Shark Attacks:
A Deeper Dive

By: Bradley Newell, Grant Hagen,
Kelsey Mersinas, and Lindsey Krempa

Contents

Data Breakdown and Cleaning

Selecting Our Columns

Our dataset tracked shark attacks starting from 1845 until 2018. In the dataset, we were given 22 columns. Those being: Case Number, Date, Year, Type, Country, Area, Location, Activity, Name, Sex , Age, Injury, Fatal (Y/N), Time, Species, Investigator or Source, pdf, href formula, href, Case Number, Case Number, and original order. After some deliberation, we decided to keep 9 of those columns. Those being: date, type, country, activity, sex, age, injury, fatal, and species.

Dropping Unnecessary Rows of Data

After selecting our columns we then started the process of cleaning the data. First, we noticed that there were unnecessary blank columns that were being recognized as legit rows of data. We dropped those rows. Second, there was another row of data thousands of rows down in the data that only had the value “xx”, and that was also being recognized as a row of data. We also dropped that row. Lastly, we found a few instances where there were rows of data that only had the value “0” and nothing else. We dropped those rows as well.

Reorganizing Values and Creating Categories

The most qualitative data was found in the columns 'Species', 'Activity', and 'Injury'. The age of this dataset meant that many values were pulled from historical documents and research in various formats, being reported as entire sentences or descriptions of situations. While these values contained valuable information, they needed to be converted into more general categories. For all three of these columns, a dictionary could be used to contain the key terms we wanted to search for in each line and the corresponding values we would replace them with. Using a .contains method in a for loop simplified the process of narrowing down our diverse data into a few recognizable categories. For standardization, we filled any blank cells with 'Unknown' and categorized any activities that were so unique they did not fit into any predefined category as 'Other'. This process resulted in data that was generalized yet still conveyed a clear story about our shark attack data.

There were several quantitative series that held simple values describing the shark attacks. The gender of the victim was intended to be represented as 'M' for male or 'F' for female. Similarly, the 'fatal' column was set up to hold either 'Y' (yes) or 'N' (no) to indicate whether the attack was fatal or not. However, both of these columns contained typos and errors along with many null values. By displaying the unique values of each column, we were able to use .loc to search for values that did not belong and print the corresponding rows to assess the values in these rows. Since there were just a few typos, they could be safely removed. Additionally, some values were correct but had extra spaces or were not capitalized, which could be remedied with .replace. This process resulted in two columns holding clear information that could be applied to our analysis.

Reorganizing the Age Column of the Dataset

Another column in our dataset that required significant cleaning was the age column. Upon initial review of this data, we noticed that there were many missing values for the age column, indicating that the age of these shark attack victims was probably unknown. We knew that we did not want to drop these missing values because it would have substantially decreased our total dataset. The age column also included an immense amount of unique characters that would have been very tedious to group or categorize, such as “mid-30s”, “10 or 12”, “teen”, “young”. With the help of Prof Booth, we used a .loc on a specific section of the age data that did contain any special characters to create a new data frame. This cleaned up data frame allowed for a more effective analysis of the age of the shark attack victims in the dataset.

Cleaning Up Countries

Another column that needed to be cleaned was “country”. In the dataset there were a few instances of a country being misspelled. Other instances where an inadvertent blank space was present. For instance, “France” and “ France”. Lastly, there was a few instances where the same country was referred to, but used different names (i.e. Great Britain the United Kingdom). To combat these issues, I utilized the package “fuzzywuzzy” to perform fuzzy matches on incorrect or differently named countries. However, even after performing fuzzy matches, there were still instances where the country was returned as a null. There were rows of data where the attack was in the middle of the ocean or only a continent was shown. In those instances, country was labeled “Unknown” and we proceeded.

Adding Additional Columns

After selecting our data, dropping the instances of unnecessary rows, and cleaning our dataset, we then added some columns to aid in our analysis. The first of which being the latitude and longitude. With the help of the package “country info” we were able to find the latitude and longitude of each country. For each row of data that isn't labeled “Unknown”, we appended the latitude and longitude to their respective columns: lat and lon.

Subsequently, once the coordinates were found for our countries, we were then able to find which hemisphere each country was in. We created a function that took in the country column, found the latitude/longitude and then sorted each row into 1 if it was in the northern hemisphere, 2 if it was located in the southern hemisphere, and 3 for anything that couldn't be sorted (these instances will later be labeled as unknown). Now we have the hemisphere where each country is in.

Next, we can create dictionaries that will help us see what season each attack was in. See **appendix entry A** to see the dictionaries. With the dictionaries created, we then ran a for loop to append the seasons utilizing the hemisphere column created earlier. See **appendix B** to see the code.

Is There a Relationship Between Gender and Activity?

We thought that it would be interesting to look for any trends while comparing the gender of the shark attack victims. We ultimately decided to look for any significant relationships between gender and activity, such as “surfing”, “swimming” or “fishing” for example. We wanted to know if there were any activity types that had notably higher female victims, and vice versa for male victims.

We hypothesized that the overall number of male victims will be higher than female victims when grouping by activity. To our surprise, our horizontal bar graph showed that the number of male victims was higher than females in almost every activity type, with the exception of “walking”. Even though “walking” had more female victims, it was not by a significant amount. See **Appendix C** for our horizontal bar graph comparing the number of male and female victims for each activity type.

We did ponder upon these results. Are males more likely to be risk-takers and participate in more dangerous activities such as surfing and diving? Do women just have more common sense? Either way, we were correct that the overall number of male victims were higher than female victims. Looking at the overall distribution of male and female victims (regardless of activity type), the number of male victims was a significant 87.3% while females were 12.7%. See **Appendix D** for the pie chart comparing the distribution of female and male shark attack victims.

What were the Most Frequent Activities for Each Gender?

We wanted to dive deeper into this research question by finding out what the most frequent activities were for both males and females. Would there be remarkable differences between the activity types of males and females? Or would they essentially be the same? Initially, we tried to create a bar graph that had “activity” on the x-axis and the number of victims on the y-axis, with two different colored bars for each gender. However, this bar graph was very cluttered due to the high number of different activities on the x-axis and did not call for an effective analysis.

We ultimately decided to create two different donut graphs, one for males and one for females, that focused specifically on the top five most frequent activities for each gender. See **Appendix E** for these donut graphs. These donut charts showed that the five most frequent activities were the same for both males and females, with the exception of “fishing” for males and “wading” for females. The rest of the activities, which included “swimming”, “surfing”, “diving”, and “not recorded”, were included in both donut charts in different frequencies for each gender.

Overall, focusing specifically on gender as a variable yielded some interesting results and had us wondering if there are any other worthwhile relationships to explore in our dataset. What will we find if we compare injury and gender? Did one gender have a significantly higher fatality rate than the other? Will the ages of male and female victims produce any compelling trends, or will these results be scattered? These are just a few examples of what we can investigate in the future.

Is there an activity related to shark attacks?

The combination of our variables and their impacts on our shark attack data can be interpreted in multiple ways. When displaying the activities of the shark attack victims we wanted to be able to incorporate other factors into these plots to see other relationships. To interpret the data we started with visualizing the activities of the victims to see if there were any trends we could use in predicting shark attacks. We represented the distribution of activities involved in shark attacks using a bar chart, categorizing each activity by fatal and non-fatal encounters. This visual representation confirmed our initial hypothesis that activities such as surfing and swimming are significant predictors of shark encounters. However, due to the variability of our activities and the volume of data presented, this chart ends up with both very large and very small numbers that can be difficult to read when all together.

To use a more detailed approach, we wanted to get a closer look at the activities per shark, but this graph was difficult to read due to the varied amount of activities. While it provided a picture of the entirety of the data set, there was too much on the plot to be able to read and understand. And while this plot was showing the total number of attacks, it couldn’t give any insights into fatalities. To provide a more complete and focused visual, we identified the top five sharks responsible for the highest fatality counts, as these represented the most dangerous species. We created a plot depicting the most common activities associated with attacks by these selected sharks. This approach yielded clearer insights into the activities most commonly linked with incidents, especially concerning the sharks most frequently encountered in our data set. This plot highlights activities perceived as posing higher risks, particularly in encounters with the most fatal sharks. For a comparison between the comprehensive graph and the filtered version, see **Appendix F**.

Attacks by Species

The species of the shark encountered was a variable that we could investigate for relationships with our other values such as fatality of the attack and use to add to the picture of our shark attack data. Looking into how species of the shark itself relates to the frequency of shark attacks we made a basic graph to display the total number of fatalities attributed to each species. This revealed that the Great White Shark had the highest count of fatal attacks in our dataset. While a few other species also showed high fatality counts, many species had little to no documented fatal attacks.

We developed a stacked bar graph to compare the total encounters of each species of shark with the percentage of fatal attacks. This visualization provided us a complete view of both the raw counts and the percentage for each shark species. This gave us a clear view of the counts of attacks per species and the percentages of fatal attacks attributed to the species. It highlighted that Bull sharks, despite lower overall encounters, had the highest percentage of fatal attacks.

In categorizing our original data set into manageable groups, we also observed a significant number of unreported incidents. Approximately half of our shark attacks had not reported on the species of the shark involved. This limitation in our data should be considered as while great white sharks were over represented, their greater visibility and society's awareness of them could have led to more frequent identification compared to less well-known shark species.

Is there seasonality to shark attacks?

Another question that we pondered was that of seasonality. Was there seasonality to shark attacks? Were there a disproportionate amount of shark attacks in the summer months when more people are visiting beaches? From a holistic view, we can glean that there appears to be more shark attacks in the summer months, garnering 41% of all attacks. Fall and spring were fairly similar with only a 6% difference in the total number of shark attacks, winter coming last with 12% of shark attacks see **appendix G**. All of this makes intuitive sense. The warmer the season, the more people swimming in shark infested waters. Leading to more shark attacks.

A Global Look at Attacks

Diving in further we can see the impact of seasonality. We can see the difference in seasonality using bubble maps see **appendix H**. The size of each red dot signifies the number of shark attacks per country, the bigger the dot, the more shark attacks there are. I should note that the location of the dots are not the area where the attacks occurred, just where the countries are located. We can also glean that there are some unfortunate heavy lifters when it comes to shark attacks. The United States, South Africa, and Australia have the highest number of reported shark attacks.

Top Countries Per Season

If we take the top three countries by total shark attacks, and create a stacked bar chart broken up by season (**appendix I**), we can see further that there is seasonality to shark attacks. The United States reported the highest number of shark attacks, coming in at a whopping 2400 reported shark attacks, Australia in second coming in at about 1200, and South Africa with approximately 500. With the three countries shown here (**appendix I**), we can glean further that shark attacks are at their peak in Summer, slows down in the Fall, bottoms out in the Winter, and then ramps up in the Spring.

Hemisphere Breakdown

Once we had organized our data by location we could further break down our data into locations to see the relationships in our data frame. Knowing that the weather and time would be different per the different hemispheres, we wanted to apply these considerations with a line graph to get a clearer picture of our shark attack data over time.

When plotting attacks over time it became apparent that there is a significant contrast between the northern and southern hemispheres, notably due to their opposite timing of the seasons. We had made a graph to show the count of attacks per month, but this led us to further analyze shark attack data by breaking down the attack counts according to hemispheres. This approach revealed a distinct pattern where summer months consistently showed the highest rates of attacks. For instance, January, corresponding to peak summer in the southern hemisphere (particularly in Australia and South Africa), emerged as a notable period for shark incidents.

However, we recognized a potential bias in our data as during summer months there tends to be a higher number of people in the ocean and at beaches, which may be contributing to increased reported shark encounters. As our data was primarily collected as reactionary reports following shark encounters, it is not declarable from this data if the influx of people is to blame for the surge of attacks in summer months or if sharks become more active and territorial in warmer weather.

Is there a relationship between age and fatality?

One variable in our dataset was the age of the victim. We wanted to investigate any potential relationship between age and the likelihood of a shark attack.. We looked only at fatal attacks, and used .groupby for each age to display the data on a scatter plot. This revealed a notable peak on fatal attacks around 20 years old. Original scatter plot in **Appendix J**. To further analyze this relationship, we split the ages at the peak of the attacks and we applied a regression line using seaborn to show any correlation. Although this revealed a strong correlation, our dataset disproportionately represented individuals around 20 years old. This skew is likely because younger individuals are more active in recreational beach activities compared to other age groups, prompting us to examine this relationship more closely to determine causality.

With our age groups divided at 20 years old, we had two groups to compare to determine if age is in fact a strong predictor of the chances of being attacked by a shark. With a T Test we can compare the means of these two age groups and return whether the results we were seeing in our data were results of a cause and effect relationship or if it was attributed to chance. Conducting a T test between these groups, we have a Null Hypothesis that there is no relationship between the the age of a person and the likelihood they will be attacked by a shark. We found a p-value of 0.27, being higher then the expected .05, indicating that we should accept our Null Hypothesis as there is no statistically significant difference in fatality rates between the age groups. This suggests that age alone does not increase the likelihood of being a victim of a shark attack. T Test in **Appendix K**.

Although we initially noted a concentration of incidents around 20 years old, the overrepresented sample size of this age group may have inaccurately skewed age as a correlated factor with fatality rates. Further exploration into the distribution of different ages in our study reinforces that age alone does not significantly influence the likelihood of a fatal shark attack.

Bias, Limitations, and Future Work

A few notes about our dataset. First biases we noticed. For starters, there are omitted variables all across the dataset. Over half of the dataset is missing at least one value, and also the shark attack reporting relies on third party news outlets. This will provide the issue of unstandardized entries. For instance, whether or not someone was provoking a shark may be interpreted differently by different outlets, and some of the rows of data have null values for their source.

Similarly we had some serious limitations. The dataset provides very inconsistent structure. Date values have inconsistent formatting. Location (which is basically cities) have how far away they are from shore but most don't. Species values sometimes have the size of the shark, but most of the time it doesn't. There were many null values spread out across the dataset.

Finally, we have some constructive notes for the kind folks that keep track of shark attacks. For starters, we would recommend that there be more consistency in their data structure (species, dates, locations). We would also add a few columns to the dataset for even further analysis. Latitude and longitude of the attack would be super fun to have bubble map charts. Weather would be interesting to delve into further analysis. How far away the attack was from shore could provide some additional clarity to our analysis.

Citation

Data From

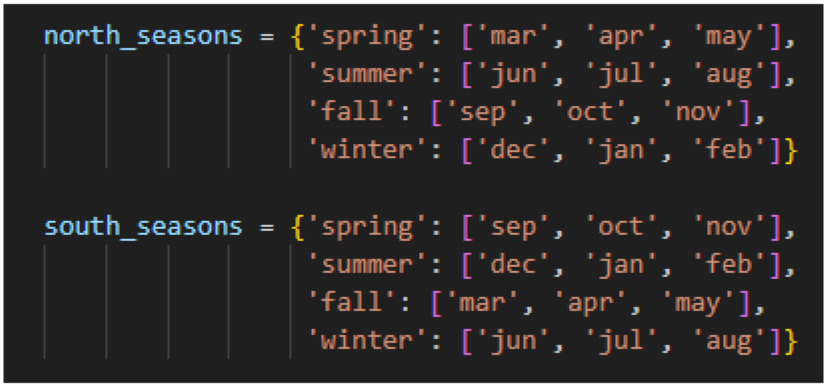
"Global Shark Attack File." Global Shark Attack File, sharkattackfile.net/index.htm

Color Pallette

"Color Palette 112421." Color-Hex, www.color-hex.com/color-palette/112421

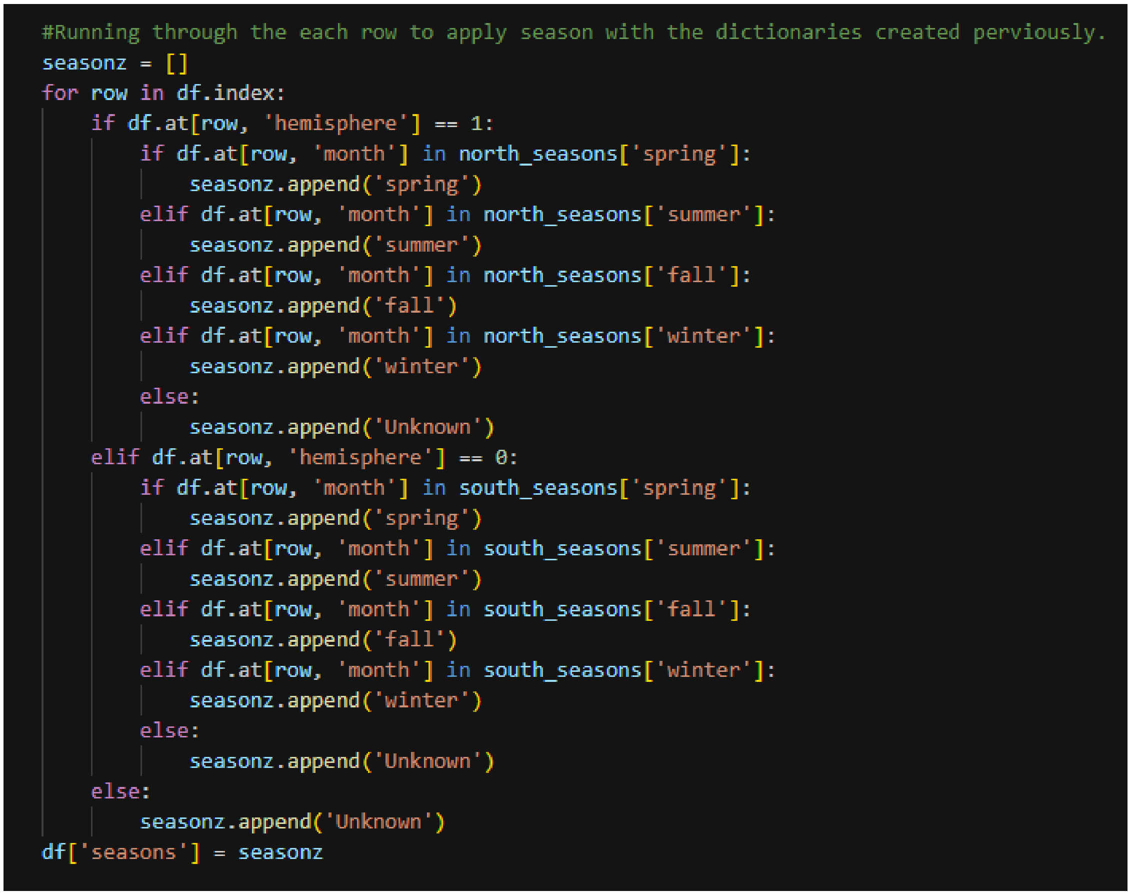
Appendix

Appendix A



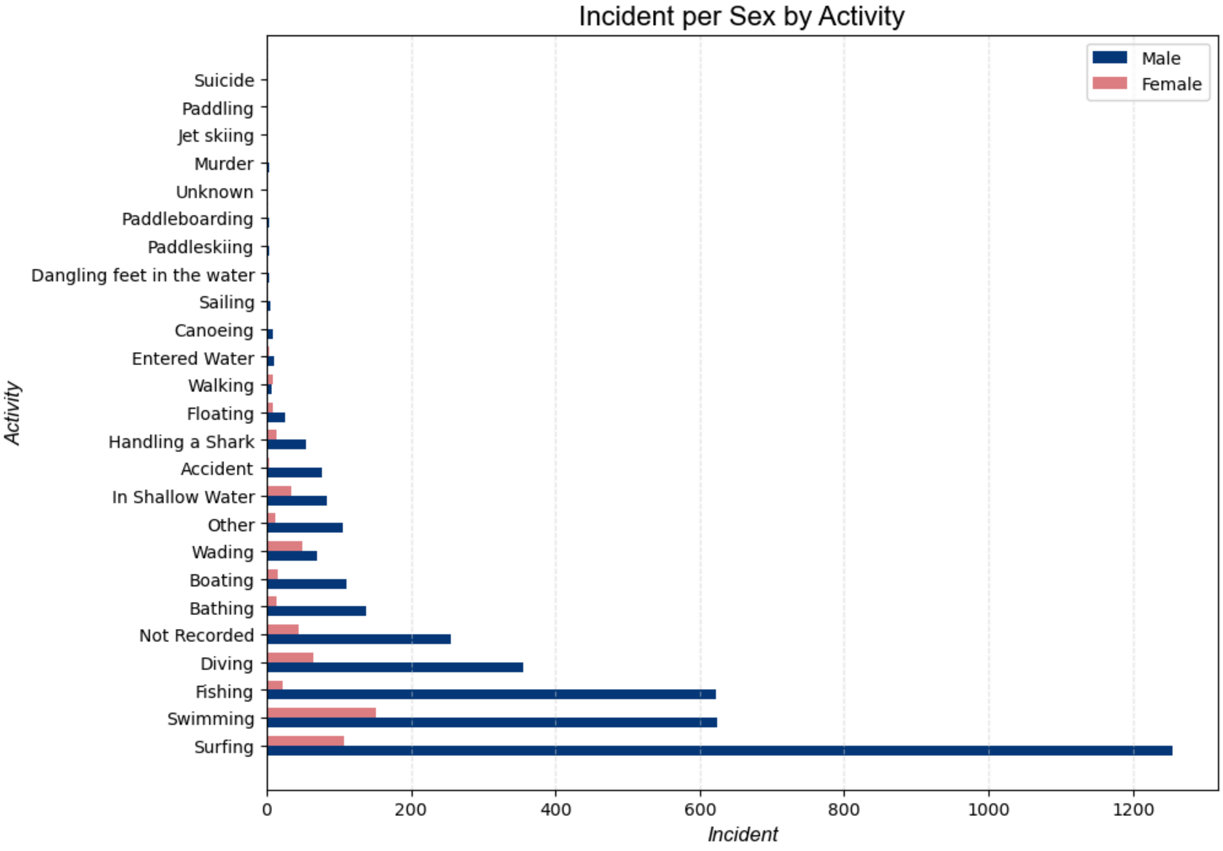
```
north_seasons = {'spring': ['mar', 'apr', 'may'],
                 'summer': ['jun', 'jul', 'aug'],
                 'fall': ['sep', 'oct', 'nov'],
                 'winter': ['dec', 'jan', 'feb']}

south_seasons = {'spring': ['sep', 'oct', 'nov'],
                 'summer': ['dec', 'jan', 'feb'],
                 'fall': ['mar', 'apr', 'may'],
                 'winter': ['jun', 'jul', 'aug']}
```

Appendix B

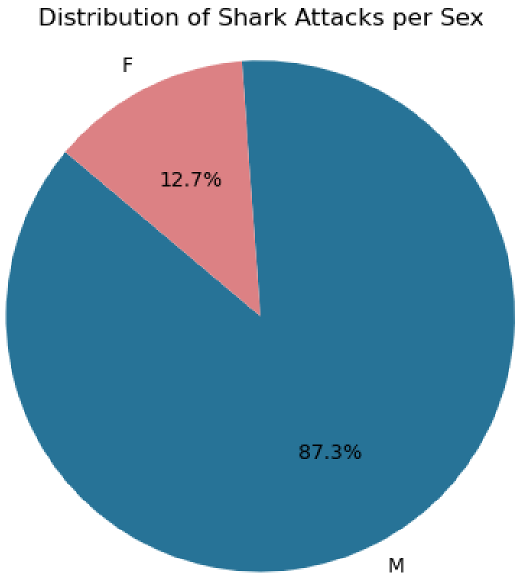


```
#Running through the each row to apply season with the dictionaries created perviously.
seasonz = []
for row in df.index:
    if df.at[row, 'hemisphere'] == 1:
        if df.at[row, 'month'] in north_seasons['spring']:
            seasonz.append('spring')
        elif df.at[row, 'month'] in north_seasons['summer']:
            seasonz.append('summer')
        elif df.at[row, 'month'] in north_seasons['fall']:
            seasonz.append('fall')
        elif df.at[row, 'month'] in north_seasons['winter']:
            seasonz.append('winter')
        else:
            seasonz.append('Unknown')
    elif df.at[row, 'hemisphere'] == 0:
        if df.at[row, 'month'] in south_seasons['spring']:
            seasonz.append('spring')
        elif df.at[row, 'month'] in south_seasons['summer']:
            seasonz.append('summer')
        elif df.at[row, 'month'] in south_seasons['fall']:
            seasonz.append('fall')
        elif df.at[row, 'month'] in south_seasons['winter']:
            seasonz.append('winter')
        else:
            seasonz.append('Unknown')
    else:
        seasonz.append('Unknown')
df['seasons'] = seasonz
```

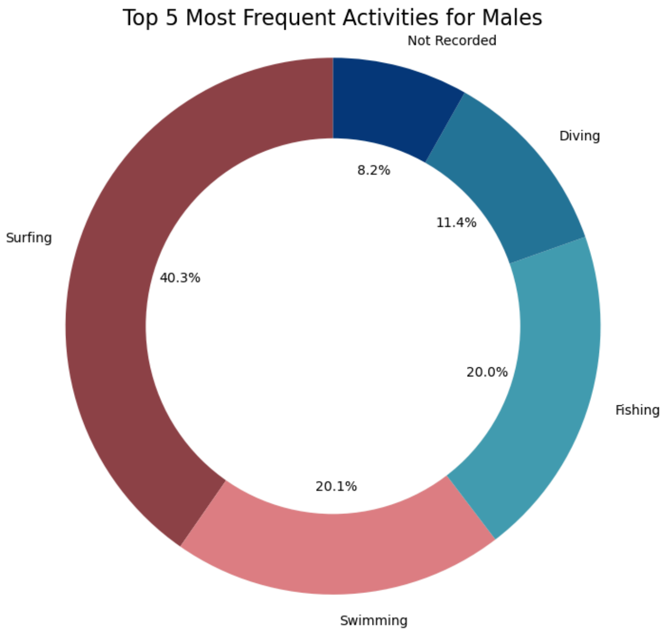
Appendix C

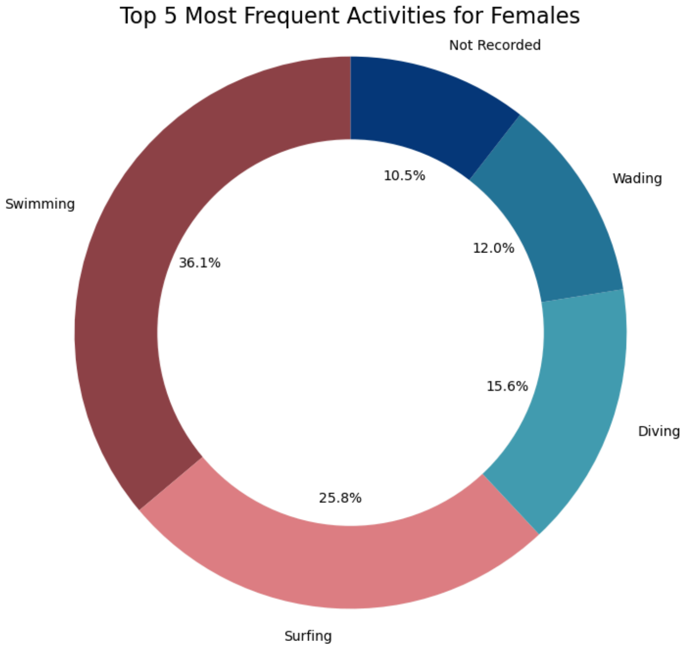

Incident per Sex by Activity
Male
Female
Activity
Suicide
Paddling
Jet skiing
Murder
Unknown
Paddleboarding
Paddleskiing
Dangling feet in the water
Sailing
Canoeing
Entered Water
Walking
Floating
Handling a Shark
Accident
In Shallow Water
Other
Wading
Boating
Bathing
Not Recorded
Diving
Fishing
Swimming
Surfing
0
200
400
600
800
1000
1200
Incident

Appendix D

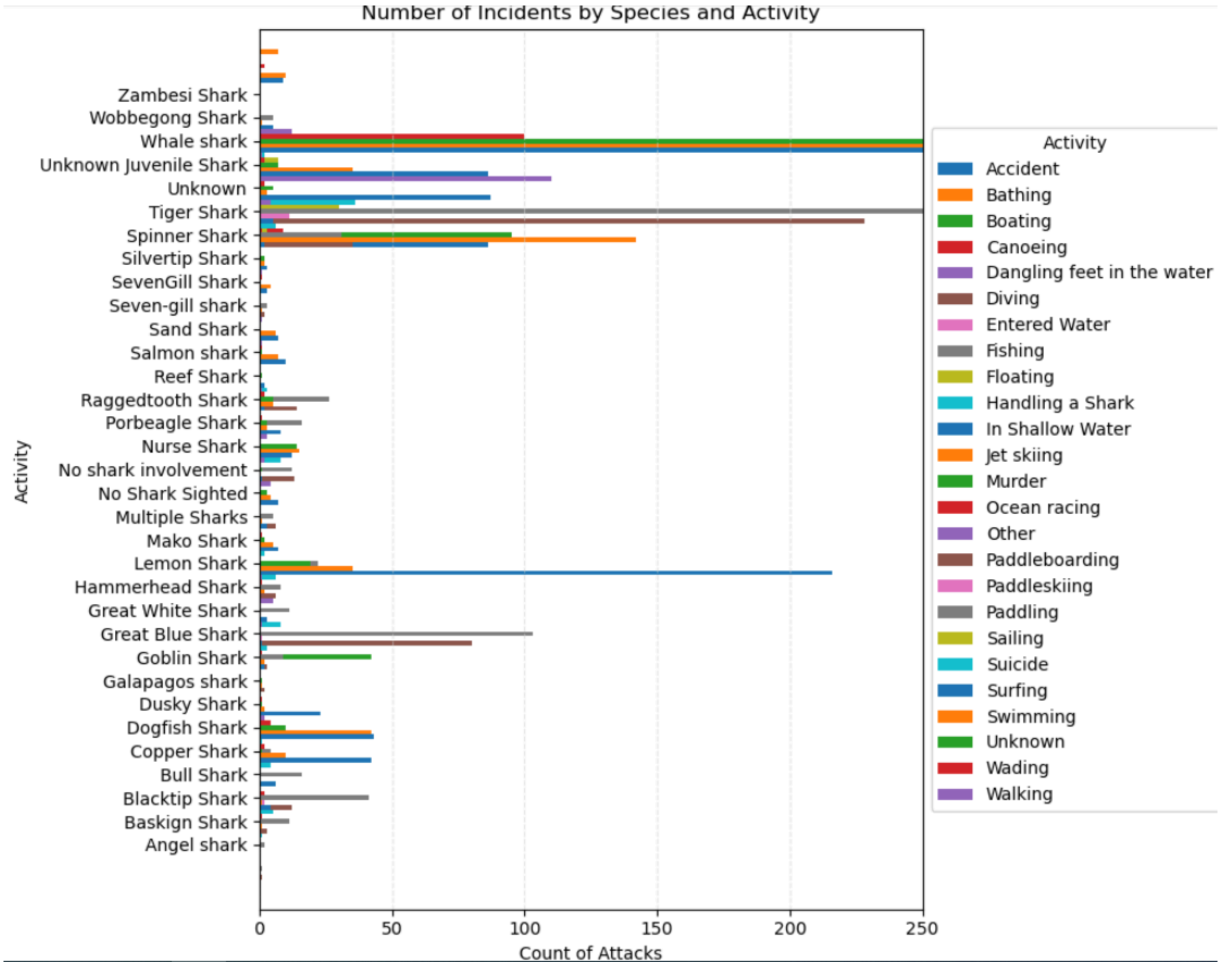


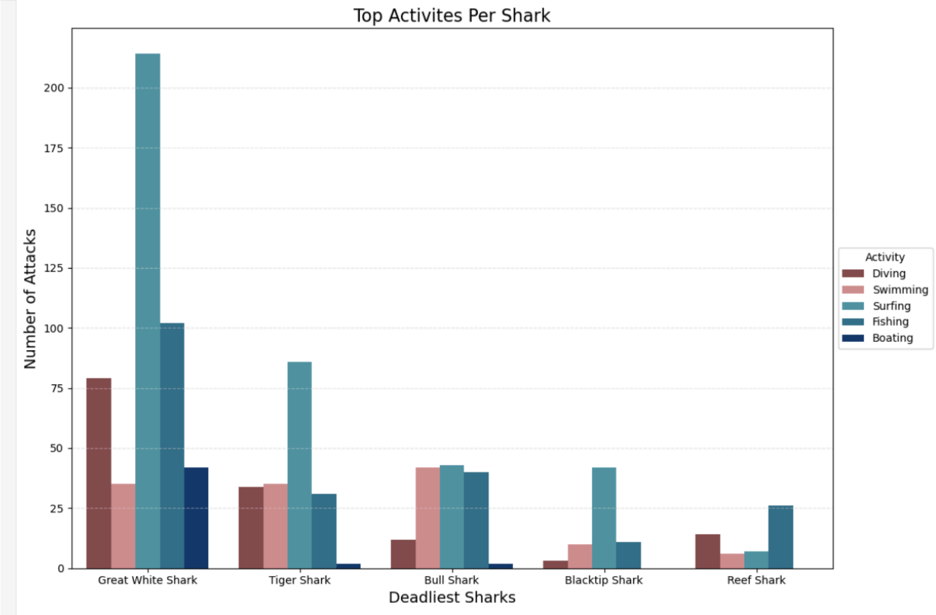
Appendix E


Top 5 Most Frequent Activities for Males
Not Recorded
8.2%
Diving
11.4%
Surfing
40.3%
20.0%
Fishing
20.1%
Swimming

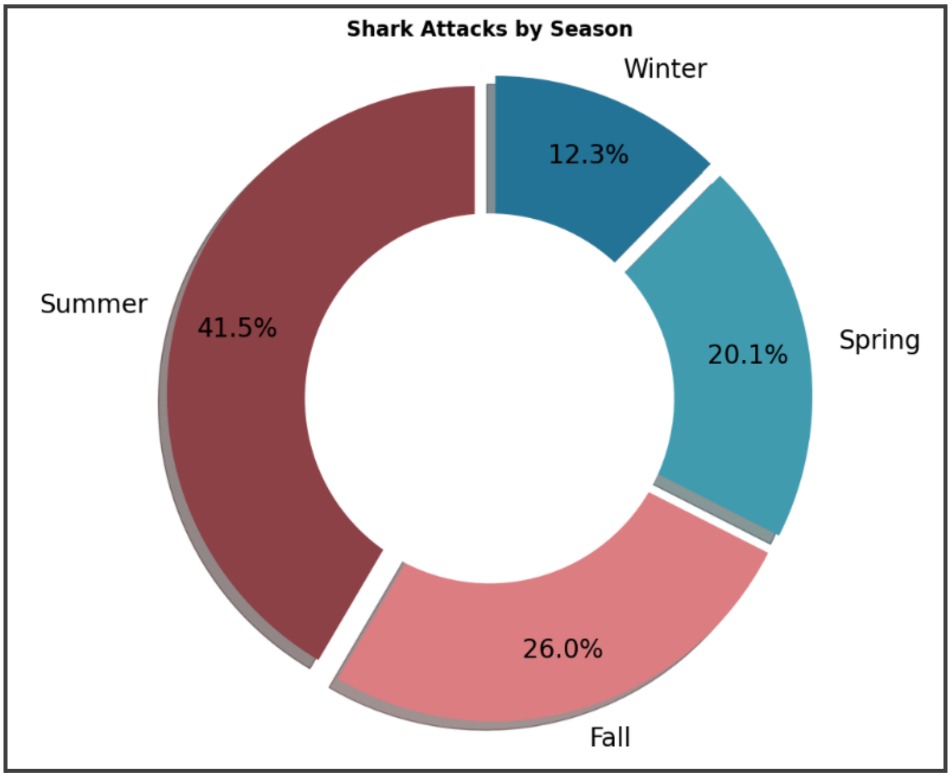

Top 5 Most Frequent Activities for Females
Not Recorded
10.5%
Wading
Swimming
36.1%
12.0%
15.6%
Diving
25.8%
Surfing

Appendix F

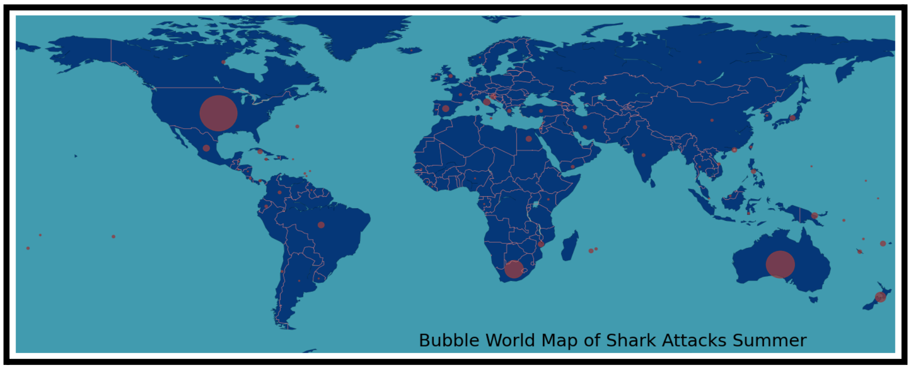

Number of Incidents by Species and Activity
Zambesi Shark
Wobbegong Shark
Whale shark
Unknown Juvenile Shark
Unknown
Tiger Shark
Spinner Shark
Silvertip Shark
SevenGill Shark
Seven-gill shark
Sand Shark
Salmon shark
Reef Shark
Raggedtooth Shark
Porbeagle Shark
Nurse Shark
No shark involvement
No Shark Sighted
Multiple Sharks
Mako Shark
Lemon Shark
Hammerhead Shark
Great White Shark
Great Blue Shark
Goblin Shark
Galapagos shark
Dusky Shark
Dogfish Shark
Copper Shark
Bull Shark
Blacktip Shark
Baskign Shark
Angel shark
Activity
0
50
100
150
200
250
Count of Attacks
Activity
Accident
Bathing
Boating
Canoeing
Dangling feet in the water
Diving
Entered Water
Fishing
Floating
Handling a Shark
In Shallow Water
Jet skiing
Murder
Ocean racing
Other
Paddleboarding
Paddleskiing
Paddling
Sailing
Suicide
Surfing
Swimming
Unknown
Wading
Walking


Top Activites Per Shark
0
25
50
75
100
125
150
175
200
Number of Attacks
Great White Shark
Tiger Shark
Bull Shark
Blacktip Shark
Reef Shark
Deadliest Sharks
Activity
Diving
Swimming
Surfing
Fishing
Boating

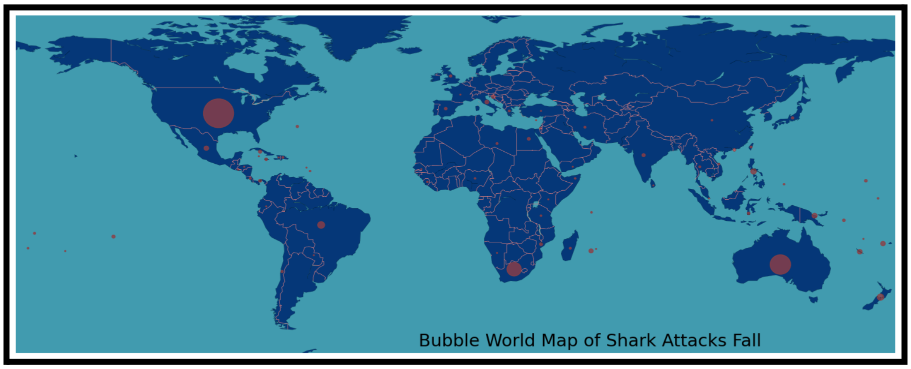
Appendix G


Shark Attacks by Season
Winter
12.3%
Summer
41.5%
20.1%
Spring
26.0%
Fall

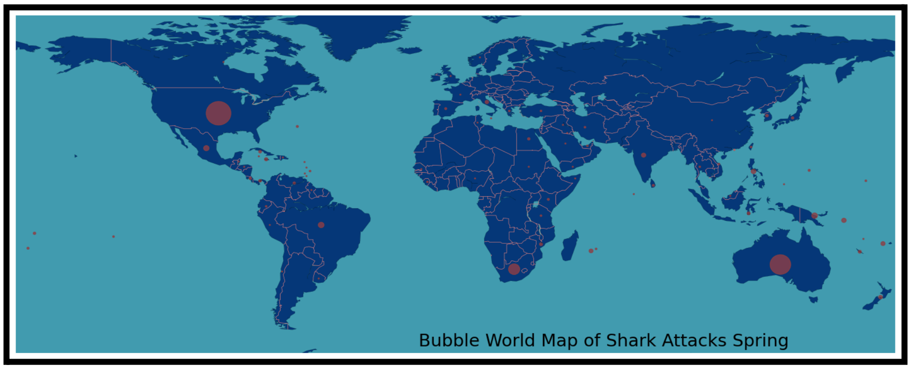
Appendix H



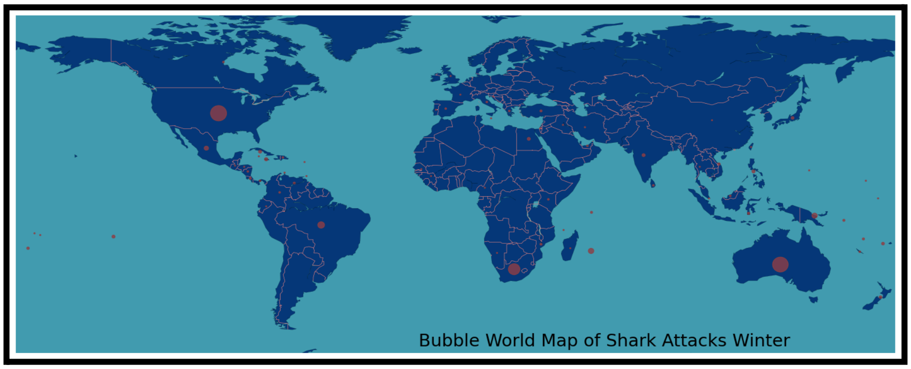
Bubble World Map of Shark Attacks Summer



Bubble World Map of Shark Attacks Fall

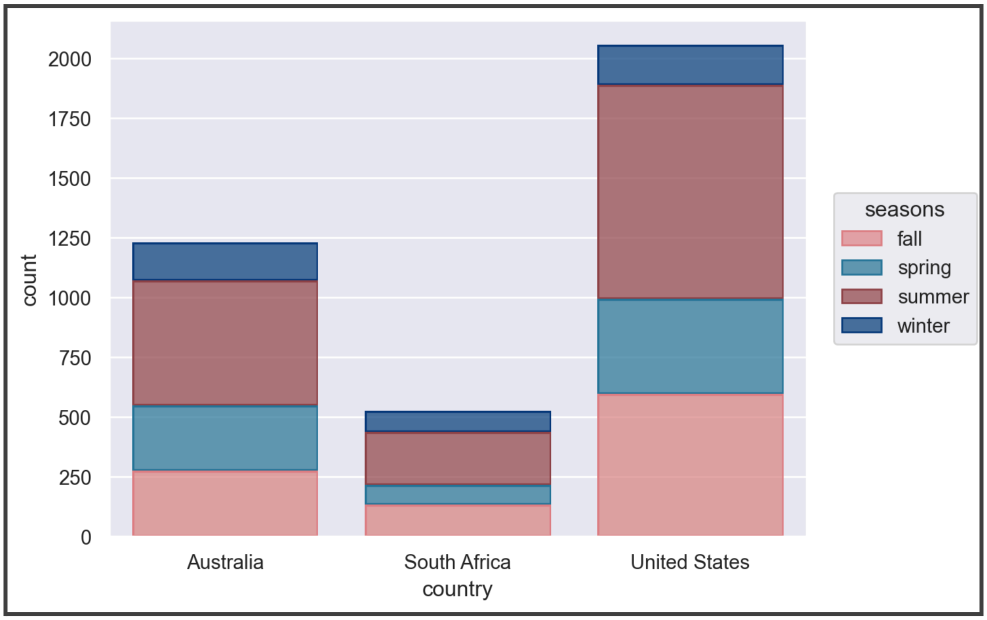


Bubble World Map of Shark Attacks Spring

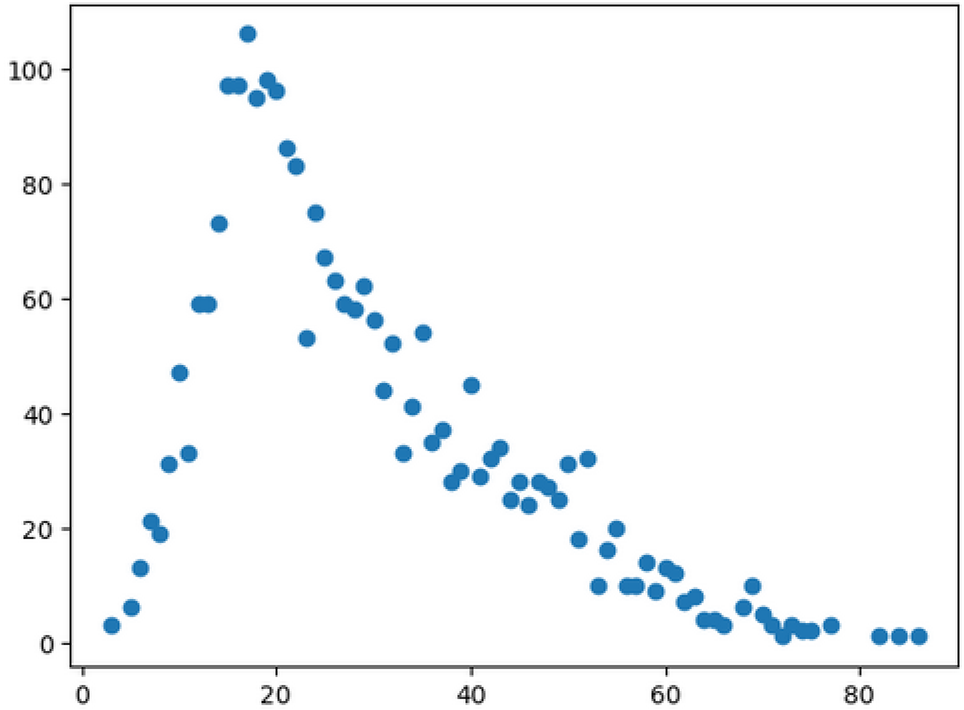


Bubble World Map of Shark Attacks Winter

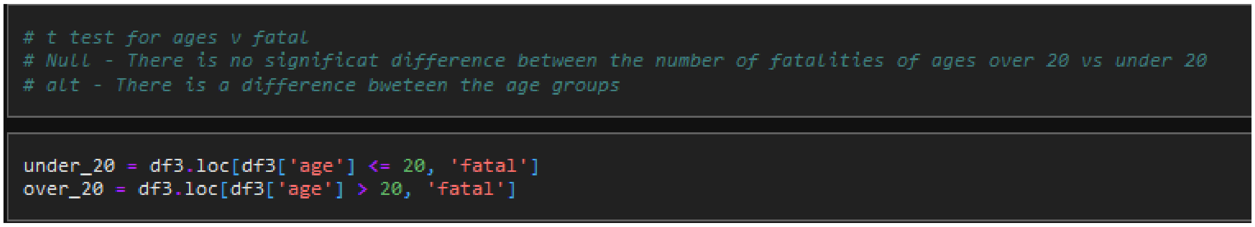
Appendix I


2000
1750
1500
1250
1000
750
500
250
0
count
Australia
South Africa
United States
country
seasons
fall
spring
summer
winter

Appendix J

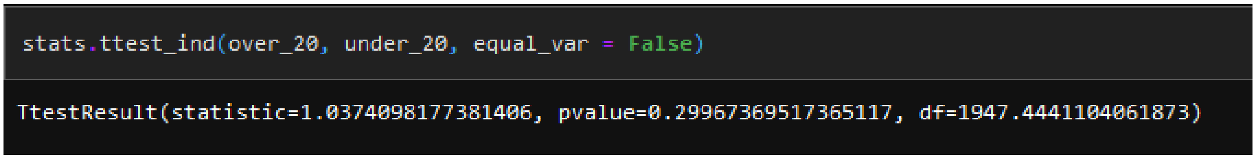


Appendix K



```
# t test for ages v fatal
# Null - There is no significat difference between the number of fatalities of ages over 20 vs under 20
# alt - There is a difference bweteen the age groups
```

```
under_20 = df3.loc[df3['age'] <= 20, 'fatal']
over_20 = df3.loc[df3['age'] > 20, 'fatal']
```



```
stats.ttest_ind(over_20, under_20, equal_var = False)
```

```
TtestResult(statistic=1.0374098177381406, pvalue=0.29967369517365117, df=1947.4441104061873)
```