Georgy Golovanov

Clustering new real estate properties around metro stations

Jun 7, 2020

Background

Modern cities constantly grow. New properties are being built at any time and require the urban infrastructure to be consistent.

Moscow being a modern permanently growing city is chosen as a subject for the analysis.

The analysis is focused on the real estate properties currently being under construction by clustering them around metro stations to reveal perspective areas.

The main purpose of the analysis is a study of new real estate developments by clustering them around metro stations.

Since the real estate properties development is growing faster than the city infrastructure the analysis can be useful for real estate marketing.

Data acquisition and cleaning

The analysis is based on the following data:

1. **List of capital construction objects in Moscow;**

— consists of the full list of ongoing projects under development in Moscow;

— The data are available on the official city administration website as a JSON file

— Feature selection: identification number, property region, type,

and its geolocation.

2. **Moscow Metro stations geolocation:**

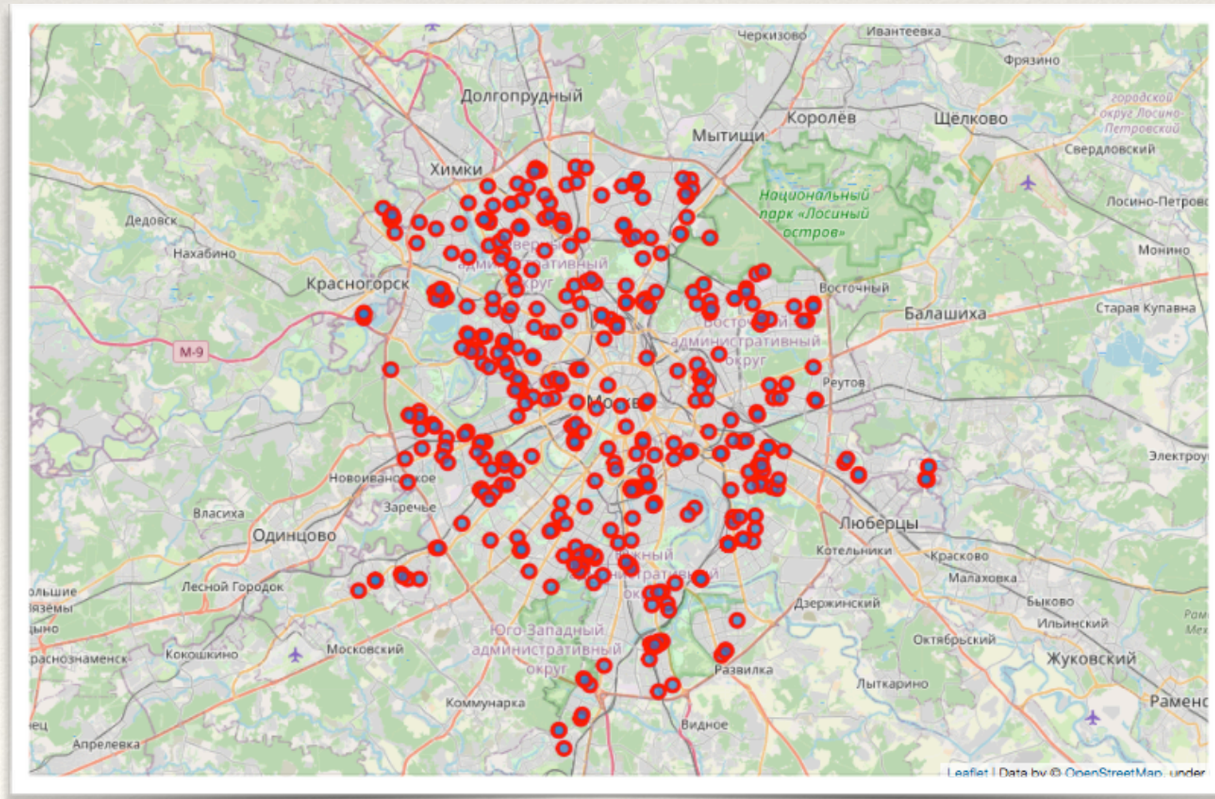
— The metro station location dataset was obtained by using '`search`' endpoint;

— Feature selection: coordinates of the object, main function, administrative area it

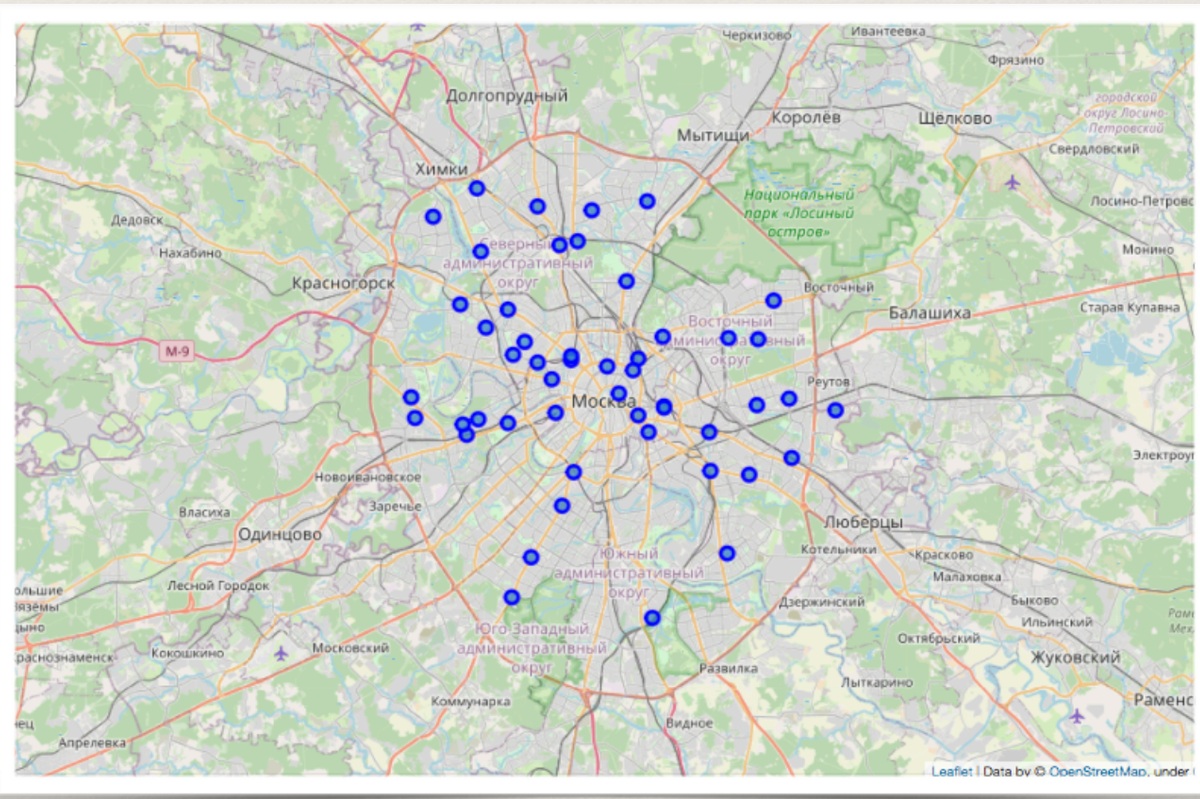
belongs to and global-id.

Data visualization

- New real estate facilities under construction



- Metro stations



Visualization is made using `folium` *library*

Methodology

The analysis is based on unsupervised machine learning algorithm **K-means**. The main goal is to form a cluster of developing buildings around a metro station. For that purpose we exploit two methods.

- **Method 1**: the clusters are formed by a single run of K-means algorithm with metro stations geolocation coordinates as fixed initial centroids. It allows to go through a list of metro stations and assign a clusters of buildings to each of them.
- **Method 2**: the clusters are formed by a series of K-means runs with random initial centroids and assign a closest metro station from a list to the found clusters' centroid.

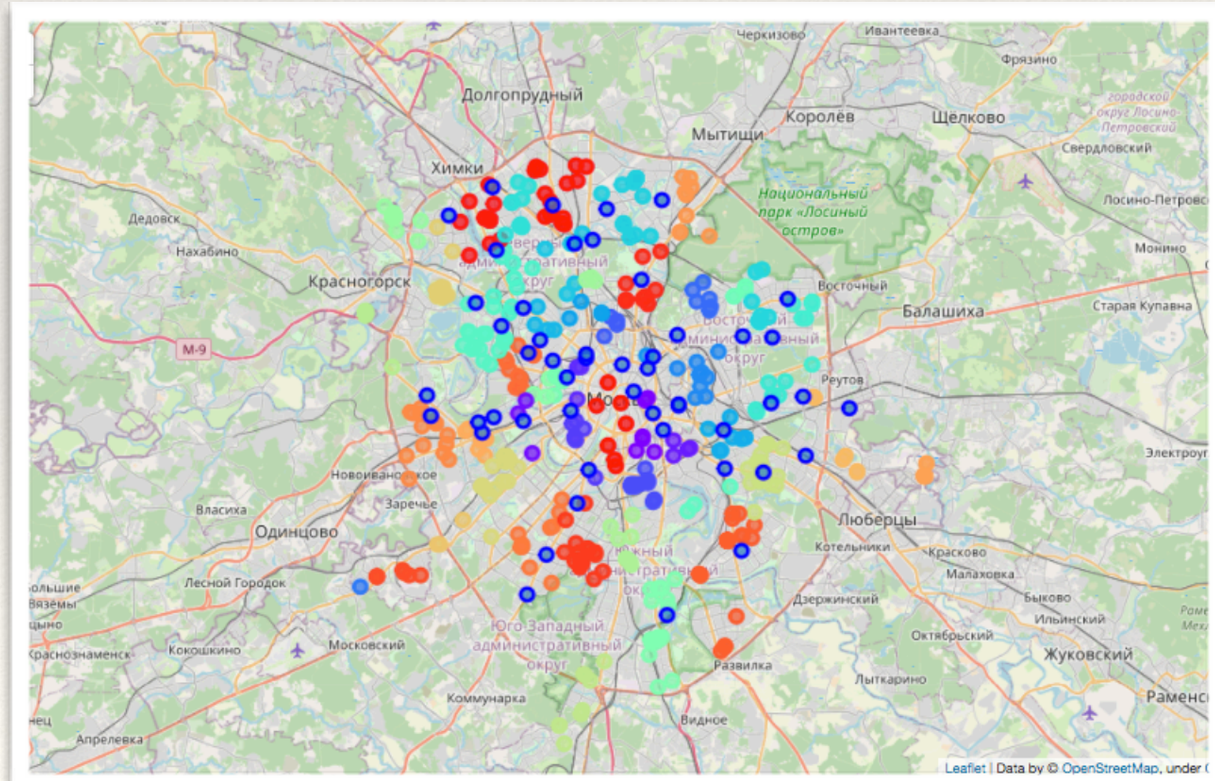
Both methods are used to analyze number of developing properties assigned to each cluster as well as parameters of developing buildings distribution within a cluster such as average distance and variance to the closest metro station

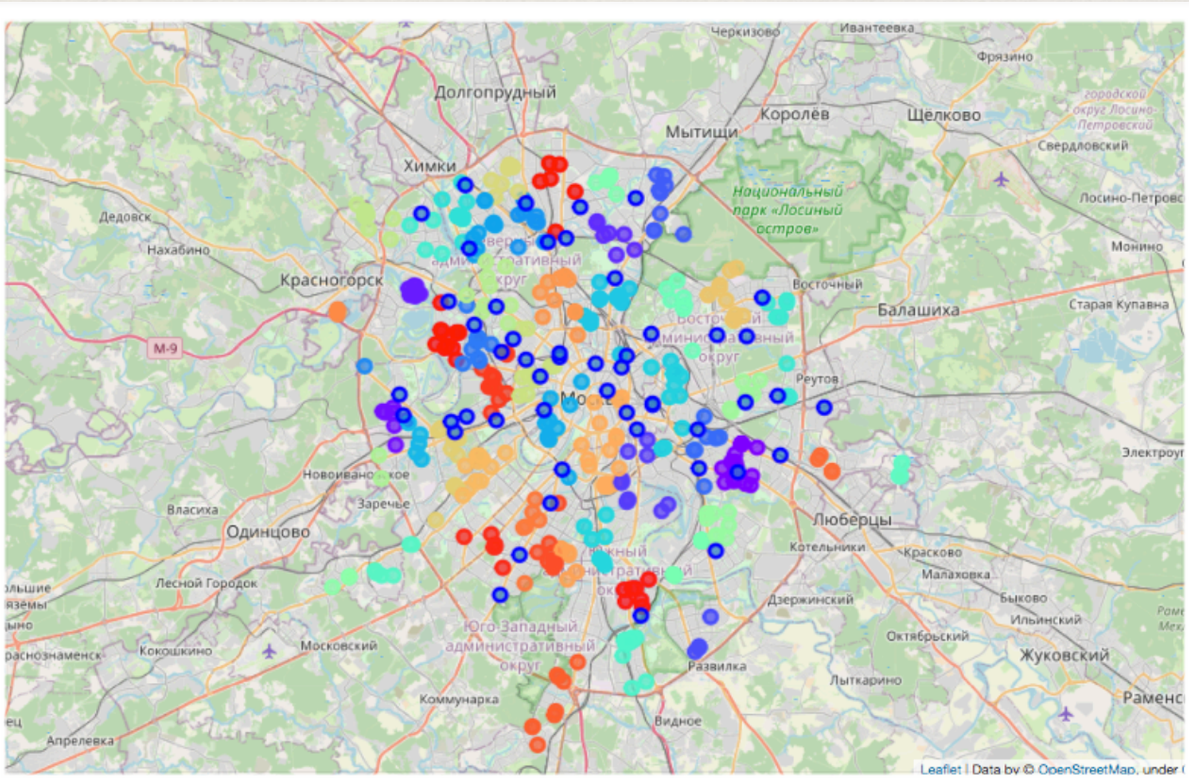
Clustering

- The clustering procedure is performed by `KMeans` class from `sklearn` library.
- *Method 1*: the centroid initialization is done by assigning metro stations latitude and longitude values.
- *Method 1*: each cluster is labeled with the same number as the metro station.
- *Method 2*: uses the same approach with the only difference of random centroids initialization, while the number of clusters is kept the same.
- *Method 2*: after formation of new clusters the geolocation coordinates of new centroids are available and therefore one can find the closest metro station using the **Foursquare API** request.

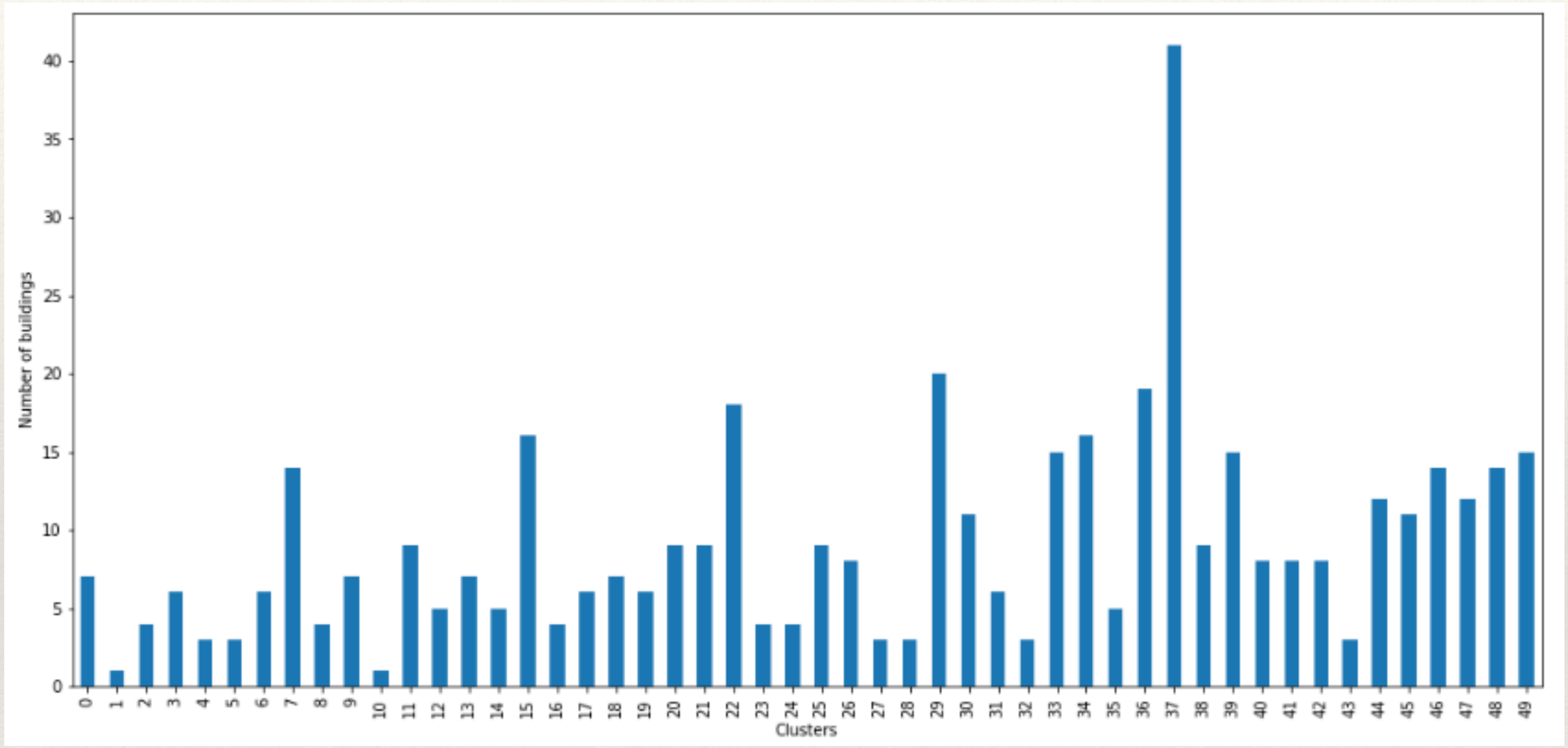
Clustering (cont'd)

- **Method 1**: metro stations coordinates as pre-fixed initial centroids
- **Method 2**: random initial centroids





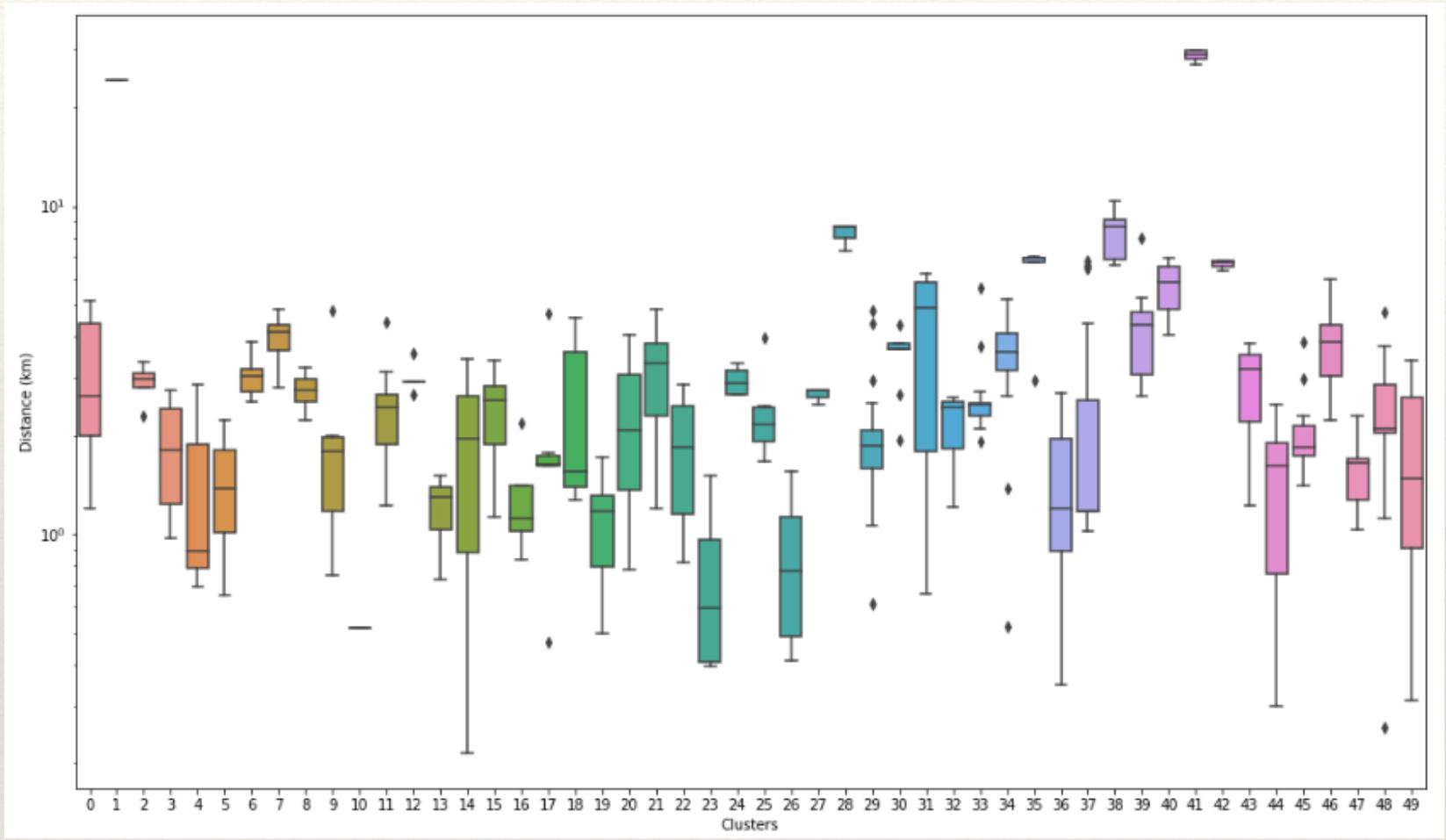
Results



- New real estate buildings within a cluster

- Average number of buildings per cluster: 9.06 ± 6.72

Results (cont'd)



Distance distribution of buildings within clusters

- The average distance doesn't exceed few kilometers and there are some clusters that have developing buildings within 1 km away from metro stations.

Discussion

The procedure has some advantages and downsides:

— A limitation: the algorithm *doesn't take into account closeness* of the adjacent metro stations. Therefore a clustering of adjacent metro stations into one can help in improvement of the algorithm.

— The methods used in the analysis also *suffer from the limitation of the number of stations* provided by Foursquare. Using the full list of stations will improve performance since it will correspond to the real coverage.

— In order to calculate distance between two points the *Haversine* distance was used. It is quite excessively within the current study ⇒ can be replace by *Euclidean* distance.

— A possible advantage of the analysis could be a usage of a list of projected metro stations that *opens a possibility to estimate values* of developing buildings in the near future.

Conclusions

The performed analysis is focused on the real estate properties currently being under construction by clustering them around metro stations.

The analysis shows that one can reveal perspective areas based on the calculated distance to a nearest metro station.

Since the real estate properties development is growing faster than the city infrastructure the analysis can be useful for real estate marketing needs.