Data Visualization

Assignment 1
Report

Submitted By
Muhammad Talha 20I-2212

Course Instructor
Dr. Muhammad Faisal Cheema

Contents

List of Figures

List of Tables

City of Los Angeles

Los Angeles is a city in California, United States. It is known as the home of Hollywood and is famous for its iconic Hollywood sign and walk of fame. Los Angeles is often written and spoken as LA. LA is the second most populated city in the United States and the most populated city in California with 3.9 million residents. The economy of the state of California is the largest among the United States of America and has a GDP of $3.3 Trillion. California independently would be the 5th largest economy in the world. [1]

Los Angeles being the most populated city of California significantly contributes to the economy of California. Just as the city invites opportunities and growth, the success of the city also attracts criminals who would like to take advantage of people.

To promote the democratic values of openness and transparency, the city provides a comprehensive list of datasets that are available for the public to download and perform analysis on. These datasets are available on LA open data website. [2]

Dataset

For our analysis, we downloaded and performed exploratory data analysis(EDA) on the Crime Data. This dataset is provided to the public by Los Angeles Police Department. The ownership of the dataset lies with the LAPD. The dataset is updated on weekly basis.

The dataset is divided into two parts. Crime data from 2010 to 2019 and crime data from 2020 to the present. The dataset has 26 features. The description of each feature is given below. [3]

Columns Description

1. DR_NO
   This column ‘division of records number’ serves as the row identifier. This column consists of 2 digits made up of the year, area code, and 5 digits.
2. Date Rptd
   This column keeps the information of the date when the crime was reported.
3. DATE OCC
   This column keeps information on the date when the crime occurred.
4. TIME OCC
   This column keeps information of the time in military hours when the crime occurred.
5. AREA
   The city is divided among 21 community police stations as geographical areas within the police department. These areas are numbered from 1 to 21.
6. AREA NAME
   This column includes the area name associated with the area.
7. Rpt Dist No
   An area is sub-divided into districts. It is represented by a four-digit code.
8. Crm Cd
   This column keeps information on the type of crime committed.
9. Crm Cd Desc

This column defines the criminal code.

10. Mocodes
    This column keeps information on the modus operandi of the criminal.
11. Vict Age
    This column keeps information on the age of the victim.
12. Vict Sex
    This column keeps the information on the gender of the victim.
13. Vict Descent
    This column keeps information on the descent of the victim.
14. Premis Cd
    This column keeps information on the type of structure where the crime was committed.
15. Premise Desc
    This column keeps information on the description of the type of structure where the crime was committed.
16. Weapon Used Cd
    This column keeps information on the type of weapon used in the crime in a code.
17. Status
    This column keeps the information on the status of the case.
18. Status Desc
    This column keeps information on the status description.
19. Crm Cd 1
    This column keeps information on the type of crime committed. This is the same as Crm code
20. Crm Cd 2
    This column keeps information on the type of additional crime committed if any.
21. Crm Cd 3
    This column keeps information on the type of additional crime committed if any.
22. Crm Cd 4
    This column keeps information on the type of additional crime committed if any.
23. LOCATION
    This column keeps the information on the street address of the crime incident.
24. Cross Street
    This column keeps information on the cross street of crime incident
25. LAT
    This column keeps information on the latitude of where the crime was committed.
26. LON
    This column keeps information on the longitude of where the crime was committed.

Data preprocessing

For data handling and manipulation, we used a well-known Industry-leading python library ‘Pandas’. Pandas is feature-rich and provides excellent efficiency and fast processing while handling large datasets.

Dataset Merging

As the dataset was divided into two sub-datasets. Datasets were merged and saved into a ‘.parquet’ file format. This file encoding shrank the dataset size on disk from 578MB to only 98MB. Later, this file was used for any further processing.

Dropping unused columns

We dropped columns ‘DR_NO’ and ‘Part 1-2’. As ‘DR_NO’ was just an ID and no description of ‘Part 1-2’ was provided.

Column renaming

Dataset columns were renamed to human-readable format. This helped in the analysis of the data.

Columns Datatypes

Since Pandas import the columns in the ‘.csv’ files as ‘objects’. To provide accurate data type to each column, we converted the data type of each column from generic ‘object’ to its specific type.

Table 1 shows new column names and datatypes

Column Name	**Panda Data Type**
Crime Date Reported	datetime64[ns]
Crime Date Occurred	datetime64[ns]
Crime Time Occurred	datetime64[ns]
Crime Area Code	category
Crime Area Name	category
Crime Reported Dist No	category
Crime Code	int64
Crime Code Desc	category
Mocodes	category
Victim Age	int64
Victim Gender	category
Victim Descent	category
Premis Code	category
Premis Description	category
Weapon Used Code	float64
Weapon Description	category
Status	category
Status Desc	category
Crime Code 1	float64
Crime Code 2	float64
Crime Code 3	float64
Crime Code 4	float64
Crime Location	category
Cross Street	object
LAT	object
LON	object

Dataset Analysis

Null values analysis

Null values analysis was done to understand the completeness of the dataset. We can observe that the dataset is largely complete and the missing values in certain columns attribute to the column itself.

Table 2 shows missing values in each column

Column Name	**Percentage empty**
Crime Date Reported	0%
Crime Date Occurred	0%
Crime Time Occurred	0%
Crime Area Code	0%
Crime Area Name	0%
Crime Reported Dist No	0%
Crime Code	0%
Crime Code Desc	0%
Mocodes	11%
Victim Age	0%
Victim Gender	9.7%
Victim Descent	9.7%
Premis Code	0.0023%
Premis Description	0.01%
Weapon Used Code	66%
Weapon Description	66%
Status	0%
Status Desc	0%
Crime Code 1	0%
Crime Code 2	93%
Crime Code 3	99.8%
Crime Code 4	99.9%
Crime Location	0%
Cross Street	0%
LAT	0%
LON	0%

Correlation matrix of Dataset.

A correlation matrix of the dataset was computed.

Table 3 correlation matrix of the dataset

	Crime Code	Victim Age	Weapon Used Code	Crime Code 1	Crime Code 2	Crime Code 3	Crime Code 4
Crime Code	1	-0.03408	0.413897	0.999482	0.039519	0.113764	0.049008
Victim Age	-0.03408	1	0.080628	-0.03394	-0.02685	-0.02871	-0.06771
Weapon Used Code	0.413897	0.080628	1	0.414439	-0.13938	-0.02753	-0.07971
Crime Code 1	0.999482	-0.03394	0.414439	1	0.053658	0.145966	0.033754
Crime Code 2	0.039519	-0.02685	-0.13938	0.053658	1	0.295848	0.237815
Crime Code 3	0.113764	-0.02871	-0.02753	0.145966	0.295848	1	0.430893
Crime Code 4	0.049008	-0.06771	-0.07971	0.033754	0.237815	0.430893	1

Feature Distribution

Feature distribution shows the range of the feature, with its values on the x-axis and the frequency on the y-axis.

Crime Date Reported and Crime date occurred

From these two distributions, we can observe that both features have similar curves distributions.

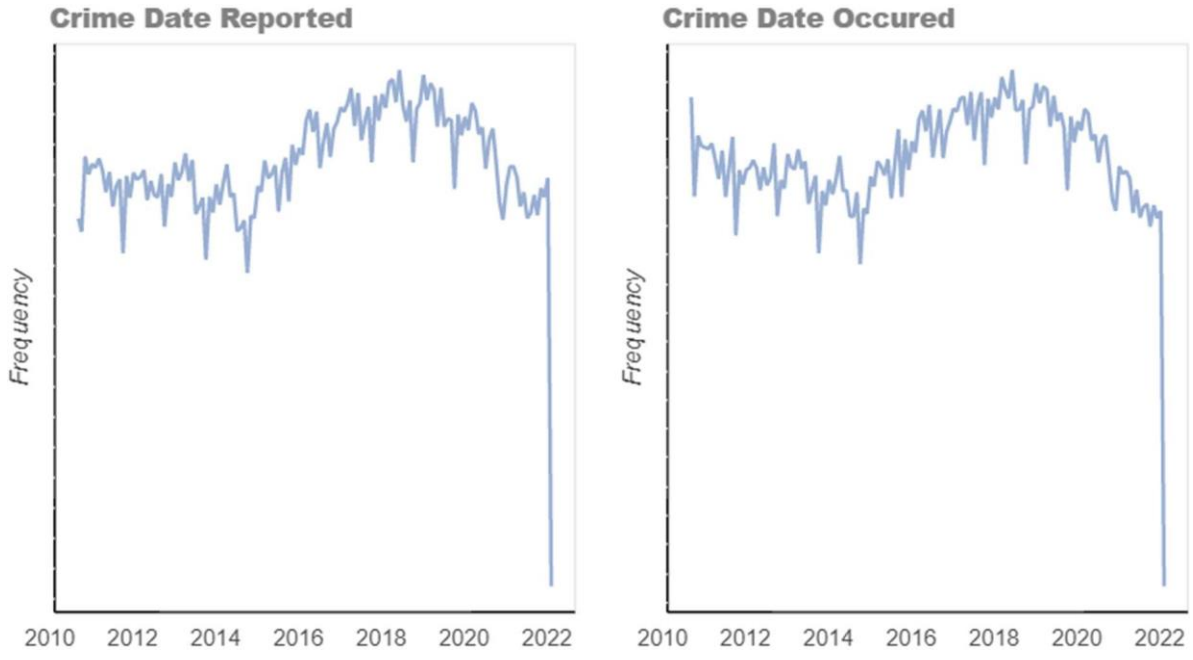

Figure 1 shows the distribution of Crime date reported left and crime date occurred right.

Crime Time occurred

We can see from this distribution that most crimes occurred at midnight.

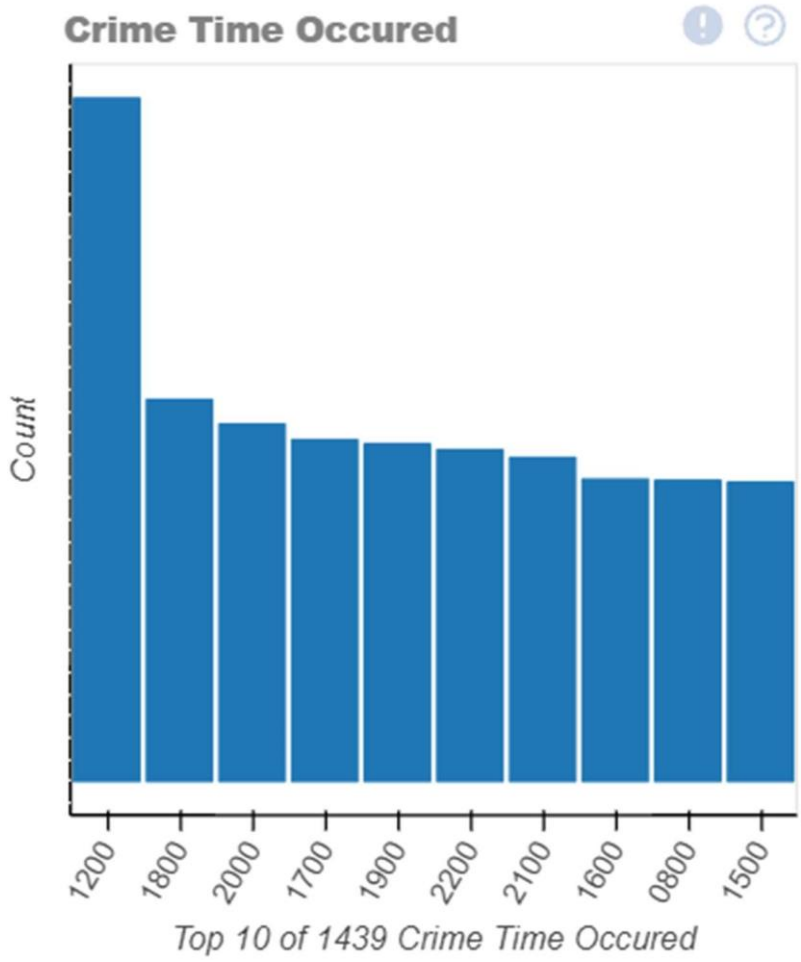

Figure 2 shows the distribution of crime time in the city

Crime Area code and description

From both these distributions, we can see 77th street and southwest are the most affected areas.

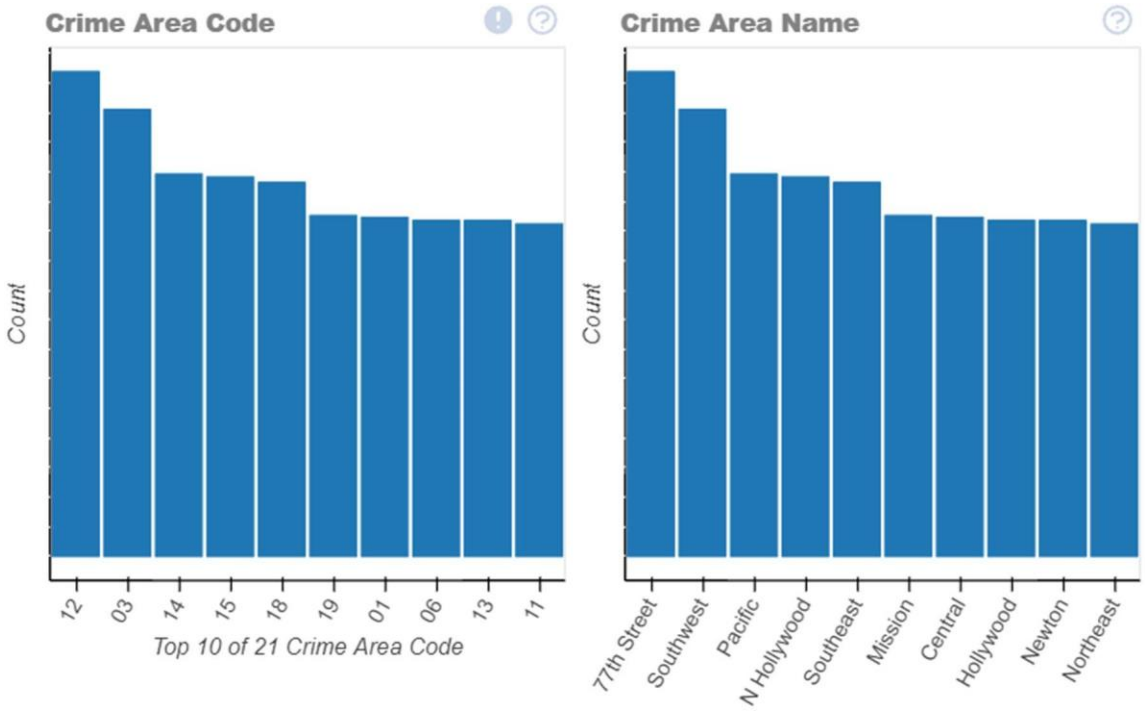

Figure 3 shows the distribution of crime areas codes and names.

Crime Code and description

Crime code distribution shows the severity of crimes that are occurred. Codes in the range of 100 to 200 are the most severe crime that includes gun violence.

Crime code description shows battery-simple assault as the most occurred crime with grand theft auto and burglary as second and third.

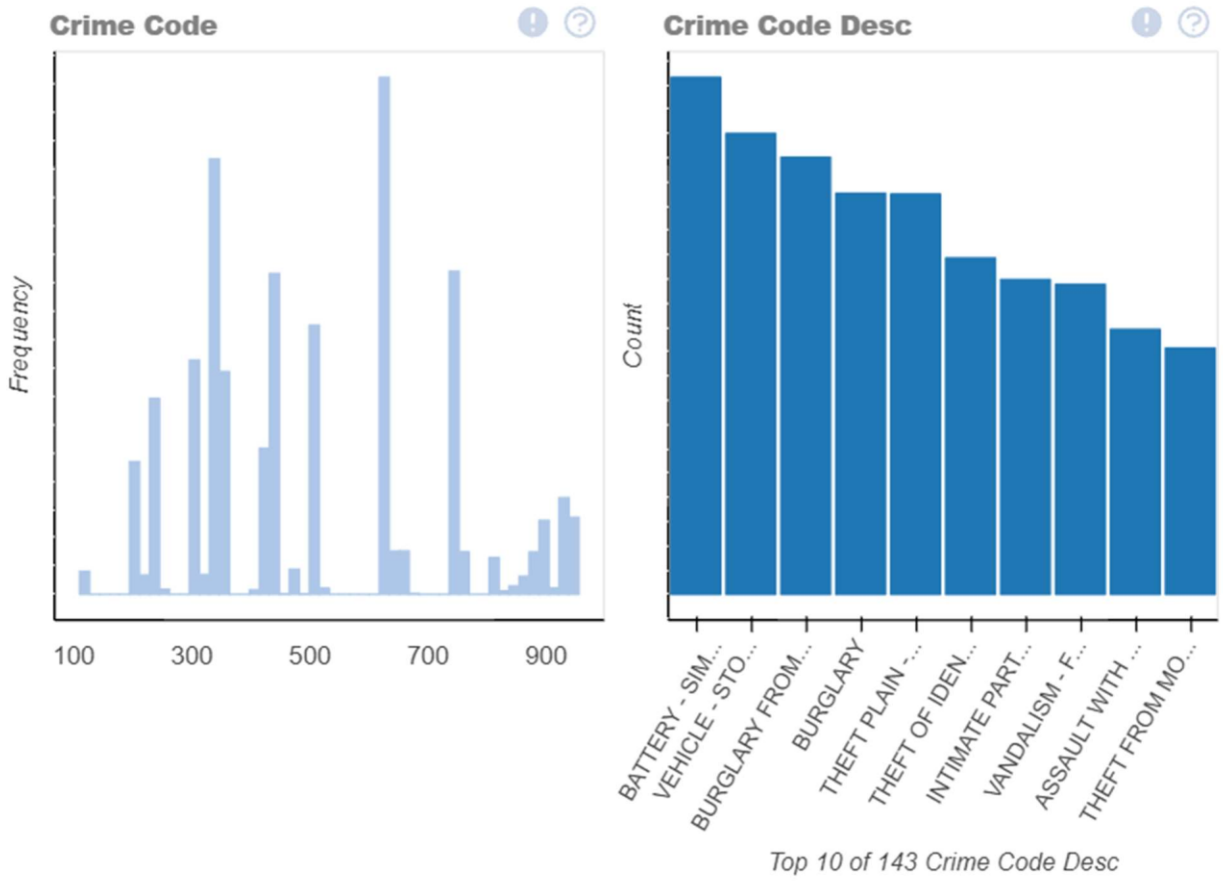

Figure 4 shows the distribution of crime codes left and top ten most occurred crimes in LA right.t

Victim Age

The distribution of victim's age shows right-skewed distribution indicating, younger people are most likely to be the victim of the crimes. While distribution also shows a spike on 0 values, that indicates, the age of the victim was not available.

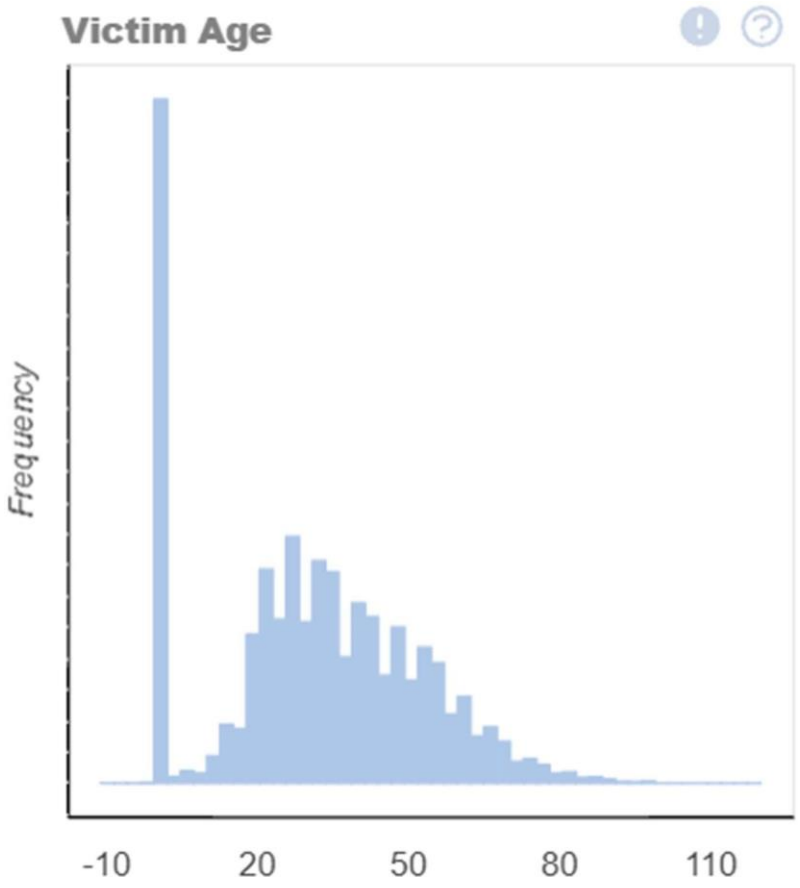

Figure 5 shows the distribution of the victim's age.

Victim’s Gender

This distribution shows an almost equal distribution of gender.

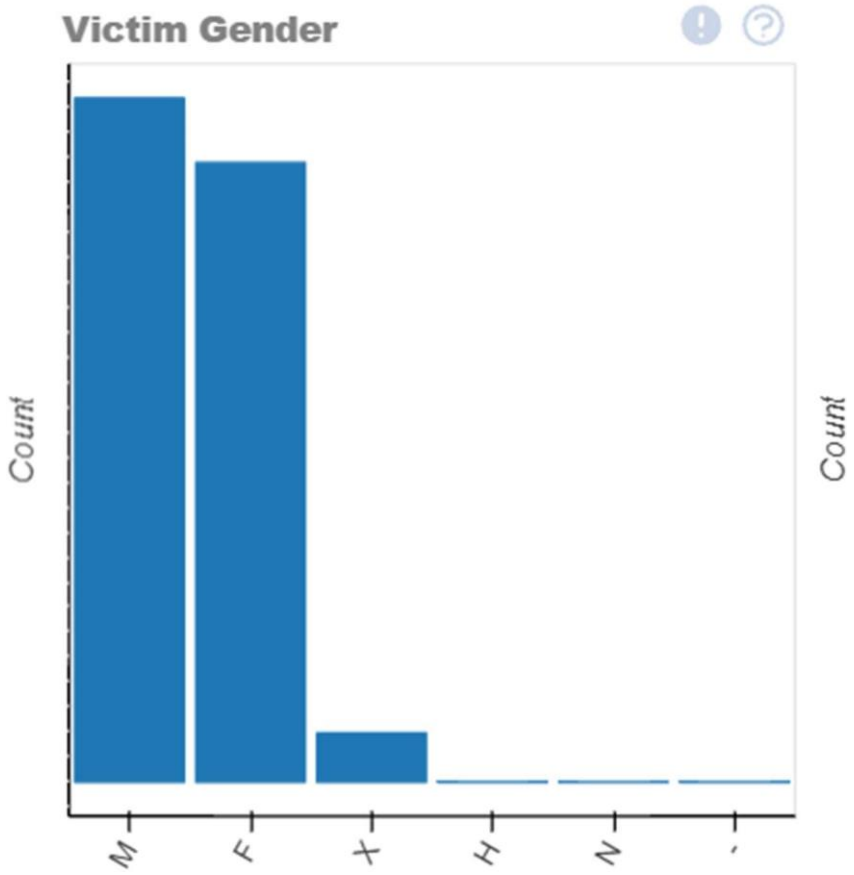

Figure 6 shows the distribution of the victim's gender.

Victims Descent

This distribution shows the race of people affected by the criminals.

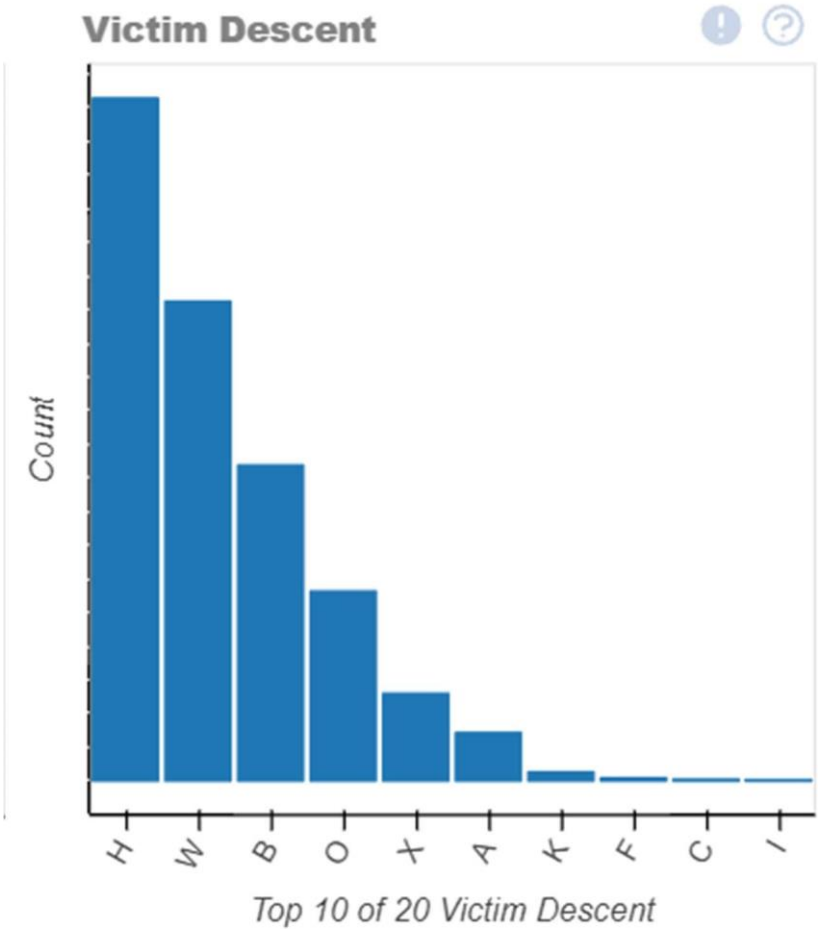

Figure 7 shows the distribution of victim's descent

Weapon Description

The distribution of weapons used in the crimes indicates most people were the victim of strong-arming and verbal threats on third and handgun as third.

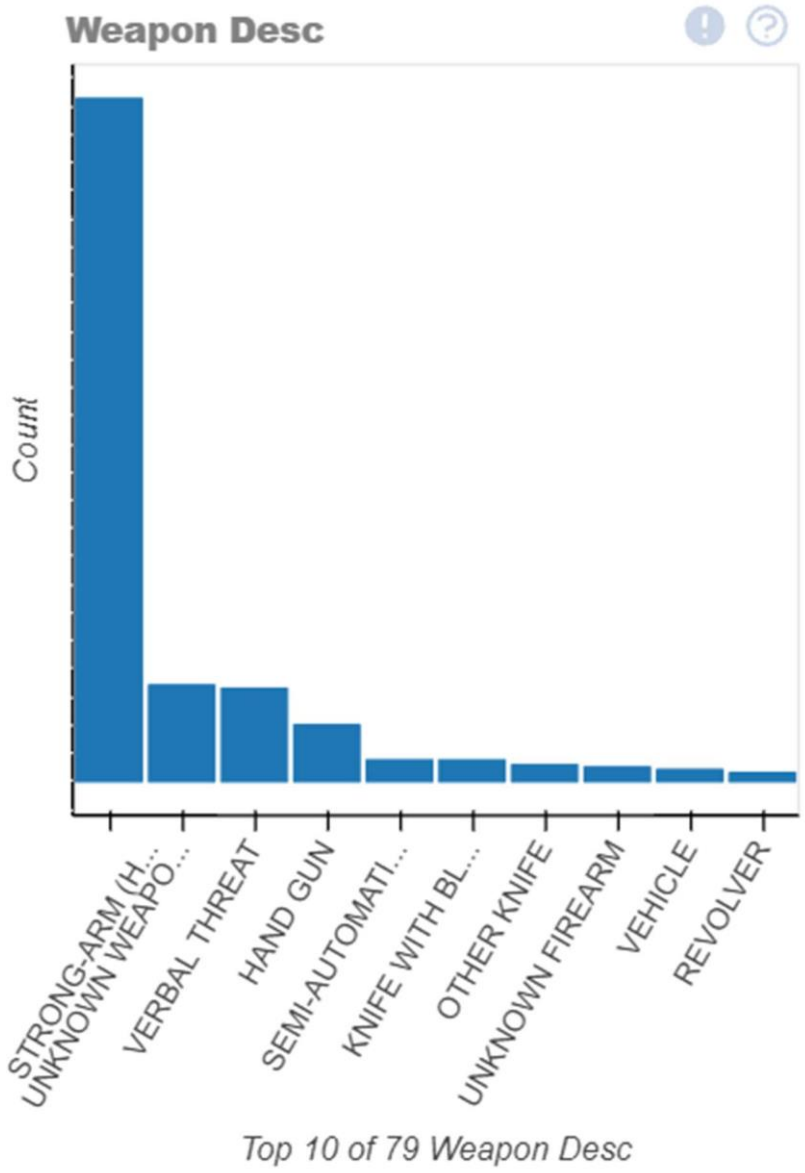

Figure 8 shows the top 10 most used weapons

Premise Description

The premise distribution shows most of the crimes occurred in streets and on second in single-family residence and third in multi-family residence.

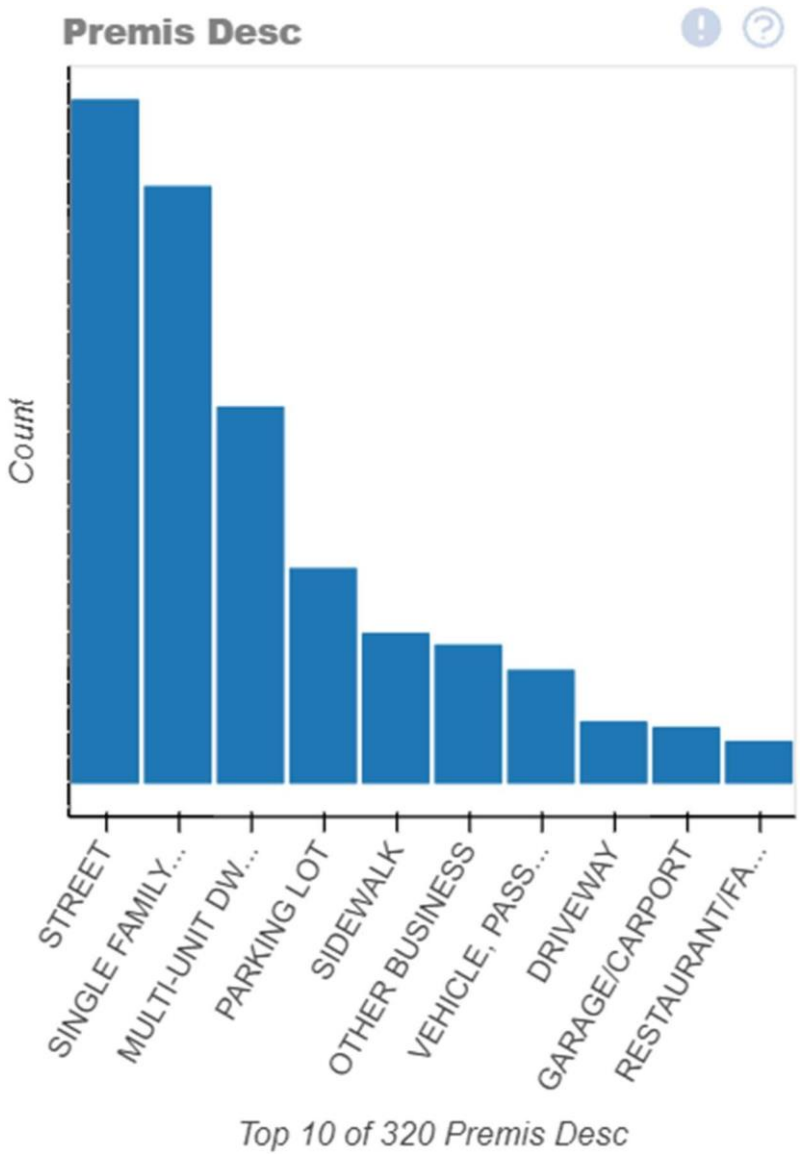

Figure 9 shows the distribution of locations of crimes

Crime Location

The distribution of crime locations indicates 6th street and 7th street being the top crime locations in the city.

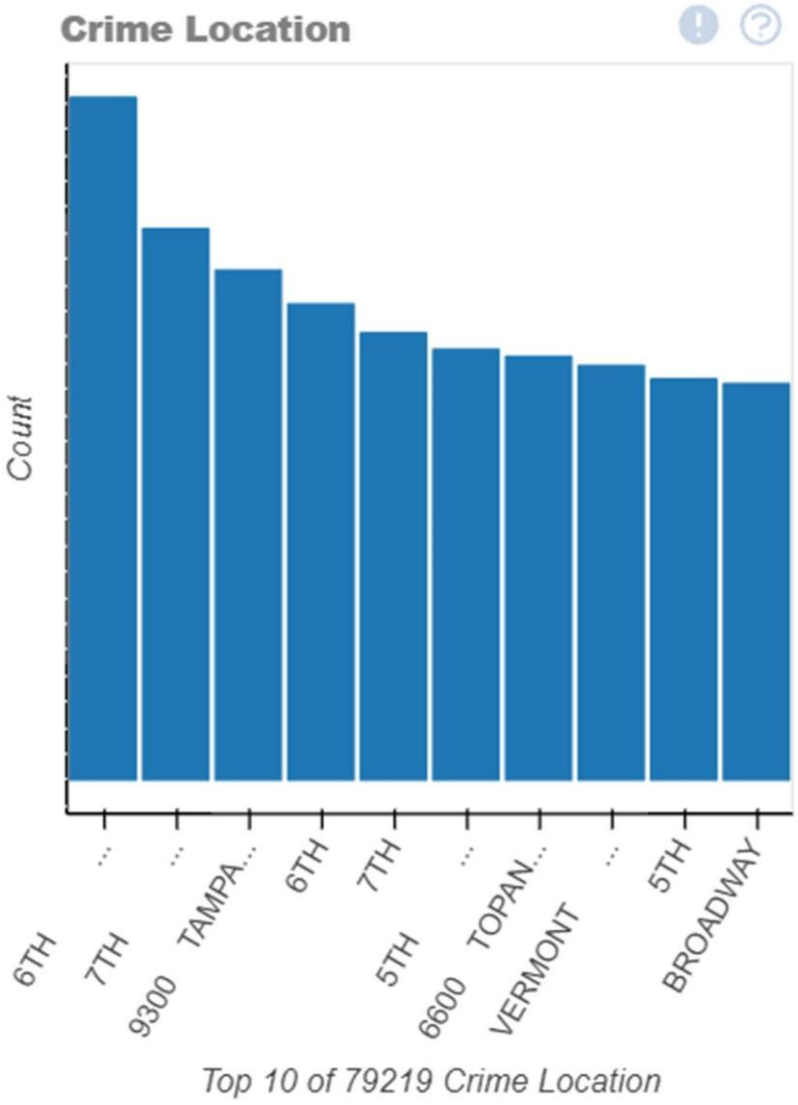

Figure 10 shows the distribution of crime location

Correlation Plots of the Dataset

The correlation graph of the dataset was created using the seaborn library function pair plot.

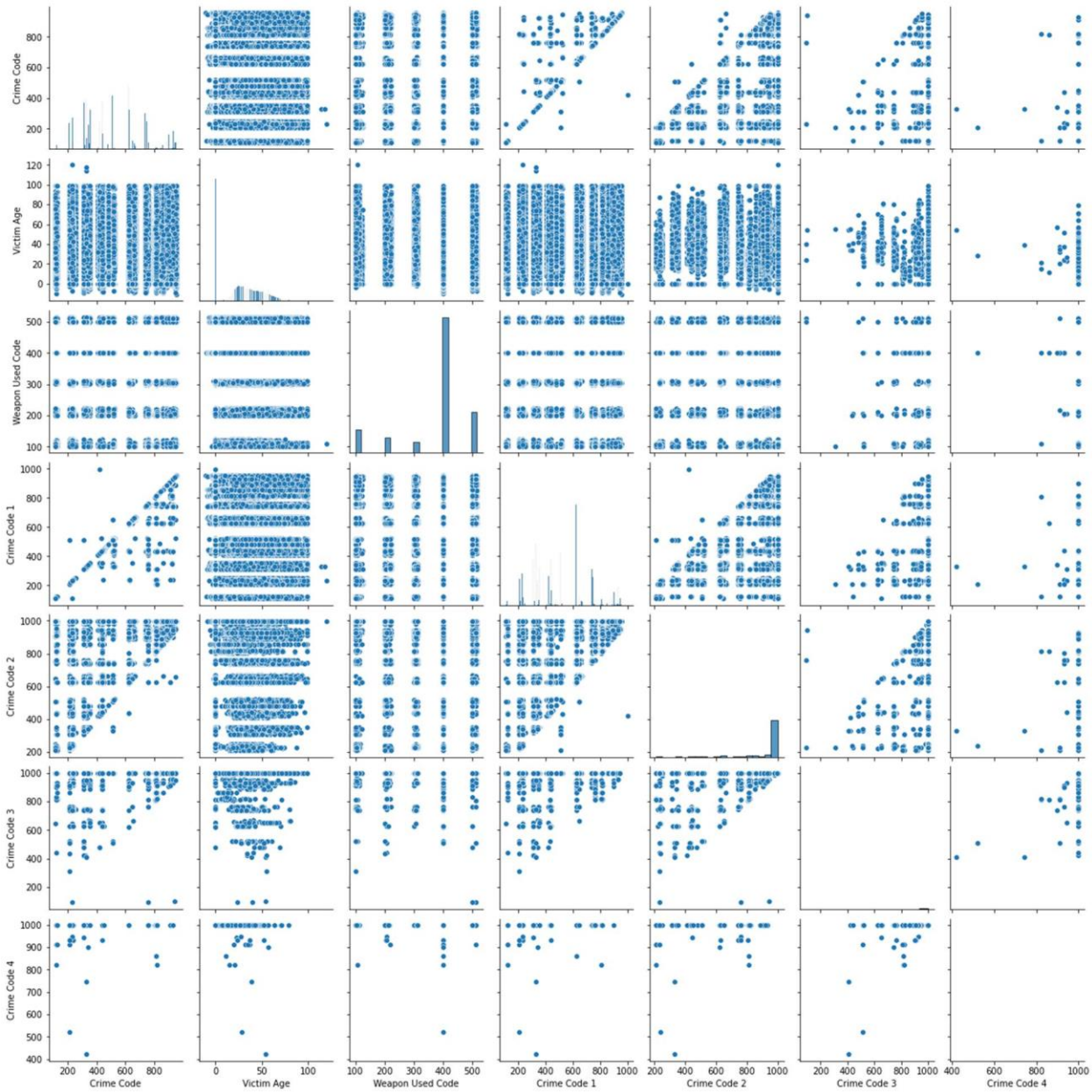

Figure 11 shows correlation graphs

Feature Engineering

Feature engineering was performed for a better understanding of the dataset.

- Reported Date
  From this feature, 3 new features based on day month, and year were created.
- Date Occurred
  From this feature, 3 new features based on day month, and year were created.
- Difference of 'Crime Occurred' and 'Crime Reported' (in Days)
  From these two features, a new feature was added to see the difference in crime that occurred and was reported in days
- Crime Time
  From this feature, a new feature was added, and time was distributed in different categories.
- Victim Age
  From this feature, a new feature was added, and age groups were formed.

Table 4 shows new features that were added

Feature name	**Pandas' datatype**
Crime Reported Day	Int64
Crime Reported Month	Int64
Crime Reported Year	Int64
Crime Occurred Day	Int64
Crime Occurred Month	Int64
Crime Occurred Year	Int64
Crime Occurred Reported difference (days)	Int64
Crime Occurred Hour	Int64
Crime Occurred Time Description	category

Data Cleaning

Data cleaning was performed for a better understanding of data.

- Data Cleaning on "Victim Gender" Column
  Data cleaning was performed on the Victim Gender column and values such as 'X', 'H', 'N','–' were converted to 'unknown'.
- Renaming "Victim Descent" Data
  Different letters representing different races were converted to their full name.
  'H','W','B','A','O','X',"K",'F','C','I','L','P','J','V','U','G','D','S','Z' to 'Hisp|Ltn|Mxcn', 'White', 'Black', 'Other Asian', 'Other', 'Unknown', 'Korea', 'Filipino', 'Chinese', 'AmInd|AlskNtv',' Laotian', 'Islander', 'Jpn', 'Vietnam', 'Unknown', 'Guaman', 'Cambod', 'Samoan', 'Asn-Indian'

Questions (Inter crime dataset)

Question 1

Hypothesis: Crimes are most likely to happen at night. [4]

Question: Do most crimes in LA happen at night. Further, are females most likely to get affected by crime at night?

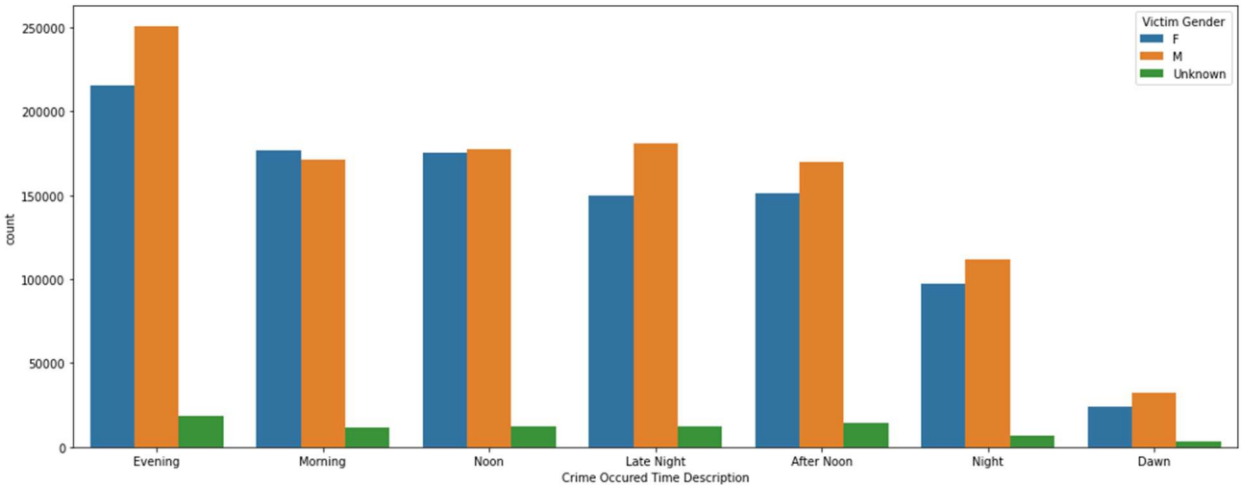

Figure 12 shows the relationship between crimes frequency, crime time, with gender segregation.

From the graph we can observe, most crimes happen in the Evening(from 7 pm to 10 pm) and at Late night (10 pm to 1 am), this verifies the constructed hypothesis. However, by looking at the data we cannot conclude that females are likely to get more affected at night. Therefore, there is no direct relation of females specifically affected by time in LA.

Question 2

Hypothesis: Cases of crimes against children and child trafficking is increasing all over the world [5] [6]

Question: What are the trends of crimes against children and what is the reporting time difference in LA.

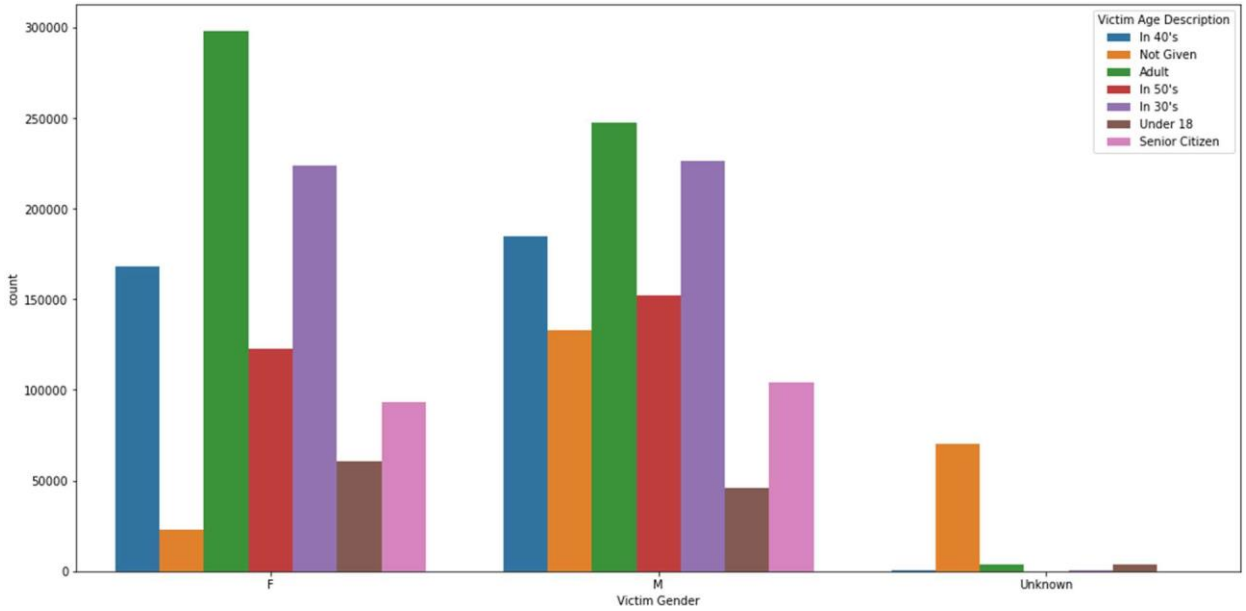

Figure 13 shows the number of cases distributed by victims age and segregated by victim's gender

In Figure 13 we can observe adults are most victims of the crimes, while interestingly we can also see many males have not reported their age.

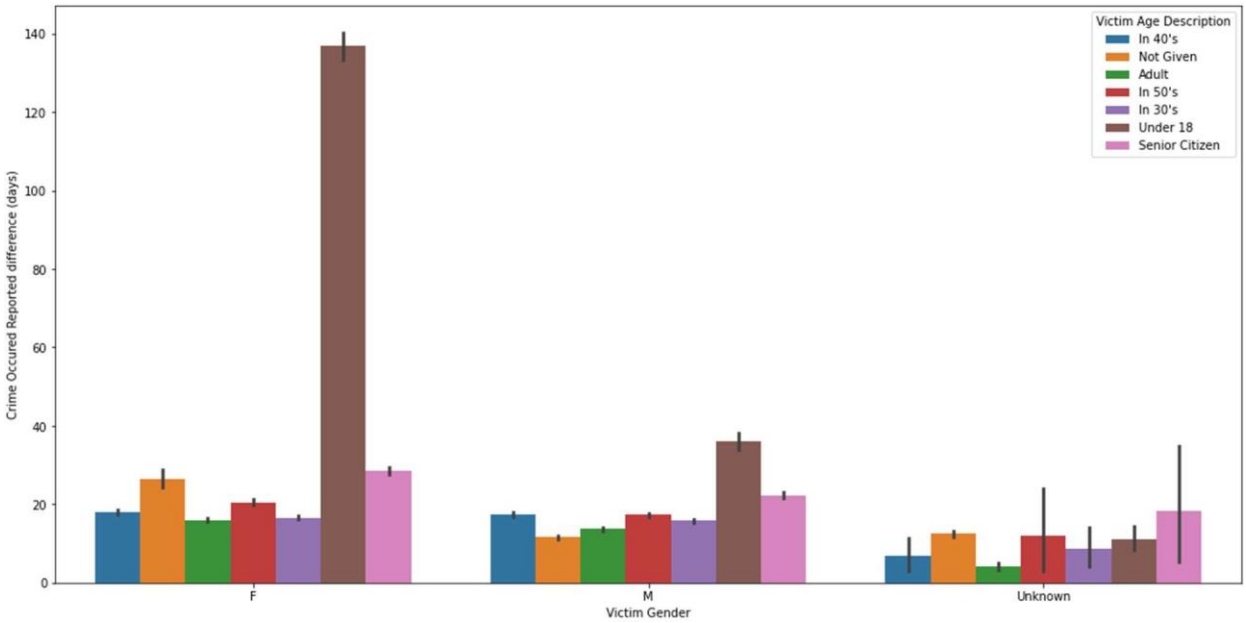

Figure 14 shows the number of days after the crime was reported, segregated by the victim's gender.

In Figure 14 we can observe that crimes against children have the highest number of days from the day the crime occurred vs the day when the crime is reported. The reluctance to report the crime is generally higher among females but it's very high among underage females.

Question 2. Intra crime and other datasets

Question 1 - Covid 19 Dataset

Hypothesis: After COVID-19, there is an increase in crime due to daily wagers losing their jobs because of lockdown. [7]

Question: Did the lockdown influence the overall crime ratio in the city?

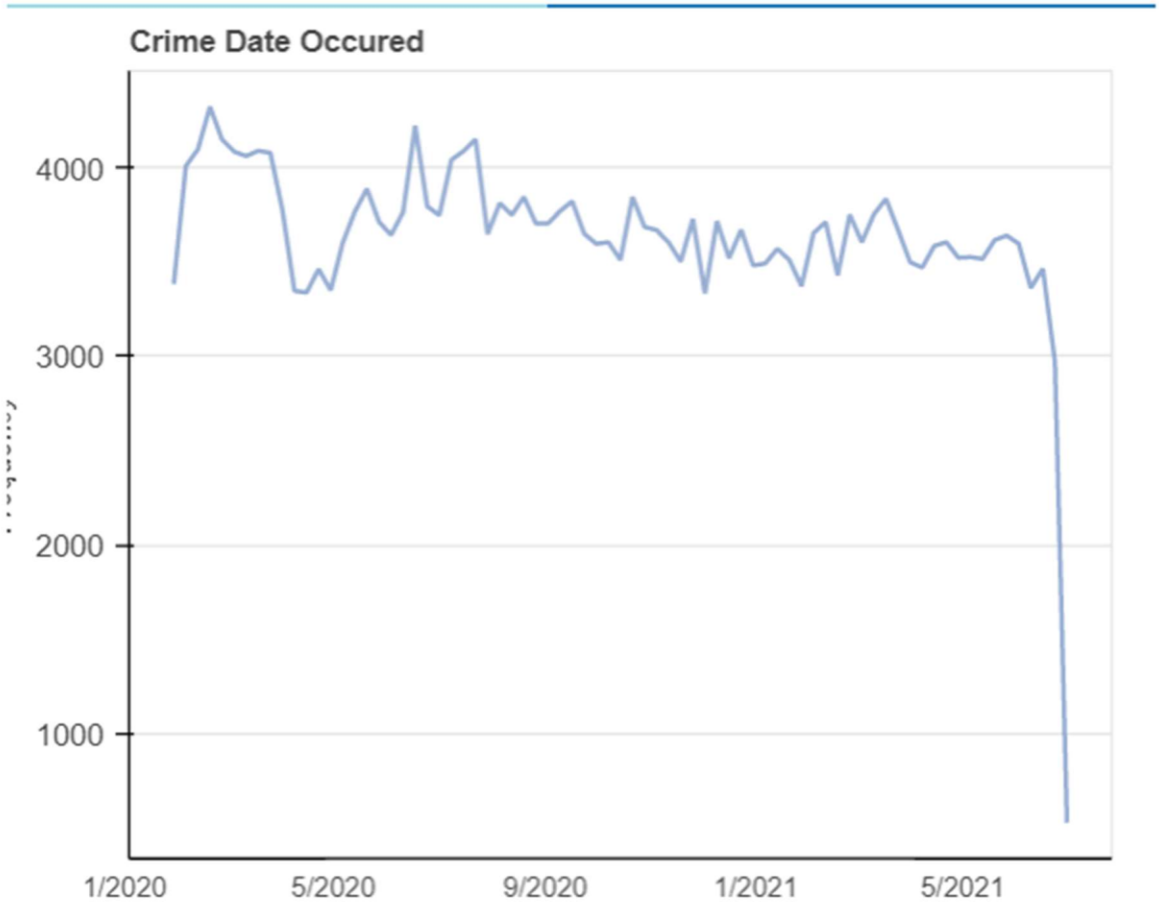

Figure 15 shows the number of crimes from January 2020 to present

In figure 15 we can observe that when the city of LA announced lockdown in March 2020, there is a sharp decline in the number of crimes as expected because streets were empty, and movement was not allowed but the number of crimes rises back up in a V shape recovery. This V-shape curve can be attributed to much better financial condition of the state of California as it can take care of its people.

Question 2 – Arrest Dataset

Hypothesis: Criminals are likely to commit crimes close to their living place. [8]

Question: What is the relationship between Crime places and criminal arrests?

Figure 16 shows a word cloud of crime area and Arrest area

From figure 16 we can observe that there are a few regions are highlighted by the word cloud in both figures. It cannot be concluded that criminals are living in the same places where they are conducting the crimes.

Question 3 – weather Dataset

Hypothesis: Winter stops crime [9]

Question: What is the relationship of Winters with the crimes?

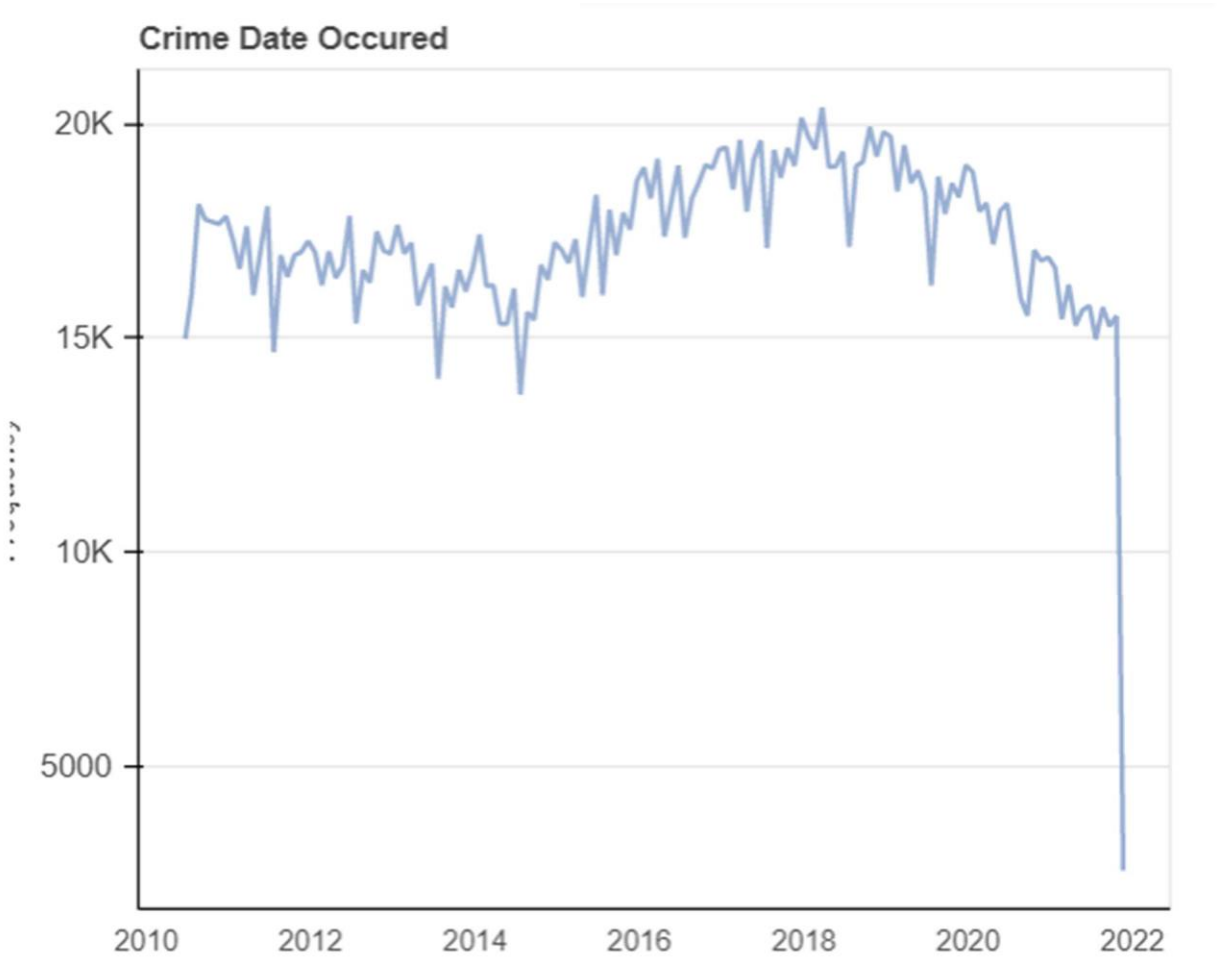

Figure 17 shows the number of crimes over the years

From figure 17 we can observe that every year there is a drop in crimes. This V shape curve usually starts at the start of the winter season and goes back to normal with the start of summer. From this data, this can be concluded that winter does stop crime.

References

[1] "California Rankings and Facts," [Online]. Available: https://www.usnews.com/news/best-states/california.

[2] "Data LA," [Online]. Available: https://data.lacity.org/.

[3] [Online]. Available: https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z.

[4] [Online]. Available: https://www.securitymagazine.com/articles/90384-murder-robbery-and-driving-while-impaired-happen-at-night.

[5] [Online]. Available: https://www.iom.int/sites/g/files/tmzbdl486/files/our_work/DMM/MAD/A4-Trafficking-External-Brief.pdf .

[6] [Online]. Available: https://www.dailynews.com/2019/10/20/la-police-failed-to-investigate-4000-serious-child-abuse-reports-in-2018-and-2019-why/.

[7] [Online]. Available: https://www.dailynews.com/2021/08/22/report-coronavirus-pandemic-key-factor-in-las-spiking-homicide-rate/.

[8] [Online]. Available: https://core.ac.uk/download/pdf/232845703.pdf.

[9] [Online]. Available: https://www.nbcnews.com/news/us-news/does-cold-stop-crime-it-seems-so-n309856.