CHICAGO
POLICE
URBS IN HORTO

Chicago crime dataset analysis

Chicago has been gaining a reputation as of late for being a very dangerous place to live. “Chicago police records recorded 506 murders in 2012, with an estimated 80 percent being gang related” (NBA). The main reason why I chose Chicago is because it is considered as one of the cities in Illinois with the highest crime rate. The dataset was collected from the ‘city of chicago’ official website which has many public datasets.

Dataset Description

For this particular project I have used three datasets ,the first dataset is the actual crime dataset which contains information about the crime that is time, location, category and primary type. The second dataframe is about the districts in chicago which contains district name and chicago district number which I will further use for getting an comprehensive idea about the crimes in chicago based on different districts. The dataset consists of 22 columns and 7.5 Million rows (dataset size 1.78GB). The dataset has information about the reported crimes from 2001 in the city of chicago except for the murders. The third dataset is about the community area names and their respective codes. Further I will discuss how I used these datasets for analysis.

- The dataset about the crime and IUCR can be found in the Chicago's official website.
- The dataset which had information about districts was also from the same portal.

Aim :

The aim of the project is to find out the crimes that occurred more often, the crime rate trend across the years, and during which month and hour of the day the crime rate was highest. Using the second dataset, I want to get the district name and join the two dataframes so that we can evaluate the crimes based on district. Initially, the dataset had many null values and even the schema was not properly defined, so I had to perform EDA before I did any transformation on the data. Spark doesn't have any built-in libraries for plotting, so I will use matplotlib to plot and I will also use SQL functions to work with the data frame.

How to run the notebook:

To run the notebook, we must have SQL functions, matplotlib, and other basic libraries installed, as mentioned in the notebook. The whole code needs to be run step by step from the beginning, and the environment should have all the two datasets attached. Several pertinent issues include, but are not limited to, the annual trend in Chicago's crime rates, the sorts of crime that are widespread in Chicago, as well as the city's most notorious crime locales. We intend to begin the study with these objectives in mind, although it is worthwhile to deviate if an interesting event is detected.

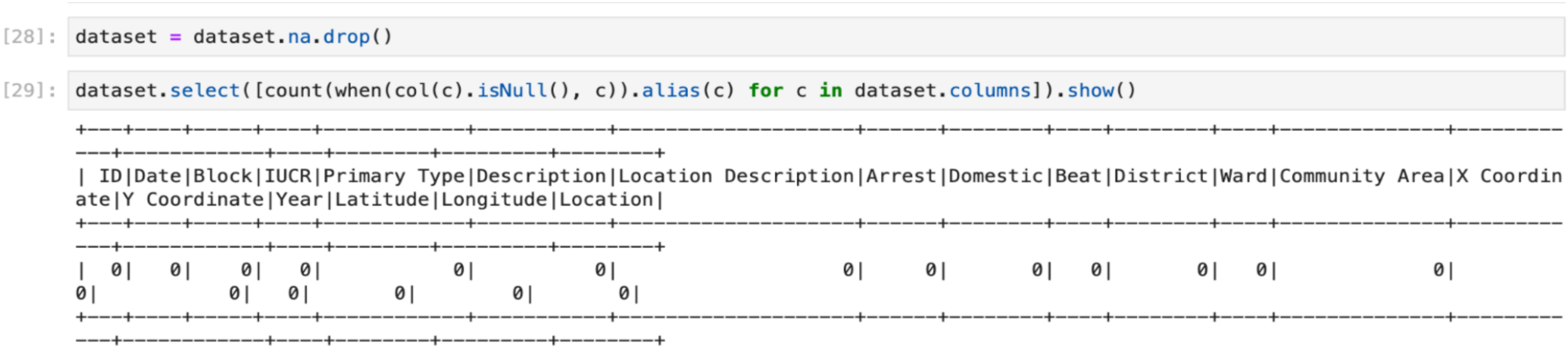
In the next slides we will try to follow the steps used for exploring and analysing the data :-

Step-1 (Exploring and Cleaning Data)

For the exploration, I started with defining the schema and then dropping all the columns which were unnecessary, the dataset consisted of 22 columns and 7.4 million rows, but for this particular analysis I considered only last ten years data so I have trimmed the data according to year. Moreover, the data had many null values which had to be removed. I removed them and also there where certain records which looked non relatable for example some crimes were described as "NON-CRIMINAL" I removed all of them using 'filter' function. These are some of the columns that are used in analyzing.

- **ID**: Unique identifier for a record.
- **Case Number** : Chicago Police Division Record Number.
- **Date**: Date of the incident.
- **Block**: The abbreviated address for the criminal activity.
- **IUCR**: Internal Uniform Chicago Crime Reporting Code.
- **Primary Type**: Type of crime
- **Description**: Little more details about the criminal activity
- **Location Description**: Location where the crime occured.
- **Arrest**: Boolean indicating whether arrest was made.
- **Domestic**: Indicates whether incident was domestic related.
- **District**: Indicates Police District where incident occurred.
- **Community Area**: The Community where incident occurred..
- **Year**: Year the incident occured.
- **Latitude & Longitude**: Location of the incident

To begin the process of cleaning my data in preparation for analysis, I wanted to check to see whether I had any nulls. I began by determining which columns had null values. Several did, including the following: Location Description, District, Community Area, X and Y coordinates, and both Latitude and Longitude.



```
[28]: dataset = dataset.na.drop()
```

```
[29]: dataset.select([count(when(col(c).isNull(), c)).alias(c) for c in dataset.columns]).show()
```

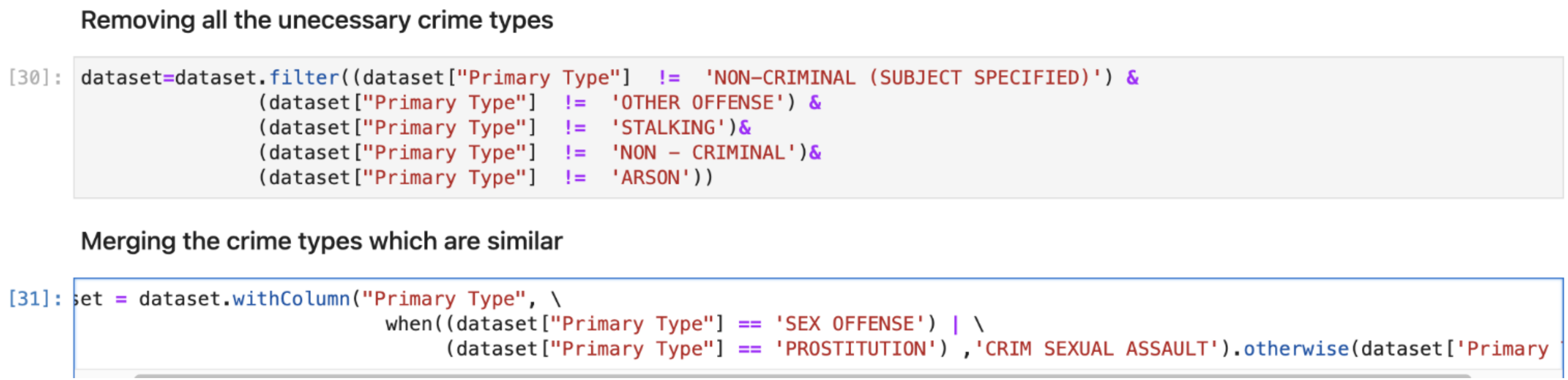
```
+---+----+-----+----+------------+-----------+--------------------+------+--------+----+--------+----+--------------+---------
---+------------+----+--------+---------+--------+
| ID|Date|Block|IUCR|Primary Type|Description|Location Description|Arrest|Domestic|Beat|District|Ward|Community Area|X Coordin
ate|Y Coordinate|Year|Latitude|Longitude|Location|
+---+----+-----+----+------------+-----------+--------------------+------+--------+----+--------+----+--------------+---------
---+------------+----+--------+---------+--------+
|  0|   0|    0|   0|           0|          0|                   0|     0|       0|   0|       0|   0|             0|
0|           0|   0|       0|        0|       0|
+---+----+-----+----+------------+-----------+--------------------+------+--------+----+--------+----+--------------+---------
---+------------+----+--------+---------+--------+
```

The original dataset had close to 7 million records and 22 features. One of the first things to check for is completeness of the records and how to handle missing values. It was observed that approx. 1/2 million rows were not completely populated. Since original data is 7 million records, it does not hurt to delete all rows with missing values. Fortunately, we do not lose much information here.

Step-2 (Checking relevant data for analyzing)

For further process firstly I removed all the crime types that are not relevant or not significant crimes then merged the similar looking crimes; the crime type that had multiple records which had similar meaning for example:-

To be able to analyze by date, I wanted to change the date column to DateTimeIndex.



Removing all the unecessary crime types

```
[30]: dataset=dataset.filter((dataset["Primary Type"]  !=  'NON-CRIMINAL (SUBJECT SPECIFIED)') &
                     (dataset["Primary Type"]  !=  'OTHER OFFENSE') &
                     (dataset["Primary Type"]  !=  'STALKING')&
                     (dataset["Primary Type"]  !=  'NON - CRIMINAL')&
                     (dataset["Primary Type"]  !=  'ARSON'))
```

Merging the crime types which are similar

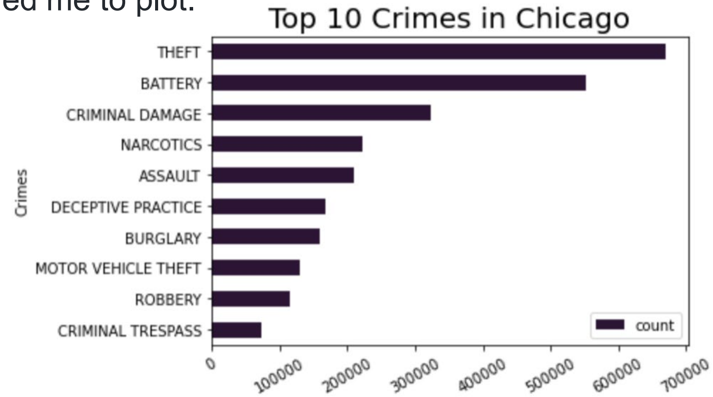
```
[31]: set = dataset.withColumn("Primary Type", \
                               when((dataset["Primary Type"] == 'SEX OFFENSE') | \
                                    (dataset["Primary Type"] == 'PROSTITUTION') ,'CRIM SEXUAL ASSAULT').otherwise(dataset['Primary
```

Step-3 (Analysis)

After thoroughly examining the entire dataset, I determined that I wanted to examine the relationship between various variables and the arrest rate. Although this was a boolean variable, I thought it could be interesting to investigate how factors such as primary type, domestic status, community region, and year affected the arrest rate. Additionally, I wanted to determine which year, month, and hour had the greatest crime rate; the relationship between arrested and non-arrested crimes; and the community area and district with the highest crime rate.

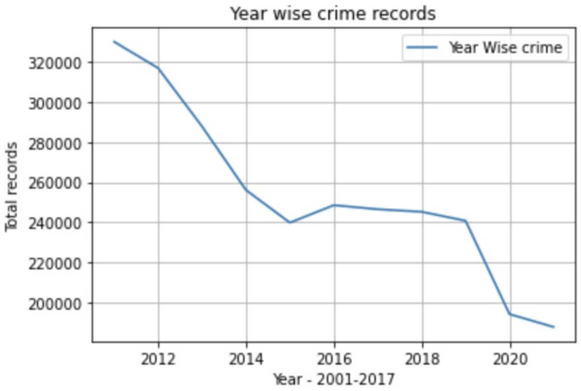
3.1 Visualizing Primary Type

To begin, I narrowed my dataframe into a few columns I was interested to look into. Primary type and count then plotted the dataset based on crime rate. For plotting I have used matplotlib and before actually plotting the dataset I had to transform it into pandas dataframe which allowed me to plot.



3.2 Crime Trend across the years

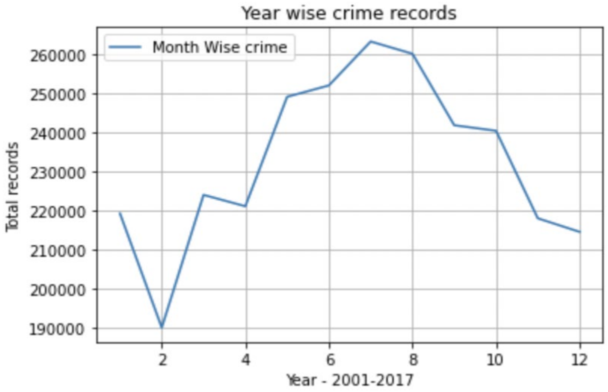
To ascertain the trend throughout time I conducted groupby on the year column and estimated the number of offenses committed in each year. I then sorted the results by year and presented them in a time-series graph to enhance the comprehension.



I wanted to verify that the outcome was accurate, so I searched for publications about crime in Chicago and discovered that other articles contained the same facts. This is one of the articles that details the most serious crimes in Chicago. Link

3.3 Crime trend across months

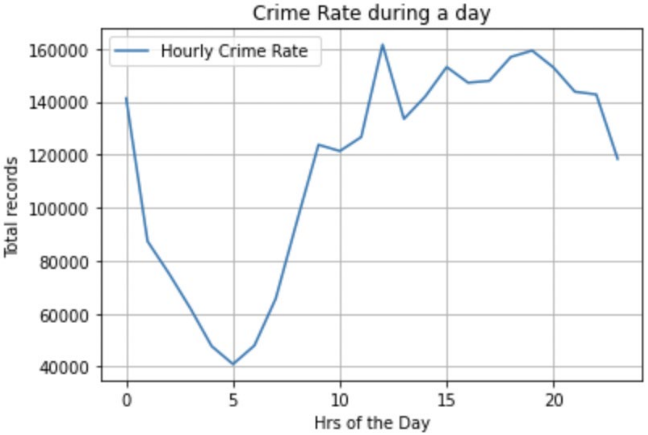
I wanted to check during which month of the year the crime rate was high for this I had to extract month from date&time column. For this first I extracted month from converted time column using SQL command then used groupby to group them based on months and then counted before plotting.



We can observe that during June, July, and August, the highest crime rates are recorded. That is usually in summer. This may seem strange, but according to various articles, the crime rate peaks during the summer, and the same is true for the state of Illinois.

3.4 Crime trend during a day

It is same as analyzing crime across month but here I wanted to check how crime rate varied across different hours of the day. I followed the similar steps, that is using SQL to extract hour and then using groupby and sorting them before plotting.

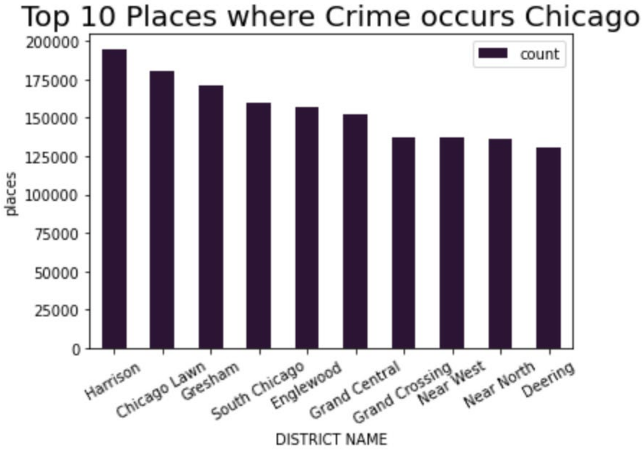


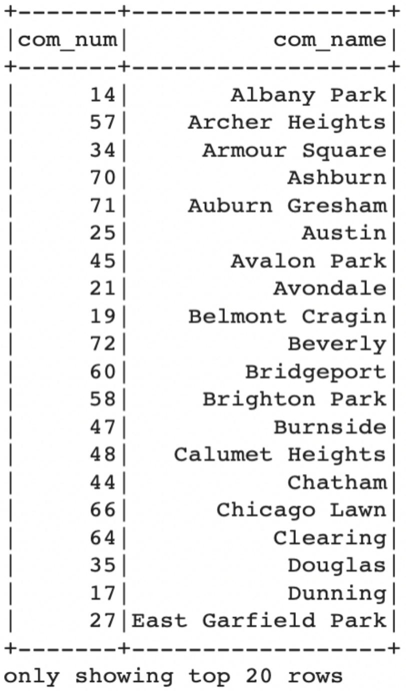
Violence committed by peaks in the afterschool hours and in the evenings on non-summer days. On Summer days, the incidence of violence committed increases through the afternoon and early evening hours, peaking between 6 p.m. and 9 p.m.

Step-4 (Joining different datasets for analysis)

As the dataset had lot of codes which were not comprehensive , I have used other two datasets for joining the original dataset. The columns I used for joining was District and community, initially the dataset had only respective codes by combining them I extracted respective names. For joining the dataframes I have used inner join -

We can have a look at the data. Initially, it had simply district numbers, which were unintelligible. I integrated them using another dataset that had district numbers and their associated district names, which simplified comprehension.





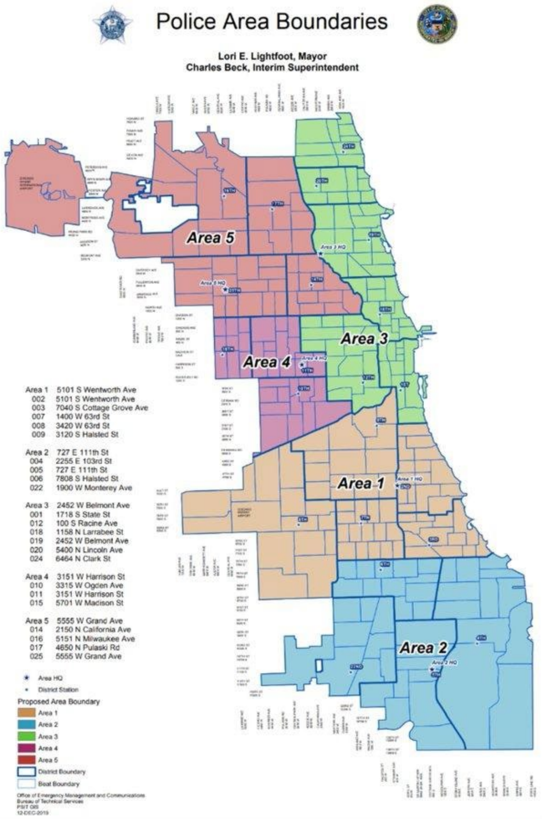
```
+-------+-----------------+
|com_num|         com_name|
+-------+-----------------+
|     14|      Albany Park|
|     57|   Archer Heights|
|     34|    Armour Square|
|     70|          Ashburn|
|     71|   Auburn Gresham|
|     25|           Austin|
|     45|      Avalon Park|
|     21|         Avondale|
|     19|   Belmont Cragin|
|     72|          Beverly|
|     60|       Bridgeport|
|     58|    Brighton Park|
|     47|         Burnside|
|     48|  Calumet Heights|
|     44|          Chatham|
|     66|     Chicago Lawn|
|     64|         Clearing|
|     35|          Douglas|
|     17|          Dunning|
|     27|East Garfield Park|
+-------+-----------------+
only showing top 20 rows
```

I repeated the previous procedures, grouping them and plotting them with matplotlib. According to the graph, the Harrison district had the most offenses, followed by Chicago Lawn and Greesham, which are all located inside Area 4 on the Chicago map. This explains why area 4 is one of the crime hotspots.

The area is on the south side of the city. The possible reason for the crime rate being too high may be the proximity of Riverdale to the freeway and the streets. Sadly, the economic conditions in the region are also quite poor, with most people making less than $8,000 a year.

Meanwhile, the area has an employment rate of 40%, with the **median household income** 71% less than the rest of the county. It may also contribute to the higher crime rate.

To get a better understanding of the crime rate based on area, I tried to calculate the crime count based on community areas as well, this fetched me the community areas with highest crime rate. As I plotted them I found out that places like Austin, Near North and South Shore were actually part of Area 4.

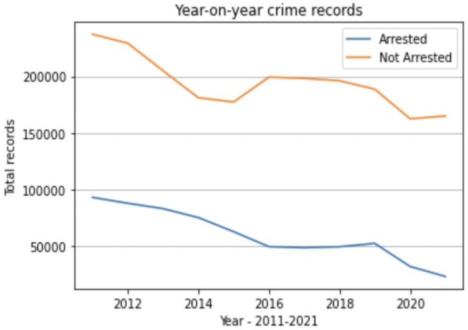


4.1 Comparison between arrested and non-arrested crimes

To compare the association between arrested and non-arrested crimes, I first filtered the crimes by label, that is, whether an arrest was made or not during the commission of the crime. Then I counted and plotted the counts for each of them in a single graph for comparison. We can deduce from the graph that the disparity between those arrested and those not arrested remained constant throughout the time period.

5. Conclusion :-

We may observe the most frequently occurring crimes and the frequently occurring places where crimes occurred based on the analysis and visualization results. According to these records, the most frequently committed crimes were theft, battery, criminal damage, and narcotics, accountt ting for 65.7 percent of all reported offenses. The most common locations for crimes to occur are on the street, sidewalks, in a residence, or an apartment, as these are the areas where the majority of people congregate. Despite the enormous volume of reported crimes in Chicago each year, the arrest rate was not even close to 50%, leading us to think that the city's police arrest and investigative tactics were ineffective. We believe that if our data analytics can provide us with all of this information about the security status of the city of Chicago, a larger data analytics project will provide significantly more valuable information that can be used as a powerful source for taking prudent actions to improve our cities' security status.



THANK YOU