Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 11 Issue 2 Year 2021 Pages 71-80
DOI: http://dx.doi.org/10.31940/matrix.v11i2.2449
URL: http://ojs.pnb.ac.id/index.php/matrix

Exploratory data analysis of crime report

Irwan Setiawan [1]*, Suprihanto [2]

[1,2] *Department of Computer Engineering and Informatics, Politeknik Negeri Bandung, Indonesia*

***Corresponding Author: *irwan@jtk.polban.ac.id*

Abstract: Visualization of data is the appearance of data in a pictographic or graphical form. This form facilitates top management to understand the data visually and get the messages of difficult concepts or identify new patterns. The approach of the personal understanding to handle data; applying diagrams or graphs to reflect vast volumes of complex data is more comfortable than presenting over tables or statements. In this study, we conduct data processing and data visualization for crime report data that occurred in the city of Los Angeles in the range of 2010 to 2017 using R language. The research methodology follows five steps, namely: variables identification, data pre-processing, univariate analysis, bivariate analysis, and multivariate analysis. This paper analyses data related to crime variables, time of occurrence, victims, type of crime, weapons used, distribution, and trends of crime, and the relationship between these variables. As the result shows, by using those methods, we can gain insights, understandings, new patterns, and do visual analytics from the existing data. The variations of crime variables presented in this paper are only a few of the many variations that can be made. Other variations can be performed to get more insights, understandings, and new patterns from the existing data. The methods can be performed on other types of data as well.

Keywords: data visualization, exploratory analysis, visual analytics, data analysis, crime report

History Article: Submitted 28 March 2021 | Revised 29 April 2021 | Accepted 8 June 2021

How to Cite: I.Setiawan and Suprihanto, "Exploratory data analysis of crime report," *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 11, no. 2, pp. 72-81, 2021.

Introduction

Data visualization is the display of data in the form of images or graphics that can help decision-makers to be able to understand data visually and get new patterns hidden in the data. Visualization of complex and large amounts of data is more manageable for humans to understand when using pictures or graphics compared to being displayed in tabular or written form.

In a modern digital era, visions used in critical organization decision making gathered from Exploratory Data Analysis (EDA). EDA is the technique of studying one or more datasets to recognize the underlying structure of the data carried there [1]. EDA can be used to identify hidden patterns and correlations among variables in the data and assist people in confirming predictions from the data. Over the last few decades, academics have introduced various tools and techniques to visualize hidden correlations among data variables using simplistic diagrams and charts [2]–[8]. Visual data analysis aid domain-specific data interpretation such as analysis of CRISPR/Cas9 screens [2], analysis of container shipping slot bookings [9], analysis of executive functions during childhood [10], analysis of kindergarten students log data [11], sodium and potassium coronate stability [12], fault injection campaigns [13], employee demographics and earnings [14], airport waiting times [15], analysis of medical data [16] to perform analytics tasks, and analysis of Airbnb's super host profile [17]. Crime is a risk that must be faced and managed. The results from the EDA can be used as input for performing identification, analysis, and plans for handling potential risks that exist in the city [18].

Unemployment, poverty, urbanization, and rapid population growth are the primary causes of social di-lemmas. One of these problems implicit in every city is a crime. For example, as reported in [19], Indonesian police reported that the crime rate per 100,000 population in 2017 is 129 people. Although it experienced a decline from 2016, which numbered 140 people, the decline occurred less than 10 percent to lessen criminality rates, police have collected a large

amount of data to analyze. The study of criminal activity and the forecast of the number of crimes remains one of the most exciting problems for researchers. Research related to crime has been widely carried out [20], [21], [22], [23].

In this study, data processing and visualization were carried out for crime report data that occurred in the city of Los Angeles in the range of 2010 to 2017. Visualization of data related to crime variables, time of occurrence, victims, types of crime, weapons used, distribution, and crime trends, and the relationship between these variables is elaborated to be further used by decision-makers to conduct further analysis.

Methodology

The research methodology, as shown in Figure.1, follows five steps, namely: variables identification, da-ta pre-processing, univariate analysis, bivariate analysis, and multivariate analysis.

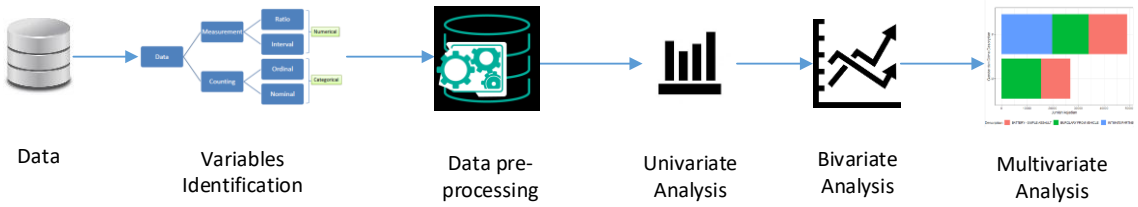

Figure 1. EDA steps

Variables identification: this is an essential step to clearly distinguish and understand the meaning of each variable in a dataset before analyzing the data. Datasets commonly have numerical, ordinal, or nominal variables [1]. An essential characteristic of numerical data is that we can apply many mathematical operations to it. A nominal, categorical, or factor variable cannot apply to mathematical operations. Ordinal variables, also referred to as ordered categorical variables or ordered factors, is a non-numeric value but possess an inherent order.

Data pre-processing: this is the second step of the EDA process. This process performs data integration (such as finding redundant attributes and tuple duplication and inconsistency), data cleaning, imputation of missing values [24], dealing with noisy data, and data reduction [25].

Univariate data analysis: the objective of the univariate analysis is to get a better understanding of each attribute. In this step, we analyze each attribute to understand how each attribute looks like. We use the ggplot2 package to visualize the data. Bivariate data analysis: the objective of the bivariate analysis is to analyze relationships between two attributes. In this step, we compare two attributes to analyze the correlation between them. We use the ggplot2 package to visualize the data. Multivariate data analysis: the objective of multivariate analysis is to get a more in-depth investigation from more than two attributes. In this step, we compare three or more attributes to analyze the correlation between them. We use the ggplot2 package to visualize the data.

Results and Discussions

The raw data are collected from the Los Angeles Police Department. The dataset reflects incidents of crime in the City of Los Angeles from 2010. The dataset represents a transcribed report from the original crime report, which is typed on paper. The original data includes over one point nine million data points for the period of 1st January 2010 to 25th November 2019. The crime report attribute includes division of records number made up of a two-digit year and five digits area ID, date reported, date occurred, time occurred, an area which referred to as geographic areas within the department, area name which represents a name designation that references a landmark of the surrounding community that is responsible for, reporting district number made up of a four-digit code that represents a sub-area within a geographic area, crime code which indicates the crime committed, modus operandi, victim age, and sex, victim descent, premise code which represents the type of structure, vehicle, or location where the crime took place, the weapon used, the status of the case, criminal code, the location which represents the

street address of crime incident rounded to the nearest hundred blocks to maintain anonymity, cross street, latitude, and longitude.

The data pre-processing step consists of removing the missing data, changing the data type of some at-tributes, rename the name of attributes, and finding redundant attributes and tuple duplication and inconsistency. In this study, because there are many NULL values in the data range 2018 to 2019, the range of data to be explored is from 1st January 2010 to 31st December 2017. From 1,900,312 crime report data will only be used 1,895,619 data.

Using R programming language and charts, we can analyze the crime data according to its variables, time of occurrence, victims, type of crime, weapons used, distribution of incidents, and trends of crime. Figures 2, 3, 4, and 5 show the distribution of crime incidents per year, per month, per day, and date respectively, from 2010 through 2017. Figures 2, 3, 4, 5, 6, 7, and Table 1 are an example of the results of univariate analysis. Figures 9, 12, and 14 are an example of the results of bivariate analysis. Figures 10, 11, and 13 are an example of the results of multi-variate analysis.

Figure. 2 shows the number of crime incidents distributed between 2010 and 2017. At the end of 2010, the Los Angeles Police Department recorded 208,883 crime reports. The number of crime incidents decreases significantly and reaches the minimum at the end of 2013. Since then, the number of crime incidents increases and reaches the maximum at the end of 2017. Signifi-cantly from 2014 to 2015, the number of crime incidents increased by almost ten percent.

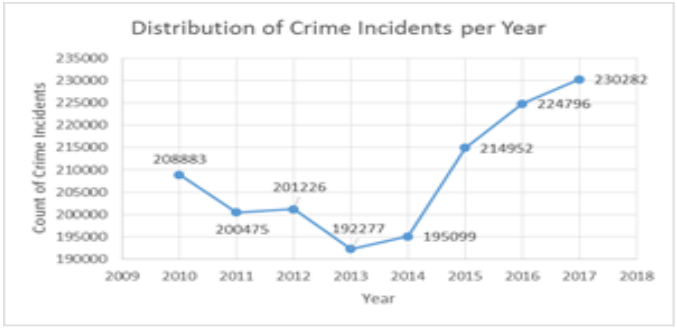

Figure 2. Distribution of crime incidents per year between 2010 and 2017

Figure 3 illustrates the distribution of crime per month from 2010 through 2017. The chart shows that February is the lowermost month of crime incidents (141,088) in Los Angeles. For the other months, it fluctuates between 150,000 and 165,000.

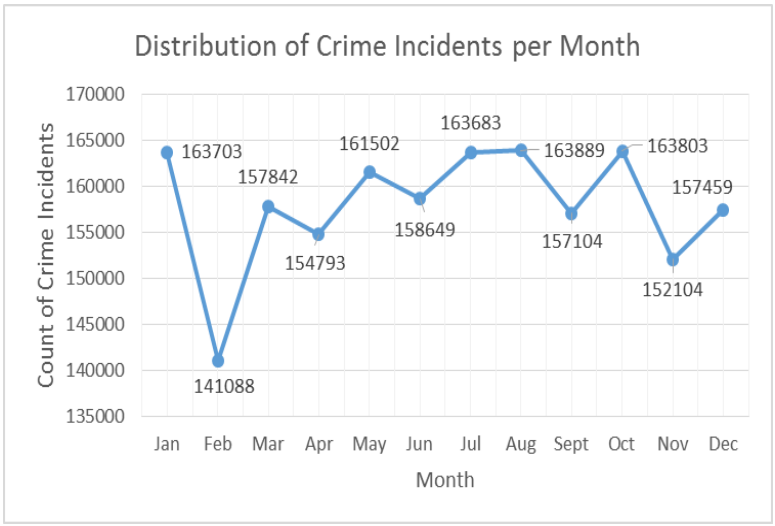

Figure 3. Distribution of crime incidents per month from 2010 through 2017

Figure 4 illustrates the distribution of crime incidents per day from 2010 through 2017. From Monday to Thursday, crime incidents fluctuate between 26,500 to 27,000, and it increases

surprisingly to 291,092 incidents on Friday. Then, it goes down to the lowest on Sunday (260,735).

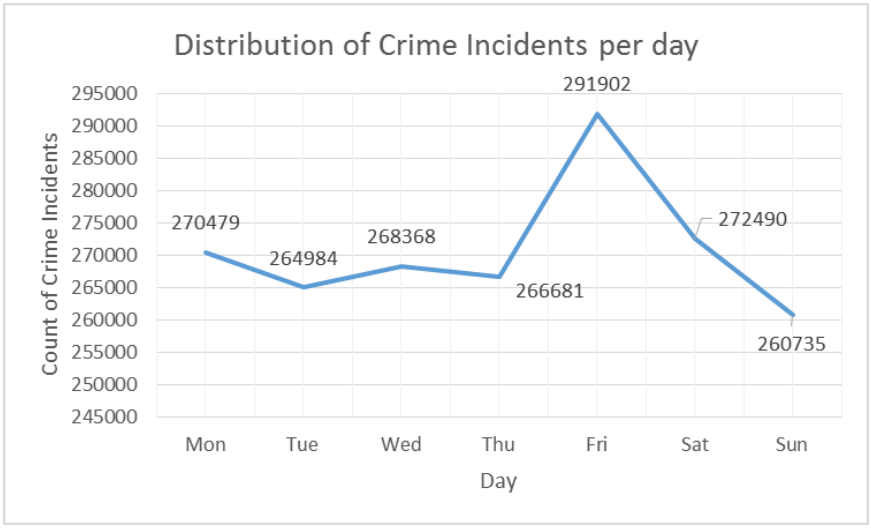

Figure 4. Distribution of crime incidents per day from 2010 through 2017

Figure 5 shows the highest ten of count of incidents according to the time when the incidents occur for the years 2010 through 2017. Interestingly enough, we can see from Figure 5. that crimes mostly happen at 12 o'clock.

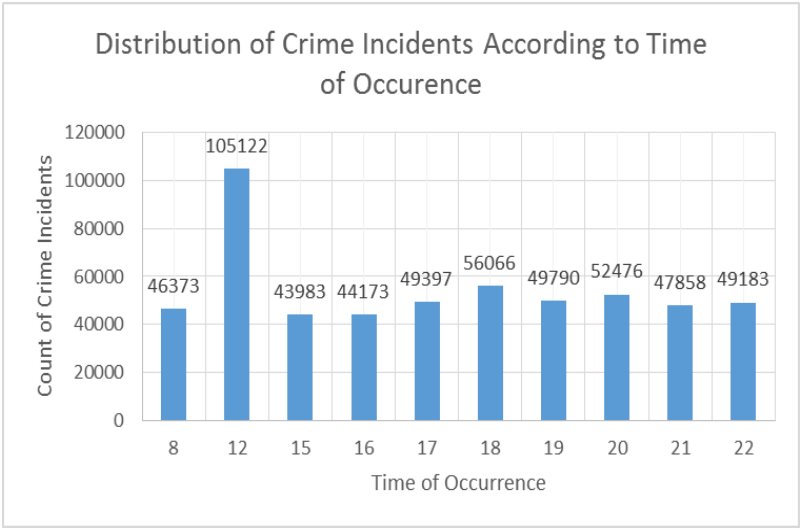

Figure 5. Distribution of crime incidents according to the time when the incidents occur during 2010 through 2017

Figure 6 shows that crime incidents from the 2nd through the 30th day of every month fluctuates between 56,000 and 66,000 incidents. Surprisingly, we can see that most crimes occur on the 1st of every month (96,879 incidents), and the lowest is on the 31st (36,851 incidents).

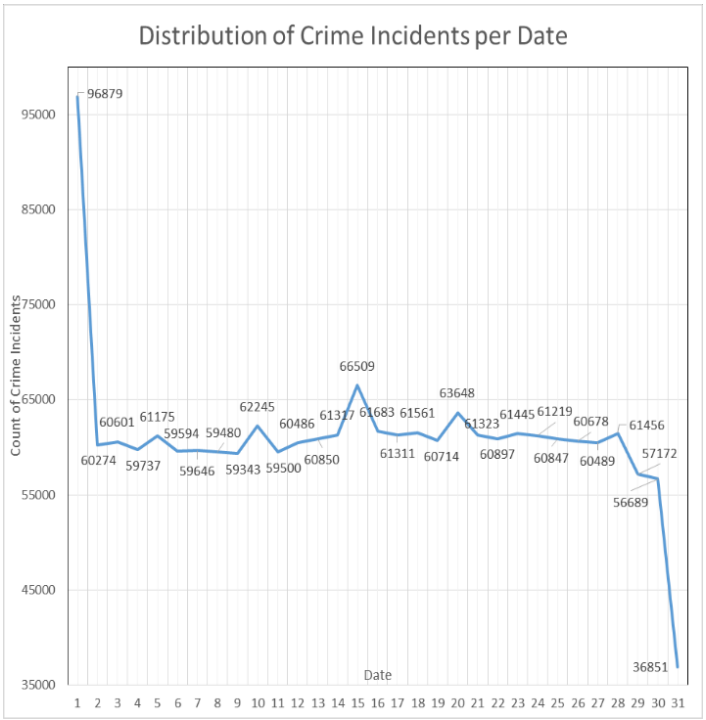

Figure 6. Distribution of crime incidents per date during 2010 through 2017

Figure 7 takes a detailed look at what type of crimes happen at 12 o'clock. It is found that most of the crimes that happen at that time are theft of identities, such as theft of name, theft of identifying number, theft of credit card number, and theft of social security number.

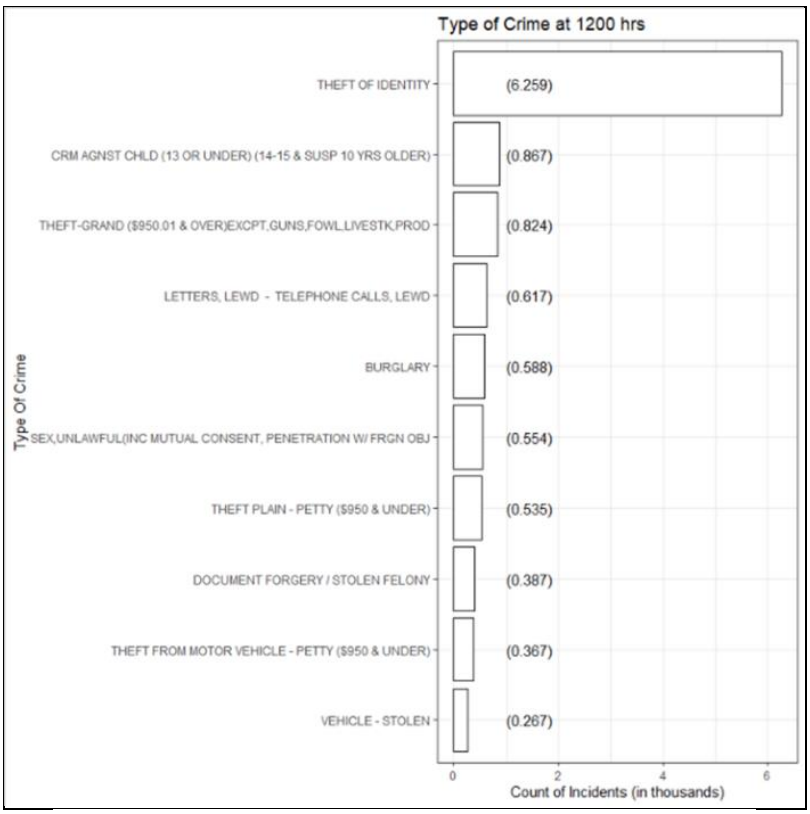

Figure 7. The top ten types of crimes happen at 12 o'clock

Figure 8 describes the distribution of crime incidents according to the gender of the victim from 2010 until 2017. It shows that through those years, the count of crime incidents is nearly the same for the male victim and female victim, with males suffers more crimes than females.

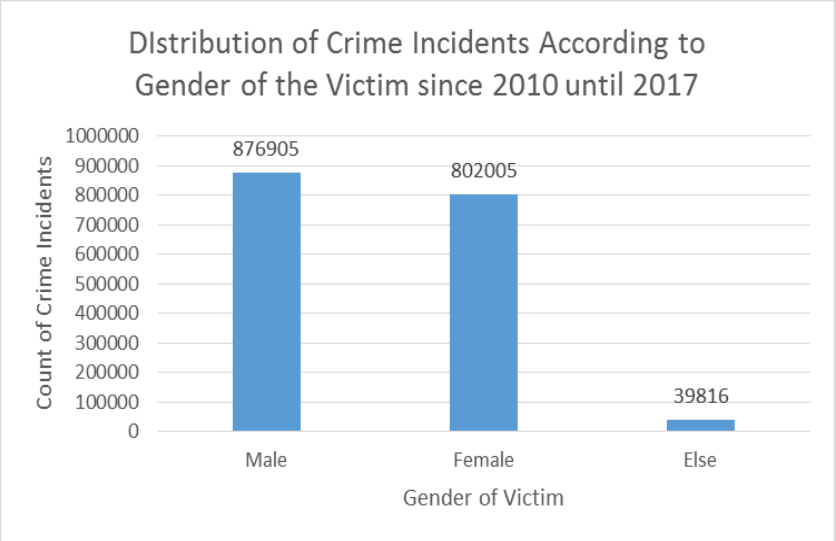

Figure 8. Distribution of Crime Incidents According to Gender of the Victim from 2010 until 2017

Figure 9 takes a look at the distribution of crime incidents according to the gender of the victim and their ages from 2010 until 2017. It shows that males and females between ages 20 and 55 suffer more crimes. Fe-males age 20 and 35 suffer most crimes. Taking a more in-depth look at what type of crimes happens to females age between 20 and 35, as shown in Figure 10, we find that they suffer the most from intimate partner-simple assault. As for males of that range of age, we find that most crime that happens to them is burglary from a vehicle.

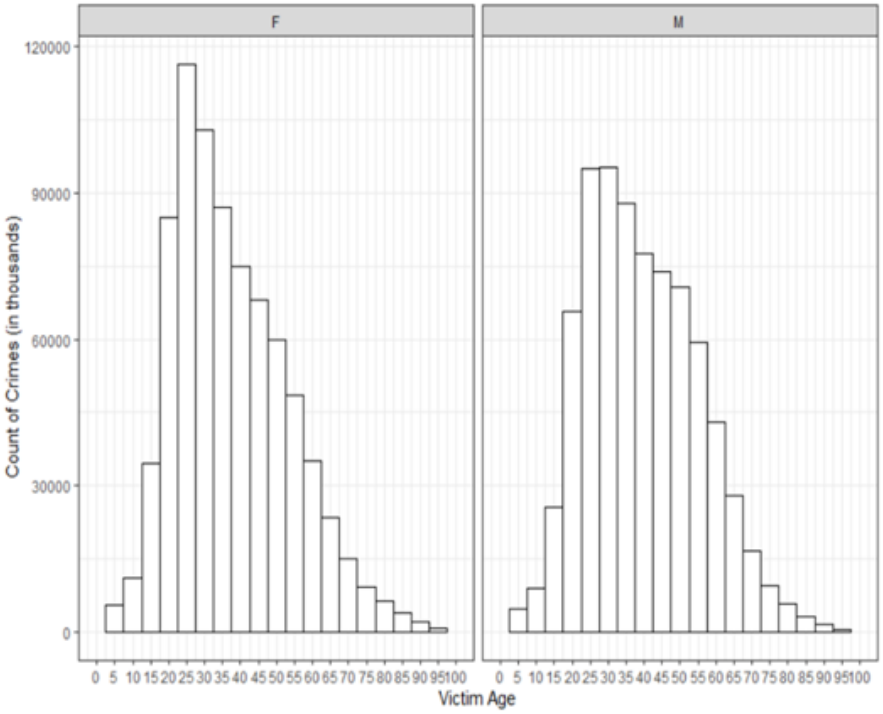

Figure 9. Distribution of Crime Incidents According to Gender of the Victim and Their Age from 2010 to 2017

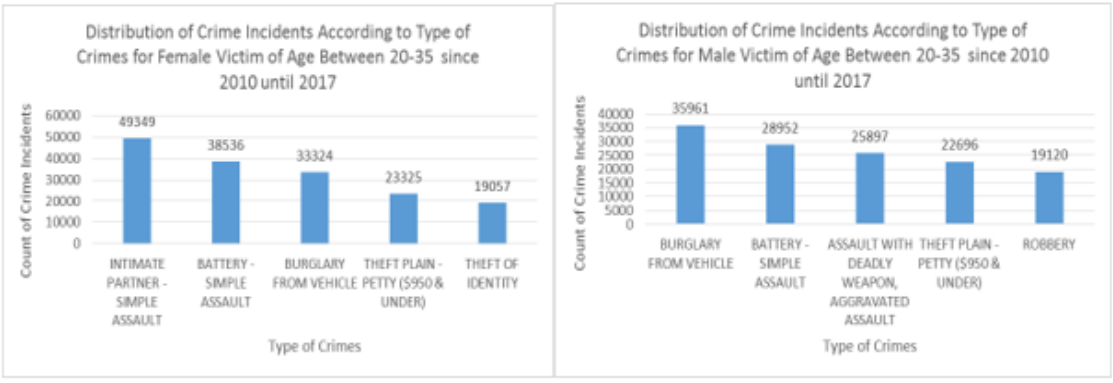

Figure 10. Top 5 of crime incidents according to the type of crimes for female victims and male victims of age between 20 and 35 from 2010 until 2017

Figure 11 shows the top 5 crime incidents according to the type of crimes for female victims and male victims from 2010 until 2017. It is interesting to note that both victims suffer the most from battery-simple as-sault. It is also interesting to know that, within the top 5 types of crime, female victims suffer intimate partner-simple assault, while male victims do not. However, in these top 5 types of crime, male victims suffer assault with a deadly weapon, while female victims do not.

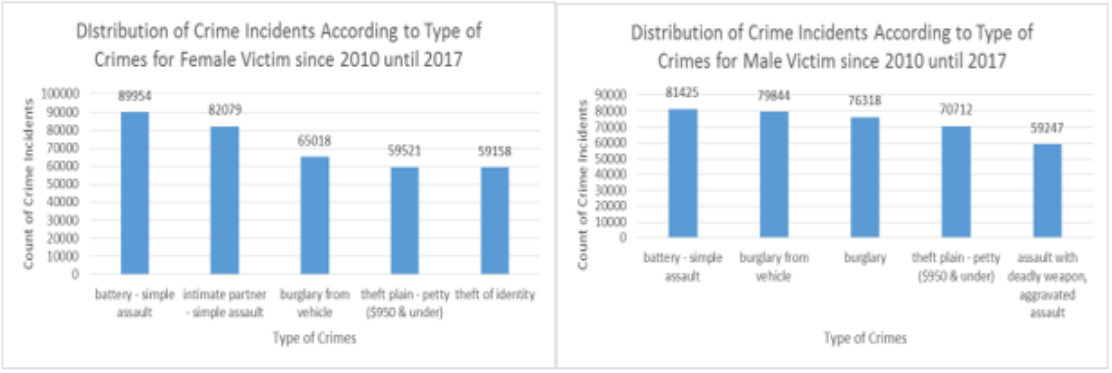

Figure 11. Top 5 of Crime Incidents According to Type of Crimes and Gender since 2010 until 2017

Figure 12 shows the distribution of crime incidents according to the premises of happening from 2010 until 2017. From the figure.12, we find that most crimes happen on the street. The second place and third place of happening are, respectively, a single-family dwelling and multi-unit dwelling such as apartments, duplex, etc.

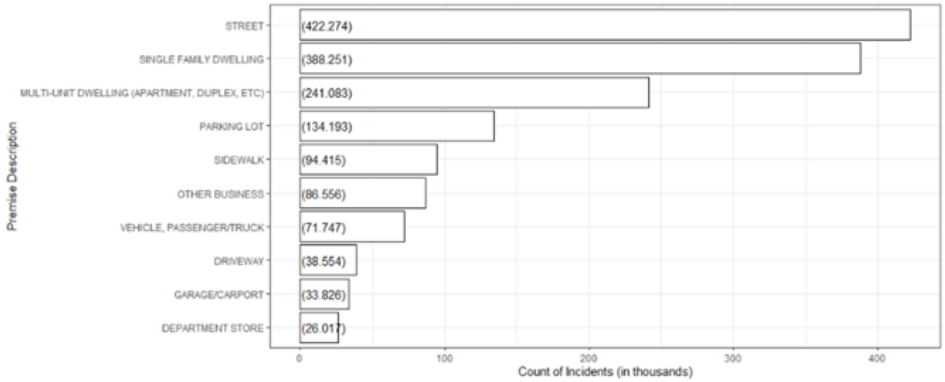

Figure 12. Distribution of Crime Incidents According to Premises since 2010 until 2017

Figure 13 shows the type of crimes that happens on those premises. Table 1 shows the count of a type of crime that happens on those premises.

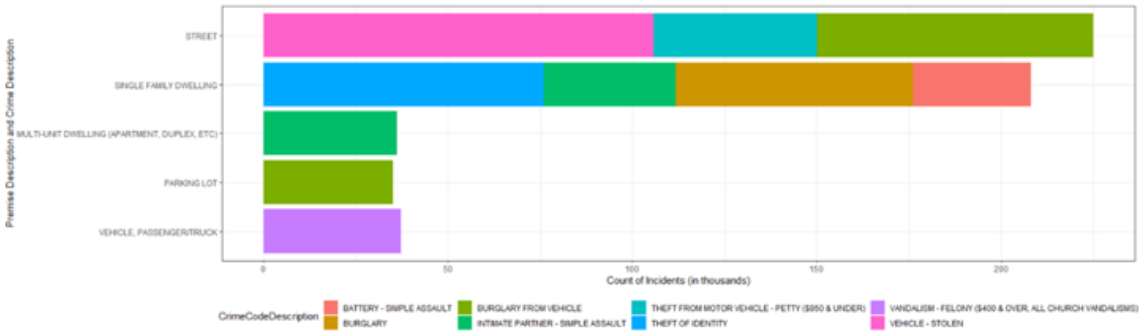

Figure 13. Type of crimes happens on the premises

Table 1. Count of type of crimes happens on the premises

Premises	Crime Type	Count
Street	Vehicle - Stolen	105,757
Street	Burglary From Vehicle	74,978
Street	Theft From Motor Vehicle - Petty ($950 & Under)	44,282
Single Family Dwelling	Theft Of Identity	75,914
Single Family Dwelling	Burglary	64,177
Single Family Dwelling	Intimate Partner - Simple Assault	35,915
Single Family Dwelling	Battery - Simple Assault	32,059
Multi-Unit Dwelling (Apartment, Duplex, etc.)	Intimate Partner - Simple Assault	36,125
Parking Lot	Burglary From Vehicle	35,015
Vehicle, Passenger/Truck	Vandalism - Felony ($400 & Over, All Church Vandalisms)	37,214

Figure 14 shows the map of the top ten locations of crimes from 2010 until 2017. The Southwest area with reporting district number 0363 is the most unsafe area with 9,609 incidents.

Figure 14. Map of Top Ten Location of Crimes (circled) since 2010 until 2017

Conclusion

This paper presents the result of Exploratory Data Analysis (EDA) using univariate analysis, bivariate analysis, and multivariate analysis. R programming language applied to 1,895,619 rows and 28 columns of Los Angeles Crime Report Data from 2010 until 2017. As the result shows, by using those methods, we can gain insights, understandings, and new patterns from the existing data. By performing EDA we can analyze the data using tables and various types of charts such as line charts, bar charts, stacked charts, and geo charts.

The variations of crime variables presented in this paper are only a few of the many variations that can be made. Other variations can be performed to get more insights, understandings, and new patterns from the existing data. The methods can be performed on other types of data as well.

References

[1] R. K. Pearson, *Exploratory Data Analysis Using R*, 1st Editio. CRC Press/Taylor & Francis Group, 2018.

[2] J. Winter, M. Breinig, F. Heigwer, D. Brügemann, S. Leible, O. Pelz, T. Zhan, and M. Boutros."CaRpools: An R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens," Bioinformatics, vol. 32, pp. 632–634, 2016.

[3] R. L. Nuzzo, "Histograms: A Useful Data Analysis Visualization," *PM R*, 2019, doi: 10.1002/pmrj.12145.

[4] P. Vermeesch, "Exploratory analysis of provenance data using R and the provenance package," *Minerals*, vol. 9, no. 3, 2019, doi: 10.3390/min9030005.

[5] A. Perer and B. Shneiderman, "Integrating statistics and visualization," p. 265, 2008, doi: 10.1145/1357054.1357101.

[6] N. Verbeeck, R. M. Caprioli, and R. Van de Plas, "Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry," *Mass Spectrom. Rev.*, p. mas.21602, Oct. 2019, doi: 10.1002/mas.21602.

[7] J. Camacho, R. A. Rodríguez-Gómez, and E. Saccenti, "Group-Wise Principal Component Analysis for Exploratory Data Analysis," *J. Comput. Graph. Stat.*, vol. 26, no. 3, pp. 501–512, 2017, doi: 10.1080/10618600.2016.1265527.

[8] A. Watson, S. Bateman, and S. Ray, "PySnippet: Accelerating exploratory data analysis in Jupyter Notebook through facilitated access to example code," *CEUR Workshop Proc.*, vol. 2322, pp. 6–9, 2019.

[9] H. Zhao, Q. Meng, and Y. Wang, "Exploratory data analysis for the cancellation of slot booking in intercontinental container liner shipping: A case study of Asia to US West Coast Service," *Transp. Res. Part C Emerg. Technol.*, vol. 106, pp. 243–263, Sep. 2019, doi: 10.1016/j.trc.2019.07.009.

[10] M. M. Richard's, S. Vernucci, F. Stelzer, I. Introzzi, and J. Guàrdia-Olmos, "Exploratory data analysis of executive functions in children: a new assessment battery," *Curr. Psychol.*, pp. 1–8, May 2018, doi: 10.1007/s12144-018-9860-4.

[11] M. Feng, D. Brenner, and A. Coulson, "Using exploratory data analysis to support implementation and improvement of education technology product," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11626 LNAI, pp. 79–83, doi: 10.1007/978-3-030-23207-8_15.

[12] N. V. Bondarev, "Classification and Prediction of Sodium and Potassium Coronates Stability in Aqueous-Organic Media by Exploratory Data Analysis Methods," *Russ. J. Gen. Chem.*, vol. 89, no. 2, pp. 281–291, Feb. 2019, doi: 10.1134/S1070363219020191.

[13] F. Cerveira *et al.,* "Exploratory data analysis of fault injection campaigns," in *Proceedings - 2018 IEEE 18th International Conference on Software Quality, Reliability, and Security, QRS 2018*, Aug. 2018, pp. 191–202, doi: 10.1109/QRS.2018.00033.

[14] K. Hara, A. Adams, K. Milland, S. Savage, B. V. Hanrahan, J. P. Bigham, and C. Callison-Burch. "Worker demographics and earnings on amazon mechanical turk: An exploratory analysis," *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 1–6, 2019.

[15] H. B. Sankaranarayanan, G. Agarwal, and V. Rathod, "An exploratory data analysis of airport wait times using big data visualisation techniques," *2016 Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2016*, pp. 324–329, 2016, doi: 10.1109/CSITSS.2016.7779379.

[16] S. Yamada, Y. Yamamoto, K. Umezawa, S. Asai, H. Miyachi, M. Hashimoto, and S. Inokuchi. "Exploratory data analysis for medical data using interactive data visualization," *2016 14th International Conference on ICT and Knowledge Engineering (ICT&KE),* 2016, pp. 7-11, doi: 10.1109/ICTKE.2016.7804091.

[17] I. Setiawan, "Analisis Eksplorasi Dan Visualisasi Profil Superhost Airbnb Kota Madrid Dan Amsterdam," *JTT (Jurnal Teknol. Ter.*, vol. 6, no. 2, p. 156, 2020, doi: 10.31884/jtt.v6i2.274.

[18] I. Setiawan, "Pengembangan Prototipe Aplikasi Manajemen Risiko Berbasis ISO 31000," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 10, no. 1, pp. 26–33, 2020, doi: 10.31940/matrix.v10i1.1817.

[19] S. S. P. & Keamanan, "Statistik Kriminal 2018," 2018.

[20] S. Pradhan, Divyansh, M. Pandey, and S. S. Rautaray, "Analysis of Suicides in India—A Study Using the Techniques of Big Data," in *Lecture Notes in Networks and Systems*, vol. 41, Springer, 2019, pp. 327–338.

[21] C. Catlett, E. Cesario, D. Talia, and A. Vinci, "A data-driven approach for spatio-Temporal crime predictions in smart cities," in *Proceedings - 2018 IEEE International Conference on Smart Computing, SMARTCOMP 2018*, Jul. 2018, pp. 17–24, doi: 10.1109/SMARTCOMP.2018.00069.

[22] M. Feng, J. Zheng, Y. Han, J. Ren, and Q. Liu, "Big Data Analytics and Mining for Crime Data Analysis, Visualization and Prediction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 10989 LNAI, pp. 605–614, doi: 10.1007/978-3-030-00563-4_59.

[23] J. Zhao and C. L. Rogalin, "Heinous Crime or Unfortunate Incident: Does Gender Matter?," *Soc. Psychol. Q.*, vol. 80, no. 4, pp. 330–341, 2017, doi: 10.1177/0190272517728923.

[24] I. Setiawan, "Time series air quality forecasting with R Language and R Studio," *J. Phys. Conf. Ser.*, vol. 1450, p. 12064, Feb. 2020, doi: 10.1088/1742-6596/1450/1/012064.

[25] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, 1st Editio., vol. 72. Springer International Publishing, 2015.