Clustering the Dataset for Better Crime Understanding in NY

Havan Patel
havangpatel@lewisu.edu
DATA-51000-002 Fall- 2022
Data Mining and Analytics
Lewis University

I Introduction

In this project we are going to use and analyze the open data about crime of New York City- one of the largest and influential American metropolis. New York is composed of many neighborhoods scattered among the city's five boroughs: Manhattan, Brooklyn, The Bronx, Queens, and Staten Island. New York is also one of the populous cities in the country and it is evident that lots of crimes will occur. This report, we will look at the complaints made to NYPD.[1] All the data in this dataset are from 2006 to 2019.

Th intent of the report is to find the groups of complaints by different such features that were provided in the dataset. By clustering or grouping such types like age, sex, location, crimes etc. we can identify the criminal activities. Each cluster contains different characteristics, and behavior and in some case, they have similarities with each other. We are also going to lower the dimensions of the space and work with that data for our cluster analysis. We will be looking at two different techniques for clustering the data and we will compare results with each other.

The future sections of this report shall describe the preprocessing of data, two different cluster techniques, results of both methods and conclusion. In section II we will process of data cleaning and analysis to better understand. Section III will describe how the dataset was used after preprocessing, where we utilize the two different techniques. In Section IV, the result obtained from the techniques are discussed. Finally, section V Provides the conclusion.

II Data Processing (cleaning of data)

In the raw dataset it included more than 5+ million instance and 30+ features [1]. Because this is too much data to process all at once and there may also be some information that we may not need. Therefore, we had to clean the data and we only used data that was dated in 2019 and complaints that were made only in Brooklyn. As we know that in real world, we collection so many data and they all are not always needed for the purpose in which they are using the data. In this dataset we have 30+ features and a lot of them are not needed for our purpose because we know they will not be of any value for us. Therefore, we only selected features that were important to use and for our use case. For this dataset we used and filtered on data from 2019 and so on. We also selected complaints there were made in Brooklyn because it contains the highest number of data as per Fig 1. For this report we used python to do our preprocessing and clustering techniques. The preprocessing is broken in to 3 parts.

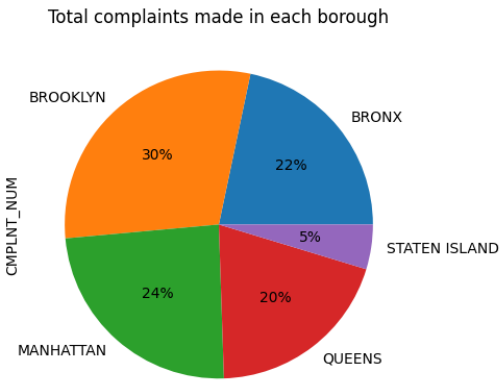


Fig 1. Complaints made in each borough

The first process is to analyze the dataset and extract the features that made the most sense for the report and all the data were mostly nominal features. We then drop the other columns that were constant, and some unnecessary columns were removed like Ids, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, etc.

The second process was to remove any rows that contained null, or values that didn't make sense for that column. For example, some values in Age column contained negative numbers and we know we can't have negative age so columns like those were removed from the dataset.

The last process was to finally reduce the number of rows by filtering the dataset and observing the dataset using different graphs. The reducing of the dataset was very necessary because there were 5+ million rows and contained way too many features. By doing this we can also reduce the overhead on our machine for computation and for the different clustering algorithms computations. For example, we can see below in Fig 2 and Fig. 3 that what are the top 10 most offense and location of complaints

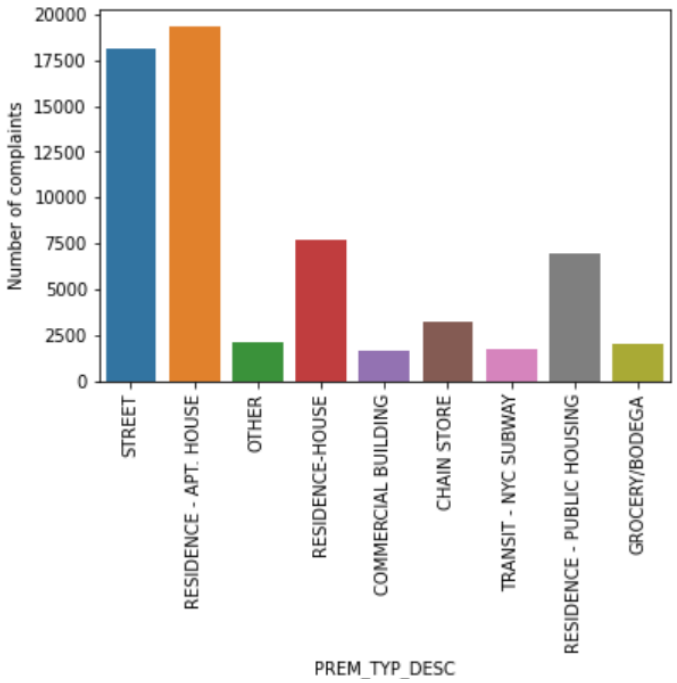


Fig. 2 Where Total Complaints were made in Brooklyn

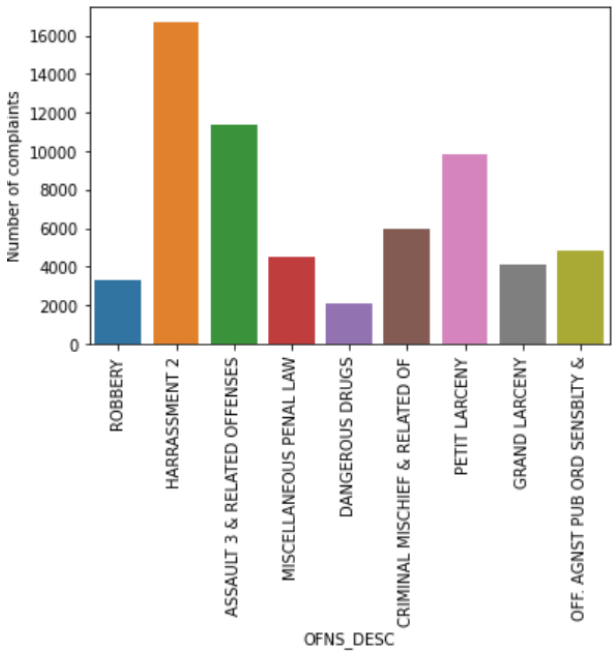


Fig. 3 Total Complaints made for offense type in Brooklyn

III DATA DESCRIPTION

As mentioned above that this data includes all valid felony, misdemeanor, and violation crimes reported to the New York Police Department(NYPD) from 2006 to 2019. "This data was manually extracted every quarter and reviewed by the Office of management Analysis and Planning".[1] Below are the following attributed that were extracted that were important for clustering analysis.

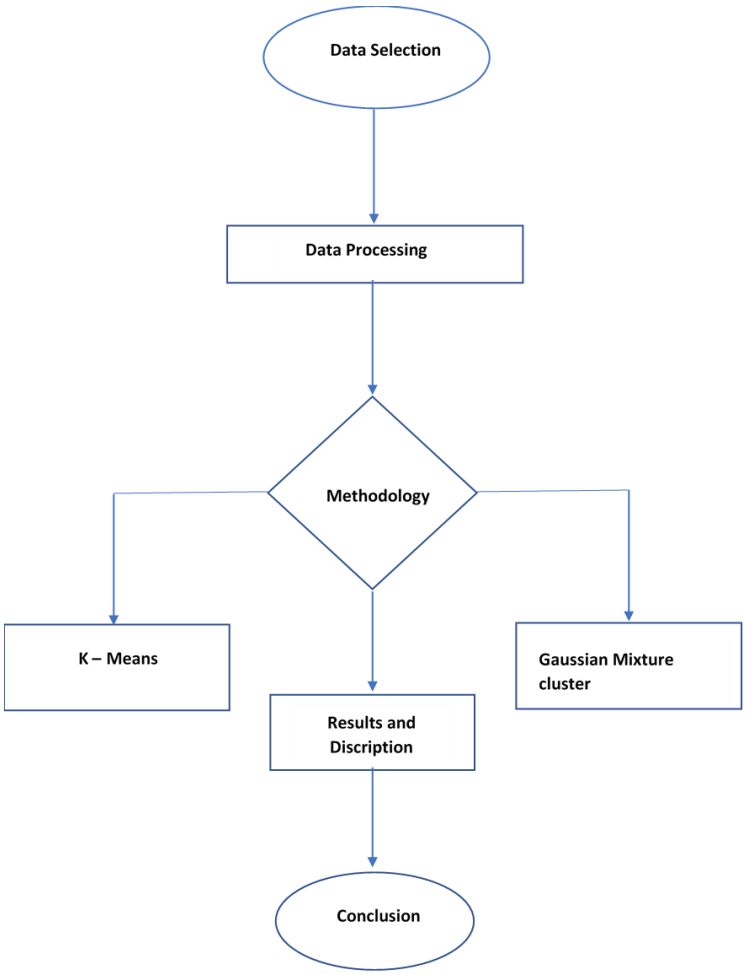
TABLE I. DATASET ATTRIBUTES

Attribute	Type	Example Value	Descripton
OFNS_DESC	Nominal (string)	Robbery	Offense Description
CRM_ATPT_CPTD_CD	Nominal (string)	Completed	Was crime completed or not
LAW_CAT_CD	Nominal (string)	FELONY	Offense level
BORO_NM	Nominal (string)	BROOKLYN	Where the incident occured
PREM_TYP_DESC	Nominal (string)	Street	Specific description of premises.
SUSP_AGE_GROUP	Nominal (string)	18-24	Suspect's age
SUSP_RACE	Nominal (string)	ASIAN / PACIFIC ISLANDER	Suspect's race

Attribute	Type	Example Value	Descripton
SUSP_SEX	Nominal (string)	M	Suspect's sex
PATROL_BORO	Nominal (string)	PATROL BORO BKLYN SOUTH	patrol borough in which the incident occurred
VIC_AGE_GROUP	Nominal (string)	25-44	Victim's age
VIC_RACE	Nominal (string)	White	Victim's race
VIC_SEX	Nominal (string)	F	Victim's sex

Again, if we refer to Fig. 2 and Fig. 3, we can conclude that a lot of crimes are happening in Streets and in Residential apartment, and with that a lot of harassment and level 3 assaults are occurring. And we know there are many different types of offense and location at which such horrible events occur. Therefore, it only makes sense to filter and cluster on data that occurs the most, so we took top 10 different events that occur the most.

Pictorial Representation Of The Project



IV METHODOLOGY

IV(I) K-MEANS

In this report one of the methods, we are going to utilize is the K-Means techniques to form out clusters. Since our dataset contains most of the data in nominal type it is hard for this algorithm to provide us with best results. The k-means algorithm works only with numeric data sets. [2] Therefore, we must transform our nominal dataset into a new feature space. By doing this it will increase our features from (95162, 13) to (95162, 177). This is one of the cons where it increases our space which would produce not good results. This can lead to curse of dimensionality as we learned in week 3. However, there are such methods where we can utilize to combat this issue. One such technique is called PCA, which reduces the dimension of large dataset.

To utilize K-Means in our dataset we need to find out the right amount of cluster to use. To figure that our we will use The Elbow Method and The Silhouette Method. For this method, we Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow. [3] In fig. 4 we can see the optimal K is 5 because it forms the elbow at that point.

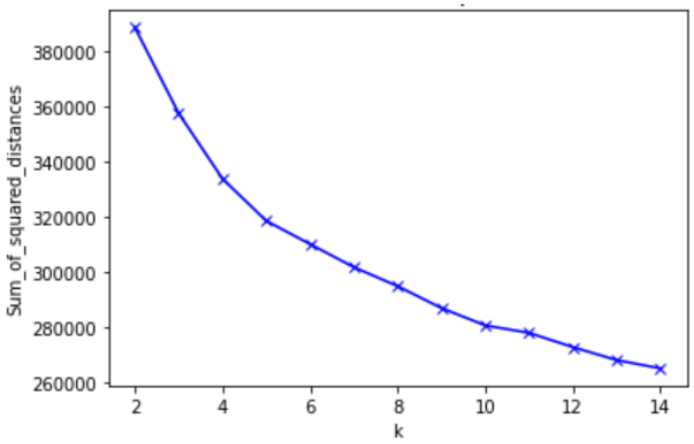


Fig. 4 Elbow Method for Optimal k

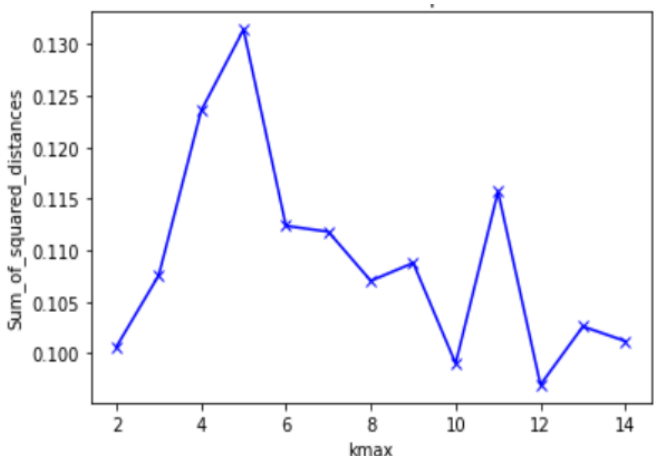


Fig. 5 silhouette Method for Optimal k

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is put in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters. In this case we look at the peak of the points and in Fig 5 we can clearly see the optimal K is we choose is 5. Therefore, we can apply K-Means on our dataset with cluster size of 5. However, we still need to reduce the dimension of our dataset so we will have to use PCA. Again, for PCA we still need to find the best optimal number of components to be able to provide us with the best result. To find optimal number of clusters for PCA we will need to compute the cumulative explained variance of our dataset. We can see that in Fig. 6 the best optimal cluster to size is 30 because it covers 95% of variance.

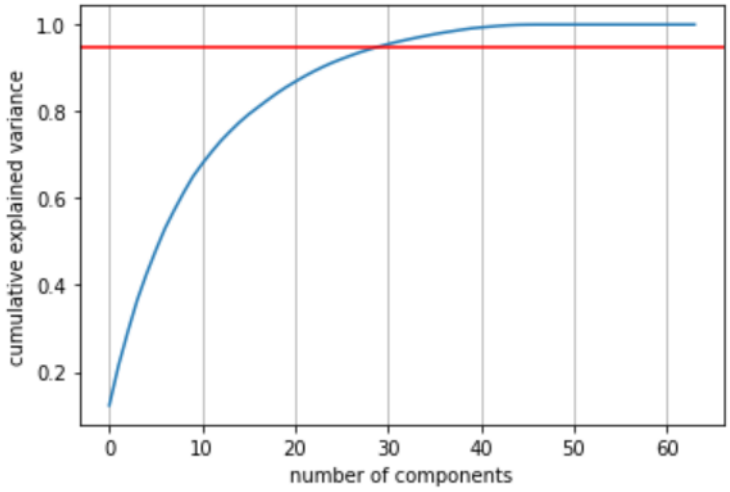


Fig. 6 PCA for Optimal k

After figuring out all the optimal amount of cluster for K-Means and PCA we were able to form this cluster show in Figure 7. We will discuss the result in later section of the report.

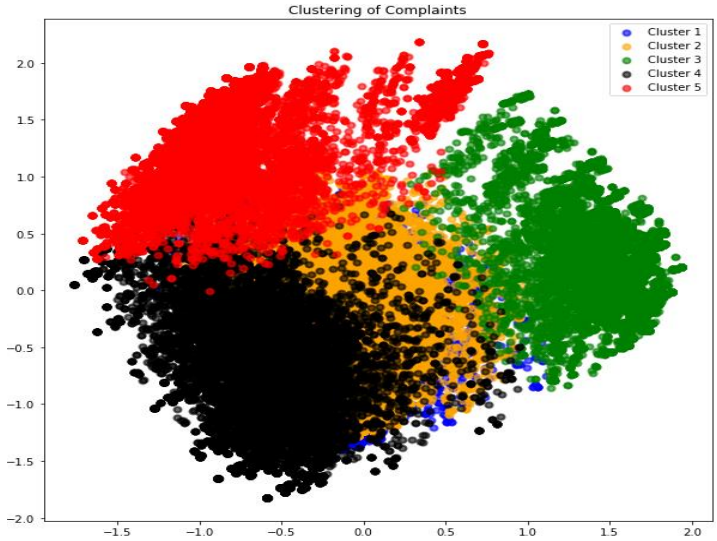


Fig. 7 Using K-Means to form a cluster of complaints

IV(II) GAUSSIAN MIXTURE CLUSTER

The Gaussian Mixture Model(GMM) can be viewed as an extension of the idea behind K-Means, but it can be a powerful tool for estimation beyond only clustering. GMM tries to find mixture of multi-dimensional Gaussian probability distribution that best model any dataset. Gaussian Mixture uses an expectation-maximization approach which chooses starting guesses for the location and shape. Then, repeats to find E for each point and finds the weights encoding the probability of membership in each cluster. Next it tries to find M- for each cluster and update the location, normalization and shape based on all data points making used of the weights.[4] Sci-Kit also has example on how to use this technique and more about the library.[5]

Since we know the from previous steps what the optimal number of clusters will be; we don't need to recalculate it again. This process will be very similar to what the K-Means process was.

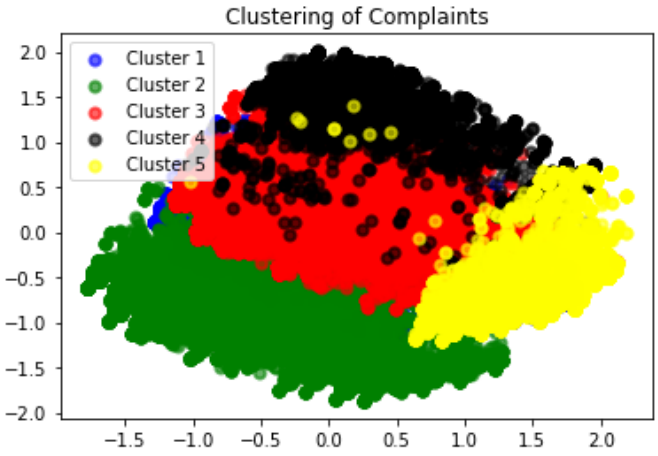


Fig.8

Fig. 8 Using GMM to form a cluster of complaints

V Results and Discussion

For both clustering techniques due to our data being mostly nominal we first had to transform it to numeric encoding. Mentioned above Fig.4 and 5 we can clearly see the optimal value for K by measuring the elbow technique. As this have been explained above, we won't go into too much detail in here.

Let's look at the cluster from figure 7 and 8. We can clearly see the difference here between two clusters. Fig. 7, which is formed using K-Means, we see a lot of overlap and lots of overfitting as well. Compared that to Fig. 8 which is formed using GMM, we also see overfitting and overlapping but not as much as we see in Fig.7.

Now let's see how this clusters are plotted with different features in the dataset. Due to having many features we will only look at 2 features, but the rest of the features are plotted in the Jupyter Notebook. From both fig. 9 and 10 we see similar results it's just that some traits are put into different clusters. If we look at cluster 4 from fig. 9 and cluster 3 from fig 10, they are both the same. They are just grouped into different clusters. Cluster 0 from left figure and 2 from right figure we see it includes Residence area. Cluster 0 on below on left and cluster 2 from below figure from right has cluster mostly formed with Assault 3, Petit Larceny. One more time we see both cluster techniques form similar cluster but different groups it in different cluster. However, Cluster 1 in both figures below is formed of Harassment 2. By looking at this we can form the group of sex, location, assault, race, etc. in a cluster to better get an idea. By doing this we can form an idea of how and what are the areas where crimes are completed? Also, we can get better understanding of what kind of crime will mostly occur(Assault, robbery, etc.) if crime happened at that location. We can also get a sense of who would be most likely to commit a crime in that location, suspects average age, race and who thy mostly target.

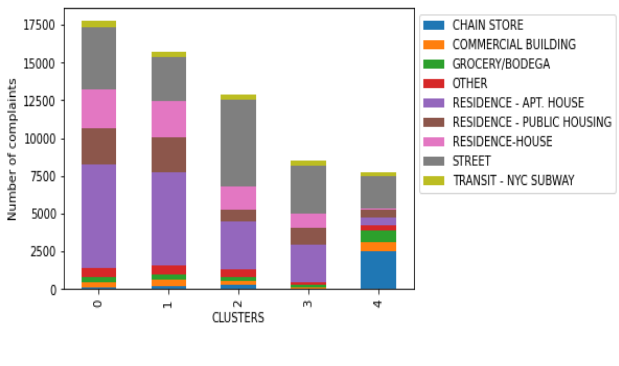


Fig. 9 Complaints clustered on Location categories with K-Means

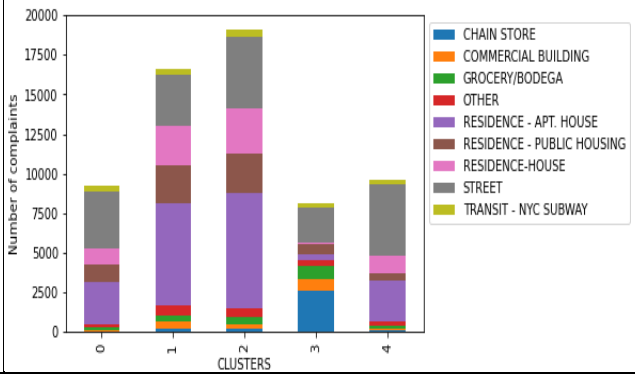


Fig. 10 Complaints clustered on Location categories with GMM

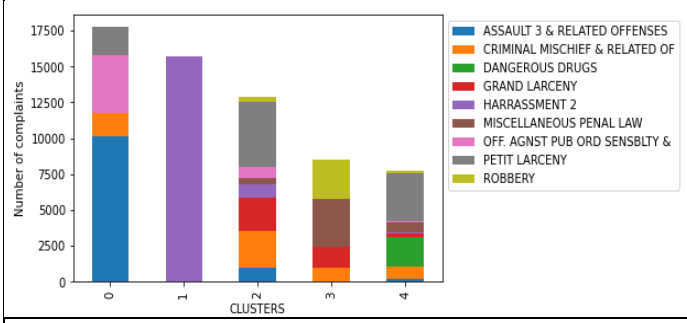


Fig. 11 Complaints clustered on offense categories with K-Means

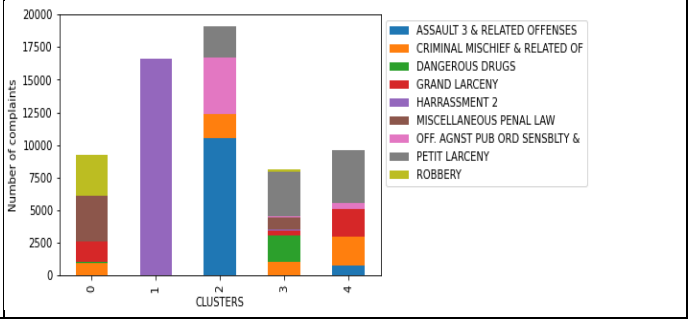


Fig. 12 Complaints clustered on offense categories with GMM

VI CONCLUSIONS

In this paper, we have analyzed the NYPD Crime Complaint Data Historic and then we have transformed the dataset so that we can use to form clusters that are meaningful because in real world data is not always perfect and we have to modify it in meaningful way to fit out needs. We looked at K-Means and Gaussian Mixture Model to form the clusters or groupings of complaints on different features. To use K-Means cluster we first must figure out what the optimal k value should be, and we figure that out by using either the elbow method or the silhouette value measure. But we utilized both techniques for comparison and surety that it will gives close to same optimal k cluster to use. Then we used PCA on it to reduce the dimensions as we see before that due to our data being nominal it would be increase the size of the feature when applying the encoding.

The process of clustering can help the NYPD a lot by knowing in which city/borough the most criminal activities occur. They can also find out the demographics such as, what's the age, sex, or race of suspect at that location. Based on all the facts, NYPD can create a meaningful infrastructure and budget to deploy the forces in those areas to stop such criminal activities from happening. This way NYPD can provide a safety and safe environment for the common public. This way not only it will benefit the public to have a piece of mind, but the police will get to know their community and get closer to them and create a bond. This will also change the law and order for that area so that it becomes livable area for everyone.

Lastly, we can utilize this kind of information to better understand the situation and asses it to make difference in the community. If this tool is used right, it is only for the benefit of the people and in a way if you look at it, it will help the economy as more business will open in those crime area if the crime rates are controlled.

REFERENCES

[1] Bruna Mendes, "NYPD Crime Complaint Data Historic (2006-2019)," 16 February 1955. *https://www.kaggle.com/datasets/brunacmendes/nypd-complaint-data-historic-20062019*

[2] https://go.documentation.sas.com/doc/en/emhpprcref/14.2/emhpprcref_hpclus_details06.htm#:~:text=The%20k%2Dmeans%20algorithm%20works,does%20not%20produce%20good%20results. I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Mahendru, Khyati. "How to Determine the Optimal K for K-Means?" *Medium*, Analytics Vidhya, 17 June 2019, https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb.

[4] VanderPlas, Jake. "In Depth: Gaussian Mixture Models." *In Depth: Gaussian Mixture Models | Python Data Science Handbook*, https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html.

[5] "2.1. Gaussian Mixture Models." *Scikit*, https://scikit-learn.org/stable/modules/mixture.html.