Assignment2

Ziya YİĞİT

19 01 2022

### Take a random sample of 1000 observations from this data (you can use sample() function), and save this sample as a new data. Answer the following questions using this sample.

### Q1 : Calculate the five-number summary statistics (minimum-Q1-median-Q3-maximum) of covid-19 daily new cases for each country within each month. (you can use quantile() function to get the quartiles)

| month | location | min | q1 | med | q3 | max |
| --- | --- | --- | --- | --- | --- | --- |
| 2 | Belgium | 0 | 0.0 | 0 | 0.0 | 0 |
| 2 | Croatia | 2 | 2.0 | 2 | 2.0 | 2 |
| 2 | Czechia | NA | NA | NA | NA | NA |
| 2 | Estonia | NA | NA | NA | NA | NA |
| 2 | France | 0 | 0.0 | 0 | 1.0 | 2 |
| 2 | Germany | 0 | 0.5 | 1 | 1.5 | 2 |
| 2 | Iceland | NA | 0.0 | 0 | 0.0 | NA |
| 2 | Italy | 0 | 0.0 | 0 | 0.0 | 0 |
| 2 | Netherlands | 0 | 0.0 | 0 | 0.0 | 0 |
| 2 | North Macedonia | 0 | 0.0 | 0 | 0.0 | 0 |
| 2 | Norway | 1 | 3.0 | 5 | 7.0 | 9 |
| 2 | Portugal | NA | NA | NA | NA | NA |
| 2 | Romania | 0 | 0.0 | 0 | 0.0 | 0 |
| 2 | Russia | 0 | 0.0 | 0 | 0.0 | 0 |
| 2 | Slovenia | NA | NA | NA | NA | NA |

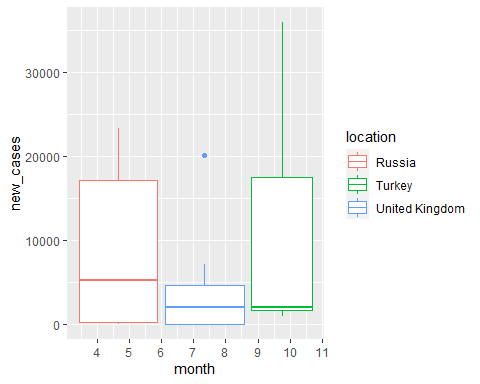
### Q2 : Find the highest daily cases and deaths separately for each country.

| location | max\_case | max\_death |
| --- | --- | --- |
| Italy | 37249 | 889 |
| Turkey | 36008 | 196 |
| Germany | 33825 | 910 |
| France | 29837 | 1231 |
| Russia | 28701 | 457 |
| Spain | 25886 | 718 |
| Sweden | 22319 | 291 |
| United Kingdom | 20078 | 433 |
| Czechia | 15731 | 220 |
| Ukraine | 15623 | 254 |
| Poland | 13632 | 544 |
| Romania | 10108 | 177 |
| Netherlands | 10029 | 170 |
| Serbia | 6903 | 57 |
| Portugal | 6640 | 95 |
| Belgium | 6124 | 230 |
| Hungary | 5097 | 99 |
| Austria | 4954 | 132 |
| Switzerland | 4946 | 88 |
| Croatia | 4396 | 66 |
| Lithuania | 3984 | 56 |
| Slovakia | 3707 | 104 |
| Denmark | 3337 | 14 |
| Bulgaria | 3146 | 189 |
| Belarus | 1893 | 9 |
| Slovenia | 1767 | 46 |
| Moldova | 1715 | 26 |
| Greece | 1678 | 98 |
| Bosnia and Herzegovina | 1541 | 81 |
| Luxembourg | 1325 | 17 |
| North Macedonia | 1161 | 35 |
| Ireland | 1025 | 77 |
| Montenegro | 767 | 9 |
| Kosovo | 728 | 17 |
| Estonia | 654 | 6 |
| Albania | 589 | 12 |
| Latvia | 587 | 6 |
| Norway | 518 | 9 |
| Finland | 460 | 43 |
| Cyprus | 424 | 5 |
| Andorra | 246 | 3 |
| Malta | 218 | 3 |
| Iceland | 91 | 2 |
| Liechtenstein | 45 | 0 |
| San Marino | 31 | 3 |
| Monaco | 22 | 0 |
| Vatican | 7 | -Inf |

### Q3 : Identify the month in which the mean daily cases is the highest for each country.

| location | month | mean\_of\_dailycases |
| --- | --- | --- |
| Turkey | 12 | 31896.00 |
| France | 10 | 29837.00 |
| Russia | 12 | 27801.50 |
| Germany | 11 | 27474.00 |
| Italy | 11 | 27207.00 |
| Spain | 11 | 22932.50 |
| United Kingdom | 11 | 20078.00 |
| Czechia | 11 | 15731.00 |
| Poland | 10 | 13632.00 |
| Ukraine | 12 | 13135.00 |
| Netherlands | 10 | 8972.80 |
| Romania | 11 | 8706.00 |
| Sweden | 12 | 7130.80 |
| Serbia | 12 | 6903.00 |
| Belgium | 11 | 6124.00 |
| Portugal | 11 | 5212.50 |
| Switzerland | 11 | 4946.00 |
| Austria | 11 | 4811.50 |
| Hungary | 11 | 4512.50 |
| Croatia | 12 | 3579.50 |
| Lithuania | 12 | 3335.25 |
| Bulgaria | 11 | 3146.00 |
| Denmark | 12 | 2950.00 |
| Slovakia | 10 | 2430.00 |
| Belarus | 12 | 1791.00 |
| Slovenia | 11 | 1735.50 |
| Greece | 11 | 1435.50 |
| Moldova | 11 | 1342.50 |
| Bosnia and Herzegovina | 11 | 1278.00 |
| North Macedonia | 12 | 953.50 |
| Ireland | 10 | 808.00 |
| Montenegro | 11 | 767.00 |
| Kosovo | 11 | 631.00 |
| Luxembourg | 11 | 602.00 |
| Latvia | 12 | 587.00 |
| Albania | 12 | 585.00 |
| Norway | 11 | 518.00 |
| Estonia | 12 | 498.00 |
| Cyprus | 12 | 413.00 |
| Finland | 12 | 382.67 |
| Andorra | 10 | 246.00 |
| Malta | 11 | 151.00 |
| Iceland | 3 | 80.00 |
| Liechtenstein | 11 | 45.00 |
| San Marino | 12 | 28.00 |
| Monaco | 11 | 9.75 |
| Vatican | 10 | 7.00 |

### Q4 : Select 3 country and plot the distribution of daily cases by month. Use location as clusters (i.e., color=location) to show the difference between countries.



### Take a random sample of 100 sentences from this data, then split these sentences into words and take each word as a member of a vector. After removing duplicated words (you can use unique() function to remove duplicated words), save this sample as a new data.

sampfile<- sample(c(sentences),size=100)  
sampfile <- sampfile %>%  
 str\_split(" ") %>%  
 simplify()  
  
sampfile<-str\_replace\_all(sampfile,"\\.","") #deleting ",.?"  
sampfile<-str\_replace\_all(sampfile,"\\,","")  
sampfile<-str\_replace\_all(sampfile,"\\?","")  
sampfile<-str\_to\_lower(sampfile)   
sampfile<-unique(sampfile)

### Q1 : Find words which are starting with “a” and ending with “e”.

Words which starts with a and ends with e are : are

### Q2 : Calculate the number of words which have more than 3 vowels.

There are 2 words which have more than 3 wovels. The words which have more than 3 vowels are : lonesome, sausage

### Q3 : List the five longest word in your data

Longest five words are : background, shortened, pleasant, lonesome, contents

### Q4 : Try to find word(s) which contain any of these words: age, any, day, exp, her, pro, the.

Words which contain age, any, day, exp, her, pro, the are : the, others, her, their, here, many, other, either, they, sausage, any, sunday, passage, gathered, pages, garbage, clothes, brothers