MINSK CITY GUIDE

Siarhei Fialka
B E L A R U S
February 18, 2020

Introduction

Let’s consider the case when somebody come to Minsk (BELARUS) and not familiar with the city. This person could be travelling or could come for a business and he/she would like to have a good time during this stay. If this person doesn’t have lot of free time and doesn’t want to spend lot of money, then he/she certainly needs some general recommendations about places in Minsk, which are easy to reach and have lot of interesting venues, where no one will get bored.

In our case, metro is a good choice for transportation. The Minsk metro is the most important part of urban transport system. It is low-cost high-capacity public transport, which is a safer and faster alternative to other public transport means.

Purpose of this project is to recommend some metro stations in Minsk, where in 5 minutes walking distance there is lot of shops, food places, attractions etc.

Data

At first, a location of metro stations across the city is required. Foursquare can be used to obtain geographical coordinates of Minsk Metro Stations through the API Endpoint "search". Because of Foursquare Regular Call Limit (up to 50 results for ‘search’ call) it is not possible to make search across all city at once, therefore search has to be done by neighborhoods (districts). Dataset of the Minsk neighborhoods can be scraped from multiple web-sources.

Next step is to explore 500 meters area around each metro station and look for venues like shops, food places, attractions etc. Foursquare can be used to obtain venues location and additional information through the API Endpoint "explore". Based on analysis of this data it is possible to make estimation which area is better to visit.

Methodology

Raw dataset of metro stations has 367 rows, where 204 rows are unique. The Minsk metro has 29 stations (https://en.wikipedia.org/wiki/List_of_Minsk_Metro_stations). Cleaned and formatted dataset has 28 rows because two stations are connected and too close to each other, therefore they can be joined.

In our case, attractiveness of each metro station depends on kind of venues around and how many of them there. In other words, the main characteristics of the data are venue category and quantity of venues of one category.

From Foursquare I got a dataset of venues that are around metro stations within a radius of 500 meters. Obtained dataset includes 788 venues which correspond to 193 unique categories. Most of these categories very similar to each other, therefore it is better to group venues by more general categories. From “Venue icon” feature I extracted 8 general categories: 'shops', 'food', 'arts_entertainment', 'travel', 'parks_outdoors', 'building', 'nightlife', 'education'.

Venues of such general categories as 'education', 'building', 'travel' were removed from dataset, because 'education' category is not correspond to our case at all, 'travel' mostly contains transport infrastructure and hotels, 'building' mostly contains gym.

To gain better understanding of the data set I grouped venues by general categories and plotted horizontal bar graph (Figure 1).

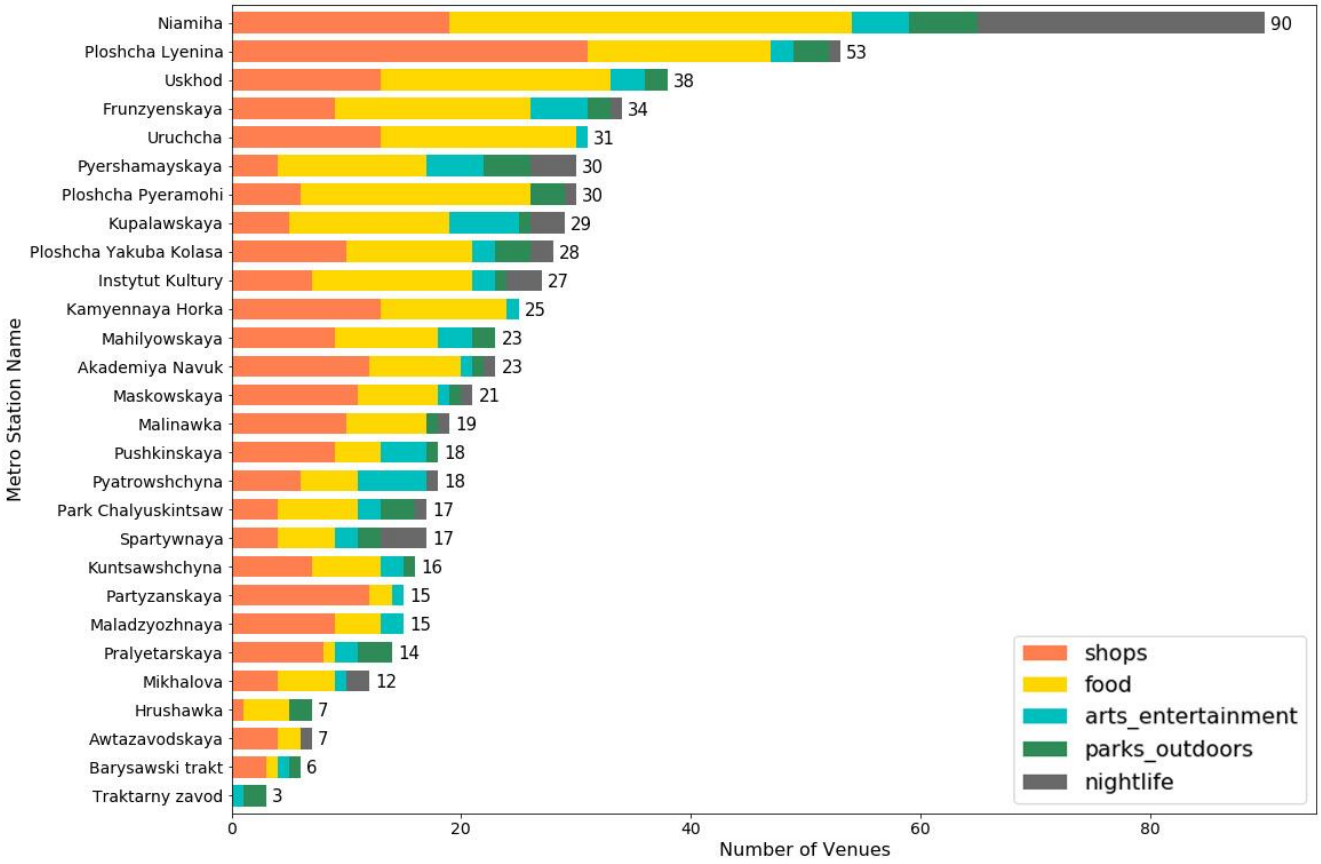

Figure 1. Rating of Minsk Metro Stations by total number of found venues in the 500m area around

K-mean clustering algorithm was applied to see groups of the metro stations that have similar characteristics. This algorithm was chosen because Density-based clustering algorithm define the most important stations and the least important stations as outliers thereby mix them; Hierarchical clustering algorithm didn't give clear result.

I included total amount of venues with increased weight for 'arts_entertainment' and 'parks_outdoors' categories in dataset for clustering, because such venues more attractive in average in comparison with other categories. Thereafter the dataset was normalized.

To choose right number of clusters, I plotted dependence of distances of samples to their closest cluster center ("inertia_ attribute") on number of clusters (Figure 2).

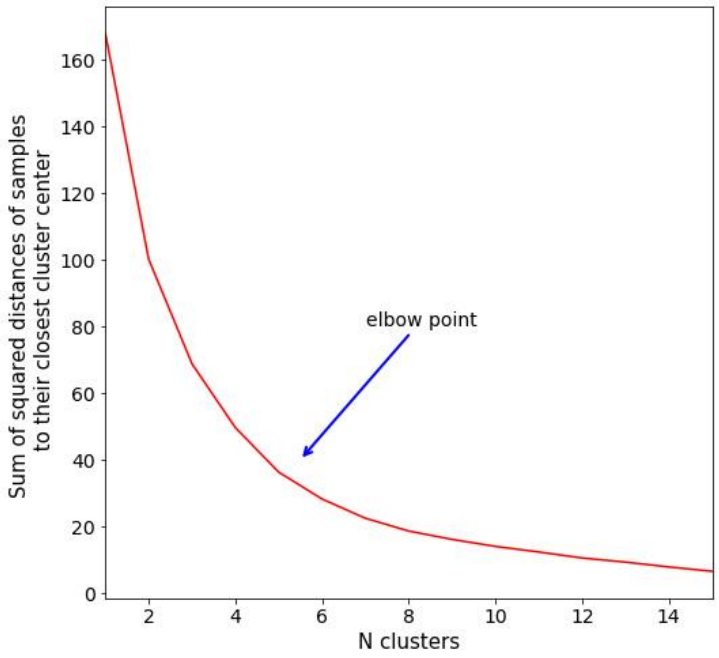

Figure 2. Choosing number of clusters for K-mean algorithm

Eventually, I have chosen 5 clusters for K-mean algorithm. We can check the centroid values by averaging the features in each cluster. Result of such calculation represented in Table 1.

Table 1. Averaged features in each cluster

Cluster	shops	food	arts_entertainment	parks_outdoors	nightlife
2	19.000000	35.000000	5.000000	6.000000	25.000000
4	36.000000	16.000000	2.000000	3.000000	1.000000
1	6.666667	14.666667	5.666667	2.000000	3.000000
3	8.083333	9.500000	2.250000	1.583333	1.083333
0	7.454545	5.636364	0.818182	0.636364	0.727273

Results

I used the Folium library to visualize the Minsk metro stations and their clusters (Figure 3). Each station illustrated by Minsk metro logo. Radius of circles corresponds to 500m in map scale. Color of circles corresponds to the different clusters. By clicking on the circles we can see name of the corresponding metro station.

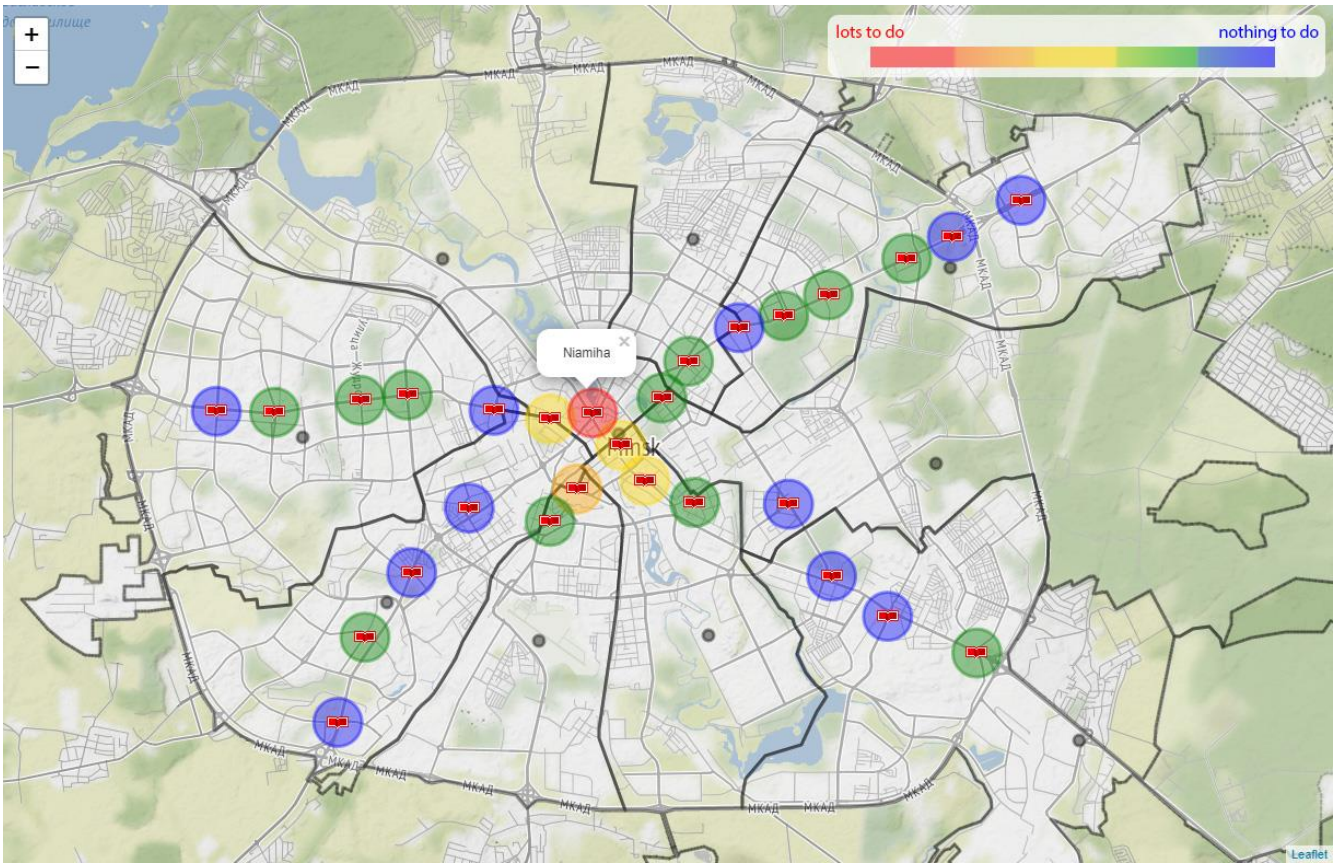

Figure 3. Map of Minsk supplemented by information about metro stations

Discussion

According to obtained results, the most interesting 500m areas around metro stations are in the city center.

The Niamiha station has the most interesting area around and is a must visit place. There are plenty of different venues, so no one will get bored there.

If the person is already familiar with area around the Niamiha station, then he/she can go to the Ploshcha Lyenina station, where are the most amount of shops in 500m area around, or he/she can go to one of the yellow cluster stations, where the most amount of arts-entertainment venues in 5 minutes walking distance.

In our case, blue cluster stations should be avoided, because it will be difficult to find what to do there.

Conclusion

In this project, I widely used ‘pandas’ library to perform data analysis. The **Foursquare** API was used here for different purposes: to search for a specific type of venues, to explore a geographical location, to explore a particular venue. K-means clustering was used for metro station segmentation. Also, Folium and matplotlib libraries were used to visualize the results.

In spite of rude data analysis, obtained results are consistent with tripadvisor.com results and the well-known concept - city center is the most interesting place to visit. To improve project, venues size, rating and number of tips should be taken into account.