# U6614: Subway Fare Evasion Arrests and Racial Bias

Sample Solution

2020-10-12

***Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on Monday, October 5th.***

## 1 Load libraries

```
library(tidyverse)
library(weights)
library(lmtest)
library(sandwich)
```

## 2 Subway station-level arrest totals

### 2.1 Load cleaned/appended arrest microdata (arrests_all.csv) w/all strings as factors.

```
arrests_all <- read.csv("arrests_all.csv",
                        stringsAsFactors = TRUE,
                        na.strings = c("")) #can specify string values to read in as NA (here blanks)
```

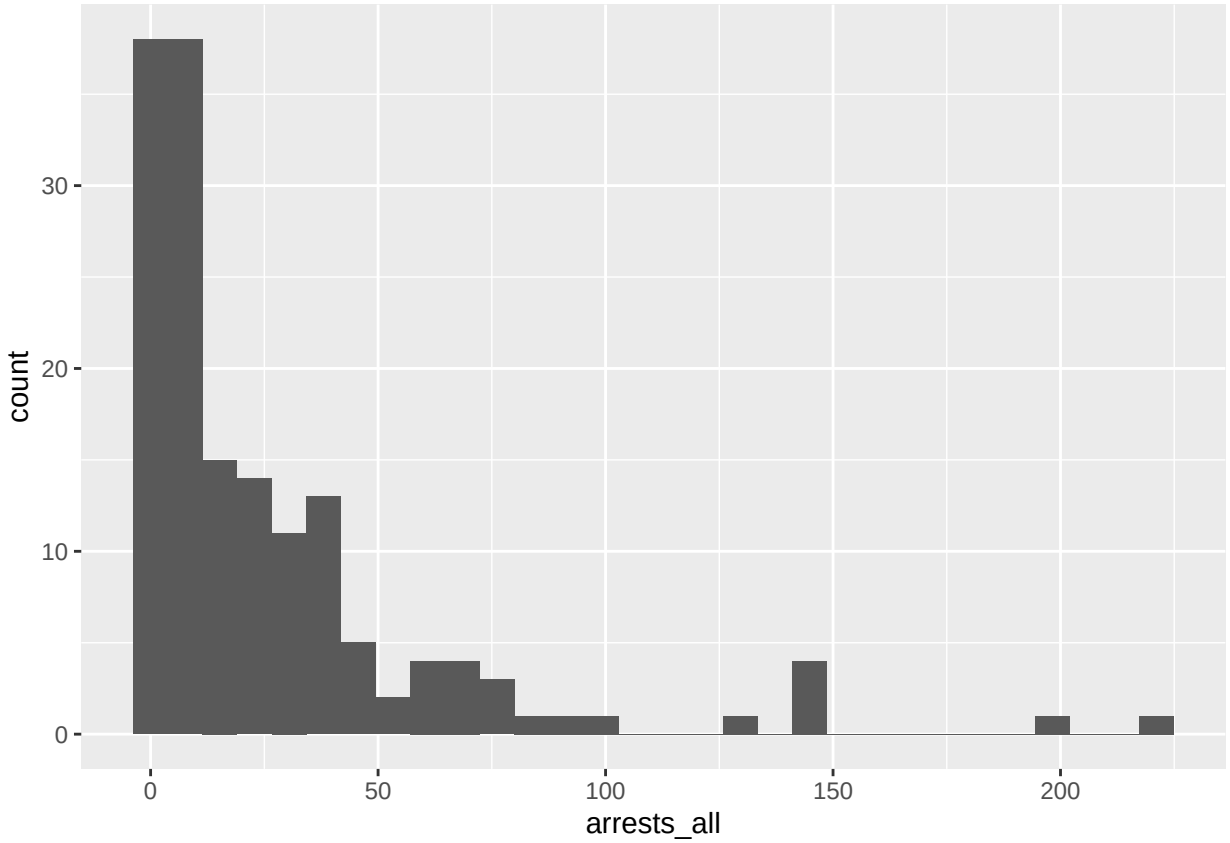
### 2.2 Create new data frame (`st_arrests`) that aggregates microdata to station-level observations including the following information.

- *st_id, loc2, total arrests*

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

### 2.3 Plot histogram of arrests and briefly describe the distribution of arrests across stations.

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




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.

# 3 Joining subway ridership and neighborhood demographic data

## 3.1 Import and inspect secondary data sources (*station_povdataclean_2016.csv* and *Subway Ridership by Station - BK.csv*).

```
st_poverty <- read.csv("station_povdataclean_2016.csv",
                       stringsAsFactors = TRUE)

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

str(st_poverty)
```

```
## 'data.frame':    157 obs. of  10 variables:
##  $ x             : num  -74 -74 -74 -74 -74 ...
##  $ y             : num  40.7 40.6 40.6 40.6 40.6 ...
##  $ st_id         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ mta_name      : Factor w/ 157 levels "15 St-Prospect Park F subway G subway",..: 1 2 3 4 5 6 7 8 9
##  $ pop_black_2016: int  306 184 232 202 104 33 237 192 125 841 ...
##  $ pov_black_2016: int  51 110 161 2 71 1 84 24 41 211 ...
```

```
##  $ pop_all_2016  : int  14094 17846 17528 15560 14788 14811 18356 8719 8090 11657 ...
##  $ pov_all_2016  : int  1060 3325 5685 4125 2738 3341 3342 1674 2600 1390 ...
##  $ povrt_all_2016: num  0.0752 0.1863 0.3243 0.2651 0.1852 ...
##  $ shareblack    : num  0.02171 0.01031 0.01324 0.01298 0.00703 ...
```

```
str(st_ridership)
```

```
## 'data.frame':    157 obs. of  8 variables:
##  $ st_id     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ mta_name  : Factor w/ 157 levels "15 St-Prospect Park F subway G subway",..: 1 2 3 4 5 6 7 8 9 10
##  $ swipes2011: int  1449241 1654305 1306839 1828462 1370001 1220065 1266146 1243737 3659139 4020665
##  $ swipes2012: int  1852346 1661770 1294568 1819255 1377648 1245145 1206526 1198226 3772420 4300400
##  $ swipes2013: int  2005424 1954729 1313344 1641928 1547029 1286961 1421090 1218842 3826400 4210873
##  $ swipes2014: int  2017347 1996205 1277024 1775695 1564060 1305487 1488192 1274939 3970318 4120946
##  $ swipes2015: int  2011237 1925239 1392552 1784992 1592525 1279525 1508412 1221949 4138929 4199208
##  $ swipes2016: int  1958444 2003777 1344286 723187 1656437 409416 1992945 1188884 4235509 4138758 ..
```

### 3.2 Join ridership and demographic data to `st_arrests` and inspect results (store new data frame as `st_joined`).

```
st_joined <- st_poverty %>%
  inner_join(st_ridership) %>%
  inner_join(st_arrests)
```

- *Inspect results from joins, drop unnecessary columns from the ridership data, and group* ***`st_joined`*** *by st_id and mta_name.*

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

st_joined_grouped <- st_joined %>%
  select(!drop_vars) %>%
  group_by(st_id, mta_name)

#display structure of ungrouped data frame to avoid lengthy output listing every group
st_joined_grouped %>% ungroup() %>% str()
```

```
## tibble [157 x 13] (S3: tbl_df/tbl/data.frame)
##  $ x             : num [1:157] -74 -74 -74 -74 -74 ...
##  $ y             : num [1:157] 40.7 40.6 40.6 40.6 40.6 ...
##  $ st_id         : int [1:157] 1 2 3 4 5 6 7 8 9 10 ...
##  $ mta_name      : Factor w/ 157 levels "15 St-Prospect Park F subway G subway",..: 1 2 3 4 5 6 7 8 9
##  $ pop_black_2016: int [1:157] 306 184 232 202 104 33 237 192 125 841 ...
##  $ pov_black_2016: int [1:157] 51 110 161 2 71 1 84 24 41 211 ...
##  $ pop_all_2016  : int [1:157] 14094 17846 17528 15560 14788 14811 18356 8719 8090 11657 ...
##  $ pov_all_2016  : int [1:157] 1060 3325 5685 4125 2738 3341 3342 1674 2600 1390 ...
##  $ povrt_all_2016: num [1:157] 0.0752 0.1863 0.3243 0.2651 0.1852 ...
##  $ shareblack    : num [1:157] 0.02171 0.01031 0.01324 0.01298 0.00703 ...
##  $ swipes2016    : int [1:157] 1958444 2003777 1344286 723187 1656437 409416 1992945 1188884 4235509
##  $ loc2          : Factor w/ 157 levels "15 st prospect park f g line",..: 1 4 154 5 7 6 10 9 11 12
##  $ arrests_all   : int [1:157] 2 8 2 2 11 2 7 8 22 7 ...
```

```
summary(st_joined_grouped)
```

```
##        x                y             st_id
##  Min.   :-74.03   Min.   :40.58   Min.   :  1
##  1st Qu.:-73.98   1st Qu.:40.63   1st Qu.: 40
```

```
##  Median :-73.96   Median :40.67   Median : 79
##  Mean   :-73.96   Mean   :40.66   Mean   : 79
##  3rd Qu.:-73.93   3rd Qu.:40.69   3rd Qu.:118
##  Max.   :-73.87   Max.   :40.72   Max.   :157
##
##                                    mta_name   pop_black_2016  pov_black_2016
##  15 St-Prospect Park F subway G subway:  1   Min.   :    0   Min.   :   0
##  18 Av  D subway                       :  1   1st Qu.:  398   1st Qu.:  86
##  18 Av  F subway                       :  1   Median : 2399   Median : 616
##  18 Av N subway                        :  1   Mean   : 4579   Mean   :1166
##  20 Av  D subway                       :  1   3rd Qu.: 7504   3rd Qu.:1644
##  20 Av  N subway                       :  1   Max.   :20739   Max.   :6706
##  (Other)                               :151
##   pop_all_2016    pov_all_2016   povrt_all_2016      shareblack
##  Min.   : 2721   Min.   :  308   Min.   :0.03321   Min.   :0.00000
##  1st Qu.:11303   1st Qu.: 1959   1st Qu.:0.16348   1st Qu.:0.02366
##  Median :15167   Median : 3206   Median :0.22859   Median :0.15819
##  Mean   :15250   Mean   : 3775   Mean   :0.24010   Mean   :0.29328
##  3rd Qu.:18371   3rd Qu.: 5241   3rd Qu.:0.30144   3rd Qu.:0.54165
##  Max.   :31071   Max.   :11612   Max.   :0.57800   Max.   :0.88898
##
##    swipes2016                                 loc2      arrests_all
##  Min.   :  406793   15 st prospect park f g line:  1   Min.   :  2.00
##  1st Qu.: 1188884   1523 Avenue U               :  1   1st Qu.:  4.00
##  Median : 1863036   1778 w 7th st               :  1   Median : 13.00
##  Mean   : 2449301   18th av and 85th st         :  1   Mean   : 26.82
##  3rd Qu.: 3027658   18th ave and 64th st        :  1   3rd Qu.: 36.00
##  Max.   :13818168   20th ave and 64th st        :  1   Max.   :223.00
##                     (Other)                     :151
```

### 3.3 Print the top 10 stations by total arrest counts, along with the poverty rate and share black for each subway station area shown.

```
st_joined_grouped %>%
  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          0.580
##  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
```

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

## 4.1 Compute arrest intensity and prep data.

- *Drop the observation for the Coney Island station and very briefly explain your logic*
- *Create new column of data representing fare evasion enforcement intensity (arrperswipe_2016 = arrests per 100,000 swipes)*
- *Create dummy variable indicating whether a station is high poverty (highpov = 1 if pov rate is > median pov rate across all Brooklyn station areas)*
- *Create new dummy for majority Black station areas (nblack = 1 if shareblack > .5)*
- *coerce new dummy variables into factors w/category labels*
- *Assign results to new data frame called* ***stations***

```
stations <- st_joined_grouped %>%
  filter(st_id != 66) %>%
  mutate(arrperswipe = round(arrests_all / (swipes2016 / 100000), 2),
         highpov = as.numeric(povrt_all_2016 > median(st_joined_grouped$povrt_all_2016)),
         nblack = as.numeric(shareblack > .5),
         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")),
         shareblack = round(shareblack, 2),
         povrt_all_2016 = round(povrt_all_2016, 2))

#validation: now check top 10 stations by arrest intensity
stations %>%
  arrange(desc(arrperswipe)) %>%
  select(st_id, mta_name, arrperswipe, arrests_all, shareblack, povrt_all_2016, highpov, nblack) %>%
  head(n = 10)
```

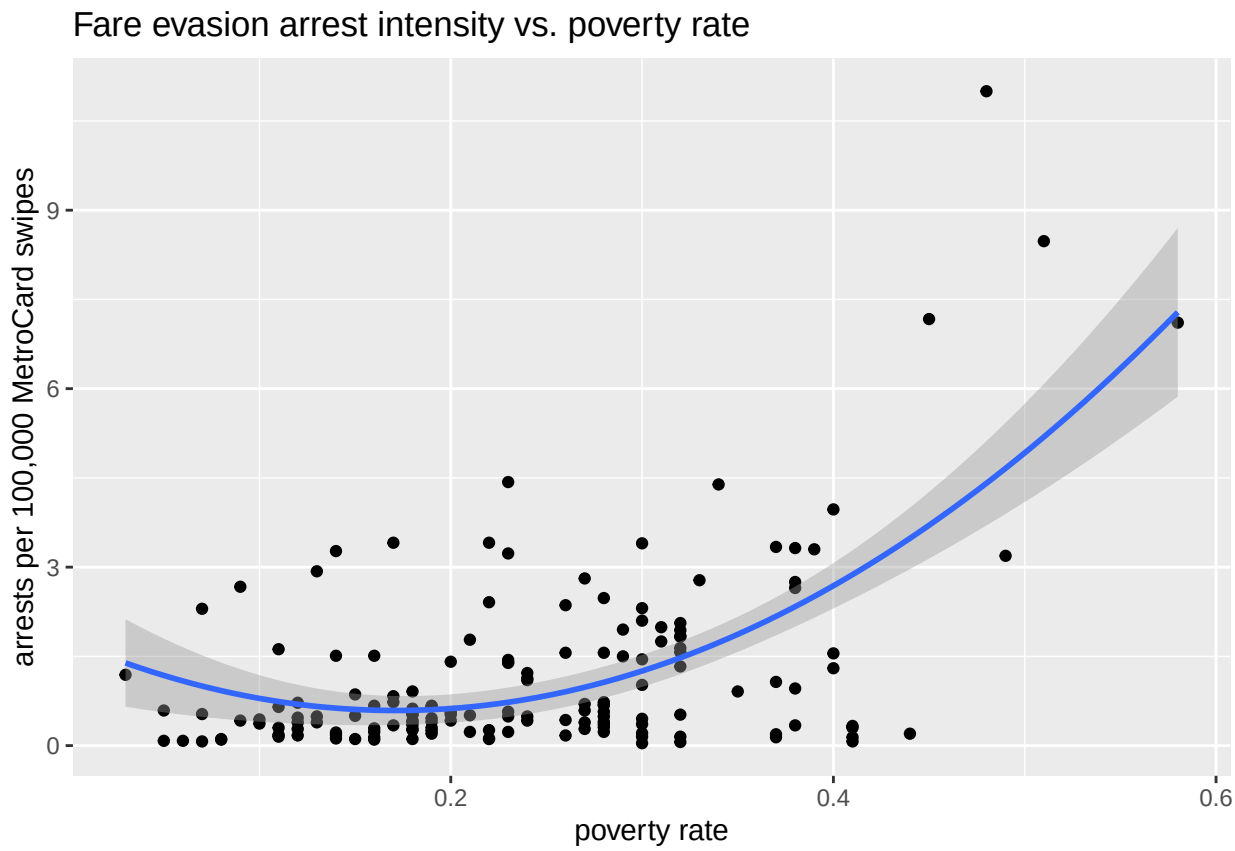
```
## # A tibble: 10 x 8
## # Groups:   st_id, mta_name [10]
##    st_id mta_name arrperswipe arrests_all shareblack povrt_all_2016 highpov
##    <int> <fct>          <dbl>       <int>      <dbl>          <dbl> <fct>
##  1   101 Junius ~       11             75       0.78          0.48  High p~
##  2    26 Atlanti~        8.48          37       0.66          0.51  High p~
##  3   111 Livonia~        7.17          75       0.83          0.45  High p~
##  4   147 Sutter ~        7.11         102       0.75          0.580 High p~
##  5   106 Kingsto~        4.43          90       0.78          0.23  High p~
##  6   112 Lorimer~        4.39          70       0.15          0.34  High p~
##  7   140 Rockawa~        3.97          61       0.78          0.4   High p~
##  8    54 Canarsi~        3.41         133       0.83          0.17  Not hi~
##  9   141 Rockawa~        3.41          61       0.8           0.22  Not hi~
## 10   144 Shepher~        3.4           36       0.61          0.3   High p~
## # ... with 1 more variable: nblack <fct>
```

## 4.2 Examine the relationship between arrest intensity and poverty rates

- *Show a scatterplot of arrest intensity vs. poverty rates along with the regression line you think best fits this relationship*
- *Which regression specification do you prefer: linear or quadratic? Be clear about your logic and if applicable cite statistical evidence to support your decision.*

- *Report diff in mean arrest intensity between high/low pov areas and assess statistical significance*
- *For this assignment: weight observations by the number of MetroCard swipes when calculating difference in group means, but not for your regressions*

```
#quadratic
ggplot(stations, #specify data frame to use
       aes(x = povrt_all_2016, y = arrperswipe)) + #specify columns to use
  geom_point() + #specify plot geometry
  ggtitle('Fare evasion arrest intensity vs. poverty rate') + #add title
  labs(x = 'poverty rate', y = 'arrests per 100,000 MetroCard swipes') + #change axis labels
  geom_smooth(method = 'lm', formula = y ~ x + I(x^2)) #add regression line
```




```
ols1l <- lm(arrperswipe ~ povrt_all_2016, data = stations)
summary(ols1l)
coeftest(ols1l, vcov = vcovHC(ols1l, type="HC1")) #get robust SEs

ols1q <- lm(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2), data = stations)
summary(ols1q)
coeftest(ols1q, vcov = vcovHC(ols1q, type="HC1"))
```

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. We 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.

If you prefer the linear specification because it is a bit simpler to interpret without changing the substantive conclusions, that is a reasonable justification.

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

```r
#regress arrest intensity on highpov dummy to implement diff in means test (weighted, robust SEs)
ols_diff1 <- lm(formula = arrperswipe ~ highpov, data = stations, weights = swipes2016)
ols_diff1_robSE <- coeftest(ols_diff1, vcov = vcovHC(ols_diff1, type="HC1"))
```

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.

# 5 How does neighborhood racial composition mediate the relationship between poverty and arrest intensity? Examine the relationship between arrest intensity & poverty by Black vs non-Black station area (nblack).

## 5.1 Present a table showing the difference in mean arrests per swipe for each group in a 2x2 table of highpov vs nblack.

- *Remember to weight by the number of MetroCard swipes at each station*
- *Could the difference in arrest intensity be explained by differences in poverty rate?*

```r
t1_arrper_wtd <- with(stations,
                      tapply(arrperswipe * swipes2016,
                             list("High Poverty" = highpov, "Predominantly Black" = nblack),
                             mean)/
                      tapply(swipes2016,
                             list("High Poverty" = highpov, "Predominantly Black" = nblack),
                             mean))

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

round(t1_arrper_wtd, 2)
```

```
##                   Predominantly Black
## High Poverty       Majority non-Black Majority Black
```

```
##   Not high poverty               0.66           1.19
##   High poverty                   0.82           2.49
```

```
round(t1_povrt_wtd, 2)
```

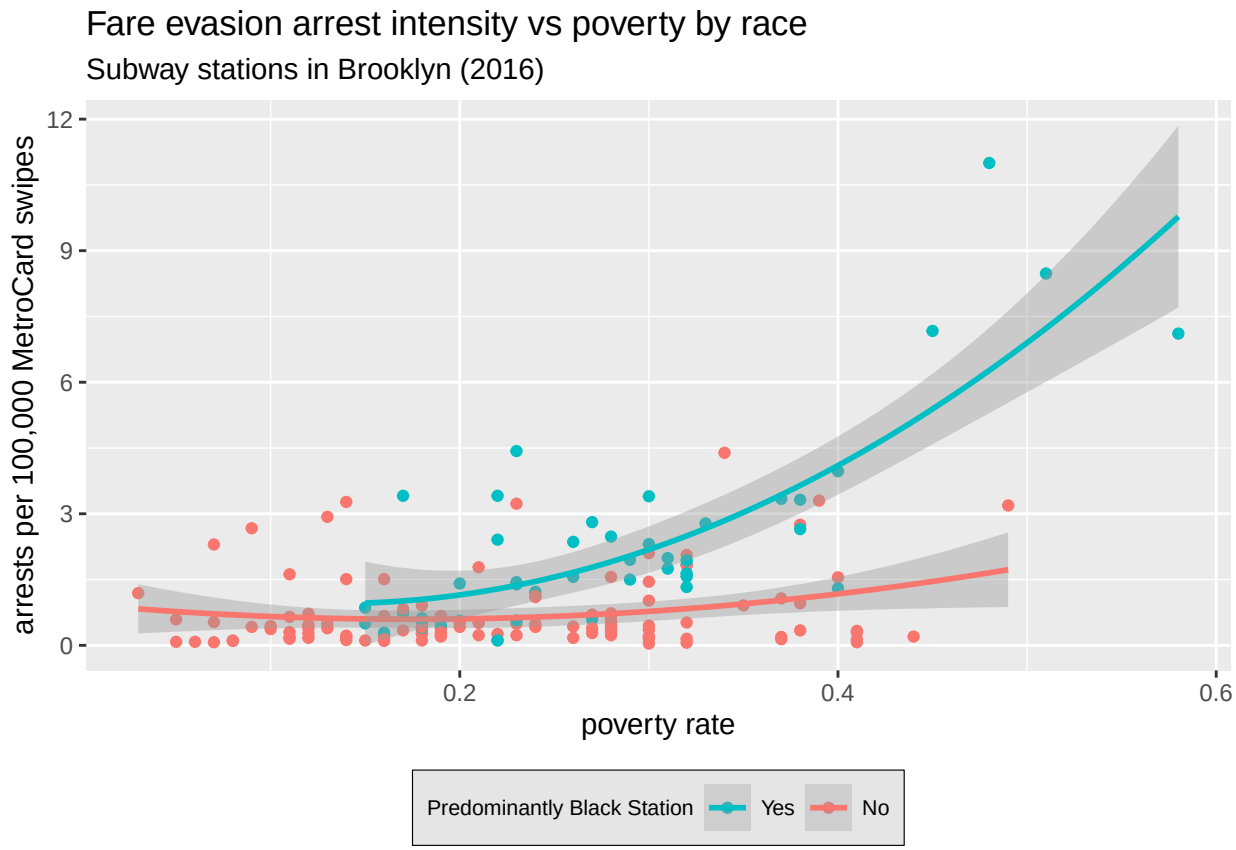
```
##                   Predominantly Black
## High Poverty       Majority non-Black Majority Black
##   Not high poverty               0.13           0.19
##   High poverty                   0.32           0.32
```

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 this is not a likely explanation for the difference in fare evasion arrest intensity (but we can use regression analysis to explore how the relationship between poverty rates and fare evasion differs based on neighborhood racial composition).

## 5.2 Show a scatterplot of arrest intensity vs. poverty rates along with the regression line you think best fits this relationship.

- *Which regression specification do you prefer: linear or quadratic? Be clear about your logic and if applicable cite statistical evidence to support your decision.*
- *Interpret your preferred regression specification (carefully!).*

```
#quadratic
ggplot(stations, aes(x = povrt_all_2016, y = arrperswipe, 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 ~ povrt_all_2016, data = stations_black)

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

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

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

Visual inspection of the fitted regression lines reveal 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. Said another way, the result suggest that 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: it's a bit more straightforward to provide a numerical interpretation of coefficient estimates 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 (say, 40%) compared to a non-Black station area with the same poverty rate. If you prefer the linear specification because it is a bit simpler to interpret without changing the substantive conclusions, that

is a reasonable justification.

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.

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).

Regardless of functional form, poverty is only a statistically significant determinant of fare evasion arrest intensity at subway stations in predominantly Black station areas.

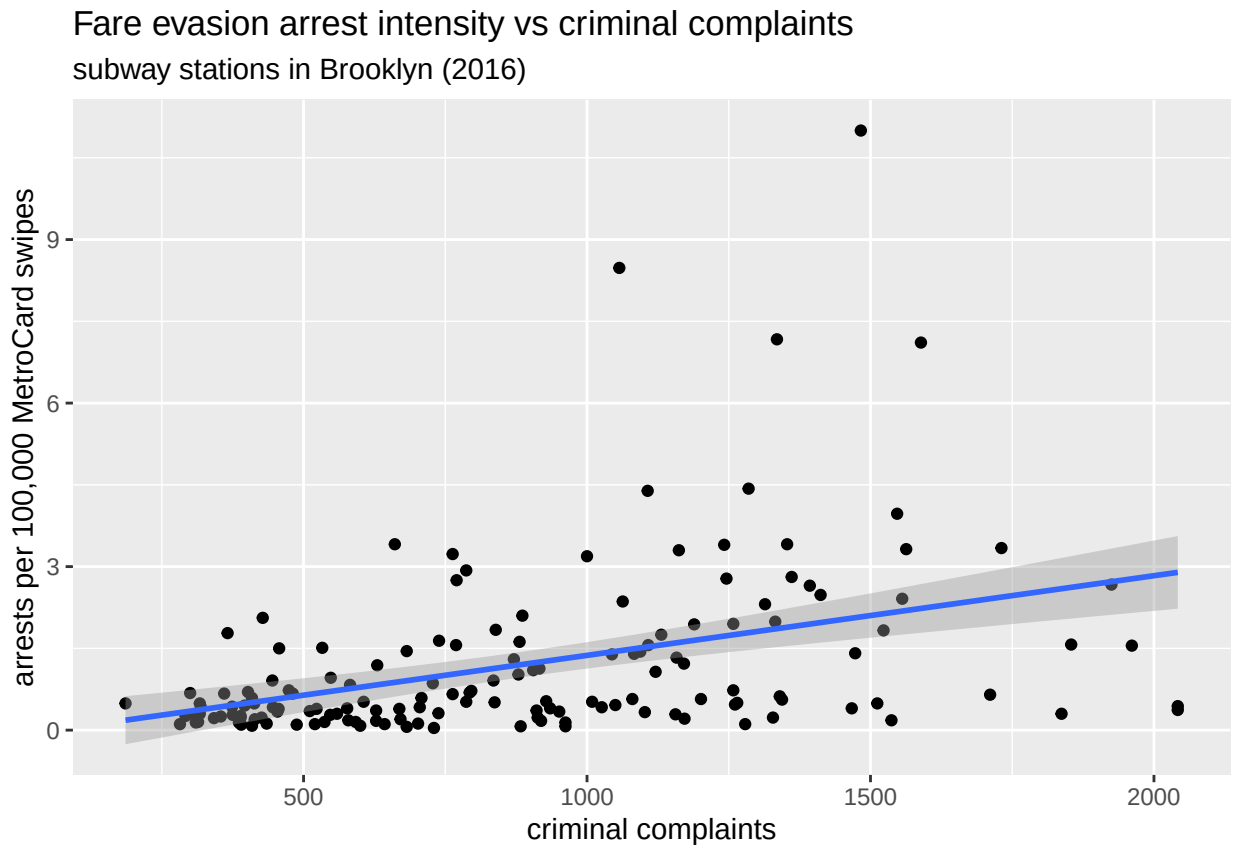
# 6 Examine the relationship between arrest intensity and crime

## 6.1 Load the crime data and join to the existing stations data frame (*nypd_criminalcomplaints_2016.csv*)

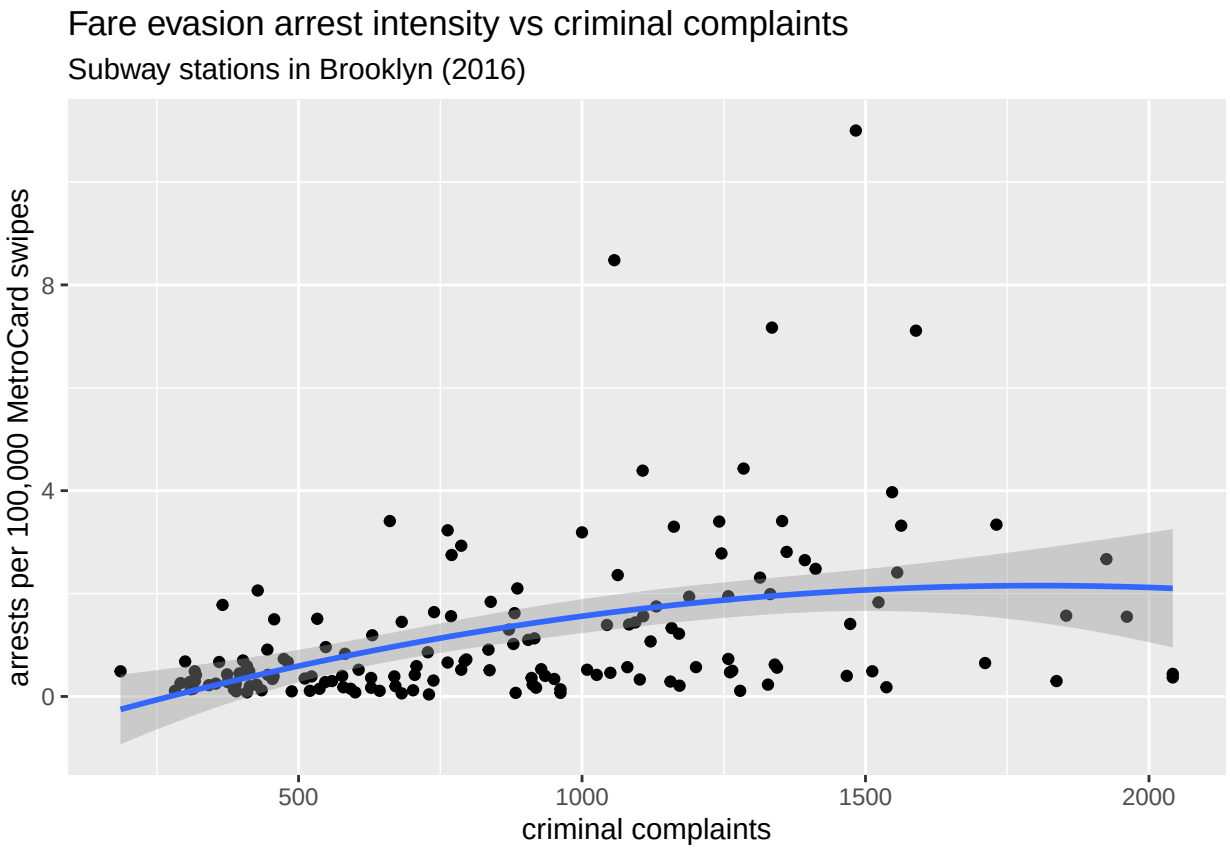
```
st_crime <- read.csv("nypd_criminalcomplaints_2016.csv")

stations_wcrime <- stations %>%
  inner_join(st_crime) %>%
  arrange(desc(crimes)) %>%
  filter(crimes < 2367) #exclude the stations with the 4 highest counts of criminal complaints
```

## 6.2 Examine the overall relationship between arrest intensity and crime (without taking neighborhood racial composition or poverty into account) (comparable to Section 4.2).

```
#linear
ggplot(stations_wcrime, aes(x = crimes, y = arrperswipe)) +
    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 = "soli
          legend.direction = "horizontal",
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8))
```




```
#quadratic
ggplot(stations_wcrime, aes(x = crimes, y = arrperswipe)) +
    geom_point()  +
    geom_smooth(method = 'lm', formula = y ~ x + I(x^2)) +
    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 = "soli
          legend.direction = "horizontal",
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8))
```




```
ols_c_l <- lm(arrperswipe ~ crimes, data = stations_wcrime)
ols_c_l_robSE <- coeftest(ols_c_l, vcov = vcovHC(ols_c_l, type="HC1")) #get robust SEs

ols_c_q <- lm(arrperswipe ~ crimes + I(crimes^2), data = stations_wcrime)
ols_c_q_robSE <- coeftest(ols_c_q, vcov = vcovHC(ols_c_q, type="HC1")) #get robust SEs
```

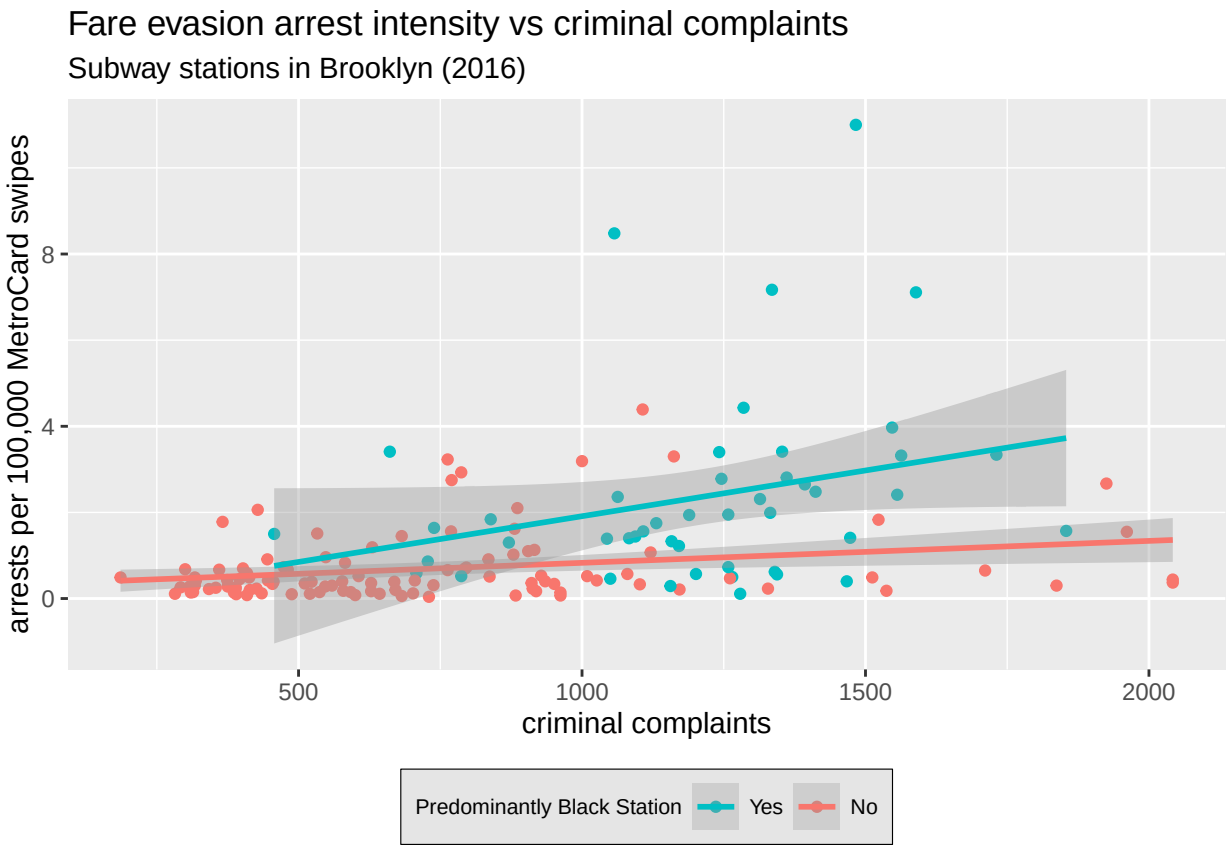
Regardless of the functional form, criminal complaints explain about 16% of the variation in fare evasion arrest intensity across subway stations in Brooklyn (0.166 and 0.156 for quadratic and linear models, respectively).

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).

### 6.3 Examine how neighborhood racial composition mediates the relationship between arrest intensity and crime (comparable to Section 5.2).

```
#linear
ggplot(stations_wcrime, aes(x = crimes, y = arrperswipe, 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 = "soli
          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 ~ crimes, data = stations_wcrime_black)
ols_c_b_l_robSE <- coeftest(ols_c_b_l, vcov = vcovHC(ols_c_b_l, type="HC1"))

#nblack == 1: quadratic model with station observations
ols_c_b_q <- lm(arrperswipe ~ crimes + I(crimes^2), data = stations_wcrime_black)

#nblack == 0: linear model with station observations
ols_c_nb_l <- lm(arrperswipe ~ crimes, data = stations_wcrime_nonblack)
ols_c_nb_l_robSE <- coeftest(ols_c_nb_l, vcov = vcovHC(ols_c_nb_l, type="HC1"))

#nblack == 0: quadratic model with station observations
ols_c_nb_q <- lm(arrperswipe ~ crimes + I(crimes^2), data = stations_wcrime_nonblack)
```

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 less

pronounced differences.

Focusing on the linear model for ease of interpretation: 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 0.001) and statistically significant at the 5% level (p-value = 0.0127).

# 7 Summarize and interpret your findings with respect to subway fare evasion enforcement bias based on race

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 to confirm 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 confirms 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; 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: do police deployment decision explain these disparities, implicit and/or explicit bias in who is stopped and what enforcement action is taken (arrest vs summons), or some combination of these mechanisms? 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?