SharkTrack

The Shark Tale Project

Sheryll Dumapal

Business Case & Opportunity for SharkTrack

Shark attacks pose a risk to public safety, tourism, and local economies, with no real-time detection solutions currently in place. SharkTrack addresses this gap by leveraging AI and data analytics to predict and prevent shark encounters.

Opportunity:

- Use real-time tracking and historical data to identify high-risk areas.
- Provide governments, surfers, and businesses with actionable safety insights and products.
- Enhance ocean safety while promoting coexistence with marine life through non-invasive monitoring.

Overview

Original Dataset & Hypothesis

The dataset contains global shark attack records from the 19th to 21st centuries, including details such as date, location, species, activity, and outcomes. I formulated three hypotheses:

1. Surfing increases the likelihood of shark encounters.
2. Attacks are more frequent in the afternoon due to increased human activity.
3. Certain regions experience more attacks due to environmental conditions.

Data Cleaning & Analysis Structure

- Preprocessing: Removed duplicates, handled missing values, and standardized species names.
- Filtering: Focused on relevant time periods and attack types.
- Exploratory Data Analysis (EDA): Grouped data by activity, time, and location to validate hypotheses.
- Visualization: Used bar charts, time-series analysis, and heatmaps to detect patterns.

Unique Data Cleaning Techniques

- String Normalization: Removed extra spaces and standardized species names (`df["Species_Types"].str.strip()`).
- Custom Shark Classification: Used `str.contains()` to categorize attacks by species.
- Decade-Based Analysis: Aggregated data to eliminate yearly noise and reveal long-term trends.
- Handling Missing Data: Used logical imputation for unknown values and excluded unreliable records.

Data Wrangling and Cleaning

Missing Data: Many records lacked species identification, time, or activity details.

Solution: Imputed logical values where possible; excluded unreliable records.

Duplicates & Inconsistencies: Inconsistent species names and formatting issues.
Solution: Standardized names using `.str.strip()` and grouped similar entries.

Time Formatting Issues: Attack timestamps were incomplete or inconsistent.
Solution: Categorized attacks into time-of-day bins (Morning, Midday, Afternoon, etc.).

Historical Data Variability: Older records had inconsistent reporting.
Solution: Aggregated data by decade to ensure reliability and remove yearly noise.

Exploratory Data Analysis (EDA) & Insights

Methods Used:

- Grouping & Aggregation: Analyzed attack trends by activity, time, and location.
- Visualization: Used bar charts, time-series plots, and heatmaps for pattern detection.
- Statistical Analysis: Calculated attack frequencies and distributions.

Key Insights:

- Surfing had the highest attacks (1,150 cases), confirming it’s high-risk.
- Afternoon saw the most attacks (659 cases), aligning with peak human activity.

Obstacles & Key Learnings

- **Obstacle:** Missing and inconsistent data, especially for species identification and attack times.
- **Mistake:** Not building a structure for cleaning data, which has led me to extra working hours on the project.
- **Solution:** Following standardized process steps, creating a template that I can follow use it for several projects in the future.
- **Lesson Learned:** Data quality and following steps is critical. Proper cleaning and preprocessing are essential for accurate, reliable insights, shaping a more structured approach for future analyses.

Hypothesis Evaluation & Key Findings

Supported Hypotheses:

- Surfing increases shark encounters – Confirmed (1,150 cases).
- Attacks peak in the afternoon – Confirmed (659 cases).
- Certain regions have more attacks – Confirmed (Florida, Australia, South Africa).

Surprising Insights:

- Most shark attacks are non-fatal, contrary to common fear.
- Many cases lack species identification (2,580 unknown species), highlighting data gaps.
- New Smyrna Beach, Florida has the highest attacks globally (182 cases).

Implications:

- Better tracking technology is needed for real-time risk assessment.
- Governments and tourism industries can use data-driven insights to improve safety measures.
- Education and awareness campaigns can reduce panic and promote coexistence with sharks.

Github Overview

Explanation:

1. Data cleaning was performed in the cleaning file.
2. The wrangling file contains visualizations, findings, and conclusions.
3. The images folder contains all the charts and plots
4. The Shark Attack folder contains the csv file, the species_counts.csv and the cleaned_data.pkl

Questions?

The End

Thank you for your time and listening!