Predicting the best area for a new construction building in New York

Konstantinos Bousios

February 26, 2019

1. Data acquisition and cleaning

1.1 Data sources

The value for buildings in New York can be found from Kaggle dataset here. This dataset, however, lacks data for more information about benefits in the neighbourhoods. For example, venues in neighbourhoods. Exploring more about the neighbourhoods of New York I used the dataset from here. Also, after processing from the above datasets, I will use Foursquare API for the Borough of New York which I find out from the analysis of the other two datasets, to explore the benefits of each neighbourhood. For example, our findings into the exploration of neighbourhoods, give us information about the category of each venue and the exact location in the neighbourhood of the appropriate borough.

1.2 Data cleaning

Data downloaded from multiple sources. The processing of data from kaggle done in IBM Watson Studio. From first dataset, we kept columns for the borough, value of buildings and information about the locations. After this processing, I cleared all the nan values and created the dataset with all the data I need for my research. I had chosen the Watson studio for the processing of data from the first dataset for the reason that the dataset was too big and I had a problem with the memory to run the whole procession in local. By this choice, I achieved to create one dataset in csv form and continue the project. There were several problems with the first dataset like I had a lot of missing values and I tried to find as many possible information in the missing values for the reason, I wanted to keep the dataset with a lot of information and not minimize it, with the conviction that it will give me better conclusions. After fixing these problems, my data are ready for the analysis. For the second dataset, I created columns with values of Borough, Neighbourhood, Latitude, and Longitude from New York. After data cleaning in the first dataset, there were 27,312 samples and 8 features in the data. In the second dataset, there were 306 samples and 4 features in the data. By completing the first analysis of the two datasets, it will be followed by API Foursquare by the import of new data, to make a deeper exploration in the problem and give us the best solution.