**Objective 1 Descriptions**

What were mentioned in last meeting:

**1.Current Issues:**

1.Data Types & Structures:

-Some date fields are stored as text and need proper formatting.

-Inconsistencies in database fields require standardization.

2.Data Storage & Retrieval:

-Unclear update frequency.

-Data sources (internal, external, API, manual) need documentation.

3.Data Sensitivity:

-Some data may contain sensitive information requiring careful handling.

**2.Expected Outcomes:**

1. Standardize Data Formats & Types

-Recommend proper date formats and data cleaning strategies.

2.Document Data Sources

-Record update frequency, origins, storage format, and API availability.

3.Evaluate Data Accessibility & Storage

-Assess suitability for data warehousing and integration.

**Questions for next meeting:**

API

**Initial Findings**

1. **Stranding Data Sources Work：IWC\_StrandingsData\_FromNPR(Data)**

**Dataset summary[[1]](#endnote-0):**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Non-Null Count** | **Data Type** |
| id | 5883 | object |
| Data Year | 5883 | int64 |
| Year Submitted | 5883 | int64 |
| Large Area | 5831 | object |
| Species | 5732 | object |
| Country | 5883 | object |
| Local Area | 3929 | object |
| Local Taxonomy | 1256 | object |
| Local Area (Long/Lat) | 2408 | object |
| Females | 5623 | object |
| Males | 869 | float64 |
| Unknown | 1665 | object |
| Live or Dead | 370 | object |
| Outcomes | 370 | object |
| Information | 2021 | object |
| Contacts | 5424 | object |
| References | 1682 | object |

**Data Issues and Recommendations:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Sub-category** | **Current status** | **Issues identified** | **Recommendations**  **(Using python)** |
| Standardize Data Formats & Types | ID Format | **‘id’** is stored as object (text) | **‘id’** needs to be numeric | Standardize format |
| Date Format | **‘Data Year’** & ‘**Year Submitted**’are int64 | \ | \ |
| Numeric Data | **‘Females’, ‘Males’, ‘Unknown’**stored as object | Cannot perform numerical calculations | Convert to numeric *=>using pd.to\_numeric()* |
| Geolocation Data | **‘Local Area (Long/Lat)’** is object | Latitude/Longitude not separated，making it difficult for spatial analysis (e.g., mapping, GIS, and gap analysis). | Split into Latitude &Longitude columns and convert them into numeric format  *Convert to string=>astype(str),*  *Split using=>str.split(',', expand=True)*  *Convert to numeric =>pd.to\_numeric* |
| Missing Values | **Live or Dead** has only 370out of 5883 total records | Data completeness issue | Needs to be considered for its impact on analysis |
| Document Data Sources | Data Source | From NPR | \ | \ |
| Update Frequency[[2]](#endnote-1) | Data is available from 2010-2023, but submissions occur irregularly from 2013-2023. | Some years (e.g., 2013-2015) have relatively few records, may lead to missing trends and potential bias. | Standardize submission frequency (e.g., annual updates). Regular reporting could be encouraged. Supplement missing data with external sources. |
| Storage Format[[3]](#endnote-2) | .csv | \ | \ |
| API [[4]](#endnote-3)Availability | ?not mentioned |  |  |
| Evaluate Data Accessibility & Storage | Data Storage[[5]](#endnote-4) | File Size: 1.70 MB  Memory Usage in Pandas: 5.02 MB | CSV format is fine for now but may become inefficient as data grows. | If data volume increases, consider storing in a relational database (SQL) or cloud storage. |
| Data Integration[[6]](#endnote-5) | **‘id’**column exists, no duplicates, no missing values. | **‘id’**could be the primary key for merging | \ |

1. **Vessel Strike Data Sources Work**-may need to consider the IWC Ship Strikes Database Table Relationships

**（1）AustraliaData\_working.xlsx**

**（2）IWC\_ShipStrikes\_PublicData.xlsx**

**（3）Marine strike log - Share - IWC.xlsx**

**Codes**

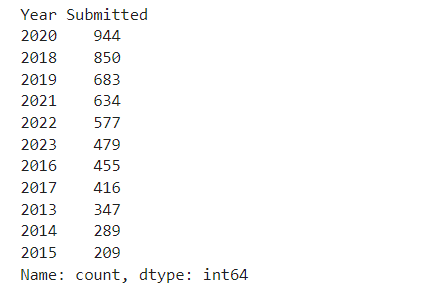
1. Data set summary:

   print(df.head())

   print(df.info()) [↑](#endnote-ref-0)
2. Document Data Sources

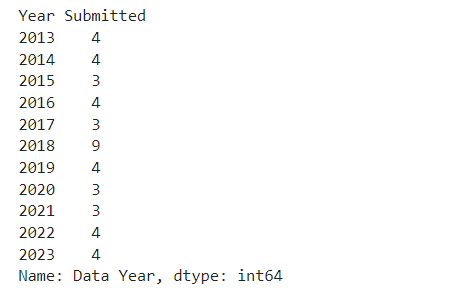
   # Count records by "Year Submitted"

   print(df['Year Submitted'].value\_counts())

   # Compare "Data Year" and "Year Submitted" to see how data is reported

   print(df.groupby('Year Submitted')['Data Year'].nunique())

   # Check the range of years covered

   print(f"Data Year Range: {df['Data Year'].min()} - {df['Data Year'].max()}")

   print(f"Submission Year Range: {df['Year Submitted'].min()} - {df['Year Submitted'].max()}")

   b4b9a1197c4406fc2b08b80c6ea10c9 [↑](#endnote-ref-1)
3. Storage Format

   import os

   file\_path = r"...\IWC\_StrandingsData\_FromNPR(Data).csv"

   file\_format = os.path.splitext(file\_path)[1]

   print(f"Storage Format: {file\_format}") [↑](#endnote-ref-2)
4. Check API Availability

   import requests

   api\_url = "https://example.com/api/stranding\_data" # Replace with actual API

   response = requests.get(api\_url)

   if response.status\_code == 200:

   print("API is available!")

   else:

   print("No API access found.") [↑](#endnote-ref-3)
5. Data Storage

   import os

   # Get file size in MB

   file\_path = r"C:\Users\BFSU\Desktop\UoM\semester2\Apply data science\Objective1-DBs\IWC\_StrandingsData\_FromNPR(Data).csv"

   file\_size = os.path.getsize(file\_path) / (1024 \* 1024) # Convert bytes to MB

   # Get dataframe memory usage

   memory\_usage = df.memory\_usage(deep=True).sum() / (1024 \* 1024) # Convert bytes to MB

   print(f"File Size: {file\_size:.2f} MB")

   print(f"Memory Usage in Pandas: {memory\_usage:.2f} MB") [↑](#endnote-ref-4)
6. Data Integration

   print(f"Total Records: {len(df)}")

   print(f"Unique IDs: {df['id'].nunique()}")

   print(f"Duplicate IDs: {len(df) - df['id'].nunique()}") [↑](#endnote-ref-5)