MINI PROJECT

Member 1: Dharssini K

Reg No : 2033009

Member 2: Tanushree R

Reg No : 2033036

Covid 19-Data Analysis

We have taken a small dataset od Covid 19,The data used here is the record as on 29-April-2020

The datasource is downloaded from Kaggle as a CSV file

We are going to analyze this data using Pandas DataFrame

In [155...
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import os
```

In [156...
```
data=pd.read_csv(r"C:\Users\Lenovo\Desktop\dataset.csv")
```

In [157...
```
data
```

Out[157...

	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

	Date	State	Region	Confirmed	Deaths	Recovered
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

1.Show the number of confirmed and recovered cases in each region

In [158...
```
data.head(2)
```

Out[158...

	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 [159...
```
data.groupby('Region')['Confirmed','Recovered'].sum()
```

```
<ipython-input-159-20fd7b835859>: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[159...

Region	Confirmed	Recovered
Afghanistan	1939	252
Albania	766	455

Region	Confirmed	Recovered
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 [160…
```
data.head(2)
```

Out[160…

	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 [161…
```
data = data[~(data.Confirmed<10)]
```

In [162…
```
data.head(20)
```

Out[162…

	Date	State	Region	Confirmed	Deaths	Recovered

	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 minimum number of Deaths cases where recorded

In [163…

```
data.groupby('Region').Deaths.sum().sort_values(ascending= True).head(50)
```

Out[163…

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
```

4.In which region maximum number of confirmed cases where recorded

In [164…
```
data.groupby('Region').Confirmed.sum().sort_values(ascending= False).head(20)
```

Out[164…
```
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
```

5. How many confirmed,deaths and recovered cases were reported from India till April 29

In [165…
```
data[data.Region=='India']
```

Out[165…

	Date	State	Region	Confirmed	Deaths	Recovered

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

Q 6-A) Sort the entire data wrt No. of Confirmed cases in ascending order

In [166…
```
data.sort_values(by=['Confirmed'],ascending= True)
```

Out[166…

	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 Confirmed cases in descending order

In [167…
```
data.sort_values(by=['Confirmed'],ascending= False)
```

Out[167...

	Date	State	Region	Confirmed	Deaths	Recovered
265	4/29/2020	New York	US	299691	23477	0
153	4/29/2020	NaN	Spain	236899	24275	132929
80	4/29/2020	NaN	Italy	203591	27682	71252
168	4/29/2020	NaN	UK	165221	26097	0
57	4/29/2020	NaN	France	165093	24087	48228
...	...	...	...	...	...	...
144	4/29/2020	NaN	Seychelles	11	0	6
27	4/29/2020	NaN	Burundi	11	1	4
59	4/29/2020	NaN	Gambia	10	1	8
156	4/29/2020	NaN	Suriname	10	1	8
70	4/29/2020	NaN	Holy See	10	0	2

304 rows × 6 columns

7.Average number of confirmed,Deaths and recovered cases on Apr 29 Worldwide

In [168...

```
data[['Confirmed','Deaths','Recovered']].mean()
```

Out[168...

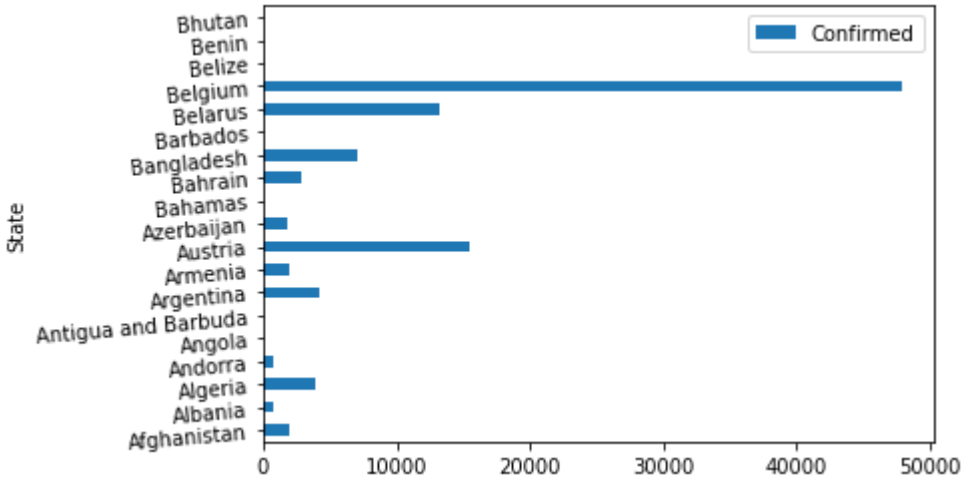
```
Confirmed    10505.944079
Deaths         748.792763
Recovered     2735.651316
dtype: float64
```

8.Analyse Confirmed cases through visualization

In [169...

```
datanew={'State':['Afghanistan','Albania','Algeria','Andorra','Angola','Antigua and Barbuda','Argentina','Armenia','Austr
         'Confirmed':[1939,766,3848,743,27,24,4285,1932,15402,1766,80,2921,7103,80,13181,47859,18,64,7]}
datanew = pd.DataFrame(datanew,columns=['State','Confirmed'])
```

```
datanew.plot(x ='State', y='Confirmed', kind = 'barh',rot=5, fontsize=10)
plt.show()
```

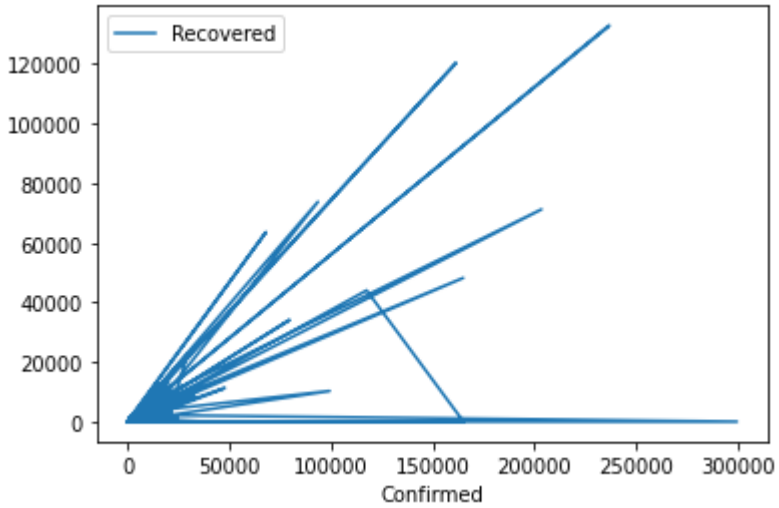


From this we conclude that Belgium has more number of confirmed cases on Apr 29 from first 20 datas

In [170…

```
data.plot(x='Confirmed',y='Recovered')
```

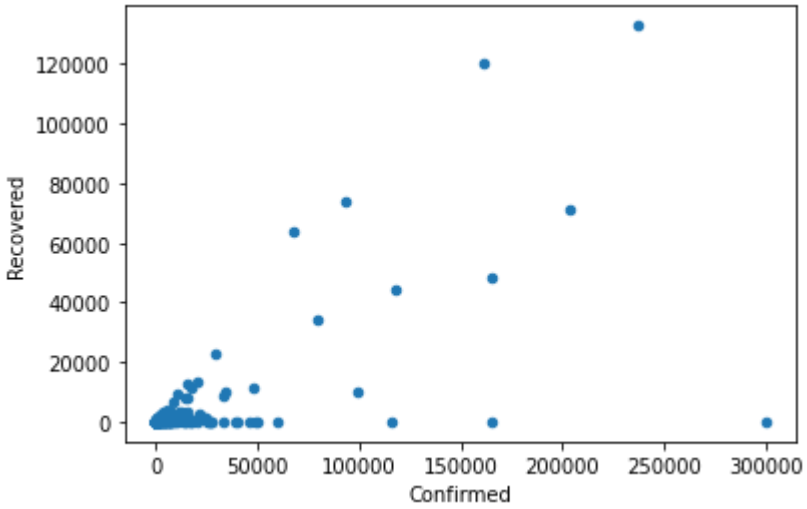
Out[170… <AxesSubplot:xlabel='Confirmed'>



9.Finding Correlation between Confirmed and Recovered cases

In [171…
```
data.plot(x="Confirmed", y="Recovered", kind="scatter")
```

Out[171… <AxesSubplot:xlabel='Confirmed', ylabel='Recovered'>



From this we could conclude that there is no correlation between Confirmed and Recovered Cases

10.Summary Statistics of entire data

In [172…
```
data.describe()
```

Out[172…

	Confirmed	Deaths	Recovered
count	304.000000	304.000000	304.000000
mean	10505.944079	748.792763	2735.651316
std	32717.761818	3321.228882	13064.686914
min	10.000000	0.000000	0.000000
25%	138.750000	3.000000	4.000000
50%	776.500000	14.500000	91.500000
75%	5288.500000	170.500000	603.500000

	Confirmed	Deaths	Recovered
max	299691.000000	27682.000000	132929.000000

11.Regression between Confirmed and Death cases

In [173…

```
#collecting x and y value
X=data['Confirmed'].values
Y=data['Deaths'].values
%matplotlib inline
```

In [174…

```
#mean X and Y
mean_x=np.mean(X)
mean_y=np.mean(Y)
#total no of values
m=len(x)
#using this formula to calculate a(slope) and b(intercept)
numer=0
denom=0
for i in range (m):
    numer +=(X[i]-mean_x)*(Y[i]-mean_y)
    denom += (X[i]-mean_x)**2
a=numer/denom #a=Σ[(x(i)-x)(y(i)-y))  /  (Σ(x(i)-x)^2) )]
b=mean_y-(a*mean_x) #[y-ax]
print(a,b)
```

```
0.08452962223128024 -139.27072101848182
```

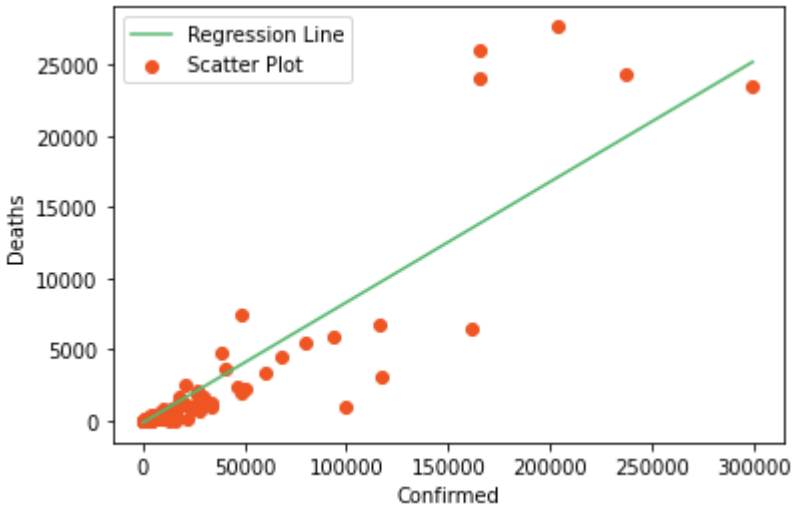
In [175…

```
#plotting values and regression line
max_x= np.max(X)
min_x= np.min(X)
#Calculating line values x and y
x=np.linspace(min_x,max_x)
y=b+a*x

#plotting the line

plt.plot(x,y,color='#58b970',label='Regression Line')
#plotting scatter points
plt.scatter(X,Y,c='#ef5423',label='Scatter Plot')
```

```
plt.xlabel('Confirmed')
plt.ylabel('Deaths')
plt.legend()
plt.show()
```



In the above graph we have fitted a simple linear regression line.

For a unit increase in x, y value will be increased by 0.08452962223128024 units.