Data Science for Policy IA7514 Assignment 3

Subway Fare Evasion Arrests: Exploring Racial Disparities

your uni here

2026-01-31

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

Load libraries

1 Aggregating to subway station-level arrest totals

1a) Load full set of cleaned arrest microdata (arrests.clean.rdata).

1b) Using tidyverse functions, create a new data frame (`st_arrests`) that aggregates the microdata to station-level observations. For `st_arrests`, the unit of analysis should be the station, with columns for `st_id`, `loc2` and `total arrests`.

1c) Plot histogram of arrests and briefly describe the distribution of arrests across stations.

2 Joining subway ridership and neighborhood demographic data and prepping data for analysis.

2a) Read in poverty and ridership csv files with strings as factors (`station_povdataclean_2016.csv` and `Subway Ridership by Station - BK.csv`).

2b) Join both data frames from 3a to `st_arrests` and inspect results (store new data frame as `st_joined`).

- Inspect results from joins, drop unnecessary ridership columns ("swipes") from the ridership data, and group `st_joined` by `st_id` and `mta_name`.
- Only display ungrouped version of `st_joined` for compactness.

2c) Print the top 10 stations by total arrest counts. Only display `st_id`, `mta_name`, `arrests_all`, `shareblack`, `povrt_all_2016` (no other columns). Round percentages to 2 decimal points for this question and all subsequent questions.

- For better looking tables, we recommend passing your table into the `kable()` function from the `knitr` package. Just add `%>% kable()` at the end of your pipe.

2d) Compute arrest intensity and other explanatory variables for analysis.

- Drop the observation for the Coney Island station and very briefly explain your logic
- Create new column of data for the following:
  - fare evasion arrest intensity: `arrperswipe_2016` = arrests per 100,000 ridership ('swipes')
  - a dummy indicating if a station is high poverty: `highpov` = 1 if pov rate is > median pov rate across all Brooklyn station areas
  - a dummy for majority Black station areas: `nblack` = 1 if `shareblack` > 0.5
- Coerce new dummy variables into factors with category labels
- Assign results to new data frame called `stations`
- Display top 10 stations by arrest intensity using `kable()` in the `knitr` package

2e) How do the top 10 stations by arrest intensity compare to the top 10 stations by arrest count?

3 Explore relationship between arrest intensity and poverty rates across subway station areas.

3a) Examine the relationship between arrest intensity and poverty rates

- Show a scatterplot of arrest intensity vs. poverty rates along with your preferred regression line (linear or quadratic, not both!). Weight observations by ridership, and label your axes appropriately. **Only show one plot with your preferred specification!**
- Which regression specification do you prefer: linear or quadratic? Be clear about your logic and cite statistical evidence to support your decision.
- Interpret your preferred regression specification (carefully!). Remember to test for statistical significance for any estimates you choose to emphasize.

3b) Estimate and test the difference in mean arrest intensity between high/low poverty areas

- Report difference and assess statistical significance
- Weight observations by ridership

4 How does neighborhood racial composition (`nblack`) moderate the relationship between poverty and arrest intensity?

4a) Present a table showing the difference in mean arrests intensity for each of the four groups defined by the interaction of `highpov` and `nblack`. Remember to weight by ridership.

- HINT: use `group_by()` and `summarise()`
- BONUS: can you report this information in a 2x2 table?

4b) Does the difference in mean arrest intensity between high-poverty majority Black and high-poverty majority non-Black stations appear to be explained by differences in the mean poverty rate?

- **Step 1**: calculate the difference in mean arrest intensity between high poverty Majority Black and Majority Non-Black station areas. Make sure to calculate statistics weighted by ridership. Test whether this difference is statistically significant.
- **Step 2**: Repeat above steps for the poverty rate instead of arrest intensity.
- **Step 3**: Don't forget to answer the question above in your own words!

4c) Present and interpret a scatterplot of arrest intensity vs. poverty rates broken down by majority Black vs. majority non-Black (`nblack`), including different regression lines for each group of stations.

- use separate aesthetics for Black and non-Black station areas
- include the regression lines that you think best capture this relationship:
- show linear or quadratic specifications (not both!)
- weight observations by ridership, and label your axes appropriately
- remember to carefully interpret your preferred regression specification (carefully!)

4d) BONUS: Next let's let's think about how measurement error might impact results from 4c. Do you think measurement error could bias your estimates of neighborhood racial gaps in the effect of poverty on enforcement intensity from 4c? Explain, carefully. Do you have any creative ideas to address any concerns you have about potential bias due to measurement error?

- One source of measurement error owes to the fact that we're using racial-ethnic composition and poverty rates for the neighborhood surrounding each station to proxy for characteristics of riders at each station. These variables are measured with *non-random* error; demographic measures for the surrounding neighborhood will tend to be a less accurate proxy for the demographics of riders at that station for busier stations that are destinations for commuters, tourists and others who may not live in very vicinity close to the station.
- Tip: this is a very tricky issue! In order to think through the measurement error problem and it's consequences you will probably want to consult your Quant II notes and/or my Quant II video lecture 4 on the course website.
- Can you think of any other measurement error problems that might affect your results from 5b?
- Do you have any creative ideas for addressing any concerns you have about potential bias due to this source of measurement error, using this data or other data you think might exist?

We will discuss your answers and the issue of measurement error during class.

5 Is the differential effect of poverty in majority Black station areas explained by differences between stations in crime?

One determinant of fare evasion enforcement is police presence: when more police are present, the greater the chances they will encounter fare evasion. Moreover, the NYPD has often claimed they go where the crime is.

In the absence of data on police deployment across the subway system, we can use the number of crimes as a proxy for police presence.

5a) Load `nypd_criminalcomplaints_2016.csv` and join to stations by `st_id` and `mta_name`.

5b) Are there more crimes reported in high-poverty Majority Black station areas than in high-poverty Majority non-Black station areas? Report the difference in crimes and assess statistically significance.

5c) Does the difference in crimes that you found in 5b explain the finding from 4c that poverty has a stronger positive effect on arrest intensity in majority Black station areas than in majority non-Black station areas?

- start with your preferred specification from 4c
- next, control for the the number of crimes and see if the conclusions from 4c change
- do your conclusions change if you consider different functional forms for the relationship between crime and arrest intensity?

6. Summarize and interpret your findings with respect to racial disparities in subway fare evasion arrest intensity. Be very careful about how you frame and justify any claims of racial bias; any such claims should be supported by the analysis you present.

- Is there any additional analysis you'd like to do with the data at hand?
- Are there any key limitations to the data and/or analysis affecting your ability to examine racial disparities in enforcement?
- Is there any additional data you'd like to see that would help strengthen your analysis and interpretation?
- For this question, try to be very specific and avoid vaguely worded concerns.