2. Data Acquisition and Cleaning

2.1 Data Used and Source
- **London Crime Data**
- **List of London Boroughs**
- **List of districts in the Royal Borough of Kingston upon Thames**

London Crime Data

London crime data shows the crime per borough in London is taken from Kaggle . The dataset contains the following columns:

- **lsoa_code: Area code**
- **borough : Common name of Borough**
- **major_category: High level crime**
- **minor_category : Low level crime**
- **Value   : count of crime**
- **Year   : year of record**
- **Month: Month of record**

List of London Boroughs

List of London boroughs is scrapped from the wiki page, this page contains additional information about the boroughs

- **Borough: The names London boroughs.**
- **Inner: Categorizing the borough as an Inner or Outer London Borough.**
- **Status: Categorizing the borough as Royal, City or other.**
- **Local authority: The local authority of borough.**
- **Political control: The political party ruling**
- **Headquarters: HQ of the Boroughs.**
- **Area : Area of the borough .**
- **Population : The population in the borough recorded**
- **Co-ordinates: The latitude and longitude of the boroughs.**
- **Nr. in map: The number assigned to each borough to represent visually on a map.**

List of districts in the Royal Borough of Kingston upon Thames

List of districts in the Royal Borough of Kingston upon Thames is wiki page , we tweaked it to find the below columns

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

2.2 Data Cleaning

From the London crime data, the crimes during the year 2016 is only selected. The major categories of crime are pivoted to get the total crimes per borough as per the category.

Beautiful Soup Library is used to scrap List of London boroughs of wiki to obtain borough details. Many string manipulation and removal is required to obtain data in desired format.

These two datasets are merged to on Borough Names to get all the required information inside a single dataset , which can help us visualize crimes count against each borough so that we can find which borough has highest and which have lowest crimes against them

Then we will create the third dataset from List of districts in the Royal Borough of Kingston upon Thames   using pandas and set the Neighborhood and will fill the latitude and longitude from **geocoding API**

Then we will use this dataset to generate venues of each neighborhood using **FourSquare API**