After running our analysis on the CCRB dataset we do not see any major privacy concerns with the release of this dataset. We have reached this conclusion because of the following reasons:

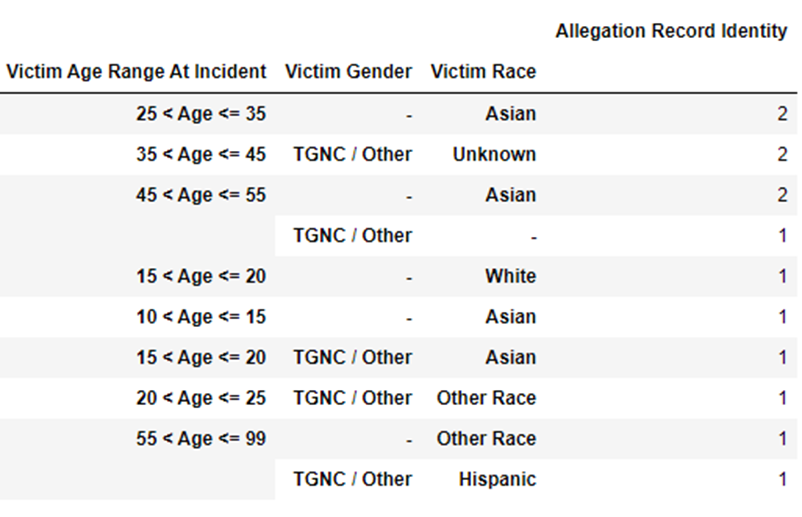
* Complainants are not required to report the victim’s demographic information (age, gender, and race), and as a result these fields are not very well populated (40-50% empty). This practice provides a layer of privacy protection in the data collection process.
* This dataset only includes the police precinct of the reported incident. It does not include the exact location or address information for the victim. Even at the smallest level of grouping based on a victim’s demographic information (ex: Asian, TGNC/Other, 10 <= Age <= 15) the overall populations for these groups are large enough in police precincts that it would be very hard to identify an individual with this dataset.
  + [Seminal work](https://dataprivacylab.org/projects/identifiability/paper1.pdf) in this area was done by Dr. Latanya Sweeney in 2000 and still is very relevant (about page [here](https://aboutmyinfo.org/identity/about)). She was able to link de-identified health records that contained birthdate, gender, and zipcode to the voter registration (which has name and address). In this case, we don’t have birthdate or even birth year (only age range) which provides a layer of privacy protection. We don’t have a zip code. The use of police precinct,a larger geographic area, provides another layer of privacy protection. Furthermore, the precinct where the complaint took place may or may not be the precinct that the victim resides in.

More details on the results of the assessment can be found below.

If we ignore age (given that we only have age range and not exact age at time of incident), we are left with gender and race. Victim gender is mostly populated with “Male” or “Female” and a very small subset is (~0.05%) “TGNC/Other”. Assuming a roughly 50:50 distribution for Male:Female in any given precinct, let’s take the smallest gender grouping (“TGNC/Other”) and look at the race characteristics for these individuals. The identifying race values are “Asian”, “Hispanic”, “Black”, “White”, and “American Indian” and the precincts with the smallest populations for these race groups have the following populations:

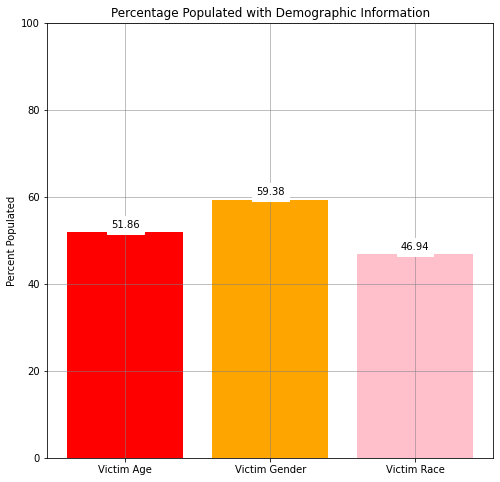
* White-only: 3000 (precinct 113)
* Black or African American-only: 1300 (precinct 6)
* Asian-only: 450 (precinct 42)
* These categories have precincts with very low populations:
  + American Indian or Alaskan Native-only: 5 precincts with less than 100, the lowest is 56.

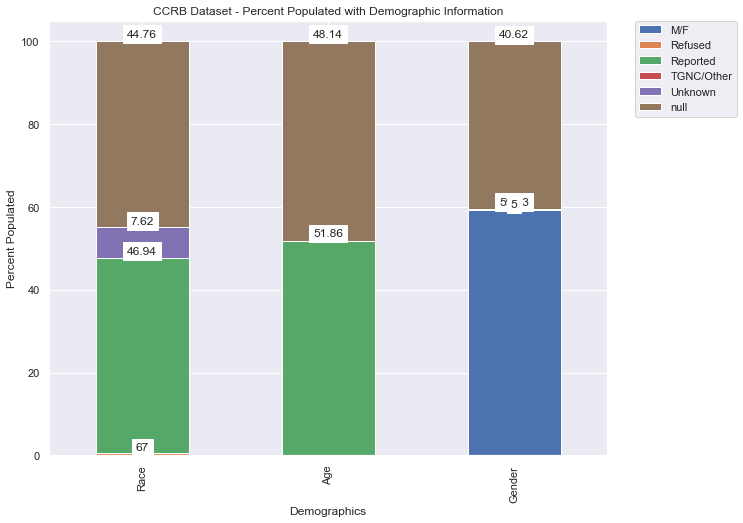
Even for this small group of “TGNC/Other” persons, it would be very difficult to identify the individual person based on their race at the police precinct level.



Here are the detailed results of our privacy assessment for the CCRB dataset:

**Unknown/Missing Demographic Information**





* There are a substantial number of unknowns in the Victim Race column – only 46.94% of the values in this field are identifying (“Asian”, “Black”, “White”, “Hispanic”, “American Indian”). For this exercise, we grouped together “Other”, “Refused”, “Unknown”, and nulls into one bucket representing all unknown values, which comprises 53.05% of the data.
* When we look at the number of people that provided identifying information for race we only have 59.32% of individuals that identified as ‘Male’ or ‘Female’, and 0.05% of individuals are ‘TGNC/Other’. The number of nulls or missing values for gender comprise 40.62% of the data.
* Only 51.86% of ages provided in the Victim Age column. This field is already grouped into buckets of age ranges instead of exact age which provides additional privacy protection. The age groupings are as follows:

25 < Age <= 35

35 < Age <= 45

20 < Age <= 25

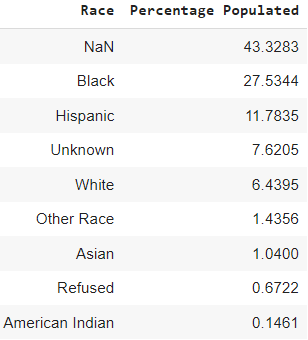
15 < Age <= 20

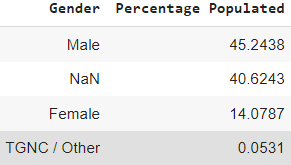
45 < Age <= 55

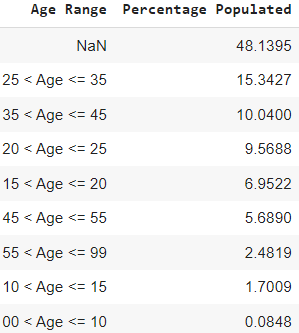
55 < Age <= 99

10 < Age <= 15

00 < Age <= 10



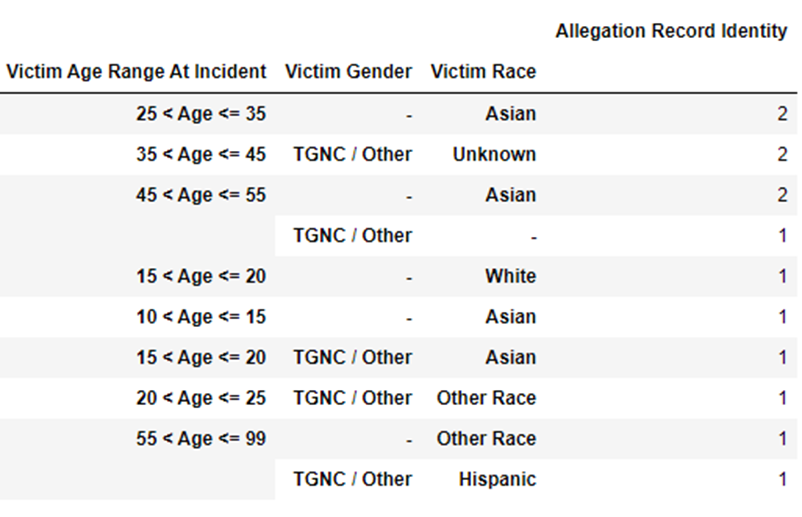




**Identification at the Police Precinct level:**

This dataset does not include a victim's location or address. The location identifiers in this dataset include police precinct of incident, borough, and location type (ex: park, police building, etc.). Given this and past research conducted on precinct level populations by race (see end of this email), it would be very difficult to identify an individual at the precinct level.

We also considered the possibility of the missing/unknown values discussed above becoming more populated over time (something that is 50% populated this year may be 90% in the next 10-15 years). But, that also means that the smallest grouping of these demographic characteristics for an individual will also grow in size. For example, the table below, represents the smallest groupings of age, gender, and race in the dataset, and the size of these groups will also grow with time making it harder to identify an individual in a given precinct.



**Past Research on Precinct Level Populations by Race:**

Precinct 22 has 25 people total (this is likely Central Park). Removing this from consideration, here are the lowest population precincts:

* The lowest total population is 20k (precinct 14)
* The lowest population for single race demographics
  + White-only: 3000 (precinct 113)
  + Black or African American-only: 1300 (precinct 6)
  + Asian-only: 450 (precinct 42)
  + These categories have precincts with very low populations:
    - American Indian or Alaskan Native-only: 5 precincts with less than 100, the lowest is 56.
    - Native Hawaiian Pacific Islander-only: very few precincts with more than 100.