DSPC7514 Assignment 2: Subway Fare Evasion Microdata Analysis

Arnav Sahai (as7483)

2026-02-04

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on the due date.

Do note hardcode any statistics in your write-up, make sure to use inline code references. Round any decimals for readability when appropriate.

1 Load libraries.

```
library(fastDummies)
library(tidyverse)
getwd()
```

```
## [1] "/Users/arnavsahai/Desktop/Data Analysis for Policy Research using R/Lecture 2"
```

2 Load, inspect and describe the two public defender client datasets (BDS & LAS).

2a) Load datasets using `read_csv()` and inspect.

- Get a good look at the data, but don't print long, clunky output here; one approach is to call the str() function for each dataset but to suppress the included list of attributes by including the option give.attr = FALSE.

```
arrests_bds <- read_csv("microdata_BDS_inclass.csv", na = "")
arrests_las <- read_csv("microdata_LAS_inclass.csv", na = "")
str(arrests_bds, give.attr = FALSE)
```

```
## spc_tbl_ [2,246 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ client_zip: num [1:2246] 11205 11385 11226 11207 11225 ...
##  $ age       : num [1:2246] 25 20 19 17 21 52 59 32 22 19 ...
##  $ ethnicity : chr [1:2246] "Hispanic" "Hispanic" "Non-Hispanic" "Non-Hispanic" ...
##  $ race      : chr [1:2246] "White" "Black" "Black" "Black" ...
```

```
##  $ male      : num [1:2246] 1 1 0 1 1 1 1 1 0 1 ...
##  $ loc2      : chr [1:2246] "jefferson st l line station" "myrtle - wyckoff avs station" "winthrop s
##  $ st_id     : num [1:2246] 100 119 156 156 156 156 156 156 156 156 ...
##  $ year      : num [1:2246] 2016 2016 2016 2016 2016 ...
```

```
str(arrests_las, give.attr = FALSE)
```

```
## spc_tbl_ [1,965 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ client_zip   : num [1:1965] 11222 10016 11236 11236 NA ...
##  $ las_race_key : chr [1:1965] "Black" "Asian or Pacific Islander" "Black" "Black" ...
##  $ hispanic_flag: chr [1:1965] "N" "N" "N" "N" ...
##  $ age          : num [1:1965] 32 47 20 64 23 29 26 52 52 22 ...
##  $ year         : num [1:1965] 2016 2016 2016 2016 2016 ...
##  $ male         : num [1:1965] 1 0 1 1 1 1 0 1 1 1 ...
##  $ dismissal    : num [1:1965] 0 1 0 0 0 0 1 0 0 1 ...
##  $ loc2         : chr [1:1965] "kingston - throop avs" "avenue h q subway" "nostrand ave and fulton s
##  $ st_id        : num [1:1965] 106 28 131 150 131 27 68 44 85 31 ...
```

```
arrests_bds <- arrests_bds %>% mutate(race = as.factor(race), ethnicity = as.factor(ethnicity))
arrests_las <- arrests_las %>% mutate(race = as.factor(las_race_key), ethnicity = as.factor(hispanic_fla
```

2b) Give a brief overview of the data. The aim is not be exhaustive, but to paint a picture of they key features of the data with respect to the policy questions you'll be exploring.

The data are two administrative records files from Brooklyn public defenders: Brooklyn Defender Services (BDS) and Legal Aid Society (LAS). Each row is one person arrested for subway fare evasion in Brooklyn. Both datasets include demographics (age, sex, race/ethnicity) and arrest location (subway station name and id). BDS and LAS use different variable names and coding for race and ethnicity. LAS also includes a dismissal indicator (case outcome). The data are from 2016.

2c) For each dataset, what is the unit of observation and population represented by this "sample"? Do you think this sample does a good job representing the population of interest? Why or why not?

The unit of observation in each dataset is one arrest—one client of that public defender organization. The population represented is fare evasion arrestees in Brooklyn who were represented by BDS or by LAS, not all fare evasion arrests in Brooklyn. Whether this sample represents the population of all subway fare evasion enforcement in Brooklyn depends on how cases are assigned or chosen: if BDS and LAS together cover most such cases in Brooklyn, the combined sample may be fairly representative; if many arrestees use other defenders or no defender, or if the two organizations serve different areas or types of clients, the sample could over- or under-represent certain groups or locations. So the sample is useful for describing enforcement among these clients but may not generalize to all fare evasion arrests.

2d) Inspect and describe the coding of race and ethnicity in each dataset.

BDS:Race is in a single variable with categories such as Black, White, Asian/Pacific Islander, Am Indian, plus "0" and "Unknown." Ethnicity is separate, with Hispanic, Non-Hispanic, Other, and "0." So race and Hispanic identity are two different columns and must be combined for a single race/ethnicity measure.

LAS:Race is in `las_race_key` (e.g., Black, White, Asian or Pacific Islander, Latino, Unknown) and Hispanic identity is in `hispanic_flag` (Y/N). So Hispanic identity appears in two places: as "Latino" in the race key and as the Hispanic flag. To match BDS, we build a single `race_eth` that prioritizes Hispanic identity and then uses race for non-Hispanic individuals, with consistent category names across both datasets.

2e) From the outset, are there any data limitations you think are important to note?

Several limitations matter for interpretation: (1) Coverage: The data include only clients of BDS and LAS, not all fare evasion arrests in Brooklyn, so the combined sample may not represent all enforcement. (2) No benchmark: We observe arrests but not the demographic composition of riders or stations, so we cannot directly compare arrest shares to "expected" shares. (3) Missing outcomes in BDS: Dismissal is available only in the LAS data, so dismissal rates by race/ethnicity are only partly observed. (4) Missing race/ethnicity: Some records have missing or unknown race/ethnicity, which we handle by excluding them from some denominators. (5) Single year: The data are from 2016, so we cannot speak to trends over time.

3 Clean BDS race and ethnicity data (insert code chunks that only include code you used to recode and very briefly validate your recoding).

3a) BDS: race data (generate column `race_clean`).

```
arrests_bds.clean <- arrests_bds %>%
  mutate(race_clean = case_when(race == "0" ~ NA_character_, race == "Unknown" ~ NA_character_, race ==
  mutate(race_clean = fct_relevel(as.factor(race_clean), "Asian/Pacific Islander", "Black", "White", "O
arrests_bds.clean %>% count(race_clean, sort = TRUE)
```

```
## # A tibble: 5 x 2
##   race_clean                n
##   <fct>                 <int>
## 1 Black                  1465
## 2 White                   533
## 3 <NA>                    194
## 4 Other                    33
## 5 Asian/Pacific Islander   21
```

3b) BDS: ethnicity data (generate column `ethnicity_clean`).

```
arrests_bds.clean <- arrests_bds.clean %>%
  mutate(hispanic = recode(ethnicity, "0" = NA_character_, "Other" = "Non-Hispanic"))
table(arrests_bds.clean$hispanic, arrests_bds.clean$ethnicity, useNA = "always")
```

```
##
##                    0 Hispanic Non-Hispanic Other <NA>
##   Hispanic         0      493            0     0    0
##   Non-Hispanic     0        0         1558     5    0
##   <NA>            33        0            0     0  157
```

3c) Generate a single race/ethnicity factor variable `race_eth` with mutually exclusive categories.

```
arrests_bds.clean <- arrests_bds.clean %>%
  mutate(race_eth = case_when(
    hispanic == "Hispanic" ~ "Hispanic",
    hispanic == "Non-Hispanic" & race_clean == "White" ~ "Non-Hispanic White",
    hispanic == "Non-Hispanic" & race_clean == "Black" ~ "Non-Hispanic Black",
    hispanic == "Non-Hispanic" & race_clean == "Asian/Pacific Islander" ~ "Asian/Pacific Islander",
    hispanic == "Non-Hispanic" & race_clean == "Other" ~ "Other",
    TRUE ~ NA_character_
  )) %>%
  mutate(race_eth = fct_drop(factor(race_eth, levels = c("Asian/Pacific Islander", "Hispanic", "Non-Hisp
arrests_bds.clean %>% count(race_eth, sort = TRUE)
```

```
## # A tibble: 6 x 2
##   race_eth                   n
##   <fct>                  <int>
## 1 Non-Hispanic Black      1359
## 2 Hispanic                 493
## 3 <NA>                     195
## 4 Non-Hispanic White       165
## 5 Asian/Pacific Islander    21
## 6 Other                     13
```

4 Clean LAS race and ethnicity data

4) Follow your own steps to end up at a comparably coded `race_eth` variable for the LAS data.

- create race_eth in arrests_las with the same coding as for BDS
- note that Hispanic identity is included in two columns, not one: las_race_key and hispanic_flag
- Make sure you end up with a data frame with the following variable names and identical coding as in arrests_bds_clean:
  - race_eth, age, male, dismissal (not in the BDS data), st_id, loc2

```
arrests_las.clean <- arrests_las %>%
  mutate(race_clean = case_when(
    race == "Asian or Pacific Islander" ~ "Asian/Pacific Islander",
    race == "Unknown" | race == "0" ~ NA_character_,
    TRUE ~ as.character(race)
  )) %>%
  mutate(race_clean = fct_relevel(as.factor(race_clean), "Asian/Pacific Islander", "Black", "White", "O
  mutate(hispanic = recode(ethnicity, "Y" = "Hispanic", "N" = "Non-Hispanic")) %>%
  mutate(race_eth = case_when(
    hispanic == "Hispanic" ~ "Hispanic",
    hispanic == "Non-Hispanic" & race_clean == "White" ~ "Non-Hispanic White",
    hispanic == "Non-Hispanic" & race_clean == "Black" ~ "Non-Hispanic Black",
    hispanic == "Non-Hispanic" & race_clean == "Asian/Pacific Islander" ~ "Asian/Pacific Islander",
    hispanic == "Non-Hispanic" & race_clean == "Other" ~ "Other",
    TRUE ~ NA_character_
  )) %>%
```

```
  mutate(race_eth = fct_drop(factor(race_eth, levels = c("Asian/Pacific Islander", "Hispanic", "Non-Hisp
  select(race_eth, age, male, dismissal, st_id, loc2)
arrests_las.clean %>% count(race_eth, sort = TRUE)
```

```
## # A tibble: 6 x 2
##   race_eth                  n
##   <fct>                 <int>
## 1 Non-Hispanic Black     1201
## 2 Non-Hispanic White      294
## 3 <NA>                    259
## 4 Hispanic                189
## 5 Asian/Pacific Islander   11
## 6 Other                    11
```

5 Combining (appending) the BDS and LAS microdata

5a) Create a column (`pd`) to identify public defender data source.

5b) Append `arrests_bds.clean` and `arrests_las.clean` using `bind_rows()`. Store as new data frame `arrests.clean` and inspect for consistency/accuracy.

```
arrests_bds.clean <- arrests_bds.clean %>%
  mutate(pd = "bds") %>%
  select(pd, race_eth, age, male, st_id, loc2) %>%
  mutate(dismissal = NA_real_, pd = factor(pd), race_eth = as.factor(race_eth))
arrests_las.clean <- arrests_las.clean %>%
  mutate(pd = "las") %>%
  select(pd, race_eth, age, male, dismissal, st_id, loc2) %>%
  mutate(pd = factor(pd), race_eth = as.factor(race_eth))
arrests.clean <- bind_rows(arrests_bds.clean, arrests_las.clean) %>%
  select(pd, race_eth, age, male, dismissal, st_id, loc2)
summary(arrests.clean$race_eth)
```

```
## Asian/Pacific Islander               Hispanic     Non-Hispanic Black
##                     32                    682                   2560
##     Non-Hispanic White                  Other                   NA's
##                    459                     24                    454
```

5c) What is the total number of subway fare evasion arrest records?

The combined dataset contains 4211 arrest records.

```
nrow(arrests.clean)
```

```
## [1] 4211
```

5d) Save `arrests.clean` as an .RData file, in a folder for next class called Lecture4.

```
dir.create("Lecture4", showWarnings = FALSE)
save(arrests.clean, file = "Lecture4/arrests.clean.RData")
```

6 Descriptive statistics by race/ethnicity

6a) Print the number of arrests for each race/ethnicity category (a frequency table).

```
arrests.clean %>% count(race_eth, sort = TRUE)
```

```
## # A tibble: 6 x 2
##   race_eth                  n
##   <fct>                 <int>
## 1 Non-Hispanic Black     2560
## 2 Hispanic                682
## 3 Non-Hispanic White      459
## 4 <NA>                    454
## 5 Asian/Pacific Islander   32
## 6 Other                    24
```

6b) Print the proportion of total arrests for each race/ethnicity category. How does excluding NAs change the results?

```
arrests.clean %>% count(race_eth, sort = TRUE) %>% mutate(prop = n / sum(n))
```

```
## # A tibble: 6 x 3
##   race_eth                  n    prop
##   <fct>                 <int>   <dbl>
## 1 Non-Hispanic Black     2560 0.608
## 2 Hispanic                682 0.162
## 3 Non-Hispanic White      459 0.109
## 4 <NA>                    454 0.108
## 5 Asian/Pacific Islander   32 0.00760
## 6 Other                    24 0.00570
```

```
arrests.clean %>% filter(!is.na(race_eth)) %>% count(race_eth, sort = TRUE) %>% mutate(prop = n / sum(n
```

```
## # A tibble: 5 x 3
##   race_eth                  n    prop
##   <fct>                 <int>   <dbl>
## 1 Non-Hispanic Black     2560 0.681
## 2 Hispanic                682 0.182
## 3 Non-Hispanic White      459 0.122
## 4 Asian/Pacific Islander   32 0.00852
## 5 Other                    24 0.00639
```

6c) Report the average age, share male, and dimissal rate for each race/ethnicity category. Include the total sample size (all observations). Include the sample size for the dismissal variabe as well (just the number of non-NA observations).

```
race_eth_stats <- arrests.clean %>%
  group_by(race_eth) %>%
  summarise(
    n = n(),
    avg_age = mean(age, na.rm = TRUE),
    share_male = mean(male, na.rm = TRUE),
    dismissal_rate = mean(dismissal, na.rm = TRUE),
    n_dismissal = sum(!is.na(dismissal))
  )
race_eth_stats
```

```
## # A tibble: 6 x 6
##   race_eth                   n avg_age share_male dismissal_rate n_dismissal
##   <fct>                  <int>   <dbl>      <dbl>          <dbl>       <int>
## 1 Asian/Pacific Islander    32    28.9      0.938          0.636          11
## 2 Hispanic                 682    29.6      0.908          0.528         176
## 3 Non-Hispanic Black      2560    29.1      0.875          0.514        1117
## 4 Non-Hispanic White       459    29.7      0.898          0.587         276
## 5 Other                     24    28.3      0.833          0.444           9
## 6 <NA>                     454    27.3      0.621          0.720          93
```

6d) Describe any noteworthy findings from the table you presented in 6c.

Average age differs across race/ethnicity groups; some groups have a younger average age than others. The share male is high in most groups, consistent with fare evasion enforcement being concentrated among men. Where dismissal is observed (mainly LAS clients), dismissal rates vary by race/ethnicity; comparing those rates can inform discussion of differential outcomes. The total number of arrests in the table is 4211, while the number of observations with non-missing dismissal is 1682 (smaller because BDS has no dismissal variable).

7 Subway-station level analysis

7a) Create dummy variables for each race/ethnicity category and show summary statistics only for these dummy variables.

```
arrests.clean <- dummy_cols(arrests.clean, select_columns = "race_eth")
arrests.clean %>% summarise(across(starts_with("race_eth_"), ~ mean(., na.rm = TRUE)))
```

```
## # A tibble: 1 x 6
##   `race_eth_Asian/Pacific Islander` race_eth_Hispanic race_eth_Non-Hispanic Bl~1
##                               <dbl>             <dbl>                      <dbl>
## 1                           0.00852             0.182                      0.681
## # i abbreviated name: 1: `race_eth_Non-Hispanic Black`
## # i 3 more variables: `race_eth_Non-Hispanic White` <dbl>,
## #   race_eth_Other <dbl>, race_eth_NA <dbl>
```

7b) Aggregate to station-level observations and show a table with the top 10 stations by arrest totals, including the following information for each station:

- station name (loc2)
- station id
- total number of arrests at each station
- total number of arrests for each race_eth category at each station
- sort in descending order by total number of arrests
- remember to only show the top 10 stations
- use kable() in the knitr package for better formatting

```
arrests_stations <- arrests.clean %>%
  group_by(loc2) %>%
  summarise(
    st_id = first(st_id),
    n_arrests = n(),
    across(starts_with("race_eth_"), sum),
    .groups = "drop"
  ) %>%
  arrange(desc(n_arrests)) %>%
  head(10)
knitr::kable(arrests_stations)
```

loc2	st_id	n_arrests	race_eth_Asian/Pacific Islander	race_eth_Hispanic	race_eth_Non-Hispanic Black	race_eth_Non-Hispanic White	race_eth_Other	race_eth_NA
coney island-stillwell ave	66	223	NA	NA	NA	NA	NA	11
jay st - metrotech	99	198	NA	NA	NA	NA	NA	12
utica ave and fulton st	150	143	NA	NA	NA	NA	NA	7
utica ave and eastern parkway	70	142	NA	NA	NA	NA	NA	6
marcy ave j m z line	114	141	NA	NA	NA	NA	NA	7
nostrand ave and fulton st a c station	131	141	NA	NA	NA	NA	NA	7
canarsie rockaway pkwy	54	133	NA	NA	NA	NA	NA	6
sutter avenue station l line	147	102	NA	NA	NA	NA	NA	6
kingston - throop avs	106	90	NA	NA	NA	NA	NA	3
nevins st 2 3 4 5 lines	123	86	NA	NA	NA	NA	NA	5

7c) Aggregate to station-level observations (group by loc2), and show a table of stations with at least 50 arrests along with the following information:

- station name (loc2)

- station arrest total
- combined total number of Black and Hispanic arrests
- total number of arrests with race/ethnicity coded as NA
- share of arrests that are Black and Hispanic (excluding race_eth = NA from denominator)
- sorted in ascending order by Black and Hispanic arrest share
- remember to only show stations with at least 50 total arrests
- use kable() in the knitr package for better formatting

```
arrests_stations_top <- arrests.clean %>%
  group_by(loc2) %>%
  summarise(
    n_total = n(),
    n_bh = sum(race_eth == "Non-Hispanic Black" | race_eth == "Hispanic", na.rm = TRUE),
    n_na = sum(is.na(race_eth)),
    .groups = "drop"
  ) %>%
  mutate(denom = n_total - n_na, sh_bh = if_else(denom > 0, n_bh / denom, NA_real_)) %>%
  filter(n_total >= 50) %>%
  arrange(sh_bh) %>%
  select(loc2, n_bh, n_na, sh_bh, n_total)
knitr::kable(arrests_stations_top)
```

loc2	n_bh	n_na	sh_bh	n_total
marcy ave j m z line	97	7	0.7238806	141
myrtle av and broadway station	52	4	0.8000000	69
coney island-stillwell ave	171	11	0.8066038	223
graham ave l line	39	6	0.8125000	54
broadway and lorimer st j m station	56	2	0.8235294	70
clinton - washington avs station	48	5	0.8275862	63
jay st - metrotech	154	12	0.8279570	198
hoyt-schermerhorn a c g line	54	7	0.8437500	71
myrtle - willoughby avs g line	39	5	0.8666667	50
canarsie rockaway pkwy	113	6	0.8897638	133
nevins st 2 3 4 5 lines	74	5	0.9135802	86
kingston - throop avs	81	3	0.9310345	90
hoyt st 2 3	69	3	0.9324324	77
sutter avenue station l line	90	6	0.9375000	102
nostrand ave and fulton st a c station	126	7	0.9402985	141
utica ave and fulton st	129	7	0.9485294	143
livonia ave l line	68	4	0.9577465	75
junius st 3 line	70	2	0.9589041	75
utica ave and eastern parkway	131	6	0.9632353	142
court st r subway/borough hall 2 subway 3 subway 4 subway 5 subway	53	4	0.9636364	59
rockaway ave c line	56	4	0.9824561	61
sutter av - rutland rd 3 line	64	3	0.9846154	68
rockaway ave 3 line	56	5	1.0000000	61

7d) Briefly summarize any noteworthy findings from the table you just generated.

Stations with at least 50 arrests show a wide range in the share of arrests that are Black or Hispanic (the `sh_bh` column). Stations at the bottom of the sorted table have the *lowest* share Black/Hispanic; those

at the top have the *highest*. It is worth noting whether any stations with very high total arrest volume have a relatively low Black/Hispanic share (i.e., a relatively high share of Non-Hispanic White or other groups)—such stations would be especially relevant for the policy question. Conversely, stations with both high volume and a very high Black/Hispanic share suggest enforcement concentrated in certain areas and demographic groups. You can cite specific stations or ranges from the table using inline R if helpful.

8 (OPTIONAL) Visualize the distribution of arrests by race/ethnicity at stations with more than 100 arrests.

- Hint: see R code from class, section 8

```
arrests_stations_race_top <- arrests.clean %>%
  group_by(loc2) %>%
  mutate(st_arrests = n()) %>%
  ungroup() %>%
  group_by(loc2, race_eth) %>%
  summarise(arrests = n(), st_arrests = first(st_arrests), .groups = "drop") %>%
  arrange(desc(st_arrests)) %>%
  filter(st_arrests > 100)
arrests_stations_race_top %>%
  ggplot(aes(x = reorder(loc2, -st_arrests), y = arrests, fill = race_eth)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(x = "Station (loc2)", y = "Arrests", fill = "Race/ethnicity")
```

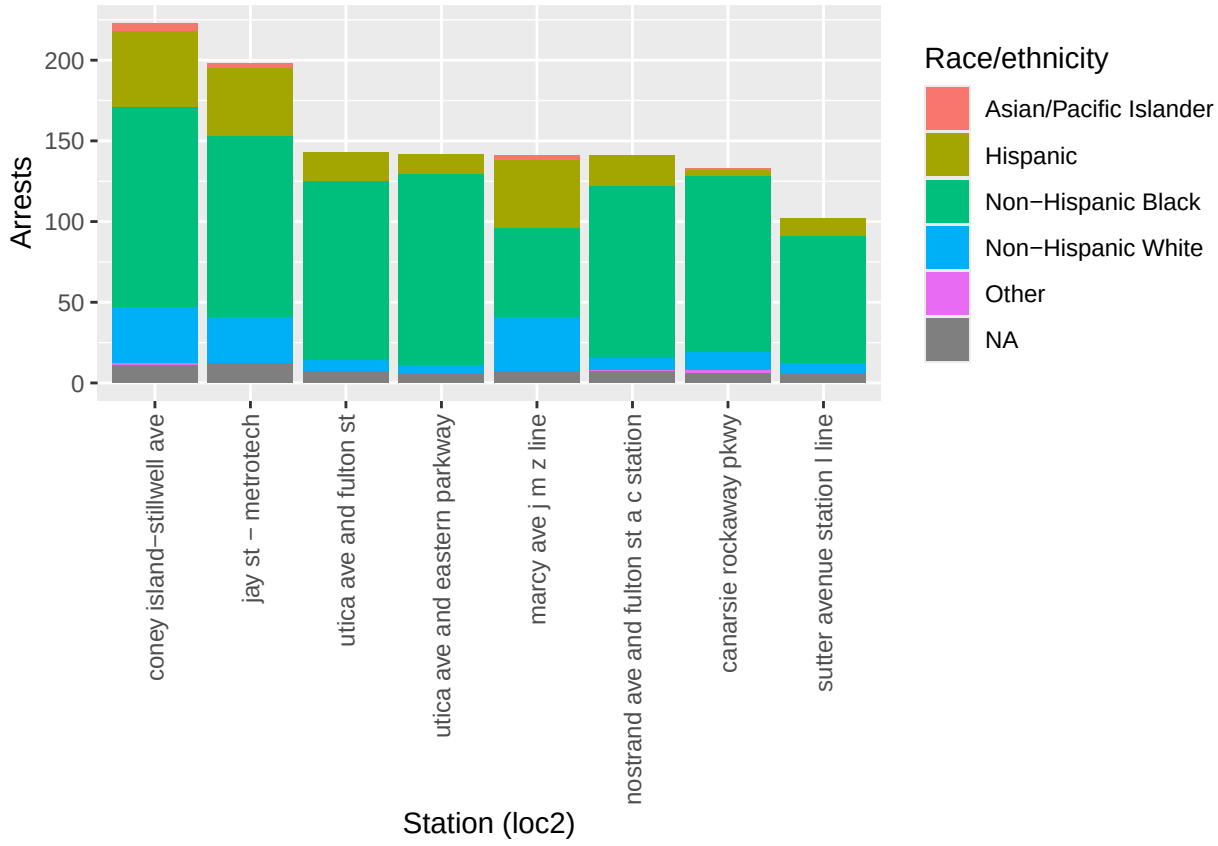

Figure 1: Race/ethnicity breakdown at stations with >100 arrests.