**IWC Data Source Documentation**

**1. Stranding Data Sources**

**Dataset Name:** IWC\_StrandingsData\_FromNPR

**1. Data overview**

* Source or Provider: NPR
* File Type: CSV (.csv)
* Number of Records: 5,883
* Number of Columns: 17

**2. Description**

**Purpose:** This dataset is used to track and analyze whale stranding events, with a focus on understanding the causes, locations, and whale demographics.

**Context:** The data provides insight into whale mortality patterns and potential environmental threats. It may help researchers, conservationists, and policymakers improve marine protection strategies.

**Use Case:** Can be used for Marine conservation efforts, environmental policy-making, biodiversity monitoring, identifying patterns in whale strandings, scientific research, including studies on climate change impacts, human interaction, and ecosystem health.

**3. Data details**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id | object | Unique identifier for each record |
| Data Year | int64 | Year the stranding occurred |
| Year Submitted | int64 | Year the data was reported |
| Large Area | object | General geographical area |
| Species | object | Species of the stranded whale |
| Country | object | Country where stranding occurred |
| Local Area | object | More specific stranding location |
| Local Taxonomy | object | Taxonomic classification in local terms |
| Local Area (Long/Lat) | object | Latitude and Longitude of the event |
| Females | object | Number of female whales in the event |
| Males | float64 | Number of male whales in the event |
| Unknown | object | Unspecified number of whales |
| Live or Dead | object | Status of the whale(s) at the time of stranding |
| Outcomes | object | Consequence of the stranding |
| Information that may help to explain data | object | Additional context or notes |
| Contacts | object | Contact person or organization reporting the event |
| References | object | Sources used in reporting the event |

**4. Data Issues and Recommendations**

**4.1 Inconsistent Data**

**4.1.1 Data Type Format**

**Issues:**

* The **‘id’** column is stored as an object instead of numeric.
* **‘Data Year’ & ‘Year Submitted’** are in **int64** but should be in **datetime** format.
* **‘Females’ and ‘Unknown’** are stored as text, making numerical calculations difficult.
* **‘Local Area (Long/Lat)’** is stored as an object rather than separate numeric fields for **latitude and longitude**.

**Recommendations:**

* Convert **id** to a numeric format if it represents an integer-based identifier.
* Convert **Data Year & Year Submitted** to datetime.
* Convert **Females & Unknown** to numeric.
* Split **Local Area (Long/Lat)** into **two separate columns (Latitude, Longitude)** and store them as float.

**4.1.2 Inconsistent Values**

**Issues:**

Different ways of recording ‘unknown’ values: ‘Unknown’, ‘Not known’, ‘not reported’, blank spaces, and missing values are inconsistently used.

Some species names are written inconsistently or classified differently.

**Recommendations:**

Standardize unknown values (use either “Unknown” for unclear data or NaN for truly missing values).

Standardize species naming conventions using a controlled vocabulary.

**4.2 Missing values**

**Observation:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Missing Values (Count)** | **Missing Values (%)** |
| Outcomes | 5513 | 93.71 |
| Live or Dead | 5513 | 93.71 |
| Males | 5014 | 85.23 |
| Local Taxonomy | 4627 | 78.65 |
| Unknown | 4218 | 71.7 |
| References | 4201 | 71.41 |
| Information that may help to explain data | 3862 | 65.65 |
| Local Area (Long/Lat) | 3475 | 59.07 |
| Local Area | 1954 | 33.21 |
| Contacts | 459 | 7.8 |
| Females | 260 | 4.42 |
| Species | 151 | 2.57 |
| Large Area | 52 | 0.88 |
| id | 0 | 0 |
| Country | 0 | 0 |
| Year Submitted | 0 | 0 |
| Data Year | 0 | 0 |

**Issues:**

* Key variables such as **Live or Dead, Outcomes, and Males** have over 85% missing values, making certain analyses difficult.
* Geolocation data **(Local Area (Long/Lat))** is missing for over 50% of records, limiting spatial analysis.
* **Species** are missing in some cases, which can impact species-specific research.

**Recommendations:**

* Assess the impact of missing data on analysis objectives.
* Consider imputation methods (e.g., filling missing location data based on nearby records).
* For highly incomplete fields (**Outcomes, Live or Dead**), consider whether they should be retained or excluded from analysis.

**4.3 Data Sources Documentation**

* **Data Source:** From NPR
* **Update Frequency:** Data is available from 2010-2023, but submissions are irregular, with fewer records in certain years (2013-2015).
* **Storage Format:** CSV (.csv)
* **API Availability:** No API is available, but alternative sources include:

**Mediterranean Database of Cetacean Strandings (MEDACES)**: <http://medaces.uv.es/>

**Indian Ocean Network for Cetacean Research**: <https://indocet.org/en/research/report-a-stranding/>

**Recommendations:**

* If possible, establish a standardized update schedule (e.g., annual reporting).
* Consider external data sources to supplement missing records.

**4.4 Data Storage & Integration**

**Observation:**

* **File Size:** **1.70 MB**
* Memory Usage in Pandas: 5.02 MB
* Data Structure Suitability:
* CSV format is manageable for now, but as data grows, it may become inefficient.
* Consider storing in a relational database (SQL) or cloud-based storage for scalability.

**Issues:**

* Limited geolocation data affects spatial analysis.
* High missing values in critical fields reduce completeness.

**Recommendations:**

* Ensure that **id** remains a unique and complete primary key for database merging.
* Improve data collection methods to reduce missing location data.

**Summary of key issue and recommendations**

|  |  |  |
| --- | --- | --- |
| **Category** | **Sub-category** | **Recommendations** |
| Standardize Data Formats & Types | ID Format | Standardize format |
| Date Format | convert into ‘datetime’format |
| Numerical Data | Convert into numerical data format |
| Geolocation Data | - Split into Latitude & Longitude columns  - Convert them into numeric format |
| Missing Values | / | - Assess the impact of missing values  - Use appropriate imputation techniques |
| Data sources documentation | Update Frequency | * Standardize submission frequency (e.g., annual updates). Regular reporting could be encouraged. * Supplement missing data with external sources. |
| Data Storage & Integration | Data Storage | If data volume increases, consider storing in a relational database (SQL) or cloud storage. |

**2. Vessel Strike Data Sources**

**2.1 AustraliaData\_working**

There are three sheets in this dataset, which are FromSSDB, DavePeelData, and ammc-shipstrike-2022-09-19\_4385.

**2.1.1 FromSSDB**

**1. Data Overview**

* **Dataset Name:** FromSSDB
* **Source or Provider:** FromSSDB dataset was extracted fromAustraliaData\_working.xlsx
* **File Type:** Excel (.xlsx)
* **Number of Records:** 68
* **Number of Columns:** 116

**2. Description**

**Purpose:** It focuses on whale-vessel strike incidents, documenting cases where ships have collided with whales. The data is essential for understanding the frequency, severity, and contributing factors to these incidents, providing detailed records of whale species affected, environmental conditions, vessel characteristics, and observed injuries.

**Context:** The dataset is combined from several tables in a database, including tables like ss\_vessels, user, ss\_contacts, global\_species, ss\_cases, ss\_case\_evidence\_links, global\_largearea, ss\_evidences, ss\_evidence\_pins, ss\_evidence\_pin\_data, and ss\_question\_options

**Use Case:** The dataset can be used in marine conservation, maritime safety, policy-making, and scientific research to analyze whale-vessel collisions. It helps assess species vulnerability, optimize shipping routes, develop environmental regulations, and support conservation efforts. Moreover, it can be used for trend analysis, risk prediction e.g. predicting high-risk areas and contributes to educational and advocacy campaigns for marine wildlife conservation.

**3. Data Details**

|  |  |  |
| --- | --- | --- |
| **Column** | **Data type** | **Description** |
| EvidPinID | int64 | Unique identifier for evidence records |
| CaseLinkID | int64 | Identifier linking related cases |
| EvidLinkID | int64 | Identifier linking related pieces of evidence |
| CaseID | int64 | Unique case identifier |
| EvidencesID | int64 | Unique identifier assigned to each piece of evidence associated with a case |
| Conclusion | object | Classification of the case (Definite, Probable, or Possible) |
| CaseTtitle | object | Title or summary of the case in the format of YYYY-MM-DD – [collision/at sea/shore] – global area |
| CaseNotes | object | Additional details or assessments related to the case |
| CaseDate | datetime64[ns] | Date of incident |
| GlobalArea | object | General geographic area where the incident occurred. It consists of 5 unique values which are Pacific Ocean - Coral Sea, Pacific Ocean – South, Indian Ocean, Pacific Ocean – North, and Pacific Ocean - Tasman Sea |
| ScientificName | object | Scientific name of the species involved |
| CommonName\_1 | object | Common name of the species involved |
| conclusion\_1 | object | Additional conclusion classification (Definite, Probable, or Possible) |
| date\_evidence | datetime64[ns] | Date when evidence was recorded |
| at\_sea | int64 | Indicator if the incident occurred at sea |
| collision | int64 | Indicator if a collision occurred |
| Stuck | int64 | Indicator if the vessel got stuck |
| near\_miss | int64 | Indicator if a near-miss occurred |
| Status | object | Status of the case investigation e.g. assigned |
| additional\_required | int64 | Indicates if further investigation is required |
| ReportID | int64 | Unique identifier for incident reports |
| DateInc | datetime64[ns] | Date of the incident occurrence |
| DateDesc | object | Description of the incident date details |
| LargeArea | object | Broad geographical classification. It consists of 5 areas which are Pacific Ocean - Coral Sea, Pacific Ocean – South, Indian Ocean, Pacific Ocean – North, and Pacific Ocean - Tasman Sea |
| LocatDesc | object | Specific location description of the incident |
| Latitude | object | Latitude coordinate of the incident location |
| Longitude | object | Longitude coordinate of the incident locatio |
| OtherInfo | object | Miscellaneous information about the case |
| Witness | object | Names of witnesses of the event |
| LastPort | float64 | Last port the vessel departed from |
| DepartDate | float64 | Date the vessel departed |
| DestPort | object | Destination port of the vessel |
| ArriveDate | datetime64[ns] | Arrival date of the vessel |
| Route | object | Route taken by the vessel |
| SciSpecies | object | Scientific classification of species involved |
| CommonName | object | Common name of the species involved |
| CommSpecies | object | Commercial species classification |
| ConfidenceSpp | object | Confidence level in species identification (Confirmed, Indeterminate, Unconfirmed, Tentative) |
| SppChara | object | Describes the characteristics or identification confidence of the species. |
| SppIDer | object | The person or method used to identify the species |
| AnimalCondition | object | Condition of the whale post-incident e.g. Alive |
| CarcassArrive | object | Indicates whether the whale carcass arrived on shore or was recovered from the vessel e.g. Stranded dead |
| NecropLevel | object | Level of necropsy performed |
| NecropReport | object | Report associated with necropsy findings (Yes, No, Not Known) |
| NecropContact | object | Name of the person responible for necropsy |
| Sex | object | Sex of the whale involved |
| EstLength | float64 | Estimated length of the whale |
| EstLegthIDer | float64 | Person or method used to estimate whale length |
| MeasLength | float64 | Measured length of the whale |
| AgeClass | object | Age classification of the whale e.g. Subadult, and Calf |
| AgeMethod | float64 | Method used for estimating the whale’s age |
| Tissue | object | Indicates whether a tissue sample was collected from the whale (Yes, No, Not Known) |
| TissueContact | float64 | Contact person related to tissue sampling. |
| WhaleSightPre | object | Observations of the whale before the strike |
| TimeBeforeStrike | float64 | Time interval before the strike occurred |
| WhaleBehaviorPre | object | Behavior of the whale before the strike |
| OtherCetacean | object | Other cetaceans present during the incident |
| GroupSize | float64 | Number of whales or cetaceans involved |
| CollisionEvid | object | Evidence confirming the collision |
| WhalePost | object | Condition of the whale after the strike |
| BloodSight | object | Whether blood was observed post-strike (Yes, No, Not Known) |
| WhaleSightPost | object | Indicates whether the whale was observed after the vessel strike (Yes, No, Not Known) |
| WhaleInjury | object | Description of injuries sustained by the whale |
| WhaleLife | object | Describes the whale’s survival status after the strike. |
| WhaleBehaviorPost | object | Whale behavior after the strike |
| PropellerStrike | object | Indicates whether the whale was struck by the vessel’s propeller |
| DeadPreStrike | object | Indicates whether the whale was already dead before the vessel strike |
| StrandReport | object | Indicates whether the whale was stranded after the strike |
| BowEvidence | object | Evidence of a whale strike on the vessel’s bow |
| BowPosition | object | Specific position on the bow where the whale strike occurred |
| VesselSpeed | float64 | Speed of the vessel at the time of the incident |
| VesselID | float64 | Unique identifier for the vessel |
| VesselUserID | float64 | Identifier for the person or organization that owns the vessel |
| VesselType | object | Type of vessel involved in the incident |
| VesselName | object | Name of the vessel |
| VesselPort | object | Port associated with the vessel |
| VesselIMO | float64 | International Maritime Organization (IMO) number for the vessel |
| VesselLength | float64 | Length of the vessel |
| VesselDraft | float64 | Draft depth of the vessel |
| VesselTonnage | float64 | Weight or tonnage of the vessel |
| VesselYrBuilt | float64 | Year the vessel was built |
| VesselHullCon | object | Hull construction type of the vessel |
| VesselHullMat | float64 | Material of the vessel’s hull |
| VesselHullCol | float64 | Color of the vessel’s hull |
| VesselPropulsion | object | Type of propulsion system used by the vessel |
| VesselEngType | float64 | Type of engine used by the vessel |
| VesselEngBHP | float64 | Engine power in brake horsepower (BHP) |
| VesselCruiseSpeed | float64 | Cruising speed of the vessel |
| VesselPropNum | float64 | Number of propellers on the vessel |
| VesselPropDia | float64 | Diameter of the vessel’s propellers |
| VesselPropPitch | float64 | Pitch angle of the vessel’s propellers |
| VesselPropBlade | float64 | Number of blades on the vessel’s propeller |
| VesselShaftDis | float64 | Distance of the shaft from a reference point |
| VesselCompany | float64 | Company that owns or operates the vessel |
| VesselMaster | float64 | Name or identifier of the vessel's captain or master |
| VesselOwner | float64 | Name of the person or entity that owns the vessel |
| VesselOther | float64 | Additional vessel-related details that do not fit in other categories |
| PropRPM | float64 | Rotations per minute (RPM) of the vessel’s propeller |
| VesselDamage | object | Indicates whether the vessel sustained damage due to the incident |
| DamageDesc | object | Description of the damage sustained by the vessel |
| AvoidanceAction | object | Actions taken by the vessel to avoid the whale |
| VesselDoing | object | The vessel's activity or purpose at the time of the incident |
| WatchKeepers | float64 | Number of watchkeepers on duty at the time of the strike |
| OtherNavInfo | object | Additional navigational details related to the incident |
| WeatherDesc | float64 | Description of weather conditions at the time of the incident |
| VisType | float64 | Visibility conditions during the incident |
| WindForce | float64 | Strength of the wind at the time of the strike |
| SeaState | float64 | Condition of the sea at the time of the incident |
| WaterDepth | float64 | Depth of the water where the strike occurred |
| LightCond | float64 | Lighting conditions at the time of the incident |
| EvidenceType | float64 | Type of evidence available to confirm the incident |
| StuckBow | float64 | Indicates whether the whale was stuck on the vessel’s bow after impact |
| PersonInjury | float64 | Indicates whether any human injuries occurred due to the incident |
| InjuryDesc | float64 | Description of injuries sustained by individuals involved |
| UserID | int64 | Identifier for the person who recorded the data |
| UserName | object | Name of the person who recorded the data |

**4. Data Issues and Recommendations**

**4.1. Inconsistent Data**

**4.1.1 Data Type Format**

**Issues:**

* The DateDesc column is incorrectly formatted as an object due to an invalid date format.
* Columns containing only missing values are stored as float64.

**Example fields and values:**

All values in DepartDate and VesselOwner are missing, but they are stored as float64 due to missingness instead of being properly classified as datetime and object, respectively.

**Recommendation:**

* If a column represents dates, store it in the datetime datatype.
* If a column contains textual data, store it as an object (or string in newer Pandas versions).
* If a column contains numeric values, use float for decimal numbers and int for whole numbers.

**4.1.2 Inconsistent Values**

**Issues:** The dataset contains various ways of recording ‘unknown’ values.

**Example fields and values:**

* ‘Unknown’ : Route, Sex, AgeClass, VesselDoing
* ‘unknown’ : WhaleBehaviorPost, DeadPreStrike, SppChara
* ‘Not known’ : NecropLevel, NecropReport, Tissue, WhaleSightPre, WhaleBehaviorPre, WhalePost, BloodSight, WhaleSightPost, VesselDamage, AvoidanceAction
* Others e.g. not reported, not indicated, non, none, and blank

**Recommendation:**

* Standardize all ‘unknown’ values by converting them into a single consistent term (e.g., "Unknown" for ambiguous values and NaN for truly missing data).
* Define a clear distinction between "Unknown" and missing values
* Use NaN for missing values where data was never recorded.
* Use "Unknown" where information is ambiguous but exists.

**4.2 Missing Values**

**Observation:**

|  |  |  |
| --- | --- | --- |
| **Column** | **Missing Values**  **(Count)** | **Missing Values (%)** |
| EvidPinID | 0 | 0.00 |
| CaseLinkID | 0 | 0.00 |
| EvidLinkID | 0 | 0.00 |
| CaseID | 0 | 0.00 |
| EvidencesID | 0 | 0.00 |
| Conclusion | 22 | 32.35 |
| CaseTtitle | 0 | 0.00 |
| CaseNotes | 11 | 16.18 |
| CaseDate | 7 | 10.29 |
| GlobalArea | 4 | 5.88 |
| ScientificName | 20 | 29.41 |
| CommonName\_1 | 17 | 25.00 |
| conclusion\_1 | 22 | 32.35 |
| date\_evidence | 10 | 14.71 |
| at\_sea | 0 | 0.00 |
| collision | 0 | 0.00 |
| Stuck | 0 | 0.00 |
| near\_miss | 0 | 0.00 |
| Status | 0 | 0.00 |
| additional\_required | 0 | 0.00 |
| ReportID | 0 | 0.00 |
| DateInc | 10 | 14.71 |
| DateDesc | 57 | 83.82 |
| LargeArea | 4 | 5.88 |
| LocatDesc | 3 | 4.41 |
| Latitude | 23 | 33.82 |
| Longitude | 23 | 33.82 |
| OtherInfo | 3 | 4.41 |
| Witness | 18 | 26.47 |
| LastPort | 68 | 100.00 |
| DepartDate | 68 | 100.00 |
| DestPort | 67 | 98.53 |
| ArriveDate | 67 | 98.53 |
| Route | 67 | 98.53 |
| SciSpecies | 20 | 29.41 |
| CommonName | 17 | 25.00 |
| CommSpecies | 12 | 17.65 |
| ConfidenceSpp | 11 | 16.18 |
| SppChara | 58 | 85.29 |
| SppIDer | 66 | 97.06 |
| AnimalCondition | 67 | 98.53 |
| CarcassArrive | 49 | 72.06 |
| NecropLevel | 57 | 83.82 |
| NecropReport | 56 | 82.35 |
| NecropContact | 66 | 97.06 |
| Sex | 57 | 83.82 |
| EstLength | 61 | 89.71 |
| EstLegthIDer | 68 | 100.00 |
| MeasLength | 66 | 97.06 |
| AgeClass | 57 | 83.82 |
| AgeMethod | 68 | 100.00 |
| Tissue | 53 | 77.94 |
| TissueContact | 68 | 100.00 |
| WhaleSightPre | 45 | 66.18 |
| TimeBeforeStrike | 67 | 98.53 |
| WhaleBehaviorPre | 51 | 75.00 |
| OtherCetacean | 61 | 89.71 |
| GroupSize | 65 | 95.59 |
| CollisionEvid | 28 | 41.18 |
| WhalePost | 35 | 51.47 |
| BloodSight | 40 | 58.82 |
| WhaleSightPost | 34 | 50.00 |
| WhaleInjury | 41 | 60.29 |
| WhaleLife | 47 | 69.12 |
| WhaleBehaviorPost | 51 | 75.00 |
| PropellerStrike | 54 | 79.41 |
| DeadPreStrike | 0 | 0.00 |
| StrandReport | 66 | 97.06 |
| BowEvidence | 66 | 97.06 |
| BowPosition | 66 | 97.06 |
| VesselSpeed | 65 | 95.59 |
| VesselID | 60 | 88.24 |
| VesselUserID | 60 | 88.24 |
| VesselType | 60 | 88.24 |
| VesselName | 61 | 89.71 |
| VesselPort | 67 | 98.53 |
| VesselIMO | 68 | 100.00 |
| VesselLength | 65 | 95.59 |
| VesselDraft | 68 | 100.00 |
| VesselTonnage | 68 | 100.00 |
| VesselYrBuilt | 68 | 100.00 |
| VesselHullCon | 65 | 95.59 |
| VesselHullMat | 68 | 100.00 |
| VesselHullCol | 68 | 100.00 |
| VesselPropulsion | 65 | 95.59 |
| VesselEngType | 68 | 100.00 |
| VesselEngBHP | 68 | 100.00 |
| VesselCruiseSpeed | 68 | 100.00 |
| VesselPropNum | 68 | 100.00 |
| VesselPropDia | 68 | 100.00 |
| VesselPropPitch | 68 | 100.00 |
| VesselPropBlade | 68 | 100.00 |
| VesselShaftDis | 68 | 100.00 |
| VesselCompany | 68 | 100.00 |
| VesselMaster | 68 | 100.00 |
| VesselOwner | 68 | 100.00 |
| VesselOther | 68 | 100.00 |
| PropRPM | 68 | 100.00 |
| VesselDamage | 58 | 85.29 |
| DamageDesc | 58 | 85.29 |
| AvoidanceAction | 60 | 88.24 |
| VesselDoing | 67 | 98.53 |
| WatchKeepers | 68 | 100.00 |
| OtherNavInfo | 64 | 94.12 |
| WeatherDesc | 68 | 100.00 |
| VisType | 68 | 100.00 |
| WindForce | 68 | 100.00 |
| SeaState | 68 | 100.00 |
| WaterDepth | 68 | 100.00 |
| LightCond | 68 | 100.00 |
| EvidenceType | 68 | 100.00 |
| StuckBow | 68 | 100.00 |
| PersonInjury | 68 | 100.00 |
| InjuryDesc | 68 | 100.00 |
| UserID | 0 | 0.00 |
| UserName | 54 | 79.41 |

**Recommendation:**

* There are 35 columns where all values are missing which should be dropped because they are unmeaningful data
* The other 45 columns have more than 50% missing values, considering as having severe missing data issues. They should be considered whether to be dropped, imputed, or find external data sources if we want to analyse the data or handle redundancy
* Only 17 columns have missing values less than 5%, which is ignorable missing rate.

**4.3 Data Quality and Completeness:**

**Evaluation:**

There are 116 columns in this dataset. Some of them may be unnecessary, and the large number of columns can be considered a data structure issue, as some may be redundant, irrelevant, or overly complex for analysis. For example, features like Conclusion and conclusion\_1 may contain the same information, making them duplicates.

Moreover, several columns have only one unique value (excluding NaN), meaning they lack variability. For instance, features like stuck and near\_miss contain only 0, offering no meaningful distinction. Some columns have only one unique value along with missing data (NaN). For example, VesselDoing contains only Unknown and missing values, suggesting that the missing data may carry the same meaning as Unknown.

**Recommendation:**

Unnecessary columns should be removed, especially those with excessive missing valuee or columns with low variance, meaning that those columns have little variability and may not contribute useful information.

**4.4 Storage and Integration:**

**Current State:** The dataset is stored as a sheet in the Excel file named AustraliaData\_working.xlsx

**Recommendation:**

* For scalability and better integration into a data warehouse, consider migrating the data to a SQL database or cloud-based storage.
* In SQL databases, consider separating categories (e.g., vessel details, location details) into separate tables to improve structure and efficiency.

**Summary of key issues and recommendations:**

|  |  |  |
| --- | --- | --- |
| **Category** | **Sub-category** | **Recommendation** |
| **Inconsistent Data** | **Data Type Format** | Standardize data types: store dates as datetime, text as object/string, and numbers as int/float. |
| **Inconsistent Values** | Standardize "unknown" values by using "Unknown" for uncertain entries and NaN for actual missing data. |
| **Missing Values** | **-** | - Drop 35 fully empty columns if they are unnecessary  - Assess 45 columns with >50% missing data  - Ignore missing values in 17 columns (<5% missingness). |
| **Data Quality and Completeness** | **-** | - Remove redundant/low-variance columns  - Drop duplicates (e.g., Conclusion and conclusion\_1)  - Eliminate single-value columns. |
| **Storage and Integration** | **-** | - Migrate to SQL/cloud storage  - Separate data categories (e.g., vessel details, location details) for better structure. |

**Limitation of this analysis:**

Some columns contains only missing values, so some data descriptions may be misinterpreted.

**2.1.2. DavePeelData**

**Dataset Name:** DavePeelData

**1. Data overview**

* Source or Provider: International Whaling Commission (IWC)
* File Type: Excel (.xlsx)
* Number of Records: 196
* Number of Columns: 16

**2. Description**

**Purpose:** The dataset records collisions between vessels and whale species, including event ID, year, source of evidence, geographical location (large area and small area), scientific name of the whale species, common name, evidence type, collision outcome, and vessel type, among 16 attributes in total. The primary data source is newspaper reports, and the recorded data spans a long period, with the earliest entry dating back to 1857. The main objective of this dataset is to collect and analyze historical data on whale-vessel collisions to study trends, influencing factors, and potential risks associated with this phenomenon.

**Context:** The dataset falls within the fields of marine ecological conservation and whale research and was collected by the International Whaling Commission (IWC). It covers geographical regions such as the South Pacific Ocean, the South Indian Ocean, and waters near Australia. The data documents collisions between various types of vessels (such as sailing ships, military vessels, and passenger steamers) and whales, along with descriptions of whale injuries. This dataset is of significant value for studying the interactions between whales and human maritime activities, assessing the risks of vessel collisions, and promoting marine conservation policies.

**Use Case:**

1. Marine Conservation Policies: Analyze whale-vessel collision patterns to support governments and environmental organizations in developing navigation restrictions, low-speed zones, and avoidance strategies to reduce the threat posed by vessels to whales.
2. Whale Research: Utilize collision data to study the movement patterns of different whale species (e.g., sperm whales) in specific sea areas and their reactions to vessels.
3. Data Visualization and Analysis: Use maps and time-series analysis to identify high-risk areas, predict future collision-prone regions, and propose mitigation measures.
4. Historical Data Modeling: Apply machine learning techniques (e.g., time-series forecasting, classification models) to analyze the key factors contributing to whale collisions, aiming to improve maritime management and conservation strategies.
5. Marine Incident Investigation: Provide historical case studies for marine biologists and policymakers to assess whether vessel collisions have led to declines in whale populations and to explore potential countermeasures.

**3. Data details**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| ID | Object  (e.g., A192, A1) | Unique identifier |
| IWC ID | object  (e.g., NA, 271, P) | Unique identifier (IWC) |
| Year | int64  from 1857 to 2019  (e.g., 2003, 2006) | Year when the event occurred |
| EvidenceDate | object  (e.g., 4-Jun-1857, 1888, Jul-1904) | Date when evidence was recorded |
| Source | object  ('Newspaper', 'IWC', 'Shipwreck database', 'Website',  'ProgRepAustralia', 0, 'IWC DATA', 'Facebook') | Source of evidence |
| LargeArea | object  (e.g., Indian Ocean - South) | Large geographical area |
| SmallArea | object  (e.g., 10 miles North of Wollongong, NSW) | Specific location |
| ScientificName | object  (e.g., Physeter macrocephalus, Unidentified cetacean ) | Scientific name of the whales |
| CommonName | object  (e.g., Blue whale, Unidentified cetacean, 0) | Common name of the whales |
| IDQuality | object  ('Indeterminate', 'Confirmed', 'Unconfirmed',  'Tentative dwarf/antarctic') | Quality of the identification |
| EvidenceType | object  ('Collision', 'Shore', 'Stranded', 'Post mortem', 0, 'Not known',  'Washed up', 'At Sea', 'Found at Sea', 'Found on Shore', 'Photo', 'Probable collision', 'Washed UP', 'Post mortem, seen floating in ferry route', 'Post mortem, washed up') | Type of evidence |
| CollisionEvidence | object  (e.g., ‘Witnesse’, ‘Unknown’, ‘likely fatal’,‘Passengers’) | Detail on collision evidence |
| StrikeResult | object  (e.g., ‘Unknown’, ‘Whale almost cut in half’, ‘Whale swam away’) | Outcome of the strike |
| OtherInfo | object  (e.g., ‘Newspaper report’, ‘SC/57/ProgRepAustralia’, ‘A whale 'attacking' boat’) | Additional relevant information |
| VesselType | object  (e.g., ‘Clipper’, ‘unknown’, ‘Passenger steamer’) | Type of vessel involved |
| CSIRO Flagged Issue | object  (e.g., ‘Described as an 'attack’, ‘Possibly a whale’) | Flags for issues identified by CSIRO (Commonwealth Scientific and Industrial Research Organization) |

**4. Data Issues and Recommendations**

**4.1 Inconsistent Data**

**4.1.1 Data Type Format**

**Issues:**

* The **ID field** is not continuous and is stored as an Object
* Missing values are stored in different formats, such as **NA, null, not known, and unknown**.

**Recommendations:**

* Converting the **ID field** to a standardized numeric format.
* standardizing the representation of missing values to **NA**.

**4.1.2 Unclear Data Meaning**

**Issues:**

* In the **IWC ID field**, the value **'P'** indicates that the record originates from **ProgRepAustralia**, but its meaning is unclear in the IWC ID.

**Recommendations:**

* If **ProgRepAustralia** and **IWC** are not clearly related, convert all 'P' values to NA.

**4.1.3 Duplicate Fields**

**Issues:**

* **Year** and **EvidenceDate** contain overlapping

**Recommendations:**

* Unify them into a single column with a standardized date format for **EvidenceDate** (e.g., 1857-06-24, 1888, 1904-07).

**4.1.4 Inconsistent Categorical Classifications**

**Issues:**

* Some categorical fields **(Source, IDQuality, EvidenceType, StrikeResult， VesselType)** contain multiple names for the same category.

**Recommendations:**

* Standardize category names across these fields.

**4.2 Missing values**

**Observation:**

|  |  |  |
| --- | --- | --- |
| **Column** | **Missing Values**  **(Count)** | **Missing Values (%)** |
| ID | 0 | 0.00 |
| IWC ID | 110 | 57.59 |
| Year | 0 | 0.00 |
| EvidenceDate | 0 | 0.00 |
| Source | 4 | 2.09 |
| LargeArea | 2 | 1.05 |
| SmallArea | 0 | 0.00 |
| ScientificName | 0 | 0.00 |
| CommonName | 2 | 1.05 |
| IDQuality | 132 | 69.11 |
| EvidenceType | 4 | 2.09 |
| CollisionEvidence | 36 | 18.85 |
| StrikeResult | 15 | 7.85 |
| OtherInfo | 70 | 36.65 |
| VesselType | 18 | 9.42 |
| CSIRO Flagged Issue | 173 | 90.58 |

**Issues:**

* Missing values found in columns: IWC ID, Source, LargeArea, CommonName, IDQuality, EvidenceType, CollisionEvidence, StrikeResult, OtherInfo, VesselType, CSIRO Flagged Issue.

**Recommendations:**

* Complete the missing information as much as possible using other datasets.

**4.4 Data Quality and Completeness:**

**Issues:**

* Incomplete Dates
* Lack of Geospatial Data

**Recommendation:**

* Some records in 'EvidenceDate' contain incomplete information.
* Location information is not precise; latitude and longitude data are missing.

**4.5 Data Storage & Integration**

**Issues:**

* Dataset is currently in Excel format, which may not be optimal for large-scale analysis.

**Recommendations:**

* Storing the data in a structured format like SQL or cloud-based storage for better accessibility and scalability.

**4.3 Data Sources Documentation**

* Data Source: From IWC
* Update Frequency: Not mentioned
* Storage Format: .xlsx
* API Availability: Not mentioned

**Issues:**

* The update frequency is not specified, making it difficult to track changes or assess data reliability.
* There is no API available for direct data retrieval.

**Recommendations:**

* Contact data providers to clarify the update frequency of this dataset.
* Explore alternative data sources or structured data retrieval processes (e.g., scheduled updates, web scraping if permitted).

**Summary of key issue and recommendations**

|  |  |  |
| --- | --- | --- |
| **Category** | **Sub-category** | **Recommendations** |
| Inconsistencies | Formatting issues | Converting the ID field to a standardized numeric format |
| Standardize the representation of missing values to NA |
| Unclear Data Meaning | If ProgRepAustralia and IWC are not clearly related, convert all 'P' values to NA |
| Duplicate Fields | Unify Year and EvidenceDate into a single column with a standardized date format for EvidenceDate (e.g., 1857-06-24, 1888, 1904-07). |
| Inconsistent Categorical Classifications | Standardize category names across fields. |
| Missing Values | Null Entries | Complete the missing information as much as possible using other datasets |
| Data Quality & Completeness | Incomplete Dates | Complete the missing Dates as much as possible using other datasets |
| Lack of Geospatial Data | Enhance geospatial accuracy by including latitude and longitude coordinates where possible |
| Storage & Integration | Data Structure | Storing the data in a structured format like SQL or cloud-based storage for better accessibility and scalability. |

**2.1.3. ammc-shipstrike-2022-09-19\_4385**

**Dataset Name:** ammc-shipstrike-2022-09-19\_4385

**1. Data overview**

* Source or Provider: International Whaling Commission (IWC)
* File Type: Excel (.xlsx)
* Number of Records: 20
* Number of Columns: 116

**2. Description**

**Purpose:** The dataset primarily records interactions between vessels and whales or other marine organisms, including vessel strikes, whale sighting reports, event time and location, witness information, and vessel types. It contains 116 attributes, covering event ID, geographic location, timestamps, vessel types, vessel strike types, and evidence submission dates, among others. The core objective is to study the impact of maritime activities on marine life and provide scientific support for marine conservation policies and maritime management.

**Context:** This dataset falls within the fields of marine ecological conservation, whale research, and maritime safety management and is collected by the International Whaling Commission (IWC). It covers multiple marine regions, including the South Pacific Ocean, South Indian Ocean, Tasman Sea, and Timor Sea, with a primary focus on interactions between vessels and whales or other marine organisms, particularly vessel strikes. The collection of this data helps assess the impact of maritime activities on marine ecosystems and supports the development of relevant laws, regulations, and conservation measures.

**Use Case:**

1. Marine Conservation Policies: Analyze whale-vessel collision patterns to support governments and environmental organizations in developing navigation restrictions, low-speed zones, and avoidance strategies to reduce the threat posed by vessels to whales.
2. Whale Research: Utilize collision data to study the movement patterns of different whale species (e.g., sperm whales) in specific sea areas and their reactions to vessels.
3. Data Visualization and Analysis: Use maps and time-series analysis to identify high-risk areas, predict future collision-prone regions, and propose mitigation measures.
4. Historical Data Modeling: Apply machine learning techniques (e.g., time-series forecasting, classification models) to analyze the key factors contributing to whale collisions, aiming to improve maritime management and conservation strategies.
5. Marine Incident Investigation: Provide historical case studies for marine biologists and policymakers to assess whether vessel collisions have led to declines in whale populations and to explore potential countermeasures.

**3. Data details**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| **SHIPSTRIKE\_ID** | int64 | Unique identifier |
| **SHIPSTRIKE\_TYPE\_ID** | int64 | Unique identifier |
| SHIPSTRIKE\_TYPE | object | Type of shipstrike |
| **EVENT\_ID** | int64 | Unique identifier |
| **WHEN\_LOCATION\_ID** | float64 | Unique identifier |
| WHEN\_LOCATION | object | Location information |
| **WHEN\_WITNESS\_CONTACT\_ID** | float64 | Unique identifier |
| WHEN\_WITNESS\_CONTACT\_FIRST\_NAME | object | Witness Contact Info |
| WHEN\_WITNESS\_CONTACT\_LAST\_NAME | object | Witness Contact Info |
| WHEN\_LAST\_PORT | object | Last Departure Port |
| WHEN\_DEPARTURE\_DATE | datetime64[ns] | Departure Date |
| WHEN\_DESTINATION\_PORT | float64 | Destination Port |
| WHEN\_ARRIVAL\_DATE | datetime64[ns] | Arrival Date |
| WHEN\_ROUTE | object | Navigation Route |
| **SPECIES\_IDENTIFIED\_CONTACT\_ID** | float64 | Unique identifier |
| SPECIES\_IDENTIFIED\_CONTACT\_FIRST\_NAME | object | Species Identification Contact info |
| SPECIES\_IDENTIFIED\_CONTACT\_LAST\_NAME | object | Species Identification Contact info |
| SPECIES\_LOCAL\_NAME | float64 | Species Local Name |
| SPECIES\_CHARACTERISTICS | object | Species Characteristics |
| WHAT\_SEEN\_BEFORE | float64 | What was seen before the collision |
| WHAT\_SEEN\_TIME\_COLLISION | float64 | What was seen at the time of collision |
| WHAT\_SEEN\_OTHER\_CETACEANS | float64 | Were other cetaceans seen |
| WHAT\_SEEN\_GROUP\_SIZE | float64 | Size of the observed whale group |
| WHAT\_HAPPENED\_INDICATION\_DESC | object | Incident description |
| WHAT\_HAPPENED\_TO\_WHALE\_DESC | object | What happened to the whale |
| WHAT\_HAPPENED\_BLOOD\_SEEN | float64 | Was blood seen |
| WHAT\_HAPPENED\_WHALE\_SEEN\_AFTER | float64 | Was the whale seen after the incident |
| WHAT\_HAPPENED\_WHALE\_INJURIES | object | Whale injuries |
| WHAT\_HAPPENED\_DOA\_INDICATIONS | object | Records whether there were signs that the whale was already dead when discovered |
| WHAT\_HAPPENED\_PROPELLER\_DESC | object | Propeller-related description |
| WHAT\_HAPPENED\_DEAD\_BEFORE\_DESC | object | Was the whale dead before the incident |
| WHAT\_HAPPENED\_WAS\_REPORTED | object | Was the incident reported |
| WHAT\_HAPPENED\_ALERTED\_STUCK | object | Was there an alert about the whale being stuck |
| WHAT\_HAPPENED\_POSITION\_STUCK | object | Position where the whale was stuck |
| WHAT\_HAPPENED\_BEHAVIOUR\_AFTER | object | Whale behavior after the incident |
| **WEATHER\_VISIBILITY\_ID** | float64 | Unique identifier |
| VISIBILITY | float64 | Visibility |
| WEATHER\_LIGHT\_CONDITIONS\_ID | float64 | Weather Light Conditions ID |
| WEATHER\_LIGHT\_CONDITIONS | float64 | Weather Light Conditions |
| WEATHER\_WATER\_DEPTH | float64 | Water Depth |
| WEATHER\_BEAUFORT\_WIND\_FORCE | float64 | Beaufort Wind Force |
| WEATHER\_DESCRIPTION | float64 | Weather Description |
| **VESSEL\_ID** | float64 | Unique identifier |
| **VESSEL\_ACTIVITY\_DAMAGE\_ID** | float64 | Unique identifier |
| VESSEL\_DAMAGE | object | Vessel Damage |
| **VESSEL\_ACTIVITY\_COLLISION\_ID** | float64 | Unique identifier |
| VESSEL\_ACTIVITY\_COLLISION | object | Vessel Activity During Collision |
| VESSEL\_ACTIVITY\_STRIKE\_SPEED | float64 | Vessel Strike Speed |
| VESSEL\_ACTIVITY\_DAMAGE\_DESC | object | Vessel Damage Description |
| VESSEL\_ACTIVITY\_TOTAL\_WATCH | float64 | Total Vessel Observation Time |
| VESSEL\_ACTIVITY\_NAV\_DESC | object | Vessel Navigation Description |
| **WHALE\_DETAILS\_NECROPSY\_ID** | float64 | Unique identifier |
| WHALE\_DETAILS\_NECROPSY\_FIRST\_NAME | object | Whale Necropsy Examiner info |
| WHALE\_DETAILS\_NECROPSY\_LAST\_NAME | object |  |
| **WHALE\_DETAILS\_AGE\_ID** | float64 | Unique identifier |
| WHALE\_DETAILS\_AGE\_FIRST\_NAME | float64 | Whale Age Assessment Examiner info |
| WHALE\_DETAILS\_AGE\_LAST\_NAME | float64 |  |
| **WHALE\_DETAILS\_TISSUE\_ID** | float64 | Unique identifier |
| WHALE\_DETAILS\_TISSUE\_FIRST\_NAME | object | Whale Tissue Sample Collector info |
| WHALE\_DETAILS\_TISSUE\_LAST\_NAME | object |  |
| **WHALE\_DETAILS\_ESTIMATED\_ID** | float64 | Unique identifier |
| WHALE\_DETAILS\_ESTIMATED\_FIRST\_NAME | object | Whale Estimation Examiner info |
| WHALE\_DETAILS\_ESTIMATED\_LAST\_NAME | object |  |
| WHALE\_DETAILS\_CARCASS\_DESC | object | Whale Carcass Description |
| WHALE\_DETAILS\_SHORE\_DESC | object | Whale Stranding Description |
| WHALE\_DETAILS\_NECROPSY\_DESC | object | Whale Necropsy Description |
| IS\_DELETED | int64 | Is Deleted |
| **CREATED\_BY\_ID** | float64 | Unique identifier |
| CREATED\_DATE | datetime64[ns] | Created Date |
| **MODIFIED\_BY\_ID** | float64 | Unique identifier |
| MODIFIED\_DATE | datetime64[ns] | Modified Date |
| ROWID | object | Row ID |
| vessel\_activity\_propeller\_rpm | float64 | Vessel Activity Propeller RPM |
| vessel\_activity\_action\_to\_avoid | object | Vessel Action to Avoid Collision |
| event\_class | object | Event Class |
| platform\_type | object | Platform Type |
| **event\_id.1** | int64 | Unique identifier |
| sighting\_cue | float64 | Sighting Cue |
| species\_category | object | Species Category |
| species\_confidence\_level | float64 | Species Confidence Level |
| behaviour | float64 | Records the whale’s behavior during the event |
| scientific\_name | object | The scientific name of the whale |
| common\_name | object | The commonly used name of the whale |
| observation\_method | float64 | Observation Method |
| total\_count\_certainty | float64 | Total Count Certainty |
| iwc\_report | float64 | IWC Report |
| **guid** | object | Unique identifier |
| location | object | Location |
| location\_wkt | object | Location WKT |
| **organisation\_id** | int64 | Unique identifier |
| organisation\_name | object | Organisation Name |
| organisation\_code | object | Organisation Code |
| location\_description | object | Provides a detailed description of the event location |
| notes | float64 | Records additional remarks |
| vessel\_name | float64 | Vessel Name |
| total\_count | int64 | Records the total number of whales, sightings, or related events. |
| created\_by\_id.1 | float64 | Records the user ID of the person who created this data. |
| is\_approved | float64 | Indicates whether this record has been reviewed or approved |
| is\_secure | int64 | Indicates whether this record is protected |
| whaling\_type | float64 | Records the type of whaling activity |
| whaling\_country | float64 | Records the country conducting whaling activities |
| date\_observed | datetime64[ns] | Records the date when the whale was observed. |
| date\_reported | datetime64[ns] | Records the date when the incident was officially reported. |
| date\_hotline\_call | float64 | Records the date of a hotline report or inquiry related to the incident |
| date\_investigated | float64 | Records the date when the incident was officially investigated. |
| date\_assessed | float64 | Records the date when experts or relevant organizations assessed the incident. |
| date\_confirmed | float64 | Records the date when the incident or data was officially confirmed. |
| date\_submitted | datetime64[ns] | Records the date when data or reports were submitted. |
| date\_lodged | float64 | Records the date when the data or report was archived. |
| date\_unlodged | float64 | Records the date when the data or report was removed from the archive. |
| date\_verified | float64 | Records the date when the data or event information was verified. |
| date\_unverified | float64 | Records the date when the data or event was marked as unverified. |
| date\_compliance\_notified | float64 | Records the date when a compliance notification was issued regarding the event or data. |
| date\_caught | float64 | Records the date when the whale was caught |
| created\_date.1 | datetime64[ns] | Records the timestamp when the data entry was created. |
| modified\_date.1 | datetime64[ns] | Records the timestamp when the data entry was last modified |

**4. Data Issues and Recommendations**

**4.1 Inconsistent Data**

**4.1.1 Data Type Format**

**Issues:**

* Some ID fields such as **WHEN\_WITNESS\_CONTACT\_ID, WEATHER\_VISIBILITY\_ID, and VESSEL\_ID** are stored as float due to missing values.

**Recommendations:**

* Convert ID fields to integer format and ensure missing values are handled appropriately.

**4.1.2 Unclear Data Meaning**

**Issues:**

* Some fields contain values whose meanings are not explicitly defined, such as **IS\_DELETED(0)** and **is\_approved (0,1)**.

**Recommendations:**

* Provide documentation or a data dictionary to clarify the meanings of categorical values.

**4.1.3 Duplicate Unique Identifiers**

**Issues:**

* Some fields, such as EVENT\_ID and event\_id.1, appear to be duplicate identifiers.

**Recommendations:**

* Investigate whether both fields are necessary or if one should be removed.

**4.1.4 Inconsistent Categorical Classifications**

**Issues:**

* Some categorical fields (SPECIES\_CHARACTERISTICS, WHALE\_DETAILS\_CARCASS\_DESC) contain multiple names for the same category.

**Recommendations:**

* Standardize category names across these fields.

**4.2 Missing values**

**Observation:**

|  |  |  |
| --- | --- | --- |
| **Column** | **Missing Values**  **(Count)** | **Missing Values (%)** |
| **SHIPSTRIKE\_ID** | 0 | 0 |
| **SHIPSTRIKE\_TYPE\_ID** | 0 | 0 |
| SHIPSTRIKE\_TYPE | 0 | 0 |
| **EVENT\_ID** | 0 | 0 |
| **WHEN\_LOCATION\_ID** | 7 | 35 |
| WHEN\_LOCATION | 7 | 35 |
| **WHEN\_WITNESS\_CONTACT\_ID** | 18 | 90 |
| WHEN\_WITNESS\_CONTACT\_FIRST\_NAME | 18 | 90 |
| WHEN\_WITNESS\_CONTACT\_LAST\_NAME | 18 | 90 |
| WHEN\_LAST\_PORT | 18 | 90 |
| WHEN\_DEPARTURE\_DATE | 19 | 95 |
| WHEN\_DESTINATION\_PORT | 20 | 100 |
| WHEN\_ARRIVAL\_DATE | 19 | 95 |
| WHEN\_ROUTE | 17 | 85 |
| **SPECIES\_IDENTIFIED\_CONTACT\_ID** | 15 | 75 |
| SPECIES\_IDENTIFIED\_CONTACT\_FIRST\_NAME | 15 | 75 |
| SPECIES\_IDENTIFIED\_CONTACT\_LAST\_NAME | 15 | 75 |
| SPECIES\_LOCAL\_NAME | 20 | 100 |
| SPECIES\_CHARACTERISTICS | 12 | 60 |
| WHAT\_SEEN\_BEFORE | 20 | 100 |
| WHAT\_SEEN\_TIME\_COLLISION | 20 | 100 |
| WHAT\_SEEN\_OTHER\_CETACEANS | 20 | 100 |
| WHAT\_SEEN\_GROUP\_SIZE | 20 | 100 |
| WHAT\_HAPPENED\_INDICATION\_DESC | 14 | 70 |
| WHAT\_HAPPENED\_TO\_WHALE\_DESC | 7 | 35 |
| WHAT\_HAPPENED\_BLOOD\_SEEN | 20 | 100 |
| WHAT\_HAPPENED\_WHALE\_SEEN\_AFTER | 20 | 100 |
| WHAT\_HAPPENED\_WHALE\_INJURIES | 3 | 15 |
| WHAT\_HAPPENED\_DOA\_INDICATIONS | 15 | 75 |
| WHAT\_HAPPENED\_PROPELLER\_DESC | 13 | 65 |
| WHAT\_HAPPENED\_DEAD\_BEFORE\_DESC | 15 | 75 |
| WHAT\_HAPPENED\_WAS\_REPORTED | 16 | 80 |
| WHAT\_HAPPENED\_ALERTED\_STUCK | 19 | 95 |
| WHAT\_HAPPENED\_POSITION\_STUCK | 19 | 95 |
| WHAT\_HAPPENED\_BEHAVIOUR\_AFTER | 18 | 90 |
| **WEATHER\_VISIBILITY\_ID** | 20 | 100 |
| VISIBILITY | 20 | 100 |
| WEATHER\_LIGHT\_CONDITIONS\_ID | 20 | 100 |
| WEATHER\_LIGHT\_CONDITIONS | 20 | 100 |
| WEATHER\_WATER\_DEPTH | 20 | 100 |
| WEATHER\_BEAUFORT\_WIND\_FORCE | 20 | 100 |
| WEATHER\_DESCRIPTION | 20 | 100 |
| **VESSEL\_ID** | 20 | 100 |
| **VESSEL\_ACTIVITY\_DAMAGE\_ID** | 14 | 70 |
| VESSEL\_DAMAGE | 15 | 75 |
| **VESSEL\_ACTIVITY\_COLLISION\_ID** | 15 | 75 |
| VESSEL\_ACTIVITY\_COLLISION | 15 | 75 |
| VESSEL\_ACTIVITY\_STRIKE\_SPEED | 16 | 80 |
| VESSEL\_ACTIVITY\_DAMAGE\_DESC | 13 | 65 |
| VESSEL\_ACTIVITY\_TOTAL\_WATCH | 19 | 95 |
| VESSEL\_ACTIVITY\_NAV\_DESC | 19 | 95 |
| **WHALE\_DETAILS\_NECROPSY\_ID** | 19 | 95 |
| WHALE\_DETAILS\_NECROPSY\_FIRST\_NAME | 19 | 95 |
| WHALE\_DETAILS\_NECROPSY\_LAST\_NAME | 19 | 95 |
| **WHALE\_DETAILS\_AGE\_ID** | 20 | 100 |
| WHALE\_DETAILS\_AGE\_FIRST\_NAME | 20 | 100 |
| WHALE\_DETAILS\_AGE\_LAST\_NAME | 20 | 100 |
| **WHALE\_DETAILS\_TISSUE\_ID** | 19 | 95 |
| WHALE\_DETAILS\_TISSUE\_FIRST\_NAME | 19 | 95 |
| WHALE\_DETAILS\_TISSUE\_LAST\_NAME | 19 | 95 |
| **WHALE\_DETAILS\_ESTIMATED\_ID** | 19 | 95 |
| WHALE\_DETAILS\_ESTIMATED\_FIRST\_NAME | 19 | 95 |
| WHALE\_DETAILS\_ESTIMATED\_LAST\_NAME | 19 | 95 |
| WHALE\_DETAILS\_CARCASS\_DESC | 4 | 20 |
| WHALE\_DETAILS\_SHORE\_DESC | 18 | 90 |
| WHALE\_DETAILS\_NECROPSY\_DESC | 19 | 95 |
| IS\_DELETED | 0 | 0 |
| **CREATED\_BY\_ID** | 20 | 100 |
| CREATED\_DATE | 0 | 0 |
| **MODIFIED\_BY\_ID** | 20 | 100 |
| MODIFIED\_DATE | 0 | 0 |
| ROWID | 0 | 0 |
| vessel\_activity\_propeller\_rpm | 19 | 95 |
| vessel\_activity\_action\_to\_avoid | 16 | 80 |
| event\_class | 0 | 0 |
| platform\_type | 0 | 0 |
| **event\_id.1** | 0 | 0 |
| sighting\_cue | 20 | 100 |
| species\_category | 0 | 0 |
| species\_confidence\_level | 20 | 100 |
| behaviour | 20 | 100 |
| scientific\_name | 0 | 0 |
| common\_name | 0 | 0 |
| observation\_method | 20 | 100 |
| total\_count\_certainty | 20 | 100 |
| iwc\_report | 20 | 100 |
| **guid** | 0 | 0 |
| location | 0 | 0 |
| location\_wkt | 0 | 0 |
| **organisation\_id** | 0 | 0 |
| organisation\_name | 0 | 0 |
| organisation\_code | 0 | 0 |
| location\_description | 1 | 5 |
| notes | 20 | 100 |
| vessel\_name | 20 | 100 |
| total\_count | 0 | 0 |
| created\_by\_id.1 | 20 | 100 |
| is\_approved | 1 | 5 |
| is\_secure | 0 | 0 |
| whaling\_type | 20 | 100 |
| whaling\_country | 20 | 100 |
| date\_observed | 0 | 0 |
| date\_reported | 0 | 0 |
| date\_hotline\_call | 20 | 100 |
| date\_investigated | 20 | 100 |
| date\_assessed | 20 | 100 |
| date\_confirmed | 20 | 100 |
| date\_submitted | 0 | 0 |
| date\_lodged | 20 | 100 |
| date\_unlodged | 20 | 100 |
| date\_verified | 20 | 100 |
| date\_unverified | 20 | 100 |
| date\_compliance\_notified | 20 | 100 |
| date\_caught | 20 | 100 |
| created\_date.1 | 0 | 0 |
| modified\_date.1 | 0 | 0 |

**Issues:**

* A total of 27 fields have complete records (SHIPSTRIKE\_ID, SHIPSTRIKE\_TYPE\_ID, SHIPSTRIKE\_TYPE, EVENT\_ID, IS\_DELETED, CREATED\_DATE, MODIFIED\_DATE, ROWID, event\_class, platform\_type, event\_id.1, species\_category, scientific\_name, common\_name, guid, location, location\_wkt, organisation\_id, organisation\_name, organisation\_code, total\_count, is\_secure, date\_observed, date\_reported, date\_submitted, created\_date.1, modified\_date.1).
* A total of 42 fields contain only null values.

**Recommendations:**

* Consider deleting some fields that contain only null values based on their meanings.

**4.4 Data Quality and Completeness:**

**Issues:**

* The field **VESSEL\_ACTIVITY\_STRIKE\_SPEED** lacks a specified unit.

**Recommendation:**

* Define and document the unit of measurement.

**4.5 Data Storage & Integration**

**Issues:**

* The dataset appears to concatenate all data into a single table.
* Dataset is currently in Excel format, which may not be optimal for large-scale analysis.

**Recommendations:**

* Consider decomposing the dataset into multiple tables based on field meanings, using primary keys (PK) and foreign keys (FK) to establish relationships. Remove redundant fields to improve data integrity and query efficiency.
* Storing the data in a structured format like SQL or cloud-based storage for better accessibility and scalability.

**4.3 Data Sources Documentation**

* Data Source: From IWC
* Update Frequency: Not mentioned
* Storage Format: .xlsx
* API Availability: Not mentioned

**Issues:**

* The update frequency is not specified, making it difficult to track changes or assess data reliability.
* There is no API available for direct data retrieval.

**Recommendations:**

* Contact data providers to clarify the update frequency of this dataset.
* Explore alternative data sources or structured data retrieval processes (e.g., scheduled updates, web scraping if permitted).

**Summary of key issue and recommendation:**

|  |  |  |
| --- | --- | --- |
| **Category** | **Sub-category** | **Recommendations** |
| Inconsistencies | Formatting issues | Convert ID fields to integer format and ensure missing values are handled appropriately. |
| Unclear Field Meanings | Provide documentation or a data dictionary to clarify the meanings of categorical values. |
| Duplicate Unique Identifiers | Investigate whether both fields are necessary or if one should be removed |
| Inconsistent Categorical Classifications | Standardize category names across these fields |
| Missing Values | Null Entries | Consider deleting some fields that contain only null values based on their meanings. |
| Data Quality & Completeness | Missing Unit | Define and document the unit of measurement |
| Storage & Integration | Database Normalization | Consider decomposing the dataset into multiple tables based on field meanings, using primary keys (PK) and foreign keys (FK) to establish relationships. Remove redundant fields to improve data integrity and query efficiency. |
| Data Structure | Storing the data in a structured format like SQL or cloud-based storage for better accessibility and scalability. |

**2.2. IWC\_ShipStrikes\_PublicData**

**1. Data Overview**

* **Dataset Name:** IWC\_ShipStrikes\_PublicData
* **Source or Provider:** IWC
* **File Type:** Excel (.xlsx)
* **Number of Records:** 1,697
* **Number of Columns:** 9

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| CaseID | float64 | Identifier for each case |
| ReportID | float64 | Identifier for each report |
| DateInc | object | Incident date |
| LargeArea | object | Geographical area where the incident occurred |
| Latitude | object | Latitude coordinate of the incident |
| Longitude | object | Longitude coordinate of the incident |
| CommSpecies | object | Common species involved in the incident |
| Sex | object | Sex of the whale involved |
| AgeClass | object | Age classification of the whale |

**2. Description**

**Purpose:** This dataset tracks incidents involving ships and whales, documenting collisions, potential contributing factors, and demographic details of the affected whales.

**Context:** Understanding ship strikes is essential for whale conservation, maritime safety, and policymaking. This dataset can support research on the frequency of vessel-whale collisions, assess potential risk factors, and inform mitigation strategies to protect whale populations.

**Use Case: Conservation and marine protection** – Identifying high-risk areas for ship strikes; **Policy development** – Supporting regulations for reducing whale-ship collisions; **Data visualization and analysis** – Conducting spatial risk analysis to optimize shipping routes.

**3. Data details**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| CaseID | float64 | Unique identifier for each case |
| ReportID | float64 | Unique identifier for each report |
| DateInc | object | Date of the incident |
| LargeArea | object | General oceanic region where the incident occurred |
| Latitude | object | Latitude coordinate of the incident |
| Longitude | object | Longitude coordinate of the incident |
| CommSpecies | object | Common species of the affected whale |
| Sex | object | Sex of the whale |
| AgeClass | object | Age category of the whale |

**4. Data Issues and Recommendations**

**4.1 Inconsistent Data**

**4.1.1 Data Type Format**

**Issues:**

* **'DateInc'** is stored as an object (text) instead of a proper datetime format.
* **Latitude and Longitude** are stored as text rather than numerical values, affecting geospatial analysis.
* Categorical variables (**'LargeArea', 'CommSpecies', 'Sex', 'AgeClass'**) are stored as objects, which might require encoding for modeling.

**Recommendations:**

* Convert 'DateInc' to **datetime** format.
* Split 'LargeArea' into two separate columns: **Main Ocean/Sea Name** and **Specific Sub-Region**.
* Convert 'Latitude' and 'Longitude' to float64 for better geospatial analysis.
* Consider One-Hot Encoding for categorical columns if needed for modeling.

**4.1.2 Inconsistent Values**

**Issues:**

* Whale species names appear inconsistently. Some records indicate "Probable" species, while others use different naming conventions.
* Various formats for unknown values exist, including ‘**Unknown’, ‘Not Reported’, ‘Not Indicated’,** and blank values.

**Recommendations:**

* Standardize species naming conventions to ensure consistency.
* Unify unknown values into one standard label, such as "Unknown" or NaN for missing data.

**4.2 Missing values**

**Observation:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Missing Values （Count）** | **Missing Values (%)** |
| CaseID | 354 | 20.86 |
| ReportID | 1 | 0.06 |
| DateInc | 372 | 21.92 |
| LargeArea | 450 | 26.52 |
| Latitude | 1012 | 59.63 |
| Longitude | 1014 | 59.75 |
| CommSpecies | 839 | 49.44 |
| Sex | 764 | 45.02 |
| AgeClass | 892 | 52.56 |

**Issues:**

* Over 50% of missing values in key fields like **Latitude, Longitude, and AgeClass**, which may impact spatial analysis.
* CaseID has missing values, which could affect data linking and integrity.

**Recommendations:**

* Assess whether missing location data can be filled using external sources or geospatial interpolation.
* Verify the impact of missing **AgeClass** and **CommSpecies** on analysis.

**4.3 Data Sources Documentation**

* Data Source: From IWC
* Update Frequency: Not mentioned
* Storage Format: .xlsx
* API Availability: Not mentioned

**Issues:**

* The update frequency is not specified, making it difficult to track changes or assess data reliability.
* There is no API available for direct data retrieval.

**Recommendations:**

* Contact data providers to clarify the update frequency of this dataset.
* Explore alternative data sources or structured data retrieval processes (e.g., scheduled updates, web scraping if permitted).

**4.4 Data Storage & Integration**

**Observation:**

* File Size: 0.08 MB
* Potential Scalability Issues: While the file is small now, growth in records may require a relational database or cloud storage.
* CaseID and ReportID Relationships:
* Total Records: 1,697
* Unique CaseIDs: 1,135
* Duplicate CaseIDs: 562
* Missing CaseIDs: 354
* Unique ReportIDs: 1,696
* Duplicate ReportIDs: 1
* CaseID has a many-to-one relationship with ReportID.

**Issues:**

* Missing **CaseIDs** could indicate data integrity issues.

**Recommendations:**

* Consider using a relational database if data volume increases significantly.

**Summary of key issue and recommendations**

|  |  |  |
| --- | --- | --- |
| **Category** | **Sub-category** | **Recommendations** |
| Standardize Data Formats & Types | Date Format | Convert the format to Datetime format |
| Categorical Data | * May apply One-Hot Encoding if needed * Split the columns   For **'LargeArea'**:  -Split into Two Columns:Column 1: Main Ocean/Sea Name  Column 2: Specific Sub-Region  For **'CommSpecies'**:  -The situation ‘probable’ can be tag within One-Hot Encoding  -Standardize name of species |
| Missing Values | / | * Assess the impact of missing values * Use appropriate imputation techniques |
| Data Sources Documentation | Update Frequency | Contact data providers toclarify the update cycle. |
| API Availability | Consider developing a structured data retrieval process (e.g., scheduled updates or web scraping if permitted). |
| Data Storage & Integration | Data Integration | Ensure the uniqueness and completeness of Primary key. |

**2.3. USA\_sampledataset**

### 1. Data Overview

* **Dataset Name:** USA\_sampledataset.xlsx
* **File Type:** Excel (.xlsx)
* **Number of Records:** 291
* **Number of Columns:** 25
* **Schema Definition:**
  + **Data types:**
    - **int64:** Index, Event.Index, Source.Confirmed, Record.Used.Garrison2024, Record.Used.Conn.Silber.2013
    - **float64:** Vessel.Speed.knots, Interaction.Confirmed, Observed.Post.Interaction, Mortality.Injury.Observed, Followup.Uninjured
    - **object:** Event.Type, Date, Species, Country, Region, Location, Vessel.Length.meters, Vessel.Size.Category, Vessel.Type, Whale Sex, Whale Age, Notes, Fate, Fate.Individual, Data Sources
  + **Relationships:**
    - The primary key is likely the **Index** field, with **Event.Index** serving as an additional event identifier.

### 2. Description

* **Overview:**  
   This dataset contains records of individual events related to marine incidents (such as vessel strikes or whale interactions) with details including date, species, location, vessel characteristics, and outcomes.
* **Context:**  
   Data is collected for research purposes in marine incident analysis. It is used in conservation studies, impact assessments, and wildlife monitoring.
* **Use Cases:**  
   Intended for statistical analysis, data quality assessment, and integration into larger data systems (e.g., data warehouses) for marine conservation research and policymaking.

### 3. Data Details

**Columns and Meanings:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Format** | **Description** |
| **Index** | int64 | Unique record identifier. |
| **Event.Index** | int64 | Identifier for the event. |
| **Event.Type** | object | Type of event (e.g., "Individual"). |
| **Date** | object | Event date; inconsistent formats observed (e.g., "1961-09-01 00:00:00" vs. "5/1/1885"). |
| **Species** | object | Whale species information (mostly "Unid." indicating unknown species). |
| **Country** | object | Country where the event occurred (e.g., "UNK", "Cook Islands", "U.S."). |
| **Region** | object | Geographic region of the event. |
| **Location** | object | Specific location details. |
| **Vessel.Speed.knots** | float64 | Vessel speed in knots. |
| **Vessel.Length.meters** | object | Vessel length in meters; may include numeric values or "UNK" for unknown. |
| **Vessel.Size.Category** | object | Category of vessel size. |
| **Vessel.Type** | object | Type of vessel (only 1 missing value observed). |
| **Whale Sex** | object | Sex of the whale. |
| **Whale Age** | object | Age information for the whale. |
| **Notes** | object | Additional notes (~5.5% missing). |
| **Fate** | object | Outcome of the event. |
| **Fate.Individual** | object | Individual outcome details. |
| **Source.Confirmed** | int64 | Indicator if the source is confirmed. |
| **Interaction.Confirmed** | float64 | Indicator of confirmed interaction (~4.8% missing). |
| **Observed.Post.Interaction** | float64 | Post-interaction observation value (~6.5% missing). |
| **Mortality.Injury.Observed** | float64 | Observation of mortality/injury (~25% missing). |
| **Followup.Uninjured** | float64 | Follow-up on uninjured individuals; extremely high missing rate (~96.2%). |
| **Record.Used.Garrison2024** | int64 | Record indicator for Garrison2024. |
| **Record.Used.Conn.Silber.2013** | int64 | Record indicator for Conn Silber 2013. |
| **Data Sources** | object | Data source references (fully populated). |

### 4. Data Issues and Recommendations

#### *Inconsistencies*

* **Issue:**
  + The **Date** field shows inconsistent formats (e.g., "1961-09-01 00:00:00" vs. "5/1/1885").
  + **Recommendation:** Convert the Date field to a uniform datetime format (e.g., YYYY-MM-DD).
* **Issue:**
  + **Vessel.Length.meters** is stored as an object and may contain non-numeric values such as "UNK".
  + **Recommendation:** Convert this field to a numeric type where possible and treat non-numeric entries as missing.

#### *Missing Values*

* **Observation:**
  + **Interaction.Confirmed:** ~4.8% missing
  + **Observed.Post.Interaction:** ~6.5% missing
  + **Mortality.Injury.Observed:** ~25% missing
  + **Followup.Uninjured:** ~96.2% missing
* **Recommendation:**
  + For fields with low to moderate missing values, consider imputation methods (mean/median for numeric fields, mode for categorical fields) as appropriate.
  + For **Followup.Uninjured**, given the extremely high missing rate, consider dropping this field or investigating further if it is critical for analysis.

#### *Data Quality & Completeness*

* **Evaluation:**  
   Overall, most fields have low missing rates and good completeness; however, the extremely high missing rate in **Followup.Uninjured** may bias any analysis if included.
* **Recommendation:**
  + Review the necessity of fields with very high missing rates.
  + Assess potential biases and decide whether to exclude such fields from further analysis.

#### *Storage & Integration*

* **Current State:**  
   The dataset is stored as an Excel file (.xlsx).
* **Recommendation:**
  + For scalability and better integration into a data warehouse, consider migrating the data to a SQL database or a cloud-based storage solution.
  + Ensure that a unique key (e.g., **Index**) is maintained for seamless integration with other datasets.

### Summary of Issue Categories and Recommendations

|  |  |  |
| --- | --- | --- |
| **Issue Category** | **Sub-Category** | **Recommendation** |
| **Format** | Date and Numeric Fields | Standardize the Date field to a uniform datetime format; convert Vessel.Length.meters to numeric, treating non-numeric entries as missing. |
| **Missing Data** | Followup.Uninjured | Consider dropping the Followup.Uninjured field due to its ~96% missing rate, or investigate imputation methods if the field is deemed essential. |
| **Data Consistency** | Categorical Fields | Standardize text cases and labels (e.g., unify "UNK" to "Unknown") to ensure consistency across the dataset. |
| **Storage** | File Format | Consider migrating from Excel to a SQL database or cloud storage solution for improved scalability and integration capabilities. |

**2.4. Marine strike log – Share – IWC**

### **1. DATA OVERVIEW**

· **Dataset Name:** Marine strike log – Share – IWC

· **Version Number:** v1.0

· **Date of Creation and Last Update:** Unknown (historical records span 1819 - 2022)

· **Source or Provider:** International Whaling Commission (IWC)

· **File Type:** Excel (.xslx)

· **Number of Records:**

3 worksheets:

‘Piv - All’ : 67 records

‘Strike Log’: 1322 records

‘List’: 27 records

· **Number of Columns:**

‘Piv - All’: 24 columns

‘Strike Log’: 68 columns

‘List’: 27 columns

· **Schema Definition:**

o Data types : String, Integer, Float , Categorical

o Relationships: Species IDs and Injury descriptions appear in both ‘Strike Log’ and ‘List’. ‘Piv - All’ is a pivoted summary based on raw strike log data

### **2. DESCRIPTION**

· **Overview:** The Marine Strike Log dataset documents interactions between marine vessels and fauna (e.g. whales, sharks, sunfish), including proximity incidents, physical strikes, and injuries. It tracks both quantitative (counts) and qualitative (descriptions) data about each incident. Three worksheets are present. ‘Strike log’ gives raw, detailed strike-level data (primary source). ‘Piv-All’ provides a summary view of aggregated insights and ‘List’ is a reference table for species/ID/injury descriptions.

· **Context:** This data is collected by the International Whaling Commission to monitor the impact of marine traffic and incidents, as well as to inform conservation policies and maritime safety regulations.

· **Use Cases:**

* Trend analysis of marine strikes over time and regions.
* Identification of high-risk zones and vessel types
* Species-specific risk modelling and conservation strategy
* Evaluation of injury severity to fauna and vessel damage outcomes

### **3. DATA DETAILS**

### **A. Pivot Summary (‘Piv - All’)**

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| Year | Object | Year of recorded data summary |
| Count of Proximity | Float64 | Total numbers of incidents that year |
| Proximity | Object | Type of event (e.g. Collision, Observation) |
| Count of Proximity.1 | Float64 | Number of that specific proximity type |
| Ocean | Object | Region or ocean body |
| Count of Proximity.2 | Float64 | Number of incidents by ocean |
| Type of Position | Object | GPS/Ocean area etc. |
| Count of Proximity.3 | Float64 | Count by type of position |
| Boat type | Object | Vessel type involved |
| Count of Proximity.4 | Float64 | Count per boat type |
| What was hit/Seen | Object | Object hit or fauna seen |
| Count of Proximity.5 | Float64 | Count for hit object/animal |
| ID | Object | Common Species name |
| Count of Proximity.6 | Float64 | Count by species |
| Injury to fauna | Object | Description of injury outcome |
| Count of Proximity.7 | Float64 | Count per injury type |

**B. Detailed Records (‘Strike Log’)**

This sheet contains parent columns explaining how each sub-column contributes to the strike log. The table below describes each column listed.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| Document version | Unknown - Spreadsheet shows #REF! error | Indication for user of current version of the dataset |
| Editor | Object | Editor of current version |
| Total | Int64 | Number of strike logs |
| Marine Strike log code | Object | Used to log specific marine strikes in the following format:  *‘SkipperBoat(Or IWC Code)-time-date-No’* |
| ORG CODE | Object | Code to identify which Organisation strike log is related to |
| CODE | Object | First column with unique ID code |
| CODE | Int64 | Second column with unique ID code |
| Surveyor Contact name | Object | Name of surveyor related to specific log |
| Contact | Object | Contact info for surveyor |
| Time UTC | Object | Time of strike log |
| Day | Int64 | Day of strike log |
| Month | Object | Month of strike log |
| Year | Int64 | Year of strike log |
| 1st, 2nd, 3rd party obvs | Int64 | Indication of who observed strike |
| Source | Object | Origin of Strike log |
| Confirmed | Object | Name of confirming party |
| Survey Date | Object | Date survey carried out |
| Confidential | Object | Classifying whether strike is ‘Open’ or ‘Anonymous’ |
| Source Link | Object | Reference, Report, or link to encounter |
| Source Link 2 | Object | Second Reference, Report, or link to encounter |
| Skipper | Unknown - Empty Column | Possible name of skipper |
| Boat name/MMSI | Unknown - Empty Column | Name of boat/Vessel |
| Boat type | Object | Type of Boat |
| Event | Unknown - Empty Column | Type of Event |
| Event leg | Object | Location leg of specific strike event |
| Ocean | Object | Ocean where event occurred |
| Ocean area | Object | Specific area of ocean where event occurred |
| Type of Position | Object | Classification of position logged |
| GPS Lat | Float64 | Latitude of Incident |
| GPS Long | Float64 | Longitude of Incident |
| Ref Source Dead Reckoning Position Lat | Float64 | Unknown - Empty Column |
| Ref Source Dead Reckoning Position Long | Float64 | Unknown - Empty Column |
| Distance to Dead reckoning position | Int64 | Distance based on previously known location, course and speed. |
| Bearing to Dead Reckoning position | Int64 | Bearing based on previously known location, course and speed. |
| Radius of Dead Reckoning circle | Int64 | Radius based on previously known location, course and speed. |
| Dead Reckoning Lat | Float64 | Latitude based on previously known location, course and speed. |
| Dead Reckoning Long | Float64 | Longitude based on previously known location, course and speed. |
| Estimated zone | Unknown - Empty Column | GeoJson for estimated position zone |
| Number incidents leg | Int64 | How many times did the boat have an incident on this race leg |
| Proximity | Object | Type of incident, e.g. attack, collision, close cal, etc. |
| Speed | Int64 | Speed in Kts |
| Amount of damage | Object | Description of severity of damage to vessel, abandon or performance impaired |
| Type of damage to boat | Object | Description of type of damage to vessel |
| Injured crew | Object | Description of injuries to crew |
| Seen before | Boolean | Indication if seen before incident |
| Seen after | Boolean | Indication if seen after incident |
| Injury to fauna | Object | Description of injury to fauna |
| What was hit/seen | Object | Observation or estimation of what was hit/seen |
| ID | Object | Confirmed exact ID of what was hit |
| Photo | Empty Column - Presumably jpg, pdf | Visual record of incident |
| Reported to IWC | Boolean | Whether reported by IWC |
| Tech | Object | What tech is onboard for collision avoidance |
| Tech/on/used | Object | Indication if tech worked or not |
| Notes | Object | Miscellaneous notes on each incident |
| Quotes | Object | Quotes from those on board or observing incident |
| Other1 | Object | Any other notable info |
| Other2 | Object | Any other notable info |
| Other3 | Object | Any other notable info |

**C. Reference List (‘List’)**

This table provides a reference for the larger marine strike log worksheet above, indicating the possible values that may fill certain column rows.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| Party | Int64 | Indication of whether 1st, 2nd |
| Proximity | Object | Shows different possible labels for proximity |
| Large area | Object | Provides list of areas in Oceans to be used in strike log |
| Type | Object | Classification (e.g. Sharl, Cetacean, Megafauna) |
| Whale Names | Object | Scientific name |
| ID | Object | Common name (e.g. Fin whale) |
| Damage to vessel | Object | Recorded consequence to ship |
| Injured crew | Object | Type of injury crew sustained |
| Fauna injury | Object | Injury description for the animal |
| Confidential | Object | Types of confidentiality for classification |
| Tech | Object | Types of tech possibly utilised |
| Tech worked | Object | Vocabulary used to determine how and if tech succeeded |
| Y/N | Boolean | Yes = TRUE  No = FALSE |
| Full or partial survey | Object | Indication of whether survey is ‘Full’ or ‘Partial’ |
| Position | Object | Types of position used in data log, e.g. estimated zone, dead reckoning position, GPS, etc. |

### **4. DATA ISSUES AND RECOMMENDATIONS**

·

|  |  |  |  |
| --- | --- | --- | --- |
| **Issue Category** | **Sub-Category** | **Issues Identified** | **Recommendation** |
| Inconsistencies | Column naming | Unnamed columns, repeating headers (e.g. Count of Proximity.1, 2…) | Rename columns with descriptive labels |
| Inconsistencies | Format | Strike Log has wide-format export with numeric column names | Reshape to long-format table |
| Missing Values | No/Null values | Many null or vague entries like “Don’t know”, “Unknown” in following columns:  Fauna injury, ID, photo, estimated zone, event, boat name/MMSI, skipper | Use standardised categories like “Unknown” or “Unverified”.  Consider removal of any unnecessary columns with no data logged at all |
| Completeness | Data Coverage | Some years and regions are underrepresented | Mark coverage gaps for future data collection |
| Storage & Integration | File Consistency | Encoding Issues and irregular delimiters in ‘Strike Log’ sheet | Convert to UTF-8, validate schema, store in SQL or cleaned Excel format |
| Integration | Schema linkage | ID and injury descriptions vary slightly between worksheets | Link tables via consistent ID and injury-type keys |
|  |  |  |  |

**3. Bycatch Data Sources**

Data Overview：

* **Dataset Name:** IWC\_LargeWhaleBycatch.csv / ICW\_SmallCetaceanBycatch.csv
* **Version Number:** [Not Specified]
* **Date of Creation and Last Update:** [Not Specified]
* **Source or Provider:** International Whaling Commission (IWC)
* **File Type:** .csv
* **Number of Records:** 1001
* **Number of Columns:** 29
* **Schema Definition:**
  + **Data types:**
    - Int64: id, Data Year, Year Submitted
    - Float: Total Females: Dead, Total Females: Seriously Injured, Total Females: Injured, Total Females: Unknown, Total Males: Dead, Total Males: Seriously Injured, Total Males: Injured, Total Males: Unknown, Total Unknown: Dead, Total Unknown: Seriously Injured, Total Unknown: Injured, Individuals - "Fate" Unknown: Unknown, Individuals data are available?
    - Object: Large Area, Species, Country, Local Area, Local Taxonomy, Local Area (Long/Lat), Targeted fishery species, Fishing Gear, Other Fishing Gear, How Observed, Contacts, References, Explanation
  + **Relationships:**
    - Primary key: id
    - Foreign key: Large Area, Species, Total Females,Total Males

Description

Overview

This dataset contains 1001 data records related to the bycatch of small cetaceans. The records cover basic information such as the data year, submission year, large area, species, and country, as well as detailed records of the casualties of small cetaceans by gender (female, male, unknown) in bycatch incidents, as well as information about the target species of the fishery, the type of fishing gear, the method of observation, the contact information, and the reference information.

Context

Judging from the content of the data, the dataset can be used for research on marine environmental protection and fisheries management. By collecting these data, the bycatch of small cetaceans in different regions and fisheries can be analyzed, providing a scientific basis for the formulation of sound fisheries policies and the protection of small cetacean species.

Use Cases

* The data can be used to count the number of small cetaceans by-catches, species distributions and mortalities in different regions and years to assess the impact of fishing activities in different regions on small cetaceans.
* A data quality assessment will be conducted to verify the integrity, accuracy and consistency of the data, which will provide a basis for subsequent, more in-depth data analysis.
* The data can be integrated into a larger marine ecological data system to support comprehensive marine ecological research and policy formulation.

Data details

|  |  |  |
| --- | --- | --- |
| **Column name** | **Data type** | **Description** |
| id | int64 | Unique identifier |
| Data Year | int64 | Year of incident |
| Year Submitted | DATETIME | Year in which data were submitted to the database |
| Large Area | object | Broad geographic area in which the incident occurred |
| Species | object | Species of the cetacean |
| Country | object | Country in which the incident occurred |
| Local Area | object | Specific geographic location |
| Local Taxonomy | object | Local classification information for cetaceans |
| Local Area (Long/Lat) | object | geographic coordinate |
| Total Females: Dead | float64 | Whale Casualty Status by Sex |
| Total Females: Seriously Injured | float64 | Whale Injured status by Sex |
| Total Females: Injured | float64 | Whale Injured status by Sex |
| Total Females: Unknown | float64 | Whale Casualty Status by Sex |
| Total Males: Dead | float64 | Whale Casualty Status by Sex |
| Total Males: Seriously Injured | float64 | Whale Injured status by Sex |
| Total Males: Injured | float64 | Whale Injured status by Sex |
| Total Males: Unknown | float64 | Whale Casualty Status by Sex |
| Total Unknown: Dead | float64 | Whale Casualty Status by Sex |
| Total Unknown: Seriously Injured | float64 | Whale Injured status by Sex |
| Total Unknown: Injured | float64 | Whale Injured status by Sex |
| Total Unknown: Unknown | float64 | Whale status by Sex |
| Individual data are available from the contact below. | object | Marking of availability for contact for further data |
| Targeted fishery species | object | Fisheries target species |
| Fishing Gear | object | Type of fishing gear used |
| Other Fishing Gear | object | Type of fishing gear used |
| How Observed | object | observer identity |
| Contacts | object | contact information |
| References | object |  |
| Explanation | object |  |

Missing value

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Missing Values(Count)** | **Missing Values(%)** |
| id | 0 | 0 |
| Data Year | 0 | 0 |
| Year Submitted | 0 | 0 |
| Large Area | 0 | 0 |
| Species | 26 | 2.6 |
| Country | 0 | 0 |
| Local Area | 466 | 46.55 |
| Local Taxonomy | 877 | 87.61 |
| Local Area (Long/Lat) | 667 | 66.63 |
| Total Females: Dead | 178 | 17.78 |
| Total Females: Seriously Injured | 215 | 21.48 |
| Total Females: Injured | 215 | 21.48 |
| Total Females: Unknown | 211 | 21.08 |
| Total Males: Dead | 169 | 16.88 |
| Total Males: Seriously Injured | 215 | 21.48 |
| Total Males: Injured | 215 | 21.48 |
| Total Males: Unknown | 212 | 21.18 |
| Total Unknown: Dead | 95 | 9.49 |
| Total Unknown: Seriously Injured | 196 | 19.58 |
| Total Unknown: Injured | 206 | 20.58 |
| Individuals - "Fate" Unknown: Unknown | 183 | 18.28 |
| Individuals data are available? | 1001 | 100 |
| Targeted fishery species | 567 | 56.64 |
| Fishing Gear | 51 | 5.09 |
| Other Fishing Gear | 987 | 98.6 |
| How Observed | 94 | 9.39 |
| Contacts | 102 | 10.19 |
| References | 770 | 76.92 |
| Explanation | 478 | 47.75 |

Data issues and recommendations

|  |  |  |  |
| --- | --- | --- | --- |
| **Issue Category** | **Sub - Category** | **Issues Identified** | **Recommendation** |
| Data Completeness | High Missing Rate Fields | Some fields have a large number of missing values, such as Local Taxonomy(only 124 non-empty values), Other Fishing Gear(only 14 non-empty values), | Communicate with the data collector to try to supplement the data, delete fields that have little impact, make reasonable assumptions, or fill in important fields. |
| Data Completeness | Low Missing Rate Fields | Persons data available?(0 non-empty values), etc.; many fields have different degrees of missing values, such as Species with a missing rate of around 2.6%, Local Area with a missing rate of around 46.6% | Consult the information or contact the vendor to complete missing values |
| Data Quality | Overall Data Quality | There are many missing values in the overall data, which may bias the analysis based on complete data. In particular, the lack of some key information (such as geographic location and target species) may affect the comprehensive assessment of small cetacean bycatch | Establish a data quality review mechanism to reduce the impact of missing values by pre-processing the data. |
| Storage | File Format | The data is stored as a CSV file. | Migrate to database storage and build a data warehouse for data integration. |