SHARK ATTACKS

Luna Tissera & Rori Chávez

Hypothesis

In coastal regions with high recreational activity, such as the east coast of the United States, Australia and South Africa, shark attacks are likely to be more frequent during the summer months due to the greater influx of people in the water and increased recreational water activities during this season.

BRIEFLY DESCRIBE THE ORIGINAL DATASET

The original dataset had 23 columns and 6969 rows, in which shark attacks over the past centuries were recorded.
Most of the columns contained object and float of data type.

To observe the records of each column, we used the unique method as follows: df[Gender].unique().

This allowed us to observe that the data were not standardized, and within each column, the data were mostly with different values.

Following the example of the Gender column, we had 'F', 'Female', 'Masc', 'M', etc.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6969 entries, 0 to 6968
Data columns (total 23 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            6944 non-null   object
 1   Year            6942 non-null   float64
 2   Type            6926 non-null   object
 3   Country         6894 non-null   object
 4   State           6462 non-null   object
 5   Location        6379 non-null   object
 6   Activity        6358 non-null   object
 7   Name            6724 non-null   object
 8   Sex             6365 non-null   object
 9   Age             3950 non-null   object
 10  Injury          6909 non-null   object
 11  Unnamed: 11     6382 non-null   object
 12  Time            3418 non-null   object
 13  Species         3812 non-null   object
 14  Source          6925 non-null   object
 15  pdf             6799 non-null   object
 16  href formula    6819 non-null   object
 17  href            6796 non-null   object
 18  Case Number     6798 non-null   object
 19  Case Number.1   6797 non-null   object
 20  original order  6799 non-null   float64
 21  Unnamed: 21     1 non-null      object
 22  Unnamed: 22     2 non-null      object
dtypes: float64(2), object(21)
memory usage: 1.2+ MB
```

STRUCTURE AND PROCESS OF OUR DATA CLEANING AND ANALYSIS

We decided to focus on the last 20 years, so we created a filter to remove the years that we were not going to use. We did the same with the columns that we were not going to use and only kept the columns that were relevant to test our hypothesis, so we ended up with 11 columns and 2324 rows.

Once we had selected the columns we wanted, we began standardizing the values within each column.

- **Column Type:** We created a dictionary with four categories (keys) that we used to standardize the registe into four categories such as 'Unprovoked', 'Provoked', 'Unconfirmed', 'Exogens'. Using the replace method on the column of our dataframe.

- **Column Country:** We used the same method as in the previous case and converted the records to lowercase using the str.title method, which retains the first letter in uppercase.

- **Column Activity:** For this case, we used Regex 101 to obtain activity patterns and categorized them into 5 activities. We used the regex pattern in a dictionary and then created a function that we reused for the rest of the cases.

- **Column Species:** We followed the same procedure as in the previous column. But in this case, as there were many records of different sharks, we decided to only create a regex pattern to retain the 6 most well-known ones.

- **Column Date & Column Year.**

Data Wrangling and Cleaning

Significant data cleaning challenges

- Transform "Date" to Date format.
- Use regex and make dictionaries to define categories.
- Define the function

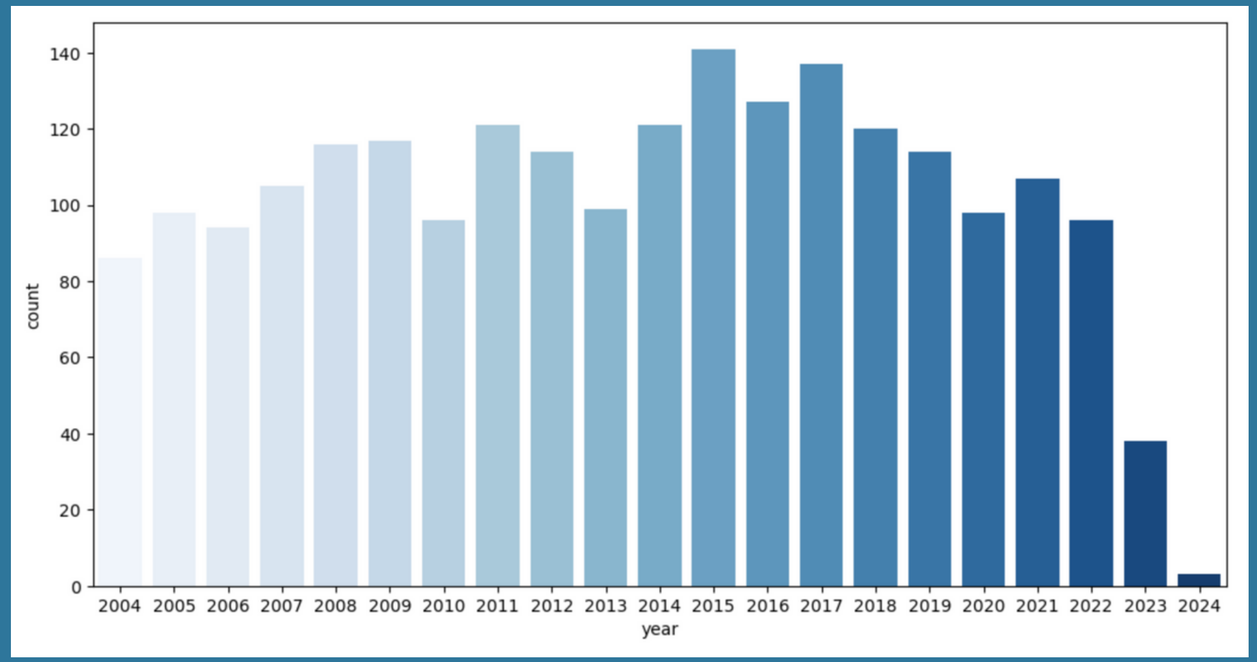
Explain how you resolved these challenges

- With the first was to leave it last because when we tried to get started we didn't even know how to start. And then, we decided to delete the 100 rows that were giving us problems because they were null.
- We asked how to use it and defined the categories to make it clearer. What didn't fit in, we marked as "others".
- Here as the problem was that they were not being deleted because we could not fill in everything. We decided that before applying the function we should change everything that was empty for str so that when we could not find how to replace it, we could replace it with another one.

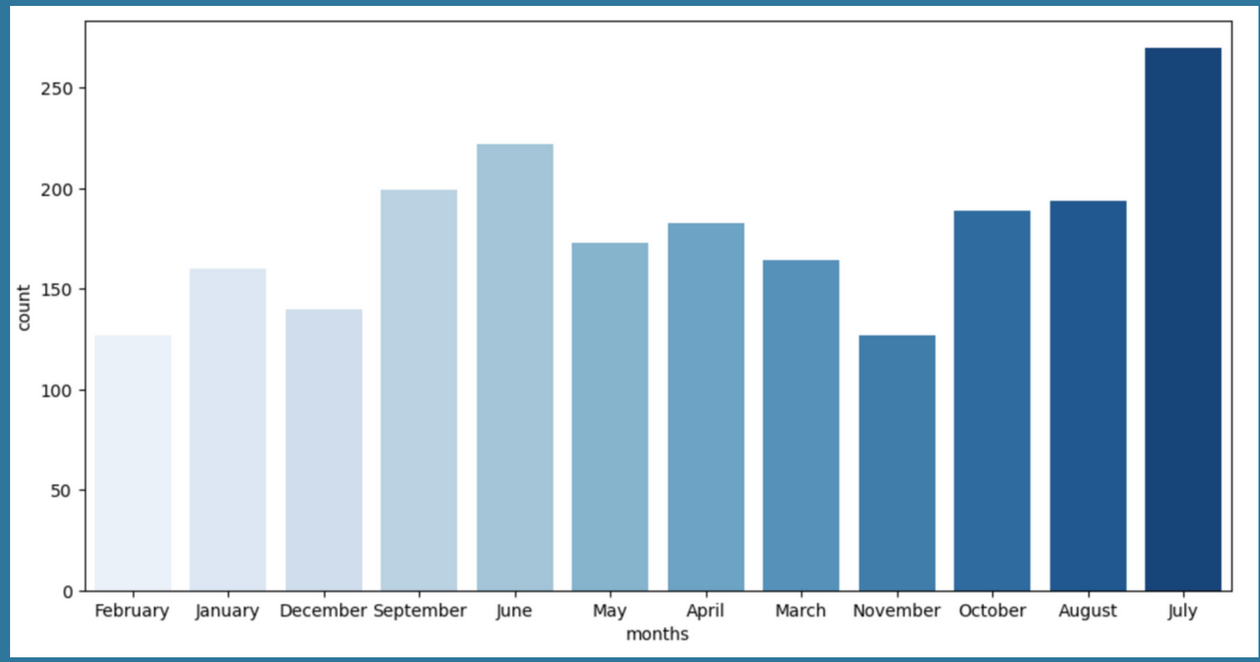
Exploratory data analysis + Conclusions

Graphics used: Bar charts

What was the year with the most attacks?

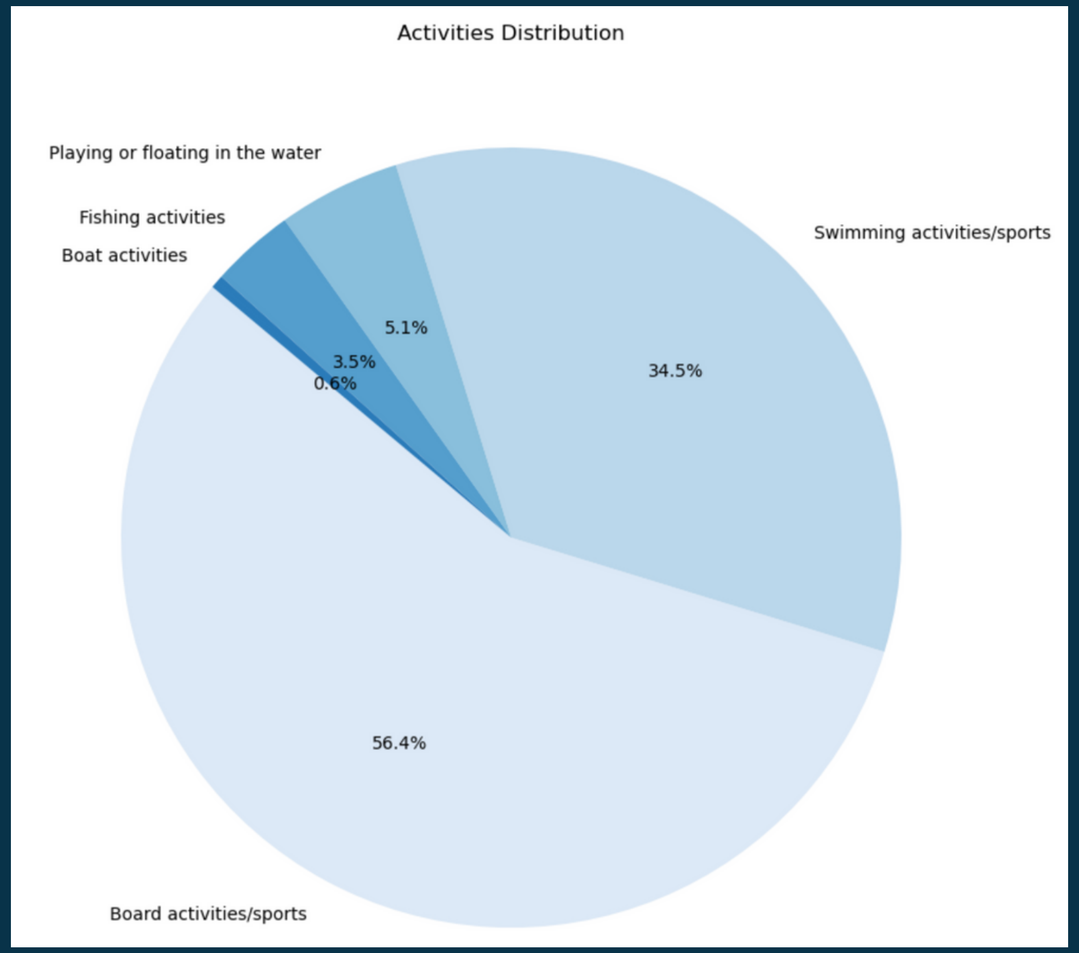


What was the month with the most attacks?

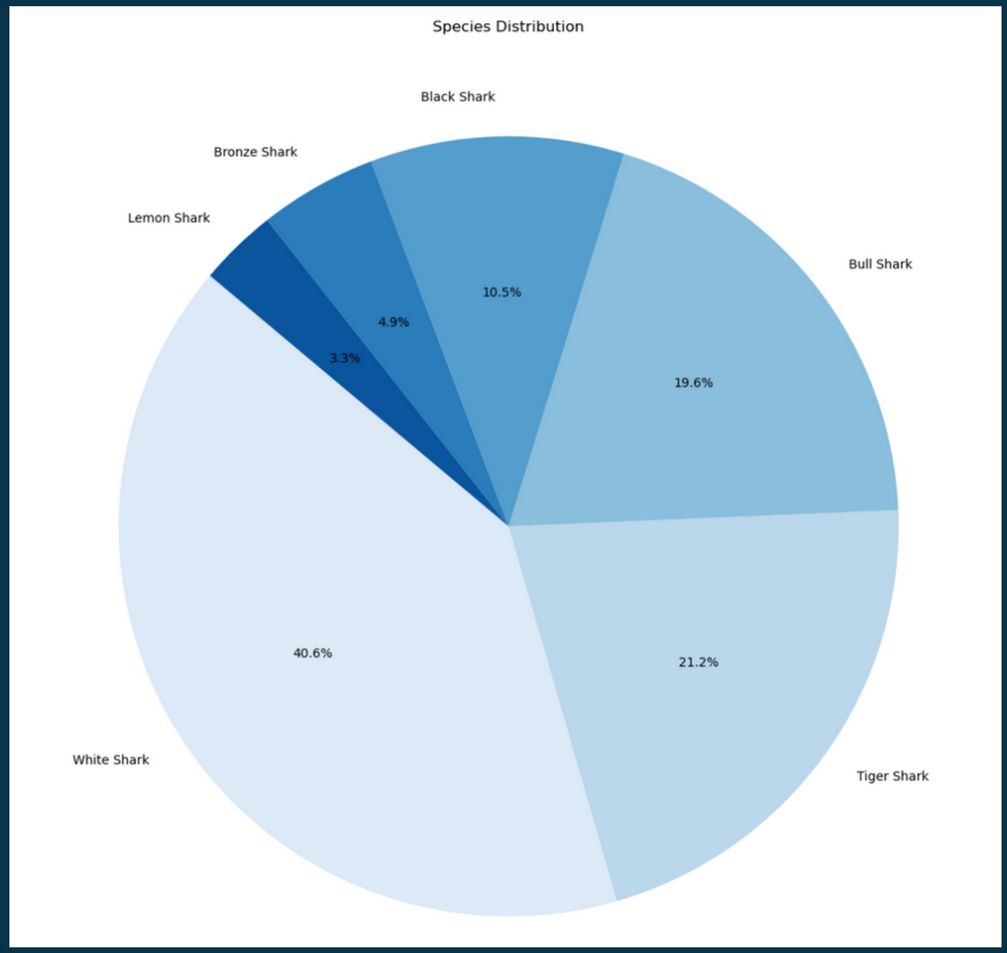


Exploratory data analysis + Conclusions

Graphics used: Pie Chart



What kind of activities were the people who suffered shark attacks doing?

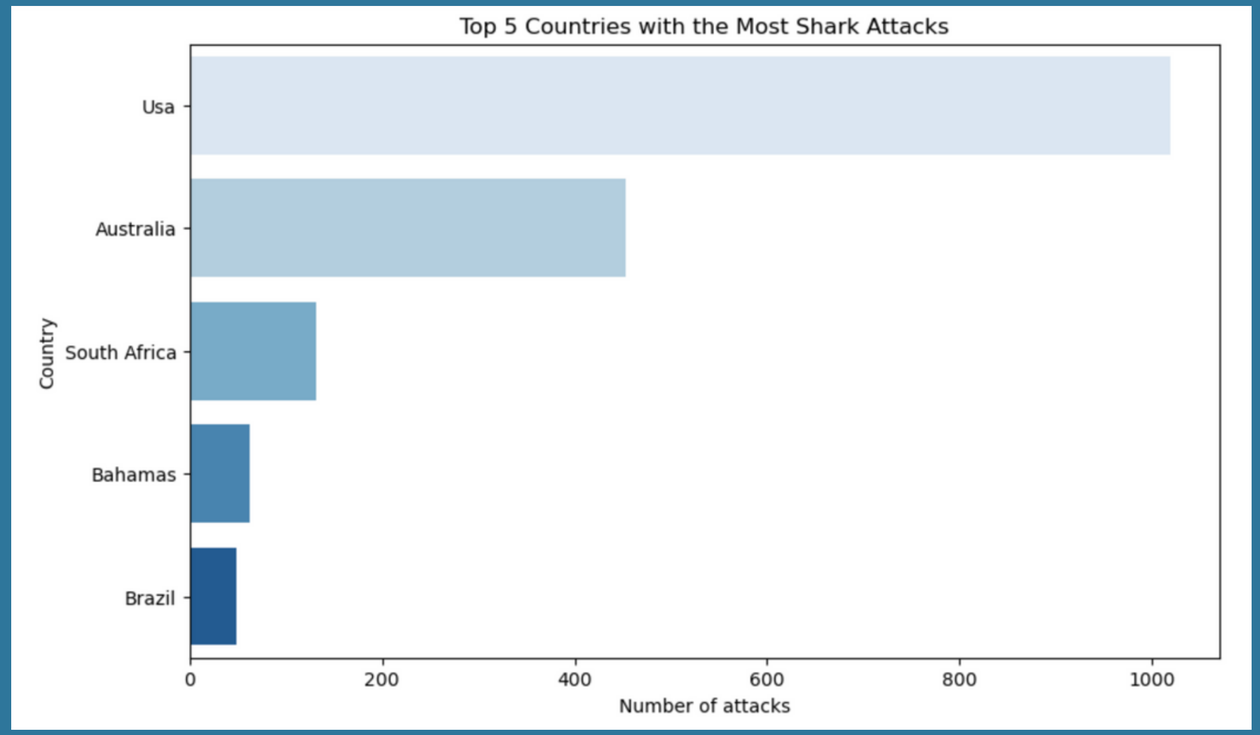


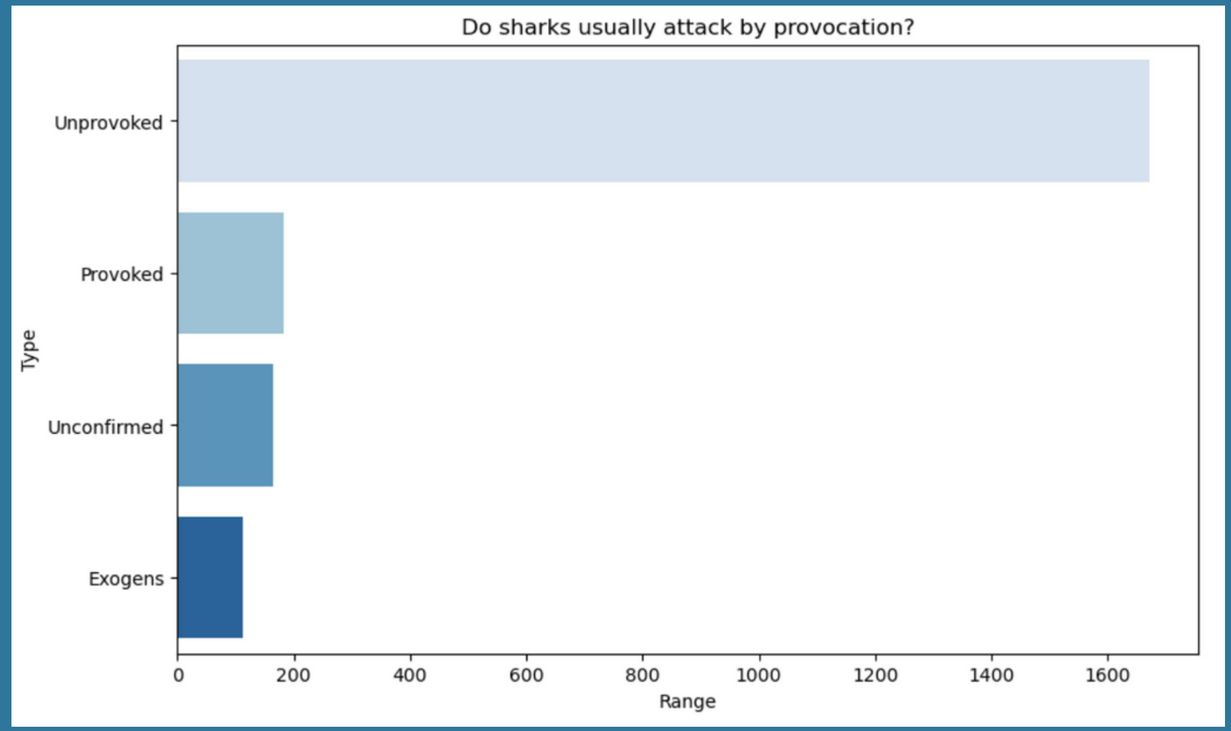
What species of sharks do they attack most often?

Exploratory data analysis + Conclusions

Graphics used: Bar charts

Top 5 countries with the most shark attacks





OUR BONUS:

Do sharks usually attack by provocation?

Major obstacle

What we learned?

- How to get started.
- Getting organized is the key.
- That the work is long sometimes is not so complex, so calm down.

THANK YOU!

SHARK ATTACKS

Luna Tissera & Rori Chávez