2. Data Acquisition and Cleaning

2.1 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 London crime data 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)[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.

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, the following are columns:

● Neighbourhood: Name of the neighbourhood in the Borough.

● Borough: Name of the Borough.

● Latitude: Latitude of the Borough.

● Longitude: Longitude of the Borough.