Relationship between median income and shooting incidents in NYC ZIP codes

Ray Botha (University of Colorado Boulder, raymond.botha@colorado.edu)

8/14/2021

Project Datasets

This is an analysis of shooting incidents that occurred in New York City going back to 2006. Let's explore the relationship between the frequency of shooting incidents and median income in different NYC ZIP codes.

For more details on datasets see: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic https://data.census.gov/cedsci/table?q=median%20income&tid=ACSST5Y2019.S1901

First, we'll import the data.

```
library(tidyverse)
library(ggmap)
library(tigris)

# NYPD shootings data
shootings_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
shooting_rows <- read_csv(shootings_url) %>%
  rename(lat = Latitude,
         lon = Longitude) %>%
  select(lat, lon)

# Income and population data by ZIP Code Tabulation Area, from ACS 5-year (2019)
income_file <- "./data/2019-census-incomes.csv"
income_rows <- read_csv(income_file, skip = 1) %>%
  rename(zip_code="Geographic Area Name",
         median_income="Estimate!!Households!!Median income (dollars)") %>%
  select(zip_code, median_income) %>%
  mutate(median_income = na_if(median_income, "-")) %>%
  drop_na() %>%
  mutate(zip_code = str_sub(zip_code, 7))
population_file <- "./data/2019-census-population.csv"
population_rows <- read_csv(population_file, skip=1) %>%
  rename(zip_code="Geographic Area Name",
         population="Estimate!!Total") %>%
  select(zip_code, population) %>%
  mutate(population = na_if(population, 0)) %>%
  drop_na() %>%
  mutate(zip_code = str_sub(zip_code, 7))

# NYC map from Stamen
basemap <- get_stamenmap(bbox = c(left = -74.3,
                                  right = -73.6,
```

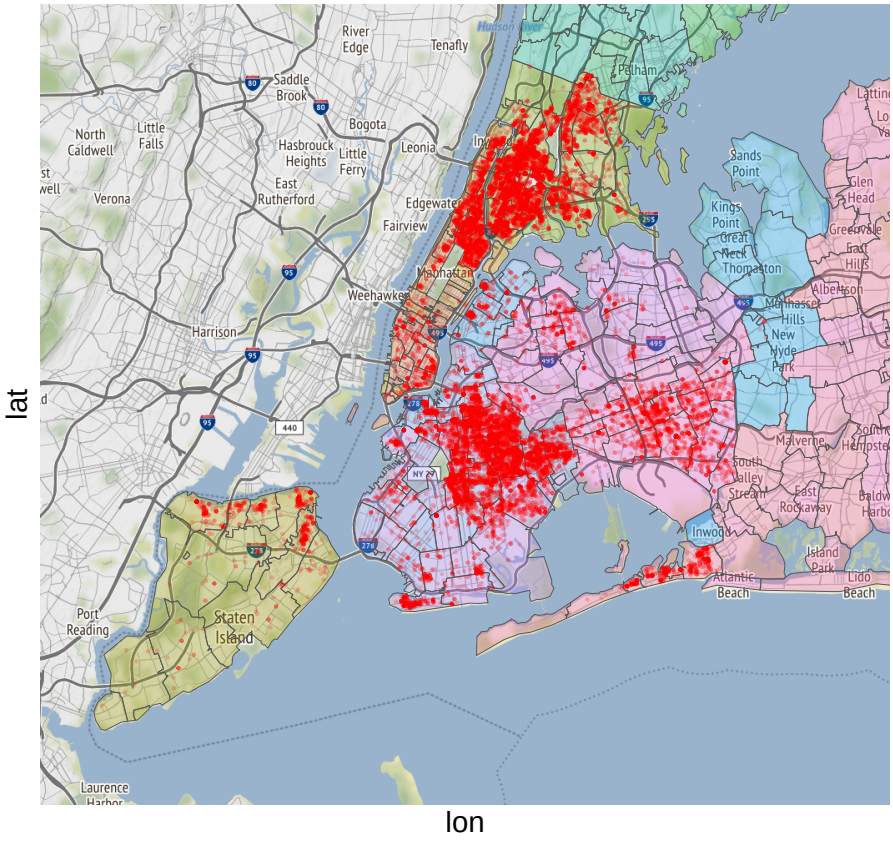
```
                                  bottom = 40.45,
                                  top = 40.95),
                          zoom = 11)

# ZIP Code Tabulation Area coordinate polygons
zips_sf <- zctas(cb = TRUE, starts_with = "1", class = "sf") %>%
  select(zip = ZCTA5CE10, geometry) %>%
  filter(zip > 10000 & zip < 11700)
```

Now, let's plot the ZIP codes of NYC, to make sure we know the bounding latitude/longitude polygons of relevant ZIP codes. Here we can add the shooting incidents too.

```
ggmap(basemap, extent = "panel") +
  geom_sf(aes(fill = zip),
          data = zips_sf,
          inherit.aes = F,
          size = 0,
          alpha = 0.3) +
  geom_point(data = shooting_rows, col="red", size=.1, alpha=.2) +
  coord_sf(ndiscr = F) +
  theme(legend.position = "none")
```

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one
```



Now we find the intersections of shooting incident coordinates with the ZIP code polygons.

```
library(sf)
```

```
## Linking to GEOS 3.8.1, GDAL 3.2.1, PROJ 7.2.1
```

```
# Convert data to SF objects
shootings_sf = st_as_sf(shooting_rows, coords=c("lon", "lat"), crs=4326)

# Reproject spatial data to planar coordinate system
shootings_planar = st_transform(shootings_sf, crs=3857)
zips_planar = st_transform(zips_sf, crs=3857)

# Find planar intersects of incidents with ZIP polygons
zip_codes = zips_planar[["zip"]]
intersect_indices = as.integer(st_intersects(shootings_planar, zips_planar))
shootings_w_zip <- shooting_rows %>% add_column(zip_code = zip_codes[intersect_indices])
```

Now that our incidents have associated ZIP codes, we can combine our data for income, incidents, and population.

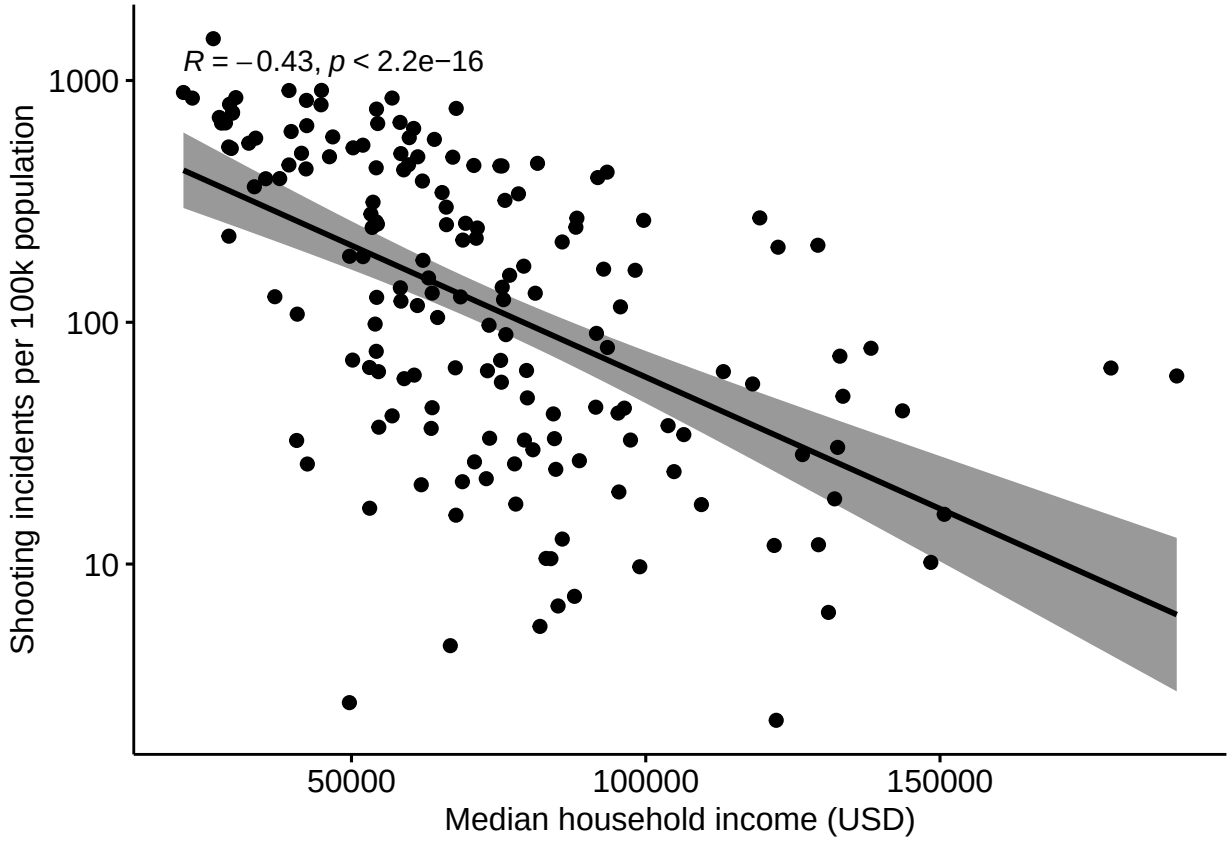
```
zips_summary <- shootings_w_zip %>%
  group_by(zip_code) %>%
  summarize(incidents=n()) %>%
  inner_join(population_rows) %>%
  mutate(incidents_per_100k = (incidents / population) * 100000) %>%
  inner_join(income_rows) %>%
  mutate(across(median_income, as.double))
```

```
## Joining, by = "zip_code"
## Joining, by = "zip_code"
```

Finally, we can look at the correlation between shooting incidents per 100k population and median household income, for each ZIP Code Tabulation Area.

```
library(ggpubr)
ggscatter(zips_summary,
          x = "median_income",
          y = "incidents_per_100k",
          add = "reg.line",
          conf.int = TRUE,
          cor.coef = TRUE,
          cor.method = "kendall",
          xlab = "Median household income (USD)",
          ylab = "Shooting incidents per 100k population") + yscale("log10")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Conclusion

There seems to be a clear correlation between median income in ZIP codes, and the frequency of shooting incidents. You may be inclined to extrapolate that the socioeconomic environment of low-income households are more likely to be involved in violent crime, because of factors such as access to education, healthcare, and state resources, presence of organized crime, and lack of opportunity. However, to draw conclusions from the data requires further facts and research to avoid bias.