Georgios Karyotis, geo.kariotis@gmail.com

A picture containing clipart

Description automatically generated

Data Analytics PROJECT FOR the Certificate IN Data ANaLytics- November 2020 Class

COVID-19 and Economic Complexity Data Analysis

# GitHub URL

https://github.com/gpkaryotis/UCDPA\_Karyotis

# Abstract

COVID-19 pandemic has affected everyone’s life during last and current year. In this project, data related to COVID-19 will be investigated to i) find out which continents have been affected more from COVID-19 pandemic, ii) find out possible relationships between COVID-19 total cases and total deaths and economic complexity, iii) see possible differences of the COVID-19 pandemic waves in Ireland and in Greece.

# Introduction

Analysing data and make useful observations based on this data have been made easier with the current state of technology and tools. Many sources of valid data exist in the world web, i.e. kaggle.com, and many languages like R and Python support many features to transform, analyse and visualize the data. In this project, an effort has been made to make insightful observations about the effects that COVID-19 pandemic has across the different continents and compare the Ireland’s and Greece’s experience.

# Dataset

In this project, two data sources were used, the following table has some description of the datasets.

|  |  |  |
| --- | --- | --- |
| **Table 1: Project’s Data Sets Descriptions** | | |
| **Data** | **Source** | **Description** |
| COVID-19 | https://github.com/owid/covid-19-data/tree/master/public/data | This COVID-19 related data is gathered by the “Our World In Data” organization and there are many details there and for all the countries. Name of the file: "owid-covid-data.csv" |
| Economic Complexity | [https://atlas.cid.harvard.edu/rankings](about:blank) | The website [https://atlas.cid.harvard.edu/](about:blank) has reliable data about the economic complexity of several countries, in total 133. The economic complexity ranking is interesting because we can investigate how effective countries have been on fighting the COVID-19 pandemic based on their rank on economic complexity. Name of the file:  "Country Complexity Rankings 1995 - 2018.csv" |

# Implementation Process

The project was implemented in Python and in PyCharm Python IDE. The main program is in the Python script “ucd\_covid\_karyotis\_project.py”. The key elements of the implementation process were the following: i) develop functions to read the csv files, ii) transform, merge, and clean the data, iii) visualize the data to find insights and improve the plots. The main packages used at this project were the following i) pandas, ii) matplotlib.pyplot, iii) seaborn. Next the benefits of the used python packages will be described, and the main elements of the Python script.

|  |  |
| --- | --- |
| **Table 2: Python Packages’ Feature Use** | |
| **Package** | **Main Package elements in ucd\_covid\_karyotis\_project.py** |
| pandas | Used in i) reading csv files either from local source or from web, ii) create DataFrames, iii) keeping only specific columns from the initial DataFrames, iv) merging two different DataFrames via an inner-join merge, v) group the DataFrames based on the continent location and selection the total cases and total deaths at a specific date. |
| matplotlib.pyplot | Used to plot bar chats and time-series chats |
| seaborn | Used to develop a heatmap diagram to investigate the relationship among economic complexity and COVID-19 response effectiveness |

The developed python script has three main code parts:

1. Create the necessary DataFrames by reading two csv files, one file is being read from the web location https://covid.ourworldindata.org/data/owid-covid-data.csv, the other has been already downloaded,
2. Clean and keep only the necessary columns from the previous stage DataFrames,
3. Visualize the DataFrames to make useful observations.

Next the second step will be described more, as the second step is the one where the key data transformations occur.

|  |  |
| --- | --- |
| **Table 3: Columns kept from the “owid-covid-data.csv” file at DataFrame dt\_covid** | |
| **Kept Column** | **Reason** |
| location | It has the country name, the column name is changed to Country |
| continent | It has the Continent location of the country |
| date | The date of the data |
| new\_cases\_per\_million | The daily new COVID-19 cases per million at the country |
| new\_deaths\_per\_million | The daily new COVID-19 deaths per million at the country |
| total\_cases\_per\_million | The total COVID-19 cases per million at the country until the date specified in the data column |
| total\_deaths\_per\_million | The total COVID-19 deaths per million at the country until the date specified in the data column |

|  |  |
| --- | --- |
| **Table 4: Columns kept from the "Country Complexity Rankings 1995 - 2018.csv" file at DataFrame dt\_comp\_econ** | |
| **Kept Column** | **Reason** |
| Country | It has the country name, and it is used to merge the Economic Complexity DataFrame with the COVID DataFrame |
| ECI Rank 2018 | It holds the rank of Economic Complexity Index for the year 2018. There are no rankings for the year 2019 or 2020. |

After the appropriate columns were kept then the two DataFrames were merged on the common column “Country”, and the merge was an inner-join. It is important to mention that the “dt\_comp\_econ” DataFrame had fewer countries and not all the countries had the same name in the two DataFrames, thus the produced DataFrame had fewer rows than the initial DataFrames.   
 The next table has the main DataFrames used in the script for visualization with a basic description.

|  |  |
| --- | --- |
| **Table 5: Description of the process which created the major Pandas variables  and other Data Structures of ucd\_covid\_karyotis\_project.py script** | |
| **Variable** | **Process** |
| dt\_covid | Read the relevant csv file, it is a Pandas DataFrame. |
| dt\_compl\_econ | Read the relevant csv file, it is a Pandas DataFrame. |
| acc\_dt\_covid | Kept the rows of dt\_covid which have date value “29-01-2021”. It is a Pandas DataFrame. |
| time\_series\_dt\_covid\_gr\_irl | Kept all the rows which either have “Greece” or “Ireland” as their value at the “Country” column of the dt\_covid DataFrame. It is a Pandas DataFrame. |
| grouped\_by\_continent\_total\_cases | Groups the acc\_dt\_covid rows based on the continent column value and assigns the mean value of total\_cases\_per\_million of each country to that continent. It is a Pandas DataFrame |
| grouped\_by\_continent\_total\_deaths | Groups the acc\_dt\_covid rows based on the continent column and assigns the mean value of total\_deaths\_per\_million of each country to that continent. It is a Pandas DataFrame |
| acc\_dt\_covid\_w\_complex\_econ | Merged the acc\_dt\_covid and dt\_compl\_econ on the Country column. It has also a new column 'Complexity\_Quarter', which organizes each country into 4 possible Quarters of Economic Complexity, the countries with the most complex economies are in the Q1 quarter. It is a Pandas DataFrame. |
| covid\_deaths\_complexity\_group | Groups the acc\_dt\_covid\_w\_complex\_econ rows based on the 'Complexity\_Quarter' column and assigns the mean value of total\_deaths\_per\_million of each country to that Quarter. It is a Pandas DataFrame |
| covid\_deaths\_complexity\_crosstab | A crosstab structure which is created based on acc\_dt\_covid\_w\_complex\_econ DataFrame. The crosstab’s rows group the countries based on their “Complexity\_Quality” rank, its columns group the countries based on their “continent”. The values of the crosstab are the normalised mean total deaths per million. It is a Pandas crosstab structure. |

|  |  |
| --- | --- |
| **Table 6: Purpose of the major variables of ucd\_covid\_karyotis\_project.py script** | |
| **Variable** | **Key Purpose** |
| dt\_covid | Holds the key COVID-19 data |
| dt\_compl\_econ | Holds the key Economic Complexity Index data |
| acc\_dt\_covid | Holds the data of all the countries for only one date so the total numbers of cases and deaths being processed easily for up to that date |
| time\_series\_dt\_covid\_gr\_irl | This DataFrame has only the COVID-19 data for Ireland and Greece to plot the daily cases and deaths in a time-series diagram. |
| grouped\_by\_continent\_total\_cases | Used to plot a bar diagram which shows which continent’s mean total cases per million of COVID-19 cases |
| grouped\_by\_continent\_total\_deaths | Used to plot a bar diagram which shows which continent’s mean total deaths per million of COVID-19 cases |
| acc\_dt\_covid\_w\_complex\_econ | An intermediate data structure which is the base for the next data structures covid\_deaths\_complexity\_group,  covid\_deaths\_complexity\_crosstab |
| covid\_deaths\_complexity\_group | It is used to plot a bar plot which shows the mean total cases per million of COVID-19 cases for countries organized into 4 Economic Complexity Quarter. |
| covid\_deaths\_complexity\_crosstab | Used to plot a heatmap where one axis is the Economic Complexity Rank, the other is the Continent location of each country and the presented values are the normalized mean total COVID-19 related deaths per million. |

The next section of this report has the visualizations, majority of the visualizations use the matplotlib library, the heatmap plot was developed by using the seaborn library.

# Results

In this project five key plots were created and are displayed below.

Figure 1: Mean of Total COVID-19 Cases at each Continent

Chart, bar chart

Description automatically generated

Figure 2: Mean of Total COVID-19 Deaths at each Continent.

Chart, bar chart

Description automatically generated

Figure 3: Mean of Total COVID-19 Deaths at each Quarter of Economic Complexity Country RankChart, bar chart

Description automatically generated

Figure 4: Normalized Mean of Total COVID-19 Deaths per Quarter of Economic Complexity Rank and Continent

# Chart, treemap chart Description automatically generated

Figure 5: Time Series of Daily New Cases and Deaths for Ireland and Greece

Graphical user interface, application

Description automatically generated

# Insights

Some of the key insights from the above diagrams are:

* European countries were affected more from the COVID-19 pandemic than countries from the rest continents as the first two diagrams show,
* The diagram of mean total deaths of countries grouped by Economic Complexity Quarter rank shows that the countries with most complex economies had more deaths, and the least economic complex countries have fewer deaths,
* Based on heatmap of the normalized mean of the total deaths the most affected countries were countries which have the most complex economies and are in North America or in Europe. This insight is counter intuitive as someone could expect that countries with complex economies are advanced economies with good strategies and health care systems thus more capable to fight a health pandemic,
* Greece had low COVID-19 cases and deaths during the first wave, but the second COVID-19 wave affected the people more, as there have been more cases and deaths during the second wave,
* Greece does not seem to have a third COVID-19 wave yet,
* Ireland managed better the second COVID-19 wave than the first one and the third one,
* Ireland’s third COVID-19 wave is the hardest wave for the Irish people.