Applied Data Science Capstone Project

Segmenting and Clustering Neighborhoods- London and Paris

Lavenya Mohanasundaram

06 May 2021

IBM APPLIED DATA SCIENCE CAPSTONE

BATTLE OF NEIGHBORHOODS – LONDON AND PARIS

USING MACHINE LEARNING WITH PYTHON

1) Introduction

London is a leading global city. London is the capital of England and the United Kingdom; it is also the largest city within the country. It exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation. London has a diverse range of people and cultures, and more than 300 languages are spoken in the region. The London metropolitan area is the third-most populous in Europe, after Istanbul and the Moscow Metropolitan Area, with 14,040,163 inhabitants in 2016.

Paris is the capital and most populous city of France, located in the north-central part of the nation. Since the 17th century, Paris has been one of Europe's major centers of finance, diplomacy, commerce, fashion, gastronomy, science and arts. The City of Paris is part of Île-de-France region, and it is considered as one of economic centers in Europe. It is multicultural city and provides many business opportunities. It was ranked as the second most visited travel destination in the world in 2019, after Bangkok and just ahead of London.

Both London and Paris are found at the heart of two great European nations. London and Paris are quite the popular tourist and vacation destinations for people all around the world. They are diverse and multicultural and offer a wide variety of experiences that is widely sought after.

This project can be useful for those who moves to these cities, to find a good area to build and grow prosperously. In order to get a very good location details that meet this need, the London and Paris are explored through clustering and segmentation based on the London and Paris Post code and proximity to supplies. We try to group the neighborhoods of London and Paris respectively and draw insights to what they look like now.

2) Business Problem

Besides the two being great cities, each of them has their unique winning points as compared to the other. So, if you are planning to embark on a trip or change your residence, and can't quite choose between the two, don't get all stressed up. The aim of this project is to help people to choose their destinations depending on the experiences that the neighborhoods have to offer and what they would want to have. The goal is to help stakeholders and globetrotters to make informed decisions and address any concerns they have including the different kinds of cuisines, provision stores and what the city has to offer.

2.1) Target Audience

The purpose of this project is to help people in exploring better facilities around their neighborhoods. It will help people making smart and efficient decision on selecting great neighborhoods out number of other postal area in both the cites London and Paris. Lots

of people are migrating from various cities and needed lots of research for good housing prices, new business and reputed professional places for their children. The tourists can plan accordingly by choosing the neighborhoods in both cities.

This project is for those people who are looking for better neighborhoods and businesses. It will help people to get the awareness of area and neighborhood before visiting these big cities.

3) Data Acquisition

3.1) Data Description

This project will rely on geolocation data for both London and Paris. Postal codes in each city serve as a starting point. Using Postal codes, we use can find out the neighborhoods, boroughs, venues and their most popular venue categories.

For this project we need the following data:

3.1.1) London

To derive our solution, we scrape our data from web source

Data Source : https://en.wikipedia.org/wiki/List_of_areas_of_London

This Wikipedia page has information about all the neighborhoods, we limit it London.

1.borough: Name of Neighborhood
2.town: Name of borough
3.post_code: Postal codes for London.

This Wikipedia page lacks information about the geographical locations. To solve this problem, we use ArcGIS API.

3.1.2) Paris

To derive our solution, we leverage JSON data available from web source

Data Source : https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e

The JSON file has data about all the neighborhoods in France, we limit it to Paris.

1.postal_code: Postal codes for France
2.nom_comm: Name of Neighborhoods in France
3.nom_dept: Name of the boroughs, equivalent to towns in France
4.geo_point_2d: Tuple containing the latitude and longitude of the Neighborhoods.

3.2) Infrastructures Description

Different kinds of infrastructures in each neighborhood in London and Paris

Data Source:

- ✓ ArcGIS API
- ✓ Foursquare API

3.2.1) ArcGIS API

ArcGIS Online enables you to connect people, locations, and data using interactive maps. Work with smart, data-driven styles and intuitive analysis tools that deliver location intelligence. Share your insights with the world or specific groups.

More specifically, we use ArcGIS to get the geo locations of the neighborhoods of London. The following columns are added to our initial dataset which prepares our data.

1.latitude: Latitude for Neighborhood

2.longitude: Longitude for Neighborhood

3.2.2) Foursquare API

Venue Data

The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in London and Paris; it is used to study the popular venues of different neighborhoods as well as build the unsupervised learning model to cluster neighborhoods.

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside every neighborhood. For each neighborhood, we have chosen the radius to be 500 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1.Neighborhood: Name of the Neighborhood
2.Neighborhood Latitude: Latitude of the Neighborhood
3.Neighborhood Longitude: Longitude of the Neighborhood
4.Venue: Name of the Venue
5.Venue Latitude: Latitude of Venue
6.Venue Longitude: Longitude of Venue
7.Venue Category: Category of Venue

Using these data collected for both London and Paris will allow exploration and examination to build our model. This is a project that will make use of many data science skills, from web scraping, working with API (ArcGIS and Foursquare), data cleaning, data wrangling and map visualization (Folium), Exploratory Data Analysis to perform unsupervised Machine Learning using K-means clustering and Natural Language Processing using word cloud.