Final Project: Visualizing Coronavirus Data

Name: Sal Figueroa

2025-05-15

Contents


```
#### Load necessary library ####
packages <- c("knitr", "kableExtra", "magrittr", "readr", "geosphere")

install_me <- packages[!(packages %in% installed.packages()[, "Package"])]
if (length(install_me)) install.packages(install_me)

library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.4.3
```

```
library(magrittr)
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.4.3
```

```
library(geosphere)
```

```
## Warning: package 'geosphere' was built under R version 4.4.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(treemapify)
```

```
## Warning: package 'treemapify' was built under R version 4.4.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
## Warning: package 'tibble' was built under R version 4.4.3
## Warning: package 'tidyr' was built under R version 4.4.3
## Warning: package 'purrr' was built under R version 4.4.3
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
## Warning: package 'forcats' was built under R version 4.4.3
```

```
## Warning: package 'lubridate' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v stringr   1.5.1
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggfittext)
```

```
## Warning: package 'ggfittext' was built under R version 4.4.3
```

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.4.3
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(dplyr)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.4.3
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.4.3
```

```r
library(scales)
library(stringr)

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

Data Load

```r
#print("Lab represents data download May 4th, 2025")

#time_series_covid19_confirmed_US.csv 1154 columns, 289 rows.
#Confirmed_Global <- as.data.frame(read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/

#Serve as data frame length and width variables if needed
#rowmax <- nrow(Confirmed_Global[,]) #limit Qty of Rows 289
#colmax <- ncol(Confirmed_Global[,]) #limit Qty of Rows 1147

#loads column headers in data.frame
#NameCol <- as.data.frame(colnames(Confirmed_Global))

#Var holds the length of the NameCol data.frame
#MAXNameCol <- nrow(NameCol)

#Name of the final column name
#FinalColName <- NameCol[MAXNameCol,]

#Loads the URL of the .RData file in the GitHub repository
#GitHuburl <- "https://github.com/Figgs0bit/CSIT165-ModData/raw/refs/heads/master/Module-13/human_prote
#names the download for the *.Rdata file.
#RdataFile <- "human_proteins.RData"

#Download the file into the working directory, Note: mode-WB needed to download binary .rdata file.
#download.file(GitHuburl, RdataFile, mode = "wb")
#Load the .RData file from working directory into the R environment
#load(RdataFile)
```

Github Repository

Repository holds all related files to Project

github: Figgs0bit-Final Project (https://github.com/Figgs0bit/CSIT165-CovidGroupProj-2)

Required data sets

Data for 2019 Novel Coronavirus is operated by the John Hopkins University Center for Systems Science and Engineering (JHU CSSE). Data includes daily time series CSV summary tables, including confirmations, recoveries, and deaths. Country/region are countries/regions hat conform to World Health Organization (WHO). Lat and Long refer to coordinates references for the user. Date fields are stored in MM/DD/YYYY format.

For this project, we will use ALL of the data sets provided in this GitHub repository. These include global data sets for COVID-19 associated confirmations and deaths as well as COVID-19 data sets associated with confirmations and deaths for cities in the US.

This lab represents data downloaded on 04/xx/2025

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

-time_series_covid19_deaths_global.csv

-time_series_covid19_deaths_US.csv

-time_series_covid19_confirmed_global.csv

-time_series_covid19_confirmed_US.csv

Instructions

Before beginning your objectives in your final document, please state which day you downloaded the data sets on for analysis. The objectives for this lab will cumulatively cover many subjects discussed in this course and will also contain an objective for manipulating strings. 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.

Objectives

This project will encompass many of the lessons we have learned throughout the course, including interactive visualizations. RMarkdown files must be written such that each time you render the document it will download the necessary data sets for analysis. Please render the RMarkdown file the day it is due to reflect the most recent data sets. With this added functionality, your code must be able to analyze the datasets regardless of the date you render your document. Unlike others projects in the past, you will have the ability to solve these problems using any method you choose. Be careful, however, with the methods you use as you will be graded on the appropriateness of your solution and how well you execute your desired algorithm. If there appears to be a logic fail in how you executed your code, you will be penalized. This is an opportunity to really showcase your new found skills in data science with R!

Objective 1 - *Global Map*

You are tasked to create a world map to gain an appreciation for where the most occurrences of COVID-19 confirmations and deaths are located.

Create this map using leaflet for the most recent date as shown below. For this map, sum the confirmations and deaths of provinces into one value to depict the total number for the country they belong to. When creating a marker for each country in the map, calculate lat and long as the mean values for the provinces that make up each country.

Customize the map to reflect the differences in magnitude for confirmations and deaths. In the example map below, circle markers that are blue represent low values, gray represents neutral values, and red represents high values. Low, middle, and high values were categorized to aesthetically map the markers based on their probabilistic distribution using the quartile function. You may use any method you like so that it is logical and allows visualization of value intensity. As well, customize the map to include hover labels that indicate country names and popup labels to show the value of confirmations and deaths for that country. For extra help using leaflet, consult this website along with the information provided in your textbooks.

Objective 2 - *Narrowing Down Hot Spots*

Seeing the global map of COVID-19 cases results in the stark realization that some countries are more affected than others. In order to narrow down your studies, create a table using kable from knitr listing the top countries as far as confirmations and deaths (sum values for provinces of the same country into one value and show the country only). Now that we are using RMarkdown to create HTML files, we have much more options for how we display our table. For reference on how to customize tables using knitr, visit this website. Consult the table below for an example of a customized table ranking cases by country. While it is not required to replicate this table exactly, it would be a fantastic challenge to show off your knitr prowess.

##Objective 3 - *Zooming Into Our State After reading the top tables, you are stunned! The US overtakes every other country in terms of COVID-19 confirmations. As such, you are concerned about the state you live in and would like to understand how COVID-19 events have shaped the trajectory of the disease. Create two scatter plots to gain a better understanding. The first scatter plot should be California's trajectory for confirmations. The second scatter plot should show California's top three city trajectories for confirmations. You are interested in studying how the vaccine affected the number of confirmations. The Moderna vaccine was first available as an emergency use authorized (EUA) vaccine and required two shots spaced six weeks apart. Indicate on the plots the day the second dosage was given to those that received the first dosage the day Moderna was EUA (January 29th, 2021). As a diligent scientist that knows that new COVID variants have mutations in the spike protein (the region that the vaccine was developed for), you also want to study how confirmation rates change as new variants become the dominant infectious strain. Indicate on the plots when the delta and omicron variants became the dominant strain in California (May 11th, 2021 and November 26th, 2021 respectively). In the example below, the function plot_grid from the R package cowplot was to organize the graphs into a grid to more easily compare statewide vs top city plots.*

Objective 4 - ***Digging Deeper***

Although these plots do not tell the whole story, they are great for helping us determine where to look. Different cities may have different populations, population densities, cultural discrepancies, compliance, and city regulations to name a few. We will explore the role of population on these metrics using visualizations. Arrange two scatter plots using cowplot's plot_grid to show the relationship between population and confirmed counts as well as death counts and confirmed counts. You will need to use a log transform on all variables to show such a relationship. Please consult the example below for an idea of what this may look like. From these graphs we can see that population greatly affects confirmations and deaths. This coincides with our plots above as the population of Los Angeles is 301% greater than San Diego's population and 406% greater than Riverside's population!

Instructions *The Global Health Initiative has recently employed a new data science response team, CSIT-165, that uses data science to characterize pandemic diseases. CSIT-165 disseminates data driven analyses to global and local decision makers.*

CSIT-165 is a conglomerate comprised of two fabricated entities: World Health Organization (WHO) and U.S. Pandemic Response Team (USPRT). Your and your partner's role is to play a data scientist from one of these two entities. Discuss with your partner to decide who will be part of WHO and USPRT.

Getting Started *One project member per group must create a new repository on GitHub. Initialize this repository with a readme.md file that lists each member of the group. If your group decides to collaborate using a centralized workflow (recommended), then the project member that created the repository must declare their partners as collaborators in GitHub. Each project member will clone this repository onto their machine using RStudio. In RStudio, create a project from version control with GitHub using the HTTP address of the repository created by project member.*

All project members must first contribute to analyses by uploading data sets respective to the entity they belong to in the CSIT-165 data science response team. If you belong to WHO then you are responsible for providing code necessary to download the global data sets and if you are assigned to USPRT then you are responsible for providing code necessary to download the US data sets.