Subway Fare Evasion Arrests and Racial Bias

Myung Eun Hyeon

2023-09-18

1 Load libraries.

install.packages("plyr") install.packages("weights")

```
library(plyr)
library(tidyverse)
library(fastDummies)
library(weights)
library(lmtest)
library(sandwich)
library(knitr)
```

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

```
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 st
##  $ 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 ...
```

The BDS data includes 2246 observations (client arrest records), and the LAS data includes 1965 observations. Both datasets include basic demographic information on age, sex, race, ethnicity (coded differently in each dataset), as well as information on the location/subway station where the arrest occurred. The LAS data also includes information on the case dismissal rates.

In each raw dataset, the unit of observation is the arrested individual (client). The representative population is all individuals arrested in Brooklyn during 2016 by the NYPD for subway fare evasion and were represented by public defenders. If nearly all individuals arrested for fare evasion were represented by public defenders, then this sample represents all subway fare evasion arrests in Brooklyn in 2016. Although this difficult to argue convincingly without additional information, it is supported by court observers.

```
#recode race/ethnicity information from character to factors
arrests_bds <- arrests_bds %>%
  mutate(race = as.factor(race),
         ethnicity = as.factor(ethnicity) )

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

summary(arrests_bds$race)
summary(arrests_las$race)

summary(arrests_bds$ethnicity)
summary(arrests_las$ethnicity)
```

First, used the mutate(race =) function to make the column names of the column containing the race values that we are referring to in both datasets are the same. The coding of race in the datasets are different in that the bds dataset puts together Native Americans (American Indians) and Asians and Pacific Islanders together under one race category, while the las dataset has only the Asian and Pacific Islanders as one category and does not have one for Native Americans. Second, used the mutate(ethnicity =) function to make the column names of column containing the race values are the same. The bds dataset has four categories: Hispanic, Non-Hispanic, Other, and NA's. Meanwhile, the las dataset has three categories: N, Y, NA's, possibly answers to the question: Are you Hispanic?

A data limitation is that there are people that are categorized under NA's and Unknowns, and we cannot know from this data whether this is possibly due to the incapacity to identify bi- or multi-racial categories or if there are any other reasons. This makes it difficult to determine whether we should remove the NA's or unknowns when cleaning the dataset or these are important information that must be kept. It's also important to emphasize what information this data does not include that might be relevant to the question of biased fare evasion enforcement: 1) fare evasion that resulted in a summons (ticket + fine) rather than an arrest, 2) fare evasion enforcement on buses.

2 Data Cleaning

BDS: race data

```
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")))
arrests_bds.clean %>%
  count(race_clean, sort = TRUE)
```

```
## # A tibble: 5 x 2
##   race_clean                n
##   <fct>                 <int>
## 1 Black                  1465
## 2 White                   533
## 3 <NA>                    194
## 4 Other                    33
## 5 Asian/Pacific Islander   21
```

Using = "NA", combined the data that are marked as 0 and unknown under NA. By categorizing Am Indian under the "Other", ensured that the race categories for the bds and las dataset are the same. Then used the levels function to set levels on the values.

```
arrests_bds.clean <- arrests_bds.clean %>%
  mutate(hispanic = recode(ethnicity,
                           "0" = "NA",
                           "Other" = "Non-Hispanic",)) %>%
  mutate(hispanic = factor(hispanic,
                           levels = c("Hispanic", "Non-Hispanic")))

summary(arrests_bds.clean$hispanic)
```

```
##     Hispanic Non-Hispanic         NA's
##          493         1563          190
```

First, created a "hispanic" column that has three categories: hispanic, NA, and Non-hispanic by recoding the 0 values into NA and Other values into Non-Hispanic. Then used the levels function to set levels on the values.

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

arrests_bds.clean %>%
  count(race_eth, sort = TRUE)
```

```
## # A tibble: 6 x 2
##   race_eth                  n
##   <fct>                 <int>
## 1 Black                  1361
```

```
## 2 Hispanic                 493
## 3 <NA>                     193
## 4 Non-Hispanic White       165
## 5 Asian/Pacific Islander    21
## 6 Other                     13
```

First, started by converting the factors into characters to use the ifelse function to create mutually exclusive categories, as the ifelse function does not preserve factors the same way. Then, created a race_eth factor variable by storing the hispanic values to the Hispanic column and all other race and ethnic categories to their appropriate factors in the hispanic_char and race_char. Then, categorized the race "White" as "Non-Hispanic White", combining ethnicity and race.

3 Clean LAS race and ethnicity data

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

Here, first started by renaming the race categories in the las dataset to align with the race_eth column in the bds dataset. Then, used ifelse function to allocate "Y" values to the Hispanic column and the factor and levels functions to set levels

4 Combining (appending) the BDS and LAS microdata

```
arrests_bds.clean <- arrests_bds.clean %>% mutate(pd = "bds")
arrests_las.clean <- arrests_las.clean %>% mutate(pd = "las")
```

Here, used the mutate function to create the column "pd". For the bds dataset, the "pd" column will display the value "bds" for all rows. For the las dataset, the "pd" column will display the value "las" for all rows.

```
# Append `arrests_bds.clean` and `arrests_las.clean`
arrests_all <- rbind.fill(arrests_las.clean, arrests_bds.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_all)
```

```
##    pd                          race_eth         age             male
##  bds:2246   Hispanic               : 704   Min.   : 0.00   Min.   :0.0000
##  las:1965   Non-Hispanic White     : 459   1st Qu.:20.00   1st Qu.:1.0000
##             Asian/Pacific Islander:  32   Median :26.00   Median :1.0000
##             Black                  :2562   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
```

Here, used the rbind.fill function to append two datasets where the las dataset has one extra column than the bds dataset. Then, set the character columns as factors and selected 7 columns to inspect for consistency and accuracy in the new data frame.

4.0.1 Total number of subway fare evasion arrest records?

```
nrow(arrests_all)
```

```
## [1] 4211
```

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

```
# Save `arrests_all` as an .RData file
save(list = "arrests_all", file = "arrests_all.RData")
```

5 Descriptive statistics by race/ethnicity

```
arrests_all %>%
  count(race_eth, sort = TRUE)
```

```
##                 race_eth    n
## 1                  Black 2562
## 2               Hispanic  704
## 3     Non-Hispanic White  459
## 4                   <NA>  430
## 5 Asian/Pacific Islander   32
## 6                  Other   24
```

```r
# Proportion of total arrests for each race/ethnicity category
prop.table(table(arrests_all$race_eth, useNA = "always")) %>%
  round(2) %>%
  as.data.frame() %>%
  arrange(desc(Freq)) %>%
  rename(race_eth = Var1)
```

```
##                 race_eth Freq
## 1                  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
```

```r
prop.table(table(arrests_all$race_eth)) %>%
  round(2) %>%
  as.data.frame() %>%
  arrange(desc(Freq)) %>%
  rename(race_eth = Var1)
```

```
##                 race_eth Freq
## 1                  Black 0.68
## 2               Hispanic 0.19
## 3     Non-Hispanic White 0.12
## 4 Asian/Pacific Islander 0.01
## 5                  Other 0.01
```

Here, Excluding the NAs increase the proportion of arrests for Hispanic, Black, and Non-Hispanic White by 2%, 7%, and 1%. These add up to the 10% of arrests under the NA category.

```r
# Average age, share male, and dismissal rate for each race/ethnicity category
race_eth_stats <- arrests_all %>%
  group_by(race_eth) %>%
  summarise(avg_age = mean(age, na.rm=TRUE),
            avg_male = mean(male, na.rm=TRUE),
            avg_dismissal = mean(dismissal, na.rm=TRUE))

race_eth_stats %>%
  group_by(race_eth) %>%
  count(avg_age, avg_male, avg_dismissal, sort = TRUE)
```

```
## # A tibble: 6 x 5
## # Groups:   race_eth [6]
##   race_eth               avg_age avg_male avg_dismissal     n
##   <fct>                    <dbl>    <dbl>         <dbl> <int>
## 1 Hispanic                  29.7    0.901         0.538     1
## 2 Non-Hispanic White        29.7    0.898         0.587     1
## 3 Asian/Pacific Islander    28.9    0.938         0.636     1
## 4 Black                     29.1    0.875         0.514     1
## 5 Other                     28.3    0.833         0.444     1
## 6 <NA>                      26.0    0.603         0.75      1
```

The average age is the highest for Non-Hispanic White, which is 29.7 years old, and the lowest for N/A, which is 26 years old. The percentage of male is high for all categories, around 90%, and NA has the lowest percentage of male at 60% and the Asian/Pacific Islander has the highest at 94%. The average dismissal rate is the highest for NA at 75%, second highest is Asian/Pacific Islander at 60%, and the lowest is for Other at 44%.

6 Subway-station level analysis

```
# Creating dummy variables for each race/ethnicity category
arrests_dummy <- dummy_cols(arrests_all, select_columns = "race_eth") %>%
  select(-pd, -race_eth, -age, -male, -st_id, -loc2, -dismissal) %>%
  summarise(n = n(),
            avg_Black = mean(race_eth_Black, na.rm = TRUE),
            avg_Hispanic= mean(race_eth_Hispanic, na.rm = TRUE),
            avg_API = mean(`race_eth_Asian/Pacific Islander`, na.rm = TRUE),
            avg_NHW = mean(`race_eth_Non-Hispanic White`, na.rm = TRUE),
            avg_Other = mean(race_eth_Other, na.rm = TRUE)) %>%
    arrange(desc(n))
knitr::kable(arrests_dummy)
```

n	avg_Black	avg_Hispanic	avg_API	avg_NHW	avg_Other
4211	0.6775985	0.1861941	0.0084634	0.1213965	0.0063475

Top 10 stations by arrest totals

```
arrests_dummy <- dummy_cols(arrests_all, select_columns = "race_eth")

arrests_stations <- arrests_dummy %>%
    group_by(loc2) %>%
    summarise(st_id = first(st_id),
              n = n(),
              n_black = sum(race_eth_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

loc2	st_id	n	n_black	n_hisp	n_api	n_nhw	n_oth
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

Stations with at least 50 arrests

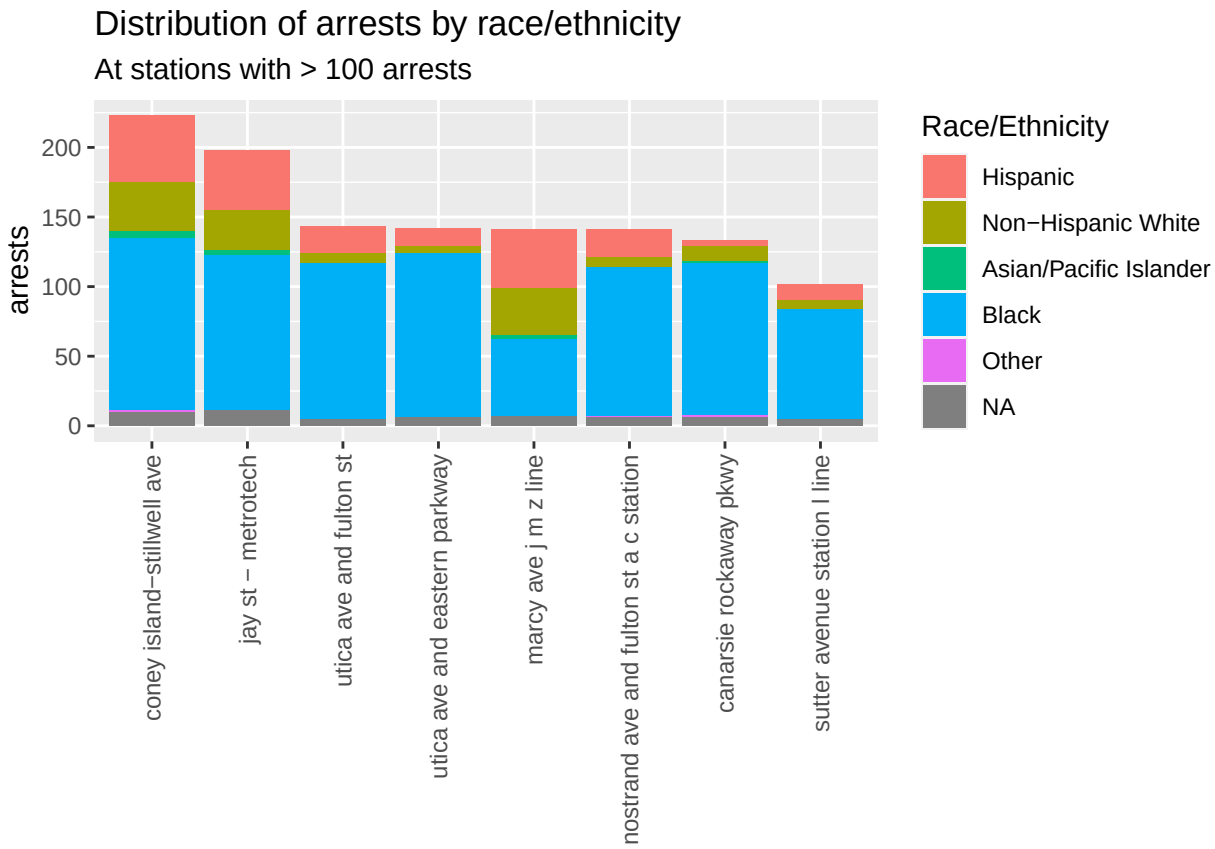
```
arrests_stations_top <- arrests_stations %>%
  group_by(loc2) %>%
  summarise(st_id = first(st_id),
            n = n(),
            n_station_arrest = n_black + n_hisp + n_api + n_nhw + n_oth,
            n_BH = n_black + n_hisp,
            n_NA = sum(is.na(n)),
            prop_BH = n_BH/(n_station_arrest-n_NA)) %>%
  arrange(prop_BH) %>%
  filter(n_station_arrest > 50)
knitr::kable(arrests_stations_top)
```

loc2	st_id	n	n_station_arrest	n_BH	n_NA	prop_BH
marcy ave j m z line	114	1	134	97	0	0.7238806
myrtle av and broadway station	117	1	66	53	0	0.8030303
coney island-stillwell ave	66	1	213	172	0	0.8075117
broadway and lorimer st j m station	112	1	68	56	0	0.8235294
clinton - washington avs station	64	1	58	48	0	0.8275862
jay st - metrotech	99	1	187	155	0	0.8288770
hoyt-schermerhorn a c g line	98	1	65	55	0	0.8461538
canarsie rockaway pkwy	54	1	127	113	0	0.8897638
nevins st 2 3 4 5 lines	123	1	81	74	0	0.9135802
kingston - throop avs	106	1	87	81	0	0.9310345
hoyt st 2 3	97	1	75	70	0	0.9333333
sutter avenue station l line	147	1	97	91	0	0.9381443
nostrand ave and fulton st a c station	131	1	135	127	0	0.9407407
utica ave and fulton st	150	1	138	131	0	0.9492754
livonia ave l line	111	1	72	69	0	0.9583333
junius st 3 line	101	1	73	70	0	0.9589041
utica ave and eastern parkway	70	1	136	131	0	0.9632353
court st r subway/borough hall 2 subway 3 subway 4 subway 5 subway	68	1	55	53	0	0.9636364
rockaway ave c line	141	1	58	57	0	0.9827586
sutter av - rutland rd 3 line	148	1	65	64	0	0.9846154
rockaway ave 3 line	140	1	57	57	0	1.0000000

Among the stations that have higher than 50 total arrests, Jay S t- Metrotech station had the highest percentage of arrests of Black and Hispanic individuals at 83%. Generally, the percentage is very high, ranging from 72% to 83%. This implies that there may be discrimination in the policing activities of the NYPD for fare evasion, but this needs to be further explored through regression analysis.

7 Visualizing the distribution of arrests by race/ethnicity at stations with more than 100 arrests.

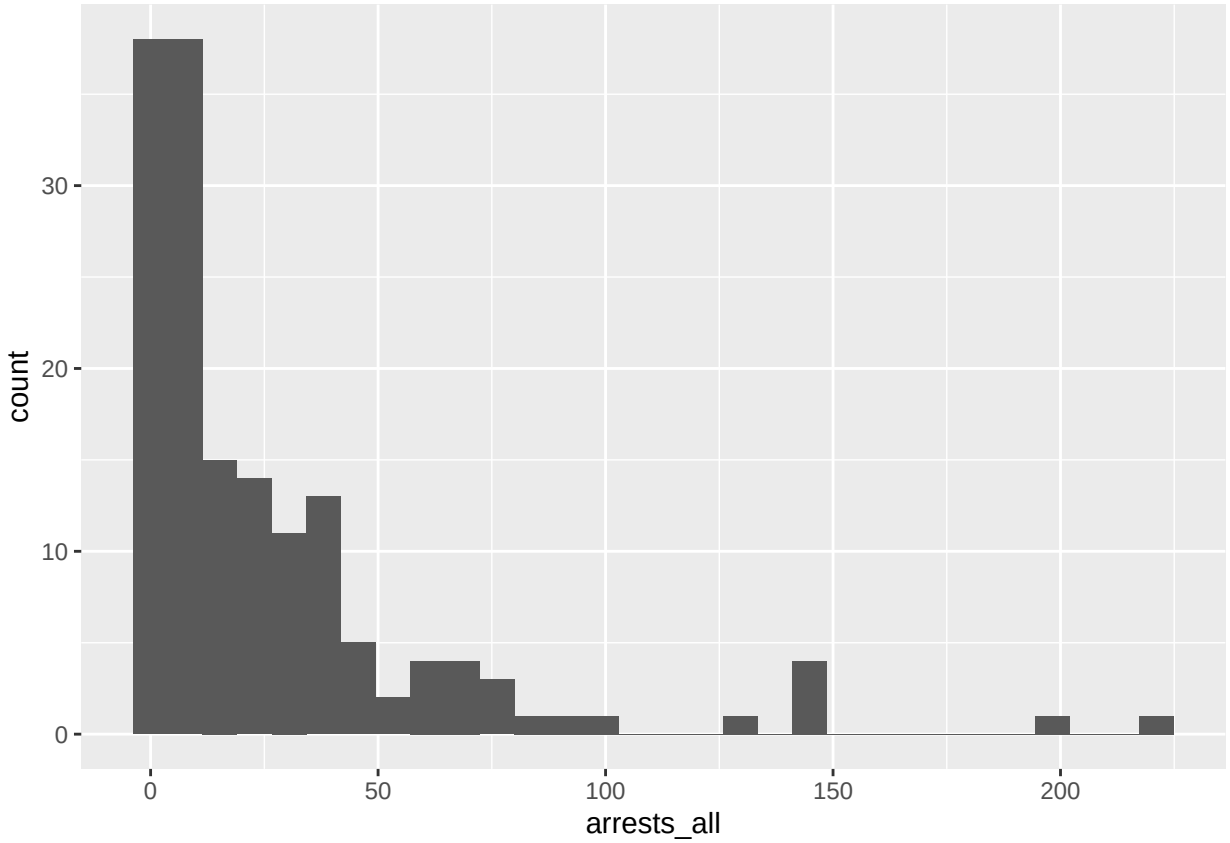
```
arrests_stations_race <- arrests_all %>%
  group_by(loc2) %>%
  mutate(st_arrests = n()) %>%
  ungroup() %>%
  group_by(loc2, race_eth) %>%
  summarise(arrests = n(), st_arrests = first(st_arrests)) %>%
  arrange(desc(st_arrests)) %>%
  filter(st_arrests > 100)

g <- ggplot(arrests_stations_race,
            aes(x = reorder(loc2, -st_arrests),
                y = arrests, fill = race_eth)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "right",
        axis.title.x=element_blank(),
        axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 1)) +
  scale_fill_discrete(name = "Race/Ethnicity") +
  ggtitle("Distribution of arrests by race/ethnicity",
          subtitle = "At stations with > 100 arrests")
g
```



8 Aggregating to subway station-level arrest totals

```
load("arrests.clean.RData")
```

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

ggplot(data = st_arrests, aes(x = arrests_all)) + geom_histogram()
```



The distribution of arrests across stations is extremely skewed to the right. This histogram shows that the majority of subway stations had a relatively small number of fare evasion arrests. The median station arrest total is 13 compared to a mean of 26.82, with 8 stations home to more than 100 arrests.

9 Joining subway ridership and neighborhood demographic data

```
st_poverty <- read.csv("station_povdataclean_2016.csv",
                       stringsAsFactors = TRUE)
st_ridership <- read.csv("Subway Ridership by Station - BK.csv",
                         stringsAsFactors = TRUE)

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 = c("st_id" = "st_id")) %>%
    inner_join(st_ridership, by = c("st_id" = "st_id",
                                    "mta_name" = "mta_name")) %>%
  select(-all_of(drop_vars)) %>%
  group_by(st_id, mta_name)

st_joined %>%
```

```
  ungroup() %>%
  str(give.attr = FALSE)
```

```
## tibble [157 x 14] (S3: tbl_df/tbl/data.frame)
##  $ st_id         : int [1:157] 66 99 150 70 114 131 54 147 106 123 ...
##  $ loc2          : Factor w/ 157 levels "15 st prospect park f g line",..: 66 100 149 148 110 129 54
##  $ arrests_all   : int [1:157] 223 198 143 142 141 141 133 102 90 86 ...
##  $ x             : num [1:157] -74 -74 -73.9 -73.9 -74 ...
##  $ y             : num [1:157] 40.6 40.7 40.7 40.7 40.7 ...
##  $ mta_name      : Factor w/ 157 levels "15 St-Prospect Park F subway G subway",..: 66 99 150 70 114
##  $ pop_black_2016: int [1:157] 36 1939 14825 13135 1542 10311 5624 11804 16176 2698 ...
##  $ pov_black_2016: int [1:157] 2 677 4592 3796 483 2437 900 6706 3832 306 ...
##  $ pop_all_2016  : int [1:157] 5186 12437 18556 17561 23711 15934 6753 15751 20610 13654 ...
##  $ pov_all_2016  : int [1:157] 1329 1939 6149 5565 9182 3511 1156 9104 4809 1221 ...
##  $ povrt_all_2016: num [1:157] 0.256 0.156 0.331 0.317 0.387 ...
##  $ shareblack    : num [1:157] 0.00694 0.15591 0.79893 0.74796 0.06503 ...
##  $ nblack        : int [1:157] 0 0 1 1 0 1 1 1 1 0 ...
##  $ swipes2016    : int [1:157] 5025598 13091255 5152649 9051970 4272443 5861658 3897784 1435112 2031
```

First, I used the inner_join() function to join st_arrests and st_poverty data frames on st_id column that exists in both data frames. Then, using a pipe, I joined the new combined data frame to the st_ridership data frame on st_id and mta_name columns that exist in both data frames. Then, since the st_poverty data frame only has data concering the year 2016, I first created a variable with columns of swipes per station from 2011 to 2015 and then used the select() function with all_of() to make clear that we are dropping all the column names in the variable drop_vars. Ungrouping showed all of the 13 columns that are now in the data frame and 157 observations.

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

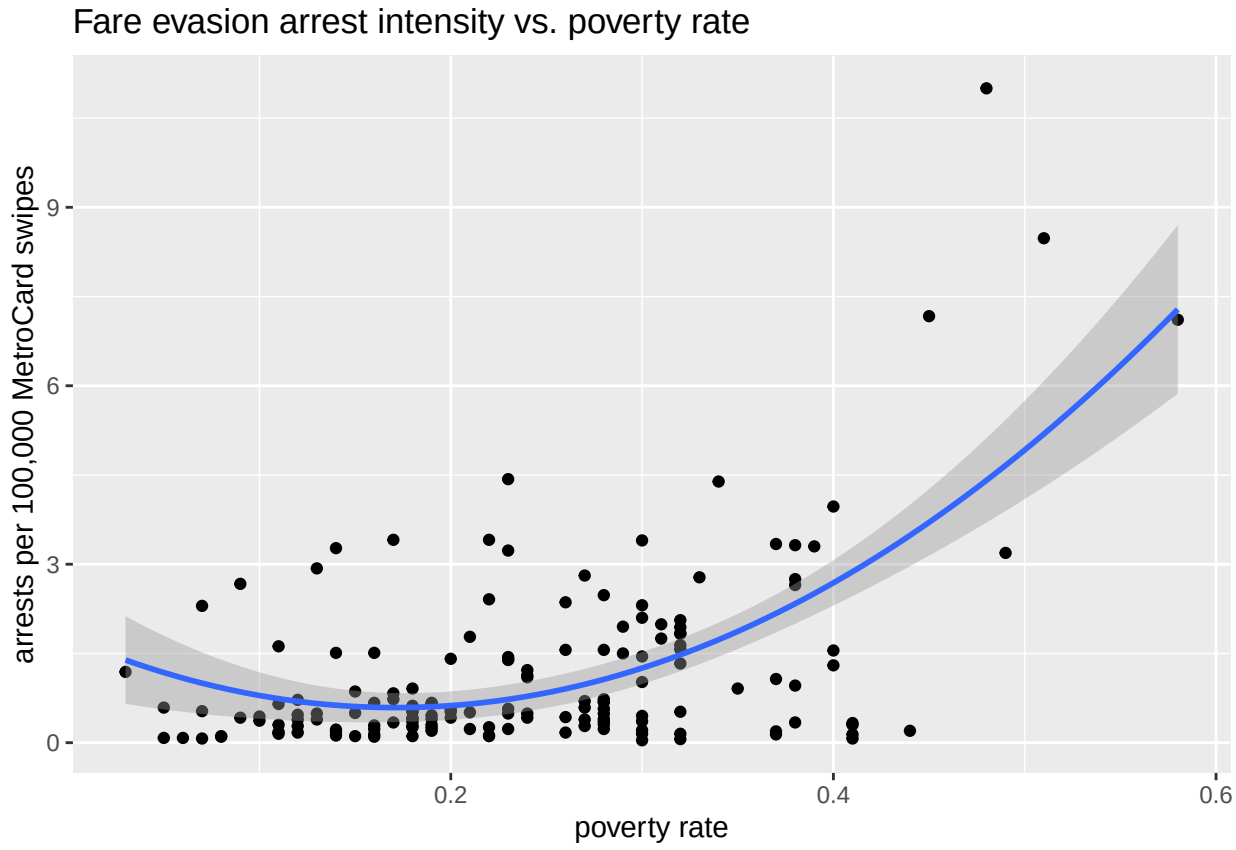
```
## # A tibble: 10 x 5
## # Groups:   st_id, mta_name [10]
##    st_id mta_name                          arrests_all shareblack povrt_all_2016
##    <int> <fct>                                   <int>      <dbl>          <dbl>
##  1    66 "Coney Island-Stillwell Av D sub~         223       0.01           0.26
##  2    99 "Jay St-MetroTech A subway C sub~         198       0.16           0.16
##  3   150 "Utica Av  A subway C subway "             143       0.8            0.33
##  4    70 "Crown Heights-Utica Av 3 subway~         142       0.75           0.32
##  5   114 "Marcy Av  J subway M subway Z s~         141       0.07           0.39
##  6   131 "Nostrand Av  A subway C subway"          141       0.65           0.22
##  7    54 "Canarsie-Rockaway Pkwy L subway"         133       0.83           0.17
##  8   147 "Sutter Av  L subway"                     102       0.75           0.58
##  9   106 "Kingston-Throop Avs C subway"             90       0.78           0.23
## 10   123 "Nevins St  2 subway 3 subway 4 ~          86       0.2            0.09
```

10 Explore relationship between arrest intensity and poverty rates across subway station (areas)

```
stations <- st_joined %>%
  filter(st_id != 66) %>%
  mutate(arrperswipe_2016 = round(arrests_all / (swipes2016 / 100000), 2),
         highpov = as.numeric(povrt_all_2016 > median(st_joined$povrt_all_2016)),
         nblack = as.numeric(shareblack > .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"))) %>%
  arrange(desc(arrperswipe_2016)) %>%
  select(st_id, mta_name, arrperswipe_2016, arrests_all, shareblack, povrt_all_2016, highpov, nblack, s
knitr::kable(head(stations, n = 10))
```

st_id	mta_name	arrperswipe_2016	arrests_all	shareblack	povrt_all_2016	highpov	nblack	swipes2016
101	Junius St 3 subway	11.00	75	0.78	0.48	High poverty	Majority Black	681935
26	Atlantic Av L subway	8.48	37	0.66	0.51	High poverty	Majority Black	436300
111	Livonia Av L subway	7.17	75	0.83	0.45	High poverty	Majority Black	1045593
147	Sutter Av L subway	7.11	102	0.75	0.58	High poverty	Majority Black	1435112
106	Kingston-Throop Avs C subway	4.43	90	0.78	0.23	High poverty	Majority Black	2031710
112	Lorimer St J subway M subway	4.39	70	0.15	0.34	High poverty	Majority non-Black	1595826
140	Rockaway Av 3 subway	3.97	61	0.78	0.40	High poverty	Majority Black	1534978
54	Canarsie-Rockaway Pkwy L subway	3.41	133	0.83	0.17	Not high poverty	Majority Black	3897784
141	Rockaway Av C subway	3.41	61	0.80	0.22	Not high poverty	Majority Black	1786802
144	Shepherd Av C subway	3.40	36	0.61	0.30	High poverty	Majority Black	1059144

```
ggplot(stations,
       aes(x = povrt_all_2016,
           y = arrperswipe_2016)) +
  geom_point() +
  ggtitle('Fare evasion arrest intensity vs. poverty rate') +
  labs(x = 'poverty rate', y = 'arrests per 100,000 MetroCard swipes') +
  geom_smooth(method = 'lm', formula = y ~ x + I(x^2))
```




```
ols1l <- lm(arrperswipe_2016 ~ povrt_all_2016, data = stations)
summary(ols1l)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ povrt_all_2016, data = stations)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4635 -0.7814 -0.1709  0.2944  8.0472
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.5190     0.2789  -1.861   0.0646 .
## povrt_all_2016   7.2330     1.0659   6.786 2.34e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.391 on 154 degrees of freedom
## Multiple R-squared:  0.2302, Adjusted R-squared:  0.2252
## F-statistic: 46.05 on 1 and 154 DF,  p-value: 2.339e-10
```

```
coeftest(ols1l, vcov = vcovHC(ols1l, type="HC1"))
```

```
##
```

```
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    -0.51903    0.38846 -1.3361 0.1834816
## povrt_all_2016  7.23300    1.88762  3.8318 0.0001848 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#quadratic model(all stations)
ols1q <- lm(arrperswipe_2016 ~ povrt_all_2016 + I(povrt_all_2016 ^ 2),
            data = stations)

summary(ols1q)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ povrt_all_2016 + I(povrt_all_2016^2),
##     data = stations)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2840 -0.5613 -0.2361  0.4207  6.5904
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.7648     0.4722   3.738 0.000262 ***
## povrt_all_2016       -13.7231     3.7798  -3.631 0.000385 ***
## I(povrt_all_2016^2)   40.0690     6.9852   5.736 5.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.266 on 153 degrees of freedom
## Multiple R-squared:  0.3664, Adjusted R-squared:  0.3582
## F-statistic: 44.25 on 2 and 153 DF,  p-value: 6.859e-16
```

```
coeftest(ols1q, vcov = vcovHC(ols1q, type="HC1"))
```

```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)            1.76476    0.43906  4.0194 9.131e-05 ***
## povrt_all_2016       -13.72307    4.50874 -3.0437  0.002752 **
## I(povrt_all_2016^2)   40.06902   10.55736  3.7954  0.000212 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on visual inspection, both the linear and quadratic models appear to fit the relationship between fare evasion arrest intensity and poverty rates across all stations fairly well. However, I prefer the quadratic model because it explains more of the variation in arrest intensity than the linear model; the quadratic model has an adjusted R-squared of 0.36 compared to 0.23 for the linear model. Here, I chose not to weight station observations by the number of MetroCard swipes, so that each station area is equally weighted in

the regression analysis. When computing statistics for groups of stations, I will weight by swipes so that statistics are representative of the ridership in each group.

```
stations %>%
ungroup() %>%
group_by(highpov) %>%
summarise(n = n(),
mean_pov = weighted.mean(povrt_all_2016, swipes2016),
mean_arrper = weighted.mean(arrperswipe_2016, swipes2016))
```

```
## # A tibble: 2 x 4
##   highpov              n mean_pov mean_arrper
##   <fct>            <int>    <dbl>       <dbl>
## 1 Not high poverty    79    0.146       0.783
## 2 High poverty        77    0.319       1.42
```

```
ols_diff1 <- lm(formula = arrperswipe_2016 ~ highpov, data = stations,
weights = swipes2016)
ols_diff1_robSE <- coeftest(ols_diff1, vcov = vcovHC(ols_diff1, type="HC1"))
ols_diff1_robSE
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          0.78293    0.11127  7.0365 6.077e-11 ***
## highpovHigh poverty  0.63316    0.19953  3.1732  0.001821 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference in average fare evasion arrest intensity between high- and low-poverty subway stations (weighted by MetroCard swipes) is 0.63 with a p-value of 0.0018. Thus we can conclude that this difference is statistically significant beyond the 1% level.

11 Neighborhood racial composition and the relationship between poverty and arrest intensity

```
t1_arrper_wtd <- tapply(stations$arrperswipe_2016 * stations$swipes2016,
                        list(stations$highpov, stations$nblack),
                        sum) /
  tapply(stations$swipes2016,
         list(stations$highpov, stations$nblack),
         sum)
round(t1_arrper_wtd, 2)
```

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

After weighting by the number of MetroCard swipes at each station, I calculated the difference in mean arrests per swipes at each station. The arrests intensity (per 100,000 swipes) is highest for stations around neighborhoods where the majority are Black residents and high poverty rates, at 2.49 arrests per 100,000 swipes. It is lowest for stations around neighborhoods where the majority are non-Black residents and not high poverty rates, at 0.66 arrests per 100,000 swipes. The differences in poverty rate could not explain the differences in arrest intensities because The arrest intensity per 100,000 swipes is higher for neighborhoods with majority Black residents regardless of whether poverty rate was high or not. Poverty rate may be a factor that influences the arrest intensity, but it cannot be the only variable that explains the differences in arrest intensity.

```
t1_povrt <- with(stations,
                 tapply(povrt_all_2016,
                        list("High Poverty" = highpov,
                             "Predominantly Black" = nblack),
                        mean))
t1_povrt_wtd <-
  tapply(stations$povrt_all_2016 * stations$swipes2016,
         list(stations$highpov,
              stations$nblack),
         sum) /
  tapply(stations$swipes2016,
         list(stations$highpov,
              stations$nblack),
         sum)

round(t1_povrt_wtd, 2)
```

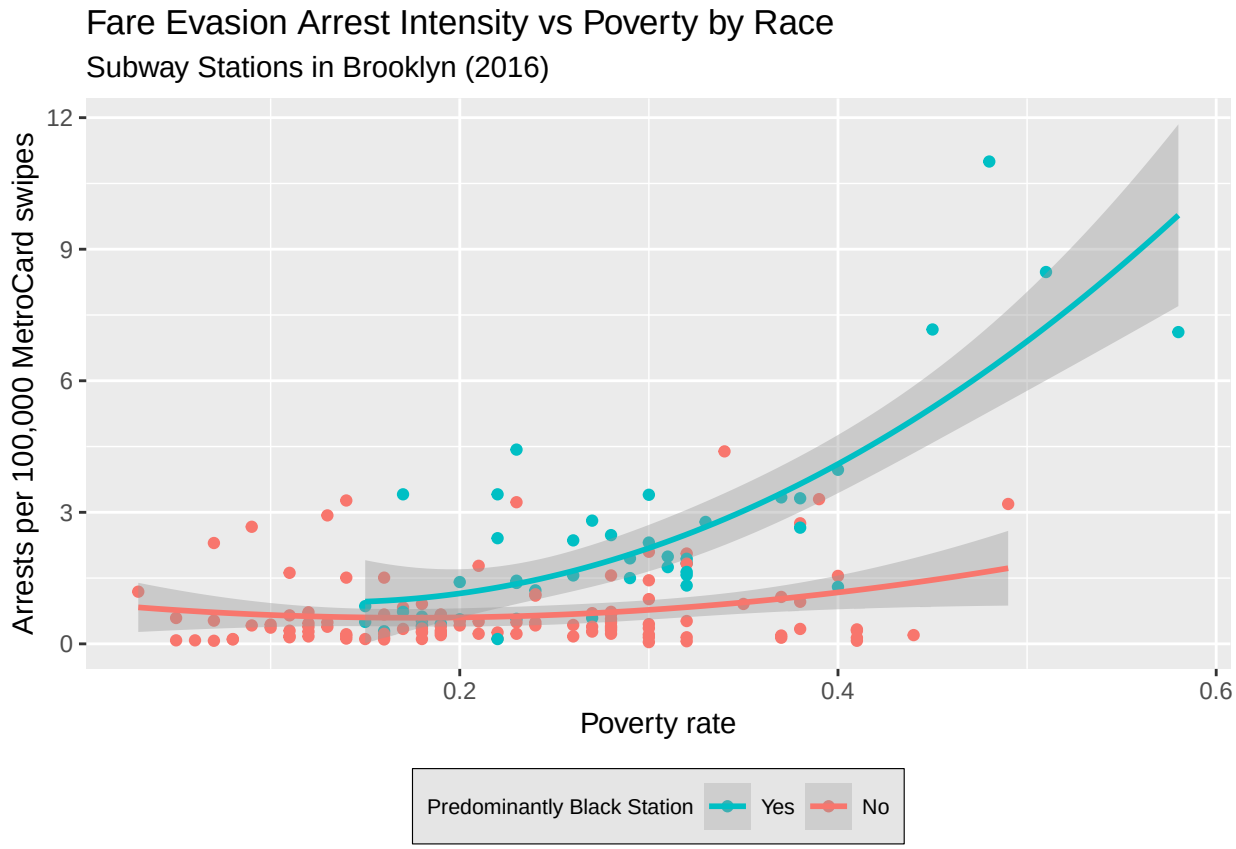
```
##                  Majority non-Black Majority Black
## Not high poverty               0.13           0.19
## High poverty                   0.32           0.32
```

Here, I calculated the mean differences of poverty rates and compared it with the mean differences in arrests per 100,000 swipes between majority non-Black and majority Black areas. The poverty rates are similar for both areas, and thus cannot be a great explanation of arrest intensity. On the other hand, the arrest intensity for majority Black areas is much higher than majority non-Black areas.

The above tables show that mean arrests per 100,000 MetroCard swipes are more than 3 times as high at subway stations in majority Black areas compared to non-Black areas. Poverty rates, on the other hand, are very similar between majority-Black and non-Black high-poverty subway station areas, suggesting it is not a likely explanation for the difference in fare evasion arrest intensity. A regression analysis could help explore how the relationship between poverty rates and fare evasion differs based on neighborhood racial composition.

```
ggplot(stations, aes(x = povrt_all_2016, y = arrperswipe_2016, color = nblack)) +
  geom_point()  +
  geom_smooth(method = 'lm', formula = y ~ x + I(x^2)) +
  ylab("Arrests per 100,000 MetroCard swipes") +
  xlab("Poverty rate") +
  ggtitle("Fare Evasion Arrest Intensity vs Poverty by Race",
          subtitle = "Subway Stations in Brooklyn (2016)") +
  scale_color_discrete(name = "Predominantly Black Station",
                       labels=c("No", "Yes"),
                       guide = guide_legend(reverse=TRUE)) +
  theme(legend.position = "bottom",
```

```
        legend.background = element_rect(color = "black",
                                         fill = "grey90",
                                         size = .2,
                                         linetype = "solid"),
        legend.direction = "horizontal",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8) )
```




```
#get separate data frames by predominantly Black stations to estimate separate models
stations_black <- stations %>% filter(nblack == "Majority Black")
stations_nonblack <- stations %>% filter(nblack == "Majority non-Black")

#nblack == 1: linear model with station observations
ols_b_l <- lm(arrperswipe_2016 ~ povrt_all_2016,
data = stations_black)
summary(ols_b_l)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ povrt_all_2016, data = stations_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0929 -0.8625 -0.3157  0.4367  5.2176
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.5547     0.6589  -3.877 0.000349 ***
## povrt_all_2016  17.3689     2.2057   7.875 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.448 on 44 degrees of freedom
## Multiple R-squared:  0.5849, Adjusted R-squared:  0.5755
## F-statistic: 62.01 on 1 and 44 DF,  p-value: 6.111e-10
```

```
#nblack == 1: quadratic model with station observations
ols_b_q <- lm(arrperswipe_2016 ~ povrt_all_2016 + I(povrt_all_2016^2),
data = stations_black)
summary(ols_b_q)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ povrt_all_2016 + I(povrt_all_2016^2),
##     data = stations_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8004 -0.6455 -0.2596  0.2268  4.7304
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.731      1.644   1.053  0.29813
## povrt_all_2016       -11.737     10.560  -1.112  0.27253
## I(povrt_all_2016^2)   44.150     15.713   2.810  0.00743 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.346 on 43 degrees of freedom
## Multiple R-squared:  0.6493, Adjusted R-squared:  0.633
## F-statistic: 39.81 on 2 and 43 DF,  p-value: 1.642e-10
```

```
#nblack == 0: linear model with station observations
ols_nb_l <- lm(arrperswipe_2016 ~ povrt_all_2016,
data = stations_nonblack)
summary(ols_nb_l)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ povrt_all_2016, data = stations_nonblack)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9411 -0.4738 -0.2912  0.0913  3.4798
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.4200     0.1935   2.171   0.0321 *
```

```
## povrt_all_2016   1.4418     0.7913   1.822   0.0712 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8498 on 108 degrees of freedom
## Multiple R-squared:  0.02982,    Adjusted R-squared:  0.02084
## F-statistic:  3.32 on 1 and 108 DF,  p-value: 0.07123
```

```
#nblack == 0: quadratic model with station observations
ols_nb_q <- lm(arrperswipe_2016 ~ povrt_all_2016 + I(povrt_all_2016^2),
data = stations_nonblack)
summary(ols_nb_q)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ povrt_all_2016 + I(povrt_all_2016^2),
##     data = stations_nonblack)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1954 -0.4587 -0.2425  0.0692  3.4834
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.9399     0.3716   2.530   0.0129 *
## povrt_all_2016        -3.9510     3.3919  -1.165   0.2467
## I(povrt_all_2016^2)   11.3320     6.9337   1.634   0.1051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8433 on 107 degrees of freedom
## Multiple R-squared:  0.05345,    Adjusted R-squared:  0.03576
## F-statistic: 3.021 on 2 and 107 DF,  p-value: 0.05293
```

Quadratic results are shown here because it explains a greater share of the variation in fare evasion arrest intensity for predominantly Black station areas than the linear model (0.63 compared to 0.58), but the same substantive conclusion holds regardless of functional form.

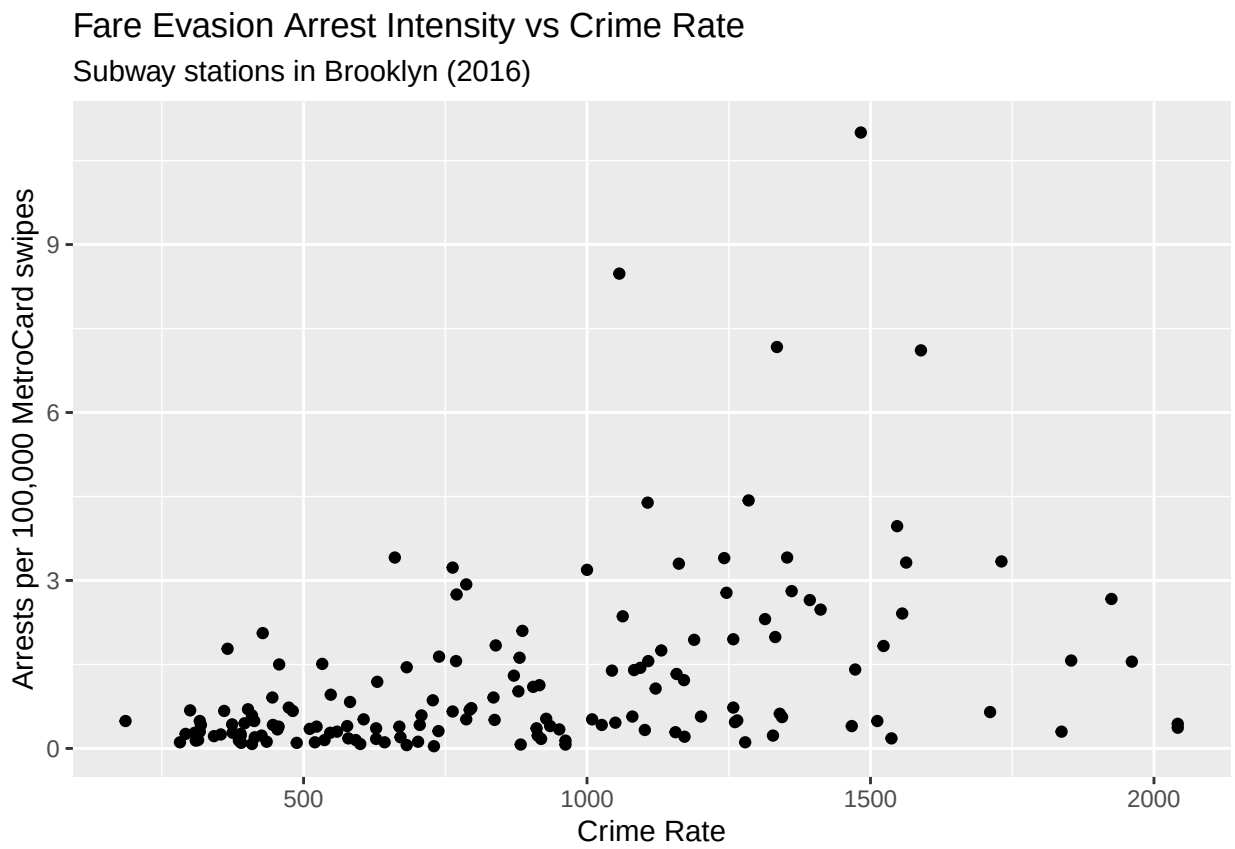
The fitted regression lines show a clear pattern for both the linear and quadratic specifications: fare evasion arrest intensity increases (at an increasing rate) along with poverty rates at subway stations in predominantly Black areas, but not at other stations. In other words, a predominantly Black station area tends to experience significantly higher arrest intensity than a non-Black station with a similarly high poverty rate. Note that the above interpretation is qualitative in nature because providing numerical interpretation of coefficient estimates is easier with a linear model. Alternatively, it would be informative to compare predicted fare evasion arrest intensity for a predominantly Black station area with a specified poverty rate i.e. 40%, compared to a non-Black station area with the same poverty rate.

For both quadratic and linear models, poverty rates explain very little of the variation in arrest intensity among non-Black station areas in Brooklyn (0.04 and 0.02, respectively). This may be because, regardless of the functional form, poverty is only a statistically significant determinant of fare evasion arrest intensity at subway stations in predominantly Black station areas.

12 Relationship between arrest intensity and crime

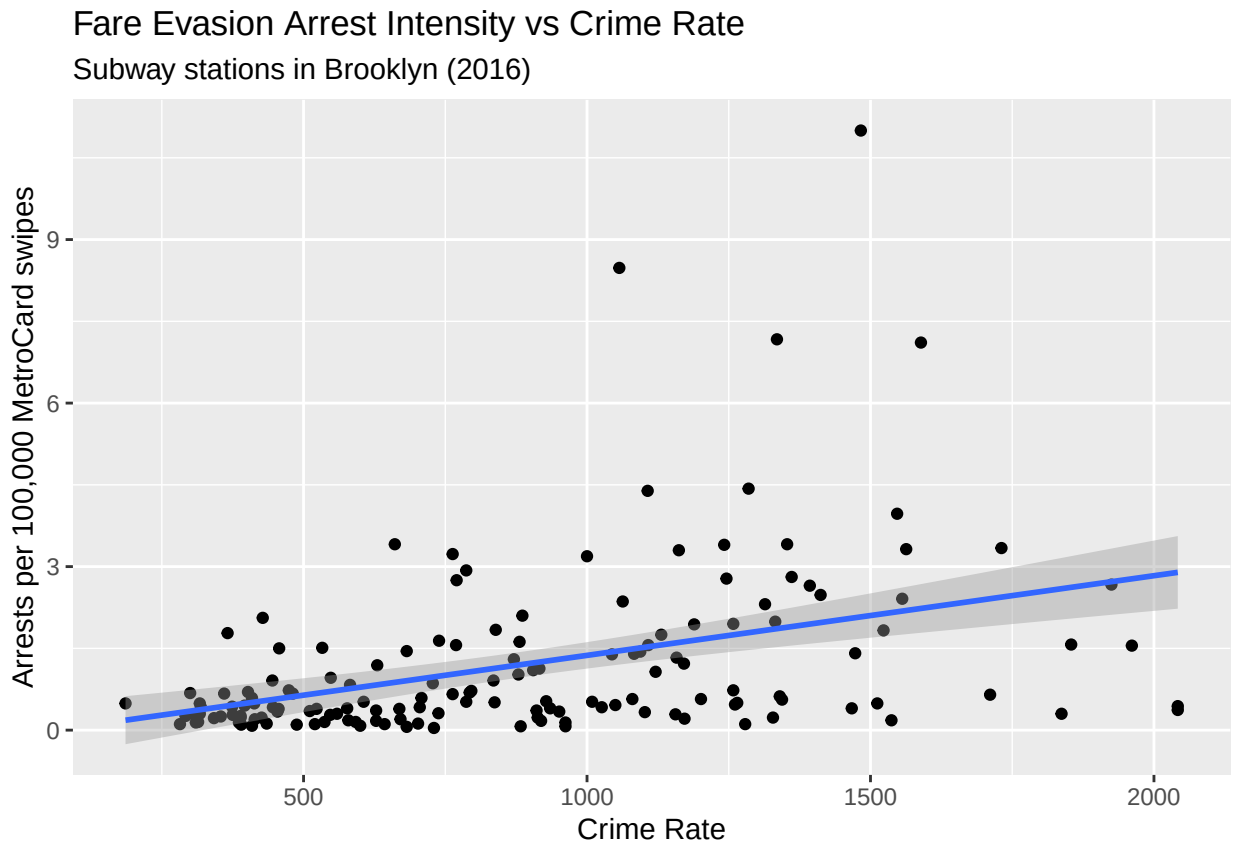
```
st_crime <- read.csv("nypd_criminalcomplaints_2016.csv")
```

```
stations_wcrime <- stations %>%
  inner_join(st_crime) %>%
  arrange(desc(crimes))

cutoffs <- stations_wcrime %>%
  select(crimes)

#exclude the stations with the 4 highest counts of criminal complaints
stations_wcrime <- stations_wcrime %>%
  filter(crimes < cutoffs$crimes[4])

ggplot(stations_wcrime, aes(x = crimes, y = arrperswipe_2016)) +
  geom_point() +
  ylab("Arrests per 100,000 MetroCard swipes") + xlab("Crime Rate") +
  ggtitle("Fare Evasion Arrest Intensity vs Crime Rate",
          subtitle = "Subway stations in Brooklyn (2016)") +
  theme(legend.position = "bottom",
        legend.background = element_rect(color = "black",
                                         fill = "grey90",
                                         size = .2),
        legend.direction = "horizontal",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8))
```

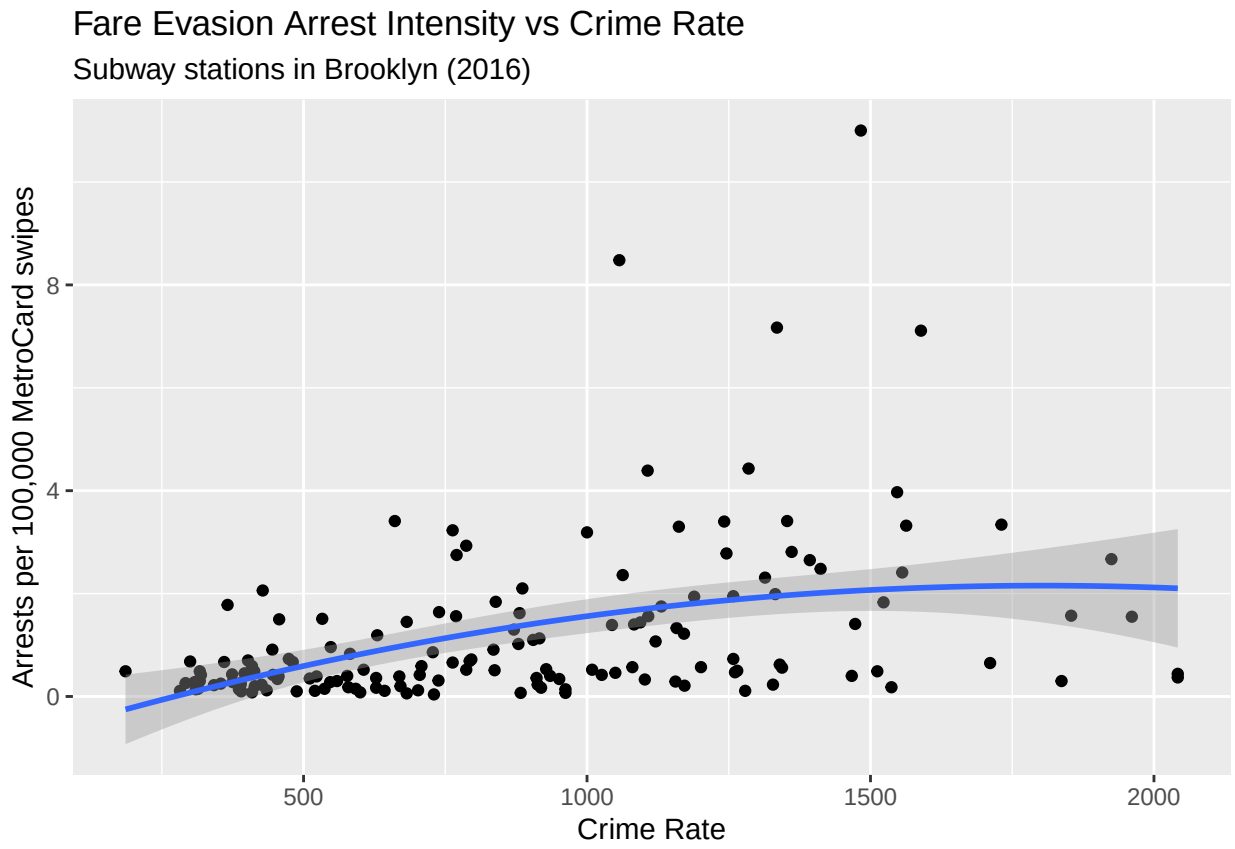


The scatterplot shows a clear curvature shape, with an upward sloping pattern until it reaches the peak near 1,500 crime complaints then a downward sloping pattern. I will now add a line of fit to scatter plots with linear and quadratic models to see which fits best.

```
#scatter plot that does not vary by nblack with linear plots
ggplot(stations_wcrime, aes(x = crimes, y = arrperswipe_2016)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x) +
  ylab("Arrests per 100,000 MetroCard swipes") + xlab("Crime Rate") +
  ggtitle("Fare Evasion Arrest Intensity vs Crime Rate",
          subtitle = "Subway stations in Brooklyn (2016)") +
  theme(legend.position = "bottom",
        legend.background = element_rect(color = "black",
                                         fill = "grey90",
                                         size = .2,
                                         linetype = "solid"),
        legend.direction = "horizontal",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8))
```




```
#w/ quadratic plots
ggplot(stations_wcrime, aes(x = crimes, y = arrperswipe_2016)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x + I(x^2)) +
  ylab("Arrests per 100,000 MetroCard swipes") + xlab("Crime Rate") +
  ggtitle("Fare Evasion Arrest Intensity vs Crime Rate",
          subtitle = "Subway stations in Brooklyn (2016)") +
  theme(legend.position = "bottom",
        legend.background = element_rect(color = "black",
                                         fill = "grey90",
                                         size = .2,
                                         linetype = "solid"),
        legend.direction = "horizontal",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8))
```




```
ols_c_l <- lm(arrperswipe_2016 ~ crimes, data = stations_wcrime)
summary(ols_c_l)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ crimes, data = stations_wcrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5242 -0.6949 -0.2219  0.2597  8.9226
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0896487  0.2672311  -0.335    0.738
## crimes       0.0014612  0.0002719   5.374 2.88e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.46 on 150 degrees of freedom
## Multiple R-squared:  0.1615, Adjusted R-squared:  0.1559
## F-statistic: 28.88 on 1 and 150 DF,  p-value: 2.882e-07
```

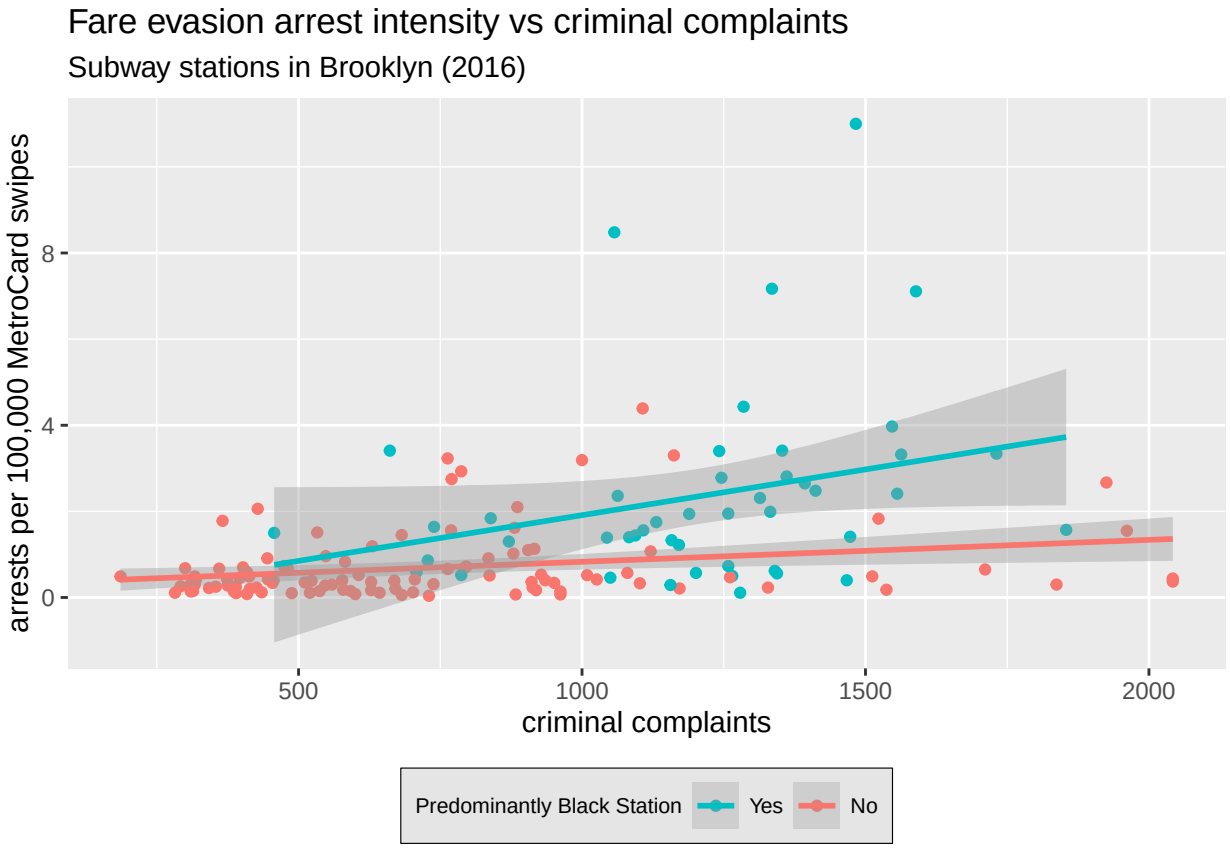
```
ols_c_q <- lm(arrperswipe_2016 ~ crimes + I(crimes^2), data = stations_wcrime)
summary(ols_c_q)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ crimes + I(crimes^2), data = stations_wcrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9093 -0.7507 -0.2241  0.3555  8.9398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.362e-01  5.221e-01  -1.602  0.11138
## crimes       3.316e-03  1.149e-03   2.886  0.00448 **
## I(crimes^2) -9.192e-07  5.534e-07  -1.661  0.09883 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.451 on 149 degrees of freedom
## Multiple R-squared:  0.1767, Adjusted R-squared:  0.1657
## F-statistic: 15.99 on 2 and 149 DF,  p-value: 5.116e-07
```

Regardless of the functional form, criminal complaints explain about 16% of the variation in fare evasion arrest intensity across subway stations in Brooklyn (0.166 and 0.156 for quadratic and linear models, respectively). From the linear model, we can see that the effect of criminal complaints on arrest intensity (0.0015) is statistically significant beyond the 1% level (p-value = 0).

Now, I will examine how neighborhood racial composition mediates the relationship between arrest intensity and crime rates using the linear regression model for the ease of interpretation.

```
ggplot(stations_wcrime, aes(x = crimes, y = arrperswipe_2016, color = nblack)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x) +
  ylab("arrests per 100,000 MetroCard swipes") + xlab("criminal complaints") +
  ggtitle("Fare evasion arrest intensity vs criminal complaints",
          subtitle = "Subway stations in Brooklyn (2016)") +
  scale_color_discrete(name = "Predominantly Black Station",
                       labels=c("No", "Yes"),
                       guide = guide_legend(reverse=TRUE)) +
  theme(legend.position = "bottom",
        legend.background = element_rect(color = "black", fill = "grey90",
                                         size = .2, linetype = "solid"),
        legend.direction = "horizontal",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8))
```




```
#get separate data frames by predominantly Black stations to estimate separate models
stations_wcrime_black <- stations_wcrime %>%
filter(nblack == "Majority Black")
stations_wcrime_nonblack <- stations_wcrime %>%
filter(nblack == "Majority non-Black")

#nblack == 1: linear model with station observations
ols_c_b_l <- lm(arrperswipe_2016 ~ crimes, data = stations_wcrime_black)
summary(ols_c_b_l)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ crimes, data = stations_wcrime_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5037 -1.0261 -0.4570  0.3065  8.0623
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.212305   1.381094  -0.154   0.8785
## crimes       0.002124   0.001112   1.910   0.0627 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.16 on 44 degrees of freedom
```

```
## Multiple R-squared:  0.07653,    Adjusted R-squared:  0.05554
## F-statistic: 3.646 on 1 and 44 DF,  p-value: 0.06272
```

```r
#nblack == 0: linear model with station observations
ols_c_nb_l <- lm(arrperswipe_2016 ~ crimes, data = stations_wcrime_nonblack)
summary(ols_c_nb_l)
```

```
##
## Call:
## lm(formula = arrperswipe_2016 ~ crimes, data = stations_wcrime_nonblack)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9898 -0.4468 -0.2539  0.1684  3.5064
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.3197155  0.1589837   2.011  0.04691 *
## crimes      0.0005093  0.0001879   2.711  0.00784 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7969 on 104 degrees of freedom
## Multiple R-squared:  0.06601,    Adjusted R-squared:  0.05703
## F-statistic: 7.351 on 1 and 104 DF,  p-value: 0.007844
```

Estimating separate linear models for the relationship between criminal complaints and arrest intensity for predominantly Black and non-Black station areas reveals a similar pattern as with poverty rates, but with less pronounced differences. The linear relationship between criminal complaints and arrest intensity explains under 6% of the variation regardless of neighborhood racial composition, but the estimated positive effect is four times as large in predominantly Black station areas (0.002 compared to 0.0005) and statistically significant at the 5% level (p-value = 0.0627).

13 Conclusion

The results presented here are consistent with race-based enforcement of fare evasion at subway stations in Brooklyn. As the poverty rate for a subway station area increases, fare evasion arrest intensity tends to increase in predominantly Black station areas (and the association is statistically significant) but not in non-Black station areas.

A similar pattern holds for criminal complaints and fare evasion arrest intensity, though the disparities based on neighborhood racial composition are far less pronounced.

One additional test worth doing is confirming that the positive association between poverty rates and fare evasion arrest intensity in predominantly Black neighborhoods is still statistically significant when simultaneously controlling for criminal complaints (but not in non-Black neighborhoods). This test will confirm that regardless of where the NYPD enforcement of other crimes is more prevalent, higher poverty Black neighborhoods face considerably higher fare evasion arrests than similarly higher poverty neighborhoods that are not predominantly Black.

The results of this analysis are consistent with disproportionately enforcing fare evasion as a crime of poverty in Black communities. In other words, the totality of NYPD policing decisions result in heightened enforcement of fare evasion in higher-poverty, predominantly Black neighborhoods. This analysis does not, however,

inform the relative importance of different mechanisms driving these patterns: policy deployment decision, implicit and/or explicit bias in the decision to stop people and the subsequent enforcement action (arrest vs summons), or some combination. There may also be other differences in subway rider characteristics and behavior that could explain the observed relationship between neighborhood racial composition and fare evasion enforcement intensity, but disparate impact by race is clear even if the all of the underlying mechanisms are not.

Analyzing differences in fare evasion summonses compared to arrests would also be informative: are there significant differences in the demographics of individuals who are stopped for fare evasion, in addition to differences in the enforcement action taken once they are stopped?