1. Data Acquisition and Cleaning:

1.1. Data Acquisition:

The data acquired for this project is a combination of data from three sources. The first data source of the project uses a kaggle datasets download -d 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 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

The second source of data is scraped from a Wikipedia page that contains the 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.): 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.

The third data source is the list of Neighbourhoods in the Royal Borough of Kingston upon Thames as found on a Wikipedia page. This dataset is created from scratch using the list of neighbourhood available on the site.

- Neighbourhood: Name of the neighbourhood in the Borough.
- Borough: Name of the Borough.
- Latitude: Latitude of the Borough.
- Longitude: Longitude of the Borough.

1.2 Data Cleaning:

The data preparation for each of the three sources of data is done separately. From the London crime data, the crimes during the most recent year (2016) are only selected. 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. This is important because 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.
After visualizing the crime in each borough we can find the borough with the lowest crime rate and hence tag that borough as the safest borough. The third source of data is acquired from the list of neighbourhoods in the safest borough on Wikipedia. This dataset is created from scratch, the pandas data frame is created with the names of the neighbourhoods and the name of the borough with the latitude and longitude are left blank

The coordinates of the neighbourhoods is be obtained using Google Maps API geocoding to get the final dataset. The new dataset is used to generate the venues for each neighbourhood using the Foursquare API.