Table of Contents

small dataset

In [1]:

```
import pandas as pd
```

In [14]:

```
data = pd.read_excel('/Users/Downloads/covid.xlsx')
```

In [15]:

```
data
```

Out[15]:

	Date	State	Region	Confirmed	Deaths	Recovered
0	4/29/2020	NaN	Afghanistan	1939	60	252
1	4/29/2020	NaN	Albania	766	30	455
2	4/29/2020	NaN	Algeria	3848	444	1702
3	4/29/2020	NaN	Andorra	743	42	423
4	4/29/2020	NaN	Angola	27	2	7
...	...	...	...	...	...	...
316	4/29/2020	Wyoming	US	545	7	0
317	4/29/2020	Xinjiang	Mainland China	76	3	73
318	4/29/2020	Yukon	Canada	11	0	0
319	4/29/2020	Yunnan	Mainland China	185	2	181
320	4/29/2020	Zhejiang	Mainland China	1268	1	1263

321 rows × 6 columns

In [ ]:

```
# 1.
# df.count()
# df.isnull().sum()
```

In [16]:

```
data.count()
```

Out[16]:

```
Date         321
State        140
Region       321
Confirmed    321
Deaths       321
Recovered    321
dtype: int64
```

In [18]:

```
data.isnull().sum()
```

Out[18]:

```
Date           0
State        181
Region         0
Confirmed      0
Deaths         0
Recovered      0
dtype: int64
```

In [ ]:

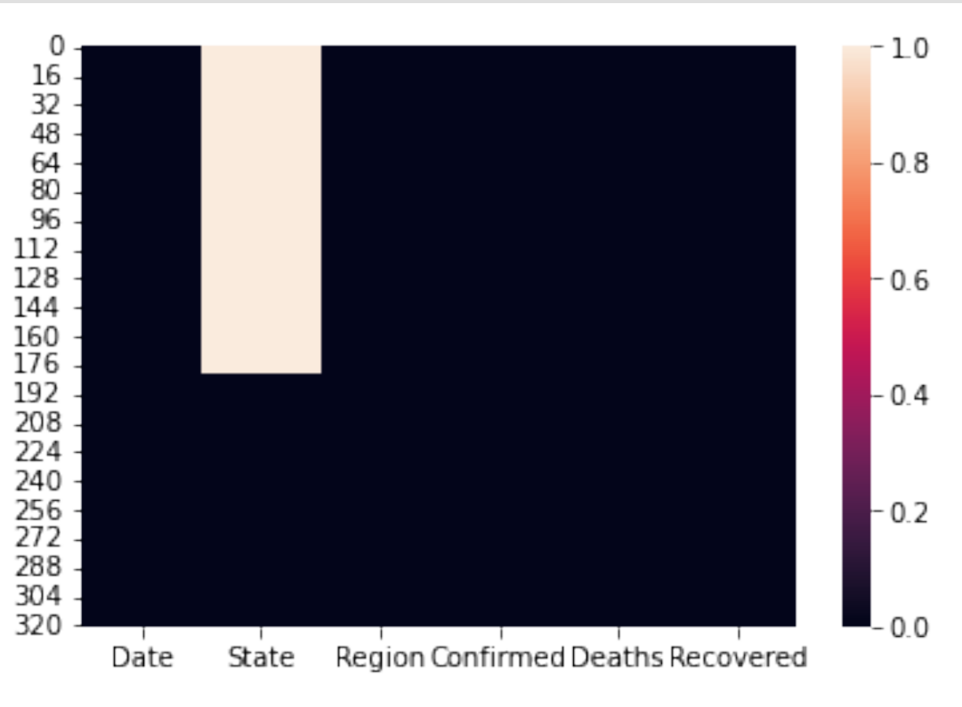
```
# 2.
# import seaborn as sns
# import matplotlib.pyplot as plt
# sns.heatmap(df.isnull())
# plt.show()
```

In [19]:

```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [20]:

```
sns.heatmap(data.isnull())
plt.show()
```



Q1) show the number of confirmed, deaths and recovered cases in each region

In [ ]:

```
# df.groupby('Region').sum().head(50)
# df.groupby('Region').['Confirmed'].sum().sort_values(ascending=False).head(20)
# df.groupby('Region'),[['Confirmed','Recovered']].sum()
```

In [21]:

```
data.head(3)
```

Out[21]:

	Date	State	Region	Confirmed	Deaths	Recovered
0	4/29/2020	NaN	Afghanistan	1939	60	252
1	4/29/2020	NaN	Albania	766	30	455
2	4/29/2020	NaN	Algeria	3848	444	1702

In [30]:

```
data.groupby('Region').sum().head(10)
```

Out[30]:

	Confirmed	Deaths	Recovered
Region			
Afghanistan	1939	60	252
Albania	766	30	455
Algeria	3848	444	1702
Andorra	743	42	423
Angola	27	2	7
Antigua and Barbuda	24	3	11
Argentina	4285	214	1192
Armenia	1932	30	900
Australia	6752	91	5715
Austria	15402	580	12779

In [29]:

```
data.groupby('Region')['Confirmed'].sum().sort_values(ascending=False).head(10)
```

Out[29]:

```
Region
US                1039909
Spain              236899
Italy              203591
France             166543
UK                 166441
Germany            161539
Turkey             117589
Russia              99399
Iran                93657
Mainland China      82862
Name: Confirmed, dtype: int64
```

In [32]:

```
data.groupby('Region')[['Confirmed','Recovered']].sum()
```

Out[32]:

	Confirmed	Recovered
Region		
Afghanistan	1939	252
Albania	766	455
Algeria	3848	1702
Andorra	743	423
Angola	27	7
...	...	...
West Bank and Gaza	344	71
Western Sahara	6	5
Yemen	6	1
Zambia	97	54
Zimbabwe	32	5

187 rows × 2 columns

Q2) remove all the records where confirmed cases is less than 10

In [ ]:

```
# df.Confirmed < 10
# df[df.Confirmed < 10]
# df[~(df[df.Confirmed < 10)]
# df = df[~(df[df.Confirmed < 10)]
```

In [33]:

```
data.head(3)
```

Out[33]:

	Date	State	Region	Confirmed	Deaths	Recovered
0	4/29/2020	NaN	Afghanistan	1939	60	252
1	4/29/2020	NaN	Albania	766	30	455
2	4/29/2020	NaN	Algeria	3848	444	1702

In [38]:

```
data = data[~(data['Confirmed'] < 10)]
```

In [41]:

```
data.Confirmed.min()
```

Out[41]:

```
10
```

Q3) in which region, maximum number of confirmed cases were recorded?

In [ ]:

```
# df.groupby('Region').Confirmed.sum().sort_values(ascending=False).head(20)
```

In [43]:

```
data.groupby('Region')['Confirmed'].sum().sort_values(ascending=False)
```

Out[43]:

```
Region
US             1039909
Spain           236899
Italy           203591
France          166536
UK              166432
                ...
Seychelles          11
Burundi             11
Suriname            10
Gambia              10
Holy See            10
Name: Confirmed, Length: 180, dtype: int64
```

Q4) in which region, minumum number of deaths cases were corded?

In [ ]:

```
# df.groupby('Region').Confirmed.sum().sort_values(ascending=True).head(20)
```

In [50]:

```
data.groupby('Region').Deaths.sum().sort_values(ascending=True).head(30)
```

Out[50]:

```
Region
Cambodia                            0
Seychelles                          0
Saint Lucia                         0
Central African Republic            0
Saint Kitts and Nevis               0
South Sudan                         0
Rwanda                              0
Grenada                             0
Macau                               0
Madagascar                          0
Nepal                               0
Namibia                             0
Saint Vincent and the Grenadines    0
Mozambique                          0
Holy See                            0
Timor-Leste                         0
Mongolia                            0
Uganda                              0
Laos                                0
Eritrea                             0
Vietnam                             0
Fiji                                0
Dominica                            0
Gambia                              1
Equatorial Guinea                   1
Eswatini                            1
Cabo Verde                          1
Maldives                            1
Guinea-Bissau                       1
Liechtenstein                       1
Name: Deaths, dtype: int64
```

Q5) how many confirmed, deaths & recovered cases were reported from India till 29 April 2020?

In [ ]:

```
# df[df.Region == 'Country_name']
```

In [49]:

```
data[data.Region=='India']
```

Out[49]:

	Date	State	Region	Confirmed	Deaths	Recovered
74	4/29/2020	NaN	India	33062	1079	8437

In [52]:

```
data[data.Region=='Canada']
```

Out[52]:

	Date	State	Region	Confirmed	Deaths	Recovered
183	4/29/2020	Alberta	Canada	5165	87	0
193	4/29/2020	British Columbia	Canada	2087	109	0
215	4/29/2020	Grand Princess	Canada	13	0	0
246	4/29/2020	Manitoba	Canada	275	6	0
259	4/29/2020	New Brunswick	Canada	118	0	0
266	4/29/2020	Newfoundland and Labrador	Canada	258	3	0
273	4/29/2020	Nova Scotia	Canada	935	28	0
276	4/29/2020	Ontario	Canada	16978	1153	0
279	4/29/2020	Prince Edward Island	Canada	27	0	0
282	4/29/2020	Quebec	Canada	26610	1762	0
290	4/29/2020	Saskatchewan	Canada	383	6	0
318	4/29/2020	Yukon	Canada	11	0	0

Q6 A) sort the entire data wrt No. of confirmed cases in ascending order

In [53]:

```
# df.sort_values(by = ['Confirmed'], ascending = True)
```

In [54]:

```
data.sort_values(by=['Confirmed'], ascending=True)
```

Out[54]:

	Date	State	Region	Confirmed	Deaths	Recovered
156	4/29/2020	NaN	Suriname	10	1	8
70	4/29/2020	NaN	Holy See	10	0	2
59	4/29/2020	NaN	Gambia	10	1	8
318	4/29/2020	Yukon	Canada	11	0	0
217	4/29/2020	Greenland	Denmark	11	0	11
...	...	...	...	...	...	...
57	4/29/2020	NaN	France	165093	24087	48228
168	4/29/2020	NaN	UK	165221	26097	0
80	4/29/2020	NaN	Italy	203591	27682	71252
153	4/29/2020	NaN	Spain	236899	24275	132929
265	4/29/2020	New York	US	299691	23477	0

304 rows × 6 columns

Q6 B) sort the entire data wrt No. of recovered cases in descending order

In [55]:

```
data.sort_values(by=['Recovered'],ascending=False)
```

Out[55]:

	Date	State	Region	Confirmed	Deaths	Recovered
153	4/29/2020	NaN	Spain	236899	24275	132929
61	4/29/2020	NaN	Germany	161539	6467	120400
76	4/29/2020	NaN	Iran	93657	5957	73791
80	4/29/2020	NaN	Italy	203591	27682	71252
229	4/29/2020	Hubei	Mainland China	68128	4512	63616
...	...	...	...	...	...	...
258	4/29/2020	Nevada	US	4934	230	0
257	4/29/2020	Nebraska	US	3851	56	0
255	4/29/2020	Montana	US	451	16	0
254	4/29/2020	Missouri	US	7660	338	0
274	4/29/2020	Ohio	US	17303	937	0

304 rows × 6 columns

In [ ]: