Stats 140XP Group Project Data Transformation Code

Charlie Hoose, JJ Svenson, Kendall Keely, Blair Warren, Mackenzie Lindhold, Claire Nabours

2025-12-02

We are looking at drone strike data of US drones on Yemen, Pakistan, and Somalia. The following code is all of the data preprocessing and manipulation used to create the final dataset used for EDA, visualizations, and statistical analysis.

```r
#install.packages("readxl")
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(ggplot2)
library(stringr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v tibble  3.2.1
## v purrr   1.0.2     v tidyr   1.3.1
## v readr   2.1.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Somalia dataset

```
excel_sheets("us-strikes-in-somalia-2007-to-present.xlsx")
```

```
## [1] "Notes"                             "Year-by-year summary"
## [3] "All US actions"                    "Chart US strikes"
## [5] "Chart US strike casualties"        "Chart US strikes and people kil"
## [7] "Form responses 1"
```

```
df_somalia <- read_excel("us-strikes-in-somalia-2007-to-present.xlsx", sheet = 3)
head(df_somalia)
```

```
## # A tibble: 6 x 19
##   ‘Strike ID‘ Date                Location  Confirmed/\r\npossib~1 ‘Air strike?‘
##   <chr>       <dttm>              <chr>     <chr>                          <dbl>
## 1 SOM001      2007-01-07 00:00:00 Ras Kamb~ Confirmed                          1
## 2 SOM002      2007-01-09 00:00:00 Hayo, Ga~ Confirmed                          1
## 3 SOM003      2007-01-09 00:00:00 Hayo      Possible                           1
## 4 SOM004      2007-01-23 00:00:00 Waldena   Confirmed                          1
## 5 SOM005      2007-06-01 00:00:00 Bargal    Confirmed                          0
## 6 SOM006      2008-03-03 00:00:00 Dhobley   Confirmed                          0
## # i abbreviated name: 1: ‘Confirmed/\r\npossible US strike‘
## # i 14 more variables: ‘Drone strike‘ <dbl>, ‘Strike type‘ <chr>,
## #   ‘Minimum strikes‘ <dbl>, ‘Maximum strikes‘ <dbl>,
## #   ‘Minimum people killed‘ <dbl>, ‘Maximum people killed‘ <dbl>,
## #   ‘Minimum civilians killed‘ <dbl>, ‘Maximum civilians killed‘ <dbl>,
## #   ‘Minimum children killed‘ <dbl>, ‘Maximum children killed‘ <dbl>,
## #   ‘Minimum people injured‘ <dbl>, ‘Maximum people injured‘ <dbl>, ...
```

Yemen dataset

```
excel_sheets("us-strikes-in-yemen-2002-to-present.xlsx")
```

```
## [1] "Notes"                  "Year-by-year summary"   "All US actions"
## [4] "Chart US strikes"       "Chart people killed"    "Chart civilians killed"
```

```
df_yemen <- read_excel("us-strikes-in-yemen-2002-to-present.xlsx", sheet = 3)
head(df_yemen)
```

```
## # A tibble: 6 x 20
##   ‘Strike ID‘ Date                Location   Province ‘Type of attack‘
##   <chr>       <dttm>              <chr>      <chr>    <chr>
## 1 YEM001      2002-11-03 00:00:00 Unknown    Marib    "Drone strike"
## 2 YEM002      2009-12-17 00:00:00 Al Majala  Abyan    "Cruise missile strike"
## 3 YEM003      2009-12-17 00:00:00 Arhab      Sanaa    "US-Yemen ground operatio~
## 4 YEM004      2009-12-24 00:00:00 Rafd       Shabwa   "Airstrike\r\nPossible cr~
## 5 YEM005      2010-01-12 00:00:00 Unknown    Shabwa   "Yemen ground operation\r~
## 6 YEM006      2010-01-15 00:00:00 Al Ajashir Saada    "Airstrike"
## # i 15 more variables: ‘Confirmed/\r\npossible US attack?‘ <chr>,
## #   ‘Air operation?‘ <dbl>, ‘Drone strike‘ <dbl>, ‘Minimum strikes‘ <dbl>,
## #   ‘Maximum strikes‘ <dbl>, ‘Minimum people killed‘ <dbl>,
## #   ‘Maximum people killed‘ <dbl>, ‘Minimum civilians killed‘ <dbl>,
## #   ‘Maximum civilians killed‘ <dbl>, ‘Minimum children killed‘ <dbl>,
## #   ‘Maximum children killed‘ <dbl>, ‘Minimum people injured‘ <dbl>,
## #   ‘Maximum people injured‘ <dbl>, ‘Strike link‘ <chr>, extra <chr>
```

Pakistan dataset

```
excel_sheets("cia-and-us-military-drone-strikes-in-pakistan-2004-to-present.xlsx")
```

```
## [1] "Notes"                             "Summary tables and casualty rat"
## [3] "Drone strikes data"                "Chart Drone strikes"
## [5] "Chart people killed"               "Chart civilians killed"
## [7] "Chart US strikes and minimum pe" "Chart summary figures, by Presi"
## [9] "Chart casualty rates"
```

```
df_pakistan <- read_excel("cia-and-us-military-drone-strikes-in-pakistan-2004-to-present.xlsx", sheet =
head(df_pakistan)
```

```
## # A tibble: 6 x 15
##   ‘Strike ID‘ Date                Location  Area          Minimum people kille~1
##   <chr>       <dttm>              <chr>     <chr>                          <dbl>
## 1 B1          2004-06-17 00:00:00 Wana      South Waziri~                      6
## 2 B2          2005-05-08 00:00:00 Toorikhel North Waziri~                      2
## 3 B3          2005-11-05 00:00:00 Mosaki    North Waziri~                      8
## 4 B4          2005-12-01 00:00:00 Haisori   North Waziri~                      6
## 5 B5          2006-01-13 00:00:00 Damadola  Bajaur Agency                     13
## 6 B6          2006-10-30 00:00:00 Chenegai  Bajaur Agency                     81
## # i abbreviated name: 1: ‘Minimum people killed‘
## # i 10 more variables: ‘Maximum people killed‘ <dbl>,
## #   ‘Minimum civilians killed‘ <dbl>, ‘Maximum civilians killed‘ <dbl>,
## #   ‘Minimum children killed‘ <dbl>, ‘Maximum children killed‘ <dbl>,
## #   ‘Minimum people injured‘ <dbl>, ‘Maximum people injured‘ <dbl>,
## #   ‘Strike link‘ <chr>, extra <chr>, Index <chr>
```

Now the political dataset, data is from: https://history.house.gov/Institution/Presidents-Coinciding/Party-Government/

```
df_party <- read_excel("political_swing.xlsx")
head(df_party)
```

```
## # A tibble: 6 x 5
##    Year ‘House Majority Party‘ ‘Senate Majority Party‘ ‘Presidential Party‘
##   <dbl> <chr>                  <chr>                   <chr>
## 1  2000 Republican             Republican              Republican
## 2  2001 Republican             Democrat                Republican
## 3  2002 Republican             Democrat                Republican
## 4  2003 Republican             Republican              Republican
## 5  2004 Republican             Republican              Republican
## 6  2005 Republican             Republican              Republican
## # i 1 more variable: ‘Unified?‘ <chr>
```

At this point, we can join the political dataset onto each of the drone datasets, so we have the political parties in control of the House, Senate, and Presidency during each strike.

```
# pakistan dataset
df_pakistan <- df_pakistan %>%
  mutate(Year = year(Date)) %>%
  left_join(df_party, by = "Year")
head(df_pakistan,5)
```

```
## # A tibble: 5 x 20
##   ‘Strike ID‘ Date                Location  Area          Minimum people kille~1
##   <chr>       <dttm>              <chr>     <chr>                           <dbl>
## 1 B1          2004-06-17 00:00:00 Wana      South Waziri~                       6
## 2 B2          2005-05-08 00:00:00 Toorikhel North Waziri~                       2
## 3 B3          2005-11-05 00:00:00 Mosaki    North Waziri~                       8
## 4 B4          2005-12-01 00:00:00 Haisori   North Waziri~                       6
## 5 B5          2006-01-13 00:00:00 Damadola  Bajaur Agency                      13
## # i abbreviated name: 1: ‘Minimum people killed‘
## # i 15 more variables: ‘Maximum people killed‘ <dbl>,
## #   ‘Minimum civilians killed‘ <dbl>, ‘Maximum civilians killed‘ <dbl>,
## #   ‘Minimum children killed‘ <dbl>, ‘Maximum children killed‘ <dbl>,
## #   ‘Minimum people injured‘ <dbl>, ‘Maximum people injured‘ <dbl>,
## #   ‘Strike link‘ <chr>, extra <chr>, Index <chr>, Year <dbl>,
## #   ‘House Majority Party‘ <chr>, ‘Senate Majority Party‘ <chr>, ...
```

```
# yemen dataset
df_yemen <- df_yemen %>%
  mutate(Year = year(Date)) %>%
  left_join(df_party, by = "Year")
head(df_yemen,5)
```

```
## # A tibble: 5 x 25
##   ‘Strike ID‘ Date                Location  Province ‘Type of attack‘
##   <chr>       <dttm>              <chr>     <chr>    <chr>
## 1 YEM001      2002-11-03 00:00:00 Unknown   Marib    "Drone strike"
## 2 YEM002      2009-12-17 00:00:00 Al Majala Abyan    "Cruise missile strike"
## 3 YEM003      2009-12-17 00:00:00 Arhab     Sanaa    "US-Yemen ground operation~
## 4 YEM004      2009-12-24 00:00:00 Rafd      Shabwa   "Airstrike\r\nPossible cru~
## 5 YEM005      2010-01-12 00:00:00 Unknown   Shabwa   "Yemen ground operation\r\~
## # i 20 more variables: ‘Confirmed/\r\npossible US attack?‘ <chr>,
## #   ‘Air operation?‘ <dbl>, ‘Drone strike‘ <dbl>, ‘Minimum strikes‘ <dbl>,
## #   ‘Maximum strikes‘ <dbl>, ‘Minimum people killed‘ <dbl>,
## #   ‘Maximum people killed‘ <dbl>, ‘Minimum civilians killed‘ <dbl>,
## #   ‘Maximum civilians killed‘ <dbl>, ‘Minimum children killed‘ <dbl>,
## #   ‘Maximum children killed‘ <dbl>, ‘Minimum people injured‘ <dbl>,
## #   ‘Maximum people injured‘ <dbl>, ‘Strike link‘ <chr>, extra <chr>, ...
```

```
# somalia dataset
df_somalia <- df_somalia %>%
  mutate(Year = year(Date)) %>%
  left_join(df_party, by = "Year")
head(df_somalia,5)
```

```
## # A tibble: 5 x 24
##   ‘Strike ID‘ Date                Location  Confirmed/\r\npossib~1 ‘Air strike?‘
##   <chr>       <dttm>              <chr>     <chr>                          <dbl>
## 1 SOM001      2007-01-07 00:00:00 Ras Kamb~ Confirmed                          1
## 2 SOM002      2007-01-09 00:00:00 Hayo, Ga~ Confirmed                          1
## 3 SOM003      2007-01-09 00:00:00 Hayo      Possible                           1
## 4 SOM004      2007-01-23 00:00:00 Waldena   Confirmed                          1
## 5 SOM005      2007-06-01 00:00:00 Bargal    Confirmed                          0
## # i abbreviated name: 1: ‘Confirmed/\r\npossible US strike‘
## # i 19 more variables: ‘Drone strike‘ <dbl>, ‘Strike type‘ <chr>,
## #   ‘Minimum strikes‘ <dbl>, ‘Maximum strikes‘ <dbl>,
## #   ‘Minimum people killed‘ <dbl>, ‘Maximum people killed‘ <dbl>,
## #   ‘Minimum civilians killed‘ <dbl>, ‘Maximum civilians killed‘ <dbl>,
## #   ‘Minimum children killed‘ <dbl>, ‘Maximum children killed‘ <dbl>,
## #   ‘Minimum people injured‘ <dbl>, ‘Maximum people injured‘ <dbl>, ...
```

Now we can convert these to csvs for later use in our project

```
write.csv(df_pakistan, "df_pakistan.csv", row.names = FALSE)
write.csv(df_yemen, "df_yemen.csv", row.names = FALSE)
write.csv(df_somalia, "df_somalia.csv", row.names = FALSE)
```

We can continue the data pre-processing step by combining the separate datasets into one grand dataset:

```
df_yemen <- df_yemen %>% mutate(Country = "Yemen")
df_pakistan <- df_pakistan %>% mutate(Country = "Pakistan")
df_somalia <- df_somalia %>% mutate(Country = "Somalia")

# To combine the 3 tables into 1, we need to standardize the column names between the datasets so they
df_yemen2 <- df_yemen %>% rename(
  Area = Province,
  `Confirmed/Possible US Strike` = `Confirmed/\r\npossible US attack?`,
  `Type of Attack` = `Type of attack`,
  `Air strike?` = `Air operation?`,
  `Min Strikes` = `Minimum strikes`,
  `Max Strikes` = `Maximum strikes`,
  `Min People Killed` = `Minimum people killed`,
  `Max People Killed` = `Maximum people killed`,
  `Min Civilians Killed` = `Minimum civilians killed`,
  `Max Civilians Killed` = `Maximum civilians killed`,
  `Min Children Killed` = `Minimum children killed`,
  `Max Children Killed` = `Maximum children killed`,
  `Min People Injured` = `Minimum people injured`,
  `Max People Injured` = `Maximum people injured`,
  `Strike Link` = `Strike link`
)

df_pakistan2 <- df_pakistan %>% rename(
  `Min People Killed` = `Minimum people killed`,
  `Max People Killed` = `Maximum people killed`,
  `Min Civilians Killed` = `Minimum civilians killed`,
  `Max Civilians Killed` = `Maximum civilians killed`,
  `Min Children Killed` = `Minimum children killed`,
  `Max Children Killed` = `Maximum children killed`,
  `Min People Injured` = `Minimum people injured`,
  `Max People Injured` = `Maximum people injured`,
  `Strike Link` = `Strike link`
)

df_somalia2 <- df_somalia %>% rename(
  `Confirmed/Possible US Strike` = `Confirmed/\r\npossible US strike`,
  `Type of Attack` = `Strike type`,
  `Min Strikes` = `Minimum strikes`,
  `Max Strikes` = `Maximum strikes`,
  `Min People Killed` = `Minimum people killed`,
  `Max People Killed` = `Maximum people killed`,
  `Min Civilians Killed` = `Minimum civilians killed`,
  `Max Civilians Killed` = `Maximum civilians killed`,
  `Min Children Killed` = `Minimum children killed`,
  `Max Children Killed` = `Maximum children killed`,
  `Min People Injured` = `Minimum people injured`,
  `Max People Injured` = `Maximum people injured`,
  `Strike Link` = `Strike link`
)


combined_df <- bind_rows(df_yemen2, df_pakistan2, df_somalia2)
combined_df <- combined_df %>% select(Country, everything()) # to get country in first row since that i
head(combined_df)
```

```
## # A tibble: 6 x 27
##   Country ‘Strike ID‘ Date                Location   Area   ‘Type of Attack‘
##   <chr>   <chr>       <dttm>              <chr>      <chr>  <chr>
## 1 Yemen   YEM001      2002-11-03 00:00:00 Unknown    Marib  "Drone strike"
## 2 Yemen   YEM002      2009-12-17 00:00:00 Al Majala  Abyan  "Cruise missile str~
## 3 Yemen   YEM003      2009-12-17 00:00:00 Arhab      Sanaa  "US-Yemen ground op~
## 4 Yemen   YEM004      2009-12-24 00:00:00 Rafd       Shabwa "Airstrike\r\nPossi~
## 5 Yemen   YEM005      2010-01-12 00:00:00 Unknown    Shabwa "Yemen ground opera~
## 6 Yemen   YEM006      2010-01-15 00:00:00 Al Ajashir Saada  "Airstrike"
## # i 21 more variables: ‘Confirmed/Possible US Strike‘ <chr>,
## #   ‘Air strike?‘ <dbl>, ‘Drone strike‘ <dbl>, ‘Min Strikes‘ <dbl>,
## #   ‘Max Strikes‘ <dbl>, ‘Min People Killed‘ <dbl>, ‘Max People Killed‘ <dbl>,
## #   ‘Min Civilians Killed‘ <dbl>, ‘Max Civilians Killed‘ <dbl>,
## #   ‘Min Children Killed‘ <dbl>, ‘Max Children Killed‘ <dbl>,
## #   ‘Min People Injured‘ <dbl>, ‘Max People Injured‘ <dbl>,
## #   ‘Strike Link‘ <chr>, extra <chr>, Year <dbl>, ...
```

```
#write.csv(combined_df, "df_combined.csv", row.names = FALSE)
# since we rewrite over this later, we can comment out this write.csv
```

Continuing, we can do some error handling and erroneous data checking. For max and min strikes, there are some rows where max - min is negative, which obviously means the data is messed up. Furthermore, for max and min people injured, there is a negative instance of max - min. Because there are so few cases for this, we can just remove the data without worrying about losing the shape of our data.

```r
# see the distribution of max - min to identify negatives
table(combined_df$`Max Strikes`-combined_df$`Min Strikes`)
```

```
##
##  -5  -2   0   1   2   3   4   5   6
##   1   1 487  14   6   2   1   1   1
```

```r
table(combined_df[,19]-combined_df[,18])
```

```
## Max People Injured
##  -1   0   1   2   3   4   5   6   7   8   9  10  11  12  13  20  22  25  28  41
##   1 744  62  47  22  16  17   8   4   4   3   5   1   1   3   1   1   1   1   1
##  55
##   1
```

```r
# remove the incorrect rows
rows_to_remove <- which((combined_df[,19] - combined_df[,18]) < 0)
combined_df[rows_to_remove, ]
```

```
## # A tibble: 1 x 27
##   Country  ‘Strike ID‘ Date                Location      Area   ‘Type of Attack‘
##   <chr>    <chr>       <dttm>              <chr>         <chr>  <chr>
## 1 Pakistan Ob193       2011-02-24 00:00:00 Mohammad Khel North~ <NA>
## # i 21 more variables: ‘Confirmed/Possible US Strike‘ <chr>,
## #   ‘Air strike?‘ <dbl>, ‘Drone strike‘ <dbl>, ‘Min Strikes‘ <dbl>,
## #   ‘Max Strikes‘ <dbl>, ‘Min People Killed‘ <dbl>, ‘Max People Killed‘ <dbl>,
## #   ‘Min Civilians Killed‘ <dbl>, ‘Max Civilians Killed‘ <dbl>,
## #   ‘Min Children Killed‘ <dbl>, ‘Max Children Killed‘ <dbl>,
## #   ‘Min People Injured‘ <dbl>, ‘Max People Injured‘ <dbl>,
## #   ‘Strike Link‘ <chr>, extra <chr>, Year <dbl>, ...
```

```r
combined_df <- combined_df[-rows_to_remove, ]

rows_to_remove_2 <- which((combined_df$`Max Strikes`-combined_df$`Min Strikes`) < 0)
combined_df[rows_to_remove_2,]
```

```
## # A tibble: 2 x 27
##   Country ‘Strike ID‘ Date                Location Area         ‘Type of Attack‘
##   <chr>   <chr>       <dttm>              <chr>    <chr>        <chr>
## 1 Yemen   YEM266      2017-03-06 00:00:00 Unknown  Multiple pr~ US air or drone~
## 2 Yemen   YEM279      2017-09-16 00:00:00 Unknown  Unknown      US air or drone~
## # i 21 more variables: ‘Confirmed/Possible US Strike‘ <chr>,
## #   ‘Air strike?‘ <dbl>, ‘Drone strike‘ <dbl>, ‘Min Strikes‘ <dbl>,
## #   ‘Max Strikes‘ <dbl>, ‘Min People Killed‘ <dbl>, ‘Max People Killed‘ <dbl>,
## #   ‘Min Civilians Killed‘ <dbl>, ‘Max Civilians Killed‘ <dbl>,
## #   ‘Min Children Killed‘ <dbl>, ‘Max Children Killed‘ <dbl>,
## #   ‘Min People Injured‘ <dbl>, ‘Max People Injured‘ <dbl>,
## #   ‘Strike Link‘ <chr>, extra <chr>, Year <dbl>, ...
```

```
combined_df <- combined_df[-rows_to_remove_2, ]
```

To continue our data processing, we can create some new columns for better analysis later on. We can make a middle column between min and max for many of the values, like finding the middle between min and max number of strikes. We will use the median for this value, although technically the mean would be the same since n=2.

```
combined_df$`Med Strikes` <- (combined_df$`Min Strikes` + combined_df$`Max Strikes`) / 2

combined_df$`Med People Killed` <- (combined_df$`Min People Killed` + combined_df$`Max People Killed`)

combined_df$`Med Civilians Killed` <- (combined_df$`Min Civilians Killed` + combined_df$`Max Civilians

combined_df$`Med Children Killed` <- (combined_df$`Min Children Killed` + combined_df$`Max Children Kil

combined_df$`Med People Injured` <- (combined_df$`Min People Injured` + combined_df$`Max People Injured

head(combined_df)
```

```
## # A tibble: 6 x 32
##   Country ‘Strike ID‘ Date                Location   Area   ‘Type of Attack‘
##   <chr>   <chr>       <dttm>              <chr>      <chr>  <chr>
## 1 Yemen   YEM001      2002-11-03 00:00:00 Unknown    Marib  "Drone strike"
## 2 Yemen   YEM002      2009-12-17 00:00:00 Al Majala  Abyan  "Cruise missile str~
## 3 Yemen   YEM003      2009-12-17 00:00:00 Arhab      Sanaa  "US-Yemen ground op~
## 4 Yemen   YEM004      2009-12-24 00:00:00 Rafd       Shabwa "Airstrike\r\nPossi~
## 5 Yemen   YEM005      2010-01-12 00:00:00 Unknown    Shabwa "Yemen ground opera~
## 6 Yemen   YEM006      2010-01-15 00:00:00 Al Ajashir Saada  "Airstrike"
## # i 26 more variables: ‘Confirmed/Possible US Strike‘ <chr>,
## #   ‘Air strike?‘ <dbl>, ‘Drone strike‘ <dbl>, ‘Min Strikes‘ <dbl>,
## #   ‘Max Strikes‘ <dbl>, ‘Min People Killed‘ <dbl>, ‘Max People Killed‘ <dbl>,
## #   ‘Min Civilians Killed‘ <dbl>, ‘Max Civilians Killed‘ <dbl>,
## #   ‘Min Children Killed‘ <dbl>, ‘Max Children Killed‘ <dbl>,
## #   ‘Min People Injured‘ <dbl>, ‘Max People Injured‘ <dbl>,
## #   ‘Strike Link‘ <chr>, extra <chr>, Year <dbl>, ...
```

Although we already have a column examining if the majority party in control of the Senate and House of Representatives matches that of the presidency, it also might be nice to see if there is unification between the House and Senate. We can create that column below.

```
combined_df$`Legislative Unified?` <- ifelse(
  combined_df$`House Majority Party` == combined_df$`Senate Majority Party`,
  combined_df$`House Majority Party`, "No")
head(combined_df)
```

```
## # A tibble: 6 x 33
##   Country ‘Strike ID‘ Date                Location   Area   ‘Type of Attack‘
##   <chr>   <chr>       <dttm>              <chr>      <chr>  <chr>
## 1 Yemen   YEM001      2002-11-03 00:00:00 Unknown    Marib  "Drone strike"
## 2 Yemen   YEM002      2009-12-17 00:00:00 Al Majala  Abyan  "Cruise missile str~
## 3 Yemen   YEM003      2009-12-17 00:00:00 Arhab      Sanaa  "US-Yemen ground op~
## 4 Yemen   YEM004      2009-12-24 00:00:00 Rafd       Shabwa "Airstrike\r\nPossi~
## 5 Yemen   YEM005      2010-01-12 00:00:00 Unknown    Shabwa "Yemen ground opera~
## 6 Yemen   YEM006      2010-01-15 00:00:00 Al Ajashir Saada  "Airstrike"
## # i 27 more variables: ‘Confirmed/Possible US Strike‘ <chr>,
## #   ‘Air strike?‘ <dbl>, ‘Drone strike‘ <dbl>, ‘Min Strikes‘ <dbl>,
## #   ‘Max Strikes‘ <dbl>, ‘Min People Killed‘ <dbl>, ‘Max People Killed‘ <dbl>,
## #   ‘Min Civilians Killed‘ <dbl>, ‘Max Civilians Killed‘ <dbl>,
## #   ‘Min Children Killed‘ <dbl>, ‘Max Children Killed‘ <dbl>,
## #   ‘Min People Injured‘ <dbl>, ‘Max People Injured‘ <dbl>,
## #   ‘Strike Link‘ <chr>, extra <chr>, Year <dbl>, ...
```

After doing some basic EDA (not included in this file), we have come to the realization that the current way that the columns are titled is annoying, inefficient, and frankly we don’t like using “ every time we want to reference a column. Therefore, we can use some simple regex manipulation to get the column titles into a more simple and reference-able state, as shown below.

```
colnames(combined_df) <- colnames(combined_df) %>%
  tolower() %>%
  str_replace_all(" ", "_") %>%
  str_replace_all("[/?]", "_")
head(combined_df)
```

```
## # A tibble: 6 x 33
##   country strike_id date                location   area   type_of_attack
##   <chr>   <chr>     <dttm>              <chr>      <chr>  <chr>
## 1 Yemen   YEM001    2002-11-03 00:00:00 Unknown    Marib  "Drone strike"
## 2 Yemen   YEM002    2009-12-17 00:00:00 Al Majala  Abyan  "Cruise missile strik~
## 3 Yemen   YEM003    2009-12-17 00:00:00 Arhab      Sanaa  "US-Yemen ground oper~
## 4 Yemen   YEM004    2009-12-24 00:00:00 Rafd       Shabwa "Airstrike\r\nPossibl~
## 5 Yemen   YEM005    2010-01-12 00:00:00 Unknown    Shabwa "Yemen ground operati~
## 6 Yemen   YEM006    2010-01-15 00:00:00 Al Ajashir Saada  "Airstrike"
## # i 27 more variables: confirmed_possible_us_strike <chr>, air_strike_ <dbl>,
## #   drone_strike <dbl>, min_strikes <dbl>, max_strikes <dbl>,
## #   min_people_killed <dbl>, max_people_killed <dbl>,
## #   min_civilians_killed <dbl>, max_civilians_killed <dbl>,
## #   min_children_killed <dbl>, max_children_killed <dbl>,
## #   min_people_injured <dbl>, max_people_injured <dbl>, strike_link <chr>,
## #   extra <chr>, year <dbl>, house_majority_party <chr>, ...
```

Continuing, we can also create columns for the amount of militant individuals affected by US drone strikes. Since we have the fields for people and civilians, we can assume that the difference between those two would be for military-related individuals. Hence, we can create the same min, medium, and max fields for militants.

```
combined_df$min_military_killed <- combined_df$min_people_killed - combined_df$min_civilians_killed

combined_df$max_military_killed <- combined_df$max_people_killed - combined_df$max_civilians_killed

combined_df$med_military_killed <- (combined_df$max_military_killed  + combined_df$min_military_killed)

head(combined_df)
```

```
## # A tibble: 6 x 36
##   country strike_id date                location   area   type_of_attack
##   <chr>   <chr>     <dttm>              <chr>      <chr>  <chr>
## 1 Yemen   YEM001    2002-11-03 00:00:00 Unknown    Marib  "Drone strike"
## 2 Yemen   YEM002    2009-12-17 00:00:00 Al Majala  Abyan  "Cruise missile strik~
## 3 Yemen   YEM003    2009-12-17 00:00:00 Arhab      Sanaa  "US-Yemen ground oper~
## 4 Yemen   YEM004    2009-12-24 00:00:00 Rafd       Shabwa "Airstrike\r\nPossibl~
## 5 Yemen   YEM005    2010-01-12 00:00:00 Unknown    Shabwa "Yemen ground operati~
## 6 Yemen   YEM006    2010-01-15 00:00:00 Al Ajashir Saada  "Airstrike"
## # i 30 more variables: confirmed_possible_us_strike <chr>, air_strike_ <dbl>,
## #   drone_strike <dbl>, min_strikes <dbl>, max_strikes <dbl>,
## #   min_people_killed <dbl>, max_people_killed <dbl>,
## #   min_civilians_killed <dbl>, max_civilians_killed <dbl>,
## #   min_children_killed <dbl>, max_children_killed <dbl>,
## #   min_people_injured <dbl>, max_people_injured <dbl>, strike_link <chr>,
## #   extra <chr>, year <dbl>, house_majority_party <chr>, ...
```

Finally, we can rewrite our csv to export for EDA, data analysis, statistical testing, etc.

```r
write.csv(combined_df, "df_combined.csv", row.names = FALSE)
```