MINING AND MANAGING BIG DATA FOR PREDICTING STATISTICAL SOLUTIONS OF REAL-TIME ISSUES USING R LANGUAGE

Pravina Bhatt
Master of Applied Computing
University of Windsor
Windsor, ON
bhatt9@uwindsor.ca

Laveen Vasnani
Master of Applied Computing
University of Windsor
Windsor, ON
vasnani@uwindsor.ca

Yixian Hao
Master of Applied Computing
University of Windsor
Windsor, ON
hao119@uwindsor.ca

ABSTRACT

This paper is an initiative for providing application based solution for the real-time issues faced in our society every now and then. We have used the statistical methodology for deriving proof based result with the help of large amount of data collected for research. These facts are the representation of Big Data set, accumulated on the basis of day to day observations for respective authorities which includes the list starting from civilians to experts in certain sector. In this paper, we utilized the advanced functionality of R language for depicting graphical visualization of clustered data set. Furthermore, this paper covers the proof grounded depiction of mining, clustering and manipulating huge data set of criminal records generated by Chicago police department for the retrieval of period oriented investigation. This research is the pillar of prediction of upcoming possibilities of crimes which are stated in vast categories. Moreover link predication of this gathered study can be further used by researchers for portraying generation based evolution of crimes with respect to geography.

Keywords

Big Data, data set, data frame, data mining, RStudio, R language, statistical computing

1. INTRODUCTION

Among the currently evolving latest terminology, Big Data ensures an extensive area of impact on our society and our business life cycle [8]. Big Data has perceptively transformed our society and will continue to draw diverse attentions from technical experts as well as community in broad-spectrum [12]. Information and their utilization varies as per requirement based on the purposes. This well-known term Big Data, not only deals with the structure and storage related aspects but also renders around the operations implemented to visualize specific estimations. As the erection of facts escalates, their clustering and enlisting categories also gets broaden. All this elaboration is done with the motto of retrieving crucial prediction related to domain specific researches. Various implemented algorithms challenges fruitful results for human development [13], but the consequences of user specific query recovery is only possible using specific procedural algorithms [7]. To widen the range of such possibilities tools can be used as boon for working with this vast data set in very user friendly and easily understandable manner[2].

Figure 1: Big Data Platform[1]

Under this operation we have used R language which is widely recognized programming language that supports effective data analysis for statistical computing and graphical representation. In this paper we have demonstrated hands-on with dealing real-time data set collected on large scale recording and fetching precise statistic among clustered pattern, which can be later used for prediction of future forecasts.

The later part of this paper covers declaration of proposed problem statement in immediate section. Related works are briefly reviewed in section 3 which will be followed by enlisting requirements for setting up our project in section 4. Section 5 will let our users to get familiarize with the

[1]https://thecustomizewindows.com/2014/01/big-data-and-r-programming-language/

analysis tool used for achieving desired goal. Whereas section 6 and 7 will cover experiment on instantaneous dataset and its corresponding data retrieval. Finally we will inscribe future scope of our project and conclude our thought in the last portion of this paper under sections 8.

2. PROBLEM STATEMENT

Due to the extensive range of diverse data sources and the huge volume, it is too difficult to collect, integrate and analyze Big Data with scalability from scattered locations [12]. Managing huge set of records is perhaps considered to be most tedious in address with complex information gathering and their access [18] [13]. Thus, here is our initiation of overcoming those drawbacks in very efficient and attractive manner. We selected illustration of massive facts that were collected in a raw form, which managed and manipulated using statistical programming language. We consumed the strength of R language for storing series complex data-analysis with the help of its highly extensible and powerful scripting functionality. This detail oriented scripting language will be encoded for displaying graphical notation of grouped output depending upon time span and locality. Processing the crime record collected by Chicago police department, which is an instance of Big Data we intended to gaze the regular pattern of crimes evolving based on several constraints. This pattern recognition may further utilized for futuristic prediction of similar crimes and implement step ahead precautions for the same.

3. RELATED RESEARCHES

Before moving towards practicing with concrete figures we need to understand the methodologies used by peer investigators who dealt with it. This section is all about listing correlated works under the roof of Big Data and their own strategy of regulate them.

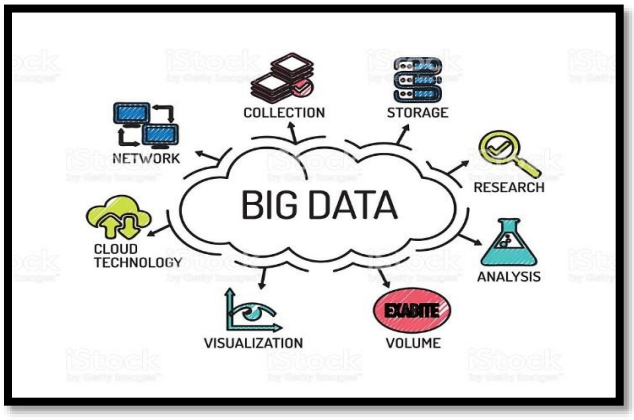

Figure 2: Relative research works in Big Data[2]

3.1 Analyses of Phylogenetic and Evolution with R [21]

Phylogenetic covers a wide are of related issues whose analyzing includes tree estimation, dating divergence times and estimating speciation rates. The implementation of these methods in R enhances their integration under single user interface. Under this work author claims about advantage of R's interpreted language functionality which helps to do analyzing with a single line. R stands apart from all other programming languages due to its vast range of rewards which is reasonable for efficient statistics analysis.

3.2 Predicting students' performance [15]

A system which can predict the performance of students from their previous performances using concepts of data mining techniques under Classification. In this privilege author applied classification algorithms on students personal information and predicted the general and individual performance of freshly admitted students in future examinations.

3.3 Business intelligence and analytics [5][11]

BI&A includes business-centric practices and methodologies that can be applied to various high-impact applications such as e-commerce, market intelligence, e-government, healthcare, and security. It suggests a sea change in the fundamental ways that managers will scan their business environments, recognize opportunities, problems and crises, diagnose problems, search for solutions, and establish the criteria to select among alternatives. Scholars claim that the emerging Business Intelligence /Business Analytics concepts particularly Big Data concepts will impact some or all of these phases. According to them the concepts evolved since decades about this phase holds to be true even after Big Data concepts emerged. Further it is depicted that it is necessary to adapt the latest terms and technologies for polishing its vast capabilities of application, as Business Intelligence and analytics is all about growth of market.

3.4 In Government sector [6]

Many white papers, journal articles, and business reports have proposed ways governments can use big data to help them serve their citizens and overcome national challenges (such as rising health care costs, job creation, natural disasters, and terrorism).

Success depends on their ability to integrate and analyze information, develop supporting systems (such as big-data control towers), and support decision making through analytics. Here writer articulated about challenges faced by governments of top countries and opportunities they found in big data to overcome it.

[2]https://www.istockphoto.com/ca/vector/big-data-chart-with-keywords-and-icons-sketch-gm598053934-102462261

3.5 Mobile system targeted to high transmission rate [12]

This is one of revolutionary review in the sector of telecommunication and similar field which demands high performance. This review is inspired by support vector machines for analyzing data using classification and regression associated with algorithms. In the later part author concludes that Big Data analytics is still in the initial stage of development, since existing Big Data techniques and tools are very limited to solve the real Big Data problems completely, in which some of them even cannot be viewed as Big Data tools in the true sense. Therefore, more scientific investments from both governments and enterprises should be poured into this scientific paradigm to capture huge values from Big Data.

3.6 Dealing with Big Data for Health Care [22]

Big data, including analytics, is a powerful tool that will be as useful in health care as it has been in other industries. Under this investigation, scholars had examined into the depth of functionality portion of Big Data. This is the analytical invention of identifying and managing high-risk and high-cost patients. Moreover, according to this ethics it is claimed that such methodology will be among those that deliver the greatest value for health care organizations in the near term. This general approach has great potential for improving value in health care and organizations that employ it will benefit, especially under payment reform.

4. SYSTEM REQUIREMENTS

Until now we were familiar with analyzing and matrix computation feature of R language but on the other side it consists of vivid vibrant functionalities for restoring data and their recovery. R belongs to family of file restoring utilities. It mends files both on local disks and on disks on remote computers over network, even if their partition structures are impaired. To utilize the hidden strength of this coding language we need to get into more hardware part, which will be the basic step for configuration.

RStudio itself doesn't require a ration of computational power, so our requirements are going to be dependent on how we are using R. The number of cores, speed of the cores and the amount of RAM that we need is highly dependent on the work/analysis we will be doing. R itself is single threaded, and as such, we won't benefit from additional cores unless we are familiar with the various libraries that parallelize work and are then able to leverage multiple cores. It is unlikely for beginners to get into this much depth, who use R for only academic purpose. But in this application based process we need to focus on huge data set which will eat maximum potential of hardware, so it is must to set up accordingly for exploration purpose. However, if user intend to be analyzing larger data sets (>1GB) then it would be wise to invest in more RAM. Moreover below listed are the basic requirements which is must setting up the instance of data set mentioned in later section of this paper.

a) A processor-compatible platform running for various types of operating system either Windows, Mac OS X or Linux.
b) At least 1 GB of RAM and enough disk space for recovered files, image files, etc.
c) The administrative privileges are required to install and run R-Studio utilities for Windows, Mac OS X or Linux.
d) A network connection for data recovering over network.

5. ANALYSIS TOOLS

Here comes strategies and aspects that is to be considered for dealing with analysis tools and using them. Our main emphasis for implementing R language led us to experience interactive graphical user interface tool that provides working environment for R named R studio. It is a free integrated development environment for R language that supports functional retrieval of immediate information. Moreover, it wires analyzing constraint specific grouping, multiple work space, interactive output visualization and code suggestions.

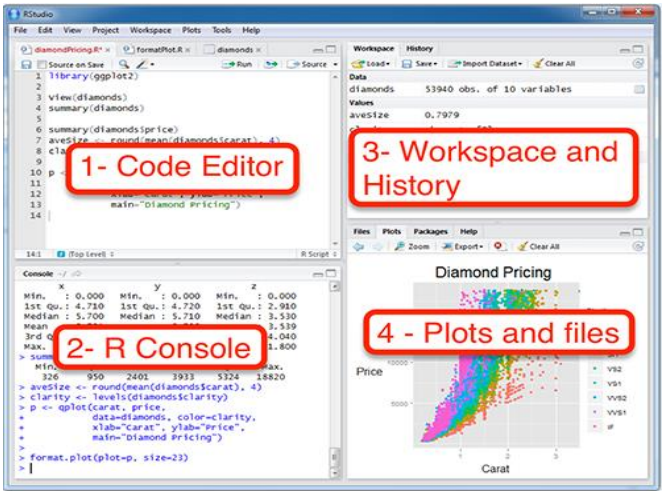

Figure 3: View of RStudio Screen[3]

Talking about dealing with datasets using R language, all the objects will be imported to the memory. Such a design makes it possible to interact with R in a faster way for most of the users. But in terms of Big Data it will decrease the running speed and sometimes even causes memory-related errors.

[3]http://www.rstudio.com/Unformat_Help/index2.html?systemrequirements.html

Mainly, there are three aspects need to be taken care of: 1.Improving the efficiency and executing speed; 2. Storing data externally for potential memory limitations; 3.Using specialized methods to deal with the large amount of data. Details will be discussed below.

5.1 Efficiency

a) Try to use vectored computations. Using built-in functions to deal with vectors, matrices and lists (such as sapply, lapply and mapply) and avoiding loops (while and for).
b) Use data frames only when necessary because matrices cost less.
c) Set colClasses and nrows explicitly when importing external data set to data frame with read. table() which would reduce the costs of memory. Use scan () when importing external dataset to matrices.
d) Before dealing the whole dataset, use the subset to do some testing in order to optimize as well as find out bugs.
e) Delete temporary and unused objects regularly. Use rm(list=ls()) to clean the memory and rm(object) to remove a specific object.
f) Profile your programs and check the time spent on every function. Functions like Rprof(), summaryRprof(), system.time() and various similar one would help.

5.2 External Storage

Instead of importing all the data into memory, one better strategy of Big Data is external storage. Users can store the data in disks or databases and they are only visited when necessary which would reduce the limitations of memory. There are several packages can be used to store data out of memory.

a) ff --------- Provides a data structure stored on disks which allows user to manipulate it as it is in memory.
b) bigmemory -------- Supports CRUD operations of matrices of large scales. The matrices are stored in shared memory or memory-mapped files.
c) filehash ------- Implements a simple Key-Value database of which the key is related to the data stored on disks.
d) ncdf, ncdf4 ------ Provides an interface to Unidata netCRF data files.
e) RODBC, RMySQL, ROracle, RPostgreSQL, and RSQLite --------- These packages help to operate data from external RDMS.

5.3 Special Algorithms

a) Biglm and speedglm provide effective algorithms to fit linear models or generalized linear models against Big Data. Functions like lm() and glm() could help.
b) Users can create matrices of large scales by using the package bigmemory. Package biganalytics includes k-means clustering, column statistics and a packaging of biglm(). Package bigabulate provides functions like table(), split(), tapply(). Package bigalgebra supports functions of advanced linear algebra.
c) Package biglars includes least-angle regression, lasso, and stepwise regression against large datasets.
d) Package brobdingnag can be used to manipulate big numbers(greater than 2^1024)

6. TESTING USING DATASET

On the basis of previous discussion users might got familiarized with how to set up R language and to embed it into RStudio. Here comes the illustrative step for experiment those knowledge for managing day to day issues that subsists around our society. In this section we have covered main motive of this research using example of Big Data chunk prepared by Chicago police department. This detail is about all crimes happened in the year 2017 which is categorized on the basis of various constraints like case number, location of crime, type of crime and many more. This data set is a rich collection of important aspects that can affect the reputation of government if ignored. Our initiative is to refine this information process, scrutinize and extract the desired outcome using R programming language.

As R language is capable of read data from various sources like csv, excel and json, we selected csv file that better fits the requirement. Thus below given steps includes uploading the project on working environment of RStudio:

a) Data import

```
# import data from csv
data = data.frame(fread("Crimes2001.csv",sep=",",
                        header = T,
                        data.table=F,
                        verbose = T,
                        integer64 = 'numeric',
                        stringsAsFactors = T))
# remove all the lines with NA
data.clean = na.omit(data)
```

Figure 4: Code snippet for data import in R

b) Coding: Includes all default functions depending upon the structure of output
c) Graphical plotting of output: Use pie() for pie chart, plot() for line chart, barplot() for bar chart, dotchart() for dot chart and other similar kind of library functions for representation.

As we mentioned earlier that we experimented on gigantic file worth 1.44 GB of size, which is commanded to retrieve

attractive visualization of upshots within 26 seconds. This file consists of 6571632 tuples and 22 attribute which may take more than 1000 minutes for other tools [12]. This testing is applied on system with following configuration:

a) Processor: 7th Gen mobile i7 CPU
b) CPU cores: 2.7 GHz
c) Operating System: Windows 10
d) RAM: 16 GB

And the result is displayed below:

```
> data = fread("Crimes2001.csv",sep=',',header = TRUE)
Read 6571636 rows and 22 (of 22) columns from 1.442 GB file in 00:00:26
>
```

Figure 5: Efficiency encountered as a result

7. STATISTICAL RESULTS

The below given graphical representations are the outcome of interviewing potential civilians those may contribute in fetching condition specific output. Moreover when we came across the constraints to be applied for results there comes the processing advanced and powerful capability of R language. Below pictured view are the rudimentary structure using different test cases for numerous ways of representing graph.

7.1 Distribution of primary types of crimes:

For nominated record, under test case 1 we classified it on the basis of repetitive practice of similar crimes. This study will help to come up with frequent kind of crimes in Chicago for year 2017.

```
########################################
# Number of Crimes for Different Type of Crimes
########################################
summary_primaryType = summary(data.clean$Primary.Type,
                              maxsum = 10)
# calculate the percentage of each category
piepercent = paste(round(100*summary_primaryType/sum(summary_primaryType),2),
                   "%")
# draw pie chart
pie(summary_primaryType,
    piepercent,
    col=rainbow((length(summary_primaryType))),
    main = "Distribution of Primary Type of Crimes")
# draw legend
legend("topright",
       names(summary_primaryType),
       cex=0.8,
       fill=rainbow(length(summary_primaryType)))
```

Figure 6: Code for distribution of primary types of crimes

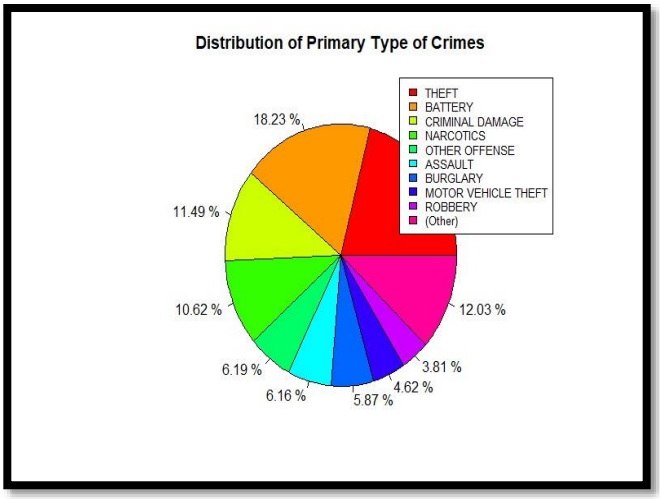

Figure 7: Pie Chart for distribution of primary types of crimes

```
########################################
# Arrest Rate for Different Type of Crimes
########################################
arrest_type_true = rep(0,10)
arrest_type_true_percentage = rep(0,10)
for(i in 1:10){
  arrest_type_true[i] = dim(data.clean[data.clean$Arrest == "true" &
          data.clean$Primary.Type == names(summary_primaryType[i]),])[1]

  arrest_type_true_percentage[i] =
    paste(round(arrest_type_true[i]*100 / summary_primaryType[i],2),"%")
}
# draw pie chart
pie(arrest_type_true,
    arrest_type_true_percentage,
    col=rainbow((length(arrest_type_true))),
    main = "Arrest rate for different type of crimes")
# draw legend
legend("topright",
       names(summary_primaryType),
       cex=0.8,
       fill=rainbow(length(arrest_type_true)))
```

Figure 8: Code for estimating arrest rate

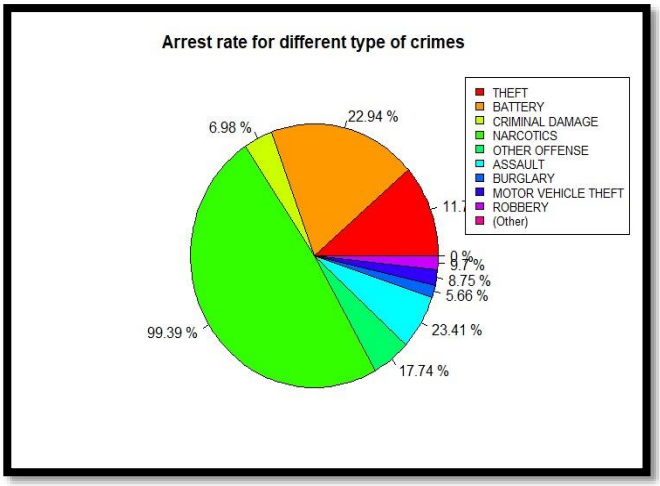

Figure 9: Pie chart for Arrest rate for various crimes

```
##########################################
# Number of Crimes for Different Type of Crimes for Area 25
##########################################
data.clean.25 = data.clean[data.clean$Community.Area == 25,]
summary_primaryType_25 = summary(data.clean.25$Primary.Type,maxsum = 10)
# calculate the percentage of each category
piepercent_25 = paste(
  round(100*summary_primaryType_25/sum(summary_primaryType_25),2)
  ,"%")
# draw pie chart
pie(summary_primaryType_25,
    piepercent_25,
    col=rainbow((length(summary_primaryType_25))),
    main = "Distribution of Primary Type of Crimes for Area 25")
# draw legend
legend("topright",
       names(summary_primaryType_25),
       cex=0.8,
       fill=rainbow(length(summary_primaryType_25)))
```

Figure 10: Code snippet for crimes for selected area 25

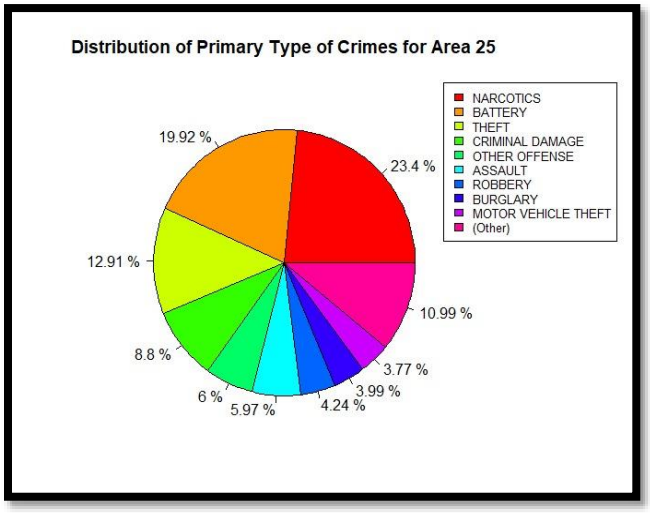

Figure 11: Pie chart for distribution of primary type of crimes for specific area

7.2 Growth rate of crime count per month:

Adding into types of representation of outcomes, next we encountered line chart for displaying the growth rate of crime count going by month. Thus, above given is the product of this trial. Crimes that remain unreported can often skew rates meaning it can generally be assumed that occurrences of crime are more prevalent than testified crime statistics advocate.

```
##############################
# number of crimes per month
##############################

c = cbind(data.clean,
          month = format(as.Date(data.clean$Date,"%m/%d/%Y"),"%m"))
plot(names(summary(c$month)),
     summary(c$month),
     type = "b",
     main = "Number of Crimes per month",
     xlab="month",
     ylab="c-count")
```

Figure 12: Code snippet for line chart

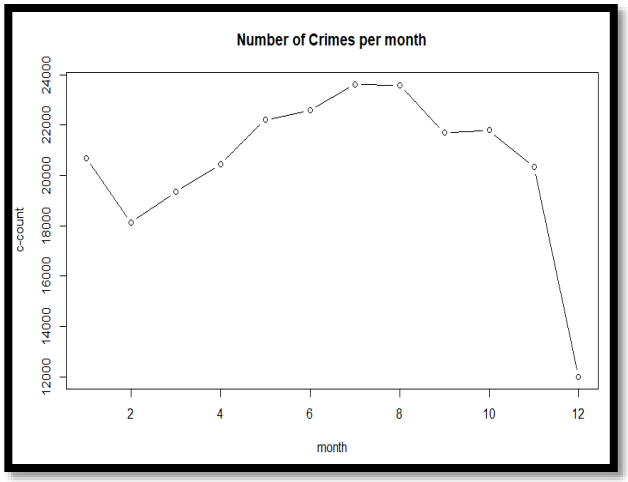

Figure 13: Line Chart for demonstrating growth rate of number of crimes per month

7.3 Scrutinizing count of crimes based on zones:

```
####################################
# Number of Crimes for Different Areas
####################################
summary_communityArea = table(data.clean$Community.Area)
# sort areas according to number of crimes
summary_communityArea = sort(summary_communityArea,
                             decreasing = TRUE)
# get the 20 areas with most number of cimes
summary_communityArea = summary_communityArea[1:20]
barplot(height = summary_communityArea,
        names.arg = names(summary_communityArea),
        xlab = "Area Code",
        ylab = "Number of Crimes",
        main="Crimes for Each Area",
        col="yellow")
```

Figure 14: Code for bar plot tested

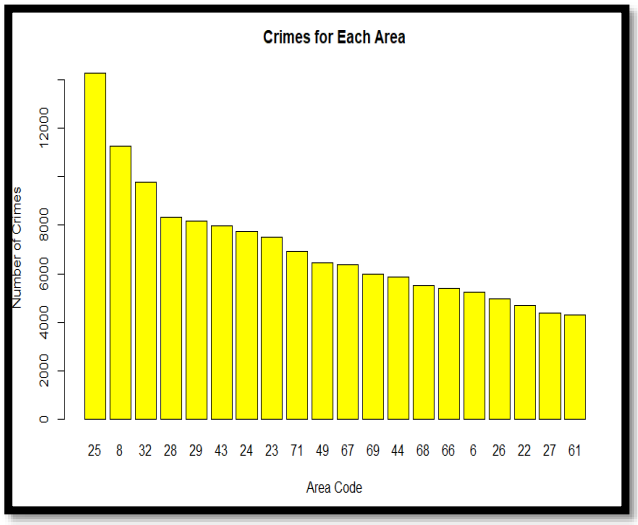

Figure 15: Bar Chart for crimes for each area

Moreover, focusing into intension of this test cases, we believe that this test cases are originally redirected towards protecting society. Thus, alternative test case is to extract location based crime rate, for planning to fight against it. If we get possibilities of types of crimes, in future government might lay its step to take precautions and convenience people for their safeguard.

7.4 Analyzing top three crimes among selected locations:

This final illustrative test case focuses on finding more filtered class of refinement of crimes by selecting top three crimes among 5 locations.

```
###################################
# top 3 types of crimes for 5 areas with most crimes
###################################areas = names(summary_communityArea[1:5])
areas = names(summary_communityArea[1:5])
y_array = rep(0,15)
label_array = rep(0,15)
area_array = rep(0,15)
color_array = rep(0,15)
color_pool=c('red','purple','blue','black','brown')
counter = 0
for (i in areas){
  tmp = data.clean[data.clean$Community.Area == i,]
  summary_tmp = sort(summary(tmp$Primary.Type),decreasing = TRUE)[1:3]
  for(j in seq(1:3)){
    area_array[counter*3+j] = i
    y_array[counter*3+j] = summary_tmp[j]
    label_array[counter*3+j] = names(summary_tmp[j])
    color_array[counter*3+j] = color_pool[counter+1]
  }
  counter = counter + 1
}
df = data.frame(area_array,
                y_array,
                label_array,
                color_array)
df$area_array = factor(df$area_array)
df$color_array = as.character(df$color_array)
dotchart(x=df$y_array,
         label=df$label_array,
         groups= df$area_array,
         xlab="C-Count",
         color=df$color_array,
         cex=.7,
         main="Top 3 types of crimes for 5 areas with most crimes")
```

Figure 16: Code implemented for dot chart

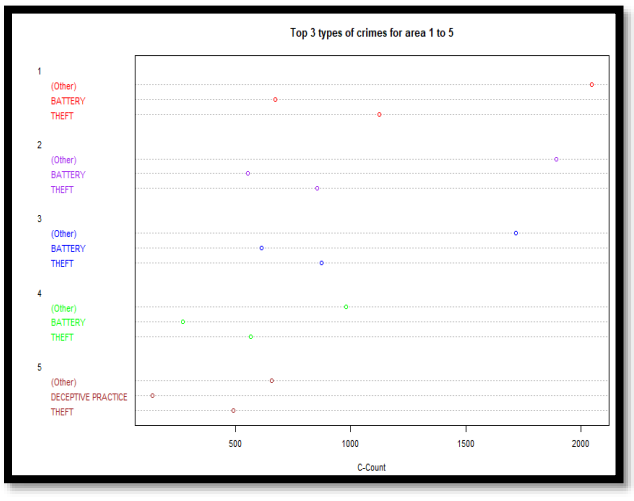

Figure 17: Dot chart for extracting top three crimes for selected area

This locations are not static, user can change it as per their necessity of investigation. As conversed in prior segment of this article location based evaluation is inevitable due to diversity in localities. Hence, the below stated dot chart is the exemplification of 5 locations organized in the database and corresponding top 3 crimes prevailing in that whereabouts.

8. CONCLUSION

Since its inception, Big Data is the new business and social science frontier. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to manage and process data. Moreover, it has become clear that "more data is not just more data", but that "more data is different" [7]. It is considered to be widely diversified, fast growing, challenging and vast stream of technology. Bid Data is an arena that always seeks additional revision, statistics and efforts to explore it which is going to mandate more along with technological advancement. We discussed in this paper some visions about the theme, and what we reflected the main concerns and challenges supposed to be faced by future.

This paper initiates a collaborative research effort to begin examining massive records using test cases which practically derives the stimulating critical evidences. We recognized some of the major subjects in Big Data storage, management, and processing with the help of R language. Latest technologies are striving for efficient data fetching in less time strap due to which effective coding with powerful programming languages like R is being motivated in this study. Moreover, RStudio is a supportive tool that provides hands-on for commendably working with R language. We elaborated administrative centric test cases which falls under boundaries of locality, frequency, temperament and background consideration, proportional to crime rate in Chicago. Adding to it the forte of R programming is confirmed by enlisting ways of graphical representation in this programming language. Basically, R is command line scripting language that can be made more interactive using RStudio tool.

Finally, four test cases chosen for this superior problems which are themselves a miniature of current state of government needs. These extinct of testimonials showcase how high quality academic research can address real-world problems and contribute solutions that are relevant and long lasting—exactly the challenge that our regime continues to face.

9. REFERENCES

[1] Zhao Y. R and data mining: Examples and case studies. Academic Press; 2012 Dec 31.

[2] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016 Oct 1.

[3] Driscoll K. From Punched Cards to" Big Data": A Social History of Database Populism. communication+ 1. 2012;1(1):1-33.

[4] Feuerlicht G. Database Trends and Directions: Current Challenges and Opportunities. InDATESO 2010 Apr 21 (pp. 163-174).

[5] Chen H, Chiang RH, Storey VC. Business intelligence and analytics: from big data to big impact. MIS quarterly. 2012 Dec 1:1165-88.

[6] Kim GH, Trimi S, Chung JH. Big-data applications in the government sector. Communications of the ACM. 2014 Mar 1;57(3):78-85.

[7] Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. InSystem sciences (HICSS), 2013 46th Hawaii international conference on 2013 Jan 7 (pp. 995-1004). IEEE.

[8] Katal A, Wazid M, Goudar RH. Big data: issues, challenges, tools and good practices. InContemporary Computing (IC3), 2013 Sixth International Conference on 2013 Aug 8 (pp. 404-409). IEEE.

[9] Wamba SF, Akter S, Edwards A, Chopin G, Gnanzou D. How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics. 2015 Jul 1; 165:234-46.

[10] Michael K, Miller KW. Big data: New opportunities and new challenges [guest editors' introduction]. Computer. 2013 Jun; 46(6):22-4.

[11] Lee J, Bagheri B, Kao HA. Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics. InInternational proceeding of int conference on industrial informatics (INDIN) 2014 Jul (pp. 1-6).

[12] Jatain A, Ranjan A. Statically Analysis on Big Data Using R Programming & SVM; 2017.

[13] Shafi A, Shah SF. Big Data Analysis And Deterministic Encryption Challenges. International Journal. 2017 Jul 1;8(7).

[14] Lichtenwalter RN, Lussier JT, Chawla NV. New perspectives and methods in link prediction. InProceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining 2010 Jul 25 (pp. 243-252). ACM.

[15] Adhatrao K, Gaykar A, Dhawan A, Jha R, Honrao V. Predicting students' performance using ID3 and C4. 5 classification algorithms. arXiv preprint arXiv:1310.2071. 2013 Oct 8.

[16] Fan W, Bifet A. Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter. 2013 Apr 30;14(2):1-5.

[17] Kumar P. Analysis and Exploring of different recent trends in processing of Big data. International Journal of Research and Engineering. 2016 Nov 4;3(8):62-5.

[18] Bifet A. Mining big data in real time. Informatica. 2013;37(1).

[19] Jadhav DK. Big data: the new challenges in data mining. Int J Innov Res ComputSci & Technol. 2013;1(2):39-42.

[20] Leung CK, MacKinnon RK, Jiang F. Reducing the search space for big data mining for interesting patterns from uncertain data. Inbig data (BigData congress), 2014 IEEE international congress on 2014 Jun 27 (pp. 315-322). IEEE.

[21] Orme D. The caper package: comparative analysis of phylogenetics and evolution in R. R package version. 2013 Nov 29;5(2):1-36.

[22] Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs. 2014 Jul 1;33(7):1123-31.