

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light greenish-blue. They are positioned diagonally, with the blue one in front of the green one.

Introduction to Data Mining Predicting Crime Type in NYC

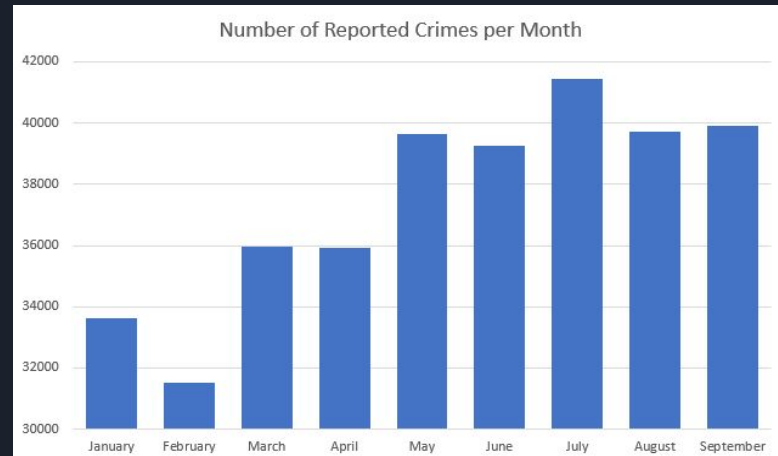
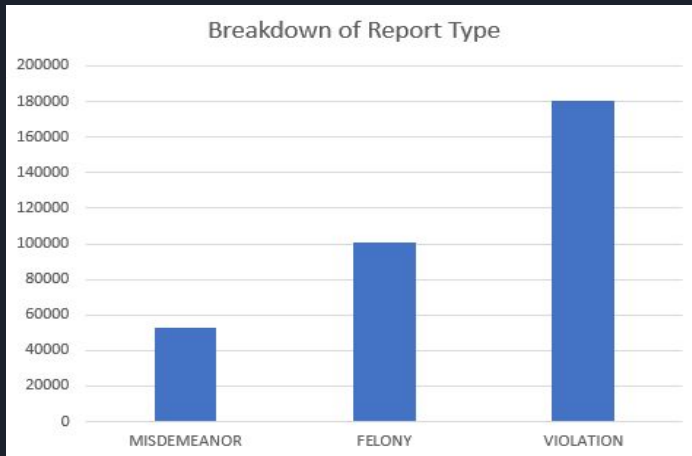
Omar Qusous, Sertan Akinci, Sam SR



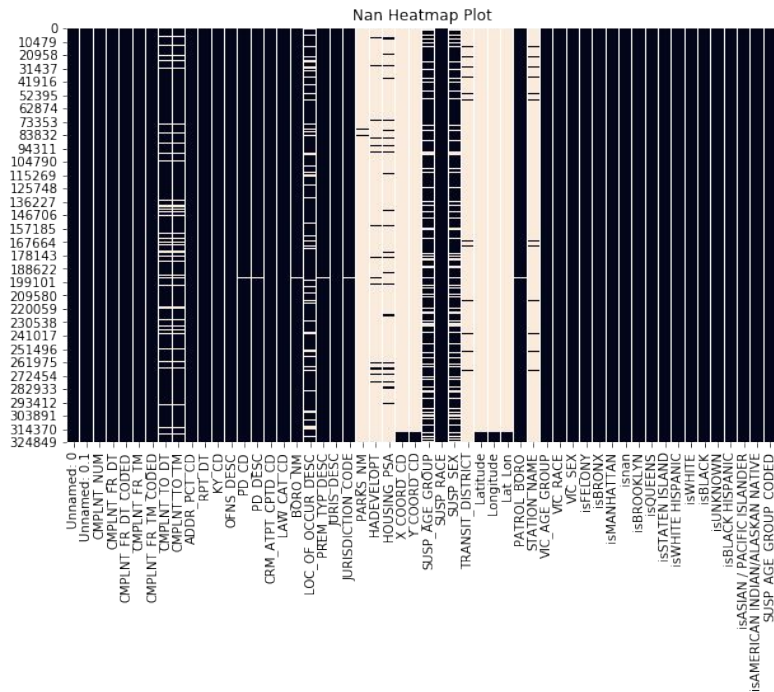
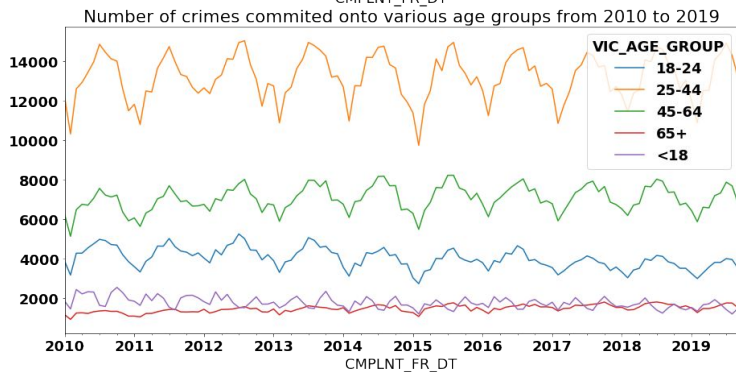
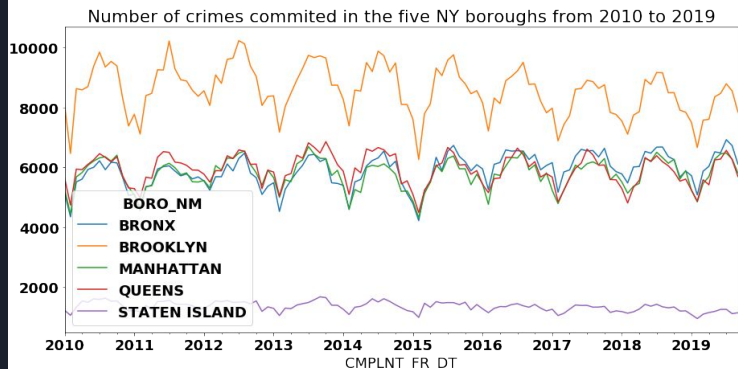
Our Dataset

- NYPD Complaint Data Historic
- Data provided by NYC Open Data and the New York Police Department
- Contains every felony, misdemeanor, and violation crime that the New York City Police Department received a report for from 2006 to 2019
- Contains information such as race / age / sex of victims and suspects, the time of day the crime was committed, and information about the location
- Over 6 million observations
- 35 columns - mostly categorical
- Updated Annually

Exploratory Data Analysis



Exploratory Data Analysis





Cleaning and Preprocessing

- Limited our data to 2019, giving us about 350,000 reports
- Encoded numerical variables (date/time) to a continuous number
 - Nulls -> average
- Encoded ordinal categorical variables (age) to continuous number
 - Nulls -> average
- One-hot encoded categorical (not ordinal) variables (borough, sex, etc) to separate features for each possible
 - Nulls -> randomly assigned matching the given distribution of possibilities



Available Features

- Date of reported event (start and end)
- Time of reported event (start and end)
- Precinct in which the event occurred
- Date of report
- Offense classification code / description
- Internal classification code / description
- Indicator of whether or not the crime was successfully completed
- Level of offense
- Borough in which the offense occurred
- Specific Location around the premises
- Description of the premises
- Jurisdiction code / description
- Victim Race / Sex / Age
- Suspect Race / Sex / Age
- Specific location as X and Y coordinate on the NY State Plane Coordinate System
- Latitude and Longitude
- Patrol borough



Decision Tree Models: Targets and Features

- For simplicity, separate models targeting each crime type
- Targets: felony, misdemeanor, violation
- Features:
 - Time of reported event
 - Date of reported event
 - Victim Age / Sex / Race
 - Borough in which the event took place
 - Description of the premises
 - Specific location on the premises
 - Jurisdiction code description
- Max leaf nodes = 50



Decision Tree Models: Results

- Tweaking the test_size and the max_leaf_nodes, our best results were:

- Violation:

- Accuracy: 84%
- F1: 11%
- Confusion Matrix:

28134	14
5243	7

- Felony:

- Accuracy: 70%
- F1: 3%
- Confusion Matrix:

22903	561
9343	591

- Misdemeanor:

- Accuracy: 58%
- F1: 61%
- Confusion Matrix:

8497	6832
7034	11035



Other Modeling Attempts

- Random Forest
- Multinomial Logistic Regression
- Support Vector Machine
- Random Forest
- These methods did not see as much success as decision tree



App Demo

- Developed an app with a user interface which can let you play around with our decision tree model
- Choose parameters and try out the classification