SciencePG
Science Publishing Group

Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya

Stephen Mangara Wainana[1], Joseph Njuguna Karomo[2, *], Rachael Kyalo[1], Noah Mutai[1]

[1]Department of Mathematics and Informatics, Taita Taveta University, Nairobi, Kenya

[2]Department of Pure and Applied Sciences, Kirinyaga University, Nairobi, Kenya

Email address:
stephenmangara93@gmail.com (S. M. Wainana), josekaromo@gmail.com (J. N. Karomo), rachealkate471@yahoo.com (R. Katwa),
ncheruiyot3@gmail.com (N. Mutai)
[*]Corresponding author

Abstract: Crimes have been the most dangerous threat to peace, development, human right, social, political and economic stability in Kenya. There is a great need to eradicate crime to facilitate development and counter all vices that are caused by crime. Efficient management of crime requires an adequate understanding of the patterns in which crime occur to put the appropriate measures in place for crime prevention. Crime has been in existence since the beginning of time hence will remain, and one of the solutions is to identify the pattern in which it occurs to prevent or counter it effectively as it occurs. The main objective of the study was to find out how different crimes are related. The study considered a number of data mining techniques which included; clustering, specifically k-means algorithm, mapping and APRIORI algorithm to analyze how different crimes are related and how often they occur. Crime cases were found to be decreasing over the years under study and counties with a high population reported higher number of crimes as compared to those with low population. The study suggested that these crimes could be controlled by directing more resources in the highly populated counties. The study leaves a research gap where the same crime data could be analyzed using time series methods since observed crime offenses are recorded alongside the time they occur.

Keywords: Crime, Clustering, Data Mining Techniques, Specifically K-means Algorithm,
Mapping and APRIORI Algorithm, Shiny App

1. Introduction

Data mining is widely used in many domains, such as retail, finance, telecommunication and social media [1]. It is, therefore, important as it empowers individuals to make valuable revelations autonomously without depending on analysts for their organizations. Reference [2] defines data mining as the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data to discover meaningful patterns. In other words, it is a process to identify interesting knowledge from large amounts of data [3]. There is often information "hidden" in a data set and is not evident hence human analysts analyzing the data manually can take a lot of time to identify useful information and in most cases, much of the data may not be analyzed.

When data mining techniques are used efficiently using standard analysis tools, much shorter time and labor is used than in traditional methods of data analysis [4].

Various data mining techniques have previously been used including association rules (relation) in making a correlation between more than one items of the same kind to discover patterns and classification which is used in describing multiple items to discover particular classes of clusters with correlating outcomes [5]. Several softwares are available for data mining, for instance, R Development Core Team, 2012, SPSS, Python and Perl. The purpose of this study was to use data mining techniques to analyze crime cases in Kenya.

2. Related Studies

There is a solid relationship that occur naturally between

crime and data mining techniques [6]. Many scholars explain that this is due to three factors. First, large volumes of data exist since security organs and agencies record a lot of data as possible while investigating crimes [7]. This enables them to monitor and investigate possible crimes and also prevent crimes that have already happened from happening again. Secondly, they have an ability to discover patterns in huge data sets which makes data mining a tool that is reliable as compared to manual processing of data which is time consuming and inappropriate for large data sets [8]. Thirdly, escalating economy see security organs facing a tuff budget which is not sufficient for them to pay their personnel and do proper investigation on crime. Data mining makes it easier and efficient for them to reveal crime patterns which are useful to them within their small budget. Reference [9] used clustering technique particularly, K means algorithm to detect crime patterns and speed up the process of solving crime. He also used semi-supervised learning technique for knowledge discovery from the crime records and to help increase the predictive accuracy.

An outlier-based data association technique for connecting criminal occurrences was used by [10]. They utilized this technique and clarified that an irregularity or exception stamp capacity is utilized to gauge perception extremeness. They connected this technique to the theft data from Virginia and Richmond and contrasted the outcome and a comparability based affiliation strategy. Their outcomes demonstrate that the outlier based data association technique is favorable.

Reference [11] combined the association rule and clustering were into a productive exploratory apparatus for the disclosure of spatial-transient patterns. They presented two techniques for this exploratory investigation and the detail calculations to successfully investigate geo-referenced information. They show the calculations with genuine crime data.

Reference [8] portrays a product system for building and applying information mining calculations to crime investigation issues. This structure gives particular concentrate on spatial information mining. The researcher gives a few motivations to legitimize this concentration including spatial questions are additional tedious, spatial examination is harder to do than investigations in light of quality matching. Spatial information mining can yield imperative prompt advantages for crime examination as violations have an innately spatial segment, and spatial investigation is a key to law requirement asset distribution [12].

A general casing work that demonstrates the relationship between data mining techniques connected in criminal and knowledge examination investigation and the crime sorts was explained by [7]. They recognize and organize eight crime sorts (criminal traffic offenses, sex crime, theft, fraud, arson, gang/drug offenses, violent crime, and cybercrime) in expanding request of open mischief on the level pivot. On the vertical hub, they orchestrate the methods in expanding request of investigation ability. They recognized four noteworthy classes of crime data mining methods: substance extraction, affiliation, forecast, and example perception. Every class speaks to an arrangement of systems for use in specific sorts of crime examination. They then distinguished the convergence

of the systems with the crime sorts signifying where every procedure could be successfully utilized for every crime type.

3. Methodology

3.1. Data

The study used secondary data retrieved from ICT authority website. The data used were; offenses per county for the year 2015, crime figure for the year 2012 to 2015 and monthly crime figures for the year 2012 to 2015.

3.2. Mapping

The study mapped the most frequent crime in all counties on the map of Kenya. Each county was shaded by a color which signified the level of crime in that county. For instance, red represented the county with the highest cases of crime

3.3. K-means Clustering

Reference [13] explains k-means clustering as a data mining algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. In this technique, observations are clustered into k groups where k is provided as an input parameter. The algorithm then assigns each and every observation to a cluster based on its proximity to the mean of that cluster. The mean of the cluster is computed again and the process continues.

The algorithm aims at minimizing an objective function which is:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where;

i. $\left\| x_i^{(j)} - c_j \right\|^2$ represents distance amongst a number of observations $x_i^{(j)}$ and the cluster mean.

ii. c_j is the cluster mean representing the distance of n data points from their cluster means.

iii. k represents a number of cluster means.

3.4. Association Rules

Reference [14] describes association rules as if-then statements that help uncover relationships between seemingly unrelated data in a relational database or another information repository. The study used Apriori algorithm which is a classic algorithm used in data mining for learning association rules in determining frequent crime and their relationship [15].

4. Data Analysis and Presentation

4.1. Crime Volume by Year

We are interested to see how the crime figures fared within the four years from 2012 to 2015 by presenting their totals
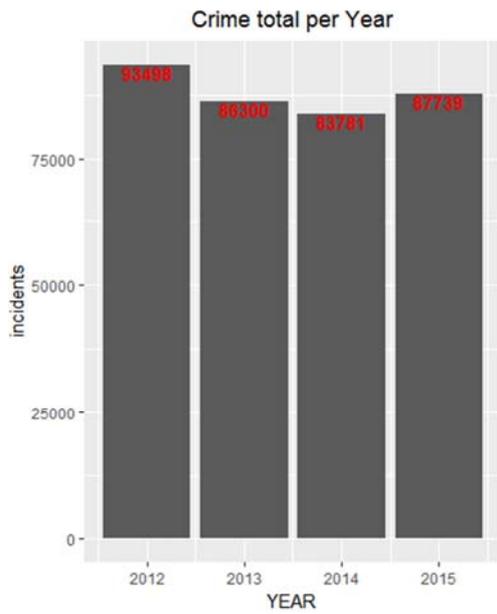
and their differences graphically.

Figure 1. Total crime by year.

From the bar graph above, it is noticed that the crime totals went at a decreasing rate from 2012 to 2014 but rose at a small but considerable margin in 2015. Though there was an increase on the year 2015, we can conclude that the crime volume per year generally went at a decreasing rate since the previous years never hit the volume of the year 2012.
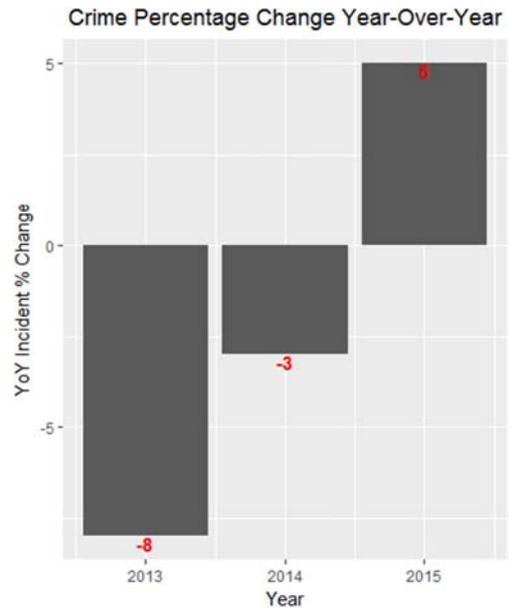

Figure 2. Percentage change in crime by year.

On crime analysis by percentage change year over year, the figures decreased by 8% in 2013 from 2012. The figures also decreased by 3% in 2014 from 2013 and there was an increase of 8% on the year 2015 from 2014.

4.1.1. Crime Figures over Years by Category

We present the crime figures over the four years graphically by categories which combine related crimes. The results are presented in the figure 3.
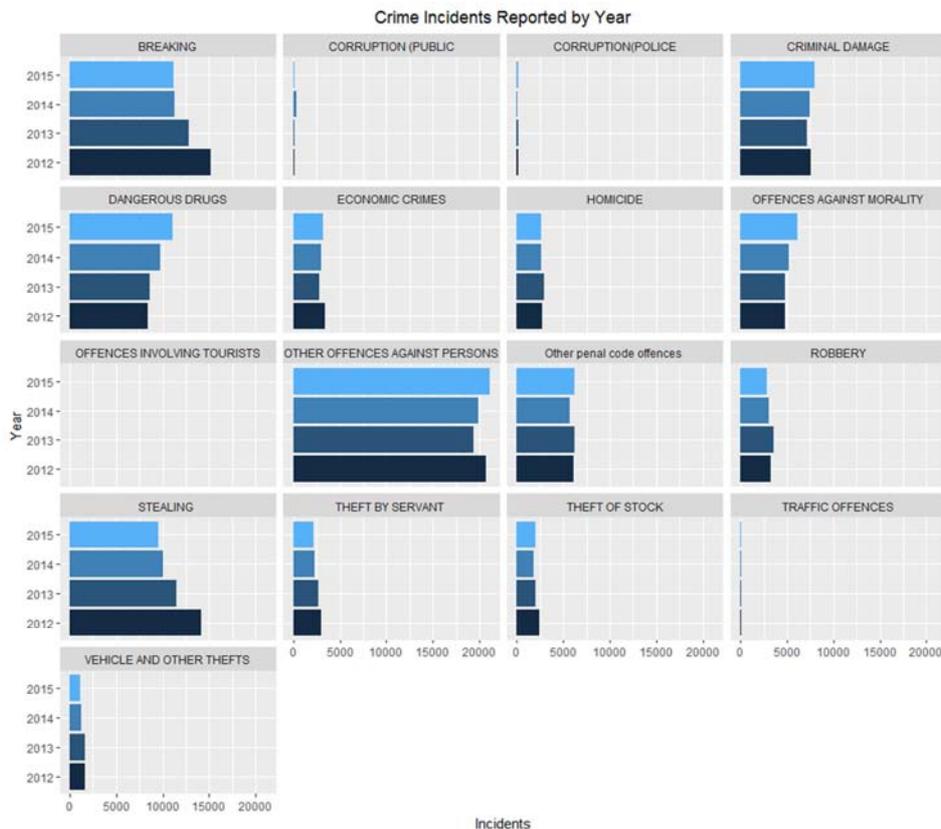

Figure 3. Crime incident category reported by year.

Statistics show that offenses against person recorded the highest figures continuously over the four years. The category included crimes like; assault, creating disturbance and affray. The year 2015 and 2012 recorded figures more than 20000 while the year 2014 and 2013 recorded crime figures above 18000 in offenses against person category. It was hard to explain the reason why this category recorded the highest figures but citing the types of crime in the category we may conclude that maybe there was a political influence on the numbers.

It is also noticed that breaking and stealing recorded a decreasing trend in figures over the period under review.

Dangerous drugs recorded an increasing trend from 2012 to 2015 with all the years recording figures above 7500 nationally. Offenses involving tourists recorded the minimum crime figures over the year with traffic offense recording little but much higher figures.

4.1.2. 2015 Crime Figures Conferring to Dominance

The most dominance or crime prevalent Counties in the year 2015 were recorded as Kiambu 4768 incidents, Nakuru 4384 incidents, Nairobi 4383 incidents, Meru 4215 incidents, Mombasa 3194 incidents, Bungoma 2852 incidents, Kakamega 2514 incidents and Muranga 2353 incidents.
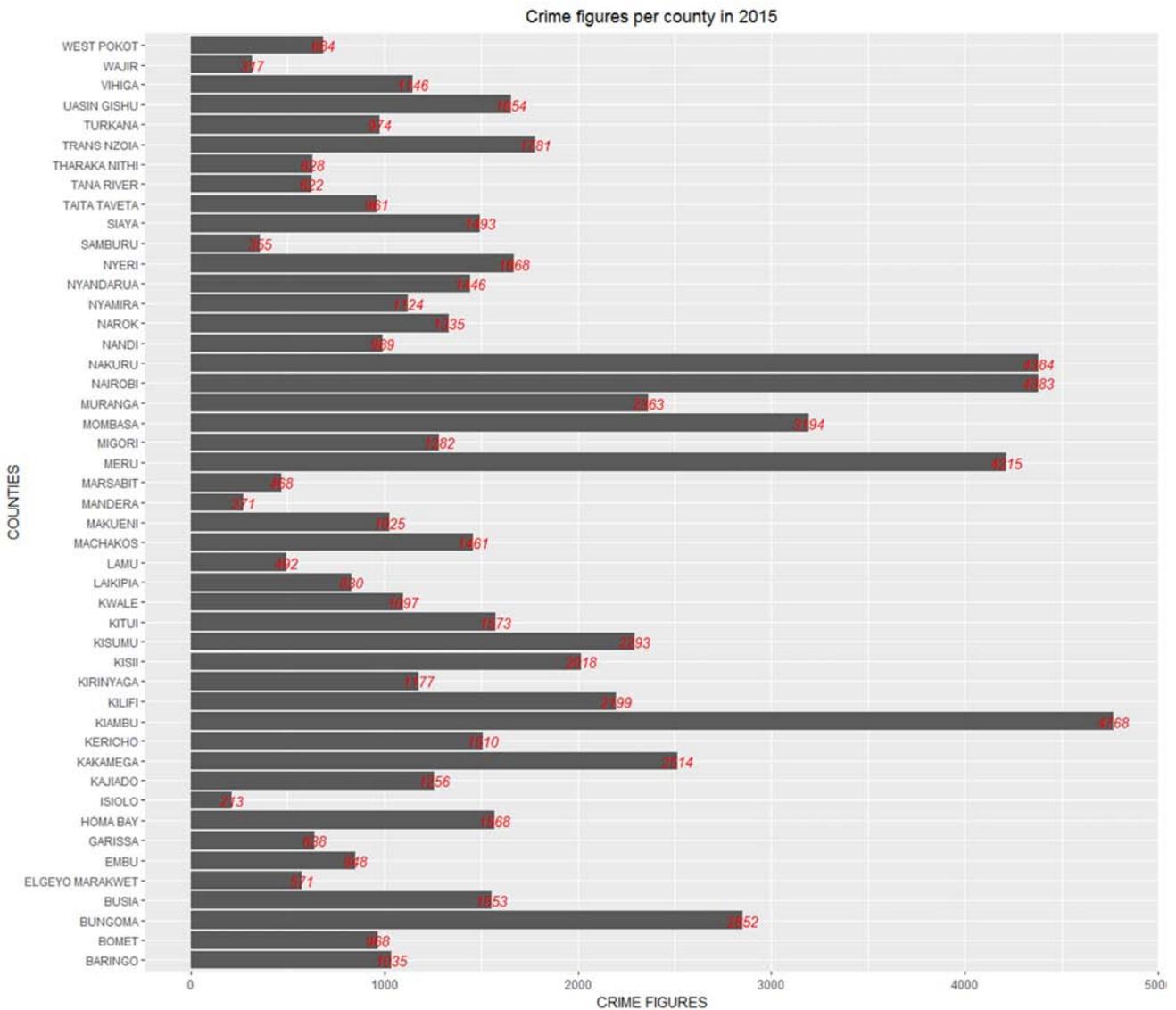

Figure 4. Crime figures per county for year 2015.

Among the counties that recorded least incidents in 2015 are Isiolo 213 incidents, Mandera 271 incidents, Wajir 317 incidents, Samburu 356 incidents and Marsabit 468 incidents.

4.1.3. Crime Analysis by Frequency

This section analyzes crime data for the year 2015 in all forty-seven counties in Kenya by histograms. The analysis is presented on a shiny app which is coded in R. The application represents a histogram of the crime category which the user selects from a drop-down list, a column of the data of selected variable, the structure of the data and the summary of the data in different tab panels.
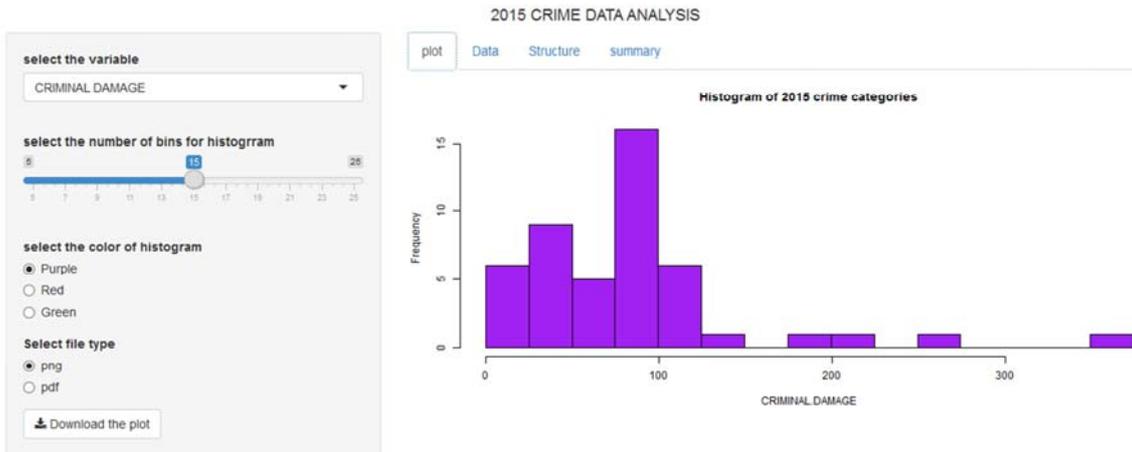
Figure 5. *A histogram of criminal damage in 2015.*

The selected category, criminal damage, on the shiny app shows that more than 15 counties had the crime figures ranging between 75 and 100. It is also clear that majority of the counties recorded less than 125 incidents while the minority of the counties recorded figures above 125 incidents with some recording up to 350 incidents.

Other variables can be chosen from the drop-down list where their histogram appear upon selection. The user can also perform any of the operations shown in the sidebar panel of the application. The operations include reducing or increasing the number of the bins or the bars on the histogram to optimize contrast, choosing the color among the provided options on the radio buttons and downloading the appearing plot of the histogram.

Figure 6. *Criminal damage data on shiny app in 2015.*

The data tab shows the data of the selected variable from the dropdown list. As it is in the histogram above it is also clear majority of the counties are recording figures below 125 incidents on criminal damage.
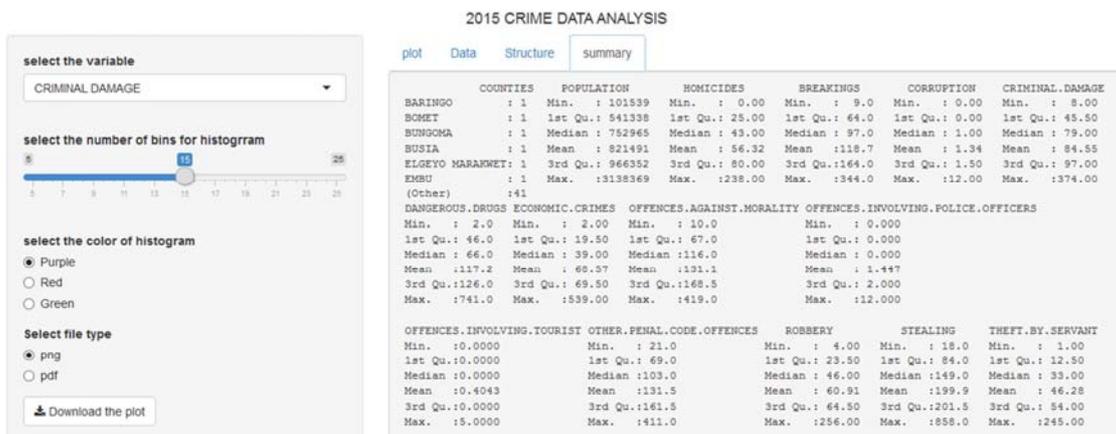

Figure 7. *Summary of the crime data in all county in 2015.*

The summary tab shows the summarized data on population for all the 47 counties. For instance, in Baringo County; the minimum population reads 101,539, the 1st quartile (lower quartile) is 541,338, the median is 752,965, the mean is 821,429, and the 3rd quartile (upper quartile) is 966,352, whereas the maximum population is 3,138,369. The summary data for the

rest of the counties is shown clearly on the shiny app.

4.2. 2015 Crime Analysis by Mapping

This section presents the mapping of population and crime totals in all counties.
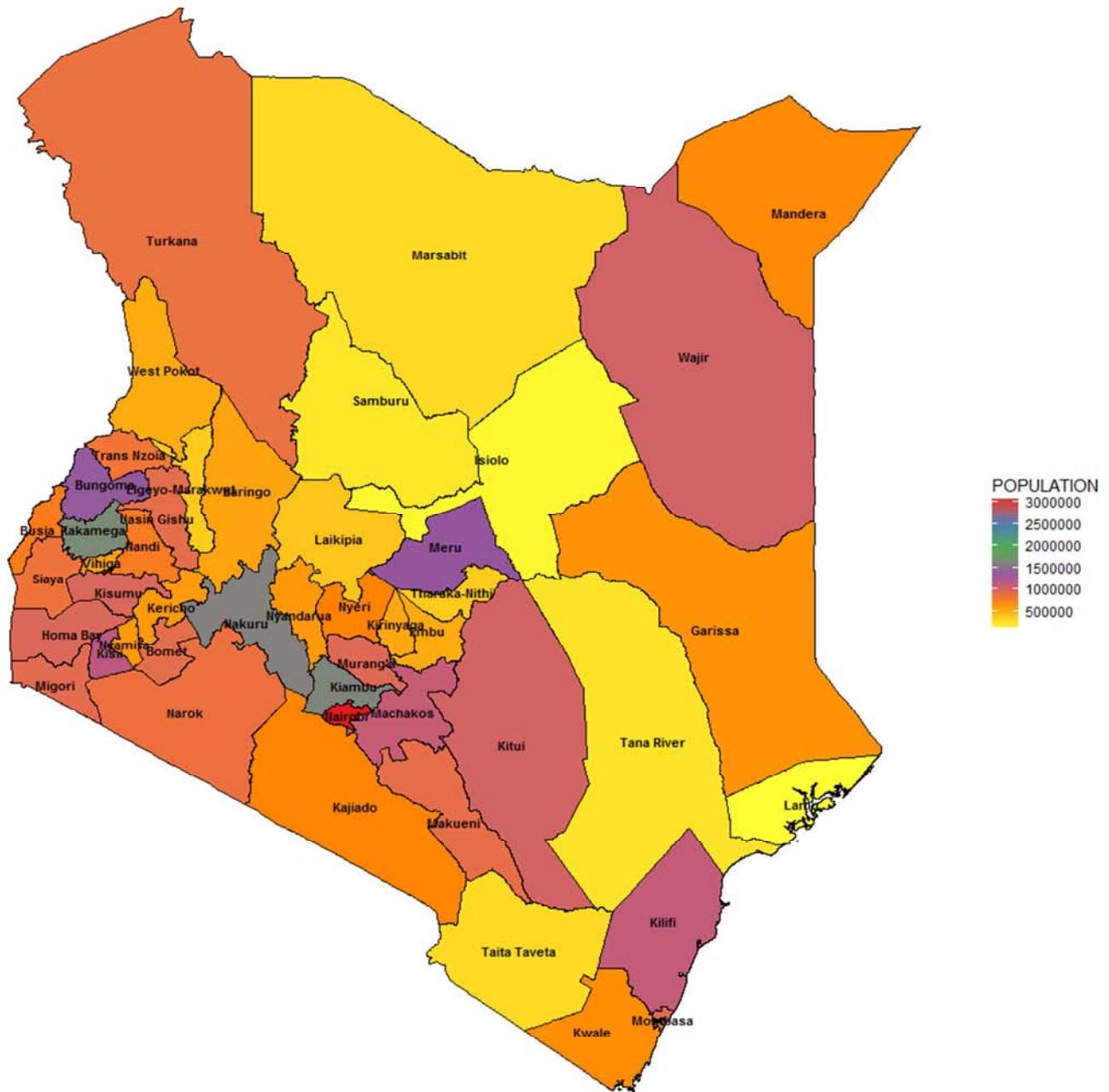
MAPPING COUNTIES POPULATION

Figure 8. *Mapping of the population in each county in year 2015.*

On mapping the population of all counties in the year 2015, the map show that Nairobi county has the highest population of about 3,000,000 people. Kiambu, Nakuru and Kakamega counties also had a big population of about

1,600,000 people.

Lamu, Isiolo, Samburu, Taita Taveta, Tana River, Tharaka Nithi and Elgeyo Marakwet counties had the lowest population of below 500,000 people.
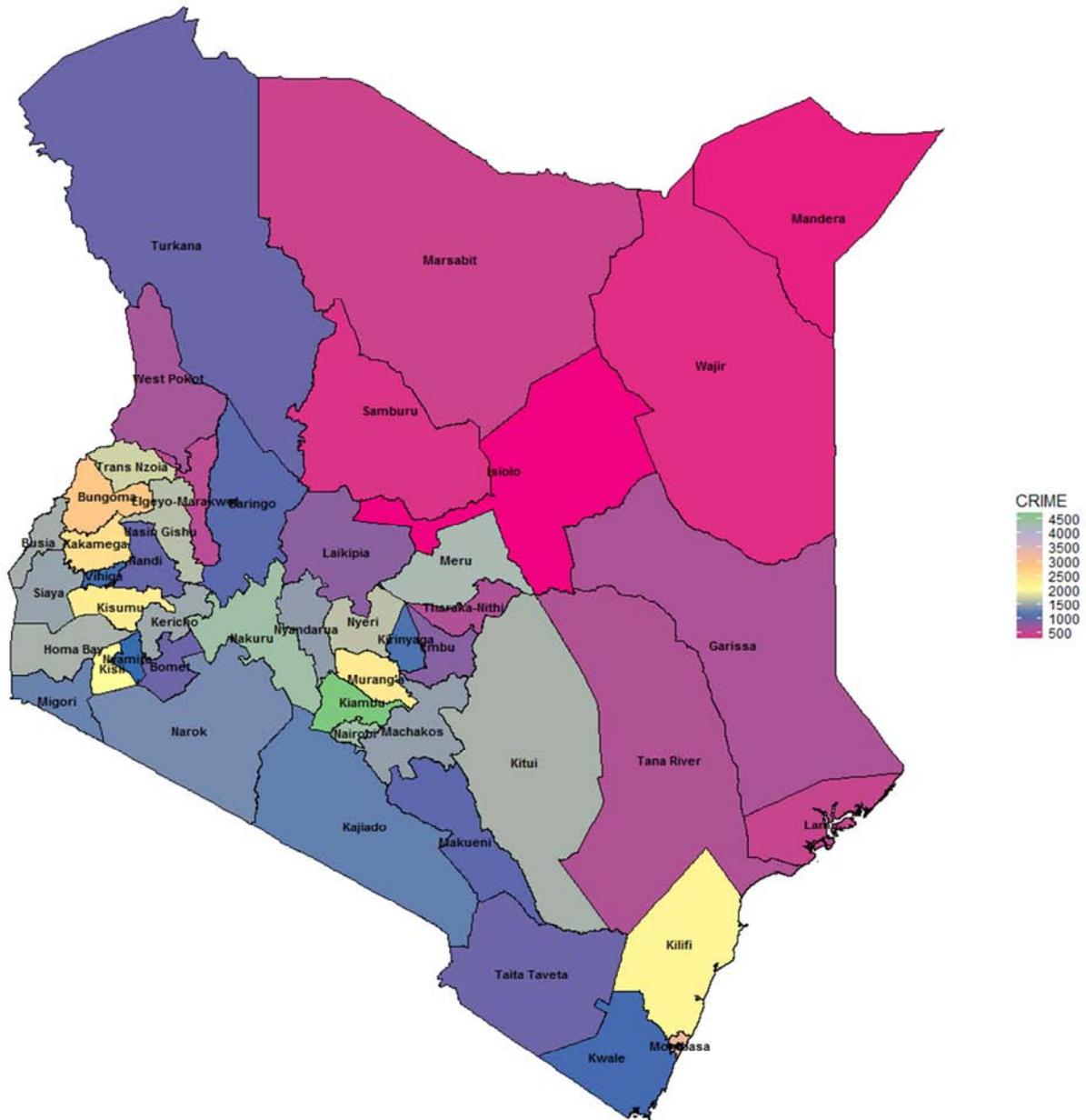
mapping crime totals for year 2015

Figure 9. *Crime figures for the year 2015 in all 47 counties.*

On mapping crime data on 2015, statistics show that Kiambu County recorded the highest crime figures followed by Counties like Nakuru, Nairobi and Meru recording crime figures above 3500. Among the Counties which recorded least crime figures from the map above are Isiolo, Mandera, Samburu, and Wajir recording figures below 1000 according to the scale on the map.

The maps show that crime is higher in Counties which got high population.

4.3. K-means Clustering

On clustering population and crime figures on the year 2015 in all 47 counties using k-means clustering algorithm on a shiny app, statistics show that there is formation of two different groups when k=2. The two groups have their centroids at (65000, 1600) and (1900000, 3900) as shown on the chat above. The two groups are on the basis of the counties with low population and the counties with higher population and it is clear that as the population increases there is an increase in crime figures. Also, the clusters show that most of the counties among the 47 counties of Kenya fall under the group of counties with low population as the first cluster on the chat got many points which are clustered together. The second cluster shows that only a few counties got very high population as the points in the clusters are a bit scattered.
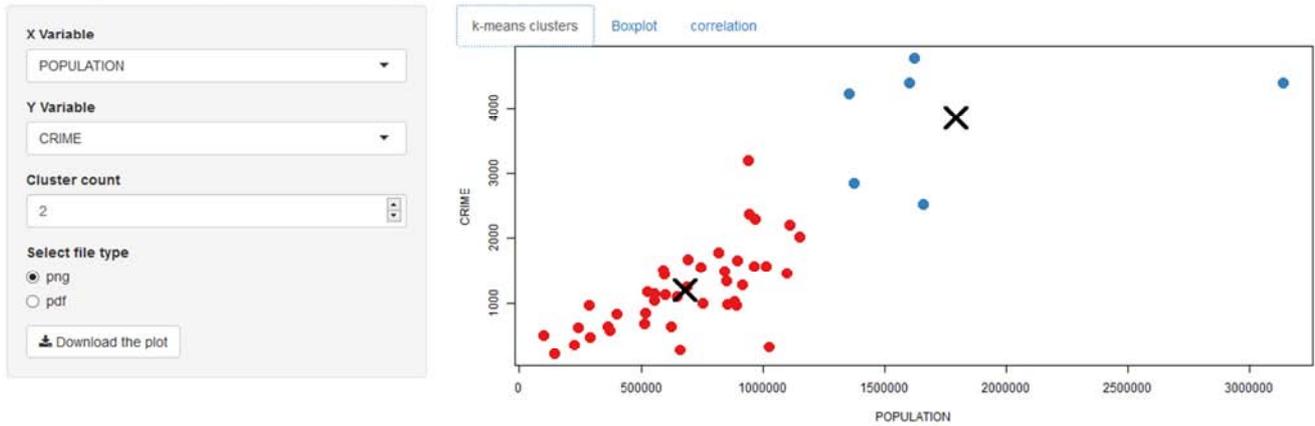
Figure 10. *Population and crime figures for 2015 when K=2.*

On increasing number of k to k=3, three distinct groups are formed with their centroids at around (480000, 900), (1000000, 1800) and (2000000, 4000) according to the chart shown below. Though there is change on the number of clusters the chat maintains its general view that the crime figures increases with an increase in population.
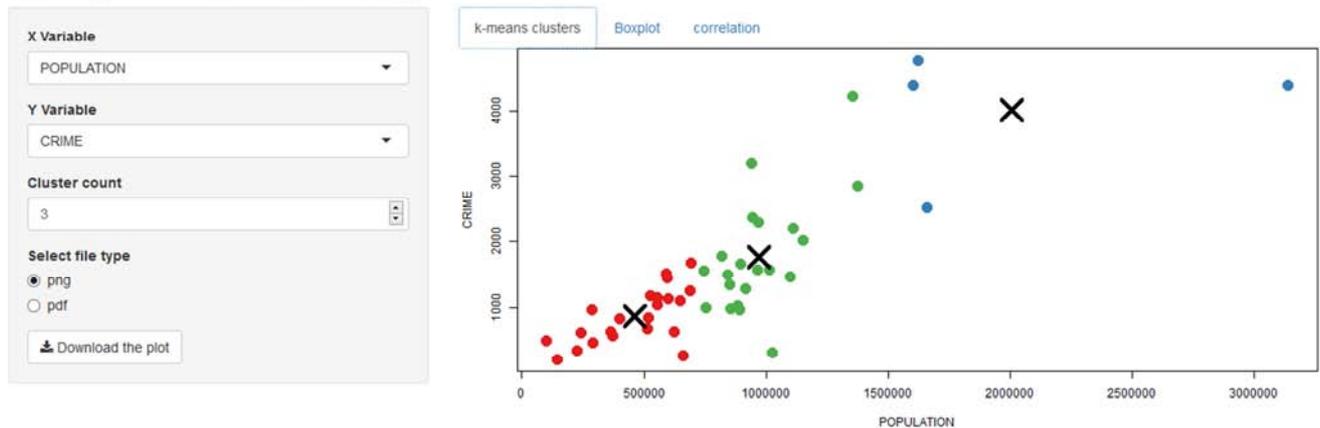

Figure 11. *Population and crime clustering for 2015 when K=3.*

The counties with the highest population are very few as shown on the cluster with blue points. Since the points represent the number of counties (47 counties), it is then clear that only four counties got population higher than 1500000 and all of which recorded crime figures above 2000 incidents in the year 2015.

The relationship between the two variables selected from drop-down lists on the shiny app can further be explained by their correlation coefficient by selecting correlation tab panel on the app as shown on the figure below.

Figure 12. *Correlation of population and crime figures on 2015.*

The correlation between the crime figures and population is 0.8079845 which is close to positive one (+1). An increase in population leads to an increase in crime figures.
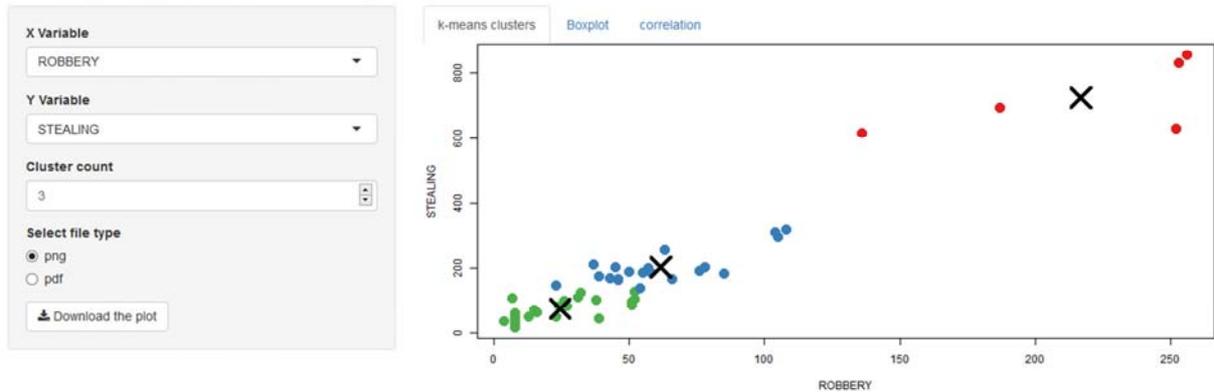

Figure 13. *Robbery and stealing clustering for 2015 when K=3.*

When k=3, statistic show that there are three clusters with their centroids at different points on the plot. The relationship between the two variables can be described as linear as increase in one crime causes an increase to the other.
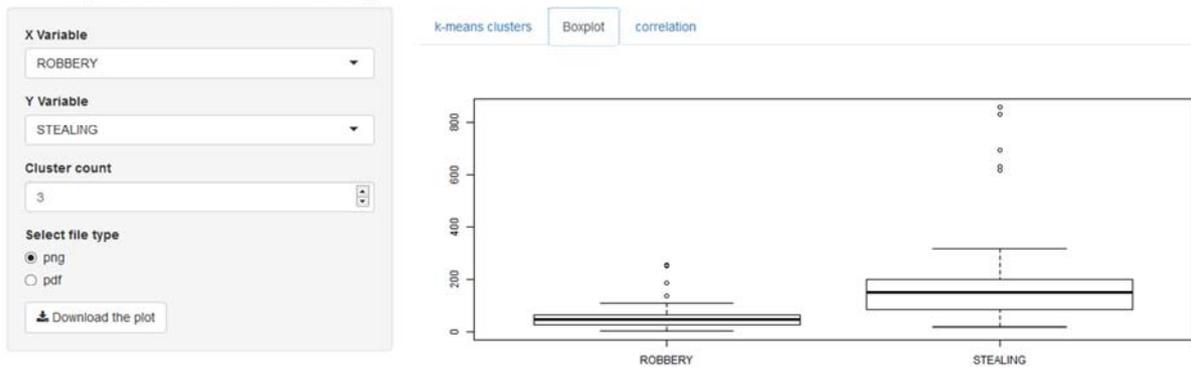

Figure 14. *Robbery and stealing quartiles.*

On the second tab panel of the app is a boxplot of the two selected crimes, in this case robbery and stealing. The boxplots shows all the quartiles of the crimes selected including 1st quartile, median and 3rd quartile. The bold horizontal line represents the median while the lower and upper parts of the box represents the 1st and 3rd quartile respectively. The lower line below the box and the upper line above the lines box represents the minimum and the maximum crime figures recorded on the year 2015.

In this case the median of robbery is 46 incidents while median of stealing is 149 incidents. The minimum value for robbery incidents recorded was 4 incidents and maximum of 256 incidents while the minimum value for stealing was 18 incidents and a maximum of 858 incidents.
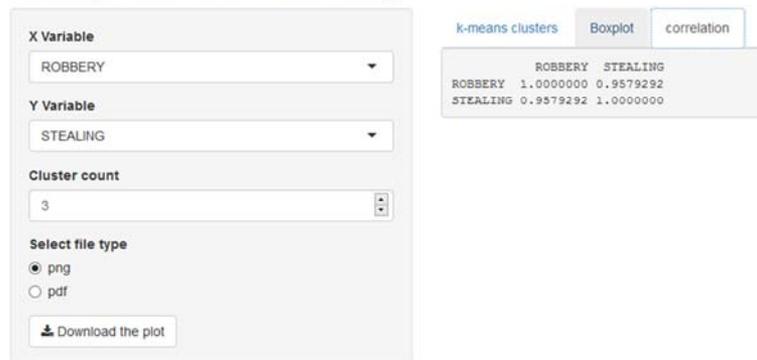

Figure 15. *Correlation of robbery and sealing.*

The two crime categories (robbery and stealing) have a strong positive correlation of 0.9579292 hence an increase in one crime leads to an increase of the other.
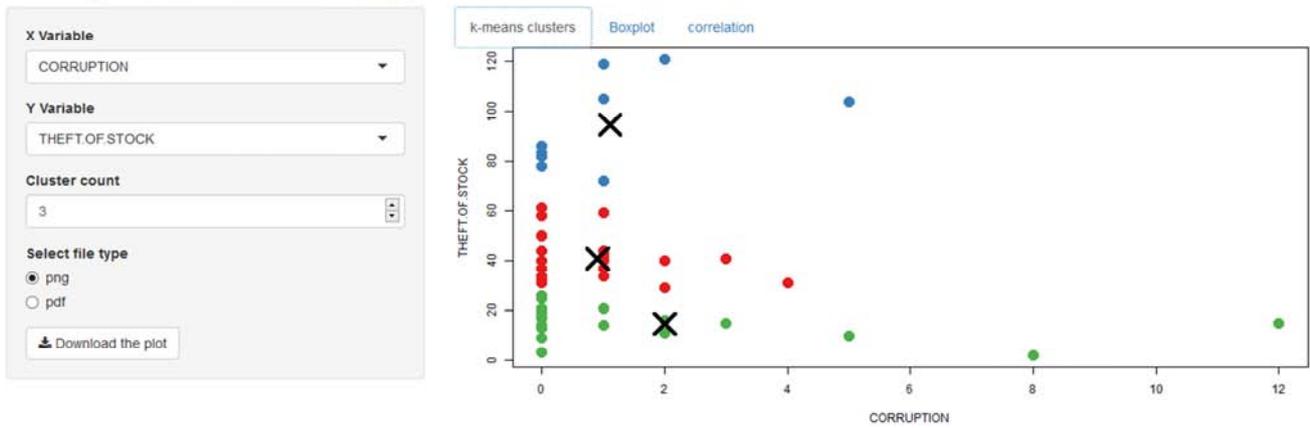

Figure 16. *Clustering of corruption and theft of stock when k=3.*

The two crime category forms thee distinct groups on application of k-means clustering with k=3. Though the groups are distinct, they appear to have no linear relationship or we cannot be able to explain one category inferring the other.

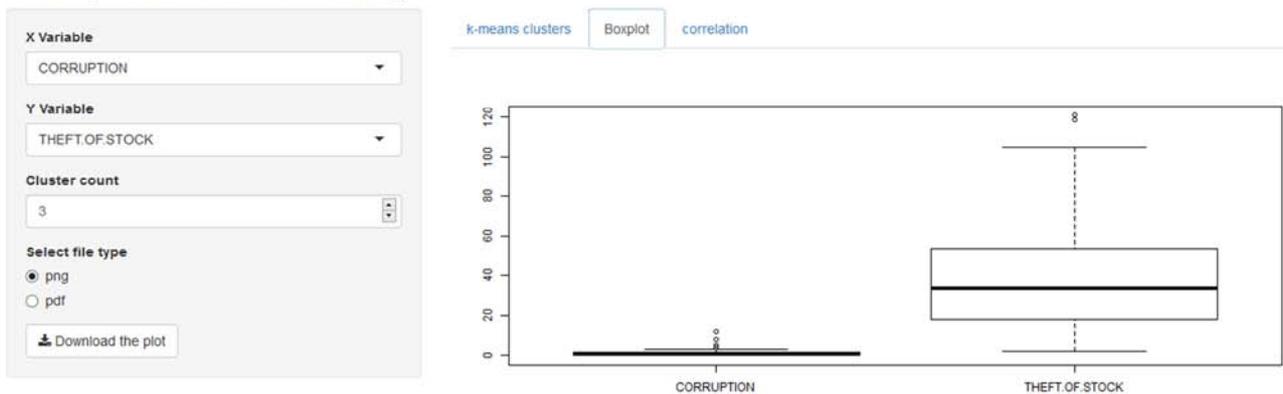***Figure 17.*** *Corruption and theft of stock boxplots.*

Above are the medians of the two crime categories of the selected crimes, that is, corruption and theft of stock on the boxplot tab panel.
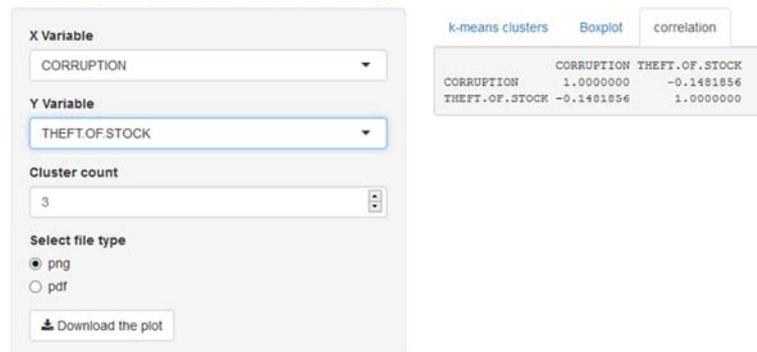

Figure 18. *Corruption and theft of stock correlation.*

The correlation of the two crime categories is -0.1481856 which is a close to 0. The two crime categories have very small negative correlation.

4.4. Association Rules Using APRIORI Algorithms

	rules	support	confidence	lift
254	{ Assault , Creating Disturbance } => { Affray }	0.05882353	1.0	17.0
255	{ Burglary , Other Breaking } => { House Breaking }	0.05882353	1.0	17.0
256	{ Burglary , House Breaking } => { Other Breaking }	0.05882353	1.0	17.0
257	{ House Breaking , Other Breaking } => { Burglary }	0.05882353	1.0	17.0
258	{ Stealing by Agents , Stealing by Directors } => { Ste...	0.05882353	1.0	17.0
259	{ Stealing by Agents , Stealing by employee/servant ...	0.05882353	1.0	17.0
260	{ Stealing by Directors , Stealing by employee/serva...	0.05882353	1.0	17.0
261	{ Theft from M/V , Theft of M/V } => { Theft of M/V par...	0.05882353	1.0	17.0
262	{ Theft from M/V , Theft of M/V parts } => { Theft of M...	0.05882353	1.0	17.0
263	{ Theft of M/V , Theft of M/V parts } => { Theft from M...	0.05882353	1.0	17.0
264	{ Theft from M/V , Theft of M/V } => { Theft of Motor c...	0.05882353	1.0	17.0
265	{ Theft from M/V , Theft of Motor cycle } => { Theft of...	0.05882353	1.0	17.0
266	{ Theft of M/V , Theft of Motor cycle } => { Theft from ...	0.05882353	1.0	17.0
267	{ Theft from M/V , Theft of M/V parts } => { Theft of M...	0.05882353	1.0	17.0
268	{ Theft from M/V , Theft of Motor cycle } => { Theft of...	0.05882353	1.0	17.0
269	{ Theft of M/V parts , Theft of Motor cycle } => { Thef...	0.05882353	1.0	17.0
270	{ Theft of M/V , Theft of M/V parts } => { Theft of Mot...	0.05882353	1.0	17.0
271	{ Theft of M/V , Theft of Motor cycle } => { Theft of M...	0.05882353	1.0	17.0
272	{ Theft of M/V parts , Theft of Motor cycle } => { Thef...	0.05882353	1.0	17.0
273	{ Obtaining by False Pretense , Other Fraud/Forgery ...	0.05882353	1.0	17.0
274	{ False Accounting , Other Fraud/Forgery Offences } ...	0.05882353	1.0	17.0

***Figure 19.** Sample APRIORI rules.*

Figure 19 shows some set of rules that were generated after application of APRIORI algorithm on crime figure for the year 2012-2015 data set. In nature, association rules are probabilistic or rather if (antecedent) and then (consequent) statements in which antecedent and consequent are item sets and have no any item in common as shown in figure 19. The table also shows support of rules on this sample as 0.05882353, confidence as 1.0 and lift as 17.0.

The table also shows that lift of every rule is greater than one which implies that there is a very strong relationship between crimes in antecedent and consequent part of different rules generated.
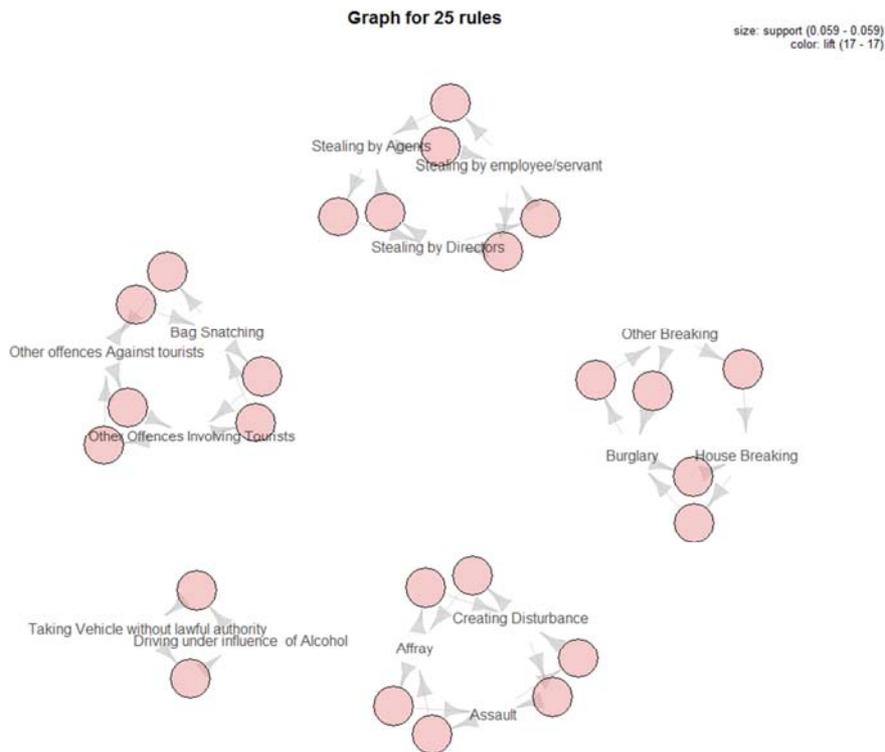

***Figure 20.** Crimes happening together frequently.*

The frequent set of crime was mined from the data set using APRIORI algorithm for association rules. Among the 25 rules plotted above shows frequent sets of crime that are associated or rather related and happens together frequent. For example, taking vehicle without lawful authority are associated and happens together frequently. Also, there is association between stealing by agents, stealing by employee/servant and stealing by directors and these crimes happens together frequently or in other words they are highly correlated.

5. Conclusions and Recommendations

5.1. Conclusions

Crime figures were decreasing from the year 2012 through 2014, but an increment was noted in the year 2015. The crime category with the highest crime figures throughout the period under study was 'other offenses against persons' followed by 'breaking' while 'stealing' and 'dangerous drugs' followed at a close margin. 'Offenses involving tourists' had the list crime figures throughout the years. Kiambu, Nakuru, Nairobi and Meru counties had the highest crime figures in the year 2015 with counties like Isiolo, Wajir and Mandera having the least crime numbers in the same year also, Nairobi county had the highest population, followed by Kiambu, Nakuru, Kakamega and Nakuru counties respectively while counties that registered the least population were Lamu, Isiolo, and Samburu as per the descriptive statistics analysis findings.

Upon applying the k-means algorithm on 2015 crime data set, crime categories like 'robbery' and 'stealing' have distinct groups that have a very strong linear relationship. Also, population and crime total for the year 2015 form groups that are strongly related. There are also crime categories that are not strongly related and they form k groups that don't have a linear relationship. For example, 'corruption' and 'theft of stock' have a low negative correlation.

APRIORI algorithm shows that many crimes are associated.

5.2. Recommendations

The ministry of internal security should consider directing more resources in counties with the high population as compared to the lowly populated counties since it is evident from the conclusions that highly populated counties report a high number of crimes as compared to those with low population.

The ministry of internal security should be aware of the crimes that are related this will help them successfully minimize crime incidents in their counties since, after the occurrence of a certain crime, they will be able to prevent a related crime from happening.

5.3. Suggestions for Further Study

Time series analysis can further be used to analyse crime data since observed crime offenses are recorded together with time they are observed. Methodologies such as box-Jenkins methodology for time series can be used to predict the expected number of crimes in the future from past crime observations.

Text mining techniques can be used in text articles to identify crimes that are frequently mentioned in social media platforms. There is a high probability that the crimes mentioned frequently also occur frequently and maybe depending on other crimes.

References

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

[2] P.-N. Tan, M. Steinbach and. V. Kumar, "Data mining cluster analysis: basic concepts and algorithms," *Introduction to data mining,* pp. 487--533, 2013.

[3] M. Kaufmann, J. Han and J. Pei, Data mining: concepts and techniques, Morgan Kaufmann, 2000.

[4] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to data mining, Pearson Education India, 2016.

[5] M. Brown, "Data mining techniques.," *Developer Works, IBM Corporation,* pp. 1-16, 11 December 2012.

[6] C. C. Yang and o. D. Ng, "Terrorism and crime related weblog social network: Link, content analysis and information visualization," in *2007 IEEE Intelligence and Security Informatics,* 2007.

[7] H. Chen, W. Chung,. J.. J. Xu, G. Wang,. Y. Qin and M. Chau, "Crime data mining: a general framework and some examples," *computer,* vol. 4, pp. 50--56, 2004.

[8] D. E. Brown, "The regional crime analysis program (RECAP): a framework for mining data to catch criminals," in *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218),* 1998.

[9] S. V. Nath, "Crime pattern detection using data mining," in *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops,* 2006.

[10] S. Lin and. D. E. Brown, "An outlier-based data association method for linking criminal incidents," *Decision Support Systems,* vol. 41, no. 3, pp. 604--615, 2006.

[11] V. Estivill-Castro and I. Lee, "Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data," in *Proc. of the 6th International Conference on Geocomputation,* 2001.

[12] P. L. Brantingham and P. J. Brantingham, "Environment, routine and situation: Toward a pattern theory of crime," *Advances in criminological theory,* vol. 5, no. 2, pp. 259--94, 1993.

[13] A. Shafeeq and K. Hareesha, "Dynamic clustering of data with modified k-means algorithm," in *Proceedings of the 2012 conference on information and computer networks,* 2012.

[14] A. Bhardwaj, A. Sharma and V. Shrivastava, "Data mining techniques and their implementation in blood bank sector--a review," *International Journal of Engineering Research and Applications (IJERA),* vol. 2, no. 4, pp. 1303--1309, 2012.

[15] C. Li, N. Ding, G. Zhang and L. Li, "Association Analysis of Serial Cases Based on Apriori Algorithm," in *Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence,* 2019.