KC crime data 2017-2022

665k rows of data to begin with (incidents)

Removed all duplicate rows based on same report no & date & time & location & IBRS/description—these dupes are almost certainly the opposite report from the victim/perp and so would be double counting.

Cleaning address/location

This leaves ~330k rows.

There are only 4 rows that have NO location and NO address, so we will just delete those. (location and address are basically redundant, but some rows have one and not the other so we can use them to fill in the others)  
another 300 rows have ‘unknown’ listed as the address and/or location with a ZIP of 99999, removing those as well.

Another ~110 rows are from cities in other states or other parts of Missouri outside of Kansas city and the 6-county area. Removing.

this leaves 329k rows. There still may be some to remove here if we are narrowing in on only KC proper and not the other cities (nkc, gladstone, liberty, independence, etc)

Cleaning crimes

Recoded based on text ‘offense’ field to capture crimes that ought to have NIBRS part I codes. Captured approx. 3k part I crimes this way. Then removed all remaining rows with missing NIBRS codes. This dropped 30k rows

299k rows remaining

Removed warrants and suicides from data (~1k)

298k remaining

Pulling out report numbers with multiple entries of part I crimes to clean that to match UCR reporting methodology.

980 rows need to have hierarchy rule applied

First removed duplicates based on same report no, date, and IBRS code (so if there were two agg. Assault on the same night but 5 min apart in reporting it’s only one row now). then remove dupes based on same report no and same date after sorting by IBRS because it’s already hierarchical conveniently. (so murder + agg assault on the same date same report is condensed to just the murder row)

Then finally we remove multiples that are on different days but the same month, as the UCR summarizing happens at the month level. This removes about 200 more rows, leaving 561, meaning we’re dropping about 500 rows through this process. (if there was a report with 2 stolen autos, one on jan 4 and one on jan 10, only the single jan 4 row stays. If one is in jan and one in feb they both stay even if report# is the same).

Next questions:

clusters of reports at locations (police stations, courthouse, etc)—are there clusters, how many, how to handle

duplicate report nos but on different days/times/locations—likely adding incidents to known perpetrator? Inclined to keep all bc it’s ambiguous what the right move would be. 4k rows affected by this decision.