Classification of Moscow Metro stations using Foursquare data

Introduction

Moscow Metro has 264 stations and is one of the largest public transit systems in the world. It is used by more than 6 million people daily.

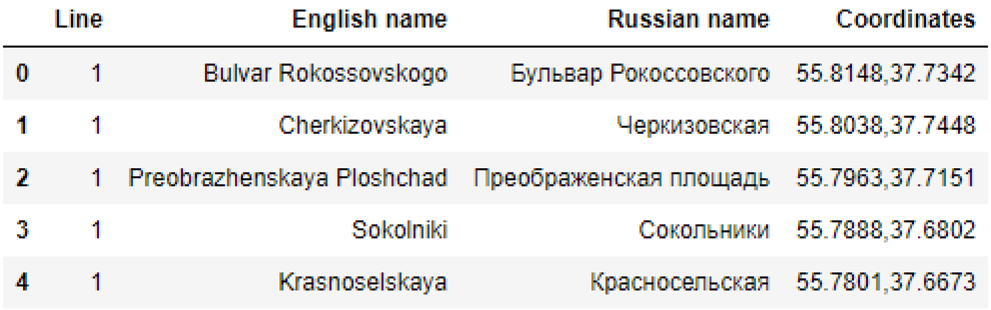
For this project, we want to look at the neighborhoods surrounding metro stations and classify them. Some neighborhoods are mostly residential, others have more business or commercial spaces surrounding them. The venues closest to a station determine why and how people use it. E.g. if there are no professional places in a neighborhood its residents are likely to travel to other areas for work. This creates daily migrations of people.

By analyzing this data we can classify stations by their primary usage. This data can be useful for city planners to determine where from and where to people are most likely to travel for work and leisure, plan further extension of the network and find places for new development.

Data

We'll need data on the location of stations and on the venues closest to them.

1. List of stations and their geographical coordinates — scraped from this Wikipedia page.



	Line	English name	Russian name	Coordinates
0	1	Bulvar Rokossovskogo	Бульвар Рокоссовского	55.8148,37.7342
1	1	Cherkizovskaya	Черкизовская	55.8038,37.7448
2	1	Preobrazhenskaya Ploshchad	Преображенская площадь	55.7963,37.7151
3	1	Sokolniki	Сокольники	55.7888,37.6802
4	1	Krasnoselskaya	Красносельская	55.7801,37.6673

Figure 1. Stations data

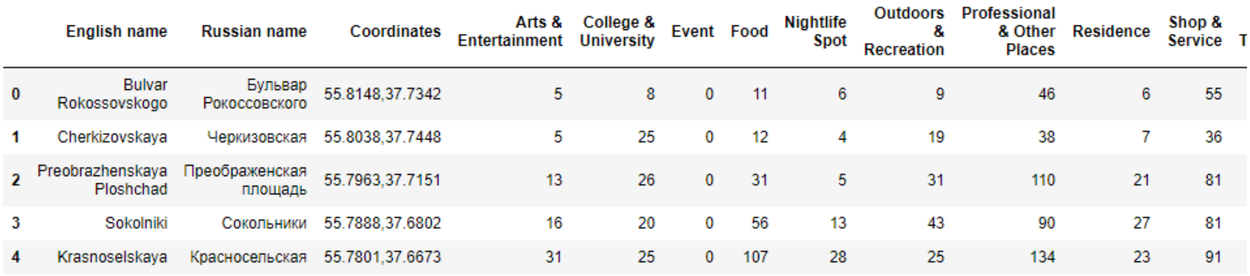
2. Foursquare API to explore venue types surrounding each station. Foursquare outlines these high-level venue categories with more sub-categories.
   - Arts & Entertainment (4d4b7104d754a06370d81259)
   - College & University (4d4b7105d754a06372d81259)
   - Event (4d4b7105d754a06373d81259)
   - Food (4d4b7105d754a06374d81259)
   - Nightlife Spot (4d4b7105d754a06376d81259)
   - Outdoors & Recreation (4d4b7105d754a06377d81259)
   - Professional & Other Places (4d4b7105d754a06375d81259)
   - Residence (4e67e38e036454776db1fb3a)
   - Shop & Service (4d4b7105d754a06378d81259)
   - Travel & Transport (4d4b7105d754a06379d81259)

We'll be querying the number of venues in each category in a 1000m radius around each station. This radius was chosen because 1000m is a reasonable walking distance.

Methodology

We can use the Foursquare explore API with category ID to query the number of venues of each category in a specific radius. The response contains a totalResults value for the specified coordinates, radius and category.

We've obtained this data for each station. The full csv is available on Github.



	English name	Russian name	Coordinates	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	T
0	Bulvar Rokossovskogo	Бульвар Рокоссовского	55.8148,37.7342	5	8	0	11	6	9	46	6	55	
1	Cherkizovskaya	Черкизовская	55.8038,37.7448	5	25	0	12	4	19	38	7	36	
2	Preobrazhenskaya Ploshchad	Преображенская площадь	55.7963,37.7151	13	26	0	31	5	31	110	21	81	
3	Sokolniki	Сокольники	55.7888,37.6802	16	20	0	56	13	43	90	27	81	
4	Krasnoselskaya	Красносельская	55.7801,37.6673	31	25	0	107	28	25	134	23	91	

Figure 2. Count of venues of each category in a 1000m radius for each station

Exploratory analysis & basic cleanup

Let's look at the data. We can see for example that Turgenevskaya station has the highest number of Professional & Other Places (192) while Belokamennaya station has 0.

Let's display the number of venues as boxplots (showing the average count, spread and outliers).

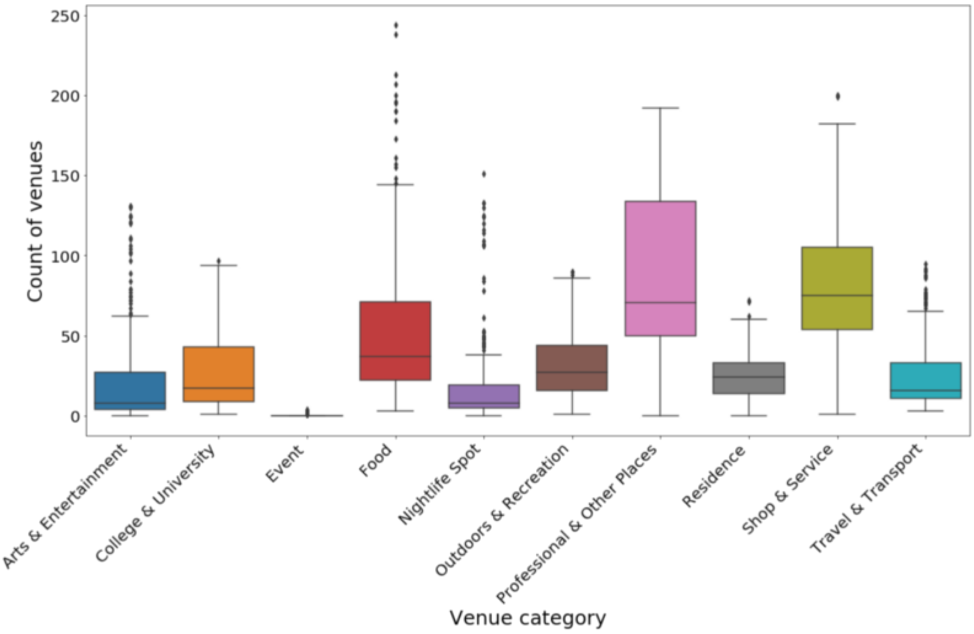


Figure 3. Boxplots of number of venues in each category

We can see that the most frequent venue categories are Food, Shop & Service and Professional & Other Places. Event has very little data, so we'll discard it.

Data preparation

Let's normalize the data using min-max scaling (scale count of venues from 0 to 1 where 0 is the lowest value in a set and 1 is highest). This both normalizes the data and provides an easy to interpret score at the same time. The scaled diagram looks like this:

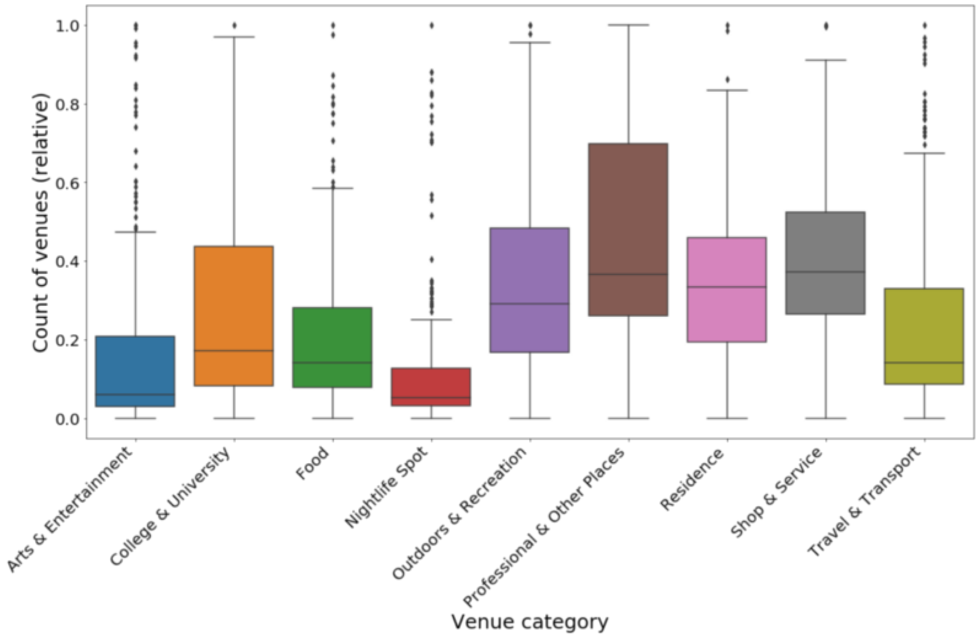


Figure 4. Boxplots of scaled number of venues in each category

Clustering

We'll be using k-means clustering. These were the preliminary results with different number of clusters:

- 2 clusters show the uptown/downtown divide
- 3 clusters add clustering within the downtown
- 4 clusters also identify neighborhoods with very low number of venues
- 5 and more clusters are difficult to interpret

For the final analysis let's settle on 4 clusters (0 to 3). Let's visualize the clusters profiles using boxplots.

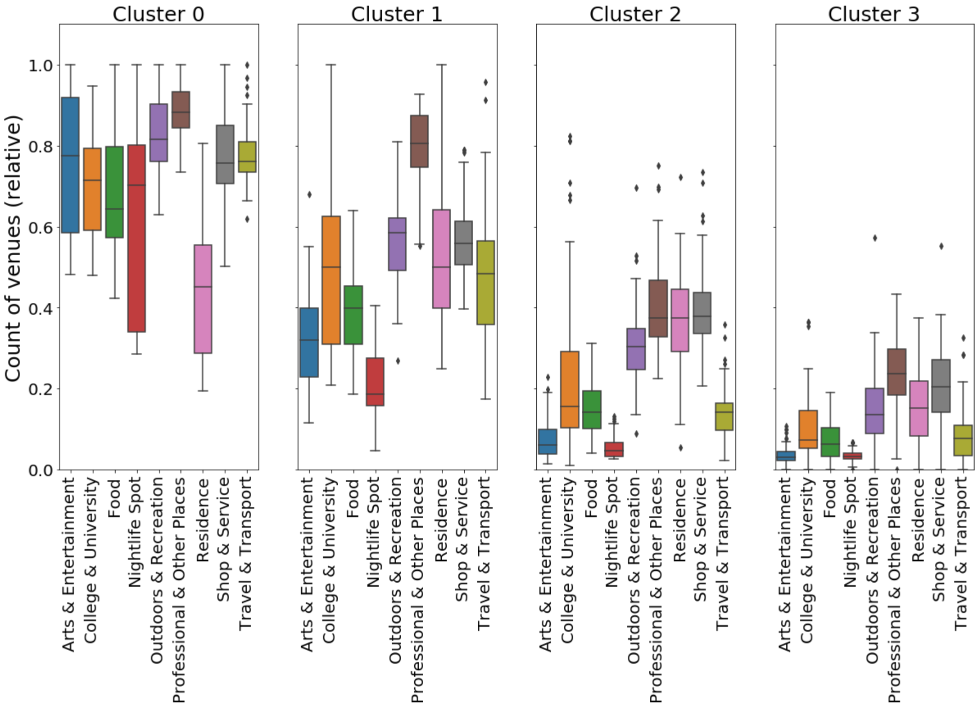


Figure 5. Clusters and their relative count of venues

And plot them on a map (full interactive map available at https://theptyza.github.io/map_moscow_metro_foursquare/map/).

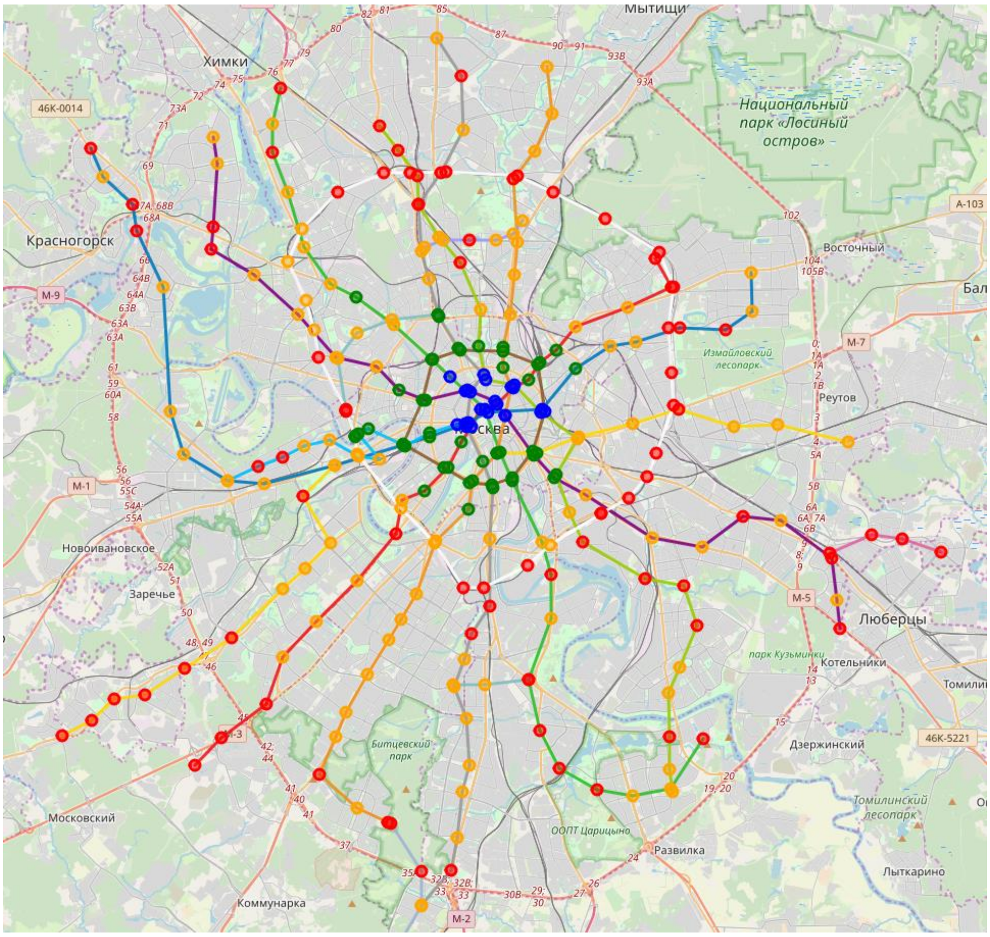


Figure 6. Clusters map. Cluster 0 is Blue, 1 is Green, 2 is Yellow and 3 is Red.

For each station we'll display top 3 venue categories and their 0 to 1 scores for this category.

Results

Here is how we can characterize the clusters by looking at venue scores

- Cluster 0 (Blue) has consistently high scores for all venue categories. This is the most diversely developed part of the city
- Cluster 1 (Green) has highest marks for Professional & Other Places. This is the business part of the city.
- Cluster 2 (Orange) has lower marks with best scores in Profiessional, Residence and Shop & Service.
- Cluster 3 (Red) has low marks across the board. These appear to be underdeveloped areas.

Plotting the clusters on a map shows us that

- Cluster 0 is the oldest and central part of the city
- Cluster 1 is also downtown. Most of these stations are inside or near the Circle Line and have excellent transit accessibility.
- Clusters 2 and 3 aren't so clearly geographically distributed. Cluster 3 areas tend to be at the outskirts but some are more centrally located.

Some stations were classified as Cluster 4 despite being more centrally and accessibly located. This could be a legacy of the “Rust Belt” of closed and abandoned factories (https://www.vesti.ru/doc.html?id=3030139&cid=4441). Many stations of the recently opened Moscow Central Circle railway fall into this category. These are prime areas for business and residential development.

Discussion

To be fair, Foursquare data isn’t all-encompassing. The highest number of venues are in the Food and Shop & Service categories. Data doesn’t take into account a venue’s size (e.g. a university building attracts a lot more people that a hot dog stand – each of them is still one Foursquare “venue”).

Conclusion

Foursquare data is limited but can provide insights into a city’s development. This data could be combined with other sources (e.g. city data on number of residents) to provide more accurate results.