Meeting 1

Rui Li

9/7/2020

# Require packages

library(utils)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

library(maps)

# Import the newest COVID-19 Dataset

#these libraries need to be installed  
#library(utils)  
  
#read the Dataset sheet into “R”. The dataset will be called "data".  
data <- read.csv("https://opendata.ecdc.europa.eu/covid19/casedistribution/csv", na.strings = "", fileEncoding = "UTF-8-BOM")

# Data cleaning and Summary

#library("dplyr") require "dplyr" package for the use of %>%  
covid.data = data %>%  
 rename(Date="dateRep") %>%  
 rename(Region="countriesAndTerritories") %>%  
 mutate(Date=as.Date(Date,format="%d/%m/%y"))  
  
summary(covid.data)

## Date day month year   
## Min. :2020-01-01 Min. : 1.00 Min. : 1.000 Min. :2019   
## 1st Qu.:2020-04-13 1st Qu.: 8.00 1st Qu.: 4.000 1st Qu.:2020   
## Median :2020-06-01 Median :16.00 Median : 6.000 Median :2020   
## Mean :2020-05-28 Mean :15.78 Mean : 5.413 Mean :2020   
## 3rd Qu.:2020-07-21 3rd Qu.:24.00 3rd Qu.: 7.000 3rd Qu.:2020   
## Max. :2020-12-31 Max. :31.00 Max. :12.000 Max. :2020   
##   
## cases deaths Region geoId   
## Min. :-2461.0 Min. :-1918.00 Australia: 252 AT : 252   
## 1st Qu.: 0.0 1st Qu.: 0.00 Austria : 252 AU : 252   
## Median : 8.0 Median : 0.00 Belgium : 252 BE : 252   
## Mean : 658.8 Mean : 21.58 Brazil : 252 BR : 252   
## 3rd Qu.: 138.0 3rd Qu.: 2.00 Canada : 252 CA : 252   
## Max. :90802.0 Max. : 4928.00 China : 252 CH : 252   
## (Other) :39698 (Other):39698   
## countryterritoryCode popData2019 continentExp   
## AUS : 252 Min. :8.150e+02 Africa : 9591   
## AUT : 252 1st Qu.:1.395e+06 America: 8932   
## BEL : 252 Median :8.519e+06 Asia : 9303   
## BRA : 252 Mean :4.325e+07 Europe :11804   
## CAN : 252 3rd Qu.:2.916e+07 Oceania: 1516   
## (Other):39886 Max. :1.434e+09 Other : 64   
## NA's : 64 NA's :64   
## Cumulative\_number\_for\_14\_days\_of\_COVID.19\_cases\_per\_100000  
## Min. :-145.3020   
## 1st Qu.: 0.3261   
## Median : 4.1755   
## Mean : 31.2202   
## 3rd Qu.: 24.5463   
## Max. :1058.2259   
## NA's :2781

head(covid.data)

## Date day month year cases deaths Region geoId countryterritoryCode  
## 1 2020-09-07 7 9 2020 74 2 Afghanistan AF AFG  
## 2 2020-09-06 6 9 2020 20 0 Afghanistan AF AFG  
## 3 2020-09-05 5 9 2020 16 0 Afghanistan AF AFG  
## 4 2020-09-04 4 9 2020 45 1 Afghanistan AF AFG  
## 5 2020-09-03 3 9 2020 38 3 Afghanistan AF AFG  
## 6 2020-09-02 2 9 2020 9 0 Afghanistan AF AFG  
## popData2019 continentExp  
## 1 38041757 Asia  
## 2 38041757 Asia  
## 3 38041757 Asia  
## 4 38041757 Asia  
## 5 38041757 Asia  
## 6 38041757 Asia  
## Cumulative\_number\_for\_14\_days\_of\_COVID.19\_cases\_per\_100000  
## 1 1.0488475  
## 2 0.8543244  
## 3 1.0777630  
## 4 1.1355942  
## 5 1.2722861  
## 6 1.5929864

# Data Distribution of population

#get a list of dates, countries, population, and continents  
dates = unique(covid.data$Date)  
regions = unique(covid.data$Region)  
pops = unique(covid.data$popData2019)  
continents = unique(covid.data$continentExp)  
  
#population dataframe  
pop.data = unique(data.frame(covid.data$Region,covid.data$popData2019,covid.data$continentExp))  
  
names(pop.data)=c("Region","Population2019","Continents")  
  
summary(pop.data)

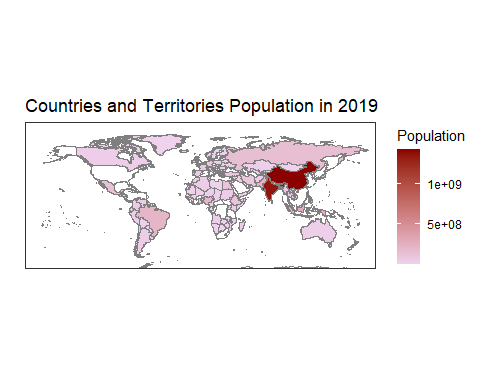
## Region Population2019 Continents  
## Afghanistan: 1 Min. :8.150e+02 Africa :55   
## Albania : 1 1st Qu.:9.736e+05 America:49   
## Algeria : 1 Median :6.964e+06 Asia :43   
## Andorra : 1 Mean :3.669e+07 Europe :54   
## Angola : 1 3rd Qu.:2.572e+07 Oceania: 8   
## Anguilla : 1 Max. :1.434e+09 Other : 1   
## (Other) :204 NA's :1

head(pop.data)

## Region Population2019 Continents  
## 1 Afghanistan 38041757 Asia  
## 243 Albania 2862427 Europe  
## 426 Algeria 43053054 Africa  
## 673 Andorra 76177 Europe  
## 851 Angola 31825299 Africa  
## 1021 Anguilla 14872 America

#get the world polygon  
worldData = map\_data("world")  
#plot world population based on color  
ggplot() +  
 geom\_map(data = worldData, map = worldData,  
 aes(x = long, y = lat, group = group, map\_id=region),  
 fill = "white", colour = "#7f7f7f", size=0.5) +   
 geom\_map(data = pop.data, map=worldData,  
 aes(fill=Population2019, map\_id=Region),  
 colour="#7f7f7f", size=0.5) +  
 coord\_map("rectangular", lat0=0, xlim=c(-180,180), ylim=c(-60, 90)) +  
 scale\_fill\_continuous(low="thistle2", high="darkred", guide="colorbar") +  
 scale\_y\_continuous(breaks=c()) +  
 scale\_x\_continuous(breaks=c()) +  
 labs(fill="Population", title="Countries and Territories Population in 2019", x="", y="") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 theme\_bw()

## Warning: Ignoring unknown aesthetics: x, y



# Data Distribution of cases

#cases dataframe  
case.data = data.frame(covid.data$Region,covid.data$cases)   
names(case.data) = c("Region","cases")   
case.data=aggregate(case.data$cases, by=list(Category=case.data$Region), FUN=sum)  
names(case.data) = c("Region","cases")  
  
summary(case.data)

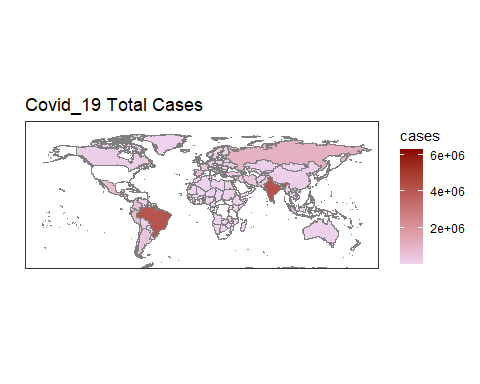
## Region cases   
## Afghanistan: 1 Min. : 3   
## Albania : 1 1st Qu.: 1190   
## Algeria : 1 Median : 6255   
## Andorra : 1 Mean : 129290   
## Angola : 1 3rd Qu.: 51697   
## Anguilla : 1 Max. :6276421   
## (Other) :204

head(case.data)

## Region cases  
## 1 Afghanistan 38398  
## 2 Albania 10255  
## 3 Algeria 46364  
## 4 Andorra 1215  
## 5 Angola 2935  
## 6 Anguilla 3

#get the world polygon  
worldData = map\_data("world")  
#plot world population based on color  
ggplot() +  
 geom\_map(data = worldData, map = worldData,  
 aes(x = long, y = lat, group = group, map\_id=region),  
 fill = "white", colour = "#7f7f7f", size=0.5) +   
 geom\_map(data = case.data, map=worldData,  
 aes(fill=cases, map\_id=Region),  
 colour="#7f7f7f", size=0.5) +  
 coord\_map("rectangular", lat0=0, xlim=c(-180,180), ylim=c(-60, 90)) +  
 scale\_fill\_continuous(low="thistle2", high="darkred", guide="colorbar") +  
 scale\_y\_continuous(breaks=c()) +  
 scale\_x\_continuous(breaks=c()) +  
 labs(fill="cases", title="Covid\_19 Total Cases", x="", y="") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 theme\_bw()

## Warning: Ignoring unknown aesthetics: x, y



# Data Distribution of death

#cases dataframe  
death.data = data.frame(covid.data$Region,covid.data$deaths)   
names(death.data) = c("Region","deaths")   
death.data=aggregate(death.data$deaths, by=list(Category=death.data$Region), FUN=sum)  
names(death.data) = c("Region","deaths")  
  
summary(death.data)

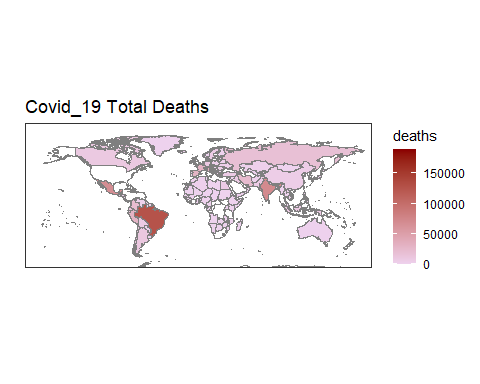
## Region deaths   
## Afghanistan: 1 Min. : 0.0   
## Albania : 1 1st Qu.: 17.5   
## Algeria : 1 Median : 118.0   
## Andorra : 1 Mean : 4234.6   
## Angola : 1 3rd Qu.: 815.0   
## Anguilla : 1 Max. :188941.0   
## (Other) :204

head(death.data)

## Region deaths  
## 1 Afghanistan 1412  
## 2 Albania 316  
## 3 Algeria 1556  
## 4 Andorra 53  
## 5 Angola 117  
## 6 Anguilla 0

#get the world polygon  
worldData = map\_data("world")  
#plot world population based on color  
ggplot() +  
 geom\_map(data = worldData, map = worldData,  
 aes(x = long, y = lat, group = group, map\_id=region),  
 fill = "white", colour = "#7f7f7f", size=0.5) +   
 geom\_map(data = death.data, map=worldData,  
 aes(fill=deaths, map\_id=Region),  
 colour="#7f7f7f", size=0.5) +  
 coord\_map("rectangular", lat0=0, xlim=c(-180,180), ylim=c(-60, 90)) +  
 scale\_fill\_continuous(low="thistle2", high="darkred", guide="colorbar") +  
 scale\_y\_continuous(breaks=c()) +  
 scale\_x\_continuous(breaks=c()) +  
 labs(fill="deaths", title="Covid\_19 Total Deaths", x="", y="") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 theme\_bw()

## Warning: Ignoring unknown aesthetics: x, y



#Other tries  
centroids <- summarize(group\_by(worldData, region),  
 x = mean(range(long)), y = mean(range(lat)))  
names(centroids) <- c("Region","long","lat")  
  
centroids <- centroids[centroids$Region%in%regions,]  
head(centroids)

## # A tibble: 6 x 3  
## Region long lat  
## <chr> <dbl> <dbl>  
## 1 Afghanistan 67.7 33.9  
## 2 Albania 20.2 41.2  
## 3 Algeria 1.64 28.0  
## 4 Andorra 1.58 42.5  
## 5 Angola 17.9 -11.2  
## 6 Anguilla -63.1 18.2

ggplot() +   
 geom\_polygon(data = worldData,aes(x = long, y = lat, fill = region, group = group), color = "white") +   
 coord\_fixed(1.3) +  
 guides(fill=FALSE) + # do this to leave off the color legend  
 geom\_point(aes(x=long, y=lat), data = centroids, color = "red")

