코로나 데이터 전처리

- 공공데이터 포털에서 제공하는 "보건복지부 코로나19 시·도 발생 현황" 외부데이터 활용
- 기준년월 기준 전국/제주도 누적확진자수, 월별확진자수 합 추출

In [1]:
```
# 필요 라이브러리 로드
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#출력 그림 크기 조절
sns.set(rc={'figure.figsize':(12,12)})
plt.style.use('ggplot')
from matplotlib import font_manager, rc
font_path = "C:/Windows/Fonts/NGULIM.TTF"
font = font_manager.FontProperties(fname=font_path).get_name()
rc('font', family=font)

import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
#코로나 데이터 로드
df = pd.read_csv('covidencoding.csv', encoding = 'cp949')
df = df.drop("Unnamed: 0",axis=1)
df
```

Out[2]:

	createdt	defcnt	gubun	incdec	qurrate
0	2021-08-31 09:43:52.857	5836	검역	13	0
1	2021-08-31 09:43:52.857	2602	제주	9	385.69
2	2021-08-31 09:43:52.857	9980	경남	56	298.78
3	2021-08-31 09:43:52.857	7100	경북	27	269
4	2021-08-31 09:43:52.857	2600	전남	14	140.42
...	...	...	...	...	...
10375	2020-03-04 19:44:27.27	9	인천	2	NaN
10376	2020-03-04 19:44:27.27	4006	대구	405	NaN
10377	2020-03-04 19:44:27.27	93	부산	3	NaN
10378	2020-03-04 19:44:27.27	99	서울	1	NaN
10379	2020-03-04 19:44:27.27	5328	합계	516	NaN

10380 rows × 5 columns

In [3]:
```
#기간 데이터 월별로 변환
df["y_m"] = df.createdt.apply(lambda x: x[:7])
df
```

Out[3]:

	createdt	defcnt	gubun	incdec	qurrate	y_m

	createdt	defcnt	gubun	incdec	qurrate	y_m
0	2021-08-31 09:43:52.857	5836	검역	13	0	2021-08
1	2021-08-31 09:43:52.857	2602	제주	9	385.69	2021-08
2	2021-08-31 09:43:52.857	9980	경남	56	298.78	2021-08
3	2021-08-31 09:43:52.857	7100	경북	27	269	2021-08
4	2021-08-31 09:43:52.857	2600	전남	14	140.42	2021-08
...	...	...	...	...	...	...
10375	2020-03-04 19:44:27.27	9	인천	2	NaN	2020-03
10376	2020-03-04 19:44:27.27	4006	대구	405	NaN	2020-03
10377	2020-03-04 19:44:27.27	93	부산	3	NaN	2020-03
10378	2020-03-04 19:44:27.27	99	서울	1	NaN	2020-03
10379	2020-03-04 19:44:27.27	5328	합계	516	NaN	2020-03

10380 rows × 6 columns

In [4]:
```
# 제주 지역 데이터만 추출
df_j = df.loc[df.gubun == "제주"]
df_j.drop("createdt",axis=1)
```

Out[4]:

	defcnt	gubun	incdec	qurrate	y_m
1	2602	제주	9	385.69	2021-08
20	2593	제주	12	384.36	2021-08
39	2581	제주	13	382.58	2021-08
58	2568	제주	21	380.65	2021-08
77	2547	제주	23	377.54	2021-08
...	...	...	...	...	...
10290	4	제주	0	NaN	2020-03
10308	4	제주	0	NaN	2020-03
10326	4	제주	0	NaN	2020-03
10344	4	제주	1	NaN	2020-03
10362	3	제주	0	NaN	2020-03

547 rows × 5 columns

In [5]:
```
# 제주지역 월별 누적 확진자 수 추출
for i in df_j["y_m"].unique():
    print(i,"기간 누적 확진자 수")
    r = df_j[df_j["y_m"] == i]["defcnt"].max()
    print(r)
    df_j.loc[df_j["y_m"] == i,"defcnt"] = r
#    df_j["defcnt"] = df_j.apply(lambda x : x[x["y_m"] == i]["defcnt"].max(),axis = 1)
```

```
2021-08 기간 누적 확진자 수
2602
2021-07 기간 누적 확진자 수
1732
2021-06 기간 누적 확진자 수
1264
2021-05 기간 누적 확진자 수
1030
2021-04 기간 누적 확진자 수
707
2021-03 기간 누적 확진자 수
625
2021-02 기간 누적 확진자 수
570
2021-01 기간 누적 확진자 수
522
2020-12 기간 누적 확진자 수
416
2020-11 기간 누적 확진자 수
80
2020-10 기간 누적 확진자 수
59
2020-09 기간 누적 확진자 수
59
2020-08 기간 누적 확진자 수
45
2020-07 기간 누적 확진자 수
26
2020-06 기간 누적 확진자 수
19
2020-05 기간 누적 확진자 수
15
2020-04 기간 누적 확진자 수
13
2020-03 기간 누적 확진자 수
9
```

In [6]:
```
# 제주지역 월별 확진자수 추출 및 제주 지역 코로나 데이터 생성
df_j_g = df_j.groupby(['y_m','gubun','defcnt'])['incdec'].sum()
f_j = df_j_g.reset_index()
f_j = f_j.rename(columns={"defcnt" : "제주_누적확진자","incdec": "제주_월별확진자"})
f_j = f_j.drop("gubun",axis=1)
f_j
```

Out[6]:

	y_m	제주_누적확진자	제주_월별확진자
0	2020-03	9	6
1	2020-04	13	4
2	2020-05	15	2
3	2020-06	19	4
4	2020-07	26	7
5	2020-08	45	19
6	2020-09	59	14
7	2020-10	59	0
8	2020-11	80	21
9	2020-12	416	336

	y_m	제주_누적확진자	제주_월별확진자
10	2021-01	522	106
11	2021-02	570	48
12	2021-03	625	55
13	2021-04	707	82
14	2021-05	1030	323
15	2021-06	1264	234
16	2021-07	1732	468
17	2021-08	2602	870

In [7]:
```
# 전국 데이터 추출
df_a = df.loc[df.gubun == "합계"]
df_a.drop("createdt",axis=1)
df_a
```

Out[7]:

	createdt	defcnt	gubun	incdec	qurrate	y_m
18	2021-08-31 09:43:52.852	251416	합계	1370	485.1	2021-08
37	2021-08-30 09:45:42.502	250046	합계	1485	482.45	2021-08
56	2021-08-29 09:50:47.776	248561	합계	1619	479.59	2021-08
75	2021-08-28 09:54:06.998	246944	합계	1791	476.47	2021-08
94	2021-08-27 09:47:47.644	245153	합계	1838	473.01	2021-08
...	...	...	...	...	...	...
10307	2020-03-08 14:56:02.02	7134	합계	367	NaN	2020-03
10325	2020-03-07 15:29:59.59	6767	합계	483	NaN	2020-03
10343	2020-03-06 15:09:04.04	6284	합계	518	NaN	2020-03
10361	2020-03-05 15:29:39.39	5766	합계	438	NaN	2020-03
10379	2020-03-04 19:44:27.27	5328	합계	516	NaN	2020-03

546 rows × 6 columns

In [8]:
```
# 전국 월별 누적확진자수 추출
for i in df_a["y_m"].unique():
    print(i,"기간 누적 확진자 수")
    r = df_a[df_a["y_m"] == i]["defcnt"].max()
    print(r)
    df_a.loc[df_a["y_m"] == i,"defcnt"] = r
#    df_j["defcnt"] = df_j.apply(lambda x : x[x["y_m"] == i]["defcnt"].max(),axis = 1)
```

```
2021-08 기간 누적 확진자 수
251416
2021-07 기간 누적 확진자 수
198339
2021-06 기간 누적 확진자 수
156961
```

```
2021-05 기간 누적 확진자 수
140338
2021-04 기간 누적 확진자 수
122007
2021-03 기간 누적 확진자 수
103084
2021-02 기간 누적 확진자 수
89674
2021-01 기간 누적 확진자 수
78203
2020-12 기간 누적 확진자 수
60740
2020-11 기간 누적 확진자 수
34201
2020-10 기간 누적 확진자 수
26511
2020-09 기간 누적 확진자 수
23812
2020-08 기간 누적 확진자 수
19947
2020-07 기간 누적 확진자 수
14305
2020-06 기간 누적 확진자 수
12800
2020-05 기간 누적 확진자 수
11468
2020-04 기간 누적 확진자 수
10765
2020-03 기간 누적 확진자 수
9786
```

In [9]:
```
# 전국 월별 확진자수 추출 및 전국 코로나 데이터 생성
df_a_g = df_a.groupby(['y_m','gubun','defcnt'])['incdec'].sum()
f_a = df_a_g.reset_index()
f_a = f_a.rename(columns={"defcnt" : "전국_누적확진자","incdec": "전국_월별확진자"})
f_a = f_a.drop("gubun",axis=1)
f_a
```

Out[9]:

	y_m	전국_누적확진자	전국_월별확진자
0	2020-03	9786	4974
1	2020-04	10765	1018
2	2020-05	11468	703
3	2020-06	12800	1334
4	2020-07	14305	1506
5	2020-08	19947	5642
6	2020-09	23812	3865
7	2020-10	26511	2616
8	2020-11	34201	7689
9	2020-12	60740	26541
10	2021-01	78203	17492
11	2021-02	89674	11468
12	2021-03	103084	13400

	y_m	전국_누적확진자	전국_월별확진자
13	2021-04	122007	18921
14	2021-05	140338	18307
15	2021-06	156961	16603
16	2021-07	198339	41382
17	2021-08	251416	53077

In [10]:
```
# 데이터 병합
new_df = pd.merge(f_a,f_j,how = "left",on = "y_m")
new_df
```

Out[10]:

	y_m	전국_누적확진자	전국_월별확진자	제주_누적확진자	제주_월별확진자
0	2020-03	9786	4974	9	6
1	2020-04	10765	1018	13	4
2	2020-05	11468	703	15	2
3	2020-06	12800	1334	19	4
4	2020-07	14305	1506	26	7
5	2020-08	19947	5642	45	19
6	2020-09	23812	3865	59	14
7	2020-10	26511	2616	59	0
8	2020-11	34201	7689	80	21
9	2020-12	60740	26541	416	336
10	2021-01	78203	17492	522	106
11	2021-02	89674	11468	570	48
12	2021-03	103084	13400	625	55
13	2021-04	122007	18921	707	82
14	2021-05	140338	18307	1030	323
15	2021-06	156961	16603	1264	234
16	2021-07	198339	41382	1732	468
17	2021-08	251416	53077	2602	870

In [11]:
```
new_df.to_csv("코로나데이터_전처리.csv",index = False, encoding = "cp949")
```