The Neighborhoods of New York

Carlos Jimenez

April 07, 2020

1. Introduction

1.1. Background

New York City, officially the City of New York. It is the largest and most influential American metropolis. New York City is in reality a collection of many neighborhoods scattered among the city's five boroughs which are Bronx, Brooklyn, Manhattan, Queens, and Staten Island each exhibiting its own characteristics and ways of life. They say that moving from one city neighborhood to the next may be like passing from one country to another. Therefore, it is advantageous to know how each neighborhood is similar one to another in each borough.

1.2. Problem

The data that contribute to the classification of each neighborhood include the top most common venues found in each of the neighborhood. This project aims to cluster by similarity each cluster in all boroughs and finally give a general analysis of the complete New York City, comparing all the neighborhood in the city.

1.3. Interest

It works for everyone trying to move (for many reasons, be it moving because of job or trying to move closer to a specific school), this information can be useful to know more about those similar neighborhoods. Other who may be interested can be real estate agent looking to improve their options when offering a more similar property in terms of venues nearby.

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.

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, 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.