# Data Wrangling Project: Sharks Attack Dataset

## Instructions

Welcome to the final project of this data wrangling module! In this project, you will get a chance to work through the entire data wrangling workflow while preparing the shark\_attacks.csv file for analysis. This dataset contains very dirty data and will require a lot of work! This project is broken down into key steps of the data wrangling process to help guide you along the process. When you are finished, save the wrangled dataset as a final\_project.csv file. Submit the final project as a zip folder named final\_project.zip. Make sure the zipped folder has both your wrangled dataset and this word document within it. Best of luck!

## Step 1: Decide which tool to use

This dataset contains around 1100 rows. Discuss which tool (BigQuery/Python/Google Sheets) is best suited for the data cleaning task for this dataset. Mention the relevant advantages and disadvantages of each tool. Finally, state which tool you think is best suited for the task and why. (6 marks)

|  |
| --- |
| I personally think Google Sheets is best suited for this task. The dataset is relatively small and requires a lot of manual cleaning, which is more suitable to be performed using Google Sheets as compared to BigQuery and Python.  Below are some of the advantages and disadvantages of all the 3 tools mentioned in the question:  **BigQuery**  Advantages: Useful for querying and extracting specific data from large datasets efficiently  Disadvantages: Since it’s a query tool, it may not provide a user-friendly interface for data cleaning tasks like we can’t really manually amend the values in each columns.  **Python**  Advantages: Suitable for performing automated data cleaning tasks and for handling large datasets efficiently.  Disadvantages: Unable to amend the information for each columns manually as per Google Sheets, every action needs to be written in code so it’s time consuming.  **Google Sheets**  Advantages: Easy data visualization and the ability to perform manual data cleaning tasks column by column.  Disadvantages: Time consuming for large datasets as Spreadsheet has it’s limitations, whereas Python is designed to handle large amount of data. Hence it may not handle large datasets efficiently. |

## Step 2: Data Inspection

Inspect the dataset. In the box below, discuss the following:

* Are there any irrelevant columns? Which ones?
* Are there any duplicates?
* Which columns have missing data?
* For each column with missing data, describe what you think the best way to handle that missing data is, and why?
* Are there any errors? Describe any you find.
* Is there anything else that requires data cleaning attention?

(12 marks)

|  |
| --- |
| **Are there any irrelevant columns? Which ones?**  Yes there are a few columns which are irrelevant to the dataset.   * PDF, Href and href formula columns as these 3 columns are basically just showing the pdf file name and the URLs, which is not value adding to our project. * Case number as the information is basically same as the date column, which is the date of shark attacks * Case number.1 and case number.2. The data in both columns are duplicated, and both columns consists of the same data as the Case number column.   **Are there any duplicates?**  Yes. There are 4 duplicated rows within the dataset.  Columns: Case Number is duplicated with Case Number.1 and Case Number.2; href formula and href column are almost the same.  **Which columns have missing data?**  Year, Type, Country, Area, Location, Activity, Name, Sex, Age, Injury, Fatal(Y/N), Time, Species, Investigator or Source  **For each column with missing data, describe what you think the best way to handle that missing data is, and why?**  To replace the missing data with “N/A”. If we delete the columns, the data deleted might come across useful when we are carrying out further analysis. It would be too late at that point for us to restore our deleted columns when we need it in the future.  **Are there any errors? Describe any you find.**  Yes. Some of the data within the columns are actually outliers. For example, there’s a “2017” value in the Fatal (Y/N) column.  **Is there anything else that requires data cleaning attention?**  Checking for outliers within every single column, and changing them to the correct information that matches the remaining information within the column.  For example, adding “http://sharkattackfile.net/spreadsheets/pdf\_directory/ “ to the rows for href and href format if they’re missing.  The country name should be only capital letters for the first letter instead of all capital letters (excluding USA) |

## 

## Step 3: Data Cleaning

Following on from Step 2, clean the dataset. Document all the changes you make in the box below. Before data cleaning, make sure to check every column thoroughly (audit the data). List all the actions to take so that you don’t overlook anything. (12 marks)

|  |
| --- |
| 1. Dropping duplicated columns (Case Number.1 and Case Number.2) and the 4 duplicated rows identified. 2. Sort the data according to “Original order” column in ascending order 3. Replacing all the missing values in the dataset with “N/A” 4. Clean up data (eg whitespaces) using Data clean-up> clean up suggestions function 5. Update the data format for each column accordingly. 6. Checking the values for each column by using column stats and pivot table, updating any outliers identified. 7. Change the data format for the “Time” column from 12h00 to 12:00 using the time format. For the remaining time such as evening/ between 6h00 to 8h00, I have summarised them as either dusk, dawn, morning, midday, afternoon, evening, night, midnight. 8. Replace all the values in the “Age” column that contains non-specific numerics with “N/A” (eg 10s, 20s) 9. Changing the data in the “Country “ column to only capital letter for the first letter and small letters for the remaining letters (excluding USA) using “=PROPER()” formula. 10. Ensuring that all the column names start with capital letters, except for PDF and href as they’re formatted as it is. |

## 

## Step 4: Data Cleaning Validation

Go through the data cleaning checklist and make sure there is no dirty data remaining! List below all the data validation steps you take. (3 marks)

|  |
| --- |
| 1. Check if there are any missing values, ensure there are no more blanks columns. 2. Check for duplicates, and remove the duplicates if there are any. 3. Check if all the outliers have been updated by using the pivot table. To update the outliers with the correct data if any of them have been missed out. |

## Step 5: Data Enrichment

With the dataset cleaned it’s time to enrich the data:

* Make an address column, by combining the Location, Area and Country columns together (this might affect your missing value strategy!).
* Add a new column, call it “Shark”. Extract information from the Species column. If the species text mentions the word “white”, make the “Shark” column value “Great White”. If the text mentions “bull”, make the “Shark” column value “Bull”. Otherwise, if neither of the words found, make the value “Other”. (Hint: make sure the species column is all lowercase).

## Step 6: Publish the dataset

Export the data as csv file. Call it final\_project.csv. Submit the file in a zip folder called final\_project.zip. Make sure the zip folder contains both your wrangled dataset and this word document with your answers!

## 