2. Data acquisition and cleaning

2.1 Data sources

A list of Moscow Metro stations with coordinates can be found on Wikipedia here. Passenger traffic data can be found here. Additionally, we use Foursquare to list venues near each metro station.

2.2 Data cleaning

Data downloaded or scraped from multiple sources was combined into one table. First, I took the table with coordinates of metro stations and data with the values of average daily traffic at metro stations for 2017 from Wikipedia. Then I corrected some inaccuracies and typos and merged into one data frame. The stations with the type ‘surface’ and ‘elevated, open’ are discarded. These are the stations of Moscow Monorail and Moscow Central Circle, which are very different from regular stations.

Stations, located closer than 250 meters to each other are essentially one station. They just have different names on different lines. Therefore, I combine them into one station, and the passenger traffic values for these stations are added up (for several stations the traffic values coincide completely, in this case the value is taken).

Using the Foursquare service, I get a list of venues located within the radius of 500 meters for each metro station. Then, I group some categories. All categories ending in “Restaurant” and “Pizza Place” are renamed to “Restaurant” to analyze the restaurants as a whole piece. I also group categories ending in ‘Bar’ or ‘Pub' into one category 'Bar / Pub'.

Thus, I get the final table consisting of 178 entries for metro stations. It contains information on the coordinates, passenger traffic (for 119 stations) and the number of venues of each category near each station.