The Neighborhoods of New York

Shibbu Joseph

March 2021

2. Data acquisition and cleaning process

2.1. Data sources

The data was acquired thanks to the city of New York open data that can be found in this link clicking https://opendata.cityofnewyork.us/. The dataset that we used was the one of Neighborhood Names GIS that the link can be found https://data.cityofnewyork.us/City-Government/Neighborhood-Names-GIS/99bc-9p23. The data contains the geolocation, object id, name, stacked, borough, Annoline1, Annoline2, Annoline3, AnnoAngle of each neighborhood. We will use this data to classify each of the neighborhoods. We will be using the foursquare API to get the list of venues

2.2. Data cleaning

The data was downloaded but to work with the dataset I had to make a few changes in the dataset. First of all I had to eliminate some columns that wouldn't contribute at all with the analysis like stacked, Annoline1, Annoline2, Annoline3, and AnnoAngle. After doing the dropping of each of the columns there was another problem to work with, the_geom (the geolocation column) in each row had a format of POINT (Longitude, Latitude) so I had to eliminate first of all the point and the parentheses and after that separate the longitude and the latitude in different columns for each row. After doing all this the data frame was ready so I could work with it. I had some problems in the Staten Island analysis with the foursquare API with a neighborhood that wasn't returning close by venues so in the analysis I had to drop it. And in the New York analysis with a neighborhood of name 'Chelsea' that was getting duplicated.