Urban Technology - Crime data analysis and prediction

5th December 2023

Amin Suaad

Goal of the project

- Analysis of San Francisco Crime dataset from kaggle. This dataset consists incidents from 1/1/2003 to 5/13/2015. Dataset: https://www.kaggle.com/competitions/sf-crime/data
- Visualizing and understanding top crime categories.
- Understanding the effect of date and time on crime category.
- Understanding the effect of location on crime category.
- Checking the correctness of the dataset.
- Predicting severity of the crime using different Machine Learning methods.
- Comparing the performances of different Machine Learning methods.

Data fields

- **Dates** - timestamp of the crime event
- **Category** - type of the crime event
- **Descript** - description in details of the crime event
- **DayOfWeek** - the day of the week
- **PdDistrict** - name of the Police Department District. Example: NORTHERN, SOUTHERN
- **Resolution** - the solution after the crime event. Example: ARREST, BOOKED
- **Address** - the approximate street address of the crime event
- **X** - Longitude of the crime event
- **Y** - Latitude of the crime event

Dataset

	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
...	...	...	...	...	...	...	...	...	...
878044	2003-01-06 00:15:00	ROBBERY	ROBBERY ON THE STREET WITH A GUN	Monday	TARAVAL	NONE	FARALLONES ST / CAPITOL AV	-122.459033	37.714056
878045	2003-01-06 00:01:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Monday	INGLESIDE	NONE	600 Block of EDNA ST	-122.447364	37.731948
878046	2003-01-06 00:01:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Monday	SOUTHERN	NONE	5TH ST / FOLSOM ST	-122.403390	37.780266
878047	2003-01-06 00:01:00	VANDALISM	MALICIOUS MISCHIEF, VANDALISM OF VEHICLES	Monday	SOUTHERN	NONE	TOWNSEND ST / 2ND ST	-122.390531	37.780607
878048	2003-01-06 00:01:00	FORGERY/COUNTERFEITING	CHECKS, FORGERY (FELONY)	Monday	BAYVIEW	NONE	1800 Block of NEWCOMB AV	-122.394926	37.738212

878049 rows × 9 columns

Unique crime categories

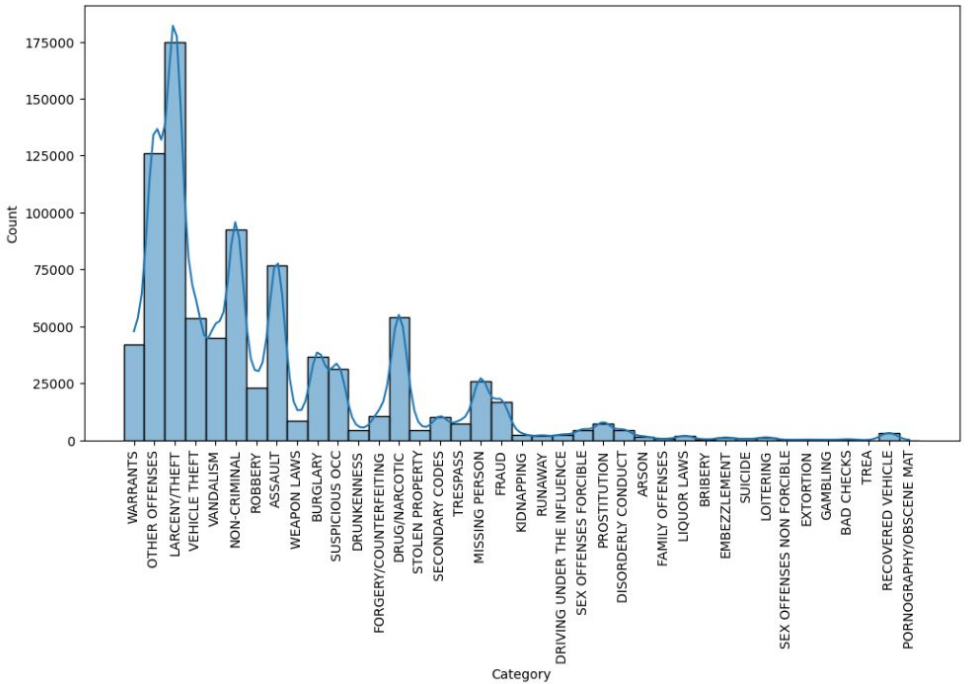
- There are 39 different crime categories.
- Some of them are severe like “ROBBERY” and others are less severe like “DRUNKENNESS”.
- Crime categories can be divided into less severe and more severe type depending on the severity of the crime.

```
Crime_Categories = df["Category"].unique()
Crime_Categories
```

```
array(['WARRANTS', 'OTHER OFFENSES', 'LARCENY/THEFT', 'VEHICLE THEFT',
       'VANDALISM', 'NON-CRIMINAL', 'ROBBERY', 'ASSAULT', 'WEAPON LAWS',
       'BURGLARY', 'SUSPICIOUS OCC', 'DRUNKENNESS',
       'FORGERY/COUNTERFEITING', 'DRUG/NARCOTIC', 'STOLEN PROPERTY',
       'SECONDARY CODES', 'TRESPASS', 'MISSING PERSON', 'FRAUD',
       'KIDNAPPING', 'RUNAWAY', 'DRIVING UNDER THE INFLUENCE',
       'SEX OFFENSES FORCIBLE', 'PROSTITUTION', 'DISORDERLY CONDUCT',
       'ARSON', 'FAMILY OFFENSES', 'LIQUOR LAWS', 'BRIBERY',
       'EMBEZZLEMENT', 'SUICIDE', 'LOITERING',
       'SEX OFFENSES NON FORCIBLE', 'EXTORTION', 'GAMBLING', 'BAD CHECKS',
       'TREA', 'RECOVERED VEHICLE', 'PORNOGRAPHY/OBSCENE MAT'],
      dtype=object)
```

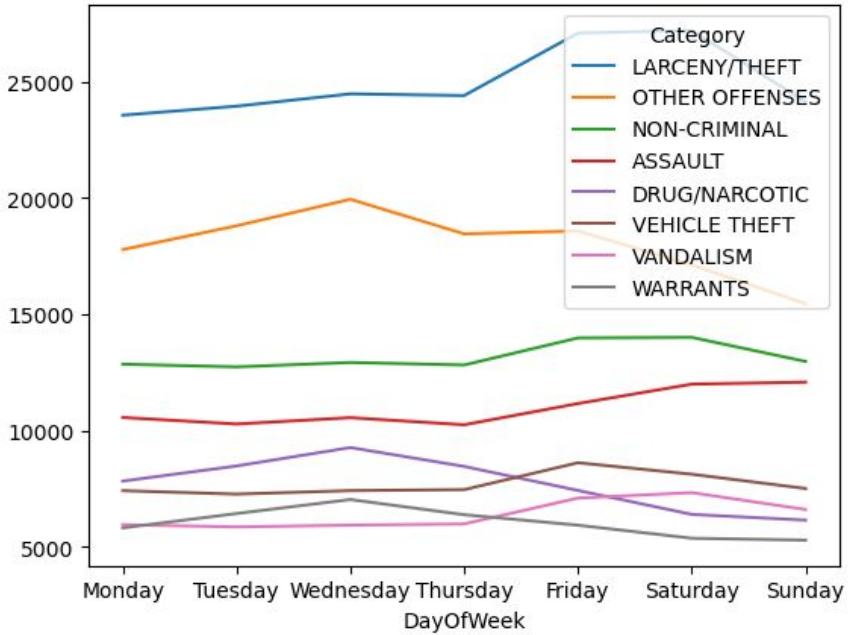
Most frequent crimes

- Approximately 20% (8 out of 39, which is 20%) of the categories add up to (nearly) 75% of the crimes.
- Top 8 categories are 'LARCENY/THEFT', 'OTHER OFFENSES', 'NON-CRIMINAL', 'ASSAULT', 'DRUG/NARCOTIC', 'VEHICLE THEFT', 'VANDALISM', 'WARRANTS'

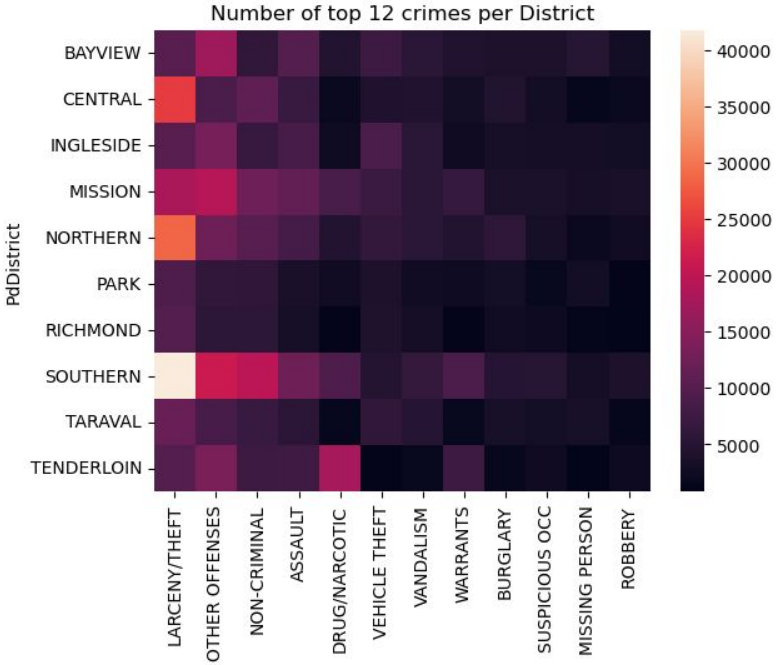


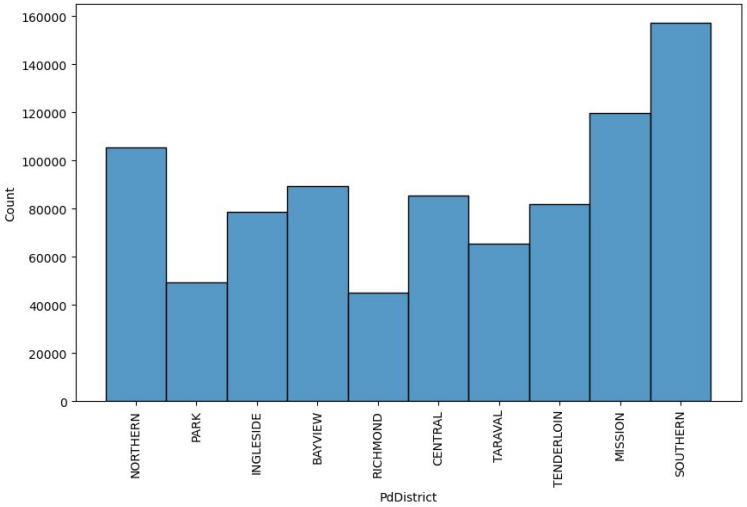
Weekdays and weekend

- Some of the crimes increases on Friday. Those are LARCENCY/THEFT, NON-CRIMINAL, VEHICLE THEFT, VANDALISM, ASSAULT
- Some wednesday crimes, which are OTHER OFFENSES, DRUG/NARCOTIC, WARRANTS
- Day of the week might be interesting in prediction task.



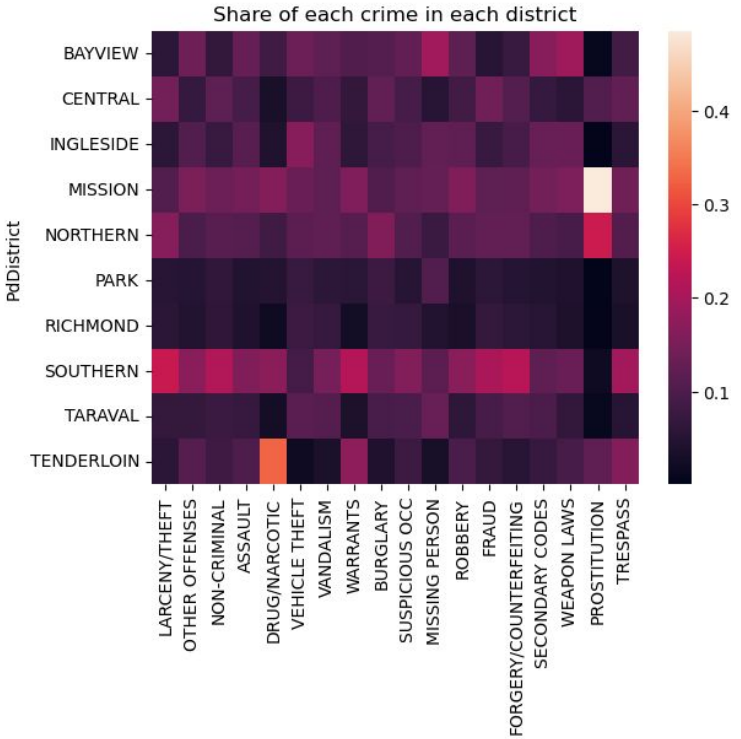
Number of crimes in each district





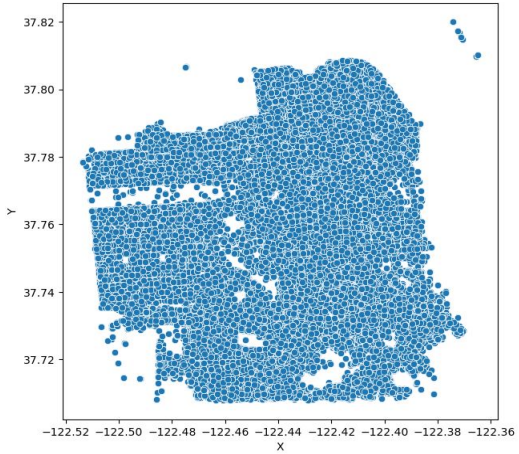
Share of each crime in each district

- Top 17 crimes are shown.
- Tenderloin district is problematic with drug/narcotic.
- Mission district has a large share of offences related to prostitution.
- Larceny/theft seems to be equally distributed among all except southern district.

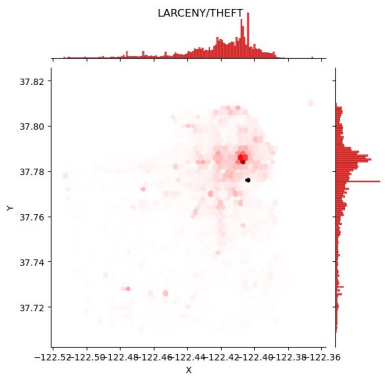


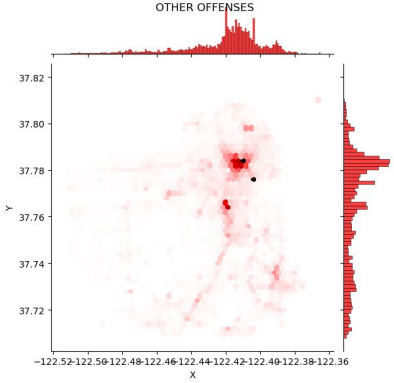
Longitude and latitude

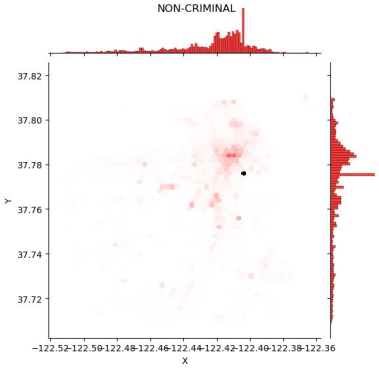
- The shape of the scatter plot represents the shape of the city.
- It is really important to check the correctness of longitude and latitude column because wrong data might be problematic for a prediction task.
- At this point, we can see similarity and with a lots of data the shape becomes more and more clear.

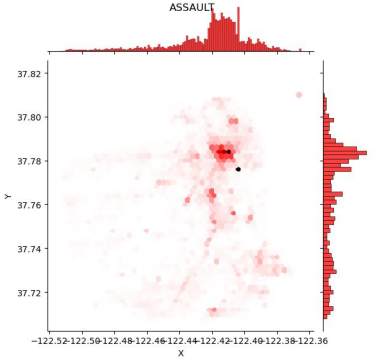


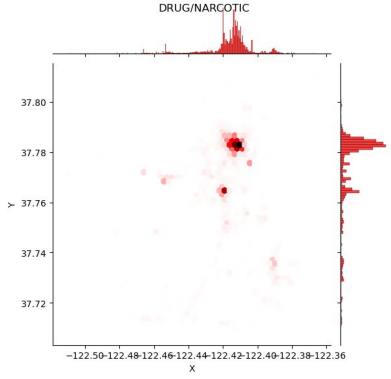
Hotspot of top 8 crimes

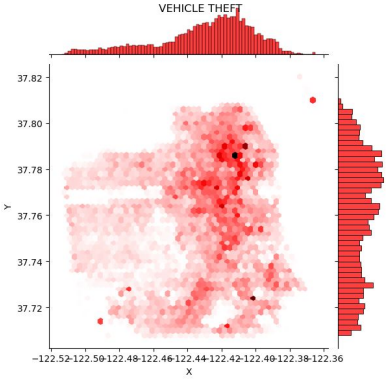


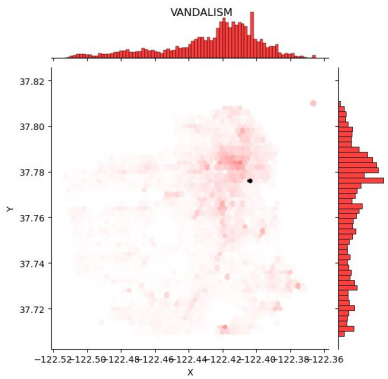


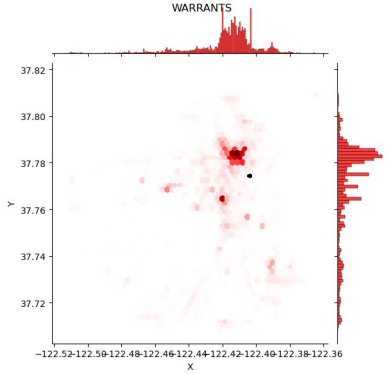












Predicting the severity of the crime

- Client wanted a prediction of severity of crime based on location and time.
- We defined what is severe and what is less severe. For example: Robbery goes to severe category and Prostitution goes to less severe category.
- Some wrong longitude and latitude in the dataset.
- 80% training and 20% test set.
- Features used: "DayOfWeek", "X", "Y", "PdDistrict", "Hour", "Month"
- Algorithms: KNN, Random Forest, ~~XGBoost~~
- GridSearchCV for hyperparameter tuning.
- SMOTE for class balancing.

Severe and less severe

Severe:

WARRANTS, LARCENY/THEFT, VEHICLE THEFT, VANDALISM, ROBBERY, ASSAULT, WEAPON LAWS, BURGLARY, FORGERY/COUNTERFEITING, DRUG/NARCOTIC, STOLEN PROPERTY, MISSING PERSON, FRAUD, KIDNAPPING, SEX OFFENSES FORCIBLE, ARSON, FAMILY OFFENSES, BRIBERY, SUICIDE, EXTORTION

Less severe:

OTHER OFFENSES, NON-CRIMINAL, SUSPICIOUS OCC, DRUNKENNESS, SECONDARY CODES, TRESPASS, RUNAWAY, DRIVING UNDER THE INFLUENCE, PROSTITUTION, DISORDERLY CONDUCT, LIQUOR LAWS, EMBEZZLEMENT, LOITERING, SEX OFFENSES NON FORCIBLE, GAMBLING, BAD CHECKS, TREA, RECOVERED VEHICLE, PORNOGRAPHY/OBSCENE MAT

KNN Classifier

- 1 represents severe class and 0 represents less severe class.
- Features: "DayOfWeek", "X", "Y", "PdDistrict"
- n_neighbors=9
- Training accuracy: 0.69
- Test accuracy: 0.65
- Recall : 0.65
- Precision : 0.61
- F1 Score : 0.61
- Confusion Matrix: [[15324 43713]

  [18590 97983]]

	precision	recall	f1-score	support
0	0.45	0.26	0.33	59037
1	0.69	0.84	0.76	116573
accuracy			0.65	175610
macro avg	0.57	0.55	0.54	175610
weighted avg	0.61	0.65	0.61	175610

Random Forest Classifier

- 1 represents severe class and 0 represents less severe class.
- Features: "DayOfWeek", "X", "Y", "PdDistrict"
- n_estimators=150
- Training accuracy: 0.73
- Test accuracy: 0.65
- Recall : 0.65
- Precision : 0.62
- F1 Score : 0.62
- Confusion Matrix: [[14920 44117]

  [16903 99670]]

	precision	recall	f1-score	support
0	0.47	0.25	0.33	59037
1	0.69	0.86	0.77	116573
accuracy			0.65	175610
macro avg	0.58	0.55	0.55	175610
weighted avg	0.62	0.65	0.62	175610

SMOTE (Synthetic Minority Oversampling Technique)

- SMOTE is an oversampling technique which is used to solve class imbalance problem in the dataset.
- The main goal is to balance class distribution by replicating the minority class examples.
- SMOTE generates new data points by linear interpolation.
- $x' = x + rand(0, 1) * |x - xk|$, x is a data point from the minority class and xk is one of the k-nearest neighbors of x.

KNN Classifier (After SMOTE)

- 1 represents severe class and 0 represents less severe class.
- Features: "DayOfWeek", "X", "Y", "PdDistrict"
- n_neighbors=6
- Training accuracy: 0.66
- Test accuracy: 0.58
- Recall : 0.58
- Precision : 0.60
- F1 Score : 0.59
- Confusion Matrix: [[28185 30852]

  [43246 73327]]

	precision	recall	f1-score	support
0	0.39	0.48	0.43	59037
1	0.70	0.63	0.66	116573
accuracy			0.58	175610
macro avg	0.55	0.55	0.55	175610
weighted avg	0.60	0.58	0.59	175610

Random Forest Classifier (After SMOTE)

- 1 represents severe class and 0 represents less severe class.
- Features: "DayOfWeek", "X", "Y", "PdDistrict"
- n_estimators=100
- Training accuracy: 0.73
- Test accuracy: 0.59
- Recall : 0.59
- Precision : 0.62
- F1 Score : 0.60
- Confusion Matrix: [[30340 28697]

  [42582 73991]]

	precision	recall	f1-score	support
0	0.42	0.51	0.46	59037
1	0.72	0.63	0.67	116573
accuracy			0.59	175610
macro avg	0.57	0.57	0.57	175610
weighted avg	0.62	0.59	0.60	175610

Sum up

- SMOTE did not work as I expected. Without SMOTE the results are better.
- Without SMOTE, KNN and Random Forest performed almost similar.
- Some useful features could be interesting. For example: average education of the area, average age of the area, financial situation of the people living in the area of the crime incident.
- One further possibility: Scoring the category column depending on the severity of crime and fit a regression model.

Resources

- https://www.kaggle.com/competitions/sf-crime
- https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/
- https://www.kaggle.com/code/nitinvijay23/predict-the-crime-category-knn-logistic