Location for a new restaurant in the Moscow

Recommender system

Alexander Popov

08.05.2020

1. Introduction

1.1 Background

The choice of location for opening a restaurant is very important. Along with a good menu and a team of professionals, a well-chosen location is one of the main factors for the success of a new restaurant. When choosing a place, you need to pay attention to such parameters as demographics, access, location of other restaurants nearby, crime rates, environmental situation, and many others.

1.2 Problem

There are about 14 thousand different food service establishments in Moscow. The question posed is which of metro stations areas are most suitable for opening a fusion restaurant, based on the above selection factors. However, the price segment of the restaurant is not defined, so it is impossible to find the target demographic.

1.3 Interest

Basically, this project will be addressed to those who want to open a restaurant in Moscow. The results of the project are not completely accurate, but they open up opportunities for further analysis and allow narrow the search area.

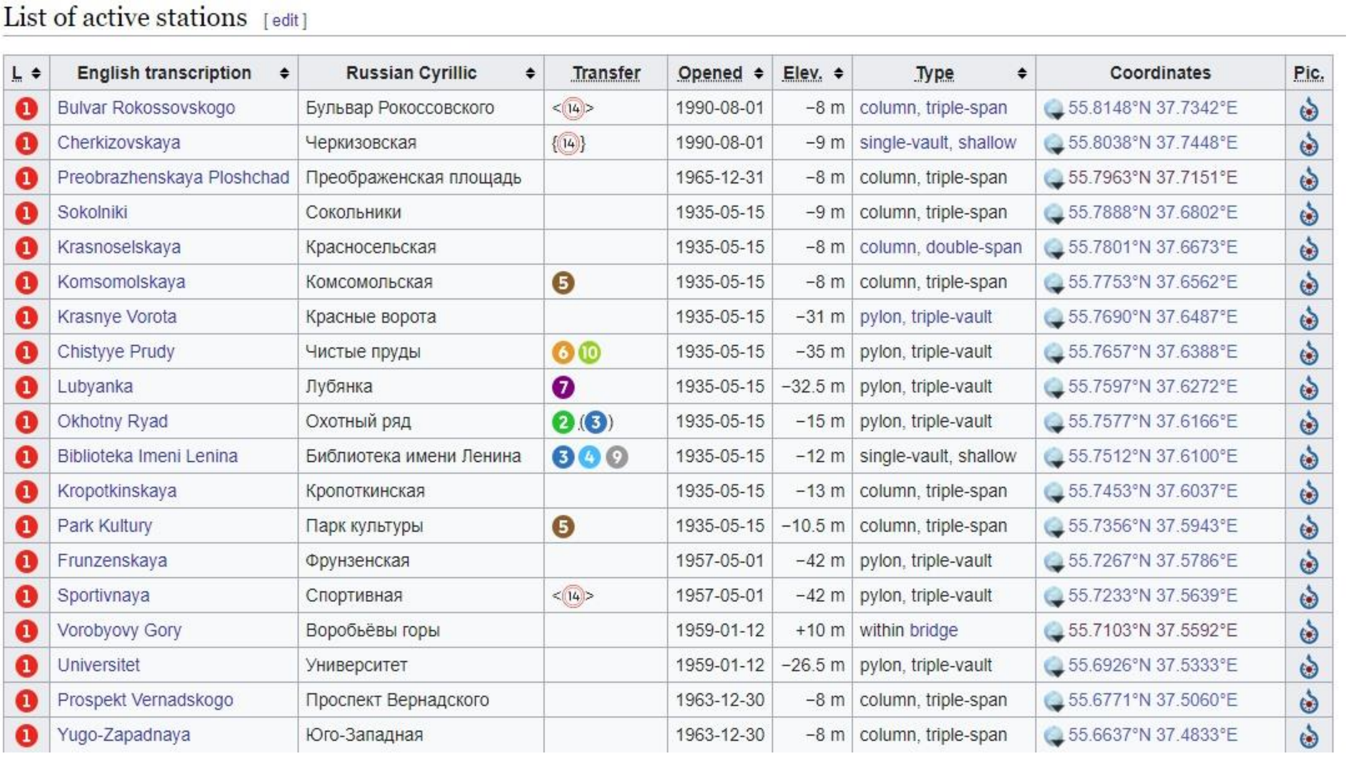
2. Data acquisition and cleaning

2.1 Data sources

Unfortunately, I couldn't find any ready to use datasets, so I used web scraping techniques. Also, no data about the crime rating was found**.**

First of all, need information about metro stations, which can be found in table on the Wikipedia page. On this page, there are a variety of data for all stations. For example, for Sokolniki Station there is service line number, directly name of station in

English transcription, name in Russian Cyrillic, transfer to other lines, date of open, elevation below grade, station structure type, coordinates and picture link at Wiki Commons. An example of data can be seen in Figure 1. But some data may not exist for some stations. Some of this information is necessary for further search of the nearest establishments, which will be performed using the Foursquare API.



List of active stations [edit]

L	English transcription	Russian Cyrillic	Transfer	Opened	Elev.	Type	Coordinates	Pic.
1	Bulvar Rokossovskogo	Бульвар Рокоссовского	<14>	1990-08-01	−8 m	column, triple-span	55.8148°N 37.7342°E	
1	Cherkizovskaya	Черкизовская	{14}	1990-08-01	−9 m	single-vault, shallow	55.8038°N 37.7448°E	
1	Preobrazhenskaya Ploshchad	Преображенская площадь		1965-12-31	−8 m	column, triple-span	55.7963°N 37.7151°E	
1	Sokolniki	Сокольники		1935-05-15	−9 m	column, triple-span	55.7888°N 37.6802°E	
1	Krasnoselskaya	Красносельская		1935-05-15	−8 m	column, double-span	55.7801°N 37.6673°E	
1	Komsomolskaya	Комсомольская	5	1935-05-15	−8 m	column, triple-span	55.7753°N 37.6562°E	
1	Krasnye Vorota	Красные ворота		1935-05-15	−31 m	pylon, triple-vault	55.7690°N 37.6487°E	
1	Chistyye Prudy	Чистые пруды	6 10	1935-05-15	−35 m	pylon, triple-vault	55.7657°N 37.6388°E	
1	Lubyanka	Лубянка	7	1935-05-15	−32.5 m	pylon, triple-vault	55.7597°N 37.6272°E	
1	Okhotny Ryad	Охотный ряд	2 (3)	1935-05-15	−15 m	pylon, triple-vault	55.7577°N 37.6166°E	
1	Biblioteka Imeni Lenina	Библиотека имени Ленина	3 4 9	1935-05-15	−12 m	single-vault, shallow	55.7512°N 37.6100°E	
1	Kropotkinskaya	Кропоткинская		1935-05-15	−13 m	column, triple-span	55.7453°N 37.6037°E	
1	Park Kultury	Парк культуры	5	1935-05-15	−10.5 m	column, triple-span	55.7356°N 37.5943°E	
1	Frunzenskaya	Фрунзенская		1957-05-01	−42 m	pylon, triple-vault	55.7267°N 37.5786°E	
1	Sportivnaya	Спортивная	<14>	1957-05-01	−42 m	pylon, triple-vault	55.7233°N 37.5639°E	
1	Vorobyovy Gory	Воробьёвы горы		1959-01-12	+10 m	within bridge	55.7103°N 37.5592°E	
1	Universitet	Университет		1959-01-12	−26.5 m	pylon, triple-vault	55.6926°N 37.5333°E	
1	Prospekt Vernadskogo	Проспект Вернадского		1963-12-30	−8 m	column, triple-span	55.6771°N 37.5060°E	
1	Yugo-Zapadnaya	Юго-Западная		1963-12-30	−8 m	column, triple-span	55.6637°N 37.4833°E	

Figure 1. Data on the Wikipedia page.

Data on the average cost of real estate at metro stations areas that will allow perform a comparative analysis will be taken from this article. Information presented as a metro station table with mean price (RUR/square meter), mean flat price (RUR), annual change of these prices. Please note that the station names are in Russian.

Passenger traffic per station data will be taken from this site. Traffic values in thousands of people per day. In this case, the stations are separated by lines, so the information is located on several pages. Each page represents a specific line. This data is also needed for further comparative analysis. And again, it’s a Russian language site.

Since there is no data on the level of pollution for each metro station, I can use the ecology rate by district data. This requires the list of metro stations by district and the pollution rating of districts. Rating values are represented as numbers from 1 (high level of pollution) to 4 (low level of pollution).

2.2 Data cleaning

After data is collected in dataframes, need to prepare them for use. However, it should be noted that many stations do not have any data. The missing data can be found in various other sources, but this is difficult to implement for a large number of stations. Therefore, merging will occur at the final stage of analysis, when the number of stations is minimal and possible to add the missing data manually. Also, because of the variety of Russian-language spelling variants, it is worth bringing all Russian-language names to one variant. For example, ‘Troparyovo’ = ‘Тропарево’ = Тропарёво.

The dataframe of metro stations contains a large number of features, while the only interesting columns are the names in English and Russian and coordinates. The column in Russian is needed as a key when later merging with other frames, since the sources of other data are in Russian. The other columns can be discarded. After that, need to remove extra characters from the rows, such as Byte order Mark or degree signs in the coordinate column. Then, need to divide the coordinates column into the latitude and longitude columns and fill in the missing data using the Geocoder library. In the end, need to delete duplicate stations.

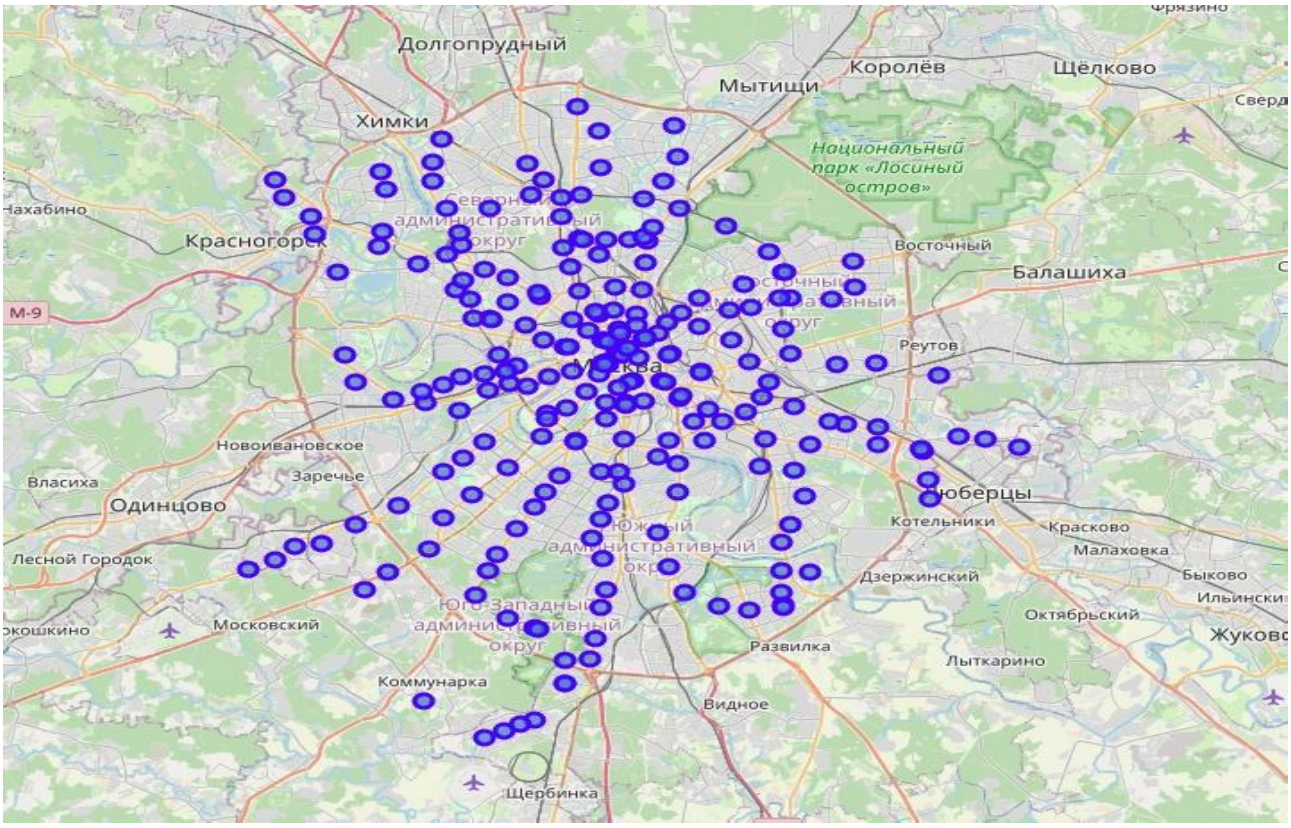


Figure 2. Location of Moscow metro stations on the map

The rest of the frames are generally ready to work. Only need to remove extra characters from some columns, correct the Russian-language names of stations and districts, and delete lines with missing values.

The final stage of data collection and clearing is the collection and processing of information about establishments in metro station areas using the Foursquare API. The radius of the area is 900 meters. Since the fusion restaurant includes various cuisines of the world, restaurants of certain cuisines will compete. Therefore, need to make edits to this frame - replace all lines containing the word 'restaurant' with 'restaurant' ('Asian restaurant' => 'Restaurant')

3. Methodology.

3.1 Algorithm selection.

To solve this problem, I use unsupervised machine learning algorithm, specifically the k-means clustering algorithm. As mentioned above, there is no data available for many stations, so the cluster analysis will be based on data about the nearest establishments. In other words, the idea is to find similar areas of stations and identify areas with a predominant number of restaurants. To use the clustering algorithm, need to group and convert categorical data of restaurant types into numerical data using one-hot encoding. Along the way, I can find the most common places (Figure 3).

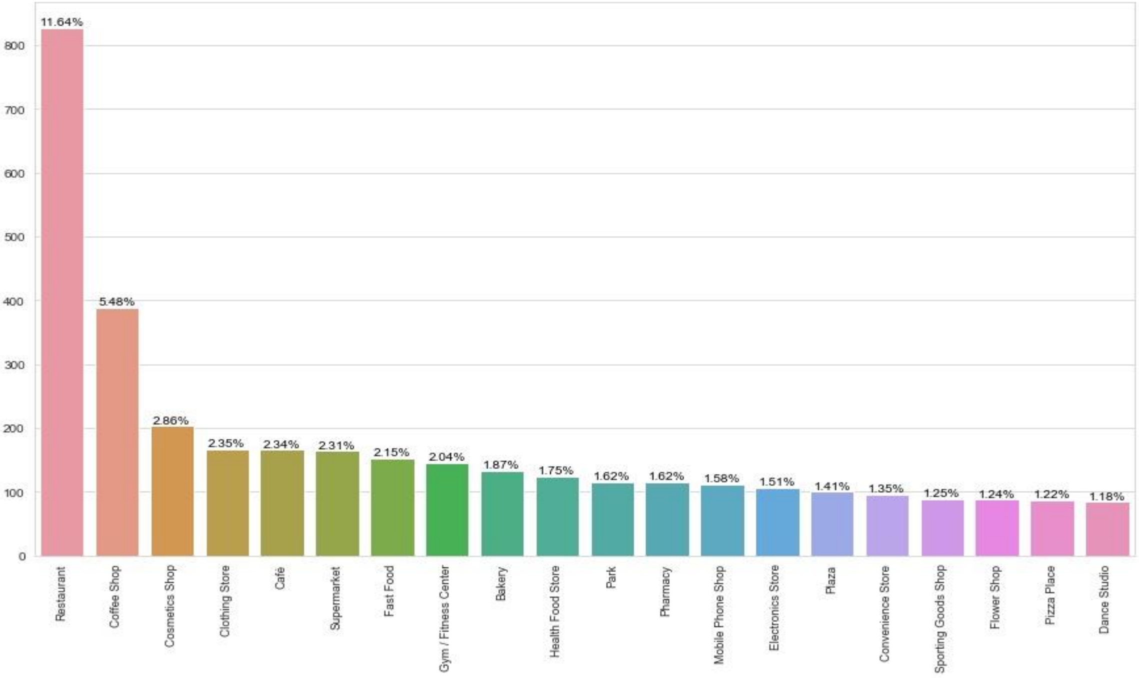


Figure 3. Number of most common venues

The most common venues in the city are restaurants, which account for about 12% of the total number of establishments. I can assume that the number of metro station areas suitable for opening a new restaurant will be small.

The search for the optimal value of k with Elbow method did not yield any good results (Figure 4), so the silhouette score method was used (Figure 5).

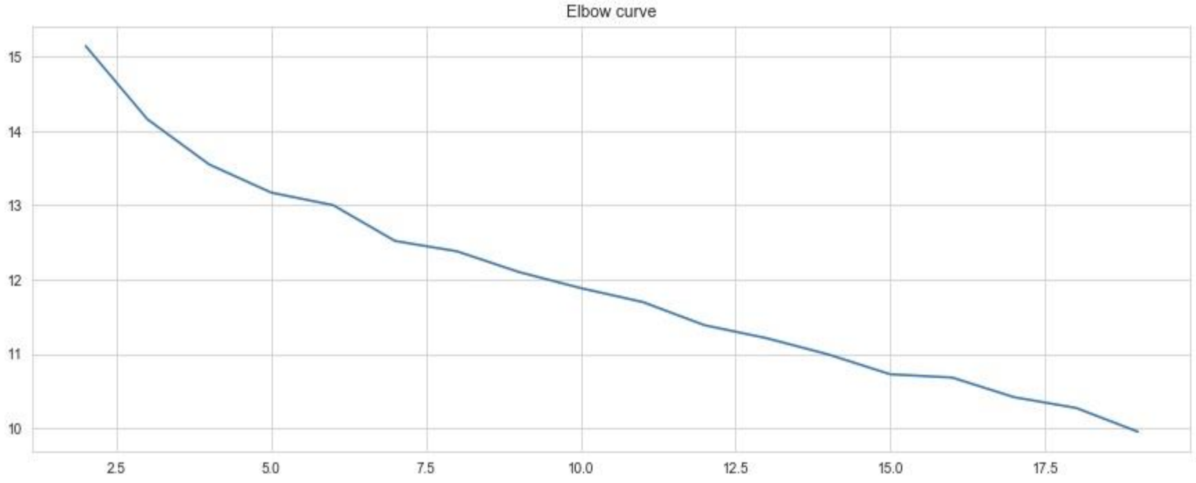
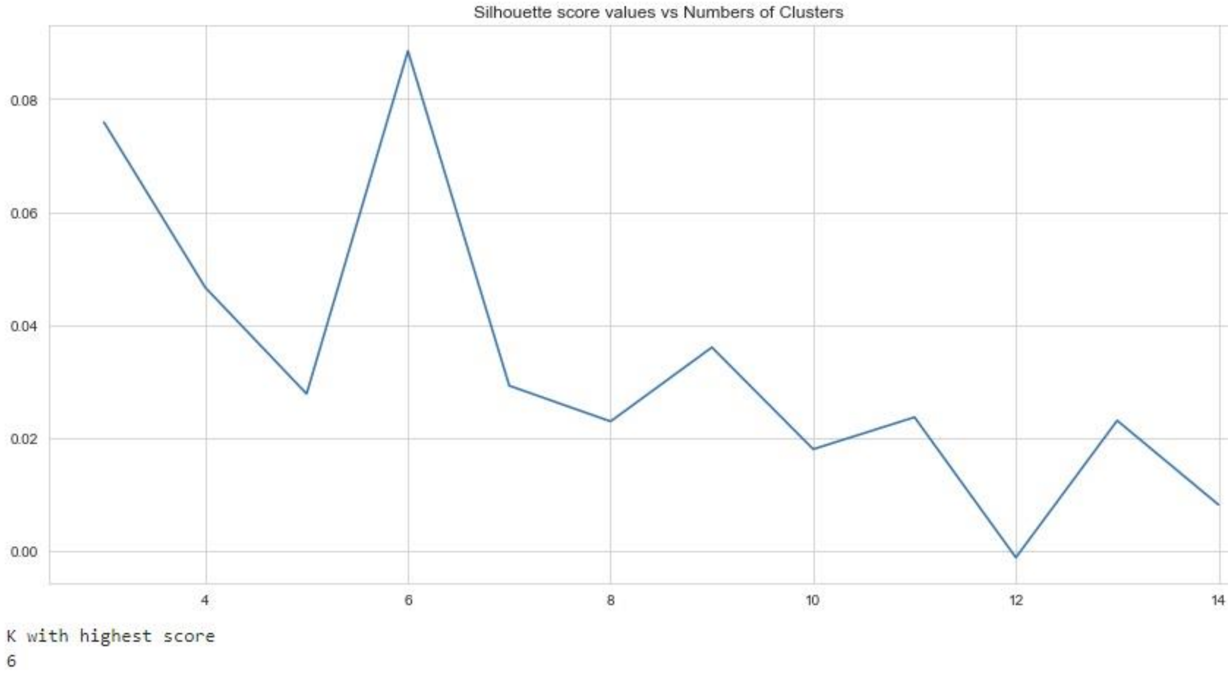


Figure 4. Elbow method



```
K with highest score
6
```

Figure 5. Silhouette score method

The graph shows that the optimal value of k = 6. Other parameters of the model were not changed.

3.2 Cluster analysis.

As a result of using the k-means clustering algorithm, stations are divided into clusters as follows

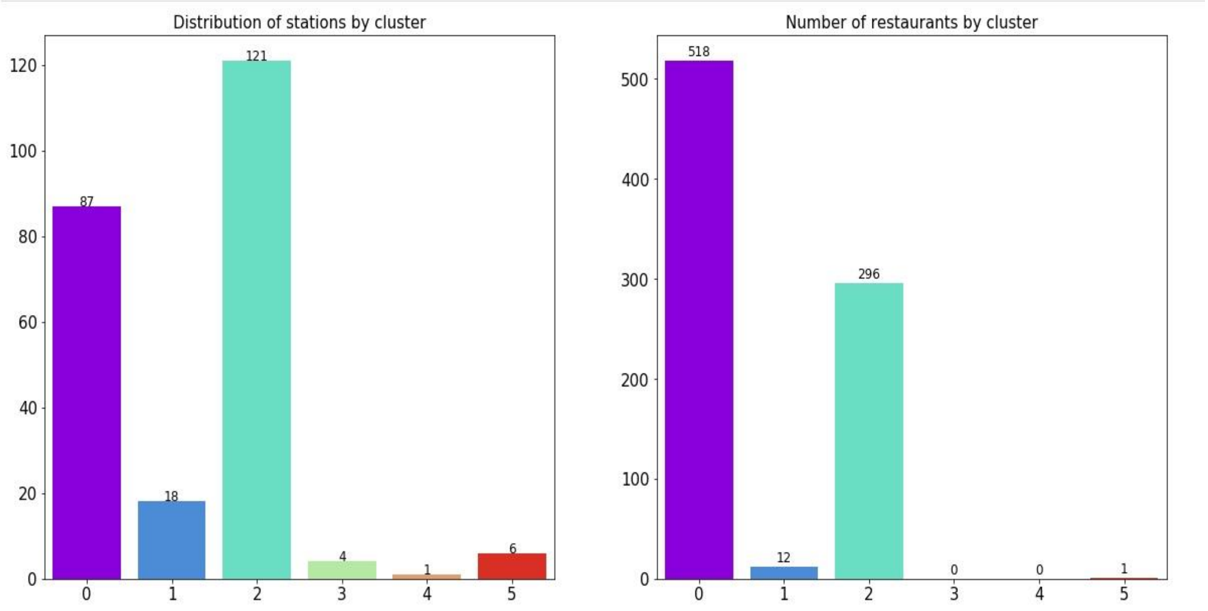


Figure 6. Results of clustering

Looking at the charts, I can note one cluster with a huge number of restaurants. This is cluster 0, which consists of only 87 stations, but contains 518 restaurants. It is obvious that the metro station areas of this cluster are not suitable for opening a restaurant, so this cluster will not be analyzed in detail. For cluster 2, which contains 121 stations and 296 restaurants, need to find out the number of other establishments.

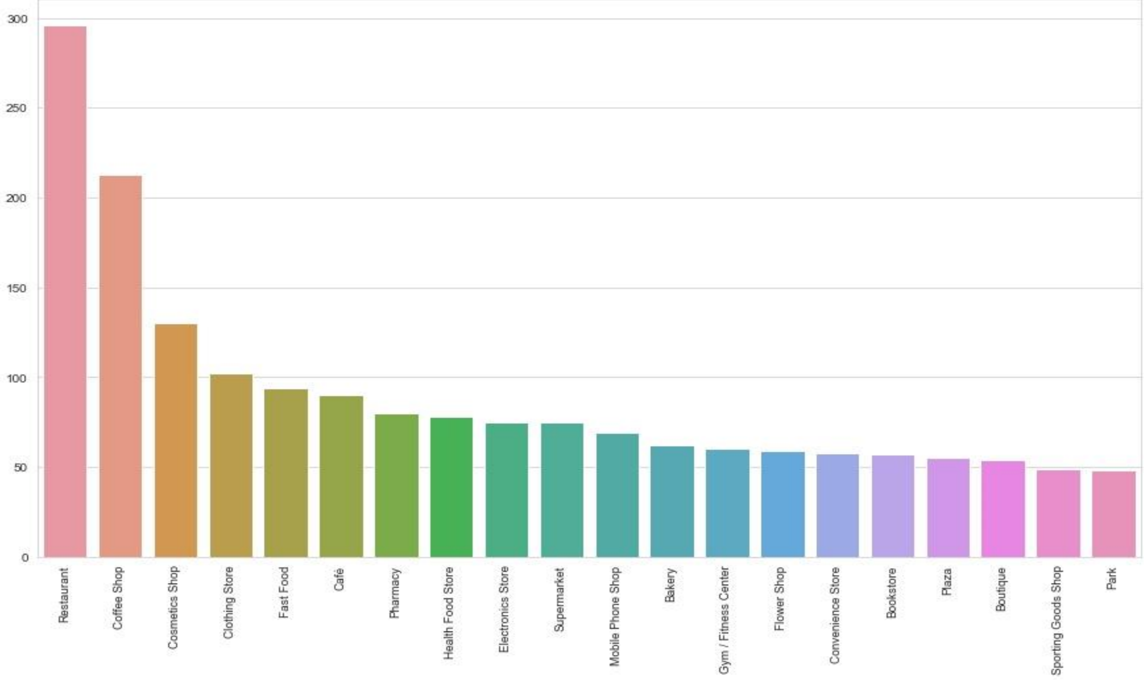


Figure 7. The most common venues of cluster 2

Despite the smaller number of restaurants, the number of other establishments, such as cafes and coffee shops, makes the station areas of this cluster unsuitable for opening a restaurant.

Cluster 1 has only 18 stations and 12 restaurants. The number of other food venues is also small. I think it's worth selecting stations from the cluster that don't have restaurants in their vicinity

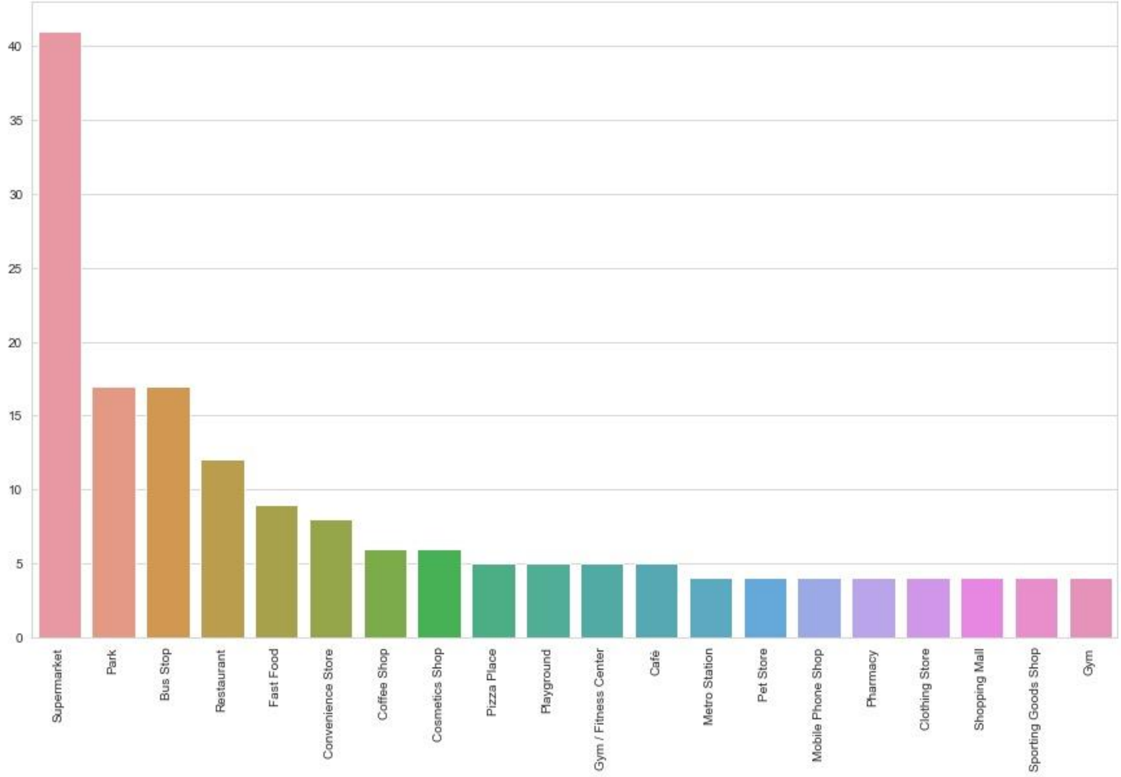


Figure 8. The most common venues of cluster 2

Cluster 3 consists of many parks, two historical sites, and no restaurants. These areas of the metro station are suitable for opening a new restaurant

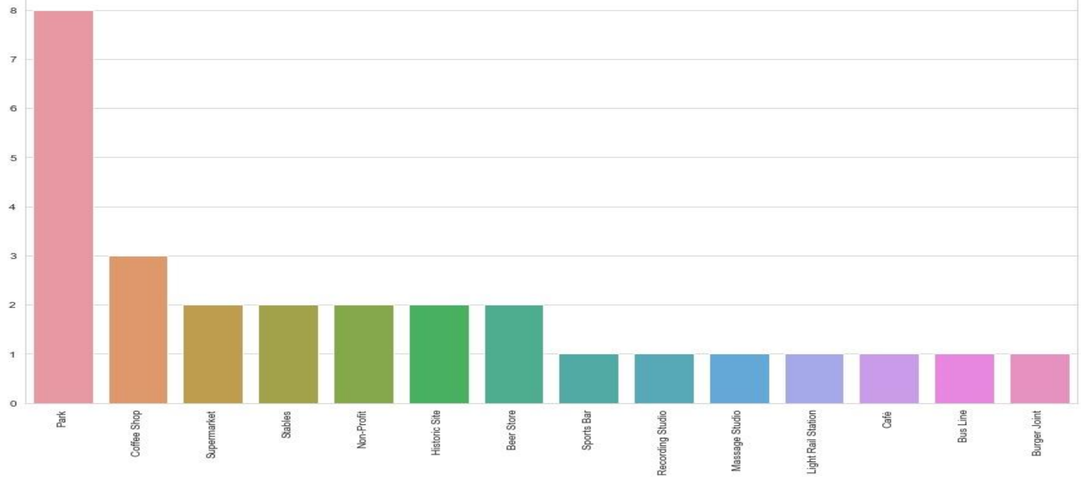
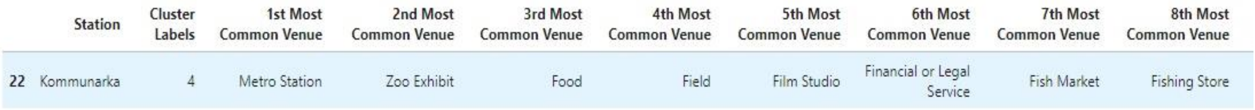


Figure 9. The most common venues of cluster 3

Cluster 4, which consist of only one station – Kommunarka metro station.



	Station	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
22	Kommunarka	4	Metro Station	Zoo Exhibit	Food	Field	Film Studio	Financial or Legal Service	Fish Market	Fishing Store

Figure 10. The most common venues of cluster 4

No any cafes or restaurants. This station area can be a good place for open new restaurant.

And cluster 5, which includes many sports facilities, but there are also museums and just one restaurant, which is located at Spartak station. I think, it is possible to exclude this metro station area from the final selection.

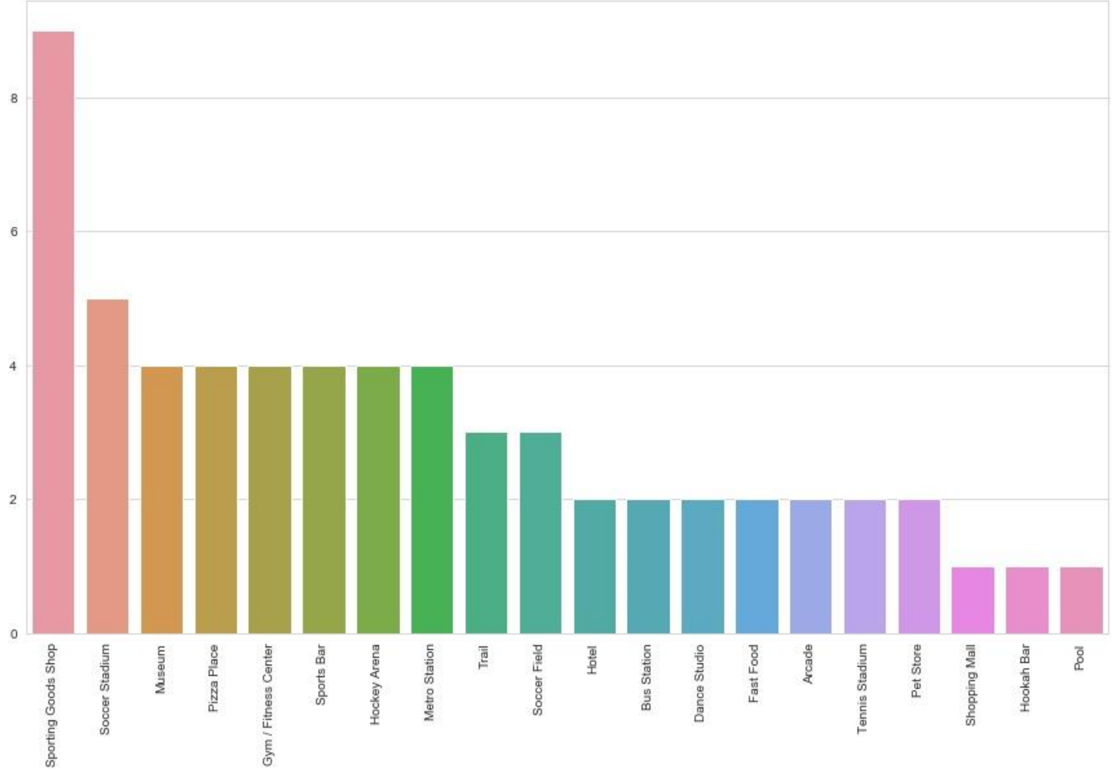
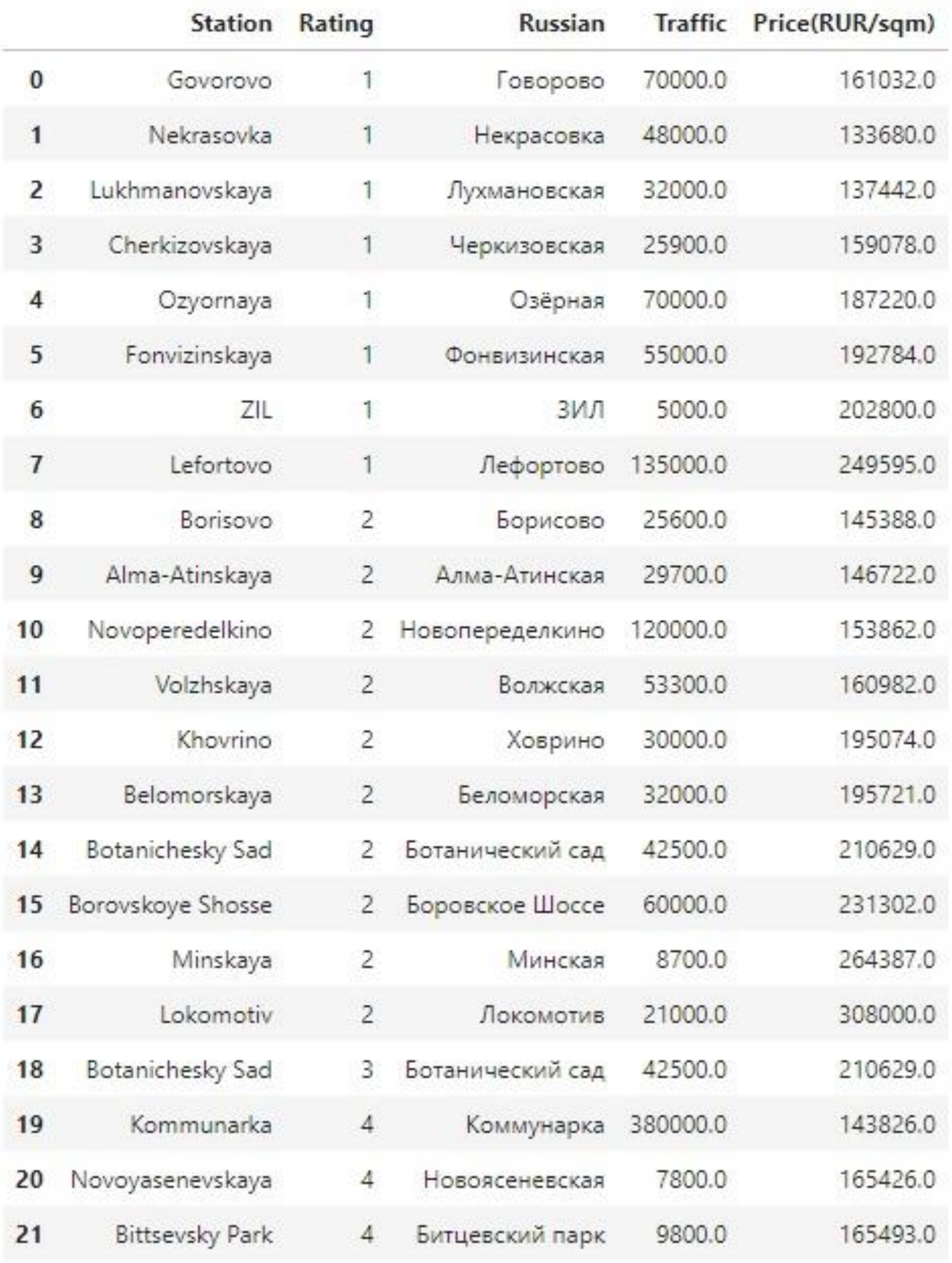


Figure 11. The most common venues of cluster 5

Now I can combine the selected stations into a single frame and add price, traffic, and eco-rating values. The average price per square meter will be used as the price value.

4. Results.



	Station	Rating	Russian	Traffic	Price(RUR/sqm)
0	Govorovo	1	Говорово	70000.0	161032.0
1	Nekrasovka	1	Некрасовка	48000.0	133680.0
2	Lukhmanovskaya	1	Лухмановская	32000.0	137442.0
3	Cherkizovskaya	1	Черкизовская	25900.0	159078.0
4	Ozyornaya	1	Озёрная	70000.0	187220.0
5	Fonvizinskaya	1	Фонвизинская	55000.0	192784.0
6	ZIL	1	ЗИЛ	5000.0	202800.0
7	Lefortovo	1	Лефортово	135000.0	249595.0
8	Borisovo	2	Борисово	25600.0	145388.0
9	Alma-Atinskaya	2	Алма-Атинская	29700.0	146722.0
10	Novoperedelkino	2	Новопеределкино	120000.0	153862.0
11	Volzhskaya	2	Волжская	53300.0	160982.0
12	Khovrino	2	Ховрино	30000.0	195074.0
13	Belomorskaya	2	Беломорская	32000.0	195721.0
14	Botanichesky Sad	2	Ботанический сад	42500.0	210629.0
15	Borovskoye Shosse	2	Боровское Шоссе	60000.0	231302.0
16	Minskaya	2	Минская	8700.0	264387.0
17	Lokomotiv	2	Локомотив	21000.0	308000.0
18	Botanichesky Sad	3	Ботанический сад	42500.0	210629.0
19	Kommunarka	4	Коммунарка	380000.0	143826.0
20	Novoyasenevskaya	4	Новоясеневская	7800.0	165426.0
21	Bittsevsky Park	4	Битцевский парк	9800.0	165493.0

Figure 12. Resulting dataframe

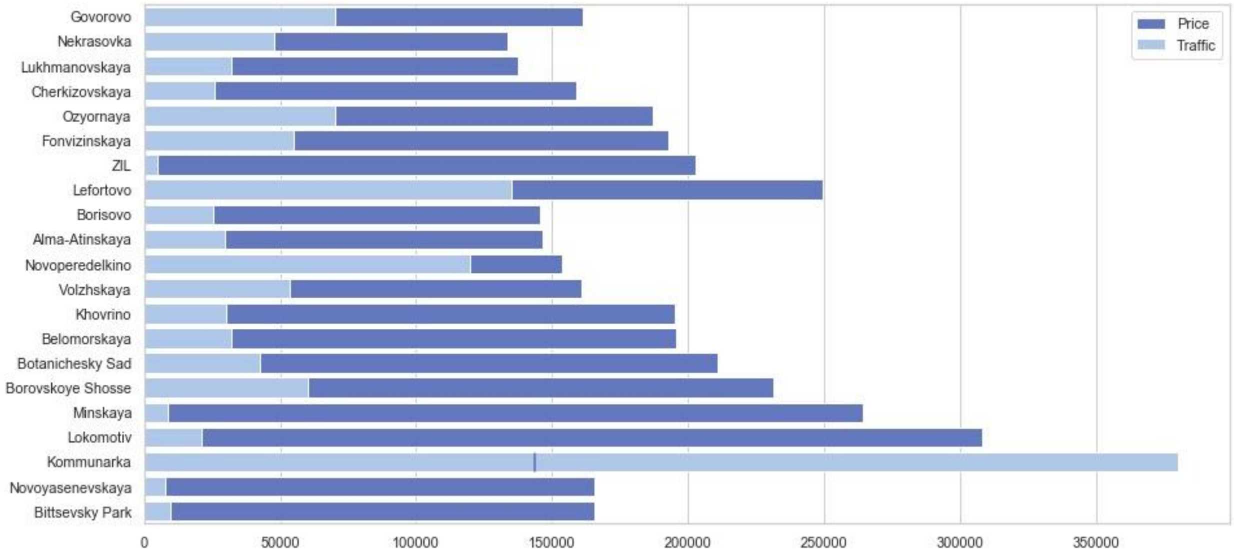


Figure 13. Prices and traffics of stations

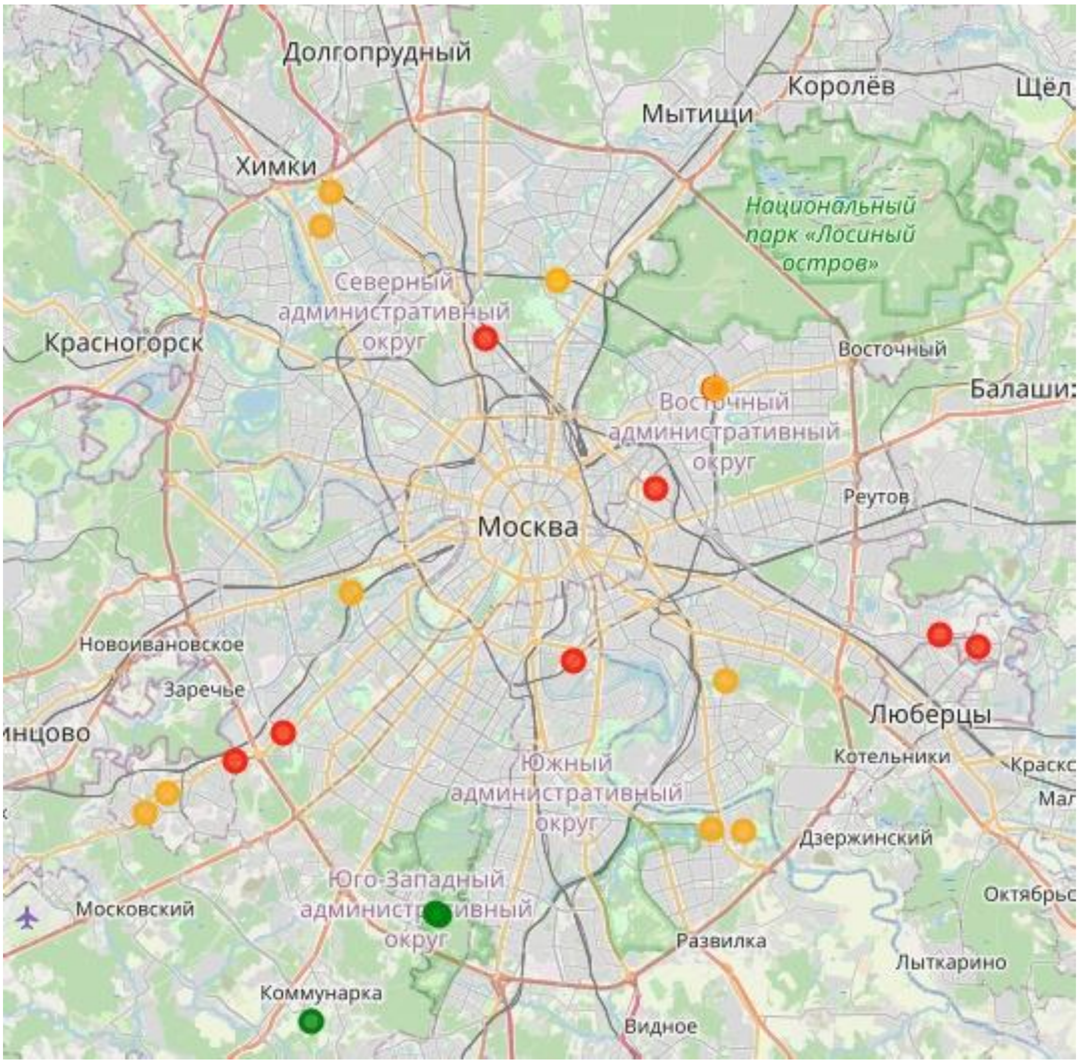


Figure 14. Location of best metro stations

5. Discussion.

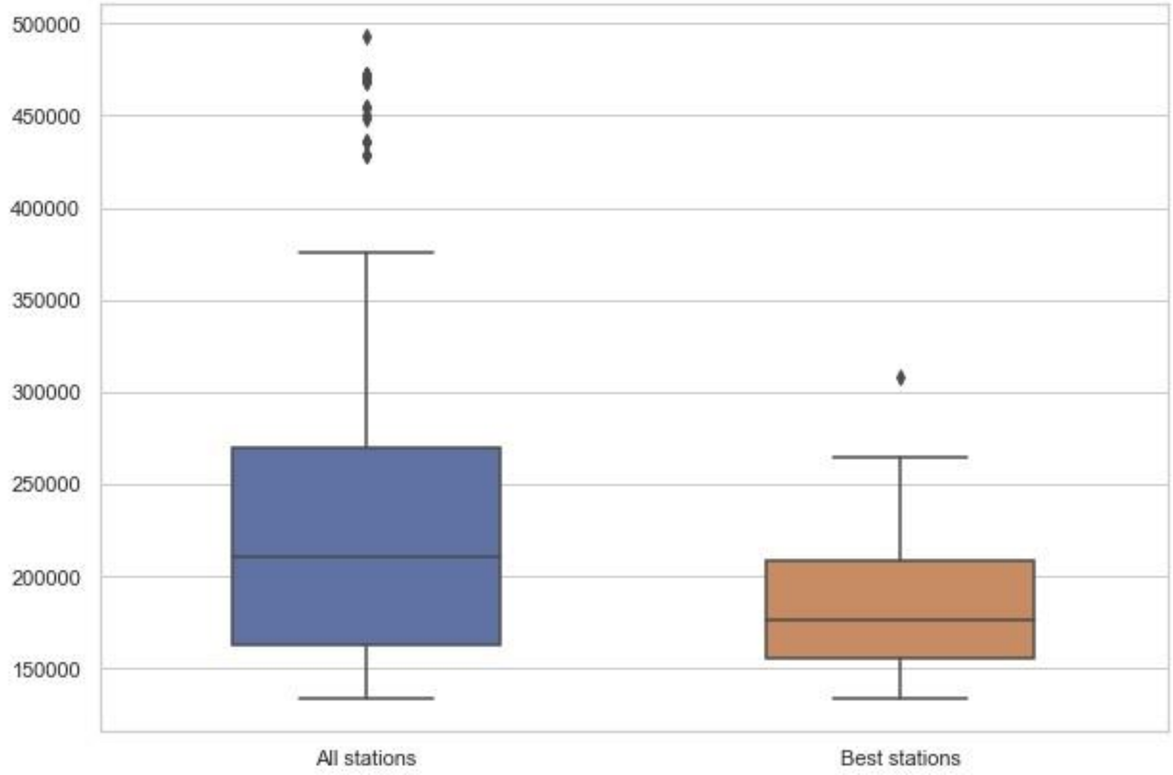


Figure 15. Prices comparison

Prices per square meter in the final sample are quite small. So the average price per square meter for all stations is 235753 RUR, and for stations in the final sample - 187321 RUR.

Despite the fact that all areas of the stations from the final sample are somehow suitable for opening a new restaurant, you can select the most suitable area - this is the area of the Kommunarka metro station. This station has the highest traffic and rating values and the lowest price.

When you look at each station area in detail, you can find factors that reduce the attractiveness of the area for opening a restaurant. For example, the area of the ZIL metro station, which is located next to the factory of the same name (but it closed now). On the one hand, the passenger traffic is 5,000 people per day, the average price is about 200,000 RUR, and the eco-rating is 1.While on the other hand, in the former factory has already under construction a large number of residential complexes, the presence of which can raise the popularity of the area.

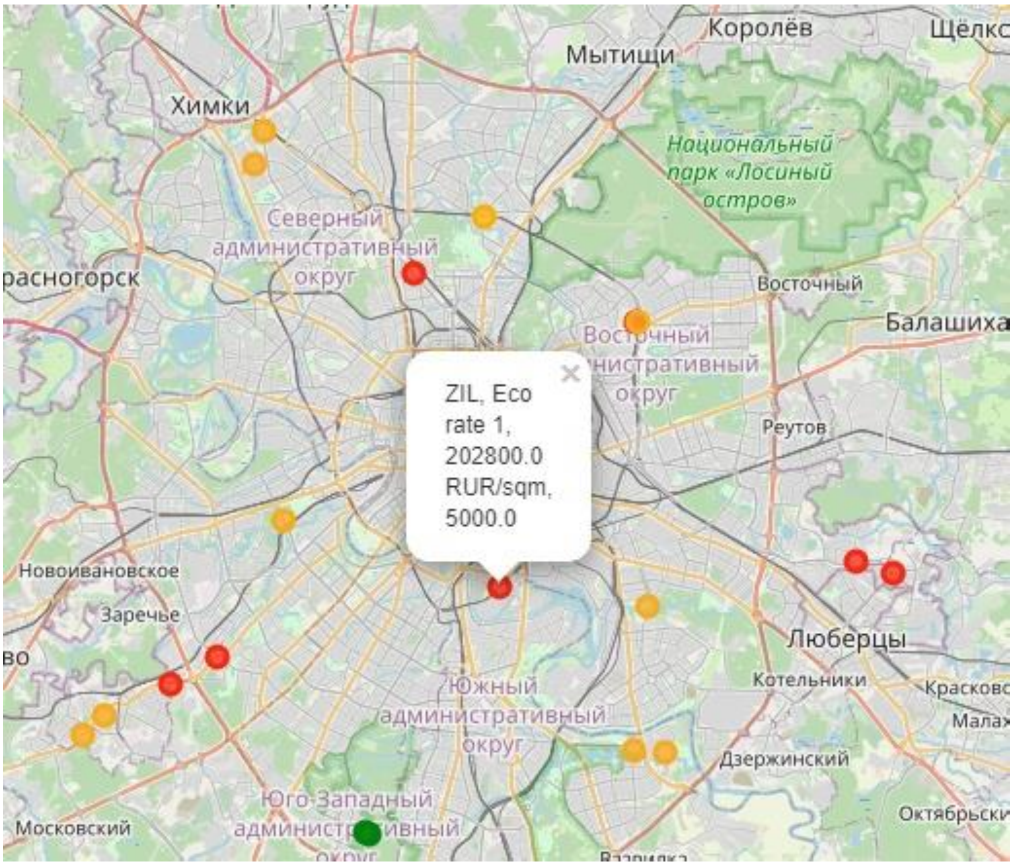

Королёв
Щёлко
Мытищи
Химки
Национальный парк «Лосиный остров»
Северный административный округ
Красногорск
Восточный
Балашиха
Восточный административный округ
ZIL, Eco rate 1, 202800.0 RUR/sqm, 5000.0
Реутов
Новоивановское
Заречье
Люберцы
Котельники
Красково
Малах
Южный административный округ
Дзержинский
Юго-Западный административный округ
Московский
Октябрьский

Figure 16. ZIL station on a map

You can also see that there are two pairs of stations located close to each other. It's a Bittsevsky park\Novoyasenevskaya metro stations and Cherkizovskaya or Lokomotiv stations. If the first pair can be considered as one station, then the second pair is not easy. These stations have different values for prices, traffics, and eco ratings.

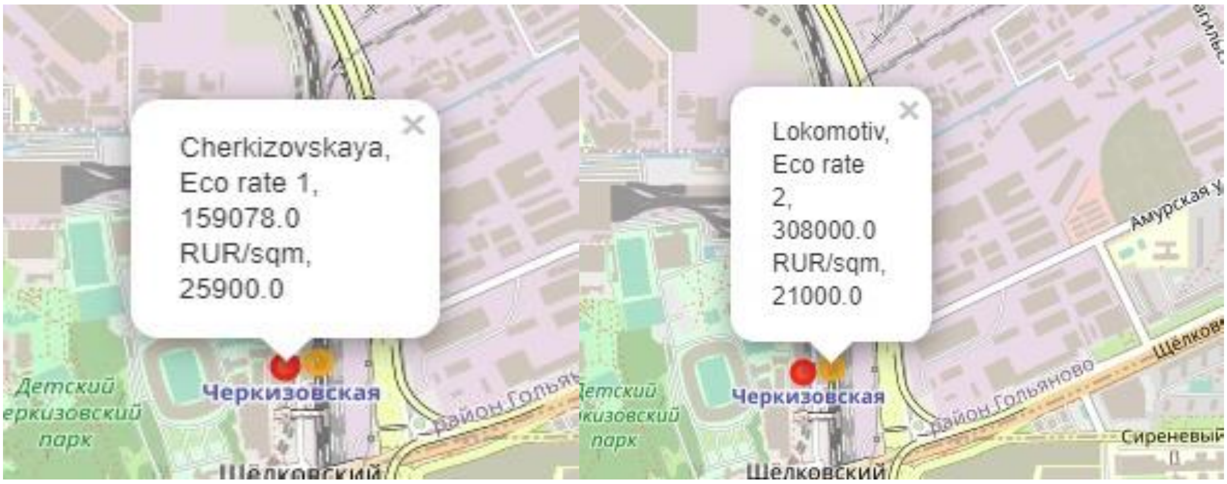

Cherkizovskaya, Eco rate 1, 159078.0 RUR/sqm, 25900.0
Lokomotiv, Eco rate 2, 308000.0 RUR/sqm, 21000.0
Черкизовская
Детский Черкизовский парк
район Гольяново
Щёлковский
Амурская ул.
Щёлковское
Сиреневый

Figure 17. Cherkizovskaya and Lokomotiv Stations on a map

6. Conclusion.

In this project, I divided Moscow metro station areas into clusters using the K-means algorithm. After that, I analyzed each cluster for the suitability of the cluster's metro station areas for opening a new restaurant. As a result, two clusters were completely excluded and one partially. Then the information about the metro station areas was updated with information about the passenger traffic of the stations, the average cost per square meter of real estate in the area of these stations, and the values of the station's eco-rating. As a result, the final sample included 20 station areas, considering paired stations. Again, the results obtained are not absolutely accurate and require in-depth analysis. The search methodology used in this project can be applied to many other cities around the world.