 1500 meters from Metro and 1000 meters from School, we can conclude that the center point of the Home area should be on the line between (M)etro and (S)chool. Knowing the locations of Metro station and Schools we can calculate the center points of Home areas (with radius 1000 meters) for each of the school:

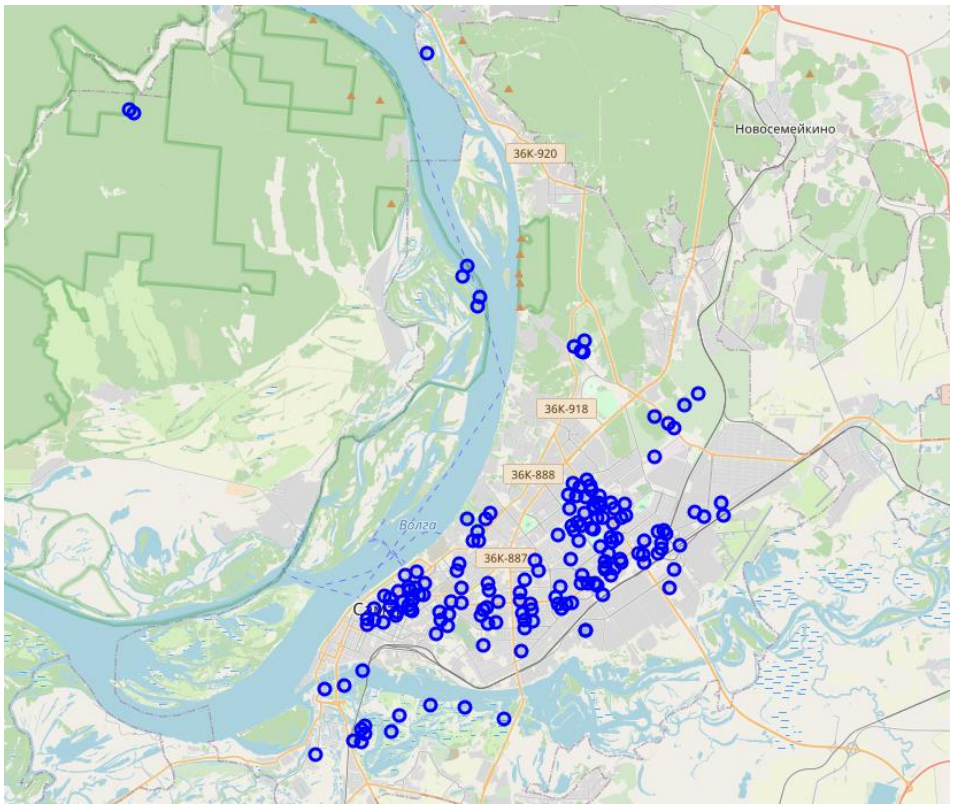


This way for the school dataset were found the new information:

1. The nearest metro station for each school
2. The distance from Metro station to each school
3. The center points (H) for areas of possible home locations around each school
4. The distance between a center point (H) of possible home locations to each school

	id	District	Institution	In-depth study	Pre-school education	schoolLat	schoolLong	nearestStation	station_school_Distance	homeLat	homeLong	home_school_distance
0	0	Railway	MBOU Lyceum Classic	0	0	53.19387	50.13831	Moskovskaya	1.76	53.197470	50.146975	0.70
1	1	Railway	MBOU School № 18	0	0	53.17531	50.17021	Gagarinskaya	2.80	53.185265	50.172923	1.12
2	2	Railway	MBOU School № 37	0	0	53.19948	50.16072	Moskovskaya	0.38	53.200836	50.160421	0.15

We can visualize all home points which we received during the operation:



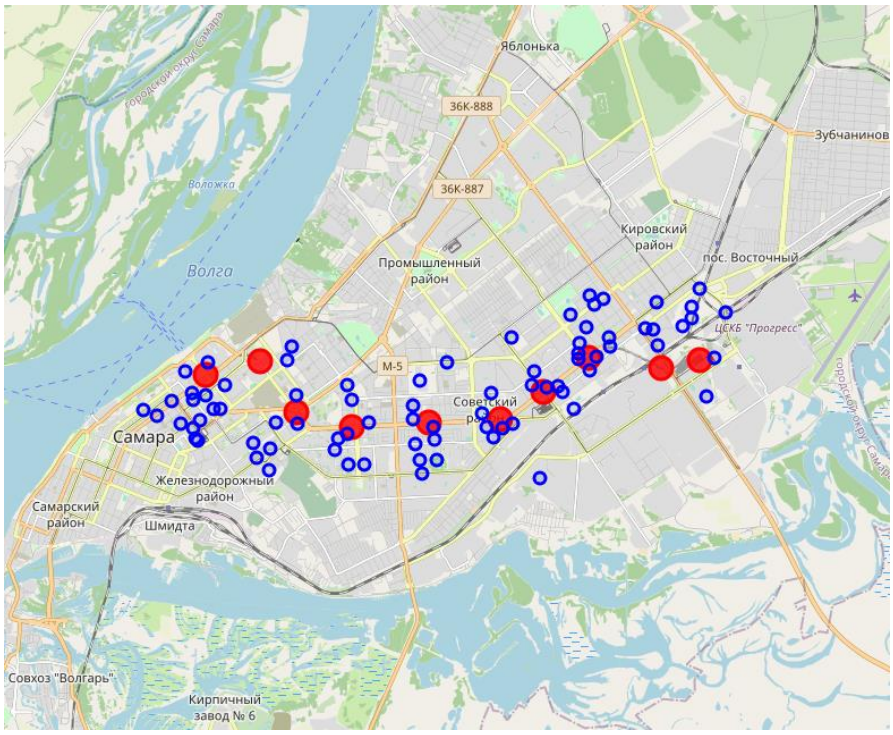
Now we can filter out the schools which do not meet our requirements:

This operation leaves only 80 out of 167 records in our dataframe we had before:

```
suitable_schools.shape
```

```
(80, 12)
```

If we visualize the Home points (blue circles) and metro stations (red circles) we can clearly see that all the points are located nearby metro stations.



After all manipulations with geocoding, calculations and filtering we receive the following dataset:

	id	District	Institution	In-depth study	Pre-school education	nearestStation	station_school_Distance	homeLat	homeLong	home_school_distance
0	0	Railway	MBOU Lyceum Classic	0	0	Moskovskaya	1.76	53.197470	50.146975	0.70
1	1	Railway	MBOU School № 37	0	0	Moskovskaya	0.38	53.200836	50.160421	0.15
2	2	Railway	MBOU School number 40 named after twice Hero of the Soviet Union Marshal A.M. Vasilevsky	0	0	Moskovskaya	2.14	53.192388	50.151979	0.85

Interesting places

The third part of my investigation will be an analysis of venues which are located nearby of defined home areas. If there is no interesting places around the places – then no need to consider them for buying apartments as it will be very boring.

For these purposes we will use Foursquare API which will allow us to find venues nearby of possible home locations. We will analyze venues not further than **1000** meters from center points of home locations.

Foursquare API returned the dataset with **3336 records in 159 unique categories**. It's a significant amount of venues, but not some of the categories are useful or interesting for daily life (like 'Auto, 'Factory' etc). So we will filter not interesting categories from the dataset.

Non interesting categories for us:

'ATM|Auto|Bar|Breakfast|Boutique|Burger|Bus|Cafe|Café|Car|Diner|Factory|Gastropub|Hotel|Intersection|Lounge|Market|Pharmacy|Place|Pub|Service|Shop|Station|Store|Supermarket|Restaurant'

After applying filtering, our dataframe includes only 898 records in 61 unique categories.

Now, if we count a number of venues per each location, we will discover that some of them almost do not have venues:

```
ss_venues.groupby('Institution').count().head(5)
```

Institution	homeCenterLatitude	homeCenterLongitude	Venue	Venue Latitude	Venue Longitude	Venue Category
MAOU Samara Lyceum of Information Technologies	33	33	33	33	33	33
MAOU Samara Medical and Technical Lyceum	29	29	29	29	29	29
MBOU Classical gymnasium number 54 Sunday	22	22	22	22	22	22
MBOU Evening school number 8	2	2	2	2	2	2
MBOU Gymnasium Perspective	7	7	7	7	7	7

So, we will exclude places for home locations, where were found less than **5 nearby venues**. As these locations will be considered places without interesting places:

Institution	homeCenterLatitude	homeCenterLongitude	Venue	Venue Latitude	Venue Longitude	Venue Category
MAOU Samara Lyceum of Information Technologies	33	33	33	33	33	33
MAOU Samara Medical and Technical Lyceum	29	29	29	29	29	29
MBOU Classical gymnasium number 54 Sunday	22	22	22	22	22	22
MBOU Gymnasium Perspective	7	7	7	7	7	7
MBOU Gymnasium № 11	23	23	23	23	23	23

After filtering possible locations, we have 60 places (80 – 20 places) for further consideration.

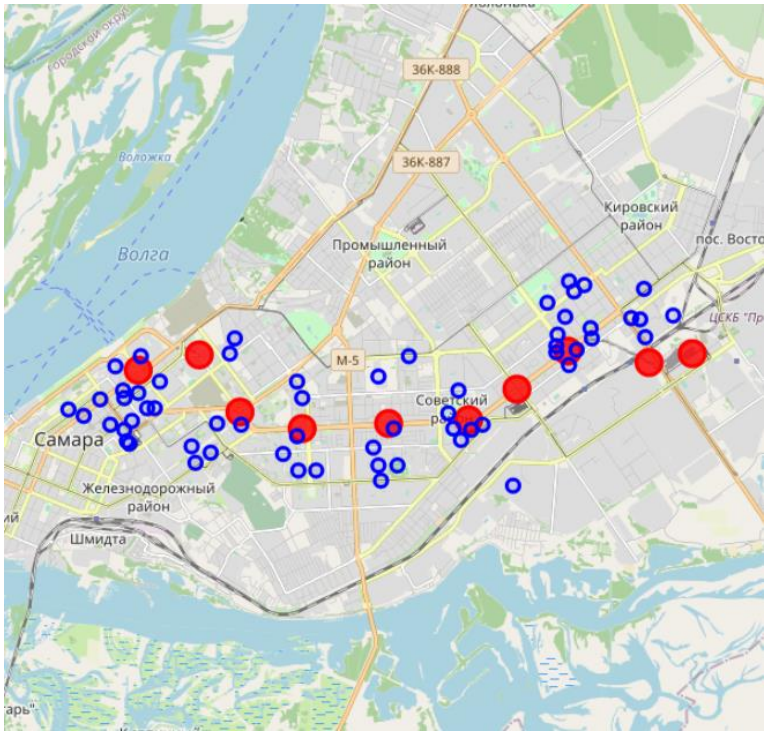
```
#let's find a list of boring places:
counts = ss_venues['Institution'].value_counts()
boring_places = ss_venues[~ss_venues['Institution'].isin(counts[counts > 5].index)]['Institution'].unique()
#convert resuling ndarray into a list
boring_places = boring_places.tolist()

print(suitable_schools.shape)
suitable_schools = suitable_schools[~suitable_schools['Institution'].isin(boring_places)]
print(suitable_schools.shape)
```

```
(80, 10)
(60, 10)
```

After applying filtering on this step for venues, our dataframe includes **only 857 records in 60 unique categories** which is only 25.7% of the initial venues dataframe.

And center points of locations which will be interested for us is now the following:



4 RESULTS

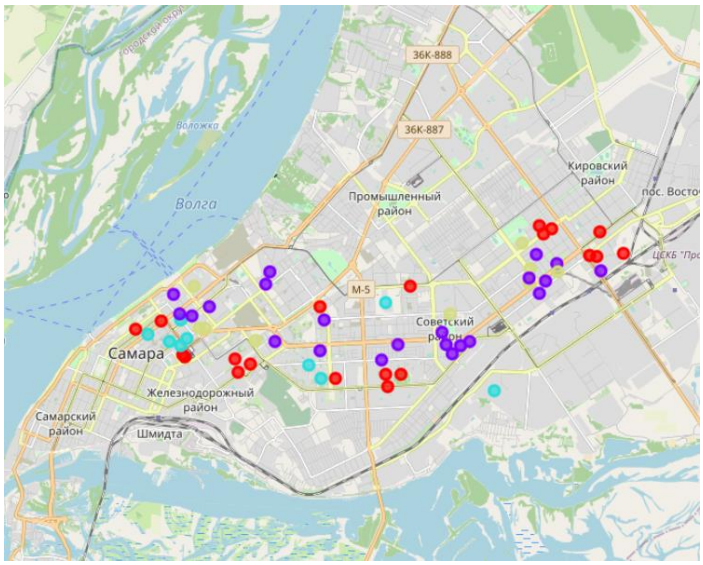
After we filtered possible home locations which located in boring places, we now can make a cluster analysis in order to find similar areas for searching the best location for buying an apartment.

After the completion of clustering analysis, we received the following clusters which we have to consider for buying apartments.

We receive the following distribution of locations by clusters:

Cluster Labels	
0	20
1	22
2	9
3	9

And the following visualization on a map:

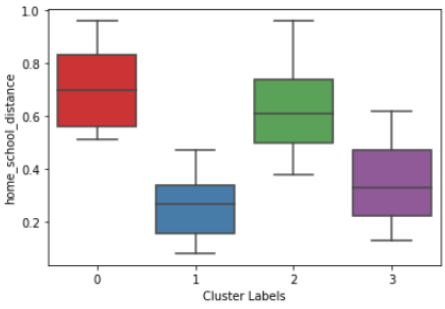


Let’s look at the data of resulting clusters:

Mean values of each clusters:

Cluster Labels	homeLong	homeLat	In-depth study	Pre-school education	station_school_Distance	home_school_distance
0	50.196885	53.206724	0.150000	0.000000	1.744000	0.697000
1	50.198840	53.206816	0.000000	0.090909	0.634545	0.253636
2	50.156770	53.199926	0.111111	1.000000	1.593333	0.638889
3	50.196575	53.210068	1.000000	0.333333	0.864444	0.345556

Distribution of distances between possible home locations in clusters:



School distribution by clusters based on 'In-depth study' and ‘Pre-school education’ features:

Cluster Labels	In-depth study	Institution
0	**0**	17
	1	3
1	**0**	22
2	**0**	8
	1	1
3	**1**	9

Cluster Labels	Pre-school education	Institution
0	**0**	20
1	**0**	20
	1	2
2	**1**	9
3	**0**	6
	1	3

5 CONCLUSIONS

Analyzing data from each cluster, we can make the following descriptions of the received clusters:

Cluster	Number of places	Description
0	20	The remotest schools from home (mean: 0.697 km) The remotest schools from metro stations (mean: 1.744 km) No pre-school education (0 vs 20 schools) Almost no In-depth education (3 vs 17 schools)
1	22	The closest schools from home (mean: 0.253 km) The closest schools from metro stations (mean: 0.634 km) Almost no pre-school education (2 vs 20 schools) No In-depth education (0 vs 22 schools)
2	9	The almost most remote schools from home (mean: 0.639 km) The almost most remote schools from metro stations (mean: 1.593 km) With pre-school education (9 vs 0 schools) Almost no In-depth education (1 vs 8 schools)
3	9	The almost closest schools from home (mean: 0.346 km) The almost closest schools from metro stations (mean: 0.864 km) Some schools with pre-school education (3 schools have vs 6) With In-depth education (9 vs 0 schools)

Based on the resulting description, I would exclude from the further consideration:

Cluster 0 – as it's the possible home locations with remotest schools and without Pre-school and In-depth education

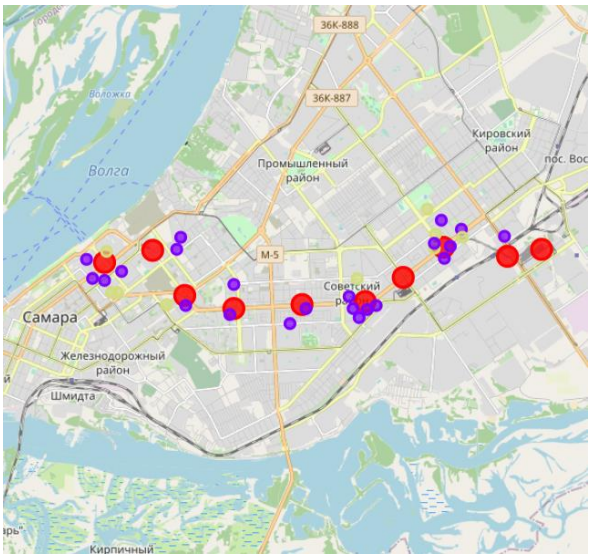
Cluster 2 – as it's also locations with almost remotest schools and no In-depth education. Although the school have pre-school educations, but due to Covid-19 restrictions it is not an available option for the schools

I would leave for the further consideration:

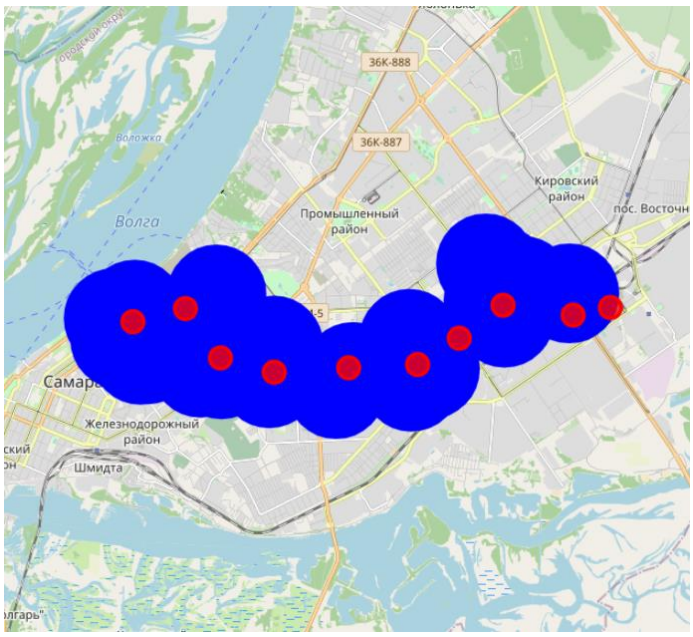
Cluster 1 – as it's the closest schools. At the same time the schools do not have pre-school or In-depth education.

Cluster 3 - I will definitely leave the cluster as the locations with the school located closely and all the school have In-depth education, which is definitely an important point for a good education.

After excluding data from Cluster 0 and 2 we will receive the following map with the possible center home locations:



And let's transform the center points of possible home locations into areas with the radius = 1000 meters (the max allowed distance from home to school) which will define areas for my further searches of apartments (where red circles are metro stations):



6 DISCUSSION

After all our manipulations we have reduced a search area from 167 records to only 31 records (only 18.5%). And although it could significantly simplify further searches of new apartments, I see further steps for improving the results with a future development of the approach.

I would select the feature directions for further reducing and improvement received data:

1. Every school graduate of Russian schools pass exams (Unified State Exams). They are the same for all schools of our country. Aggregated results of these exams should be published on sites of schools. We can use the results of the exams as a criterion for selecting schools – the higher results – the better. It's a good feature for selecting schools
2. In different districts of Samara city are different crime rates. We can use the criminal stats by districts in order to define the safe areas in Samara as a criterion for buying apartments.
3. We have areas on the map which we have to consider for buying apartments. Now we can make a search of apartment sales advertisements on Internet.