Task3- Analysing Covid-19 Data (DataAnalytics) DEBANSHU SAHA

Data cleaning, Data transformation using python

Importing all the necessary libraries

In [3]:
```
import os
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
from math import sqrt
from datetime import datetime

%matplotlib inline
```

Importing dataset

In [2]:
```
raw_data_confirmed = pd.read_csv('C:\\Users\\DELL\\Downloads\\COVID-19-master\\COV
raw_data_deaths = pd.read_csv('C:\\Users\\DELL\\Downloads\\COVID-19-master\\COVID-
raw_data_Recovered = pd.read_csv('C:\\Users\\DELL\\Downloads\\COVID-19-master\\COV

print("The Shape of confirmed is: ", raw_data_confirmed.shape)
print("The Shape of deaths is: ", raw_data_deaths.shape)
print("The Shape of recovered is: ", raw_data_Recovered.shape)

raw_data_confirmed.head()
```

```
The Shape of confirmed is:  (271, 350)
The Shape of deaths is:  (271, 350)
The Shape of recovered is:  (256, 350)
```

Out[2]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	
1	NaN	Albania	41.15330	20.168300	0	0	0	0	
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	

5 rows × 350 columns

Un-Pivoting the data

In [3]:
```
raw_data_confirmed2 = pd.melt(raw_data_confirmed, id_vars=['Province/State', 'Coun
                              var_name=['Date'])
raw_data_deaths2 = pd.melt(raw_data_deaths, id_vars=['Province/State', 'Country/Re
                           var_name=['Date'])
raw_data_Recovered2 = pd.melt(raw_data_Recovered, id_vars=['Province/State', 'Coun
```

```
                                var_name=['Date'])


print("The Shape of Confirmed is: ", raw_data_confirmed2.shape)
print("The Shape of deaths is: ", raw_data_deaths2.shape)
print("The Shape of recovered is: ", raw_data_Recovered2.shape)


raw_data_confirmed2
```

```
The Shape of Confirmed is:  (93766, 6)
The Shape of deaths is:  (93766, 6)
The Shape of recovered is:  (88576, 6)
```

Out[3]:

	Province/State	Country/Region	Lat	Long	Date	value
0	NaN	Afghanistan	33.939110	67.709953	1/22/20	0
1	NaN	Albania	41.153300	20.168300	1/22/20	0
2	NaN	Algeria	28.033900	1.659600	1/22/20	0
3	NaN	Andorra	42.506300	1.521800	1/22/20	0
4	NaN	Angola	-11.202700	17.873900	1/22/20	0
...	...	...	...	...	...	...
93761	NaN	Vietnam	14.058324	108.277199	1/1/21	1474
93762	NaN	West Bank and Gaza	31.952200	35.233200	1/1/21	139223
93763	NaN	Yemen	15.552727	48.516388	1/1/21	2101
93764	NaN	Zambia	-13.133897	27.849332	1/1/21	20997
93765	NaN	Zimbabwe	-19.015438	29.154857	1/1/21	14084

93766 rows × 6 columns

In [4]:
```
raw_data_confirmed2.head()
```

Out[4]:

	Province/State	Country/Region	Lat	Long	Date	value
0	NaN	Afghanistan	33.93911	67.709953	1/22/20	0
1	NaN	Albania	41.15330	20.168300	1/22/20	0
2	NaN	Algeria	28.03390	1.659600	1/22/20	0
3	NaN	Andorra	42.50630	1.521800	1/22/20	0
4	NaN	Angola	-11.20270	17.873900	1/22/20	0

Converting the new column to dates

In [5]:
```
raw_data_confirmed2['Date'] = pd.to_datetime(raw_data_confirmed2['Date'])
raw_data_deaths2['Date'] = pd.to_datetime(raw_data_deaths2['Date'])
raw_data_Recovered2['Date'] = pd.to_datetime(raw_data_Recovered2['Date'])
```

Renaming the Values

In [6]:
```
raw_data_confirmed2.columns = raw_data_confirmed2.columns.str.replace('value', 'Co
raw_data_deaths2.columns = raw_data_deaths2.columns.str.replace('value', 'Deaths')
```

```
raw_data_Recovered2.columns = raw_data_Recovered2.columns.str.replace('value', 'Re
```

In [7]:
```
raw_data_confirmed2.head()
```

Out[7]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed
0	NaN	Afghanistan	33.93911	67.709953	2020-01-22	0
1	NaN	Albania	41.15330	20.168300	2020-01-22	0
2	NaN	Algeria	28.03390	1.659600	2020-01-22	0
3	NaN	Andorra	42.50630	1.521800	2020-01-22	0
4	NaN	Angola	-11.20270	17.873900	2020-01-22	0

Investigating the NULL values

In [8]:
```
raw_data_Recovered2.isnull().sum()
```

Out[8]:
```
Province/State    65394
Country/Region        0
Lat                   0
Long                  0
Date                  0
Recovered             0
dtype: int64
```

Dealing with NULL values

In [9]:
```
raw_data_confirmed2['Province/State'].fillna(raw_data_confirmed2['Country/Region'],
raw_data_deaths2['Province/State'].fillna(raw_data_deaths2['Country/Region'], inpla
raw_data_Recovered2['Province/State'].fillna(raw_data_Recovered2['Country/Region'],
```

In [10]:
```
raw_data_Recovered2.isnull().sum()
```

Out[10]:
```
Province/State    0
Country/Region    0
Lat               0
Long              0
Date              0
Recovered         0
dtype: int64
```

printing shapes before the join

In [11]:
```
print("The Shape of confirmed is: ", raw_data_confirmed2.shape)
print("The Shape of deaths is: ", raw_data_deaths2.shape)
print("The Shape of recovered is: ", raw_data_Recovered2.shape)
```

```
The Shape of confirmed is:  (93766, 6)
The Shape of deaths is:  (93766, 6)
The Shape of recovered is:  (88576, 6)
```

In [12]:
```
raw_data_confirmed2
```

Out[12]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed
0	Afghanistan	Afghanistan	33.939110	67.709953	2020-01-22	0
1	Albania	Albania	41.153300	20.168300	2020-01-22	0
2	Algeria	Algeria	28.033900	1.659600	2020-01-22	0
3	Andorra	Andorra	42.506300	1.521800	2020-01-22	0
4	Angola	Angola	-11.202700	17.873900	2020-01-22	0
...	...	...	...	...	...	...
93761	Vietnam	Vietnam	14.058324	108.277199	2021-01-01	1474
93762	West Bank and Gaza	West Bank and Gaza	31.952200	35.233200	2021-01-01	139223
93763	Yemen	Yemen	15.552727	48.516388	2021-01-01	2101
93764	Zambia	Zambia	-13.133897	27.849332	2021-01-01	20997
93765	Zimbabwe	Zimbabwe	-19.015438	29.154857	2021-01-01	14084

93766 rows × 6 columns

In [13]:
```
full_join = raw_data_confirmed2.merge(raw_data_deaths2[['Province/State','Country/
                                      how = 'left',
                                      left_on = ['Province/State','Country/Region'
                                      right_on = ['Province/State', 'Country/Regio

full_join.head()
```

Out[13]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths
0	Afghanistan	Afghanistan	33.93911	67.709953	2020-01-22	0	0
1	Albania	Albania	41.15330	20.168300	2020-01-22	0	0
2	Algeria	Algeria	28.03390	1.659600	2020-01-22	0	0
3	Andorra	Andorra	42.50630	1.521800	2020-01-22	0	0
4	Angola	Angola	-11.20270	17.873900	2020-01-22	0	0

In [14]:
```
full_join = full_join.merge(raw_data_Recovered2[['Province/State','Country/Region'
                                      how = 'left',
                                      left_on = ['Province/State','Country/Region'
                                      right_on = ['Province/State', 'Country/Regio

full_join.head()
```

Out[14]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered
0	Afghanistan	Afghanistan	33.93911	67.709953	2020-01-22	0	0	0.0
1	Albania	Albania	41.15330	20.168300	2020-01-22	0	0	0.0
2	Algeria	Algeria	28.03390	1.659600	2020-01-22	0	0	0.0
3	Andorra	Andorra	42.50630	1.521800	2020-01-22	0	0	0.0
4	Angola	Angola	-11.20270	17.873900	2020-01-22	0	0	0.0

Adding Month and Year as a new Column

In [15]:
```
full_join['Month-Year'] = full_join['Date'].dt.strftime('%b-%Y')
```

In [16]:
```
full_join.head()
```

Out[16]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered	M
0	Afghanistan	Afghanistan	33.93911	67.709953	2020-01-22	0	0	0.0	
1	Albania	Albania	41.15330	20.168300	2020-01-22	0	0	0.0	
2	Algeria	Algeria	28.03390	1.659600	2020-01-22	0	0	0.0	
3	Andorra	Andorra	42.50630	1.521800	2020-01-22	0	0	0.0	
4	Angola	Angola	-11.20270	17.873900	2020-01-22	0	0	0.0	

In [17]:
```
full_join2 = full_join.copy()

#creating a new date columns - 1
full_join2['Date - 1'] = full_join2['Date'] + pd.Timedelta(days=1)
full_join2.rename(columns={'Confirmed': 'Confirmed - 1', 'Deaths': 'Deaths - 1', '
                           'Date': 'Date Minus 1'}, inplace=True)
```

In [18]:
```
full_join2.head()
```

Out[18]:

	Province/State	Country/Region	Lat	Long	Date Minus 1	Confirmed - 1	Deaths - 1	Recovered - 1
0	Afghanistan	Afghanistan	33.93911	67.709953	2020-01-22	0	0	0.0
1	Albania	Albania	41.15330	20.168300	2020-01-22	0	0	0.0
2	Algeria	Algeria	28.03390	1.659600	2020-01-22	0	0	0.0
3	Andorra	Andorra	42.50630	1.521800	2020-01-22	0	0	0.0
4	Angola	Angola	-11.20270	17.873900	2020-01-22	0	0	0.0

In [19]:
```
full_join3 = full_join.merge(full_join2[['Province/State', 'Country/Region','Confi
                            'Recovered - 1', 'Date - 1', 'Date Minus 1']], how = '
                             left_on = ['Province/State','Country/Region','Date'],
                             right_on = ['Province/State', 'Country/Region','Date
full_join3.head()
```

Out[19]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered
0	Afghanistan	Afghanistan	33.93911	67.709953	2020-01-22	0	0	0.0
1	Albania	Albania	41.15330	20.168300	2020-01-22	0	0	0.0
2	Algeria	Algeria	28.03390	1.659600	2020-01-22	0	0	0.0
3	Andorra	Andorra	42.50630	1.521800	2020-01-22	0	0	0.0
4	Angola	Angola	-11.20270	17.873900	2020-01-22	0	0	0.0

In [20]:
```
full_join3['Confirmed Daily'] = full_join3['Confirmed'] - full_join3['Confirmed - 1
full_join3['Deaths Daily'] = full_join3['Deaths'] - full_join3['Deaths - 1']
full_join3['Recovered Daily'] = full_join3['Recovered'] - full_join3['Recovered - 1

print(full_join3.shape)
```

```
(93766, 17)
```

In [21]:
```
full_join3.head()
```

Out[21]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered	N
0	Afghanistan	Afghanistan	33.93911	67.709953	2020-01-22	0	0	0.0	
1	Albania	Albania	41.15330	20.168300	2020-01-22	0	0	0.0	
2	Algeria	Algeria	28.03390	1.659600	2020-01-22	0	0	0.0	
3	Andorra	Andorra	42.50630	1.521800	2020-01-22	0	0	0.0	
4	Angola	Angola	-11.20270	17.873900	2020-01-22	0	0	0.0	

In [ ]:

In [22]:
```
##########################################################################################
######################### Braking the numbers by Day #####################################
##########################################################################################


#creating a new df
full_join2 = full_join.copy()

#creating a new date columns - 1
full_join2['Date - 1'] = full_join2['Date'] + pd.Timedelta(days=1)
full_join2.rename(columns={'Confirmed': 'Confirmed - 1', 'Deaths': 'Deaths - 1', 'R
                           'Date': 'Date Minus 1'}, inplace=True)

#Joing on the 2 DFs
full_join3 = full_join.merge(full_join2[['Province/State', 'Country/Region','Confi
                             'Recovered - 1', 'Date - 1', 'Date Minus 1']], how = '
                              left_on = ['Province/State','Country/Region','Date'],
                              right_on = ['Province/State', 'Country/Region','Date

#minus_onedf.rename(columns={'Confirmed': 'Confirmed - 1', 'Deaths': 'Deaths - 1',

full_join3.head()

# Additional Calculations
full_join3['Confirmed Daily'] = full_join3['Confirmed'] - full_join3['Confirmed - 1
full_join3['Deaths Daily'] = full_join3['Deaths'] - full_join3['Deaths - 1']
full_join3['Recovered Daily'] = full_join3['Recovered'] - full_join3['Recovered - 1

print(full_join3.shape)
```

```
(93766, 17)
```

In [23]:
```
full_join3.head()
```

Out[23]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered	N
0	Afghanistan	Afghanistan	33.93911	67.709953	2020-01-22	0	0	0.0	
1	Albania	Albania	41.15330	20.168300	2020-01-22	0	0	0.0	
2	Algeria	Algeria	28.03390	1.659600	2020-01-22	0	0	0.0	
3	Andorra	Andorra	42.50630	1.521800	2020-01-22	0	0	0.0	
4	Angola	Angola	-11.20270	17.873900	2020-01-22	0	0	0.0	

In [24]:
```
# Additing manually the numbers for first day

full_join3['Confirmed Daily'].loc[full_join3['Date'] == '2020-01-22'] = full_join3
full_join3['Deaths Daily'].loc[full_join3['Date'] == '2020-01-22'] = full_join3['D
full_join3['Recovered Daily'].loc[full_join3['Date'] == '2020-01-22'] = full_join3

# deleting columns
del full_join3['Confirmed - 1']
del full_join3['Deaths - 1']
del full_join3['Recovered - 1']
del full_join3['Date - 1']
del full_join3['Date Minus 1']
```

```
C:\Users\DELL\anaconda3\lib\site-packages\pandas\core\indexing.py:1637: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_block(indexer, value, name)
```

Removing Negative values

In [25]:
```
full_join3[full_join3["Deaths Daily"]<0]
```

Out[25]:

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recove
14778	Iceland	Iceland	64.963100	-19.020800	2020-03-16	180	0	
15653	Philippines	Philippines	12.879721	121.774017	2020-03-19	217	17	
15862	Iceland	Iceland	64.963100	-19.020800	2020-03-20	409	0	
16134	India	India	20.593684	78.962880	2020-03-21	330	4	2
16311	Quebec	Canada	52.939900	-73.549100	2020-03-22	219	4	N
...	...	...	...	...	...	...	...	
87140	Ireland	Ireland	53.142400	-7.692100	2020-12-08	74682	2097	2336
87259	Yemen	Yemen	15.552727	48.516388	2020-12-08	2078	606	138
88203	France	France	46.227600	2.213700	2020-12-12	2350793	57210	15255
88585	Tajikistan	Tajikistan	38.861000	71.276100	2020-12-13	12704	88	1213
92439	Bosnia and Herzegovina	Bosnia and Herzegovina	43.915900	17.679100	2020-12-28	109911	3942	7612

83 rows × 12 columns

In [26]:
```
full_join3['Deaths Daily']=np.where(full_join3['Deaths Daily']<0 ,0,full_join3['De
```

In [27]:
```
full_join3['Confirmed Daily']=np.where(full_join3['Confirmed Daily']<0 ,0,full_joi
```

In [28]:
```
full_join3['Recovered Daily']=np.where(full_join3['Recovered Daily']<0 ,0,full_joi
```

Exporting output file

In [30]:
```
path = "C:\\Users\\DELL\\Desktop\\"

# Changing my CWD
os.chdir(path)

full_join3.to_csv('CoronaVirus Data.csv')
```