Shark Attacks: An Exploratory Data Analysis

Brenna Wallace

Omer Zaher

Jennifer Wilkerson

SMU – Boot Camp Data Science

Introduction

Our group analyzed data on shark attacks to see if there were any correlation in these attacks across the world and then more specifically in the US. About 400 million Americans visit the beach every year. Beaches offer us a range of activities like boating, fishing, swimming, and surfing. Most people enjoy these coastal waters and seldom think of encountering a shark let alone being attacked by one. Although these instances are what most would consider rare; they still happen. Most attacks are classified as provoked and unprovoked attacks. However, we will be taking a deeper look at the data on these attacks to get a better understanding of: if we should be more cautious of the activities we like to enjoy on our coastal shores and do they affect the likelihood of a shark attack. We came up with 4 primary questions:

1. Where do reported shark attacks happen most in the world?
2. What reported activities are being done in the USA during a Shark attack?
3. How severe was the USA shark attack?
4. What reported month of the year do shark attacks occur most often in the USA?

* Which countries have the most shark attacks
* What time of the year do most shark attacks happen
* Which states have the most attacks?
* Where do shark attacks in the US happen?

Shark Attacks: An Exploratory Data Analysis

# Data Cleaning

The first thing we did was import our dependencies followed by reading in the csv file. One of the main obstacles we faced is that there is very limited data regarding shark attacks, for example some countries may not necessarily report shark attacks due to restrictive governments. We needed to start by removing unnecessary columns and ended up with the following for our new data frame columns, “Case #, Date, Year, Type, Country, Area, Location, Activity, Sex, Age, Injury, Fatal(y/n) [figure 1.1].

Then we renamed our columns to be lower case and removed any spaces and dropped any duplicate rows [Figure 1.2 & Figure 1.3]. We had to decide how we wanted to explore the information so we started by narrowing down and cleaning our data so that we were only looking at the reported shark attacks across the world ranging from 2000-2018 because that is where we had the most data to analyze [Figure 1.4].

Most of our data stems from the USA so we created a new data frame ranging from 2000-2018 just for the USA alone [Figure 1.5].

The next thing done was to organize the attacks by activity type and injury type. Our activity categories were; “on\_water, in\_water, fishing, science, other”. We created an excel file to read in that organized our parent activities accordingly [Figure 1.6]. Our injury categories were; “Laceration, Bite, No Injury, Puncture, Severe Injury, and Other”. We created a new column for parent injury by creating a mask/filter for each injury type.

The next thing we wanted to look at was injury type by month so we could see if there was a change over time this will also address the time of year that has the most reported attacks. There was some more data wrangling that needed to be done so we used the string replace function in our date column. This allowed us to remove the common text or symbols that would put the entries in the date column back to a convertible date. Then, we used string strip to remove the leading and trailing white spaces to make sure our date column was clean [Figure 1.7].

## Reported Attack Locations

Since our data was limited, we decided to make a bar chart of the top 7 countries with the most reported shark attacks [Figure 2.1], we also added a map of the world attacks [Figure 2.2]. After looking at our data at we found that more than 50% of all reported attacks happened in the USA. To investigate the USA further, we created a horizontal bar chart for the top 7 United States [Figure 2.3] and another corresponding map [Figure 2.4]. From this we found that most reported attacks happen in Florida, Hawaii, and California with 53%, 13%, and 11%, respectively. This information aligned with what we expected.

**Reported Activities During Attack**

We categorized the activities and created a bar chart that shows on\_water activities had the most instances of a shark attack [Figure 3.1].

We generated three different maps using the data from the API, but the heat map with the marker layer added to it was the most useful [Figure 3.2]. Most of the attacks occur in Florida which was known since it was shown in the bar graph charting the attack amount in the top seven states with the most attacks. The heatmap reinforced what was charted in the bar graph by showing where exactly in Florida the attacks occurred [Figure 3.3]. New Smyrna Beach, Florida had the largest hotspot of attacks. Other popular beach areas in Florida with clusters of attacks include Melbourne, Vero Beach, Port St. Lucie, and West Palm Beach. All these beaches are on the Eastern side of Florida.

**Severity of Attack Reported**

Since we now had categories of injury for all USA shark attacks, we created a bar chart indicating the highest frequency of injury type (severity) which is lacerations in the USA to address the severity of the injury [Figure 4.1]. For fatality, less than 2% of attacks were reported as fatal.

**Reported by Month**

For the attacks reported by month, we were able to make a line graph showing the instances by injury type over time (every month) and concluded most injuries happen from July and September with lacerations leading the way. Using a violin chart as well, we were able to visualize the reported attacks by month per year.

**Gender and Ages Reported**

For Gender reported by injury type, we created a new data frame that did a group by on gender and parent injury and the data showed that males were leading females in all parent injuries and reported attacks as well at around 3 times more frequently. We realized there was some correlation in ages and activity, so we created a new column labeled age group and created a new data frame that was filtered by age group and activity. We were able to graph it with subsets of each activity and the data showed that most people between the ages of 0-19 are at most risk of being attacked regardless of what activity they were doing with over 300 reports. At over 59 years old we saw a large drop off in the number of attacks, which aligned with our expectations.

References

Last Name, F. M. (Year). Article Title. *Jounal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Footnotes

1[Add footnotes, if any, on their own page following references. For APA formatting requirements, it’s easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. (Note: If you delete this sample footnote, don’t forget to delete its in-text reference as well. That’s at the end of the sample Heading 2 paragraph on the first page of body content in this template.)]

Figures

Figure 1.1

Text

Description automatically generated with low confidence

Figure 1.2

Table

Description automatically generated

Figure 1.3

Graphical user interface, text, application, email

Description automatically generated

Figure 1.4

Graphical user interface, text, application, email

Description automatically generated

Figure 1.5

Table

Description automatically generated with medium confidence

Figure 1.6

Graphical user interface, text, application

Description automatically generated

Figure 1.7

Text

Description automatically generated

Figure 2.1

Chart, bar chart

Description automatically generated

Figure 2.2

Map

Description automatically generated

Figure 2.3

Chart

Description automatically generated

Figure 2.4

Map

Description automatically generated

Figure 3.1

Chart, sunburst chart

Description automatically generated

Figure 3.2

Map

Description automatically generated

Figure 3.3

Map

Description automatically generated

Figure 3.4

Chart, bar chart

Description automatically generated

Figure 3.5

Chart

Description automatically generated

Figure 4.1

Chart, bar chart

Description automatically generated

Figure 4.2

Chart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generated

Figure 5.1