Lab 4: 2019 Novel Coronavirus

Name: Sal Figueroa

2025-04-19

Contents

Github Repository

Holds all related files

github: Figgs0bit-Lab4 (https://github.com/Figgs0bit/CSIT165-Lab4.git)

Required data sets

2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by John Hopkins CSSE

The following data set is operated by the John Hopkins University Center for Systems Science and Engineering (JHU CSSE). Data sets include daily time series CSV summary tables for confirmations and deaths associated with COVID-19. Lat and Long refer to coordinate references for the data field. Date fields are stored in MM/DD/YYYY format. For this laboratory, we will be use confirmed and deaths data sets for the US.

Project Objectives

Before beginning your objectives in your final document, please state which day you downloaded the data sets on for analysis. For this laboratory, use the packages described in Module 10 and 11 to accomplish each objective where appropriate. The surgeon general for the United States recently created a new data science initiative, CSIT-165, that uses data science to characterize pandemic diseases. CSIT-165 disseminates data driven analyses to state governors. You are a data scientist for CSIT-165 and it is up to you and you alone to manipulate and visualize COVID-19 data for disease control.

Objective 1

Which state has the most confirmed cases of COVID-19? Which state has the most deaths?

```
#cat("Test VAR - Confirmed(row, col):",rowmax_Conf,",",colmax_Conf,", Deaths(row, col):",rowmax_Deaths,
#[j,6] County column
#[j,7] State column
#[j,1155] last column
#rowmax_Conf <- nrow(Confirmed_US[,]) #limit Qty of Rows 289
```

```
#colmax_Conf <- ncol(Confirmed_US[,]) #limit Qty of Rows 289
#rowmax_Deaths <- nrow(Deaths_US[,]) #limit Qty of Rows 289
#colmax_Deaths <- ncol(Deaths_US[,]) #limit Qty of Rows 289

#Deaths_US[i,j] #[3342, 1154]
#Confirmed_US[i,j] #[3342, 1155]


# removes NA values
Deaths_US[is.na(Deaths_US)] <- 0
Confirmed_US[is.na(Confirmed_US)] <- 0

V_State <- c() #Will Hold all US States and territories.
V_USconfirmed <- c() #Will Hold all entries of confirmed_US
V_USdeaths <- c() #Will Hold all entries of confirmed_US

sumC <- 0 #resets confirmed_global Sum C after each country multiple
sumD <- 0 #resets deaths_global Sum D after each country multiple
CountyCol <- 7


#creates a data frame with country and either deaths_global, confirmed_global values.
df_comfirmed <- c()
df_deaths <- c()
#Data is broken down by county or territory, script must now sort and consolidate common states.
#The following if else ladder sorts and filters the duplicate
#-counties and combines the data for a common region.
for(j in 1:rowmax) #Iterate through rows 3342
{
  #first if conditional is need as the the first row will
  #-not fall through the remaining if statements.


  if(j == 1)
    {
      #Adds the incremented value into each respective vector
      V_State <- c(V_State, Confirmed_US[j,CountyCol])#adds State or territory
      V_USconfirmed <- c(V_USconfirmed, Confirmed_US[j,colmax])#adds US confirmed
      V_USdeaths <- c(V_USdeaths, Deaths_US[j,colmax])#adds US Deaths

    }
  #The following (2) else if statements test and sort through duplicate country
  #-entries and then sums them together, respectively.
  else if(Confirmed_US[j,CountyCol] %in% Confirmed_US[j+1,CountyCol])#conditional for strings
    {
      #When a duplicate country entry is encountered values
      #-are summed up for Confirmed and death cases
      sumC <- sumC + Confirmed_US[j,colmax]
      sumD <- sumD + Deaths_US[j,colmax]

    }
  #conditional statement for strings
  else if((Confirmed_US[j,CountyCol] %in% Confirmed_US[j-1,CountyCol]) && (!(Confirmed_US[j,CountyCol])
```

```
    {
      #When a duplicate country entry is encountered values
      #-are summed up for Confirmed and death cases
      sumC <- sumC + Confirmed_US[j,colmax]
      sumD <- sumD + Deaths_US[j,colmax]

      #loads the shorten results with the common countries combined
      V_State <- c(V_State, Confirmed_US[j,CountyCol])#adds State or territory
      V_USconfirmed <- c(V_USconfirmed, sumC)#adds US confirmed
      V_USdeaths <- c(V_USdeaths, sumD)#adds US Deaths

      #resets the following sum variables to zero for the next duplicate country
      sumC <- 0
      sumD <- 0

    }
  #The last else statement is where countries without duplicates are loaded
  else
    {
      #adds deaths_global, confirmed_global to new vectors
      V_State <- c(V_State, Confirmed_US[j,CountyCol])#adds State or territory
      V_USconfirmed <- c(V_USconfirmed, Confirmed_US[j,colmax])#adds US confirmed
      V_USdeaths <- c(V_USdeaths, Deaths_US[j,colmax])#adds US Deaths

    }

}
#creates and loads data frame with the shorten and consolidated vectors
df_comfirmed <- data.frame(V_State, V_USconfirmed)
df_deaths <- data.frame(V_State, V_USdeaths)

#Reorders the Data-frame in decreasing order
#df_C_decreasing <- df_comfirmed[order(df_comfirmed$V_USconfirmed, decreasing = TRUE),]
#df_D_decreasing <- df_deaths[order(df_deaths$V_USdeaths, decreasing = TRUE),]

#View(df_C_decreasing)
#View(df_D_decreasing)
```

Objective 2

Create a new table called current_tally by merging confirmed and deaths containing only the columns iso2, Province_State, Admin2, Lat, Long_, and today's count for confirmed and deaths. Rename current_tally's columns as Country, State, City, Lat, Long, Confirmed, and Deaths respectively. Filter current_tally for data points where the country is the US. Please print current_tally as a tibble.

```
#current_tally <- Confirmed_US %>%
#                 merge(Deaths_US, by = c("iso2", "Province_State", "Admin2", "Lat", "Long_", "X2.29.20

current_tally <- cbind(Confirmed_US[2],Confirmed_US[7],Confirmed_US[6],Confirmed_US[9],Confirmed_US[10]
colnames(current_tally) <- c("Country","State","City","Lat","Long","Confirmed","Deaths")
current_tally <- tibble(current_tally)
#print(current_tally)
```

Objective 3

In current_tally, create a new variable, unaccounted, as the absolute difference between confirmed and deaths for each city. Determine the number of unaccounted for each state using grouped operation.

```
current_tally <- current_tally %>%
        mutate('unaccounted' = abs(Confirmed - Deaths))

#view(current_tally)

unaccounted_tally <- current_tally %>%
        group_by(State) %>%
        summarise(Total_unaccounted = sum(unaccounted))

#view(unaccounted_tally)
```

using the code above, use the package kable and create a table showing the top 5 state with the most Total_unaccounted values

Objective 4

Using the package kable, create a table showing the top 5 unaccounted states.

```
# Display the top 5 states

Decrease_unaccounted <- unaccounted_tally[order(unaccounted_tally$Total_unaccounted, decreasing = TRUE)

top5_unaccounted <- head(Decrease_unaccounted, 5)

kable(top5_unaccounted, caption = "Top 5 States by Total Unaccounted COVID Cases")
```

Table 1: Top 5 States by Total Unaccounted COVID Cases

State	Total_unaccounted
California	12028540
Texas	8372830
Florida	7487740
New York	6718707
Illinois	4051794

Objective 5

Create any visualization you like using the data sets used in this laboratory. Provide your visualization with a title and custom axis labels. Describe how this visualization conveys information that is not readily apparent from looking at the data alone.

```
DF_Scatter <- cbind(Confirmed_US[7],Confirmed_US[6],PopDeaths_US[12],Confirmed_US[colmax],Deaths_US[colr
colnames(DF_Scatter) <- c("State","City","Population","Confirmed","Deaths")

#view(DF_Scatter)

DF_Scatter <- DF_Scatter %>%
        mutate('unaccounted' = abs(Confirmed - Deaths))

#view(DF_Scatter)
```

```r
unaccounted_DF_Scatter <- DF_Scatter %>%
        group_by(State) %>% #, Population
        summarise(Total_unaccounted_DF_Scatter = sum(unaccounted),Total_StatePopulation = sum(Populatio

view(unaccounted_DF_Scatter)
```

```r
# Example dataset (you can skip this chunk if using your own data)
data <- data.frame(unaccounted_DF_Scatter)

StateTeritory <- data[1]

#State, Total_unaccounted_DF_Scatter, Total_StatePopulation

ggplot(data, aes(x = Total_unaccounted_DF_Scatter, y = Total_StatePopulation, color = State)) +
  geom_point(alpha = 0.7, size = 1) +
  labs(
    title = "Decrease in Unaccounted Cases vs County Population",
    x = "County Population (popCounties)",
    y = "Decrease in Unaccounted Cases",
    color = "State/Territory") + theme(
    # Title alignment. Number from 0 (left) to 1 (right)
    legend.title.align = 1,
    # Text label alignment. Number from 0 (left) to 1 (right)
    legend.text.align = 0,
    # Legend position: right, left, bottom, top, none
    legend.position = "top",
    # Margin around each legend
    legend.margin = margin(0.2, 0.2, 0.2, 0.2, "cm"),
    # Layout of items in legends ("horizontal" or "vertical")
    legend.direction = NULL,
    # Positioning legend inside or outside plot
    # ("center" or two-element numeric vector)
    legend.justification = "center",

    legend.key.size = unit(1.2, "lines"),     # key size (unit)
    legend.key.height = NULL,                 # key height (unit)
    legend.key.width = NULL,                  # key width (unit)

    legend.spacing = unit(0.4, "cm"),
    legend.spacing.x = NULL,                  # Horizontal spacing
    legend.spacing.y = NULL,                  # Vertical spacing

    # Arrangement of multiple legends ("horizontal" or "vertical")
    legend.box = NULL,
    # Margins around the full legend area
    legend.box.margin = margin(0, 0, 0, 0, "cm"),
    # Background of legend area: element_rect()
    legend.box.background = element_blank(),
    # The spacing between the plotting area and the legend box
    legend.box.spacing = unit(0.4, "cm"))
```

```
## Warning: The `legend.title.align` argument of `theme()` is deprecated as of ggplot2
## 3.5.0.
## i Please use theme(legend.title = element_text(hjust)) instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The `legend.text.align` argument of `theme()` is deprecated as of ggplot2
## 3.5.0.
## i Please use theme(legend.text = element_text(hjust)) instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```