2. Data Acquisition and Data Cleaning

2.1. Data Acquisition

This project is implemented using a combination of multiple data sources. Three datasets are used of which some are directly available while some are scraped from Wikipedia.

2.1.1. London Crime Data

This data covers the number of criminal reports by month, LSOA borough, and major/minor category from Jan 2008-Dec 2016. The crimes per borough in London are shown in the dataset. The columns in the dataset are as follows:

- **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

This is the main dataset which consists a total of 13M rows. Each entry/row describes a crime in such a way that we can understand the name of the London borough in which the crime has taken place along with the LSOA code for the area. The further attributes mention the major and minor category in which the crime can be classified. The value which is the monthly count of reported crime for that category of crime in the mentioned borough. The year of the reported entry (2008-2016) along with the month (1-12) are the last two attributes.

2.1.2. London Boroughs

The specific information about the boroughs present in London. This data is scraped from a Wikipedia page. The columns in this dataset are as follows:

- **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.

This data consists of detailed information regarding the 33 boroughs in London with a row for each borough with 10 attributes/features for each, which we did not have in the previous dataset. The borough name will be a common column in this and the London Crime Data datasets. The Inner/Outer category of the borough is mentioned along with the status of the borough as Royal, City or Other Borough. The other attributes include the local authority and political control for each borough along with the headquarter location and the total area of the borough. The co-ordinates on the map is an important attribute for each borough. Population for the borough is also an attribute. Nr. In map is a unique number given to each borough just for the purpose of representation on the map.

2.1.3. Neighbourhoods in the Royal Borough of Kingston upon Thames

The list of neighbourhoods data is scraped from Wikipedia page for the Royal Borough of Kingston upon Thames. The columns in this dataset are as follows:

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

The neighbourhoods in the borough which is identified safe are explored using this scraped data. The dataset consists or multiple entries, each associated with a neighbourhood in the Royal Borough of Kingston upon Thames. The neighbourhood name, borough name are the first two attributes. The other attributes mention the latitude and longitude values for that neighbourhood entry.

2.2. Data Cleaning

2.2.1. London Crime Data

From the entire dataset, the data associated with the latest year (2016) is selected for further processing. The major categories of crimes are used and pivoted to get the total crimes in each borough based on the category.

	Borough	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	Total
0	Barking and Dagenham	1287	1949	919	378	534	5607	6067	16741
1	Barnet	3402	2183	906	499	464	9731	7499	24684
2	Bexley	1123	1673	646	294	209	4392	4503	12840
3	Brent	2631	2280	2096	536	919	9026	9205	26693
4	Bromley	2214	2202	728	417	369	7584	6650	20164

Fig. London Crime Data (after Pre-Processing)

Considering the first entry in the above data, we can see that 'Barking and Dagenham' borough is the borough name. The following attributes give us the detailed total summary for the count of crimes based on the specific categories. For instance, total crimes under the burglary category are recorded as 1287, while under Criminal Damage are 1949. Similarly the values for Drugs, Other Notifiable Offences, Robbery, Theft and Handling, Violence Against the Person have recorded values as 919, 378, 534, 5607 and 6067 respectively. The last column named as 'Total' gives us the count of total crimes for that borough. Understanding such attribute values for each borough, we can identify the borough with maximum number of crimes and the total counts for each category as well. We can understand the overall characteristics of each borough as far as the safety factor is concerned. The conclusion which can be drawn from this data is valuable, considering the objective of this project.

2.2.2. London Boroughs

The data from Wikipedia page is scraped using the **Beautiful Soup** python library to extract it in a tabular format. String manipulation is required to ensure that the names of Boroughs match with that in the previously collected data. We need the names to match exactly as we will be merging the datasets further.

	Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est)[1]	Co-ordinates	Nr. in map
0	Barking and Dagenham	NaN	NaN	Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	194352	51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E /...	25
1	Barnet	NaN	NaN	Barnet London Borough Council	Conservative	North London Business Park, Oakleigh Road South	33.49	369088	51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W /...	31
2	Bexley	NaN	NaN	Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	236687	51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E /...	23
3	Brent	NaN	NaN	Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	317264	51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W /...	12
4	Bromley	NaN	NaN	Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	317899	51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E /...	20

Fig. List of London Boroughs

The scraped data gives us detailed description about various attributes of the boroughs in London. Consider any such entry, for example 'Barking and Dagenham' borough. The various attribute values give us the description that includes information such as local authority which is Barking and Dagenham London Borough Council. The political control is in the category 'Labour'. The headquarter address is also available for each borough along with the

geographical co-ordinates for the borough. The area for this borough is 13.93 (sq. mi) and the population is 194,352. The unique number associated is 25 (used to represent on the map). The additional features that this dataset can be used to briefly study any of the boroughs while deriving the conclusions.

2.2.3. Merged Dataset

Visualizing the most crimes in each borough using the dataset created by merging the two previous datasets.

	Borough	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est)[1]	Co-ordinates	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	Total
0	Barking and Dagenham	Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	194352	51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E /...	1287	1949	919	378	534	5607	6067	16741
1	Barnet	Barnet London Borough Council	Conservative	North London Business Park, Oakleigh Road South	33.49	369088	51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W /...	3402	2183	906	499	464	9731	7499	24684
2	Bexley	Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	236687	51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E /...	1123	1673	646	294	209	4392	4503	12840
3	Brent	Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	317264	51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W /...	2631	2280	2096	536	919	9026	9205	26693
4	Bromley	Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	317899	51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E /...	2214	2202	728	417	369	7584	6650	20164

Fig. London Crimes per Borough

The previous datasets are merged using the common column ‘Borough’ to combine all the available data together to collectively analyse the data for every borough. The example of ‘Barking and Dagenham’ here now has all the attributes/features combined which we separately observed in the previous datasets.

2.2.4. Neighbourhoods in the Royal Borough of Kingston upon Thames

The neighbourhoods in the borough which is specifically identified as safe, the Royal Borough of Kingston upon Thames, are explored. The data (name of borough and neighbourhoods) is scraped from the Wikipedia page and the latitude and longitude values are obtained using the Google Maps API geocoding to form the final dataset. Further, Foursquare API will be used to generate venues for each neighbourhood.

	Neighborhood	Borough	Latitude	Longitude
0	Berrylands	Kingston upon Thames	51.393781	-0.284802
1	Canbury	Kingston upon Thames	51.417499	-0.305553
2	Chessington	Kingston upon Thames	51.358336	-0.298622
3	Coombe	Kingston upon Thames	51.419450	-0.265398
4	Hook	Kingston upon Thames	51.367898	-0.307145

Fig. Neighbourhoods of the safest boroughs

Focusing on the identified borough, each entry in this dataset now refers to a unique neighbourhood. Each neighbourhood record has a latitude and longitude value associated with it. We can use this location data to use FourSquare API to explore the various venues in these neighbourhoods.