CHICAGO CRIME DATA ANALYSIS

Group 3: -
Deepti Khatri
Preethi Kannan
Rakesh Jain
Sanket Bhaud

Contents

ABSTRACT

PROBLEM

There has an abundance of crimes committed on a daily basis in Chicago, one of the most populated states in the United States. In this project, we use Chicago Crime dataset which reflects reported incidents of crime that occurred in the City of Chicago from 2012 to 2016. Our goal is to help the Chicago police department and local public by identifying the most occurring crimes, observe the trend over the years, and find the highly prone crime areas.

METHODS USED

The methods used focus on cleaning the data followed by preliminary analysis, comprising temporal and spatial analysis. Also, the data contains mostly categorical variables, so multi class classification techniques are used for predictive modelling which predicts the crime type.

RESULTS OBTAINED

Our analysis mainly focuses on temporal and spatial visualization along with finding a suitable model to predict the primary crime type.

MATERIALS AND METHODS

DATASET DESCRIPTION

Our dataset is taken from Kaggle website() spanning five years of data. It contains around a million records and 23 columns.

1. unique_id: - Unique identifier for the record.
2. date: - Date when the incident occurred. this is sometimes a best estimate.
3. primary_type: - The primary description of the IUCR code.
4. description: - The secondary description of the IUCR code, a subcategory of the primary description.
5. loc_desc: - Description of the location where the incident occurred.
6. arrest: - Indicates whether an arrest was made.

7. beat: - Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.
8. district: - Indicates the police district where the incident occurred.
9. ward: - The ward (City Council district) where the incident occurred.
10. community_area: - Indicates the community area where the incident occurred. Chicago has 77 community areas.
11. year: - Year the incident occurred.
12. latitude: - The latitude of the location where the incident occurred.
13. longitude: - The longitude of the location where the incident occurred.
14. location: - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal.

TOOLS AND TECHNIQUES

The methods used to build the model are based on predictive modelling and exploratory analysis.

DATA PRE-PROCESSING

The Steps used for data pre-processing are as follows: -

- Removed 2017 data since it was incomplete.
- Feature Extraction: Date column categorised into morning, afternoon, evening and night.
- Feature Grouping: Crimes types into crime categories.
- Removed missing values from latitude and longitude columns.
- Dropped unused levels from Primary Type
- Categorized crimes into Violence and Non - Violence and location description by using “%n%” method.
- Using POSIX function to transform the datetime into standard format

EXPLORATORY ANALYSIS

- Performed Spatial analysis using Leaflet to visualize crime hotspots and its data on interactive maps.
- Temporal analysis was performed to see the crime patterns through the day using plots and graphs.

PREDICITVE MODELLING

Based on the preliminary findings, the problem reduced to a classification problem to predict the primary crime type. The various methods we used for predictive modelling are Multinomial Logistic Regression, Random Forest and Support Vector Machines.

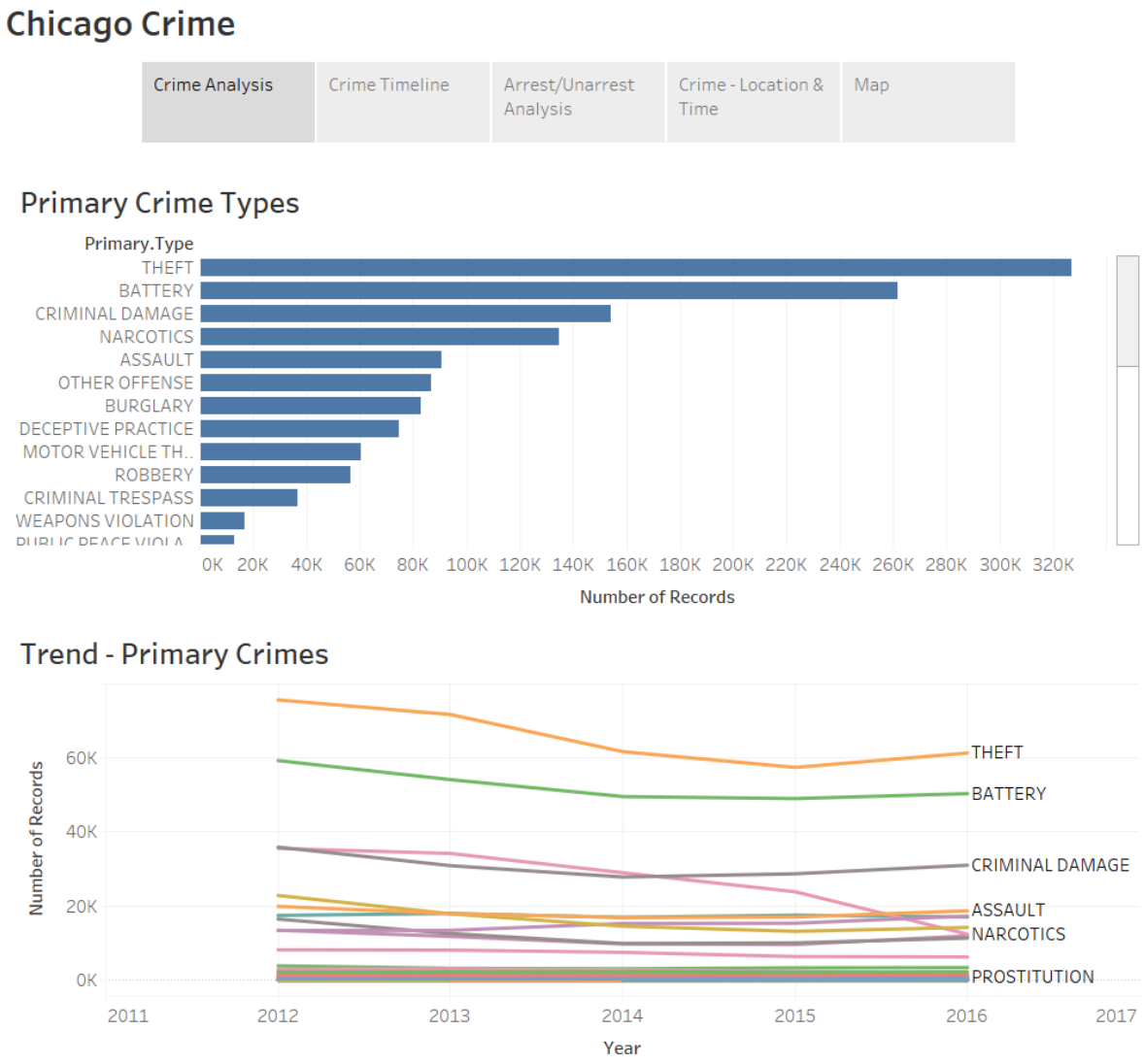
TOOLS USED

R and Tableau**.**

RESULTS AND DISCUSSION

RESULTS

Following are the snapshots of the stories for data exploration**: -**

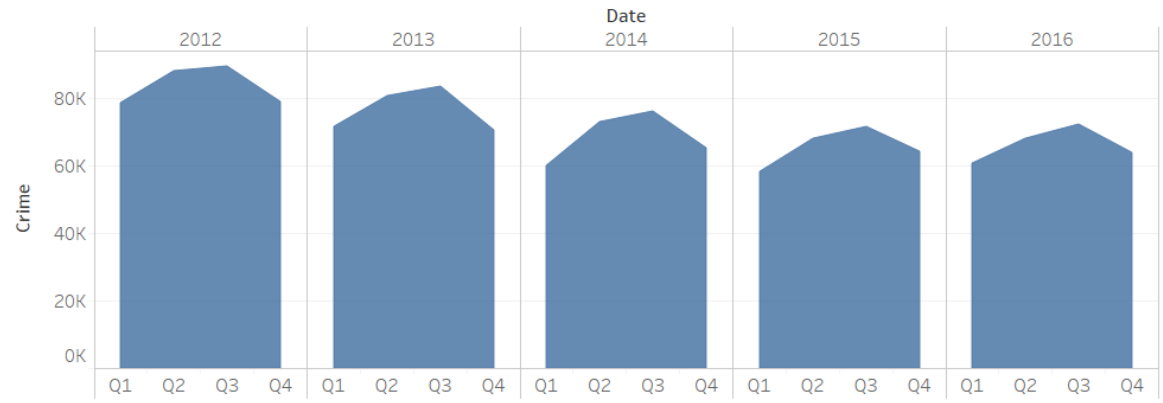


- Above plots show that crime rate of theft, battery, criminal damage, assault and narcotics are highest with the narcotics decreasing drastically over the years.

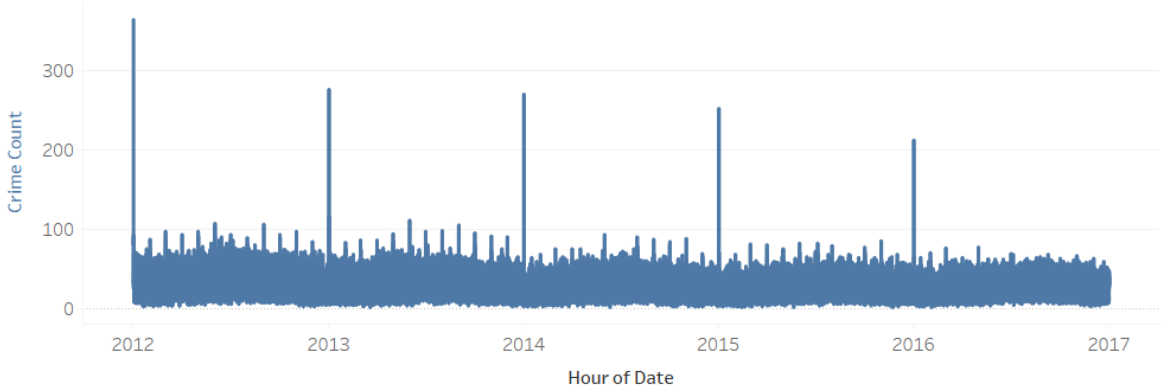
Chicago Crime

Crime Analysis	Crime Timeline	Arrest/Unarrest Analysis	Crime - Location & Time	Map

Quarterly Crime Analysis



Crimes By Hour

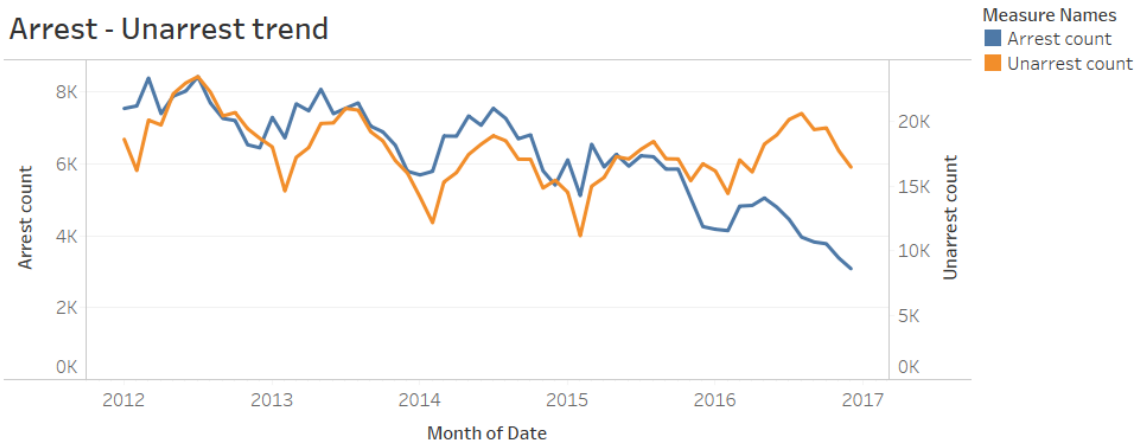


- Most crimes happen in Quarter 2 and 3, that is when weather is good, in summer and fall. When people are out, criminals get better chance of committing crimes.
- Crimes by hour plot shows that during every new year i.e. Jan 1, 1 AM crimes are at its peak.

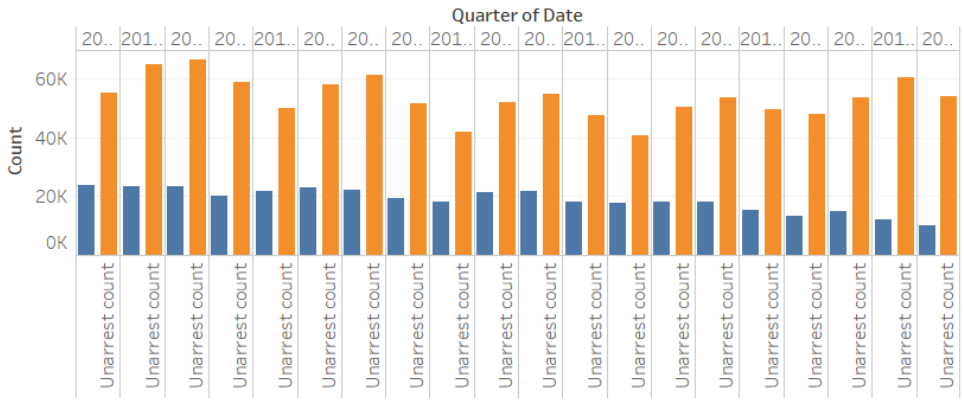
Chicago Crime

Crime Analysis | Crime Timeline | Arrest/Unarrest Analysis | Crime - Location & Time | Map

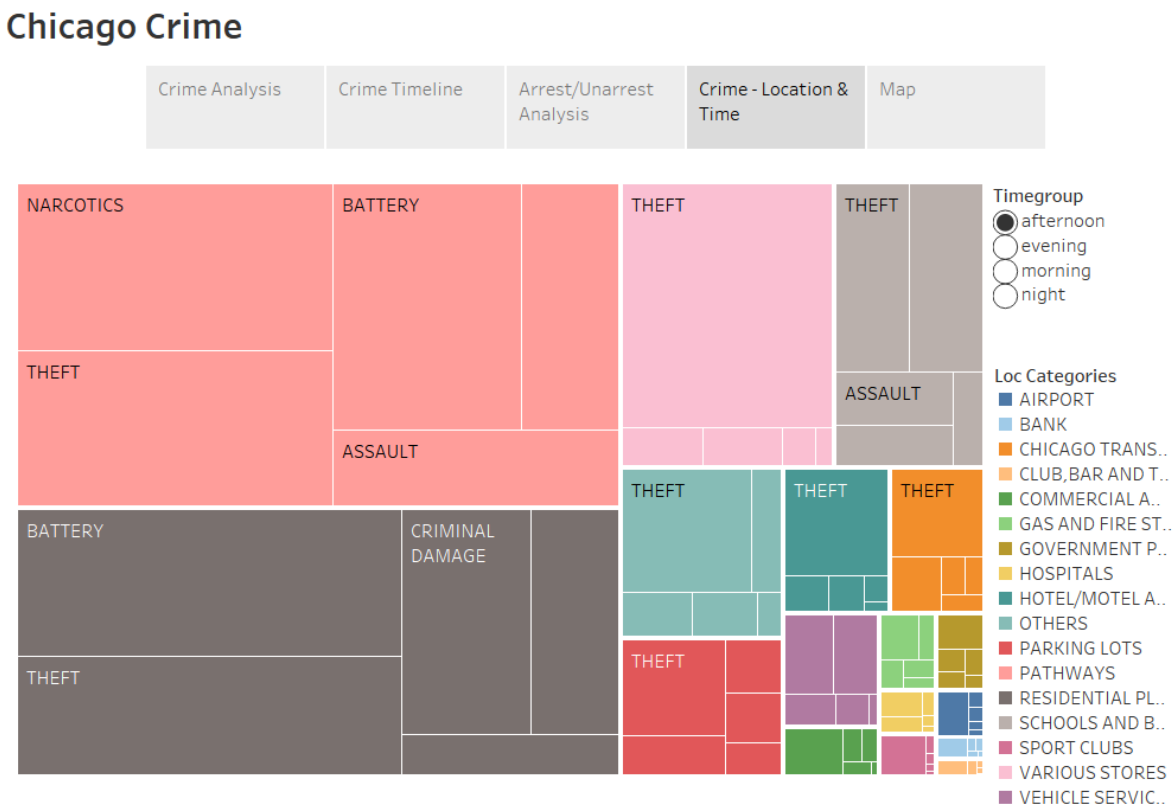
Arrest - Unarrest trend



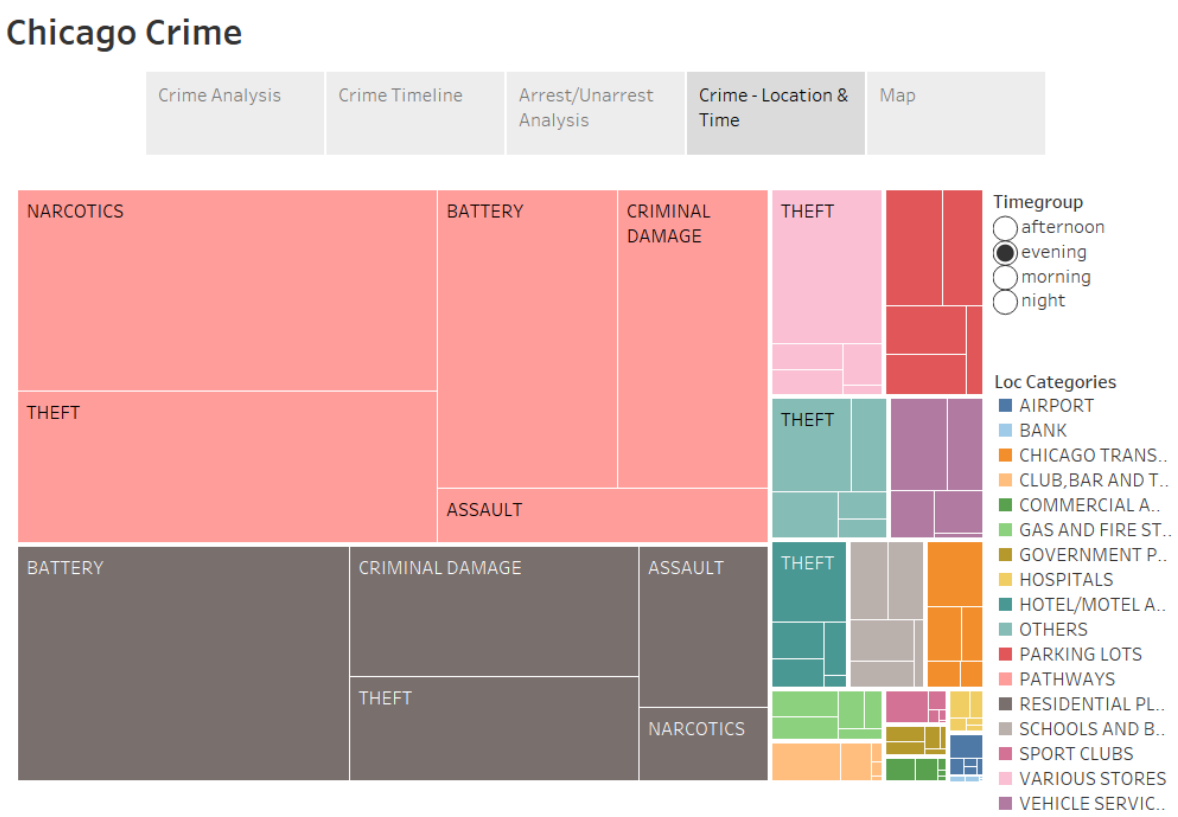
Arrest - Unarrest Count



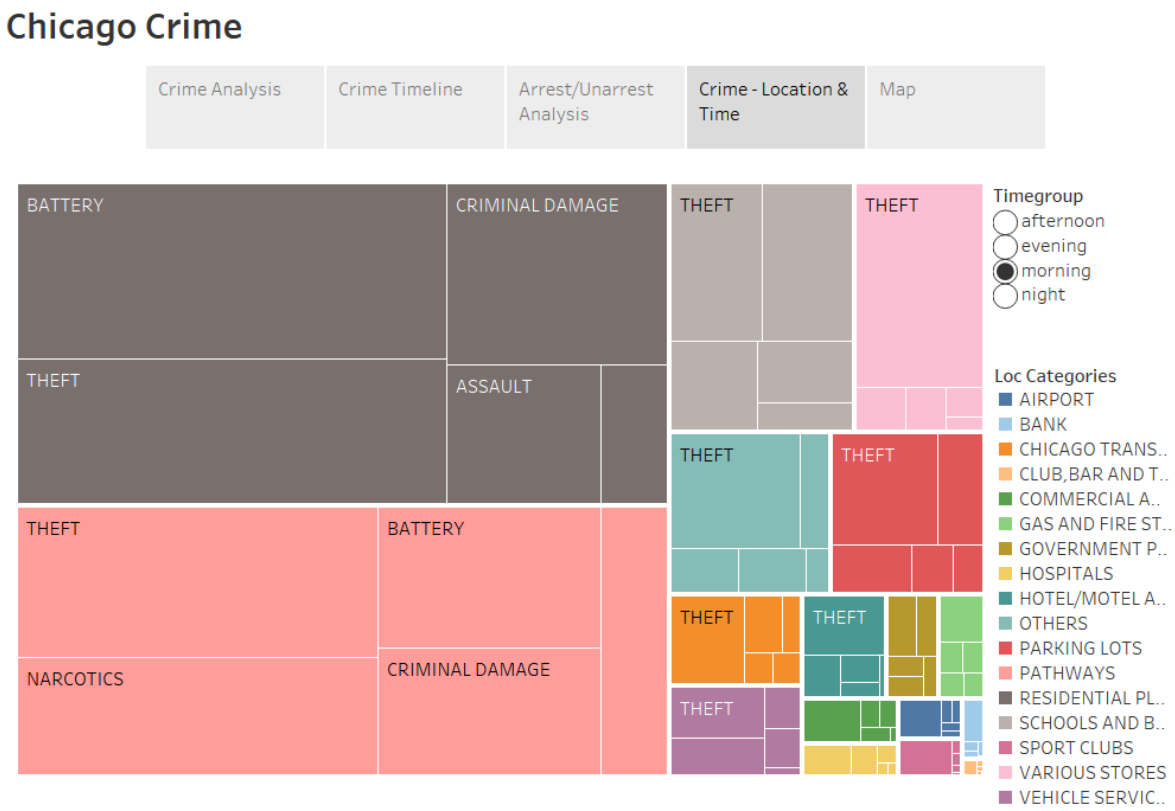
- Non arrest count is higher than the arrest count. Police department need to take serious action to arrest the criminals, as we can see non arrest count is decreasing in year 2016.



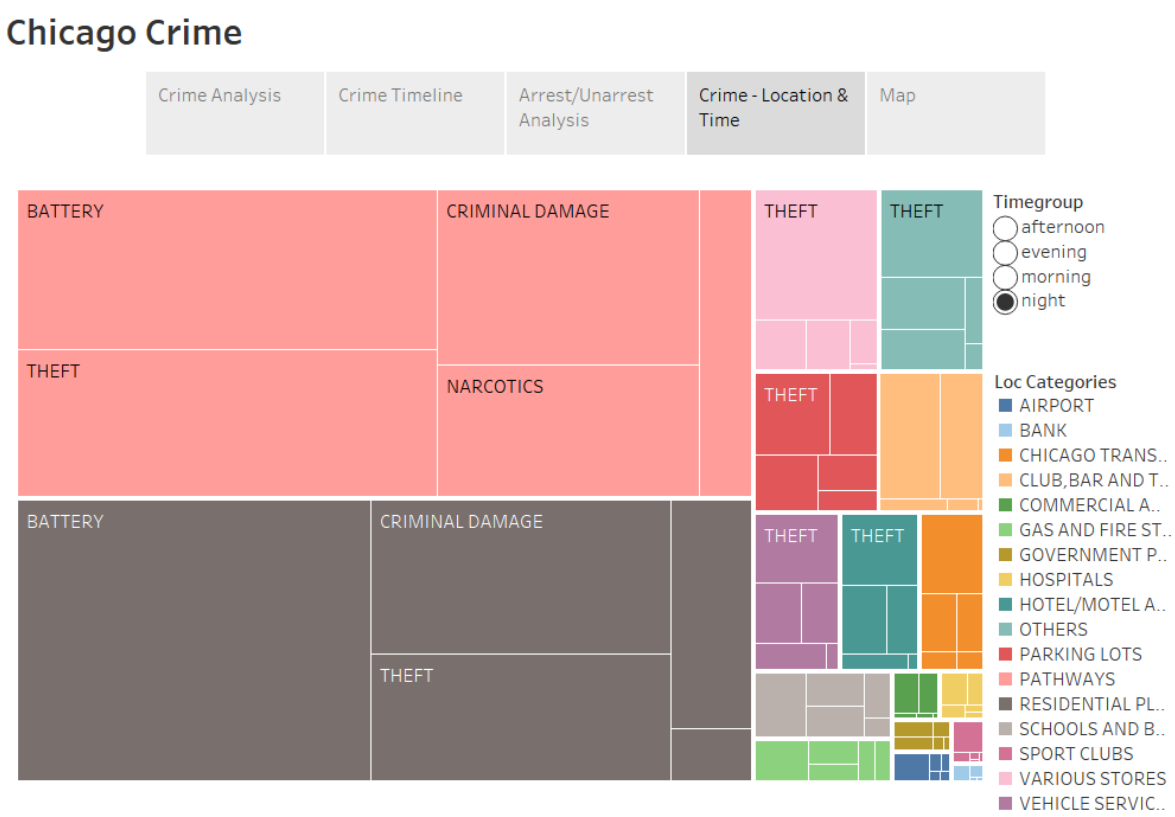
- Plot shows crime and its location for afternoon. During afternoon most crimes happen on pathways.



- Plot shows crime and its location every evening. During evening most crimes happens on pathways.



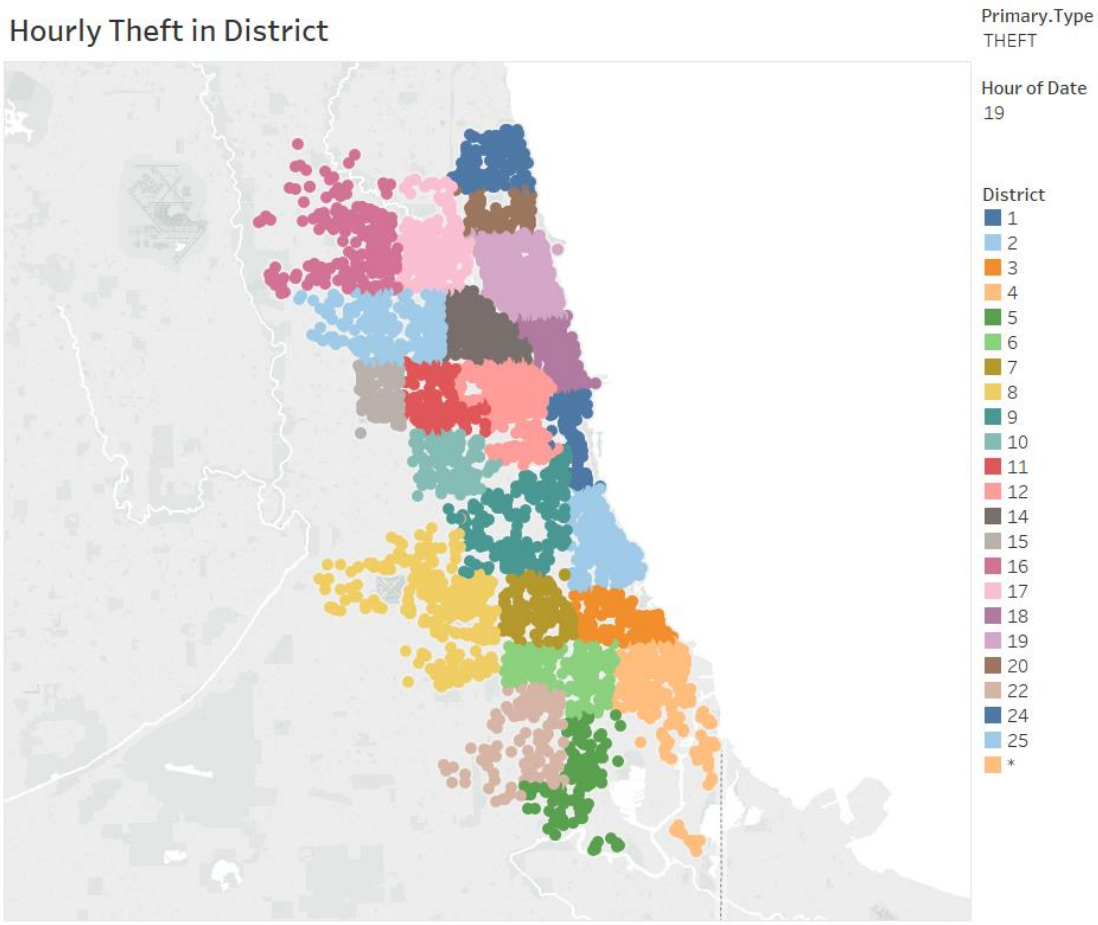
- Plot shows crime and its location every morning. During morning most crimes happens in Residential places.



- Plot shows crime and its location every night. During night most crimes happens on pathways.

Chicago Crime

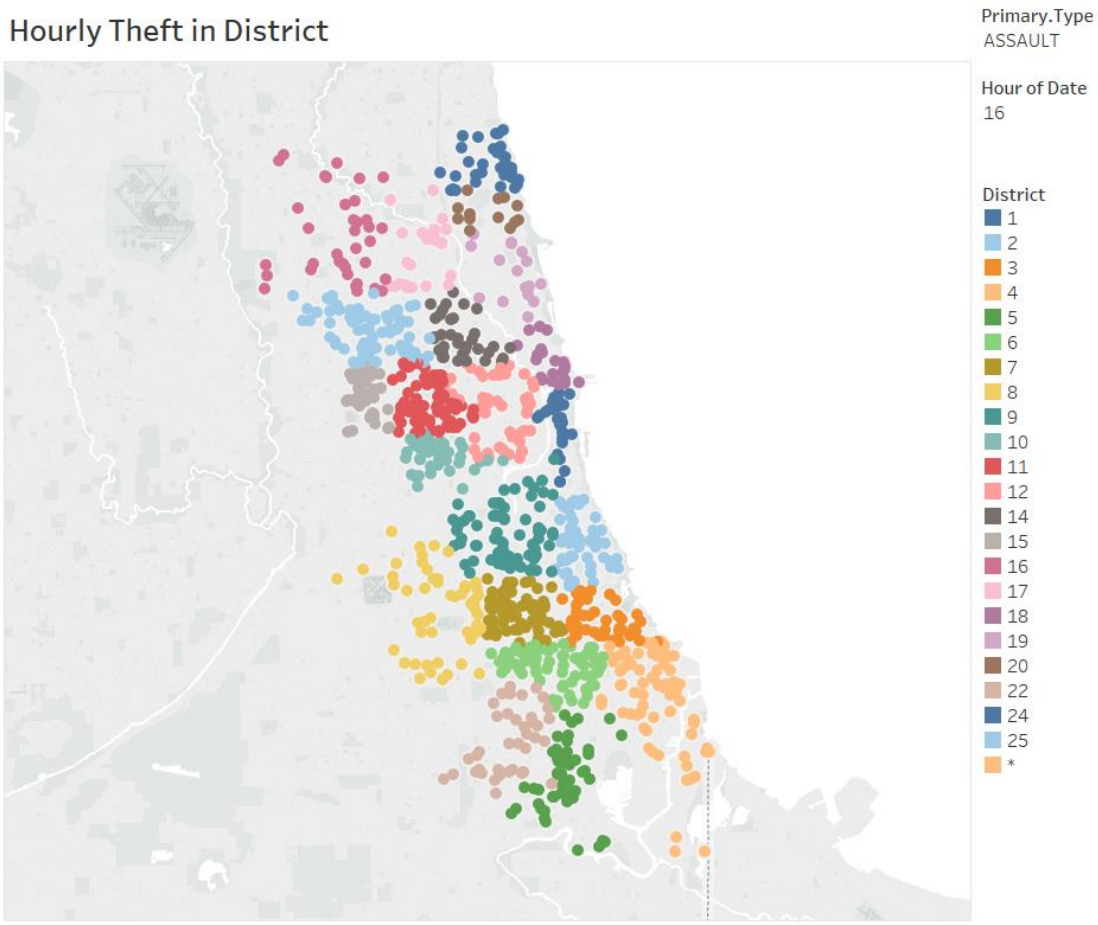
Crime Analysis	Crime Timeline	Arrest/Unarrest Analysis	Crime - Location & Time	Map



- Maps show theft count district wise in Chicago city at 1900 Hrs.

Chicago Crime

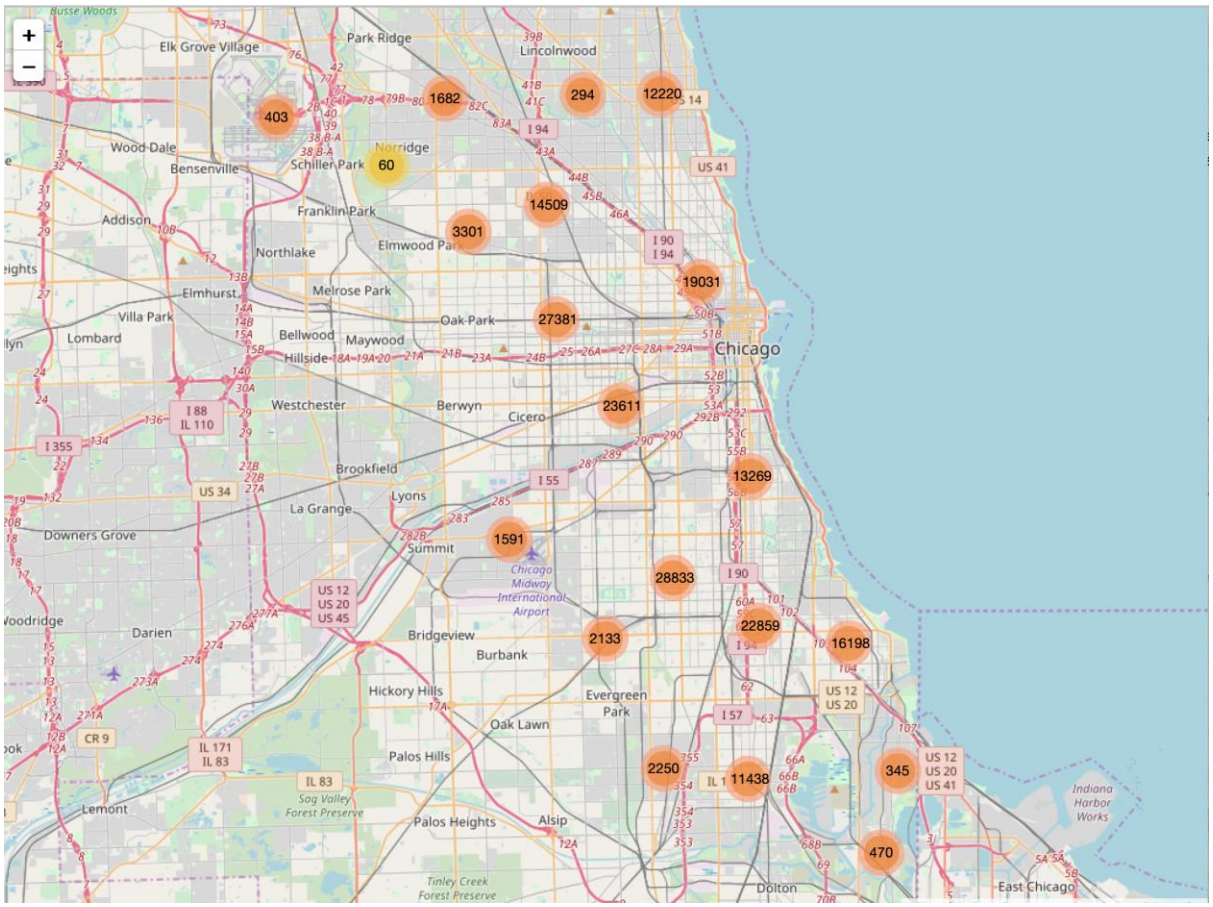
Crime Analysis	Crime Timeline	Arrest/Unarrest Analysis	Crime - Location & Time	Map



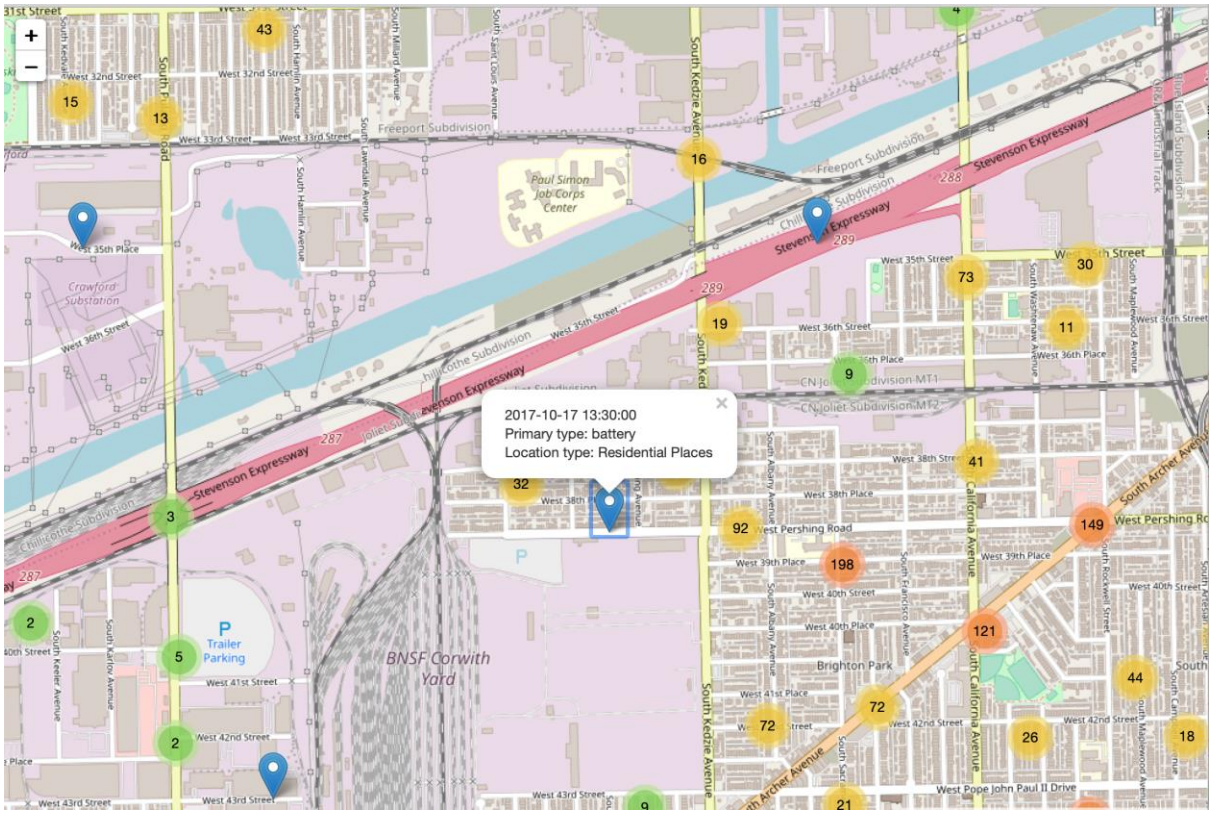
- Maps show assault count district wise in Chicago city at 1600 Hrs.

Chicago Crime

Leaflet (maps)



- This is the spatial map showing the distribution of the total number of crimes committed in the Chicago city.
- We see that most of the crime distribution is in the Illinois region towards the bay area.



- As we zoom in, we can scroll to the actual location where the crime happened with its specific details.
- We can also see the colour coded distribution highlighting highest number of crimes in red while the least number of crimes in green.

Following are the results of Predictive Analysis :-

- **Multinomial Logistic regression:** -

```
multinom(formula = Primary.Type ~ . - Date - Crimehour - Block,
    data = TrainSet, maxit = 500)
```

```
Confusion Matrix and Statistics

                 Reference
Prediction        ASSAULT BATTERY CRIMINAL DAMAGE NARCOTICS THEFT
  ASSAULT              26      17               0         0     0
  BATTERY             756    2260               0         0     0
  CRIMINAL DAMAGE       0       0             362         3   276
  NARCOTICS             0       0              61      1181    74
  THEFT                 0       0             972        22  2580

Overall Statistics

               Accuracy : 0.7461
                 95% CI : (0.7368, 0.7553)
    No Information Rate : 0.3411
    P-Value [Acc > NIR] : < 2.2e-16
```

- **Random Forest**

```
                 Reference
Prediction        ASSAULT BATTERY CRIMINAL DAMAGE NARCOTICS THEFT
  ASSAULT             128     193               0         0     0
  BATTERY             654    2084               0         0     0
  CRIMINAL DAMAGE       0       0             524         8   518
  NARCOTICS             0       0              57      1163    70
  THEFT                 0       0             814        35  2342

Overall Statistics

               Accuracy : 0.7265
                 95% CI : (0.717, 0.7359)
    No Information Rate : 0.3411
    P-Value [Acc > NIR] : < 2.2e-16
```

- **SVM**

```
> model = svm(Primary.Type ~ .- Date - Crimehour - Block, data = train1)
> model

Call:
svm(formula = Primary.Type ~ . - Date - Crimehour - Block, data = train1)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.03225806

Number of Support Vectors:  11839
```

```
> accuracy = sum(diag(table(test1$Primary.Type,pred)))/nrow(test1)
> accuracy
[1] 0.7454016
```

```
> confusionMatrix(data=pred,reference=test1$Primary.Type,positive='yes')
Confusion Matrix and Statistics

                 Reference
Prediction        ASSAULT BATTERY CRIMINAL DAMAGE NARCOTICS THEFT
  ASSAULT               0       0               0         0     0
  BATTERY             775    2314               0         0     0
  CRIMINAL DAMAGE       0       0             149         0    90
  NARCOTICS             0       0              81      1169    95
  THEFT                 0       0            1128        18  2771

Overall Statistics

               Accuracy : 0.7454
                 95% CI : (0.736, 0.7546)
    No Information Rate : 0.3441
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6465
 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: ASSAULT	Class: BATTERY	Class: CRIMINAL DAMAGE
Sensitivity	0.00000	1.0000	0.10972
Specificity	1.00000	0.8765	0.98756
Pos Pred Value	NaN	0.7491	0.62343
Neg Pred Value	0.90978	1.0000	0.85523
Prevalence	0.09022	0.2694	0.15809
Detection Rate	0.00000	0.2694	0.01735
Detection Prevalence	0.00000	0.3596	0.02782
Balanced Accuracy	0.50000	0.9383	0.54864

	Class: NARCOTICS	Class: THEFT
Sensitivity	0.9848	0.9374
Specificity	0.9762	0.7966
Pos Pred Value	0.8691	0.7074
Neg Pred Value	0.9975	0.9604
Prevalence	0.1382	0.3441
Detection Rate	0.1361	0.3226
Detection Prevalence	0.1566	0.4560
Balanced Accuracy	0.9805	0.8670

MODEL	Logistic Regression	Random Forest	SVM
TEST ERROR RATE	75%	73%	75%

- So, we can say that the validation error rate of Multinomial Logistic Regression and SVM is the highest.
- However, since SVM almost gave the same error rate after tuning the hyperparameters, we can multinomial logistic regression is the best model for prediction.

DISCUSSION

- From the interactive spatial plots, we see that residential areas were the highly crime prone areas.
- The temporal analysis showed most crimes happening between spring and fall every year.
- The above results would help the local public who wish to transfer to different areas in the city or to be cautious when they are out on the roads.
- Based on the experimental results, multinomial logistic regression model performed well in prediction. The proposed model finds its use in law enforcement department to deploy more police forces in highly crime prone areas.

LIMITATIONS

- Data imbalance issue is present since there were more records for theft than any other crime types.
- Influences of crimes based on the presence of beat in that location.
- Not able to generate heat maps on geospatial map as google api's were removed.

FUTURE SCOPE

- Use stratified sampling and feature selection to get better models for predicting crime type.
- Include demographic data and use age group to find better crime patterns
- Perform k-means clustering to find natural groupings of crime locations.
- Detailed analysis on why arrest rates are decreasing over the years.

BUSINESS OBJECTIVE ACCOMPLISHED

1. The non-violence crimes are decreasing over the years but the same is not the case for violence crimes
2. Serious Offenses crimes are more as compared to other crimes
3. Most crimes take place at Residential and Pathways
4. Most theft crimes take place at 6 pm in the evening and less at 5 in the morning
5. The number of arrest are very less over the years

REFERENCES

- ***https://trendct.org/2015/06/26/tutorial-how-to-put-dots-on-a-leaflet-map-with-r/***
- ***https://nycdatascience.com/blog/student-works/r-shiny/analysis-and-visualization-of-crime-in-chicago/***
- ***https://www.rdocumentation.org/packages/leaflet/versions/2.0.2/topics/leaflet***