Analysis of Global COVID-19 Pandemic Data

In [4]:
```
getwd()
```

'C:/Users/OYENIRAN MATTHEW'

In [51]:
```
library(httr)
library(rvest)
```

In [52]:
```
get_wiki_covid19_page <- function() {

  # Our target COVID-19 wiki page URL is: https://en.wikipedia.org/w/index.p
  # Which has two parts:
  # 1) base URL `https://en.wikipedia.org/w/index.php
  # 2) URL parameter: `title=Template:COVID-19_testing_by_country`, seperate

  # Wiki page base
  wiki_base_url <-  "https://en.wikipedia.org/w/index.php"

  # You will need to create a List which has an element called `title` to sp
  # in our case, it will be `Template:COVID-19_testing_by_country`
  wiki_params <- list(title = "Template:COVID-19_testing_by_country")


  # - Use the `GET` function in httr library with a `url` argument and a `qu
  wiki_response <- GET(wiki_base_url, query = wiki_params)

  # Use the `return` function to return the response
  return(wiki_response)
}
```

In [53]:
```
# Call the get_wiki_covid19_page function and print the response
wiki_covid19_page_response <- get_wiki_covid19_page()
print(wiki_covid19_page_response)
```

```
Response [https://en.wikipedia.org/w/index.php?title=Template%3ACOVID-19_testing_by_country]
  Date: 2022-07-27 21:57
  Status: 200
  Content-Type: text/html; charset=UTF-8
  Size: 414 kB
```

In [55]:
```
# Get the root html node from the http response in task 1
wiki_covid19_page_root_node <- read_html(wiki_covid19_page_response)
```

In [56]:
```
# Get the table node from the root html node
wiki_covid19_page_table_node <- html_node(wiki_covid19_page_root_node, "tabl
```

In [57]:
```
# Read the table node and convert it into a data frame, and print the data f
wiki_covid19_page_data_frame <- html_table(wiki_covid19_page_table_node )
wiki_covid19_page_data_frame
```

Country or region	Date[a]	Tested	Units[b]	Confirmed(cases)
Afghanistan	17 Dec 2020	154,767	samples	49,621
Albania	18 Feb 2021	428,654	samples	96,838
Algeria	2 Nov 2020	230,553	samples	58,574
Andorra	23 Feb 2022	300,307	samples	37,958
Angola	2 Feb 2021	399,228	samples	20,981
Antigua and Barbuda	6 Mar 2021	15,268	samples	832
Argentina	16 Apr 2022	35,716,069	samples	9,060,495
Armenia	29 May 2022	3,099,602	samples	422,963
Australia	22 Jul 2022	75,558,764	samples	9,019,965
Austria	26 Jul 2022	192,947,042	samples	4,753,092
Azerbaijan	11 May 2022	6,838,458	samples	792,638

In [58]:
```
# Print the summary of the data frame
summary(wiki_covid19_page_data_frame)
```

```
 Country or region    Date[a]             Tested            Units[b]
 Length:173         Length:173         Length:173         Length:173
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Confirmed(cases)   Confirmed /tested,% Tested /population,%
 Length:173         Length:173          Length:173
 Class :character   Class :character    Class :character
 Mode  :character   Mode  :character    Mode  :character
 Confirmed /population,%     Ref.
 Length:173              Length:173
 Class :character        Class :character
 Mode  :character        Mode  :character
```

In [59]:
```
preprocess_covid_data_frame <- function(data_frame) {

  shape <- dim(data_frame)
```

```
  # Remove the World row
  data_frame<-data_frame[!(data_frame$`Country or region`=="World"),]
  # Remove the last row
  data_frame <- data_frame[1:172, ]

  # We dont need the Units and Ref columns, so can be removed
  data_frame["Ref."] <- NULL
  data_frame["Units[b]"] <- NULL

  # Renaming the columns
  names(data_frame) <- c("country", "date", "tested", "confirmed", "confirm

  # Convert column data types
  data_frame$country <- as.factor(data_frame$country)
  data_frame$date <- as.factor(data_frame$date)
  data_frame$tested <- as.numeric(gsub(",","",data_frame$tested))
  data_frame$confirmed <- as.numeric(gsub(",","",data_frame$confirmed))
  data_frame$'confirmed.tested.ratio' <- as.numeric(gsub(",","",data_frame$
  data_frame$'tested.population.ratio' <- as.numeric(gsub(",","",data_frame
  data_frame$'confirmed.population.ratio' <- as.numeric(gsub(",","",data_fr

  return(data_frame)
}
```

In [60]:
```
# call `preprocess_covid_data_frame` function and assign it to a new data fr
new_covid_data_frame <- preprocess_covid_data_frame(wiki_covid19_page_data_

head(new_covid_data_frame)
```

country	date	tested	confirmed	confirmed.tested.ratio	tested.population.ratio	confirmed
Afghanistan	17 Dec 2020	154767	49621	32.1	0.40	
Albania	18 Feb 2021	428654	96838	22.6	15.00	
Algeria	2 Nov 2020	230553	58574	25.4	0.53	
Andorra	23 Feb 2022	300307	37958	12.6	387.00	
Angola	2 Feb 2021	399228	20981	5.3	1.30	
Antigua and Barbuda	6 Mar 2021	15268	832	5.4	15.90	

In [61]:
```
# Print the summary of the processed data frame again
```

```
                country             date         tested
 Afghanistan        :  1   22 Jul 2022: 15   Min.   :     3880
 Albania            :  1   23 Jul 2022:  8   1st Qu.:   512037
 Algeria            :  1   21 Jul 2022:  4   Median :  3029859
 Andorra            :  1   1 Mar 2021 :  3   Mean   : 30563584
 Angola             :  1   17 Jul 2022:  3   3rd Qu.: 11797975
 Antigua and Barbuda:  1   18 Jul 2022:  3   Max.   :925534224
 (Other)            :166   (Other)    :136
   confirmed        confirmed.tested.ratio tested.population.ratio
 Min.   :       0   Min.   : 0.00          Min.   :   0.0065
 1st Qu.:   37618   1st Qu.: 5.00          1st Qu.:   9.3500
 Median :  281196   Median : 9.85          Median :  46.5000
 Mean   : 2387071   Mean   :11.01          Mean   : 168.2240
 3rd Qu.: 1181110   3rd Qu.:15.25          3rd Qu.: 145.5000
 Max.   :89824190   Max.   :42.80          Max.   :3003.0000

 confirmed.population.ratio
 Min.   : 0.000
 1st Qu.: 0.425
```

In [62]:
```
# Export the data frame to a csv file
```

In [93]:
```
df= read.csv("C:/Users/OYENIRAN MATTHEW/Documents/new_covid_data.csv")
subset <- subset(df, select=c('country', 'confirmed'))
```

	country	confirmed
5	Angola	20981
6	Antigua and Barbuda	832
7	Argentina	9060495
8	Armenia	422963
9	Australia	9019965
10	Austria	4753092

Get a subset of the extracted data frame

In [66]:
```
# Get the total confirmed cases worldwide
total_confirmed <- subset(df, select = c('confirmed'))
x = sum(total_confirmed)
```

410576153

In [67]:
```
# Get the total tested cases worldwide
total_tested <- subset(df, select = c('tested'))
y = sum(total_tested)
```

5256936385

In [68]:
```
# Get the positive ratio (confirmed / tested)
positive_ratio = x/y
positive_ratio
```

0.0781017921714873

TASK 6: Get a country list which reported their testing data

In [69]:
```
# Get the `country` column
country <- df[,'country']
```

Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria Azerbaijan Bahamas Bahrain Bangladesh Barbados Belarus Belgium Belize Benin Bhutan Bolivia Bosnia and Herzegovina Botswana Brazil Brunei Bulgaria Burkina Faso Burundi Cambodia Cameroon Canada Chad Chile China[c] Colombia Costa Rica Croatia Cuba Cyprus[d] Czechia Denmark[e] Djibouti Dominica Dominican Republic DR Congo Ecuador Egypt El Salvador Equatorial Guinea Estonia Eswatini Ethiopia Faroe Islands Fiji Finland France[f][g] Gabon Gambia Georgia[h] Germany Ghana Greece Greenland Grenada Guatemala Guinea Guinea-Bissau Guyana Haiti Honduras Hungary Iceland India Indonesia Iran Iraq Ireland Israel Italy Ivory Coast Jamaica Japan Jordan Kazakhstan Kenya Kosovo Kuwait Kyrgyzstan Laos Latvia Lebanon Lesotho Liberia Libya Lithuania Luxembourg[i] Madagascar Malawi Malaysia Maldives Mali Malta Mauritania Mauritius Mexico Moldova[j] Mongolia Montenegro Morocco Mozambique Myanmar Namibia Nepal Netherlands New Caledonia New Zealand Niger Nigeria North Korea North Macedonia Northern Cyprus[k] Norway Oman Pakistan Palestine Panama Papua New Guinea Paraguay Peru Philippines Poland Portugal Qatar Romania Russia Rwanda Saint Kitts and Nevis Saint Lucia Saint Vincent San Marino Saudi Arabia Senegal Serbia Singapore Slovakia Slovenia South Africa South Korea South Sudan Spain Sri Lanka Sudan Sweden Switzerland[l] Taiwan[m] Tanzania Thailand Togo Trinidad and Tobago Tunisia Turkey Uganda Ukraine United Arab Emirates United Kingdom United States Uruguay Uzbekistan Venezuela Vietnam Zambia Zimbabwe

▸ **Levels**:

In [70]:
```
# Check its class (should be Factor)
```

'factor'

In [71]:
```
# Conver the country column into character so that you can easily sort them
country_new = as.character(country)
```

'character'

In [77]:
```
#Sort the countries AtoZ
x = sort(country_new)
```

'Afghanistan' 'Albania' 'Algeria' 'Andorra' 'Angola' 'Antigua and Barbuda' 'Argentina' 'Armenia' 'Australia' 'Austria' 'Azerbaijan' 'Bahamas' 'Bahrain' 'Bangladesh' 'Barbados' 'Belarus' 'Belgium' 'Belize' 'Benin' 'Bhutan' 'Bolivia' 'Bosnia and Herzegovina' 'Botswana' 'Brazil' 'Brunei' 'Bulgaria' 'Burkina Faso' 'Burundi' 'Cambodia' 'Cameroon' 'Canada' 'Chad' 'Chile' 'China[c]' 'Colombia' 'Costa Rica' 'Croatia' 'Cuba' 'Cyprus[d]' 'Czechia' 'Denmark[e]' 'Djibouti' 'Dominica' 'Dominican Republic' 'DR Congo' 'Ecuador' 'Egypt' 'El Salvador' 'Equatorial Guinea' 'Estonia' 'Eswatini' 'Ethiopia' 'Faroe Islands' 'Fiji' 'Finland' 'France[f][g]' 'Gabon' 'Gambia' 'Georgia[h]' 'Germany' 'Ghana' 'Greece' 'Greenland' 'Grenada' 'Guatemala' 'Guinea' 'Guinea-Bissau' 'Guyana' 'Haiti'

'Honduras' 'Hungary' 'Iceland' 'India' 'Indonesia' 'Iran' 'Iraq' 'Ireland' 'Israel'
'Italy' 'Ivory Coast' 'Jamaica' 'Japan' 'Jordan' 'Kazakhstan' 'Kenya' 'Kosovo'
'Kuwait' 'Kyrgyzstan' 'Laos' 'Latvia' 'Lebanon' 'Lesotho' 'Liberia' 'Libya'
'Lithuania' 'Luxembourg[i]' 'Madagascar' 'Malawi' 'Malaysia' 'Maldives' 'Mali'
'Malta' 'Mauritania' 'Mauritius' 'Mexico' 'Moldova[j]' 'Mongolia' 'Montenegro'
'Morocco' 'Mozambique' 'Myanmar' 'Namibia' 'Nepal' 'Netherlands'
'New Caledonia' 'New Zealand' 'Niger' 'Nigeria' 'North Korea' 'North Macedonia'
'Northern Cyprus[k]' 'Norway' 'Oman' 'Pakistan' 'Palestine' 'Panama'
'Papua New Guinea' 'Paraguay' 'Peru' 'Philippines' 'Poland' 'Portugal' 'Qatar'
'Romania' 'Russia' 'Rwanda' 'Saint Kitts and Nevis' 'Saint Lucia' 'Saint Vincent'
'San Marino' 'Saudi Arabia' 'Senegal' 'Serbia' 'Singapore' 'Slovakia' 'Slovenia'
'South Africa' 'South Korea' 'South Sudan' 'Spain' 'Sri Lanka' 'Sudan' 'Sweden'
'Switzerland[l]' 'Taiwan[m]' 'Tanzania' 'Thailand' 'Togo' 'Trinidad and Tobago'
'Tunisia' 'Turkey' 'Uganda' 'Ukraine' 'United Arab Emirates' 'United Kingdom'
'United States' 'Uruguay' 'Uzbekistan' 'Venezuela' 'Vietnam' 'Zambia' 'Zimbabwe'

In [73]:
```
# Sort the countries ZtoA
sorted = sort(country_new, decreasing = TRUE)

#Print the sorted ZtoA list
```

'Zimbabwe' 'Zambia' 'Vietnam' 'Venezuela' 'Uzbekistan' 'Uruguay' 'United States'
'United Kingdom' 'United Arab Emirates' 'Ukraine' 'Uganda' 'Turkey' 'Tunisia'
'Trinidad and Tobago' 'Togo' 'Thailand' 'Tanzania' 'Taiwan[m]' 'Switzerland[l]'
'Sweden' 'Sudan' 'Sri Lanka' 'Spain' 'South Sudan' 'South Korea' 'South Africa'
'Slovenia' 'Slovakia' 'Singapore' 'Serbia' 'Senegal' 'Saudi Arabia' 'San Marino'
'Saint Vincent' 'Saint Lucia' 'Saint Kitts and Nevis' 'Rwanda' 'Russia' 'Romania'
'Qatar' 'Portugal' 'Poland' 'Philippines' 'Peru' 'Paraguay' 'Papua New Guinea'
'Panama' 'Palestine' 'Pakistan' 'Oman' 'Norway' 'Northern Cyprus[k]'
'North Macedonia' 'North Korea' 'Nigeria' 'Niger' 'New Zealand' 'New Caledonia'
'Netherlands' 'Nepal' 'Namibia' 'Myanmar' 'Mozambique' 'Morocco' 'Montenegro'
'Mongolia' 'Moldova[j]' 'Mexico' 'Mauritius' 'Mauritania' 'Malta' 'Mali' 'Maldives'
'Malaysia' 'Malawi' 'Madagascar' 'Luxembourg[i]' 'Lithuania' 'Libya' 'Liberia'
'Lesotho' 'Lebanon' 'Latvia' 'Laos' 'Kyrgyzstan' 'Kuwait' 'Kosovo' 'Kenya'
'Kazakhstan' 'Jordan' 'Japan' 'Jamaica' 'Ivory Coast' 'Italy' 'Israel' 'Ireland' 'Iraq'
'Iran' 'Indonesia' 'India' 'Iceland' 'Hungary' 'Honduras' 'Haiti' 'Guyana'
'Guinea-Bissau' 'Guinea' 'Guatemala' 'Grenada' 'Greenland' 'Greece' 'Ghana'
'Germany' 'Georgia[h]' 'Gambia' 'Gabon' 'France[f][g]' 'Finland' 'Fiji'
'Faroe Islands' 'Ethiopia' 'Eswatini' 'Estonia' 'Equatorial Guinea' 'El Salvador'
'Egypt' 'Ecuador' 'DR Congo' 'Dominican Republic' 'Dominica' 'Djibouti'
'Denmark[e]' 'Czechia' 'Cyprus[d]' 'Cuba' 'Croatia' 'Costa Rica' 'Colombia'
'China[c]' 'Chile' 'Chad' 'Canada' 'Cameroon' 'Cambodia' 'Burundi' 'Burkina Faso'
'Bulgaria' 'Brunei' 'Brazil' 'Botswana' 'Bosnia and Herzegovina' 'Bolivia' 'Bhutan'
'Benin' 'Belize' 'Belgium' 'Belarus' 'Barbados' 'Bangladesh' 'Bahrain' 'Bahamas'
'Azerbaijan' 'Austria' 'Australia' 'Armenia' 'Argentina' 'Antigua and Barbuda'
'Angola' 'Andorra' 'Algeria' 'Albania' 'Afghanistan'

Identify countries names with a specific pattern

In [74]:
```
#Identify countries names with a specific pattern
# Print the matched country names
```

```
x_match <- regmatches(country_new, regexpr("United.+", country_new))
```

'United Arab Emirates' 'United Kingdom' 'United States'

Pick two countries you are interested, and then review their testing data

In [76]:
```
# Select a subset (should be only one row) of data frame based on a selecte
select = subset(df, select=c('country', 'confirmed', 'confirmed.population.
selected_1 = select[1,]
```

country	confirmed	confirmed.population.ratio
Afghanistan	49621	0.13

In [83]:
```
select = subset(df, select=c('country', 'confirmed', 'confirmed.population.
selected_2 = select[5,]
```

	country	confirmed	confirmed.population.ratio
5	Angola	20981	0.067

Compare which one of the selected countries has a larger ratio of confirmed cases to population

In [85]:
```
# Use if-else statement

if (selected_1$confirmed.population > selected_2$confirmed.population) {
    print("Afghanistan has higher COVID-19 infection risk")
  } else {
    print("Angola has higher COVID-19 infection risk")
```

```
[1] "Afghanistan has higher COVID-19 infection risk"
```

Find countries with confirmed to population ratio rate less than a threshold

In [94]:
```
```

	country	date	tested	confirmed	confirmed.tested.ratio	tested.population.ratio
1	Afghanistan	17 Dec 2020	154767	49621	32.100	0.4000
3	Algeria	2 Nov 2020	230553	58574	25.400	0.5300
5	Angola	2 Feb 2021	399228	20981	5.300	1.3000
6	Antigua and Barbuda	6 Mar 2021	15268	832	5.400	15.9000

	country	date	tested	confirmed	confirmed.tested.ratio	tested.population.ratio
14	Bangladesh	24 Jul 2021	7417714	1151644	15.500	4.5000
19	Benin	4 May 2021	595112	7884	1.300	5.1000
25	Brunei	2 Aug 2021	153804	338	0.220	33.5000
27	Burkina Faso	4 Mar 2021	158777	12123	7.600	0.7600
28	Burundi	5 Jan 2021	90019	884	0.980	0.7600
29	Cambodia	1 Aug 2021	1812706	77914	4.300	11.2000
30	Cameroon	18 Feb 2021	942685	32681	3.500	3.6000
32	Chad	2 Mar 2021	99027	4020	4.100	0.7200
34	China[c]	31 Jul 2020	160000000	87655	0.055	11.1000
45	DR Congo	28 Feb 2021	124838	25961	20.800	0.1400
47	Egypt	23 Jul 2021	3137519	283947	9.100	3.1000
52	Ethiopia	24 Jun 2021	2981185	278446	9.300	2.6000
57	Gabon	23 Jul 2021	958807	25325	2.600	3.1000
58	Gambia	15 Feb 2021	43217	4469	10.300	2.0000
61	Ghana	3 Jul 2021	1305749	96708	7.400	4.2000
64	Grenada	11 May 2021	28684	161	0.560	25.7000
66	Guinea	21 Jul 2021	494898	24878	5.000	3.8000
67	Guinea-Bissau	7 Jul 2022	145231	8400	5.800	7.7000
69	Haiti	6 Jul 2022	210836	31980	15.200	1.8000

	country	date	tested	confirmed	confirmed.tested.ratio	tested.population.ratio
80	Ivory Coast	3 Mar 2021	429177	33285	7.800	1.6000
82	Japan	1 Mar 2021	8487288	432773	5.100	6.7000
85	Kenya	5 Mar 2021	1322806	107729	8.100	2.8000
89	Laos	1 Mar 2021	114030	45	0.039	1.6000
93	Liberia	17 Jul 2021	128246	5396	4.200	2.5000
97	Madagascar	19 Feb 2021	119608	19831	16.600	0.4600
98	Malawi	15 Jul 2022	596813	86900	14.600	3.1000
101	Mali	7 Jul 2021	322504	14449	4.500	1.6000
103	Mauritania	16 Apr 2021	268093	18103	6.800	6.1000
104	Mauritius	22 Nov 2020	289552	494	0.170	22.9000
110	Mozambique	22 Jul 2021	688570	105866	15.400	2.2000
111	Myanmar	16 Sep 2021	4047680	440741	10.900	7.4000
115	New Caledonia	3 Sep 2021	41962	136	0.320	15.7000
117	Niger	22 Feb 2021	79321	4740	6.000	0.3500
118	Nigeria	28 Feb 2021	1544008	155657	10.100	0.7500
119	North Korea	25 Nov 2020	16914	0	0.000	0.0660
124	Pakistan	5 Mar 2021	9173593	588728	6.400	4.2000
127	Papua New Guinea	17 Feb 2021	47490	961	2.000	0.5300

	country	date	tested	confirmed	confirmed.tested.ratio	tested.population.ratio
136	Rwanda	6 Oct 2021	2885812	98209	3.400	22.3000
142	Senegal	12 Jul 2021	624502	46509	7.400	3.9000
148	South Korea	1 Mar 2021	6592010	90029	1.400	12.7000
149	South Sudan	26 May 2021	164472	10688	6.500	1.3000
151	Sri Lanka	30 Mar 2021	2384745	93128	3.900	10.9000
152	Sudan	7 Jan 2021	158804	23316	14.700	0.3600
156	Tanzania	18 Nov 2020	3880	509	13.100	0.0065
157	Thailand	4 Mar 2021	1579597	26162	1.700	2.3000
158	Togo	23 Jul 2022	765539	37956	5.000	8.9000

In [ ]:

In [ ]: