

**전세계 코로나19 추이 분석을 통한**

**글로벌 피해도 및 소외 지역 파악**

**- CSSE 데이터셋을 바탕으로**

2023.03.15

**B1팀**

**김예지, 서영석, 이현빈, 전국림**

분석용 데이터 구축

**<목차>**

1. **서론**
2. 데이터 분석 배경 p. 3
3. 데이터 분석 설명 p. 3
4. **본론**
5. 국가별 코로나 바이러스 확진자 및 사망자 분석 p. 4
6. 코로나 환자 미발생 국가 분석 p.11
7. 코로나 환자 발생 최대 국가 분석 p.12
8. **분석 결과 및 결론**
9. 데이터 분석 결과 및 결론 도출 p.15

**참고 자료**

**1. 서론**

**1) 데이터 분석 배경**

2023년 현재, 지난 2020년부터 세계를 강타한 코로나 19 팬더믹이 위드 코로나(With Covid19) 시대로 완전히 넘어가고 있다. 이로 인해 국경 이동의 자율성과 높은 경제순환율 등이 세계적으로 높아지고 있다. 하지만 여전히 위드 코로나 정책에 대한 불안함을 지닌 의견들도 존재한다. 코로나19는 높은 치사율과 변이성을 보이고 있는 특이체라는 이유가 주된 근거이다.

이러한 불안감을 최소화하기 위해, 각 국가는 지속적인 코로나 바이러스 대응 구축이 필수적이다. 따라서 이 보고서에서는 코로나 바이러스에 대한 장기/지속적인 대응 플랜을 우선적으로 마련해야 하는 국가를 나타내고자 한다. 즉, 이 보고서를 통해 코로나19 바이러스에 영향을 크게 받은 국가를 확인할 수 있다. 또한, 데이터 결과 분석을 통해 코로나 이외의 제 3의 신종 바이러스가 발생할 경우, 능동적인 대응책 수립이 가능하도록 하는, 국가 의료 체계 재확립의 필요성을 직관적으로 인지할 수 있다.

동시에 세계의 코로나 확진자 및 사망자 수를 파악함으로써 의료, 통계적으로 소외된 국가의 존재 여부를 확인하고자 한다. 소외된 국가라는 판단은 '코로나19와 관련된 통계 수치가 집계되지 않아 데이터셋에 존재하지 않거나, 집계 수치가 '0'으로 기록되고 있을 경우'로 한정하여 가정한다.

**2) 데이터 분석 설명**

이 보고서에서는 R을 활용하여 세계 코로나19 확진자 및 사망자 데이터를 분석할 예정이다. 데이터셋은 일관성을 위해 Johns’ Hopkis 대학 내 The Center For Systems Science and Engineering(CSSE) 에서 업데이트 하는 것으로 제한한다.

아울러 보다 정확한 분석 및 수치 계산을 위해 기간을 고정하도록 했다. 따라서 ‘2021년 8월 ~ 2022년 7월’로 기간을 한정하여, 해당 1년 동안의 데이터만 활용한다. 데이터를 분석하는 과정에서 파악된 데이터 자체의 기재 오류는 ‘0’ 으로 변환하여 분석했다. 또한 이 보고서에서는 중심내용과 연관된 코드를 간략화하여 기재했다. 따라서 세부 코드 확인을 위해서는 추가 첨부한 'EUC\_KR용 1번문제' 파일을 활용하면 된다.

**2. 본론**

**1) 국가별 코로나 바이러스 확진자 및 사망자 분석**

세계 코로나 바이러스 추이 파악을 위해, 가장 먼저 국가별 코로나 바이러스 확진자와 코로나 바이러스로 인한 사망자를 분석하고자 한다. 이 보고서의 토대가 되는 CSSE 데이터 셋은 누적 데이터로 구성되어 있다. 따라서 옳은 누적 값이 입력되어 있는 지를 선행해서 파악해야한다. 누적된 값을 기준으로 한 분석값과, 일별 확진자 및 사망자를 일자별로 추출한 분석값을 비교해 보도록 한다.

**1-1. 누적 데이터 기준 분석**

**1-1-1. 전체 국가 데이터 분석**

누적된 데이터 기준으로 먼저 분석하고자 한다. 이 경우, 대상 기간인 2021년 8월~2022년 7월까지의 코로나 확진자 및 사망자를 파악하기 위해서는 2021년 7월 31일 자료와 2022년 7월 31일 자료를 기준으로 분석을 진행하면 된다. 이를 위해 두 자료 파일을 불러온다.

|  |
| --- |
| covid21 <- covidAll2021 %>% filter(date == "07-31-2021.csv")  covid21  covid22 <- covidAll2022 %>% filter(date == "07-31-2022.csv")  covid22 <- subset(covid22, select=c(Country\_Region, Deaths, Confirmed))  covid22 |

covid21은 2021년 7월 31일 자료이며, covid22는 2022년 7월 31일 자료다. 두 자료를 불러온 이후, 아래 코드와 같이, 분석 기간 동안의 전체 사망자 수와 전체 확진자 수를 구한다.

|  |
| --- |
| #21년 데이터에 -값  library(dplyr)  micovid21 <- covid21 %>% summarise(miDeath = -Deaths, miConfirmed = -Confirmed)  #21년 데이터에 -한 값을 컬럼 추가해서 붙이기 & -한 값만 남기기  setcovid21 <- cbind(covid21,micovid21)  setcovid21 <- subset(setcovid21, select=c(Country\_Region, miDeath, miConfirmed))  names(setcovid21) <- c("Country\_Region","Deaths","Confirmed")  # 22년 데이터 - 21년 데이터  library(stringr)  total <- rbind(covid22,setcovid21)  total<-aggregate(total[,c('Deaths','Confirmed')], by=list(total$Country\_Region), FUN=sum) |

해당 코드로 정제한 누적 데이터를 활용해, 원하는 값을 분석한다. '글로벌 기준 사망자, 확진자, 일평균 사망자, 일평균 확진자'의 수는 아래 사진과 같이 확인할 수 있다.

|  |
| --- |
|  |
| global <- total %>% summarise(allDeath = sum(Deaths), allConfirmed = sum(Confirmed),  dayDeath = sum(Deaths)/365, dayConfirmed = sum(Confirmed)/365) |

글로벌 기준 총 사망자, 총 확진자, 일평균 사망자, 일평균 확진자에 대한 값은 각각 'allDeath', 'allConfirmed', 'dayDeath', 'dayConfirmed'라는 칼럼으로 나타나고 있다.

**1-1-2. 개별 국가 데이터 분석**

누적 데이터로 분석한 '개별 국가 기준 사망자, 확진자, 일평균 사망자, 일평균 확진자'의 수는 total 변수를 활용하였고, 해당 값들을 일자 수(365)로 나누는 방법으로 구할 수 있다. 아래 사진과 같이, 개별 국가에 대한 수치 중 13개 국가만 아래와 같이 보고서에 작성했다.

|  |
| --- |
|  |
| each <- total %>% mutate(total, dayDeath = round(Deaths/365,2), dayConfirmed = round(Confirmed/365,2))  names(each) <- c("Country\_Region","Deaths","Confirmed","dayDeath","dayConfirmed") |

**1-2. 일별 데이터 기준 분석**

누적 데이터 값이 정상적으로 기재되어 있는지 확인하고자, 일별 데이터 기준으로 한 재분석을 실시한다. 이미 데이터셋에 누적되어 값이 입력되어 있다. 따라서 for문을 통해 일별 데이터 자료를 순차적으로 불러 값의 차이를 구하는 방식으로 일별 확진자 및 사망자 수를 구했다. 이때, 2021년과 2022년의 데이터로 나누어 불러왔다. 이로써 보다 변수 및 데이터 확인이 용이하게 했다. 이 보고서에서는 2021년 파일을 불러온 코드만 아래에서 볼 수 있다. 그 외 코드는 별첨한 파일로 확인 가능하다.

|  |
| --- |
| #2021 파일 가져오기  for (i in 2:length(src\_file)){  covid2021\_temp <- read.csv(  paste0(src\_dir, "/", src\_file[i]),  sep=",",  header=T,  stringsAsFactors=F)  covid2021\_temp %>% arrange(Country\_Region)  # 새로운 컬럼 date 생성  covid2021\_temp$date <- src\_file[i]  # 2021년 정보를 위의 과정으로 차례차례 담음  covid2021 <- rbind(covid2021, covid2021\_temp)  print(i)  } |

또한, Raw Data의 입력 오류로 마이너스 값(음수 값)이 출력되는 경우를 확인할 수 있었다. 때문에 일별 데이터를 추출하여 생성한 분석 데이터가 보다 정확하다는 결론을 내릴 수 있었다. 따라서 마이너스로 추출되는 값을 모두 '0'으로 정제한, 국가별 일별 데이터를 기준으로 삼아 이후 분석을 진행한다.

|  |
| --- |
| #21년일일 확진자와 사망자 음수값을 0으로 변경  covidAll2021final$dayDeath <- ifelse(covidAll2021final$dayDeath < 0,0,covidAll2021final$dayDeath)  covidAll2021final$dayConfirmed<-  ifelse(covidAll2021final$dayConfirmed<0,0,covidAll2021final$dayConfirmed) |
| #22년일일 확진자와 사망자 음수값을 0으로 변경  covidAll2022final$dayDeath <- ifelse(covidAll2022final$dayDeath < 0,0,covidAll2022final$dayDeath)  covidAll2022final$dayConfirmed<-  ifelse(covidAll2022final$dayConfirmed<0,0,covidAll2022final$dayConfirmed) |

음수 값이 출력되고 있는 국가 및 일자는 아래와 같이 확인할 수 있다. 먼저 2021년 데이터를 바탕으로 일일 확진자와 사망자 음수값 출력되는 부분이 존재하는 지를 확인하는 코드를 구성했다. 코드 내용은 아래와 같다.

|  |
| --- |
| #21년 일일 확진자와 사망자 음수값 리스트 출력  covidAll2021final[covidAll2021final$dayConfirmed<0,]  covidAll2021final[covidAll2021final$dayDeath<0,] |

해당 코드를 실행하였을 경우, 아래와 같은 값을 확인할 수 있다. 사망자는 15개의 데이터가, 확진자는 28개의 데이터가 음수값이 출력이 된다. 여기서는 전체를 추출했다.

|  |
| --- |
|  |
| **2021년 일일 확진자 음수값 리스트** |
|  |
| **2021년 일일 사망자 음수값 리스트** |

2022년 데이터를 바탕으로 일일 확진자와 사망자 음수값 출력되는 부분이 존재하는 지 확인하는 코드를 구성했다. 코드 내용은 아래와 같다.

|  |
| --- |
| #22년 일일 확진자와 사망자 음수값 리스트 출력  covidAll2022final[covidAll2022final$dayConfirmed<0,]  covidAll2022final[covidAll2022final$dayDeath<0,] |

해당 코드를 실행하였을 경우, 아래와 같은 값을 확인할 수 있다. 사망자는 51개의 데이터가, 확진자는 47개의 데이터가 음수값이 출력이 된다. 여기서는 일부만 추출했다.

|  |
| --- |
|  |
| **2022년 일일 확진자 음수값 리스트** |
|  |
| **2022년 일일 사망자 음수값 리스트** |

**1-2-1. 전체 국가 데이터 분석**

음수값에 대한 데이터 정제 후, 일별 데이터로 분석한 '글로벌 기준 사망자, 확진자, 일평균 사망자, 일평균 확진자'의 수는 아래 사진과 같이 확인할 수 있다.

|  |
| --- |
|  |
| totalcovid <- rbind(covidAll2021total,covidAll2022total)  allR <- totalcovid %>% summarise(allDeath = sum(Deaths), allConfirmed = sum(Confirmed), dayDeath = sum(Deaths)/365, dayConfirmed = sum(Confirmed)/365) |

글로벌 기준 총 사망자, 총 확진자, 일평균 사망자, 일평균 확진자에 대한 값은 각각 'allDeath', 'allConfirmed', 'dayDeath', 'dayConfirmed'라는 칼럼으로 나타나고 있다.

**1-2-2. 개별 국가 데이터 분석**

일별 데이터로 분석한 '개별 국가 기준 사망자, 확진자, 일평균 사망자, 일평균 확진자'의 수는 아래 사진과 같이 확인할 수 있다. 여기서는 개별 국가에 대한 수치 중 13개 국가만 추출하였다.

|  |
| --- |
|  |
| covidtotal<-aggregate(totalcovid[,c('Deaths','Confirmed')], by=list(totalcovid$Country\_Region), FUN=sum)  names(covidtotal) <- c("Country\_Region","Deaths","Confirmed")  covidtotal$dayDeath <- covidtotal$Deaths/365  covidtotal$dayConfirmed <- covidtotal$Confirmed/365  covidtotal |

6~8페이지에서 음수값 출력이 발생한 것을 확인할 수 있었다. 즉 누적 데이터 기준과 일별 데이터 기준의 결과값이 차이가 존재하는 것으로 판단된다. 실제 데이터 값을 비교했을 때, 총 사망자 수는 40,244명, 총 확진자 수는 2,816,447명, 일 평균 사망자 수는 약 110명, 일 평균 확진자 수는 7,717명 차이가 난다.

따라서 이후에 이어지는 분석에서는 일별 데이터를 기준으로 정제 및 통계 분석을 진행할 예정이다. 또한, 파트별 높은 수치를 기록한 국가 확인에 대한 상세 분석은 '본론3) 코로나 환자 발생 최대 국가 분석'에서 확인할 수 있다.

**2) 코로나 환자 미발생 국가 분석**

데이터셋에서 코로나 환자가 발생하지 않은 국가를 파악하면, 소외된 국가 존재 여부를 확인할 수 있다. 코로나 환자의 수가 지속적으로 0이 나오는 경우가 이에 해당한다. 코로나 환자에 대한 해당 국가의 집계가 정확하게 이루어지지 않고 있기 때문이다.

또한 코로나 환자의 수가 0인 국가를 파악하기에 앞서, 집계가 전혀 되지 않고 있는 국가도 존재할 수 있다. 따라서 전처리 과정을 진행하여, 집계가 되지 않아 데이터가 NA인 국가를 선행하여 확인했다. 해당 코드와 결과는 아래와 동일하다.

|  |  |
| --- | --- |
|  | table(is.na(covid2021$Confirmed))  table(is.na(covid2022$Confirmed)) |

is.na 함수를 활용하여 NA값을 추출했을 때, 결과창(Consol)에서 해당 데이터는 존재하지 않음을 확인할 수 있다. 데이터가 존재하지 않은 국가를 먼저 분류하였으므로, 데이터가 0으로 입력된 국가를 파악했다. 코드와 실행 결과는 아래와 같다.

|  |  |
| --- | --- |
|  | **#사망자가 0인 경우**  deathzero<-covidtotal%>% filter(Deaths == 0) |
|  | **#확진자가 0인 경우**  confirmzero<-covidtotal%>% filter(Confirmed == 0) |

결과값을 통해 Antarctica, Diamond Princess를 비롯한 총 7개 지역의 코로나 환자 사망자 집계가 0으로 입력되고 있음을 확인할 수 있다. 하지만 이 중 Diamond Princess, Holy See, MS Zaandam, Summer Olympics 2020, Winter Olympics 2022는 지역이 아니다. 이 외의 Antarctica(남극), Marshall Islands(마셸제도)만 지역(국가)이다. Antarctica(남극), Marshall Islands(마셸제도)는 지리적 특성상 집계가 원활히 이루어 지지 않아, 코로나19 환자가 존재하지 않는 것으로 결과가 나타나고 있는 가능성이 크다. 그러나 해당 지역들의 집계 누락 및 소외 지역 취급에 대한 유의 필요성은 여전히 존재한다. 이러한 집계 결과가 지속적으로 누적된다면, 해당 지역의 거주자들에 대한 의료 소외로도 이어질 수 있기 때문이다. 즉, 해당 지역 거주자는 범국가적 의료 지원 등을 받을 수 없는 가능성이 높다.

**3) 코로나 환자 발생 최대 국가 분석**

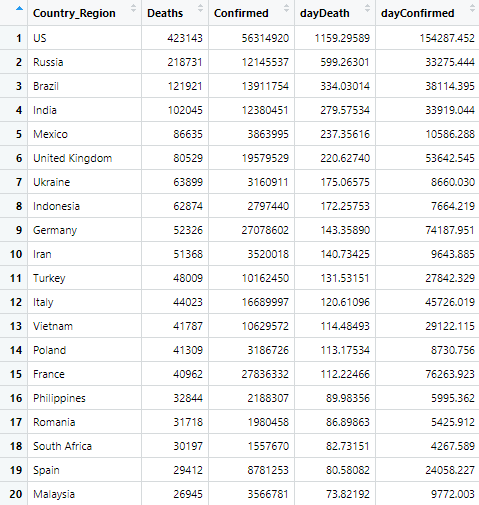
마지막으로 코로나19의 영향을 가장 크게 받은 국가를 파악하고자 한다. 이를 통해 바이러스 영향에 가장 높은 영향을 받는 국가들을 확인할 수 있다. 해당 국가들은 코로나19를 비롯한 타 바이러스에도 영향을 받을 가능성이 높기에, 데이터 분석 결과를 능동적인 대응을 위한 장기적인 플랜 및 기반을 형성하는 근거 자료로 활용할 수 있다.

여기에서는 '총 사망자, 총 확진자, 일평균 사망자, 일평균 확진자'를 확인하여 상위 20개 국가들에 대해 통계를 출력했다. 해당 값들을 확인하기 위한 코드는 아래와 동일하다. 'Confirmdesc' 에는 총 확진자 기준, 'Deathdesc'에는 총 사망자 기준, 'Confirmdaydesc'에는 일평균 확진자 기준, 'Deathdaydesc'에는 일평균 사망자 기준으로 값을 정렬하였다.

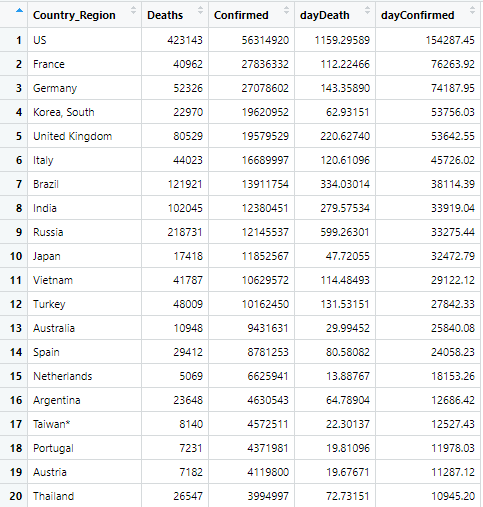
|  |
| --- |
| Confirmdesc <- covidtotal %>% arrange(desc(Confirmed))  Deathdesc <- covidtotal %>% arrange(desc(Deaths))  Confirmdaydesc <- covidtotal %>% arrange(desc(dayConfirmed))  Deathdaydesc <- covidtotal %>% arrange(desc(dayDeath))  head(Confirmdesc,20) %>% View()  head(Deathdesc,20)%>% View()  head(Confirmdaydesc,20)%>% View()  head(Deathdaydesc,20)%>% View() |

위와 같은 코드를 실행했을 때, 출력되는 국가 순위는 차이가 존재한다. 이에 대해서는 아래에 첨부한 파트별 사진 이후에 설명하고자 한다.

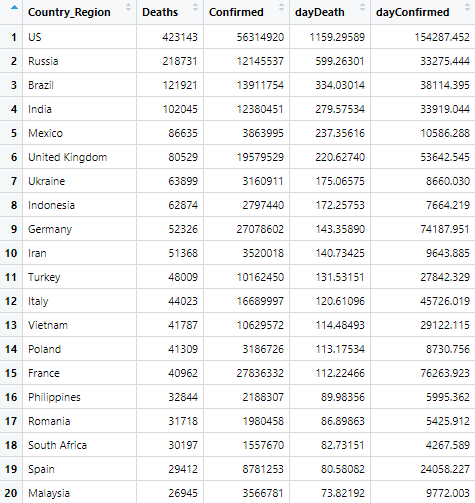
**3-1. 총 사망자 기준 상위 20개 국가**



**3-2. 총 확진자 기준 상위 20개 국가**



**3-3. 일평균 사망자 기준 상위 20개 국가**



**3-4. 일평균 확진자 기준 상위 20개 국가**



네가지 데이터 분석 결과, 공통적으로 미국(US)이 가장 많은 사망자, 확진자 수를 지닌 것으로 확인할 수 있다. 2번 째로 높은 국가는 사망자, 확진자에 따라 차이가 존재하는데, 사망자의 경우 러시아(Russia), 확진자의 경우 프랑스(France)이다. 또한 한국, 일본, 네덜란드를 비롯한 총 9개국은 확진자 상위 20개 국에는 랭크되어 있지만, 사망자 상위 20개 국에는 존재하지 않는다. 즉 확진자 대비 사망자 수가 낮은 것으로 추측 가능하다. 반면 멕시코, 우크라니아, 인도네시아를 비롯한 총 9개국은 사망자 상위 20개 국에는 랭크되어 있지만, 확진자 상위 20개 국에는 존재하지 않는다. 즉 확진자 대비 사망자 수가 높은 것으로 추측할 수 있다.

**3. 분석 결과 및 결론**

**1) 데이터 분석 결과 및 결론 도출**

분석 대상 기간 동안의 전세계 코로나19 확진자 및 사망자 수가 가장 높은 국가는 미국이다. 세계의 핵이라고 할 수 있는 국가이기에, 높은 빈도의 국경 이동 등의 영향이 강하게 존재했을 것으로 추측된다. 또한 확진자 상위 국가와 사망자 상위 국가가 약 50% 차이가 존재하는 것을 알 수 있다. 전체적으로 보았을 때, 국가의 범주는(지리적 위치, 선진화 정도 등) 유사하다. 따라서 개별 국가의 방역 정책에 따른 차이의 결과라고 볼 수 있다.

데이터 분석 결과, Diamond Princess, MS Zaandam은 1년 동안 코로나 확진자 수가 0명, 사망자 수가 0명 인 것으로 확인된다. 그러나 해당 지역은 크루즈 선박으로 해양을 일정 기간동안 순항하므로, 집계가 되지 않는 것으로 판단할 수 있다. 반면 Antarctica(남극), Marshall Islands(마셸제도)는 지속적으로 인구가 거주하고 있는 지역인 동시에, 코로나 환자에 대한 낮은 집계 수치가 기록되고 있다.

이 두 지역은 특수 지역인 만큼, 소수 거주자로 인한 집계 결과라고 치부할 수도 있다. 하지만 전염 및 변이가 빠른 코로나19 바이러스의 특성과, 해당 지역의 지리적 특성상 미확인 확진자 및 미확인 사망자가 존재할 가능성도 존재한다. 즉 Antarctica(남극), Marshall Islands(마셸제도)의 경우, 1) 통계적 사각지대 2) 집계는 되고 있으므로, 확진자 및 사망자를 파악할 수 없는 의료 사각지대 라는 사실 중 최소 하나는 존재한다고 여겨진다. 이는 Antarctica(남극), Marshall Islands(마셸제도)의 거주자들에 대한 인권 문제와도 이어질 수 있다. 또한, 코로나19 이외의 바이러스가 발생하게 된다면, 해당 지역은 또다시 집계에서 누락될 수 있다. 따라서 해당 지역에 대한 특수한 인지 및 지속적인 관리의 필요성을 재고해야 한다.

**참고 자료**

1. **데이터셋: Johns’ Hopkis 대학 내 The Center For Systems Science and Engineering(CSSE) :**

<https://github.com/CSSEGISandData/COVID-19/tree/770dafdb73e9dc31140db77b13b1b92cfd8241f9/csse_covid_19_data/csse_covid_19_daily_reports>