UK & Essex Crime Analysis: Exploratory Data Analysis and Retrieval-Augmented Generation (RAG) Chatbot Implementation

By Christopher Final

Abstract

This capstone project explores the application of data science methodologies to crime analysis in England and Wales, using the county of Essex as a regional case study. With rising crime levels and declining public trust in policing, there is a critical need for localised, transparent, and data-driven solutions. This project addresses this need by integrating exploratory data analysis (EDA), time-series forecasting, machine learning, and natural language processing (NLP) to provide meaningful, accessible insights from publicly available crime data.

The core analytical framework includes the use of Seasonal ARIMA (SARIMA) models to forecast crime trends, Random Forest classifiers for anomaly detection, and clustering techniques (K-Means and DBSCAN) to identify geographic crime hotspots. These techniques revealed temporal surges in crime, regional disparities, and strong correlations with socioeconomic deprivation. Notably, the project highlights how deprived areas experience consistently higher crime rates, confirming insights from academic research on inequality and criminal behaviour.

To bridge the gap between complex analytics and public engagement, a Retrieval-Augmented Generation (RAG) chatbot was developed using LangChain and OpenAI embeddings. This tool allows users to query the data conversationally, increasing transparency and usability. An interactive dashboard, hosted on Render, complements the chatbot by visualising crime counts by Lower Super Output Areas (LSOAs). The dashboard is accessible by the URL https://crime-map-chatbot.onrender.com/ and GitHub repositories https://github.com/cfinal15.

Despite challenges related to API access, computing limitations, and district boundary effects, the project demonstrates the feasibility and value of integrating machine learning, GIS, and NLP in public sector applications. Ultimately, it offers a replicable model for data-driven regional policing that promotes both strategic decision-making and improved community trust and cohesion.

May 2025

Table of Contents

1 - Introduction

Overview of project

The rise in recorded crime incidents, alongside the erosion of public trust in policing, presents a formidable challenge for the government and police forces across the UK. This capstone project responds to this dual crisis by demonstrating how data science techniques, such as exploratory data analysis (EDA), machine learning (ML), and retrieval-augmented generation (RAG), can be used to transform crime data into actionable insights.

Using Essex as a regional case study, this project applies advanced analytics to publicly available datasets from data.police.uk. By employing SARIMA for time series forecasting, clustering algorithms like K-Means and DBSCAN for spatial analysis, and Random Forest models for anomaly detection, the project aims to surface both known and novel crime patterns. These include hotspot identification, seasonal crime surges, and outlier behaviours.

The project goes further than just technical discovery. By integrating results into a RAG-based chatbot built using LangChain and OpenAI embeddings. A Flask web application can deliver a conversational interface for public-facing insights. This dashboard provides interactive maps and a user-friendly portal for residents and policymakers alike, bridging the gap between advanced analytics and accessible, transparent communication.

The ultimate goal is to demonstrate how data-driven crime analysis can inform smarter policing, promote resource efficiency, and foster greater public trust.

Background and context

In the year ending September 2024, England and Wales recorded 9.5 million headline crimes, a 12% increase from the previous year (ONS, 2024). While long-term statistics show a downward trend over decades, this recent uptick reveals underlying vulnerabilities. Communities are increasingly exposed to crimes such as violence, antisocial behaviour, and theft, while systemic failures, notably in tackling domestic abuse and knife crime, have made headlines.

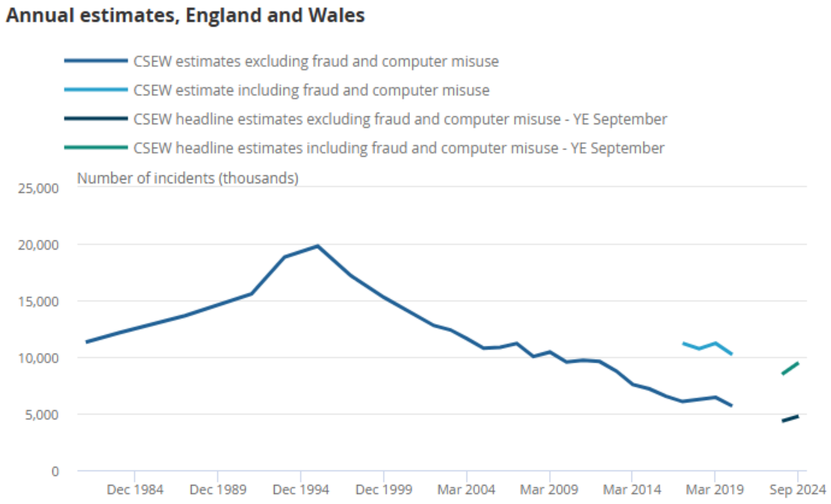


Figure 1A – Annual Crime Estimates for England and Wales (ONS, 2024)

Compounding this problem is a significant decline in public confidence in UK policing. According to YouGov (2024), 52% of UK adults now express "little or no confidence" in the police to effectively manage local crime, up from 39% in 2019. This sentiment has been inflamed by high-profile incidents, including the murder of Sarah Everard by a serving police officer, which further fuelled debates around accountability and reform (End Violence Against Women, 2024).

Economically, crime imposes a significant burden. The Home Office estimates the total cost of crime at over £59 billion annually across individuals and businesses (Heeks et al., 2018). Such pressures demand scalable, localised, and transparent solutions.

Adding urgency is the UK government's National AI Strategy (HM Government 2021), which outlines a roadmap to become a global AI leader by 2030. One of its pillars is "AI for public good", highlighting AI's potential in policing, healthcare, and public services. This includes not just predictive analytics, but also tools that promote public trust through transparency and explainability.

These trends make it clear that there is a growing national and institutional appetite to leverage AI and machine learning in crime analysis. However, caution is needed with rigorous ethical safeguards and community engagement.

Problem Statement

Tackling over 9 million crimes annually across 44 police forces in England and Wales presents a logistical and analytical challenge. Broad-scale strategies often fail to address localised crime patterns, leaving communities vulnerable to persistent safety issues. A localised approach to tackling crime is essential due to the overwhelming scale of the issue, with individual police forces facing unique challenges.

Essex provides an ideal testbed for localised predictive policing. As one of the largest non-metropolitan counties in England, it encompasses both urban centres (e.g., Southend, Basildon) and rural areas (e.g., Uttlesford, Maldon), offering a diverse landscape for spatial and socioeconomic analysis.

In 2024, Essex Police recorded 158,403 crimes. Violence and sexual offences accounted for nearly 40% of total cases between 2022 and 2024. Crime here is seasonal, with spikes in March, July, and December. A pattern often linked to public holidays, weather, and school breaks.

Moreover, Essex is a known innovator in crime prevention. The Knife Crime and Violence Model (KCVM) is a machine learning algorithm jointly developed by Essex Police and the University of Essex. It demonstrated the power of AI to identify high-risk individuals and intervene early. This spirit of innovation, combined with data availability, makes Essex a prime candidate for a multi-method case study.

The region also faces socio-economic challenges. According to the Essex County Council (2024), over 188,000 residents live in the most deprived quintile of the UK, concentrated in coastal towns such as Clacton and Harwich. These areas also show higher crime concentrations, making Essex a rich context for evaluating the relationship between deprivation, trust, and public safety.

However, working with crime data poses several key challenges:

- High data complexity - large volumes of semi-structured CSV data with inconsistencies, missing fields, and inconsistent geocoding.

- Hotspot identification - Accurately identifying dynamic or seasonal crime clusters is difficult using static models.
- Trust and transparency - Data use must be both effective and ethically sound to avoid reinforcing bias or mistrust.
- Limited accessibility of insights - Raw data is often inaccessible or unintuitive to the public.

This project proposes an integrated solution combining spatial-temporal crime forecasting, clustering, and anomaly detection, wrapped into a user-accessible RAG-based chatbot and visual dashboard, specifically for the Essex region.

Objectives

This capstone project aims to apply modern data science techniques to crime data in Essex to deliver both predictive insights and accessible tools for residents. The key objectives are:

- **Collect and analyse crime data from Police UK archives & API -** The UK Police open data archives provide a comprehensive repository of crime and policing data across England and Wales, shared with the public since 2011 to enhance transparency and provide valuable insights. The platform offers detailed information on street-level crime, outcomes, and stop-and-search activities, available as semi-structured CSV files or through an API. The data includes attributes such as location, jurisdiction, crime type, and outcome, enabling granular analysis and supporting targeted crime prevention strategies.

- **Perform exploratory data analysis (EDA) and visualise crime trends -** Performing exploratory data analysis (EDA) and visualising crime trends involves several key steps. EDA ensures that the data is clean, consistent, and well-structured, which is crucial for accurate analysis and reliable machine learning models. It includes calculating descriptive statistics to summarise the data and identify anomalies, creating visual representations like bar charts and scatter plots to spot trends and patterns, and using interactive maps for geospatial analysis to identify crime hotspots. Trend analysis helps understand the dynamics of crime over time, while correlation analysis reveals potential causes and contributing factors. By performing EDA and visualising crime trends, we gain a comprehensive understanding of the current crime landscape in Essex, which is foundational for developing targeted strategies and implementing advanced machine learning models to predict and prevent crime.

- **Build machine learning models for crime prediction and anomaly detection -** To understand and anticipate crime patterns more effectively, this project builds a series of machine learning models tailored for both prediction and anomaly detection. Supervised learning techniques, such as Random Forest and Linear Regression, are employed to predict crime counts based on features including time of year, crime type, and geographic area. These models help identify key predictors of crime and enable future-oriented resource planning. In parallel, unsupervised techniques like Isolation Forest are implemented to detect anomalies, such as localised increases in violence. Together, these models support a proactive approach to crime analysis by offering insights into both expected trends and unexpected deviations, helping stakeholders respond more efficiently to evolving public safety concerns.

- **Integrate data insights into a Retrieval-Augmented Generation (RAG) chatbot -** To improve the accessibility and interpretability of crime data, the project integrates analytical outputs into a Retrieval-Augmented Generation (RAG) chatbot, enabling users to interact with structured insights using natural language. Using the LangChain framework, the chatbot combines a semantic retriever, which searches embedded documents for relevant context, with a language model that generates fluent and accurate responses. Crime statistics, trends, and hotspot summaries are transformed into text-based documents and indexed using vector stores like FAISS. When a user poses a question, the system retrieves relevant content and uses the LLM to generate a context-aware, conversational answer grounded in real data. This integration not only enhances public engagement with crime analytics but also supports transparent, on-demand access to insights that might otherwise be locked behind technical interfaces or static dashboards.

Scope of project

This project focuses on the analysis of crime data in Essex between January 2022 and December 2024, using a combination of geospatial, temporal, and machine learning techniques. The analytical scope centres on identifying crime trends and hotspots through clustering algorithms (K-Means and DBSCAN), forecasting crime patterns with SARIMA models, and detecting anomalous behaviour using Random Forest and Isolation Forest classifiers. Beyond analysis, the project investigates the correlation between crime rates and socioeconomic deprivation using Index of Multiple Deprivation (IMD) scores. To make the insights accessible and actionable, the project delivers a web-based dashboard incorporating interactive crime maps and a Retrieval-Augmented Generation (RAG) chatbot. Built with Python-based tools including Pandas, Scikit-learn, LangChain, and Folium, and deployed via Flask and Render, the solution aims to improve public transparency and decision-making by enabling real-time interaction with processed crime data. This scope aligns with broader national objectives around digital innovation in policing and public sector transparency.

2 - Literature Review

Introduction to Crime Data Analytics

Crime analysis has evolved into a cornerstone of modern policing, driven by the rapid advancements in big data technologies and the ongoing Fourth Industrial Revolution, which has significantly transformed how data is captured, processed, and utilised across sectors (Petrillo et al 2018). The integration of large-scale data collection, machine learning, and real-time analytics has empowered policing bodies to detect patterns, forecast criminal activity, and allocate resources more strategically than ever before. With the growing availability of public datasets, such as those provided by the data.police.uk, crime data can now be analysed at local, regional, and national levels, offering deep insights into criminal trends. As data becomes increasingly granular and accessible, the potential for evidence-based crime prevention, policymaking, and community engagement continues to expand.

Through data-driven crime mapping, trend forecasting, and statistical modelling, researchers and policymakers gain critical tools to inform decisions, allocate funding, and improve community safety. Crime analytics also empowers residents, offering visibility into crime activity and policing responses in their communities, thereby supporting the push for greater transparency and accountability within their communities.

This section reviews the evolution of crime data analysis techniques. From traditional time-series forecasting and regression modelling to clustering, spatial analysis, and the more recent integration of natural language models such as Retrieval-Augmented Generation (RAG). Each of these approaches plays a distinct role in understanding the causes, distribution, and prediction of crime. By grounding the project in these established methods, we ensure its academic and practical validity.

Crime Prediction and Machine Learning Models

One of the common statistical approaches to crime analysis is time series techniques. Autoregressive Integrated Moving Average (ARIMA) and its seasonal extension, Seasonal ARIMA (SARIMA), are foundational statistical models widely used in time series forecasting. These models are particularly effective when working with structured or semi-structured data, such as monthly or daily crime counts, due to their ability to capture linear trends, autocorrelations, and seasonal fluctuations. ARIMA models rely on historical observations and lagged error terms to predict future values, making them well-suited for datasets that exhibit stationarity after differencing (Box et al 2015). The SARIMA model enhances this capability by incorporating seasonal components into the ARIMA framework. This is especially relevant in crime analytics, where temporal patterns often repeat annually due to social and environmental factors (Chatfield, 2003). Furthermore, these models are computationally efficient and require relatively less preprocessing compared to more complex machine learning models, making them a practical choice for structured and semi-structured datasets commonly retrieved from open data portals like Police UK. Their interpretability and diagnostic transparency also support easy integration into decision-making workflows within public safety contexts.

Supervised machine learning (ML) provides an opportunity to detect hidden patterns in historical data and predict future criminal activity with greater accuracy than traditional methods. In contrast to statistical models like ARIMA, which rely heavily on time-series structures, supervised ML methods can incorporate a wider variety of features such as location, crime type, time of day, demographic data, and socio-economic indicators. Which allow for multidimensional forecasting and classification.

Commonly used algorithms in crime forecasting include linear regression, decision trees, random forests, support vector machines (SVM), and neural networks. These models are trained on labelled datasets, where both input variables (e.g., time, location, deprivation score) and target outputs (e.g., crime count, category) are known, to learn complex relationships that may be non-linear or interaction heavy. For instance, regression models have been used to forecast crime volumes in specific regions by analysing past crime trends alongside auxiliary data such as unemployment rates and population density (Wang et al., 2013). Classification models, on the other hand, are particularly useful in predicting the type of crime likely to occur in a given area or time frame, based on contextual features (e.g, a burglary happening at 3 PM in a wealthy neighbourhood might have different causes than one at midnight in a deprived area). Contextual features such as time of day, district location, and deprivation scores will help machine learning models to understand the situational factors contributing to crime patterns

One of the key strengths of supervised Machine Learning in this domain is its ability to handle noisy or high-dimensional data, making it suitable for working with semi-structured datasets such as those derived from the Police UK archives. Random Forests algorithms have demonstrated strong performance in crime classification tasks due to their robustness against overfitting and their ability to rank feature importance, which is a critical aspect when explaining model decisions to policymakers (Bowers and Johnson, 2014). Additionally, the application of ensemble methods and hyperparameter tuning has allowed for improved accuracy in crime forecasting deployments.

Despite their predictive power, Machine Learning models in crime analytics are not without limitations. Issues such as data imbalance, bias in historical crime reporting, and over-reliance on past patterns can introduce ethical concerns, particularly when models inadvertently reinforce over-policing in marginalised communities (Lum and Isaac, 2016). To mitigate these risks, recent studies voice the importance of incorporating fairness-aware Machine Learning techniques and conducting regular audits of model performance across different population subgroups.

Overall, supervised machine learning offers a powerful framework for crime forecasting, particularly when paired with open datasets and geographic granularity. When implemented responsibly, these models can support evidence-based decision-making, enhance operational policing, and improve public trust through transparency and accountability. Unsupervised machine learning should also be explored. Unsupervised techniques such as clustering offer critical insights into the spatial distribution of crime. Clustering algorithms like K-Means, DBSCAN, and HDBSCAN are commonly used to identify crime hotspots by grouping together geographic areas that exhibit similar patterns of criminal activity. These methods are especially effective when working with geocoded crime data, enabling analysts to detect emergent clusters across different districts and regions. Unlike supervised models, clustering does not rely on labelled outputs, making it ideal for exploratory analysis in unknown or dynamic environments. Tools such as CrimeStat have long demonstrated the power of spatial statistics in hotspot detection and tactical deployment of police resources (Levine, 2015). When applied to Essex crime data, such techniques can support resource allocation, police patrol planning, and community safety interventions by revealing where crimes are most likely to recur.

Spatial and Environmental Factors in Crime

Understanding the spatial and environmental dimensions of crime is central to both criminology and data-driven policing. Criminal activity is not distributed randomly but is influenced by geographic, social, and economic condition. Theories such as routine activity theory and crime pattern theory suggest that the convergence of motivated offenders, suitable targets, and a lack of capable guardianship contributes to the formation of crime hotspots (Brantingham and Brantingham, 1995). These hotspots often correlate with socio-environmental factors such as poor lighting, high population density, and low levels of social cohesion.

Recent research has reinforced the idea that socioeconomic inequality plays a significant role in shaping crime distribution. For instance, Courson and Nettle (2021) modelled how individuals in economically deprived areas may be more prone to engage in criminal behaviour due to both desperation and the potential for variable high rewards. Their findings indicate that persistent inequality not only exacerbates crime rates but also reduces social trust, leading to a feedback loop of disadvantage and insecurity.

The spatial resolution of modern datasets, like Lower Super Output Areas (LSOAs) in the UK, has enabled the opportunity for more granular analysis of environmental risk factors. Studies have shown that districts with higher levels of unemployment, lower educational attainment, and poor housing conditions consistently report elevated crime rates (Kurland and Piquero, 2022). By incorporating these environmental features into crime models, analysts can better predict where criminal activity is likely to occur and design more targeted intervention strategies.

Moreover, geographic information systems (GIS) and spatial data science tools have made it possible to visualise and quantify these patterns. When combined with machine learning, spatial attributes can significantly improve the performance of crime forecasting models. These contextual and environmental layers are particularly important for understanding not just where crime happens, but why it occurs in specific places, allowing for place-based policing strategies that are both proactive and equitable.

Public Trust and Ethical Implications of Predictive Policing

The adoption of predictive policing technologies, while promising in terms of efficiency and resource optimisation, raises profound ethical concerns and questions about public trust in police bodies. Trust in UK policing has been eroded over the past decade, particularly following high-profile incidents involving misconduct, institutional bias, and failures in accountability. According to a 2024 YouGov survey, over half of UK adults (52%) reported having "little or no trust" in the police to effectively address crime, a figure significantly higher than 39% in 2019. High-profile cases such as the Sarah Everard tragedy, where a serving police officer was convicted of murder, have intensified scrutiny around the accountability, transparency, and ethical standards of policing institutions (End Violence Against Women, 2024).

Predictive policing and the use of algorithms to forecast where crimes are likely to occur or who may be involved has the potential to be a technological solution to declining public confidence. However, critics argue that such systems may reinforce existing biases and perpetuate inequality if built on flawed or unbalanced data. Lum and Isaac (2016) demonstrated how predictive models trained on biased datasets can lead to over-policing in marginalised communities, creating a feedback loop that exacerbates distrust and fuels systemic discrimination.

To mitigate these risks, researchers, policymakers and institutions, should advocate for the use of fairness-aware machine learning techniques and transparent algorithmic governance. This includes bias detection in input data, performance evaluation across demographic subgroups, and public accountability mechanisms such as meaningful community engagement. In the UK, where police data is open-access and regulated by data protection frameworks like the GDPR, there is also a legal obligation to ensure that algorithmic decisions remain explainable and challengeable.

Ultimately, while predictive policing holds potential for more strategic and effective crime prevention, its success hinges not only on technical accuracy but also on ethical integrity and social legitimacy. Embedding transparency, fairness, and public dialogue into the development and deployment of these technologies is essential to restoring and sustaining trust in modern policing.

International Approaches to Data-Driven Policing

Globally, policing agencies are increasingly adopting data-driven technologies to enhance crime prediction, improve resource allocation, and support operational planning. These approaches vary in scale, transparency, and public impact. One of the earliest and most widely known predictive policing tools was PredPol, developed in the United States. Designed to forecast crime hotspots based on historical data, the tool was implemented across several U.S. cities. However, it was later discontinued amid criticism over racial bias and lack of transparency, highlighting the potential for algorithmic reinforcement of existing inequalities (Lum & Isaac, 2016). Despite its technical capabilities, PredPol's downfall highlighted the importance of ethical safeguards, community consultation, and explainability.

On the other hand, CompStat, a data-driven accountability system introduced by the New York Police Department, has been more enduring. While not predictive in nature, it uses weekly crime data to track trends and hold precincts accountable for outcomes (Weisburd et al 2003). CompStat's success has influenced similar systems in Chicago, Los Angeles, and London, where geographic clustering and real-time dashboards now support hotspot policing strategies.

Despite these advancements, concerns persist about data bias and the role of "black-box" algorithms in public safety, where inputs go in and outputs come out, but the process in between is opaque and unclear to understand. A common limitation in many global approaches is the lack of public-facing transparency. The majority of systems are developed for internal use, offering little visibility or interpretability for the general population.

This project distinguishes itself by not only applying predictive and spatial techniques but also embedding retrieval-augmented generation (RAG) and open data into a publicly accessible dashboard and chatbot. This enables Essex residents, local authorities, and community stakeholders to interact with insights in an intuitive and transparent manner. In doing so, it bridges the gap between technical capability and public accountability. An area where global models have often fallen short.

Natural Language Processing and RAG in the Public Sector

The rise of natural language processing (NLP) has transformed how information is accessed, processed, and communicated within public services. NLP technologies enable systems to understand, interpret, and generate human language, offering tools that can bridge the communication gap between the public and complex data systems. In the context of crime analytics, NLP can be used to summarise reports, classify incidents, or, more recently, support conversational interfaces through chatbots and large language models (LLMs).

One of the most promising developments in this field is the emergence of Retrieval-Augmented Generation (RAG), a framework that combines document retrieval with language generation. Unlike standalone language models, which generate responses based on pre-trained internal knowledge, RAG architectures are designed to pull relevant information from an external knowledge base (e.g., police reports, crime data, or policy documents) and integrate it into generated responses. This ensures that outputs are not only linguistically coherent but also factually grounded in up-to-date, domain-specific content (Lewis et al., 2020).

For the public sector, RAG-powered chatbots are gaining traction for their ability to deliver evidence-based, context-aware responses to public queries. For example, residents can query about crime statistics in their area, seek guidance on police procedures, or ask for updates on safety trends, receiving responses that are both natural and tailored to their needs. This is particularly valuable in sectors like policing, where accessibility, transparency, and public confidence are central concerns. By enabling connections with crime datasets, a RAG chatbot serves as an interface that transforms static open data into actionable knowledge for communities.

The benefits of RAG are further amplified when paired with open-source tools such as LangChain, which supports dynamic integration of various data formats and retrieval sources. These systems allow developers to index structured datasets, like CSV files, or unstructured text, and build conversational agents that can reason over and respond with relevant context. However, deploying NLP-based systems for the public sector organisations also comes with challenges. These include ensuring data privacy, limiting misinformation, and building mechanisms to verify the factual consistency of generated responses.

Despite these challenges, RAG presents a valuable opportunity to enhance transparency, accessibility, and responsiveness in public services. In the context of this project, integrating RAG with local crime data offers a novel method to democratise access to complex datasets, empower residents with localised insights, and foster greater trust in data-driven public systems.

Filling the Gap: This Project's Role in the Literature

While the literature demonstrates significant progress in crime forecasting, spatial analysis, and the use of NLP in public services, many existing predictive policing systems rely solely on statistical models or supervised machine learning for forecasting, without combining them with real-time data exploration or natural language interfaces that improve public accessibility.

One notable gap is the lack of localised, explainable crime analysis tools that bridge predictive analytics with user-friendly interaction. While police forces may use internal dashboards, these are often not publicly available, nor do they leverage natural language processing to allow everyday users to explore crime trends conversationally. In addition, most crime analysis studies focus on national-level trends or aggregated regional data, limiting the granularity needed to inform district-level or LSOA-specific interventions.

Furthermore, the literature on Retrieval-Augmented Generation (RAG) in public services is still emerging, with few applications tailored to structured datasets like police reports or open government APIs. Most RAG implementations focus on unstructured documents (e.g., articles, legislation), but rarely on structured CSV-based data with time-series or geospatial dimensions.

This presents an opportunity to expand the capabilities of RAG systems into new domains where data context and numerical interpretation are essential.

This project addresses these gaps by offering a multi-layered solution that combines:

- Time-series forecasting (e.g., SARIMA) to capture seasonality in crime
- Clustering and regression models to understand spatial and socioeconomic patterns
- A LangChain-powered RAG chatbot that delivers real-time insights from structured Essex crime data
- An interactive dashboard that visualises key trends by location and crime type

By uniting predictive analytics, spatial modelling, and conversational NLP into a cohesive system, this project contributes a novel framework for accessible, explainable, and localised crime intelligence. It also demonstrates how data science can be applied ethically and transparently to rebuild public trust and enable informed decision-making at both the institutional and community levels.

3 - Methodology

Research Design

This project adopts a case study-based applied research design, focusing on the use of data science techniques to analyse and forecast crime trends within Essex from 2022 to 2024. The methodology integrates quantitative analysis, machine learning, and natural language processing (NLP) to create both predictive insights and a user-accessible crime information system.

The core of the research involves the exploration, modelling, and communication of structured and semi-structured crime data. A combination of exploratory data analysis (EDA), time-series forecasting (using SARIMA), clustering, and supervised learning models is used to uncover patterns and make predictions about future criminal activity. These techniques are particularly suited to understanding spatial and temporal trends in crime, as well as testing the relationship between socio-economic factors, like deprivation deciles, and crime occurrence.

In addition to the data modelling components, the project includes the development of a Retrieval-Augmented Generation (RAG) system using the LangChain framework. This NLP-based component enables users to interact with the Essex crime dataset in natural language, supported by a backend system that retrieves relevant data and generates interpretable, factual responses. This hybrid AI approach provides an innovative pathway for making crime data more accessible and actionable. To support transparency and stakeholder engagement, the project is also deployed as an interactive dashboard and chatbot, built using Python-based tools such as Pandas, Scikit-learn, Folium, Plotly, and Flask.

Data Collection from Police UK data files and API

The UK Police open data archives provide a comprehensive repository of crime and policing data across England and Wales. Since 2011, this information has been shared with the public to enhance transparency and provide valuable insights into crime and policing. The platform offers detailed information on street-level crime, outcomes, and stop-and-search activities, which can be downloaded as simple, semi-structured CSV files. Additionally, data can be accessed through an API that employs a leaky bucket methodology. This rate-limiting algorithm ensures that requests are processed at a constant rate, preventing server overload by maintaining a steady flow of information.

The data is broken down by police force and lower layer super output area (LSOA), making it highly detailed and useful for analysis. The UK Police API offers a rich data source, including information on neighbourhood team members, upcoming events, street-level crime and outcomes, and the nearest police stations. This combination of open data archives and an accessible API provides helpful resources for researchers, developers, and the public to explore and utilise detailed crime and policing information. The detailed breakdown of data by police force and LSOA allows for granular analysis, making it possible to identify specific areas with higher crime rates and understand the factors contributing to these patterns. This level of detail is invaluable for developing targeted crime prevention strategies and evaluating the impact of policing efforts.

The platform's ability to provide street-level crime data is particularly valuable for local communities and policymakers. By understanding where crimes are occurring, authorities can allocate resources more effectively and implement targeted interventions to reduce crime rates.

The outcome data helps in assessing the effectiveness of policing strategies and identifying areas that require improvement.

Data Features and Preprocessing Techniques

Monthly data from the police archives comes with 11 attributes or columns, including detailed information such as the location where the reported crime took place, the jurisdiction of the police force responsible for the area, the type of crime (e.g., theft, assault, vandalism), and the outcome of the crime report (e.g., investigation ongoing, case closed, suspect arrested). These attributes provide a comprehensive view of crime incidents, allowing for in-depth analysis and understanding of crime patterns.

A description for the 11 attributes is listed below -

Variable	**Description**
Crime ID	Unique 64-digit Crime reference
Month	Month of when the crime occurred in YYYY-MM format
Reported by	Which of the 44 police forces reported the crime
Falls within	The jurisdiction of where the crime took place, in one of the 44 police force areas
Longitude	Longitude coordinates of where the crime occurred
Latitude	Latitude coordinates of where the crime occurred
Location	Road location of where the crime took place or near
LSOA code	Lower Super Output Area (LSOA) code location of where the crime took place
LSOA name	Lower Super Output Area (LSOA) name of where the crime took place
Crime type	Type of crime described by the 14 different crime categories
Last outcome category	Outcome status by the police force to confirm if it's under investigation, or if the investigation has been completed whilst confirming if a suspect had been identified or convicted

Crime analysis relies heavily on high-quality data to identify patterns, predict trends, and support law enforcement decision-making. The UK police collect vast amounts of crime-related data that is usually cleaned, but just like any other raw dataset, they often require preprocessing to ensure accuracy and consistency. For example,

- **Standardising data formats** – raw datasets often contain variables recorded in inconsistent formats (e.g., dates as DD/MM/YYYY, YYYY-MM-DD, or unstructured text). To enable accurate temporal analysis—such as identifying trends by day, month, or year—these formats must be standardised. Python libraries like datetime and pandas can parse and convert dates into a uniform structure. This ensures compatibility with analytical tools and facilitates operations like time-series aggregation and seasonal crime pattern detection.
- **Handling missing values and outliers** – Before analysing the data, checking for missing values and outliers helps to ensure accurate insights. Missing data can be handled through deletion (if records are incomplete) or imputation (replacing missing values with statistical estimates such as the median or mode). Additionally, context-based replacements can be applied—for example, deriving a missing LSOA (Lower Layer Super Output Area) code from available geographical coordinates. Outliers must also be carefully examined, particularly to

ensure the dataset reflects only Essex-based crimes by filtering out incidents logged by Essex Police that occurred outside the county. Further checks should identify statistical anomalies, such as unusually high crime counts, which may indicate data entry errors or genuine spikes in criminal activity. By systematically managing missing data and outliers, the dataset becomes more reliable for crime trend analysis and predictive modelling.

- **Encoding categorial variables** - Our dataset contains categorical variables (e.g., crime types, locations, and outcomes) that need to be converted into numerical representations for machine learning analysis. Since most algorithms cannot interpret raw text labels directly, techniques like label encoding (assigning integer values) or one-hot encoding (creating binary columns per category) are applied. The choice depends on the variable's nature—label encoding suits ordinal data (e.g., severity levels), while one-hot encoding avoids false ordinal relationships in nominal data (e.g., borough names). Encoding will ensure the models we use, can effectively identify patterns in categorical crime features.
- **Normalisation and standardisation scaling** - To prevent variables with larger ranges from dominating machine learning models, techniques like min-max scaling (resizing values to a 0–1 range) or standardisation (transforming data to have mean=0 and variance=1) can looked to be applied. For example, normalizing socioeconomic indices and crime frequencies ensures equal weighting in predictive models. Robust scaling (using median/IQR) may also be preferred for datasets with outliers, such as rare high-crime spikes. Proper scaling improves model convergence in algorithms like k-nearest neighbours or neural networks while maintaining interpretability for crime trend analysis.

Modelling and Analysis Techniques

To extract actionable insights and forecast crime trends across Essex, the project implements a range of modelling and analysis techniques, such as time-series forecasting, machine learning, clustering, and anomaly detection. These models are selected to explore the spatial-temporal nature of crime data, support localised predictions, and contribute to the development of a retrieval-augmented question-answering system.

- **Time-Series Forecasting (SARIMA) -** For forecasting crime trends over time, particularly seasonal patterns in categories such as shoplifting and public disorder, the project employs the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. Prior to model selection, the time series is evaluated using the Augmented Dickey-Fuller (ADF) test to assess stationarity. Based on these findings, SARIMA parameters (p,d,q)(P,D,Q,s) are tuned using grid search and AIC minimisation. The model is trained on 2022–2023 data and validated on 2024 values using RMSE and MAPE as performance metrics. This approach captures both trend and seasonal components, providing interpretable and reliable forecasts.
- **Supervised Machine Learning -** To explore relationships between crime and social factors (e.g., Index of Multiple Deprivation), the project applies regression and classification models, including Linear Regression, Decision Trees, and Random Forests. Feature inputs include crime category, LSOA location, time variables (month, year), and deprivation scores. For classification tasks, such as predicting the most likely crime type in a given district, models are evaluated using accuracy, F1-score, and confusion matrices, with cross-validation applied to ensure generalisability. Feature importance outputs from Random Forests are used to identify key predictors in Essex's crime landscape.

- **Clustering for Hotspot Identification -** To detect spatial concentrations of crime, the project uses unsupervised learning algorithms such as K-Means and DBSCAN. Geospatial features like longitude, latitude, and LSOA groupings are used to identify crime hotspots. K-Means is evaluated using the elbow method and silhouette scores, while DBSCAN is tested for its ability to capture irregular density-based clusters, which may indicate emerging hotspots. Outputs are visualised using Folium maps to provide stakeholders with interpretable, map-based views of high-risk zones.
- **Anomaly Detection -** The project also includes anomaly detection to highlight unusual spikes in crime, which might represent rare events or data inconsistencies. The Isolation Forest algorithm is used to detect outliers based on multidimensional inputs (e.g., date, location, category). Detected anomalies are manually reviewed and visualised in the dashboard to support proactive investigations or data cleaning efforts.

Each model is developed, tuned, and evaluated using Python libraries such as Scikit-learn and Statsmodels. Together, these techniques provide a comprehensive analysis pipeline that supports accurate forecasting, hotspot detection, and the creation of interpretable insights suitable for both public engagement and internal policing strategy.

NLP and RAG System Development

To make crime data more accessible and understandable to a non-technical audience, this project incorporates a Retrieval-Augmented Generation (RAG) system powered by natural language processing (NLP). The RAG framework enables a conversational interface through which users can query the Essex crime dataset and receive contextualised, data-backed responses in natural language. This component complements the analytical side of the project by enhancing user engagement, interpretability, and transparency.

The RAG system is built using the LangChain framework, which facilitates the combination of document retrieval and generative language models (LLMs). The first step in this pipeline involves preprocessing the crime dataset into retrievable documents. The structured data files containing monthly crime statistics or LSOA summaries, will be transformed into text-based chunks that retain key information such as location, crime type, trends, and statistical summaries. These chunks are then embedded using sentence-transformer models and stored in a vector database, like FAISS or Chroma, to enable fast semantic search.

When a user inputs a natural language query input like, “In 2024, what is the most common crime in Chelmsford?”, the system performs query understanding using basic NLP techniques like named entity recognition (NER), keyword extraction, and embedding generation. The retriever module uses these embeddings to search the vector store and return the top-k most relevant chunks of information. These results are passed to the generator module, which uses a language model to formulate a coherent and grounded response that integrates the retrieved content.

This architecture allows the chatbot to deliver responses that are both factually grounded in the data and conversationally fluent. It ensures that the system remains updated with real crime data rather than relying solely on pre-trained knowledge. To improve usability, the RAG component is embedded within a Flask-based web interface that mimics a chat experience. Users can ask questions, receive visual summaries, or be redirected to relevant dashboard views, depending on the query type.

By combining retrieval and generation, the RAG chatbot not only simplifies access to structured crime data but also offers a scalable and ethical method for engaging with public datasets to support transparency and community awareness.

4 - Data Presentation and Analysis

Introduction to Analysis

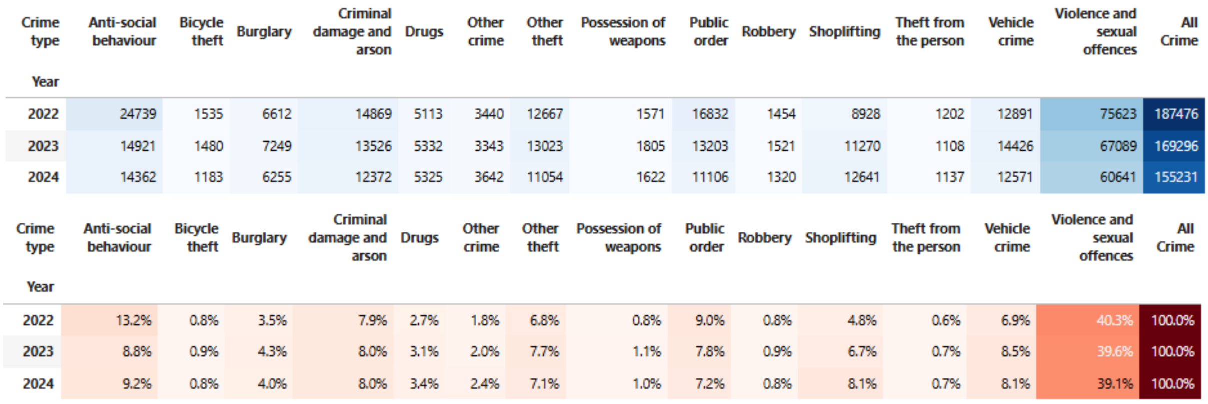
This chapter presents the results of the data analysis and modelling processes applied to the Essex crime dataset. A combination of exploratory data analysis (EDA), time-series forecasting, machine learning, and clustering techniques were used to uncover trends, patterns, and spatial crime hotspots between 2022 and 2024. In addition, the integration of a Retrieval-Augmented Generation (RAG) chatbot demonstrates how natural language interfaces can make crime data more accessible. Each section details the analytical approach taken, highlights key findings, and sets the foundation for the critical discussion and evaluation later in the paper.

Exploratory Data Analysis (EDA) of Crime Data

Between 2022 and 2024, Essex Police recorded a total of 518,461 reported crimes, averaging around 14,400 offences per month or 473 per day. While the data shows a slight year-on-year decrease in overall crime, the consistently high monthly and daily figures point to a persistent level of criminal activity across the county. These patterns offer key insights into long-term crime trends and fluctuations in public safety demand.

A closer look at the breakdown by crime type reveals significant variation in prevalence and behaviour over time. "Violence and sexual offences" remain the most common crime category, accounting for 40.3% of all reported crimes in 2022 (75,623 incidents). This follows a similar pattern seen across the UK, where the police flagged 827,609 offences as domestic abuse-related in YE September 2024 (ONS 2024). The dominance across the three-year period signals a longstanding challenge for law enforcement, potentially influenced by social, economic, and cultural factors. It also highlights areas where public protection strategies and offender management may require sustained attention.

In contrast, "Anti-social behaviour", while less frequent overall, still constitutes a notable portion of the crime landscape, with 54,022 incidents recorded between 2022 and 2024. Although this category shows less volatility than others, it has implications for community cohesion and perceptions of safety, particularly in urban centres and areas with younger populations.

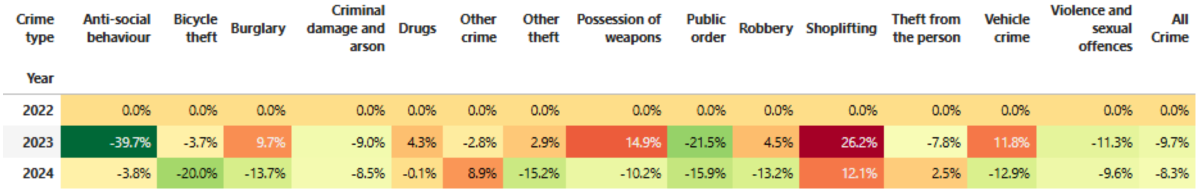


Crime type	Anti-social behaviour	Bicycle theft	Burglary	Criminal damage and arson	Drugs	Other crime	Other theft	Possession of weapons	Public order	Robbery	Shoplifting	Theft from the person	Vehicle crime	Violence and sexual offences	All Crime
Year															
2022	24739	1535	6612	14869	5113	3440	12667	1571	16832	1454	8928	1202	12891	75623	187476
2023	14921	1480	7249	13526	5332	3343	13023	1805	13203	1521	11270	1108	14426	67089	169296
2024	14362	1183	6255	12372	5325	3642	11054	1622	11106	1320	12641	1137	12571	60641	155231

Crime type	Anti-social behaviour	Bicycle theft	Burglary	Criminal damage and arson	Drugs	Other crime	Other theft	Possession of weapons	Public order	Robbery	Shoplifting	Theft from the person	Vehicle crime	Violence and sexual offences	All Crime
Year															
2022	13.2%	0.8%	3.5%	7.9%	2.7%	1.8%	6.8%	0.8%	9.0%	0.8%	4.8%	0.6%	6.9%	40.3%	100.0%
2023	8.8%	0.9%	4.3%	8.0%	3.1%	2.0%	7.7%	1.1%	7.8%	0.9%	6.7%	0.7%	8.5%	39.6%	100.0%
2024	9.2%	0.8%	4.0%	8.0%	3.4%	2.4%	7.1%	1.0%	7.2%	0.8%	8.1%	0.7%	8.1%	39.1%	100.0%

Figure 4A – Crime count and share across Essex by year and crime type

One crime type that deviates from the general downward or fluctuating trends is shoplifting, which has exhibited consistent year-on-year increases across Essex from 2022 through 2024. This sustained rise suggests a growing issue that may be linked to wider social or economic pressures, such as the cost-of-living crisis and/or reduced staffing in retail security. In contrast, possession of weapons experienced a notable 14% increase in 2023 compared to the previous year, followed by a decline in 2024. Although this spike appears to be short-term, it could indicate a temporary surge in violence-related incidents or proactive policing initiatives, such as increased stop-and-search activity. Regardless, this trend may warrant closer examination by crime analysts and policymakers at Essex Police, as even brief escalations in weapon-related offences can have significant implications for public safety. Monitoring such changes can help inform targeted interventions, community outreach, or legislative adjustments in the future.



Crime type Year	Anti-social behaviour	Bicycle theft	Burglary	Criminal damage and arson	Drugs	Other crime	Other theft	Possession of weapons	Public order	Robbery	Shoplifting	Theft from the person	Vehicle crime	Violence and sexual offences	All Crime
2022	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2023	-39.7%	-3.7%	9.7%	-9.0%	4.3%	-2.8%	2.9%	14.9%	-21.5%	4.5%	26.2%	-7.8%	11.8%	-11.3%	-9.7%
2024	-3.8%	-20.0%	-13.7%	-8.5%	-0.1%	8.9%	-15.2%	-10.2%	-15.9%	-13.2%	12.1%	2.5%	-12.9%	-9.6%	-8.3%

Figure 4B – Year on year percentage change in crime by crime type in Essex

The distribution of crime across Essex is uneven and varies significantly from district to district, highlighting regional differences in the prevalence and types of offences committed. Data from 2022 to 2024 indicates that Southend-on-Sea, Basildon, and Colchester consistently report the highest levels of crime in the county. In contrast, districts such as Brentwood, Rochford, and Maldon report some of the lowest levels of recorded crime.

It's important to note that variations in crime levels across districts are influenced by a complex interplay of social, economic, and environmental factors. For example, districts with higher levels of deprivation or unemployment may experience more frequent reports of certain crimes, while areas with robust community engagement and proactive policing may see lower figures. Additionally, differences in reporting practices, public trust in law enforcement, and policing resources can all affect the number of recorded incidents.

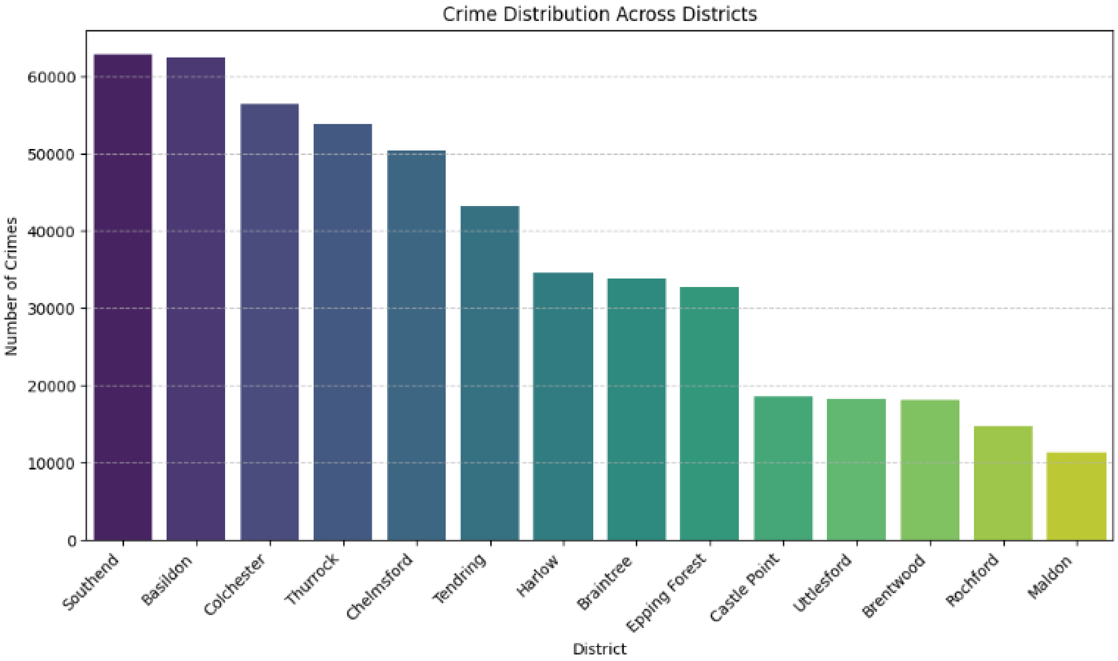
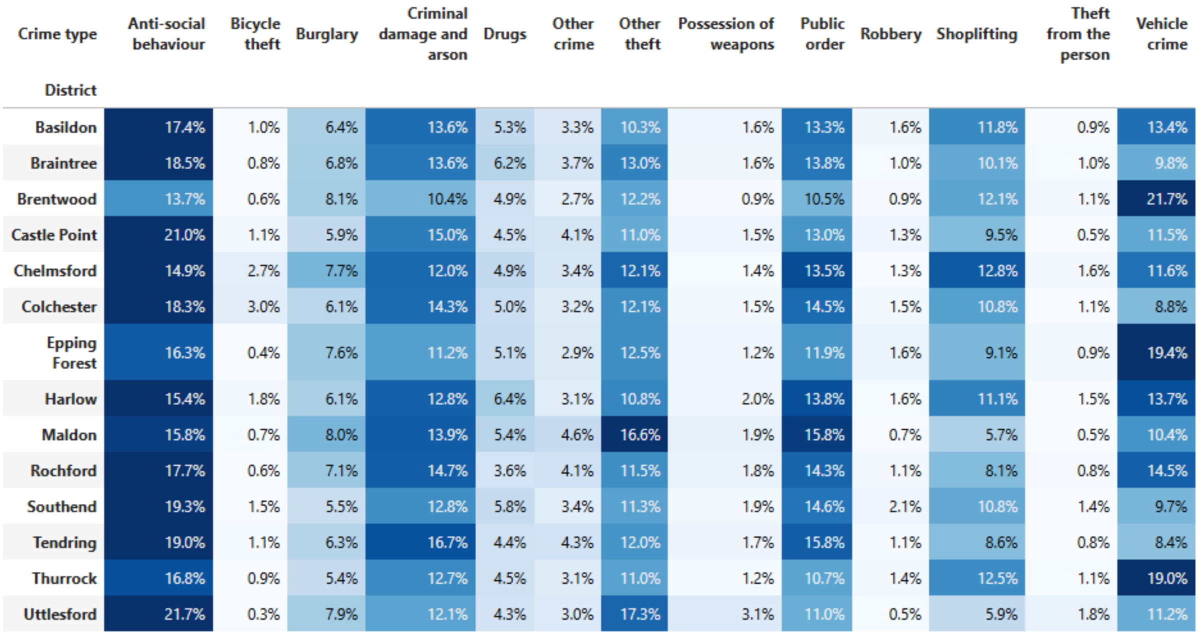


Figure 4C – Crime distributions across the districts of the Essex police force region

Although *Violence and sexual offences* consistently dominate as the most reported crime type across all districts in Essex, a deeper analysis reveals notable variation in the distribution of other crime types at the district level. When this dominant category is omitted from the analysis, new patterns emerge that highlight localised crime dynamics.

Figure Y illustrates with the darker shades, that *Anti-social Behaviour* becomes significantly more prominent, particularly in the district of Uttlesford, where it accounts for 21.7% of reported crimes. This suggests that community-related disorder may be a greater concern in this more rural area. In contrast, the share of *Anti-social Behaviour* is relatively lower in districts such as Brentwood, Epping Forest, and Thurrock, where *Vehicle crime* makes up a larger proportion of total crime. This is likely due to these districts neighbouring London so greater commuter traffic into the city makes it more vulnerable to Vehicle crime.

Additional district-level crime spikes include Maldon, which has an unusually high share of *Public Order* offences at 15.8%. Similarly, Tendring stands out with *Criminal damage and arson* comprising 16.7% of its crime profile—suggesting property-related violence is more of a localised concern there than elsewhere in the county.



Crime type	Anti-social behaviour	Bicycle theft	Burglary	Criminal damage and arson	Drugs	Other crime	Other theft	Possession of weapons	Public order	Robbery	Shoplifting	Theft from the person	Vehicle crime
District													
Basildon	17.4%	1.0%	6.4%	13.6%	5.3%	3.3%	10.3%	1.6%	13.3%	1.6%	11.8%	0.9%	13.4%
Braintree	18.5%	0.8%	6.8%	13.6%	6.2%	3.7%	13.0%	1.6%	13.8%	1.0%	10.1%	1.0%	9.8%
Brentwood	13.7%	0.6%	8.1%	10.4%	4.9%	2.7%	12.2%	0.9%	10.5%	0.9%	12.1%	1.1%	21.7%
Castle Point	21.0%	1.1%	5.9%	15.0%	4.5%	4.1%	11.0%	1.5%	13.0%	1.3%	9.5%	0.5%	11.5%
Chelmsford	14.9%	2.7%	7.7%	12.0%	4.9%	3.4%	12.1%	1.4%	13.5%	1.3%	12.8%	1.6%	11.6%
Colchester	18.3%	3.0%	6.1%	14.3%	5.0%	3.2%	12.1%	1.5%	14.5%	1.5%	10.8%	1.1%	8.8%
Epping Forest	16.3%	0.4%	7.6%	11.2%	5.1%	2.9%	12.5%	1.2%	11.9%	1.6%	9.1%	0.9%	19.4%
Harlow	15.4%	1.8%	6.1%	12.8%	6.4%	3.1%	10.8%	2.0%	13.8%	1.6%	11.1%	1.5%	13.7%
Maldon	15.8%	0.7%	8.0%	13.9%	5.4%	4.6%	16.6%	1.9%	15.8%	0.7%	5.7%	0.5%	10.4%
Rochford	17.7%	0.6%	7.1%	14.7%	3.6%	4.1%	11.5%	1.8%	14.3%	1.1%	8.1%	0.8%	14.5%
Southend	19.3%	1.5%	5.5%	12.8%	5.8%	3.4%	11.3%	1.9%	14.6%	2.1%	10.8%	1.4%	9.7%
Tendring	19.0%	1.1%	6.3%	16.7%	4.4%	4.3%	12.0%	1.7%	15.8%	1.1%	8.6%	0.8%	8.4%
Thurrock	16.8%	0.9%	5.4%	12.7%	4.5%	3.1%	11.0%	1.2%	10.7%	1.4%	12.5%	1.1%	19.0%
Uttlesford	21.7%	0.3%	7.9%	12.1%	4.3%	3.0%	17.3%	3.1%	11.0%	0.5%	5.9%	1.8%	11.2%

Figure 4D – Crime share across each district by crime type

While crime patterns vary significantly across different districts and crime groups, a consistent and concerning outcome for Essex County is the overall conviction rate for reported crimes. Over the three-year period analysed, Essex Police recorded approximately half a million reported crimes. However, only around 7.3% of these incidents led to a potential conviction, whether through suspects awaiting court proceedings or offenders receiving official cautions. As illustrated by the pie chart in Figure 4E, approximately 82.3% of reported crimes resulted in no conviction outcome. These cases typically concluded with outcomes such as investigations being closed without identifying a suspect, a decision not to pursue further action, or other administrative closures. Additionally, about 10% of crimes had either a blank or unknown last recorded outcome, suggesting gaps in the reporting or updating of case statuses.

The low conviction proportion underscores the importance of targeted interventions, improved investigative resources, and continuous monitoring of case outcomes to ensure more crimes are successfully prosecuted or resolved.

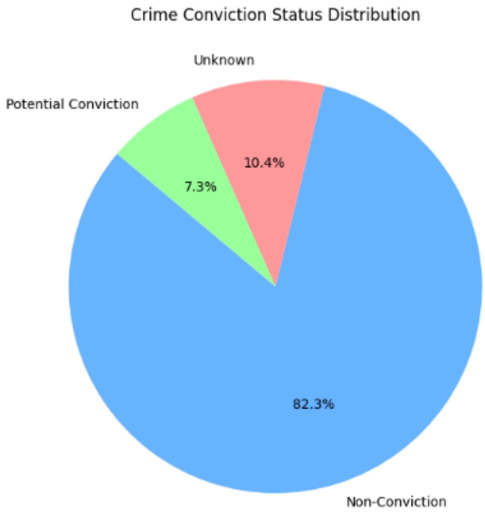


Figure 4E – Crime conviction percentage for Essex crime

Time-Series Analysis of Crime Rates

Crime in Essex does not occur at a consistent rate throughout the year. Analysis of data from 2022 to 2024 reveals seasonal patterns, with noticeable fluctuations in crime volumes depending on the month, as shown in Figure 4F. Notably, crime levels tend to spike during March, July, and August. The rises in these months could be down to a few different factors. For example, warmer weather and school holidays in July and August often correlate with increased outdoor activity, larger public gatherings, and more opportunities for certain types of crime such as anti-social behaviour, theft, and public order offences. In contrast, the winter months of November and December typically show a noticeable decline in total crime numbers. This could be attributed to factors such as shorter daylight hours, colder weather, and reduced public activity, all of which can serve as natural deterrents to some types of crimes, especially those committed in public spaces.

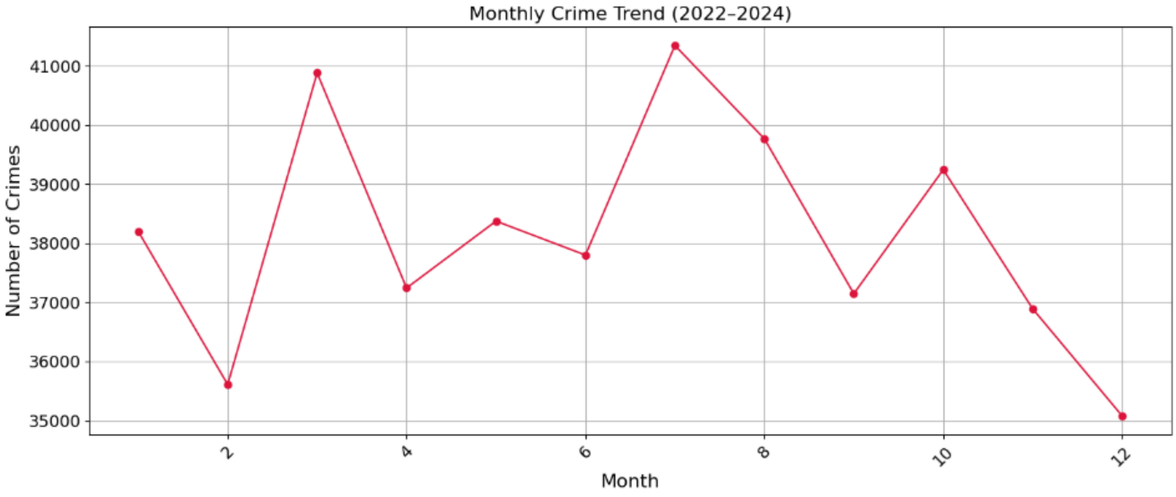


Figure 4F – 2022 to 2024 Crime count by month for Essex

Spikes in crime also vary significantly across different crime types. While the majority of offences generally follow the overall trend of declining or stabilising crime levels in Essex, shoplifting stands out as an exception. Unlike other crime types, shoplifting has demonstrated a consistent upward trajectory over the years. In 2022, Essex Police recorded an average of approximately 700 shoplifting offences per month, which steadily increased to around 1,300 offences per month by the latter half of 2024 — nearly doubling within two years.

As illustrated in Figure 4G, the monthly distribution of shoplifting incidents reveals recurring seasonal spikes, with the most pronounced increases occurring during the winter months of November and December. This goes against the 'whole picture' which showed spikes in the month of March, July, and August. These insights highlight the importance of disaggregating crime data by type and time period to better understand specific patterns that might be obscured in aggregated trends.

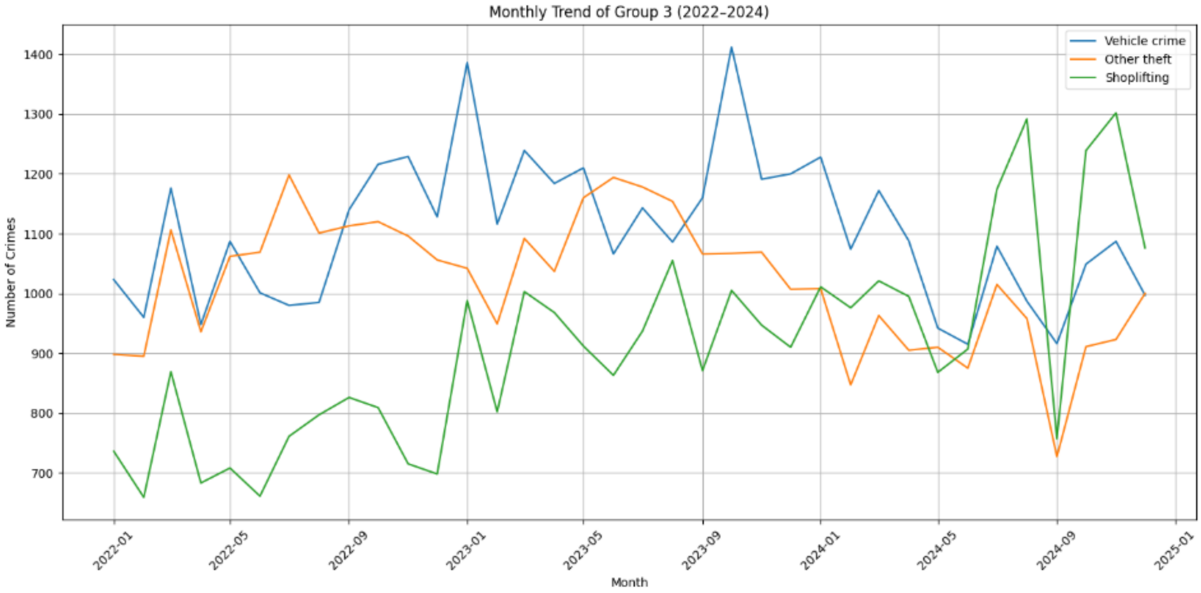


Figure 4G – 2022 to 2024 Crime count by crime types, Vehicle crime, Other theft, and Shoplifting

Crime Prediction Models

Socio-economic conditions play a significant role in shaping crime patterns. Courson and Nettle (2021) developed a model demonstrating that societies marked by greater inequality tend to experience higher levels of crime and lower levels of social trust. Their model also suggests that individuals may still choose to engage in exploitative behaviour—even when the expected outcome is negative due to strict penalties—driven by two key factors: the potential for high, variable rewards and a critical threshold of desperation that individuals seek to avoid falling below.

In Essex, these dynamics are particularly relevant. Approximately 188,000 people across the Essex Police Force area live in the most deprived 20% of England, with concentrations of deprivation especially notable in coastal towns such as Southend-on-Sea, Harwich, and Clacton (Essex County Council, 2024).

To explore the relationship between deprivation and crime more formally, I conducted a regression analysis examining crime counts against the Index of Multiple Deprivation (IMD) deciles. These deciles rank areas from 1 (most deprived) to 10 (least deprived) within the Essex region, offering insight into how socio-economic disadvantage may correlate with levels of reported crime.

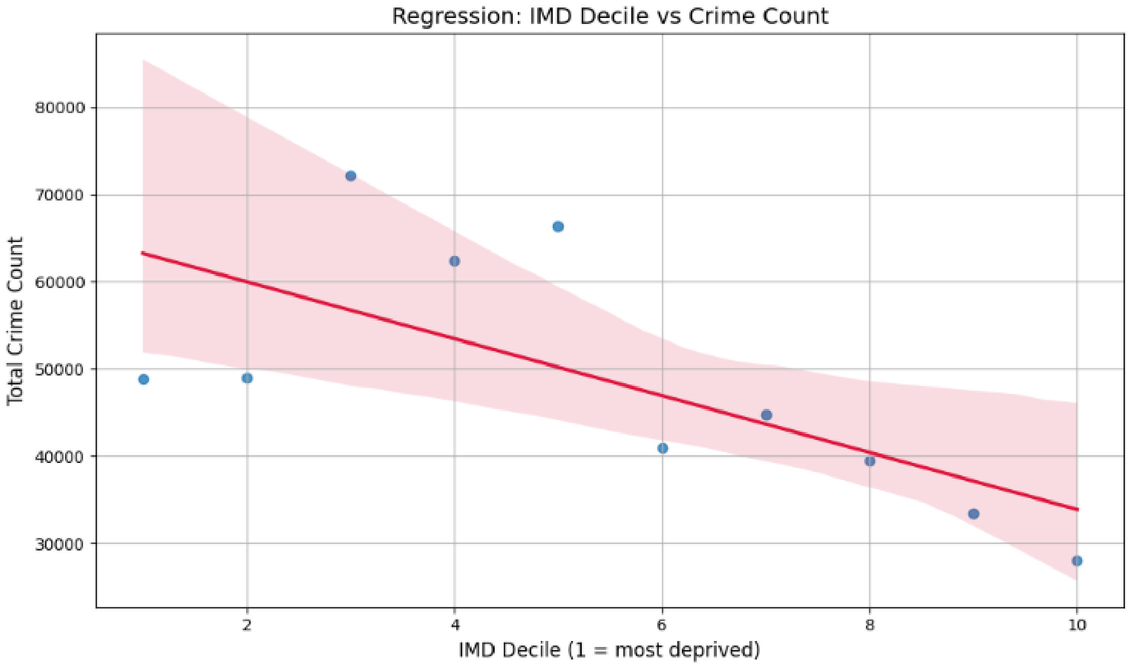
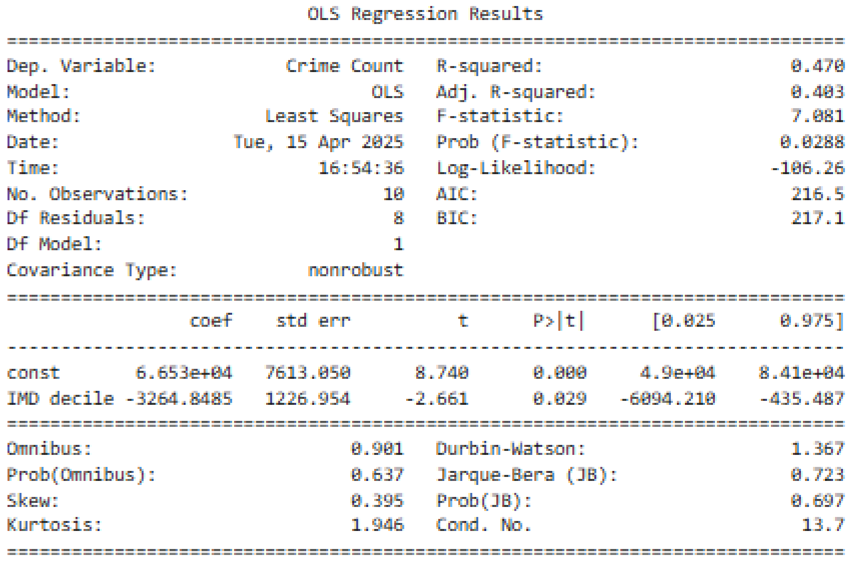


Figure 4H – Regress graph for deprivation and crime count relationship



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Crime Count   R-squared:                       0.470
Model:                            OLS   Adj. R-squared:                  0.403
Method:                 Least Squares   F-statistic:                     7.081
Date:                Tue, 15 Apr 2025   Prob (F-statistic):             0.0288
Time:                        16:54:36   Log-Likelihood:                -106.26
No. Observations:                  10   AIC:                             216.5
Df Residuals:                       8   BIC:                             217.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       6.653e+04   7613.050      8.740      0.000     4.9e+04    8.41e+04
IMD decile -3264.8485   1226.954     -2.661      0.029   -6094.210    -435.487
==============================================================================
Omnibus:                        0.901   Durbin-Watson:                   1.367
Prob(Omnibus):                  0.637   Jarque-Bera (JB):                0.723
Skew:                           0.395   Prob(JB):                        0.697
Kurtosis:                       1.946   Cond. No.                         13.7
==============================================================================
```

Figure 4I – OLS Regression results for deprivation and crime count

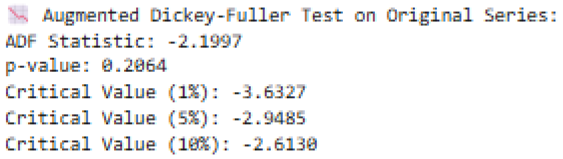
Figure 4H shows a negative correlation between IMD decile and crime count: lower-decile (more deprived) areas consistently experience higher crime levels than affluent ones. This is supported by the OLS regression in Figure 4I, which yields an R^2 of 0.47, indicating deprivation explains 47% of the variation in crime. The p-value (0.029) confirms this relationship is statistically significant. The regression coefficient suggests that each 1-point increase in IMD decile is associated with roughly 3,265 fewer crimes over the 2022–2024 period.

To forecast future crime, a SARIMA model was used. SARIMA extends ARIMA by incorporating seasonality through a second set of parameters:

SARIMA(p,d,q) × (P,D,Q,s), where seasonal and non-seasonal components capture repeating patterns. It combines:

- AR: past values
- I: differencing to remove trend
- MA: past error smoothing
- plus
- SAR, SI, SMA for seasonal effects (e.g., annual cycles)

Observed spikes in March, July, and August (Figure 4F) suggest seasonal trends, which were confirmed via the Augmented Dickey-Fuller test (p = 0.2064). As the result is above 0.05, we fail to reject the null hypothesis, indicating non-stationarity and the presence of seasonality in the data.

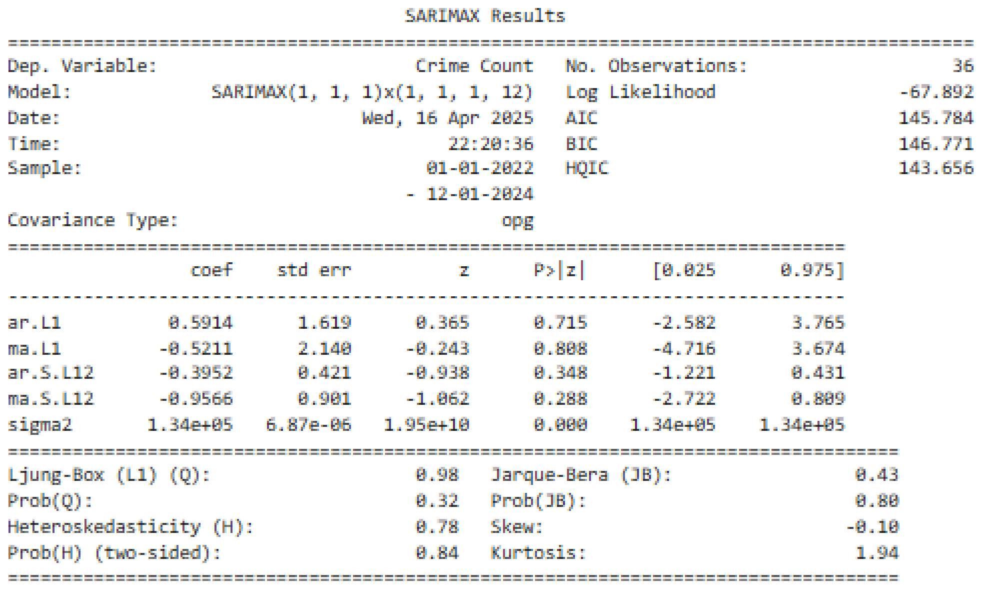


```
📉 Augmented Dickey-Fuller Test on Original Series:
ADF Statistic: -2.1997
p-value: 0.2064
Critical Value (1%): -3.6327
Critical Value (5%): -2.9485
Critical Value (10%): -2.6130
```

Figure 4J – Augmented Dickey-Fuller test results for Seasonality

The SARIMA(1,1,1)(1,1,1,12) model applied one level of differencing to remove trends and included autoregressive and moving average terms to account for temporal dependencies. The seasonal parameters indicate yearly seasonality with annual differencing. Although individual coefficients were not statistically significant (p > 0.05), the model achieved a low AIC (145.78), suggesting a good fit. Residual diagnostics (Ljung-Box p = 0.32, Jarque-Bera p = 0.80) confirmed that residuals behaved like white noise, indicating no significant autocorrelation or deviation from normality (Figure 4K).

Figure 4L shows the SARIMA forecast for the first half of 2025, predicting moderate fluctuations with a potential peak of 15,000 crimes in one month. The widening confidence intervals reflect increasing uncertainty, but the model effectively captures the underlying seasonal trends without projecting major deviations.



```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                       Crime Count   No. Observations:                   36
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                 -67.892
Date:                           Wed, 16 Apr 2025   AIC                            145.784
Time:                                   22:20:36   BIC                            146.771
Sample:                               01-01-2022   HQIC                           143.656
                                    - 12-01-2024
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.5914      1.619      0.365      0.715      -2.582       3.765
ma.L1         -0.5211      2.140     -0.243      0.808      -4.716       3.674
ar.S.L12      -0.3952      0.421     -0.938      0.348      -1.221       0.431
ma.S.L12      -0.9566      0.901     -1.062      0.288      -2.722       0.809
sigma2       1.34e+05   6.87e-06   1.95e+10      0.000    1.34e+05    1.34e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.98   Jarque-Bera (JB):                 0.43
Prob(Q):                              0.32   Prob(JB):                         0.80
Heteroskedasticity (H):               0.78   Skew:                            -0.10
Prob(H) (two-sided):                  0.84   Kurtosis:                         1.94
===================================================================================
```

Figure 4K – SARIMAX results for seasonality and goodness of fit for Essex crime data 2022 to 2024

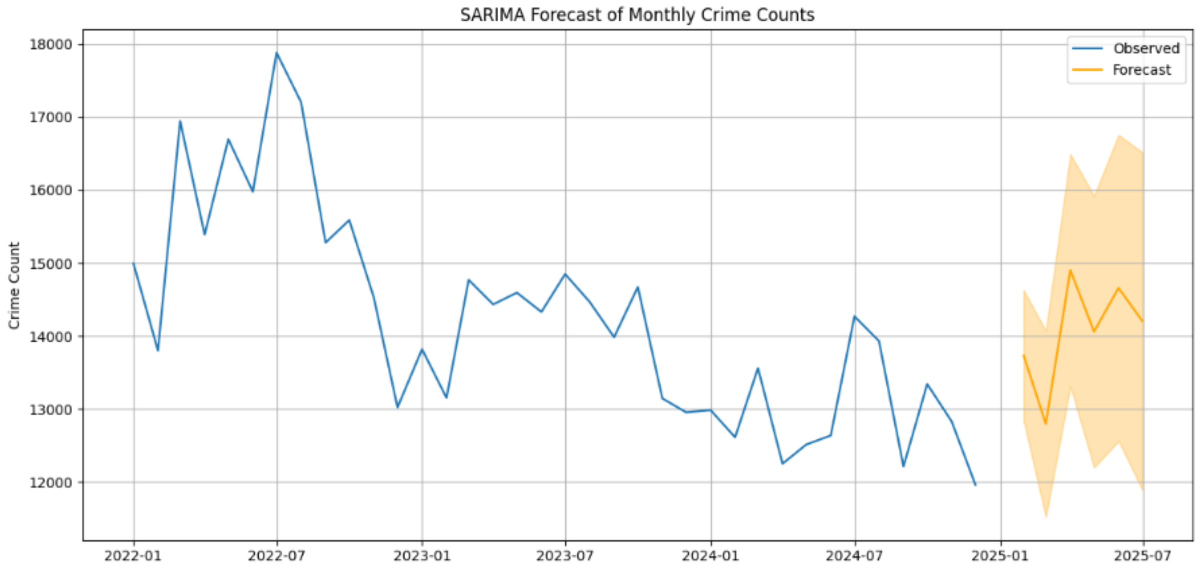


Figure 4L – SARIMA Crime forecast for future months in 2025 based off 2022 to 2024 crime data for Essex

Hotspot Identification with Clustering

Identifying geographic concentrations of crime, or hotspots, is important for understanding spatial crime dynamics. In this section, clustering algorithms are applied to the Essex crime dataset to detect patterns and highlight areas with elevated crime intensity. Both K-Means and DBSCAN methods are utilised to reveal spatial groupings, with each technique offering distinct advantages. K-Means provides an overview of broad crime zones across the county, while DBSCAN captures more nuanced, irregular clusters that may otherwise be obscured. The outputs are visualised using interactive maps, providing valuable insights into the distribution of crime types and informing targeted policing efforts.

Initially, crime counts per LSOA were analysed to visualise the spatial distribution of crime across the Essex police area. Using libraries such as GeoPandas and Folium, an interactive map was created to identify clusters more effectively. As shown in Figure 4M, the choropleth map highlights LSOA regions with higher crime rates in darker shades of red, with the darkest areas representing more than 500 reported crimes during the 2022–2024 period. Notable clusters emerge in the southwest of the county, particularly near neighbouring London boroughs, as well as in seaside towns such as Clacton-on-Sea (northeast) and Southend-on-Sea. These high-concentration zones align with earlier findings from the EDA, where the LSOA Southend-on-Sea 015B recorded over 6,000 reported crime incidents between 2022 and 2024.

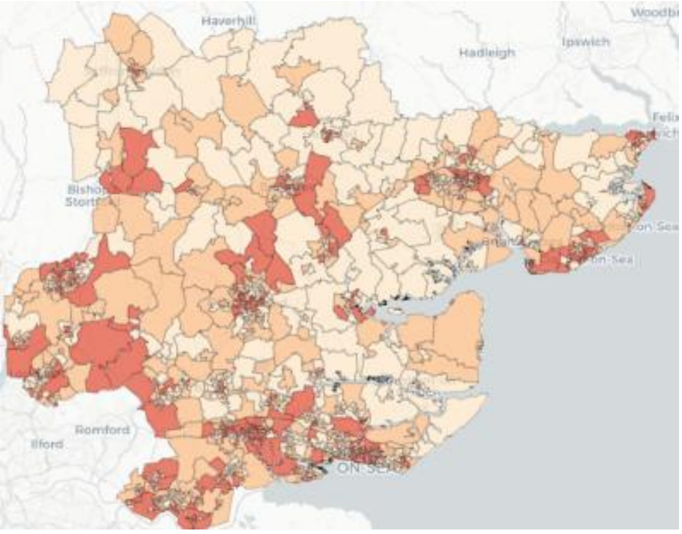
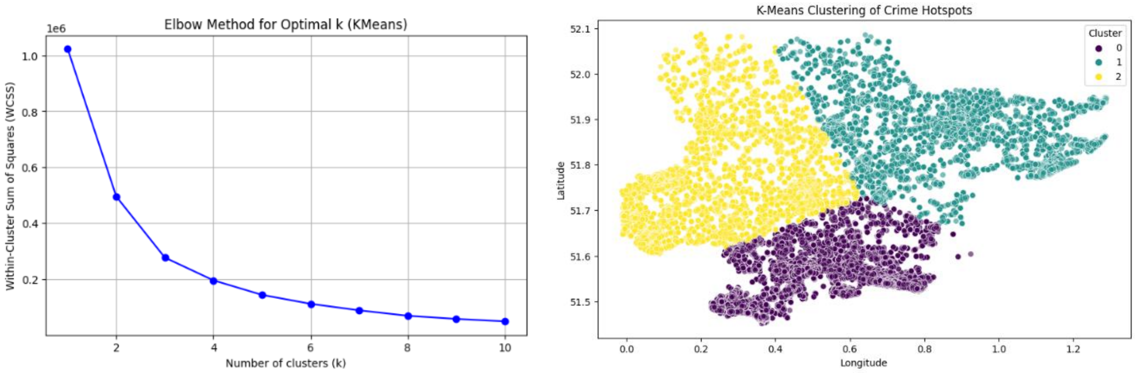


Figure 4M - Crime distribution across Essex LSOAs (2022–2024), with darker red areas indicating higher reported crime concentrations.

To accompany the crime distribution by LSOA, we can use K-Means clustering to uncover spatial patterns by segmenting geographic locations into zones with similar activity levels. K-Means clustering is an unsupervised machine learning technique used to group data points into a predefined number of clusters based on similarity. K-Means aims to minimise intra-cluster variance, ensuring that all points within a cluster are geographically or statistically similar to one another and positioned closely around the cluster's centroid

To help identify the optimal number of clusters, the Elbow Method (Figure 4N) was used to calculate the within-cluster sum of squares (WCSS). The WCSS sharply declined until k = 3, after which the curve levelled off. This inflection point indicates that three clusters provide an optimal balance between model complexity and intra-cluster variance reduction.

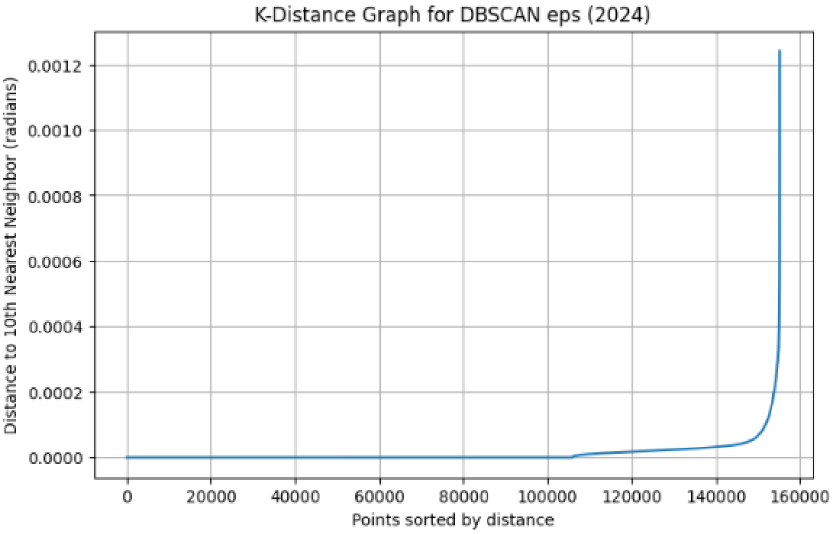
The K-Means clustering output, shown in Figure 4O, reveals three dominant spatial crime zones across Essex, with each cluster representing areas of similar crime density and geographic proximity. These clusters broadly align with the previously mapped LSOA-level crime concentrations. For instance, the southern region of Essex—captured in Cluster 0—includes high-density areas such as Southend-on-Sea and Basildon, which were also highlighted as dark red zones in the LSOA choropleth map. Similarly, Cluster 1 in the northeast corresponds to more moderately active districts like Colchester and Tendring, while Cluster 2 spans the west and northwest, encompassing areas adjacent to the London border that also displayed elevated crime counts. The correspondence between K-Means groupings and LSOA-defined hotspots strengthens the validity of the clustering model, showing that K-Means successfully abstracts administrative crime volumes into coherent spatial patterns.



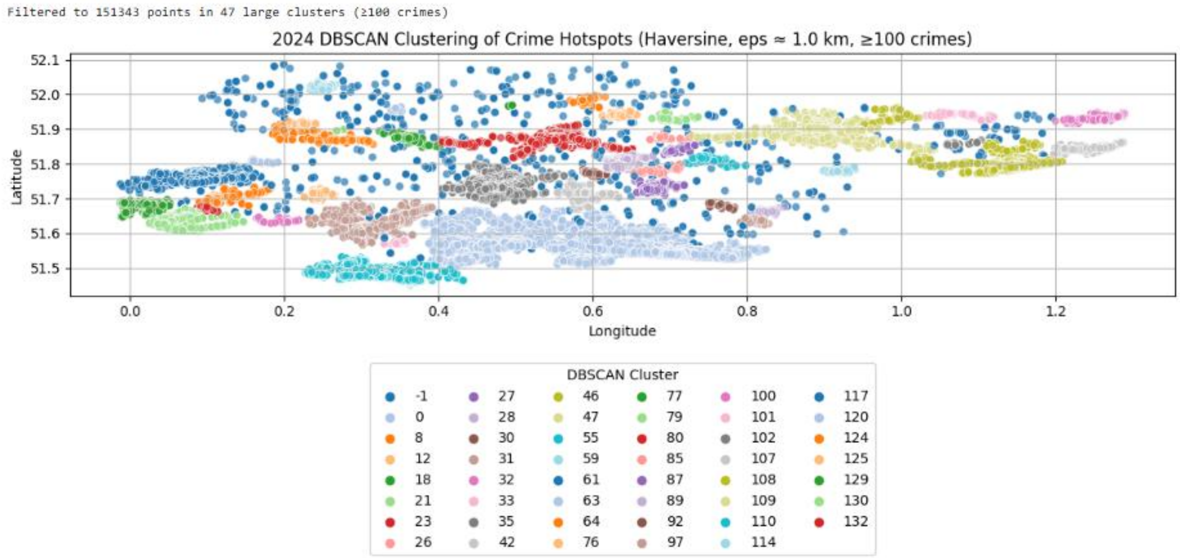
Figures 4N & 4O - Elbow method optimality graph using within-cluster sum of squares, & K-Means clustering graph for Essex

To focus into clusters further, DBSCAN algorithms were used. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised machine learning algorithm that groups together data points that are closely packed, marking points in low-density areas as outliers or noise. This method was particularly suited to crime data analysis, where the goal was to discover areas with unusually high concentrations of criminal incidents without prior assumptions about the number of clusters.

Using the DBSCAN algorithm with a radius of approximately 1.0 km and a minimum cluster size threshold of 100 crimes, spatial clusters of crime hotspots in Essex during 2024 were effectively identified. The K-Distance graph, calculated with the 10th nearest neighbour, guided the selection of an appropriate clustering eps value of 0.00015 or ~1km. The resulting DBSCAN model produced 47 significant clusters, each representing a concentrated area of criminal activity. These clusters were visualised on a scatter plot, in figure 4Q, with unique colours and a legend to differentiate between them, offering a clear view of the geographic distribution and density of crime. Following this, the cluster data was integrated into a Folium map, where each cluster was plotted interactively with color-coded markers. This map enables users to zoom in on individual hotspots, click on markers for context (e.g., cluster ID or crime type), and explore spatial crime patterns with greater clarity and utility than static plots allow.



Figures 4P – K-Distance graphs for choosing an appropriate eps value for clustering



```
Filtered to 151343 points in 47 large clusters (≥100 crimes)
```

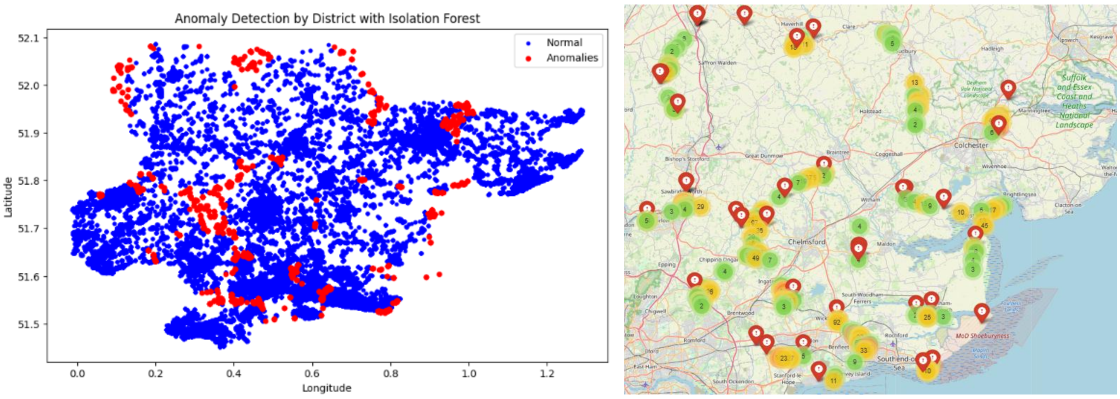
Figures 4Q – DBSCAN scatterplot for crime clusters in Essex 2024

Anomaly Detection in Crime Data

Anomalies and outliers are unavoidable in real everyday data, and crime data is no exception. Whilst they may be a small representation of the dataset, the impact can be significant (Majzoub and Saleh, 2025). By identifying patterns or behaviours that deviate significantly from normal historical trends, police forces can begin to understand hidden patterns. Anomaly analysis will highlight emerging crime hotspots, unexpected spikes in particular offenses, or geographic areas experiencing unusual levels of criminal activity. Detecting these outliers early can provide valuable insights for proactive intervention, resource allocation, and policy decision-making. Instead of relying solely on predefined patrol routes or historical averages, anomaly detection enables dynamic, evidence-driven responses to evolving threats.

With the dataset cleaned and longitude/latitude coordinates being standardised to help create a non-bias model. Isolation Forest is then applied with a contamination rate of 1%, meaning it is expected that approximately 1% of points in each district represent anomalies. Crimes flagged as anomalies are separated from normal crimes and visualized both on a scatter plot and an interactive Folium map. This allows for a clear comparison between areas of expected crime patterns and potential emerging crime hotspots.

As shown in figure 4T, we can see the anomalies for each districted highlighted red. A pattern emerges slightly with anomalies being highlighted on the bordering or districts and the county of Essex itself. We then have Figure 4T showing the anomalies in an interactive folium map, which has labels for the data points to explain what crime type it is and what district it is located in.

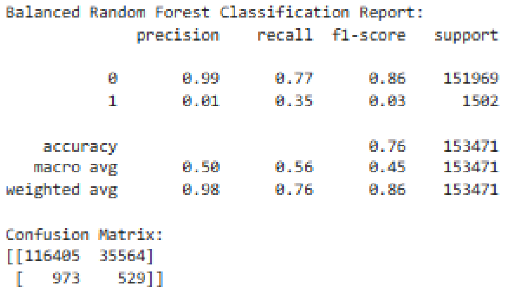


Figures 4R & 4S – Static Anomaly Detection by District with Isolation Forest and interactive folium map

After detecting anomalies, a Random Forest Classifier was trained to predict whether a crime instance was anomalous (1) or normal (0). This binary classification used features like Crime Type, District, and Month, which were label-encoded for model compatibility. The dataset was split into training and testing sets to evaluate performance. A balanced Random Forest with 100 trees was trained using class weights to account for the rarity of anomalies.

The model achieved a recall of roughly 35% for anomalies and correctly identified 529 anomalous cases, despite some false positives. A classification report and confusion matrix were used for evaluation. Figure 4U presents feature importance: Month was the strongest predictor (50%), suggesting seasonal influence, followed by District (25%) and Crime Type (20%).

This approach offers practical insights into when and where anomalies occur, supporting data-driven decision-making for patrol planning and resource allocation.

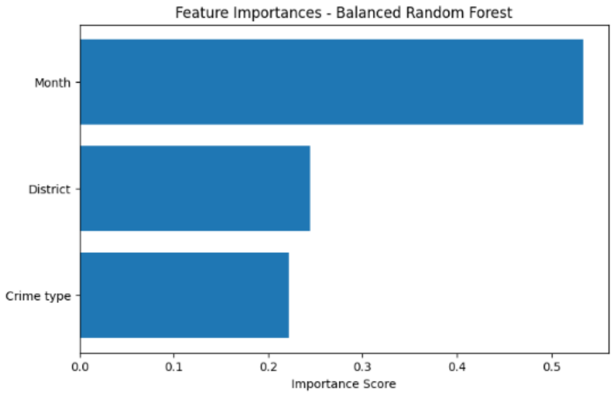


```
Balanced Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.77      0.86    151969
           1       0.01      0.35      0.03      1502

    accuracy                           0.76    153471
   macro avg       0.50      0.56      0.45    153471
weighted avg       0.98      0.76      0.86    153471

Confusion Matrix:
[[116405  35564]
 [   973    529]]
```



Figures 4T & 4U – Random Forest Classification report for significance and feature importance by variables Month, District, and Crime Type

RAG System Development & Dashboard

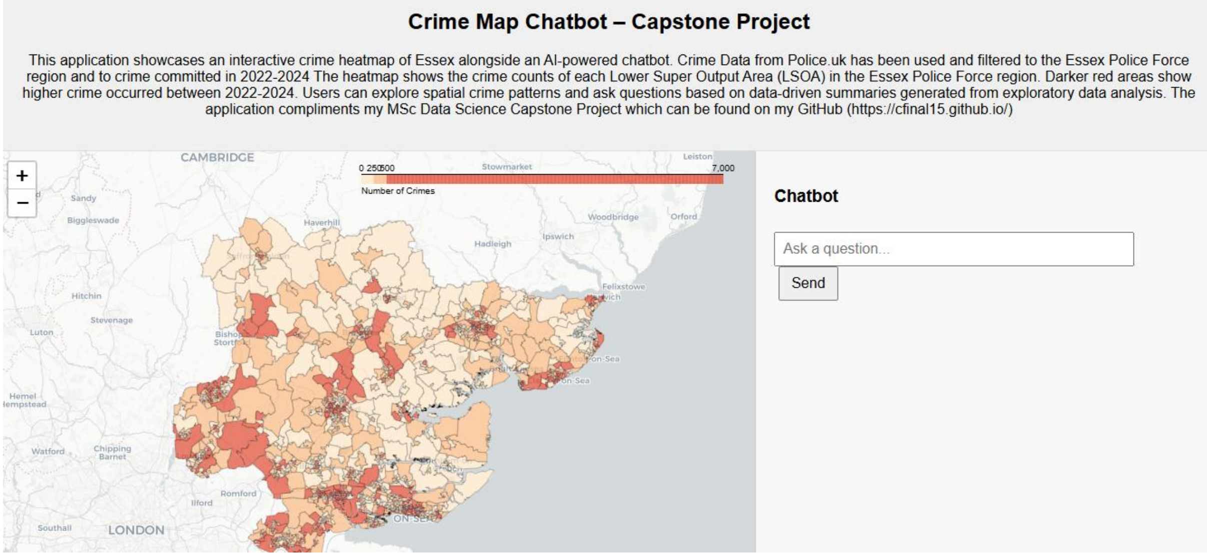
Public access to performance data significantly enhances trust in policing institutions (Mason et al 2014). To make the crime data used in this project more accessible and useful, a Retrieval-Augmented Generation (RAG) system was developed using natural language processing (NLP) techniques. This system allows users to ask questions about crime trends in plain language, instead of having to interpret complex charts or reports.

The RAG pipeline uses pre-processed summaries of key findings—such as monthly crime trends, types of offences, and hotspot areas—which were structured into a JSON format. These summaries were then stored in a vector database using FAISS for fast and relevant search results. Code for this was stored in the publicly available GitHub repository, https://github.com/cfinal15/crime_map.

The system was built using LangChain, a Python framework for creating AI-driven applications. It uses OpenAI's language model combined with a vector retriever through the RetrievalQA chain. Embeddings were generated by OpenAIEmbeddings and stored into the publicly available GitHub repository. From there, embeddings were converted into LangChain-compatible objects.

To make the system usable by the public, a simple Flask web application was built (Figure 4V). This app manages the user interface and connects the chatbot to a backend endpoint. The front end, made with HTML and CSS, includes a responsive layout that combines an interactive Folium map, showing crime by LSOA, with a chat window. Users can ask questions like "What are the top crime hotspots?" or "How has crime changed over time?" and receive instant, tailored answers, as shown in figure 4W.

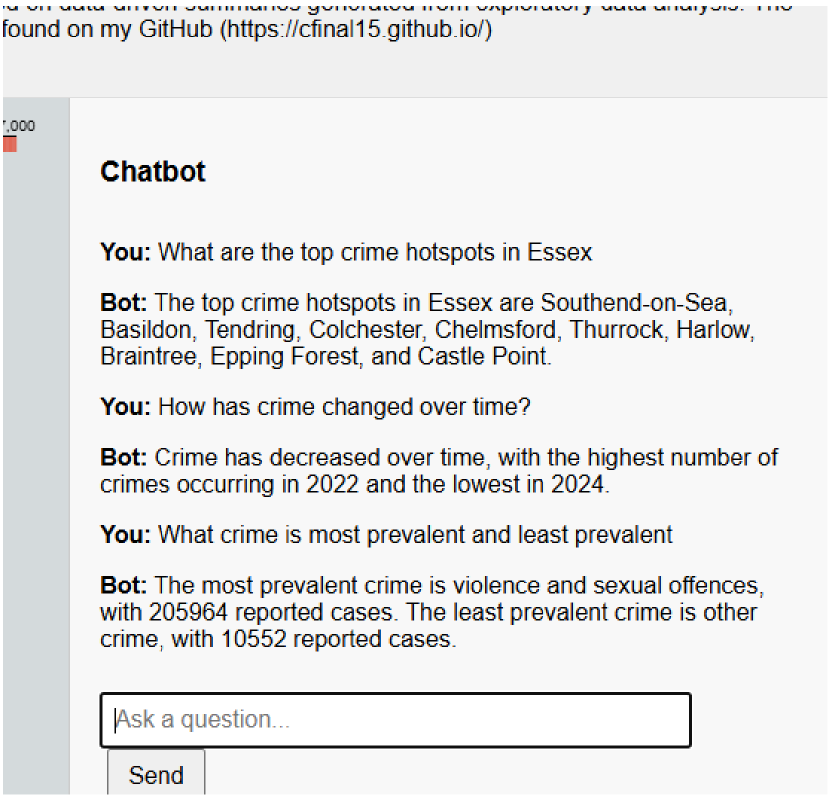
The full system was deployed on Render, a cloud hosting service connected to the project's GitHub repository. The publicly available URL https://crime-map-chatbot.onrender.com is automatically updated whenever new code is pushed to GitHub. Render handles dependency installation and runs the Flask server using Gunicorn. This streamlined deployment workflow enables rapid iteration and continuous delivery, resulting in a user-friendly, interactive dashboard that combines geospatial data with AI-powered crime insights that are accessible to both analysts and the general public.



Crime Map Chatbot – Capstone Project

This application showcases an interactive crime heatmap of Essex alongside an AI-powered chatbot. Crime Data from Police.uk has been used and filtered to the Essex Police Force region and to crime committed in 2022-2024 The heatmap shows the crime counts of each Lower Super Output Area (LSOA) in the Essex Police Force region. Darker red areas show higher crime occurred between 2022-2024. Users can explore spatial crime patterns and ask questions based on data-driven summaries generated from exploratory data analysis. The application compliments my MSc Data Science Capstone Project which can be found on my GitHub (https://cfinal15.github.io/)

Figures 4V –Render Crime Dashboard for public accessibility available at: https://crime-map-chatbot.onrender.com/



found on my GitHub (https://cfinal15.github.io/)

Chatbot

You: What are the top crime hotspots in Essex

Bot: The top crime hotspots in Essex are Southend-on-Sea, Basildon, Tendring, Colchester, Chelmsford, Thurrock, Harlow, Braintree, Epping Forest, and Castle Point.

You: How has crime changed over time?

Bot: Crime has decreased over time, with the highest number of crimes occurring in 2022 and the lowest in 2024.

You: What crime is most prevalent and least prevalent

Bot: The most prevalent crime is violence and sexual offences, with 205964 reported cases. The least prevalent crime is other crime, with 10552 reported cases.

Figures 4W –Crime Dashboard chatbot outputs for questions relating to Essex crime data 2022 to 2024

5 - Discussion, Conclusion and Recommendations

Data analysis overview and objective recap

For my discussion, conclusion, and evaluation, I have critically evaluated the results presented in the previous section, exploring their implications, limitations, and alignment with the project's original objectives. The aim of the project was to apply data science techniques to understand, forecast, and communicate crime trends in Essex using a combination of exploratory analysis, predictive modelling, and natural language generation. The following discussion reflects on the effectiveness of the methods used, assesses their real-world relevance, and considers opportunities for improvement and further research.

Key Insights and Evaluation of Results

Spatial and temporal patterns –

The exploratory and modelling phases of this project uncovered several important spatial and temporal trends in crime across Essex. The EDA revealed that violence and sexual offences consistently made up the largest proportion of reported crimes, with over 200,000 violent and sexual offences report to Essex Police. This accounts for around 40% of total crime across the three-year period. This aligns with national trends but also highlights a persistent pressure point for Essex Police. Temporal analysis identified seasonal peaks in crime, particularly in March, July, and August, suggesting a link with school holidays, weather patterns, and social activity. These recurring patterns were later validated by the SARIMA model, which showed strong seasonal cycles and predicted moderate monthly fluctuations into 2025.

Spatially, the LSOA-level mapping revealed distinct geographic disparities in crime concentration, with areas such as Southend-on-Sea, Basildon, and Tendring emerging as persistent hotspots.

Clustering models –

The patterns identified spatially were further validated by the K-Means clustering, which grouped these high-incident zones into a single dominant cluster. The alignment between LSOA-level density and K-Means clustering reinforces the reliability of both methods and underlines the importance of regional socio-economic context in shaping crime distribution.

In contrast, the DBSCAN clustering approach added greater granularity by detecting micro-hotspots and isolating noise points. This was particularly useful for identifying less obvious pockets of concentrated crime, such as peripheral urban zones and rural fringes with isolated spikes. These findings suggest that broad administrative zones (e.g., districts) may underrepresent localised risk areas, and that point-based clustering methods can support more nuanced tactical decisions, such as targeted patrols or temporary enforcement zones.

Anomaly detection –

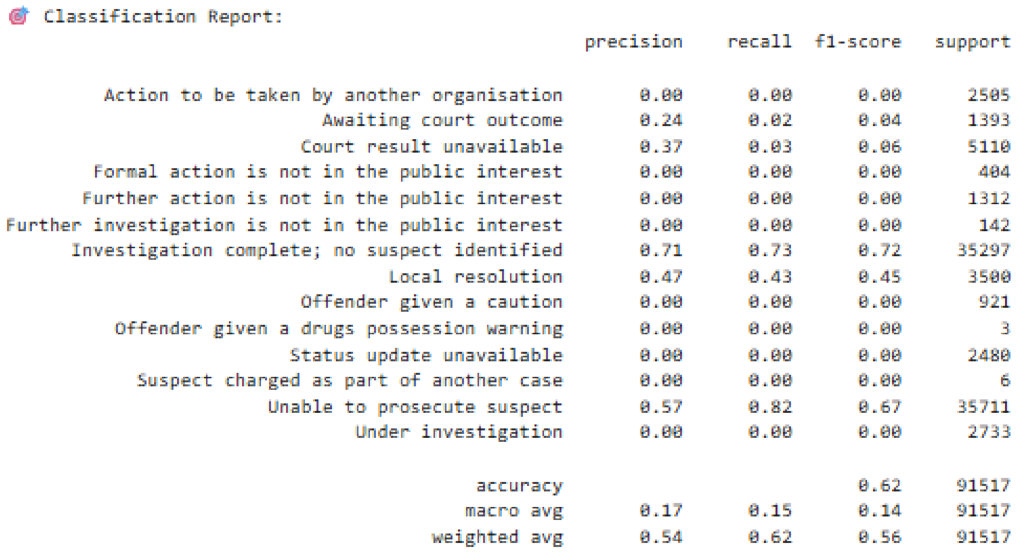
As shown in Figure 4T, the Random Forest model was able to detect 529 crime anomalies, with a recall rate of around 35%. This means it successfully identified about one-third of all unusual crime incidents. The model used features like the month, district, and crime type to make its predictions. When we looked into which features mattered most, Month and District came out as more important than Crime type. This suggests that crime anomalies are more often linked to when and where crimes happen, rather than the specific type of offence.

However, we noticed that a lot of these detected anomalies occurred along the edges of police districts. This wasn't because those areas necessarily had more unusual crimes, but rather because the model was only looking at data inside each district. It couldn't "see" what was happening just over the border. As a result, crimes near district boundaries sometimes looked more isolated than they really were, leading the model to mistakenly mark them as anomalies. This issue is common when models don't account for nearby areas outside the immediate boundary.

To improve this in the future, it might help to include surrounding areas when analysing a district, or to create a 'buffer zone' so the model has more context. Even though this version of the model had its limitations, it still shows how anomaly detection can help identify unusual patterns in crime, if care is taken to avoid false alarms caused by how the data is split or structured.

Random forest for crime outcome prediction –

Interest in crime data from the public comes from convictions outcome curiosity. When carrying out Random Forest model developed for predicting crime outcomes showed potential but also clear limitations. As shown in Figure 5A, while the overall accuracy reached 62%, closer inspection of the classification report and confusion matrix revealed that the model strongly favoured high-frequency classes such as "Investigation complete; no suspect identified" and "Unable to prosecute suspect", achieving F1-scores of 0.72 and 0.67, respectively. However, most other outcome categories were poorly predicted, with F1-scores close to 0.00, indicating that the model failed to generalise across less frequent classes. This imbalance is reflected in the low macro-average F1-score (0.14), suggesting that future work should address class imbalance more explicitly. Approaches such as class weighting could help achieve better distributional learning. Although the current model offers useful insights for common outcomes, further refinement is needed to make Random Forest a robust and equitable tool for predicting crime outcomes across the full spectrum of case types.



```
🎯 Classification Report:
                                                     precision    recall  f1-score   support

         Action to be taken by another organisation       0.00      0.00      0.00      2505
                             Awaiting court outcome       0.24      0.02      0.04      1393
                           Court result unavailable       0.37      0.03      0.06      5110
        Formal action is not in the public interest       0.00      0.00      0.00       404
       Further action is not in the public interest       0.00      0.00      0.00      1312
Further investigation is not in the public interest       0.00      0.00      0.00       142
      Investigation complete; no suspect identified       0.71      0.73      0.72     35297
                                   Local resolution       0.47      0.43      0.45      3500
                           Offender given a caution       0.00      0.00      0.00       921
          Offender given a drugs possession warning       0.00      0.00      0.00         3
                          Status update unavailable       0.00      0.00      0.00      2480
            Suspect charged as part of another case       0.00      0.00      0.00         6
                        Unable to prosecute suspect       0.57      0.82      0.67     35711
                                Under investigation       0.00      0.00      0.00      2733

                                           accuracy                           0.62     91517
                                          macro avg       0.17      0.15      0.14     91517
                                       weighted avg       0.54      0.62      0.56     91517
```

Figure 5A –Random Forest model developed for predicting crime outcomes and its classification report

Regression and socioeconomic relationships –

The analysis confirms that socioeconomic deprivation is strongly associated with higher crime levels across the Essex region. The regression results demonstrated a clear negative correlation between IMD decile and crime count, where more deprived areas (lower deciles) experienced significantly higher volumes of reported crime. With an R^2 value of 0.47, the model explains nearly half of the variation in crime based solely on deprivation, suggesting that structural inequality is a substantial driver of criminal activity. The statistically significant p-value (0.029) reinforces this finding. The regression coefficient indicates that each upward shift in affluence (e.g., one decile increase) corresponds to approximately 3,265 fewer crimes between 2022 and 2024.

These results align with broader theoretical models, such as Courson and Nettle (2021), who argue that inequality can fuel exploitative behaviour and erode social trust. In the context of Essex, where large pockets of deprivation exist in towns like Southend-on-Sea, Harwich, and Clacton, the findings underscore the need for targeted social and economic interventions alongside policing efforts. Moreover, the relationship between deprivation and crime highlights the importance of integrating place-based social data into crime forecasting and resource allocation models. While the current analysis is limited to linear regression and reported crime, future work could incorporate additional contextual variables, such as employment rates and educational attainment. As a result, it could help to build a more comprehensive understanding of the social determinants of crime.

RAG chatbot and NLP integration –

To help make crime analysis data more accessible, the implementation of a Retrieval-Augmented Generation (RAG) chatbot was create. Unlike traditional dashboards that rely on static charts, the RAG system allows users to ask natural language questions and receive dynamic, context-aware responses. This significantly lowers the barrier to insight extraction, especially for members of the public or frontline staff who may not have data science expertise. The chatbot draws from structured exploratory data summaries—including crime trends, hotspot rankings, and crime-type distributions, which are embedded into a searchable vector database. In evaluation, the system successfully returned relevant answers to queries such as “Where are the top hotspots?” or “How has crime changed over time?”, demonstrating its utility as a transparent and interpretable communication layer. While not yet connected to real-time data feeds, the chatbot presents a scalable foundation for future public-facing crime intelligence tools, with the potential to enhance engagement, awareness, and trust between police forces and the communities they serve.

Limitations and Ethical Considerations

API data limitations

The original intention was to utilise API access to the Police.uk database. However, challenges emerged due to the API’s leaky bucket rate-limiting system, which restricted large or frequent requests. As a result, I had to manually download monthly crime data files for Essex from 2022 to 2024 and combine them using data wrangling techniques. While this workaround did not compromise the quality of the analysis, a fully functional and scalable API connection would have eliminated the need for manual processing and enabled a more seamless, automated data pipeline.

Data and Boundary Issues -

Limitations also arose during the clustering and anomaly detection phases, particularly in the misclassification of crimes near district boundaries. In anomaly detection, because the analysis was restricted to administrative borders, the model lacked visibility of incidents in neighbouring districts, causing border crimes to appear isolated. Standardising coordinates within each district introduced spatial distortions, and naturally low activity in border areas (e.g., parks, rivers) further exaggerated this effect. Together, these factors led to false anomaly inflation near edges, highlighting the need for future models to incorporate cross-boundary spatial context.

Additionally, in the clustering analysis, the relationship between Essex and neighbouring counties (e.g., Suffolk, London, Cambridgeshire) was not captured, as cross-border crime is not recorded in the dataset. As a result, hotspots near county boundaries may be underrepresented, limiting the model's regional accuracy.

Model limitations -

Despite producing valuable insights, the models used in this project exhibited several limitations. The SARIMA model, while effective in capturing seasonal trends and short-term fluctuations, showed increasingly wide confidence intervals in future projections, which reduced its reliability for long-term forecasting. The Random Forest classifier used for anomaly detection was affected by class imbalance, as anomalous crime instances were underrepresented in the dataset. This likely impacted the model's recall and precision when identifying true anomalies. Lastly, the DBSCAN clustering algorithm was highly sensitive to its parameter settings. More specifically, the epsilon and minimum sample values, which required manual tuning. Small changes to these parameters significantly altered cluster shapes and counts, making consistent replication across regions more challenging without adaptive methods.

Ethical and practical risks

The use of predictive models in crime analysis introduces several important ethical and practical considerations. One key concern is the potential for bias embedded in training data to be reflected or amplified in model outputs, potentially resulting in discriminatory outcomes. Such bias can disproportionately affect certain communities and may further erode public trust in law enforcement if not carefully addressed. Additionally, false positives in anomaly detection can lead to misallocation of police resources or unfairly label certain areas or individuals as high risk, influencing both public perception and policy decisions. These risks highlight the need for transparency and explainability, especially in public-facing tools such as RAG-based chatbots. To build trust and ensure accountability, such systems must clearly communicate how responses are generated and what data sources are used. Future development should prioritise the integration of explainable AI (XAI) techniques to make decision-making processes interpretable and auditable, helping to mitigate unintended consequences and support ethical deployment.

Crime map dashboard and computing constraints

While deploying the dashboard through Render for public access, several limitations affected its functionality and scalability. Render's base tier offers limited memory and storage, which created issues when serving large data files and model embeddings generated during preprocessing. Attempts to load full JSON summaries or detailed Folium maps often led to timeouts or memory exhaustion at startup. As a result, the deployed version was scaled back to include only a single Folium map and a compressed Json file containing high-level insights.

More advanced dashboard features, such as multi-layer geospatial maps, dynamic visualisations, or real-time data interaction, were not feasible due to computing constraints. These would require greater processing power and memory, especially to maintain responsiveness under heavier usage. Overall, these challenges highlight the trade-off between public accessibility and computational resources and reinforce the need for more robust infrastructure in future deployments.

Recommendations for Practice and Future Work

This project has demonstrated the value of integrating exploratory data analysis, time-series forecasting, and clustering to develop a multi-dimensional understanding of crime dynamics in Essex. The findings not only validated established trends, such as the higher prevalence of crime in more deprived and urban areas, but also uncovered new patterns, including temporal surges during specific months and cross-boundary crime hotspots. These insights offer practical value for informing predictive policing, targeted resource deployment, and long-term strategic planning, not only for Essex Police but also for other forces aiming to identify similar trends. However, it is important to recognise that a "one-size-fits-all" approach may not be appropriate across all police jurisdictions. This was evident in the district-level analysis, where varying socioeconomic conditions significantly influenced crime patterns. Understanding these localised factors is crucial for designing effective and context-sensitive policy interventions.

Alongside advanced data science techniques, ranging from K-Means clustering to machine learning algorithms, there is a critical need to pair these tools with public transparency and accessibility. By integrating machine learning, geographic information systems (GIS), and natural language processing (NLP), this project bridged technical capability with real-world public service relevance. The development of a Retrieval-Augmented Generation (RAG) chatbot illustrates the potential to close the gap between police forces and the communities they serve, particularly for non-technical users seeking accessible and understandable crime insights. This aligns with the findings of Mason et al (2014) who showed that public access to performance data significantly enhances trust in policing institutions.

As mentioned previously, advancing effective policing strategies through data science methodologies, while also improving community trust and encouraging accurate crime reporting, requires a shift toward more robust and scalable data infrastructures. Specifically, the adoption of advanced databases with high-performance API access is crucial. Enhanced APIs would support real-time or near-real-time data retrieval, automate data integration pipelines, and enable dynamic updates to analytical systems, including forecasting models and public-facing tools like crime dashboards or RAG-based chatbots. Such infrastructure not only improves operational efficiency for analysts and police authorities but also supports greater transparency, responsiveness, and accountability to the public, ultimately reinforcing trust in policing institutions.

Conclusion and final reflections

Throughout this project, I have developed a strong foundation in several advanced data science techniques, including time-series forecasting with SARIMA, spatial clustering using K-Means and DBSCAN, and anomaly detection with Random Forest models. I also gained hands-on experience in developing natural language processing applications, particularly through the use of LangChain and vector databases to build a Retrieval-Augmented Generation chatbot. These technical capabilities were complemented by broader competencies in data cleaning, geospatial mapping, API integration, and deployment using cloud platforms such as Render as illustrated in Appendix C. Beyond the technical skills, perhaps the most valuable outcome has been the deeper appreciation I gained for the ethical implications and social responsibilities involved in applying predictive technologies within sensitive domains such as crime and public safety.

This project demonstrates the real-world utility of combining machine learning, geospatial analysis, and natural language interfaces to extract meaningful insights from publicly available crime data. By leveraging these techniques in tandem, the project was able to map spatial crime clusters, forecast future crime trends, and develop a user-friendly chatbot that translated complex analyses into accessible language. This holistic approach bridged the often-wide gap between technical modelling and community-facing application, revealing the power of data science not just for analytical discovery, but also for resident engagement and accountability.

Working with live crime data presented numerous challenges, including incomplete records, spatial distortions near district borders, and limitations in computing resources. Yet, these obstacles served as critical learning points. They highlighted the importance of transparency in algorithm design, the necessity of ethical safeguards, and the vital need for robust and interoperable data infrastructure in public institutions. The experience underscored that technological innovation alone is insufficient without an accompanying framework of ethical and operational oversight.

Importantly, the tools and methodologies explored here align with the UK Government's National AI Strategy (HM Government, 2021), which advocates for using artificial intelligence for public good. This project illustrates how that vision can be realised through transparent, localised, and accessible AI systems. By applying machine learning to crime forecasting and embedding the results within a public-facing RAG chatbot, the project offers a replicable model for ethical AI integration in regional policing. Unlike many international systems that function as opaque, internal-only tools, this solution places public communication and interpretability at its core. In doing so, it reflects a growing demand for AI systems that are not only intelligent, but also accountable and socially responsive.

Ultimately, this capstone project offers a blueprint for how data science can contribute to safer, more transparent communities. It lays the groundwork for future iterations that could incorporate real-time data feeds, broader geographic coverage, or expanded user interaction. The project also encourages future collaboration between technologists, public institutions, and local communities, illustrating the transformative potential of responsible AI in modern governance. As AI continues to evolve, projects like this provide a tangible path forward for applying it with fairness, clarity, and public value from evidence-based insights.

6 - References

Bowers, K.J. and Johnson, S.D., 2014. A critique of the design of predictive crime mapping experiments. *Environment and Planning A*, 46(1), pp.197–211.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., 2015. *Time series analysis: Forecasting and control*. 5th ed. Hoboken, NJ: Wiley.

Brantingham, P.J. and Brantingham, P.L., 1995. Criminality of place: Crime generators and crime attractors. *European Journal on Criminal Policy and Research*, 3(3), pp.5–26.

Chatfield, C., 2003. *The analysis of time series: An introduction*. 6th ed. Boca Raton, FL: Chapman and Hall/CRC.

Courson, K. and Nettle, D., 2021. Exploitative strategies in unequal societies: A computational model. *Evolution and Human Behaviour*, 42(2), pp.161–169.

De Courson, B. and Nettle, D., 2021. Why do inequality and deprivation produce high crime and low trust?. *Scientific Reports*, 11(1), p.1937.

End Violence Against Women Coalition (EVAW), 2024. *Inquiry finds major police failings to stop known perpetrator Wayne Couzens*. Available at: https://www.endviolenceagainstwomen.org.uk/inquiry-finds-major-police-failings-to-stop-known-perpetrator-wayne-couzens/ [Accessed 18th February 2025].

Essex County Council, 2024. *Greater Essex Trends 2024*. Available at: https://data.essex.gov.uk/dataset/e5lox/greater-essex-trends-2024 [Accessed 10th April 2025].

Essex Police, 2023. *Inside Essex Police's Battle Against Knife Violence*. Available at: https://science.police.uk/delivery/case-studies/inside-essex-polices-battle-against-knife-violence/ [Accessed 6th April 2025].

Heeks, M., Reed, S., Tafsiri, M. and Prince, S., 2018. *The economic and social costs of crime: Second edition*. Home Office Research Report 99. Available at: https://www.gov.uk/government/publications/the-economic-and-socialcosts-of-crime

HM Government, 2021. *UK National AI Strategy*. Department for Digital, Culture, Media & Sport. Available at: https://www.gov.uk/government/publications/national-ai-strategy [Accessed 10th May 2025].

Indices of Multiple Deprivation (IMD), 2019. [online] Available at: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [Accessed 10th April 2025].

Kurland, J. and Piquero, A.R., 2022. Understanding the spatial and temporal dynamics of crime in urban areas. *Applied Geography*, 142, p.102681. https://doi.org/10.1016/j.apgeog.2022.102681

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems*, 33, pp.9459–9474.

Lum, K. and Isaac, W., 2016. To predict and serve? *Significance*, 13(5), pp.14–19.

Mason, D., Hillenbrand, C. and Money, K., 2014. Are informed citizens more trusting? Transparency of performance data and trust towards a British police force. *Journal of Business Ethics*, 122, pp.321–341.

Office for National Statistics (ONS), 2024. *Crime in England and Wales: Year ending September 2024*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingseptember2024 [Accessed 12th February 2025].

Petrillo, A., De Felice, F., Cioffi, R. and Zomparelli, F., 2018. Fourth industrial revolution: Current practices, challenges, and opportunities. *IntechOpen*.

Saleh, R.A., Majzoub, S. and Saleh, A.M.E., 2025. *Fundamentals of Robust Machine Learning: Handling Outliers and Anomalies in Data Science*. Hoboken, NJ: John Wiley & Sons.

Spencer, D. and Tait, A., 2025. *A Portrait of Modern Britain: Crime and Closing the 'Toughness Gap'*. Policy Exchange. Available at: https://policyexchange.org.uk/wp-content/uploads/A-Portrait-of-Modern-Britain-Crime-and-closing-the-Toughness-Gap.pdf [Accessed 4th March 2025].

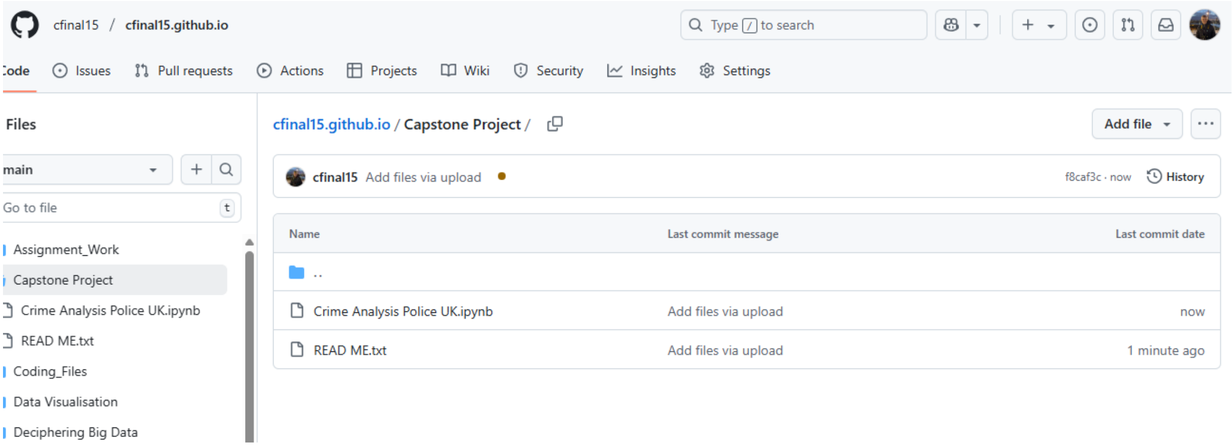
Wang, T., Rudin, C., Wagner, D. and Sevieri, R., 2013. Learning to detect patterns of crime. In: *Machine Learning and Knowledge Discovery in Databases*, pp.515–530.

Weisburd, D., Mastrofski, S.D., McNally, A.M., Greenspan, R. and Willis, J.J., 2003. Reforming to preserve: Compstat and strategic problem solving in American policing. *Criminology & Public Policy*, 2(3), pp.421–456.

YouGov, 2024. *How much confidence do Brits have in the police to deal with crime?* [online] Available at: https://yougov.co.uk/topics/politics/trackers/how-much-confidence-brits-have-in-police-to-deal-with-crime [Accessed 11th December 2024].

7 - Appendices

Appendix A – Capstone Project GitHub Repository with Jupyter Notebook containing Python code for all analysis and mapping (https://github.com/cfinal15/cfinal15.github.io/tree/main/Capstone%20Project)



- Assignment_Work
- Capstone Project
- Crime Analysis Police UK.ipynb
- READ ME.txt
- Coding_Files
- Data Visualisation
- Deciphering Big Data

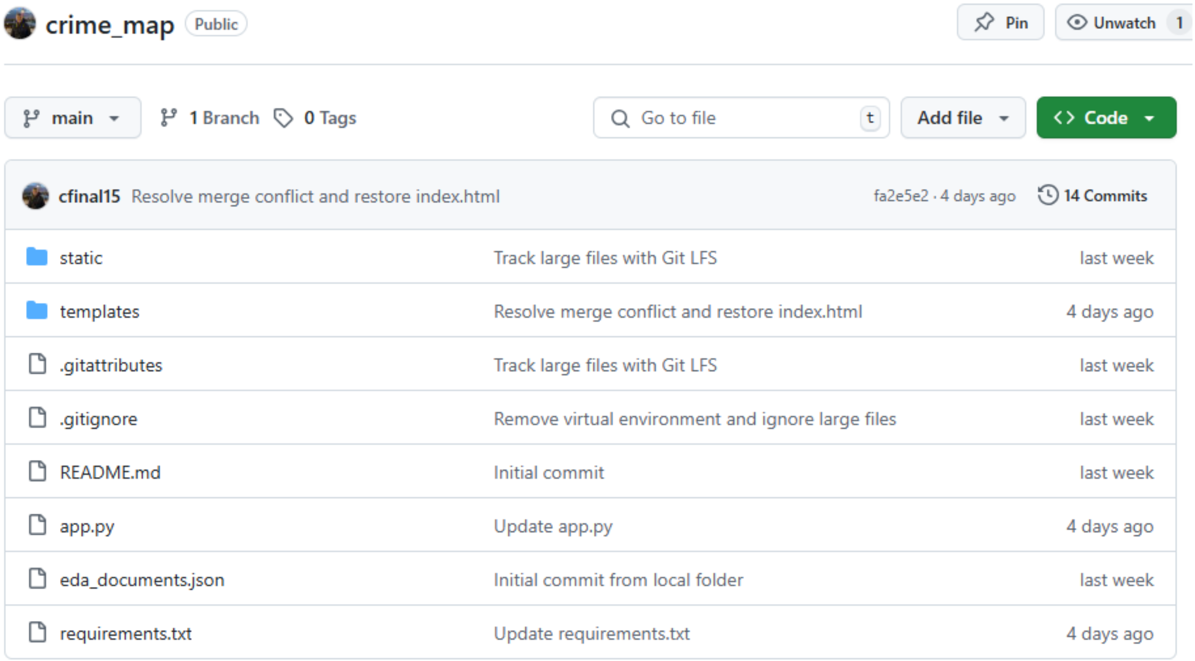
cfinal15.github.io / Capstone Project /

cfinal15 Add files via upload · f8caf3c · now

Name	Last commit message	Last commit date
..		
Crime Analysis Police UK.ipynb	Add files via upload	now
READ ME.txt	Add files via upload	1 minute ago

Appendix B – Dashboard and Chatbot Repository with python code for render application

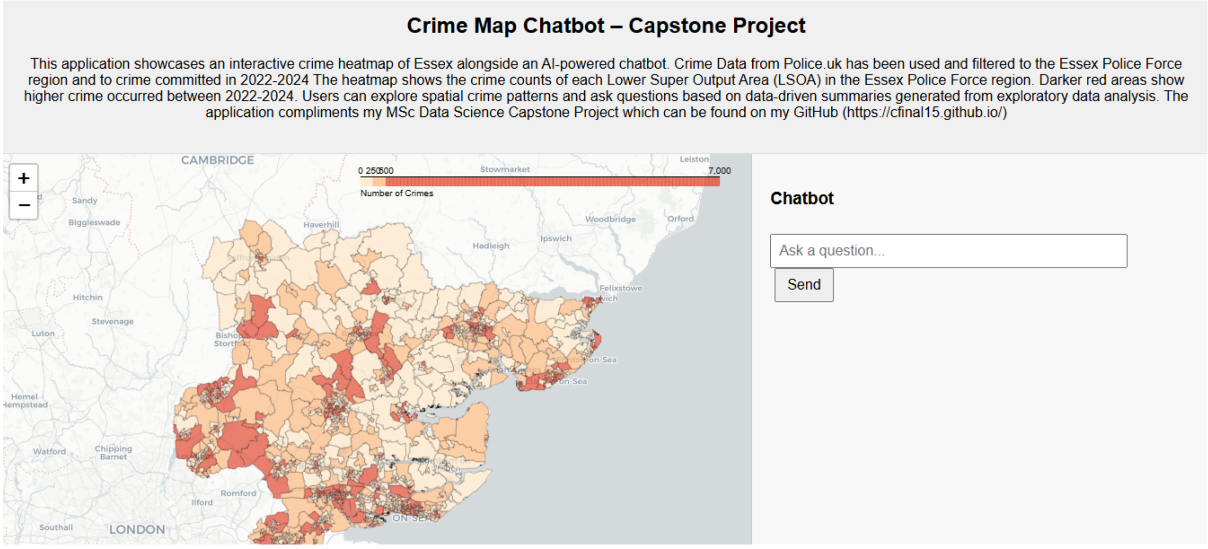
(https://github.com/cfinal15/crime_map)



cfinal15 Resolve merge conflict and restore index.html · fa2e5e2 · 4 days ago · 14 Commits

static	Track large files with Git LFS	last week
templates	Resolve merge conflict and restore index.html	4 days ago
.gitattributes	Track large files with Git LFS	last week
.gitignore	Remove virtual environment and ignore large files	last week
README.md	Initial commit	last week
app.py	Update app.py	4 days ago
eda_documents.json	Initial commit from local folder	last week
requirements.txt	Update requirements.txt	4 days ago

Appendix C – Live render site for Crime Dashbaord with chatbot (https://crime-map-chatbot.onrender.com/)



Crime Map Chatbot – Capstone Project

This application showcases an interactive crime heatmap of Essex alongside an AI-powered chatbot. Crime Data from Police.uk has been used and filtered to the Essex Police Force region and to crime committed in 2022-2024 The heatmap shows the crime counts of each Lower Super Output Area (LSOA) in the Essex Police Force region. Darker red areas show higher crime occurred between 2022-2024. Users can explore spatial crime patterns and ask questions based on data-driven summaries generated from exploratory data analysis. The application compliments my MSc Data Science Capstone Project which can be found on my GitHub (https://cfinal15.github.io/)

Chatbot