CHICAGO CRIME DATA ANALYSIS

T CROSS CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CROSS
SCENE DO NOT CROSS CRIME SCENE

By,
Deepti Khatri
Preethi Kannan
Rakesh Jain
Sanket Bhaud

Data

- Our dataset : Chicago Crime Data from 2012 to 2016 (5 years) from Kaggle
- Referred Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
- It has around million records and 23 columns.
- Tools for cleaning and analysing : R and Tableau

Attributes

1. primary_type :- Primary Crime Type.
   - **crime_categories** :- Categorised crime into violence and non-violence
2. arrest :- Indicates whether an arrest was made.
3. beat :- Indicates the beat where the incident occurred.
4. location - district, ward, community_area, latitude and longitude
   - **loc_categories** :- Categorised locations where the incident occurred.
5. year :- Year the incident occurred.
6. date :- Date when the incident occurred
   - **Crimehour** :- Round off value extracted from date in 24 hour format.
   - **Timegroup** :- Categorised time hour into morning, afternoon, evening and night.

Goals

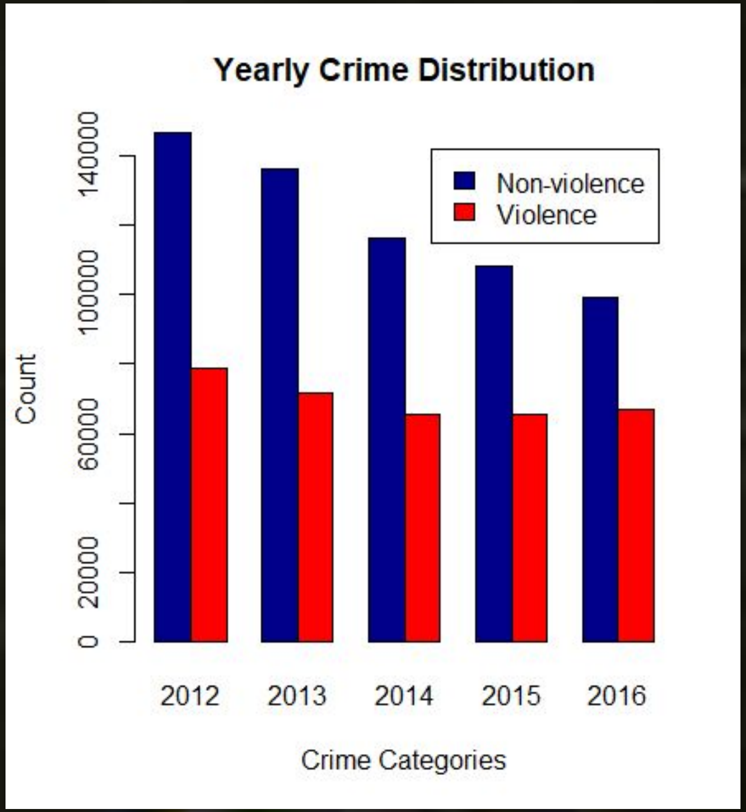
- Identify the trend in the crime rate in Chicago over the years.
- Identify the most occurring type of crime(drugs/narcotics violation or theft).
- Find the highly prone crime areas (apartments, office, roads etc)
- Find the time of the day when most of the crimes occur.
- Use heat maps to show the crime distribution in various regions.
- Create models for predicting crime type.

Data Cleaning

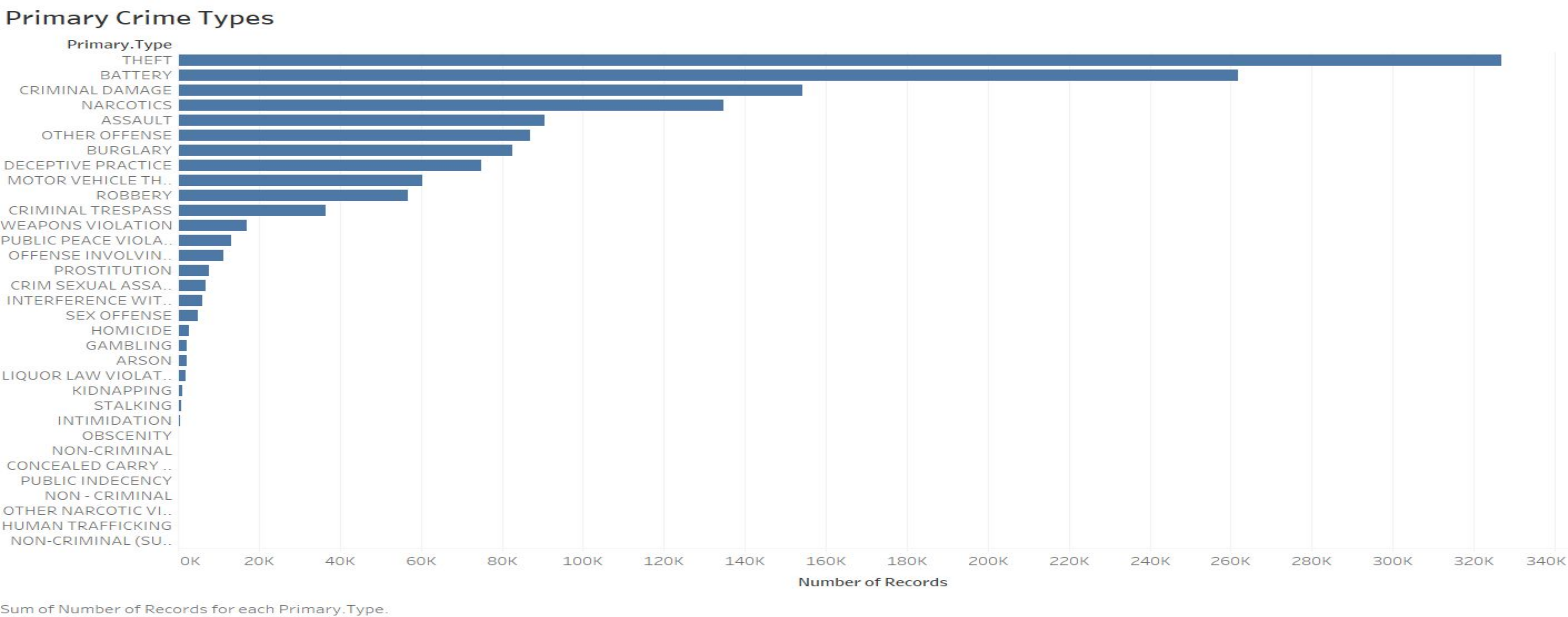
- Removed 2017 data since it was incomplete.
- Feature Extraction : Date column categorised into morning, afternoon, evening and night.
- Feature Grouping : Crimes types into crime categories.
- Removed missing values from latitude and longitude columns .
- Dropped unused levels from Primary Type.
-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

Preliminary Analysis

- Yearly trend in crime categories
- Non - Violent crimes decreasing over the years
- Violent crimes are almost the same

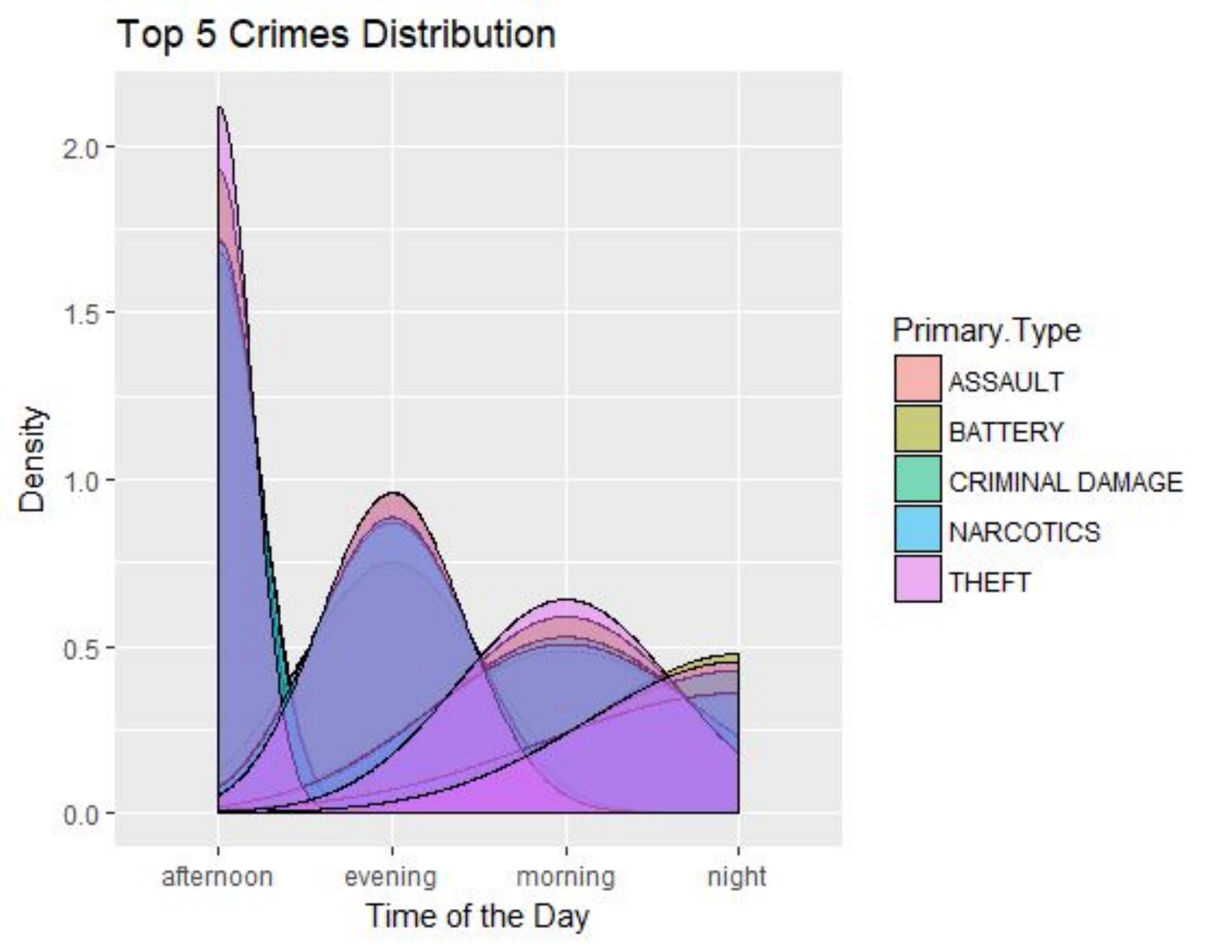


Preliminary Analysis



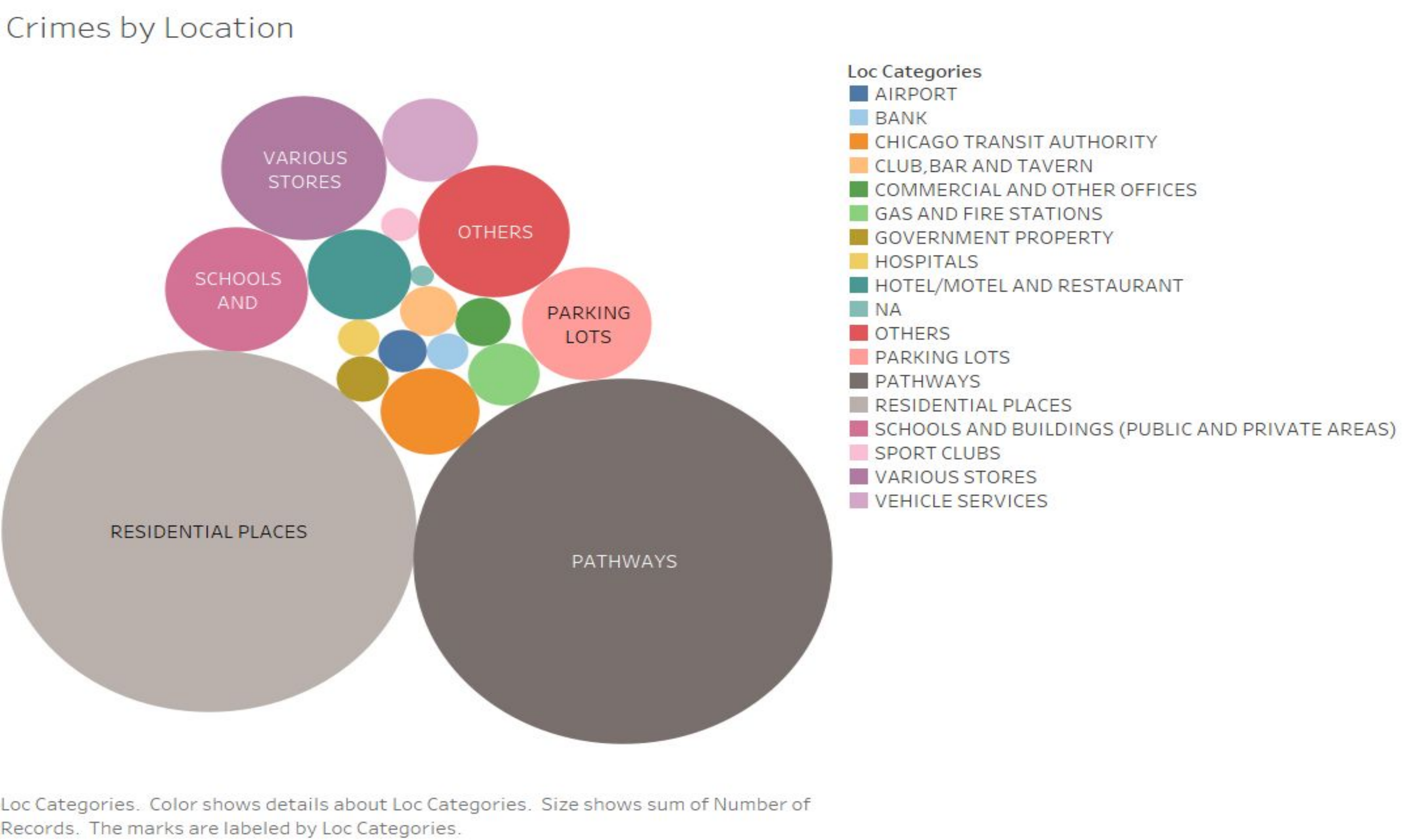
Preliminary Analysis

- Top five crimes take place mostly during afternoon with theft being the highest
- Lowest is night



Preliminary Analysis

- Residential places and pathways are the highly crime prone areas


Crimes by Location

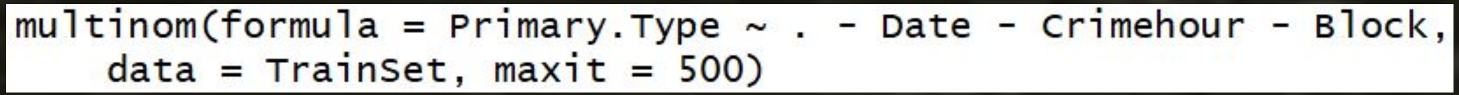
Loc Categories. Color shows details about Loc Categories. Size shows sum of Number of Records. The marks are labeled by Loc Categories.

Predictive Modeling

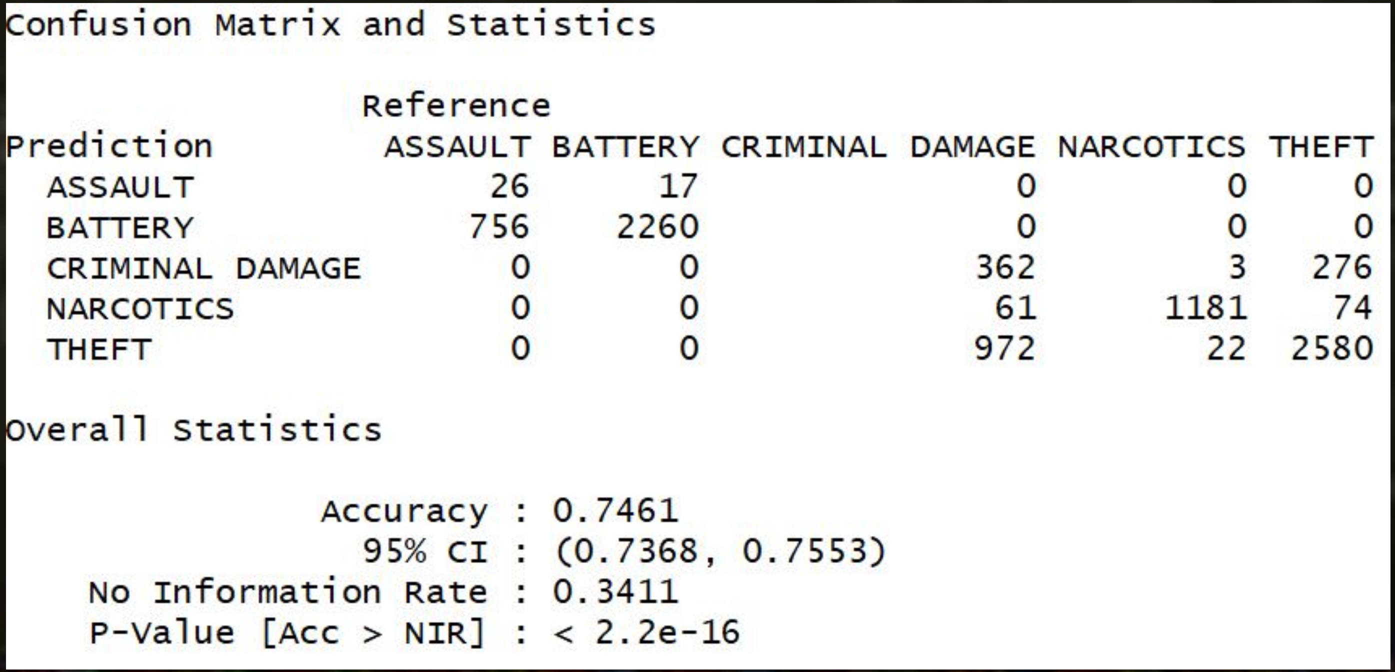
- Predict crime type using data filtered for top five crimes
- Features - Timegroup, Location, Arrest, Beat
- Training Set - 20000
- Test Set - 8600
- Classification methods

Predict Crime Type - Multinomial Logistic Regression

- **Accuracy on test data- 74.6%**
- No information rate - 34%



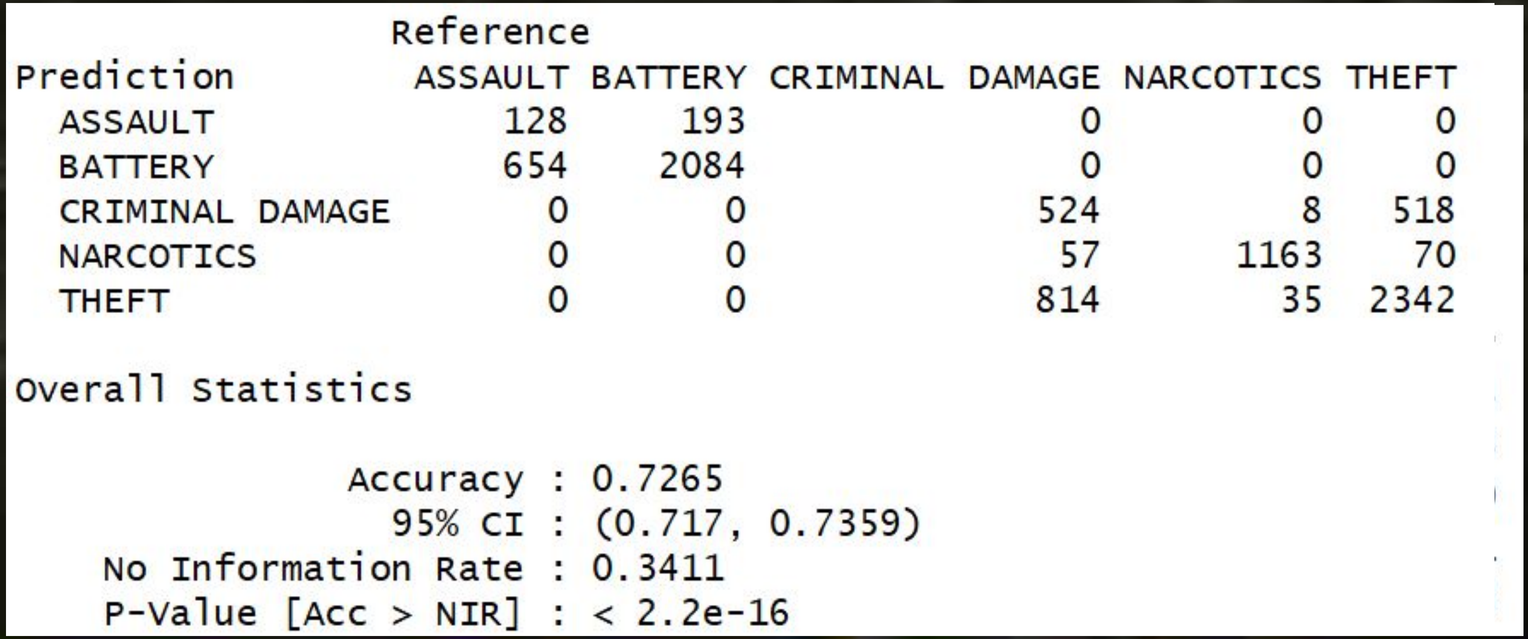
```
multinom(formula = Primary.Type ~ . - Date - Crimehour - Block,
    data = TrainSet, maxit = 500)
```



```
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
```

Predict Crime Type - Random Forest

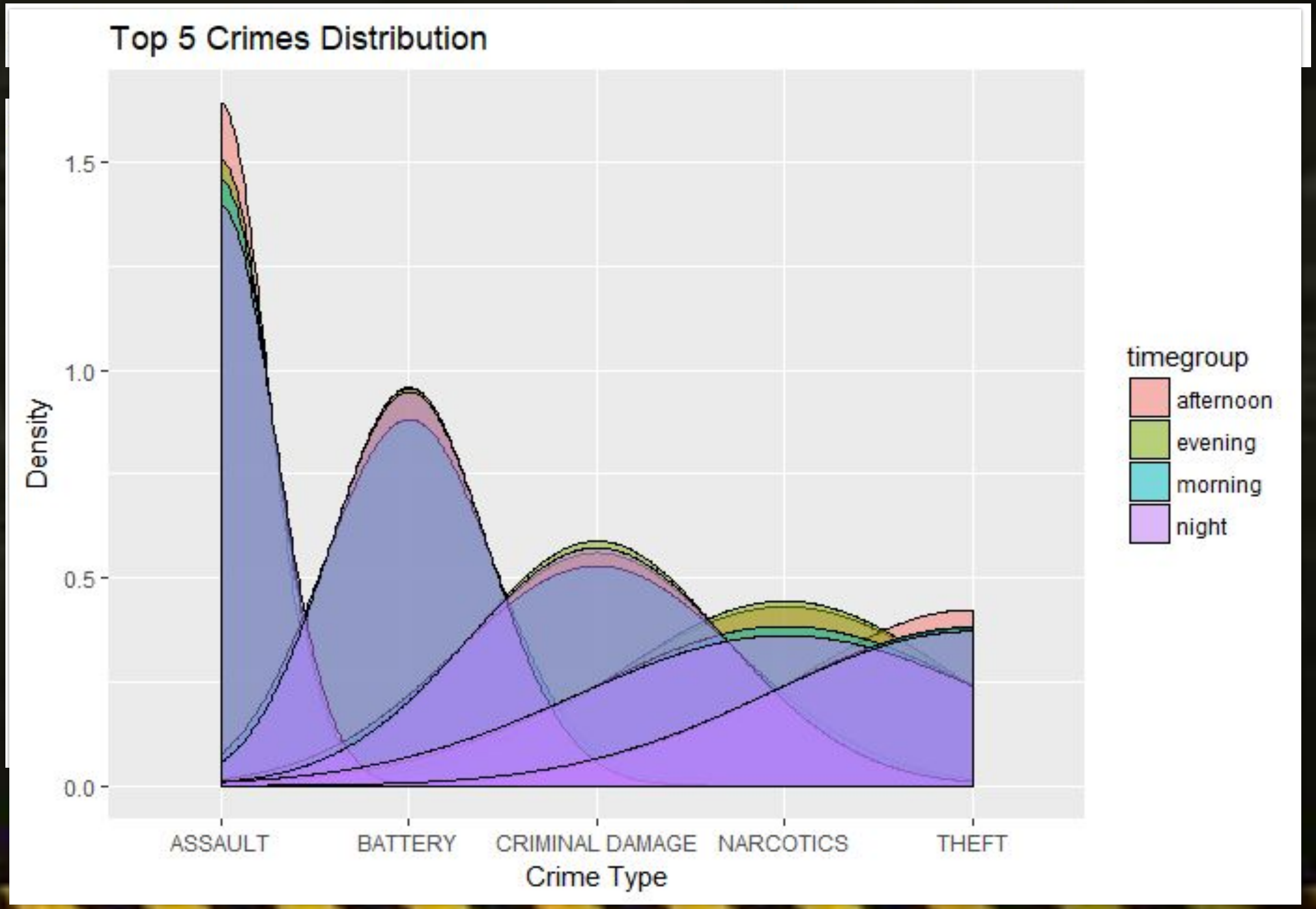
- Tuned mtry and ntree using CV
- mtry = 4 and ntree=600
- OOB error rate = 27%
- **Accuracy = 72.7%**



```
                  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
```

Predict Crime Type - SVM

- Classes were not clearly separable
- Tuned hyperparameters
- **Accuracy - 74.6%**



Challenges

- Multinomial regression difficult to train a dataset that contains 33 categories for Primary type.
- Data imbalance - there are more records for theft.
- All combinations of tuning parameters for SVM almost gave same accuracy.
- Spatial Visualization using Google maps is paid!

Conclusion

- **Multinomial Logistic regression** performed well in prediction
- Higher number of crimes between **spring and fall** every year.
- From our spatial plot, most of the crimes occur in **residential areas**.
- So, we recommend to deploy more police forces in those crime hotspots.

Future Scope

- Perform stratified sampling
- K means clustering to get more insights into data
- Spatial analysis using maptools

THANKS!!!