**ANALYTICAL STUDY ON THE CRIME RATE IN CHICAGO.**

Masters in Science in Data Analytics (MSCDAD\_B)

Database and Analytics Programming

ANNJOYS ROBERT

*Database and Analytics Programming*

*National College of Ireland*

*X22137459@student.ncirl.ie*

ALIASGAR ABDULHUSAIN WADHWANWALA

*Database and Analytics Programming*

*National College of Ireland*

*X21236381@student.ncirl.ie*

SHAJO VARGHESE ERAMPLACKAL

*Database and Analytics Programming*

*National College of Ireland*

*X22115943@student.ncirl.ie*

**I. ABSTRACT:**

In this project, three crime datasets from Chicago are analyzed. The datasets are loaded into MongoDB, and Pandas library in Python is used to carry out an ETL (Extract, Transform, Load) procedure. The three datasets are combined and added to Postgres SQL. This merged dataset will be extracted from Postgres SQL, and we will be using the Plotly library in Python to create visualizations.

The research looks at variables including location, time of day, and kind of crime to uncover the trends and patterns in crime rates across the city. Our aim is to create visualizations to summarize our data and to gain insight into our data to help in the decision-making process. The project's main objectives are to examine strategies to increase public safety using data-driven methods and provide insights into crime trends in Chicago city.

**II. INTRODUCTION:**

1. ***Motivation:***

Significant concern has been raised by the rising crime rates in the largest American cities. Crime data has been used to identify crime trends and pinpoint areas requiring more law enforcement presence. By studying crime records, we can discover more about the factors that contribute to crime, such as geography, demographics, and the time of day. This can help law enforcement agencies build effective crime prevention strategies and increase public safety.

1. ***Objective:***

The objective of this project is to acquire insight into the factors that lead to crime in one of the largest cities in the US, Chicago. This will help all the stakeholders and policymakers to make informed decisions and help to improve public safety.

Our specific objectives include:

1. Analysis of the most prevalent crime categories in the city of Chicago.

2. Analysis of whether the injury was caused due to gunshots or not.

3. Overview of the number of Arrests based on Different Races in Chicago and the Number of Victims based on Different Races.

4. Knowing the frequency of domestic and non-domestic crimes that have been committed in the city.

5. Analyzing the percentage of victims based on their age categories.

6. Analysis of the number of incidents where the victim is Juvenile.

The overall objective of this project is to use different Python libraries, relational and non-relational databases, and visualizations to give a thorough examination of crime trends in Chicago. By doing this, we seek to further the creation of efficient crime prevention techniques and enhance public security in Chicago.

**III. RELATED WORK:**

[1]. The "Exploratory Data Analysis of Crime in Chicago Using Python" article by Jhonatan Oliveira on Medium examines the Chicago crime dataset using Python. The article highlights the usage of libraries for data purification, exploratory data analysis, and visualization, including pandas, matplotlib, and seaborn.

[2]. "Chicago Crime Analysis and Visualization" by Shun Zhang - This GitHub repository contains a Jupyter notebook with Python code for processing and visualizing the Chicago crime dataset. The notebook covers data cleansing, exploratory data analysis, and visualization using tools like pandas, NumPy, and Matplotlib.

[3]. Mounika Namburu's "Analyzing Crime in Major US Cities with Python" focuses on crime statistics from places like New York, Chicago, and Houston. Before integrating the datasets and putting them in PostgreSQL, the author imports the data into MongoDB and performs ETL using Pandas. Using Python tools like Matplotlib and Seaborn, the post discusses data preparation, cleaning, and visualization.

[4]. Michael Huttner's "Exploratory Data Analysis of Crime Data Using Python" examines crime statistics from several US cities, including Atlanta, Denver, and Seattle. Before integrating the datasets and putting them in PostgreSQL, the author imports the data into MongoDB and performs ETL using Pandas. Additionally, the article discusses data preparation and cleaning as well as data visualization with Python tools like Plotly and Bokeh.

[5]. Ashley Cooper's "Visualizing Crime in US Cities with Python" uses Python and Pandas to examine crime data from cities including Philadelphia, Dallas, and Detroit. Before integrating the datasets and putting them in PostgreSQL, the author imports the data into MongoDB and performs ETL using Pandas. The article provides details on data preparation and cleaning as well as data visualization using Matplotlib and Plotly, two Python packages.

[6]. Marcin Kowalski's "Crime Data Analysis with Python and PostgreSQL" uses Python and Pandas to examine crime data from Los Angeles. Before putting the data in PostgreSQL, the author imports the data into MongoDB and conducts ETL using Pandas. The article discusses data preparation and cleaning as well as data visualization using Python modules like Seaborn and Folium.

**IV. METHODOLOGY**

1. ***Dataset Description***

Three datasets, each pertaining to crimes in Chicago, have been selected for analysis.

**Chicago Arrests dataset**:

Every arrest made in Chicago between 2001 to the present is listed in the Chicago Arrests dataset. It contains information about the arrestee's race, the case number, date of the arrest, and arrest number. Additionally, the dataset includes information about the charges brought against the defendant, including the law code, the charge's classification, type, and description for up to four charges. Along with geographic information like ZIP codes, community areas, census tracts, and wards, the dataset also includes information that can be used to examine how arrests are distributed throughout the city.

* sid - a unique identifier for each record
* id - another identifier for each record
* position - location data for each record
* created\_at - the time when the record was created
* created\_meta - metadata for the creation of the record
* updated\_at - the time when the record was last updated
* updated\_meta - metadata for the update of the record
* meta - additional metadata for the record
* CB\_NO - the arrest number
* CASE NUMBER - the case number for the arrest
* ARREST DATE – the date of the arrest
* RACE – the race of the accused
* CHARGE 1 STATUTE - the statute code for the first charge
* CHARGE 1 DESCRIPTION - the description of the first charge
* CHARGE 1 TYPE - the type of the first charge
* CHARGE 1 CLASS - the class of the first charge
* CHARGE 2 STATUTE - the statute code for the second charge
* CHARGE 2 DESCRIPTION - the description of the second charge
* CHARGE 2 TYPE - the type of the second charge
* CHARGE 2 CLASS - the class of the second charge
* CHARGE 3 STATUTE - the statute code for the third charge
* CHARGE 3 DESCRIPTION - the description of the third charge
* CHARGE 3 TYPE - the type of the third charge
* CHARGE 3 CLASS - the class of the third charge
* CHARGE 4 STATUTE - the statute code for the fourth charge
* CHARGE 4 DESCRIPTION - the description of the fourth charge
* CHARGE 4 TYPE - the type of the fourth charge
* CHARGE 4 CLASS - the class of the fourth charge
* CHARGES STATUTE - the statute code for all charges combined
* CHARGES DESCRIPTION - the description of all charges combined
* CHARGES TYPE - the type of all charges combined
* CHARGES CLASS - the class of all charges combined
* Boundaries - ZIP Codes - information about the ZIP code boundaries
* Community Areas - information about the community area boundaries
* Zip Codes - information about the ZIP codes
* Census Tracts - information about the census tracts
* Wards - information about the wards.

**Chicago Violence dataset:**

Statistics about violent episodes that occurred in Chicago between 2014 and the present are included in the Chicago Violence dataset. It includes information on the incident's chronology, type, victims' ages, sexes, and races, and whether it was a domestic incident. A binary indicator for the incident's total number of victims, the victim's age, and whether they were shot is also included in the data. The dataset makes it feasible to comprehend the kind and frequency of violent occurrences in Chicago.

* sid - a unique identifier for each record
* id - another identifier for each record
* position - location data for each record
* created\_at - time for when the record was created
* created\_meta - metadata for the creation of the record
* updated\_at - the time when the record was last updated
* updated\_meta - metadata for the update of the record
* meta - additional metadata for the record
* TIME\_PERIOD - the time period of the incident
* TIME\_PERIOD\_START - the start time of the time period
* TIME\_PERIOD\_END - the end time of the time period
* PRIMARY\_TYPE - the primary type of the crime
* AGE - the age of the victim
* SEX - the sex of the victim
* RACE - the race of the victim
* JUVENILE\_I - a binary indicator of whether the victim was a juvenile or not
* DOMESTIC\_I - a binary indicator of whether the incident was domestic or not
* GUNSHOT\_INJURY\_I - a binary indicator of whether the victim suffered a gunshot injury or not
* NUMBER\_OF\_VICTIMS - the number of victims in the incident.

**Chicago Crime Dataset:**

Data on crimes reported in Chicago from 2022 to the present are included in the Chicago Crime dataset. The dataset offers detailed information on the crimes that have been reported in Chicago as well as their locations, enabling the study of crime trends over time and across the city's many neighbourhoods.

* sid - a unique identifier for each record
* id - another identifier for each record
* position - location data for each record
* created\_at - time for when the record was created
* created\_meta - metadata for the creation of the record
* updated\_at - time for when the record was last updated
* updated\_meta - metadata for the update of the record
* meta - additional metadata for the record
* CASE# - the case number of the reported crime
* DATE OF OCCURRENCE - the date on which the crime occurred
* BLOCK - the location of the crime, typically the block number
* IUCR - the Illinois Uniform Crime Reporting code for the crime
* PRIMARY DESCRIPTION - the primary description of the crime
* SECONDARY DESCRIPTION - the secondary description of the crime
* LOCATION DESCRIPTION - the location where the crime occurred
* ARREST - a binary indicator of whether an arrest was made for the crime
* DOMESTIC - a binary indicator of whether the crime was domestic in nature
* BEAT - the police beat where the crime occurred
* WARD - the ward in which the crime occurred
* FBI CD - the FBI code for the crime that occurred
* X COORDINATE - the X coordinate of the location where the crime occurred
* Y COORDINATE - the Y coordinate of the location where the crime occurred
* LATITUDE - the latitude of the location where the crime occurred
* LONGITUDE - the longitude of the location where the crime occurred
* LOCATION - the location where the crime occurred
* Historical Wards 2003-2015 - the historical ward in which the crime occurred
* Zip Codes - the ZIP code where the crime occurred
* Community Areas - the community area where the crime occurred
* Census Tracts - the census tract where the crime occurred
* Wards - the ward where the crime occurred
* Boundaries - ZIP Codes - the boundary of the ZIP code where the crime occurred

1. ***Process Flow***

The JSON data was stored in MongoDB and three datasets were stored in three different collections in a database. To read data from Mongo DB, the PyMongo library was used. All three datasets were read and transformed using pipelines. An Extract function was defined which takes input from the Mongo DB collection name and returns the entire dataset as a pandas Data Frame. This pandas Data Frame will then be passed to a user-defined Transform function that contains the code for pre-processing. This user-defined function is passed as input to the pipe function which returns the initial processed dataset which will be used for merging.

1. ***Pre-processing:***

*Pre-processing for Dataset1 (Chicago Arrest Dataset):*

There were a lot of redundant columns in this particular dataset. A huge number of columns that would not be useful in the final analysis were dropped. The only columns retained were the Case number, Arrested Date, Race of the arrested person and a derived month column from the Date. The data was filtered and the data for only 2022 was taken into the processed dataset.

*Pre-processing for Dataset2 (Chicago Crime Dataset):*

The pre-processing for the dataset commenced with removing unwanted columns and filtering 2022 data from the dataset. The time column in the dataset contained information about the time the incident occurred. But it would not be beneficial for later analysis. Hence, feature engineering was carried out on the column and a new column “Time Period” was created that stores information about the time of day the incident occurred. (For ex: morning,night, etc ).The columns “Primary desc” contained a lot of unique values that described the nature of the crime. A lot of unique values implies visualizing the column would be difficult. The column was also categorized and a new column “Crime Category ” was built that had only 7 unique values .

*Pre-processing for Dataset3 (Chicago Violence Dataset):*

The pre-processing steps for the dataset was similar to the former ones. The age column in the dataset was categorized according to different age groups to get an idea on which age group is  more likely  to become  victim of a crime.

The crime and arrest dataset were joined on the “Case number” and the resultant merged data was merged with the Violence dataset on the “Date” column. The final merged data was stored in Postgres SQL. Later on , they were retrieved from the Postgres database to create Visualizations.

1. ***Visualization:***

The data visualization has been done using the graphical representation of data and information from all three datasets using visual elements such as charts and graphs. The data visualization has been done to simplify the complex data in an easy-to-understand manner that enables the lawmakers to make informed decisions and identify trends, patterns, and relationships of the crimes in Chicago that might not be clearly understood from raw data. Python libraries like Plotly helped us to create visualizations directly from PostgreSQL data.

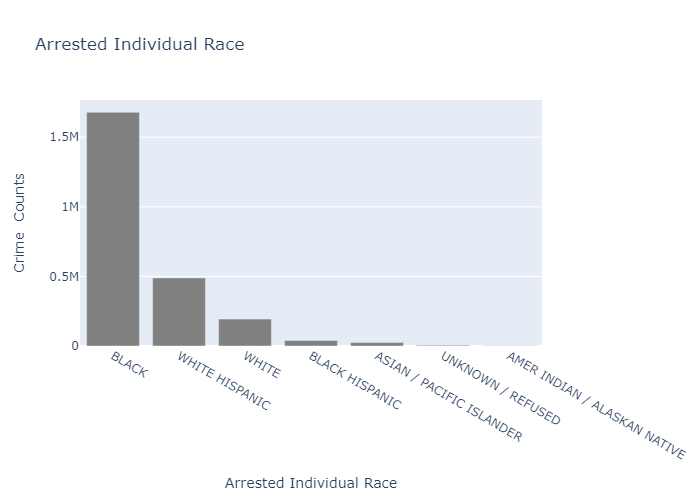


Fig 4.1: Number of Arrests based on Different Races.

The bar graph shown in fig 4.1 compares the number of arrests for specific Race categories in the city of Chicago. The x-axis displays the categories of Race (such as Black, white, white Hispanic, black Hispanic, Asian/pacific islander, American Indian/Alaskan native), while the y-axis represents the number of arrests in each category.

The graph allows easy comparison between the number of arrests and the Race of the convict. Black and Hispanics were the most common race among the individuals arrested in Chicago.

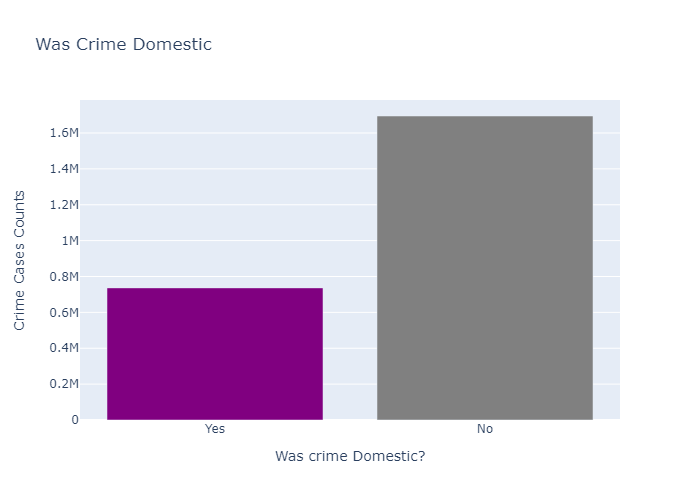


Fig 4.2: Frequency of domestic and non-domestic crimes.

In Fig 4.2 represents the frequency of domestic and non-domestic crimes in the city of Chicago. The x-axis displays the categories of crime, with one category indicating domestic crimes that is “Yes” and the other indicating non-domestic crimes as “No”. Here the y-axis represents the number of incidents. In legal contexts, domestic violence is depicted by purple colour. Almost 30 percent of the crimes committed in 2022 in Chicago  were Domestic crimes.

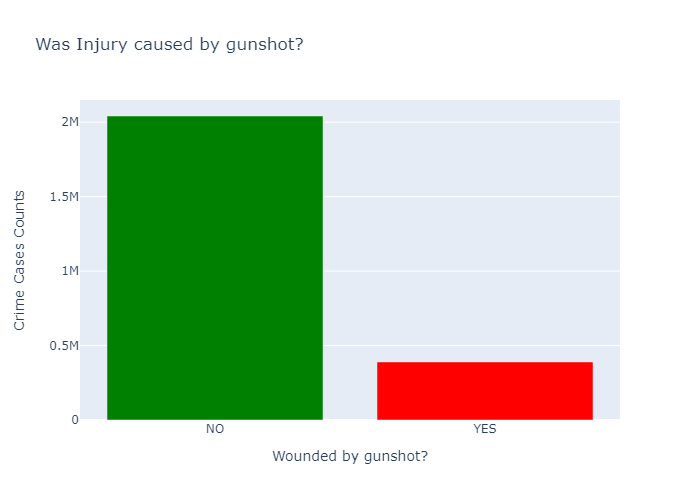


Fig 4.3 Representing the injury caused due to gunshots.

The above bar graph fig 4.3 is visually representing injury caused due to gunshots, with two bars representing 'yes' or 'no' for injury. The x-axis displays the categories 'yes' and 'no', while the y-axis represents the number of incidents. The graph shows the total number of incidents where injuries occurred due to gunshots and how many did not result in injuries. The height of each bar represents the number of incidents for each category, allowing for easy comparison between the 'yes' and 'no' categories. Almost 16 percent of the cases in Chicago involved a gunshot wound in 2022.

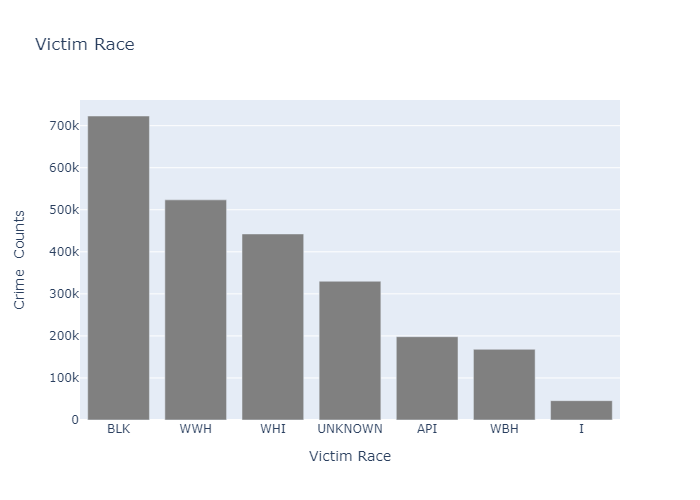


Fig 4.4: Number of Victims based on Different Races.

The bar graph shown in fig 4.4 compares the number of victims in specific Race categories in the city of Chicago. The x-axis displays the categories of Race (such as Black, white, white Hispanic, black Hispanic, Asian/pacific islander, American Indian/Alaskan native), while the y-axis represents the number of victims in each category. The graph allows easy comparison between the number of victims and the Race of the victims. Most of the victims in crime cases in Chicago in 2022 were Black.

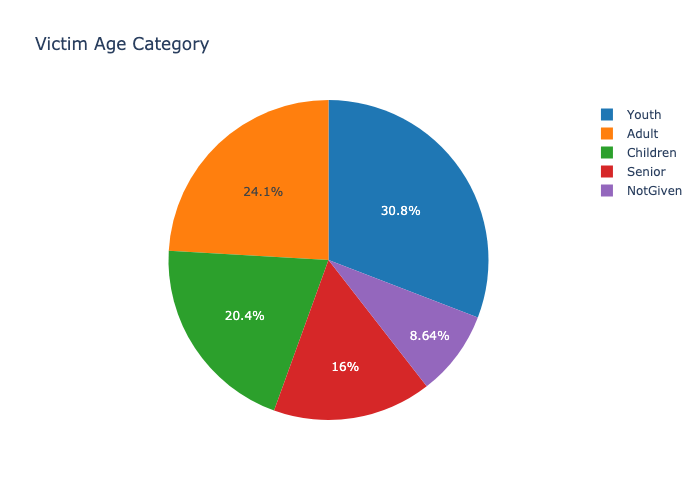


Fig 4.5: Representing the percentage of victims based on their age categories.

The pie chart shows us in visualization the differences in the percentage of victim age categories between New York and Los Angeles. The victims are categorised into youth, adult, children, and seniors different colours in the pie chart represent the victims belonging to various categories.

Most of the victims in Chicago fell under youth category in the year 2022.

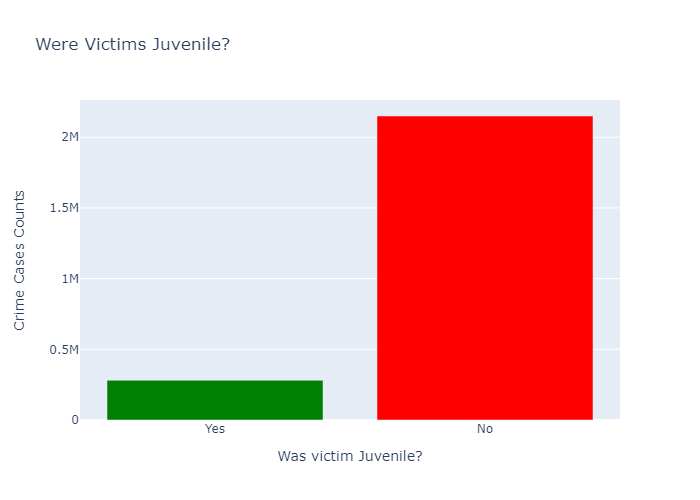


Fig 4.6: Representing whether the victim was Juvenile or Not.

The above bar graph fig 4.6 is visually representing whether the victim was a Juvenile or not, with two bars representing 'yes' or 'no' for juveniles. The x-axis displays the categories 'yes' and 'no', while the y-axis represents the number of incidents. The graph shows the total number of incidents where the victim is Juvenile. The height of each bar represents the number of incidents for each category, allowing for easy comparison between 'the yes' and 'no' categories. In legal contexts, juveniles are depicted by typically Green colour. In the year 2022, 12-16 percent of the victims were juvenile in Chicago.

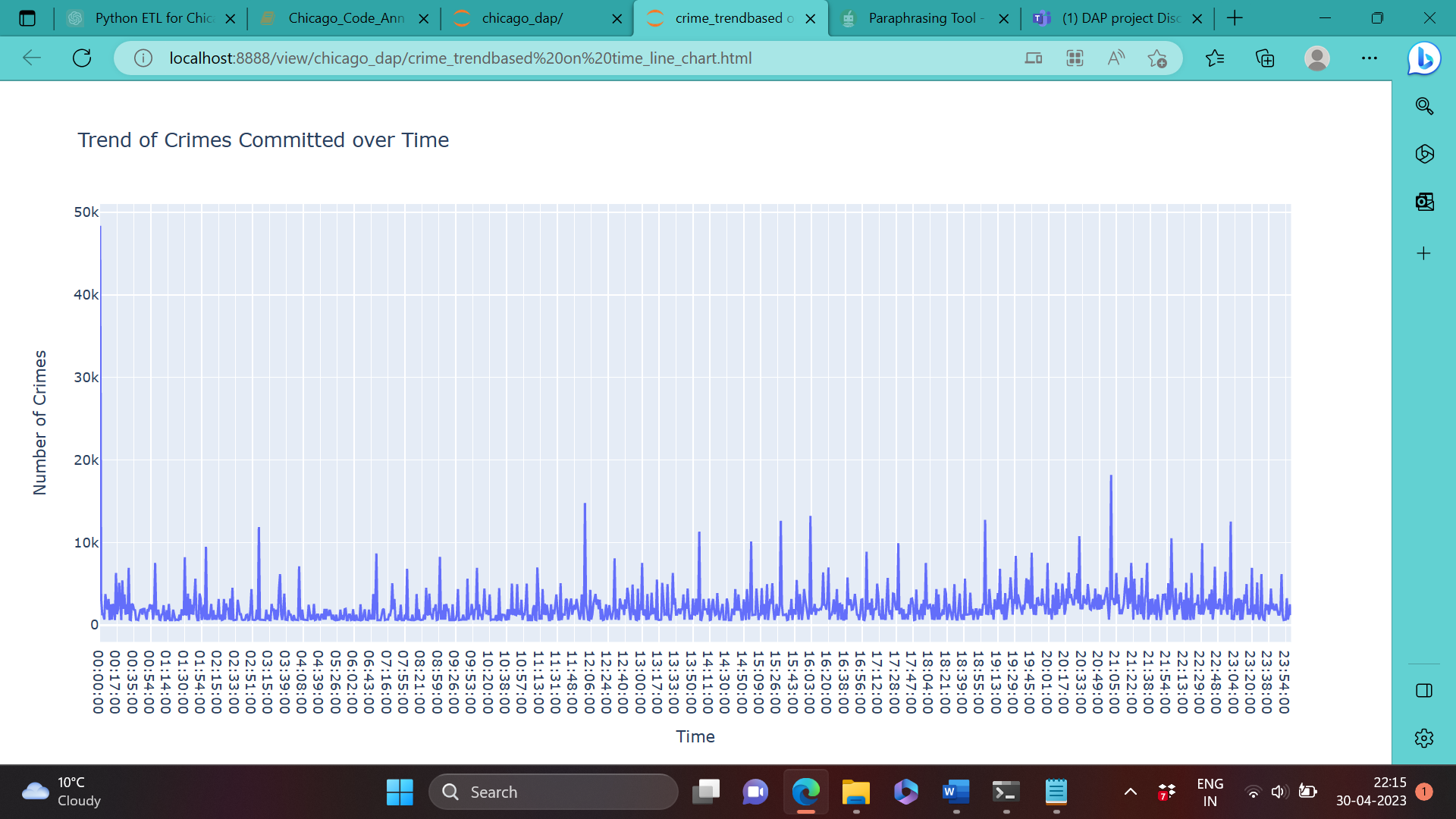


Fig 4.7: Trend on the crime committed over time.

The line chart Fig 4.7 represents the trend of crimes committed over time in a day. The line graph shows the patterns of crimes committed during a particular time period in a day and it is visible that the peak time for the increase in crime is at 21:05, where the x-axis indicates the time and the y-axis indicates the number of crimes committed during a particular time.

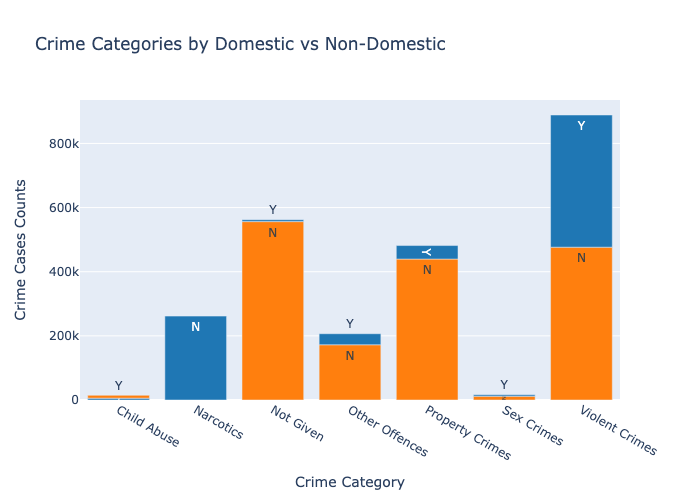


Fig 4.8: Crime Categories by Domestic vs non-domestic.

The above bar chart Fig 4.8 displays the distribution of different types of crimes, categorized by domestic and non-domestic settings. Each bar in the graph is divided into two i.e Yes or No based on the different crime categories. The x-axis represents the different types of crimes, while the y-axis represents the number of cases reported.

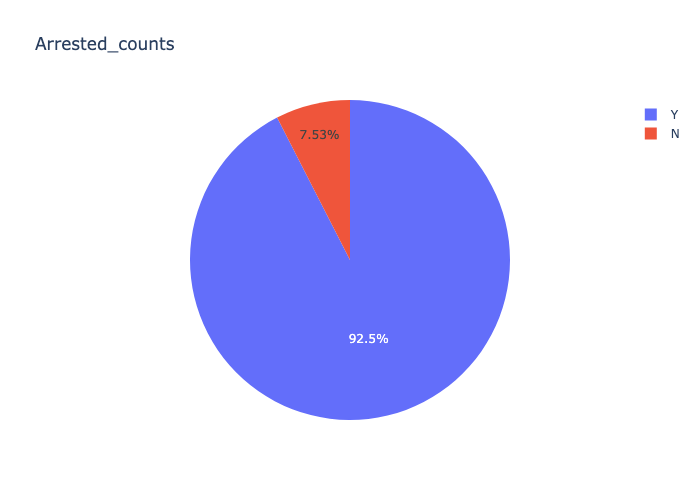


Fig 4.9: Arrests Made (Y/N)

The figure 4.9 gives us an idea of whether the person convicted or suspected of a crime was arrested or not. In 92.5 per cent of the cases, there were arrests made by the police department. No arrests were made in 7.5 per cent of the total cases. The police department should thrive to reduce this percentage so that the individuals committing crimes be brought to justice as soon as possible.

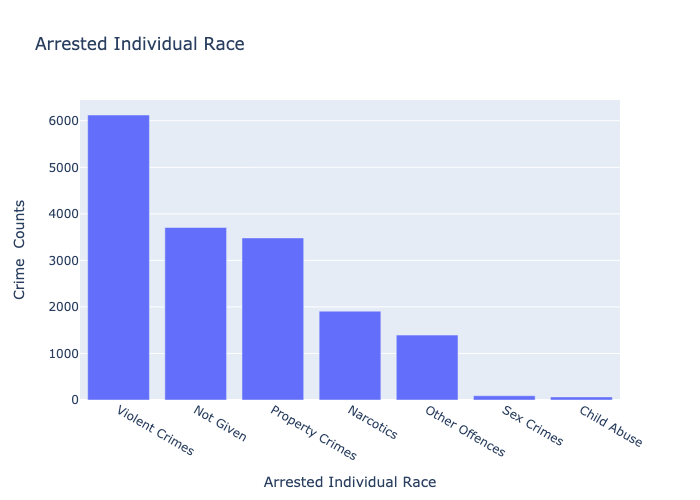


Fig 4.10 Representing Top Crime Categories

The bar graph represents the prevalent crime categories in Chicago. Violent crimes which include theft, robbery, assault had the highest frequency among all crime categories, while Sex crimes and child abuse had low counts.. The categories are listed on the vertical axis of the graph, while the horizontal axis displays the number of reported crimes. Each bar represents the number of incidents reported for a particular crime category.

**V. RESULTS AND EVALUATION**:

The results of this study have a number of consequences. Positive changes include the decline in reported crimes in Chicago over time. To lessen crime in the areas most impacted, more work must be done. The distribution of crimes in Chicago demonstrates the need for increased funding for the city's core and the hardest-hit neighbourhoods. Law enforcement agencies in Chicago can use the project's findings to guide their resource allocation and policy decisions.

The analysis of three Chicago crime datasets and the creation of visuals using Python's Plotly package were all successful outcomes of this project. The data can be utilized to guide policy decisions by the law since they offer insightful information about how crimes are distributed in Chicago.

**VI. CONCLUSION AND FUTURE WORK:**

***1. Conclusion:***

In conclusion, our Python data analytics project on Chicago crime patterns and trends has given us meaningful insights. We were able to determine the most typical crime kinds, the places with the greatest crime rates, and the times of day in the year 2022 when crimes are most likely to occur through the help of data visualization.

Additionally, our effort illustrated the potency of using Python and its numerous libraries for data analysis and visualization, including Pandas, and Plotly. With the use of these technologies, we were able to quickly evaluate big datasets and effectively communicate our conclusions.

This effort lays a strong platform for future investigation and study of Chicago's criminal activity. Policymakers and law enforcement can use these insights to guide choices and create plans to lower crime density and improve public safety in the city.

***2. Future Work:***

Although the Chicago Police Department's crime statistics were the focus of our project, there are other data sources that could offer insightful information on crime trends and patterns. For instance, including information on the climate, and demographics could give us an understanding of the underlying causes of crime.

By Increasing the geographic focus, our effort concentrated on crime in the city of Chicago, the analysis might also include other cities in the US., which can give more grasp of crime in the larger areas around Chicago by broadening the geographic reach.

Creating a system that provides up-to-date information on crime patterns and trends can play a vital role in guiding law enforcement operations and public safety initiatives. This system could be created by integrating crime data with real-time information on police activity and other factors.

**VII. BIBLIOGRAPHY**

[1]. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython by Wes McKinney

[2]. Data Science from Scratch: First Principles with Python by Joel Grus

[3]. Python Data Science Handbook: Essential Tools for Working with Data by Jake VanderPlas

[4]. Pandas Cookbook: Recipes for Scientific Computing, Time Series Analysis and Data Visualization using Python by Theodore Petrou

[5]. Hands-On Data Analysis with Pandas: Efficiently perform data collection, wrangling, analysis, and visualization using Python by Stefanie Molin

[6]. Machine Learning for Algorithmic Trading: Predictive models to extract signals from market and alternative data for systematic trading strategies with Python by Stefan Jansen

[7]. Applied Data Science with Python Specialization on Coursera (a series of courses) by University of Michigan

[8]. Data Analysis with Python: A Comprehensive Guide to Understanding Data Analytics Techniques with Python by Ivan Idris

[9]. Learning IPython for Interactive Computing and Data Visualization by Cyrille Rossant

[10]. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More by Matthew A. Russell

[11]. Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow by Sebastian Raschka

[12]. Data Wrangling with Python: Tips and Tools to Make Your Life Easier by Jacqueline Kazil and Katharine Jarmul

[13]. Data Visualization with Python and Matplotlib: Master 80+ interactive plots and learn techniques to display your data in compelling ways by Benjamin Root and Chadwick Boulay

[14]. Statistics for Data Science: Leverage the power of statistics for Data Analysis and Data Science by James D. Miller

[15]. Bayesian Analysis with Python: Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ by Osvaldo Martin

[16]. Applied Predictive Modeling by Max Kuhn and Kjell Johnson

[17]. Deep Learning with Python by François Chollet

[18]. Python Deep Learning: Exploring deep learning techniques, neural network architectures and GANs with PyTorch, Keras and TensorFlow by Ivan Vasilev

[19]. Exploratory Data Analysis with Pandas and Python 3.x by Dr. Tomasz Lelek

[20]. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists by Alice Zheng and Amanda Casari

**.**