WEEK8 Assignment-1:Data Wrangling with Pandas

#letslearnstolearn

Data Wrangling with Pandas Assignment problems

Problem-1

The 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. Download the latest version of the dataset from January 1, 2020 to December 31, 2020 in CSV format. Save this file in a code folder (but DON'T submit it).

Visit https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

```
countriesAndTerritories  Argentina  Brazil  China  Colombia  India  Iraq  Mexico  Peru  Russia  Spain  Turkey  UK  USA
dateRep                                                                                                                
2020-12-26               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
2020-12-27               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
2020-12-28               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
2020-12-29               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
2020-12-30               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
2020-12-31               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

[366 rows x 12 columns]
```

1.Read in the covid19_cases.csv file.

```
In [ ]: import pandas as pd
        covid19_cases = pd.read_csv("https://opendata.ecdc.europa.eu/covid19/casedistribution/csv/data.csv")
        covid19_cases.to_csv("covid19_cases.csv")
        covid19_cases.head()
```

```
Out[ ]:
   dateRep  day  month  year  cases  deaths  countriesAndTerritories  geoId  countryterritoryCode  popData2019  continentExp  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
0  31/12/2020   31      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
1  30/12/2020   30      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
2  29/12/2020   29      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
3  28/12/2020   28      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
4  27/12/2020   27      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
```

```
In [ ]: covid19_cases.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59443 entries, 0 to 59442
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   dateRep                 59443 non-null  object 
 1   day                     59443 non-null  int64  
 2   month                   59443 non-null  int64  
 3   year                    59443 non-null  int64  
 4   cases                   59443 non-null  int64  
 5   deaths                  59443 non-null  int64  
 6   countriesAndTerritories 59443 non-null  object 
 7   geoId                   59183 non-null  object 
 8   countryterritoryCode    59308 non-null  object 
 9   popData2019             59308 non-null  float64
 10  continentExp            59443 non-null  object 
 11  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000  58885 non-null  float64
dtypes: float64(2), int64(5), object(5)
memory usage: 5.4 MB
```

Create a date column and set the index to the date/day column and the pct_change() function.

```
In [ ]: covid19_cases['dateRep'] = pd.to_datetime(covid19_cases['dateRep'])
        covid19_cases.head()
```

```
Out[ ]:
   dateRep  day  month  year  cases  deaths  countriesAndTerritories  geoId  countryterritoryCode  popData2019  continentExp  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
0 2020-12-31   31      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
1 2020-12-30   30      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
2 2020-12-29   29      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
3 2020-12-28   28      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
4 2020-12-27   27      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
```

```
In [ ]: covid19_cases.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59443 entries, 0 to 59442
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype         
---  ------                  --------------  -----         
 0   dateRep                 59443 non-null  datetime64[ns]
 1   day                     59443 non-null  int64         
 2   month                   59443 non-null  int64         
 3   year                    59443 non-null  int64         
 4   cases                   59443 non-null  int64         
 5   deaths                  59443 non-null  int64         
 6   countriesAndTerritories 59443 non-null  object        
 7   geoId                   59183 non-null  object        
 8   countryterritoryCode    59308 non-null  object        
 9   popData2019             59308 non-null  float64       
 10  continentExp            59443 non-null  object        
 11  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000  58885 non-null  float64       
dtypes: datetime64[ns](1), float64(2), int64(5), object(4)
memory usage: 5.4 MB
```

```
In [ ]: covid19_cases.set_index('dateRep', inplace=True)
        covid19_cases.head()
```

```
Out[ ]:
            day  month  year  cases  deaths  countriesAndTerritories  geoId  countryterritoryCode  popData2019  continentExp  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
dateRep                                                                                                                                                                                    
2020-12-31   31      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
2020-12-30   30      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
2020-12-29   29      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
2020-12-28   28      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
2020-12-27   27      12  2020      0       0              Afghanistan     AF                   AFG    38041754.0          Asia                                                  0.141
```

```
In [ ]: covid19_cases.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 59443 entries, 2020-12-31 to 2020-01-01
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   day                     59443 non-null  int64  
 1   month                   59443 non-null  int64  
 2   year                    59443 non-null  int64  
 3   cases                   59443 non-null  int64  
 4   deaths                  59443 non-null  int64  
 5   countriesAndTerritories 59443 non-null  object 
 6   geoId                   59183 non-null  object 
 7   countryterritoryCode    59308 non-null  object 
 8   popData2019             59308 non-null  float64
 9   continentExp            59443 non-null  object 
 10  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000  58885 non-null  float64
dtypes: float64(2), int64(5), object(4)
memory usage: 5.4 MB
```

Using the countriesAndTerritories column, filter the data down to Argentina, Brazil, China, Colombia, India, Iraq, Mexico, Peru, Russia, Spain, Turkey, the UK, and the USA.

```
In [ ]: countries = ['Argentina', 'Brazil', 'China', 'Colombia', 'India', 'Iraq', 'Mexico', 'Peru', 'Russia', 'Spain', 'Turkey', 'United_Kingdom', 'United_States_of_America']
        covid19_cases = covid19_cases[covid19_cases['countriesAndTerritories'].isin(countries)]
        covid19_cases.head()
```

```
Out[ ]:
            day  month  year  cases  deaths countriesAndTerritories  geoId  countryterritoryCode  popData2019  continentExp  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
dateRep                                                                                                                                                                                   
2020-12-31   31      12  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-12-30   30      12  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-12-29   29      12  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-12-28   28      12  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-12-27   27      12  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
```

```
In [ ]: covid19_cases.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 4270 entries, 2020-12-31 to 2020-01-01
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   day                     4270 non-null   int64  
 1   month                   4270 non-null   int64  
 2   year                    4270 non-null   int64  
 3   cases                   4270 non-null   int64  
 4   deaths                  4270 non-null   int64  
 5   countriesAndTerritories 4270 non-null   object 
 6   geoId                   4270 non-null   object 
 7   countryterritoryCode    4270 non-null   object 
 8   popData2019             4270 non-null   float64
 9   continentExp            4270 non-null   object 
 10  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000  4270 non-null   float64
dtypes: float64(2), int64(5), object(4)
memory usage: 400.3+ KB
```

See the date column as the index and sort the index.

```
In [ ]: covid19_cases.sort_index(inplace=True)
        covid19_cases.head()
```

```
Out[ ]:
            day  month  year  cases  deaths countriesAndTerritories  geoId  countryterritoryCode  popData2019  continentExp  Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
dateRep                                                                                                                                                                                   
2020-01-01    1       1  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-01-02    2       1  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-01-03    3       1  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-01-04    4       1  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
2020-01-05    5       1  2020      0       0                   Spain     ES                   ESP    46937060.0        Europe                                                 0.000
```

Find the dates 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 [ ]: covid19_cases = covid19_cases.pivot(columns='countriesAndTerritories', values='cases').fillna(0)
        covid19_cases.head()
```

```
Out[ ]:
countriesAndTerritories  Argentina  Brazil  China  Colombia  India  Iraq  Mexico  Peru  Russia  Spain  Turkey  United_Kingdom  United_States_of_America
dateRep                                                                                                                                                      
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    0.0       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    0.0       0.0    0.0   0.0     0.0   0.0     0.0    0.0     0.0             0.0                       0.0
```

```
In [ ]: covid19_cases.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 366 entries, 2020-01-01 to 2020-12-31
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Argentina                366 non-null    float64
 1   Brazil                   366 non-null    float64
 2   China                    366 non-null    float64
 3   Colombia                 366 non-null    float64
 4   India                    366 non-null    float64
 5   Iraq                     366 non-null    float64
 6   Mexico                   366 non-null    float64
 7   Peru                     366 non-null    float64
 8   Russia                   366 non-null    float64
 9   Spain                    366 non-null    float64
 10  Turkey                   366 non-null    float64
 11  United_Kingdom           366 non-null    float64
 12  United_States_of_America 366 non-null    float64
dtypes: float64(13)
memory usage: 40.0 KB
```

Plot the data set.

```
In [ ]: covid19_cases.plot(figsize=(15,10))
```

```
Out[ ]: <AxesSubplot:xlabel="dateRep">
```

Problem-2

In order to determine the case totals per country efficiently we need the aggregation table. As 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 cases per country. Read in the covid19_total_cases.csv file into DataFrame. Note: Encoding='utf8' does not work for this file.

https://raw.githubusercontent.com/prasertcbs/basic-dataset/master/covid19/COVID-19%20Dataset%20-%20covid19_total_cases.csv

When loading in the CSV file, pass in index_col='Unnamed: 0' Note that it will be helpful to manipulate the data before loading the countries.

```
In [ ]: pd.read_csv("https://raw.githubusercontent.com/prasertcbs/basic-dataset/master/covid19/COVID-19%20Dataset%20-%20covid19_total_cases.csv", index_col='Unnamed: 0')
```

```
Out[ ]:
                         cases
Unnamed: 0                    
US                    19216052
India                 10021107
Brazil                 7491160
Russia                 2871421
France                 2627647
Spain                  2167520
Argentina              1827557
Italy                  1813744
Colombia               1662971
United Kingdom         1647899
Mexico                 1440691
Peru                   1000078
Iran                    978195
South Africa            827827
Poland                  818602
Ukraine                 734274
Netherlands             646219
Chile                   630227
Belgium                 614303
Israel                  529722
Germany                 521838
Canada                  521677
Philippines             478852
Indonesia               475698
Turkey                  471782
Sweden                  456902
Czechia                 456702
Pakistan                455274
Romania                 441035
Portugal                426999
Iraq                    418977
Switzerland             401116
Japan                   380276
Hungary                 378118
Bangladesh              377972
Saudi Arabia            370133
Austria                 368222
Nepal                   267010
Dominican Republic      265131
Panama                  263305
Kazakhstan              260284
Ecuador                 258146
Serbia                  256140
Bolivia                 249439
Kuwait                  245194
Morocco                 243342
Ireland                 240136
Moldova                 239124
Honduras                237391
Oman                    236124
Guatemala               234183
Georgia                 232118
Belarus                 229139
Costa Rica              227135
Lebanon                 225911
Azerbaijan              225772
Libya                   224922
Tunisia                 223124
Slovakia                221124
Bulgaria                220124
Croatia                 219124
Uzbekistan              218124
Slovenia                217124
Lithuania               216124
Paraguay                215124
Jordan                  214124
Cuba                    213124
Greece                  212124
Egypt                   211124
Nigeria                 210124
Venezuela               209124
Afghanistan             208124
Algeria                 207124
Kyrgyzstan              206124
Uruguay                 205124
Mongolia                204124
Armenia                 203124
Estonia                 202124
Latvia                  201124
Malaysia                200124
Cyprus                  199124
Luxembourg              198124
Maldives                197124
Qatar                   196124
Bahrain                 195124
Finland                 194124
Norway                  193124
Denmark                 192124
Australia               191124
New Zealand             190124
Singapore               189124
Thailand                188124
Vietnam                 187124
Sri Lanka               186124
Myanmar                 185124
Kenya                   184124
Ethiopia                183124
Ghana                   182124
Cameroon                181124
Ivory Coast             180124
Sudan                   179124
Zambia                  178124
Zimbabwe                177124
Uganda                  176124
Tanzania                175124
Senegal                 174124
Rwanda                  173124
Mozambique              172124
Namibia                 171124
Botswana                170124
Malawi                  169124
Angola                  168124
Mauritius               167124
Eswatini                166124
Gabon                   165124
Guinea                  164124
Mali                    163124
Burkina Faso            162124
Niger                   161124
Chad                    160124
Somalia                 159124
Haiti                   158124
Fiji                    157124
Papua New Guinea        156124
Timor-Leste             155124
Laos                    154124
Cambodia                153124
Brunei                  152124
Diamond Princess        151124
MS Zaandam              150124
Taiwan                  149124
South Korea             148124
Vanuatu                 147124
Micronesia              146124
Kiribati                145124
Samoa                   144124
Tonga                   143124
Solomon Islands         142124
Western Sahara          141124
Yemen                   140124
Syria                   139124
Libya                   138124
Sierra Leone            137124
Liberia                 136124
Guinea-Bissau           135124
Gambia                  134124
Togo                    133124
Benin                   132124
Mauritania              131124
Comoros                 130124
Lesotho                 129124
Burundi                 128124
South Sudan             127124
Eritrea                 126124
Central African Republic 125124
Sao Tome and Principe   124124
```

```
In [ ]: covid19_total_cases = pd.read_csv("https://raw.githubusercontent.com/prasertcbs/basic-dataset/master/covid19/COVID-19%20Dataset%20-%20covid19_total_cases.csv", index_col='Unnamed: 0')
        covid19_total_cases.head()
```

```
Out[ ]:
               cases
Unnamed: 0          
US          19216052
India       10021107
Brazil       7491160
Russia       2871421
France       2627647
```

```
In [ ]: covid19_total_cases.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 215 entries, US to Sao Tome and Principe
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   cases   215 non-null    int64
dtypes: int64(1)
memory usage: 3.4+ KB
```