The Battle of Neighborhoods- London

Presentation by Nicholas

Introduction

- **Background**: Safety is a top concern when moving to a new area. If you don't feel safe in your own home, you're not going to be able to enjoy living there.

- **Problem**: This project aims to select the safest borough in London based on the total crimes, explore the neighborhoods of that borough to find the 10 most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

- **Interest**: The project can help investors and other social categories to get better understanding of the neighborhoods in the Boroughs of London.

Data Acquisition and Cleaning

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 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 third data source is the list of Neighborhoods in the Royal Borough of Kingston upon Thames as found on a wikipedia page.

Data Acquisition and Cleaning

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 the boroughs for each major 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.
- The two datasets are merged on the Borough names to form a new 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.
- The new dataset is used to generate the 10 most common venues for each neighborhood using the Foursquare API, finally using k means clustering algorithm to cluster similar neighborhoods together.

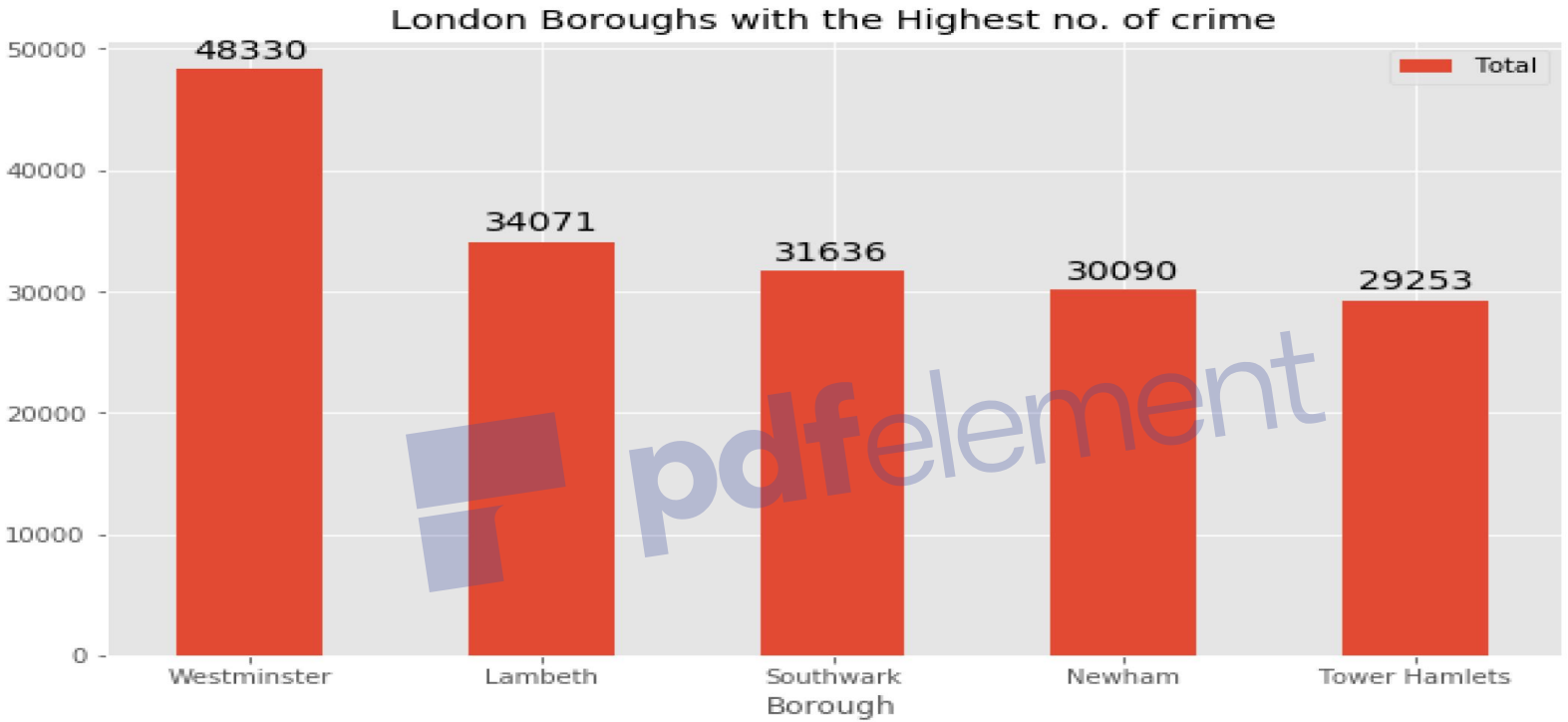
Methodology

Exploratory Data Analysis:

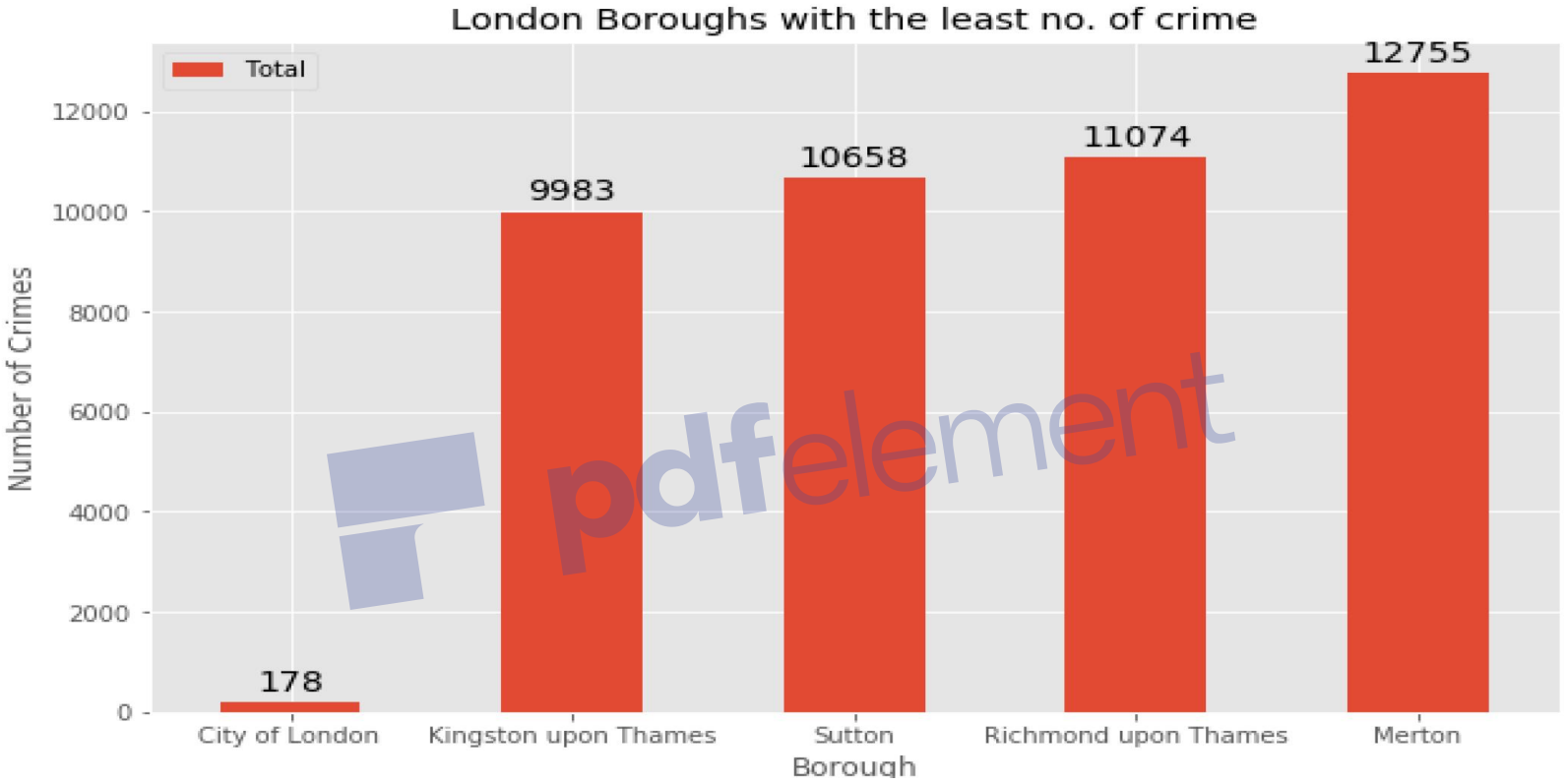
Statistical summary of crimes

	Burglary	Criminal Damage	Drugs	Notifiable Offences	Robbery	Theft and Handling	Violence Against the person	Total
count	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000
mean	2069.242424	1941.545455	1179.212121	479.060606	682.666667	8913.121212	7041.848485	22306.696970
std	737.448644	625.207070	586.406416	223.298698	441.425366	4620.565054	2513.601551	8828.228749
min	2.000000	2.000000	10.000000	6.000000	4.000000	129.000000	25.000000	178.000000
25%	1531.000000	1650.000000	743.000000	378.000000	377.000000	5919.000000	5936.000000	16903.000000
50%	2071.000000	1989.000000	1063.000000	490.000000	599.000000	8925.000000	7409.000000	22730.000000
75%	2631.000000	2351.000000	1617.000000	551.000000	936.000000	10789.000000	8832.000000	27174.000000
max	3402.000000	3219.000000	2738.000000	1305.000000	1822.000000	27520.000000	10834.000000	48330.000000

The count for each of the major categories of crime returns the value 33 which is the number of London boroughs. ‘Theft and Handling’ is the highest reported crime during the year 2016 followed by ‘Violence against the person’, ‘Criminal damage’. The lowest recorded crimes are ‘Drugs’, ‘Robbery’ and ‘Other Notifiable offenses’.

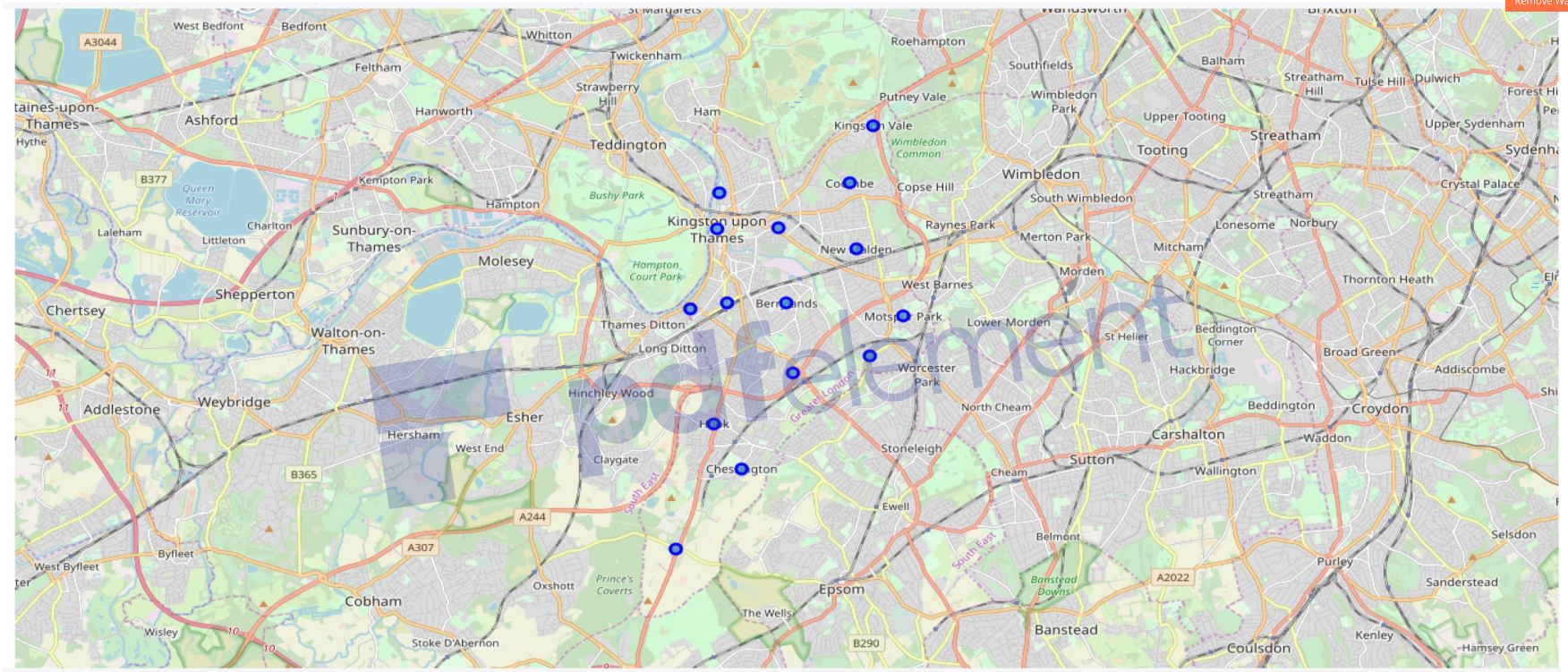


According to the data above we can observe that the boroughs with the highest number of crimes are Westminster, Lambeth, Southwark, Newham and Tower Hamlets.



According to the data graph above we can acknowledge that the Boroughs of London with least number of crime are City of London, Kingston upon Thames, Sutton, Richmond upon Thames,Merton.

Neighborhoods in London



These are the 15 neighborhoods in the royal borough of Kingston upon Thames. They are visualised on a map using folium on python

Modelling

- Using the final dataset containing the neighborhoods in Kingston upon Thames along with the latitude and longitude, we can find all venues within a 1000 meter radius of each neighborhood by connecting to the Foursquare API.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berrylands	51.393781	-0.284802	Surbiton Racket & Fitness Club	51.392676	-0.290224	Gym / Fitness Center
1	Berrylands	51.393781	-0.284802	Jean's Cafe	51.393214	-0.296070	Coffee Shop
2	Berrylands	51.393781	-0.284802	Prince of Wales	51.389642	-0.297580	Pub
3	Berrylands	51.393781	-0.284802	The Berrylands Pub	51.398169	-0.280365	Pub
4	Berrylands	51.393781	-0.284802	K2 Bus Stop	51.392302	-0.281534	Bus Stop

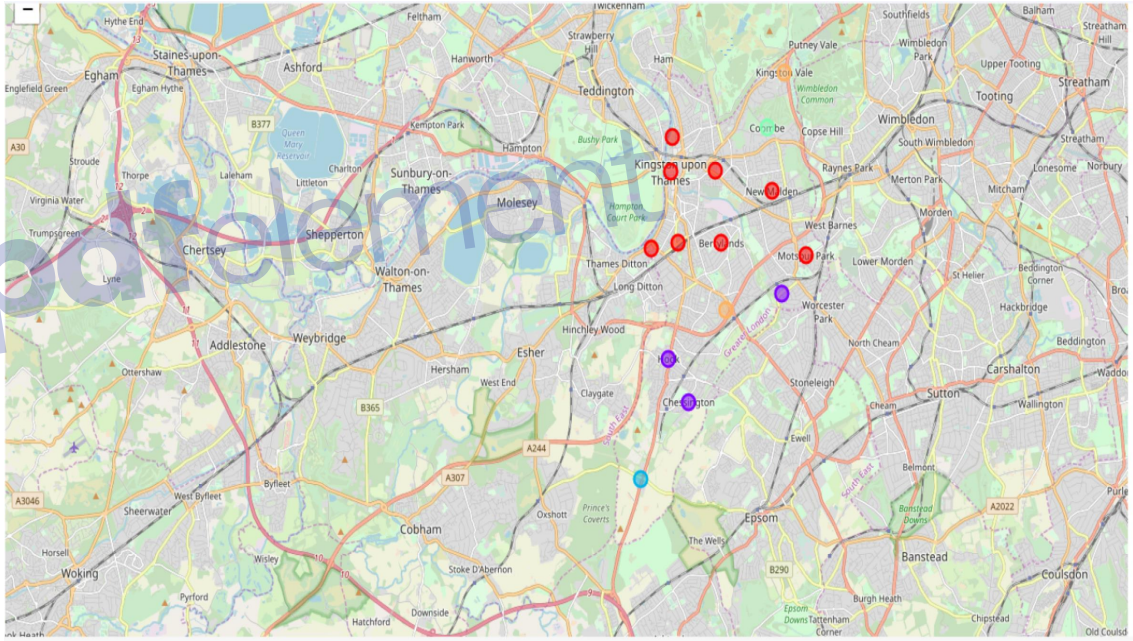
- One hot encoding is done on the venues data. The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

- We will use a cluster size of 5 for this project that will cluster the 15 neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

Results

After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. Visualising the clustered neighborhoods on a map using the folium library.

Every cluster is color coded for the ease of presentation, we can see that majority of the neighborhood falls in the red cluster which is the first cluster.Three neighborhoods have their own cluster (Blue, Purple and Orange), these are clusters two, three, and five. The green cluster consists of one neighborhood which is the fourth cluster



Discussion

The aim of the project is to help investors, students, or other social categories who want to develop new business projects, to relocate, to study, or to live in the safest and secure borough in London . Therefore , if someone is looking for a neighborhood with good connectivity and public transportation we can see that Clusters one, two and five have train stations and bus stops.

Furthermore, the first cluster has neighborhoods most suitable for shopping, dining, coffee meetings, pubs or for someone eager to have a taste of international cuisines. This cluster is most suitable for family, students or potential new residents seeking to relocate. For new business opportunities, clusters three, four and five are the best choice. The obvious choice for football fans is cluster five. Clusters one, four and five are most suitable for new investments.

Conclusion

The project can help investors and other social categories to get betterunderstanding of the neighborhoods in the Boroughs of London. Therefore, we can acknowledge that technology is helpful to stay ahead in finding out more about areas before moving, living, study or invest in new business opportunities

The primary goal and focus of the project was on safety and security. The future of this project resides in taking other factors into account by adding new data in order to improve the decision making of companies, students or new residents.