Serena Bonci - Dataset sanitization
Kathryn Van Etten - Algorithm handling
Marc Angelo Acebedo - Report Write-up

DATASET

The dataset used in this project depicts all the crimes reported in New York City between 2013-2015 (with a few outliers from 1997-2005), compiled directly by the NYPD. Our aim was to implement the Apriori algorithm in order to see frequent sets that occur among different types of sex-related crimes. These outliers were included in the study since the time of the commission of the crime was not important to our aim.

For data sanitization, we removed the following seven columns that contained unnecessary data: *latitude, longitude, latitude_longitude, x-coordinates, y-coordinates, addresses,* and *parks_nvm* (representing crimes that took place in a park). These columns were unnecessary in our aim since we sought to find frequency and variance among the *type* of sex-crime and not their geographical locations. Afterward, every row with missing data was removed to account for greater accuracy. We then used R to filter the data and show only sex crimes, excluding sex crimes committed against minors due to potentially triggering content. Grep was used to include and omit entries by a given string per row.

After the process of data sanitization, the dataset went from over 100,000 rows and 24 columns to less than 5,000 rows and 17 columns. After more sanitization, the number of columns went down to 7 after removing the column that corresponded to row *number, x, x.1, x.2, x.3*, (unclear meaning), *complaint number, report date, report time, resolution date,* and *resolution time*, *ky_cd* (represents which kind of crime took place), *pd_cd* (code corresponding to specific type of crime)*, jurisdiction, location of occurrence,* and *addr_pct_cd* (code corresponding to whether crime took place inside or in front of something).

ALGORITHM IMPLEMENTATION

Python was used to code the Apriori Algorithm in question. The minimum confidence used was 0.5 and, as a final step in an analysis of the result, an upper bound of 0.9 was implemented to avoid too many redundant and irrelevant results. Since we wanted to turn every grouped frequency into a Python set, *frozenset()* was used to make those sets immutable.

RESULTS

An interesting variance of frequent itemsets displayed after implementing the Apriori Algorithm using Python.

Of all the boroughs, only Manhattan and Brooklyn showed up in correspondence to *completed*. Similarly, out of all the locations in which the commission of the sex crime could occur, the only

type that showed up was *residence* (e.g. apartment/house), implying that sex crimes occur mostly in or around residential areas in New York City. The rules extracted from these frequent itemsets are summarized below.

Rules
Brooklyn implies misdemeanor, completed, and sexual abuse.
Manhattan implies misdemeanor, completed, sex crimes
Residence implies sex crimes, misdemeanor, and felony.

Overall, there were more completed than attempted sex crimes.
Overall, there were more misdemeanors than felonies.

The items appearing in correspondence to the two boroughs in their respective frequent itemsets is perfectly consistent with the overall pattern of other frequent itemsets that do not include those boroughs. Therefore, Manhattan and Brooklyn show consistent patterns in frequency in sexual crimes that are misdemeanors and are completed.

As for residential places, there is no difference in the implication of a felony or misdemeanor taking place.

RECOMMENDATIONS
Cluster analysis could have been a better approach for this study. Next time, I would advise the NYPD to explicitly show data committed by NYPD officers as well as any other law enforcement officer, in order to aid in any future related studies covering analyses of police brutality vs. civilian criminal activity.