Replicate Ba et al. 2021, Cleaning Police Stops

Officer stops

- **stop_id**: Unique identifier for each stop.
- **time**: Time of stop.
- **date**: Date of stop.
- **district**: Police district where the stop took place.
- **po_first**: Was the focal officer the first to respond to the scene?
- **stop_type**: What was the type of the stop?
- **contact_type**: Collapsed version of **stop_type** (less categories).
- **civ.race**: Race of the civilian.
- **civ.gender**: Sex of the civilian.
- **civ.age**: Age of the civilian at the time of the stop.
- **lat**: Latitude of the stop.
- **lon**: Longitude of the stop.
- **officer_id**: Unique identifier for the officer.
- **month**: Month of stop. **Dropped as it can be recreated**.
- **civilian_race_short**: Collapsed version of **civ.race** (less categories). **Dropped as it can be recreated**.
- **hour**: Hour of the day when the stop took place.
- **stop_officer_id**: Unique identifier for each entry.
- The unit of analysis is a unique officer involved in a stop. Each row can be uniquely identified by **officer_id** and **stop_id** or by **stop_officer_id**.
- Each stop involves only one civilian, but they can involve multiple officers. It is possible multiple stops are all a part of one larger incident involving multiple civilians. This can be investigated by examining stops that took place in the same location at the same time involving the same officers.
- **Number of stops**: 1703158
- **Number of unique stops**: 946912

Ba et al. 2021 Replication - Merging Stops to Officer Assignment

Ba et al. (2021) are only concerned with stops which occur during an officer's shift so their data cleaning and merging process reflects that.

1. Right join stops and assignments using officer_id and date. I.e. keep every assignment and only stops which occurred during the start date of the shift.
2. Keep any stops which occurred during the hours of the shift.
    - Round the stop time to the lowest hour.
    - Round the shift start time to the lowest hour.
    - Round the shift end time to the highest hour.
3. Right join stops and assignments using officer_id and date (incremented by one day). This captures stops which occurred the **following day** during an overnight shift.
4. Filter out all stops which do not occur during the hours of the shift using the same logic as in step 2.
5. Merge the data sets (stops merged by date and stops merged by the next day) by row, and join them back with the shift assignment data set to capture any shift assignments which did not have a stop associated with them.

- Number of resulting stop-shift assignment observations: 3665923
- Number of shift assignments: 2932680

- Number of stop-officer observations: 1338755
  - Percentage of stop-officer observations occurring during an officer's shift: 78.6042751%.
- Number of officers who made stops during their shift: 6947
  - Percentage of officers who made at least one stop during **any** of their shifts: 100%

Changing Ba et al. 2021 Data Cleaning and Merging Processes

Missing Starting and Ending Shift Times

- Of the 2932680 shift assignments, 40064 are missing their start time and end time.
  - This represents 1.3661225% of all shift assignments.
- 7850 shift assignments matched with at least one stop **without having** a start time and an end time.
  - This represents 0.2676733% of all shift assignments.
- Ba et al. 2021 implicitly assume the matched stops occurred during these shifts which may or may not be plausible. The **date** matches at the very least (i.e. the date of the stop and the date of the shift). One can compare the same beat assignments during the same days of the week to see how regular shift start times and end times are. Assuming regularity, the start times and end times can be imputed. A more conservative solution would be to just drop all shift assignments which do not have a start time and an end time. Both will be explored.

Duplicate Entries

- There are 315 stops which match to more than one shift assignment.
  - These **duplicate** stops represent 0.0184951% of all stops (1703158).
  - Overall, 630 records are affected which represents 0.0171853% of all stop-shift assignment records (3665923).
- How can a stop match to more than one shift assignment?
  - The officer in question could have had overlapping shift assignments, and the stop occurred during the intersection of their shifts.
  - One of the shifts had a missing start time or end time, and the stop matched with that shift (in addition to matching with at least one another shift assignment).

Potential Solutions

- One easy fix which immediately comes to mind is to remove all stop-shift assignments where the start time and end time are missing. This doesn't eliminate all duplicates, but it does eliminate most of them. This issue may also be partially fixed upstream when I implement the solution for dealing with missing start times and end times.
  - Number of duplicate stops after removing shift assignments with missing start and end times: 69
  - Percentage reduction after removing shift assignments with a missing start and end time: 78.0952381%. (Going from 315 to 69)
- What should be done about the remaining duplicate entries? Well the stop did occur during both shifts if we are to believe the officer worked on overlapping shifts. It is accurate to say a stop occurred during both shifts. Care will just need to be taken when aggregating the number of stops to an officer across shifts to avoid over counting (i.e. counting the same stop twice).

Create outcomes

Aggregate across all shift assignments:

- The total number of stops for that shift.
- The total number of stops by contact type for that shift.
- The total number of stops by civilian race for that shift.