W1 Project
Data cleaning & wrangling

Paula Hernández / Yu Ting Hu

21-May-2022

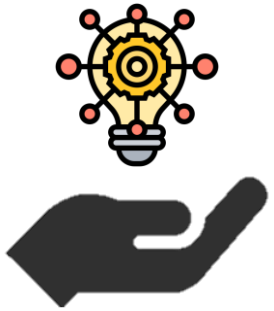
Project Overview

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

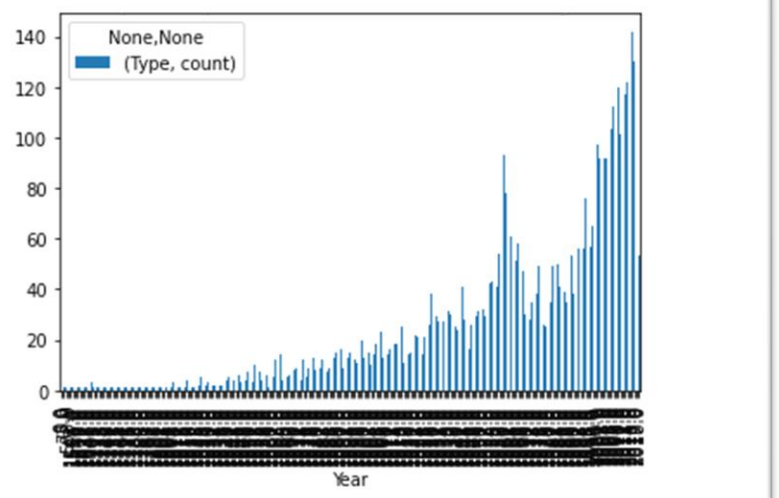
Exploring the Data

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.

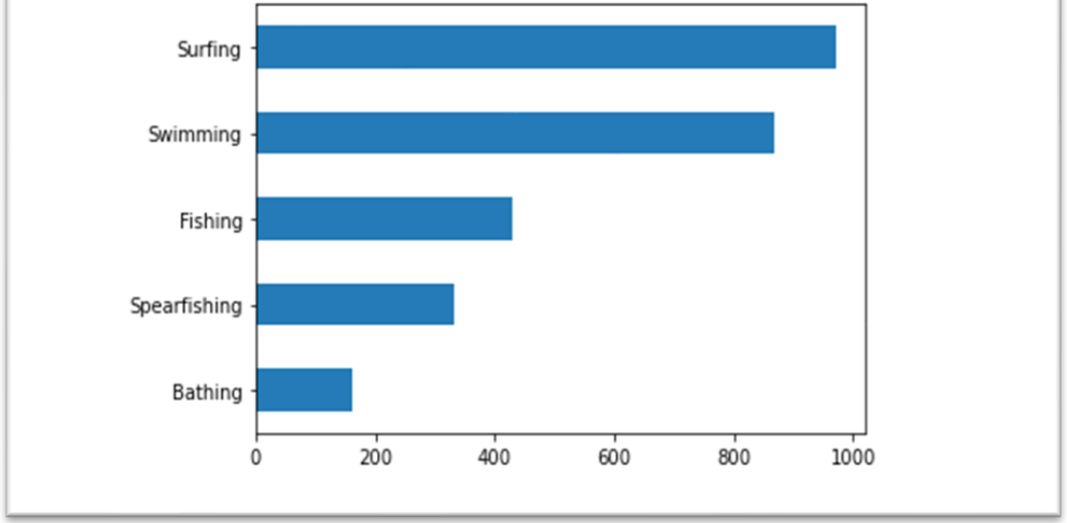
```
[22]: # Registers by Year

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

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




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

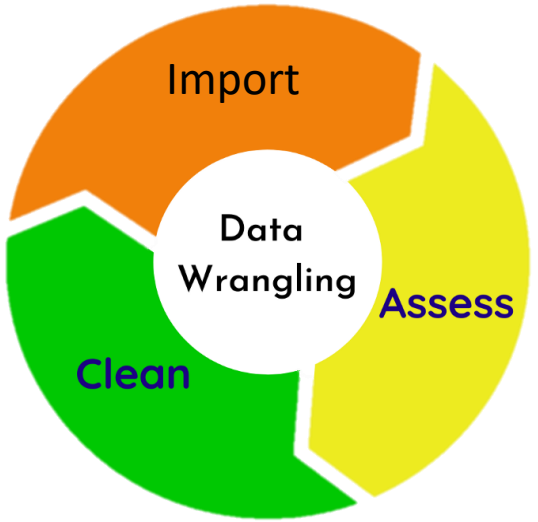


Cleaning and formatting the Data

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

The resulting DataFrame after cleaning has a total of 4120 row x 9 columns:

	Case Number	Year	Type	Country	Activity	Sex	Age	Fatal (Y/N)	Species
0	2018.06.25	2018.0	Boating	USA	Paddling	F	57	N	White shark
1	2018.06.18	2018.0	Unprovoked	USA	Standing	F	11	N	NaN
3	2018.06.08	2018.0	Unprovoked	AUSTRALIA	Surfing	M	NaN	N	2 m shark
4	2018.06.04	2018.0	Provoked	MEXICO	Free diving	M	NaN	N	Tiger shark, 3m
5	2018.06.03.b	2018.0	Unprovoked	AUSTRALIA	Kite surfing	M	NaN	N	NaN
...	...	...	...	...	...	...	...	...	...
4493	1950.00.00.e	1950.0	Unprovoked	GREECE	Swimming	NaN	NaN	Y	NaN
4494	1950.00.00.d	1950.0	Unprovoked	SINGAPORE	Diving for coins	M	NaN	Y	NaN
4495	1950.00.00.c	1950.0	Unprovoked	NEW CALEDONIA	Spearfishing, but walking carrying fish on end...	M	NaN	N	NaN
4496	1950.00.00.b	1950.0	Unprovoked	NEW CALEDONIA	Helmet diving, collecting trochus shell	M	NaN	N	NaN
4497	1950.00.00.a	1950.0	Unprovoked	FIJI	NaN	M	NaN	N	NaN

4120 rows × 9 columns

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:

- ✓ Format column names and delete blank spaces
- ✓ Format dates (Case_Number and Year)
- ✓ Get month in order to know the season
- ✓ Format type of attack
- ✓ Format country names and convert to standardized country codes per ISO-3166
- ✓ Get coordinates (latitude and longitude) from standardized country codes
- ✓ Format type of Activity
- ✓ Format data of Sex and Age
- ✓ Format Fatality data
- ✓ Format Shark Species

Data Visualization

Attacks by Activity

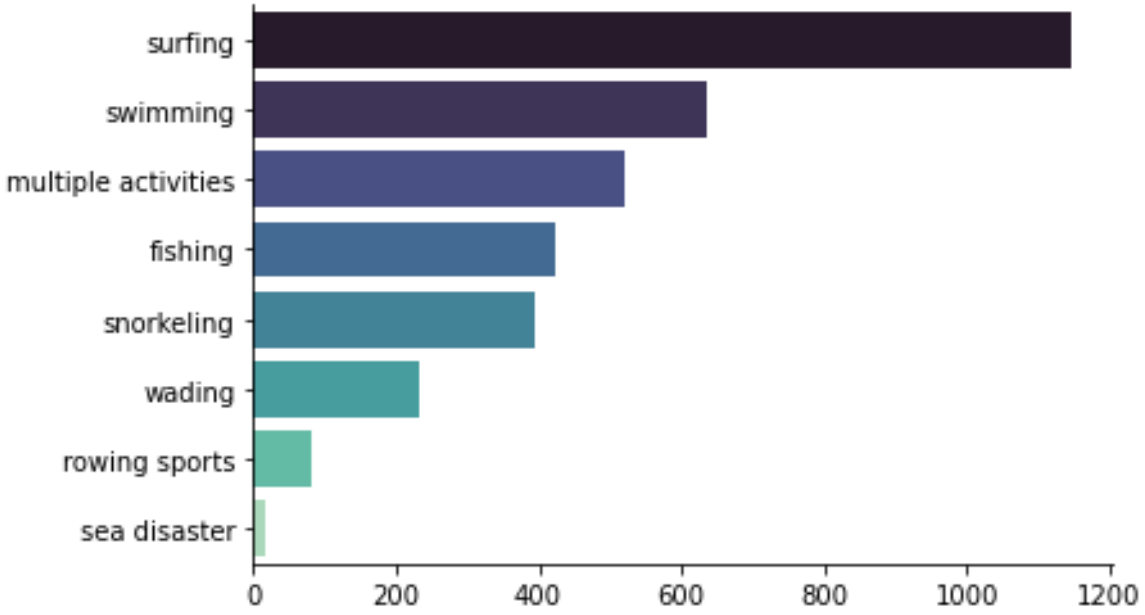


Fig.1 Total number of Attacks by Activity

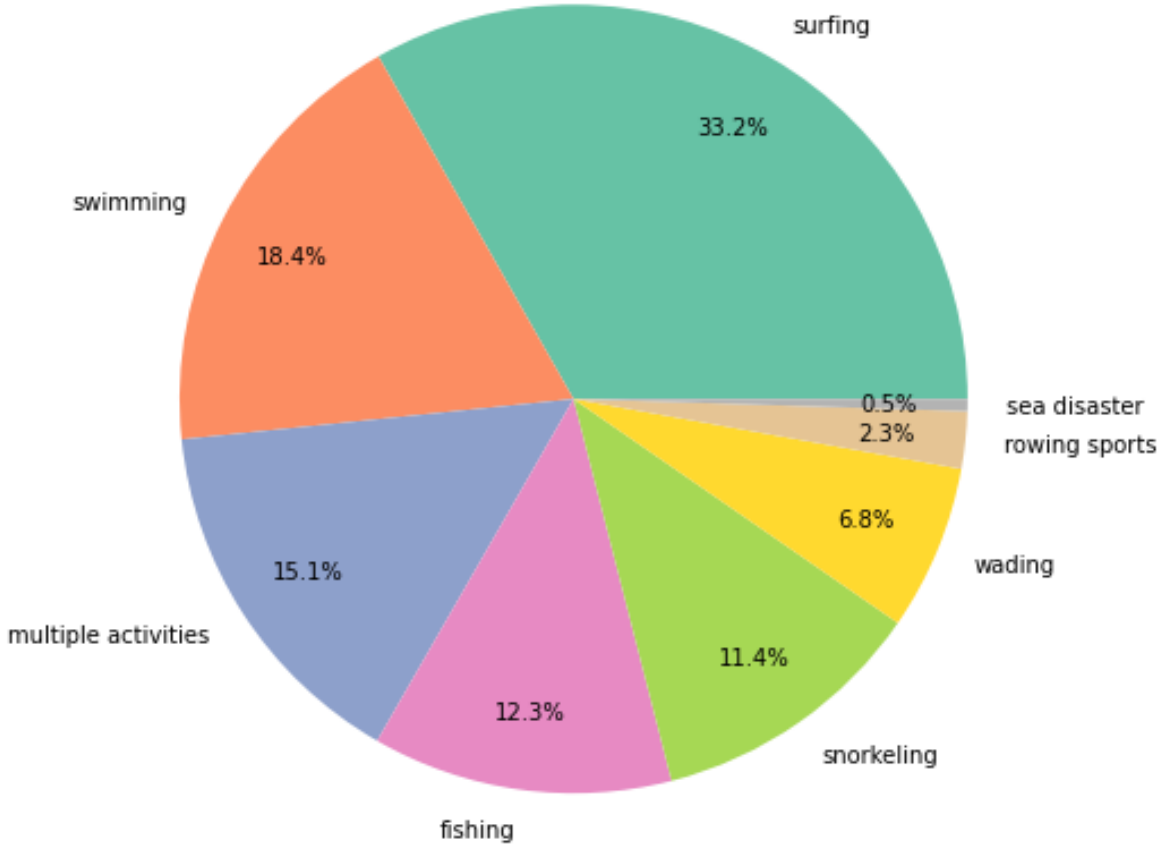


Fig.2 Percentage Attacks by Activity

Data Visualization

Fatality by Activity

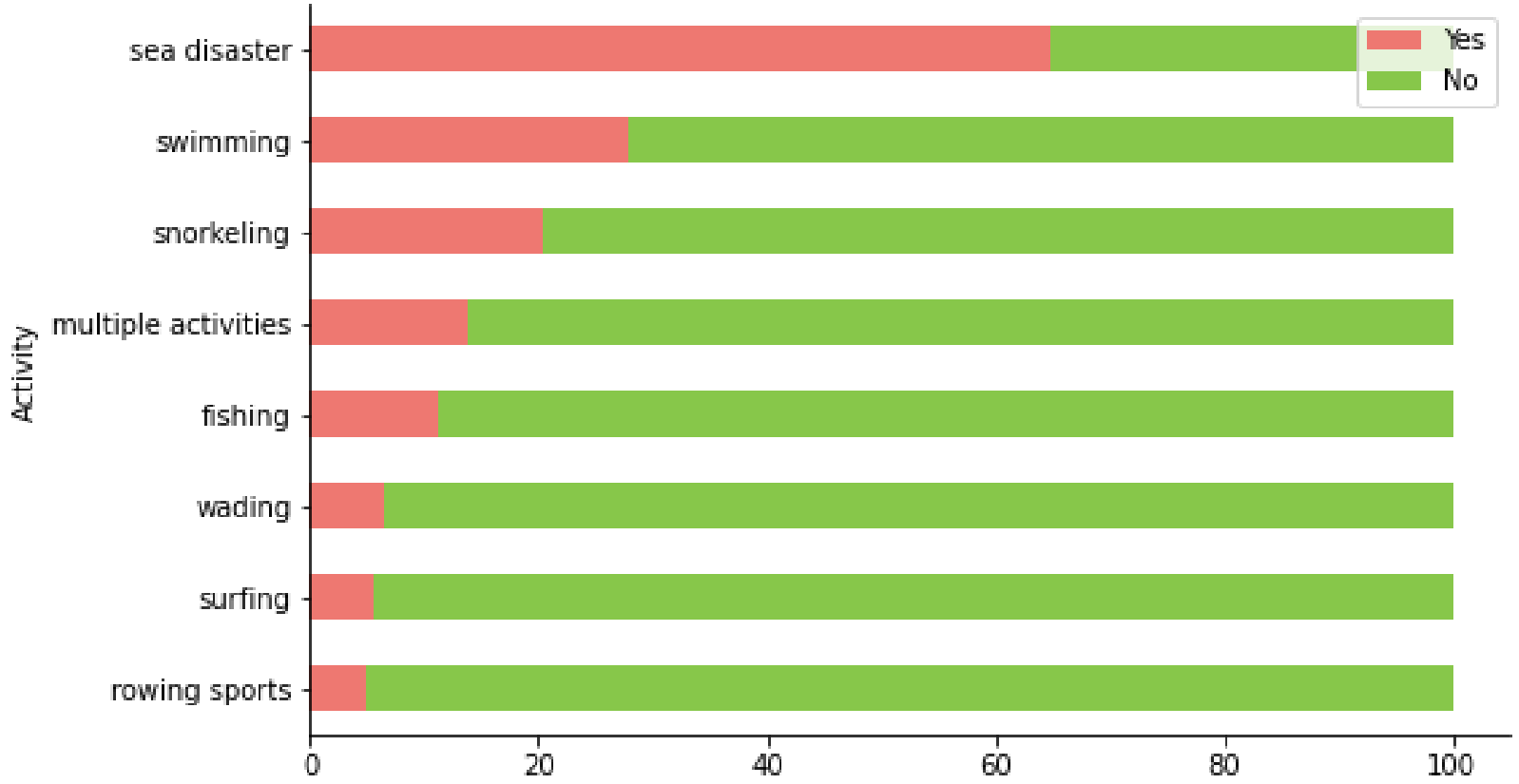


Fig.3 Fatality % per Activity

Data Visualization

Fatality by Type of Event

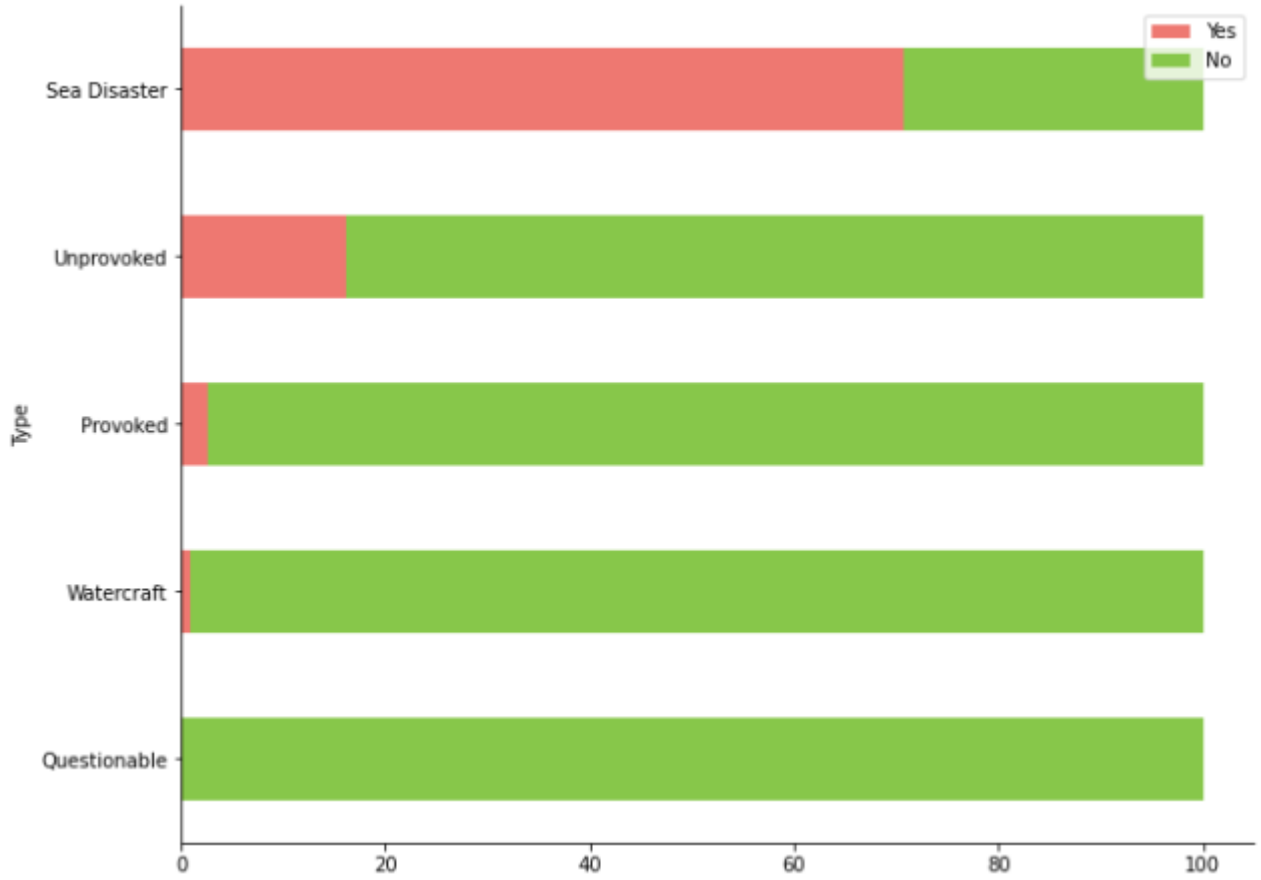


Fig.4 Fatality % per Type of Event

Data Visualization

Social-demographic Analysis

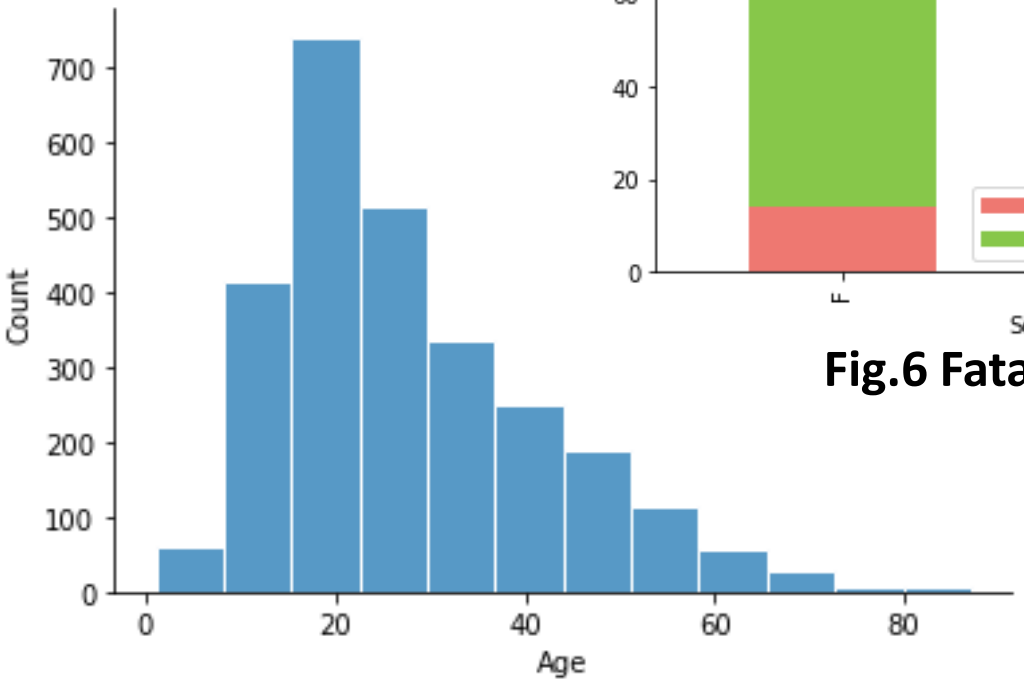


Fig.5 Attacks by Age Histogram

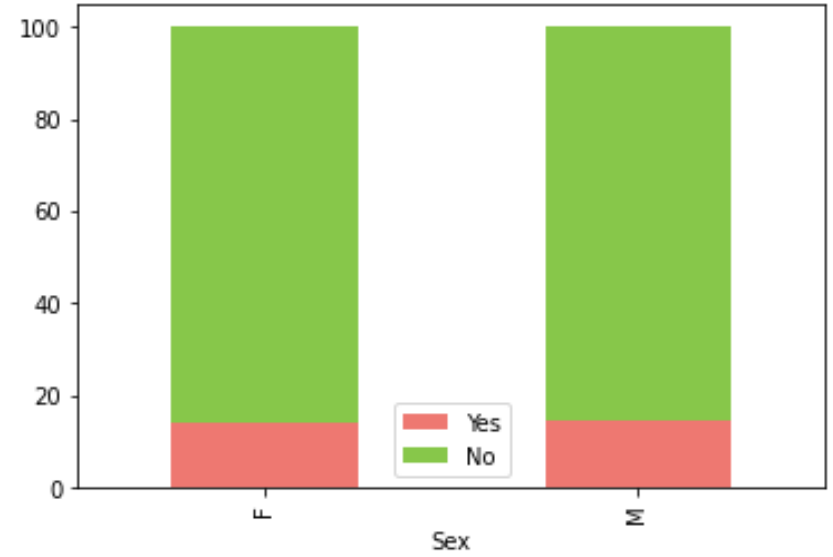


Fig.6 Fatality by Sex

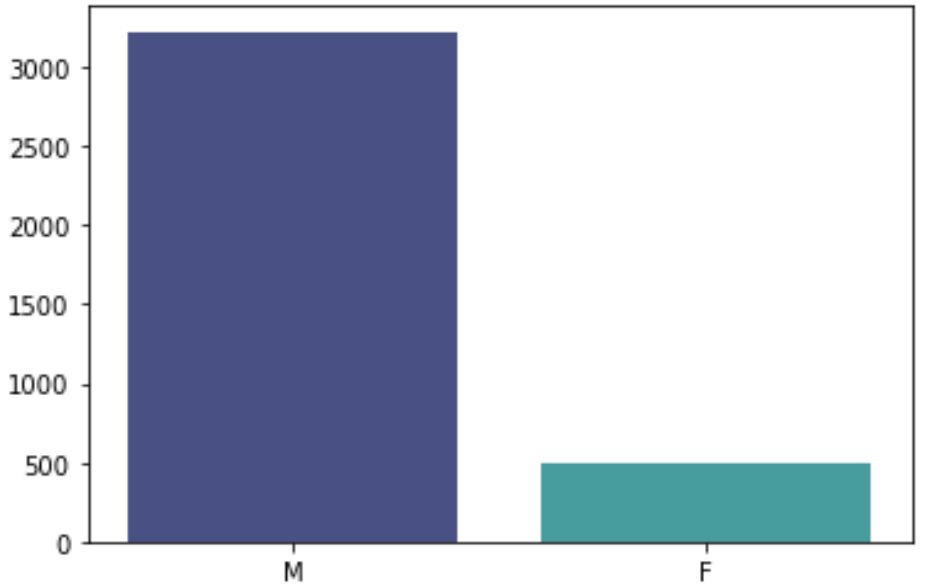


Fig.7 Total attacks by Sex

Data Visualization

Attacks by Shark Species

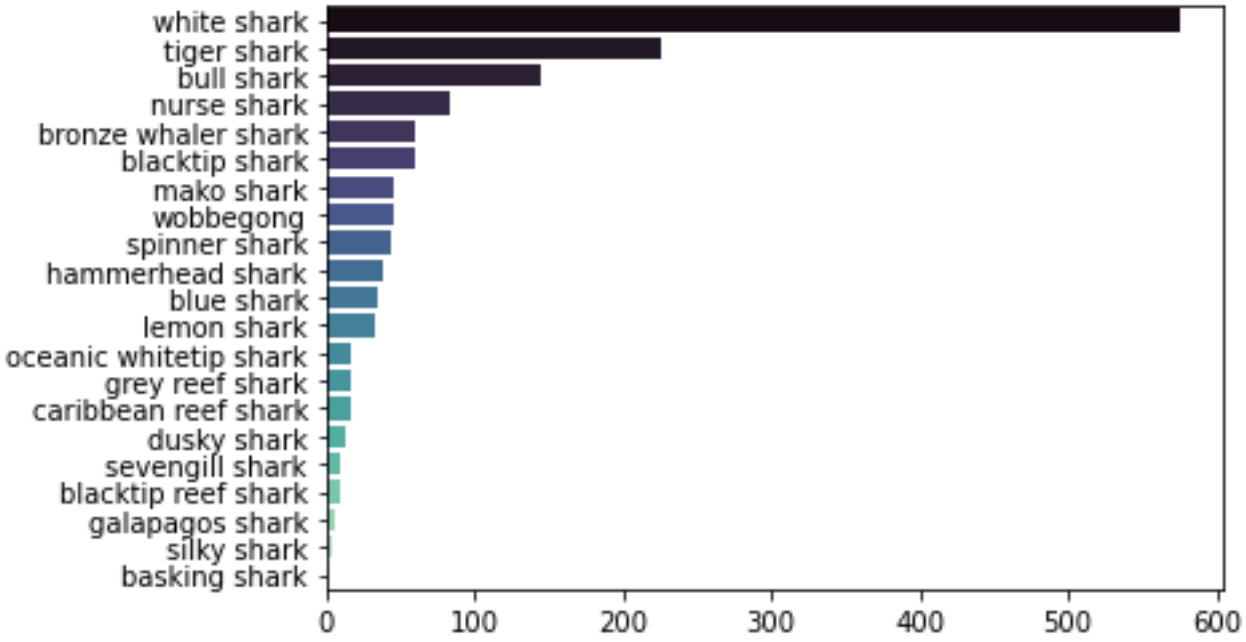


Fig.8 Total no. of Attacks by Shark

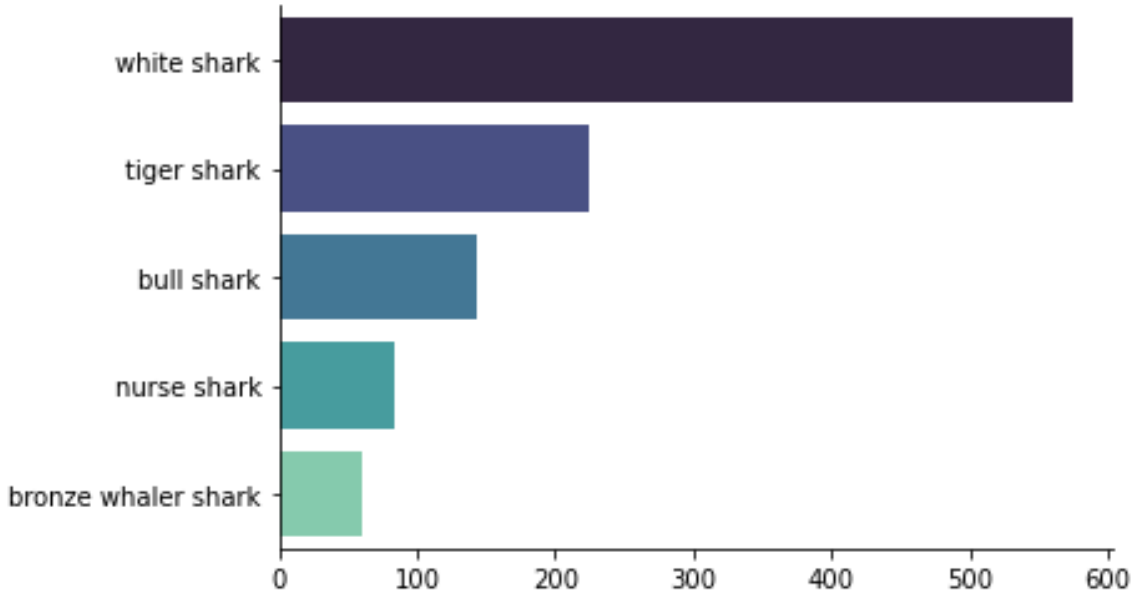


Fig.9 Total no. of Attacks by Shark (Top 5)

Data Visualization

Fatality by Shark Species

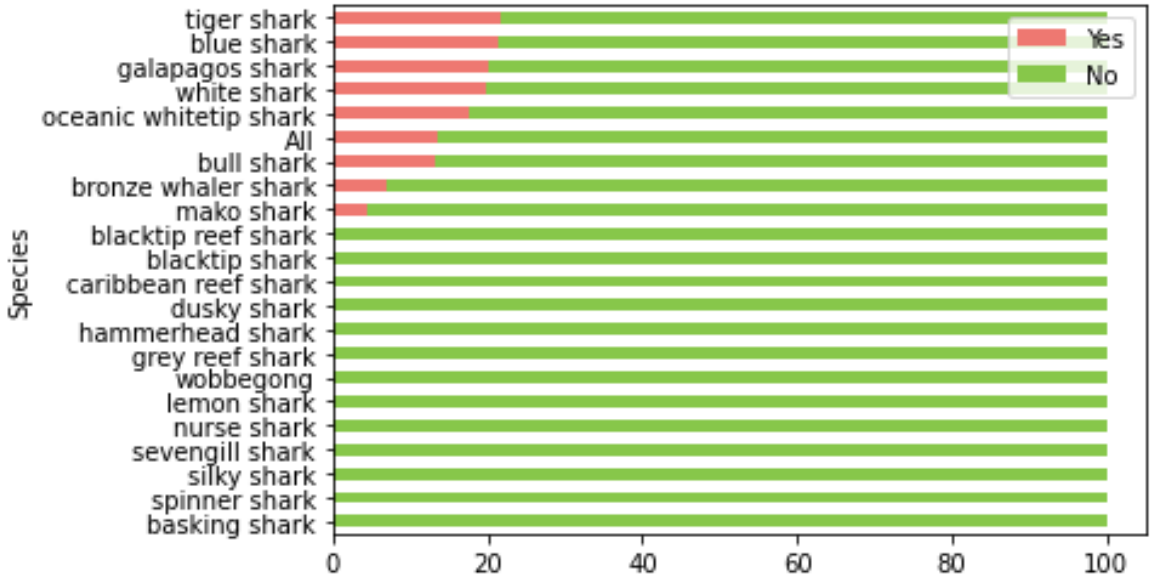


Fig.10 Fatality % for all Shark Species

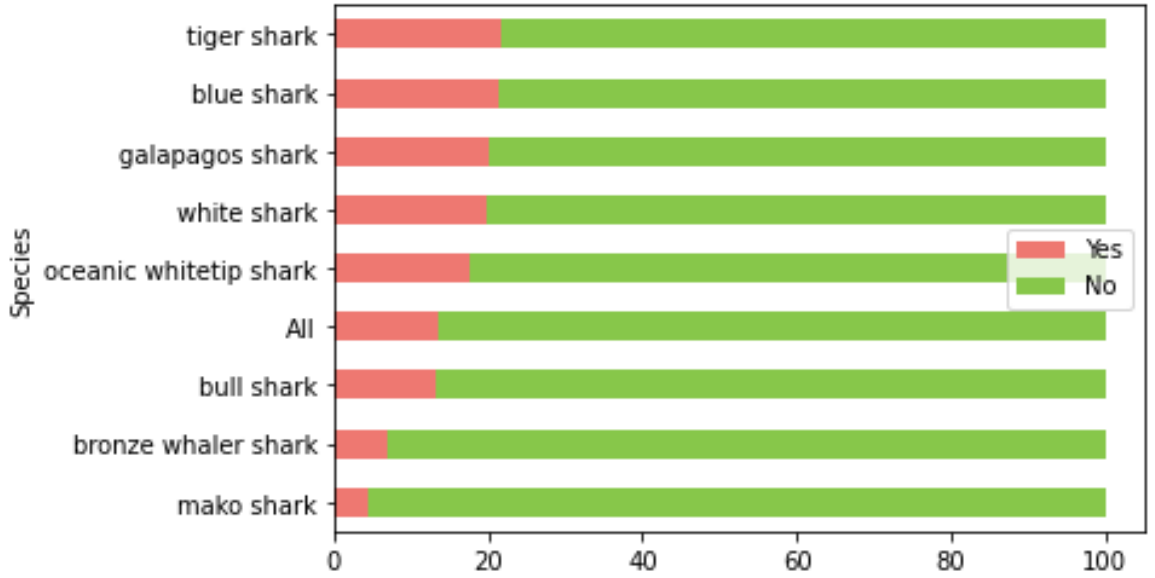


Fig.11 Fatality % for Shark Species w/fatality

Data Visualization

Attacks by Season

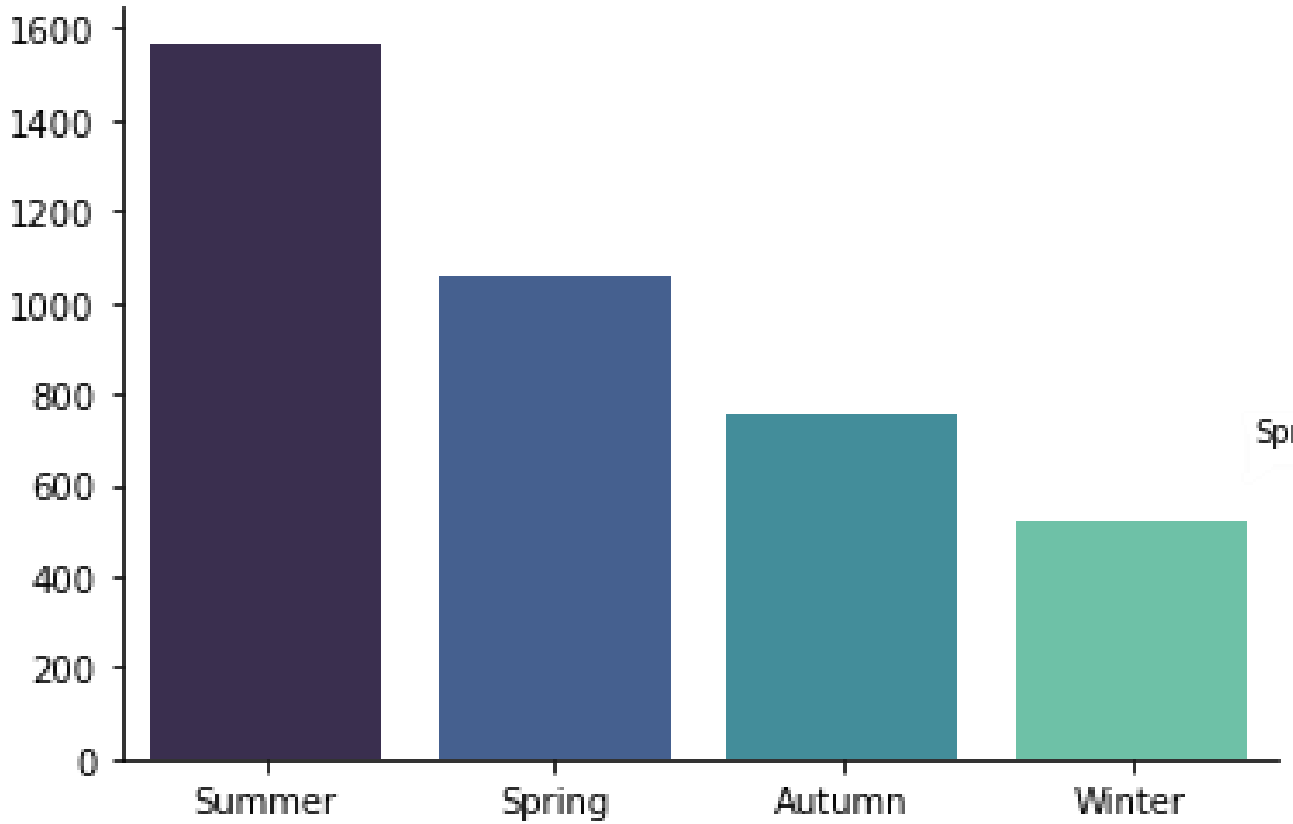


Fig.12 Total No. Attacks by Season

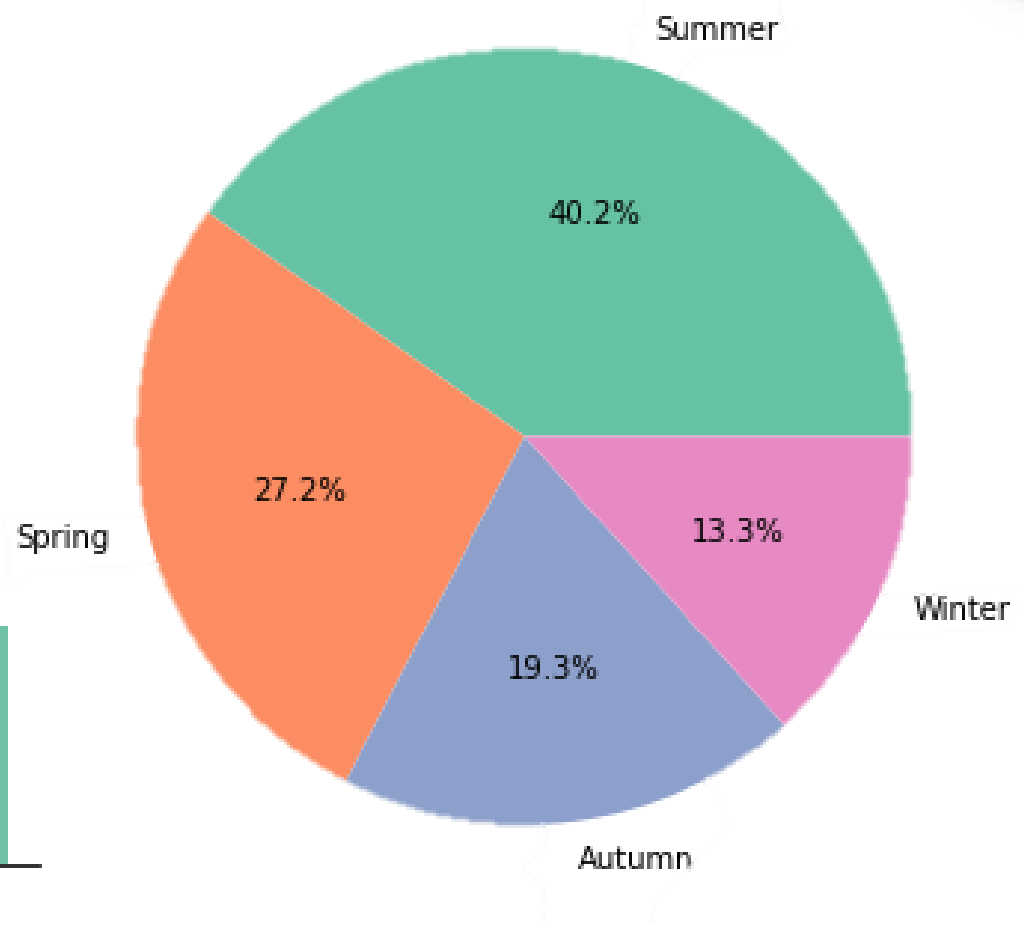


Fig.13 Total % Attacks by Season

Data Visualization

Fatality by Season

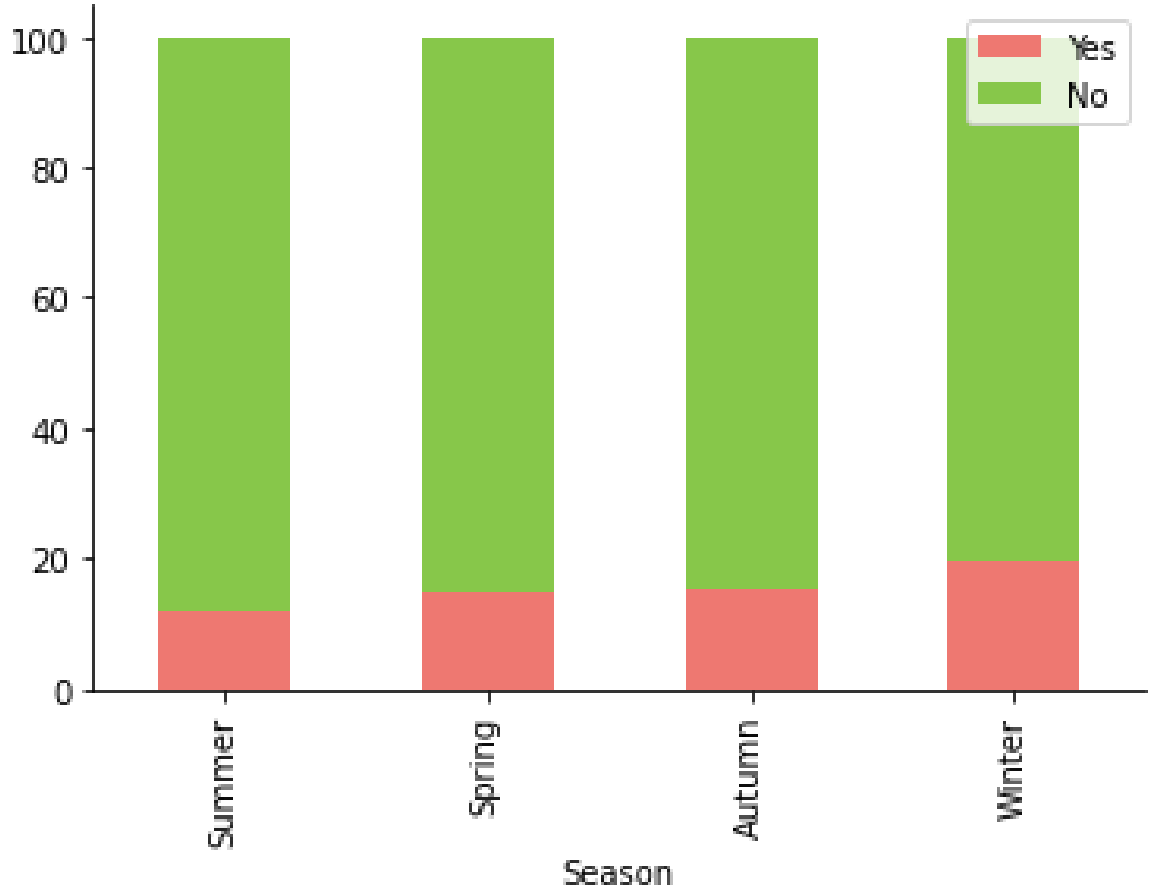


Fig.14 Fatality % by Season

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.15 Total No. Attacks by Country (Top 5)

Interactive World Map

Interactive World Map

Conclusions

- Surfing is the sport with the highest probability of shark attack (33.2%)
- Despite this, Surfing is in the 7th position of activity that causes fatality.
- Top 3 activities that cause fatalities are: sea disaster, swimming and snorkelling (activities where people do not take without additional equipment).
- Following Sea Disaster, unprovoked events are the most common type of event that causes fatality.
- The maximum number of attacks is in people between 15 and 25 years old.
- The majority (>80%) 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.
- Only 18% of all shark species have been registered to cause fatality.
- The shark species with highest number of attacks is the White shark (40%).
- 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.
- Almost half (40.2%) of the attacks occur in the summer season, followed by 27.2% in the spring season.
- The top3 countries having shark attacks are: USA, Australia, South Africa.

Questions