DATA WRANGLING

In [1]:
```
import pandas as pd
df=pd.read_csv('C:/Users/USER/Downloads/covid_19_data (1).csv')
```

In [2]:
```
df.head() #prints 1st 5 columns
```

Out[2]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0

In [3]:
```
df.tail() #prints last 5 columns
```

Out[3]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
306424	306425	05/29/2021	Zaporizhia Oblast	Ukraine	2021-05-30 04:20:55	102641.0	2335.0	95289.0
306425	306426	05/29/2021	Zeeland	Netherlands	2021-05-30 04:20:55	29147.0	245.0	0.0
306426	306427	05/29/2021	Zhejiang	Mainland China	2021-05-30 04:20:55	1364.0	1.0	1324.0
306427	306428	05/29/2021	Zhytomyr Oblast	Ukraine	2021-05-30 04:20:55	87550.0	1738.0	83790.0
306428	306429	05/29/2021	Zuid-Holland	Netherlands	2021-05-30 04:20:55	391559.0	4252.0	0.0

In [4]:
```
df.shape  ##shows 306429 rows and 8 columns
```

Out[4]:
```
(306429, 8)
```

In [5]:
```
df.columns
```

Out[5]:
```
Index(['SNo', 'ObservationDate', 'Province/State', 'Country/Region',
       'Last Update', 'Confirmed', 'Deaths', 'Recovered'],
      dtype='object')
```

In [6]:
```
df.info()
```

Out[6]:
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306429 entries, 0 to 306428
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   SNo             306429 non-null  int64  
 1   ObservationDate  306429 non-null  object 
 2   Province/State  228329 non-null  object 
 3   Country/Region  306429 non-null  object 
 4   Last Update     306429 non-null  object 
 5   Confirmed       306429 non-null  float64
 6   Deaths          306429 non-null  float64
 7   Recovered       306429 non-null  float64
dtypes: float64(3), int64(1), object(4)
memory usage: 14.0+ MB
```

"INFO" tells that there are 306429 data which is equal to actual data so there are no null values.Where as in PROVINCE/STATE there are 228329 clouns only that is there are some unfilled data.

In [7]:
```
df.describe()
```

Out[7]:

	SNo	Confirmed	Deaths	Recovered
count	306429.000000	3.064290e+05	306429.000000	3.064290e+05
mean	153215.000000	8.567091e+04	2036.403268	5.042029e+04
std	88458.577156	2.775516e+05	6410.938048	2.015124e+05
min	1.000000	-3.028440e+05	-178.000000	-8.544050e+05
25%	76608.000000	1.042000e+03	13.000000	1.100000e+01
50%	153215.000000	1.037500e+04	192.000000	1.751000e+03
75%	229822.000000	5.075200e+04	1322.000000	2.027000e+04
max	306429.000000	5.863138e+06	112385.000000	6.399531e+06

In [8]:
```
df['Country/Region'].value_counts()
```

Out[8]:
```
Russia                  30251
US                      26740
Japan                   18059
Mainland China          15758
India                   13182
                        ...  
Azerbaijan                  1
North Ireland               1
Republic of Ireland         1
Cape Verde                  1
East Timor                  1
Name: Country/Region, Length: 229, dtype: int64
```

VALUE COUNTS tells about frequency i.e how many times the particular country or region is repeated

In [9]:
```
df.rename(columns={'Deaths':'Death'})
```

Out[9]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Death	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
306424	306425	05/29/2021	Zaporizhia Oblast	Ukraine	2021-05-30 04:20:55	102641.0	2335.0	95289.0
306425	306426	05/29/2021	Zeeland	Netherlands	2021-05-30 04:20:55	29147.0	245.0	0.0
306426	306427	05/29/2021	Zhejiang	Mainland China	2021-05-30 04:20:55	1364.0	1.0	1324.0
306427	306428	05/29/2021	Zhytomyr Oblast	Ukraine	2021-05-30 04:20:55	87550.0	1738.0	83790.0
306428	306429	05/29/2021	Zuid-Holland	Netherlands	2021-05-30 04:20:55	391559.0	4252.0	0.0

306429 rows × 8 columns

In [10]:
```
print(df.dtypes)
```

Out[10]:
```
SNo                 int64
ObservationDate    object
Province/State     object
Country/Region     object
Last Update        object
Confirmed         float64
Deaths            float64
Recovered         float64
dtype: object
```

In [11]:
```
df.loc[:,df.all()] # prints all non zero columns
```

Out[11]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00
...	...	...	...	...	...
306424	306425	05/29/2021	Zaporizhia Oblast	Ukraine	2021-05-30 04:20:55
306425	306426	05/29/2021	Zeeland	Netherlands	2021-05-30 04:20:55
306426	306427	05/29/2021	Zhejiang	Mainland China	2021-05-30 04:20:55
306427	306428	05/29/2021	Zhytomyr Oblast	Ukraine	2021-05-30 04:20:55
306428	306429	05/29/2021	Zuid-Holland	Netherlands	2021-05-30 04:20:55

306429 rows × 5 columns

In [12]:
```
#selecting columns with any one non-zero
df.loc[:,df.any()]
```

Out[12]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
306424	306425	05/29/2021	Zaporizhia Oblast	Ukraine	2021-05-30 04:20:55	102641.0	2335.0	95289.0
306425	306426	05/29/2021	Zeeland	Netherlands	2021-05-30 04:20:55	29147.0	245.0	0.0
306426	306427	05/29/2021	Zhejiang	Mainland China	2021-05-30 04:20:55	1364.0	1.0	1324.0
306427	306428	05/29/2021	Zhytomyr Oblast	Ukraine	2021-05-30 04:20:55	87550.0	1738.0	83790.0
306428	306429	05/29/2021	Zuid-Holland	Netherlands	2021-05-30 04:20:55	391559.0	4252.0	0.0

306429 rows × 8 columns

In [13]:
```
#select columns with null values
df.loc[:, df.isnull().any()]
```

Out[13]:

	Province/State
0	Anhui
1	Beijing
2	Chongqing
3	Fujian
4	Gansu
...	...
306424	Zaporizhia Oblast
306425	Zeeland
306426	Zhejiang
306427	Zhytomyr Oblast
306428	Zuid-Holland

306429 rows × 1 columns

In [14]:
```
#select columns without null values
df.loc[:, df.notnull().any()]
```

Out[14]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
306424	306425	05/29/2021	Zaporizhia Oblast	Ukraine	2021-05-30 04:20:55	102641.0	2335.0	95289.0
306425	306426	05/29/2021	Zeeland	Netherlands	2021-05-30 04:20:55	29147.0	245.0	0.0
306426	306427	05/29/2021	Zhejiang	Mainland China	2021-05-30 04:20:55	1364.0	1.0	1324.0
306427	306428	05/29/2021	Zhytomyr Oblast	Ukraine	2021-05-30 04:20:55	87550.0	1738.0	83790.0
306428	306429	05/29/2021	Zuid-Holland	Netherlands	2021-05-30 04:20:55	391559.0	4252.0	0.0

306429 rows × 8 columns

In [15]:
```
df.dropna()
```

Out[15]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
306424	306425	05/29/2021	Zaporizhia Oblast	Ukraine	2021-05-30 04:20:55	102641.0	2335.0	95289.0
306425	306426	05/29/2021	Zeeland	Netherlands	2021-05-30 04:20:55	29147.0	245.0	0.0
306426	306427	05/29/2021	Zhejiang	Mainland China	2021-05-30 04:20:55	1364.0	1.0	1324.0
306427	306428	05/29/2021	Zhytomyr Oblast	Ukraine	2021-05-30 04:20:55	87550.0	1738.0	83790.0
306428	306429	05/29/2021	Zuid-Holland	Netherlands	2021-05-30 04:20:55	391559.0	4252.0	0.0

228329 rows × 8 columns

In [16]:
```
df.loc[:, df.isnull().any()]
```

Out[16]:

	Province/State
0	Anhui
1	Beijing
2	Chongqing
3	Fujian
4	Gansu
...	...
306424	Zaporizhia Oblast
306425	Zeeland
306426	Zhejiang
306427	Zhytomyr Oblast
306428	Zuid-Holland

306429 rows × 1 columns

In [ ]:
```

```