HDL REPORT of [Violent-Crime-Data-Analysis-of-Montgomery-County-Area](https://github.com/BhargaviKalaparty/Violent-Crime-Data-Analysis-of-Montgomery-County-Area)

# 

……

# Abstract

With the advancement in the field of data science and analysis, the new data collection and analysis techniques are making remarkable progress in making human life easier and more predictable. Data set collected by FBI for Montgomery County of Maryland police department can be very helpful in understanding the trends and frequency of crimes and effectively reducing the casualties. For that extensive data analysis is required on the given data set. The goal of this research is to deduce effective data from this open-source data and implementing different techniques of Data analysis using Python Libraries. For this specific report group, NIBRS A category crimes are in under consideration and research questions are created in accordance with the same. All research questions are answered by suitable visualization techniques. Exploratory data analysis is performed on this data set after cleaning to extract maximum useful information and with the help of different visualization techniques many real-life problems can be managed with the help of this research report.

……………

# Table of Contents

1. [INTRODUCTION 1](#_bookmark0)
   1. [Problem statement 1](#_bookmark1)
   2. [Scope of Research 2](#_bookmark2)
   3. [Research Questions 2](#_bookmark3)
   4. [Workflow 2](#_bookmark4)
2. [PRELIMINARY DATA ANALYSIS 4](#_bookmark5)
   1. [Data Set 4](#_bookmark6)

[Dataset Description 4](#_bookmark7)

* 1. [Programming Language and Tools 6](#_bookmark8)

[Required Tools for the Implementation 7](#_bookmark9)

[Proposed Implementation Steps: 7](#_bookmark10)

* 1. [Data Quality Initial Assessment 7](#_bookmark11)

[Data Cleaning: 7](#_bookmark12)

[Data Reduction 8](#_bookmark13)

[Data Integration 8](#_bookmark14)

[Data Imputation 8](#_bookmark15)

[Data Discretization 8](#_bookmark16)

1. [EXPLORATORY DATA ANALYSIS 9](#_bookmark17)
   1. [Introduction 9](#_bookmark18)
   2. [Descriptive Statistics: 9](#_bookmark19)
   3. [Data Visualization 11](#_bookmark20)

[Implementation 13](#_bookmark21)

[Research Question 13](#_bookmark22)

[Research Question 14](#_bookmark23)

[Research Question 14](#_bookmark24)

1. [CONCLUSION 16](#_bookmark25)
2. [REFERENCES 17](#_bookmark26)

**iv**

……………

# List of Figures and Tables

|  |  |
| --- | --- |
| Figure 1 | Work flow of the course work |
| Figure 2 | Data Types representation using Python |
| Figure 3 | Implementation Steps |
| Figure 4a | Crime Category analysis using Pie Chart |
| Figure 4b | Crime Analysis based on NIBRS Code - Pie Chart |
| Figure 5 | Crime Occurrence by Police District - Bar chart |
| Figure 6 | Heat Map |
| Table 1 | Data types in the data set |

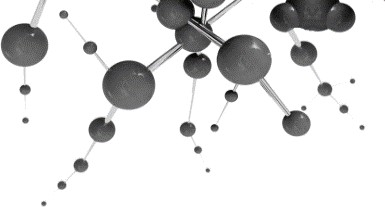
**v**

……………

# List of Abbreviations

|  |  |  |
| --- | --- | --- |
|  |  |  |
| NIBRS | National Incident based reporting system |
| ID | Identification |
| CR | Crime Rate |
| UCR | Uniform Crime Reporting |
| EDA | Exploratory Data Analysis |
| CIJS | Criminal Justice Information Services |
| PD | Police Department |
|  |  |

**vi**



**1**

**Introduction**

* 1. **PROBLEM STATEMENT**

As time goes, there always will be an increase in population and crime rate, hence, the crime data will be ever increasing. In a way to accommodate this FBI came up with an efficient way to segregate and maintain the quality of the crime data. NIBRS captures all these details and provides law enforcement agencies with a standardized, electronic blueprint for storing the NIBRS data within their individual records management systems. The goal of NIBRS is to enhance the quantity, quality, and timeliness of crime data collection and to improve the methodology used in compiling, analysing, auditing, and publishing the collected crime statistics. There are two categories of offenses reported in the NIBRS: Group A and Group B. Group A defines violent crimes out of the two. In the crime data, the NIBRS codes are mentioned but they aren’t in any order. The rates of violent crime from NIBRS were on average less than 1% higher than rates from data defined by Summary UCR. NIBRS provides individual records of incidents with details on offense, offender, Victim, and property. With the help of data analysis and all these details we can achieve viable inferences on crime analysis. Number of criteria such as seriousness, frequency, prevalence, probability etc. are used to categorize the crime data into Group A and B. [3], [7]

The data set provided is crime data of Montgomery County of Mary land. This is an open-source database which can be accessed by the public at any point of time, and it keeps updating every day. This data is time lined from 1st July 2016. In the crime data of Montgomery County of Maryland, the NIBRS codes are mentioned, however, they aren’t categorized in any order. [10],[11]

The variables in the dataset are both quantitative and qualitative in nature. Quantitative data contains numeric values, and these are further divided into continuous and discrete. Continuous values are the ones which can be fractional, on the other hand, discrete values are only finite. In the given data set, the columns such as Incident ID, offense code and CR number contain discrete values and the columns such as longitude and latitude contain continuous values. Coming

to qualitative data, it can be defined as a non- numeric data. This kind of data is further categorized to ordinal and nominal, ordinal having values which contain specific order like NIBRS code in the given data and nominal doesn’t have any specific order to it, for instance, columns such as Crime Name 1,2,3, Police District, City and State. [8]

* 1. **SCOPE OF RESEARCH**

The main goal of this research is to deeply analyze and understand the trends of this data set and increasing the credibility of the data set by implementing different statistical methods for analyzing and auditing the data set and highlighting the useful crime statistics. The provided data is an ongoing research and is being updated regularly. Initial data analysis and EDA is performed on the data set to answer the research questions.

* 1. **RESEARCH QUESTIONS**

Following research questions are created after initial analysis of the data set and comparing them with real life problems and their solutions.

* + 1. What are the percentages of occurrence of each crime type and NIBRS A category crime data. (Crime type: ex: shoplifting, theft from building etc)?
    2. What is the graphical representation of which place has the most crime rate?
    3. What is the frequency of crime in different police districts (which PD name has the most and which has the least number of cases)?
    4. Map representation of crime name in Montgomery area using latitude and longitude values from crime data?
    5. What is the monthly analysis of crime data according to the crime category? (what months has more crime rate than the others.)
    6. Analyzing of crime data across the years 2016-2022 based on offense code.
    7. Identifying the time of occurrence of crime in a day( Morning/Evening)?
    8. Which Street type (AL, AVE, BLV etc.) is having greater number of crimes?
    9. Analyze and visualize which police agency has more number of Violent crimes ?
    10. What is the average number of Victims for the crime data?. Visualize with average count, NIBRS code and the description of type of crime associated to the code.
  1. **WORKFLOW**

The following workflow represents the working of this research project. First step is to analyze the crime data set then designing the research questions based on the analysis and then data cleaning and removal of irrelevant data and null values

to clean data, for this assisgnment we are taking a subset of the data and making relevant visualisations.

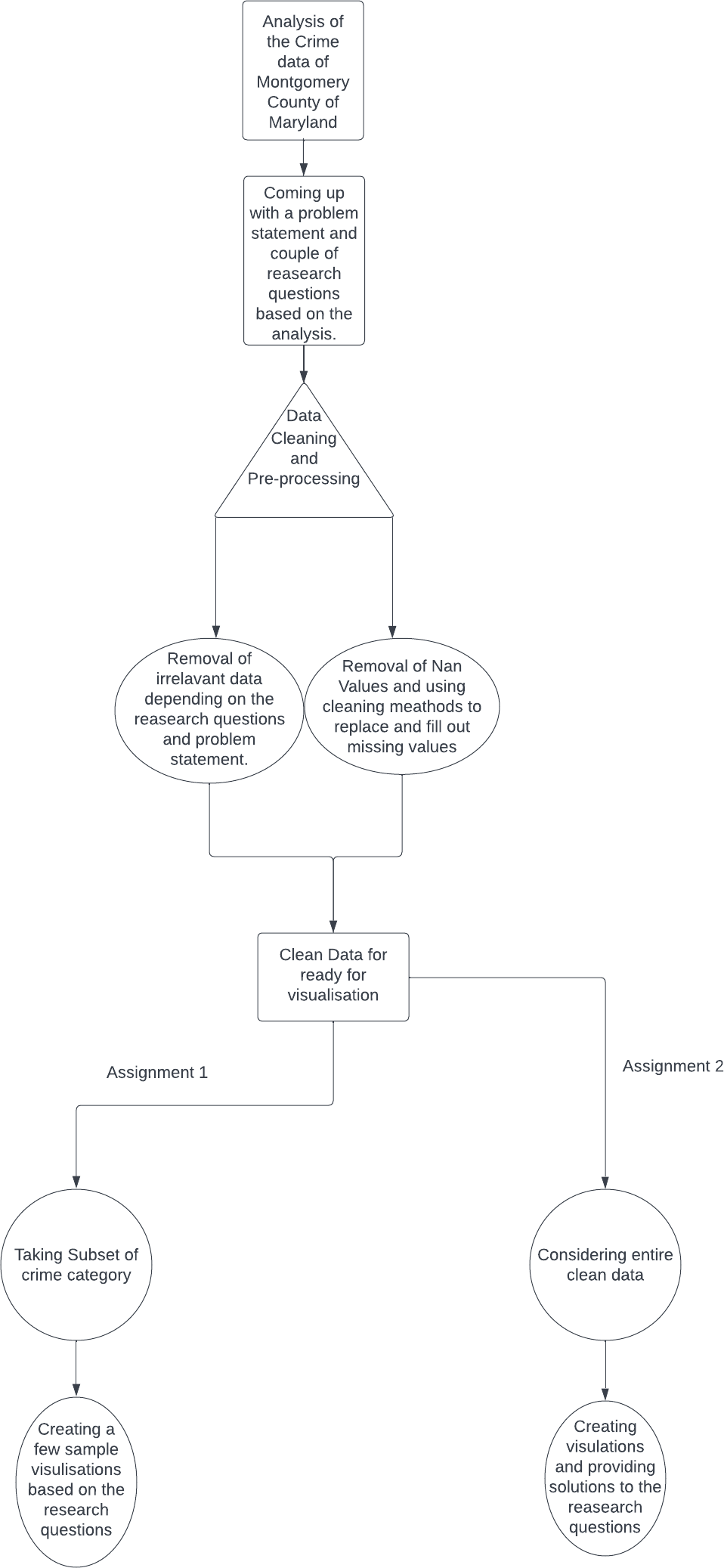
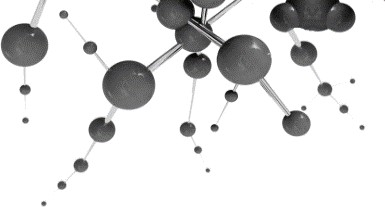


Figure1:Workflow



**2**

**Preliminary Data Analysis**

* 1. **DATA SET**

The data given was obtained from authorized police incident reports and reported offences that were categorized utilizing the UCR Program of the Criminal Justice Information Services (CJIS) Division. The NIBRS. The data is structured data (CSV). The dataset has 30 attributes, with the offence type and its category—such as robbery, public disturbance, and sexual assault—being important features. Along with the district, locality, and exact time the offence happened, the dataset consists of the following attributes: [1]

## Dataset Description:

Following are the attributes that are part of our Dataset.

|  |  |
| --- | --- |
| **Attribute** | **Type** |
| Incident ID | numeric |
| Offence Code | String |
| CR Number | numeric |
| Dispatch Date / Time | String |
| NIBRS Code | String |
| Victims | numeric |
| Crime Name1 | String |
| Crime Name2 | String |
| Crime Name3 | String |
| Police District Name | String |
| Block Address | String |
| City | String |
| State | String |
| Zip Code | Float |
| Agency | String |
| Place | String |
| Sector | String |

|  |  |
| --- | --- |
| Beat | String |
| PRA | String |
| Address Number | Float |
| Street Prefix | String |
| Street Name | String |
| Street Suffix | String |
| Street Type | String |
| Start\_Date\_Time | String |
| End\_Date\_Time | String |
| Latitude | Float |
| Longitude | Float |
| Police District Number | String |
| Location | String |

Table 1: Data Types in the data set

Python has standard built in datatypes. String means a collection of characters/ words this is the most frequent data type in the this particular crime data. Float is nothing but a floating point variable and numeric has 3 types in it: float, integer (just the digit without float). [5]

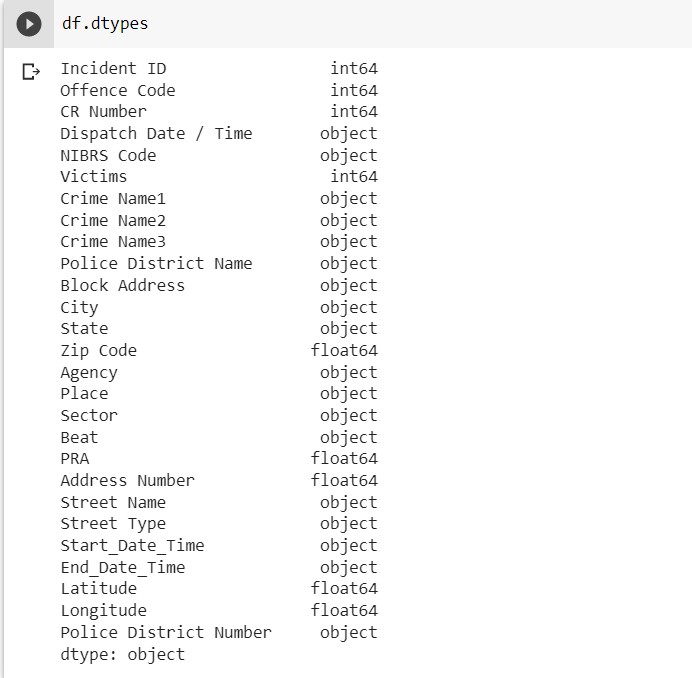


Figure 2: Data Types representation using Python

* 1. **PROGRAMMING LANGUAGE AND TOOLS**

Our Analysis uses various data mining techniques to analyze the data e.g., Correlation which can be used to find the relationship between the different attribute’s e.g., Crime rates, district and reporting date and time. A low correlation suggests that the variables are not actually connected to one another, whereas a high correlation suggests that at least two elements have a strong association with one another.

To carry out data analysis, Different analysis tools will be used to perform statistical analysis and data mining, Python is a highly utilized tool, it’s an open- source object-oriented Programming Language. Python is used to analyze, manipulate, clean and crunch the data and to examine the correlation and regression.

Lastly some python libraries is used to import and visualize the data e.g., pandas to import the data, NumPy to perform statistical analysis and matplotlib and Seaborn to draw the graph and for better visual analysis of dataset.

## Required Tools for the Implementation:

* + - Anaconda/ Google Co-LAB
    - Windows
    - Python
    - Excel
    - Python Libraries (Pandas, Numpy, Matplotlib, Seaborn )

## Proposed Implementation Steps:

Following figure represents steps of analysis using tools and programing languages

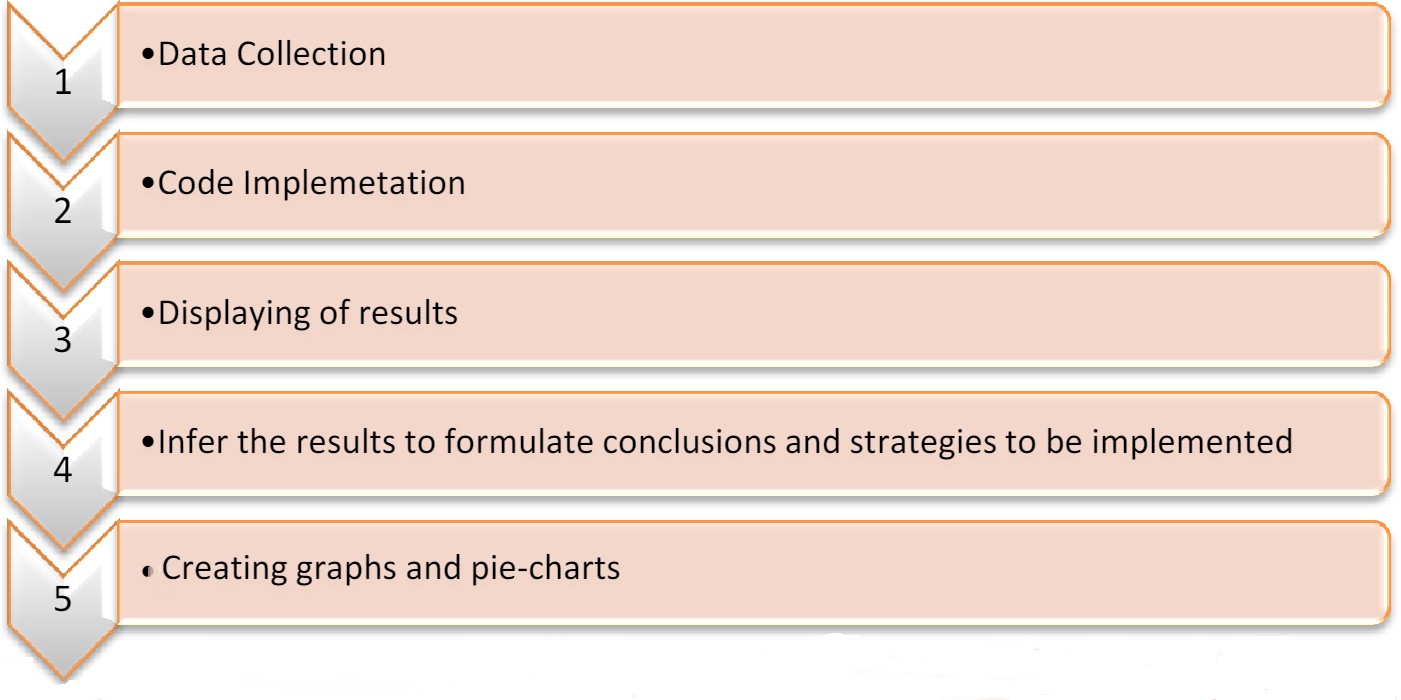


Figure 3: Implementation Steps

* 1. **DATA QUALITY INITIAL ASSESSMENT**

We are certain that finding the linkage between criminal activities can greatly aid in the detection of problematic hotspot. As a result, our strategical approach concentrates on three key factors of the dataset i.e., type, location and time of occurrence of the crime. Based on these characteristics we attempted to identify any potentially noteworthy frequent patterns. [12]

## Data Cleaning:

After the initial overview and observation, we have discovered that there are lot of missing values in the dataset e.g., Crime occurrence Date, location, address, Dispatch Date / Time etc. but these are not part of primary/important attributes of dataset and as a result we conclude that we don’t need to clean them. All the

key attributes are there with clear and complete values. Furthermore, we noticed no noisy or inconsistent values in these features. [6], [14]

## Data Reduction:

Another approach for data cleaning is dataset reduction where non-important attributes are removed from the dataset. For example, in the given dataset among the 30 attributes, some attributes having duplicate values e.g., Location and latitude and longitude columns having same values, similarly street prefix, street suffix and other columns that are not part of research questions. All the irrelevant characteristics will be removed, and only relevant data will be used for data mining and visualization purposes. On the other hand, we can also reduce the data from the columns as well e.g., we are just focusing on group of crimes i.e., Shoplifting, mobile snatching, etc. we can group the data accordingly and then remove all the other columns accordingly.

## Data Integration:

For this dataset, we can use multiple data integration techniques, first is to harmonize the key attributes name i.e., naming convention for the key attributes Crime Name1, Crime Name2, Crime Name3 to crime category, type and description. Our research requires to examine of date and time attributes to display the correlation between the crimes and months or years to represent the data in the form of scatter graph. Therefore, we used Dispatch Date / Time field to generate three new properties i.e., Crime Day, Month, and year.

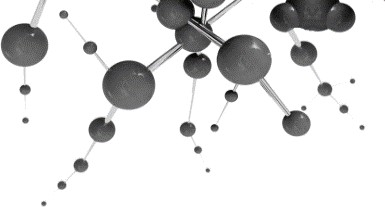
Another approach is to use military time system where we can generate another attribute Crime Hour without focusing on minutes to find the appropriate patterns. We can use the same approach for the field Start\_Date\_Time depending on the data and requirements.

## Data Imputation:

Data Imputation is also a technique used to fill the missing values in the dataset. This technique is very useful to fill the missing values or null values in the data set specially when the dataset is very huge. There are several techniques of data imputation like mean, mode, median. The implementation is discussed in the section 3.

## Data Discretization:

Discretization is very useful when working with large dataset and data is too noisy. Once Data integration is done, we can apply data transformation on crime time and crime type attributes to minimize the diversity of the dataset. We can apply data discretization on smaller groups to get more frequent patterns.



**3**

**Exploratory Data Analysis**

* 1. **INTRODUCTION:**

Exploratory data analysis (EDA) is a technique used to explore, examine, and research datasets using various statistical and visualization approaches. Data cleaning is required to analyze trends and patterns of crimes committed. There are four reasons to further clean this data [2],[4]

* + - Identify Missing information
    - Querying and Indexing
    - Identify Irrelevant data and features
    - Recognizing Inconsistent data and outliers
    - Discover patterns
  1. **DESCRIPTIVE STATISTICS:**

There are different types of exploratory data analysis methods the most suitable for this type of data set is using Multivariate graphical EDA Techniques which allows data to be examined and shows relationship between data attributes. Similarly, there are many different statical method which can be used to understand data sets and extract maximum information from the data set. Methods used for this report are

* + - Sample Size Determination
    - Mode
    - Mean

Statistical method – ‘mean’ allows data to understand trend with respect to different attributes in the data set. Mode works around the data set and infers the most repeated value/ frequent. Sampling can be used to extract some data from the whole data set and then analyze its relationship with other aspects in the data set. Python libraries (NumPy, Pandas) simple functions form mean averaging, mode and sampling will be used to analyze this data set accordingly. [13]

For pre-processing some columns are dropped which are not used in specified research questions such as Street Prefix, Street Suffix, location etc. The data set provided has a combination of data which needs to be put through couple of

statistical methods to make useful inferences through visualization. This section will explain the statistical methods used for each research.

The first research question will be looking at percentage of occurrence of crimes by crime category and NIBRS code. For this, we need to use the pandas library and apply the **cd.info ()** function to understand the information of each column in the data, especially crime Name 1 which is renamed as crime category and NIBRS code**. cd.isna (). sum()** to get the number of missing values in each category. Crime Category has 60 missing values and Crime type and Crime Description columns are correlated to each other, and it is checked that the missing values for all three columns are on the same index, mode can’t be used so dropped them. Next, NIBRS code has no null values so the entire dataset for this question is ready for visualization.

The second research question will be looking at the place of occurrence of crime and their count. This question will be using the column ‘Place’ to make the necessary inferences. The place column has almost 95 categories of places in which the crimes took place. Some categories also had a subcategory. For example, the place ‘Street’ has further sub columns such as ‘Street- Alley’, ‘Street- Bus stop’, ‘Street-commercial’, ‘Street - in vehicle’, ‘Street-other’ and ‘Street -residential’. All of them are separated by a ‘– ‘and they can be separated by using **split()** function to access the initial part of string that is ‘street’ in this case which is enough to get an understanding of where the crime is happening. The visualization can now be performed on the data which is shorter and necessary.

The next question is on the police district column, **isna().sum()** gave that there are about 62 null values in the set. The **value\_counts()** function will give the ascending order of most repeated police district names. The **mode ()[0]** command in Pandas library will give the first category name of most frequent names. Mode in python statistical analysis is a function which categorizes the most appeared values in the set of data. Now, the nan values are imputed with most frequent one. Hence, the column is now ready for visualization.

The other question is on the monthly analysis of crime data this requires an analysis into Dispatch date. The dispatch date consists of couple of null values which are similar to values in start date, hence, by using copy function we fill in the available values. Then with the help of **split function**, the date is further separated, here we consider the first value which is the month and delete the rest. Now the date has all month values, which can be visualised according to the research question.

The next question will focus on crime occurrences during which part of the day (morning or evening). This will be using the time column which was initially split from date. Using **if condition** in NumPy, values less than 12 considered as morning, values between 12-5 as afternoon and after 5 as night. The data after categorizing will now be ready for visualisation.

Eighth research question will focus on which street type has more number of crimes and this would be using the street type column. For the nan values in street type, similar method police district name nan values is used. Finding out the ascending order of most repeated street types by using **value\_counts()** function and **mode()** concept to impute the nan data.

Last research question has the Statistics part in visualization rather than cleaning. This research question will be analyzing the average number of victims which is nothing but mean and will be using estimator as np.mean to achieve the result. [9]

* 1. **DATA VISUALIZATION**

Data Visualization is the process of graphical representation of given data to make valid interpretations to understand the data correctly. Python libraries such as Matplotlib and Seaborn can be used to perform these visualizations. [15]

There are many techniques to visualize data. Some of them are as follows:

**Scatterplot:** This type of plot is usually used to represent correlations.

**Histogram:** This is used to show distribution of the variable.

**Bar chart:** This is used for representing categorical data with help of vertical and horizontal bars.

**Pie Chart:** This is again used to represent categorical data but in divided proportions and percentages.

**Count plot:** This is similar to bar chart but the y-axis represents the number of occurrences of variable represented in the x-axis.

**Boxplot:** This is a quartile distributed representation of data. It can compare number of variable and also the level of the variable.

**Heat map:** This is a matrix plot which will plot the data into color coded matrices.

**Dis plot:** Similar to bar graph but represents the distribution of single variant data.

**Joint plot:** As the name suggests this is a combination of Scatter and Displot, used to compare the distribution of one variable to the other(Scatter to Displot).

**Heatmap:** This visualizes the distribution and clustering of the scattered plot variable values.

Research questions are designed in a way to analyze and extract maximum information from the data set which can help in identifying pattern and reduction of overall crime rate. Some of the research questions need representation in the form of bar chart, some in the form of pie chart, heat map and simple graphs to identify plots.

The first research question is representation of the crime occurrence based on NIBRS A category code and type. The type of visualization chosen for this is a pie chart. The reason behind the choice is that pie chart represents categorical data, the two variables we are considering for plotting are categorical and we need to represent the percentage of occurrence which is best done with splitting the data into categories and showing the percentages. Pie chart can be plotted by using Matplotlib library in Python.

The second research question asks for a graphical representation of places where occurrence of crime is more. For this representation, Displot will do justice as it is displays univariate data (Place) in bars and also give the count on the y-axis.

The third research question is on the representation which will help make inferences on which police district has more number of crimes. For this representation Displot is chosen as the data again is univariate and will give the count of repetition of every police district mentioned in the data.

The fourth research question is map representation of crime data. To justify this, a scattered plot with heat map can be used as it not only gives the location but also clusters of crime occurrences to understand which region has the most.

The fifth question is on monthly analysis of crimes according to the crime category, this can be achieved with using a simple bar graph. Since, both the variables are categorical and allows aggregation, bar graph would be the best choice for this.

The sixth question is based on the analysis of crime data with offense code over a time line of 2016-2022. This can be best achieved by a scattered plot as the data is bivariate and numerical and the scattered dots help in categorizing which year had most offenses committed.

The seventh question is an analysis on the what time of day (ex: morning or night) has most crime rate. This can be easily achieved by using count plot. As the data has single variable and it also provides the count (repetition) on the y-axis. Hence it makes the analysis easy as we can also look at the number of crimes associated with each time of the day.

The eight question is on the street type, which has more occurrences if crimes. This can be plotted using a displot, this will do justice as it is displays univariate data in bars and also give the count (occurrences) on the y-axis.

The nineth question will focus on which police agency has the most number of crimes. This can be achieved by representing the data on the hisplot with KDE as it is single variant and gives density as well.

In the last question we are looking at average victims for the crime data with crime category. Since this has 3 variable, we will be representing the data with a barplot and a hue.

## Implementation:

For all the below implementations we have considered only partial data (offence code 23A-F).

## Research Question:

The first question is to plot a pie chart to represent the percentages of occurrence of each crime type and NIBRS A category crime, we plotted two pie charts: (5a) representing the crime type and (5b) representing NIBRS code.

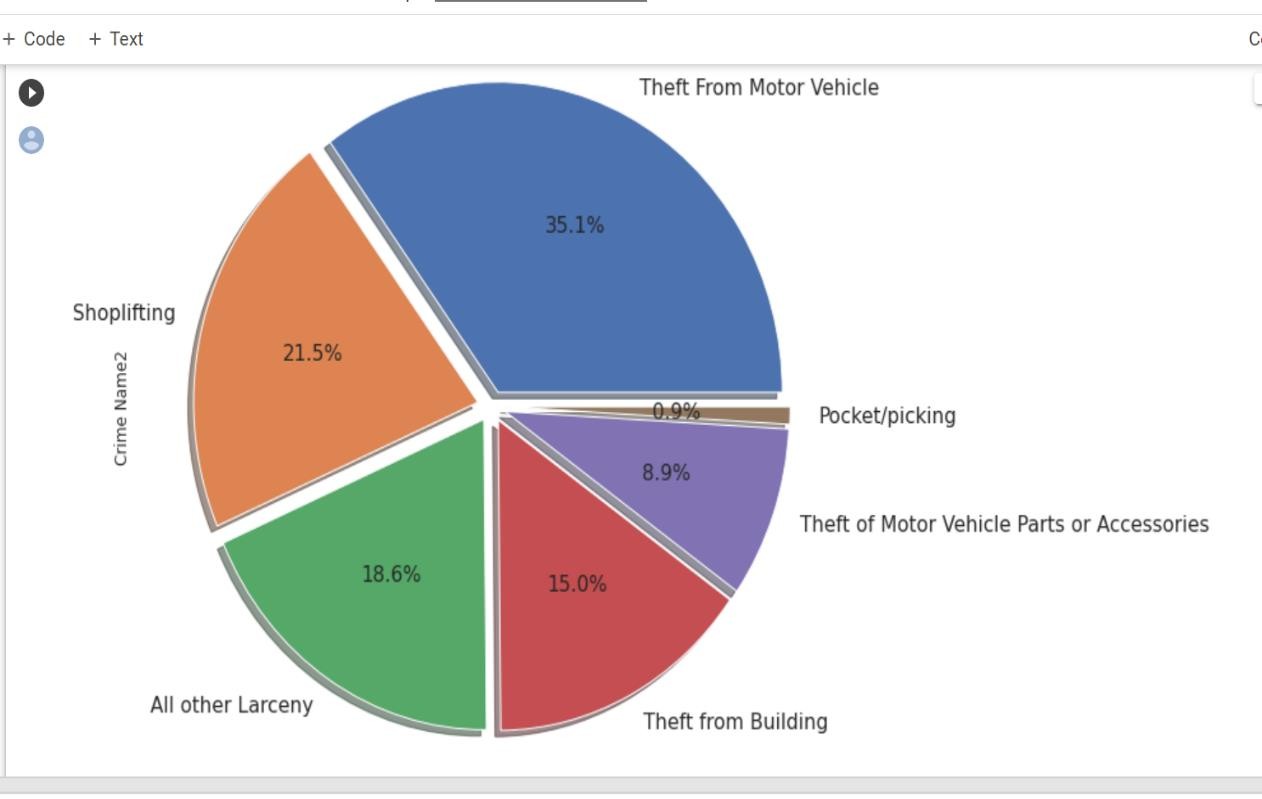


Figure 4a: Crime category analysis using pie chart

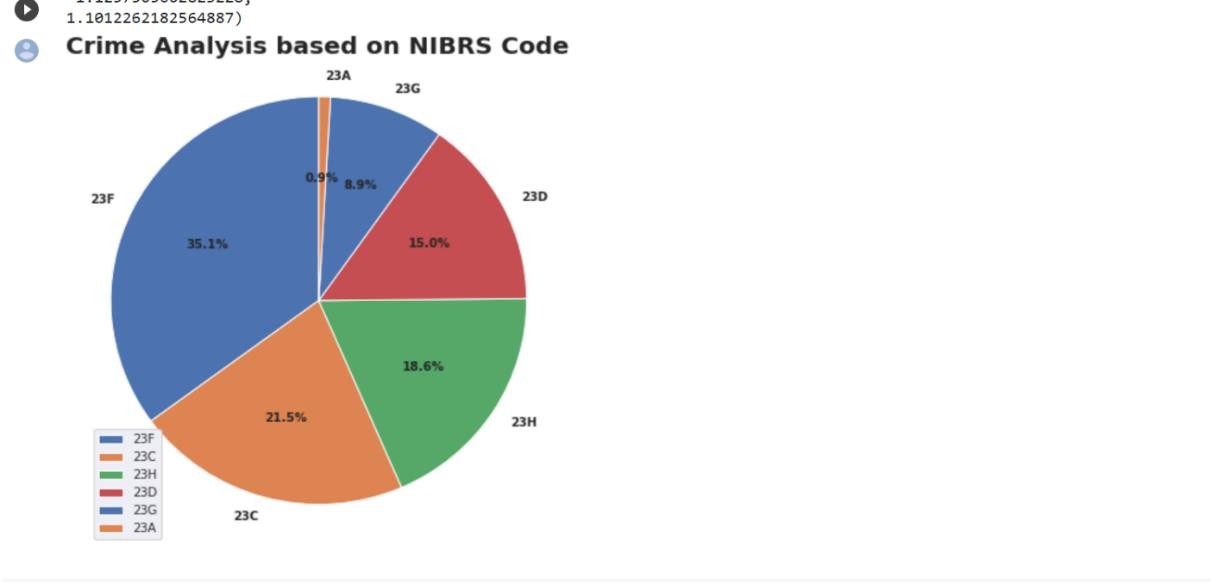


Figure 4b: Crime Analysis based on NIBRS Code

## Research Question:

The second research question is on the representation of a bargraph to understand which police district has what ratio of crimes. Bar can be plotted using Seaborn library. We have considered X label as police district y as crimes which denotes the count and set the bar color to maroon, added title, dpi and font sizes. Below is the visualization.

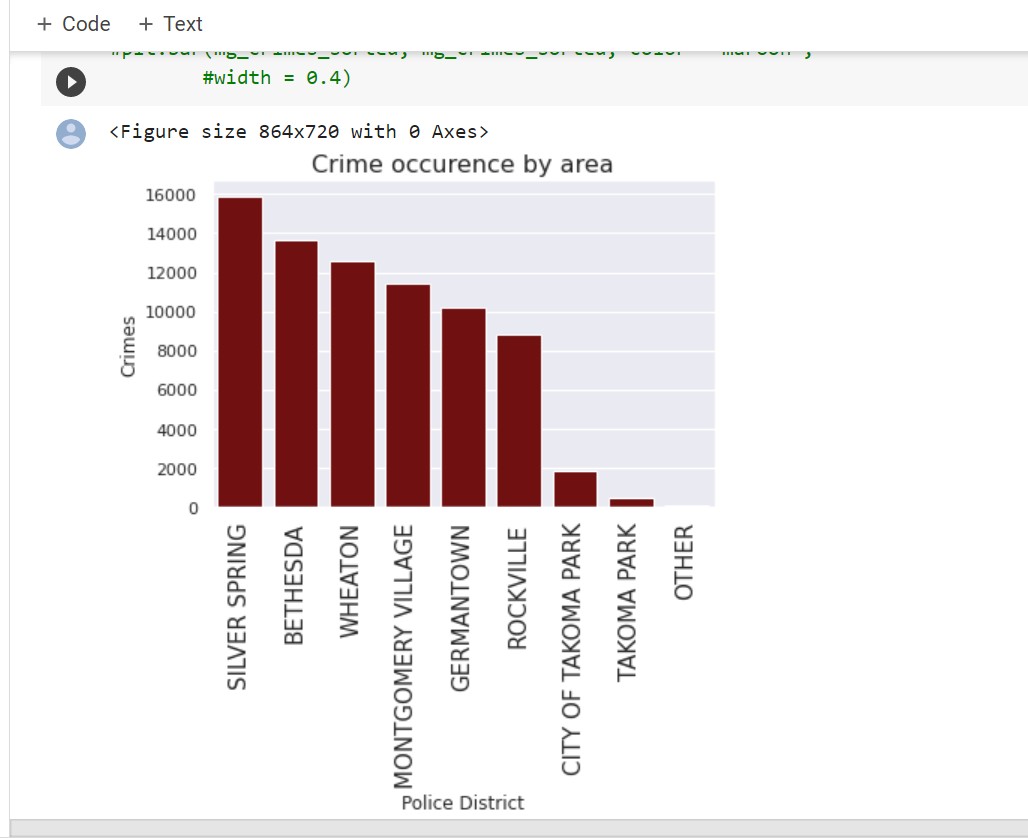


Figure 5: Crime Occurrence Police District Bar chart

## Research Question:

For the fourth research questions, we are considering longitude and latitude variables plotted a scatter graph with a heat map to understand the hot spots of crimes in the area. This can be done using Seaborn. Below is the representation.

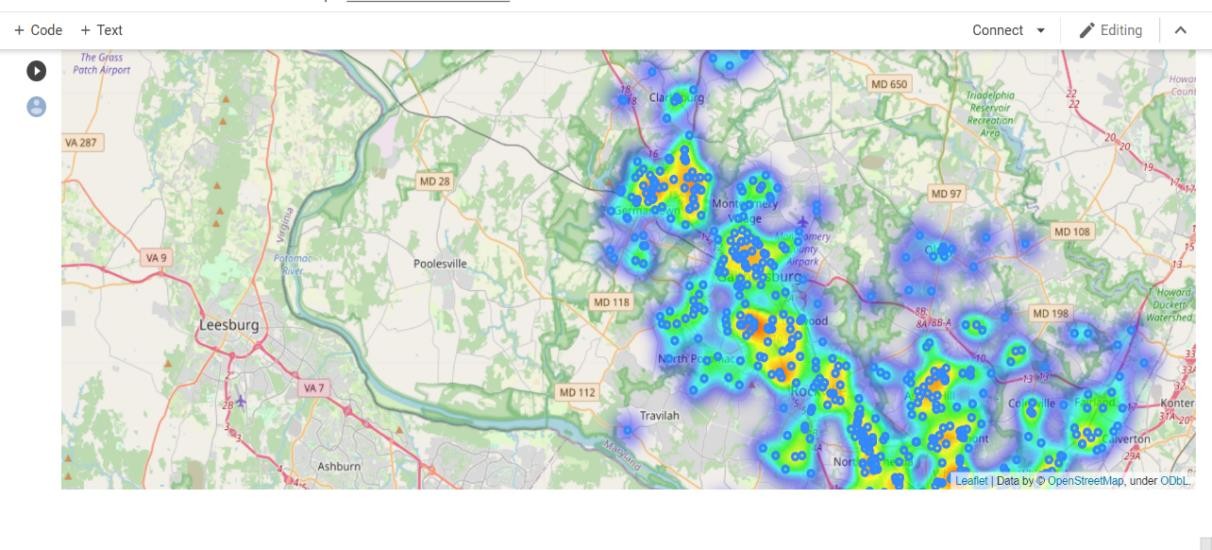
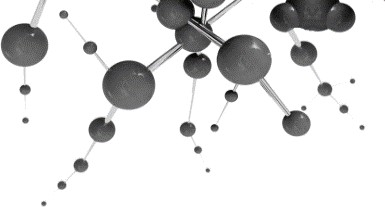


Figure 6: Heat map



**4**

**Conclusion:**

In conclusion, the above used research questions and visualization techniques of Violent crime category (NIBRS A) will help the law enforcement agencies to recognize patterns and take precautions. For example, with the second visualization of police districts, its easier to understand that most crimes occurred in Silver Spring PD and actions like employing more staff here can make a difference in crime rate.