Data Science for Policy IA7514 Assignment 3

Subway Fare Evasion Arrests: Exploring Racial Disparities

your uni here

2026-02-09

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

```
# Install once in R Console if needed (do not run when knitting):
# install.packages(c("tidyverse", "knitr", "estimatr", "ggpmisc", "modelsummary"))
library(tidyverse)
library(knitr)
library(estimatr)
library(ggpmisc)
library(modelsummary)
```

1 Aggregating to subway station-level arrest totals

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

```
getwd()
```

```
## [1] "/Users/arnavsahai/Desktop/Data Analysis for Policy Research using R/Lecture 3"
```

```
load("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`.

```
st_arrests <- arrests.clean %>%
  group_by(st_id, loc2) %>%
  summarise(arrests_all = n()) %>%
  arrange(desc(arrests_all)) %>%
  ungroup()
```

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

```
ggplot(data = st_arrests, aes(x = arrests_all)) +
  geom_histogram(bins = 20) +
  labs(x = "Total arrests per station", y = "Count of stations",
       title = "Distribution of arrests across stations")
```

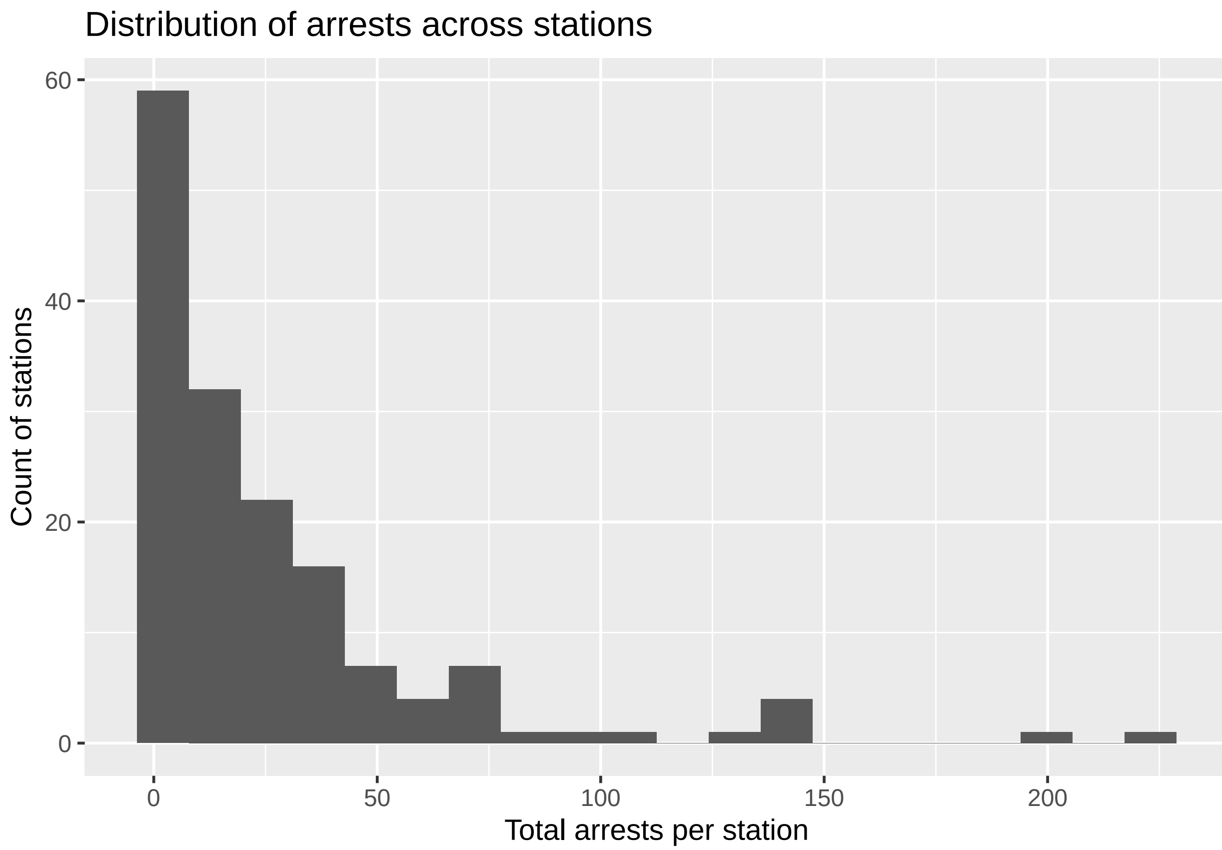

Figure 1: Distribution of arrests across stations

The distribution of arrests across stations is right-skewed: most stations have relatively few arrests, while a smaller number of stations have high arrest counts. The bulk of the mass is at low-to-moderate arrest totals.

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`).

```
st_poverty <- read.csv("station_povdataclean_2016.csv", stringsAsFactors = TRUE) %>%
  select(-nblack, -mta_name)

st_ridership <- read.csv("Subway Ridership by Station - BK.csv", stringsAsFactors = TRUE)
```

2b) Join both data frames from 2a 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.

```
# Match types: st_id must be integer for join with poverty/ridership
st_arrests <- st_arrests %>% mutate(st_id = as.integer(st_id))

drop_vars <- c("swipes2011", "swipes2012", "swipes2013", "swipes2014", "swipes2015")

st_joined <- st_arrests %>%
  inner_join(st_poverty, by = "st_id") %>%
  inner_join(st_ridership, by = "st_id") %>%
  select(-all_of(drop_vars)) %>%
  group_by(st_id, mta_name)

# Display ungrouped for compactness (e.g. first rows)
ungroup(st_joined) %>% head(10) %>% kable()
```

st_id	loc2	arrests_all	x	y	pop_black_2016	pov_black_2016	pop_2016	pov_2016	povrt_all_2016	shareblack	mta_name	swipes2016
66	coney island-stillwell ave	223	-73.98174	40.57810	36	2	5186	1329	0.2562669	0.0069418	Coney Island-Stillwell Av D subway F subway N subway Q subway	5025598
99	jay st - metrotech	198	-73.98726	40.69235	1939	677	12437	1939	0.1559058	0.1559058	Jay St-MetroTech A subway C subway F subway R subway	13091255
150	utica ave and fulton st	143	-73.92907	40.67914	14825	4592	18556	6149	0.3313753	0.7989330	Utica Av A subway C subway	5152649
70	utica ave and eastern parkway	142	-73.93113	40.66876	13135	3796	17561	5565	0.3168954	0.7479643	Crown Heights-Utica Av 3 subway 4 subway	9051970

st_id	loc2	arrests_all	x	y	pop_black_2016	pov_black_2016	pop_all_2016	pov_all_2016	povrt_all_2016	shareblack	mta_name	swipes2016
114	marcy ave j m z line	141	-73.95791	40.70855	5642	483	23711	9182	0.3872464	0.0650331	Marcy Av J subway M subway Z subway	4272443
131	nostrand ave and fulton st a c station	141	-73.94979	40.68053	15311	2437	15934	3511	0.2203464	0.6471068	Nostrand Av A subway C subway	5861658
54	canarsie rockaway pkwy	133	-73.90321	40.64556	5624	900	6753	1156	0.1711803	0.8328150	Canarsie-Rockaway Pkwy L subway	3897784
147	sutter avenue station l line	102	-73.90198	40.66917	13804	6706	15751	9104	0.5779951	0.7494527	Sutter Av L subway	1435112
106	kingston - throop avs	90	-73.94086	40.67962	19176	3832	20610	4809	0.2333333	0.7848617	Kingston-Throop Avs C subway	2031710
123	nevins st 2 3 4 5 lines	86	-73.98049	40.68826	2698	306	13654	1221	0.0894204	0.1975978	Nevins St 2 subway 3 subway 4 subway 5 subway	3220745

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.

```
st_joined %>%
  mutate(shareblack = round(shareblack, 2), povrt_all_2016 = round(povrt_all_2016, 2)) %>%
  arrange(desc(arrests_all)) %>%
  select(st_id, mta_name, arrests_all, shareblack, povrt_all_2016) %>%
  head(10) %>%
  kable()
```

st_id	mta_name	arrests_all	shareblack	povrt_all_2016
66	Coney Island-Stillwell Av D subway F subway N subway Q subway	223	0.01	0.26
99	Jay St-MetroTech A subway C subway F subway R subway	198	0.16	0.16
150	Utica Av A subway C subway	143	0.80	0.33
70	Crown Heights-Utica Av 3 subway 4 subway	142	0.75	0.32
114	Marcy Av J subway M subway Z subway	141	0.07	0.39
131	Nostrand Av A subway C subway	141	0.65	0.22
54	Canarsie-Rockaway Pkwy L subway	133	0.83	0.17
147	Sutter Av L subway	102	0.75	0.58
106	Kingston-Throop Avs C subway	90	0.78	0.23
123	Nevins St 2 subway 3 subway 4 subway 5 subway	86	0.20	0.09

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

```r
stations <- st_joined %>%
  mutate(arrperswipe = round(arrests_all / (swipes2016 / 100000), 2),
         highpov = as.numeric(povrt_all_2016 > median(st_joined$povrt_all_2016)),
         nblack = as.numeric(shareblack > 0.5),
         shareblack = round(shareblack, 2),
         povrt_all_2016 = round(povrt_all_2016, 2)) %>%
  mutate(highpov = factor(highpov, levels = c(0, 1),
                          labels = c("Not high poverty", "High poverty")),
         nblack = factor(nblack, levels = c(0, 1),
                          labels = c("Majority non-Black", "Majority Black"))) %>%
  filter(st_id != 66)

# Top 10 by arrest intensity
stations %>%
  arrange(desc(arrperswipe)) %>%
  select(st_id, mta_name, arrperswipe, arrests_all, shareblack, povrt_all_2016, highpov, nblack) %>%
  head(10) %>%
  kable()
```

st_id	mta_name	arrperswipe	arrests_all	shareblack	povrt_all_2016	highpov	nblack
101	Junius St 3 subway	11.00	75	0.78	0.48	High poverty	Majority Black
26	Atlantic Av L subway	8.48	37	0.66	0.51	High poverty	Majority Black
111	Livonia Av L subway	7.17	75	0.83	0.45	High poverty	Majority Black
147	Sutter Av L subway	7.11	102	0.75	0.58	High poverty	Majority Black
106	Kingston-Throop Avs C subway	4.43	90	0.78	0.23	High poverty	Majority Black
112	Lorimer St J subway M subway	4.39	70	0.15	0.34	High poverty	Majority non-Black
140	Rockaway Av 3 subway	3.97	61	0.78	0.40	High poverty	Majority Black
54	Canarsie-Rockaway Pkwy L subway	3.41	133	0.83	0.17	Not high poverty	Majority Black
141	Rockaway Av C subway	3.41	61	0.80	0.22	Not high poverty	Majority Black

st_id	mta_name	arrperswipe	arrests_all	shareblack	povrt_all_2016	highpov	nblack
144	Shepherd Av C subway	3.40	36	0.61	0.30	High poverty	Majority Black

Coney Island (st_id 66) is dropped because it is an extreme outlier: very high ridership (tourist destination) that is unrepresentative of typical station-area demographics and enforcement, and would distort arrest intensity and its relationship with poverty and race.

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

```r
top10_count <- st_joined %>%
  ungroup() %>%
  mutate(shareblack = round(shareblack, 2), povrt_all_2016 = round(povrt_all_2016, 2)) %>%
  arrange(desc(arrests_all)) %>%
  select(st_id, mta_name, arrests_all, shareblack, povrt_all_2016) %>%
  head(10)

top10_intensity <- stations %>%
  ungroup() %>%
  arrange(desc(arrperswipe)) %>%
  select(st_id, mta_name, arrperswipe, arrests_all, shareblack, povrt_all_2016) %>%
  head(10)

# Overlap: stations in both top-10 lists
inner_join(top10_count %>% select(st_id), top10_intensity %>% select(st_id), by = "st_id")
```

```
## # A tibble: 3 x 1
##   st_id
##   <int>
## 1    54
## 2   147
## 3   106
```

The top 10 by arrest count are mostly high-ridership stations, so raw arrests can be high even when intensity (arrests per 100k swipes) is moderate. The top 10 by arrest intensity are stations with high arrests relative to ridership—often lower-ridership or higher-enforcement stations. There can be limited overlap: high-count stations are not necessarily high-intensity, and vice versa.

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.

```r
ggplot(stations, aes(x = povrt_all_2016, y = arrperswipe, weight = swipes2016)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x) +
  labs(x = "Poverty rate", y = "Arrests per 100,000 ridership",
       title = "Fare evasion arrest intensity vs. poverty rate")
```

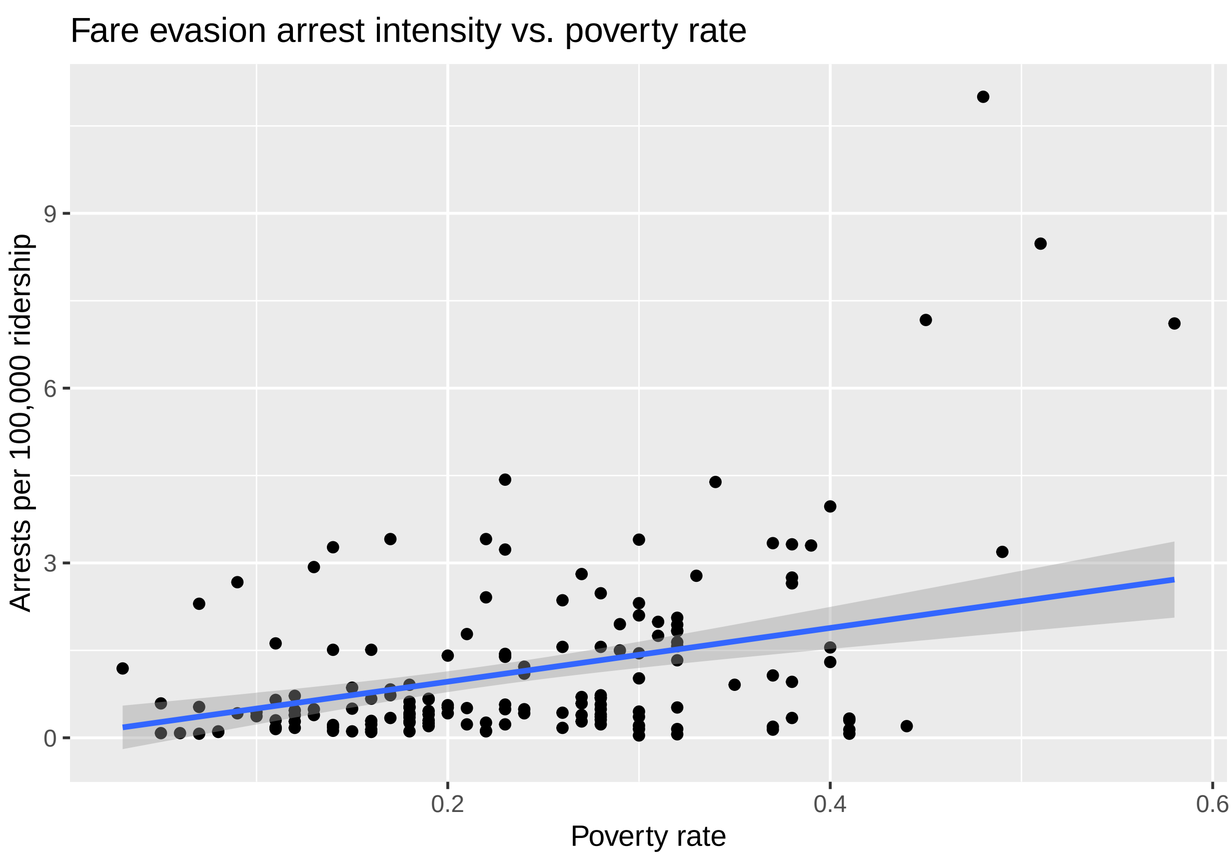

Figure 2: Arrest intensity vs. poverty rate (weighted by ridership)

```r
# Linear (preferred for simplicity and interpretability; quadratic can be tested for improvement)
m_3a_l <- lm_robust(arrperswipe ~ povrt_all_2016, data = stations, weight = swipes2016)
m_3a_q <- lm_robust(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
                    data = stations, weight = swipes2016)
summary(m_3a_l)
```

```
##
## Call:
## lm_robust(formula = arrperswipe ~ povrt_all_2016, data = stations,
##     weights = swipes2016)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)     0.04023     0.2445  0.1646 0.8695097  -0.4428   0.5232 154
## povrt_all_2016  4.61213     1.1940  3.8627 0.0001648   2.2534   6.9709 154
##
## Multiple R-squared:  0.1507 ,    Adjusted R-squared:  0.1452
## F-statistic: 14.92 on 1 and 154 DF,  p-value: 0.0001648
```

```r
summary(m_3a_q)
```

```
##
## Call:
## lm_robust(formula = arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
##     data = stations, weights = swipes2016)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)            1.286      0.416   3.093 0.002359   0.4646   2.1083 153
## povrt_all_2016        -8.144      4.221  -1.929 0.055521 -16.4828   0.1946 153
## I(povrt_all_2016^2)   26.632      9.613   2.771 0.006290   7.6417  45.6225 153
##
## Multiple R-squared:  0.2249 ,    Adjusted R-squared:  0.2148
## F-statistic: 9.575 on 2 and 153 DF,  p-value: 0.0001208
```

```r
# F-test for quadratic: is improvement in fit significant?
pf(m_3a_q$fstatistic[1], m_3a_q$fstatistic[2], m_3a_q$fstatistic[3], lower.tail = FALSE)
```

```
##       value
## 0.000120778
```

Specification choice: I prefer the linear specification. The quadratic term may improve fit slightly, but the linear model is simpler and the marginal effect of poverty is constant and easy to interpret. If the F-test for the quadratic specification is not significant (or the adjusted R-squared gain is small), the linear model is adequate.

Interpretation: A one-percentage-point increase in the station-area poverty rate is associated with about 4.612 more arrests per 100,000 ridership. The coefficient is statistically significant ($0.0016 < 0.05$). So higher poverty is associated with higher fare-evasion arrest intensity.

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

```
stations %>%
  ungroup() %>%
  group_by(highpov) %>%
  summarise(n = n(),
            mean_arrper = round(weighted.mean(arrperswipe, swipes2016), 2))
```

```
## # A tibble: 2 x 3
##   highpov              n mean_arrper
##   <fct>            <int>       <dbl>
## 1 Not high poverty    79        0.78
## 2 High poverty        77        1.42
```

```
m_3b <- lm_robust(arrperswipe ~ highpov, data = stations, weights = swipes2016)
summary(m_3b)
```

```
##
## Call:
## lm_robust(formula = arrperswipe ~ highpov, data = stations, weights = swipes2016)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                     Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)           0.7829     0.1122   6.981 8.213e-11   0.5614    1.004 154
## highpovHigh poverty   0.6332     0.2005   3.157 1.916e-03   0.2370    1.029 154
##
## Multiple R-squared:  0.06832 ,   Adjusted R-squared:  0.06227
## F-statistic: 9.969 on 1 and 154 DF,  p-value: 0.001916
```

The difference in mean arrest intensity between high-poverty and not-high-poverty areas equals the coefficient on "High poverty" 0.6332 arrests per 100k ridership. This difference is statistically significant at conventional levels ($p < 0.05$), so high-poverty station areas have higher arrest intensity on average.

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?

```
stations %>%
  group_by(nblack, highpov) %>%
  summarise(mean_arrint = round(weighted.mean(arrperswipe, swipes2016), 2), .groups = "drop")
```

```
## # A tibble: 4 x 3
##   nblack             highpov          mean_arrint
##   <fct>              <fct>                  <dbl>
## 1 Majority non-Black Not high poverty        0.66
## 2 Majority non-Black High poverty            0.82
## 3 Majority Black     Not high poverty        1.19
## 4 Majority Black     High poverty            2.49
```

```
# 2x2 table (BONUS)
stations %>%
  group_by(nblack, highpov) %>%
  summarise(mean_arrint = round(weighted.mean(arrperswipe, swipes2016), 2), .groups = "drop") %>%
  pivot_wider(names_from = highpov, values_from = mean_arrint) %>%
  kable()
```

nblack	Not high poverty	High poverty
Majority non-Black	0.66	0.82
Majority Black	1.19	2.49

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!

	Majority non-Black (N=48)		Majority Black (N=29)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
arrperswipe	0.82	0.93	2.49	1.72	1.68	0.33
povrt_all_2016	0.32	0.05	0.32	0.07	0.00	0.01

```r
stations <- stations %>% mutate(weights = swipes2016)

# Step 1 & 2: difference in means (arrest intensity and poverty rate) by nblack, high-pov only
datasummary_balance(~ nblack, fmt = 2,
  data = subset(stations, highpov == "High poverty",
              select = c("nblack", "arrperswipe", "povrt_all_2016", "weights")))
```

```r
# OLS for inference (weighted, robust SEs)
m_4b_arr <- lm_robust(arrperswipe ~ nblack,
  data = subset(stations, highpov == "High poverty"), weights = swipes2016)
m_4b_pov <- lm_robust(povrt_all_2016 ~ nblack,
  data = subset(stations, highpov == "High poverty"), weights = swipes2016)
summary(m_4b_arr)
```

```
##
## Call:
## lm_robust(formula = arrperswipe ~ nblack, data = subset(stations,
##     highpov == "High poverty"), weights = swipes2016)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                       Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## (Intercept)             0.8171     0.1565   5.222 1.529e-06   0.5054    1.129 75
## nblackMajority Black    1.6755     0.3303   5.072 2.761e-06   1.0174    2.334 75
##
## Multiple R-squared:  0.2917 ,    Adjusted R-squared:  0.2822
## F-statistic: 25.73 on 1 and 75 DF,  p-value: 2.761e-06
```

```r
summary(m_4b_pov)
```

```
##
## Call:
## lm_robust(formula = povrt_all_2016 ~ nblack, data = subset(stations,
##     highpov == "High poverty"), weights = swipes2016)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                       Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## (Intercept)            0.31729   0.009147 34.6862 6.222e-48  0.29907  0.33551 75
## nblackMajority Black   0.00383   0.014557  0.2631 7.932e-01 -0.02517  0.03283 75
##
## Multiple R-squared:  0.0009764 , Adjusted R-squared:  -0.01234
## F-statistic: 0.06922 on 1 and 75 DF,  p-value: 0.7932
```

Step 3: The difference in mean arrest intensity between high-poverty Majority Black and Majority non-Black areas is given by the coefficient on nblack (Majority Black) in the first regression; the difference in mean poverty rate is given by the same in the second. If the poverty-rate difference is small or insignificant relative to the arrest-intensity difference, the gap in arrest intensity is not well explained by poverty alone. If poverty is substantially higher in Majority Black high-poverty areas and the arrest-intensity gap moves in the same direction, some of the gap may be explained by the fact that poverty is associated with arrest intensity.

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!)

```
ggplot(stations, aes(x = povrt_all_2016, y = arrperswipe, color = nblack, weight = swipes2016)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x) +
  labs(x = "Station area poverty rate", y = "Arrests per 100,000 ridership",
       title = "Fare evasion arrest intensity vs. poverty by race",
       subtitle = "Subway stations in Brooklyn (2016)") +
  scale_color_discrete(name = "Majority Black station area", labels = c("No", "Yes"),
                       guide = guide_legend(reverse = TRUE)) +
  theme(legend.position = "bottom")
```

```
m_4c_interact <- lm_robust(arrperswipe ~ povrt_all_2016 * nblack,
                           data = stations, weights = swipes2016)
summary(m_4c_interact)
```

```
##
## Call:
## lm_robust(formula = arrperswipe ~ povrt_all_2016 * nblack, data = stations,
##     weights = swipes2016)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                          0.4889     0.2119   2.308 2.236e-02
## povrt_all_2016                       1.1473     1.0382   1.105 2.708e-01
## nblackMajority Black                -1.8531     0.7142  -2.594 1.040e-02
## povrt_all_2016:nblackMajority Black  11.3601     2.7594   4.117 6.278e-05
##                                    CI Lower CI Upper  DF
## (Intercept)                         0.07036   0.9075 152
## povrt_all_2016                     -0.90373   3.1984 152
## nblackMajority Black               -3.26418  -0.4420 152
## povrt_all_2016:nblackMajority Black  5.90848  16.8118 152
##
```

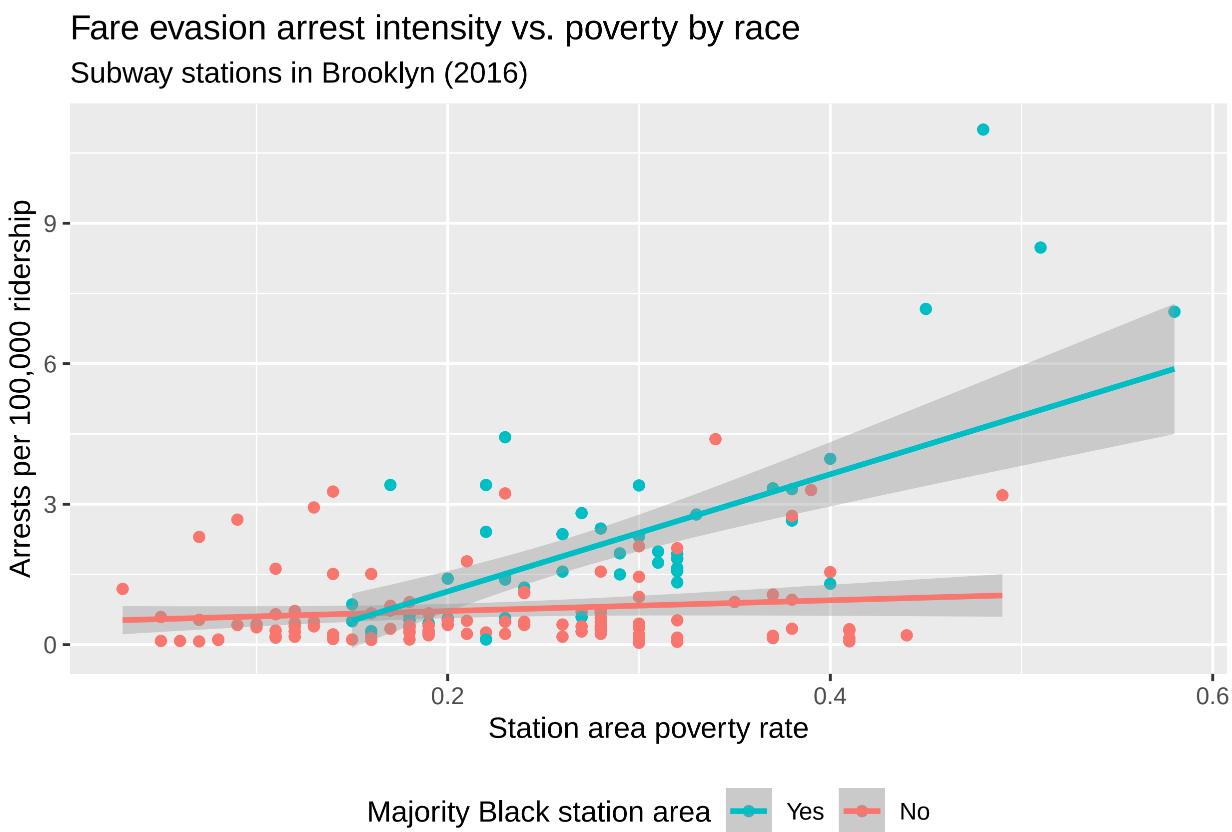

Figure 3: Arrest intensity vs. poverty by majority Black vs. non-Black station area

```
## Multiple R-squared:  0.4149 ,    Adjusted R-squared:  0.4033
## F-statistic: 18.09 on 3 and 152 DF,  p-value: 4.342e-10
```

Interpretation: The interaction model allows the effect of poverty on arrest intensity to differ by racial composition. The coefficient 1.1473 on povrt_all_2016 is the effect of poverty in Majority non-Black areas; the coefficient 11.3601 on povrt_all_2016:nblack Majority Black is the additional effect of poverty in Majority Black areas. If that interaction term is positive and significant, poverty has a stronger positive association with arrest intensity in Majority Black station areas than in Majority non-Black areas—consistent with enforcement being more responsive to poverty in Majority Black neighborhoods.

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?

BONUS answer (brief): Station-area race and poverty are proxy measures for the demographics of who actually uses each station; for busier or destination stations, riders may come from elsewhere, so the proxy is measured with non-random error. Classical measurement error in a regressor typically attenuates its coefficient; with non-random error, bias can go either way and may affect the interaction term. Other measurement issues could include crime under-reporting or classification differences by area. Creative ideas: use rider surveys or origin–destination data by station to better measure rider demographics; use station fixed effects or instruments if panel or instrument availability allows.

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`.

```
st_crime <- read.csv("nypd_criminalcomplaints_2016.csv")
# Crime file has st_id and crimes only; join by st_id
stations <- stations %>%
  inner_join(st_crime, by = "st_id") %>%
  group_by(st_id, 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.

```
stations_highpov_only <- stations %>% ungroup() %>% filter(highpov == "High poverty")

stations_highpov_only %>%
  group_by(nblack) %>%
  summarise(n = n(),
            mean_crimes = round(weighted.mean(crimes, swipes2016), 2),
            .groups = "drop") %>%
  kable()
```

nblack	n	mean_crimes
Majority non-Black	48	NA
Majority Black	29	NA

```
m_5b <- lm_robust(crimes ~ nblack, data = stations_highpov_only, weights = swipes2016)
summary(m_5b)
```

```
##
## Call:
## lm_robust(formula = crimes ~ nblack, data = stations_highpov_only,
##     weights = swipes2016)
##
```

```
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                      Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## (Intercept)             815.0      56.68  14.379 2.966e-23    702.1    928.0 75
## nblackMajority Black    496.1     117.51   4.222 6.725e-05    262.0    730.2 75
##
## Multiple R-squared:  0.3283 ,    Adjusted R-squared:  0.3193
## F-statistic: 17.82 on 1 and 75 DF,  p-value: 6.725e-05
```

The coefficient on 496.1 nblack (Majority Black) is the ridership-weighted difference in mean crimes between high-poverty Majority Black and Majority non-Black station areas. If it is positive and statistically significant, there are more crimes reported in high-poverty Majority Black areas than in high-poverty Majority non-Black areas.

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?

```r
# Baseline: 4c interaction model (no crime control)
summary(m_4c_interact)
```

```
##
## Call:
## lm_robust(formula = arrperswipe ~ povrt_all_2016 * nblack, data = stations,
##     weights = swipes2016)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                           0.4889     0.2119   2.308 2.236e-02
## povrt_all_2016                        1.1473     1.0382   1.105 2.708e-01
## nblackMajority Black                 -1.8531     0.7142  -2.594 1.040e-02
## povrt_all_2016:nblackMajority Black  11.3601     2.7594   4.117 6.278e-05
##                                     CI Lower CI Upper  DF
## (Intercept)                          0.07036   0.9075 152
## povrt_all_2016                      -0.90373   3.1984 152
## nblackMajority Black                -3.26418  -0.4420 152
## povrt_all_2016:nblackMajority Black  5.90848  16.8118 152
##
## Multiple R-squared:  0.4149 ,    Adjusted R-squared:  0.4033
## F-statistic: 18.09 on 3 and 152 DF,  p-value: 4.342e-10
```

	4c baseline	+ crimes (level)	+ log(crimes)	+ crimes per 100k
(Intercept)	0.49	−0.16	−2.95	0.24
	(0.21)	(0.24)	(1.05)	(0.20)
povrt_all_2016	1.15	2.05	1.86	0.89
	(1.04)	(1.02)	(1.02)	(1.00)
nblackMajority Black	−1.85	−1.63	−1.76	−1.58
	(0.71)	(0.73)	(0.74)	(0.68)
povrt_all_2016 × nblackMajority Black	11.36	9.95	10.20	9.67
	(2.76)	(2.79)	(2.80)	(2.50)
crimes		0.00		
		(0.00)		
log(crimes + 1)			0.49	
			(0.15)	
crimes_per_100k				0.01
				(0.00)
Num.Obs.	156	156	156	156
R2	0.415	0.462	0.460	0.470
R2 Adj.	0.403	0.448	0.446	0.456

```r
# Add crimes (linear)
m_5c_linear <- lm_robust(arrperswipe ~ povrt_all_2016 * nblack + crimes,
                         data = stations, weights = swipes2016)
# Log(crimes)
m_5c_log <- lm_robust(arrperswipe ~ povrt_all_2016 * nblack + log(crimes + 1),
                      data = stations, weights = swipes2016)
# Crimes per 100k swipes
stations <- stations %>% mutate(crimes_per_100k = (crimes / swipes2016) * 100000)
m_5c_intensity <- lm_robust(arrperswipe ~ povrt_all_2016 * nblack + crimes_per_100k,
                            data = stations, weights = swipes2016)

modelsummary(list("4c baseline" = m_4c_interact,
                  "+ crimes (level)" = m_5c_linear,
                  "+ log(crimes)" = m_5c_log,
                  "+ crimes per 100k" = m_5c_intensity),
             fmt = 2, gof_map = c("nobs", "r.squared", "adj.r.squared"))
```

Interpretation: Compare the interaction term (povrt × Majority Black) across models. If it shrinks and becomes insignificant when we add crimes, the stronger effect of poverty in Majority Black areas may be partly explained by higher crime (and thus police presence) there. If the interaction remains significant after controlling for crime, the racial gap in the poverty–enforcement relationship is not fully explained by differences in crime. Different functional forms (level, log, per 100k swipes) can be compared to check robustness.

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.

Summary of findings

- Poverty and arrest intensity: Arrest intensity (arrests per 100k swipes) is positively associated with station-area poverty; high-poverty areas have higher arrest intensity on average (Section 3).
- Moderation by race: The relationship between poverty and arrest intensity differs by racial composition (Section 4c): in Majority Black station areas, the slope of poverty on arrest intensity is larger than in Majority non-Black areas. Among high-poverty stations, Majority Black areas have higher mean arrest intensity than Majority non-Black areas (Section 4b). Whether that gap is "explained" by higher average poverty in those Black areas depends on comparing the difference in mean poverty to the difference in mean arrest intensity (Section 4b Step 3).
- Role of crime:Controlling for reported crimes (Section 5c) tests whether the poverty-by-race interaction is explained by differential police presence (proxied by crime). If the interaction shrinks when adding crimes, part of the gap may be explained by where police are deployed; if it persists, the disparity is not fully explained by crime.

Framing: We observe disparities in outcomes (arrest intensity) by station-area race and poverty. We do not observe officer intent or individual-level discrimination; any discussion of "bias" should be limited to what the data support (e.g., differential intensity by area demographics) and include appropriate caveats.

Additional analysis with current data: Sensitivity checks (e.g., excluding other outliers, different poverty/race cutoffs, alternative definitions of arrest intensity); subgroup analyses by line or zone; robustness of key results to linear vs quadratic specification.

Key limitations: (1) Station-area demographics proxy for rider demographics with non-random measurement error (busier stations draw commuters from elsewhere). (2) Crime is a proxy for police presence; we do not observe actual deployment. (3) Cross-sectional design; unobserved confounders (e.g., fare evasion rates, reporting differences) that vary by race and poverty cannot be ruled out. (4) Brooklyn only; results may not generalize to other boroughs.

Additional data that would help: Rider-level or station-level demographics of who uses each station (to reduce measurement error in race/poverty); actual police deployment or staffing by station/time; fare evasion rates or incidents by station (to separate enforcement from underlying behavior); panel data (years before/after policy changes) to strengthen causal inference.