Global Shark Attack Activity

Daniel Fleck, Austin Mann, Thane Miller

Introduction

The Global Shark Attack data set covers reported incidents involving human altercations with sharks. The data set goes back as far as to include historical records from the 1500s. However the bulk of the data comes from much more recent attacks. This can be attributed to the increased global communication and ease of reporting attacks as technology has developed. The data set includes an abundance of details about each attack that can be used to analyze causes of shark attacks. These attack details are categorized into variables that include important information like; date, activity, country, injury, and species. The Shark Research Institute maintains this file and continues to update it with new attacks. The mission of the Global Shark Attack file as stated on their website is to "...provide current and historical data on shark/human interactions for those who seek accurate and meaningful information and verifiable references." By providing this open source data they hope to create a better understanding of shark and human interactions and how to prevent them, and through our analysis we hope to do just that. As a starting point for our analysis we wanted to see how commonly held beliefs about shark attacks compared to the actual data. When looking for a source of these beliefs we found the Florida Museum, which includes information about shark attacks and its section for advice to avoid attacks addressed many of the beliefs about shark attacks. Common beliefs about shark attacks included avoiding the water at dawn and dusk, staying close to shore, avoid water that is murky or appears to have a lot of runoff, and to avoid using jerky movement while swimming. Since weather came up and that was not present in the current data the National Oceanic and Atmospheric Administration, NOAA, seemed like a good source for finding this information.

Questions

The commonly held beliefs about shark attacks lead us to forming questions about shark attacks that we wanted to use the data to address. The first question was about location. What countries seemed to attract the highest number of shark attacks? After location, activity was the next question we wanted to look into. Of all the different ocean and beach activities people participate in which seemed to attract the greatest number of shark attacks? Weather has a big impact on ocean conditions which should have an impact on sharks. Do shark attacks correlate with any particular daily weather conditions? In addition to weather is there any particular time of day where sharks seem to be the most active? After the shark craze created by jaws the idea of sharks like the great white being manhunters became popular. Are there any species of shark that are the most likely to be responsible for attacks? As we worked with the data and attempted to find answers to our questions we found a few more areas that we were curious about. How does age affect the outcome of attacks? After studying sharks by species we also wanted to find if any certain species were responsible for any particular type of injury.

The body of the report will go into further details about the detail sets that were used and the information that they contain. It will also address the python code that was used to get information from the data sets and the logic behind the methods that were used. The bulk of the body will be devoted to presenting the graphics and analysis of the data as it was used to answer our main questions.

Data Set

The Global Shark Attack File provided the majority of our information. The data set contains 6300 instances of shark attacks from around the world. The earliest incidents go back to the 1500s and the set is kept current being updated each month. The attacks that have happened more recently are more complete in their information. The older attacks were taken from historical records so they do not always include as much information. Each variable in the data set provided relevant information about the individual attacks. The first being the case number which was used as an index. The next column was dedicated to the date of each attack. Some were down to the day but many of the older attacks simply had the year or a time frame that it occured. After date there was a separate column for year, which was filled for almost all of the entries whereas date was missing about 300.

The next variable was type, this was split into provoked and unprovoked attacks but sometimes boating was included as a category. The next three variables were devoted to locations getting more specific. The largest category of location was the country that the attack occurred in. The next was area that was filled in with the state or province depending on the country. The final location variable was location of the attack which narrowed it down to the city or beach that the attack occurred in. The country variable was the most complete only missing about 50 entries but as it narrowed to area and location a couple hundred rows were incomplete. Activity was also an important variable for our analysis, While it has a lot of entries there was no standard format and the data needed a lot of cleaning.

The next few variables had to do with the person that was attacked. This included name, sex, age, injury, and fatal (Y/N). The name was not useful but age and sex provided good information for seeing the demographics of attack victims. Injury and fatality also allowed us to study how severe attacks from each species were as well as comparing age to fatality. Age was only present in a little over half of the entries, but all the other variables were filled out for the majority of entries. Time was included when present, but only in a little under half of the cases. When it was included it was not very organized with some entries being a numerical time and others being a categorical phrase for the time of day. The last useful variable was species. It was present in 3,462 of the entries but it allowed use to answer a lot of the questions we posed earlier.

The last variables were investigator or source, pdf, href formula, href, case number.1, case number.2, and original order. These variables provided outside information about the shark attacks but were not in a format that could be used for our analysis. Overall the data set contained a lot of very useful information, there were some gaps but every variable had enough information that we could use it for our analysis. Despite this abundance of information a lot of the variables we wanted to use had no uniform format. To get the data set to a useful point we had to clean a lot of it.

The weather data set was pulled in as an addition to the shark attack file to see the effects of weather on shark attacks. To get weather data from NOAA we had to request it from their website. Daily weather recordings from around the world create enormous datasets so we narrowed the scope down to places in the top three countries for attacks. We requested data records from present day going back to January 1st 2000 as a way to lower the data amount and also to see if there were any current trends we could find. From the United States we chose North Carolina and Florida to study as well as bringing in records from Australia and South Africa. After merging the weather data to the shark attack data on date and country we checked to see which variables were the most complete to use to study weathers effects. We decided to study precipitation, observed temperature, maximum temperature, and minimum temperature. These variables had the most categories and also weather effects that would have a significant impact on the ocean.

Code/Methods

Weather

The weather data sets had to be created through NOAA's search tools on https://www.ncdc.noaa.gov/cdo-web/. The initial weather data set mentioned in our proposal was too large to download and required you to work through Kaggle to analyze it. After getting the data set the next step was finding the best way to merge with the shark attack data set. Location and date were the most important factors that needed to be matched. The original weather data had date in dd/mm/yyyy format while the shark data was dd/mon/yyyy. To get these different formats to match pandas to_datetime function was used. This created uniformity across the date columns for both data sets. The next challenge was to merge across location. The weather date set had no location column only name which had the name of the weather station where the data was collected. Luckily the name included the state and country abbreviations. Using regular expressions we were able to pull out the country abbreviations and put them into a new column called country. After that the string replace function to replace the abbreviations with their full names. After cleaning up the two sets we used a pandas merge function to outer join the data sets on country and date. After checking to make sure the merge work it was found that all of the

attacks in the US were brought in regardless of state. We then used a boolean statement to only include rows where the area variable was North Carolina or FLorida.

Location

The location information provided by the Shark Attack was detailed in that it included "Location", "Area" and "Country" for over 90% of the recorded attacks. However, in order to be able to map the locations of these attacks latitude and longitude are needed. The "Location" variable provides county, city, beach and other micro level location information. The "Area" variable provides state, province, and other more macro level location information. While "Country" includes the Country information for all the attacks. In order to get the most accurate location information we combined all three columns into one "Full Location" column in the following order: "Location, Area, Country." After creating this new variable we used Geocoders in the Geopy package to get these Latitude and Longitude points. Because of constrictions with the Geocoders package API only approx 2500 requests were allowed per day and a limit of the amount per minute before the there was a timeout error. Therefore we ran approx 2000 locations per day, and set a sleep on the the for loop to prevent a timeout error from occurring.

The Geocoders package allowed us to get coordinates for only 3303 of the 6302 attacks in the dataset. Although this is only a little over half it allowed us to get a good visual representation of the locations for all attacks in the dataset. In order to map the locations we used the folium package in python. The folium package provides an interactive mapping visualization. It allows you to zoom in on certain areas and locations like google maps. It also provides an option to include a pop-up with your plotted points, we chose to include the point coordinates here. Along with the folium package we used the external application ArcGIS to create more interactive and detailed visualizations with the data. ArcGIS made it simple to import our data choose what variables to focus on and display in each map. It provided great interaction similar to before with the capability to zoom, but also provided the option to search by city and location. The pop-ups included with each mapped point gave us an area to show further detail on each attack. We included details on name, sex, fatality, activity, species, size, time, etc. and even provided a reference link for those attacks that included a pdf write up of the attack. We used ArcGIS to map our data and visualize by activity, size, and fatality. The weather data we were able to pull in by location also gave us an opportunity to map attacks and visualize based on daily temperature minimums and maximums on the day of each attack.

Sex

Cleaning the sex column for better analysis was simple as the data came with very few unique sex classifications. Extracting "M" and "F" using string extract was the extent of what had to be done for the sex variable.

Fatal

The Fatal Variable included "Y" if the attack was fatal, and "N" if it was not. The Fatal column included different variations of upper and lowercase "Y" and "N" but also included "Unknown." We did not make any assumptions that "Unknown" should be classified as Yes or No and therefore just extracted "Y" or "N" from the strings. All unknown rows were now missing, we have data for the Fatal variable for 90% of attacks.

Activity

The Activity variable was not very clean in its initial format from the Shark Attack File. While 5,758 of the entries were completed there was no uniform style of entry. Many were one word activities like surfing, fishing or paddling. However, others were complete sentences like, 'swimming on sandbar adjacent to channel'. Since this would fall into a swimming category but not be recognized by panda functions we needed to clean up the activity entries. Using regular expressions to search for keywords in all the entries we were able to condense all the variables into a smaller number of categories and put them into a new column, activity_type. Initially surfing had 971 entries into the activity column. However other entries like 'surfing, but treading water','surfing, but standing beside the board', and 'surfng' could all fit into the activity type of surfing. After pulling all the variables that included 'surf' the count went up to 1,194. By selecting other key words like swim, boat, and diving we were able to fit all of the activities into 22 activity types that would allow for much more meaningful analysis.

Time

The time variable mainly came in two forms, a 24 hour time or a part of the day. For example, one observation might have a time entry of "17h00" and another might have "afternoon." We chose to format time into the part of day. We don't think that the loss in specific time information for a shark attack is too serious since the light conditions depend on location latitude and time of year. Also, we would have to skip or change the times listed as a part of day.

We used regular expressions to extract the digits from the 24 hour times, and then used pandas' cut function to bin the numbers from 0 to 2400 into intervals we picked for the different parts of the day. The times 0-500 became "Night", >500-800 are "Early Morning", >800-1100 are "Morning, >1100-1400 are "Midday", >1400-1700 are "Afternoon", >1700-2100 are "Evening", and >2100-2400 are "night." One potential issue is that some entries had a range for the time, for example "14h00 to 1500." The regular expression matching only pulled the first time, "14h00." As we were going to put the times into arbitrary categories anyway, this didn't seem too problematic. Another issue is that the time ranges for parts of the day are not standardized, so our categories might not match up perfectly with the intent of the original database.

We then extracted the parts of the day that matched our time category names and combined the two columns. Around 100 of the 3000 entries in time could not be processed into our time categories.

Size

A quick glance at the shark species column showed that it could have two pieces of information, the species and the size of the shark. Some observations had only one of those, some had nothing, and some had more detail about the incident. A complication was that the format was inconsistent, even for entries with shark species, size, and nothing else. Size could be in meters or feet and could be listed before or after the species. We used regular expressions to extract digits that were followed by a single quote for feet or an "m" for meters. The sizes in meters were converted to feet, and then combined with the feet column. We also put the sizes into categories. We did notice one problematic size extraction that interpreted a bite size in millimeters as the shark size, so that was removed from the size category column.

Species

To find the shark species, we extracted three or more letter words followed by "shark" and, optionally, a three or more letter word before that. Some shark species like bronze whaler shark have three words in their name where others, like bull shark, only have two. We only extracted three letter and up strings to not capture the "m" from the size. We did try another method to extract shark species that would involve searching for each species individually but was less likely to miss or cut off names. Testing this for a few of the species showed an addition in less than ten to the count for the species. One problem with grouping the shark species names is that a species may be known by different names based on region. We caught one example of this where bull sharks are known as Zambesi sharks in Africa, but there are likely more examples.

Injury

The injury column has, overall, the most complex phrasing of the dataset. Many entries have phrases detailing type and location of one or multiple injuries. There is no consistent or formulaic style to the column that would make extracting information using regular expressions easy. We chose to perform very simple regular expression matches to keywords for injury type and place. This process loses a lot of the information from the injury descriptions but also creates categorical variables that can be compared with other variables.

Age

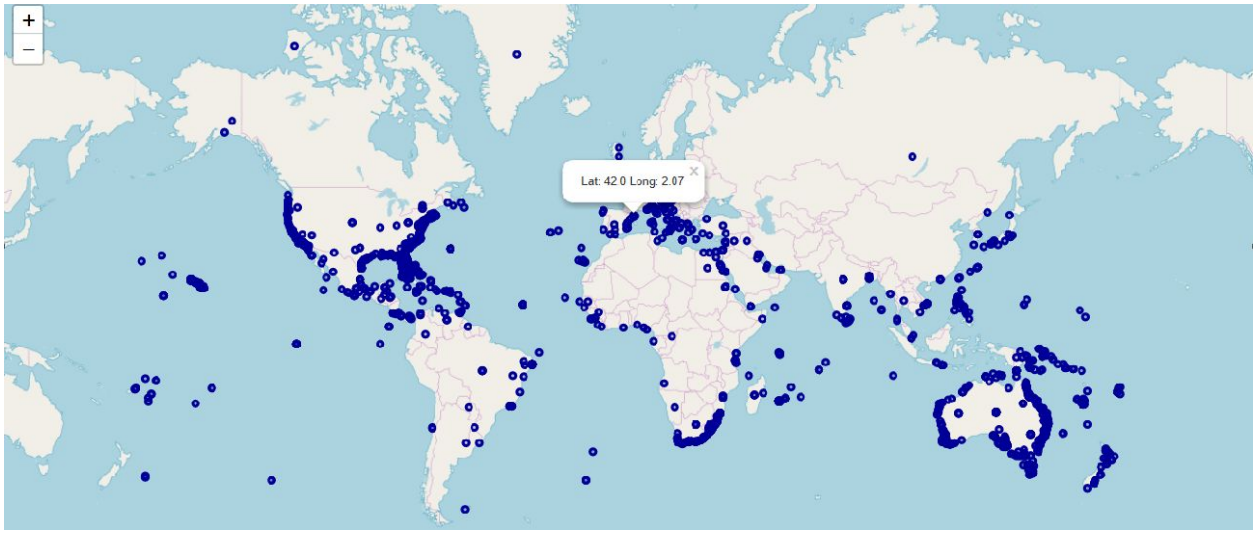
The Age variable had data in both numeric and string format and the ages were classified a number of ways making it quite difficult to clean up. Some were submitted as single numeric ages, others included multiple ages separated by "&" "to" and "or", and others were even written as age classifications with "mid-20s", "teen", and "adult" being examples. Because of the

varying differences between a lot of the data in the Age variable this cleanup took multiple steps in order to get maximum data for our analysis. We replaced all rows that included character information that were not specific enough to be classified into an age as missing. We made the assumption that teen could be classified as 18, however did not want to make an assumption on the classifications such as "young" "adult" and "elderly" because the age ranges for these are far too broad to assume. For any records that included two numbers separated by "&" "or" and "to" we created a function to average out the two numbers in the record. We did not want to choose one of the numbers as we would have to decide on choosing the smallest or largest number and this could skew my data in either direction. So instead we believed averaging the two numbers would be best for the integrity of the data. We also created another function in order to classify all records that included "mid" into a numeric value. For any value that included "mid" we added 5 to the age that followed the word "mid" in the record. For example, if the record was "mid-30s" we extracted 30 and added 5 making the "mid-30's" record 35.

Graphs/Analysis

Location

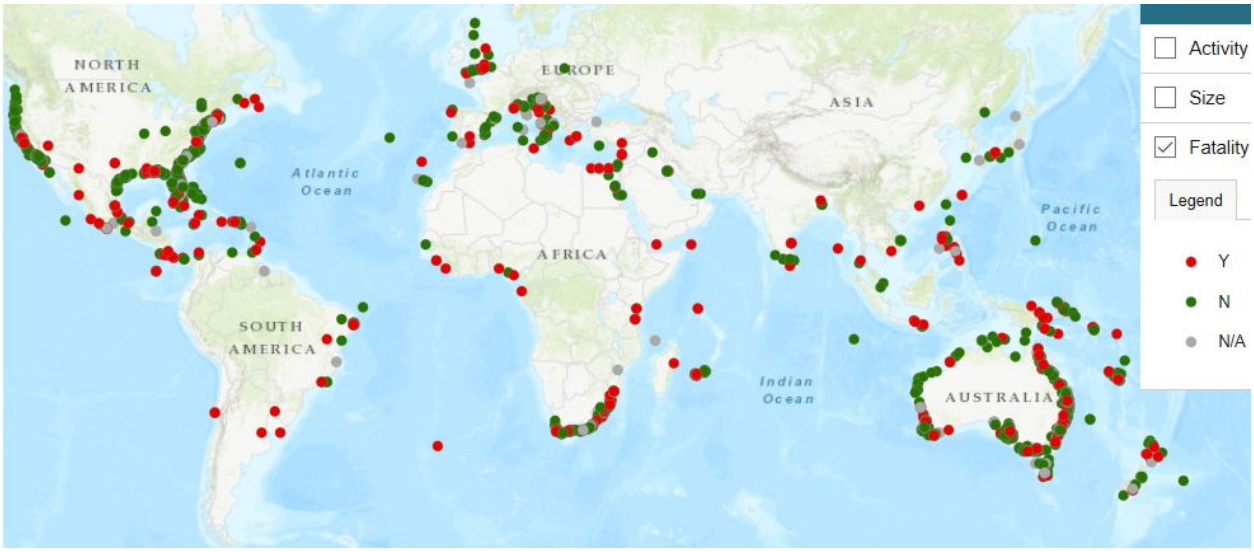
Plotting the shark attacks on an interactive map can give us a good visual representation of the shark attack data and hot spots of shark attack activity.



The map above shows a large volume of attacks occur up and down the East and West coast of the United States, along with the entire coast of Florida. The tip of South Africa going up the east coast of the country includes another large cluster of attacks. Australia is similar to that of the United States where cluster of attacks occur up and down the East and West coast. The tip of Australia off the coast of Melbourne and Adelaide are also highly concentrated locations with shark attack activity. As the mapping suggests, the top 3 countries that included the most shark

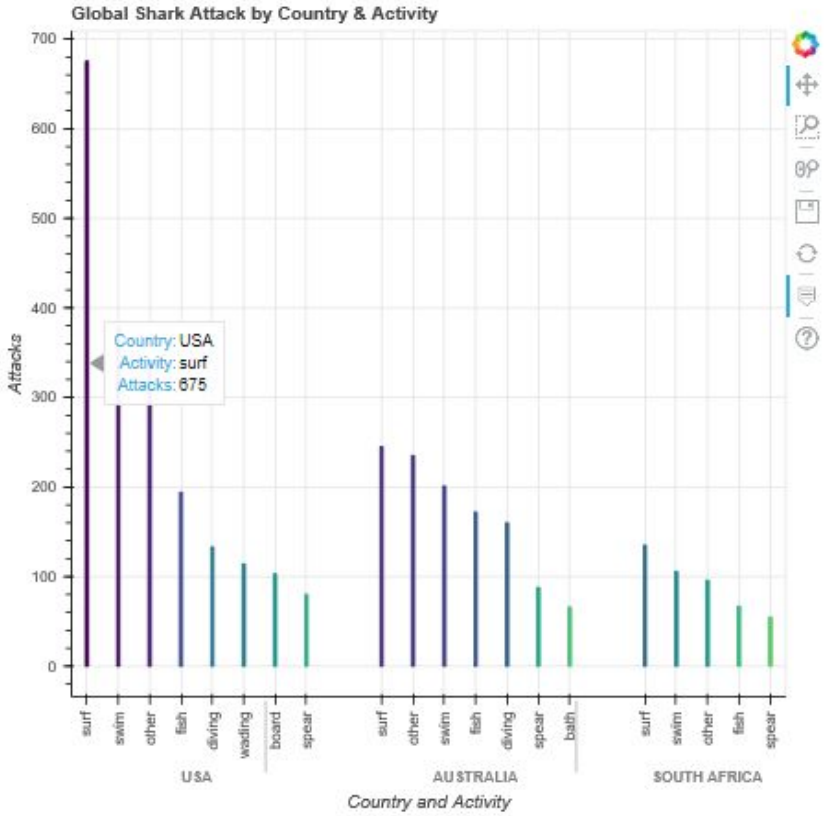
attacks were in order the United States, Australia, and South Africa. 2,229 attacks in the United States, 1,338 in Australia, and 578 in South Africa.

Mapping attacks by other variables including fatality, activity and size also provides a good summary of the potential impact location can have on a number of shark attack variables. The map below plots the shark attacks by fatality, red being fatal and green being not. Focusing in on a few regions we can see the number of fatal attacks are greater in some regions than others. The west coast of the United States has a large number of attacks with very few being fatal. Stretching into Northern California and Oregon we see no attacks reported in this region have been fatal, and there is a large cluster of non fatal attacks off the coast of San Diego as well. The East Coast of the US is similar, with large clusterings of the attacks being non fatal as you can see in the image below. The biggest concentration of fatal attacks appears to be off the coast of South Africa, most notably around the eastern coast of the country. Because we do not have demographic data on the number of people in the water it is hard to definitively say whether shark attacks in South Africa are more likely to end in fatality compared to other areas across the world. A big reason for this concentration of fatal attacks could also be due to the increased likelihood of reporting a fatal attack versus a non fatal attack. Shark attack data in countries outside the United States may not be as reported especially in regions across the South African coast that may not be in close proximity to a major city.

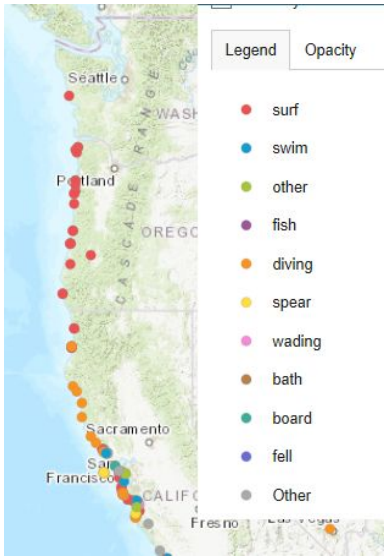


Based on a number of factors the activity reported during shark attacks differs by country and region. The leading activity reported during a shark attack in the dataset was surfing, closely followed by other, a list of activities that could not be categorized into one activity, and swimming. 1194 attacks occurred while surfing, 1157 during other activities, and 1095 while swimming. The graph below shows activity by the top 3 countries reporting, this follows a similar trend to the overall numbers. Surfing is the leading activity in all three countries followed by some order of swimming, other, and fishing. In the United States the number of

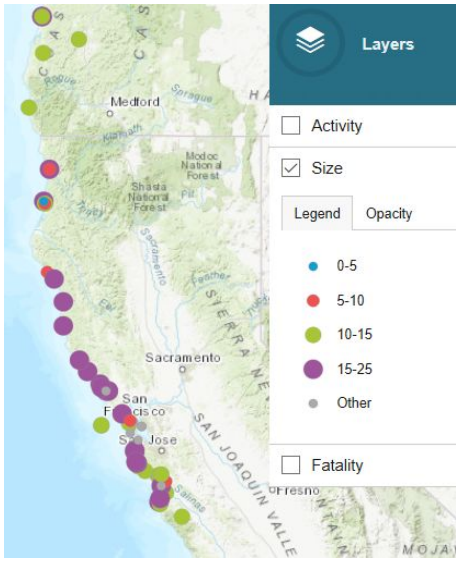
surfing attacks is far greater than any other activity as the surfing culture continues to get more popular. A 2016 article in Newsweek stated the number of surfers in the United States had increased from an estimated 1.8 million in 2002 to 2.5 million in 2016 according to a study from the Surf Industry Manufacturers Association. The number of surfing attacks in Australia and South Africa are not as significantly higher than the other activities like that of the United States. Mapping shark attacks by activity can show us why this may be so.

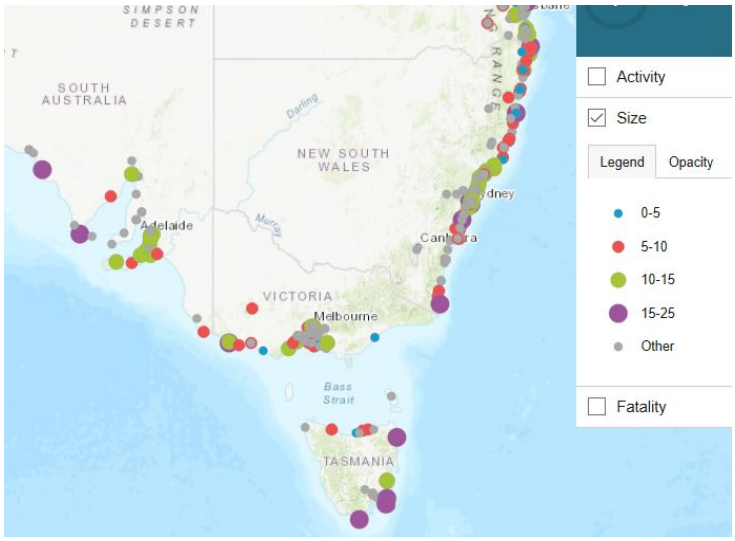


Below we focus on a few regions where the map shows a higher proportion of one activity. The full map of shark attacks can be seen in Figure 4 in the Appendix, however it is hard see anything notable without zooming in closer at a regional level. First, the Northern West Coast of the United States is a region where we see the activity during attack is only surfing or diving. Most notably off the coast of Washington and Oregon we don't see any attacks that occured while doing an activity other than surfing. This may be because the cold water temperatures in this area keep swimmers and more casual beach goers out of the water. South Africa has a high concentration of swimming and fishing attacks in the Cape Town area. With Cape Town being a major port city and fishing being a big part of its economy, the number of fishing attacks makes sense. The map of Australia by attack did not show any notable cluster of activity type in a particular region. It showed an even concentration of all attack types along the entire coast of the Country.

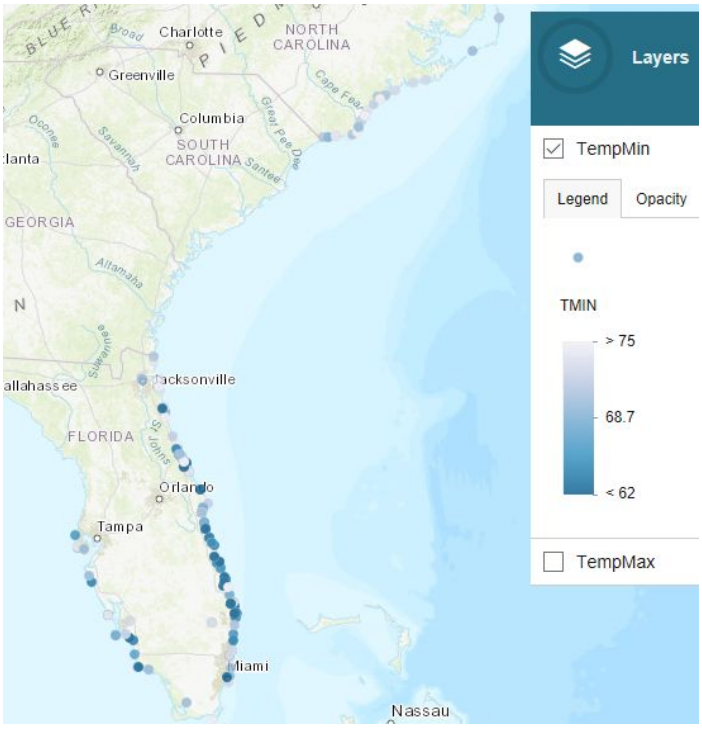


Lastly, mapping shark attacks by shark size may give us a good idea of size characteristic of the shark population by region. The coast of California most notably in the San Francisco region has a large proportion of shark attacks by sharks sized 15-25 feet, the largest category of shark size, and the second largest size in the 10-15 foot range. South Africa had a few attacks in the 15-25 foot range, while up the east coast of the country near the City of Durban there was a high concentration of attacks by 5-10 foot sharks. Next to California, the SE tip of Australia seemed to have the largest concentration of 10-15 and 15-25 foot shark attacks. Sydney and Adelaide have a large concentration of 10-15 foot shark attacks with the southern tip of Tasmania having a cluster of 15-25 foot shark attacks.



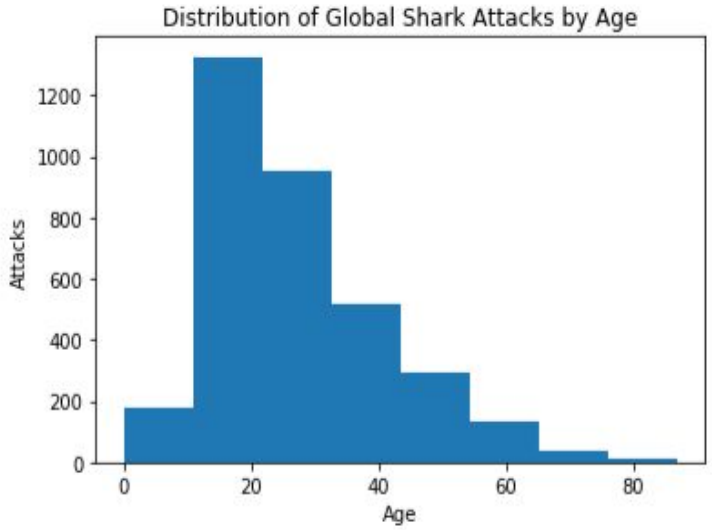


Being able to map weather variables as a secondary source with our shark attack coordinates helped show a relationship between shark attacks and observed water temperatures by region. Focusing on Florida and North Carolina we see a difference in the daily temperature minimums recorded on the day of the attacks. The daily temperature minimums along the coast of Florida were all on the lower end, with a majority of the coordinates being a dark blue point which signifies a daily minimum temperature of less than 62 degrees Fahrenheit. Meanwhile the points along the coast of North Carolina were all very light shades of blue, indicating a higher minimum daily temperature in the range of 75 degrees Fahrenheit and above. A major reason for this may be because of the high outside temperatures Florida experiences year round. Florida experiences far milder winters than NC, therefore beach activities like surfing and swimming can continue later in the winter and earlier into the spring than in North Carolina.



Age

The histogram below shows the distribution of ages for all shark attack victims in our dataset. This graph shows the majority of attacks occur with victims in the 18-30 year old range, teens to young adults. As stated earlier, the top activities during the time of attack were surfing, swimming, and fishing. These are all cardiovascular intensive activities, that would be more common in the teen to adult range where people are more active and in shape.



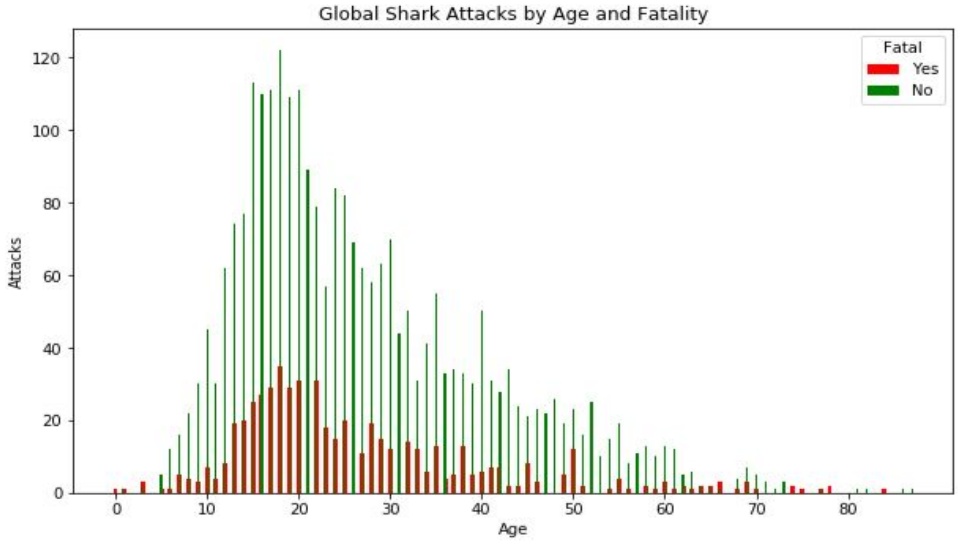
Age distribution by Gender

Studying the gender ratio of the shark attacks was not pivotal to answering any of our initial questions but we still felt it would be interesting. Surprisingly men are attacking in disproportionately high numbers to women. One possible reason for this is surfing being such a large portion of attack activity and a majority of surfers are male. Also because the data set goes so far back in history a majority of the sailing and fishing attacks would have male victims since they made up the bulk of sailors at the time. The graph included in the appendix, Figure 6, shows the distribution across age and gender. The distribution across age are relatively similar except for the counts for men being much higher. Simply breaking down the attacks into male and female victims the number of males attacked is 5,096 and only 637 of the victims were female.

Age distribution by Fatality:

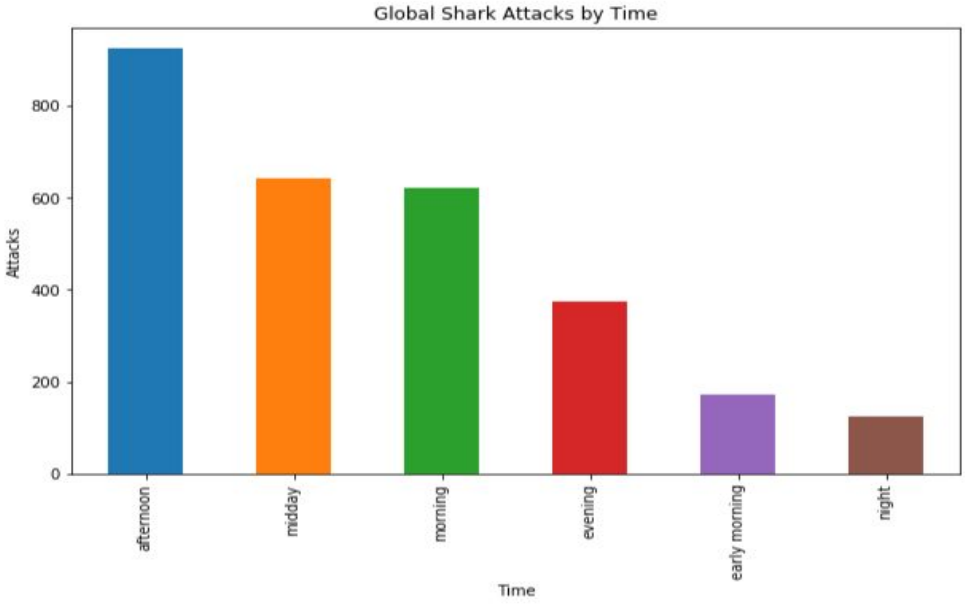
Graphing the age of attacks for fatal vs non-fatal victims we hoped to see if there was a difference in distribution between the two. In the graph below we see the distribution of non-fatal attacks in green, and fatal attacks in red. The distribution across ages is very similar in the middle age ranges, but the proportion of fatal to non-fatal attacks appears to increase as age goes up. Although the number of attacks for victims aged 60 and up are not a high number the number of fatal attacks in this range is close to the same amount or even higher than the number

of non-fatal attacks. With such a small sample size it is hard to say with certainty that older victims are more likely to have an attack end in fatality than younger victims, however the data does to appear to suggest that.



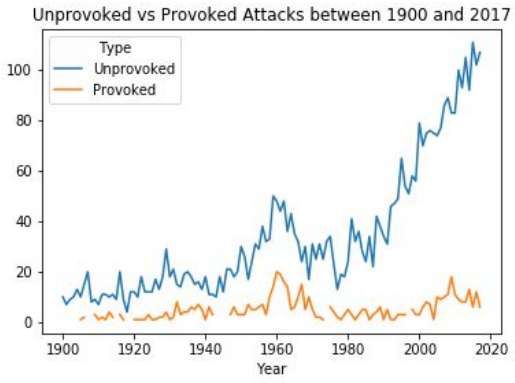
Time of Day

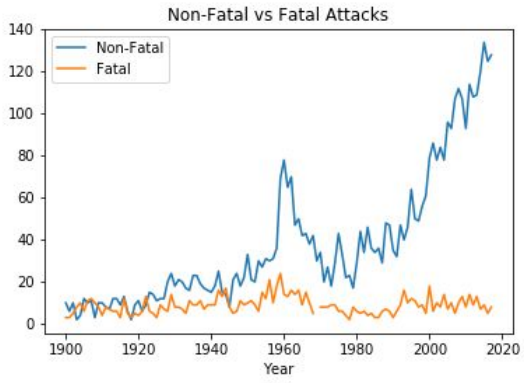
We find that shark attacks occur most often in the afternoon, followed by midday and morning. This finding conflicts with advice to avoid shark attacks by staying out of the water around dusk and dawn. It is likely that afternoon, midday, and morning have the highest counts of shark attacks because they are when people are more likely to be in the water, instead of representing shark behaviour.



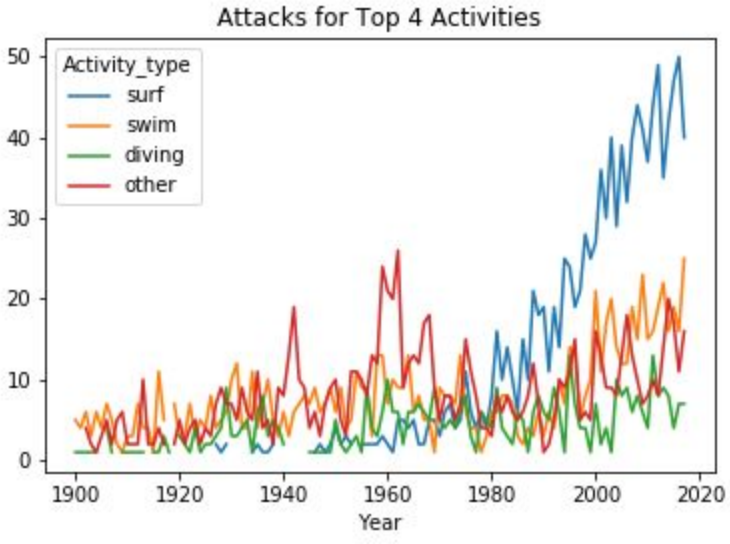
Shark Attack Activity over Time

Looking at the graphs of shark attacks between 1900 and 2017, there are two features that immediately jump out. There is a spike in the number of attacks around 1960, and there is an increase in attacks since around 1990. It is interesting that these increases are for non-fatal and unprovoked attacks, not for fatal or provoked attacks. The Global Shark Attack File defines a provoked attack as one where the shark has been speared, hooked, captured, or in general where the human purposefully brings themselves in contact with the shark. It's possible that the lack of growth in fatal attacks is due to better medical care for attack victims. It could also be due to an increase in certain kinds of shark attacks.



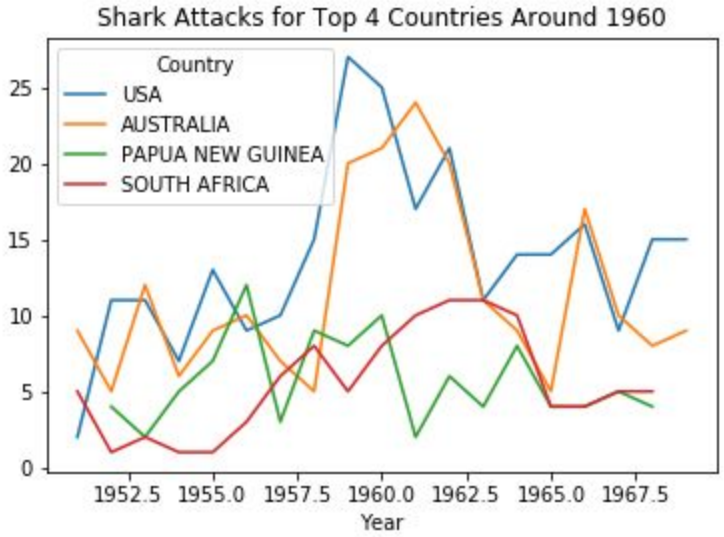


The increase in attacks since 1990 is matched quite well by the increase in attacks on surfers. Increases in the number of surfers and other beach-goers likely causes the increase in shark attacks.

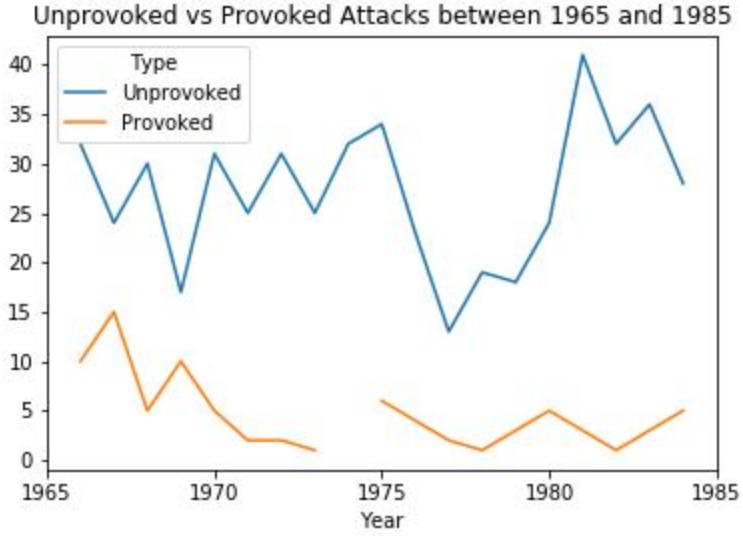


One possible explanation in the spike in attacks around 1960 is that South Africa suffered a string of fatal shark attacks in late 1957 to early 1958, causing people around the world to be more wary of sharks. It could also be that the increase in beach tourism that led to the increased

attacks in South Africa was also mirrored in places like the US and Australia. It's possible that the high number of attacks led to anti-shark measures, which then reduced the number of attacks.



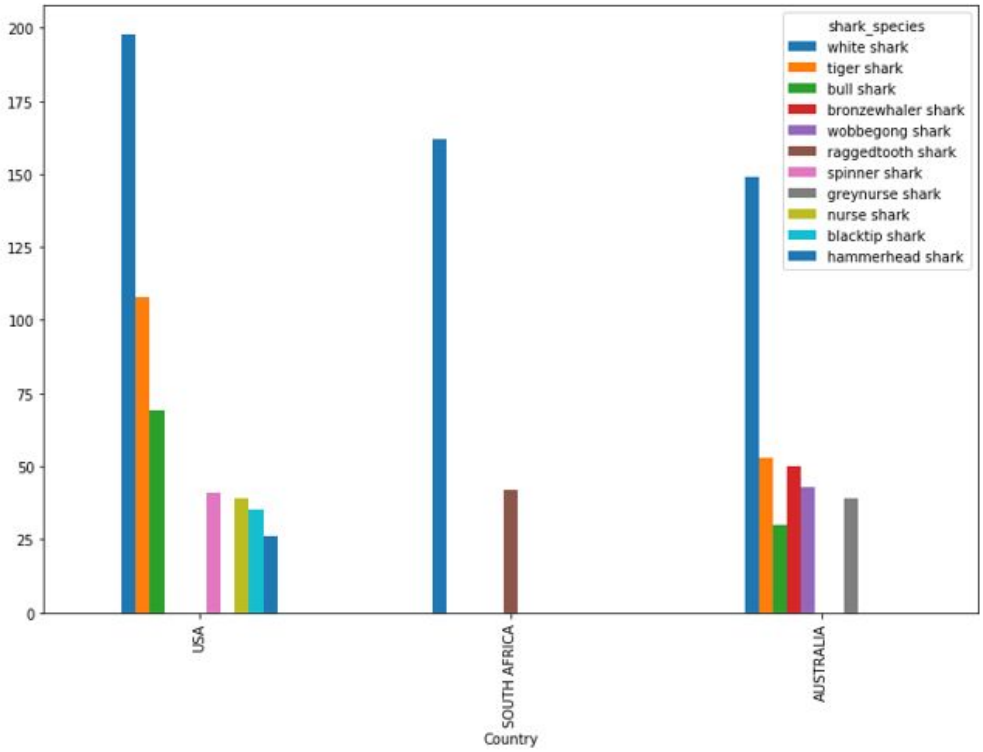
Although not as obvious on the graphs as the large increases, a drop in attacks just before 1980 also caught our attention. The drop appears to occur after 1975, which corresponds with the release of the movie *Jaws*. Given the popularity and impact of *Jaws*, we find it plausible that the movie had an impact on shark attacks, whether due to anti-shark measures or reduced beach-going tourism.



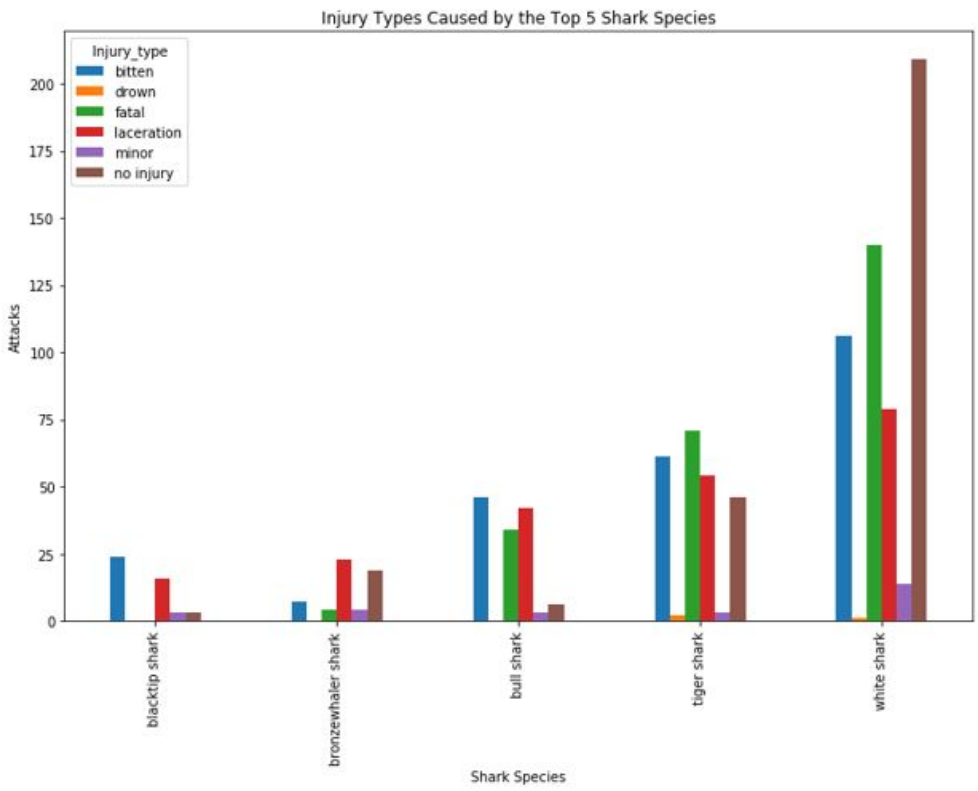
Species

The most significant shark species were white, tiger, and bull sharks because the were responsible for the largest numbers of attacks. White sharks were responsible for 621 attacks, followed by tiger sharks with 286. Bull shark was the last species with a significant number of attacks and that was 171. While looking at species we felt it was important so see if location had

any impact. Looking at the 15 most abundant species of shark grouped by country gave us an idea of which sharks were active where. White sharks made up the most attacks for each separate country. After that the data started to vary. Tiger and bull sharks were the next two major species for both the United States and Australia while the ragged tooth shark came in second for South Africa. The US was unique in the amount of spinner and nurse sharks that were active along its coasts. Australia on the other hand had a lot of activity from bronze whaler sharks, wobbegong sharks, and hammer grey nurse sharks.



After looking at species across country we wanted to investigate the severity of attacks by each species. Figure 7 in the appendix gives a quick overview of how injuries were broken down and their frequencies. Going down from most frequent of fatal, bitten, laceration, no injury, minor, and final drowning. White sharks have an unusual high amount of attacks resulting in no injury while for other species it is only their third most common result. Tigers and White attacks result in a high number of fatalities followed by bites and then lacerations. The injury pattern is from most severe to less severe. Bull sharks on the other hand have more bites then fatal attacks but a very small number of no injuries. This leads to the conclusion that while whites and tiger sharks can be very fatal there is a good chance you will be unharmed after an encounter with one of those two species. Bull sharks on the other hand are much more likely to cause minor damage without being fatal, but an encounter with a bull shark almost always ends with some amount of injury.



Conclusion

We found some relationships that surprised us and some that ran counter to common shark attack myths, but there was a common pattern. For many of the questions we answered, we found a relationship reflecting human behavior around going into the water, leaving us less certain about the effect of shark behavior. We found that the United States, Australia, and South Africa were the top three locations of shark attacks. The most common activities people attacked by sharks were participating in were surfing, swimming, and fishing. The effects of weather on attacks were dependent on location as well. While the sample size was not as large as we hoped it still gave us an idea about shark attack trends. In Florida the daily minimum temperatures during attacks were much lower than in North Carolina. We attribute this to Florida having milder winters so people are more likely to be in the water year round. In North Carolina the summer months are packed with beach goers providing much more opportunities for shark attacks to occur. This same increase in human activity could also be the explanation for the afternoon being the time of day that most people were attacked by sharks. It is entirely possible that while these factors have the largest amount of attacks you are more likely to be attacked during other times or weather conditions but since less people are in the ocean the attack counts do not have the same numbers. White sharks, tiger sharks, and bull sharks were responsible for the largest

number of attacks and they all seemed to have a fairly consistent pattern of injuries. By looking at age it is clear that while the very young and very old are involved in shark attacks at much lower rates, the percentage to sharks attacks that are fatal are much higher. We can not definitively say whether a combination of these factors will lead to a shark attack or keep you safe from one, but they do give a good idea of trends and patterns that can be associated with shark attacks.

Future Work

To further the project, it would be important to control for the number of people surfing, swimming, or otherwise potentially coming into contact with sharks. In addition to changes in population, countries and coastal communities have had changes in tourist numbers that probably put more people in the water than ever before. An additional control for shark attack risk could be approximate time in the water per person. If surfers spend more time in the water per person than swimmers, we would expect more shark attacks on surfers if all else was the same.

A potential limitation we had in the data that we could work to fix to further our project was in the analysis of our injury variable. The injury variable was categorized using simple methods that might misrepresent the full picture available in the data. Using more advanced regular expression captures, or even natural language processing, might make the injury variable more useable for analysis without losing detail.

We were limited in the amount of weather data we could bring in because of the time density of that data. Ideally we would be able to combine weather data for as many locations and dates as we could find. Other limitations of the weather data is that it was missing values for some variables and did not have information that we thought might be important for shark attacks like sky cover and water surface roughness.

Appendix

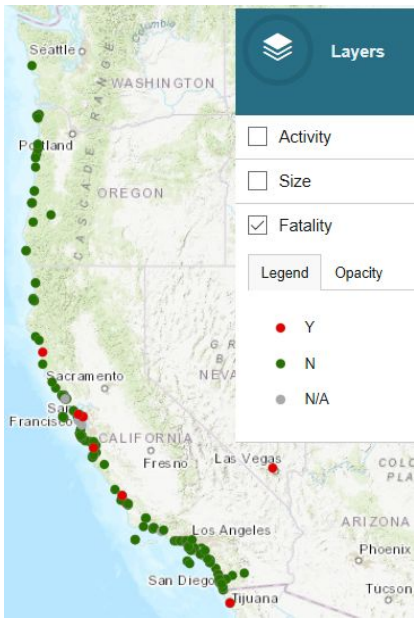


Figure 1 - Shark attacks by fatality - West Coast United States

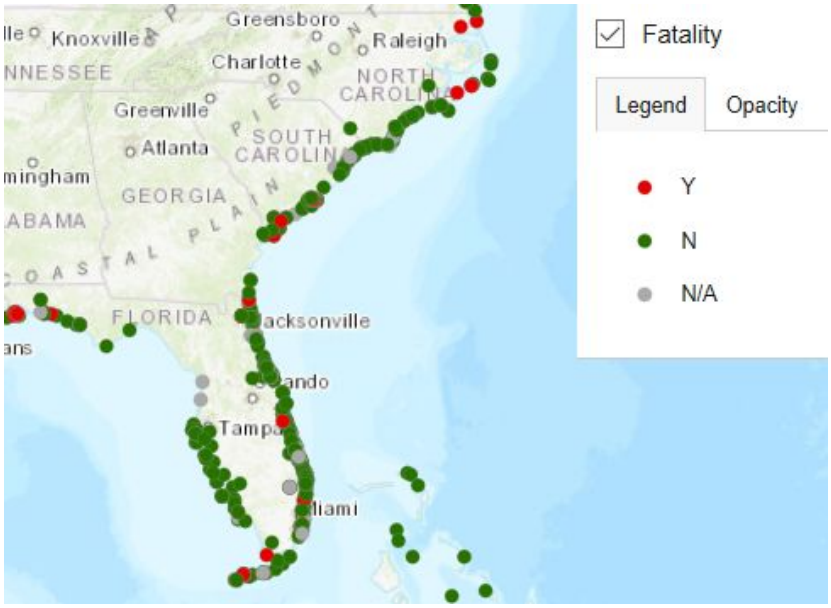


Figure 2 - Shark attacks by fatality - East Coast United States

Figure 3 - Shark attacks by fatality - South Africa

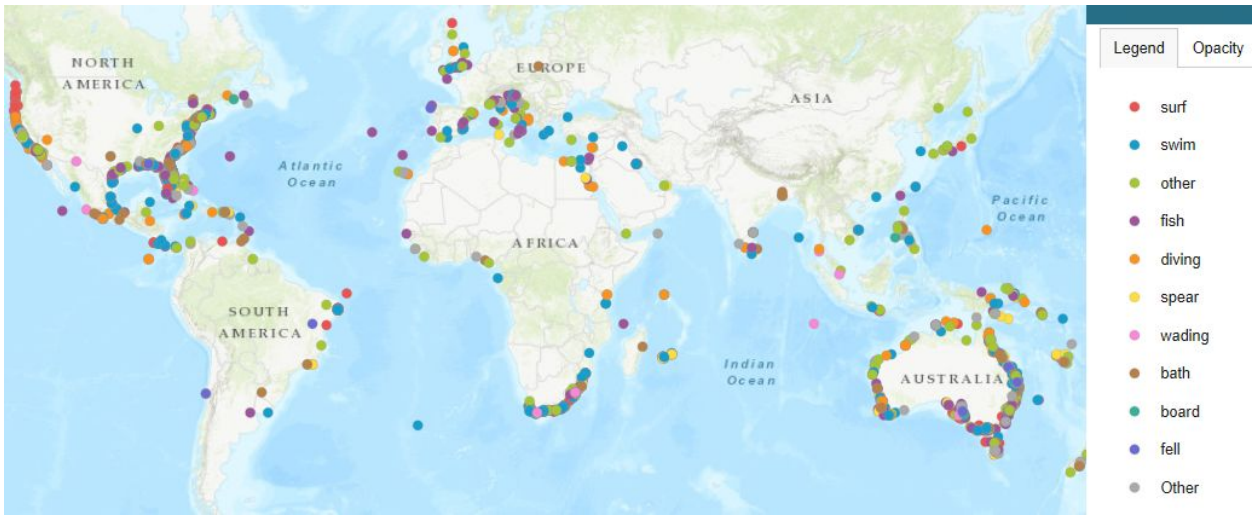


Figure 4 - Shark attacks by activity

Figure 5 - Shark Attack by Size - South Africa

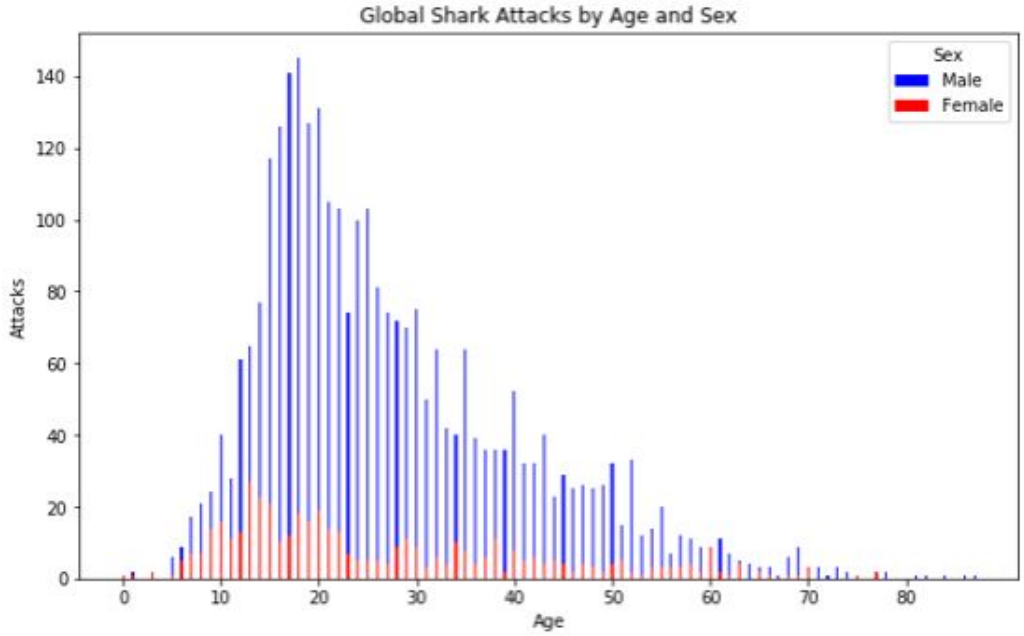


Figure 6 - Global Shark Attacks by Age and Sex

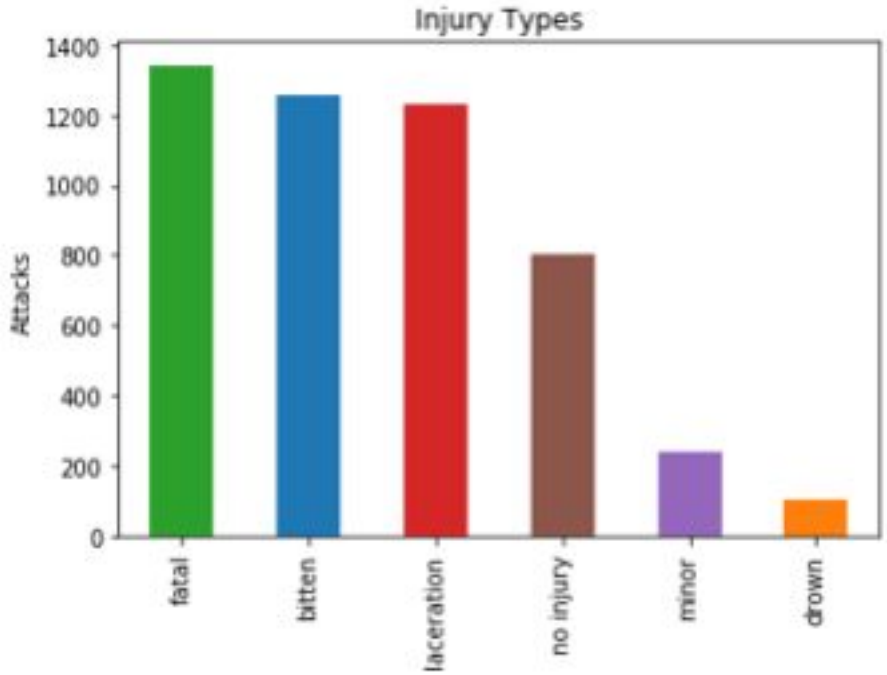


Figure 7- Injury Types

Reference

https://www.newsweek.com/surfer-increase-waves-crowded-congested-trestles-477005
https://www.sharkattackfile.net/incidentlog.htm
https://en.wikipedia.org/wiki/Black_December