Task -02 Data Cleaning and EDA

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
data = pd.read_csv('covid.csv')
```

In [4]:
```
data
```

Out[4]:

	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 [5]:
```
data.count()
```

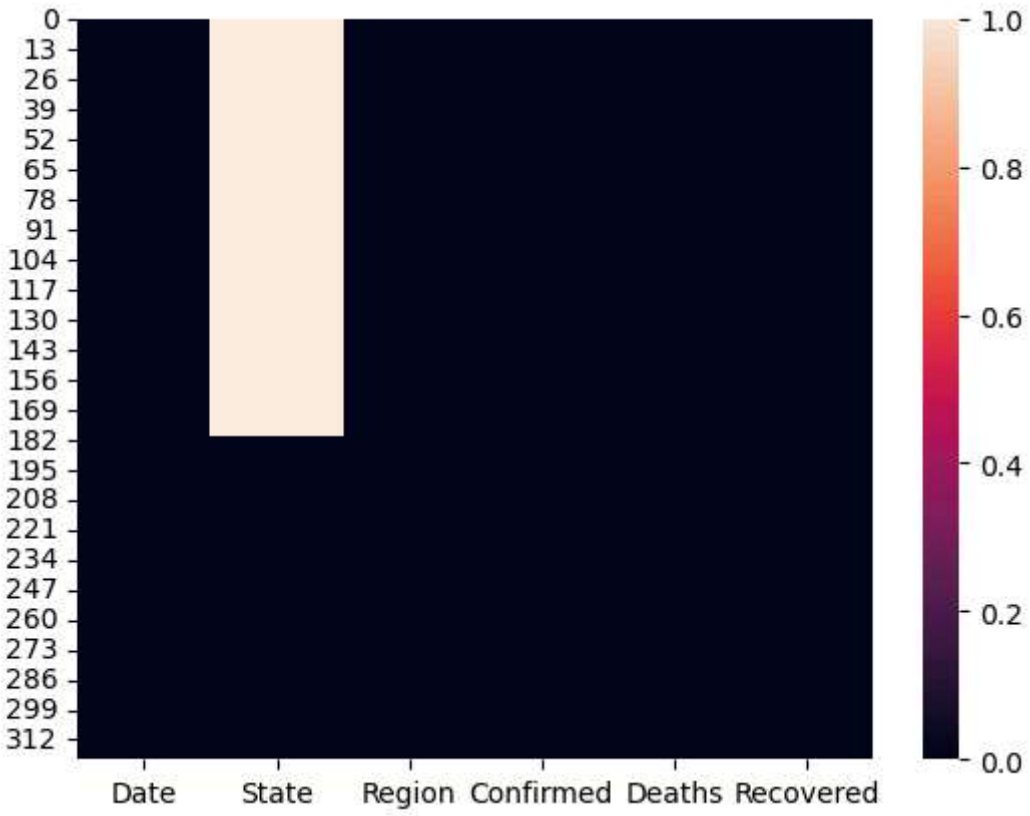
Out[5]:
```
Date         321
State        140
Region       321
Confirmed    321
Deaths       321
Recovered    321
dtype: int64
```

In [6]:
```
data.isnull().sum()
```

Out[6]:
```
Date           0
State        181
Region         0
Confirmed      0
Deaths         0
Recovered      0
dtype: int64
```

In [7]:
```
# We can see this null values results in the form of heatmap

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



Performing EDA on the covid dataset

1. Show the number of confirmed, Deaths and Recovered cases in each Region

In [8]:
```
#df.groupby('Region').sum().head(50)
#df.groupby('Region')['Confirmed'].sum().sort_Values(ascending=False).head(20)
#df.groupby('Region')['Confirmed, 'Recovered].sum()
```

In [9]:
```
data.head(2)
```

Out[9]:

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

In [10]:
```
data.groupby('Region').sum().head(20)
```

```
C:\Users\91958\AppData\Local\Temp\ipykernel_20372\3787432426.py:1: FutureWarning: The
default value of numeric_only in DataFrameGroupBy.sum is deprecated. In a future vers
ion, numeric_only will default to False. Either specify numeric_only or select only c
olumns which should be valid for the function.
  data.groupby('Region').sum().head(20)
```

Out[10]:

Region	Confirmed	Deaths	Recovered
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
Azerbaijan	1766	23	1267
Bahamas	80	11	23
Bahrain	2921	8	1455
Bangladesh	7103	163	150
Barbados	80	7	39
Belarus	13181	84	2072
Belgium	47859	7501	11283
Belize	18	2	9
Benin	64	1	33
Bhutan	7	0	5

In [11]:
```
data.groupby('Region')['Confirmed'].sum().sort_values(ascending = False).head(10)
```

Out[11]:
```
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 [12]:
```
data.groupby('Region')['Confirmed', 'Recovered'].sum()
```

```
C:\Users\91958\AppData\Local\Temp\ipykernel_20372\581960954.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  data.groupby('Region')['Confirmed', 'Recovered'].sum()
```

Out[12]:

Region	Confirmed	Recovered
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

2. Remove all the records where Confirmed Cases is Less Than 10.

In [13]:
```
#df.Confirmed < 10
#df[df.Confirmed < 10]
#df[~(df.Confirmed < 10)]
#df = df{~(df.Confirmed < 10)}
```

In [14]:
```
data.head(2)
```

Out[14]:

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

In [16]:
```
data[data.Confirmed < 10]
```

Out[16]:

	Date	State	Region	Confirmed	Deaths	Recovered
18	4/29/2020	NaN	Bhutan	7	0	5
98	4/29/2020	NaN	MS Zaandam	9	2	0
105	4/29/2020	NaN	Mauritania	8	1	6
126	4/29/2020	NaN	Papua New Guinea	8	0	0
140	4/29/2020	NaN	Sao Tome and Principe	8	0	4
177	4/29/2020	NaN	Western Sahara	6	0	5
178	4/29/2020	NaN	Yemen	6	0	1
184	4/29/2020	Anguilla	UK	3	0	3
192	4/29/2020	Bonaire, Sint Eustatius and Saba	Netherlands	5	0	0
194	4/29/2020	British Virgin Islands	UK	6	1	3
203	4/29/2020	Diamond Princess cruise ship	Canada	0	1	0
272	4/29/2020	Northwest Territories	Canada	5	0	0
284	4/29/2020	Recovered	Canada	0	0	20327
285	4/29/2020	Recovered	US	0	0	120720
288	4/29/2020	Saint Barthelemy	France	6	0	6
289	4/29/2020	Saint Pierre and Miquelon	France	1	0	0
305	4/29/2020	Tibet	Mainland China	1	0	1

In [17]:
```
data[(data.Confirmed < 10)]
```

Out[17]:

	Date	State	Region	Confirmed	Deaths	Recovered
18	4/29/2020	NaN	Bhutan	7	0	5
98	4/29/2020	NaN	MS Zaandam	9	2	0
105	4/29/2020	NaN	Mauritania	8	1	6
126	4/29/2020	NaN	Papua New Guinea	8	0	0
140	4/29/2020	NaN	Sao Tome and Principe	8	0	4
177	4/29/2020	NaN	Western Sahara	6	0	5
178	4/29/2020	NaN	Yemen	6	0	1
184	4/29/2020	Anguilla	UK	3	0	3
192	4/29/2020	Bonaire, Sint Eustatius and Saba	Netherlands	5	0	0
194	4/29/2020	British Virgin Islands	UK	6	1	3
203	4/29/2020	Diamond Princess cruise ship	Canada	0	1	0
272	4/29/2020	Northwest Territories	Canada	5	0	0
284	4/29/2020	Recovered	Canada	0	0	20327
285	4/29/2020	Recovered	US	0	0	120720
288	4/29/2020	Saint Barthelemy	France	6	0	6
289	4/29/2020	Saint Pierre and Miquelon	France	1	0	0
305	4/29/2020	Tibet	Mainland China	1	0	1

In [18]:
```
data = data[~(data.Confirmed < 10)]  # To remove the records satisfying a particular c
```

In [19]:
```
data
```

Out[19]:

	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

304 rows × 6 columns

In [20]:
```
data.head(20)
```

Out[20]:

	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
5	4/29/2020	NaN	Antigua and Barbuda	24	3	11
6	4/29/2020	NaN	Argentina	4285	214	1192
7	4/29/2020	NaN	Armenia	1932	30	900
8	4/29/2020	NaN	Austria	15402	580	12779
9	4/29/2020	NaN	Azerbaijan	1766	23	1267
10	4/29/2020	NaN	Bahamas	80	11	23
11	4/29/2020	NaN	Bahrain	2921	8	1455
12	4/29/2020	NaN	Bangladesh	7103	163	150
13	4/29/2020	NaN	Barbados	80	7	39
14	4/29/2020	NaN	Belarus	13181	84	2072
15	4/29/2020	NaN	Belgium	47859	7501	11283
16	4/29/2020	NaN	Belize	18	2	9
17	4/29/2020	NaN	Benin	64	1	33
19	4/29/2020	NaN	Bolivia	1110	59	117
20	4/29/2020	NaN	Bosnia and Herzegovina	1677	65	710

3. In which Region, maximum number of Confirmed cases were recorded?

In [21]:
```
#df.groupby('Region').Confirmed.sum().sort_values(ascending = False).head(20)
```

In [22]:
```
data.head(2)
```

Out[22]:

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

In [23]:
```
data.groupby('Region').Confirmed.sum().sort_values(ascending = False).head(20)
```

Out[23]:
```
Region
US                  1039909
Spain                236899
Italy                203591
France               166536
UK                   166432
Germany              161539
Turkey               117589
Russia                99399
Iran                  93657
Mainland China        82861
Brazil                79685
Canada                52860
Belgium               47859
Netherlands           38993
Peru                  33931
India                 33062
Switzerland           29407
Ecuador               24675
Portugal              24505
Saudi Arabia          21402
Name: Confirmed, dtype: int64
```

4. In which Region, minimum number of Deaths cases were recorded ?

In [25]:
```
data.head(2)
```

Out[25]:

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

In [26]:
```
data.groupby('Region').Deaths.sum().sort_values(ascending = True).head(50)
```

Out[26]:

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
Brunei	1
Burundi	1
Botswana	1
Suriname	1
Benin	1
Djibouti	2
Angola	2
Libya	2
Chad	2
West Bank and Gaza	2
Belize	2
Zambia	3
Malawi	3
Nicaragua	3
Syria	3
Ethiopia	3
Antigua and Barbuda	3
Gabon	3
Hong Kong	4
Zimbabwe	4

Name: Deaths, dtype: int64

5. How many Confirmed, Deaths and Recovered cases were reported from India till 29 April 2020?

In [27]:
```
data.head(2)
```

Out[27]:

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

In [28]:
```
data[data.Region == 'India']
```

Out[28]:

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

6.a Sort the entire data wrt No. of Confirmed cases in ascending order.

In [29]:
```
#df.sort_values(by = ['Confirmed'], ascending = True)
```

In [30]:
```
data.head(2)
```

Out[30]:

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

In [31]:
```
data.sort_values(by = ['Confirmed'],ascending = True)
```

Out[31]:

	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

6.b Sort the entire data wrt No. of Recovered cases in descending order

In [32]:
```
data.sort_values(by = ['Recovered'], ascending = False)
```

Out[32]:

	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 [ ]:
```

```