TASK 4:

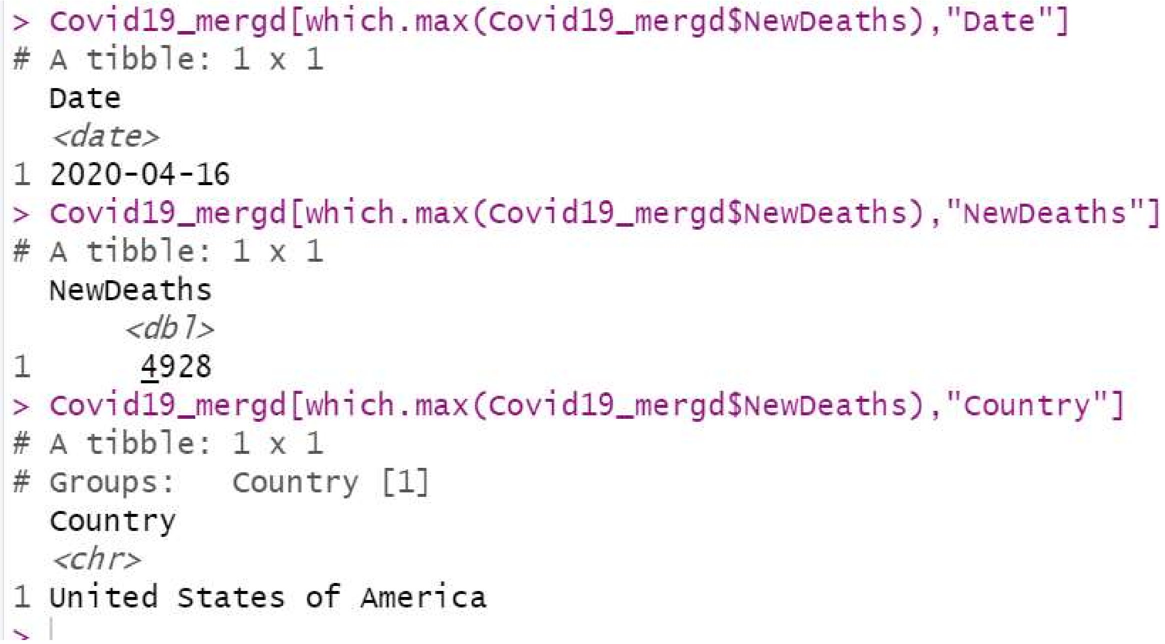
Results of Task 2:

Question 4. Find the day with the highest reported death toll across the world. Print the date and the death toll of that day.

Answer:

Highest reported death toll is on **16th April 2020** which is **4928** deaths in a single day.

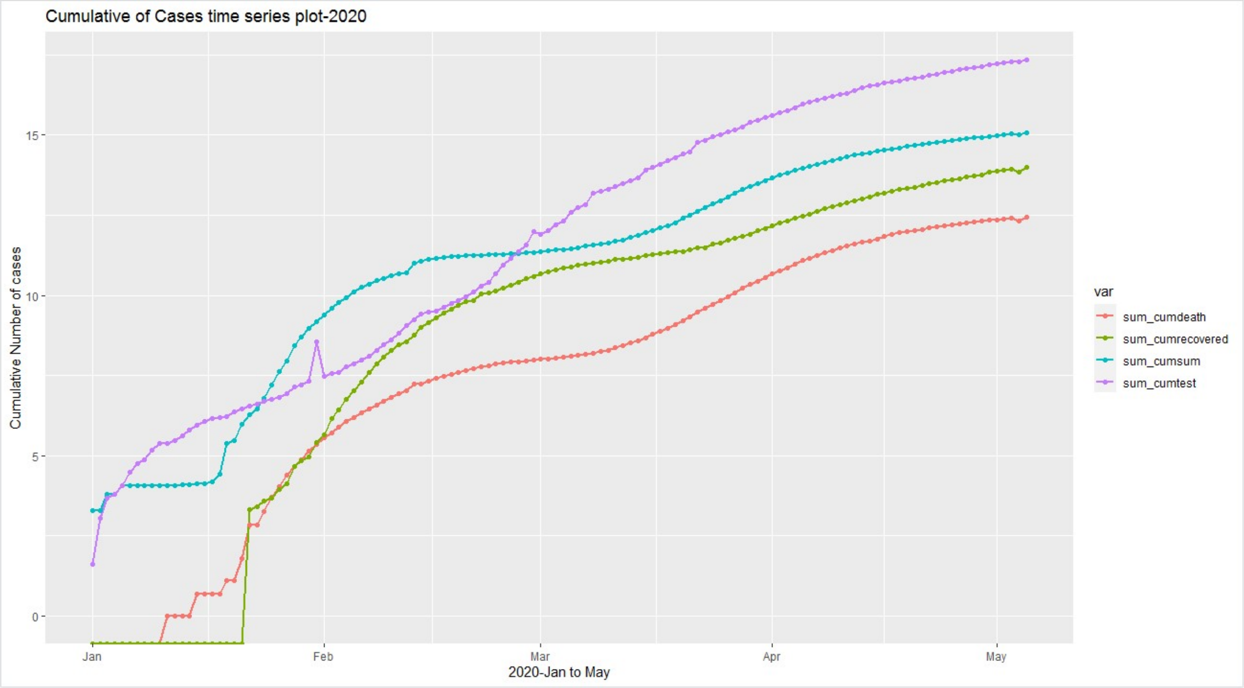
This is reported in **United States of America.**



```
> Covid19_mergd[which.max(Covid19_mergd$NewDeaths),"Date"]
# A tibble: 1 x 1
  Date
  <date>
1 2020-04-16
> Covid19_mergd[which.max(Covid19_mergd$NewDeaths),"NewDeaths"]
# A tibble: 1 x 1
  NewDeaths
      <dbl>
1      4928
> Covid19_mergd[which.max(Covid19_mergd$NewDeaths),"Country"]
# A tibble: 1 x 1
# Groups:   Country [1]
  Country
  <chr>
1 United States of America
>
```

5. Build a graph to show how the cumulative data of (Infected Cases, Deaths, Recovered, Tests) change over the time for the whole world collectively. [Hint: Use geom_line, use log for Y axis for better presentation, Use different colour to distinguish between new cases, deaths, and recovered]

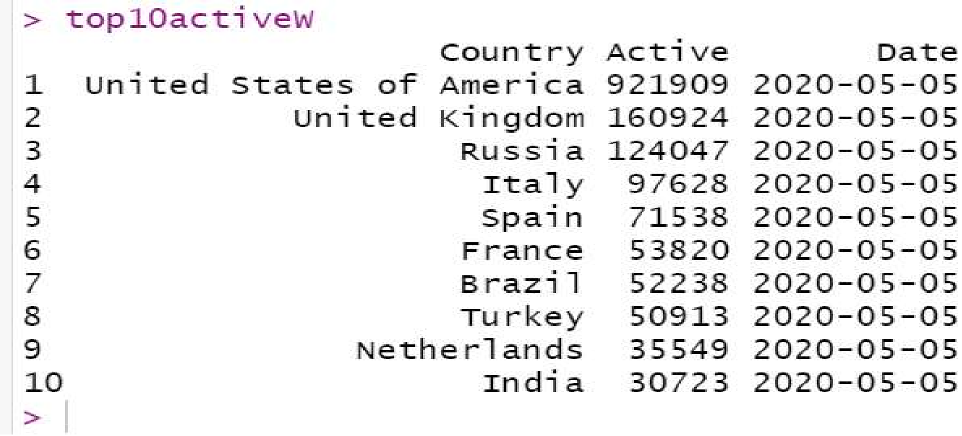
Answer: There is gradual increase in cases. Similarly, gradual increase in deaths . recovered and and tests has improved over time.



7. Based on the last day data, extract the whole records of the top 10 countries worldwide that have current active cases, total confirmed cases, and fatality rate in separate dataframes (i.e. top10activeW, top10casesW, top10fatalityW, top10testsMW). [Hint: you can use head(arranged_data, n=10) to get the top 10 records]

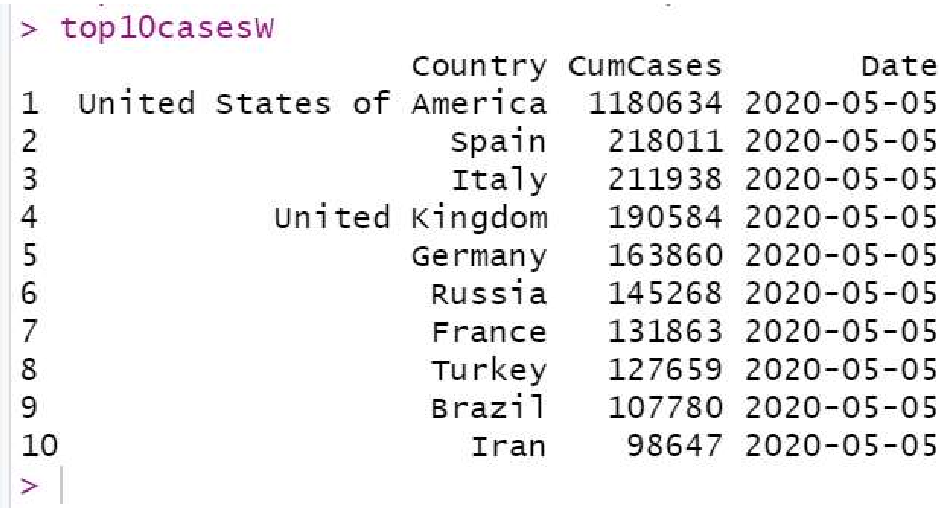
Answer:

Top ten countries with highest active cases on May 05th 2020 are as shown here.



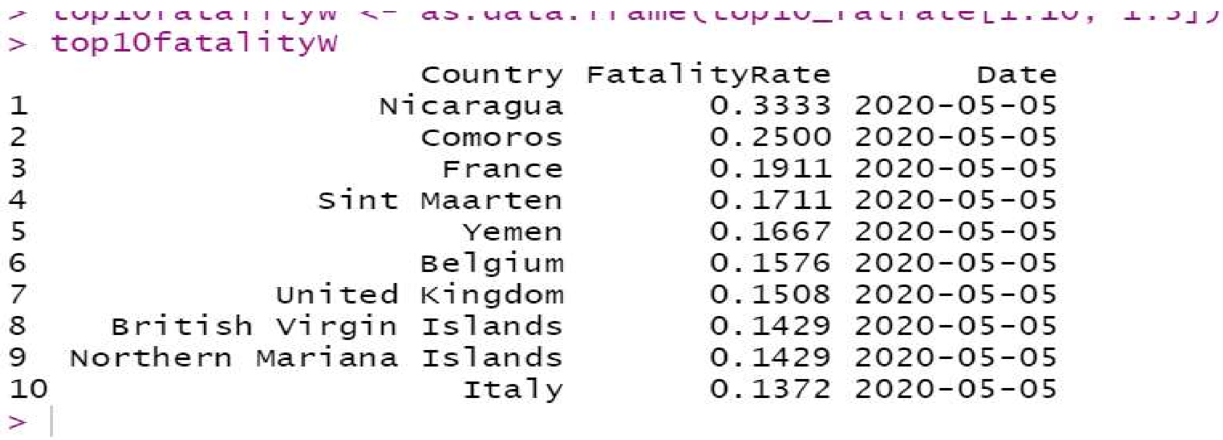
```
> top10activeW
                   Country Active       Date
1  United States of America 921909 2020-05-05
2           United Kingdom 160924 2020-05-05
3                   Russia 124047 2020-05-05
4                    Italy  97628 2020-05-05
5                    Spain  71538 2020-05-05
6                   France  53820 2020-05-05
7                   Brazil  52238 2020-05-05
8                   Turkey  50913 2020-05-05
9              Netherlands  35549 2020-05-05
10                   India  30723 2020-05-05
> 
```

Top ten countries with highest total number of cases on May 05th 2020 are as shown here.



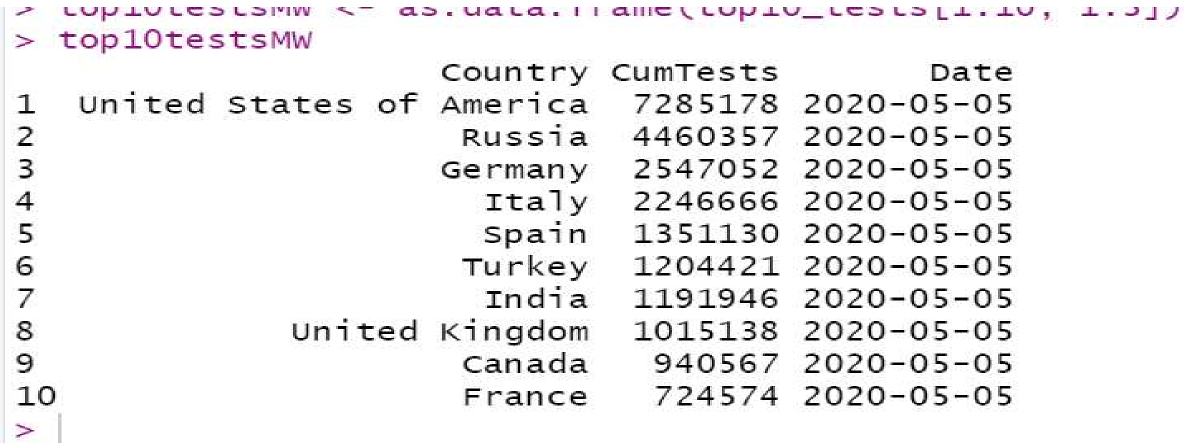
```
> top10casesW
                   Country CumCases       Date
1  United States of America  1180634 2020-05-05
2                     Spain   218011 2020-05-05
3                     Italy   211938 2020-05-05
4            United Kingdom   190584 2020-05-05
5                   Germany   163860 2020-05-05
6                    Russia   145268 2020-05-05
7                    France   131863 2020-05-05
8                    Turkey   127659 2020-05-05
9                    Brazil   107780 2020-05-05
10                     Iran    98647 2020-05-05
> 
```

Top ten countries with highest on May 05th 2020 are as shown here.



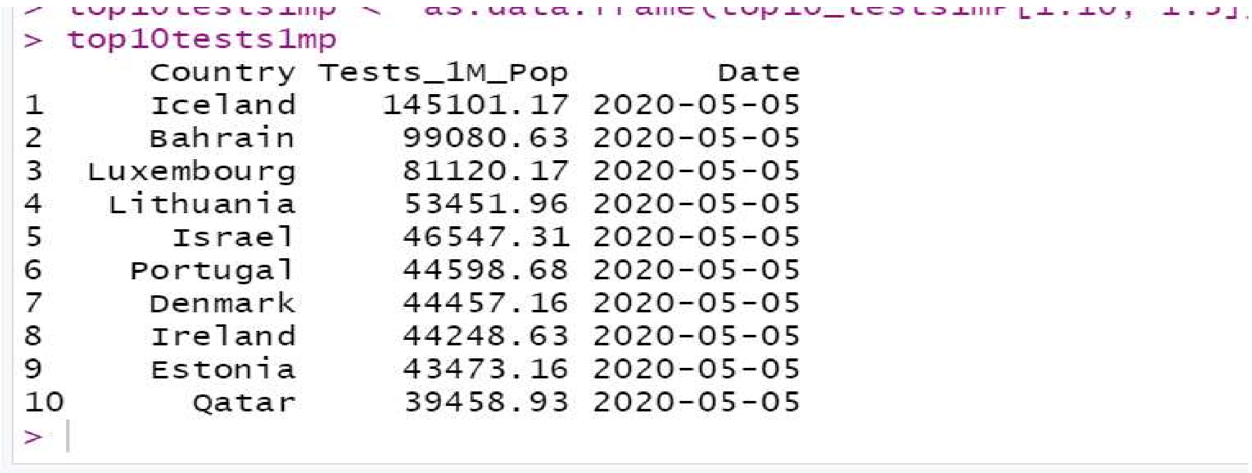
```
> top10fatalityW <- as.data.frame(top10_fatrate[1:10, 1:3])
> top10fatalityW
                    Country FatalityRate       Date
1                 Nicaragua       0.3333 2020-05-05
2                   Comoros       0.2500 2020-05-05
3                    France       0.1911 2020-05-05
4              Sint Maarten       0.1711 2020-05-05
5                     Yemen       0.1667 2020-05-05
6                   Belgium       0.1576 2020-05-05
7            United Kingdom       0.1508 2020-05-05
8    British Virgin Islands       0.1429 2020-05-05
9  Northern Mariana Islands       0.1429 2020-05-05
10                    Italy       0.1372 2020-05-05
> 
```

Top ten countries with highest number of tests as May 05th 2020 are as shown here.



```
> top10testsMW <- as.data.frame(top10_tests[1:10, 1:3])
> top10testsMW
                   Country CumTests       Date
1  United States of America  7285178 2020-05-05
2                    Russia  4460357 2020-05-05
3                   Germany  2547052 2020-05-05
4                     Italy  2246666 2020-05-05
5                     Spain  1351130 2020-05-05
6                    Turkey  1204421 2020-05-05
7                     India  1191946 2020-05-05
8            United Kingdom  1015138 2020-05-05
9                    Canada   940567 2020-05-05
10                   France   724574 2020-05-05
> 
```

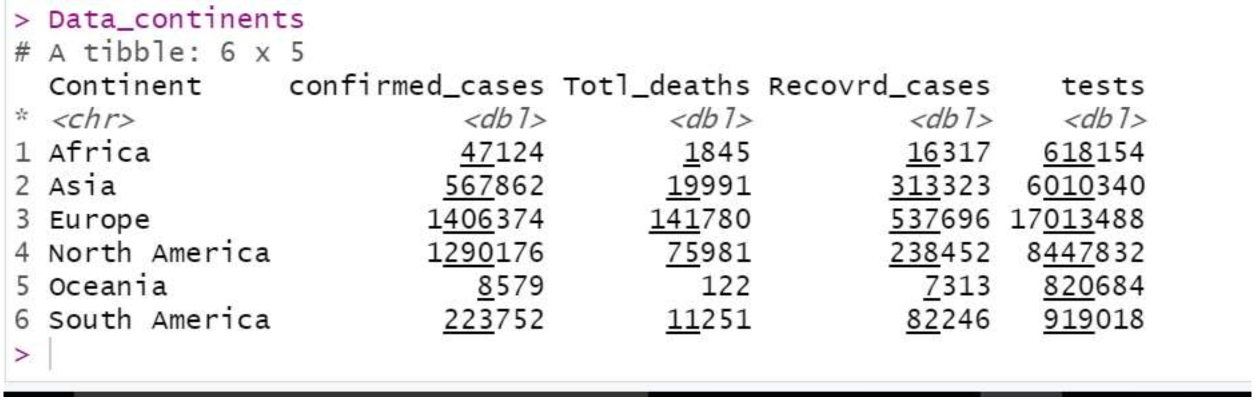
Top ten countries with highest number of tests per million population as May 05th 2020 are as shown here.



```
> top10tests1mp <- as.data.frame(top10_tests1mP[1:10, 1:3]
> top10tests1mp
      Country Tests_1M_Pop       Date
1     Iceland    145101.17 2020-05-05
2     Bahrain     99080.63 2020-05-05
3  Luxembourg     81120.17 2020-05-05
4   Lithuania     53451.96 2020-05-05
5      Israel     46547.31 2020-05-05
6    Portugal     44598.68 2020-05-05
7     Denmark     44457.16 2020-05-05
8     Ireland     44248.63 2020-05-05
9     Estonia     43473.16 2020-05-05
10      Qatar     39458.93 2020-05-05
> 
```

8. Based on the last day data, print the up to date confirmed, death, recovered cases as well as the tests for every continent.

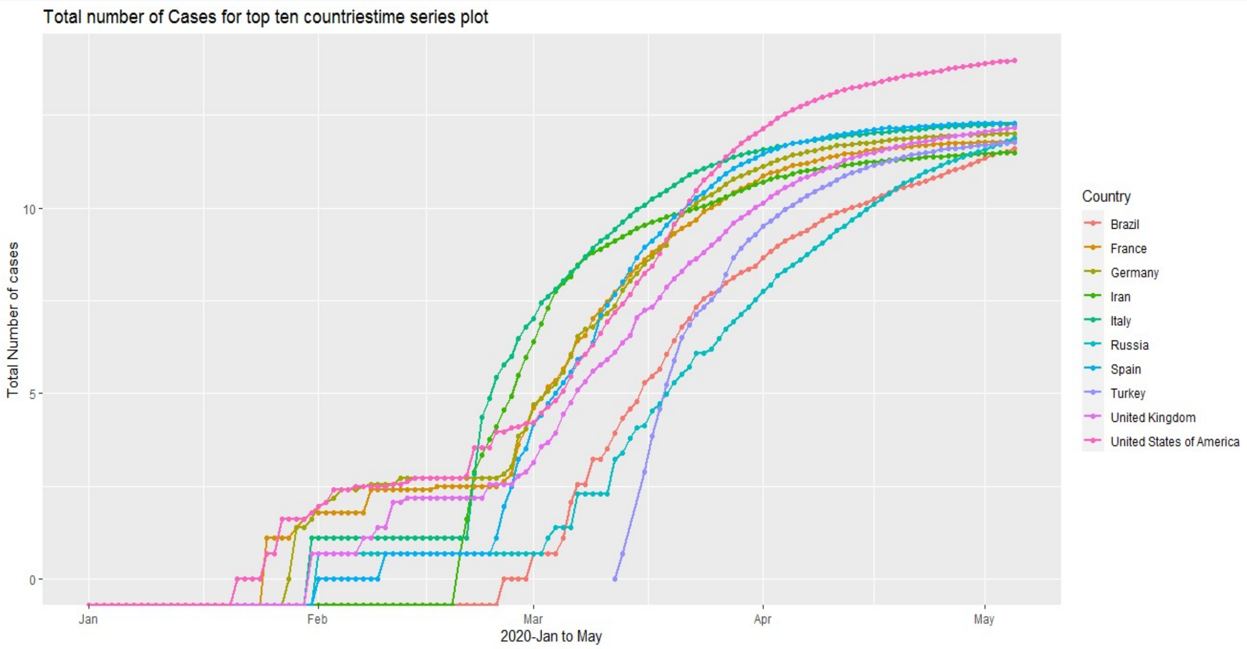
Answer:



```
> Data_continents
# A tibble: 6 x 5
  Continent     confirmed_cases Totl_deaths Recovrd_cases    tests
* <chr>                   <dbl>       <dbl>         <dbl>    <dbl>
1 Africa                  47124        1845         16317   618154
2 Asia                   567862       19991        313323  6010340
3 Europe                1406374      141780        537696 17013488
4 North America         1290176       75981        238452  8447832
5 Oceania                  8579         122          7313   820684
6 South America          223752       11251         82246   919018
> 
```

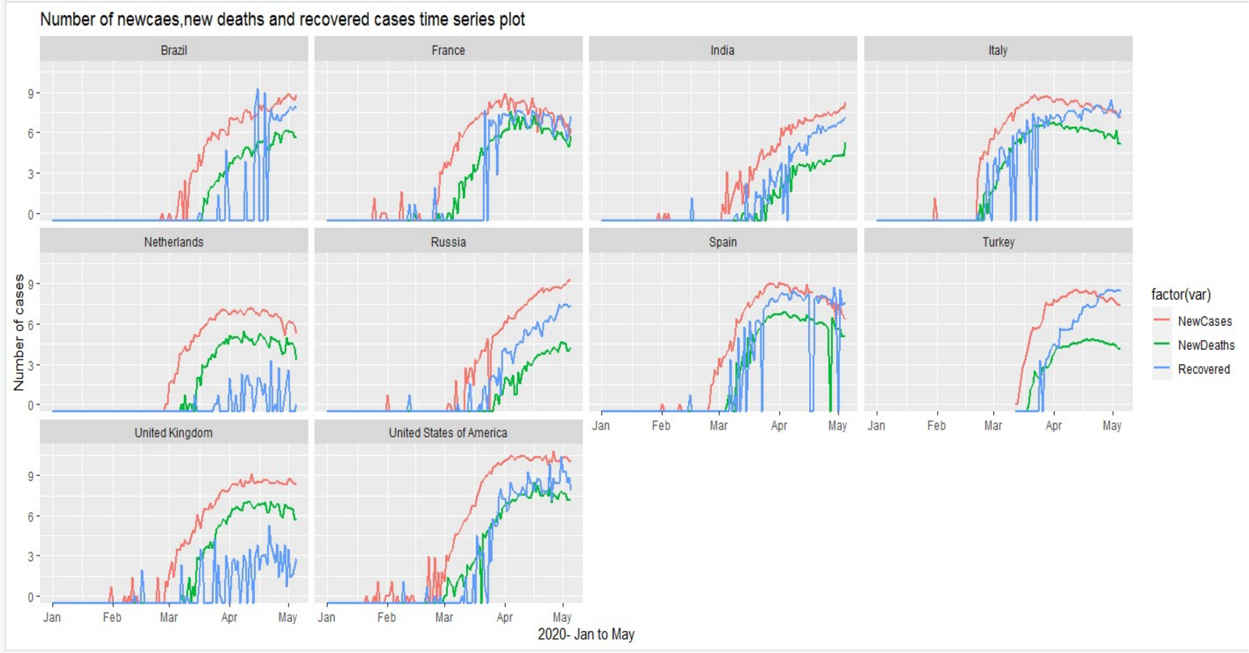
9. Build a graph to show the total number of cases over the time for the top 10 countries that have been obtained in question 7 (Use log for Y axis for better presentation). [Hint: first you need to get the data of the top-10 countries and then plot their lines]

Answer:



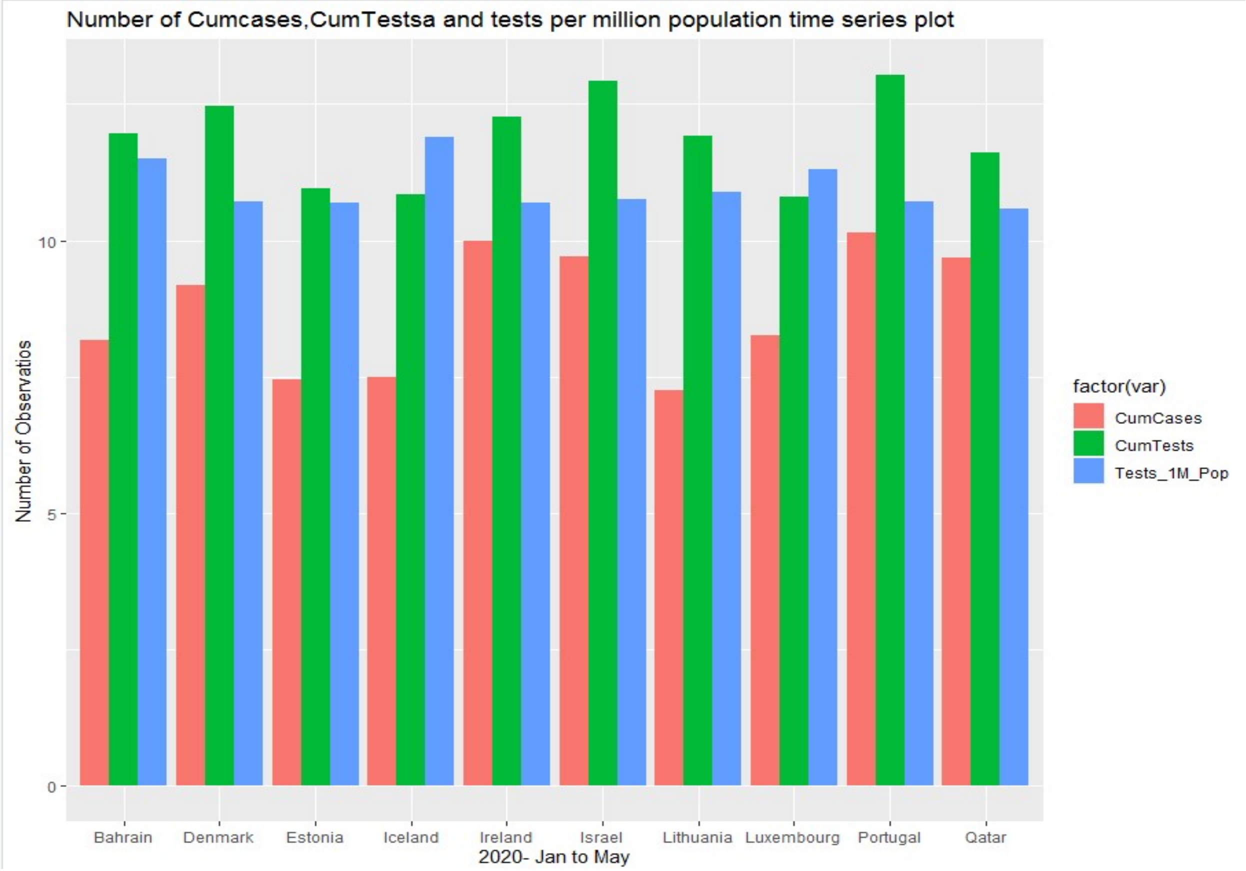
10. Build a graph for the top 10 countries with current highest active cases which was obtained previously in question 7. The graph should have one subgraph (i.e. using facet function) for each of these countries, every subgraph should show how the new cases, new deaths, and new recovered cases were changing over time (Use log for Y axis for better presentation, Use different colour to distinguish between new cases, deaths, and recovered). [hint: geom_line function with date on x_axis and each of the values of the variables in y_axis]

Answer:

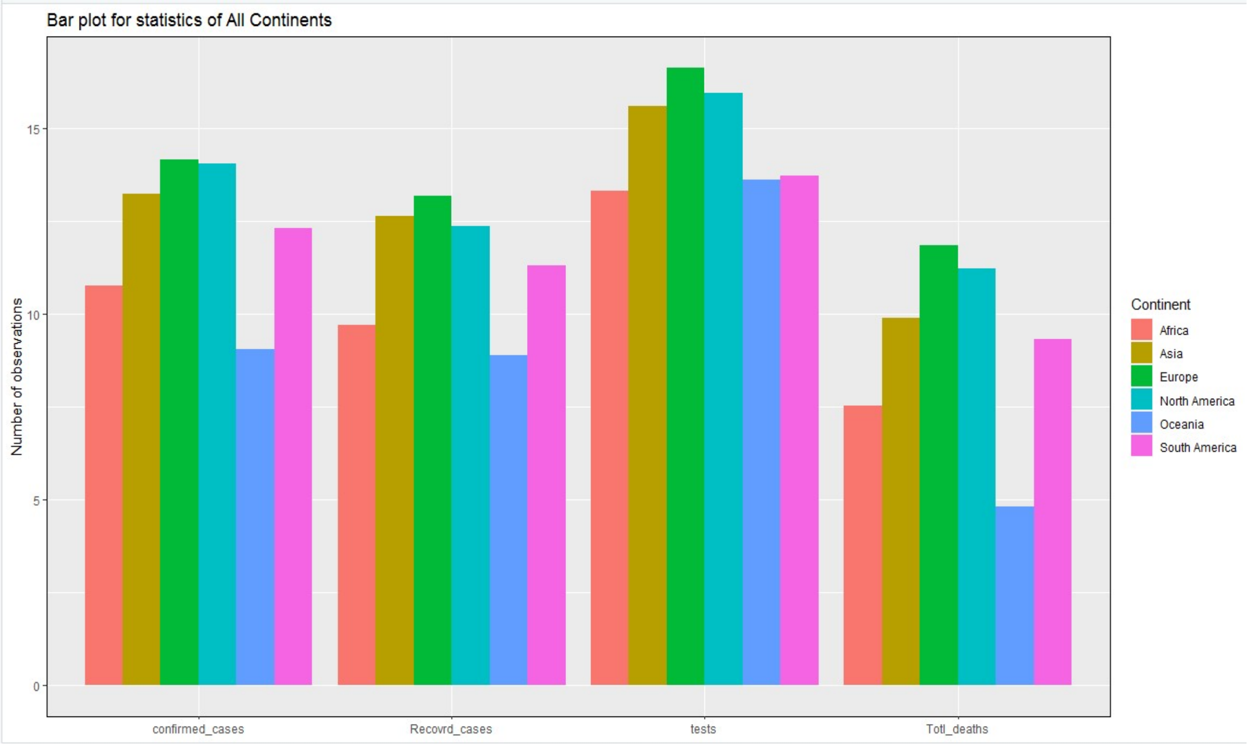


11. Build a graph for the top 10 countries with current highest total tests per one million of the population which was obtained previously in question 7. This graph should present total number of infected cases, total tests so far, and the total tests per million of the population for each country. [hint: you can use bar chart to achieve this task]

Answer:



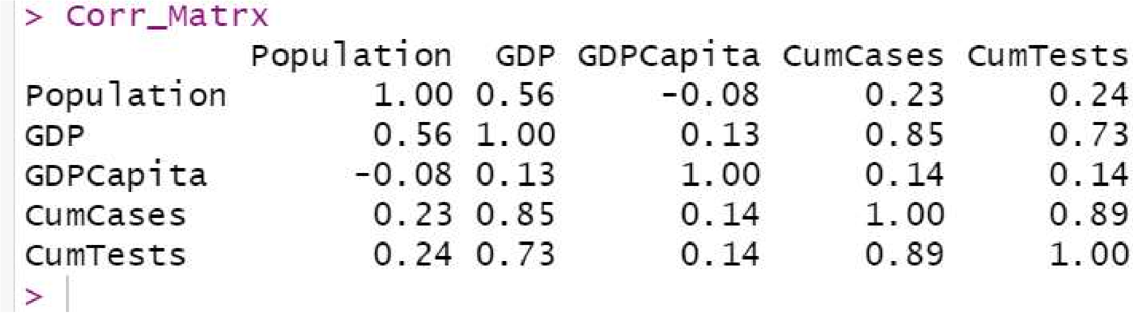
12. Build a graph to present the statistics of all continents which was obtained previously in question 8 (Use log for Y axis for better presentation, Use Continent in the legend, make sure x-axis labels does not overlap).



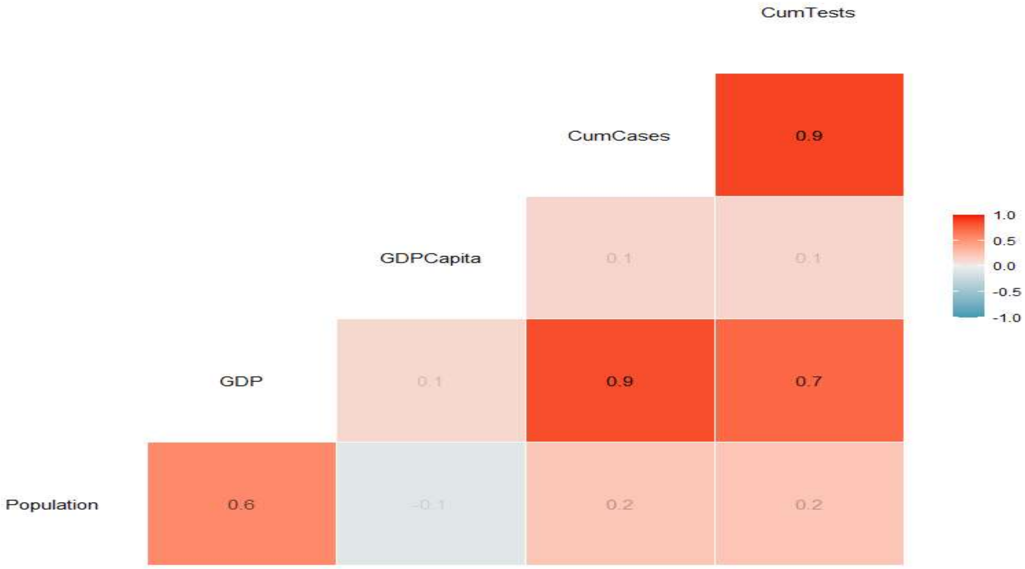
##

Task3:

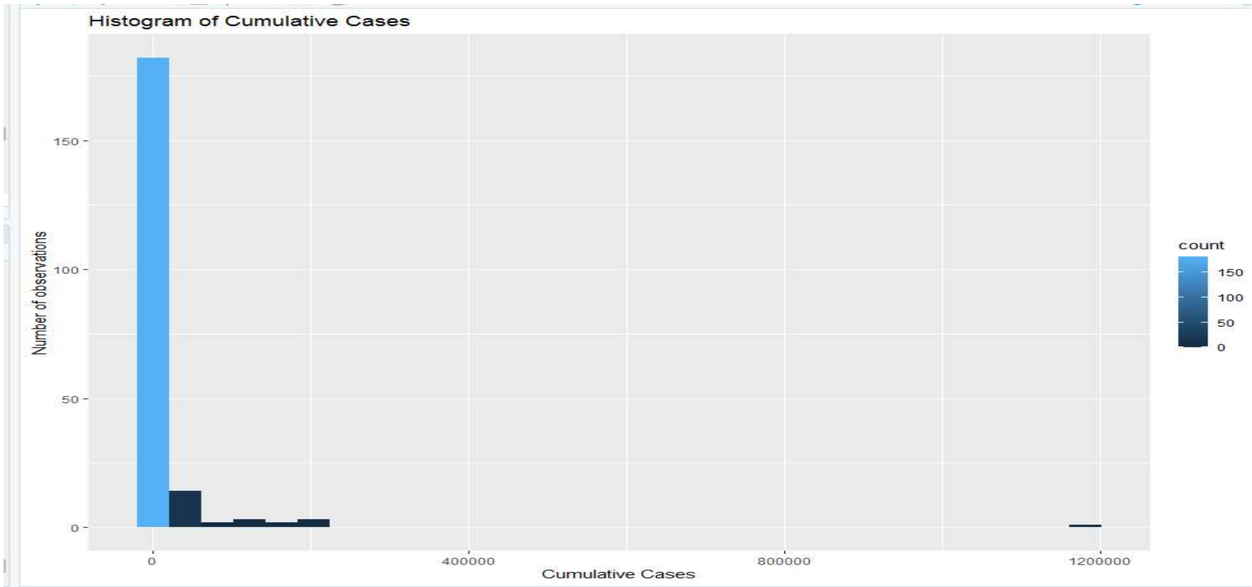
2. Compute the correlation matrix between the variables of the "cor_data" and visualise this correlation matrix.

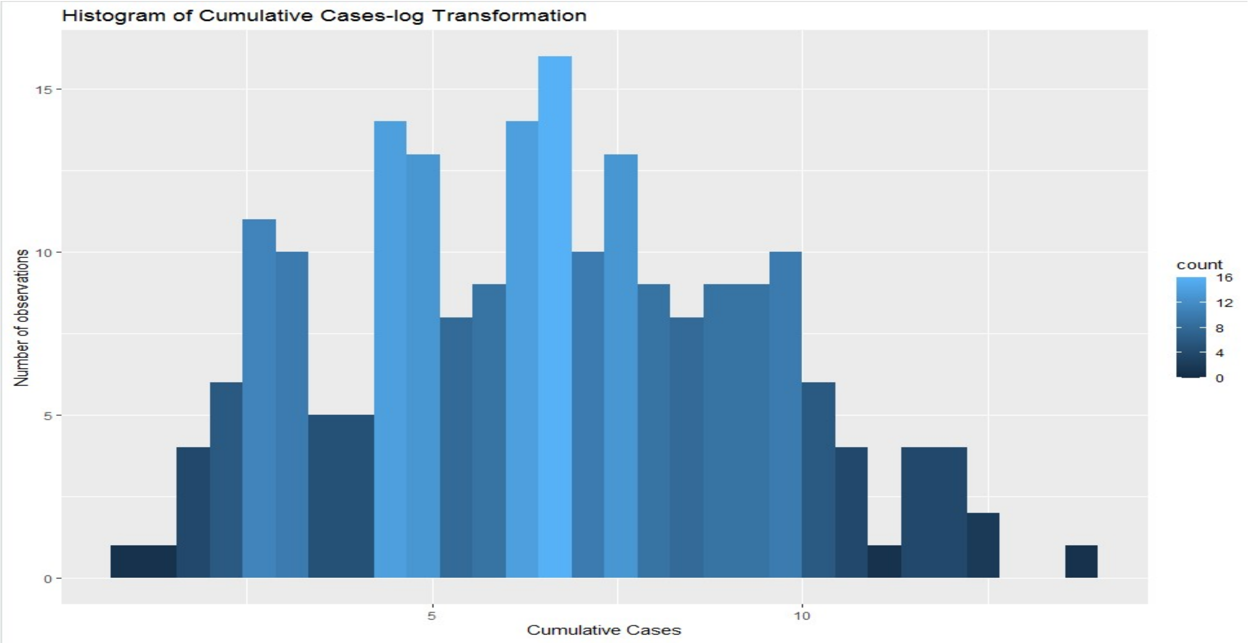


```
> Corr_Matrx
           Population  GDP GDPCapita CumCases CumTests
Population       1.00 0.56     -0.08     0.23     0.24
GDP              0.56 1.00      0.13     0.85     0.73
GDPCapita       -0.08 0.13      1.00     0.14     0.14
CumCases         0.23 0.85      0.14     1.00     0.89
CumTests         0.24 0.73      0.14     0.89     1.00
> 
```

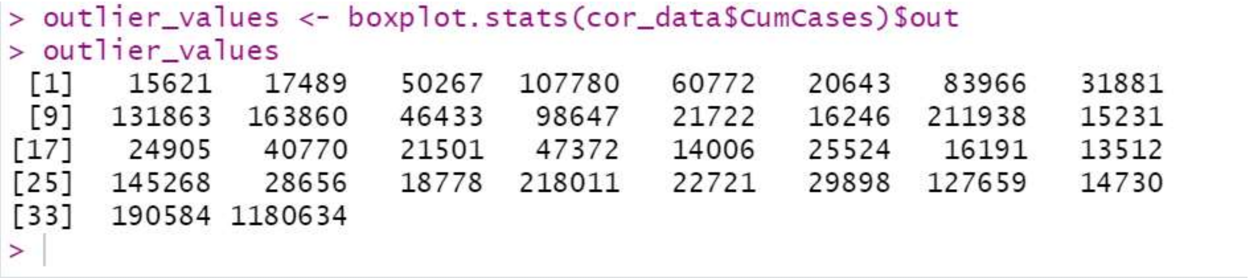


3. visualise the distribution of the cumulative cases in the cor_data with and without changing the scale of the x axis to log transformation. [Hint: you can use the geom_histrogram function]



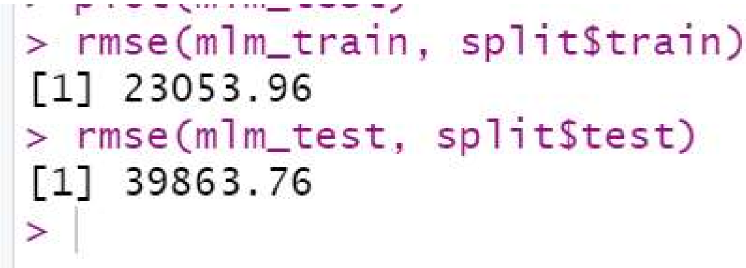


4. **Print the outlier values of the cumulative cases in “cor_data”.**



```
> outlier_values <- boxplot.stats(cor_data$CumCases)$out
> outlier_values
 [1]   15621   17489   50267  107780   60772   20643   83966   31881
 [9]  131863  163860   46433   98647   21722   16246  211938   15231
[17]   24905   40770   21501   47372   14006   25524   16191   13512
[25]  145268   28656   18778  218011   22721   29898  127659   14730
[33]  190584 1180634
> 
```

6. **Train a linear regression model to predict cumulative cases from the GDP of the countries. Then, evaluate this model on the test data and print the root mean square error value.**

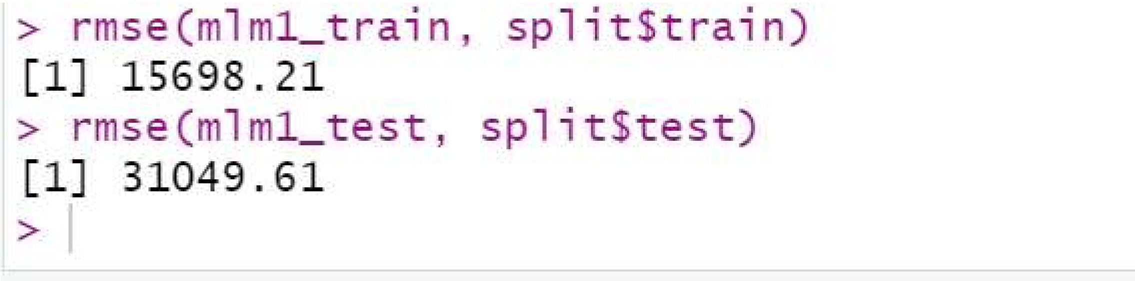


```
> rmse(mlm_train, split$train)
[1] 23053.96
> rmse(mlm_test, split$test)
[1] 39863.76
> 
```

7. Train another linear regression model to predict cumulative cases from all the other variables. Then, evaluate this model on the test data and print the root mean square error value.

Answer:

As the RMSE value lesser when considering all the explanatory variable **m1m1_train** model is the best fitted one.



```
> rmse(mlm1_train, split$train)
[1] 15698.21
> rmse(mlm1_test, split$test)
[1] 31049.61
> 
```