Identification of Crime Prone Areas

Irfan Siddavtam
Information Technology, KJSCE
irfansiddavatam@somaiya.edu
Mumbai, India

Labdhi Jain
Information Technology, KJSCE
labdhi.jain@somaiya.edu
Mumbai, India
1814015

Dhruv Doshi
Information Technology, KJSCE
dhruv.doshi@somaiya.edu
Mumbai, India
1814002

Shubham Bhakuni
Information Technology, KJSCE
shubham.bhakuni@somaiya.edu
Mumbai, India
1814006

Kunj Gala
Information Technology, KJSCE
kunj.gala@somaiya.edu
Mumbai, India
1814021

***Abstract*— This paper soughts to identify areas in New York City that are prone to crimes. This paper also identifies the type of crime which has the highest frequency in a particular area. The paper discusses the dataset attributes, preprocessing required on the data, and different clustering techniques like K means, DBSCAN, HDBSCAN, Birch, etc. Finally the paper then compares and uses the most optimal algorithm for our problem. Lastly, detailed statistics on crimes across different regions are provided.**

Keywords*— *Crime, NYPD, Kmeans, DBSCAN

I. INTRODUCTION

Crime rate is increasing day by day and even though crimes can occur anywhere, they are more in some places than others. When selecting a neighborhood to move to, crime rate is one of the important factors people consider. For authorities too, knowing the high crime rate areas is crucial to ensure the community's safety.Hence, the identification of crime prone areas will not only benefit the citizens but also the authorities to make better and informed decisions. This project aims to utilize the power of Machine Learning to tackle the increasing criminal activity in our society. We can identify the crime prone areas by analysing the historic data from emergency services. Further, the aim is to build a solution for predictive policing that is able to predict the type of crime, given the location. This solution will help the authorities continually identify areas prone to crimes and policing-related incidents, making deployments and patrolling in those areas more effective and thereby preempt and prevent any untoward incidents that could result in an emergency.

The dataset used in the project is by the New York Police Department(NYPD). The data is collected by the NYPD from the ICAD system. It consists of 5 million rows and 19 columns. Latitude, longitude, incident description, incident data and time, radio code are some of the important columns. The radio code attribute consists of a unique code for each crime type listed in the dataset. The spread of the data is shown in Fig.1

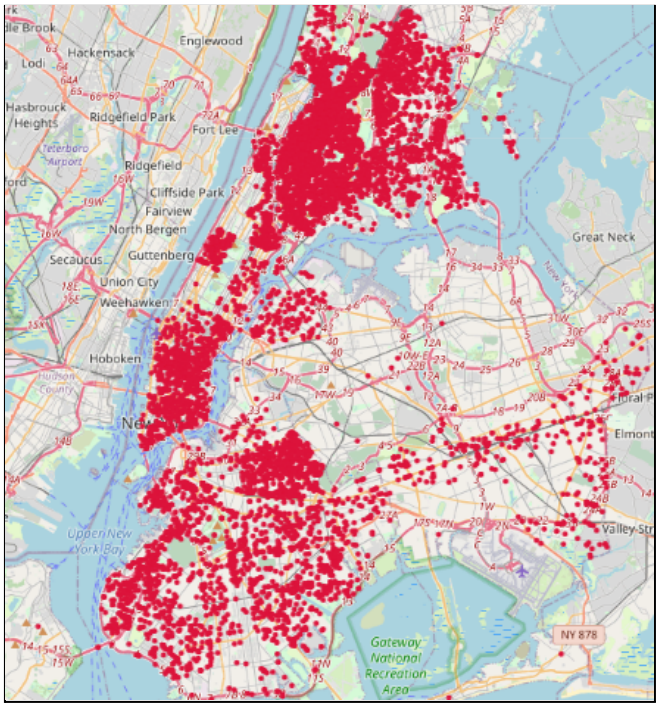

Fig. 1 The spread of the data points

The further sections discuss the methodology used i.e, data preprocessing and model building, comparison of clustering algorithms and the results obtained.

II. LITERATURE REVIEW

Sr. No.	Name	Publication Type	Publication Year	Publication Agency	Aim	Conclusion
1.	Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis	Conference	2021	IEEE	To predict Spatio-temporal criminal patterns within the New York City neighbourhoods.	Classified and predicted the criminal hotspots using XGboost, SVM and Random Forest classifiers; XGboost model outperforms.
2.	Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques	Journal	2021	IEEE	To Improve the predictive accuracy by Logistic Regression, SVM, Naïve Bayes, etc; time-series analysis by LSTM; visual summary through EDA; crime forecasting.	XGBoost achieves the maximum accuracy in chicago dataset and KNN in Los Angeles dataset.
3.	Crime Analysis and Prediction Using Data Mining	Conference	2014	IEEE	To predict regions which have high probability for crime occurrence and visualize crime prone areas.	Tested the accuracy of classification and prediction based on different test sets; used Bayes theorem which showed more than 90% accuracy.
4.	Discovering Crime Trends and Patterns Using Three-Dimensional Visual Analytics	Conference	2019	IEEE	To Introduce a 3D visual analytics framework that interactively visualizes crime data and other relevant datasets on a highly accurate 3D model of the City of Vancouver, Canada.	Used three major visualization techniques - building highlights, street sections, and kernel density; gained insights into the data which may assist in developing innovative strategies for crime prevention.
5.	Data visualization of crimes in a city using machine learning	Conference	2020	IEEE	Visualize the NY City crime data in form of various plots	After step-by-step implementation, attributes used for plotting graphs were year, area, premise code and age. These were the important parameters required for the plots.
6.	Area-Specific Crime Prediction Models	Conference	2016	IEEE	Crime prediction models based on hierarchical and multi-task statistical learning.	Developed area-specific crime prediction models using hierarchical and multi-task learning. Mitigate sparseness by sharing information across areas,

7.	Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas	Conference	2020	IEEE	Built model for news headline classification, crime location extraction, and crime report generation where the text was in bengali language.	The overall accuracy of the model was 82% ; automated the data for computing district wise crime rate with data in Indian language
8.	Crime Prediction Using Auto Regression Techniques for Time Series Data	Conference	2018	IEEE	Explored the Auto Regression Techniques to accurately predict the crime with minimum error for time series data by identifying the relationship among crimes attributes.	Generalized Linear Model (GLM) for Crime Site Selection (CSS) using Big Data deliver better results and forecast spatio-temporal crime events with certainty.

III. PROPOSED METHODOLOGY

A. Data Preprocessing

The first step of data preprocessing was **Data Cleaning.** It is a way of identifying incomplete or inaccurate parts of the data and then modifying, replacing or deleting them based on the necessity. The column 'CAD_EVNT_ID' had 3.2% repeated values. The rows with repeated values were dropped. Later all the missing values were dropped. After this step, 4,871,204 rows were left.

As a part of the **Data Reduction** step, the OBJECTID column was dropped as it had 68% missing values, CREATE_DATE column was dropped as it was irrelevant for our project, BORO_NM and PATRL_BORO_NM were highly correlated so PATRL_BORO_NM was retained as it provided more detailed information,GEO_CD_X and GEO_CD_Y were also dropped as they were highly correlated to the Latitude and Longitude columns respectively. Next, the RADIO_CODE column had 420 distinct events but not all of them were crimes. Hence, radio codes which did not represent crimes were manually detected and dropped. Also, different radio codes that depicted the same crime were merged into a single radio code. Finally data was reduced to 10 unique crime types including harassment, robbery, assault, etc. as shown in table 1.

	TYP_DESC
22Q2	Larceny
29H1	Harassment
29Q1	Other
39T1	Trespassing
20R	Roberry
53D	Dispute
24Q2	Assault
50G2	Disorderly Conduct
10S2	Shooting
33	Explosion

Table 1 Crime types

The third step of data pre-processing was **Data Integration.** Data integration is the process of combining data from different sources into a single, unified view. In this step INCIDENT_DATE and INCIDENT_TIME were merged into the INCIDENT_DATE_TIME column as shown in Fig. 2.

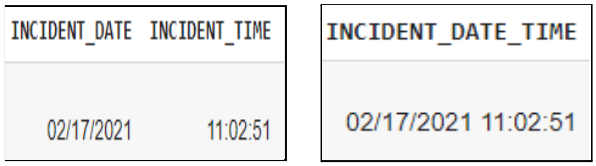

Fig. 2 Data integration

The fourth data pre-processing step is **Data Conversion**. Data conversion means changing the data in one format to another format. The datatype of INCIDENT_DATE_TIME was originally String. But when in String, date/time operations could not be performed on them. Hence, this attribute was converted into Pandas DateTime (Timestamp) datatype. Similarly ADD_TS, DISP_TS, ARRVD_TS and CLOSNG_TS were also converted into the Pandas DateTime (Timestamp) datatype.

The next step in data preprocessing was **Data Sampling**. Sampling is performed to reduce the size of the dataset by selecting only the samples that would represent the whole dataset. Since the dataset selected was too huge (1064806 rows and 13 columns), a sample containing 25% data using the random sampling method was created which could be used in further model building. After performing this step the dataset obtained had 500459 rows.

Then lastly **One Hot Encoding** was implemented on the dataset to perform binarization of the categories in order to use those columns in our model. 'PTRL_BORO_CD' column had 8 categories as there are 8 boroughs in New york. This column was one hot encoded to convert it to 1/0 data for each category. Similarly, 'CIP_JOBS' containing 4 categories was also one hot encoded. So after the entire data preprocessing the dataset obtained had 500455 rows and 22 columns.

B. Model Building

Clustering is a technique of grouping data into different clusters, consisting of similar data points. In this section different clustering algorithms or models are used on the dataset to finally find the most optimal one.

a. K Means

K means clustering algorithm was implemented to find the custers with high crime prone areas. K represents the number of clusters. The algorithm continuously computes the centroids of clusters until it finds the optimal clusters. The data points were assigned to a cluster by computing squared distances between the centroid and the data point and choosing the one with the least distance. This ensures that the similar data points belong to the same clusters. To determine the optimal value of k the elbow method was used from which k=5 was obtained.

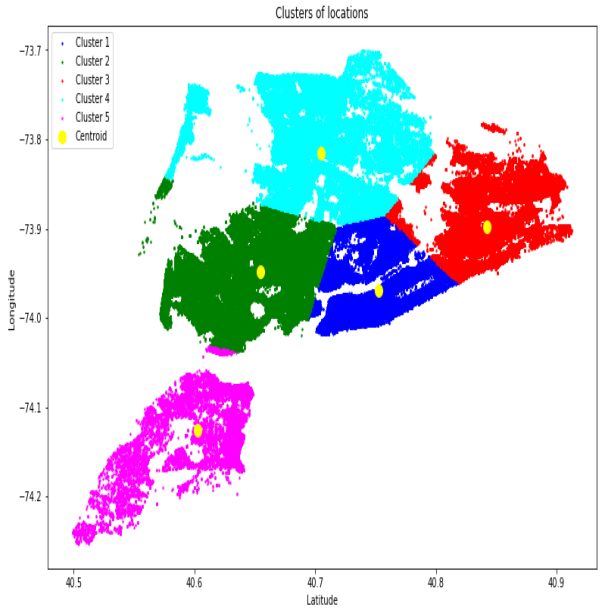

Fig. 3 K means clustering result

Fig 3. Shows the result obtained. Here the data points have been divided into 5 clusters and these clusters represent high crime prone areas.

b. Dbscan

Dbscan is a density based clustering algorithm. DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It clusters the data based on the density and the algorithm is able to correctly identify the outliers. It also does not need the value of clusters to be declared beforehand. Two main parameters used in dbscan algorithm are epsilon and minpoints. Epsilon indicates the radius of the circle that needs to be created around a center point to make clusters and minpoints indicates the least number of points that must be included in the circle to declare it as one cluster. eps=0.005 and minpoints=800 was set and for these values 23 labels/clusters were obtained.

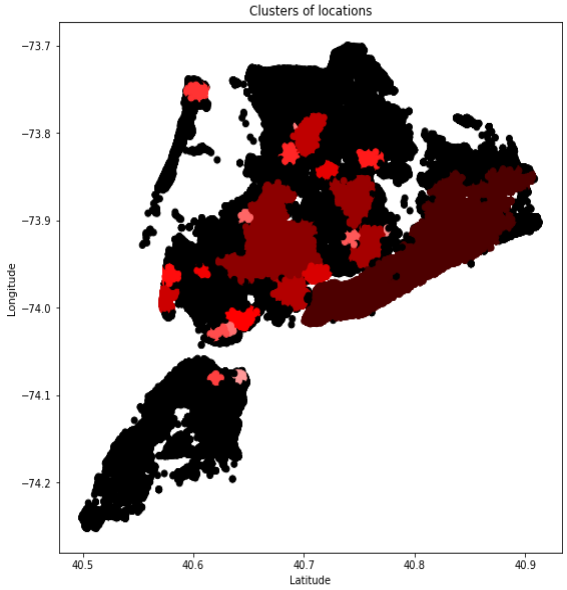

Fig. 4 DBSCAN result

As seen in Fig 4. Dbscan gave better results than k means clustering as the clusters are well defined. The algorithm also identified outliers which are shown in black color. Here, the locations where the number of crimes occurring is high, can be seen clearly.

c. Hdbscan

HDBSCAN stands for Hierarchical Density-Based Spatial Clustering of Applications with Noise. K means performs best when the clusters are round or spherical, equally dense, equally dense, not contaminated by noise, etc. But in practical situations clusters are of random shapes, different sizes and densities. HDBSCAN extends DBSCAN by converting it into a hierarchical clustering algorithm. The 2 main tuning parameters used in HDBSCAN are min_cluster_size and min_samples. The min_cluster_size specifies the minimum number of points a cluster should have. Any linkage with lesser points than the min_cluster_size is considered to be falling out of the cluster. The min_samples parameter specifies the minimum number of points in the neighborhood for a point to be considered a core point. The min_cluster_size and min_samples were set to 1000 and 50 respectively to get 15 labels. The results obtained using this method are shown in Fig. 5. The results were not satisfactory as clear clusters were not observed,

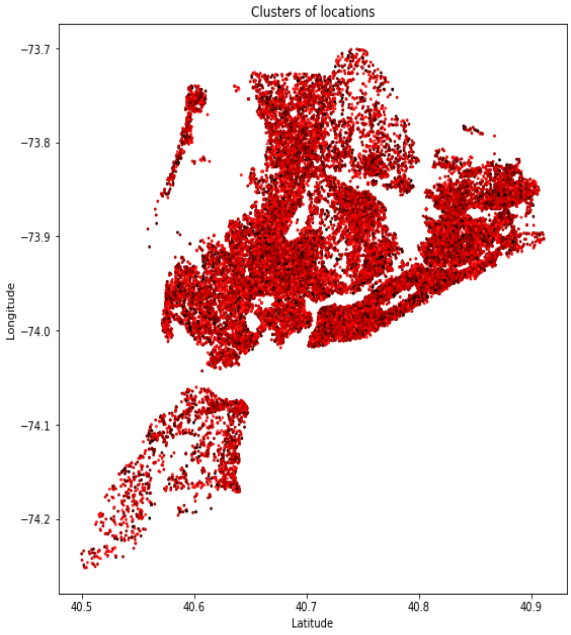

Fig. 5 HDBSCAN result

d. Optics

OPTICS Clustering stands for Ordering Points To Identify Cluster Structure. Dbscan is sensitive to parameter setting hence to alleviate this problem. Optics clustering algorithm was developed. It does not actually produce dataset clustering. Instead it outputs the dataset cluster ordering. It visualizes the reachability distance and uses that to extract clusters. So points with dense clusters are listed close to one another.

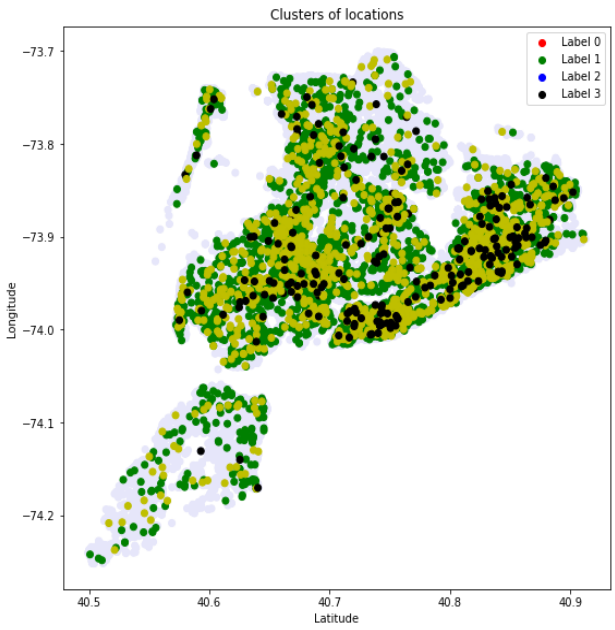

Fig. 6 OPTICS result

The result obtained from the optics algorithm can be seen in fig 6. Here 3 clusters were obtained and outliers were depicted using black color. As seen, the clusters obtained are overlapping and there are no defined boundaries between them. Hence the results from this algorithm were not satisfactory.

e. Birch

Birch stands for Balanced Iterative Reducing and Clustering using Hierarchies. This algorithm generates a small summary of large datasets that retains most of the aspects and then performs clustering on this summary. These smaller regions created using the summary are known as clustering features. Two parameters were used in the birch algorithm. One was threshold which is the maximum number of data points a cluster can hold and the other one was n_clusters which indicated the number of clusters to be formed. threshold = 0.01 and n_clusters = 10 were set forimplementation.

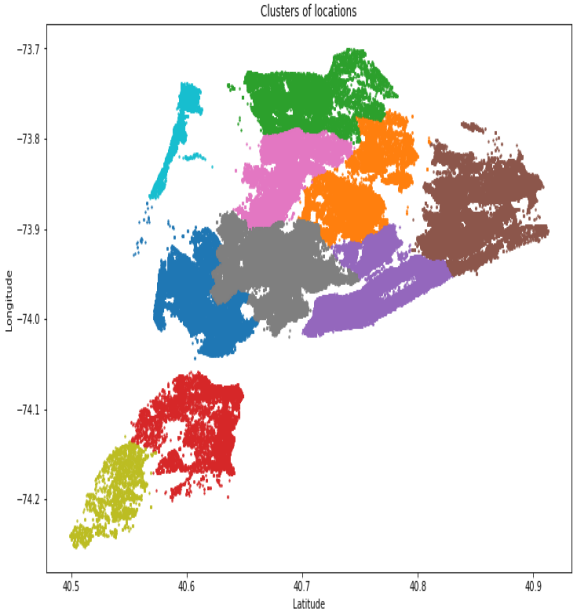

Fig. 7 Birch result

The above fig 7 shows the plot after implementing the birch algorithm. The data points were clustered into well defined 10 clusters. By seeing this plot, the areas with high crime rates can be recognized hence the results were better than the optics algorithm.

IV. COMPARISON OF CLUSTERING ALGORITHMS

The above 5 algorithms were compared using two evaluation metrics 'DB Index' and 'CH Index'. Davies-Bouldin Index is the ratio of distances within the clusters to distances between the clusters. Lower the values of db index, better is the clustering performance. Calinski-Harabasz is the ratio of the dispersion within the cluster to the dispersion between the clusters. Higher the values of ch index better is the clustering performance. The values of DB Index and CH Index for all the 5 algorithms are listed in table 2.

Algorithm	**Db Index**	**CH Index**
K Means	0.694	552559.3
DBScan	996.5	9.5067
HDBScan	4.907	135343
Optics	66.7	2.25
Birch	0.798	422257.1

Table 2 Comparison of algorithms

From the above table it can be seen that K means clustering algorithm has lowest DB Index and highest CH Index followed by Birch algorithm. Thus it can be concluded that K means clustering algorithm performs best among all the other algorithms.

V. IDENTIFICATION OF CRIME PRONE AREAS USING CRIME TYPE

Till now, the identification process of crime areas based on crime location was implemented. Now the identification of crime prone areas based on crime type will be done. As seen in table 1 there were 10 types of crime mapped to their unique radio codes. If the user wants to see statistics of a particular type he/she will enter the radio code for that particular crime and he/she will be displayed the plots accordingly. Here, the plots for the radio code '50G2' which is 'Disorderly Conduct' will be shown.

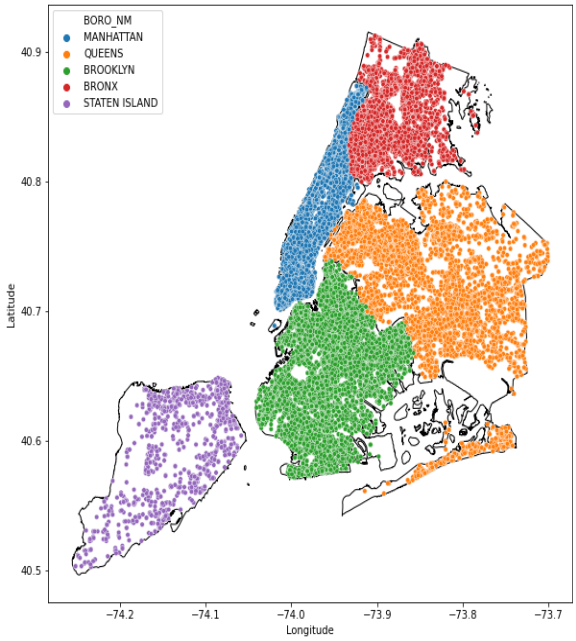

Fig. 8 Distribution of Disorderly Conduct crime data points

Geopandas was used to plot the shapefile of New York. From fig 8, it can be seen that the data points for the crime 'Disorderly Conduct' have been plotted on the New York map according to boroughs. Each borough has been indicated a different color. From this it can be inferred that Brooklyn and Manhattan are densely plotted which means disorderly conduct is high in those boroughs. Similarly Staten Island has relatively less numbers of the same crime. The same statistics were plotted using a bar graph (Fig 9) again with x axis indicating the name of boroughs and y axis indicating the number of crimes as disorderly conduct.

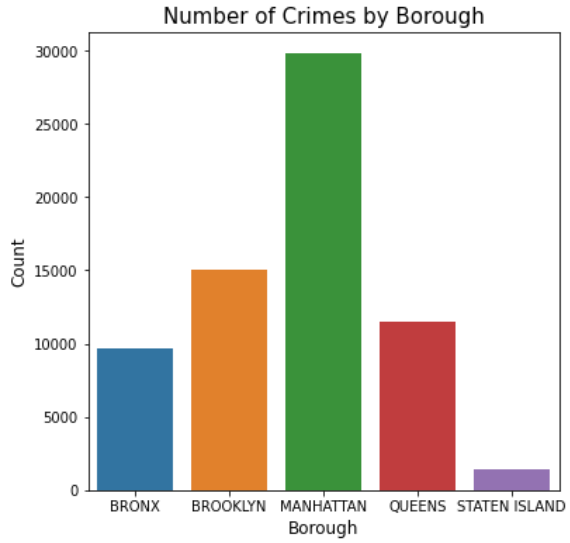

Fig. 9 Borough wise distribution of disorderly conduct crime type

The next plot (Fig 10) is the more informative plot as along with the density of crimes we can also see the total number of that crime in each borough.

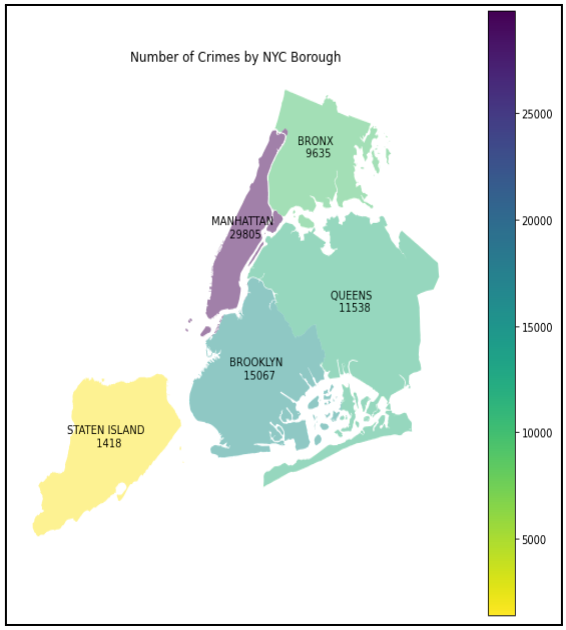

Fig. 10 Detailed distribution of disorderly conduct in NYC boroughs

As shown in the above plot, The purple color indicates high crime numbers and the yellow color indicates lowest crime numbers. Using the same color scale, the boroughs have been assigned colors with Manhattan being the borough with highest disorderly conduct crime rates followed by Brooklyn and Queens and Staten Island being the least one.

VI. RESULTS

As noted earlier, the K-means clustering algorithm outperformed all other algorithms hence the clusters obtained in K-means were used for further analysis.

First, the locations of the center of each cluster was found to identify the crime prone area names. The list of the cluster areas can be seen in Fig. 11. Next, the no of crimes in each cluster was calculated. As seen in Fig. 12, it was found that cluster 2 has the highest number of crimes followed by cluster 0 and so on.

```
No. of crimes in cluster  0 :  131367
No. of crimes in cluster  1 :  126581
No. of crimes in cluster  2 :  138149
No. of crimes in cluster  3 :  85471
No. of crimes in cluster  4 :  18887
```

Fig. 12 No. of crimes in each cluster

Later, the top three crimes in each cluster were found by taking the cluster number as an input from the user. As seen in Fig. 13 Dispute is the highest occurring crime followed by larceny and robbery in cluster number 2. This means that Croes Avenue, The Bronx, New York has the highest cases of Dispute.

```
Enter category number:2
1.  TYP_DESC    Dispute
Name: 53D, dtype: object
2.  TYP_DESC    larceny
Name: 22Q2, dtype: object
3.  TYP_DESC    robbery
Name: 20R, dtype: object
```

Fig. 13 Top 3 crimes in cluster 2

Lastly, the time between the call received by the NYPD and the time the cops arrived at the incident location was calculated. The average response time of each cluster was calculated. The results can be seen in Fig. 14.

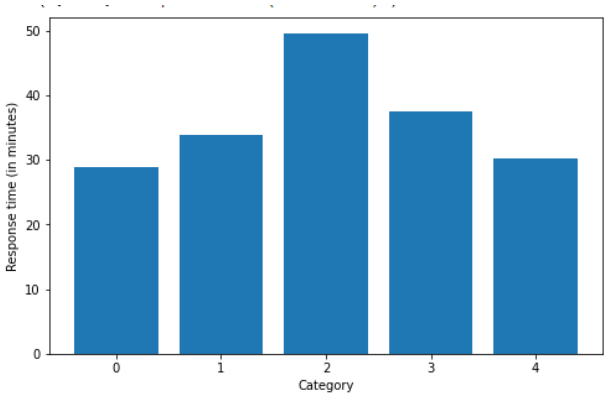

Fig. 14 Average response time of police in each cluster

It can be observed that cluster 2 has the highest response time of around 50 minutes and cluster 0 has the lowest response time of approximately 28 minutes. From this it can be inferred that the police departments are probably located far from cluster 2 and at a closer distance from cluster 0.

VII. CONCLUSION

The K-means clustering algorithm gave a DB index score 0.694 which was the best of all other algorithms implemented, hence the clusters obtained using K means were used for further crime analysis. In the analysis it was found that cluster 2, i.e, Croes Avenue, The Bronx, New York, 10472, United States, has the highest number of crimes among all the clusters. Further, the highest occurring crime in cluster 2 was Dispute.

In future, the same algorithms can be used on another dataset to compare results of both and find out which attributes are more beneficial in identifying crime prone areas. Other clustering algorithms can also be used in the future which provide a better accuracy than K means clustering algorithm.

```
Queensboro Bridge, Ed Koch Queensboro Bridge Outer Roadway, Queensbridge, Queens, Queens County, New York, 11101, United States
821, East 38th Street, Brooklyn Community District 17, New York, 11210, United States
1418, Croes Avenue, The Bronx, New York, 10472, United States
111-32, 147th Street, Cedar Manor, Queens, Queens County, New York, 11435, United States
429, Saint Andrews Road, Staten Island, Richmond County, New York, 10306, United States
```

Fig. 11 Location of center of clusters

REFERENCES

[1] A. A. Almuhanna, M. M. Alrehili, S. H. Alsubhi and L. Syed, "Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis," 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), 2021, pp. 23-30, doi: 10.1109/CAIDA51941.2021.9425120.

[2] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[3] S. Sathyadevan, M. S. Devan and S. S. Gangadharan, "Crime analysis and prediction using data mining," 2014 First International Conference on Networks Soft Computing (ICNSC2014), 2014, pp. 406-412, doi: 10.1109/CNSC.2014.6906719.

[4] A. J. Park, V. Spicer, H. H. Tsang, K. Behiels and J. Song, "Discovering Crime Trends and Patterns Using Three-Dimensional Visual Analytics," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0545-0549, doi:10.1109/IEMCON.2019.8936251.

[5] P. Sharma, S. Khater and S. Sharma, "Data visualization of crimes in a city using machine learning," 2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH), 2020, pp. 55-60, doi: 10.1109/INBUSH46973.2020.9392166.

[6] M. A. Boni and M. S. Gerber, "Area-Specific Crime Prediction Models," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 671-676, doi: 10.1109/ICMLA.2016.0118.

[7] S. Mukherjee and K. Sarkar, "Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas," 2020 IEEE Calcutta Conference (CALCON), 2020, pp. 444-450, doi: 10.1109/CALCON49167.2020.9106554.

[8] R. Yadav and S. Kumari Sheoran, "Crime Prediction Using Auto Regression Techniques for Time Series Data," 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), 2018, pp. 1-5, doi: 10.1109/ICRAIE.2018.8710407.

[9] Hana Anber, Akram Salah and A. A. Abd El-Aziz, “A Literature Review on Twitter Data Analysis,” 2016 International Journal of Computer and Electrical Engineering, 2016, doi: 10.17706/ijcee.2016.8.3.241-249

[10] Heidi Cohen, “How reliable is Twitter?”, 2013 Accessed on: 01/12/2022 [Online], Available: https://heidicohen.com/reliable-twitter-research/

[11] N Schwitter, “GOING DIGITAL Web data collection using Twitter as an example”, 2020, Accessed on: 01/12/2022 [Online] Available: www.oxfam.org [12] “Twitter API documentation” 2022, Twitter Developer Platform, Accessed on: 01/15/2022 [Online] Available: https://developer.twitter.com/en/docs/developerportal/overview

[13] Zhang AJ, Albrecht L, Scott SD, “Using Twitter for Data Collection With Health-Care Consumers: A Scoping Review”, 2017 International Journal of Qualitative Methods, 2017, doi: 10.1177/1609406917750782

[14] Hana Anber, Akram Salah, A. A. Abd El-Aziz, “A Literature Review on Twitter Data Analysis” International Journal of Computer and Electrical Engineering 2016, doi: 10.17706/ijcee.2016.8.3.241-24

[15] D. M. Raza and D. B. Victor, "Data mining and Region Prediction Based on Crime Using Random Forest," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 980-987, doi: 10.1109/ICAIS50930.2021.9395989.