Question 1

Refer to these Stata DTA file: ***Countries dependencies area***. This is a list of the world's countries and their dependencies by land, water, and total area. For more details, visit the Wikipedia page

- Calculate the percentage of water in relation to the total land area measured in kilometers:

$$\text{Water } \% = \left(\frac{\text{Total Area in km}^2 - \text{Land Area in km}^2}{\text{Total Area in km}^2} \right) \times 100 \tag{1}$$

- Calculate the percentage of water in relation to the total land area measured in miles:

$$\text{Water } \% = \left(\frac{\text{Total Area in miles}^2 - \text{Land Area in miles}^2}{\text{Total Area in miles}^2} \right) \times 100 \tag{2}$$

- Find out which 10 countries have the largest land (in km^2) with water in the world.
- Find out which 10 countries have the water (%) in the world.

Question 2

Refer to the ***ISO 3166 Countries Codes CSV file***. This file contains official state names, sovereignty status—indicating if a country is UN member, disputed, or under another state's control—and TLD (Top-Level Domain), which is the Internet domain suffix for each country (e.g., .pk for Pakistan). For more details, visit the Wikipedia page.

- Read the CSV and assign the Name `sovereign_UN_memb_df` to this Dataframe.
- **Clean the Data**: Remove any unnecessary notes, references, or sources attached to the values in any column. For example remove the [e] from **Antarctica**[e] or [c][d] from **Åland**[c][d]
- **Filter UN Member Countries**: Eliminate any countries from the dataset that are not UN members. You can determine UN membership by checking the "Sovereignty Status" column.
- **Reset index** Remember, there are currently 193 sovereign states that are members of the United Nations.Ensure that the index is reset properly.

Question 3

Please refer to the ***world bank country groups CSV file***. This file contains the The World Bank Group classification of the world's economies, classified by income, region, and World Bank lending status. Visit the interactive version on the World Bank Country Groups

- Read the CSV and assign the Name `wb_groups_df` to this Dataframe.
- **Merge Datasets**: Combine `wb_groups_df` with the `sovereign_UN_memb_df`.
- Name the merged dataframe as `countries_df`.
- Include "merge indicator" column to show the status of each row from both datasets.
- Consider using the column **ISO Alpha3 Code** while merging the data frames.
- Verify that all 193 UN member countries are matched by cross-tabulating the "merge indicator" and "sovereignty status" columns.
- **Drop and Reset**: Keep only the rows that are present in both datasets, resulting in a total of 193 rows. Drop the "merge indicator" and "Code" columns. Ensure that the index is reset properly.

Question 4

Please refer to the ***internet speedtest global index CSV file***. The Speedtest Global Index includes internet speed Ranking of mobile and fixed broadband speeds from around the world. For country wise Speedtest Intelligence, you can visit the Speedtest Global Index by Ookla.

- Read the CSV and assign the Name `ookla_df` to this Dataframe.
- **Merge Datasets**: Combine `ookla_df` with the `countries_df`.
- Name the merged dataframe as `speedtest_index_df`.
- Consider using the column **ISO Alpha3 Code** while merging the data frames.
- Find out which countries have the fastest, and slowest internet speeds in the world the in terms of Fixed Broadband speed
- Find out which countries have the fastest, and slowest internet speeds in the world the in terms of Mobile speed
- Compute the minimum and maximum Fixed Broadband speed Mbps for each 1) region, 2) income group?
- Compute the minimum and maximum Mobile Speed Mbps for each 1) region, 2) income group?
- Print top 25 countries based on Fixed Broadband Mbps.
- Print top 25 countries based on Mobile Speed Mbps.
- Filter the region as "South Asia", Print three columns country, Mobile , and Fixed Broadband Speeds
- Additional optional interactive resources to explore: Cloud Infrastructure Map around the World, Submarine Cable Map, and more. And of course, let's try not to get too distracted by where the sharks are dining on Pakistan's ocean cables!
- Can you write a few lines about the state of the internet, technology, and data? Reflect on where we stand in this relentless race.

Question 5

Please refer to the ***COLDAT Colonies CSV File***. This data is sourced from [Becker, 2019] and has been simplified to include indicator variables for major European colonizers: Belgium, Italy, Germany, France, Britain, the Netherlands, Spain, and Portugal. Note that a country may have multiple colonizers, while the column "last colonizer" indicates the single colonizer from which the country later gained independence.

- Convert the following columns to categorical type: Belgium, Italy, Germany, France, Britain, the Netherlands, Spain, Portugal, and "last colonizer."
- Determine the number of countries that remain colonies using the "last colonizer" column.
- Calculate the total number of colonies for each colonizer: Belgium, Italy, Germany, France, Britain, the Netherlands, Spain, and Portugal.
- Compare the number of colonies listed in the "last colonizer" column with those indicated by the individual colonizer columns. If there are differences, identify how many countries fall under both categories. For example, consider countries that were colonized by both Britain and France.

Question 6

Refer to these CSV files: ***1) Africa, 2) Americas, 3) Asia, 4) Europe, 5) Australia Oceania, 6) Transcontinental States*** . These files contain country names, the date of sovereignty acquisition, and descriptions of how sovereignty was acquired. For more details, visit the Wikipedia page. Each of these data frames has three same columns 1) Country Name, 2) Sovereignty Acquisition Date, 3) Sovereignty Acquisition Description.

- Read the CSV files and assign corresponding name such as: `Africa_df`, `Americas_df`, `Asia_df`, `Australia_Oceania_df`, `Europe_df`, and `Transcontinental_states_df`
- Create a column -"continent_name" in each of these 6 files, fill with appropriate name of the content.
- Combine all six DataFrames into a single unified DataFrame and name it `continents_df`.
- Ensure that the resulting DataFrame retains all relevant columns and number of rows from each continent.
- **Clean the Data**: Remove any unnecessary notes, references, or sources attached to the values in any column.
- Extract valid four-digit years starting with "15", "16", "17", "18", "19", or "20" from the `Sovereignty Acquisition Date` column and store them in a new column named `Year`.
- Fill any missing values in the `Year` column with the default value of 1400 and convert the column to a numeric data type, ensuring that any conversion errors result in `NaN`.
- This default is used because some countries may have acquisition dates that are earlier or in BC (Before Christ), as indicated by entries such as "600 BC", "218 BC", and "3 September 301". The value 1400 serves as a placeholder for these cases where specific years cannot be accurately extracted.
- Extract colonizer: Belgium, Italy, Germany, France, Britain, the Netherlands, Spain, Portugal, other using the column Sovereignty Acquisition Description.
- A country may have multiple entries for different dates of sovereignty acquisition, Therefore remove duplicate entries based on Country_Name, ensuring that you keep the row with the latest Year sovereignty acquisition for each country.
- Convert the colonizer columns to categorical type. Compare the number of colonies for each colonizer using this data alongside the COLDAT dataset. Don't worry if the numbers don't match exactly; the goal is to achieve a relatively comparable analysis.

References

Bastian Becker. Colonial Dates Dataset (COLDAT), 2019. URL https://doi.org/10.7910/DVN/T9SDEW.