Water shutoffs in Detroit

Myung Eun Hyeon

2023-09-18

Load libraries

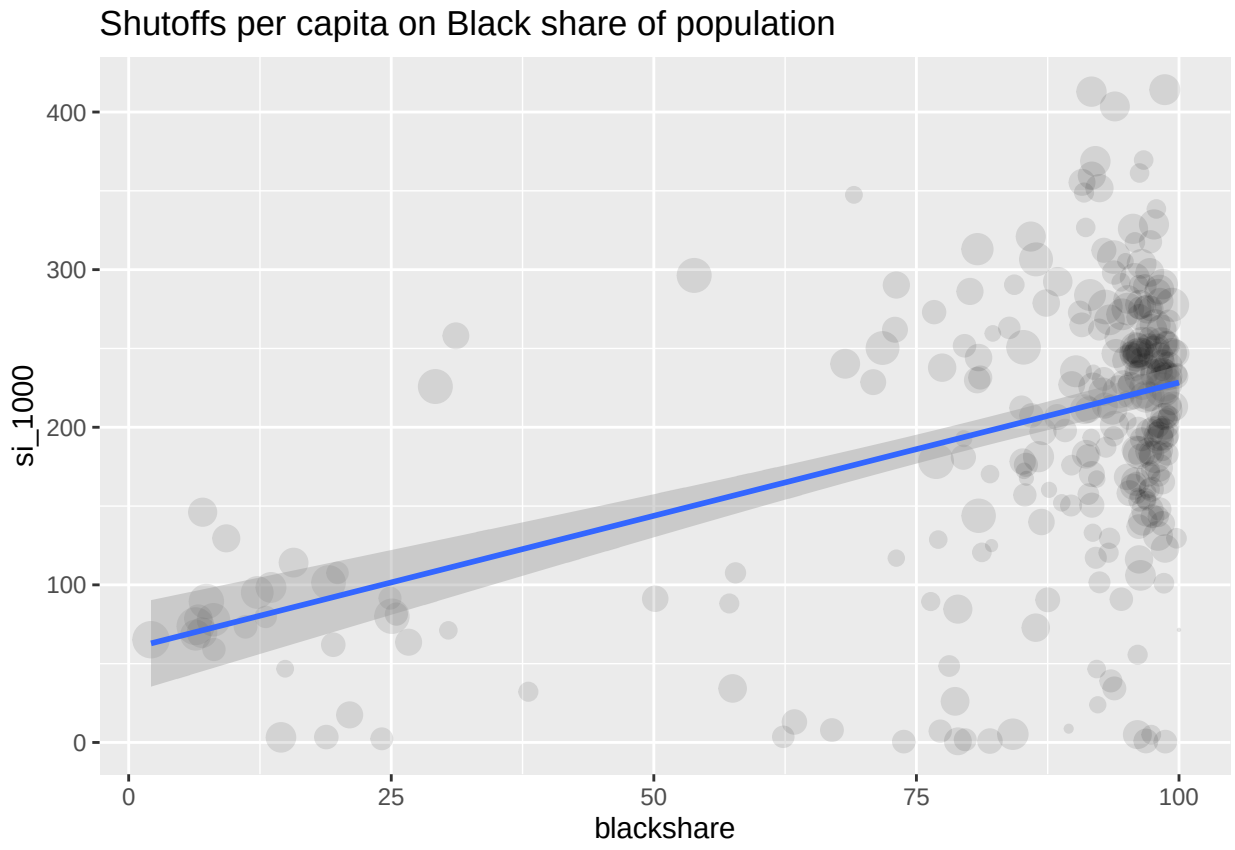
```
library(readstata13)
library(tidyverse)
library(lubridate)
library(weights)
```

1 "Cross-sectional" analysis

```
input_acs_tract <- read.dta13("ACS_10_17_5YR_CensusTract.dta")

acs_tract.clean <- input_acs_tract %>%
  select(census_tract_long, year, num_pop_total, num_income_median,
         per_race_black_alone_or_combo, geodisplaylabel) %>%
  rename(tractid = census_tract_long,
         pop = num_pop_total,
         medianinc = num_income_median,
         blackshare = per_race_black_alone_or_combo) %>%
  mutate(black75 = as.numeric(blackshare >= 75), #create a dummy variable
         inc_above_median = as.numeric(medianinc > 26884.59)) %>%
  arrange(tractid, year)

input_si <- read.dta13("si_1017_cleaned.dta")

si.clean <- input_si %>%
  select(si_order_number, census_tract_long, year, month) %>%
  rename(tractid = census_tract_long) %>%
  arrange(tractid, year, month)

si_tract_ym <- si.clean %>%
  group_by(tractid, year, month) %>%
  summarise(si_count = n_distinct(si_order_number)) %>%
  arrange(tractid, year, month)

tract_ym <- left_join(si_tract_ym, acs_tract.clean,
```

```
                      by = c("tractid", "year")) %>%
      mutate(date = make_date(year, month, 1)) %>%
      arrange(tractid, year, month) %>%
      filter(date != "2017-11-01")

tract <- tract_ym %>%
  group_by(tractid) %>%
  summarise(si_count = sum(si_count),
            pop = mean(pop, na.rm = TRUE),
            blackshare = mean(blackshare, na.rm = TRUE),
            black75 = round(mean(black75, na.rm = TRUE), 0),
            medianinc = mean(medianinc, na.rm = TRUE),
            inc_above_median = round(mean(inc_above_median, na.rm = TRUE), 0) ) %>%
  mutate(si_1000 = si_count / (pop / 1000) ) %>%
  arrange(tractid)
```

```
ggplot(data = tract, aes(x = blackshare,
                         y = si_1000,
                         weight = pop,
                         size = pop)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = 'lm', formula = y ~ x)  +
  scale_size(range = c(0.1, 6), guide = "none") +
  ggtitle("Shutoffs per capita on Black share of population")
```



The above scatterplot plots the relationship between shutoffs per capita and the share of residents that are

Black across Census tracts (census tracts are weighted by population, with larger markers for more populous tracts). While the majority of tracts are predominantly Black, the concentration of tracts that are more than 75% Black have markedly higher shutoffs per capita. Also note that all Census tract statistics based on the American Community Survey (for share Black, median household income, and population) are calculated as averages from 2010 through 2017.

```
cor(tract$blackshare, tract$si_1000, use = "pairwise.complete.obs")
```

```
## [1] 0.4679621
```

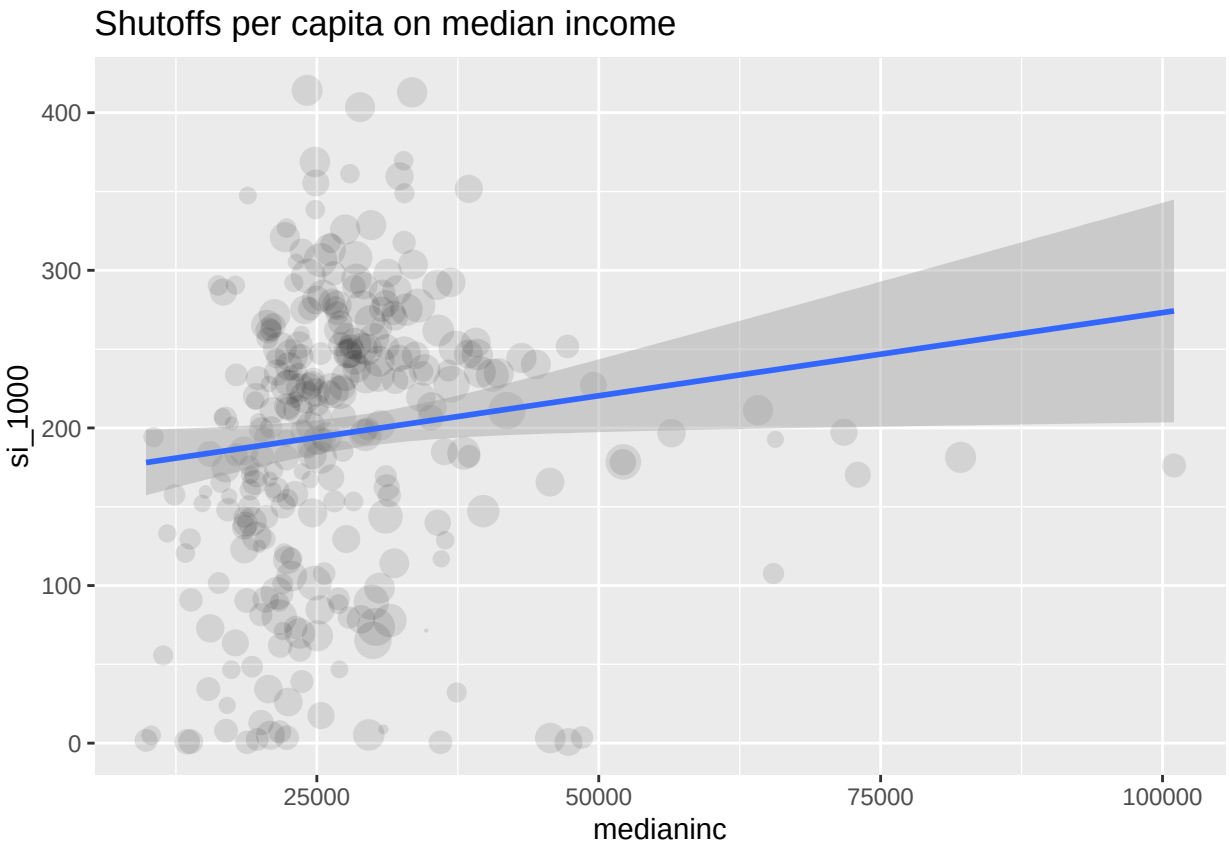
```
wtd.cors(tract$blackshare, tract$si_1000, weight = tract$pop)
```

```
##           [,1]
## [1,] 0.5147538
```

Here, I computed correlations between share Black and shutoffs per capita, unweighted and weighted by the number of population, to see if what I interpreted above is correct. Both unweighted and weighted correlation shows a positive association, with the weighted correlation demonstrating a stronger association.

2 Relationship between median income and shutoffs per capita across census tracts in Detroit

```
ggplot(data = tract, aes(x = medianinc,
                         y = si_1000,
                         weight = pop,
                         size = pop)) +
     geom_point(alpha = 0.1) + #alpha adjusts the transparency of points
     geom_smooth(method = 'lm', formula = y ~ x)  +
     scale_size(range = c(0.1, 6), guide = "none") +
  ggtitle("Shutoffs per capita on median income")
```



The above scatterplot plots the relationship between shutoffs per capita and median household income across Census tracts (census tracts are again weighted by population). The overwhelming majority of tracts have a median income below 50,000, and the relationship between shutoffs per capita and median income is positive.

```
cor(tract$medianinc, tract$si_1000, use = "pairwise.complete.obs")
```

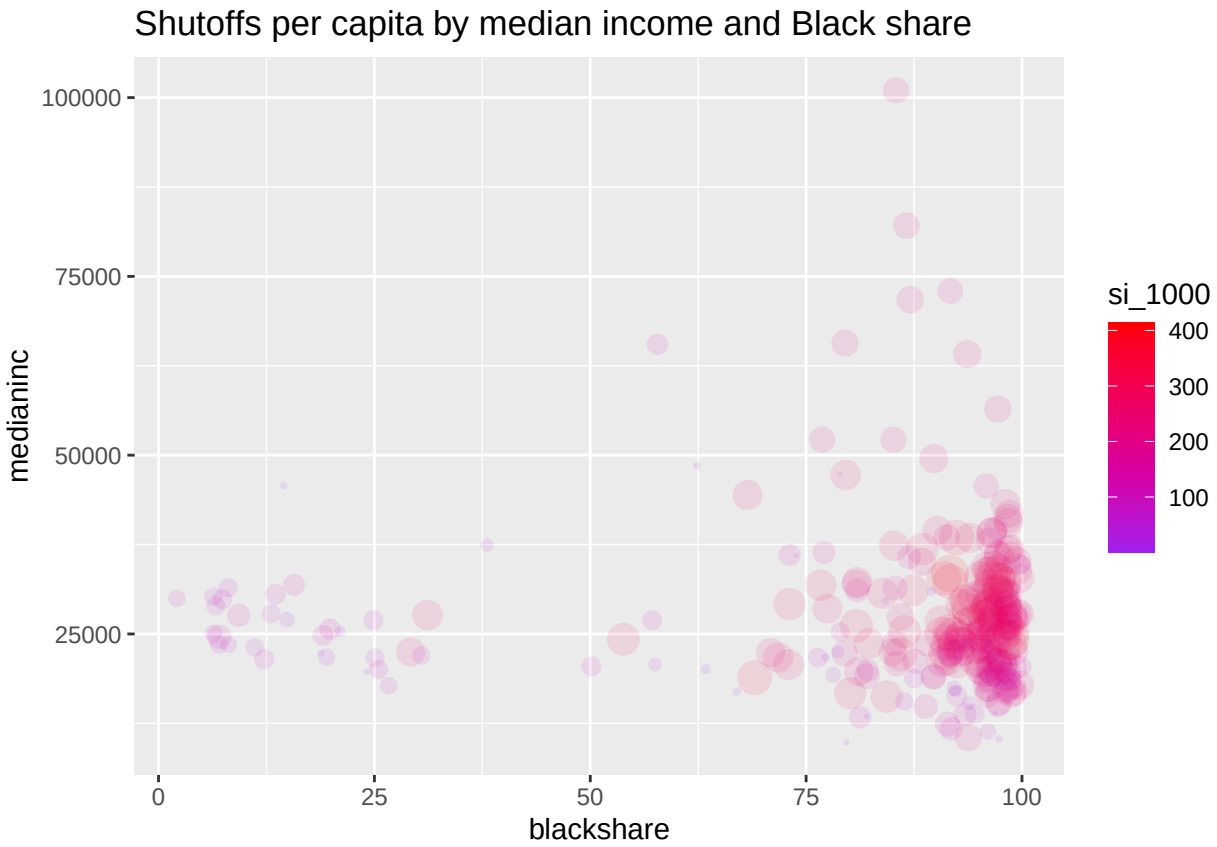
```
## [1] 0.1388281
```

```
wtd.cors(tract$medianinc, tract$si_1000, weight = tract$pop)
```

```
##           [,1]
## [1,] 0.1253117
```

Again, I computed unweighted and weighted correlation between the two variables to confirm my above interpretation. Both unweighted and weighted correlation is positive, although the number is much closer to zero, meaning that the association is quite weak.

```
ggplot(data = tract,
       aes(x = blackshare, y = medianinc, size = si_1000, color = si_1000)) +
  geom_point(alpha = 0.1) +
  scale_size(range = c(0.1, 6), guide = "none") +
  scale_color_gradient(low="purple", high="red") +
  ggtitle("Shutoffs per capita by median income and Black share")
```



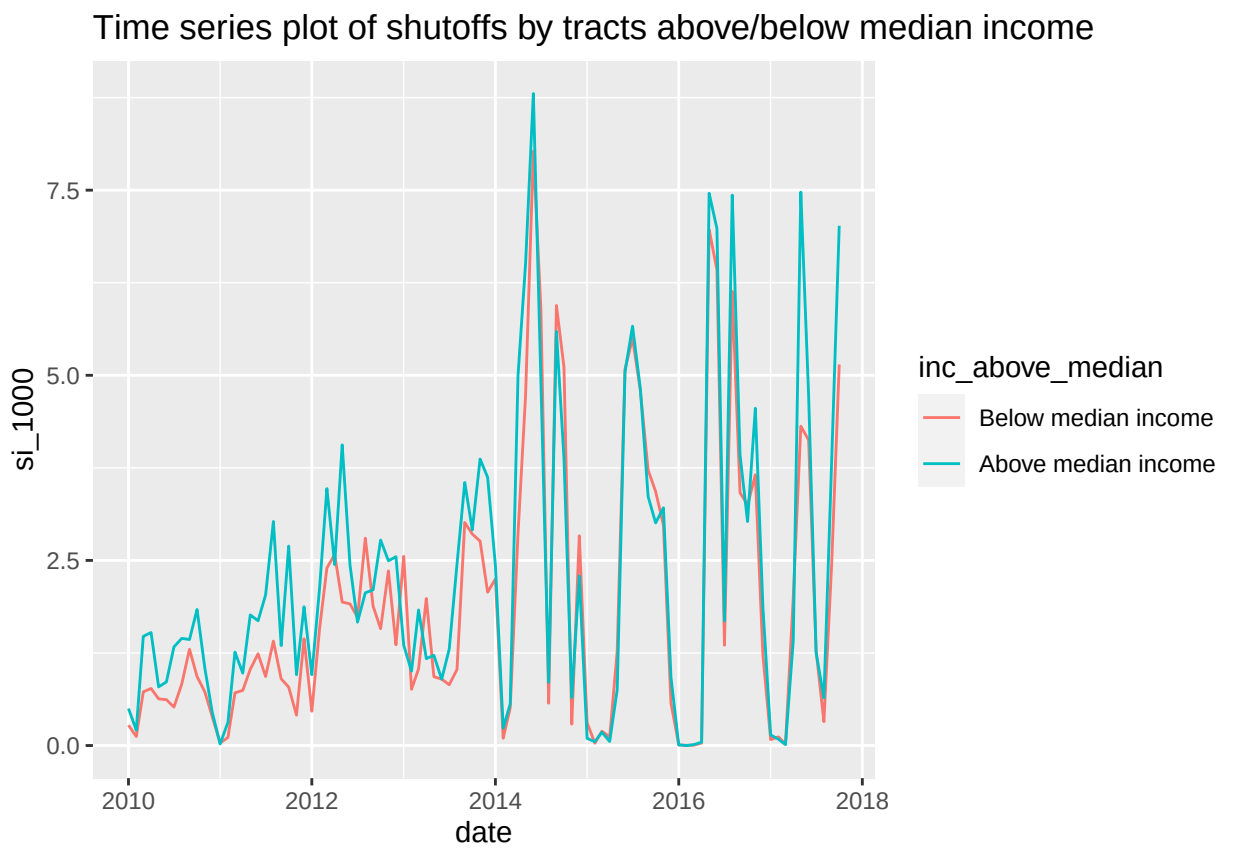
The above scatterplot shows how shutoffs per capita vary along with both share Black and median household income: tracts with more shutoffs per capita appear as larger, redder circles. While most tracts in Detroit are relatively low income and over 75% Black, it's clear that the tracts with the highest shutoffs per capita tend to be those with the highest share Black, regardless of median income.

3 Time-series analysis

```
detroit_pop <- sum(tract$pop)

ym <- tract_ym %>%
  group_by(date) %>%
  summarise(si_count = sum(si_count)) %>%
  mutate(si_1000 = si_count / (detroit_pop / 1000))

detroit_pop_hi_inc <- tract %>%
    filter(inc_above_median == 1) %>%
    summarise(sum(pop)) %>%
    as.numeric()

detroit_pop_lo_inc <- tract %>%
  filter(inc_above_median == 0) %>%
  summarise(sum(pop)) %>%
```

```
  as.numeric()

ym_inc <- tract_ym %>%
  group_by(date, inc_above_median) %>%
  summarise(si_count = sum(si_count)) %>%
  na.omit() %>%
  ungroup() %>%
  complete(date,
           inc_above_median,
           fill = list(si_count = 0)) %>%
  mutate(pop = if_else(inc_above_median == 1,
                       detroit_pop_hi_inc,
                       detroit_pop_lo_inc),
         si_1000 = si_count / (pop / 1000))

ym_inc$inc_above_median <- factor(ym_inc$inc_above_median,
                                  levels = c(0,1),
                                  labels = c("Below median income", "Above median income"))
```

```
ggplot(ym_inc,
       aes(x = date, y = si_1000, color = inc_above_median)) +
  geom_line() +
  ggtitle("Time series plot of shutoffs by tracts above/below median income")
```



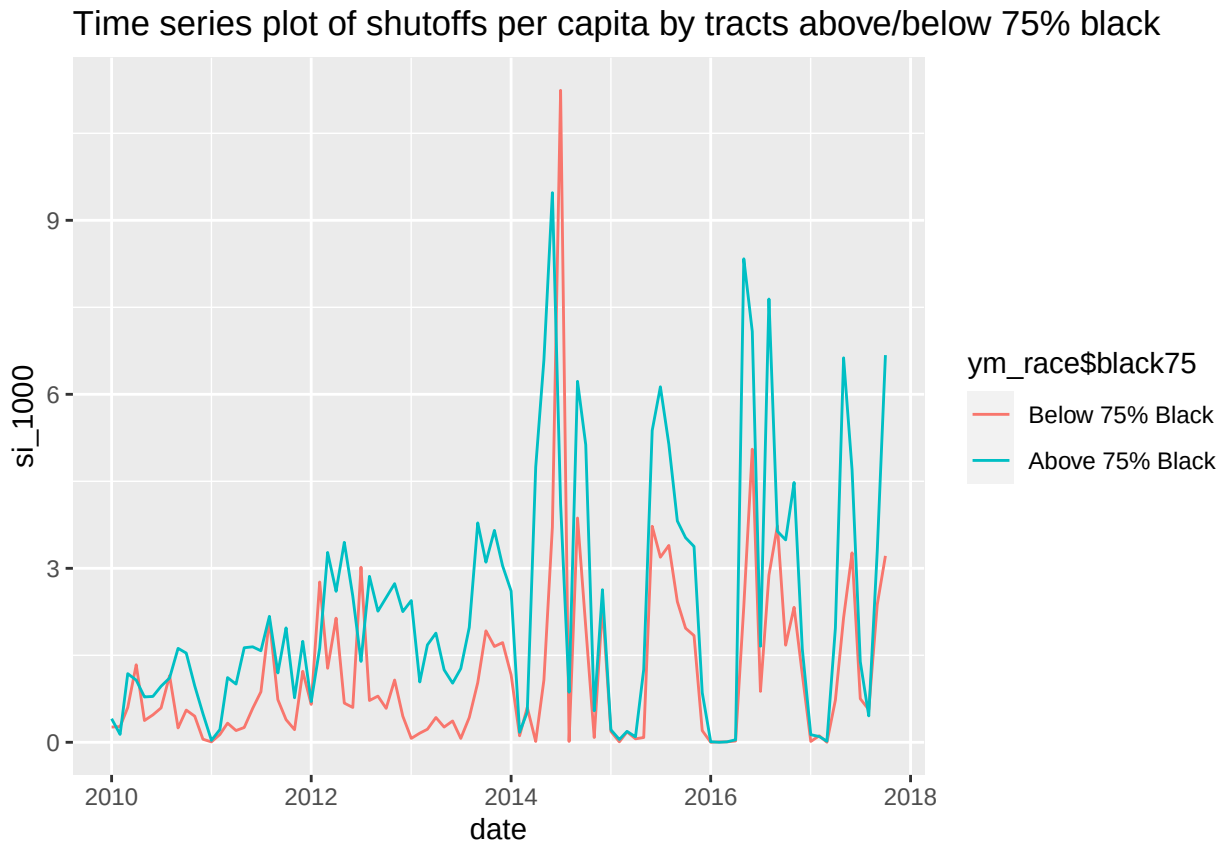
The variation between water shutoff rates per capita over time for tracts above median income and below median income is larger before 2014 and after 2017. This is possibly due to the default by Detroit government

in 2014 that led to restructuring which may have de-prioritized funds for water supply and thus shutoffs occurred more frequently for both groups regardless of income. Prior to the default, the shutoff rates for census tracts that are above median income is generally higher, except for the period between mid 2012 to mid 2013, where the shutoff rates for below median income is higher. These variations further imply that the association between median income and water shutoff rates is quite weak, since there is no consistency in the direction of variation – there are time periods where either income group experienced higher number of water shutoffs.

4 Shutoffs per capita over time for tracts that are at least 75% Black and those that aren't

```
detroit_pop_black <- tract %>%
  filter(black75 == 1) %>%
  summarise(sum(pop)) %>%
  as.numeric()

detroit_pop_nblack <- tract %>%
  filter(black75 == 0) %>%
  summarise(sum(pop)) %>%
  as.numeric()

ym_race <- tract_ym %>%
  group_by(date, black75) %>%
  summarise(si_count = sum(si_count)) %>%
  na.omit() %>%
  ungroup() %>%
  complete(date,
           black75,
           fill = list(si_count = 0)) %>%
  mutate(pop = if_else(black75 == 1,
                       detroit_pop_black,
                       detroit_pop_nblack),
         si_1000 = si_count / (pop / 1000))

ym_race$black75 <- factor(ym_race$black75,
                          levels = c(0,1),
                          labels = c("Below 75% Black", "Above 75% Black"))
```

```
ggplot(ym_race,
       aes(x = date, y = si_1000, color = ym_race$black75)) +
  geom_line() +
  ggtitle("Time series plot of shutoffs per capita by tracts above/below 75% black")
```



When comparing shutoffs per capita over time for Census tracts that are at least 75% Black to other tracts, pronounced differences are visible: monthly shutoffs per capita are noticeably higher in predominantly Black Census tracts compared to other tracts in nearly every month.

5 Conclusion

Based on both cross-sectional and time series analysis, race appears to be a more important factor for explaining the type of households most affected by public water shutoffs. The cross-sectional analysis showed that the association between race and water shutoffs per capita is stronger, with the high positive correlation between the two variables and the graph that was moderated to look more like the association between share black and water shutoffs per capita when plotted together with median income. Further, through time series analysis, we learned that the variation between census tracts with at least of greater than 75% black residents consistently experienced higher level of water shutoffs per capita, unlike the variation between median income and water shutoffs per capita which told an inconsistent story.