Project Deliverable I

Nicholas Occhipinti
https://github.com/nick-occ/dsba-6520-project

Dataset & Business Use Case

For the project I have chosen a dataset from NYC Open Data that has both NYPD historic complaint data and NYPD year to date complaint data. These datasets include felony, misdemeanor and violation crimes that have been reported to the NYPD. Each record in this dataset is a complaint that has occurred between 2006 to 2021. This is a very large dataset containing over 7 million records. There are 35 features in the dataset containing information about the type of complaint/offense, the level of offense (felony, misdemeanor, etc), the police jurisdiction, the location of the complaint/offense such as latitude/longitude, transit district and station names if applicable and lastly there is information about both the victim and the suspect such as age groups, race and sex.

For the purpose of the analysis I will be filtering the data based on dates ranging between 2019 and the latest available data on record. This will reduce the dataset from 7.4 million complaints to under 1 million. These complaints occurred in 77 precincts within the 5 boroughs of New York City. The complaints in the datasets represent 65 categories of offenses that can further be broken down into 369 more specific classifications such as different levels of offenses.

The business use case is one of a crime analyst working for New York City that has been tasked to work on a project identifying patterns of recurring crime around the city. Since New York City has such a large population of people and a large area for police to patrol, it is important to spot trends of high crime locations to better allocate police resources. This historical data can help tell a story about how areas of New York City have changed over time and can analyze if these areas are seeing an increase or decrease of crime. The main goal of this project is using the results to develop actionable items that can help the NYPD make better decisions with the data. This can lead to increasing the police presence in different areas, better communication between NYPD and communities about crime patterns, and where to allocate funds towards technology such as security cameras.

Another use case from this project is studying how crime has changed during the COVID-19 pandemic. One reason I chose the timeline of the data from 2019 to 2021, is because it presents three different periods; pre-pandemic (2019), pandemic (2020) and vaccine/return to normal period (2021) By studying these time periods I can analyze if certain types of offenses are higher or lower than normal as a result of different factors such as places being closed and people being quarantined in their home.

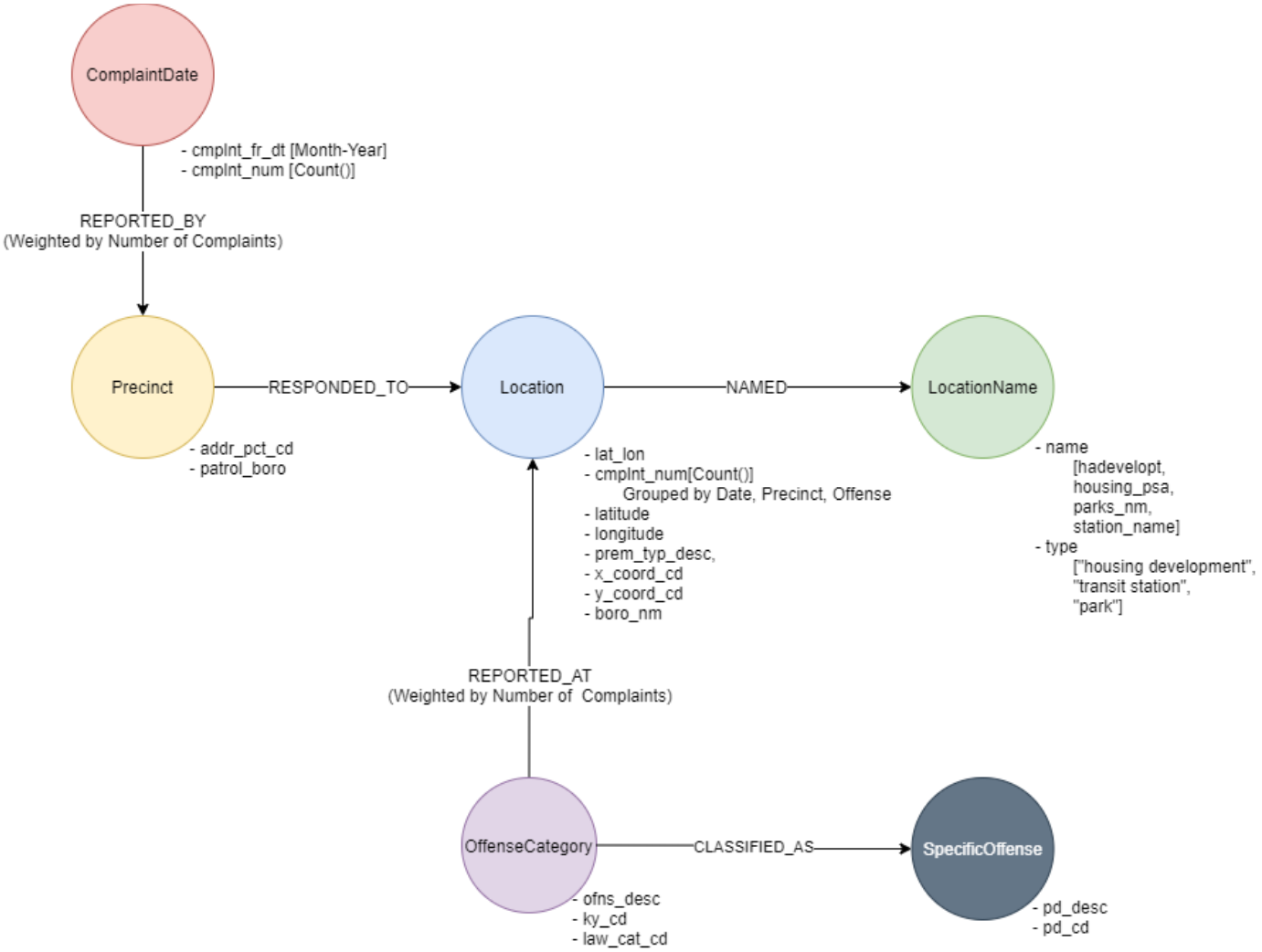
Graph Data Model

In this data model there are 6 node types, **ComplaintDate**, **Precinct**, **Location**, **LocationName**, **OffenseCategory** and **SpecificOffense**. The ComplaintDate node uses the "cmplnt_fr_dt" field which is when the complaint was first reported. The dates are a key to determining how complaints have changed over time. To avoid having too many nodes and capturing the data daily, I resampled it monthly. These nodes would then be created for every relevant Month-Year combination and the count of complaints are being stored as a property.

The Precinct node is the precinct where the complaint was reported and is represented by the "addr_pct_cd" field. The relationship between ComplaintDate and Precinct is weighted by the number of complaints that occurred in the precincts during that month and year. At a high level this can tell us if complaints are going up or down over time in the precincts.

Locations are where the complaints occur, and precincts respond to the complaints at these locations. The smallest level of granularity are latitude and longitude coordinates which represent the point of the middle of the block. In the future I will be researching if these coordinates can be aggregated at a higher level such as by neighborhood. Some locations may also have names that are represented in the LocationName node. This is where I am storing the name of the location and if the type of location is a housing development, transit station or a park.

Complaints are categorized into types of offenses in the "ky_cd" field. These offenses can be broken down further into more specific categories of offenses in the "pd_cd" field. For this reason I created them as two separate nodes, OffenseCategory and SpecificOffense. Based on the requirements of the analysis the user could just work with the OffenseCategory node to get the general offense name. However, if more information is needed, the SpecificOffense node offers that extra level of detail.


ComplaintDate
- cmplnt_fr_dt [Month-Year]
- cmplnt_num [Count()]
REPORTED_BY
(Weighted by Number of Complaints)
Precinct
- addr_pct_cd
- patrol_boro
RESPONDED_TO
Location
- lat_lon
- cmplnt_num[Count()]
Grouped by Date, Precinct, Offense
- latitude
- longitude
- prem_typ_desc,
- x_coord_cd
- y_coord_cd
- boro_nm
NAMED
LocationName
- name
[hadevelopt,
housing_psa,
parks_nm,
station_name]
- type
["housing development",
"transit station",
"park"]
REPORTED_AT
(Weighted by Number of Complaints)
OffenseCategory
- ofns_desc
- ky_cd
- law_cat_cd
CLASSIFIED_AS
SpecificOffense
- pd_desc
- pd_cd

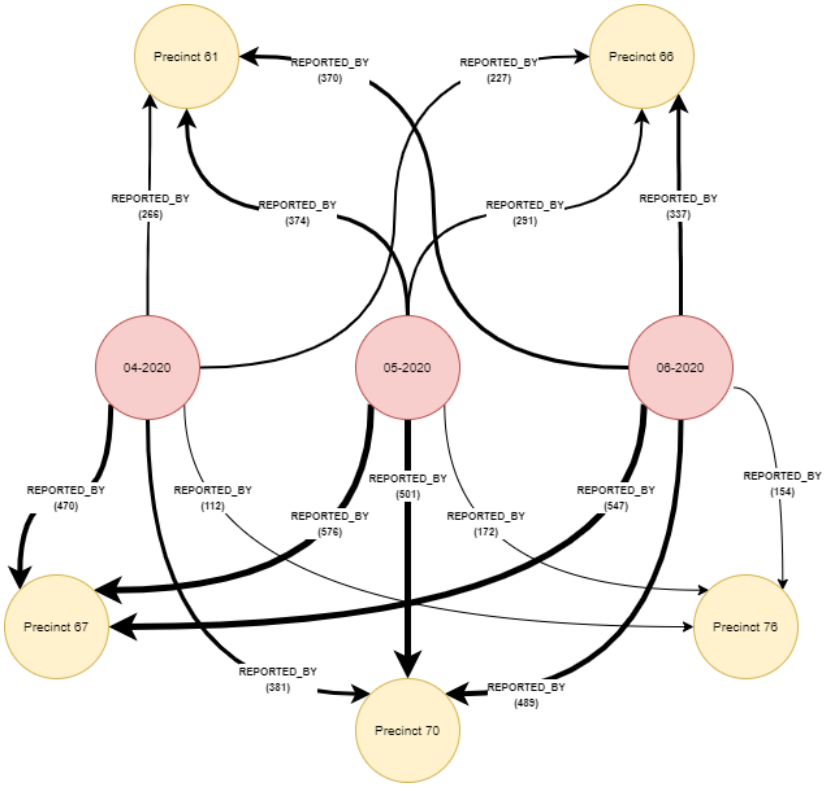
Graph Projections

Number of Precinct Complaints During a Time Period

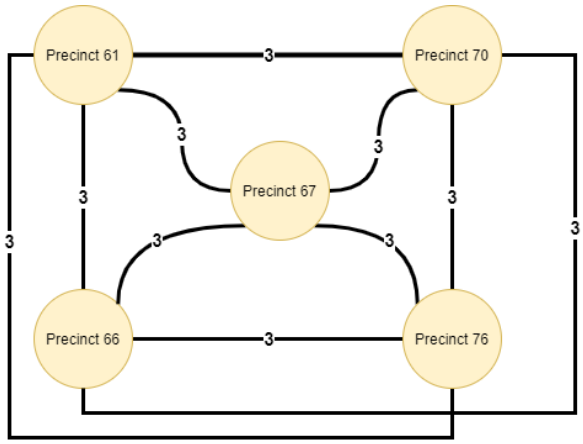
In this projection I am showing the number of complaints that occurred in sample precincts I have chosen during the months of April, May and June of 2020. The dates were resampled by month and the data is aggregated to count the number of complaints during that month by the precinct. The relationships are weighted and the number in the relationship represents the number of complaints.

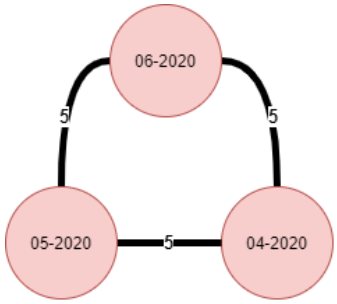
ADDR_PCT_CD	CMPLNT_DATE	
61	2020-04-30	266
	2020-05-31	374
	2020-06-30	370
66	2020-04-30	227
	2020-05-31	291
	2020-06-30	337
67	2020-04-30	470
	2020-05-31	576
	2020-06-30	547
70	2020-04-30	381
	2020-05-31	501
	2020-06-30	489
76	2020-04-30	112
	2020-05-31	172
	2020-06-30	154

Bipartite Projection



Monopartite Projections



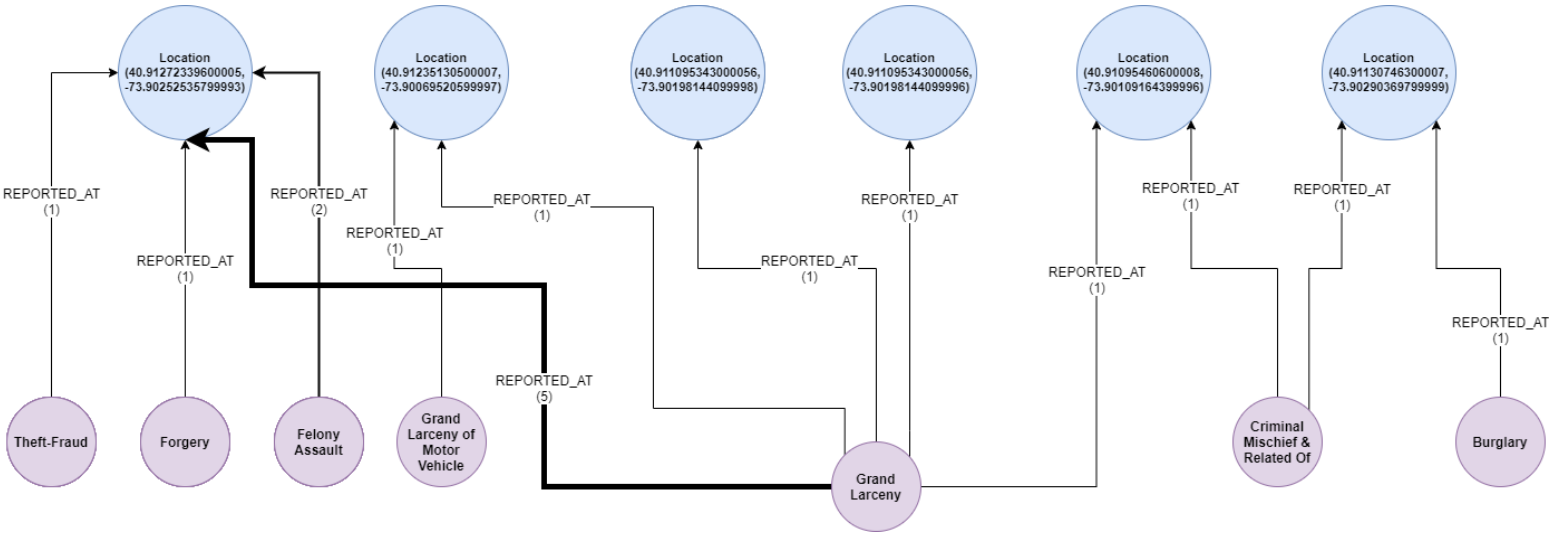


Relationship Between Offenses in the Same Location

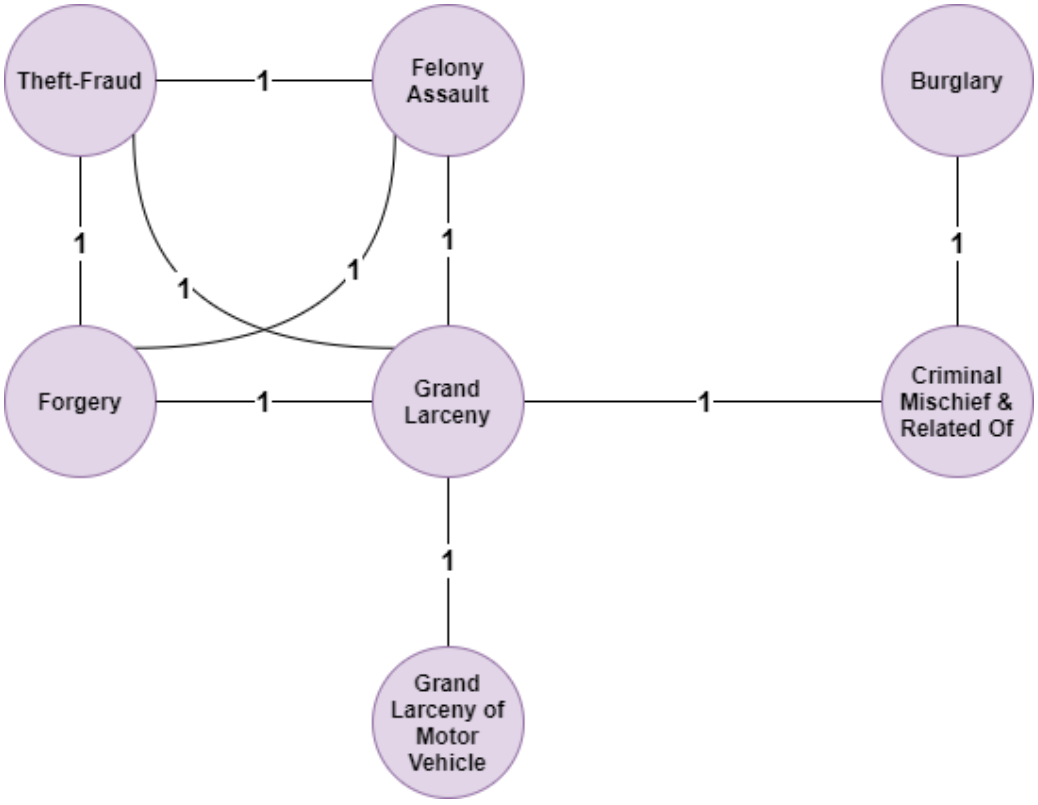
This projection shows the relationship of what offenses occur in the same location. This can also be used to show patterns of similar crimes that occur in nearby locations. In this example I am showing sample locations in the Bronx through the latitude and longitude combinations and what offenses have been linked to that location and the number of complaints. The relationships are weighted and the number in the relationship represents the number of complaints.

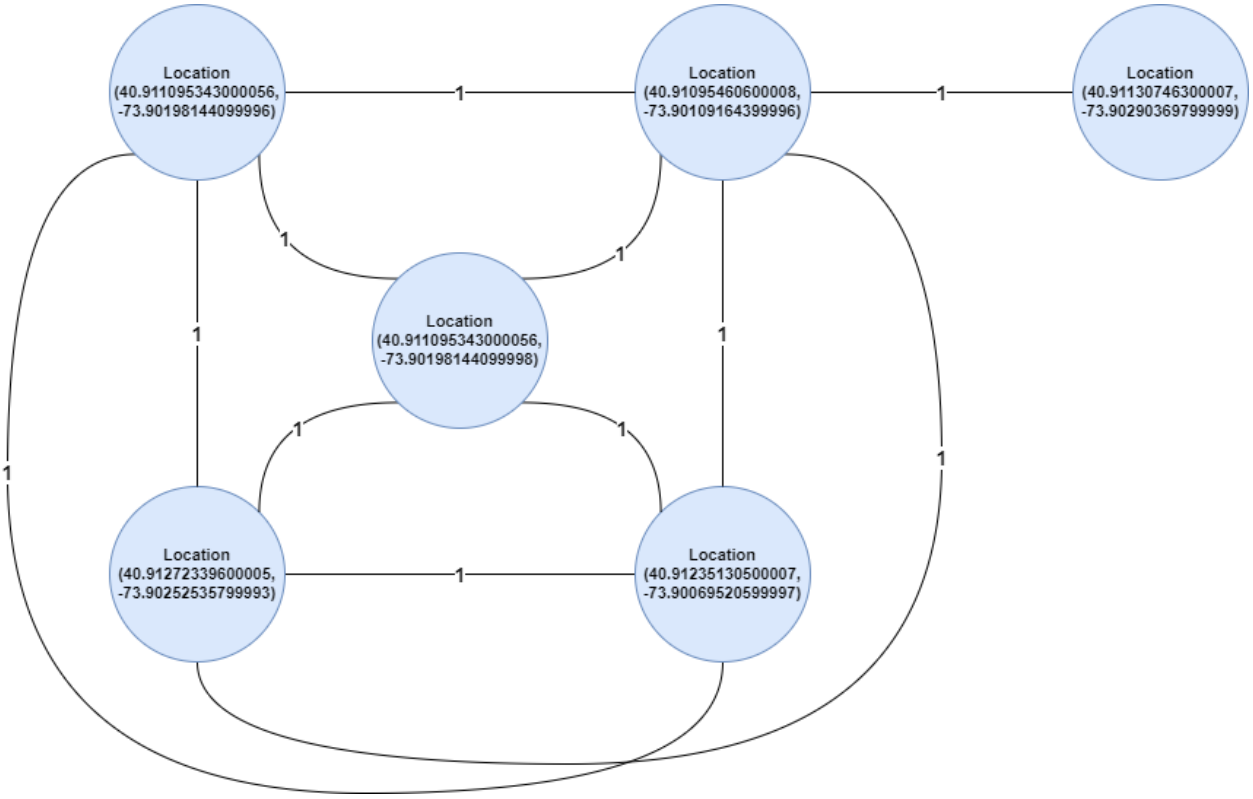
BORO_NM	Lat_Lon	OFNS_DESC	CMPLNT_NUM
BRONX	(40.91272339600005, -73.90252535799993)	THEFT-FRAUD	1
		GRAND LARCENY	5
		FORGERY	1
		FELONY ASSAULT	2
	(40.91235130500007, -73.90069520599997)	GRAND LARCENY OF MOTOR VEHICLE	1
		GRAND LARCENY	1
	(40.91130746300007, -73.90290369799999)	CRIMINAL MISCHIEF & RELATED OF	1
		BURGLARY	1
	(40.911095343000056, -73.90198144099998)	GRAND LARCENY	1
	(40.911095343000056, -73.90198144099996)	GRAND LARCENY	1
	(40.91095460600008, -73.90109164399996)	GRAND LARCENY	1
		CRIMINAL MISCHIEF & RELATED OF	1

Bipartite Projection



Monopartite Projections




Location
(40.911095343000056,
-73.90198144099996)
Location
(40.91095460600008,
-73.90109164399996)
Location
(40.91130746300007,
-73.90290369799999)
Location
(40.911095343000056,
-73.90198144099998)
Location
(40.91272339600005,
-73.90252535799993)
Location
(40.91235130500007,
-73.90069520599997)
1