Shark Attacks

An Exploratory Data Analysis

PREPARED BY:

BRENNA WALLACE

OMAR ZAHER

JENNIFER WILKERSON

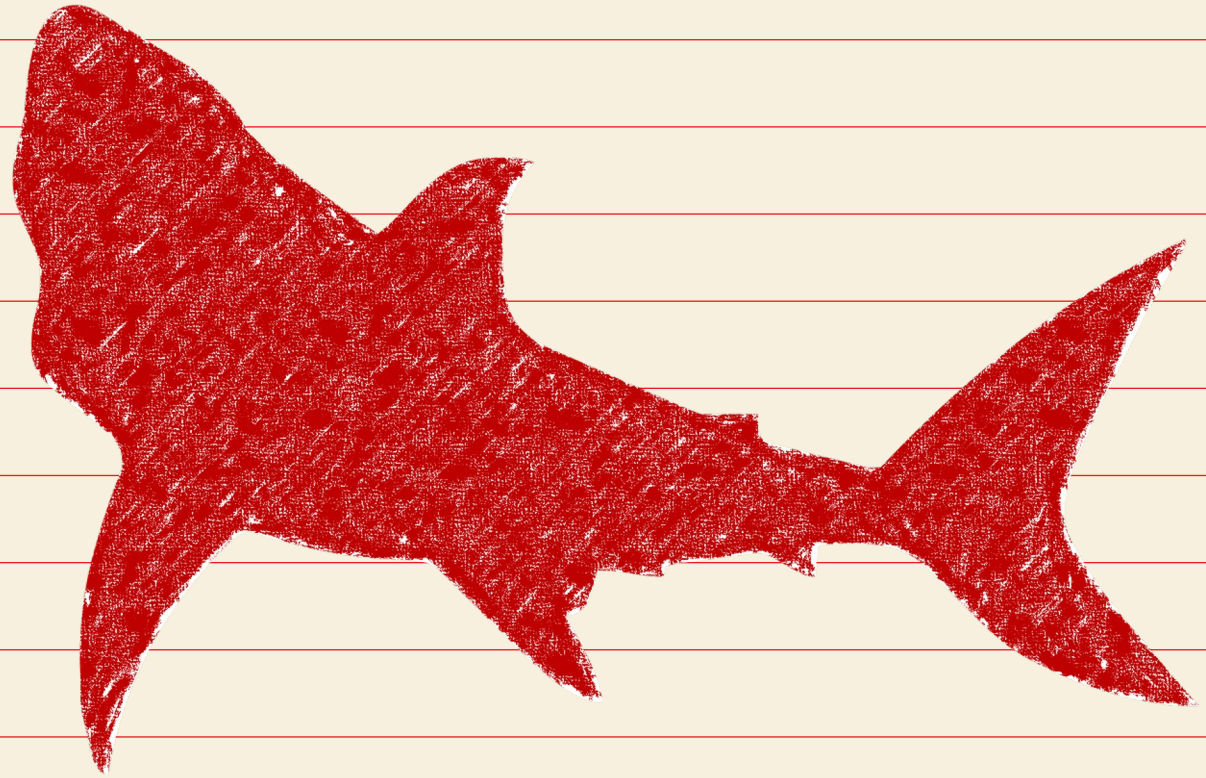
Introduction

Today we would like to talk about Shark attacks.

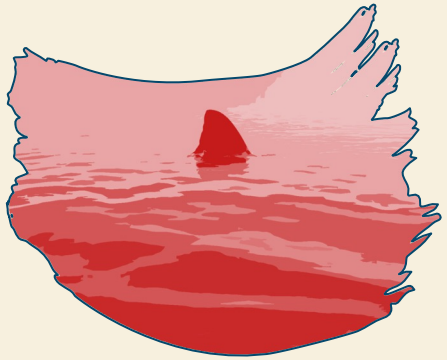
- Where do they occur?
- How dangerous is a shark attack?
- Does it necessarily mean it's the end if you are attacked?
- And what are most people doing whenever they are attacked by these underwater predators?

Most attacks are classified as provoked and unprovoked attacks. However, we will be taking a deeper "dive" into the data on these attacks to get a better understanding on; If we as a people, 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?

List of Contents

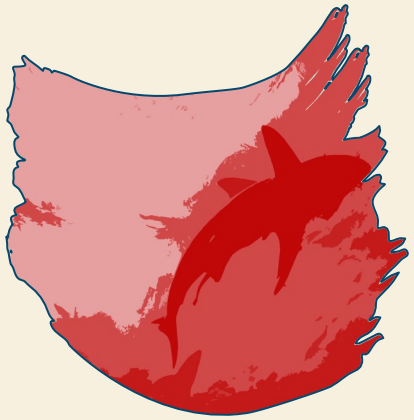


Research Questions

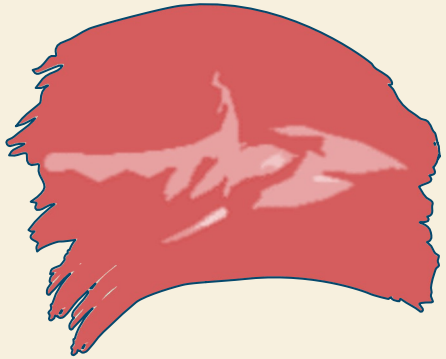


Where do reported shark attacks happen most in the world?

What reported activities are being done during a Shark attack?

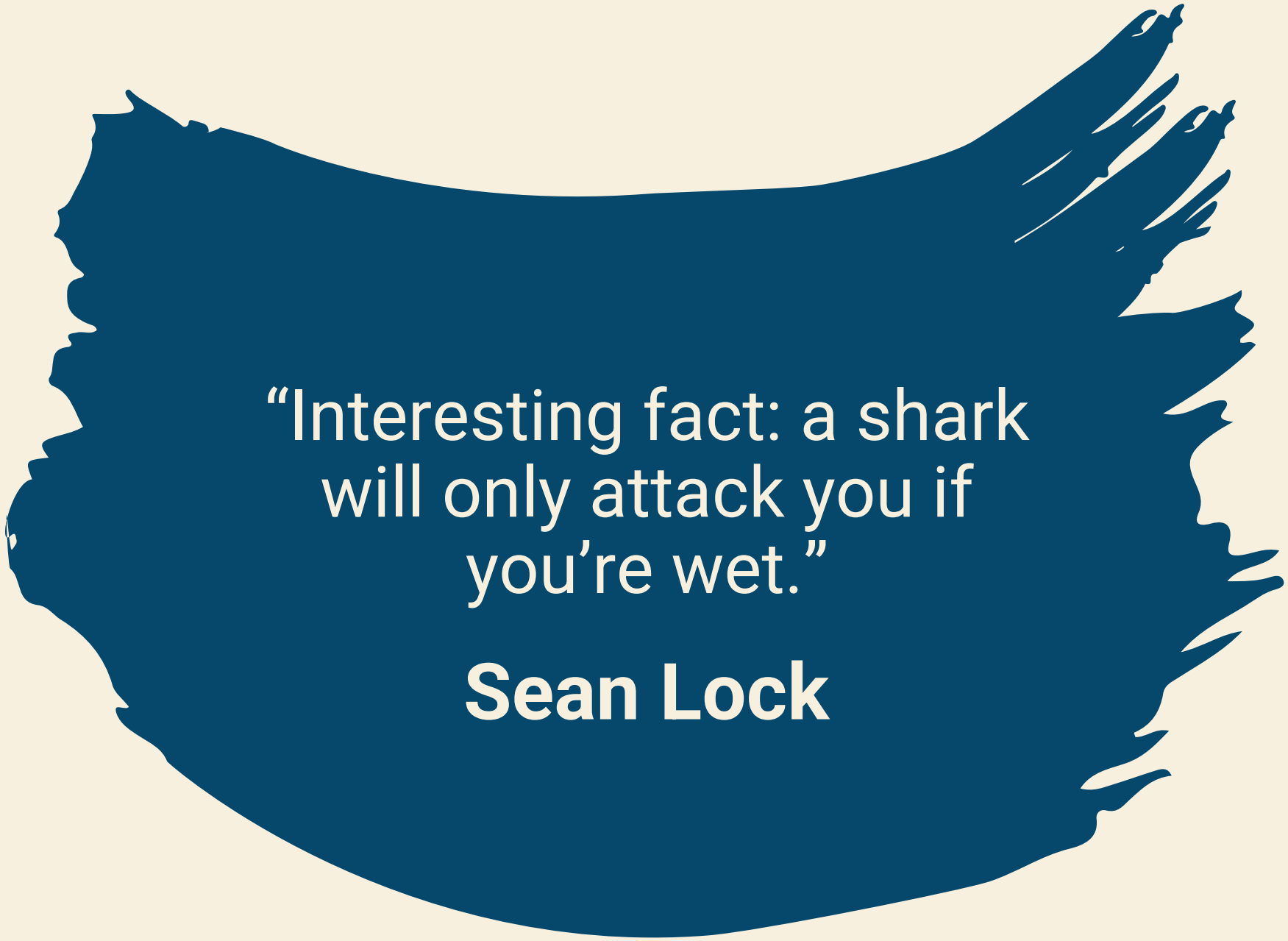


How severe was the attack?



What reported month of the year do shark attacks occur most often?

Hypothesis: As population increases over time, reported shark attacks will also increase.


“Interesting fact: a shark will only attack you if you’re wet.”
Sean Lock

Introduction to data

- The dataset we used was found on Kaggle: https://www.kaggle.com/datasets/mysarahmadbhat/shark-attacks?select=attacks.csv
- The data documents global shark attacks from 1900 to 2018. These reported incidents include categories such as the activity being done when the attack happened, the location of the attack, the date & time of the attack, and if the attack was fatal.
- One of our biggest tasks in this research project was to clean the data and group categories into data that could be evaluated

Data Cleaning

DROPPING UNUSED COLUMNS

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.

Data Cleaning

Continued

Rename Columns

```
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

Drop Duplicates

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

Remove duplicate case numbers/incident reports

Dropped Years Before 2000

```
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

Data Cleaning

Continued

Narrowed Down Data to USA

- The largest amount of shark attack data to work with is from the USA.
- We dropped countries other than the USA and had between 815 to 1011 rows of data to work with

Code

```
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
```

Data Cleaning

For sorting the various activities into Parent Activity groups

Parent Activity Explanation

We went through our USA data frame in excel and created a parent activity group column that contained: On Water, In Water, Fishing, Science, Other. Then we read in the new csv file and merged it into our USA dataframe.

Code

```
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

Data Cleaning

For sorting the various activities into Parent Activity groups

Parent Injury Explanation

We used string contains() masks to group injuries into parent categories.

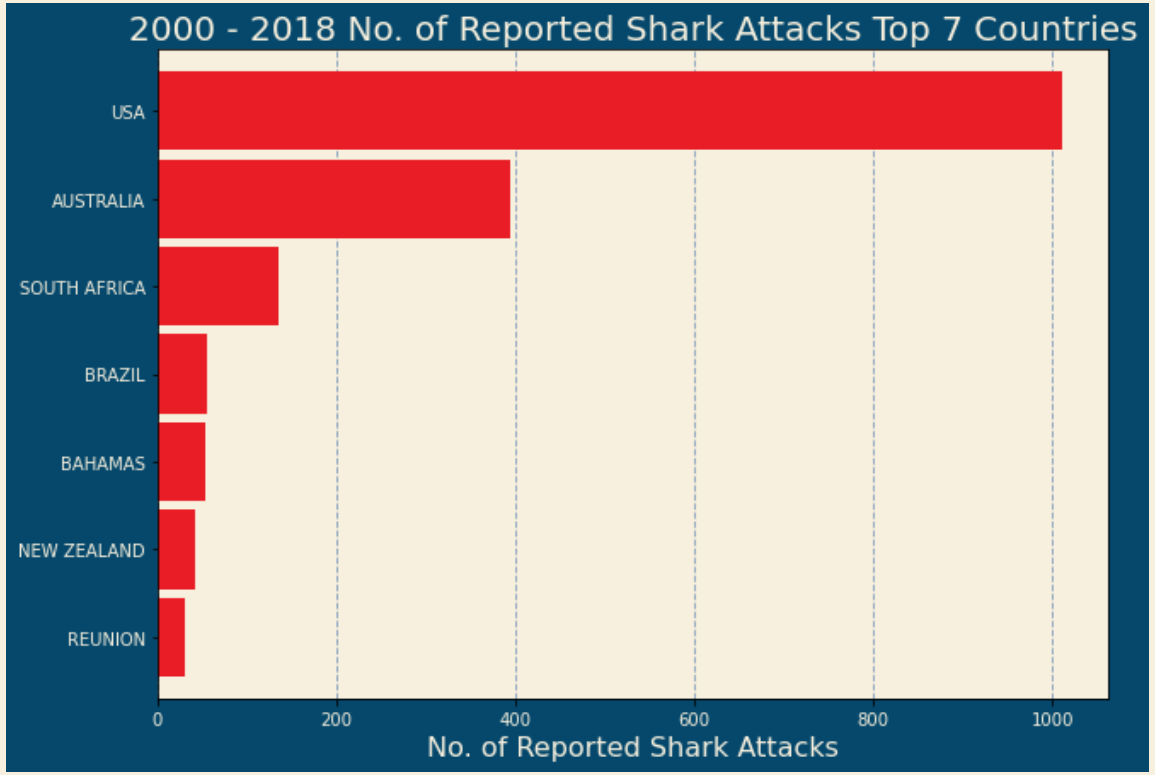
```
df_usa["parent_injury"] = None

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()
```

Where do reported shark attacks happen most in the world?

Reported Shark Attacks

Top 7 Countries

2000-2018

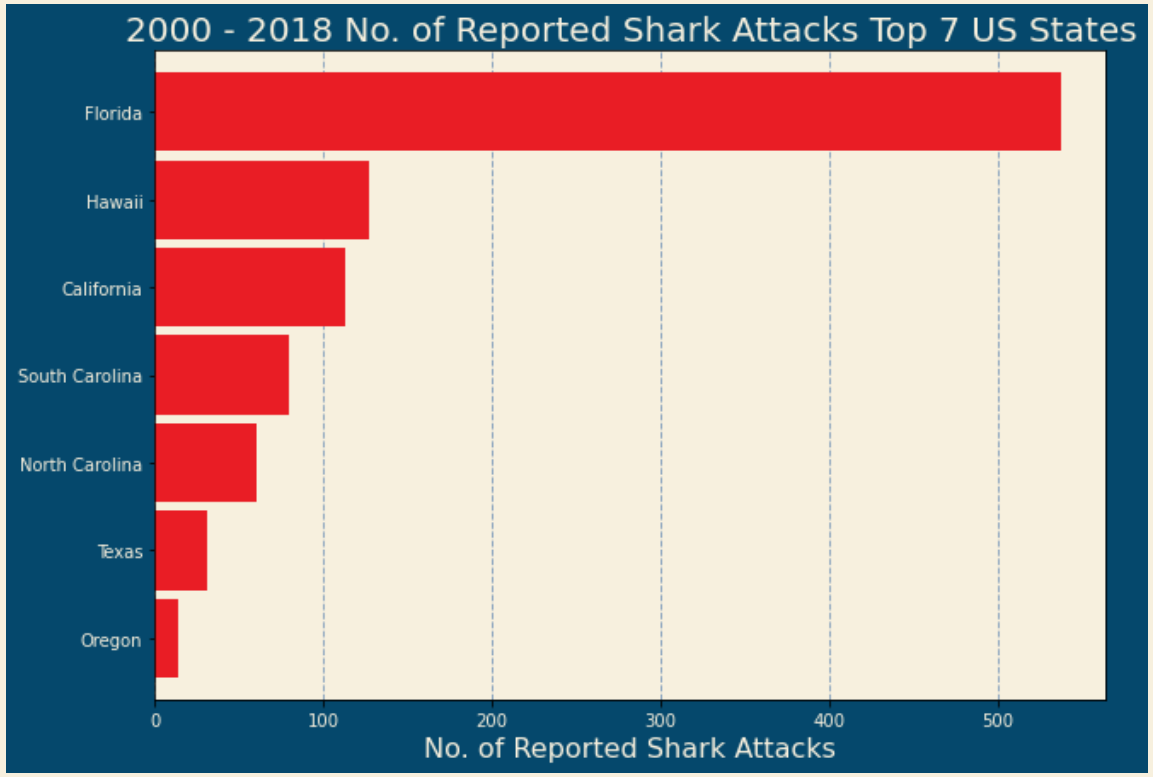


Shark Attacks are reported most often in the United States, followed by Australia and South Africa.

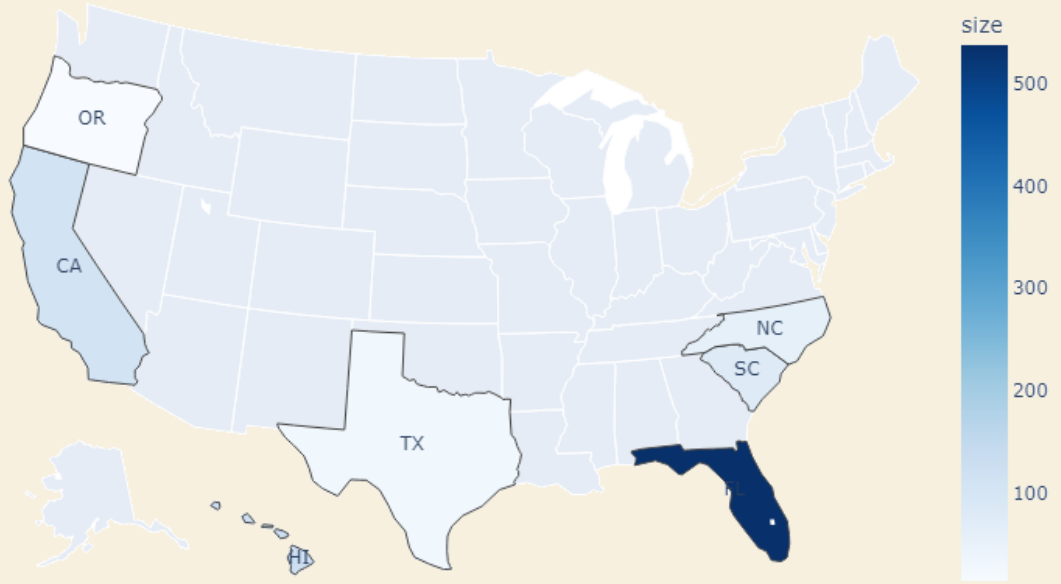
Reported Shark Attacks

Top 7 States in the USA

2000-2018

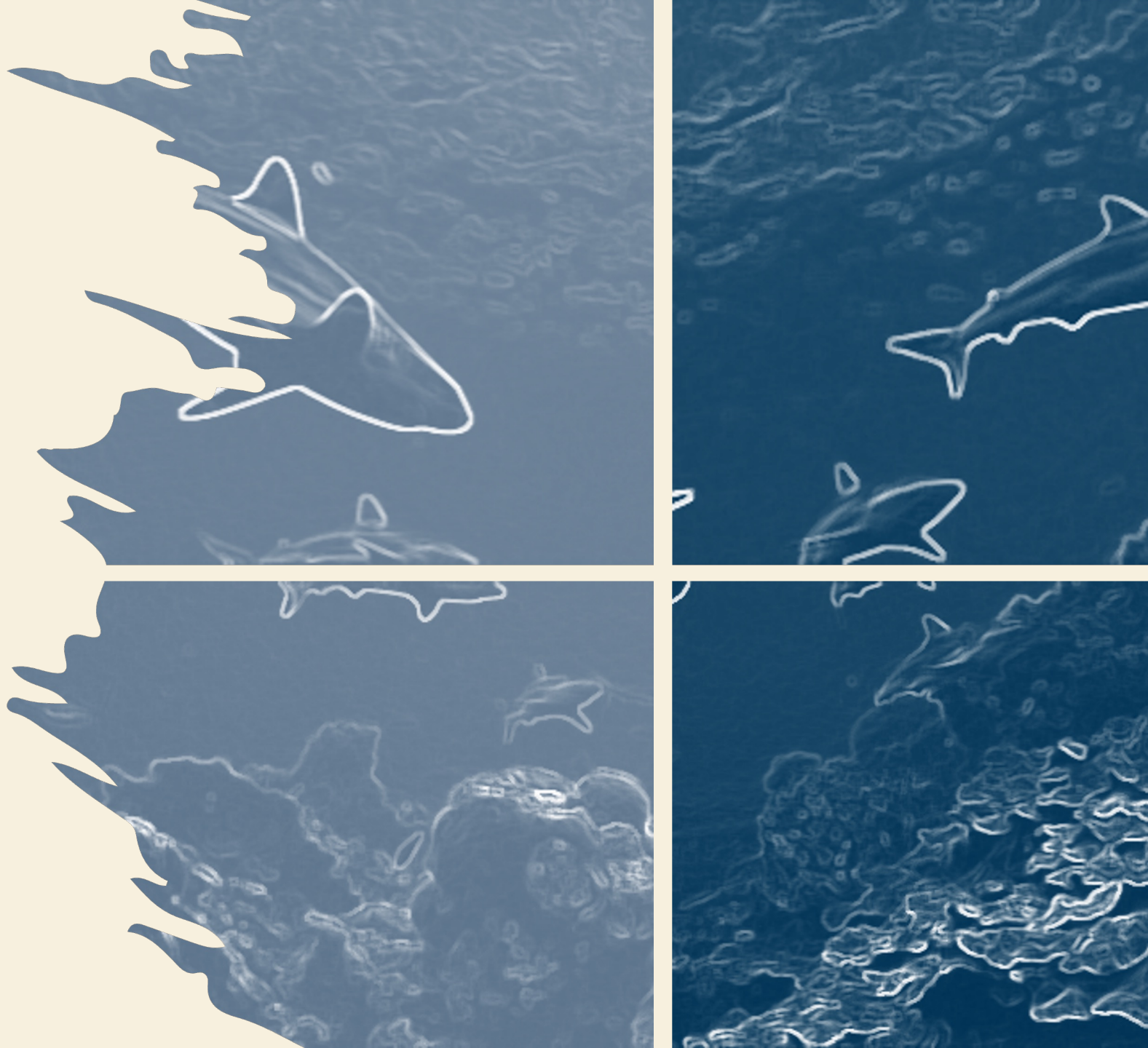


Looking at the United States only, Florida has more than 50% of the reported shark attacks, followed by Hawaii and California.



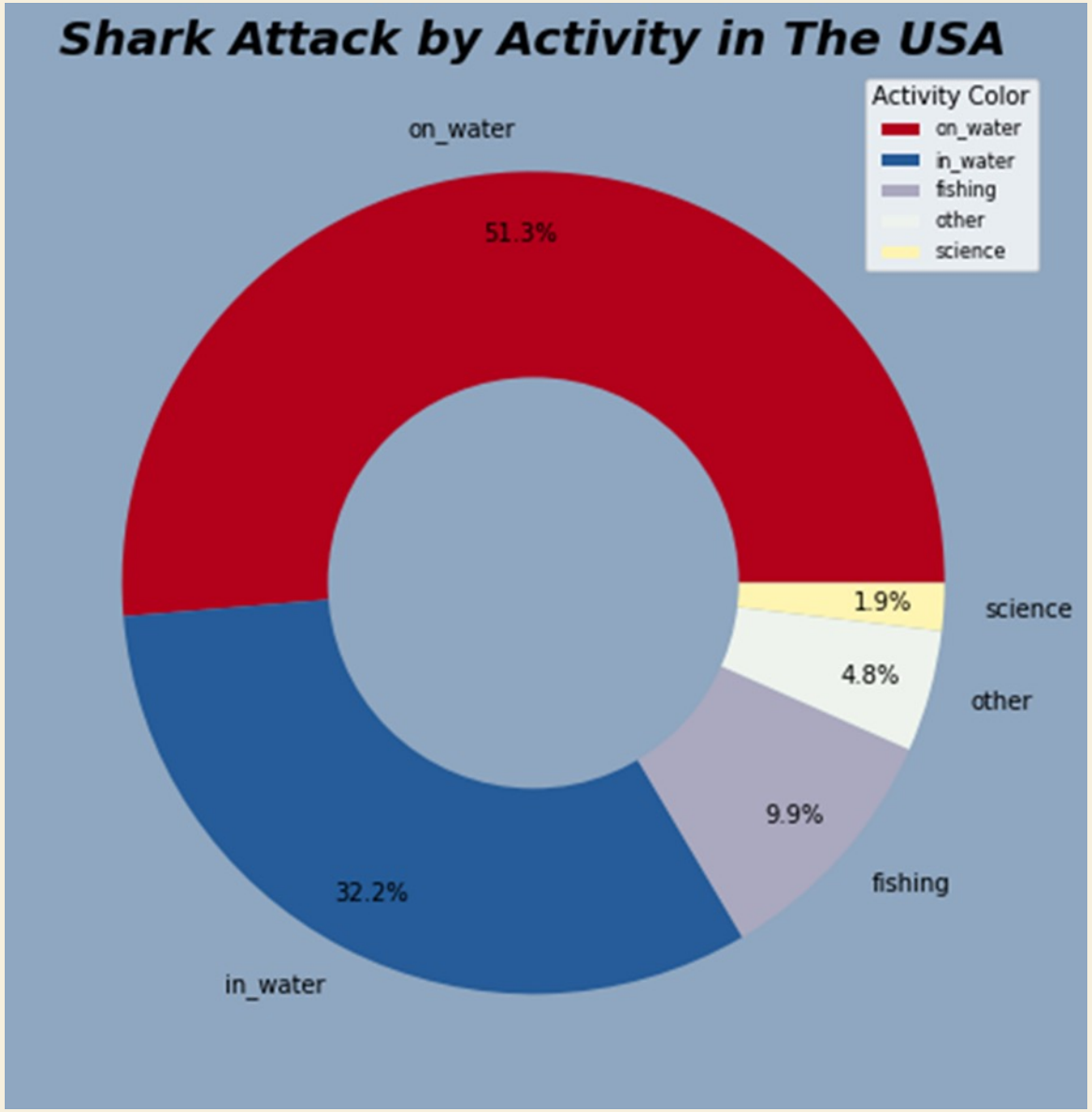
What reported activities are being done during a Shark attack?

Data is also explored by looking at age group, gender, and injury type



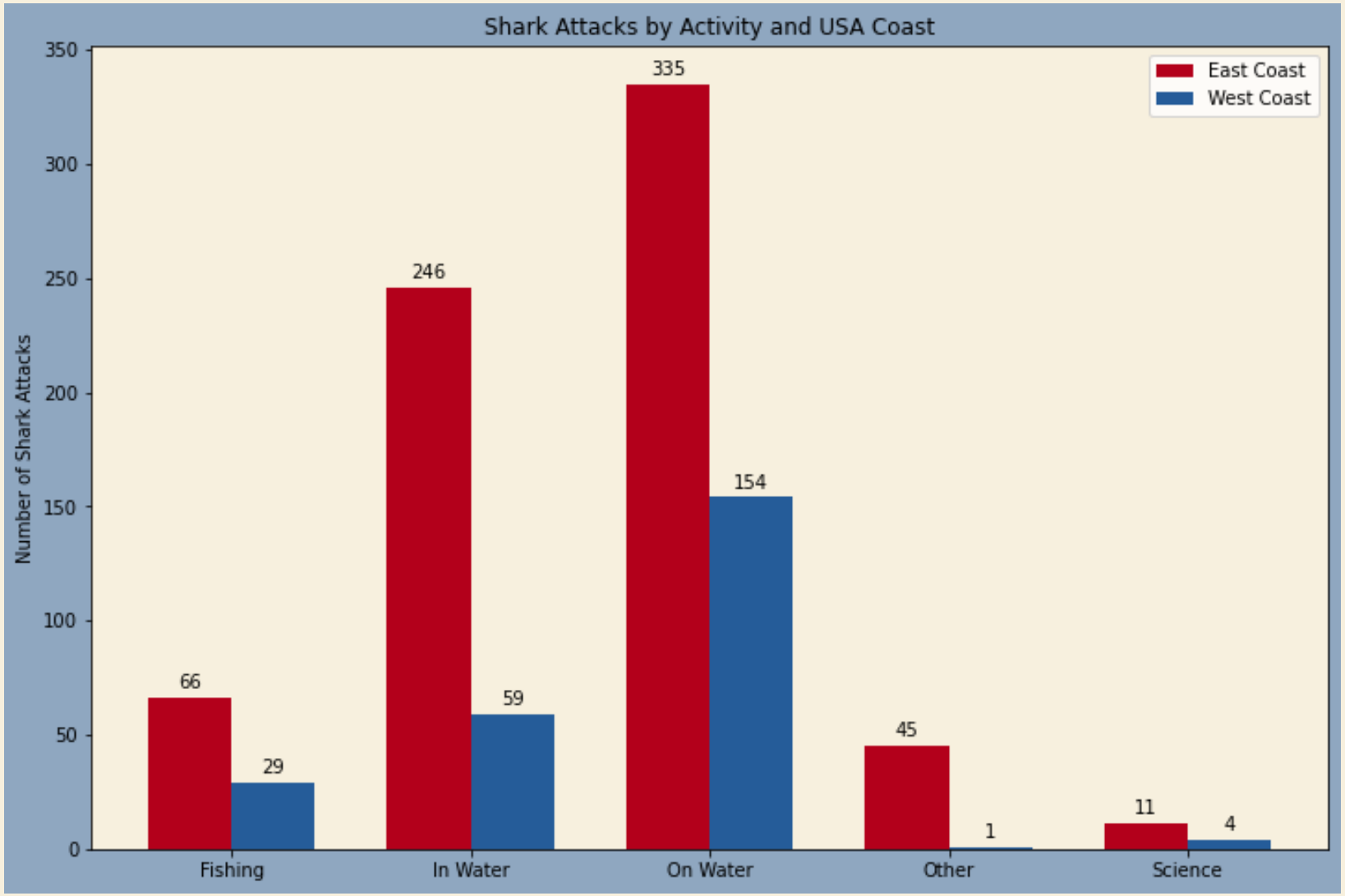
Shark Attacks by Activity

Shark Attacks in the USA



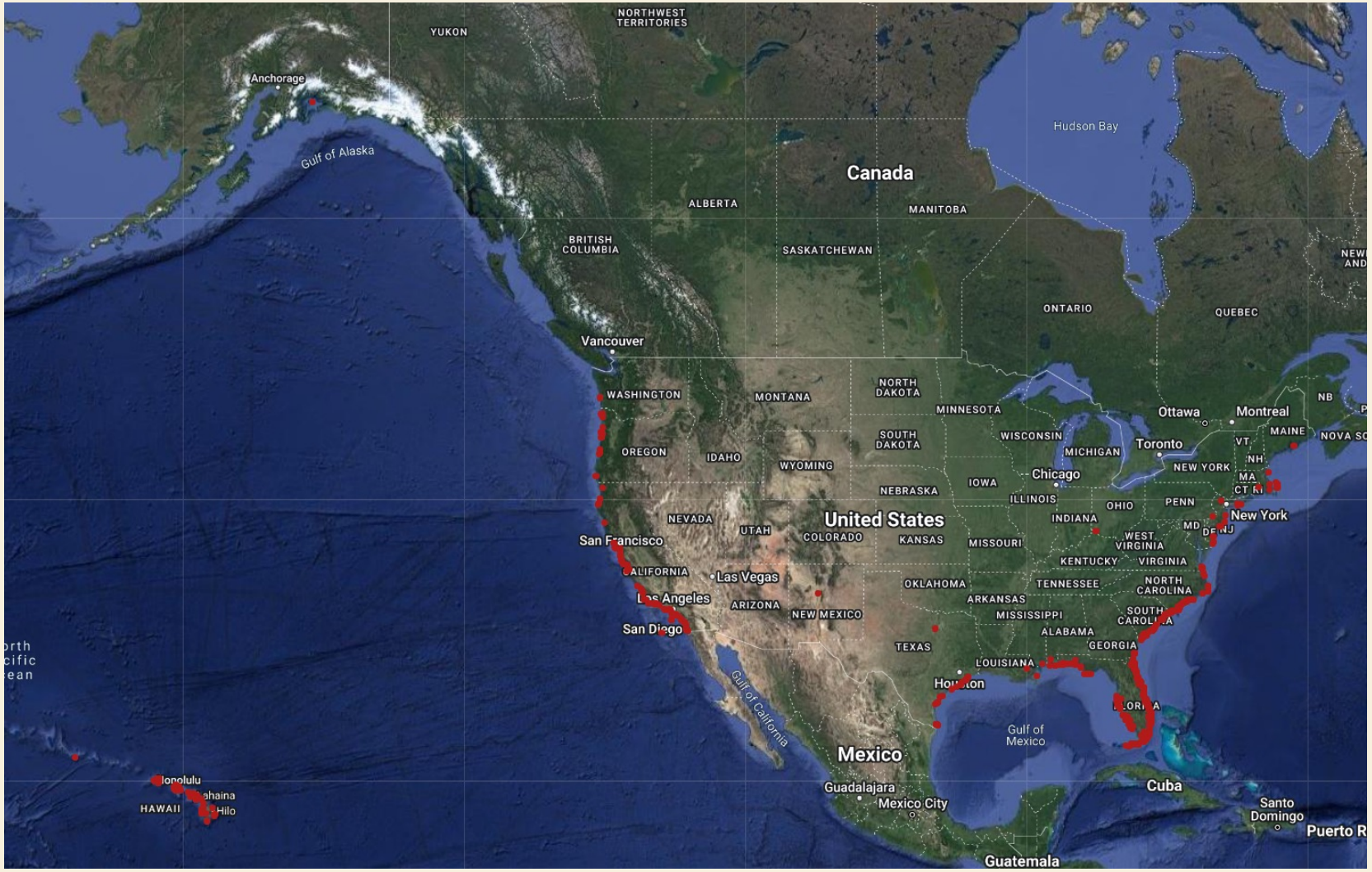
Shark Attacks by Activities on East Coast vs West Coast

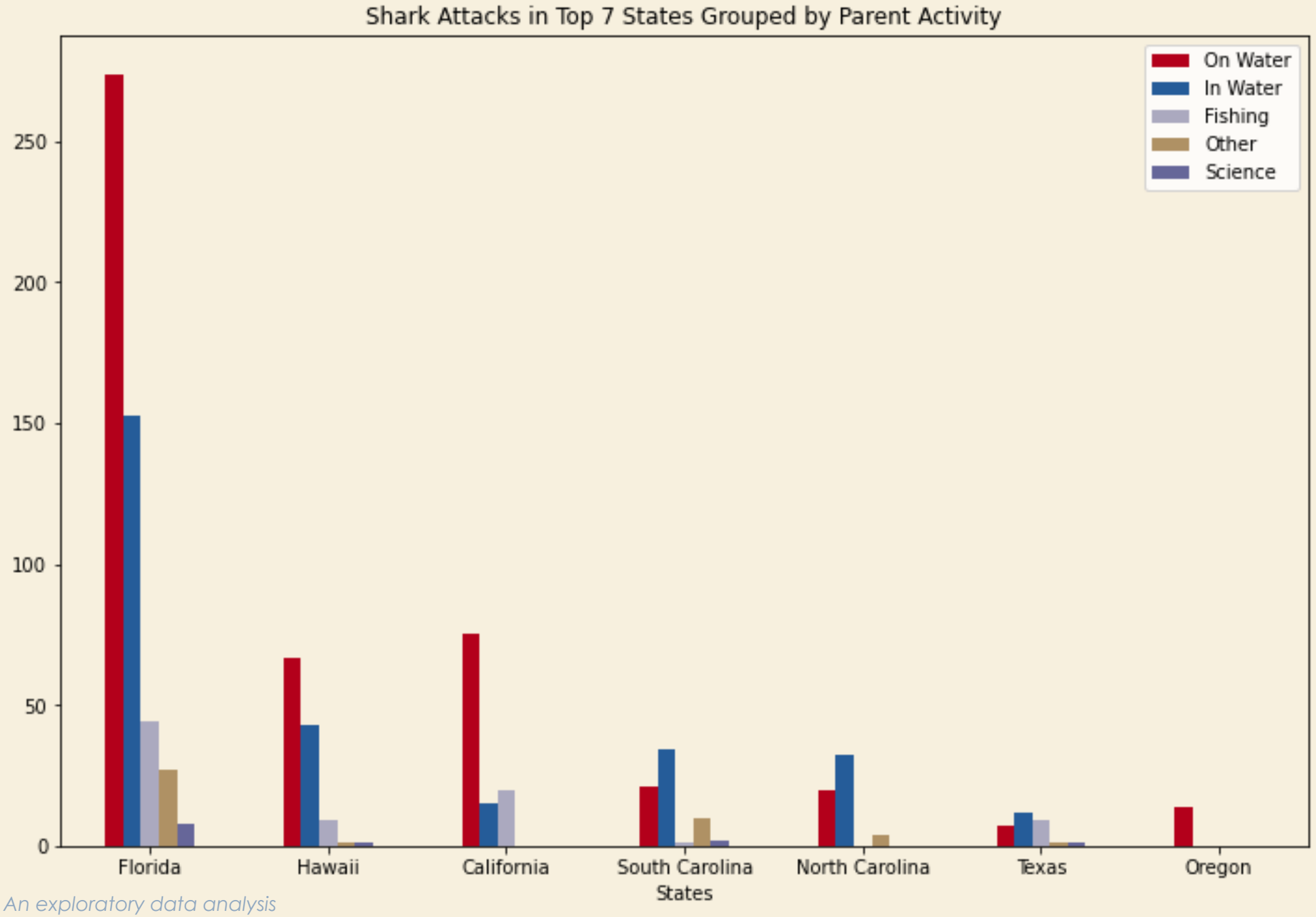
Shark Attacks in the USA



Reported Shark Attacks in the USA Map

2000-2018

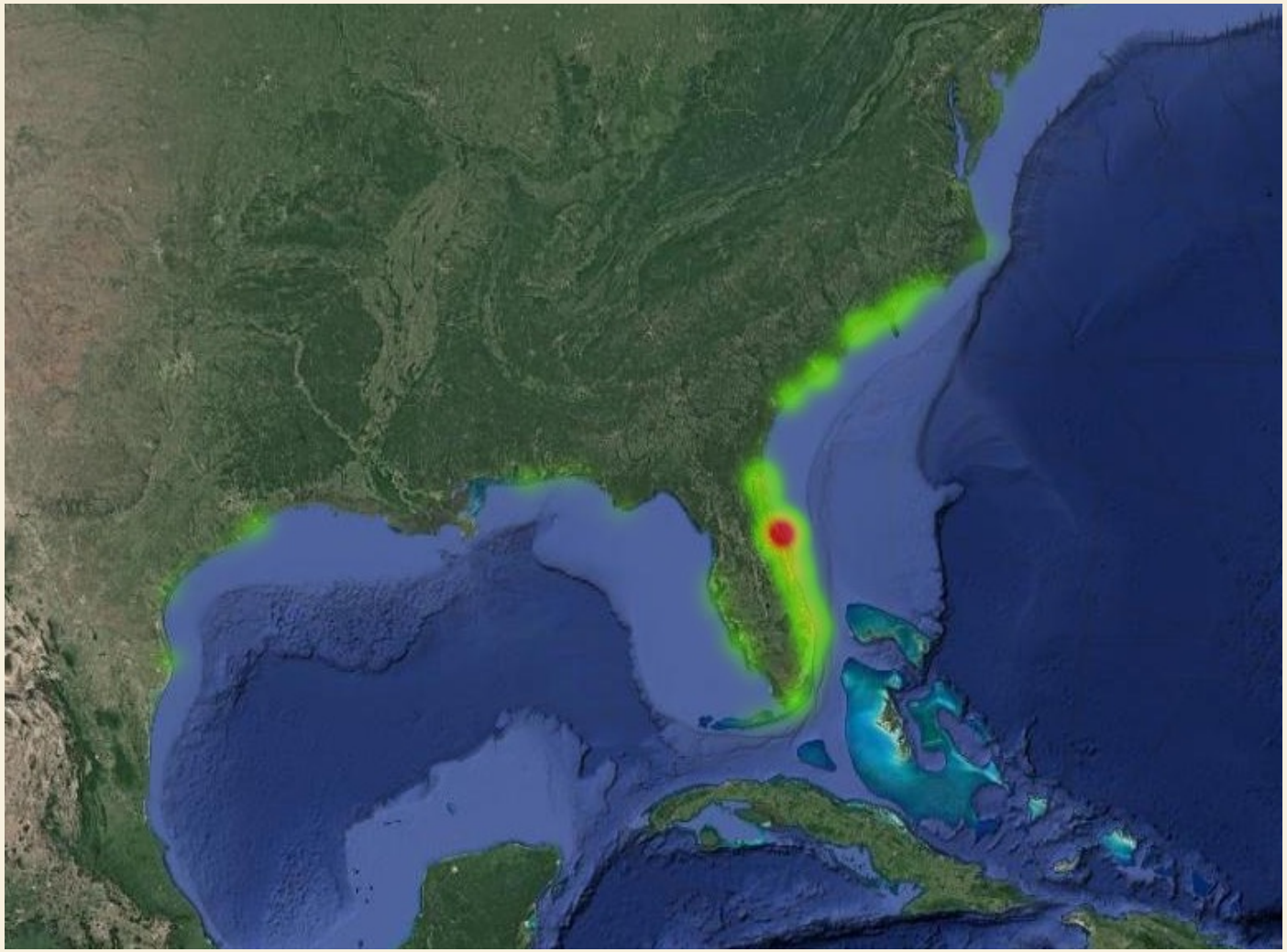



Shark Attacks in Top 7 States Grouped by Parent Activity
On Water
In Water
Fishing
Other
Science
250
200
150
100
50
0
Florida
Hawaii
California
South Carolina
North Carolina
Texas
Oregon
States

Reported Shark Attacks

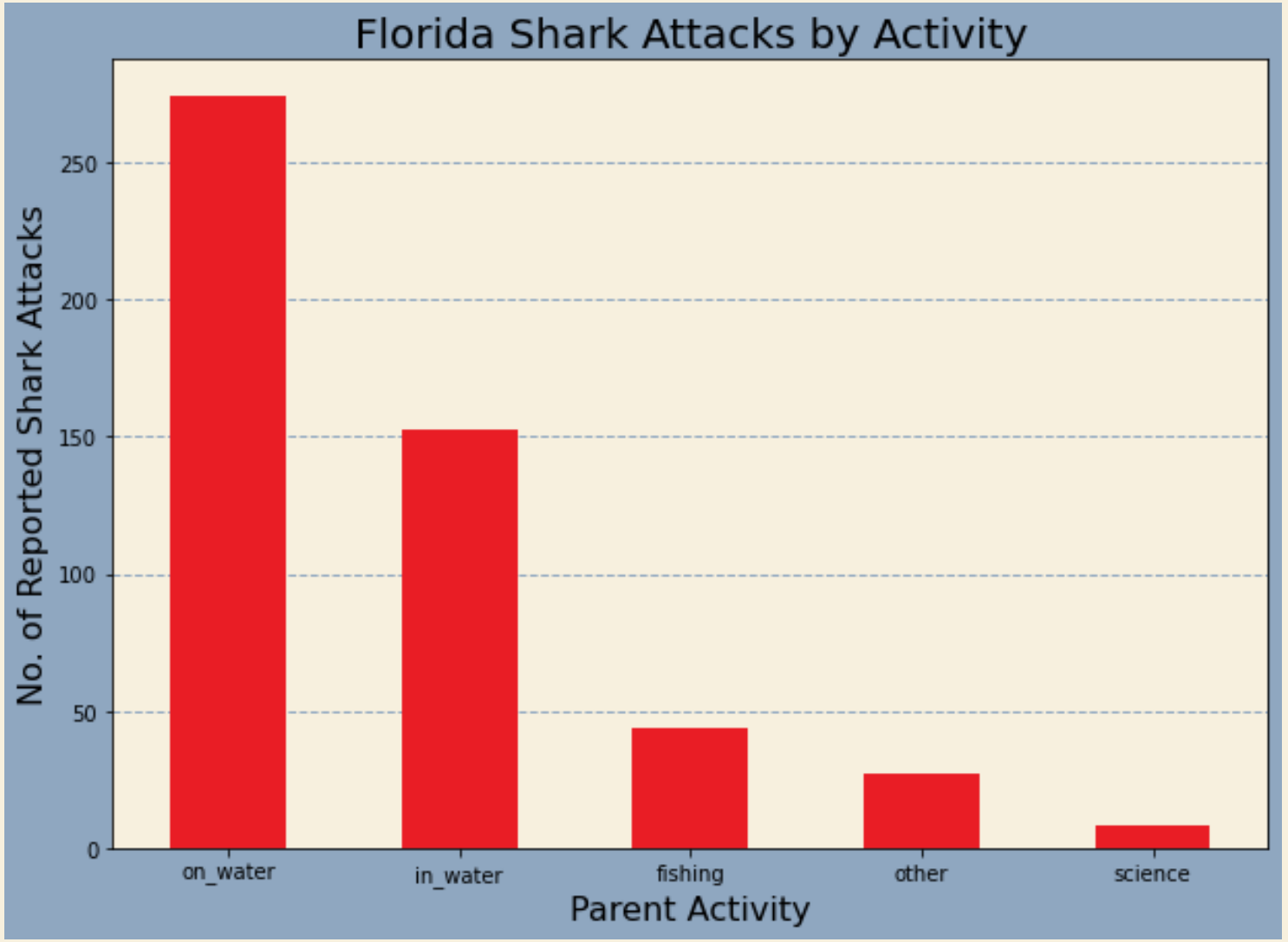
Heatmap of the East Coast Line

2000-2018



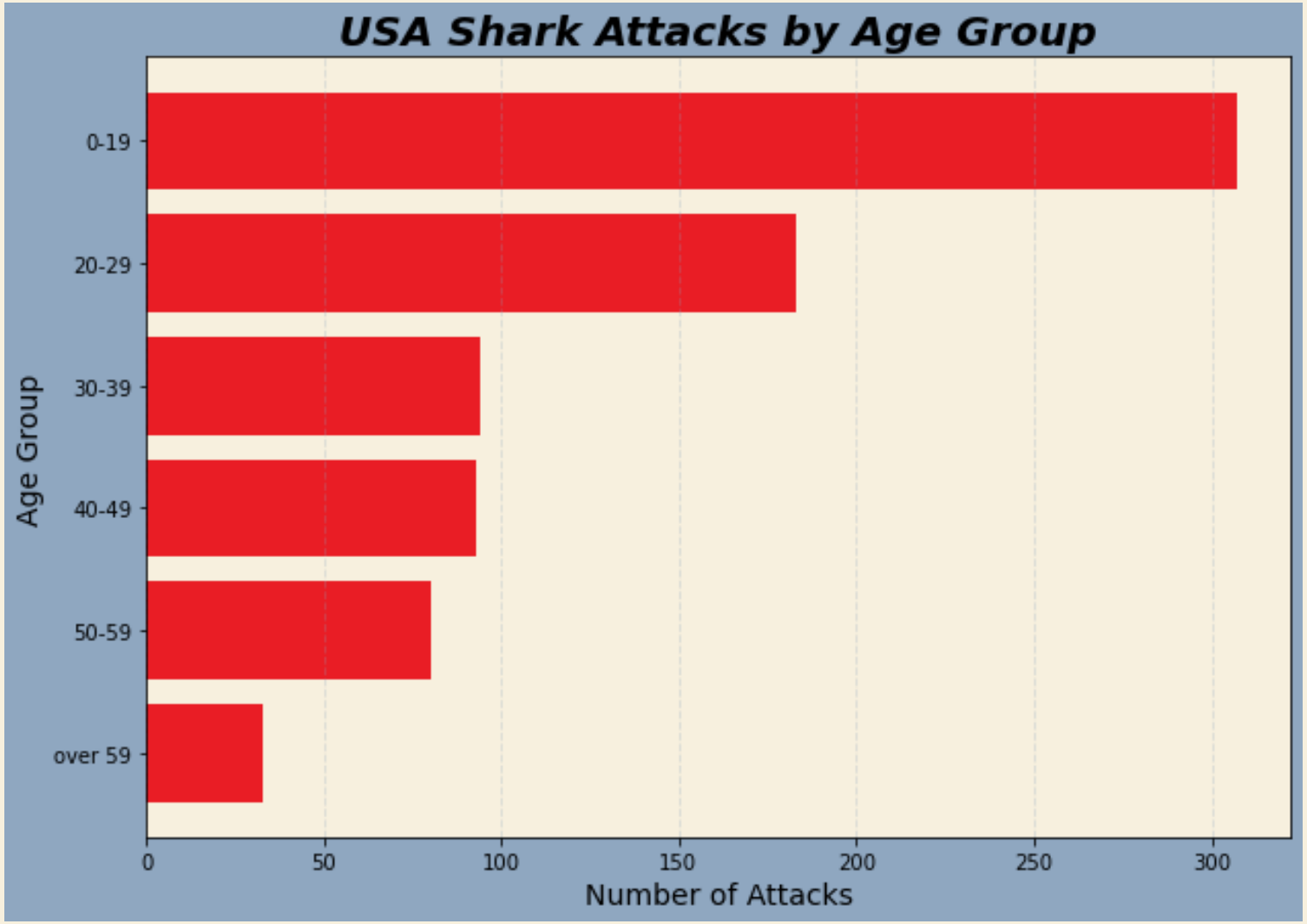
Shark Attacks by Activities in Florida

Shark Attacks in the USA



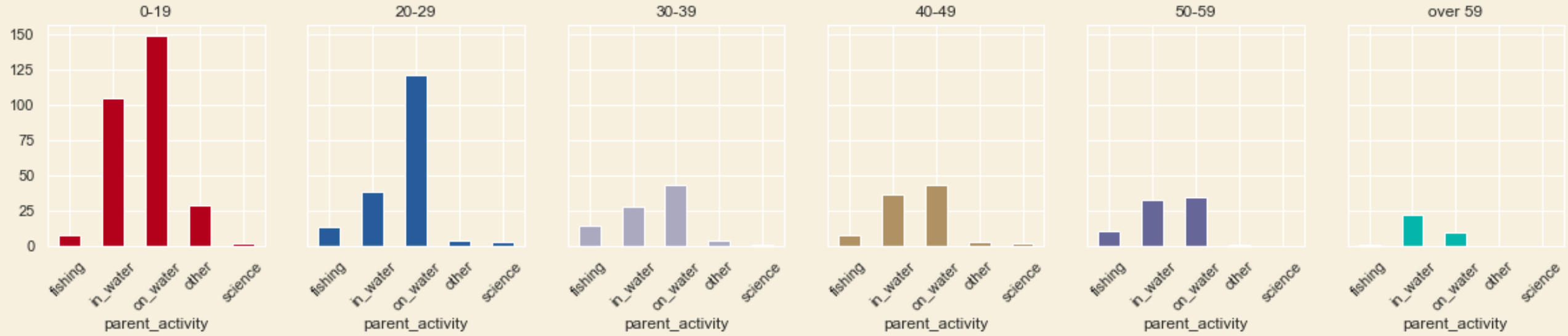
Shark Attacks by Age Group

Shark Attacks in the USA



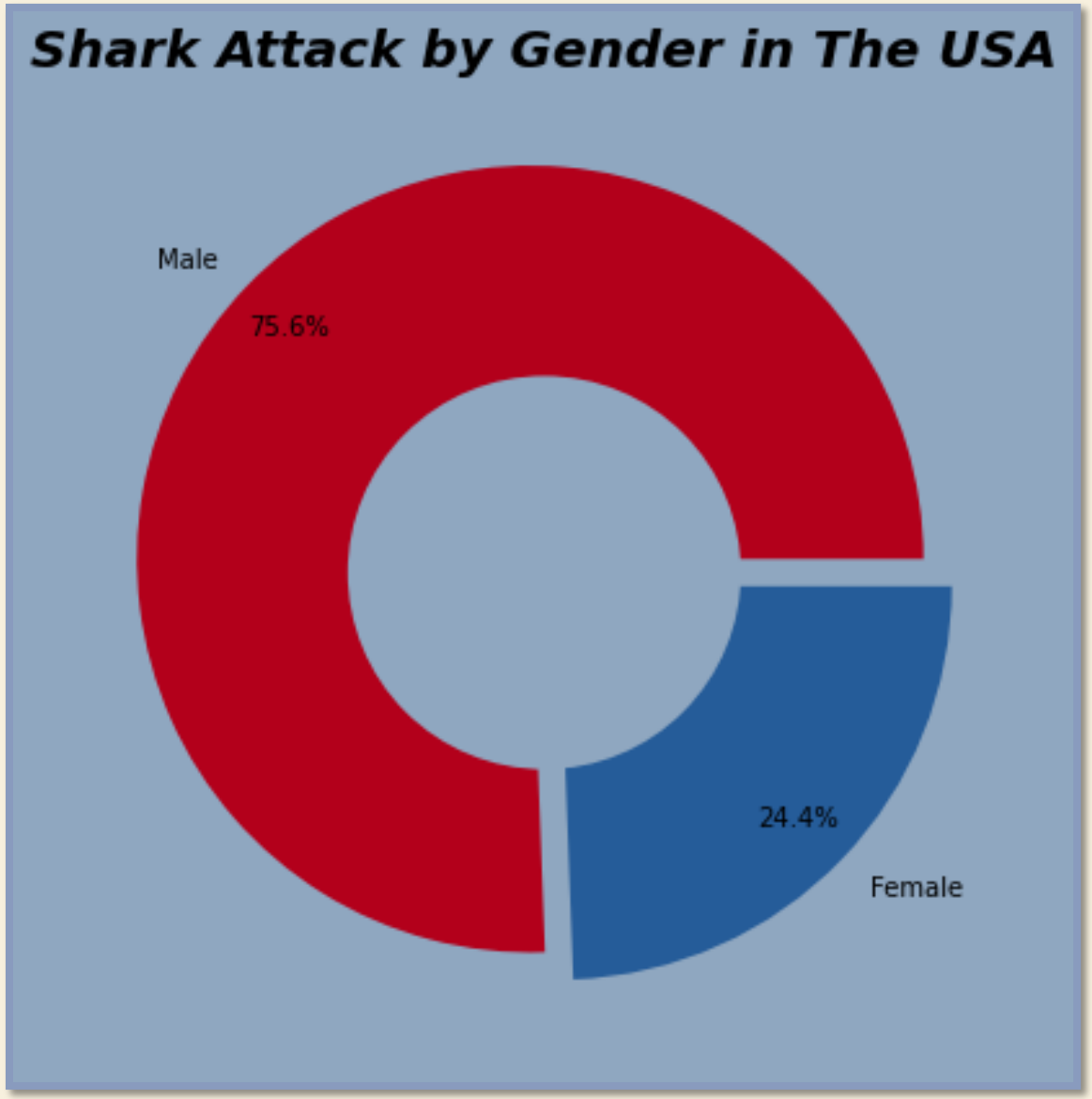
Shark Attacks by Age Group and Activities

Shark Attacks in the USA



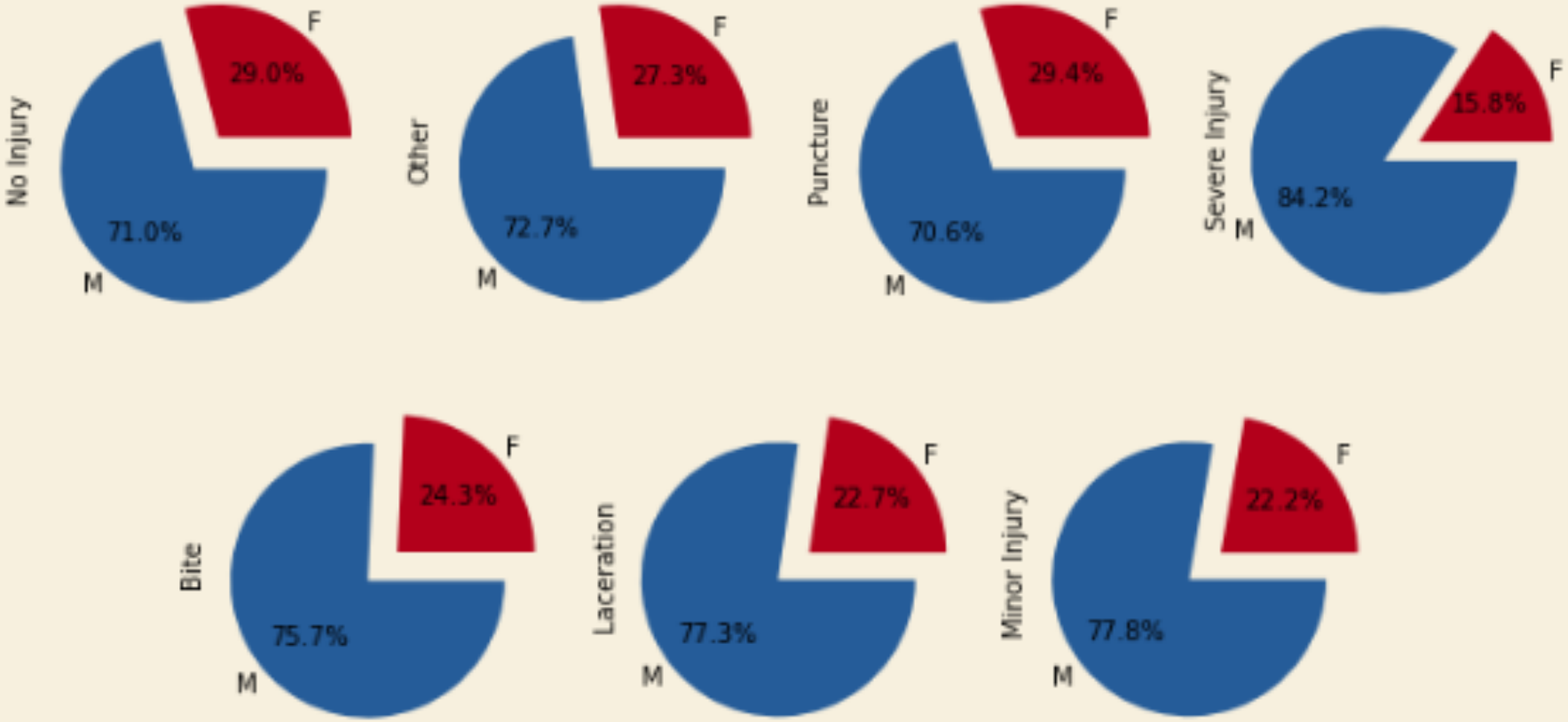
Shark Attacks by Gender

Shark Attacks in the USA



Shark Attacks by Gender and Activities

Shark Attacks in the USA

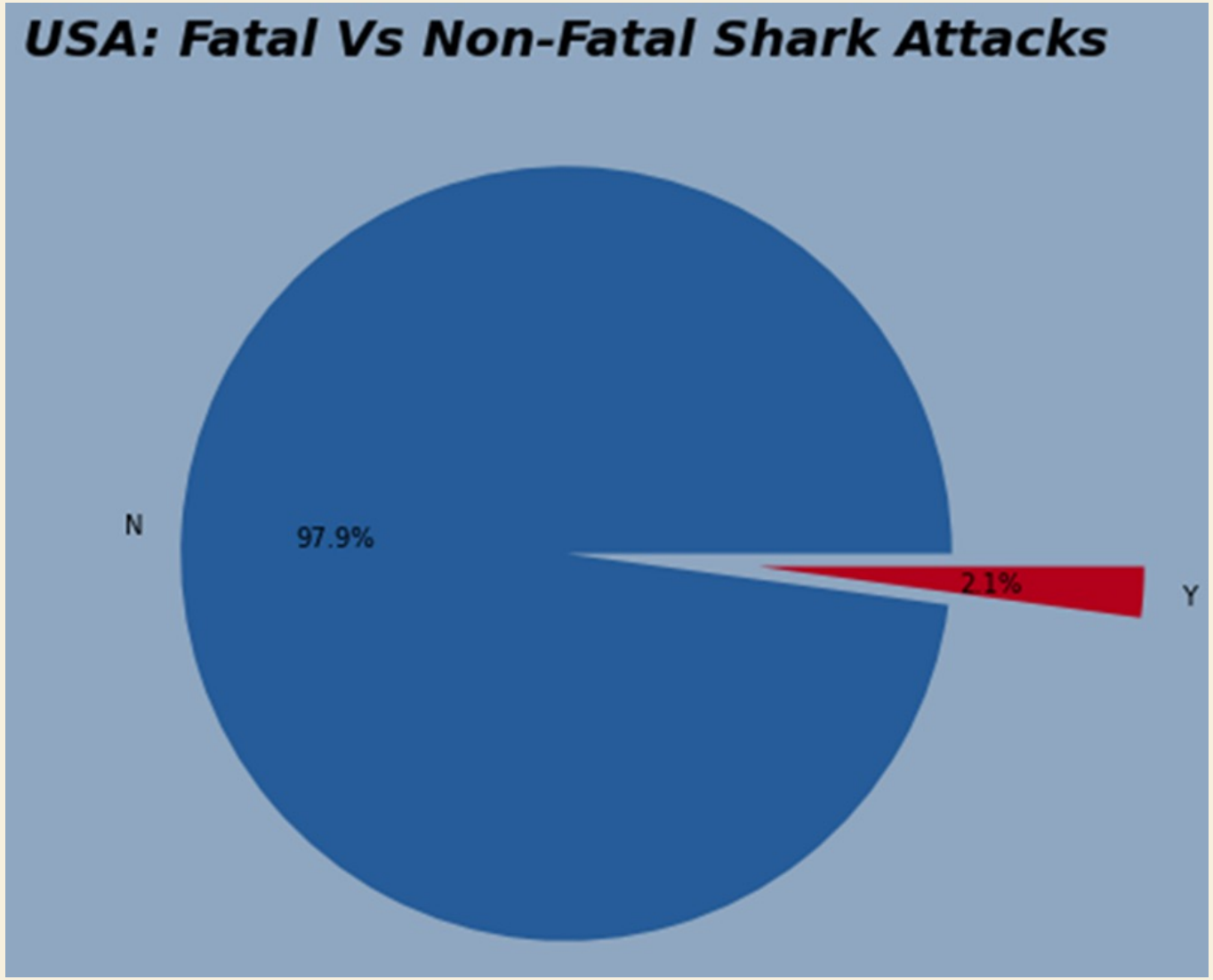


How severe was the attack?

Shark Attacks

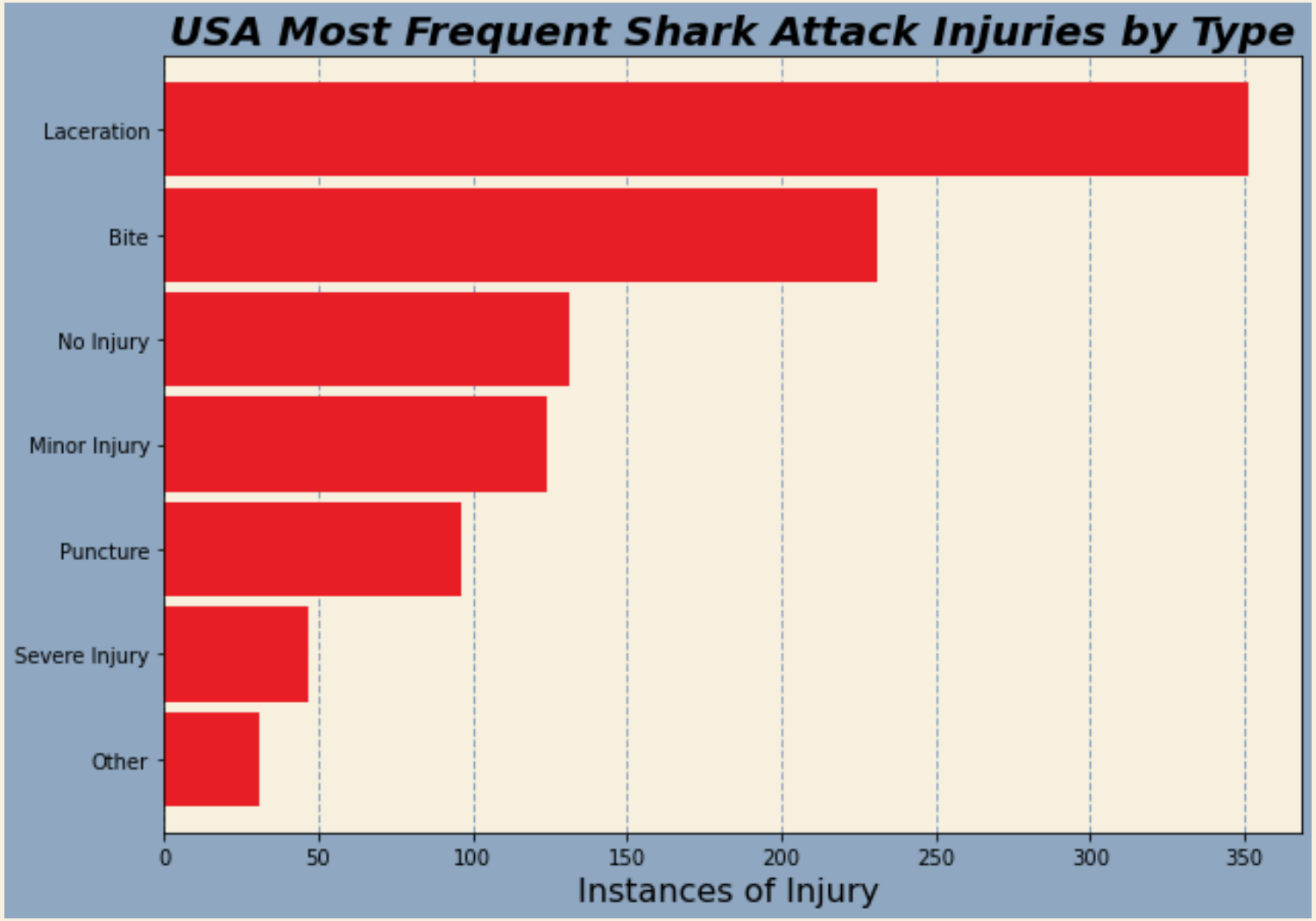
Fatal Vs. Non-Fatal

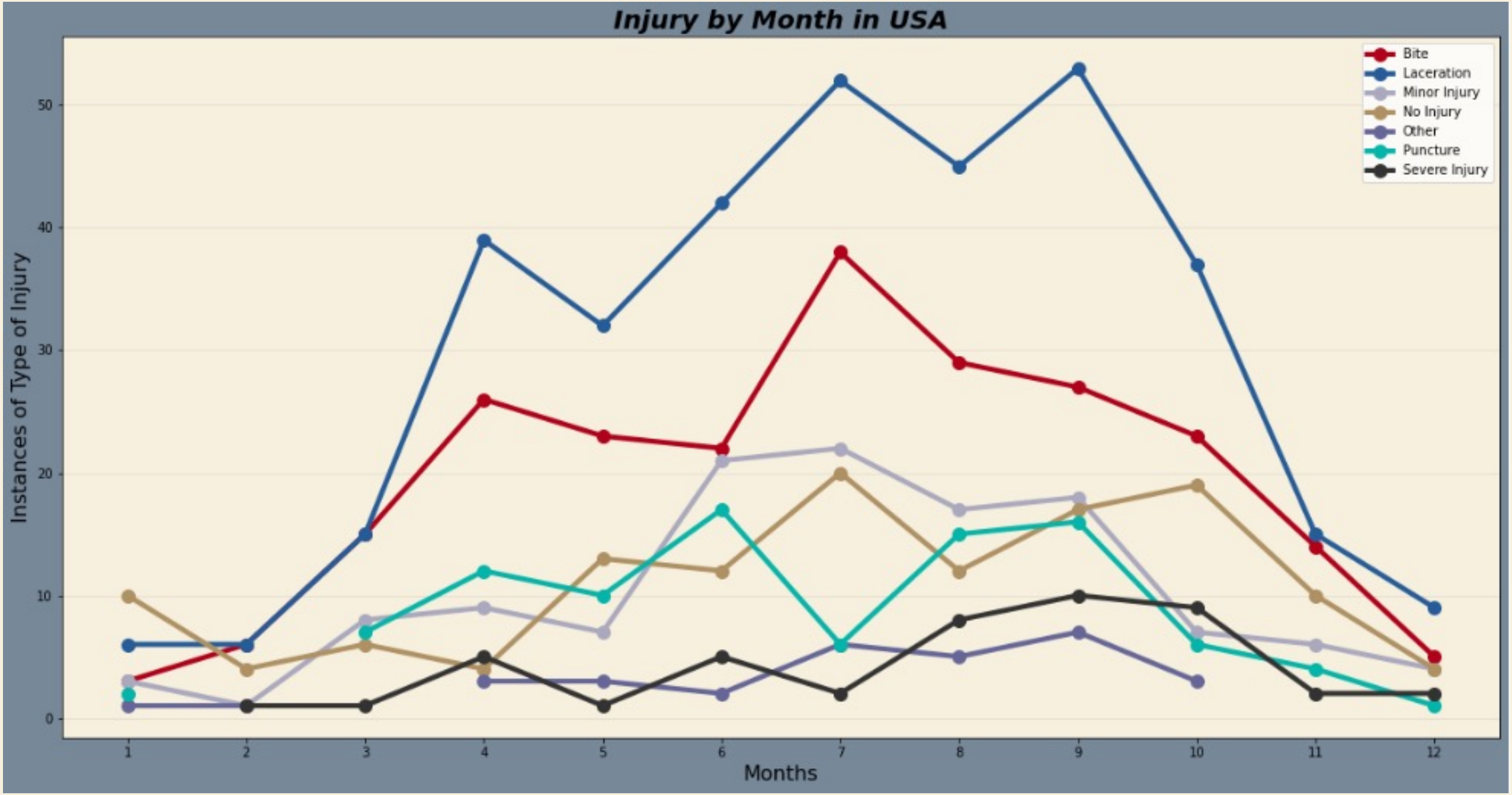
Shark Attacks in the USA



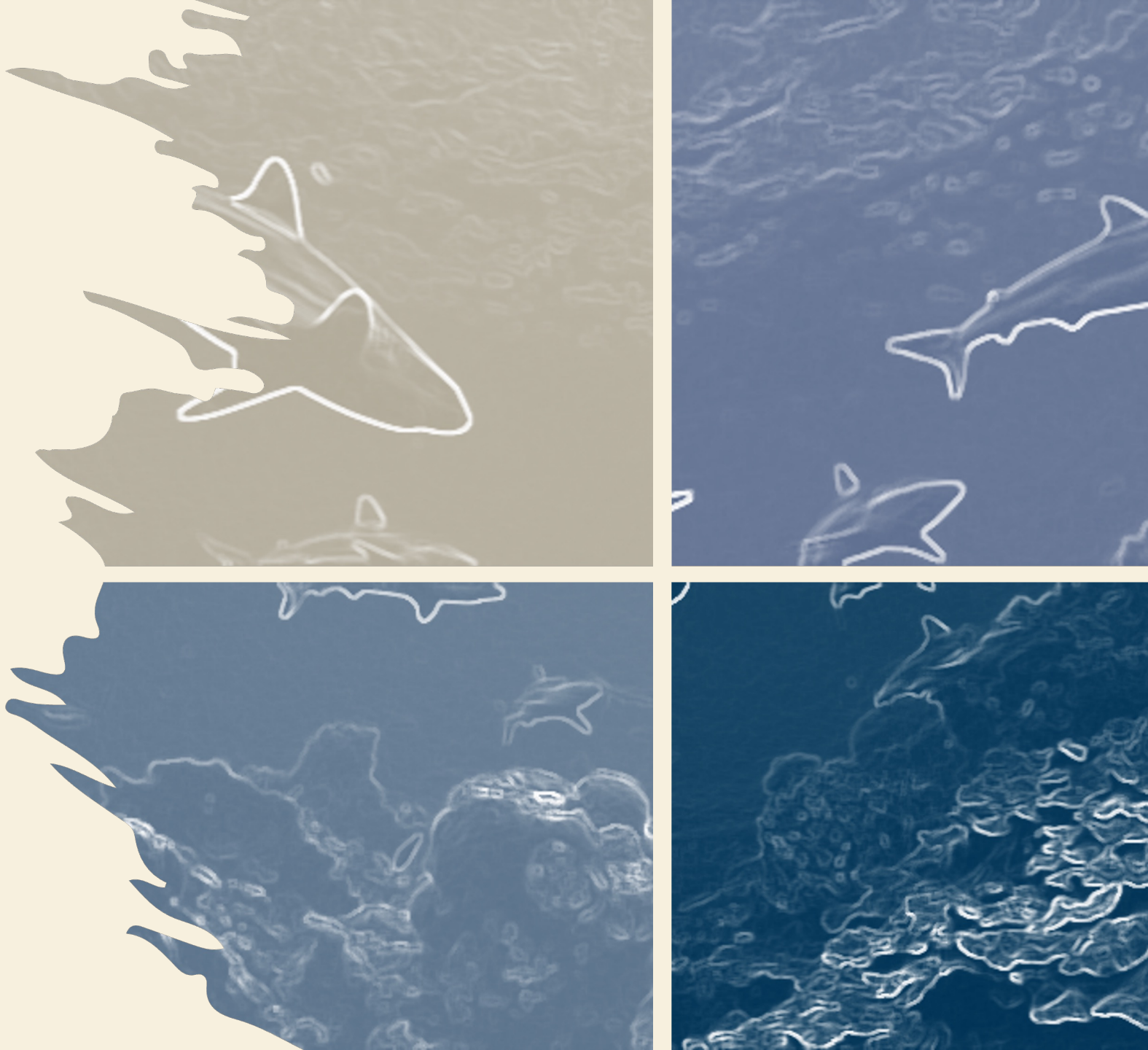
Shark Attacks by Injury Type

Shark Attacks in the USA



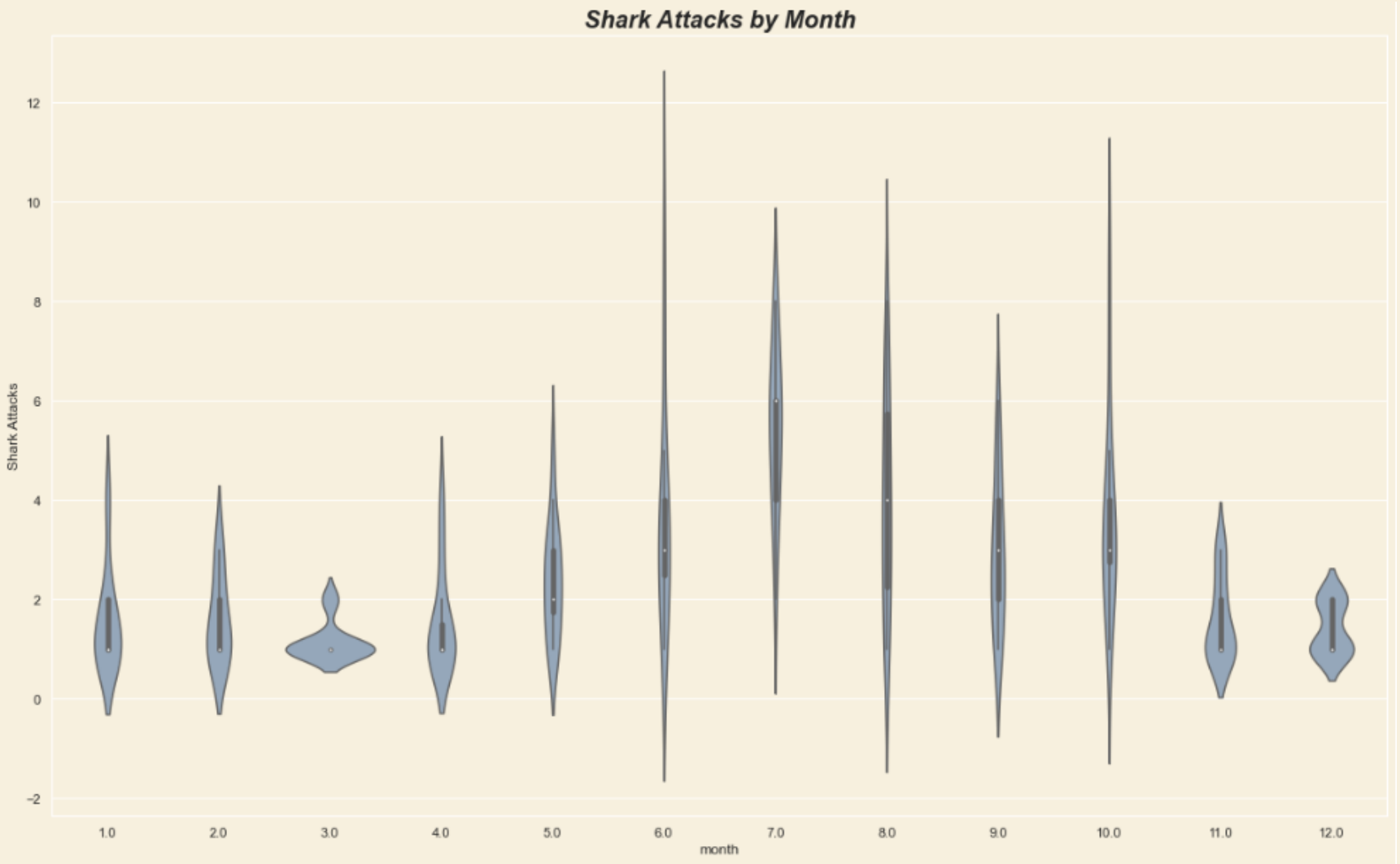

Injury by Month in USA
Bite
Laceration
Minor Injury
No Injury
Other
Puncture
Severe Injury
Instances of Type of Injury
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
Months

When have shark attacks been most reported?



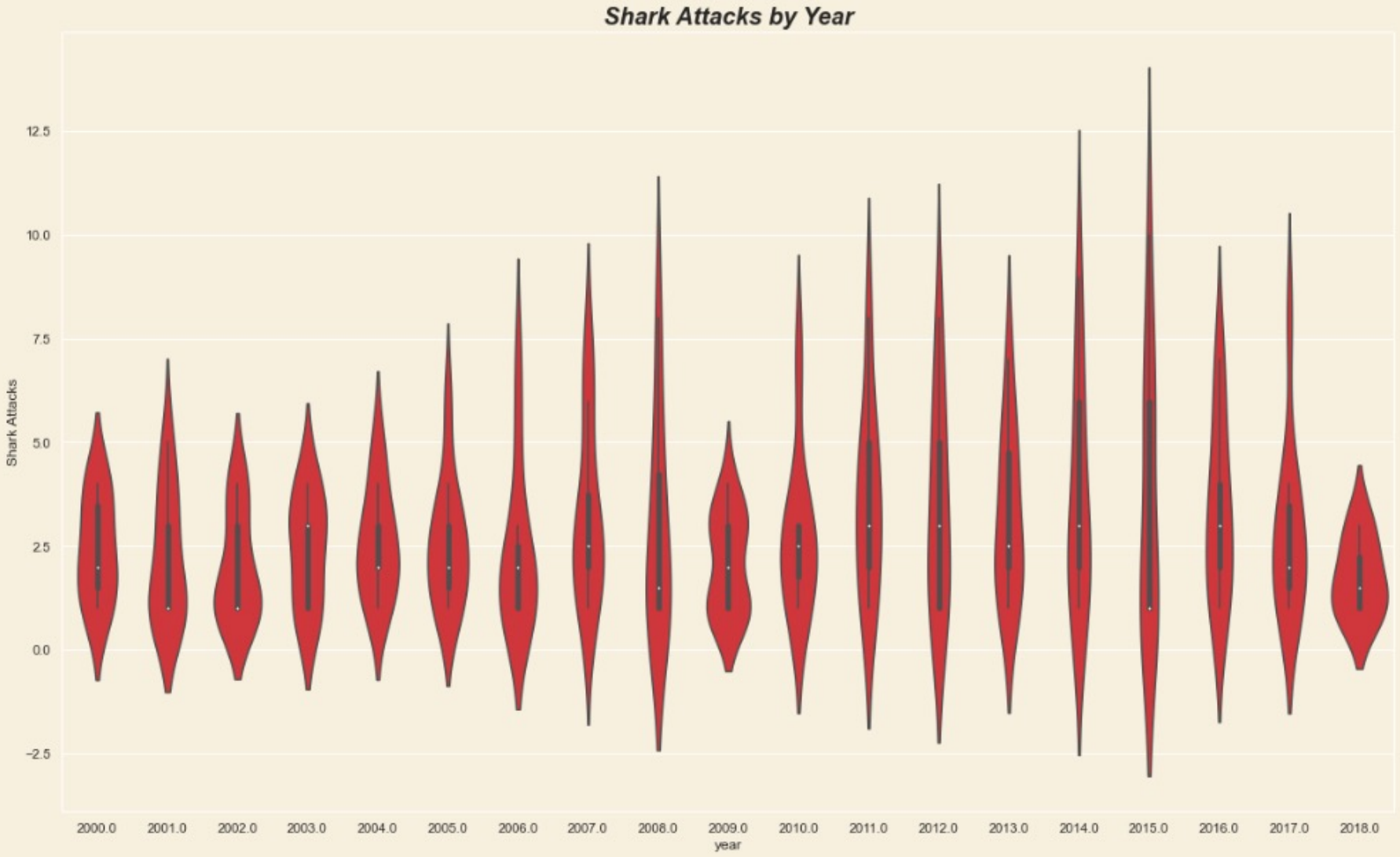
Shark Attacks by Month

Shark Attacks in the USA



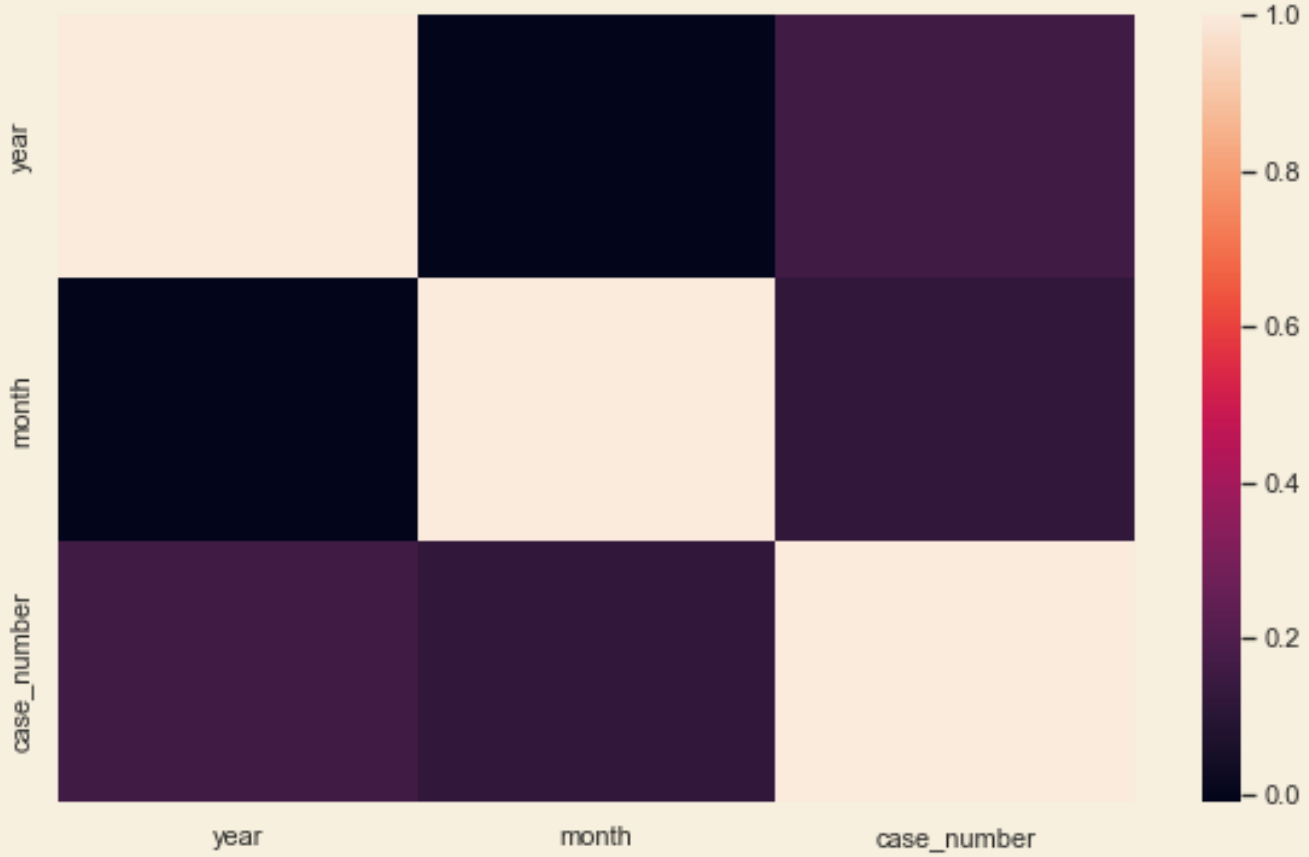
Shark Attacks by Year

Shark Attacks in the USA



Exploring Shark Attack Data

Shark Attacks in Florida



OLS Regression Results

Dep. Variable:	case_number	**R-squared:**	0.042
Model:	OLS	**Adj. R-squared:**	0.030
Method:	Least Squares	**F-statistic:**	3.590
Date:	Sun, 13 Nov 2022	**Prob (F-statistic):**	0.0298
Time:	18:35:30	**Log-Likelihood:**	-351.16
No. Observations:	168	**AIC:**	708.3
Df Residuals:	165	**BIC:**	717.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-121.5698	57.920	-2.099	0.037	-235.930	-7.209
year	0.0616	0.029	2.138	0.034	0.005	0.119
month	0.0806	0.049	1.637	0.103	-0.017	0.178

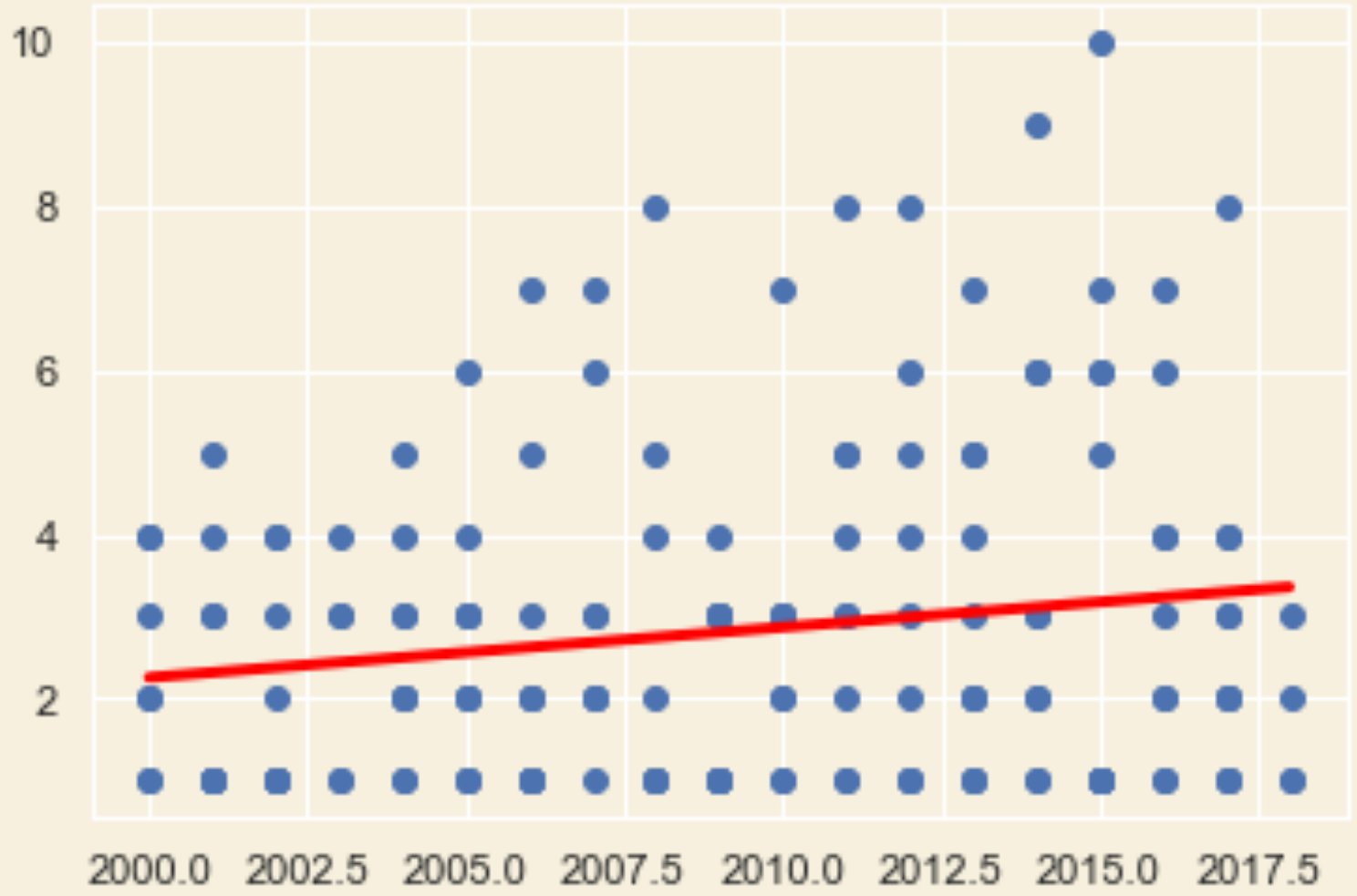
Omnibus:	28.185	**Durbin-Watson:**	1.542
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36.338
Skew:	1.082	**Prob(JB):**	1.29e-08
Kurtosis:	3.710	**Cond. No.**	7.64e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.64e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Linear Regression

Florida Reports by year

Given our data so far, we can see that our variance of attacks is incoherent. This makes it difficult to accurately predict the consistency of attacks in Florida. However, because the regression is at the very least, increasing over time. This backs our hypothesis that as population increases over time then so will the amount of reported Shark Attacks



Linear Regression

Florida Reports by year – Predictions

Our hypothesis that reported shark attacks will continue to rise with increased population shows to be true based on the predictions shown for 2022 and 2030.

```
In [137]: predict = [2022]

          predict = np.array(predict)
          predict = predict.reshape(1, -1)
          reg.predict(predict)

Out[137]: array([3.60260757])
```

```
In [138]: predict = [2030]

          predict = np.array(predict)
          predict = predict.reshape(1, -1)
          reg.predict(predict)

Out[138]: array([4.09169892])
```

Data Limitations

- Case Reports have missing data like the age, gender, and type of shark
- Lack of reporting around the world
- What is considered a shark attack is subjective, there were several that may not have been shark related

Conclusion

Based on the data that was analyzed:

- Fatal shark attacks are very rare, making up for ~2% of reported shark attacks
- Reported attacks are slightly increasing
- Florida has the most reported shark attacks
- Men are 3x more likely than women to be reported attacked by a shark
- July – September have the most reported shark attacks, and November – March with the least.

Data Location

- file:///C:/Users/brenn/SMU_Bootcamp/Projects/project_1_group_3/smu_project1_sharkattacks/smu_project1_sharkattacks/bee_work/export.html


Thank you!
Questions?