Trends of Collective Action in U.S. Cities (1955-1995)

Oliver Breese

4/20/2020

PURPOSE

The right to assemble in the United States is a major reason for our consistent standing as a liberal nation since our conception. If a large enough portion of the population is unhappy with the way things are going, we have the ability to change it through public display. The ability to disagree with the government in the face of officials is rare in the world, and time and time again pushed the U.S. to affect change in its laws.

This analysis attempts to find factors that correlate with protests in the United States to better understand, and possibly predict future trends in collective action and counterculture.

VALUE

Why is this valuable? The possible ability to predict, or at least understand better, protests and demonstrations could help us learn a great deal about the american public.

First and foremost, what makes people desire change, and what are they willing to do to achieve it? For examples, what does the government do that tends to make people the most likely to take matters into their own hands. Is it things that personally affect them? Is it the most egregious violation of civil rights? Now, in this analysis, these questions specifically won't be addressed by data that much, but hopefully researchers, citizens, or even the government could look at the data here to help answer those questions for themselves.

Aditionally, are there inherent charactaristics in people that make them more likely to demonstrate? Do demographics affect in any measureable way likelihood to participate in collective action? This topic pertains more directly to the analysis to be done in this paper, because demographic data is much more readily availible, and much more applicable per city. What could be gained from knowing this? If we have correlations, even loose ones, we can be given a city, and in looking at its unique collection of people, have some idea of how much it will participate in a protest.

Another possible valueable use of this analyis would be the ability to analyze different topics and causes in demonstrations. Given similar code to that used here, and a dataset investigating locations of kinds of protests, perhaps we could discover factors in demographics that relate to class of movement. For example, I would predict that younger groups protest about the environment vastly more than any other age group. If this could be done, then given a city, its demographics, and some recent political changes regarding a subject, we might be able to predict the cities with the most response.

PREVIOUS RESEARCH

From what I could find, there isn't too much research in this specific area. I found one Boston Universtiy project (https://www.bu.edu/articles/2017/counting-american-protests/ (https://www.bu.edu/articles/2017/counting-american-protests/)) about what they call "Charlottesville-related protest and the national division of people. To allow them to summarize for themselves:"Perkins and Leung have confirmed 351 Charlottesville-related protests involving 64,198 people, and while many have occurred in the northeast, the pair point out that "Charlottesville protests, like protests about healthcare, immigration, and civil rights, appear all throughout the country." For them, the national distribution of the Charlottesville protests—and others—emphasizes that people throughout the country care about the same issues. "We are not as different or divided as is often portrayed in the national

media," says Perkins." The major similarities between my project and theirs were the "dot maps", but instead of citywide, they measured number of people at the gathering. While I was researching other versions of these maps, I saw many like theirs but none on the city scale. I believe that my analysis will add unique value to the research community because it is comparing cities' protest on a much larger scale.

RESEARCH QUESTIONS

When conducting this research, the following were the most guiding questions for me:

- Do demographic or societal factors correlate with and influence frequency and location of demonstrations in the United States?
- What is the relationship between age, gender, income, and race with likelihood to protest.
- What is the relationship between the presence of violence, arrests, and deaths of the protests and the protests' frequency?
- How do the four decades covered differ in amount and type of protest?

DATASETS

- Dynamics of Collective Action:
  (https://web.stanford.edu/group/collectiveaction/cgi-bin/drupal/node/21
  (https://web.stanford.edu/group/collectiveaction/cgi-bin/drupal/node/21))
  "Welcome, user! Here you can access data from an ongoing project about collective action in the United States."

- Major Urban Areas 1to500k (https://catalog.data.gov/dataset/blm-rea-cop-2010-usa-major-urbanv-areas-1to500k-poly (https://catalog.data.gov/dataset/blm-rea-cop-2010-usa-major-urbanv-areas-1to500k-poly))
  Description: "U.S. Census Urbanized Areas represents the Census 2000 Urbanized Areas (UA) and Urban Clusters (UC). A UA consists of contiguous, densely settled census block groups (BGs) and census blocks that meet minimum population density requirements (1000ppsm /500ppsm), along with adjacent densely settled census blocks that together encompass a population of at least 50,000 people. A UC consists of contiguous, densely settled census BGs and census blocks that meet minimum population density requirements, along with adjacent densely settled census blocks that together encompass a population of at least 2,500 people, but fewer than 50,000 people. The dataset covers the 50 States plus the District of Columbia within United States."

```
########################## Preliminary Data ##########################
library(haven)
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
setwd("C:/Users/brees/OneDrive/NU_Yr_I_Sem_II/Bostonography/Data/Final_Data")

df <- read_dta(file = 'final_data_v10.dta', encoding='latin1')

#slice columns of interest
df.a <- df[c('evyy', 'evmm', 'viold', 'arrestd', 'deaths')]

#make own mode function
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

#aggregate demonstration by year and month
(aggregate(df.a,
           by = list(df$newsnm),
           FUN = getmode))
```

```
##   Group.1 evyy evmm viold arrestd deaths
## 1     NYT 1965    5     0      NA      0
```

```
#aggregate violent demonstration by year and month
(aggregate(df.a,
           by = list(df$viold),
           FUN = getmode))
```

```
##   Group.1 evyy evmm viold arrestd deaths
## 1       0 1965    5     0      NA      0
## 2       1 1967    7     1       0      0
```

```
#aggregate if arrests were made by year and month
(aggregate(df.a,
           by = list(df$arrestd),
           FUN = getmode))
```

```
##   Group.1 evyy evmm viold arrestd deaths
## 1       0 1969    5     0       0      0
## 2       1 1973    7     0       1      0
```

```
#aggregate if people were killed by year and month
(aggregate(df.a,
           by = list(df$deaths),
           FUN = getmode))
```

```
##   Group.1 evyy evmm viold arrestd deaths
## 1       0 1965    5     0      NA      0
## 2       1 1968    7     1       0      1
```

Most importantly from these findings, there are some patterns to be found. Before diving into too much research, we at least know that there's likely somethig to be discorvered here in terms of the relation between time period and collective action.

To interpret this: - First table: total number of demonstrations (most in May 1965) - Second table: number of violent demonstrations (most nonviolent in May 1965, most violent in July 1967) - Third table: number of demonstrations where arrests were made (most where arrests were not made in May 1969, most where arrests were made in July 1973) - Fourth table: number of demonstrations where people were killed (most where people were not killed in May 1965, most where people were killed in July 1968)

There are some obvious patterns here. It seems that the mid to late 60s had the most protests of all kinds, and the mid 60s had fewer instances of violence, arrests, and deaths. The measures are almost certainly skewed by the heavy concentration of protests around 1965, and in the next section percentages (rather than sheer number) will be investigated.

The first measure, for one, isn't surprising. In 1965, the Vietnam war was quickly becoming less and less favorable to the American public, and the protests in May of this year were becoming more agressive, with draft card burnings becoming more commonplace. So much so, that in August, the Draft Card Mutilation Act was passed. The same kind of peaceful protests continued for the next few years, justifying our next few numbers of the years which had the least violence, arrests, and deaths.

Later years, however, outline different tensions in the country. In the late 60s, with the assissination of MLK sparking violent confrontations between civil rights protestors and police. Again, our numbers that have the most violence, arrents, and deaths around this time make perfect sense.

Another interesting observation is the heavy correlation between month and these measures. First, it seems that early summer (May-July) is a popular time to protest. Even further than that, however, for some reason, the least violence, arrests, and deaths were all in May, while the most of those same things were all in July. Upon first glance, it might seem as if there were just specific events going on in those months, but this occurs over the course of at least five years. Perhaps there is some difference in national sentiment during these close, but different times in the year.

The following code initializes all subsections of the collective action data to be used in the future visualizations and analysis.

```
library(tidyverse)
```

```
## -- Attaching packages -----------
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

```
## -- Conflicts --------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)

# how many total entries are there for boston?
df_city <- df[c('city1')]

city_freq <- table(df_city)

# the entire dataset
all_city_freq <- data.frame(city_freq)
all_city_freq <- data.frame("city" = tolower(all_city_freq$df_city), "freq" = all_city_freq$Fre
q)
all_city_freq$city <- str_to_title(all_city_freq$city)

# datasets by decade
df_city_year <- df[c('city1', 'evyy')]
city_year <- table(df_city_year)

# 1955 - 1965
city_year_55_65 <- df_city_year[df_city_year$evyy < 1966,]
city_freq_55_65 <- table(data.frame(city_year_55_65$city1))
all_city_freq_55_65 <- data.frame(city_freq_55_65)
all_city_freq_55_65 <- data.frame("city" = tolower(all_city_freq_55_65$Var1), "freq" = all_city_
freq_55_65$Freq)
all_city_freq_55_65$city <- str_to_title(all_city_freq_55_65$city)

# 1966 - 1975
city_year_66_75 <- df_city_year[df_city_year$evyy > 1966 & df_city_year$evyy < 1976,]
city_freq_66_75 <- table(data.frame(city_year_66_75$city1))
all_city_freq_66_75 <- data.frame(city_freq_66_75)
all_city_freq_66_75 <- data.frame("city" = tolower(all_city_freq_66_75$Var1), "freq" = all_city_
freq_66_75$Freq)
all_city_freq_66_75$city <- str_to_title(all_city_freq_66_75$city)

# 1976 - 1985
city_year_76_85 <- df_city_year[df_city_year$evyy > 1975 & df_city_year$evyy < 1986,]
city_freq_76_85 <- table(data.frame(city_year_76_85$city1))
all_city_freq_76_85 <- data.frame(city_freq_76_85)
all_city_freq_76_85 <- data.frame("city" = tolower(all_city_freq_76_85$Var1), "freq" = all_city_
freq_76_85$Freq)
all_city_freq_76_85$city <- str_to_title(all_city_freq_76_85$city)

# 1986 - 1995
city_year_86_95 <- df_city_year[df_city_year$evyy > 1985,]
city_freq_86_95 <- table(data.frame(city_year_86_95$city1))
all_city_freq_86_95 <- data.frame(city_freq_86_95)
all_city_freq_86_95 <- data.frame("city" = tolower(all_city_freq_86_95$Var1), "freq" = all_city_
freq_86_95$Freq)
all_city_freq_86_95$city <- str_to_title(all_city_freq_86_95$city)

# datasets by type
df_city_type <- df[c('city1', 'viold')]
city_type <- table(df_city_type)
```

```
# viold (violent)
city_type_viold <- df_city_type[df_city_type$viold == 1,]
city_type_viold <- table(data.frame(city_type_viold$city1))
all_city_type_viold <- data.frame(city_type_viold)
all_city_type_viold <- data.frame("city" = tolower(all_city_type_viold$Var1), "freq" = all_city_
type_viold$Freq)
all_city_type_viold$city <- str_to_title(all_city_type_viold$city)

# non-violent
city_type_nviold <- df_city_type[df_city_type$viold == 0,]
city_type_nviold <- table(data.frame(city_type_nviold$city1))
all_city_type_nviold <- data.frame(city_type_nviold)
all_city_type_nviold <- data.frame("city" = tolower(all_city_type_nviold$Var1), "freq" = all_cit
y_type_nviold$Freq)
all_city_type_nviold$city <- str_to_title(all_city_type_nviold$city)

# 1955 - 1965 viold
city_year_55_65_viold <- df_city_year[df_city_year$evyy < 1966 & df_city_type$viold == 1,]
city_freq_55_65_viold <- table(data.frame(city_year_55_65_viold$city1))
all_city_freq_55_65_viold <- data.frame(city_freq_55_65_viold)
all_city_freq_55_65_viold <- data.frame("city" = tolower(all_city_freq_55_65_viold$Var1), "freq"
= all_city_freq_55_65_viold$Freq)
all_city_freq_55_65_viold$city <- str_to_title(all_city_freq_55_65_viold$city)

# 1966 - 1975 viold
city_year_66_75_viold <- df_city_year[df_city_year$evyy > 1966 & df_city_year$evyy < 1976 & df_c
ity_type$viold == 1,]
city_freq_66_75_viold <- table(data.frame(city_year_66_75_viold$city1))
all_city_freq_66_75_viold <- data.frame(city_freq_66_75_viold)
all_city_freq_66_75_viold <- data.frame("city" = tolower(all_city_freq_66_75_viold$Var1), "freq"
= all_city_freq_66_75_viold$Freq)
all_city_freq_66_75_viold$city <- str_to_title(all_city_freq_66_75_viold$city)

# 1976 - 1985 viold
city_year_76_85_viold <- df_city_year[df_city_year$evyy > 1975 & df_city_year$evyy < 1986 & df_c
ity_type$viold == 1,]
city_freq_76_85_viold <- table(data.frame(city_year_76_85_viold$city1))
all_city_freq_76_85_viold <- data.frame(city_freq_76_85_viold)
all_city_freq_76_85_viold <- data.frame("city" = tolower(all_city_freq_76_85_viold$Var1), "freq"
= all_city_freq_76_85_viold$Freq)
all_city_freq_76_85_viold$city <- str_to_title(all_city_freq_76_85_viold$city)

# 1986 - 1995 viold
city_year_86_95_viold <- df_city_year[df_city_year$evyy > 1985 & df_city_type$viold == 1,]
city_freq_86_95_viold <- table(data.frame(city_year_86_95_viold$city1))
all_city_freq_86_95_viold <- data.frame(city_freq_86_95_viold)
all_city_freq_86_95_viold <- data.frame("city" = tolower(all_city_freq_86_95_viold$Var1), "freq"
= all_city_freq_86_95_viold$Freq)
all_city_freq_86_95_viold$city <- str_to_title(all_city_freq_86_95_viold$city)
```

The following code creates union dataframes with the spatial map data and the collective action data per city. I filtered the spatial data for cities with population over 100000 for two reasons. The first is that very small cities that only appeared in the collective action data for either 1 or 0 protests were not in the spatial dataset, and therefore

causes errors and double counting in the results. The second reason is that only counting the cities above that mark made maps easier to read and less cluttered with fairly irrelevant data to our purpose.
Additionally, the blank, all gray map is to display which cities are to be shown, and what the default size of a marker is (important when changing it to scale later).

In terms of format, for the rest of this paper, the goal of visualization and explanation will be stated before, then there will be all the maps, then there will be analysis at the end. I chose this format to facilitate reading the analysis all at once while being able to refer back to the visualizations easily.

```
library(sf)
```

```
## Linking to GEOS 3.6.1, GDAL 2.2.3, PROJ 4.9.3
```

```
library(maptools)
```

```
## Loading required package: sp
```

```
## Checking rgeos availability: TRUE
```

```
library(ggspatial)
```

```
## Warning: package 'ggspatial' was built under R version 3.6.3
```

```
library(rgdal)
```

```
## rgdal: version: 1.4-8, (SVN revision 845)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
##  Path to GDAL shared files: C:/Users/brees/Documents/R/win-library/3.6/rgdal/gdal
##  GDAL binary built with GEOS: TRUE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files: C:/Users/brees/Documents/R/win-library/3.6/rgdal/proj
##  Linking to sp version: 1.3-2
```

```
require(fuzzyjoin)
```

```
## Loading required package: fuzzyjoin
```

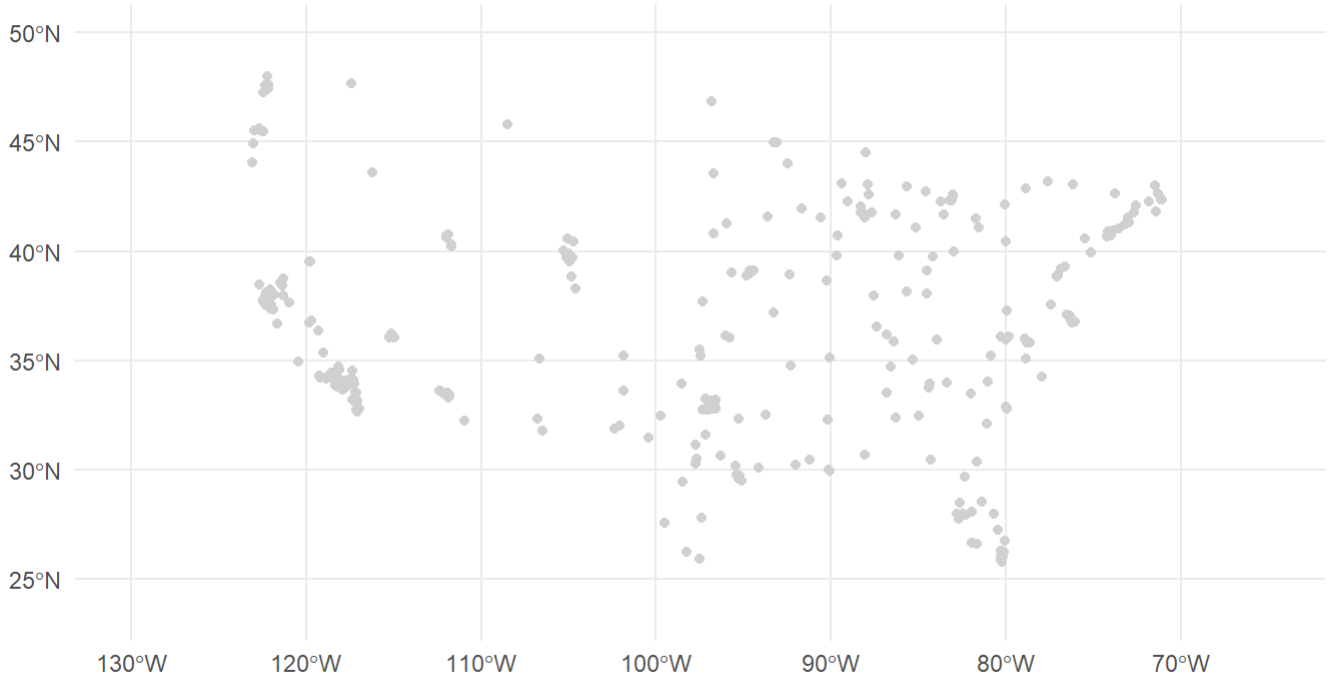
```
## Warning: package 'fuzzyjoin' was built under R version 3.6.3
```

```
setwd("C:/Users/brees/OneDrive/NU_Yr_I_Sem_II/Bostonography/Data/Final_Data")

us_spatial_cities <- st_read(
  dsn= ".",
  layer="unnamed.gdb-point",
  stringsAsFactors = FALSE
)
```

```
## Reading layer `unnamed.gdb-point' from data source `C:\Users\brees\OneDrive\NU_Yr_I_Sem_II\Bo
stonography\Data\Final_Data' using driver `ESRI Shapefile'
## Simple feature collection with 3886 features and 48 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: -159.3191 ymin: 19.58272 xmax: -68.77234 ymax: 64.86928
## epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_defs
```

```
us_spatial_cities <- us_spatial_cities[us_spatial_cities$POPULATION > 100000, ]

# whole dataset
df_plus_spatial <- full_join(us_spatial_cities, all_city_freq, by = c('NAME' = 'city'))
df_plus_spatial[is.na(df_plus_spatial)] <- 0
df_plus_spatial <- df_plus_spatial[df_plus_spatial$POPULATION > 0, ]

# 1955 - 1965
df_plus_spatial_55_65 <- full_join(us_spatial_cities, all_city_freq_55_65, by = c('NAME' = 'cit
y'))
df_plus_spatial_55_65[is.na(df_plus_spatial_55_65)] <- 0
df_plus_spatial_55_65 <- df_plus_spatial_55_65[df_plus_spatial_55_65$POPULATION > 0, ]

# 1966 - 1975
df_plus_spatial_66_75 <- full_join(us_spatial_cities, all_city_freq_66_75, by = c('NAME' = 'cit
y'))
df_plus_spatial_66_75[is.na(df_plus_spatial_66_75)] <- 0
df_plus_spatial_66_75 <- df_plus_spatial_66_75[df_plus_spatial_66_75$POPULATION > 0, ]

# 1976 - 1985
df_plus_spatial_76_85 <- full_join(us_spatial_cities, all_city_freq_76_85, by = c('NAME' = 'cit
y'))
df_plus_spatial_76_85[is.na(df_plus_spatial_76_85)] <- 0
df_plus_spatial_76_85 <- df_plus_spatial_76_85[df_plus_spatial_76_85$POPULATION > 0, ]

# 1986 - 1995
df_plus_spatial_86_95 <- full_join(us_spatial_cities, all_city_freq_86_95, by = c('NAME' = 'cit
y'))
df_plus_spatial_86_95[is.na(df_plus_spatial_86_95)] <- 0
df_plus_spatial_86_95 <- df_plus_spatial_86_95[df_plus_spatial_86_95$POPULATION > 0, ]

# viold
df_plus_spatial_viold <- full_join(us_spatial_cities, all_city_type_viold, by = c('NAME' = 'cit
y'))
df_plus_spatial_viold[is.na(df_plus_spatial_viold)] <- 0
df_plus_spatial_viold <- df_plus_spatial_viold[df_plus_spatial_viold$POPULATION > 0, ]
df_plus_spatial_viold <- head(df_plus_spatial_viold, nrow(df_plus_spatial_viold) - 1)

# nviold
df_plus_spatial_nviold <- full_join(us_spatial_cities, all_city_type_nviold, by = c('NAME' = 'ci
ty'))
df_plus_spatial_nviold[is.na(df_plus_spatial_nviold)] <- 0
df_plus_spatial_nviold <- df_plus_spatial_nviold[df_plus_spatial_nviold$POPULATION > 0, ]
df_plus_spatial_nviold <- head(df_plus_spatial_nviold, nrow(df_plus_spatial_nviold) - 1)

# 1955 - 1965 viold
df_plus_spatial_55_65_viold <- full_join(us_spatial_cities, all_city_freq_55_65_viold, by = c('N
AME' = 'city'))
df_plus_spatial_55_65_viold[is.na(df_plus_spatial_55_65_viold)] <- 0
df_plus_spatial_55_65_viold <- df_plus_spatial_55_65_viold[df_plus_spatial_55_65_viold$POPULATIO
N > 0, ]

# 1966 - 1975 viold
```

```
df_plus_spatial_66_75_viold <- full_join(us_spatial_cities, all_city_freq_66_75_viold, by = c('N
AME' = 'city'))
df_plus_spatial_66_75_viold[is.na(df_plus_spatial_66_75_viold)] <- 0
df_plus_spatial_66_75_viold <- df_plus_spatial_66_75_viold[df_plus_spatial_66_75_viold$POPULATIO
N > 0, ]

# 1976 - 1985 viold
df_plus_spatial_76_85_viold <- full_join(us_spatial_cities, all_city_freq_76_85_viold, by = c('N
AME' = 'city'))
df_plus_spatial_76_85_viold[is.na(df_plus_spatial_76_85_viold)] <- 0
df_plus_spatial_76_85_viold <- df_plus_spatial_76_85_viold[df_plus_spatial_76_85_viold$POPULATIO
N > 0, ]

# 1986 - 1995 viold
df_plus_spatial_86_95_viold <- full_join(us_spatial_cities, all_city_freq_86_95_viold, by = c('N
AME' = 'city'))
df_plus_spatial_86_95_viold[is.na(df_plus_spatial_86_95_viold)] <- 0
df_plus_spatial_86_95_viold <- df_plus_spatial_86_95_viold[df_plus_spatial_86_95_viold$POPULATIO
N > 0, ]


#blank map
ggplot() + geom_sf(data = df_plus_spatial, color = "#D0D0D0") + coord_sf(xlim = c(-130, -65), yl
im = c(23.5, 50)) + theme_minimal()
```



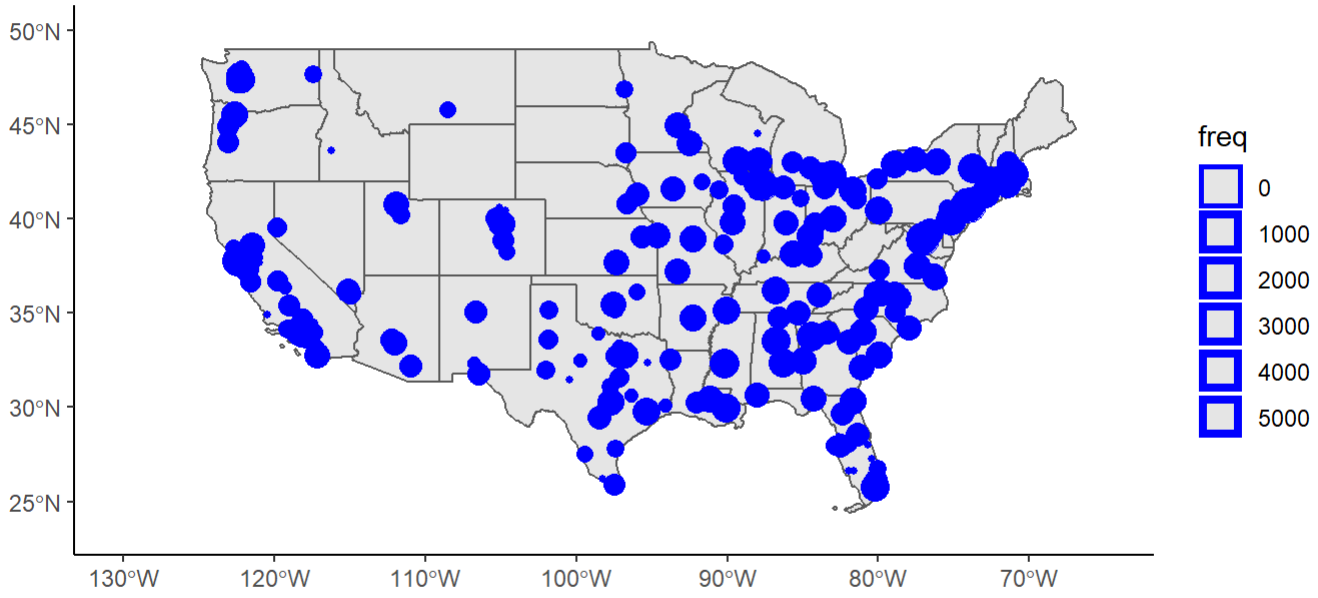
These next four maps represent frequencies of demonstrations in the aformentioned cities. The first two measure the raw number of demonstrations, while the last two represent number of demonstration per-captia (a more enlightening, but harder to differentiate map). I chose to use both the size-only and the mixed size and color gradient maps because they have different strenghts. The size-only is more comparable to later maps that use gradient for other measures, and the size-only per-capita map is easier to read than the mixed one. However, the raw number map with gradient is much more clear than its predecessor.

To discuss types of measurement The first maps use a log scale (to help account for population differences), while those on the right represent per capita, with no log scale, as to see the outliers more clearly. In practice, we are really accounting for population in both methods, but in different ways as to gain different angles from which to view the data.

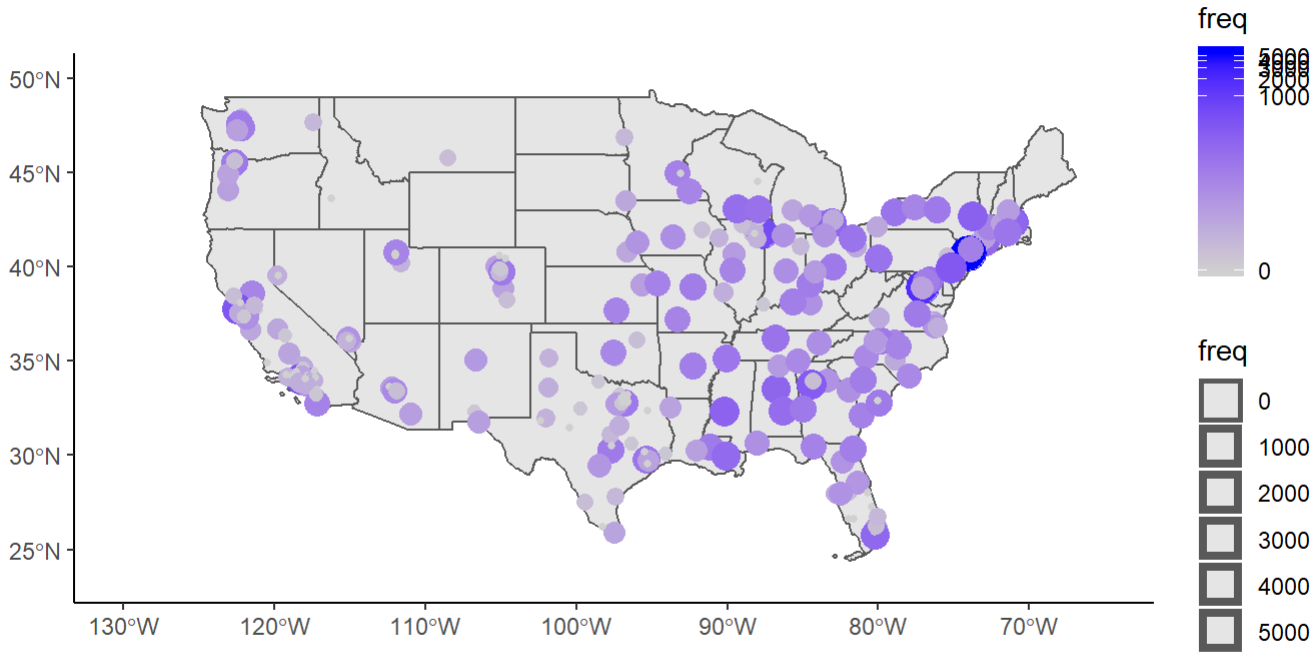
```
setwd("C:/Users/brees/OneDrive/NU_Yr_I_Sem_II/Bostonography/Data/Final_Data")

state_map <- st_read(
  dsn= ".",
  layer="tl_2015_us_state",
  stringsAsFactors = FALSE
)
```

```
## Reading layer `tl_2015_us_state' from data source `C:\Users\brees\OneDrive\NU_Yr_I_Sem_II\Bos
tonography\Data\Final_Data' using driver `ESRI Shapefile'
## Simple feature collection with 56 features and 14 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -179.2311 ymin: -14.60181 xmax: 179.8597 ymax: 71.44106
## epsg (SRID):    4269
## proj4string:    +proj=longlat +datum=NAD83 +no_defs
```

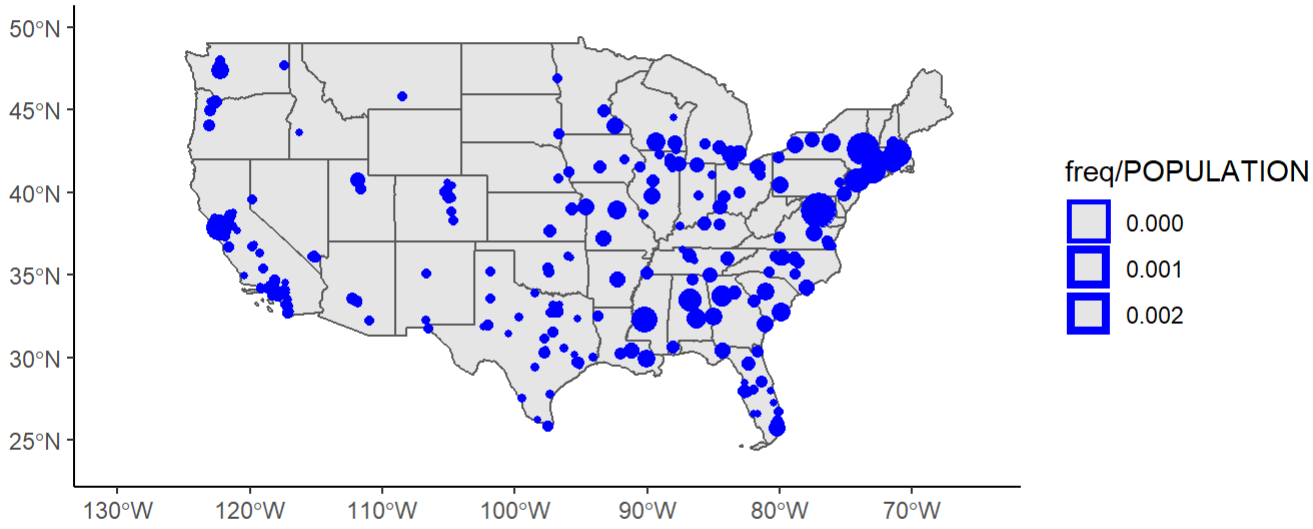
```
#freq uncolored
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq), color =
"blue") + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans = "pse
udo_log") + theme_classic()
```



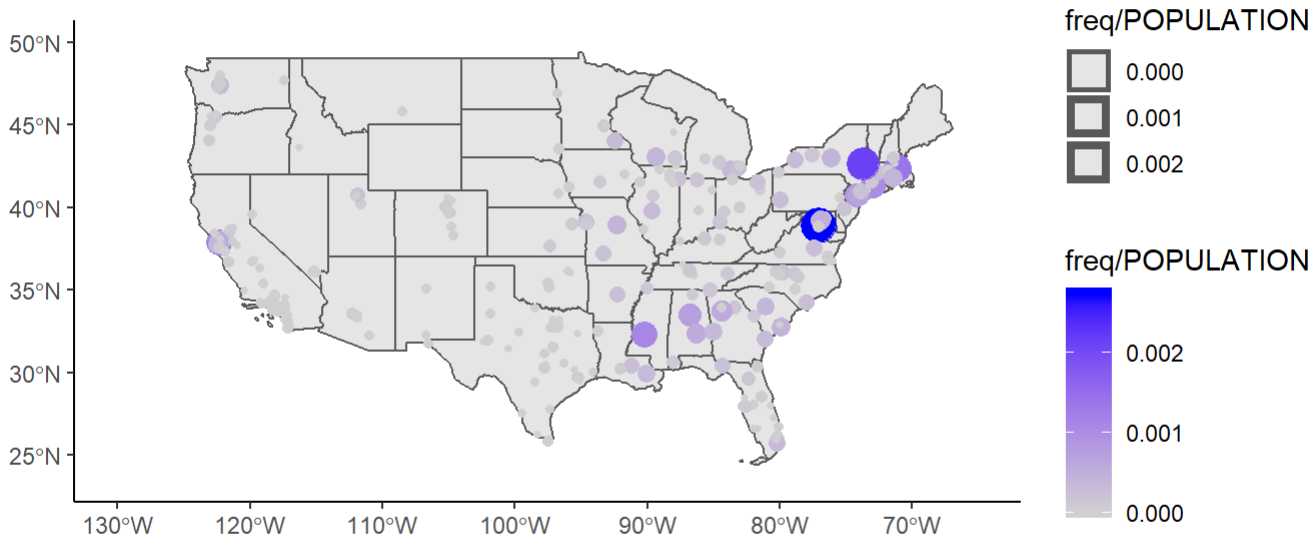

```
#freq colored
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq, color =
 freq)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans = "pse
udo_log") + scale_color_gradientn(colors = c("#D0D0D0", "blue"), trans = "pseudo_log") + theme_c
lassic()
```




```
#freq plain
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq / POPULAT
ION), color = "blue") + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuou
s(trans = "pseudo_log") + theme_classic()
```



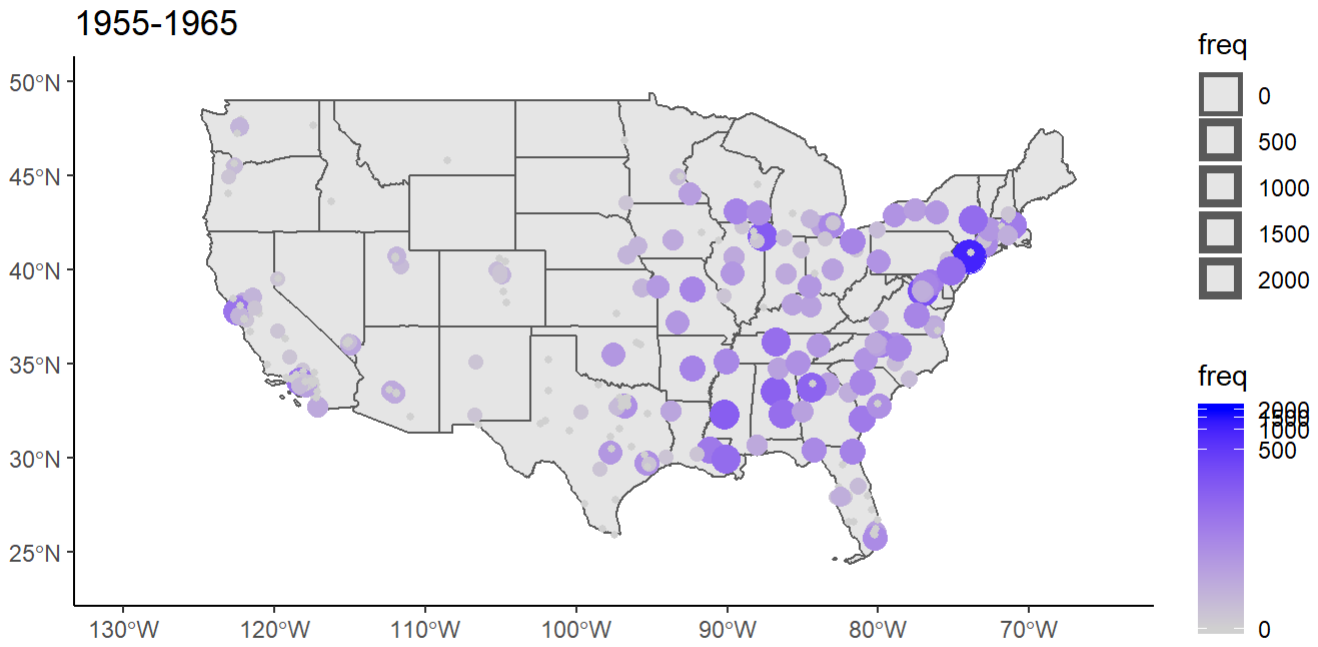

```
#freq gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq / POPULAT
ION, color = freq / POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_siz
e_continuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("#D0D0D0", "blue"), trans
 = "pseudo_log") + theme_classic()
```



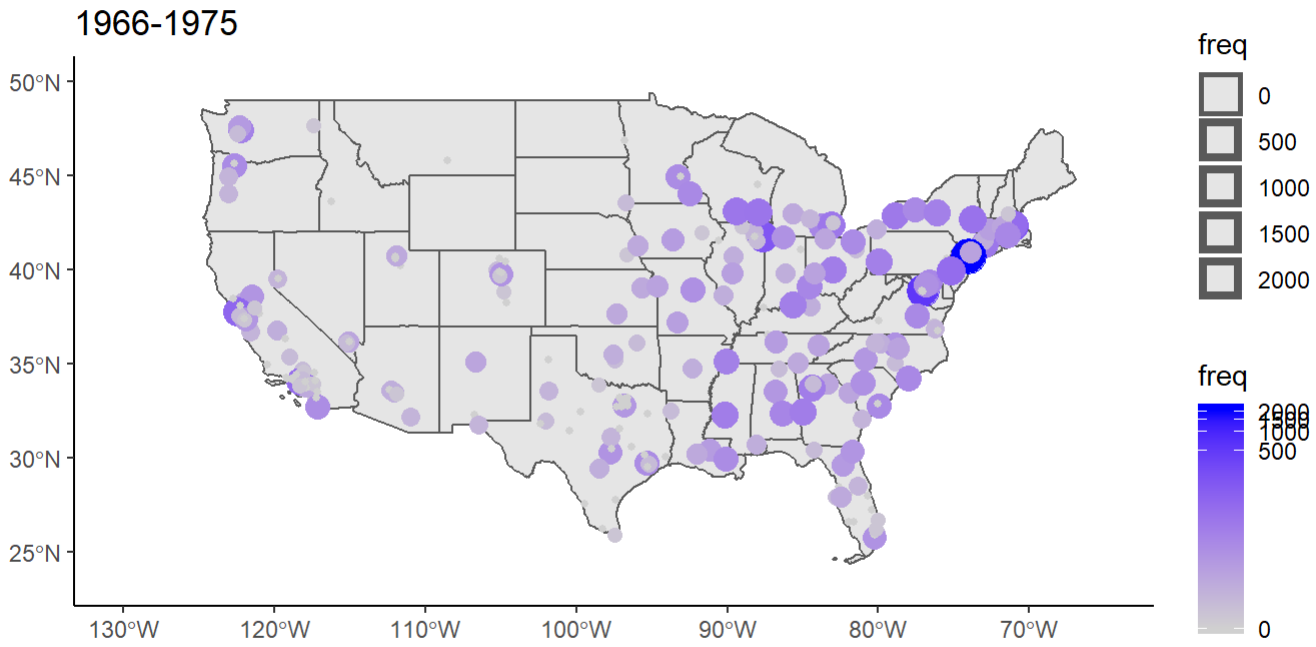
For the results, nothing here is too surprising, but it is certainly valuable. First, it makes sense that the coasts tend to protest more due to their population centers, but even when that's accounted for, we see a few strong outliers on the final map, especially that dark blue circle, Washington D.C.. This makes sense that our capital which historically has a tradition of being very activist would have a vastly larger per capita demonstration than other cities.

Moving on to the separations by decade. I decided to separate each of the four decades the data covers to better understand the relationship between politcal unrest and collective action. The Stanford collective action dataset studies the perfect time period, because the 60s and 70s were so full of civil clashes, and the 80s and 90s were relatively very calm. Additionally, the 60s focused more on civil rights, while the 70s focused more on war, so perhaps there are differences in protests when fighting for different kinds of change.

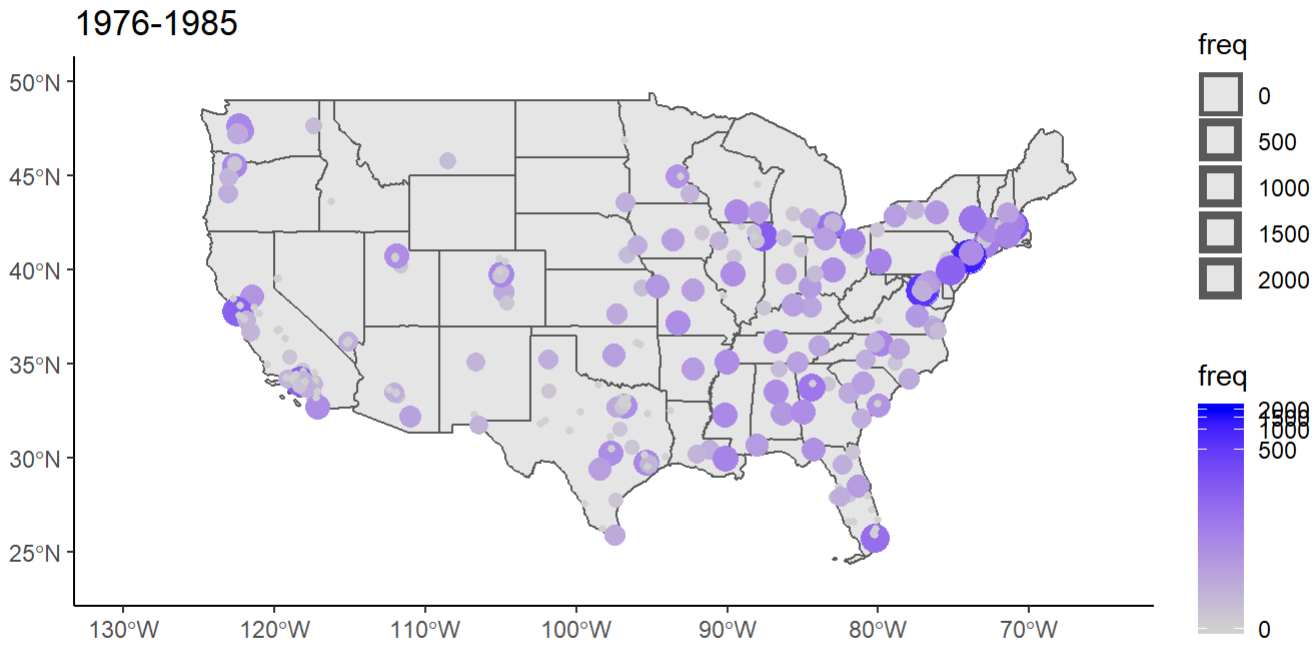
```
#decades
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_55_65, aes(size = freq, co
lor = freq)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans =
"pseudo_log", limits = c(0, 2000)) + scale_color_gradientn(colors = c("#D0D0D0", "blue"), trans
 = "pseudo_log", limits = c(0, 2000)) + theme_classic() + ggtitle("1955-1965")
```



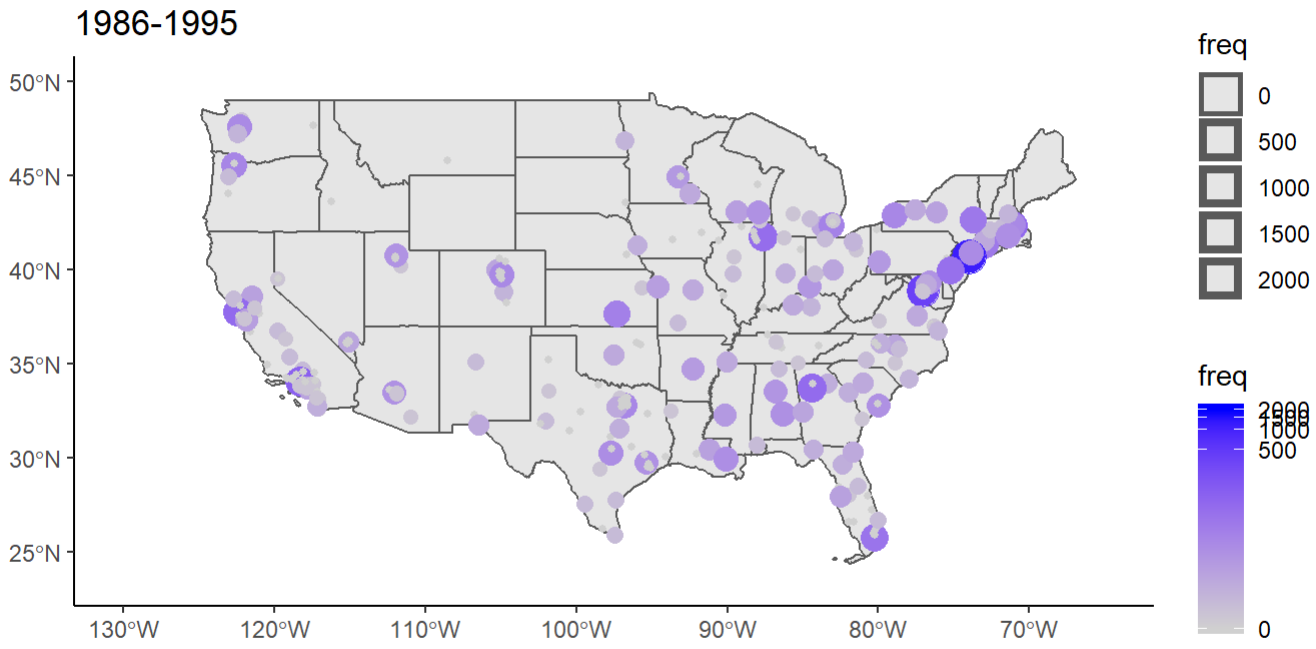

```
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_66_75, aes(size = freq, co
lor = freq)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans =
"pseudo_log") + scale_color_gradientn(colors = c("#D0D0D0", "blue"), trans = "pseudo_log") + the
me_classic() + ggtitle("1966-1975")
```




```
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_76_85, aes(size = freq, co
lor = freq)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans =
"pseudo_log", limits = c(0, 2000)) + scale_color_gradientn(colors = c("#D0D0D0", "blue"), trans
 = "pseudo_log", limits = c(0, 2000)) + theme_classic() + ggtitle("1976-1985")
```



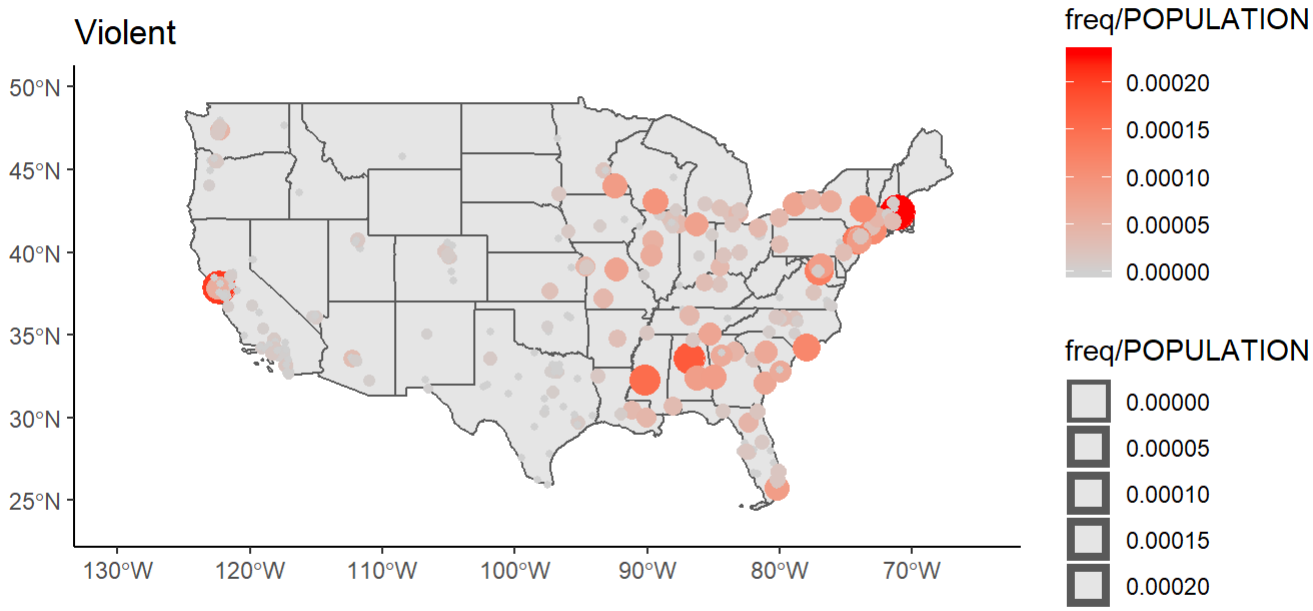

```
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_86_95, aes(size = freq, co
lor = freq)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans =
"pseudo_log", limits = c(0, 2000)) + scale_color_gradientn(colors = c("#D0D0D0", "blue"), trans
 = "pseudo_log", limits = c(0, 2000)) + theme_classic() + ggtitle("1986-1995")
```



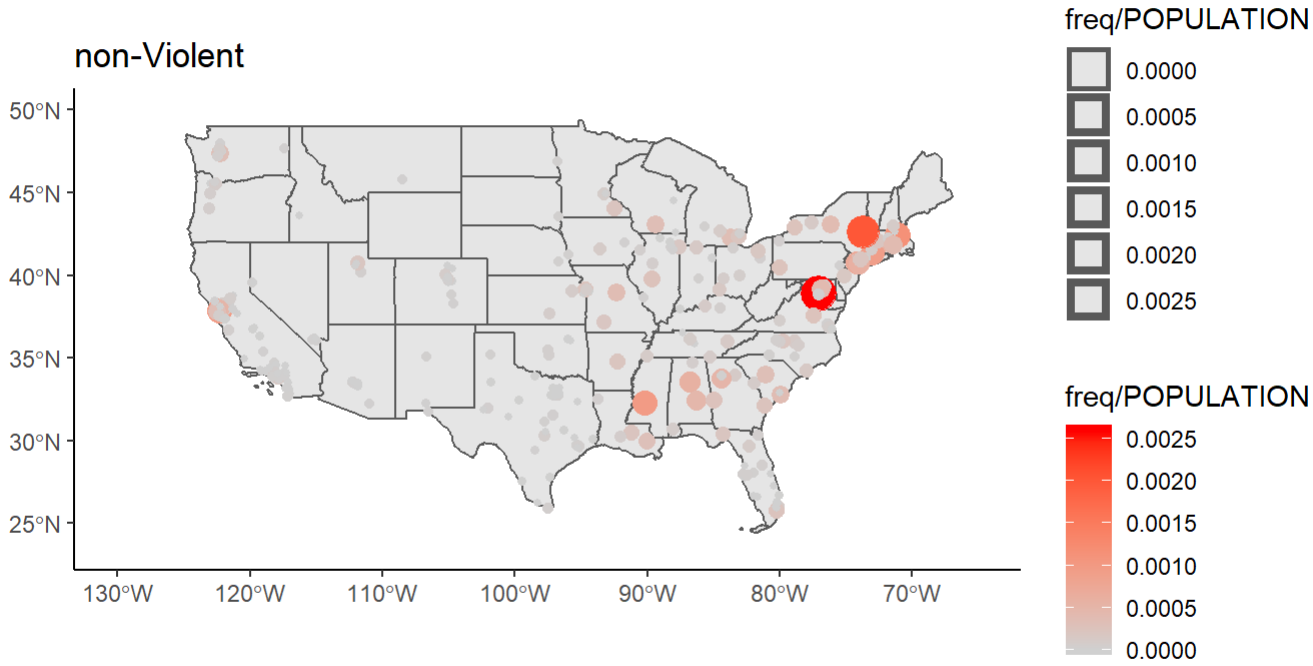
The difference in the four decades (1965-1995) is interesting, but not too surprising. In the late 50s and early 60s, segregation protests were occurring up until it was outlawed at the end of this period in 1964. Notice the preponderance of protests in the south in this decade compared to the others. Additionally, notice from 66-75 the concentration on the coasts and north. Vietnam protests at this time were decidedly more of a coastal liberal issue, so that can certainly explain what we see on this map. In the next two, the demonstrations are less, and from 86 to 95, decidedly more spread out. I would be interested to see possibly what was primarily being protested at the time, but no specific regional ones come to the mind of a CS (not history) major.

Contining to the first metric of the demonstrations, violence in protests is probably one of the most important ways in which to judge them. Originally, before looking too deeply into this dataset, I had planned to look into arrests and deaths as well, but those are either much more infrequent or at least much more infrequently documented. Luckily, violence typically encompasses both of the former measures as a subset in itself. The following two maps are simply nationwide violent protests, followed by nonviolent ones.

```
#viold
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_viold, aes(size = freq / P
OPULATION, color = freq / POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + sca
le_size_continuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("#D0D0D0", "red"), t
rans = "pseudo_log") + theme_classic() + ggtitle("Violent")
```

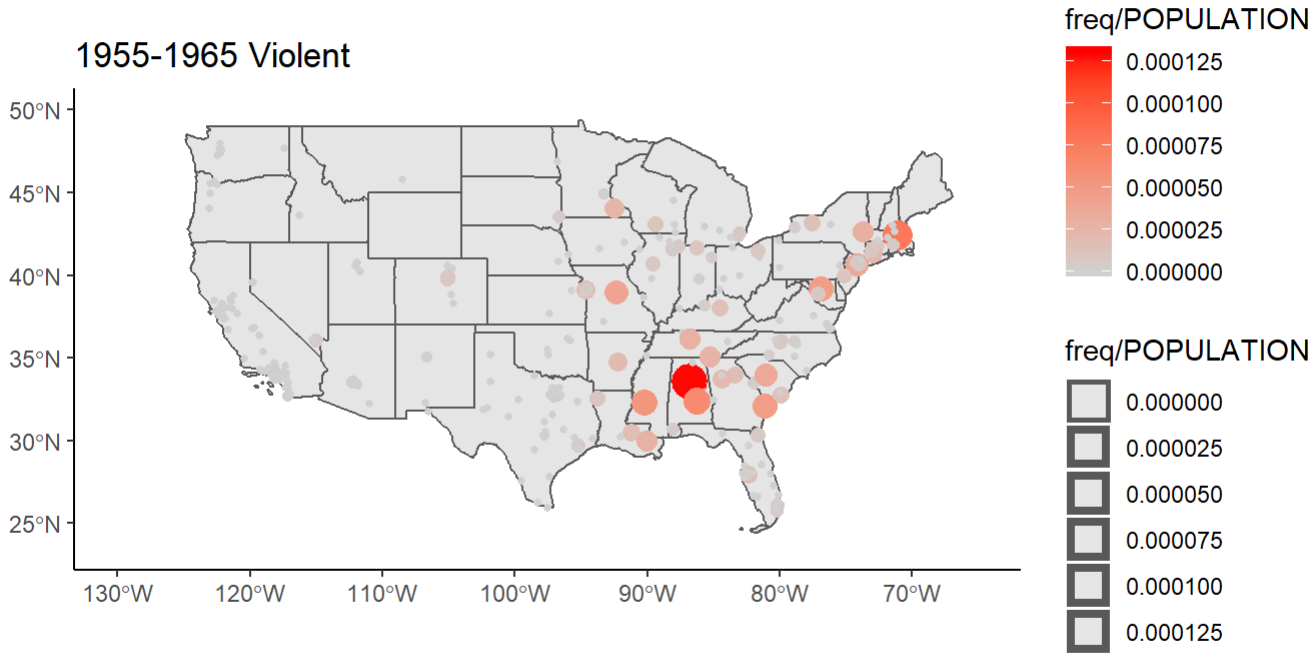



```
#viold
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_nviold, aes(size = freq /
 POPULATION, color = freq / POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + s
cale_size_continuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("#D0D0D0", "red"),
trans = "pseudo_log") + theme_classic() +  ggtitle("non-Violent")
```

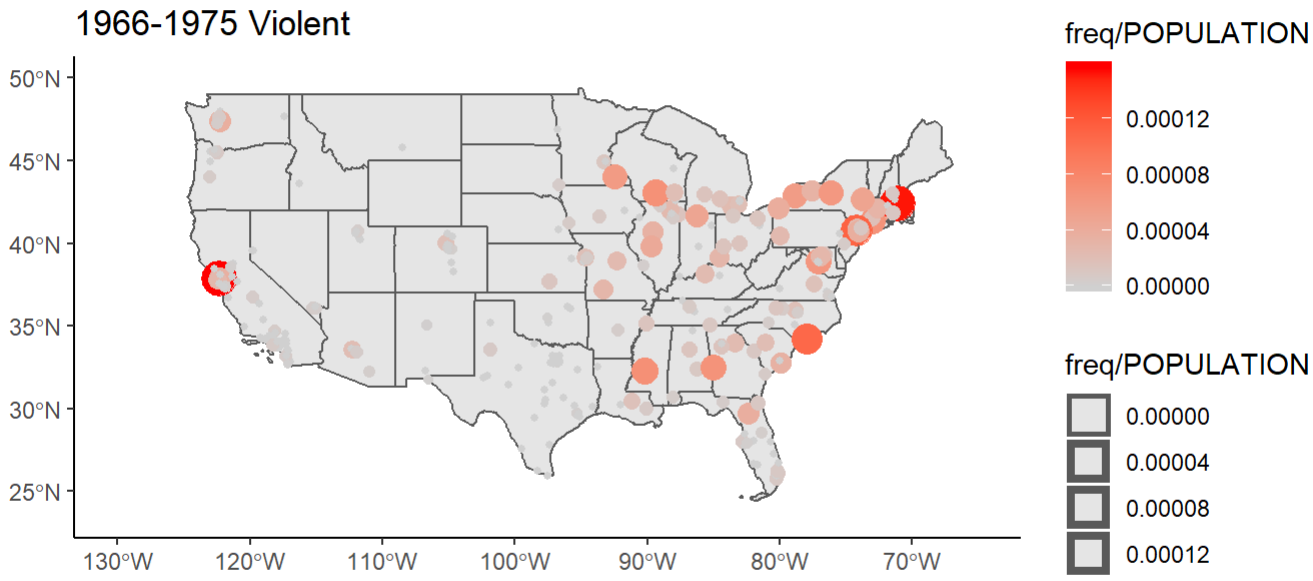


Interestingly, the non-violent map looks strikingly similar to the per-capita frequency map, so it seems that places arent “more peaceful”, but looking to the first map, they can certainly be more violent. The bigger population centers and the south tend to have much more violent protests, but breaking them down into decade will allow us to get a much better idea as to what causes violence.

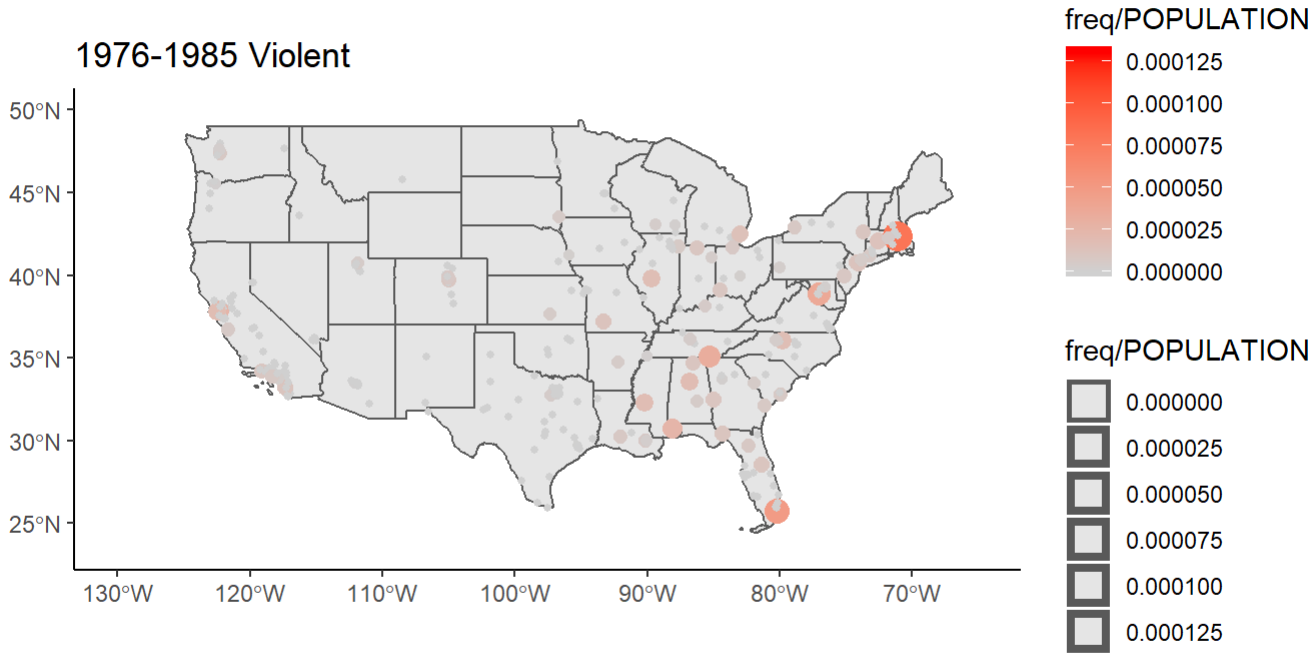
```
#decades viold
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_55_65_viold, aes(size = fr
eq / POPULATION, color = freq / POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50))
+ scale_size_continuous(trans = "pseudo_log", limits = c(0, 0.00013)) + scale_color_gradientn(co
lors = c("#D0D0D0", "red"), trans = "pseudo_log", limits = c(0, 0.00013)) + theme_classic() + gg
title("1955-1965 Violent")
```



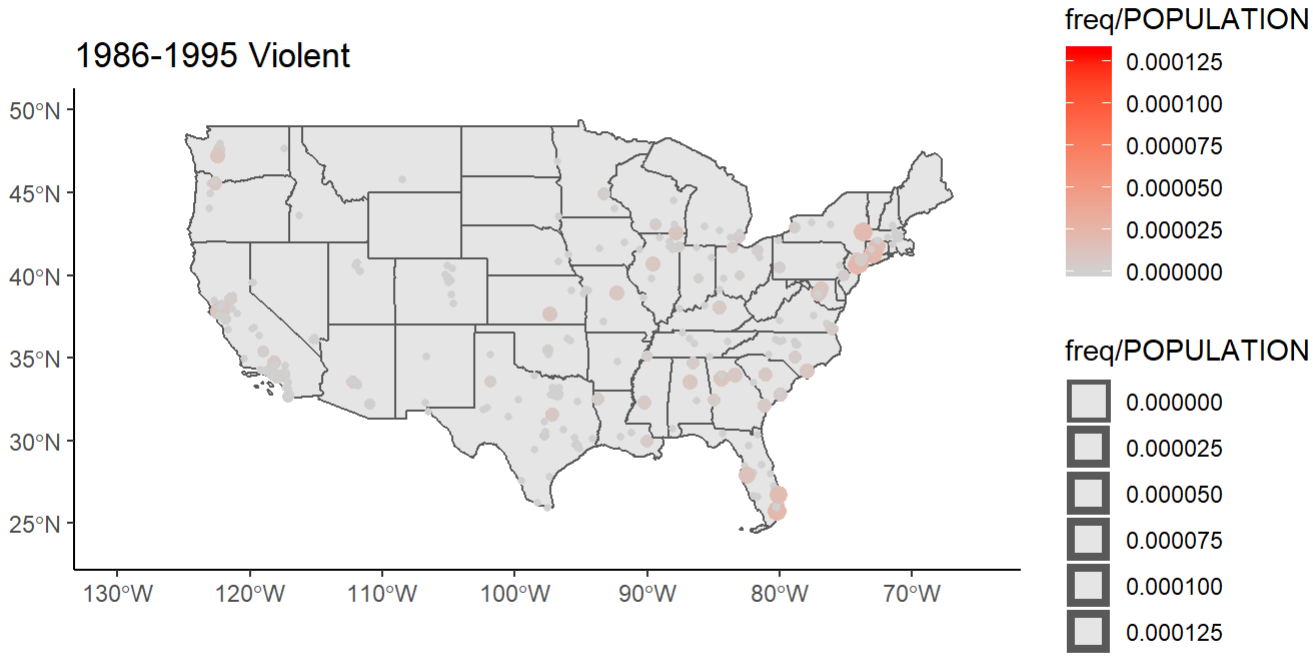

```
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_66_75_viold, aes(size = fr
eq / POPULATION, color = freq / POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50))
+ scale_size_continuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("#D0D0D0", "re
d"), trans = "pseudo_log") + theme_classic() + ggtitle("1966-1975 Violent")
```




```
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_76_85_viold, aes(size = fr
eq / POPULATION, color = freq / POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50))
+ scale_size_continuous(trans = "pseudo_log", limits = c(0, 0.00013)) + scale_color_gradientn(co
lors = c("#D0D0D0", "red"), trans = "pseudo_log", limits = c(0, 0.00013)) + theme_classic() + gg
title("1976-1985 Violent")
```



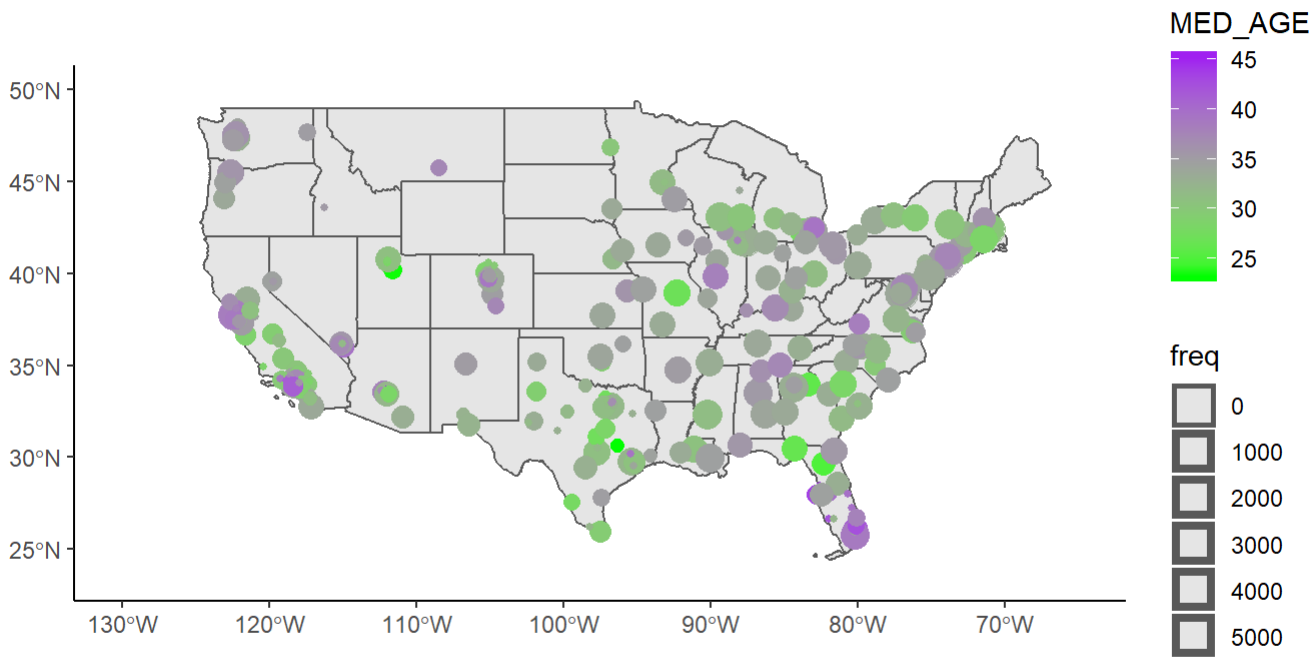

```
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial_86_95_viold, aes(size = fr
eq / POPULATION, color = freq / POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50))
+ scale_size_continuous(trans = "pseudo_log", limits = c(0, 0.00013)) + scale_color_gradientn(co
lors = c("#D0D0D0", "red"), trans = "pseudo_log", limits = c(0, 0.00013)) + theme_classic() + gg
title("1986-1995 Violent")
```



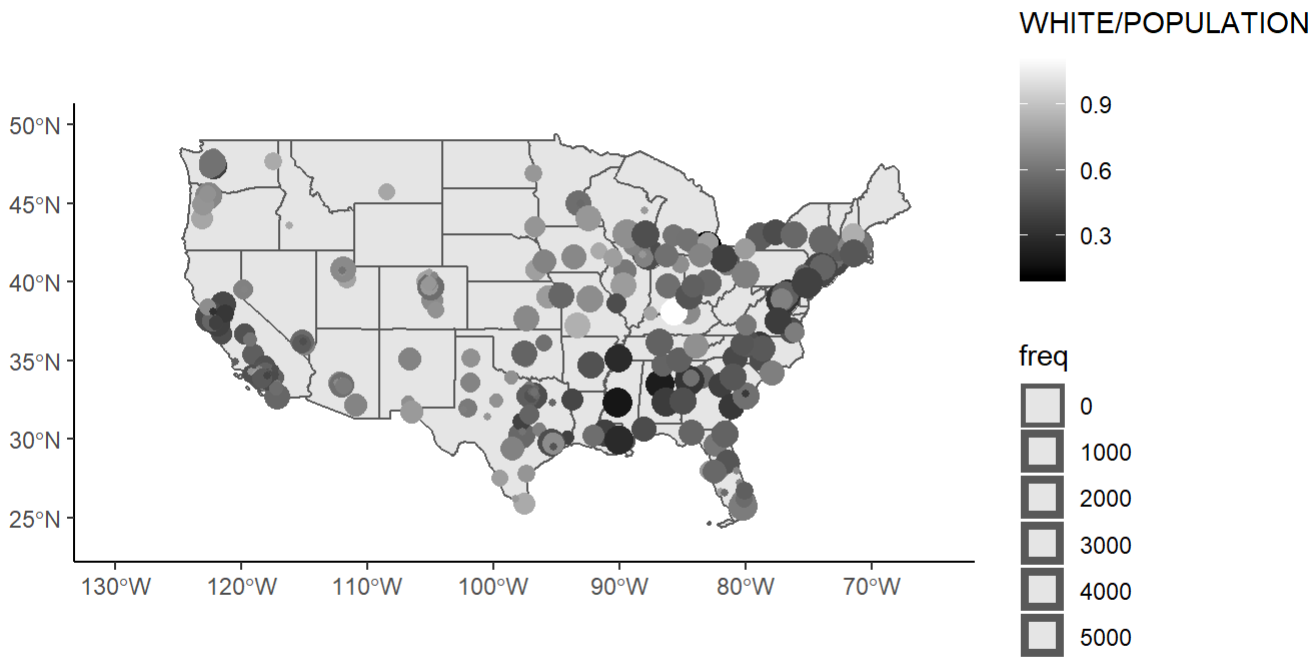
These results are extremely illuminating. What political movement definited the late 50s and all of the 60s? Civil rights. There's a huge concentration of violence in the deep south, right where you would expect it to be in the first map (and carried over to the second). What movement defined the late 60s and early 70s? Vietnam. Notice the preponderance of violence on the coasts, where the war was least popular. The following two decades, being very peaceful and the U.S. being very united show a huge dropoff in violent protests, which makes perfect sense.

In the following maps, I plotted the gradient as demographic data taken from the spatial city data. I used, median age, white / population, black / population, average family size, and gender. There are reasons for these to be the metrics. First, age is something typically associated with collective action, especially when discussing events like Vietnam protests. This might be because younger people have more progressive views, or because they're more out in the world, but it would be interesting to see if this perception is true. Race is always an important factor when analyzing demographics, mainly because it correlates to so many different things like opportunity and socioeconomic status. Additionally, the civil rights movement is covered in our data, which is obvioulsy a major race issue. The average family size was a piece of data that I found when looking through the spatial dataset, and was interested in. It could possibly correlate to religion or socioeconomics. If it provides a meaningful correlation to protests, I would definitely look more into it (spoiler alert: it doesn't have any measurable impact, and simply seems location-based). Lastly, gender is another important demographic, and I, as a researcher, am not aware of any perceptions that one gender protests more, so I was interested to look at that as well.

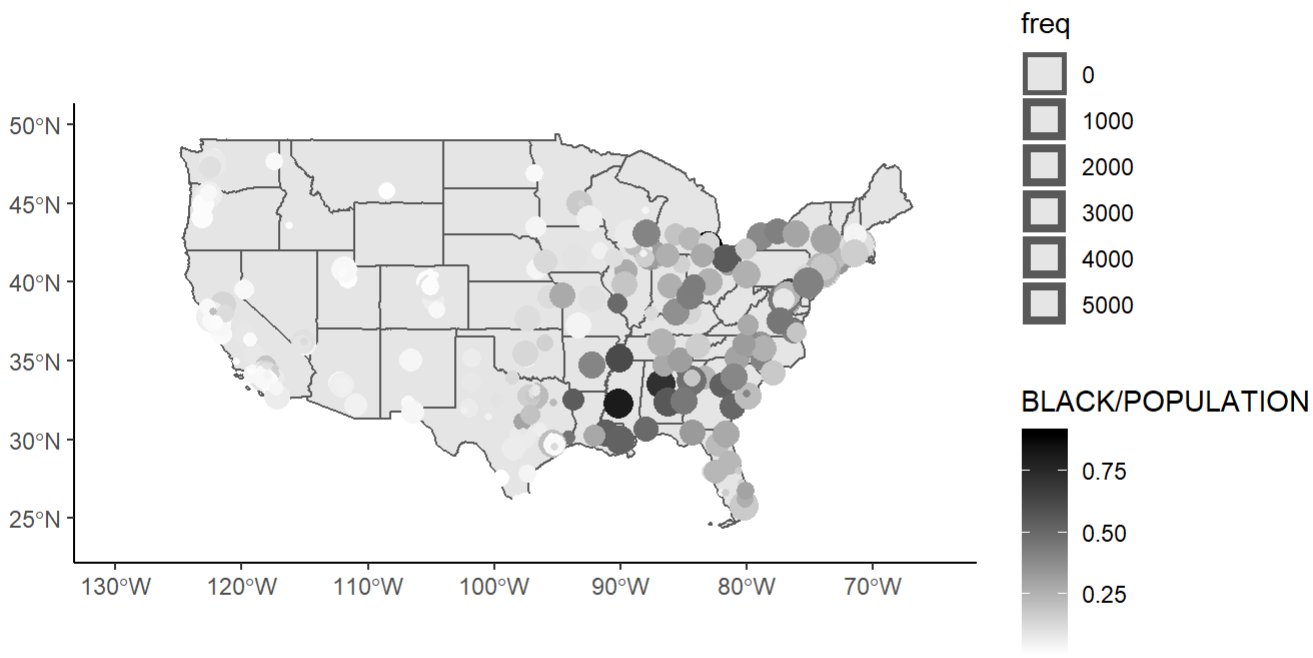
```
#med_age gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq, color =
 MED_AGE)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans =
"pseudo_log") + scale_color_gradientn(colors = c("green", "purple")) + theme_classic()
```



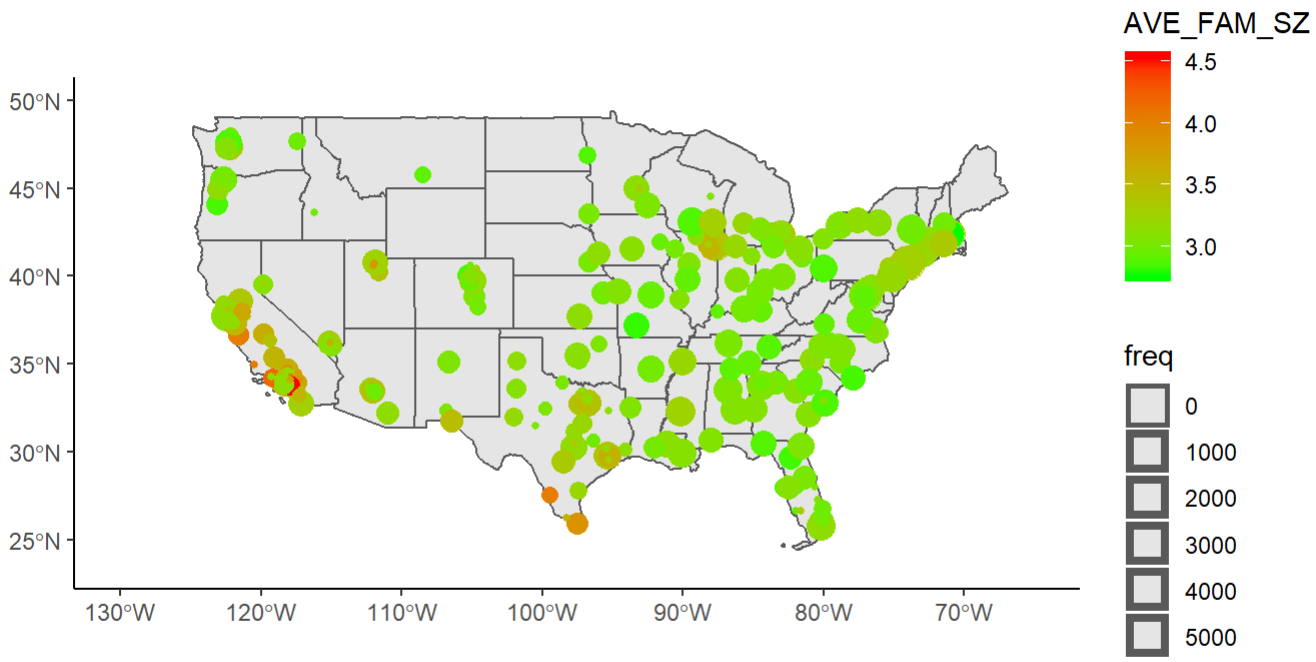

```
#white gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq, color =
 WHITE/POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous
(trans = "pseudo_log") + scale_color_gradientn(colors = c("black", "white")) + theme_classic()
```



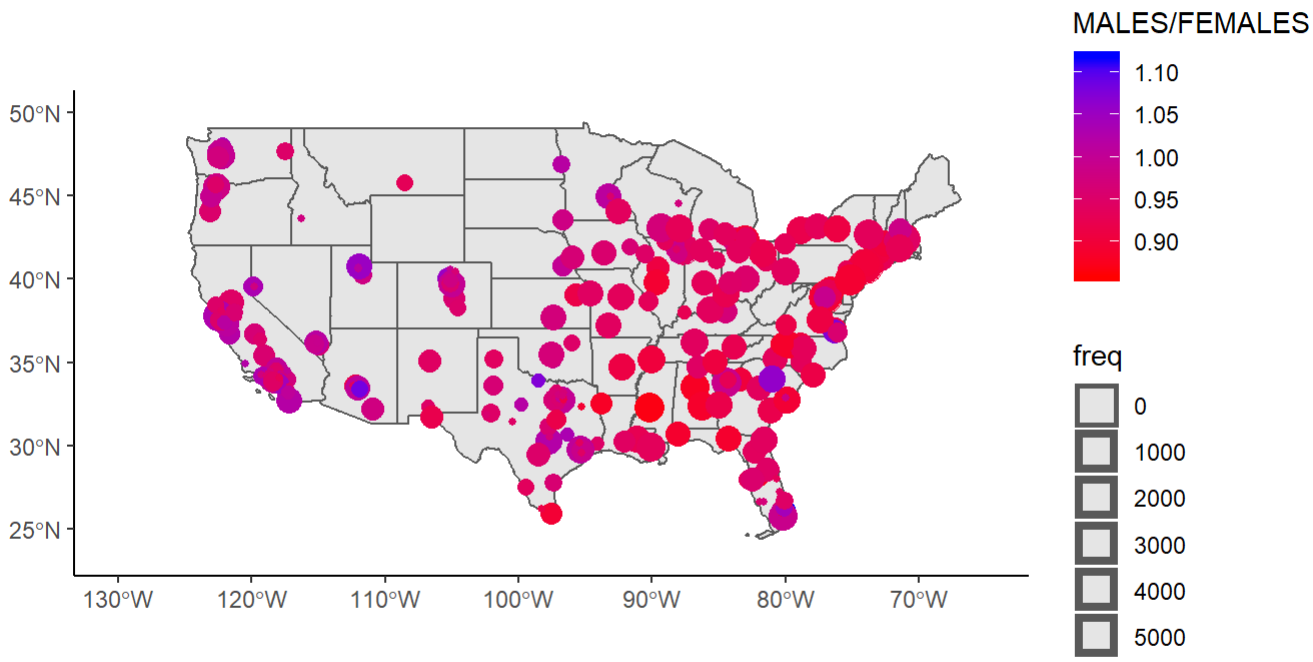

```
#black gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq, color =
 BLACK/POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous
(trans = "pseudo_log") + scale_color_gradientn(colors = c("white", "black")) + theme_classic()
```




```
#avg_fam_sz gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq, color =
 AVE_FAM_SZ)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(trans
 = "pseudo_log") + scale_color_gradientn(colors = c("green", "red")) + theme_classic()
```




```
#male - female gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq, color =
 MALES/FEMALES)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuous(tra
ns = "pseudo_log") + scale_color_gradientn(colors = c("red", "blue")) + theme_classic()
```



Unfortunately, none of these are particularly enlightening.

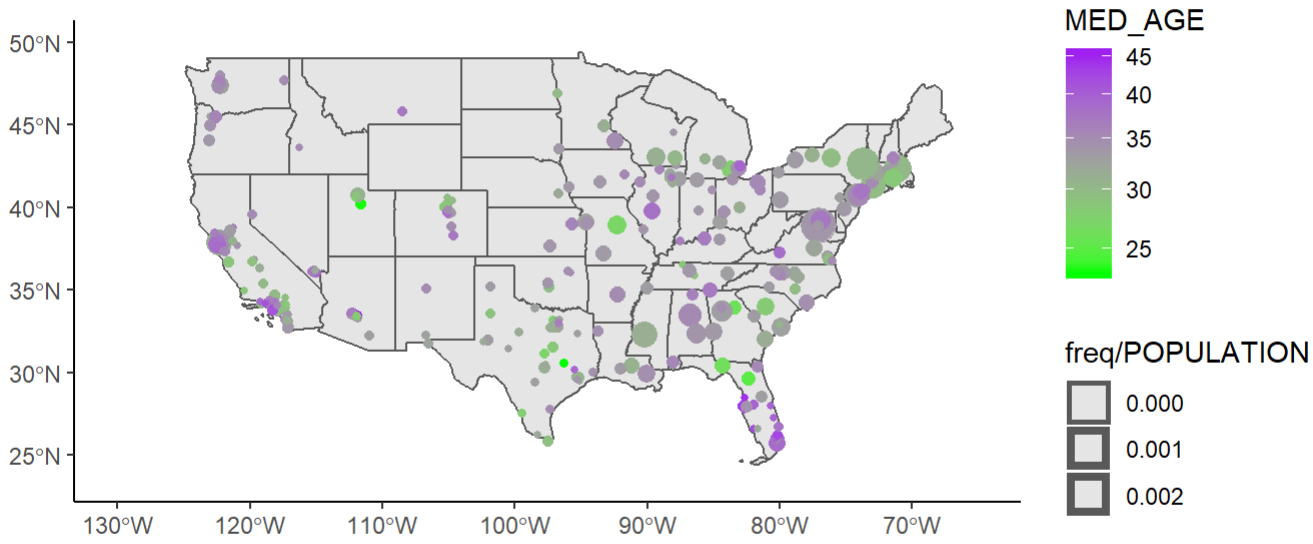
The median age map definitely shows that if a city has a median age of above 40, it will not have a high amount of protest, but the amount of these cities were lacking. There does not seem to be much correlation between a younger population and more protests, as one would maybe expect.

Looking at the race map, perhaps there is a slight edge towards diverse cities in terms of who protests more, but if there is, it's again a very weak correlation.

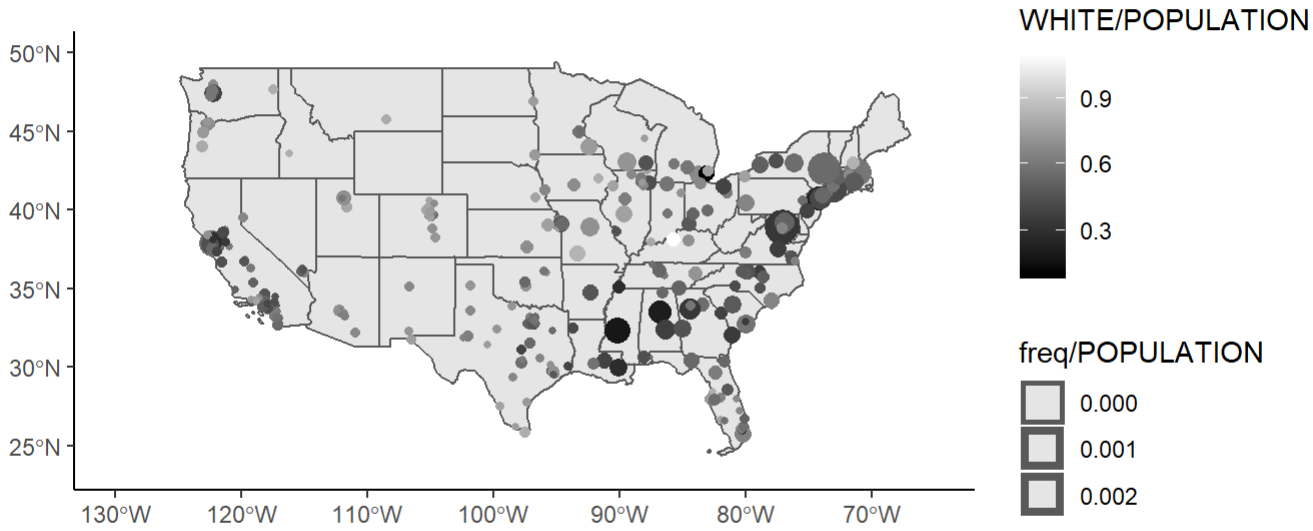
The last two measures seem to have little to no effect on protets.

It struck me that while these maps and their absolute measures are good, perhaps a freqency of protest to population map would be more effective, as these are probably putting too much weight into city size. The following are the same maps, but adjusted per capita.

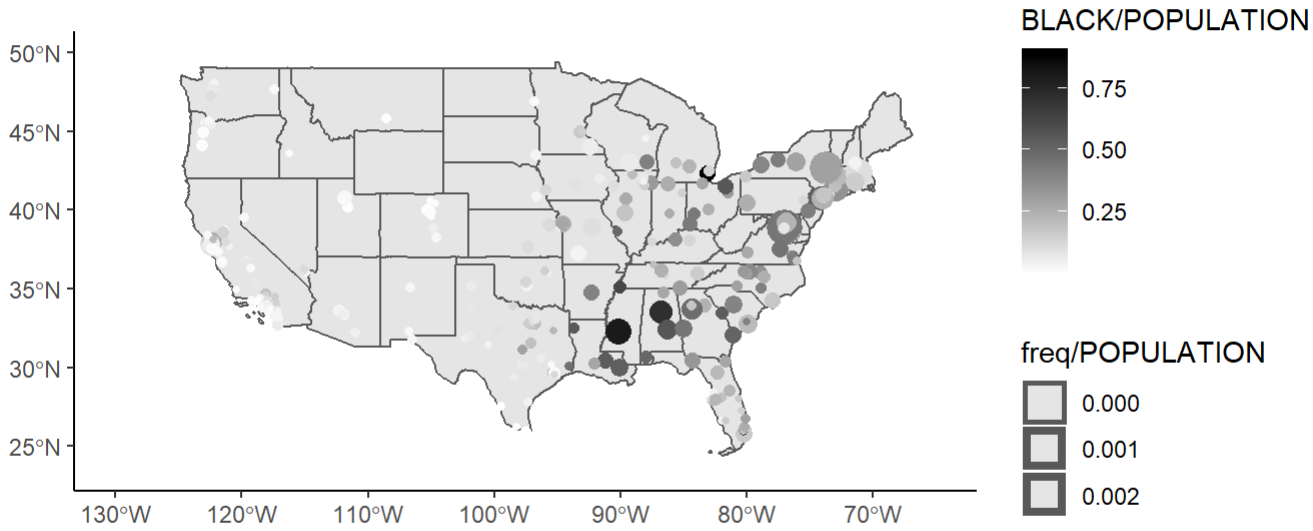
```
#med_age gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq / POPULAT
ION, color = MED_AGE)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_continuo
us(trans = "pseudo_log") + scale_color_gradientn(colors = c("green", "purple"), trans = "pseudo_
log") + theme_classic()
```



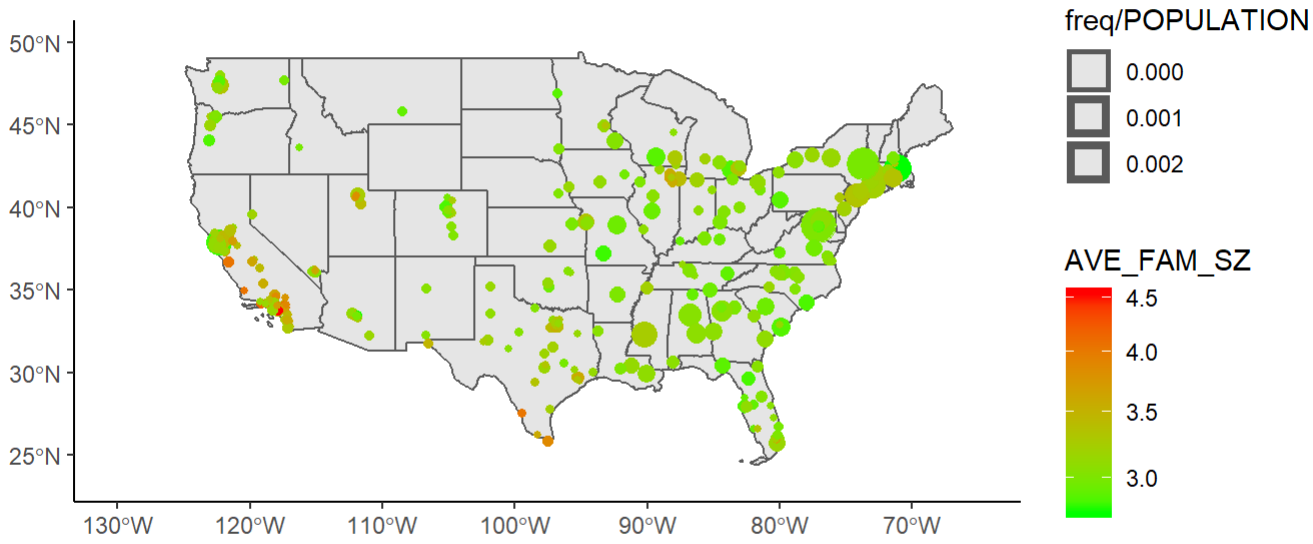

```
#white gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq / POPULAT
ION, color = WHITE/POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size
_continuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("black", "white"), trans =
"pseudo_log") + theme_classic()
```



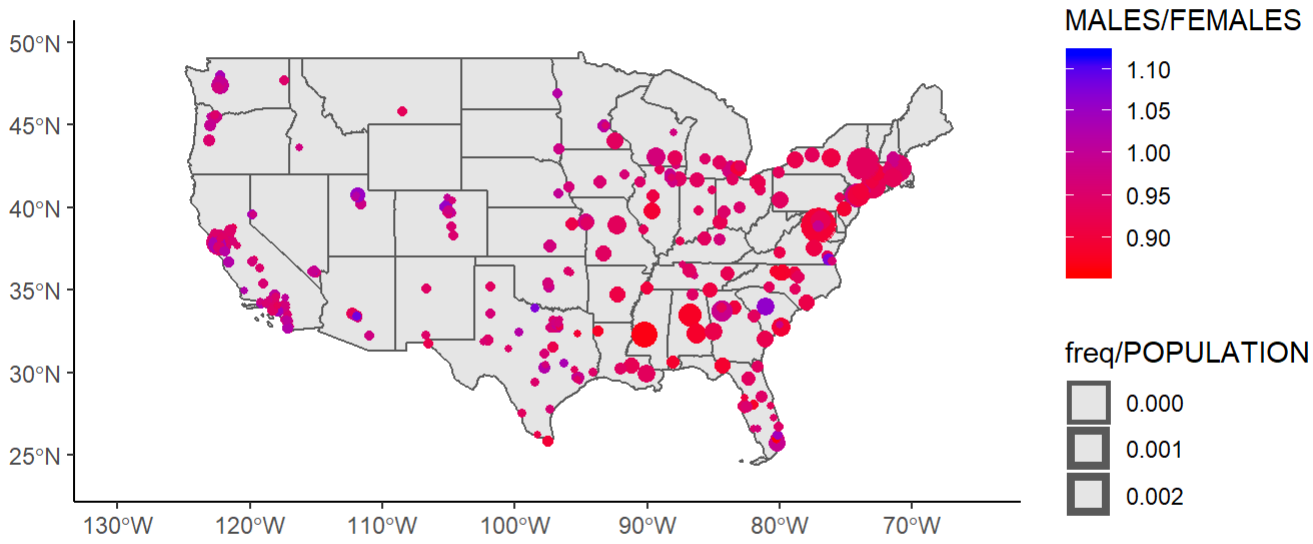

```
#black gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq / POPULAT
ION, color = BLACK/POPULATION)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size
_continuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("white", "black"), trans =
"pseudo_log") + theme_classic()
```




```
#avg_fam_sz gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq / POPULAT
ION, color = AVE_FAM_SZ)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_conti
nuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("green", "red"), trans = "pseudo_
log") + theme_classic()
```




```
#male - female gradient
ggplot() + geom_sf(data = state_map) + geom_sf(data = df_plus_spatial, aes(size = freq / POPULAT
ION, color = MALES/FEMALES)) + coord_sf(xlim = c(-130, -65), ylim = c(23.5, 50)) + scale_size_co
ntinuous(trans = "pseudo_log") + scale_color_gradientn(colors = c("red", "blue"), trans = "pseud
o_log") + theme_classic()
```



When adjusting these numbers to a per capita scale, correlations definitely became more clear. Age and average household remained uncertain, so it is reasonable to claim that there isn't much of a relationship. Gender was very interesting. When adjusting per person, we see that all the major protest cites are very red. I would be interested to see if this relationship is due to city size and perhaps a greater female population in larger cities, or if more female cities actually tending to demonstrate more. With race, a similar result, especially on the WHITE/POPULATION map. We can see that all the biggest protest sites are dark grey or black. That being said, it could also be the bigger cities are simply more diverse.

Moving away from maps, another interesting piece of data to have is top cities on a smaller scale. You can pick out the main ones, like Washington D.C. and New York, but after that it becomes hard to distinguish. Looking at the following table, there are the top 25 cities by per-capita protest frequency.

```
top_ten_freq_per_capita <- head(df_plus_spatial[order(-df_plus_spatial$freq / df_plus_spatial$PO
PULATION), ], n = 25)


top_ten_freq_per_capita <- data.frame(top_ten_freq_per_capita$NAME, top_ten_freq_per_capita$ST,
 top_ten_freq_per_capita$freq, top_ten_freq_per_capita$POPULATION, top_ten_freq_per_capita$freq
 / top_ten_freq_per_capita$POPULATION)

top_ten_freq_per_capita
```

```
##    top_ten_freq_per_capita.NAME top_ten_freq_per_capita.ST
## 1                    Washington                          DC
## 2                        Albany                          NY
## 3                     Cambridge                          MA
## 4                      Berkeley                          CA
## 5                       Jackson                          MS
## 6                     New Haven                          CT
## 7                        Newark                          NJ
## 8                    Birmingham                          AL
## 9                        Boston                          MA
## 10                     New York                          NY
## 11                     Hartford                          CT
## 12                      Atlanta                          GA
## 13                     Richmond                          CA
## 14                   Montgomery                          AL
## 15                     Columbia                          MD
## 16                     Syracuse                          NY
## 17                     Columbia                          MO
## 18                   Charleston                          SC
## 19                      Madison                          WI
## 20                   Providence                          RI
## 21                San Francisco                          CA
## 22                     Columbia                          SC
## 23                    Ann Arbor                          MI
## 24                         Kent                          WA
## 25                     Columbus                          GA
##    top_ten_freq_per_capita.freq top_ten_freq_per_capita.POPULATION
## 1                          1881                               688642
## 2                           214                               100753
## 3                           160                               116577
## 4                           141                               120662
## 5                           191                               169848
## 6                           138                               133467
## 7                           243                               284054
## 8                           164                               216500
## 9                           473                               674913
## 10                         5875                              8679888
## 11                           81                               127175
## 12                          269                               477371
## 13                           62                               111298
## 14                          115                               207148
## 15                           53                               106531
## 16                           64                               145627
## 17                           53                               122007
## 18                           62                               143224
## 19                          110                               258275
## 20                           77                               182386
## 21                          343                               878294
## 22                           53                               137578
## 23                           45                               123301
## 24                           48                               132665
## 25                           66                               193432
##    top_ten_freq_per_capita.freq.top_ten_freq_per_capita.POPULATION
```

```
## 1                                                    0.0027314628
## 2                                                    0.0021240062
## 3                                                    0.0013724834
## 4                                                    0.0011685535
## 5                                                    0.0011245349
## 6                                                    0.0010339635
## 7                                                    0.0008554711
## 8                                                    0.0007575058
## 9                                                    0.0007008311
## 10                                                   0.0006768521
## 11                                                   0.0006369176
## 12                                                   0.0005635030
## 13                                                   0.0005570630
## 14                                                   0.0005551586
## 15                                                   0.0004975078
## 16                                                   0.0004394789
## 17                                                   0.0004344013
## 18                                                   0.0004328883
## 19                                                   0.0004259026
## 20                                                   0.0004221815
## 21                                                   0.0003905298
## 22                                                   0.0003852360
## 23                                                   0.0003649605
## 24                                                   0.0003618136
## 25                                                   0.0003412052
```

These all make reasonable sense. D.C. is first, as usual, Albany and Berkeley, both notorious for demonstratins are 2 and 4. Massachusetts, one of the most progressive states, has both Cambridge and Boston in the top 10. Southern cities like Jackson and Birmingham due to the huge civil rights protests in them. One very interesting observation is that these are all large cities, but with per-capita measurements, population is taken out of the equation. This means it's true that individuals who live in larger cities are themselves more likely to protest, not just that more protests happen in larger cities due to population. Now, I'm sure there are more opportunities and it is likely easier to demonstrate, and that would be someone better than me at researching to look into, but there is definitely more counterculture the larger a city is.

ANSWERING QUESTIONS

Returning back to the motivating questions of the research:

- Do demographic or societal factors correlate with and influence frequency and location of demonstrations in the United States? After looking at all the results, I found that demographic groups really aren't as different as many people and the media like to believe. Any correlations that were found were weak or nonexistant.
- What is the relationship between age, gender, income, and race with likelihood to protest. Similar to the last question, correlations were fairly weak or nonexistant. Both age and household size (possibly representative of religion or income) were not visibly connected to likelohood to demonstrate. gender and race were more correlated, but I was unable to separate these numbers from city size, so there is room for doubt.
- What is the relationship between the presence of violence, arrests, and deaths of the protests and the protests' frequency? The most interesting analysis came from the violence analysis when separated by decade. Violent protests were certainly more rare, but also more indicative of the time. When looking at

peaceful demonstrations, it looked very similar to the original map, essentially just scaled down a small bit for every city. However, the violence maps very clearly described the concentrations of each decade's protests.

- How do the four decades covered differ in amount and type of protest? Following from the last answer, we can definitely see the difference in the four decades clearly in the violence data. There are certainly changes in collective action depending on events happening in the country, which makes sense, but the degree of separation was surprising. Especially regarding civil rights, the violence was concentrated almsot entirely in the South.

FINAL THOUGHTS

Collective action in the United States is a huge topic. I couldn't hope to cover all its intricacies with surface-level data visualization. The anlaysis done here is great to begin to get an idea for trends and patterns in protests. It accomplished the original goal of future prediction as well as I could have hoped for the methods used. The most reliable result was where violent protests occurred. It seems that violence occurs vastly more in areas where the given issue is more important or directly affects the citizens in the area. This may seem trivial, but it was interesting that the non-violent protests didn't change in the same way, so people do tend to become more violent when it matters more. Unfortunately, this method of prediction relies on knowledge of the event or issue, and would require much more analysis of "who cares about what".

Ideally, a researcher in the future could pick up where I left off and continue on the demographic size. I'm sure there are better datasets out there, and people who can properly incorperate them into this kind of analysis. Given more time, it would be possible to go city by city comparing very specific measures to get a good idea of what factors cause more protests (or at least heavily correlate). With that kind of analysis, it would be possible to get numerical answers instead of visual. While visualation is good for the passing observer, as I'm sure my analysis is, if the government or researchers really wanted to predict or obtain absolute results, numbers are better.

Going even deeper, if the data is out there, I would be very interested in smaller portions of the city. Does it matter if you live downtown or in the suburbs? Additionally, I'm sure this analysis already exists, what demographic changes do we see as we move through different parts of the city, especially in terms of age, income, and race. If correlations could be found within the city, the analysis would take on a another purpose. Theroretically, police forces as well as the media and journalists could adapt to known protest areas in order to be at the right place at the right time while politcal unrest is high.

As we can see, there are a huge number of ways this could go. My analysis shows with a reasonable degree of certainty that correlations, at least loose ones, do exist. While it would take a talented team and more time, I hope that this could be expanded upon to find even more interesting results.