Chicago Crime Data Analysis

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

GitHub: https://github.com/sandeshgupta/crime-data-analysis

Dataset: Crimes - 2001 to Present | City of Chicago | Data Portal

1. Objective

This project is aimed at performing an analysis of the crime of the past 20 years in the city of Chicago. Our goal is to mine the underlying hidden patterns, similarities and find relationships between the factors involved in the crime.

2. System design and Implementation

Association Rule Learning: We used this technique to find the relationship between different attributes in a dataset. Apriori and FP growth were the algorithms used here.

Clustering: We used this method to determine patterns in the crime occurring in Chicago based on the locations, the time and type of crimes, and the time of year.

3. Experiments / Proof of concept evaluation

3.1 Preprocessing

The original dataset contained more than 7 million records of crimes from 2001 till present. With crime data, it is important to understand the most recent crime patterns so as to take proactive measures to avoid it. Hence, the analysis was performed on current year 2021 data. We also performed analysis on Pre and Post Covid data that has data from the past 3 years. As part of this, this step involved a one-time effort of segregating the relevant data into individual files so as to avoid loading the whole 7M records into memory.

3.2 Implementation

The execution was performed as a two-part process. First part involved Association Rule Mining and the second was Clustering.

3.2.1 Association Rule Mining

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. It identifies the pattern of a consequent being present when an antecedent is present. In this scenario, association rule mining was performed on two different subsets of the Crime data. The methodology for rule mining was the same for both the subsets.

Preprocessing

Following preprocessing strategies were applied:

1. Drop redundant or unuseful columns: Following columns were dropped from the dataset as they were redundant or provided unuseful data: *ID, Case Number, IUCR, Beat, Updated On, Latitude, Longitude, Location, FBI Code, Block, X Coordinate, Y Coordinate, Date, Year*
2. Add new column: A new formatted *Month* column was added instead of *Date* as to limit the groups formed by grouping on Date. If there are 365 distinct dates, it would create 365 different groups and wouldn't be able to associate with other records.
3. Convert the record data to transaction data: Association analysis is usually performed on transaction data. For this, the record data was converted to tuples. Every value of a column should be uniquely identifiable in a transaction. Directly converting a record to a tuple wouldn't work. For eg: Value *12* in a tuple could be *District*, *Ward* or even *Month*. Hence, to uniquely identify individual column values, column name was prefixed to the value. For example, *12* in *District* was converted to *District-12*, *12* in *Month* was converted to *Month-12*.
4. Preprocessing for Clustering: Dropping the unnecessary columns such as Beat, Case number, FBI code which we felt would not be useful to the clustering analysis. Next, we limited the data to crimes that had occurred in 2021. This resulted in a dataframe containing 2,516,752 non-null values. We also split the dates into hours, months and dates. Lastly, we normalized the values of the district, hour and month values so that they could be used by our K-Means algorithm.

Methodology

In an effort to generate the rules, we tried 3 different algorithms for pattern mining from crimes of 2021.

1. Apriori algorithm: Apriori is an algorithm for frequent item set mining and association rule learning over transaction data. It proceeds by identifying the frequent individual items in the dataset and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the dataset. The preprocessed crime data was passed as an input to the Apriori implementation. It also takes hyperparameter *min_support* as an input. *Support* defines the frequency of the occurrence of an itemset. In this dataset, the attributes of the crime data were widely distributed. Hence, the support was kept to be in range of [0.05, 0.5]. *Confidence* gives the likeliness of the consequent being present, when an antecedent is present. The 2021 data was processed for multiple combinations of support, lift and confidence. Following image shows the number of rules generated for each combination.

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

Table 1.1

2. Efficient Apriori: It provides an widely used efficient implementation of Apriori as a Python package. The runtime in comparison to plain Apriori was remarkable. The only drawback for this was it didn’t return the exact support, confidence and lift of each itemset. This was required in this case to draw out the exact conclusions. The rules generated from this implementation were the same as that of the Apriori. The performance of this algorithm can be seen in table 1.2.

3. FP Growth: FP-growth is an improved version of the Apriori Algorithm which is widely used for frequent pattern mining. It optimizes the Apriori by reducing the disk I/O and subsequently computing power. It generates the FP-tree and uses it to directly generate large itemsets. It uses a recursive divide-and-conquer approach to mine frequent patterns from the large itemsets. The rules generated from this implementation were the same as that of the Apriori. The performance of this algorithm can be seen in table 1.2.

4. Comparison of the algorithms: Table 1.2 shows the comparison between the runtime of the different algorithms. Efficient Apriori was the fastest but lacked some data in the returned rules. FP Growth was approximately 10x faster than plain Apriori.

	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

Table 1.2

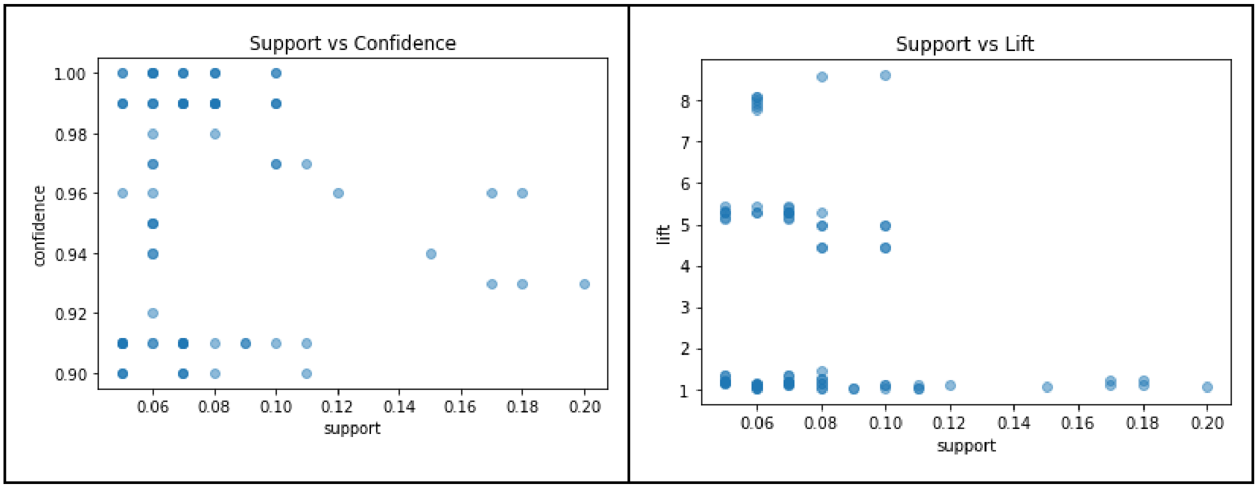
Analysis

Implementation was performed on 2 different types of data. First, the data of 2021 crimes was analysed to understand the recent patterns of crime. Second, the data was divided into 2 parts: Pre-Covid and Post-Covid. The intention was to find the difference in the crime patterns at both periods.

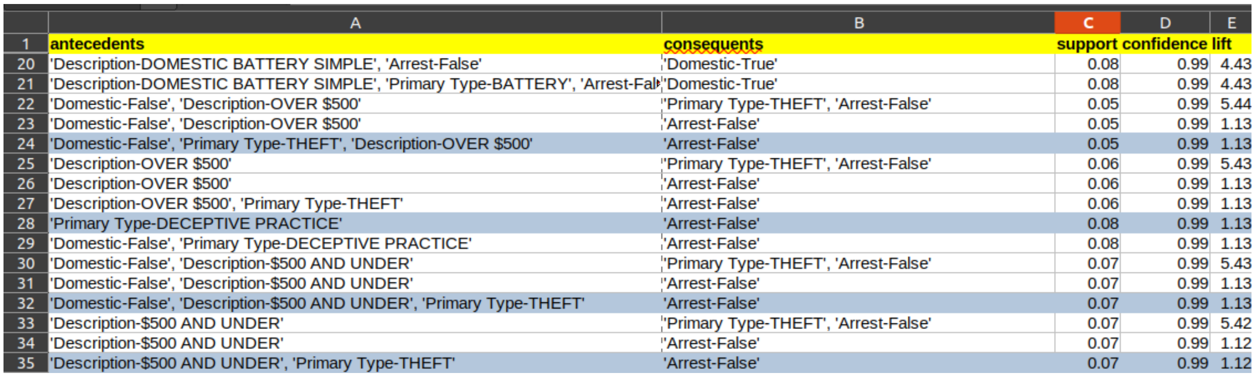
A. 2021 crime data

Dataset : Crime data of **01/01/2021 - 11/10/2021** - Total **158084 records**

Graphs: Fig 1.2 shows the distribution of Support vs Confidence and Support vs Lift.



Frequent patterns: In the Apriori analysis, 85 frequent patterns were generated for support 0.05, confidence 0.9 and lift 1. Some of the samples of the rules generated are shown in fig 1.2. Out of 85 rules generated, around 10 rules were meaningful(highlighted in blue), Rest others provided relations which were obvious.



	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

Fig 1.2

Result: After analysing all the rules, some of the interesting conclusions that can be drawn are as follows:

- 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 are 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.

B. Pre and Post Covid data

Dataset: Pre-Covid: Crime data of **03/20/2018 - 03/20/2020**. Total **528391 records**
Post-Covid: Crime data of **03/21/2020 - 10/11/2021**. Total **319531 records**

Frequent Patterns:

- Pre Covid: 82 frequent patterns were generated for support 0.05, confidence 0.5, and lift 1.
- Post-Covid: 95 frequent patterns were generated for support 0.05, confidence 0.5, and lift 1.

Out of these rules generated, we tried to analyze and compare them to generate any pattern which may have changed.

Result: After analyzing the rules generated we came to a conclusion as

- 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
- Thefts increased post covid significantly
- Pre-Covid, the months of August and September resulted in 15% of the crime whereas Post-Covid it increased to 20%, and yet no arrests were made in both the years.
- 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%

3.2.2 Clustering

Algorithms selected for clustering

We tried the following 4 clustering algorithms on the data -
i) K means clustering
K means clustering is one of the most popular clustering algorithms which uses Euclidean distance to find the similarity between data points and group them into clusters. We decided to use this algorithm as a benchmark for our analysis.

ii) K modes clustering
K modes algorithm is an extension of the kmeans algorithm and works on categorical data by comparing the dissimilarity measure between the data points using hamming distance. We used this algorithm as we felt that the results of our kmeans clustering were slightly biased by the fact that the attributes we wanted to use for our analysis were serialized (districts, crime IUCR codes and dates).

iii) Hierarchical Agglomerative Clustering
One of the issues with K-Means is that we need to determine the value of k beforehand; whereas for hierarchical clustering there is no such restriction. We decided to use this algorithm to have much deeper and granular insights about the distribution of crime across various dimensions. This was further possible by plotting the distribution as a set of heat maps.

iv) K prototype clustering
We wanted to go one step beyond kmodes clustering to see if when we used both numerical and categorical data of our dataset, we could achieve some more useful results. Kprototype clustering supports this.

Technologies and tools used

The following libraries were used for the clustering part of the clustering analysis-
Pandas, Numpy, Gensim, Scikit learn, Yellowbrick,Matplotlib

Basic analysis of data

We plotted the frequency of each crime type, the overall frequency of occurrence of crimes in the year of 2021 and the frequency of crimes on days of the week. This helped us see that the frequency of crimes over the year was varied and highest in September. The frequency over days of the week did not give us any further information so we decided to not pursue analysis using dates. Also, it was observed that the top 10 crimes were occuring at a much higher frequency than the rest of the crimes so we tried to do kmeans clustering on just those rows.

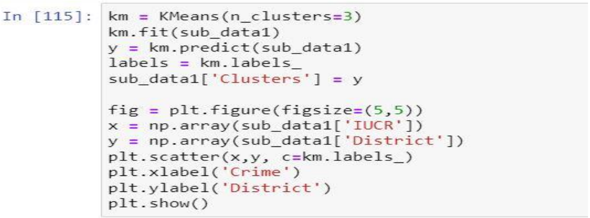
Clustering

The 4 algorithms listed above - kmeans, kmodes, kprototype and Hierarchical Agglomerative clustering were applied to the data and the results were plotted as scatter plots and heat maps. We also printed out the results of kmodes clustering to compare them with our finding from kmeans clustering.

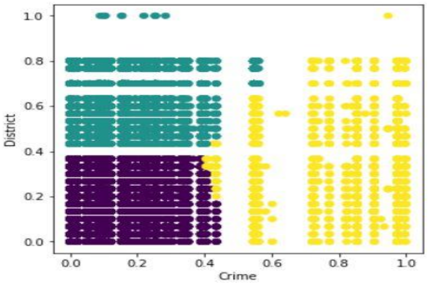
Evaluation of methods used: Silhouette score, Calinski harabasz score

Results and analysis

K means clustering on type of crime and districts

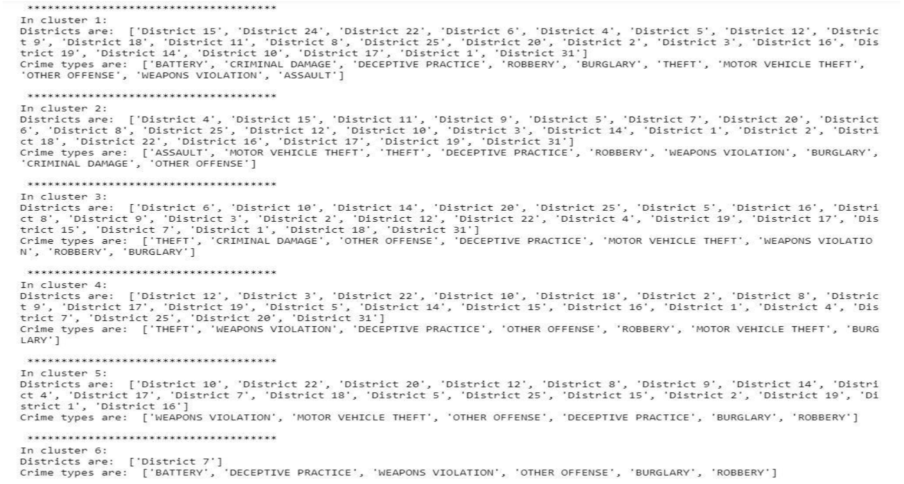


```
In [115]: km = KMeans(n_clusters=3)
          km.fit(sub_data1)
          y = km.predict(sub_data1)
          labels = km.labels_
          sub_data1['Clusters'] = y

          fig = plt.figure(figsize=(5,5))
          x = np.array(sub_data1['IUCR'])
          y = np.array(sub_data1['District'])
          plt.scatter(x,y, c=km.labels_)
          plt.xlabel('Crime')
          plt.ylabel('District')
          plt.show()
```



Our analysis on this was that most of the crimes seemed to occur in almost all districts but the frequency of certain crimes denoted by the yellow cluster was much lesser than other crimes. Example - 5131 (other offence - violent offender: failed to register new address)

K modes clustering on type of crime and districts



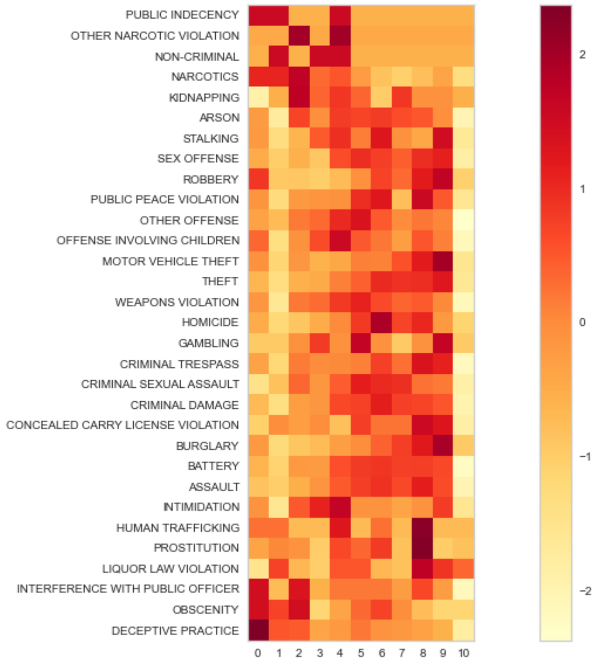
```
 *************************************
In cluster 1:
Districts are:  ['District 15', 'District 24', 'District 22', 'District 6', 'District 4', 'District 5', 'District 12', 'Distric
t 9', 'District 18', 'District 11', 'District 8', 'District 25', 'District 20', 'District 2', 'District 3', 'District 16', 'Dis
trict 19', 'District 14', 'District 10', 'District 17', 'District 1', 'District 31']
Crime types are:  ['BATTERY', 'CRIMINAL DAMAGE', 'DECEPTIVE PRACTICE', 'ROBBERY', 'BURGLARY', 'THEFT', 'MOTOR VEHICLE THEFT',
'OTHER OFFENSE', 'WEAPONS VIOLATION', 'ASSAULT']

 *************************************
In cluster 2:
Districts are:  ['District 4', 'District 15', 'District 11', 'District 9', 'District 5', 'District 7', 'District 20', 'District
6', 'District 8', 'District 25', 'District 12', 'District 10', 'District 3', 'District 14', 'District 1', 'District 2', 'Distri
ct 18', 'District 22', 'District 16', 'District 17', 'District 19', 'District 31']
Crime types are:  ['ASSAULT', 'MOTOR VEHICLE THEFT', 'THEFT', 'DECEPTIVE PRACTICE', 'ROBBERY', 'WEAPONS VIOLATION', 'BURGLARY',
'CRIMINAL DAMAGE', 'OTHER OFFENSE']

 *************************************
In cluster 3:
Districts are:  ['District 6', 'District 10', 'District 14', 'District 20', 'District 25', 'District 5', 'District 16', 'Distri
ct 8', 'District 9', 'District 3', 'District 2', 'District 12', 'District 22', 'District 4', 'District 19', 'District 17', 'Dis
trict 15', 'District 7', 'District 1', 'District 18', 'District 31']
Crime types are:  ['THEFT', 'CRIMINAL DAMAGE', 'OTHER OFFENSE', 'DECEPTIVE PRACTICE', 'MOTOR VEHICLE THEFT', 'WEAPONS VIOLATIO
N', 'ROBBERY', 'BURGLARY']

 *************************************
In cluster 4:
Districts are:  ['District 12', 'District 3', 'District 22', 'District 10', 'District 18', 'District 2', 'District 8', 'Distric
t 9', 'District 17', 'District 19', 'District 5', 'District 14', 'District 15', 'District 16', 'District 1', 'District 4', 'Dis
trict 7', 'District 25', 'District 20', 'District 31']
Crime types are:  ['THEFT', 'WEAPONS VIOLATION', 'DECEPTIVE PRACTICE', 'OTHER OFFENSE', 'ROBBERY', 'MOTOR VEHICLE THEFT', 'BURG
LARY']

 *************************************
In cluster 5:
Districts are:  ['District 10', 'District 22', 'District 20', 'District 12', 'District 8', 'District 9', 'District 14', 'Distri
ct 4', 'District 17', 'District 7', 'District 18', 'District 5', 'District 25', 'District 15', 'District 2', 'District 19', 'Di
strict 1', 'District 16']
Crime types are:  ['WEAPONS VIOLATION', 'MOTOR VEHICLE THEFT', 'OTHER OFFENSE', 'DECEPTIVE PRACTICE', 'BURGLARY', 'ROBBERY']

 *************************************
In cluster 6:
Districts are:  ['District 7']
Crime types are:  ['BATTERY', 'DECEPTIVE PRACTICE', 'WEAPONS VIOLATION', 'OTHER OFFENSE', 'BURGLARY', 'ROBBERY']
```

These results helped us validate the results of our kmeans clustering (most crimes occurring in most districts) and also led to a new insight - District 7 is more prone to Crimes like 'BATTERY', 'DECEPTIVE PRACTICE', 'WEAPONS VIOLATION', 'OTHER OFFENSE', 'BURGLARY', 'ROBBERY'.

Hierarchical Agglomerative Clustering

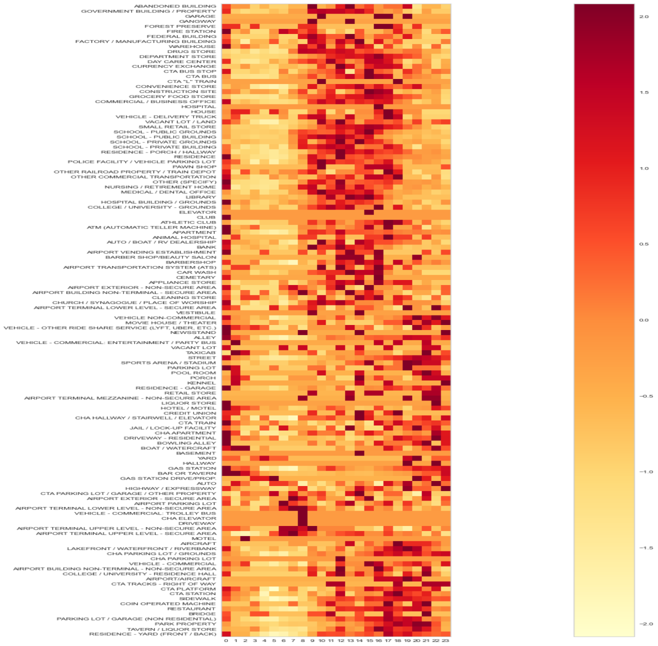
We used HEAT MAPS to plot HAC results. Following were our findings:

<u>1. Crime type VS Month</u>



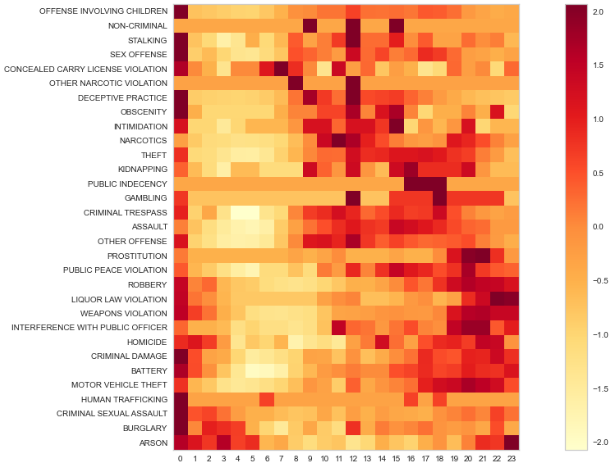
Deceptive Practice and Public Indecency occur more frequently in the months of January and February. Crimes such as Assault, Burglary, crimes involving Battery and Homicide have the highest occurrences between June and October. Relatively fewer crimes have occurred in March and October.

2. Hour of the day VS Location of the crime



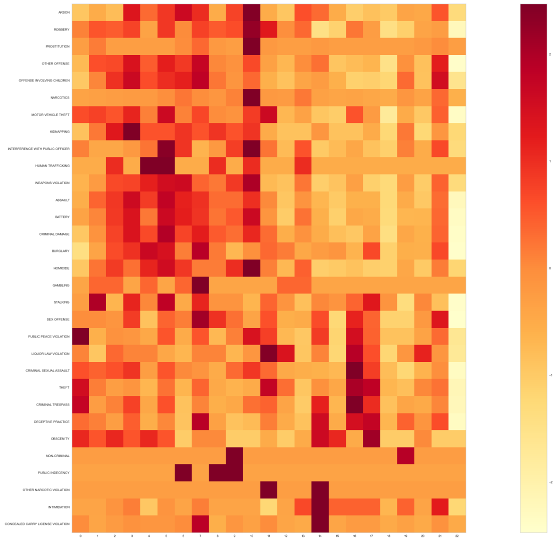
College grounds, retail stores, schools, libraries, bus stops and residences are the most vulnerable from in the afternoons till 5 pm. Theatres and CTA tracks are more vulnerable from 11 pm to 1 am.

3. Hour of the day VS type of the crime



Arson, liquor law violation, crimes involving children, obscenity, homicides and sex offences occur the most frequently between 10 pm and 1 am. Public indecency and other offences occur more frequently in the afternoon. 6-8 am seem to be safer hours for the city.

4. District VS type of the crime



Concealed carry violation, narcotic violations and intimidation occur the most frequently in District 14. Criminal sexual assult occurs the most in District 16. Human trafficking occurs most in Districts 4 & 5.

4 Discussion & Conclusions

4.1 Difficulties faced

The initial file size was huge(1.8 GB) and was killing the kernel. Splitting the file according to the required helped to smoothen the further analysis process.

4.2 Things that worked

Association rule mining led to identification of meaningful patterns related to arrests in some specific locations, months of the year and the crimes. In Clustering, HAC, KMeans and KModes worked very well with our dataset and we were able to mine useful clustering insights from them.

4.3 Things that didn’t work well

There were expectations that Pre Covid and Post Covid analysis would provide some interesting patterns around the different types of crimes. We tried K-Means for three dimensional data (with attributes like Crime, District, and Time(Seconds and Month)), but the clustering was not coherent, within cluster distances were too far off; and the clustering result did not make much sense. K-Prototypes results were not convincing either since major portions of the dataset were equally present in the clusters and the clusters were not actually distinct. We also tried converting categorical data to numeric data by using ‘gensim’, Word2Vec using Google's pre-trained model to convert textual crime types into numeric values and perform k-Means on that data.

4.5 Conclusion

Association rule mining led to believe that arrest rates are very low in districts 8 and 9. It is also low for the most common occurring crimes like Theft, Damages to vehicle, Crimes committed at residences and apartments.

Clustering led us to believe that even though a lot of the top crimes in Chicago occur in almost all the districts, there are some that occur much less frequently than others and most prominently at certain hours of the day such as 11pm - 1am and afternoon hours. There is also a correlation between crimes & months and districts & arrests. We also found out that sensitive places such as schools and libraries are the most vulnerable to crimes during the afternoon hours when they must be the busiest. Also, we saw that there are more occurrences of arrests not happening for crimes like 'ASSAULT', 'CRIMINAL DAMAGE', 'DECEPTIVE PRACTICE', 'ROBBERY', 'BURGLARY', 'MOTOR VEHICLE THEFT', 'OTHER OFFENSE', ‘BATTERY’, ‘THEFT’ out of all the 31 unique crime types.

5. Project Plan / Task Distribution

Sandesh Gupta

- Generate all data required for association rule mining: last 2 years, last 4 years,2021 data, Pre covid and post covid
- Pre-process data for Association Rule Mining

Decide relevant columns, Drop irrelevant columns, Format columns, Convert records to transaction data

- Research on how to run Apriori, Efficient Apriori, FP Growth
- Run Association analysis on 2021 data

  Run the algorithms, format itemsets and write to CSV
- Analyse and formulate the rules post Association Rule Mining. Interpret the rules and derive conclusions.

Aomkar Mathakar

- HAC
- K-Means with 3 dimensional data visualization

Mugdha Gumphekar

- K-Means
- K-Modes
- K-Prototypes
- Gensim, Word2Vec to convert categorical attribute data to numeric values and perform k-means on it.

Venkata Prithvi Raj Namburi

- Run association analysis on pre and post covid data
- Analyse and formulate the rules post Association Rule Mining. Interpret the rules and derive conclusions.

6. References

https://medium.com/analytics-vidhya/association-analysis-in-python-2b955d0180c
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data
https://www.geeksforgeeks.org/implementing-apriori-algorithm-in-python/
https://matplotlib.org/stable/gallery/color/colormap_reference.html
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
https://ai.plainenglish.io/k-means-and-k-modes-clustering-algorithm-4ff51395fa8d