

U6614: Assignment 3: Subway Fare Evasion Microdata

Your Name (your-uni)

2020-09-21

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on Monday, September 28th.

Before knitting your rmd file as a pdf, you will need to install TinyTex for Latex distribution by running the following code:

```
tinytex::install_tinytex()
```

Please visit [this](#) link for more information on TinyTex installation.

1 Load libraries

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

- 2.1 For each dataset, what is the unit of observation and representative population?
- 2.2 Inspect and describe the coding of race/ethnicity in each dataset.
- 2.3 From the outset, are there any data limitations you think are important to note?

3 Clean BDS race and ethnicity data

- 3.1 BDS: race data (generate column `race_clean`).
- 3.2 BDS: ethnicity data (generate column `ethnicity_clean`).
- 3.3 Generate a single race/ethnicity factor variable `race_eth` with mutually exclusive categories.

4 Clean LAS race and ethnicity data

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

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

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

5.2 Append `arrests_bds.clean` and `arrests_las.clean` using `rbind()`. Store as new data frame `arrests_all` and inspect for consistency/accuracy.

5.3 What is the total number of subway fare evasion arrest records?

5.4 Export `arrests_all` as `.csv`, and save as `.rds` file.

6 Descriptive statistics by race/ethnicity

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

6.2 Print the proportion of total arrests for each race/ethnicity category.

6.3 Show the average age, share male, and dismissal rate for each race/ethnicity category. Describe any noteworthy findings.

7 Subway-station level analysis

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

7.2 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 Black and Hispanic arrest share*
- *remember to only show stations with at least 50 total arrests*

- 7.3 Briefly summarize any noteworthy findings from the table you just generated.
- 7.4 (OPTIONAL) Visualize the distribution of arrests by race/ethnicity at stations with > 100 arrests.

Hint: see R code from class, section 8