Covid19 vs US county to MMSA

Amanda Kimball

11/3/2020

This markdown will review and clean the Center for Disease Controls Covid19 by county datasets (2020). Here I will read in the data and then review the first 5 elements. It shows that the County FIPS code, County Name, State, and State FIPS code are the first 4 elements. This is followed by a date (Jan. 22, 2020 is displayed Only) with the cumulative total cases for the county in question. The death data is in the same format.

covid19cases <- read.csv("covid\_confirmed\_usafacts.csv", stringsAsFactors = FALSE)  
covid19deaths <- read.csv("covid\_deaths\_usafacts.csv", stringsAsFactors = FALSE)  
  
str(covid19cases[1:5])

## 'data.frame': 3195 obs. of 5 variables:  
## $ ï..countyFIPS: int 0 1001 1003 1005 1007 1009 1011 1013 1015 1017 ...  
## $ County.Name : chr "Statewide Unallocated" "Autauga County" "Baldwin County" "Barbour County" ...  
## $ State : chr "AL" "AL" "AL" "AL" ...  
## $ stateFIPS : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ X1.22.20 : int 0 0 0 0 0 0 0 0 0 0 ...

I review to see if any of the values are N/A and find the data set fully intact.This was expected since it was downloaded from the CDC’s visualization page.

covid19deaths[is.na(covid19deaths),]

## <0 rows> (or 0-length row.names)

covid19cases[is.na(covid19cases),]

## <0 rows> (or 0-length row.names)

I will create a data frame with all of the necessary attributes from the 2 data sets. I choose to select the last day in September 2020 for the analysis, so that most of the death data is fully updated. This avoids most of the delay associated with the collection of death certificate data.

covid19 <- data.frame("County" = as.character(covid19cases$County.Name),  
 "State" = as.character(covid19cases$State),   
 "ST\_FIPS" = covid19cases$stateFIPS,   
 "Cases" = covid19cases$X9.30.20,   
 "Deaths" = covid19deaths$X9.30.20,  
 stringsAsFactors = FALSE)  
summary(covid19)

## County State ST\_FIPS Cases   
## Length:3195 Length:3195 Min. : 1.00 Min. : 0   
## Class :character Class :character 1st Qu.:18.00 1st Qu.: 127   
## Mode :character Mode :character Median :29.00 Median : 422   
## Mean :30.25 Mean : 2246   
## 3rd Qu.:45.00 3rd Qu.: 1279   
## Max. :56.00 Max. :270299   
## Deaths   
## Min. : 0.00   
## 1st Qu.: 1.00   
## Median : 7.00   
## Mean : 64.21   
## 3rd Qu.: 28.00   
## Max. :7329.00

I downloaded the delineation data from US Census Bureau (2020) to obtain a cross reference between county data in the CDC dataset and MMSA data in the BRFSS dataset that I will be using at a later date. I renamed the headers to match up with the CDC dataset and created a new column for the MMSA without the state which is consistent with the BRFSS dataset.

county\_MMSA <- read.csv("County to MMSA list1\_2020.csv", stringsAsFactors = FALSE)  
names(county\_MMSA) <- c("MMSA\_State", "County", "State", "ST\_FIPS", "County\_FIPS", "Central\_Outlying")  
county\_MMSA$MMSA <- gsub(",.\*","", county\_MMSA$MMSA\_State)  
str(county\_MMSA)

## 'data.frame': 1916 obs. of 7 variables:  
## $ MMSA\_State : chr "Aberdeen, SD" "Aberdeen, SD" "Aberdeen, WA" "Abilene, TX" ...  
## $ County : chr "Brown County" "Edmunds County" "Grays Harbor County" "Callahan County" ...  
## $ State : chr "South Dakota" "South Dakota" "Washington" "Texas" ...  
## $ ST\_FIPS : int 46 46 53 48 48 48 40 26 72 72 ...  
## $ County\_FIPS : int 13 45 27 59 253 441 123 91 3 5 ...  
## $ Central\_Outlying: chr "Central" "Outlying" "Central" "Outlying" ...  
## $ MMSA : chr "Aberdeen" "Aberdeen" "Aberdeen" "Abilene" ...

I summed up the cases and deaths as a check to insure that everything was in the right order of magnitude (compared with current reported case and death data for the US) and so I could further manipulate the data and then verify that nothing was lost or changed during the analysis. These values are consistent with the current reports of covid19 on the CDC website (Covid-19, 2020).

sum(covid19$Cases)

## [1] 7174631

sum(covid19$Deaths)

## [1] 205148

I will first merge the covid-19 counts with the MMSA delineation data. I do that on county and also state FIPS code since those are the only 2 in common columns. Note state is abbreviated in the Covid-19 (2020) dataset. This will give me covid-19 counts by MMSA region.

covid\_location <- merge(covid19, county\_MMSA, all.x = TRUE, by = c("County","ST\_FIPS"))  
summary(covid\_location)

## County ST\_FIPS State.x Cases   
## Length:3195 Min. : 1.00 Length:3195 Min. : 0   
## Class :character 1st Qu.:18.00 Class :character 1st Qu.: 127   
## Mode :character Median :29.00 Mode :character Median : 422   
## Mean :30.25 Mean : 2246   
## 3rd Qu.:45.00 3rd Qu.: 1279   
## Max. :56.00 Max. :270299   
##   
## Deaths MMSA\_State State.y County\_FIPS   
## Min. : 0.00 Length:3195 Length:3195 Min. : 1.00   
## 1st Qu.: 1.00 Class :character Class :character 1st Qu.: 33.00   
## Median : 7.00 Mode :character Mode :character Median : 75.00   
## Mean : 64.21 Mean : 97.82   
## 3rd Qu.: 28.00 3rd Qu.:127.00   
## Max. :7329.00 Max. :840.00   
## NA's :1377   
## Central\_Outlying MMSA   
## Length:3195 Length:3195   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

The counts remain the same after the merge, but there are some N/As in the dataset.

sum(covid\_location$Cases)

## [1] 7174631

sum(covid\_location$Deaths)

## [1] 205148

There are 1377 counties which are not included in the MMSA units and will be removed from my listing. These will not align with the data from the BFRSS dataset and can not be used for this analysis. They will be removed during the location aggregation that I perform next.

covid\_location[is.na(covid\_location$County\_FIPS),]

## Cases Deaths MMSA\_State State.y County\_FIPS Central\_Outlying MMSA  
## 1 591 13 <NA> <NA> NA <NA> <NA>  
## 3 1182 19 <NA> <NA> NA <NA> <NA>  
## 5 74 1 <NA> <NA> NA <NA> <NA>  
## 6 316 19 <NA> <NA> NA <NA> <NA>  
## 8 639 10 <NA> <NA> NA <NA> <NA>  
(removed for brevity)  
## 3188 332 6 <NA> <NA> NA <NA> <NA>  
## 3190 81 2 <NA> <NA> NA <NA> <NA>  
## 3192 125 1 <NA> <NA> NA <NA> <NA>  
## 3194 369 16 <NA> <NA> NA <NA> <NA>  
## 3195 63 0 <NA> <NA> NA <NA> <NA>

The aggregation functions will sum the covid cases and deaths by location, then merge them back together by MMSA region.

covid\_case\_location <- aggregate(covid\_location$Cases~covid\_location$MMSA, FUN = sum)  
covid\_death\_location <- aggregate(covid\_location$Deaths~covid\_location$MMSA, FUN = sum)  
  
covid\_data <- merge(x = covid\_case\_location, y = covid\_death\_location, by = "covid\_location$MMSA")  
names(covid\_data) <- c("MMSA\_NAME", "CASES", "DEATHS")

The sum of the cases and deaths are now decreased by the counties not listed in the MMSA regions.

sum(covid\_data$CASES)

## [1] 6688798

sum(covid\_data$DEATHS)

## [1] 194570

I calculate the death\_rate as deaths per cases in the region.

covid\_data <- covid\_data[covid\_data$CASES>0,] #otherwise the death rate blows up due to divide by 0.  
covid\_data$death\_rate <- 100\*covid\_data$DEATHS/covid\_data$CASES  
summary(covid\_data)

## MMSA\_NAME CASES DEATHS death\_rate   
## Length:856 Min. : 22.0 Min. : 0.00 Min. :0.000   
## Class :character 1st Qu.: 636.2 1st Qu.: 8.00 1st Qu.:1.031   
## Mode :character Median : 1415.0 Median : 27.00 Median :1.757   
## Mean : 7814.0 Mean : 227.30 Mean :2.197   
## 3rd Qu.: 4343.2 3rd Qu.: 98.25 3rd Qu.:2.947   
## Max. :556299.0 Max. :43438.00 Max. :9.655

I will export the data so that I can create GIS Maps in Tableau.

write.csv(covid\_data, 'covid\_MMSA.csv')

References:

Unknown.(2020). Coronavirus disease 2019 (COVID-19): United States COVID-19 cases and deaths by County. Centers for Disease COntrol and Prevention. Retrieved October 28, 2020 from: <http://covid.cdc.gov/covid-data-tracker/#county-map>

Unknown. (2020). Core based statistical areas (CBSAs), metropolitan divisions, and combined statistical areas (CSAs), March 2020. United States Census Bureau. Obtained November 3, 2020 from: <https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html>