Project Report on Chicago Crime Data

1. Introduction

In this project, we study the Chicago Crime dataset, which is one of the greatest open source data sets in this field, to gain a better knowledge of the city's security situation.As this is an initial phase of the project there are no significant insights, I have tried to perform EDA prior to actually working with the data. While working with the data the major thing which was hindering was the time taking to process data, so my initial idea for EDA was to consider the data based on certain time frame.

As the dataset was very large and had data from the year 2012-2022, I decided to consider only recent 10 years of the dataset. During the EDA, I was able to figure out many features which were interesting. I was able to figure out the solution for few of the problem statements as well.

2. Dataset Description

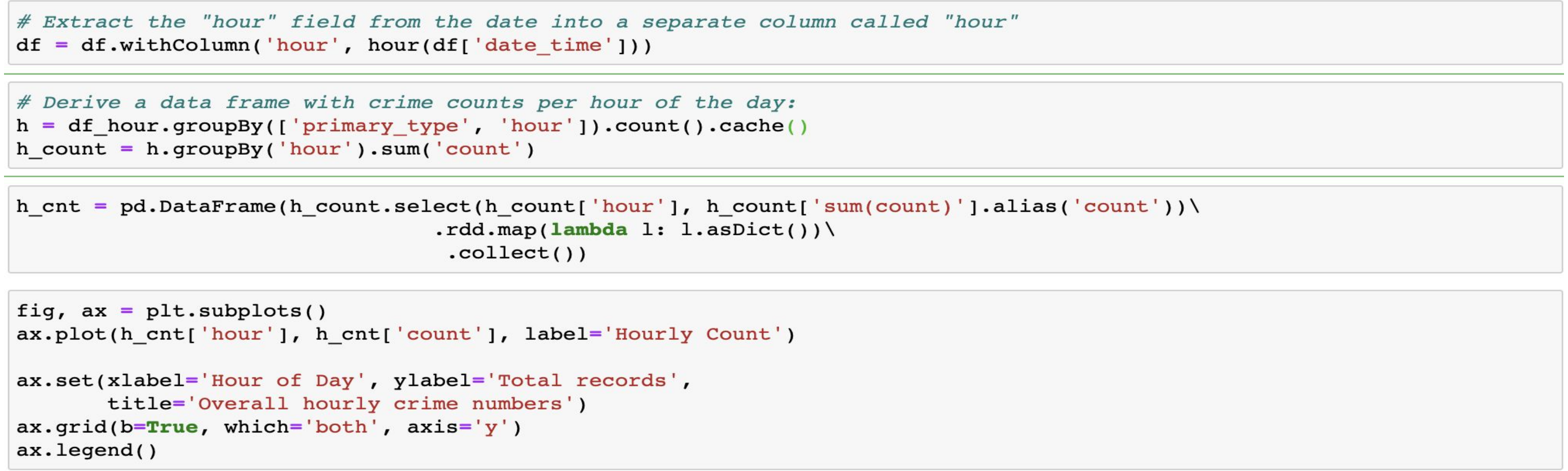
The dataset consists of 22 columns and 7.5 Million rows . The dataset has information about the reported crimes from 2001 in the city of chicago except for the murders. In general, the data includes information such as the date/time of the incident, the block where the crime occurred, the kind of offense, a description of the place, if an arrest was made, and geographic coordinates. In the following sections, we will discuss the specific properties of our data. The list of the names of each column from left to right are as follow:

ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location.

3. Problem Definition

1. How has the number of various crimes changed over time in Chicago?
2. How have the number arrests corresponded to the crimes changed over time in Chicago?
3. Which crimes are most frequently committed?
4. How have arrests evolved over the 16 years?
5. What time of the day are criminal the busiest?
6. Which locations are these frequent crimes being committed to?

To answer these problems I have followed steps such as, first comparing the unique variable count in each column and then grouped them, later plotted them to get better understanding of data.



```
# Extract the "hour" field from the date into a separate column called "hour"
df = df.withColumn('hour', hour(df['date_time']))
```

```
# Derive a data frame with crime counts per hour of the day:
h = df_hour.groupBy(['primary_type', 'hour']).count().cache()
h_count = h.groupBy('hour').sum('count')
```

```
h_cnt = pd.DataFrame(h_count.select(h_count['hour'], h_count['sum(count)'].alias('count'))\
                                 .rdd.map(lambda l: l.asDict())\
                                  .collect())
```

```
fig, ax = plt.subplots()
ax.plot(h_cnt['hour'], h_cnt['count'], label='Hourly Count')

ax.set(xlabel='Hour of Day', ylabel='Total records',
       title='Overall hourly crime numbers')
ax.grid(b=True, which='both', axis='y')
ax.legend()
```

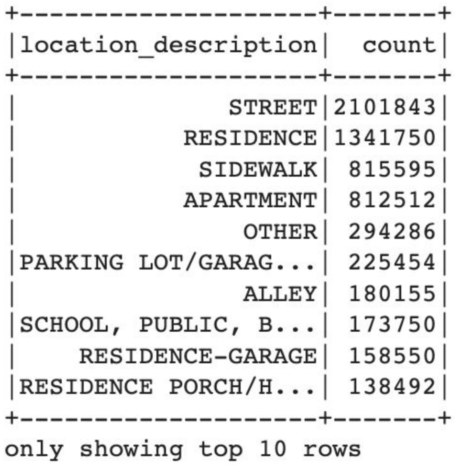
4. Analysis

In the initial phase of the EDA I found out many interesting results such as, the most occurred crimes, which community are had more number of crimes, these all are were done using EDA, however, something I found out was the features which can be used for applying Machine Learning model. But before performing the ML model I wanted to merge two dataframes which would allow me to give more features, however, in the process of doing it I was not able to find out how to merge two Dataframes as the Dataframes were not having equal number of columns.

So, I wanted to reach you regarding this problem. As for now I am trying to perform analysis on the current data and using some of the features I wanted try using Machine Learning model if it fits the problem. Here I have attached one of the output from the analysis which shows the location where most number of crimes are happening.
I also looked at the data for a few individual crimes, like theft, manslaughter, and sexual harassment. The findings show that the number of killings has been steadily decreasing between 2001 and 2016.

5.Conclusion

By the end of the project I want to perform thorough analysis on the data and describe all the information from dataset, during this process I want to clearly visualize the whole method which would make more easy to understand the data. I even want to join the two different tables and extract features which would help me in further analysis and will also help me to extract features which could probably fit with a Machine Learning model. This is the idea for right now but as I progress, I would like to describe comprehensively about the chicago crime.



```
+--------------------+-------+
|location_description|  count|
+--------------------+-------+
|              STREET|2101843|
|           RESIDENCE|1341750|
|            SIDEWALK| 815595|
|           APARTMENT| 812512|
|               OTHER| 294286|
|PARKING LOT/GARAG...| 225454|
|               ALLEY| 180155|
|SCHOOL, PUBLIC, B...| 173750|
|    RESIDENCE-GARAGE| 158550|
|RESIDENCE PORCH/H...| 138492|
+--------------------+-------+
only showing top 10 rows
```