# Predict Crime: A comparative study between Chicago and Seattle using spatial and temporal approach

## Methods and Initial Results

**Name: Yangyang Dai**

Research question:

How to predict crime using spatial and temporal approach in machine learning and how the results of this method differ between city Chicago and Seattle.

Data:

The first Chicago crime data set is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system (https://data.cityofchicago.org/Public- Safety/Crimes-2001-to-present/ijzp-q8t2).
The data is recorded from 2001 to present with 22 attributes including 'ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary Type', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Beat', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', 'Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Location'. There are in total 6.57M instances of crimes. The following table is a snapshot of the raw data.

Fig 1. Chicago crime data

| ID | Case Numb | Date | Block | IUCR | Primary Type | Description | Locati |
|---|---|---|---|---|---|---|---|
| 11267718 | JB201047 | 03/26/2018 11:55:00 PM | 066XX S WESTERN AVE | 0860 | THEFT | RETAIL THEFT | GAS ST. |
| 11267724 | JB201048 | 03/26/2018 11:45:00 PM | 007XX S WELLS ST | 1330 | CRIMINAL TRESPASS | TO LAND | CONST |
| 11267725 | JB201049 | 03/26/2018 11:45:00 PM | 117XX S THROOP ST | 0910 | MOTOR VEHICLE THEFT | AUTOMOBILE | STREET |
| 11267755 | JB201053 | 03/26/2018 11:40:00 PM | 064XX W IRVING PARK RD | 1330 | CRIMINAL TRESPASS | TO LAND | GROCE |
| 11267705 | JB201038 | 03/26/2018 11:35:00 PM | 068XX S ASHLAND AVE | 502P | OTHER OFFENSE | FALSE/STOLEN/ALTERED TRP | STREET |
| 11267737 | JB201040 | 03/26/2018 11:30:00 PM | 031XX S KEELER AVE | 0326 | ROBBERY | AGGRAVATED VEHICULAR HIJACKING | STREET |
| 11267784 | JB201062 | 03/26/2018 11:30:00 PM | 040XX W MAYPOLE AVE | 031A | ROBBERY | ARMED: HANDGUN | STREET |
| 11267753 | JB201044 | 03/26/2018 11:19:00 PM | 039XX W 63RD ST | 1812 | NARCOTICS | POSS: CANNABIS MORE THAN 30GMS | STREET |
| 11267813 | JB201046 | 03/26/2018 11:15:00 PM | 003XX W 109TH PL | 0486 | BATTERY | DOMESTIC BATTERY SIMPLE | RESIDE |
| 11268448 | JB202026 | 03/26/2018 11:15:00 PM | 007XX W 47TH ST | 1320 | CRIMINAL DAMAGE | TO VEHICLE | STREET |
| 11267714 | JB201023 | 03/26/2018 11:10:00 PM | 071XX S ASHLAND AVE | 0320 | ROBBERY | STRONGARM - NO WEAPON | STREET |
| 11267792 | JB201032 | 03/26/2018 11:08:00 PM | 077XX S CORNELL AVE | 1310 | CRIMINAL DAMAGE | TO PROPERTY | RESIDE |
| 11267701 | JB201028 | 03/26/2018 11:05:00 PM | 033XX W WARNER AVE | 1320 | CRIMINAL DAMAGE | TO VEHICLE | STREET |
| 11268384 | JB201839 | 03/26/2018 11:05:00 PM | 056XX S WABASH AVE | 0620 | BURGLARY | UNLAWFUL ENTRY | APARTM |

The second Seattle crime dataset is obtained from all the Police responses to 9-1-1from Seattle police department. (https://data.seattle.gov/Public- Safety/Seattle-Police Department- 911-Incident-Response/3k2p- 39jp?category=Public Safety&view_name=Seattle-Police-Department-911-Incident-Response) The data ranges from 2010 to 2017 and contains 19 attributes, 1.47M instances in total. The variables include 'CAD CDW ID', 'CAD Event Number',

'General Offense Number', 'Event Clearance Code', 'Event Clearance Description', 'Event Clearance SubGroup', 'Event Clearance Group', 'Event Clearance Date', 'Hundred Block Location', 'District/Sector', 'Zone/Beat', 'Census Tract', 'Longitude', 'Latitude', 'Incident Location', 'Initial Type Description', 'Initial Type Subgroup', 'Initial Type Group', 'At Scene Time'. The following table is a snapshot of the raw data.

Fig 1. Seattle crime data

| Hundred Block Location | District/Sec | Zone/B | Census Tr | Longitude | Latitude | Incident Location |
|---|---|---|---|---|---|---|
| 3 AV S / S WASHINGTON ST | K | K3 | 9200.2014 | -122.330271593 | 47.600875809 | (47.600875809°, -122.330271593°) |
| 20XX BLOCK OF 15 AV W | Q | Q1 | 5802.2003 | -122.37613941 | 47.636336049 | (47.636336049°, -122.37613941°) |
| 6 AV / YESLER WY | K | K3 | 9200.1002 | -122.326350868 | 47.601708802 | (47.601708802°, -122.326350868°) |
| 86XX BLOCK OF 24 AV SW | F | F2 | 11401.2005 | -122.363172642 | 47.525585666 | (47.525585666°, -122.363172642°) |
| 135XX BLOCK OF 23 AV NE | L | L1 | 200.6017 | -122.304248161 | 47.727498035 | (47.727498035°, -122.304248161°) |
| 63XX BLOCK OF 29 AV SW | F | F1 | 10700.4001 | -122.369833395 | 47.546493546 | (47.546493546°, -122.369833395°) |
| MARTIN LUTHER KING JR WY S / S GENESEE ST | R | R2 | 10001.3006 | -122.295370641 | 47.563805602 | (47.563805602°, -122.295370641°) |
| CALIFORNIA AV SW / SW ALASKA ST | W | W2 | 10500.4003 | -122.386778535 | 47.561104368 | (47.561104368°, -122.386778535°) |
| 24XX BLOCK OF AURORA AV N | Q | Q2 | 6000.2045 | -122.346642846 | 47.641419741 | (47.641419741°, -122.346642846°) |
| 43XX BLOCK OF S FERDINAND ST | R | R3 | 10300.1002 | -122.278843236 | 47.558197565 | (47.558197565°, -122.278843236°) |
| 68XX BLOCK OF 30 AV NE | U | U3 | 3800.1004 | -122.29516869 | 47.678505415 | (47.678505415°, -122.29516869°) |
| 1XX BLOCK OF NW 81 ST | J | J2 | 2900.1014 | -122.359311741 | 47.687664142 | (47.687664142°, -122.359311741°) |
| 26XX BLOCK OF S DEARBORN ST | G | G3 | 8900.4011 | -122.298174668 | 47.595534943 | (47.595534943°, -122.298174668°) |
| 30XX BLOCK OF E UNION ST | C | C3 | 8800.1000 | -122.293157653 | 47.612937729 | (47.612937729°, -122.293157653°) |

## Models:

Random forests:
This is an ensemble learning method for classification, regression and other tasks. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. Random decision forests tend to correct for decision trees' habit of overfitting to their training set.

Decision Tree with KNN filter:
In Order to achieve higher accuracy, a combination of two classifiers are used with decision tree model being created every time as data is entered KNN being implemented when the input is requested to be tested.

To be specific, A decision tree is a supervised learning algorithm that builds regression models in the form of a tree structure. It splits a dataset into smaller and smaller subsets and thus simultaneously develops associated decision trees incrementally. The final result is a tree with decision nodes and leaf nodes. While KNN does not use available training data to establish any general rules and it is non-parametric.

## Variables:

*Chicago Crime Data:*

The original dataset with 22 variables and their data types are shown in the diagram below.

Fig. Chicago crime data variables table

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6591364 entries, 0 to 6591363
Data columns (total 22 columns):
ID                      int64
Case Number             object
Date                    object
Block                   object
IUCR                    object
Primary Type            object
Description             object
Location Description    object
Arrest                  bool
Domestic                bool
Beat                    int64
District                float64
Ward                    float64
Community Area          float64
FBI Code                object
X Coordinate            float64
Y Coordinate            float64
Year                    int64
Updated On              object
Latitude                float64
Longitude               float64
Location                object
dtypes: bool(2), float64(7), int64(3), object(10)
memory usage: 1018.3+ MB
```

Among 22 attributes, I picked variables ID, DATE, YEAR, DISTRICT, AND PRIMARY TYPE that are highly relevant to the center of research interests and transformed the dataset to the new data frame as shown below for model building purposes.

Fig. Chicago crime reduced variables

| | ID | Year | District | month | day | time | tod | t | Primary Type |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 10000092.0 | 2015.0 | 11.0 | 3.0 | 18.0 | 07:44:00 | PM | T7 | BATTERY |
| **1** | 10000094.0 | 2015.0 | 7.0 | 3.0 | 18.0 | 11:00:00 | PM | T8 | OTHER OFFENSE |
| **2** | 10000095.0 | 2015.0 | 2.0 | 3.0 | 18.0 | 10:45:00 | PM | T8 | BATTERY |
| **3** | 10000096.0 | 2015.0 | 2.0 | 3.0 | 18.0 | 10:30:00 | PM | T8 | BATTERY |
| **4** | 10000097.0 | 2015.0 | 11.0 | 3.0 | 18.0 | 09:00:00 | PM | T8 | ROBBERY |

*Seattle Dataset:*
The original dataset with 19 variables and their data types are shown in the diagram below.

Fig. Seattle crime data variables

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1475610 entries, 0 to 1475609
Data columns (total 19 columns):
CAD CDW ID                        1475610 non-null object
CAD Event Number                  1475610 non-null int64
General Offense Number            1475610 non-null int64
Event Clearance Code              1467647 non-null float64
Event Clearance Description       1467646 non-null object
Event Clearance SubGroup          1467646 non-null object
Event Clearance Group             1467646 non-null object
Event Clearance Date              1467541 non-null object
Hundred Block Location            1472123 non-null object
District/Sector                   1474545 non-null object
Zone/Beat                         1475609 non-null object
Census Tract                      1472791 non-null float64
Longitude                         1475609 non-null float64
Latitude                          1475609 non-null float64
Incident Location                 1475609 non-null object
Initial Type Description          897797 non-null object
Initial Type Subgroup             897797 non-null object
Initial Type Group                897797 non-null object
At Scene Time                     434811 non-null object
dtypes: float64(4), int64(2), object(13)
memory usage: 213.9+ MB
```

Among 22 attributes, I picked variables CAD CDW ID, DISTRICT/SECTOR, EVENT CLEARANCE GROUP, DATE that are highly relevant to the center of research interests and transformed the dataset to the new data frame as shown below for model building purposes.

Fig. Seattle crime reduced variables

| | CAD CDW ID | District/Sector | Event Clearance Group | month | day | time | tod | t | year |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 15736 | M | DISTURBANCES | 7 | 17 | 08:49:00 | PM | T7 | 2010 |
| **1** | 15737 | Q | OTHER PROPERTY | 7 | 17 | 08:50:00 | PM | T7 | 2010 |
| **2** | 15738 | M | NUISANCE, MISCHIEF | 7 | 17 | 08:55:00 | PM | T7 | 2010 |
| **3** | 15739 | D | TRAFFIC RELATED CALLS | 7 | 17 | 09:00:00 | PM | T8 | 2010 |
| **4** | 15740 | D | NUISANCE, MISCHIEF | 7 | 17 | 09:00:00 | PM | T8 | 2010 |

Methods:

The methods I use would be the same for the two data sets for the two cities.

*Exploratory analysis:*

The preliminary analysis includes visualizing relations between different variables and crime incidents. For example, the relation between year and crime rates, month and crime rates, day and crime rates, and crime types varieties.

*Data preprocessing:*

Based on the original dataset, I extracted the variables that are highly relevant to the center of research interests as shown in the above section. The steps I did to preprocess the raw data include:

- Dropped rows with Null values that appeared in the variables of interest
- Extracted Year, Month, Day, and Time from time series data
- Transformed time data from 12am to 12pm to eight categories of time slot (represented as 't' column in the new data frame)
    - T1: 12am-3am
    - T2: 3am-6am
    - T3: 6am-9am
    - T4: 9am-12pm
    - T5:12pm-3pm
    - T6: 3pm-6pm
    - T7: 6pm-9pm
    - T8: 9pm-12am
- Factorize Crime Type and Time Slot so that these two data can be used in random forests and decision trees

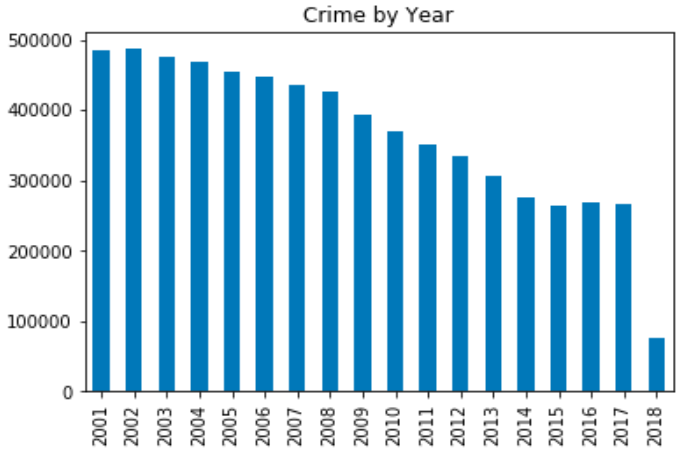
*Model building:*

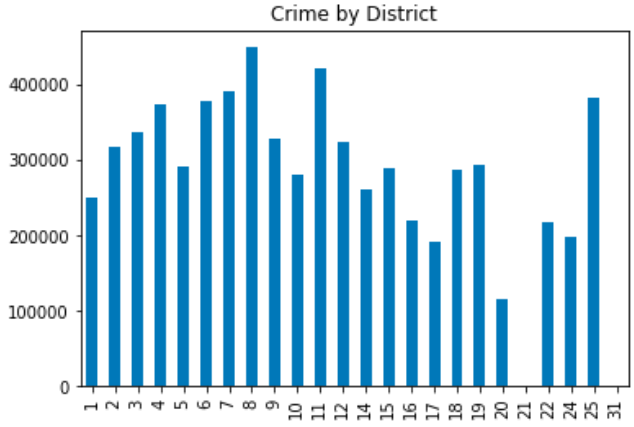
- Split train and test data sets for two cities
- Build random forest classification models
- Test the models and find accuracy scores

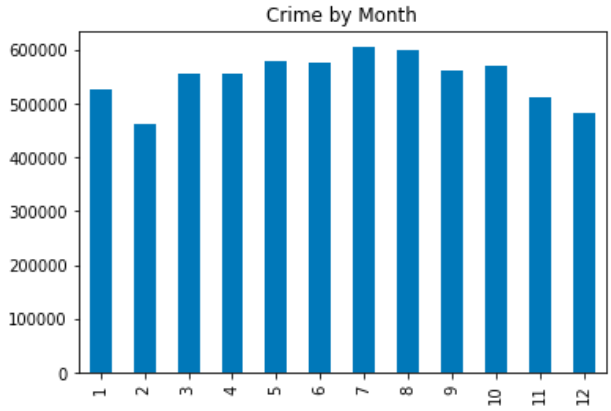
## Initial results:

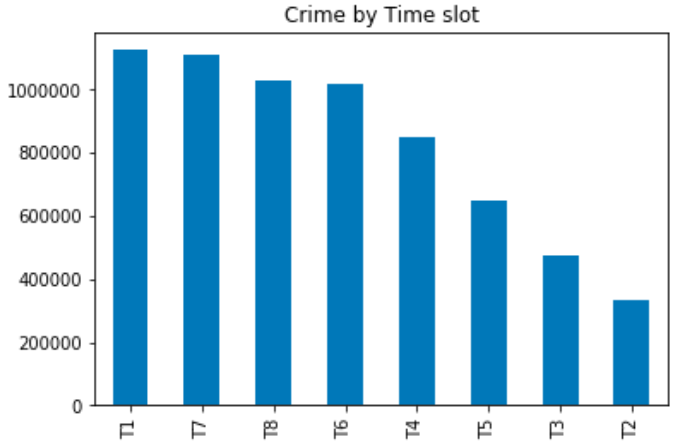
### *Exploratory analysis*

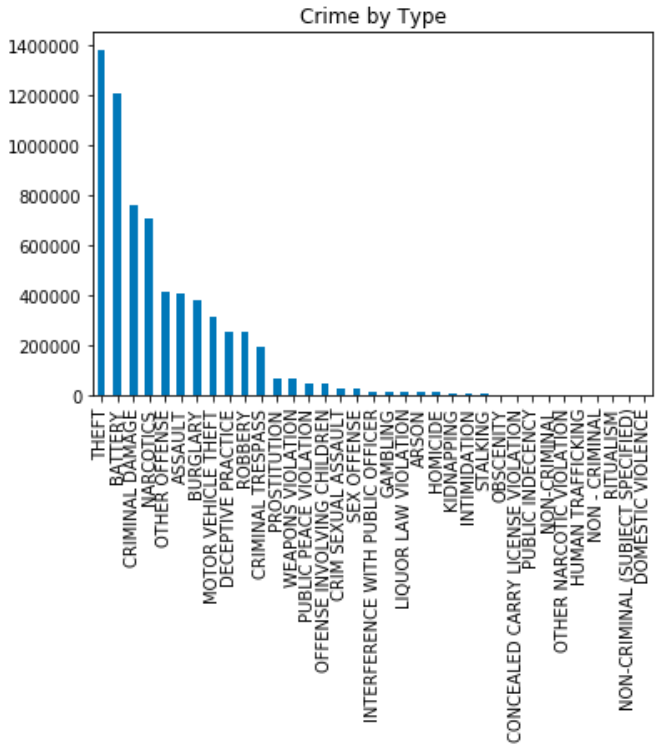
*Chicago crime:*


Crime by Year
500000
400000
300000
200000
100000
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018


Crime by District
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
11
12
14
15
16
17
18
19
20
21
22
24
25
31


Crime by Month
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
11
12

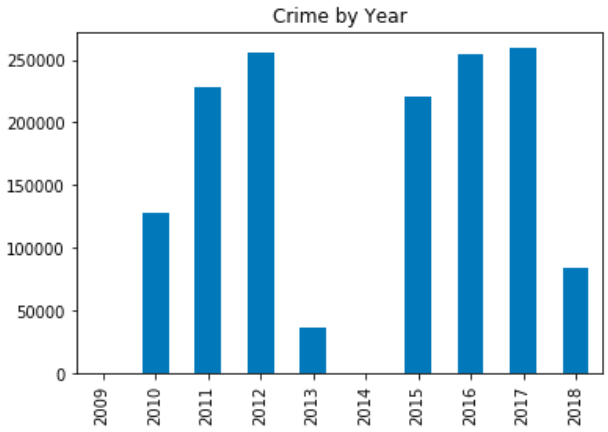

Crime by Time slot
1000000
800000
600000
400000
200000
0
T1
T7
T8
T6
T4
T5
T3
T2

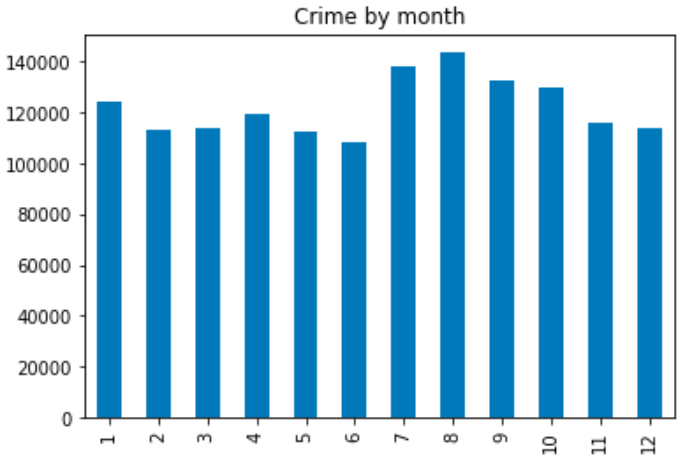

Crime by Type
1400000
1200000
1000000
800000
600000
400000
200000
0
THEFT
BATTERY
CRIMINAL DAMAGE
NARCOTICS
OTHER OFFENSE
ASSAULT
BURGLARY
MOTOR VEHICLE THEFT
DECEPTIVE PRACTICE
ROBBERY
CRIMINAL TRESPASS
PROSTITUTION
WEAPONS VIOLATION
PUBLIC PEACE VIOLATION
OFFENSE INVOLVING CHILDREN
CRIM SEXUAL ASSAULT
SEX OFFENSE
INTERFERENCE WITH PUBLIC OFFICER
GAMBLING
LIQUOR LAW VIOLATION
ARSON
HOMICIDE
KIDNAPPING
INTIMIDATION
STALKING
OBSCENITY
CONCEALED CARRY LICENSE VIOLATION
PUBLIC INDECENCY
NON-CRIMINAL
OTHER NARCOTIC VIOLATION
HUMAN TRAFFICKING
NON - CRIMINAL
RITUALISM
NON-CRIMINAL (SUBJECT SPECIFIED)
DOMESTIC VIOLENCE

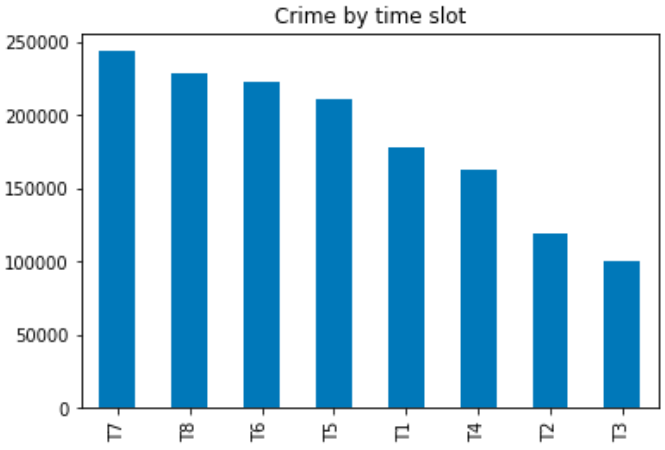
we can see from the above that:

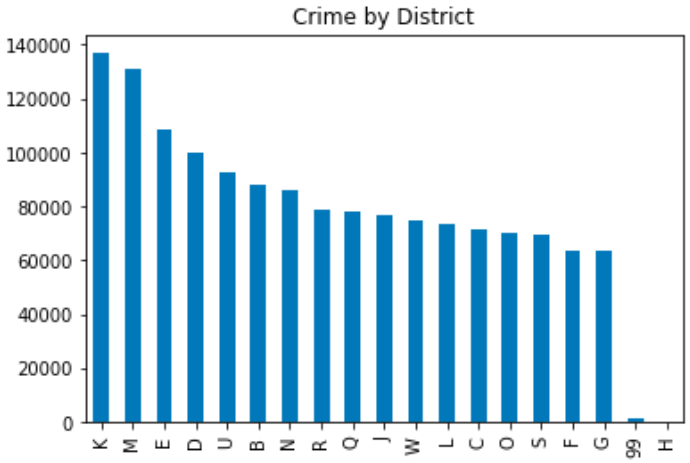
- The total number of crimes tend to decrease slowly over the years in Chicago. The data in 2018 is incomplete so the volume is relatively small.
- In Chicago, districts 8, 12, 25 seem to have more crimes than the other districts from the graph.
- Throughout the year, month August and September seem to have relatively higher crimes than other months while February, November and December tend to have lower crimes.
- From the time slot graph, crimes occur more often within T1 and T7, which indicates 12am-3am and 6pm-9pm are two time frames that tend to more have crimes.
- Theft and Battery are two most common crime types in Chicago.

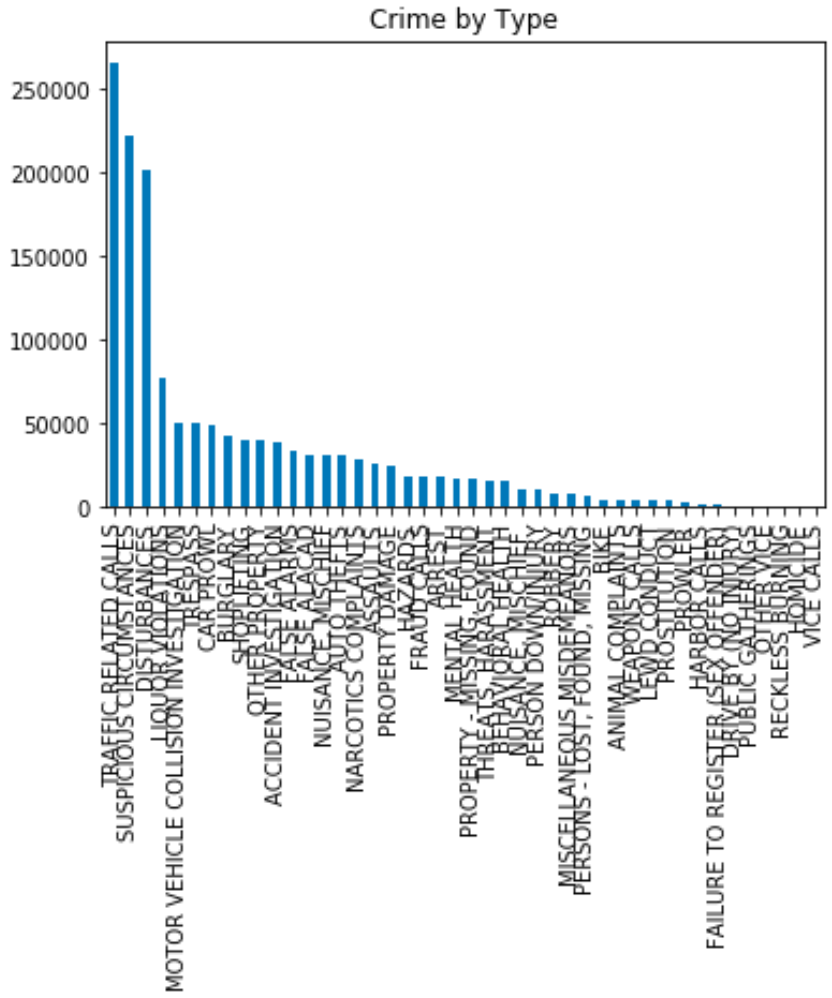
*Seattle crime:*






Compared to Chicago, we can see from the Seattle date set that:

- There is no particular pattern in crime counts distribution in Seattle over the years. And due to lack of data, the volumes of 2012 and 2018 are very small.
- Districts K and M seem to have more crimes than the other districts from the graph.
- Throughout the year, month August and July seem to have relatively higher crimes than other months while February and December tend to have lower crimes.
- From the time slot graph, crimes occur more often within T8 and T7, which indicates 9pm-12am and 6pm-9pm are two time frames that tend to more have crimes.
- Traffic related calls, suspicious circumstances, and disturbances are two most common crime types in Seattle.

*Models:*

The random forest model used to classify the Chicago crime data obtained the following results:

```
Train Accuracy ::  0.26939407655020875
Test Accuracy  ::  0.2543163790850615
Confusion matrix  [[     0      4   2081 ...      0   1365      0]
 [     0    197  66310 ...      0  68462      0]
 [     0    464 211241 ...      0 179450      0]
 ...
 [     0      1    461 ...      0    630      0]
 [     0    393 144535 ...      0 339216      0]
 [     0     36  12673 ...      0   8406      0]]
```

And the random forest model used to classify the Seattle crime data obtained the following results:

```
Train Accuracy ::  0.24107531501060928
Test Accuracy  ::  0.21843350181982457
Confusion matrix  [[511   0   2 ...   7   0   0]
 [ 13   0   0 ...   2   0   0]
 [ 70   0  22 ...  49   0   0]
 ...
 [ 63   0   0 ... 447   0   0]
 [  0   0   0 ...   0   0   0]
 [ 11   0   1 ...   8   0   0]]
```

The above is only initial model exploration, I expect the model accuracy would increase after a complete transformation of data and incorporation of both decision trees and knn clustering methods.