Analysing Crime Patterns and Classifying Crime Hotspots: A Temporal and Spatial Analysis Approach

Abstract

Crime pattern recognition is crucial for effective law enforcement and public safety management. In this study, we present a comprehensive framework for crime pattern recognition utilizing spatial analysis and forecasting techniques. The methodology consists of seven key steps: (1) Crime Data Collection, (2) Pre-processing including Data Transformation and Filtering, (3) Feature Selection, (4) Data Analysis incorporating Spatial Analysis, Prioritization of Crime Activities, and Forecasting Models, (5) Forecasting using ARIMA (Auto Regressive Integrated Moving Average) model for Regions of High, Low, and Moderate Crime rates, (6) Presentation of Forecasted Output, and (7) Visualization of Data using Graphs, Maps, and Charts.

Spatial analysis plays a pivotal role in understanding the geographical distribution of crime incidents. Utilizing geospatial libraries such as GeoPandas and Folium, we visualize crime data on maps, enabling law enforcement agencies to identify crime hotspots and allocate resources effectively. Additionally, we employ marker clustering and heatmap techniques to discern patterns and trends in crime activities. We demonstrate the effectiveness of forecasting models, particularly ARIMA, in predicting future crime rates based on historical data. By categorizing regions into high, low, and moderate crime zones, law enforcement agencies can devise targeted intervention strategies to mitigate criminal activities. The proposed framework is applied to real-world crime datasets, yielding valuable insights into crime patterns and trends. Through comprehensive data analysis and visualization, our approach empowers law enforcement agencies with actionable intelligence for proactive crime prevention and management.

Introduction:

Crime remains a pervasive societal challenge, posing significant threats to public safety and well-being. Effectively combating crime requires a deep understanding of its patterns and dynamics, enabling law enforcement agencies to devise proactive strategies for prevention and intervention. Crime pattern recognition, facilitated by advancements in data analytics and geospatial technologies, has emerged as a vital tool in this endeavour. By leveraging data-driven approaches, law enforcement agencies can identify crime hotspots, prioritize resource allocation, and forecast future crime trends.

In this context, we present a comprehensive Crime Pattern Recognition Project aimed at developing an integrated framework for analysing and forecasting crime patterns. Our methodology encompasses several key stages, beginning with the collection and pre-processing of crime data. We emphasize the importance of robust data transformation and filtering techniques to ensure data quality and consistency. Feature selection is performed to identify relevant variables that contribute to crime patterns. Leveraging advanced data analysis techniques, including spatial analysis and prioritization of crime activities, we gain insights into the spatial distribution and temporal dynamics of crime incidents. By harnessing the power of geospatial libraries such as GeoPandas and Folium, we visualize crime data on interactive maps, facilitating intuitive interpretation and decision-making.

A significant aspect of our project involves the application of forecasting models to predict future crime rates. We employ the Autoregressive Integrated Moving Average (ARIMA) and XGBoost model to forecast crime occurrences in regions characterized by high, low, and moderate crime rates. This predictive capability enables law enforcement agencies to anticipate emerging crime trends and allocate resources pre-emptively. Our project emphasizes the importance of data visualization as a means of communicating findings effectively. Through the use of graphs, maps, and charts, we present compelling visual representations of crime patterns and trends, enabling stakeholders to grasp complex information at a glance. The effectiveness of our proposed framework is demonstrated through its application to real-world crime datasets. By analysing historical crime data and forecasting future trends, we provide actionable insights that empower law enforcement agencies to formulate evidence-based strategies for crime prevention and management.

Literature Review

Paper Title	*Scope*	*Methodology*	*Accuracy*
Crime prediction based on crime types and using spatial and temporal crime hotspot	*Planning to apply more classifier to increase the accuracy of the model*	*Naïve bayes classifier, decision tree, apriori algorithm*	*51% in Denver ds, 54% in Los Angeles*
Crime hotspot detection using statistical and geospatial methods	*Stat scan can also be explored for wide range of areas*	*KDE, Getis-Order Gi statistics, SPTM*	*92%*
Temporal Crime Analysis Using KDE and ARIMA Models in the Indian Context	*With the help of these insights, regions with high levels of crime can be selected for intense observation as a preventative method for reducing crime rates*	*Geospatial analysis and virtualization*	*75%*
Crime analysis and prediction using data mining	*Our software predicts crime prone regions in India on a particular day. It will be more accurate if we consider a particular state/region*	*Naïve bayes, apriori algorithm, decision tree, NER, mongo db, Neo4j db, graph db*	*80%*
Crime Analysis Using Data Mining Techniques and Algorithms	*The future work is to use new tool to analyze and minimize the criminal activities*	*Data mining, Naïve bayes, predictive approach*	*90%*
Crime Analysis and Prediction using Optimized K-Means Algorithm	*the result of crime analysis can be used to make various strategies for crime control and the optimal deployment of resources in crime avoidance*	*Clustering, optimized K-means*	*Improved accuracy*
Crime Hotspot Prediction based on Dynamic spatial analysis	*Aims to use proactive approach as a crime prediction models can be used to predict crime rates and crime hotspots.*	*Linear regression, machine learning spatial analysis*	*Improved accuracy*
Analysing crimes of indian datasets based on machine learning methods	*Attempt to discover various factorrs affecting crimes in india.*	*KNN, Decision tree, Random forest*	*95.23%*
Spatio-temporal crime analysis using KDE and ARIMA model in indian context	*Prediction of crime prone areas*	*ARIMA model, KDE, predictive analysis*	*78%*
Crime analysis using K-means clustring	*Prediction of crime based on different data mining techniques*	*Cluster, rapid minier*	*Not mentioned*

A Geo-spatial approach for crime hot spot prediction	*Sparse matrx analysis spatial clustring method is used for crime prediction.*	*Sparse matrix*	*Not mentioned*
Crime analysis and hotspot prediction	*Aims to use the power of algorithm like RNN, STNN*	*RNN, STNN*	*Not mentioned*
Analysis of crime pattern using data mining techniques	*Aims to help the law enforsement agiencies*	*Data mining, RICIS system*	*Not mentioned*
Crime prediction and monitoring framework based on spatial analysis	*Aims to use web mapping and visualization based crime predictin tool which is built in R.*	*Crime analysis, web mapping, R, map visualization*	*Not mentioned*
Crime pattern analysis, visualization and predictin using data mining	*Aims to provide solution to provide enhanced process of crime nalysis*	*K-means, cluster, correlation*	*Not mentioned*
Identifying the appropriate spatial resolution for the analysis of crime patterns	*Aime to develop a general method that is capable of identifying the most appropriate spatial unit for the anlysis of spatial patterns*	*Clustring, R*	*Not mentioned*
Crime pattern detection, analysis and prediction	*Used the supervised and semi-supervised learning techniques for knowledge discoverie*	*K-means, Naive bayes, regression, apriori*	*Not mentioned*
Crime pattern detection using data mining	*We also developed a weighting scheme for attributes here to deal with limitations of various out of the box clustering tools and techniques*	*K-means, clustring, learning techniques*	*Not mentioned*
Crime analysis using k-means clustring	*purpose of this paper is to analyze the crime which entails theft, homicide and various drug offences which also include suspicious activities, noise complaints and burglar alarm by using qualitative and quantitative approach*	*K-means, statistical methods, data mining*	*Not mentioned*
An overview on crime prediction methods	*Our objective is to identify current implementations of crime prediction method and the possibility to enhance it for future needs*	*Crime analysis, prediction*	*Improved*

A comprehensive analysis of crime analysis using data miing techniques	*This paper illustrates about the techniques and discussed about the recent related works that can be used to perform crime analysis*	*Prediction, pattern identification*	*Not mentioned*
Tools and techniques implemented in crime dataset	*paper proposes the use of optimization data mining techniques for developing such a crime analysis tool.*	*R tools, data mining*	*Not mentioned*

A multitude of studies have delved into the intricate realm of crime hotspot prediction and analysis, employing an array of methodologies, and drawing from varied datasets. Some researchers have opted for dynamic spatial analysis techniques, allowing for the exploration of how crime hotspots evolve and shift over time. In contrast, others have homed in on specific crime types, leveraging spatial and temporal hotspots to enhance predictive accuracy. Furthermore, machine learning methods have been applied to analyse Indian crime datasets, showcasing the potential of advanced algorithms in discerning complex patterns within crime data. Additionally, statistical, and geospatial approaches have been utilized to detect crime hotspots, shedding light on the spatial clustering of criminal activities, and enabling more targeted law enforcement strategies. Temporal crime analysis, facilitated by techniques like Kernel Density Estimation (KDE) and ARIMA models, has offered insights into the temporal dynamics of crime occurrences, allowing for better anticipation of future trends. Geo-statistical methods and clustering techniques have also been instrumental in hotspot prediction and crime pattern analysis, providing valuable tools for understanding spatial relationships and identifying crime clusters. Moreover, studies have explored the impact of spatial resolution on the accuracy of crime pattern analysis, recognizing the importance of fine-tuning analytical parameters for optimal results. Some researchers have provided comprehensive overviews of crime prediction methods, synthesized various approaches, and highlighted their respective strengths and limitations. Finally, the analysis of criminal spatial events in Geographic Information Systems (GIS) has emerged as a promising avenue for hotspot prediction, leveraging specific crime data to identify geographic areas prone to heightened criminal activity. Collectively, these studies contribute to a nuanced understanding of crime pattern recognition and prediction, showcasing the diversity of methodologies employed and the breadth of insights gained across different contexts.

Architecture Diagram

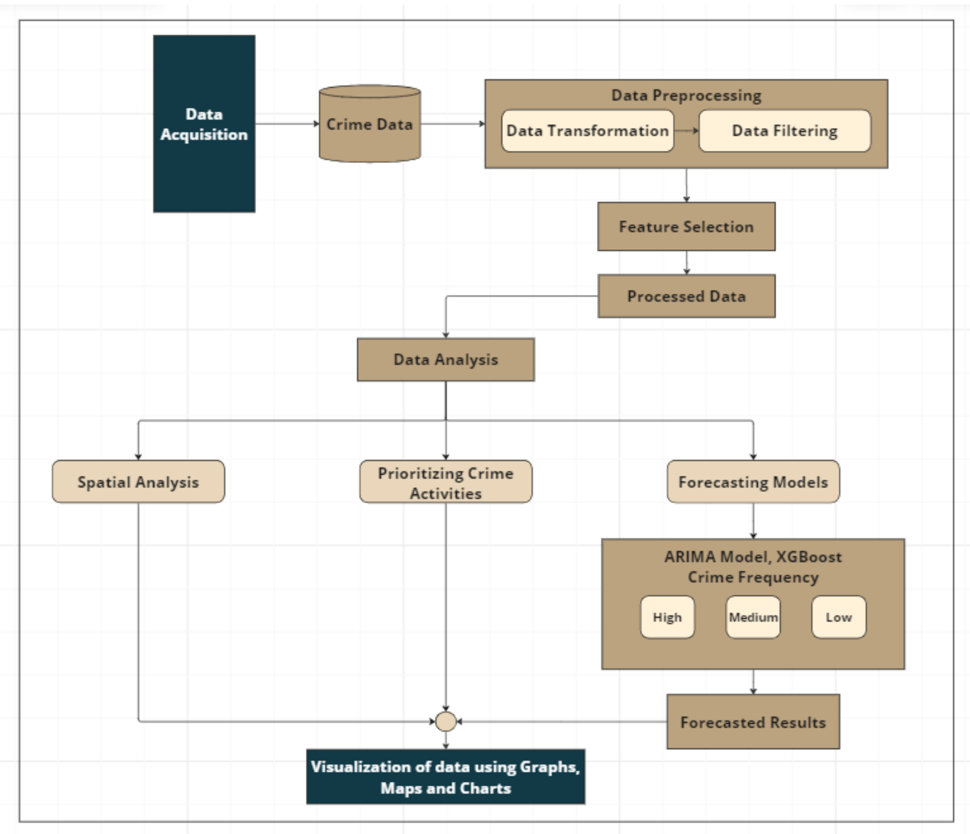


Figure 1. Architecture diagram

The diagram represents a systematic approach to crime pattern recognition, illustrating the various steps involved from data collection to the visualization of crime analysis results. Here is a breakdown of the flow diagram. The process starts with collecting crime data. This data could come from various sources such as police reports, public records, or other relevant databases. Data Transformation step involves converting the raw crime data into a suitable format for analysis. It might include normalizing the data, handling missing values, or encoding categorical variables. Data filtering includes irrelevant or redundant information from the dataset to ensure that the data is clean and relevant for further analysis. The feature selection step involving the most significant variables or features in the dataset that will be used for the analysis. Feature selection helps reduce the complexity of the data and improves the performance of the analysis. Processed data results with data preprocessing and feature selection create a refined dataset that is ready for analysis.

Data Analysis is the core step where various analytical techniques are applied to the processed data to uncover patterns and insights related to crime activities. Spatial analysis involves examining the geographical distribution of crime incidents. It helps in identifying crime hotspots and understanding the spatial patterns of crime. Prioritizing Crime Activities based on the data analysis and certain crime activities are prioritized. This step may involve ranking crime types or incidents based on their frequency, severity, or other criteria. Forecasting models are developed to predict future crime trends. These models help in anticipating potential increases in crime rates and preparing accordingly. The ARIMA (Auto-Regressive Integrated Moving Average) model and XGBoost model defines a specific type of forecasting method used to predict the frequency of crimes over time. It categorizes crime frequency into High, Medium, and Low. Forecasted Output of the forecasting models provides predictions about future crime patterns. This information is critical for law enforcement and policymaking. Visualization of Data using Graphs, Maps, and Charts results of the analysis and forecasting are visualized using various tools like graphs, maps, and charts. This step helps in communicating the findings effectively to stakeholders, allowing for better decision-making and resource allocation.

In summary, the diagram outlines a comprehensive workflow for crime pattern recognition, starting from data collection, moving through preprocessing and analysis, and culminating in the visualization of insights. This process enables authorities to understand crime patterns better, prioritize resources, and develop strategies to mitigate crime effectively.

Data Collection and Preprocessing

We collected the dataset from National Crime Records Bureau (NCRB) website, we have considered all over India crime rates so where we expected a dataset to be at least containing 750 rows, as we got to know that, including the territories, we have 806 districts all over India, as no crime dataset contains all city names. We have collected the data from 2017 to 2022 where our data includes minimum 720 rows of districts and it also takes all the union territories of India into consideration. As our Architecture Diagram includes three phases Spatial Analysis (1), Prioritization of Crime Events (2) and Forecasting the crime events (3) for future endeavours. For the forecasting, it is obvious that it includes time series data, where at least 6 datasets will be better for the model training purpose. In the dataset which we have collected, it includes 933 rows and 145 columns, of which 12 are main crime columns and those crimes also include sub-columns. In terms of rows, it includes ludes state names above each state district row, a total for each state and some miscellaneous rows too. The miscellaneous rows are Narcotics, Bureau of Investigation, Intelligence Wing, NRI Wing, Special Task Force, Railway Police, SCRB, SOB, SSG, CID, GRP, BIEO and Unnamed city North and South crime rows, etc.,

S. No	State/UT/District	Murder (Sec.302 IPC)	Culpable Homicide not amounting to Murder (Sec.304 IPC)	Causing Death by Negligence							
					Deaths due to Negligence relating to Road Accidents						
				Causing Death by Negligence (Sec.304-A IPC) (Col.6+Col 9 to 12)	Deaths due to Negligence relating to Road Accidents (Total) (Col.7+Col.8)	Hit and Run	Other Accidents (other than Hit and Run)	Deaths due to Negligence relating to Rail Accidents	Deaths due to Medical Negligence	Deaths due to Negligence of Civic Bodies	Deaths due to other Negligence
1	2	3	4	5	6	7	8	9	10	11	12
State: Andhra Pradesh											
1	Anantapur	113	4	569	553	50	503	0	0	0	16
2	Chittoor	70	5	529	497	36	461	0	0	0	32
3	Cuddapah	88	10	487	469	32	437	0	0	0	18
4	East Godavari	69	14	664	641	138	503	0	0	0	23
5	Guntakal Railway	11	0	4	0	0	0	0	0	0	4
6	Guntur	100	6	611	595	62	533	0	0	0	16
7	Guntur Urban	32	9	266	266	3	263	0	0	0	0
8	Krishna	35	8	370	367	35	332	0	0	0	3
9	Kurnool	93	22	597	568	41	527	0	1	0	28
10	Nellore	67	10	494	475	52	423	0	0	0	19
11	Prakasham	91	46	485	485	33	452	0	0	0	0

Figure 2. Raw dataset

In terms of preprocessing, it was done manually using MS Excel platform. We combined all 145 columns to 12 main columns where these columns represent the crimes which are Districts name, Causing Death by Negligence, Hurt, Assault on Women with Intent to Outrage her Modesty, Kidnapping and Abduction, Rioting, Offences promoting enmity between different groups, Theft, Burglary, Dacoity, Counterfeiting, Forgery Cheating and Fraud, Rash Driving on Public Way. We combined the sub columns of these main columns into one by considering the average of the sub columns. These sub columns also included the combining of 2 to 3 columns where we excluded those. After combining all the sub columns to the main frame, we converted the data to Comma Delimited (CSV file Format) for better implementation purpose.

S. No	State/UT/District												
1	2	Causing Death by Neg	Hurt	Assault or	Kidnapping and Abduction	Rioting (Sec.147	Offences prc	Theft	Burglery	Dacoity	Counter fe	Forgery Cl	Rash Driving on Public Way
State: Andhra Pradesh													
1	Anantapur	211.375	174.1667	52.125	9.555555556	4.352941176	2.5	303	183	3	2	65	967
2	Chittoor	194.375	72.25	21	1.555555556	2.647058824	0.5	274	71	4	1	27	162
3	Cuddapah	180.375	207.8333	61.875	4.111111111	1.941176471	0.5	347	156	0	1	47	3398
4	East Godavari	246.125	86.08333	63.375	4.888888889	0.470588235	3	640	248	1	1	55	1197
5	Guntakal Railway	1	1.916667	0.375	0	0	0	445	0	0	1	0	0
6	Guntur	227.125	142.5	41	10.11111111	2.529411765	0	439	136	0	1	66	3014
7	Guntur Urban	99.75	48.83333	19.75	13.88888889	0.352941176	1.5	529	133	1	0	66	338
8	Krishna	138.375	130.5	55.125	11.77777778	1.411764706	4.5	291	175	0	1	52	458
9	Kurnool	220.25	107.75	42.375	6	2.176470588	2	326	196	2	2	69	586
10	Nellore	182.875	131.6667	37	5.333333333	2.823529412	0	555	251	7	0	40	2792
11	Prakasham	181.875	122.5	46.625	6	3.647058824	0	345	181	5	1	28	546
12	Rajahmundry	60.5	37.41667	12.25	1.666666667	0.294117647	0.5	321	89	2	1	11	233
13	Srikakulam	114	73.58333	12.125	1.222222222	0.705882353	0	68	61	0	0	20	483
14	Tirupathi Urban	94.75	35.08333	10.375	3.444444444	4.058823529	0	391	152	3	2	46	937
15	Vijayawada City	133.5	53.25	34.375	3.888888889	0.117647059	3.5	1153	205	2	0	97	652
16	Vijayawada Railway	0	2.333333	0.875	0.111111111	0.058823529	0	750	1	1	1	0	2
17	Visakha Rural	124.5	33.41667	18.5	2	1.235294118	0	87	77	0	0	14	384
18	Visakhapatnam	126.875	88.41667	39.25	17	0.294117647	0	897	251	3	1	119	432
19	Vizianagaram	165.875	117.3333	12.75	1.666666667	0.529411765	0	127	75	1	0	18	1968
20	West Godavari	250.5	147.75	60	8.888888889	0.352941176	9.5	580	285	3	0	72	664
State: Arunachal Pradesh													

Figure 3. Dataset after partial pre-processing

After combining all the sub columns into one main column using the Average and Round method in MS Excel where crimes should be considered in whole values. We also removed the total values of each state which was not needed in our consideration. We combined the miscellaneous which represents above 60 crime rate valued rows for those which are represented with the district name in the dataset like New Delhi etc., to the unknown districts we added these rows previous district present in the dataset which results below 60 crime rate value in all 6-year datasets, as it will also be considered to our view where if the value is higher and if it is added to the city's crime rate, it may result in enormous changes when it is visualized in spatial analysis'. We also got to know that few districts were divided into 2 or more rows including the north, south, east and west parts of the districts, and some districts were also having 2 or more rows including rural, urban, city parts of the districts in the dataset where all those were combined into 1 single row for easy identification in the visualization parts. We also manually included two more columns Latitude and Longitude in all 6-year datasets where for the Geospatial visualization, these columns were needed for several purposes for marking each district in Geospatial maps. We also create done more dataset which includes every state name and its latitude and longitude values.

Districts	Latitude	Longitude	Causing Death by	Hurt	Assault on	Kidnappin	Rioting	Offences	Theft	Burglery	Dacoity	Counter feiting	Forgery Cl	Rash Driving on Public Way	Target
Anantapur	14.7899	77.5985	211	174	52	10	4	3	303	183	3	2	65	967	1977
Chittoor	13.2257	79.0909	194	72	21	2	3	1	274	71	4	1	27	162	832
Cuddapah	14.5621	78.826	180	208	62	4	2	1	347	156	0	1	47	3398	4406
East Godavari	17.4663	81.8329	246	86	63	5	0	3	640	248	1	1	55	1197	2545
Guntakal	15.1707	77.38	1	2	0	0	0	0	445	0	0	1	0	0	449
Guntur	16.3096	80.4298	327	192	61	24	3	2	968	269	1	1	132	3352	5332
Krishna	16.4826	80.9351	138	131	55	12	1	5	291	175	0	1	52	458	1319
Kurnool	15.8317	78.0392	220	108	42	6	2	2	326	196	2	2	69	586	1561
Nellore	14.668	79.9639	183	132	37	5	3	0	555	251	7	0	40	2792	4005
Prakasham	15.5875	79.4813	182	123	47	6	4	0	345	181	5	1	28	546	1468
Rajahmundry	17.0057	81.8083	61	37	12	2	0	1	321	89	2	1	11	233	770
Srikakulam	18.2953	83.8975	114	74	12	1	1	0	68	61	0	0	20	483	834
Tirupathi	13.7238	79.3865	95	35	10	3	4	0	391	152	3	2	46	937	1678
Vijayawada	16.5772	80.628	134	53	35	4	0	4	1903	206	3	1	97	654	3094
Visakhapatnam	17.5931	83.2048	252	121	58	19	1	0	974	328	3	1	133	816	2706
Vizianagaram	18.1107	83.3969	166	117	13	2	1	0	127	75	1	0	18	1968	2488
West Godavari	16.8573	81.4286	251	148	60	9	0	10	580	285	3	0	72	664	2082
Anjaw	28.0611	96.8317	0	1	0	0	0	0	3	0	0	0	0	0	4
Changlang	27.3979	96.2567	3	1	0	1	0	0	17	11	1	0	1	2	37
Dibang Valley	28.8688	95.8998	0	0	0	0	0	0	1	1	0	0	0	0	2
Kameng East	27.5868	92.9345	1	3	1	0	0	0	13	7	0	0	1	1	27
Kameng West	27.3154	92.4033	7	0	1	0	0	0	11	15	0	0	1	7	42
Kurung Kumey	28.0126	93.2206	0	1	0	0	0	0	5	1	0	0	0	0	7

Figure 4. Fully pre-processed dataset

Exploratory Data Analysis

The next phase in our methodology is the Exploratory Data Analysis (EDA), which aims to uncover initial insights, detect anomalies, and visualize patterns in the crime data. For this study, the crime data for various districts was loaded into a Tableau application, facilitating efficient data manipulation and analysis. The dataset comprises multiple years, allowing for a temporal analysis of crime trends.

To begin, we inspected the dataset using the column header in Tableau to review the first few records and understand the structure of the data. This preliminary inspection confirmed the presence of essential variables such as 'Districts' and 'Target' (representing crime rates or counts).

We then proceeded to visualize the distribution of crime incidents across different districts using bar plots. Initially, a basic bar plot was generated to display the 'Target' variable across 'Districts', providing a straightforward comparison of crime rates between districts. This visualization was enhanced by customizing the plot with labels for the x-axis (Districts) and y-axis (Target), and a title to describe the content of the graph.

To improve the clarity and readability of the visualization, we resized the plot, significantly increasing its width and height. This adjustment allowed for a more detailed examination of the differences in crime rates across districts. By iterating this process for data from different years, we could identify temporal trends and changes in crime patterns, thereby gaining a deeper understanding of how crime rates evolved over time.

These visualizations revealed key insights into the geographical distribution of crime, highlighting districts with notably high or low crime rates. The temporal analysis further provided evidence of trends, such as increasing or decreasing crime rates in specific districts over the years. These findings are critical for law enforcement agencies to strategically allocate resources and implement targeted crime prevention measures.

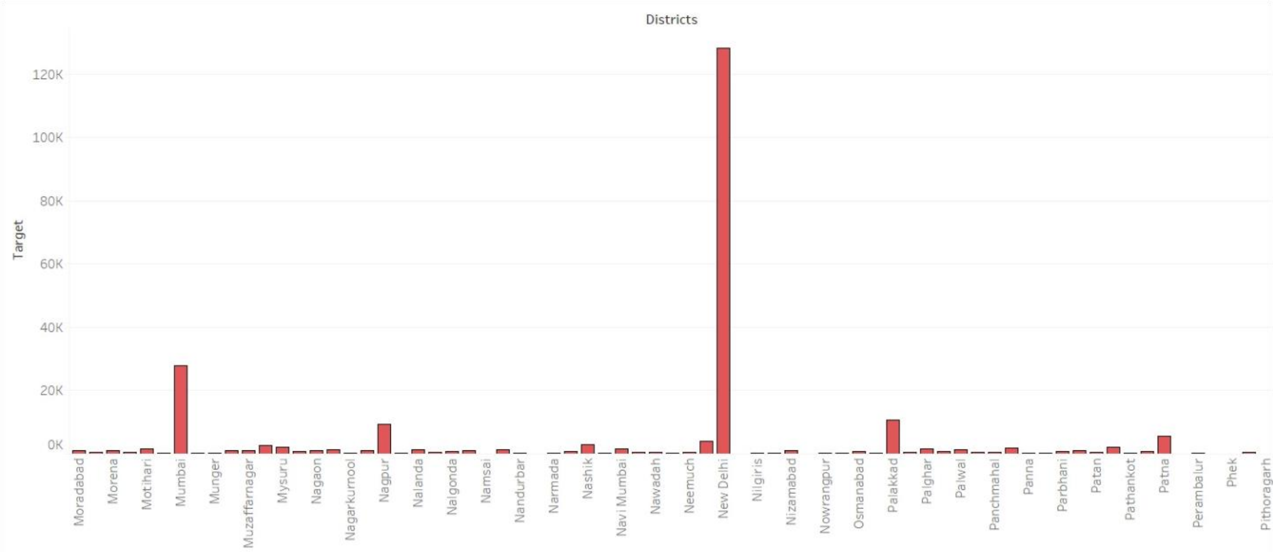


Figure 5. Bar chart representation of target values for district (Year 2017)

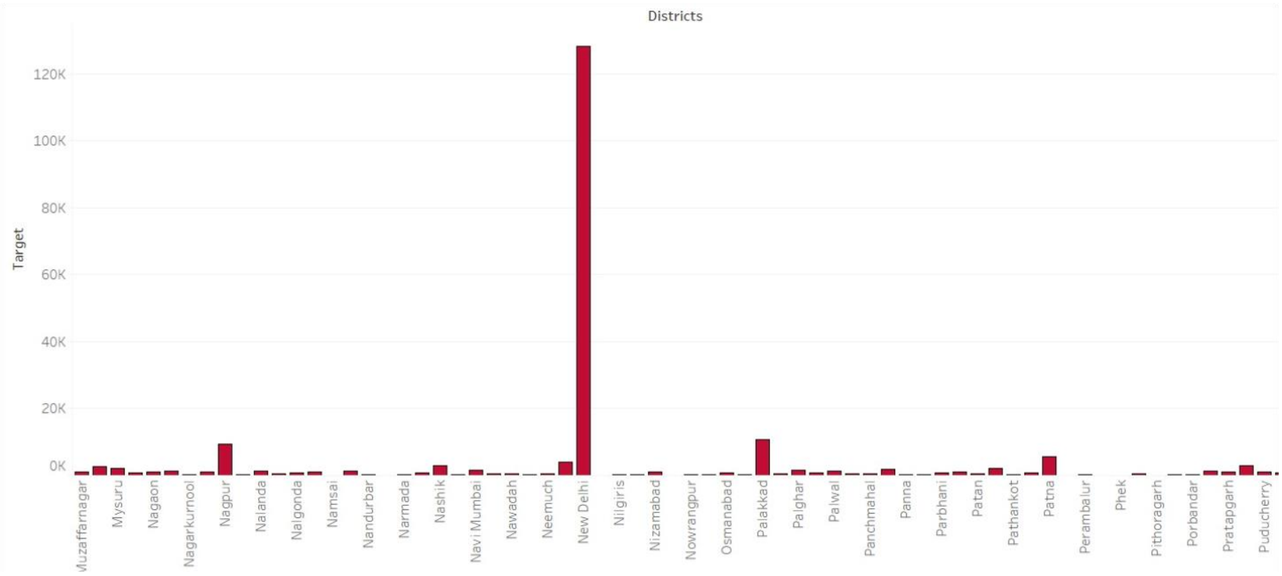


Figure 6. Bar chart representation of target values for district (Year 2018)

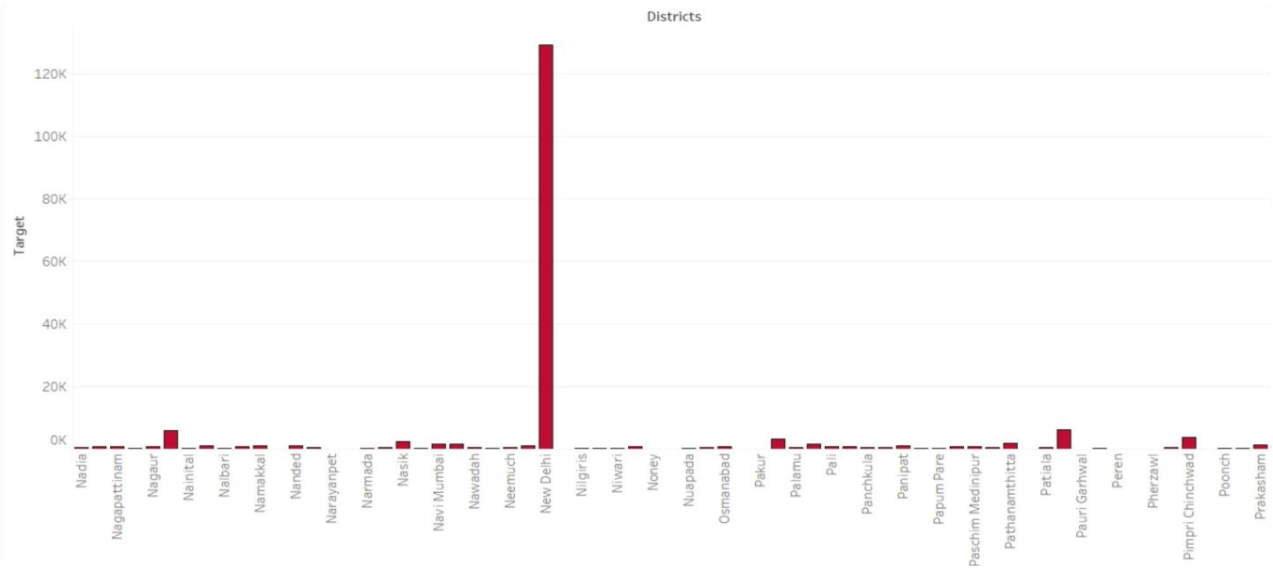


Figure 7. Bar chart representation of target values for district (Year 2019)

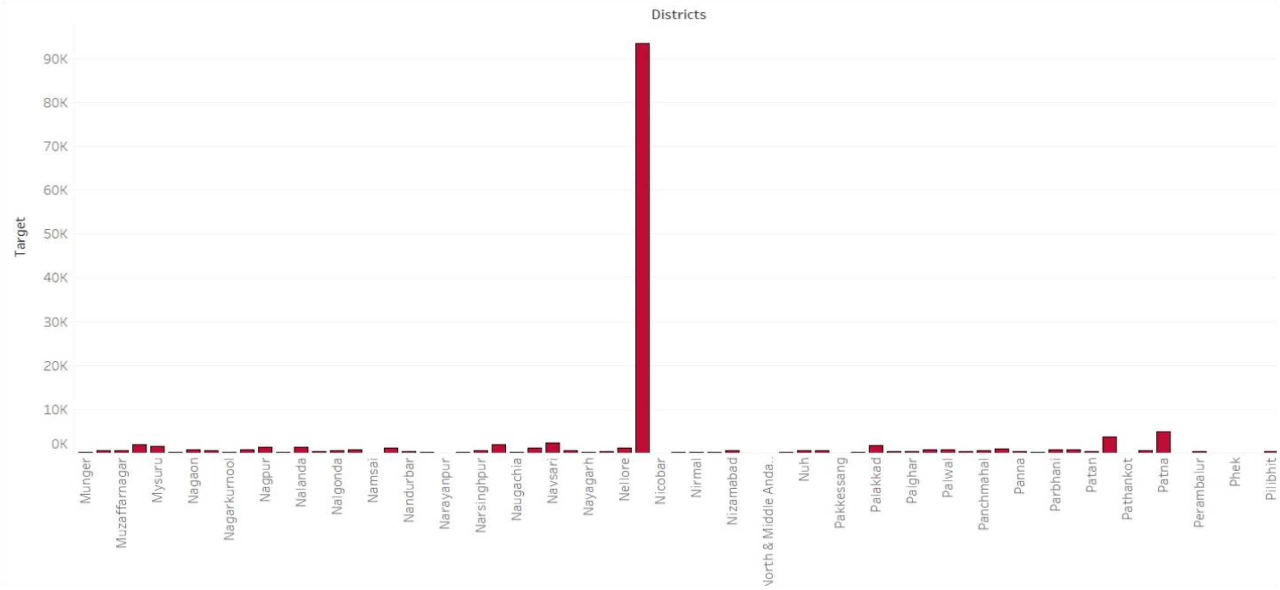
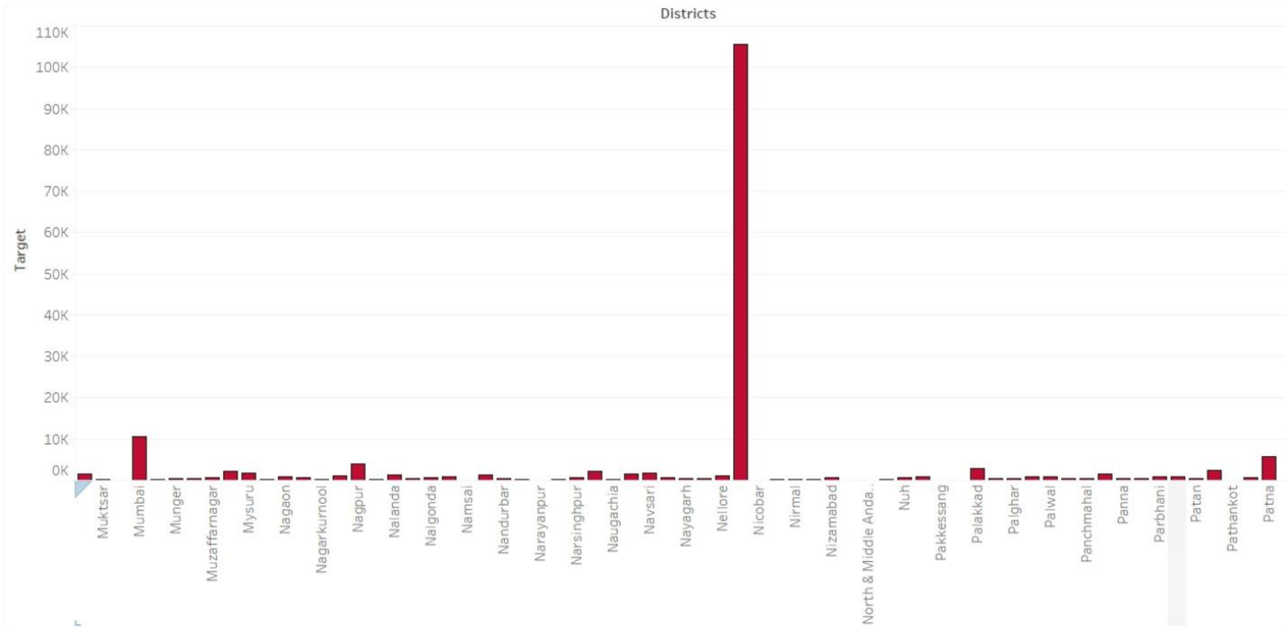


Figure 8. Bar chart representation of target values for district (Year 2020)



Figure

Figure 9. Bar chart representation of target values for district (Year 2021)

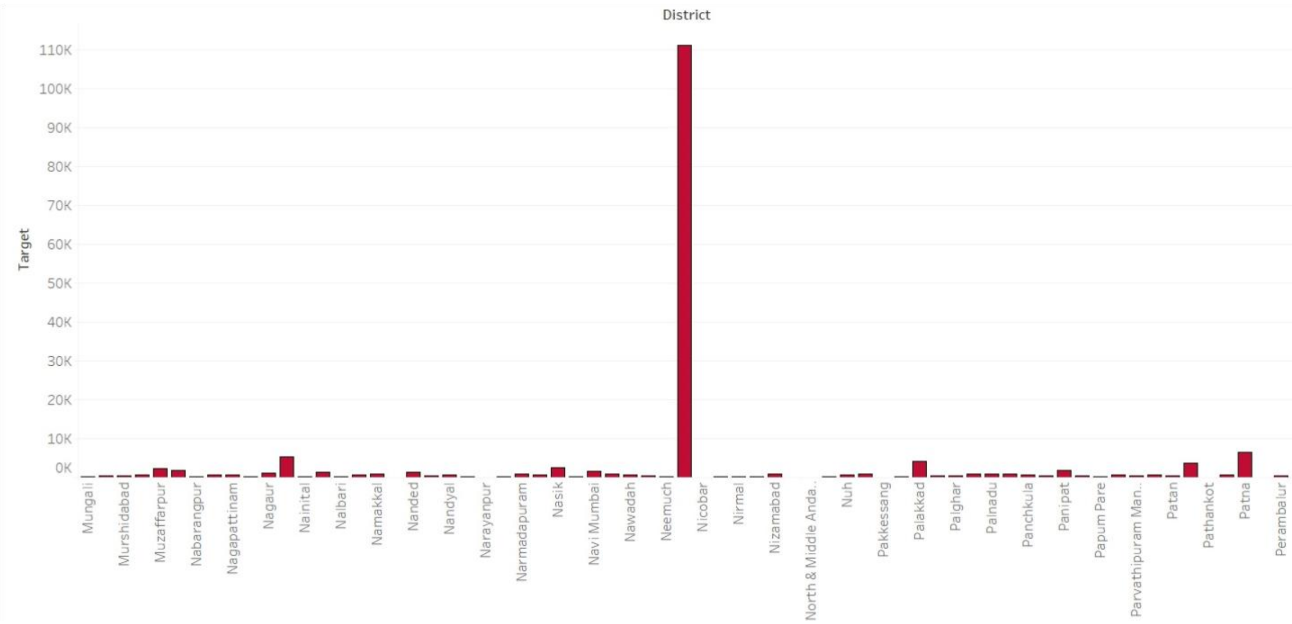


Figure 10. Bar chart representation of target values for district (Year 2022)

Spatial Analysis

Spatial analysis is integral to understanding the geographical distribution and dynamics of crime incidents. By leveraging spatial data, law enforcement agencies can identify crime hotspots, discern patterns, and allocate resources more effectively. In this study, we employ geospatial libraries such as GeoPandas, Folium, and Plotly to facilitate the visualization and analysis of spatial data. GeoPandas extends the data structures of pandas to allow for the manipulation of geometric data types, making it suitable for spatial operations and analysis. Folium is used for creating interactive maps, while Plotly enhances data visualization with interactive plots and charts. GeoPandas simplifies working with geospatial data in Python, integrating seamlessly with other libraries like pandas and shapely. It enables the reading, writing, and manipulation of geospatial data, allowing for efficient spatial operations such as buffering, spatial joins, and aggregations. In our analysis, we convert crime data into a GeoDataFrame for enhanced spatial processing and visualization. This conversion allows us to perform sophisticated spatial operations and integrate various geospatial datasets for comprehensive analysis.

Folium is a powerful library for visualizing geospatial data, creating interactive maps that can be easily embedded in web applications or Jupyter notebooks. Folium supports various mapping features, including tile layers, marker clusters, and heatmaps, which are crucial for identifying crime hotspots. For instance, we initialize the map centered on the geographic area of interest and use marker clustering to manage the visualization of numerous crime incidents. This technique groups nearby incidents into clusters, simplifying the map and preventing overplotting. As the map is zoomed in, clusters break apart to reveal individual incidents, offering a detailed view at different zoom levels. Identifying crime hotspots is fundamental for effective law enforcement. This approach enables law enforcement agencies to focus their resources on areas with the highest crime, thereby improving the efficiency and effectiveness of their interventions.

To provide an alternative visualization of crime density, we use Plotly's density mapbox. This method visualizes crime density with a continuous colour scale, offering an intuitive understanding of crime intensity across different regions. Plotly’s interactive capabilities allow users to explore the data dynamically, adjusting views and gaining deeper insights into the spatial distribution of crimes. By analysing spatial patterns, we gain insights into the underlying causes and trends of criminal activities. This analysis provides a quantitative basis for understanding the spatial characteristics of crime, enabling more informed decision-making.

Spatial analysis not only identifies hotspots but also aids in prioritizing crime activities based on severity and frequency. By combining spatial data with temporal trends, law enforcement agencies can focus on high-risk areas and time periods, optimizing patrol routes and intervention strategies. This targeted approach enhances the ability of law enforcement to prevent crime and improve public safety.

The folium package which is used for the analysis required latitude and longitude values in the data where we tried to create the data in a JSON file format. The latitude and longitude values of states and districts which were considered in the data was created separately and also these longitude and latitude data were included in the crime data too. These data were helped in marking the cities which were included th the dataset using the markers in folium package. The latitude and longitude values of each States and Union Territories of India was also needed in order to get the good insights from the python folium package.

We have used several methods in Folium package such as Map, Markercluster, Marker, Plugins and GeoJson. Folium's MarkerCluster, plugins, GeoJson, and marker methods are essential tools for creating detailed and interactive maps. These methods and plugins enhance the visualization and analysis of geospatial data in various applications. The MarkerCluster method is particularly useful for handling large datasets of geographical points. When numerous markers are plotted on a map, it can become cluttered and difficult to interpret. MarkerCluster groups these markers into clusters based on their proximity, which dynamically adjusts as the user zooms in and out of the map. The GeoJSON method in the Folium package enables the integration of complex geographical boundaries and features into interactive maps. This feature not only improves map readability but also enhances performance by reducing the number of individual markers displayed simultaneously. This capability is essential for spatial analysis, providing context for the data and facilitating the identification of patterns and trends within specific geographic areas. As users zoom in, clusters break apart to reveal the individual markers, allowing for detailed inspection of densely populated areas. This method is ideal for applications like crime mapping, where it's crucial to visualize high-density point data clearly.

Utilizing the MarkerCluster, plugins, marker, and GeoJson methods from the Folium package offers profound benefits for spatial analysis, particularly in the realm of crime pattern recognition. The MarkerCluster method enhances map readability and performance by aggregating nearby crime incidents into clusters, reducing visual clutter and allowing users to discern patterns more easily. As users zoom in, clusters decompose into individual markers, facilitating a detailed examination of crime distributions. Folium's plugins extend its functionality with advanced tools such as HeatMap for identifying hotspots and GeoJson for visualizing temporal changes in crime data. These plugins enable comprehensive trend analysis and provide insights into the effectiveness of interventions over time. The marker methods allow for precise placement and customization of individual crime incidents on the map, with interactive popups and tooltips offering detailed information like the type of crimes.

For Heatmap visualization, we have considered plotly package, where it benefitted us in several ways for the visualization purpose like dropdown box containing District name, latitude and longitude values and the Target value which is considered for the Heatmap visualization. The major feature which benefited us by using Plotly heatmaps were their ability to represent data density and multiple variations through colour gradients, and it was easier to handle large datasets efficiently, ensuring smooth interaction even with extensive data points. Making it easier to identify hotspots and anomalies at a glance. This visual representation is particularly effective for large datasets, where traditional charts may fail to highlight subtle patterns. The usage of plotly package, the darker intensity of the colours in the map helped us to identify the highest crime occurred city, where we were able to differentiate cities which were popping up with red spots in the heatmap where from all of those red spots we could able to identify New Delhi consisting of a greater number of crimes in all 6 years data with more that 1 lakh crime counts in total.

With the help of Folium package, we tried to plot the Geospatial map using markers which was integrated with MarkerCluster method. The package helped us to mark the cities with different range of crimes with 3 distinct colours namely green, orange and red, and black colour for the outlier. In the heatmap we plotted only with New Delhi district, we faced an issue was it was consisting of more than 1 lakh crime rates, and no other districts were competent to that district, it was required to cross check once again with the machine for identification of crimes with the same value seen with New Delhi. It allowed us to define our own function for the separate set of colours for separate set of ranges in the crime. We considered below 500 crime rates to be in green, above from 500 to 2000 orange colour was given and from 2000 and above the red spots will be visible, where those were the cities which were considered as the crime hotspots from our datasets. New Delhi was alone coloured with black as we got to know with our heatmap visualization which was considered as the outlier from the plotted map.

Figure 11. Map visualization with state borders

Figure 12. Map visualization including MapCluster

Figure 13. Map visualization including MapCluster (2nd zoom level)

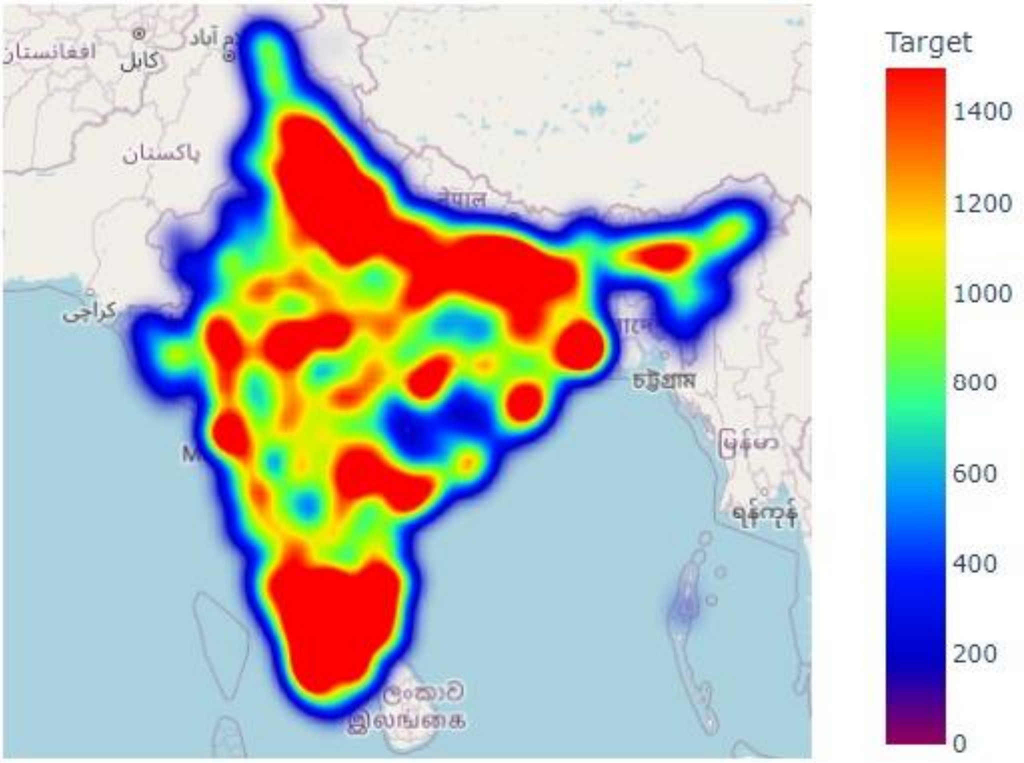


Figure 14. Plotly heat map visualization for the values of attribute "Target"

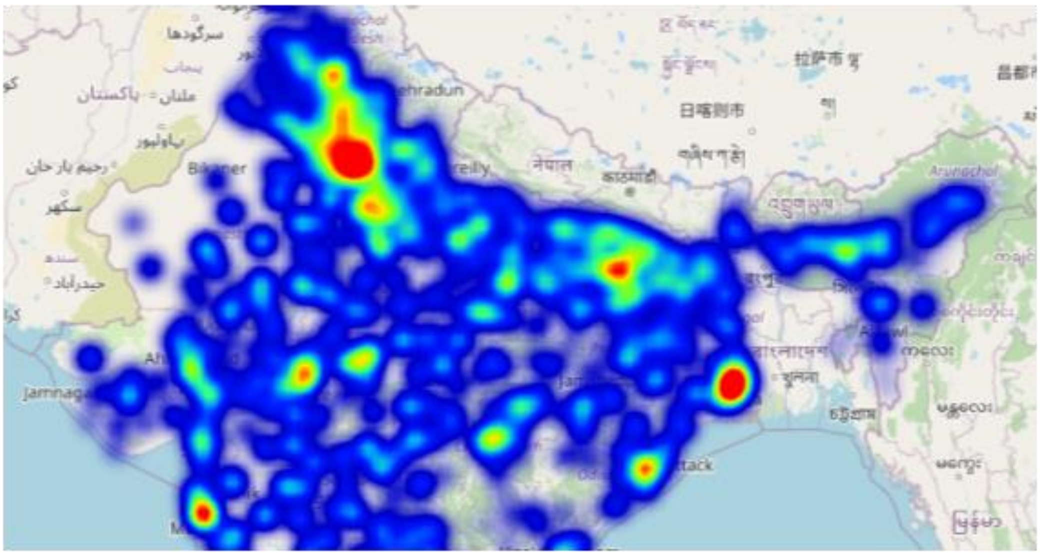


Figure 15. Plotly heat map visualization for the values of attribute "Target" (2nd zoom level)

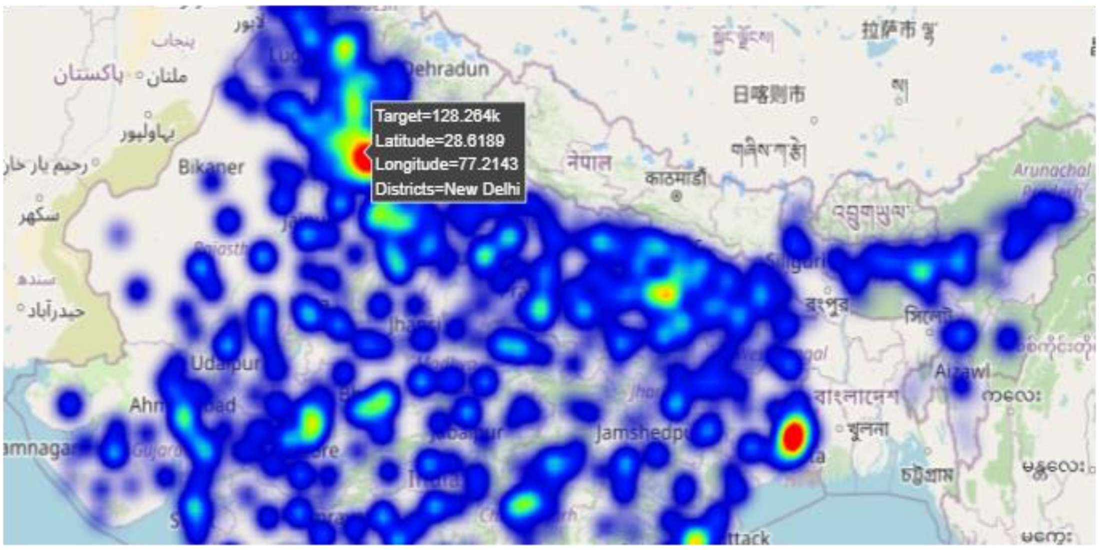


Figure 16. Plotly heat map visualization for the values of attribute "Target" (3rd zoom level)

Figure 17. Map visualization for showing the different marker color according to specific range

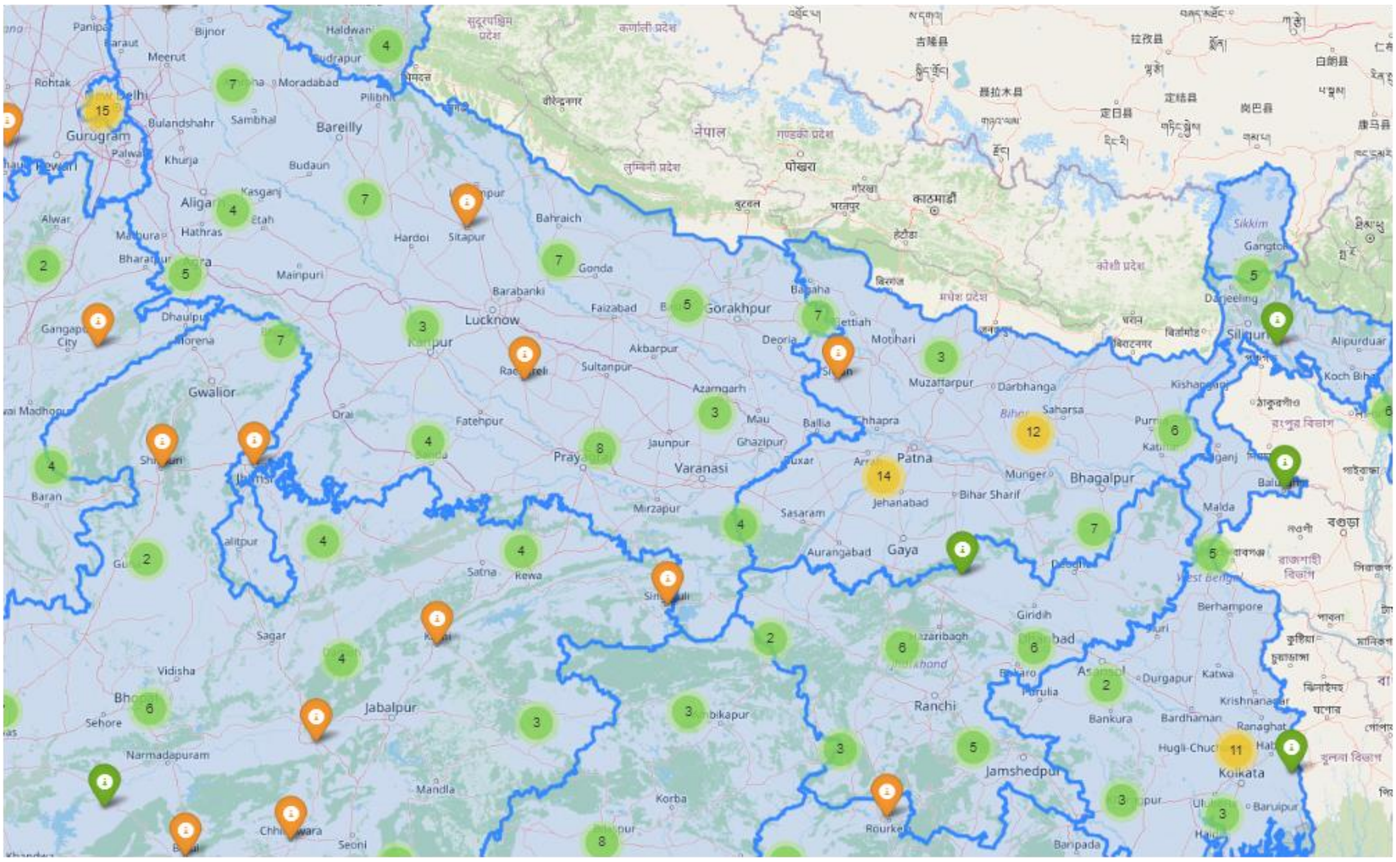


Figure 18. Map visualization for showing the different marker color according to specific range (2nd zoom level)

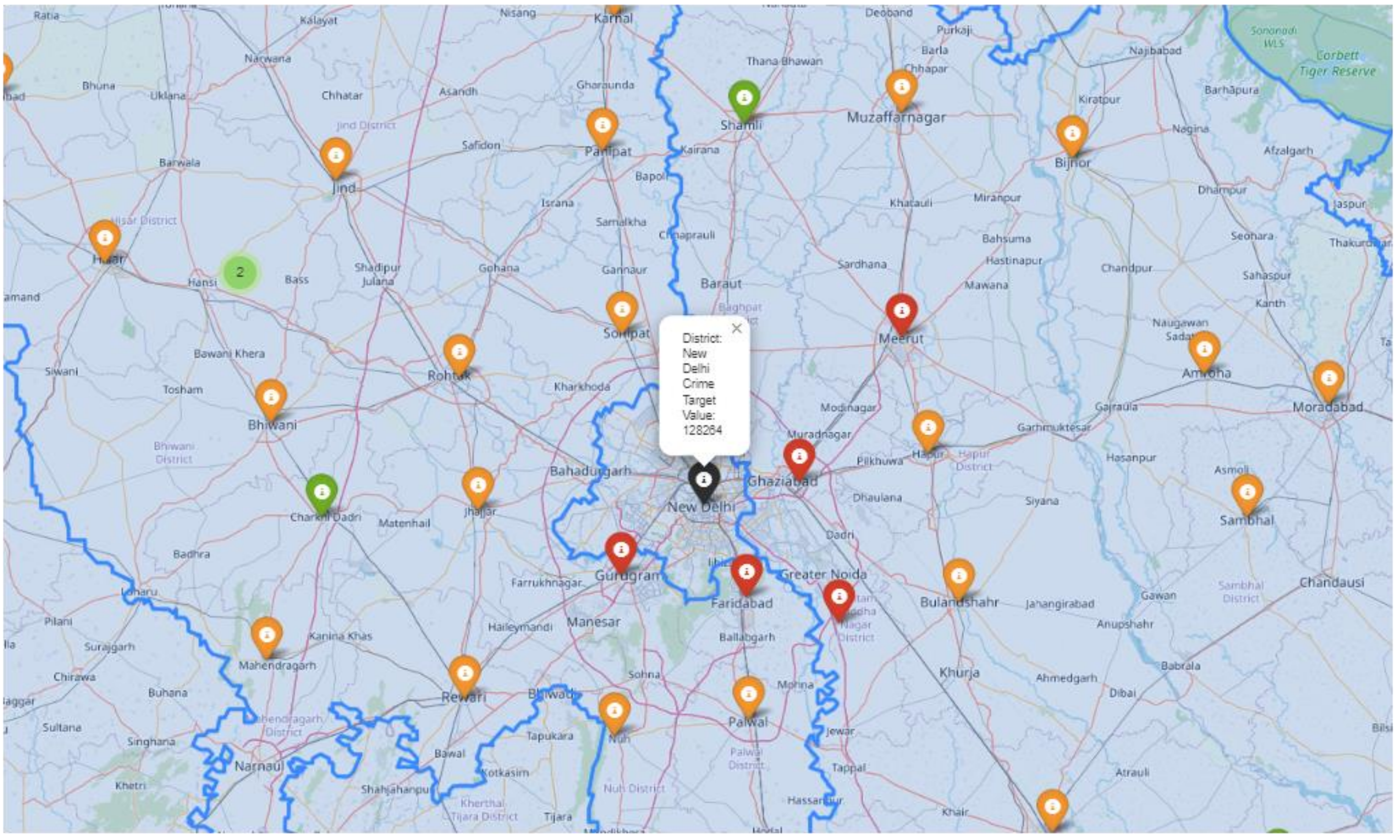


Figure 19. Map visualization for showing the different marker color according to specific range (3rd zoom level)

Prioritization of Crime Events

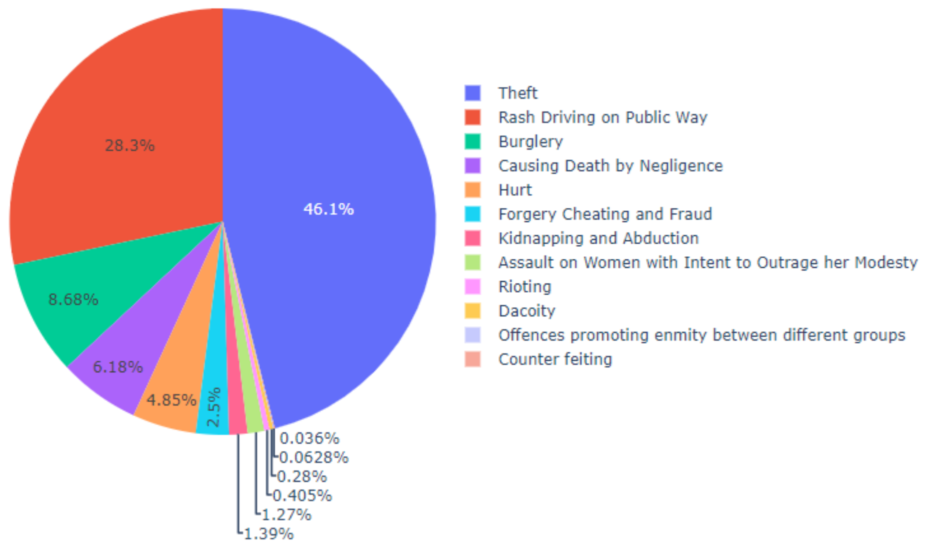


Figure 20. Pie chart of the values for each attribute (2017 dataset)

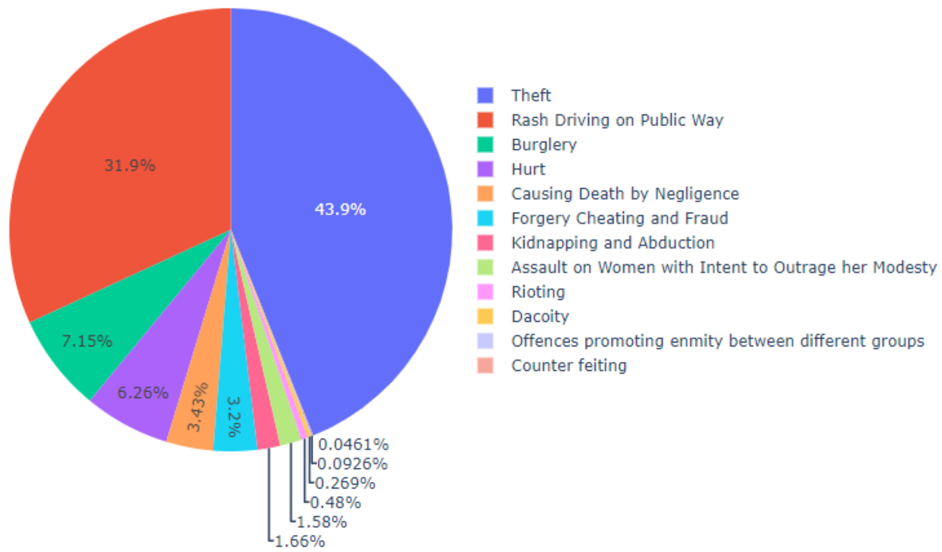


Figure 21. Pie chart of the values for each attribute (2018 dataset)

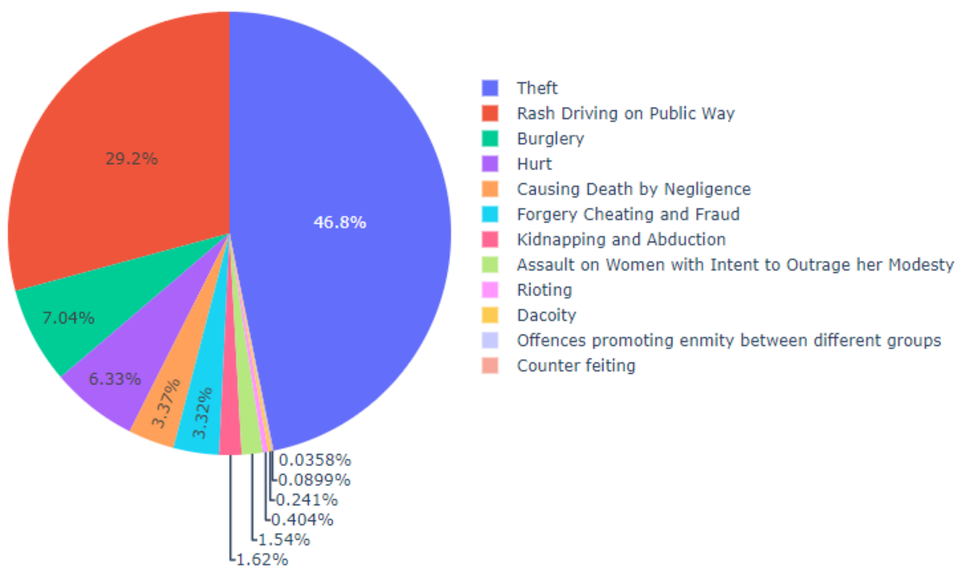


Figure 22. Pie chart of the values for each attribute (2019 dataset)

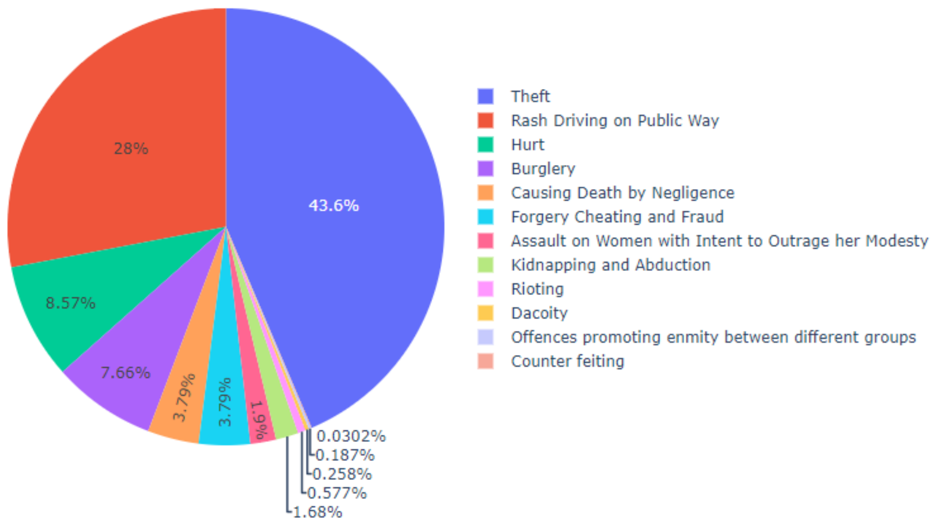


Figure 23. Pie chart of the values for each attribute (2020 dataset)

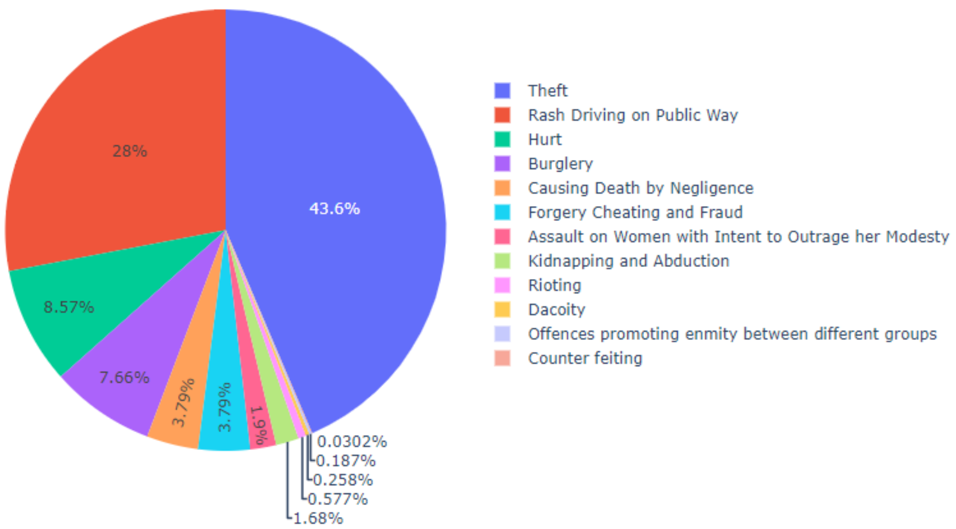


Figure 24. Pie chart of the values for each attribute (2021 dataset)

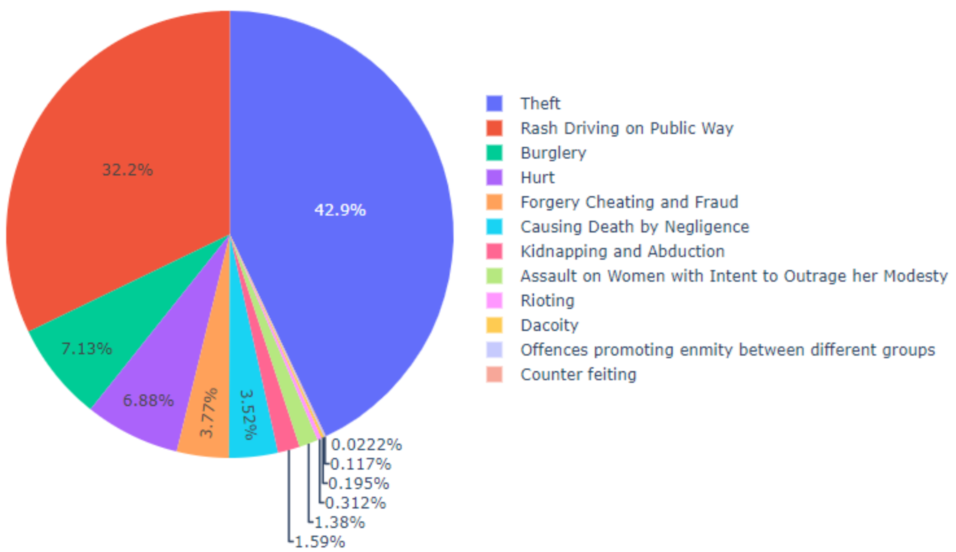


Figure 25 Pie chart of the values for each attribute (2022 dataset)

Prioritizing crime events is crucial for effective law enforcement and resource management. This process involves identifying and focusing on crimes that have the highest impact on public safety and community well-being. After the analysis done on spatial view of crime events in each district, we analysed that by including the target variable to the heatmap visuals, there was an ambiguity to understand which of the crime is contributing the most. With that instinct in mind the prioritization helped in several ways where each crime rate was keenly taken into consideration.

Factors influencing prioritization include the severity of the crime, frequency of occurrence, potential for harm, and community concerns. Prioritization is done basically with the idea with respect to all the crimes which is been considered in the dataset. After spatial analysis, the question arising was to get to know which crime is contributing the most for the target variable which is been considered. As per the graphical visualization it was easier with Pie chart visualization where it makes the viewer to easily understand the dataset. As we did the exploratory data analysis for all 6-year dataset, it was necessary to get the prioritization charts for all the years. To visualize the prioritization of crime events, pie charts were created for each year's dataset, providing a clear and immediate representation of the proportion of various crimes.

These pie charts highlight the distribution of diverse types of crimes, with theft crime consistently occupying a massive portion of the chart across all years. This visual approach facilitates an intuitive understanding of the dominance and persistence of theft crime within the overall crime landscape. The pie charts reveal trends and changes in crime distribution over the six-year period, enabling law enforcement to track shifts in crime patterns. By observing the relative proportions of theft crime compared to other crimes, strategic decisions can be made to prioritize theft crime due to its higher incidence and impact on the community.

This method of prioritization, grounded in data visualization, ensures that the most pressing issues receive focused attention. By consistently monitoring and analysing the proportion of theft crimes, law enforcement can allocate resources effectively, implement targeted interventions, and develop policies aimed at reducing the incidence of theft. Additionally, this approach helps in communicating priorities to stakeholders and the community, fostering transparency and collaborative efforts in crime prevention and safety enhancement.

Forecasting of Crime Events using ARIMA and XGBoost Model

Forecasting models are preferred in crime pattern recognition when dealing with time-series data due to several significant advantages. Firstly, forecasting models are specifically designed to handle and analyse sequential data, capturing temporal dependencies and trends that are crucial for accurate predictions in time-series contexts. ARIMA model excel in identifying patterns over time, including seasonality and cyclic behaviour, which are often present in crime data. By leveraging the temporal structure, forecasting models can provide more precise and timely predictions, allowing law enforcement to anticipate and respond to crime trends effectively. Forecasting models can incorporate historical crime data to project future occurrences, providing a dynamic and forward-looking approach that is essential for proactive policing and resource allocation. By utilizing forecasting models, law enforcement agencies can gain a deeper understanding of crime trends and enhance their ability to prevent and mitigate criminal activities. The data which we considered during the phase 1 part was considered here, where the data needed some of the preprocessing techniques. All 6 years data was combined as forecasting model must be trained using a single data file for easy learning purpose. In the dataset which was given for the machine to learn consisted of 14 columns where Theft crime was focused on, and the dataset consisted of 4403 rows in total. The 6 years data which was previously considered for spatial analysis and for prioritization of crime events included Latitude and longitude values too, which was not necessary for forecasting model.

Year	Causing D	Hurt	Assault or	Kidnappir	Rioting	Offences	Theft	Burglery	Dacoity	Counter fe	Forgery Cl	Rash Drivi	Target
2017	211	174	52	10	4	3	303	183	3	2	65	967	1977
2017	194	72	21	2	3	1	274	71	4	1	27	162	832
2017	180	208	62	4	2	1	347	156	0	1	47	3398	4406
2017	246	86	63	5	0	3	640	248	1	1	55	1197	2545
2017	1	2	0	0	0	0	445	0	0	1	0	0	449
2017	327	192	61	24	3	2	968	269	1	1	132	3352	5332
2017	138	131	55	12	1	5	291	175	0	1	52	458	1319
2017	220	108	42	6	2	2	326	196	2	2	69	586	1561
2017	183	132	37	5	3	0	555	251	7	0	40	2792	4005
2017	182	123	47	6	4	0	345	181	5	1	28	546	1468
2017	61	37	12	2	0	1	321	89	2	1	11	233	770
2017	114	74	12	1	1	0	68	61	0	0	20	483	834
2017	95	35	10	3	4	0	391	152	3	2	46	937	1678
2017	134	53	35	4	0	4	1903	206	3	1	97	654	3094
2017	252	121	58	19	1	0	974	328	3	1	133	816	2706
2017	166	117	13	2	1	0	127	75	1	0	18	1968	2488
2017	251	148	60	9	0	10	580	285	3	0	72	664	2082
2017	0	1	0	0	0	0	3	0	0	0	0	0	4
2017	3	1	0	1	0	0	17	11	1	0	1	2	37
2017	0	0	0	0	0	0	1	1	0	0	0	0	2
2017	1	3	1	0	0	0	13	7	0	0	1	1	27
2017	7	0	1	0	0	0	11	15	0	0	1	7	42
2017	0	1	0	0	0	0	5	1	0	0	0	0	7
2017	1	0	0	1	0	0	19	10	0	0	1	6	38
2017	1	1	0	1	0	0	11	5	1	0	0	2	22
2017	2	0	0	0	0	0	6	5	0	0	1	1	15
2017	7	8	4	4	0	0	177	26	0	0	5	26	257
2017	4	2	1	1	0	0	50	14	1	0	2	14	89

Figure 26. Processed dataset, excluding longitude and latitude values

ARIMA and XGBoost forecasting models were employed to recognize crime patterns using a six-year dataset, focusing on 12 selected attributes. The analysis specifically targeted theft crime identified as the most prevalent crime in the dataset. By prioritizing theft crime events, the study aimed to enhance predictive accuracy and provide actionable insights for law enforcement. The ARIMA model, known for its effectiveness in handling time series data with trends and seasonality, complemented the XGBoost model, which excels in capturing complex nonlinear relationships and interactions among variables. Together, these models offered a robust framework for anticipating theft crime trends, thereby aiding in the strategic allocation of resources and proactive crime prevention measures.

```
[1] "accuracy = 98.2958329482919 %"
```

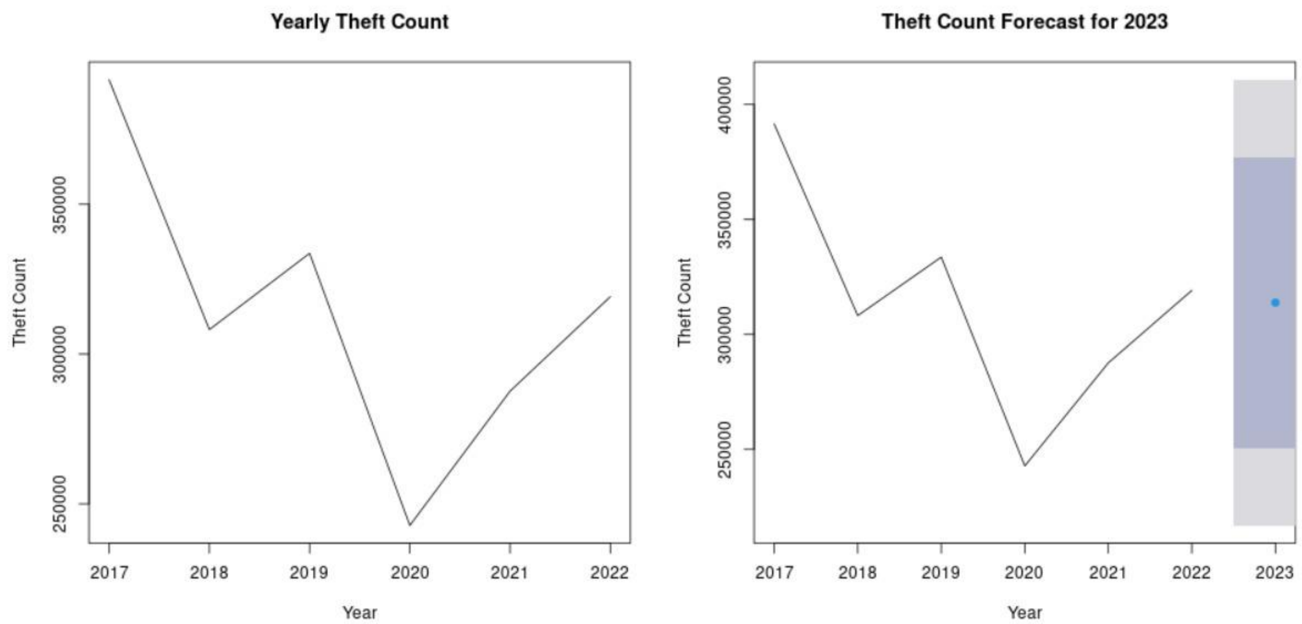


Figure 27. ARIMA model

The provided images display the forecasting results of crime counts using two different models: ARIMA (Auto Regressive Integrated Moving Average) and XGBoost (Extreme Gradient Boosting). Both models are applied to a six-year dataset containing consistent attributes across the years. This comparative study aims to evaluate the performance and effectiveness of these models in predicting crime patterns, particularly focusing on the year 2023. The first image illustrates the output from the ARIMA forecasting model. ARIMA is well-known for its capability to handle time-series data by modelling the data points in terms of past values and past errors. The historical data trend from 2017 to 2022 shows significant fluctuations in crime count, with notable peaks and troughs. The ARIMA model forecasts the crime count for 2023 as a discrete red point, suggesting an expected value of around 320,000 crimes. This point prediction aligns with the increasing trend observed from 2020 to 2022, indicating that the model anticipates a continuation of the upward trend in crime counts.

The second image represents the XGBoost forecasting model results. XGBoost is a powerful machine learning algorithm designed for high predictive accuracy and efficiency. The left panel displays the historical yearly theft count from 2017 to 2022, similar to the ARIMA plot, showing a pattern of significant variability over the years. The right panel provides a forecast for 2023, incorporating a range of uncertainty. The forecasted point for 2023 is around 313,819 crimes, with an 80% confidence interval ranging from approximately 250,456 to 377,182, and a 95% confidence interval extending from about 216,914 to 410,724. This broader range indicates the model's consideration of variability and uncertainty in its predictions.

The comparative analysis highlights several key differences and similarities between the two models. While both models project a similar point forecast for 2023, the ARIMA model provides a narrower and more precise forecast, reflecting its nature of capturing linear relationships and being influenced by latest trends. In contrast, the XGBoost model, with its machine learning foundation, accounts for a wider range of potential outcomes, reflecting the complexity and uncertainty inherent in crime data. This broader forecast range could be more useful in planning and resource allocation, where accounting for a wider array of possibilities is crucial.

In conclusion, both ARIMA and XGBoost models offer valuable insights into future crime trends, but they cater to distinct aspects of forecasting. ARIMA is beneficial for its simplicity and effectiveness in time-series analysis, while XGBoost excels in capturing complex patterns and providing detailed uncertainty estimates. The choice between these models depends on the specific needs of the analysis, whether the focus is on straightforward, interpretable predictions or on a comprehensive understanding of prediction uncertainty and potential variability.

```
Fallback number of boosting rounds: 100
[1] 333621.9
[1] "Mean Absolute Error (MAE): 14454.90625"
[1] "accuracy = 95.3938715469745 %"
```

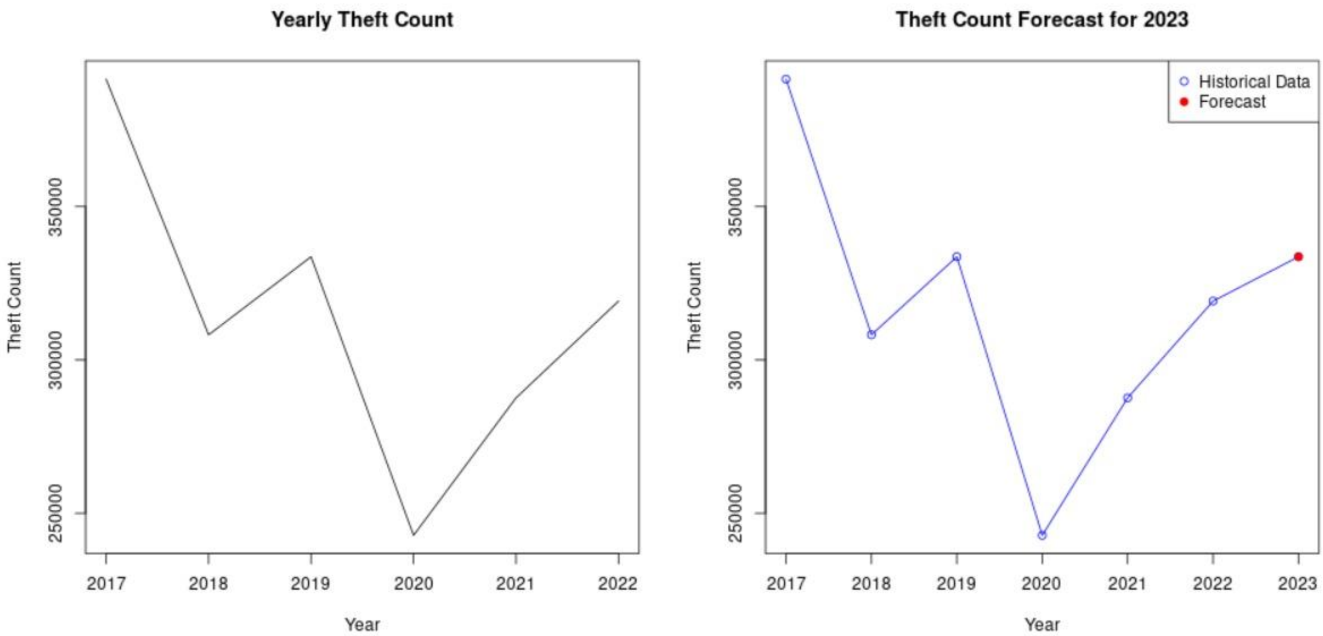


Figure 28. XGboost model

Validation and Evaluation

The two images provided illustrate the validation and evaluation of crime forecasting models, specifically comparing the XGBoost and ARIMA models. The first image presents the performance of the XGBoost model, indicating a theft count forecast for 2023. The model demonstrates a high level of accuracy at 95.39% with a Mean Absolute Error (MAE) of 14,454.91. The plot on the left shows the historical theft counts from 2017 to 2022, exhibiting a fluctuating trend. The plot on the right compares the historical data with the 2023 forecast, indicating a notable increase in theft counts, consistent with the observed trend. In contrast, the second image showcases the ARIMA model's forecasting performance. This model yields an even higher accuracy rate of 98.29%, which is remarkable for time series forecasting.

The left plot again depicts the historical theft counts, while the right plot represents the forecast for 2023, including a confidence interval. The ARIMA model predicts a slight increase in theft counts for 2023, and the inclusion of the confidence interval provides a visual representation of the uncertainty in the forecast, which is relatively narrow, indicating a high level of confidence in the prediction. When comparing the XGBoost and ARIMA models, both demonstrate strong forecasting capabilities with high accuracy rates. However, the ARIMA model slightly outperforms the XGBoost model in terms of accuracy. The XGBoost model's MAE indicates it has a relatively small average error in its predictions, but the ARIMA model's tighter confidence interval suggests it may provide more reliable forecasts.

This reliability is crucial in crime forecasting as it allows for better planning and resource allocation by law enforcement agencies. The graphical representation of historical data and forecasts highlights the importance of visual analytics in understanding and interpreting model predictions. The trend analysis from 2017 to 2022 shows significant variations in theft counts, which both models successfully capture and project into the future. The XGBoost model's forecast for 2023 shows a more pronounced increase in theft counts compared to the ARIMA model, suggesting it might be more sensitive to latest trends. Conversely, the ARIMA model's forecast, with its confidence intervals, provides a balanced view of expected trends while accounting for possible variability.

In conclusion, both the XGBoost and ARIMA models offer valuable insights for crime forecasting, each with its strengths. These models are instrumental in aiding law enforcement agencies in making data-driven decisions, improving crime prevention strategies, and optimizing resource deployment. The choice between these models may depend on specific needs, such as the importance of capturing recent trends versus accounting for forecast uncertainty.

Ethical Considerations

Ethical considerations are crucial in any data analysis project to ensure the responsible use of data and the protection of individuals' rights. For our project involving the analysis of district-wise IPC crime data for the years 2017-2022, several ethical considerations were considered, Firstly, we ensured that our analysis was conducted with respect for the communities represented in the data, avoiding any stigmatization or negative labelling of districts. Additionally, we ensured that the data collection processes adhered to standards of informed consent where applicable, making sure that individuals were aware of and agreed to the use of their data. Robust data security measures were also put in place to prevent unauthorized access, use, or dissemination of the data. We paid careful attention to identifying and addressing any biases in the data to ensure that the findings were fair and did not disproportionately impact any particular group. Finally, we maintained transparency in our data processing and analysis methods to allow for scrutiny and verification by external parties, fostering trust and accountability in our work.

Conclusion

This paper has utilized 6-year crime data (2017-2022) that are specific to Indian conditions. The research findings in this paper have used Spatial Analysis for crime hotspot identification, ARIMA model and XGBoost to forecast future crime behaviour. This helps in identifying the probable factors responsible for causing the crime. The forecasting rate accuracy for both ARIMA and XGBoost model are about 98% and 95%, respectively. With the help of these insights, regions with high levels of crimes can be selected for intense observation as a preventive method for reducing crime rates. 12 types of crimes are considered, and the data is considered based on them. Along with the present scope of the project, which is forecasting of crime in the future years, we can also do the forecasting rate of crimes on the geospatial areas takes place in future scope. Along with this, one can also try to use the data precisely with specific data size. As a result, by examining crime patterns, the system will automatically learn to recognize changing patterns. Also crime factors change over time. As these results will be a assisting factor to the law agencies who handle the crime activities throughout the country.

References

1. Hajela, G., Chawla, M., & Rasool, A. (2021). Crime hotspot prediction based on dynamic spatial analysis. ETRI Journal, 43(6), 1058-1080.

2. Almanie, T., Mirza, R., & Lor, E. (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. arXiv preprint arXiv:1508.02050.

3. Sarah, J., Danny, A. M., Deen, J. M., Dongre, L., SV, C., & Ramchandani, H. (2021). Analysing Crimes of Indian Datasets Based on Machine Learning Methods. Turkish Online Journal of Qualitative Inquiry, 12(8).

4. Mondal, S., Singh, D., & Kumar, R. (2022). Crime hotspot detection using statistical and geospatial methods: a case study of Pune City, Maharashtra, India. GeoJournal, 87(6), 5287-5303.

5. Boppuru, P. R., & Ramesha, K. (2020). Spatio-temporal crime analysis using KDE and ARIMA models in the Indian context. International Journal of Digital Crime and Forensics (IJDCF), 12(4), 1-19.

6. Agarwal, J., Nagpal, R., & Sehgal, R. (2013). Crime analysis using k-means clustering. International Journal of Computer Applications, 83(4).

7. Das, S., & Choudhury, M. R. (2016). A geo-statistical approach for crime hot spot prediction.

8. Sharma, R., Kaned, Y., Singh, S., Lund, A., & Goplani, B. (2019). Crime Analysis and Hotspot Prediction. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 11(SUP), 425-429.

9. Chikodili, H. U., Ogbobe, P. O., & Okoronkwo, M. C. (2021). Analysis of Crime Pattern using Data Mining Techniques. International Journal of Advanced Computer Science and Applications, 12(12).

10. ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. Procedia computer science, 132, 696-705.

11. Sonawanev, T., Shaikh, S., Shaikh, S., Shinde, R., & Sayyad, A. (2015). Crime pattern analysis, visualization and prediction using data mining. IJARIIE, 1(4), 681-686.

12. Malleson, N., Steenbeek, W., & Andresen, M. A. (2019). Identifying the appropriate spatial resolution for the analysis of crime patterns. PloS one, 14(6), e0218324.

13. Sathyadevan, S., Devan, M. S., & Gangadharan, S. S. (2014, August). Crime analysis and prediction using data mining. In 2014 First international conference on networks & soft computing (ICNSC2014) (pp. 406-412). IEEE.

14. Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 1, pp. 225-230). IEEE.

15. Nath, S. V. (2006, December). Crime pattern detection using data mining. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops (pp. 41-44). IEEE.

16. Chauhan, C., & Sehgal, S. (2017, May). A review: crime analysis using data mining techniques and algorithms. In 2017 International Conference on Computing, Communication and Automation (ICCCA) (pp. 21-25). IEEE.

17. Krishnendu, S. G., Lakshmi, P. P., & Nitha, L. (2020, March). Crime analysis and prediction using optimized K-means algorithm. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 915-918). IEEE.

18. Joshi, A., Sabitha, A. S., & Choudhury, T. (2017, October). Crime analysis using K-means clustering. In 2017 3rd International conference on computational intelligence and networks (CINE) (pp. 33-39). IEEE.

19. Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017, May). An overview on crime prediction methods. In 2017 6th ICT International Student Project Conference (ICT-ISPC) (pp. 1-5). IEEE.

20. Vijayarani, S., Suganya, E., & Navya, C. (2020). A comprehensive analysis of crime analysis using data mining techniques. International Journal of Computer Science Engineering (IJCSE), 9(1).

21. Begam, M. R., Sengottuvelan, P., & Ramani, T. (2015). Survey: Tools and Techniques implemented in Crime Data Sets. vol, 2, 707-710.

22. Jabeen, N., & Agarwal, P. (2021). Data mining in crime analysis. In Proceedings of Second International Conference on Smart Energy and Communication: ICSEC 2020 (pp. 97-103). Springer Singapore.

23. *Mohammed, A. F., & Baiee, W. R. (2020, November). Analysis of criminal spatial events in GIS for predicting hotspots. In IOP conference series: materials science and engineering (Vol. 928, No. 3, p. 032071). IOP Publishing.*