코로나 데이터 분석
(2021.8.1. ~ 2022.7.31)

팀원
이지훈, 이현빈, 전국림

(1) 일별 국가별 코로나 발생자수와 사망자 수를 기준으로 전처리하시오. 일부 국가는 지역별로 코로나 발생자수와 사망자 수가 분리되어 있으니 국가별로 집계하고 국가, 총발생자수, 총사망자수, 일평균 발생자수, 일평균사망자수 리스트를 제시하시오.

```
#국가별 그룹
g_data = data.groupby(['Country/Region']).sum(numeric_only=True)
d_g_data = d_data.groupby(['Country/Region']).sum(numeric_only=True)

#2021년 7월 31일 ~ 2022년 7월 31일 데이터 선택
c_period = g_data.loc[:,'2021-07-31':'2022-07-31']
d_period = d_g_data.loc[:,'2021-07-31':'2022-07-31']

def convert_to_daily(df):
    daily_df = df.diff(axis=1).drop(df.columns[0], axis=1)
    return daily_df

confirmed = convert_to_daily(c_period)
deaths = convert_to_daily(d_period)
```

원데이터에서 국가별로 그룹하고 인덱싱으로 분석 기간의 데이터를 추출하였다. 그 다음에 diff 함수를 사용해 일별 데이터를 구하였다. 일평균 데이터는 소수점 2자리까지 표시하였다.

<알파벳순 10개국 데이터>

	A	B	C	D	E
1		총확진자	일평균확진자	총사망자	일평균사망자
2	Afghanistan	38595	105.74	1040	2.85
3	Albania	179016	490.45	1088	2.98
4	Algeria	96062	263.18	2622	7.18
5	Andorra	30830	84.47	25	0.07
6	Angola	59524	163.08	901	2.47
7	Antarctica	11	0.03	0	0
8	Antigua and	7470	20.47	101	0.28
9	Argentina	4630543	12686.42	23648	64.79
10	Armenia	195255	534.95	4019	11.01

누적 데이터와 일별 데이터 계산 결과 비교

```
(c_period['2022-07-31'] - c_period['2021-07-31'] != result['총확진자']).sum()
0
(d_period['2022-07-31'] - d_period['2021-07-31'] != result['총사망자']).sum()
0
```

c_period, d_period 는 각각 해당 날짜까지의 누적 데이터이며 result는 일별 데이터로 계산한 데이터이다. 다른 값의 개수는 0이므로 실수가 없었음을 확인할 수 있다.

(2) 데이터가 0인 경우(코로나 환자 0)와 데이터가 없는 경우를 구분하여 전처리하고 전처리 시 data가 없는 국가는 제외하고 제외된 국가 리스트를 제시하시오.

```
c_period.isnull().sum().sum()
0
d_period.isnull().sum().sum()
0
```

기간 내의 데이터프레임에서 결측치는 없음을 확인했다.

(3) 1년동안 코로나 총 발생자수, 총 사망자수, 일평균 발생자수, 일평균사망자 수 기준으로 가장 많은 20개 국가를 내림차순으로 정렬(4가지 기준각각 sorting) 하고 각 기준 별 기술통계량을 산출하여 리포트 하시오.

```
#확진자 상위 20
result.sort_values(by = '총확진자', ascending=False).head(20)
#사망자 상위 20
result.sort_values(by = '총사망자', ascending = False).head(20)
```

sort_value 함수와 ascending = False 파라미터를 사용해서 확진자, 사망자를 내림차순으로 정렬했다. 모두 기간이 동일함으로 총 확진자와 일평균 확진자, 총사망자와 일평균 사망자의 순위는 같다.

<확진자 기준 상위 5개국>

	총확진자	일평균확진자	총사망자	일평균사망자
US	56238021	154076.77	420608	1152.35
France	27806603	76182.47	40962	112.22
Germany	27020544	74028.89	52218	143.06
Korea, Sout	19620952	53756.03	22970	62.93
United Kingc	17632481	48308.17	48637	133.25

<사망자 기준 상위 5개국>

	총확진자	일평균확진.	총사망자	일평균사망자
US	56238021	154076.8	420608	1152.35
Russia	12145537	33275.44	218731	599.26
Brazil	13911754	38114.39	121838	333.8
India	12380451	33919.04	102045	279.58
Mexico	3863595	10585.19	86619	237.31

(4) 1년동안 대한민국에서 발생한 코로나 발생자수 및 사망자 수 데이터 대상으로 전처리를 실시하시오.

```
korea_c = confirmed.loc[confirmed.index == 'Korea, South']
korea_d = deaths.loc[deaths.index == 'Korea, South']
korea_c = korea_c.T
korea_d = korea_d.T
korea = pd.DataFrame(columns=['Confirmed', 'Deaths'])
korea['Confirmed'] = korea_c
korea['Deaths'] = korea_d
```

일별 데이터에서 인덱스가 Korea, South인 값만 불러와서 행과 열을 바꾼 후 새로운 데이터프레임에 값을 넣어주었다.

	Confirmed	Deaths
2021-08-01	1215	1
2021-08-02	1201	5
2021-08-03	1723	2
2021-08-04	1776	3
2021-08-05	1704	4
2021-08-06	1822	3
2021-08-07	1728	5
2021-08-08	1492	4
2021-08-09	1539	9
2021-08-10	2219	1
2021-08-11	1986	3
2021-08-12	1990	6
2021-08-13	1929	4
2021-08-14	1817	8
2021-08-15	1553	11

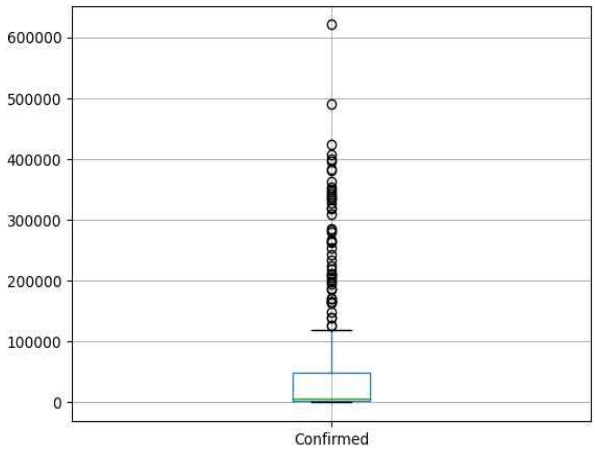
(5) 1년동안 대한민국에서 발생한 코로나 발생자수 및 사망자 수 데이터 대상으로 기술통계량(평균, 중앙값, 최빈값, 표준편차, 분산, 첨도, 왜도, 범위, 최소값, 최대값, 합, 관측수)을 구하시오

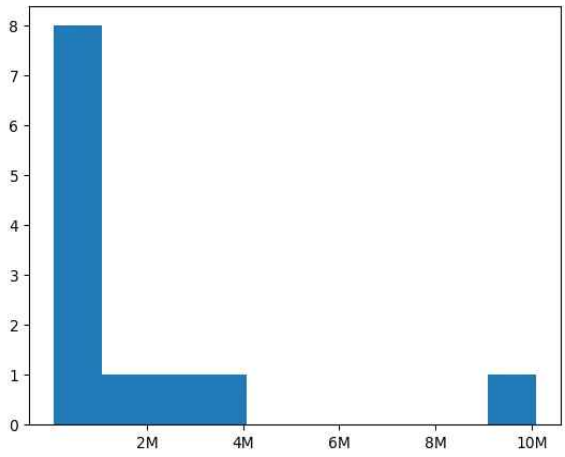
	sample_size	sum	mean	median	mode	std
Confiremd	365	19620952	53756.03	7210	1617, 1803, 35883 (2회)	99022.34
Deahts	365	22970	62.93151	23	6, 7, 9 (13회)	93.13477

I	J	K	L	M
skew	kurtosis	min	max	range
2.544092	6.550822	0	621317	621317
2.269093	4.467725	0	470	470

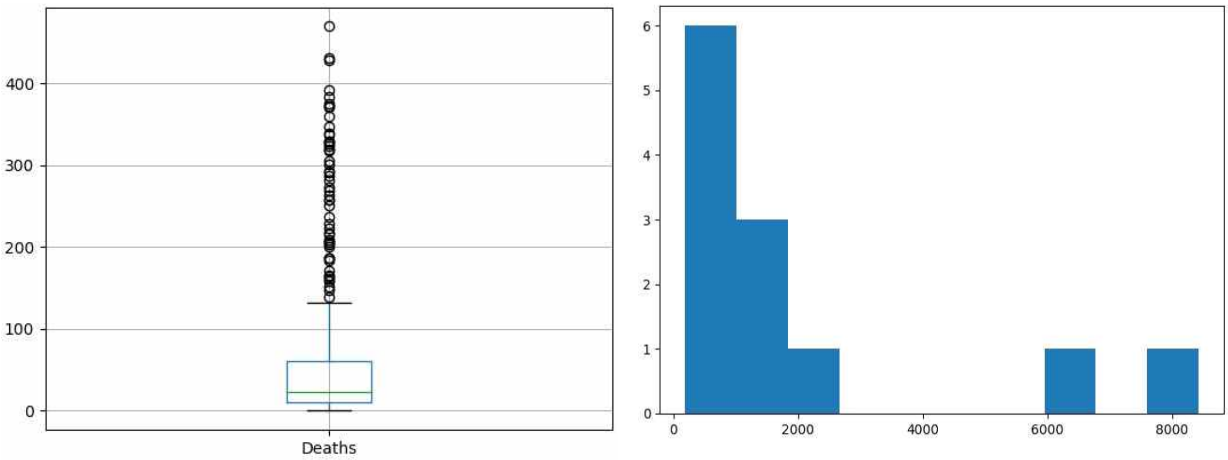
(6) 1년동안 대한민국에서 발생한 코로나 발생자수 및 사망자 수 데이터 대상으로 pandas package 내 함수로 box plot과 월별로 코로나 발생자수 및 사망자수의 히스토그램으로 시각화하시오

<월별 확진자 boxplot과 histogram>





<월별 사망자 boxplot과 histogram>



히스토그램의 가로축은 월별 확진자와 사망자 수를, 세로축은 해당 구간 내 포함된 개월 수를 나타낸다. 예를 들어 월별 사망자 히스토그램에서 가로 8000, 세로 1인 것은 월별 사망자가 8천 명 이상인 달이 1번 있었음을 의미한다. 박스플롯에서는 위쪽 수염이 아래쪽 수염보다 더 길게 나타나며, 히스토그램은 왼쪽 끝이 높고 중간은 비어 있으며 오른쪽 끝은 낮다.

이를 통해 알 수 있는 것은 코로나 환자 수가 점진적으로 증가하는 것이 아니라 급격하게 증가한다는 점이다. 예를 들어, 2022년 1월의 확진자 수는 20만 명쯤이지만 3월 확진자는 천만 명에 달한다는 것을 확인할 수 있다. 이러한 급격한 증가는 이전에는 경험하지 못한 상황이 발생할 수 있다는 것을 시사한다.

이러한 급격한 확진자 증가는 의료인력이 부족하거나 병상이 부족하여 최선의 치료를 받지 못하는 환자들이 많아지는 등 부정적인 영향을 미치기도 한다. 이러한 사실은 다음 펜데믹에 대비하여 필요한 대책을 마련하는 데 중요한 정보가 된다.