Shark Attack

David, Ishwarya, Josean, Vanessa

Problem/Opportunity

A shark safety startup seeks government contracts with coastal towns.

- they focus on safety through education, using shark cage diving experiences as an education tool
- they want to know where and when to focus their initial efforts to maximize revenue with their limited starting budget
- they have a limited amount of trained staff, which limits how many locations they can manage at once

Hypothesis

We hypothesize that the following 2 states and months will be the best places and times to focus their efforts on marketing to secure contracts and maximize revenue opportunities:

- California, USA from May to September
- New South Wales, Australia from Oct to April

Project Overview

Our data cleaning process included:

a. Eliminating rows and columns of all null values and columns not relevant to our research

b. Standardizing categorical columns and converting dates to datetime

c. Cleaning strings by eliminating spaces, letters, punctuation etc.

d. Analyzing null values and filling null values as we deemed appropriate

```
shark_attack['Sex'].value_counts()

# Number of NaNs in the 'Sex' column
nans_to_fill = shark_attack['Sex'].isna().sum()

# Calculate how many should be 'M' (Male) and 'F' (Female)
n_males = int(0.88 * nans_to_fill)
n_females = nans_to_fill - n_males

values_to_fill = ['M'] * n_males + ['F'] * n_females

# Shuffle the list to randomize the distribution
np.random.shuffle(values_to_fill)

# Fill NaN values in the 'Sex' column
shark_attack.loc[shark_attack['Sex'].isna(), 'Sex'] = values_to_fill

shark_attack['Sex'].value_counts()
```

```
Sex
M     6113
F      852
na       4
Name: count, dtype: int64
```

Data Wrangling and Cleaning

The most significant data cleaning challenge we faced was converting our dates to datetime type.

```
formats = [
    "%d %b%Y",    # Example: 14Jun2023
    "%d %b %Y",   # Example: 14 Jun 2023
    "%Y%m%d",     # Example: 20230614
    "%d %B %Y",   # Example: 14 June 2023
    "%d %b%Y",    # Example: 14Jun'23
    "%d %b %Y",   # Example: 14 Jun '23
    "%d%b%Y",     # Example: 14Jun'23
    "%d%b %Y",    # Example: 14Jun '23
    "%d %b%y",    # Example: 14Jun23
    "%d %b %y",   # Example: 14 Jun 23
    "%d%b%y",     # Example: 14Jun23
    "%d%b %y",    # Example: 14Jun 23
]

for fmt in formats:
    try:
        shark_attack['Date'] = pd.to_datetime(shark_attack['Date'], format=fmt)
        break  # Stop trying once successful
    except ValueError:
        continue  # Try the next format if ValueError is raised

shark_attack['Date'] = pd.to_datetime(shark_attack['Date'], errors='coerce')
shark_attack['Date'][500:550]
```

Data Wrangling and Cleaning

A secondary challenge was the large variety of misspellings across countries, states and locations that needed to be cleaned up.

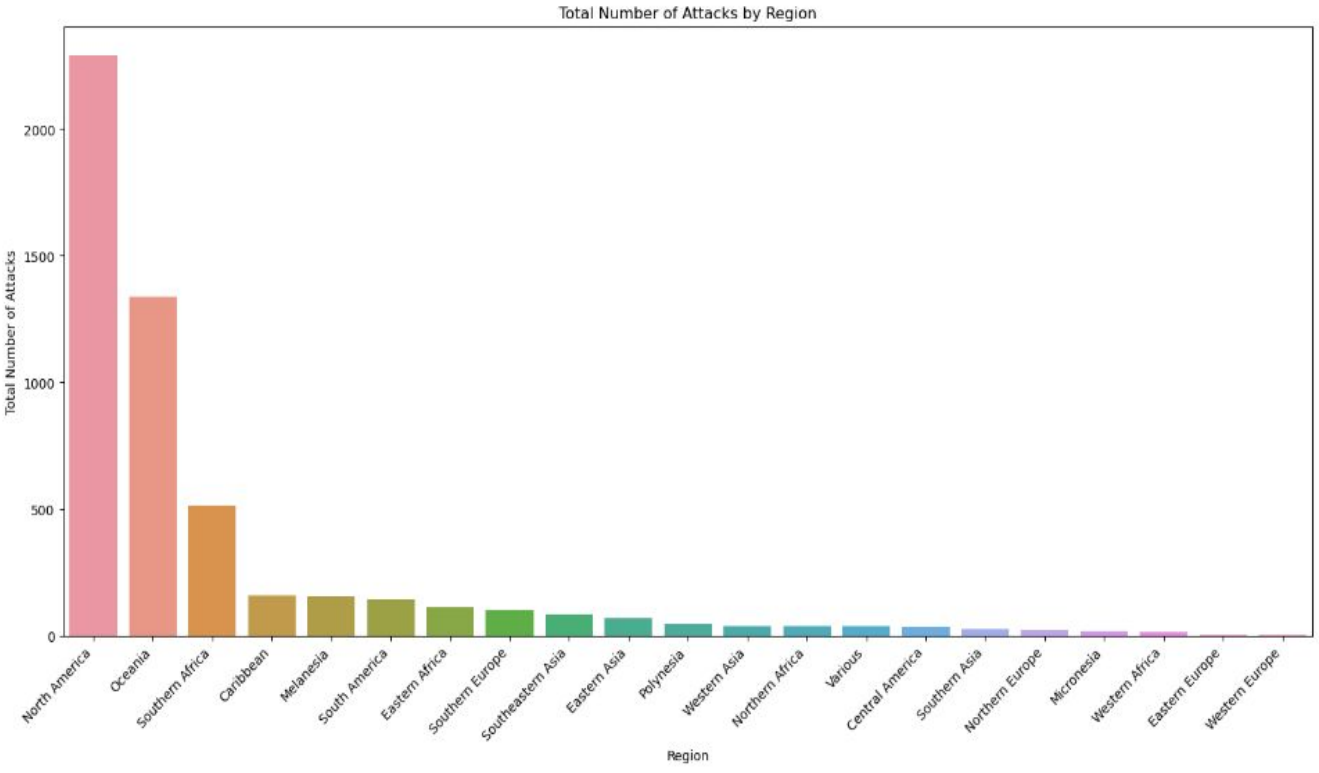
```
"""
We have corrected the countries that were misspelled on our first review of the country list.
"""

corrections = {'Mexico':'MEXICO',' PHILIPPINES':'PHILIPPINES',' TONGA':'TONGA','CEYLON':'CEYLON (SRI LANKA)'
               ,'COLUMBIA':'COLOMBIA','EGYPT ':'EGYPT','Fiji':'FIJI','INDIAN OCEAN?':'INDIAN OCEAN',
              'MALDIVE ISLANDS':'MALDIVES','MEXICO ':'MEXICO','Maldives':'MALDIVES','MeXICO':'MEXICO',
              'NICARAGUA ':'NICARAGUA','NORTH ATLANTIC OCEAN ':'NORTH ATLANTIC OCEAN', 'New Zealand':'NEW ZEALAND',
              'OKINAWA':'JAPAN','PACIFIC OCEAN ':'PACIFIC OCEAN', 'RED SEA?':'RED SEA','REUNION':'REUNION ISLAND',
              'SAN DOMINGO':'SANTO DOMINGO','ST HELENA, British overseas territory':'ST HELENA','ST. MAARTIN':'ST MARTIN',
              'ST. MARTIN':'ST MARTIN','SUDAN?':'SUDAN','UNITED ARAB EMIRATES (UAE)':'UNITED ARAB EMIRATES','ENGLAND':'UNIT
              'ST KITTS / NEVIS':'ST KITTS AND NEVIS','RED SEA / INDIAN OCEAN':'RED SEA','ANDAMAN / NICOBAR ISLANDAS':'ANDA
              'IRAN / IRAQ':'IRAN','SOUTHWEST PACIFIC OCEAN':'SOUTH PACIFIC OCEAN','ITALY / CROATIA':'ITALIA', 'YEMEN ':'YE
              'CEYLON (SRI LANKA)':'SRI LANKA','ASIA?':'ASIA',
             }

shark_attack['Country']=shark_attack['Country'].replace(corrections)
```

Exploratory Data Analysis

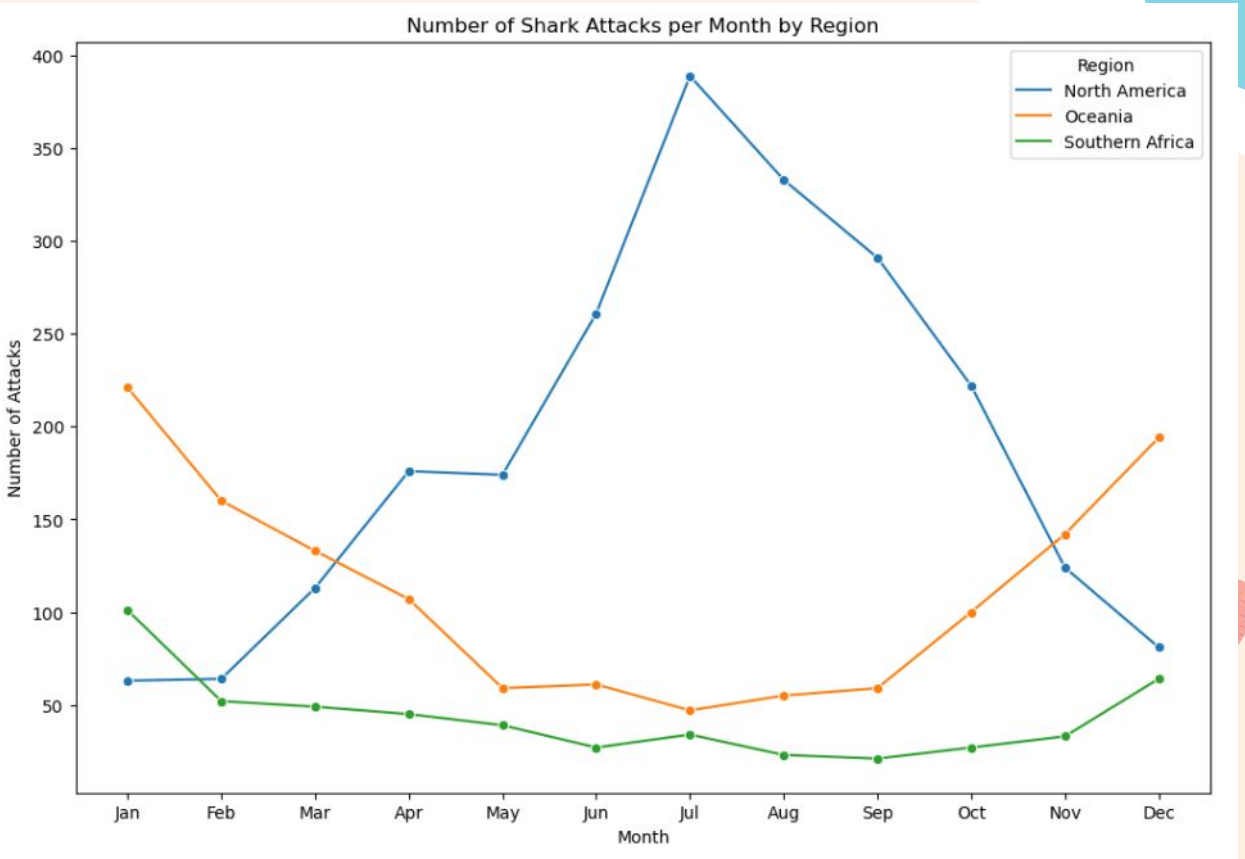
- The methods that we used were:
  a. **Summarize statistics**
     i. We filtered the data and created pivot tables to create comparison tables
  b. **Relationship between variables**
     i. We merged duplicates and dropped some rows with empty information
  c. **Visualization on graphics**
     i. Linear Graphic Chart
     ii. Bar Chart Graphic



North America, Oceania and Southern Africa are our top regions for shark attacks.

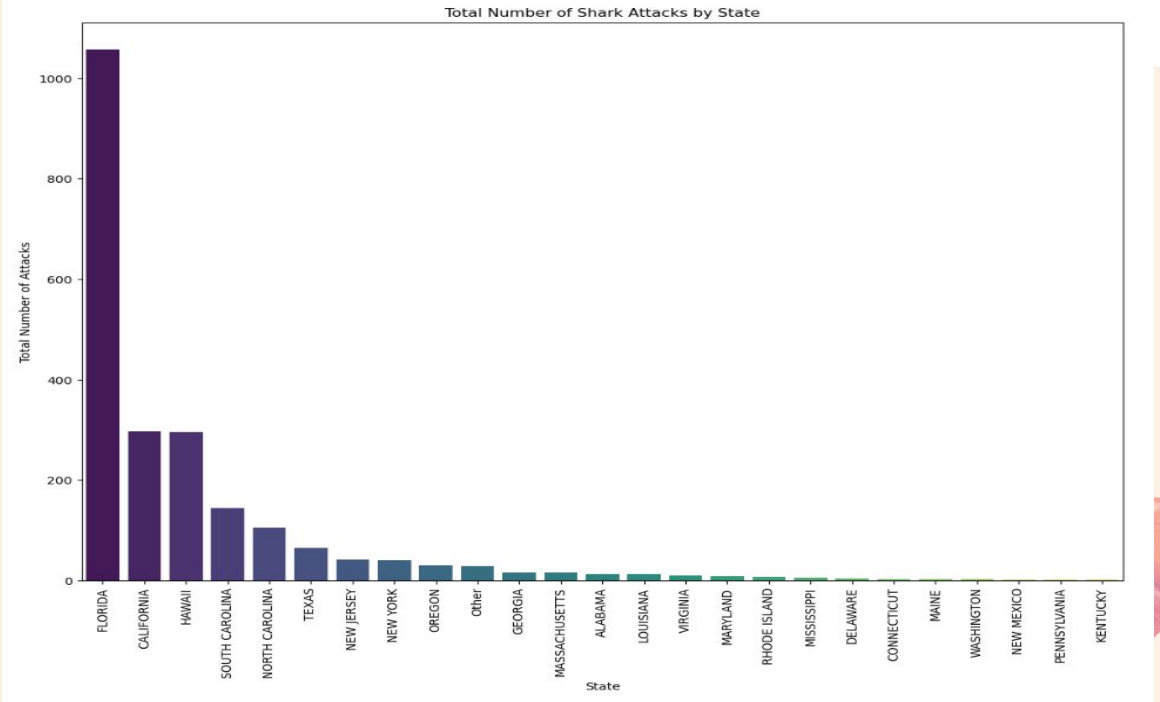
Our analysis will then focus on the top 2.

- Summer time in North America: June to August
- Summer time in Oceania: December to February
- Summer time in Southern Africa: November to March



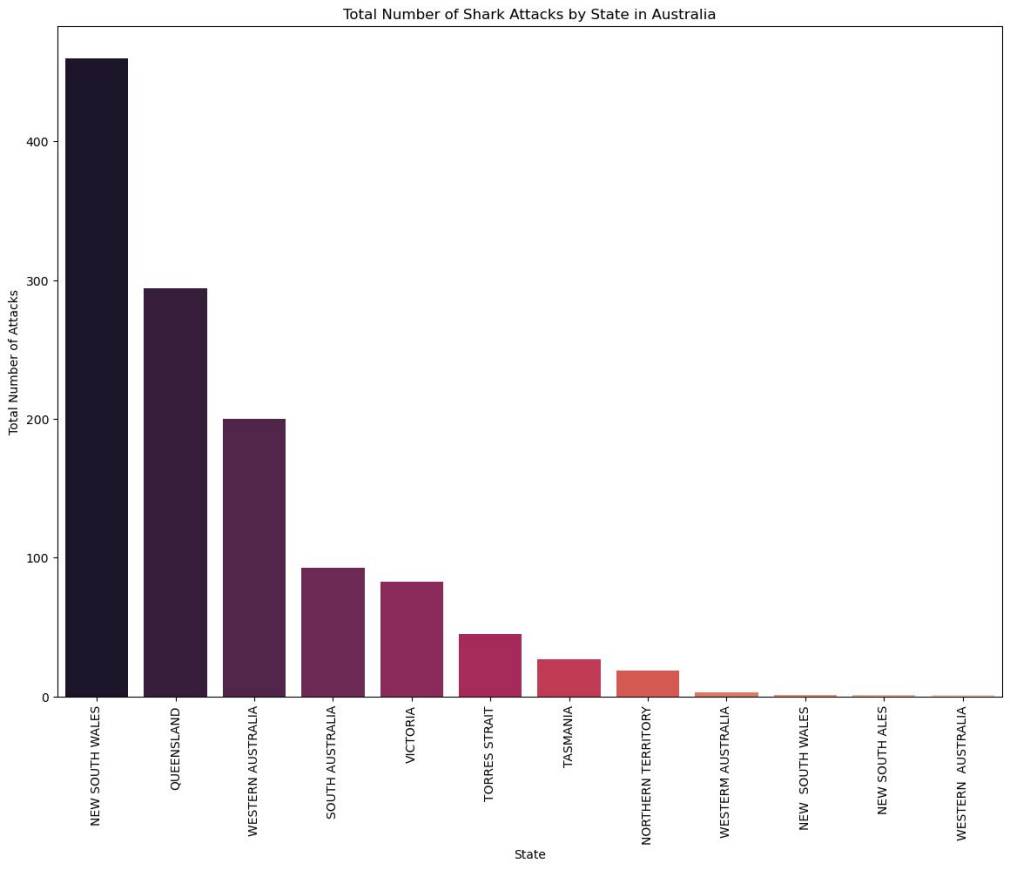
USA is the #1 region in North America.

Florida, California and Hawaii are the top 3 states, with Florida at #1 by a very high margin.



Australia is the #1 region in Oceania.

New South Wales, Queensland, and Western Australia are the top 3 states, with New South Wales at #1 by a very high margin.



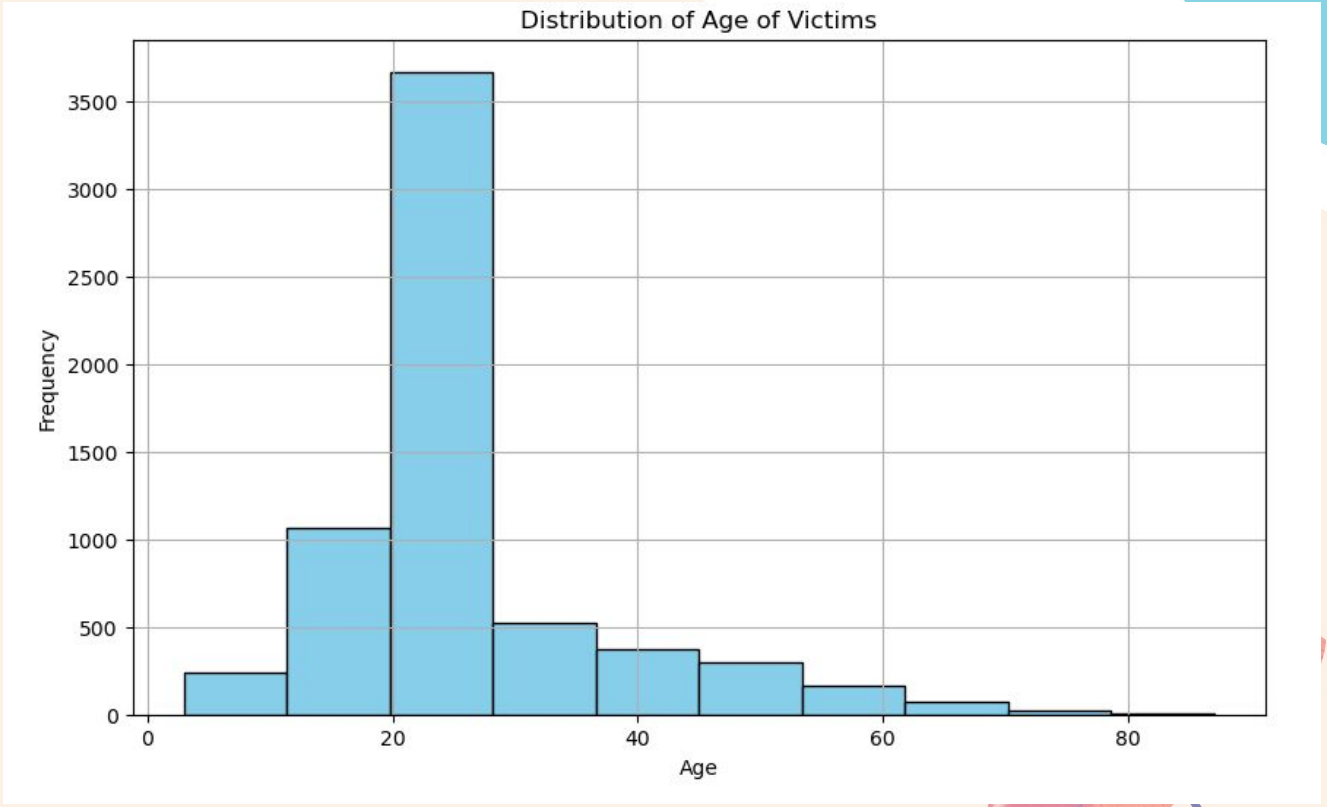
```
Location
New Smyrna Beach, Volusia County                          191
Daytona Beach, Volusia County                              31
Ponce Inlet, Volusia County                                28
Melbourne Beach, Brevard County                            20
Cocoa Beach, Brevard  County                               18
                                                          ...
Quarter mile south of Ponce de Leon Inlet, Volusia County   1
Off Zelda Boulevard, Daytona Beach, Volusia Countyy         1
Tigertail Beach, Collier County                             1
Marco Island, Collier County                                1
Mosquito Inlet (Ponce Inlet), Volusia County                1
Name: count, Length: 550, dtype: int64
```

In Florida, New Smyrna Beach in Volusia County was the top location for shark attacks.

```
Location
nan                                  15
Sydney Harbor                        10
Sydney                               10
Brisbane River                        9
Near Thursday Island                  9
                                     ..
Opal Reef                             1
Stockton Beach                        1
7 km off Trigg surf beach, Perth      1
Lefthanders Beach, Margaret River     1
Roebuck Bay                           1
Name: count, Length: 1092, dtype: int64
```

In New South Wales, Sydney was the top location for shark attacks.

The vast majority of shark attack victims are in their 20s. This could be due to a variety of factors, such as they are more likely to engage in water-related activities, like surfing, swimming, and diving, which increase their likelihood of encountering sharks. Additionally, they might be more adventurous and willing to explore deeper waters where shark encounters are more common.



Conclusion and Insights

- **Our initial hypothesis:** California, USA from May to September and New South Wales, Australia from Oct to April
- **The data shows:** Florida,Volusia County June through September. Australia, New South Wales (where Sydney is)
- We were surprised to see that the vast majority of victims were male, by a shocking margin.
- While we didn't expect it to be the top result, we were surprised the Caribbean did not have more reported incidents.
- One major insight we discovered is that shark attack incident peaks lined up with the summer months of each region. This could be due to warmer waters during the summer months, which might attract certain shark species closer to shore, thereby increasing the likelihood of interactions with humans engaging in water activities

Implications of findings

- **Opportunity for lean operation:** The company can optimize resources by not maintaining permanent bases in high-risk areas simultaneously. There can be flexibility in resource allocation based on seasonal fluctuations in shark attack frequency.
- **Customized educational materials:** Further research can be done into the types of incidents to influence the educational materials sold by the company and the experiences they provide to their clients as well as their staff training.
- **Diversification of experiences:** periods of lower risk of shark attacks can offer a wider range of water-based experiences. Alternative activities can be arranged during peak seasons, where there is higher susceptibility to shark encounters
- **Continuous Monitoring and Adaptation:** Ongoing monitoring of shark activity trends is essential. Proactive adjustment of operations and risk management practices can be made based on evolving patterns.

Thank you!

David, Ishwarya, Josean, Vanessa