

APPLIED DATA SCIENCE CAPSTONE – FINAL REPORT

Introduction

The first step in approaching a data science problem is problem understanding. This step is very important since it allows us to know the kind of decisions we want to make, the information or data that will be needed to inform those decisions and finally, the kind of analysis that will be used to arrive at those decisions. In a nutshell, developing a mental model of the problem allows us to properly structure potentially relevant information needed to solve the problem.

We will begin the problem understanding with the problem statement.

Problem Statement

London is one of the most popular places for tourists and vacations. This may be due to the fact that London is made up of diverse people, multicultural and offers a wide variety of experiences for tourists. In this project, we will try to group the neighborhoods of London and draw insights to what they look like now. This is to help tourists and people who have intentions of migrating make good decisions based on the insights that will be drawn from this project.

Project Goal

The goal of this project is to explore city of London and find out what makes this city one of the popular destinations for tourists.

Data Acquisition and Understanding

This project faces several constraints and opportunities. There are not so many well-known or readily available datasets to help explore the city of London to arrive at the goals described above. To complete the project in full, we need the geographical location data for London. Postal codes in the city serve as a starting point. Using Postal codes, we can easily find out the neighborhoods, boroughs, venues and their most popular venue categories that can be used for this purpose.

For this project we decided to acquire information from [Wikipedia](#)

Finally, we will use the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 500 meters.

Data Collected

The information acquired from this site includes:

- Location
- London Borough
- Post Town
- Postcode District

- Dial Code
- OS Grid ref

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

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.

- Latitude: Latitude for Neighborhood
- Longitude: Longitude for Neighborhood

Methodology

The goal for this project as mentioned earlier is to explore city of London and find out what makes this city one of the popular destinations for tourists. Based on this stated goal, we will create a folium map to obtain insights that can help us arrive at some plausible conclusions.

We will then go ahead and summarize our findings in a short presentation.

To achieve the first goal, the we will have to answer the question below:

- What are some of the features that this city possesses?

The steps taken to derive the map that shows the various neighborhoods as well as the top 10 venues and corresponding features include:

1. Data Collection

To attain the neighborhoods in London, we start by scraping the list of areas of London from Wikipedia.

2. Data Preprocessing

Now we will clean our data and make it suitable for our analysis

The first step taken in the data cleaning process was to change the column names to lower case and replace all spaces with underscores to allow easy access. Then we removed all numeric values from the London borough column.

3. Feature Engineering

We observed that dataset contains all neighborhoods which were not need. Therefore, we only selected values containing only London.

4. London Map Visualization

We then used folium to visualize the neighborhoods located in London just to have a clear idea of where these neighborhoods are located.

5. One Hot Encoding

We then went ahead and performed one-hot encoding on the categorical values in our dataset by converting them to numerical values. This allowed us to calculate the top 10 common venues based on similarities in features.

6. Modeling and Analysis

We used KMeans Clustering ML algorithm to cluster similar neighborhoods together. We chose the clusters of 5.

Results and Discussion

The neighborhoods of London are very multicultural. There are a lot of different cuisines including Indian, Italian, Turkish and Chinese. London seems to take a step further in this direction by having a lot of Restaurants, bars, juice bars, coffee shops, Fish and Chips shop and Breakfast spots. It has a lot of shopping options too with that of the Flea markets, flower shops, fish markets, Fishing stores, clothing stores. The main modes of transport seem to be Buses and trains. For leisure, the neighborhoods are set up to have lots of parks, golf courses, zoo, gyms and Historic sites.

Overall, the city of London offers a multicultural, diverse and certainly an entertaining experience.

Conclusion

The purpose of this project was to explore the city of London and find out what makes this city one of the popular destinations for tourists and migrants. We explored the city based on their postal codes and then extrapolated the common venues present in each of the neighborhoods finally concluding with clustering similar neighborhoods together.

We could see that each of the neighborhoods in London has a wide variety of experiences to offer which is unique in its own way. The cultural diversity is quite evident which also gives the feeling of a sense of inclusion.