U.S. Exploratory Data Analysis

Joseph Lavicka

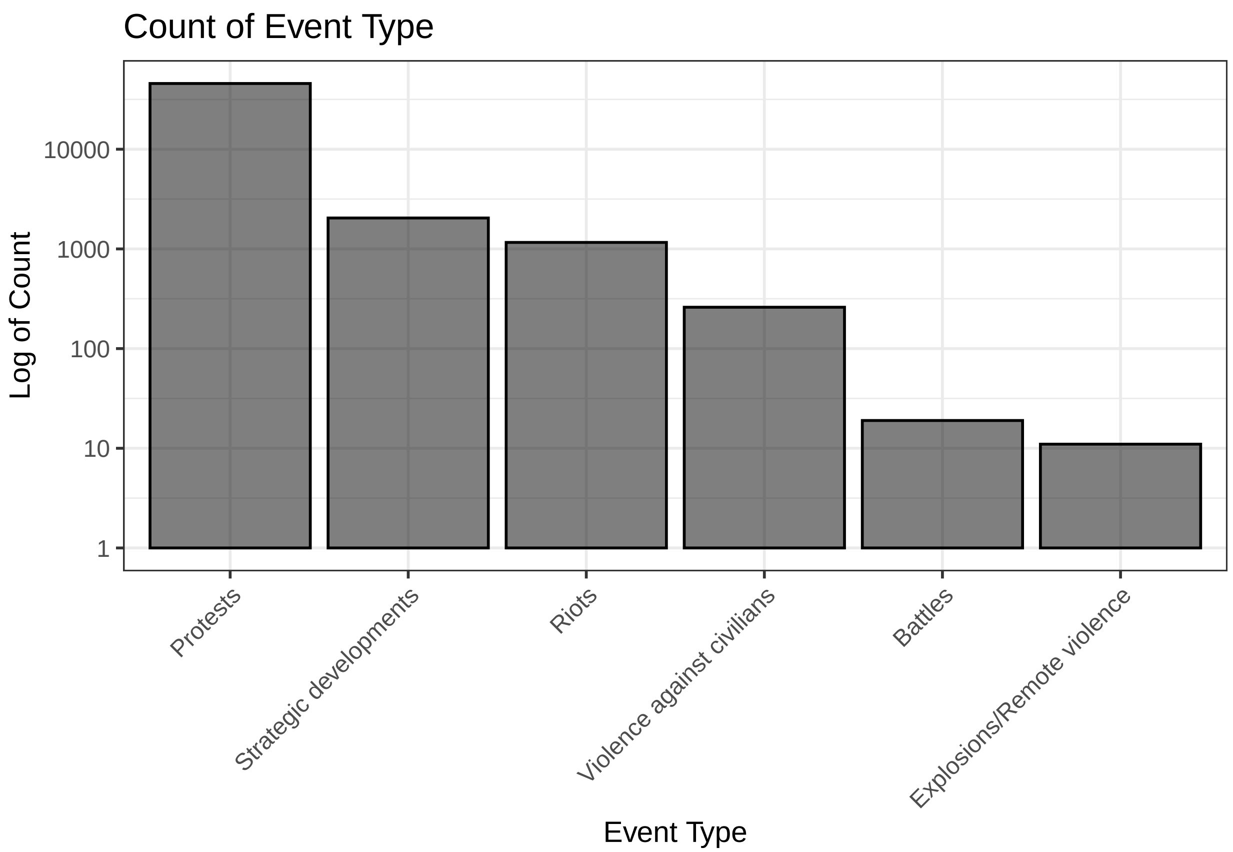
2023-01-06

packages

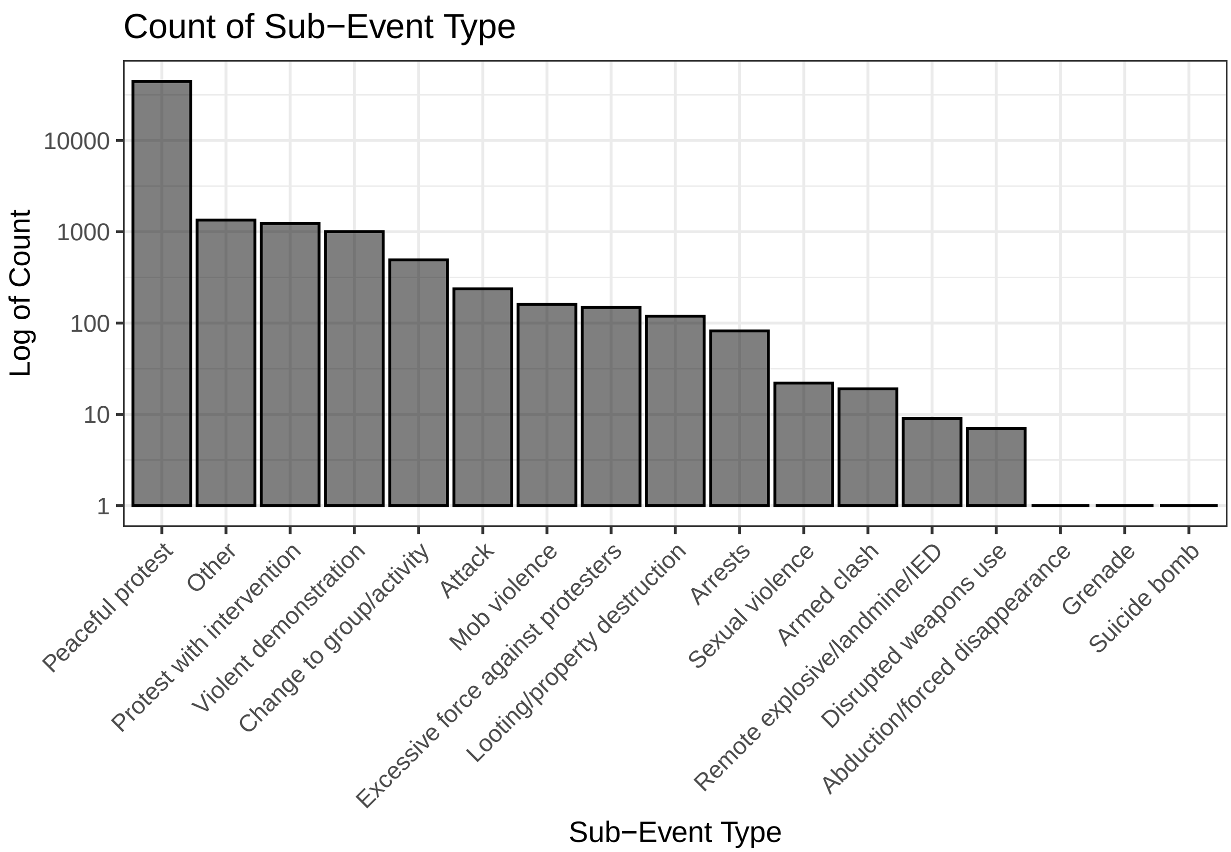
```
library(tidyverse)
```

explore ACLED dataset

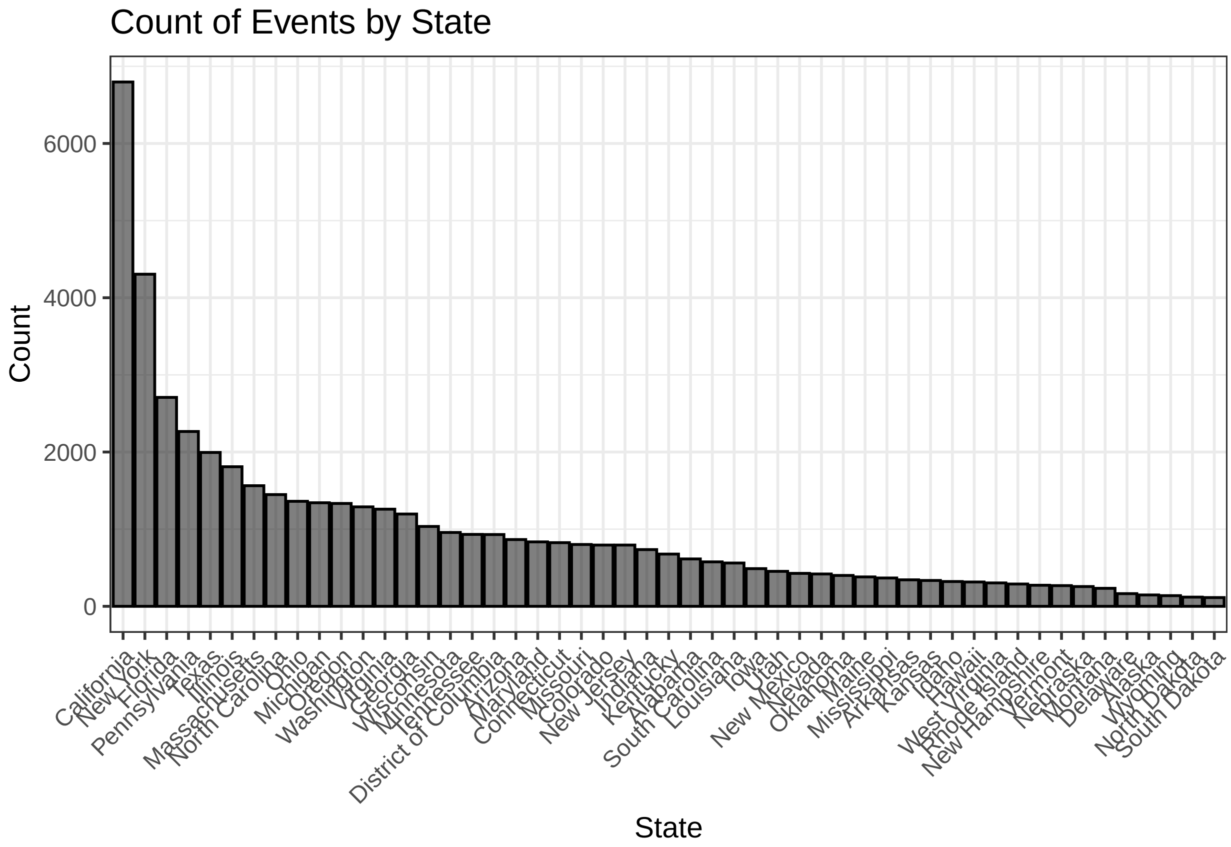
```
acled <- read_csv("../data/acled_us/2012-01-01-2022-11-30-United_States.csv")

acled %>%
  group_by(event_type) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  ggplot(aes(x = reorder(event_type, -n), y = n)) +
  geom_bar(stat = 'identity', fill = "black", color = "black", alpha = .5) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Event Type") +
  ylab("Log of Count") +
  ggtitle("Count of Event Type") +
  scale_y_log10()
```



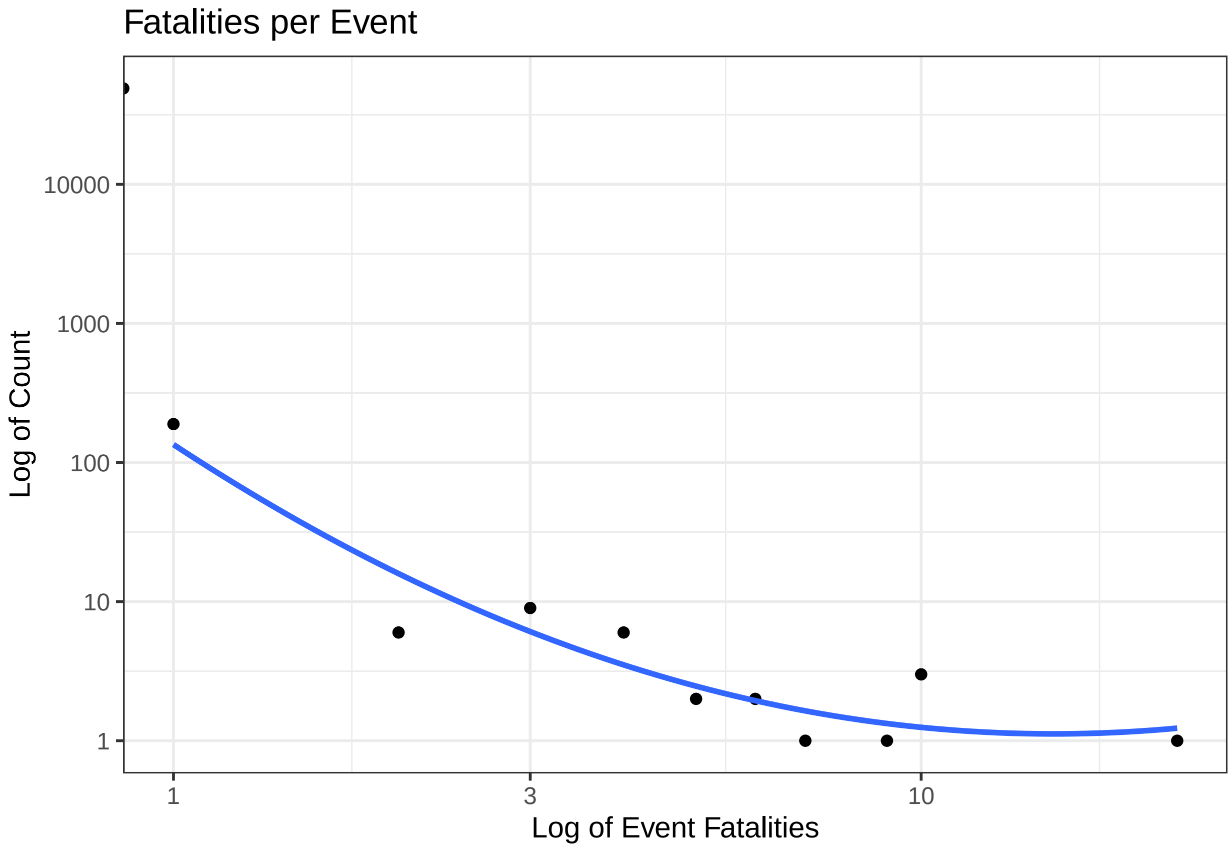
```
acled %>%
  group_by(sub_event_type) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  ggplot(aes(x = reorder(sub_event_type, -n), y = n)) +
  geom_bar(stat = 'identity', fill = "black", color = "black", alpha = .5) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Sub-Event Type") +
  ylab("Log of Count") +
  ggtitle("Count of Sub-Event Type") +
  scale_y_log10()
```



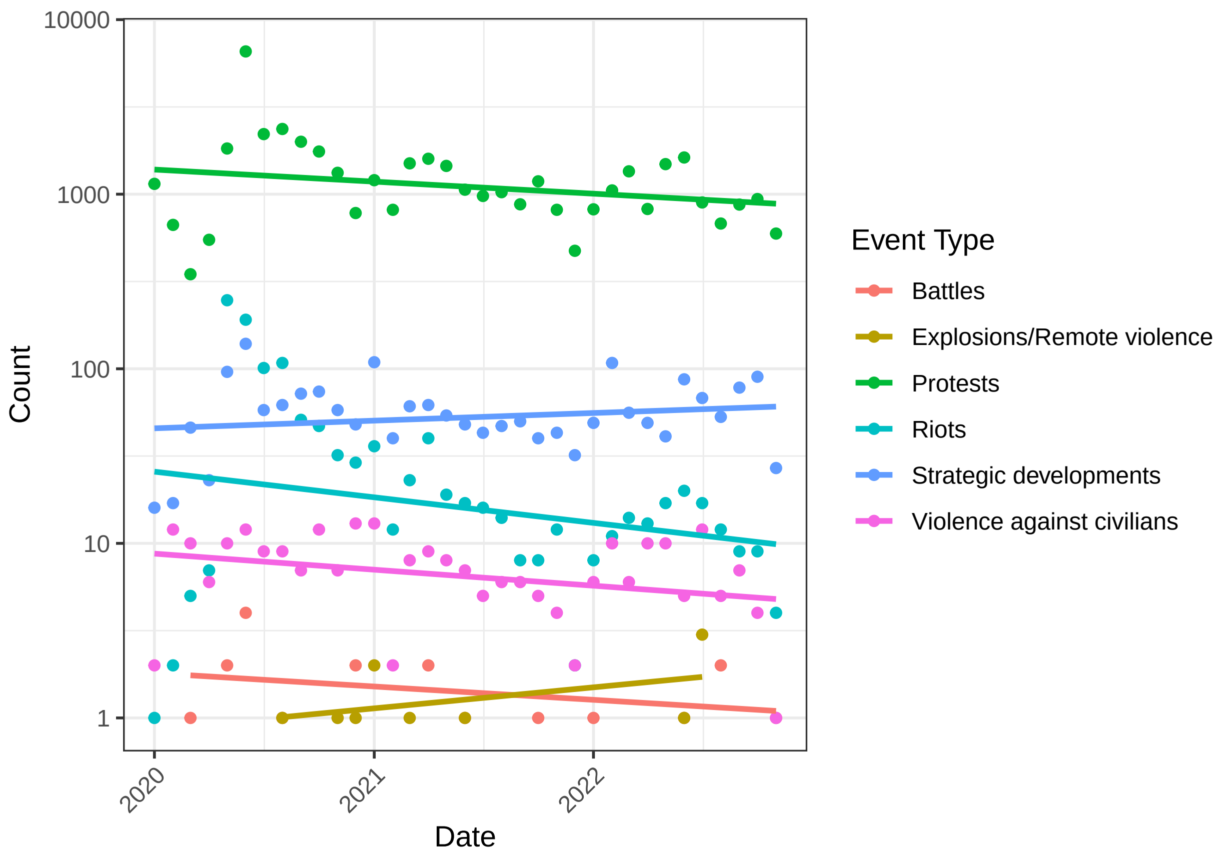
```
acled %>%
  group_by(admin1) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  ggplot(aes(x = reorder(admin1, -n), y = n)) +
  geom_bar(stat = 'identity', fill = "black", color = "black", alpha = .5) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("State") +
  ylab("Count") +
  ggtitle("Count of Events by State")
```



```r
acled %>%
  group_by(fatalities) %>%
  summarise(n = n()) %>%
  arrange(fatalities) %>%
  ggplot(aes(y = n, x = fatalities)) +
  geom_point() +
  scale_y_log10() +
  scale_x_log10() +
  stat_smooth(method = lm, formula = (y ~ poly(x, 2)), se = FALSE) +
  theme_bw() +
  xlab("Log of Event Fatalities") +
  ylab("Log of Count") +
  ggtitle("Fatalities per Event")
```



```
acled %>%
  select(event_date, event_type) %>%
  mutate(event_date = lubridate::dmy(event_date)) %>%
  group_by(date = lubridate::floor_date(event_date, 'month'), event_type) %>%
  summarise(n = n()) %>%
  ggplot(aes(x = date, y = n, color = event_type)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_log10() +
  scale_color_discrete(name = "Event Type") +
  ylab("Count") +
  xlab("Date")
```



```
acled %>%
  group_by(admin1, admin2) %>%
  summarise(n = n()) %>%
  arrange(desc(n))
```

```
## # A tibble: 1,882 x 3
## # Groups:   admin1 [51]
##    admin1               admin2                   n
##    <chr>                <chr>                <int>
##  1 California           Los Angeles           1566
##  2 New York             New York              1235
##  3 District of Columbia District of Columbia   930
##  4 Illinois             Cook                   894
##  5 California           San Diego              691
##  6 Oregon               Multnomah              581
##  7 California           San Francisco          515
##  8 California           Alameda                511
##  9 Massachusetts        Suffolk                485
## 10 Pennsylvania         Philadelphia           483
## # ... with 1,872 more rows
```

explore final dataset

```
final <- read_csv("../data/final.csv")

dim(final)
```

```
## [1] 110005     23
```

```
summary(final)
```

```
##       year         month            month_abv            region
##  Min.   :2020   Length:110005      Length:110005      Length:110005
##  1st Qu.:2020   Class :character   Class :character   Class :character
##  Median :2021   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2021
##  3rd Qu.:2022
##  Max.   :2022
##    FIPS_code        admin1           admin1_abv           admin2
##  Min.   : 1001   Length:110005      Length:110005      Length:110005
##  1st Qu.:18175   Class :character   Class :character   Class :character
##  Median :29175   Mode  :character   Mode  :character   Mode  :character
##  Mean   :30375
##  3rd Qu.:45081
##  Max.   :56045
##  admin2_full               n                n_bool            pop
##  Length:110005      Min.   :  0.0000   Min.   :0.0000   Min.   :      57
##  Class :character   1st Qu.:  0.0000   1st Qu.:0.0000   1st Qu.:   10829
##  Mode  :character   Median :  0.0000   Median :0.0000   Median :   25752
##                     Mean   :  0.4454   Mean   :0.1381   Mean   :  105549
##                     3rd Qu.:  0.0000   3rd Qu.:0.0000   3rd Qu.:   68397
##                     Max.   :138.0000   Max.   :1.0000   Max.   :10014009
##   less_than_hs    hs_diploma     some_college    pres_election
##  Min.   : 1.4   Min.   : 6.50   Min.   : 5.90   Min.   :0.00000
##  1st Qu.: 7.9   1st Qu.:29.30   1st Qu.:27.50   1st Qu.:0.00000
##  Median :11.2   Median :34.30   Median :31.00   Median :0.00000
##  Mean   :12.4   Mean   :33.93   Mean   :31.06   Mean   :0.02857
##  3rd Qu.:15.9   3rd Qu.:39.20   3rd Qu.:34.50   3rd Qu.:0.00000
##  Max.   :78.1   Max.   :55.00   Max.   :81.80   Max.   :1.00000
##   mid_election       unemp_rate       PCTPOVALL_      PCTPOV017_
##  Min.   :0.00000   Min.   : 0.900   Min.   : 3.00   Min.   : 2.60
##  1st Qu.:0.00000   1st Qu.: 3.800   1st Qu.: 9.90   1st Qu.:12.60
##  Median :0.00000   Median : 5.000   Median :12.80   Median :17.60
##  Mean   :0.02857   Mean   : 5.378   Mean   :13.73   Mean   :18.68
##  3rd Qu.:0.00000   3rd Qu.: 6.600   3rd Qu.:16.60   3rd Qu.:23.40
##  Max.   :1.00000   Max.   :22.800   Max.   :43.90   Max.   :59.70
##    MEDHHINC_         infl_all       infl_food
##  Min.   : 22901   Min.   :101.7   Min.   :102.6
##  1st Qu.: 47825   1st Qu.:105.1   1st Qu.:106.4
##  Median : 55152   Median :109.5   Median :109.5
##  Mean   : 57473   Mean   :110.8   Mean   :111.9
##  3rd Qu.: 64104   3rd Qu.:116.2   3rd Qu.:117.3
##  Max.   :160305   Max.   :126.1   Max.   :128.1
```

```
str(final)
```

```
## spc_tbl_ [110,005 x 23] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ year         : num [1:110005] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ month        : chr [1:110005] "January" "January" "January" "January" ...
##  $ month_abv    : chr [1:110005] "Jan" "Jan" "Jan" "Jan" ...
##  $ region       : chr [1:110005] "East South Central" "East South Central" "East South Central" "Eas
##  $ FIPS_code    : num [1:110005] 1001 1003 1005 1007 1009 ...
##  $ admin1       : chr [1:110005] "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ admin1_abv   : chr [1:110005] "AL" "AL" "AL" "AL" ...
##  $ admin2       : chr [1:110005] "Autauga" "Baldwin" "Barbour" "Bibb" ...
##  $ admin2_full  : chr [1:110005] "Autauga County" "Baldwin County" "Barbour County" "Bibb County" ..
##  $ n            : num [1:110005] 0 0 0 0 0 0 0 0 0 0 ...
##  $ n_bool       : num [1:110005] 0 0 0 0 0 0 0 0 0 0 ...
##  $ pop          : num [1:110005] 58805 231767 25223 22293 59134 ...
##  $ less_than_hs : num [1:110005] 11.3 9.5 25.3 19.1 17.2 25.1 13.6 14.9 17.4 17.2 ...
##  $ hs_diploma   : num [1:110005] 31.4 27.2 35.7 45.1 35.1 41.4 46.5 34.4 37.1 39.2 ...
##  $ some_college : num [1:110005] 29 31.4 27.4 24.5 34.5 23.3 23.9 31.8 31 30.7 ...
##  $ pres_election: num [1:110005] 0 0 0 0 0 0 0 0 0 0 ...
##  $ mid_election : num [1:110005] 0 0 0 0 0 0 0 0 0 0 ...
##  $ unemp_rate   : num [1:110005] 5.4 6.2 7.8 7.3 4.6 6 9.6 7.8 7.5 5.1 ...
##  $ PCTPOVALL_   : num [1:110005] 11.2 8.9 25.5 17.8 13.1 30.8 20.6 14.5 16.3 14.7 ...
##  $ PCTPOV017_   : num [1:110005] 14.9 12.4 37.5 21.9 18.9 38.7 30.8 16.7 26.2 23.3 ...
##  $ MEDHHINC_    : num [1:110005] 67565 71135 38866 50907 55203 ...
##  $ infl_all     : num [1:110005] 103 103 103 103 103 ...
##  $ infl_food    : num [1:110005] 103 103 103 103 103 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   year = col_double(),
##   ..   month = col_character(),
##   ..   month_abv = col_character(),
##   ..   region = col_character(),
##   ..   FIPS_code = col_double(),
##   ..   admin1 = col_character(),
##   ..   admin1_abv = col_character(),
##   ..   admin2 = col_character(),
##   ..   admin2_full = col_character(),
##   ..   n = col_double(),
##   ..   n_bool = col_double(),
##   ..   pop = col_double(),
##   ..   less_than_hs = col_double(),
##   ..   hs_diploma = col_double(),
##   ..   some_college = col_double(),
##   ..   pres_election = col_double(),
##   ..   mid_election = col_double(),
##   ..   unemp_rate = col_double(),
##   ..   PCTPOVALL_ = col_double(),
##   ..   PCTPOV017_ = col_double(),
##   ..   MEDHHINC_ = col_double(),
##   ..   infl_all = col_double(),
##   ..   infl_food = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
sum(is.na(final))
```

```
## [1] 0
```