Data Visualization Assignment



|  |  |
| --- | --- |
| **Name:** | Shivani Bhat |
| **Student Number:** | 20211408 |
| **Program:** | MCM |
| **Module Code:** | CA682 |
| **Assignment Title:** | Data Visualization |
| **Submission Date:** | 17 Dec 2020 |
| **Module Coordinator:** | Dr. Suzanne Little |

Declaration on Plagiarism:

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at <http://www.dcu.ie/info/regulations/plagiarism.shtml>, <https://www4.dcu.ie/students/az/plagiarism> and/or recommended in the assignment guidelines

Name: Shivani Bhat   Date: 17 Dec, 2020

**Topic: Analysis of Crime Data (Los Angeles) from 2010-2019**

*“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.” — Arthur Conan Doyle,*a Scandal in Bohemia

**Abstract:**

Crime, as stated by Wikipedia, is an unlawful act that is punishable by the State. Crime Statistics are important to understand if the legislative procedures are working or if they require any changes and amendments. Crime analysis plays a major role in supporting the operations of the police department. It helps the department strategize and improve measures in the area where crime is prone to happen. It assists them in problem-solving and evaluates the performance of their department in various zones. If the rate of crime is falling, it would indicate that what the system is doing seems to work. It helps them plan further actions. Here, we figure out patterns of crime, the common areas, and the type of crime that takes place. This helps detectives identify and apprehend suspects. Additionally, crime analysis benefits the police to focus their resources on high-crime places to reduce crime in their community. It reduces the frequency of the crime and its severity. In this course work, I would be working on handling crime data of Los Angeles from 2010-2019 that would include cleaning, processing, exploring, and visualizing the data to find answers to a few research questions. The primary research question here would be the trend in crime occurrence vs. crime reported. The work would include finding if there is any delay in crime reported. Also, which area has the highest crime occurrence rate? What is the most common time a crime is bound to occur and at what time over the years?

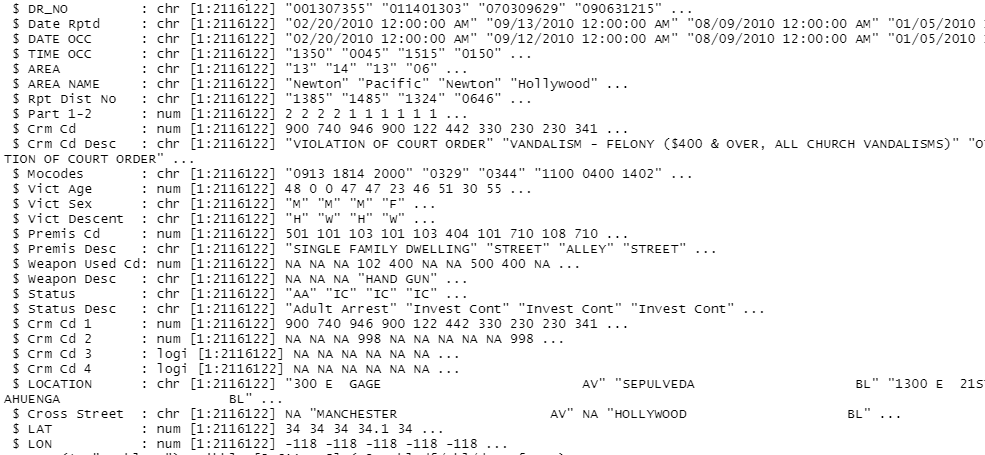
**Data Description:**

The dataset used here is Los Angeles Crime Data (2010 - 2019). Like any other major cities there are different regions that are crime-ridden and according to a survey L.A is ranked as 123rd most crime ridden city.

Link to the dataset: <https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z>

The dataset belongs to the LAPD (Los Angeles Police Department), California. It reflects incidents that have occurred in LA over nine years. It is an open data set with over 2.12 million rows (instances) and 28 columns (attributes). The data structure is as follows:

Note: str(DataFrame) command in R provides the structure of the DataFrame.



Characteristics of **Big Data** included here:

The dataset is refreshed weekly. The last update was performed on 9 December, 2020, hence included the parameter of “Velocity”. Additionally, we can also partially include the characteristic “Volume” as the entire dataset cannot be processed in an excel sheet.

**Data Cleaning, Data Processing & Data Exploration:**

The dataset here is open raw data. The data is transcribed from original paper reports, hence they contain inconsistencies. To retrieve quality data, we initially need to perform Data Cleaning.

I have performed data cleaning and dimensionality reduction using R studio.

Steps Performed for Data Cleaning:

Primary quality checks should be initially performed. We must check the date format (done by str() function), the missing values, null data (done by is.na()), trailing spaces in the column header. The data cleaning procedure followed here are as follows:

* The date format was inconsistent, data was in the format of dd-mm-yyyy 12:00:00 am, with a default time added to all instances. The default time is not required and should be removed. The date was also in “char” format. This is transcribed to date (to.Date() using lubridate package). This is done for both Date rptd and Date occ columns.
* Time occurrence of crime is in an incorrect format ex: 230, 45, and 1530. This should be translated to 02:30, 00:45 and 15:30 respectively. strptime() function helps to perform the task of formatting.
* Sex of criminals can be either M/F. But as seen in the data using the function table(), the sex is of various categories (M F H N X). X here is stated as Unknown in data description. All sex types other than M and F are replaced with NA.
* Data for age contains various 0 and NA values. NA would indicate “no data”. But age can never be 0. Hence, this must be handled. Values of age for females which are mentioned as 0 are replaced with the mean of age for females. Similarly, mean of age for all men is integrated instead of 0. The age values which do not have any corresponding sex category are replaced with NA.
* DR\_No and Area code which as seen in the structure are char values, should be transformed to Int value by as.numeric() function.
* Apply() function with is.na() provides a number of NA values. All these must be replaced with ‘-‘since mathematical operations cannot work on attributes with NA values present.

Steps involved in Data Exploration:

* One of the research questions that are examined is the relation between the date of crime reported and the date of crime occurrence. This relation would indicate if there is any deviation and whether the cases are reported immediately after its occurrence. For this, delay of time should be calculated. Therefore, an additional column “delay” is added using *mutate*(). This variable will be used for analysis.

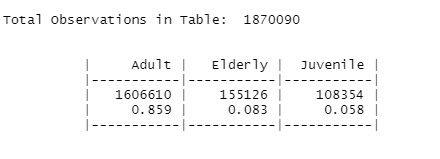


Here, we can **analyze** that cases have taken an average of 20 days to report from the occurrence date. The maximum delay to report is approximately 11 years. There are cases which have been reported on the same date as well (minimum delay is zero).

* Exploration on Age variable using crossTable() function evaluated that there were age instances which are serving as outliers in the dataset ex: negative values, age >100.

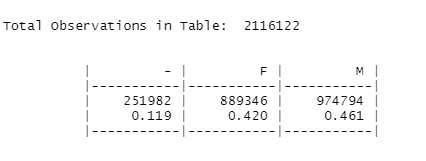
The age values which were less than 10 and greater than 80 (assumption: only age group 10-80 commit a crime) were replaced with NA and the process to replace them with them mean is similar to that in data cleaning. Further, age instances were grouped into “Juvenile (10-17), Adult (18-60) and Elderly (60+).

The results were as shown:



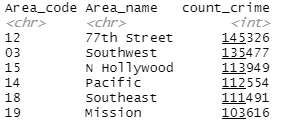
This depicts the count and proportion from the population. On **analysis**, we can see that the age group called as Adult has been involved in crimes at a great extent.

* CrossTable() performed on Sex variable provided a brief **analysis** that males are involved in more cases than females.

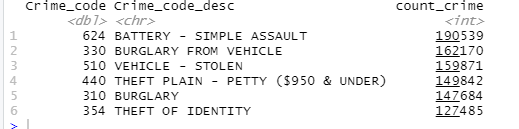


* Area-wise count of cases reported were grouped (group\_by()) for better readability and clarity.

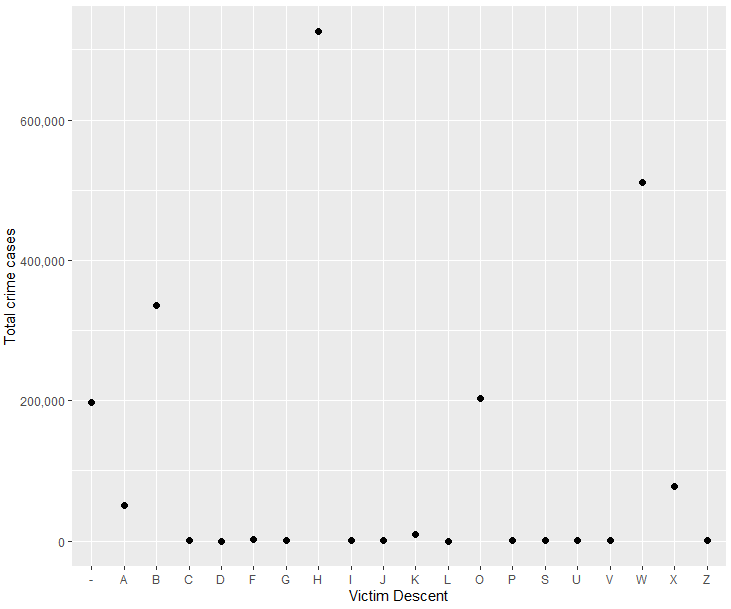
The **analysis** done on the result states that 77th Street (Area code= 12) has the highest number of cases occurred.

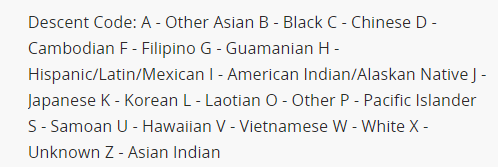


* The type of crime is another attribute present which can be explored. The top few rows of the result ( head()) of grouping together all the crime description is :



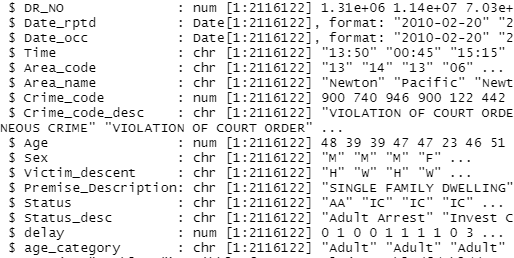
We can **analyze** that a lot of crimes occurred have been that of simple assault, followed by burglary and vehicles being stolen.

* Victim Descent accommodates a diverse range of categories analysis on which gave the following plot.



**Analysis** shows here that Victim descent code =H (Latin/ Hispanic/ Mexican) are victims of a high number of crimes.

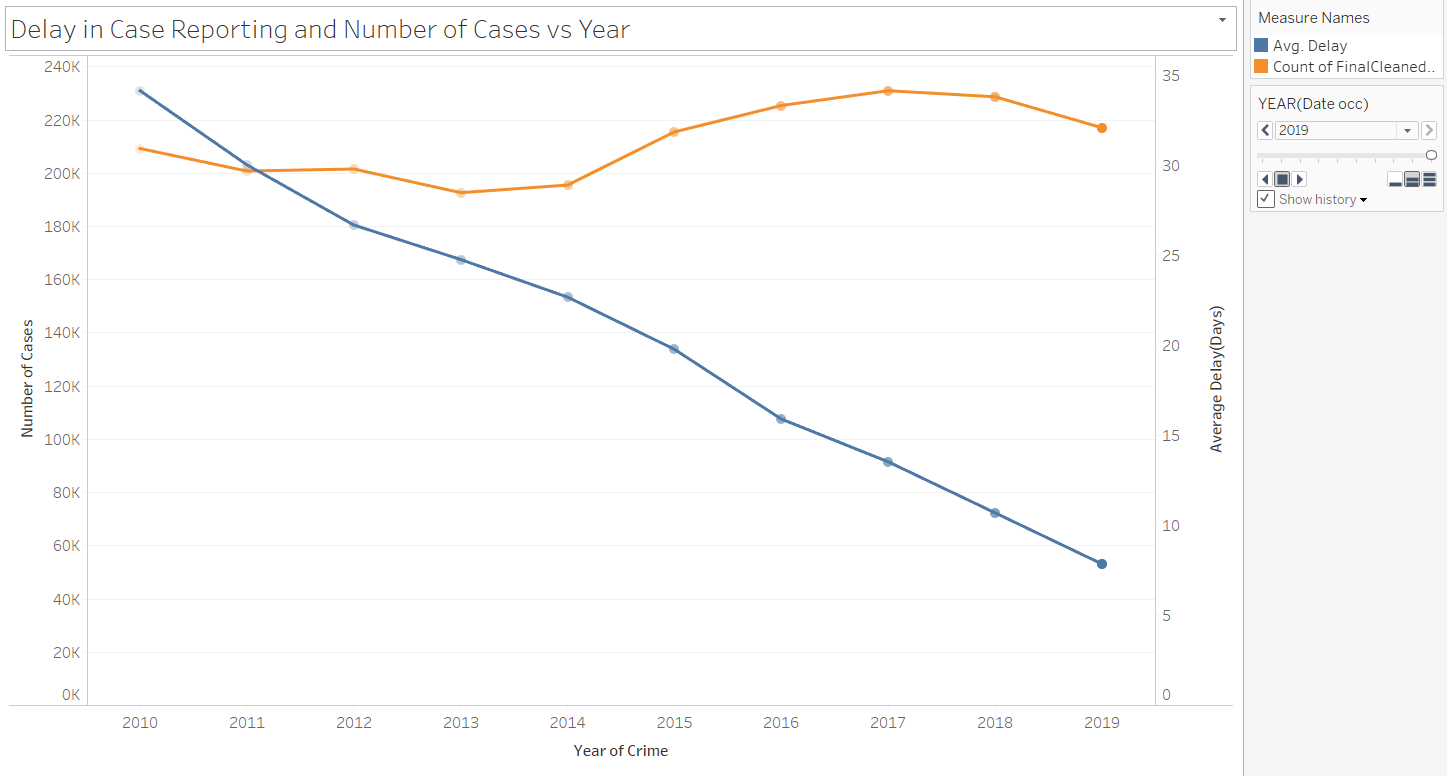
The final data which would be used for visualization now contains the following attributes:



**Data Visualization:**

Visualization 1: Crime occurrence and Crime report Delay over the years

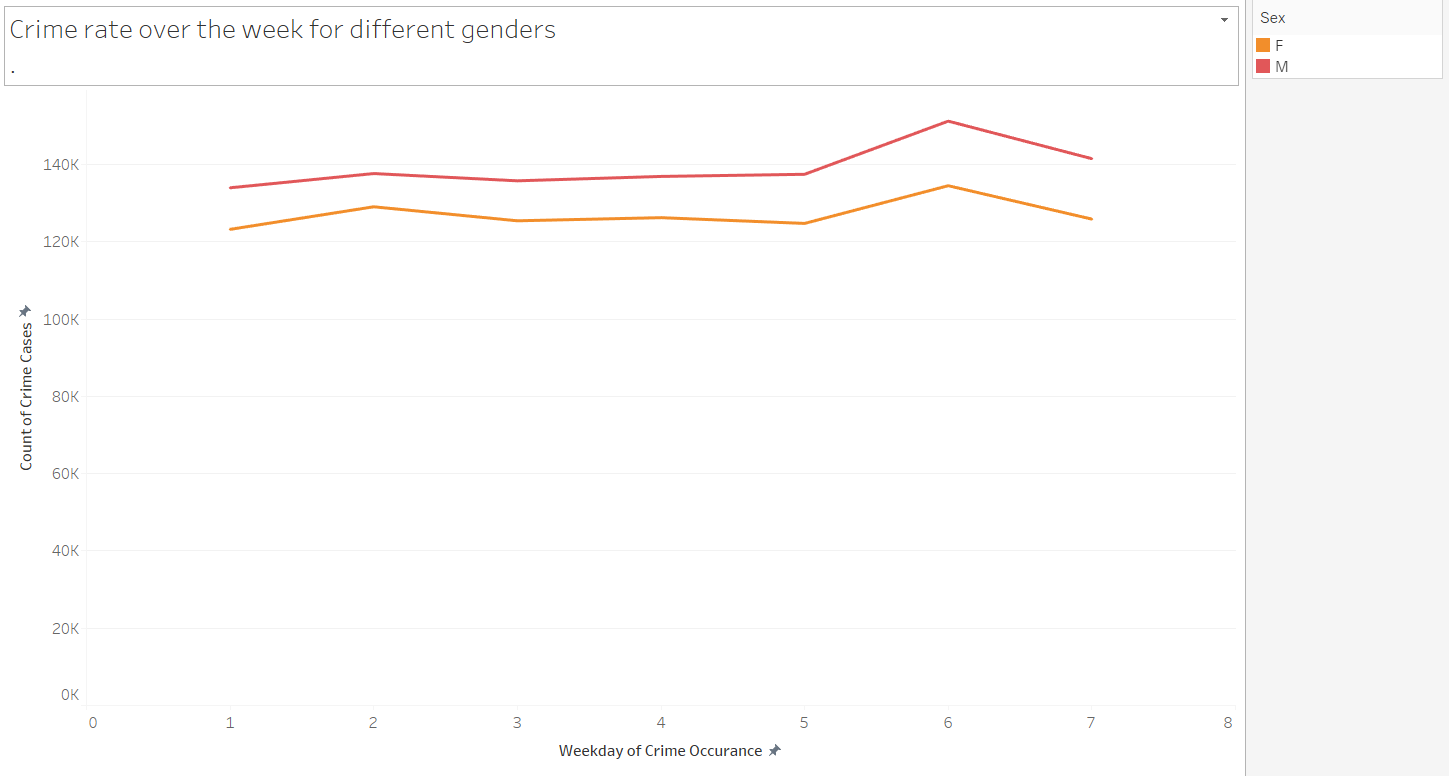
After taking into consideration the number of cases that have occurred over 9 years, we have observed that the crime count over the years has shown a slightly increasing model. When we now compare it with the delay of crime reporting, we observe that a maximum delay of 35 days on average is illustrated in 2010. Delay of case reporting has shown a decline over the years. The LAPD department has shown progress in outlining its goals and tracking their progress over the years.



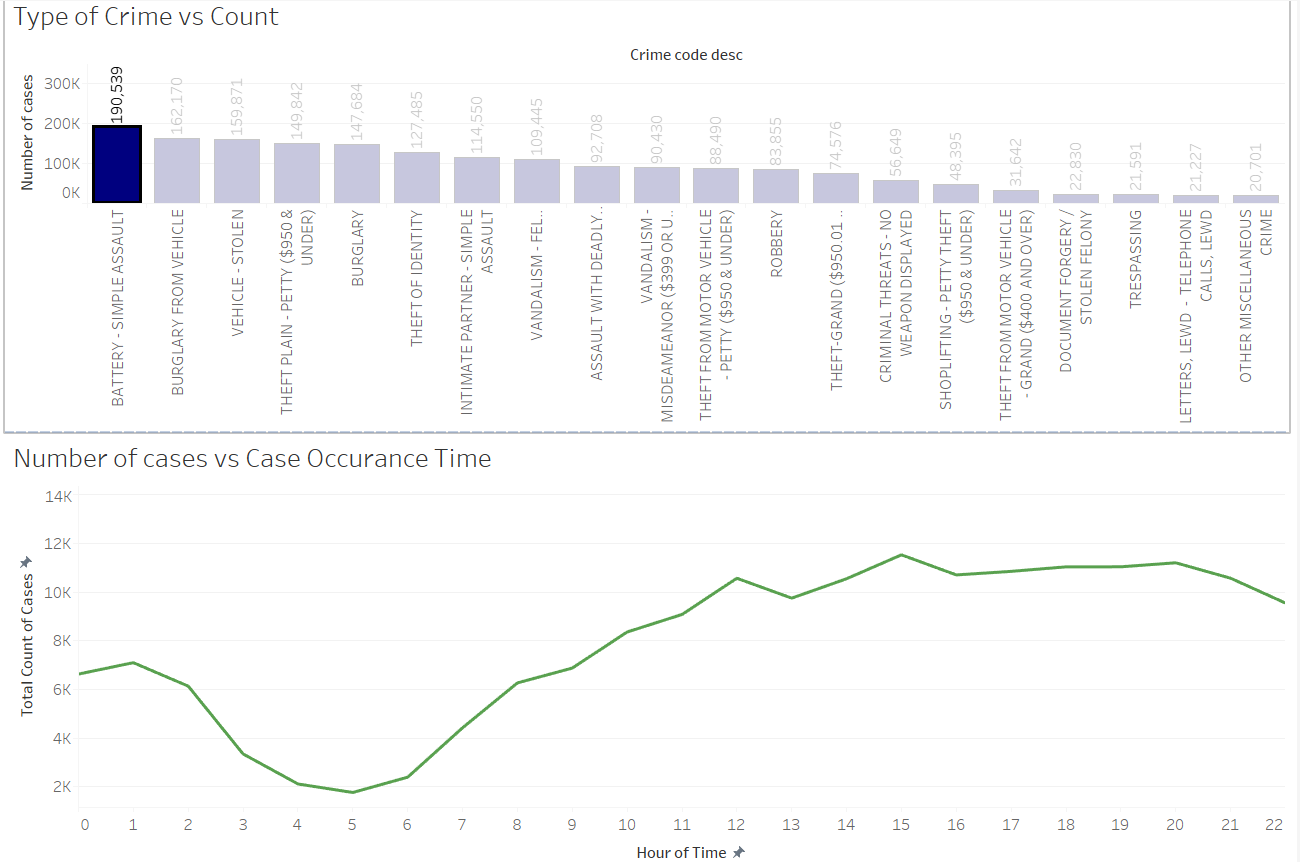
The characteristic feature of the attributes varies over time. Hence, a time-series graph would suit best. Two different colors are used to show each trend line. The graph is animated to show the trend of both lines over time. Additionally, we can also select the year from which we would like to see the trend.

Visualization 2: When Crime Happens?

The frequency of the crime adds another dimension to the analysis and helps monitor the team to understand the proportion of crime through the weekdays. The frequency of beginning of crime occurs most common on weekdays especially Friday (here the days are mapped to the day number 1 being Sunday). A closer look at the attributes, most of them count to noon, which can also indicate the noon, can be a default time assigned for unknown start times. Among all the weekdays, Fridays have more crime expected. We can also see that Saturday follows. Other observations happen to peak around Monday, which could be around holidays like Christmas and Thanksgiving. Hence, this would suggest that holidays may influence the crime rate. The plot is divided based on gender to analyze how the crime rate over the week gets influenced by different gender (here NA values have been ignored). The trend is almost similar for both genders. The only difference would be the number of cases which are higher for males. They are more likely to be victims of crimes.



Similar to the previous plot, the trend moves over time and hence a time series plot. The trend lines are drawn with different colors to recognize gender. An interactive graph between type of crime and time of crime is also done. This can be seen from my dashboard the link to which is given below.



We can see that “Battery-Simple assault” crime type occurs at a highest rate at 3 pm and the least at 5 am. This could suggest that daylight is a factor affecting this type of crime. Another example would be that “Burglary from Vehicle” takes place at the highest rate after 6 pm suggesting that after dawn, as vehicle users would be at home, vehicle burglary would be the easiest.

 Tools and Libraries used:

Major cleaning is performed on R studio including libraries like *lubridate*(time and date formatting), *dplyr* (for filtering content), ggplot(for plots) and some baseR functions. The major challenge was to refactor the time attribute as all libraries support only date-time or date format. Visualization is done on Tableau which supports a varied range of graphs.

Link to my dashboard: [Shivani Bhat - Profile | Tableau Public](https://public.tableau.com/profile/shivani.bhat1640#!/)

Link to the video: <https://drive.google.com/file/d/1qLJw1ySQ1-oGJrHPptN6V5-szk2WtntT/view?usp=sharing>

**Conclusion:**

Notwithstanding the different bad behaviors and their examples, the day of the week can be used to envision the crime rate while sex and character can be used as precursors to choose setbacks. The crime rate is high all through the closures of the week while Male and Hispanic fall will undoubtedly be victims. Economics are furthermore connected with bad behavior plans as explicit spots are assessed higher in infringement. Ordinarily, the crime rate will be high on Fridays. Of course, it will, in general, be depended upon to be low on days around Christmas or Thanksgiving; Sunday is needed to have the least crime rate.

Further examination could join in separating bad behavior occasions for each sort in different areas. This analysis is just a starting point to what we can further explore. The scope of crime extends far and wide.

**References:**

[1] Statistical Data - Los Angeles Police Department, https://www.lapdonline.org/statistical\_data

[2] How the LAPD Uses Data to Predict Crime https://www.wired.com/story/los-angeles-police-department-predictive-policing/

[3] C. S. Marzan, M. J. C. Baculo, R. de Dios Bulos, and C. Ruiz, “Time Series Analysis and Crime Pattern Forecasting of City Crime Data,” in *Proceedings of the International Conference on Algorithms, Computing and Systems*, New York, NY, USA, Aug. 2017, pp. 113–118, doi: 10.1145/3127942.3127959.