INSTRUCTIONS

- **Question 1 (1 Mark)**

  Perform the following preprocessing steps:

  - Drop all rows from the "exposure" dataset without country name
  - Join the two datasets (exposure.csv and Countires.csv) based on the "country" columns in the datasets, keeping the rows as long as there is a match between the country columns of both dataset (do not concatenate the datasets)
    You must ensure the countries with name issues match (e.g., USA and United States) but ignore if a country does not exist in either of datasets (e.g., Sudan and "South Sudan" are not the same)
  - Keep only a single country column
  - Set the index of the resultant dataframe as 'Country'
  - Sort the dataset by the index (ascending)

- **Question 2: (based on the dataframe created in Question-1) (1 Mark)**
  The "Cities" column is a complex string, containing information about cities (e.g., latitude and longitude) of the corresponding country; you should **explore** the content of this column for each country and ***add two new columns*** to the dataframe called: avg_latitude and avg_longitude ."avg_latitude" is the average latitude for all cities of the corresponding country, and "avg_longitude" is the average longitude for all cities of the same country.

- **Question 3: (based on the dataframe created in Question-2) (1 Mark)**
  Given that the first case of COVID has been found in Wuhan with the following coordinates: (30.5928° N, 114.3055° E), sort the dataframe based on how close they are to Wuhan . You can use "avg_latitude" and "avg_longitude"; the countries close to Wuhan should be ranked first in the final dataset. The final dataset should also contain ***a new column*** called distance_to_Wuhan , showing the distance to Wuhan in km. Update: You can assume the earth's radios is R = 6373.

- **Question 4: (based on the dataframe created in Question-2) (1 Mark)**
  Using the continent dataset **(Countries-Continets.csv)** calculate the average covid_19_Economic_exposure_index for each continent. The output should be a dataframe with two columns: " Continent ", " average_covid_19_Economic_exposure_index ". Rank the continents based on average "Covid_19_Economic_exposure_index" (ascending), with "Continent" being the index.

- **Question 5: (based on the dataframe created in Question-2) (1 Mark)**
  What is the the average "Foreign direct investment" and "Net_ODA_received_perc_of_GNI" for each income class (e.g., HIC, MIC, LIC). The output should be a dataframe with three columns: " Income Class ", " Avg Foreign direct investment ", " Avg_ Net_ODA_received_perc_of_GNI ", with "Income Class" being the index. If a country has no value for the given column, ignore the row. UPDATE: the column name changed by adding "Avg"

- **Question 6: (based on the dataframe created in Question-2) (1 Marks)**
  List top 5 most **Populous** cities located in Low Income Countries; ignore cities without population information. The output is a **python list** , containing the names of the cities.

- **Question 7: (based on the dataframe created in Question-2) (2 Marks)**
  Find cities which are located in different countries but have the same name. The result dataset should contain 2 columns: " city ", " countries ", with city being also the index. The result dataset should not have duplicate records. For instance, either "Melbourne, {Florida, Australia}" or "Melbourne, {Australia, Florida}" should be present in the final dataframe, not both. The countries column is a list of countries separated by comma (',').

- **Question 8: (based on the dataframe created in Question-2) (2 Marks)**
  - In a visualization show what percentage of the world population is living in each South American country. You can use the continent dataset **(Countries-Continets.csv)** to answer this question.
  - Choose an appropriate visualization type, presenting the requested information in the best way (check the lecture for Data Visualization about selecting the right paradigm)
  - Plot human-readable visualization; it should be self-explanatory and its elements (e.g., labels, legends) must be clear.

- **Question 9 : (based on the dataframe created in Question-2) (2 Marks)**
  - Plot a visualization to compare the high, middle, and low income level countries based on the following metrics:
  Covid_19_Economic_exposure_index_Ex_aid_and_FDI Covid_19_Economic_exposure_index_Ex_aid_and_FDI_and_food_import
  Foreign direct investment, net inflows percent of GDP
  Foreign direct investment
  - Choose an appropriate visualization type, presenting the requested information in the best way.
  - Plot human-readable visualization; it should be self-explanatory and its elements (e.g., labels, legends) must be clear.

- **Question 10: (based on the dataframe created in Question-2) (3 Marks)**
  - Plot a scatter plot with y axis being "avg_latitude" and x axis being "avg_longitude". Each point in this plot indicates a labelled country.
  - Ink each country (e.g., red, green) based on its continent (e.g, Asia, Africa). You can pick any colour for each continent. You can use the continent dataset **(Countries-Continets.csv)** to answer this question.
  - The size of each point must represent the population of its country. For example, the points representing China and India should be bigger than that of Australia.
  - Add a legend showing the name of continents and their associated colours.
  - Plot human-readable visualization; it should be self-explanatory and its elements (e.g., labels, legends) must be clear.