U6614: Assignment 3

Philip Crane (plc2137)

2024-02-05

1 Load libraries.

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

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

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

- 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.

```
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 ...
```

2a) 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, comprised of public defender records from the Bronx Defenders (BDS) and Legal Aid Society (LAS), contains information on individuals arrested for subway fare evasion. The variables most relevant to us codify race, ethnicity, age, location, and dismissal status.

2b) 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 both datasets is the individual arrestee. The sample does not do a good job representing the population of interest, as it is limited to individuals who have been arrested for subway fare evasion and have sought public defender services, leaving many questions about the individuals who were not arrested, evaded fare on the bus, or hired a private attorney. This sample is likely to be biased, as it does not include individuals who were not arrested, as well as those who were arrested but did not seek public defender services. The data also has trouble definitively identifying the arrestee's race/ethnicity in a consistent way.

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

```r
summary(arrests_bds$race)
```

```
##    Length     Class      Mode 
##      2246 character character
```

```r
summary(arrests_bds$ethnicity)
```

```
##    Length     Class      Mode 
##      2246 character character
```

```r
summary(arrests_las$las_race_key)
```

```
##    Length     Class      Mode 
##      1965 character character
```

```r
summary(arrests_las$hispanic_flag)
```

```
##    Length     Class      Mode 
##      1965 character character
```

In the BDS dataset, both "race" and "ethnicity" are character variables. While "race" codifies conventional non-hispanic race categories, "ethnicity" only refers to a "Hispanic"/"Non-Hispanic" dichotomy.

"Race" indicates the same in the LAS data, while "hispanic_flag" serves as the "Hispanic"/"Non-Hispanic" identifier.

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

We don't know if these race/ethnicity identifiers are self-reported or the work of NYPD/public defender offices. This way of codifying what are often realistically blurry lines can distort the narrative. It is also important to consider how the "NA" indicator is used, whether in instances of no response from the arrestee or ambiguity by the police, both have implications on how reflective the data is to reality.

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

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

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

```r
arrests_bds <- arrests_bds %>%
    mutate(race = as.factor(race))

arrests_bds.clean <- arrests_bds %>%
      mutate(race_clean = recode(race, "0" = "NA",
                                       "Unknown" = "NA",
                                       "Am Indian" = "Other" ) )%>%
      mutate(race_clean = factor(race_clean,
                                 levels = c("Black", "White", "Asian/Pacific Islander", "Other")))
```

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

```r
arrests_bds <- arrests_bds %>%
    mutate(ethnicity = as.factor(ethnicity))
```

```
arrests_bds.clean <- arrests_bds.clean %>%
      mutate(ethnicity_clean = recode(ethnicity,
                                 'O' = 'NA',
                                 'Other' = 'Non-Hispanic')) %>%
      mutate(ethnicity_clean = factor(ethnicity_clean,
                                 levels = c('Hispanic', 'Non-Hispanic')))
```

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

```
arrests_bds.clean <- arrests_bds.clean %>%
      mutate(race_clean_char = as.character(race_clean),
             ethnicity_clean_char   = as.character(ethnicity_clean)) %>%
      mutate(race_eth = ifelse(ethnicity_clean_char %in% "Hispanic",
                                 ethnicity_clean_char,
                                 race_clean_char) ) %>%
      mutate(race_eth = as.factor(recode(race_eth,
                                           "White" = "Non-Hispanic White",
                                           "Black" = "Non-Hispanic Black"))) %>%
      select(-race_clean_char, -ethnicity_clean_char)
```

4 Clean LAS race and ethnicity data

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

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

arrests_las.clean <- arrests_las %>%
  mutate(race_clean = recode(las_race_key,
                             "Asian or Pacific Islander" = "Asian/Pacific Islander",
                             "Unknown" = "NA",
                             "Latino" = "Hispanic",
                             "White" = "Non-Hispanic White",
                             "Black" = "Non-Hispanic Black")) %>%
  mutate(race_eth = ifelse(hispanic_flag %in% "Y", "Hispanic", race_clean)) %>%
  mutate(race_eth = factor(race_eth, levels = c("Non-Hispanic Black",
                                                 "Hispanic",
                                                 "Non-Hispanic White",
                                                 "Asian/Pacific Islander",
                                                 "Other")))
```

NOTE: you may be able to do everything in a single pipe, depending on your approach (but you certainly don't have to).

5 Combining (appending) the BDS and LAS microdata

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

```
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.

```
arrests.clean <- bind_rows(arrests_bds.clean, arrests_las.clean) %>%
    mutate(pd = as.factor(pd),
           st_id = as.factor(st_id),
           loc2 = as.factor(loc2)) %>%
    select(pd, race_eth, age, male, st_id, loc2, dismissal)
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     
##      st_id                                      loc2        dismissal     
##  66     : 223   coney island-stillwell ave           : 223   Min.   :0.0000  
##  99     : 198   jay st - metrotech                   : 198   1st Qu.:0.0000  
##  150    : 143   utica ave and fulton st              : 143   Median :1.0000  
##  70     : 142   utica ave and eastern parkway        : 142   Mean   :0.5392  
##  114    : 141   marcy ave j m z line                 : 141   3rd Qu.:1.0000  
##  131    : 141   nostrand ave and fulton st a c station: 141   Max.   :1.0000  
##  (Other):3223   (Other)                              :3223   NA's   :2529    
```

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

The total number of subway fare evasion arrest records is 4211.

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

```
#save(list="arrests.clean",
 #    file = "C:\\Users\\philc\\OneDrive\\Desktop\\Spring
  #          2024\\R\\Lectures\\Lecture4\\arrests.clean.RData")
##Commenting this out as it caused issues when knitting##
```

6 Descriptive statistics by race/ethnicity

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

```
table(arrests.clean$race_eth, useNA = "always")
```

```
##
## Asian/Pacific Islander               Hispanic     Non-Hispanic Black
##                     32                    704                   2562
##     Non-Hispanic White                  Other                   <NA>
##                    459                     24                    430
```

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

```
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
```

```
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
```

Excluding NAs adds additional weight to three categories: "Hispanic", "Non-Hispanic Black", and "Non-Hispanic White". "Non-Hispanic Black" is impacted the most, increasing from 0.61 to 0.68 when NAs are excluded.

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

```
race_eth_stats <- arrests.clean %>%
  group_by(race_eth) %>%
  summarise(mean_age = mean(age, na.rm = TRUE),
            share_male = mean(male, na.rm = TRUE),
            dismissal_rate = mean(dismissal, na.rm = TRUE),
            total_n = n(),
            dismissal_n = n_distinct(dismissal, na.rm = TRUE))

print(race_eth_stats)
```

```
## # A tibble: 6 x 6
##   race_eth               mean_age share_male dismissal_rate total_n dismissal_n
##   <fct>                     <dbl>      <dbl>          <dbl>   <int>       <int>
## 1 Asian/Pacific Islander     28.9      0.938          0.636      32           2
## 2 Hispanic                   29.7      0.901          0.538     704           2
## 3 Non-Hispanic Black         29.1      0.875          0.514    2562           2
## 4 Non-Hispanic White         29.7      0.898          0.587     459           2
## 5 Other                      28.3      0.833          0.444      24           2
## 6 <NA>                       26.0      0.603          0.75      430           2
```

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

While the average age of arrestees is relatively uniform across all race_eth catagories (28.3 to 29.7), the arrest rate of Non-Hispanic Black individuals is notably higher than the other categories at 2562, while the second highest, Hispanic, is 704.

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 = mean(`race_eth_Non-Hispanic Black`, na.rm = TRUE),
            mean_hispanic = mean(`race_eth_Hispanic`, na.rm = TRUE),
            mean_white = mean(`race_eth_Non-Hispanic White`, na.rm = TRUE),
            mean_asianpi = mean(`race_eth_Asian/Pacific Islander`, na.rm = TRUE),
            mean_other = mean(`race_eth_Other`, na.rm = TRUE))
```

```
## # A tibble: 1 x 5
##   mean_black mean_hispanic mean_white mean_asianpi mean_other
##        <dbl>         <dbl>      <dbl>        <dbl>      <dbl>
## 1      0.678         0.186      0.121      0.00846    0.00635
```

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

```r
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
- 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

```r
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_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),
            n_na    = sum(`race_eth_NA`, na.rm = TRUE)) %>%
  mutate(n_bh = n_black + n_hisp,
         new = 1-n_bh,
         n_bh = sum(n_bh, na.rm = TRUE),
         share_bh = n_bh / (n - n_na)) %>%
  filter(n >= 50) %>%
  arrange(share_bh)

knitr::kable(arrests_stations_top)
```

loc2	st_id	n	n_black	n_hisp	n_api	n_nhw	n_oth	n_na	n_bh	new	share_bh
coney island-stillwell ave	66	223	124	48	5	35	1	10	3266	-171	15.33333
jay st - metrotech	99	198	112	43	3	29	0	11	3266	-154	17.46524
utica ave and fulton st	150	143	112	19	0	7	0	5	3266	-130	23.66667
utica ave and eastern parkway	70	142	118	13	0	5	0	6	3266	-130	24.01471
nostrand ave and fulton st a c station	131	141	107	20	0	7	1	6	3266	-126	24.19259
marcy ave j m z line	114	141	55	42	3	34	0	7	3266	-96	24.37313
canarsie rockaway pkwy	54	133	109	4	1	11	2	6	3266	-112	25.71654
sutter avenue station l line	147	102	79	12	0	6	0	5	3266	-90	33.67010
kingston - throop avs	106	90	69	12	0	6	0	3	3266	-80	37.54023
nevins st 2 3 4 5 lines	123	86	63	11	0	6	1	5	3266	-73	40.32099
hoyt st 2 3	97	77	58	12	0	5	0	2	3266	-69	43.54667
junius st 3 line	101	75	60	10	1	2	0	2	3266	-69	44.73973
livonia ave l line	111	75	56	13	0	3	0	3	3266	-68	45.36111
broadway and lorimer st j m station	112	70	34	22	0	11	1	2	3266	-55	48.02941
myrtle av and broadway station	117	69	38	15	0	13	0	3	3266	-52	49.48485
hoyt-schermerhorn a c g line	98	71	46	9	0	10	0	6	3266	-54	50.24615
sutter av - rutland rd 3 line	148	68	61	3	0	0	1	3	3266	-63	50.24615
clinton - washington avs station	64	63	42	6	0	10	0	5	3266	-47	56.31034

loc2	st_id	n	n_black	n_hisp	n_api	n_nhw	n_oth	n_na	n_bh	new	share_bh
rockaway ave c line	141	61	50	7	0	1	0	3	3266	-56	56.31034
rockaway ave 3 line	140	61	49	8	0	0	0	4	3266	-56	57.29825
court st r subway/borough hall 2 subway 3 subway 4 subway 5 subway	68	59	42	11	0	2	0	4	3266	-52	59.38182
graham ave l line	88	54	28	11	0	9	0	6	3266	-38	68.04167
myrtle - willoughby avs g line	118	50	27	12	0	5	1	5	3266	-38	72.57778

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

My code here is not correct–spent a lot of time on this and simply cannot figure this out on my own. It is impossible to make any inferences based on the combined Black/Hispanic variable as the number makes no sense. What is illustrated, however is that the stations with the most arrests disproportionately lie outside of Manhattan. This suggests that NYPD are far more likely to arrest for fare evasion in Brooklyn, or at least those arrested outside of Manhattan tend to use public defender services more so than those in Manhattan.

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