Chicago Crime Data Analysis

Aomkar Mathakar - 015925013
Mugdha Gumphekar - 015786238
Sandesh Gupta - 015649036
Venkata Prithvi Raj Namburi - 015944994

Dataset

Number of records: 7,424,694

Number of attributes: 22

Sample Attributes:

Crime Type	Description	Location Description	Arrest	Domestic	District
Ward	Community area	Date	Latitude	Longitude	Block

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

Data mining methods used

1. Association rule mining
    a. Apriori algorithm
    b. Efficient Apriori
    c. FP Growth
2. Clustering methods
    a. K-means clustering
    b. K-modes clustering
    c. K-Prototype clustering
    d. Hierarchical Agglomerative Clustering

Association Rule Mining - Analysis

1. Analyse 2021 crime data
    a. Dates: 01/01/2021 - 11/10/2021
    b. Total 158084 records

2. Analyse Pre and Post Covid data
    a. Dates: Pre-covid (03/20/2018 - 03/20/2020), Post- covid (03/21/2020 - 10/11/2021)
    b. Total records: Pre-covid: 528391 and Post-covid: 319531

Association Rule Mining - Preprocessing

Preprocessing methods:

1. Drop redundant or unuseful columns
2. Add new column
3. Convert the record data to transaction data

Date	Block	IUCR
11/21/2021 11:50:00 ...	050XX W MONROE ST	0930
11/21/2021 11:45:00 ...	061XX S MONITOR AVE	0497
11/21/2021 11:40:00 ...	006XX E 67TH ST	1305
11/21/2021 11:39:00 ...	065XX S HARVARD AVE	0820
11/21/2021 11:30:00 ...	079XX S GREENWOOD A...	0486

```
[('Primary Type-DECEPTIVE PRACTICE',
  'Description-ILLEGAL USE CASH CARD',
  'Location Description-ATM (AUTOMATIC TELLER MACHINE)',
  'Arrest-False',
  'Domestic-False',
  'District-18',
  'Ward-27.0',
  'Community Area-8.0',
  'Month-1'),
```

Association Rule Mining - Algorithms used

1. Apriori algorithm
2. Efficient Apriori
3. FP Growth

Concepts:

- Support
- Confidence
- Lift

Association Rule Mining - 2021 data analysis

Support	Confidence	Lift	Number of rules generated	Time taken
0.05	0.9	1	85	16s
0.05	0.8	1	110	14s
0.10	0.8	1	18	2s
0.10	0.9	1	11	2s
0.20	0.9	1	1	50ms
0.5	0.9	1	0	34ms

Association Rule Mining - Algorithm Runtime Comparison

	Apriori	Efficient Apriori	FP Growth
Support	0.05	0.05	0.05
Confidence	0.9	0.9	0.9
Lift	1	1	1
Time taken	16s	0.84s	1.8s

Association Rule Mining - Itemset Generation (2021 data)

Min_support = 0.05, min_confidence = 0.9, min_lift = 1

	A	B	C	D	E
1	antecedents	consequents	support	confidence	lift
20	'Description-DOMESTIC BATTERY SIMPLE', 'Arrest-False'	'Domestic-True'	0.08	0.99	4.43
21	'Description-DOMESTIC BATTERY SIMPLE', 'Primary Type-BATTERY', 'Arrest-Fal	'Domestic-True'	0.08	0.99	4.43
22	'Domestic-False', 'Description-OVER $500'	'Primary Type-THEFT', 'Arrest-False'	0.05	0.99	5.44
23	'Domestic-False', 'Description-OVER $500'	'Arrest-False'	0.05	0.99	1.13
24	'Domestic-False', 'Primary Type-THEFT', 'Description-OVER $500'	'Arrest-False'	0.05	0.99	1.13
25	'Description-OVER $500'	'Primary Type-THEFT', 'Arrest-False'	0.06	0.99	5.43
26	'Description-OVER $500'	'Arrest-False'	0.06	0.99	1.13
27	'Description-OVER $500', 'Primary Type-THEFT'	'Arrest-False'	0.06	0.99	1.13
28	'Primary Type-DECEPTIVE PRACTICE'	'Arrest-False'	0.08	0.99	1.13
29	'Domestic-False', 'Primary Type-DECEPTIVE PRACTICE'	'Arrest-False'	0.08	0.99	1.13
30	'Domestic-False', 'Description-$500 AND UNDER'	'Primary Type-THEFT', 'Arrest-False'	0.07	0.99	5.43
31	'Domestic-False', 'Description-$500 AND UNDER'	'Arrest-False'	0.07	0.99	1.13
32	'Domestic-False', 'Description-$500 AND UNDER', 'Primary Type-THEFT'	'Arrest-False'	0.07	0.99	1.13
33	'Description-$500 AND UNDER'	'Primary Type-THEFT', 'Arrest-False'	0.07	0.99	5.42
34	'Description-$500 AND UNDER'	'Arrest-False'	0.07	0.99	1.12
35	'Description-$500 AND UNDER', 'Primary Type-THEFT'	'Arrest-False'	0.07	0.99	1.12

Association Rule Mining - Itemset Generation (Pre-Covid)

Min_support = 0.05, min_confidence = 0.5, min_lift = 1

	antecedents	consequents	support	confidence	lift
2	Description-DOMESTIC BATTERY SIMPLE'	Primary Type-BATTERY'	0.09	1	5.31
3	Description-$500 AND UNDER'	Primary Type-THEFT'	0.09	1	4.15
4	Description-OVER $500'	Primary Type-THEFT'	0.06	1	4.15
5	Primary Type-NARCOTICS'	Arrest-True'	0.05	1	4.79
6	Primary Type-NARCOTICS'	Domestic-False'	0.05	1	1.2
7	Primary Type-NARCOTICS'	Domestic-False', 'Arrest-True'	0.05	1	5.55
8	Primary Type-DECEPTIVE PRACTICE'	Domestic-False'	0.07	0.99	1.19
9	Ward-42.0'	Domestic-False'	0.06	0.97	1.16
10	Description-OVER $500'	Arrest-False', 'Primary Type-THEFT'	0.05	0.97	4.45
11	Description-OVER $500'	Arrest-False'	0.05	0.97	1.22
12	Primary Type-THEFT'	Domestic-False'	0.23	0.96	1.15
13	Description-$500 AND UNDER'	Arrest-False', 'Primary Type-THEFT'	0.09	0.96	4.42
14	Description-$500 AND UNDER'	Arrest-False'	0.09	0.96	1.22

Association Rule Mining - Itemset Generation (Post Covid)

Min_support = 0.05, min_confidence = 0.5, min_lift = 1

	antecedents	consequents	support	confidence	lift
2	Description-TO PROPERTY'	Primary Type-CRIMINAL DAMAGE'	0.06	1	8.08
3	Description-$500 AND UNDER'	Primary Type-THEFT'	0.08	1	5.34
4	Description-OVER $500'	Primary Type-THEFT'	0.05	1	5.34
5	Description-DOMESTIC BATTERY SIMPLE'	Primary Type-BATTERY'	0.1	1	5.01
6	Month-1'	Year-2021'	0.05	1	2.02
7	Month-11'	Year-2020'	0.05	1	1.98
8	Month-12'	Year-2020'	0.05	1	1.98
9	Primary Type-DECEPTIVE PRACTICE'	Domestic-False'	0.08	0.99	1.25
10	Primary Type-DECEPTIVE PRACTICE'	Arrest-False'	0.08	0.99	1.14
11	Description-OVER $500'	Primary Type-THEFT', 'Arrest-False'	0.05	0.99	5.51
12	Description-OVER $500'	Arrest-False'	0.05	0.99	1.14
13	Description-$500 AND UNDER'	Primary Type-THEFT', 'Arrest-False'	0.08	0.98	5.49
14	Description-$500 AND UNDER'	Arrest-False'	0.08	0.98	1.13
15	Primary Type-DECEPTIVE PRACTICE'	Domestic-False', 'Arrest-False'	0.08	0.98	1.42
16	Description-TO VEHICLE'	Arrest-False'	0.06	0.97	1.11
17	Primary Type-CRIMINAL DAMAGE'	Arrest-False'	0.12	0.96	1.11
18	Primary Type-THEFT'	Arrest-False'	0.18	0.96	1.1

Association Rule Mining - 2021 data conclusion

- In districts 4 and 8, no arrests have been made for more than 90% of the crimes.
- The months of August and September contribute to 20% of the annual crime. And yet no arrests were made for more than 90% of the cases.
- In following crimes, arrests rates were very low:
  - Theft over $500 - 1%
  - Theft under $500 - 1%
  - Deceptive practice - 1%
  - Criminal Damage to vehicle - 2%
  - Simple Assault - 6%
  - Crimes committed at residences, apartments - 6%, contributes to 25% of the crimes.

Association Rule Mining - Pre/Post Covid data conclusion

- Pre-Covid, the arrest rate for the case of Narcotics was 100% whereas no cases were recorded post covid with the same rate of confidence.
- Post-Covid thefts increased about 40%
- Pre-Covid, the months of August and September resulted in 15% of the crime whereas Post-Covid it increased to 20%, and yet the arrest was only about 1%
- In the following crimes, arrests rates are very low:
    - Theft over $500 - 1%
    - Theft under $500 - 1%
    - Deceptive practice - 1%
    - Criminal Damage to vehicle - 2%

Clustering - Data Preprocessing

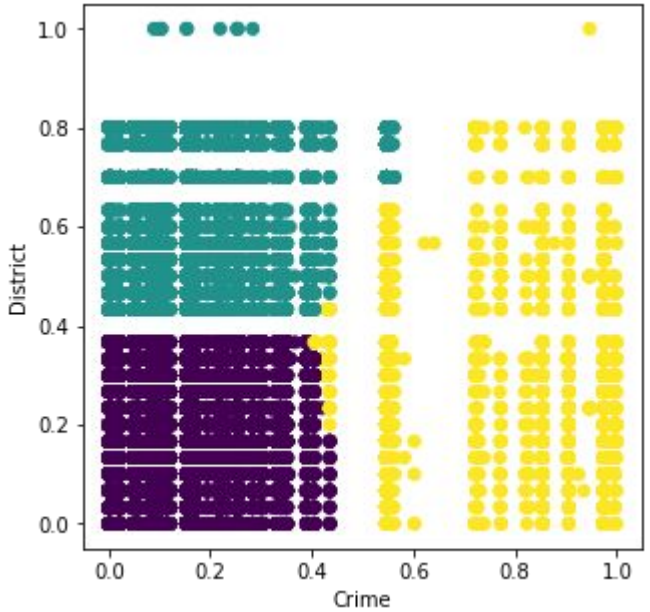
1. Dropping unnecessary columns like Beat, Case Number, FBI Code, Location co-ordinates, etc.
2. Splitting the date type data into date and time, and the time data into seconds, hours and months
3. We then normalized all these values to bring them to a scale of 0 to 1.

Clustering - Techniques used

1. K-Means clustering
2. K-Modes clustering
3. Hierarchical Agglomerative Clustering

K-Means

1. We implemented k-means for multiple combinations of dimensions
2. The most insightful one was for IUCR and District
3. Main finding - crimes such as IUCR Codes 5131 (other offence - violent offender), 4387 (Violation of Order of Protection), 3731 (Obstructing Identification) etc **occur much less frequently** than others



K-Modes

1. Extension of K-Means algorithm - works on categorical data, uses Hamming distance as the distance measure
2. Performed clustering on **Types of Crimes and Districts;** and **Crime types and Arrests**
3. The results helped validate results of kmeans clustering - most crimes occurring in most districts, led to new insights
4. District 7 is more prone to Crimes like 'Battery', 'Deceptive practice', 'Weapons violation', ‘Burglary'
5. Very few arrests were made for crimes such as 'Assault', 'Criminal Damage', 'Motor Vehicle Theft' out of all 31 unique crime types

Hierarchical Agglomerative Clustering (HAC)

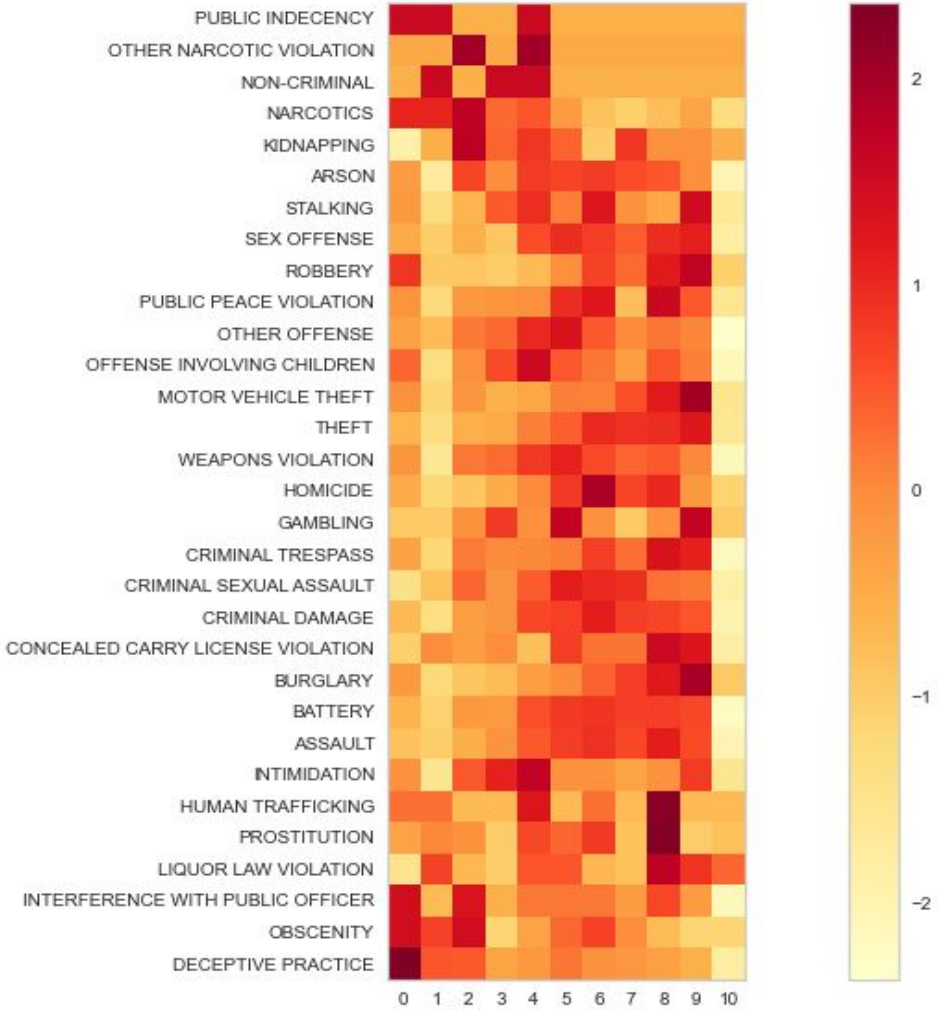
1. HAC was used to get much deeper and granular insights about the distribution of crime across various dimensions

2. Months and Crime types Findings -

   i) Deceptive Practice and Public Indecency occurred more frequently in January and February

   ii) Assault, Burglary, Homicide had the highest occurrences between June and October

   iii) Relatively fewer crimes have occurred in March and October



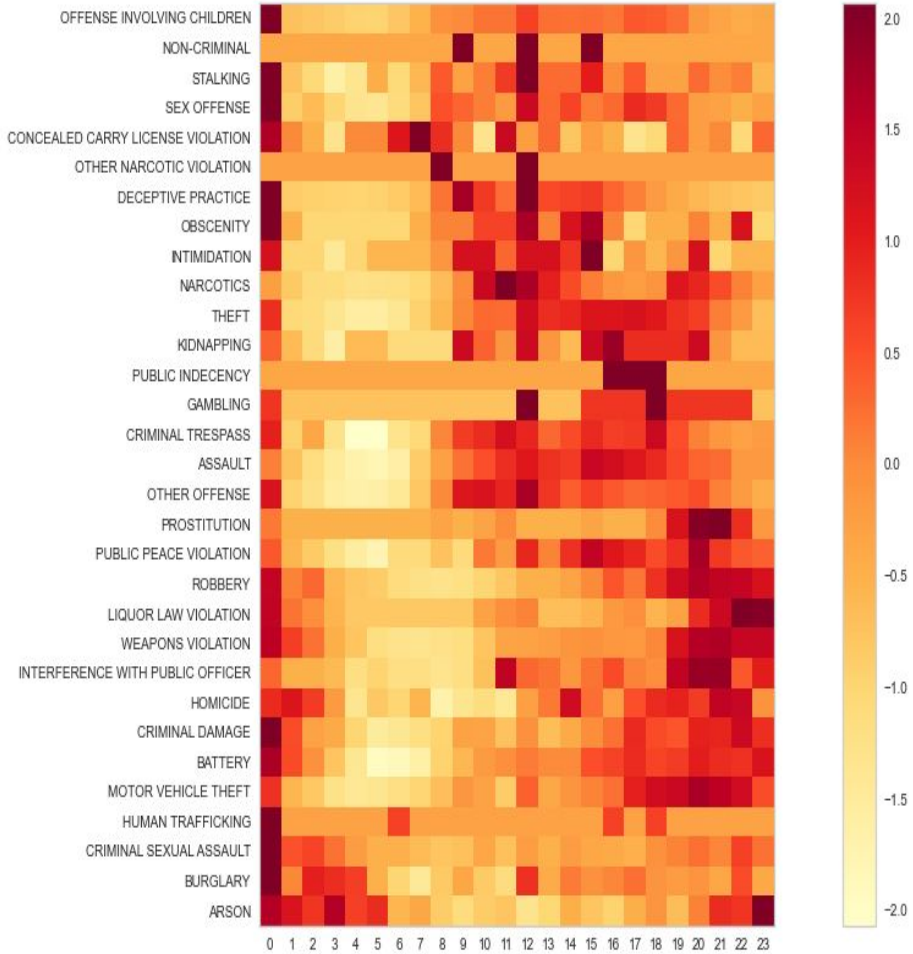
Hierarchical Agglomerative Clustering (HAC)

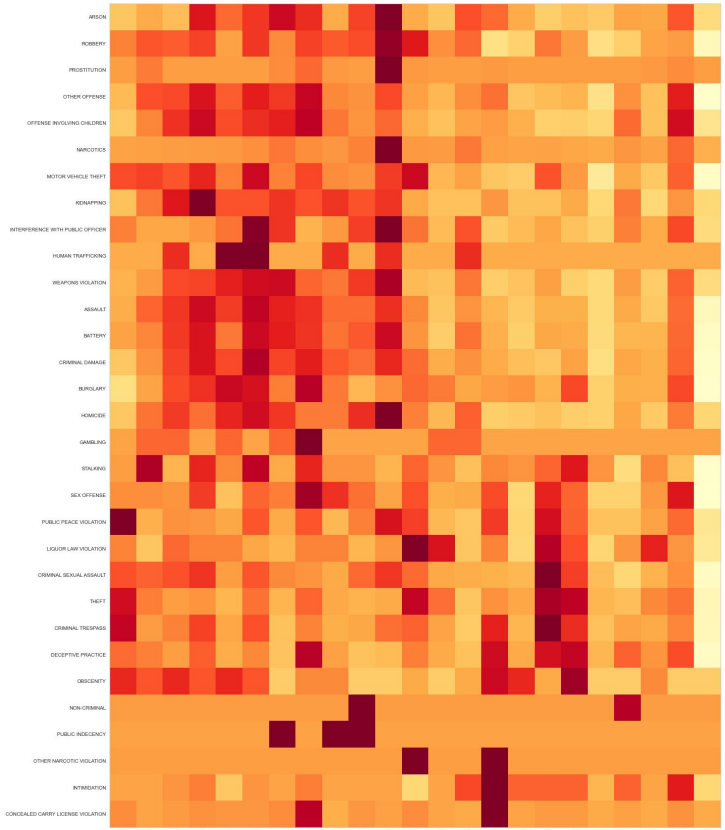
1. Hour and crime types findings -

   i) Arson, liquor law violation, crimes involving children, obscenity, homicides and sex offences occurred the most frequently between 10 pm and 1 am

   ii) Public indecency and other offences occurred more frequently in the afternoon

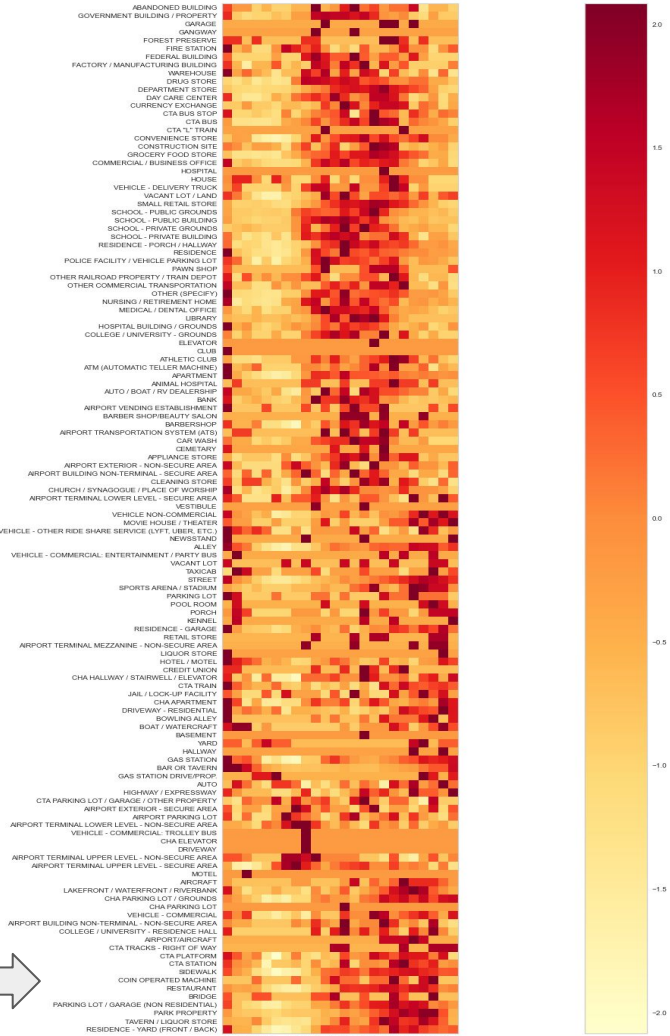
   iii) 6-8 am seem to be safer hours for the city





District and Crime types

Hour and Location of Crime



Thank you