W1 Project
Data cleaning & wrangling

Paula Hernández / Yu Ting Hu

21-May-2022

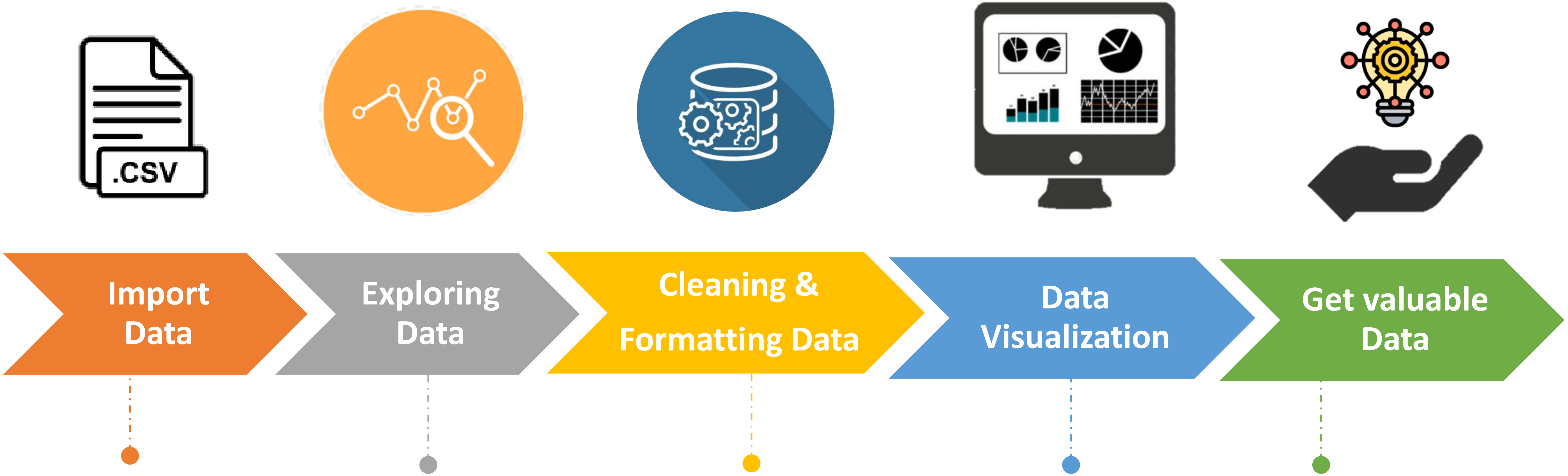
Project Overview

In this Project we have used Python3 to deal with a messy data set, and analyze and process it into valuable data, from which we have been able to extract valuable insights and information.

The data set was extracted from Global Shark Attack File. It consists of current and historical data of shark/human interactions, with the aim of better understanding these interactions, and minimize the risk of being injured by a shark, while contributing in the conservation of shark species worldwide.

Data source:
https://www.kaggle.com/datasets/teajay/global-shark-attacks?resource=download

Data Processing



Import Data

Import dataset and required libraries

Exploring Data

General data set exploration and EDA exploration

Cleaning & Formatting Data

Establish hypotheses for our scope, eliminate unneeded data and format the data

Data Visualization

Plot data with aid of plots/graphs and geographical map

Get valuable Data

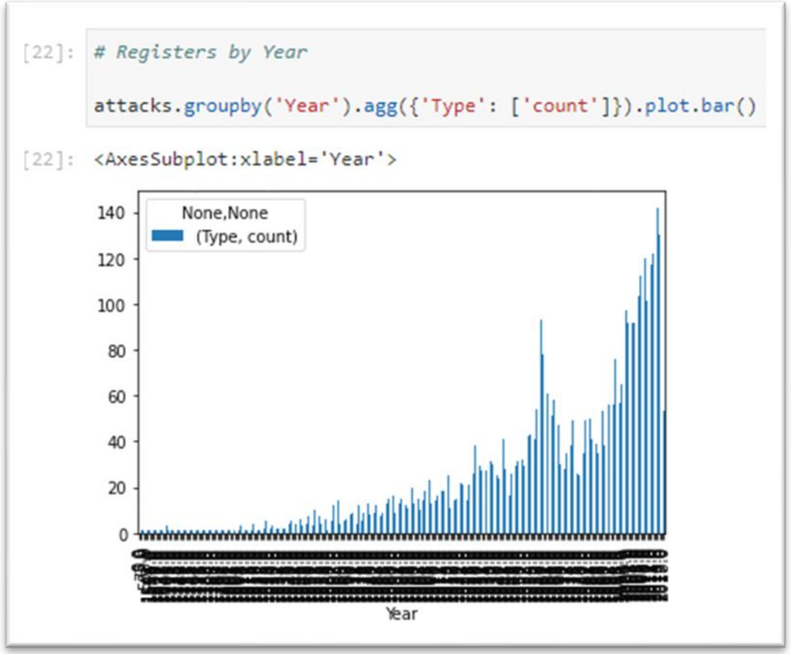
Get valuable information/insights from data obtained

Exploring the Data

To explore the data, we used df.describe(), df[“column”], and also exploration by columns using some EDA techniques.

The main purpose of this exploration is to have a general overview of the data's distribution and to filter the scope in which we want to focus our analysis.

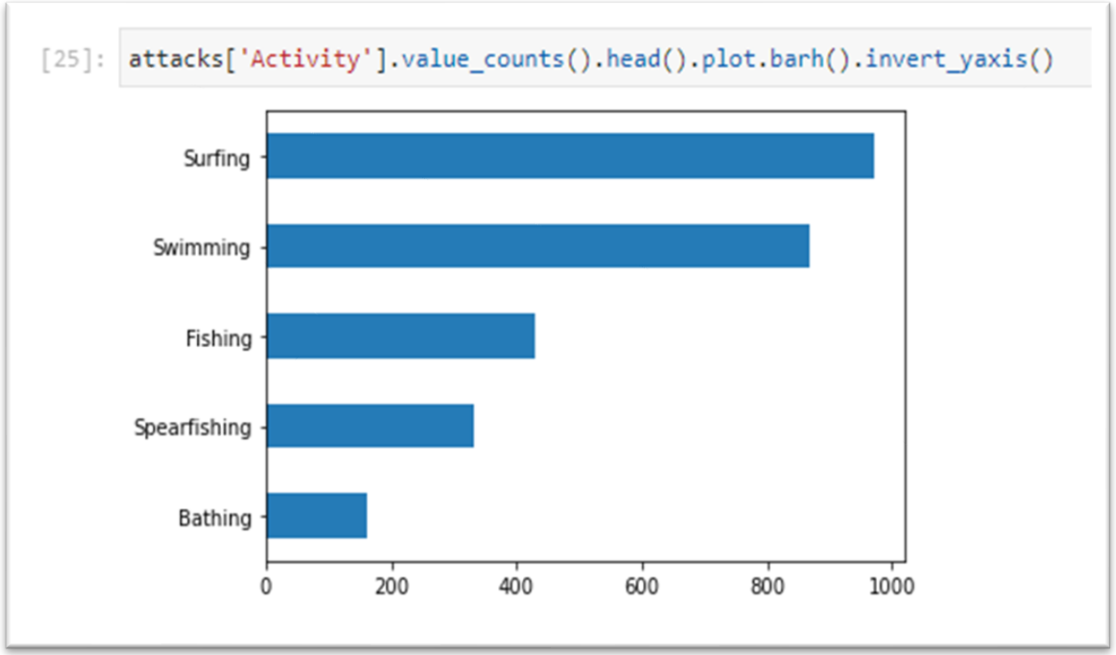
This exploration allowed us to have a general overview of the data's distribution, being able to filter the scope in which we decided to focus our analysis.



```
[22]: # Registers by Year

      attacks.groupby('Year').agg({'Type': ['count']}).plot.bar()

[22]: <AxesSubplot:xlabel='Year'>
```



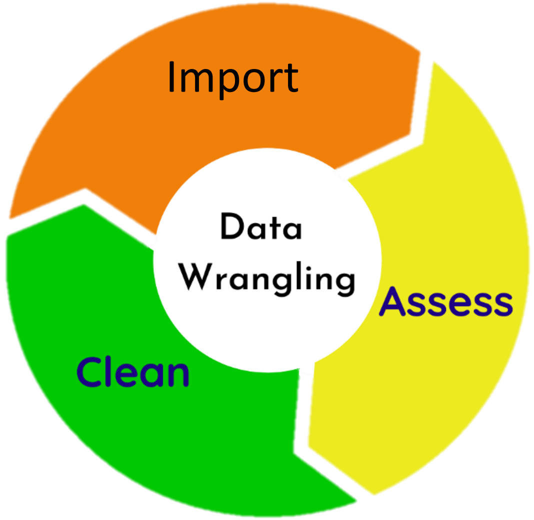
```
[25]: attacks['Activity'].value_counts().head().plot.barh().invert_yaxis()
```

Cleaning and formatting the Data

After exploring the data, we have cleaned the data based on the following hypotheses:

- ➢ Keep the columns that are relevant to our study (Case Number, Date, Year, Type, Country, Activity, Sex, Species).
- ➢ Focus the analysis on the data registered after 1950.
- ➢ Eliminate registers with empty/not valid data.

Some of the cleaning techniques and methods used are:
Drop columns, drop null values, string manipulation, dropna, isnull, map, filter, rename, replace, regex, lambda, datetime, append, etc.



Cleaning and formatting the Data

In order to understand and analyze the data correctly, we need to format the data to have standardized type of data and meaning.

After cleaning the data according to the above criteria, we will:

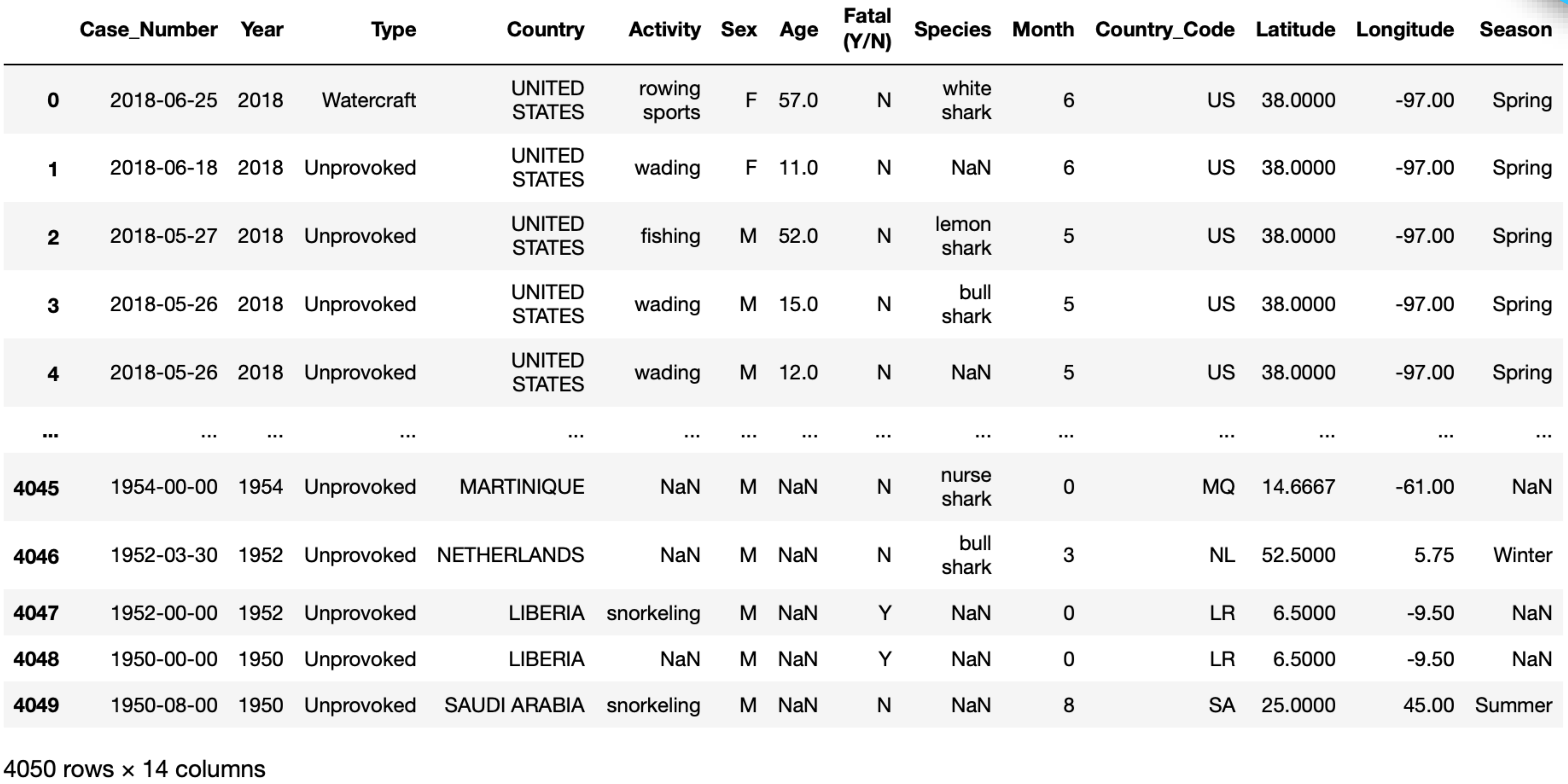
✓ Describe each variable
✓ Analyze which factors may contribute to the fatality rate of shark at
✓ Analyze the information by country

In addition, we included new columns:

✓ Country Code (Alpha-2 code per ISO 3166), by using pycountry
✓ Coordinates (Latitude and Longitude)
✓ Month
✓ Season in each geographical area

The final data frame has a total of 4050 registers and 14 columns.

Cleaning and formatting the Data



	Case_Number	Year	Type	Country	Activity	Sex	Age	Fatal (Y/N)	Species	Month	Country_Code	Latitude	Longitude	Season
0	2018-06-25	2018	Watercraft	UNITED STATES	rowing sports	F	57.0	N	white shark	6	US	38.0000	-97.00	Spring
1	2018-06-18	2018	Unprovoked	UNITED STATES	wading	F	11.0	N	NaN	6	US	38.0000	-97.00	Spring
2	2018-05-27	2018	Unprovoked	UNITED STATES	fishing	M	52.0	N	lemon shark	5	US	38.0000	-97.00	Spring
3	2018-05-26	2018	Unprovoked	UNITED STATES	wading	M	15.0	N	bull shark	5	US	38.0000	-97.00	Spring
4	2018-05-26	2018	Unprovoked	UNITED STATES	wading	M	12.0	N	NaN	5	US	38.0000	-97.00	Spring
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4045	1954-00-00	1954	Unprovoked	MARTINIQUE	NaN	M	NaN	N	nurse shark	0	MQ	14.6667	-61.00	NaN
4046	1952-03-30	1952	Unprovoked	NETHERLANDS	NaN	M	NaN	N	bull shark	3	NL	52.5000	5.75	Winter
4047	1952-00-00	1952	Unprovoked	LIBERIA	snorkeling	M	NaN	Y	NaN	0	LR	6.5000	-9.50	NaN
4048	1950-00-00	1950	Unprovoked	LIBERIA	NaN	M	NaN	Y	NaN	0	LR	6.5000	-9.50	NaN
4049	1950-08-00	1950	Unprovoked	SAUDI ARABIA	snorkeling	M	NaN	N	NaN	8	SA	25.0000	45.00	Summer

4050 rows × 14 columns

Data Visualization

Attacks by Activity

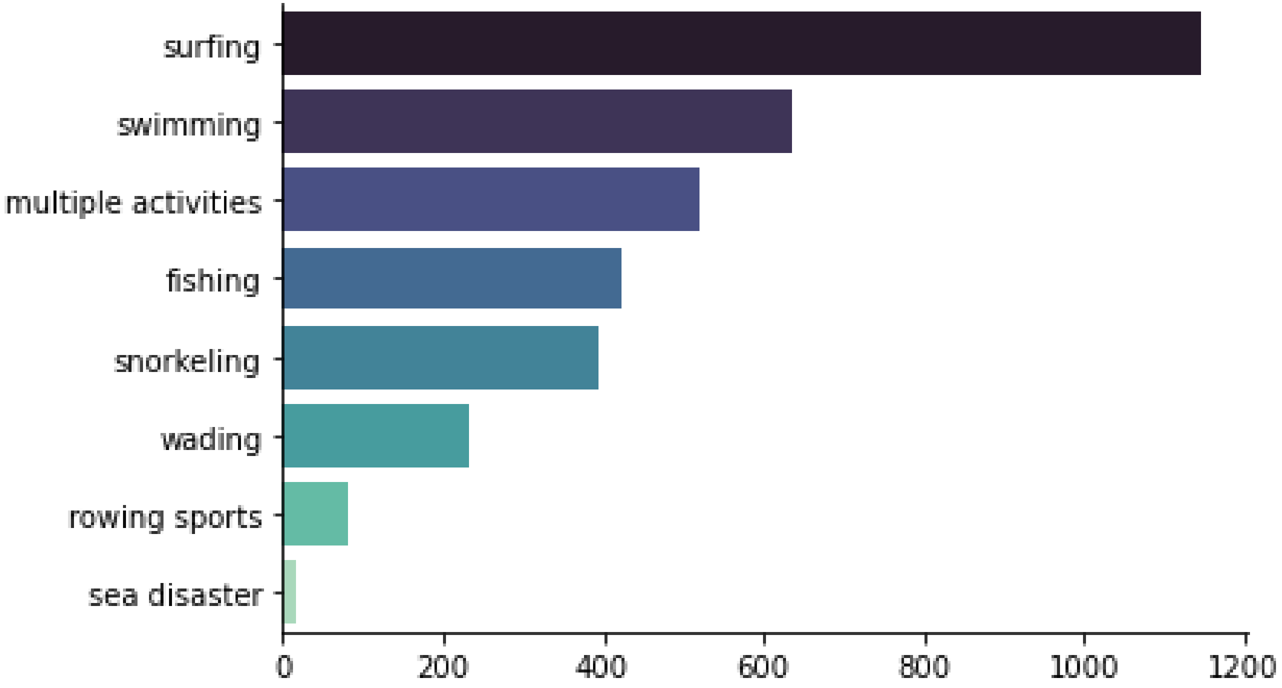


Fig.1 Total number of Attacks by Activity

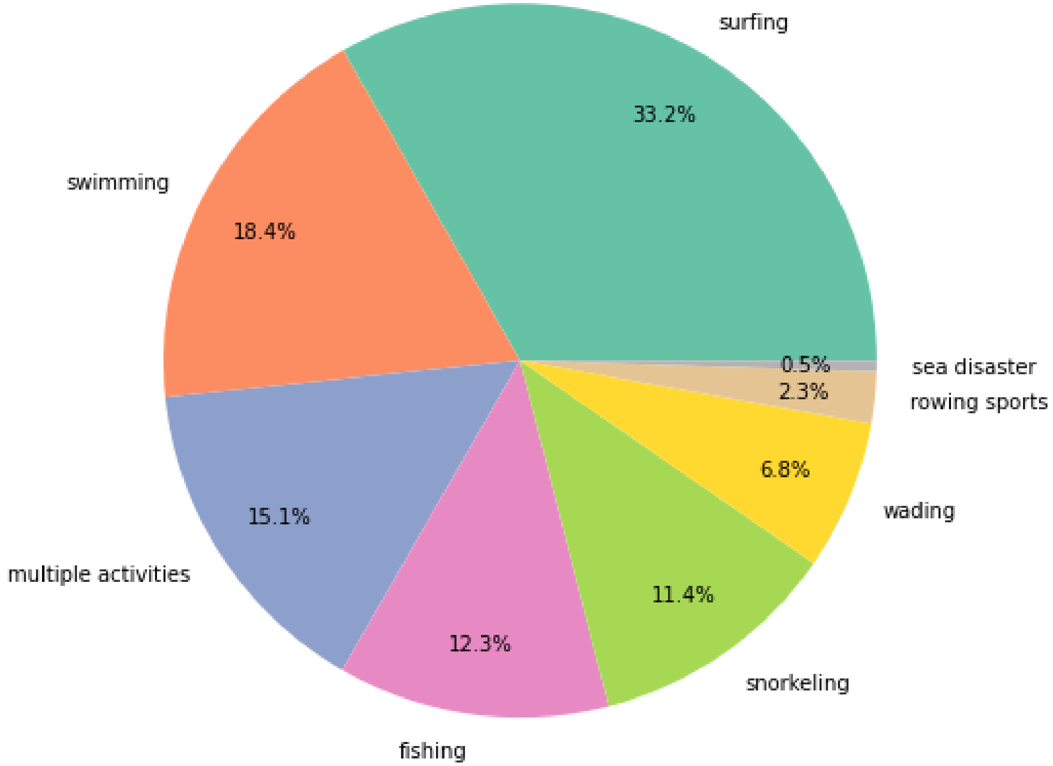


Fig.2 Percentage Attacks by Activity

Data Visualization

Fatality by Activity

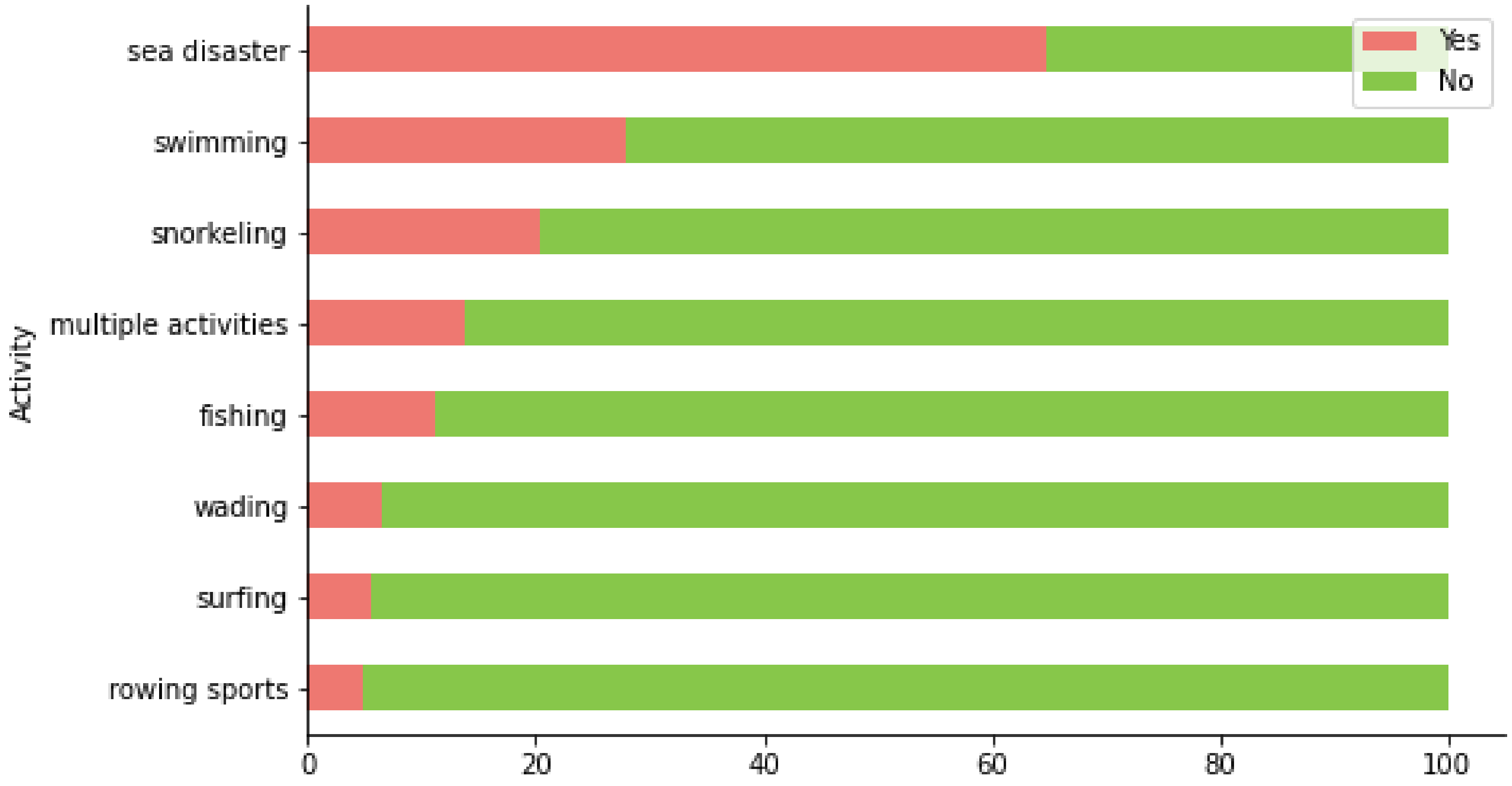


Fig.3 Fatality % per Activity

Data Visualization

Fatality by Type of Event

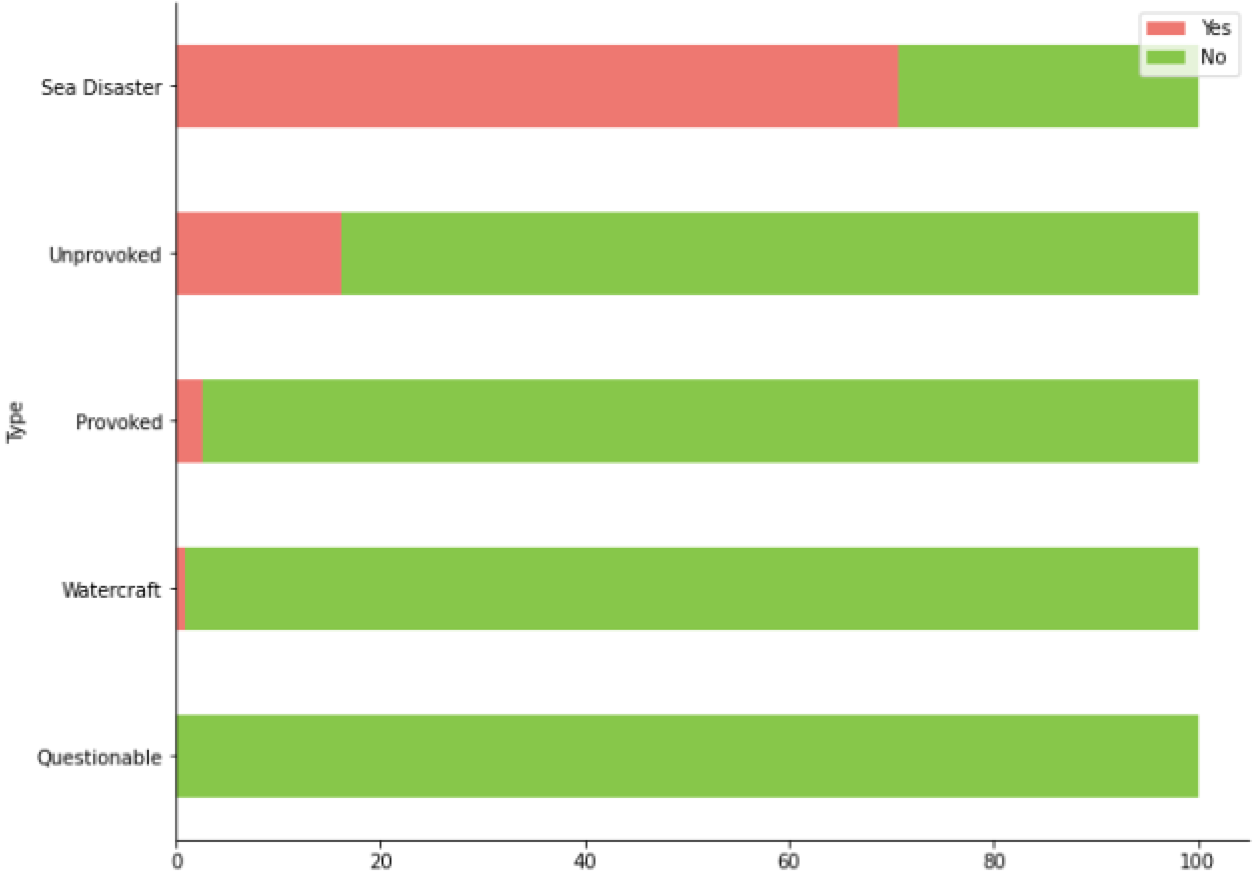


Fig.4 Fatality % per Type of Event

Data Visualization

Social-demographic Analysis

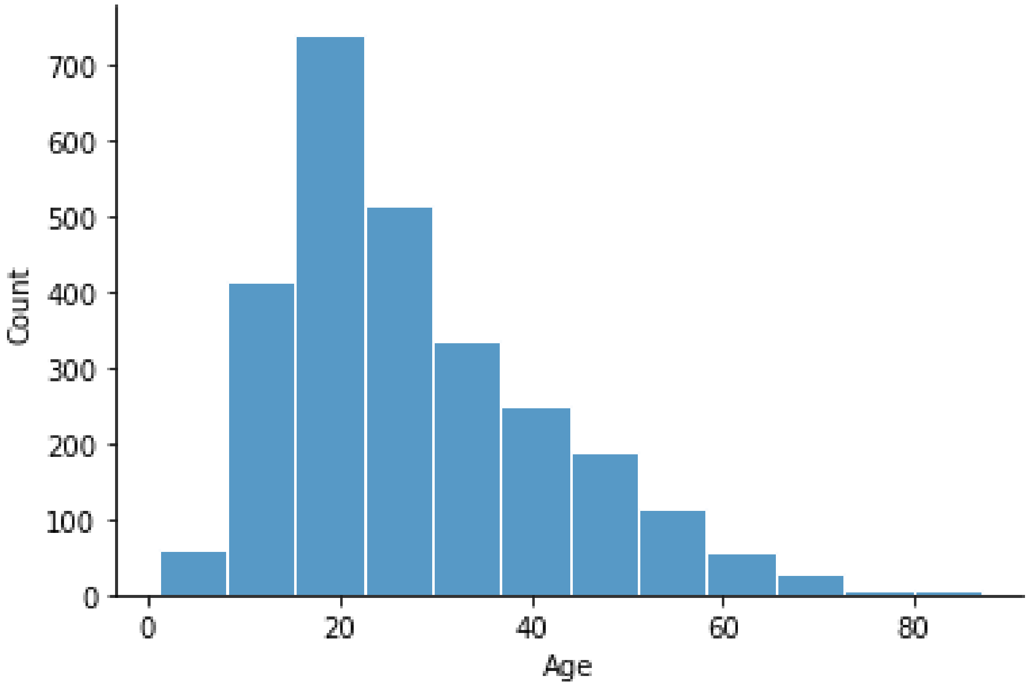


Fig.5 Attacks by Age Histogram

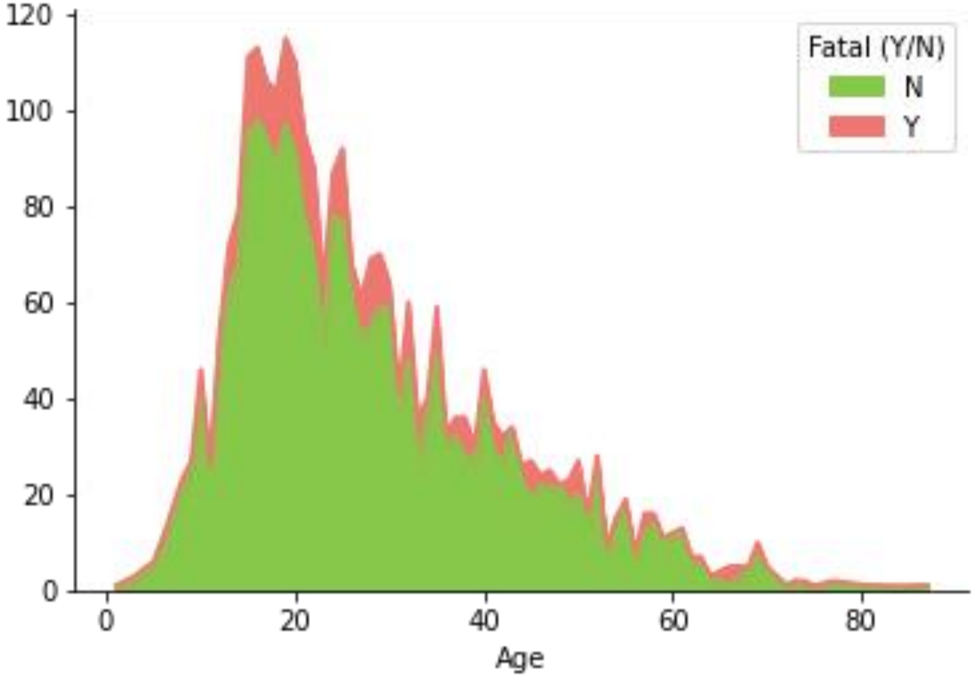


Fig.6 Fatality according Age Area

Data Visualization

Social-demographic Analysis

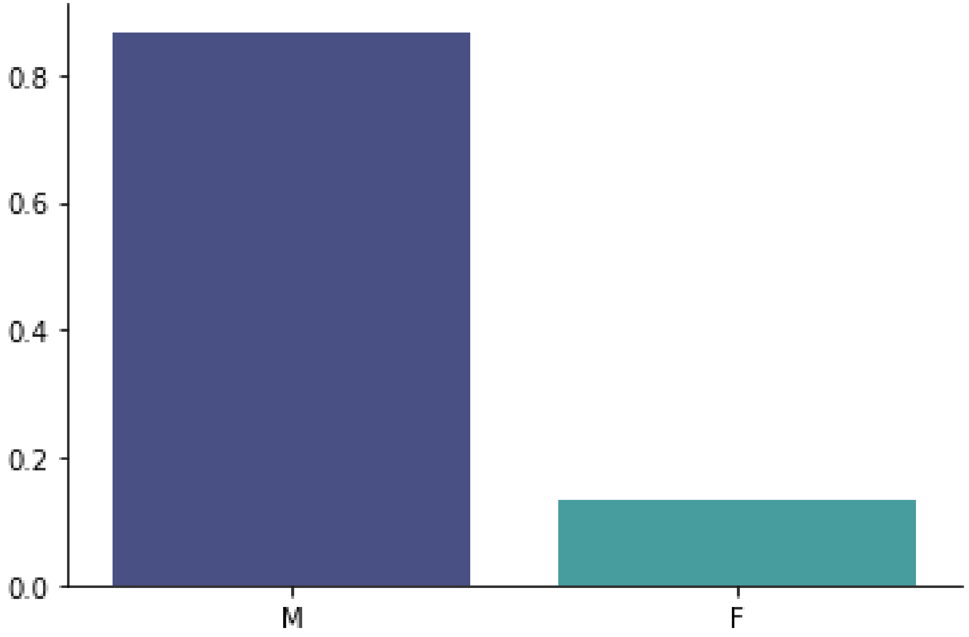


Fig. 7 Total attacks by Sex

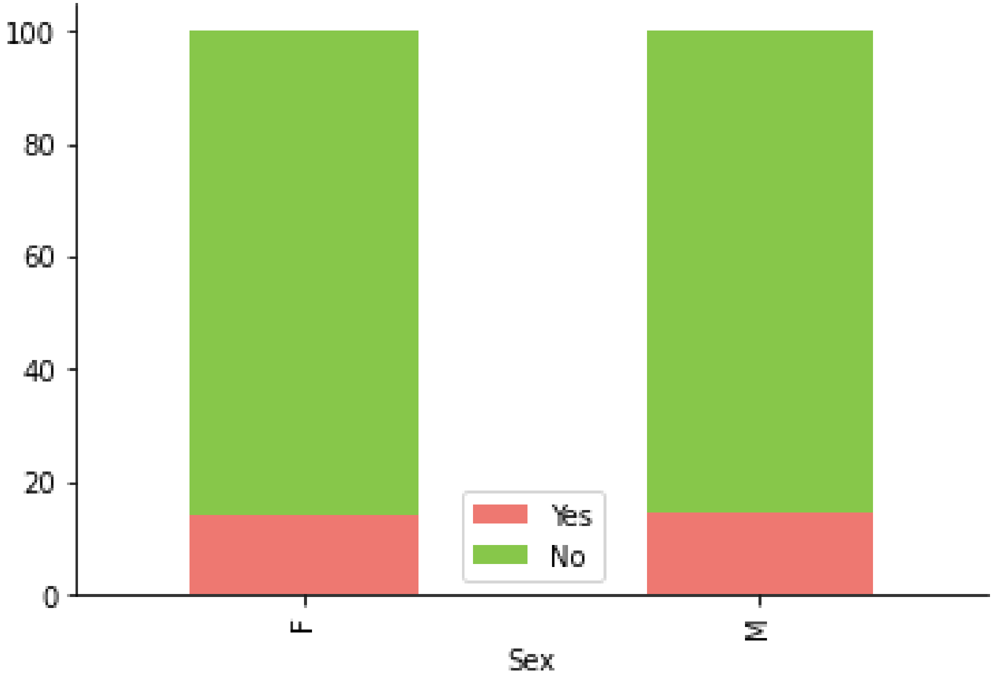
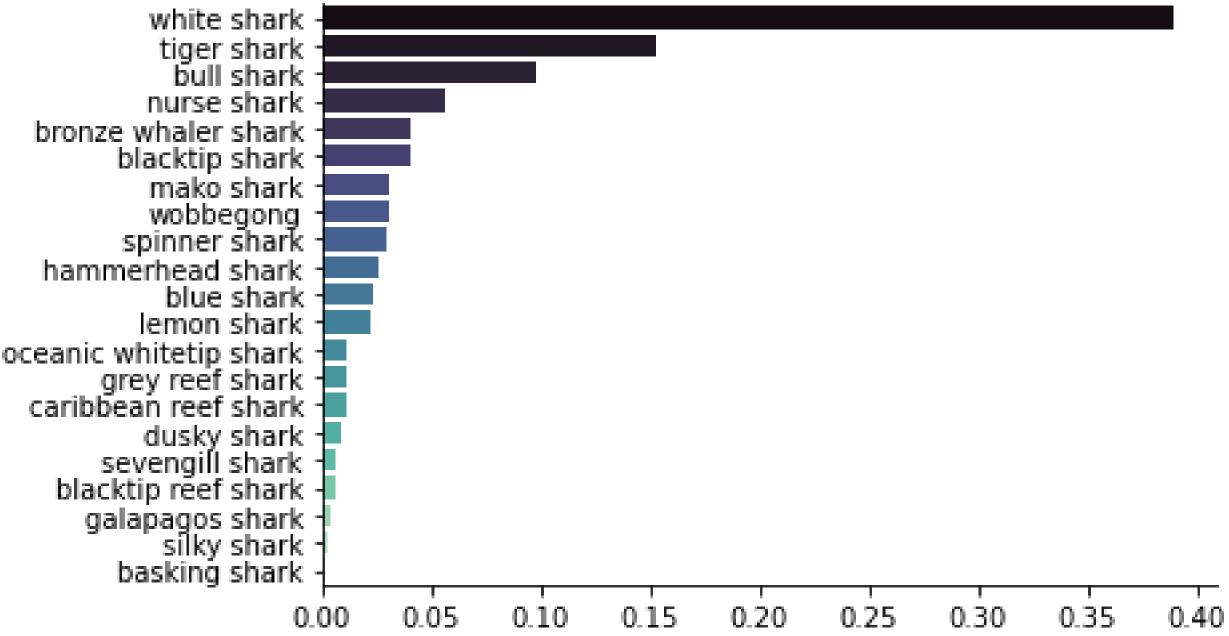


Fig.8 Fatality by Sex

Data Visualization

Attacks by Shark Species



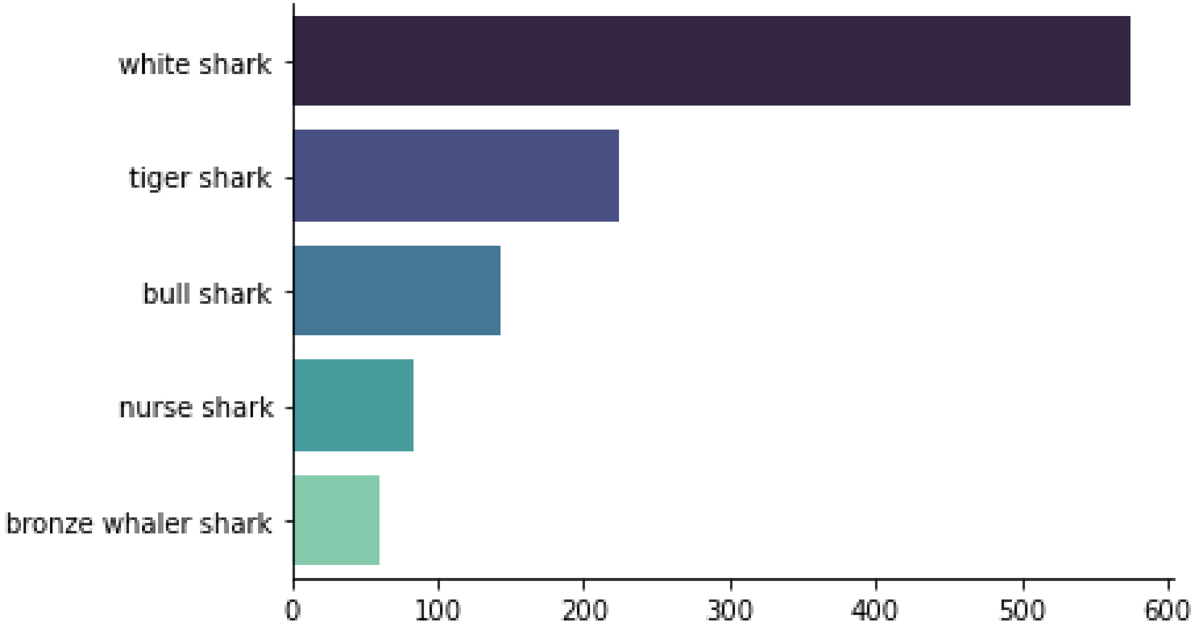


Fig.9 Total no. of Attacks by Shark

Fig.10 Total no. of Attacks by Shark (Top 5)

Data Visualization

Fatality by Shark Species

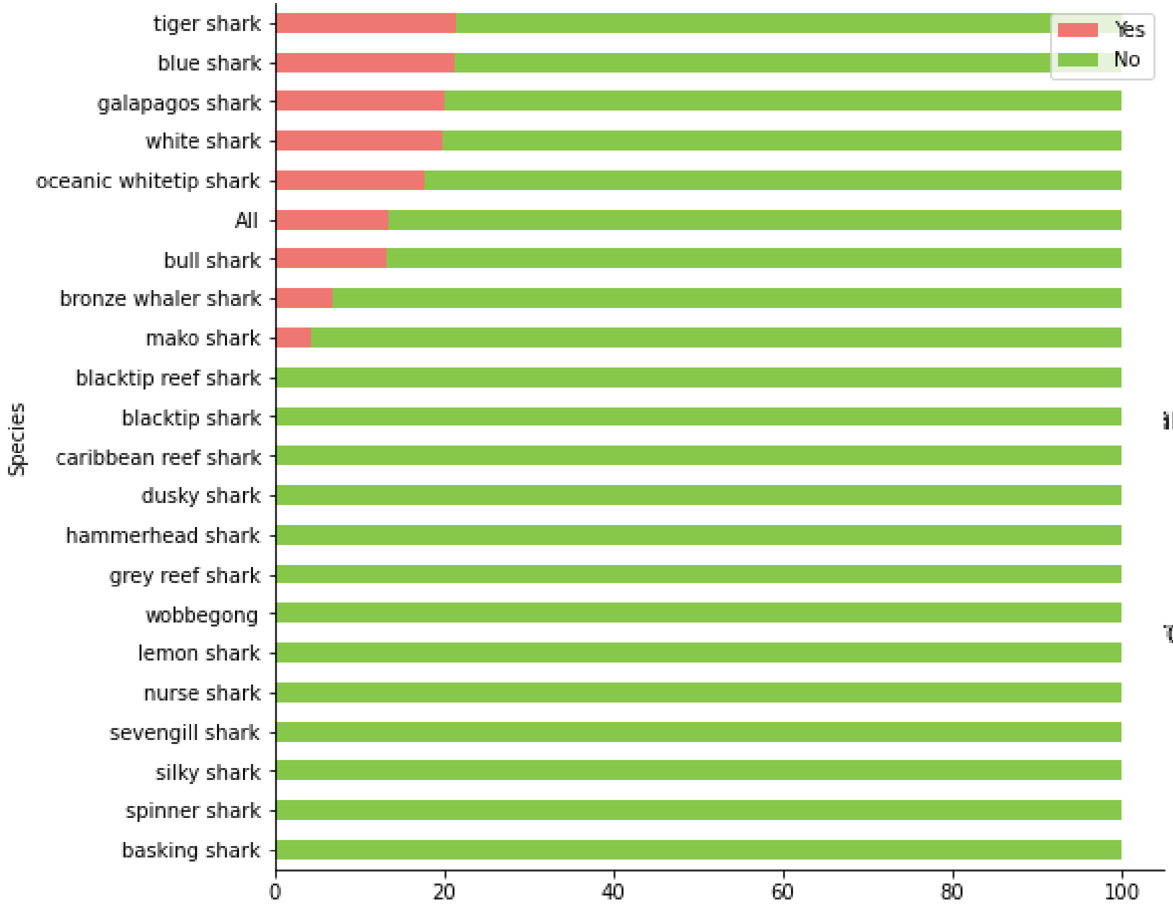


Fig.11 Fatality % for all Shark Species

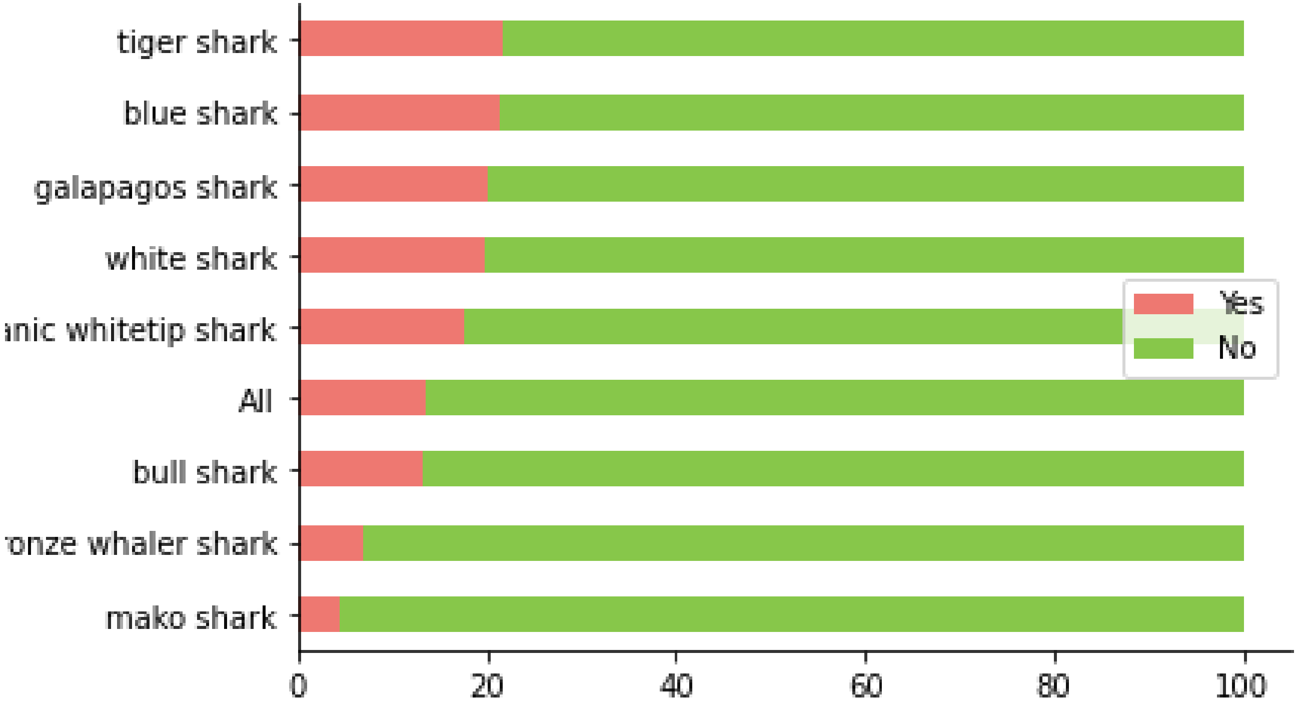


Fig.12 Fatality % for Shark Species w/fatality

Data Visualization

Attacks by Season

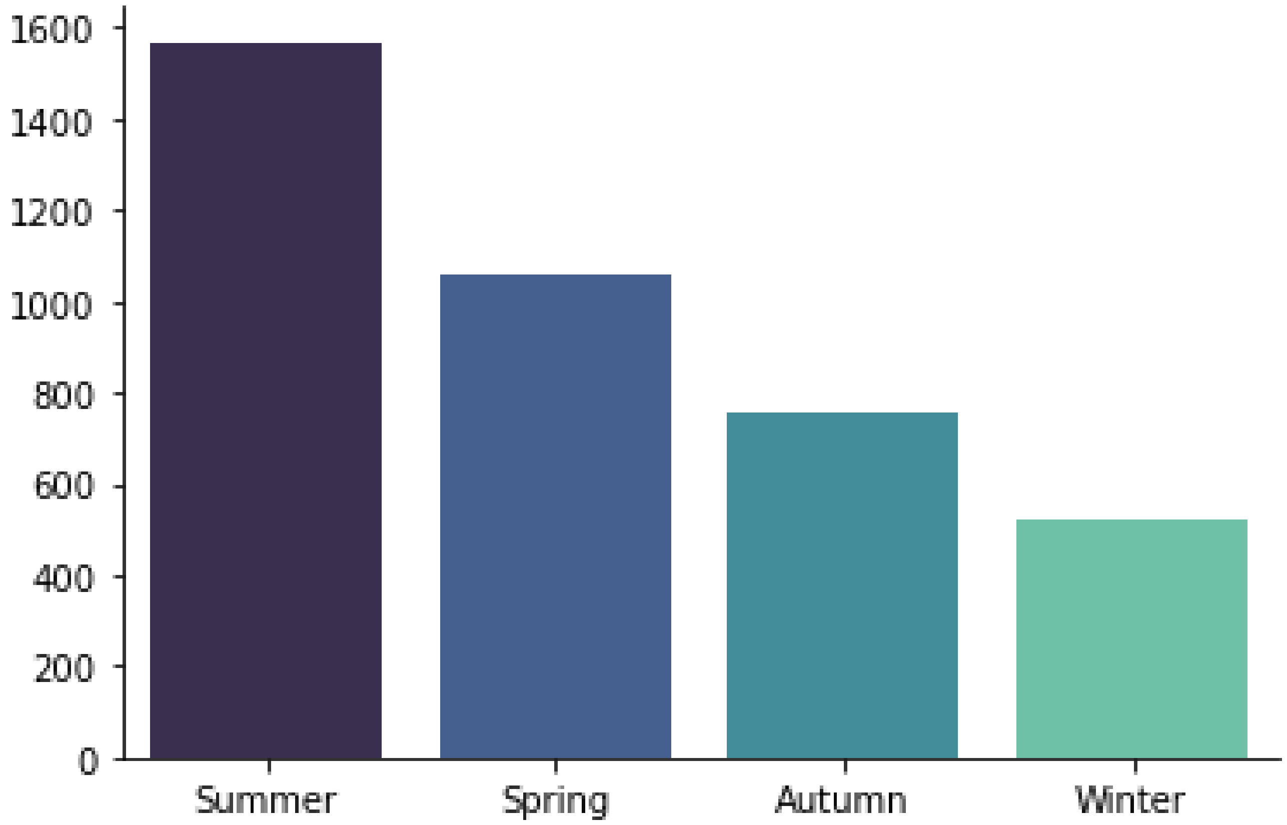


Fig.13 Total No. Attacks by Season

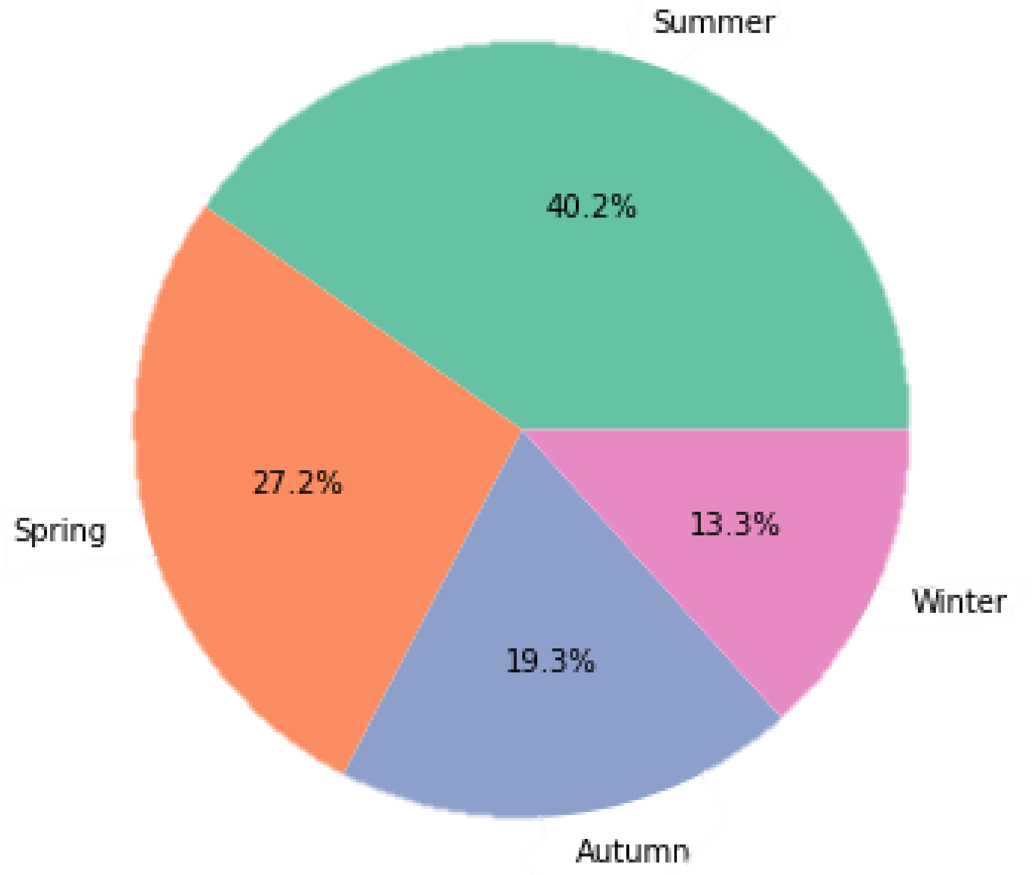


Fig.14 Total % Attacks by Season

Data Visualization

Fatality by Season

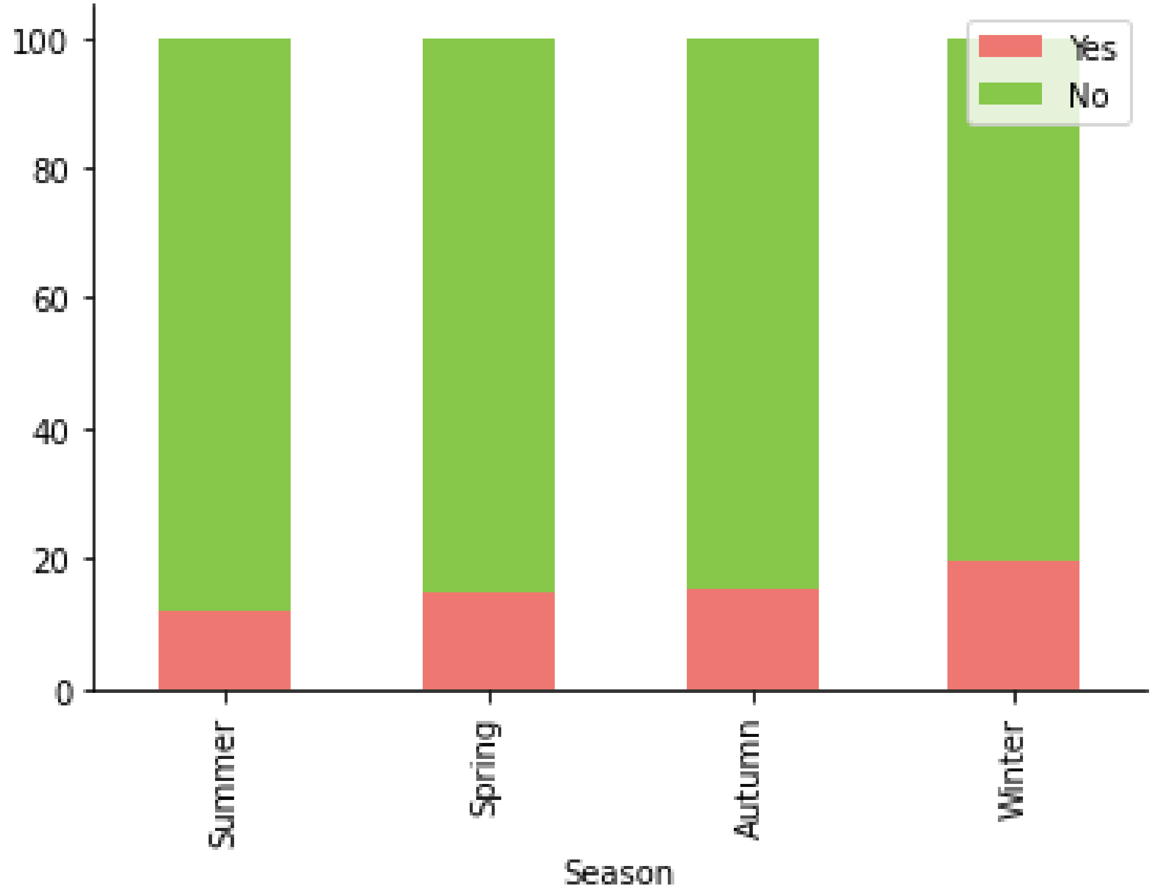
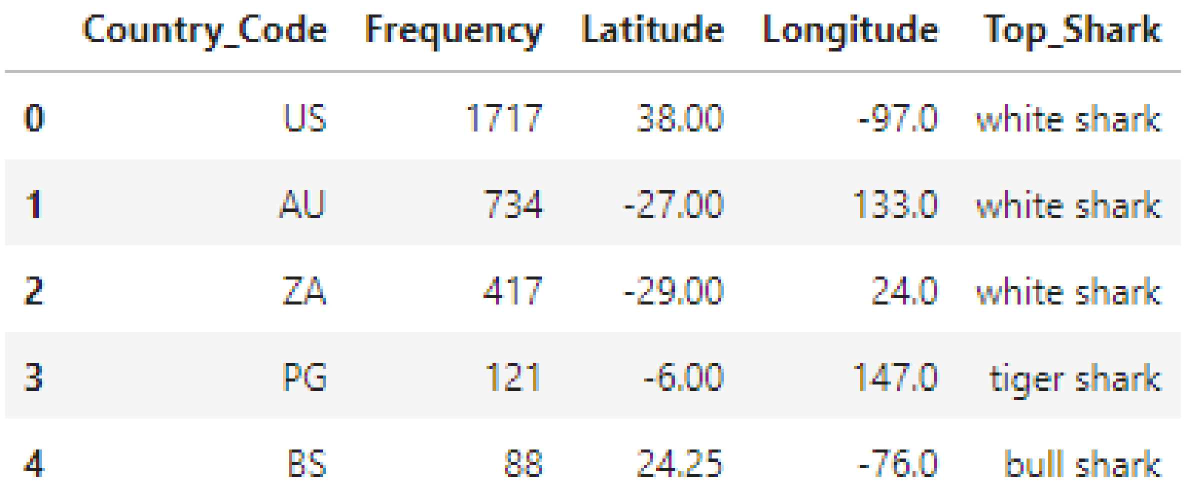


Fig.15 Fatality % by Season

Data Visualization

Attacks by Country



	Country_Code	Frequency	Latitude	Longitude	Top_Shark
0	US	1717	38.00	-97.0	white shark
1	AU	734	-27.00	133.0	white shark
2	ZA	417	-29.00	24.0	white shark
3	PG	121	-6.00	147.0	tiger shark
4	BS	88	24.25	-76.0	bull shark

Fig.16 Total No. Attacks by Country (Top 5)

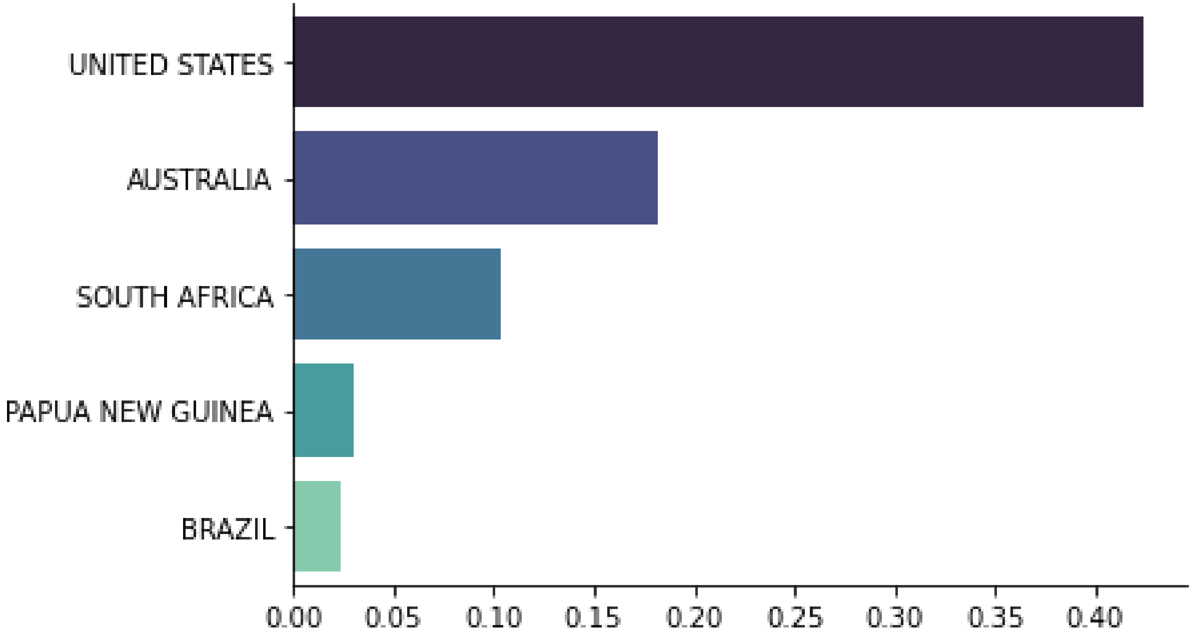
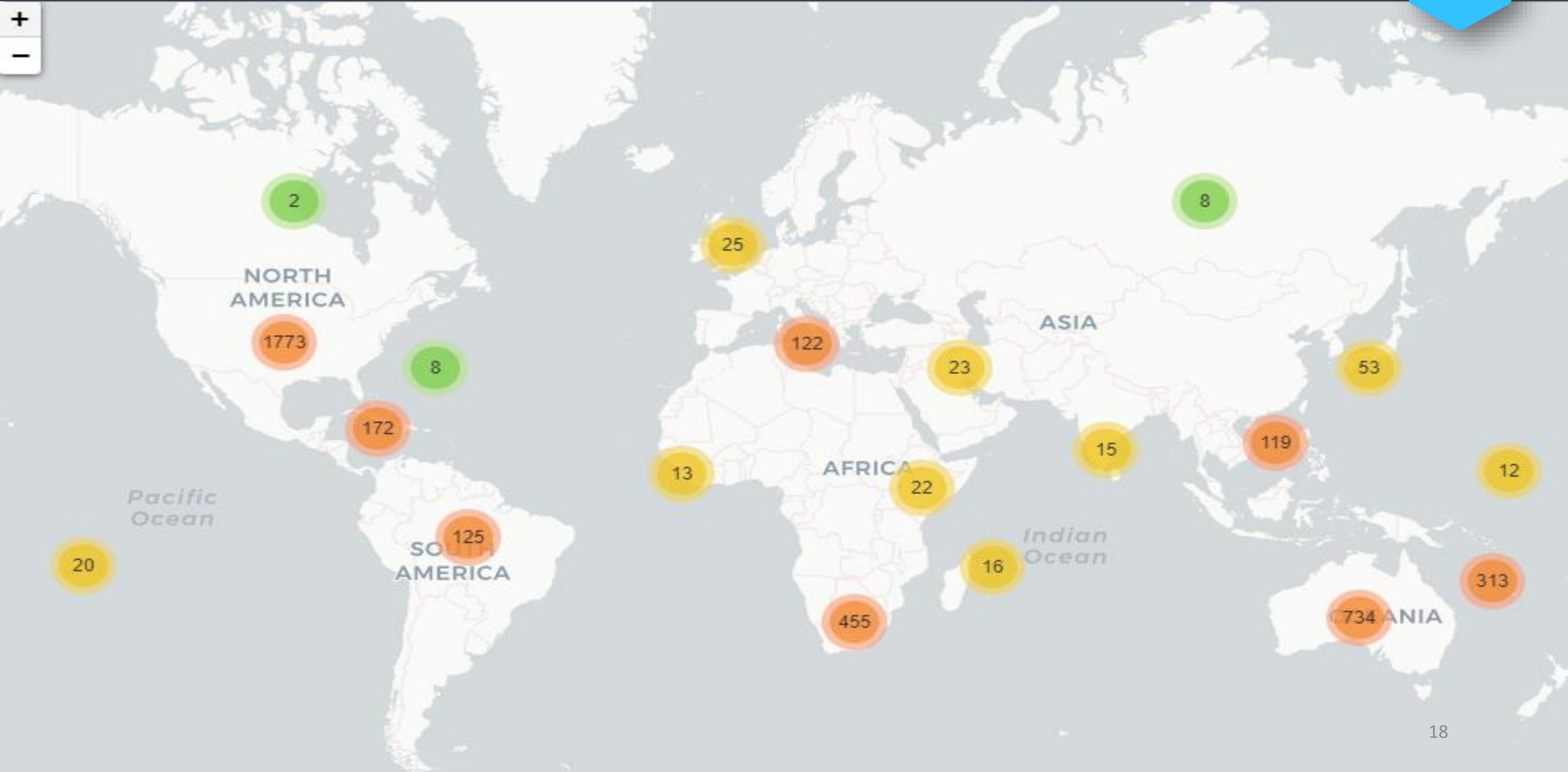
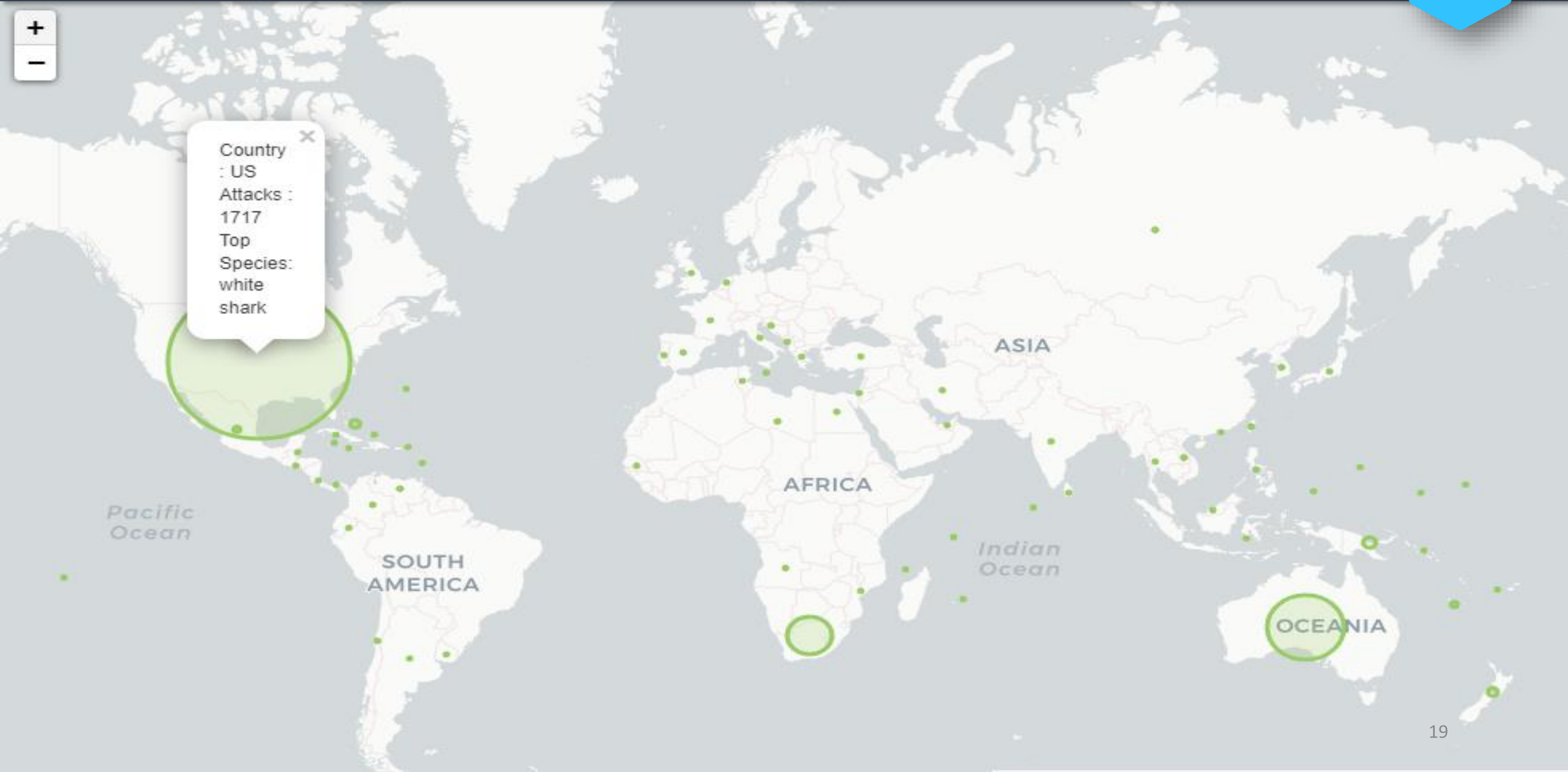


Fig.17 Total No. Attacks by Country (Top 5)

Interactive World Map



Interactive World Map



Country : US
Attacks : 1717
Top Species: white shark

Conclusions

After all the above, we have been able to extract following conclusions:

1. Surfing is the sport with the highest number of registered shark attacks (33.2%)
2. Despite this, Surfing is in the 7th out of 8 positions of activities that cause fatality.
3. Top 3 activities that cause fatalities are: sea disaster, swimming and snorkeling (activities where people do not take additional equipment).
4. Following Sea Disaster, unprovoked events are the most common type of event that causes fatality.
5. 40% of attacks occur to people between 15 and 25 years old.
6. 63% of attacks occur to people between 10 and 30 years old.
7. The majority (86.7%) of attacks occur in people of male gender. However, although most attacks are registered in male gender, the % fatality is almost identical for both men and women.
8. Only 15% of all shark attacks have been registered to cause fatality.
9. The shark species with highest number of attacks is the White shark
10. The most dangerous sharks, who attacked the most are: White shark, tiger shark, bull shark, nurse shark, bronze whaler shark. Except nurse shark, all of them caused fatalities.
11. 40.2% of the attacks occur in the summer season, followed by 27.2% in the spring season.
12. The top 3 countries with most registered shark attacks are: USA, Australia, South Africa.

Questions