DSPC 7514 Assignment 3: Subway Fare Evasion Microdata Analysis

Your Name (your-uni)

2025-09-13

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.

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.

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.

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?

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

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

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`).

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

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

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

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.

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

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

6 Descriptive statistics by race/ethnicity

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

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

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).

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

7 Subway-station level analysis

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

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

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

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

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