Covid data analysis - Africa

Isha Doshi

2022-10-21

Loaded a single month of African data. Loaded the list of African countries countries-africa.csv. There are 58 countries listed here.

```
Africa <- read.delim("covid/countries-africa.csv.bz2")
head(Africa,5)
```

```
##   rank                          country  population             year
## 1    1                          Nigeria 206,139,589             2020
## 2    2                         Ethiopia 109,224,414             2018
## 3    3 Democratic Republic of the Congo 102,561,403     July 1, 2020
## 4    4                            Egypt 101,334,404 December 8, 2020
## 5    5                     South Africa  59,956,820 December 1, 2020
##                         source
## 1                Worldometers[3]
## 2 UN population projections[4][5]
## 3      National annual projection
## 4       National population clock
## 5               Official estimate
```

```
nrow(Africa)
```

```
## [1] 58
```

```
any(is.na(Africa$country))
```

```
## [1] FALSE
```

```
# checking if any country is NA. There are 58 African countries listed in the database.
```

Collected all the names of covid data files into a character vector. There are 21 files with names starting with covid data.

```
vec= list.files(path = "covid/", pattern = "^covid-global", all.files = FALSE,
           full.names = FALSE, recursive = FALSE,
           ignore.case = FALSE, include.dirs = FALSE, no.. = FALSE)
vec
```

```
##  [1] "covid-global_01-01-2021.csv.bz2" "covid-global_02-01-2020.csv.bz2"
##  [3] "covid-global_02-01-2021.csv.bz2" "covid-global_03-01-2020.csv.bz2"
##  [5] "covid-global_03-01-2021.csv.bz2" "covid-global_04-01-2020.csv.bz2"
##  [7] "covid-global_04-01-2021.csv.bz2" "covid-global_05-01-2020.csv.bz2"
##  [9] "covid-global_05-01-2021.csv.bz2" "covid-global_06-01-2020.csv.bz2"
## [11] "covid-global_06-01-2021.csv.bz2" "covid-global_07-01-2020.csv.bz2"
## [13] "covid-global_07-01-2021.csv.bz2" "covid-global_08-01-2020.csv.bz2"
## [15] "covid-global_08-01-2021.csv.bz2" "covid-global_09-01-2020.csv.bz2"
## [17] "covid-global_09-01-2021.csv.bz2" "covid-global_10-01-2020.csv.bz2"
## [19] "covid-global_10-01-2021.csv.bz2" "covid-global_11-01-2020.csv.bz2"
## [21] "covid-global_12-01-2020.csv.bz2"
```

```
#list.files produces a character vector so we can find the length directly.
length(vec)
```

```
## [1] 21
```

Loaded the COVID data file for October 2021.Since the global data file contains not just African countries, I just selected the African ones from the list. Unfortunately not all the names match.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library (dplyr)
october2021= read.delim("covid/covid-global_10-01-2021.csv.bz2")
head(october2021)
```

```
##   FIPS Admin2 Province_State      Country_Region          Last_Update       Lat
## 1   NA                            Afghanistan 2021-10-02T04:21:25Z  33.93911
## 2   NA                                Albania 2021-10-02T04:21:25Z  41.15330
## 3   NA                                Algeria 2021-10-02T04:21:25Z  28.03390
## 4   NA                                Andorra 2021-10-02T04:21:25Z  42.50630
## 5   NA                                 Angola 2021-10-02T04:21:25Z -11.20270
## 6   NA                     Antigua and Barbuda 2021-10-02T04:21:25Z  17.06080
##       Long_ Confirmed Deaths Recovered Active        Combined_Key Incident_Rate
## 1  67.70995    155191   7206        NA     NA         Afghanistan      398.6581
## 2  20.16830    170778   2705        NA     NA             Albania     5934.3248
## 3   1.65960    203517   5815        NA     NA             Algeria      464.1098
## 4   1.52180     15222    130        NA     NA             Andorra    19701.0289
## 5  17.87390     58076   1567        NA     NA              Angola      176.7040
## 6 -61.79640      3336     81        NA     NA Antigua and Barbuda     3406.5844
##   Case_Fatality_Ratio
## 1           4.6433105
```

```
## 2           1.5839277
## 3           2.8572552
## 4           0.8540271
## 5           2.6981886
## 6           2.4280576
```

```
# checking the names of the columns in Oct 2021 data
names(october2021)
```

```
##  [1] "FIPS"                "Admin2"              "Province_State"
##  [4] "Country_Region"      "Last_Update"         "Lat"
##  [7] "Long_"               "Confirmed"           "Deaths"
## [10] "Recovered"           "Active"              "Combined_Key"
## [13] "Incident_Rate"       "Case_Fatality_Ratio"
```

49 countries can be found in Oct 2021 covid data from African countries list

```
library(tidyverse)
library (dplyr)
# filtering oct data based on African countries

AfricaOct2021<- october2021[october2021$Country_Region %in% Africa$country,]

# tried two methods

AfricaOct2021 <- filter(october2021,october2021$Country_Region %in% Africa$country)
head(AfricaOct2021)
```

```
##   FIPS Admin2 Province_State Country_Region          Last_Update      Lat
## 1   NA                            Algeria 2021-10-02T04:21:25Z  28.0339
## 2   NA                             Angola 2021-10-02T04:21:25Z -11.2027
## 3   NA                              Benin 2021-10-02T04:21:25Z   9.3077
## 4   NA                           Botswana 2021-10-02T04:21:25Z -22.3285
## 5   NA                       Burkina Faso 2021-10-02T04:21:25Z  12.2383
## 6   NA                            Burundi 2021-10-02T04:21:25Z  -3.3731
##     Long_ Confirmed Deaths Recovered Active Combined_Key Incident_Rate
## 1  1.6596    203517   5815        NA     NA      Algeria     464.10983
## 2 17.8739     58076   1567        NA     NA       Angola     176.70397
## 3  2.3158     23890    159        NA     NA        Benin     197.06021
## 4 24.6849    179220   2368        NA     NA     Botswana    7621.11306
## 5 -1.5616     14262    184        NA     NA Burkina Faso      68.22853
## 6 29.9189     17979     38        NA     NA      Burundi     151.20117
##   Case_Fatality_Ratio
## 1           2.8572552
## 2           2.6981886
## 3           0.6655504
## 4           1.3212811
## 5           1.2901416
## 6           0.2113577
```

```
nrow(AfricaOct2021)
```

```
## [1] 49
```

The 9 countries below are not matched in covid data.

```
`%notin%` <- Negate(`%in%`)
MissingAfricanCountries <- filter(Africa,Africa$country %notin% october2021$Country_Region)
MissingAfricanCountries
```

```
##   rank                                        country  population
## 1    3                 Democratic Republic of the Congo 102,561,403
## 2   16                                    Ivory Coast  22,671,331
## 3   39                              Republic of the Congo   3,697,490
## 4   NA                                   Réunion (France)     840,974
## 5   NA                                     Western Sahara     510,713
## 6   52                                         Cape Verde     491,875
## 7   NA                                   Mayotte (France)     212,600
## 8   53                              São Tomé and Príncipe     201,784
## 9   NA Saint Helena, Ascension and Tristan da Cunha (UK)       5,633
##                year                                          source
## 1      July 1, 2020                  National annual projection
## 2      May 15, 2014               Preliminary 2014 census result
## 3    April 28, 2007                           2007 census result
## 4   January 1, 2013                            Official estimate
## 5 September 2, 2014               Preliminary 2014 census result
## 6     June 16, 2010 Final 2010 census result[permanent dead link]
## 7   August 21, 2012                           2012 census result
## 8              2018                            Official estimate
## 9         June 2016                           2016 census result
```

We should care more about these countries as their population is higher than the smaller islands. If these countries are ignored, the number of cases and deaths due to covid in Africa could be miscalculated entirely. The data we would be using further would not stay accurate.

Next, find how are the names of these three countries (Two Congos and Ivory Coast) written in the covid data. After looking through the Oct 2021 data's list of countries, I found that these 3 countries are written as Congo (Kinshasa), Congo (Brazzaville) and Cote d'Ivoire.

```
names<-october2021 %>% filter(str_detect(october2021$Country_Region, "^Congo|^Cote"))
names
```

```
##   FIPS Admin2 Province_State      Country_Region          Last_Update     Lat
## 1   NA                       Congo (Brazzaville) 2021-10-02T04:21:25Z -0.2280
## 2   NA                          Congo (Kinshasa) 2021-10-02T04:21:25Z -4.0383
## 3   NA                             Cote d'Ivoire 2021-10-02T04:21:25Z  7.5400
##     Long_ Confirmed Deaths Recovered Active        Combined_Key Incident_Rate
## 1 15.8277     14359    197        NA     NA Congo (Brazzaville)     260.21676
## 2 21.7587     56997   1084        NA     NA    Congo (Kinshasa)      63.64014
## 3 -5.5471     60335    631        NA     NA       Cote d'Ivoire     228.72989
##   Case_Fatality_Ratio
## 1            1.371962
## 2            1.901854
## 3            1.045827
```

I replaced Ivory Coast in the list of African countries with Cote d'Ivoire, Democratic Republic of the Congo with Congo (Kinshasa) and Republic of the Congo with Congo (Brazzaville).

```
Africa$country[Africa$country =='Ivory Coast']<-"Cote d'Ivoire"
Africa$country[Africa$country=='Democratic Republic of the Congo']<-"Congo (Kinshasa)"
Africa$country[Africa$country=='Republic of the Congo']<-"Congo (Brazzaville)"
head(Africa)
```

```
##   rank          country  population             year
## 1    1          Nigeria 206,139,589             2020
## 2    2         Ethiopia 109,224,414             2018
## 3    3 Congo (Kinshasa) 102,561,403     July 1, 2020
## 4    4            Egypt 101,334,404 December 8, 2020
## 5    5     South Africa  59,956,820 December 1, 2020
## 6    6         Tanzania  59,734,218             2020
##                            source
## 1                Worldometers[3]
## 2 UN population projections[4][5]
## 3      National annual projection
## 4       National population clock
## 5               Official estimate
## 6                    Worldometers
```

```
# I am only left with "Réunion (France)", "Western Sahara", "Cape Verde", "Mayotte (France)", "São Tomé

`%notin%` <- Negate(`%in%`)
MissingAfricanCountries <- filter(Africa,Africa$country %notin% october2021$Country_Region)
MissingAfricanCountries
```

```
##   rank                                        country population
## 1   NA                                Réunion (France)    840,974
## 2   NA                                  Western Sahara    510,713
## 3   52                                      Cape Verde    491,875
## 4   NA                                Mayotte (France)    212,600
## 5   53                           São Tomé and Príncipe    201,784
## 6   NA Saint Helena, Ascension and Tristan da Cunha (UK)      5,633
##                year                                          source
## 1   January 1, 2013                            Official estimate
## 2 September 2, 2014                Preliminary 2014 census result
## 3     June 16, 2010 Final 2010 census result[permanent dead link]
## 4   August 21, 2012                           2012 census result
## 5              2018                            Official estimate
## 6         June 2016                           2016 census result
```

The file name is written as "covid-global_–.csv.bz2", and date always "01" in these files. Extracting the date part from the first file name as Date object.

Answer: Removing the date part of file "covid-global_01-01-2021.csv.bz2" and storing it in a object of date type.

```
filename="covid-global_01-01-2021.csv.bz2"
filename=gsub("^[^_]*_","", filename)
filename=gsub("\\..*","",filename)
filename
```

```
## [1] "01-01-2021"
```

```
firstDate= as.Date(filename, "%m-%d-%Y")
firstDate
```

```
## [1] "2021-01-01"
```

```
class(firstDate)
```

```
## [1] "Date"
```

Now I merge all the data files into one.

```
library(tidyverse)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
#creating a merged df by binding data from each file after performing processes to extract date, select
mergedDf<-NULL
for (item in vec) {
  df <- read.delim(paste0("covid/", item))
 item=gsub("^[^_]*_","", item)
 item=gsub("\\..*","",item)
 firstDate= as.Date(item, "%m-%d-%Y")
  df$year=year(firstDate)
  df$month=month(firstDate)
  selectedDf=df%>%select(starts_with("Country"),Deaths, year, month)
  selectedDf=selectedDf%>%rename(country=starts_with("Country"))
  mergedDf=rbind(mergedDf,selectedDf)
}
mergedDf<- mergedDf[mergedDf$country %in% Africa$country,]
head(mergedDf)
```

```
##         country Deaths year month
## 3       Algeria   2762 2021     1
## 5        Angola    405 2021     1
## 37        Benin     44 2021     1
## 41     Botswana     42 2021     1
## 71 Burkina Faso     85 2021     1
## 73      Burundi      2 2021     1
```

```
nrow(mergedDf)
```

```
## [1] 986
```

Extracted the population size from the dataset of African countries.

```
AfricaPopulationDf= Africa%>%select(country,population)
head(AfricaPopulationDf)
```

```
##            country  population
## 1          Nigeria 206,139,589
## 2         Ethiopia 109,224,414
## 3 Congo (Kinshasa) 102,561,403
## 4            Egypt 101,334,404
## 5     South Africa  59,956,820
## 6         Tanzania  59,734,218
```

For each country, computed the death rate: number of deaths per 1M population.

```
#Merging AfricanPopulationDf with Africa's covid data
combinedDf=merge(mergedDf,AfricaPopulationDf,by.x="country", by.y="country", all=TRUE)
combinedDf$population <- as.numeric(gsub(",","",combinedDf$population))
combinedDf$deathRate<-(combinedDf$Deaths*100000)/as.numeric(combinedDf$population)
head(combinedDf)
```

```
##   country Deaths year month population deathRate
## 1 Algeria   2762 2021     1   43000420  6.423193
## 2 Algeria   5302 2021     9   43000420 12.330112
## 3 Algeria   3261 2021     5   43000420  7.583647
## 4 Algeria   1518 2020     9   43000420  3.530198
## 5 Algeria    453 2020     5   43000420  1.053478
## 6 Algeria   4291 2021     8   43000420  9.978972
```

Analyzing which 10 countries have the largest death rate? (As of the latest date in the data, Oct 1st, 2021).

```
topTen=combinedDf[order(combinedDf$deathRate, decreasing = TRUE), ]
topTen=topTen%>%filter(topTen$year=="2021" & topTen$month=="10")
head(topTen,10)
```

```
##         country Deaths year month population deathRate
## 1       Tunisia  24901 2021    10   10982754 226.72820
## 2       Namibia   3514 2021    10    2280700 154.07550
## 3  South Africa  87705 2021    10   59956820 146.28027
## 4    Seychelles    112 2021    10      90945 123.15136
## 5      Botswana   2368 2021    10    2024904 116.94382
## 6      Eswatini   1223 2021    10    1119375 109.25740
## 7         Libya   4664 2021    10    5298152  88.03069
## 8       Morocco  14290 2021    10   37034729  38.58540
## 9      Zimbabwe   4624 2021    10   13061239  35.40246
## 10      Lesotho    633 2021    10    2007201  31.53645
```

```
topTenCountries=head(topTen$country,10)
topTenCountries
```

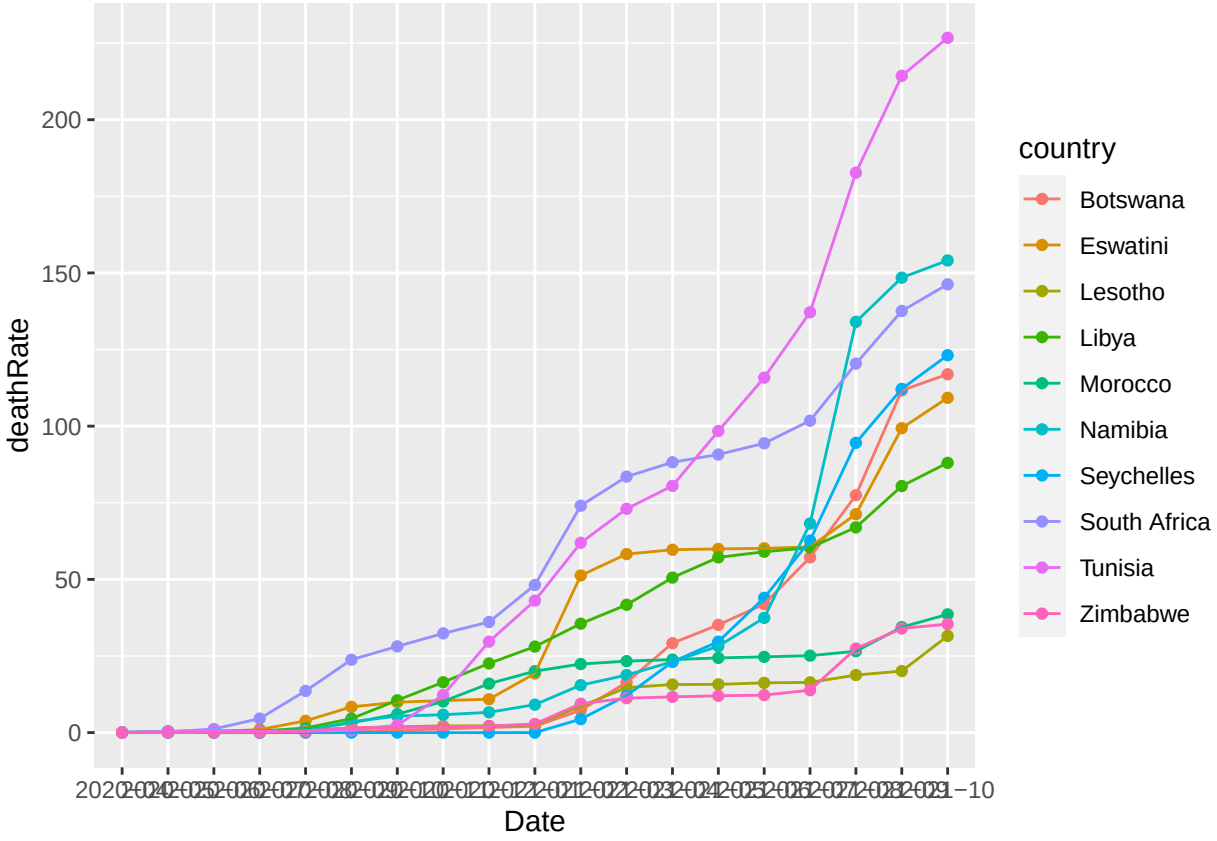
```
##  [1] "Tunisia"      "Namibia"      "South Africa" "Seychelles"   "Botswana"
##  [6] "Eswatini"     "Libya"        "Morocco"      "Zimbabwe"     "Lesotho"
```

Plotting the growth in death rate in these 10 countries over time.

```
dfToVisualize= combinedDf[combinedDf$country %in% topTenCountries,]
dfToVisualize=within(dfToVisualize, Date <- sprintf("%d-%02d", year, month))
head(dfToVisualize)
```

```
##     country Deaths year month population   deathRate    Date
## 59 Botswana     42 2021     1    2024904  2.07417241 2021-01
## 60 Botswana    712 2021     5    2024904 35.16216077 2021-05
## 61 Botswana      1 2020     6    2024904  0.04938506 2020-06
## 62 Botswana   1158 2021     7    2024904 57.18789632 2021-07
## 63 Botswana      6 2020     9    2024904  0.29631034 2020-09
## 64 Botswana   1569 2021     8    2024904 77.48515485 2021-08
```

```
ggplot(dfToVisualize, aes(x=Date, y=deathRate, fill=country, group=country)) + geom_point(aes(color = c
```



Computing the number of new monthly deaths (per 1M population) and displaying it on a similar plot.

```
#correcting the order of data for using lag
dfToVisualize=dfToVisualize[order(dfToVisualize$country, dfToVisualize$Date),]
#calculating the difference in the deaths per month, grouped by country
dfToVisualize$monthly_Deaths <- ave(dfToVisualize$Deaths, factor(dfToVisualize$country), FUN=function(x
head(dfToVisualize)
```

```
##     country Deaths year month population  deathRate    Date monthly_Deaths
## 71 Botswana      1 2020     4    2024904 0.04938506 2020-04             NA
## 65 Botswana      1 2020     5    2024904 0.04938506 2020-05              0
## 61 Botswana      1 2020     6    2024904 0.04938506 2020-06              0
## 77 Botswana      1 2020     7    2024904 0.04938506 2020-07              0
## 69 Botswana      2 2020     8    2024904 0.09877011 2020-08              1
## 63 Botswana      6 2020     9    2024904 0.29631034 2020-09              4
```

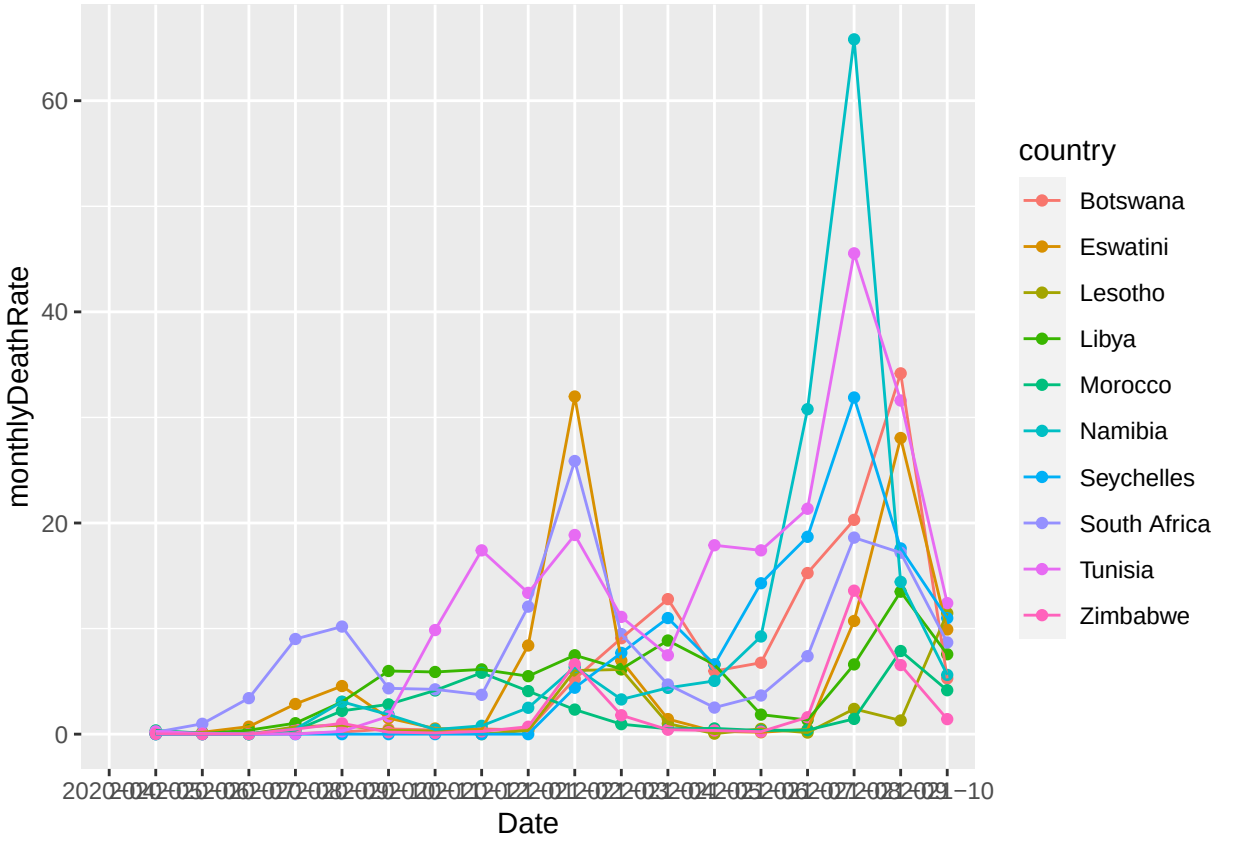
```
# now that we have monthly deaths, we can calculate monthly death rate and append that column to the df
dfToVisualize$monthlyDeathRate=dfToVisualize$monthly_Deaths*100000/dfToVisualize$population
head(dfToVisualize)
```

```
##     country Deaths year month population  deathRate    Date monthly_Deaths
## 71 Botswana      1 2020     4    2024904 0.04938506 2020-04             NA
## 65 Botswana      1 2020     5    2024904 0.04938506 2020-05              0
## 61 Botswana      1 2020     6    2024904 0.04938506 2020-06              0
## 77 Botswana      1 2020     7    2024904 0.04938506 2020-07              0
## 69 Botswana      2 2020     8    2024904 0.09877011 2020-08              1
## 63 Botswana      6 2020     9    2024904 0.29631034 2020-09              4
##    monthlyDeathRate
## 71               NA
## 65       0.00000000
## 61       0.00000000
## 77       0.00000000
## 69       0.04938506
## 63       0.19754023
```

```
# now we display this data
ggplot(dfToVisualize, aes(x=Date, y=monthlyDeathRate, fill=country, group=country)) + geom_point(aes(co
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```

```
## Warning: Removed 10 row(s) containing missing values (geom_path).
```



Namibia experienced the highest peak in the new monthly deaths. It was in 2020-08 (Aug-2020). I can see three waves of covid in the graph.