Data Wrangling with Pandas Assignment problems

Problem-1

he European Centre for Disease Prevention and Control (ECDC) provides an open dataset on COVID-19 cases called, daily number of new reported cases of COVID-19 by country worldwide. This dataset is updated daily, but we will use a snapshot that contains data from January 1, 2020 through September 18, 2020. Clean and pivot the data so that it is in wide format: (Get covid19_cases.csv file using this link:https://raw.githubusercontent.com/svkarthik86/Advanced-python/main/WEEK-8%20Assignment/covid19_cases.csv)

Read in the covid19_cases.csv file. Create a date column using the data in the dateRep column and the pd.to_datetime() function. Set the date column as the index and sort the index. Replace occurrences of United_States_of_America and United_Kingdom with USA and UK, respectively. Using the countriesAndTerritories column, filter the data down to Argentina, Brazil, China, Colombia, India, Italy, Mexico, Peru, Russia, Spain, Turkey, the UK, and the USA. Pivot the data so that the index contains the dates, the columns contain the country names, and the values are the case counts in the cases column. Be sure to fill in NaN values with 0.

In [1]:

```
import pandas as pd
url="https://raw.githubusercontent.com/svkarthik86/Advanced-python/main/WEEK-8%20Assignment/covid19_cases.csv"
covid_data=pd.read_csv(url)
covid_data["date"]=pd.to_datetime(covid_data[["year","month","day"]])
covid_data_pivot=covid_data.pivot(index="date",columns="countriesAndTerritories",values="cases").fillna(0)
covid_data_pivot.columns=covid_data_pivot.columns.str.replace("United_States_of_America","USA")
covid_data_pivot.columns=covid_data_pivot.columns.str.replace("United_Kingdom","UK")
covid_data_pivot.filter(["Argentina", "Brazil", "China", "Colombia", "India", "Italy", "Mexico", "Peru", "Russia", "Spain", "Turkey", "UK", "USA"])
```

Out[1]:

countriesAndTerritories	Argentina	Brazil	China	Colombia	India	Italy	Mexico	Peru	Russia	Spain	Turkey	UK	USA
date													
2020-01-01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020-01-02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020-01-03	0.0	0.0	17.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020-01-04	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020-01-05	0.0	0.0	15.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2020-09-14	10778.0	14768.0	29.0	7355.0	92071.0	1456.0	4408.0	6787.0	5449.0	27404.0	1527.0	3330.0	33871.0
2020-09-15	9056.0	15155.0	22.0	5573.0	83809.0	1008.0	3335.0	4241.0	5509.0	9437.0	1716.0	2621.0	34841.0
2020-09-16	9908.0	36653.0	24.0	6698.0	90123.0	1229.0	4771.0	4160.0	5529.0	11193.0	1742.0	3103.0	51473.0
2020-09-17	11893.0	36820.0	7.0	7787.0	97894.0	1452.0	4444.0	6380.0	5670.0	11291.0	1771.0	3991.0	24598.0
2020-09-18	11674.0	36303.0	44.0	7568.0	96424.0	1583.0	3182.0	5698.0	5762.0	14389.0	1648.0	3395.0	43567.0

262 rows × 13 columns

Problem-2

In order to determine the case totals per country efficiently, we need the aggregation skills , so the ECDC data in the covid19_cases.csv file has been aggregated for us and saved in the covid19_total_cases.csv file. It contains the total number of case per country. Use this data to find the 20 countries with the largest COVID-19 case totals. Hints: (Get covid19_total_cases.csv file using this link: https://raw.githubusercontent.com/svkarthik86/Advanced-python/main/WEEK-8%20Assignment/covid19_total_cases.csv)

When reading in the CSV file, pass in index_col='index' Note that it will be helpful to transpose the data before isolating the countries.

In [3]:

```
covid_cases=pd.read_csv("https://raw.githubusercontent.com/svkarthik86/Advanced-python/main/WEEK-8%20Assignment/covid19_total_cases.csv",index_col="index").T

covid_cases.sort_values(by="cases", ascending=False)[:20]
```

Out[3]:

index	cases
USA	6724667
India	5308014
Brazil	4495183
Russia	1091186
Peru	756412
Colombia	750471
Mexico	688954
South_Africa	657627
Spain	640040
Argentina	601700
Chile	442827
France	428696
Iran	416198
UK	385936
Bangladesh	345805
Saudi_Arabia	328720
Iraq	311690
Pakistan	305031
Turkey	299810
Italy	294932

In [ ]: