Capstone project – Report

Defining the problem

*Many people have to do relocations for their jobs or for their studies. Finding new places to live can be a tedious and challenging experience and many considerations are usually overlooked. People usually only broadly look where they would like to live and give a lot of attention to finding appartments that fit their costs and answer a few criteria.*

*These criteria may be whether it's close to work, whether there's a grocery store and school nearby. These considerations are of course only a few of many, perhaps, less important considerations. However, how could you possibly consider everything and still make a good decision? Many neigfhbourhoods have schools, grocery markets, coffee shops and bars. Comparing neighbourhoods while considering tens of venues in the near vicinity can therefore be an almost impossible job to do on your own.*

*One solution could be counceling to people who know the neighbourhoods, but their knowledge is also limited and usually costs a lot of money and trust. Therefore, a more robust and easily applicable method should be created to help families find their perfect neighbourhood to live in!*

## The Data I'll be using to solve this problem

##### *To this end, I'll be using data acquired from the Foursquare database and input from the user, which is a customer interested in moving to Manhattan and wants to get help finding the best Neighbourhood for his needs. To acquire the data acquired of the Foursquare database, we use the geolocational data belonging to New York City. This geolocational data consists of the names of the different neighbourhoods and their longitude and altitude values.*

##### *These coordinates can help us acquire from Foursquare certain user acquired data. For my purposes I'll be using data that describes the top 100 most popular venues in a typical neighbourhood. This popularity is calculated in the Foursquare application. Since this dataset contains the most popular venues in certain neighbourhoods, we can crosslink this dataset to a dataset containing the importance of each venue for a certain customer that wants to know the best neighbourhood for him specifically.*

##### *To this end, I created a list of all possible different venues found in all the neighbourhoods in Manhattan. For each different venue, I request an input from the user, which will rate it from 1 to 10, whether it is important for him to have it in the vicinity of 1km near to its future house. As many neighbourhoods have schools and grocery stores and coffee shops, this cross reference between the two datasets will allow for the calculation of the most fitting neighbourhoods and clustering them. Not only will the application give you the more fitting neighboorhouds, but as the neighboorhoods are expected to be a little clustered into certain areas of the city, this clustering is supposed to give certain flexibility in the choice of the user to find a correct neighbourhood.*

Methodology

*Business understanding:* As many people usually move houses at least a few times in their life, many people will find this a difficult task as many considerations have to be made, involving money, jobs, mobility and venues around. Money considerations and proximity are considerations that are usually looked at in high detail. Venues in the near vicinity of any interesting apartment is hard to observe as you’ll have to consider many different kinds of venues which all are of different relevance to each different person. Creating a tool that will help people finding the perfect neighborhood, venue-wise, is therefore a smart solution for people that want to be able to get smart insights within 10 minutes of using this application. As people usually invest hours, days and even weeks looking for apartments, devoting 10 minutes to also consider popular venues in the area is therefore energy-saving and adds a very important attribute to finding a new place to live!

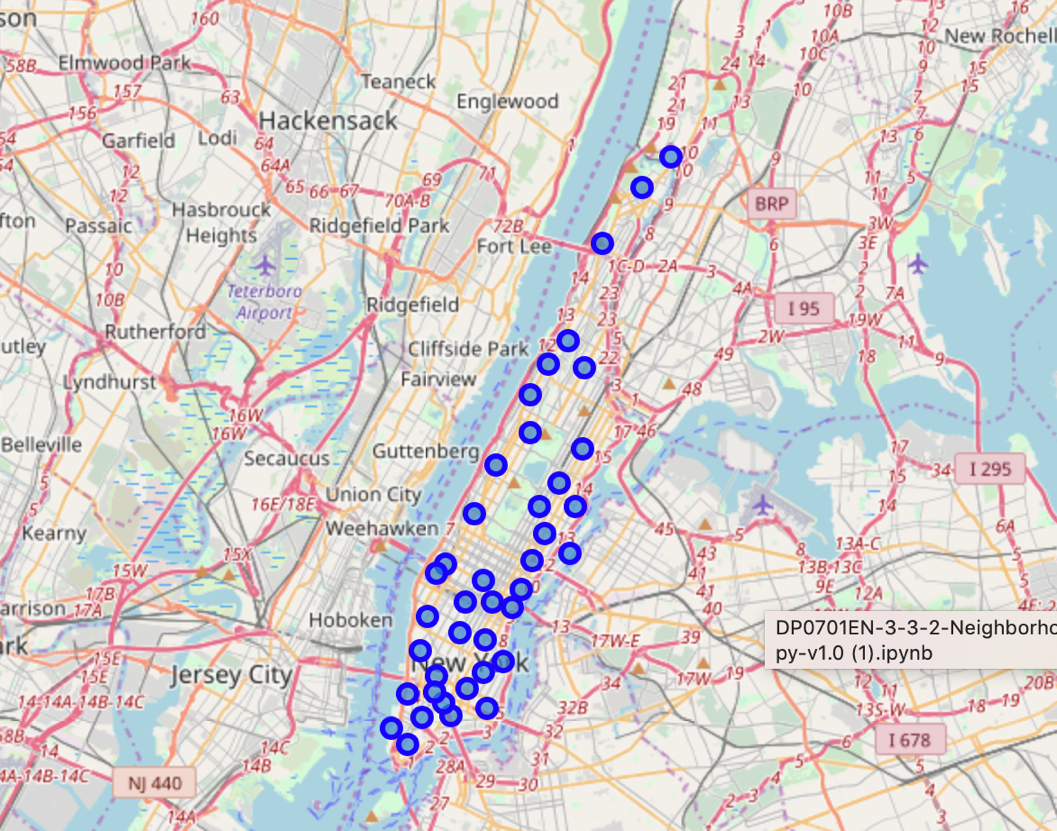
*Analytic approach and data considerations:*

* This notebook is supposed to get a city or borough of the users choice. In this case, the notebook uses the choice “Manhattan” as the borough. As the borough is chosen, the notebook looks for all different neighborhoods in the Manhattan borough, together with their coordinates, being longitude and altitude. These coordinates are important as they will help me achieve foursquare data belonging to each of these neighborhoods.
* Foursquare is a well known application in which users can upload their opinions on different venues around the world. Users can also follow the opinions of any of the other millions of users, therefore creating a rich database of venues, opinions and ratings. Foursquare allows developers to acquire some of this data by making calls to their database. This notebook uses these calls to acquire the top 100 venues for each neighborhood in the near vicinity of choice (this notebook chooses a radius of 1km). The top 100 venues are ranked according to popularity of these venues as calculated from user input. For further analysis, this notebook only relates to the top 20 venues only.
* These top 20 venues consist of restaurants, museums, barbershops and many other kinds of venues. Imagine a customer that does not like museums at all, but enjoys restaurants a lot and goes to the barber shop at least ones every two weeks. This preference for specific venues can be quantified by acquiring user input of the relevance of each specific venue. When combining this biased preference for different venues and the popularity of the venues (marked 1 to 20, according to the popularity of Foursquare users), we acquire a weighed dataset that can be used to find the perfect neighborhood or cluster of neighborhoods for a specific user. These weighed calculations are made by multiplying the rating (1-10) of the user multiplied by popularity rating (0-1). This weighed dataset is then analyzed by kmeans clustering, indicating neighborhoods that are similar in their weighed importance to the user.
* The user gets a map of all neighborhoods and to which clusters they belong. This map is supposed to help them decide where to look for apartments or to give them another consideration when looking for an apartment. More so, a new dataset is created which has the mean values of all weighed preferences for each neighborhood, which can help the user help to find the best neighborhood in each different cluster.
* Future plans include getting feedback back from the users that defines their level of happiness with the data created by the notebook. This can help the notebook acquire patterns in all data and make more sophisticated decisions in the future for users that have similar preferences as other users. All this data can therefore be used to create a model, allowing for additional machine learning approaches to be used.

Results

As the user defines which borough or city he’s interested of living in, all neighborhoods are found for this specific city or borough and the coordinates are found:

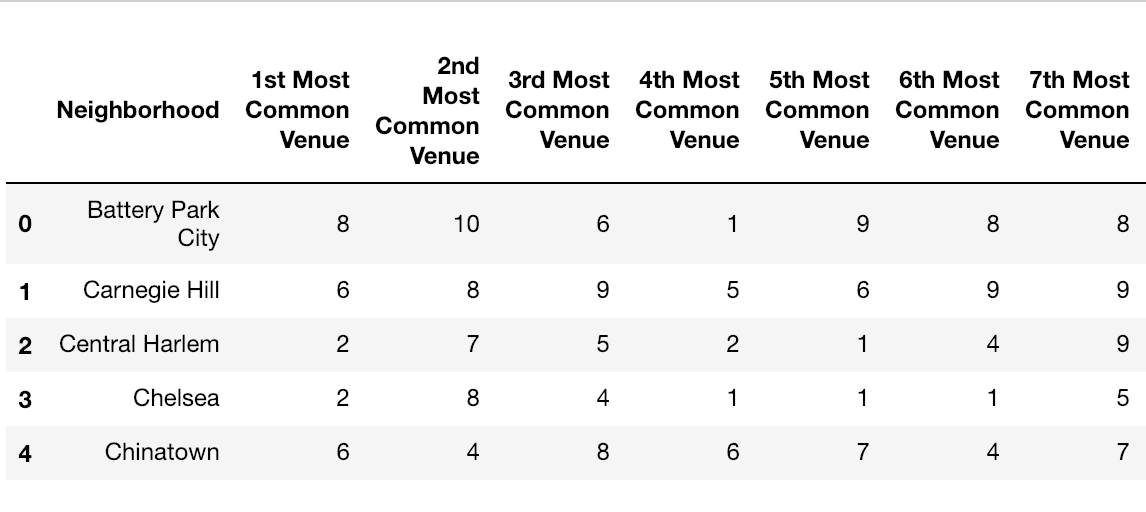
Afterwards, all the neighborhoods are mapped on a folium map. This visualization can be seen in the next figure.



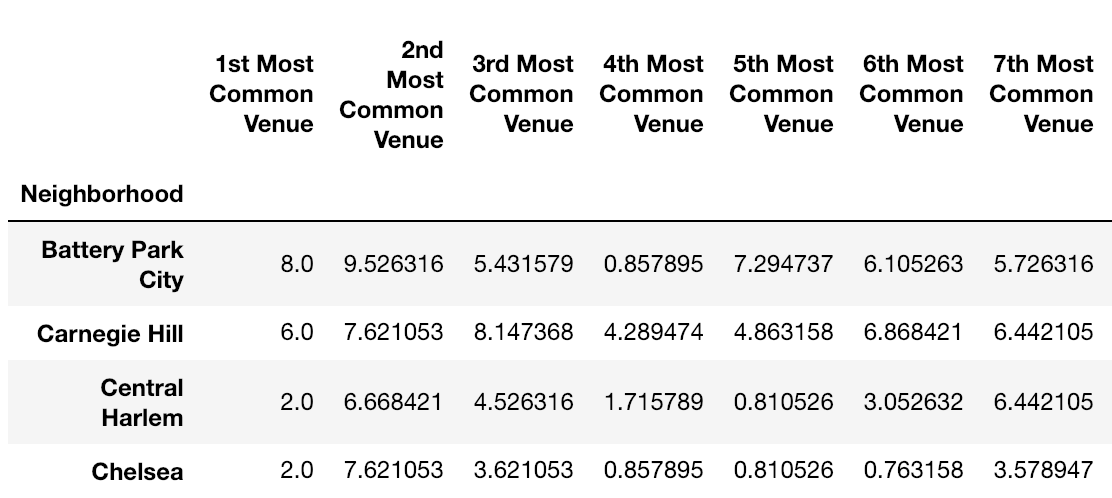
As the foursquare database is called for to acquire the top 100 venues found in each neighborhoods, the top 20 neighborhoods are noted for each different neighborhood, as can be seen in the table that’s created next:



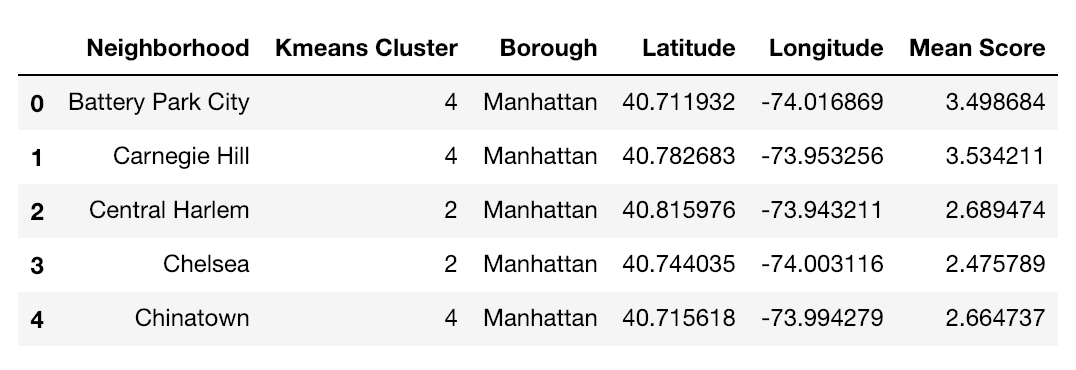
For each venue, the users rating is acquired as input to rate the importance of each venue, which will therefore change the above table to the one below, now holding the ratings for each venue:



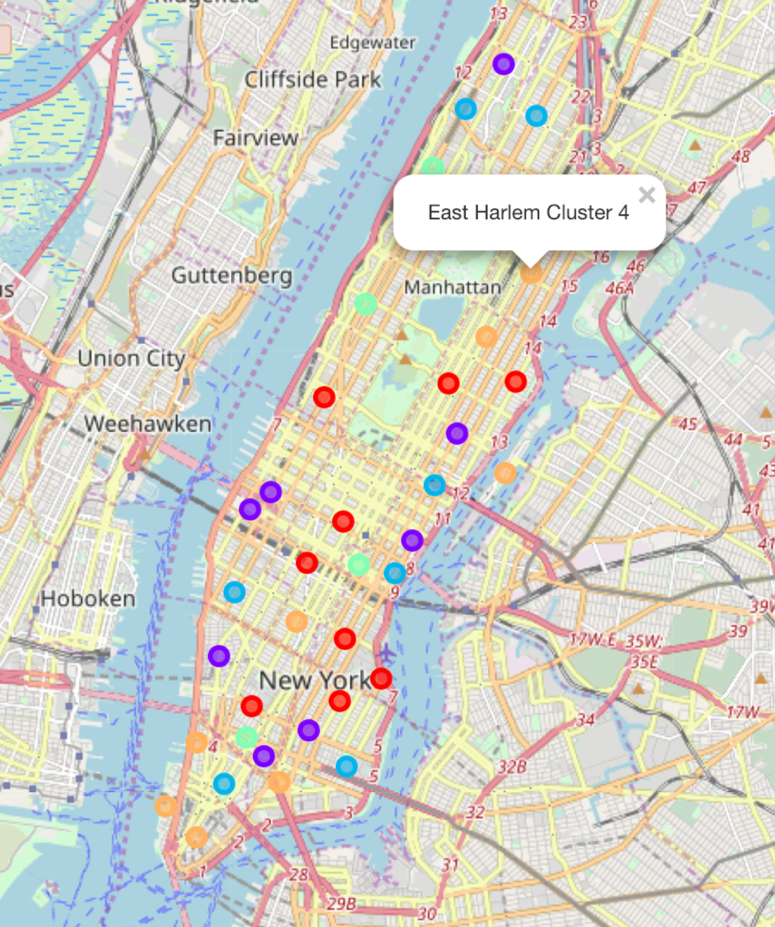
This table is then multiplied with a array that goes from 1 to 0 (as popularity decreases), resulting in the next dataset:

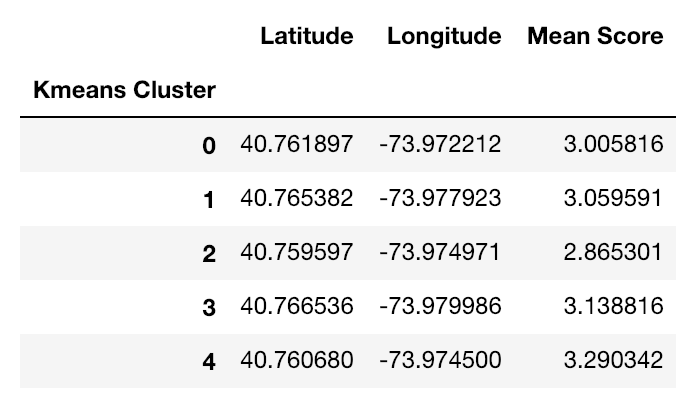


This dataset is then used to cluster the neighborhoods according to their weighed importance. Kmeans clustering with 5 clusters was used, but this can change according to feedback and modeling in future use. The resulting dataframe is created:



These neighborhoods are then mapped again, but this time with the clusters color coded into a folium map. Another table is shown to the user to see which cluster has the highest preference mean, as to understand which clusters fit the most to the user:





Discussion

This notebook is made to help people make the right choice of finding a living. To this end, foursquare data is used as it indicates which venues are most popular in certain neighborhoods. As this approach is a relatively good idea of helping users finding good neighborhoods, the foursquare data is biased towards venues that attract feedback from foursquare users, being more prominently venues made for leisure and less associated with important living considerations, such as laundry places, grocery stores or cheap shops in the area. Even though also these venues are rated in the foursquare application, they are less popular as foursquare users tend to rate less how their supermarket experience was.

This is a serious loophole in the notebook and needs to be addressed if this plan is to become a business offer. This can be done by adding additional importance weights according to the specific family situation of the user, whilst also considering other factors. These additional importance weights might therefore cancel any foursquare bias and give a more accurate recommendation to future users.

Another point of discussion is the input required from the user. Even though only 10 minutes are likely to be needed to rate all the different venues found in the neighborhoods of chosen cities or boroughs, this rating can be simplified by categorizing different venues into categories. For example, if restaurants are not important for users, there’s no need to ask for the rating of 30 different restaurants that only differ in the kitchen.

Many parts of this notebook can therefore be optimized if found necessary for any future business plan.

Conclusion

This notebook helps people to find neighborhoods that fit their preferences of known venues. Considering popular venues is of great importance when looking for new apartments to move to and this consideration is usually greatly overlooked. As it is hard to consider all venues in different neighborhoods, this application takes the ratings for different venues, supplied by the user, and therefore calculates which neighborhoods and clusters of neighborhoods fits better the preferences of the specific user.