**Advanced Python Programming**

**Winter Semester 2021 – 2022**

**Computational Assignment**

**Dr. Dionisios Sotiropoulos**

**Covid-19 Data Analysis**

This computational assignment focuses on the development of a wide range of data preprocessing, visualization and modelling routines that aim at extracting meaningful insights concerning the COVID-19 pandemic. Initially, you should download the original dataset from <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset> keeping only the following .csv files:

1. **time\_series\_covid\_19\_confirmed.csv:** storing the cumulative number of confirmed cases by date for each Province/State.
2. **time\_series\_covid\_19\_deaths.csv:** storing the cumulative number of deaths by date for each Province/State.
3. **time\_series\_covid\_19\_recovered.csv:** storing the cumulative number of recovered cases by date for each Province/State.

Your task is to develop a Python project that complies with the following implementation directives:

1. The given time series data sets contain Province/State – related information that should be aggregated at the Country/Region level. Aggregating the available data by Country/Region can be accomplished by summing the confirmed cases, deaths and recovered cases over each distinct Country/Region and subsequently averaging the corresponding latitudes and longitudes.
2. Report the unique Country/Regions idenfied.
3. Report the complete time span of the given dataset by providing the corresponding dates.
4. Provide a distinct world map-based visualization of the confirmed cases, deaths and recovered cases for the first and last dates in the dataset. This visualization could be achieved by plotting on a world map object the individual geolocation of each country where the size and/or color of each data point can be scaled according to the corresponding number of the daily volume of confirmed cases, deaths or recovered cases. Focusing on the time series of confirmed cases, the required graphs could be of the following form:

 

Apparently, 3 pairs of graphs should be provided. For better visualization results you could consider plotting only Country/Regions – related datapoints for which the corresponding daily cumulative volume exceeds a given threshold. Thus, you need to define 3 threshold variables, namely: **confirmed\_cases\_thres**, **deaths\_thres** and **recovered\_cases\_thres**. Moreover, better visualization results could be attained by providing a different visualization for each continent (e.g. North America, South America, Europe, Asia, Africa, Australia)

1. Provide three animated visualizations that combine all the intermediate figures of the previous plots for the confirmed cases, deaths and recovered cases. You could, for example, generate three .gif images each of which would be composed from the individual graphs for each distinct date.
2. Provide a Country/Region – based ranking according the total number of confirmed cases, deaths and recovered cases for the top 10 Countries – Regions in the dataset. Visualize the obtained results. The corresponding chart for the total number of confirmed cases could be of the following form:



1. Compute the number of **new** confirmed cases, **new** deaths and **new** recovered cases for the top 10 Countries/Regions. Provide a graphical representation of the acquired results. The corresponding graph for the number of new confirmed cases could be of the following form:



Based on the number of new confirmed cases per day, develop an algorithm that can detect the number of distinct covid waves and their duration for each one of the top 10 Countries/Regions. Provide a graphical representation of your results.