

San Francisco Crime Classification

A Review of crimes from 2003 to 2017

Randall Lueck | IST 652 | 12/10/2018

Contents

[Obtaining the Data 2](#_Toc532066635)

[Data Dictionary for SF Crime Information 2](#_Toc532066636)

[Data Dictionary for SF Police District Shapefile 2](#_Toc532066637)

[Scrub 3](#_Toc532066638)

[Duplicates and “NA” Values 3](#_Toc532066639)

[Incorrect Coordinates and Formats 3](#_Toc532066640)

[Datetime Format 3](#_Toc532066641)

[Examine 3](#_Toc532066642)

[Crimes by Year 4](#_Toc532066643)

[Seasonal Crime 5](#_Toc532066644)

[Time of Day 5](#_Toc532066645)

[Days of the Week 6](#_Toc532066646)

[Crime Categorization 6](#_Toc532066647)

[Model 10](#_Toc532066648)

[Interpret 11](#_Toc532066649)

# Obtaining the Data

The data in this paper comes from San Francisco’s city data website (data.sfgov.org). The original data package consists of crime information from 2003 to 2018. Over this period the department recorded over 2 million crimes and incidents ranging from non-criminal activities, to misdemeanors, to major federal offenses. They’ve made the information public in a comma-separated values (csv) file so that the public can easily view and manipulate the information. For each crime, the police collected information under one of the following thirteen categories:

## Data Dictionary for SF Crime Information

|  |  |  |
| --- | --- | --- |
| **Term** | **Description** | **Example** |
| Incident Number | A unique number to identify the incident | 50436712 |
| Category | Type of crime | Assault |
| Description | An in-depth explanation of the incident | Assault with knife |
| DayOfWeek | The day of the incident | Monday |
| Date | Day/Month/Year of the incident | 2007-04-21 |
| Time (Military) | Time of the incident | 18:00 |
| PdDistrict | Police district where the incident occured | Tenderloin |
| Resolution | How the police resolved the incident | Got cat down from tree |
| Address | Address or rough location of the occurrence | 1100 |
| X | Latitude | -122.4468378 |
| Y | Longitude | 37.76225503 |
| Location | Dictionary of location data | {'longitude': '-122.435002864271', 'needs\_recoding': False, 'latitude': '37.7608878061245', 'human\_address': '{"address":"","city":"","state":"","zip":""}'} |
| PdId | Numeric identifier of the police department. | 5043671204134 |

Additionally, the city data website has shapefiles consisting of the ten police districts of San Francisco. These files contain geographic information with data linked to the name, size, and border locations for each district. The shapefile contains the following information.

## Data Dictionary for SF Police District Shapefile

## 

|  |  |  |
| --- | --- | --- |
| **Term** | **Description** | **Example** |
| District | Name of the police department district outlined | Tenderloin |
| Company | Department in charge of the district ranges from A to J | B |
| Shape\_Leng | Length of the district in meters measured horizontally | 87550.275142 |
| Shape\_Le\_1 | Height of the district in meters measured vertically | 100231.353916 |
| Shape\_Area | Area covered by the district measured in meters | 9.13414e+07 |
| Geometry | Shape of the district defined by an array of latitude/longitude pairs | (POLYGON((-122.391861389 37.7942468), (…)) |

# Scrub

Both data sets had only a few incorrect, duplicated, or missing values. Most of these corrections were easily made using Python’s built-in libraries and functions. However, some additional libraries were necessary. Additional libraries added for viewing and scrubbing the data sets were:

* Numpy
* Pandas
* Datetime
* GeoPandas

## Duplicates and “NA” Values

Upon viewing the layout of the above thirteen columns for crime data, it appeared that some of the data was either missing, duplicated, or formatted incorrectly. The first step was to remove any duplicate or NA values found in the data using the “dropna” and “drop\_duplicate” commands in Python.

## Incorrect Coordinates and Formats

San Francisco has a latitude and longitude coordinates of 37.773972 and -122.431297 respectively. However, over one hundred records appeared in locations far north of these boundaries in the middle of the Arctic Ocean. Since this only accounts for about 100 records out of over 2 million, these records were deleted. One could easily hypothesize that the police didn’t record the location associated with these crimes or expected to correct these data later.

For the shapefile, the producer of the file used a coordinate reference system (CRS) other than lat/lon. To make the map work, the CRS data was changed to the proper format using the geopandas library to match the data in the crime records file. Once changed, the coordinates appeared in the correct format and plotted correctly with the crime data.

## Datetime Format

Two of the categories for crime, time and date, are listed as strings in the original data. To make these two categories measurable in terms of time, they were converted to the datetime format using the datetime library. Also, all crime data from 2018 was removed due to incomplete data for the year. This was done to minimize any skewness possibly caused by data recorded from a partial year.

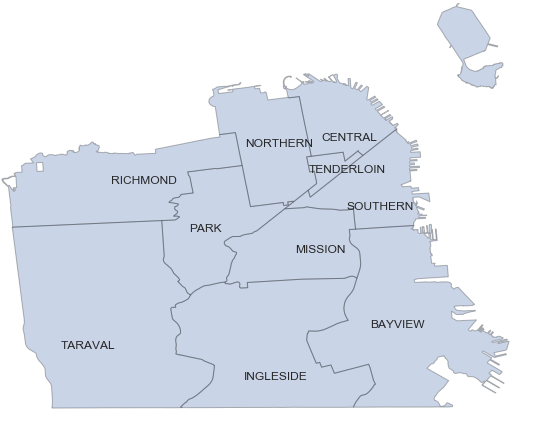
These two columns, once corrected, helped to parse the data into shorter, more readable bits of information. To make the information more general in terms of time, the data was grouped into years, seasons, and time-of-day. The finalized data set contains all of the original columns from the years 2003 to 2017 plus the four seasons, and time of day categorized as morning, afternoon, evening, and night.

# Examine

To visualize the data, a few additional libraries were imported. These include matplotlib and seaborn for graphing and traditional visualization. This examination attempts to answer the following:

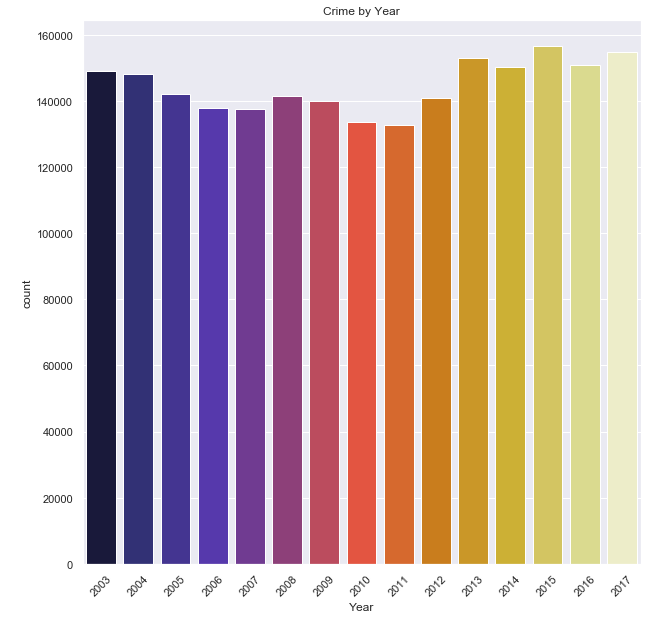
1. Which areas have the highest crime rate by time of day?
2. Which areas have the lowest?
3. Is there a way to use this data to help police decide where to patrol for certain times of the day?
4. Is crime data consistent for each season?

The below map illustrates the nine police districts of San Francisco. The northeast corner represents the downtown and financial districts consisting of North, Central, Southern, and Tenderloin. This map should make it easier to follow along with the graphical output.



## Crimes by Year

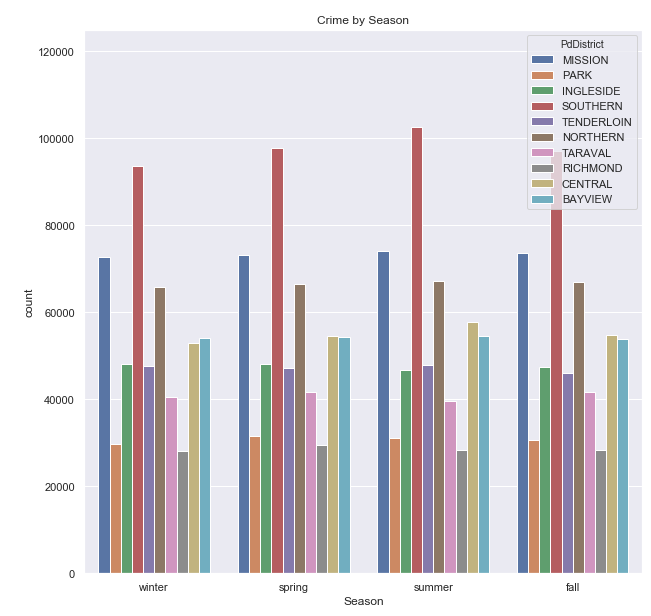
Using both seaborn and built-in libraries, the crimes by year were easily counted. From 2003 to 2017 there were 144547 crimes per year on average. Crimes rates peaked in 2015 with 156526 crimes. In 2011 San Francisco had its lowest crime count over this period with only 132697 crimes. These numbers show that crime could be predicted within reason since there haven’t been a lot of outliers in the data over the period analyzed. Here’s a visualization of crime per year:



## Seasonal Crime

Since crime appears to maintain a trend over a yearly period, I next looked at the crime rate by season in San Francisco. The data was split into seasonal information using calendar days to evenly split them into three-month sections. Spring covers days 80 to 172, Summer covers days 172 to 264, Fall covers days 264 to 355, and Winter covers all of the other days.

Splitting the data seasonally shows that, overall the seasons have no effect on crime in San Francisco. Per season crimes average to 542053 with every season in the data set having between 530000 and 550000 crimes overall. However, splitting the districts by season shows that some districts are represented with a much higher frequency in the data set.



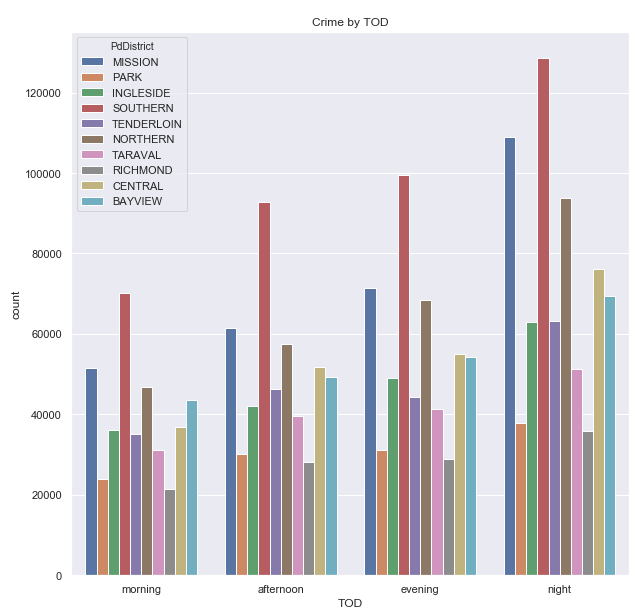
While most districts appear to have an even spread across all four seasons, crime in the Central and Southern districts tend to have slightly higher crime in Summer.

## Time of Day

In contrast to the previous two categories, time of day shows a wide spread of crime frequency. The range of crime frequency around time of day is 331547 with the following crime frequencies counted per time of day over the 15-year period:

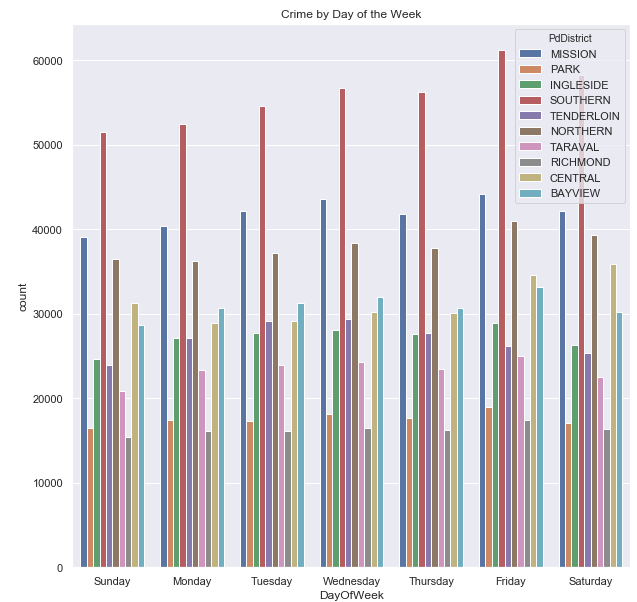
* Morning: 396,851
* Afternoon: 499,310
* Evening: 543,653
* Night: 728,398

Graphically, this data shows an obvious trend of a rise in crime from morning to night:



## Days of the Week

Similarly, the days of the week show a trend in crime rising from Sunday to Friday overall with crime trending downward on Saturday. For all the data so far, the highest crime rates tend to belong to the downtown and financial centers around the northeast corner and the east side of the city. The highest crime rates are found in Southern, Mission, Northern, Central, and Bayview regardless of the period. This trend becomes even more apparent when the crimes are split into smaller categories and when modeling later.



## Crime Categorization

This dataset contains a total of 39 types of crimes recorded across 15 years. The 39 types of crimes along with number of occurrences over the period are:

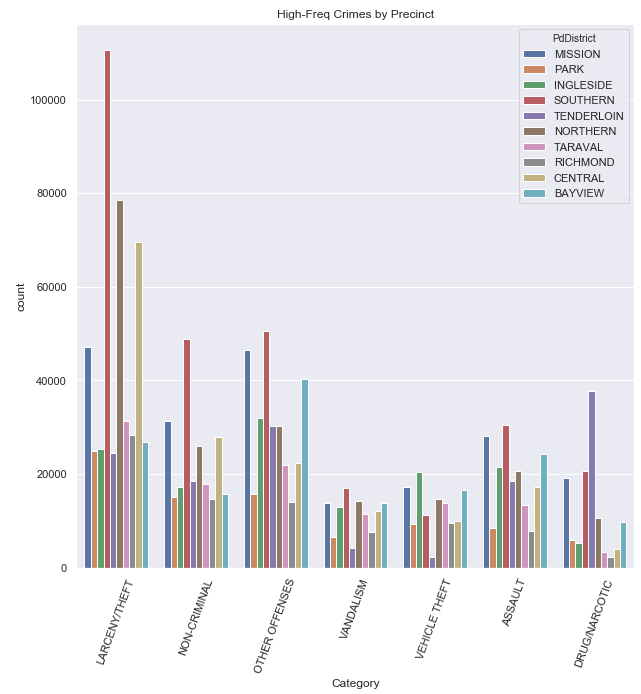
|  |  |
| --- | --- |
| Arson: 3834 | Non-criminal: 233066 |
| Assault: 19008 | Other offenses: 303752 |
| Bad checks: 916 | Pornography/obscene mat: 56 |
| Bribery: 796 | Prostitution: 16651 |
| Burglary: 89444 | Recovered vehicle: 8716 |
| Disorderly conduct: 9943 | Robbery: 54793 |
| Driving under the influence: 5580 | Runaway: 4353 |
| Drug/narcotic: 118139 | Secondary codes: 25188 |
| Drunkenness: 9742 | Sex offenses, forcible: 11422 |
| Embezzlement: 2943 | Sex offenses, non-forcible: 420 |
| Extortion: 729 | Stolen property: 11626 |
| Family offenses: 1170 | Suicide: 1274 |
| Forgery/counterfeiting: 22855 | Suspicious occ: 78727 |
| Fraud: 40731 | Trea: 14 |
| Gambling: 339 | Trespass: 18933 |
| Kidnapping: 5272 | Vandalism: 113262 |
| Larceny/theft: 466787 | Vehicle theft: 125112 |
| Liquor laws: 4069 | Warrants: 99747 |
| Loitering: 2414 | Weapon laws: 21614 |
| Missing person: 63697 |  |

Because there are so many crimes, the following graphs categorize them into three different bins:

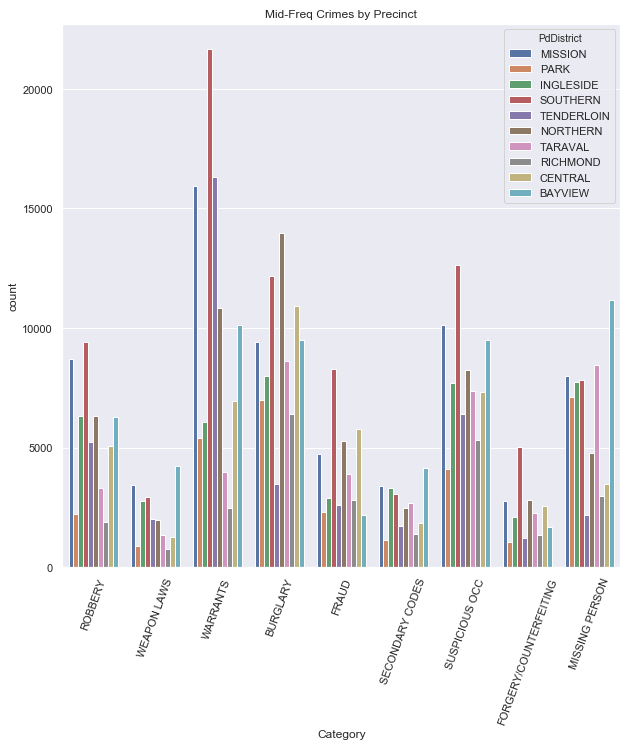
* High-frequency crimes
* Mid-frequency crimes
* Low-frequency crimes.

These graphs only represent the 25 most frequent crimes from the whole data set. They were

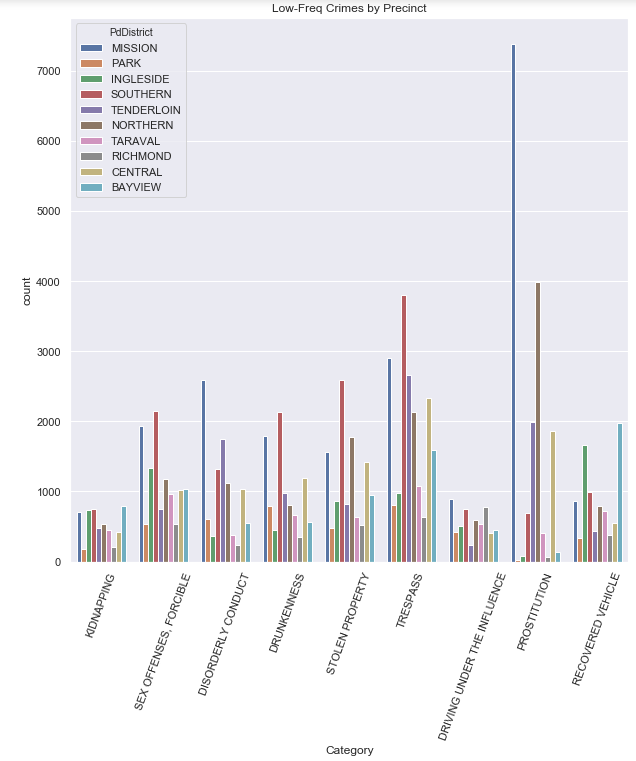
chosen based on frequency, with only crimes that occur more often than 5000 times appearing in the graphed data.



The high-frequency crimes data set encapsulates any crimes that happened 100,000 times or more over the 15-year period represented in the data set. In this subset of the data, larceny and theft appear most often while vandalism just barely makes the cut at about 113,000 occurrences. This graph also shows that most crimes are more likely to occur in Southern and Mission districts in the northeast corner represented by the red and blue bars in the graph. Tenderloin district has the highest rating for drug/narcotic crimes and is also located in the northeast corner of the city. Northern, Central, and Bayview districts located in the same vicinity of the city also appear among some of the higher crime rates for a few categories.

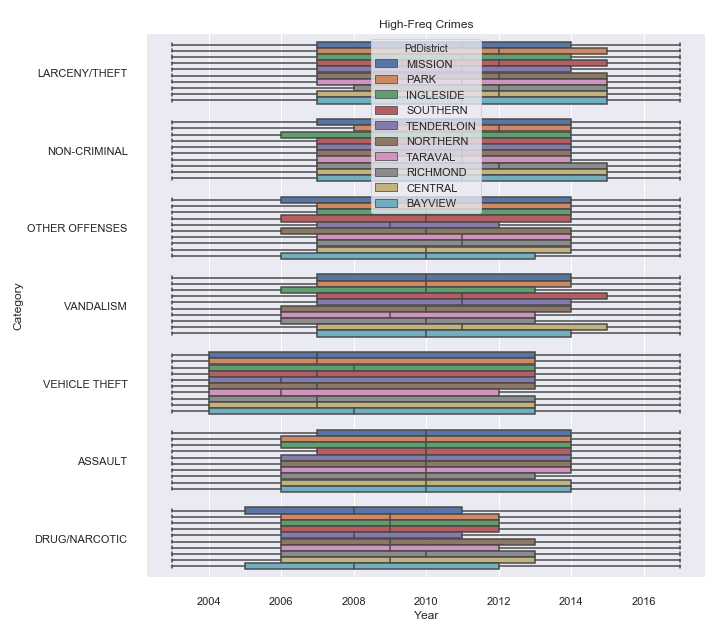


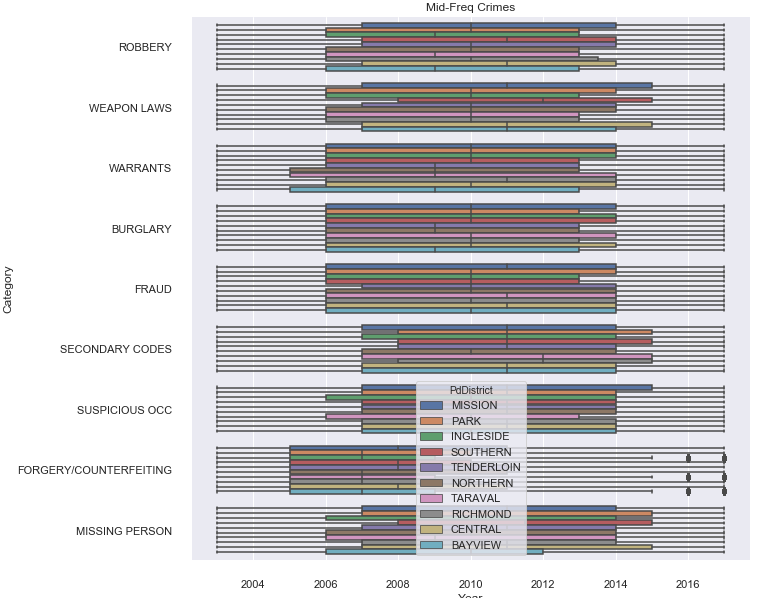
The mid-frequency crimes represent crimes that appear in the data set between 20,000 and 100,000 times. Once again, most crime appears to happen in the northeast part of the city with only a couple of exceptions. A quick glance shows that Southern, Northern, and Bayview stand out in most of these categories. In this category, warrants outweigh the other crimes by a majority.

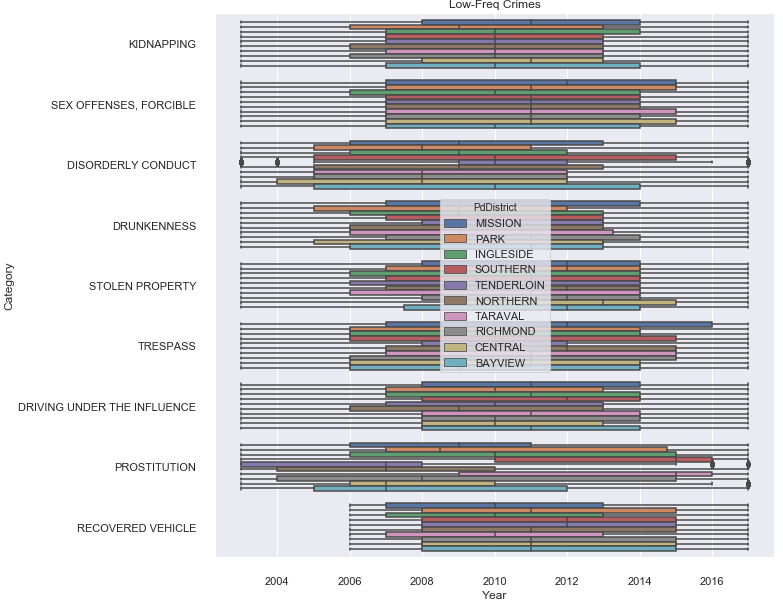


Lastly, the low-frequency graph represents crimes that appear between 5,000 and 20,000 times in the data set. This graph makes apparent that Mission district has a problem with prostitution that overshadows all crimes in the rest of this category. However, the same districts still outweigh most of the city outside of recovered vehicles in Inglewood.

The following graphs represent how crime changed over time using the above categories. While most crimes look to have spiked between 2006 and 2015, prostitution, forgery/counterfeiting, and disorderly conduct appear to have remained steady throughout the 15-year period. In most of the city’s police districts, crime has slightly risen between 2003 and 2017. This trend can be seen in the way most crimes lean toward the left side of graphs. Crime that has remained steady tends to show its median range toward the middle of the graphs. Any crimes that are not as prominent anymore, such as vehicle theft, lean more toward the right side of the graphs.

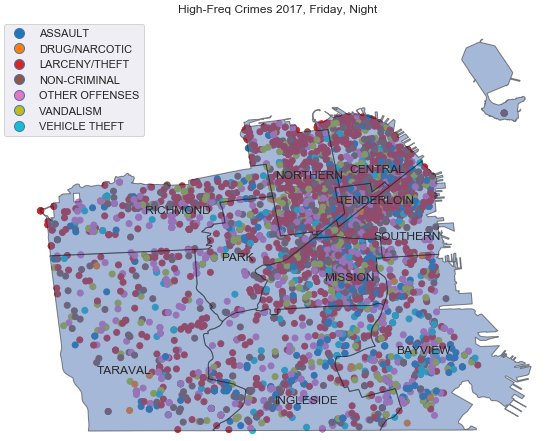






# Model

To model this data, I plotted all crimes on a map of the city. However, over 2 million dots on a city map make it difficult to comprehend how the different crimes are represented. For this reason, I created a function to view the different types of crime depending on a user’s interest. The function takes the arguments of the factors named above to be expected to influence category the most. They are year, time of day, and day of the week. Using this function, the below output represents high-frequency crimes for 2017:



Similar maps can be summoned for mid-frequency and low-frequency crimes for the same period for comparison. These two models show that even lower frequency crimes tend to happen in the same 3 -4 districts consistently.

|  |  |
| --- | --- |
|  |  |

# Interpret

While a machine learning algorithm may provide some insight into when these crimes happen, most researchers agree that predicting random crime is extremely difficult. The above visualizations and models are based on previous data collected over fifteen years. While the visualizations and models may not predict where the next crime will happen, they do shed light on the locations where criminal activity collects over time. Many of the above crimes could take place behind closed doors and so may be more difficult to predict. Also, without demographic or population data for each arrest, topics such as racial profiling or police bias can’t be weeded out for each district. Given these possibilities, the visualizations do show that certain types of crime happen less frequently now in certain areas of the city when viewed by year. Looking outside of the four downtown districts of Northern, Southern, Tenderloin, and Central, which all have an overwhelming amount of data for these visualizations, one can determine that Richmond district has a major problem with theft and larceny. We can also easily see a group of weapons laws violations in Bayview and of prostitution in Mission. Police departments responsible for these areas could focus on those handful of crimes that appear to group on the map and possibly eradicate at least that specific sort of crime in their area using visualizations similar to those above. While this probably won’t eradicate crime completely in any city, it could have an impact in a large metropolitan area like San Francisco.