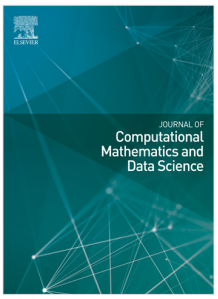
Journal Pre-proof



London street crime analysis and prediction using crowdsourced dataset

Ahmed Yunus, Jonathan Loo

PII: S2772-4158(23)00016-0
DOI: https://doi.org/10.1016/j.jcmds.2023.100089
Reference: JCMDS 100089

To appear in: *Journal of Computational Mathematics and Data Science*

Received date : 3 November 2023
Revised date : 27 November 2023
Accepted date : 11 December 2023

Please cite this article as: A. Yunus and J. Loo, London street crime analysis and prediction using crowdsourced dataset. *Journal of Computational Mathematics and Data Science* (2023), doi: https://doi.org/10.1016/j.jcmds.2023.100089.

London Street Crime Analysis and Prediction using Crowdsourced Dataset

Ahmed Yunus
School of Computing and Engineering
University of West London, London
ahmedyunuspilot@gmail.com

Professor Jonathan Loo
School of Computing and Engineering
University of West London, London
jonathan.loo@uwl.ac.uk

Abstract

To effectively prevent crimes, it's vital to anticipate their patterns and likely occurrences. Our efforts focused on analyzing diverse open-source datasets related to London, such as the Met police records, public social media posts, data from transportation hubs like bus and rail stations etc. These datasets provided rich insights into human behaviors, activities, and demographics across different parts of London, paving the way for a machine learning-driven prediction system. We developed this system using unique crime-related features extracted from these datasets. Furthermore, our study outlined methods to gather detailed street-level information from local communities using various applications. This innovative approach significantly enhances our ability to deeply understand and predict crime patterns. The proposed predictive system has the potential to forecast potential crimes in advance, enabling government bodies to proactively deploy targeted interventions, ultimately aiming to prevent and address criminal incidents more effectively.

Keywords

Crime; CRISP-DM; Linear Regression; XGB Regression; Machine Learning; Latitude; Longitude; Stop and Search; NodeJs; Flutter; Database; DynamoDb etc.

I. INTRODUCTION

A crime by definition refers to an act that is punishable by the state or other recognized authority in that demographic area **[1]**. A street crime is one of the criminal offense categories that tend to happen in public places which includes robbery, theft, anti-social behavior, murder etc. **[2]**. Street crime has become a real world problem across all the countries, thus it has become an absolute necessity to interpret criminal offense patterns, analyze and provide recommendations and mitigations procedures to the respected authority.

London, capital city of England, is one of the most gorgeous and expensive cities having one the most diverse population sets in the whole world. People from different regions and demographics come and live here to lead their life. London has faced a drastic menace regarding the street crime offenses that shows only a trend to be upwards except some exceptional years like Covid in 2020-21. The crime rate has hit 102.4 offenses per one thousand people in the time period of 2019 to 2020 **[3]**. From each consecutive year from 2015 to 2019 the crime rate showed an upward trend and only a sudden drop in 2020-21 as this was due to pandemic effect.

The United Kingdom is facing a spike in street crime that leads to a necessity of taking proper actions by the government and social entities in order to make it under control for the ease of people's lives. In order to understand the pattern and sequences of crime activities, it is vital to rely on different perspectives of community messages through different mediums such as social, government entities etc. Social media along with different government entities' provide open source dataset could act as the source for analyzing crime patterns. The most notable dataset could be twitter dataset on criminal offenses, Metropolitan police dataset on crime offenses, education rate dataset, Bus/Rail stations dataset, unemployment rate dataset etc. which would pave a way to find a pattern in the crime activities.

This paper primarily centers on analyzing crime patterns in London utilizing the aforementioned crowdsourced datasets. It aims to develop predictive capabilities for future crime events in specific locations. Additionally, to gather highly detailed street-level data, a mobile application has been introduced. This app enables users to convey their safety perceptions about particular geolocations, thereby contributing to the creation of a precise dataset.

A. Aims and purpose

The principal objective of this study is to examine the crime patterns of a specific city using open-source datasets that could have broader applicability to other cities worldwide. London was selected as the focal city for analysis due to the availability of open and crowdsourced data, as well as the research's geographical context within London. Upon concluding the analysis, our aim is to assess the dataset using various machine learning algorithms to uncover potential patterns and ascertain the dataset's predictive capacity regarding future crime activities.

B. Research objective

Our primary research objective is to conduct an in-depth review of empirical studies focusing on crime pattern analysis and prediction. We aim to identify and address significant research gaps by tackling complex challenges within this field. Leveraging various datasets pertaining to London, our endeavor involves uncovering street-level crime patterns and subsequently establishing correlations between these patterns and other associated features. This systematic exploration will culminate in the design of a real-time crime analysis and prediction system, enabling the forecast of criminal activity within specific geographic areas.

C. Methodologies

This paper adheres to the CRISP-DM structure, recognized as the industry-standard data mining process outlined in reference **[24]**. CRISP-DM, which stands for CRoss Industry Standard Process for Data Mining, comprises six phases designed to manage intricate data mining and machine learning processes effectively.

1. **Business understanding**

   Business understanding refers to achieving research objectives by defining the necessary requirements. The main objective of this paper is to conduct a thorough crime analysis on London city premises and predict future crime activities.

2. **Data Understanding**

   Data understanding refers to collection of different dataset based on business requirements. In this paper, we have gone through different open sourced dataset such as Metropolitan police dataset on crime offenses, education rate dataset, taxi information dataset, unemployment rate dataset, weather dataset etc. to make correlations among these to achieve our business requirements.

3. **Data Preparation**

   Data preparation tends to be the most complex process among all the other processes as we have to make the correlation among different features that are collected in different contexts. We have to merge them precisely so that these data points could provide some prediction power for crime patterns.

4. **Modeling**

   Crime activity prediction could be labeled both as classification or regression problem for our context. Labeling as classification means that we would predict whether an area is safe or not safe for a specific timestamp. Labeling as regression indicates that we would calculate the number of crime activities that could happen in a specific timestamp frame.

5. **Evaluation**

   To evaluate the predicting power among different features and to achieve the best results, we would evaluate our unique custom dataset to a number of machine learning models.

6. **Deployment**

   The complete system would be deployed to fulfill business requirements and to provide predictions on future crime event numbers.

II. RELATED WORKS

Crime has long been a subject of study within the realms of statistics and mathematical modeling. In our initial research phase, we extensively examined various crime modeling methodologies proposed by **Perc et al. [39]** and **Dorsogna et al. [38]**, which predominantly focus on mathematical modeling. Their research emphasizes leveraging applied mathematics and statistical physics to understand criminal activities. These studies explore a wide array of modeling approaches, encompassing partial differential equations, self-exciting point processes, agent-based modeling, spatial inspection game (Monte Carlo Simulation), evolutionary games, network science etc.. The primary objective of these approaches is to illuminate complex phenomena within crime dynamics, including the identification of crime hotspots, understanding gang formation, and comprehending the intricate networks associated with organized crime.

Crime modeling has changed to incorporate aggregated, anonymized human behavioral data as a result of the accessibility of online data. Through term-frequency analysis researched by **Williams et al. [4]**, in particular, geotagged Twitter data has been utilized to comprehend how social media themes connect to crime. The modeling of crime has also made use of dynamic data on human behavior. In a separate study conducted by **Traunmueller et al. [5]**, they delved into the connections between human activity attributes, inferred from mobile phone data, and the monthly crime rates. This research aimed to establish correlations or potential relationships between patterns of human behavior, as deduced from mobile phone data, and the fluctuations observed in crime rates on a monthly

basis. Another research by **Wang et al. [6]** and **Lal et al. [7]** recommended an automated apparatus of wrongdoing and non-crime classification of tweets that could be significantly important noteworthiness. The proposed way began with extraction of certain keywords and fed these into tf-idf vectorization in order to identify wrong-doing activities in a specific geographic zone. In spite of the fact that, although it achieved great level of extraction power in the certain datasets, but it did not added the essence of parts-of-speech labels neither it did not include and geo-location information which truly nullifies our main motivation of pinpointing street level crime activity patterns.

A research by **Sandagiri et al. [8]** used a non-linear methodology by using artificial neural networks. They used tweets and extracted the information on whether that twitter post contains any crime event information. With additional features such as adding weather in that timestamp they predicted the future occurrence of the crime event. The neural network achieved an excellent remarkable score of 92% in order to predict the crime event. Similar prediction system was built by other researchers such as **Lal et al. [7]**.

A great breakthrough in crime pattern detection happened in the research by **Zhang et al. [9]** using Google Street View Images (GSV). Along with Twitter and Foursquare which is an independent data location platform, the researchers used GSV which provides information about human behavioral information along with geolocation. But this systematic process has its own flaws such as the algorithms always think that in every place with more greenish, there might not happen any crime offense but it is a very common situation that usually a place that might be more greener than other area could be quite and less busier and which could lead to an incredible chance of theft or robbery and even worse that sexual violence act could happen.

There is another research done on crime pattern analysis by **Chen et al. [10]** which made an adjustment on the twitter dataset by adding climate related datapoints as natural components to check if there is a correlation in these features. The main rationale is that the environment could play a vital factor for the human's behavior. Another researcher **Anderson et al. [11]** showed a direct correlation between crime active recurrence and weather. The research that is done by **Chen et al. [10]** achieved 67% AUC for the experiment. But in any of their cases, they did not relate correlation in between climate information and Twitter against a particular crime event.

There is a handful of spatial analysis-based prediction research done. The work by **Rosser et al. [12]** is notable for its analysis of criminal occurrences at the street segment level as opposed to the grid or census unit level. However, given there is no analysis of traffic movement used to construct the forecasts, or density of pedestrians on specific routes, human dynamics are not expressly taken into account.

A novel approach was taken by utilizing cellular phone information in a geographic area span which took human behavior as the key factor and in this case it is the network movement **[13]**. The research had provided recommendations on a data-driven approach which is achieved by mobile network movement. The mobile phone utilization has hit approximately 6.8 billion numbers **[14]** which surely broadcasts human behavioral patterns. In the research **[13]**, a number of machine learning algorithms were used such as Logistic Regression, SVM and many more tree classifiers to understand the pattern of mobile activity in relation with crime hotspots. The dataset was very small, only 3 weeks of mobile network data and achieved a 53% accuracy score to predict crime events with the movement of mobile networks of geographical behavior. Moreover the dataset was provided in a hackathon competition and it is not even publicly released by the mobile network operator.

A particular research with a similar objective of this research **[13]** was conducted by **[15]** on making prediction of crimes using geographical location specifically in Manchester and other surrounding areas. They mainly focused on police crime record as the only point of interest and designated crime prediction as classification problem. They divided areas in hotspots which is definitely a milestone but there was no mention of timestamp neither there we could find any other relevant point of interest data that would make this prediction real time.

III. METHODOLOGY

From the empirical studies detailed in **Section II**, we've pinpointed substantial gaps and challenges that have prompted us to explore simple but effective approaches. Initially, we encountered the following research limitations from previous papers:

1. The application of mathematical models in crime modeling **[38][39]**, often involves complexity that might pose limitations in its effectiveness. Real-world crime phenomena present intricate dynamics, and the availability and quality of data further impact the efficacy of these models. The influence of societal nuances and cultural factors on criminal behavior might not be entirely encapsulated within mathematical models, thereby constraining their relevance in various social contexts.
2. The majority of the research we surveyed relied heavily on single-source datasets like Twitter or weather records. Unfortunately, these studies lacked comprehensive correlations among various features and patterns of criminal activity.
3. The research conducted did not encompass street-level information; instead, it primarily focused on broader geographical areas such as Wards or Boroughs, omitting the detailed geo-coordination of streets.
4. Regrettably, our investigation did not uncover any research utilizing victim information to discern crime sequences. This facet is deemed essential for a comprehensive understanding of crime patterns.
5. The majority of studies predominantly utilized classifiers to determine the likelihood of crime occurrence rather than focusing on forecasting future criminal activities.
6. Past research has notably lacked classification regarding specific crime events such as antisocial behavior or sexual violence within the broader crime categories.
7. The primary limitation we identified revolves around the researchers' inability to conduct real-time predictions of future crime events due to the absence of timestamps. While diverse features were considered, the lack of synchronization in timestamps across these features might result in diminished predictive accuracy. Our proposal advocates for a systematic approach to real-time crime prediction, emphasizing the imperative alignment of every feature with specific timestamps corresponding to event occurrences.

In pursuit of addressing the research gaps discussed earlier and constructing a real-time crime prediction system, this paper delineates four distinct processes outlined below that fulfills the CRISP-DM structure:

A. Analytical approach

Our analytical approach primarily focuses on the selection of open-source datasets based on specific points of interest. Subsequently, we evaluate these datasets to determine their suitability for the subsequent phases. Through this analysis, we aim to outline the design and modeling strategies for our machine learning system, emphasizing meticulous data preparation and rigorous model evaluation.

B. In depth exploratory data analysis and data preparation

During this phase, our focus lies in navigating through the chosen dataset, selecting relevant features, and consolidating them into a unified custom dataset. This amalgamated dataset serves as the input for our machine learning system, crucial for predicting crime patterns and anticipating future criminal activities.

C. Application design for victim data collection

As previously mentioned, we observed a lack of timestamp features across all crime activity pattern detections. Additionally, we noted the absence of victim-centric features within the datasets. To address this, our proposal involves the creation of features through mobile applications, offering a perspective from the victim's vantage point regarding the safety of crime hotspots. These features will also

encompass timestamps corresponding to data reception. This section will extensively elaborate on this proposed approach.

D. Dataset preparation

This section encompasses the completion of data preparation derived from various open-source datasets selected based on points of interest. It offers a comprehensive summary contextualizing the entirety of the collected data.

A. Analytical approach

The primary emphasis during this stage revolves around the selection of specific datasets predicated on their features, deliberating upon their suitability for subsequent feature extraction processes. This phase encompasses a comprehensive understanding of the dataset structures, which facilitates the formulation and modeling of our machine learning system. This understanding is pivotal in delineating precise data preparation strategies and conducting thorough model evaluations. The sequential delineation of these phases is elaborated upon below for comprehensive clarity.

1. Basic exploratory data analysis

The foundational comprehension of our system commences with an initial phase of exploratory data analysis. Various open-source datasets were meticulously examined to establish connections aligning with our research objectives. Through a comprehensive investigation of the dataset features, we compiled a tabulated summary encapsulating the key findings derived from this analysis.

Dataset	Motivation	Dataset selection context	Decision
Births by Borough, Ward, MSOA & LSOA **[28]**	The underlying motivation is to investigate whether there exists any correlation between crime activity and birth rates within specific geographic areas.	This dataset contains data based on Borough level which is basically a large area span.	Not Feasible and also not applicable.
London Schools Atlas **[29]**	The number of schools indicated a demographic movement in a particular area.	We want to look into if there is a correlation between the number of schools and crime activities.	Feasible
TfL Bus Stop Locations and Routes **[19]**	It has always become a fact that a station is usually a hotspot of crime events.	We would look into the distance of the crime incident against the bus and tube station and would try to make a connection.	Feasible
Jobs and Job Density, Borough **[21]**	This dataset contains the number of jobs living in a specific area span.	To determine whether crime rates are influenced by employment rates.	Not feasible as this is Borough-level information rather than street-level data.
Recorded Crime: Geographic Breakdown	The dataset includes the monthly total of crimes at	Unfortunately, this dataset does not include any GPS	Not feasible

[25]	each of London's three geographic levels (borough, ward, and LSOA), broken down by crime type.	locations; instead, it aggregates monthly criminal events. We had assumed that this dataset would serve as the base dataset.	
UK Police Crime Data Archive **[26]**		The entire crime dataset of the UK along with the archives are kept on this website.	Feasible
Stop and Search; Year - 2019; Met Police **[26]**	This dataset contains all different stop and search related data by Metropolitan police force in the UK	We can relate on if there are many stop and search happening, the place could prone to more crime activities	Feasible
On Street Crime In Camden **[23]**	This is a subset of the UK Police Crime Dataset, but this contains street level information.	Given that it contains all of a crime's Latitude and Longitude, this might serve as the foundational dataset for all of our research and prediction.	Feasible
Postcode Directory for London **[30]**	The dataset encompasses all postcodes, accompanied by their corresponding LSOA and MSOA designations. Moreover, it provides geolocation data for each postcode entry.	This dataset stands as the second most pivotal resource alongside the On Street Crime In Camden, facilitating the amalgamation of various datasets. Its significance lies in the potential to augment crime forecasting by incorporating supplementary features specific to each neighborhood, thereby enhancing our ability to anticipate and address criminal activity.	Feasible
HMO Licensing Register Dataset **[22]**	The dataset comprises information on all registered licenses within the Camden area.	This could have a correlation on the many licensed entities there the more crime activities could happen.	Feasible
Camden Markets And Kiosks dataset **[31]**	This dataset holds information on market and kiosk data in Camden area	Just like HMO license register dataset we would like to see the correlation between crime activity hotspots and number of market in that location	Feasible

Lower layer Super Output Area population estimates dataset, Mid-2019: SAPE22DT2 edition **[32]**	This dataset contains population information within the defined boundaries of Lower Layer Super Output Areas (LSOAs) for a specific geographical area.	It provides valuable data regarding the population within LSOAs, enabling correlations with specific crime incidents in distinct geographical locations. This correlation offers insights into potential relationships between population demographics and occurrences of various crime types within those specific areas.	Feasible
Ward -LSOA Lookup **[33]**		To enhance the integration of supplementary datasets with the On Street Crime In Camden dataset, we established a mapping system between Ward and Lower Layer Super Output Area (LSOA) codes. This mapping facilitated the connection between datasets by aligning the geographic boundaries represented by Ward codes with their corresponding LSOA codes. Such integration enables a more comprehensive analysis by linking diverse datasets and leveraging their collective insights within the context of specific geographic areas or administrative divisions within Camden.	Feasible
Twitter and Weather Dataset used in past research **[8]**	Tweets give a definite sign of a safe driving scenario. Different types of people tweet on various topics		Feasible
GSV dataset **[9]**	This might be an excellent tool for using images to study a certain setting.	Processing an image dataset requires a lot of CPU and GPU power. Additionally, it would not be possible to add this feature to our chosen dataset.	Not feasible for our research
Mobile Dataset from previous research **[13]**			Not possible given that there may be no publicly available mobile dataset and

			that this data only spans three weeks.
Custom designed anonymous user dataset		The research methodology involves the utilization and examination of custom-built applications specifically designed to address road safety concerns. By deploying these tailored applications, the study aims to delve deeply into user attitudes towards road safety. Through user engagement, feedback collection, and analysis of user interactions within these customized platforms, the research aims to discern nuanced perspectives, preferences, and concerns of individuals regarding road safety measures. This approach enables a targeted investigation into user attitudes, offering insights into how tailored solutions impact perceptions and behaviors related to road safety.	Feasible

Table 1: Open source dataset that we have examined for crime analysis. Among these datasets, we've reviewed and selected specific ones for further analysis and correlation in the subsequent stages of our research.

The dataset explored in **Section III.A.1** forms the basis of our exploratory data analysis. These datasets serve the purpose of comprehending diverse correlations within the crime dataset. While numerous street crime datasets are available on the UK Police website **[26]**, a predominant challenge lies in the absence of uniform features across these datasets. This disparity entails that certain features present in one dataset may not be available in another, hindering direct comparisons and analysis.

Therefore, we rely on the “On Street Crime In Camden”dataset as the pivotal foundation for our study. Within this dataset, the crucial points of interest revolve around crime categories, approximate street-level geolocations, and the precise dates and timestamps. These elements hold paramount significance for our analysis and subsequent research endeavors.

However, it's evident that while we have crime data, the most crucial components—characteristics and catalysts contributing to crime—are absent. Without accounting for these influential factors, predicting the occurrence of crime in a specific region becomes an impossible task. To integrate the Camden crime dataset into a machine learning model effectively, it's imperative to find and merge another dataset focusing on various points of interest, specifically emphasizing geolocation and timestamps. This amalgamation will enhance the dataset's depth and allow for a more comprehensive analysis.

2. Systematic approach to dataset preparation

Having explored the dataset earmarked for crime analysis and prediction, this paper now outlines the proposed steps to construct a unified dataset. This comprehensive dataset will encompass timestamps, geolocations, and an array of additional features relevant to potential crime incidents.

1. To acquire a unique timestamp, the approach involves filtering the crime dataset—specifically, the On Street Crime in Camden dataset **[23]**—to encompass solely the data from one or two designated years, such as consolidating all information from either 2019 or 2020.
2. To forecast forthcoming crime incidents, a viable approach involves creating an output label that predicts the number of expected crime events in a specific area within the upcoming days. This predictive labeling can be executed across various categories such as Antisocial behavior, Burglary, Violence, and sexual offenses, among others.
3. It's imperative to establish a mapping between our primary crime dataset and multiple open-source datasets that operate at either the LSOA (Lower Layer Super Output Area) or postcode level. This mapping process ensures integration and correlation between these disparate datasets, enhancing the depth and scope of our analysis.

Therefore, this is the dataset's basic preparation, which includes all required features. We will now examine more datasets to add those traits as the primary causes of crime events. So that we might feed the machine learning model with this new dataset. We would go through a few significant datasets and discuss how we intended to supplement them with the "On Street Crime in Camden" dataset **[23]**.

1. For instance, consider the weather dataset .We might be able to determine from this dataset whether the weather significantly affects crime incidents. To add this as a feature to the crime dataset, we choose the GPS position of each crime data set and compare it to the local weather data, such as Sunny, Rainy, Cloudy, etc.
2. If we want to incorporate the Bus and Tube station distance information, we would once more search through the criminal GPS and Bus and Tube station GPS, estimate the distance, and add this as a feature. Therefore, it would be a factor if the distance between a bus stop and a subway station had an impact on a crime.

Essentially, these are adaptable model processes that evolve with their utilization. We aim to leverage various datasets as necessary to introduce new features, ensuring flexibility and relevance throughout the analysis.

3. Model evaluation

Our key objective can be achieved as a regression problem. The subsequent section cover the main steps in developing machine learning models:

1. As previously discussed, the output label's purpose is to forecast the count of crime incidents within a particular location over a specified number of days. This predictive label serves as a key element in our crime occurrence prediction model.
2. To handle non-numerical categories, they need to be encoded into an n-class format, where 'n' represents the total number of categories, typically enumerated from 0 to n-1. Utilizing tools like LabelEncoder facilitates this encoding process, enabling the conversion of categorical data into a numerical format suitable for analysis.
3. The subsequent phase involves selecting several machine learning models and conducting a comparative analysis of their outcomes. This process allows for the assessment of model performance and the identification of the most effective approach for our predictive crime analysis.

4. To attain optimal results, fine-tuning hyperparameters for the machine learning models becomes essential. One approach to achieve this is by utilizing GridSearch from the SkLearn library **[37]**. This method enables systematic exploration and optimization of model parameters, enhancing the performance and accuracy of the predictive models.

The focal point of our prediction strategy lies in selecting the most effective machine learning models. Given that this constitutes a regression problem, our primary candidates include models such as **Linear Regression, Logistic Regression, and XGBRegressor**. These models are well-suited for regression tasks and will be evaluated to determine their suitability and performance in predicting crime occurrences.

B. In depth exploratory data analysis and data preparation

This section places significant emphasis on analyzing the datasets highlighted in **Section III.A.1**. Our primary research objective being the prediction of street-level crime activities, it's imperative to choose a dataset containing latitude and longitude coordinates. This spatial information allows us to analyze other relevant features corresponding to specific locations and times, facilitating the connection between these features and instances of crime activity.

Within each subsection below, we'll conduct both univariate and multivariate analyses of various datasets. This exploration aims to identify and select the features necessary for data preparation, ensuring a comprehensive understanding of their individual and combined impacts on crime activities at the street level.

1. On Street Crime In Camden - analysis and preparation

The dataset titled "On Street Crime In Camden" **[23]** holds street-level information, aligning perfectly with our research objectives. Our focus on a specific subset within London makes the Camden dataset an ideal choice for this study. Derived from the UK Police Crime Data Archives (data.police.uk, 2019, Met Police) **[26]**, "On Street Crime In Camden" dataset serves as our primary selection to advance our research objectives. Our initial step involves a comprehensive review of the dataset's features.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 328168 entries, 0 to 328167
Data columns (total 20 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Category          328168 non-null  object
 1   Street ID         328168 non-null  int64
 2   Street Name       328168 non-null  object
 3   Context           0 non-null       float64
 4   Outcome Category  250475 non-null  object
 5   Outcome Date      250475 non-null  object
 6   Service           328168 non-null  object
 7   Location Subtype  15697 non-null   object
 8   ID                328168 non-null  int64
 9   Persistent ID     232918 non-null  object
 10  Epoch             328168 non-null  object
 11  Ward Code         328168 non-null  object
 12  Ward Name         328168 non-null  object
 13  Easting           328168 non-null  float64
 14  Northing          328168 non-null  float64
 15  Longitude         328168 non-null  float64
 16  Latitude          328168 non-null  float64
 17  Spatial Accuracy  328168 non-null  object
 18  Last Uploaded     328168 non-null  object
 19  Location          328168 non-null  object
dtypes: float64(5), int64(2), object(13)
memory usage: 50.1+ MB
```

Figure 1: The "On Street Crime In Camden" dataset comprises 20 features, encompassing crime category details, incident time and location, alongside specific identifiers like Ward name, Ward Code, Street Name, and Street Id. This array of features enables us to establish correlations with other datasets, facilitating a deeper understanding of concurrent events during the recorded period. Furthermore, the Epoch field offers insight into the month and year of

the crime event, while Latitude and Longitude parameters provide finer granularity regarding street-level information, enhancing the dataset's spatial context.

There are about 3,28,168 rows of information contained in the dataset. In order to check for the missing values in the dataset so used **missingno** python library to visualize it.

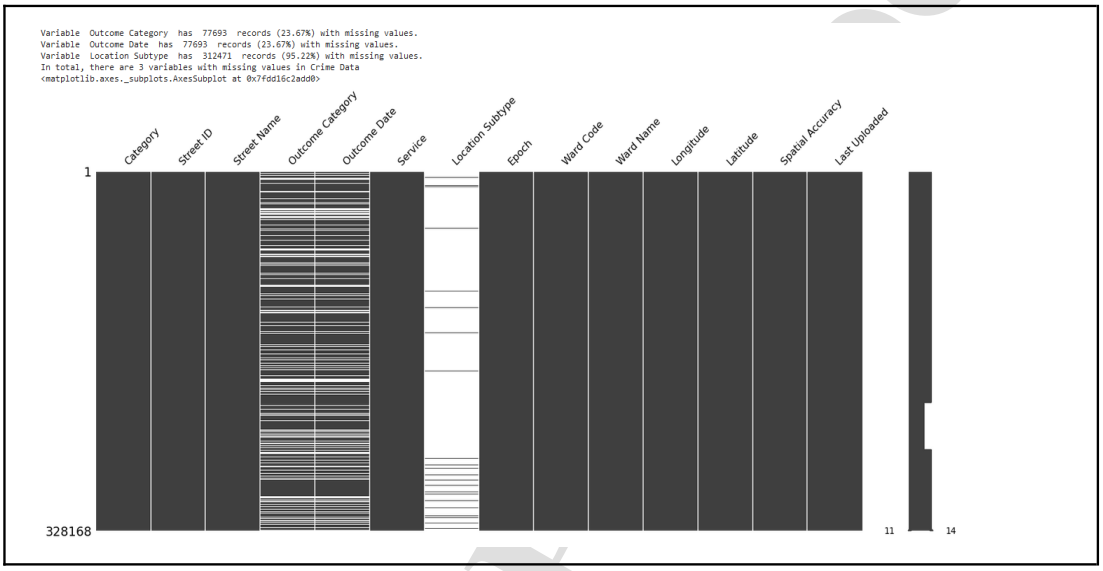


Figure 2: Visualizing missing values within the dataset reveals that certain fields contain numerous missing entries. To maintain data integrity, these fields, which are not pertinent to our analysis, will be removed from consideration. This process involves eliminating those fields with substantial missing values, ensuring a more focused dataset for our analysis.

Identifying a significant number of missing values, it becomes crucial to discern which features might contribute additional noise to our prediction model. Consequently, after removing duplicates, null values, and unnecessary features, we streamline the Camden crime dataset for future analysis. This refined dataset will consist of the selected attributes deemed essential for our predictive analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 328168 entries, 0 to 328167
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Category     328168 non-null  object
 1   Street ID    328168 non-null  int64
 2   Street Name  328168 non-null  object
 3   Service      328168 non-null  object
 4   Epoch        328168 non-null  object
 5   Ward Code    328168 non-null  object
 6   Ward Name    328168 non-null  object
 7   Longitude    328168 non-null  float64
 8   Latitude     328168 non-null  float64
dtypes: float64(2), int64(1), object(6)
memory usage: 22.5+ MB
```

Figure 3: The primary features selected from the "On Street Crime In Camden" dataset for further analysis serve as crucial points for merging additional dataset information into this comprehensive source. These selected features act as pivotal connectors for integrating and augmenting the dataset's richness and relevance.

Now that we've narrowed down the number of fields or features, our initial focus is to explore the potential predictive capacity of the "On Street Crime In Camden" dataset from various perspectives. To initiate this exploration, we've embarked on analyzing the frequency of crimes across different categories within this dataset. This approach allows us to assess the predictive power of the dataset based on the occurrence of various crime categories.

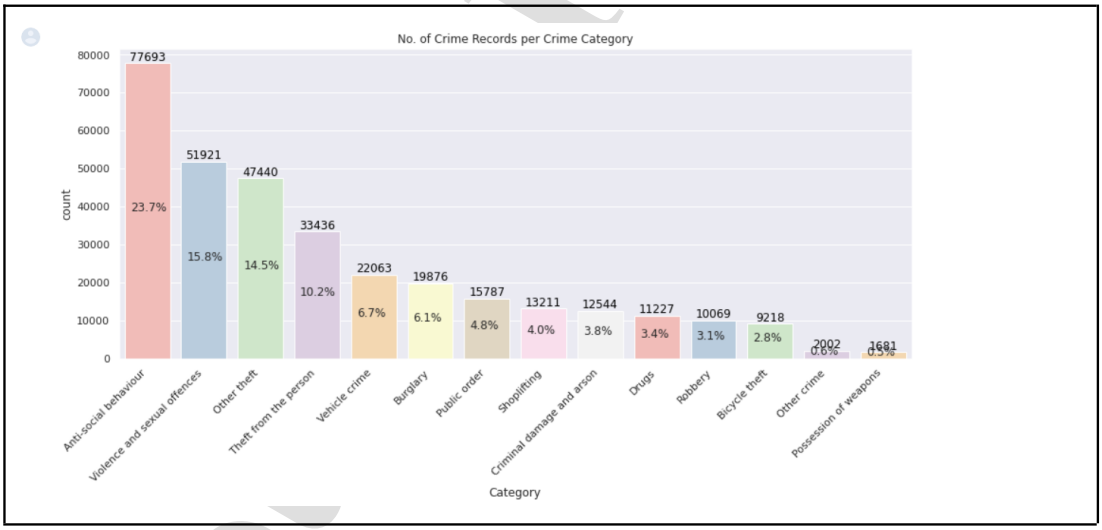


Figure 4: The analysis of the number of crimes per category reveals that Anti-social behavior stands out prominently with 77,693 occurrences, occupying the top position within the dataset. On the other hand, Possession of weapons ranks last in terms of frequency. This overview provides a clear scenario of the prevalent crime events observed in the Camden area and its surroundings, showcasing the predominant types of incidents recorded.

We also looked into how many crime incidents happened per ward.

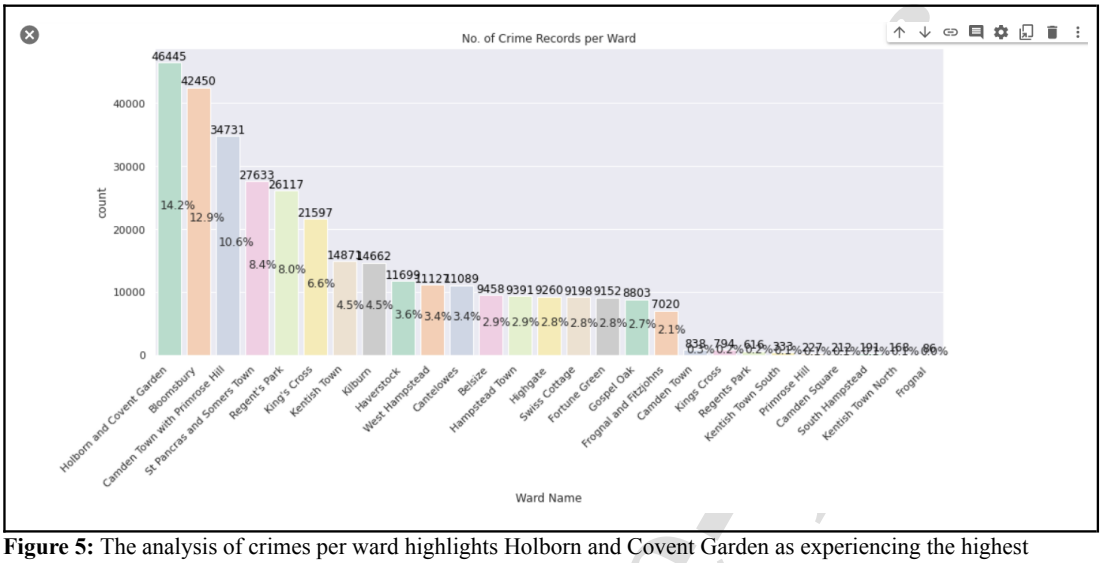


Figure 5: The analysis of crimes per ward highlights Holborn and Covent Garden as experiencing the highest number of incidents. This insight prompts a deeper examination into the surroundings of these specific wards, offering an opportunity to explore the contextual factors contributing to the heightened occurrence of crimes in these areas. Delving deeper into each ward can unveil intricate details about the localities and their potential impact on crime rates.

In preparing for real-time crime prediction, it's crucial to correlate various features from distinct datasets with the On Street Crime In Camden dataset. As discussed earlier in **Section III.A.2**, the approach involves analyzing key factors surrounding crime events at specific timestamps to ascertain catalysts for these incidents. To ensure relevance and accuracy, the decision has been made to focus on the most recent and appropriate timestamp frames for this analysis.

Understanding the context surrounding the timestamp frames, our main candidates—2019, 2020, 2021, and 2022—highlight significant periods. Considering the ongoing research in 2022, the current year's frame was eliminated. Given the exceptional impact of the Covid-19 pandemic on societal activities, particularly in 2020 and 2021, which significantly altered normalcy, we've opted to focus on the year 2019. Analyzing crime events and their relevant sequences from this period provides a more consistent and representative basis for our analysis.

So leveraging the 'Epoch' feature within the Camden dataset, we performed a filter operation, selecting data entries ranging from December 1st, 2018, to January 31st, 2020. This extended time frame enables a thorough analysis encompassing additional months, spanning from before January 1st, 2019, to after December 31st, 2019. Following this filtering process, we revisited the examination of crime incidents per ward specifically for the year 2019.

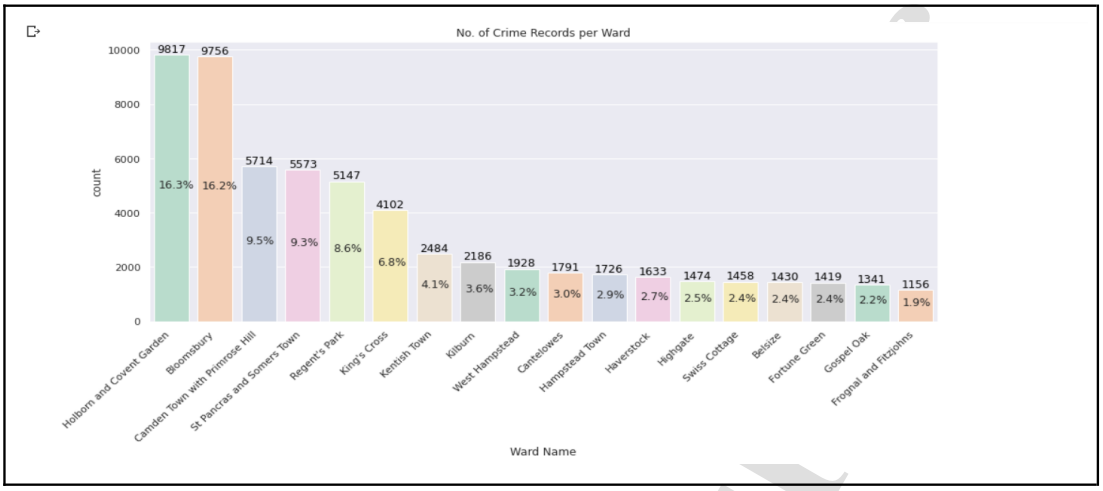


Figure 6: The consistent prominence of Holborn in recording the highest number of reported crimes in 2019 serves as a compelling reason to selectively filter specific wards. This strategic filtering process enables us to effectively merge additional datasets, particularly those that align closely with the dynamics and characteristics of Holborn, enhancing the accuracy and relevance of our analysis.

Continuing our analysis, visualizing the fluctuation of crime records in a time series format for the year 2019 unveils the periodic ebbs and flows in reported incidents. This graphical representation offers a comprehensive view of the temporal trends in crime occurrences, accentuating the peaks and troughs across various periods throughout 2019.

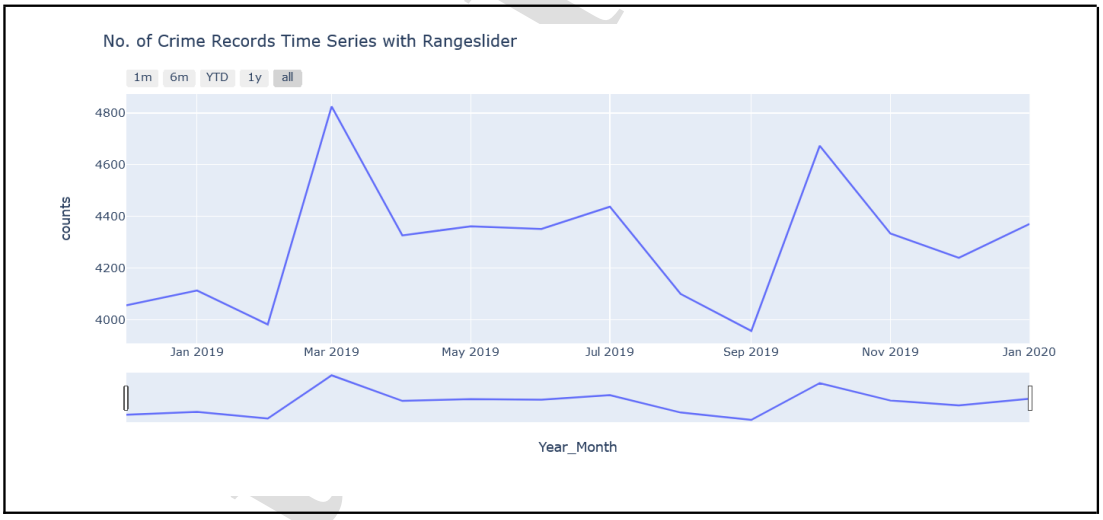


Figure 7: The time series analysis of Camden crime incidents in 2019 underscores a distinct correlation between crime occurrences and the time of the year. Observations reveal an increase in incidents during March, which aligns with expectations as temperatures normalize. Conversely, colder periods such as January and December exhibit relatively lower levels of reported crime incidents, indicating a potential weather-related influence on crime rates.

Moving forward, we delved into visualizing the crime heatmap for Camden, representing crime activities across the area. This visualization technique offers a spatial overview, presenting a graphical representation that highlights the concentration and distribution of crime incidents throughout Camden.

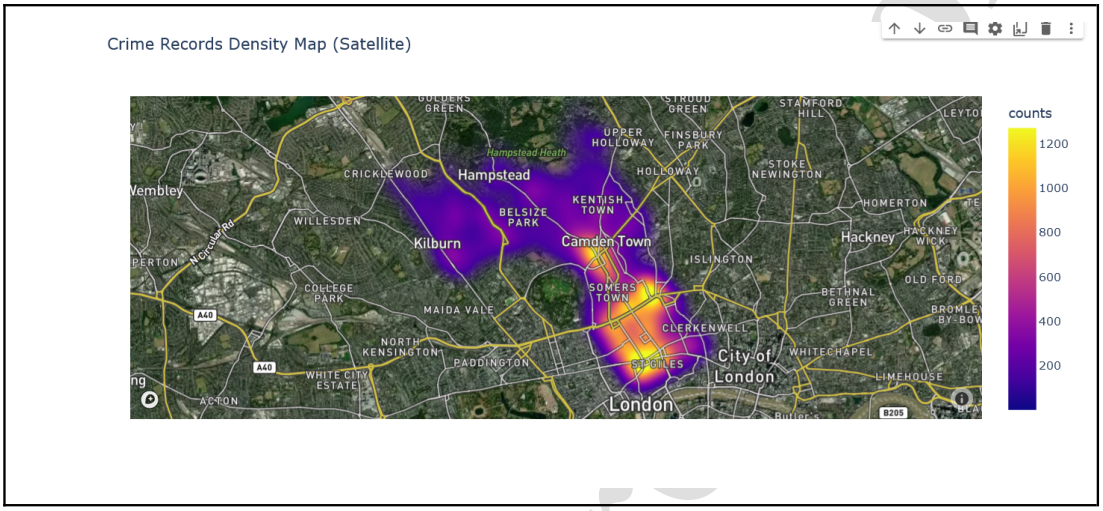


Figure 8: The Camden crime heatmap for the year 2019 showcases a concentrated area of crime incidents on one side of the region. This visualization provides valuable insight, indicating a density of reported crimes in specific locations within Camden, allowing for a spatial understanding of where these incidents predominantly occur.

It seems like crime is dense in some particular places. So this heatmap made us think out of the box and we looked into another dataset to verify our hypothesis. We looked into **HMO Licensing Register [22]**, to find out how many licenses registered in a particular ward has.

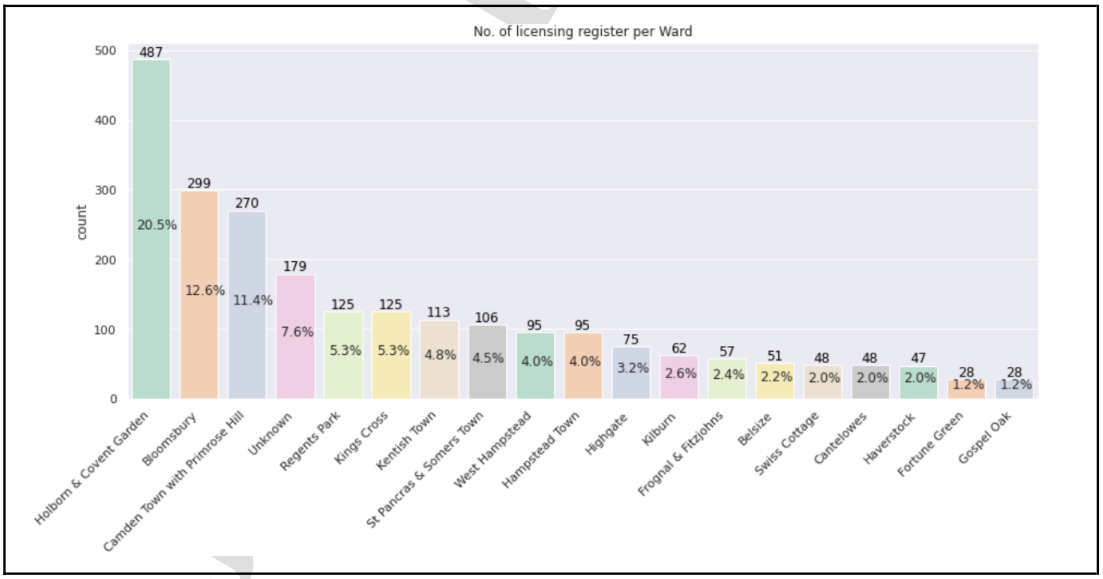


Figure 9: The correlation between the number of wards identified as crime hotspots and the count of registered licenses in Camden is evident. Wards experiencing higher crime rates seem to coincide with those having a greater

number of registered businesses. This observation suggests a potential relationship between the density of business establishments and the occurrence of crimes, indicating that wards with more businesses also tend to face a higher number of reported incidents.

Filtering the existing Camden crime dataset based on the top 3 wards with the highest crime occurrences and the bottom 3 wards with the least crime incidents in the year 2019 was a strategic step to simplify analysis. By categorizing the top 3 wards as the most active crime hotspots and the bottom 3 wards as areas with the least crime occurrences, we've created a subset that allows for a focused examination of these distinct categories, facilitating a more streamlined and targeted analysis of safety and crime trends within these specific wards.

	Ward Name	Ward Code	CrimeCount
7	Holborn and Covent Garden	E05000138	9817
4	Bloomsbury	E05000129	9756
9	Camden Town with Primrose Hill	E05000130	5714
10	Fortune Green	E05000132	1419
13	Gospel Oak	E05000134	1341
5	Frognal and Fitzjohns	E05000133	1156

Figure 10: The wards in the Camden area with the highest and lowest counts of crime events in 2019, both the top and bottom 3.

Given the refined dataset now consisting of crime information from six specific wards, our focus shifted to simulating crime incidents into real-time data. Despite the 'Epoch' feature in the On Street Crime In Camden dataset providing only month and year details for crime events, the absence of precise timestamps prompted us to strategize a simulation approach for crime occurrence times. This simulated process aims to fill the gap in timestamp data, enabling a more comprehensive real-time analysis as previously discussed.

1. The 'Epoch' field in the dataset comprises values formatted as month names followed by the corresponding year number, such as "Aug-2019" or "Sep-2019". This structure predominantly includes the month name and the associated year, providing a chronological indication for each recorded crime event.
2. Through Python scripting, we standardized the format to MM-DD-YYYY. As the 'Epoch' values lacked specific day information, we assigned the 1st day for each month. Consequently, "Aug-2019" was transformed to "08-01-2019," while "Sep-2019" became "09-01-2019," maintaining consistency across the dataset by setting the day as the 1st for all converted dates.
3. We added a random number between 1 and 30 into the 'Epoch' feature, creating a new attribute labeled 'timestamp'. Essentially, this addition simulates the day within the month when the crime event purportedly occurred, enriching the dataset with a simulated timeline for each recorded incident.
4. From this ‘timestamp’ feature we extracted the week number and month number and added these as new features.
5. The week number also simulates another feature which is weather. Basically the weather changes on week numbers so we do not need to add additional features such as weather in our dataset whereas our week number could act as a weather indicator in a particular location.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29203 entries, 0 to 29202
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Category     29203 non-null  object 
 1   Street ID    29203 non-null  int64  
 2   Street Name  29203 non-null  object 
 3   Epoch        29203 non-null  object 
 4   Ward Code    29203 non-null  object 
 5   Ward Name    29203 non-null  object 
 6   Longitude    29203 non-null  float64
 7   Latitude     29203 non-null  float64
 8   timestamp    29203 non-null  object 
 9   monthNumber  29203 non-null  int64  
 10  weekNumber   29203 non-null  int64  
dtypes: float64(2), int64(3), object(6)
memory usage: 2.5+ MB
```

Figure 11: The updated features in the On Street Crime In Camden dataset now include the addition of timestamp, monthNumber, and weekNumber attributes. These new features augment the dataset, providing valuable information related to the timing and temporal aspects of the recorded crime incidents.

To establish relationships between the On Street Crime In Camden dataset and other datasets, we're leveraging postcodes or LSOAs as common identifiers. We utilized https://findthatpostcode.uk, providing latitude and longitude coordinates of the crime event locations to retrieve the respective postcode and LSOA information associated with each location. This process enables us to link these geographical coordinates to specific postcodes and LSOAs, facilitating integration with other datasets based on these shared identifiers.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Category	Street ID	Street Name	Epoch	Ward Code	Ward Name	Longitude	Latitude	timestamp	monthNumber	weekNumber	PostCode	LSOA
2	Other theft	960645	On or near Whit	Aug-19	E05000129	Bloomsbury	-0.13839	51.52366	8/15/2019	8	33	W1T 5JU	E01000854
3	Shoplifting	965228	On or near Colle	Aug-19	E05000133	Frognal and Fitzjohns	-0.17515	51.54487	8/15/2019	8	33	NW3 5DR	E01000883
4	Other theft	960532	On or near Quee	Aug-19	E05000129	Bloomsbury	-0.13545	51.52222	8/29/2019	8	35	W1T 7NZ	E01000854
5	Theft from the pe	1488506	Holborn (lu Stati	Aug-19	E05000138	Holborn and Covent Gar	-0.12021	51.5174	8/2/2019	8	31	WC1V 7DZ	E01000914
6	Anti-social behavi	960515	On or near Parki	Aug-19	E05000129	Bloomsbury	-0.12611	51.52347	8/12/2019	8	33	WC1H 0NH	E01000855
7	Other theft	960589	On or near Furth	Aug-19	E05000129	Bloomsbury	-0.12932	51.52152	8/1/2019	8	31	WC1B 5DG	E01000855
8	Criminal damage	960513	On or near Parki	Aug-19	E05000129	Bloomsbury	-0.12651	51.52394	8/12/2019	8	33	WC1N 1HE	E01000852
9	Violence and sexu	956624	On or near New	Aug-19	E05000129	Bloomsbury	-0.12969	51.51651	8/19/2019	8	34	WC1A 1HL	E01000855
10	Public order	964929	On or near Thea	Aug-19	E05000130	Camden Town with Prim	-0.14432	51.54059	8/11/2019	8	32	NW1 7BW	E01000863
11	Bicycle theft	967549	On or near Lyncr	Aug-19	E05000132	Fortune Green	-0.19248	51.55348	8/13/2019	8	33	NW6 1LB	E01000874
12	Anti-social behavi	956453	On or near St Gil	Aug-19	E05000138	Holborn and Covent Gar	-0.12814	51.5145	8/15/2019	8	33	WC2H 8AL	E01000919
13	Other theft	960588	On or near Furth	Aug-19	E05000129	Bloomsbury	-0.13219	51.52542	8/18/2019	8	33	WC1H 0AJ	E01000852
14	Other theft	956620	On or near Percy	Aug-19	E05000129	Bloomsbury	-0.1338	51.51786	8/7/2019	8	32	W1T 1DE	E01000850
15	Public order	964928	On or near Night	Aug-19	E05000130	Camden Town with Prim	-0.143	51.53968	8/25/2019	8	34	NW1 8QP	E01000863
16	Anti-social behavi	964929	On or near Thea	Aug-19	E05000130	Camden Town with Prim	-0.14432	51.54059	8/19/2019	8	34	NW1 7BW	E01000863
17	Shoplifting	961055	On or near Night	Aug-19	E05000138	Holborn and Covent Gar	-0.11196	51.51849	8/2/2019	8	31	WC1V 6DF	E01000914
18	Anti-social behavi	961004	On or near Red L	Aug-19	E05000138	Holborn and Covent Gar	-0.11706	51.51921	8/19/2019	8	34	WC1R 4PS	E01000914
19	Anti-social behavi	964854	On or near Prow	Aug-19	E05000130	Camden Town with Prim	-0.14058	51.54231	8/5/2019	8	32	NW1 9PN	E01000856
20	Vehicle crime	967906	On or near Mack	Aug-19	E05000134	Gospel Oak	-0.16002	51.55488	8/2/2019	8	31	NW3 2LT	E01000888
21	Other theft	960700	On or near Hosp	Aug-19	E05000129	Bloomsbury	-0.13703	51.52486	8/24/2019	8	34	NW1 2BU	E01000854
22	Public order	956712	On or near Shaft	Aug-19	E05000138	Holborn and Covent Gar	-0.12711	51.51505	8/24/2019	8	34	WC2H 8JR	E01000919
23	Anti-social behavi	960558	On or near Endsl	Aug-19	E05000129	Bloomsbury	-0.1318	51.52637	8/15/2019	8	33	WC1H 0EB	E01000852
24	Violence and sexu	960749	On or near Gooc	Aug-19	E05000129	Bloomsbury	-0.13679	51.51897	8/6/2019	8	32	W1T 4NE	E01000851
25	Anti-social behavi	967771	On or near Arkw	Aug-19	E05000133	Frognal and Fitzjohns	-0.18002	51.55212	8/23/2019	8	34	NW3 6BG	E01000879
26	Theft from the pe	956555	On or near Thea	Aug-19	E05000138	Holborn and Covent Gar	-0.12612	51.51628	8/6/2019	8	32	WC2H 8DF	E01000919

Figure 12: This snapshot of the On Street Crime In Camden dataset encompasses all the newly integrated features discussed earlier. The timestamp feature provides simulated real-time information, while PostCode, LSOA, Latitude, and Longitude attributes offer comprehensive street-level granularity, enhancing the dataset's depth and spatial context.

2. License register, Market and Station datasets - analysis and preparation

In continuation of our research, we'll proceed to explore the next three datasets (**[22], [31], [19]**), each containing relatively smaller amounts of data as previously reviewed in **Section III.A** . These datasets will undergo traversal, followed by a merging process with the On Street Crime In Camden dataset based on distinct conditions or

criteria. This integration aims to enrich the Camden dataset, leveraging information from these diverse sources to augment our analytical capabilities and insights.

The HMO License register dataset **[22]** holds critical information about Camden's permitted HMOs (Houses in Multiple Occupation) and their management arrangements. This dataset plays a pivotal role in establishing a link between crime activities and license registers. It encompasses crucial features that enable this linkage and includes essential details regarding permitted HMOs and their management arrangements within the Camden area.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2370 entries, 0 to 2369
Data columns (total 21 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Licence Reference                   2370 non-null   object
 1   Premises Name                       2336 non-null   object
 2   Premises Address                    2330 non-null   object
 3   Status                              2370 non-null   object
 4   Ward Name                           2338 non-null   object
 5   Correspondence Address              1813 non-null   object
 6   Designated Premises Supervisor      1755 non-null   object
 7   Licensee                            2114 non-null   object
 8   Managing Agent                      167 non-null    object
 9   Licence Type                        2365 non-null   object
 10  Granted Or Refused Date             1394 non-null   object
 11  Expiry Date                         80 non-null     object
 12  Cumulative Impact Policy Area Name  556 non-null    object
 13  Easting                             2194 non-null   float64
 14  Northing                            2194 non-null   float64
 15  Longitude                           2194 non-null   float64
 16  Latitude                            2194 non-null   float64
 17  Location                            2194 non-null   object
 18  Spatial Accuracy                    2370 non-null   object
 19  Last Uploaded                       2370 non-null   object
 20  Organisation URI                    2370 non-null   object
dtypes: float64(4), object(17)
memory usage: 389.0+ KB
```

Figure 13: In the HMO License register dataset, the Ward Name stands out as the most crucial feature. Leveraging this feature allows us to conduct a grouping operation based on Ward Name, enabling a count of HMOs within each ward. This count can then be merged with the On Street Crime In Camden dataset, facilitating an enriched analysis by associating HMO counts with crime incidents across various wards in Camden.

While we previously explored and visualized this relationship in **Section III.B.1**, the correlation demonstrates the association between different wards and the number of registered licenses. This relationship serves as a critical factor in understanding the distribution of HMO licenses' across various wards within Camden.

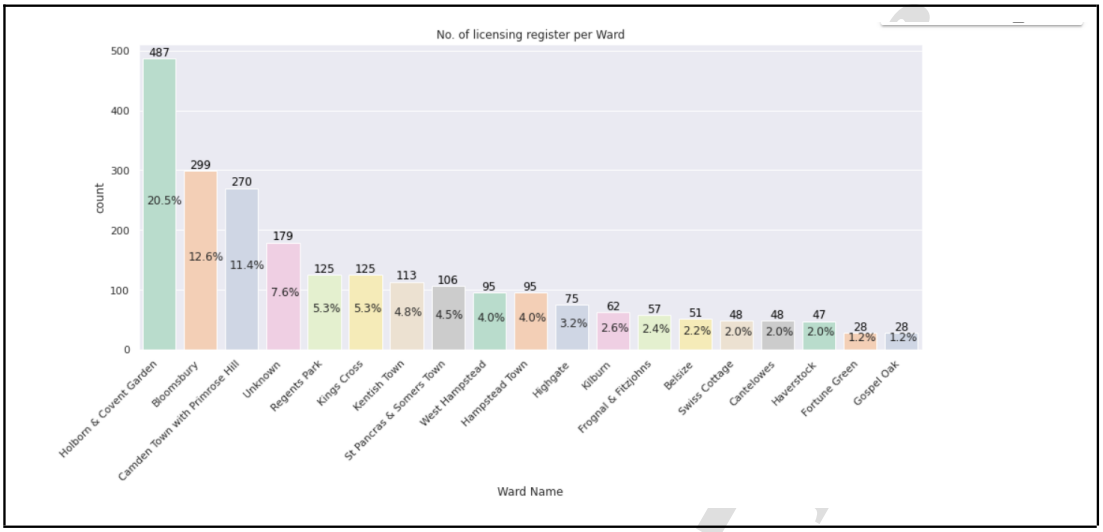


Figure 14: The correlation between Wards and the count of license registers has already been established. Utilizing this correlation, we've segmented the dataset, focusing on the top and bottom 3 wards in terms of license registers. This segmentation strategy aids in examining the association between these specific wards and the occurrences of crime incidents.

The provided density heatmap showcases the distribution of HMOs across the city of Camden.

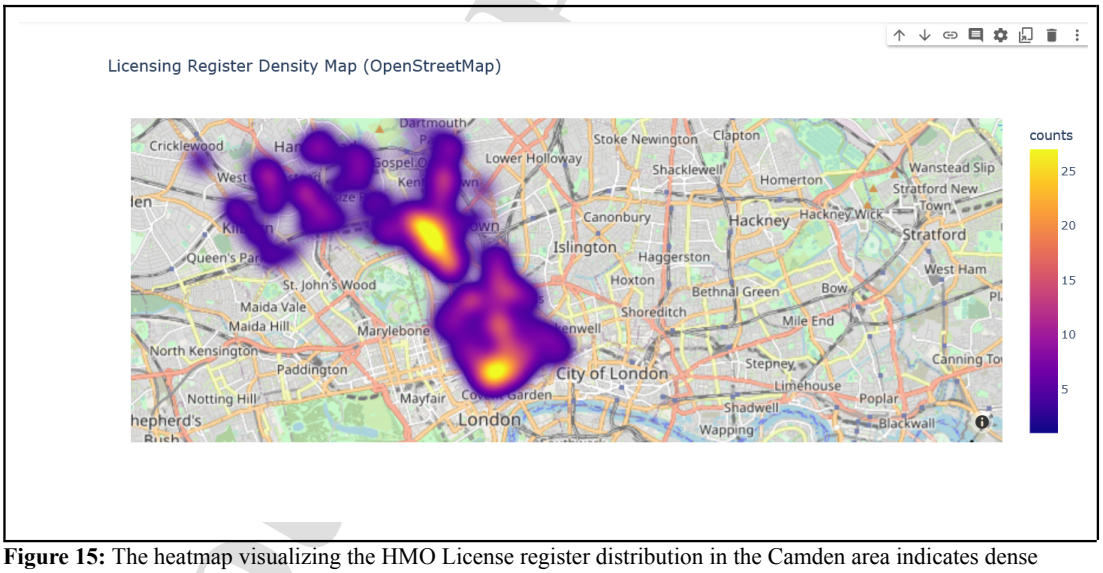


Figure 15: The heatmap visualizing the HMO License register distribution in the Camden area indicates dense clusters of license registers in specific locations. This visual representation offers insights into the concentrated areas where these license registers are prominently situated within Camden.

Next, The Market kiosk dataset **[31]** is similar to HMO License register dataset **[22]**. It shows the number of markets in the Camden area. These are the features of the Market-Kiosk dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 227 entries, 0 to 226
Data columns (total 10 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Type                         227 non-null    object
 1   Location Name                227 non-null    object
 2   Commodity                    61 non-null     object
 3   Trading Days                 61 non-null     object
 4   Trading Hours                227 non-null    object
 5   Number Of Pitches Or Kiosks  61 non-null     float64
 6   Category                     227 non-null    int64
 7   Spatial Accuracy             61 non-null     object
 8   Location                     227 non-null    object
 9   Last Uploaded                227 non-null    object
dtypes: float64(1), int64(1), object(8)
memory usage: 17.9+ KB
```

Figure 16: Within the Market-Kiosk dataset, the "Location" field harbors geospatial data that presents an opportunity for aggregation. This information can facilitate the counting of markets within specific Wards or locations. The resulting count can then be integrated into the On Street Crime In Camden dataset, enriching it with data pertaining to market occurrences in different areas or wards..

Then, we looked into different market types into the visualization.

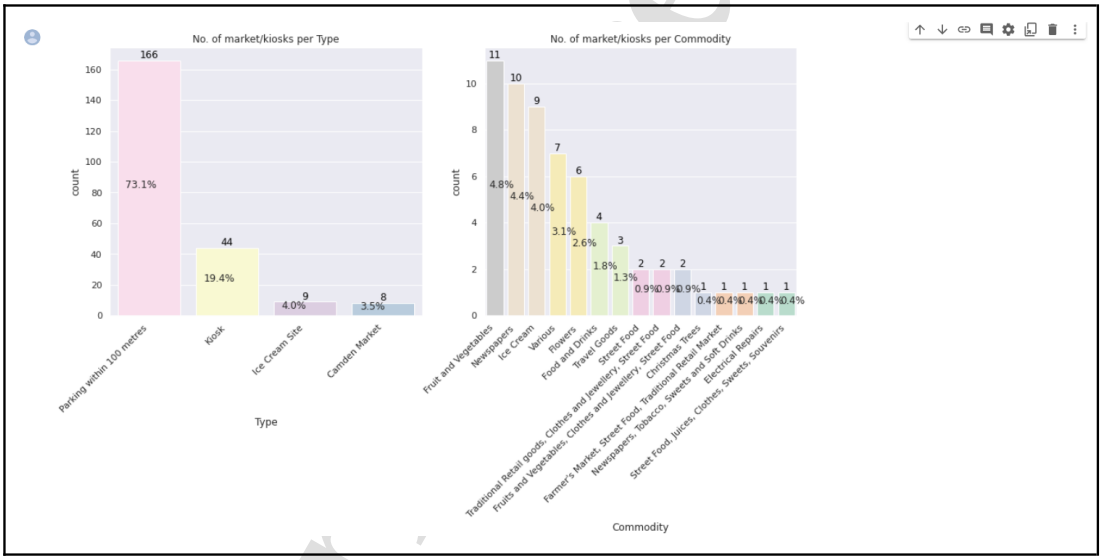


Figure 17: The count of market types within the Camden area presents an opportunity to gain insights into human behaviors based on diverse market categories. Analyzing these market type counts allows for a deeper understanding of consumer trends and preferences, shedding light on varied human behaviors associated with different market types in the region.

Moving on to the Tfl stations dataset **[19]**, our objective was to explore a potential correlation between the count of stations and crime activities. This dataset comprises specific features related to the stations, which we examined to establish any potential associations with crime incidents.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3872 entries, 0 to 3871
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   UniqueId         3872 non-null   object
 1   StationUniqueId  3872 non-null   object
 2   AreaName         3872 non-null   object
 3   AreaId           3872 non-null   int64
 4   Level            3872 non-null   int64
 5   Lat              3872 non-null   float64
 6   Lon              3872 non-null   float64
 7   FriendlyName     3872 non-null   object
dtypes: float64(2), int64(2), object(4)
memory usage: 242.1+ KB
```

Figure 19: The TFL Station dataset features were explored with a focus on aggregation, aiming to determine the proximity of stations to crime incidents. Utilizing the "Latitude" and "Longitude" attributes (referred to as Lat and Lon in this dataset), we performed an aggregation based on geographical coordinates to ascertain the number of stations in close proximity to crime incidents. This analysis aimed to unveil any correlations between station placement and crime occurrences.

The subsequent phase involves the integration of these datasets into the existing "On Street Crime In Camden" dataset. The following steps are proposed for this process:

1. Our analysis involves examining each instance of crime activity in Camden. During each occurrence, we calculate the proximity of markets, registered licenses, and stations in the vicinity.
2. The primary objective is to establish a correlation between the count of specific entities and the frequency of crime incidents. Various options were considered, including calculating the presence of markets, registered licenses, and stations within proximity ranges of 150/200/300/500 meters around each crime incident.
3. After deliberation, a radius of 300 meters was chosen, considering it a feasible walking distance for individuals. This parameter selection aims to assist in discerning patterns within the crime activities in that specific area.

After incorporating these features and applying filters in the primary Camden crime dataset, the current set of features comprises the following, with a corresponding number of rows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28531 entries, 0 to 28530
Data columns (total 17 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Category                            28531 non-null  object
 1   Street ID                           28531 non-null  int64
 2   Street Name                         28531 non-null  object
 3   Epoch                               28531 non-null  object
 4   Ward Code                           28531 non-null  object
 5   Ward Name                           28531 non-null  object
 6   Longitude                           28531 non-null  float64
 7   Latitude                            28531 non-null  float64
 8   timestamp                           28531 non-null  object
 9   monthNumber                         28531 non-null  int64
 10  weekNumber                          28531 non-null  int64
 11  PostCode                            28531 non-null  object
 12  LSOA                                28531 non-null  object
 13  numberOfMarketKioskWithin300M       28531 non-null  int64
 14  numberOfTubePointWithin300M         28531 non-null  int64
 15  numberOfBusStopWithin300M           28531 non-null  int64
 16  numberOfLicenceRegisterWithin300M   28531 non-null  int64
dtypes: float64(2), int64(7), object(8)
memory usage: 3.7+ MB
```

Figure 20: The current feature set of the Camden crime dataset, subsequent to the integration of Market, License Register, and Station data, encompasses the initial 13 features previously discussed, along with three additional features denoted by indices **[22]**, **[31]**, and **[19]**. These newly added features include:

- numberOfMarketKioskWithin300M: Represents the count of markets within a 300-meter radius of each crime incident.
- numberOfTubePointWithin300M: Indicates the number of train stations within a 300-meter radius of each crime incident.
- numberOfBusStopWithin300M: Signifies the count of bus stations within a 300-meter radius of each crime incident.
- numberOfLicenseRegisterWithin300M: Denotes the number of Houses in Multiple Occupation (HMOs) within a 300-meter radius of each crime incident.

3. Stop and Search dataset - analysis and preparation

The Stop and Search dataset for 2019**,** sourced from the Metropolitan Police **[26]**, compiles all recorded stop and search activities conducted by the Metropolitan Police in the UK. This combined dataset amalgamates street-level crime data with its respective outcomes, offering a detailed geographical perspective. To commence the analysis, data spanning from 2015 to 2019 was gathered. The consolidation of these datasets resulted in a single dataset comprising 1,434,065 rows of data and encompassing 15 distinct features. Here are the features present within this dataset:

```
 #   Column                                  Dtype
---  ------                                  -----
 0   Type                                    object
 1   Date                                    object
 2   Part of a policing operation            float64
 3   Policing operation                      float64
 4   Latitude                                float64
 5   Longitude                               float64
 6   Gender                                  object
 7   Age range                               object
 8   Self-defined ethnicity                  object
 9   Officer-defined ethnicity               object
 10  Legislation                             object
 11  Object of search                        object
 12  Outcome                                 object
 13  Outcome linked to object of search      bool
 14  Removal of more than just outer clothing bool
dtypes: bool(2), float64(4), object(9)
memory usage: 89.5+ KB
```

Figure 21: The dataset encompasses comprehensive records of stop-and-search incidents within a specific geographical area, emphasizing crucial features such as Latitude and Longitude. These geographic coordinates serve as pivotal markers, providing detailed insights into the precise locations where each stop-and-search occurrence transpired. By leveraging this dataset, it becomes feasible to map and analyze the spatial distribution and clustering of these incidents, enabling a deeper understanding of the geographical patterns and hotspots associated with stop-and-search activities. The Latitude and Longitude features stand as fundamental elements, facilitating spatial analysis and geographical context crucial for comprehensive assessments of stop-and-search trends within the specified location.

There are many missing values in the dataset.

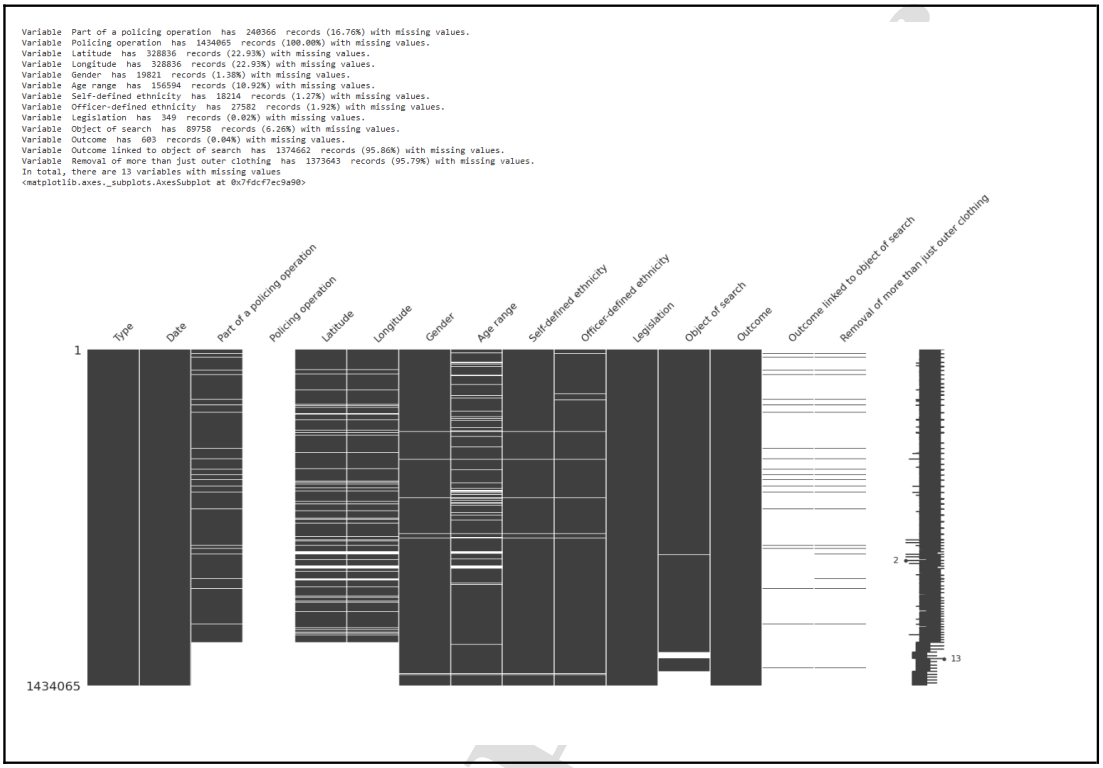


Figure 22: In the Stop-and-Search dataset, the presence of missing values, visually represented as white marks in graphical representations, is a significant consideration. These gaps in data, when visually apparent, often signify areas where information is absent or incomplete. Recognizing that certain data points might not contribute meaningfully to the analysis, the approach of selectively removing these instances becomes essential. By identifying and excluding these specific data points represented by white marks in the graphical representation, the dataset's integrity can be preserved, ensuring a more accurate and focused analysis. This strategic removal of irrelevant or incomplete data enhances the reliability and robustness of subsequent analyses performed on the Stop-and-Search dataset.

To start the dataset analysis, we first looked into age-range vs stop and search, gender vs stop and search etc.

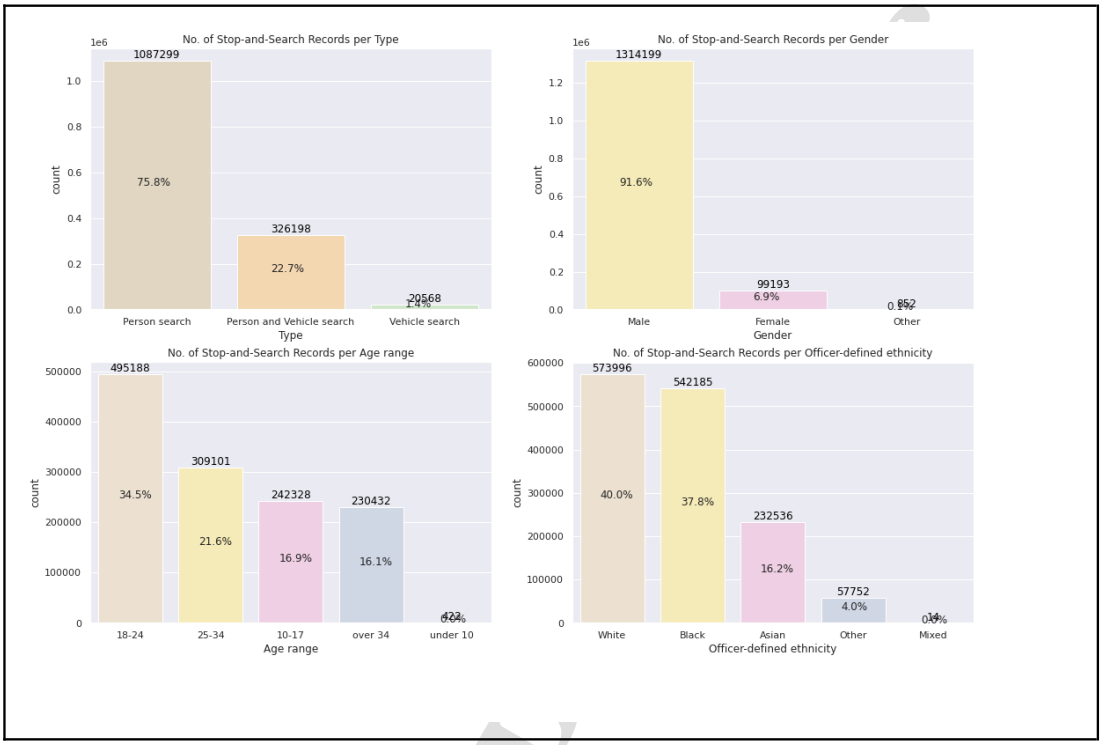


Figure 23: The univariate analysis of the Stop-and-Search dataset reveals compelling insights illustrated across four figures. In the top-left depiction, the breakdown of Stop-and-Search types displays a percentage distribution, highlighting the varied categories of searches conducted. Moving to the top-right figure, a predominant trend emerges as males constitute the primary demographic subjected to Stop-and-Search instances. The bottom-left graph indicates a noticeable trend where individuals within the 18-24 age bracket are more frequently subjected to these searches, signifying a higher incidence among the younger population. Lastly, the bottom-right figure portrays a relatively equitable distribution between white and black demographics regarding their involvement in Stop-and-Search incidents, showcasing a comparable rate of encounters for both racial groups. These visual representations collectively illuminate key demographic patterns and distributions within the Stop-and-Search dataset, offering crucial insights into the demographics most impacted by such interventions.

The comprehensive analysis of the Stop-and-Search dataset underscores prominent trends: a significant proportion of Stop-and-Search incidents primarily involve males, with a predominant focus on personal searches. Moreover, the concentration of these incidents among individuals aged 18-24 stands out as a notable observation. These findings serve as a catalyst, reinforcing our motivation to validate the initial hypothesis guiding the creation of the mentioned application. The intention behind this application lies in eliciting responses from diverse gender and age demographics, aiming to glean insights into their perceptions of safety and insecurity within specific locations. By engaging various demographics and gathering their perspectives, the application aims to collect valuable feedback on why certain areas evoke feelings of safety or unease among different gender and age cohorts. This reinforces the importance of understanding the nuanced concerns and perceptions of various demographic groups, thereby informing strategies to address and improve safety perceptions within those locations.

We also looked into the object of search vs the number of stops and searches.

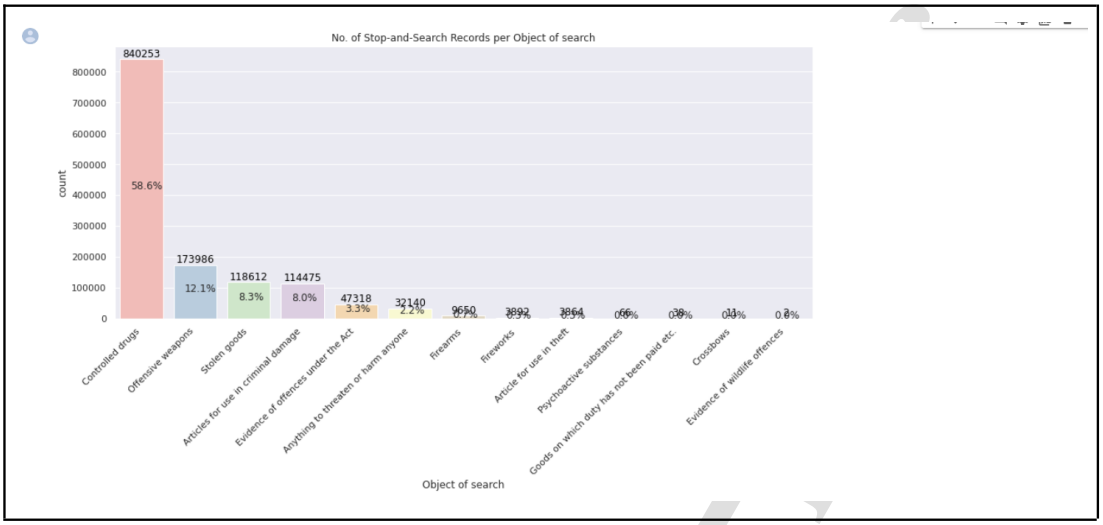


Figure 24: The analysis comparing the number of Stop-and-Search instances with the object of search reveals a distinct trend: a prevalent focus on drug-related searches. The data indicates a notable correlation between the frequency of Stop-and-Search occurrences and the object of search, highlighting a substantial emphasis on searches related to drugs.

The exploration of ethnicity profiles in correlation with stop-and-search numbers within the 2019 timespan yields crucial insights into demographic patterns associated with these interventions. Analyzing the relationship between ethnicity and the frequency of stop-and-search incidents during this timeframe provides a comprehensive understanding of how different ethnic groups are impacted by these interventions. This examination sheds light on potential disparities or disproportionate representations of specific ethnicities within the stop-and-search data, offering valuable context to assess the equity and fairness of these interventions across different demographic groups. Understanding these ethnic profiles within the scope of stop-and-search activities is instrumental in evaluating the effectiveness and fairness of law enforcement practices and aids in identifying areas for potential improvement or targeted intervention strategies to address any disparities observed among ethnic groups.

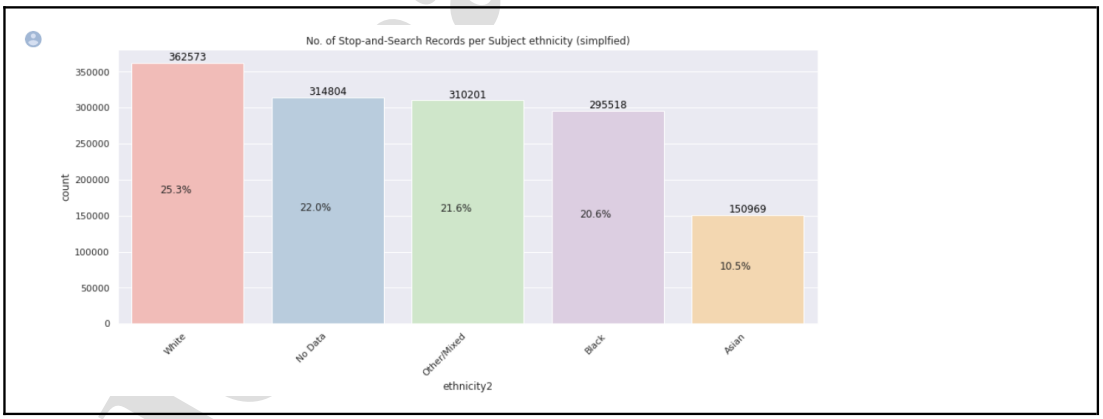


Figure 25: Ethnicity vs number of stops and searches. This also verifies that, in a particular location, different ethnic people can feel differently.

Let's have a look into the heatmap of correlation among all the features.

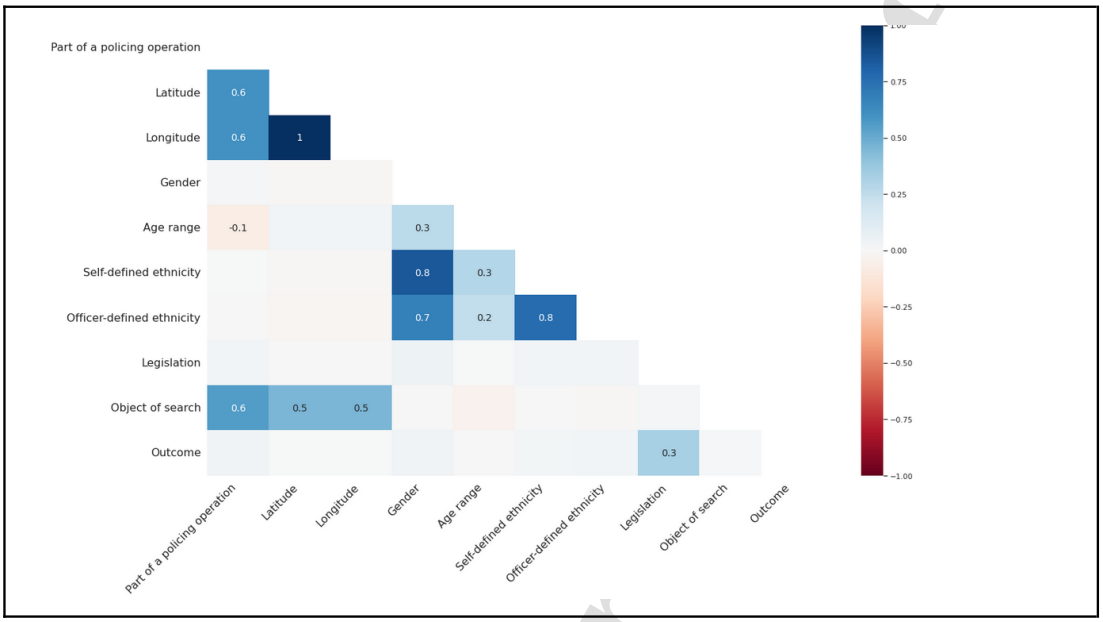


Figure 26: The correlation heatmap generated from the Stop-and-Search dataset indicates a strong influence of geolocation on policing operations, particularly concerning the object of search. The heatmap reveals notable correlations between specific geographic locations and the nature or type of items being searched for during these interventions. This suggests that certain areas or geolocations might exhibit a higher incidence or focus on particular objects of search within the context of policing operations.

We also looked into the gender percentage of the Stop and Search.

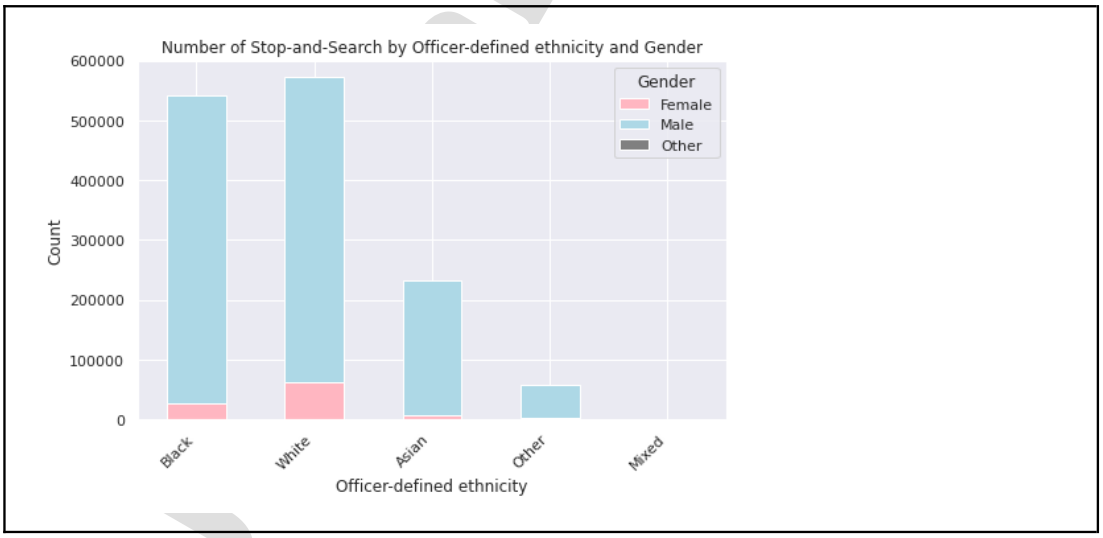


Figure 27: The analysis of gender and ethnicity concerning Stop-and-Search incidents highlights a significant predominance of males subjected to these interventions. This substantial male representation underscores the

urgency and motivation to design an application that facilitates the collection of such information more effectively. Understanding the disproportionate representation of males in these encounters underscores the need for a platform that can efficiently gather diverse demographic data, including gender and ethnicity, to comprehensively capture and analyze the nuances surrounding Stop-and-Search incidents. Such an application could significantly contribute to gathering detailed insights into the experiences and perceptions of different demographic groups, enhancing efforts to address disparities and improve the fairness and efficacy of law enforcement engagements.

The temporal analysis of stop-and-search incidents throughout 2019 reveals intriguing fluctuations, indicating certain periods characterized by notable spikes in these interventions. These spikes denote specific instances or periods within the year when there's a discernible increase in the frequency or intensity of stop-and-search activities. Understanding these fluctuations in the time series data is pivotal as it provides insights into the temporal dynamics and potential factors influencing the escalation of stop-and-search incidents during those particular periods. Further investigation into the underlying causes or contextual factors driving these spikes could offer valuable context to comprehend the fluctuating patterns in stop-and-search activities throughout the observed year.

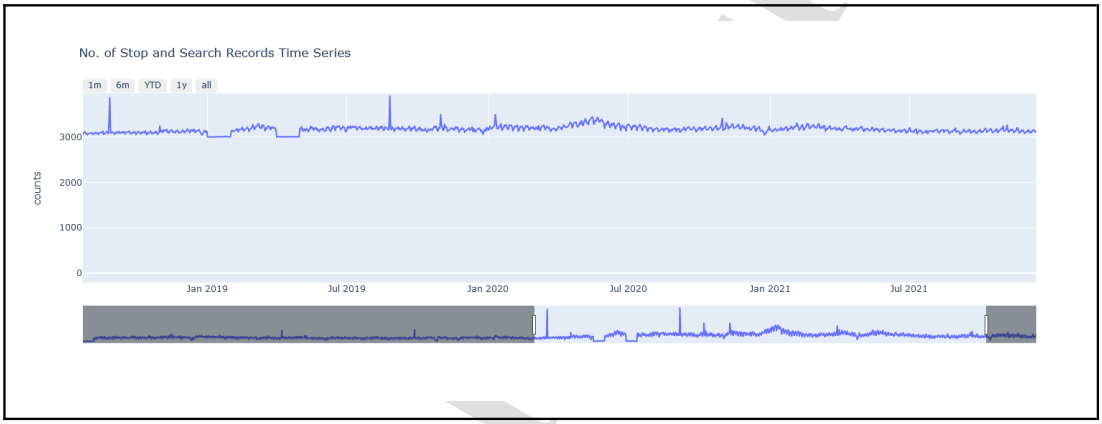


Figure 28: Time Series of stop and search in 2019. This basically verifies our hypothesis that, just like crime events, stop and search could differ based on time.

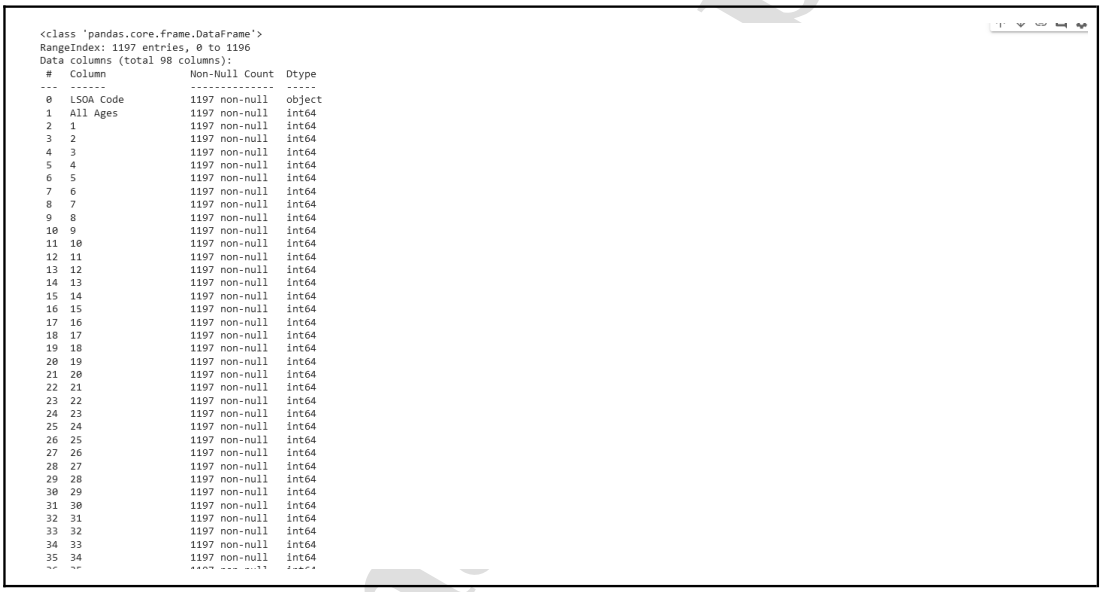
We also found out that stop and search mainly focused near different stations which basically verifies our other hypothesis on, tube and bus station may play a vital role in crime activity.

Figure 29: Camden area stop and search density map.

In our analysis of the complete Stop-and-Search dataset, we devised an aggregation method to integrate this information into the Camden crime dataset as a new feature. For every recorded crime event in the Camden dataset, we utilized the simulated timestamp to retrospectively examine the Stop-and-Search data within a one-month timeframe preceding each event. Subsequently, we added this derived count of Stop-and-Search incidents to the respective rows in the Camden crime dataset, thereby augmenting the dataset with this pertinent information for each recorded crime event.

4. Population estimate dataset - analysis and preparation

The Lower Layer Super Output Area population estimates dataset, Mid-2019: SAPE22DT2 edition **[32]**, offers valuable insights into population figures over recent years. Analyzing this dataset enables a comprehensive understanding of population trends and changes within specific geographic areas, providing a historical perspective on population dynamics. Population estimate dataset has these features.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1197 entries, 0 to 1196
Data columns (total 98 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   LSOA Code           1197 non-null   object
 1   All Ages            1197 non-null   int64
 2   1                   1197 non-null   int64
 3   2                   1197 non-null   int64
 4   3                   1197 non-null   int64
 5   4                   1197 non-null   int64
 6   5                   1197 non-null   int64
 7   6                   1197 non-null   int64
 8   7                   1197 non-null   int64
 9   8                   1197 non-null   int64
 10  9                   1197 non-null   int64
 11  10                  1197 non-null   int64
 12  11                  1197 non-null   int64
 13  12                  1197 non-null   int64
 14  13                  1197 non-null   int64
 15  14                  1197 non-null   int64
 16  15                  1197 non-null   int64
 17  16                  1197 non-null   int64
 18  17                  1197 non-null   int64
 19  18                  1197 non-null   int64
 20  19                  1197 non-null   int64
 21  20                  1197 non-null   int64
 22  21                  1197 non-null   int64
 23  22                  1197 non-null   int64
 24  23                  1197 non-null   int64
 25  24                  1197 non-null   int64
 26  25                  1197 non-null   int64
 27  26                  1197 non-null   int64
 28  27                  1197 non-null   int64
 29  28                  1197 non-null   int64
 30  29                  1197 non-null   int64
 31  30                  1197 non-null   int64
 32  31                  1197 non-null   int64
 33  32                  1197 non-null   int64
 34  33                  1197 non-null   int64
 35  34                  1197 non-null   int64
```

Figure 30: The Population Estimate dataset features detailed demographic information for each Lower Layer Super Output Area (LSOA) boundary, presenting the population count based on different age groups. This comprehensive dataset offers a breakdown of the number of individuals residing within each LSOA boundary, categorized by age ranges. The granular segmentation by age allows for a nuanced understanding of the population distribution within specific geographic areas, facilitating analyses related to age demographics and population dynamics across various regions.

We conducted an analysis to determine the total population per ward by leveraging the Ward-LSOA Lookup **[33]** dataset, which facilitated the conversion of each Lower Layer Super Output Area (LSOA) of Camden to its corresponding Ward. This conversion process allowed us to aggregate and calculate the combined population figures for each ward, providing a comprehensive understanding of the total population within specific administrative divisions or wards within the Camden area.

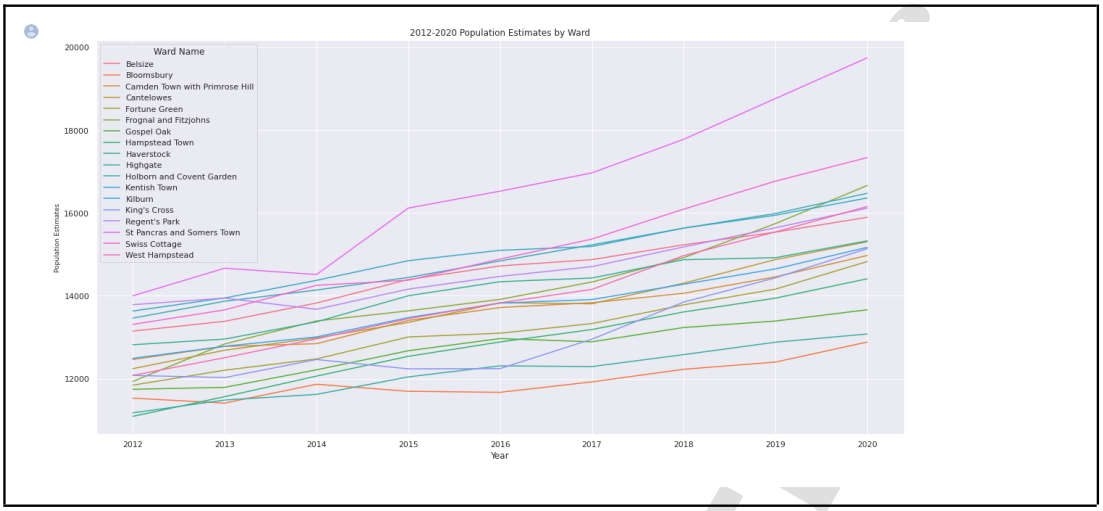


Figure 31: Estimating the population per ward in Camden is a valuable addition, especially when correlating it with the aggregated data of various age groups within the selected wards from the On Street Crime in Camden dataset. This integration provides a more comprehensive view of the demographics, enabling a deeper analysis of crime trends and patterns within specific population segments.

In utilizing this dataset, we segmented the population data into two distinct groups: "young_population," encompassing individuals aged between 15 and 40, and "total_population," representing the entire spectrum of age groups. Our hypothesis centers around exploring whether the total population and the concentration of the young population within a specific geolocation exert an influence on the frequency or occurrence of crime activities. This segmentation allows us to investigate potential correlations between the overall population size and the concentration of the young population within certain geographic areas and the incidence of crime events.

C. Application development for victim data collection

It's apparent that our detection of crime activity patterns lacks timestamp data, posing a challenge due to the absence of victim-perspective features. Meeting our business requirements heavily relies on timestamped crime activities. To address this gap, we have considered developing features that encapsulate the victim's viewpoint in safety data collection of crime hotspots, incorporating timing as a crucial factor in data acquisition. This broader initiative aims to encompass these aspects within our analysis, enhancing the depth and relevance of our findings concerning crime patterns and safety perceptions.

In this paper, our proposal centers on the development of a novel application aimed at facilitating the real-time publication of street safety reports by active users. We outline our strategy to create this application concurrently across Android and iOS platforms, encompassing both mobile devices and smartwatches. Additionally, we introduced Amazon DynamoDB, a NoSQL database, as the storage solution for this application. This paper serves as a comprehensive guide, detailing the meticulous steps involved in the design and construction of this innovative application, elucidating the process from inception to implementation.

1. Frontend development

The applications we're developing for various platforms—Android and iOS mobile devices, Android watches, and iOS watches—are collectively referred to as the frontend. To ensure cross-platform compatibility, we've opted for the Flutter framework **[34]**, enabling us to create unified apps for Android, iOS, and Android Watch. The primary motivation behind watch apps is to simplify user engagement, allowing easier submission of street

safety opinions compared to mobile apps, which often involve cumbersome unlocking, opening, and data transmission processes. The upcoming sections will delineate the primary phases of the application, providing users with an insight into its functionalities and usage.

1. Application would start up and check to see if a random Id had previously been produced and stored locally. If not, it would create one and save it in order to identify the specific device and application that is transmitting the data.
2. The user would get a drop-down form after the initial check. To comprehend the persona of the user, we would pose the following questions:
    a. Your age?
        i. 15-25
        ii. 25-40
        iii. 40+
    b. Your gender?
        i. Male
        ii. Female
        iii. Not to mention
    c. Your profession?
        i. Engineer
        ii. Doctor
        iii. Entrepreneur
3. We would then show them a screen where they could tap to transmit information on whether or not they felt secure in a certain geographic location or in a street.

Essentially, while the core process remains the same, the user interface will vary across both mobile and watch platforms. We'll provide an in-depth exploration of these steps within the app, detailing every action. Our application spans four different platforms, and we aim to showcase each step comprehensively across all platforms, offering a detailed insight into the unique user experiences tailored for each device.

The initiation would be the welcome message.

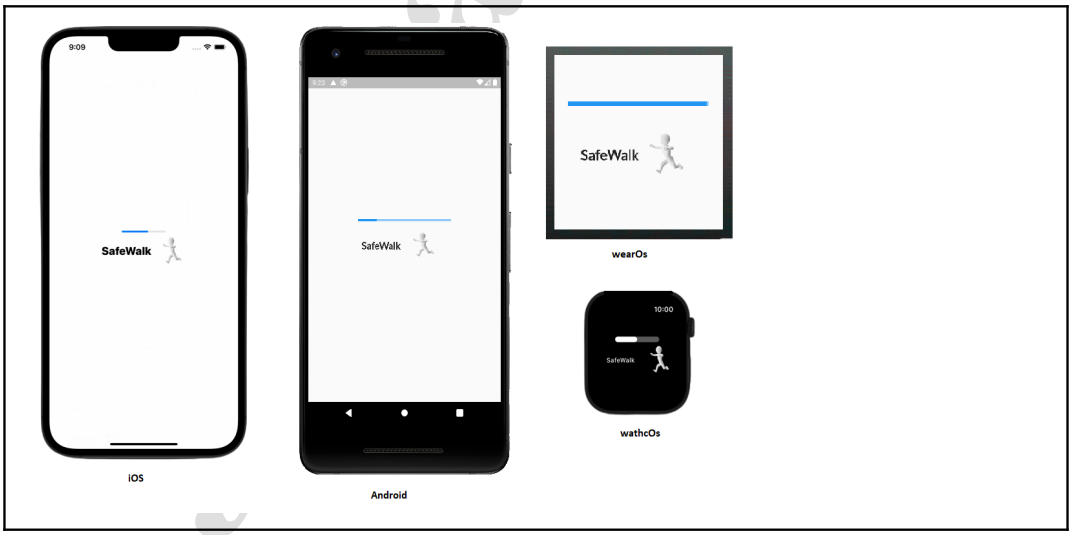


Figure 35: The initial screen users encounter upon launching the safety app, SafeWalk, serves as the entry point into the application. This screen acts as a gateway, providing users with their first interaction and navigation options within the app's interface.

After the initial welcome message, the app would check if any random id exists as the user id. If not the app would detect that this is the first the app is installed and will proceed to the page where we designed some questionnaire that would be asked to the user in order to understand the user's persona.

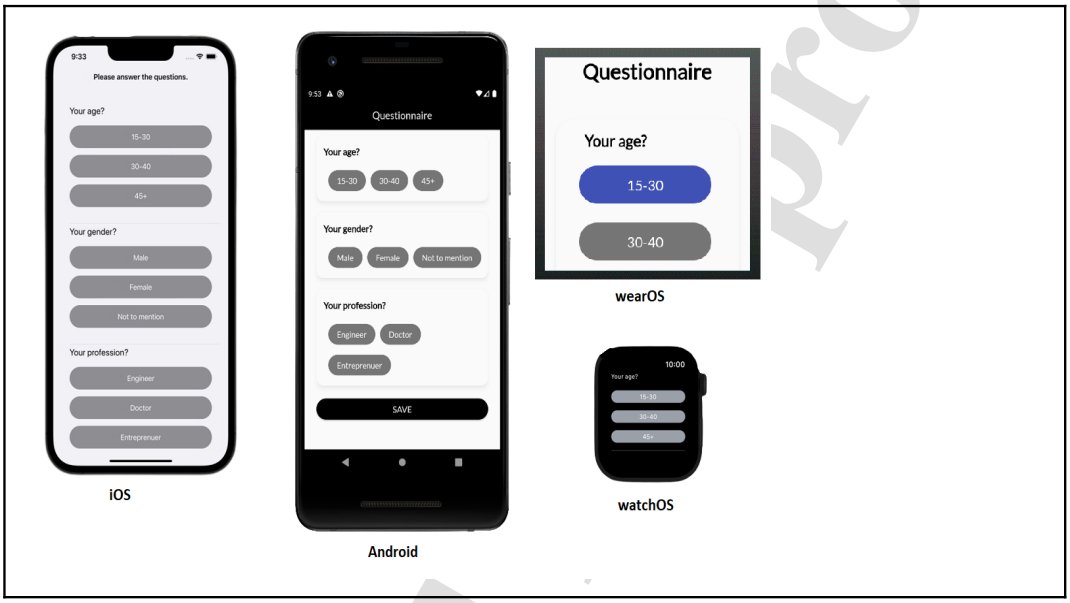


Figure 36: A questionnaire page has been designed to collect user information, providing valuable insights into the demographics of app users. This enables us to create human personas, enhancing our ability to correlate user profiles with crime or stop-and-search incidents more accurately. This data-driven approach enables a more precise understanding of how different user profiles relate to various incidents, facilitating targeted analyses and improvements within the app's functionalities.

Upon successful completion of the questionnaire, the obtained results are directed to the backend NodeJS server **[35]**. Initially, the server generates a random UUID and stores the acquired user information in Amazon DynamoDB **[36]**. Subsequently, upon receiving the response, the server sends a 200 response status along with the generated random ID, which is stored within the application as the user ID. This unique random ID serves as the user's identifier, linking all subsequent data transmitted to the server. This robust connection between data and the assigned random ID ensures coherent association and tracking of user-generated information within the application's ecosystem.

Following a successful response, the app navigates the user to the main page, facilitating the submission of their feelings for a specific location. To access location details, the app requires explicit permission for GPS Location across various platforms. This permission enables the app to accurately fetch the user's current location, ensuring that users can effectively share their feelings or safety assessments for specific locations within the application.

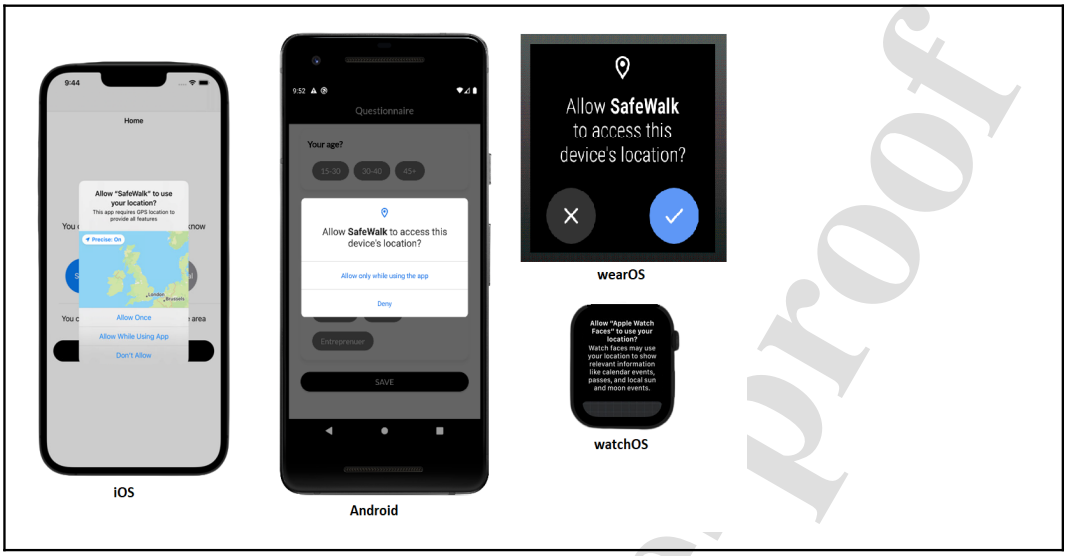


Figure 37: Indeed, the application necessitates GPS Location permission access on diverse platforms, including Android, iOS, wearOS, and watchOS. Users must explicitly grant consent for the application to transmit their location data to the server. This consent ensures that the app can securely and appropriately access and relay location information, complying with user preferences and privacy considerations while sending these details to the server for processing.

After giving the precise GPS location permission, the app would show the user the main screen.

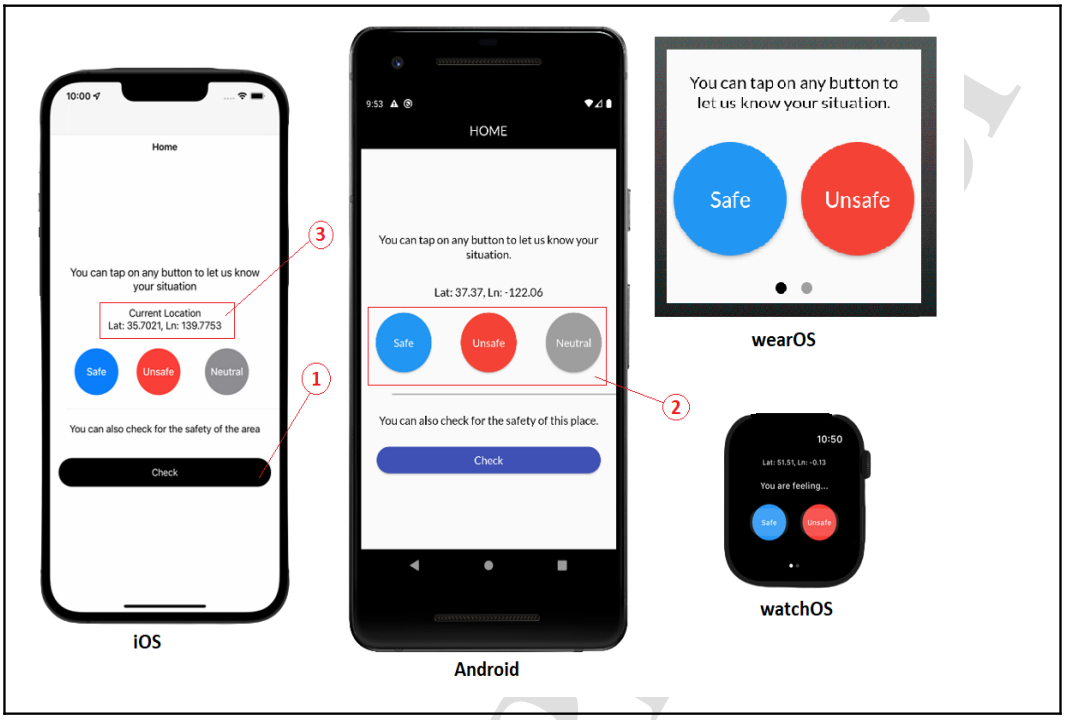


Figure 38: Each platform's main screen is meticulously designed to suit the device's characteristics and user interaction patterns, ensuring a cohesive experience while leveraging the unique capabilities of each platform.

The main screen is organized into distinct sections, each identified and delineated by chronological numbers, providing a clear and structured layout for users to navigate through various elements or features within the interface.

1. This section has a button that connects the server with the machine learning model that would take the user's location and make a prediction whether that place is safe or not. Figure shows the response message back.
2. This section consists of multiple buttons in different platforms that would send the safety feeling of the user in the server and database.
3. This section shows the current location of the user.

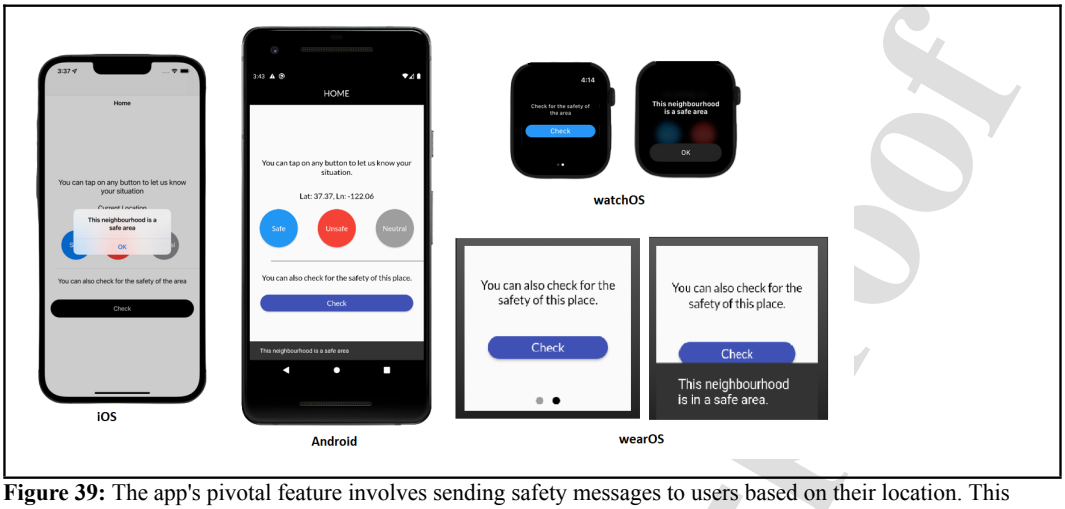


Figure 39: The app's pivotal feature involves sending safety messages to users based on their location. This functionality is closely integrated with a proposed machine learning system that operates using a pre-fed dataset. This system analyzes the location-based data and responds with assessments of whether the current location is deemed safe or not, providing users with crucial safety insights in real-time.

The Reactive design approach integrates a fallback page within the application structure to manage error occurrences across various layers—frontend, backend, and database. This page acts as a safety net, ensuring that whenever an error arises within any of these components, a standardized fallback interface is displayed to the user. This design strategy aims to provide a seamless user experience by gracefully handling errors in different scenarios, mitigating disruptions and maintaining the app's usability.

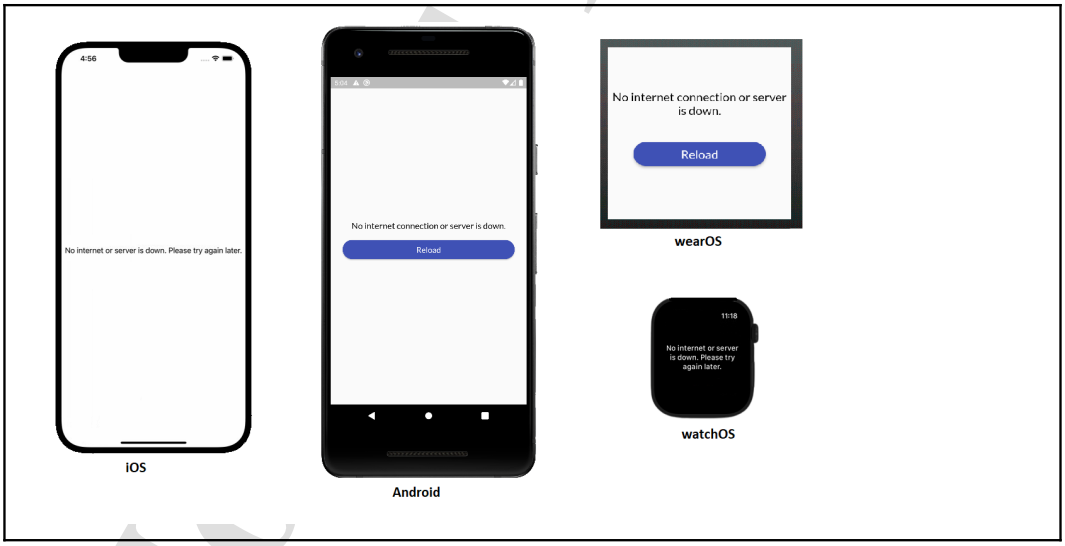


Figure 40: Implementing error handling within the framework of a Reactive design pattern ensures that users encountering errors receive accurate and context-specific messages. This approach holds immense importance as it facilitates the gathering of crucial user information related to encountered errors. By displaying the appropriate error messages, this pattern not only enhances user experience but also aids in acquiring valuable insights into the nature and frequency of errors, contributing significantly to the improvement and optimization of the application.

The architecture and design patterns of the app align with the SOLID design principle, pioneered by Robert C **[18]**. Martin, emphasizing single responsibility and modularity in code composition. This approach prioritizes code that is easily extendable for integrating new features while maintaining a clear and distinct purpose for each module. Additionally, the app has integrated the Reactive design pattern, which emphasizes resilience, reactivity, and elasticity regardless of the app's state, including instances of failure responses. This design philosophy ensures that the app consistently provides some form of response to users, even amidst challenging situations, fostering a robust and adaptable user experience.

2. Backend development

Node.js, often known as Node, is a framework built based on the Google Chrome V8 JavaScript runtime engine. **[17]** defined Node.js is an event-driven (event-based), non-blocking (non-blocking), and I/O platform built on the Google Chrome v8 engine. Asynchronous Input/Output along with an event-driven architectural design are the main characteristics of NodeJS. In the typical architecture, which deploys additional resources mainly as a multi-threaded paradigm in order to achieve high concurrency solutions, whereas in event-driven architecture, a single thread is provided and deployed to lessen the synchronous I/O requests' latency overhead. By retaining a single-threaded architecture and asynchronous request call for all I/O, Node.js restricts frequent task switching. By making use of the processing advantages of callback functions, its novel event loop system design enhances task access efficiency.

Incorporating a NodeJS server streamlines request handling, enhancing scalability and performance. Its JavaScript-based ecosystem enables rapid development and flexibility for efficient application management. Furthermore, it fulfills diverse requirements:

1. The NodeJS framework is incredibly modular and scalable.
2. Learning curve is really obvious.
3. It employs Event Driven Architecture, which may be installed on any slow machine.
4. It is a framework with only one thread.
5. It may effectively handle CPU-intensive activities because it only uses one thread for them.
6. A large user base and libraries to address many complicated issues.

Our NodeJS and database server in conjunction with the frontend application also follows the MVC design pattern. According to MVC**[16]** is an architectural design pattern that is typically applied to web-based applications. Model, view, and controller are the three main levels that are offered. MVC is a widely used design pattern among developers. It serves as a full framework. In our context our backend and frontend applications follow a complete MVC design pattern when our frontend acts as View, the backend acts as the controller and then the database acts as the Model for our system.

3. Database selection

We have a constantly dynamic data structure that must be able to fit in our database, so we would like to utilize a NoSQL database which can easily fit any data and scale horizontally. At any size, query speed for Amazon DynamoDB's NoSQL database system is in the single-digit millisecond range **[36]**. The primary explanation is that it always returns data in O(1) run time complexity queries because it stores data using a hash key.

In addition to using DynamoDB, we adopted a single-table design in which all data, including user and GPS data, is contained in a single table. It would be expensive in terms of performance to utilize a relational database and have to join numerous tables. The issue of completely forbidding the join system was resolved by DynamoDB. As there would be less overhead table creation, single table designs are less expensive. We moved into this architecture for that reason.

Relational databases don't need to worry about remembering the data access pattern because they focus primarily on the normalization process. However, while designing a single table, one must decide which key to use as a search query and how to build the access pattern. To do that, we added one Local Secondary Index called "accessPattern" which identifies a row depending on whether it contains user data or street safety data with GPS position.

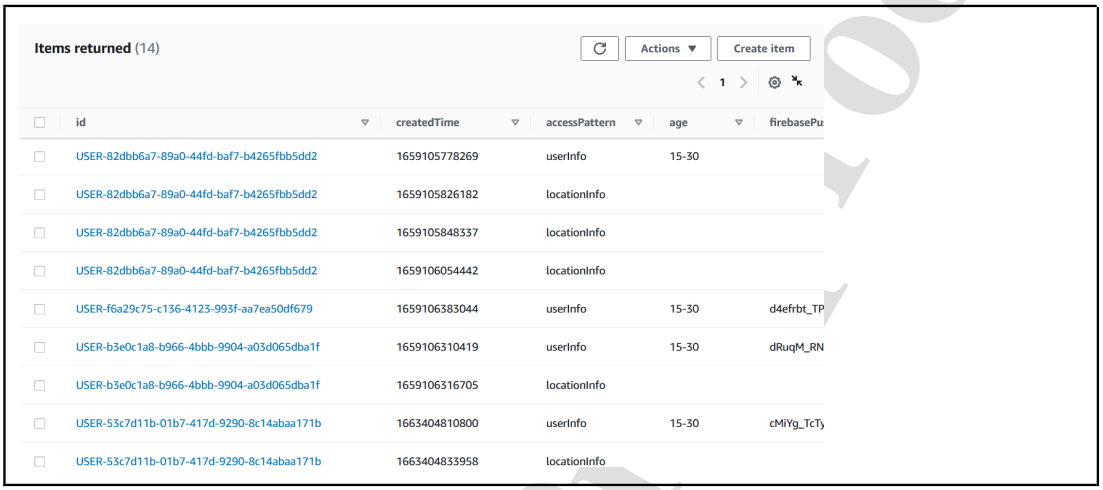


Figure 41: The database comprises collections from the application's data. The 'id' serves as a distinctive identifier for individual devices, while the 'accessPattern' categorizes the unique device data, specifying whether it pertains to user information or the user's location data.

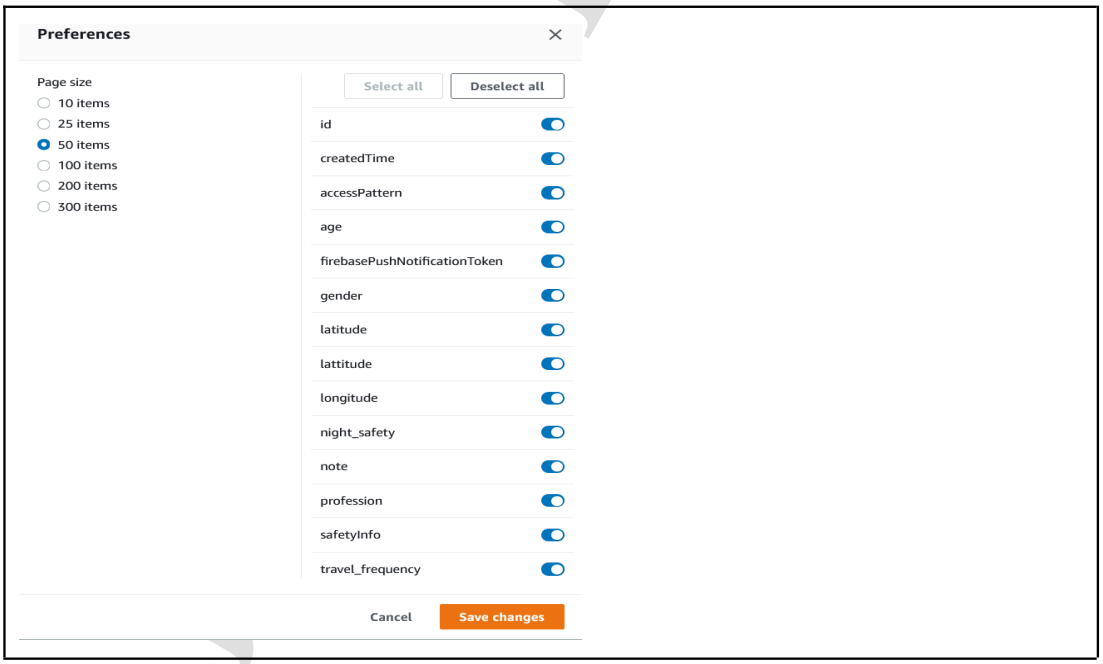


Figure 42: The app's features are centered around storing user-provided information in the database. Initially, users input personal details followed by expressing safety sentiments regarding specific geolocations. These inputs are stored and analyzed to facilitate predictive insights and analysis within the application.

We can find out that one single table holding different types of user data and these are distinguished by **accessPattern** attribute. Based on the accessPattern we store and fetch the user personal information or location information.

The table has the primary key consists of id and createdTime attribute. The id is generated from the backend from different types of data and stored in the table. Same userId can have multiple rows as createdTime would make the key very unique hashing. There is a Local Secondary Index (LSI) is created which gives us the chance to make another key consisting of the id and the accessPattern so that we can easily fetch data based on accessPattern. Also Amazon DynamoDB gives the opportunity to define any other key based on the demand which is defined as Global Secondary Index to search any data in O(1) runtime.

4. System security and privacy

The server employs UUID v4, generating completely random IDs for one-to-one user mapping, ensuring no feasible way to trace back to real individuals. Data transmission between the app and server utilizes the HTTPS protocol, encrypting data during transmission. Additionally, robust measures are in place to address vulnerabilities like SQL injection and XSS attacks, ensuring comprehensive mitigation of potential security risks within the system.

D. Dataset preparation

This section focuses solely on our unique dataset, from the analysis done on other datasets in **Section III.B**. Our custom dataset is carefully crafted to predict future crime activities using specific features we've gathered. It's the foundation for our predictive models, helping us understand and forecast potential crimes based on the specific data we've collected.

1. Initially, we selected the "On Street Crime In Camden" dataset and filtered it specifically for the data from the year 2019. The primary reason behind choosing this dataset was its inclusion of geolocation information for each recorded crime event.
2. We narrowed down the dataset by selecting the top three wards with the highest recorded crime events and the bottom three wards with the least recorded incidents. This step was taken to reduce the dataset size for further analysis.
3. We generated a simulated timestamp using the "epoch" feature, initially containing month and year details. By selecting a random number between 1 and 30, we simulated the day of the crime event and combined it with the "epoch" feature, creating a new feature termed "timestamp." This process aimed to simulate varied timings for crimes occurring in different locations.
4. We retrieved postcodes, wards, and LSOA (Lower Layer Super Output Area) details from https://findthatpostcode.uk/ by utilizing the GPS location of each recorded crime. This approach enables us to merge additional datasets based on postcodes, wards, or LSOA, facilitating a comprehensive analysis based on geographical and administrative divisions.
5. Utilizing the geolocation data, we merged information regarding the count of bus and tube stations within a 300-meter radius from the crime location in the Camden crime dataset. This step aimed to incorporate proximity-based details of transportation hubs to enrich the dataset for further analysis.
6. We incorporated data on the count of markets and license registers within a 300-meter radius from the crime hotspots, specifically based on postcodes. This merge aimed to augment the dataset with details concerning local markets and HMO licensed establishments in close proximity to the crime locations for comprehensive analysis.

7. For each entry in the Camden crime dataset, we iterated through the rows. During this process, for every recorded crime event, we examined the simulated timestamp. We then assessed the number of crime events that occurred within the dataset during the preceding one-month timeframe from that specific timestamp. Subsequently, we appended this information to the respective rows within the dataset, providing additional context regarding the frequency of crime occurrences within a month's span relative to each event.
8. Similarly, for the Stop-and-Search dataset, we undertook a comparable process. However, in this instance, we specifically focused on the last two weeks of stop-and-search activities related to the recorded crime events. This allowed us to contextualize each crime event with the stop-and-search activities occurring within a more immediate time frame, providing a snapshot of law enforcement activities closer to the events.
9. We designed a simulated feature within the app that aggregates the total number of unsafe reports received for a specific location. This process involved aggregating data from the past two weeks and appending this cumulative count to each corresponding location. Essentially, this feature provides an overview of the total unsafe reports recorded for individual locations within the specified timeframe. As the app hasn't been launched yet, we've temporarily assigned the same count as the stop-and-search numbers. This placeholder measure helps approximate potential unsafe reports until the app is deployed and starts collecting actual data.

Our primary dataset comprises a combination of open-sourced features and custom application-generated simulated features. These collectively form the comprehensive set of attributes within our dataset, blending external data with internally generated simulated characteristics for comprehensive analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24555 entries, 0 to 24554
Data columns (total 23 columns):
 #   Column                                  Non-Null Count  Dtype  
---  ------                                  --------------  -----  
 0   crime_category                          24555 non-null  object 
 1   street_id                               24555 non-null  int64  
 2   street_name                             24555 non-null  object 
 3   epoch                                   24555 non-null  object 
 4   ward_code                               24555 non-null  object 
 5   ward_name                               24555 non-null  object 
 6   longitude                               24555 non-null  float64
 7   latitude                                24555 non-null  float64
 8   timestamp                               24555 non-null  object 
 9   month_number                            24555 non-null  int64  
 10  week_number                             24555 non-null  int64  
 11  postcode                                24555 non-null  object 
 12  lsoa                                    24555 non-null  object 
 13  total_crime_count_before_one_month      24555 non-null  int64  
 14  total_crime_count_after_two_weeks       24555 non-null  int64  
 15  number_of_market_kiosk_within_300M      24555 non-null  int64  
 16  number_of_tube_point_within_300M        24555 non-null  int64  
 17  number_of_bus_stop_within_300M          24555 non-null  int64  
 18  number_of_license_register_within_300M  24555 non-null  int64  
 19  total_stop_and_search_in_past_two_weeks 24555 non-null  int64  
 20  total_population                        24555 non-null  int64  
 21  young_population                        24555 non-null  int64  
 22  total_unsafe_report_in_past_two_weeks   24555 non-null  int64  
dtypes: float64(2), int64(13), object(8)
memory usage: 4.3+ MB
```

Figure 43: The custom dataset features we've curated are specifically tailored for our prediction system. These features are meticulously selected and designed to enhance the accuracy and effectiveness of our predictive model, catering to the specific requirements of our system for crime activity forecasting. The breakdown of each feature listed below, despite prior discussions on dataset preparation, is as follows.

	Feature name	Description
1	crime_category	Includes the crime category associated with each recorded incident.
2	stree_id	Specifies the street where the crime event occurred.
3	stree_name	Indicates the specific street name where the crime event took place.
4	epoch	Holds details regarding the month and year of the crime incident.
5	ward_code	We retrieved the ward code where the crime occurred using the GPS location of the incident. Previously we tagged this as “Ward Code” in the dataset.
6	ward_name	Ward name of that ward. Previously we tagged this as “Ward Name” in the dataset.
7	longitude	The longitude geo-information denotes the specific longitudinal coordinate of the crime incident's location. Previously we tagged this as “Longitude” in the dataset.
8	latitude	The latitude geo-information represents the precise latitudinal coordinate of the crime incident's location. Previously we tagged this as “Latitude” in the dataset.
9	timestamp	Simulated time information is generated by augmenting the epoch feature with a random day number between 1 and 30.
10	month_number	Derived from the timestamp, it indicates the month in which the crime occurred. Previously we tagged this as “monthNumber” in the dataset.
11	week_number	Obtained from the timestamp, it specifies the week during which the crime occurred. Previously we tagged this as “weekNumber” in the dataset.
12	post_code	Similarly, we added the postcode by retrieving it using the same method as the ward code from the GPS location of the crime. Previously we tagged this as “PostCode” in the dataset.
13	lsoa	Similar to retrieving the ward code, we obtained the Lower Layer Super Output Area (LSOA) of the crime event from latitude and longitude data. This LSOA information is utilized for merging with other datasets, enabling comprehensive analysis. Previously we tagged this as “LSOA” in the dataset.
14	total_crime_count_before_one_month	This represents the count of crimes that occurred at this geolocation within the month preceding the day of this specific crime incident.

15	total_crime_count_after_two_weeks	This indicates the count of crimes that occurred at this geolocation within the two weeks following the day of this specific crime incident.
16	number_of_market_kiosk_within_300M	This value represents the count of markets and kiosks within a 300-meter radius from the vicinity of the crime incident. Previously we tagged this as "numberOfMarketKioskWithin300M" in the dataset.
17	number_of_tube_point_within_300M	This indicates the count of underground railway stations (tubes) within a 300-meter radius from the vicinity of the crime incident. Previously we tagged this as "numberOfTubePointWithin300M" in the dataset.
18	number_of_bus_stop_within_300M	This represents the count of bus stations located within a 300-meter radius from the vicinity of the crime incident. Previously we tagged this as "numberOfBusStopWithin300M" in the dataset.
19	number_of_license_register_within_300M	This signifies the count of House in Multiple Occupation (HMO) license registrations within a 300-meter radius from the area of the crime incident. Previously we tagged this as "numberOfLicenseRegisterWithin300M" in the dataset.
20	total_stop_and_search_in_past_two_weeks	This signifies the cumulative count of stop-and-search activities conducted in the two weeks preceding the crime incident.
21	total_population	The total population residing around the vicinity of the crime incident's geolocation.
22	young_population	The young population aged 15-40 residing around the vicinity of the crime incident's geolocation.
23	total_unsafe_report_in_past_two_weeks	The simulated feature generated by our application, at present, doesn't provide additional information. However, we devised this feature using the following formula: total_unsafe_report_in_past_two_weeks= total_stop_and_search_in_past_two_weeks + 2 * total_crime_count_before_one_month Our hypothesis is that the total number of unsafe reports will exceed the combined sum of actual stop-and-search instances and the total count of crimes.

Table 2: These are the features we've extracted for our prediction system's dataset from all the open source dataset we analyzed.

Our primary goal is to predict the number of potential crime activities in a specific location. To address this, we've created the key output label "total_crime_count_after_two_weeks," forecasting the expected number of crime incidents from the current day to the subsequent two weeks. To optimize the dataset for machine learning models, we're considering additional preparation steps. While Street ID, postcodes, latitude, and longitude serve as unique identifiers, they might introduce noise into the dataset for the models. However, through our dataset preparation,

we've observed that even if we remove these identifiers, the combination of other features, such as "total_crime_count_before_one_month" and "number_of_market_kiosk_within_300M," allows us to still pinpoint a location effectively.

Our dataset was partitioned into a training set and a test set. The training set contains 80% of our primary prepared dataset, while the test set holds the remaining 20%. Our primary predictive focus revolves around forecasting the occurrences or the count of crime activities expected in the next two weeks, considering all available features. Here are the final features selected for our predictive models:.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19644 entries, 9064 to 2915
Data columns (total 11 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   month_number                          19644 non-null  int64
 1   week_number                           19644 non-null  int64
 2   total_crime_count_before_one_month    19644 non-null  int64
 3   number_of_market_kiosk_within_300M    19644 non-null  int64
 4   number_of_tube_point_within_300M      19644 non-null  int64
 5   number_of_bus_stop_within_300M        19644 non-null  int64
 6   number_of_license_register_within_300M  19644 non-null  int64
 7   total_stop_and_search_in_past_two_weeks  19644 non-null  int64
 8   total_population                      19644 non-null  int64
 9   young_population                      19644 non-null  int64
 10  total_unsafe_report_in_past_two_weeks  19644 non-null  int64
dtypes: int64(11)
memory usage: 1.8 MB
```

Figure 44: The ultimate set of features chosen to be inputted into the machine learning model. Our end goal is to find the value of "total_crime_count_after_two_weeks".

We also tried to find all the best features among these 11 features. We used the SelectKBest method from SkLearn to find out the best features among these. These are the best features found

1. total_crime_count_before_one_month
2. number_of_license_register_within_300M
3. total_unsafe_report_in_past_two_weeks

But we came to the conclusion that in order to prevent overfitting we should go for all the features to be fed in the model and find out the outcome. And finally it is clear that this is a regression problem, so we will be using different regression models for achieving our objective.

IV. RESULT ANALYSIS

In this section, we'll delve into the outcomes derived from our prepared dataset utilized for crime activity prediction using Regression Models. To comprehensively evaluate these results, it's essential to define certain terms that aid in understanding the predictive capability of these models. These terms will serve as benchmarks for assessing the effectiveness and accuracy of our predictive models.

1. **Absolute Error**
   The difference between the actual value and the predicted value is referred to as the absolute error. The lower the better.

2. **Mean Absolute Error (MAE)**
   This represents the average of all absolute errors known as the Mean Absolute Error. The lower the better.

3. **Mean Squared Error (MSE)**
   This checks the measurement of how closely the line generated by the regression model resembles a set of points. The lower the better.

4. **Root Mean Squared Error (RMSE)**
   RMSE acts similar to MSE, but it is square rooted to make it easier to understand. The lower the better.

5. **Max Error**
   Highest Absolute Error. The lower the better.

6. **R-squared (R^2)**
   This is the metrics that defines the goodness-of-fit measurement for any regression model. It ranges from 0 to 1 but can be presented in the percentage. The higher the better.

A. Linear Regression

The initial model employed for training and testing is the Linear Regression model using default parameters. By fitting the test dataset into this model, we obtained specific results that reflect the model's performance in handling the data. These results provide valuable feedback regarding the fitting process and the model's efficacy.

Mean Absolute Error	1.817
Mean Squared Error	6.811
Root Mean Squared Error	2.609
R2 Score	70.2%
Max Error	20.971

Table 3: Linear Regression metrics. These evaluation metrics provide a comprehensive overview of a machine learning model's performance. The Mean Absolute Error (MAE) of 1.817 signifies the average absolute discrepancy between predicted and actual values. Meanwhile, the Mean Squared Error (MSE) of 6.811, along with its Root Mean Squared Error (RMSE) counterpart at 2.609, indicates the average squared and square root of the differences, respectively, emphasizing the model's sensitivity to larger errors. The R2 Score of 70.2% showcases the model's capability in explaining about 70.2% of the variance in the dependent variable, reflecting its overall explanatory power. Lastly, the Max Error of 20.971 highlights the largest observed difference between predicted and actual values, offering insight into the model's maximum prediction discrepancy. These metrics collectively aid in understanding the model's accuracy, precision, and explanatory ability on the dataset under consideration.

The results indicate that even without fine-tuning the prepared dataset, the model exhibits significant prediction capability. The notable R2 Score of approximately 70% reflects the model's ability to explain and predict crime activities based on the dataset provided. Additionally, examining sample data showcases the accuracy of calculated actual and predicted values, providing further insight into the model's performance.

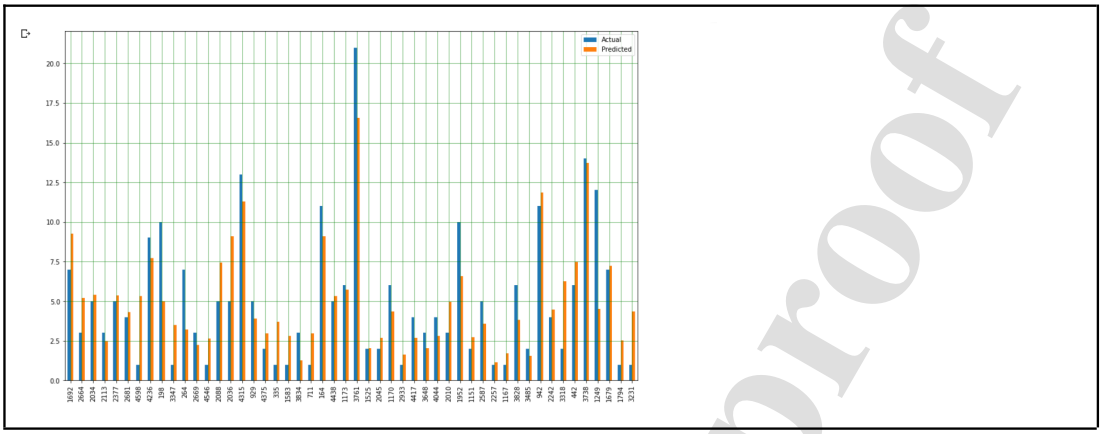


Figure 45: Displaying the actual and predicted values in the test dataset through bars visually illustrates the model's predictive ability. The comparison between these values provides a clear insight into how well the model aligns with the actual observed data, highlighting its predictive performance across various instances within the dataset.

We also found out our regression line could fit nicely among the data points.

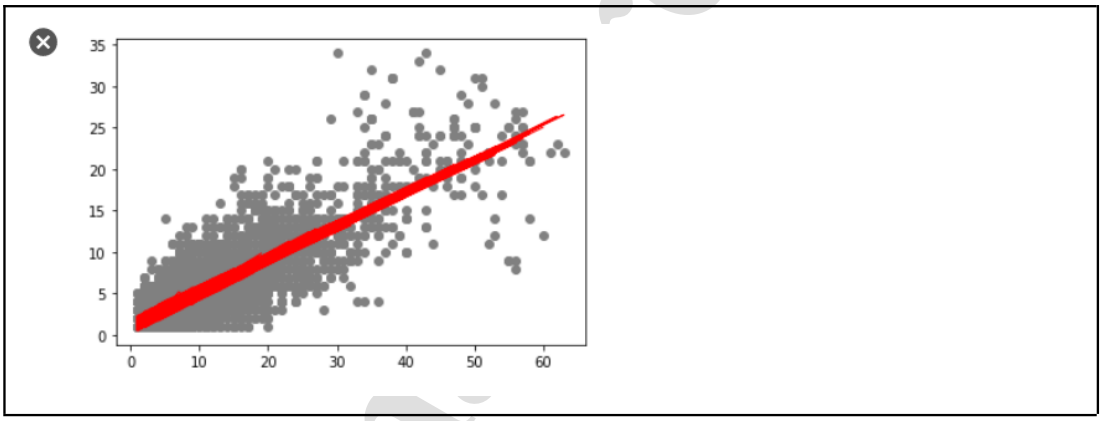


Figure 46: The regression interceptions exhibit a smooth and well-fitted alignment within the model. This fitting indicates a coherent relationship between the predicted values and the observed data, showcasing the model's accuracy in capturing the underlying patterns within the dataset.

After this we also tried to look if the results could be any better if we normalize the features using the StandardScaler method from SkLearn library. These are the results we received.

Mean Absolute Error	1.821
Mean Squared Error	6.835
Root Mean Squared Error	2.614
R2 Score	70.1%

Max Error	21.146

Table 4: The Linear Regression model's performance metrics with normalized data closely resemble those without normalization. This similarity in metrics suggests that normalization hasn't significantly altered the model's predictive ability. The model maintains a consistent level of performance even with the application of normalization to the dataset.

B. Ridge Regression

Ridge regression is a variation of linear regression that focuses on optimizing the loss function to minimize complexity. To make this adjustment, a penalty parameter that equals the square of the coefficients' magnitude is introduced here for the Ridge Regression. Below are the results we got from Ridge regression.

Mean Absolute Error	1.817
Mean Squared Error	6.811
Root Mean Squared Error	2.609
R2 Score	70.2%
Max Error	20.971

Table 5: The metrics for Ridge Regression closely resemble those of Linear Regression. This similarity in performance metrics indicates a comparable predictive ability between the two models. Both Ridge Regression and Linear Regression showcase analogous levels of performance across the evaluation metrics.

C. Lasso Regression

Lasso regression is also another extension of Linear Regression where mainly the loss function is tuned for reducing complexities. By restricting the sum of the absolute values of the model coefficients, Lasso modifies the loss function to reduce the model's complexity. The results are below.

Mean Absolute Error	1.816
Mean Squared Error	6.813
Root Mean Squared Error	2.610
R2 Score	70.2%
Max Error	20.959

Table 6: The metrics for Lasso Regression showcase a resemblance to those of both Linear Regression and Ridge Regression. Similar performance metrics suggest a comparable predictive capability among the three models. All three—Linear Regression, Ridge Regression, and Lasso Regression—demonstrate analogous levels of performance across the evaluation metrics used.

It's noteworthy that for various linear regression techniques, the penalty parameter, often referred to as alpha, is carefully tuned to attain optimal results. Tuning this parameter is crucial as it significantly impacts the performance and predictive accuracy of different linear regression models. Achieving the best results often involves fine-tuning this parameter to strike a balance between model complexity and performance.

D. XGBoost Regression

Utilizing the XGBoost Regression as our subsequent model, we initiated by normalizing the dataset to suit the model's requirements. For evaluation, we employed the Root Mean Squared Error (RMSE) as the metric to assess the model's performance. Below is the convergence graph depicting the model's progress during training.

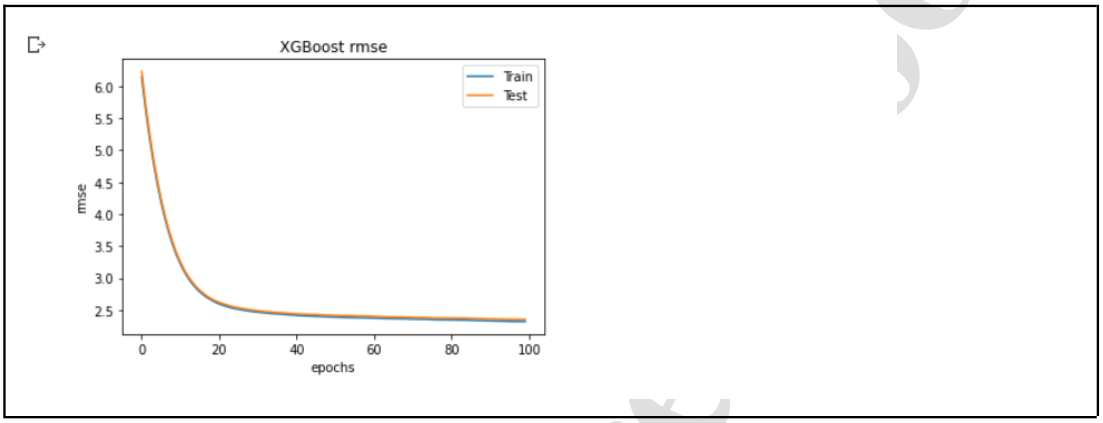


Figure 47: The convergence graph for XGBoost's RMSE demonstrates a gradual convergence towards a plateau. This convergence trend suggests a stabilizing pattern in the model's performance, indicating that further iterations might not significantly improve the model's predictive ability.

We also looked into other default measurements of this regression model.

Mean Absolute Error	1.706
Mean Squared Error	5.517
Root Mean Squared Error	2.348
R2 Score	75.8%
Max Error	16.717

Table 7: The XGBoost Regression model exhibited a notable performance enhancement compared to other linear regression methods. Specifically, our XGBoost Regressor showcased superior performance, outperforming the default Linear Regression model by achieving an impressive R2 score of approximately 76%. This substantial improvement signifies the superior predictive capability and effectiveness of the XGBoost Regression model in this scenario.

V. RESEARCH APPLICATION

The outcomes derived from our model represent a crucial aspect in implementing our research to address real-world challenges. To reiterate our research goal, it revolves around the analysis and prediction of crime activities within the broader London region. Despite the unavailability of a comprehensive dataset covering the entire London geography, we successfully acquired datasets from the Camden area, enabling us to construct a tailored dataset capable of predicting crime occurrences in specific geolocations over defined periods. This strategy holds applicability for two distinct entities in implementation.

1. **Application Users**

   In **Section III.A**, we laid out the foundation of the application to collect user's data. We also introduced a critical feature for the app that lets a user find out whether a place is safe to visit or not. If we look into the figure 39, we would find out they would have to tap the button that would take the GPS location of the user, convert it into the nearest street id we collected from Camden and then could reply back to the safety of the street based on the total crime that would happen on the next few weeks. Although this is in week level granularity, the system could be modified further which is discussed in **Section VI** to get a real time analysis of the street condition. After getting the GPS location of the user, basically the next steps are all almost the same as the one we discussed for the Government Organization section above.

2. **Government organization**

   Government organizations like the London Metropolitan Police can directly utilize this research to forecast crime activities in specific geolocations, focusing on the Camden area. For instance, to predict events in the WC2H-8AH postcode in week 25 of 2022, they would input the postcode and week number into the system. The system would then derive unchanging features—such as bus and tube station counts, market numbers, and similar data—from historical records. These would be integrated into the input features, forming a structured input for analysis. A simple test data could look like this:

month_number	7
week_number	25
total_crime_count_before_one_month	130
number_of_market_kiosk_within_300M	20
number_of_tube_point_within_300M	3
number_of_bus_stop_within_300M	5
total_stop_and_search_in_past_two_weeks	200
total_unsafe_report_in_past_two_weeks	250

Table 8: With new, unseen data, we aim to leverage these input features to forecast the crime counts in the near future. This fresh set of data serves as the basis for predicting forthcoming crime occurrences using our established model.

VI. LIMITATIONS AND FUTURE WORK

The system we've proposed and developed primarily centers on analyzing crime patterns. To enable real-time analysis, our aim is to condense the time frame involved. At present, we compute past crime occurrences and stop-and-search activities for individual dates, forecasting events for the upcoming two weeks. To elevate the precision of real-time analysis, our goal is to narrow down the timeframe for calculating past crimes and enhance the accuracy of predicting events specifically for the current day, rather than a broader two-week projection. While this transition presents a challenge, it holds promise for substantially improved results.

A prospective area for future work involves refining the focus on crime category activities. Presently, we've considered past crime activities collectively, irrespective of their specific crime categories. We aggregated all the past month's crime activities to create a single feature. Moving forward, a refinement could involve associating each crime category as an individual outcome. For instance, instead of summarizing all crime activities for the past month, we could predict specific crime categories, such as total sexual assaults in the upcoming weeks, defining this prediction as "twoWeeksSexualAssaultCrimes." However, fine-tuning the model to predict these specific outcomes requires meticulous adjustments and enhancements.

In our upcoming work, our focus will predominantly revolve around the dataset collected from the mobile application. While our primary model is built on the prepared dataset, it might lack the depth needed to predict various scenarios accurately under different conditions. Furthermore, we haven't incorporated the data collected from the application into our foundational model. Therefore, upon reviewing the collected data:

1. User’s age
2. Gender
3. Profession
4. Commute frequently etc.
5. Safety/Unsafe message from the app.

In our current approach, we've solely utilized a single feature—an aggregation of total unsafe messages received from the user's application. Integrating additional features without context into our primary dataset lacks meaningful correlation. Therefore, future work may involve the creation of an independent dataset, with each new feature appended in an aggregated form to the base dataset. Below, we illustrate a potential scenario:

Indeed, certain locations might pose greater safety concerns for women, especially during nighttime, compared to men. In scenarios where users, particularly women, inquire about safety levels through the app, the gender of the user becomes a crucial factor in making accurate predictions. Considering the gender of the user is essential for ensuring a more tailored and precise prediction, especially in contexts where safety concerns vary based on gender-specific experiences in certain areas.

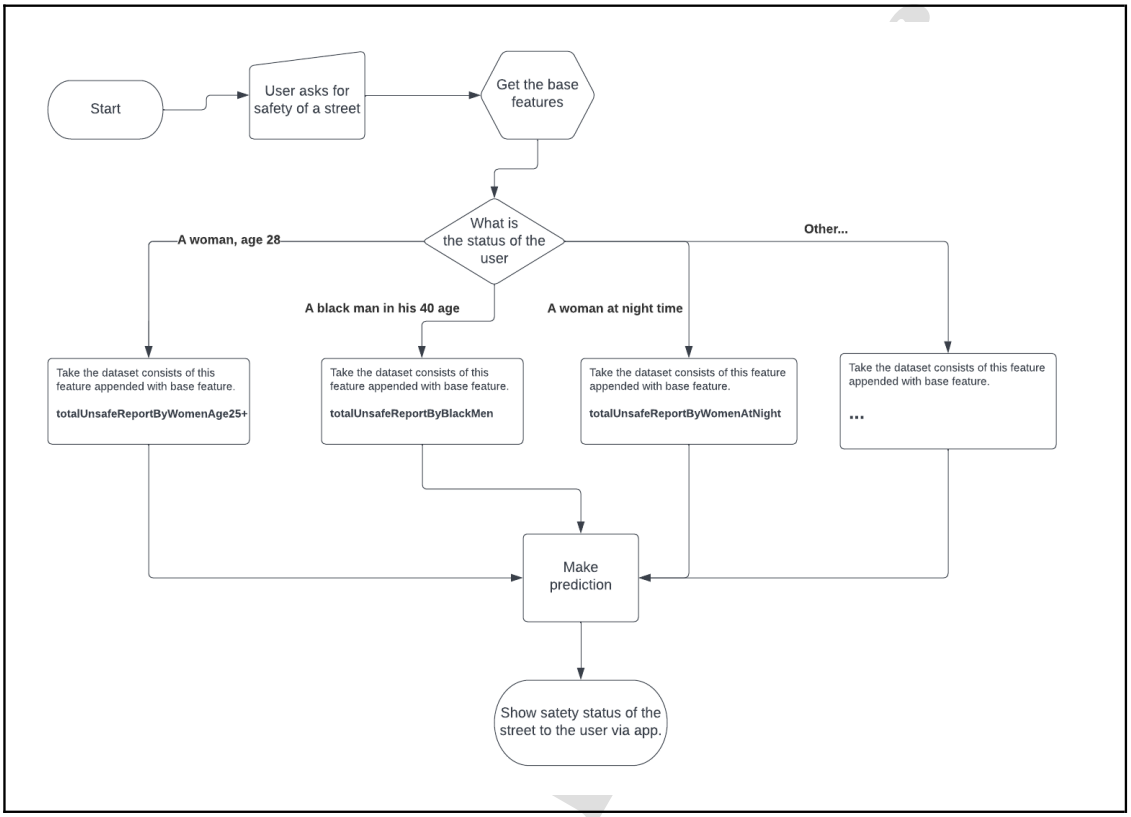


Figure 48: Our vision involves the app delivering tailored safety messages to diverse user personas. When a pedestrian inquires about a location's safety, the system retrieves recent crime statistics, station records etc. as base features, and uses the previously provided user persona. Then, it selects the appropriate model to predict the safety of that street. This task necessitates a substantial volume of data points to ensure accurate safety assessments for locations.

VII. CONCLUSION

Our primary aim is to analyze crime activities and facilitate real-time predictions for any location using available open-source data. We specifically chose the London Camden area due to its diverse range of accessible datasets, including Metropolitan police data and Stop and Search data. This work demonstrates the potential of leveraging these open-source datasets to construct a comprehensive dataset. This dataset can be employed in various regression models to forecast future crime activities effectively.

This methodology holds applicability to any location equipped with similar open-source features. However, the dataset used in this study had certain limitations, lacking actual timestamps and victim-specific data points. To address these limitations, our research introduced an innovative approach: the development of an application. This application aims to enhance the dataset by offering additional data points crucial for analyzing crime patterns. It includes accurate timestamps, victim perspectives, personas, and other pertinent details to enrich and augment the dataset for more comprehensive analysis.

Our research illustrated that with a dataset formulated around defined points of interest, a straightforward linear regression model is suitable for this scenario. Moving forward, our focus lies in data collection through the application and expanding predictions to accommodate diverse conditions.

The system has been thoroughly developed and is ready for deployment within the Camden area. Its primary goal is to provide users with crime safety awareness and equip administrators with actionable insights. By enabling proactive measures, the system aims to enhance overall safety and effectively tackle security issues. Additionally, a partial release of the resource materials related to this research has been made available **[40]** to support further advancements in this research field.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used chatGPT as a Generative-AI assisted tool in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

VIII. REFERENCES

[1] Crime, https://en.wikipedia.org/wiki/Crime [Accessed November 16, 2023].

[2] Elvia, Street Crime: Definition, Types, Examples & Powerful Deterrent. Reolink. https://reolink.com/blog/street-crime-how-security-cameras-prevent-it/ [Accessed August 31, 2022].

[3] Crime rate per 1,000 population in London from 2015/16 to 2022/23. Statista. https://www.statista.com/statistics/380963/london-crime-rate/ [Accessed August 31, 2022].

[4] Williams, Matthew & Burnap, Pete & Sloan, Luke. (2016). Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns. British Journal of Criminology. 57. azw031. 10.1093/bjc/azw031.

[5] M. Traunmueller, G. Quattrone, L. Capra, Mining mobile phone data to investigate urban crime theories at scale, International Conference on Social Informatics, 2014, pp. 396–411.

[6] Wang, X., Gerber, M.S., Brown, D.E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. In: Yang, S.J., Greenberg, A.M., Endsley, M. (eds) Social Computing, Behavioral - Cultural Modeling and Prediction. SBP 2012. Lecture Notes in Computer Science, vol 7227. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29047-3_28

[7] Sangeeta Lal, Lipika Tiwari, Ravi Ranjan, Ayushi Verma, Neetu Sardana, Rahul Mourya, Analysis and Classification of Crime Tweets, Procedia Computer Science, Volume 167, 2020, Pages 1911-1919, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.211.

[8] S. P. C. W. Sandagiri, B. T. G. S. Kumara and B. Kuhaneswaran, "ANN Based Crime Detection and Prediction using Twitter Posts and Weather Data," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain, 2020, pp. 1-5, doi: 10.1109/ICDABI51230.2020.9325660.

[9] Yihong Zhang, Panote Siriaraya, Yukiko Kawai, and Adam Jatowt. 2020. Analysis of street crime predictors in web open data. J. Intell. Inf. Syst. 55, 3 (Dec 2020), 535–559. https://doi.org/10.1007/s10844-019-00587-4

[10] X. Chen, Y. Cho and S. Y. Jang, "Crime prediction using Twitter sentiment and weather," 2015 Systems and Information Engineering Design Symposium, Charlottesville, VA, USA, 2015, pp. 63-68, doi: 10.1109/SIEDS.2015.7117012.

[11] Anderson, C. A., & Anderson, D. C. (1984). Ambient temperature and violent crime: Tests of the linear and curvilinear hypotheses. Journal of Personality and Social Psychology, 46(1), 91–97. https://doi.org/10.1037/0022-3514.46.1.91

[12] Rosser, G., Davies, T., Bowers, K.J. et al. Predictive Crime Mapping: Arbitrary Grids or Street Networks?. J Quant Criminol 33, 569–594 (2017). https://doi.org/10.1007/s10940-016-9321-x

[13] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. 2014. Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data. In Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14). Association for Computing Machinery, New York, NY, USA, 427–434. DOI:https://doi.org/10.1145/2663204.2663254

[14] Wen Dong, Bruno Lepri, and Alex (Sandy) Pentland. 2011. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (MUM '11). Association for Computing Machinery, New York, NY, USA, 134–143. https://doi.org/10.1145/2107596.2107613

[15] P. Cichosz, "Urban Crime Risk Prediction Using Point of Interest Data," ISPRS International Journal of Geo-Information, vol. 9, no. 7, p. 459, Jul. 2020, doi: 10.3390/ijgi9070459.

[16] Majeed, A. and Rauf, I., 2022. *MVC Architecture: A Detailed Insight to theModern Web Applications Development*. [online] Crimsonpublishers.com. Available at: <https://crimsonpublishers.com/prsp/fulltext/PRSP.000505.php> [Accessed 17 September 2022].

[17] X. Huang, "Research and Application of Node.js Core Technology," *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, 2020, pp. 1-4, doi: 10.1109/ICHCI51889.2020.00008.

[18] Robert C. Martin. "Principles Of OOD". butunclebob.com. Retrieved 2014-07-17.

[19] TfL Bus Stop Locations and Routes, https://data.london.gov.uk/dataset/tfl-bus-stop-locations-and-routes, [Accessed August 30, 2022].

[20] Detection and Prediction using Twitter Posts and Weather Data," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-5, doi: 10.1109/ICDABI51230.2020.9325660.

[21] Jobs and Job Density, Borough, https://data.london.gov.uk/dataset/jobs-and-job-density-borough, [Accessed August 29, 2022]

[22] HMO Licensing Register, https://opendata.camden.gov.uk/Housing/HMO-Licensing-Register/x43g-c2rf, [Accessed August 13, 2022]

[23] On Street Crime In Camden, https://opendata.camden.gov.uk/Crime-and-Criminal-Justice/On-Street-Crime-In-Camden/qeje-7ve7/data, [Accessed August 15, 2022]

[24] Schröer, Christoph & Kruse, Felix & Marx Gómez, Jorge. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science. 181. 526-534. 10.1016/j.procs.2021.01.199.

[25] Recorded Crime: Geographic Breakdown, https://data.london.gov.uk/dataset/recorded_crime_summary, [Accessed August 17, 2022]

[26] UK Police Crime Data Archive, https://data.police.uk/data/, [Accessed August 15, 2022]

[27] Sangeeta Lal, Lipika Tiwari, Ravi Ranjan, Ayushi Verma, Neetu Sardana, Rahul Mourya. Analysis and Classification of Crime Tweets, Procedia Computer Science, Volume 167,2020, Pages 1911-1919, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.211.

[28] Births by Borough, Ward, MSOA & LSOA, https://data.london.gov.uk/dataset/births-borough-ward-msoa-lsoa , [Accessed August 13, 2022]

[29] London School Atlas, https://data.london.gov.uk/dataset/london-schools-atlas, [Accessed August 14, 2022]

[30] Postcode Directory for London dataset, https://data.london.gov.uk/dataset/postcode-directory-for-london, [Accessed August 17, 2022]

[31] Camden Markets And Kiosks dataset, https://opendata.camden.gov.uk/Community/Camden-Markets-And-Kiosks/ikye-tidm/data, [Accessed August 18, 2022]

[32] Lower layer Super Output Area population estimates dataset, Mid-2019: SAPE22DT2 edition https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates, [Accessed August 14, 2022]

[33] Ward -LSOA Lookup, https://opendata.camden.gov.uk/People-Places/Ward-LSOA-Lookup-Maps/n4v5-ijg8, [Accessed August 14, 2022]

[34] Flutter, https://flutter.dev/, [Accessed November 21, 2023]

[35] NodeJS, https://www.w3schools.com/nodejs/nodejs_intro.asp, [Accessed November 21, 2023]

[36] Amazon DynamoDB, https://aws.amazon.com/dynamodb/, [Accessed November 21, 2023]

[37] SkLearn, https://scikit-learn.org/stable/, [Accessed November 24, 2023]

[38] Dorsogna, Maria & Perc, Matjaz. (2014). Statistical physics of crime: A review. Physics of Life Reviews. 12. 10.1016/j.plrev.2014.11.001.

[39] Perc M, Donnay K, Helbing D (2013) Understanding Recurrent Crime as System-Immanent Collective Behavior. PLoS ONE 8(10): e76063. https://doi.org/10.1371/journal.pone.0076063

[40] A. Yunus, “Crime-Analysis-Resource”. Zenodo, Nov. 27, 2023. doi: 10.5281/zenodo.10208462

Highlights (for review)

1. Open-source datasets like police data, station records etc. can serve crime analysis and prediction purposes
2. Crime prediction relies on timestamps and locations of previous crimes
3. Merging diverse open datasets on crime facilitates basic regression for predicting criminal activities
4. Street-level data collection relies on mobile apps where pedestrians can share their safety sentiments about streets

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: