2 Data

According to the prior Labs in the IBM data-science course, this study will based on location data provided by Foursquare. To make the location data alive, some socio-economic statistics will be taken in account also. Once again NYC acts as the central location of interest.

2.1 Data requirements

For this study geo-locational information about the boroughs and the particular neighborhoods of NYC is needed. Also the socio-economic statistics are needed on borough level. For quality reasons the needed data should be actual and accurate. So, the last US Census data will be gathered for this project.

2.2 Data collection

The project uses Foursquare API as primarily data source. By doing so, features of nearby places of the neighborhoods will be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100. As mentioned, data from US Census Bureau (https://www.census.gov/) is used to supplement the Foursquare data. To explore socio-economic statistics, this study focus on "QuickFacts" provided by US Census Bureau.

QuickFacts data are derived from: Population Estimates, American Community Survey, Census of Population and Housing, Current Population Survey, Small Area Health Insurance Estimates, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits.

2.3 Data preparation

Data will be prepared directly in the Jupyter notebook with the help of some essential python libraries. These libraries are numpy, pandas, geopy, matploblib, sklearn and folium. For handling JSON files, XLSX files and requests adequate modules will be imported also.

For data preparation some functions are defined to handle data more convenient. For example there is a function that extracts the category of the venue, a function that extracts nearby venues, and one function for return most common venues. These functions are adapted from prior labs in this course.