Data Acquisition and Cleaning

Data Acquisition

The data acquired for this project is a combination of data from three sources.

1. London crime data (https://www.kaggle.com/jboysen/london-crime) that shows the crime per borough in London. The dataset contains the following columns:

    - **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
    - **value**: monthly reported count of categorical crime in given borough
    - **year**: Year of reported counts, 2008-2016
    - **month**: Month of reported counts, 1-12

2. The list of London boroughs (https://en.wikipedia.org/wiki/List_of_London_boroughs). This page contains additional information about the boroughs, the following are the columns:

    - **Borough**: The names of the 33 London boroughs.
    - **Inner**: Categorizing the borough as an Inner London borough or an Outer London Borough.
    - **Status**: Categorizing the borough as Royal, City or other borough.
    - **Local authority**: The local authority assigned to the borough.
    - **Political control**: The political party that control the borough.
    - **Headquarters:** Headquarters of the Boroughs.
    - **Area (sq mi)**: Area of the borough in square miles.
    - **Population (2013 est)[1]**: The population in the borough recorded during the year 2013.
    - **Co-ordinates**: The latitude and longitude of the boroughs.
    - **Nr. in map**: The number assigned to each borough to represent visually on a map.

3. The list of Neighborhoods in the Royal Borough of Kingston upon Thames (https://en.wikipedia.org/wiki/List_of_districts_in_the_Royal_Borough_of_Kingston_upon_Thames). This dataset is created from scratch using the list of neighborhood available on the site, the following are columns:

    - **Neighborhood:** Name of the neighborhood in the Borough.
    - **Borough:** Name of the Borough.
    - **Latitude:** Latitude of the Borough.
    - **Longitude**: Longitude of the Borough.

Data Cleaning

From the London crime data we use only the crimes during the most recent year (2016). The major categories of crime are pivoted to get the total crimes per borough as per the category.

The second data is scraped from a wikipedia page using the Beautiful Soup library in python. Using this library we can extract the data in the tabular format as shown in the website. After the web scraping, string manipulation is required to get the names of the boroughs in the correct form (we will be merging the two datasets together using the Borough names).

The two datasets are merged on the Borough names to form a new dataset that combines the necessary information in one dataset. The purpose of this dataset is to visualize the crime rates in each borough and identify the borough with the least crimes recorded during the year 2016.

Once crime is obtained in each district, we can determine the district with the lowest crime rate, and therefore label that district as the safest district. The third source of data comes from Wikipedia's list of neighborhoods in the safest district. This data set is created from scratch, the pandas data frame is created with the names of the neighborhoods and the name of the municipality with the latitude and longitude are left blank.

The neighborhood coordinates have been obtained using the Google Maps API geocoding to obtain the final data set. The new dataset is used to generate the locations for each neighborhood using the Foursquare API.