Documentation

Problem definition and analysis

Analysis of shark attack incidents. Creating a program - a pipeline for data download, preparation, analysis, and visualization to get the top 10 countries by the number of shark attack incidents in a year selected by the user.

Errors/limitations: For the purposes of this assignment, only basic data cleaning was performed (all country names converted to upper case). Country duplicates have not been checked/corrected in case of differences in spelling. The number of incidents has not been normalized per country's population. The minimum year 0 might be caused by a missing value. If more countries have the same number of incidents, pandas prioritizes according to the order of appearance, so some countries might not be displayed in the plot even when having the same number of incidents as the displayed countries. The solution to the abovementioned issues is outside the scope of this exercise.

Input: Shark attacks data: A dataset freely available at the Global Shark Attack File webpage as an .xls file: https://www.sharkattackfile.net/incidentlog.htm. This data is regularly updated based on reported cases. To ensure up-to-date information, one of the requirements is that the program downloads the current version from the web before the analysis.

Output: A bar plot of the top 10 countries ordered from the highest to the lowest number of shark attack incidents, shown on the screen and saved as .png.

Algorithms/mathematical relationships:

Minimum_year in the dataset <= Year selected by the user <= Maximum_year in the dataset

Number of incidents = number of times each unique value in the "Country" variable appears after subsetting the data for a selected year.

Steps:

1. Preparation – downloading data, loading libraries
2. Reading the data
3. Cleaning and subsetting
4. Calculating numbers of incidents
5. Plotting the data and generating outcome

Design

PREPARATION:

- Load libraries
- Download the Shark attacks data
- Save the data to the folder

READING THE DATA:

- Read the data as a pandas DataFrame

CLEANING AND SUBSETTING:

- Unify the country names – convert all country names in the "Country" variable to upper case
- Convert the "Year" variable to integer
- Ask the user which year they are interested in (FUNCTION):
    - Get minimum and maximum of the "Year" variable to inform about the range of years available in the dataset
    - Print the range for the user
    - Ask the user to enter a year within this range (WHILE LOOP)
    - Create a subset of data according to the user's selection (CONDITION)

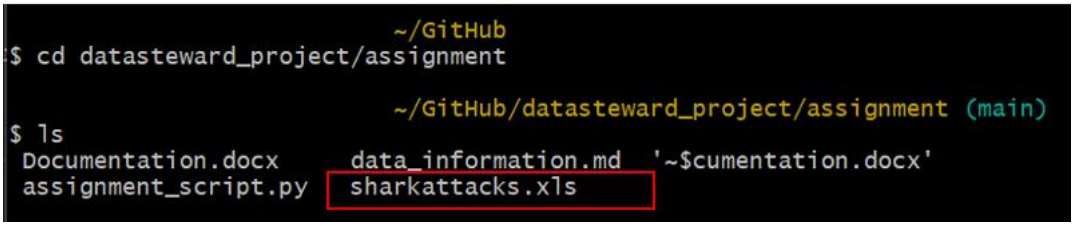
CALCULATING NUMBER OF INCIDENTS FOR THE SELECTED YEAR AND PLOTTING THE DATA (FUNCTION):

- From the subset of data, calculate the number of incidents in each country
- Select the top 10 countries by the number of incidents and order from highest to lowest
- Make a bar plot of countries with the number of incidents, ordered from highest to lowest
- Show the plot on the screen
- Save the plot as .png

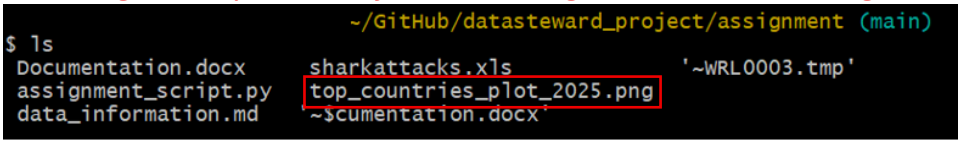
Verification / testing

Steps:

1. Downloading data: *Checking working directory:*



2. Reading the data: *Printing the first 6 rows of the dataset using the head() command.*
3. Cleaning and subsetting:
   3.1. Converting country names to upper case:
   Checking the list of unique values of the variable: print(df["variable_name"].unique())
   3.2. Converting the "Year" variable to integer: *Checking type: df["variable_name"].dtype)*
   3.3. Calculating the minimum and maximum value in the "Year" variable:
   Calculating "manually": print(df["variable_name"].min()), print(df["variable_name"].max())
   3.4. Testing loop for inserting the year value in which the user is interested in:
   Condition for inserted value:
   Minimum_year in the dataset <= Year selected by the user <= Maximum year in the dataset
   Testing the loop by inserting a number that violates the condition, i.e.:
   i) Year selected by the user < Minimum_year in the dataset
   ii) Year selected by the user > Maximum year in the dataset
   3.5. Creating a subset of data for the selected year
   Printing list of unique values of the "Year" variable in the selected subset:
   print(df_subset["variable_name"].unique())
4. Calculating numbers of incidents per country in selected countries:
   Printing top ten countries by number of incidents per country in a selected year:
   print(top_countries).
   Comparing the values with a pivot table in Excel, eg. for year 2025.
   For this, the subset df was saved as xls. Idealy, the original file with all values would be used, but because the data are not cleaned properly and to simplify, the subset was saved. See the test.xls for comparison of Python outcome and pivot table.
5. Plotting the data and generating outcome.
   Checking the output visually and checking the file in the working directory:



Implementation

See the source code in assignment_script.py file.

Declaration of AI use

I have used ChatGPT to generate the script for data download, to modify and consult scripts for functions which I have further modified and developed, and to consult errors and the meaning of the script. I have checked and understood all the script in the source code.

Link to GitHub repository

https://github.com/KatarinaRi/datasteward_project