



# Trabajo Práctico 01

## Limpieza de datos

13 de noviembre de 2024

Calidad de Datos

**Grupo : 12**

Integrante	LU	Correo electrónico
Navarro, Solana	906/22	solanan3@gmail.com
Suarez, Ines	890/22	ine.suarez22@gmail.com
Wittmund Montero, Lourdes	1103/22	lourdesmonterochiara@gmail.com



**Facultad de Ciencias Exactas y Naturales**  
Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (+54 +11) 4576-3300

<http://www.exactas.uba.ar>

# 1. Introducción

La página web Shark Attack File lleva un registro detallado y actualizado de los incidentes ocurridos por ataques de tiburones alrededor del mundo. Esta base de datos recopila información relevante sobre los ataques, proporcionando un archivo accesible de incidentes que puede descargarse para su análisis. El log de incidentes disponible en el sitio contiene los siguientes datos, entre otros:

Columna	Informacion
Case #	ID del ataque
Date	Fecha en que ocurrió el ataque
Country	País en donde ocurrió el ataque
Area	Provincia/estado/zona donde ocurrió el ataque
Location	Ciudad/localidad mas cercana
Type	Tipo de accidente
Activity	Actividad que estaba haciendo la víctima
Name	Nombre de la víctima
Sex	Sexo de la víctima
Age	Edad de la víctima
Injury	Tipo de daños que sufrió la víctima
Time	Momento del día en el que ocurrió el ataque
Species	Especie de tiburón involucrado
Source	Investigador que reporto el ataque o fuente de donde se conocen los datos del ataque

Tabla 1: Informacion del dataset

Los accidentes los vamos a catalogar en 5 tipos distintos:

**Unprovoked vs Provoked** - El GSAF define un incidente como **Provoked** cuando el tiburón fue herido con una lanza, enganchado, capturado o cuando un humano causó el "primer sangrado". Aunque estos incidentes no son de gran interés para los expertos en comportamiento de tiburones, cuando se conoce la especie involucrada y se dispone de fotos previas a la lesión, los patrones de mordeduras resultan útiles para determinar la especie de tiburón en otros casos donde no se pudo identificar la especie por parte de la víctima o testigos. Sabemos que un tiburón rara vez percibe a un humano vivo como una presa. Muchos incidentes son impulsados por curiosidad, otros pueden ocurrir cuando el tiburón percibe al humano como una amenaza o un competidor por una fuente de alimento, y podrían clasificarse como **Provoked** si se observa desde la perspectiva del tiburón. Los **Unprovoked** se encuentran en color "Tan" mientras que los **Provoked** en naranja.

**Warcraft/Attacks on boats** – Los incidentes en los que un tiburón muerde o embiste una embarcación están marcados en verde. Sin embargo, en los casos en los que el tiburón fue enganchado, atrapado en redes o golpeado, la entrada se marca en naranja, ya que se clasifican como **Provoked**.

**Air/Sea Disasters** - Los tiburones mantienen el equilibrio del ecosistema marino al eliminar a los animales muertos o heridos. Muchos incidentes ocurren porque, al igual que otros animales que no dependen solo del instinto, los tiburones exploran su entorno. Al no tener manos, pueden investigar un objeto desconocido con sus bocas. A diferencia de los humanos, los tiburones no actúan con malicia; simplemente hacen lo que la naturaleza les ha diseñado para hacer. Los desastres aéreos/marítimos son accidentes que colocan a las personas en el territorio diario de los tiburones. Las pérdidas humanas por tiburones durante las guerras resultan de la crueldad del ser humano hacia otro ser humano. Los **Sea Disasters** se encuentran en amarillo.

**Questionable** - Incidentes en los que no se dispone de suficientes datos para determinar si la lesión fue causada por un tiburón o si la persona se ahogó y su cuerpo fue posteriormente devorado por tiburones. En algunos casos, a pesar de los informes de los medios, la evidencia indicó que no hubo ninguna intervención de tiburones en absoluto. Estos incidentes se marcan en azul.

En este informe, analizaremos los datos provenientes de este archivo con el objetivo de obtener una comprensión más profunda de las tendencias y patrones relacionados con los ataques de tiburones.

A continuación, procederemos a analizar y limpiar los datos para asegurar la calidad y precisión de la información disponible. Nuestro objetivo es obtener una visión clara de la cantidad de incidentes clasificados por tipo y la distribución de estos incidentes en distintos países.

Mediante este enfoque, podremos conocer no solo cuántos incidentes pertenecen a cada categoría, sino también cómo se distribuyen geográficamente, observando patrones regionales o posibles concentraciones en países específicos.

## 2. Análisis

Para este análisis, comenzaremos realizando una descripción detallada de las columnas de interés, específicamente "Type" y "Country". En primer lugar, analizaremos cada columna de forma individual, observando los valores presentes en cada una.

- **Type:** Este atributo contiene información sobre el tipo de accidente registrado. Como se mencionó previamente, existen cinco tipos oficialmente definidos en el dataset. Los registros que corresponden a alguno de estos tipos serán considerados como válidos o correctos. Estos valores representan, lógicamente, la mayoría de los datos en la columna. Sin embargo, también existen algunos registros que no se ajustan a esta categorización y, por lo tanto, no cumplen con los tipos establecidos.

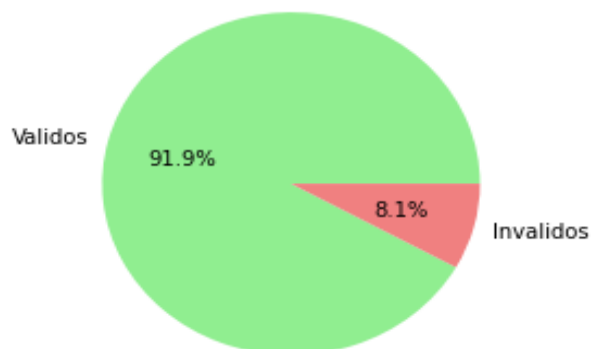


Figura 1: Porcentaje de datos que cumplen o no con la categorización

En el gráfico de torta se observa que el 8.1 % de los datos no corresponden a ninguno de los cinco tipos principales, conteniendo valores inválidos o vacíos. Esto demuestra la importancia de contar con datos limpios y correctamente clasificados. Si bien un 8.1 % puede parecer una proporción relativamente baja, en datasets amplios esta cantidad de datos incorrectos puede impactar en el análisis general, ya que representa una porción significativa de casos que no se pueden clasificar dentro de los tipos definidos. Este porcentaje de valores incorrectos o vacíos puede llevar a una interpretación sesgada o limitar la precisión al comparar la incidencia entre los tipos de accidente.

Existen varias posibles causas para estos datos inconsistentes: errores en el ingreso, formatos inadecuados, datos incompletos o incluso ambigüedad en la clasificación original. Reconocer estas causas es fundamental, ya que revela áreas que podrían optimizarse en el proceso de recopilación de datos y ayuda a comprender las limitaciones del análisis actual.

Vamos a ver mejor que ocurre con estos tipos distintos.



Figura 2: Distribución de datos inválidos

Al observar este gráfico, destaca que la mayoría de los datos que no pertenecen a las cinco categorías principales de tipos de accidentes están etiquetados como Invalid (94.5 %), mientras que el resto representa un porcentaje mucho menor, con valores null o vacíos (3.3 %) y una pequeña proporción de otros términos como Unverified o palabras con significados similares. Dada la alta proporción de registros con la etiqueta Invalid, decidimos estudiar estos casos manualmente para identificar patrones o posibles razones detrás de esta categorización.

En nuestro análisis, descubrimos que una gran cantidad de los registros etiquetados como Invalid incluían en la columna "Species" descripciones del estilo "no se sabe si hubo presencia de tiburón antes de la muerte" o expresiones similares las cuales se ajustan a la definición de los casos clasificados como **Questionable**, una de las cinco categorías válidas en el dataset. Esto sugiere que varios de estos registros etiquetados como Invalid podrían en realidad reclasificarse bajo la categoría **Questionable**, lo que mejoraría la precisión y coherencia del dataset. El resto de ellos dejaba en duda la mera presencia de un tiburón en el accidente, generándonos la pregunta de si realmente debían estar en este dataset.

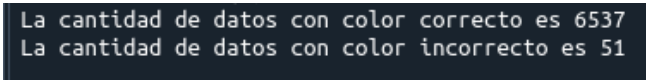
Los restantes de los registros inválidos representan una proporción significativamente menor. Algunos de estos valores son datos vacíos o null, lo cual puede interpretarse como falta de información que dificulta su clasificación. Estos datos vacíos también se consideran en el proceso de perfilado de datos, donde se identifican valores ausentes o anómalos que requieren análisis adicional. La falta de valores en estos registros reduce la cantidad de información útil disponible, lo que puede limitar la capacidad para extraer conclusiones significativas en el análisis, pero al ser un porcentaje tan bajo en este caso no parecería traernos problemas evidentes.

Además, en esta pequeña porción restante encontramos términos como Unverified y palabras derivadas, que también podrían asociarse con la categoría **Questionable**, pues indican una incertidumbre en la clasificación del incidente. Incluso encontramos algunos casos con espacios adicionales, por ejemplo el Provoked que observamos en el gráfico que es un " Provoked". Estas cosas afectan la calidad de los datos y reflejan una necesidad de estandarización y limpieza de los mismos.

Después de hacer algunos procesos de limpieza respecto a lo descrito anteriormente y explicados en la sección de decisiones tomadas, procedimos a verificar la coherencia entre los tipos y los colores asignados por la página.

Para lograrlo, agregamos una columna adicional al dataframe, en la cual registramos el color de fondo de cada celda en la columna "Type". Había algunas celdas con otros colores que no eran ninguno de los 5 definidos así que estas las dejamos vacías. Luego, guardamos la cantidad total de datos presentes, ya que este dato sería útil para los cálculos posteriores.

A continuación, eliminamos todas las filas con datos vacíos o incorrectos (aquellos que no coincidían con los tipos de color esperados). Con esta depuración, pudimos calcular cuántos datos coincidían correctamente con el color asignado y cuántos presentaban un color incorrecto. Esto lo obtuvimos restando el número de coincidencias correctas de la cantidad total de datos.



```
La cantidad de datos con color correcto es 6537
La cantidad de datos con color incorrecto es 51
```

Figura 3: Cantidad de datos que cumplen o no con la categorización

Como se puede observar en la imagen, la cantidad de datos con el tipo asignado incorrectamente o con un color que no se identifica con ninguno de los 5 descriptos (51) es muy baja en comparación con los datos correctamente asignados (6,537). Esto representa un porcentaje menor en el conjunto total de datos y sugiere que, en general, el sistema logra clasificar los tipos de manera bastante precisa en función del color de fondo.

Para entender mejor estos casos, realizamos una revisión manual de los datos que mostraban un color incorrecto en la columna "Type" (excluyendo aquellos vacíos). Al analizar las descripciones de cada caso, comprobamos que casi todos estos datos deberían haber tenido asignado el tipo correspondiente a su color de fondo, en lugar del tipo registrado inicialmente en la columna.

En la sección decisiones tomadas volvemos con estos temas y mostramos cual fue nuestra solución.

- **Country:** Este atributo contiene información geográfica sobre el lugar donde ocurrió cada ataque de tiburón. Al igual que en la columna "Type", encontramos tanto categorías válidas como no válidas. Sin embargo, en este caso, los problemas fueron más complejos debido a que había una mayor variedad de opciones válidas en comparación. Esta amplia gama de opciones válidas hace que la clasificación y el control de los datos incorrectos sea más desafiante.

Para obtener una visión preliminar de la información disponible antes de realizar la limpieza de esta columna, visualizamos la distribución de los ataques por país. Este análisis nos permitió observar las cantidades de incidentes registrados en cada ubicación.

Es evidente el contraste entre los cuatro países con el mayor número de accidentes registrados, ya que se observa una

	Países	Cantidad
0	USA	2552
1	AUSTRALIA	1481
2	SOUTH AFRICA	597
3	NEW ZEALAND	144
4	PAPUA NEW GUINEA	136
5	...	...
6	EGYPT / ISRAEL	1
7	MEXICO	1
8	Seychelles	1
9	GRAND CAYMAN	1
10	CEYLON (SRI LANKA)	1

Figura 4: Cantidad de ataques por país sin limpieza

gran diferencia en las cifras de cada uno. El país con el mayor valor es Estados Unidos, que tiene casi el doble de incidentes que Australia, el segundo país en la lista. Esta disparidad en los datos resalta la notable concentración de incidentes en ciertos países, lo que podría indicar factores geográficos, socioeconómicos o de población que influyen en la ocurrencia de estos accidentes.

Por otro lado, es interesante observar que todos los países que ocupan las últimas posiciones en la lista tienen solo un ataque registrado. Aunque este dato no es visible directamente en el gráfico, en realidad, son muchos los países que comparten esta misma cifra, además de los que ya se muestran en la tabla. Este patrón sugiere que, en muchos casos, los ataques de tiburones ocurren de forma aislada en diversas partes del mundo, lo que puede reflejar una menor incidencia o una falta de registros en ciertas áreas geográficas. Este tipo de información es relevante para comprender la distribución global de los incidentes y la posible variabilidad en los datos disponibles.

Para determinar cuáles eran o cómo clasificar los registros incorrectos, analizamos a mano y descubrimos que muchos de ellos contenían errores tipográficos que afectaban la consistencia de los datos. Entre los errores más comunes se encontraban el uso inapropiado de barras, puntos, comas y signos de pregunta, lo que generaba inconsistencias en la clasificación y dificultaba la correcta asignación de los registros a las categorías geográficas adecuadas. Además, observamos que algunos registros indicaban que los ataques ocurrieron en océanos o en ubicaciones ambiguas, como áreas situadas entre dos países, lo que complicaba aún más su categorización precisa.

Al igual que en la columna "Type" encontramos datos incompletos o con espacios adicionales, lo que generaba más dificultades para asegurar la exactitud de la información. Estos problemas reflejan la necesidad de un proceso de validación y limpieza de los datos más riguroso para mejorar la calidad y fiabilidad de la información geográfica contenida en el dataset.

Un ejemplo claro de esta problemática es la presencia de registros donde 10 personas escribieron "COLUMBIA" y 2 personas escribieron "COLOMBIA". Aunque ambos términos se refieren al mismo país, las diferencias en la escritura impiden una correcta agrupación de los datos. Este tipo de variabilidad, similar a los errores tipográficos mencionados previamente, refleja las inconsistencias que encontramos durante el análisis manual.

Country	count
COLUMBIA	10
COLOMBIA	2

Figura 5: Cantidad de accidentes registrados en Colombia

Otra observación importante fue la inconsistencia en el uso de mayúsculas y minúsculas. Al no contar con un criterio estandarizado para la entrada de datos, muchos países aparecían en diferentes formatos: algunos tenían el nombre completo en mayúsculas, otros en minúsculas, y algunos más con solo la primera letra en mayúscula. Esta falta de uniformidad generaba inconsistencias que dificultaban la correcta clasificación y análisis de los datos, ya que las variaciones en la capitalización no eran reconocidas como equivalentes, incluso si se referían al mismo país. La estandarización de este aspecto es crucial para asegurar una correcta agrupación y facilitar la comparación de los registros.

La imagen proporciona un ejemplo claro de las inconsistencias en la capitalización mencionadas anteriormente. Podemos

Country	count
MEXICO	103
Mexico	1
MeXICO	1

Figura 6: Cantidad de accidentes registrados en Mexico

observar que el nombre de México aparece registrado de varias maneras: todo en mayúsculas ("MEXICO"), con solo la primera letra en mayúscula ("Mexico") e incluso con una combinación inusual de mayúsculas y minúsculas ("MeXICO"). Estas variaciones en el uso de mayúsculas y minúsculas no solo dificultan la consistencia visual, sino que también complica la correcta agrupación de registros, ya que el sistema no las reconoce como equivalentes a pesar de que se refieren al mismo país.

Este ejemplo ilustra la importancia de la estandarización de capitalización en los datos, como se describió en el texto anterior. La estandarización es, por lo tanto, un paso esencial para mejorar la calidad y precisión del dataset, garantizando que los datos de un mismo país se agrupen y analicen correctamente.

En la sección de decisiones tomadas volvemos con estos temas y mostramos cuál fue nuestra solución.

### 3. Decisiones tomadas

Una vez terminado el análisis de ambas columnas de interés tuvimos que decidir cómo resolver las cuestiones de limpieza de datos innecesarios, para eso tomamos las siguientes decisiones.

- **Eliminación de filas con datos vacíos en la columna "Type":** Se descartaron todas las filas en las que el dato de la columna "Type" estaba vacío. Dado que este atributo es fundamental para nuestro análisis y que los datos nulos no aportan información relevante, optamos por su eliminación para evitar distorsiones en los resultados.
- **Reclasificación de filas con el valor "Invalid" en la columna "Type":** En los casos donde el valor en la columna "Type" era Invalid, pero la descripción proporcionada indicaba un contexto coherente con la categoría **Questionable**, reemplazamos el valor Invalid por **Questionable**. Esto permitió una clasificación más precisa de los datos y mejoró la representatividad de los registros dentro de las categorías principales.
- **Descartar filas con valor Invalid en "Type":** Las filas con el valor Invalid en la columna "Type" que dejaban en duda la presencia de un tiburón antes de la muerte de la víctima fueron eliminadas. Consideramos que estos casos no calificaban como ataques de tiburones confirmados y, por lo tanto, no cumplían con los criterios necesarios para pertenecer al dataset.
- **Eliminación de filas con valores no reconocidos:** Se eliminaron las filas que no contenían ni el valor Invalid ni alguna de las cinco categorías principales en la columna "Type". Dado que estas filas representaban un porcentaje extremadamente bajo y carecían de relevancia para el análisis global, su eliminación permitió concentrar el análisis en los datos clave y evitar posibles sesgos debidos a información marginal.
- **Clasificación de filas con valores vacíos en la columna "Color":** Con el fin de mantener la prolijidad y evitar valores nulos en el dataframe, se decidió asignar el color correspondiente al tipo indicado en la columna "Type" para cada fila en la que la columna "Color" había quedado vacía. Esta situación se produjo debido a una falla en el sistema de asignación de colores, que aplicó un color incorrecto en ciertos casos. Como resultado, al ejecutar el código para agregar la columna con el color de cada celda, estos datos quedaron sin valor. Esta solución permitió completar los valores vacíos de manera consistente con los tipos indicados.
- **Reclasificación de filas con valor incorrecto en la columna "Type":** Durante nuestro análisis, concluimos que en las filas donde el valor de la columna "Type" no coincidía con el color de la celda, la información correcta debía derivarse del color asignado. Por este motivo, decidimos ajustar estos valores y reemplazamos el contenido de la columna "Type" con el tipo indicado por el color de fondo de la celda, asegurando así una clasificación más precisa y consistente.
- **Eliminación de filas con datos vacíos en la columna "Country":** Al igual que en la columna "Type", se descartaron todas las filas en las que el dato de la columna "Country" estaba vacío. Nuevamente consideramos que esta era la opción más conveniente para facilitar el manejo y análisis efectivo de los datos.
- **Estandarización de nombres de países en mayúsculas:** Se convirtieron todos los nombres de los países a mayúsculas para unificar y asegurar que los registros referidos al mismo país fueran identificados correctamente. Este ajuste facilita el análisis, evitando que variaciones en la escritura generen duplicaciones innecesarias en los datos.

- **Eliminación de espacios adicionales:** Se eliminaron todos los espacios innecesarios ubicados al inicio o al final de cada nombre de país, ya que estos generaban inconsistencias en los datos. Sin embargo, se conservaron los espacios internos en los nombres de países que los incluyen, como por ejemplo South Africa, para respetar la correcta denominación geográfica.
- **Eliminación de registros con símbolos ambiguos:** Se eliminaron los datos de la columna "Country" que contenían símbolos como signos de interrogación o barras, ya que estos representaban ubicaciones ambiguas o indeterminadas, lo cual dificultaba una categorización confiable.
- **Eliminación de datos con las palabras Ocean y Sea:** Los registros que solo especificaban ubicaciones generales como "Ocean" o "Sea" fueron eliminados. Al no indicar un país específico decidimos excluir estos datos para evitar que obstruyeran el análisis de ubicaciones más precisas y relevantes.

Una vez completado este proceso de limpieza y tomando las decisiones de depuración necesarias, logramos obtener un dataset mucho más estructurado y confiable para abordar nuestras preguntas de investigación. A lo largo de esta etapa de limpieza, enfrentamos múltiples problemas que suelen aparecer en datasets complejos: valores nulos, errores tipográficos, variaciones en la escritura, datos ambiguos y categorías inconsistentes. Cada uno de estos problemas fue cuidadosamente evaluado y solucionado para asegurar la calidad y homogeneidad de los datos.

La limpieza de datos es una tarea fundamental cuando se trabaja con información que proviene de fuentes heterogéneas y que, con frecuencia, contiene errores o inconsistencias que pueden sesgar los resultados. La estandarización de categorías y la eliminación de registros irrelevantes son pasos clave que permiten que el análisis sea más robusto y preciso.

Este enfoque no solo mejora la calidad de los datos, sino que también facilita su interpretación, permitiéndonos obtener conclusiones más fundamentadas y reducir la posibilidad de resultados erróneos debidos a información defectuosa. Con este dataset limpio, podemos ahora proceder a responder nuestras preguntas de análisis con la certeza de que los resultados serán lo más precisos posible, habiendo minimizado los efectos de las problemáticas iniciales del dataset.

## 4. Respuestas

### 4.1.

En este trabajo, se nos pide analizar los datos relacionados con los ataques de tiburones, enfocandonos específicamente en las categorías registradas en la columna "Type" la cual nos indica el tipo de ataque. Nuestro objetivo es obtener la cantidad de incidentes correspondientes a cada categoría, para así comprender la distribución de los distintos tipos de ataques.

Una vez ya hecha la limpieza de los datos obtuvimos la siguiente información:

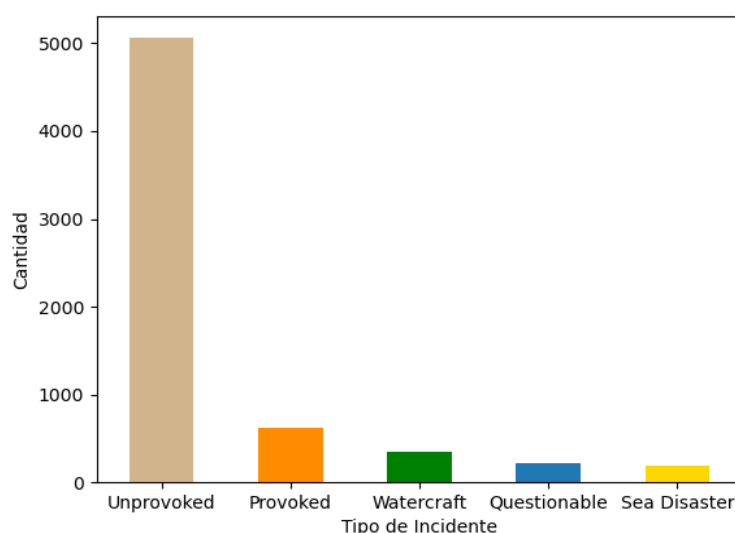


Figura 7: Distribucion de datos por tipo de accidente

Como se puede observar, la gran mayoría de los casos registrados corresponden a ataques **Unprovoked**, con un total de 5,061 incidentes, mientras que las demás categorías se encuentran considerablemente por debajo: 622 ataques **Provoked**, 348 **Watercraft**, 213 **Questionable** y 194 **Sea Disaster**.

Se cree que la gran diferencia en la cantidad de ataques **Unprovoked** se debe a que muchos de estos incidentes ocurren en las zonas de mayor riesgo, que suelen ser las más cercanas a la costa, donde las actividades recreativas (como surf, natación y buceo) son frecuentes. Esta genera una mayor exposición al riesgo, lo que a su vez incrementa la cantidad de ataques registrados. A partir de la información proporcionada en la página, podemos concluir que la mayoría de los ataques de tiburones no ocurren por una intención depredadora como se podría imaginar. En realidad, al compartir el espacio marítimo con los seres humanos en hábitats tan frecuentados y al ser una especie extremadamente sensible a ciertos estímulos, el tiburón puede reaccionar de manera violenta sin que necesariamente haya una provocación directa. Su comportamiento se activa por reflejos instintivos, situaciones de estrés o hasta la confusión con sus presas naturales.

Además, desde la perspectiva de los tiburones, muchos de estos incidentes podrían clasificarse como **Provoked**, ya que un tiburón, enfrentado a lo que considera amenazas o a movimientos inusuales en su territorio, responde con el instinto de protegerse. Visto así, estos accidentes surgen más de la interacción natural y la vulnerabilidad compartida en su entorno que de un propósito agresivo de los tiburones hacia los humanos.

Por otro lado, los ataques categorizados como **Provoked** suelen ocurrir cuando los humanos interactúan de manera directa o agresiva con los tiburones. Esto puede suceder en situaciones como la pesca o captura de estos animales, o cuando una persona intenta defenderse ante una posible amenaza. Sin embargo, este tipo de ataques son considerablemente menos comunes, ya que la mayoría de las personas evita el contacto directo con los tiburones y, en general, procura no provocarlos.

No obstante, algunos tiburones pueden interpretar la presencia humana como un riesgo o una competencia por recursos. En estos casos, el tiburón podría atacar, no como una acción depredadora, sino en un intento de proteger su territorio o sus fuentes de alimento. Es decir, aunque los humanos no estén buscando una confrontación, la percepción del tiburón y sus instintos defensivos pueden llevarlo a responder de manera agresiva. Esta respuesta, basada en la defensa de su espacio y recursos, subraya que los ataques no siempre son resultado de un comportamiento hostil, sino de un instinto de autopreservación y territorialidad en un ambiente que ambos, humanos y tiburones, deben compartir.

Los ataques a embarcaciones, clasificados en la categoría **Watercraft**, se producen generalmente debido al movimiento y las vibraciones generadas en el agua, que los tiburones pueden interpretar como señales de posibles presas o intrusos en su territorio. Estos ataques suelen ser una respuesta instintiva, donde el tiburón reacciona a lo que percibe como actividad anómala en su entorno. En algunos casos, los tiburones confunden los motores, remos o partes de las embarcaciones con animales heridos u otros objetos de interés, lo que los lleva a aproximarse e incluso morder.

Aunque estos incidentes pueden parecer alarmantes, son menos frecuentes que los ataques directos a personas. Esto se debe a que los tiburones no asocian las embarcaciones con presas de la misma forma que pueden hacerlo con seres humanos en el agua. Para ellos, los objetos flotantes carecen de las características que generalmente despiertan sus instintos de caza, como los movimientos erráticos y las señales bioeléctricas que emiten otros seres vivos. Además, los tiburones tienden a alejarse de elementos que no les son familiares o que no se comportan como presas naturales.

Sin embargo, en aquellas ocasiones en las que un tiburón se siente atraído por el movimiento o las vibraciones de una embarcación, su curiosidad o instinto territorial puede llevarlo a aproximarse y a realizar lo que se conoce como "mordiscos exploratorios". Este tipo de interacción ocurre más frecuentemente en áreas donde los tiburones están acostumbrados a la actividad humana o donde pueden confundir embarcaciones pequeñas con presas en movimiento, especialmente si las condiciones del agua o la visibilidad son limitadas.

La categoría **Questionable** comprende aquellos casos en los que existen dudas significativas sobre las circunstancias del ataque o sobre si realmente un tiburón fue el causante de la lesión. Estos incidentes presentan un nivel de ambigüedad considerable, ya que la falta de testigos presenciales o de pruebas concluyentes complica la verificación de los hechos. En algunos de estos casos, no queda claro si la persona ya había fallecido antes del supuesto ataque, lo que abre la posibilidad de que el tiburón interactuara con el cuerpo sin intención depredadora, sino como parte de su comportamiento natural de exploración.

Además, las condiciones del entorno y el estado de los restos suelen dificultar el análisis, ya que factores como el tiempo transcurrido desde el incidente, la exposición al agua y el posible deterioro de la evidencia limitan la capacidad de los investigadores para establecer una secuencia clara de los eventos. Dado que estos casos no cuentan con pruebas concluyentes y a menudo se basan en conjeturas o hallazgos incompletos, resulta complicado clasificarlos con certeza en otras categorías de ataques. Por esta razón hay pocos accidentes catalogados en esta categoría al ser la más ambigua.

Finalmente, los incidentes clasificados como **Sea Disaster** son relativamente raros y difíciles de documentar con precisión debido a la naturaleza compleja y caótica de estos eventos. A diferencia de los ataques **Unprovoked**, que ocurren en circunstancias donde las interacciones con los tiburones pueden observarse o estudiarse, los casos de **Sea Disaster** suelen desarrollarse en situaciones extremas, como naufragios, accidentes marítimos o en aguas abiertas donde las personas quedan vulnerables y expuestas en el hábitat natural de los tiburones durante un periodo prolongado.

Este tipo de incidentes es complicado de registrar y clasificar, ya que en muchos casos no es posible determinar la cantidad exacta de ataques o el comportamiento del tiburón involucrado. Muchas veces estos accidentes son clasificados como **Questionable** o **Unprovoked** dado que es difícil determinar la verdadera naturaleza del incidente y como realmente actuó o influyó en él la presencia del tiburón. Las condiciones suelen ser incontrolables y caóticas, con poca visibilidad y, a menudo, en áreas remotas o en alta mar. Estas dificultades hacen que los datos disponibles sobre los **Sea Disasters** sean limitados y, en ocasiones, se basen en testimonios o en reconstrucciones posteriores de los hechos.



En muchos **Sea Disasters**, los tiburones pueden sentirse atraídos por las señales emitidas por la situación misma, como los sonidos de una embarcación dañada, el olor o la presencia de restos en el agua, o las vibraciones causadas por el pánico y el movimiento de las personas afectadas. Sin embargo, no siempre está claro si los ataques ocurren por un comportamiento depredador, un instinto de exploración, o como resultado de un fenómeno oportunista.

## 4.2.

En esta sección, se nos pide analizar la cantidad de incidentes de cada tipo por país. El objetivo principal es identificar como varían los tipos de ataques según la ubicación geográfica, lo que podría revelar diferencias en el comportamiento de los tiburones y en las interacciones humanas en distintas regiones del mundo.

Más allá de simplemente contar los incidentes por país y tipo, este análisis permitirá entender patrones geográficos que podrían estar influenciados por factores como la actividad humana, las condiciones climáticas, o la biodiversidad marina de cada región. Esto nos ayudará a descubrir si ciertos países son más propensos a un tipo específico de ataque y qué factores podrían contribuir a esas tendencias.

Una vez ya hecha la limpieza de los datos obtuvimos lo siguiente:

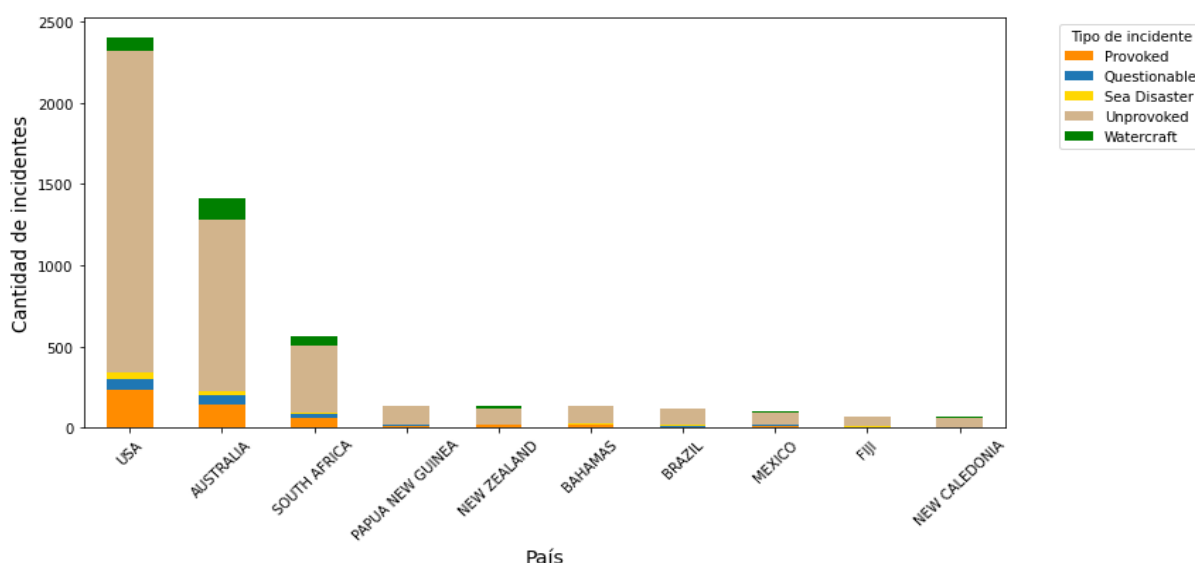


Figura 8: Cantidad de incidentes de cada tipo por país

Vamos a enfocarnos y analizar los resultados de los 10 países con mayor número de incidentes. Podemos notar que Estados Unidos destaca con la mayor cantidad de incidentes, seguido por Australia y Sudáfrica. Otros países, como Papúa Nueva Guinea, Nueva Zelanda, Bahamas, Brasil, México, Fiji y Nueva Caledonia, tienen significativamente menos incidentes en comparación con los tres primeros y una cantidad muy parecida entre sí, indicando que, excepto por los tres países con gran cantidad, la mayoría rondan el mismo número de accidentes, con la cantidad bajando de una manera muy lenta y siendo cada vez más aproximado entre sí.

Los incidentes **Unprovoked** representan la mayor parte de los incidentes en la mayoría de los países, siguiéndose con la categoría **Provoked**, los relacionados con embarcaciones **Watercraft**, y por último las categorías **Sea Disasters** y **Questionable** en menor proporción, coincidiéndose con lo analizado en el inciso anterior.

Estados Unidos cuenta con una de las líneas costeras más largas del mundo, que abarca desde las costas del Atlántico, el Golfo de México y el Pacífico hasta las aguas tropicales de Hawái. Esta diversidad de ambientes marinos crea un hábitat adecuado para varias especies de tiburones y aumenta las probabilidades de encuentros con humanos. Las costas estadounidenses, especialmente en estados como Florida, California y Hawái, son destinos populares tanto para turistas como para residentes. Esto significa que, en cualquier época del año, hay un gran número de personas en el agua, lo que incrementa las posibilidades de incidentes debido a la alta interacción entre humanos y tiburones.

Además, Estados Unidos cuenta con avanzados sistemas de registro y documentación de ataques de tiburones, liderados por organizaciones como el Archivo Internacional de Ataques de Tiburón (ISAF). Esto significa que los incidentes tienden a ser registrados de manera precisa, lo cual podría contribuir a que Estados Unidos encabece la lista en términos de cantidad total de incidentes. En áreas como Florida, la corriente del Golfo proporciona aguas cálidas y ricas en nutrientes que atraen a muchas especies marinas, incluidos los tiburones. Estas condiciones favorables para la vida marina crean un ambiente donde los tiburones y las personas coinciden frecuentemente.

Igualmente, podemos destacar que, si bien Estados Unidos posee la mayor cantidad de incidentes en general, Australia cuenta con una mayor cantidad de ataques de tipo **Watercraft**. Esto puede ser consecuencia de que Australia alberga algunas de las especies de tiburones más grandes y potencialmente peligrosas, como el tiburón blanco, el tiburón tigre y el tiburón

toro, que suelen estar activos en áreas cercanas a la costa y en zonas de navegación. Estos tiburones son particularmente curiosos y a veces tienden a investigar objetos en movimiento, como embarcaciones pequeñas y tablas de surf, lo que puede llevar a más incidentes de tipo **Watercraft**.

En Australia, las actividades acuáticas como el surf, el remo y el kayak son extremadamente populares y forman parte del estilo de vida en muchas ciudades costeras, esto se debe a que tiene una costa que es particularmente adecuada para deportes acuáticos en aguas abiertas, con temperaturas favorables durante gran parte del año. Esta alta frecuencia de personas en embarcaciones pequeñas aumenta las probabilidades de interacciones con tiburones, especialmente en áreas conocidas por la presencia de estas especies. En cambio, en los Estados Unidos, aunque los deportes acuáticos son también populares, no tienen la misma concentración en zonas donde los tiburones grandes se encuentren con tanta frecuencia.

En los Estados Unidos, por otro lado, los incidentes **Unprovoked** suelen ser más comunes y están asociados con la gran cantidad de bañistas en playas de estados como Florida y California, donde también hay presencia de tiburones. Estos incidentes ocurren cuando los tiburones confunden a los humanos con sus presas.

Además, puede haber diferencias en la manera en que los incidentes son documentados y clasificados en cada país. Australia tiene una infraestructura de monitoreo y registro altamente desarrollada en sus costas debido a la frecuencia de estos encuentros, lo que podría resultar en una mayor precisión al identificar incidentes específicos de tipo **Watercraft**. En cambio, en los Estados Unidos, algunos incidentes pueden quedar registrados como **Unprovoked** cuando el contexto del ataque no es claro o cuando la interacción ocurre en situaciones más ambiguas.

Por otro lado, Sudafrica ocupa el tercer lugar en la cantidad total de incidentes de ataques de tiburón por varias razones clave de su geografía y biodiversidad, entre otras cosas. Las aguas sudafricanas albergan algunas de las especies de tiburones más grandes y potencialmente peligrosas, estos tiburones son altamente activos en las costas sudafricanas, especialmente en las zonas de confluencia entre el Océano Atlántico y el Océano Índico, lo cual crea un hábitat ideal para especies marinas grandes, incluida una gran variedad de tiburones.

Sudáfrica cuenta con una numerosa población de focas, particularmente en lugares como Seal Island en Ciudad del Cabo, que es un atractivo punto de caza para los tiburones blancos. Esto genera una alta actividad de estos depredadores en áreas relativamente cercanas a la costa. También los deportes acuáticos como el surf, el buceo en jaula para observar tiburones y el remo son populares, especialmente en zonas como la costa de Ciudad del Cabo, Durban y otros destinos costeros. Estas actividades llevan a una alta exposición humana en aguas donde hay tiburones grandes, aumentando el riesgo de interacciones.

La Corriente de Agulhas, que transporta aguas cálidas desde el Océano Índico, y la Corriente de Benguela, que trae aguas frías y ricas en nutrientes desde el Atlántico Sur, se encuentran frente a la costa sudafricana. Esto crea un entorno marino rico en biodiversidad, lo cual atrae tanto a tiburones como a otros animales marinos en busca de alimento. Últimamente, Sudáfrica ha implementado importantes programas de observación de tiburones y de investigación sobre interacciones entre tiburones y humanos, como el Shark Spotters en Ciudad del Cabo. Esto significa que se llevan registros más detallados de los incidentes, incluyendo reportes de ataques y encuentros con tiburones. Al igual que en Australia, en Sudáfrica el océano y las actividades costeras son una parte importante de la vida cotidiana para muchas personas. Este alto nivel de interacción aumenta el riesgo de incidentes debido a la frecuente presencia de personas en las aguas donde también habitan tiburones.

Además de Estados Unidos, Australia y Sudáfrica, otros países con incidencias destacadas de ataques de tiburones en el gráfico incluyen a Nueva Zelanda, Brasil, México, Papúa Nueva Guinea, Fiji, y las Bahamas. Aunque estos países muestran un menor número de incidentes en comparación con los tres primeros, su presencia en el análisis es coherente con factores ambientales y de actividad humana en sus costas.

Por ejemplo, Nueva Zelanda y Fiji, ubicados en el Océano Pacífico, son destinos populares para el surf y el buceo, actividades que aumentan la posibilidad de interacciones con tiburones debido al movimiento y las vibraciones en el agua. Papúa Nueva Guinea, con una gran diversidad marina y abundantes recursos, presenta zonas donde tiburones y humanos se encuentran en áreas de pesca, lo que eleva el riesgo de incidentes. En Brasil y México, el clima tropical y las playas concurridas atraen tanto a turistas como a tiburones, especialmente en zonas como el noreste de Brasil y la península de Yucatán en México. En las Bahamas, el ecoturismo que involucra la observación de tiburones es una actividad frecuente, lo cual puede aumentar los encuentros cercanos con estas especies.

Este contexto ayuda a entender la variabilidad de incidentes en diferentes regiones y resalta cómo las condiciones locales —como la actividad turística, el clima, y la biodiversidad marina— pueden influir en el tipo y la frecuencia de ataques de tiburones en cada país.

## 5. Conclusion

En conclusión, el proceso de limpieza y análisis realizado sobre el dataset de ataques de tiburones ha sido fundamental para mejorar la calidad y precisión de la información disponible. A través de la revisión y estandarización de datos, eliminamos registros ambiguos, valores vacíos y errores tipográficos, lo cual permitió obtener una representación más fiel de los tipos de ataques y su distribución geográfica. Este trabajo de depuración fue esencial para identificar patrones claros en la ocurrencia de incidentes y comprender mejor los factores que influyen en las interacciones entre tiburones y humanos. Igualmente el dataset contenía un gran porcentaje de datos correctos, así que ya desde un principio se puede decir que se contaba con una calidad muy buena, aunque siempre hay cosas para perfeccionar y acomodar.

El análisis también reveló que la gran mayoría de los ataques corresponden a incidentes **Unprovoked** o no provocados,

los cuales suelen ocurrir en áreas cercanas a la costa, donde la actividad humana es alta. Esto sugiere que los tiburones no suelen ver a los humanos como presas, sino que los ataques se producen debido a la curiosidad o a la confusión. Los ataques **Provoked**, en cambio, reflejan situaciones en las que los humanos interactúan de manera directa con los tiburones, como durante la pesca o la captura, lo que pone de relieve la importancia de respetar su espacio natural para reducir incidentes.

Asimismo, el estudio mostró que la distribución de los ataques varía por país, siendo Estados Unidos, Australia y Sudáfrica las regiones con mayor número de incidentes, lo cual podría estar relacionado con sus extensas costas y alta concentración de actividades acuáticas. Cada país presenta particularidades en los tipos de incidentes: por ejemplo, los ataques a embarcaciones son más comunes en Australia, donde tiburones grandes y potencialmente peligrosos suelen investigar objetos en movimiento, como tablas de surf y embarcaciones pequeñas.

En definitiva, el análisis de datos de ataques de tiburones no solo contribuye a mejorar nuestra comprensión sobre el comportamiento de estos animales, sino que también proporciona información valiosa para diseñar estrategias de prevención y minimizar el riesgo de encuentros peligrosos. Con un dataset más limpio y estandarizado, los resultados obtenidos no solo permiten extraer conclusiones más precisas, sino que también abren la puerta a futuras investigaciones sobre la relación entre tiburones y humanos en distintas regiones del mundo.