Ocean Aware

Reducing Shark Incidents Among
Teen Surfers in the U.S.A.

Patricia Viladomiu
Veerpal Chattha
Marta García
Carlos Vera

Introduction

The project uses the **Global Shark Attack File (GSAF)**, a large dataset that documents **reported shark attacks worldwide**.

Problem: Many young surfers—especially beginners and teenagers—enter the ocean without sufficient knowledge about marine ecosystems, shark behavior, or safe surfing practices.

Objective: To evaluate the real level of shark-incident risk among teenagers (ages 12-20) in the USA and determine whether recreational activities such as surfing pose a significant danger. Additionally, to use the data to support the development of safe, educational, and well-informed surf programs of youth.

Hypothesis: “Among ocean activities practiced by teenagers, surfing presents the highest relative risk of shark incidents.”

🚧 Major Obstacles & Key Learnings

1. Inconsistent Text Formats

- **The Challenge:** Categorical columns (Country, Activity, Sex) contained hundreds of variations and typos (M, male, m). This **fragmented the data** and prevented accurate counting.
- **The Solution:** Implemented a robust **standardization pipeline** using Pandas methods (.strip(), .lower()) and **dictionary mapping** (.replace()) to unify all variants into single categories (e.g., 'Male', 'Surfing').
- **Key Learning: Aggressive text normalization** is the foundational step for reliable analysis.

2. Temporal Data Chaos

- **The Challenge:** The raw Date and Year data was messy and unreliable, making it impossible to perform essential **seasonal or trend analysis**.
- **The Solution:** Used **Date Coercion** (pd.to_datetime) to clean the time values and **Feature Engineering** to create a new, essential feature: **Season**.
- **Key Learning:** We demonstrated the ability to **engineer new features** from chaotic data, which was crucial for identifying high-risk seasonal periods (like Summer).

🔍 Exploratory Data Analysis

The techniques used in EDA are the following:

Aggregation / Grouping :

This technique is used to identify which states are more dangerous, which activities are more risky and which age groups are most affected. This technique identifies patterns, trends, and distributions across different dimensions.

The outcome was filtered for key outcomes fatal/ Non fatal. The conclusions are drawn from the aggregated tables.

We used following grouping to show distribution of categorical variables (number of attacks, activity).

Attacks grouped by **Activity Category**

Attacks grouped by **Season**

Attacks grouped by **State**

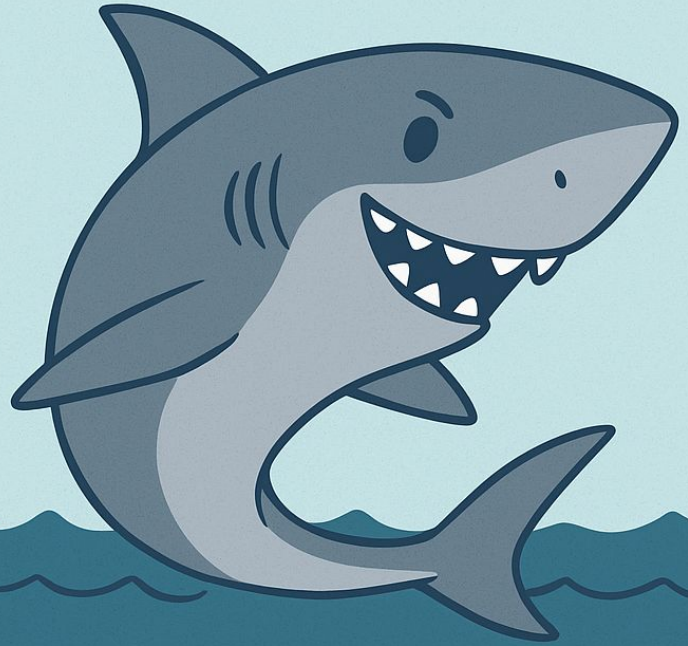
TOP 5 MOST DANGEROUS WATER ACTIVITIES

TEENS 12–20, USA

1	SURFING	253
2	SWIMMING	81
3	FISHING	26
4	SNORKELING	5
5	WADING	2

TOP 5 MOST DANGEROUS STATES

TEENS 12–20, USA



1	FLORIDA	349
2	SOUTH CAROLINA	35
3	CALIFORNIA	33
4	NORTH CAROLINA	33
5	HAWAII	30

🎯 Conclusions

Our analysis confirms that the modern risk of shark incidents is **specific, predictable, and targetable**. The findings compel a direct intervention to protect the most vulnerable demographic.

The Proposal: Florida Summer Safety Camp ⛺

Based on the evidence—concentrated risk in **Florida**, peaked seasonality in **Summer**, and high incidence during **Surfing** among the **12-20 age group**—we propose a targeted educational program.

- **Intervention:** Launch a **Shark Safety & Behavioral Education Summer Camp** in high-risk zones of Florida.
- **Goal:** To proactively **reduce incident numbers** and **change the course of the exponential risk trend** by teaching responsible coastal practices.

Final Call to Action 🧠

We must convert data into direct action: Education is the most powerful tool to reduce behavioral risk and improve the safety of our coastlines.

Thank you

Patricia Viladomiu
Veerpal CHATTHA
Marta García
Carlos Vera

Data Cleaning

Challenges faced :

There were many critical data missing. There were datasets with inconsistent date formats.

For cleaning the data frames we used removing useless or empty columns as there were many useless columns. There were also a lot of Inconsistent categorical values (e.g., "M", "male", "F", "Female") in the sex column. Many columns contained no meaningful data like age column had non-numeric characters (like “?”, “+”, “-”, ranges, or text), which prevents numerical analysis. The raw date data was messy and inconsistent.

Cleaning steps that were required:

Converted names to lowercase. Removed spaces and punctuation.

For the sex column the values were converted to lowercase, missing values are filled with unknown.

The age column was converted to numeric, regex is used to remove letters, special characters, and hyphens. A new age group column was created to categorize the age.

We Looped through each of the three columns: Country, State, Location ,converted all values to strings and removed leading/trailing spaces