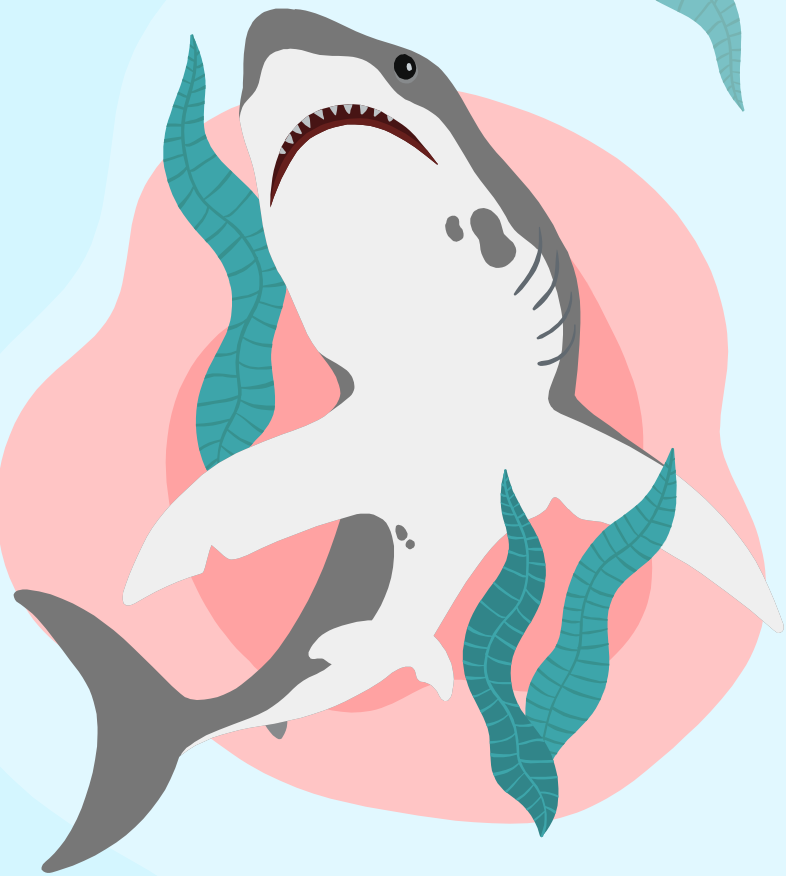
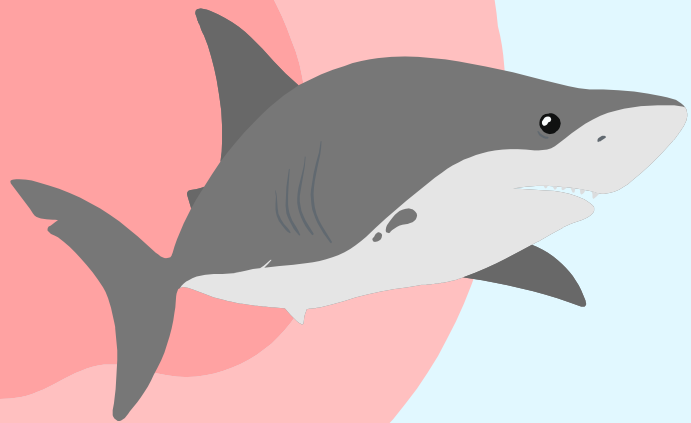


Shark Attack

Data Cleaning Project

Vanessa Dechen





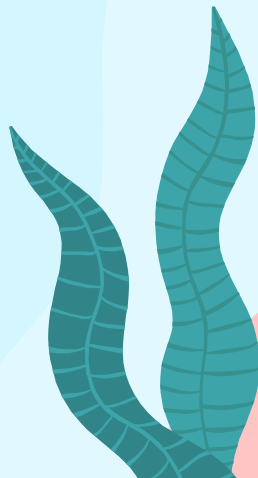
Goal and Description

This data analysis aimed at *cleaning the "Global Shark Attacks"* dataset in order to **analyse the fatality incidence in the records and the profile of these fatal attacks** considering 'sex', 'age', 'country' and 'year' information provided.

Steps

1. The original dataset was made of **25723 lines and 24 columns**.

After cleaning lines with over 22 null values in them, it was left with **6302 lines** for the analysis.



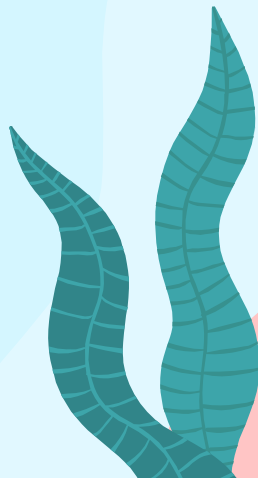
Steps

2. The 'fatal', 'sex' and 'age' **columns were cleaned**.

Unwanted categories and specific errors that represented less than 1% of data were turned into nulls or replaced by a small correction.

2.1 The **'fatal'** column was left with **'y'**, **'n'** and null values.

2.2 The **'sex'** column was left with **'m'**, **'f'** and null values.



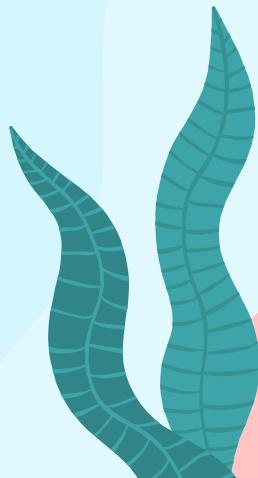
Steps

2.3 The **'age'** columns was left with **float and null values**.

If there were uncertain age values for a single person, the average of the values was used. If the record contained ages for more than a victim, each age turned into a different line and information from other columns were duplicated.

The dataset ended up with 6332 lines.

2.4 The **'age_groups'** column was created to classify ages in the following intervals: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-87.



Steps

3. **Data frames were created** to analyse the total values and percentages of categories for these columns: 'fatal', 'sex', 'age_groups', 'fatal X sex', 'fatal X age_groups'.

4. The **'country' and the 'year' columns** were grouped by the 'fatal' column and the frequency of their values.



Analysis

From the lines of 6332 records:

Most attacks are not fatal
(75.6% from 90.16% not nulls)

	fatal	t%
n	4316	75.6
y	1393	24.4

Analysis

From the lines of 6332 records:

Most attacks involve male victims
(88.9% from 90,95% not nulls)

	sex	t%
m	5120	88.9
f	639	11.1

Analysis

From the lines of 6332 records:

The attacks frequency decreases with age increase, except for young children between 0-9 (from 54,4% not nulls)

	age_group	t%
10-19	1093	31.72
20-29	1017	29.51
30-39	560	16.25
40-49	355	10.30
50-59	196	5.69
0-9	118	3.42
60-69	80	2.32
70-79	22	0.64
80-87	5	0.15

Analysis

From the lines of 6332 records:

The mean age of the attacks was 27.4 years old
(nulls not counted)

count	3446.000000
mean	27.410186
std	13.894617
min	0.300000
25%	17.000000
50%	24.000000
75%	35.000000
max	87.000000

Analysis

From the lines of 6332 records:

Most fatal cases happen among males
(22.44% from 82,83% not nulls)

	fatal	sex	cases	t%	r%
0	n	m	3491	66.56	88.13
1	n	f	470	8.96	11.87
2	y	m	1177	22.44	91.67
3	y	f	107	2.04	8.33

Analysis

From the lines of 6332 records:

Most fatal cases happen between ages 10-29
(over 65% from 50,55% not nulls)

	fatal	age_group	cases	t%	r%
9	y	10-19	198	6.19	32.62
10	y	20-29	197	6.15	32.45
11	y	30-39	91	2.84	14.99
12	y	40-49	45	1.41	7.41
13	y	50-59	32	1.00	5.27
14	y	60-69	18	0.56	2.97
15	y	0-9	18	0.56	2.97
16	y	70-79	7	0.22	1.15
17	y	80-87	1	0.03	0.16

Analysis

From the lines of 6332 records:

The top 10 countries for fatal attacks are:

AUSTRALIA	283
USA	187
SOUTH AFRICA	106
PAPUA NEW GUINEA	56
MEXICO	44
BRAZIL	38
PHILIPPINES	35
REUNION	29
NEW ZEALAND	24
CUBA	24

Analysis

From the lines of 6332 records:

The top 10 years for fatal attacks since 1960 are:

2000.0	18
1944.0	17
1993.0	16
1942.0	16
1964.0	16
1962.0	16
1966.0	15
1963.0	15
1954.0	15
1960.0	14

Therefore...

Most attacks are not fatal, but among fatal victims most are males between 10-29 years old.

Most fatal cases are from 2000 and Australia.

