Final Project Step Two

Alan Donahue

8/8/2021

# How to import and clean my data

#setting the working directory  
setwd("C:/Users/Alan Donahue/Documents/data science masters/DSC 520 Stats/GIT/dsc520")  
  
#load the libraries  
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

#load the data  
veteran\_suicide\_sex.df <- read.csv("data/project\_data/veteran\_suicide\_by\_sex.csv")  
suicide\_age.df <- read.csv("data/project\_data/suicide\_by\_age.csv")  
veteran\_suicide.df <- read.csv("data/project\_data/veteran\_suicide\_overall.csv")  
non\_veteran\_suicide.df <- read.csv("data/project\_data/non-veteran\_suicide\_overall.csv")  
recent\_VHA\_user.df <- read.csv("data/project\_data/recent\_VHA\_user.csv")  
non\_recent\_VHA\_user.df <- read.csv("data/project\_data/non-recent\_VHA\_user.csv")  
  
  
#only taking the total per year of veteran suicides split by sex  
total\_veteran\_suicide\_sex <- veteran\_suicide\_sex.df %>% filter(State\_of\_Death == "Total U.S.")  
#checking for spelling issues  
unique(total\_veteran\_suicide\_sex$Year)

## [1] 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018

unique(total\_veteran\_suicide\_sex$Geographic.Region)

## [1] " "

unique(total\_veteran\_suicide\_sex$State\_of\_Death)

## [1] "Total U.S."

unique(total\_veteran\_suicide\_sex$Sex)

## [1] "Total" "Male" "Female"

#only taking the total per year of veteran suicides split by age  
total\_veteran\_suicide\_age <- suicide\_age.df %>% filter(State\_of\_Death == "Total U.S.")  
#checking for spelling issues  
unique(total\_veteran\_suicide\_age$Year)

## [1] 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018

unique(total\_veteran\_suicide\_age$Geographic.Region..Based.on.State.of.Death)

## [1] " "

unique(total\_veteran\_suicide\_age$State\_of\_Death)

## [1] "Total U.S."

unique(total\_veteran\_suicide\_age$Age.Group)

## [1] "Total" "18–34" "35–54" "55–74" "75+"

For the data sets, in the final project step 1, I had a total of 15 data sets. Since then, I have slimmed down the data to a total of 6 data sets. I am able to slim down the data because I have refined my research question to what is the most vulnerable time for a veteran to think about committing suicide.

Originally, I had two excel files that contained multiple tables on multiple different sheets. I decided to take the 6 important data sets and separate them by file. I chose to use a .csv file for each one because that is the easiest to work with for me. Now I have 6 distinct data sets to be able to work with.

The veteran suicide by sex and by age both included breakdowns by states in their respective data sets. I decided to only focus on the yearly data because the rest of my data sets were about each year.

Otherwise, the data was already very clean. There weren’t missing data, incorrect spelling, or anything like that.

# What does the final data set look like?

#number of rows  
nrow(total\_veteran\_suicide\_sex)

## [1] 42

nrow(total\_veteran\_suicide\_age)

## [1] 70

nrow(veteran\_suicide.df)

## [1] 14

nrow(non\_veteran\_suicide.df)

## [1] 14

nrow(recent\_VHA\_user.df)

## [1] 14

nrow(non\_recent\_VHA\_user.df)

## [1] 14

#number of columns  
ncol(total\_veteran\_suicide\_sex)

## [1] 5

ncol(total\_veteran\_suicide\_age)

## [1] 8

ncol(veteran\_suicide.df)

## [1] 14

ncol(non\_veteran\_suicide.df)

## [1] 14

ncol(recent\_VHA\_user.df)

## [1] 14

ncol(non\_recent\_VHA\_user.df)

## [1] 14

#Column names for each data set  
colnames(total\_veteran\_suicide\_sex)

## [1] "Year" "Geographic.Region" "State\_of\_Death"   
## [4] "Sex" "X.Veteran.Suicides."

colnames(total\_veteran\_suicide\_age)

## [1] "Year"   
## [2] "Geographic.Region..Based.on.State.of.Death"   
## [3] "State\_of\_Death"   
## [4] "Age.Group"   
## [5] "X.Veteran.Suicides."   
## [6] "Veteran.Suicide.Rate.per.100.000"   
## [7] "X.General.Adult.Population.Suicides."   
## [8] "General.Adult.Population.Suicide.Rate.per.100.000"

colnames(veteran\_suicide.df)

## [1] "Year.of.Death"   
## [2] "Veteran.Suicide.Deaths"   
## [3] "Veteran.Population.Estimate"   
## [4] "Veteran.Crude.Suicide.Rate.per.100.000"   
## [5] "Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [6] "Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000"  
## [7] "Male.Veteran.Suicide.Deaths"   
## [8] "Male.Veteran.Population.Estimate"   
## [9] "Male.Veteran.Crude.Suicide.Rate.per.100.000"   
## [10] "Male.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [11] "Female.Veteran.Suicide.Deaths"   
## [12] "Female.Veteran.Population.Estimate"   
## [13] "Female.Veteran.Crude.Suicide.Rate.per.100.000"   
## [14] "Female.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"

colnames(non\_veteran\_suicide.df)

## [1] "Year.of.Death"   
## [2] "Non.Veteran.Suicide.Deaths"   
## [3] "Non.Veteran.Population.Estimate"   
## [4] "Non.Veteran.Crude.Suicide.Rate.per.100.000"   
## [5] "Non.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [6] "Non.Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000"  
## [7] "Male.Non.Veteran.Suicide.Deaths"   
## [8] "Male.Non.Veteran.Population.Estimate"   
## [9] "Male.Non.Veteran.Crude.Suicide.Rate.per.100.000"   
## [10] "Male.Non.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [11] "Female.Non.Veteran.Suicide.Deaths"   
## [12] "Female.Non.Veteran.Population.Estimate"   
## [13] "Female.Non.Veteran.Crude.Suicide.Rate.per.100.000"   
## [14] "Female.Non.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"

colnames(recent\_VHA\_user.df)

## [1] "Year.of.Death"   
## [2] "VHA.Veteran.Suicide.Deaths"   
## [3] "VHA.Veteran.Population.Estimate"   
## [4] "VHA.Veteran.Crude.Suicide.Rate.per.100.000"   
## [5] "VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [6] "VHA.Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000"  
## [7] "Male.VHA.Veteran.Suicide.Deaths"   
## [8] "Male.VHA.Veteran.Population.Estimate"   
## [9] "Male.VHA.Veteran.Crude.Suicide.Rate.per.100.000"   
## [10] "Male.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [11] "Female.VHA.Veteran.Suicide.Deaths"   
## [12] "Female.VHA.Veteran.Population.Estimate"   
## [13] "Female.VHA.Veteran.Crude.Suicide.Rate.per.100.000"   
## [14] "Female.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"

colnames(non\_recent\_VHA\_user.df)

## [1] "Year.of.Death"   
## [2] "Non.VHA.Veteran.Suicide.Deaths"   
## [3] "Non.VHA.Veteran.Population.Estimate"   
## [4] "Non.VHA.Veteran.Crude.Suicide.Rate.per.100.000"   
## [5] "Non.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [6] "Non.VHA.Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000"  
## [7] "Male.Non.VHA.Veteran.Suicide.Deaths"   
## [8] "Male.Non.VHA.Veteran.Population.Estimate"   
## [9] "Male.Non.VHA.Veteran.Crude.Suicide.Rate.per.100.000"   
## [10] "Male.Non.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"   
## [11] "Female.Non.VHA.Veteran.Suicide.Deaths"   
## [12] "Female.Non.VHA.Veteran.Population.Estimate"   
## [13] "Female.Non.VHA.Veteran.Crude.Suicide.Rate.per.100.000"   
## [14] "Female.Non.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000"

#head function  
head(total\_veteran\_suicide\_sex)

## Year Geographic.Region State\_of\_Death Sex X.Veteran.Suicides.  
## 1 2005 Total U.S. Total 6,056  
## 2 2005 Total U.S. Male 5,870  
## 3 2005 Total U.S. Female 186  
## 4 2006 Total U.S. Total 5,968  
## 5 2006 Total U.S. Male 5,800  
## 6 2006 Total U.S. Female 168

head(total\_veteran\_suicide\_age)

## Year Geographic.Region..Based.on.State.of.Death State\_of\_Death Age.Group  
## 1 2005 Total U.S. Total  
## 2 2005 Total U.S. 18–34  
## 3 2005 Total U.S. 35–54  
## 4 2005 Total U.S. 55–74  
## 5 2005 Total U.S. 75+  
## 6 2006 Total U.S. Total  
## X.Veteran.Suicides. Veteran.Suicide.Rate.per.100.000  
## 1 6,056 24.7  
## 2 574 25.5  
## 3 2,122 28.1  
## 4 1,970 20.1  
## 5 1,387 28.1  
## 6 5,968 24.8  
## X.General.Adult.Population.Suicides.  
## 1 31,610  
## 2 8,455  
## 3 13,541  
## 4 6,554  
## 5 3,060  
## 6 32,352  
## General.Adult.Population.Suicide.Rate.per.100.000  
## 1 14.7  
## 2 13.1  
## 3 15.9  
## 4 13.5  
## 5 18.7  
## 6 14.4

head(veteran\_suicide.df)

## Year.of.Death Veteran.Suicide.Deaths Veteran.Population.Estimate  
## 1 2005 6,056 24,546,000  
## 2 2006 5,968 24,020,000  
## 3 2007 6,174 23,597,000  
## 4 2008 6,489 23,295,000  
## 5 2009 6,455 22,914,000  
## 6 2010 6,472 22,739,000  
## Veteran.Crude.Suicide.Rate.per.100.000  
## 1 24.7  
## 2 24.8  
## 3 26.2  
## 4 27.9  
## 5 28.2  
## 6 28.5  
## Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 25.6  
## 2 25.4  
## 3 26.8  
## 4 28.7  
## 5 28.8  
## 6 29.3  
## Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000  
## 1 18.5  
## 2 17.8  
## 3 19.1  
## 4 20.9  
## 5 21.4  
## 6 21.8  
## Male.Veteran.Suicide.Deaths Male.Veteran.Population.Estimate  
## 1 5,870 22,699,000  
## 2 5,800 22,202,000  
## 3 5,992 21,820,000  
## 4 6,287 21,557,000  
## 5 6,232 21,135,000  
## 6 6,244 20,952,000  
## Male.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 25.9  
## 2 26.1  
## 3 27.5  
## 4 29.2  
## 5 29.5  
## 6 29.8  
## Male.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 27.6  
## 2 27.5  
## 3 29.1  
## 4 31.1  
## 5 31.0  
## 6 31.8  
## Female.Veteran.Suicide.Deaths Female.Veteran.Population.Estimate  
## 1 186 1,847,000  
## 2 168 1,818,000  
## 3 182 1,777,000  
## 4 202 1,738,000  
## 5 223 1,779,000  
## 6 228 1,787,000  
## Female.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 10.1  
## 2 9.2  
## 3 10.2  
## 4 11.6  
## 5 12.5  
## 6 12.8  
## Female.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 10.2  
## 2 9.1  
## 3 10.0  
## 4 11.6  
## 5 12.7  
## 6 12.6

head(non\_veteran\_suicide.df)

## Year.of.Death Non.Veteran.Suicide.Deaths Non.Veteran.Population.Estimate  
## 1 2005 25,554 189,978,444  
## 2 2006 26,384 200,628,294  
## 3 2007 27,580 203,118,104  
## 4 2008 28,556 205,606,197  
## 5 2009 29,384 208,308,799  
## 6 2010 30,876 211,398,287  
## Non.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 13.5  
## 2 13.2  
## 3 13.6  
## 4 13.9  
## 5 14.1  
## 6 14.6  
## Non.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 13.5  
## 2 13.2  
## 3 13.6  
## 4 13.9  
## 5 14.1  
## 6 14.6  
## Non.Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000  
## 1 15.5  
## 2 14.6  
## 3 15.0  
## 4 15.2  
## 5 15.4  
## 6 15.9  
## Male.Non.Veteran.Suicide.Deaths Male.Non.Veteran.Population.Estimate  
## 1 19,258 80,568,449  
## 2 19,798 86,634,862  
## 3 20,623 88,096,429  
## 4 21,439 89,405,764  
## 5 22,076 91,004,727  
## 6 23,282 92,260,564  
## Male.Non.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 23.9  
## 2 22.9  
## 3 23.4  
## 4 24.0  
## 5 24.3  
## 6 25.2  
## Male.Non.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 26.9  
## 2 24.7  
## 3 25.1  
## 4 25.5  
## 5 25.6  
## 6 26.4  
## Female.Non.Veteran.Suicide.Deaths Female.Non.Veteran.Population.Estimate  
## 1 6,296 109,409,995  
## 2 6,586 113,993,432  
## 3 6,957 115,021,675  
## 4 7,117 116,200,433  
## 5 7,308 117,304,072  
## 6 7,594 119,137,723  
## Female.Non.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 5.8  
## 2 5.8  
## 3 6.0  
## 4 6.1  
## 5 6.2  
## 6 6.4  
## Female.Non.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 5.8  
## 2 5.8  
## 3 6.1  
## 4 6.2  
## 5 6.3  
## 6 6.4

head(recent\_VHA\_user.df)

## Year.of.Death VHA.Veteran.Suicide.Deaths VHA.Veteran.Population.Estimate  
## 1 2005 1,712 5,289,086  
## 2 2006 1,799 5,377,962  
## 3 2007 1,774 5,430,509  
## 4 2008 1,923 5,501,565  
## 5 2009 1,887 5,670,577  
## 6 2010 1,914 5,877,245  
## VHA.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 32.4  
## 2 33.4  
## 3 32.7  
## 4 34.9  
## 5 33.3  
## 6 32.6  
## VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 30.0  
## 2 31.1  
## 3 30.3  
## 4 32.8  
## 5 31.3  
## 6 31.5  
## VHA.Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000  
## 1 22.7  
## 2 21.1  
## 3 22.0  
## 4 23.7  
## 5 23.3  
## 6 23.8  
## Male.VHA.Veteran.Suicide.Deaths Male.VHA.Veteran.Population.Estimate  
## 1 1,656 4,917,232  
## 2 1,765 4,988,796  
## 3 1,724 5,023,740  
## 4 1,870 5,075,978  
## 5 1,828 5,223,574  
## 6 1,848 5,408,524  
## Male.VHA.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 33.7  
## 2 35.4  
## 3 34.3  
## 4 36.8  
## 5 35.0  
## 6 34.2  
## Male.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 32.5  
## 2 35.0  
## 3 33.1  
## 4 36.6  
## 5 33.9  
## 6 34.7  
## Female.VHA.Veteran.Suicide.Deaths Female.VHA.Veteran.Population.Estimate  
## 1 56 371,854  
## 2 34 389,166  
## 3 50 406,769  
## 4 53 425,587  
## 5 59 447,003  
## 6 66 468,721  
## Female.VHA.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 15.1  
## 2 8.7  
## 3 12.3  
## 4 12.5  
## 5 13.2  
## 6 14.1  
## Female.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 13.8  
## 2 8.3  
## 3 11.9  
## 4 11.6  
## 5 13.5  
## 6 13.0

head(non\_recent\_VHA\_user.df)

## Year.of.Death Non.VHA.Veteran.Suicide.Deaths  
## 1 2005 4,344  
## 2 2006 4,169  
## 3 2007 4,400  
## 4 2008 4,566  
## 5 2009 4,568  
## 6 2010 4,558  
## Non.VHA.Veteran.Population.Estimate  
## 1 19,186,040  
## 2 18,570,944  
## 3 18,095,438  
## 4 17,722,390  
## 5 17,172,350  
## 6 16,790,705  
## Non.VHA.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 22.6  
## 2 22.4  
## 3 24.3  
## 4 25.8  
## 5 26.6  
## 6 27.1  
## Non.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 24.4  
## 2 23.8  
## 3 25.8  
## 4 27.5  
## 5 28.2  
## 6 28.7  
## Non.VHA.Veteran.Age..and.Sex.Adjusted.Suicide.Rate.per.100.000  
## 1 17.4  
## 2 17.1  
## 3 18.3  
## 4 20.2  
## 5 20.9  
## 6 21.3  
## Male.Non.VHA.Veteran.Suicide.Deaths Male.Non.VHA.Veteran.Population.Estimate  
## 1 4,214 17,713,633  
## 2 4,035 17,144,871  
## 3 4,268 16,728,061  
## 4 4,417 16,412,981  
## 5 4,404 15,843,430  
## 6 4,396 15,475,651  
## Male.Non.VHA.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 23.8  
## 2 23.5  
## 3 25.5  
## 4 26.9  
## 5 27.8  
## 6 28.4  
## Male.Non.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 26.3  
## 2 25.6  
## 3 28.0  
## 4 29.7  
## 5 30.3  
## 6 31.0  
## Female.Non.VHA.Veteran.Suicide.Deaths  
## 1 130  
## 2 134  
## 3 132  
## 4 149  
## 5 164  
## 6 162  
## Female.Non.VHA.Veteran.Population.Estimate  
## 1 1,472,407  
## 2 1,426,073  
## 3 1,367,377  
## 4 1,309,409  
## 5 1,328,920  
## 6 1,315,054  
## Female.Non.VHA.Veteran.Crude.Suicide.Rate.per.100.000  
## 1 8.8  
## 2 9.4  
## 3 9.7  
## 4 11.4  
## 5 12.3  
## 6 12.3  
## Female.Non.VHA.Veteran.Age.Adjusted.Suicide.Rate.per.100.000  
## 1 9.4  
## 2 9.4  
## 3 9.5  
## 4 11.6  
## 5 12.4  
## 6 12.3

Here is a quick breakdown of the 6 different data sets. This helps me visualize how I can move forward with combining the different data sets to find new information.

# Questions for future steps

Since the data sets are very clean to begin with, I don’t think I need to know anything else about importing or cleaning data. The hard work of the individuals who put together the initial data sets made it very simple to quickly clean up what I needed to clean up.

# What information is not self-evident?

Right now, all the information is pretty self-evident. Most of the data sets have some overlap in regard to the different variables they have like sex or age.

# What are different ways you could look at this data?

To answer the questions for this final project, I think I’m going to focus on age, sex, and whether a veteran received help from the VHA. I believe that these three topics are important to finding out when a veteran is most vulnerable to suicide. I want to look at the different trends in those three topics to help find an answer as well as look at all three together.

# How do you plan to slice and dice the data?

Yes I plan on slicing and dicing the data. Additionally, I think there might be some benefit of combining different parts of the data sets together.

I plan to slice the data sets in several different ways to see if there are any trends. I plan to slice by year (2005-2018), by age, by sex, and by whether or not they received help from the VHA. In order to do that, I will also have to take the different parts of the data sets and join them together.

# How could you summarize your data to answer key questions?

I can use the different summary statistics that R has to offer. For instance, I can find the mean of the rate of suicide by year, age, sex, and whether or not a veteran received help from the VHA.

Additionally, other functions include min, max, median, range, etc. All of these are very helpful to use to get a quick peek at the trends in the data.

# What types of plots and tables will help you illustrate the findings to your questions?

In the beginning to help explore the data more, box plots, histograms, and Q-Q plots are important. They can be used to visually look at the distribution of the data and if there are any outliers.

Additionally, frequency tables could be helpful to give a quick view of the data based on the four variables I plan to focus on.

# Do you plan to incorporate machine learning techiques?

At this point, I don’t know any machine learning techniques since that is going to be learned during week 10. If I find any benefit to utilizing machine learning techniques, I will use them in the project. Otherwise, I won’t include those techniques.

# Questions for future steps

Is there any type of machine learning technique that might help out my project?