Project 1

Naycari de Luna and Marc Robert Grabiel

2021-04-11

Contents

Background

The World Health Organization has recently employed a new data science initiative, *CSIT-165*, that uses data science to characterize pandemic diseases. *CSIT-165* disseminates data driven analyses to global decision makers.

CSIT-165 is a conglomerate comprised of two fabricated entities: *Global Health Union (GHU)* and *Private Diagnostic Laboratories (PDL)*. Your and your partner's role is to play a data scientist from one of these two entities.

Data

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

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.

Project Objectives

Objective 1

```r
confirmed_origin <- confirmed[which.max(confirmed$X1.22.20), c("Province.State", "Country.Region")]
confirmed_origin
```

```
##    Province.State Country.Region
## 72          Hubei          China
```

```r
confirmed_country <- as.character(confirmed_origin[[c("Country.Region")]])
confirmed_state <- as.character(confirmed_origin[[c("Province.State")]])
```

```r
deaths_origin <- deaths[which.max(deaths$X1.22.20), c("Province.State", "Country.Region")]
deaths_origin
```

```
##    Province.State Country.Region
## 72          Hubei          China
```

```r
deaths_country <- as.character(deaths_origin[[c("Country.Region")]])
deaths_state <- as.character(deaths_origin[[c("Province.State")]])

recovered_origin <- recovered[which.max(recovered$X1.22.20), c("Province.State", "Country.Region")]
recovered_origin
```

```
##    Province.State Country.Region
## 57          Hubei          China
```

```r
recovered_country <- as.character(recovered_origin[[c("Country.Region")]])
recovered_state <- as.character(recovered_origin[[c("Province.State")]])

if(identical(deaths_country, recovered_country) == identical(recovered_country, confirmed_country))
  {
  print("same country")
  }else
    {
      print("not same country")
    }
```

```
## [1] "same country"
```

```r
if(identical(deaths_state, recovered_state) == identical(recovered_state, confirmed_state))
{
  print("same state")
}else
{
  print("not same state")
}
```

```
## [1] "same state"
```

All three data sources from the GHU and PDL show Hubei, China, as the origin for Covid-19. The number of recovered, deaths, and confirmed cases were highest when this recording began on January 22nd, 2020. Although it is possible that Hubei had the most awareness (through testing and recording) our agencies believe that the sustained increase of all three case counts suggest it was in fact close to the region of Hubei, China.

Objective 2

```r
i <- 0
column_num <- ncol(confirmed)
column_num_values <- confirmed[column_num - i]
column_num_b4 <-confirmed[column_num - i - 1]
column_values_sum <- sum(column_num_values == 0)
column_b4_sum <- sum(column_num_b4 == 0)

for (i in 1:column_num)
{
  if(column_values_sum != column_b4_sum)
  {
    break
  }
  i <- i + 1
  column_num_values <- confirmed[column_num - i]
  column_num_b4 <-confirmed[column_num - i - 1]
  column_values_sum <- sum(column_num_values == 0)
  column_b4_sum <- sum(column_num_b4 == 0)
  column_values_sum == column_b4_sum
}
print(i)
```

```
## [1] 79
```

```
zero_values <- confirmed[confirmed$X1.20.21 == 0, ]
recent_case <- zero_values[zero_values$X1.21.21 != 0, c("Province.State", "Country.Region")]
recent_case
```

```
##     Province.State Country.Region
## 183                    Micronesia
```

The most recent case occurred 79 days before the last day entered into the dataset. The case occurred on January 21st, 2021, in Micronesia. There is no Province/State associated to our database for Micronesia. Micronesia was the only location that went from 0 cases to 1 or more cases.

Objective 3

```
locations <- confirmed[c(72, 183), c(4:1)]
Hubei <- c(locations[1,1], locations[1,2])
Micronesia <- c(locations[2,1], locations[2,2])

dist_between <- round(distm(Hubei, Micronesia)*0.000621371, digits = 2)

paste(locations[2,3], " is ", dist_between, " miles away from ", locations[1,4], ", ", locations[1,3],
```

```
## [1] "Micronesia is 2955.32 miles away from Hubei, China."
```

The origin of Covid-19 is suspected to be from Hubei, China. The most recent confirmed case we can see is from Micronesia. Micronesia is 2955.32 miles away from Hubei, China.

Objective 4

```

column_num_deaths <- ncol(deaths)
total_deaths <- subset(deaths, select = c(1, 2, column_num_deaths))
total_deaths$StateRegion <- do.call(paste0, total_deaths[1:2])
colnames(total_deaths) = c("Province.State", "Country.Region", "Total.Deaths", "StateRegion")

column_num_recovered <- ncol(recovered)
total_recovered <- subset(recovered, select = c(1, 2,column_num_recovered))
total_recovered$StateRegion <- do.call(paste0, total_recovered[1:2])
colnames(total_recovered) = c("Province.State", "Country.Region", "Total.Recovered", "StateRegion")

column_num_confirmed <- ncol(confirmed)
total_confirmed <- subset(confirmed, select = c(1, 2,column_num_confirmed))
total_confirmed$StateRegion <- do.call(paste0, total_confirmed[1:2])
```

```r
colnames(total_confirmed) = c("Province.State", "Country.Region", "Total.Confirmed", "StateRegion")

risk_score <- merge(total_deaths, total_recovered, by = "StateRegion")
risk_score <- subset(risk_score, select = c(1:4, 7))
risk_score <- merge(risk_score, total_confirmed, by = "StateRegion")
risk_score <- subset(risk_score, select = c(2:5, 8))
risk_score$Risk.Score <- risk_score$Total.Deaths / risk_score$Total.Recovered
risk_score$Death.Burden <- risk_score$Risk.Score * risk_score$Total.Confirmed

Global_Risk_Score <- sum(risk_score$Total.Deaths) / sum(risk_score$Total.Recovered)

Global_Risk_Score
```

Objective 4.1

```
## [1] 0.03825597
```

```r
head(risk_score[order(-risk_score$Risk.Score),], n = 12)
```

```
##     Province.State.x Country.Region.x Total.Deaths Total.Recovered
## 21                            Belgium        23428               0
## 163                       Netherlands        16754               0
## 165  New South Wales        Australia           54               0
## 201                            Serbia         5700               0
## 223                            Sweden        13621               0
## 242                    United Kingdom       127080               0
## 244                                US       561783               0
## 147       Martinique           France           59              98
## 254                             Yemen         1031            2027
## 218                             Spain        76328          150376
## 74                             France        97422          274401
## 160                        MS Zaandam            2               7
##     Total.Confirmed Risk.Score Death.Burden
## 21           922487        Inf          Inf
## 163         1342447        Inf          Inf
## 165            5330        Inf          Inf
## 201          639476        Inf          Inf
## 223          857401        Inf          Inf
## 242         4368045        Inf          Inf
## 244        31151495        Inf          Inf
## 147            8887  0.6020408 5.350337e+03
## 254            5276  0.5086334 2.683550e+03
## 218         3347512  0.5075810 1.699133e+06
## 74          4903965  0.3550351 1.741080e+06
## 160               9  0.2857143 2.571429e+00
```

```r
tail(risk_score[order(-risk_score$Risk.Score),], n = 22)
```

```
##                                 Province.State.x      Country.Region.x
## 6                                       Anguilla        United Kingdom
## 60                                                            Dominica
## 70                   Falkland Islands (Malvinas)        United Kingdom
## 86                                     Greenland               Denmark
## 101                                                           Holy See
## 119                                      Jiangsu                 China
## 129                                                               Laos
## 139                                        Macau                 China
## 146                                                   Marshall Islands
## 152                                                         Micronesia
## 164                                New Caledonia                France
## 170                                      Ningxia                 China
## 172                           Northern Territory             Australia
## 184                                      Qinghai                 China
## 191 Saint Helena, Ascension and Tristan da Cunha        United Kingdom
## 192                                              Saint Kitts and Nevis
## 194                    Saint Pierre and Miquelon                France
## 196                                                              Samoa
## 206                                       Shanxi                 China
## 213                                                    Solomon Islands
## 232                                        Tibet                 China
## 246                                                            Vanuatu
##     Total.Deaths Total.Recovered Total.Confirmed Risk.Score Death.Burden
## 6              0              22              25          0            0
## 60             0             159             165          0            0
## 70             0              54              60          0            0
## 86             0              31              31          0            0
## 101            0              15              27          0            0
## 119            0             708             716          0            0
## 129            0              47              49          0            0
## 139            0              48              49          0            0
## 146            0               4               4          0            0
## 152            0               1               1          0            0
## 164            0              58             121          0            0
## 170            0              75              75          0            0
## 172            0             107             112          0            0
## 184            0              18              18          0            0
## 191            0               4               4          0            0
## 192            0              44              44          0            0
## 194            0              24              24          0            0
## 196            0               2               3          0            0
## 206            0             240             248          0            0
## 213            0              18              19          0            0
## 232            0               1               1          0            0
## 246            0               1               3          0            0
```

There are 22 rows under the way the data has been defined as Province/State or Country/Region areas that have 0 risk score. They are Anguila (United Kingdom), Dominica, Falkland Islands (Malvinas, United Kingdom), Greenland (Denmark), Holy See, Jiangsu (China), Laos, Macau (China), Marshall Islands, Micronesia, New Caledonia (France), Ningxia (China), Northern Territory (Australia), Qinghai (China), Saint Helena (Ascension and Tristan da Cunha, United Kindgdom), Saint Kitts and Nevis, Saint Pierre and Miquelon (France), Samoa, Shanxi (China), Solomon Islands, Tibet (China), and Vanuatu.

The highest risk area is slightly harder to accurately define because seven areas have 0 recovered individuals reported and this makes the calculation infinite (and invalid). Most of the areas have a high number of deaths except for New South Wales (Australia). These areas are Belgium, Netherlands, New South Wales (Australia), Serbia, Sweden, United Kingdom, and US. The areas with the highest risk scores that are not infinite are Martinique (France), Yemen, Spain, France, and MS Zaandam. However, two of these areas have death and recovered counts below 100 which for me highlights problems with this parameter. These high risk areas have risk scores from 0.286-0.602. This is much higher (up to 15 times) than the global risk score, 0.038. However, because there are some countries with no reported number of recovered individuals the global risk value is slightly inflated because the global number of deaths from those countries was included. There is not a clear trend between risk score and burden score. The populations have not been standardized and so the resulting burden score can vary greatly.

One problem with this dataset is that some countries have stopped reporting or have never reported the number of individuals recovered. Sweden and the Netherlands never show a reported recovered case in the dataset. The United States, for example, stopped reporting numbers on December 14th, 2020.

```
total_deaths2 <- subset(deaths, select = c(2,column_num_deaths))
total_deaths2$Sum.total <- rowSums(total_deaths2[-1])
total_deaths2 <- aggregate(x = total_deaths2$Sum.total, by = list(total_deaths2$Country.Region), FUN = 
colnames(total_deaths2) = c("Country.Region", "Total.Deaths")

total_recovered2 <- subset(recovered, select = c(2,column_num_recovered))
total_recovered2$Sum.total <- rowSums(total_recovered2[-1])
total_recovered2 <- aggregate(x = total_recovered2$Sum.total, by = list(total_recovered2$Country.Region
colnames(total_recovered2) = c("Country.Region", "Total.Recovered")

total_confirmed2 <- subset(confirmed, select = c(2,column_num_confirmed))
total_confirmed2$Sum.total <- rowSums(total_confirmed2[-1])
total_confirmed2 <- aggregate(x = total_confirmed2$Sum.total, by = list(total_confirmed2$Country.Region
colnames(total_confirmed2) = c("Country.Region", "Total.Confirmed")

risk_score2 <- merge(total_deaths2, total_recovered2, by = "Country.Region")
risk_score2 <- merge(risk_score2, total_confirmed2, by = "Country.Region")
risk_score2$Risk.Score <- risk_score2$Total.Deaths / risk_score2$Total.Recovered
risk_score2$Death.Burden <- risk_score2$Risk.Score * risk_score2$Total.Confirmed

Global_Risk_Score2 <- sum(risk_score2$Total.Deaths) / sum(risk_score2$Total.Recovered)
Global_Risk_Score2
```

Objective 4.2

```
## [1] 0.03807731
```

```
kable(head(total_confirmed2[order(-total_confirmed2$Total.Confirmed),], n = 5))
```

	Country.Region	Total.Confirmed
184	US	31151495
24	Brazil	13445006
80	India	13358805
63	France	5001685
143	Russia	4580633

```
kable(head(total_recovered2[order(-risk_score2$Total.Recovered),], n = 5))
```

	Country.Region	Total.Recovered
80	India	12081443
24	Brazil	11739649
143	Russia	4209754
178	Turkey	3301217
86	Italy	3107069

```
kable(head(total_deaths2[order(-risk_score2$Total.Deaths),], n = 5))
```

	Country.Region	Total.Deaths
184	US	561783
24	Brazil	351334
114	Mexico	209212
80	India	169275
182	United Kingdom	127324

The top five countries with the most confirmed cases are the US, Brazil, India, France, and Russia. The top five countries with most recovered are India, Brazil, Russia, Turkey, and Italy. This list will not include the countries that never reported or stopped reporting recovered cases. The top five countries with the most deaths are the US, Brazil, Mexico, India, and the United Kingdom.

GitHub Log

```
git log --pretty=format:"%nSubject: %s%nAuthor: %aN%nDate: %aD%nBody: %b"
```

```
##
## Subject: Final version by Marc with Objective 4 draft
## Author: mrgrabiel
## Date: Sun, 11 Apr 2021 18:25:38 -0700
## Body:
##
## Subject: Time to copy, paste, and format
## Author: Naycari De Luna
## Date: Sun, 11 Apr 2021 11:32:49 -0700
## Body:
##
## Subject: NDL attempt at Obj 3
## Author: Naycari De Luna
## Date: Sun, 11 Apr 2021 11:26:19 -0700
## Body:
##
## Subject: obj 2 without loop
## Author: Naycari De Luna
## Date: Sun, 11 Apr 2021 08:09:41 -0700
## Body:
##
## Subject: Obj 2 without loops
## Author: Naycari De Luna
## Date: Sun, 11 Apr 2021 08:05:13 -0700
## Body:
##
## Subject: Draft project file for submission.  Up to obj. 3
## Author: mrgrabiel
## Date: Sat, 10 Apr 2021 16:54:34 -0700
## Body:
##
## Subject: Include all three absolute filepaths
## Author: mrgrabiel
## Date: Sat, 10 Apr 2021 14:15:40 -0700
## Body:
##
## Subject: update on obj 4
## Author: Naycari De Luna
## Date: Sat, 10 Apr 2021 10:34:19 -0700
## Body:
##
## Subject: Progress on obj 4
## Author: Naycari De Luna
## Date: Sat, 10 Apr 2021 09:37:22 -0700
## Body:
##
## Subject: complete object 1
## Author: Naycari De Luna
## Date: Thu, 8 Apr 2021 20:54:43 -0700
## Body:
##
## Subject: worked a bit on ob1 for deaths and recovery
## Author: Naycari De Luna
## Date: Wed, 7 Apr 2021 23:04:01 -0700
## Body:
##
## Subject: re-uploading correct data files
## Author: Naycari De Luna
## Date: Wed, 7 Apr 2021 21:23:16 -0700
## Body:
##
## Subject: adding covid recovered data
## Author: Naycari De Luna
## Date: Tue, 6 Apr 2021 23:29:04 -0700
## Body:
##
## Subject: adding covid death data
## Author: Naycari De Luna
## Date: Tue, 6 Apr 2021 23:27:23 -0700
## Body:
##
## Subject: Objective 3 code for distance between recent and origin
## Author: mrgrabiel
## Date: Sun, 4 Apr 2021 16:45:19 -0700
## Body:
##
## Subject: Share Objective 2 code for recent confirmed case
## Author: mrgrabiel
## Date: Sun, 4 Apr 2021 12:36:04 -0700
## Body:
##
## Subject: Share objective 1 code for confirmed cases
## Author: mrgrabiel
## Date: Sun, 4 Apr 2021 10:56:48 -0700
## Body:
##
## Subject: Add files via upload
## Author: mrgrabiel
## Date: Sat, 3 Apr 2021 18:10:22 -0700
## Body:
##
## Subject: Update README.md
## Author: ndeluna-i
## Date: Thu, 1 Apr 2021 21:54:50 -0700
## Body:
##
## Subject: Initial commit
## Author: ndeluna-i
## Date: Wed, 24 Mar 2021 19:03:15 -0700
## Body:
```