DSPC7514 Assignment 2: Subway Fare Evasion Microdata Analysis

Sample Solutions

2026-01-23

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.

```
# remember to make sure these packaged are installed before trying to load
library(tidyverse)
library(fastDummies)
```

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 ...
```

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 BDS data includes 2246 observations (client arrest records), and the LAS data includes another 1965 observations. Both datasets include basic demographic information on age, sex, race, ethnicity (coded differently in each dataset), as well as information on the location/subway station where the arrest occurred. The LAS data also includes information on case dismissal rates.

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?

In each raw dataset, the unit of observation is the arrested individual (client). On the surface the representative population is all individuals arrested by the NYPD for subway fare evasion in Brooklyn during 2016 who are represented by public defenders. If nearly all individuals arrested for fare evasion are represented by public defenders, then this sample comes close to the universe of subway fare evasion arrests in Brooklyn in 2016. This is difficult to argue convincingly without additional information, but is supported anecdotally by court observers.

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

```r
#recode race/ethnicity information from character to factors
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_flag) )
# compare race coding
summary(arrests_bds$race)
```

```
##                         0                Am Indian Asian/Pacific Islander
##                        35                        1                     21
##                     Black                    Other                Unknown
##                      1465                       32                      2
##                     White                     NA's
##                       533                      157
```

```r
summary(arrests_las$race)
```

```
## Asian or Pacific Islander                      Black                   Hispanic
##                        11                       1247                         21
##                    Latino                      Other                   Unknown
##                         2                         20                         10
##                     White                       NA's
##                       426                        228
```

```r
# compare Hispanic/ethnicity coding
summary(arrests_bds$ethnicity)
```

```
##            0     Hispanic Non-Hispanic        Other          NA's
##           33          493         1558            5           157
```

```r
summary(arrests_las$ethnicity)
```

```
##    N    Y NA's
## 1619  189  157
```

Race information is generally stored in one variable, Hispanic identity in a second variable. To work towards consistent variable names and coding in both datasets, let's first recode the raw race and ethnicity information into two separate columns of data (factors) named `race` and `ethnicity`.

While each dataset refers to similar race and ethnicity categories, there are different category names in each (including some slightly different spellings).

We also note that Hispanic identity factors into both race and Hispanic variables in the Legal Aid Society (LAS) data; in the BDS data, information on Hispanic identity is only included in the ethnicity variable.

Each dataset also contains a different set of values that seem to convey unknown race/ethnicity information, in addition to true missings (e.g. "0" and "Unknown" in addition to blank entries).

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

It's unclear what processes are used to code race and ethnicity at each public defender group. How much does the information reflect client self-identification rather than identity assigned by police and entered into arrest reports?

It's also important to emphasize what information this data does **not** include that might be relevant to the question of biased fare evasion enforcement:

- fare evasion that resulted in a summons (ticket + fine) rather than an arrest
- fare evasion enforcement on buses

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`).

```
# identify every combination of race-ethnicity in the raw data
table(arrests_bds$race,
      arrests_bds$ethnicity,
      useNA = "always")
```

```
##
##                             0 Hispanic Non-Hispanic Other <NA>
##    0                       31        1            3     0    0
##    Am Indian                0        0            1     0    0
##    Asian/Pacific Islander   0        0           21     0    0
##    Black                    2      104         1358     1    0
##    Other                    0       20           11     1    0
##    Unknown                  0        0            0     2    0
##    White                    0      368          164     1    0
##    <NA>                     0        0            0     0  157
```

```
# recode as factor in an internally consistent manner (address NAs, specify levels)
arrests_bds.clean <- arrests_bds %>%
  mutate(race_clean = recode(race,
                             "0" = NA_character_, # use NA_character
                             "Unknown" = NA_character_,
                             "Am Indian" = "Other")) %>%
  mutate(race_clean = fct_relevel(race_clean,
                                  "Asian/Pacific Islander",
                                  "Black",
                                  "White",
                                  "Other"))

# validation: confirm the recode worked as intended
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
```

```
table(arrests_bds.clean$race,
      arrests_bds.clean$race_clean,
      useNA = "always")
```

```
##
##                          Asian/Pacific Islander Black White Other <NA>
##    0                                          0     0     0     0   35
##    Am Indian                                  0     0     0     1    0
```

```
##    Asian/Pacific Islander                        21     0     0     0     0
##    Black                                          0  1465     0     0     0
##    Other                                          0     0     0    32     0
##    Unknown                                        0     0     0     0     2
##    White                                          0     0   533     0     0
##    <NA>                                           0     0     0     0   157
```

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

```
# ok now let's recode to Hispanic, Non-Hispanic, and NA
arrests_bds.clean <- arrests_bds.clean %>%
  mutate(hispanic = recode(ethnicity,
                           "0" = NA_character_,
                           "Other" = "Non-Hispanic"))

# validation: confirm the recode worked as intended
summary(arrests_bds.clean$hispanic)
```

```
##     Hispanic Non-Hispanic         NA's
##          493         1563          190
```

```
table(arrests_bds.clean$race_clean,
      arrests_bds.clean$hispanic,
      useNA = "always")
```

```
##
##                            Hispanic Non-Hispanic <NA>
##    Asian/Pacific Islander         0           21    0
##    Black                        104         1359    2
##    White                        368          165    0
##    Other                         20           13    0
##    <NA>                           1            5  188
```

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

```
# let's investigate a bit
table(arrests_bds.clean$race_clean,
      arrests_bds.clean$hispanic,
      useNA = "always")
```

```
##
##                            Hispanic Non-Hispanic <NA>
##    Asian/Pacific Islander         0           21    0
##    Black                        104         1359    2
##    White                        368          165    0
##    Other                         20           13    0
##    <NA>                           1            5  188
```

```
# create a single factor variable w/mutually exclusive groups, call it race_eth
# levels should be:
#   - Black, Non-Hispanic White, Hispanic, Asian/Pacific Islander, Other, NA
arrests_bds.clean <- arrests_bds.clean %>%
  mutate(race_eth = if_else(hispanic %in% "Hispanic",
                            hispanic,
                            race_clean)) %>%
```

```r
  mutate(race_eth = recode(race_eth,
                           "White" = "Non-Hispanic White",
                           "Black" = "Non-Hispanic Black"))


arrests_bds.clean <- arrests_bds.clean %>%
      mutate(race_eth = factor(race_eth,
                               levels = c("Asian/Pacific Islander",
                                          "Hispanic",
                                          "Non-Hispanic Black",
                                          "Non-Hispanic White",
                                          "Other")))

# validate results: joint distribution of race_eth and hispanic
table(arrests_bds.clean$race_eth,
      arrests_bds.clean$hispanic,
      useNA = "always")
```

```
##
##                          Hispanic Non-Hispanic <NA>
##   Asian/Pacific Islander        0           21    0
##   Hispanic                    493            0    0
##   Non-Hispanic Black            0         1359    2
##   Non-Hispanic White            0          165    0
##   Other                         0           13    0
##   <NA>                          0            5  188
```

Note that `race_eth` assigns individuals who identify as both Hispanic and a race other than white as Hispanic. This means, for example, that an individual who identifies as both Black and Hispanic appears as Hispanic in the `race_eth` column.

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

```
#inspect race/ethnicity coding in LAS data
table(arrests_las$las_race_key,
      arrests_las$hispanic_flag,
      useNA = "always")
```

```
##
##                                  N    Y <NA>
##   Asian or Pacific Islander   11    0    0
##   Black                     1201   46    0
##   Hispanic                    20    1    0
##   Latino                       2    0    0
##   Other                       11    9    0
##   Unknown                     10    0    0
##   White                      294  132    0
##   <NA>                        70    1  157
```

```
#generate race_eth column as a factor wth correct levels
arrests_las.clean <- arrests_las %>%
  mutate(race_eth = recode(las_race_key,
                          "Asian or Pacific Islander" = "Asian/Pacific Islander",
                          "Unknown" = NA_character_,
                          "Latino" = "Hispanic",
                          "White" = "Non-Hispanic White",
                          "Black" = "Non-Hispanic Black")) %>%
  mutate(race_eth = ifelse(hispanic_flag %in% "Y",
                          "Hispanic",
                          race_eth) ) %>%
  mutate(race_eth = factor(race_eth,
                          levels = c("Asian/Pacific Islander",
                                     "Hispanic",
                                     "Non-Hispanic Black",
                                     "Non-Hispanic White",
                                     "Other")))

#validate

  # show race_eth distribution
    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>                    237
## 4 Hispanic                211
```

```
## 5 Asian/Pacific Islander    11
## 6 Other                     11
```

```r
# show cross-tab between hispanic_flag and new race_eth variable
  table(arrests_las.clean$race_eth,
        arrests_las.clean$hispanic_flag,
        useNA = "always")
```

```
##
##                             N    Y <NA>
##   Asian/Pacific Islander   11    0    0
##   Hispanic                 22  189    0
##   Non-Hispanic Black     1201    0    0
##   Non-Hispanic White      294    0    0
##   Other                    11    0    0
##   <NA>                     80    0  157
```

```r
# show cross-tab between race and new race_eth variable
  table(arrests_las.clean$race_eth,
        arrests_las.clean$race,
        useNA = "always")
```

```
##
##                          Asian or Pacific Islander Black Hispanic Latino Other
##   Asian/Pacific Islander                        11     0        0      0     0
##   Hispanic                                       0    46       21      2     9
##   Non-Hispanic Black                             0  1201        0      0     0
##   Non-Hispanic White                             0     0        0      0     0
##   Other                                          0     0        0      0    11
##   <NA>                                           0     0        0      0     0
##
##                          Unknown White <NA>
##   Asian/Pacific Islander       0     0    0
##   Hispanic                     0   132    1
##   Non-Hispanic Black           0     0    0
##   Non-Hispanic White           0   294    0
##   Other                        0     0    0
##   <NA>                        10     0  227
```

5 Combining (appending) the BDS and LAS microdata

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

```r
arrests_bds.clean <- arrests_bds.clean %>% mutate(pd = "bds")
arrests_las.clean <- arrests_las.clean %>% mutate(pd = "las")
```

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

```r
arrests.clean <- bind_rows(arrests_las.clean,
                           arrests_bds.clean) %>%
  mutate(pd = as.factor(pd),
         st_id = as.factor(st_id),
```

```r
         loc2 = as.factor(loc2)) %>% # original station/location info is not continuous
   select(pd, race_eth, age, male, dismissal, st_id, loc2)
  summary(arrests.clean)
```

```
##    pd                         race_eth         age             male       
##  bds:2246   Asian/Pacific Islander:  32   Min.   : 0.00   Min.   :0.0000  
##  las:1965   Hispanic              : 704   1st Qu.:20.00   1st Qu.:1.0000  
##             Non-Hispanic Black    :2562   Median :26.00   Median :1.0000  
##             Non-Hispanic White    : 459   Mean   :29.18   Mean   :0.8748  
##             Other                 :  24   3rd Qu.:35.00   3rd Qu.:1.0000  
##             NA's                  : 430   Max.   :71.00   Max.   :1.0000  
##                                           NA's   :317     NA's   :314     
##    dismissal          st_id                                        loc2     
##  Min.   :0.0000   66     : 223   coney island-stillwell ave            : 223  
##  1st Qu.:0.0000   99     : 198   jay st - metrotech                    : 198  
##  Median :1.0000   150    : 143   utica ave and fulton st               : 143  
##  Mean   :0.5392   70     : 142   utica ave and eastern parkway         : 142  
##  3rd Qu.:1.0000   114    : 141   marcy ave j m z line                  : 141  
##  Max.   :1.0000   131    : 141   nostrand ave and fulton st a c station: 141  
##  NA's   :2529     (Other):3223   (Other)                               :3223  
```

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

The total number of subway fare evasion arrest records from both BDS and LAS is 4211.

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

```r
save(list = "arrests.clean",
     file = "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      2562
## 2 Hispanic                 704
## 3 Non-Hispanic White       459
## 4 <NA>                     430
## 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?

```
# including NAs
prop.table(table(arrests.clean$race_eth, useNA = "always")) %>%
  round(2) %>%
  as.data.frame() %>%
  arrange(desc(Freq)) %>%
  rename(race_eth = Var1)
```

```
##                 race_eth Freq
## 1     Non-Hispanic Black 0.61
## 2               Hispanic 0.17
## 3     Non-Hispanic White 0.11
## 4                   <NA> 0.10
## 5 Asian/Pacific Islander 0.01
## 6                  Other 0.01
```

```
# excluding NAs
prop.table(table(arrests.clean$race_eth)) %>%
  round(2) %>%
  as.data.frame() %>%
  arrange(desc(Freq)) %>%
  rename(race_eth = Var1)
```

```
##                 race_eth Freq
## 1     Non-Hispanic Black 0.68
## 2               Hispanic 0.19
## 3     Non-Hispanic White 0.12
## 4 Asian/Pacific Islander 0.01
## 5                  Other 0.01
```

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(),
```

```
            mean_age = mean(age, na.rm = TRUE),
            mean_male = mean(male, na.rm = TRUE),
            mean_dismissal = mean(dismissal, na.rm = TRUE),
            n_dismissal = sum(!is.na(dismissal)) )
race_eth_stats
```

```
## # A tibble: 6 x 6
##   race_eth                   n mean_age mean_male mean_dismissal n_dismissal
##   <fct>                  <int>    <dbl>     <dbl>          <dbl>       <int>
## 1 Asian/Pacific Islander    32     28.9     0.938          0.636          11
## 2 Hispanic                 704     29.7     0.901          0.538         197
## 3 Non-Hispanic Black      2562     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>                     430     26.0     0.603          0.75           72
```

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

Interestingly, arrested individuals with NA race/ethnicity are 3 to 4 years younger on average, and only 61% male compared to 83 to 94% male for those with race/ethnicity specified in the data.

The dismissal rate is also noticeably higher for API individuals, and lower for NA individuals. However, the sample sizes for these groups are very small by comparison, and the dismissal variable is only included in the LAS data so the samples sizes for that column are even smaller than for the other columns! With such a small number of observations for these groups it is very unlikely that we'd be able to conclude there are true differences in dismissal rates between API and NA individuals and other groups–we could do t-tests to check, more on that next week! Said another way, we can't rule out that the differences we see here are due to sampling variation, and thus should not be emphasizing them as findings at this point.

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(mean_black = round(mean(`race_eth_Non-Hispanic Black`, na.rm = TRUE), 2),
            mean_hisp = round(mean(race_eth_Hispanic, na.rm = TRUE), 2),
            mean_nhw  = round(mean(`race_eth_Non-Hispanic White`, na.rm = TRUE), 2),
            mean_api  = round(mean(`race_eth_Asian/Pacific Islander`, na.rm = TRUE), 2),
            mean_oth  = round(mean(race_eth_Other, na.rm = TRUE), 2),
            mean_NA   = round(mean(race_eth_NA, na.rm = TRUE), 2)  )
```

```
## # A tibble: 1 x 6
##   mean_black mean_hisp mean_nhw mean_api mean_oth mean_NA
##        <dbl>     <dbl>    <dbl>    <dbl>    <dbl>   <dbl>
## 1       0.68      0.19     0.12     0.01     0.01     0.1
```

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 = n(),
            n_black = sum(`race_eth_Non-Hispanic Black`, na.rm = TRUE),
            n_hisp  = sum(race_eth_Hispanic, na.rm = TRUE),
            n_api   = sum(`race_eth_Asian/Pacific Islander`, na.rm = TRUE),
            n_nhw   = sum(`race_eth_Non-Hispanic White`, na.rm = TRUE),
            n_oth   = sum(race_eth_Other, na.rm = TRUE) )   %>%
  arrange(desc(n))
knitr::kable(head(arrests_stations, n = 10))
```

loc2	st_id	n	n_black	n_hisp	n_api	n_nhw	n_oth
coney island-stillwell ave	66	223	124	48	5	35	1
jay st - metrotech	99	198	112	43	3	29	0
utica ave and fulton st	150	143	112	19	0	7	0
utica ave and eastern parkway	70	142	118	13	0	5	0
marcy ave j m z line	114	141	55	42	3	34	0
nostrand ave and fulton st a c station	131	141	107	20	0	7	1
canarsie rockaway pkwy	54	133	109	4	1	11	2
sutter avenue station l line	147	102	79	12	0	6	0
kingston - throop avs	106	90	69	12	0	6	0
nevins st 2 3 4 5 lines	123	86	63	11	0	6	1

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
- share of arrests that are Black and Hispanic (excluding race_eth = NA from denominator)
- sorted in ascending order above (of) 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(st_id = first(st_id),
            n = n(),
            n_black = sum(`race_eth_Non-Hispanic Black`, na.rm = TRUE),
            n_hisp  = sum(race_eth_Hispanic, na.rm = TRUE),
            n_bh    = sum(`race_eth_Non-Hispanic Black`, race_eth_Hispanic, na.rm = TRUE),
            n_na    = sum(race_eth_NA),
            sh_bh   = round(n_bh / (n - n_na), 2)) %>%
  select(loc2, n, n_bh, n_na, sh_bh) %>%
  filter(n >= 50) %>%
  arrange(sh_bh)
knitr::kable(arrests_stations_top)
```

loc2	n	n_bh	n_na	sh_bh
marcy ave j m z line	141	97	7	0.72
myrtle av and broadway station	69	53	3	0.80
coney island-stillwell ave	223	172	10	0.81
graham ave l line	54	39	6	0.81
broadway and lorimer st j m station	70	56	2	0.82
clinton - washington avs station	63	48	5	0.83
jay st - metrotech	198	155	11	0.83
hoyt-schermerhorn a c g line	71	55	6	0.85
myrtle - willoughby avs g line	50	39	5	0.87
canarsie rockaway pkwy	133	113	6	0.89
nevins st 2 3 4 5 lines	86	74	5	0.91
hoyt st 2 3	77	70	2	0.93
kingston - throop avs	90	81	3	0.93
nostrand ave and fulton st a c station	141	127	6	0.94
sutter avenue station l line	102	91	5	0.94
utica ave and fulton st	143	131	5	0.95
court st r subway/borough hall 2 subway 3 subway 4 subway 5 subway	59	53	4	0.96
junius st 3 line	75	70	2	0.96
livonia ave l line	75	69	3	0.96
utica ave and eastern parkway	142	131	6	0.96
rockaway ave c line	61	57	3	0.98
sutter av - rutland rd 3 line	68	64	3	0.98
rockaway ave 3 line	61	57	4	1.00

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

At every single high-arrest subway station, the majority of arrested individuals are Black or Hispanic. This isn't surprising, given that 87 percent of *all* arrested individuals with coded race/ethnicity are Black or Hispanic.

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

- Hint: see R code from class, section 8

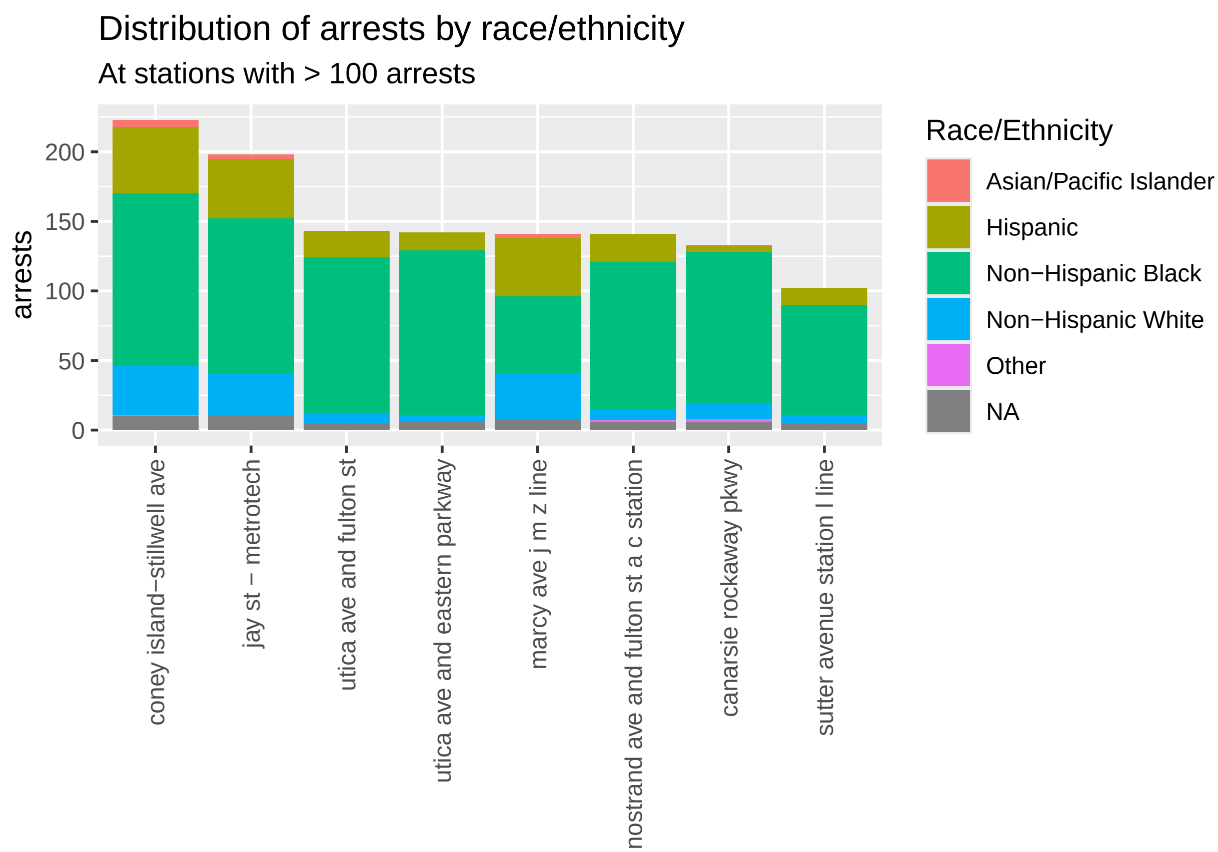
```
#get data frame with obs for every station-race_eth pairings on arrest counts
arrests_stations_race <- arrests.clean %>%
  group_by(loc2) %>%
  mutate(st_arrests = n()) %>%
  ungroup() %>%
  group_by(loc2, race_eth)  %>%
  summarise(arrests = n(), st_arrests = first(st_arrests)) %>%
  arrange(desc(st_arrests)) %>%
  filter(st_arrests > 100)

arrests_stations_race
```

```
## # A tibble: 39 x 4
## # Groups:   loc2 [8]
##    loc2                         race_eth               arrests st_arrests
##    <fct>                        <fct>                    <int>      <int>
##  1 coney island-stillwell ave Asian/Pacific Islander       5        223
##  2 coney island-stillwell ave Hispanic                    48        223
##  3 coney island-stillwell ave Non-Hispanic Black         124        223
##  4 coney island-stillwell ave Non-Hispanic White          35        223
##  5 coney island-stillwell ave Other                        1        223
##  6 coney island-stillwell ave <NA>                        10        223
##  7 jay st - metrotech         Asian/Pacific Islander       3        198
##  8 jay st - metrotech         Hispanic                    43        198
##  9 jay st - metrotech         Non-Hispanic Black         112        198
## 10 jay st - metrotech         Non-Hispanic White          29        198
## # i 29 more rows
```

```
g <- ggplot(arrests_stations_race,
       aes(x = reorder(loc2, -st_arrests),
           y = arrests, fill = race_eth)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "right",
        axis.title.x = element_blank(),
        axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 1))  +
  scale_fill_discrete(name = "Race/Ethnicity") +
  ggtitle("Distribution of arrests by race/ethnicity",
    subtitle = "At stations with > 100 arrests")

g
```




```
# alternative way, save plot and recall it
# ggsave(g, filename = "g.png")
# knitr::include_graphics("g.png")
```