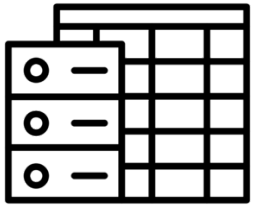
Seasafe staffing agency

Tania Zakowski
Houleye Anne
Lucie Stenger
Shashi Bhushan Singh

Project Overview

Description of the dataset and hypothesis:



- Shark attacks over different territories
- how, when, and who
- media covering

“There are more fatal shark attacks in Florida”

Our cleaning process:

1. Exploring dataset (number of unique values per column, general quality)
2. Defining our hypothesis, and deciding to focus on the USA data, and only keep specific columns
3. Cleaning the columns, re-formatting values, removing rows with outliers, adding new columns for the purpose of our analysis (date & season) using various data cleaning methods including RegEx.

Data Wrangling and Cleaning

Significant data cleaning challenges encountered:

	Date	Year	State	Location	Activity	Sex	Age	Injury	Species	New_Date	New_Date2	Season	Year_with_Season
1	04 Mar 2024	2024.0	Hawaii	Old Man's, Waikiki	surfing	M	NaN	No injury, shark bit surfboard	8' shark	04 Mar 2024	04 Mar 2024	spring	2024
2	02 Mar-2024	2024.0	Hawaii	Rainbows, Oahu	swimming	F	11	Lacerations to left foot	4' shark	02 Mar 2024	02 Mar 2024	spring	2024
10	30 Dec-2023	2023.0	Hawaii	Baby Beach, Maui	surfing	M	39	FATAL	NaN	30 Dec 2023	30 Dec 2023	winter	2023
24	05 Nov-2023	2023.0	Florida	Juno Beach, Palm Beach County	swimming	M	66	Lacerations to right forearm	NaN	05 Nov 2023	05 Nov 2023	autumn	2023
29	25 Oct 2023	2023.0	Hawaii	Pua'ena Point, Haleiwa, Oahu	surfing	M	30	Bite to right thigh	8' shark	25 Oct 2023	25 Oct 2023	autumn	2023
33	15 Oct 2023	2023.0	Hawaii	Hanalei Bay, Kauai	surfing	M	50	Left leg and hand injured	10' tiger shark	15 Oct 2023	15 Oct 2023	autumn	2023

- Many values were text entries making it difficult to categorize them and therefore analyse them. In particular, the activity the victim was doing when the attack occurred, the date of the attack, and the shark species.

Data Wrangling and Cleaning

Explain how we resolved these challenges:

- Using RegEx (re.search, re.match) and using replace function when a certain value is included in the text limiting data referring to time by using the column Date and creating new variables based on it to work on the same rows. Ex:

```
#Injury cleaning
import re

def categorize_injury(injury):
    injury = str(injury).strip().lower()

    if re.search(r'fatal|died|death|bit him in half|knocked over', injury):
        return 'Fatal'

    elif re.search(r'minor|small|superficial|laceration|bite|abrasion', injury):
        return 'Minor'

    elif re.search(r'no injury|uninjured|not injured', injury):
        return 'No Injury'

    else:
        return 'Unknown'


shark_attacks_USA_df.loc[:, 'Injury']  = shark_attacks_USA_df['Injury'].apply(categorize_injury)
print(shark_attacks_USA_df.Injury.unique())
```

Exploratory Data Analysis

Methods used:

a. value.count, describe(),
b. groupby
c. scatterplot
d. library import for analysis and visualisation (numpy, seaborn, matpotlib)

Insights and pattern found:

e. Most of the attacks happened during Summer and in Florida
f. Most of the person attacked suffered minor injuries

Major obstacle

Biggest obstacle and difficulties:

- Trying to implement a new type of data, i.e. a time serie to get timesteps logged and follow the evolution of datas through time.
- The cleaning of columns and null values could be complicated because the data type must be taken into account

What we learned from it and how it influenced your project:

- Dependencies resulting of the working conditions (tools and libraries/packages) especially when all collaborators don’t have similar settings

Conclusion and Insights

Activity	Injury	Activity
fishing	Fatal	28
	Minor	72
	No Injury	68
	Unknown	174
surfing	Fatal	15
	Minor	252
	No Injury	90
	Unknown	390
swimming	Fatal	69
	Minor	131
	No Injury	8
	Unknown	220
wading	Fatal	3
	Minor	57
	No Injury	3
	Unknown	67

- Our initial hypothesis is supported by the very large variability of the number of attacks depending on the location (particularly Florida), a seasonal effect can be seen in Florida while not in Hawaii. We observed more fatal attacks in Florida.

- Attacks were found for many categories of activity, but their number is largely dominated by surfing, followed by swimming. Nonetheless, attacked were more likely to be deadly when swimming than surfing.

- Potential implications of our findings: implementation of timestamp would be useful for our company, checking if adjustment in staff have meaningful impact and if global warming impacts the location of attacks.

Thank you!
Seasafe staffing agency

Tania Zakowski
Houleye Anne
Lucie Stenger
Shashi Bhushan Singh