raw process

June 6, 2020

[1]:
```
import pandas as pd
```

[43]:
```
covid=pd.read_csv('data/time_series_covid19_confirmed_US.csv')
```

[44]:
```
covid.head(5)
```

[44]:
```
   UID iso2 iso3  code3  FIPS Admin2            Province_State Country_Region  \
0   16   AS  ASM     16  60.0    NaN            American Samoa             US
1  316   GU  GUM    316  66.0    NaN                      Guam             US
2  580   MP  MNP    580  69.0    NaN  Northern Mariana Islands             US
3  630   PR  PRI    630  72.0    NaN               Puerto Rico             US
4  850   VI  VIR    850  78.0    NaN            Virgin Islands             US

       Lat     Long_  ... 5/26/20  5/27/20  5/28/20  5/29/20  5/30/20  \
0 -14.2710 -170.1320  ...       0        0        0        0        0
1  13.4443  144.7937  ...     167      170      171      172      172
2  15.0979  145.6739  ...      22       22       22       22       22
3  18.2208  -66.5901  ...    3324     3397     3486     3647     3718
4  18.3358  -64.8963  ...      69       69       69       69       69

   5/31/20  6/1/20  6/2/20  6/3/20  6/4/20
0        0       0       0       0       0
1      172     175     175     177     179
2       22      22      23      24      26
3     3776    3873    3935    4023    4508
4       69      70      70      70      71

[5 rows x 146 columns]
```

[45]:
```
covid.shape
```

[45]:
```
(3261, 146)
```

This looks very complicated, thus we need some cleaning here.

[46]:
```
covid['Combined_Key'].nunique()
```

[46]:
```
3261
```

We find that the location is defined by combined_key, which is a meaningful primary key of the table, and thus we can make more visuable split

[47]:
```
# time series table:
time_series=covid[covid.columns[10:]].copy()
time_series.head()
```

[47]:
```
                   Combined_Key  1/22/20  1/23/20  1/24/20  1/25/20  1/26/20  \
0            American Samoa, US        0        0        0        0        0
1                     Guam, US        0        0        0        0        0
2  Northern Mariana Islands, US        0        0        0        0        0
3               Puerto Rico, US        0        0        0        0        0
4            Virgin Islands, US        0        0        0        0        0

   1/27/20  1/28/20  1/29/20  1/30/20  ...  5/26/20  5/27/20  5/28/20  \
0        0        0        0        0  ...        0        0        0
1        0        0        0        0  ...      167      170      171
2        0        0        0        0  ...       22       22       22
3        0        0        0        0  ...     3324     3397     3486
4        0        0        0        0  ...       69       69       69

   5/29/20  5/30/20  5/31/20  6/1/20  6/2/20  6/3/20  6/4/20
0        0        0        0       0       0       0       0
1      172      172      172     175     175     177     179
2       22       22       22      22      23      24      26
3     3647     3718     3776    3873    3935    4023    4508
4       69       69       69      70      70      70      71

[5 rows x 136 columns]
```

[48]:
```
# location table:
location=covid[covid.columns[6:11]].copy()
location.head()
```

[48]:
```
             Province_State Country_Region      Lat     Long_  \
0            American Samoa             US -14.2710 -170.1320
1                      Guam             US  13.4443  144.7937
2  Northern Mariana Islands             US  15.0979  145.6739
3               Puerto Rico             US  18.2208  -66.5901
4            Virgin Islands             US  18.3358  -64.8963

                   Combined_Key
0            American Samoa, US
1                     Guam, US
2  Northern Mariana Islands, US
3               Puerto Rico, US
4            Virgin Islands, US
```

[49]:
```
#other information:
other=list(covid.columns[:6])
other.append('Combined_Key')
others=covid[other].copy()
```

```
others.head()
```

[49]:

	UID	iso2	iso3	code3	FIPS	Admin2	Combined_Key
0	16	AS	ASM	16	60.0	NaN	American Samoa, US
1	316	GU	GUM	316	66.0	NaN	Guam, US
2	580	MP	MNP	580	69.0	NaN	Northern Mariana Islands, US
3	630	PR	PRI	630	72.0	NaN	Puerto Rico, US
4	850	VI	VIR	850	78.0	NaN	Virgin Islands, US

[52]:
```
others.to_csv('data/other_information.csv', index=False)
```

[53]:
```
time_series.to_csv('data/time_series_data.csv', index=False)
```

[54]:
```
location.to_csv('data/location.csv', index=False)
```

[ ]:
```

```