ANALYSIS OF CRIMES IN LOS ANGELES

Nithish Christopher
x*23116099 @student.ncirl.ie*
Data Analytics
National College of Ireland
Dublin Ireland

Tamil Selvan Giri Moorthy
x23189975 *@student.ncirl.ie*
Data Analytics
National College of Ireland
Dublin Ireland

Kiruthika Suresh
x*23189916 @student.ncirl.ie*
Data Analytics
National College of Ireland
Dublin Ireland

***Abstract*— The big cities like Los Angeles which is the most popular city in the United State of California. It consists of 3.9 million people (about twice the population of New Mexico) residing within the city. With these millions of people in one city, crimes will also be more. A crime is a harmful illegal act that can be done to an individual and to society. In Los Angeles crimes are increasing day by day and it is reported that violent crimes in California have increased by 6.1%. Crime should be reduced to ensure the wellness of the people in our city and proper actions should be taken to reduce crime. Our research's main goal is to identify and analyze the patterns and trends of crime in Los Angeles. Using our analysis, we can observe the crime that is most occurred, and we can find the place where most of the crimes occurred. We used python programming language to evaluate the data, Dagster ETL tool to execute the ETL functions. From our analysis the Los Angeles Police Department can get an idea to reduce these crimes from happening.**

Keywords—Los Angeles, crimes in past years, arrested list from crime, victims

I. INTRODUCTION

A. PROJECT MOTIVATION

In recent days, the unacceptable crimes are happening more in all over the world. The purpose of the project is to achieve analysis of Los Angeles crime data in multiple years. Los Angeles is the second most populous city in the United States, because of population, urbanization, Infrastructure deficiencies, various illegal activities, unemployment and other more reasons crime rate increasing rapidly. For this detailed analysis we used three datasets namely, Crimes in 2013 to 2019, crimes in 2020 to present, arrested in 2020 to present. In this dataset we get details about crime happening between years (2013 to 2023), to identify the trends of crime pattern, the age of people who are affected by crime, crime happening areas and criminals arrested by crime. Here particularly we are focusing on analysing the top 10 crimes happened, to find the most affected areas because of crime and arrest rate of criminals in some past years. The pattern of crime cannot be predicted accurately since it is not systematic. But with the analysis we can find other factors that influence the crimes like place, the type of safety measures that we can use to reduce the crimes.

Analysing about crime from 2013 to present and arrest data from 2020 to present and we get to know about emerging of crime and we can focus to crime prevention activities. At the same time improving precaution through area wise and age factor. It helps to know about which crime pattern followed criminal investigated and arrested as well as we can increase community and city protection based on crime pattern from arrested data, improve security on those locations.

B. OBJECTIVES

Analyzing this crime data and arrest data we can understand the crime patterns followed from higher-level to lower-level which crime is happened most, so this can be identified by doing year wise crime comparison from 2013 to present. By evaluating age, gender category, and area data. We can determine which type of age category people, and which type of gender category people were affected lot of this crime incidents, and we can find the area where the crime occurred a lot. We are finally analyzing arrest data from 2020 to the present which age category people mostly arrested by crime.

II. RELATED WORK

[1] This study's analysis examines understanding about crime and victimization in Los Angeles. In Particularly what kind of crimes people in Los Angeles have experienced as victims. It also focused on how frequently crime victims personally know the people who harmed them, emphasizing the relationship between intimacy and crime. And another goal of the study is to understand how people live over the city, regarding safety. Also compare results with what other studies have found and look at in the current statistical data. Overall, this complete study gives information about several types of crime in big cities like Los Angeles.

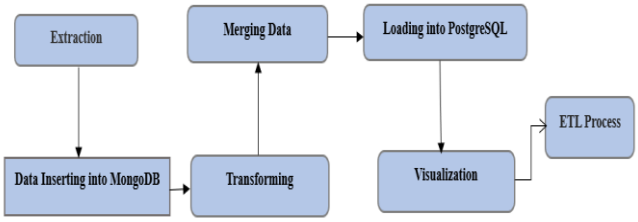
In this analysis, they used 218 participants who are all currently living in greater Los Angeles. The respondents are separated followed by differences of gender and ethnicity. And this data collected from who are all answered for conducted survey. They used multiple analyses executed on this data particularly they used descriptive and inferential analysis using through SPSS. In descriptive analysis, they focused on categorical variables, and ANOVA was used to detect differences between ages, wealth status and ethnicity. And multiple correlations are performed to identify any important between crime, victimization, type of crime. From parametric measures it was found that crimes frequently happened was robbery in the city and using summary of data found who are mostly targeted in crime by age and ethnicity. Some more measures of correlation in several important relationship between variables from crime data analyzed which respondents are feeling safe in ethnicity wise in Los Angeles. Using Crosstabulations founded who are all residents have targeted in through gender wise and wealth status or social class.

In this final discussion, using SPSS analysis on correlation part they identified the common crime was robbery and crosstabulation to analyzed Hispanics and Whites usually to

feel safe, while middle eastern residents are feeling less safe from this survey, Women are feeling less safe compare with men these are analyzed. Because of emerging crime, we are curious about to analyze these kinds of problems over the world and from this article and analysis we more interested to do crime analysis on Los Angeles. [1]

III. MEATHODOLOGY

In this project we are using 3 datasets namely "arrest data", "crime data 2013 to 2019", "crime data 2020 to present" all the datasets are collected from data gov website. The "arrest data 2020 to present" dataset describes the arrest incident that happened in the city of Los Angeles. It contains the information of people arrested by the police. The "crime data 2013 to 2019" and "crime data 2020 to present" describes the crime incident that happened in the city of Los Angeles from 2013 to 2019 and 2020 to present. It contains information about the people who are attacked by the criminals. All these datasets are transcribed from the original crime reports.



A. TOOLS AND LIBRARIES

For our work we are using Docker Desktop to create a separate container for MongoDB and PostgreSQL. We utilized MongoDB to load the dataset first. In this process we are following ETL method, first step is to extract the data in Mango DB, then transform the data for your requirement and load the data in the Dragster server. To visualize the data, we are using seaborn and matplotlib

B. OPERATIONAL FLOW

- **Arrest Data**:

  We collected this dataset from the data gov website. It is a semi-structured dataset (xml file), containing the details of people arrested by the police department. The data contents of the Arrest Data are illustrated in the below figure 1. This dataset has an equal number of integer and string variables. This dataset contains complete details about arrested people from Los Angeles, we are used as mentioned website gave much information about our dataset.

  We had many columns from which the important columns are picked, and others are removed from our dataset to achieve our objective questions. It contains details of people arrested from the year 2020 to the present.

ARREST DATA		
VARIABLES	**DESCRIPTION**	**DATATYPES**
report id	Describes the report id of the crime	int
Arrest Date	Respected arrest date	datetime
Area ID	Describes the area id with the code	int
Area Name	Describes the area name where the crime incident happened.	string
Rpt Dist No	gives reported district number	int
Age	age of the offender	int
Sex Code	Gender of the criminal	string
Address	Street address where the offender got arrested.	string

Figure 1

- **Crime Data (2013- 2019)**:

  This dataset includes semi-structured data (xml file) gathered from the data gov website. It contains brief details of the victims who were attacked by the offenders from 2013 to 2019. The content of this dataset is illustrated in Figure 2.

CRIME DATA (2013 – 2019)		
VARIABLES	**DESCRIPTION**	**DATATYPES**
DR_NO	Describes the division of records, an official file number	int
Date Rptd	Date of the incident reported	datetime
DATE OCC	Describes the area id with the code	datetime
TIME OCC	The time of the incident in 24-hour military time	string
AREA	describes community police stations of the area	int
AREA NAME	Describes the name of the geographic Area	int
Rpt Dist No	A four-digit code representing a sub-area within a geographic area	string
Crm Cd	The code indicating the crime committed	string
Crm Cd Desc	Indicates the crime committed and it denotes the primary offenses	string
Vict Age	Age of the Victim	int
Vict Sex	Gender of the Victim	String
Location	Street address of the crime incident	string

Figure 2

- **Crime Data (2020- 2024)**:

  This is structure data (csv file) that has been used for the dataset. This dataset is gathered from the data gov website. It contains brief details of the victims attacked by the offenders from 2020 to present. The content of this dataset is illustrated in Figure 3.

CRIME DATA (2020 - Present)		
VARIABLES	**DESCRIPTION**	**DATATYPES**
DR_NO	Describes the division of records, an official file number	int
Date Rptd	Date of the incident reported	datetime
DATE OCC	Describes the area id with the code	datetime
TIME OCC	The time of the incident in 24-hour military time	string
AREA	describes community police stations of the area	int
AREA NAME	Describes the name of the geographic Area	int
Rpt Dist No	A four-digit code representing a sub-area within a geographic area	string
Crm Cd	The code indicating the crime committed	string
Crm Cd Desc	Indicates the crime committed and it denotes the primary offenses	string
Vict Age	Age of the Victim	int
Vict Sex	Gender of the Victim	String
Location	Street address of the crime incident	string

Figure 3

IV. ETL

We are connecting MongoDB and PostgreSQL using Docker. Then we are converting both crime datasets (xml file) and arrest data (csv file) into dictionary format and loading them into MongoDB. For each dataset we are collecting the objects from MongoDB and store them into a data frame and proceed with the further transformation. Initially we removed all the irrelevant columns from the data frame, and we handled the null values, missing values, and outliers. We replace the null values as required for our project.

Further, as part of the transformation, we created a new variable year extracted from the date reported on crime dataset (2013 to 2019) and the crime dataset (2020 to present). Similarly, for arrest data we created a new variable year with reported arrest date. By converting the date time variable and extracting only the year which is useful to sort the crime incidents according to the year format. Also, we are creating a new column location by adding the two variables, namely location and cross street, which is combined to give an exact address of the crime and arrest incidents. The variable's location and cross street does not give any information when they are separated. But it gives valuable information on crime when they are combined. So, these variables are merged into a new variable and the separated variables are dropped from the dataset. Similarly, we do the same process for the remaining two datasets.

We created a new Data Frame called merged data where we combined both crimes from 2013 to 2019 and crimes from 2020 to present. So that it can be used to find the trends and patterns of crime. We are also finding a filtered dataset where we filter the top 10 crime types so that we can find the best trends and patterns for our objective questions. We have used the dagster for our ETL server and after loading we create a job file to run all the ese into dagster server, from the pipeline visualization we could analyse and understand the data workflow.

To load the transformed dataset, we need to connect the PostgreSQL. we have merged and the transformed two data sets “crime data 2013 to 2019”and “crime data 2020 to 2024” which are stored into the PostgreSQL database as single table named “crime data” and stored the arrest data as a separate table in the PostgreSQL “arrest data”. When it is successful we create an engine and load the transformed dataset to PostgreSQL through an SQL query. Through this method our transformed data is loaded to PostgreSQL. Once we completed all these steps, we did the analysis for the variables from both tables, and we did the visualization process to achieve our research purpose.

V. RESULTS AND EVALUTION

We can get a clear picture about the data through charts, graphs, and plots. It allows us to identify specific patterns, structures, or trends. To visualize the datasets, we are using matplotlib and seaborn libraries. We created a count plot, box plot, heatmap, line plot for our dataset using matplotlib. By using seaborn we can all the data into a single plot.

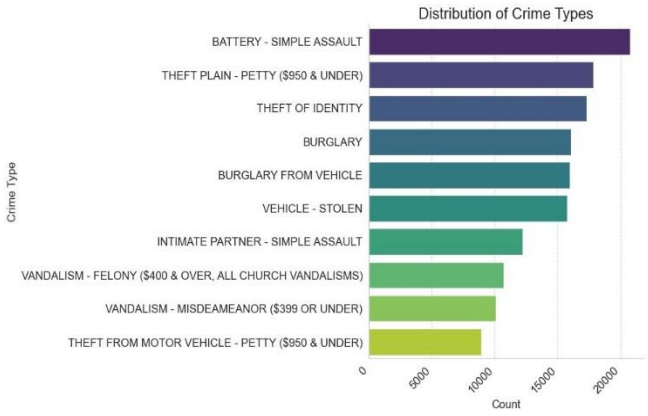


Figure 4

In this above figure 4 we are displaying the top crimes which happened from 2013 to 2023. There are lot of crimes listed in our dataset in that we have displayed the top 10 crimes that has happened, with this plot we can understand that over 20000 crimes are the Simple assault involving batteries, approximately 17000 where petty theft ($950 and under), and over 16000 has been identified as theft are the top 3 crimes listed in each year.

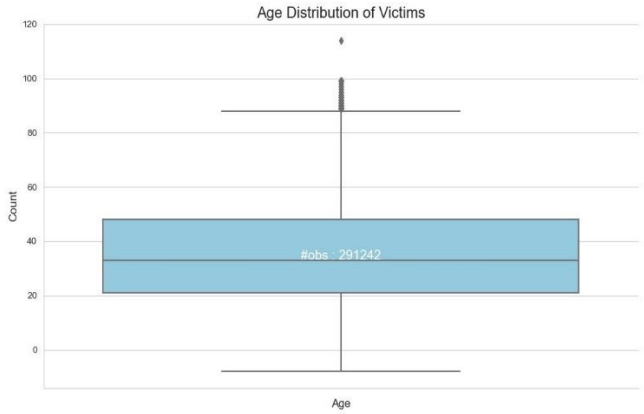


Figure 5

In Figure 5, we are using box plot to get the distribution of the victims' age so that we can understand which type of age category people are getting targeted by offenders. This plot clearly gives us an idea that around 2,91,242 people between the ages 20 to 45 are the most affected victims of crime from 2013 to 2024.

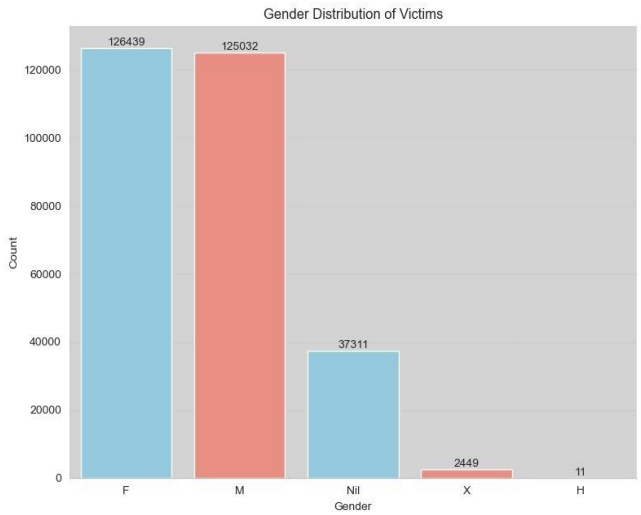


Figure 6

In Figure 6, to find the gender distribution of the victims who are affected by the offenders the most in a crime are found using the count plot. The plot clearly explains that females are the most affected victims in a crime which gives us a valuable information that around 1,26,439 women are affected, and 1,25,032 men are affected by the offenders.

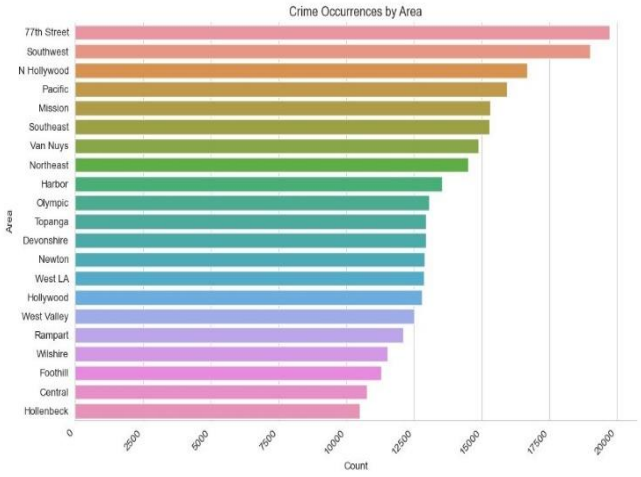


Figure 7

From Figure 7, for a big city like Los Angeles the area plays a significant role in finding the crime occurrence. Using our count plot, we could determine the most dangerous area where crimes occur in Los Angeles. From our plot, 77th street is the hotspot for crimes, southwest is the second most dangerous area, and N Hollywood is in the third place.

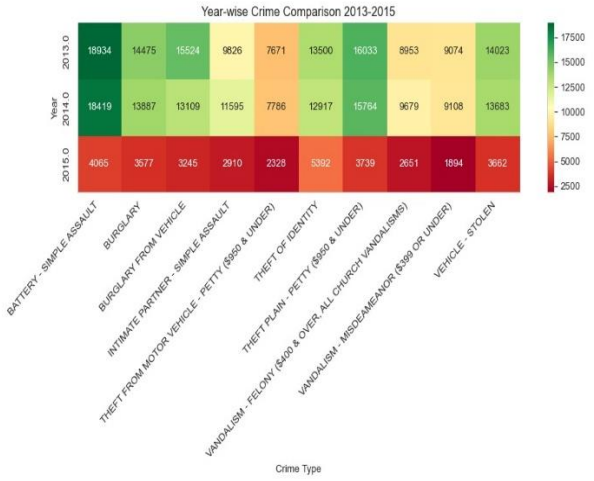


Figure 8

In Figure 8, we use heatmap to explain year wise crime comparison, we have taken 3 years from 2013 - 2015 where most crimes happened. This heatmap gives us the numerical value of total number of each crime that happened in each year. Each box represents a value of correlation between year and crime type, and we use colors to represent the strength and direction of correlation. With this plot we can get an idea that in 2013 and 2014, the crime type battery-simple assault is at the highest with 18,934 and 18,419, respectively.

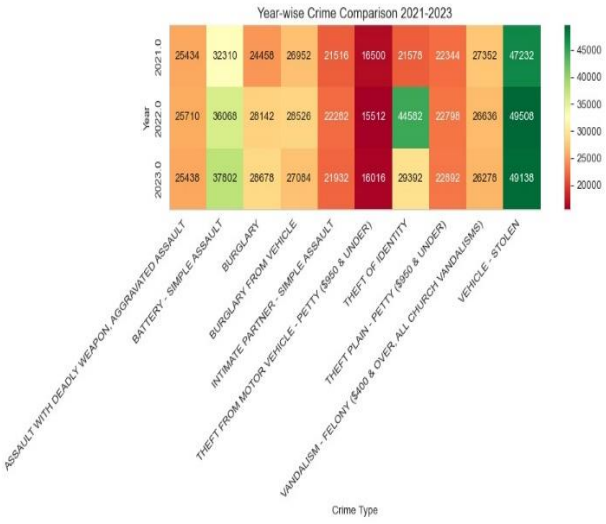


Figure 9

In figure 9, Similarly we are displaying last three-year wise crime comparison from 2021 - 2023 where most crimes happened. This heatmap gives us the numerical value of total number of each crime that happened in each year. With this plot we can get an idea that in 2023 and 2022, Vehicle-stolen crime is at the highest with 49,134 and 49,508, respectively.

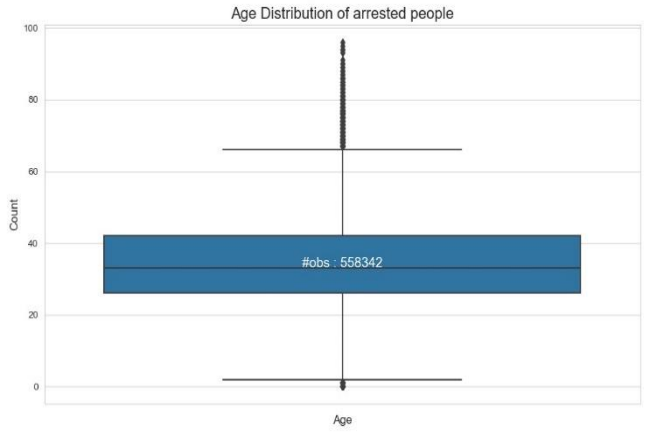


Figure 10

In figure 10 here, we used the 3rd dataset (arrest data 2019 to present) for this box plot analysis. Our plot contains different types of filed cases and people of different ages arrested, here we found that the age of 25 to 42 people mostly arrested for crime. It shows the people arrested in the last 4 years (2020 to 2023) between 25 to 42 years.

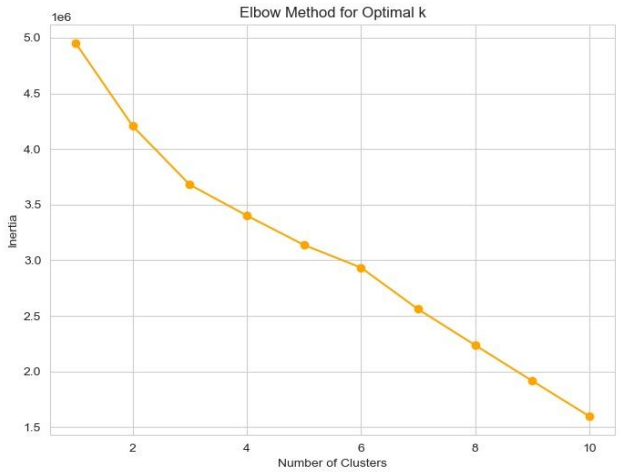


Figure 11

From figure 11, it represents the K-means Clustering on our crime dataset. We include the important columns like 'Year,' 'Victim Age,' 'Victim Sex' and 'Crime Description' as features for understanding the crime patterns over time and across different demographics. Since the 'Victim Sex' and 'Crime description' are categorical data, we encode them and convert them into numerical to determine the optimal number of clusters using elbow method. It ranges from 1 to 10 cluster numbers and computes the inertia (within clusters- sum of squares) for each number of clusters. The inertia is plotted against the number of clusters to determine the optimal K. In Figure 9 when the inertia begins to decrease at slower rate forming an elbow like shape indicates the optimal number of clusters.

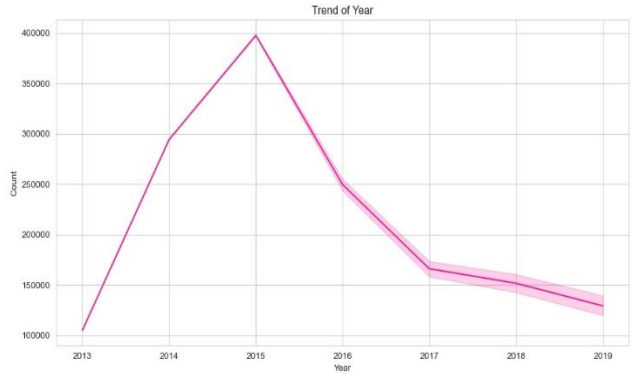


Figure 12

In Figure 12, the time series analysis shows that the crime rate tracked from 2013 to 2019 indicates that the crime rates are gradually increasing from 2013 to 2015, followed by a decrease towards 2019. In 2015, approximately 4,00,000 crimes were registered. It is at peak during the observed period of 2015. This trend suggests that there is a sudden increase in crime rates over the years analyzed with our given dataset as shown in figure 10.

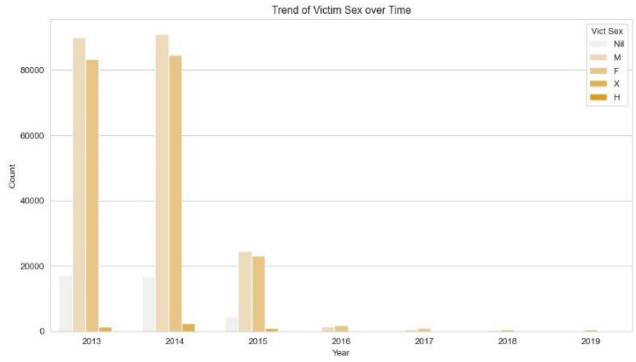


Figure 13

In Figure 13, we are analysis trend of victim sex over various time periods from 2013 to 2019 using time series analysis.

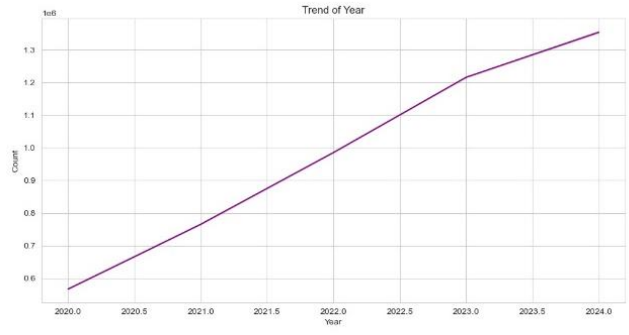


Figure 14

Figure 14 visualization indicates the crime rates that are tracked from 2020 to 2024. It shows that the crime rates are gradually increasing from 2020 to 2024.

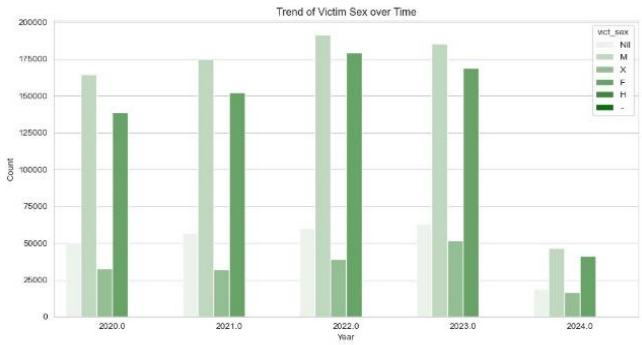


Figure 15

From figure 15, we can analyze the trend of victim sex with time series analysis, we use the 'Year' and 'Victim Sex' for plotting this plot.

VI. CONCLUSION AND FUTURE WORKS

By understanding our given 3 datasets, we have witnessed that the crime rates have increased drastically in recent years. From our analysis we have determined that the Battery-Simple Assault crime and Theft related crime are the most repetitive crimes which is happened from the year (2013 – 2023). By evaluating the age, gender category and area data, we found that people between the ages of 25 and 40 are significantly affected. By doing the gender analysis we identified that women are highly affected by offenders. 77th street is the most dangerous area in Los Angeles, nearly 19,500 crimes have occurred in that area. From the year 2020 to 2024 crimes like vehicle stolen, Assault with deathly weapon, vandalism is increased a lot compared to previous 10 years. From our final analysis on arrest data, we observed that 25 to 42 age category people are mostly registered as criminals in the year 2020 to 2024.

This detailed analysis and visualization gives us an overview of crime for upcoming years in Los Angeles. From this analysis we can help our Los Angeles Police Department to take necessary measures so that these crimes won't be repeated. In a survey we found that in 2016, 290 people were killed in the city of Los Angeles. From 2015, there is a 13 % increase in robberies. There was a 38% increase in 2014 and 10% increase over 2015 in overall violent crimes. We can say that there are 2 main crimes which needs to be reduced mainly violent crimes like Robbery, Assault and Property crime like burglary and grand theft. With our analysis the Los Angeles Police Department can take safety measures for women in our survey we get to know that women are the most affected victims, and they can conduct various awareness programs to notify that the 25 to 40 age people to be aware of these crimes and they can proceed with precautions while simultaneously ensuring their own protection. The Areas were the crimes occurred the most should be monitored, and securities can be improved by placing more security cameras. Government officials can give more safety protocols on unsafe areas from highest to lowest, whatever mentioned in analysis for more registered crime pattern they can increase security services. At the same time from the list of criminals age and crime registered areas they can conduct any volunteer counseling on the title of "Leaving Crime Behind" and should initiate on increasing employment services for those arrested people to reduce criminal activities.

VII. REFERENCES

[1] https://www.scirp.org/journal/paperinformation?paperid=85693

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.