Final Report

Group Name : **Goal Diggers**

Aparajita Satish Ramanathan
Computer Science
UC, Riverside
Riverside, CA, USA
asati004@ucr.edu
Sid: 862324503

Harshitha Sarva
Computer Science
UC, Riverside
Riverside, CA, USA
hsarv001@ucr.edu
Sid: 862323552

Malmurugan Sukumar
Computer Science
UC, Riverside
Riverside, CA, USA
msuku002@ucr.edu
Sid: 862246504

Rohit Ramesh Kumashi
Computer Science
UC, Riverside
Riverside, CA, USA
rkuma069@ucr.edu
Sid: 862324644

Spoorthi Badikala
Computer Science
UC, Riverside
Riverside, CA, USA
sbadi006@ucr.edu
Sid: 862324457

1. INTRODUCTION:

The importance of safety has always been a significant concern for the general public. Crime has hit new heights and innovations in recent decades. [1] While the volume of Big Data increases, so do the complexity and the relationships underneath the data. This project focuses on manipulating Big Data Techniques and algorithms to perform Smart Data analysis of crimes in Los Angeles. The main idea stems from promoting a better lifestyle for people to visualize their area's safety. This model can promote women's safety in unsafe neighborhoods by sending them alerts when they enter the danger zone to be more cautious in future modules.

The paper proposes developing an OLAP (Online Analytical Processing) System to summarize and comprehend the crime patterns, and we use [2] clustering algorithms to locate high-risk areas of the city. The essence of the project lies in the features that it operates from the dataset and the algorithms it uses to evaluate. The humongous data is from a set of actual crimes reported in Los Angeles, and it is then processed and uses Spark Database for storage and retrieval.

2. MOTIVATION:

In this project, our objective is to analyze the crime data containing various crimes that occurred in Los Angeles for a period of time. We want to live in a world that is ideally safe and free of crimes. Unfortunately, that is not the case in the real world.

While such crimes are out of our control, We can use the data provided by the LAPD to aid the authorities by identifying the crime hotspots in Los Angeles and studying the patterns involved in these crimes, such as the type of neighborhood where these crimes occur. Through this analysis, we identify the patterns behind these crimes and go into further detail by pinpointing the frequent type of crimes, identifying the top most dangerous areas, and specifying times with higher crime incidents. We then use clustering algorithms to highlight the areas with the highest crime rates and visualize the data in an informative manner.

There are many [2][3] Clustering algorithms to choose from, and no single best clustering algorithm for all cases. We will use multiple clustering algorithms to determine the most effective. We will also use Big Data tools such as pySpark to preprocess the data and Spark SQL for OLAP, and these tools will enable us to process complex data from large datasets quickly and efficiently, thereby saving valuable time.

3. LITERATURE SURVEY:

3.1. Crime Datasets

Studying crime and its occurrences help in driving policies about where to position law enforcement effectively. Understanding the relationship between the spatial and temporal aspects is important when doing an analysis of crime datasets [1][2]. LAPD crime data was obtained from original crime reports typed on paper, and there may be some inaccuracies within the data. The data has more than a million rows that delineate the crime incident. It has 28 attributes that contain numerical, text, and date-time data. https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z

3.2. Big data tools:

This project makes use of Spark SQL which provides Spark with more information about the structure of data and the computations[5]. Since the dataset has a multitude of different attributes and values for the same crime, this project performs a union function of the data loaded. Later the OLAP is performed through Spark SQL which will be explained in detail in the next section.

3.3 OLAP:

Analyzing the data at hand plays an important role in constructing algorithms. This helps us construct an algorithm that uses this information to produce optimal results. [8] the paper outlines how OLAP technology (Online Analytical Processing) is one of the most efficient methods for analyzing large-scale data.

[10] the paper shows how Python libraries like Pandas and Seaborn are the most powerful analytical tools available. From the above research, we proceeded to incorporate OLAP using the above libraries available in the PySpark MLib tool.

3.4. Clustering Algorithms:

The k-means algorithm, a common classification approach, is used here to group or cluster crime data with similar features. Paper [11] demonstrates that applying k-means clustering algorithms on crime datasets obtains a significant crime-prediction accuracy. In the spatial crime datasets, where the class area spans or covers more, this feature of the k-means calculation makes it more appropriate than other approaches.

Other two clustering algorithms that have been proven efficient in recent times would be Gaussian Mixture Model (GMM)[17] and Bisecting KMeans [18]. All three algorithms will be implemented and compared to choose the optimal one.

3.5. Evaluation Metrics

Previous works [14][15] on multiple clustering algorithms built upon different datasets have ensured that the Elbow method is one of the most optimal methods in deciding the number of clusters. In our project, we will use the elbow method to fix the number of clusters for our baseline and improve clustering algorithms.

4 PROJECT COMPONENTS:

4.1. OLAP:

Before we go about analyzing the data set, we performed pre-processing and cleaning of the crime dataset. Such tasks include but are not limited to loading the dataset from the provider into the Spark Session, performing data type correctness checks for each category, checking for missing values for each category, and filling/dropping the rows that have missing values.

The analytical processing that we have performed until now will be similar to what can be expected in future work. We have analyzed aspects and questions like the number of crime incidents by year, month, and time of the day, distribution of crimes committed by area or the number of crimes by category. We have utilized Spark SQL to achieve our analysis, and have employed matplotlib and seaborn to aid in the making of visualizations. Some of these questions are intuitively depicted as shown below.

4.1.1 How different ages are affected by crime?

Amongst other questions that prevail in the OLAP Analysis, this segment involves analysis based on age group. The initial dataset is thoroughly cleaned concerning people's age and then a spark connection is established. Initially, the victim's age is pulled from the dataset where for every distinct age that is present we get the count and categories the data into three sub-segment as follows:

Ages 1-17: Minors, people who are still under parents' supervision and go to school ranging from kindergarten to high school.

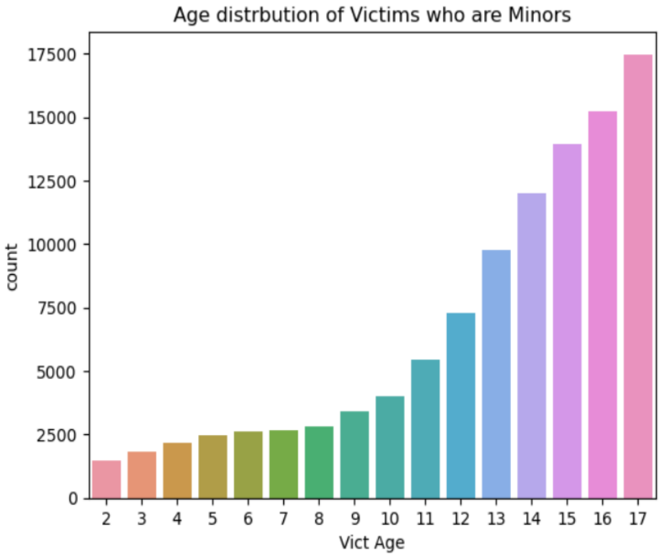


Fig: 4.1.1.a Minor Age Vs Count of Crime.

Ages 18-49: Majors or Adults, This is where the earning population lies.

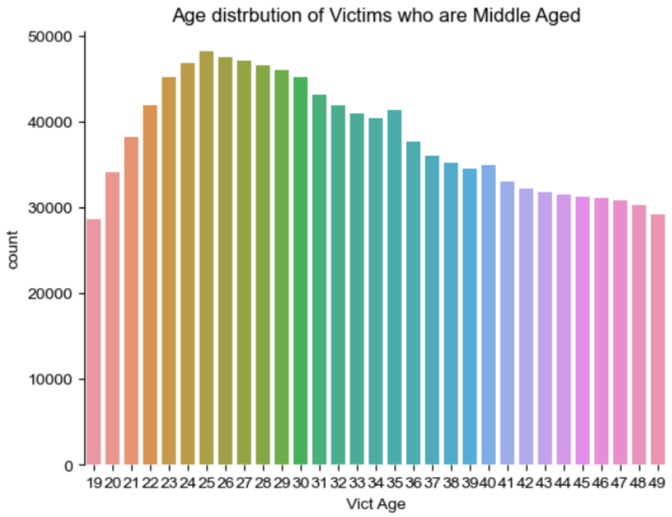


Fig: 4.1.1.b Adult Age Vs Count of Crime.

Ages 50 onwards: They are considered seniors, consisting of retired or near retirement population.

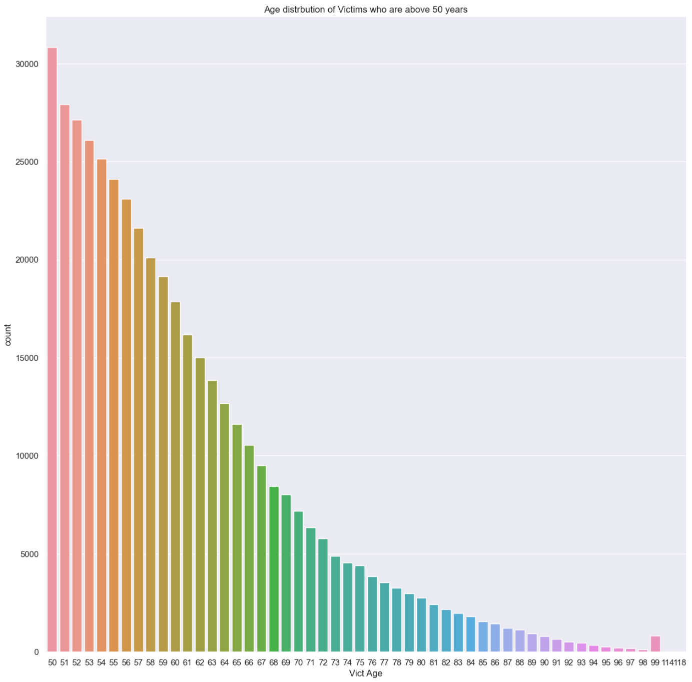


Fig: 4.1.1.c Senior Age Vs Count of Crime.

The graph below each category shows the number of crimes that occurred in Los Angeles for each category based on age.

Not so surprisingly crime among Majors or Adults is skyrocketing, but what is concerning was the number of crime that was bestowed against children under the age of 18 with parental supervision.

The next part will follow further depth analysis of crime amongst minors based on area name and age, so that we can see in which area the minors are more affected.

Minor Aged OLAP Analysis:

The top two victims' ages under the Minor category from the above graph as taken and the count is compared with the area name.

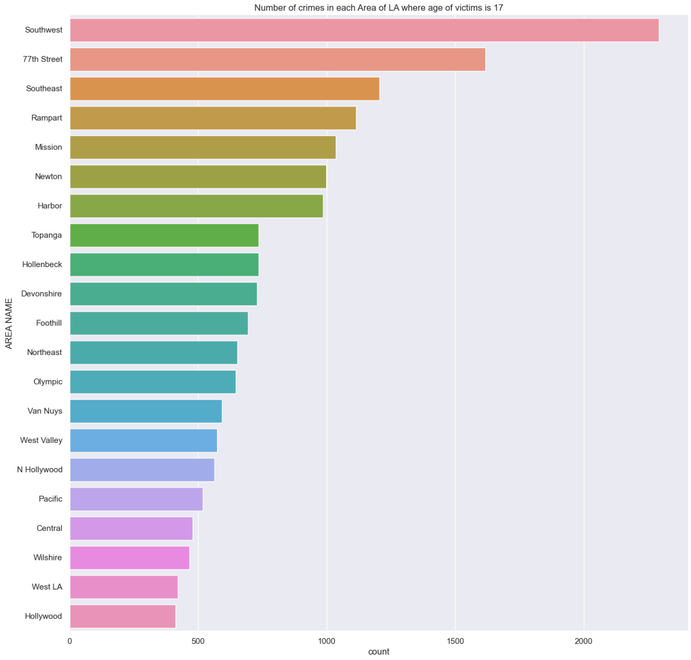


Fig: 4.1.1.d Area of crime Vs crime count for 17 yr old.

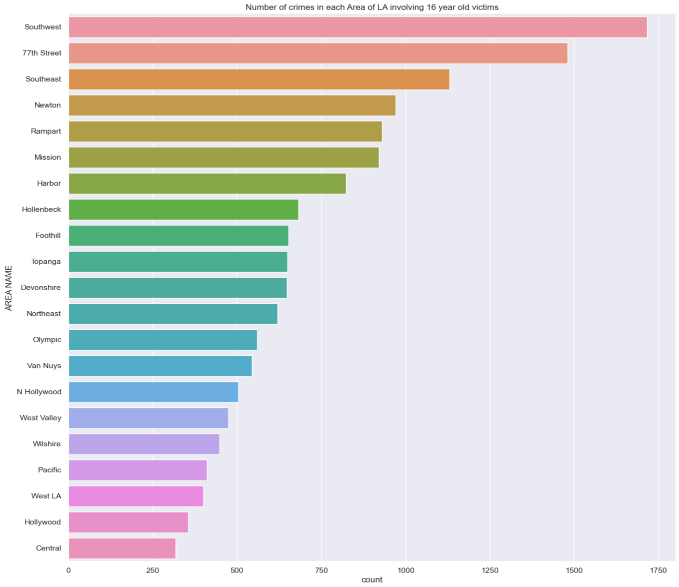


Fig: 4.1.1.b Area of crime Vs crime count for 16 yr old

The ages 17 and 16 are most affected especially in South West LA and 77 Street where the occurrence of the crime is highest. Generating these analyses is highly necessary as we can have an idea of area safety and be more conscious in these places.

4.1.2 Crime numbers by various time periods - year

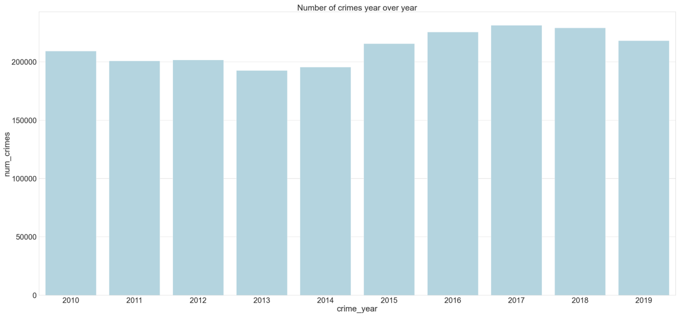


Fig 4.1.2 Crime incidents over years

The bar graph above shows that the number of crime incidents has generally increased over the passage of time. The number of crime incidents reached its nadir in 2013, following which the crime incidents have obeyed the general trend of increasing numbers. More in-depth analysis is needed over smaller time frames to understand crimes and this is explained below.

4.1.3 Crime numbers by various time periods - month

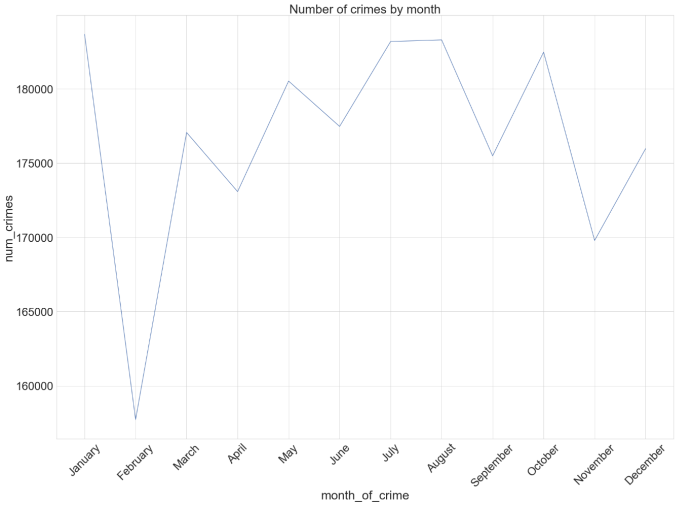


Fig 4.1.3 Crime incidents by month

The above line chart depicts the historical monthly crime incidents. It was observed in general that the period from November to March was a lean period on average in terms of the volume of criminal activities compared to the June to September period where the average was the highest. This can possibly be attributed to the fact that June to September months are climatically pleasant in comparison to other periods, leading to more interactions between victims and criminals. However, let's also explore the crime number by the time of the day.

4.1.4 Crime numbers by various time periods - time of day

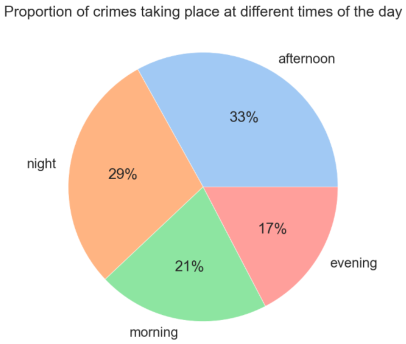
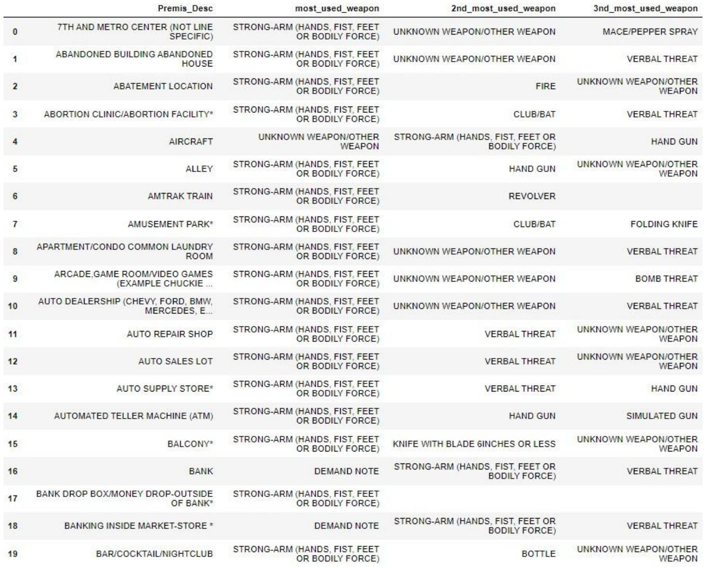


Fig 4.1.4 Proportion of crimes according to time of day

The pie chart shows that a third of the crimes take place in the afternoon, followed closely by those that occur at night times. Such times account for most crimes as these are usually when people either return from the office, travel around, or do grocery shopping.

1.5 What is the choice of weapon used on different premises



	Premis_Desc	most_used_weapon	2nd_most_used_weapon	3nd_most_used_weapon
0	7TH AND METRO CENTER (NOT LINE SPECIFIC)	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	UNKNOWN WEAPON/OTHER WEAPON	MACE/PEPPER SPRAY
1	ABANDONED BUILDING ABANDONED HOUSE	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	UNKNOWN WEAPON/OTHER WEAPON	VERBAL THREAT
2	ABATEMENT LOCATION	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	FIRE	UNKNOWN WEAPON/OTHER WEAPON
3	ABORTION CLINIC/ABORTION FACILITY*	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	CLUB/BAT	VERBAL THREAT
4	AIRCRAFT	UNKNOWN WEAPON/OTHER WEAPON	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	HAND GUN
5	ALLEY	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	HAND GUN	UNKNOWN WEAPON/OTHER WEAPON
6	AMTRAK TRAIN	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	REVOLVER	
7	AMUSEMENT PARK*	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	CLUB/BAT	FOLDING KNIFE
8	APARTMENT/CONDO COMMON LAUNDRY ROOM	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	UNKNOWN WEAPON/OTHER WEAPON	VERBAL THREAT
9	ARCADE,GAME ROOM/VIDEO GAMES (EXAMPLE CHUCKIE ...	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	UNKNOWN WEAPON/OTHER WEAPON	BOMB THREAT
10	AUTO DEALERSHIP (CHEVY, FORD, BMW, MERCEDES, E...	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	UNKNOWN WEAPON/OTHER WEAPON	VERBAL THREAT
11	AUTO REPAIR SHOP	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	VERBAL THREAT	UNKNOWN WEAPON/OTHER WEAPON
12	AUTO SALES LOT	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	VERBAL THREAT	UNKNOWN WEAPON/OTHER WEAPON
13	AUTO SUPPLY STORE*	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	VERBAL THREAT	HAND GUN
14	AUTOMATED TELLER MACHINE (ATM)	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	HAND GUN	SIMULATED GUN
15	BALCONY*	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	KNIFE WITH BLADE 6INCHES OR LESS	UNKNOWN WEAPON/OTHER WEAPON
16	BANK	DEMAND NOTE	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	VERBAL THREAT
17	BANK DROP BOX/MONEY DROP-OUTSIDE OF BANK*	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)		
18	BANKING INSIDE MARKET-STORE *	DEMAND NOTE	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	VERBAL THREAT
19	BAR/COCKTAIL/NIGHTCLUB	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	BOTTLE	UNKNOWN WEAPON/OTHER WEAPON

Fig 4.1.5 Top 3 Weapons of choice based on premise

For all the premise types available in Los Angeles, the above text table shows the 3 most used weapons while perpetrating the crime. While devising this table for some premise types, the top rank did not have a weapon used which makes little sense for the question, so we filtered those out. This knowledge can guide police to use tactics specific to the premise and weapon used. For eg. ATM machines had handguns and simulated guns used most often by criminals to commit crime in such places. Police can then man such areas more frequently and devise tactics to reduce crimes there.

4.1.6 What is the most typical crime inflicted on various victim demographics?

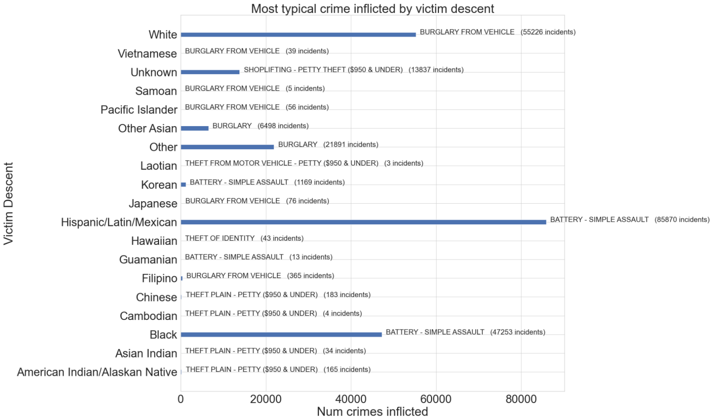


Fig 4.1.6 Most typical crime inflicted by victim descent

The bar plot on top shows the type of crime victims of varying demographics endure. People from the Black and Hispanic/Mexican community are often the target of crimes that involve battery and physical assault.

This knowledge can help the police to launch directed community programs that can help various demographics to be safe from such crimes.

4.1.7 Which Gender has more victims?

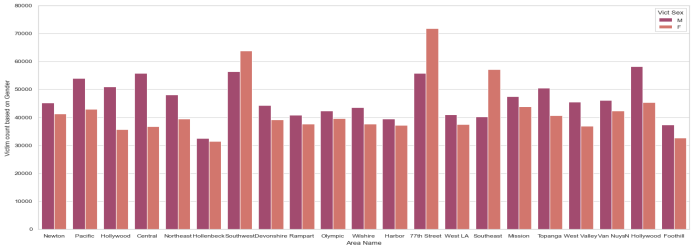


Fig 4.1.7: Gender Vs Locality

The above graph compares the number of people affected by crimes in each locality based on gender. Of the 21 localities we have sampled, 18 show that Males are generally subject to crimes compared to females, which is a significant majority. However, the dataset doesn't explicitly mention the male-to-female ratio in Los Angeles. As such, this data could be skewed if there exist more males than females in any particular locality.

4.1.8 How does race/ethnicity factor into crime victims?

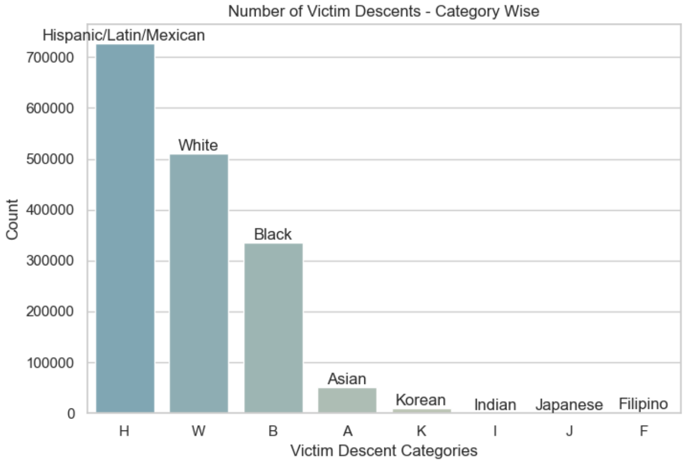


Fig 4.1.8: Number of Victim Decent - Category wise

This bar graph classifies the victims based on their race. It is clear that Hispanic/ Latin people have suffered the most number of crimes from this graph. The next two races that suffer the most crimes are White people and Black/African-American people. On the other hand, Filipinos were subject to the least number of crimes.

4.1.9 Which communities are targeted in a certain locality?

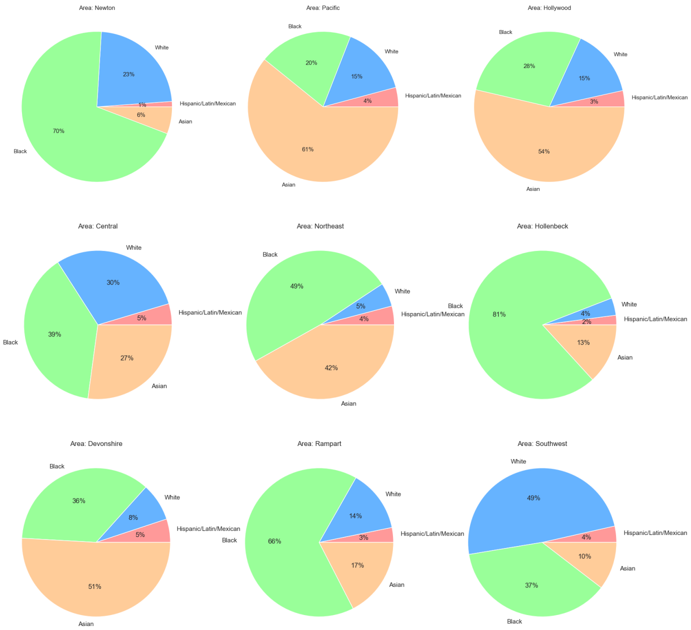


Fig 4.1.9: Number of victims vs Victim descent in each area

In this pie chart, we dig a little deeper and analyze which races suffer the most crimes in a particular locality. Black / African-American people seem to be the most targeted racial group for crimes in Newton, Central, Northeast, and Hollenbeck. Asians were victims of the most number of crimes in the Pacific and Hollywood. For instance, this data could be extremely useful for families that are looking to move into a particular neighborhood.

4.1.10 What are the most and least crime ridden areas in LA?

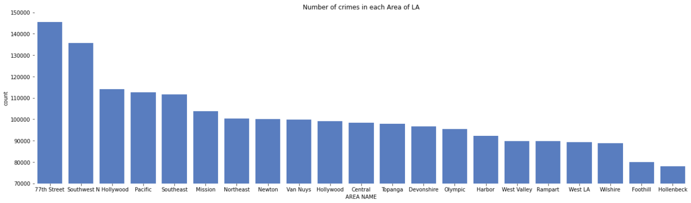


Fig 4.1.10: Total number of crimes recorded in each area from 2010-2019

The above graph shows the total number of crimes that took place in each area through 2010-2019 and orders them in descending order. The top three most crime-ridden areas in LA are

1. 77th Street
2. Southwest
3. N Hollywood

The areas which we can categorize safe in LA would be 'Foothill' and 'Hollenbeck'. We can also see that there is a difference of 67505 in the number of crimes between the safest and unsafe areas in the city. This major difference shows how crime is centered in particular areas and can be controlled.

4.1.11 Trends in crime rate by locality from 2010-2019

This plot depicts the crime rate in each locality between the years 2010 - 2019. From the graph, we can infer that the crime rate in general, rose from 2010 to 2018. The reforms brought about in 2018 seem to have worked as there was a decrease in crime rate in 2019 across most localities but we cannot confirm that this decrease in crime rate was maintained in more recent years as the dataset hasn't been updated since 2019. We can also see that a few areas have an increasing crime trend and need LAPD's attention. Concerning areas with increasing crime rate would be: 'Central', 'Pacific', 'Southeast' and 'Wilshire'.

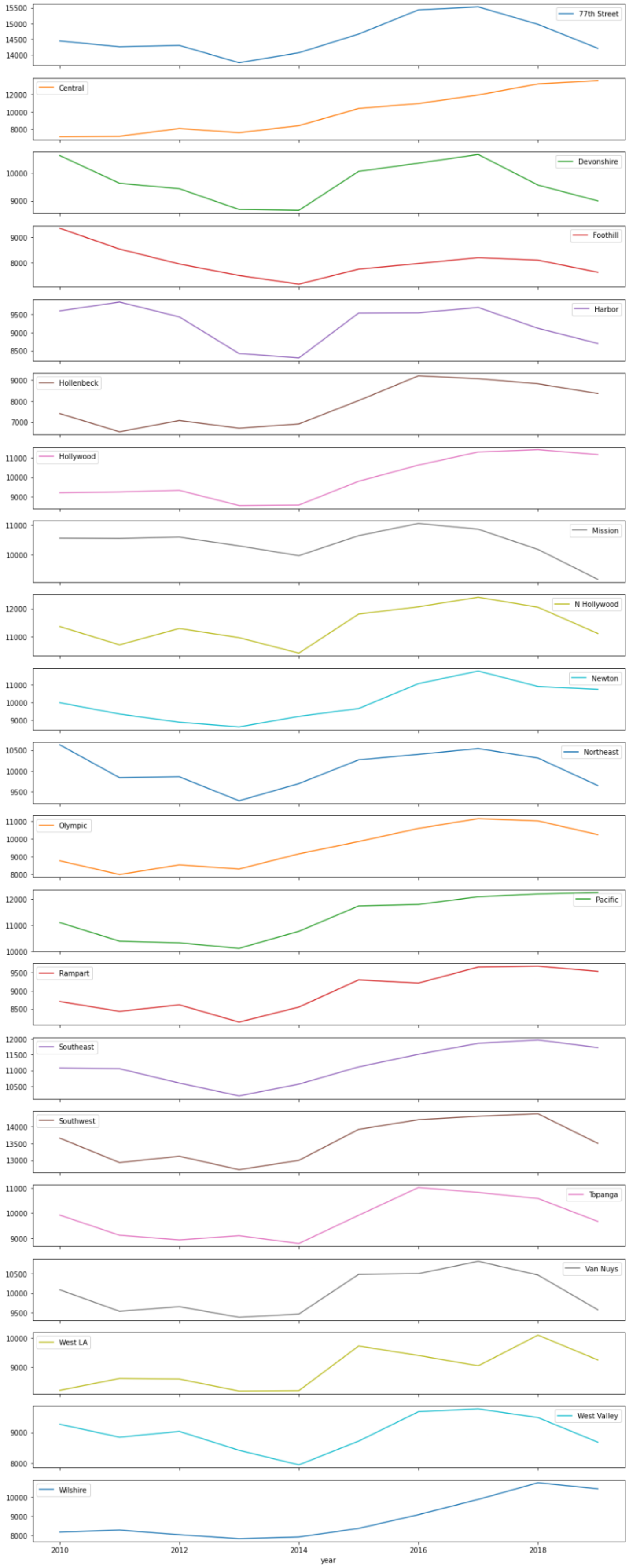


Fig 4.1.11: Number of Crimes in each area from 2010-2019

1.12 Trends in crime relative to areas:

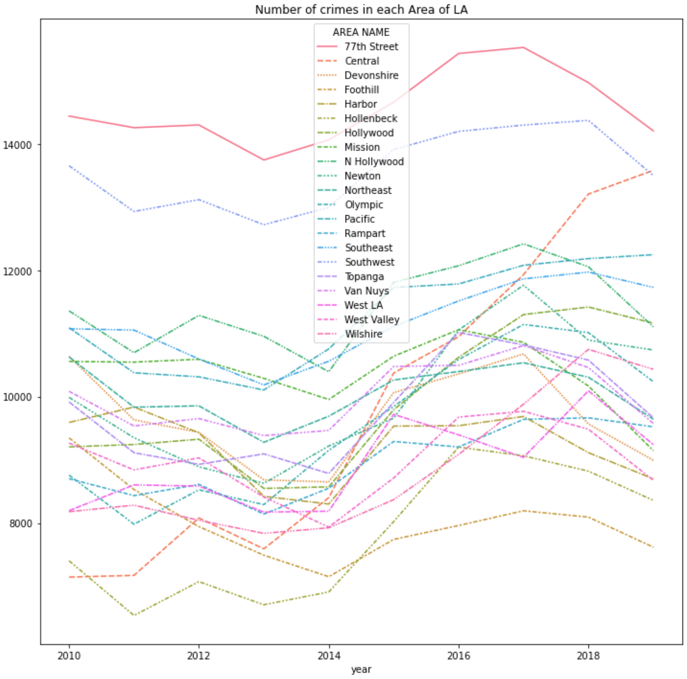


Fig 4.1.12 Crime trend in each area.

This plot shows the crime trend in each area relative to all areas in LA. We can see that two areas: '77th Street' and 'Southwest' have very high crime rates and started to show improvement after the year 2018. Even with a dip in the number of crimes, they record the highest number of crimes every year. We can also see that the 'Central' area had an alarming increase in crime rates and needs attention.

4.2. Clustering algorithms:

In this part of the project, we develop clustering algorithms to classify and visualize high crime locations in Los Angeles city.

We implement K-Means, Gaussian Mixture Model, and Bisecting K-Means clustering algorithms from Spark's MLlib library. With each of these clustering algorithms, we plan to group the crime data based on the latitude and longitude values and this will provide the top crime hotspots geographically.

4.2.1 Transforming and Scaling the data:

We use VectorAssembler to merge the latitude and longitude values and StandardScaler to scale our inputs before loading them to the machine learning models.

4.2.2 Clustering Models:

We fit all three algorithms onto the transformed and scaled data and get the predictions. We then split the output of each model based on the prediction class and then convert it into a Geo-Pandas dataframe.

After conversion, we implement the convex hull algorithm to generate polygon data for each cluster and then plot it on top of a 2D-Map layer of the LA city using Folium Maps. We also plot the cluster centers for each cluster so that we can provide some additional information about where the crimes are concentrated at each cluster.

Fig 4.2.1, 4.2.2, and 4.2.3 represent the outputs from K-Mean, GMM, and Bisecting K-Means respectively. Just from the figures, we can conclude that only the results from K-Means and Bi-Secting K-Means look promising as the areas don't overlap but the GMM clusters overlap and cannot be considered.

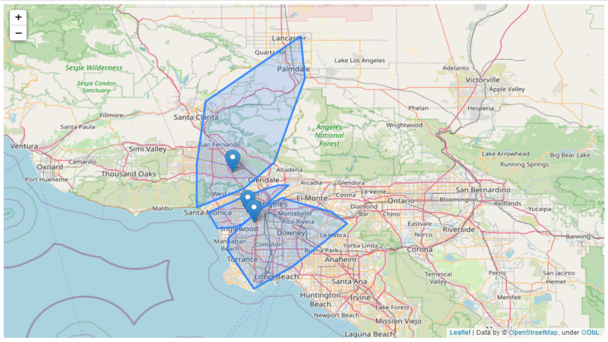


Fig 4.2.1 Cluster output: K-Means

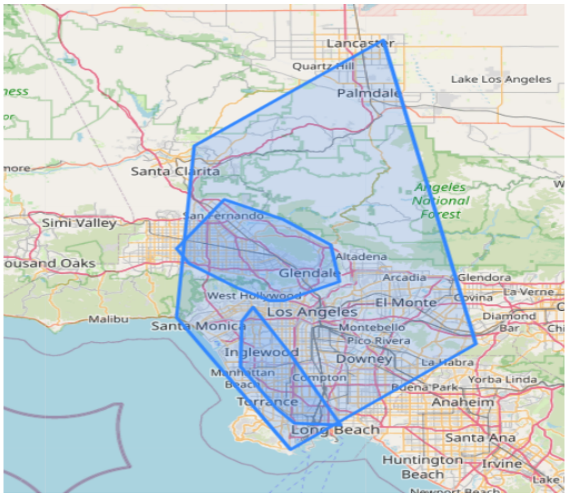


Fig 4.2.2 Cluster output: GMM

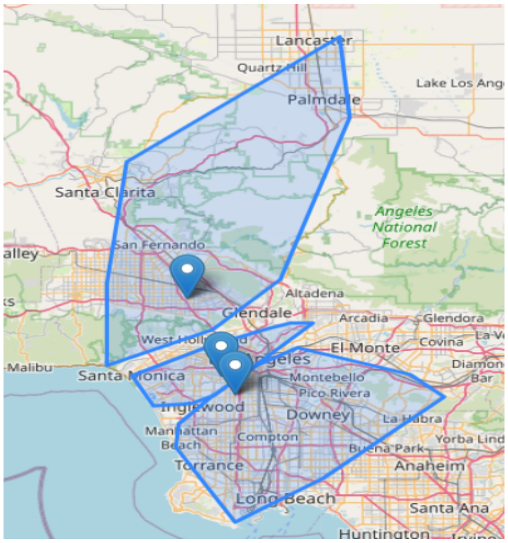


Fig 4.2.3 Cluster output : GMM

5 EVALUATION

5.1 Comparing tools for OLAP:

To get the best performing tool to do

5.2 Clustering Algorithms:

5.2.1. Deciding the optimal number of clusters.

We used the Elbow method to find our algorithm's optimal number of clusters. Elbow method calculates the sum of within-cluster variance, W, is calculated for clustering with different values of the number of clusters, k. The elbow method was used on the KMeans model and the resulting optimal value was used to build the rest of the clustering models.

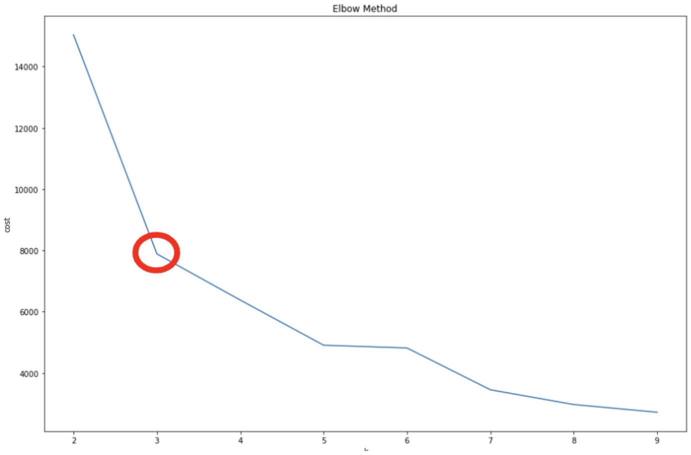


Fig 5.2.1: Elbow method for K-Means clustering

The above graph shows the cost of the model with various k values. We found our algorithm's optimal number of clusters with the Elbow method to be 3.

5.2.2. Comparing the performance of multiple algorithms

To evaluate the performance of clustering models, we compared the results of each clustering mechanism using pySpark's Clustering Evaluator. The clustering Evaluator will compute the Silhouette score which measures the goodness of a model by calculating the squared Euclidean distance between each cluster. Silhouette score ranges from [1,-1], where the scores range signifies:

1: Model's clusters are well apart from each other and clearly distinguished
0: Model's clusters are indifferent
-1: Model's clusters are assigned in the wrong way.

Time taken to build or execute each clustering model was also calculated to compare the models' performance.

Clustering Model	**Silhouette Score**	**Execution Time**
KMeans	0.67654377	3.293
GMM	0.66198917	9.219
Bisecting KMeans	0.67653679	5.562

Table 5.2.2: Clustering Models Evaluation

The above table shows the silhouette score for each clustering algorithm we built using PySpark's ml library and their execution time. Based on our results it is safe to conclude that KMeans is the best-performing model when compared to the rest.

5.3. Comparing tools performance

We evaluated the performance of the Big data tool (PySpark) used for building the clustering algorithms, by comparing it with other python tools like Pandas. We evaluated PySpark in data manipulation operations and machine learning operations

5.3.1 Data Manipulation operations:

To find the best performing dataframe, a set of data manipulation operations were performed on the Spark dataframe and Pandas dataframe. These operations were also used to perform OLAP.

Table 5.3.1 documents our results. From this, we can see that the PySpark dataframe outperformed the Pandas dataframe.

Operations	**Spark Execution Time**	**Pandas Execution Time**
Loading data from csv file	0.106	4.44
Filtering and renaming columns	0.017	0.10
Aggregate functions	0.49	0.08
Grouping and ordering data	0.016	0.36

Table 5.3.1: Spark vs Pandas in data manipulation operations

5.3.2: Machine Learning operations:

To compare PySpark and Pandas machine learning operations efficiency, three clustering models were built using both tools. The performance of each tool will be evaluated based on the time each tool took to generate a clustering model.

Model Name	**Spark Execution Time**	**Pandas Execution Time**
KMeans	3.293	1.959
Gaussian Mixture Model	9.219	1.271
Bisecting KMeaNS	5.562	2.265

Table 5.3.2: Spark vs Pandas in machine learning operations

From Table 5.3.2, we can say that Pandas ml library performed much better than Spark ml library.

CONCLUSION:

The project aims to provide a detailed analysis of the Los Angeles crime dataset. The components dictate the flow of the project where we first clean the data and process it through various PySpark operations. Then, OLAP is performed to answer some big questions which give in-depth knowledge about crimes in LA and also provide a clear visualization of each question it answers. The data clustering algorithms aim to back OLAP findings and provide clusters of crime hotspots in the city on a 2D interactive map. The correctness of the project is also verified with rigorous evaluation methods.

Through our findings, we can say that our project achieves all the goals it aimed to fulfill and provides a clear analysis that can be applied to other equivocal crime datasets. Work presented in this paper can also be extended in the future to develop applications for police departments and also to warn citizens. This can be made possible by integrating our findings in applications like google maps to warn citizens when they enter an unsafe area.

TEAM CONTRIBUTIONS:

Aparajita Satish Ramanathan: Worked on the introduction and parts of the report. Later, performed Data Cleaning and Pre-processing for OLAP based on Age. Used Spark SQL queries for Data Analysis and built a data visualization for deeper analysis. The first part of OLAP is a simple Analysis for the general Age Category using PySpark and Pandas and the Second part involves a multi-dimensional query and used matplotlib for visualization.

Malmurugan Sukumar: Worked and researched on developing clustering algorithms using PySpark MLib and

Pandas. Developed an interactive Folium map (spatial component) to visualize the clustered information on a 2D map. Proposed and implemented evaluation metrics for clustering models. Proposed a detailed evaluation plan and provided necessary code to test PySpark tool efficiency. Also documented the same in all deliverables.

Harshitha Sarva: Worked on data cleaning and preprocessing using PySpark and visualized the data using Pandas and Seaborn Libraries. Derived insights and presented visualizations on the dataset based on Victim's Sex, Race, and Locality. Later, documented the work in all deliverables.

Rohit Kumashi: Performed data cleaning and preprocessing using PySpark. Utilized Spark SQL and pandas to answer questions of interest and derive actionable insights from data pertaining to when, where, and how the crime took place. Built data visualizations using matplotlib to depict the analyses. Also documented the work in all deliverables.

Spoorthi Badikala: Cleaned and pre-processed data using PySpark. Queried cleaned data using Spark SQL and Pandas to draw insights from the data. Performed visualization using Pandas to summarize the data. Helped fellow teammates in brainstorming ideas for clustering algorithms and evaluation plans. Performed evaluation on data manipulation operations for both PySpark and Pandas dataframes. Documented the work done in all deliverables.

REFERENCES:

[1] Tompson L. , & Coupe T. (2018). Time and opportunity. In Bruinsma G. J. N. , & Johnson S. D. (Eds.), The Oxford handbook of environmental criminology (pp. 695–719). Oxford: Oxford University Press. www.doi.org/10.1093/oxfordhb/9780190279707.013.19.

[2] Steenbeek W. , & Weisburd D. (2016). Where the action is in crime? An examination of variability of crime across different spatial units in the Hague, 2001–2009. Journal of Quantitative Criminology, 32(3), 449–469. www.doi.org/10.1007/s10940-015-9276-3.

[3] Felson M. , & Poulsen E. (2003). Simple indicators of crime by time of day. International Journal of Forecasting, 19(4), 595–601. www.doi.org/10.1016/S0169-2070(03)00093-1.

[4] Lösel F. (2017). Evidence comes by replication, but needs differentiation: the reproducibility issue in science and its relevance for criminology. Journal of Experimental Criminology, 1–22. www.doi.org/10.1007/s11292-017-9297-z.

[5] Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, and Matei Zaharia. 2015. Spark SQL: Relational Data Processing in Spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15). Association for Computing Machinery, New York, NY, USA, 1383–1394.

[6] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: a unified engine for big data processing. Commun. ACM 59, 11 (November 2016), 56–65. https://doi.org/10.1145/2934664

[7] A. Yudovin. Essential Optimization Methods to Make Apache Spark Work Faster // ALTOROS. – 2019. – URL: https://www.altoros.com/research-papers/essential-optimization-methods-to-makeapache-spark-work-faster/

[8] Surajit Chaudhuri and Umeshwar Dayal. 1997. An overview of data warehousing and OLAP technology. SIGMOD Rec. 26, 1 (March 1997), 65–74. https://doi.org/10.1145/248603.248616

[9] McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics.

[10] Al-Janabi, Kadhim & Fatlawi, Hayder. (2010). Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures.

[11] Chetan G. Wadhai , Tiksha P. Kakade , Khushabu A. Bokde , Dnyaneshwari S. Tumsare, 2018, Crime Analysis Using K-Means Clustering, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 07, Issue 04 (April 2018)

[12] C. Naga Himaja1 , M. V. Ramakrishna
Analysis of Crime Data Using Kernel Based K-Nearest Neighbor(KNN)
http://www.joics.org/gallery/ics-1297_1.pdf

[13] J. R, S. Jagan, S. Khasim, L. Dhavamani, V. Mathiazhagan and D. Kumar Bagal, "Crime Visualization using A Novel GIS-Based Framework," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), 2021, pp. 1-5, doi: 10.1109/ICCICA52458.2021.969712

[14] D. Marutho, S. Hendra Handaka, E. Wijaya and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," 2018 International Seminar on Application for Technology of Information and Communication, 2018, pp. 533-538, doi: 10.1109/ISEMANTIC.2018.8549751.

[15] F. Liu and Y. Deng, "Determine the Number of Unknown Targets in Open World Based on Elbow Method," in IEEE Transactions on Fuzzy Systems, vol. 29, no. 5, pp. 986-995, May 2021, doi: 10.1109/TFUZZ.2020.2966182.

[16] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65 (1987)

[17] Bond, Stephen & Hoeffler, Anke & Temple, Jonathan. (2001). GMM Estimation and Empirical Growth Models.

[18] V. Rohilla, M. S. S. Kumar, S. Chakraborty and M. S. Singh, "Data Clustering using Bisecting K-Means," 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2019, pp. 80-83, doi: 10.1109/ICCCIS48478.2019.8974537.