Project 4: Hate Crimes

TOTAL: 20pts

PURPOSE

To understand patterns of hate crimes across the U.S. and use those patterns to understand similarities and differences to Washington, D.C.

GOAL

Analyze hate crime across the U.S. to identify areas (cities/counties/metropolitan statistical areas) similar to Washington, D.C.

DATA

The Federal Bureau of Investigation (F.B.I.) is one of multiple Department of Justice resources that collects hate crime data. The F.B.I. makes a variety of their collected and aggregated national data available via their website, which includes Hate Crime events. Regular users can get 1 year of data at a time, but larger downloads are available here. That’s what we will use - because it’s Project 4 and you’re not a ‘regular user.’

METHODS

Before we begin, it is worth mentioning that the F.B.I. cautions against doing what we are about to do.

Step 0: Load and install packages

1. You will need to install these packages (run both of these if using the AU computers):
    a. install.packages('tidyverse')
    b. install.packages('explore')
2. You will need to load these libraries, no matter the computer:
    a. library(tidyverse)
    b. library(explore)

Step 1: Get data

1. Our source data from the FBI is here. Specifically from the 'Documents & Downloads' section, and then 'Hate Crime'. It's all the reported hate crimes for 1991-2023.
    a. Download and then connect to the data, like this (depending on your computer):
        i. setwd("C:/Users/USERNAME HERE/Downloads/xxx") #Windows
        ii. setwd("/Users/USERNAME HERE/Downloads/xxx") #Mac
    b. And then read in the data:
        i. hate <- read.csv("hate_crime.csv", stringsAsFactors = FALSE)

Step 2: Inspect data

1. Look at the column names:
    a. colnames(hate)
2. Look at the unique values for the agency type field
    a. unique(hate$agency_type_name)
3. Look at the unique values for the area population groups
    a. unique(hate$population_group_desc)
4. Others?

Step 3: Subset data

1. Make a subset of the data for just D.C.:
    a. hate.dc <- subset(hate, hate$state_abbr == 'DC')
2. And now analyze it, relative to all data:
    a. explore(hate)
    b. explore(hate.dc)
        i. Focus on year; victim counts and type; offender counts, race, and ethnicity; offense name; location name; bias description
            1. ***Identify the variables that stand out for D.C.***

Step 4: Filter data

1. Filter the main table by the geography that makes the most sense to you:
    a. *hate.city <- hate %>% filter(agency_type_name == 'xxx') # replace xxx with an Agency Type*
2. Even better: filter the main table by the right geography *and* a population group
    a. hate.similar <- hate %>% filter(population_group_code == 'yyy') # replace xx with an Agency Type, replace yyy with a population grouping
    b. Based on this, generate a list of potential areas to consider:
        i. unique(hate.similar$pug_agency_name)

Step 5: Summarize data

1. Create a new table summarizing events per area:
    a. hate.events.sum <- hate.similar %>%
    b. group_by(pug_agency_name) %>%
    c. summarise(count = n()) %>%
    d. mutate(pct = count/sum(count)*100)
        i. This code generate a summary table using Step 4, grouped by name, count number of events, and adding a percentage - for percentage of all events
2. Add a 'diff' column to calculate the difference in number of events for each area relative to D.C.
    a. sum.events$diff <- sum.events$count - sum.events$count[zz] # replace zz with the row number for D.C.
3. Now convert that difference to an absolute value
    a. sum.events$abs.diff <- abs(sum.events$diff)
4. Create a new table that reorders your summary table in ascending order of absolute difference
    a. events <- sum.events[order(sum.events$abs.diff),]
5. Add a new field for ranking each area, relative to D.C.
    a. events$rank <- seq.int(nrow(events)) - 1
        i. "-1" is so D.C. becomes rank 0 instead of 1
6. **Do some analysis of the 'events' table to compare D.C. to other cities**

HERE IS THE SCRIPT WE BUILT ON CLASS ON 19 MARCH 2025

Step 6: More summarizing data

Apply a similar process to the race field…

1. Create a new table summarizing events per area per race:
    a. sum.offender.race <- hate.city.similar %>%
    b. group_by(pug_agency_name, offender_race) %>%
    c. summarise(count = n()) %>%
    d. mutate(pct = count/sum(count)*100)
2. Create a subset table for just 'Black/African American' offenders
    a. offender.black <- subset(sum.offender.race, sum.offender.race$offender_race == 'Black or African American')
3. Calculate difference and absolute difference
    a. offender.black$diff <- offender.black$pct - offender.black$pct[ww] # replace ww with the row number for D.C.
    b. offender.black$abs.diff <- abs(offender.black$diff)
4. Reorder and rank the table
    a. offender.black <- offender.black[order(offender.black$abs.diff),]
    b. offender.black$rank <- seq.int(nrow(offender.black)) -1
5. Repeat the process for 'White' offenders - subset, calculate differences, reorder, and rank
    a. offender.white <- subset(sum.offender.race, sum.offender.race$offender_race == 'White')
    b. offender.white$diff <- offender.white$pct - offender.white$pct[vv] # replace vv with the row number for D.C.
    c. offender.white$abs.diff <- abs(offender.white$diff)

   d. offender.white <- offender.white[order(offender.white$abs.diff),]
   e. offender.white$rank <- seq.int(nrow(offender.white)) -1

6. Repeat #5 for the other races… if you feel it’s necessary
7. When complete, merge the separate race tables back together:
   a. offender.race <- rbind(offender.black, offender.white, …)
8. **Do some analysis using the ‘offender.race’ table to compare D.C. to other cities**

Step 7: Even more summarizing data

- Repeat the process from Step 6 for other fields of analytic value, including: *offense*, *location*, and *bias*.
- You don’t have to calculate every value for every field, focus on the big numbers relative to D.C. (similar to Black and White for Race)
- Example:
    - sum.location <- hate.city.similar %>%
    -   group_by(pug_agency_name, location_name) %>%
    -   summarise(count = n()) %>%
    -   mutate(pct = count/sum(count)*100)
    -
    - loc.highway <- subset(sum.location, sum.location$location_name == 'Highway/Road/Alley/Street/Sidewalk')
    - loc.highway$diff <- loc.highway$pct - loc.highway$pct[33]
    - loc.highway$abs.diff <- abs(loc.highway$diff)
    - loc.highway <- loc.highway[order(loc.highway$abs.diff),]
    - loc.highway$rank <- seq.int(nrow(loc.highway)) -1
    -
    - loc.home <- subset(sum.location, sum.location$location_name == 'Residence/Home')
    - loc.home$diff <- loc.home$pct - loc.home$pct[33]
    - loc.home$abs.diff <- abs(loc.home$diff)
    - loc.home <- loc.home[order(loc.home$abs.diff),]
    - loc.home$rank <- seq.int(nrow(loc.home)) -1
    -
    - location <- rbind(loc.highway, loc.home)
- **Do some analysis using these tables to compare D.C. to other cities**

Another variable to measure:

- sum.bias <- hate.city.similar %>%
-   group_by(pug_agency_name, bias_desc) %>%
-   summarise(count = n()) %>%
-   mutate(pct = count/sum(count)*100)
-
- bias.ag <- subset(sum.bias, sum.bias$bias_desc == 'Anti-Gay (Male)')
- bias.ag$diff <- bias.ag$pct - bias.ag$pct[32]
- bias.ag$abs.diff <- abs(bias.ag$diff)
- bias.ag <- bias.ag[order(bias.ag$abs.diff),]
- bias.ag$rank <- seq.int(nrow(bias.ag)) -1

-
- bias.ab <- subset(sum.bias, sum.bias$bias_desc == 'Anti-Black or African American')
- bias.ab$diff <- bias.ab$pct - bias.ab$pct[34]
- bias.ab$abs.diff <- abs(bias.ab$diff)
- bias.ab <- bias.ab[order(bias.ab$abs.diff),]
- bias.ab$rank <- seq.int(nrow(bias.ab)) -1
- bias.at <- subset(sum.bias, sum.bias$bias_desc == 'Anti-Transgender')
- bias.at$diff <- bias.at$pct - bias.at$pct[24]
- bias.at$abs.diff <- abs(bias.at$diff)
- bias.at <- bias.at[order(bias.at$abs.diff),]
- bias.at$rank <- seq.int(nrow(bias.at)) -1
-
- bias <- rbind(bias.ab, bias.ag, bias.at)

HERE IS THE SCRIPT WE BUILT IN CLASS ON 26 MARCH 2025

Step 8: Visualize data

- Create bar graphs as necessary. Use Project 1 as a guide.
- Create scatter plots as necessary from the tables generated from steps 1, 5, 6, and 7. Use the instructions from Project 2 and Project 3 as a guide.
- **Do some analysis using the scatter plots to compare D.C. to other cities**

Step 9: Add politics?

- Here is the initial script for working with political data from class on 2/5/2025

Step 10: Develop a workflow/process

It’s highly recommended to use a clear, repeatable process for finding your similar cities. Something, potentially, like this:

1. **Find data**
    a. Use the explore function for the full dataset and the dc dataset to find what is uniquely D.C. about these data
2. **Get data**
    a. You’ve created at least three (hopefully more) summary tables: *events*, *offender race*, and *location*
3. **Finesse data**
    a. Explore the hate.dc, hate.city.similar, events, offender.race, and ***any other tables*** created
        i. Examine the counts, percentages, and comparisons to D.C
    b. Create comparisons and rankings for each table
    c. For each table, notate each cities are the highest ranked
    d. “Highest” is arbitrary, but apply a reasonable approach to identifying how many are the “highest”
        i. Finding a natural break in the data is the recommended best way
    e. Compare the lists of highest ranked cities across each table, and spot common cities

        - i. This process helps identify "under the radar" similarities - for example, a city that never ranks #1, but is always near the top

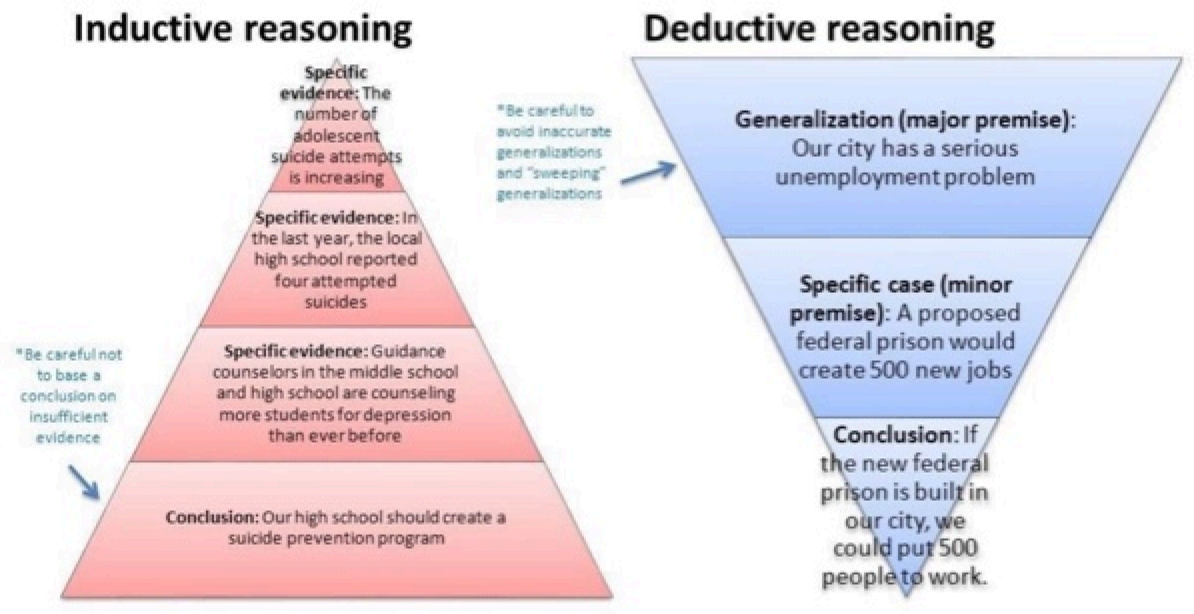
4. **Visualize data**
    - a. Subset your data based on a specific field/attribute
        - i. For example, if you queried population by race, filter your data to only look at a specific race
    - b. Subset again to look at fewer areas
        - i. This subset is based on the values within a field
        - ii. This creates a table with less rows that is likely great for visualizing
    - c. Experiment with scatterplots
        - i. Examine the location (counts), size (percentage), and color (absolute difference) of the dots

ANALYSIS

When conducting your analysis, consider this perspective:

1. State a finding
    - a. This is the new thing that you've identified from the data
2. Provide your evidence
    - a. This is typically a statistical reference
3. Add the context
    - a. This is the integration of the stat(s) and the finding
    - b. Compare your finding to the norm/average, or a similar measure in the data

When analyzing data, there's really two approaches: **Inductive** and **Deductive**


Inductive reasoning
Specific evidence: The number of adolescent suicide attempts is increasing
Specific evidence: In the last year, the local high school reported four attempted suicides
Specific evidence: Guidance counselors in the middle school and high school are counseling more students for depression than ever before
Conclusion: Our high school should create a suicide prevention program
*Be careful not to base a conclusion on insufficient evidence
Deductive reasoning
*Be careful to avoid inaccurate generalizations and "sweeping" generalizations
Generalization (major premise): Our city has a serious unemployment problem
Specific case (minor premise): A proposed federal prison would create 500 new jobs
Conclusion: If the new federal prison is built in our city, we could put 500 people to work.

- **Inductive** is like a cone, where you start your process with a known or suspected value, and work outwards thru the data to identify evidence that supports it

    - ○ Inductive is high-risk, high-reward. You will potentially find your answers either much faster, or much slower
- **Deductive** is like a funnel, where you start your process with everything, and filter thru the data to identify evidence so a specific case(s) naturally emerge
    - ○ Deductive is even-paced. You will find your answers consistently and systematically.

Your goal, for this project:

1. **Five similar "cities"**
    - a. Consider cities, counties, agencies/jurisdictions
        - i. If you use counties or agencies/jurisdictions, be clear to address what major city is within that area
2. **Three variables per similar city**
    - a. Don't use attributes from the same field for one city (i.e. race or sex). Use completely different variables.
    - b. This is potentially as many as nine different variables, unless you use a different attribute from the same dataset for different cities
3. ***Incorporate politics***
    - *a. Think about voting tendencies/preferences as more than a binary (blue/red) comparison*
        - *i. Examine trends over time and percentages of an area*
4. **Analyze locations**
    - a. Make maps
    - b. Characterize any potential similar city as either being in the same hotspot as D.C., a hotspot in general, or not in a hotspot
        - i. If a potential similar city is not located in a police shooting hotspot, it's not similar to D.C.
5. **One visual per city**
    - a. That's five total.

FORMAT

Submit your script as a usable HTML file, via RMarkdown. You create this output as a .Rmd file in RStudio, but only submit the .html file.

Use this file as a template.

SUBMISSION

Once your analysis is complete, please submit your project as either an HTML file (as the output/export/knit of your RMarkdown script), via Canvas.

GRADES

Data: **2pt**

- Provide a brief paragraph (3-4 sentences) on the data used in this project. This includes the specific source(s), the size, and the specific temporal and spatial constraints.

Methods: **2pt**

- Provide a brief paragraph (3-4 sentences) on the research methods leveraged to answer this question. This includes the software, calculations, skills, techniques, and unique workflows used to analyze your data and develop an answer. You do NOT need to describe click-by-click instructions or lines of code describing how you did things; you DO need to describe a logical process that is specific enough that a reader could replicate.

Analysis: **10pts**

- Using the data specific to this project, identify five cities *similar* to Washington, D.C. (2pts each)
    - Provide justification of similarity for each city (4-5 sentences each)
        - Each justification should include at least three specific reasons for similarity
            - Each reason should include a relevant statistical reference

Visuals: **5pts**

- Create one visual per similar city (1pt each)
- Each visual should have a direct connection to the analysis, and be clearly labeled

Formatting: **1pt**

- Error-free writing. No typos, run-on sentences, or sentence fragments. Proper punctuation. Real words.
    - Need help paraphrasing dense content? Try this. And here's a great reference, too.
- Project created in RMarkdown and submitted as an HTML file via Canvas.

Please email me with any questions.