Covid_19 Data analysis Using Python

We are taken a small dataset Covid-19, Just for understanding Purpose

The datset is available as a csv file. Download from Kaggle

We will analyze the data using the pandas data frame.

Import Library

In [28]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Uploading Csv fle

In [29]:
```
df = pd.read_csv("C:\Users\Syed Arif\Downloads\4. covid_19_data (1).csv")
```

```
  Input In [29]
    df = pd.read_csv("C:\Users\Syed Arif\Downloads\4. covid_19_data (1).csv")
                                                                            ^
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes in posi
tion 2-3: truncated \UXXXXXXXX escape
```

Remove Unicode error

We are Written Small r before Quatation

In [30]:
```
df = pd.read_csv(r"C:\Users\Syed Arif\Downloads\4. covid_19_data (1).csv")
```

In [31]:
```
df
```

Out[31]:

	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

Data Preprocessing

.head()

head is used show to the By default = 5 rows in the dataset

In [32]:
```
df.head()
```

Out[32]:

	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

.tail()

tail is used to show rows by Descending order

In [33]:
```
df.tail()
```

Out[33]:

	Date	State	Region	Confirmed	Deaths	Recovered
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

.shape

It shoe the total no of rows & Column in the dataset

In [34]:
```
df.shape
```

Out[34]:
```
(321, 6)
```

.Columns

It show the no of each Column

In [35]:
```
df.columns
```

Out[35]:
```
Index(['Date', 'State', 'Region', 'Confirmed', 'Deaths', 'Recovered'], dtype
='object')
```

.dtypes

This Attribute show the data type of each column

In [36]:
```
df.dtypes
```

Out[36]:
```
Date         object
State        object
Region       object
Confirmed     int64
Deaths        int64
Recovered     int64
dtype: object
```

.unique()

In a column, It show the unique value of specific column.

In [37]:
```
df["Region"].unique()
```

Out[37]:
```
array(['Afghanistan', 'Albania', 'Algeria', 'Andorra', 'Angola',
       'Antigua and Barbuda', 'Argentina', 'Armenia', 'Austria',
       'Azerbaijan', 'Bahamas', 'Bahrain', 'Bangladesh', 'Barbados',
       'Belarus', 'Belgium', 'Belize', 'Benin', 'Bhutan', 'Bolivia',
       'Bosnia and Herzegovina', 'Botswana', 'Brazil', 'Brunei',
       'Bulgaria', 'Burkina Faso', 'Burma', 'Burundi', 'Cabo Verde',
       'Cambodia', 'Cameroon', 'Central African Republic', 'Chad',
       'Chile', 'Colombia', 'Congo (Brazzaville)', 'Congo (Kinshasa)',
       'Costa Rica', 'Croatia', 'Cuba', 'Cyprus', 'Czech Republic',
       'Denmark', 'Diamond Princess', 'Djibouti', 'Dominica',
       'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador',
       'Equatorial Guinea', 'Eritrea', 'Estonia', 'Eswatini', 'Ethiopia',
       'Fiji', 'Finland', 'France', 'Gabon', 'Gambia', 'Georgia',
       'Germany', 'Ghana', 'Greece', 'Grenada', 'Guatemala', 'Guinea',
       'Guinea-Bissau', 'Guyana', 'Haiti', 'Holy See', 'Honduras',
       'Hungary', 'Iceland', 'India', 'Indonesia', 'Iran', 'Iraq',
       'Ireland', 'Israel', 'Italy', 'Ivory Coast', 'Jamaica', 'Japan',
       'Jordan', 'Kazakhstan', 'Kenya', 'Kosovo', 'Kuwait', 'Kyrgyzstan',
       'Laos', 'Latvia', 'Lebanon', 'Liberia', 'Libya', 'Liechtenstein',
       'Lithuania', 'Luxembourg', 'MS Zaandam', 'Madagascar', 'Malawi',
       'Malaysia', 'Maldives', 'Mali', 'Malta', 'Mauritania', 'Mauritius',
       'Mexico', 'Moldova', 'Monaco', 'Mongolia', 'Montenegro', 'Morocco',
       'Mozambique', 'Namibia', 'Nepal', 'Netherlands', 'New Zealand',
       'Nicaragua', 'Niger', 'Nigeria', 'North Macedonia', 'Norway',
       'Oman', 'Pakistan', 'Panama', 'Papua New Guinea', 'Paraguay',
       'Peru', 'Philippines', 'Poland', 'Portugal', 'Qatar', 'Romania',
       'Russia', 'Rwanda', 'Saint Kitts and Nevis', 'Saint Lucia',
       'Saint Vincent and the Grenadines', 'San Marino',
       'Sao Tome and Principe', 'Saudi Arabia', 'Senegal', 'Serbia',
       'Seychelles', 'Sierra Leone', 'Singapore', 'Slovakia', 'Slovenia',
       'Somalia', 'South Africa', 'South Korea', 'South Sudan', 'Spain',
       'Sri Lanka', 'Sudan', 'Suriname', 'Sweden', 'Switzerland', 'Syria',
       'Taiwan', 'Tanzania', 'Thailand', 'Timor-Leste', 'Togo',
       'Trinidad and Tobago', 'Tunisia', 'Turkey', 'UK', 'Uganda',
       'Ukraine', 'United Arab Emirates', 'Uruguay', 'Uzbekistan',
       'Venezuela', 'Vietnam', 'West Bank and Gaza', 'Western Sahara',
       'Yemen', 'Zambia', 'Zimbabwe', 'US', 'Canada', 'Mainland China',
       'Australia', 'Hong Kong', 'Macau'], dtype=object)
```

nuique()

It will show the total no of unque value from whole data frame

In [38]:
```
df.nunique()
```

Out[38]:
```
Date           1
State        137
Region       187
Confirmed    282
Deaths       142
Recovered    195
dtype: int64
```

.describe()

It show the Count, mean , median etc

In [39]:
```
df.describe()
```

Out[39]:

	Confirmed	Deaths	Recovered
count	321.000000	321.000000	321.000000
mean	9949.800623	709.152648	3030.277259
std	31923.853086	3236.162817	14364.870365
min	0.000000	0.000000	0.000000
25%	104.000000	2.000000	2.000000
50%	653.000000	12.000000	73.000000
75%	4655.000000	144.000000	587.000000
max	299691.000000	27682.000000	132929.000000

.value_counts

It Shows all the unique values with their count

In [40]:
```
df["Region"].value_counts()
```

Out[40]:
```
US                58
Mainland China    31
Canada            15
France            11
UK                11
                  ..
Guinea             1
Guinea-Bissau      1
Guyana             1
Haiti              1
Macau              1
Name: Region, Length: 187, dtype: int64
```

.isnull()

It shows the how many null values

In [41]:
```
df.isnull()
```

Out[41]:

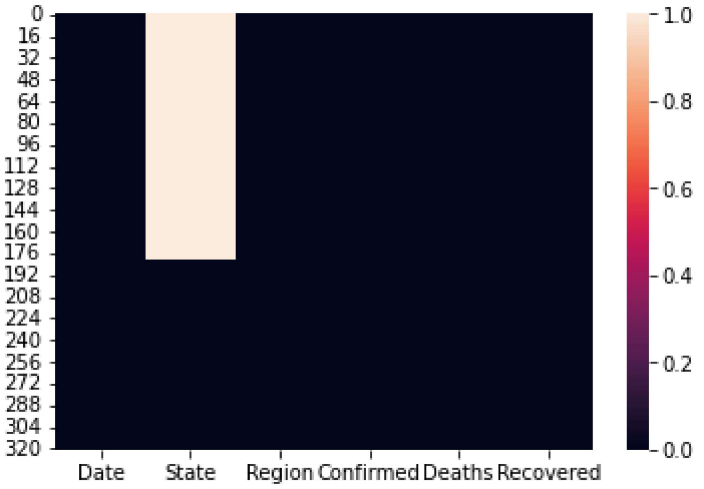
	Date	State	Region	Confirmed	Deaths	Recovered
0	False	True	False	False	False	False
1	False	True	False	False	False	False
2	False	True	False	False	False	False
3	False	True	False	False	False	False
4	False	True	False	False	False	False
...	...	...	...	...	...	...
316	False	False	False	False	False	False
317	False	False	False	False	False	False
318	False	False	False	False	False	False
319	False	False	False	False	False	False
320	False	False	False	False	False	False

321 rows × 6 columns

Data Visulaziation

In [42]:
```
sns.heatmap(df.isnull())
plt.show
```

Out[42]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



Show the number of Confirmed, Recovered & Death Cases in each region?

In [44]:
```
df.head(2)
```

Out[44]:

	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 [49]:
```
df.groupby("Region")["Confirmed"].sum().sort_values(ascending = False).head(10
```

Out[49]:
```
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 [53]:
```
df.groupby("Region")["Confirmed" , "Recovered"].sum()
```

```
C:\Users\Syed Arif\AppData\Local\Temp\ipykernel_11132\1451574403.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  df.groupby("Region")["Confirmed" , "Recovered"].sum()
```

Out[53]:

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

Remove all the records where Confirmed Cases s less then 10/

In [56]:
```
df[(df.Confirmed < 10)]
```

Out[56]:

	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

~ This sign is used to remove the values of condtion

In [57]:
```
df[~(df.Confirmed < 10)]
```

Out[57]:

	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 which region , maximum number of Confirmed cases were recorded ?

In [62]:
```
df.groupby('Region').Confirmed.sum().sort_values(ascending = True).head(10)
```

Out[62]:
```
Region
Yemen                     6
Western Sahara            6
Bhutan                    7
Papua New Guinea          8
Sao Tome and Principe     8
Mauritania                8
MS Zaandam                9
Gambia                   10
Holy See                 10
Suriname                 10
Name: Confirmed, dtype: int64
```

In which region , Minimum number of Deaths were recorded ?

In [65]:
```
df.groupby('Region').Deaths.sum().sort_values(ascending = True).head(10)
```

Out[65]:
```
Region
Laos                0
Mongolia            0
Mozambique          0
Cambodia            0
Fiji                0
Namibia             0
Nepal               0
Madagascar          0
Macau               0
Papua New Guinea    0
Name: Deaths, dtype: int64
```

How many Confirmed , Deaths Recovered cases were reported in Canada?

In [71]:
```
df[df.Region == "Canada"].sum()
```

Out[71]:
```
Date         4/29/20204/29/20204/29/20204/29/20204/29/20204...
State        AlbertaBritish ColumbiaDiamond Princess cruise...
Region       CanadaCanadaCanadaCanadaCanadaCanadaCanadaCana...
Confirmed                                                52865
Deaths                                                    3155
Recovered                                                20327
dtype: object
```