PRACTICAL REPORT

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

# Table of Contents

1. [INITAL DATA ANALYSIS 6](#_bookmark0)
   1. [Quality of the given Data: 7](#_bookmark1)

[Size of data set: 7](#_bookmark2)

[Frequency Counts 7](#_bookmark3)

[Descriptive/ summary statistics 8](#_bookmark4)

[Normality 10](#_bookmark5)

[Outlier detection 12](#_bookmark6)

* 1. [Data Transformation 13](#_bookmark7)

[Data Sorting: 13](#_bookmark8)

[Data Deletion 13](#_bookmark9)

* 1. [Characteristics of data set: 15](#_bookmark10)

[Display of data top and bottom 15](#_bookmark11)

[Basic Plots 15](#_bookmark12)

1. [EXPLORATORY DATA ANALYSIS 19](#_bookmark14)
   1. [Finding Null values and missing values 19](#_bookmark13)
   2. [Correlation Structures and Distribution Characteristics 20](#_bookmark15)
   3. Trends in data using appropriate visualization tools Error! Bookmark not

defined.

* 1. [Research questions and related visualizations 22](#_bookmark16)

[Visualization no 1 22](#_bookmark17)

[Visualization no 2 25](#_bookmark18)

[Visualization no 3 25](#_bookmark19)

[Visualization no 4 26](#_bookmark20)

[Visualization no 5 27](#_bookmark21)

[Visualization no 6 28](#_bookmark22)

[Visualization no 7 28](#_bookmark23)

[Visualization no 8 30](#_bookmark24)

[Visualization no 9 30](#_bookmark25)

[Visualization no 10 31](#_bookmark26)

[2.5 32](#_bookmark27)

1. [CONCLUSIONS 33](#_bookmark28)

**REFERENCES** ERROR! BOOKMARK NOT DEFINED.

# List of Figures

|  |  |
| --- | --- |
| Fig no 1 | Importing Libraries |
| Fig no 2 | Head of imported data |
| Fig no 3 | Size of data |
| Fig no 4 | Frequency Count |
| Fig no 5 | visualization of frequency count |
| Fig no 6 | Descriptive Statastics summary |
| Fig no 7 | Graphical representation of Descriptive summary |
| Fig no 8(a,b,c,d,e,f,g,h) | Normality |
| Fig no 9 | Outliers |
| Fig no 10 | Outliers box plot |
| Fig no 11 | Data sorting |
| Fig no 12(a,b,c,d,e) | Data deletion |
| Fig no 13 | Data after deletion |
| Fig no 14 | Crime per year bar chart |
| Fig no 15 | Crime and crime types |
| Fig no 16 | Types of weapon |

|  |  |
| --- | --- |
| Fig no 17(a,b) | Dropping null values |
| Fig no 18 | Missing values |
| Fig no 19 | Correaltion |
| Fig no 20a | Visulization 1.1 |
| Fig no 21b | Visulization 1.2 |
| Fig no 22 | Visulization 2 |
| Fig no 23 | Visulization 3 |
| Fig no 24 | Visulization 4 |
| Fig no 25 | Visulization 5 |
| Fig no 26 | Visulization 6 |
| Fig no 27 | Visulization 7.1 |
| Fig no 27 | Visulization 7.2 |
| Fig no 28 | Visulization 8 |
| Fig no 29 | Visulization 9 |
| Fig no 30 | Visulization 10 |

1. **Initial Data Analysis**

The data set that will be analyzed in this report is the crime data of Montgomery County(State of Maryland - MD) of the USA’s crime statistics. The report will focus on the most violent crimes. The FBI categorized this data based on the severity of crime into different codes called NIBRS code. This report will be performing the analysis on NIBRS category A crime. ‘A‘ category has list of crime ID’s which are related to violent crime when compared to B category.

As the first step into the data analysis imported libraries like pandas, matplotlib, seaborn, NumPy and datetime.

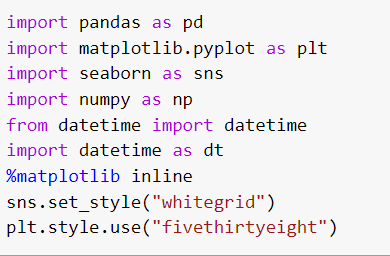


Fig no 1: importing Librareis

To upload and read the data, following code is used. head is used to display they top five rows of the given dataset.





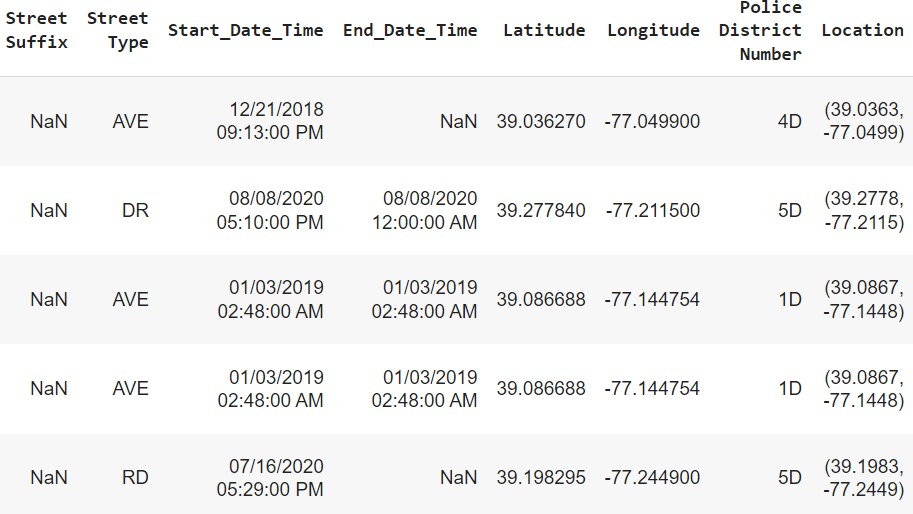


Fig no 2: Head

* 1. **QUALITY OF THE GIVEN DATA:**

**Size of data set:**

The size of the data is (322033, 30). The ‘322033’ is the data values in the data set and 30 is the variables (columns) in the data. Below is the code and the output for the size of the data.

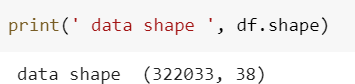


Fig no 3: Size of data

**Frequency Counts:**

Used the below code to find the missing values in each cell and displayed it as a row and the output is displayed below.

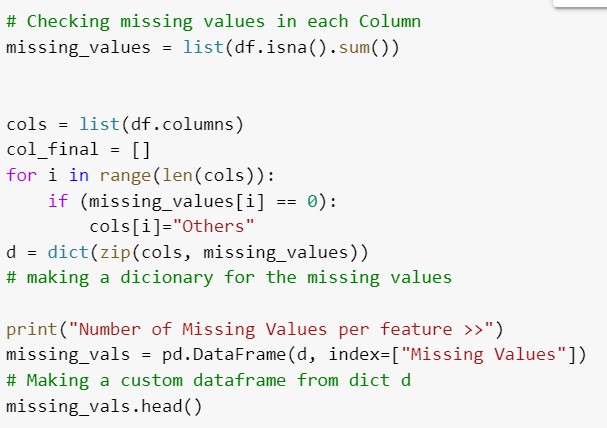


Fig no 4: Frequency Count

Above code gives following results

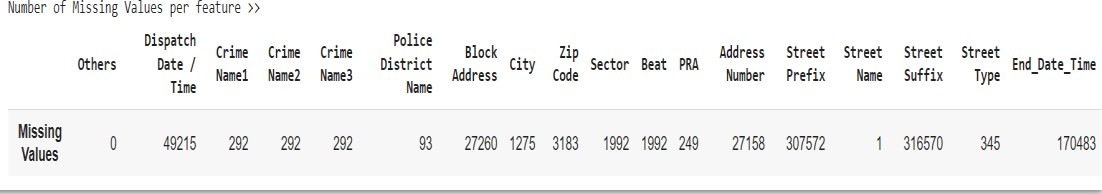
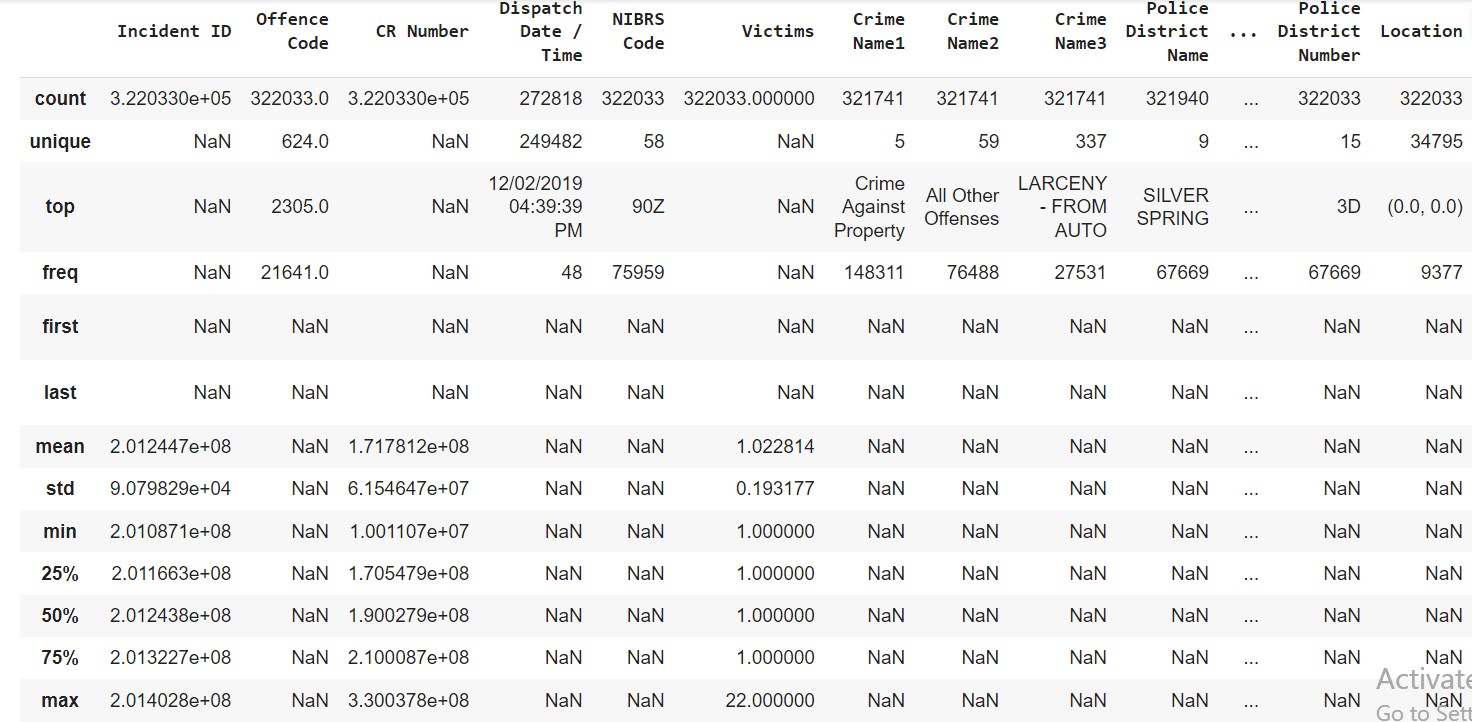


Fig no 5: visualization of frequency count

**Descriptive/ summary statistics:**

Descriptive statistics is the summary statistics which summarizes the features of the dataset. Below is the code and output.



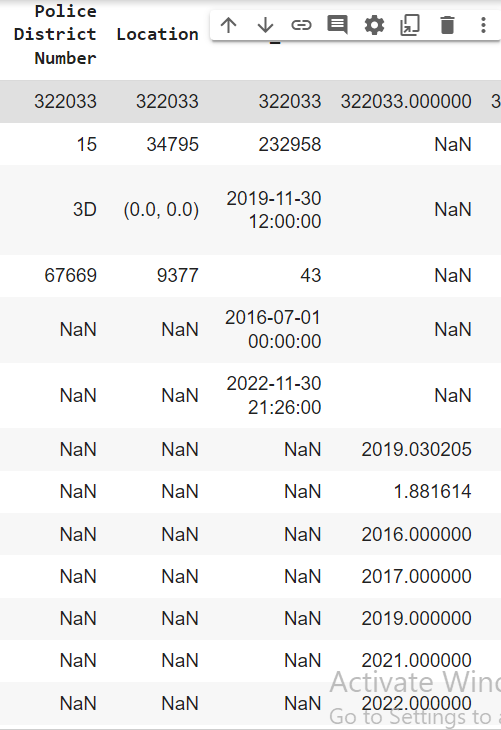
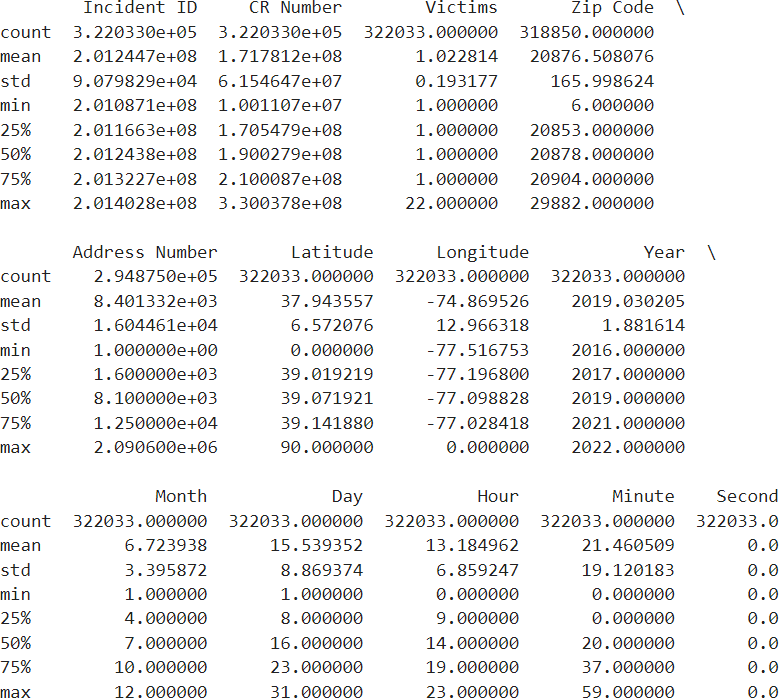


Fig no 6: Descriptive statastics summary



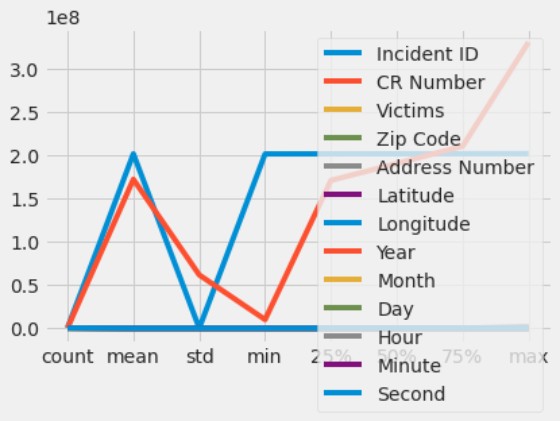
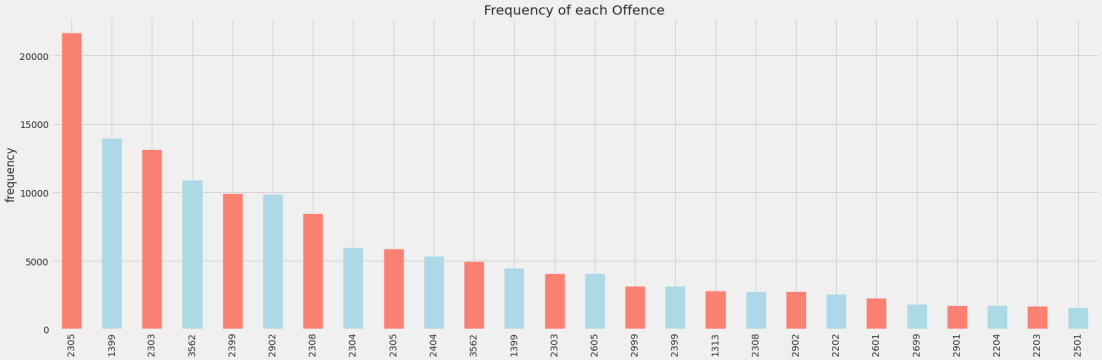
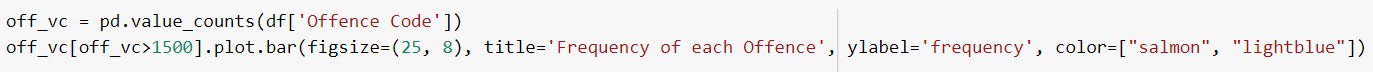
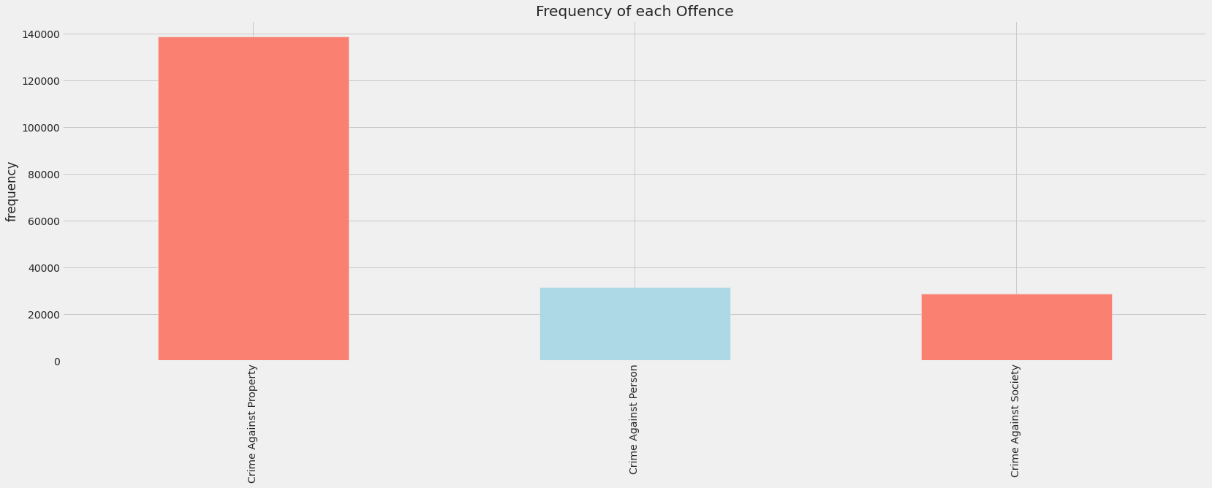


Fig no 7: Graphical representation of Descriptive summary

**Normality:**

Normality graphs (frequency histograms) detect if the data is normally distributed. The normal data gives out a bell shape frequency curve on histogram. The below is the code and curve for Offense code, CrimeName1,2,3, Police District Number column.





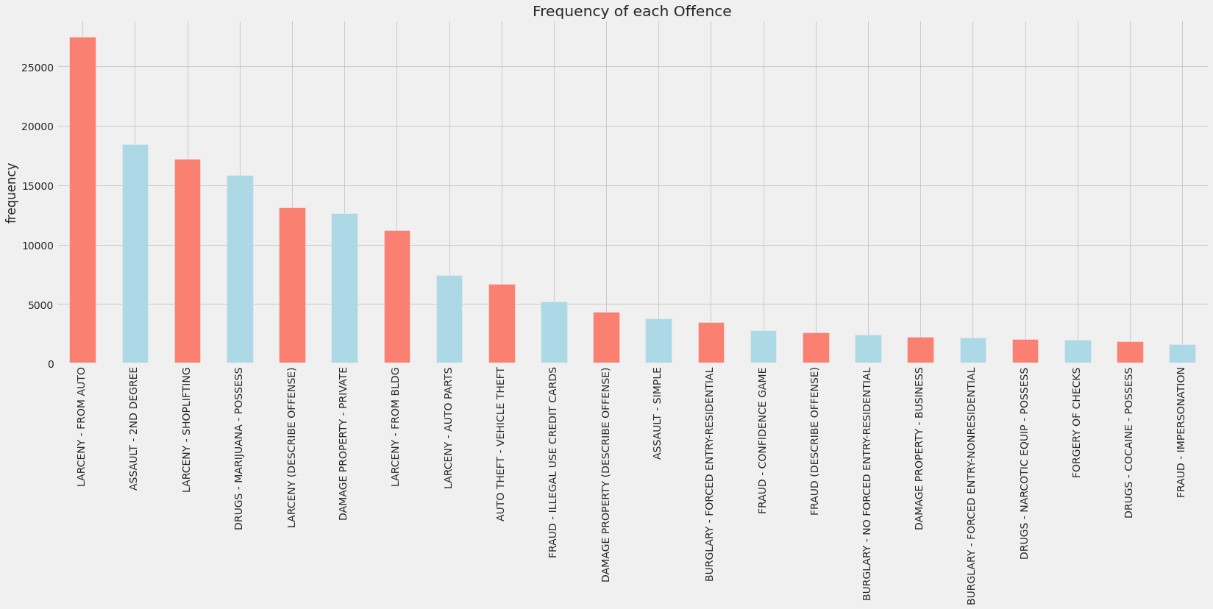
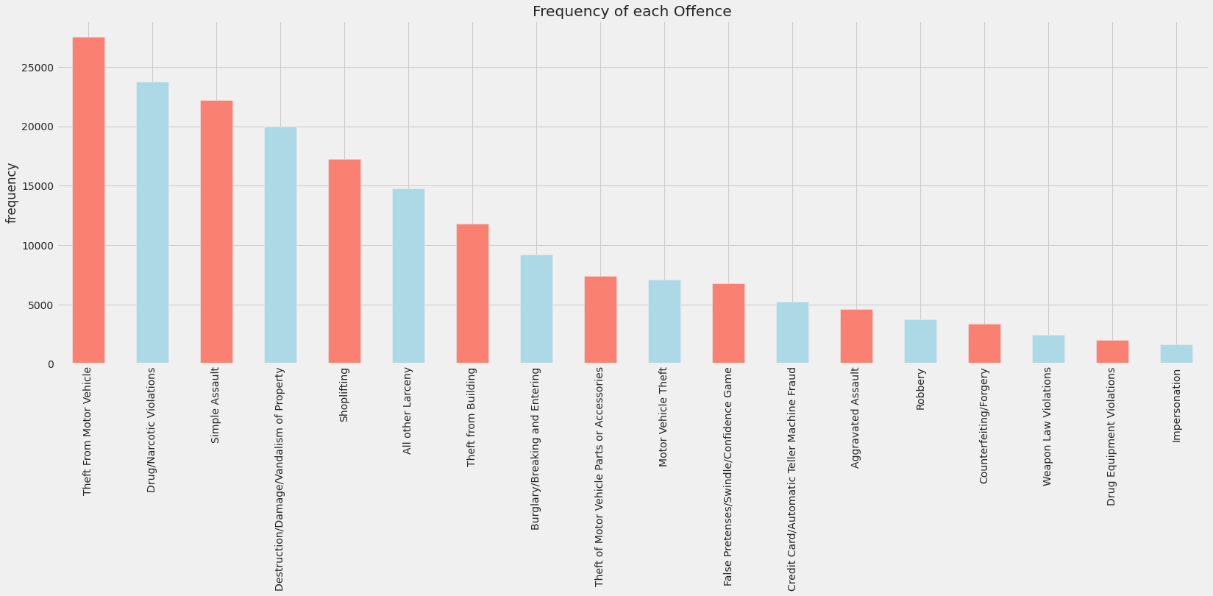


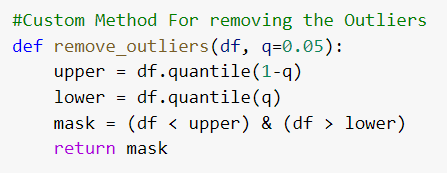
Fig no 8(a,b,c,d,e,f,g,h): Normality

The bar pattern shows no bell hence the data in these column is not normally distributed, means it has extreme values.

## Outlier detection:

The values or observation that lies at a abnormal distance from the rest of the values in data are called outliers.

The column with Victims has numerical data, we used the quantile function to understand the upper and lower limits and the code returned all values false, hence most of the data has no particular outliers and removed if any.



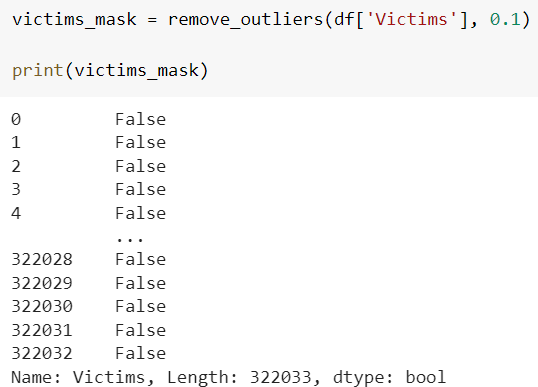
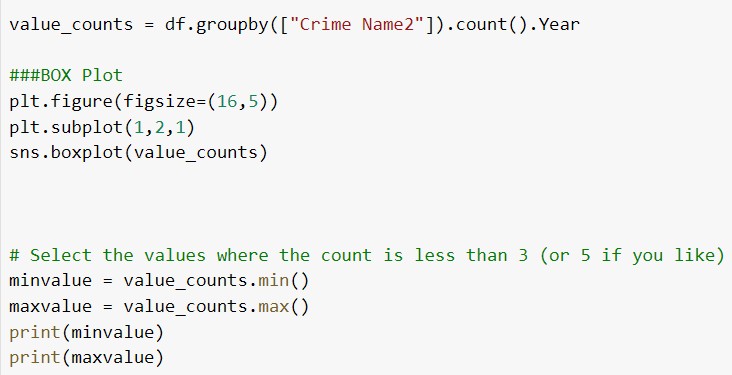


Fig no 9: Outliers

Another column, CrimeName 2 has extreme values, this column has characters, hence we initially used valuecount to count the number of repetitions and plotted the graph. The boxplot below shows the outliers as dots. We printed min and max values as it is the most extreme values from median. 1 and 72402 is output.



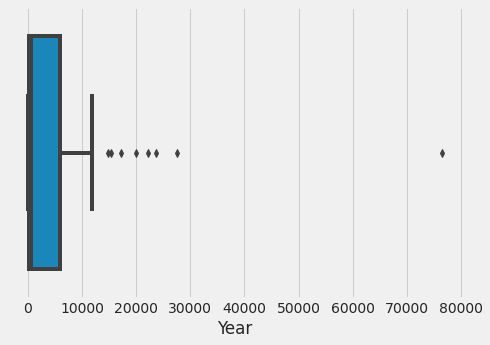


Fig no 10: Outliers box plot

* 1. **DATA TRANSFORMATION:**

**Data Sorting:**

The column Start date has the month year and date of the day that the crime happened. In our research questions there needs to be analysis made on day, time and year of crime occurrence along with other variables. For this analysis we need the start time split into month year and day separately and created 3 new properties namely: month, day, hour, minute and second. Below is the code for doing the same.

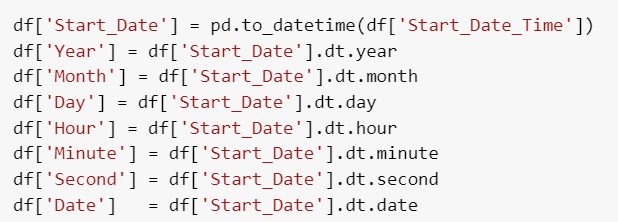
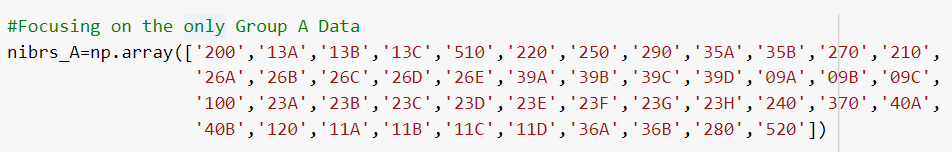


Fig no 11: Data sorting

**Data Deletion:**

The data that the report focuses on is the NIBRS crime A category data only. Hence we used a NumPy array to separate this data from the rest for the needed analysis. The code is as below.





Now, the research questions are based on only certain columns and not all of them. Hence dropping the columns which are unnecessary for the analysis and the code is a s below.



Fig no 12(a,b,c,d,e): Data deletion

The place column has lot of data which is quite repetitive after the first string. For example, it has values like Parking garage- church/commercial/county etc. However instead of plotting so many values we can split the term parking garage and exclude the rest if the string to analyse which decrease the data and gives out a more insightful information through visualisation. Below if the code for the extraction.

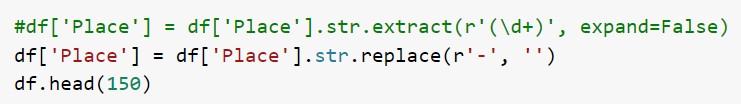




Fig no 13: Data after deletion

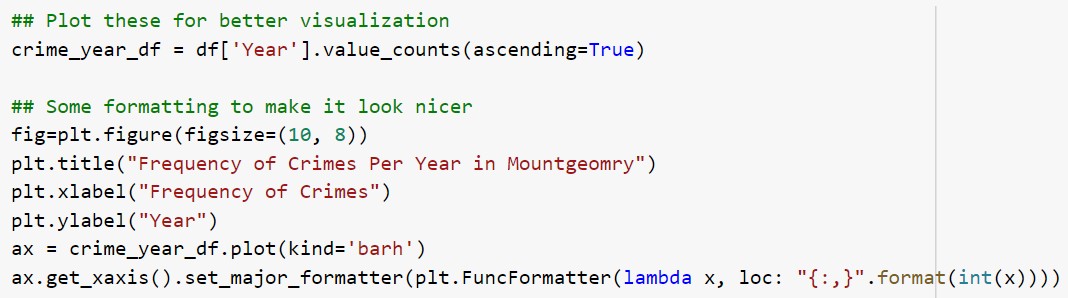
* 1. **CHARACTERISTICS OF DATA SET:**

**Display of data top and bottom**

The code for displaying top and bottom 8 rows and its outputs are as below:

**Basic Plots:**

For the first basic plot, the report looks at frequency of crime rate per year in the Montgomery County. The code for the same is below along with the visualization.



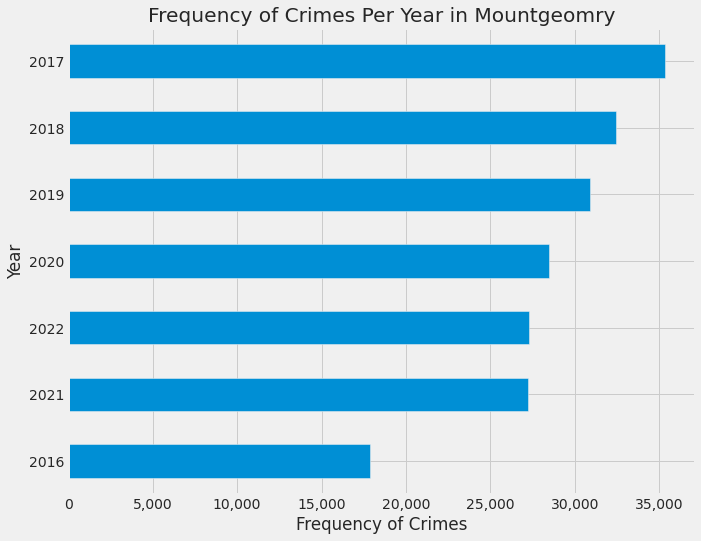
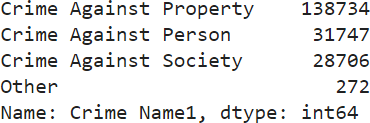
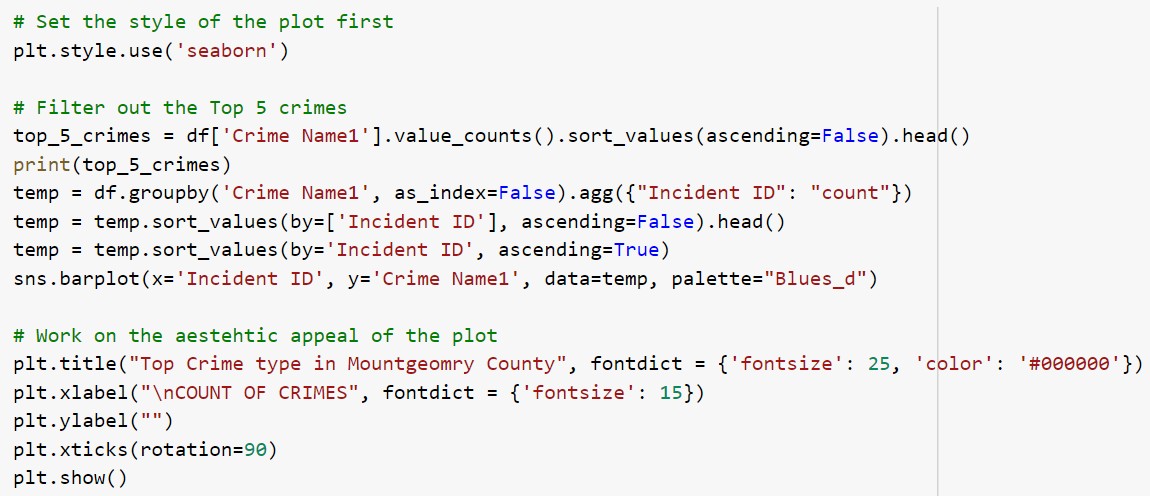


Fig no 14: crime per year bar chart

According to the plotted graphs the occurrence of crimes is the greatest in the year 2017 and with a slight decrease in 2018, 2019 and 2020 consecutively. 2021 and 2022 having equal frequency and 2016 at the lowest crime rate hence the safest year amongst all of the others.

For the next basic plot, analysis is made on the type of crime (Crime Name1), visualization looking at top crime type in Montgomery Count. The code and visualization is as below.



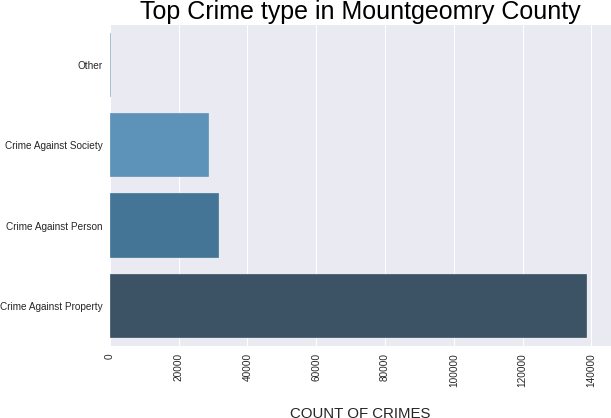
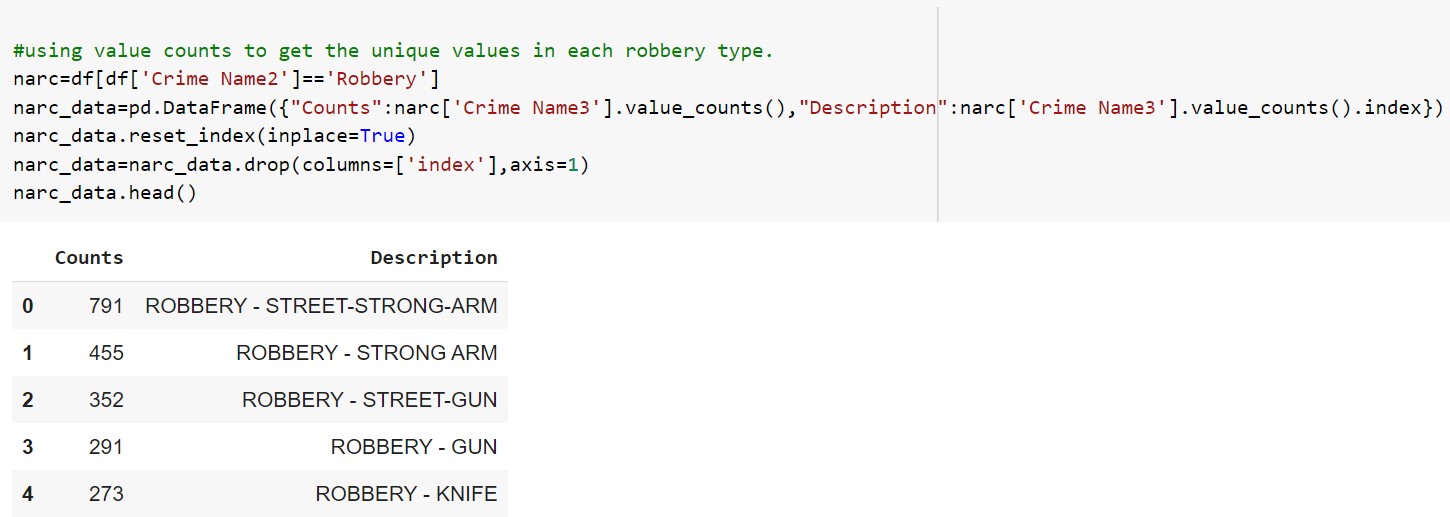


Fig no 15: crime and crime types

The bar graph clearly shows that the most occurrence is crime against property and then comes crime against person and society with a very slight difference between each other.

The next basic plot is a bar graph which represents the crime type robbery. The bar plot has the count on the x-axis and type of weapon used for robbery on the y-axis. Below is the code and visualization for the visualization.



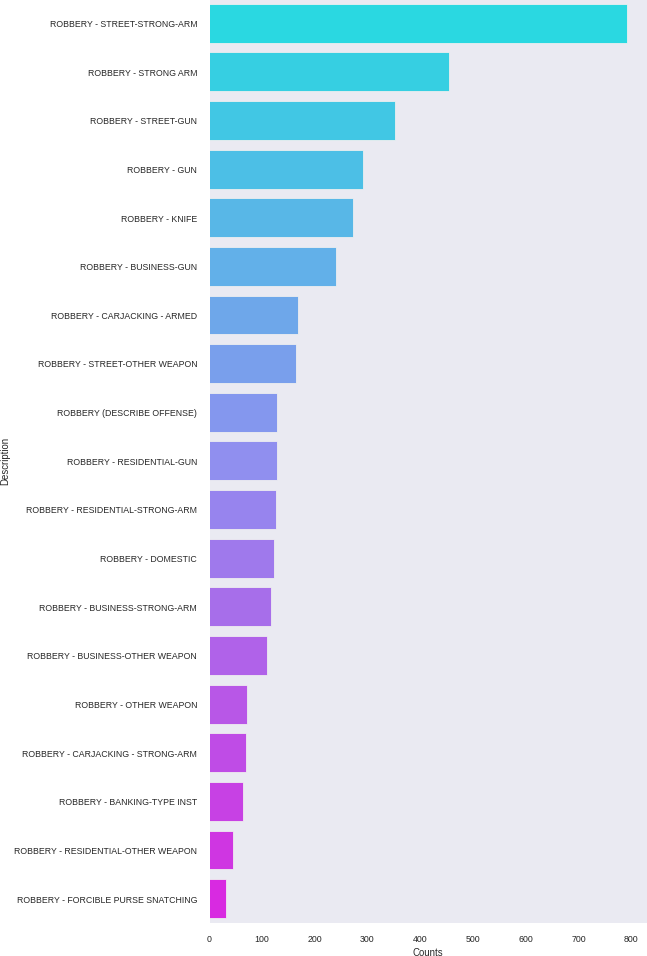


Fig no 16: Types of weapon

It is evident from thr visualization from the grapgh the most robberies are happened on thr street using strong arm, second comes the street guns and the last robbery is forcible per snatching

# Exploratory Data Analysis

* 1. **FINDING NULL VALUES AND MISSING VALUES:**

The df.info() shows the basic count of values. Anything less than 321741 in our data set represents a null value. Below is the code for checking missing values. df.isna() is the basic function used and below is the entire code and output to display the missing values in data for each column. From the above dataset, it is observed that with high number of missing values are related to the geographical location of the crime scene i.e., address, street, block address etc. We have around 865131 null values in our dataset and most of the values are not numeric.

All the null values in Crime Name1, 2 and 3 variables are the same. And around 600 values are missing which is almost less than 0.01% less than the dataset. Hence deleting these values with below code.

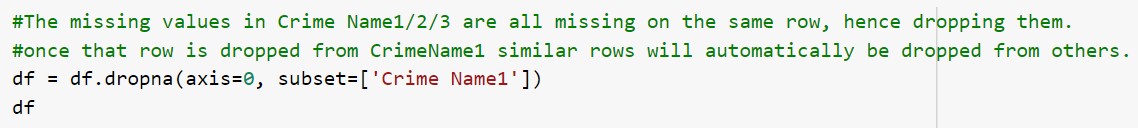


Fig no 17a: Dropping null values

df.isna().sum() is now giving 62 values in police district name and 188 values in Street Type. Since all the values are characters, mean(average), median(middle value) or mode(most repeated) can not give an appropriate value to fill in the null values. Hence going to replace these values with a character called Others, by doing this the authenticity of the dataset is not altered and the visualizations are displayed appropriately. The below code is used to do the same and the output is also displayed where there are no more null or missing values.

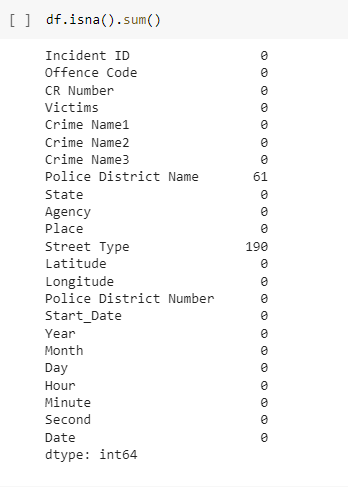


Fig no 17b: Dropping null values

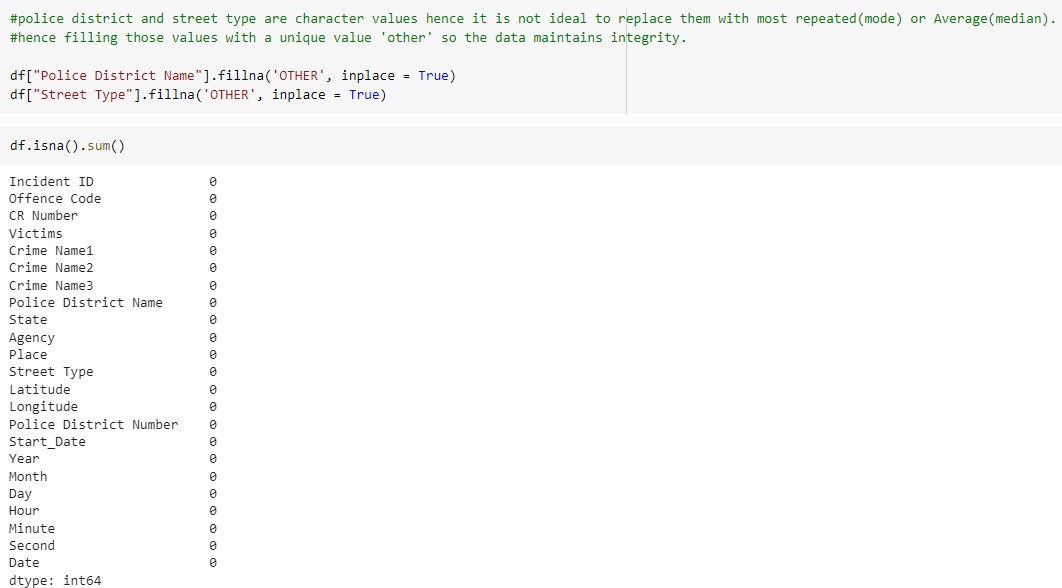


Fig no 18: Missing values

* 1. **CORRELATION STRUCTURES AND DISTRIBUTION CHARACTERISTICS:**

The code df.corr().abs() is a pandas function used to compute pairwise correlation for columns, this excludes the null values. The correlation values defines the link between 2 columns, 1 holds the strongest correlation. Below is the output table and it clearly shows when Incident ID is compared to Incident ID again value is 1 that means it has maximum correlation. It can be observed that the entire data set has all positive values when compared to each other hence there is always a certain amount of correlation between all the columns.



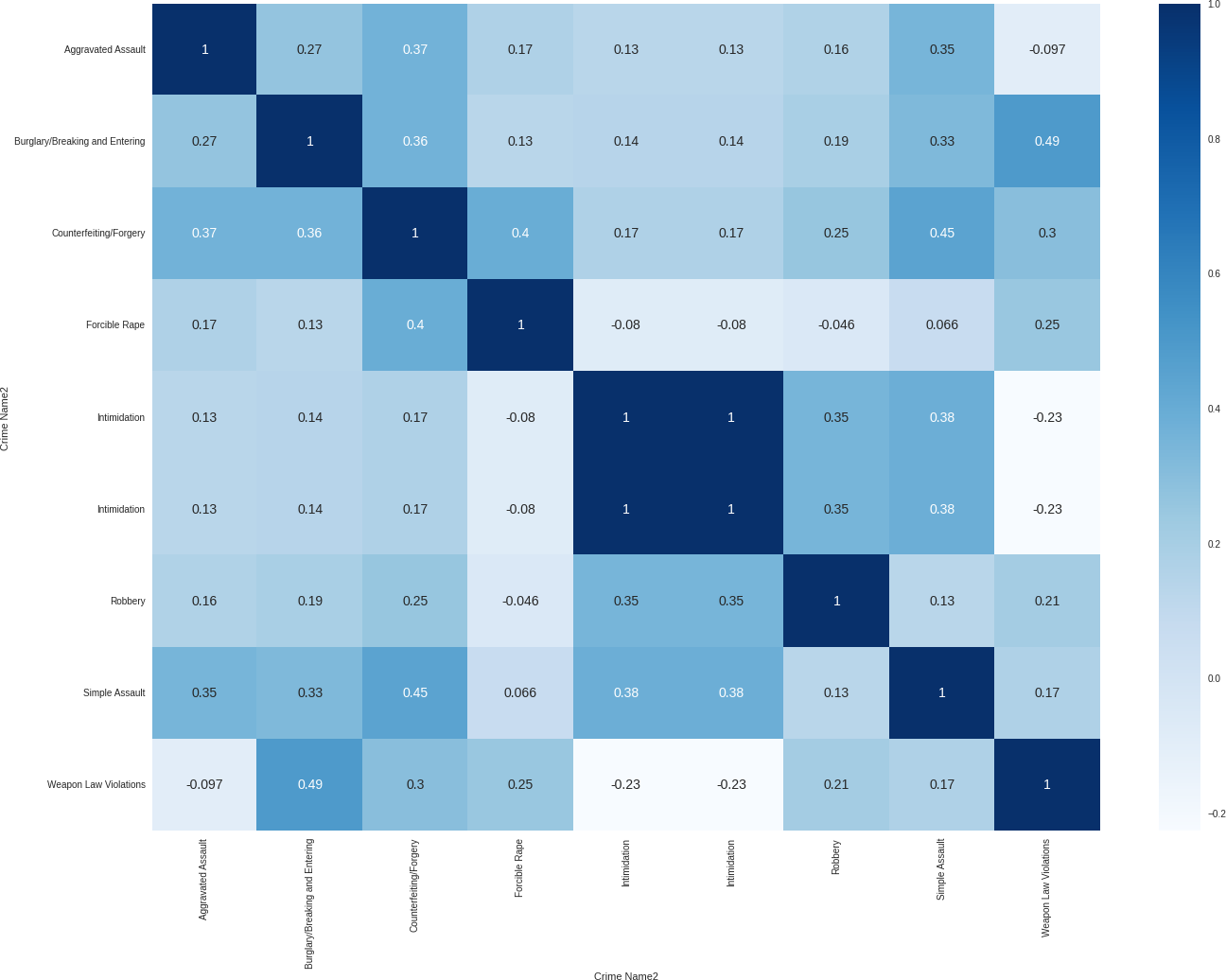
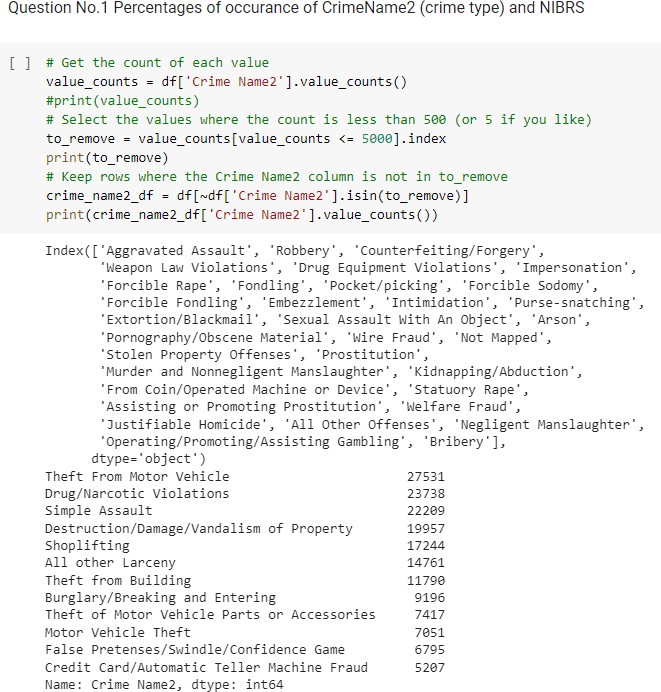


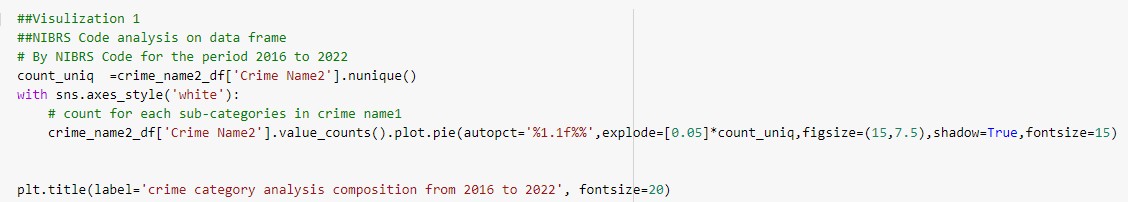
Fig no 19: Correaltion

* 1. **RESEARCH QUESTIONS AND RELATED VISUALIZATIONS:**

**Visualization no 1**

In the first research question, we look at percentages of crimes according to the crime type2 column, this column gives us information on type of crime – ex: robbery, murder etc. Below is the code and output related to the question. Also, a part of code used ‘value\_counts’ a pandas function to understand the count of each crime and display only the most occurring crimes. Here we have taken values less than more than 5000 as the graph gives more clearer and informative that way.





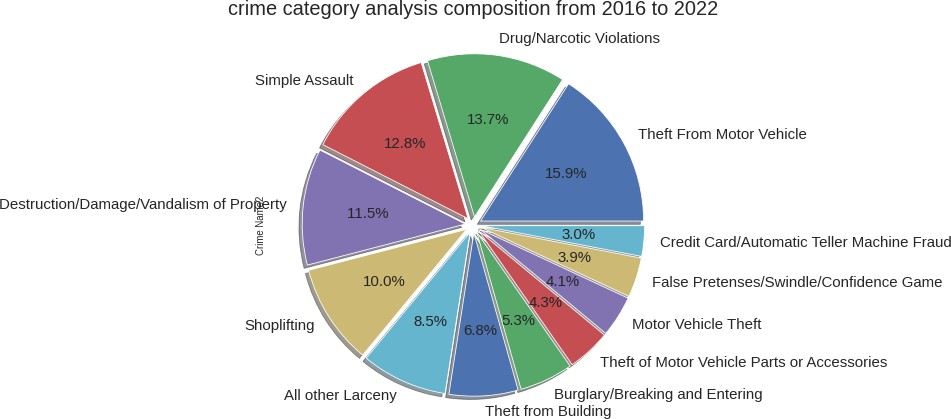


Fig no 20a: Visulization 1.1

The visualization shows the crime category analysis with each colour of the pie representing a different crime. We have used value counts for unique count of variables and plotted them using the below code. The maximum percentage is at 15.9% for theft from motor vehicles. Least at 3- credit card/ATM.

In the second part of the question, we have displayed the percentages of NIBRS A category crimes. Below is the code and the output for the same.

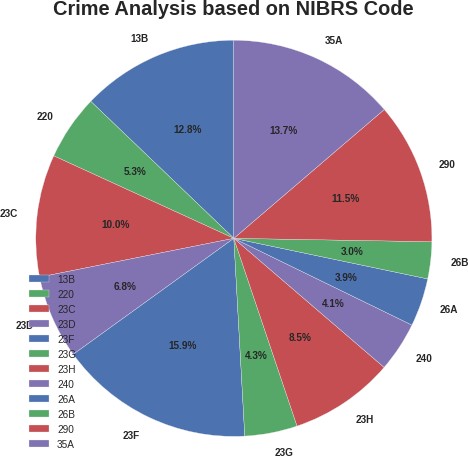
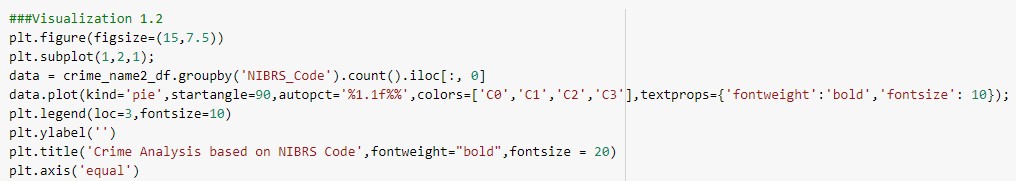


Fig no 21b: Visulization 1.2

The codes are colored with legend at the bottom and codes are displayed with percentages on a pie chart. It is clear from the visualization that the maximum percentage of crime that occurred is with the code 23F with 15.9%. 23F comes under the category crime against property and it is the theft from Motor Vehicles.

Second highest is with 13B- Crime Against Person (Simple Assault) at 12.8%. The least according to our threshold value is 3% for 26B (crime against property- credit card/ ATM fraud), this could be because these directly deal with banks and have utmost security. Both these graphs infer same information but one displays the crime names and the other just the codes.

**Visualization no 2**

The second question talks about graphical representation of which place has what number of crimes. This question uses places column, where in the values of string are separated for certain variables in EDA. Below is the code and visualization to answer the question. Here the values greater than 3000 are plotted for a better understanding.

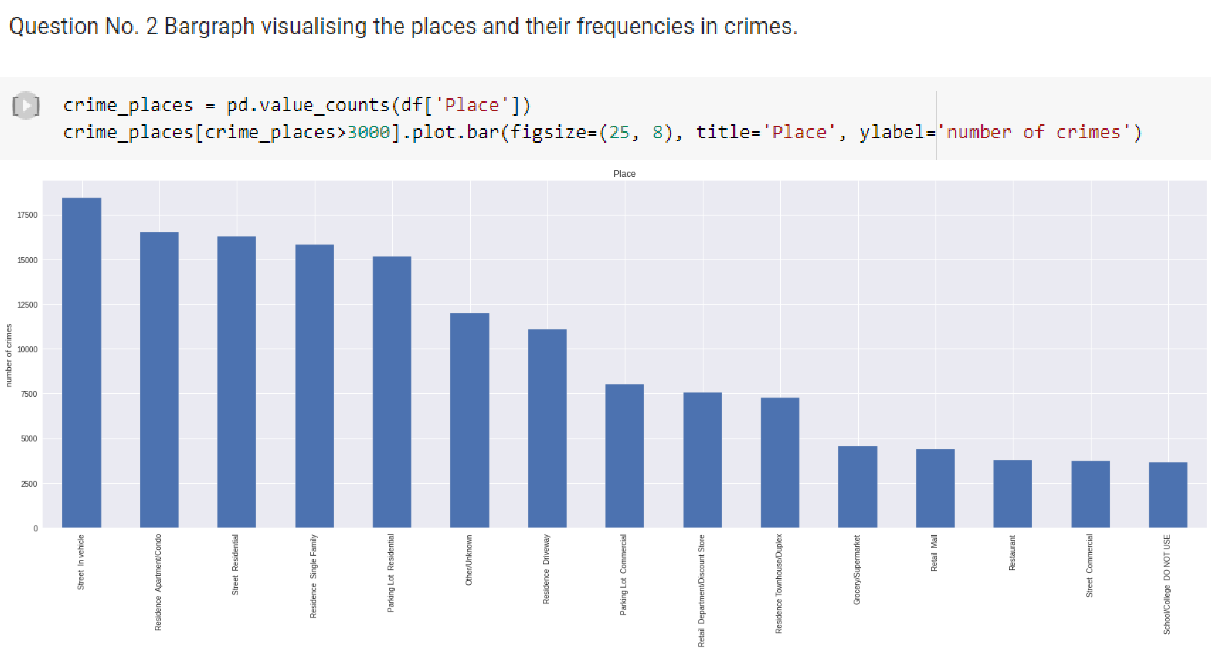


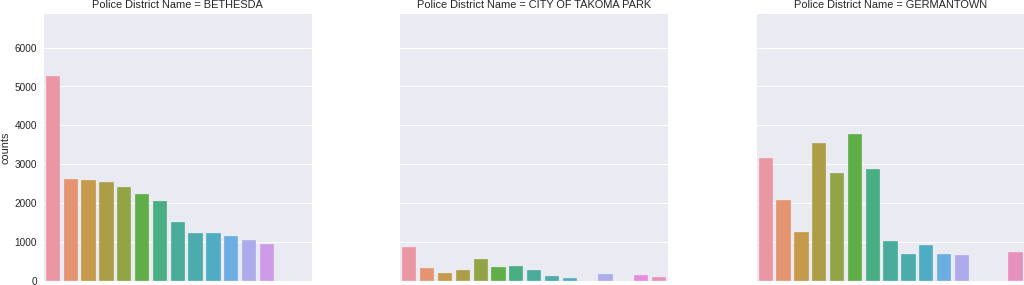
Fig no 22: Visulization 2

The graph plotted is a bar graph with place on the x-axis and number on y-axis. According to the visualization, the maximum number of crimes happened in Street- in the Vehicle (more than 17500). The next highest is with Residence- Apartment/ condo. Least according to the threshold is at the School and Restaurant and street-commercial are pretty much equal in occurrence at approx. 4500.

**Visualization no 3**

Moving on to the third question, we are looking at frequency of crimes according to the police districts. For this analysis we will be using the Police district name column comparing each crime name2 to it. The below code and visualizations explains the research question in detail.





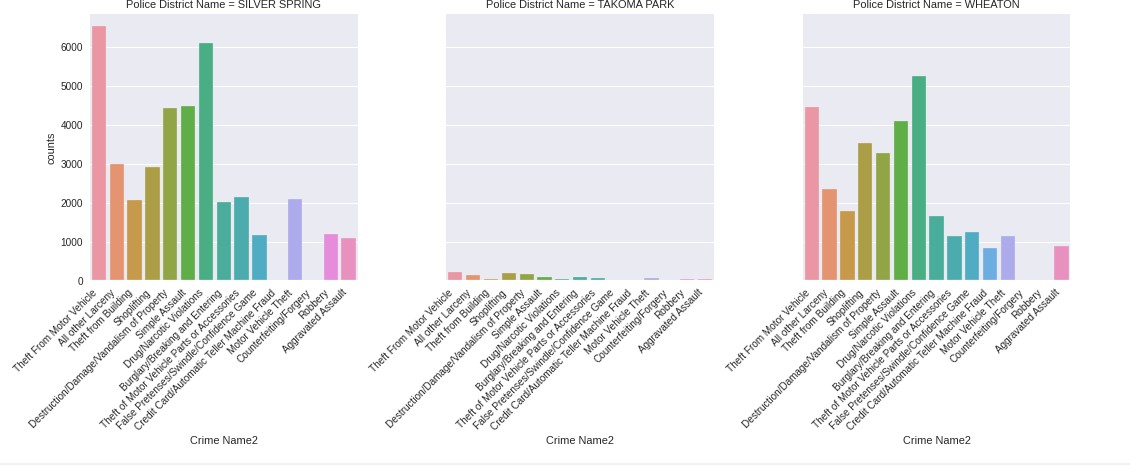


Fig no 23: Visulization 3

The visualization is a collection of different bar graphs, group by function is used to list out and sort the police district and crime names according to the count. Each graph plots crimes according to every police district. In almost all the police districts, theft from motor cycles has the highest occurrence apart from German town which has simple assault and Wheaton which has Durg/Narcotic violations as most occurred crimes. The police districts: City of Takoma Park and Takoma park very less occurrences of crime compared to the others like Bethesda, Montgomery Village and Silver spring.

**Visualization no 4**

Furthermore, the fourth questions code and output interestingly displays the geographical representation using latitude and longitude column values of each crime. Each color on the map represents different crime and when you place the cursor on the dotted color you get the information of crime along with the latitude and longitude value. Below is the code and visualization for the same. Used a plotly library and the data visualization using a scattered plot. This visualization will help look at target areas for each crime and which area has which crimes at the most and least.



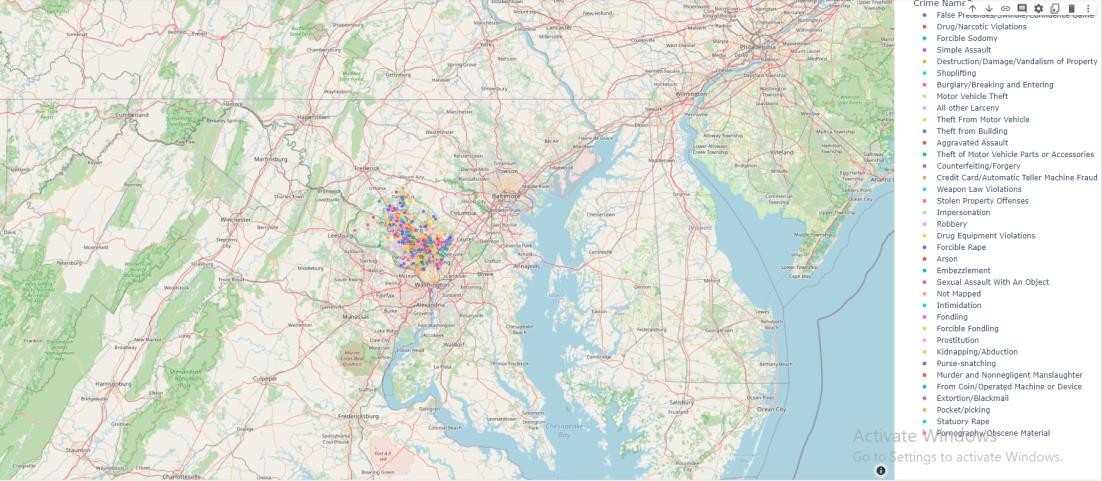


Fig no 24: Visulization 4

**Visualization no 5**

The fifth question looks at the monthly analysis of data across all the years from 2016-2022. The start date column is split into Month, day hour and minute and this data is use to answer the next few questions. For this particular question the month is taken and a bar graph is plotted. The grouped bar graph has months on the x axis (1-January, 2- February and so on) and their occurrence count on y axis, the year here is denoted by colour(hue), and the legend for the year is given on the top right corner of the graph.. The code and visualization is as below.

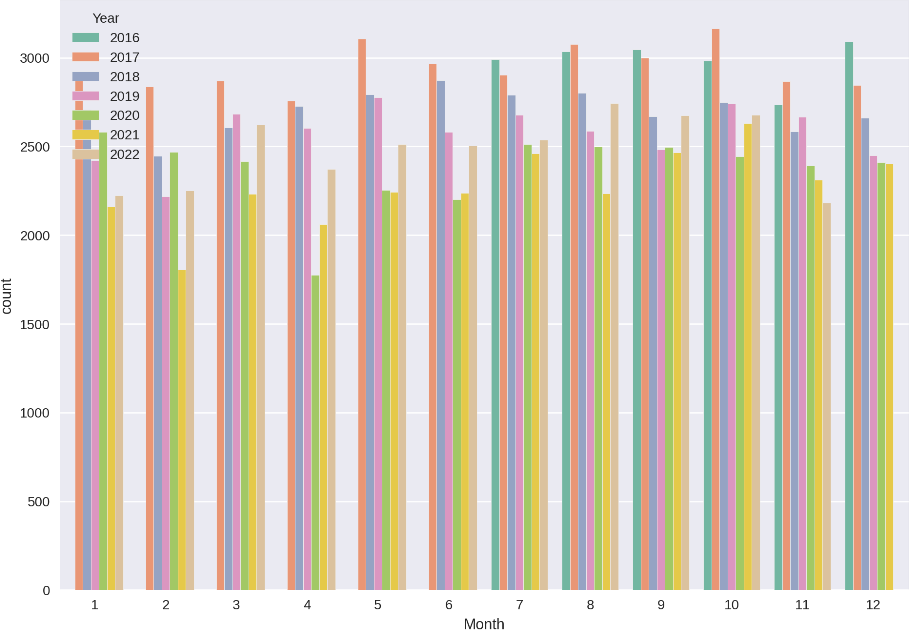
A picture containing company name  Description automatically generated

Fig no 25: Visulization 5

Overall in all the months the highest occurrence is for the year 2017, except for the December(12) month, in which the highest occurrence goes to the year 2016. The least number of occurrences goes to the months February(2) and March(4) with a count of approx. 1750 for the years 2021 and 2020 respectively.

**Visualization no 6**

The sixth question asks for analysis on years from (2016-2022) with crime. To achieve this we considered the column year and CrimeName1 (crime category). The category of crime is of 4 types : crime against person, property, society and others, which is given as a legend. Years are on the x axis and number of occurrences on the y. The code and visualization is as below.

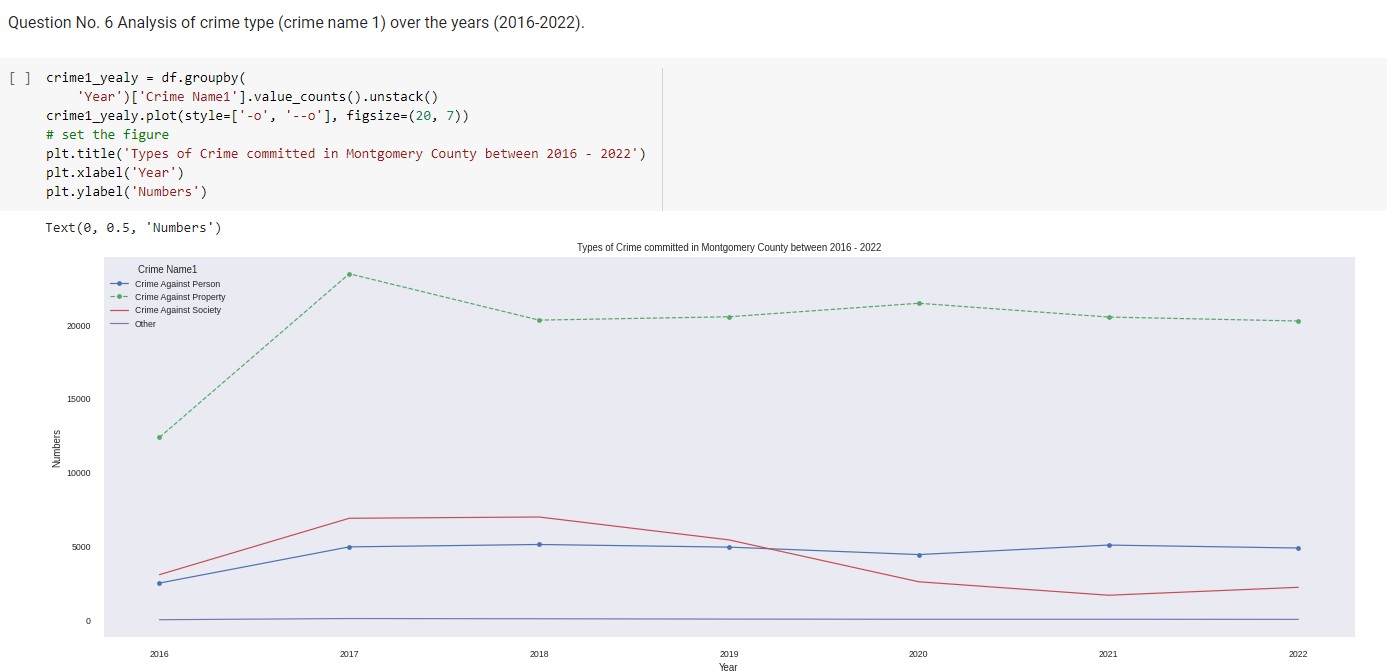


Fig no 26: Visulization 6

The line graph depicts that the highest occurrence all throughout 2016-2022 is crime against property, it rapidly grew from approx. 12000 to 25000 in just one year (2016-2017), and pretty much is stable at around same values till 2022.

Crime against person and society started at close values in 2016. 2019 is when they intersected, from then on till 2020 crime against society decreased and almost flatlined and crime against person is pretty much same overall throughout the years.

**Visualization no 7**

Seventh question talks about time of occurrences of crime in a year. For this the time column is used 0-23 being the hour in the day. The below code and visualization explains the question well.



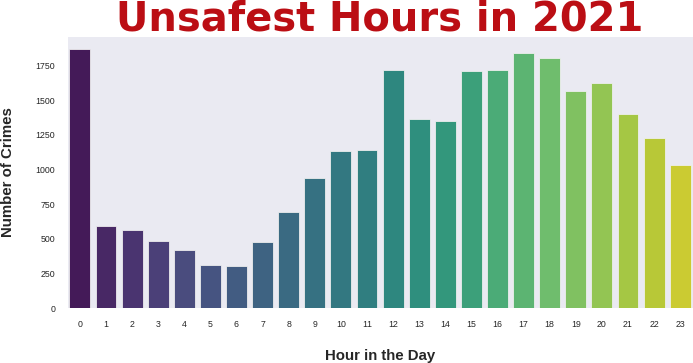


Fig no 27: Visulization 7.1

In the year 2021 the unsafe hour is at 0 (ie, 12 in the midnight) which is on par with 17 (5pm).

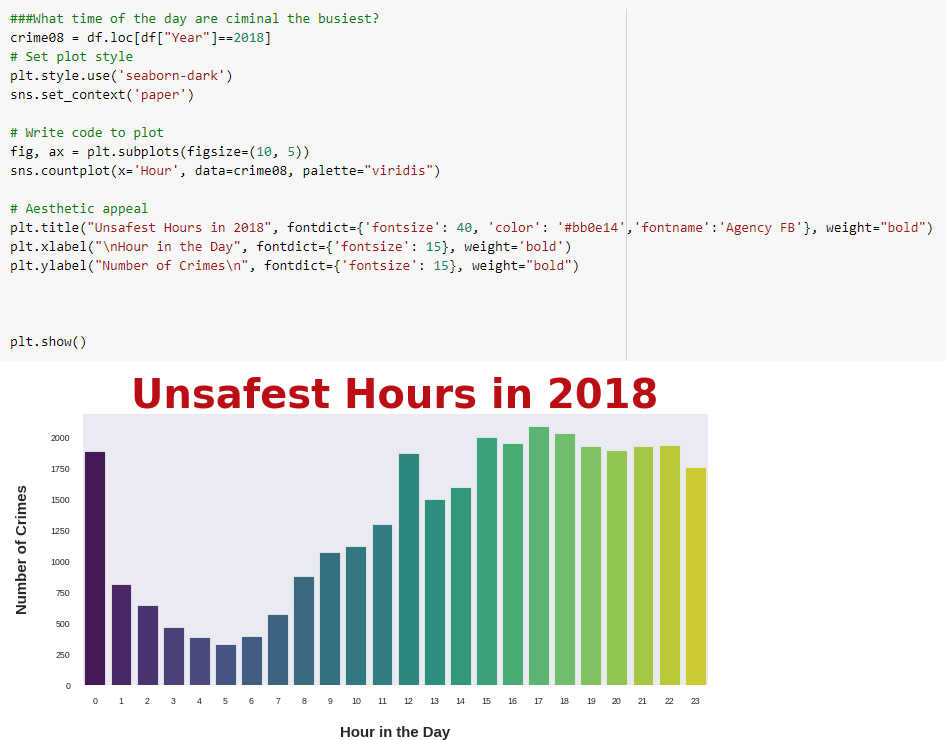


Fig no 27: Visulization 7.2

The unsafe hour in 2018 is again at 17 (5 pm). Its safe to say that early mornings (3-5 am) are the safest amongst the lot.

**Visualization no 8**

The eight question looks at which street type has busiest crime rate. To achieve this we have chosen a bar graph to plot and used groupby function to sort values. Below is the code and visualization for the same.

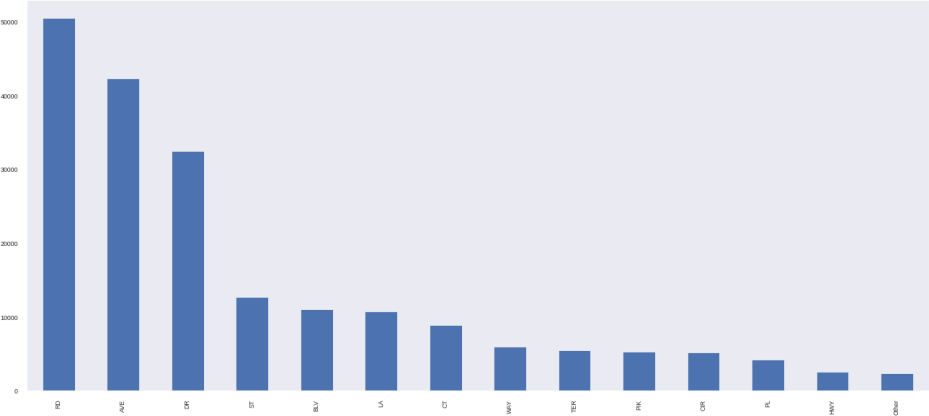
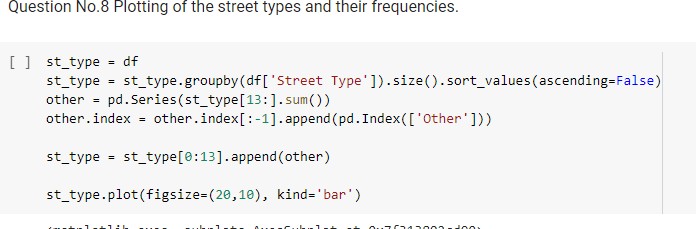


Fig no 28: Visulization 8

The bar graph has street type on the x-axis and count on Y. The column street type is used to make this analysis. According to the graph, RD (road) has most number of occurrences, next come AVE (Avenue) and the least with HWY (highway). This data can be helpful for patrolling services and surveillance.

**Visualization no 9**

Now the next research question is based on the which police agency has the most frequency. A simple Hisplot is used to solve this research question. The x- axis of the plot, marks police agency name and y plots the count.

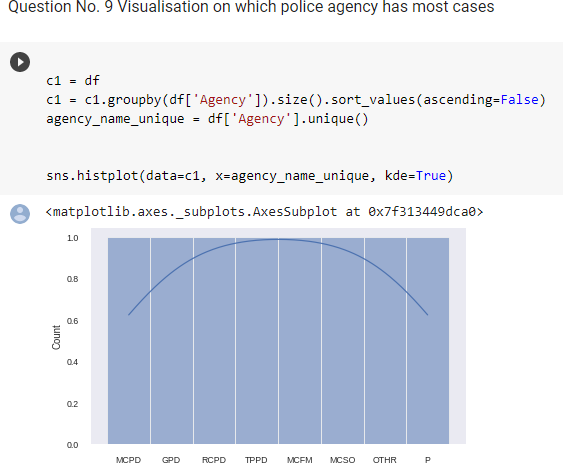
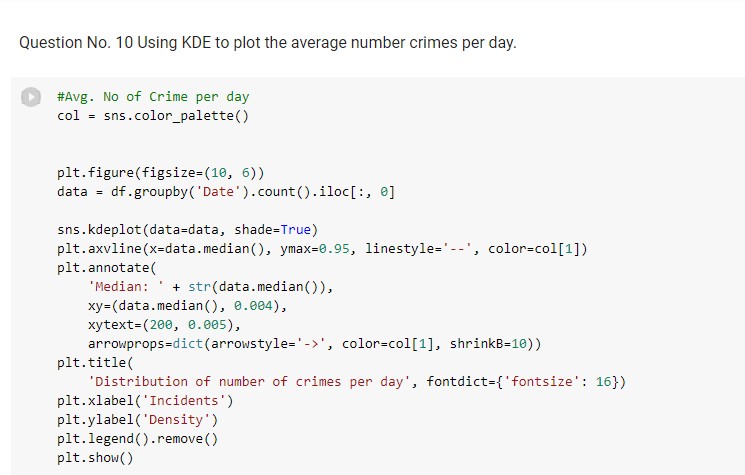


Fig no 29: Visulization 9

The frequency line is normally distributed. TTPD and MCMF have the most count when compared to others as they are at the peak of the curve.

**Visualization no 10**

The final code is as below and the graph shows a distribution of number of crimes per day, using the date column. Used groupby and count for data sorting. The plot used here is Kdeplot, Kernel distribution estimation plot. The code calculated the median and plotted the Kdeplot showing a density curve. The median of the frequency curve is a dotted line in-between the count 75-100, approx. 85. So as a conclusion there are approximately 85 crime occurrences/incidents per day.



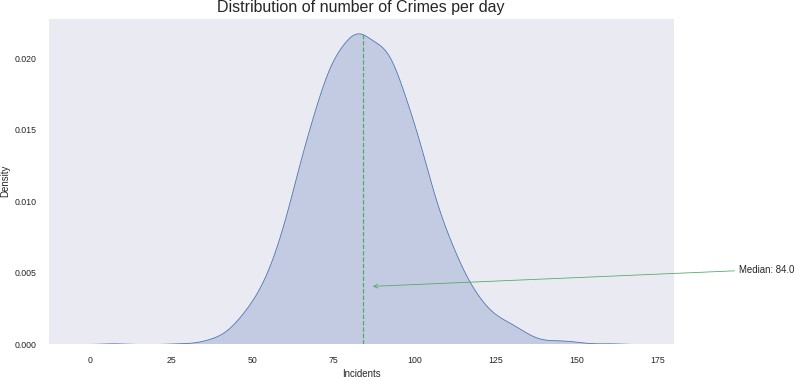


Fig no 30: Visulization 10

# Conclusions

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.