# Quest: Shark Attack

Toan, Jayashree, Giuseppe, & Theresa

Link

# Project Overview

- Briefly describe the original dataset and the hypothesis you've formulated.
  - The dataset provides information on shark attacks, including:
    - Date of incident, Year, Type of attack, Country, State, Location, Activity at the time of attack, Victim's name, sex, and age, Shark species involved, Source of information
  - Various factors influence the likelihood of shark attacks.
    - Factors to explore:
      - Activity type: Swimming, surfing, fishing, etc.
      - Geographical location: Countries, states, specific beaches
      - Shark species involved: Tiger shark, bull shark, etc.

**Hypothesis:**
Shark attack incidents are more frequent in warmer coastal regions in the US with high human activity in water-based recreational activities, compared to cooler regions

# Project Overview

- Explain the **structure and process** of your data cleaning and analysis.
  - Daily data cleaning was based on the learnings of the daily lectures and labs
  - At first impression it was clear the data was filled with null values, and unstructured inputs
  - Next, the team decided which data cleaning tools to use to make the data more digestible
- Unique data cleaning techniques or methods you've employed.

**First Procedure**

```
[16] 1 #Cleaning the column name:
     2 #1. Strip unwanted spaces
     3 #2. Replace space with underscore.
     4 shark_df.rename(columns ={col :col.strip().lower().replace(' ','_') for col in shark_df.columns}, inplace = True)
     5 print(red_color+f"After cleaning{reset_color}\n{shark_df.columns}")
```

```
After cleaning
Index(['date', 'year', 'type', 'country', 'state', 'location', 'activity',
       'name', 'sex', 'age', 'injury', 'unnamed:_11', 'time', 'species',
       'source', 'pdf', 'href_formula', 'href', 'case_number', 'case_number.1',
       'original_order', 'unnamed:_21', 'unnamed:_22'],
      dtype='object')
```

# Data Wrangling and Cleaning

1. Discuss the significant data cleaning challenges you encountered (missing data, duplicates, formatting issues, etc.).
    a. Filling the null values (sex, activity, location, and species) - achieved
    b. Changing the date column to date format - achieved
    c. Changing the time column to time format - not achieved
    d. Converting the data type of age (example 3 to 6 months) - not achieved

2. Explain how you resolved these challenges
    a. Trial & error - every data cleaning method presents its positives and negatives
        i. Some cleaning methods provided different than anticipated results

# Exploratory Data Analysis

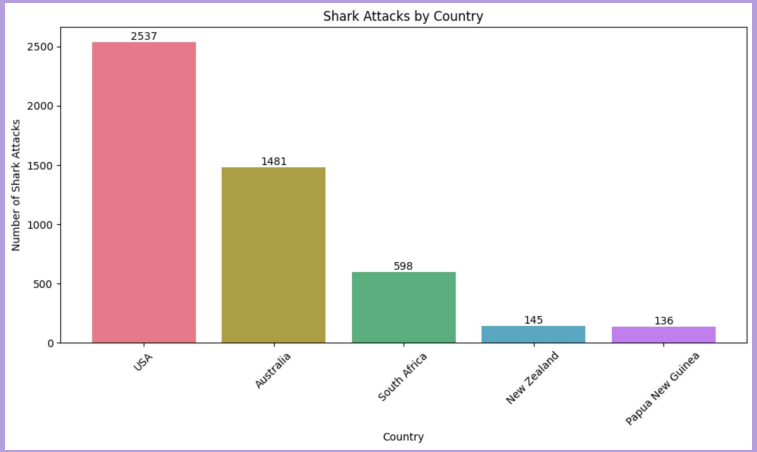
- Discuss the exploratory data analysis methods you used.
  - Groupby was used to test our hypothesis:
    - Country, Year, Sex, Species, and Activity
- Share insights and interesting patterns you found.
  - In the end we only considered the US to align with our hypothesis, and the top 5 states

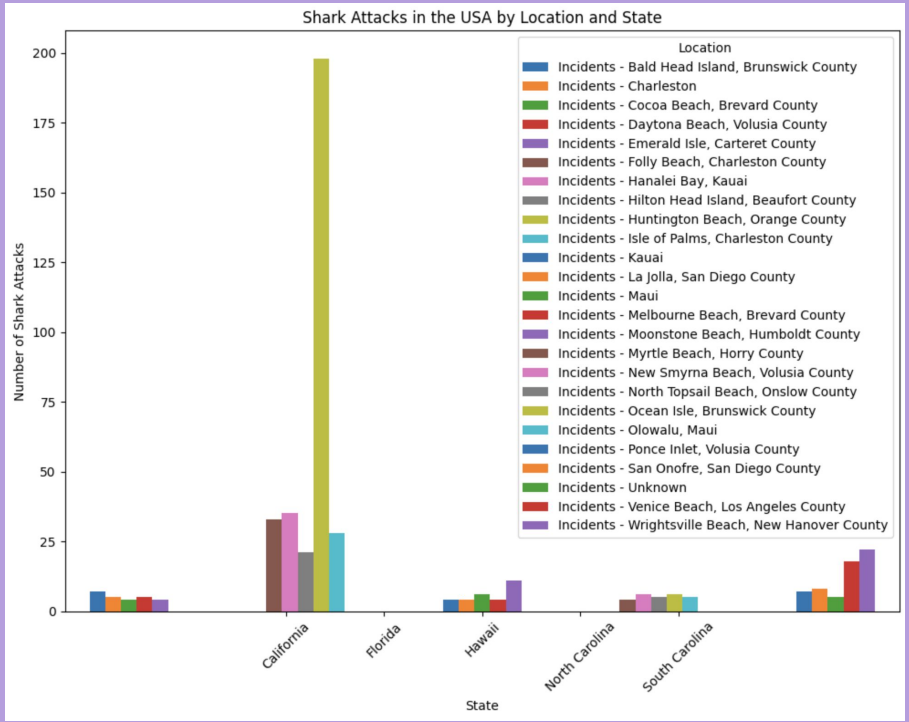
```
Top 5 of column:  Country
Country
USA              2538
AUSTRALIA        1481
SOUTH AFRICA      597
NEW ZEALAND       144
BAHAMAS           136
Name: count, dtype: int64
```

| Country | State | Location | Incidents |
|---|---|---|---|
| USA | Florida | New Smyrna Beach, Volusia County | 198 |
| | | Daytona Beach, Volusia County | 35 |
| | | Cocoa Beach, Brevard County | 33 |
| | | Ponce Inlet, Volusia County | 28 |
| | | Melbourne Beach, Brevard County | 21 |
| | Hawaii | Unknown | 11 |
| | | Maui | 6 |
| | | Hanalei Bay, Kauai | 4 |
| | | Kauai | 4 |
| | | Olowalu, Maui | 4 |
| | California | Huntington Beach, Orange County | 7 |
| | | La Jolla, San Diego County | 5 |
| | | San Onofre, San Diego County | 5 |
| | | Moonstone Beach, Humboldt County | 4 |
| | | Venice Beach, Los Angeles County | 4 |
| | South Carolina | Myrtle Beach, Horry County | 22 |
| | | Isle of Palms, Charleston County | 18 |
| | | Folly Beach, Charleston County | 8 |
| | | Charleston | 7 |
| | | Hilton Head Island, Beaufort County | 5 |
| | North Carolina | Ocean Isle, Brunswick County | 6 |
| | | Emerald Isle, Carteret County | 6 |
| | | Wrightsville Beach, New Hanover County | 5 |
| | | North Topsail Beach, Onslow County | 5 |
| | | Bald Head Island, Brunswick County | 4 |

```
Attacks by sex
sex
M          5567
F           775
Unknown     581
dtype: int64
```

# Visualization



# Major Obstacle

- Discuss the biggest obstacle or mistake you encountered during this project.
  - Finding a balance between brainstorming a hypothesis, while confidently using the data cleaning techniques
  - Writing pivot because it kept showing a ‘duplicate’ error message specifically for the index column

- Share what you learned from it and how it influenced your project.
  - How to collaborate effectively 💫 Our team learned from each other since everyone picked up different insights from the daily lectures & labs
    - Using Google Collab, Google Doc, & Slack both efficiently and effectively

# Conclusion and Insights

- Discuss whether your initial hypothesis was supported or refuted.
  a. Our initial hypothesis was **supported** by the data

- Share any surprising insights or findings.
  a. We were not surprised - our hypothesis aligned with the data the US States with the highest incidents are:
     - Florida, Hawaii, California, South Carolina, North Carolina

- Discuss potential implications of your findings.
  a. These States should offer a ‘Shark Security’ service for beach goers to feel safe, and comfortable
  b. The next step would be to deep dive into which beaches these shark attacks occurred to make the most impact with this service
     - Example in Hawaii: Maui, and Kauai

# Quest: Shark Attack

Toan, Jayashree, Giuseppe, & Theresa


Thank you!