Pete Rodrigue
Machine Learning Homework #1

You can find a link to my github repo with all the code here:
https://github.com/pete-rodrigue/ML_spring_2019.git

Problem 1

1. See code. The commented out section downloads the data by paging through a sequence of URLs.
2.

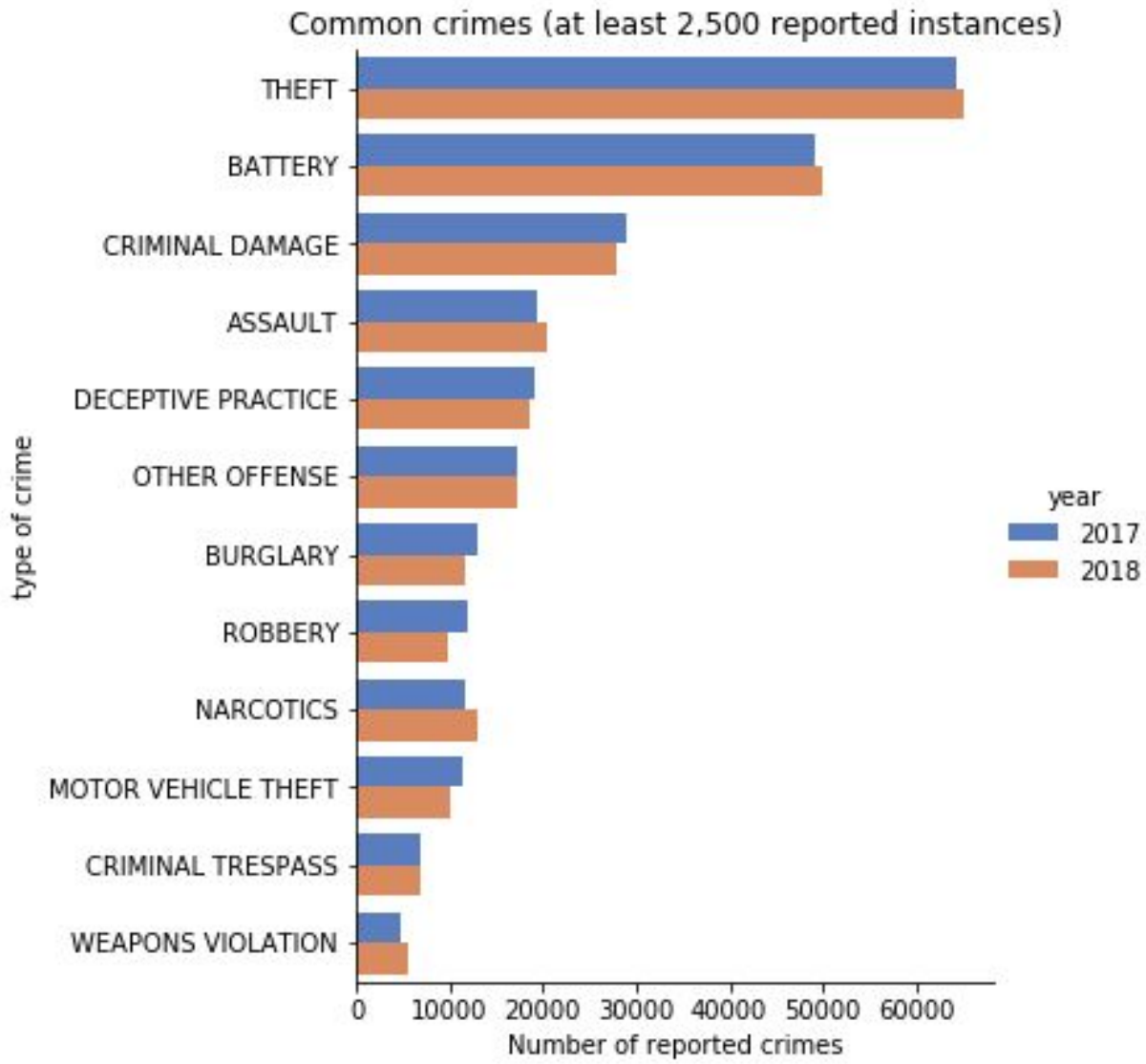
Here are the total numbers of different reported crimes in 2017 and 2018:

type of crime	**2017**	**2018**
THEFT	64344	65073
BATTERY	49214	49781
CRIMINAL DAMAGE	29042	27805
ASSAULT	19303	20375
DECEPTIVE PRACTICE	19018	18703
OTHER OFFENSE	17227	17123
BURGLARY	13001	11729
ROBBERY	11877	9683
NARCOTICS	11658	12975
MOTOR VEHICLE THEFT	11406	9987
CRIMINAL TRESPASS	6812	6904
WEAPONS VIOLATION	4686	5450
OFFENSE INVOLVING CHILDREN	2271	2229
CRIM SEXUAL ASSAULT	1628	1626
PUBLIC PEACE VIOLATION	1498	1370

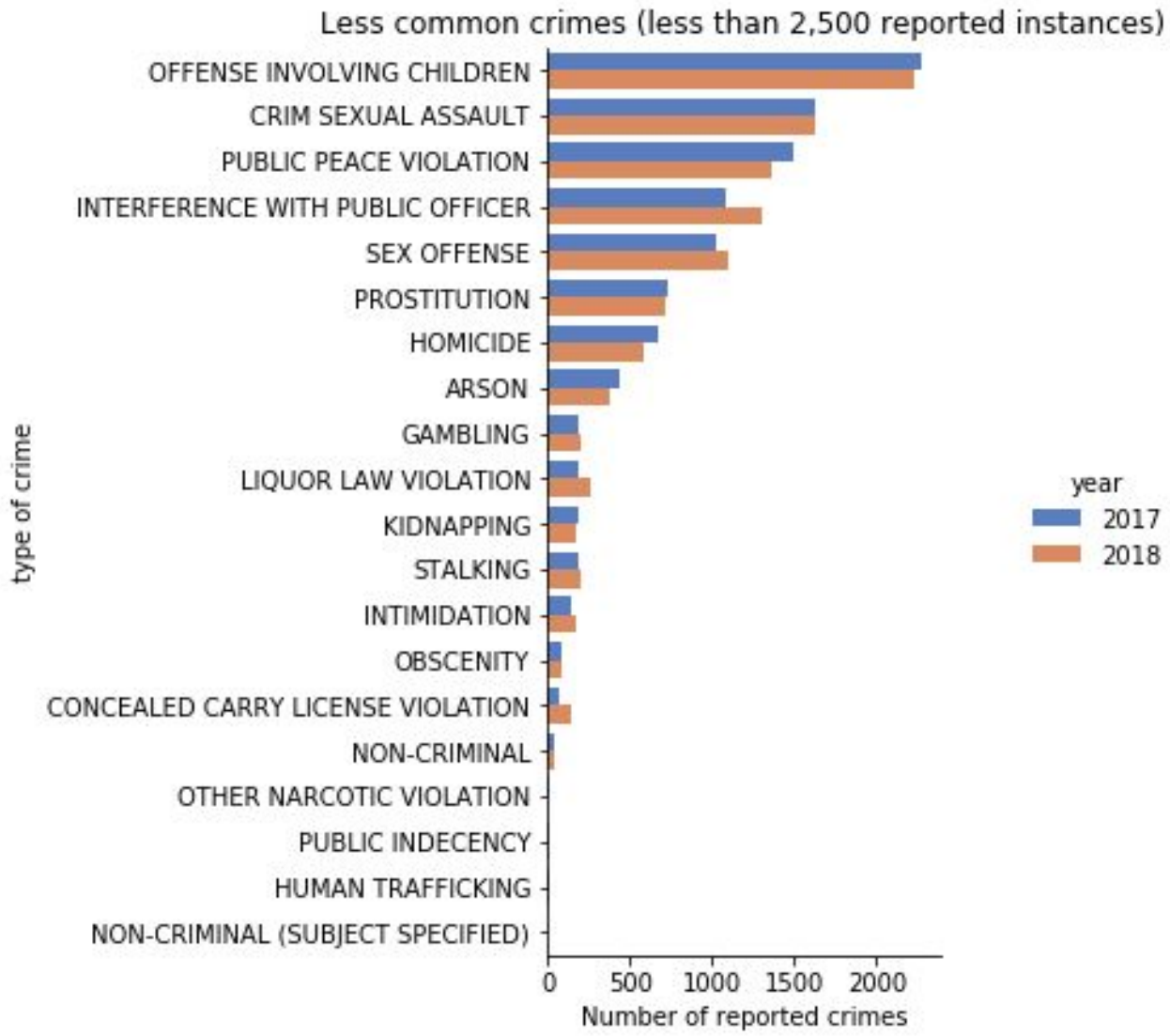
INTERFERENCE WITH PUBLIC OFFICER	1086	1305
SEX OFFENSE	1025	1100
PROSTITUTION	735	718
HOMICIDE	676	586
ARSON	444	372
GAMBLING	191	201
LIQUOR LAW VIOLATION	191	267
KIDNAPPING	190	169
STALKING	188	203
INTIMIDATION	151	168
OBSCENITY	87	86
CONCEALED CARRY LICENSE VIOLATION	69	149
NON-CRIMINAL	38	37
OTHER NARCOTIC VIOLATION	11	1
PUBLIC INDECENCY	10	14
HUMAN TRAFFICKING	9	14
NON-CRIMINAL (SUBJECT SPECIFIED)	2	3

Here is that same data, but in an easier-to-read bar graph:

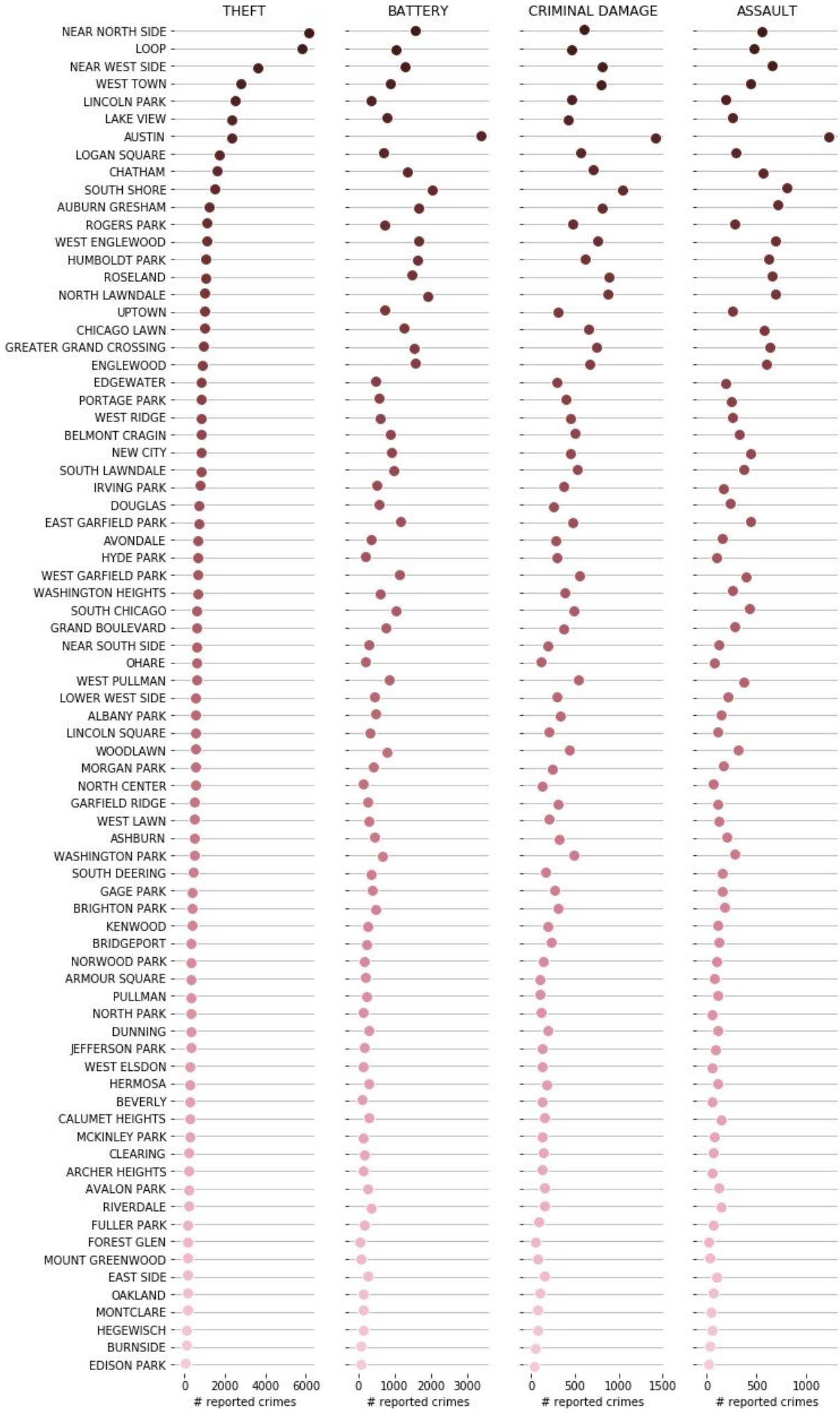
Common crimes:



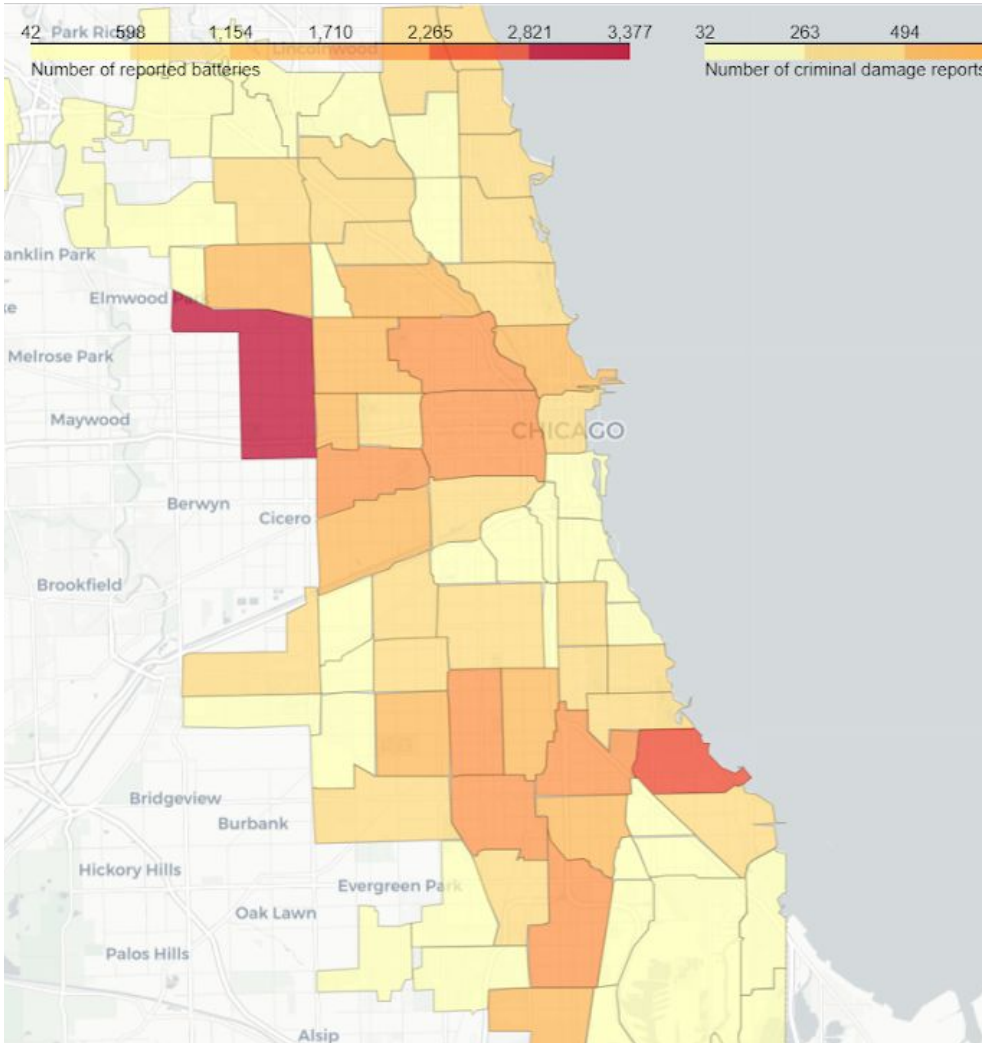
Less common crimes:



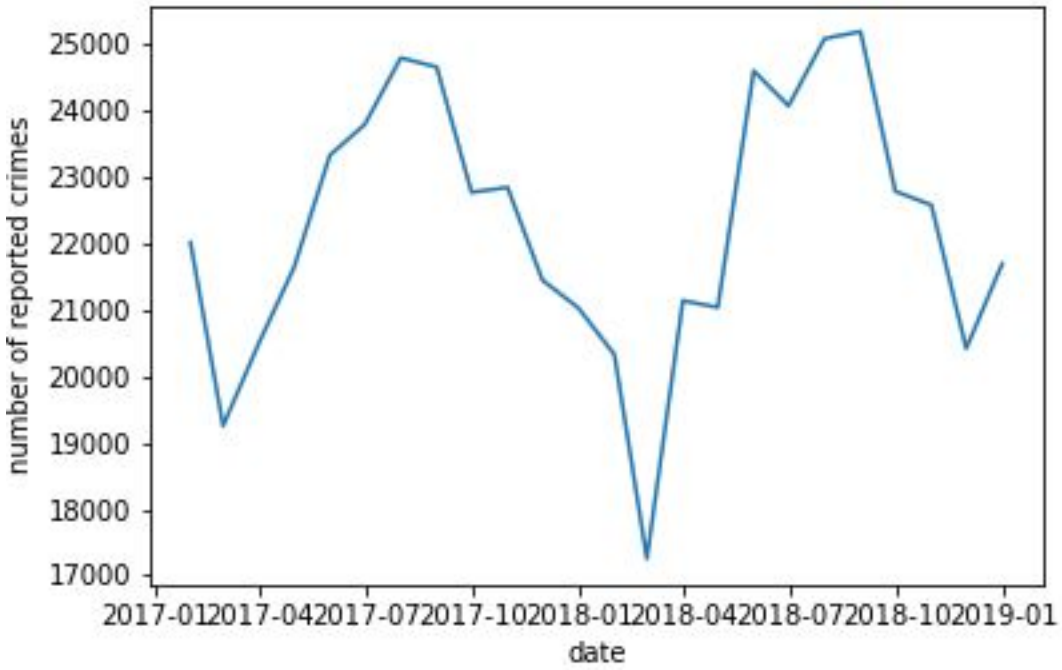
On the next page, see the number of reported crimes by community area, for the four most common crimes.


THEFT
BATTERY
CRIMINAL DAMAGE
ASSAULT
NEAR NORTH SIDE
LOOP
NEAR WEST SIDE
WEST TOWN
LINCOLN PARK
LAKE VIEW
AUSTIN
LOGAN SQUARE
CHATHAM
SOUTH SHORE
AUBURN GRESHAM
ROGERS PARK
WEST ENGLEWOOD
HUMBOLDT PARK
ROSELAND
NORTH LAWNDALE
UPTOWN
CHICAGO LAWN
GREATER GRAND CROSSING
ENGLEWOOD
EDGEWATER
PORTAGE PARK
WEST RIDGE
BELMONT CRAGIN
NEW CITY
SOUTH LAWNDALE
IRVING PARK
DOUGLAS
EAST GARFIELD PARK
AVONDALE
HYDE PARK
WEST GARFIELD PARK
WASHINGTON HEIGHTS
SOUTH CHICAGO
GRAND BOULEVARD
NEAR SOUTH SIDE
OHARE
WEST PULLMAN
LOWER WEST SIDE
ALBANY PARK
LINCOLN SQUARE
WOODLAWN
MORGAN PARK
NORTH CENTER
GARFIELD RIDGE
WEST LAWN
ASHBURN
WASHINGTON PARK
SOUTH DEERING
GAGE PARK
BRIGHTON PARK
KENWOOD
BRIDGEPORT
NORWOOD PARK
ARMOUR SQUARE
PULLMAN
NORTH PARK
DUNNING
JEFFERSON PARK
WEST ELSDON
HERMOSA
BEVERLY
CALUMET HEIGHTS
MCKINLEY PARK
CLEARING
ARCHER HEIGHTS
AVALON PARK
RIVERDALE
FULLER PARK
FOREST GLEN
MOUNT GREENWOOD
EAST SIDE
OAKLAND
MONTCLARE
HEGEWISCH
BURNSIDE
EDISON PARK
0 2000 4000 6000
reported crimes
0 1000 2000 3000
reported crimes
0 500 1000 1500
reported crimes
0 500 1000
reported crimes

[You can find a leaflet map of common reported incidents by community area here.]

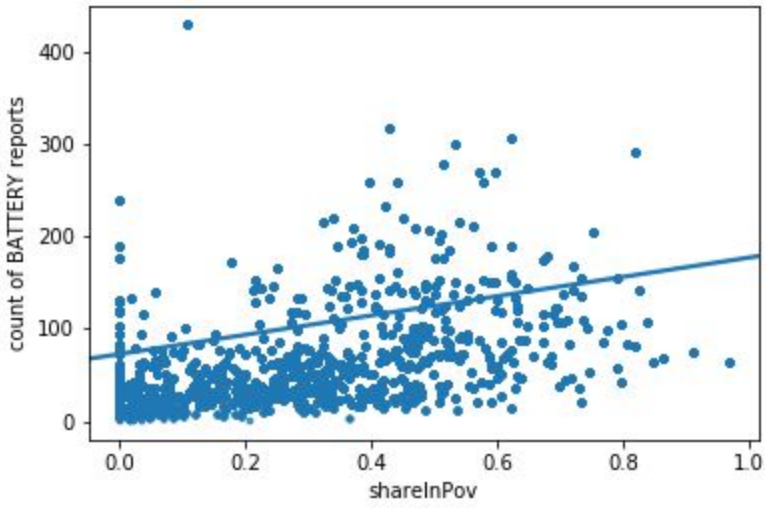


This graph shows reported crime trends over 2017 and 2018. You can see that reported crime tends to rise in the summer:

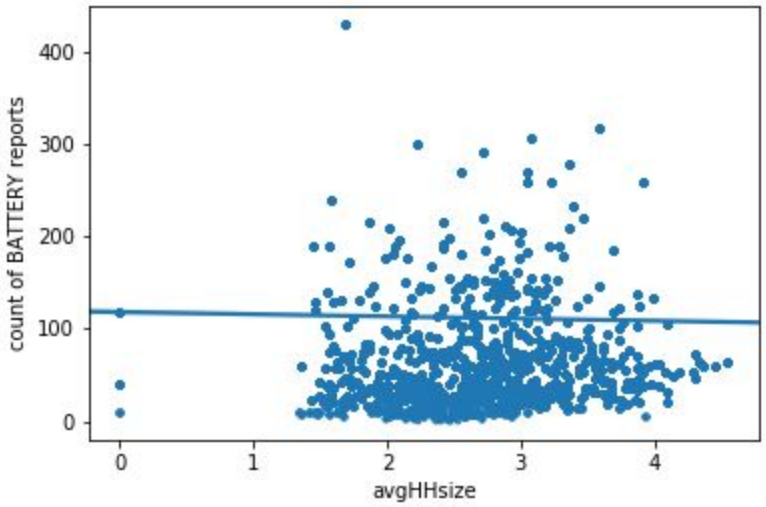


Problem 2

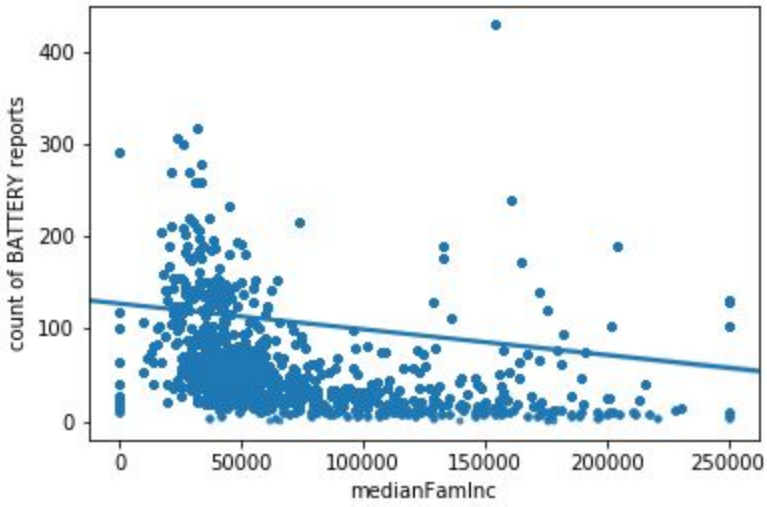
For this problem I grabbed data on median family income, average household size, and share of population in poverty, all at the census tract level.
Tracts with more battery reports also tended to be have higher poverty rates:



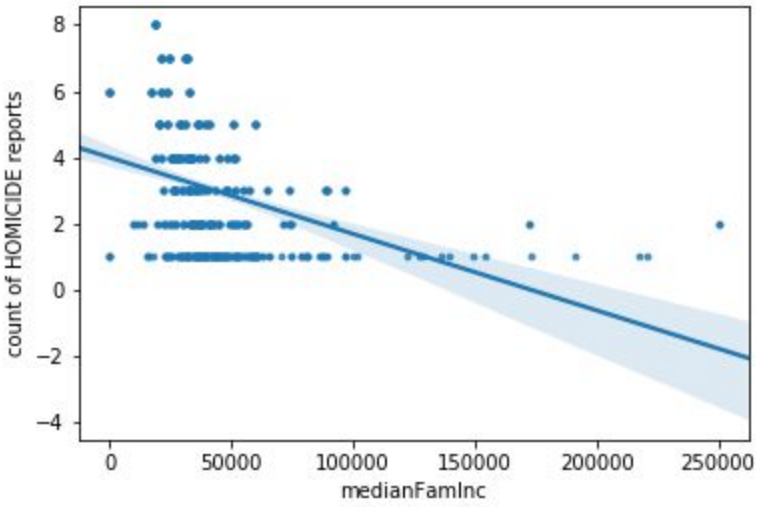
There wasn't much of a relationship between average household size in the tract and number of battery reports:



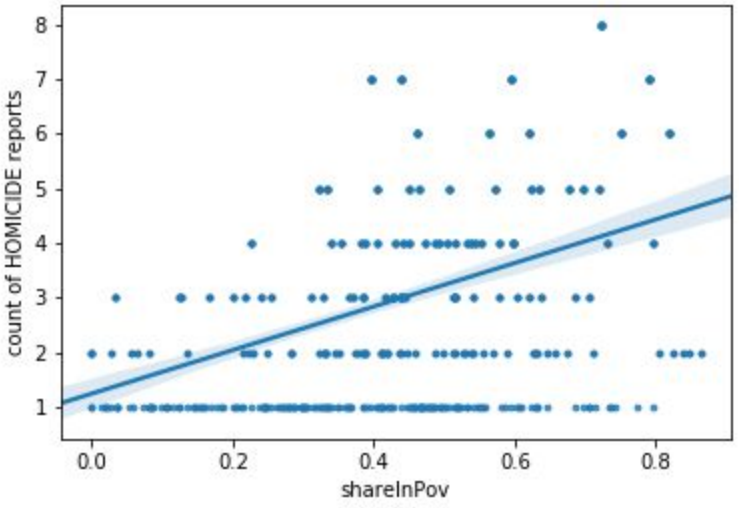
Tracts with lower median family incomes tended to have more battery reports, although there were some richer tracts that also had higher numbers of reports:



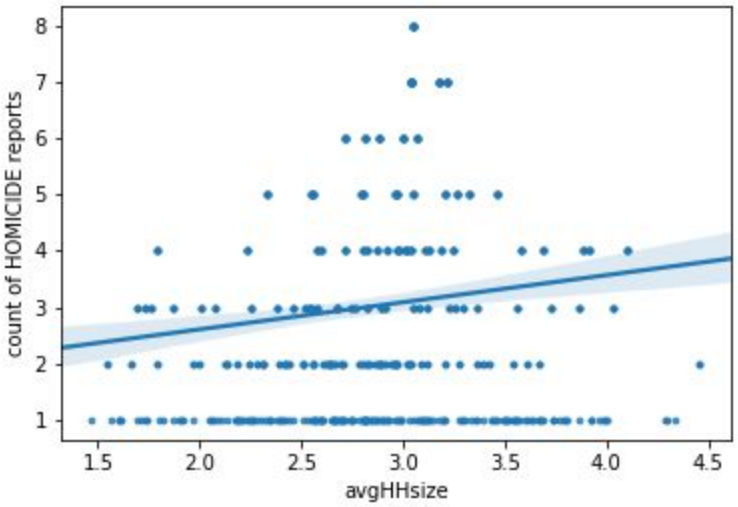
Tracts with lower family income also saw most of the homicide reports:



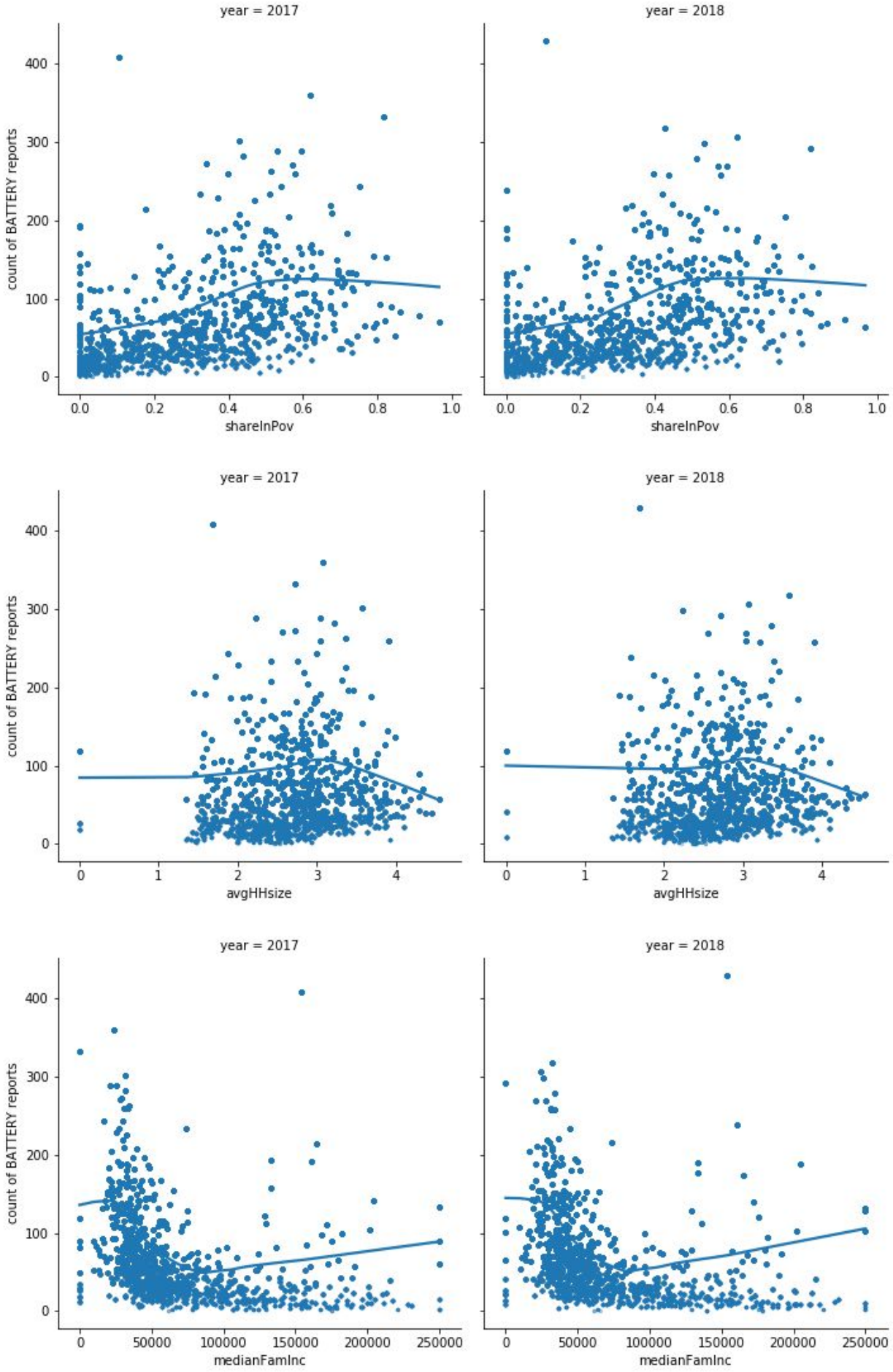
Tracts with higher poverty rates saw more homicides:



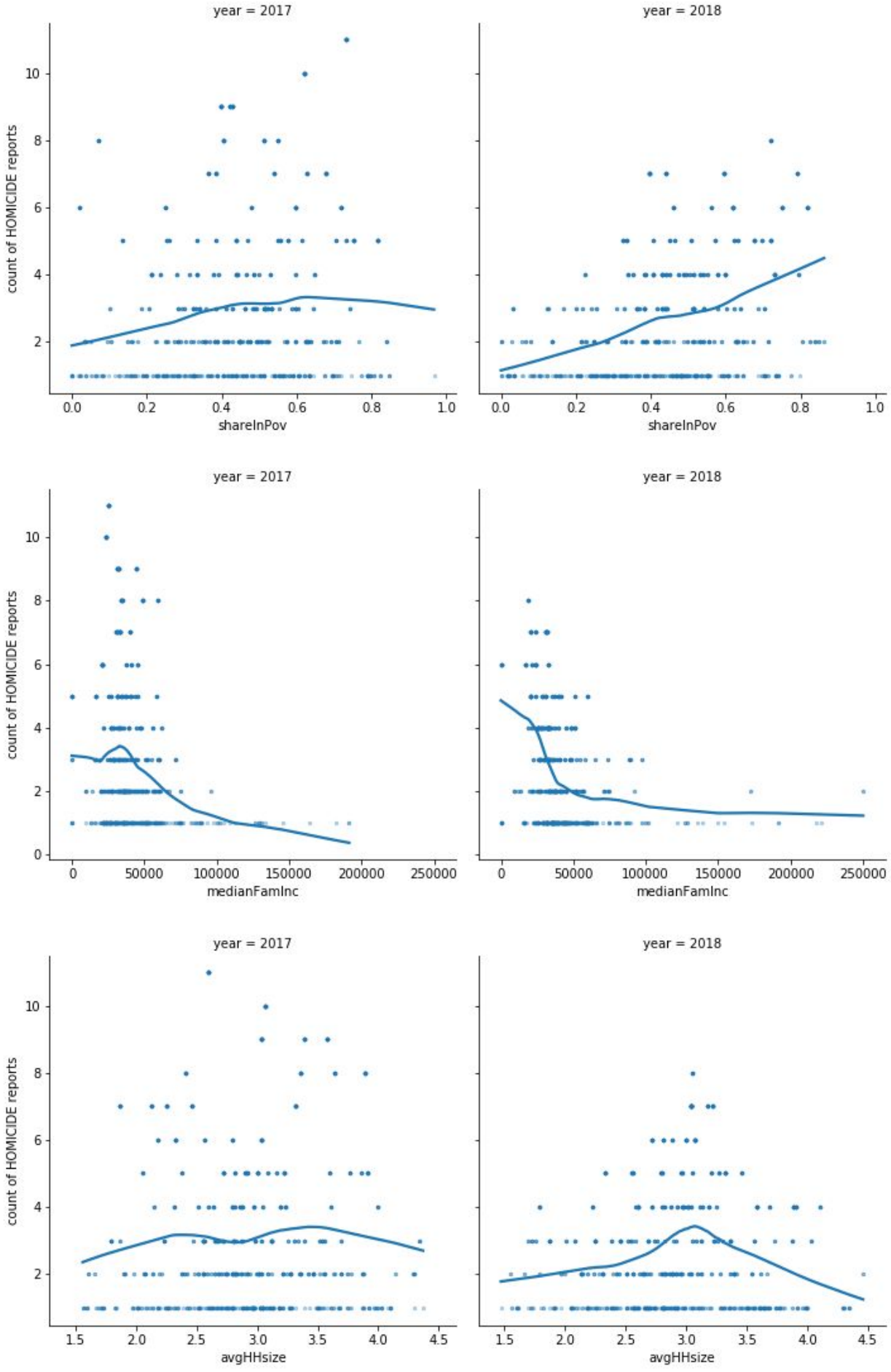
But there was less of a strong relationship between household size and homicide reports:



These relationships were pretty consistent between 2017 and 2018, especially for battery:

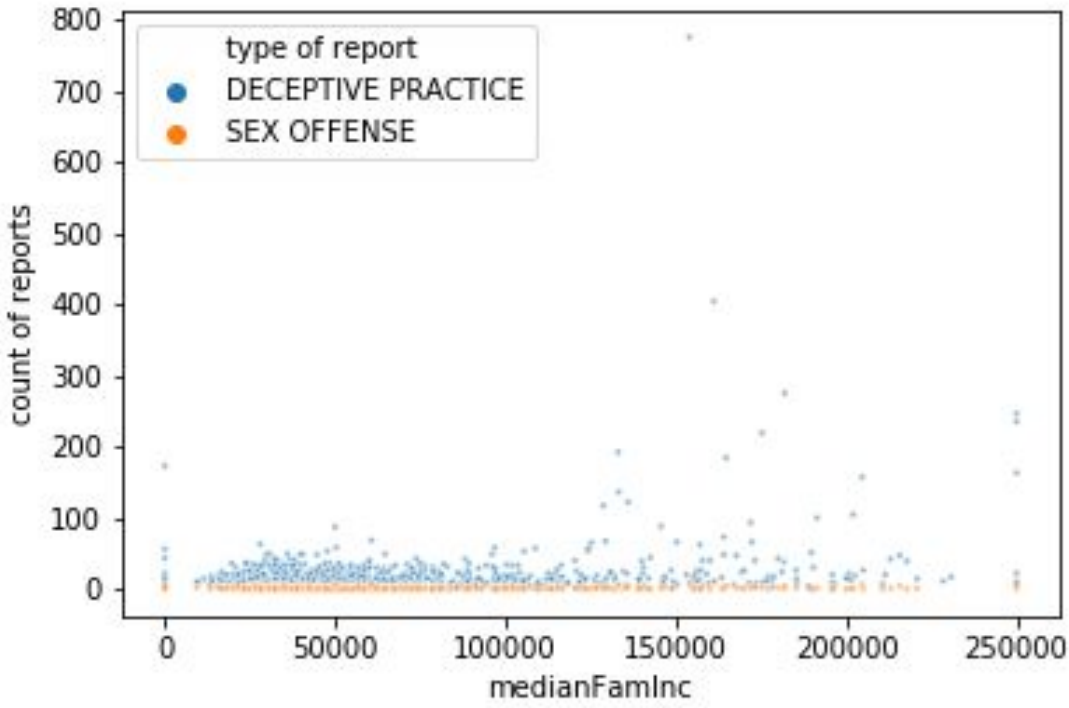


The relationship between tract income and homicide count may have strengthened over time:



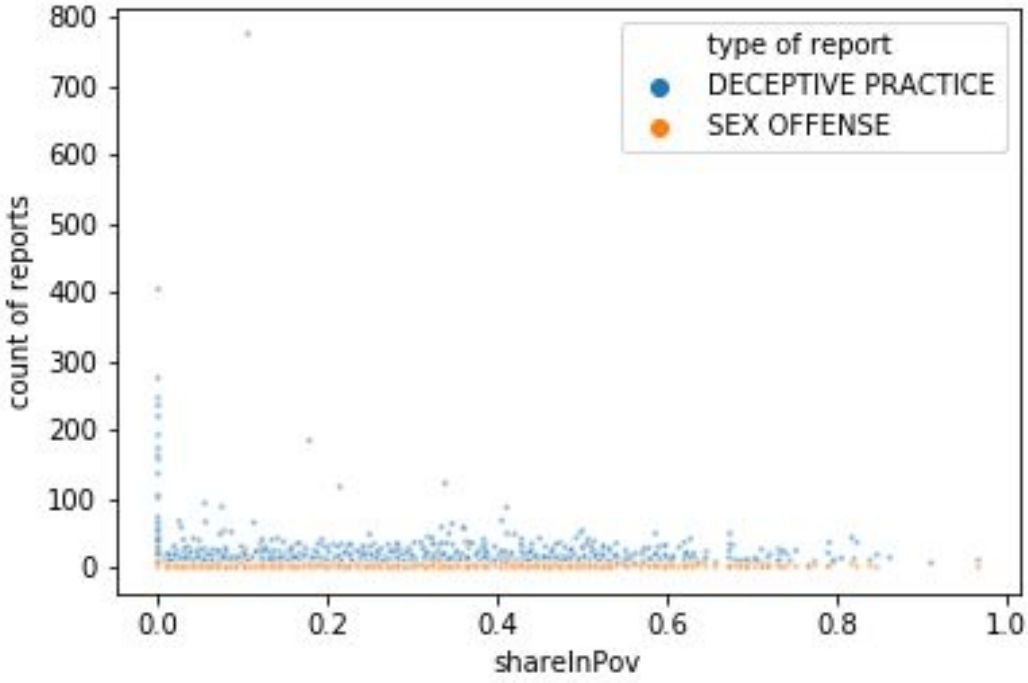
Tracts with “deceptive practice” reports versus “sex offense” reports.

A small number of richer tracts seem to have more deceptive practice reports:



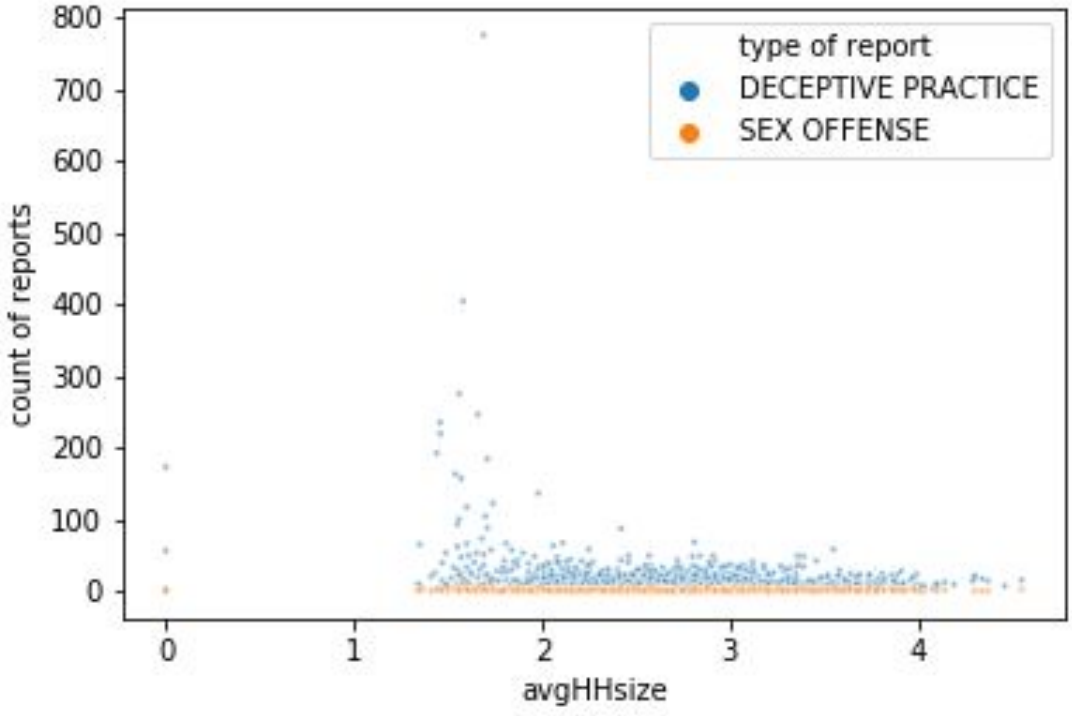
The report-weighted tract average of median family income is about $93,000 for “deceptive practice” and $72,000 for “sex offense.”

A small number of tracts with no low-income people (or no people in general) seem to have a lot of deceptive practice reports.



The report-weighted tract average of “share in poverty” is about 0.24 for “deceptive practice” and 0.34 for “sex offense.”

A small number of tracts with smaller household sizes have more deceptive practice reports:

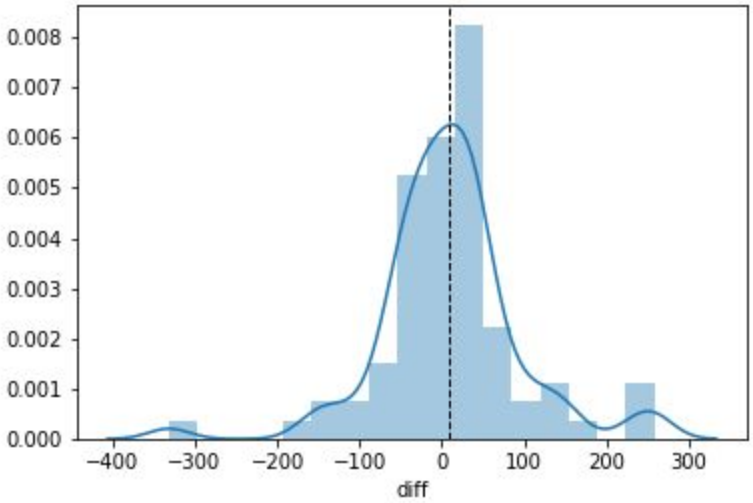


The report-weighted tract average of “average household size” is about 2.4 for “deceptive practice” and 2.6 for “sex offense.”

Problem 3

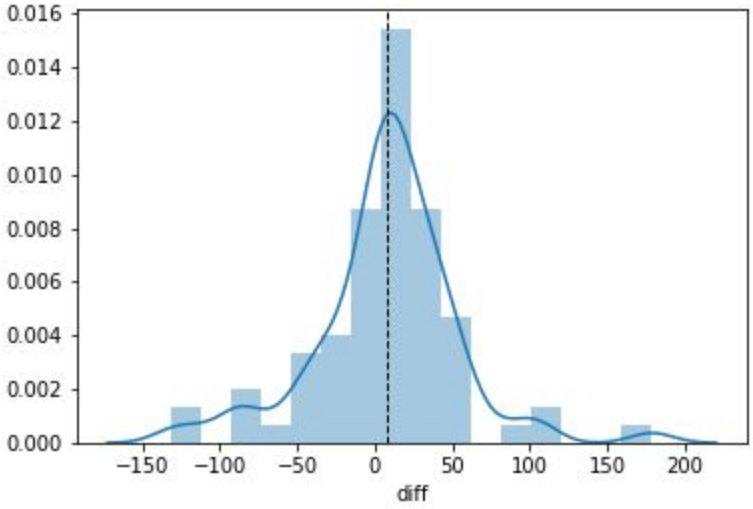
1. Overall, reported crime fell slightly, from 5,829,604 total reports in 2017 to 5,798,034 reports in 2018. That said, certain types of crimes saw an uptick citywide. While homicide, criminal damage, deceptive practice, burglary, robbery and motor vehicle theft all fell slightly, theft, battery, assault, and narcotics rose slightly. And some community areas saw large changes. Theft reports fell from 3,103 to 2,772 in West Town, for example, a decline of 331 reports. That was the biggest decline. But you can see there's wide variation below.

   This shows the distribution of the change in theft reports at the community area level:

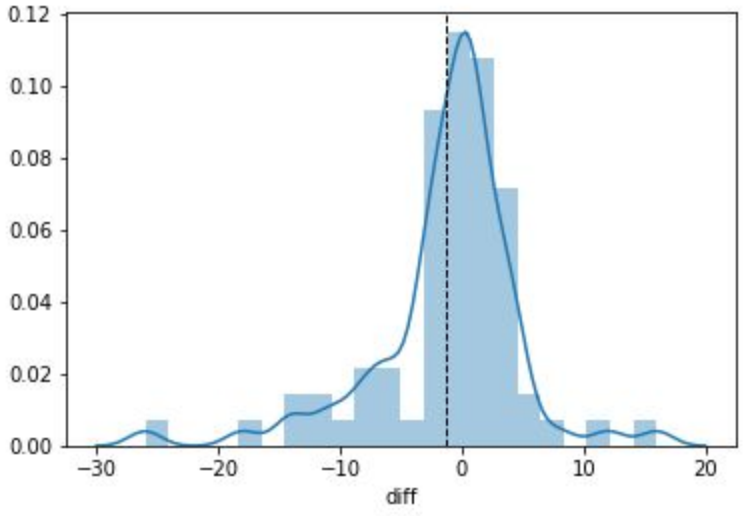


   You can see that West Town saw a large decline, while other community areas saw an increase. The mean change at the community area level (not weighted by population) is shown by the dashed line.

   Similarly, the graph below shows the change for battery reports:



And this histogram shows the change for homicides. Note that while many areas saw a decline in homicides, some saw substantial increases. West Pullman is the most extreme example. In 2017 it already had 10 homicides. In 2018 that rose to 26.



2.
    a. Honestly I was a little confused by his numbers. When I looked at the number of robberies in that month-long period, they fell from 28,623 to 25,433. If you just look at the 43rd ward, they also fell, from 493 to 319. Battery reports rose overall, from 129,050 to 137,837. But that's only a 6.8% increase. And in the 43rd ward they fell slightly, from 1102 to 1015. Burglaries also fell both citywide and in his ward. Motor vehicle thefts did increase in his ward (they doubled). So his numbers seem pretty different.
    b. I suspect he was using this online tool provided by the police department, which may explain why our numbers differ. But that seems a little iffy, because that online tool lets you set the search radius pretty small, so you might not be surprised to see large percentage year-to-year variation in the numbers of different crimes.
3. Key findings:
    a. Overall crime is slightly down.
    b. Despite that, some communities have seen substantial increases in homicides and other serious crimes.
    c. Those communities are relatively few in number. They tend to have lower average incomes. Homicides are especially concentrated in lower-income areas.
    d. Reported crimes are seasonal: they increase in the summer and fall in the winter.The most common reported crimes are theft, battery, criminal damage, and assault.
4. Caveats:
    a. This is only two years of data. We'd want to look at more years to get a better sense of trends.
    b. We've only looked at a few community characteristics (income, poverty, and household size). We'd want to look at more to get a richer sense of the places with the most reports.
    c. Reported crime is not necessarily the same thing as an actual crime.

d. Some crimes may happen in small numbers but have an outsized effect. (The number of reports isn't everything). There may be lots of thefts in the loop, but a small number of homicides in a neighborhood demands is a bigger problem.
e. We should always worry about missing data. We can't be sure if people don't report some types of crimes (like prostitution or white collar crime).

Problem 4

1. Address 2111 S Michigan Ave, Chicago, IL 60616 is in census tract 3301. Based on the data I have, which is aggregated at the census tract level, the most likely reported crime will be theft. (This is looking at both years of data combined). Here are the raw numbers and the probability of each report:

	number	probability
THEFT	794	0.31
DECEPTIVE PRACTICE	396	0.16
BATTERY	388	0.15
CRIMINAL DAMAGE	202	0.08
ASSAULT	166	0.07
OTHER OFFENSE	137	0.05
MOTOR VEHICLE THEFT	104	0.04
CRIMINAL TRESPASS	100	0.04
ROBBERY	100	0.04
BURGLARY	63	0.02
SEX OFFENSE	17	0.01
CRIM SEXUAL ASSAULT	15	0.01
WEAPONS VIOLATION	15	0.01
NARCOTICS	12	0.00
OFFENSE INVOLVING CHILDREN	10	0.00
INTERFERENCE WITH PUBLIC OFFICER	7	0.00
PUBLIC PEACE VIOLATION	4	0.00

LIQUOR LAW VIOLATION	3	0.00
OBSCENITY	3	0.00
ARSON	3	0.00
INTIMIDATION	2	0.00
NON-CRIMINAL	2	0.00
STALKING	2	0.00
CONCEALED CARRY LICENSE VIOLATION	1	0.00

2. It is more likely that the call came from Garfield Park (I used both East and West Garfield Park). There were 2498 reports of theft in East and West Garfield Park in 2017 and 2018, and only 1949 in Uptown during that time. Based on those figures we'd expect about 56 percent of theft reports from Uptown and Garfield Park to come from Garfield Park. So if a given theft call comes from one of those two places, it's about 1.27 times more likely that it came from Garfield Park.
3. 260 total calls are about battery (100 from Garfield Park and 160 from Uptown). So 38% (100 / 260) of all battery calls come from Garfield Park. That means that if a given call is about battery, it's 0.73 times less likely to come from Garfield Park than Uptown (38 / 52 = 0.73).