# Crime Patrol: Working With Crime Data

## Group Name: **The Phoenix Miners**


**Amit Rokade** **ar5618@rit.edu**

**Tejas Arya** **ta2763@rit.edu**

**Anmol Jaising** **akj9051@rit.edu**

# Introduction

- With the increase in rate in crime across the country, police agencies look to respond immediately to respond to the complaints. It happens frequently that by the time a complaint has been filed, another crime has already been committed.

- Crime Patrol is a data cleaning and preparation pipeline for crime data collected for NYC and LA and provide insights into that help police force to take preventive measures.

- Data visualization is performed on raw and clean data providing better insights into the crime trend of various cities. Such information can be further used to mitigate crime rates and ensure well being of civilians and properties.

# Datasets

The selected datasets are:

1. New York police department Complaint data.
    - Contains complaint data of crime reported between 2006 and 2017
    - Includes attributes such as complaint type, place of occurrence, data of occurrence, etc.
    - The dataset is in comma separated value (CSV) format
    - Contains 7M rows and 35 columns.

2. Los Angeles police department crime data:
    - Contains crime records occurred between 2010 and 2019 in Los Angeles.
    - The dataset is in CSV format.
    - Contains 2M crimes incidents with 28 columns.

# Motivation

- With an increase in population in New York and Los Angeles, comes an increase in crime and complaint rate as well.
- The motivation for this project comes in a manner to reduce the crimes occurring.
- As well as understanding and examining where crime and complaint records, in an effort can be made to minimize it.

# Data Collection

- Data has been collected from two different sources in the (CSV) format.

- Both data sets had similar features, although their names were inconsistent. For example:

  1. **LAPD**: ‘Date_Rptd’, ‘TIME_OCC’, ‘AREA’, ‘AREA_NAME’, ‘Crm_Cd’, ‘Crm_Cd_Desc’
  2. **NYPD**: ‘RPT_DT’, ‘CMPLNT_FR_TM’, ‘BORO_NM’, ‘KY_CD’, ‘OFNS_DESC’

  The columns listed above refer to the same things with different column names
  ( This was found out by studying the data dictionary)

  Even though the format of values in these attributes differed, we tried to make it same in the cleaning and preparation stage.

- Attributes that are irrelevant to both datasets are dropped to make merging the columns possible.

# Data Cleaning and Preparation

- The various data cleaning and preparation methods used in Python were
  - Replacing  null values
  - Handling various outliers
  - Downcasting data types.

- Attribute formats were converted so that they are homogenous, For eg. Date attribute was converted to  MM/DD/YYYY & Time attribute was converted to hh:mm:ss which helped to merge the dataset.

- There were multiple attributes having invalid values. For eg. The Victim's age being "0" in some records, made no sense. Outlier detection methods such as Tukey IQR is used to remove such values, and replace UNKNOWN records and values like "0" with their mode.

# Data Cleaning and Preparation

- Each attributes missing values percentage is calculated. Those above 50% were discarded.
- CRIME_CODE and PATROL_DIVISION were converted from float to string, to make sure that these values were not read as decimals
- Label coding was applied to the categorical data to help with visualization.
- After the cleaning, boxplots showed no outliers.
- The samples were normalized to have equal classes number of classes in both the datasets using upsampling.

# Data Storage

- Renaming attributes to a common name, the merged dataset contained:
  - DATE_REPORTED
  - DATE_OCCURRED
  - TIME_OCCURRED
  - PATROL_DIVISION
  - AREA_NAME
  - CRIME_CODE
  - CRIME_DESCRIPTION
  - VICTIM_AGE
  - VICTIM_SEX
  - VICTIM_RACE
  - PREMISE
  - TARGET
- The dataset is further uploaded to a mySQL database.

# Statistics

## RAW DATA

### Statistics

| Variable | Total Count | N | N* | CumN | Percent | CumPct | Mean | SE Mean | TrMean |
|---|---|---|---|---|---|---|---|---|---|
| DATE_REPORTED | 100000 | 100000 | 0 | 100000 | 100.000 | 100.000 | 41643 | 4.29 | 41676 |
| DATE_OCCURRED | 100000 | 99994 | 6 | 99994 | 99.994 | 99.994 | 41621 | 5.63 | 41658 |
| TIME_OCCURRED | 100000 | 100000 | 0 | 100000 | 100.000 | 100.000 | 677.87 | 2.58 | 627.65 |
| PATROL_DIVISION | 100000 | 99981 | 19 | 99981 | 99.981 | 99.981 | 37.129 | 0.114 | 34.662 |
| CRIME_CODE | 100000 | 100000 | 0 | 100000 | 100.000 | 100.000 | 401.87 | 0.668 | 389.99 |
| LAT | 100000 | 99802 | 198 | 99802 | 99.802 | 99.802 | 37.394 | 0.0107 | 37.403 |
| LON | 100000 | 99802 | 198 | 99802 | 99.802 | 99.802 | -96.165 | 0.0704 | -96.178 |

| Variable | StDev | Variance | CoefVar | Sum | Sum of Squares | Minimum | Q1 |
|---|---|---|---|---|---|---|---|
| DATE_REPORTED | 1355 | 1836844 | 3.25 | 4164268598 | 1.73595E+14 | 38718 | 40613 |
| DATE_OCCURRED | 1779 | 3166601 | 4.28 | 4161854966 | 1.73537E+14 | -321799 | 40594 |
| TIME_OCCURRED | 817.35 | 668061.84 | 120.58 | 67786595.84 | 1.12756E+11 | 0.00 | 0.61 |
| PATROL_DIVISION | 35.910 | 1289.535 | 96.72 | 3712233.000 | 2.66761E+08 | -99.000 | 10.000 |
| CRIME_CODE | 211.22 | 44615.30 | 52.56 | 40186916.00 | 2.06114E+10 | 101.00 | 310.00 |
| LAT | 3.368 | 11.343 | 9.01 | 3731949.733 | 1.40683E+08 | 0.0000 | 34.062 |
| LON | 22.252 | 495.159 | -23.14 | -9.59747E+06 | 9.72360E+08 | -118.666 | -118.330 |

| Variable | Median | Q3 | Maximum | Range | IQR | Mode | N for Mode | Skewness |
|---|---|---|---|---|---|---|---|---|
| DATE_REPORTED | 41727 | 42796 | 44028 | 5310 | 2183 | 42948 | 44 | -0.26 |
| DATE_OCCURRED | 41706 | 42777 | 43830 | 365629 | 2183 | 40544 | 83 | -85.31 |
| TIME_OCCURRED | 1.00 | 1430.00 | 2359.00 | 2359.00 | 1429.39 | 1200 | 2636 | 0.66 |
| PATROL_DIVISION | 19.000 | 63.000 | 123.000 | 222.000 | 53.000 | 14 | 3877 | 1.00 |
| CRIME_CODE | 350.00 | 578.00 | 956.00 | 855.00 | 268.00 | 341 | 10048 | 0.65 |
| LAT | 34.321 | 40.732 | 40.913 | 40.913 | 6.671 | 34.1016 | 358 | -0.24 |
| LON | -118.167 | -73.928 | 0.0000 | 118.666 | 44.402 | -118.274 | 462 | 0.02 |

| Variable | Kurtosis | MSSD |
|---|---|---|
| DATE_REPORTED | -0.91 | 1822870 |
| DATE_OCCURRED | 17396.04 | 3149568 |
| TIME_OCCURRED | -1.17 | 667207.85 |
| PATROL_DIVISION | -0.34 | 1282.032 |
| CRIME_CODE | -0.06 | 44364.31 |
| LAT | 0.70 | 11.311 |
| LON | -1.94 | 493.264 |

## CLEAN DATA

### Statistics

| Variable | Total Count | N | N* | CumN | Percent | CumPct | Mean | SE Mean | TrMean |
|---|---|---|---|---|---|---|---|---|---|
| DATE_REPORTED | 100000 | 100000 | 0 | 100000 | 100 | 100 | 41761 | 4.25 | 41805 |
| DATE_OCCURRED | 100000 | 100000 | 0 | 100000 | 100 | 100 | 41742 | 4.32 | 41788 |
| TIME_OCCURRED | 100000 | 100000 | 0 | 100000 | 100 | 100 | 0.56224 | 0.000862 | 0.57016 |
| PATROL_DIVISION | 100000 | 100000 | 0 | 100000 | 100 | 100 | 38.124 | 0.116 | 35.756 |
| CRIME_CODE | 100000 | 100000 | 0 | 100000 | 100 | 100 | 407.19 | 0.687 | 395.46 |

| Variable | StDev | Variance | CoefVar | Sum | Sum of Squares | Minimum | Q1 |
|---|---|---|---|---|---|---|---|
| DATE_REPORTED | 1343 | 1804399 | 3.22 | 4176108036 | 1.74579E+14 | 38718 | 40741 |
| DATE_OCCURRED | 1367 | 1867637 | 3.27 | 4174222211 | 1.74428E+14 | 219 | 40724 |
| TIME_OCCURRED | 0.27256 | 0.07429 | 48.48 | 56224.48681 | 39040.99535 | 0.000000 | 0.37500 |
| PATROL_DIVISION | 36.604 | 1339.862 | 96.01 | 3812394.000 | 2.79328E+08 | -99.000 | 10.000 |
| CRIME_CODE | 217.38 | 47254.54 | 53.39 | 40719408.00 | 2.13061E+10 | 102.00 | 310.00 |

| Variable | Median | Q3 | Maximum | Range | IQR | Mode | N for Mode | Skewness |
|---|---|---|---|---|---|---|---|---|
| DATE_REPORTED | 41901 | 42910 | 44027 | 5309 | 2169 | 43126 | 47 | -0.37 |
| DATE_OCCURRED | 41882 | 42894 | 43830 | 43611 | 2170 | 40179 | 91 | -1.11 |
| TIME_OCCURRED | 0.59583 | 0.79167 | 0.99931 | 0.99931 | 0.41667 | 0.5 | 4222 | -0.45 |
| PATROL_DIVISION | 19.000 | 67.000 | 123.000 | 222.000 | 57.000 | 14 | 3522 | 0.94 |
| CRIME_CODE | 348.00 | 578.00 | 956.00 | 854.00 | 268.00 | 341 | 9852 | 0.62 |

| Variable | Kurtosis | MSSD |
|---|---|---|
| DATE_REPORTED | -0.84 | 1803502 |
| DATE_OCCURRED | 19.17 | 1867806 |
| TIME_OCCURRED | -0.75 | 0.07413 |
| PATROL_DIVISION | -0.48 | 1333.931 |
| CRIME_CODE | -0.21 | 47396.56 |

# Data Mining

***What is the most frequently targeted age group amongst all the crimes...?***

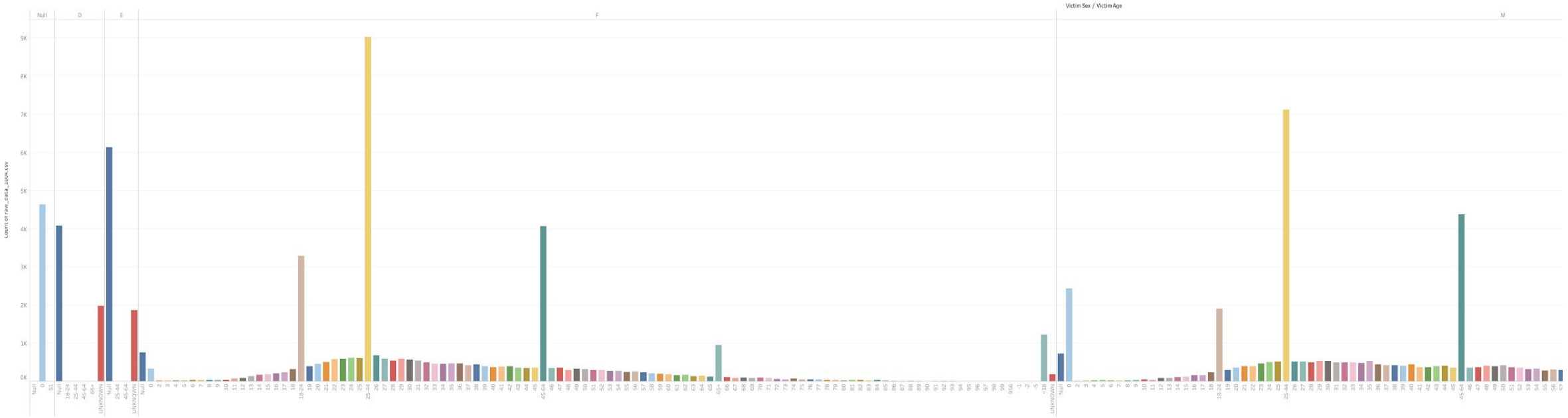
- Weka is used to perform data mining which answers this question.
- Algorithms such as J48, BayesNet, and NaiveBayes were used as classifiers.
- Clean dataset was split in a ratio of 70:30 in terms of training and testing.

Results by the classifiers on the clean dataset, showing J48 having the highest accuracy.

| Data | Algorithm | Target | Accuracy | Precision | Recall | F-Measure | ROC |
|---|---|---|---|---|---|---|---|
| 100K_Cleaned | J48 | AGE | 75.8118 | 0.758 | 0.997 | 0.861 | 0.865 |
| 100K_Cleaned | BayesNet | AGE | 62.4088 | 0.814 | 0.663 | 0.73 | 0.861 |
| 100K_Cleaned | NaiveBayes | AGE | 60.7441 | 0.819 | 0.662 | 0.732 | 0.851 |

# Data Visualization

**Victim age against Victim Sex for Raw data**




- Each color is an age group filter
- Since the format of representing age is different in LA (absolute) and NY (interval + absolute) , the charts automatically get laid side by side showing similar highs and lows.

# Data Visualization

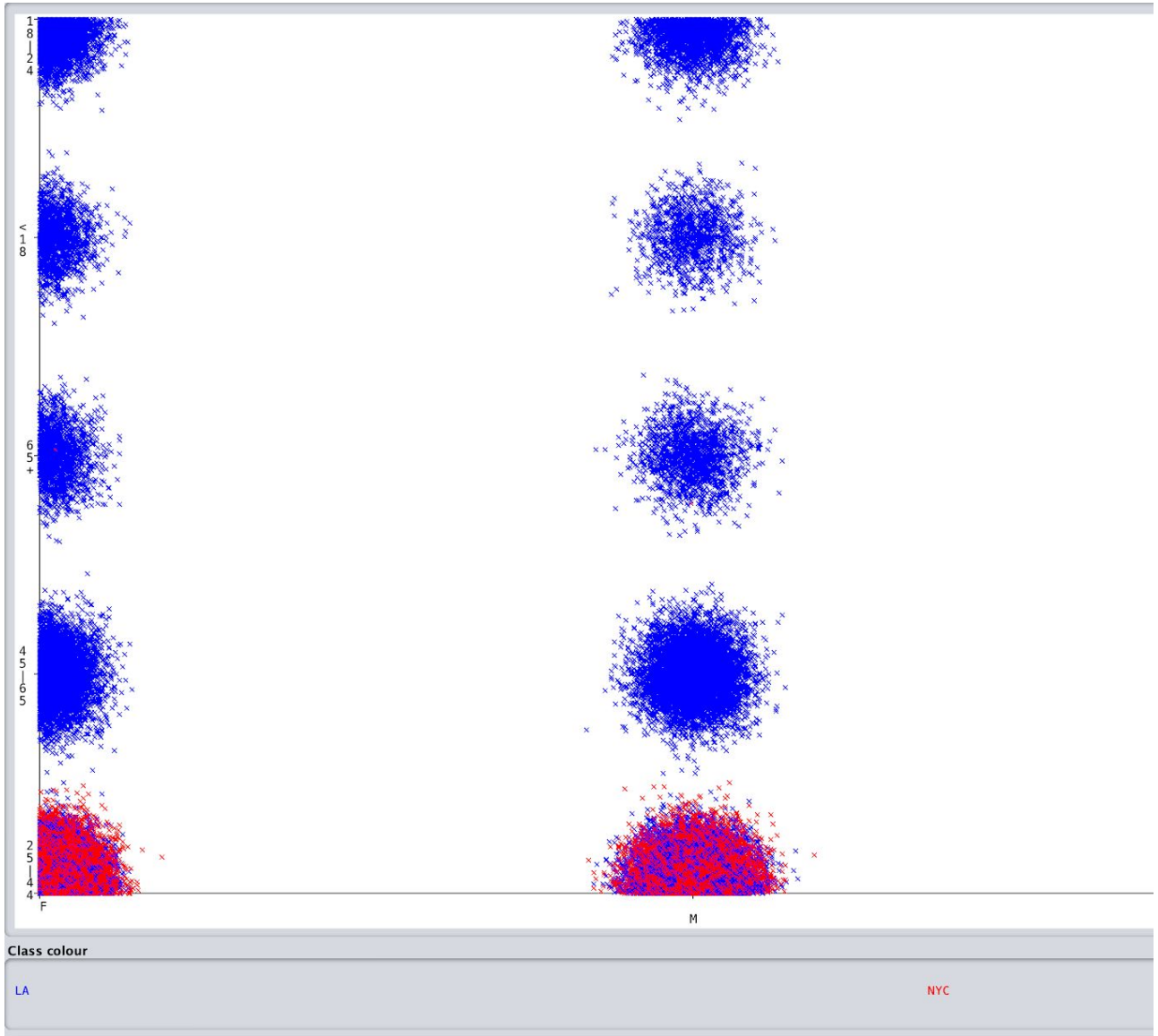
-Clusters are laid out vertically with Y axis showing the age ranges [ 18-24 , <18 , 65+ , 45-65, 25-44 ]

-X axis separates the sex Male (M) and Female (F)

-A third dimension is also used for visualization which separates **LA** and **NYC** data with color

***What can we infer from this data?***
***-> Not much data was recorded for NYC in the age ranges below 25 and above 44***

***-> Crimes are pretty consistent across all age groups***

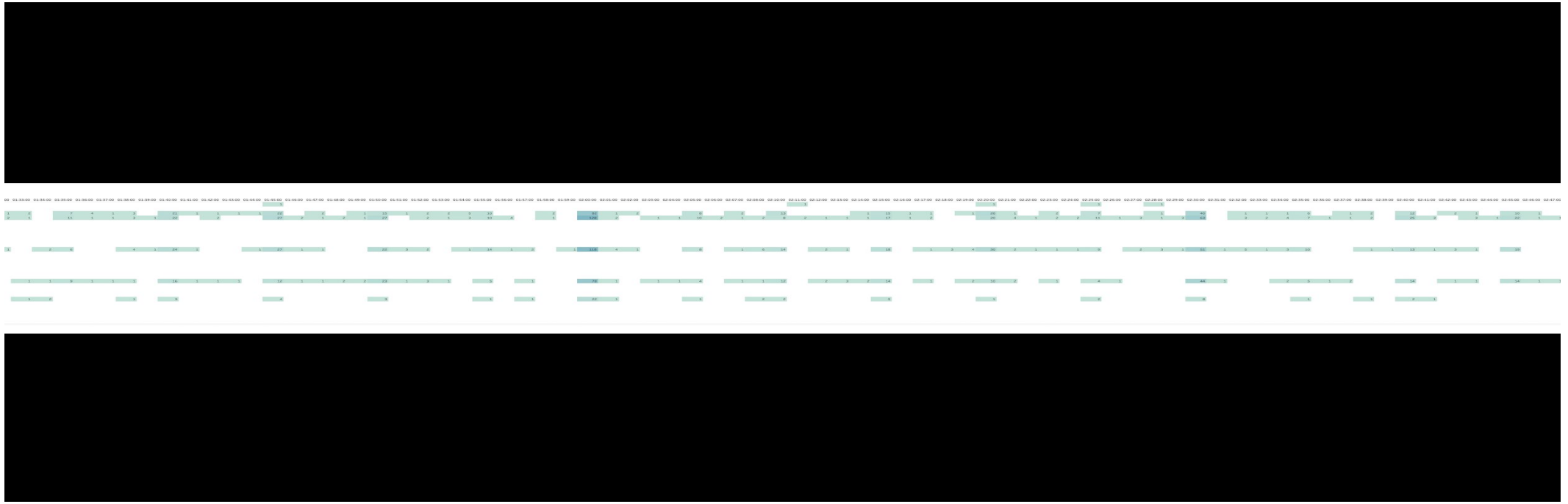


# Data Visualization

**Time_Occured against Area_Name**

-Inconsistency is seen when Time_Occured is mapped against Area_Name name for the raw data which makes it very difficult to infer any information from this data.

-This is due to the inconsistency in time recording formats



# Data Visualization

The visualization is understandable now , color shows the frequency and it can be seen that the most number of complaints were recieved for Brooklyn Area with ~900 crimes occuring around noon

<Time Occurred vs. Area Name>

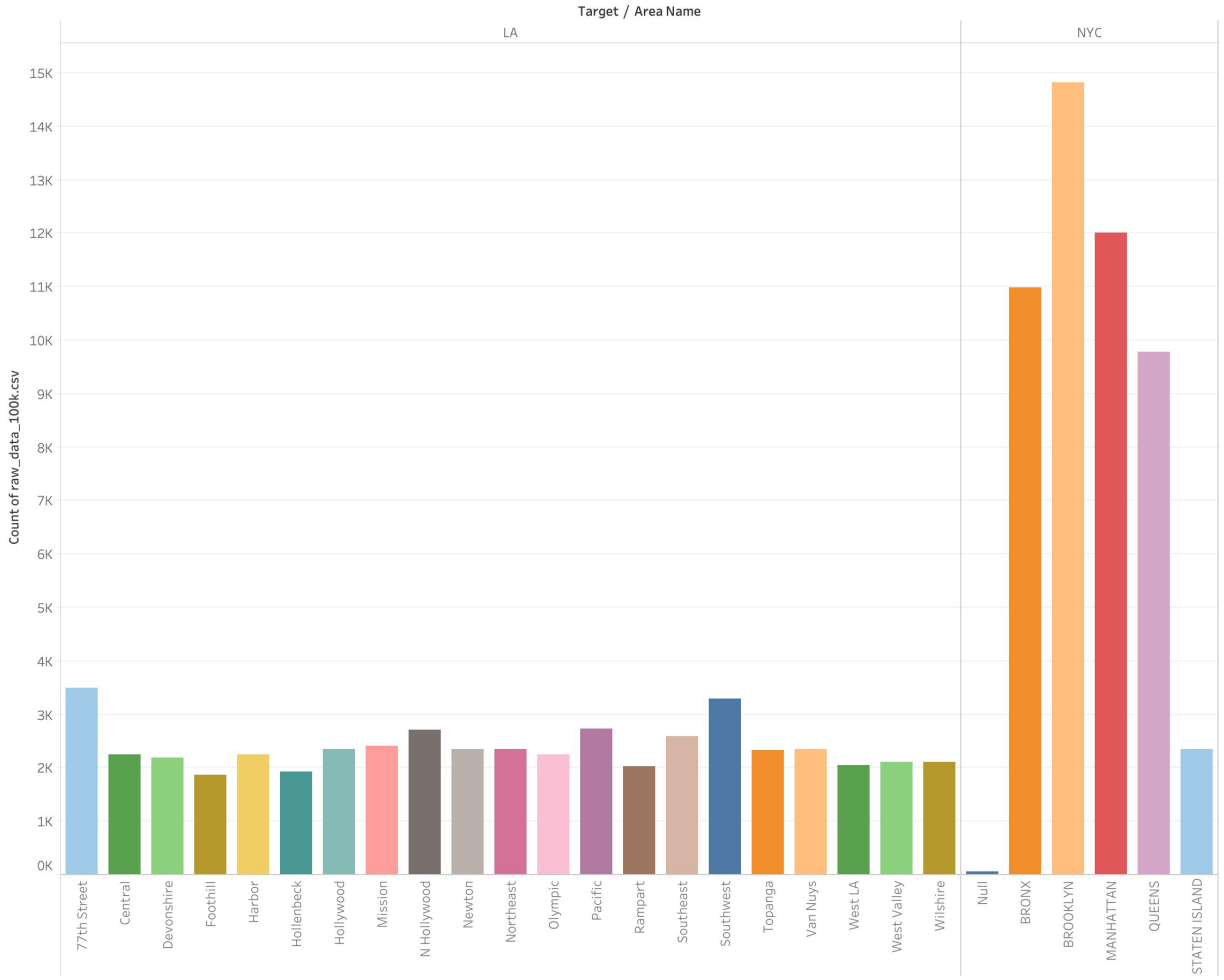
| Area Name | Time Occurred 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 77th Street | 124 | 84 | 82 | 70 | 58 | 36 | 59 | 103 | 116 | 126 | 155 | 145 | 241 | 151 | 178 | 183 | 171 | 181 | 190 | 164 | 169 | 166 | 183 | 162 |
| BRONX | 480 | 316 | 254 | 246 | 197 | 171 | 176 | 272 | 407 | 402 | 407 | 427 | 553 | 501 | 535 | 625 | 591 | 596 | 625 | 617 | 588 | 581 | 473 | 464 |
| BROOKLYN | 809 | 486 | 394 | 299 | 320 | 254 | 259 | 368 | 614 | 625 | 627 | 635 | 887 | 719 | 738 | 875 | 797 | 845 | 880 | 818 | 866 | 815 | 756 | 636 |
| Central | 103 | 72 | 80 | 55 | 50 | 24 | 62 | 64 | 104 | 110 | 120 | 114 | 152 | 114 | 132 | 136 | 146 | 106 | 133 | 126 | 130 | 116 | 105 | 111 |
| Devonshire | 95 | 63 | 49 | 34 | 25 | 28 | 43 | 45 | 101 | 92 | 102 | 79 | 227 | 119 | 121 | 137 | 140 | 148 | 149 | 155 | 115 | 111 | 95 | 102 |
| Foothill | 81 | 60 | 36 | 34 | 27 | 27 | 46 | 56 | 93 | 81 | 68 | 84 | 166 | 70 | 88 | 94 | 95 | 86 | 91 | 110 | 100 | 91 | 93 | 88 |
| Harbor | 106 | 58 | 44 | 35 | 29 | 34 | 53 | 41 | 105 | 81 | 84 | 84 | 152 | 87 | 102 | 96 | 92 | 117 | 125 | 114 | 111 | 116 | 135 | 95 |
| Hollenbeck | 80 | 37 | 34 | 31 | 30 | 19 | 29 | 43 | 62 | 62 | 73 | 65 | 128 | 85 | 75 | 96 | 100 | 81 | 92 | 88 | 81 | 97 | 101 | 80 |
| Hollywood | 171 | 127 | 113 | 60 | 42 | 31 | 34 | 48 | 86 | 64 | 111 | 93 | 155 | 86 | 96 | 114 | 85 | 99 | 137 | 112 | 143 | 148 | 126 | 109 |
| MANHATTA.. | 498 | 374 | 338 | 289 | 234 | 147 | 143 | 243 | 397 | 502 | 449 | 494 | 658 | 517 | 585 | 686 | 659 | 609 | 703 | 609 | 570 | 476 | 421 | 457 |
| Mission | 108 | 64 | 49 | 31 | 34 | 30 | 51 | 62 | 106 | 89 | 87 | 88 | 154 | 96 | 103 | 120 | 121 | 135 | 140 | 135 | 121 | 130 | 134 | 105 |
| N Hollywood | 142 | 71 | 71 | 52 | 36 | 39 | 39 | 68 | 107 | 100 | 111 | 96 | 246 | 113 | 134 | 129 | 146 | 185 | 184 | 169 | 143 | 136 | 131 | 108 |
| Newton | 99 | 70 | 57 | 52 | 43 | 38 | 32 | 64 | 120 | 89 | 120 | 98 | 149 | 132 | 123 | 109 | 119 | 119 | 122 | 132 | 121 | 115 | 107 | 112 |
| Northeast | 88 | 59 | 39 | 33 | 32 | 31 | 49 | 59 | 110 | 70 | 95 | 89 | 221 | 94 | 123 | 103 | 116 | 125 | 136 | 132 | 133 | 122 | 120 | 121 |
| Olympic | 101 | 73 | 69 | 50 | 33 | 23 | 37 | 51 | 100 | 82 | 107 | 96 | 165 | 89 | 91 | 130 | 112 | 120 | 127 | 135 | 100 | 120 | 120 | 126 |
| Pacific | 104 | 74 | 59 | 45 | 32 | 38 | 43 | 74 | 122 | 134 | 123 | 99 | 183 | 125 | 139 | 141 | 161 | 131 | 167 | 131 | 140 | 142 | 121 | 109 |
| QUEENS | 505 | 302 | 269 | 264 | 290 | 163 | 185 | 282 | 382 | 466 | 431 | 416 | 670 | 447 | 479 | 629 | 554 | 611 | 596 | 549 | 597 | 528 | 542 | 451 |
| Rampart | 96 | 78 | 79 | 31 | 33 | 26 | 30 | 49 | 75 | 83 | 80 | 71 | 119 | 78 | 85 | 106 | 89 | 107 | 139 | 130 | 117 | 134 | 133 | 117 |
| Southeast | 82 | 74 | 69 | 58 | 35 | 40 | 53 | 70 | 99 | 89 | 127 | 101 | 197 | 113 | 125 | 137 | 133 | 132 | 150 | 130 | 156 | 137 | 123 | 115 |
| Southwest | 121 | 71 | 61 | 56 | 40 | 42 | 41 | 88 | 142 | 116 | 146 | 160 | 241 | 172 | 216 | 202 | 187 | 198 | 198 | 173 | 173 | 153 | 136 | 127 |
| STATEN ISL.. | 109 | 86 | 65 | 46 | 30 | 26 | 33 | 60 | 100 | 100 | 119 | 99 | 160 | 103 | 153 | 171 | 133 | 151 | 161 | 141 | 124 | 116 | 102 | 120 |
| Topanga | 82 | 59 | 45 | 35 | 34 | 26 | 35 | 53 | 119 | 99 | 94 | 106 | 246 | 124 | 134 | 139 | 131 | 140 | 142 | 117 | 141 | 115 | 113 | 86 |
| Van Nuys | 118 | 62 | 46 | 44 | 29 | 21 | 35 | 57 | 96 | 106 | 90 | 105 | 232 | 111 | 120 | 103 | 135 | 135 | 144 | 114 | 150 | 129 | 104 | 98 |
| West LA | 93 | 44 | 38 | 35 | 21 | 25 | 27 | 40 | 104 | 104 | 93 | 104 | 214 | 102 | 124 | 130 | 132 | 129 | 121 | 131 | 130 | 99 | 98 | 77 |
| West Valley | 103 | 58 | 50 | 24 | 33 | 32 | 33 | 53 | 81 | 100 | 100 | 92 | 222 | 92 | 114 | 120 | 95 | 125 | 105 | 131 | 118 | 104 | 109 | 86 |
| Wilshire | 78 | 52 | 45 | 34 | 24 | 27 | 27 | 46 | 99 | 90 | 116 | 112 | 186 | 113 | 115 | 95 | 137 | 113 | 127 | 124 | 131 | 104 | 98 | 81 |

Count of cleaned_data..
19 887

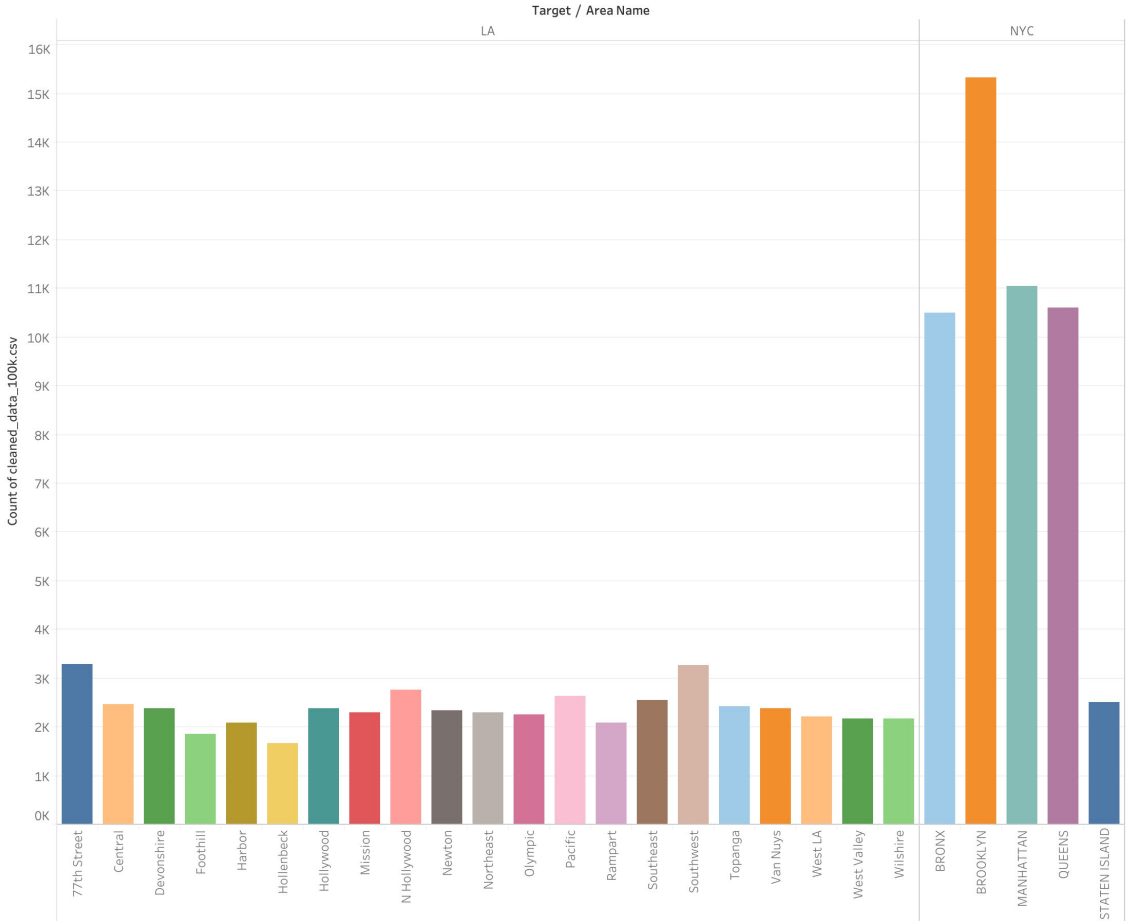
Count of cleaned_data_100k.csv broken down by Time Occurred Hour vs. Area Name. Color shows count of cleaned_data_100k.csv. The marks are labeled by count of cleaned_data_100k.csv.

# Data Visualization

**Histogram of Frequency count vs Target/Area Name**



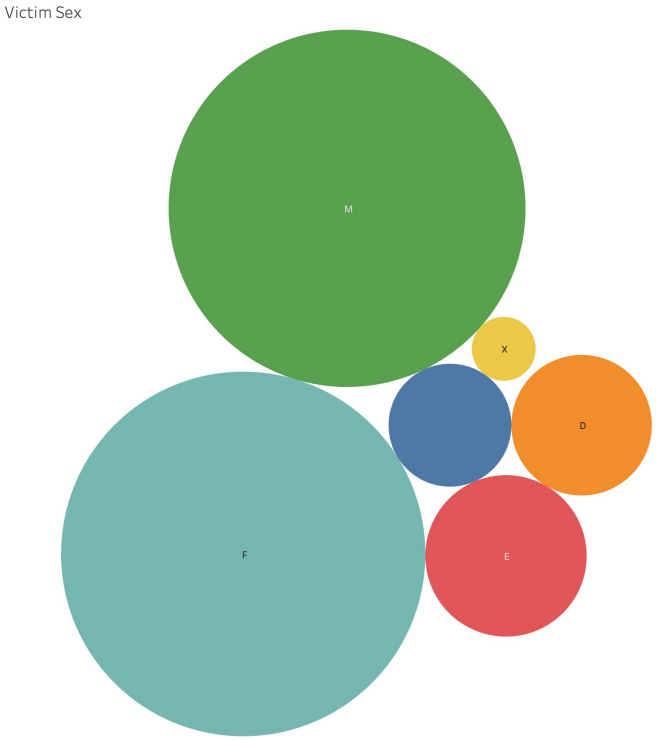

Before removing null values

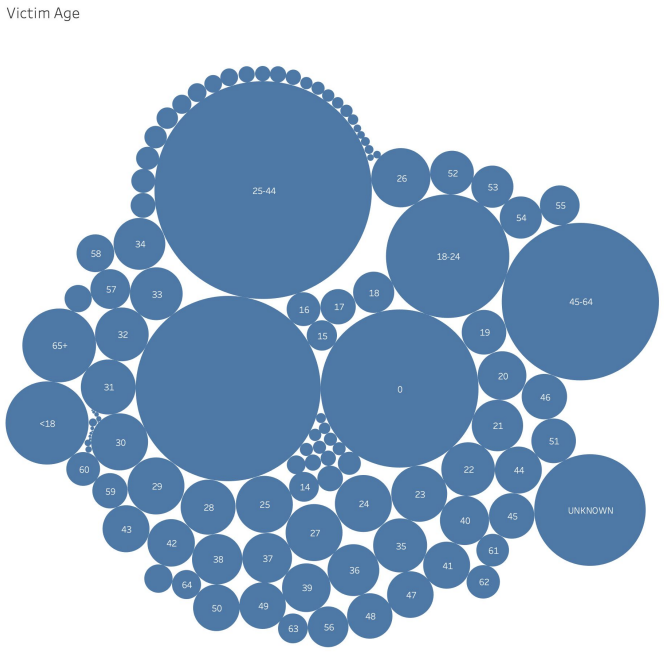


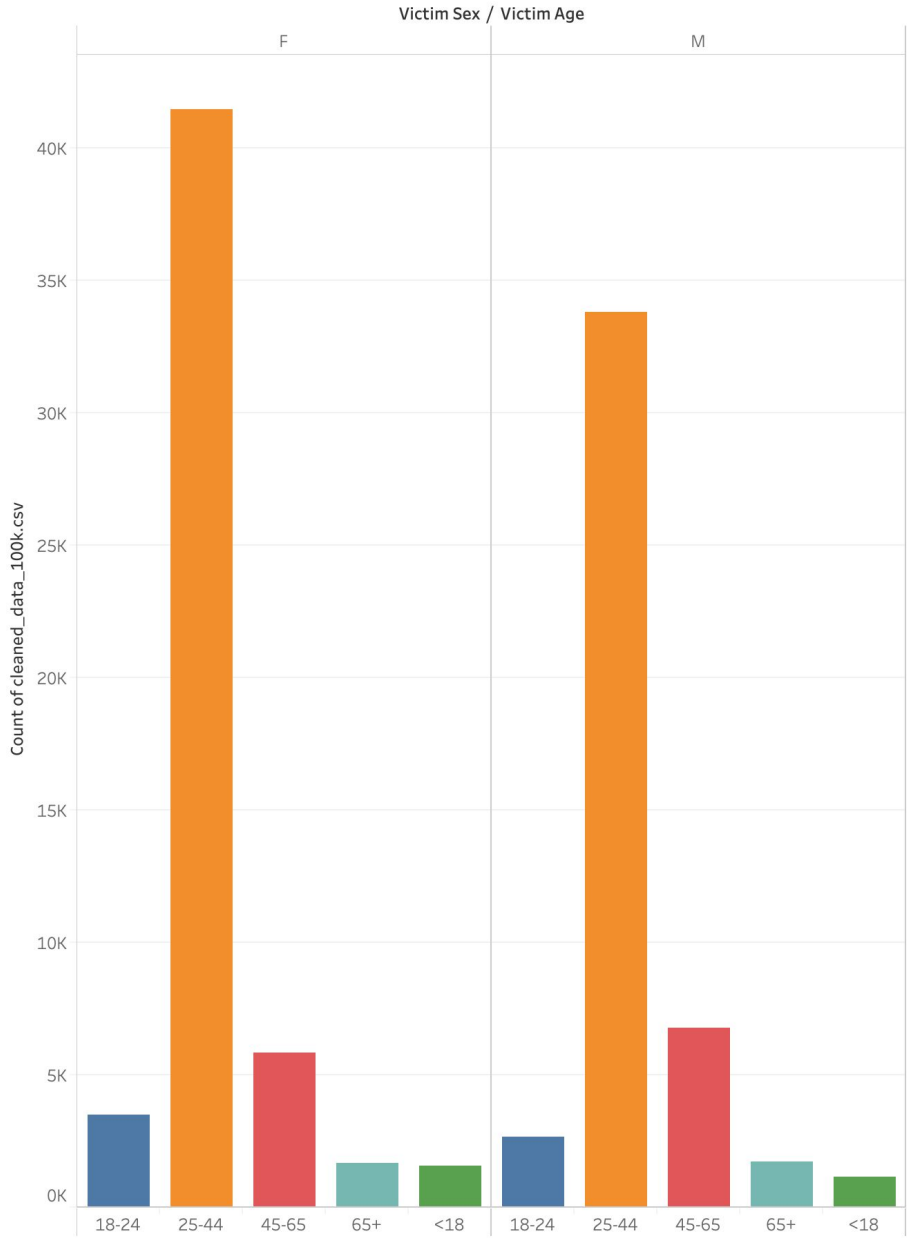

After removing null values

# Data Visualization

**Different views for victim age and sex in raw data**

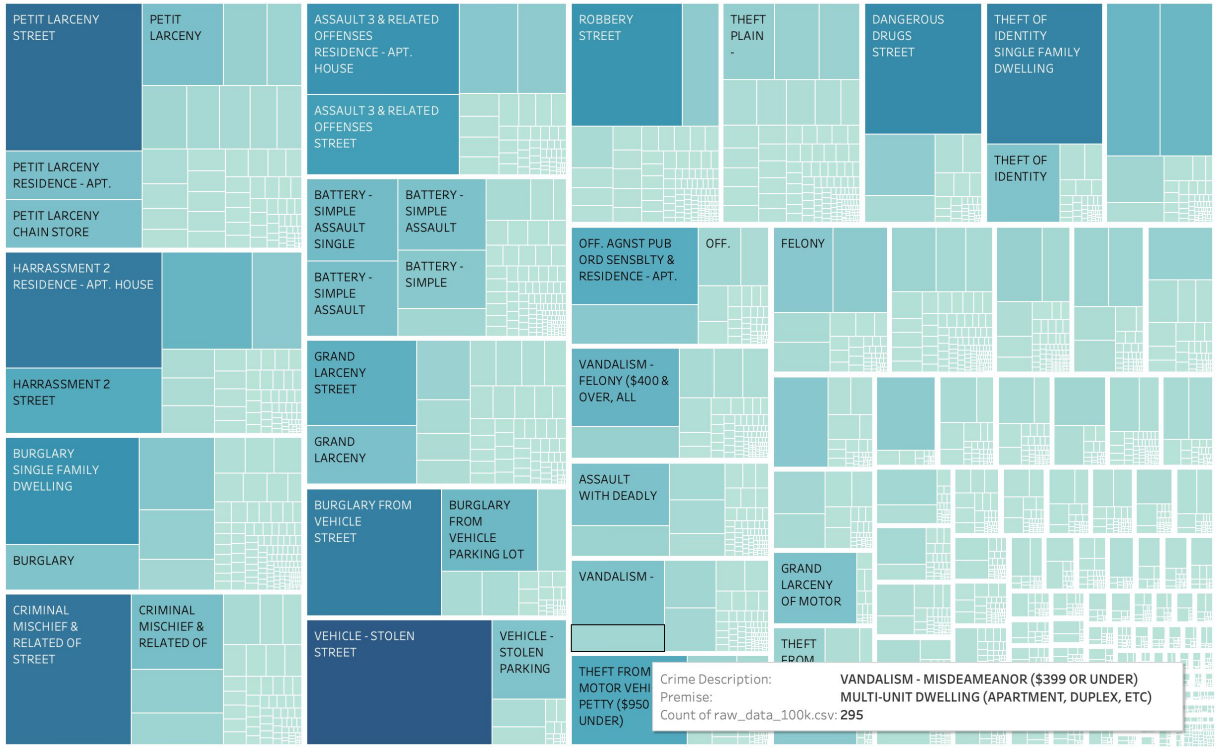


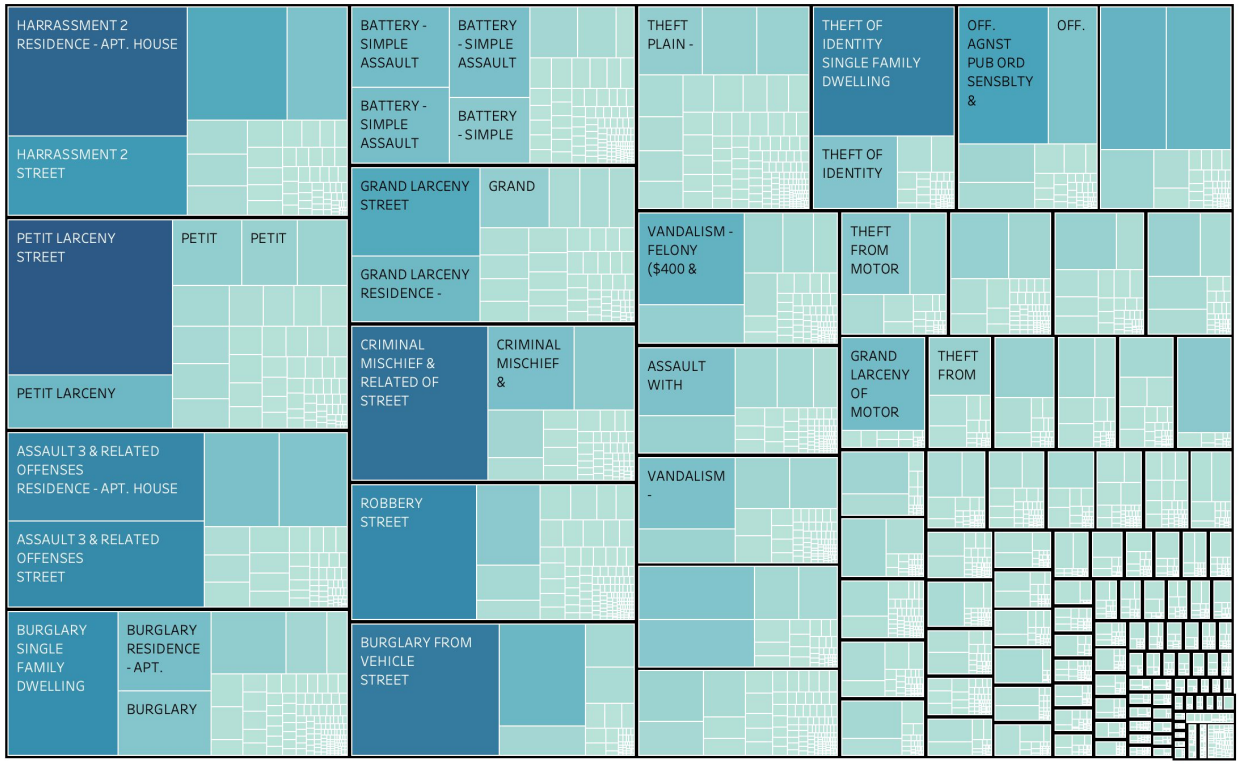



# Data Visualization

The following visualization shows block clusters based on crime reported in a particular precinct with color being used to indicate frquency.




LA




NYC

# Issues Faced

- Merging data from 2006 to 2010 and from 2017 to 2019 was difficult.
- The process of cleaning, preparing, and merging data was time consuming.
- Atleast some of the enriched data was lost in some attributes as we had to coarse grain the formats of the attributes as one of the datasets used coarse grained methods.
- There were a lot of irrelevant attributes, present in both datasets, which weren’t apparent, visualization techniques helped to find out these problems.

# Conclusion

- 3 algorithms J48 Decision Tree, Bayes Net and Naive Bayes were used to classify most targeted **victim age** with J48 showing a 74% accuracy amongst 5 classes.

- Algorithm performance was enhanced due to balancing the target classes which tackled the dataset imbalance problem

- The same algorithms was used to find most **targeted area**

- The analysis done on the data can help provide security in an efficient manner at a particular place and/or at a particular time

- Clustering was done along with visualization to infer important information like

    1) Brooklyn was found to be the most dangerous area at around 12 o'clock with most number of cases.