Name: Aruna Balasiva
Email: aruna.316@gmail.com
Course: Certificate in Introductory Data Analytics
GitHub Link: https://github.com/ArunaAR/UCDPA_Aruna

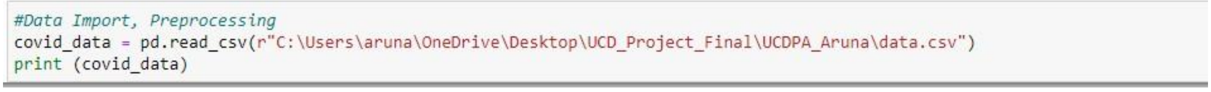
Abstract

Covid-19 ranks among the most lethal infectious diseases of the 21st century. For this assignment, I conducted an analysis using a dataset comprising daily counts of newly reported Covid-19 cases and deaths across EU countries. The dataset, sourced from the European Centre for Disease Prevention and Control (ECDC), is periodically updated between Monday and Wednesday. The analysis utilizes data spanning from January 2021 to September 2021.

Data Import and Pre-processing:

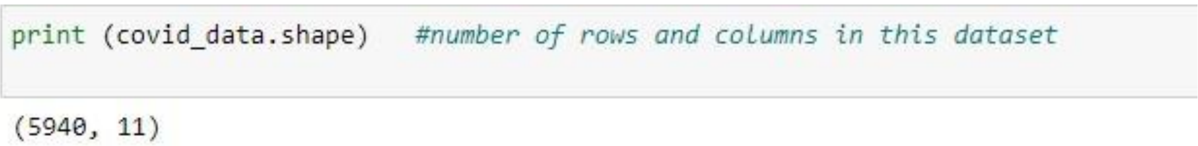
For this assignment, PyCharm was used for coding and testing, while Jupyter Notebook was utilized to display the dataset output, including tables, graphs, and visualizations.

The initial step involved importing the dataset. There is one CSV file, data.csv, which was obtained from ECDC



```
#Data Import, Preprocessing
covid_data = pd.read_csv(r"C:\Users\aruna\OneDrive\Desktop\UCD_Project_Final\UCDPA_Aruna\data.csv")
print (covid_data)
```

I examined the dataset and found that it contains 5,940 rows and 11 columns.

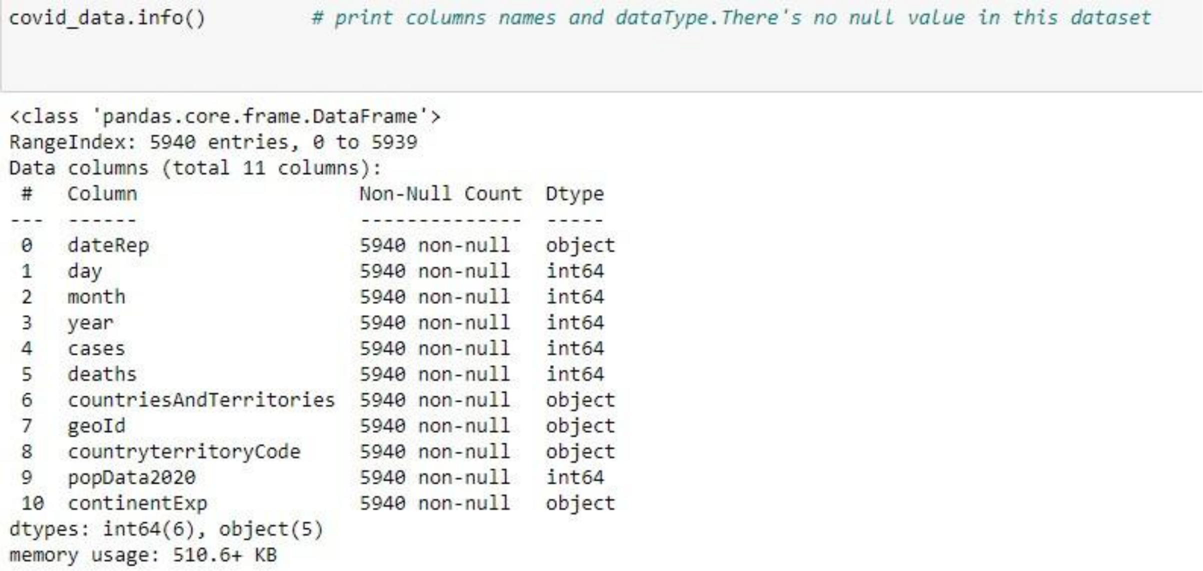


```
print (covid_data.shape)   #number of rows and columns in this dataset

(5940, 11)
```

Next, I examine the available data types in this dataset using the `.info()` method. This step is essential to determine the data type of each column and to identify any null values (NaN) present within the dataset.

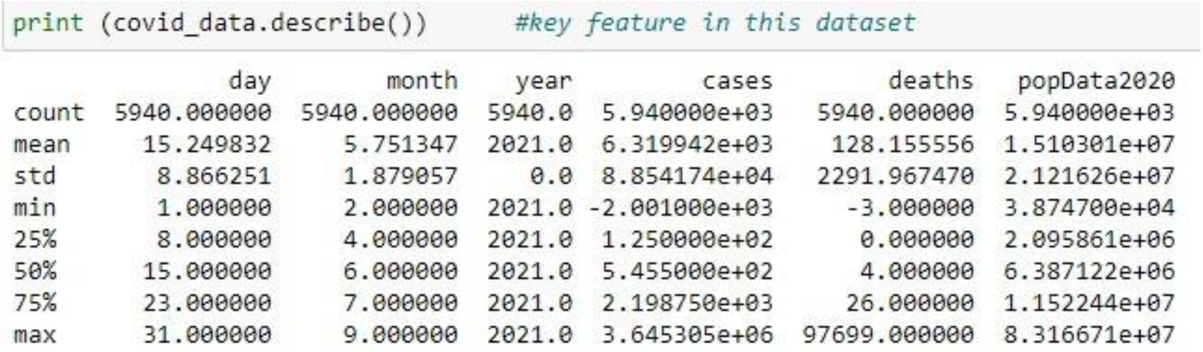
Based on the output, it can be determined that this dataset contains null values.



```
covid_data.info()        # print columns names and dataType.There's no null value in this dataset

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5940 entries, 0 to 5939
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   dateRep                  5940 non-null   object
 1   day                      5940 non-null   int64 
 2   month                    5940 non-null   int64 
 3   year                     5940 non-null   int64 
 4   cases                    5940 non-null   int64 
 5   deaths                   5940 non-null   int64 
 6   countriesAndTerritories  5940 non-null   object
 7   geoId                    5940 non-null   object
 8   countryterritoryCode     5940 non-null   object
 9   popData2020              5940 non-null   int64 
 10  continentExp             5940 non-null   object
dtypes: int64(6), object(5)
memory usage: 510.6+ KB
```

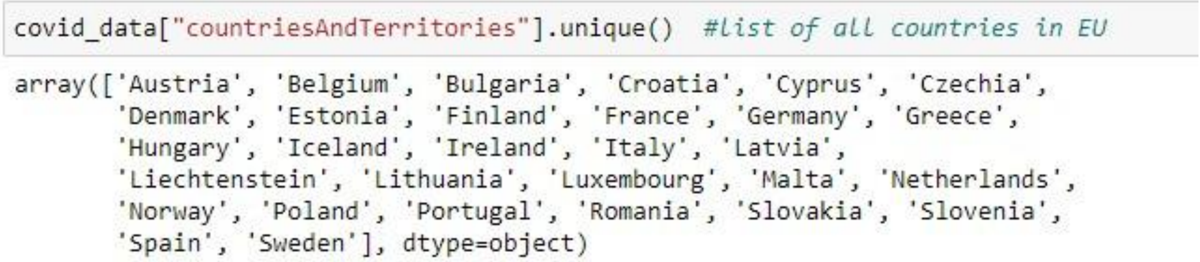
Using .describe() to display the core stats for the entire table.



```
print (covid_data.describe())      #key feature in this dataset
```

```
               day        month    year         cases        deaths   popData2020
count  5940.000000  5940.000000  5940.0  5.940000e+03   5940.000000  5.940000e+03
mean     15.249832     5.751347  2021.0  6.319942e+03    128.155556  1.510301e+07
std       8.866251     1.879057     0.0  8.854174e+04   2291.967470  2.121626e+07
min       1.000000     2.000000  2021.0 -2.001000e+03     -3.000000  3.874700e+04
25%       8.000000     4.000000  2021.0  1.250000e+02      0.000000  2.095861e+06
50%      15.000000     6.000000  2021.0  5.455000e+02      4.000000  6.387122e+06
75%      23.000000     7.000000  2021.0  2.198750e+03     26.000000  1.152244e+07
max      31.000000     9.000000  2021.0  3.645305e+06  97699.000000  8.316671e+07
```

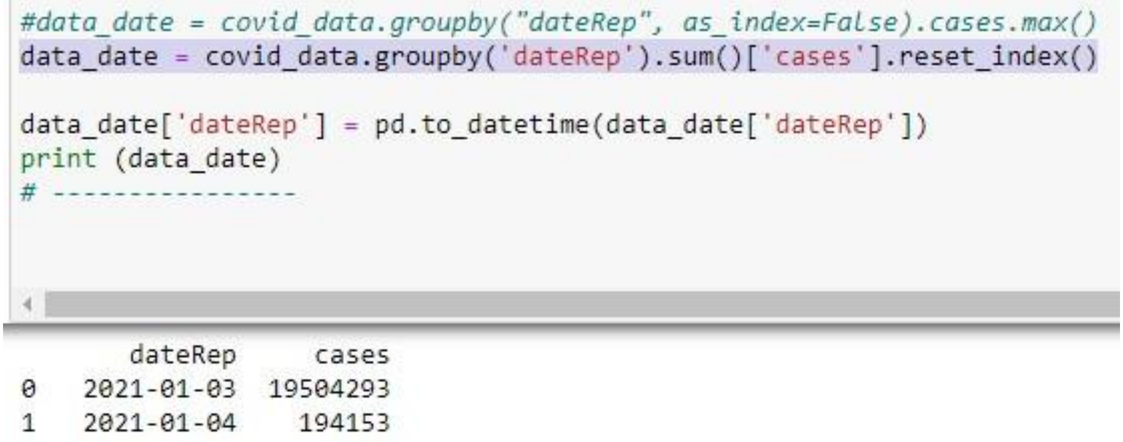
This dataset contains a list of EU countries:



```
covid_data["countriesAndTerritories"].unique()  #list of all countries in EU
```

```
array(['Austria', 'Belgium', 'Bulgaria', 'Croatia', 'Cyprus', 'Czechia',
       'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece',
       'Hungary', 'Iceland', 'Ireland', 'Italy', 'Latvia',
       'Liechtenstein', 'Lithuania', 'Luxembourg', 'Malta', 'Netherlands',
       'Norway', 'Poland', 'Portugal', 'Romania', 'Slovakia', 'Slovenia',
       'Spain', 'Sweden'], dtype=object)
```

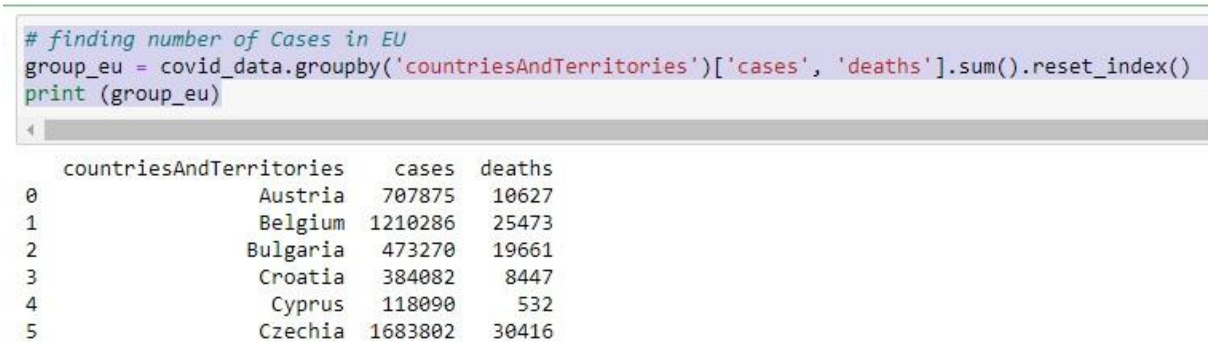
I primarily utilized the GroupBy function in Pandas, which enables efficient division of data into distinct groups.



```
#data_date = covid_data.groupby("dateRep", as_index=False).cases.max()
data_date = covid_data.groupby('dateRep').sum()['cases'].reset_index()

data_date['dateRep'] = pd.to_datetime(data_date['dateRep'])
print (data_date)
# ----------------
```

```
     dateRep     cases
0 2021-01-03  19504293
1 2021-01-04    194153
```



```
# finding number of Cases in EU
group_eu = covid_data.groupby('countriesAndTerritories')['cases', 'deaths'].sum().reset_index()
print (group_eu)
```

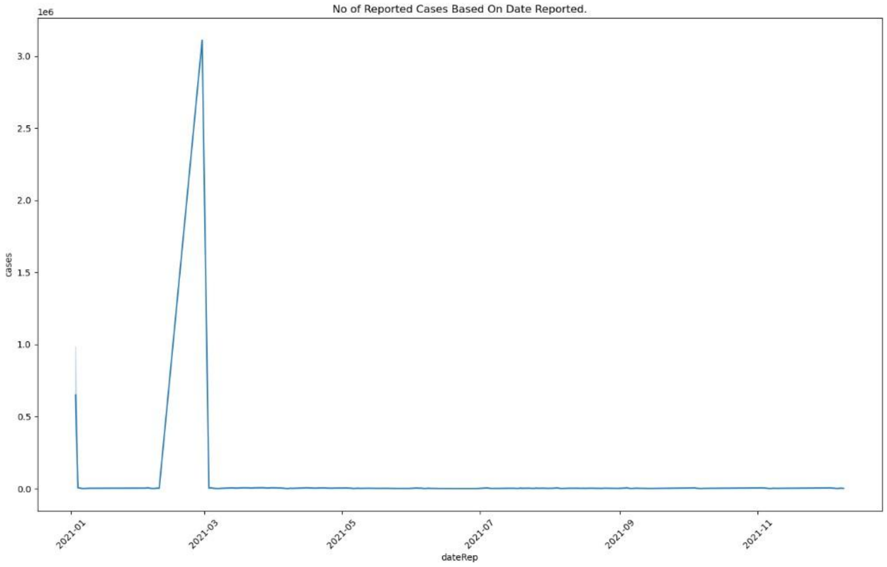
```
  countriesAndTerritories    cases  deaths
0                 Austria   707875   10627
1                 Belgium  1210286   25473
2                Bulgaria   473270   19661
3                 Croatia   384082    8447
4                  Cyprus   118090     532
5                 Czechia  1683802   30416
```

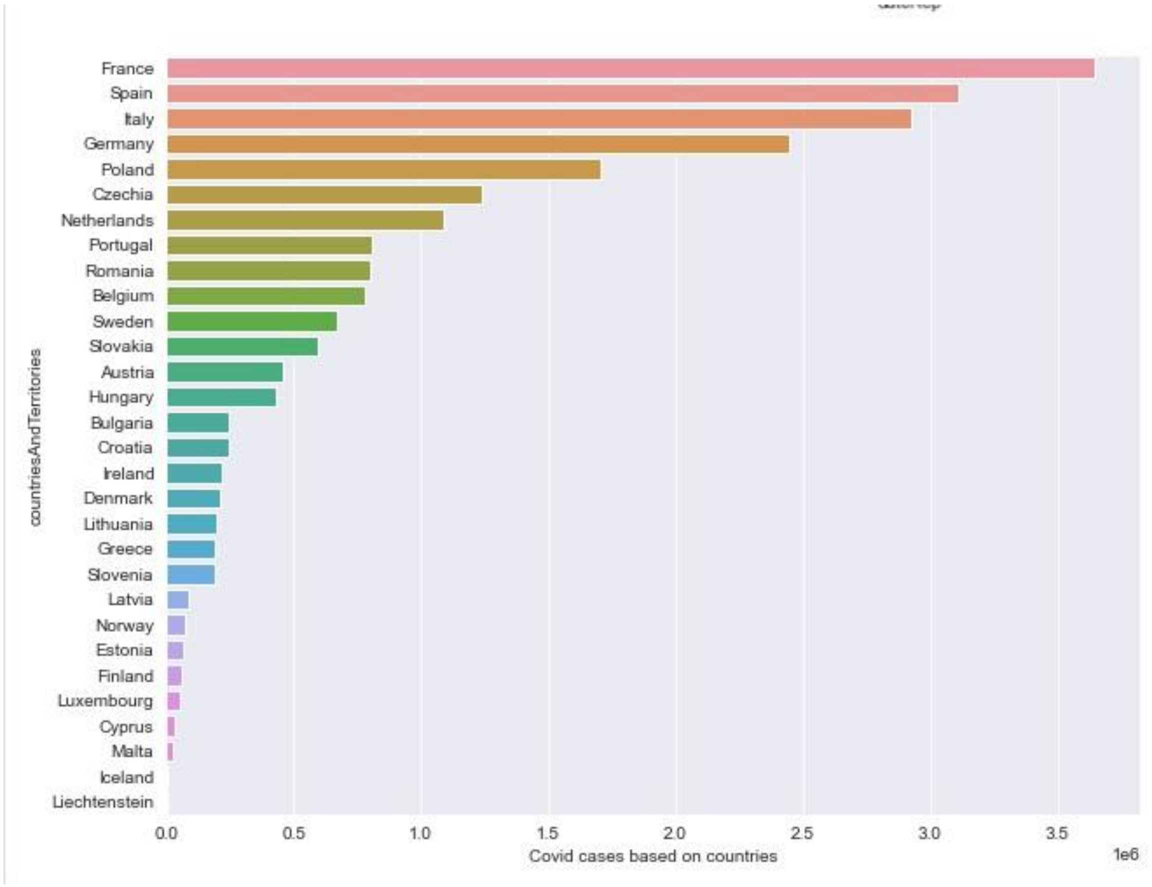
For data visualization, I utilized Matplotlib, Seaborn, and Plotly to develop bar graphs, map visualizations, and conduct trend analysis.

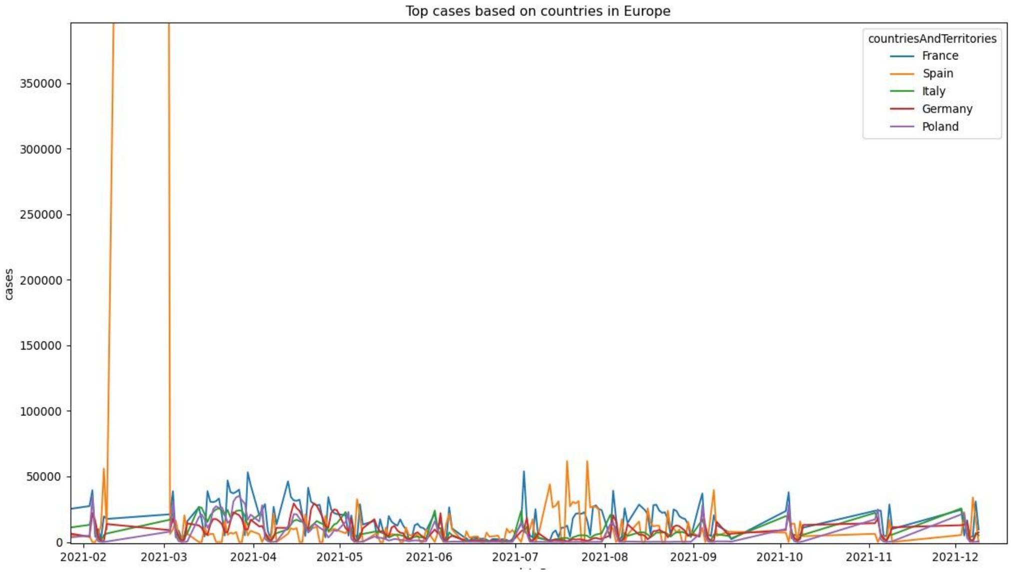
Results based on the dataset:

Cases Reported in EU

According to the dataset, the majority of cases were reported during the early part of the year.





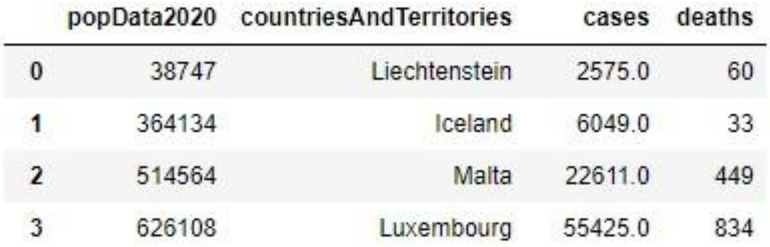


Insights

Covid-19 cases increased in the EU from January to March. According to the data, France, Spain, Italy, Germany, and Poland reported the highest numbers of cases among member countries.

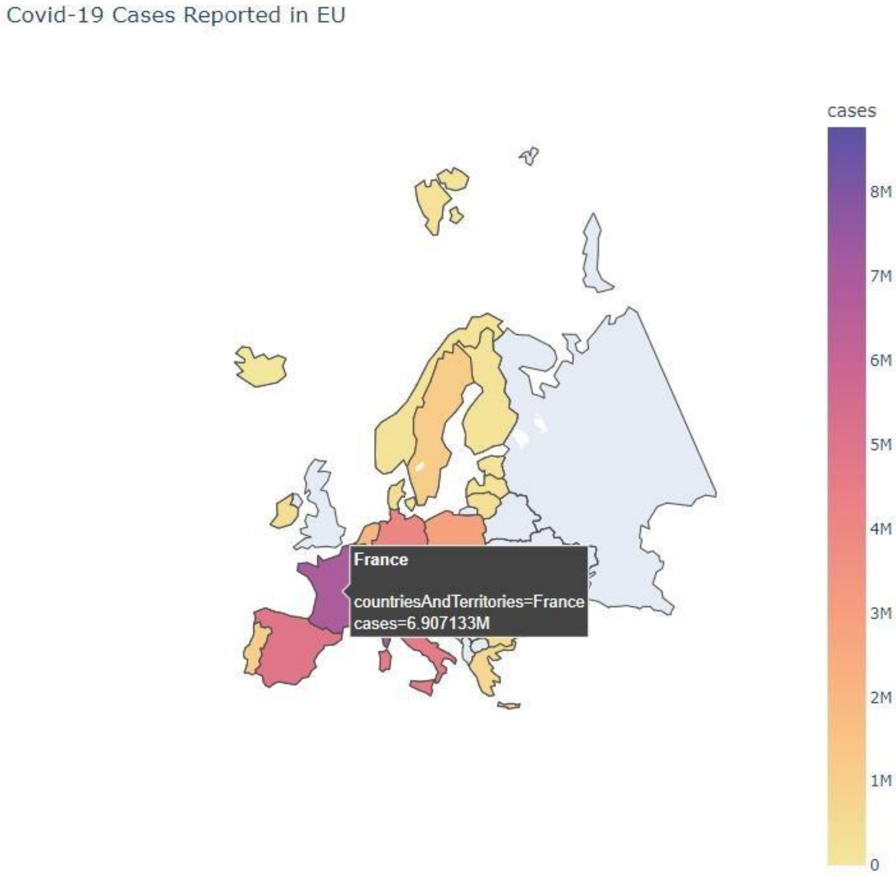
Insights

Although Liechtenstein, Iceland, Malta, and Luxembourg have comparatively smaller populations, the data indicates that these countries reported a relatively high number of Covid-19 cases.



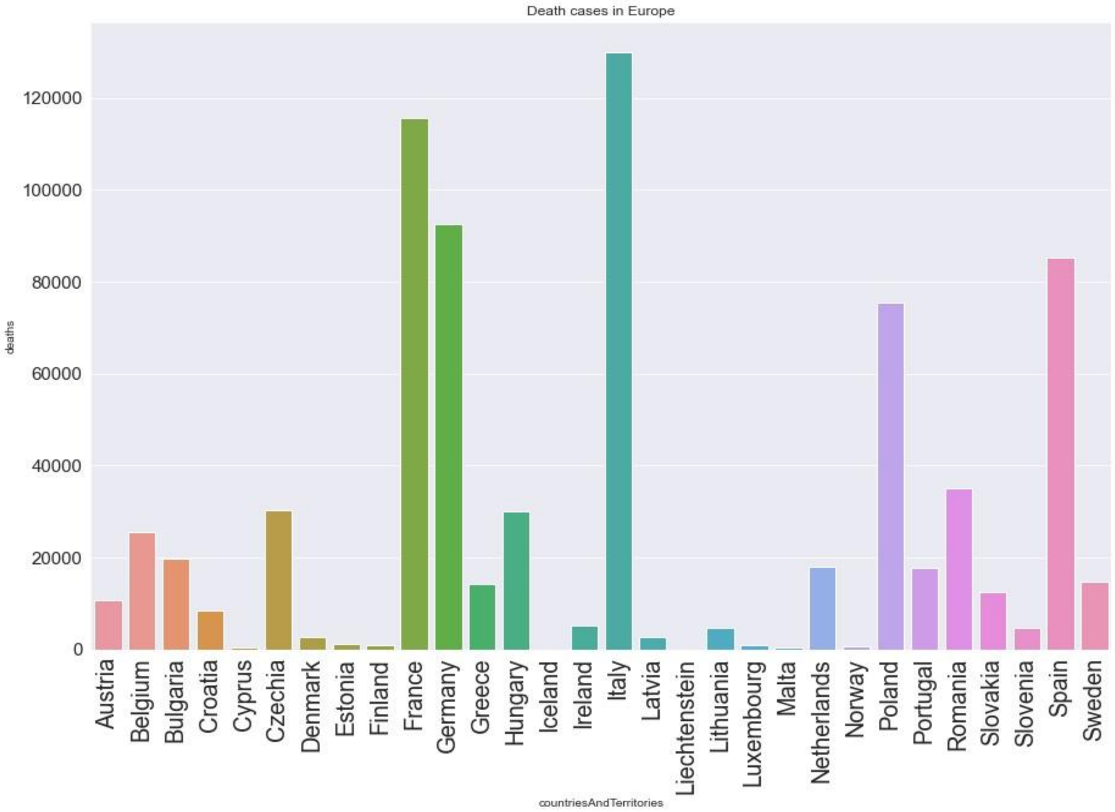
	popData2020	countriesAndTerritories	cases	deaths
0	38747	Liechtenstein	2575.0	60
1	364134	Iceland	6049.0	33
2	514564	Malta	22611.0	449
3	626108	Luxembourg	55425.0	834

A location-based visualization for the EU was created to represent the distribution of cases. For instance, when France is highlighted on the map, it indicates that over 6.9 million cases were reported between January 2021 and August 2021.



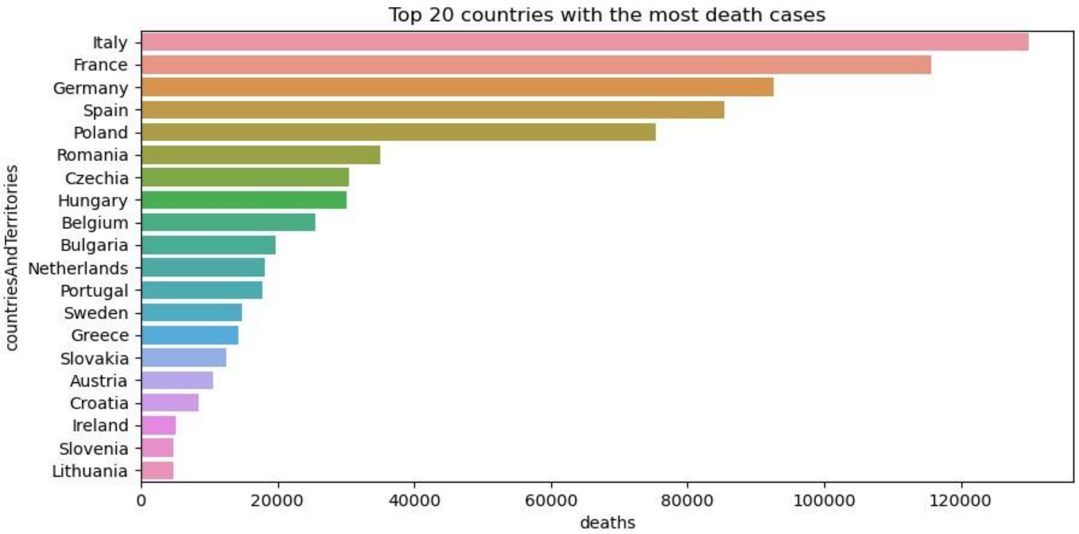
Deaths reported in EU.

The dataset indicated that Italy reported over 120,000 fatalities, with France, Germany, and Spain also registering significant numbers of deaths.

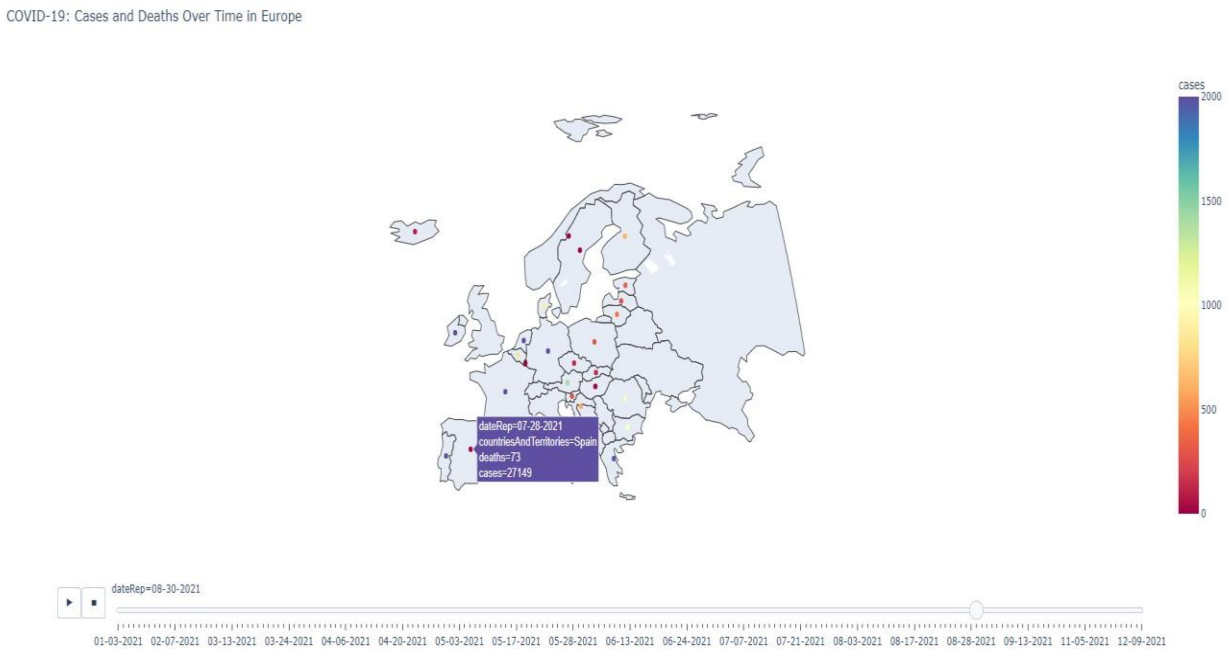


Insights

Here are the results for the top 20 death rates in the EU from January 2021 to September 2021. The data indicates that Ireland, Slovenia, and Lithuania have fewer deaths compared to other EU countries during this period.



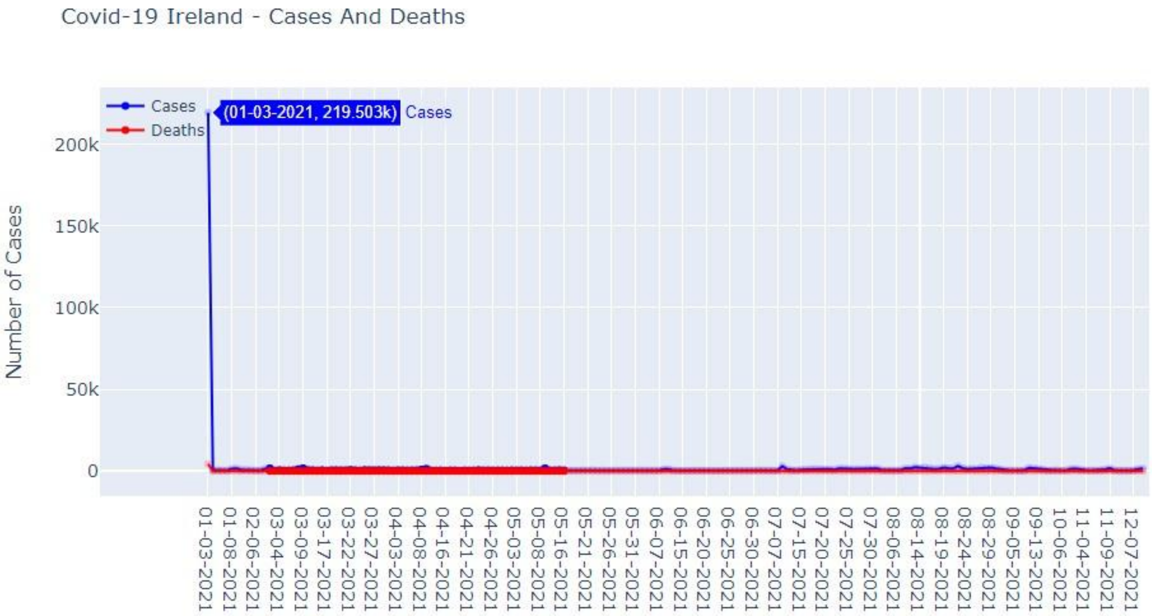
I created a time series to display the number of cases and deaths by reported date for all EU countries. For example, hovering over Spain on 28 July 2021 showed 27,149 reported cases and 73 deaths.



COVID-19: Cases and Deaths Over Time in Europe

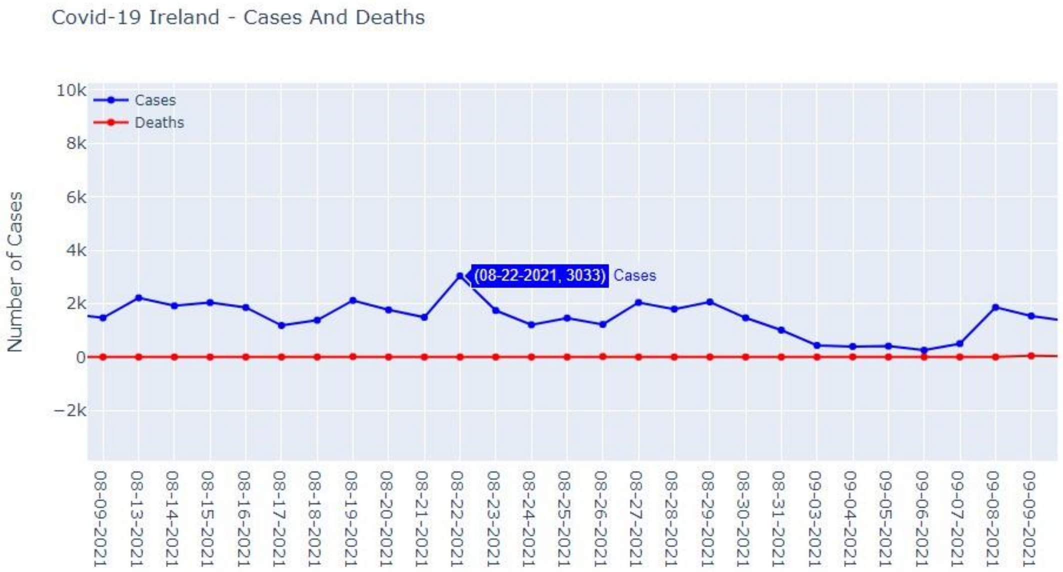
dateRep=07-28-2021
countriesAndTerritories=Spain
deaths=73
cases=27149

Ireland: Cases and Deaths

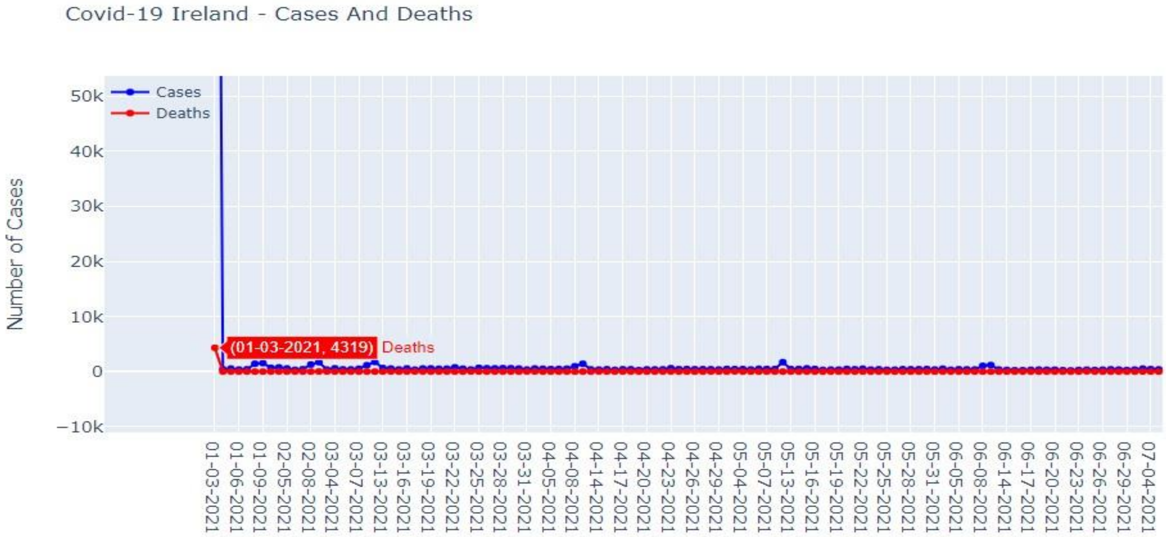


Insights

In January, there were around 219,000+ reported cases in Ireland. The number of cases gradually decreased from February to July but increased slightly in August (22 August 2021) with 3,033 cases.



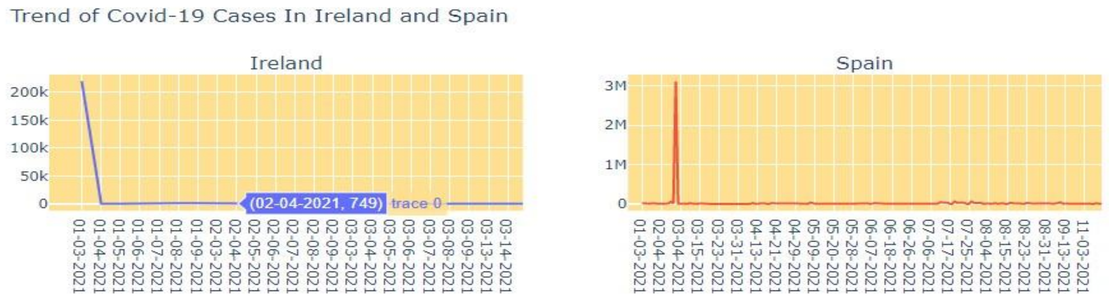
The highest death toll was recorded in January with a total of 4,319 deaths, followed by a significant decline thereafter.



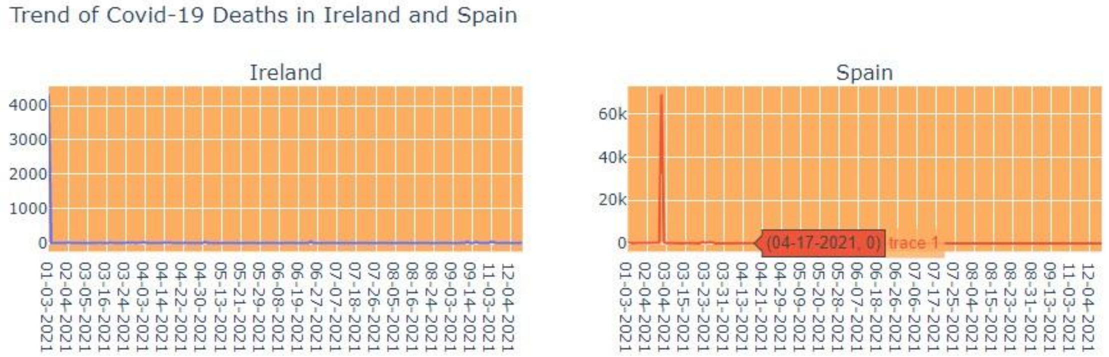
Comparing Cases and Deaths: Ireland and Spain Insights

Since Spain was one of the top five countries in terms of cases, I created a trend analysis to compare the data for cases and deaths with Ireland.

Based on the trend analysis for cases, Ireland recorded its highest number of cases in January, while Spain's cases spiked to around 3 million at the end of February.



For deaths, Ireland recorded the highest number in January, while Spain had over 60,000 deaths by the end of February.



References:

https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country

Download Date : 14-09-2021

https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-inpython/#describing-data-with-describe