“SHARK ATTACK”

Exploring Human-Shark Interactions:
Insights and Recommendations for Public Safety and Marine Conservation

Dream Team: “Yimara McMitchell, Siam Arman, Rosolino Mangano,
Marina Sofía Martín Litvinova & Kenny Huang”.

Introduction:

In this analysis, the "Dream Team", aims to explore the global shark attack data to identify patterns and factors influencing the frequency and severity of shark attacks. The goal of our project is to help the public understand human-shark interactions which can help reduce the number of fatal or non-fatal events. The data set we used can be found on the website "https://www.sharkattackfile.net /incidentlog.htm" This website documents shark attacks occurring all around the world.

The dataset includes 6,968 observations. The relevant variables of the data include date, year, type of attack, country, state, location, activity, name, sex, age, and injury. By examining this data, we will uncover insights that can inform safety measures. The objective of this project is to analyze shark data to identify patterns and factors that influence the frequency and severity of shark attacks. This will help enhance public safety and understanding of human-shark interactions.

Understanding human-shark interactions can be effective in developing safety protocols. As stated by Burgess et al. (2010) in their work about shark ecology and public safety, "understanding shark behavior and movement patterns is essential for formulating strategies that reduce the risk of shark attacks." (Burgess et al., 2010, p. 961) This shows the significance of this project and aligns with the goals of enhancing human safety.

Research Questions:

Our Analysis seeks to answer several key questions that will be used to develop strategies to decrease shark attack risks.

1. Which countries/areas experience the highest shark attacks?
2. Are most shark attacks provoked/unprovoked?
3. How have shark attacks changed over time?
4. What factors increase the likelihood of a shark attack being fatal?
5. Which species of sharks are most involved in the attacks?

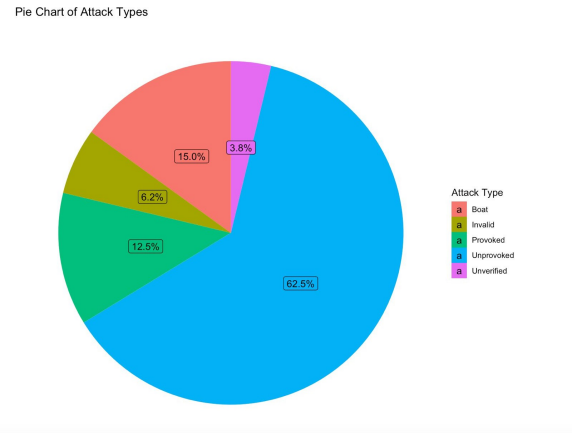
Data Cleaning Reprocessing and Exploration:

In preparing our dataset, we decided to filter our dataset by taking out the incomplete records. This was an important step in removing entries disrupting the dataset, specifically those missing the "Year" variable of when the shark attack occurred. This variable is needed to analyze the trends and patterns over time. To help enhance the quality of the dataset, we excluded all entries where the data was unavailable. This data cleaning helps enhance the accuracy of our analysis. We began exploring the dataset by checking for missing values, consistency, and completeness.

While examining the data we found 169 missing values in the original order. We also found that the minimum value is 2 meaning it's the lowest value in the data set. The 1st quartile is 1702. This means that 25% of all the observations fall below this value. The mean is 3401, the average value of the set. The median is also 3401, which is the middle of the values when data is arranged from lowest to greatest. The 3rd quartile is 5100, 75% of all observations are below this value. The maximum value is 6802, the highest value observed in the dataset. The mode is logical, thus showing that the variables contain

true/false values. This helps with the distribution of the data set and shows 6968 observations in the dataset.

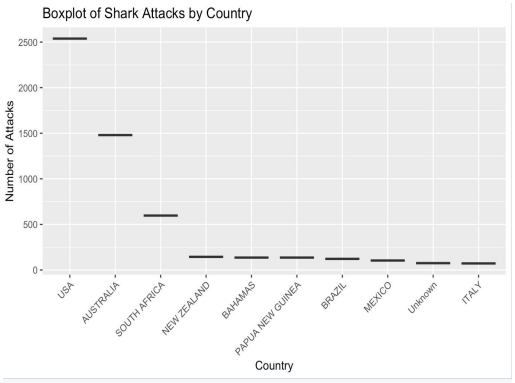
Pie Chart-Attack Types



The pie chart provides a clear visualization of different attack types confirming that the most common type of shark attacks are unprovoked. Unprovoked attacks make up 62.5% of the total dataset. Provoked attacks account for 15% while those involving boats account for 12.5%.

Unverified attacks constitute 6.2% of the data set, indicating that there are details in the dataset that may be uncertain. Lastly, invalid attack records are the least common, making up 3.8% of the dataset.

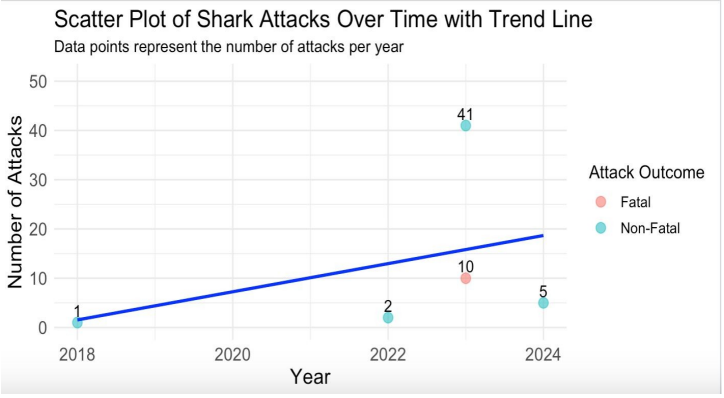
Boxplot-Attacks by Country



	Country	Num_Attacks
	<chr>	<int>
1	USA	2538
2	AUSTRALIA	1480
3	SOUTH AFRICA	597
4	NEW ZEALAND	144
5	BAHAMAS	136
6	PAPUA NEW GUINEA	136
7	BRAZIL	122
8	MEXICO	104
9	NA	75
0	ITALY	72

This box plot shows the United States has the highest number of shark attacks with a total of 2538 incidents followed by Australia with 1480 attacks and South Africa with 597 incidents. We can make a hypothesis that the country with the highest number of shark attacks has the highest number because of their high population, is near the sea where sharks are, and have a culture of going to the sea.

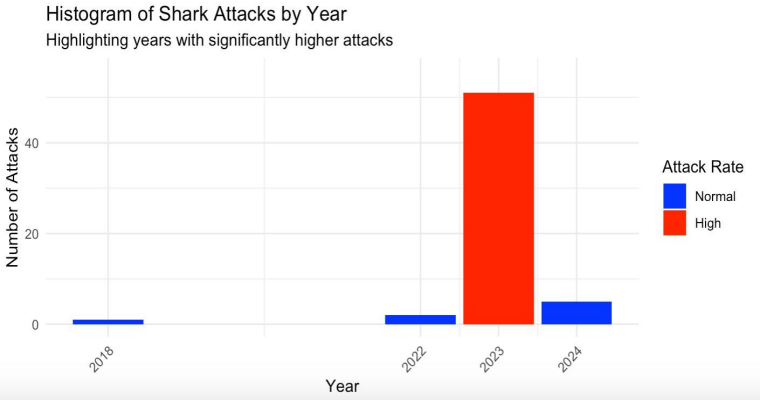
Scatter Plot-Attacks Overtime



The Scatter Plot illustrates the breakdown year by year of shark attacks by severity. Data shows that from 2018 to 2022 there was a slight increase in attacks.

Then there is a significant increase in attacks in 2023. The trend line used shows the change in the frequency of attacks over time which helps understand human-shark interaction patterns.

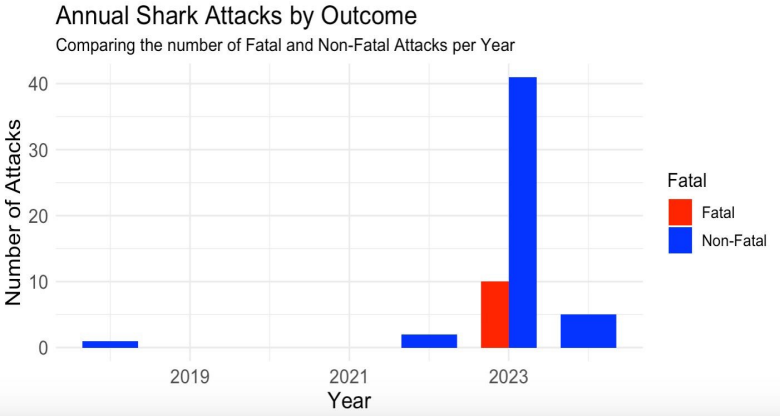
Histogram- Shark Attacks by Year



The histogram shows the distribution of shark attacks by year highlighting a spike in attacks during 2023.

Possible factors that would have influenced this were the increase of activities in shark-infested waters, or environmental changes. This can be used to identify the trends in the data over time.

Bar Graph- Annual Shark Attacks by Outcome

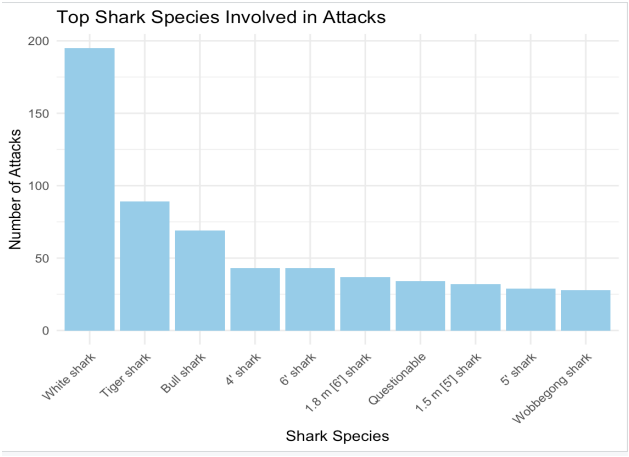


This bar graph highlights the previous histogram. It shows that although there was a spike in shark attacks in 2023, most of them were non-fatal.

The years 2019 and 2021 show that attacks were relatively low in comparison to other years. From this graph, we can see that there was a big gap between 2019 to around 2022.

We can likely guess this happened because of COVID-19 where the lockdown prevented everyone from going to the sea where the shark attacked. This shows the idea that shark attacks are not dependent on sharks to happen but dependent human activities to happen. For a human to be attacked by a shark, they need to be near an ocean, want to go in the ocean, and have an event for a shark to attack a human.

Bar Graph -Shark Species Involved in Attacks




```
> print(top_species_attacks)
# A tibble: 10 × 2
   Species            Num_Attacks
   <chr>                    <int>
 1 White shark                195
 2 Tiger shark                 89
 3 Bull shark                  69
 4 4' shark                    43
 5 6' shark                    43
 6 1.8 m [6'] shark            37
 7 Questionable                34
 8 1.5 m [5'] shark            32
 9 5' shark                    29
10 Wobbegong shark             28
>
```

The bar graph shows the frequency of shark attacks associated with different shark species, The species most involved range from Blue Shark, Bull Shark, Tiger Shark, and White Shark. According to Global Shark Attack, the White Shark is “the super-predator; it is without question the most formidable of all sharks. The White Shark swims stiffly and is capable of great speed.” (Global Shark Attack File, 2005) This characterization can explain why the White Shark tops the list of shark species involved in attacks. The graph also shows incidents in which shark involvements were not confirmed. This can illustrate the uncertainties in reporting and identifying shark species which can lead to exploration issues.

Model Development and Selection:

We used three primary models such as logistic regression for classification, linear regression for trend analysis, and k-means clustering for geographical analysis. We chose each model based on how well it suited specific research questions such as the likelihood of an attack being fatal, high-risk locations, change in shark attacks over time, high-risk species, and type of attack. The model performances were compared by accuracy and recall metrics. The logistic regression model was shown as being most effective making it the optimal model for predicting shark attack outcomes.

Classification Model

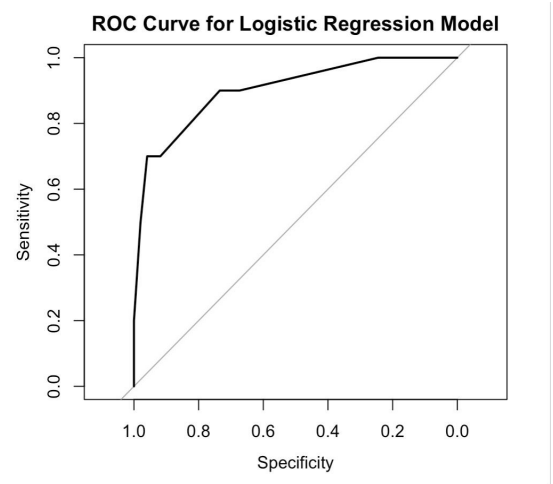
```
Call:
glm(formula = Fatal ~ Year + Country, family = binomial(), data = shark_data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               2145.721   2651.341   0.809   0.4183
Year                        -1.062      1.311  -0.810   0.4180
CountryBAHAMAS               2.051      1.342   1.529   0.1263
CountryCOLOMBIA            -17.771  10754.013  -0.002   0.9987
CountryCUBA                 21.361  10754.013   0.002   0.9984
CountryECUADOR             -17.771  10754.013  -0.002   0.9987
CountryEGYPT               -17.771  10754.013  -0.002   0.9987
CountryFRENCH POLYNESIA    -23.079  10754.015  -0.002   0.9983
CountryINDIA               -16.709  10754.013  -0.002   0.9988
CountryMexico              -17.771  10754.013  -0.002   0.9987
CountryMEXICO                2.894      1.394   2.076   0.0379 *
CountryNEW ZEALAND         -17.771   7604.236  -0.002   0.9981
CountryPHILIPPINES         -17.771  10754.013  -0.002   0.9987
CountryPORTUGAL            -17.771  10754.013  -0.002   0.9987
CountrySAMOA                21.361  10754.013   0.002   0.9984
CountrySPAIN               -17.771  10754.013  -0.002   0.9987
CountryUSA                  -1.385      1.284  -1.078   0.2810
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.699  on 58  degrees of freedom
Residual deviance: 29.167  on 42  degrees of freedom
AIC: 63.167

Number of Fisher Scoring iterations: 18
```



Evaluation of Logistic Regression Classification Model for Predicting Fatalities in Shark Attacks:

We developed a classification model to help predict the likelihood of an attack being fatal based on multiple predictors such as country, year, and whether or not the attack was fatal. Our findings conclude that except for Mexico, the country of the shark attack does not significantly impact the odds of the attack being fatal or not.

```
Confusion Matrix and Statistics

          Reference
Prediction  Non-Fatal Fatal
  Non-Fatal        47     3
  Fatal             2     7

               Accuracy : 0.9153
                 95% CI : (0.8132, 0.9719)
    No Information Rate : 0.8305
    P-Value [Acc > NIR] : 0.05094

                  Kappa : 0.6865

 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9592
            Specificity : 0.7000
         Pos Pred Value : 0.9400
         Neg Pred Value : 0.7778
             Prevalence : 0.8305
         Detection Rate : 0.7966
   Detection Prevalence : 0.8475
      Balanced Accuracy : 0.8296

       'Positive' Class : Non-Fatal
```

The model helped reveal that attacks in Mexico are more likely to be fatal, with a coefficient of 2.894 and p-value of 0.0379, suggesting that Mexico is at a higher risk than other countries. The year's coefficient is -1.062 with a p-value of 0.4180, showing that over time there isn't a significant trend in changing fatality rates from shark attacks.

This model also shows an improvement in fit for the data, reducing the null deviance from 53.699 to a residual deviance of 29.167. Said model was selected due to its ability to effectively handle binary outcomes suiting the data of whether a shark attack is fatal or non-fatal. By using the ROC Curve and Confusion Matrix we can conclude positive results in predicting the fatality of shark attack fatalities.

The model shows a high accuracy rate of 91.53%. The sensitivity rate also shows a high rate of 95.92% confirming the model's ability to predict fatal cases. However, the specificity has a rate of 70% showing moderate accuracy and a small need for improvement in identifying non-fatal entries. The positive predictive value of 94% shows reliability in predicting fatal attacks, while the negative predictive value is 77.78% and shows acceptable performance in non-fatal attacks. The Kappa of 0.6865 indicates the probability of the agreement between the model's prediction and the data. McNemar's test value of 1.0 highlights the absence of great bias in the model's predictions. Overall, this model is very effective in identifying fatal shark attacks.

Linear Regression Model

```
Call:
lm(formula = Num_Attacks ~ Year, data = attacks_per_year)

Residuals:
       1        2        3        4
 -0.7831 -13.6145  31.9277 -17.5301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6976.120  12205.989  -0.572    0.625
Year            3.458      6.037   0.573    0.625

Residual standard error: 27.5 on 2 degrees of freedom
Multiple R-squared:  0.1409,    Adjusted R-squared:  -0.2886
F-statistic: 0.328 on 1 and 2 DF,  p-value: 0.6246
```

Linear Regression Model Evaluation for Identifying Trends in Shark Attack Frequencies:

The linear regression model was used to identify the trends in the frequency of shark attacks. The residual numbers include -0.7831, -13.6145,31,9277, and -17.5301. This data shows a great variation meaning it is not suitable for the data included. The coefficient for the year is 3.458 with a standard error of 6.037, shows a t-value of 0.573. The p-value here is 0.625, which shows that the year cannot help with predicting the number of attacks significantly.

The model shows a low multiple-squared value of 0.1409 and an adjusted value of -0.2886. The model might not be a good fit for the data since it suggests that variables that aren't year might be better in identifying the trends. We selected the linear regression model to analyze trends over time. This model shows the frequency of how attacks have changed over the years. The residual standard error is 27.5 on 2 degrees of freedom which shows the average distance of the data points from the fitted line is 27.5 attacks.

The r-squared value of 0.1498 suggests that 14.09% of the variability of attacks can be explained by the year. These values indicate a poor overall fit of the model. The adjusted r-squared value is lower at -0.2886, indicating an even worse model fit. Additionally, the f-statistic is 0.328 with a p-value of 0.6246, showing that the model is not statistically significant. These results suggest that changes in the predictor "Year" may not accurately predict changes in the number of attacks, raising concerns about the reliability of the regression model.

Clustering Model

Clustering Model Evaluation for the Geographical Analysis of Shark Attack Patterns:

This clustering model categorizes clusters of shark attacks by location. The y-axis is represented by latitude and the x-axis is represented by longitude. The clusters are represented by three different colors: red, blue, and green. For cluster 1(red)it covers areas with mostly negative longitudes, this shows that many attacks were in the western hemisphere. Cluster 2(green) covers areas in both northern and southern regions. For cluster 3(blue), it is found at high positive longitudes showing attacks in the eastern hemisphere.

This clustering can help highlight geographical hotspots enhancing our understanding of where shark attacks occur most frequently. By using this clustering model, we can aim to explore the geographical distribution of shark attacks and create preventative measures. In all this model shows that shark attacks have a broad geographic range.

To conclude model evaluations, the logistic regression classification model was evaluated through the k-fold cross-validation process. The model's performance showed efficiency in distinguishing between fatal and non-fatal attacks thus making it the final model.

Limitations:

Although our analysis is comprehensive, it is not without its limitations. First, the data obtained from the Global Shark Attack File may be subject to biases and inaccuracies inherent in self-reporting and data collection methods. In addition, the removal of incomplete records during data cleaning may inadvertently remove valuable information and distort the analysis. Furthermore, inconsistent reporting and identification of shark species make it difficult to accurately assess their involvement in attacks. These limitations highlight the need for caution in interpreting results and the opportunity to improve the data collection and reporting process in future studies.

To address the noted limitations, it is recommended that standardized data collection protocols be put in place to ensure consistency and accuracy across reporting sources. Quality control measures should be implemented to verify data integrity and completeness and efforts should be made to minimize reporting bias and errors. In addition, exploring alternative data sources and incorporating advanced analytical techniques such as machine learning algorithms could provide additional insights into human-shark interactions. Prioritizing data quality and transparency will increase the reliability and robustness of future research in this area.

Policy Implications:

These findings are crucial for the development of new policies and better decision-making in shark preservation which will in turn contribute to public safety. Policymakers should consider implementing strict regulations and management strategies to minimize the risk of shark attacks in high-traffic coastal areas - such as strengthening coastal safety protocols, better and faster emergency responses, and establishing marine protected areas to protect critical shark habitats. Using scientific evidence in policy frameworks can help to achieve a sustainable coexistence between humans and sharks while protecting public safety and marine biodiversity.

Future Research Directions:

The results of our analysis will form the basis for future research to better understand human-shark interactions. This includes investigating the effects of environmental factors such as climate change and habitat degradation on shark behavior and distribution patterns. In addition, research is needed to assess the effectiveness of different deterrent techniques and management strategies to reduce shark attacks. By incorporating interdisciplinary approaches and collaborative research, complex research questions can be addressed, and knowledge generated that can be applied to conservation and public safety efforts.

Our analysis leads to implications for enhancing public safety. Emergency response teams can use this analysis to utilize frequent hotspots for quicker and more targeted responses. Swimmers and surfers also can gain insights into safer locations for water activities with lower shark activity rates. The general public can benefit from understanding low fatality rates and specific conditions that influence shark attacks. This will allow for the practice of safer behaviors. Environmental organizations can use species-specific data to create strategies that ensure human safety. Researchers and marine biologists can look at attack patterns and diverse data to further scientific studies. All our findings can educate and inform a broader audience ensuring coverage of shark-related incidents.

Conclusion:

Our Analysis has highlighted several key aspects of global shark attacks. The analysis confirms that unprovoked attacks constitute most of the shark incidents in the dataset at 62.5%; illustrating how unpredictable shark attacks can be. Geographically, The United States, Australia, and South Africa are leading in the frequency of shark attacks. This data showed significant increases in attacks from 2019 to 2023. However, despite the high frequency of attacks, the fatality rate remains low with an overall proportion of fatal attacks at 0.0718%.

The analysis also shows that doing the activity of bathing near shark-populated waters increases the likelihood of a shark attack. Species diversity attacks show that no specific shark type dominates the attack statistics however the White Shark is the most identified type. This data shows significant issues in shark species identification and missing data which can impact the accuracy of the attack findings. Notably, attacks in areas like Mexico show higher fatality rates highlighting that differences in regions can influence shark attack outcomes.

Despite our data analysis, the study shows many limitations due to biases in the Global Shark Attack File. There was a hidden data column in the data set that was irrelevant and bloated up the data set and blundered some of the analysis. There was a lot of missing data, but this can be easily fixed by using code to remove all the missing data rows. In all the data was very inconsistent.

Recommendations:

1. Improve Public Education: Inform the public about shark behavior, risk areas, and safety precautions.
2. Improve Responses: Strengthen emergency services near shark-infested waters to improve responses to shark attacks.
3. Further Research: Support and fund research into shark behavior to understand the factors leading to shark attacks.
4. Regulate High-Risk Activities: Limit certain water activities in known frequent hotspots and limit certain water activities.

References

1. **Global Shark Attack File. (2005). Incident log. Retrieved May 8, 2024, from** https://www.sharkattackfile. net/incidentlog.htm

2. **Burgess, G. H., Bruce, B. D., Cailliet, G. M., & Goldman, K. J. (2010). Shark ecology and public safety. *Journal of Coastal Research*, 26(5), 957-965.** https://doi.org/10.2112/JCOASTRES-D-10-00081.1

3. **Clua EEG, Linnell JDC. Individual shark profiling: An innovative and environmentally responsible approach for selectively managing human fatalities. *Conservation Letters*. 2019; 12: e12612.** https://doi.org/10.1111/conl.12612