ANA 515 Assignment 2

Larissa

2022-06-19

## Description of The Data

The dataset is a combination data from the Centers for Disease Control and Prevention’s Behavioral Risk Factor Surveillance System (BRFSS) and the data from Kaiser Family foundation. BRFSS includes a collection of health-related surveys of more than 400,000 Americans conducted each year, while Kaiser Family foundation provides the data on the number of hospitals and intensive care unit beds in each US county.This dataset can help identify the number of people who are at high risk of becoming seriously ill from COVID-19 per ICU bed or per hospital available in each metropolitan area, micropolitan area, or metropolitan division listed in the dataset, so they can help the area to prepare in facing COVID-19. They may take necessary actions such as adding some ICU facilities or doctors in the area.

The file is saved as csv file, and it is comma delimited.

## Reading the data into R

I used the read\_csv function from readr package to read in the data

library (readr)  
url<- "https://raw.githubusercontent.com/fivethirtyeight/data/master/covid-geography/mmsa-icu-beds.csv"   
covid<-read\_csv(url)

## Rows: 136 Columns: 7  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): MMSA, total\_percent\_at\_risk  
## dbl (5): high\_risk\_per\_ICU\_bed, high\_risk\_per\_hospital, icu\_beds, hospitals,...  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Clean the data

Separating the area name and state name into separate column.

library (tidyr)  
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

library(tidyverse)

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

## v ggplot2 3.3.6 v stringr 1.4.0  
## v tibble 3.1.7 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

covid\_clean <- covid %>%  
separate(MMSA,sep=",",into = c("Area","State"))

## Warning: Expected 2 pieces. Additional pieces discarded in 17 rows [7, 11, 14,  
## 26, 30, 32, 33, 34, 39, 41, 47, 48, 51, 59, 87, 92, 94].

## Characteristics of the data

This dataframe has 136 rows and 8 columns. The names of the columns and a brief description of each are in the table below:

Column <- colnames(covid\_clean)  
Description <- c("The name of the metropolitan area, micropolitan area or metropolitan division available in the CDC’s BRFSS","State Name","The percent of individuals in that area that are at high risk of becoming seriously ill from COVID-19, per CDC’s BRFSS","The number of high risk individuals per ICU bed in that area","The number of high risk individuals per hospital in that area","The number of ICU beds in the area, based on the Kaiser Family Foundation’s data","The number of hospitals in the area, based on the Kaiser Family Foundation’s data","The total number of high risk individuals in the area, per CDC’s BRFSS")  
Covid\_table <-data.frame(Column,Description)  
Covid\_table

## Column  
## 1 Area  
## 2 State  
## 3 total\_percent\_at\_risk  
## 4 high\_risk\_per\_ICU\_bed  
## 5 high\_risk\_per\_hospital  
## 6 icu\_beds  
## 7 hospitals  
## 8 total\_at\_risk  
## Description  
## 1 The name of the metropolitan area, micropolitan area or metropolitan division available in the CDC’s BRFSS  
## 2 State Name  
## 3 The percent of individuals in that area that are at high risk of becoming seriously ill from COVID-19, per CDC’s BRFSS  
## 4 The number of high risk individuals per ICU bed in that area  
## 5 The number of high risk individuals per hospital in that area  
## 6 The number of ICU beds in the area, based on the Kaiser Family Foundation’s data  
## 7 The number of hospitals in the area, based on the Kaiser Family Foundation’s data  
## 8 The total number of high risk individuals in the area, per CDC’s BRFSS

knitr::kable(Covid\_table,"pipe",col.name= c("Column name","Description") , align =c("l","l"))

| Column name | Description |
| --- | --- |
| Area | The name of the metropolitan area, micropolitan area or metropolitan division available in the CDC’s BRFSS |
| State | State Name |
| total\_percent\_at\_risk | The percent of individuals in that area that are at high risk of becoming seriously ill from COVID-19, per CDC’s BRFSS |
| high\_risk\_per\_ICU\_bed | The number of high risk individuals per ICU bed in that area |
| high\_risk\_per\_hospital | The number of high risk individuals per hospital in that area |
| icu\_beds | The number of ICU beds in the area, based on the Kaiser Family Foundation’s data |
| hospitals | The number of hospitals in the area, based on the Kaiser Family Foundation’s data |
| total\_at\_risk | The total number of high risk individuals in the area, per CDC’s BRFSS |

## Summary statistics

Covid\_subset <- covid\_clean %>%  
 select(icu\_beds,hospitals,total\_at\_risk)   
  
Summary <- summary(Covid\_subset)  
Summary

## icu\_beds hospitals total\_at\_risk   
## Min. : 8.0 Min. : 1.00 Min. : 17942   
## 1st Qu.: 89.5 1st Qu.: 5.00 1st Qu.: 158748   
## Median : 221.0 Median : 9.00 Median : 396082   
## Mean : 360.2 Mean : 13.76 Mean : 667189   
## 3rd Qu.: 489.5 3rd Qu.: 18.00 3rd Qu.: 932793   
## Max. :2777.0 Max. :100.00 Max. :6165102   
## NA's :1 NA's :1

icu\_mean <-mean(Covid\_subset$icu\_beds,na.rm=TRUE)  
icu\_min <-min(Covid\_subset$icu\_beds,na.rm=TRUE)  
icu\_max <-max(Covid\_subset$icu\_beds,na.rm=TRUE)  
icu\_missing <- sum(is.na(Covid\_subset$icu\_beds))  
hospital\_mean <-mean(Covid\_subset$hospitals,na.rm=TRUE)  
hospital\_min <-min(Covid\_subset$hospitals,na.rm=TRUE)  
hospital\_max <-max(Covid\_subset$hospitals,na.rm=TRUE)  
hospital\_missing <- sum(is.na(Covid\_subset$hospitals))  
at\_risk\_mean <-mean(Covid\_subset$total\_at\_risk,na.rm=TRUE)  
at\_risk\_min <-min(Covid\_subset$total\_at\_risk,na.rm=TRUE)  
at\_risk\_max <-max(Covid\_subset$total\_at\_risk,na.rm=TRUE)  
at\_risk\_missing <- sum(is.na(Covid\_subset$total\_at\_risk))

The summary statistic of the icu\_beds, hospitals and total\_at\_risk are as follow:

Icu\_beds

Min : 8, Max : 2777, Mean : 360.1851852, Missing value : 1

hospitals

Min : 1, Max : 100, Mean : 13.762963, Missing value : 1

total\_at\_risk

Min : 1.794153^{4}, Max : 6.1651017^{6}, Mean : 6.6718873^{5}, Missing value : 0