Chicago Crime Analysis & Prediction

Presented by:
Farzeem Jiwani, Agha Ahmad & Faheem Mohammed

TABLE OF CONTENTS

Problem Description

With the increase of crimes, law enforcement agencies are continuing to demand advanced geographic information systems and new data mining approaches to improve crime analytics and better protect their communities.

The primary objective is to perform predictive analysis on the dataset to comprehend whether crime is a function of locality, time, climate, or any external features, along with analyzing its trends over the years. This could be helpful to law enforcement to take appropriate measures about the When-What-Where of the crime i.e. when and what type of crime could happen in what locality.

Work Distribution!

Agha Anas Ahmad

Pyspark SQL, Hive

Farzeem Jiwani

Pyspark, EDA, Predictive Modelling

Faheem Mohammed

HDFS, Hive, Project Integration

Dataset Specification

Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system:
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

1.8 GB — Size

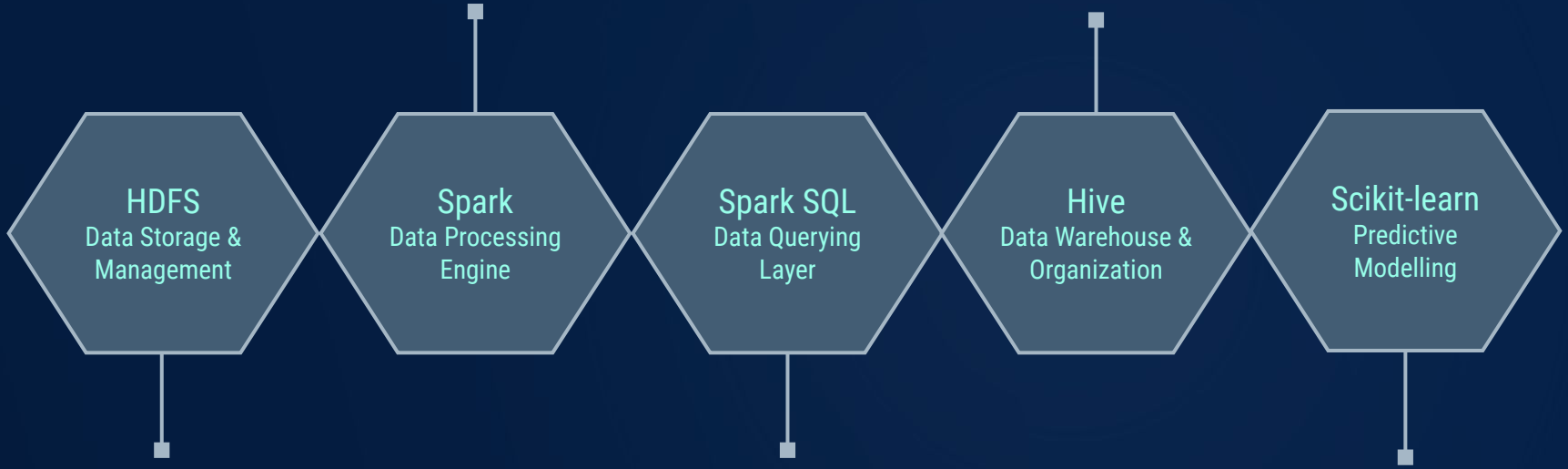
7.5 million — Rows

22 — Attributes

Key Attributes

Attribute	Description
Date	Best estimate Date when the incident occurred.
Block	Partially redacted address where the incident occurred.
IUCR	The Illinois Uniform Crime Reporting Code is used to classify criminal incidents when taking individual reports.
Primary Type	Primary description of the IUCR code.
Description	Secondary description of the IUCR code.
Location Description	Description of the location of the incident.
Arrest	Indicates whether an arrest was made.
Domestic	Indicates whether the incident was domestic-related.
District	Indicates the police district where the incident occurred.
Community Area	Indicates the Community area where the incident occurred.
Year	Year the incident occurred.
Latitude	Latitude of the location where the incident occurred.
Longitude	Longitude of the location where the incident occurred.
Location	Combination of Latitude and Longitude.

Implementation Overview



HDFS
Data Storage & Management

```
[root@sandbox-hdp Final_Project]# hadoop fs -mkdir /user/root/Project
[root@sandbox-hdp Final_Project]# hadoop fs -put Crimes_2001_to_Present.csv /user/root/Project/
[root@sandbox-hdp Final_Project]# hadoop fs -ls /user/root/Project
Found 1 items
-rw-r--r--   1 root root 1755442749 2021-11-29 20:22 /user/root/Project/Crimes_2001_to_Present.csv
[root@sandbox-hdp Final_Project]#
```

Steps for setting up Chicago Crime input file in HDFS:

1. Creating the Project Folder on HDFS:
   hadoop fs –mkdir /user/root/Project

2. Copying the dataset to HDFS:
   hadoop fs -put Crimes_2001_to_Present.csv /user/root/Project/

Spark
Data Processing Engine

Data Exploration

- PySpark DataFrames were used to read the input csv from HDFS
- Analyzed fundamental metrics and trends for the disparate types of crime
- Utilized visualization tools such as bar graphs, line graphs, and maps for appropriate illustration

Spark SQL
Data Querying Layer

Handling Missing Values
Dropped the missing values as they accounted for less than 1% of the dataset

Balancing Dataset
Balanced the dataset having extremely rare crime types and locations

Feature Selection
Factorized the dataset and filtered out relevant features from the dataset using Pearson correlation matrix

Hive
Data Warehouse

Table Structure
Partitioning on attribute primary type

```
hive> dfs -ls /apps/hive/warehouse/chicago_crimes.db/crime_part;
Found 38 items
drwxr-xr-x   - root hadoop          0 2021-11-24 02:50 /apps/hive/warehouse/chicago_crimes.db/crime_part/.hive-staging_hive_2021-11-24_02-26-52_091_4443651590611713320
1
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=ARSON
drwxrwxrwx   - root hadoop          0 2021-11-24 04:00 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=ASSAULT
drwxrwxrwx   - root hadoop          0 2021-11-24 04:00 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=BATTERY
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=BURGLARY
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=CONCEALED CARRY LICENSE VIOLATION
drwxrwxrwx   - root hadoop          0 2021-11-24 04:00 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=CRIM SEXUAL ASSAULT
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=CRIMINAL DAMAGE
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=CRIMINAL SEXUAL ASSAULT
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=CRIMINAL TRESPASS
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=DECEPTIVE PRACTICE
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=DOMESTIC VIOLENCE
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=GAMBLING
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=HOMICIDE
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=HUMAN TRAFFICKING
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=INTERFERENCE WITH PUBLIC OFFICER
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=INTIMIDATION
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=KIDNAPPING
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=LIQUOR LAW VIOLATION
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=MOTOR VEHICLE THEFT
drwxrwxrwx   - root hadoop          0 2021-11-24 04:00 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=NARCOTICS
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=NON - CRIMINAL
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=NON-CRIMINAL
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=NON-CRIMINAL (SUBJECT SPECIFIED)
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=OBSCENITY
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=OFFENSE INVOLVING CHILDREN
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=OTHER NARCOTIC VIOLATION
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=OTHER OFFENSE
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=PROSTITUTION
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=PUBLIC INDECENCY
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=PUBLIC PEACE VIOLATION
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=Primary Type
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=RITUALISM
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=ROBBERY
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=SEX OFFENSE
drwxrwxrwx   - root hadoop          0 2021-11-24 03:59 /apps/hive/warehouse/chicago_crimes.db/crime_part/primarytype=STALKING
```

Table Structure
Bucketing on attribute year

```
hive> dfs -ls /apps/hive/warehouse/chicago_crimes.db/crime_bucket;
Found 20 items
-rwxrwxrwx   1 root hadoop   82083314 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000000_0
-rwxrwxrwx   1 root hadoop  132610656 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000001_0
-rwxrwxrwx   1 root hadoop   96411543 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000002_0
-rwxrwxrwx   1 root hadoop   94913752 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000003_0
-rwxrwxrwx   1 root hadoop   93677747 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000004_0
-rwxrwxrwx   1 root hadoop   90524228 2021-11-24 03:31 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000005_0
-rwxrwxrwx   1 root hadoop   89690211 2021-11-24 03:31 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000006_0
-rwxrwxrwx   1 root hadoop   87675998 2021-11-24 03:31 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000007_0
-rwxrwxrwx   1 root hadoop   85381550 2021-11-24 03:31 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000008_0
-rwxrwxrwx   1 root hadoop   78881454 2021-11-24 03:31 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000009_0
-rwxrwxrwx   1 root hadoop   74418116 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000010_0
-rwxrwxrwx   1 root hadoop   70846272 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000011_0
-rwxrwxrwx   1 root hadoop   68065151 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000012_0
-rwxrwxrwx   1 root hadoop   62126177 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000013_0
-rwxrwxrwx   1 root hadoop   55924358 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000014_0
-rwxrwxrwx   1 root hadoop   53941382 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000015_0
-rwxrwxrwx   1 root hadoop   55186921 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000016_0
-rwxrwxrwx   1 root hadoop   55096472 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000017_0
-rwxrwxrwx   1 root hadoop   54923107 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000018_0
-rwxrwxrwx   1 root hadoop   53617219 2021-11-24 03:32 /apps/hive/warehouse/chicago_crimes.db/crime_bucket/000019_0
hive>
```

Arrests by primary type

```
NARCOTICS       723119
BATTERY 283951
THEFT   178618
CRIMINAL TRESPASS       144350
ASSAULT 96873
OTHER OFFENSE   81290
PROSTITUTION    69161
WEAPONS VIOLATION       68585
CRIMINAL DAMAGE 57013
DECEPTIVE PRACTICE      44545
PUBLIC PEACE VIOLATION  29890
MOTOR VEHICLE THEFT     27193
ROBBERY 25967
BURGLARY        23092
INTERFERENCE WITH PUBLIC OFFICER        16271
GAMBLING        14273
LIQUOR LAW VIOLATION    14066
OFFENSE INVOLVING CHILDREN      10500
SEX OFFENSE     7745
HOMICIDE        5407
CRIM SEXUAL ASSAULT     4334
ARSON   1468
CONCEALED CARRY LICENSE VIOLATION       801
KIDNAPPING      744
INTIMIDATION    643
STALKING        604
OBSCENITY       536
CRIMINAL SEXUAL ASSAULT 317
PUBLIC INDECENCY        178
OTHER NARCOTIC VIOLATION        95
NON-CRIMINAL    11
HUMAN TRAFFICKING       6
NON - CRIMINAL  6
NON-CRIMINAL (SUBJECT SPECIFIED)        3
RITUALISM       3
DOMESTIC VIOLENCE       1
Time taken: 76.652 seconds, Fetched: 36 row(s)
hive>
```

	All Data	Partitioned	Bucketing
Execution Time	76.652	72.245	65.151

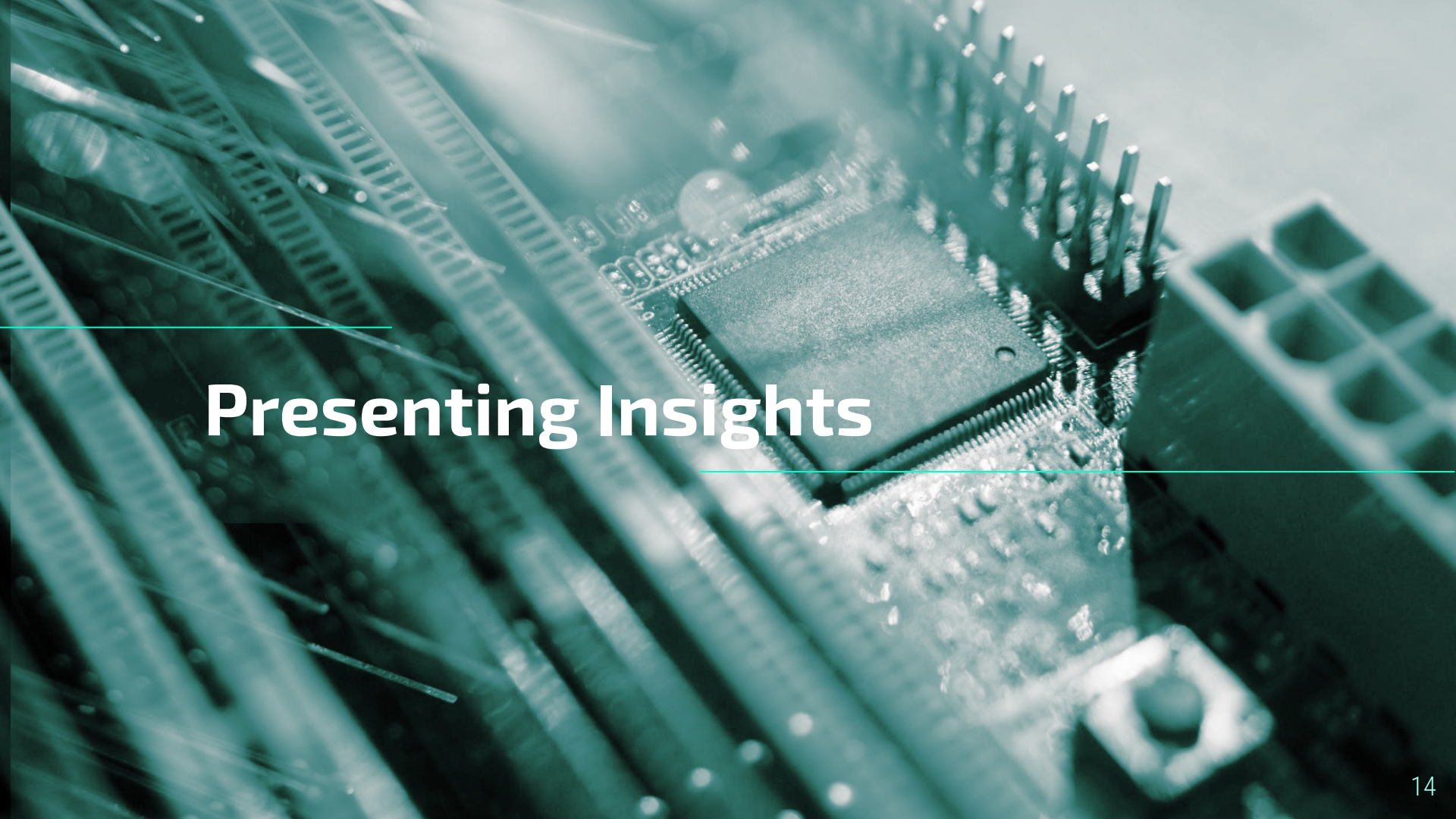
```
SELECT primarytype, COUNT(*) AS cnt FROM chicago_crimes.crime_bucket WHERE arrest = True GROUP BY primarytype ORDER BY cnt DESC
```

Districts with the most reported incidents

```
8       483013
11      465965
7       422495
6       420187
25      412906
4       401801
3       366203
9       356144
12      348159
2       338125
18      325709
19      325337
5       316648
15      313377
10      309642
1       289103
14      283447
NULL    240421
16      238526
22      229699
24      218987
17      206496
20      125176
31      216
122     131
113     106
111     33
121     19
124     12
123     8
112     5
21      4
114     4
Time taken: 32.226 seconds, Fetched: 33 row(s)
hive>
```

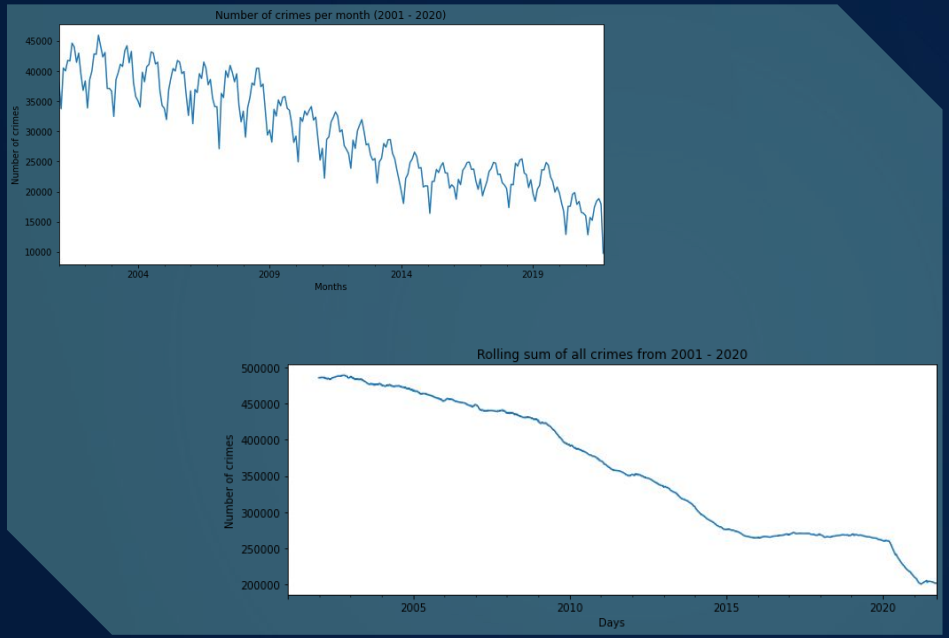
	All Data	Partitioned	Bucketing
Execution Time	43.478	39.729	29.22

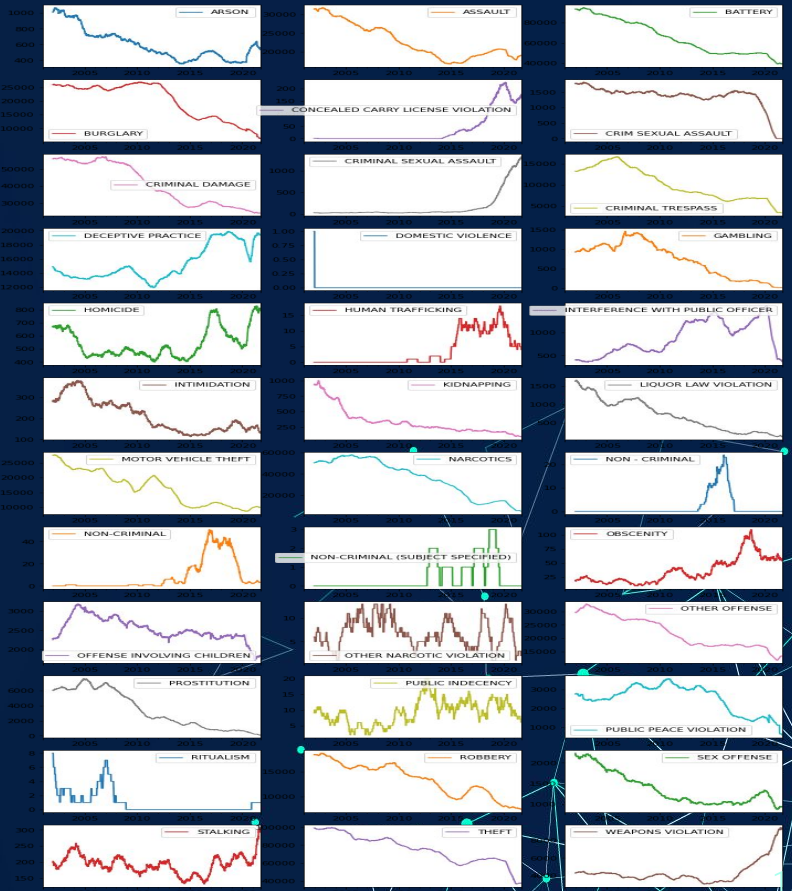
```
SELECT district, COUNT(*) AS cntdistrict FROM
chicago_crimes.crime_bucket GROUP BY
district ORDER BY cntdistrict DESC
```



Presenting Insights

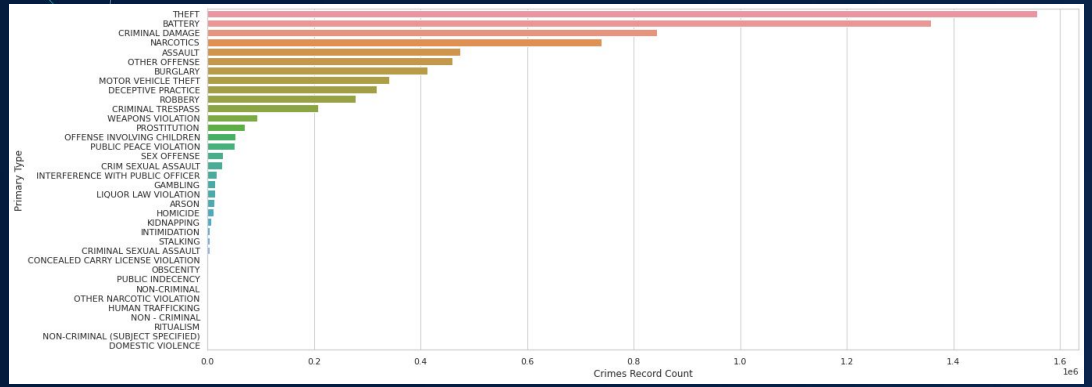
Crime Trends Over the Years





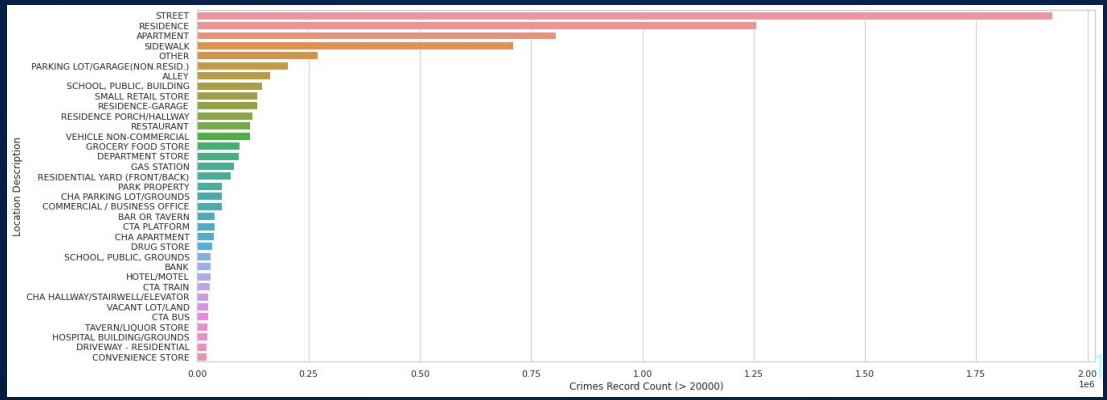
The above two graphs show a decreasing pattern of crime over the years. However, on further analysis (right graph) for each crime type, it was found not be to true as shown

Crimes based on Primary Type & Location

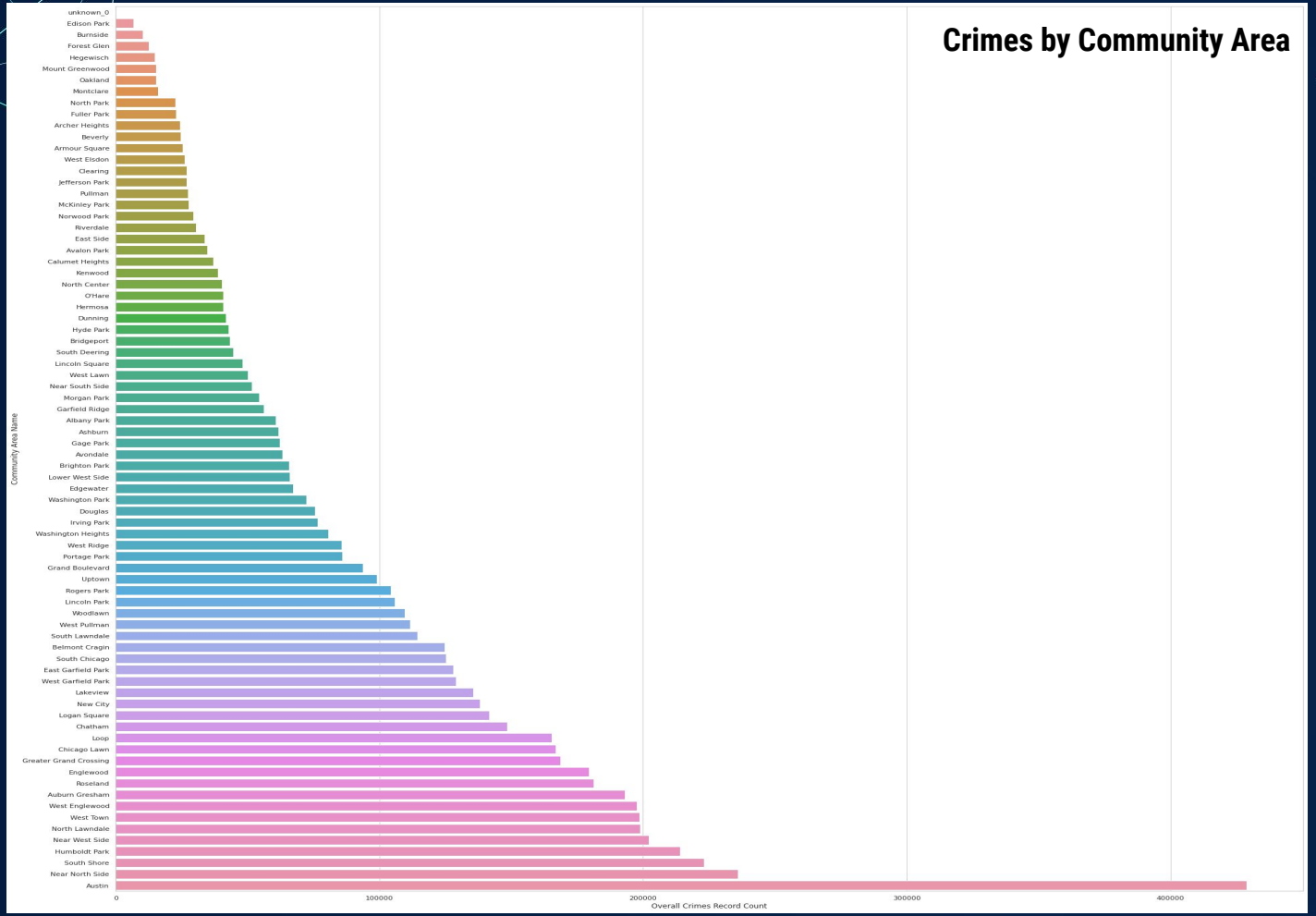


The graph depicts the distribution of crimes based on their Primary Type, among which, most crimes are Theft and Battery.

The graph depicts the distribution of crimes based on their Location, among which, most crimes occur on the Street or Residence.

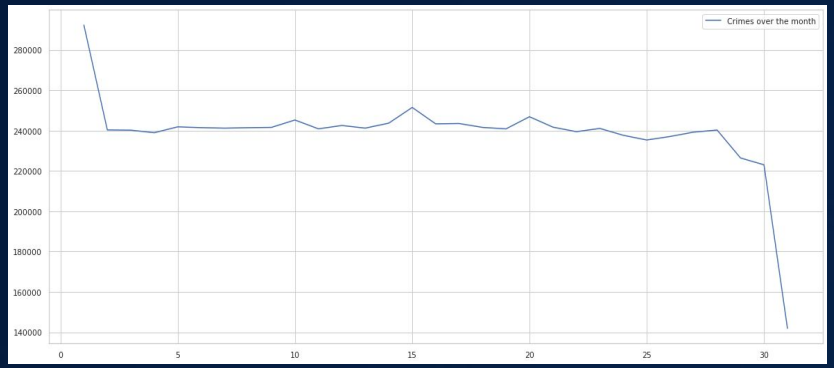


Crimes by Community Area

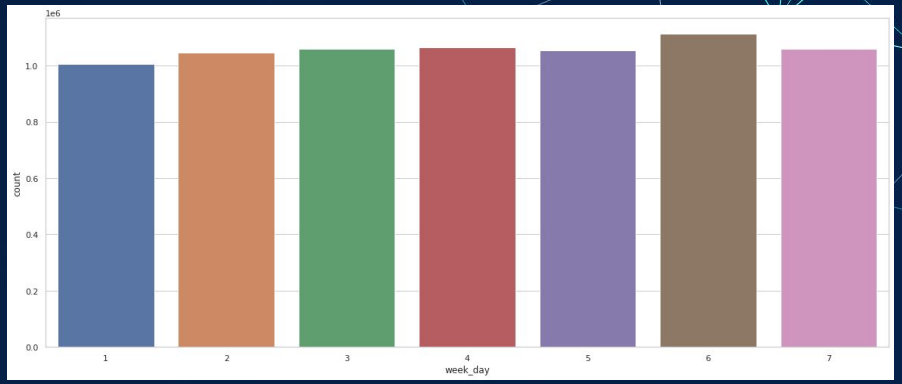


Monthly, Weekly and Hourly Crime Analysis

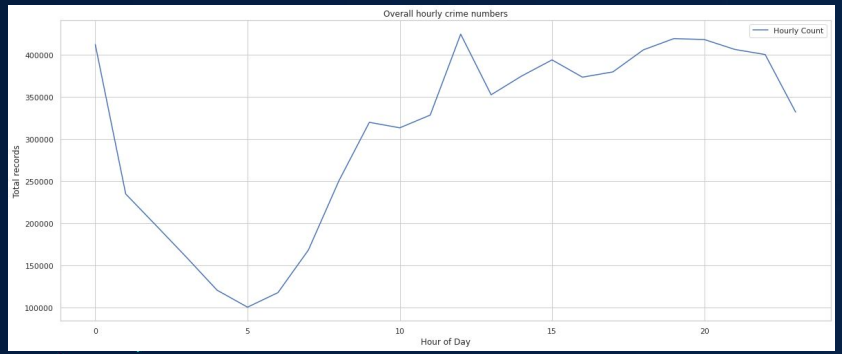
Crimes over day of the month



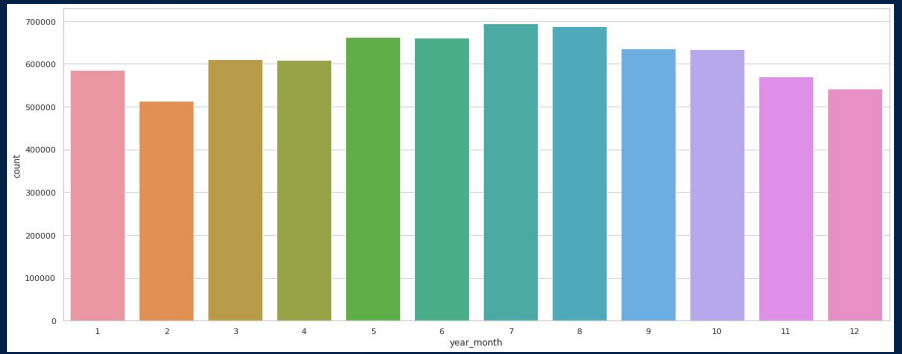
Crimes over day of the week



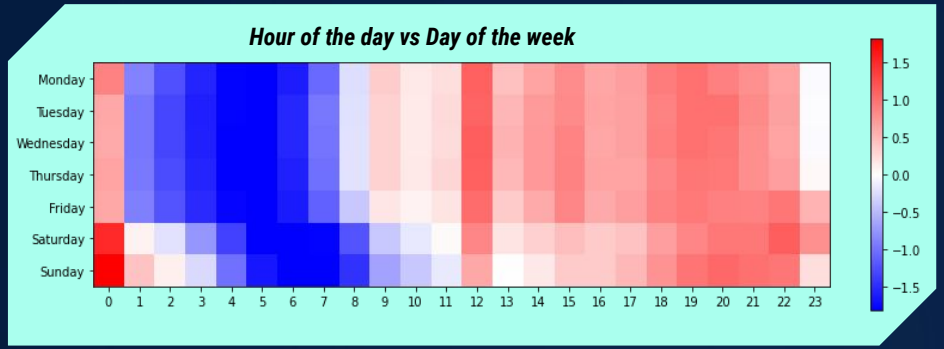
Crimes over hour of the day

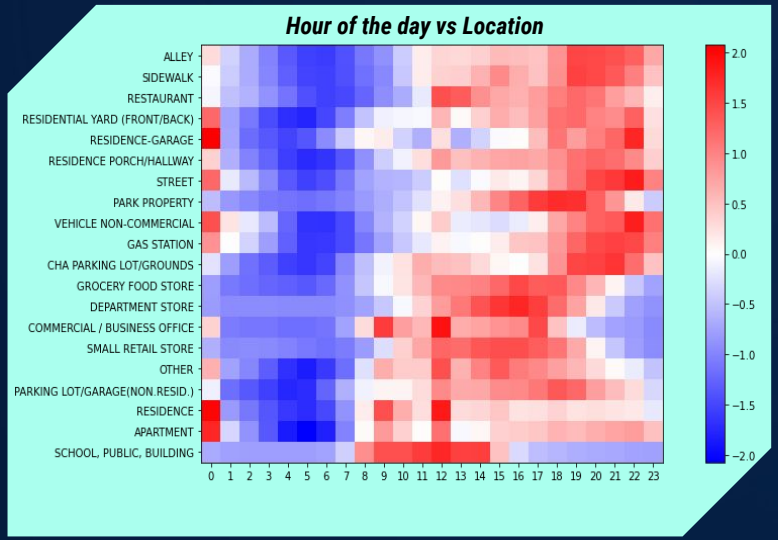


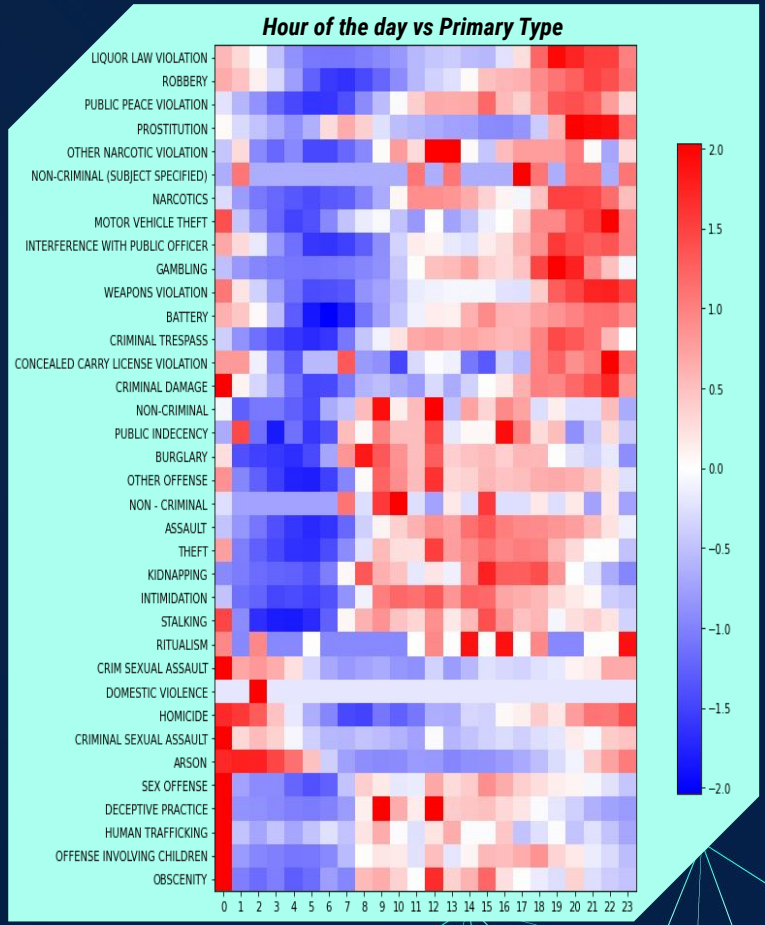
Crimes over month of the year



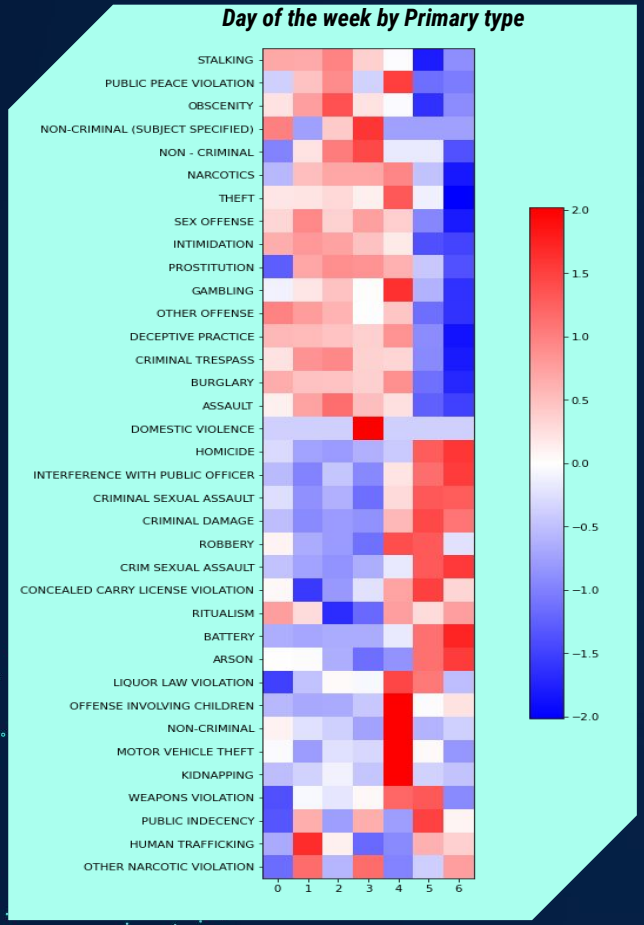
In Depth - Hourly Crime Analysis

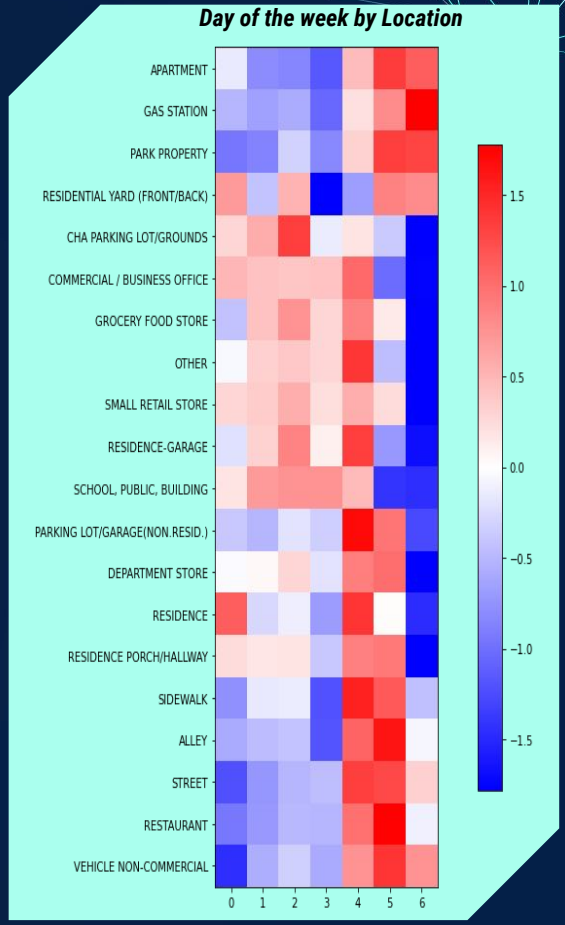




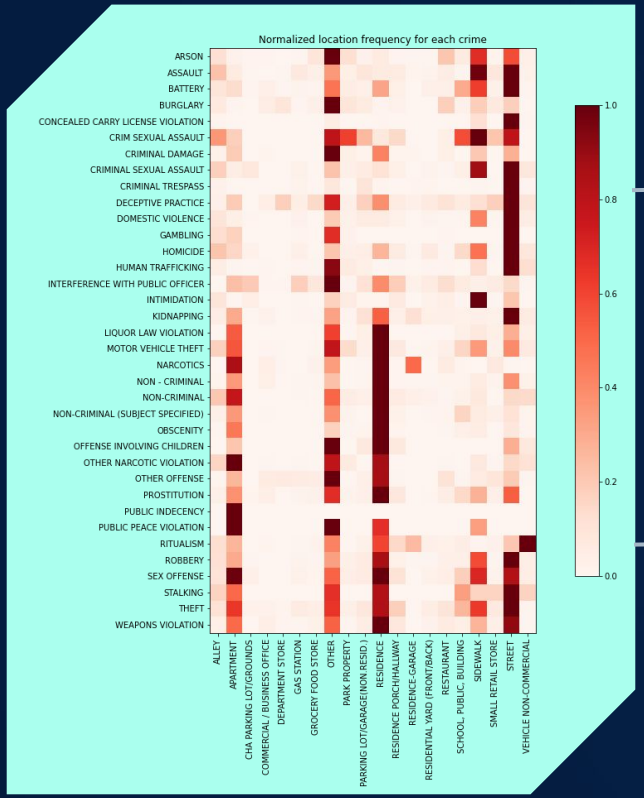


Crime Heatmap





Location vs Primary Type

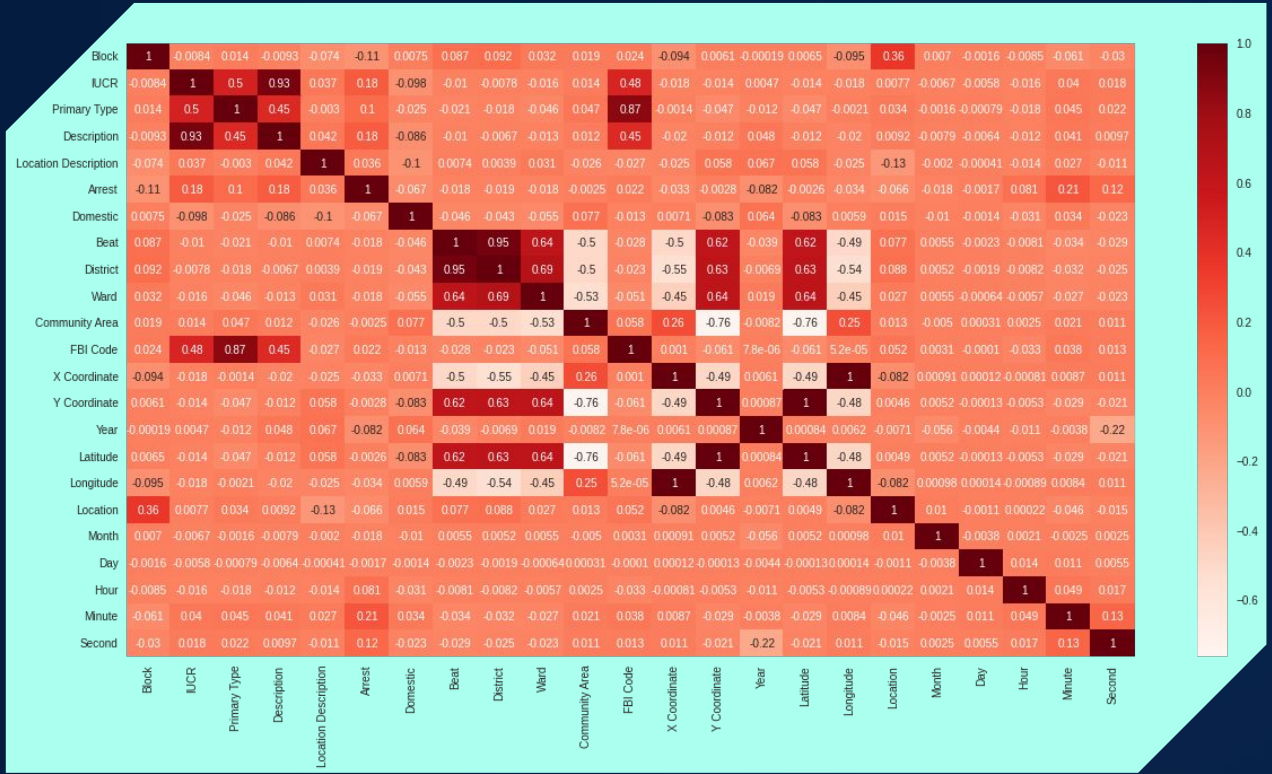


In this analysis, we have taken each crime type and re-normalized its location frequency to be between zero and one. This way, we can look at the top frequent locations of each crime type (darker red reflects a more frequent location).

It appears that most crimes occur at either apartments, residences, sidewalks, streets, or 'other'.

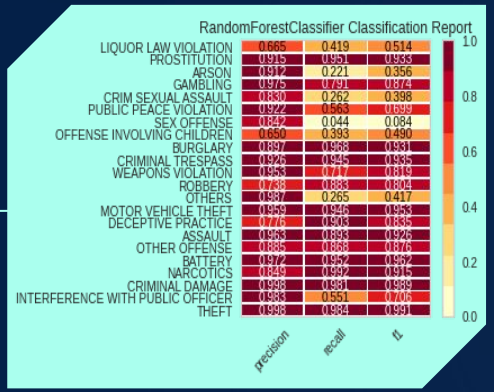
Predictive Modelling

Feature Selection



Model Building & Evaluation

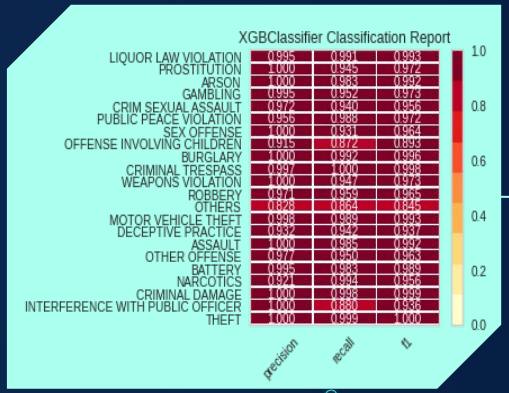
RandomForestClassifier
n_estimators=70
min_samples_split = 30
bootstrap = True
max_depth = 50
min_samples_leaf = 25



Random Forest

Accuracy : 0.93072
Recall : 0.93072
Precision : 0.9341
F1 Score : 0.93072

XGBClassifier
n_estimators=100
max_depth=3



XGBoost

Accuracy : 0.98318
Recall : 0.98318
Precision : 0.9837
F1 Score : 0.98318

Future Work & Lessons Learned

For future work, to boost the classification accuracy, it will be necessary to incorporate other information. Additionally, some events and the outcomes of the events may be associated with some crime types, for example, basketball games, baseball games, and elections. Weather information and classification of buildings can also be incorporated.

In our exploratory data analysis, we revealed that several features, for example location of the crime scene and the time of the crime, are associated with the type of crime, providing the basis for out modeling later. Hence, one single feature may not be sufficient for determining the type of crime, but a combination of various features can be powerful.

The end goal would be to create a web application portal that receives an address or location and time arguments in Chicago from the user and predicts the probability of crime (hence safety) at a specific location and a particular time.

In conclusion, we have learnt and implemented the tools to the best of our knowledge taught in the course.

Thank You!

Any Questions?