Montgomery County Crime Analysis

Research Report (Coursework)
CI7340 Applied Data Programming
Kingston University

Nazanin Zakizadeh K2144935

Sayda Rabia Altaf K2164077

Jeevan Mohan Pawar K2242210

Kantheti Dhana Shravya K2247431

Table of Contents

Abstract— **For each city's police force gathers a massive amount of information each year and keeps on various crimes that occur in various locations. This dataset has been provided by the National Incident-Based Reporting System (NIBRS) for this research project. This research aims to discuss the Preliminary Data Analysis and Exploratory Data Analysis (EDA) methods and purposes and cover the concept of data science from statistical analysis and visualisation. Through this analysis, we can discover some fascinating details and trends in crime in Montgomery County (Maryland, USA). These results will help police departments and agencies of law enforcement better comprehend crime issues and provide insights that will allow them to better track activities, estimate the likelihood of accidents, effectively deploy resources, and improve decision-making.**

Keywords*— *Exploratory Data Analysis, Preliminary Data Analysis, NIBRS, Crime, Statistical analysis, Visualisation

I. Introduction

In this research project, a dataset regarding the crimes reported in the Montgomery County of Maryland State, USA needs to be analysed to gather useful findings for the government. This data covers 2016-2022 and is obtained from the National Incident-Based Reporting System (NIBRS) [1].

Starting from the preliminary data analysis phase, we will define the research problem and related questions. In this part, we will use a wide variety of tools to implement the initial data quality assessment. After that, we will discuss different approaches to exploratory data analysis (EDA), from statistical methods to visualisation. The workflow of our research is shown in the Fig. 1.

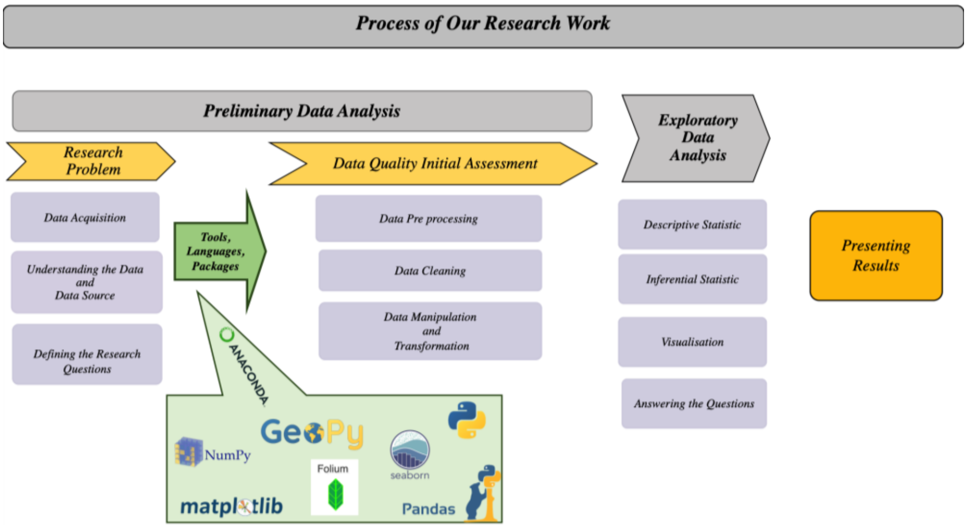

Process of Our Research Work
Preliminary Data Analysis
Research Problem
Data Acquisition
Understanding the Data and Data Source
Defining the Research Questions
Tools, Languages, Packages
Data Quality Initial Assessment
Data Pre processing
Data Cleaning
Data Manipulation and Transformation
ANACONDA
GeoPy
NumPy
Folium
seaborn
matplotlib
Pandas
Exploratory Data Analysis
Descriptive Statistic
Inferential Statistic
Visualisation
Answering the Questions
Presenting Results

Fig 1. RESEARCH PROJECT WORKFLOW

The hidden insights and patterns discovered in this study can help the authority as well as the police to coordinate and take decisions in their fight against crime. Furthermore, the research will also shed some light on the shortcomings in the quality of the dataset available at the moment and suggest ways for its improvement. A similar approach can be subsequently implemented in other counties across the nation and assist in reducing the overall crime.

This research aims to explain methods and required processes to find effective patterns and reveal critical information from the data. To have a better understanding of the problem statement, the following research questions are provided, which are going to be solved in this project:

1. When do criminals often operate? Which months and days are the busiest? Are there any patterns for these crimes in terms of specific days of the month? Is there a certain month that has the highest or the least criminal activity?
2. Which crime type has the most affected victims?
3. For each type of crime, which places are considered the most dangerous/safest?
4. Which police district deals with the highest and lowest number of crimes every year?
5. Which police departments are quickest to respond?
6. What is the frequency of each crime based on a year?
7. Which time of the day is the most active for criminals in the county?
8. What are the hotspot locations for criminal activities in Montgomery County?
9. What is the relationship between the weekdays and the number of crimes?
10. Which cities are considered the safest and the most dangerous ones in terms of the number of crimes?

II. Preliminary Data Analysis

A. *Dataset*

The dataset represents the crimes reported between Jan 2016 and Aug 2022 in the Montgomery County of Maryland State, USA. The dataset is available in a CSV file, the reported crimes are classified according to the National Incident-Based Reporting System (NIBRS) of the Criminal Justice Information Services (CJIS) Division Uniform Crime Reporting (UCR) Program and documented by approved police incident reports [2].

Montgomery County's open data website, data Montgomery, is the main source of this dataset and provides the public with direct access to crime statistic databases - including raw data and search functions - of reported county crime. The entered data follows Uniform Crime Reporting (UCR) rules where multiple offenses may appear as part of a single reported incident, and each offense may have multiple victims [2].

It is a structured dataset consisting of a total of 306,094 entries (rows) each with a total of 30 features (columns). The dataset also includes fields that are empty or that contain invalid data. A summary of these features is represented in the following table:

Table 1. DATASET SUMMARY

Column Name	Nature	Data Type	Description [2]	Sample Values
Incident ID	Identifier	Integer	Police Incident Number	201124150, 201184494
Offence Code	Categorical	String	Offense Code is the code for an offense committed inside the event, as established by the National Incident-Based Reporting System (NIBRS) of the CJIS Division Uniform Crime Reporting (UCR) Program.	9109, 9021
CR Number	Identifier	Integer	Police Report Number	17012068, 190057156
Dispatch Date / Time	Temporal	String	The actual date and time an Officer was dispatched	04/20/2018 05:37:51 PM, 07/14/2022 05:16:04 PM
NIBRS Code	Categorical	String	FBI NIBRS codes	90Z, 35A
Victims	Numerical	Integer	Number of Victims	1, 2
Crime Name1	Categorical	String	Crime against Society/Person/Property or Other	Crime Against Society, Not a Crime
Crime Name2	Categorical	String	Describes the NIBRS_CODE	All Other Offenses, Simple Assault

Crime Name3	Categorical	String	Describes the OFFENCE_CODE	DRIVING UNDER THE INFLUENCE LIQUOR, DRUGS - MARIJUANA - POSSESS
Police District Name	Categorical	String	Name of District (Rockville, Wheaton, etc.)	WHEATON, ROCKVILLE
Block Address	Geographical	String	Address in 100 block level	7300 BLK CALHOUN PL,14300 BLK LAYHILL RD
City	Geographical/ Categorical	String	City	SILVER SPRING, GAITHERSBURG
State	Geographical/ Categorical	String	State	MD, DC
Zip Code	Geographical /Categorical	Float	Zip code	20910,20902
Agency	Categorical	String	Assigned Police Department	MCPD, GPD
Place	Categorical	String	Place description	Residence - Apartment/Condo, Parking Lot - Commercial
Sector	Categorical	String	Police sector name, a subset of District	J, L
Beat	Categorical	String	Police patrol area, a subset of Sector	4J1,4L1
PRA	Categorical	String	Police response area, a subset of Beat	380,701
Address Number	Categorical	Float	House or Business Number	200,700
Street Prefix	Categorical	String	North, South, East, West	N, W
Street Name	Categorical	String	Street name	RANDOLPH, GLENMONT
Street Suffix	Categorical	String	Quadrant (NW, SW, NE, SW, etc.)	S, NW
Street Type	Categorical	String	Type of street (Ave, Drive, Highway, etc.)	RD, AVE
Start_Date_Time	Temporal	String	Occurred from date/time	03/07/2018 06:40:00 PM,04/20/2019 06:00:00 PM
End_Date_Time	Temporal	String	Occurred to date/time	03/09/2018 05:30:00 PM,04/21/2019 02:00:00 PM
Latitude	Geographical	Float	Latitude	39.0757,39.0567
Longitude	Geographical	Float	Longitude	-77.002,-77.0489
Police District Number	Categorical	String	Major Police Boundary	4D,6D
Location	Geographical	String	Location	(39.0757, -77.002),(39.0567, -77.0489)

B. *Programming Languages and Tools*

The vast amounts of data being produced, stored and consumed worldwide would increase rapidly, reaching 64.2 zettabytes in 2020. Over the next few years up to 2025, global data creation is projected to grow to more than 180 zettabytes [3].

The sheer large scale of these numbers is truly astonishing, but all this raw data would be useless unless it is analysed in some manner that can give informative insights, patterns, trends, and correlations.

Data analysis is a methodology that involves numerous steps such as collecting, cleaning, and organising the data, further it is analysed, interpreted, and visualised for information. There are two primary methods of data analysis:

1. Qualitative Analysis: It relies on data collected through interviews, surveys, and observations, which can help develop ideas or hypotheses for potential quantitative research. It is generally used to understand concepts, ideas, thoughts, or experiences following a subjective approach and doesn’t use statistics.
2. Quantitative Analysis: It involves applying statistical methods to the collected raw data to discover hidden trends and unseen patterns. It involves the use of various analytical tools and software to ascertain, research and find information.

The given dataset requires a quantitative analysis, specifically statistical analysis which tries to answer the question, “What happened?”. Statistical analysis will illustrate means and deviations in continuous data, whereas percentages and frequencies in categorical data.

Data analysis enables future prediction of trends, better customer targeting, improvements in business decisions, and much more. It has inspired a whole new era of tools and programming languages, some of the popular ones alongside their features, advantages, and disadvantages are discussed below:

Table 2. COMPARISON OF DATA ANALYSIS TOOLS / PROGRAMMING LANGUAGES

Comparison Factor	Python	R	SAS
Overview	Most versatile and fastest growing high-level programming languages today. It was initially developed for object-oriented programming and web development, later it evolved into a major tool for data analysis and scientific computing.	R, a low-level language and highly extensible is one of the leading programming languages that can be efficiently used for statistical analysis, visualisation, and modelling.	SAS is a statistical software suite widely used for BI (Business Intelligence), data management, and predictive analysis.
Used by	Spotify, Netflix, NASA, Google, and CERN are some of the popular users of Python.	Uber, Twitter, Google, and Facebook use R for advanced analysis and forecasting.	Google, Facebook, Netflix, and Twitter are a few companies that use SAS. It is also heavily used in clinical research.
Cost	It is free/open-source.	It is free/open-source.	It is proprietary software and expensive, packages cost is additional.
Learning curve	Quite a simple syntax and easy to grasp, a programming background is not required.	Difficult syntax and needs time to grasp, especially for users without any prior programming knowledge.	Easy to learn, it has good GUI and SQL knowledge that will speed up your understanding.
Speed and RAM	Fastest amongst the three and ideal for critical applications, performance is not limited by RAM size.	Slow speeds due to longer low-level language codes, and limited performance based on RAM size.	Due to drag and drop abilities, automatically creates, and optimizes code, with no dependency on RAM size.
Visualization	Rich set of libraries such as matplotlib, seaborn, vispy, etc. provides excellent graphical capabilities.	Provides the best visualisation amongst the three with its packages such as ggplot, ggplot2, lattice, etc.	Still lacks the visualisation prowess when compared with the other two, but efforts are being made to improve.
Deep Learning	Packages like Keras and TensorFlow allow advanced AI.	Packages like kerasR and Keras provide the interface to Python packages.	Still at an infancy stage in deep learning.
Customer Support	No dedicated support, but a huge online community of developers and official documentation.	No dedicated support, and a moderately large community of developers but it is lesser when compared to Python.	Excellent customer support at the corporate level to troubleshoot issues.

Python in comparison with all the other tools packs all the required features needed for data analysis. It stands out as the preferred choice for analysing the Montgomery Crime Dataset, its ease of use and wide range of libraries allows exploration, pre-processing, manipulation, and highly customizable powerful visualisations.

Python libraries that will be required:

- NumPy

  NumPy stands for Numerical Python, a powerful array processing package that can handle n-dimensional arrays. Many packages are built on NumPy due to its fast and efficient container objects to pass around and work on the homogeneous data.

- Pandas

  Pandas provides high-performance and easy-to-use data structures for working with tabular or structured datasets. It provides two important object types which are at its core, DataFrame which resembles a 2-D array or a table with rows and columns, and Series which resembles a 1-D array. It has a goal to become the most powerful and flexible tool available for data manipulation and real-world data analysis in any programming language [4].

- GeoPy

  GeoPy provides clients for geocoding web services, this allows to easily locate any addresses using latitude and longitude or vice versa. Google Maps, Bing Maps or Openstreet Map Nominatim web services have their own classes providing interfaces to the respective APIs..

- Matplotlib

  Matplotlib is one of the most popular visualisation packages. It can produce static, animated, and interactive plots, ranging from simple line graphs, histograms, and scatter plots to complex heatmap graphs. It graphs your data on Figure (e.g. Jupiter widgets, windows, etc.), each of which contains one or more Axes, an area where points can be represented in terms of coordinates [5].

  It has essentially two ways to plot any visualisation:

  i. The object-oriented style is where Figures and Axes objects are created explicitly, and then call methods on them to plot.
  ii. The function-based style where pyplot sub-package implicitly creates and manages Figures and Axes, using the pyplot functions to plot.

- Seaborn

  Seaborn is an advanced data visualisation package; it is built on top of Matplotlib and tightly integrates with Pandas. It provides a high-level interface to plot informative statistical graphics allowing easy exploration of data. It automatically performs the internal semantic mappings and aggregations to produce graphs using simple functions and its parameters, but sometimes to achieve some customizations it might be necessary to drop down to the matplotlib layer.

- Folium

  Folium combines the data wrangling strengths in Python with the geographical mapping abilities of Leaflet.js enabling various types of visualisations in maps such as heat maps, grid maps, and color maps.

C. *Initial Data Quality Analysis*

The dataset contains some useful information using which crime patterns can be observed and visualized. On the other hand, the dataset also contains null fields or fields that hold invalid data. A careful observation of the null values in the dataset is made and conclusions are in a way that few fields can be dropped/avoided but some fields which hold analytical importance are suggested to have values, which in turn will help the analysis of the criminal activity.

1. Transformation Methods

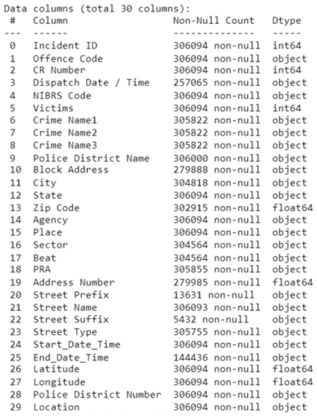
Data transformation is important to clean the dataset and make it a clean dataset for further analysis. There are multiple data transformation techniques, which are to be opted based on the requirement of the dataset. Different types of data transformation methods are listed below:

- Removing Duplicates
- Modification of Data using Function**s**
- Replacing Values
- Detecting and Filtering Outliers
- Data Pre-processing

In order to deal with data using Python packages, there are a set of steps to be followed to get a clear picture of data contained in the dataset.

Step 1: Load the dataset into the python notebook using the pandas library.

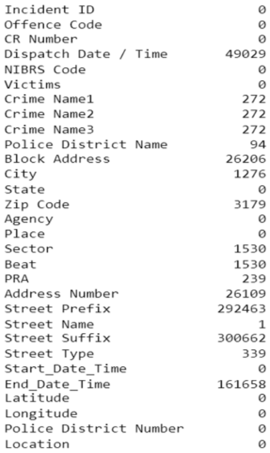
Step 2: Get the information about the dataset, using .info()



```
Data columns (total 30 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Incident ID            306094 non-null  int64
 1   Offence Code           306094 non-null  object
 2   CR Number              306094 non-null  int64
 3   Dispatch Date / Time   257065 non-null  object
 4   NIBRS Code             306094 non-null  object
 5   Victims                306094 non-null  int64
 6   Crime Name1            305822 non-null  object
 7   Crime Name2            305822 non-null  object
 8   Crime Name3            305822 non-null  object
 9   Police District Name   306000 non-null  object
 10  Block Address          279888 non-null  object
 11  City                   304818 non-null  object
 12  State                  306094 non-null  object
 13  Zip Code               302915 non-null  float64
 14  Agency                 306094 non-null  object
 15  Place                  306094 non-null  object
 16  Sector                 304564 non-null  object
 17  Beat                   304564 non-null  object
 18  PRA                    305855 non-null  object
 19  Address Number         279985 non-null  float64
 20  Street Prefix          13631 non-null   object
 21  Street Name            306093 non-null  object
 22  Street Suffix          5432 non-null    object
 23  Street Type            305755 non-null  object
 24  Start_Date_Time        306094 non-null  object
 25  End_Date_Time          144436 non-null  object
 26  Latitude               306094 non-null  float64
 27  Longitude              306094 non-null  float64
 28  Police District Number 306094 non-null  object
 29  Location               306094 non-null  object
```

Fig 2. INFORMATION OF THE DATASET FEATURES

Step 3: Fetch the details of the null values, column-wise.



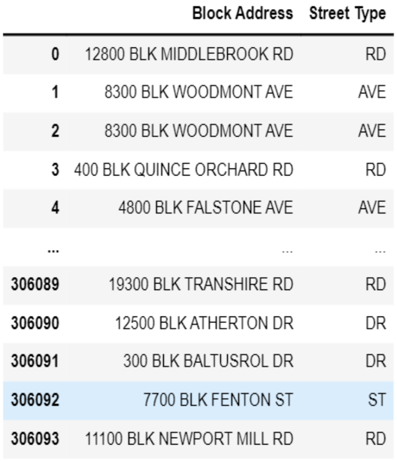
```
Incident ID                    0
Offence Code                   0
CR Number                      0
Dispatch Date / Time       49029
NIBRS Code                     0
Victims                        0
Crime Name1                  272
Crime Name2                  272
Crime Name3                  272
Police District Name          94
Block Address              26206
City                        1276
State                          0
Zip Code                    3179
Agency                         0
Place                          0
Sector                      1530
Beat                        1530
PRA                          239
Address Number             26109
Street Prefix             292463
Street Name                    1
Street Suffix             300662
Street Type                  339
Start_Date_Time                0
End_Date_Time             161658
Latitude                       0
Longitude                      0
Police District Number         0
Location                       0
```

Fig 3. A NUMBER OF NULL VALUES FOR EACH COLUMN

Step 4: Understand the relationships between the columns containing null values, and the analytical value that these columns hold.

Step 5: Based on the conclusions made, different transformation/ data manipulation techniques can then be inferred so that the data is clear of null values.

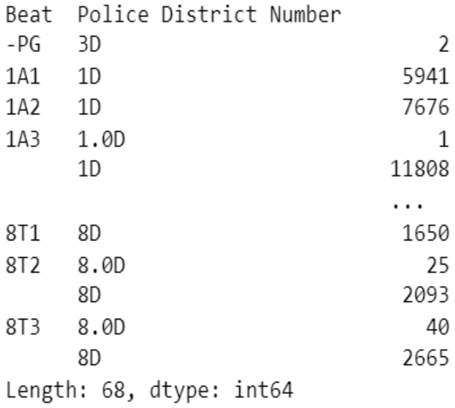
2. Data Cleaning and Methods to account for
   Analysis of the null values of the columns and ways to replace the null values
   - Street Suffix/Street Prefix has the highest number of null values, Street Prefix column holds 95.54% of its values Null/empty. Also, Street Suffix has 98.22% of its total entries null, which corresponds to all the columns' data do not have any valid entries. The approach is to drop the null values of these columns.
   - On further analysis, the 'Crime name1',' Crime Name2', and 'Crime Name3' columns have an equal number of null values, for the same records. NIBRS system of reporting data has a crime name associated with the NIBRS code mentioned in the report, and also in the data set.
   - The approach is to fill the null values of these columns/records based on the corresponding crime name given in the official documentation of NIBRS style of reporting.
   - Though latitude and longitude have no null values, there are some records in which these columns hold '0', which is an invalid value and does not have a proper location associated with it. The idea is to get the latitude and longitude for these records using other available columns like street name, block address, etc.
   - The null values in 'Block Address' can be replaced with the block address details that are obtained using the longitude and latitude details in the dataset. The python package 'geopy' helps achieve this.
   - Location column also has no null values, but the column contains data that is just a combination of latitude and longitude columns' data. This column is a redundant column, which can be dropped as it does not serve any great analytical value.
   - 'Sector' and 'Beat' are related columns, that is Beat is a subsection of Sector.
   - 'Street Type' has approximately 0.11% null entries in the data set. Also, in records where the street type column is not null, it is observed that the street type is mentioned as a part of the address block column's entry. After the address block columns are cleaned up and the missing values are filled for this column, the null values in the street type column can be filled up using the address block column data.



	Block Address	Street Type
0	12800 BLK MIDDLEBROOK RD	RD
1	8300 BLK WOODMONT AVE	AVE
2	8300 BLK WOODMONT AVE	AVE
3	400 BLK QUINCE ORCHARD RD	RD
4	4800 BLK FALSTONE AVE	AVE
...	...	...
306089	19300 BLK TRANSHIRE RD	RD
306090	12500 BLK ATHERTON DR	DR
306091	300 BLK BALTUSROL DR	DR
306092	7700 BLK FENTON ST	ST
306093	11100 BLK NEWPORT MILL RD	RD

Fig 4. COMPARING THE CONTENT OF BLOCK ADDRESS WITH STREET TYPE

- The columns ‘Beat’, ‘Sector’, ‘Police District Number’, and ‘Zip Code’ share an interesting relationship. It is observed that the combination of ‘Zip Code',' Beat’, and ‘Police District Number’ is unique and 2 of any of these columns combined have a one-on-one relationship with the third column. Also, Sector is a subset of Beat that means, and once we obtain null replacements of the ‘Beat’ column, we can use this data to replace the nulls of the Sector column. ‘Police District Number’, ‘Zip Code',' Beat’ relationship



```
Beat  Police District Number
-PG   3D                          2
1A1   1D                       5941
1A2   1D                       7676
1A3   1.0D                        1
      1D                      11808
                               ...
8T1   8D                       1650
8T2   8.0D                       25
      8D                       2093
8T3   8.0D                       40
      8D                       2665
Length: 68, dtype: int64
```

Fig 5. COUNT POLICE DISTRICT NUMBER FOR EACH BEAT

This data still contains a few invalid/redundant entries of police district number. For example, 8.0D corresponds to 8D, which is just a typographical/human error.

3. Data Wrangling Skills
Data Wrangling is a process of cleaning, organizing, and transforming data into a desirable state. In the provided dataset, the data wrangling skills used are
    - Identifying cells with null values and deciding to fill the values or drop them.
    - Deleting data that does not help or add value to the visualization techniques.
    - Identifying outliers in the dataset and deciding to use these outliers or drop them to avoid disturbances in the visualization.

After applying data wrangling skills to the dataset, the data is clean and is ready for further steps of investigation which is Explanatory data analysis. After this process, all unwanted data or data which is important to visualize the data is removed or filled with useful values to facilitate the analysis and answer the research questions more cleaner.[6]

III. Exploratory Data Analysis

A. *Introduction*

Exploratory Data Analysis (EDA) is a major step in data pre-processing. It provides us with a brief overview of the following

- contents of the data set
- its singularities, abnormalities, and main characteristics.
- relationship between features
- identifying which features are crucial for our problem

EDA is applying multiple statistical models to the data such as

Descriptive statistics: Provide some measures of the features such as mean, median, percentage, frequency, etc.

Inferential statistics: give an overview of the relationship between features using correlation, regression, and ANOVA.

Data grouping: some basic grouping using group by in pandas

EDA techniques to perform detailed data analysis

After performing Preliminary Data Analysis in the previous section [link here], we now have much cleaner data without any null or duplicated records, but we still need to figure out more about our dataset using the following EDA approaches:

- If there are any outliers in our data, if yes then how are we going to deal with them (Using inferential Statistics, IQR)
- How correlated our features are. If two features are highly correlated, can we figure out patterns in our data or can we use one of them to get better results (Statistical Correlation and regression models)
- Figuring out patterns of crimes using aggregated features by grouping two or more features

B. *Descriptive Statistics*

To make a thorough analysis of data, we can start with the univariate analysis. It gives us insights about a single variable and provides us the information such as count, percentage, frequency, etc.

There are multiple statistical methods that we use in data analysis, directly or indirectly (using plot or graphs function that are using these methods in the backend)

Measures Of the Centre

The measure of centre statistical method can be used in univariate and bivariate analysis both.

1. Mean: Mean is a measurement that represents the average of all the values in a sample.
2. Median: The median is a measurement centre value of a value set.
3. Mode: The most recurring value in a data set or feature is known as Mode.

Using descriptive Analysis, we can analyse each of the variables in the sample data set for mean, standard deviation, minimum and maximum.

Mean (): If we want to find out the mean or an average number of victims for the given years, we will check and calculate the average of all values. For example, we take the sum of the victims of each year, divided by the total number of victims following the following formula [7][8]:

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

For this purpose, we can use the mean method which will return the arithmetic mean A.

Median (): If we want to find out the centre value of a feature, we will arrange the values in ascending or descending order and choose the middle value. Sometimes grouping value by using the median is a helpful solution to see data points from a unique perspective [7][8].

$$\text{if } n \text{ is odd, } \operatorname{median}(x) = x_{(n+1)/2}$$

$$\text{if } n \text{ is even, } \operatorname{median}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

Mode (): If we want to find out the most common type of crime among the subcategory ‘Crime Name3’, we will check the value which is repeated the most number of times. This can be done using the mode method.

Measures of spread

1. Range: It is the quantification of how widely spaced out a data set's values are.
2. Inter Quartile Range (IQR): It is a measure of variability that is based on quartilizing a data set.
3. Variance: It describes how far a random variable deviates from the value that is anticipated. It involves calculating deviation squares.
4. Standard Deviation: It measures how far apart a batch of data is from the mean.

Univariate Analysis

By looking at each variable separately, we can answer some of the research questions via univariate analysis such as

- What time of the year/month/weeks/days/seasons is most active for criminals?
- What are the most common types of crimes based on this dataset?
- Which area/region/city is most affected and/or safest

To dig deep into these questions, let us see which statistical methods we can use and why.

Stdev (): if we calculate the standard deviation of a feature, it gives us a clear picture of how far the data points are scattered from the mean. This is an immensely powerful measure to analyse a feature and investigate the patterns. For example, for a feature like ‘start_year’ that indicates the year when a crime was committed, we can plot a scatter plot and see how these crimes are scattered throughout a year using the following formula [7][8]:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2}$$

Here σ is the standard deviation and μ is the mean.

variance (): For some features, we might like to see how each number is linked with another, or in our case, we want to see how each crime is linked with another. Instead of arranging these crimes in quartiles, we can use variance to assess the relationship. Variance can be calculated as follows [7][8]:

$$Var(X) = E[(X - \mu)^2]$$

Bivariate Analysis

The goal of the bivariate analysis is to find an empirical association between two variables/features.

When there are 2 features involved in the research question, we can use bivariate analysis to produce answers to those questions.

Some of the research questions can be answered using this approach such as

- Is there a connection between a particular crime and a specific time of day or month?
- Which states/places/street types (everything related to a place) are most affected and/or safest by the crimes?
- Can multiple offenses link to a single reported incident?

The following statistical methods are used to find the relationship between the two features.

covariance (): using this method we can investigate the relationship between two features whether there is a positive or negative trend, and what we can figure out from these trends using the below formula [7][8]:

$$\mathrm{Cov}(X, Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n}$$

correlation (): This method is a normalized version of covariance. It is very crucial to see the correlation between features. Covariance is a measurement that changes according to the units of the applied variables. Therefore, the digits of covariance also vary if either or both variables' units are modified. The correlation addresses this issue. Thus, this method will provide us with more normalized results to investigate the trends. Following is the formula to calculate the correlation [7][8]:

Correlation Coefficient = ∑(x(i)- mean(x))*(y(i)-mean(y)) / √ (∑(x(i)-mean(x))2 * ∑(y(i)-mean(y))2)

Where,

- x(i)= feature x
- Mean(x) = mean of feature x
- y(i) = feature y
- Mean(y) = mean of feature y

We can use Covariance to find out the relationship between two features while correlation can be used where we might want to investigate two or more features in our dataset.

linear_regression (): This is another important statistical method where we can predict the value of a variable (dependent variable) from the value of another variable (independent variable) [7][8]

$$y = a + bx$$

Here a and b can be calculated as follows:

$$b= \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a= \frac{\sum y - b(\sum x)}{n}$$

- b = Slope of the line
- a = Y-intercept of the line
- X = Feature X values
- Y =Feature Y values

C. Data Visualisation

In this section, we discuss different types of visualisations based on the statistical analysis, to answer the research questions. The Table. 3 provides the plots which can be used regarding the statistical analysis type.

Table 3. DIFFERENT TYPES OF THE PLOTS BASED ON THE STATISTICAL CONCEPTS

Univariate Analysis	**Bivariate Analysis (Categorical)**	**Bivariate Analysis (Numerical)**
Histogram Distribution plot Box plot Count plot Pie plot	Bar plot Histogram	Scatter plot Reg plot Line plot

The plots/graphs that are appropriate for the research questions are explained as below:

For research Q1.

- "Line plot" shows the relationship between two features. It can help in showing the trends (ups and downs) between two features.

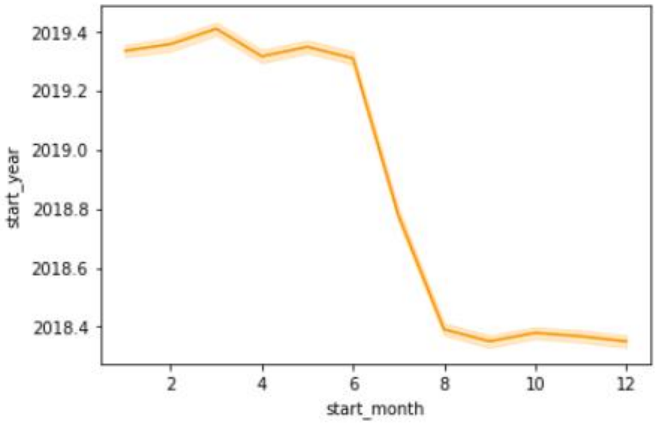


Fig 6. LINE PLOT

- "Displot" is a histogram but with added functionality to form a line that is used to plot a univariate distribution.

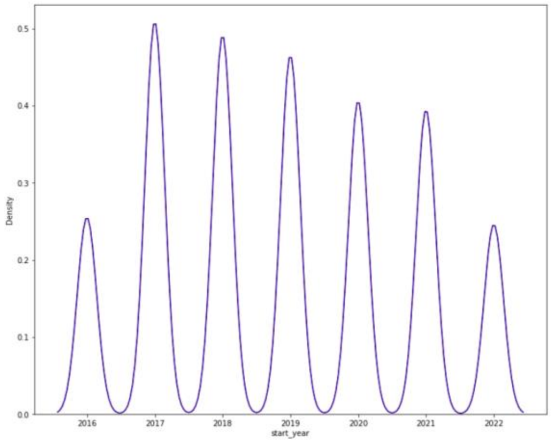


Fig 7. DISPLOT

For research Q2.

- To find out the correlation between two features, a "scatter plot" can be used. This plot is also a way to detect outliers based on the relationship between two features as well as understand the relationship whether it's negative or positive.

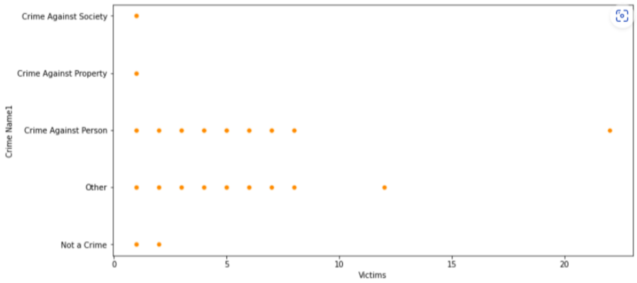


Fig 8. SCATTER PLOT

For research Q3.

- In “grouped count bar charts”, we can have two categorical features as well as the relationship in one figure, which is well suited for this question.

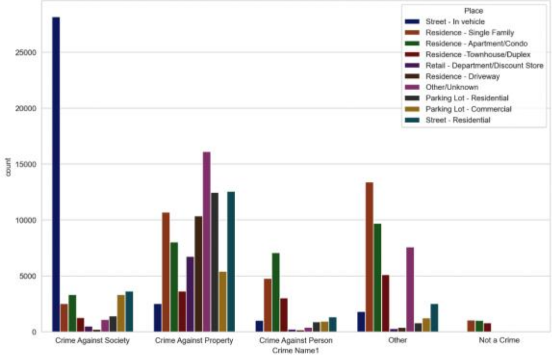


Fig 9. GROUPED COUNT BARCHART

For research Q4.

- To answer this research question, the best-suited type of graph would be a “bar plot with stacked bars”. This is because bar graphs are used to show comparisons between different categories of data.

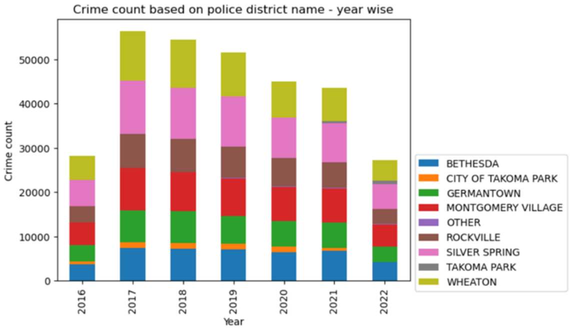


Fig 10. BAR PLOT WITH STACKED BARS

For research Q5.

- This analysis requires getting the time difference between the dispatch time of the police officers and the time crime was reported.

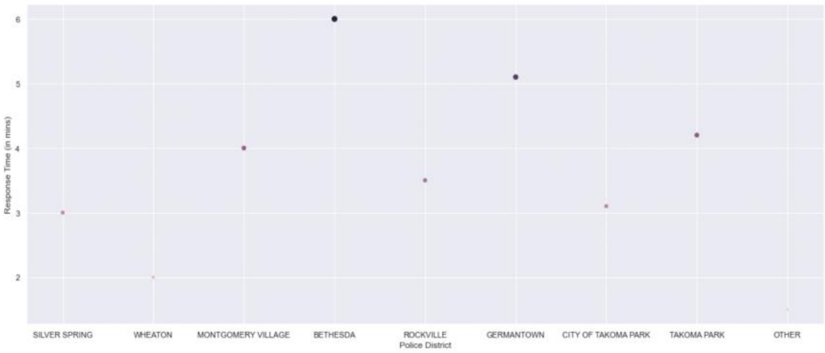


Fig 11. SCATTER PLOT

For research Q6.

- “Histogram plot” can represent how one or more variables are distributed and a parameter ‘hue’ is used to present categorical features in a single graph that can be very beneficial in figuring out patterns.

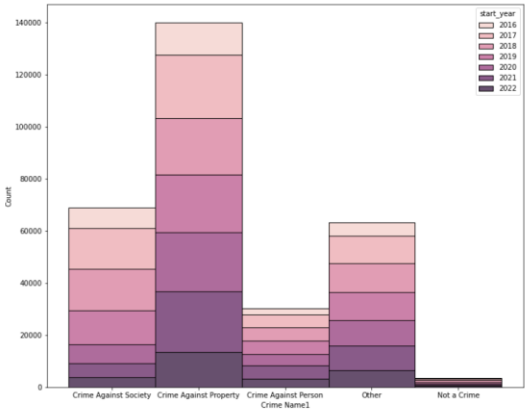
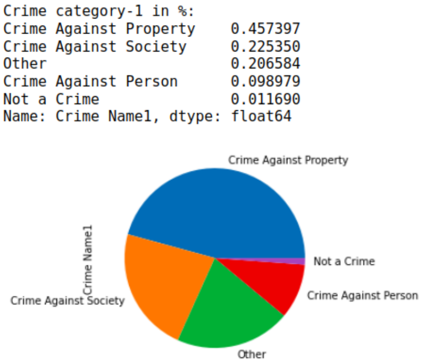


Fig 12. HISTOGRAM PLOT

- “Pie chart” along with the count_values function of pandas can be used to group a variable based on a year and then show the presentation.



```
Crime category-1 in %:
Crime Against Property    0.457397
Crime Against Society     0.225350
Other                     0.206584
Crime Against Person      0.098979
Not a Crime               0.011690
Name: Crime Name1, dtype: float64
```

Fig 13. PIE CHART

For research Q7.

- To understand this research question, a “bar graph” can be used to show the time of the day related graphs.

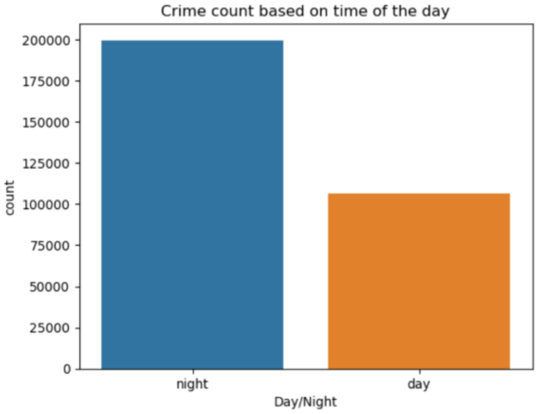


Fig 14. BAR GRAPH

For research Q8.

- A “geographical heat map” can be used to easily highlight areas on a map that have been affected by events.

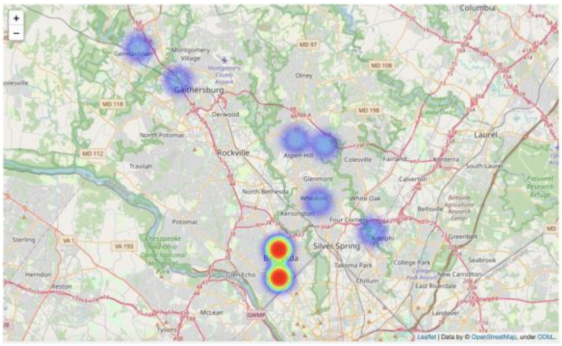


Fig 15. HEAT MAP

For research Q9.

- A “bar chart” can be used to show the distribution of the crime data in the categorical feature of the weekdays.

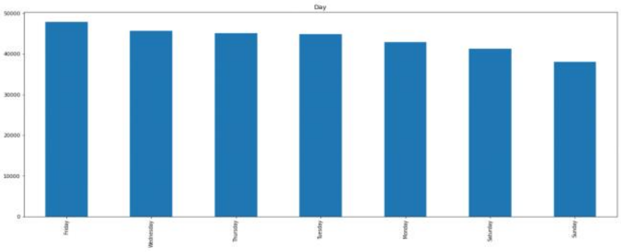


Fig 16. BAR CHART

For research Q10.

- “Bar chart” The best way to visualize the answer to this question is using a bar chart, which Compares the total number of crimes for each city.

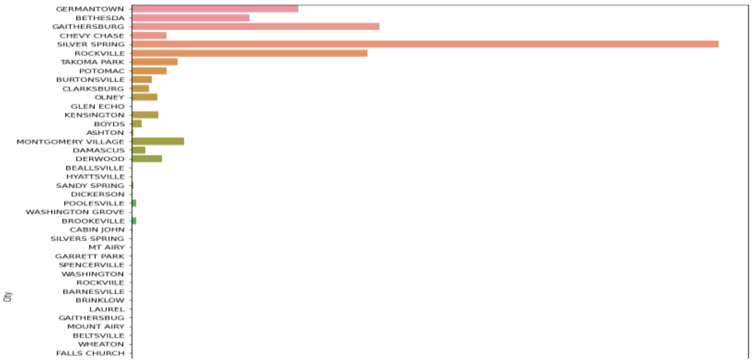


Fig 17. BAR CHART

IV. CONCLUSION

It is not surprising to see how crimes are rising with each passing day but fortunately, data science techniques can help reduce these numbers. For this project, we were given a dataset of crimes reported by Montgomery County for the years 2016-2022. An initial assessment of the dataset gave us enough insight to produce appropriate tools and technologies that were further used for the initial quality assessment of data. After analysing this data and performing initial processing we figured out that a more detailed explanatory data analysis is mandatory to work out how criminals are operating. This can be done by using inferential and descriptive analysis methods such as mean, standard deviation, correlation, and regression. We also found interesting research questions from the initial analysis such as “Is there any specific place that is being targeted by these criminals to commit crimes?”, “Is there a specific time or date that they choose to carry out some illegal activity?” To answer these questions, we need to dig deeper in data analysis and take help from data visualization to produce better explanations and suggestions. The outcome of this analysis can shed some light on the rationality and patterns behind these criminal minds and how they operate and hopefully will help police districts in mitigating these crimes.

V. References

[1] "National Incident-Based Reporting System (NIBRS)," Bureau of Justice Statistics. [Online]. Available: https://bjs.ojp.gov/data-collection/national-incident-based-reporting-system-nibrs#publications-0. [Accessed: 10-Nov-2022].

[2] "National Incident-Based Reporting System (NIBRS)," Bureau of Justice Statistics. [Online]. Available: https://bjs.ojp.gov/data-collection/national-incident-based-reporting-system-nibrs#publications-0. [Accessed: 10-Nov-2022].

[3] M. D. Montgomery County, "Crime: Open data portal," Crime | Open Data Portal, 10-Nov-2022. [Online]. Available: https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3. [Accessed: 10-Nov-2022].

[4] Published by Statista Research Department and S. 8, "Total Data Volume Worldwide 2010-2025," Statista, 08-Sep-2022. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/. [Accessed: 10-Nov-2022].

[5] "About Pandas," pandas. [Online]. Available: https://pandas.pydata.org/about/index.html. [Accessed: 10-Nov-2022].

[6] "Quick start guide," Quick start guide - Matplotlib 3.6.2 documentation. [Online]. Available: https://matplotlib.org/stable/tutorials/introductory/quick_start.html#sphx-glr-tutorials-introductory-quick-start-py. [Accessed: 10-Nov-2022].

[7] Simplilearn. "What Is Data Wrangling? Benefits, Tools, and Skills | Simplilearn." Simplilearn.com. https://www.simplilearn.com/data-wrangling-article#:~:text=Top%20Data%20Wrangling%20Skills%20Required&amp;text=A%20good%20data%20wrangler%20should,out%20to%20enrich%20the%20data. [Accessed: 10-Nov-2022].

[8] Z. Geer. "A Complete Guide To Math And Statistics For Data Science - DZone Big Data." dzone.com. https://dzone.com/articles/a-complete-guide-to-math-and-statistics-for-data-s [Accessed Nov. 10, 2022].

[9] "Statistics - mathematical statistics functions," statistics - Mathematical statistics functions - Python 3.11.0 documentation. [Online]. Available: https://docs.python.org/3/library/statistics.html. [Accessed: 10-Nov-2022].