San Francisco Crime classification

Jia Huang

Xi’an Shiyou University
Chinese Academy of Sciences

October 15, 2020

Overview

Project Overview

Project Background And Purpose

Defn

- Background

  From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay. From Sunset to SOMA, and Marina to Excelsior, this dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods.

- Purpose

  predict the category of crime that occurred, given the time and location visualize the city and crimes (see Mapping and Visualizing Violent Crime for inspiration) Content.

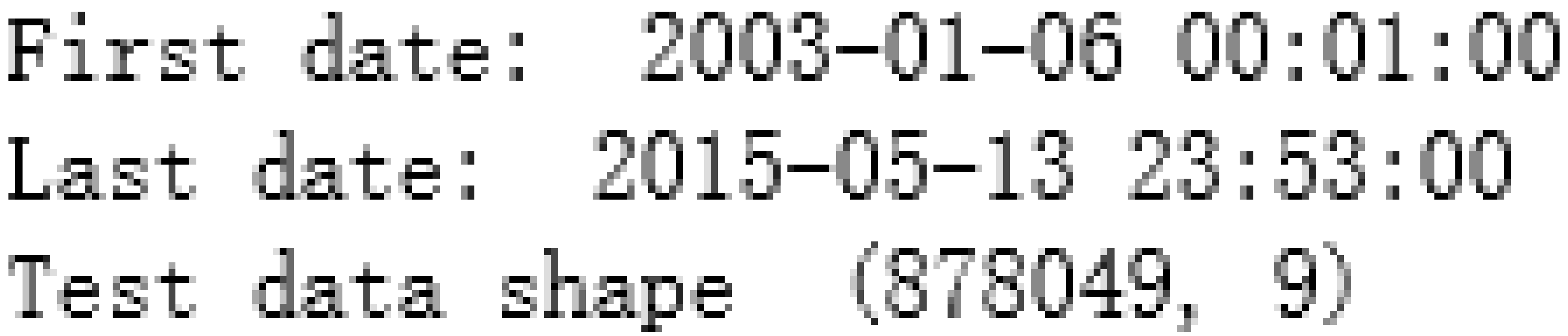
Data Pre-Processing

Date Processing

This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set. There are 9 variables.

Feature Item

By making a comprehensive analysis of the nine characteristic items in the data set mentioned above,we can reach the following conclusions:

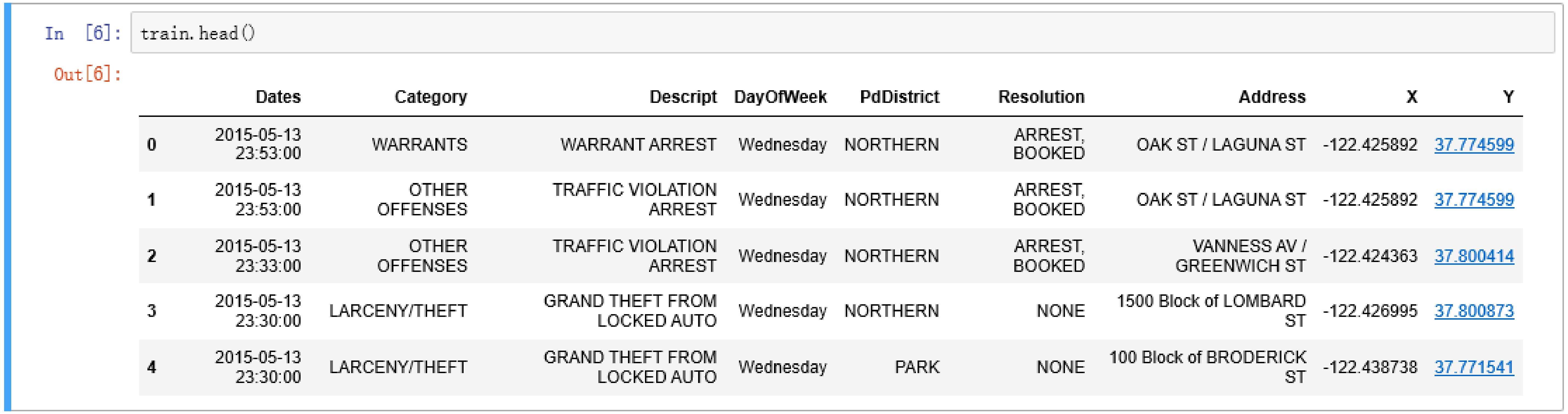


```
First date:  2003-01-06 00:01:00
Last date:  2015-05-13 23:53:00
Test data shape  (878049, 9)
```

Figure 1

The data range was from 1/1/2003 to 5/13/2015, and a data training set containing nine feature items and 87,8049 samples was created.

Features Item



```
In  [6]: train.head()
```

Out[6]:

	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
0	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
1	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.800414
3	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.426995	37.800873
4	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	PARK	NONE	100 Block of BRODERICK ST	-122.438738	37.771541

More specifically it includes the following variables.

- Date - timestamp of the crime Incident.
- Category - category of the crime incident. (This is our target variable.)
- Descript - detailed description of the crime incident
- DayOfWeek - the day of the week
- PdDistrict - the name of the Police Department District
- Resolution - The resolution of the crime incident

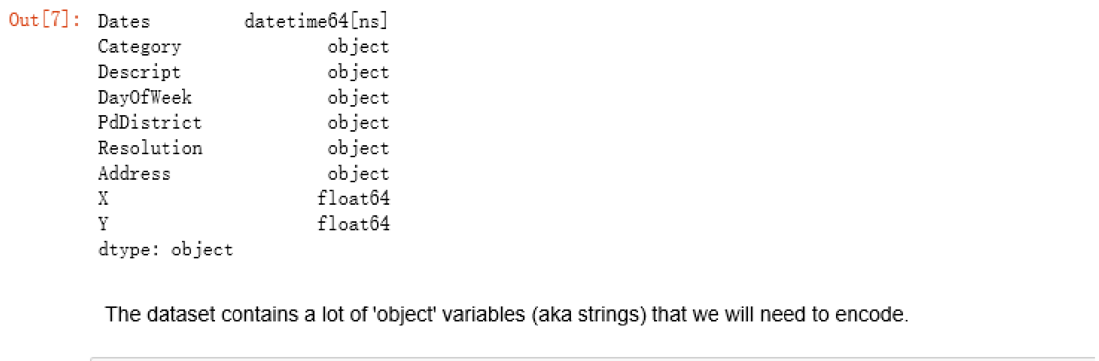
Features Item

- Address - the approximate street address of the crime incident
- X - Longitude
- Y - Latitude

Feature Analysis

Feature Analysis

- The data set contains nine eigenvalues, the data types are as follows, and we can see that the data set contains' object 'variables (also known as strings) that we need to encode.



```
Out[7]: Dates          datetime64[ns]
        Category               object
        Descript               object
        DayOfWeek              object
        PdDistrict             object
        Resolution             object
        Address                object
        X                     float64
        Y                     float64
        dtype: object
```

The dataset contains a lot of 'object' variables (aka strings) that we will need to encode.

Figure 2: Data Type

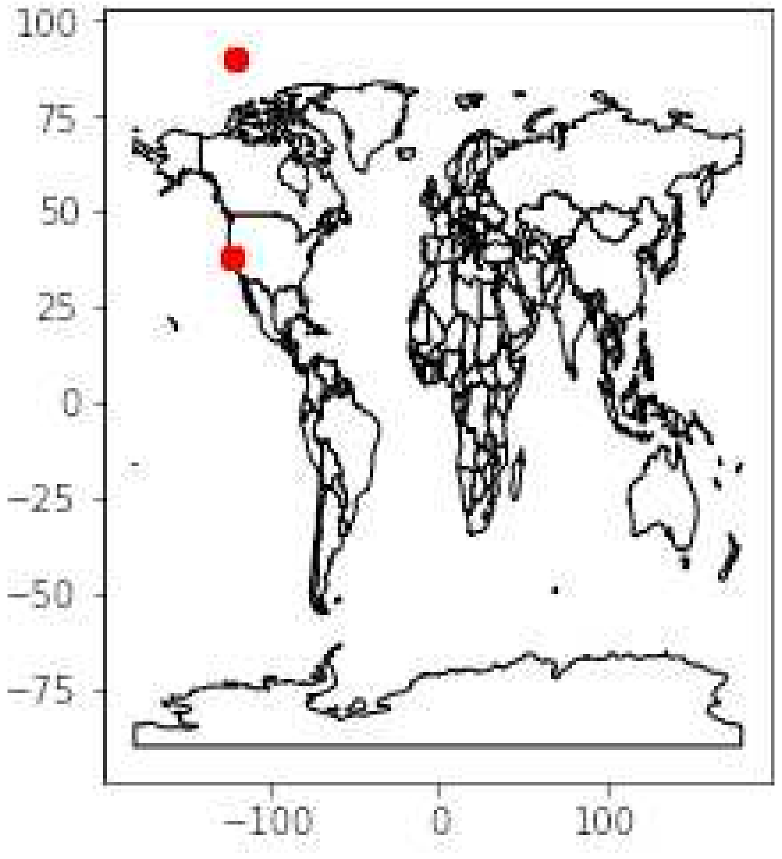
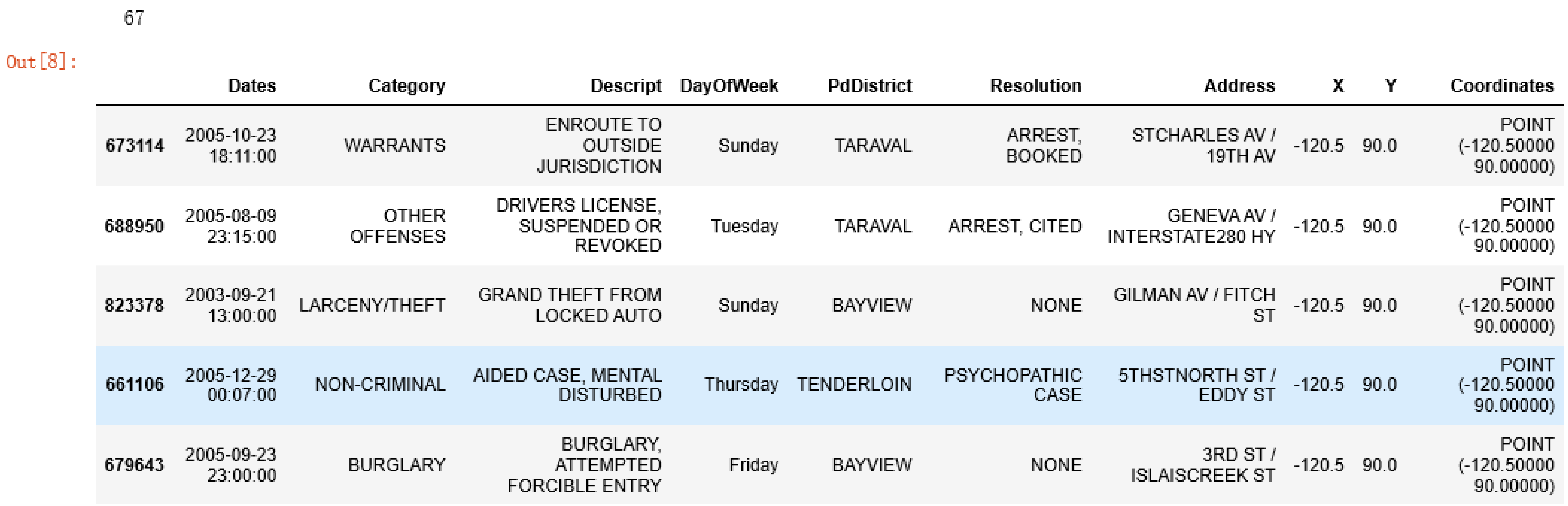


Figure 3: The Total Number of Copies

The 2,323 copies are duplicates and we need to delete them. We will also evaluate the position of the data points using the coordinates.

Feature Analysis

There are also some wrong locations in the data set. After analysis, we found a total of 67 wrong messages, which also means that we cannot use these 67 wrong messages.



```
67
Out[8]:
```

	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y	Coordinates
673114	2005-10-23 18:11:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Sunday	TARAVAL	ARREST, BOOKED	STCHARLES AV / 19TH AV	-120.5	90.0	POINT (-120.50000 90.00000)
688950	2005-08-09 23:15:00	OTHER OFFENSES	DRIVERS LICENSE, SUSPENDED OR REVOKED	Tuesday	TARAVAL	ARREST, CITED	GENEVA AV / INTERSTATE280 HY	-120.5	90.0	POINT (-120.50000 90.00000)
823378	2003-09-21 13:00:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Sunday	BAYVIEW	NONE	GILMAN AV / FITCH ST	-120.5	90.0	POINT (-120.50000 90.00000)
661106	2005-12-29 00:07:00	NON-CRIMINAL	AIDED CASE, MENTAL DISTURBED	Thursday	TENDERLOIN	PSYCHOPATHIC CASE	5THSTNORTH ST / EDDY ST	-120.5	90.0	POINT (-120.50000 90.00000)
679643	2005-09-23 23:00:00	BURGLARY	BURGLARY, ATTEMPTED FORCIBLE ENTRY	Friday	BAYVIEW	NONE	3RD ST / ISLAISCREEK ST	-120.5	90.0	POINT (-120.50000 90.00000)

Dates & Day of the week

These variables are distributed uniformly between 1/1/2003 to 5/13/2015 (and Monday to Sunday) and split between the training and the testing dataset as mentioned before. We did not notice any anomalies on these variables. The median frequency of incidents is 389 per day with a standard deviation of 48.51. Also, there is no significant deviation of incidents frequency throughout the week. Thus we do not expect this variable to play a significant role in the prediction.

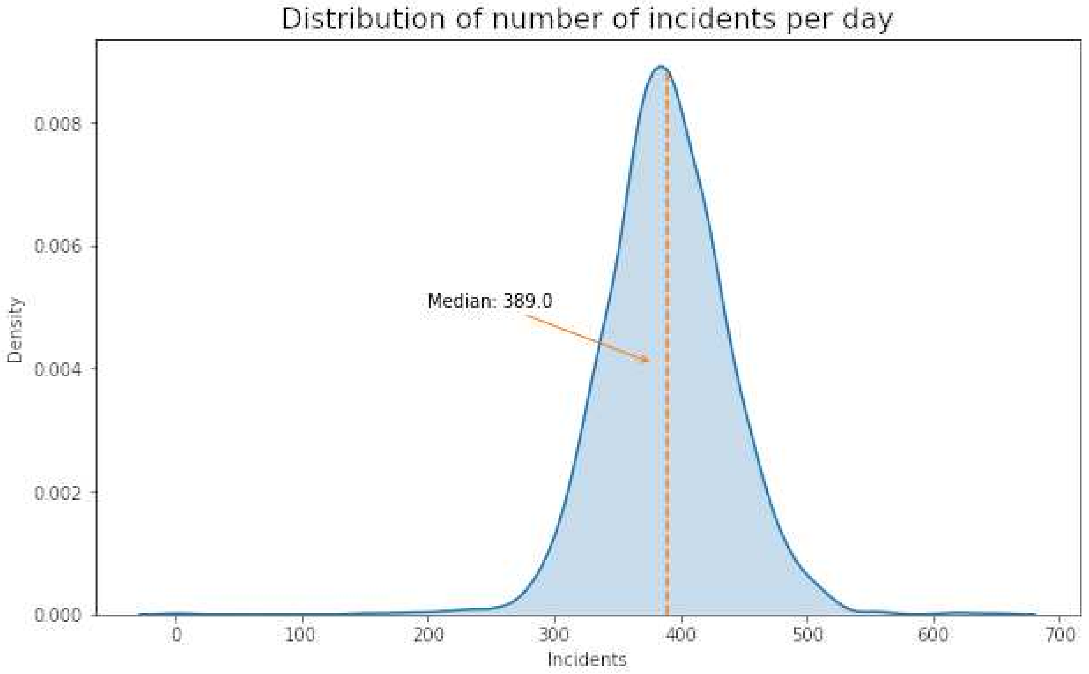


Figure 4: Dates and Day of the Week

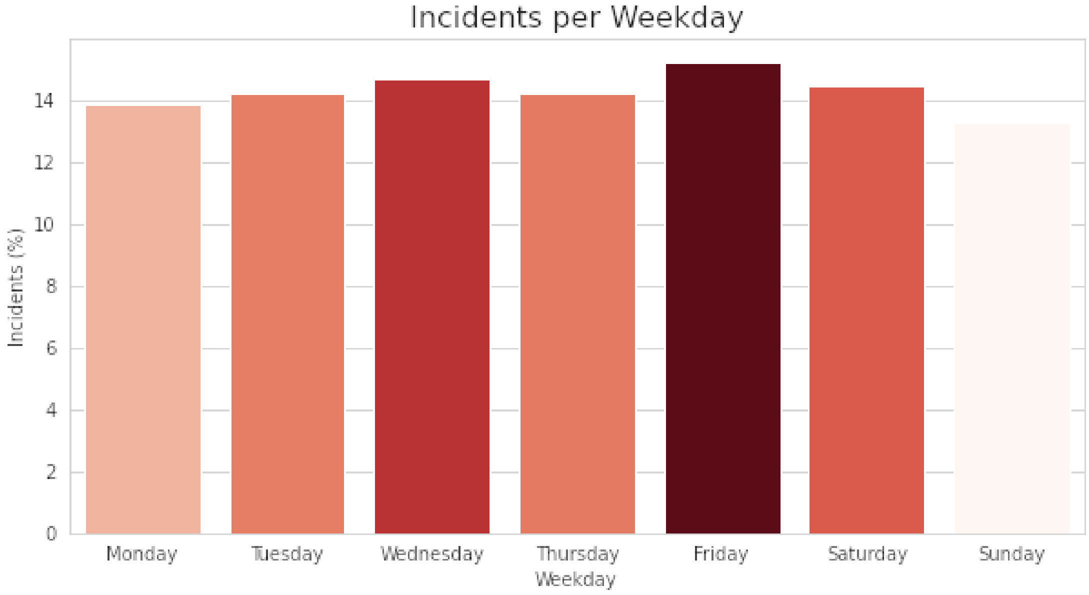


Figure 5: Per Weekday

Category & Police District

There are 39 discrete categories that the police department file the incidents with the most common being Larceny/Theft (19.91%), Non/Criminal (10.50%), and Assault(8.77%). There are significant differences between the different districts of the City with the Southern district having the most incidents (17.87%) followed by Mission (13.67%) and Northern (12.00%).

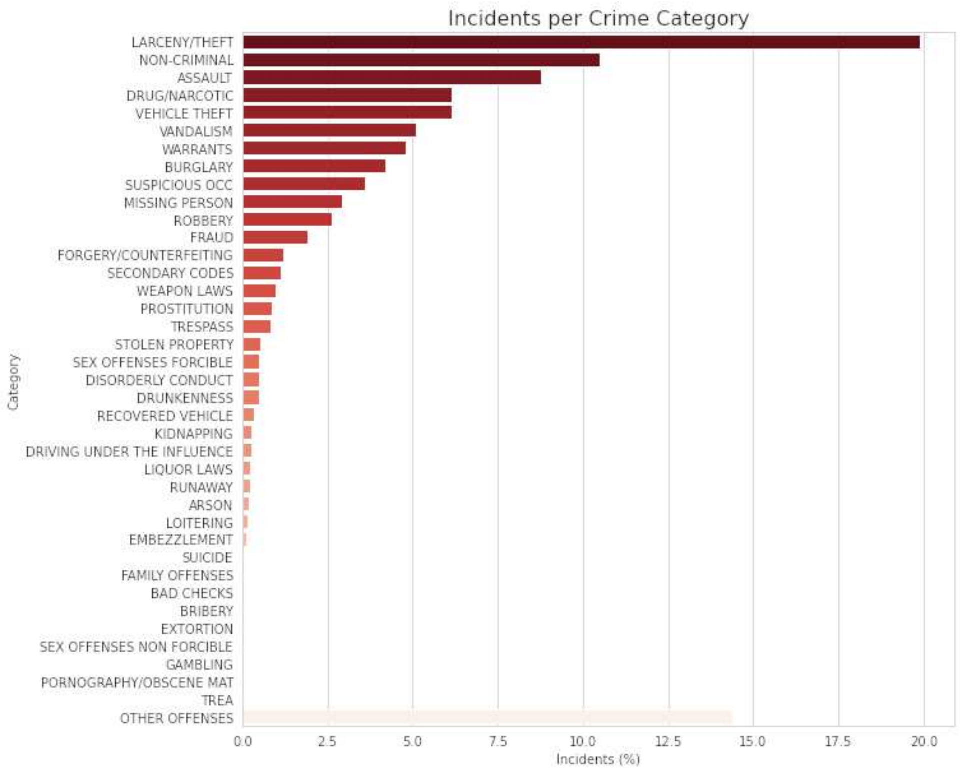


Figure 6: Category

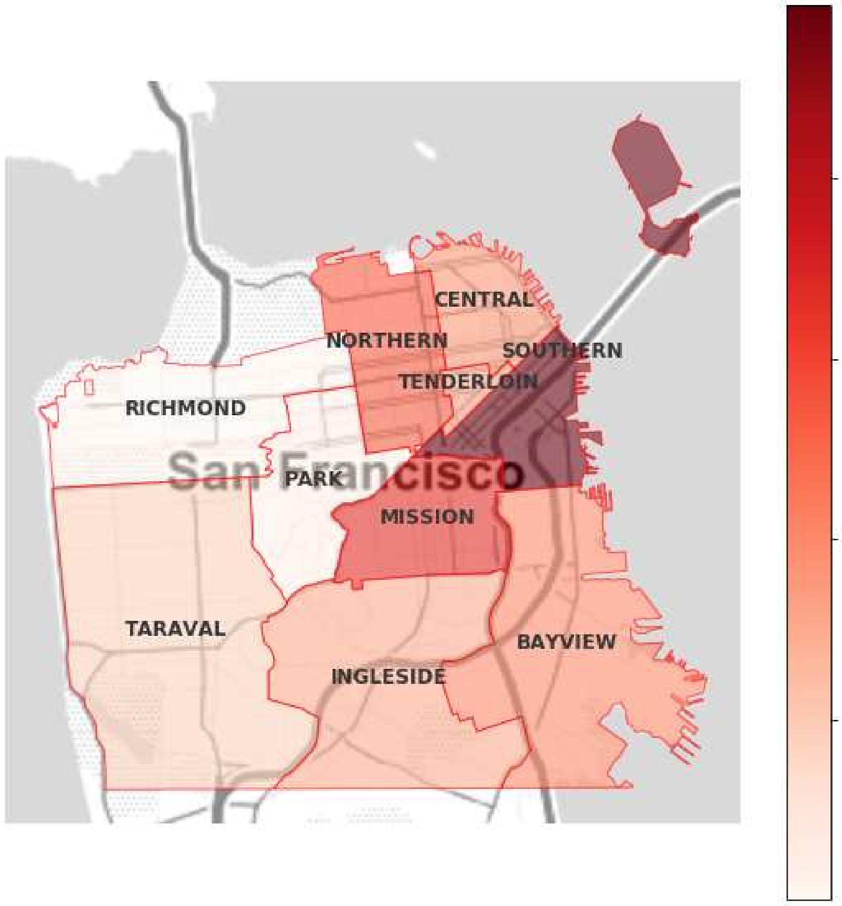


Figure 7: Police District

X & Y

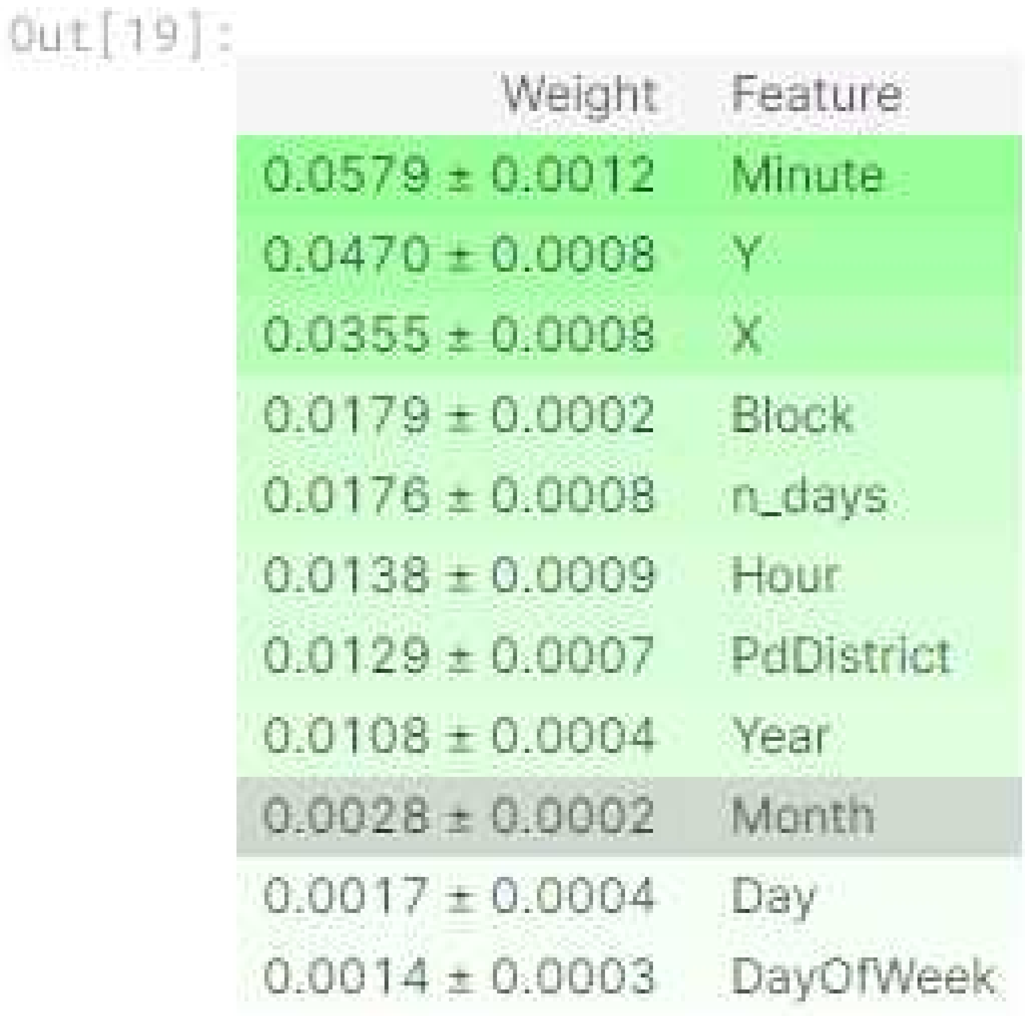
- Address Address, as a text field, requires advanced techniques to use it for the prediction. Instead in this project, we will use it to extract if the incident has happened on the road or in a building block.
- X - Longitude Y - Latitude We have tested that the coordinates belong inside the boundaries of the city. Although longitude does not contain any outliers, latitude includes some 90o values which correspond to the North Pole.

Feature Selection

Feature Engineering

Then, we created additional features. More specifically:

- From the 'Dates' field, we extracted the Day, the Month, the Year, the Hour, the Minute, the Weekday, and the number of days since the first day in the data.
- From the 'Address' field we extracted if the incident has taken place in a crossroad or on a building block.



Out[19]:

Weight	Feature
0.0579 ± 0.0012	Minute
0.0470 ± 0.0008	Y
0.0355 ± 0.0008	X
0.0179 ± 0.0002	Block
0.0176 ± 0.0008	n_days
0.0138 ± 0.0009	Hour
0.0129 ± 0.0007	PdDistrict
0.0108 ± 0.0004	Year
0.0028 ± 0.0002	Month
0.0017 ± 0.0004	Day
0.0014 ± 0.0003	DayOfWeek

Modelling

Calculate the Baseline Value For The Model

Since this is a typical multi-classification problem, we can choose to use many kinds of algorithms, including naive Bayes, KNN, decision tree and random forest.

Model Optimization

Ideas Improvement