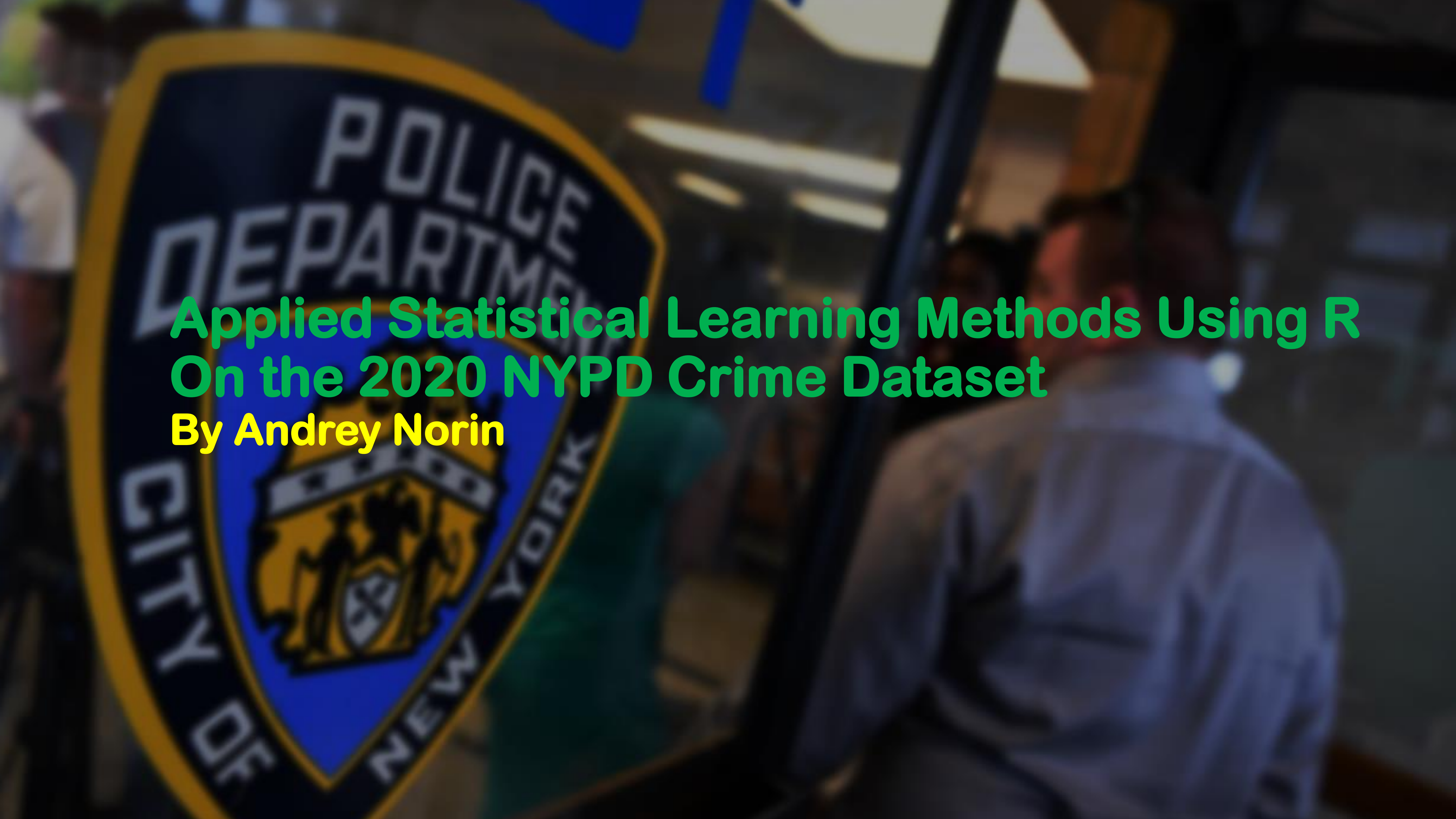


Applied Statistical Learning Methods Using R On the 2020 NYPD Crime Dataset

By Andrey Norin

Project Methodology

- Obtain and pre-process the dataset
- Understand the dataset visualization and summary statistics
- Explain outliers, collinearities and high leverage points (if any)
- Partition the dataset
- Narrow my query of interest
- Use ETL techniques to present data to algorithms in the correct format
- Apply Statistical Learning Algorithms
    - Decision Trees
    - Boosted Trees
    - Naïve Bayes
    - K-Means Clustering
- Understand and summarize findings and lessons learned

2020 NYPD Reported Crimes Dataset

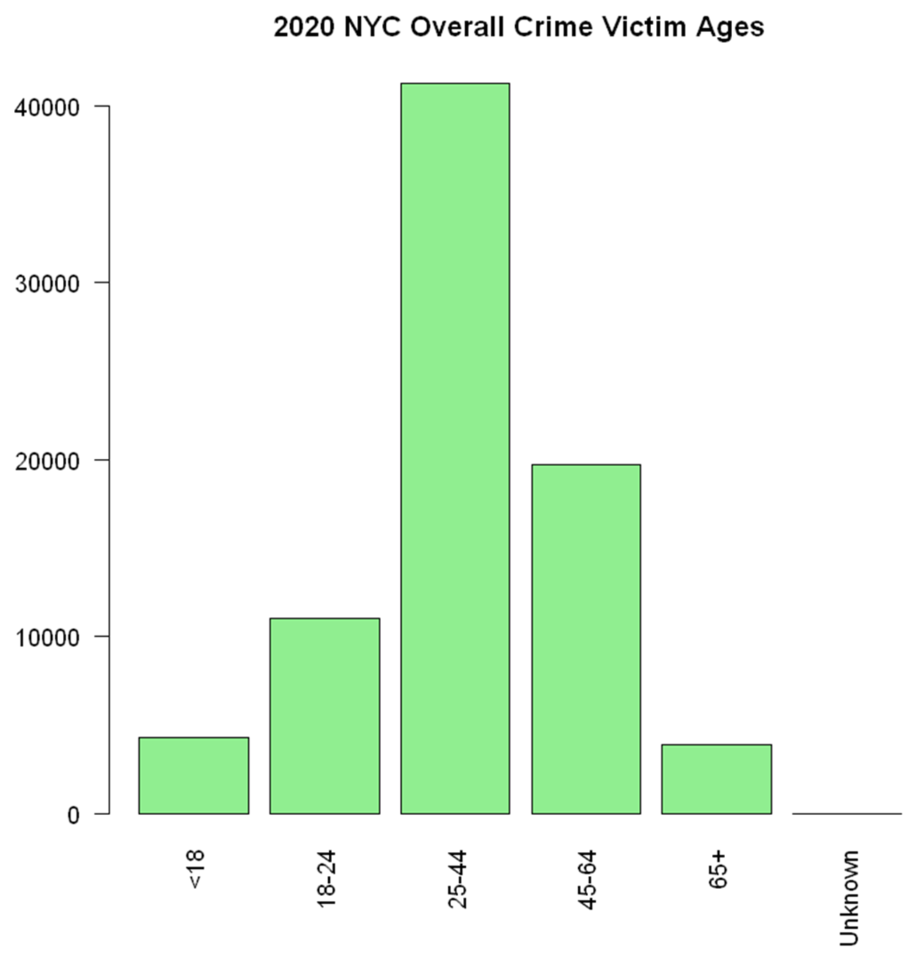
300,000 records with 36 features

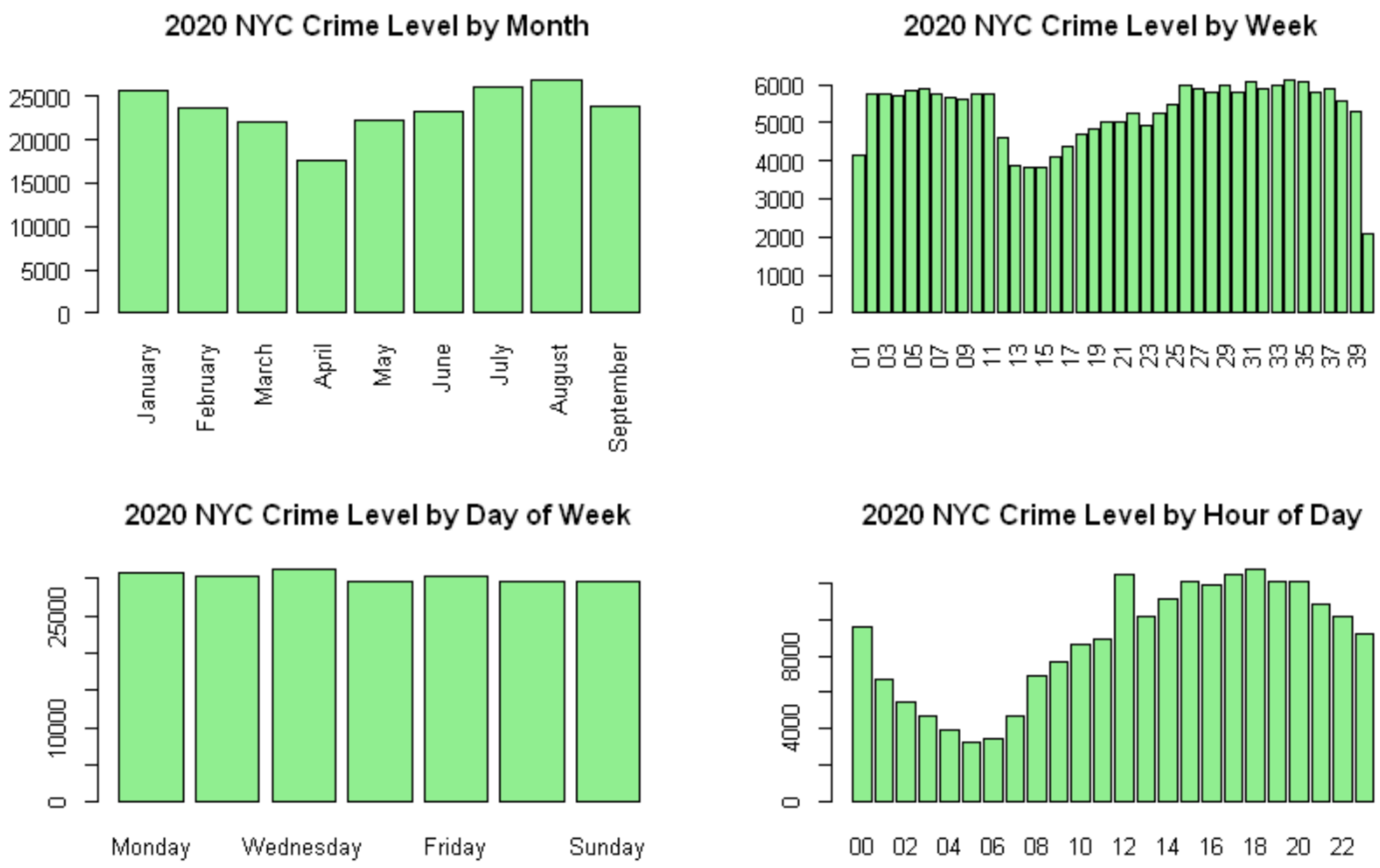
- Crime Information
  - Offense Description
  - Level of Offense
  - Classification Key Code
  - Time Reported
  - Attempted/Completed
- Location Description
  - Premises Type
  - Specific location within premises
  - Name of Park
  - Transit Station
  - Patrolling Precinct
  - Borough
  - Geographic coordinates
- Victim Information
  - Age Group
  - Race
  - Sex
- Suspect Information
  - Age Group
  - Race
  - Sex
- Complaint ID
  - Randomly generated persistent ID for each complaint

Exploratory Data Analysis

Getting to know the data through summary statistics and visualization

```
                           OffenseDesc    VicSex
HARRASSMENT 2                   :23752   F:50642
ASSAULT 3 & RELATED OFFENSES    :17824   M:29599
FELONY ASSAULT                  : 7635
OFF. AGNST PUB ORD SENSBLTY & : 5402
MISCELLANEOUS PENAL LAW         : 5100
CRIMINAL MISCHIEF & RELATED OF: 4486
(Other)                         :16042
                             VicRace      VicAgeGroup          Borough
AMERICAN INDIAN/ALASKAN NATIVE:  279   <18  : 4276   BRONX        :17718
ASIAN / PACIFIC ISLANDER      : 6871   18-24:11069   BROOKLYN     :24245
BLACK                         :31345   25-44:41270   MANHATTAN    :14925
BLACK HISPANIC                : 4587   45-64:19699   QUEENS       :19731
UNKNOWN                       : 2470   65+  : 3927   STATEN ISLAND: 3622
WHITE                         :13753
WHITE HISPANIC                :20936
                    PremisesType   SusAgeGroup
RESIDENCE - APT. HOUSE    :30076   <18  : 3512
STREET                    :15542   18-24:13556
RESIDENCE-HOUSE           :14198   25-44:45821
RESIDENCE - PUBLIC HOUSING:10188   45-64:15600
OTHER                     : 1668   65+  : 1752
GROCERY/BODEGA            : 1007
(Other)                   : 7562
                             SusRace      SusSex       Latitude
AMERICAN INDIAN/ALASKAN NATIVE:  272   F:19247   Min.   :40.50
ASIAN / PACIFIC ISLANDER      : 4798   M:60994   1st Qu.:40.67
BLACK                         :37639             Median :40.72
BLACK HISPANIC                : 5570             Mean   :40.73
UNKNOWN                       : 3362             3rd Qu.:40.81
WHITE                         :10009             Max.   :40.91
WHITE HISPANIC                :18591
  Longitude         Month               Week              DayOfWeek
Min.   :-74.25   Length:80241       Length:80241       Length:80241
1st Qu.:-73.96   Class :character   Class :character   Class :character
Median :-73.92   Mode  :character   Mode  :character   Mode  :character
Mean   :-73.92
3rd Qu.:-73.87
Max.   :-73.70

    Hour
Length:80241
Class :character
Mode  :character
```





2020 NYC Crime Level by Month
0
5000
10000
15000
20000
25000
January
February
March
April
May
June
July
August
September
2020 NYC Crime Level by Week
0
1000
2000
3000
4000
5000
6000
01
03
05
07
09
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
2020 NYC Crime Level by Day of Week
0
10000
25000
Monday
Wednesday
Friday
Sunday
2020 NYC Crime Level by Hour of Day
0
4000
8000
00
02
04
06
08
10
12
14
16
18
20
22

Partitioning Dataset

- Subdividing Dataset into
  - Crimes Against People of The State of NY
  - Crimes Against Business/Organizations
  - Crimes Against Persons
  - Test and Validation datasets using random sampling with replacement

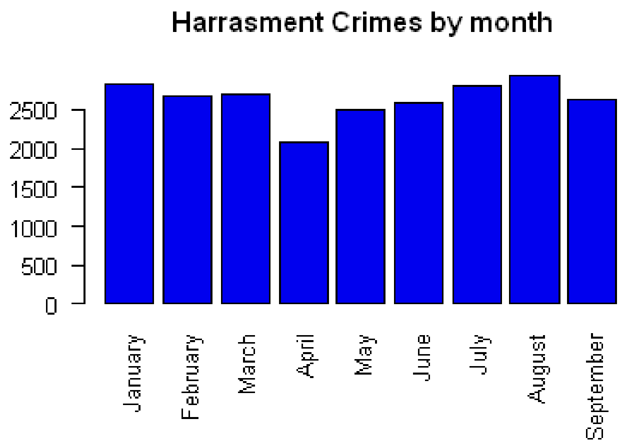
```r
CrimesAgainstBusiness     <- CD[CD$VIC_SEX == 'D',]
CrimesAgainstPeopleOfNYS <- CD[CD$VIC_SEX == 'E',]
CrimesAgainstPersons      <- CD[CD$VIC_SEX == 'M' | CD$VIC_SEX == 'F',]
```

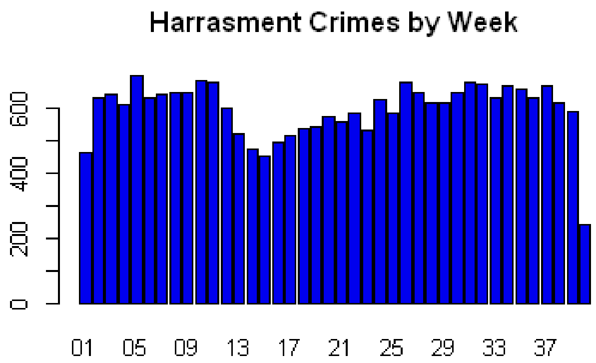
```r
## 75% of the sample size
smp_size <- floor(0.75 * nrow(NYPD))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(NYPD)), size = smp_size, replace = TRUE)

train <- NYPD[train_ind, ]
test  <- NYPD[-train_ind, ]

rownames(train) <- NULL
rownames(test) <- NULL

dim(train)
dim(test)
```

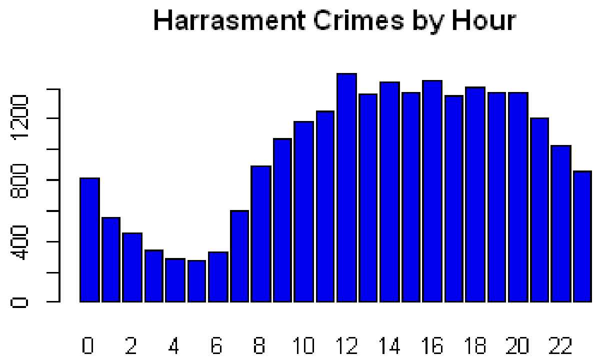
```
33099   16

20897   16
```

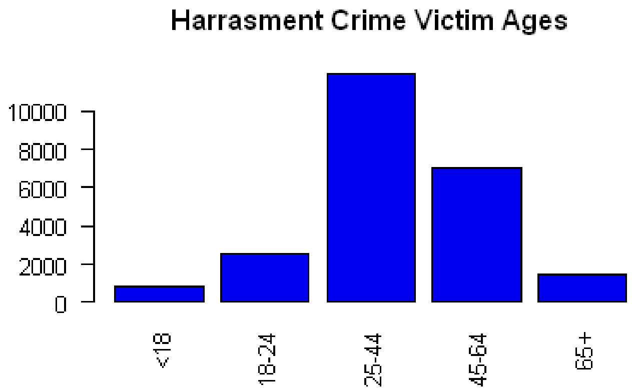
In Focus: Harassment In The 2nd Degree

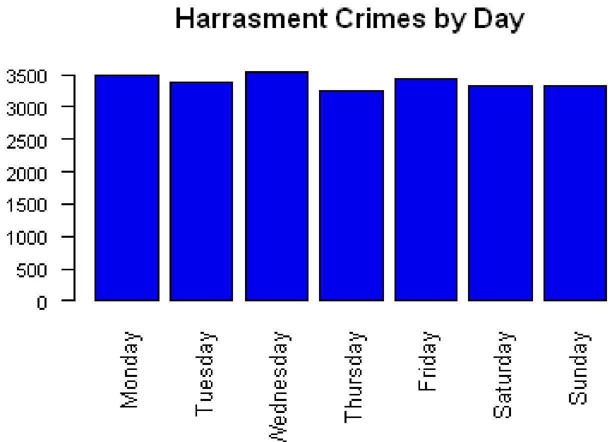
- Narrowing Down focus to Analyze Most Frequently Occurring Crime Type

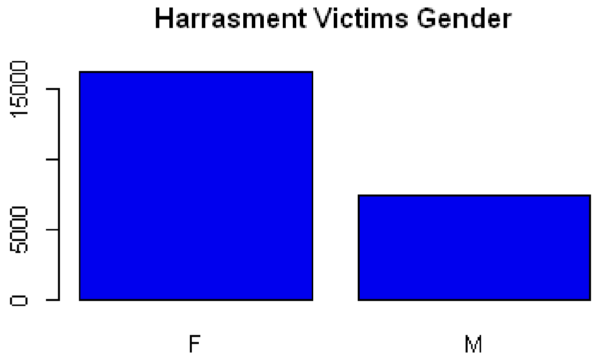

Harrasment Crimes by month
0
500
1000
1500
2000
2500
January
February
March
April
May
June
July
August
September

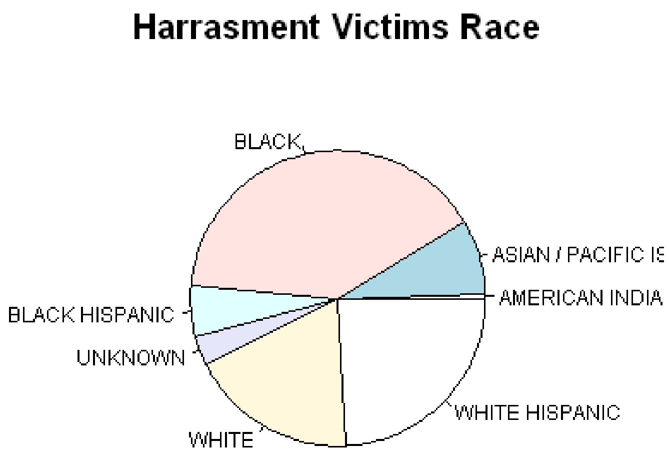

Harrasment Crimes by Week
0
200
400
600
01
05
09
13
17
21
25
29
33
37

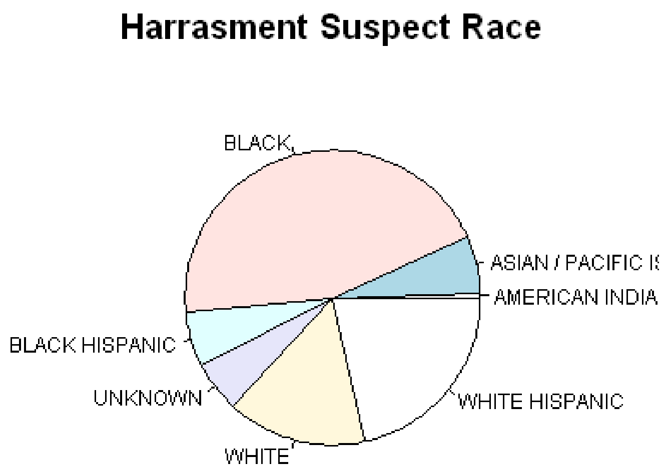

Harrasment Crimes by Hour
0
400
800
1200
0
2
4
6
8
10
12
14
16
18
20
22

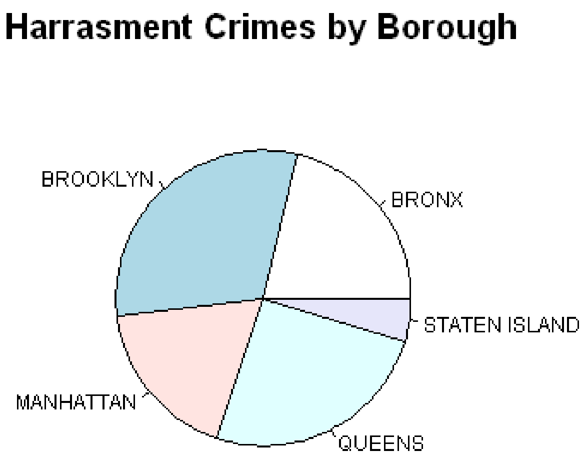

Harrasment Crime Victim Ages
0
2000
4000
6000
8000
10000
<18
18-24
25-44
45-64
65+

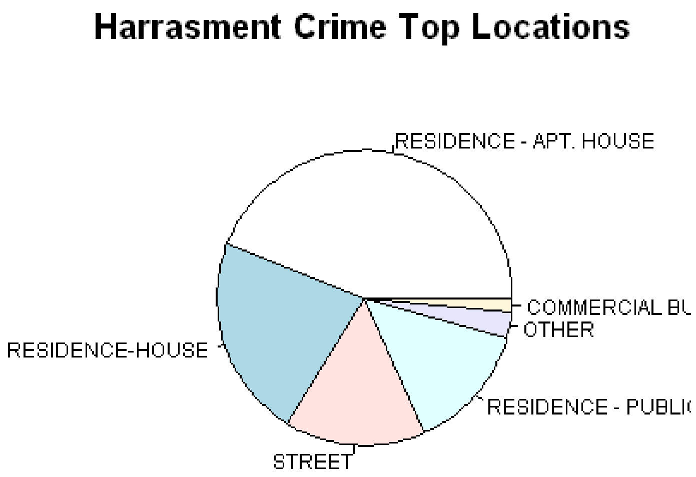

Harrasment Crimes by Day
0
500
1000
1500
2000
2500
3000
3500
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday


Harrasment Victims Gender
0
5000
15000
F
M


Harrasment Victims Race
BLACK
ASIAN / PACIFIC IS
AMERICAN INDIA
BLACK HISPANIC
UNKNOWN
WHITE
WHITE HISPANIC


Harrasment Suspect Race
BLACK
ASIAN / PACIFIC IS
AMERICAN INDIA
BLACK HISPANIC
UNKNOWN
WHITE
WHITE HISPANIC


Harrasment Crimes by Borough
BROOKLYN
BRONX
STATEN ISLAND
MANHATTAN
QUEENS

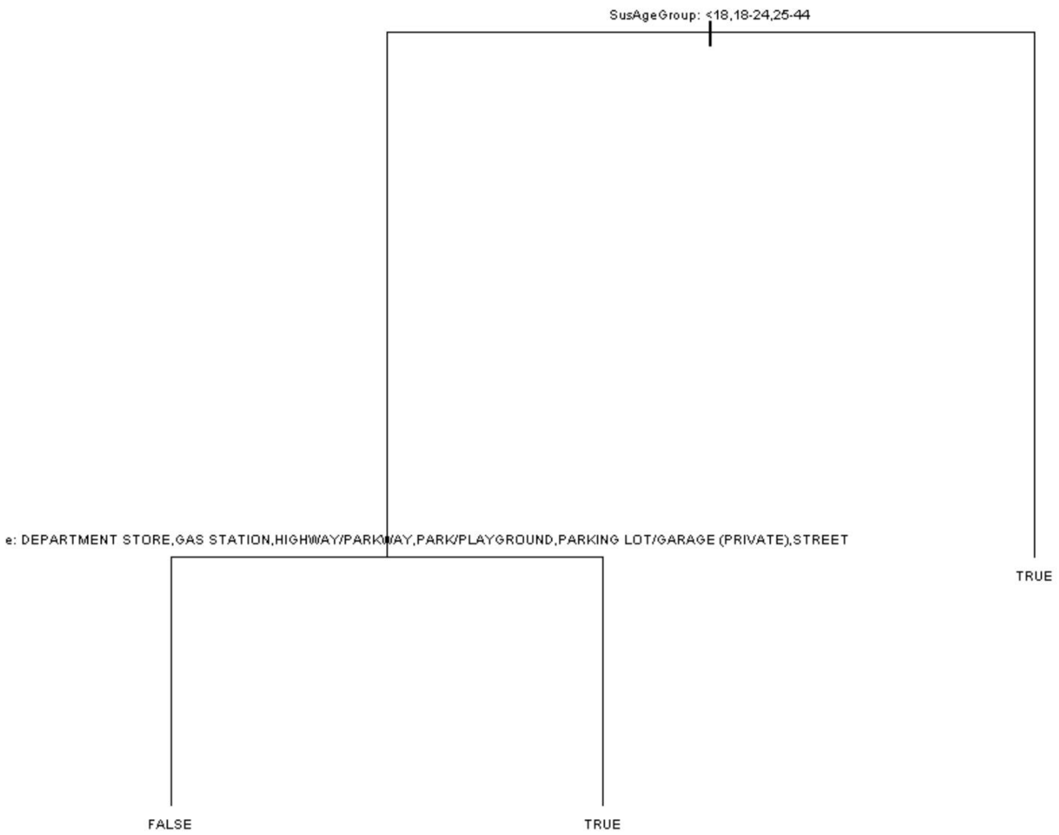

Harrasment Crime Top Locations
RESIDENCE - APT. HOUSE
COMMERCIAL BU
OTHER
RESIDENCE-HOUSE
RESIDENCE - PUBLIC
STREET

Classification Problem

- Create a statistical classifier for identifying if a crime is or is not of “Harassment 2” type
- Use Decision Tree classifier
- Improve performance of above Decision Tree classifier by using Boosting
- Learn and Deploy a Naïve Bayes classifier Against

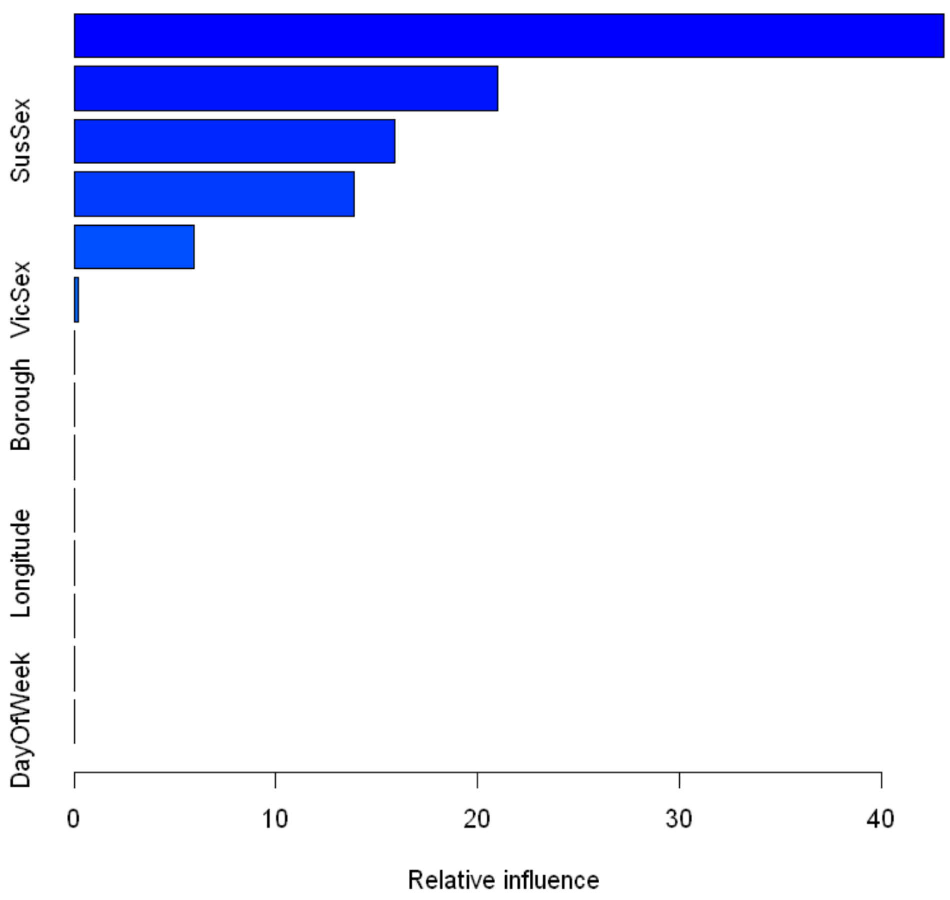
Decision Tree Classifier Performance

- Upon training a Decision Tree classifier and validating the learned model against the tree we arrived at a low prediction accuracy of 57% which is slightly better than random guessing
- Decision Tree algorithm couldn't arrive at an effective splitting criterion to classify new entries effectively
- Attempts at optimizing tree depth didn't yield tangible improvements in classification performance



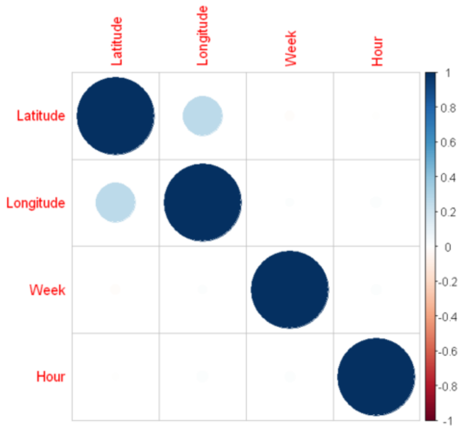
Improving Decision Tree classifier performance with Boosting

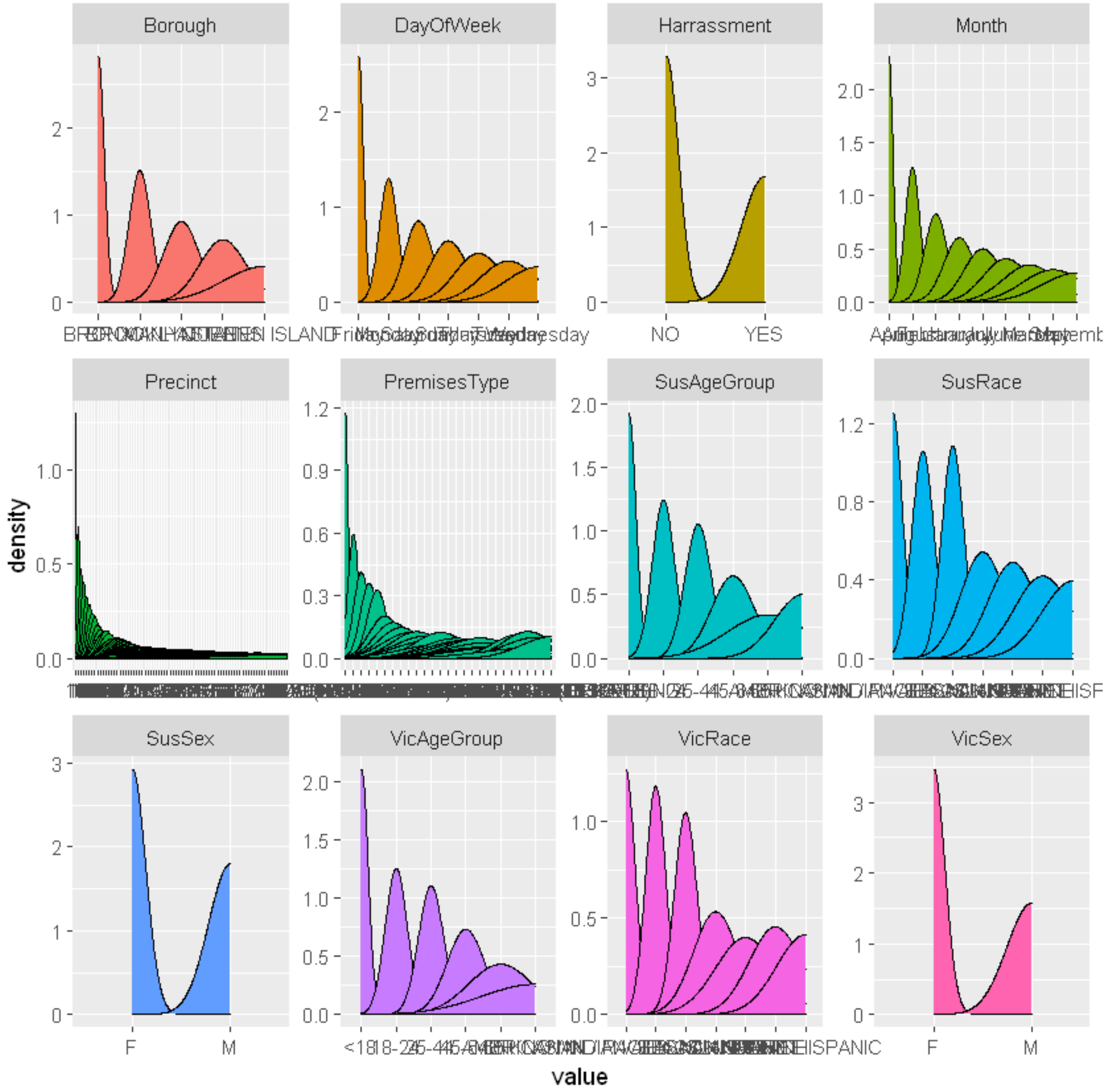
- Performing Boosting using 1,000 stump trees produced two factors with high predictive influence: **SusAgeGroup** and **Premises Type**
- Both of these features were already strongly expressed in the Decision Tree classifier, so no further prediction accuracy improvement was attained
- Weak performance of tree based classifiers against this dataset is explained by data being mostly Categorical



Naïve Bayes Classifier Performance

- Naïve Bayes classifiers as it is known to perform well against categorical data
- Density calculations of each value were performed to gain new information and to help understand Prior Probabilities
- Numeric features were checked for presence of strong correlations





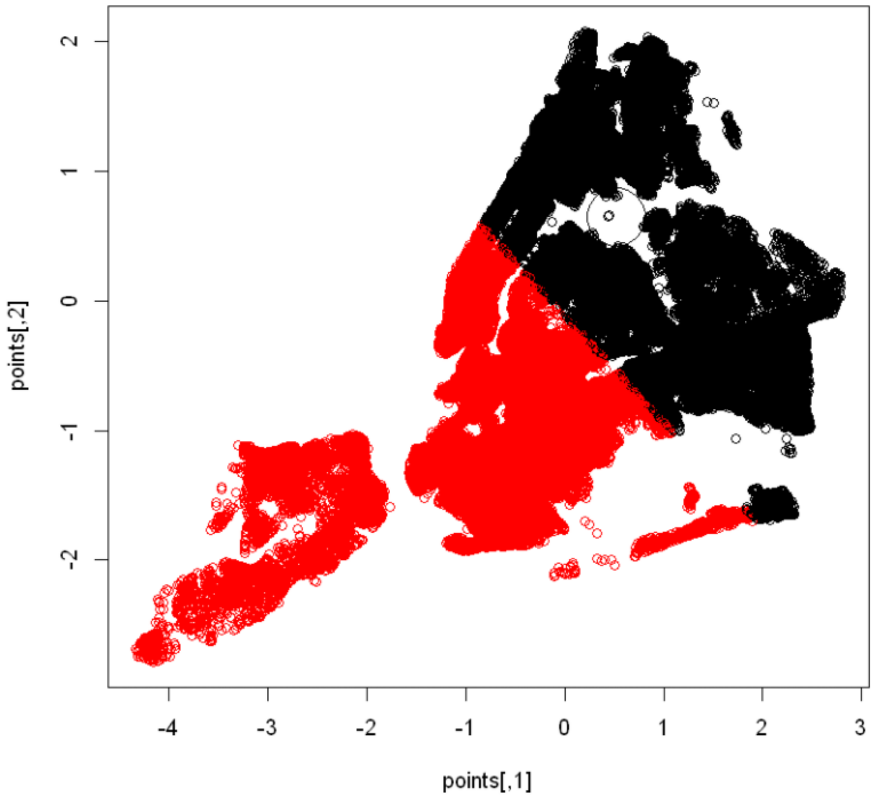
Naïve Bayes Results

- Naïve Bayes classifier improved overall prediction accuracy to 61% at 95% Confidence Level
- Classifier performs best at identifying negative tuples
- In conclusion; this dataset does not contain sufficient information to form a strong identification model
- In order to obtain further improvements in classification performance, dataset has to be enriched

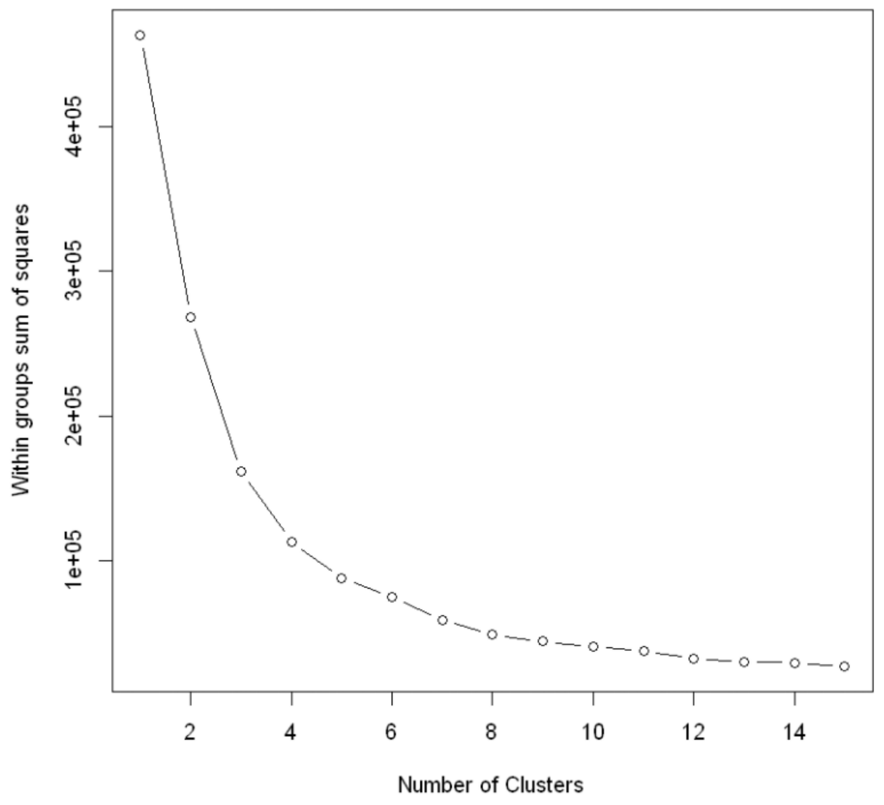
```
Confusion Matrix and Statistics

          Reference
Prediction    NO   YES
       NO   7593  5196
       YES  6277 10910

               Accuracy : 0.6173
                 95% CI : (0.6117, 0.6228)
    No Information Rate : 0.5373
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2261

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5474
            Specificity : 0.6774
         Pos Pred Value : 0.5937
         Neg Pred Value : 0.6348
             Prevalence : 0.4627
         Detection Rate : 0.2533
   Detection Prevalence : 0.4266
      Balanced Accuracy : 0.6124

       'Positive' Class : NO
```

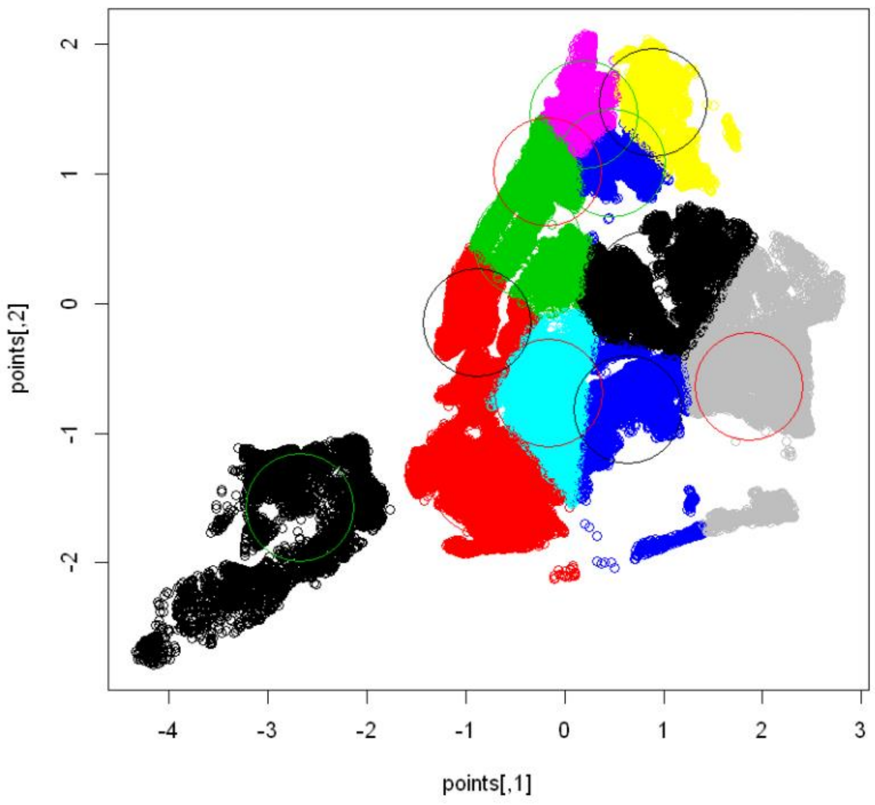
Geospatial Data Analysis

- In order to analyze the geographic data in the form of X and Y coordinates K-Means clustering algorithm was employed
- After scaling and preparing the data, initially data was partitioned into 2 clusters
- Next I performed analysis to find an optimal number of clusters
- Dividing geographical data into 8 to 10 clusters provided best results



Optimized K-Means Clustering Results





Challenges

- Real-world (non-academic) dataset
- Tightly packed fields
- Many different data types represented
  - Categorical
  - Date Time
  - Geospatial
- Missing fields and incomplete records
- Noise and Outliers
- Extensive ETL needed

Appendix

Existing works on the dataset

- https://jmc2392.github.io/exploratory2.html
- https://www.kaggle.com/adamschroeder/crimes-new-york-city/notebooks

References

- https://blog.rsquaredacademy.com/handling-date-and-time-in-r/
- https://www.datacamp.com/community/tutorials/decision-trees-R