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?

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 a grouped multibar chart by parent_activity and top seven states for reported shark attack amount [Figure 3.5]. The chart shows that for the top three states (Florida, Hawaii, and California), on_water activities had the highest amount of shark attacks. We also looked at shark attacks on the West Coast vs the East Coast and grouped by parent_activity [Figure 3.4]. In every parent_activity category, there were more attacks on the East Coast vs the West Coast which is expected since the most attacks in the United States occur around Florida.

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. All the maps generated included the info feature which made them interactive and allows the user to click on an attack to see the Case Number, Parent Activity, Parent Injury, Location, State, and Latitude and Longitude [Figure 9.1].

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 [Figure 7.3]. Using a violin chart as well, we were able to visualize the reported attacks by month per year [Figure 6.1].

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 [Figure 7.1]. 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 [Figure 7.2].

Regression

We attempted to create a linear regression model however, our model does not give us an efficient result [Figure 8.1]. The limited data that we are working with has too many variable instances and outliers for our model to properly predict any attacks.

Limitations and future work

The greatest limitation we came across was missing data within the reports. Also, because the case reporting is free response, there were many permutations of some answers. This made analyzing the data that we did have more complicated, and ultimately, we lost some reports. Other things that limit this dataset is lack of reporting worldwide and what is considered a shark attack. There were several reports that did not actual know if the shark was involved in the death or even if a shark was involved at all. For future work, it would be useful to include population rate data for beaches and then get percentages of attacks out of the population that was at the beach. This would give us a better idea of the odds of someone getting attacked based on which beach they attend.

Conclusion

In conclusion, based on the data that was analyzed, we found that shark attacks reported that are fatal are very rare and make up less than 2% of all attack types. These reported attacks happen the most in the summer months from July to September, and during the cooler month between November to March we see these number decrease significantly. Likely because more men tend to do the activities that are typically involved when a shark attack occurs, more than 3 times the number of women is reported as the gender attacked. In the world, Florida has the most reported attacks which is about 27% from 2000 – 2018. Lastly, as we hypothesized, reported attacks are slightly increasing each year.

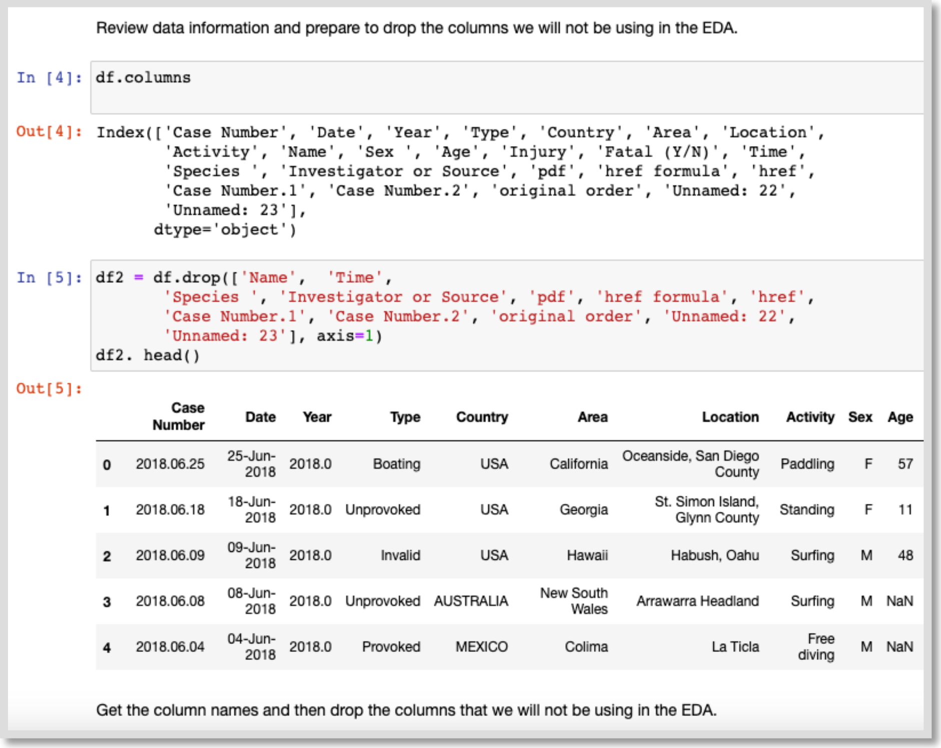
References

https://www.kaggle.com/datasets/mysarahmadbhat/shark-attacks

https://maps.googleapis.com/maps/api/geocode/json?address={text}&state={state}&key={gkey}

Figures

Figure 1.1



Review data information and prepare to drop the columns we will not be using in the EDA.

```
In [4]: df.columns
```

```
Out[4]: Index(['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.1', 'Case Number.2', 'original order', 'Unnamed: 22',
               'Unnamed: 23'],
              dtype='object')
```

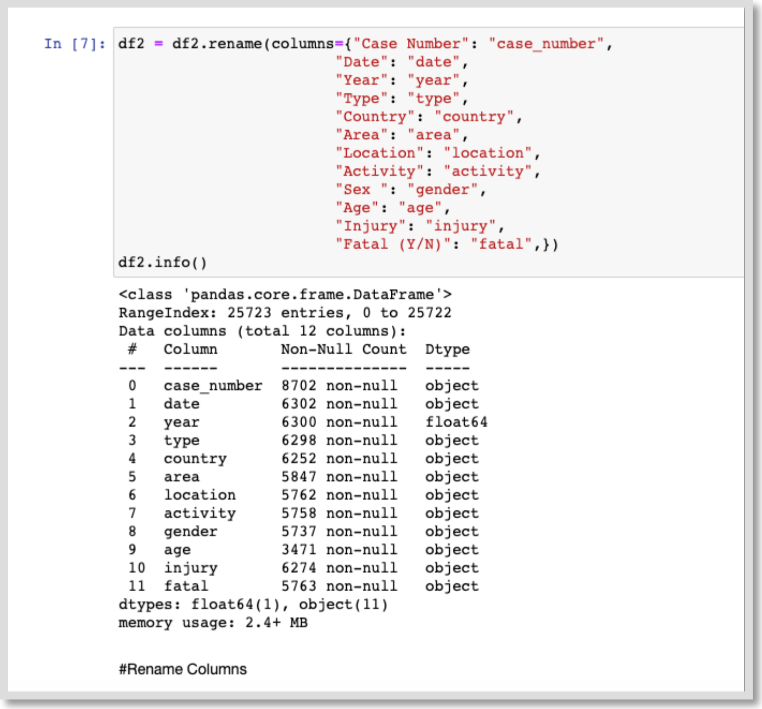
```
In [5]: df2 = df.drop(['Name',  'Time',
               'Species ', 'Investigator or Source', 'pdf', 'href formula', 'href',
               'Case Number.1', 'Case Number.2', 'original order', 'Unnamed: 22',
               'Unnamed: 23'], axis=1)
        df2. head()
```

Out[5]:

	Case Number	Date	Year	Type	Country	Area	Location	Activity	Sex	Age
0	2018.06.25	25-Jun-2018	2018.0	Boating	USA	California	Oceanside, San Diego County	Paddling	F	57
1	2018.06.18	18-Jun-2018	2018.0	Unprovoked	USA	Georgia	St. Simon Island, Glynn County	Standing	F	11
2	2018.06.09	09-Jun-2018	2018.0	Invalid	USA	Hawaii	Habush, Oahu	Surfing	M	48
3	2018.06.08	08-Jun-2018	2018.0	Unprovoked	AUSTRALIA	New South Wales	Arrawarra Headland	Surfing	M	NaN
4	2018.06.04	04-Jun-2018	2018.0	Provoked	MEXICO	Colima	La Ticla	Free diving	M	NaN

Get the column names and then drop the columns that we will not be using in the EDA.

Figure 1.2

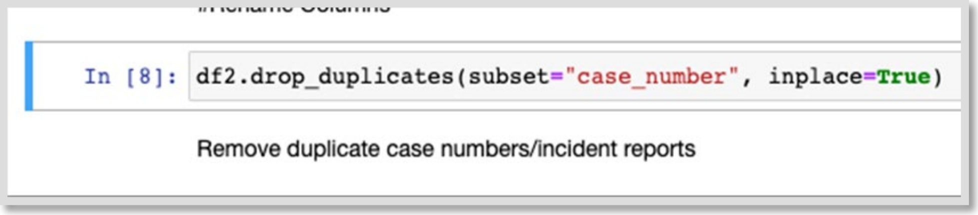


```
In [7]: df2 = df2.rename(columns={"Case Number": "case_number",
                                  "Date": "date",
                                  "Year": "year",
                                  "Type": "type",
                                  "Country": "country",
                                  "Area": "area",
                                  "Location": "location",
                                  "Activity": "activity",
                                  "Sex ": "gender",
                                  "Age": "age",
                                  "Injury": "injury",
                                  "Fatal (Y/N)": "fatal",})
        df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25723 entries, 0 to 25722
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   case_number  8702 non-null   object 
 1   date         6302 non-null   object 
 2   year         6300 non-null   float64
 3   type         6298 non-null   object 
 4   country      6252 non-null   object 
 5   area         5847 non-null   object 
 6   location     5762 non-null   object 
 7   activity     5758 non-null   object 
 8   gender       5737 non-null   object 
 9   age          3471 non-null   object 
 10  injury       6274 non-null   object 
 11  fatal        5763 non-null   object 
dtypes: float64(1), object(11)
memory usage: 2.4+ MB
```

#Rename Columns

Figure 1.3

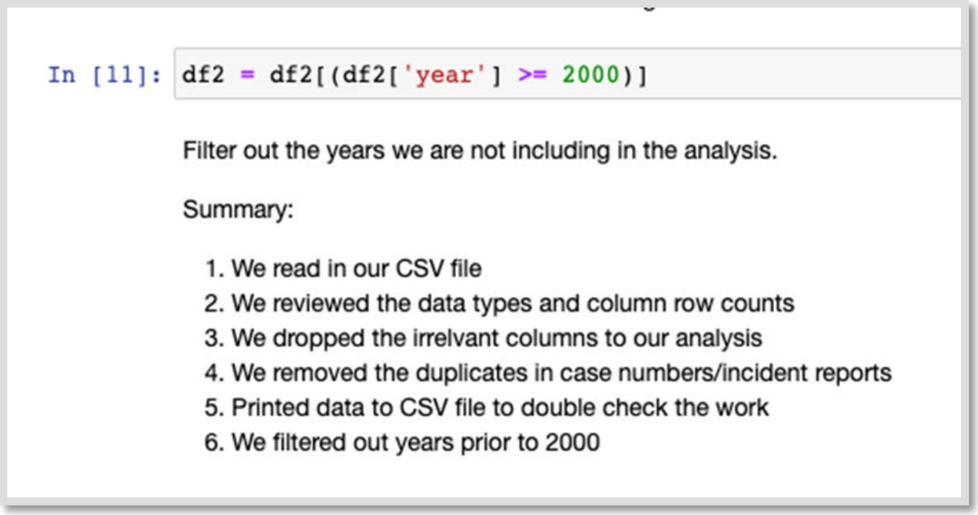


#Rename Columns

```
In [8]: df2.drop_duplicates(subset="case_number", inplace=True)
```

Remove duplicate case numbers/incident reports

Figure 1.4



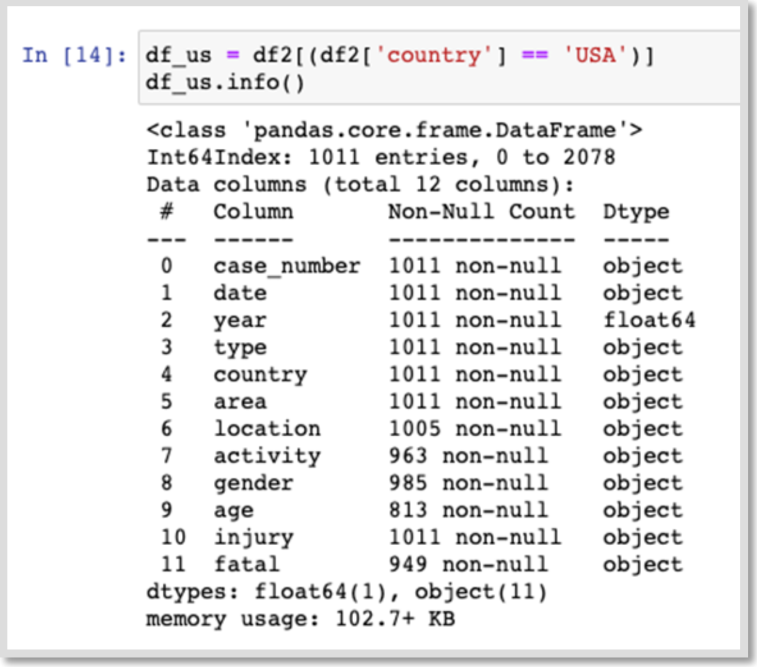
```
In [11]: df2 = df2[(df2['year'] >= 2000)]
```

Filter out the years we are not including in the analysis.

Summary:

1. We read in our CSV file
2. We reviewed the data types and column row counts
3. We dropped the irrelvant columns to our analysis
4. We removed the duplicates in case numbers/incident reports
5. Printed data to CSV file to double check the work
6. We filtered out years prior to 2000

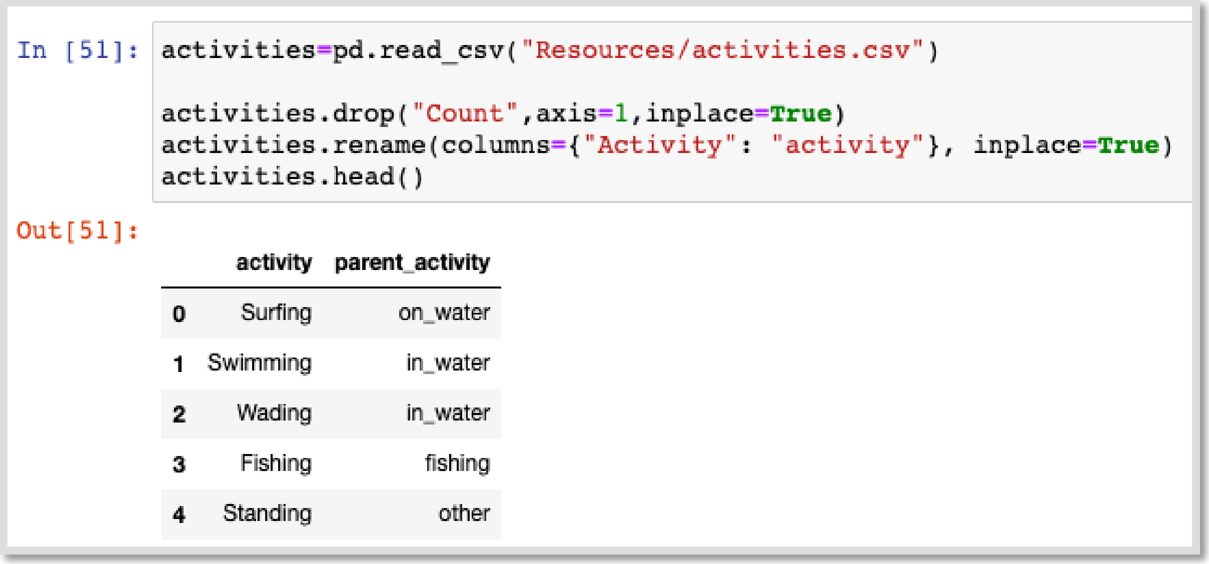
Figure 1.5



```
In [14]: df_us = df2[(df2['country'] == 'USA')]
         df_us.info()

         <class 'pandas.core.frame.DataFrame'>
         Int64Index: 1011 entries, 0 to 2078
         Data columns (total 12 columns):
          #   Column       Non-Null Count  Dtype
         ---  ------       --------------  -----
          0   case_number  1011 non-null   object
          1   date         1011 non-null   object
          2   year         1011 non-null   float64
          3   type         1011 non-null   object
          4   country      1011 non-null   object
          5   area         1011 non-null   object
          6   location     1005 non-null   object
          7   activity     963 non-null    object
          8   gender       985 non-null    object
          9   age          813 non-null    object
          10  injury       1011 non-null   object
          11  fatal        949 non-null    object
         dtypes: float64(1), object(11)
         memory usage: 102.7+ KB
```

Figure 1.6



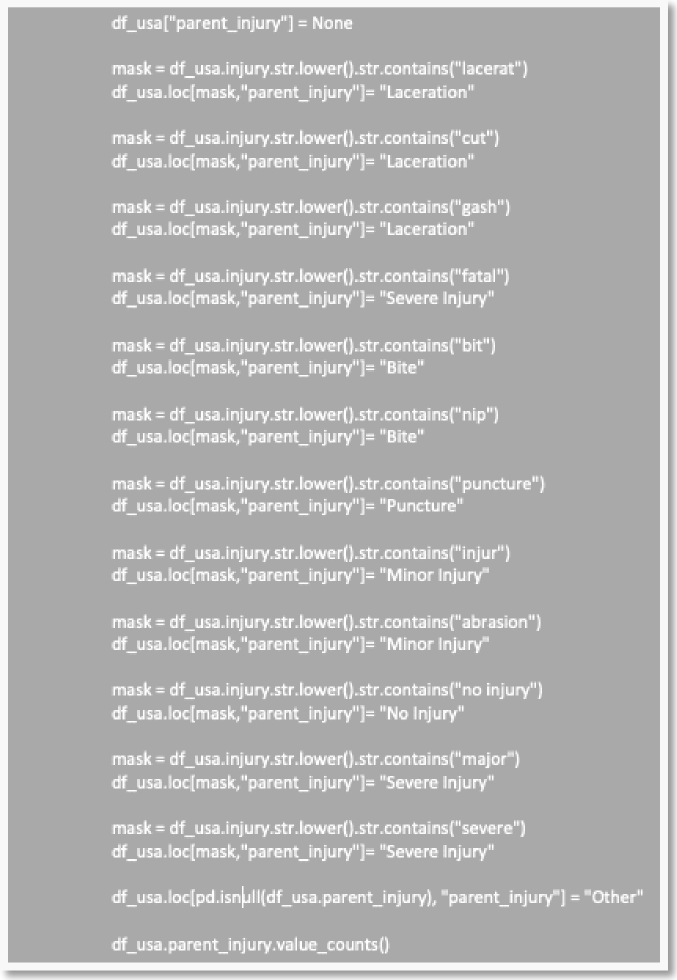
```
In [51]: activities=pd.read_csv("Resources/activities.csv")

         activities.drop("Count",axis=1,inplace=True)
         activities.rename(columns={"Activity": "activity"}, inplace=True)
         activities.head()

Out[51]:
```

	activity	parent_activity
0	Surfing	on_water
1	Swimming	in_water
2	Wading	in_water
3	Fishing	fishing
4	Standing	other

Figure 1.7



```
df_usa["parent_injury"] = None

mask = df_usa.injury.str.lower().str.contains("lacerat")
df_usa.loc[mask,"parent_injury"]= "Laceration"

mask = df_usa.injury.str.lower().str.contains("cut")
df_usa.loc[mask,"parent_injury"]= "Laceration"

mask = df_usa.injury.str.lower().str.contains("gash")
df_usa.loc[mask,"parent_injury"]= "Laceration"

mask = df_usa.injury.str.lower().str.contains("fatal")
df_usa.loc[mask,"parent_injury"]= "Severe Injury"

mask = df_usa.injury.str.lower().str.contains("bit")
df_usa.loc[mask,"parent_injury"]= "Bite"

mask = df_usa.injury.str.lower().str.contains("nip")
df_usa.loc[mask,"parent_injury"]= "Bite"

mask = df_usa.injury.str.lower().str.contains("puncture")
df_usa.loc[mask,"parent_injury"]= "Puncture"

mask = df_usa.injury.str.lower().str.contains("injur")
df_usa.loc[mask,"parent_injury"]= "Minor Injury"

mask = df_usa.injury.str.lower().str.contains("abrasion")
df_usa.loc[mask,"parent_injury"]= "Minor Injury"

mask = df_usa.injury.str.lower().str.contains("no injury")
df_usa.loc[mask,"parent_injury"]= "No Injury"

mask = df_usa.injury.str.lower().str.contains("major")
df_usa.loc[mask,"parent_injury"]= "Severe Injury"

mask = df_usa.injury.str.lower().str.contains("severe")
df_usa.loc[mask,"parent_injury"]= "Severe Injury"

df_usa.loc[pd.isnull(df_usa.parent_injury), "parent_injury"] = "Other"

df_usa.parent_injury.value_counts()
```

Figure 2.1

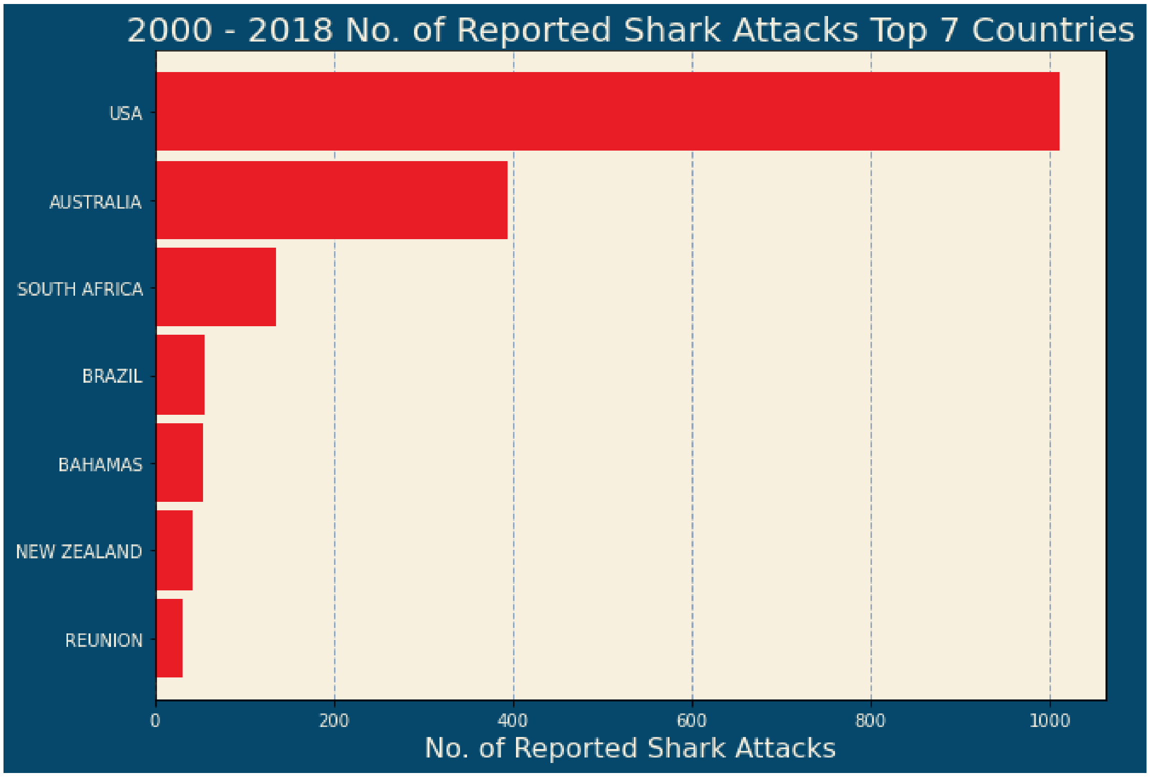


Figure 2.2

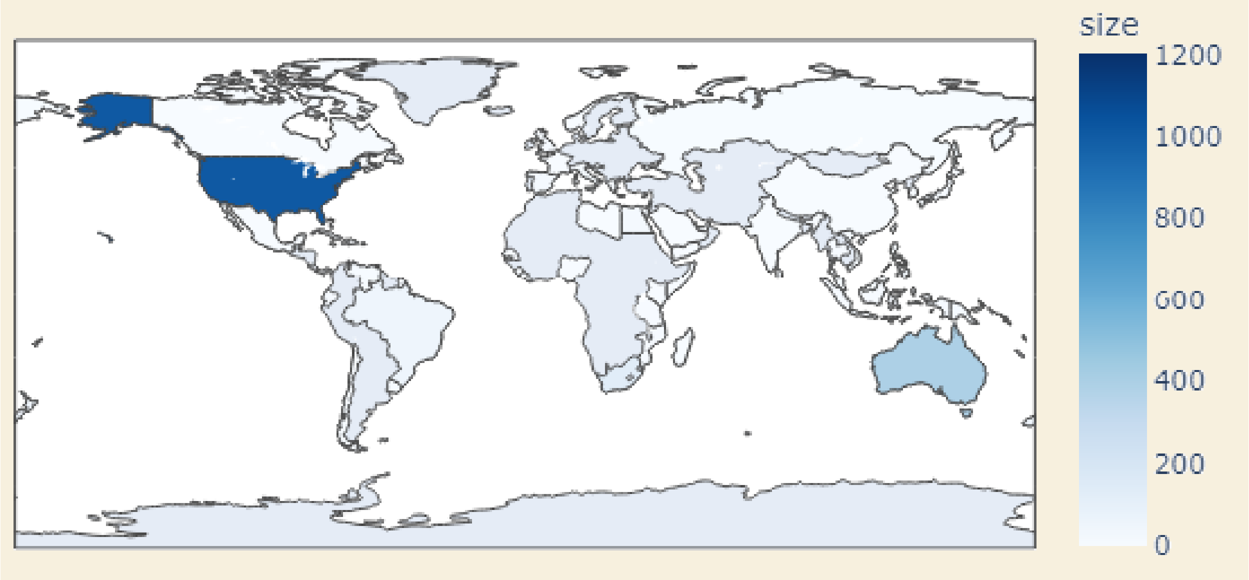


Figure 2.3

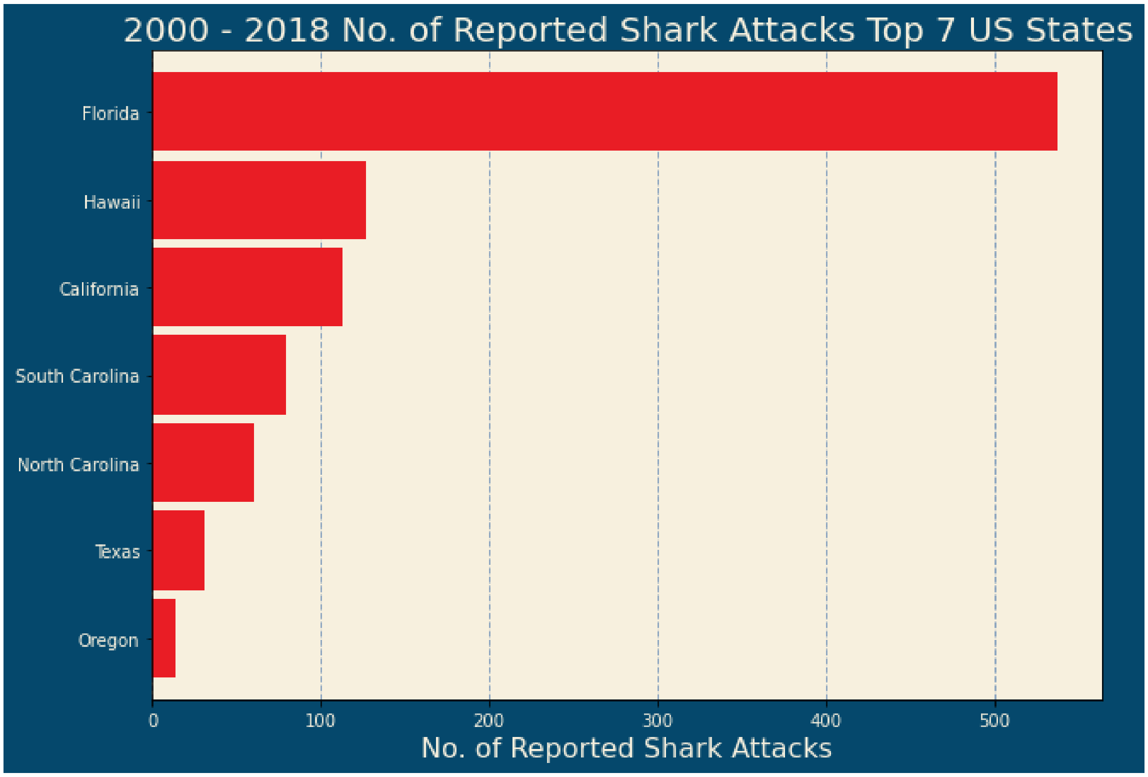


Figure 2.4

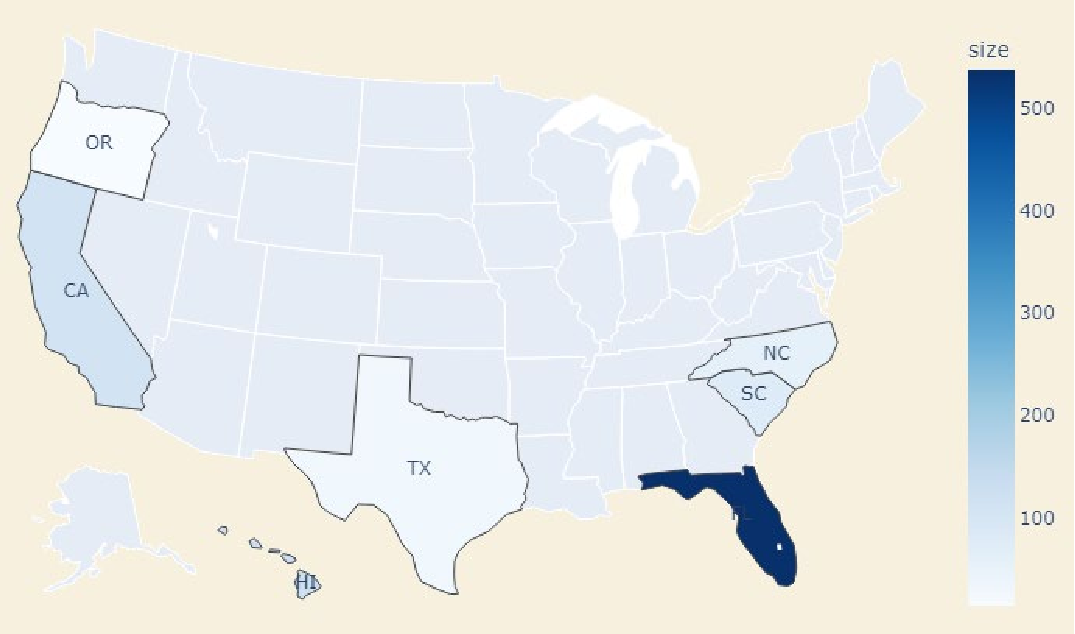


Figure 3.1

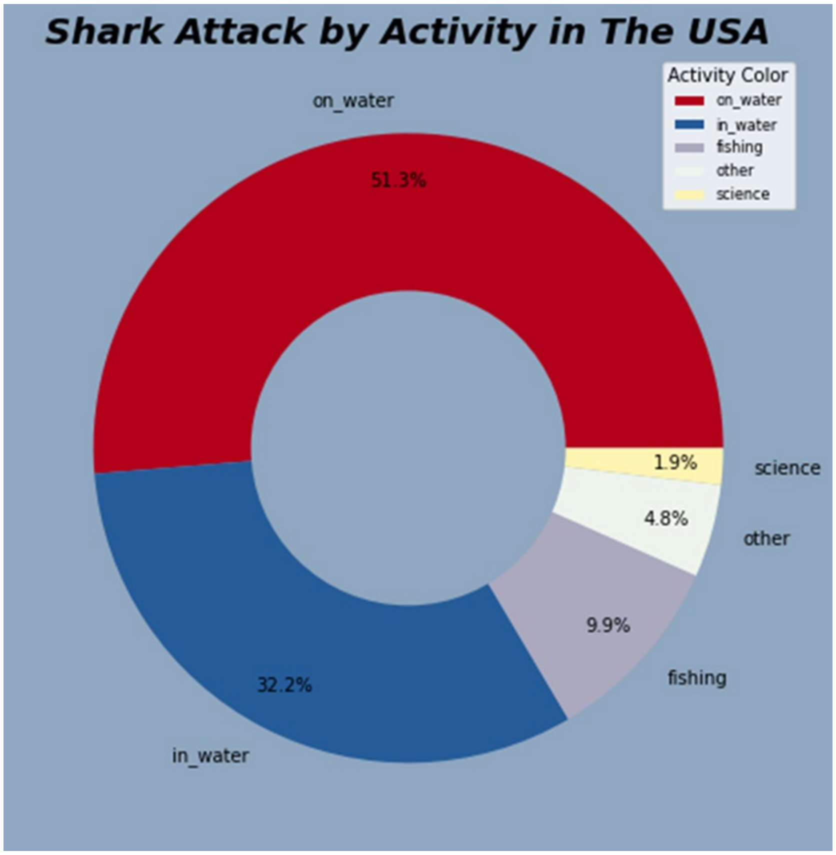


Figure 3.2

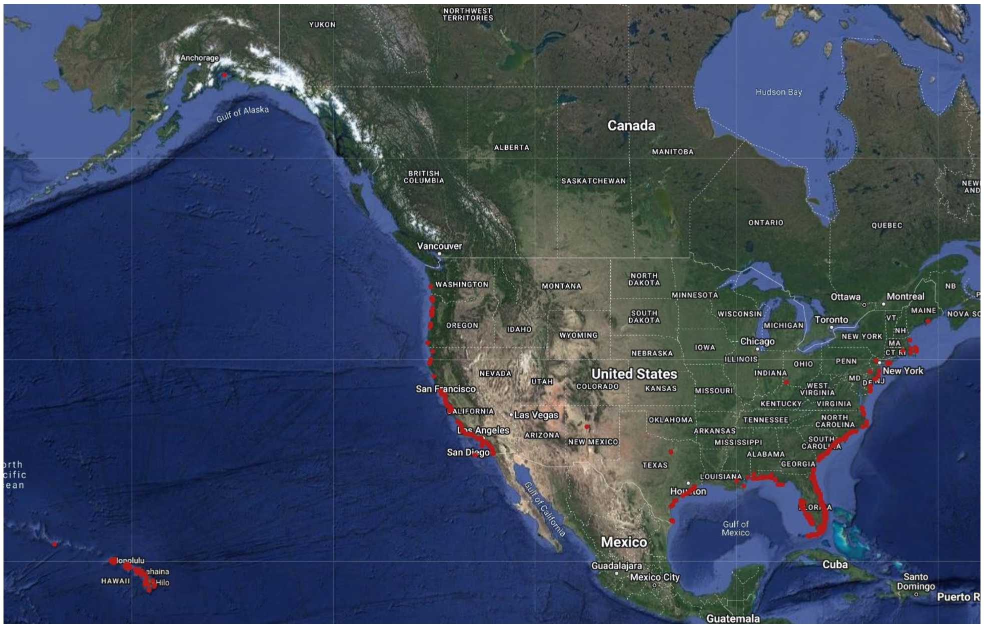


Figure 3.3

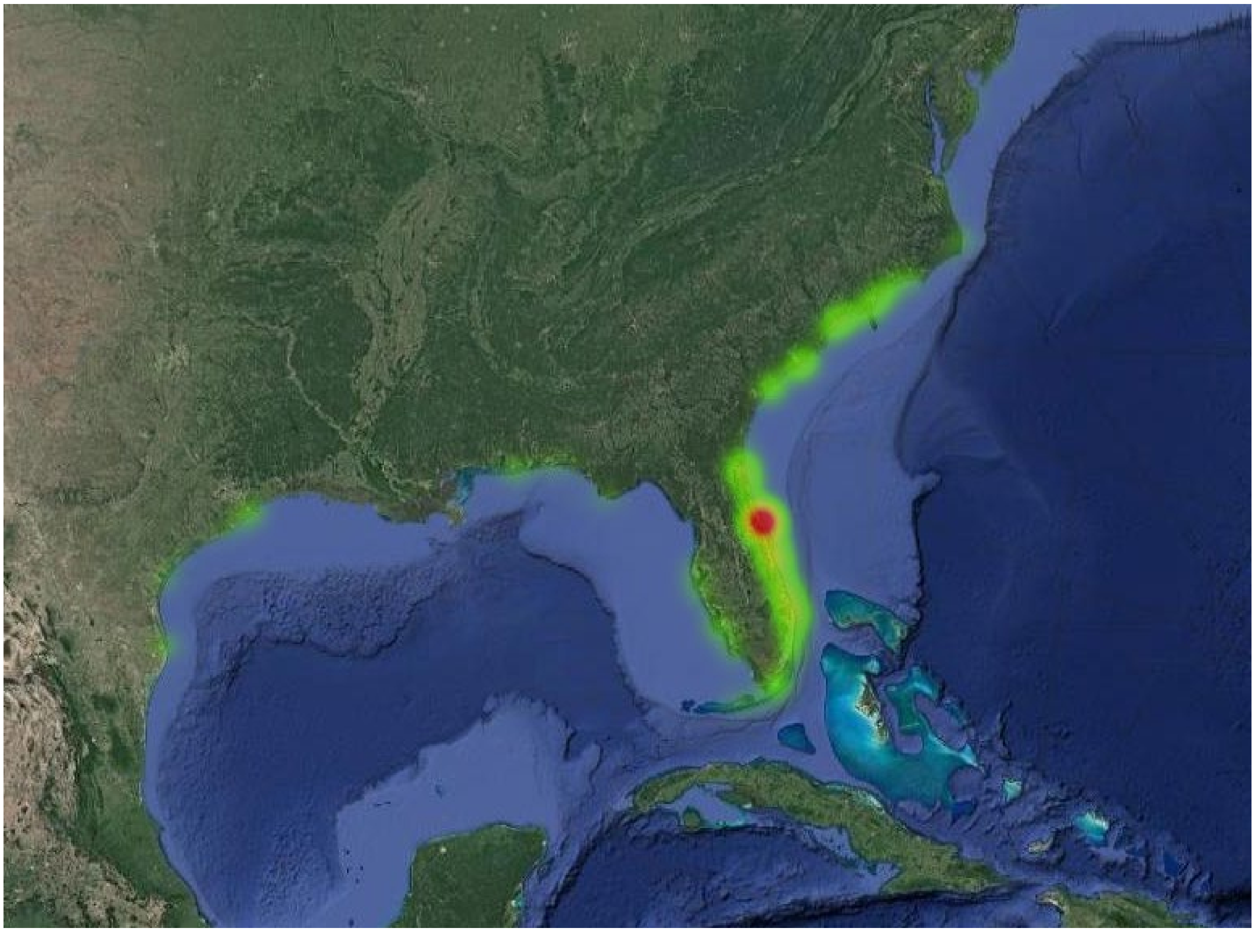


Figure 3.4

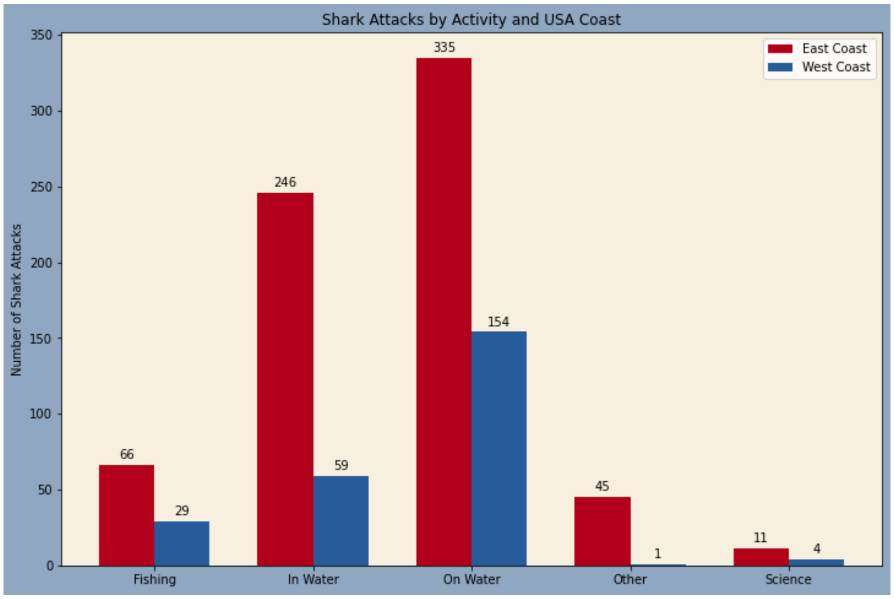


Figure 3.5

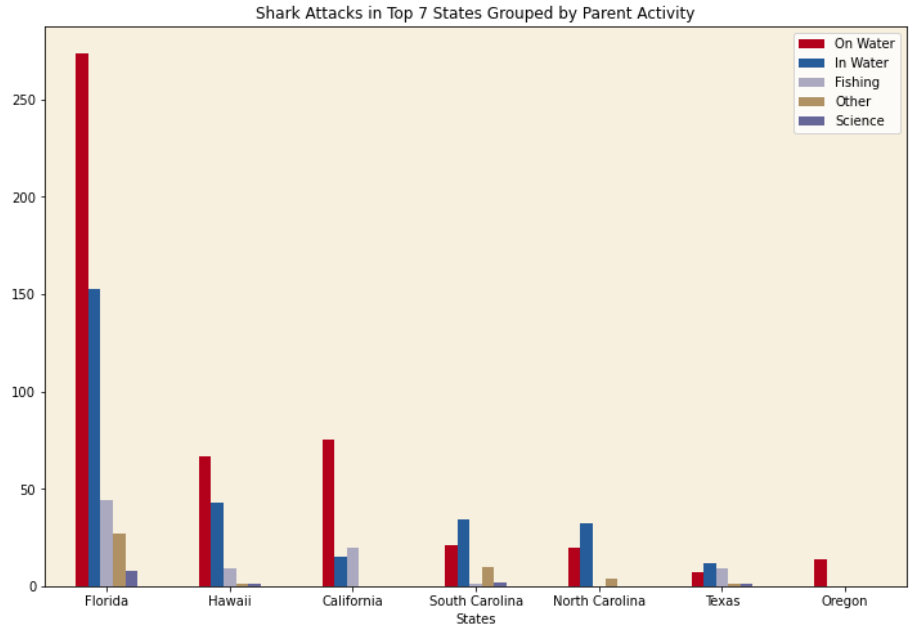


Figure 4.1

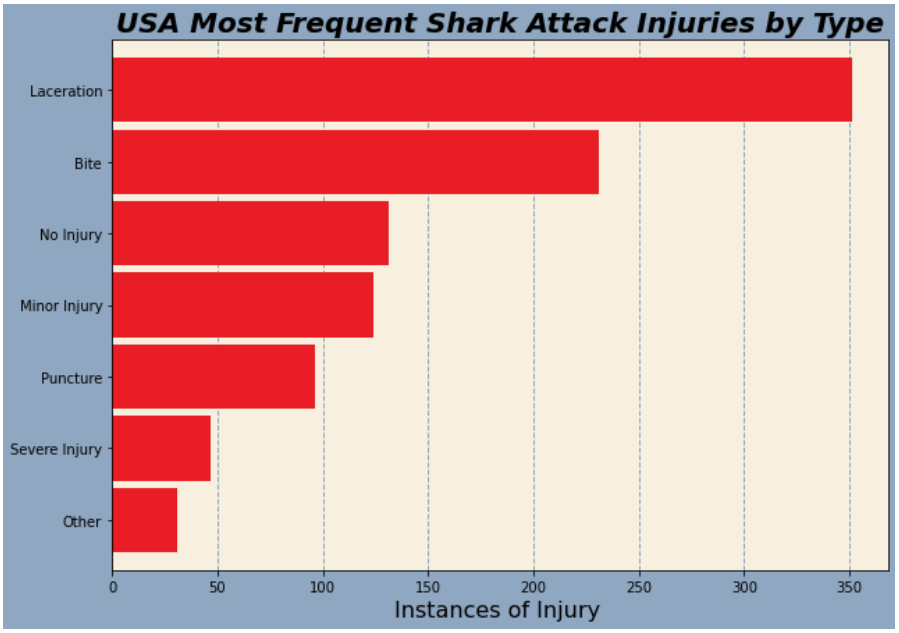


Figure 4.2

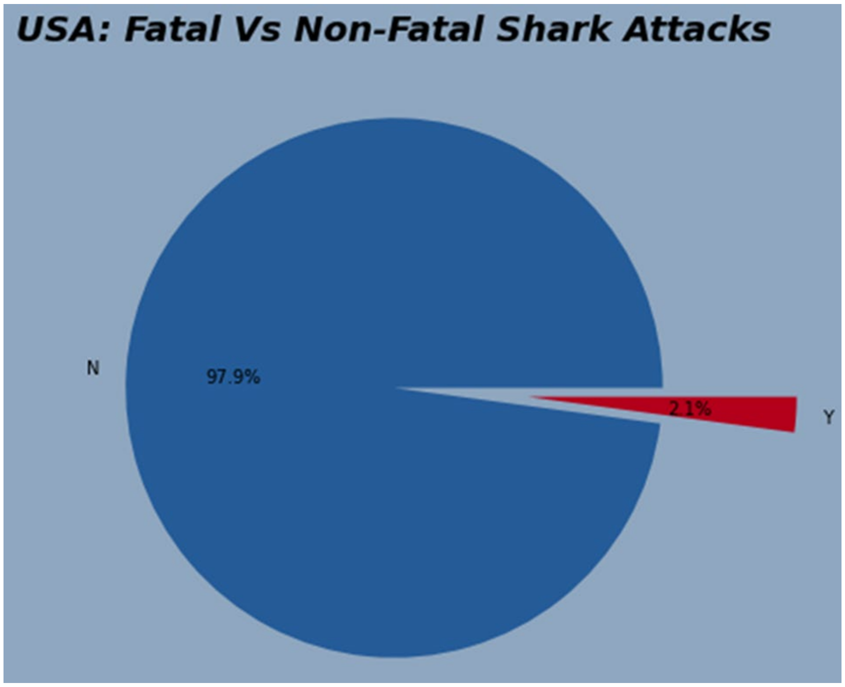


Figure 5.1

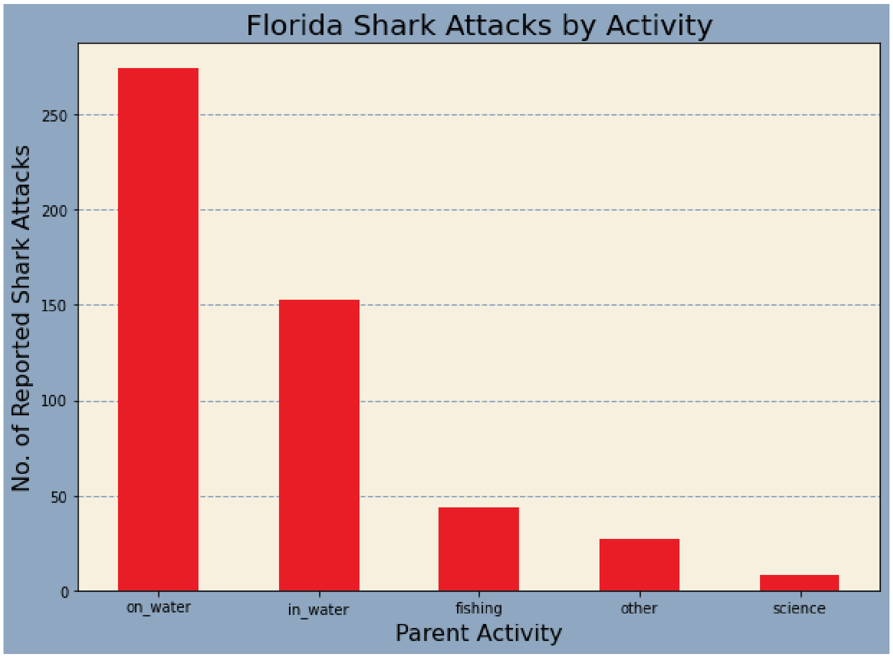


Figure 6.1

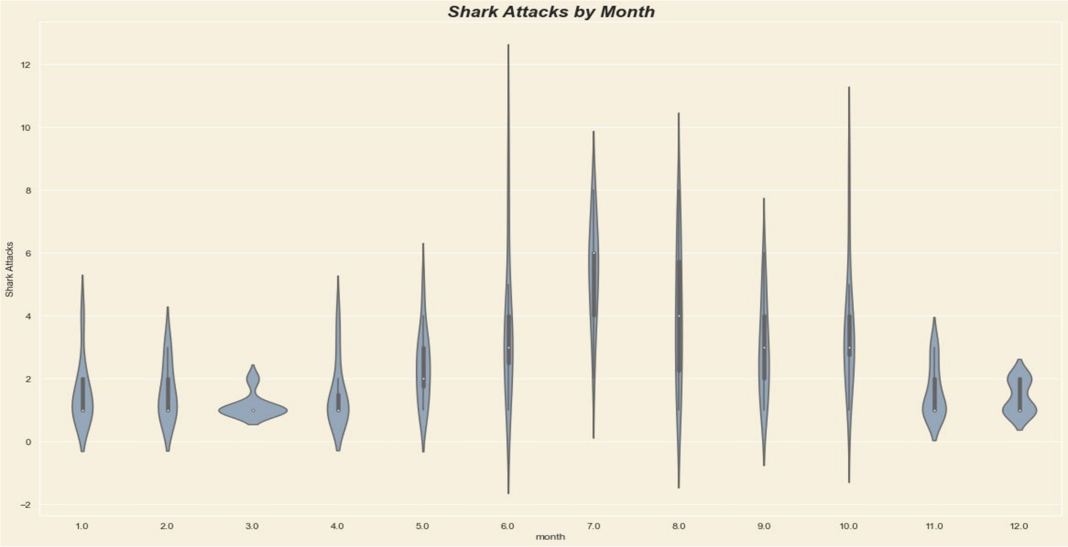


Figure 7.1

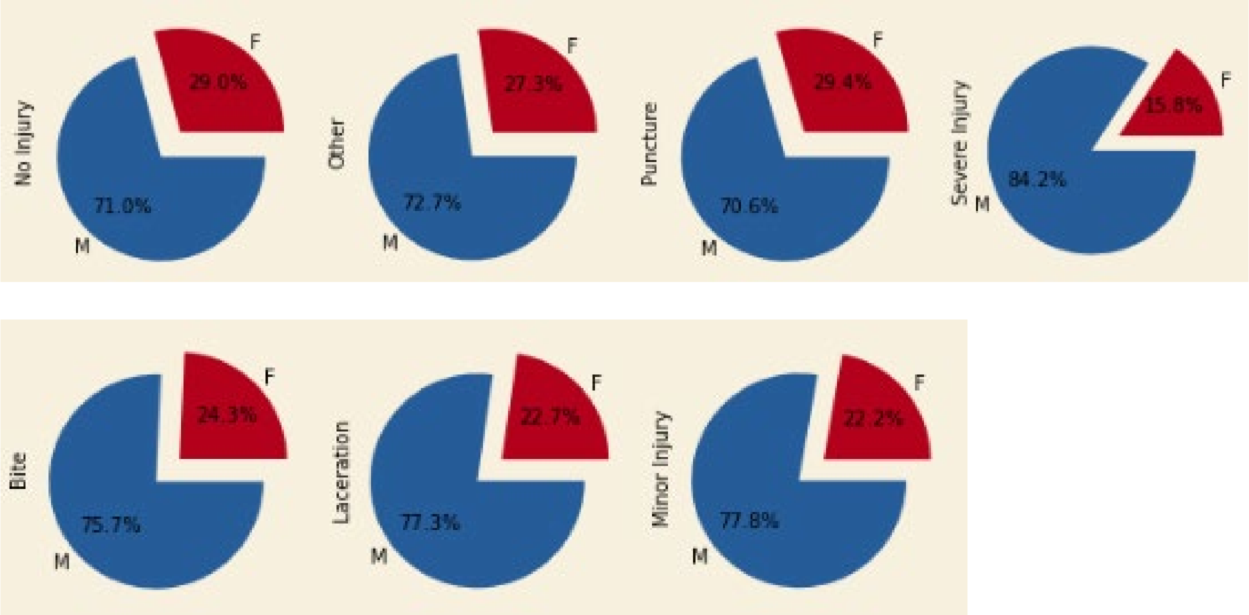


Figure 7.2

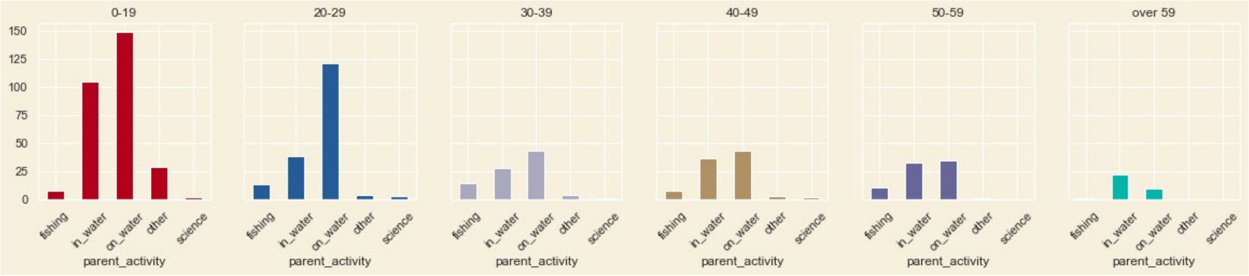


Figure 7.3

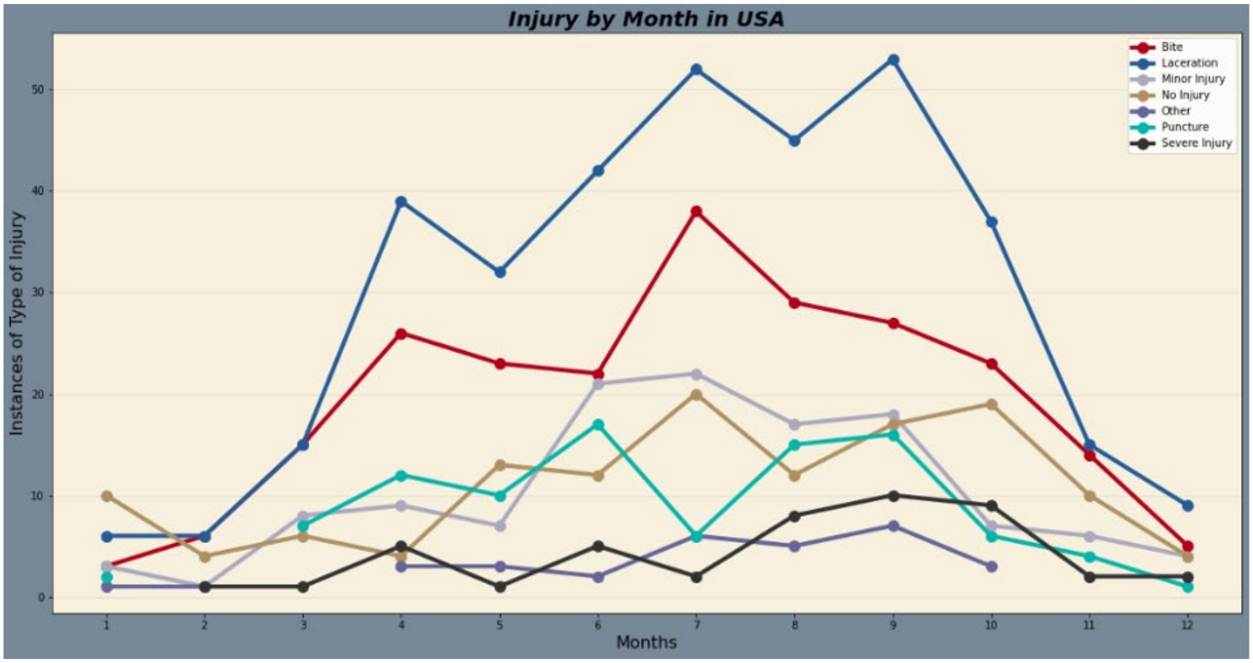


Figure 8.1

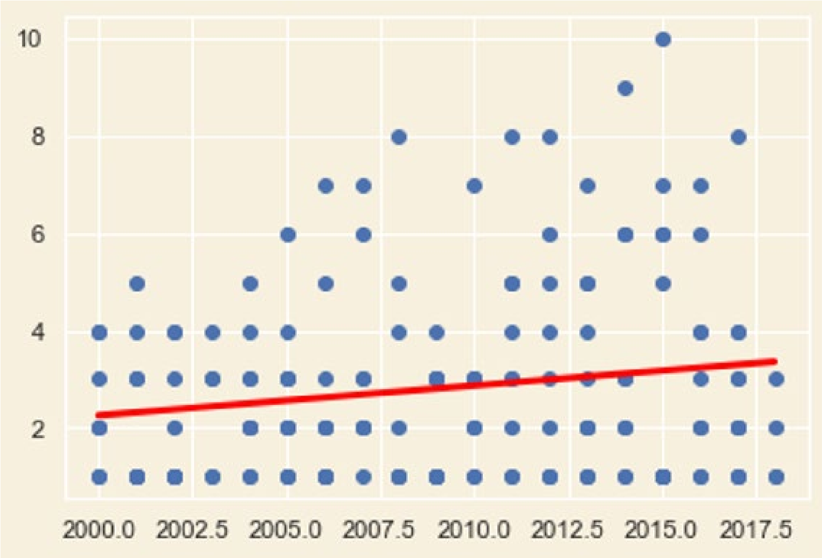
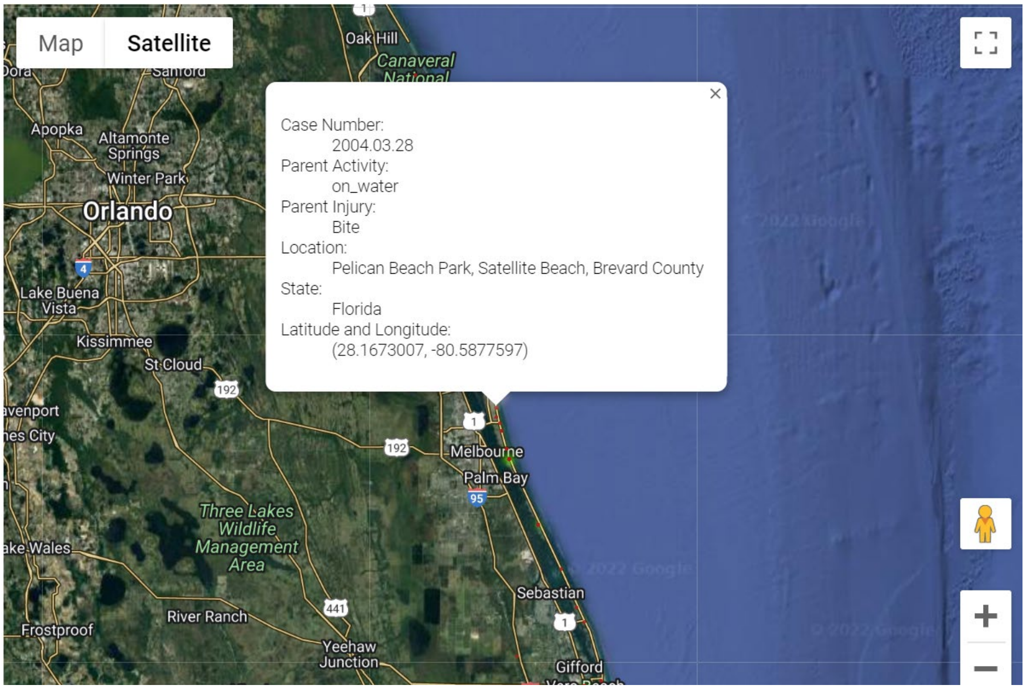


Figure 9.1



Case Number:
2004.03.28
Parent Activity:
on_water
Parent Injury:
Bite
Location:
Pelican Beach Park, Satellite Beach, Brevard County
State:
Florida
Latitude and Longitude:
(28.1673007, -80.5877597)