# **Shark\_Quest: A roadmap for shark ecotourism**

## **Introduction**

## This dataset from<https://www.sharkattackfile.net/> was used to test data cleaning, analysis, problem-solving, and teamwork skills. Our team of four rotated research and coding tasks daily to develop and implement the following business model.

## **Objective of the Analysis**

## Our objective was to creatively and strategically approach data towards data cleaning of the unwieldy and disorganised data set. The underlying yet central theme of shark-human interactions, though hostile in nature, provided a solid foundation to develop a hypothesis around shark conservation, public awareness, and education. Our data wrangling was shaped by this hypothesis.

## **Working Hypotheses**

## **Eco-tourism** is a feasible business model that can significantly contribute to shark conservation and awareness. By generating revenue, eco-tourism helps fund marine protected areas, educate tourists and reduce misconceptions about sharks as "man-eaters" and minimise actions that provoke attacks.

* **The dataset could help us** understand the nature of shark attacks, both provoked and unprovoked, identify hostile and non-hostile species, and assess endangered shark populations based on secondary research. This, in turn, could help determine significant issues and suitable entry points towards developing an eco-tourism business model

## **Data Analysis**

### Data Cleaning and Preparation

Given the project's time constraints, our goal was to make the expansive and incomplete dataset manageable while aligning with our hypotheses. The following adjustments were made to streamline the data:

1. **New Dataframe**: We analysed the columns of the original dataset and selected those most relevant to our task. Rather than cleaning and dropping columns, we created a new dataframe excluding ‘date’ ‘location’, ‘activity’, ‘age’, ‘sex’, ‘location’, ‘Unnamed:\_11’, ‘Unnamed \_21’, ‘Unnamed \_22’,"Case number""Href\_formula", "Pdf href", "Origi’al\_order", "Source", "Case number”
2. **Standardization of Columns**: All columns were standardised using functions like lower(), strip(), and replace(), while duplicates and inconsistencies were addressed.
3. **Data Cleaning**:
   1. **Year**: The column was standardised by converting it to integers, filling missing values (NaNs) with zeros, and removing gaps and float values. The data was sorted, and we chose to focus on the past 24 years, considering climate change and contemporary tourism trends.
   2. **Species**: This field was categorised, standardised, and unified based on species type.
   3. **Country and State**: We standardised different spellings and dropped rows with missing data, as precise region information was necessary.
   4. **Type**: We filtered out irrelevant categories such as "Invalid," "Watercraft," "Sea Disaster," "Under Investigation," "Unconfirmed," and retained only "Provoked" and "Unprovoked," as these offer insights into shark hostility. We also removed NaNs from this column.
   5. **Injury**: Further categorization was performed by identifying cases with lacerations, minor injuries, major injuries, fat, probably scavenging, and material damage

### Data Correlation

### After standardising the data, we examined the relationship between state, country, and injury type with the nature of shark attacks (provoked vs. unprovoked). This analysis offered key insights into both shark behaviour (unprovoked attacks) and human behaviour (provoked attacks), as well as the geographic distribution and severity of these interactions.

**Third Hypothesis**: We needed to choose one type of attack as a benchmark for measuring variables such as activity area, injury type, and shark species. Although using provoked attacks could identify less aggressive species and areas of interest, it yielded sparse data on location and species, indicating low shark density or unreliable data. In contrast, unprovoked attacks provided better insights into high-density shark areas, diverse species, and allowed us to focus on vulnerable species for eco-tourism opportunities.

### Data-Driven Insights

* **Shark ecotourism**: Feasible in areas with frequent human-shark encounters, as injuries are generally non-threatening, as supported by secondary research. The data revealed that most shark species in these locations require conservation, aligning ecotourism with preservation efforts.
* **Business proposal:** North Carolina emerged as the ideal location for shark eco-tourism due to its scenic beaches, rich marine life, and [established shark research, including the longest-running shark survey in the U.S](https://www.unc.edu/discover/the-past-present-and-future-of-shark-research/). Species in need of conservation in this area include bull, tiger, white, and blacktip sharks.

## **Conclusion**

Extracting meaningful data relies on identifying trends and testing hypotheses against relevant categories. The research question helps refine the dataset, and raw findings often require supplementary research. (Further research: Additional investigation is needed into specific locations within North Carolina, safety protocols, research programs, and the financial feasibility of the business proposal and possible sponsorships and investments.)