</talentlabs>

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!

</talentlabs>

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)

1. **BigQuery:**
    - **Advantages:**
        - Suitable for handling large datasets.
        - Powerful query capabilities.
        - Integration with other Google Cloud services.
    - **Disadvantages:**
        - Overkill for smaller datasets.
        - Requires knowledge of SQL.
        - Potential cost implications.
2. **Python:**
    - **Advantages:**
        - Highly versatile and powerful for data manipulation.
        - Extensive libraries like Pandas for data cleaning.
        - Good for automation and reproducibility.
    - **Disadvantages:**
        - Requires programming knowledge.
        - Can be less intuitive for beginners.
3. **Google Sheets:**
    - **Advantages:**
        - User-friendly interface.
        - Easy to share and collaborate.
        - Suitable for smaller datasets.
    - **Disadvantages:**
        - Limited to around 5 million cells.
        - Less powerful for complex data cleaning tasks.

Best Tool for the Task:

- For this dataset, which contains around 1100 rows, **Python** is the best-suited tool. It offers powerful data manipulation capabilities through libraries like Pandas, which are ideal for thorough data cleaning and transformation. Python also allows for easy automation and reproducibility of the data cleaning process.

</talentlabs>

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)

Irrelevant Columns:

- Yes, columns like Case Number.1, Case Number.2, original order seems redundant as they duplicate information from Case Number.

Duplicates:

- Yes, there are.

Missing Data:

- Columns with missing data:
  - Name
  - Sex
  - Age
  - Injury
  - Time
  - Species
  - Investigator or Source

Handling Missing Data:

- **Name:** Could be left as missing unless essential for analysis.
- **Sex:** Impute with 'Unknown' if missing.
- **Age:** If missing, can be left blank unless specific analysis requires it.
- **Injury:** Could use 'Unknown' or analyse contextually.
- **Time:** If not crucial, can be left missing; otherwise, consider 'Unknown'.
- **Species:** Impute with 'Unknown' or 'Not Specified'.
- **Investigator or Source:** If not critical, leave as missing.

Errors:

- Inconsistent date formats in the Date column.
- Inconsistent capitalisation and extra spaces in textual columns.
- Outliers inside age column. (e.g., Teen)

Other Data Cleaning Attention:

- Ensure consistent formatting across columns.
- Remove any irrelevant or redundant columns.

</talentlabs>

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. Remove redundant columns: Case Number.1, Case Number.2, original order.
2. Handle missing values as described.
3. Standardise date format in the Date column.
4. Normalise text data for consistent capitalization and remove extra spaces.
5. Check for and remove duplicate rows based on Case Number.

</talentlabs>

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. Verify no duplicates exist.
2. Ensure all columns are consistently formatted.
3. Confirm missing values are handled appropriately.
4. Check that irrelevant columns are removed.
5. Check for misspellings
6. Ensure all columns have the right data types

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!

</talentlabs>

visit out page:
www.talentlabs.org

contact us:
learn@talentlabs.org

19, Jalan USJ Heights 1/1B,
USJ Avenue, Subang Jaya, MY