2. Data acquisition and cleaning

2.1 Data sources

In order to perform the study described above the appropriate datasets have to be obtained, cleaned and modified. The analysis is based on the following data:

1. List of capital construction objects in Moscow;
2. Moscow Metro stations geolocation.

First dataset consists of the full list of ongoing projects currently under development in Moscow. The data are available on the official city administration website as a JSON file (https://data.mos.ru/opendata/7703742961-obekty-kapitalnogo-stroitelstva/passport?versionNumber=1&releaseNumber=886). Among others the fields of interests are property identification number, property region, type, and its geolocation. The full list of objects includes offices and residential facilities.

The geolocation data of metro stations are obtained using Foursquare API. Moscow metro stations are common meeting places, so the Foursquare provides a convenient way of getting their coordinates.

2.2 Data cleaning and manipulation

Both sources has been checked for empty data and no errors were found. The study was focused on residential real estate properties and therefore the construction objects dataset has been filtered in order to keep apartment buildings only. The dataset produced by city government used the Cyrillic encoding (the official paperwork encoding) for the most of the dataset's features. In order to be able to filter by values correctly one has to be applied 'Windows-1251' encoding.

There is a specific Moscow region Zelenograd located far outside the city boundary. This region was removed from the analysis since there was no metro station around. The feature selection for the analysis was mainly caused by the geolocation nature of the study, therefore the coordinates of the object, main function, administrative area it belongs to and id were kept. The coordinate column was split in two (latitude and longitude) columns. Finally, the dataset was represented by 453 objects.

The metro station location dataset was obtained by using ‘`search`’ endpoint of the Foursquare API around the city center. The data in a JSON format were requested and flatten. The shape of the initial data frame was (50 x 18). Despite the fact that Moscow metro system includes about three hundred stations the request output is limited by 50 stations. Only three features were kept for analysis containing station name and its latitude and longitude.