

CRIME ANALYSIS SYSTEM

A PROJECT REPORT

Submitted by

Aswatha Narayanan. S

Jai Surya R.K.

Harish. S

COLLEGE OF ENGINEERING GUINDY

ANNA UNIVERSITY : CHENNAI 600 025

OCTOBER 2018

ANNA UNIVERSITY : CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report “**CRIME ANALYSIS SYSTEM**” is the bonafide work of “**Aswatha Narayanan.S, Jai Surya R.K., Harish.S**” who carried out the project work under my supervision.

SIGNATURE

MS.V.SUGANYA

TEACHING FELLOW

COMPUTER SCIENCE & ENGINEERING

12, SARDAR PATEL RD,

ANNA UNIVERSITY,

GUINDY,

CHENNAI-600 025

Contents

1. ABSTRACT.....	4
2. INTRODUCTION	5
3. RELATED WORK.....	6
4. SYSTEM DESIGN	7
4.1. ARCHITECTURE DIAGRAM.....	7
4.2. MODULES.....	8
4.2.1.DATA TRANSFORMATION	8
4.2.2.DATA PREPROCESSING.....	8
4.2.3.CRIME DATA MINING TASKS	9
4.2.4.CRIME ANALYSIS DASHBOARD	10
4.3. ALGORITHM.....	10
5. TESTING	11
6. RESULT ANALYSIS.....	12
7. CONCLUSION AND FUTURE WORK	20
8. REFERENCES	21
9. APPENDIX.....	23

ABSTRACT

Crime Analysis System is an application which helps in the pattern detection for various criminal activities. Crime is an interesting domain where data mining plays an important role in terms of prediction and analysis. There are many pattern detection systems for crime analytics has been proposed and some have been implemented too. The existing systems all give only results with less optimization, most of the models are theoretical as well give patterns on wide scope. The k-means method is a widely used clustering technique that seeks to minimize the average squared distance between points in the same cluster. Although it offers no accuracy guarantees, its simplicity and speed are very appealing in practice. By augmenting k-means with a simple, randomized seeding technique, it obtain an algorithm that is $O(\log k)$ competitive with the optimal clustering. Experiment explains augmentation improves both the speed and the accuracy of k-means, often quite dramatically. It undertakes regional criminal activities. In this model, the crime analysis on grouping by patterns based on rate of criminal activities on the regional basis. Mostly crime analysis is often under survey study, it demonstrates the result on improvised k-means clustering and traditional k means clustering for the regional crime analysis with distortion.

INTRODUCTION

The Data mining techniques have higher influence in the fields such as, Law and Enforcement for crime problems, crime data analysis. Recent researches on these techniques link the gap between clustering theory and practice of using clustering methods on crime applications. Cluster accuracy can be improved to capture the local correlation structure by associating each cluster with the combination of the dimensions as independent weighting vector and subspace span on it. The traditional k means clustering for crime analysis used for pattern detection on the rate of criminal activities on the regional basis. It is a technique that seeks to minimize the average squared distance between points in the same cluster. Although it offers no accuracy guarantees, its simplicity and speed are very appealing in practice. The Department of police is the major organization of preventing crime. It is very difficult to find a country without crime free society. The present society has also packed with various kinds of crimes. The Police department is responsible for enhancing security, the public became feel safety, maintaining order and keeping the silence. It implemented in the form of application and user interface is to access the model. The model is based on the highly optimized algorithm which is centre based clustering algorithm this clustering algorithm is used for grouping of similar patterns used in data mining techniques.

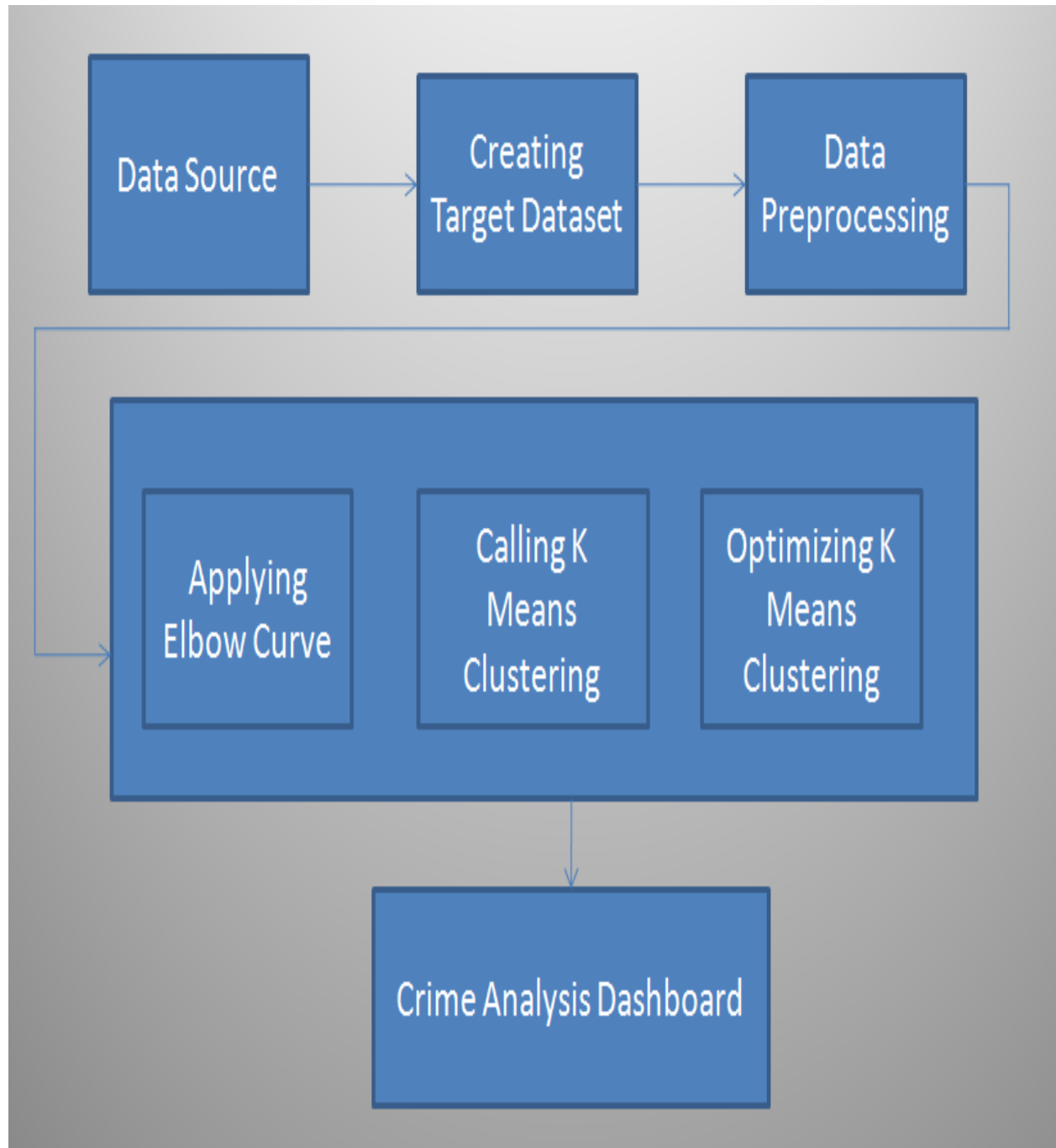
RELATED WORK:

(JYOTI AGARWAL, 2013)



SYSTEM DESIGN:

ARCHITECTURE DIAGRAM:



MODULES:

DATA TRANSFORMATION:

The *Input* for this module is the Raw Crime Dataset, which must contain comma separated values. We have used the raw UK Crime Dataset that is not preprocessed. This dataset has 4736 rows and 12 columns of information about type of crime in each state in specific UK county. The *Process* involved in this module is to transform the given dataset into a form that is suitable for further analysis tasks. The *Output* of this module is the newly transformed dataset that is appropriate to carry out the next steps.

DATA PREPROCESSING:

The *Input* for this module is the newly transformed raw dataset. This transformed data now may or may not contain outliers or which are also known as Anomalies. In order to remove the outliers, and to normalize the values obtained in the transformed dataset and find out the structure of the crime dataset, we perform Data Preprocessing. The *process* involved in the Data Preprocessing module is the process of detecting and removing outliers and normalizing the crime dataset. The *output* of this module is the preprocessed Crime Dataset that can be directly fed to the main module of the system to obtain the desired outcome.

CRIME DATA MINING TASKS:

The Crime Data Mining tasks module forms the core component of this entire Crime Analysis Application system, where all the essential tasks are carried out carefully. The input to this module is the Preprocessed Dataset obtained from the previous module. This module is further subdivided into three sub modules, namely Applying Elbow Curve, Calling K Means Clustering and finally, enhancing or optimizing K Means Clustering. The Elbow Curve plots a graph with number of clusters on one scale and the corresponding within sum of squares on other scale. This gives us the information on choosing the optimal K value, to perform K means Clustering. In the elbow curve, after a certain value of K, the within sum of squares measure would remain constant. This value can be chosen as the K value. Using this K value, we can perform the K means Clustering technique on the Cities and group Cities based on the Crime frequencies such that cities within a particular cluster are more similar to each other, than to the cities present in the other cluster. But, the efficiency of K Means Clustering can be improved by K means++ clustering algorithm, that works in a way that choosing the initial random point in dataset, is based on the probabilistic measure such that efficiency of cluster segregation is properly achieved. Comparing the efficiency ratios of the K means and K means++ clustering algorithm, we see that Kmeans++ outperforms K means clustering algorithm for slightly higher number of cluster values. Note that the total cluster size is equal to the total number of observations present in the Crime dataset. Hence, the output of this entire module is the Cluster information about each cluster and the cluster means of each crime, corresponding to each cluster and cluster assignment of each observation.

CRIME ANALYSIS DASHBOARD:

The Crime Analysis Dashboard module is the final module, that integrates crime data mining tasks module to be deployed using local server to make it available as web application , for the end user. The input to this module is the preprocessed crime dataset to be uploaded by the end user. The preprocessed dataset is applied to crime data mining tasks module, where we can also find analysis for two crime types, which gives as output the Cluster plot for each algorithm, that gives information about the cluster group for the given type of crime. The centers matrix describes the mean center for each type of crime in each Cluster. Using this information, we can see how uniform the crimes are distributed across each cluster. We can also search for specific city if it falls on given crime type and we obtain a pie chart, denoting the category level of Criticality of crimes and Tables for easy understanding. Finally, Help Guide is provided to end user on the usage of Crime Analysis Application.

ALGORITHM:

Let $X = \{ x_1, x_2, \dots x_n \}$ be set of data points and $V = \{ v_1, v_2, \dots v_k \}$ be set of centers.

- i. Randomly select 'k' cluster centers.
- ii. Calculate distance between each data point and cluster centers.
- iii. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the other cluster centers.
- iv. Recalculate new cluster using , $v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i$, where c_i represents number of data points in the i^{th} cluster.
- v. Recalculate the distance between each data point and new obtained cluster centers.
- vi. If no data point was reassigned then stop, otherwise repeat from step iii.

TESTING:

DATA TRANSFORMATION:

The Test Procedure for this module is first open dataset in MS Excel, apply “Filter and Sort” utilities appropriately. Finally, record counts of each type of crime in each city. The assumption for testing this module is that the Crime Dataset should be as comma separated values to be compatible to work with Excel. This module does not work with non-csv format dataset files.

DATA PREPROCESSING:

The Data Preprocessing module is to be tested with transformed dataset obtained from the previous module. This module correctly works when the dataset contains outliers, which it detects and removes. This module is not tested when the structure of the data is not a matrix / data frame. It also does not suit when there is no States column and Crime count columns for each type of Crime. It is tested with assumption that dataset is of only rows and columns, with each representing States and Crimes.

CRIME DATA MINING TASKS:

This module is tested with the preprocessed data from the data preprocessing module. It is tested with assumptions that the crime dataset must be strictly free from outliers, else the cluster segregation would be incorrect. Another assumption is that the dataset must strictly contain “numerical values” to be able to apply K means clustering, and this module does not work for values of K less than 0 or equal to 0.

CRIME ANALYSIS DASHBOARD:

This module is tested with Crime Dataset that is strictly free from outliers, free from categorical values, contains numerical values. When the module does not meet atleast one of the above requirements, it fails to produce correct results. It also assumes that the data must be structured so that it can be easily searchable. The test procedure is simply straightforward. Click Run the App, Choose Upload data, Click Analyze data, Click Visualize data.

RESULT ANALYSIS:

DATA TRANSFORMATION:

This module takes as input a raw Crime Dataset that is not in a suitable format for Data Mining to perform, in the first place. This process is included only when there exists Crime Data consisting of multiple same cities and the frequency of crime in each city differs. We then apply Filter and Sort utilities to group cities and their crime frequencies for each type of crime. The result is the newly transformed target dataset.

	A	B	C	D	E	F	G	H	I	J	K
1	STATES	ANTI_SOC	BICYCLE_T	BURGLARY	CRIMINAL	DRUGS	OTHER_CF	OTHER_TH	POSSESSIC	PUBLIC_O	ROBBERY
2	AMBER VA	300	2	68	65	19	5	44	3	8	2
3	ASHFIELD	4	0	1	0	1	0	0	0	0	0
4	BARNESLEY	0	0	0	0	0	1	0	0	0	0
5	BASSETLAW	0	0	0	0	0	1	0	0	0	0
6	BLACKPOOL	0	0	0	0	0	0	0	0	0	0
7	BOLSOVER	175	2	32	51	13	4	43	4	8	0
8	BRADFORD	0	0	0	0	0	0	0	0	0	0
9	BROXTOWN	5	0	0	0	1	0	0	0	0	0
10	CHESTERF	324	0	45	83	14	6	62	6	16	2

Figure 1. Transformed Target Dataset.

DATA PREPROCESSING:

The Data Preprocessing module takes as input the newly obtained transformed Target dataset, and checks if there were any outliers present in the dataset. If so, we need to remove them for correct analysis. We can also see the structure of the dataset and the value type of each observation in the crime dataset. The result is the plot of outlier present in the Crime Dataset.



Figure 2: Data Preprocessing of removing Outliers.

CRIME DATA MINING TASKS:

This core module of the Crime Analysis application consists of three sub modules namely, Applying Elbow Curve. The result of Elbow curve is the plot of Cluster Value Vs Within Sum of Square. This plot helps us in the decision of choosing an optimal value of K . The second sub module is applying K Means Clustering The result is the Cluster assignment of each observation in the dataset and the ratio of “between sum of square” and “total sum of square”, and the cluster means of each crime in each cluster, and also the total cluster size is equal to the total number of observations. The next sub module, enhancing K Means Clustering whose result also gives the same result as the previous sub module, but with “increased efficiency”.

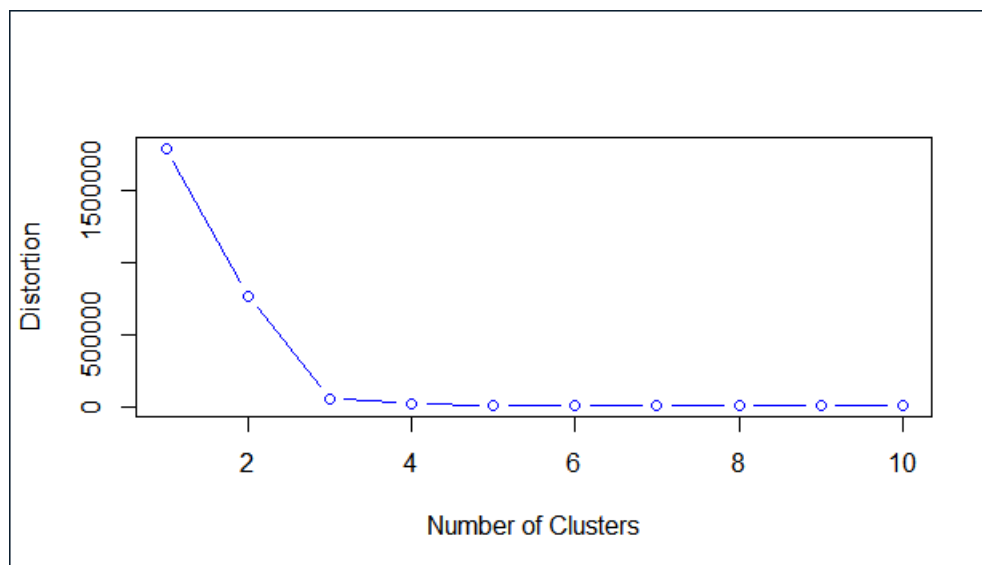
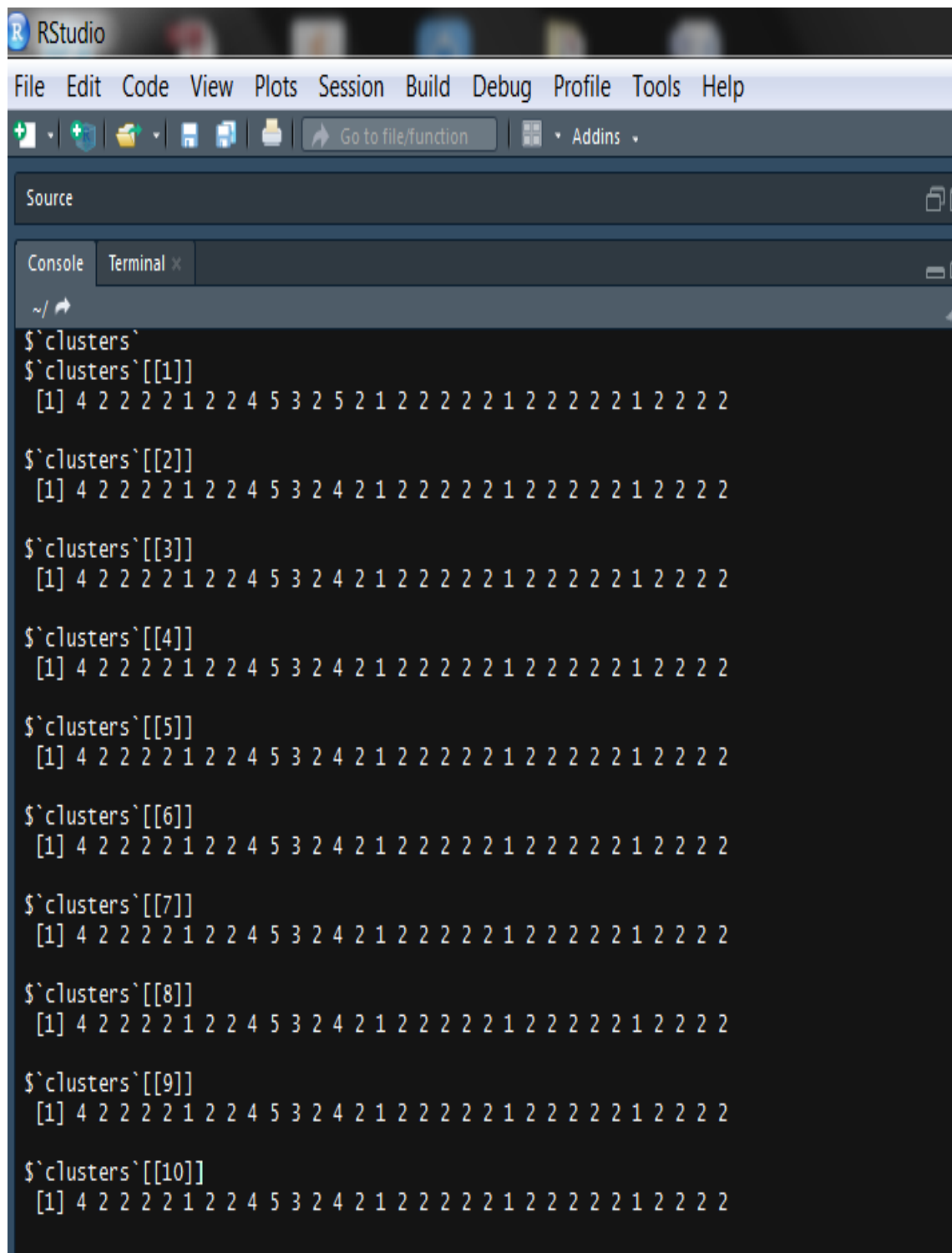


Figure 3: Applying Elbow Curve



The image shows the RStudio interface with the console pane active. The console displays the results of a K-means clustering operation over 10 iterations. The output for each iteration is a 1x10 matrix where the first element is the cluster assignment for each of the 10 data points. The clusters are labeled 1 through 4. The sequence of cluster assignments for each iteration is as follows:

```
$`clusters`  
$`clusters`[[1]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 5 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[2]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[3]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[4]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[5]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[6]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[7]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[8]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[9]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2  
  
$`clusters`[[10]]  
[1] 4 2 2 2 2 1 2 2 4 5 3 2 4 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2
```

Figure 4 a. Calling K Means Clustering

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal
~/
$centers[[8]]
  ANTI_SOCIAL_BEHAVIOUR BICYCLE_THEFT BURGLARY CRIMINAL_DAMAGE_AND_ARSON DRUGS
1          183.25         1.750000 39.500000         41.750000 8.250000
2           1.00         0.000000  0.181818         0.181818 0.181818
3          58.00         1.000000 30.000000         28.000000 1.000000
4         303.00         3.333333 65.333333         86.333333 16.000000
5        1052.00        56.000000 182.000000        247.000000 77.000000
  OTHER_CRIME OTHER_THEFT POSSESSION_OF_WEAPONS PUBLIC_ORDER ROBBERY SHOPLIFTING
1  8.250000 31.750000         2.250000         6.25 1.750000         24.5
2  0.363636 0.136363         0.000000         0.00 0.000000         0.0
3  8.000000 16.000000         0.000000         2.00 0.000000        13.0
4  6.666667 55.666667         4.666667        12.00 2.333333        65.0
5 27.000000 227.000000        15.000000        50.00 38.000000       214.0
  THEFT_FROM_PERSON VEHICLE_CRIME VIOLENCE_AND_SEXUAL_OFFENCES
1         1.750000 50.000000         71.500000
2         0.000000  0.136363         0.545454
3         1.000000 81.000000        24.000000
4         6.333333 58.000000       112.000000
5        36.000000 245.000000       430.000000

$centers[[9]]
  ANTI_SOCIAL_BEHAVIOUR BICYCLE_THEFT BURGLARY CRIMINAL_DAMAGE_AND_ARSON DRUGS
1          183.25         1.750000 39.500000         41.750000 8.250000
2           1.00         0.000000  0.181818         0.181818 0.181818
3          58.00         1.000000 30.000000         28.000000 1.000000
4         303.00         3.333333 65.333333         86.333333 16.000000
5        1052.00        56.000000 182.000000        247.000000 77.000000
  OTHER_CRIME OTHER_THEFT POSSESSION_OF_WEAPONS PUBLIC_ORDER ROBBERY SHOPLIFTING
1  8.250000 31.750000         2.250000         6.25 1.750000         24.5
2  0.363636 0.136363         0.000000         0.00 0.000000         0.0
3  8.000000 16.000000         0.000000         2.00 0.000000        13.0
4  6.666667 55.666667         4.666667        12.00 2.333333        65.0
5 27.000000 227.000000        15.000000        50.00 38.000000       214.0
  THEFT_FROM_PERSON VEHICLE_CRIME VIOLENCE_AND_SEXUAL_OFFENCES
1         1.750000 50.000000         71.500000
2         0.000000  0.136363         0.545454
3         1.000000 81.000000        24.000000

```

Figure 4 b. Applying K Means Clustering


```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
~/
+ }
>
> kmeansp2(ktest, 4)
K-means clustering with 4 clusters of sizes 23, 1, 4, 3

Cluster means:
  ANTI_SOCIAL_BEHAVIOUR BICYCLE_THEFT  BURGLARY CRIMINAL_DAMAGE_AND_ARSON    DRUGS
1          3.478261      0.04347826   1.478261          1.391304  0.2173913
2        1052.000000    56.00000000  182.000000          247.000000  77.0000000
3         183.250000     1.75000000   39.500000          41.750000   8.2500000
4         303.000000     3.33333333   65.333333          86.333333  16.0000000
  OTHER_CRIME OTHER_THEFT POSSESSION_OF_WEAPONS PUBLIC_ORDER  ROBBERY SHOPLIFTING
1    0.6956522    0.826087          0.000000    0.08695652  0.000000    0.5652174
2   27.0000000   227.000000          15.000000   50.00000000  38.000000   214.0000000
3    8.2500000   31.750000          2.250000    6.25000000   1.750000   24.5000000
4    6.6666667   55.666667          4.666667   12.00000000   2.333333   65.0000000
  THEFT_FROM_PERSON VEHICLE_CRIME VIOLENCE_AND_SEXUAL_OFFENCES
1      0.04347826      3.652174          1.565217
2     36.00000000     245.000000          430.000000
3     1.75000000     50.000000          71.500000
4     6.33333333     58.000000          112.000000

Clustering vector:
[1] 4 1 1 1 1 3 1 1 4 2 1 1 4 1 3 1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1

within cluster sum of squares by cluster:
[1] 12025.304      0.000  4730.500  4215.333
(between_SS / total_SS =  98.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"       "inicial.centers"
"
> |

```

Figure 5. Enhancing K Means Clustering

CRIME ANALYSIS DASHBOARD:

The Crime Analysis dashboard completes the Crime Analysis Application. The result of this module is the Application that end user can interact with to perform analysis. The user can upload the dataset, and analyze the data using two types of crime, which gives the Cluster plot to analyze how the data is clustered into groups. Then normal plot is also obtained for both algorithms. The center matrix gives us the information about the Cluster mean of each Crime in each Cluster. The normal plot gives the simple K Means Plot of entire Crime Dataset. The user can also search for specific city if it falls under the Crime, in which he is investigating. The pie chart gives the information about the “Category Level” of Crimes, meaning that the Category Level 1 is the most critical level of crime and the percentage value indicates that that much percentage of cities in our crime dataset are at high critical level. The user is also provided with option of downloading the table consisting of cities to local system. And finally, a Help Guide is provided with the usage of the Crime Analysis Application.

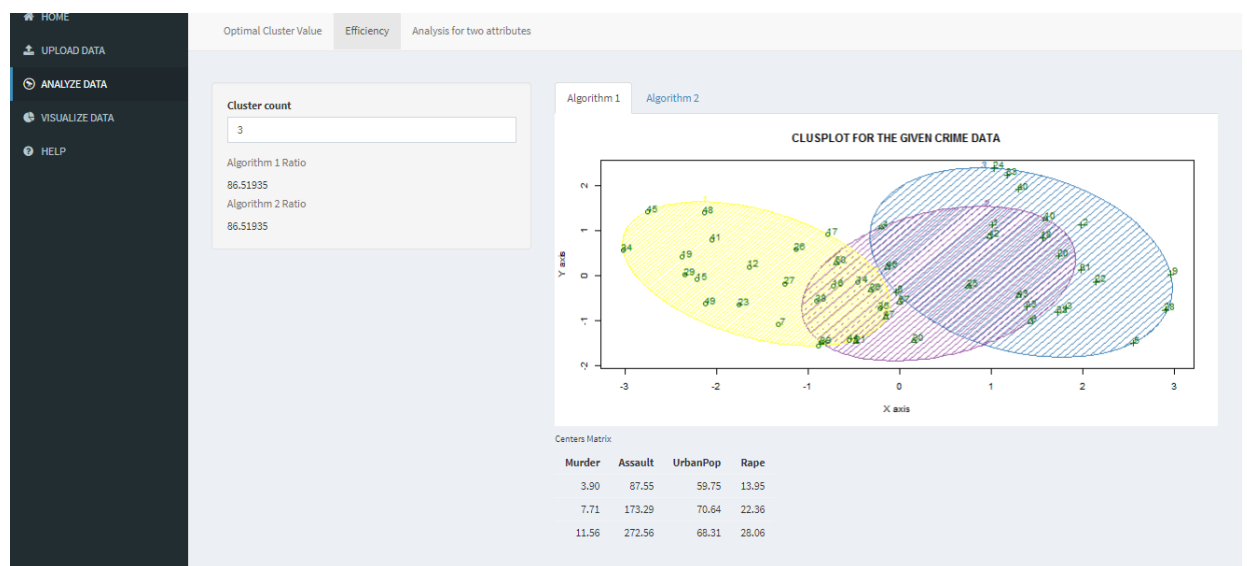


Figure 6 a. Crime Analysis Dashboard.



Figure 6 b. Crime Analysis Dashboard of Normal Plot

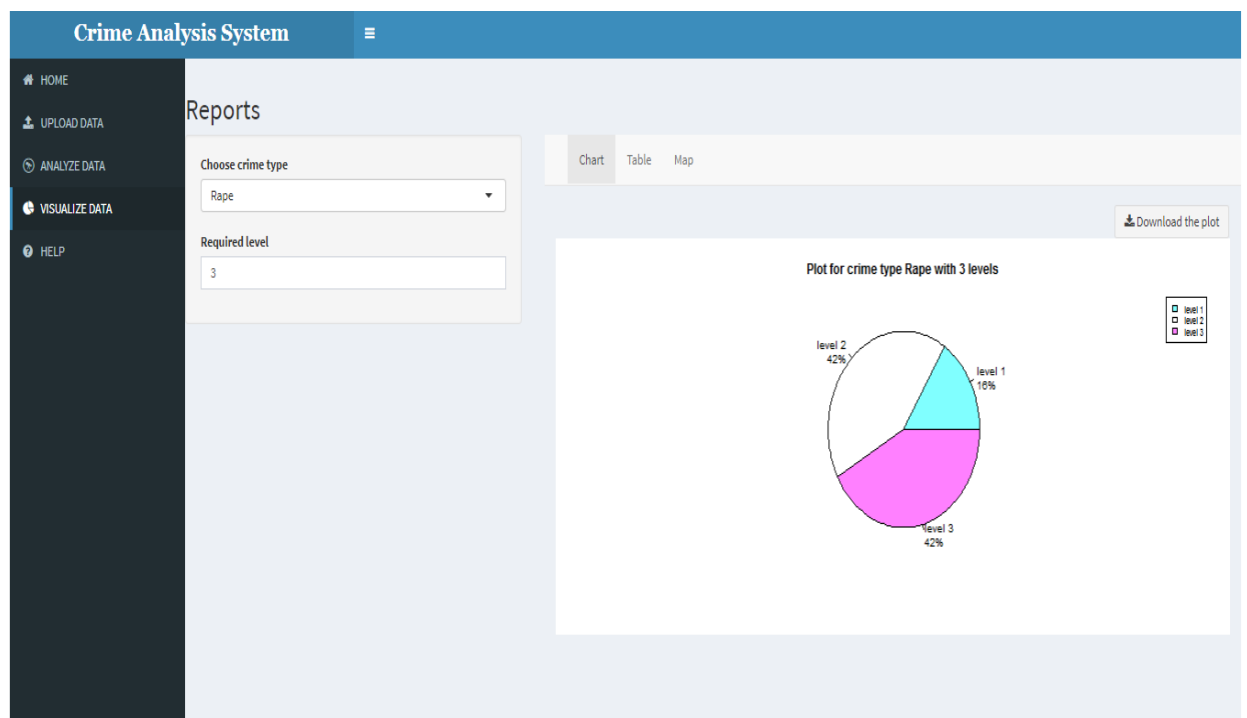


Figure 6 c. Crime Analysis Dashboard of Pie Chart of “Category Level”

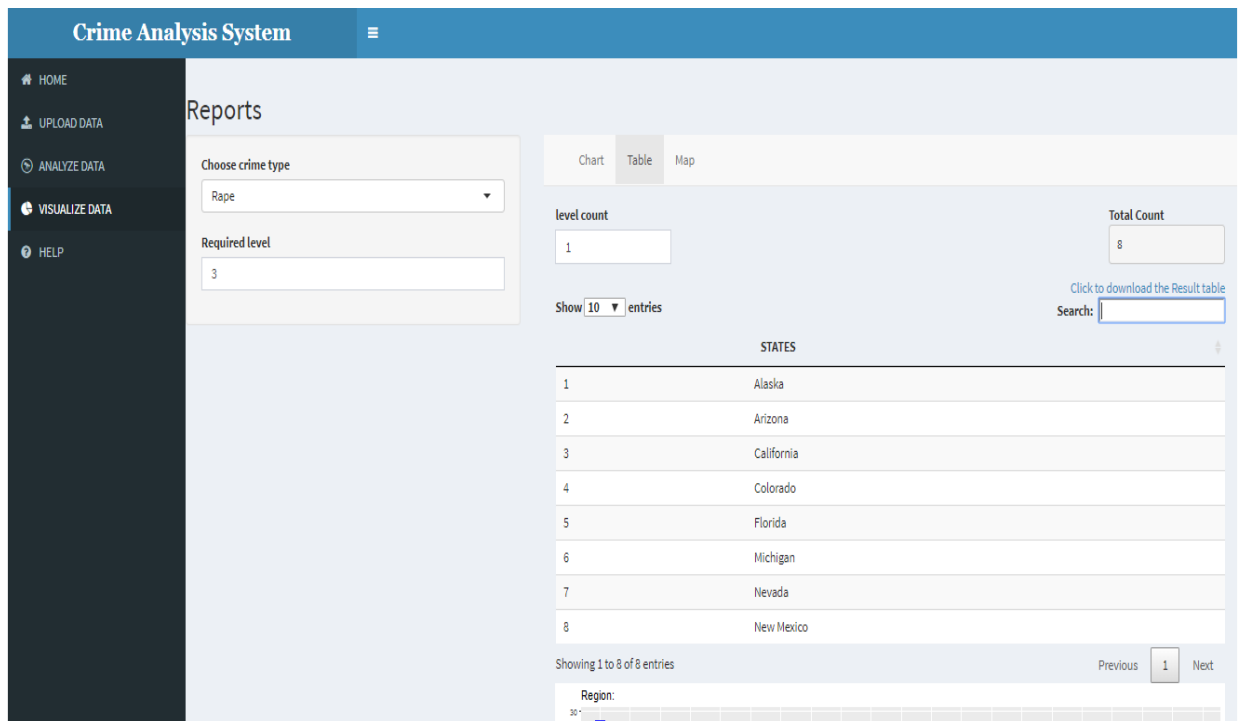


Figure 6 d. Crime Analysis Dashboard of Table of States.

CONCLUSION AND FUTURE WORK:

The user is provided with an Application portal for the Crime Analysis. It includes interactive features, user-friendly interface, that user can play around. This portal provides a dynamic clustering of the crime data points on the interface, and also visualizes the result in two forms namely, Pie Chart and Tables for easy understanding. This project can be further extended by using some advanced clustering algorithms to increase crime analysis accuracy and to enhance overall performance.

REFERENCES:

- [1] J. Agarwal ,R . Nagpal and R. Sehgal, “Crime analysis using k-means Clustering”, *International Journal of Computer Applications*, vol. 83, no. 4, pp.1-4,2013.
- [2] H.Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, “Crime data mining: a general framework and some examples”, *IEEE Computer Society* , vol. 37, no. 4, pp. 50-56, 2004.
- [3] S.B.R. Rajan, K.G.Srinivasagan, K Ramar and S.M. Sundaresan, “An automated crime pattern detection using k-means clustering”,*IEEE*, vol. 3, no.4,pp. 220, 2011.
- [4] Dr. I. Badri and S.P. Sajjan, “Location based crime detection using data mining”, *Bonfring International Journal of Software Engineering and Soft Computing*, vol. 6, no. 5, pp. 208-212,2016.
- [5] A.T.Murray, I.McGuffog, J.S.Western,P.Mullins, “Exploratory spatial data analysis techniques for examining urban crime: implications for evaluating treatment”, *The British Journal of Criminology*,vol. 41, no. 2, pp. 309-329,2001.
- [6] G.Oatley,B.Ewart, “Data mining and crime analysis”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1,no. 2,pp. 147-153,2011.
- [7] H.Hassani,X.Huang,E.S.Silva and M.Ghodsi, “A review of data mining applications in crime”, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol.9, no.3,pp. 139-154,2016.
- [8] M.Mowafy,A.Rezk,H.M.El-bakry, “General crime mining framework for unstructured crime data prediction”, *International Journal of Computer Application*,vol. 4, no. 8, pp.11-15,2018.

- [9] T.Almanie, R.Mirza, E.Lor, “Crime prediction based on crime types and using spatial and temporal criminal hotspots”,vol. 5,no. 4,pp. 1-19,2015.
- [10] K.K.Deepika,S.Vinod, “Crime analysis in India using data mining techniques”, *International Journal of Engineering and Technology*, vol. 7, no. 2.6,pp. 253, 2018.
- [11] S.Sivaranjani,S.Sivakumari,S.Maragatham, “GIS based serial crime analysis using data mining techniques”,vol.153,no. 8,pp. 19-23,2016.
- [12]W.Grohman, “Using convex sets for exploratory data analysis and visualization”, *Data Mining and Knowledge Discovery*,vol.9,no.3, pp. 275-295,2004.
- [13] I.Lee, “Exploration of massive crime datasets through data mining techniques”, *Applied Artificial Intelligence*, vol. 25,no. 5,pp. 362-379,2011.
- [14] K. Revathy, “Survey of data mining techniques on crime data analysis”, *International Journal of Data Mining Techniques and Applications*, vol.1,no.2, pp. 47-49, 2012.
- [15] M.R.Keyvanpour, “Detecting and investigating crime by means of data mining: a general crime matching framework”, *Procedia Computer Science*, vol.3,pp. 872-880,2011.

APPENDIX

Crime analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder. Information on patterns can help law enforcement agencies deploy resources in a more effective manner, and assist detectives in identifying and apprehending suspects. Crime analysis also plays a role in devising solutions to crime problems, and formulating crime prevention strategies. Quantitative social science data analysis methods are part of the crime analysis process; though qualitative methods such as examining police report narratives also play a role. The k-means algorithm has at least two major theoretic shortcomings: First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size. Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.