Analyzing Denver's Crime Data

By Neil Oza

Proposal

Denver has published its crime data for the past five years. The dataset is large, containing over 400,000 entries with nineteen columns. This is useful, as it's a large amount of data, but it is difficult to interpret in its current form. I will use some data analysis techniques to summarize the dataset and find any trends therein. In the process I will also check for and note any inconsistencies in the data so they can be corrected as appropriate.

Goals

- Inspect Denver's crime data and find any inconsistencies that exist within it
- Summarize the data by using clear and effective visualizations
- Identify trends in the data that help Denver police understand and address crime in their city more effectively

Data Collection and Cleaning

First Impressions of the Data

To start, I downloaded Denver's Crime Data as a csv from the city website. I read it into Jupyter Notebook using pandas under the name denver_data. I looked at the head of the dataset and took a look at the columns. Several of the columns confused me, so I reached out to Christine Whyckoff, the data's maintainer, who clarified some gray areas. For my and future reference, I created short summaries of each of the columns reproduced below:

OFFENSE_ID is a unique identifier for each offense. It is generated by concatenating (combining) the Incident ID, Offense Code, and Offense Code Extension. It provides a unique identifier for each offense, whereas the incident id is duplicated if multiple offenses are committed at once, and

Offense Code and Offense Code Extension will repeat whenever a particular type of offense is committed

INCIDENT_ID is a unique identifier for an occurence of offenses. most Offense ID's have unique Incident ID's, but when a person commits multiple offenses at once, e.g. liquor possession and heroine possession, multiple Offense ID's will be generated from the Incident_ID

OFFENSE_CODE is a unique identifier for a particular type of offense. things such as criminal mischief, trespassing, larceny, etc. all have different Offense Code values to identify them

OFFENSE_CODE_EXTENSION are used to describe a subset of another type crime. for example criminal_mischief- motor vehicle and criminal_mischief-other have the same Offense Code but different extensions to differentiate them

OFFENSE_TYPE_ID provides the actual name for the offense in layman's terms. each combination of Offense Code and Offense Code Extension reference a unique crime. contents of this column include things such as theft-shoplift, criminal-trespassing, and threats-to-injure

OFFENSE_CATEGORY_ID provides a more general categorization for crimes. for example, theft-shoplift and theft-from-bldg are both forms of larceny

FIRST_OCCURENCE_DATE is the first possible date/time of the offense. If Last Occurrence Date is NaN, then the first occurrence date is the time of the crime

LAST_OCCURENCE_DATE is the last possible time the offense in question could have occurred. It will be NaN if the exact time of the offense is known and will be an actual time if only a range of possible times is known.

REPORTED_DATE is the time at which the offense was reported to the police

INCIDENT_ADDRESS provides the location of the offense. not all entries have a value for this column for privacy reasons

GEO_LON and **GEO_LAT** are the latitudes and longitudes of the location of the offense

GEO_X and **GEO_Y** are the state plane (city of Denver standard projection) for the offense location. functionally similar to GEO_LON and GEO_LAT

DISTRICT_ID is the district in charge of handling the offense

PRECINCT_ID is the precinct in charge of handling the offense

NEIGHBORHOOD_ID is the neighborhood the offense occurred in

IS_CRIME states whether the offense was a crime

IS_TRAFFIC states whether the offense was a traffic incident

Looking through these columns, They can broadly be placed into four categories: columns describing location, columns identifying the crime, columns describing the type of crime, and columns describing the time the crime occurred. Though there are many columns, the dataset primarily conveys these four pieces of information. This means several things. First off, several of the columns are interchangeable. For example, GEO_X and GEO_Y express the same information as GEO_LAT and GEO_LON, and the OFFENSE_TYPE_ID and (OFFENSE_CODE and OFFENSE_CODE_EXTENSION) represent the same information. Next, some columns are more specific/general versions of other columns, such as how precincts are more specific locations within districts or how OFFENSE_TYPE_ID are more specific descriptors of crimes than OFFENSE_CATEGORY_ID. The third thing to note is that since many of the columns convey the same information, just in different ways, there actually aren’t nineteen columns worth of data. there are only four, but expressed in multiple units of measurement. This limits our ability to use machine learning on the dataset; Machine learning requires a clear target variable and features with which to identify the target variable. Offense Category ID could work as the former, but there are simply too many offense categories to effectively identify the type of crime with only location and time data.

Given that machine learning is ill-suited for this dataset, I will focus on creating effective visualizations and using hypothesis testing to identify certain trends as appropriate. More specifically, I will attempt to uncover any trends in how crimes are distributed geographically and temporally.

Data Cleaning

Before I can analyze this data, I need to make sure the data isn't missing any values and isn't full of errors. To start, I ran denver_data.info() to take a look at the null values. Most columns of the dataset had identical numbers of non-null values, which was a good sign. After some investigating, I determined that several location values were omitted for privacy considerations, and LAST_OCCURENCE_DATE had frequent missing values because the exact time of offense was often known. All of the other columns were whole and complete, so I determined the dataset was not missing information.

Once I checked for null values, I started to check for inconsistencies within the data; as mentioned earlier, many of the columns represent the same bit of data, so I wanted to make sure those pieces didn't conflict. I identified three possible errors:

1. The Offense Codes and Extensions may not reference a unique type of crime
2. The Offense ID's may not be properly constructed from the Incident ID, Offense Code, and Offense Code Extension.
3. First Occurrence, Last Occurrence, and Reported Dates may not be in chronological order

I decided to check for all three of these possible errors in turn. To start, I took a look at the Offense Codes and their extension. Each Offense Code-Extension combo should identify 1 unique type of crime; for instance code 912 should always refer to Offense Type ID: homicide-negligent. If code 912 occasionally also meant public disorder, that would be a problem and would interfere with analysis. To check, I used a groupby using the OFFENSE_CODE, OFFENSE_CODE_EXTENSION, and OFFENSE_TYPE_ID columns in that order to see if any of the offense code-extensions referenced multiple types of crimes. I printed the output, and every Offense Code-Extension combo referenced exactly one type of crime. That means there were no errors here and the Offense codes and extensions were never ambiguous.

Next I wanted to make sure all of the Offense ID's were properly constructed. As I mentioned earlier, offense ID's are created by concatenating the Incident ID, the Offense Code, and the

Offense Code Extensions. I wanted to make sure this was always true, so I reconstructed the Offense ID’s from its component parts. I created a temporary variable and converted all of the above columns into strings. I normalized the Offense Codes to 4 digits and the extensions to 2 digits by adding zeroes in front as necessary, and then I concatenated the Incident ID to the Offense Code to the Offense Code Extension in that order. I called this new value Combined ID. I then used a filter to see if the Offense ID was ever different from the Combined ID. It never was, which means all of the Offense ID’s were properly constructed. No errors here either!

Finally, I moved on to checking for errors in the reported/occurrence dates columns. As a rule, The First Occurrence Date column should always precede Last Occurrence Date, which should always precede Reported Date. I wanted to make sure that the last occurrence date never came before the First Occurrence Date and that the Reported Date never came before either Occurrence Date. I started by checking if any of the Last occurrence Dates preceded first occurrence dates. I converted the contents of the columns to datetimes, and then subtracted the Last Occurrence Date by the First Occurrence Date and printed out any negative values. This yielded about forty results, and all of them had Last Occurrence Dates before First Occurrence Dates. The relevant Offense_ID’s are reproduced below.

2014388104570700
2014399282240400
2015539465240400
2016151736355000
2017122300240400
2016820361200600
2014596887109900
2014353737299902
201526840549900
2016415601299900
2017474579299900
201764699501600
2015633506353000

2014373704299901
2014596887131302
201594326239900
2016415601220400
2017115679299900
2017694591240400
2014656045240400
2014399282353200
201526819121200
2015633506355000
2017614233240400
2016415601220200
2017759160131302

2014353737299901
20149515342240400
2015752334570700
2016751799240400
201719458220401
2017474579299901
20148047026240400
2015752334240400
2016751799549900
201624902240400
2017479079549900
2016151736359900
2017479079410400

I repeated this process to determine if either occurrence date preceded the Reported Date. This also produced approximately forty results where the Reported Dates precede either the First Occurrence Date or Last Occurrence Date. The Offense ID’s of the relevant records are once again printed below:

2014388104570700
2014399282240400
2015539465240400
2016151736355000
2017122300240400
2016820361200600
2014596887109900
2014353737299902
201526840549900
2016415601299900
2017474579299900
201764699501600
2015633506353000
2014373704299901
2014596887131302
201594326239900
2016415601220400
2017115679299900
2017694591240400
2014656045240400
2014399282353200
201526819121200
2015633506355000
2017614233240400
2016415601220200
2017759160131302
2014353737299901
20149515342240400
2015752334570700
2016751799240400
201719458220401
2017474579299901
20148047026240400
2015752334240400
2016751799549900
201624902240400
2017479079549900
2016151736359900
2017479079410400

These entries likely contain errors, and should be reviewed for accuracy. A total of circa eighty mistakes in a dataset of over 400000 entries is forgivable, so overall the dataset was well maintained and is reliable.

Filtering the Data

Once I finished checking the data for accuracy, I filtered out parts of the data i didn’t need. For starters, the dataset included traffic incidents in addition to crimes. I wanted to focus on the latter, so I created a new dataframe called denver_crime by filtering out all of the traffic incidents from denver_data. I also removed all crimes that occured in 2019 since recent crime data was less reliable and so I would have five complete years worth of data. I also filtered out the entries that had erroneous time data; the dataset was large enough to stomach the loss, and I’d rather work with data that is 100% accurate. After completing my filtering, the data was ready for analysis.

Visualizations and Analysis

The Plan

As mentioned earlier, many of the columns in the dataset overlap. In order to keep my analysis concise, I decided to use a representative element for each type of data to represent the whole. I used Reported Date to represent the time of the crime since that's the time the crime becomes pertinent to the police. I used Offense Category ID to represent the type of crime because it had a total of 14 possible values; this meant it would visualize better in bar graphs or time series plots than Offense Type ID, which has over fifty possible values. For location, I using District ID since there were a total of 7 districts, which once again visualized well. The fourth piece of data, Offense ID's, were not useful for data visualization; the values are arbitrary and used for record keeping, so it did not lend itself to EDA. My goal was to analyze crime temporally and by type first to find any overarching crime trends in the city, later combining that with district data to see if particular crimes are concentrated in specific regions. I will use pandas dataframes to manipulate the data, pyplots for visualizations, and scipy.stats for hypothesis testing.

Visualizations

I started by creating a time series plot to see how frequency of crime changed in Denver over the past five years. There has been an upward trend in crime, but the rate of increase seems to have slowed in recent years. I Next used a groupby function to create bar plots representing the total crimes committed in each category type. This showed that most crimes in Denver were theft related, public disorder, drugs and alcohol related, or fit into the miscellaneous category of all other crimes. Violent crimes are uncommon in Denver, which is reassuring.

I then combined the above two visualizations and created a time series plot illustrating how the total number of each type of crime changed over the past five years. This revealed that there has been a significant increase in thefts from motor vehicles, auto theft, and in all other crimes. The increase in theft from motor vehicles was particularly sharp, suggesting that a PSA reminding people to lock their vehicles may be warranted. For the record, public disorder was the only crime type to decrease in prevalence over the past five years.

Next, I decided to mimic the above visualizations, but instead of checking how crime types varied in preponderance year over year, I checked how they varried month to month. I used another groupby and created both a line plot to show fluctuations in each type of crime over the months and a bar graph to illustrate the fluctuation in total crime through the months. The line plot of each type of crime over the months revealed no strong trends, but the bar graph suggested that crime in general is cyclical. crimes seem to become more common during the summer and less common during the winter. I went ahead and tested this hypothesis by performing a two sample t-test comparing data from the winter and summer months. The resulting t value was over 14, which means there is a statistically significant difference in number of crimes committed in the winter vs summer months, with more crimes occurring in the summer.

In a similar vein, I broke down crime frequencies by day of the week. My goal was to discover if crime rates are different between weekdays and weekends. I once again created line and bar plots to map the trends in individual types of crime and crime in total respectively. Looking at the data, It seemed that crimes are far more common on weekdays as opposed to weekends with the exception of forms of assault, which are in fact more common on weekends. I conducted a hypothesis test to check if crimes are more common on weekdays, and the test returned a t value of over 20. This means we can definitively say crimes are more common on weekends.

To complete my temporal analysis of crimes by category, I analyzed crime types by the hour. My goal was to discover what crimes were most prevalent at what time of the day. I discovered that in general crimes occur between 8 AM and 8 PM. However, different types of crimes seem to be prevalent at different times of the day. Violent crimes such as aggravated-assault, robbery, and other-crimes-against-persons seem to be most common late at night, betwen 10 PM and 2 AM. Auto-theft, theft-from-motor-vehicle, and larceny seem to be most common from 9 AM to 5 PM, which is likely because that’s when people are at work. The crimes of murder and arson are too infrequent to come up with a meaningful trend statement.

So far I had only looked at crimes by category within various time frames. At this point, I add in the district data to see if crime varies by location. I started by creating a multi-bar graph representing total crime in each of the 7 districts over the past five years. This revealed that district 6, 3, and 1 had the most crimes while district 7 had the least by a large margin. Most districts saw steady increase in crime with the exception of district 4 and possibly district 2.

I wanted to see if certain types of crimes were more common in certain districts. To do his I created a series of multiple bar graphs for each type of crime, showing how many of each occurred in each district over the past five years. I then compared those results with the total number of crimes in each district to see if any were disproportionate. Several were. Auto-theft and thefts from motor vehicles were far more common in district 3, drug crimes were far more common in district 6, and murders were oddly common in district 2. Crimes such as aggravated assault seemed evenly distributed.

Heatmap

In order to provide a more detailed visualization of actual crime locations, I created a heatmap creator function called heatmap_creator. This heatmap creator can create heatmaps of crime locations using multiple filters. You could generate a heatmap of crime locations in 2016, a heatmap of larceny crimes over all five years, or even a heatmap of burglaries in December of 2018.

To use the tool, you will need to open the notebook in Jupyter Notebook or some related software. Open the notebook and call the function heatmap_creator in an empty cell and pass in any number of two element tuples of the form (column_name, column_value). The function will filter based on the value and create the appropriate heatmap. For example, if you pass in the tuple (‘YEAR_REPORTED’, 2017), it would produce a heatmap of all crimes committed in 2017. Similarly, if you pass in multiple tuple parameters, such as (‘MONTH_REPORTED’,6), (‘YEAR_REPORTED’, 2016), (‘DISTRICT_ID’,4) It would show all crimes committed in District 4 during June of 2016. This tool should allow police to pinpoint hotspots of any type of crime in the city, in any district. I created a few example heatmaps in the notebook myself, but I daresay the Denver’s police departments can put it to more creative than me.

Conclusion

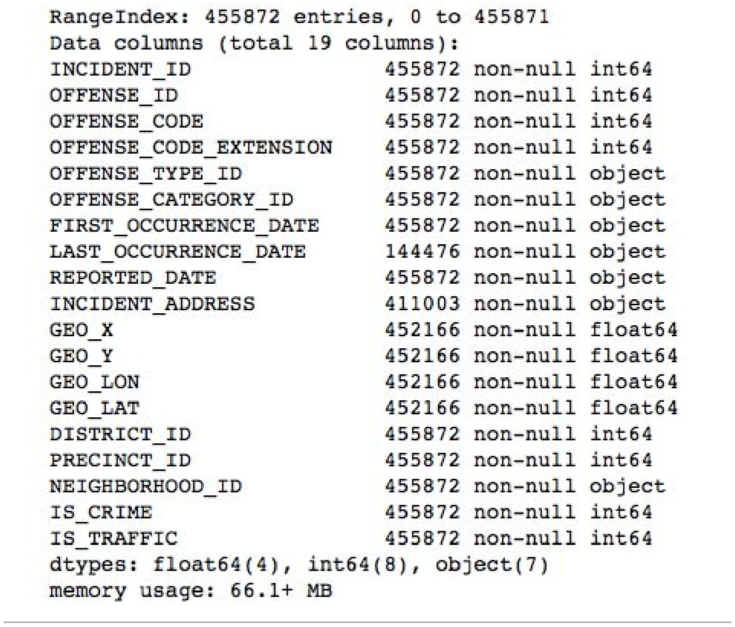
Denver’s Crime Data is a fairly clean dataset with only a handful of errors in Occurrence Date-Reported Date data listed above. Denver has experienced increase in crime rates over the past five years, though the rate of increase has slowed. Crimes are seasonal, with more crimes occurring in the summer as opposed to the winter. Crimes are also more common on weekdays than weekends, with the exception of assault related crimes. Different types of crimes are more common during different hours of the day with most forms of theft common during daylight hours while most forms of violent crime occur at night. Different districts seem to have different crime problems, with District 6 having a serious drug-alcohol problem while district 3 sports a large number of car related thefts. To provide additional visualizations, I created a heatmap generator that will allow Denver PD to generate heatmaps of different types of crimes using a variety of filters to specify time and place. I hope this EDA was useful , and I look forward to feedback and advice on how to improve the EDA. Thank you for reading.

Next Steps

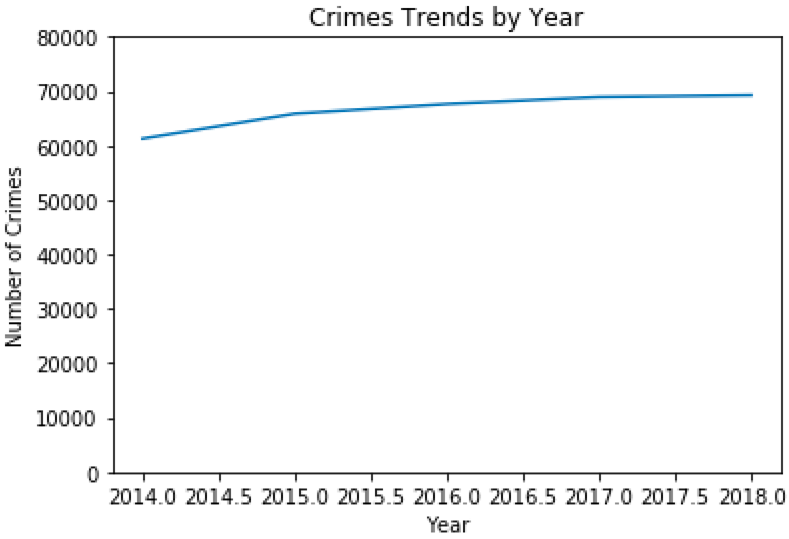
This analysis covered a lot of ground, but I believe more can be done with this data. I personally didn’t use the traffic data at all; I believe if this data were combined with Denver’s Traffic Count data, it could help the city identify problematic intersections to help reduce collisions. If combined with Denver’s Police spending data, it could be possible to determine the best ways to allocate resources to reduce crimes. If combined with data on criminals, you could determine which criminals are likely to commit multiple offenses. Finally, combining this data with housing prices could reveal how various types of crimes affect home values. The data is useful, and I daresay when combined with other datasets it will reveal more secrets.

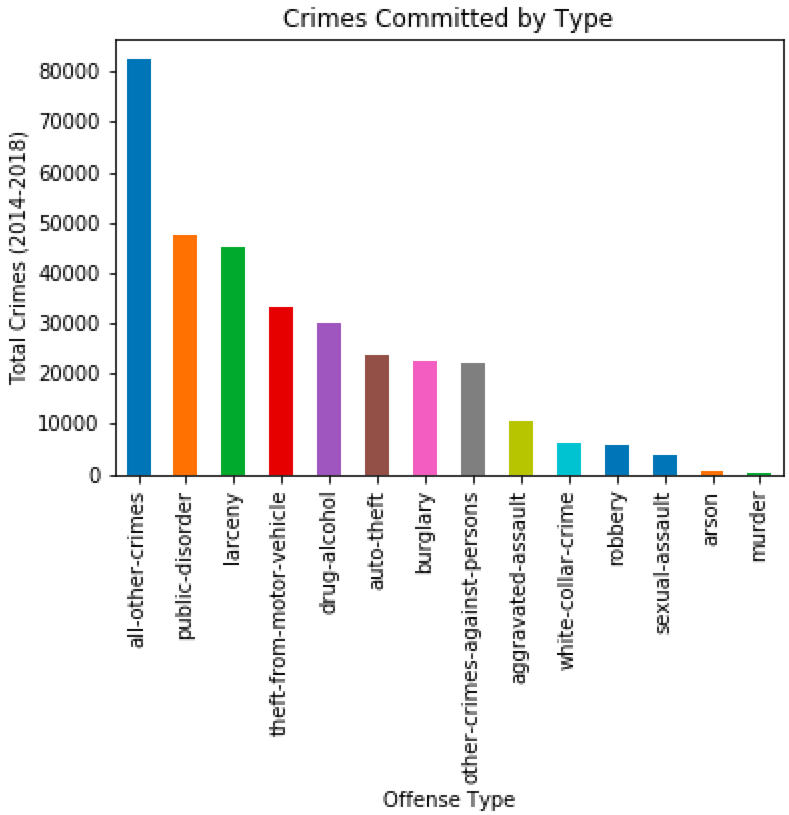
Appendix

Here I will print some of the important graphs created in this analysis. To see all visualizations, check out the full notebook at this link:
https://www.kaggle.com/neilb4yourking/analyzing-denver-s-crime-data

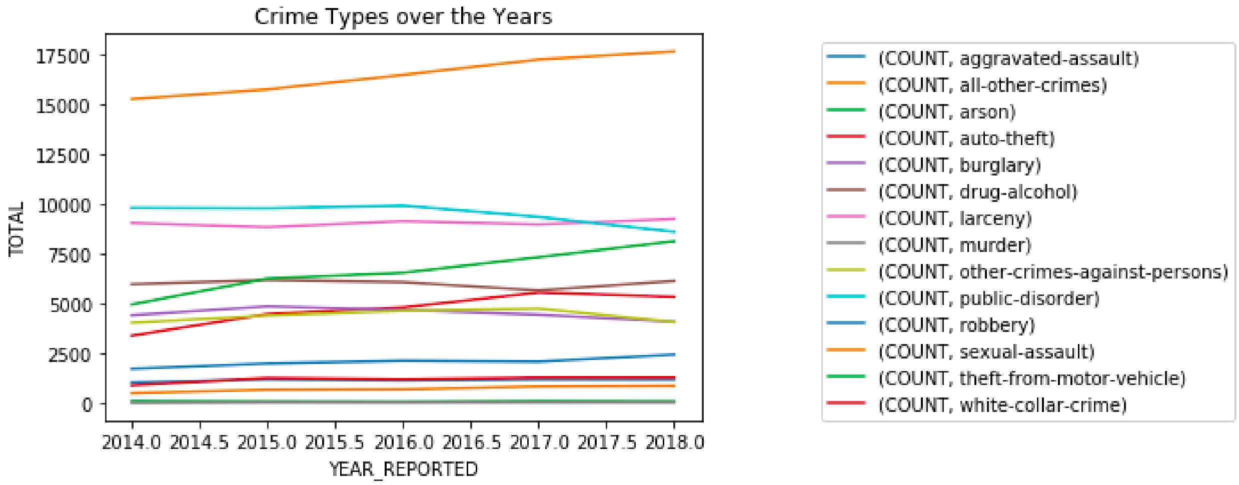


```
RangeIndex: 455872 entries, 0 to 455871
Data columns (total 19 columns):
INCIDENT_ID               455872 non-null int64
OFFENSE_ID                455872 non-null int64
OFFENSE_CODE              455872 non-null int64
OFFENSE_CODE_EXTENSION    455872 non-null int64
OFFENSE_TYPE_ID           455872 non-null object
OFFENSE_CATEGORY_ID       455872 non-null object
FIRST_OCCURRENCE_DATE     455872 non-null object
LAST_OCCURRENCE_DATE      144476 non-null object
REPORTED_DATE             455872 non-null object
INCIDENT_ADDRESS          411003 non-null object
GEO_X                     452166 non-null float64
GEO_Y                     452166 non-null float64
GEO_LON                   452166 non-null float64
GEO_LAT                   452166 non-null float64
DISTRICT_ID               455872 non-null int64
PRECINCT_ID               455872 non-null int64
NEIGHBORHOOD_ID           455872 non-null object
IS_CRIME                  455872 non-null int64
IS_TRAFFIC                455872 non-null int64
dtypes: float64(4), int64(8), object(7)
memory usage: 66.1+ MB
```

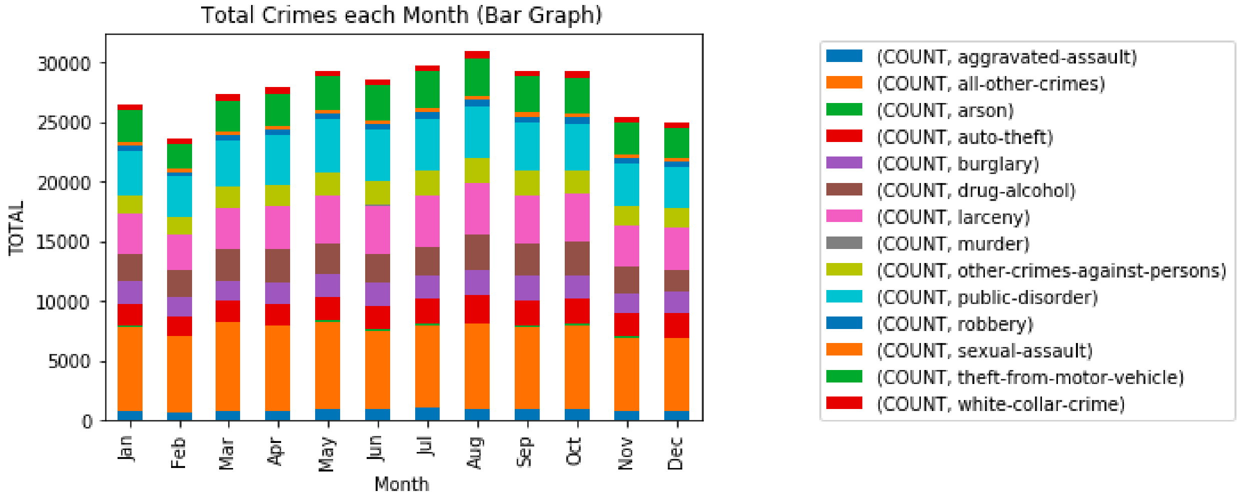


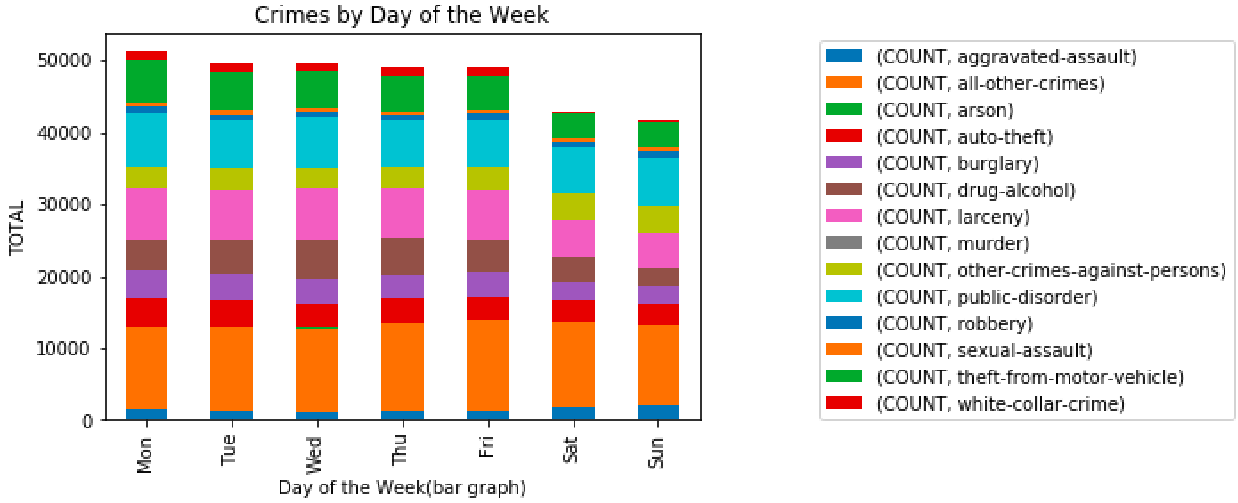


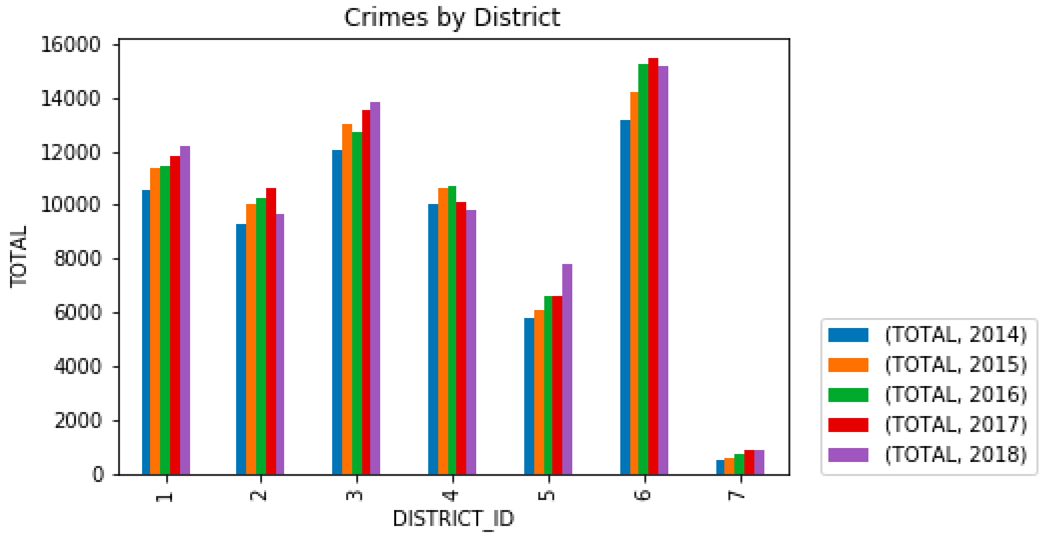
Crimes Committed by Type



Crime Types over the Years


Total Crimes each Month (Bar Graph)
TOTAL
0
5000
10000
15000
20000
25000
30000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
(COUNT, aggravated-assault)
(COUNT, all-other-crimes)
(COUNT, arson)
(COUNT, auto-theft)
(COUNT, burglary)
(COUNT, drug-alcohol)
(COUNT, larceny)
(COUNT, murder)
(COUNT, other-crimes-against-persons)
(COUNT, public-disorder)
(COUNT, robbery)
(COUNT, sexual-assault)
(COUNT, theft-from-motor-vehicle)
(COUNT, white-collar-crime)


Crimes by Day of the Week
TOTAL
0
10000
20000
30000
40000
50000
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day of the Week(bar graph)
(COUNT, aggravated-assault)
(COUNT, all-other-crimes)
(COUNT, arson)
(COUNT, auto-theft)
(COUNT, burglary)
(COUNT, drug-alcohol)
(COUNT, larceny)
(COUNT, murder)
(COUNT, other-crimes-against-persons)
(COUNT, public-disorder)
(COUNT, robbery)
(COUNT, sexual-assault)
(COUNT, theft-from-motor-vehicle)
(COUNT, white-collar-crime)


Crimes by District
TOTAL
0
2000
4000
6000
8000
10000
12000
14000
16000
1
2
3
4
5
6
7
DISTRICT_ID
(TOTAL, 2014)
(TOTAL, 2015)
(TOTAL, 2016)
(TOTAL, 2017)
(TOTAL, 2018)