Name: PRAMOD KUMAR NAGARAJ (AIS)

Intermediate python for data science

Final report of

Data explore, analyze, and visualize the CHICAGO CRIME Dataset

Introduction

Project Assigned:

Picking a real-world dataset of our choice and apply the concepts learned in the course intermediate python for data science to perform exploratory data analysis.

Dataset selected for this project:

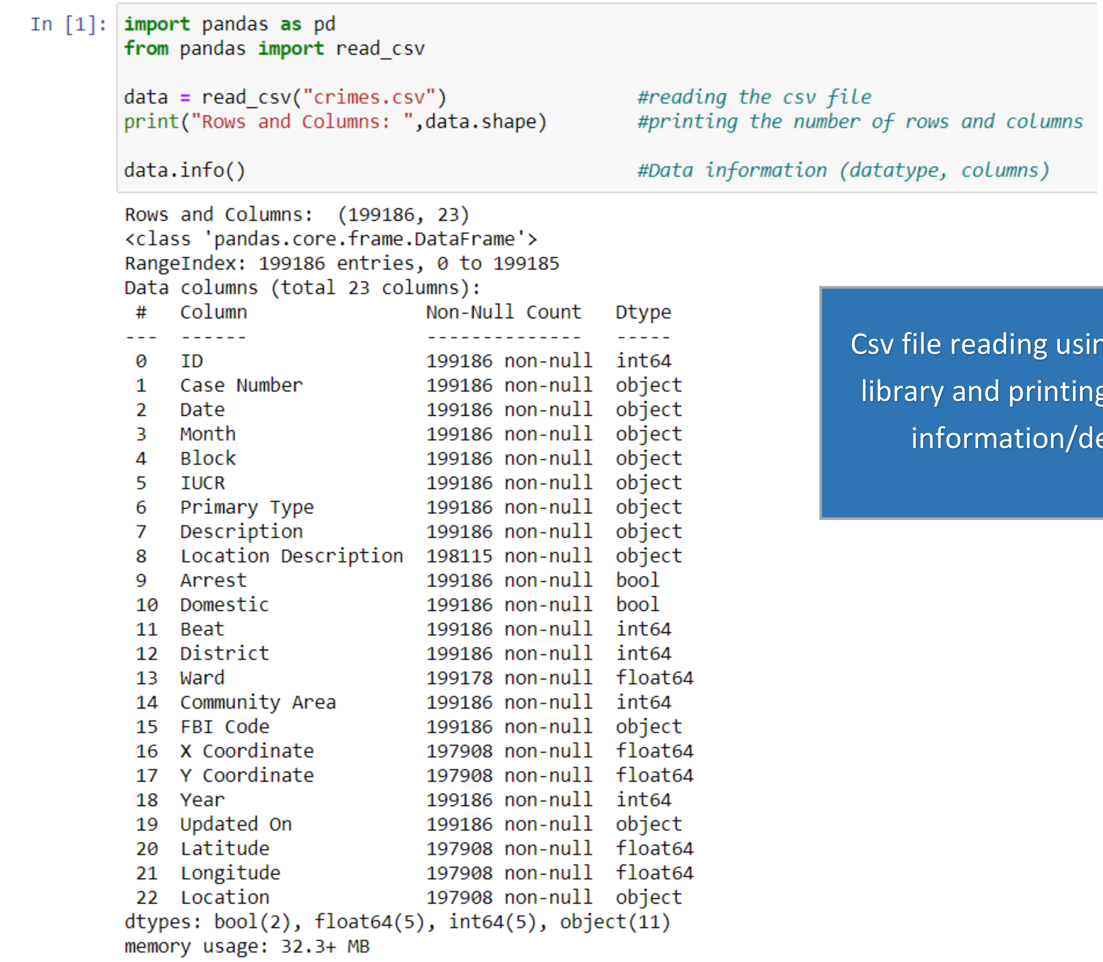
- Chicago crime dataset for the year 2020
- Format .csv
- Size 44 MB
- Original shape: Rows = 199186, Columns = 23
- Location: Chicago crime dataset provided by the Chicago data portal. The Chicago police department has registered numerous criminal cases daily since 2001 and has made this data available publicly in their website. (https://data.cityofchicago.org/Public-Safety/Crimes-2020/qzdf-xmn8)

Now, focusing on the in-depth analysis of the major types of crimes that occurred in the city, observe the trend over the months. Determine which area has the highest crimes based on crimes categories etc.

Python Libraries used:

Pandas	:	dataset read and manipulation operation.
NumPy	:	perform math operation.
Matplotlib	:	to plot graph.
Seaborn	:	to plot advance graph.
Datetime	:	to extract date, time, month, and year.

Data preparation and cleaning

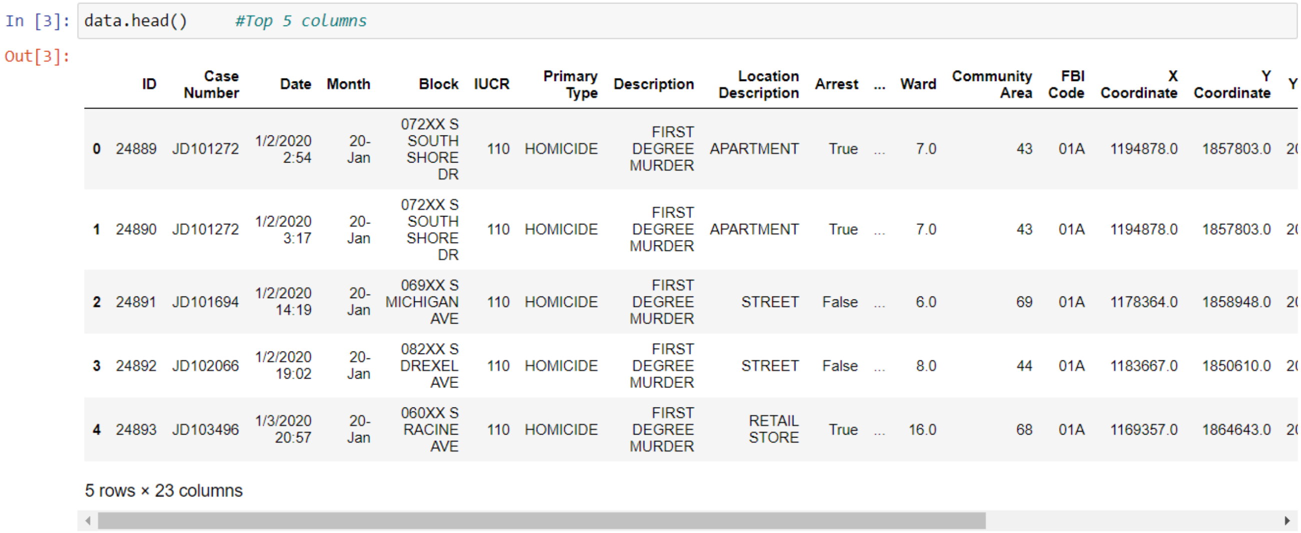


```
In [1]: import pandas as pd
        from pandas import read_csv

        data = read_csv("crimes.csv")              #reading the csv file
        print("Rows and Columns: ",data.shape)     #printing the number of rows and columns

        data.info()                                #Data information (datatype, columns)
```

```
Rows and Columns:  (199186, 23)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199186 entries, 0 to 199185
Data columns (total 23 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   ID                    199186 non-null  int64
 1   Case Number           199186 non-null  object
 2   Date                  199186 non-null  object
 3   Month                 199186 non-null  object
 4   Block                 199186 non-null  object
 5   IUCR                  199186 non-null  object
 6   Primary Type          199186 non-null  object
 7   Description           199186 non-null  object
 8   Location Description  198115 non-null  object
 9   Arrest                199186 non-null  bool
 10  Domestic              199186 non-null  bool
 11  Beat                  199186 non-null  int64
 12  District              199186 non-null  int64
 13  Ward                  199178 non-null  float64
 14  Community Area        199186 non-null  int64
 15  FBI Code              199186 non-null  object
 16  X Coordinate          197908 non-null  float64
 17  Y Coordinate          197908 non-null  float64
 18  Year                  199186 non-null  int64
 19  Updated On            199186 non-null  object
 20  Latitude              197908 non-null  float64
 21  Longitude             197908 non-null  float64
 22  Location              197908 non-null  object
dtypes: bool(2), float64(5), int64(5), object(11)
memory usage: 32.3+ MB
```

Csv file reading using Pandas library and printing dataset information/details

- Dataset first 5 rows



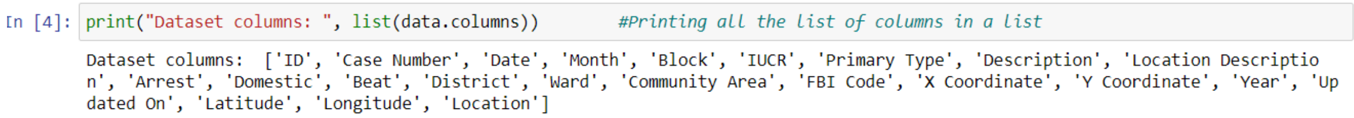
```
In [3]: data.head()      #Top 5 columns
```

Out[3]:

	ID	Case Number	Date	Month	Block	IUCR	Primary Type	Description	Location Description	Arrest	...	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Y
0	24889	JD101272	1/2/2020 2:54	20-Jan	072XX S SOUTH SHORE DR	110	HOMICIDE	FIRST DEGREE MURDER	APARTMENT	True	...	7.0	43	01A	1194878.0	1857803.0	2
1	24890	JD101272	1/2/2020 3:17	20-Jan	072XX S SOUTH SHORE DR	110	HOMICIDE	FIRST DEGREE MURDER	APARTMENT	True	...	7.0	43	01A	1194878.0	1857803.0	2
2	24891	JD101694	1/2/2020 14:19	20-Jan	069XX S MICHIGAN AVE	110	HOMICIDE	FIRST DEGREE MURDER	STREET	False	...	6.0	69	01A	1178364.0	1858948.0	2
3	24892	JD102066	1/2/2020 19:02	20-Jan	082XX S DREXEL AVE	110	HOMICIDE	FIRST DEGREE MURDER	STREET	False	...	8.0	44	01A	1183667.0	1850610.0	2
4	24893	JD103496	1/3/2020 20:57	20-Jan	060XX S RACINE AVE	110	HOMICIDE	FIRST DEGREE MURDER	RETAIL STORE	True	...	16.0	68	01A	1169357.0	1864643.0	2

5 rows × 23 columns

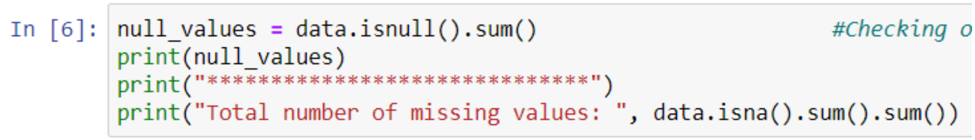
- Dataset columns



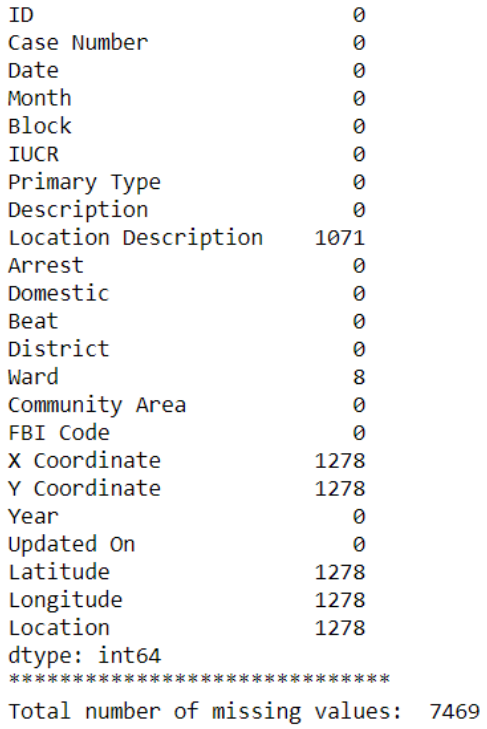
```
In [4]: print("Dataset columns: ", list(data.columns))          #Printing all the list of columns in a list
```

```
Dataset columns:  ['ID', 'Case Number', 'Date', 'Month', 'Block', 'IUCR', 'Primary Type', 'Description', 'Location Descriptio
n', 'Arrest', 'Domestic', 'Beat', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', 'Y Coordinate', 'Year', 'Up
dated On', 'Latitude', 'Longitude', 'Location']
```

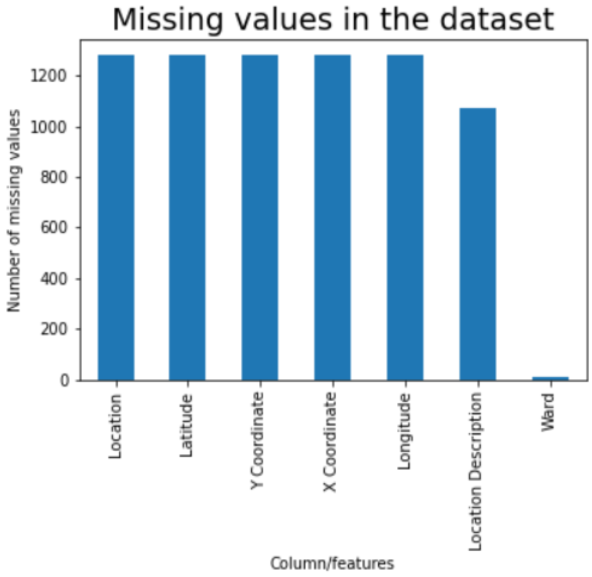
- Dataset null values



```
In [6]: null_values = data.isnull().sum()                              #Checking o
        print(null_values)
        print("******************************")
        print("Total number of missing values: ", data.isna().sum().sum())
```

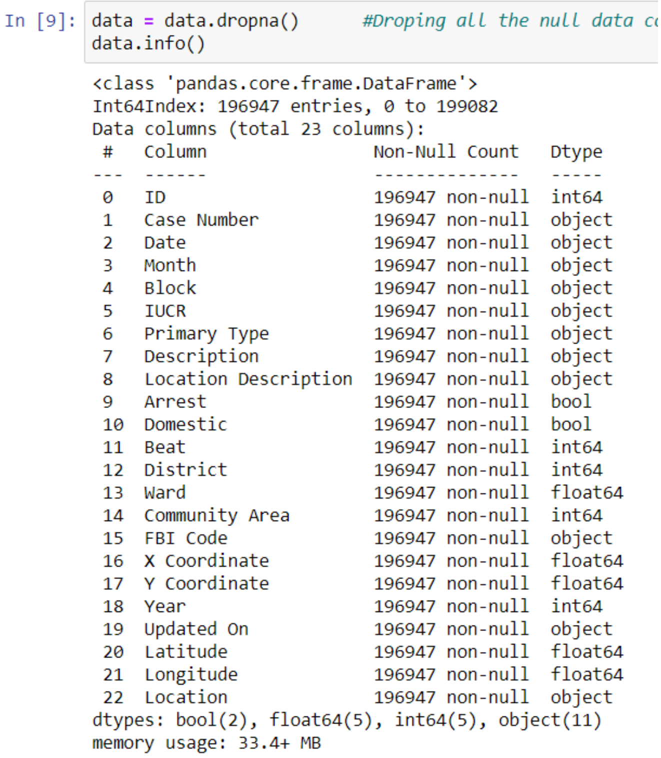


```
ID                          0
Case Number                 0
Date                        0
Month                       0
Block                       0
IUCR                        0
Primary Type                0
Description                 0
Location Description     1071
Arrest                      0
Domestic                    0
Beat                        0
District                    0
Ward                        8
Community Area              0
FBI Code                    0
X Coordinate             1278
Y Coordinate             1278
Year                        0
Updated On                  0
Latitude                 1278
Longitude                1278
Location                 1278
dtype: int64
******************************
Total number of missing values:   7469
```



- Dealing with null dataset values

Applied dropout method for null columns.

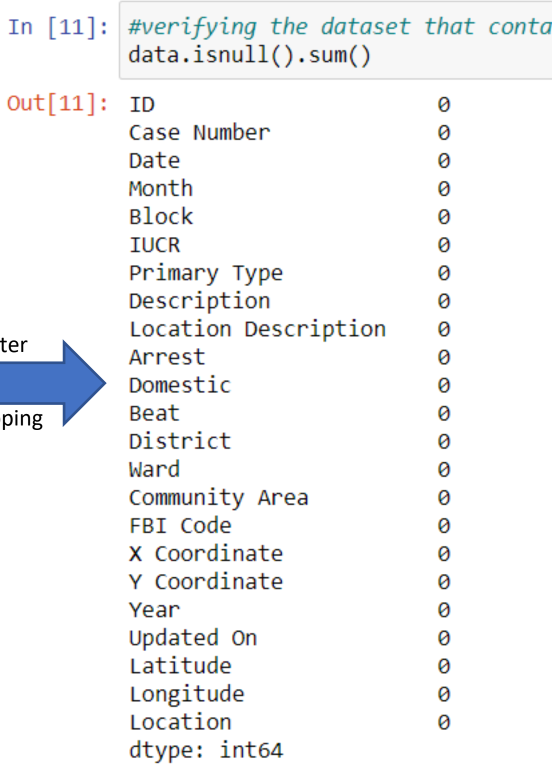


```
In [9]: data = data.dropna()         #Droping all the null data co
        data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 196947 entries, 0 to 199082
Data columns (total 23 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   ID                    196947 non-null  int64
 1   Case Number           196947 non-null  object
 2   Date                  196947 non-null  object
 3   Month                 196947 non-null  object
 4   Block                 196947 non-null  object
 5   IUCR                  196947 non-null  object
 6   Primary Type          196947 non-null  object
 7   Description           196947 non-null  object
 8   Location Description  196947 non-null  object
 9   Arrest                196947 non-null  bool
 10  Domestic              196947 non-null  bool
 11  Beat                  196947 non-null  int64
 12  District              196947 non-null  int64
 13  Ward                  196947 non-null  float64
 14  Community Area        196947 non-null  int64
 15  FBI Code              196947 non-null  object
 16  X Coordinate          196947 non-null  float64
 17  Y Coordinate          196947 non-null  float64
 18  Year                  196947 non-null  int64
 19  Updated On            196947 non-null  object
 20  Latitude              196947 non-null  float64
 21  Longitude             196947 non-null  float64
 22  Location              196947 non-null  object
dtypes: bool(2), float64(5), int64(5), object(11)
memory usage: 33.4+ MB
```

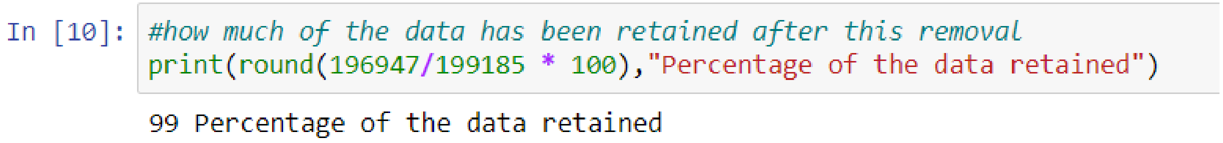
After

Dropping

Result of dropout method.



```
In [11]: #verifying the dataset that conta
         data.isnull().sum()

Out[11]: ID                      0
         Case Number             0
         Date                    0
         Month                   0
         Block                   0
         IUCR                    0
         Primary Type            0
         Description             0
         Location Description    0
         Arrest                  0
         Domestic                0
         Beat                    0
         District                0
         Ward                    0
         Community Area          0
         FBI Code                0
         X Coordinate            0
         Y Coordinate            0
         Year                    0
         Updated On              0
         Latitude                0
         Longitude               0
         Location                0
         dtype: int64
```



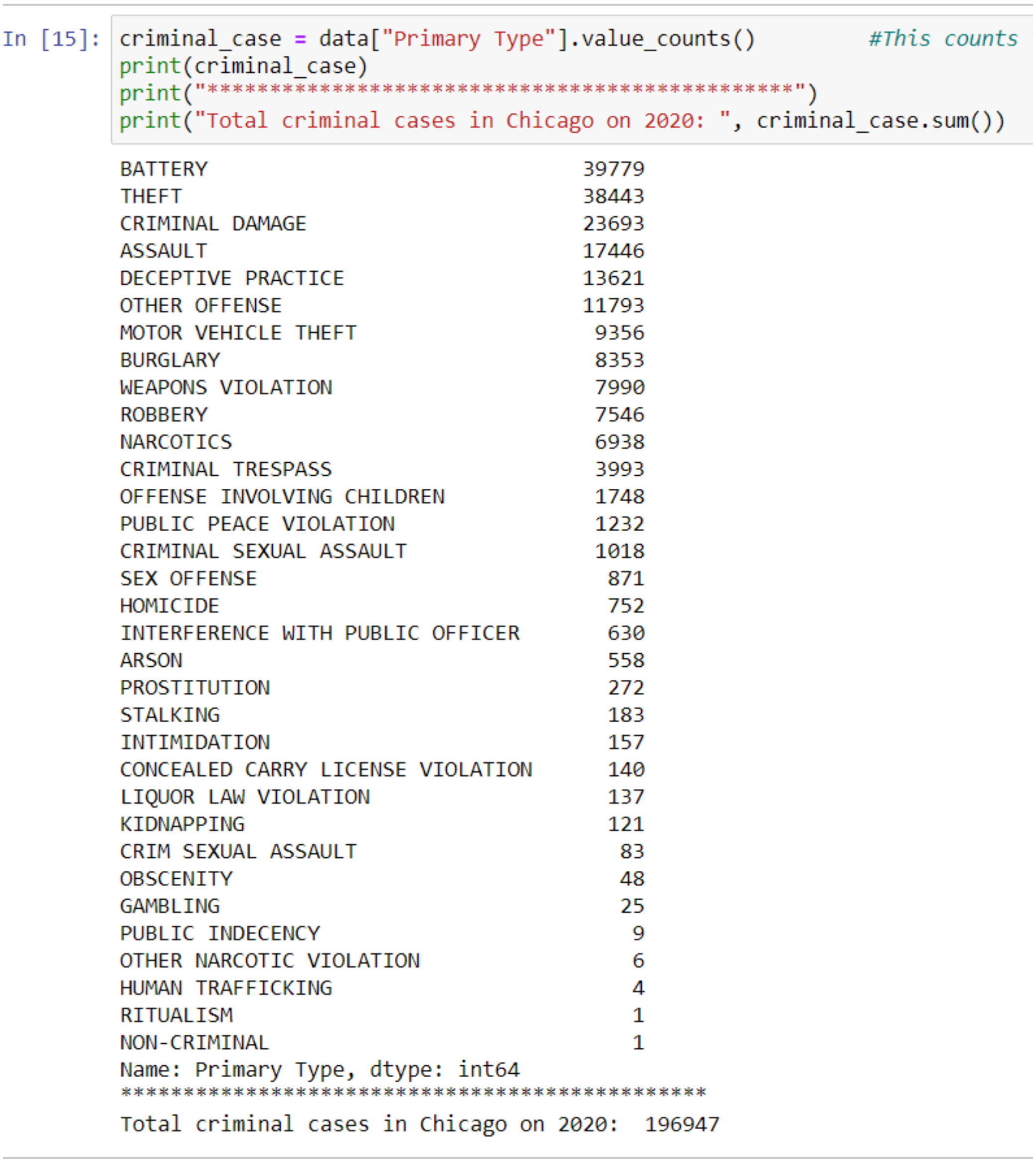
```
In [10]: #how much of the data has been retained after this removal
         print(round(196947/199185 * 100),"Percentage of the data retained")

         99 Percentage of the data retained
```

Dropping the rows will usually result in clean datasets and produce will-behaved data. But often, it removes a lot of information that reduces result acccuracy. However in our case since **99%** of the data is retained and hence there is practically no other way to work around the type of missing values we have.

- **Drop method**: dropping the row with at least one missing value.
- **Isnull()**: Pandas is null is one of package which check the null data in dataset.
- **sum():** NumPy package which help in summation (Math operation)

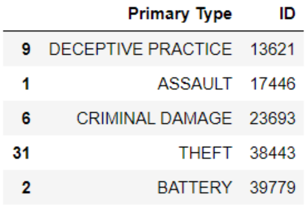
Perform exploratory analysis and visualization.

- Total criminal cases in Chicago city

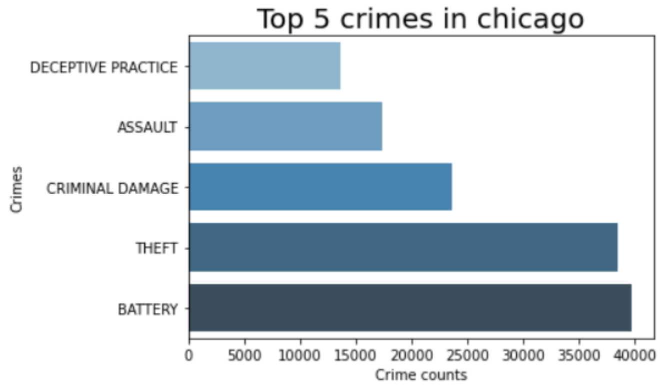


```
In [15]: criminal_case = data["Primary Type"].value_counts()        #This counts
         print(criminal_case)
         print("***********************************************")
         print("Total criminal cases in Chicago on 2020: ", criminal_case.sum())

         BATTERY                              39779
         THEFT                                38443
         CRIMINAL DAMAGE                      23693
         ASSAULT                              17446
         DECEPTIVE PRACTICE                   13621
         OTHER OFFENSE                        11793
         MOTOR VEHICLE THEFT                   9356
         BURGLARY                              8353
         WEAPONS VIOLATION                     7990
         ROBBERY                               7546
         NARCOTICS                             6938
         CRIMINAL TRESPASS                     3993
         OFFENSE INVOLVING CHILDREN            1748
         PUBLIC PEACE VIOLATION                1232
         CRIMINAL SEXUAL ASSAULT               1018
         SEX OFFENSE                            871
         HOMICIDE                               752
         INTERFERENCE WITH PUBLIC OFFICER       630
         ARSON                                  558
         PROSTITUTION                           272
         STALKING                               183
         INTIMIDATION                           157
         CONCEALED CARRY LICENSE VIOLATION      140
         LIQUOR LAW VIOLATION                   137
         KIDNAPPING                             121
         CRIM SEXUAL ASSAULT                     83
         OBSCENITY                               48
         GAMBLING                                25
         PUBLIC INDECENCY                         9
         OTHER NARCOTIC VIOLATION                 6
         HUMAN TRAFFICKING                        4
         RITUALISM                                1
         NON-CRIMINAL                             1
         Name: Primary Type, dtype: int64
         ***********************************************
         Total criminal cases in Chicago on 2020:  196947
```

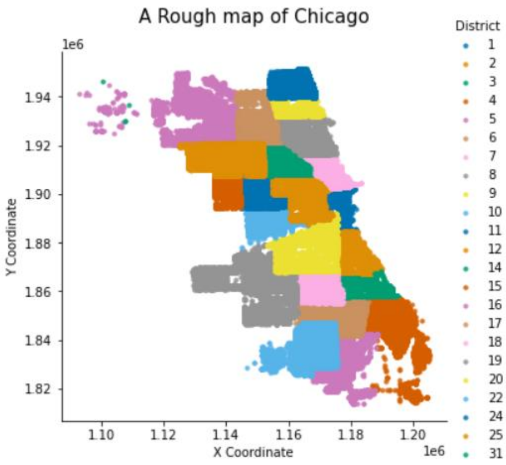
- Top 5 crimes in Chicago city in 2020

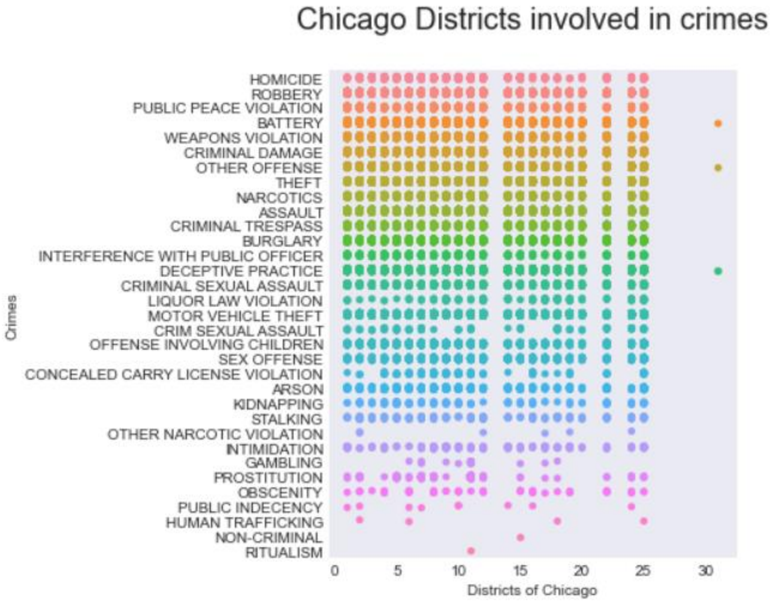


	Primary Type	ID
9	DECEPTIVE PRACTICE	13621
1	ASSAULT	17446
6	CRIMINAL DAMAGE	23693
31	THEFT	38443
2	BATTERY	39779



- Crimes in chicago city districts wise

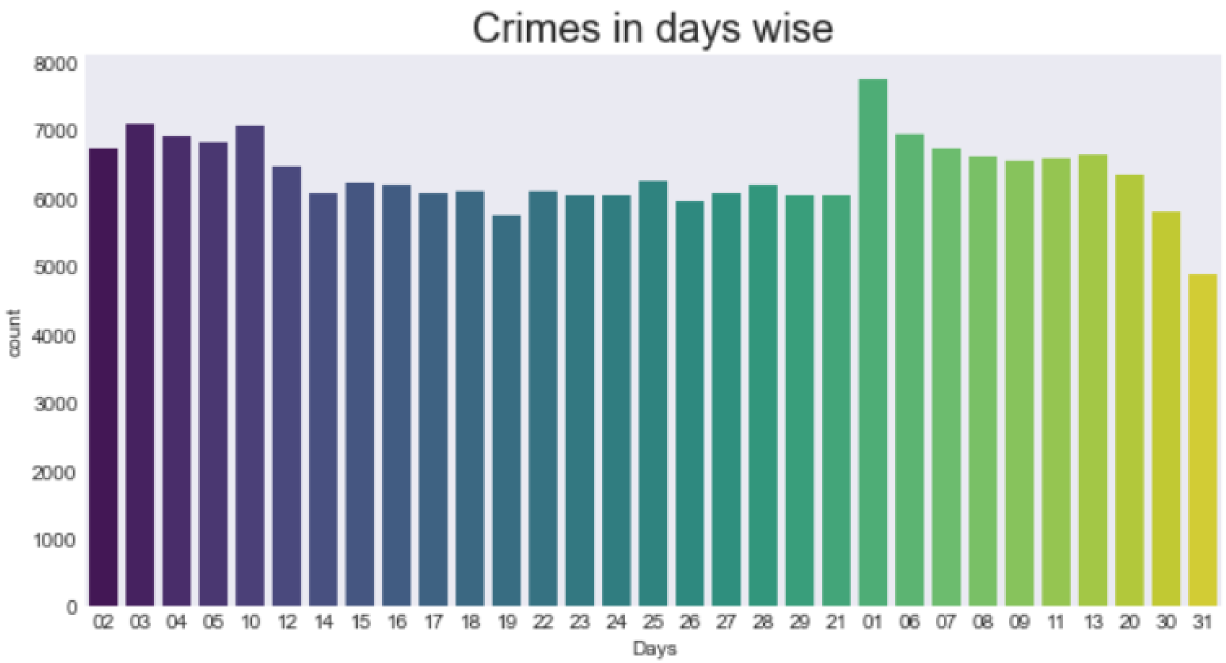




- ➢ The seaborn “cat-plot” gives the information about the Chicago districts involved in crimes.
- ➢ Graphs is clearly showing that some of Chicago districts were not involved in the crimes they are 12,13,22,24, and 26 to 30. By this we can say those districts are safest for the publics.
- ➢ And some districts are not involved in some specific crimes.

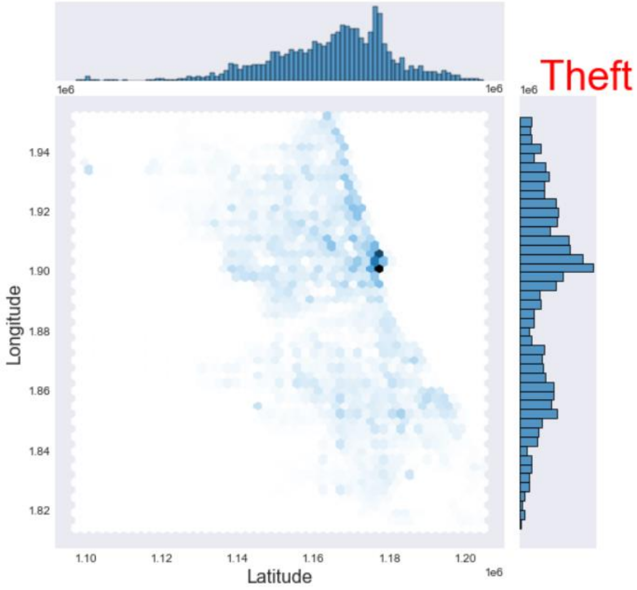
  Example: Other narcotics violation crimes are not taking place in 3 to 11 Chicago districts

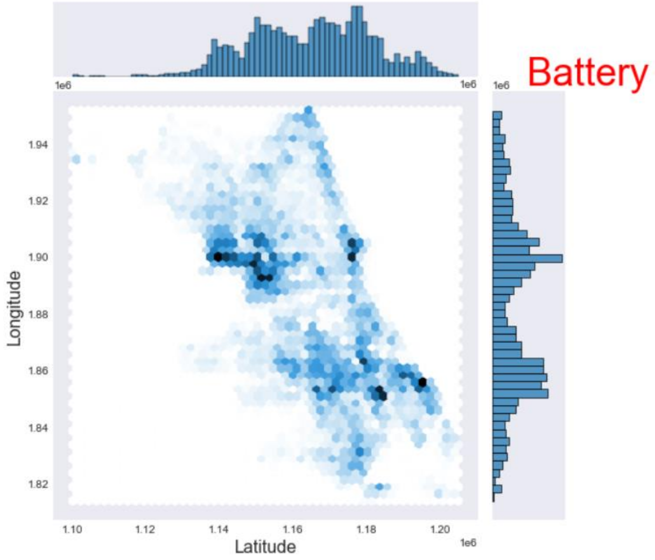
- Crimes based on hours in Chicago city.



➢ By this graph we can say that there is no safest day in Chicago city, the criminals are maintaining their consistency of crimes each day of every month. They do not have any weekends or rest day for their work.

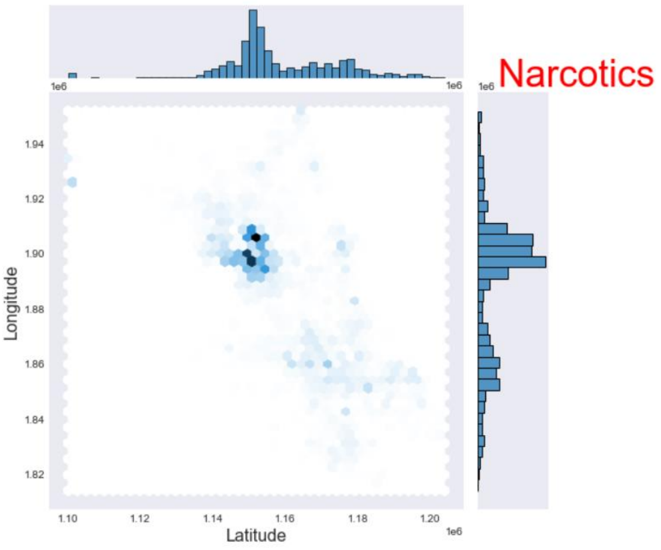
- Some specific crimes and their target location.

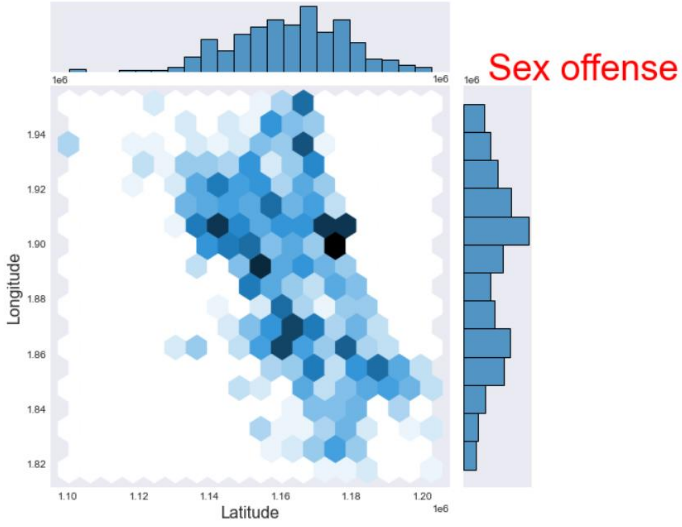




➢ Theft is spread across Chicago with a large concentration in mid-east of Chicago.

➢ Battery crimes have no exclusive localization.

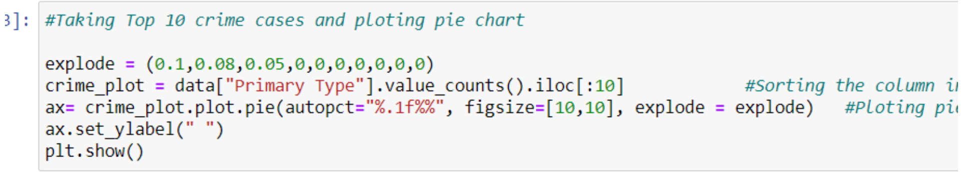




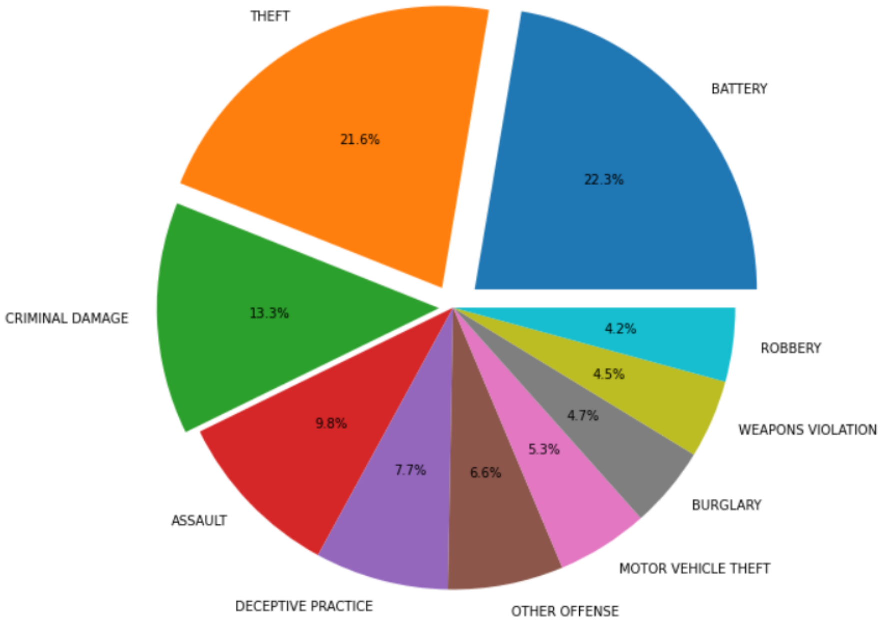
- ➢ Narcotics highly prevalent in the western part. This gives a hit of narcotics gang pin point.
- ➢ Sex offense cases are below 1000 but they are not localized.

Questions and Answers.

1. Which top 10 crimes that occurred in 2020:

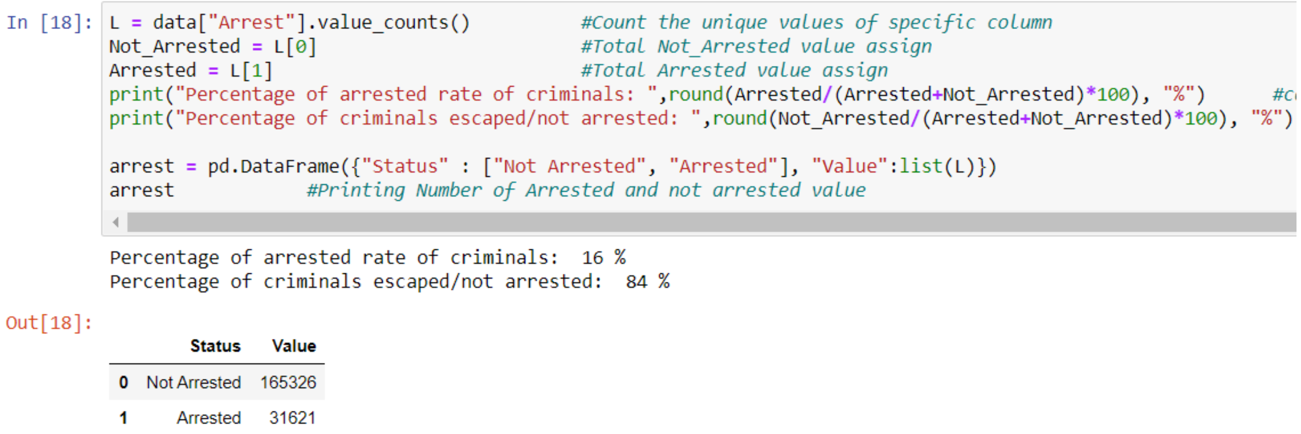


```
3]: #Taking Top 10 crime cases and ploting pie chart

    explode = (0.1,0.08,0.05,0,0,0,0,0,0,0)
    crime_plot = data["Primary Type"].value_counts().iloc[:10]          #Sorting the column i
    ax= crime_plot.plot.pie(autopct="%.1f%%", figsize=[10,10], explode = explode)   #Ploting pi
    ax.set_ylabel(" ")
    plt.show()
```



- Battery was the most occurring crime with a count of 39779 and almost 23% from total crimes
- And 2nd highest is Theft with a count of 38443 and almost 22% from total crimes
- Then followed by criminal damage, assault and deceptive practice and others.

2. Arrests in the city of Chicago by months

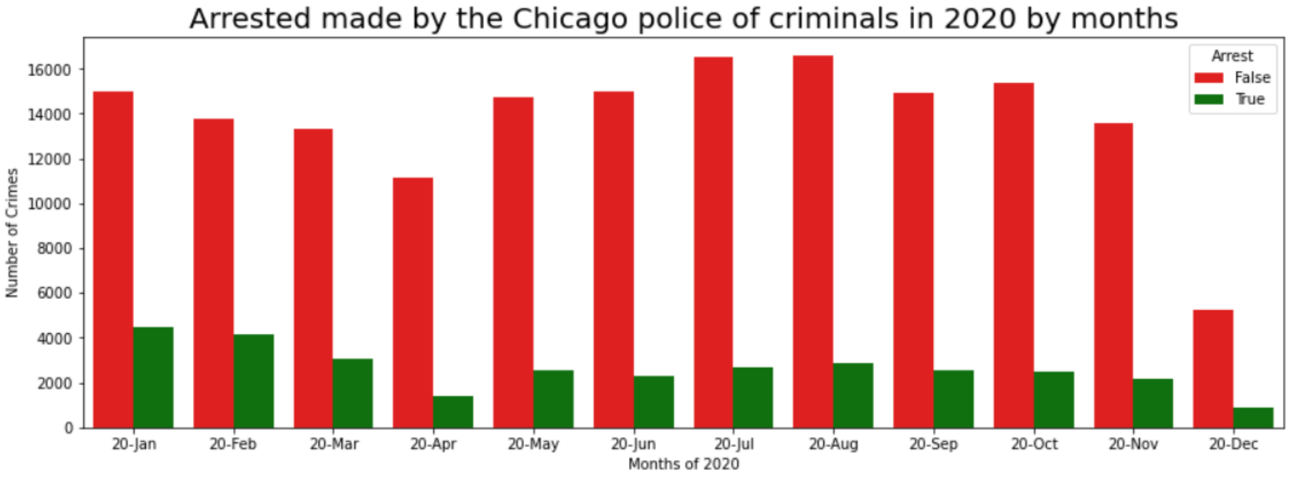


In [18]:
```
L = data["Arrest"].value_counts()          #Count the unique values of specific column
Not_Arrested = L[0]                        #Total Not_Arrested value assign
Arrested = L[1]                            #Total Arrested value assign
print("Percentage of arrested rate of criminals: ",round(Arrested/(Arrested+Not_Arrested)*100), "%")      #c
print("Percentage of criminals escaped/not arrested: ",round(Not_Arrested/(Arrested+Not_Arrested)*100), "%")

arrest = pd.DataFrame({"Status" : ["Not Arrested", "Arrested"], "Value":list(L)})
arrest            #Printing Number of Arrested and not arrested value
```

```
Percentage of arrested rate of criminals:  16 %
Percentage of criminals escaped/not arrested:  84 %
```

Out[18]:

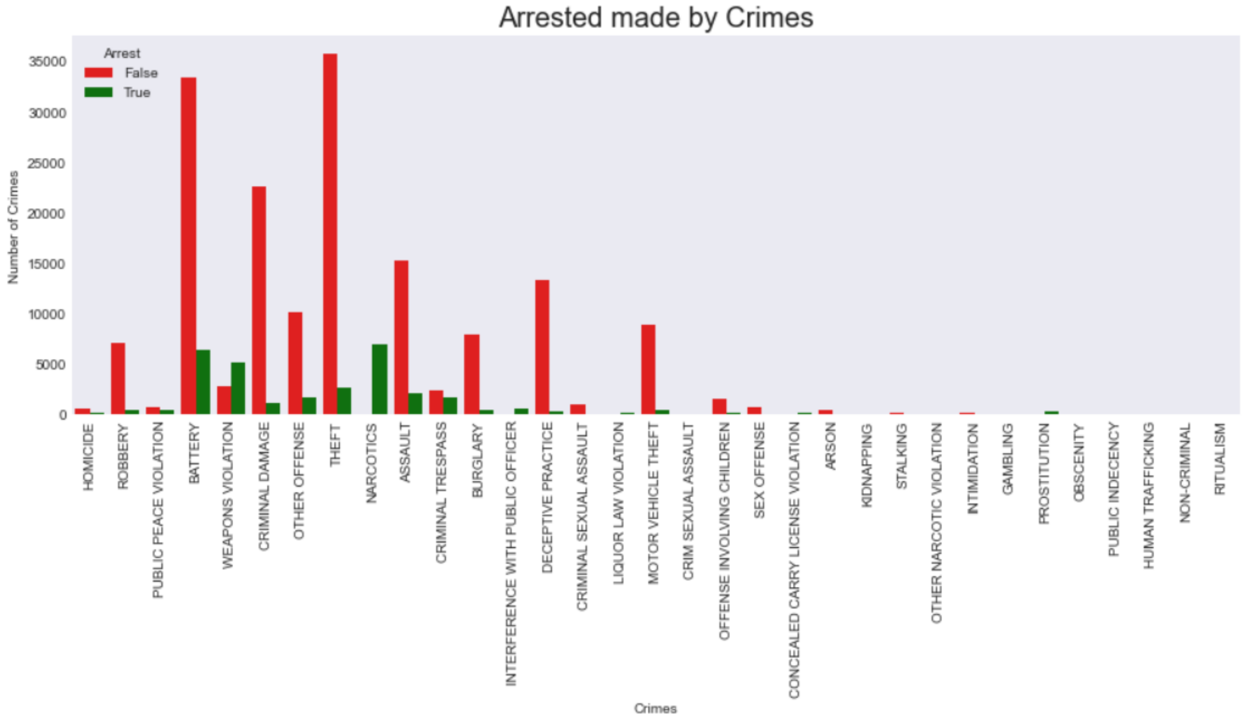
	Status	Value
0	Not Arrested	165326
1	Arrested	31621



Arrested made by the Chicago police of criminals in 2020 by months

- 84% of the crimes were not been arrested due to some reasons
- Since Arresting the criminals is very low- less than 20%, we can say this is one of the reasons for high crimes rates in Chicago city
- From the graph we can see July and August month has high crime rates and low arrest rate of criminals

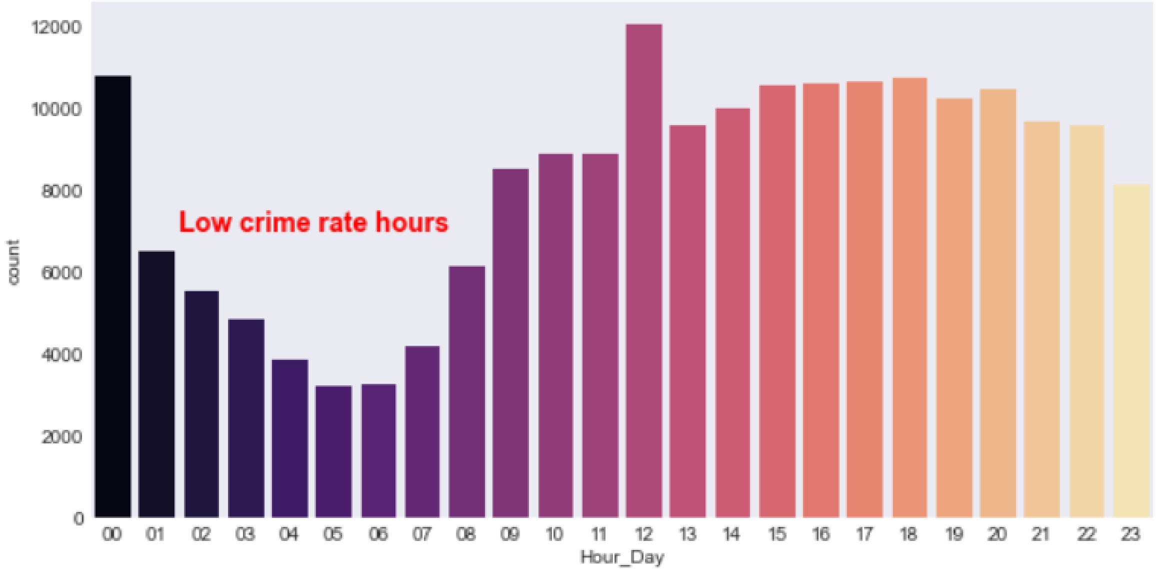
3. Arrests in the city of Chicago by crimes.



- The right-end crimes are not shown showing in the graph because the crime rate is lower than 500 so it is not visible.

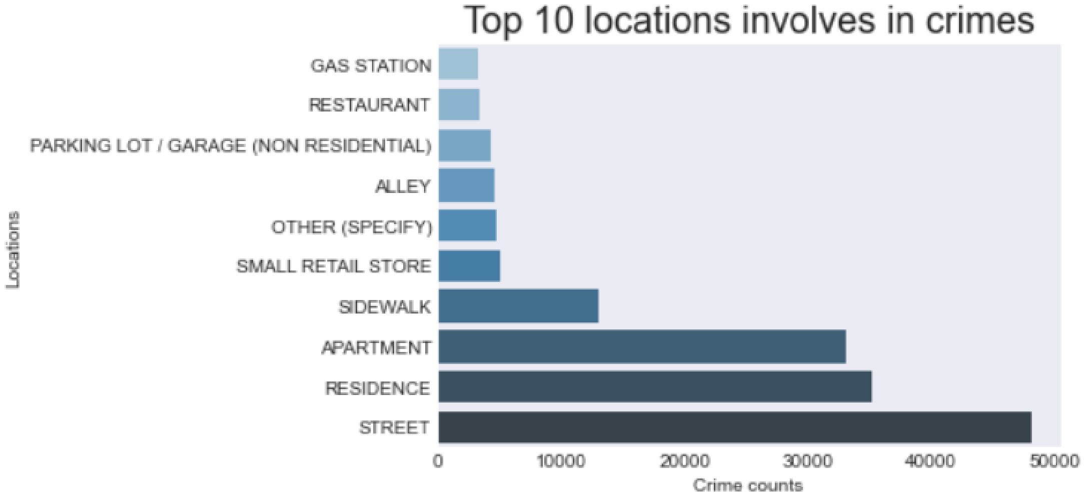
- From this graph, we see that "Narcotics" has a 100% arrest rate and even "Battery" crime has a good arrest rate comparing other crime
- None of crime arrest rate is stable and arrest rate also lower so the crime keeps on increasing every time.

4. Unsafety hours in Chicago city in 2020



- Criminal are sleeping at morning and strictly maintaining their timings at night.
- One strange thing that at 12 am the crime rate is higher then any other hours including night.

5. Top 10 locations that meant for criminal in Chicago city.



- Criminals are highly targeting on street, residence and apartment and sidewalks
- Street crime has a count of 48038 which is highest, In my opinion Chicago police are lazy or not doing their work properly or watching like an Tv show while crime attacks on streets
- This graph tells us public doesn't have safety in inside the house and outside the house too. Because the crimes are taking place mainly on roads and in houses.