Importing and 'tidying' data

Lacey Hartigan

11/22/21

```
knitr::opts_chunk$set(echo = TRUE)

library(haven)
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

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

Loading Data

Today, we're going to be working with naloxone distribution data. This is a data file that includes all naloxone distributions under TN SOR to date. I'm going to read it in from the M: Drive.

```
naloxone<-read_excel("M:/322 - State Opioid Response (SOR) Grant/11 - Quarterly and Annual Reports/Data/Na
```

Descriptives

First, let's go through some basic functions to pull descriptives from our data. Often times we want to know counts of each category, for example, for categorical data, or perhaps we want to pull the mean/median for continuous data. Let's explore some functions to do these things below. Note, this is NOT an exhaustive list. There are many many more functions than what I'll show you below.

First, let's pull a summary of the entire dataset so that I can quickly glimpse its attributes:

```
summary(naloxone)
```

```
##     region          typeofagencyutilizewhatsontheage   unitsdistr
##  Length:8781        Length:8781                       Min.   :   0.00
##  Class :character   Class :character                  1st Qu.:   3.00
##  Mode  :character   Mode  :character                  Median :  12.00
##                                                       Mean   :  24.66
##                                                       3rd Qu.:  24.00
```

```
##                                                  Max.   :1944.00
##
##    distr_date                     county            zipcode
##  Min.   :2017-01-08 00:00:00   Length:8781        Length:8781
##  1st Qu.:2019-04-08 00:00:00   Class :character   Class :character
##  Median :2020-03-10 00:00:00   Mode  :character   Mode  :character
##  Mean   :2020-02-17 03:03:04
##  3rd Qu.:2021-03-09 00:00:00
##  Max.   :2021-11-02 00:00:00
##  NA's   :3
##       FY17                FY18             FY19              FY20
##  Min.   :0.0000000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.0002278   Mean   :0.154   Mean   :0.2255   Mean   :0.2556
##  3rd Qu.:0.0000000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##
##       FY21              FY22             FY20_Q1            FY20_Q2
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.3424   Mean   :0.02198   Mean   :0.06366   Mean   :0.06605
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##     FY20_Q3            FY20_Q4            FY21_Q1            FY21_Q2
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.05523   Mean   :0.07061   Mean   :0.05683   Mean   :0.08268
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##     FY21_Q3            FY21_Q4          FY22_Q1
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.00000
##  Mean   :0.09646   Mean   :0.1065   Mean   :0.02198
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##
```

```
#for categorical variables, this doesn't tell me much; but for continuous data, it gives me a 5-number sum
```

Next, let's go ahead and pull frequency tables for categorical variables (I'm just going to do a few), so that we can see more info than the summary data above. Let's also pull a table for our "region" variable to see how well naloxone distribution varies across our state's regions.

```
table(naloxone$typeofagencyutilizewhatsontheage)
```

```
##
##             Detention/corrections facility
##                                         35
##           Emergency Medical Services (EMS)
```

```
##                                         50
##                     Faith-based organization
##                                        220
##                                Fire department
##                                        353
##                                Law enforcement
##                                       1290
##                                 Local business
##                                         29
##                               N/A (Individual)
##                                       3698
## Organization providing community resources
##                                       1143
##                                 Recovery court
##                                         88
##                                     School
##                                        339
##                   Social service organization
##                                        148
##               Syringe services program (SSP)
##                                        273
##             Treatment and/or recovery agency
##                                       1111
```

```
#This gives me a frequency count of all the types of agencies. For example, 273 of our 871 entries went to

table(naloxone$region)
```

```
##
##   R1  R2M  R2N  R2S  R3N  R3S   R4 R5N1 R5N2  R5S  R6N  R6S   R7
##  674  369  513  585  823  668 1205  570  278  615  743  668 1070
```

```
#Which region has the most naloxone entries? Which region has the least?
```

Limiting the data

It's great to see the overall picture of the data, but typically, we'll be working on a specific year/time period, rather than all data that we have. We can limit the dataset using different R commands, so that way we can pull only the data we want. Let's take the most recent fiscal year and just keep FY21.

```
#we're going to do this using tidyverse functions, which tend to be the most dynamic functions in R (in my
fy21data<-naloxone%>%
    filter(FY21==1)
    #FILTER operates on ROWS of data (think about filtering in excel); so what this will do is it will onl

#How many observations do we have for FY21? How many variables are in this new dataset?

#Now that we have a dataset with just FY21, let's re-pull our agency list to see how that distribution loo
table(fy21data$typeofagencyutilizewhatsontheage)
```

```
##
##               Detention/corrections facility
##                                           12
##             Emergency Medical Services (EMS)
```

```
##                                       10
##                     Faith-based organization
##                                       57
##                               Fire department
##                                      133
##                               Law enforcement
##                                      307
##                                Local business
##                                       27
##                              N/A (Individual)
##                                     1577
## Organization providing community resources
##                                      321
##                                Recovery court
##                                       23
##                                     School
##                                       84
##                  Social service organization
##                                       58
##               Syringe services program (SSP)
##                                       95
##             Treatment and/or recovery agency
##                                      303
```

```
#what was the most common entry for FY21?
```

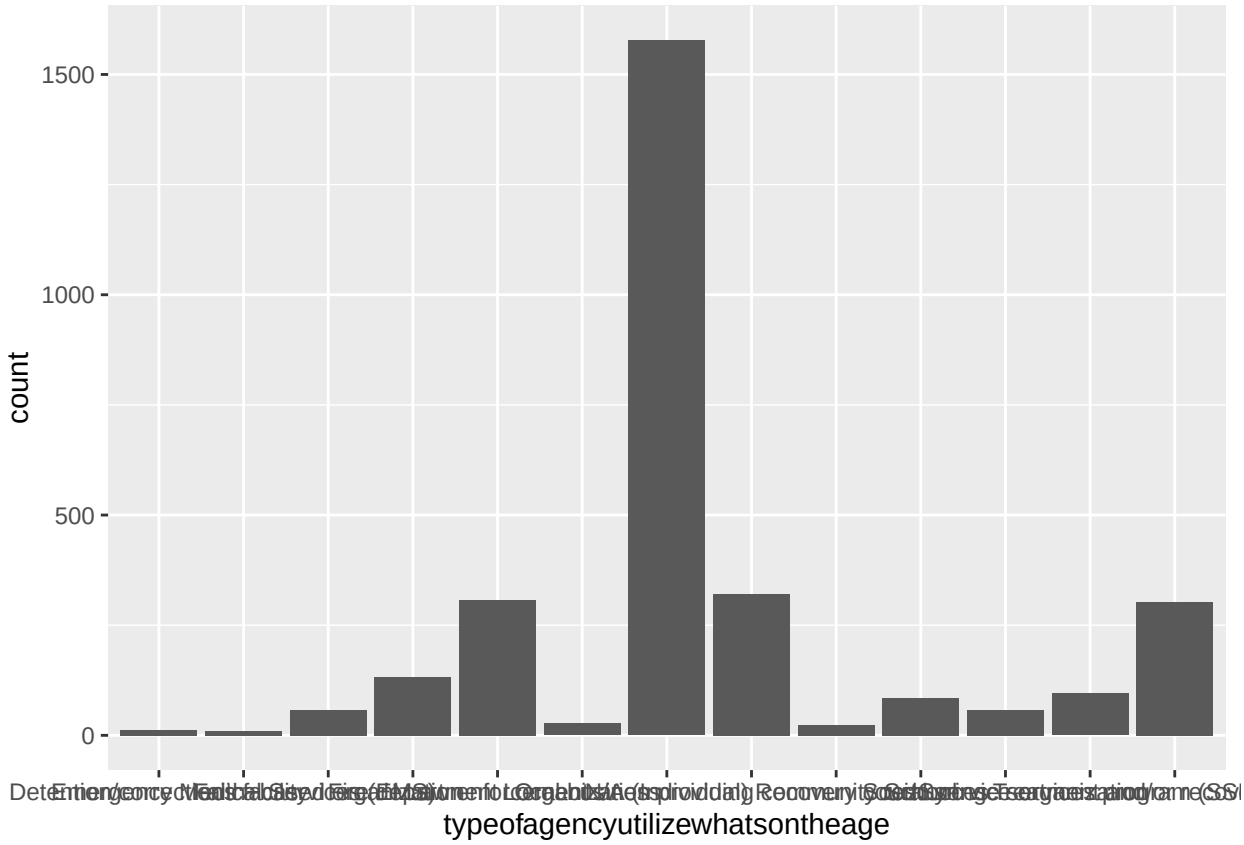
While we use `filter` to only keep (or exclude) the rows/observations we want (or don't want), we use `select` to only keep columns/variables that we want. Let's say that in our FY21 dataset we want to only keep FY21 variables.

```
fy21data<-fy21data%>%
    select(region, typeofagencyutilizewhatsontheage, unitsdistr, county, zipcode, FY21_Q1:FY21_Q4)
#LOOK at the resulting dataset. Make sure it looks how you expected it to.
```

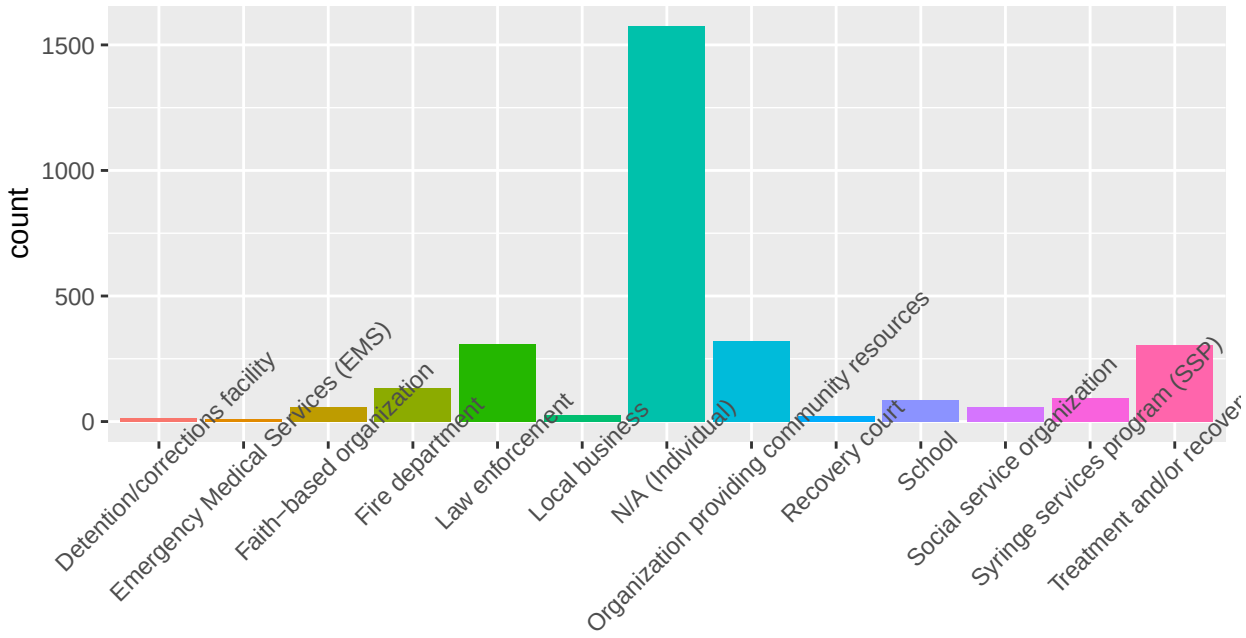
This is all good, but what if I want to plot the number of each type of agency? Then I need to prep a new dataset with the counts I pulled above. I can do that using `group_by`, `summarize`, and `n()` functions.

```
agencyplot<-fy21data%>%
    group_by(typeofagencyutilizewhatsontheage)%>%
    summarize(count=n())
#LOOK at the resulting dataset. Does it look how you want it to? If so, you can go ahead and plot it.
```

```
agenplot<-ggplot(agencyplot, aes(typeofagencyutilizewhatsontheage, y=count))+
    geom_bar(stat="identity")
agenplot
```

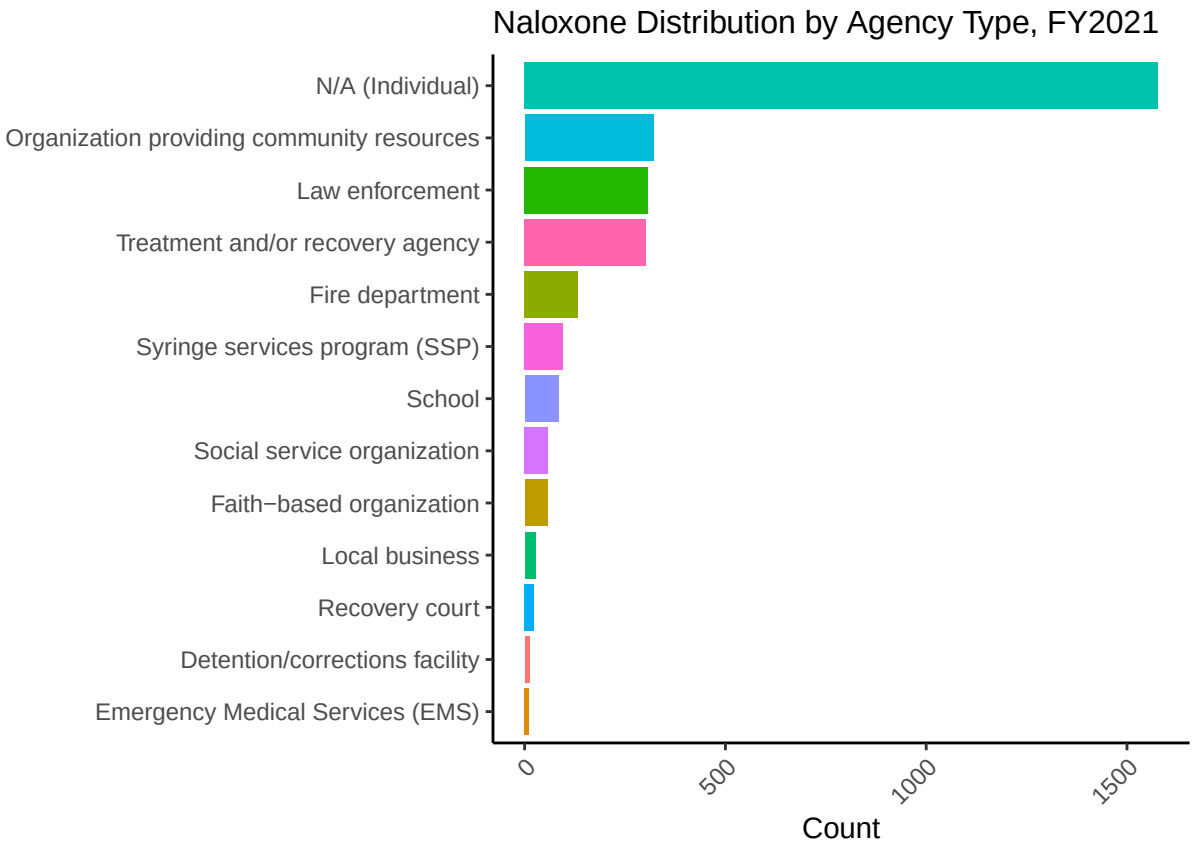



```
#what's wrong with this plot? let's fix it - and let's also make it a little prettier :)
#remove the legend and then tweak the angle of the labels so we can see them
agenplot<-ggplot(agencyplot, aes(x=typeofagencyutilizewhatsontheage, y=count, fill=typeofagencyutilizewhat
    geom_bar(stat="identity")+
    theme(legend.position="none", axis.text.x=element_text(angle=45))
agenplot
```



typeofagencyutilizewhatsontheage

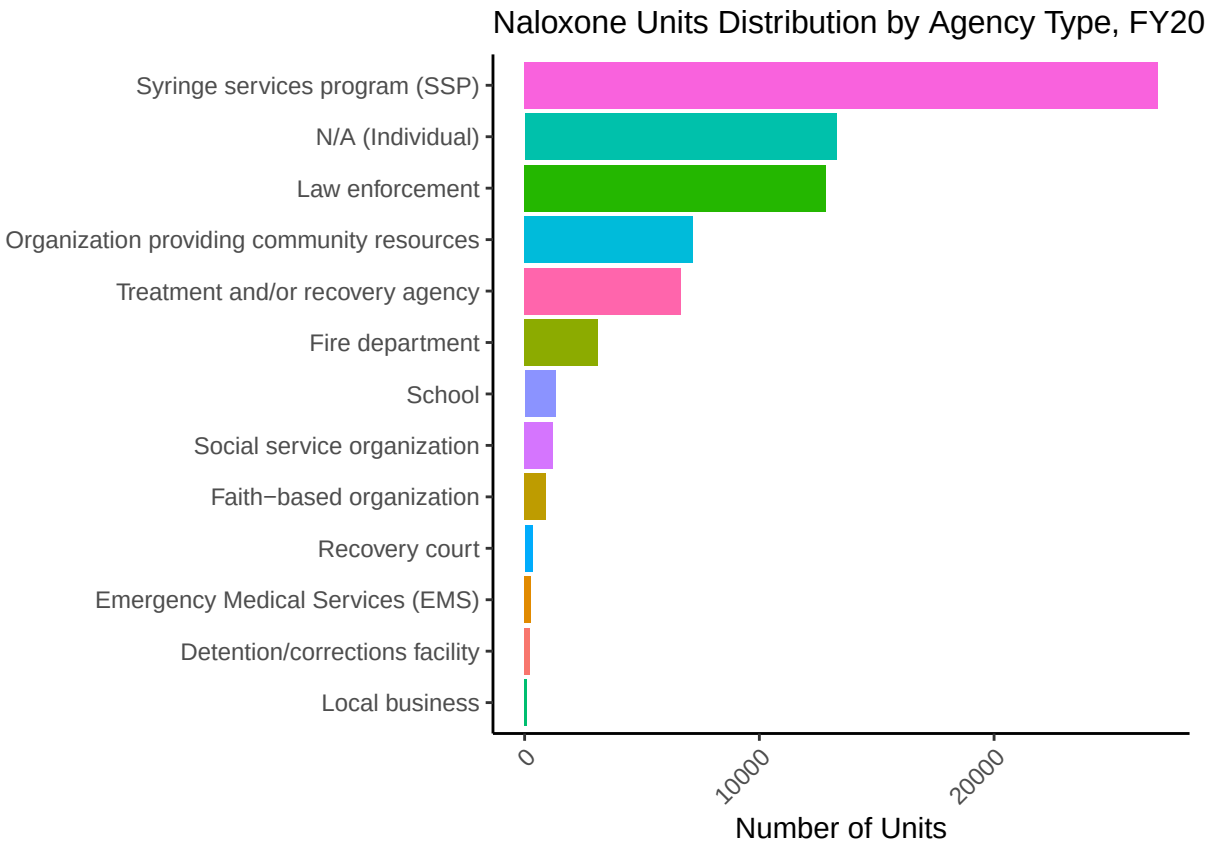
```
#getting closer - need to adjust where the labels start/end now - and remove the gray grid/background - I
agenplot<-ggplot(agencyplot, aes(x=fct_reorder(typeofagencyutilizewhatsontheage, count), y=count, fill=typ
    geom_bar(stat="identity")+
        theme_classic()+
    theme(legend.position="none", axis.text.x=element_text(angle=45, hjust=1), plot.title=element_text(siz
    xlab("")+ylab("Count")+
    ggtitle("Naloxone Distribution by Agency Type, FY2021")+
    coord_flip()
agenplot
```



These are the counts of distributions, but they don't give me any insight into just how MUCH naloxone agencies are receiving. I'm going to summarize the data in a different way, and pull that data.

```
unitsbyagency<-fy21data%>%
    group_by(typeofagencyutilizewhatsontheage)%>%
    summarize(units=sum(unitsdistr))
#LOOK at the resulting dataset. Does it look how you want it to? If so, you can go ahead and plot it.
```

```
unitsplot<-ggplot(unitsbyagency, aes(x=fct_reorder(typeofagencyutilizewhatsontheage, units), y=units, fill
    geom_bar(stat="identity")+
        theme_classic()+
    theme(legend.position="none", axis.text.x=element_text(angle=45, hjust=1), plot.title=element_text(siz
    xlab("")+ylab("Number of Units")+
    ggtitle("Naloxone Units Distribution by Agency Type, FY2021")+
    coord_flip()
unitsplot
```


Naloxone Units Distribution by Agency Type, FY20
Syringe services program (SSP)
N/A (Individual)
Law enforcement
Organization providing community resources
Treatment and/or recovery agency
Fire department
School
Social service organization
Faith−based organization
Recovery court
Emergency Medical Services (EMS)
Detention/corrections facility
Local business
0
10000
20000
Number of Units