QUEST 2:
DATA CLEANING

Team : Jaws
Beata, Samer, Sirine, Pierre

Contents

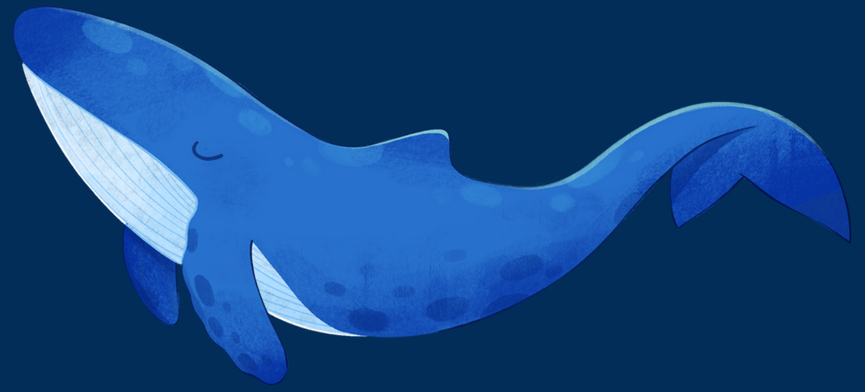
Project Overview

Ressource: open data dataset from "*sharkattackfile.net*"

Problem Statement & Hypothesis:

As shark attacking is a major & growing issue, we see the opportunity with that ressource to help us understand the phenomenon and provide for different publics a decision-making tool linked to shark attacks

Target Solution:



Decision making tool for 2 types of potential customers:

- Tourists for their trip planification
  - example: if I want to do surfing in August, which place to avoid?
- Businesses & Local Authorities for their investment planning
  - example: Increased safety checks durign specific months in Florida

MVP:

For a specific activity and period given, return a first selection of proposed safe spots and/or spots to avoid, with key insights related to help in the choice of destination.

Original Dataset

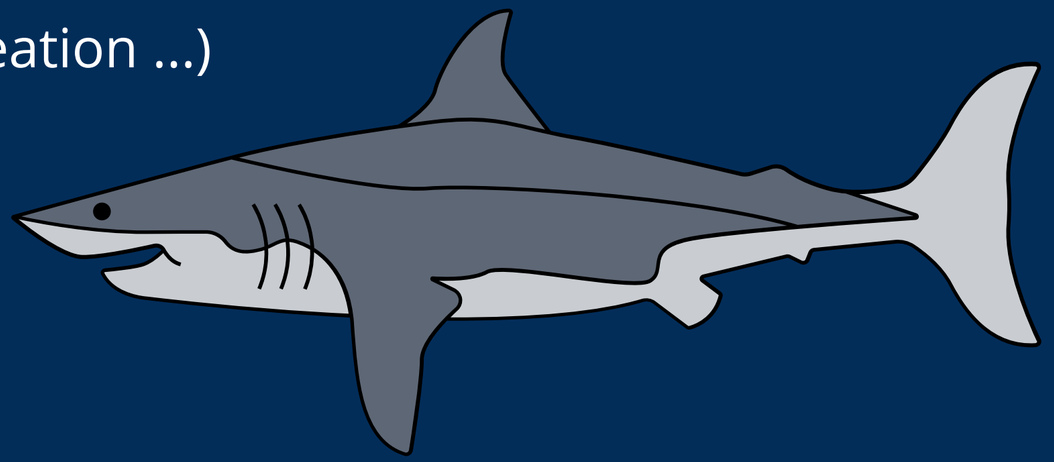
exampl

- size_shape: (6947, 23)
- columns: 'Date', 'Year', 'Type', 'Country', 'State','Location','Activity','Name','Sex','Age', 'Injury','Unnamed: 11', 'Time', 'Species ','Source', ‘pdf','href formula','href','Case Number', 'Case Number.1', 'original order', 'Unnamed: 21', 'Unnamed: 22'
- Early on we had to make a choice and drop useless columns
- The quality of the data was of the remaining columns was complicated due to bad managment

Data Cleaning

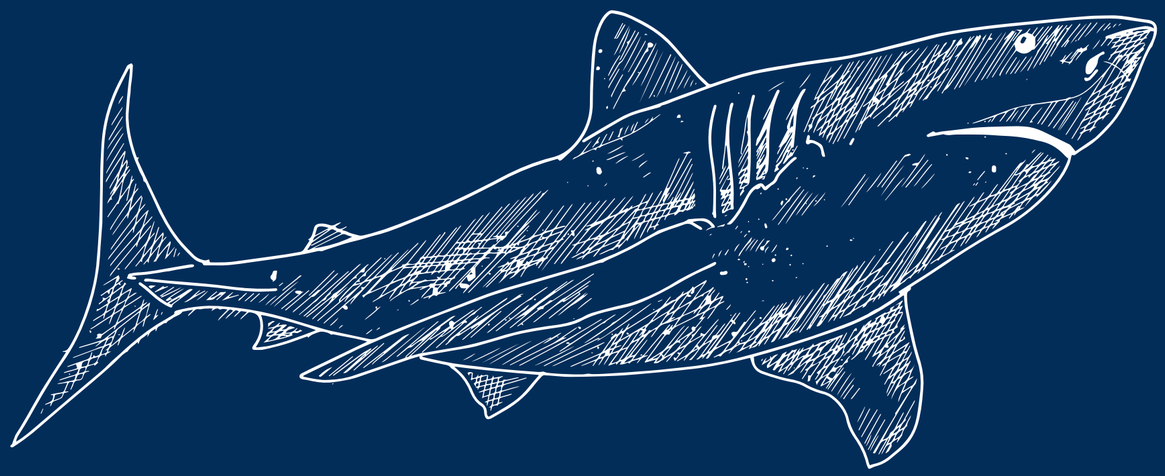
The values

- Understanding the data
- If necessary re-structure the data (category creation ...)
- Change data type if needed
- Check if values are clean :
  - Regex x in [A-Za-z]+
  - List x in [”A”, “B”, “C”]
  - Value x == “A”



Exemples

set_age_group



- Problem: the data
  - '18 months', '!2', '37, 67, 35, 27, ? & 27', ‘20's’,'31 or 33', 69, 'elderly'
- Solution: create age group to collect, store data
- How :
  - isinstance(x, int) : use of comparaison tool
  - isinstance(x, str)
    - creation of list with the data that fits into the category
    - checking the list contains x

Exemples

date

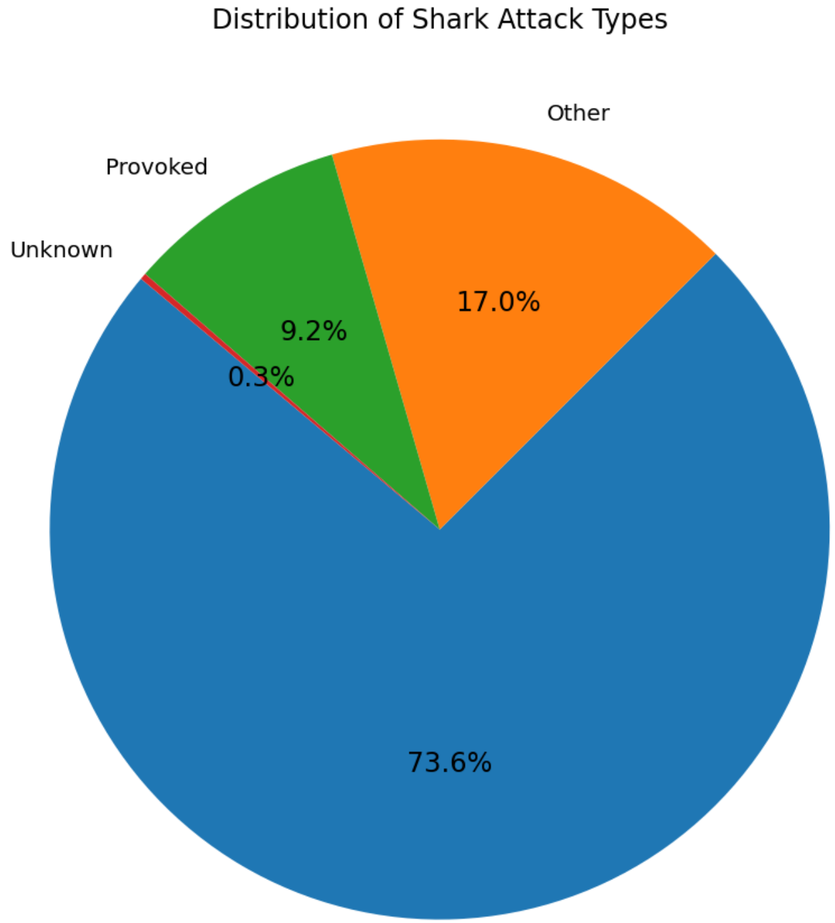
- Problem: the data
  - Date: 08 Dec-2023, Reported 02 Nov-2023, Jan-2023, 1999, Summer of 1981, 1980s
  - Year: 2023.0, ...
-
- Solution: create new column “Extracted_date” with understandable syntax

- How :
  - Fix some typos in data based on information in next rows
  - year: use of regex to catch the data and then join them
  - month: use dictionnary to rename

Exploratory Data Analysis

General Statistics Example

Out[8]:



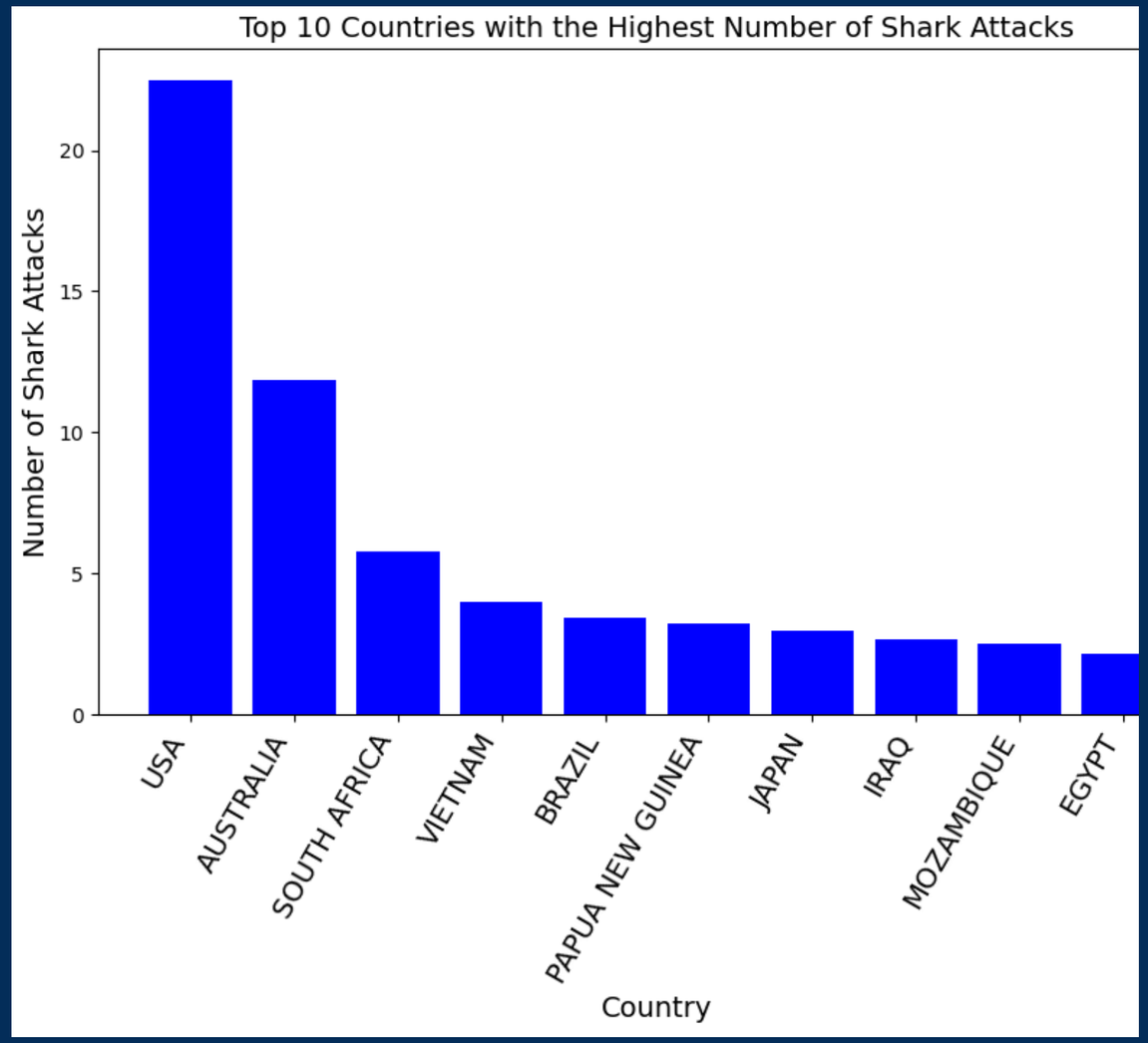
Unprovoked

SURVIVED ~87 %

FATAL ~ 10%

Exploratory Data Analysis

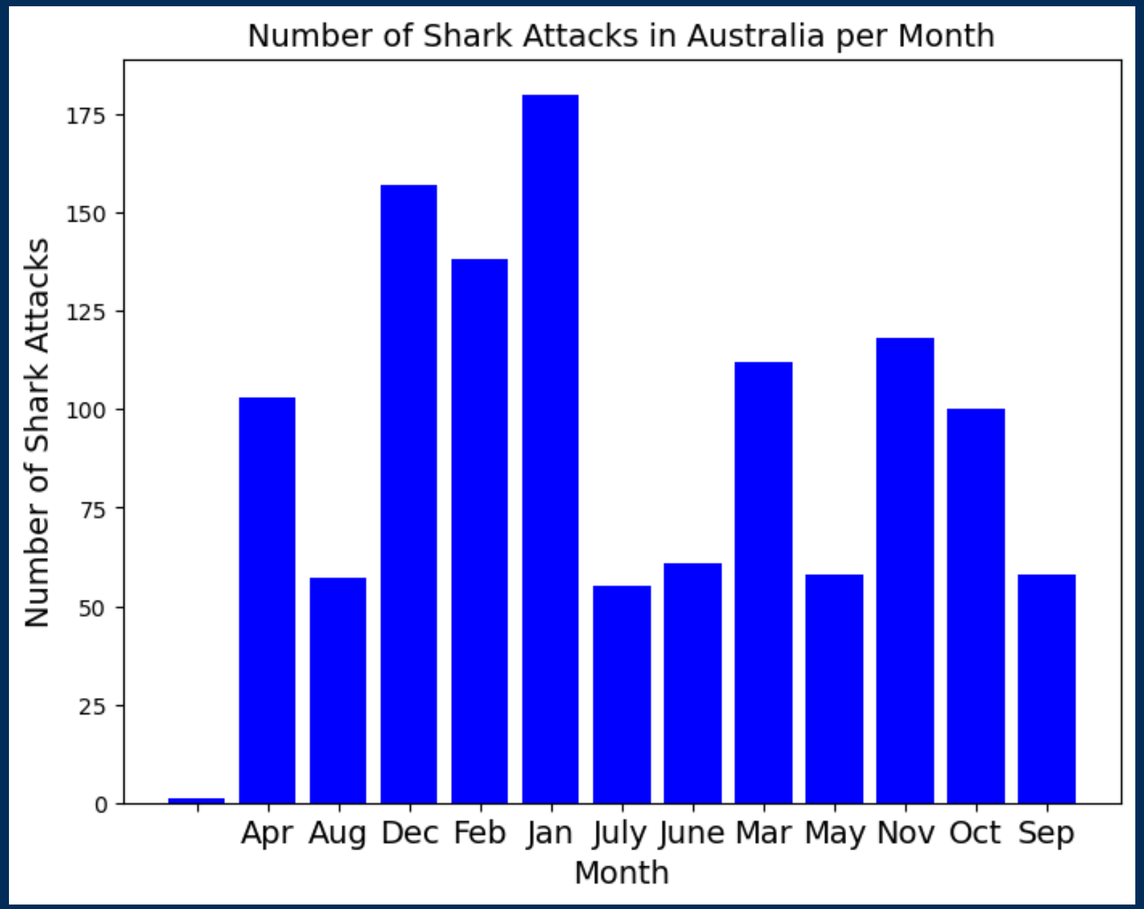
MVP related Insights

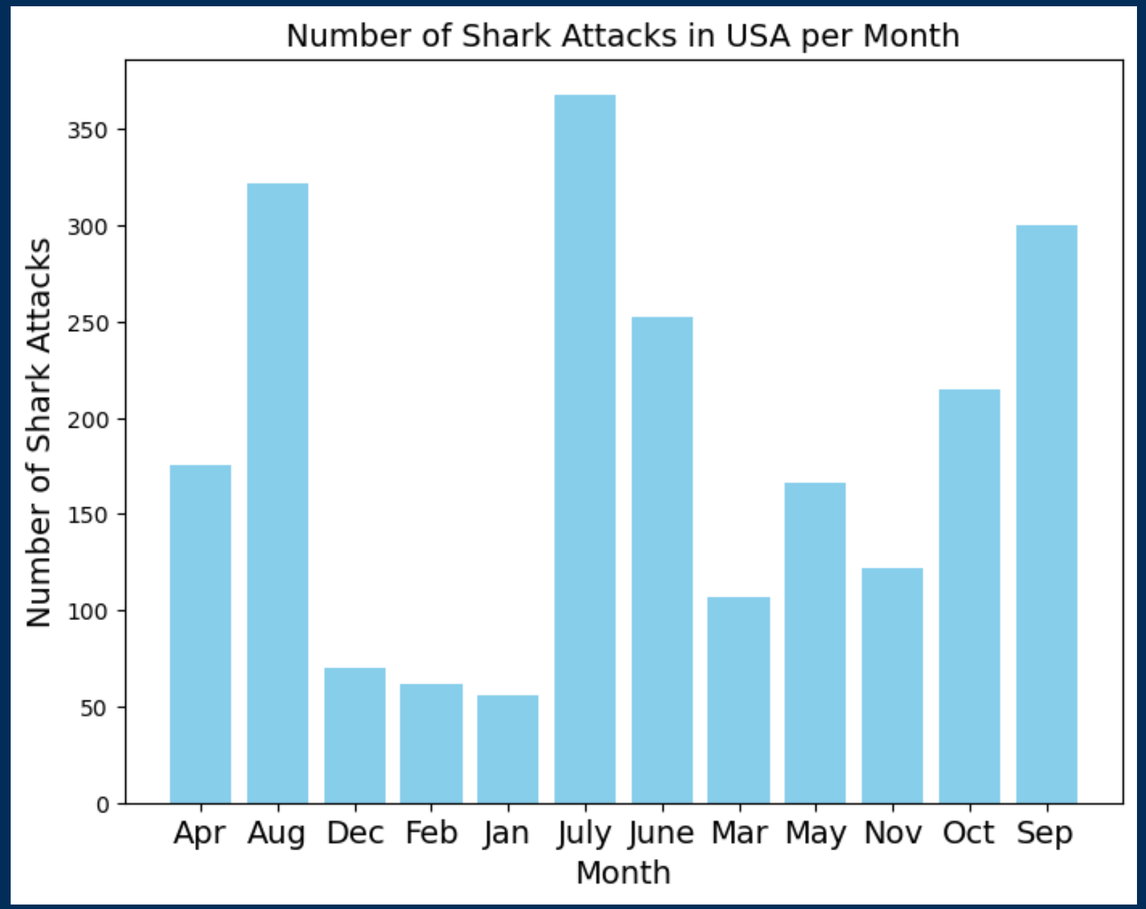


USA and Australia
have the biggest number
of shark attacks per year

Exploratory Data Analysis

MVP related Insights

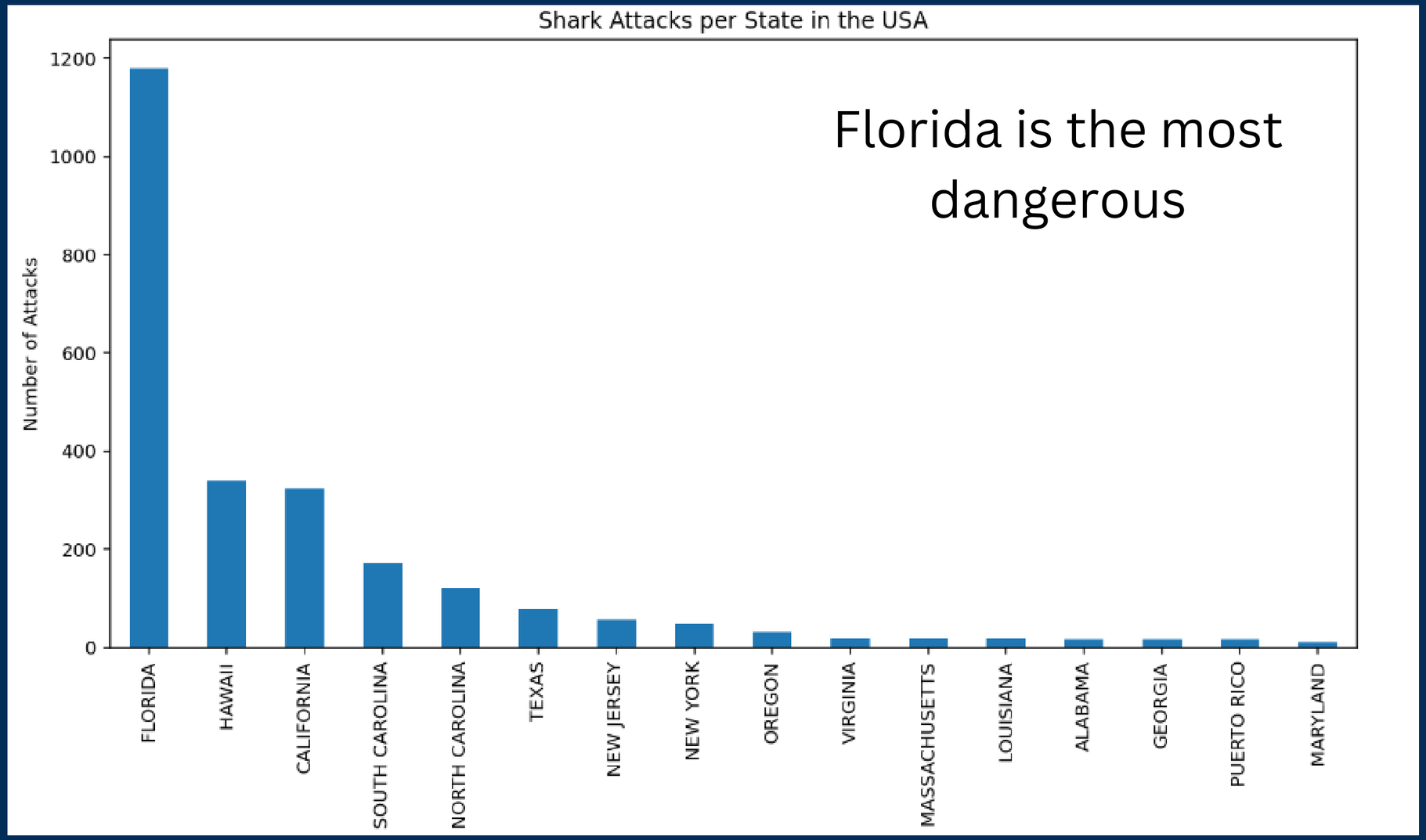




Most attacks during summer time

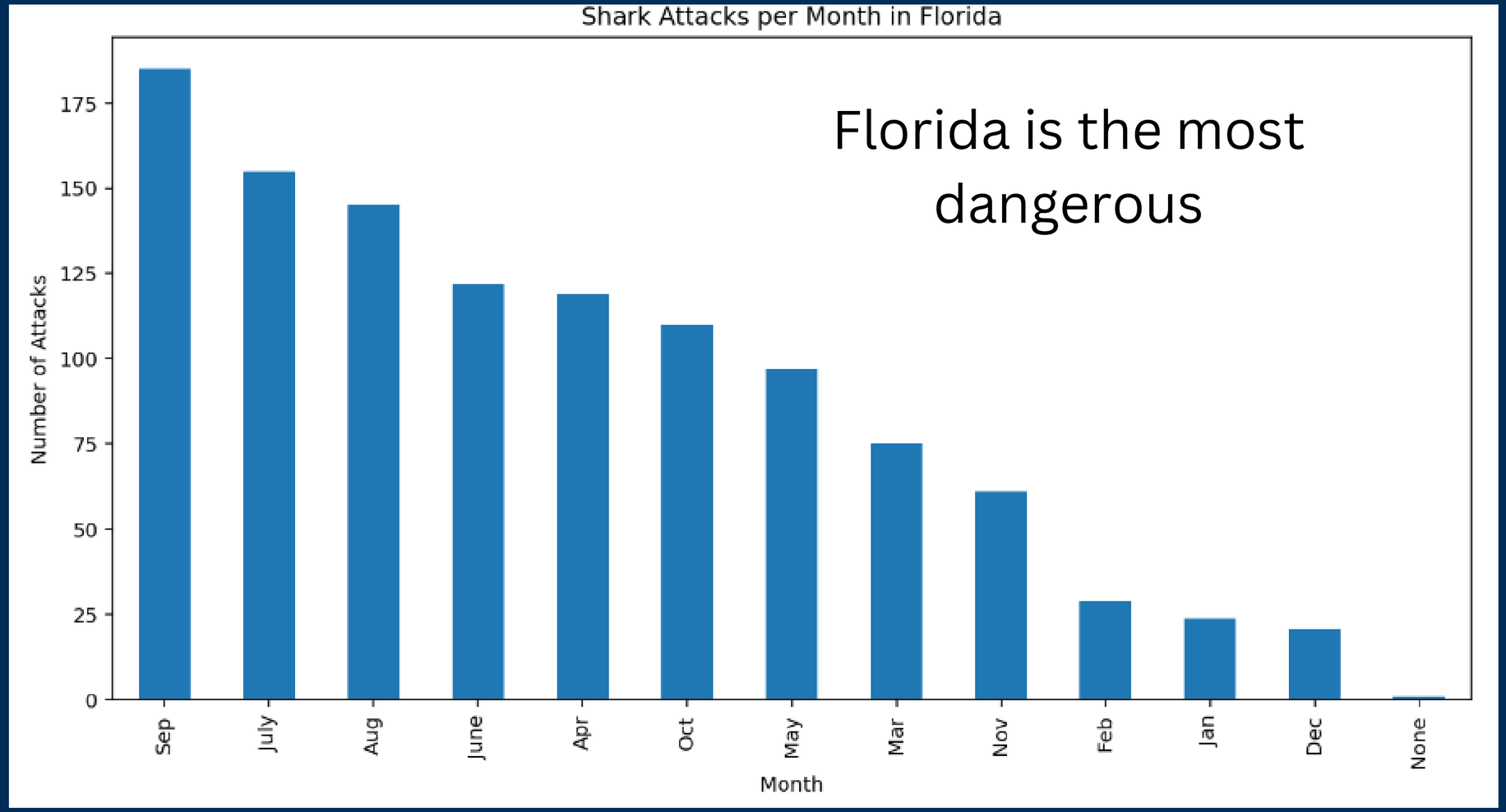
Exploratory Data Analysis

MVP related Insights



Exploratory Data Analysis

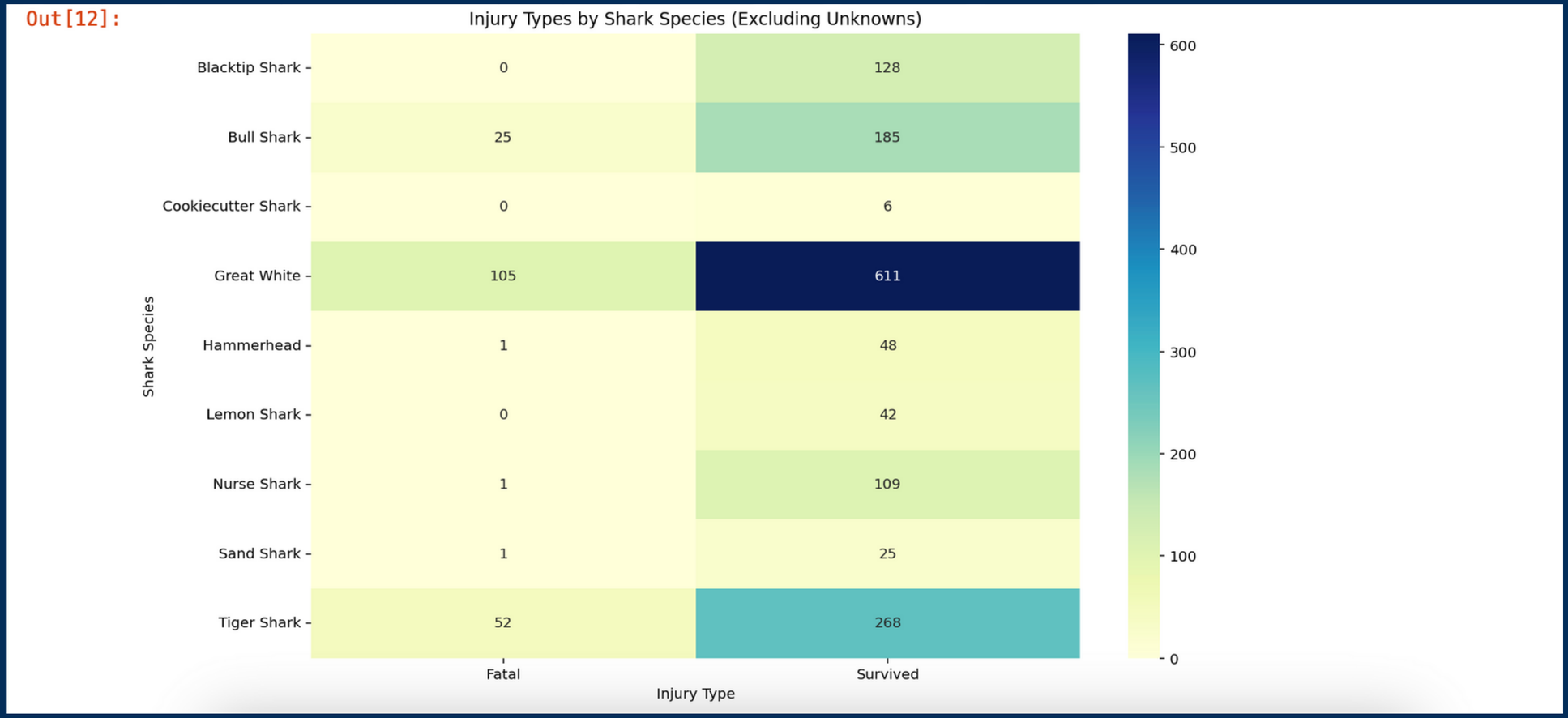
MVP related Insights



Exploratory Data Analysis

Exploratory Insights

Out[12]:



Major Obstacle & Learnings

Major Obstacles

- The way the data is input in sharkattackfile.net is very free , with lots of non-standardized inputs to deal with
- Very limited time (but that’s the game with MVP)

Key Learnings

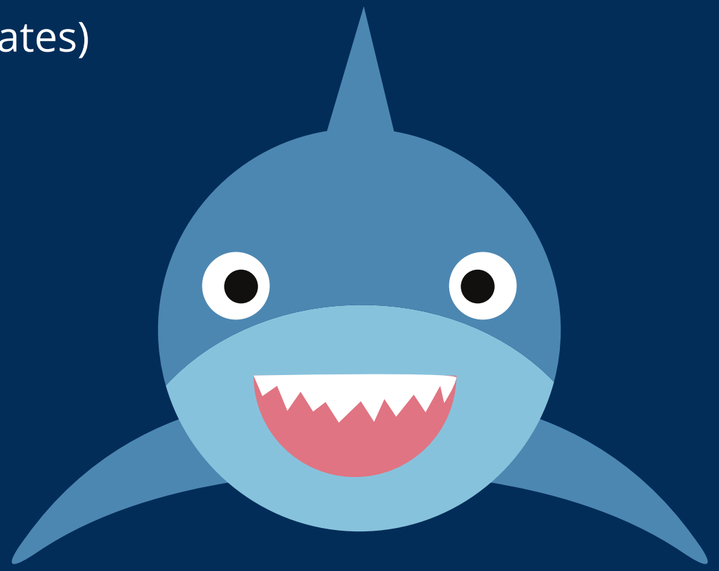
- Don’t try to clean everything and then analyse but focus early the work on the prioritary data to process

Conclusion

- **MVP Achievement:**
- Validation that it is possible to have data on locations with most shark attacks given a specific activity and month
- But:
  - data quality issues limits the insights and analysis in a short period of time
  - workaround made to concentrate on places with more attacks and better quality data
  - works have to made to pursue the data cleaning (like on duplicates)

- **Surprising insights:**

- Certain sharks species don’t kill
- Majority of people survived to shark attacks



Thank You

Resources page