

Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Table of contents

- [Introduction: Business Problem](#)
- [Data](#)
- [Methodology](#)
- [Analysis](#)
- [Results and Discussion](#)
- [Conclusion](#)

Introduction: Business Problem

This project aims to select the safest borough in London based on the **total crimes**, explore the **neighborhoods** of that borough to find the **10 most common venues** in each neighborhood and finally cluster the neighborhoods using **k-mean clustering**.

This report will be targeted to people who are looking to **relocate to London**. In order to finalise a neighborhood to hunt for an apartment, **safety** is considered as a top concern when moving to a new place. If you don't feel safe in your own home, you're not going to be able to enjoy living there. The **crime statistics** will provide an insight into this issue.

We will focus on the safest borough and explore its neighborhoods and the 10 most common venues in each neighborhood so that the best neighborhood suited to an individual's needs can be selected.

Data

Based on definition of our problem, factors that will influence our decision are:

- The total number of crimes committed in each of the borough during the last year.
- The most common venues in each of the neighborhood in the safest borough selected.

Following data sources will be needed to extract/generate the required information:

- [Part 1: Preprocessing a real world data set from Kaggle showing the London Crimes from 2008 to 2016](#): A dataset consisting of the crime statistics of each borough in London obtained from Kaggle
- [Part 2: Scraping additional information of the different Boroughs in London from a Wikipedia page](#): More information regarding the boroughs of London is scraped using the BeautifulSoup library
- [Part 3: Creating a new dataset of the Neighborhoods of the safest borough in London and generating their co-ordinates](#): Co-ordinate of neighborhood will be obtained using Google Maps API geocoding

Part 1: Preprocessing a real world data set from Kaggle showing the London Crimes from 2008 to 2016

London Crime Data

About this file

- lsoa_code: code for Lower Super Output Area in Greater London.
- borough: Common name for London borough.
- major_category: High level categorization of crime
- minor_category: Low level categorization of crime within major category.
- value: monthly reported count of categorical crime in given borough
- year: Year of reported counts, 2008-2016
- month: Month of reported counts, 1-12

Data set URL: <https://www.kaggle.com/jboysen/london-crime>

Part 2: Scraping additional information of the different Boroughs in London from a Wikipedia page

Using Beautiful soup to scrap the latitude and longitude of the boroughs in London

URL: https://en.wikipedia.org/wiki/List_of_London_boroughs

Methodology

The methodology in this project consists of two parts:

- [Exploratory Data Analysis](#): Visualise the crime rates in the London boroughs to identify the safest borough and extract the neighborhoods in that borough to find the 10 most common venues in each neighborhood.
- [Modelling](#): To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster the 15 neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

As per the wikipedia page, The City of London is the 33rd principal division of Greater London but it is not a London borough.

URL: https://en.wikipedia.org/wiki/List_of_London_boroughs

Hence we will focus on the next borough with the least crime i.e. Kingston upon Thames

Visualizing different types of crimes in the borough 'Kingston upon Thames'

We can conclude that Kingston upon Thames is the safest borough when compared to the other boroughs in London.

Part 3: Creating a new dataset of the Neighborhoods of the safest borough in London and generating their co-ordinates.

The list of Neighborhoods in the Royal Borough of Kingston upon Thames was found on a wikipedia page:

https://en.wikipedia.org/wiki/List_of_districts_in_the_Royal_Borough_of_Kingston_upon_Thames

Modelling

- Finding all the venues within a 500 meter radius of each neighborhood.
- Perform one hot encoding on the venues data.
- Grouping the venues by the neighborhood and calculating their mean.
- Performing a K-means clustering (Defining $K = 5$)

Create a function to extract the venues from each Neighborhood

Each cluster is color coded for the ease of presentation, we can see that majority of the neighborhood falls in the red cluster which is the first cluster. Three neighborhoods have their own cluster (Blue, Purple and Yellow), these are clusters two three and five. The green cluster consists of two neighborhoods which is the 4th cluster.

Analysis

Analyse each of the clusters to identify the characteristics of each cluster and the neighborhoods in them.

Results and Discussion

The aim of this project is to help people who want to relocate to the safest borough in London, expats can choose the neighborhoods to which they want to relocate based on the most common venues in it. For example if a person is looking for a neighborhood with good connectivity and public transportation we can see that Clusters 3 and 4 have Train stations and Bus stops as the most common venues. If a person is looking for a neighborhood with stores and restaurants in a close proximity then the neighborhoods in the first cluster is suitable. For a family I feel that the neighborhoods in Cluster 4 are more suitable due to the common venues in that cluster, these neighborhoods have common venues such as Parks, Gym/Fitness centers, Bus Stops, Restaurants, Electronics Stores and Soccer fields which is ideal for a family.

Conclusion

This project helps a person get a better understanding of the neighborhoods with respect to the most common venues in that neighborhood. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a

neighborhood. We have just taken safety as a primary concern to shortlist the borough of London. The future of this project includes taking other factors such as cost of living in the areas into consideration to shortlist the borough based on safety and a predefined budget.