Data Science Capstone project

USING DATA LOCATION TO IDENTIFY PLACES TO OPEN ART CAFFEE IN MOSCOW

Tatiana Shumskaya

1. Introduction

1.1 Background

Moscow is the capital and largest city of Russia. The city stands on the Moskva River in Central Russia, with a population estimated over 20 million residents in the Moscow Metropolitan Area. The city covers an area over 26,000 square kilometres (10,000 sq mi). Moscow is among the world's largest cities, being the largest city entirely within Europe. Also, Moscow has one of the largest municipal economies in Europe and it accounts more than one-fifth of Russia's gross domestic product (GDP).

1.2 Problem

Despite the economic difficulties and the unstable ruble exchange rate, the public catering market in Russia is growing steadily. In comparison with the indicators of 2014, it has grown by almost a trillion rubles: according to preliminary estimates, in 2019 its volume exceeded 2.2 trillion rubles. The industry continues to be profitable, and any catering establishment can become successful. Marketing research confirms the growing popularity of themed establishments among above-average Russians. This is due to the expanded opportunities for a good rest, which are provided in such places. These are places where you can listen to live music with friends and enjoy the taste of freshly prepared food. Thus, opening an art cafe in Moscow can be successful and bring profit to the business owner. However, Moscow is a city with a large territory and different population density. Also, the rental price for premises will vary greatly from district to district. It is important to choose a good place to open an art cafe based on what places are nearby. For example, in an industrial area, an art cafe will be unpopular due to the high cost of food, and in a business area, because of the unsuitable format of the establishment. It is also advisable to choose an area of the city with a minimum level of competition. Of course, the ideal option is to rent a room in the central or historical part of the city. However, in this case, the businessman must be ready for dumping from competitors. In addition, the rental rate in such areas may be too high. That is why it makes sense to think about opening

an art cafe in a densely populated residential area, but close to the city center and which contains establishments that may also be of interest to potential visitors to art cafes, such as theaters, museums, galleries, etc. One should note the popularity of public transport in Moscow, especially the metro. The proximity to the metro station is an important advantage for any facility, including a cafe.

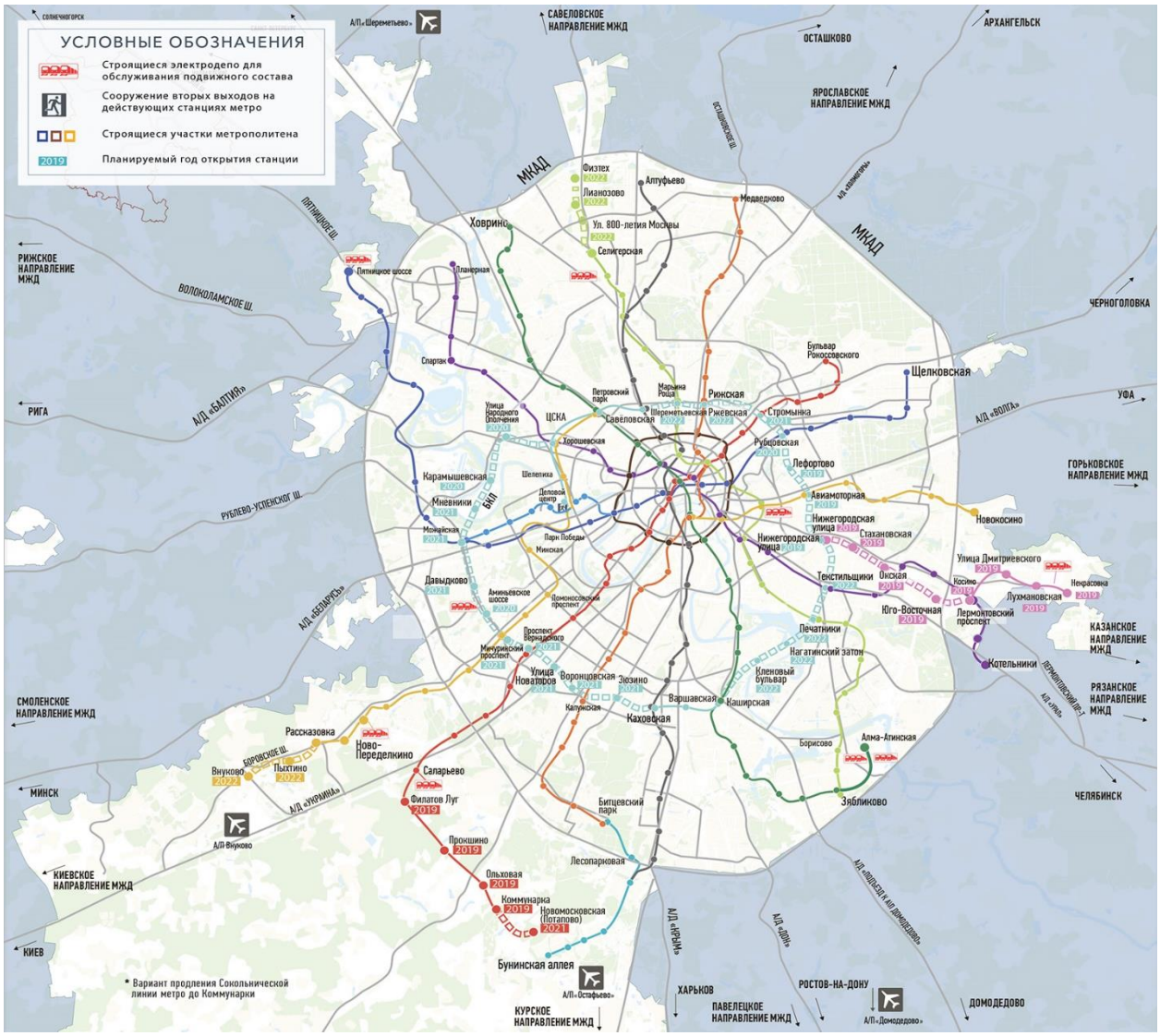


Figure 1. Moscow metro map

The aim of this project is to provide investors and business owners suitable areas, where they could potentially open a new Art café:

1. in location nearby theaters, museums, and galleries;

2. outside the city center, but as much close to the center as possible;

3. nearby one of metro stations.

2. Data acquisition and cleaning

2.1 Data sourses

In order to better understand the city of Moscow a dataset will be built through the use of web scraping and pulling venue data through the Foursquare API. A dataset with the geographical coordinates of metro stations will also be used, which will be processed using the Nominatum library. Geographic coordinates of Moscow metro stations and the location of metro stations and districts were taken from the official website of the Moscow government. Once the data is collected it will be wrangled into shape and explored using the Pandas library and Seaborn plotting library. Analysis of the data will be carried out with the Scikit-learn library, in particular K-means clustering will be applied. Results will be displayed as plots as well as maps of the city, which will be produced using the Folium library. These maps can be used to quickly narrow down potential locations for a new restaurant from a bird's eye view or 1000m perspective.

2.2 Data cleaning

Data with the binding of metro stations to Moscow districts, obtained from the website of the Moscow government, is in Russian. The district names will be transliterated after the cluster analysis of the districts. Initially, the site data contained a list of stations, line name, administrative district name, global station ID, region, status, and local ID number. In total, data were obtained for 305 Moscow metro stations.

	Station	Line	AdmArea	global_id	District	Status	ID
0	Третьяковская	Калининская линия	Центральный административный округ	58701962	район Замоскворечье	действует	136
1	Медведково	Калужско-Рижская линия	Северо-Восточный административный округ	58701963	район Северное Медведково	действует	86
2	Первомайская	Арбатско-Покровская линия	Восточный административный округ	58701964	район Измайлово	действует	41
3	Калужская	Калужско-Рижская линия	Юго-Западный административный округ	58701965	Обручевский район	действует	104
4	Каховская	Большая кольцевая линия	Юго-Западный административный округ	58701966	район Зюзино	строится	251

For further analysis, we need to delete the columns with the station line, administrative districts and identification numbers, leaving only the station names and districts.

	Station	District
0	Третьяковская	район Замоскворечье
1	Медведково	район Северное Медведково
2	Первомайская	район Измайлово
3	Калужская	Обручевский район
4	Каховская	район Зюзино

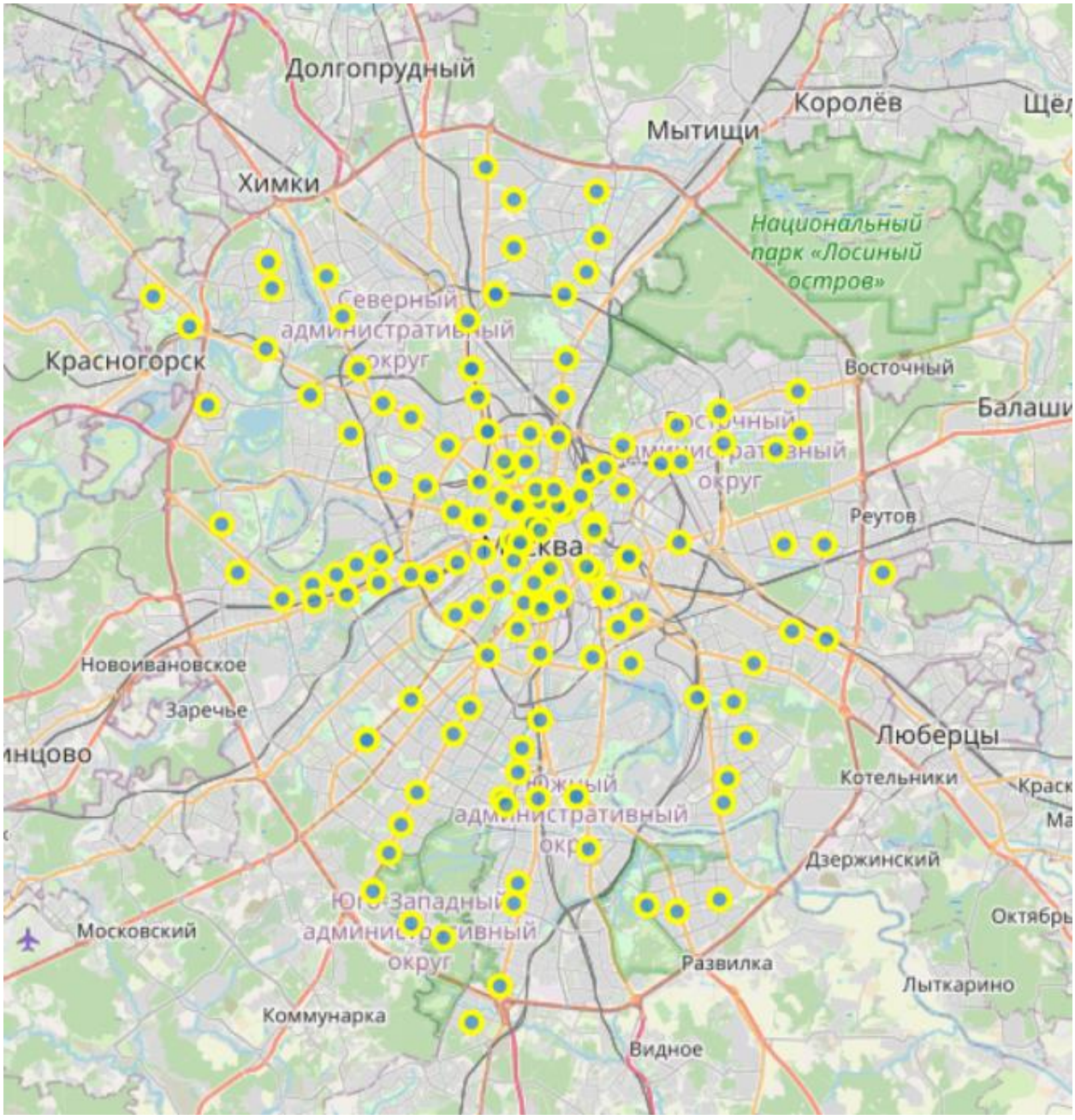
It is necessary to remove duplicate station names because some stations are on different lines (transfer stations) and are listed in the table more than once. It is also necessary to edit the name of the districts for uniformity - some names are indicated as “District Izmaylovo” or “Obruchevskuy District”, remove the word “district” from the name of the district.

	Station	District
0	Третьяковская	Замоскворечье
1	Медведково	Северное Медведково
2	Первомайская	Измайлово
3	Калужская	Обручевский
4	Каховская	Зюзино

Geographic coordinates for metro stations will also be used. Coordinates are not available for a small number of stations; they will be excluded from further analysis. Combine the Station-District and Station-Coordinates tables by the Station field.

	Station	District	Latitude	Longitude
0	Третьяковская	Замоскворечье	55.74061	37.62492
1	Медведково	Северное Медведково	55.88594	37.66120
2	Первомайская	Измайлово	55.79342	37.79979
3	Калужская	Обручевский	55.65566	37.53923
4	Каховская	Зюзино	55.65332	37.59722

After removing duplicate station names, as well as excluding stations without coordinates from the analysis, 152 unique metro stations remained in the dataframe. They are shown on the map of Moscow below:



For these stations and the areas in which they are located, data analysis will be carried out in order to find the most suitable areas and stations for opening an art cafe. For this, data from the Foursquare API will also be used.

3. Exploratory Data Analysis

To solve this problem, it is necessary to analyze the districts of Moscow in order to understand which districts are most similar to the districts in the historical center of

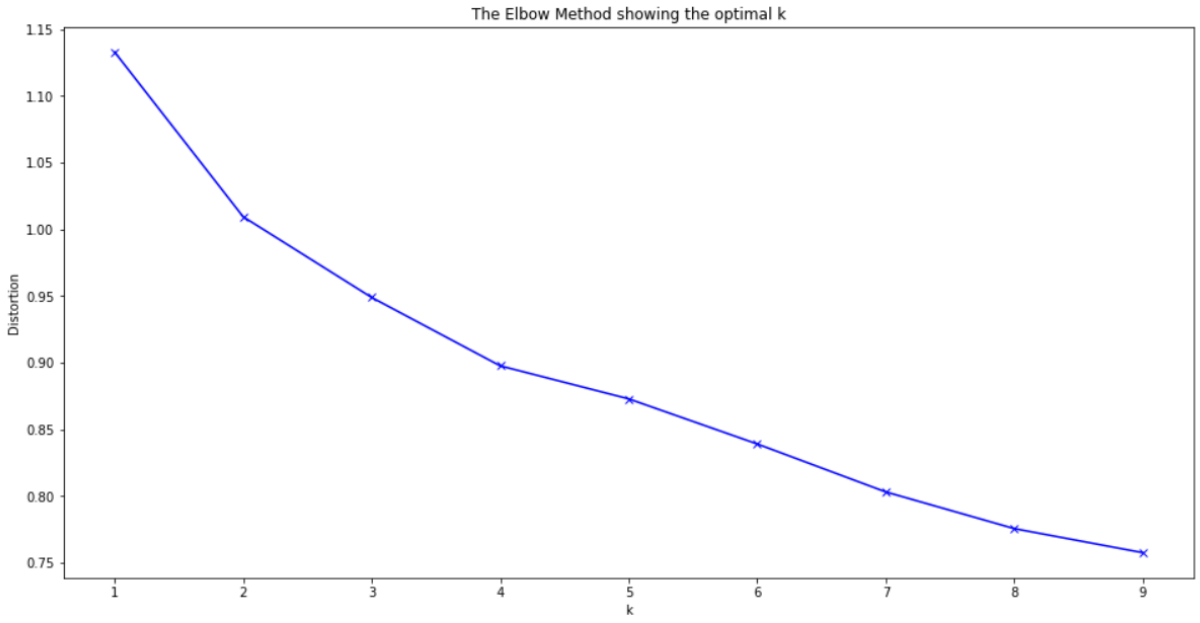
the city, but at the same time they are not in the center, but farther from it, where the rent is cheaper and the competition is lower. For this, we will use k-means cluster analysis. To analyze and cluster areas, we use Foursquare API data for each station within a radius of 1000m from the station. This will allow to understand which venues are located in the specified radius. In the future, the analysis of the venues near each station will allow to group districts. A total of 11684 unique venues were obtained using the Foursquare API.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Замоскворечье	55.74061	37.62492	Tretyakov Gallery (Третьяковская галерея)	55.741006	37.621095	Art Museum
1	Замоскворечье	55.74061	37.62492	Эссе	55.741163	37.628770	Jazz Club
2	Замоскворечье	55.74061	37.62492	Издательский дом «Самокат»	55.738188	37.626196	Bookstore
3	Замоскворечье	55.74061	37.62492	ABC Coffee Roasters	55.741158	37.622940	Coffee Shop
4	Замоскворечье	55.74061	37.62492	Республика	55.740934	37.625069	Bookstore

For each district, we will find the 10 most common categories of venus, on the basis of which we will conduct a cluster analysis of the data using the k-means method.

	District	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	9th Most Common Venue	10th Most Common Venue
0	Академический	Health Food Store	Pharmacy	Dance Studio	Café	Gym / Fitness Center	Sushi Restaurant	Supermarket	Italian Restaurant	Sporting Goods Shop	Park
1	Алексеевский	Coffee Shop	Auto Workshop	Park	Cafeteria	Mobile Phone Shop	Toy / Game Store	Health Food Store	Other Repair Shop	Hobby Shop	Electronics Store
2	Арбат	Coffee Shop	Plaza	Hotel	Museum	Cocktail Bar	Art Gallery	Russian Restaurant	Art Museum	Restaurant	Bakery
3	Аэропорт	Coffee Shop	Café	Hotel	Gym / Fitness Center	Cosmetics Shop	Park	Pharmacy	Health Food Store	Supermarket	Eastern European Restaurant
4	Бабушкинский	Supermarket	Fast Food Restaurant	Park	Gym / Fitness Center	Gym	Middle Eastern Restaurant	Cosmetics Shop	Pharmacy	Convenience Store	Food & Drink Shop

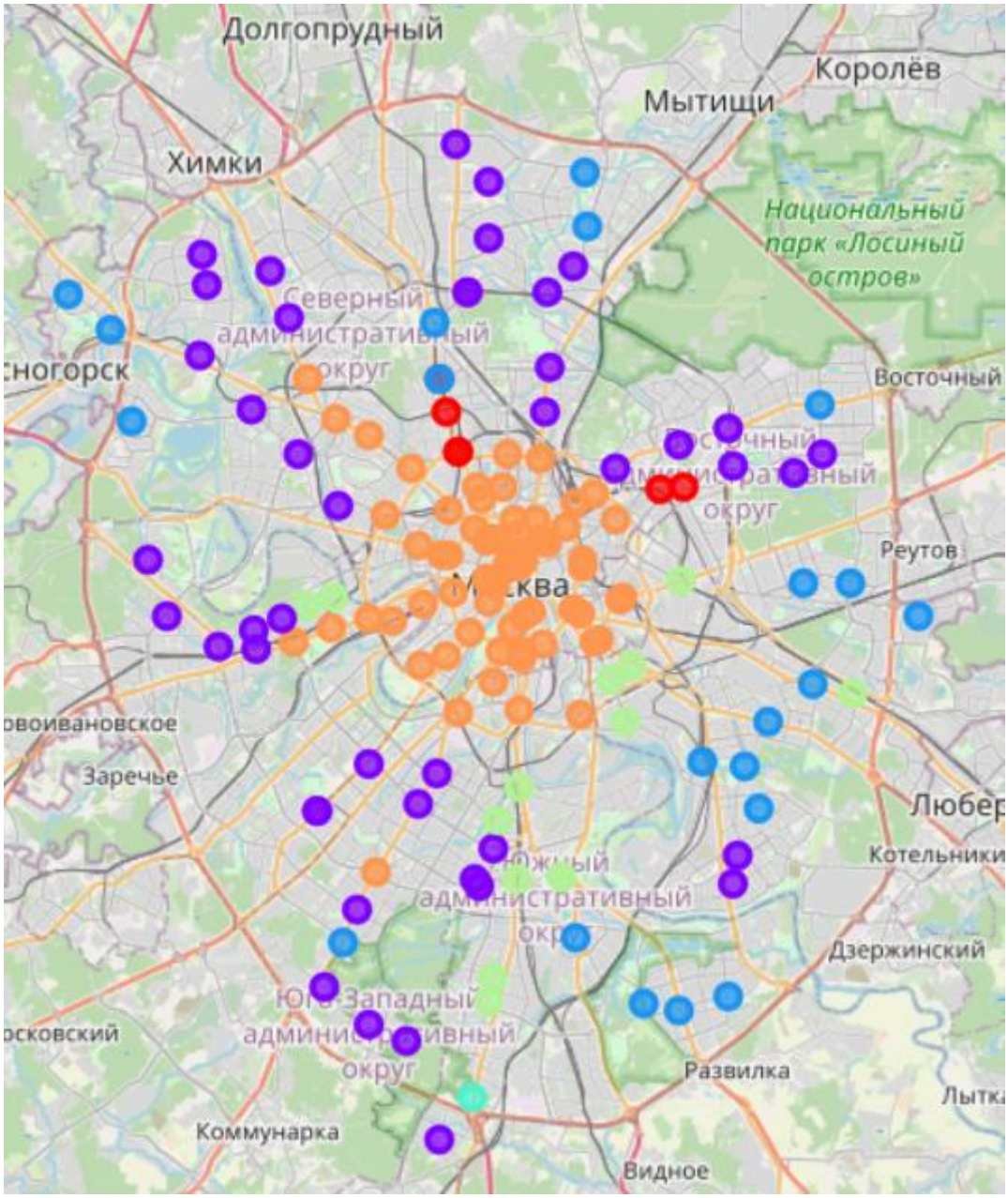
Before using the k-means method, you need to understand how many groups the data will be allocated, or, in other words, which one to use. We use the Elbow method to determine the optimal K value and the number of different clusters. It is visually seen that the optimal number of clusters is 6.



We use k-means cluster analysis for k = 6. Metro stations are distributed using the 6 clusters, the number of elements in each cluster is as follows:

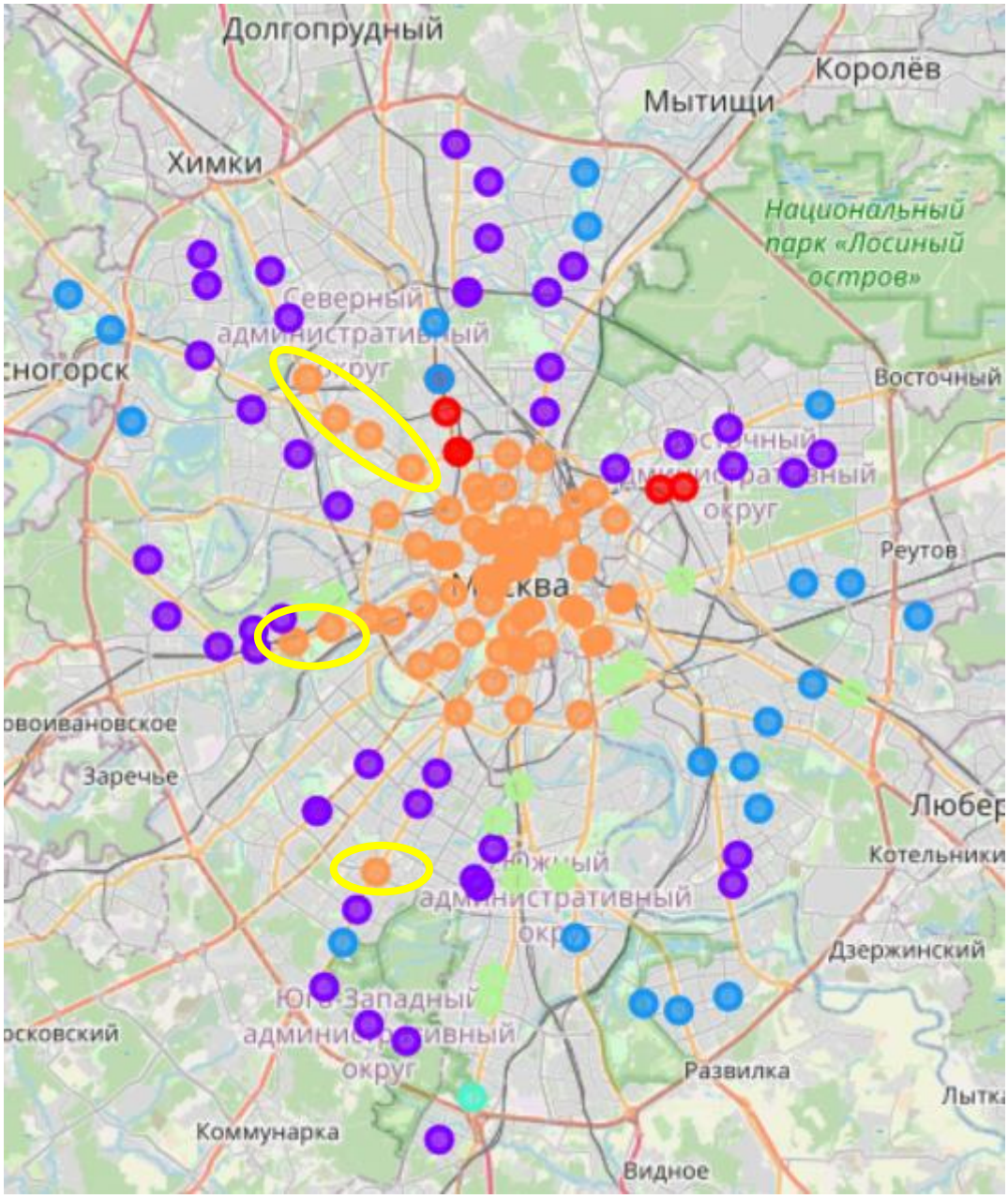
5	66
1	44
2	21
4	14
0	6
3	1

The resulting clusters of districts are shown on the map:



4. Result section

As man can see on the map, the historical center of the city is a single cluster. But also outside the center there are a number of districts that belong to the same cluster, they are highlighted on the map below.



The following areas are highlighted on the map:

• Metro Park Pobedy, Slavyansky Boulevard (Minskaya);

• Metro station Kaluzhskaya;

• Metro Dynamo, Airport, Sokol, Voikovskaya.

These districts are similar to the districts in the historical center of Moscow, but they are not in the center. According to the analysis results, they are the best suited for opening an art cafe. Also, if in the future the art cafe opens other sites, it will be optimal to place them in the areas marked on the map.

5. Discussion section

In this section, we will consider the treated clusters of Moscow districts.

Cluster 1.

Areas near the historical center of the city, which also contain industrial centers and sports facilities.

Cluster 2.

Sleeping areas with residential buildings with developed social infrastructure.

Cluster 3.

Industrial areas with poorly developed residential development and infrastructure.

Cluster 4.

Former industrial areas, in which intensive residential development has been carried out in the past 10-15 years, but the infrastructure is still underdeveloped.

Cluster 5.

An area on the outskirts of the city with poor residential development and the absence of industrial enterprises.

Cluster 6.

The historic center of the city, where social and cultural life is concentrated, as well as areas outside the center with a large number of theaters, museums, galleries, etc.

6. Conclusion

This paper considers the problem of determining the optimal area for opening an art cafe in Moscow. At the same time, the conditions of being not in the city center, but close to the metro and in an area similar to the city center, had to be met. To solve this problem, the districts of Moscow were analyzed using data from the Foursquare API and grouped into clusters, after which it was concluded that the following metro stations are the best places to open an art cafe:

- Metro Park Pobedy, Slavyansky Boulevard (Minskaya);
- Metro station Kaluzhskaya;
- Metro Dynamo, Airport, Sokol, Voikovskaya.