SDS ASSIGNMENT FINAL REPORT

NAME:ANUBUTHI

SRN:PES1UG20CS065

- we begin the assignment by first analysing the data and finding out about the various datatypes and null values
- Initially the data is as such

```
#understanding data
print(data.dtypes)
data.isnull().sum()
```

```
name                object
region              object
cases              float64
cases100k          float64
cases7days         float64
cases7days100k     float64
cases24h           float64
deaths             float64
deaths100k         float64
deaths7days        float64
deaths7days100k    float64
deaths24h          float64
dtype: object
```

```
name               0
region             0
cases              2
cases100k          2
cases7days         5
cases7days100k     3
cases24h           0
deaths             1
deaths100k         2
deaths7days        1
deaths7days100k    1
deaths24h          0
dtype: int64
```

[7]:

```
data=pd.read_csv("../input/whocovid/42.csv")
data
```

[7]:

	name	region	cases	cases100k	cases7days	cases7days100k	cases24h	deaths	deaths100k	deaths7days	deaths7days100k	deaths24h
0	Curaçao	Americas	31693962.0	9575.14	511165.0	154.43	81768.0	509219.0	153.84	8915.0	2.69	1369.0
1	Cambodia	Western Pacific	NaN	1778.68	NaN	7.81	15384.0	325772.0	23.61	1294.0	0.09	196.0
2	American Samoa	Western Pacific	15572570.0	7326.22	84982.0	39.98	12950.0	433777.0	204.07	2343.0	1.10	384.0
3	Ecuador	Americas	5823452.0	8578.28	173229.0	255.17	28754.0	99453.0	146.50	546.0	0.80	88.0
4	North Macedonia	Europe	5585316.0	3827.27	131581.0	NaN	19719.0	155230.0	106.37	4591.0	3.15	677.0
...	...	...	...	...	...	...	...	...	...	...	...	...
231	Fiji	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0
232	Slovenia	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0
233	Bahamas	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0
234	Colombia	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0
235	New Caledonia	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0

236 rows × 12 columns

INTRODUCTORY QUESTIONS

1) Create a column called fatality rate=deaths/cases*100

- New column is created using the formula
- And this is appended into the existing dataframe

```
#q1 new coulumn fatality
data['fatality_rate']=data['deaths']*100/data['cases']
data
```

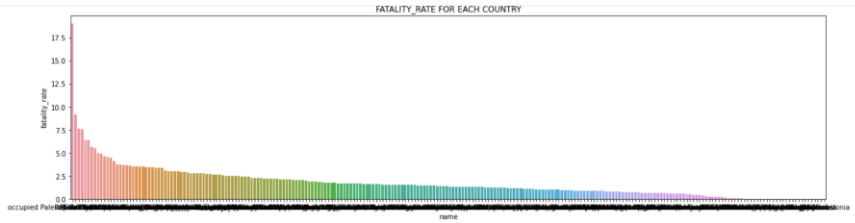
9]:

name	region	cases	cases100k	cases7days	cases7days100k	cases24h	deaths	deaths100k	deaths7days	deaths7days100k	deaths24h	fatality_rate
Curaçao	Americas	31693962.0	9575.14	511165.0	154.43	81768.0	509219.0	153.84	8915.0	2.69	1369.0	1.606675
Cambodia	Western Pacific	NaN	1778.68	NaN	7.81	15384.0	325772.0	23.61	1294.0	0.09	196.0	NaN
American Samoa	Western Pacific	15572570.0	7326.22	84982.0	39.98	12950.0	433777.0	204.07	2343.0	1.10	384.0	2.785520
Ecuador	Americas	5823452.0	8578.28	173229.0	255.17	28754.0	99453.0	146.50	546.0	0.80	88.0	1.707801
North Macedonia	Europe	5585316.0	3827.27	131581.0	NaN	19719.0	155230.0	106.37	4591.0	3.15	677.0	2.779252
...	...	...	...	...	...	...	...	...	...	...	...	...
Fiji	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN
Slovenia	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN
Bahamas	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN
Colombia	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN
New Caledonia	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN

rows × 13 columns

2) Plot highest number of fatality rates.

- We plot the graph using seaborn and sort it in descending order.
- We find the max number of cases using idxmax.and display all the details of that country
- Bar graph is ploted for the country with the maximum number of cases



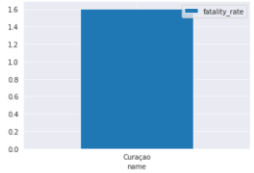
+ Code + Markdown

```
column=data['cases']
val1=column.max()
val1
```

```
31693962.0
```

```
country1=data.loc[data['cases']==val1]
```

```
bp=country1.plot.bar(x='name',y='fatality_rate',rot=0)
```




```
#q2 barchart
plt.figure(figsize=(20,5))
# make barplot and sort bars in descending order
sns.barplot(x='name', y='fatality_rate',data=data,
            order=data.sort_values('fatality_rate',ascending = False).name).set_title("FATALITY_RATE FOR EACH COUNTR
print("COUNTRY WITH HIGHEST CASES:")
data.loc[data['cases'].idxmax()]
```

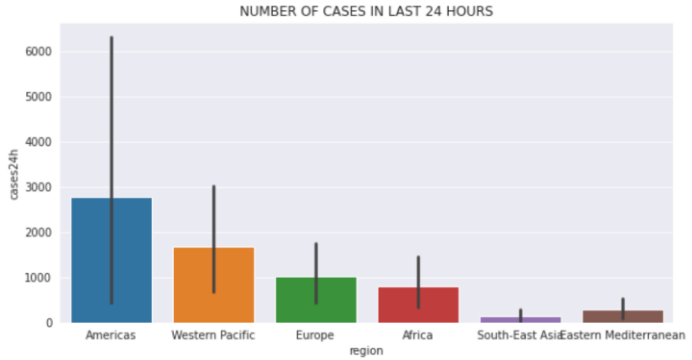
```
COUNTRY WITH HIGHEST CASES:
]: name                 Curaçao
region               Americas
cases              31693962.0
cases100k            9575.14
cases7days          511165.0
cases7days100k        154.43
cases24h            81768.0
deaths              509219.0
deaths100k            153.84
deaths7days           8915.0
deaths7days100k         2.69
deaths24h            1369.0
fatality_rate       1.606675
Name: 0, dtype: object
```

3) Bar graph to represent number of cases in the last 24 hours for all region using seaborn.

- use to analyze the trend in cases worldwide over the last 24 hours
- Using bar chart makes visualization easier

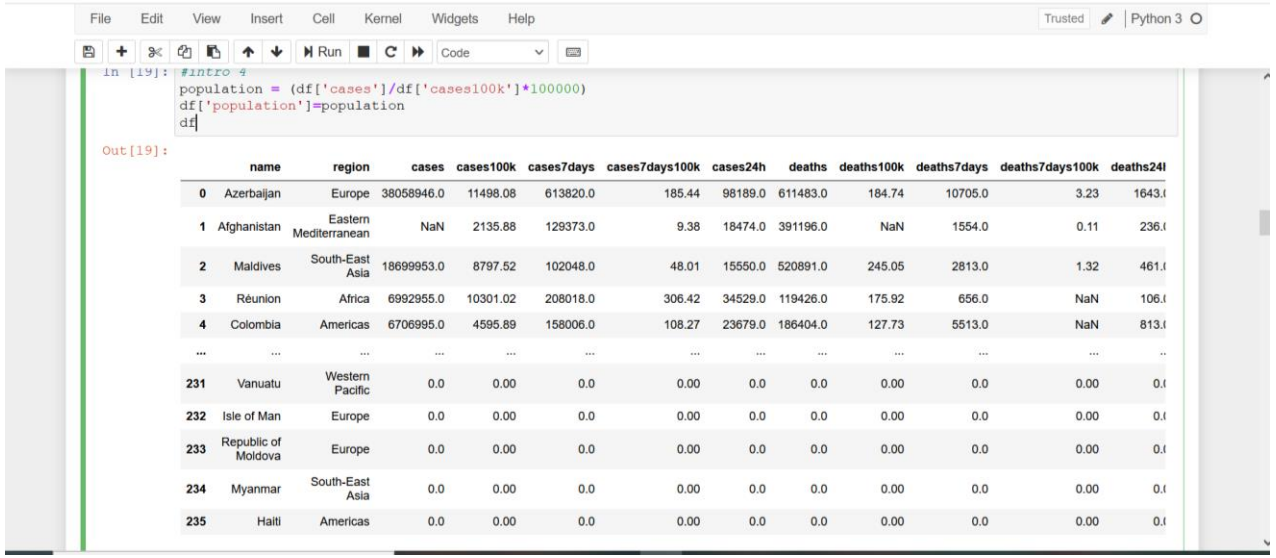
```
#q3 barchart
import matplotlib.pyplot as plt
plt.figure(figsize=(10,5))
data.groupby(['region'])
sns.set_style('darkgrid')
sns.barplot(x='region',y='cases24h',data=data).set_title("NUMBER OF CASES IN LAST 24 HOURS")
```

```
.]: Text(0.5, 1.0, 'NUMBER OF CASES IN LAST 24 HOURS')
```



4) We have to create a new column population

- We use the formula to find out the population
- Append it into the existing dataframe



File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

In [19]:
```
#intro 4
population = (df['cases']/df['cases100k']*100000)
df['population']=population
df
```

Out[19]:

	name	region	cases	cases100k	cases7days	cases7days100k	cases24h	deaths	deaths100k	deaths7days	deaths7days100k	deaths24h
0	Azerbaijan	Europe	38058946.0	11498.08	613820.0	185.44	98189.0	611483.0	184.74	10705.0	3.23	1643.
1	Afghanistan	Eastern Mediterranean	NaN	2135.88	129373.0	9.38	18474.0	391196.0	NaN	1554.0	0.11	236.
2	Maldives	South-East Asia	18699953.0	8797.52	102048.0	48.01	15550.0	520891.0	245.05	2813.0	1.32	461.
3	Réunion	Africa	6992955.0	10301.02	208018.0	306.42	34529.0	119426.0	175.92	656.0	NaN	106.
4	Colombia	Americas	6706995.0	4595.89	158006.0	108.27	23679.0	186404.0	127.73	5513.0	NaN	813.
...	...	...	...	...	...	...	...	...	...	...	...	..
231	Vanuatu	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.
232	Isle of Man	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.
233	Republic of Moldova	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.
234	Myanmar	South-East Asia	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.
235	Haiti	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.

TASK QUESTIONS

1) Data cleansing:
i) Finding the n/a values in the csv file
ii) Finding the mean
iii) Replacing the n/a values with mean for all the column
iv) We leave the 0 values as is to maintain an overall trend i.e some European and some American countries and 0 cases and deaths these values are not altered because they help in analyzing the world wide trend of spread and control

```
#task7
#identifying and displaying all the rows which have a null value
isnan=data.isnull()
rownan=isnan.any(axis=1)
rowswithnan=data[rownan]
rowswithnan
```

[15]:

	name	region	cases	cases100k	cases7days	cases7days100k	cases24h	deaths	deaths100k	deaths7days	deaths7days100k	deaths24h	fatality_rate	population
1	Cambodia	Western Pacific	NaN	1778.68	NaN	7.81	15384.0	325772.0	23.61	1294.0	0.09	196.0	NaN	NaN
4	North Macedonia	Europe	5585316.0	3827.27	131581.0	NaN	19719.0	155230.0	106.37	4591.0	3.15	677.0	2.779252	1.459347e+08
12	Jamaica	Americas	3107576.0	3736.56	40665.0	NaN	7548.0	NaN	81.90	291.0	0.35	62.0	NaN	8.316676e+07
16	Paraguay	Americas	2105891.0	3550.73	5128.0	8.65	716.0	63764.0	107.51	346.0	NaN	89.0	3.027887	5.930868e+07
20	Estonia	Europe	1578597.0	4787.71	4254.0	NaN	748.0	144428.0	438.03	NaN	0.42	28.0	9.149137	3.297186e+07
27	Congo	Africa	1192101.0	NaN	NaN	51.80	2702.0	20346.0	53.90	212.0	0.56	45.0	1.706735	NaN
53	Switzerland	Europe	391566.0	7686.64	NaN	155.18	1168.0	4743.0	93.10	148.0	2.89	22.0	1.211290	5.094111e+06
54	Jordan	Eastern Mediterranean	378911.0	1769.52	3456.0	16.14	0.0	9574.0	NaN	192.0	0.90	0.0	2.526715	2.141321e+07
68	Egypt	Eastern Mediterranean	273451.0	NaN	7354.0	25.86	1102.0	3299.0	11.60	75.0	0.27	14.0	1.206432	NaN
73	The United Kingdom	Europe	NaN	9002.30	10030.0	358.96	1775.0	3748.0	134.15	119.0	4.28	15.0	NaN	NaN
79	Holy See	Europe	219442.0	5439.86	NaN	158.16	1286.0	5033.0	124.76	128.0	3.18	20.0	2.293545	4.033964e+06
112	Sweden	Europe	77920.0	1201.32	NaN	37.00	238.0	2421.0	NaN	80.0	1.24	13.0	3.107033	6.486199e+06
224	Netherlands	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
225	Jersey	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
226	Sudan	Eastern Mediterranean	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
227	Lesotho	Africa	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
228	Uruguay	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
229	Singapore	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
230	Isle of Man	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
231	Fiji	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
232	Slovenia	Europe	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
233	Bahamas	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
234	Colombia	Americas	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN
235	New Caledonia	Western Pacific	0.0	0.00	0.0	0.00	0.0	0.0	0.00	0.0	0.00	0.0	NaN	NaN

```
#all null values have been filled
data.isnull().sum()
```

```
name              0
region            0
cases             0
cases100k         0
cases7days        0
cases7days100k    0
cases24h          0
deaths            0
deaths100k        0
deaths7days       0
deaths7days100k   0
deaths24h         0
fatality_rate     0
population        0
dtype: int64
```

+ Code + Markdown

9]:

```
data.fillna({'cases':data['cases'].mean(),'cases100k':data['cases100k'].mean(),'cases7days':data['cases7days'].mean(),
            'cases7days100k':data['cases7days100k'].mean(),'deaths':data['deaths'].mean(),
            'deaths100k':data['deaths100k'].mean(),'deaths7days':data['deaths7days'].mean(),'deaths7days100k':data['deaths7days100k'].mean(),
            'fatality_rate':data['fatality_rate'].mean(),'population':data['population'].mean()}, inplace=True)
```

```
data
```

0]:

	name	region	cases	cases100k	cases7days	cases7days100k	cases24h	deaths	deaths100k	deaths7days	deaths7days100k	deaths24h	fatality_rate	population
0	Curaçao	Americas	3.169396e+07	9575.14	511165.000000	154.430000	81768.0	509219.0	153.84	8915.0	2.69	1369.0	1.606675	3.310026e+08
1	Cambodia	Western Pacific	6.257987e+05	1778.68	8652.800866	7.810000	15384.0	325772.0	23.61	1294.0	0.09	196.0	1.861513	2.870170e+07
2	American Samoa	Western Pacific	1.557257e+07	7326.22	84982.000000	39.980000	12950.0	433777.0	204.07	2343.0	1.10	384.0	2.785520	2.125594e+08
3	Ecuador	Americas	5.823452e+06	8578.28	173229.000000	255.170000	28754.0	99453.0	146.50	546.0	0.80	88.0	1.707801	6.788601e+07
4	North Macedonia	Europe	5.585316e+06	3827.27	131581.000000	67.942189	19719.0	155230.0	106.37	4591.0	3.15	677.0	2.779252	1.459347e+08
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
231	Fiji	Western Pacific	0.000000e+00	0.00	0.000000	0.000000	0.0	0.0	0.00	0.0	0.00	0.0	1.861513	2.870170e+07
232	Slovenia	Europe	0.000000e+00	0.00	0.000000	0.000000	0.0	0.0	0.00	0.0	0.00	0.0	1.861513	2.870170e+07
233	Bahamas	Americas	0.000000e+00	0.00	0.000000	0.000000	0.0	0.0	0.00	0.0	0.00	0.0	1.861513	2.870170e+07
234	Colombia	Americas	0.000000e+00	0.00	0.000000	0.000000	0.0	0.0	0.00	0.0	0.00	0.0	1.861513	2.870170e+07
235	New Caledonia	Western Pacific	0.000000e+00	0.00	0.000000	0.000000	0.0	0.0	0.00	0.0	0.00	0.0	1.861513	2.870170e+07

236 rows × 14 columns

2) Checking which countries have suffered the most:

- This can be checked by making sure that the cases in the last 24 hours are greater than cases in the last 7 days.
- From out analysis we find that only 3 countries have had an outbreak recently

```
#task2 3 COUNTRIES HAVE AN OUTBREAK
data[['name','cases24h','cases7days']][data['cases24h']>(data['cases7days']*0.5)]
```

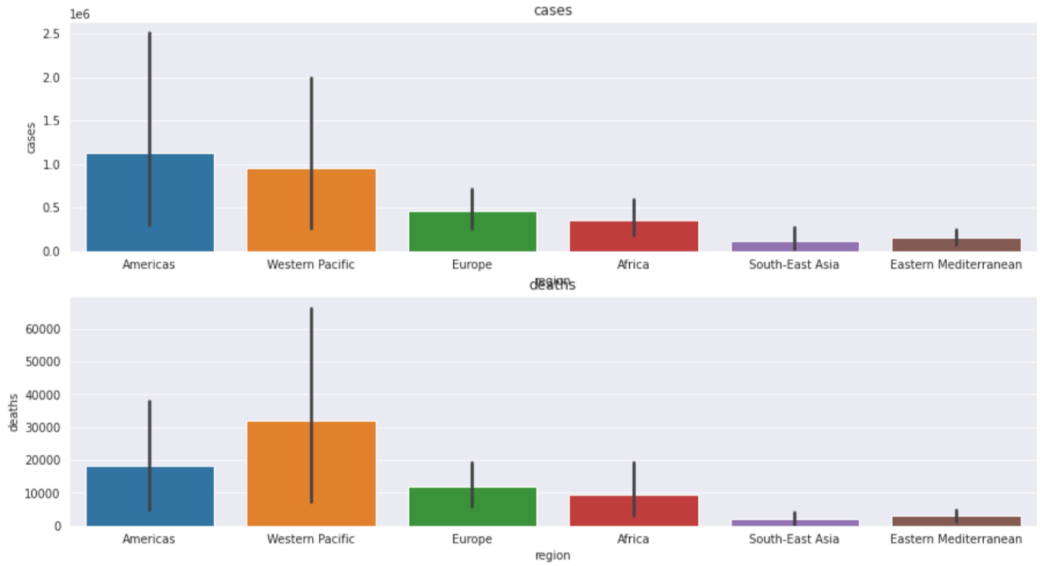
	name	cases24h	cases7days
1	Cambodia	15384.0	8652.800866
202	Falkland Islands (Malvinas)	125.0	216.000000
209	Chile	33.0	65.000000

3) Checking if Europe is affected worse than America.

- We check which country is hit worse by comparing the number of cases and the number of deaths. We can see from the graph the both the number of cases and number of deaths in America is larger then in Europ Thus I believe that the statement is **False.**

```
#task3 hypothesis testing EUROPE HAS MORE CASES THEN AMERICA
fig,ax =plt.subplots(2,1,figsize=(15,8))
sns.barplot(x=data['region'],y=data['cases'],ax=ax[0]).set_title("cases")
print("\n")
sns.barplot(x=data['region'],y=data['deaths'],ax=ax[1]).set_title("deaths")
```

```
Text(0.5, 1.0, 'deaths')
```



THE END