Most Suitable Neighborhood to Settle in Paris Who Leaves New York Using Data Science

Waruna Sanjeewa

Introduction

Today so many people leave their home country and settle in totally unknown country for different reasons such as job, medical reasons, vacation etc. When they move there they have no idea what would be the life style and facilities in that neighborhood. In this analysis I am trying to answer that question by using two cities (Paris and New York). A most suitable neighborhood in Paris to live is suggested to a person who lives in New York from this analysis.

Data

Neighborhood data of two cities for this project was collected using below two links,

- New York: https://cocl.us/new_york_dataset

- Paris: https://fr.wikipedia.org/wiki/Liste_des_quartiers_de_New_York#:~:text=Les%20quartiers%20new%2Dyorkais%20sont,le%20Bronx%20et%20Staten%20Island.

The geographical co-ordinates for these neighborhoods obtained using GEOPY library in python (For New York city geographical co-ordinates already included in Jason file downloaded from above site). Finally the nearby (within 500m radius) venue data set is obtained using https://foursquare.com .

Methodology

By using above data sets a **PYTHON** PANDAS data frame of all neighborhoods was created and then another two columns for longitude and latitude were created with the help of geo location data. Now the source data set creation is done.

By using above data set, for each neighborhood a data set of venues within the radius of 500 meter was crated with the help of FOURSQUARE.COM. Then the final data frame was crated against neighborhood with one hot encoded venue categories. Finally the data set is grouped with neighborhood by taking the mean of each venue category. Then a clustering algorithm is used to cluster these neighborhoods in to desired number of clusters.

In this analysis I used the model which is considered as one of the simplest model among them. Despite its simplicity, *k*-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data.