Title: Data Challenge #1

Authors: Brandon Wolff, Zachary Heinemann, and Stephanie Langeland

Due Date: 02/22/2017

Cleaning the data:

```
rm(list = ls(all = TRUE))   # cleans everything in the workspace

library(readr)          # easier reading of flat files
library(readxl)         # easier reading of excel files
library(dplyr)          # data manipulation functions
```

```
##
## 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
```

```
library(tidyr)          # tools for tidy datasets
```

```
## Warning: package 'tidyr' was built under R version 3.3.2
```

```
library(magrittr)       # this is not a pipe
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(lubridate)      # easier manipulation of time objects
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
## 
##     date
```

```
library(stringr)        # easier manipulation of strings
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.3.2
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: purrr
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## as.difftime(): lubridate, base
## date():        lubridate, base
## filter():      dplyr, stats
## intersect():   lubridate, base
## lag():         dplyr, stats
## setdiff():     lubridate, base
## union():       lubridate, base
```

```
path <- "/Users/StephanieLangeland/Desktop/Columbia/Applied Data Science/Git/QMSS_G5069_Applied_D_S/Data Challenges"

#path <- "C:\\Users\\Brandon\\Documents\\GitHub\\QMSS_G5069_Applied_D_S\\Data Challenges"

# define additional paths for files you will use. In each case, determine
# appropriate additions to the path:
inFileName1  <- "A-E.xlsx"           # raw data on confrontations
inFileName2  <- "ARCH535.csv"   # name equivalence tables
outFileName1 <- "ConfrontationsData_170209.csv" # output file name

# set your path to that defined above, and confirm it
setwd(path)
getwd()
```

```
## [1] "/Users/StephanieLangeland/Desktop/Columbia/Applied Data Science/Git/QMSS_G5069_Applied_D_S/Data Challenges"
```

```
# LOADING RAW DATA
# the original file uses 9999 as a sentinel value for missing values changing
# back to null upon loading

library(readxl)
Confrontations <- read_excel(inFileName1,
                             sheet = 1,
                             na = "9999"   # converting sentinel value to null
)

# rough validations that data was correctly loaded
names(Confrontations)
```

```
##  [1] "ID"        "TIMESTAMP" "DIA"       "MES"       "AÑO"
##  [6] "ESTADO"    "Municipio" "DE"        "PF"        "MIF"
## [11] "MAF"       "PFF"       "AFIF"      "PEF"       "PMF"
## [16] "PMUF"      "AMPF"      "DOF"       "CIF"       "PL"
## [21] "MIL"       "MAL"       "PFL"       "AFIFL"     "PEL"
## [26] "PML"       "PMUL"      "AMPL"      "DOL"       "CIL"
## [31] "ARL"       "ARC"       "CARG"      "CART"      "VE"
## [36] "AC"        "AP"        "DEL"       "TOR"       "DTRA"
## [41] "PRE"       "FCRU"      "ELE"       "TAX"       "DRO"
## [46] "VEH"       "VAL"
```

```
nrow(Confrontations)
```

```
## [1] 3835
```

```
summary(Confrontations)
```

```
##        ID           TIMESTAMP              DIA             MES
##  Min.   :   1.0   Min.   :1.169e+09   Min.   : 1.00   Min.   : 1.000
##  1st Qu.: 959.5   1st Qu.:1.255e+09   1st Qu.: 8.00   1st Qu.: 4.000
##  Median :1918.0   Median :1.285e+09   Median :16.00   Median : 7.000
##  Mean   :1918.0   Mean   :1.276e+09   Mean   :15.81   Mean   : 6.488
##  3rd Qu.:2876.5   3rd Qu.:1.304e+09   3rd Qu.:23.00   3rd Qu.: 9.000
##  Max.   :3835.0   Max.   :1.322e+09   Max.   :31.00   Max.   :12.000
##
##       AÑO           ESTADO        Municipio           DE
##  Min.   :2007   Min.   : 1.00   Min.   :  1.0   Min.   : 0.000
##  1st Qu.:2009   1st Qu.:12.00   1st Qu.: 13.0   1st Qu.: 1.000
##  Median :2010   Median :19.00   Median : 27.0   Median : 2.000
##  Mean   :2010   Mean   :18.95   Mean   : 35.3   Mean   : 3.563
##  3rd Qu.:2011   3rd Qu.:28.00   3rd Qu.: 39.0   3rd Qu.: 4.000
##  Max.   :2011   Max.   :32.00   Max.   :469.0   Max.   :40.000
##                                 NA's   :1       NA's   :2388
##        PF              MIF            MAF             PFF
##  Min.   : 0.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.: 1.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.000
##  Median : 2.000   Median :1.00   Median :1.000   Median :1.000
##  Mean   : 2.509   Mean   :1.31   Mean   :1.357   Mean   :1.723
##  3rd Qu.: 3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:2.000
##  Max.   :29.000   Max.   :6.00   Max.   :3.000   Max.   :8.000
##  NA's   :1669     NA's   :3748   NA's   :3821    NA's   :3788
##       AFIF           PEF             PMF             PMUF
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.25   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :2.00   Median :1.000   Median :1.000   Median :1.000
##  Mean   :2.50   Mean   :1.667   Mean   :1.667   Mean   :1.609
##  3rd Qu.:2.75   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :6.00   Max.   :6.000   Max.   :7.000   Max.   :7.000
##  NA's   :3829   NA's   :3787    NA's   :3790    NA's   :3748
##       AMPF           DOF               CIF              PL
##  Min.   : NA    Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
##  1st Qu.: NA    1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : NA    Median : 2.000   Median : 1.000   Median : 2.000
##  Mean   :NaN    Mean   : 2.459   Mean   : 1.679   Mean   : 2.272
##  3rd Qu.: NA    3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.: 3.000
##  Max.   : NA    Max.   :29.000   Max.   :10.000   Max.   :30.000
##  NA's   :3835   NA's   :1991     NA's   :3611     NA's   :2172
##       MIL             MAL            PFL             AFIFL
##  Min.   :1.000   Min.   :1.00   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.: 1.000   1st Qu.: 1.000
##  Median :1.000   Median :2.00   Median : 2.000   Median : 1.000
##  Mean   :2.003   Mean   :2.48   Mean   : 2.405   Mean   : 2.615
##  3rd Qu.:3.000   3rd Qu.:3.00   3rd Qu.: 3.000   3rd Qu.: 3.000
##  Max.   :9.000   Max.   :9.00   Max.   :16.000   Max.   :15.000
##  NA's   :3516    NA's   :3810   NA's   :3724     NA's   :3822
##       PEL             PML             PMUL            AMPL
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000   Median :1.000   Median :1.000   Median :1.000
##  Mean   :1.944   Mean   :1.908   Mean   :1.834   Mean   :1.333
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.500
##  Max.   :8.000   Max.   :7.000   Max.   :8.000   Max.   :2.000
##  NA's   :3746    NA's   :3748    NA's   :3660    NA's   :3832
##       DOL              CIL              ARL               ARC
##  Min.   : 1.000   Min.   : 1.000   Min.   :  1.000   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:  2.000   1st Qu.: 1.000
##  Median : 1.000   Median : 1.000   Median :  3.000   Median : 2.000
##  Mean   : 1.881   Mean   : 1.943   Mean   :  5.175   Mean   : 2.436
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.:  6.000   3rd Qu.: 3.000
##  Max.   :30.000   Max.   :27.000   Max.   :144.000   Max.   :34.000
##  NA's   :3052     NA's   :3499     NA's   :2139      NA's   :2781
##       CARG              CART             VE                AC
##  Min.   :   1.00   Min.   :    1   Min.   :  1.000   Min.   :0.00000
##  1st Qu.:   5.00   1st Qu.:   79   1st Qu.:  1.000   1st Qu.:0.00000
##  Median :  19.00   Median :  402   Median :  1.000   Median :0.00000
##  Mean   :  46.26   Mean   : 1171   Mean   :  2.779   Mean   :0.01904
##  3rd Qu.:  45.00   3rd Qu.: 1180   3rd Qu.:  3.000   3rd Qu.:0.00000
##  Max.   :4000.00   Max.   :86365   Max.   :354.000   Max.   :1.00000
##  NA's   :2493      NA's   :2612    NA's   :1990
##        AP             DEL               TOR                DTRA
##  Min.   :0.000   Min.   :0.00000   Min.   :0.000000   Min.   :   0.0
##  1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.: 999.5
##  Median :0.000   Median :0.00000   Median :0.000000   Median :1342.0
##  Mean   :0.261   Mean   :0.07458   Mean   :0.002086   Mean   :1239.9
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1567.0
##  Max.   :1.000   Max.   :1.00000   Max.   :1.000000   Max.   :1776.0
##
##       PRE                 FCRU             ELE
##  Min.   :0.0000000   Min.   :0.0000   Min.   :0.0000000
##  1st Qu.:0.0000000   1st Qu.:0.0000   1st Qu.:0.0000000
##  Median :0.0000000   Median :0.0000   Median :0.0000000
##  Mean   :0.0007823   Mean   :0.4931   Mean   :0.0002608
##  3rd Qu.:0.0000000   3rd Qu.:1.0000   3rd Qu.:0.0000000
##  Max.   :1.0000000   Max.   :1.0000   Max.   :1.0000000
##
##       TAX               DRO               VEH              VAL
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.01095   Mean   :0.03051   Mean   :0.5129   Mean   :0.2334
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##
```

```
# :::::: LOADING NAME CONVERSION TABLE
# the original file treats numeric codes as strings, must convert to integers
# upon loading. Also, names of municipalities are in Spanish, so must specify
# the encoding as the file is read

NameTable <- read_csv(inFileName2,
                      col_types = cols(
                          CVE_ENT = col_integer(),    # must convert to integer
                          NOM_ENT = col_character(),
                          NOM_ABR = col_character(),
                          CVE_MUN = col_integer(),    # must convert to integer
                          NOM_MUN = col_character()
                          ),
                      locale = locale(encoding = "ISO-8859-1") # to read accents properl
y
                      )

# rough validations that data was correctly loaded
names(NameTable)
```

```
## [1] "CVE_ENT" "NOM_ENT" "NOM_ABR" "CVE_MUN" "NOM_MUN"
```

```
nrow(NameTable)
```

```
## [1] 2458
```

```
summary(NameTable)
```

```
##     CVE_ENT        NOM_ENT            NOM_ABR            CVE_MUN     
##  Min.   : 1.00   Length:2458        Length:2458        Min.   :  1.0  
##  1st Qu.:14.00   Class :character   Class :character   1st Qu.: 23.0  
##  Median :20.00   Mode  :character   Mode  :character   Median : 56.0  
##  Mean   :19.26                                         Mean   :108.8  
##  3rd Qu.:24.00                                         3rd Qu.:128.8  
##  Max.   :32.00                                         Max.   :570.0  
##    NOM_MUN         
##  Length:2458       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

```
# SOME DATA PROCESSING
# as released, the database is not immediately usable, so some data processing
# is needed to start exploring the data

# 1. add actual names of states and municipalities from a Census table;
#    currently the database only has their numeric codes
# 2. rename columns from Spanish to English (not everyone speaks both languages)
# 3. convert UNIX timestamp variable to a time object; this will be useful to
#    seamlessly create a date variable, and extract month names for graphing
# 4. some additional string changes in state abbreviations that will be useful
#    when graphing
# 5. replace all missing values with 0; this will come in handy as we start to
#    explore the data futher


fullData <-
    Confrontations %>%
        # adding State and Municipality names to dataframe
        left_join(., NameTable,
                  by = c("ESTADO" = "CVE_ENT",
                         "Municipio" = "CVE_MUN")
        ) %>%
        # renaming variables to intelligible English
        rename(day.orig = DIA,
               month.orig = MES,
               #year.orig = AÃ‘O,#had to change this part
               #to run the code on windows
               year.orig = AÑO,
               state_code = ESTADO,
               mun_code = Municipio,
               state = NOM_ENT,
               state.abbr = NOM_ABR,
               municipality = NOM_MUN,
               event.id = ID,
               unix.timestamp = TIMESTAMP,
               detained = DE,
               total.people.dead = PF,
               military.dead = MIF,
               navy.dead = MAF,
               federal.police.dead = PFF,
               afi.dead = AFIF,
               state.police.dead = PEF,
               ministerial.police.dead = PMF,
               municipal.police.dead = PMUF,
               public.prosecutor.dead = AMPF,
               organized.crime.dead = DOF,
               civilian.dead = CIF,
               total.people.wounded = PL,
               military.wounded = MIL,
               navy.wounded = MAL,
               federal.police.wounded = PFL,
               afi.wounded = AFIFL,
               state.police.wounded = PEL,
               ministerial.police.wounded = PML,
               municipal.police.wounded = PMUL,
               public.prosecutor.wounded = AMPL,
               organized.crime.wounded = DOL,
               civilian.wounded = CIL,
               long.guns.seized = ARL,
               small.arms.seized = ARC,
               cartridge.sezied = CART,
               clips.seized = CARG,
               vehicles.seized = VE
        ) %>%
        # creating date by converting unix timestamp, other time-related information
        # can later be extracted from this variable
        # also modifying state abbreviations by capitalizing and droping period
        # to "beautify" graph labels later on
        mutate(date = as.Date(as.POSIXct(unix.timestamp, origin="1970-01-01")),
               state.abbr = str_to_upper(str_replace_all(state.abbr, "[[:punct:]]", ""))

        ) %>%
        # keeping only necessary variables
        select(event.id, unix.timestamp, date,
               state_code, state, state.abbr, mun_code, municipality,
               detained, total.people.dead, military.dead, navy.dead,
               federal.police.dead, afi.dead, state.police.dead,
ministerial.police.dead,
               municipal.police.dead, public.prosecutor.dead, organized.crime.dead,
               civilian.dead, total.people.wounded, military.wounded, navy.wounded,
               federal.police.wounded, afi.wounded, state.police.wounded,
               ministerial.police.wounded, municipal.police.wounded,
               public.prosecutor.wounded, organized.crime.wounded, civilian.wounded,
               long.guns.seized, small.arms.seized, cartridge.sezied, clips.seized,
               vehicles.seized
        ) %>%
        # filling in NAs with zeros, to facilitate graphing and basic computations
        # replace_na() requires a list of columns and rules to apply. Code below
        # provides that
        replace_na(
            setNames(            # creates an object with numeric column names
                lapply(          # applies a function that links numeric column names
                                 # with the asignment of 0
                    vector("list", length(select_if(., is.numeric))), # creates a list l
ength 25
                            function(x) x <- 0),  # defines assignment of 0 to numeric c
ol names
                names(select_if(., is.numeric)))  # provides numeric column names
        )
```

1) Can you replicate the 86.1% number? The overall lethality ratio?

The ratios for the Federal Police, Navy and Army?

- These figures cannot be reproduced because the dataset does not include civilians who were involved in these events who were neither wounded nor killed.
  This makes it impossible to reproduce the overall lethality figure. Additionally, the dataset does not

distinguish between civilians killed or wounded by federal police, army, or navy personnel making it impossible to reproduce the 86.1% figure and lethality ratios for the navy, army, and federal police.

1a) Provide a visualization that presents this information neatly.

- Not applicable - see response to #1 above.

1b) Please show the exact computations you used to calculate them

(most likely than not, you’ll need to do some additional munging

in the data to get there).

- Not applicable - see response to #1 above.

1c) If you could not replicate them, please show why and the difference

relative to your own computations (also, include a neat graph that summarizes

this).

```
#Group Calculations:
#civilian lethality%
fullData$Total.Civilian.Conf <- fullData$civilian.dead + fullData$civilian.wounded
civilian_lethality <- (sum(fullData$civilian.dead))/sum((fullData$Total.Civilian.Conf))
civilian_lethality
```

```
## [1] 0.3654033
```

```
fullData$Civilian.lethality <- (fullData$civilian.dead)/(fullData$Total.Civilian.Conf)
valid.cases <- 3835-sum(is.na(fullData$Civilian.lethality))
valid.cases
```

```
## [1] 495
```

```
civ_leth_by_case <- sum(fullData$Civilian.lethality, na.rm = TRUE)/495
civ_leth_by_case
```

```
## [1] 0.37937
```

```
#Total Lethality%
fullData$Total.Conf <- fullData$total.people.dead + fullData$total.people.wounded
Total_lethality <- (sum(fullData$total.people.dead))/sum((fullData$Total.Conf))
Total_lethality
```

```
## [1] 0.5898633
```

```
#organized crime lethality%
fullData$Total.organized.crime.Conf <- fullData$organized.crime.dead + fullData$organized.crime.wounded
organized_crime_lethality <- (sum(fullData$organized.crime.dead))/sum((fullData$Total.organized.crime.Conf))
organized_crime_lethality
```

```
## [1] 0.7548269
```

```
#Federal Police lethality%
fullData$Total.Federal.Police.Conf <- fullData$federal.police.dead + fullData$federal.police.wounded
Federal_Police_lethality <- (sum(fullData$federal.police.dead))/sum((fullData$Total.Federal.Police.Conf))
Federal_Police_lethality
```

```
## [1] 0.2327586
```

```
#Federal Police deaths per 1 wounded
Federal_Police_lethality2 <- (sum(fullData$federal.police.dead))/sum((fullData$federal.police.wounded))
Federal_Police_lethality2
```

```
## [1] 0.3033708
```

```
#Navy Lethality%
fullData$Total.Navy.Conf <- fullData$navy.dead + fullData$navy.wounded
Navy_lethality <- (sum(fullData$navy.dead))/sum((fullData$Total.Navy.Conf))
Navy_lethality
```

```
## [1] 0.2345679
```

```
#ARMY deaths per 1 wounded
Navy_lethality2 <- (sum(fullData$navy.dead))/sum((fullData$navy.wounded))
Navy_lethality2
```

```
## [1] 0.3064516
```

```
#Army Lethality%
fullData$Total.military.Conf <- fullData$military.dead + fullData$military.wounded
Military_lethality <- (sum(fullData$military.dead))/sum((fullData$Total.military.Conf))
Military_lethality
```

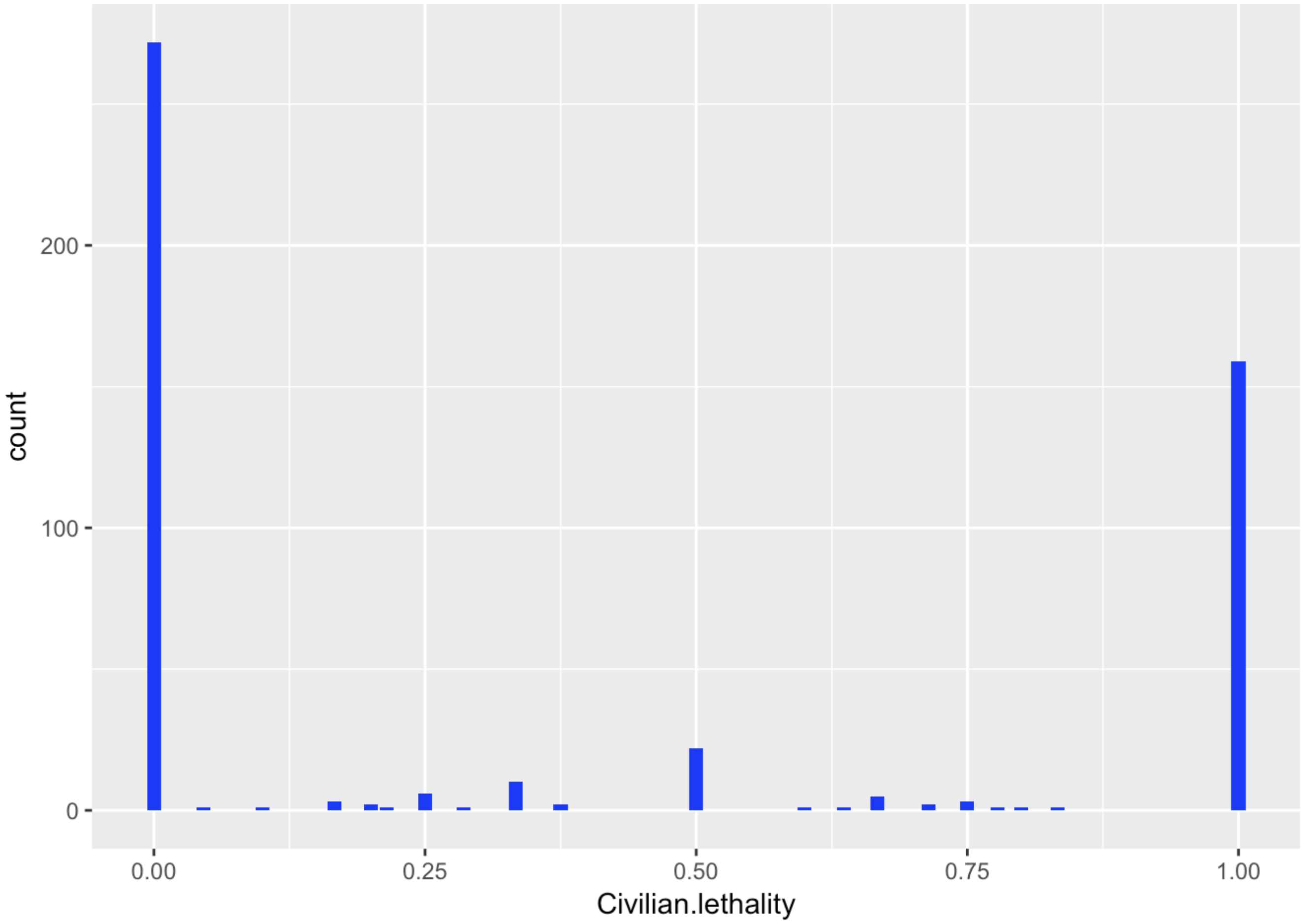
```
## [1] 0.1513944
```

```
#ARMY deaths per 1 wounded
Military_lethality2 <- (sum(fullData$military.dead))/sum((fullData$military.wounded))
Military_lethality2
```

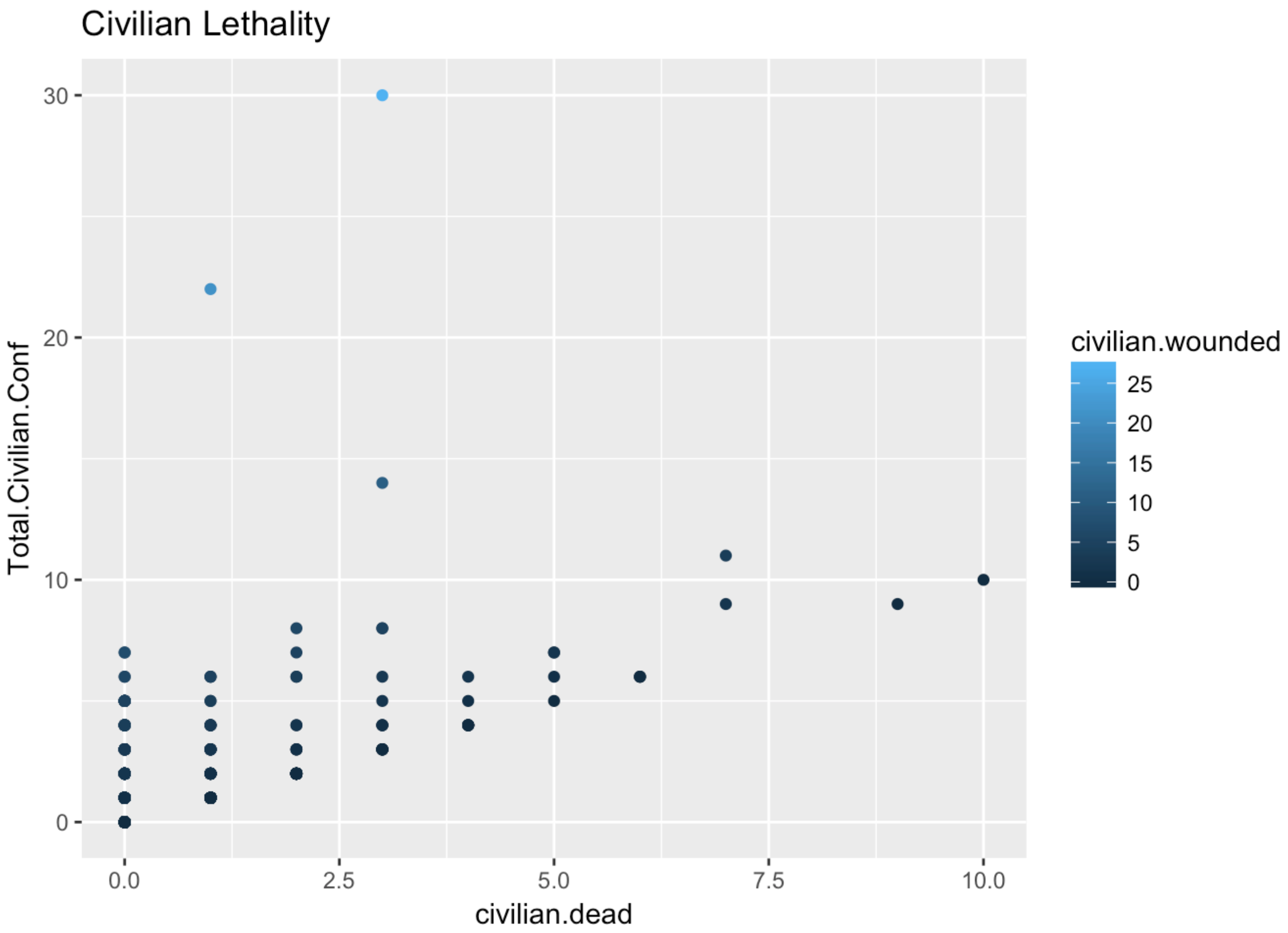
```
## [1] 0.1784038
```

```
#Visualizations:
b <- ggplot(fullData)
b <- b + geom_bar(mapping = aes(Civilian.lethality), fill = "blue")
b
```

```
## Warning: Removed 3340 rows containing non-finite values (stat_count).
```

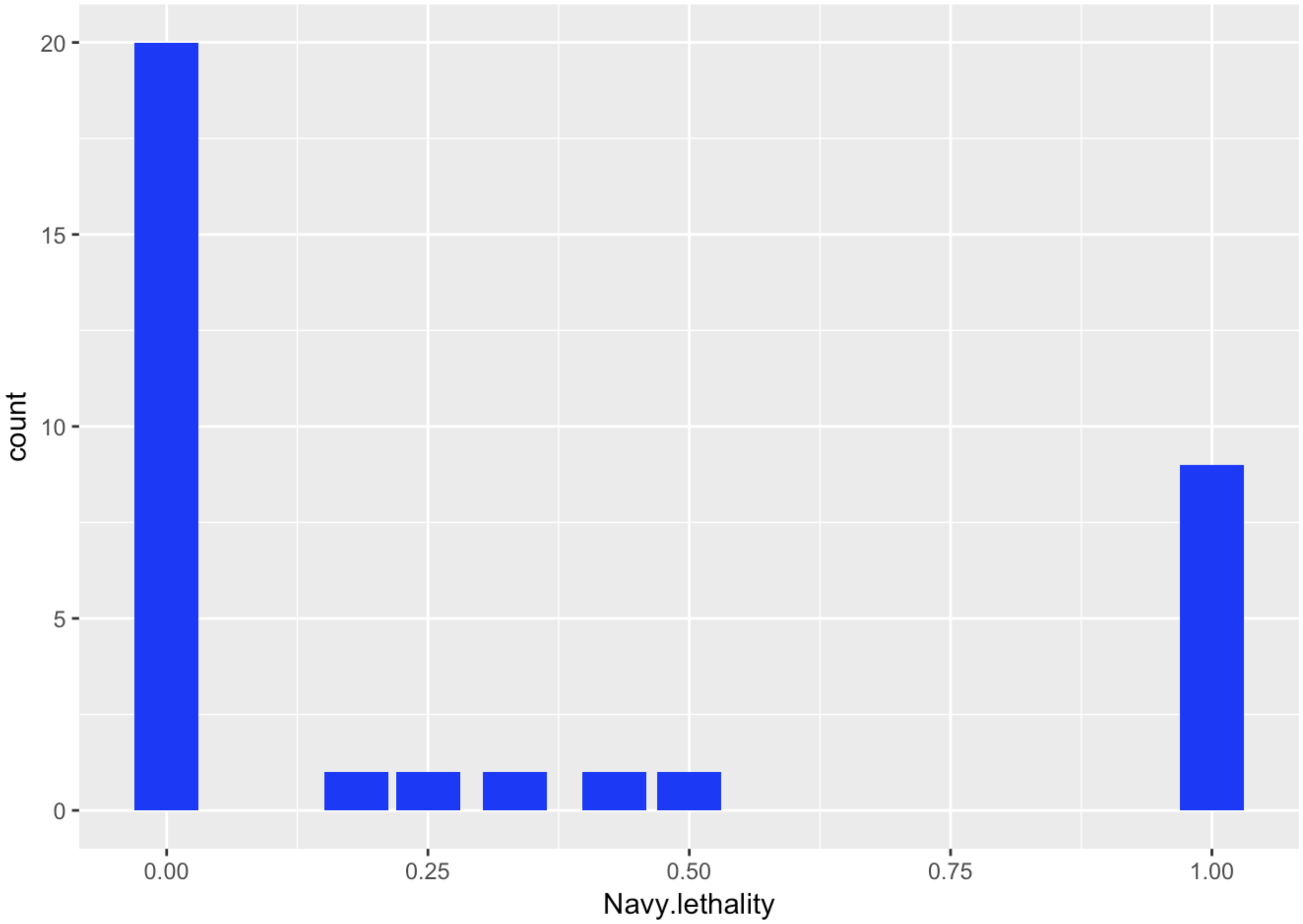


```
#in this graph 0 = wounded with no deaths and 1 = deaths with no wounded
#put this graph here to show the exterme difference of the results we came to from the 8
6.1%

#We could not replicate the results, this may be due to using different data
#or becuase we used a different method which made more logical sense to us.

B <- ggplot(fullData, aes(x =civilian.dead, y = Total.Civilian.Conf))

B + geom_point(aes(color = civilian.wounded)) +
  ggtitle("Civilian Lethality")
```



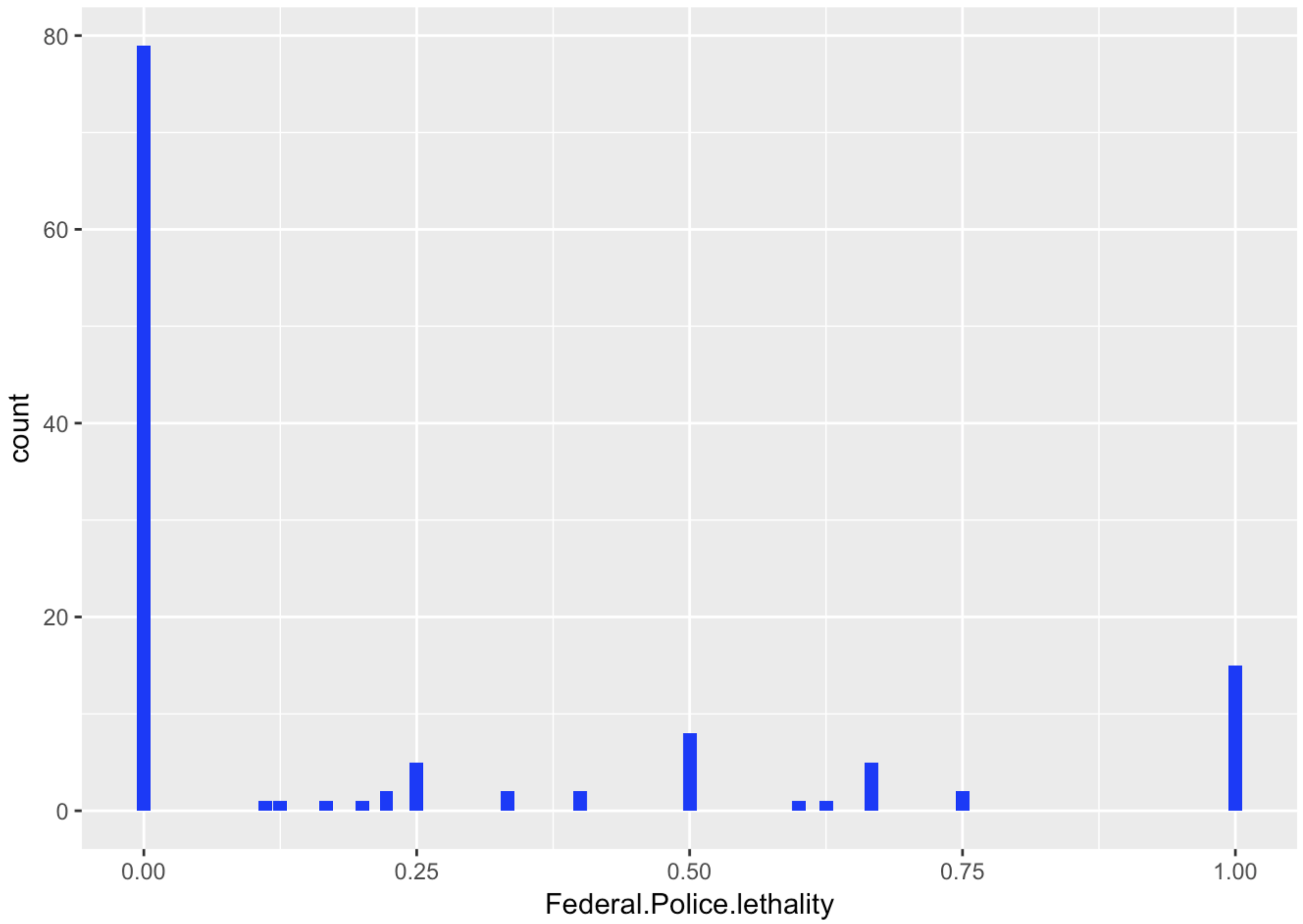
```
#Graph Navy.Lethality
fullData$Navy.lethality <- (fullData$navy.dead)/(fullData$Total.Navy.Conf)
n <- ggplot(fullData)
n <- n + geom_bar(mapping = aes(Navy.lethality), fill = "blue")
n
```

```
## Warning: Removed 3801 rows containing non-finite values (stat_count).
```



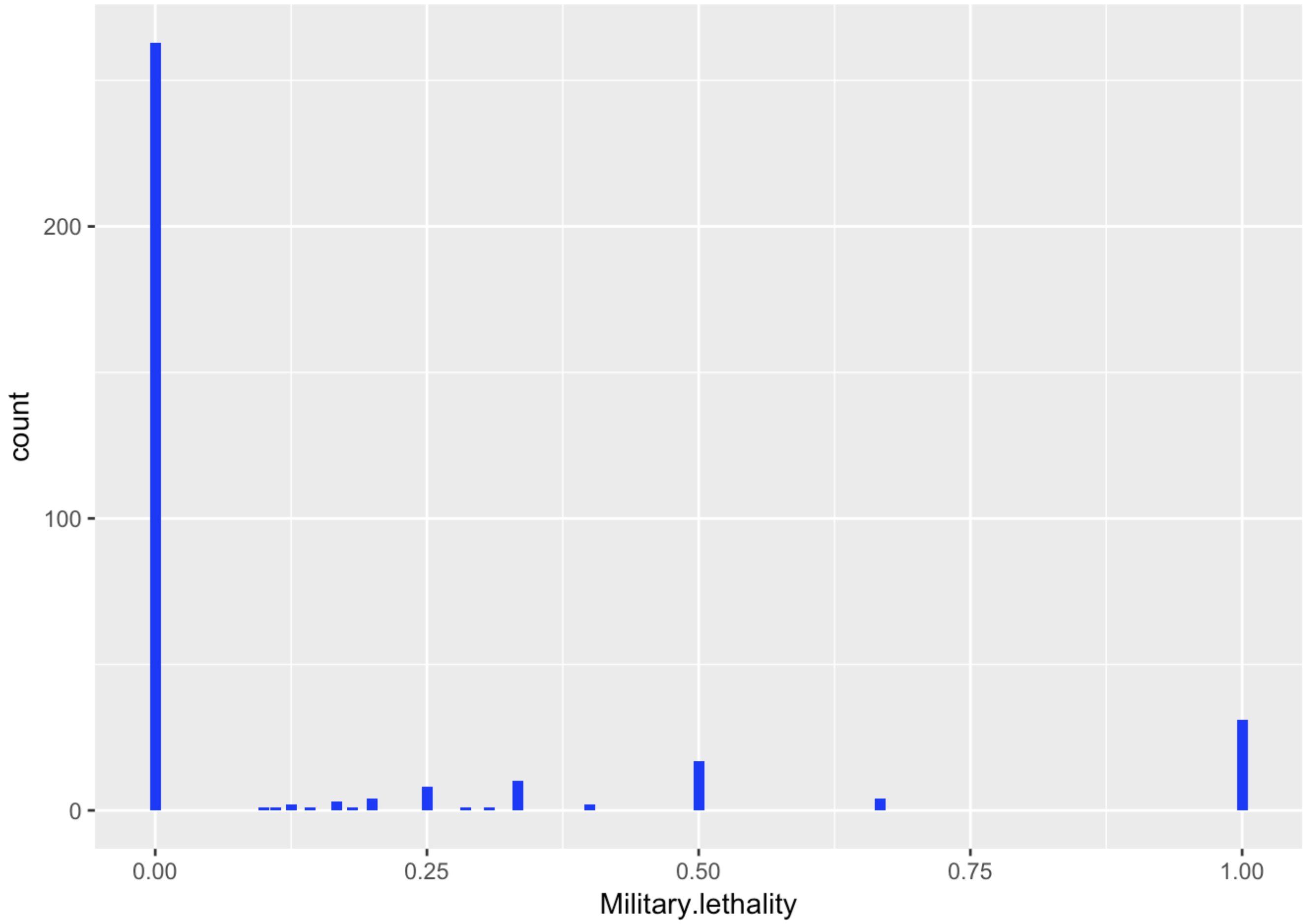
```
#Graph Fed.Police.Lethality
fullData$Federal.Police.lethality <- (fullData$federal.police.dead)/(fullData$Total.Fede
ral.Police.Conf)
fp <- ggplot(fullData)
fp <- fp + geom_bar(mapping = aes(Federal.Police.lethality), fill = "blue")
fp
```

```
## Warning: Removed 3709 rows containing non-finite values (stat_count).
```



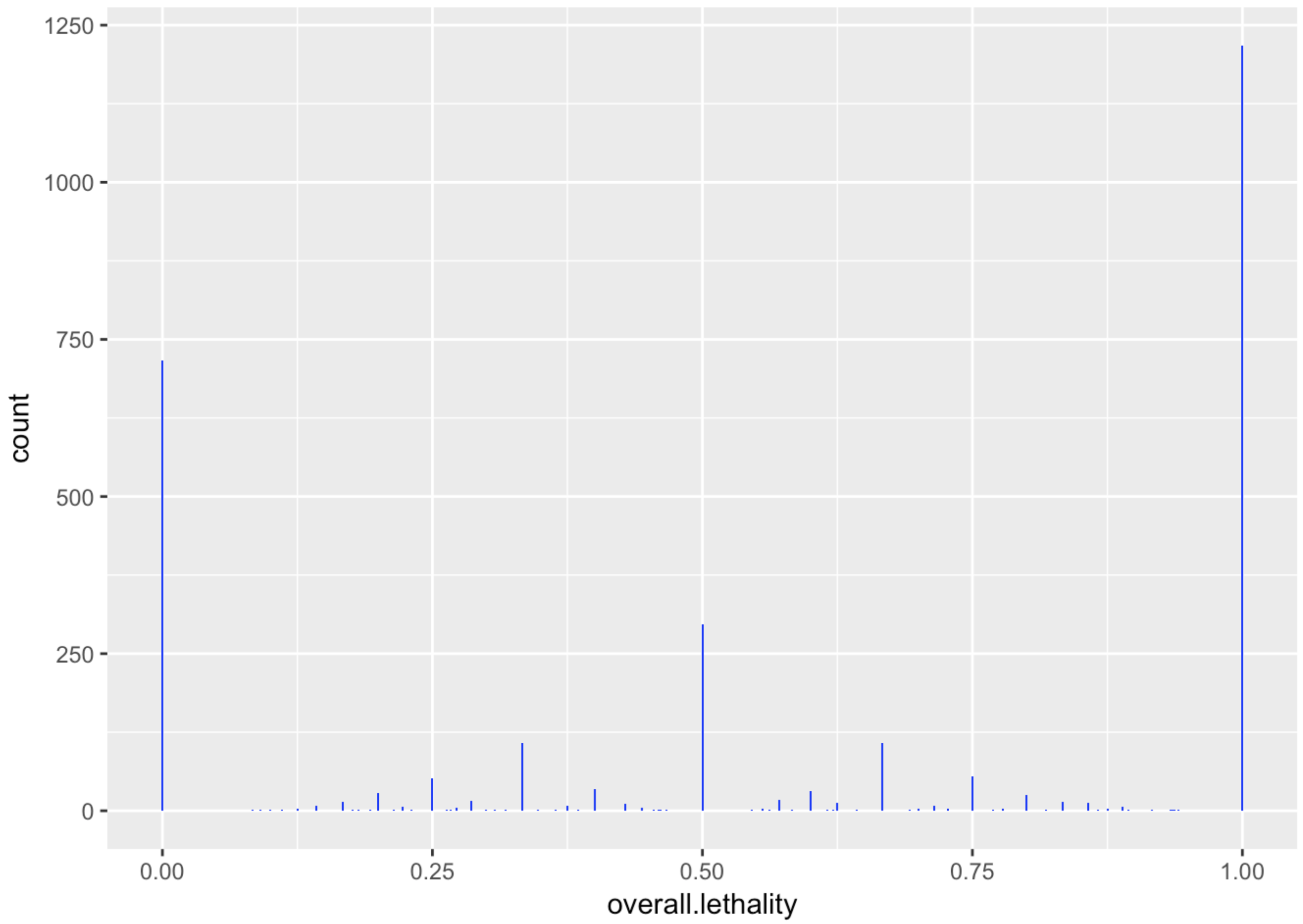
```
#Graph ARMY.Lethality
fullData$Military.lethality <- (fullData$military.dead)/(fullData$Total.military.Conf)
a <- ggplot(fullData)
a <- a + geom_bar(mapping = aes(Military.lethality), fill = "blue")
a
```

```
## Warning: Removed 3485 rows containing non-finite values (stat_count).
```



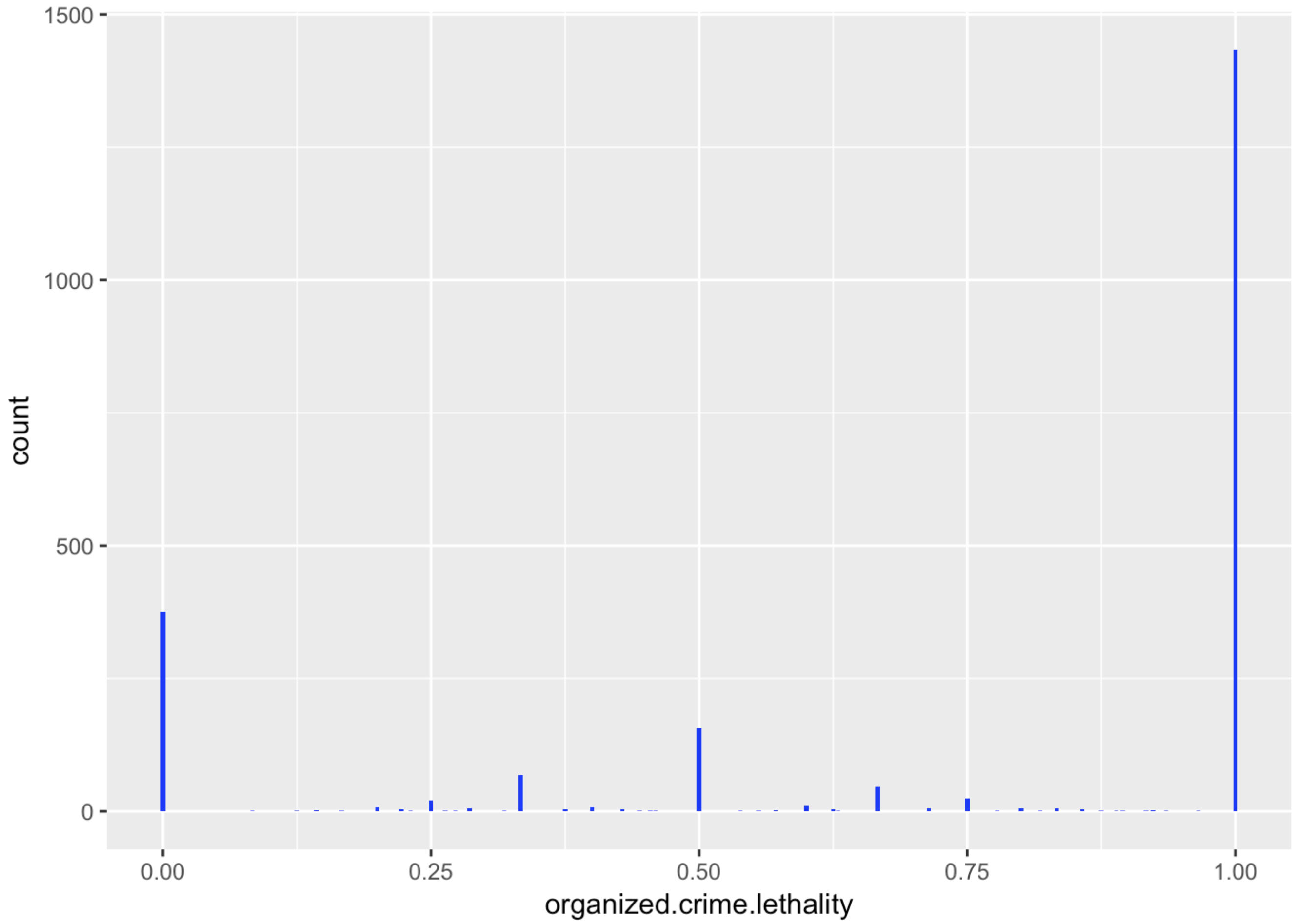
```
#overall Lethality
fullData$overall.lethality <- (fullData$total.people.dead)/(fullData$Total.Conf)
o <- ggplot(fullData)
o <- o + geom_bar(mapping = aes(overall.lethality), fill = "blue")
o
```

```
## Warning: Removed 954 rows containing non-finite values (stat_count).
```



```
#organized.crime lethality
fullData$organized.crime.lethality <- (fullData$organized.crime.dead)/(fullData$Total.or
ganized.crime.Conf)
oc <- ggplot(fullData)
oc <- oc + geom_bar(mapping = aes(organized.crime.lethality), fill = "blue")
oc
```

```
## Warning: Removed 1618 rows containing non-finite values (stat_count).
```



1d) Be very explicit: What are you assuming to generate these computations?

- The assumption is that all civilian interactions are either civilians who were wounded or killed. Therefore, no civilians survived any interaction without being killed or wounded. Additionally, in order to ascertain whether civilians were wounded or killed as a result of armed forces or federal police, we must assume that each observation (civilian.dead > 0 and/or civilian.wounded > 0) is also associated with a wounded or dead observation for the federal police, military, and navy (e.g., navy.dead > 0 and/or navy.wounded > 0).

2) Now you know the data more intimately. Think a little bit more about it,

and answer the following questions:

2a) Is this the right metric to look at? Why or why not? 2b) What is the “lethality

index" showing explicitly? What is it not showing? What is the

definition assuming?

There are quite a few critical considerations with regard to the dataset at one. First and foremost, it is worth critically assessing what exactly is being captured by this dataset. By the very definition of the dataset, a case is only ever included in this particular set if it is defined as a confrontation via the definition provided in the brief. In one regard, this is selecting on the dependent variable insofar as there are no cases of interactions with the police that do not have any violence and therefore any statistic will be likely inflating the overall percentage because it disregards the potentially large number of cases in which an interaction occurs in the absence of violence and therefore the denominator in the probability calculation is much lower than it should be.

It is worth noting that not only is the definition of confrontation constrained to violent interactions, but actually the violent interactions that do occur must be between the police and members of organized crime. Indeed, as per the formal definition of confrontation, there must "be at least three (3) organized crime members, or less if using military-grade equipment and explosives," which indicates that even instances where violence did occur but it was only between those who are not believed to or known to be organized crime members would still not be included in the dataset. Actually, via the definition of confrontation, there are a number of restraints, such as the instance must involve "violent resistance to armed forces and other government authorities" which would preclude an instance from being included in this dataset. The critical takeaway here is not even that some censoring on the dependent variable is occurring within this dataset (thereby artificially inflating the conditional probabilities calculated), but actually the tremendous amount of restraints on the dataset implemented by the narrow definition of confrontation that actually likely censor quite a large deal of cases so much so that the current statistics may not be entirely reliable.

Under this framework, it may be meaningful to conclude that the lethality index is an inappropriate measure altogether because it likely inflates the extent to which the police are acting violently and killing as opposed to wounding in the general population. In some regard, this makes good sense. If hypothetically the police killed 5 for wounding every 1, but in actuality the cases presented are only .01% of all the interactions between police and people, then the gravity of this ratio is likely sensationalized as it presented. However, this does not necessarily reduce this metric to having no value; rather, it is critical just to report the statistic exactly as it is. The interpretation presented in the piece does inflate the ratio due to the aforementioned censoring problem, but the actual index is simply a measure of the number killed compared to the number wounded given that a confrontation has occurred. The article does use the word "confrontation" explicitly, but it does not give the formal (and rather narrow, quite frankly) definition that actually precludes a number of cases that the reader may assume when reading "confrontation" as presented in the piece. Therefore, the presentation of the index and the data is not necessarily dishonest outright insofar as it does use the proper word, but it fails to address the particular definition that was used to construct the dataset, thereby biasing the interpretation presented to the readers.

Therefore, this metric is not necessarily meaningless, it can be used as an understanding of the rates at which police in general (and in comparison with other groups) kill versus wound in these particular situations, so long as it is interpreted with specific regard to exactly what it is measuring and not necessarily extrapolated to a general metric of brutality of a lack of training on behalf of the police. Whether or not this metric is appropriate, therein, is not necessarily an inherent property but instead contingent upon the question being asked. If a researcher or reporter wants to understand what groups are more adept at wounding instead of killing in confrontation situations, as per this narrow definition, then this index is perfectly acceptable; however, it does not appear that this was necessarily the question the article was interested in answering and instead there was an attempt to extrapolate this to an overall claim about the relative training of different groups in wounding versus killing, and such an extrapolation realistically cannot be supported by the data. Rather, this metric makes a very explicit assumption that a confrontation has occurred and then assesses the relative outcomes given the occurrence of that, which can certainly be meaningful, but is not able to answer the questions that the journalist presenting the piece appears to be interested in. For example, in order to extrapolate this to a broader claim about relative training of particular groups, one would need the data of non-confrontation encounters and compare those ratios to the ratios presented here; it is certainly possible that perhaps police officers have a lower lethality index in confrontation situations, but a higher one relative to their reference groups in non-confrontation situations – therefore undermining the claim of differential training that the piece suggests.

Additionally, it is worth noting that very little information about the actual event itself is gleaned besides very basic facts. This particular measure does not make recognition of the temporal aspect (does it matter when in the day and when in the year these events are occurring? Does it matter what year it is as crime rates change over time?), as well as the localization aspect (does it matter what state or municipality it is in?) and instead looks at a very globalized perspective; aggregated all of the space and time of the events to create one statistic. By definition, this may be a useful summary of the data, but it is most certainly reductive. There is reason to believe given the visualizations of the data that there is at least meaningful differences across different years and even within the same year of these dead rates and therefore lethality indices and therefore to aggregate such a tremendous amount of data into one point reduces meaningful trends and components of the relationship between the variables that helps elucidate the insight of police brutality in Mexico. Additionally, there may be reason to believe that similarly there are meaningful differences across the country that such a reductive aggregation does not capture. For example, perhaps this lethality index is actually being largely driven just by one year and only in one municipality due to an exogenous event that spiked the incidence of confrontations on the basis of organized crime in that area such that there were many killed. While the data does not necessarily appear to suggest this; this hypothetical point elucidates the possibility of significant trends being missed in the reduction of the data. Therefore, this index is not only censoring, but also reductive by definition – meaning that the insight suggested by the article may not necessarily approximate the reality.

Additionally, there are limitations outside of the dataset about the nature of the events such that the insights being gleaned are dubious. The amount of data collected about the incident is largely demographic in nature and does not include specificities of the confrontation that could impact the interpretation of the numbers. For example, if in a particular confrontation the army had to come in because of the severity of a crime and use something much more widespread (e.g., a grenade) in order to quell the threat, it would make sense that the many civilians in the area, in addition to the few dangerous criminals, would be killed and therefore the overall statistic is being driven by the severity of the threat posed. Furthermore, instances in which the army or navy are required as opposed to police force may be fundamentally different instances such that there is a tremendous need for intervention and quelling of the threat immediately so much so that more destructive (and by definition less precise) methods of quelling the threat are necessary, thereby inflating the mortality rate. Therefore, perhaps there is a fundamentally different nature to the confrontations in which the navy or army is required that inflates the statistic, and not necessarily the lack of training in quelling targets without killing civilians (or wounding without killing) – which the article appears to suggest. Therefore, the lack of substantive information about the events is another form of censoring that reduces the tangible meaning of the lethality index.

2c) With the same available data, can you think of an alternative way to capture

the same construct? Is it “better”?

The necessity or even use of a different metric is similarly contingent upon the question that the researcher wants to present and the insight that they wish to draw. It appears that the report was interested in two major insights: typically speaking in Mexico more are killed than wounded in confrontations and different groups (such as the army or navy) tend to have an even stronger effect for this than police. Under the narrow definition presented of confrontation, theoretically there is support for both claims with the existing metric. However, it is worth noting that the claim being made is likely more about the overall lethality of Mexico and not necessarily this particular metric. With the data given, the metric that was presented does make some sense (as the entire dataset itself is censored which is not necessarily the fault of the researcher) and is still providing an insight, albeit a limited one into a specific type of situation.

Theoretically, there are a few things that could perhaps improve the measurement. For one, it may be critically to provide a more dynamic measurement that incorporates municipality and time as components of the statistic (perhaps in the form of visualization or a specific statistical model). This measurement, while certainly less parsimonious, may give insight into specificities of the index as it varies and therefore provides more insight into the causes behind and severity of crime as opposed to homogenizing over all of Mexico. Indeed, it may be critical to have a nuanced understanding of the when and where such that the threat being suggested and the lack of training being accused by the article is perhaps solidified or tempered by taking into account factors related to the time and place of the confrontations.

As aforementioned, one of the major limitations of the previous measure is that it fails to incorporate (although the dataset does as well), idiosyncrasies of the confrontations that may change why a particular statistic (for example number of army or navy dead) is higher because of the extremity of the threat and subsequent necessity for quelling measures that result in quite a lot of death. However, the dataset does not necessarily provide a clear where to include this in the model. There are a number of variables about the amount of weaponry seized, which potentially could be used as a proxy for the severity or danger presented by the particular confrontation (for example if many guns were seized, it could indicate that the event was far more dangerous and therefore necessitated something like a grenade which would kill many in order to quell the threat). Incorporating this measure into the model or the calculation of the statistic may be helpful, but admittedly this is a rather weak proxy insofar as it does not necessarily give a clear depiction of the situation, just an after the fact measure of the number of guns seized (for example if 10 guns were seized, it could be 10 organized crime members who were quelled without any casualties who happened to have guns, but it also could indicate that there was a massive shoot-out which required the army or navy to intervene). Therefore, this measure may improve the statistic somewhat, but it is hard to assess by how much and there is a certain possibility that it may not necessarily improve it at all. In terms of estimating lethality, one could examine the ratio of non-civilian to civilian deaths as a means of understanding the situation (how many civilians were involved? Was it essentially just an explosion of deaths because of the severity of the situation? Was it an isolated event between the organized criminals and the armed forces?) and to give an idea of the danger presented not just to those involved, by definition, in the confrontation, but those not intentionally involved as well. However, this metric still does not capture many of the realities of the situation as the complexity of confrontations is difficult to boil down to just a comparison of some numbers; rather, it is near impossible to assess the constituents of a situation and the danger just by relative death counts – especially if the civilian count is 0 in a given confrontation, which is frequently is and could just indicate that all of the civilians in the area were kept safe by the armed forces even if many in the armed forces were harmed.

2d) What additional information would you need to better understand the data?

2e) What additional information could help you better capture the construct behind the “lethality index”

In general, the biggest considerations for what the data would need would be data that involves all interactions with the police that are not the formal definition of confrontations to supplement the current dataset that is exclusively confrontations. Specifically – in order to make the claim that police officers in Mexico are good or bad with regard to injuring versus killing, one would at least need the data of all violent interactions (whether they be with an organized crime member or not) such that one could discern if situations that necessitate some form of violence are more likely to result in wounding instead of killing those involved.

However, even if this particular dataset would be flawed insofar as it does not necessarily address the source and the necessity of the violence. If the overall claim is being made about the skills and training of the police force, such a dataset would not necessarily be able to glean if the poor training results in police officers engaging in violent confrontations more than is necessary, thereby also inflating the statistic. This exposes an important consideration: in order to make many of the important claims about the general training and lethality of Mexico, there needs to be many more variables collected about the nature of each incident. For example, knowing what exactly happened, knowing why the violence occurred, knowing what type of threat and the gravity of the threat posed to the officer are among some of the information about the interaction that would help elucidate the claim being made by the piece. However, such an ideal dataset is difficult, if not impossible, to gather. In general, one of the issues of data science is the extent to which we are bound by the natural limitations of our data.

Therefore, it is not necessarily a question of what would be the ideal dataset to answer the question, but rather what insights can be gathered from the specific data, so long as we are explicit and clear about what assumptions are or are not a part of the data. Even perhaps just something as broad as a statistic of the percentage of police interactions in this time that were confrontations could help scale the incidences and percentages such that it becomes more close to the reality of the danger presented by police officers overall in dealing with organized crime. For example, if the ratio of police officers killed to wounded is 2.6, but the number killed and number wounded are both incredibly low in the overall population (especially when compared with non-violent interactions), then the ratio may not be statistically significant and may instead be just because both numbers comparatively are quite low and just by chance happen to have one 2.6 times larger than the other (which is an even greater case for testing the statistical significance of all of these ratios). Additionally, such a situation would also indicate that even if the lethality index is high (when defined in those terms), overall it is quite safe to be a police officer and therefore this relative relationship may not necessarily be of importance. Again, this derives back to the question of what precisely constitutes the lethality index as the researchers and reporters wish to both understand and construe it to their readers at large; under the framework of wishing to define it simply as how much more likely a death is than a wounding, it is still critical to include all of the situational factors of the particular confrontation as well as the non-confrontation data in order to properly construct such an index.

This index is a very specific probability which certainly has merit to some extent, but is not necessarily as expansive as it appears. If the researcher wants to understand the index more broadly with regard to the profession, all information of interactions would be necessary. If the researcher wants to understand the index with regard to dealing with organized crime, then at least confrontation and non-confrontation data would be necessary in order to derive the probability of confrontation and then calculate the conditional probability accordingly. If the researcher is interested in understand the relative risk of dying versus being wounded under the very specific definition of a confrontation (with all of the aforementioned stipulations as per the definition), then this statistic achieves such an aim and therein a conclusion can tentatively be made. This suggests that there is not necessarily an absolute “best” way to measure the lethality index, rather it is critical to define the index in the context of the insight that the analyst wishes to draw and then evaluate the data’s ability to do so. Rather, it is not necessarily fully possible to construct a “better” measure of the lethality index without widening the definition of said index with regard to the aforementioned considerations and subsequently then evaluating what is the type of data necessary to measure such an index. This elucidates a critical data science theme: the insights one can draw are limited by the data, and the data collected should theoretically be structured with the insights desired in mind.