Project 1

Name: Naycari De Luna Partner: Marc 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.

```
confirmed_download <- getURL("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covi
confirmed <- read.csv(text=confirmed_download, stringsAsFactors = FALSE)

recovered_download <- getURL("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covi
recovered <- read.csv(text=recovered_download, stringsAsFactors = FALSE)

deaths_download <- getURL("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_
deaths <- read.csv(text=deaths_download, stringsAsFactors = FALSE)
```

Project Objectives

Objective 1

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

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

deaths_origin <- deaths[which.max(deaths$X1.22.20), c("Province.State", "Country.Region")]
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_country <- as.character(recovered_origin[[c("Country.Region")]])
recovered_state <- as.character(recovered_origin[[c("Province.State")]])

paste("Confirmed data indicates ", confirmed_state, ", ", confirmed_country, " as origin.", sep = "")
```

```
## [1] "Confirmed data indicates Hubei, China as origin."
```

```
paste("Death data indicates ", deaths_state, ", ", deaths_country, " as origin.", sep = "")
```

```
## [1] "Death data indicates Hubei, China as origin."
```

```
paste("Recovered data indicates ", recovered_state, ", ", recovered_country, " as origin.", sep = "")
```

```
## [1] "Recovered data indicates Hubei, China as origin."
```

```
if(identical(deaths_country, recovered_country) == identical(recovered_country, confirmed_country))
  {
  paste(recovered_country, " is the origin country.", sep = "")
  }else
    {
      paste(recovered_country, "is NOT the origin country.", sep = "")
    }
```

```
## [1] "China is the origin country."
```

```
if(identical(deaths_state, recovered_state) == identical(recovered_state, confirmed_state))
{
  paste(recovered_state, " is the origin state.", sep = "")
}else
{
  paste(recovered_state, " is NOT the origin state.", sep = "")
}
```

```
## [1] "Hubei is the origin state."
```

Confirmed cases, number deaths and number of recoveries data sets all suggest Hubei, China to be the origin of COVID-19. The max number of confirmed cases, deaths and recoveries occured in Hubei, China on the first day data was recorded; 01/22/2021. We are unable to conclude Hubei, China to be the true origin of COVID-19 since no data is available from earlier dates. An ideal case to determine origin would include data where number confirmed cases in countries and states are in a close proximity to one another. Right now, we have multiple locations with occurances of confirmed cases on 01/22/2021 that are not very close to eachother.

Objective 2

```
ncol <- ncol(confirmed)
confirmed_ordered <- arrange(confirmed, confirmed[5:ncol])
recent_first <- as.character(head(confirmed_ordered$Country.Region, n = 1))

recent <- confirmed_ordered[1,-c(1:4)]
```

```r
ncol_recent <- ncol(recent)

recent <- recent[,recent[,1:ncol_recent]!=0]

paste(recent_first, " had the most recent confirmed case on ", colnames(recent[1]), sep = "")
```

```
## [1] "Micronesia had the most recent confirmed case on X1.21.21"
```

The confirmed data set suggests Micronesia to have had the most recent first confirmed case of COVID-19 on 01/21/2021. No other countries appear to have had a first confirmed case on this same day. Interestingly, all countries in the data set have had confirmed cases of COVID-19 since the begining of data collection.

Objective 3

```r
nrow_origin <- as.numeric(which(grepl("Hubei", confirmed$Province.State)))
nrow_recent <- which(grepl("Micronesia", confirmed$Country.Region))

miles_between <- as.numeric(round(distm(as.numeric(confirmed[nrow_origin, 4:3]), as.numeric(confirmed[n

data <- data.frame(Case=c("Origin", "Recent"),
                   Country.Region = c(confirmed[nrow_origin, 1], confirmed[nrow_recent, 1]),
                   State.Province = c(confirmed[nrow_origin, 2], confirmed[nrow_recent, 2]),
                   Lat = c(confirmed[nrow_origin, 4], confirmed[nrow_recent, 4]),
                   Long = c(confirmed[nrow_origin, 3], confirmed[nrow_recent, 3]))

paste(data[2,3], " is ", miles_between, " miles away from ", data[1,2], ", ", data[1,3], ".", sep = "")
```

```
## [1] "Micronesia is 2955 miles away from Hubei, China."
```

Micronesia is the only location to have had the most recent first confirmed case of COVID-19. Micronesia is 2955 miles away from the suggested origin for COVID-19 in Hubei, China. No Provicne/State is associated with Micronesia.

Objective 4

Objective 4.1

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

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

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

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

paste("Global risk score is ", GRS, sep = "")
```

```
## [1] "Global risk score is 0.0381312754337771"
```

```r
head(risk_score[order(-risk_score$Risk.Score),], n = 8)
```

```
##     Province.State.x Country.Region.x Total.Deaths Total.Recovered
## 21                            Belgium        23473               0
## 163                       Netherlands        16771               0
## 165  New South Wales        Australia           54               0
## 201                            Serbia         5735               0
## 223                            Sweden        13621               0
## 242                    United Kingdom       127087               0
## 244                                US       562066               0
## 147       Martinique           France           59              98
##     Total.Confirmed Risk.Score Death.Burden
## 21           925476        Inf          Inf
## 163         1350665        Inf          Inf
## 165            5339        Inf          Inf
## 201          642208        Inf          Inf
## 223          857401        Inf          Inf
## 242         4369775        Inf          Inf
## 244        31197511        Inf          Inf
## 147            8887  0.6020408     5350.337
```

```r
tail(risk_score[order(-risk_score$Risk.Score),], n = 23)
```

```
##                                 Province.State.x      Country.Region.x
## 209                                                          Singapore
## 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
## 209           30           60335           60653 0.0004972238     30.15812
## 6              0              22              25 0.0000000000      0.00000
## 60             0             159             165 0.0000000000      0.00000
## 70             0              54              60 0.0000000000      0.00000
## 86             0              31              31 0.0000000000      0.00000
## 101            0              15              27 0.0000000000      0.00000
## 119            0             708             716 0.0000000000      0.00000
## 129            0              47              51 0.0000000000      0.00000
## 139            0              48              49 0.0000000000      0.00000
## 146            0               4               4 0.0000000000      0.00000
## 152            0               1               1 0.0000000000      0.00000
## 164            0              58             121 0.0000000000      0.00000
## 170            0              75              75 0.0000000000      0.00000
## 172            0             107             112 0.0000000000      0.00000
## 184            0              18              18 0.0000000000      0.00000
## 191            0               4               4 0.0000000000      0.00000
## 192            0              44              44 0.0000000000      0.00000
## 194            0              24              24 0.0000000000      0.00000
## 196            0               2               3 0.0000000000      0.00000
## 206            0             240             248 0.0000000000      0.00000
## 213            0              18              19 0.0000000000      0.00000
## 232            0               1               1 0.0000000000      0.00000
## 246            0               1               3 0.0000000000      0.00000
```

There are 22 areas with a low risk score of 0. These areas include provinces and states in UK, Dominica, Denmark, Holy See, China, Laos, Marshall Islands, Mmicronesia, France, Australia, Saint Kitts and Nevis, Samoa, Solom Islands and Vanuatu. A risk score of 0 is likely due to a lack of deaths data where total deaths is at 0. Singapore would be the first true low risk score at 0.00050, which is 0.13 almost 1/8 of the global risk score value.

Seven area show a high risk score of "Inf". These areas include Belgium, Netherlands, Australia (New South Wales), Serbia, Sweden, United Kingdom, and US. However, a contributing discrepency to the risk score assesment of the listed areas is likely due to a lack of recovered cases data. Having total deaths number divided by total recovered number of 0 results in Inf, which would not be repesentative of the true risk score for these areas. France (Martinique), is therefor the first listed area with a true high value risk score of 0.602. Compared to the global risk score of 0.038, Martinique's risk score is 15 times greater than the global risk score.

Death burden value of the least risky area (Singapore) is 178 times less when compared to the most risky area (Martinique). There is a large difference in death burden between then two area with opposite risk scores.

Objective 4.2

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

total_recovered2 <- subset(recovered, select = c(2,colnum_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,colnum_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

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


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


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

Note: I am unable to uninstall an old version of kableExtra package and can not knit my rmarkdown file without omitting the kable() functions.

GitHub Log

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

```
##
## Subject: Updated pdf
## Author: mrgrabiel
## Date: Sun, 11 Apr 2021 18:38:14 -0700
## Body:
##
## Subject: Load final Markdown script
## Author: mrgrabiel
## Date: Sun, 11 Apr 2021 18:33:00 -0700
## Body:
##
## 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:
```