Step 1 - Data Cleaning

Anonymous

02-01-2021

rm(list = ls(all = TRUE))  
  
# this is a function to remove missing data  
completeFun <- function(data, desiredCols) {  
 completeVec <- complete.cases(data[, desiredCols])  
 return(data[completeVec, ])  
}  
  
# need pacman package installed to run this  
# install.packages("pacman")  
pacman::p\_load("dplyr","countrycode","tidyverse","readstata13","countrycode","car","ggplot2","xlsx","tidyverse","stringi", "readxl")

## COVIDiSTRESS Data project

Andreas Lieberoth, Dominik-Borna Ćepulić, Jesper Rasmussen, Giovanni A. Travaglino, Jarno Tuominen Available at: <https://osf.io/z39us/>

The most recent file was downloaded on 27-May-2020 from the above site

“COVIDiSTRESS global survey May 4 2020 (numeric values).csv”

A comparative survey of psychological stress and pandemic attitudes. Online, not a random sample.

Original English question wording:

Concern about consequences of the coronavirus (1)…for yourself (2)…for your family (3)…for your close friends (4)…for your country (5)…for other countries across the globe

### Load Data

# The data have a strange format with the first 3 rows occupied by text  
  
# Import separate file for header  
  
cishead <- read.csv("data/numeric/COVIDiSTRESS global survey May 25 2020 (numeric values).csv", header = F, nrows=1, as.is = T)  
  
# Now import numeric data  
  
cis <- read.csv("data/numeric/COVIDiSTRESS global survey May 25 2020 (numeric values).csv")  
  
# Remove rows and rename headers   
  
cis <- cis[-c(1,2),]  
colnames(cis) <- cishead  
  
rm(cishead)  
  
# The country names are only listed in the 'choice' files, extract them here   
cisc <- read.csv("data/choice/COVIDiSTRESS global survey May 25 2020 (choice text).csv")  
  
cisc <- cisc[-c(1,2),]  
  
cisc$cname <- cisc$Country  
  
cisc <- select(cisc, ResponseId, cname)  
  
# append country name to numeric data  
cis <- left\_join(cis,cisc, by = "ResponseId")  
  
rm(cisc)  
  
# get country codes and ISO letters  
cis$cow <- countrycode(cis$cname, "country.name", "cown")  
cis$iso <- countrycode(cis$cname, "country.name", "iso3c")  
  
  
#adjust format of risk perception question  
  
cis <- cis %>%  
mutate(Corona\_concerns\_1 = as.numeric(Corona\_concerns\_1),  
 Corona\_concerns\_2 = as.numeric(Corona\_concerns\_2),  
 Corona\_concerns\_3 = as.numeric(Corona\_concerns\_3),  
 Corona\_concerns\_4 = as.numeric(Corona\_concerns\_4),  
 Corona\_concerns\_5 = as.numeric(Corona\_concerns\_5),  
 date = as.Date(EndDate))  
  
  
#get case numbers by country (drop when too low)  
cis <- cis %>%   
 group\_by(cow) %>%  
 mutate(cases = sum(!is.na(Corona\_concerns\_1))) %>%  
 ungroup()  
  
  
# Save files  
save(cis, file = "data/cis.Rdata")

### Clean Data

load(file = "data/cis.Rdata", .GlobalEnv)  
  
# Fix date & create mean of two types of risk perceptions (measurement model justifying this we find in Breznau (2020))  
  
cis <- cis %>%  
 as.data.frame() %>%  
 mutate(date = as.Date(StartDate))  
cis <- cis %>%  
 rowwise() %>%  
 mutate(concern\_self = mean(c(Corona\_concerns\_1,Corona\_concerns\_2,Corona\_concerns\_3), na.rm=T),  
 concern\_society = mean(c(Corona\_concerns\_4,Corona\_concerns\_5), na.rm=T))  
  
cis$concern\_self <- ifelse(cis$concern\_self == "NaN", NA, cis$concern\_self)  
cis$concern\_society <- ifelse(cis$concern\_society == "NaN", NA, cis$concern\_society)  
  
# S.Sudan and Panama have very few non-missing cases, plus some countries are NA. Remove.  
cis <- subset(cis, !is.na(cow) & cow!=95)  
  
# Make list of using countries (those with at least 20 cases)  
  
use\_countriesa <- as.list(c(2, 20, 55, 70, 90, 92, 94, 100, 130, 135, 140, 155, 160, 200, 205, 210, 211, 212, 220, 225, 230, 235, 255, 290, 305, 310, 316, 317, 325, 338, 339, 343, 344, 346, 349, 350, 352, 355, 360, 365, 366, 367, 368, 369, 372, 375, 380, 385, 390, 395, 560, 600, 615, 640, 651, 666, 696, 700, 703, 710, 713, 732, 740, 750, 770, 771, 816, 820, 830, 835, 840, 850, 900, 920))

###Johns Hopkins Covid-19 Tracker

A multi-source project for compiling global confirmed cases, deaths and recoveries.

Originally vailable at: <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>

Confirmed cases data downloaded May 27th. Thus, includes data through May 26th. Confirmed deaths data downloaded Jun 21st. Includes data through June 20th.

confirmed <- read.csv("data/time\_series\_covid19\_confirmed\_global.csv", header = F)  
deaths <- read.csv("data/time\_series\_covid19\_deaths\_global.csv", header = F)  
  
  
confirmed[1,1] <- "Province"  
confirmed[1,2] <- "Country"  
deaths[1,1] <- "Province"  
deaths[1,2] <- "Country"  
  
cnames <- as.list(confirmed[1,1:130])  
cnamesd <- as.list(deaths[1,1:155])  
  
colnames(confirmed) <- cnames  
colnames(deaths) <- cnamesd  
  
confirmed[,5:130] <- sapply(confirmed[5:130],as.numeric)  
deaths[,5:155] <- sapply(deaths[5:155],as.numeric)  
  
# Canada, Australia and China are by province, not a sum  
  
cca <- subset(confirmed, Country=="Australia", select = -c(Province,Country,Lat,Long))  
ccca <- subset(confirmed, Country=="Canada", select = -c(Province,Country,Lat,Long))  
ccch <- subset(confirmed, Country=="China", select = -c(Province,Country,Lat,Long))  
ccad <- subset(deaths, Country=="Australia", select = -c(Province,Country,Lat,Long))  
cccad <- subset(deaths, Country=="Canada", select = -c(Province,Country,Lat,Long))  
ccchd <- subset(deaths, Country=="China", select = -c(Province,Country,Lat,Long))  
  
  
a <- as.data.frame(colSums(cca, na.rm = T))  
b <- as.data.frame(colSums(ccca, na.rm = T))  
c <- as.data.frame(colSums(ccch, na.rm = T))  
ad <- as.data.frame(colSums(ccad, na.rm = T))  
bd <- as.data.frame(colSums(cccad, na.rm = T))  
cd <- as.data.frame(colSums(ccchd, na.rm = T))  
  
aa <- data.frame("All","Australia",0,0,t(a))  
bb <- data.frame("All","Canada",0,0,t(b))   
cc <- data.frame("All","China",0,0,t(c))  
aaa <- data.frame("All","Australia",0,0,t(ad))  
bbb <- data.frame("All","Canada",0,0,t(bd))   
ccc <- data.frame("All","China",0,0,t(cd))  
  
colnames(aa) <- cnames  
colnames(bb) <- cnames  
colnames(cc) <- cnames  
colnames(aaa) <- cnamesd  
colnames(bbb) <- cnamesd  
colnames(ccc) <- cnamesd  
  
# add the three missing countries  
confirmed <- rbind(confirmed,aa,bb,cc)  
deaths <- rbind(deaths,aaa,bbb,ccc)  
  
# remove province-level data  
confirmed <- subset(confirmed, Province=="" | Province=="All")  
deaths <- subset(deaths, Province=="" | Province=="All")  
  
rm(a,b,c,aa,bb,cc,ad,bd,cd,aaa,bbb,ccc,cca,ccca,ccch,ccad,cccad,ccchd)  
  
confirmed <- subset(confirmed, select = -c(Province,Lat,Long))  
deaths <- subset(deaths, select = -c(Province,Lat,Long))  
  
# Final data for merging, wide  
hopkins <- left\_join(confirmed, deaths, by = "Country")  
hopkins$cow <- countrycode(hopkins$Country, "country.name", "cown")  
  
colnames(confirmed) <- paste0("conf",colnames(confirmed))  
colnames(confirmed)[1] <- "Country"  
  
colnames(deaths) <- paste0("dead",colnames(deaths))  
colnames(deaths)[1] <- "Country"  
  
  
confirmed\_long <- reshape(confirmed, idvar = "Country", direction = "long", v.names = "conf", varying = 2:127)  
  
deaths\_long <- reshape(deaths, idvar = "Country", direction = "long", v.names = "dead", varying = 2:151)  
  
datel <- seq.Date(as.Date("2020-1-22"),as.Date("2020-5-26"), by = "days")  
datea <- seq.int(1,126)  
dateld <- seq.Date(as.Date("2020-1-22"),as.Date("2020-6-20"), by = "days")  
datead <- seq.int(1,151)  
   
   
datem <- data.frame(time = datea, date = datel)  
datemd <- data.frame(time = datead, date = dateld)  
datem$date <- as.Date(datem$date)  
datemd$date <- as.Date(datemd$date)  
  
deaths\_long <- left\_join(deaths\_long,datemd)  
confirmed\_long <- left\_join(confirmed\_long,datem)  
  
deaths\_long$cow <- countrycode(deaths\_long$Country, "country.name", "cown")  
confirmed\_long$cow <- countrycode(confirmed\_long$Country, "country.name", "cown")  
  
# lagged versions, 5-days   
# We include additional variables that were not used in the final analysis  
  
# sort first  
deaths\_long <- deaths\_long[order(deaths\_long$cow, deaths\_long$date),]  
confirmed\_long <- confirmed\_long[order(confirmed\_long$cow, confirmed\_long$date),]  
  
deaths\_long <- deaths\_long %>%  
 mutate(dead\_l5 = lag(dead, n = 5L),  
 dead\_lead = lead(dead, n = 18L), # create 2.5 week lead (18 days)  
 dead\_s = ifelse(dead > 0, 1, 0),# first death ID  
 dead\_s1l = lag(dead\_s, 1L),  
 dead\_dif = dead\_s - dead\_s1l,  
 dead\_dif = ifelse(cow == 710 & date == as.Date("2020-01-22"), 1, dead\_dif))  
   
# first death date  
  
deaths\_long <- deaths\_long %>%  
 group\_by(cow) %>%  
 mutate(dead\_1st = dplyr::if\_else(dead\_dif == 1, as.Date(date), as.Date("2020-06-02")),  
 dead\_1st\_date = min(dead\_1st, na.rm=T)) %>%  
 ungroup()  
  
confirmed\_long <- confirmed\_long %>%  
 mutate(conf\_l5 = lag(conf, n = 5L),  
 conf\_l10 = lag(conf, n = 10L))  
  
  
deaths\_long <- select(deaths\_long, dead, dead\_l5, dead\_lead, dead\_1st\_date, date, cow)  
confirmed\_long <- select(confirmed\_long, conf, conf\_l5, conf\_l10, date, cow)  
  
# l5 variable comes from the wrong series prior to 1-27  
deaths\_long$dead\_l5 <- ifelse(deaths\_long$date < as.Date("2020-01-27"), NA, deaths\_long$dead\_l5)  
  
confirmed\_long$conf\_l5 <- ifelse(confirmed\_long$date < as.Date("2020-01-27"), NA, confirmed\_long$conf\_l5)  
confirmed\_long$conf\_l10 <- ifelse(confirmed\_long$date < as.Date("2020-02-01"), NA, confirmed\_long$conf\_l10)  
  
# increasing or decreasing rate past week, numbers are so different that I make a trichotomy: < 1 = -1, 0-1 = 0 and > 1 = 1  
confirmed\_long$conf\_delta <- (confirmed\_long$conf - confirmed\_long$conf\_l5) - (confirmed\_long$conf\_l5 - confirmed\_long$conf\_l10)  
  
confirmed\_long$conf\_delta <- ifelse(confirmed\_long$conf\_delta < 0, -1, ifelse(confirmed\_long < 1.01, 0, 1))  
  
  
  
# Merge cases and deaths per date of survey per respondent  
  
  
# Find the moment when the curve 'flattens'  
  
deaths\_long <- deaths\_long %>%  
 mutate(dead\_lead12 = lag(dead\_lead, 12L),  
 dead\_lead11 = lag(dead\_lead, 11L),  
 dead\_lead10 = lag(dead\_lead, 10L),  
 dead\_lead9 = lag(dead\_lead, 9L),  
 dead\_lead8 = lag(dead\_lead, 8L),  
 dead\_lead5 = lag(dead\_lead, 5L),  
 dead\_lead4 = lag(dead\_lead, 4L),  
 dead\_lead3 = lag(dead\_lead, 3L),  
 dead\_lead2 = lag(dead\_lead, 2L),  
 dead\_lead1 = lag(dead\_lead, 1L),  
 dead\_lead\_past12 = (dead\_lead12 + dead\_lead11 + dead\_lead10 + dead\_lead9 + dead\_lead8)/5,  
 dead\_lead\_past5 = (dead\_lead5 + dead\_lead4 + dead\_lead3 + dead\_lead2 + dead\_lead1)/5,  
 dead\_lead\_wkchg = dead\_lead\_past12 - dead\_lead\_past5)  
  
# find the minimum point of weekly change, this is the height of the curve  
  
deaths\_long <- deaths\_long %>%  
 group\_by(cow) %>%  
 mutate(curve\_maxd = min(dead\_lead\_wkchg, na.rm=T),  
 curve\_maxs = dplyr::if\_else(curve\_maxd == dead\_lead\_wkchg, as.Date(date), as.Date("2020-6-2")),  
 curve\_max = min(curve\_maxs, na.rm=T)) %>%  
 ungroup()  
  
deaths\_long <- subset(deaths\_long, select = -c(curve\_maxd, curve\_maxs))  
  
  
rm(cis\_na)  
  
cis$date = as.Date(cis$EndDate)  
cis <- left\_join(cis,deaths\_long, by = c("cow","date"))  
cis <- left\_join(cis,confirmed\_long, by = c("cow","date"))

## Welfare State Strength

### ILO Worker Protection Coverage

ILO. 2014. “Global Programme Employment Injury Insurance and Protection | GEIP Data.” <https://www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_ent/documents/publication/wcms_573083.pdf>

### ILO - Social Spending

Public Social Expenditure as a % of GDP (Table .16) <https://www.social-protection.org/gimi/gess/ShowWiki.action?id=594#tabs-3>

socp <- read.xlsx("data/54614.xlsx", startrow = 8, sheetName = "B.16 Data (Print)")  
  
socp <- completeFun(socp, "NA..1")  
  
socp <- select(socp, NA..1, NA..18, NA..19, NA..20)  
  
colnames(socp) <- c("country","soc\_spend","year","source")  
  
socp$country <- as.character(socp$country)  
socp$soc\_spend <- as.numeric(as.character(socp$soc\_spend))  
  
socp$cow <- countrycode(socp$country, "country.name","cown")  
  
socp <- completeFun(socp, "soc\_spend")

## Maddison Data - GDP

Maddison Project data from <https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2018?lang=en>

2018 or latest data

gdpm <- read.csv("data/mpd2018.csv", header=T, stringsAsFactors = F)  
  
gdpm <- subset(gdpm, year > 2013)  
gdpm <- select(gdpm, country, cgdppc, pop)  
gdpc <- aggregate(gdpm, by = list(gdpm$country), FUN = mean)  
gdpc$cow <- countrycode(gdpc$Group.1, "country.name", "cown")  
gdpc <- gdpc %>%  
 mutate(gdp = round(cgdppc,0),  
 pop = round(pop,0))  
gdpc <- select(gdpc, cow, gdp, pop)  
  
gdpc <- completeFun(gdpc, "cow")

## Blavatnik Government Data

A severity scale from researchers at Oxford University. We take the severity of ‘lockdown’ measures by March 15th to allow for a lag for individuals to become aware of the measures and possibly see their impact. Naumann et al ([2020](https://www.uni-mannheim.de/media/Einrichtungen/gip/Corona_Studie/Schwerpunktbericht_Angstempfinden_Mannheimer_Corona_Studie.pdf)) showed that over time, the ‘panic’ subsides among the public after strong measures are taken.

<https://ourworldindata.org/grapher/covid-stringency-index>

# government response  
gov\_resp <- read.csv(file = "data/covid-stringency-index.csv", header = T)  
  
  
# get country codes  
gov\_resp$cow <- countrycode(gov\_resp$Entity, "country.name", "cown")  
  
gov\_resp <- gov\_resp %>%  
 group\_by(cow) %>%  
 mutate(drop = ifelse(Date == "15-Mar-20", 0, 1)) %>%  
 ungroup()  
  
gov\_resp <- subset(gov\_resp, drop == 0)  
gov\_resp <- subset(gov\_resp, cow %in% use\_countriesa)  
colnames(gov\_resp) <- c("1","2","3","gov\_resp","cow","4")  
gov\_resp <- select(gov\_resp, cow, gov\_resp)  
  
# impute Grenada, Malta and N Macedonia (using Wikipedia info)  
# Grenada, strong lockdown, very few cases. Score = 87  
# Malta, moderate measures. Score = 70  
# N Macedonia, very strong, near total lockdown. Score = 93  
  
gov\_resp[72,1] <- 55  
gov\_resp[72,2] <- 87  
  
gov\_resp[73,1] <- 338  
gov\_resp[73,2] <- 70  
  
gov\_resp[74,1] <- 343  
gov\_resp[74,2] <- 93  
  
  
# standardize for ease of interpretation  
  
gov\_resp$gov\_resp = as.numeric(scale(gov\_resp$gov\_resp))

## Income Concentration

As a robustness check we include the top 1% income concentration. We also leave the top 10% concentration in the data here although it has more missing values.

<https://wid.world/data/>

# income concentration  
top\_inc <- read\_xlsx("data/WID\_Data\_14082020-111941.xlsx")  
  
  
top\_inc <- top\_inc %>%  
 mutate(cow = countrycode(Country, "country.name", "cown"))  
  
top\_inc1 <- subset(top\_inc, top\_inc$Percentile == "p99p100")  
top\_inc2 <- subset(top\_inc, top\_inc$Percentile == "p90p100")  
top\_inc1 <- subset(top\_inc1, !is.na(top\_inc1$cow))  
top\_inc2 <- subset(top\_inc2, !is.na(top\_inc2$cow))  
  
# Take the mean from 2010-2019 to account for missing data  
top\_inc1 <- top\_inc1[,c(13:22,23)]  
top\_inc2 <- top\_inc2[,c(3:12,23)]  
top\_inc1$top1 <- rowMeans(top\_inc1[,1:10], na.rm = T)  
top\_inc2$top10 <- rowMeans(top\_inc2[,1:10], na.rm = T)  
top\_inc1 <- select(top\_inc1, cow, top1)  
top\_inc2 <- select(top\_inc2, cow, top10)  
# china appears multiple times  
top\_inc1 <- aggregate(top\_inc1, by = list(top\_inc1$cow), FUN = mean, na.rm = T)  
top\_inc2 <- aggregate(top\_inc2, by = list(top\_inc2$cow), FUN = mean, na.rm = T)  
rm(top\_inc)

## Solt Gini

The Solt data include multiple measures of the Gini for many countries. We take the average score provided by Solt for disposable income inequality.

Solt, Frederick. 2020. “Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database.” Social Science Quarterly [DOI](https://doi.org/10.1111/ssqu.12795)

# Solt Gini  
load("data/swiid8\_3.Rda")  
  
rm(swiid)  
swiid\_summary$cow <- countrycode(swiid\_summary$country, "country.name", "cown")  
  
#some countries do not have data since 2016, take most recent available year  
swiid\_summary <- swiid\_summary %>%  
 mutate(year = ifelse(country == "Algeria" & year == 2011 | country == "Brunei" & year == 1981 | country == "Bosnia and Herzegovina" & year == 2015 | country == "Grenada" & year == 2008 | country == "Guatemala" & year == 2014 | country == "Iceland" & year == 2015 | country == "India" & year == 2012 | country == "Japan" & year == 2015 | country == "Morocco" & year == 2014 | country == "Pakistan" & year == 2015 | country == "Philippines" & year == 2015 | country == "South Africa" & year == 2015 | country == "United Arab Emirates" & year == 2008, 2016, year))  
  
# trim Extremes ZAF BRN  
swiid\_summary$gini\_disp <- ifelse(swiid\_summary$gini\_disp > 50, 49, swiid\_summary$gini\_disp)  
  
  
swiid\_summary <- subset(swiid\_summary, year >= 2016)  
swiid\_summary <- select(swiid\_summary, cow, gini\_disp)  
swiid\_summary <- aggregate(swiid\_summary, by = list(swiid\_summary$cow), FUN = mean, na.rm = T)  
gini\_disp <- select(swiid\_summary, cow, gini\_disp)  
  
rm(swiid\_summary)

## 1st Merge Data

# The wider the date range the greater the chance of introducing global/regional period effects. Reduce range, but maximize country sample.   
  
# cis\_b is to remain at individual level  
cis\_b <- subset(cis, date < as.Date('2020-05-01'))  
  
cis\_a <- aggregate(cis\_b, by=list(cis\_b$cow),  
 FUN=mean, na.rm=T)  
  
# Calculate SE for robustness check  
cis\_asd <- aggregate(cis\_b, by=list(cis\_b$cow),  
 FUN=sd, na.rm=T)  
cis\_asd <- select(cis\_asd, Group.1, concern\_self)  
colnames(cis\_asd) <- c("cow", "concern\_self\_sd")  
  
cis\_a <- left\_join(cis\_a, cis\_asd, by = "cow")  
  
cis\_b$cases <- ifelse(cis\_b$cases<20, NA, cis\_b$cases)  
cis\_a$cases <- ifelse(cis\_a$cases<20, NA, cis\_a$cases)  
cis\_a <- completeFun(cis\_a, "cases")  
  
cis\_a$concern\_self\_se <- cis\_a$concern\_self\_sd/sqrt(cis\_a$cases)  
  
cis\_a <- select(cis\_a, cow, concern\_self, concern\_society, dead, dead\_l5, conf, conf\_l5, conf\_delta, dead\_lead, dead\_1st\_date, curve\_max, Corona\_concerns\_1, Corona\_concerns\_2, Corona\_concerns\_3, Corona\_concerns\_4, Corona\_concerns\_5, concern\_self\_se, cases)  
  
  
colnames(cis\_a) <- c("cow","concern\_self","concern\_society", "dead", "dead\_lag5", "conf", "conf\_lag5", "conf\_delta","dead\_lead", "dead\_1st\_date", "curve\_max","Corona1", "Corona2", "Corona3", "Corona4", "Corona5", "concern\_self\_se","cases")  
# add iso  
  
  
cis\_a$iso <- countrycode(cis\_a$cow, "cown", "iso3c")  
  
finaldf\_C <- full\_join(cis\_a, socp, by = "cow")  
finaldf\_C <- full\_join(finaldf\_C, geip, by = "cow")  
finaldf\_C <- full\_join(finaldf\_C, gdpc, by = "cow")  
finaldf\_C <- full\_join(finaldf\_C, gov\_resp, by = "cow")  
finaldf\_C <- full\_join(finaldf\_C, top\_inc1, by = "cow")  
finaldf\_C <- full\_join(finaldf\_C, top\_inc2, by = "cow")  
finaldf\_C <- full\_join(finaldf\_C, gini\_disp, by = "cow")

save.image(file="data/cis.RData")

## References

Breznau, Nate. 2020. “The Welfare State and Risk Perceptions: The Novel Coronavirus Pandemic and Public Concern in 70 Countries.” *European Societies*. [DOI](https://doi.org/10.1080/14616696.2020.1793215)