U6614: Assignment X

Philip Crane (plc2137)

2024-03-04

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on Monday, March 4th.

Remember to think carefully about what code you include in your knitted document. Only include code chunks that you need to generate the plots and statistics to answer the questions below. Don't include code from your working R script (that we started in class) that was only used to inspect and validate your results, and isn't necessary to answer the questions.

Load libraries

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

getwd()
```

```
## [1] "C:/Users/philc/OneDrive/Desktop/Spring 2024/R/Lectures/Lecture7"
```

```
MI_acs_tract_10_17 <- readRDS("Data/MI_acs_tract_10_17.rds")

input_si <- read.dta13("Data/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, MI_acs_tract_10_17,
                      by = c("tractid", "year")) %>%
  mutate(date = make_date(year, month, 1)) %>%
  arrange(tractid, year, month) %>%
  filter(date != "2017-11-01")
```

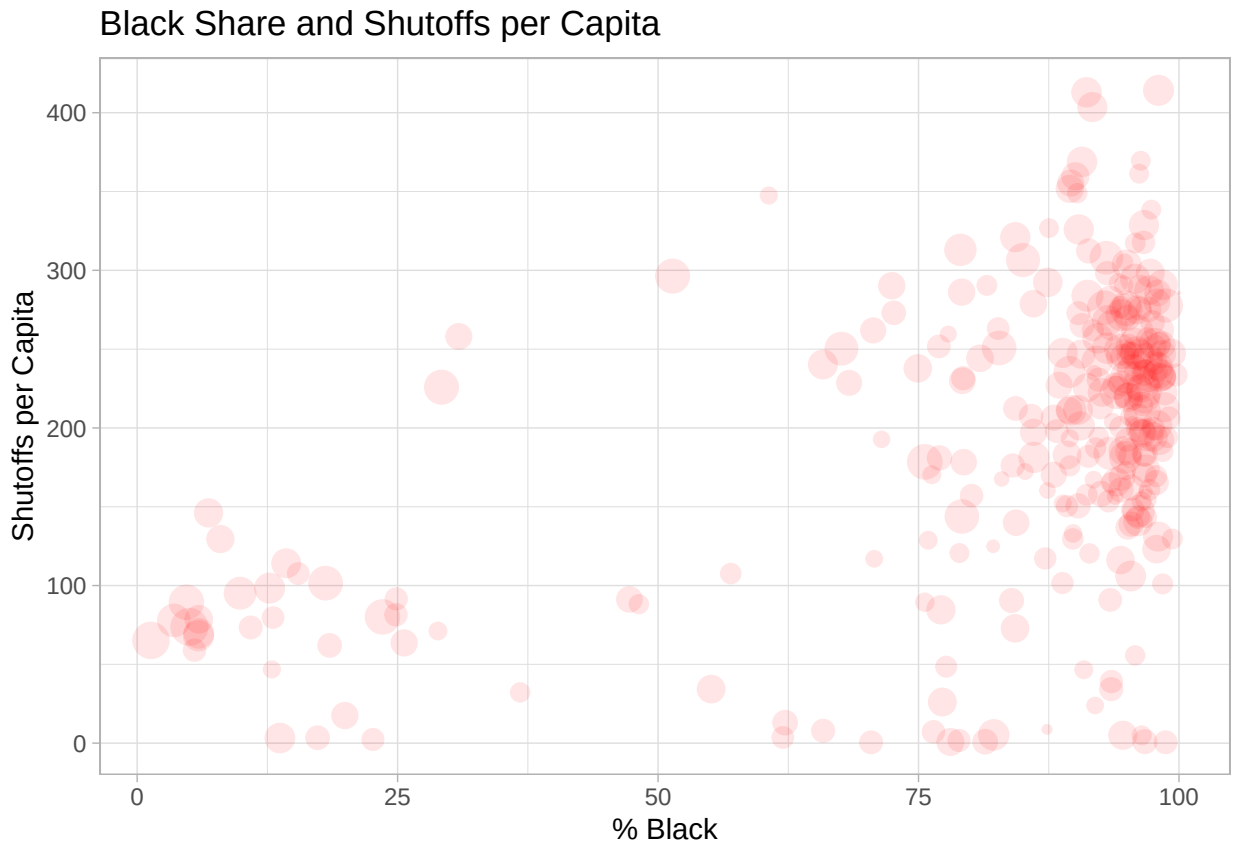
```r
tract <- tract_ym %>%
  group_by(tractid) %>%
  summarise(si_count = sum(si_count),
            pop = mean(pop, na.rm = TRUE),
            blackshare = mean(blackshare, na.rm = TRUE),
            black75 = round(mean(black75, na.rm = TRUE), 0),
            medianinc = mean(medianinc, na.rm = TRUE),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

1 "Cross-sectional" analysis

In this section we'll explore variation in shutoffs *across* Census tracts (one observation per Census tract, summing shutoffs over the whole time period).

1.1 Visualize and interpret the relationship between share Black and shutoffs per capita across census tracts in Detroit.

```r
ggplot(data = tract,
       aes(x = blackshare,
           y = si_1000,
           size = pop)) +
  geom_point(alpha = 0.1, color = "red") +
  scale_size(range = c(0.1, 6), guide = "none") +
  labs(title = "Black Share and Shutoffs per Capita", x = "% Black",
       y = "Shutoffs per Capita") +
  theme_light()
```



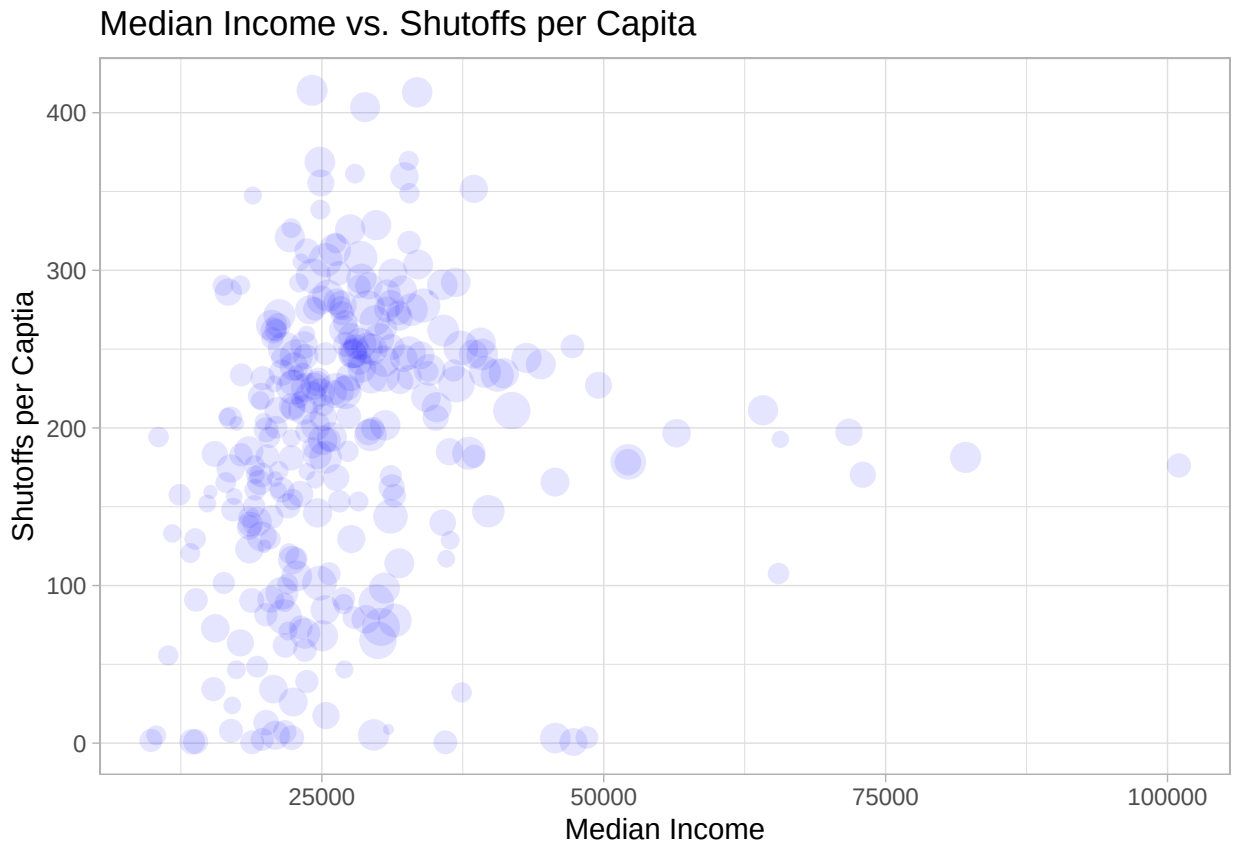

```
wtd.cor(tract$blackshare, tract$si_1000, weight = tract$pop)
```

```
##   correlation    std.err  t.value      p.value
## Y   0.5144811 0.05035423 10.21724 3.918378e-21
```

The visualization is explicitly showing a high positive relationship between areas with high black populations and water shutoffs per capita. It seems that the highest concentration of observations lies in the 90-95% black range associated with 150-300 shutoffs per capita. The weighted correlation shows a reasonably strong correlation of 0.5144 and a statistically significant p-value, further supporting

1.2 Visualize and interpret the relationship between median income and shutoffs per capita across census tracts in Detroit.

```
ggplot(data = tract,
       aes(x = medianinc,
           y = si_1000,
           size = pop)) +
  geom_point(alpha = 0.1, color = "blue") +
  scale_size(range = c(0.1, 6), guide = "none") +
  labs(title = "Median Income vs. Shutoffs per Capita", x = "Median Income",
       y = "Shutoffs per Captia") +
  theme_light()
```




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

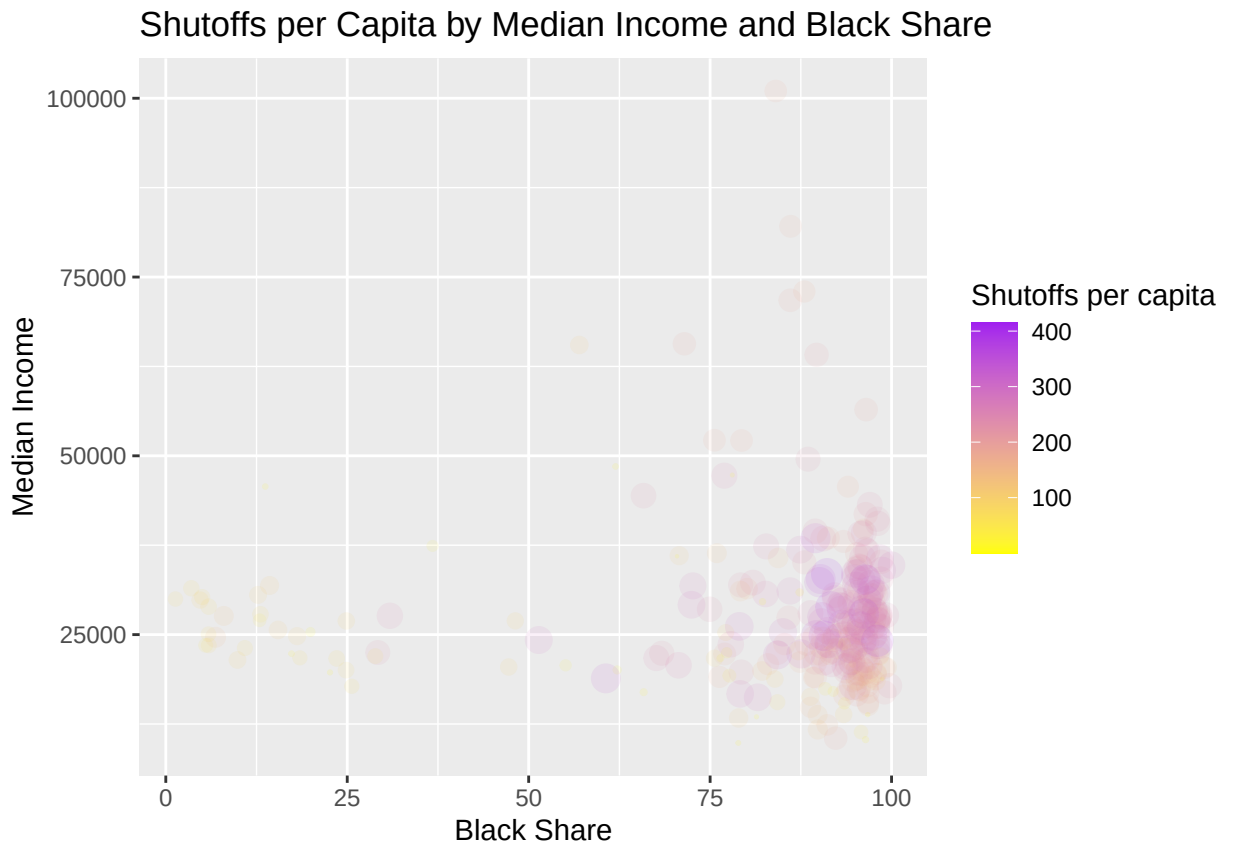
```
##   correlation    std.err  t.value    p.value
## Y   0.1254187 0.05825835 2.152802 0.03215895
```

This plot shows a wider spread with a sizable cluster of 250 shutoffs per capita for those just above the $25,000 income level. Interestingly, households earining above $25,000 generally don't experience more than 200-210 shutoffs per capita. The weighted correlation supports the weaker relationship between the variables with a value of only 0.125. The p-value is still statistically significant at <0.05, indicating statistically significant weak positive correlation between the two variables.

1.3 Visualize and interpret how shutoffs per capita relate to both Black share and median income on the *same* plot. Does race or income appear to be more salient?

```
ggplot(data = tract,
       aes(x = blackshare,
           y = medianinc,
           size = si_1000,
           color = si_1000)) +
  geom_point(alpha = 0.1) +
  scale_color_gradient(low = "yellow", high = "purple") +
  scale_size(range = c(0.1, 5), guide = "none") +
```

```
  labs(title = "Shutoffs per Capita by Median Income and Black Share", x = "Black Share", y = "Median In
       color = "Shutoffs per capita")
```




```
wtd.cor(tract$blackshare, tract$medianinc, weight = tract$pop)
```

```
##   correlation   std.err   t.value   p.value
## Y  0.02901325 0.0586973 0.4942859 0.6214785
```

Like the previous graph, the highest frequency of water shutoffs occur around the $25,000 income level, clustered around tracts of 95% Black residents. Race seems the most salient as the weighted correlation reflects a very weak positive relationship between "blackshare" and "medianinc". The high p-value confirms that this relationship is not statistically significant.

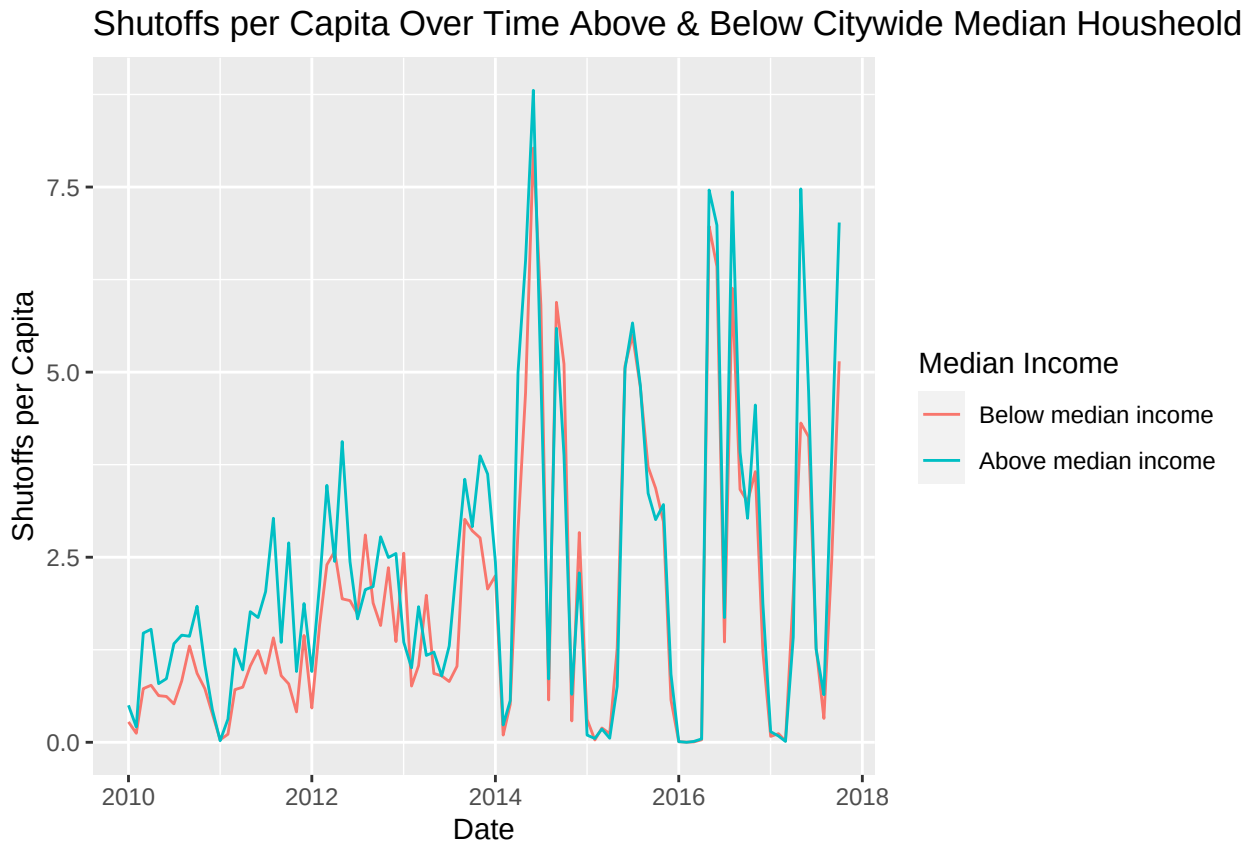
2 Time-series analysis.

In this section, we'll explore variation *between* different groups of Census tracts and over time *within* groups (with groups defined based on tract-level income and racial composition).

2.1 Plot and interpret the shutoffs per capita over time for tracts below/above citywide median housheold income (show two time series on a single plot).

```
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)) %>%
  mutate(pop = if_else(inc_above_median == 1,
                       detroit_pop_hi_inc,
                       detroit_pop_lo_inc),
         si_1000 = si_count / (pop / 1000)) %>%
  na.omit()

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() +
  labs(title = "Shutoffs per Capita Over Time Above & Below Citywide Median Housheold Income",
       x = "Date", y = "Shutoffs per Capita", color = "Median Income")
```



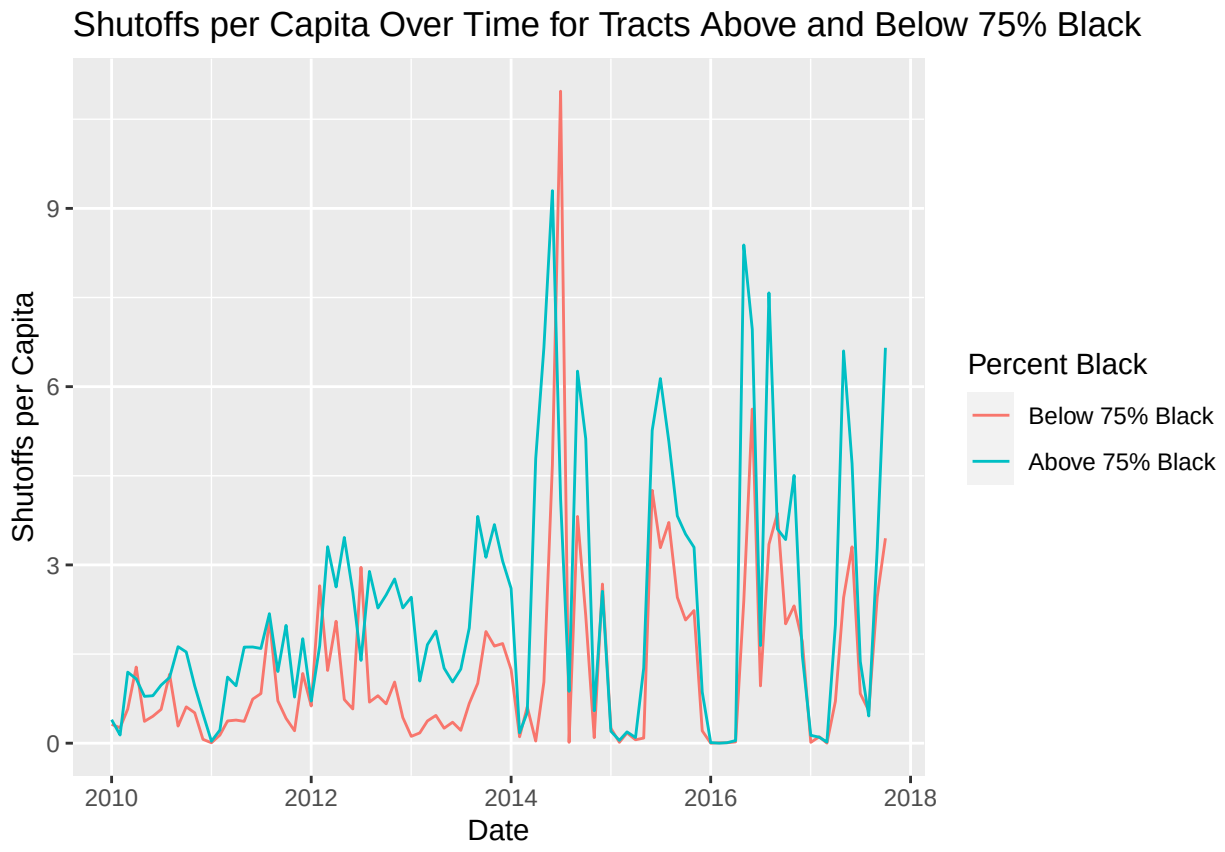
This time-series chart looks exclusively at median income versus shutoffs per capita. When measuring medium income independent of other factors, there appears to be many instances where households above the median income experience more water shutoffs than those below median income. Visually, there seems to be only three instances where below median income households experience more shutoffs (mid 2012, mid 2014, and late 2014). From 2015 to 2017, both income levels mostly align.

2.2 Plot and interpret the shutoffs per capita over time for tracts that are at least 75% Black and those that aren't (show two time series on a single plot).

```
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 = black75)) +
  geom_line() +
  labs(title = "Shutoffs per Capita Over Time for Tracts Above and Below 75% Black",
       x = "Date", y = "Shutoffs per Capita", color = "Percent Black")
```



This graph illustrates a stark contrast between tracts consisting of more than 75% Black residents and those which do not. Almost every spike in water shutoffs affect tracts with more than 75% Black residents by a fair margin. It is also interesting to note that the width of these trends are wider for these tracts, suggesting that those water shutoffs last longer in duration.

3 Conclusion

3.1 Based on the "cross-sectional" and time series analysis conducted above, does race or income appear to be a more important factor for explaining what type of households are most affected by the public water shutoffs? Explain.

Race seems to have far stronger explanatory power when looking at those affected by water shutoffs. In the cross-sectional analysis, the correlation value is far higher for race than income, with a p-value far closer to zero. The correlation loses all statistical significance when income and race are measured together, suggesting income to have a confounding effect on the relationship.

The time-series analysis supports this. While some outliers appear to show high-income households suffering from more water shutoffs (particularly pre-2012), the graph analyzing race shows far more convincing margins between tracts. The time-series graphs alone are not enough to definitively judge whether race or income is more important. However, when contextualized by the cross-sectional visualizations, race clearly displays a stronger relationship with water shutoffs.