Shark Attacks

Maybe they just want to be our friends

Angie, Hugo, Karl, Mariana

Project Overview

Original dataset: 6965 rows × 23 columns, 2 columns float and 21 columns object, duplicate rows, missing values in every column

Problem statement: Are sharks really dangerous?

Hypotheses:

1. The majority of shark attacks are not fatal.
2. The majority of shark attacks were a result of the shark being provoked.
3. There are significant differences in the gender distribution of shark attack victims.
4. There are significant differences among countries in the world regarding the number of shark attacks.
5. There are significant differences in the age distribution of shark attack victims.
6. The activities victims were engaged in are associated with the occurrence of shark attacks.

Data Wrangling and Cleaning

1. We dropped columns with no relevance for our analysis
2. We had too many Null values. We decided to only look from the year 2000 onwards
3. Variable **sex** had null values, values with no meaning and values not coherent. We homogenised the values and dropped the nulls and the ones with no meaning
4. Variable **type** had some mistakes, we cleaned it and regrouped the values
5. From the column **injury**, we created a new column with values “Fatal” and “Non-Fatal”
6. Variable **country** had some NA’s, we decided to drop them
7. Variable **age** had not coherent values and missing values. We homogenised the values and dropped the nulls
8. Variable **activity** had too many different values, we cleaned it and regrouped the values

Exploratory Data Analysis

1. Exploratory data analysis methods we used:
   - We use univariate for the statistical information
   - We use univariate and Bivariate variables for the graphicals

2. Insights and interesting patterns we found:

   - Although this is a big DataFrame, it wasn’t easy to analyse it, due to its many inconsistencies and NA’s
   - We weren’t expecting such a high frequency of attacks, especially in the more recent years
   - We found that the majority of victims were surfers

Major Obstacle

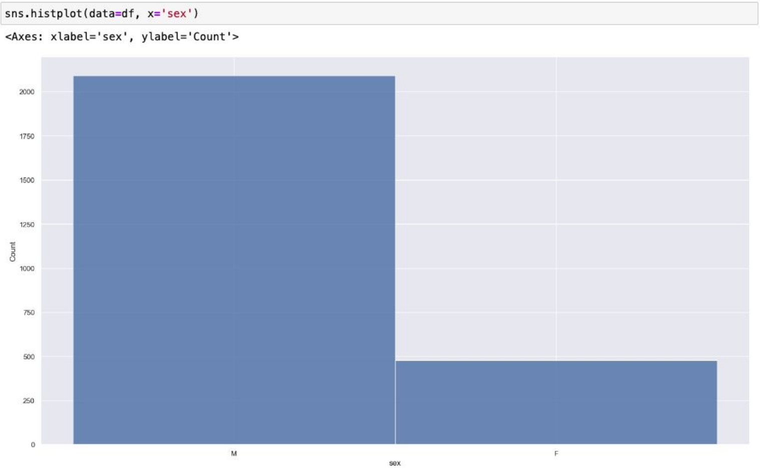
We had a lot of good ideas for hypothesis that we wanted to test, but we had very little time.

In hindsight, maybe we were too ambitious and got carried away, leading to us trying to tackle all the variables, resulting in bad time management and a lack of depth in our analysis.

Conclusion and Insights

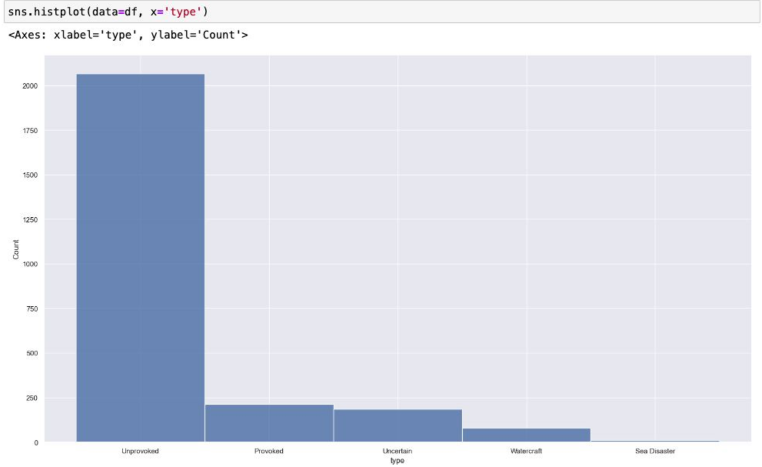
- H1: The majority of shark attacks are not fatal.
  - Supported by our analysis (Non-Fatal: 2336, Fatal: 220)

- H2: The majority of shark attacks were a result of the shark being provoked.
  - Refuted by our analysis (Unprovoked: 2067, Provoked: 213, Uncertain: 185, Watercraft: 81, Sea Disaster: 10)



```
sns.histplot(data=df, x='sex')
```

```
<Axes: xlabel='sex', ylabel='Count'>
```

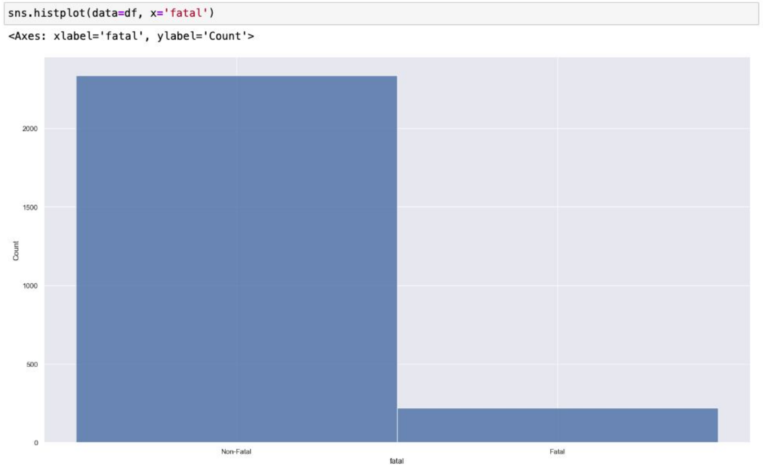


```
sns.histplot(data=df, x='type')
```

```
<Axes: xlabel='type', ylabel='Count'>
```

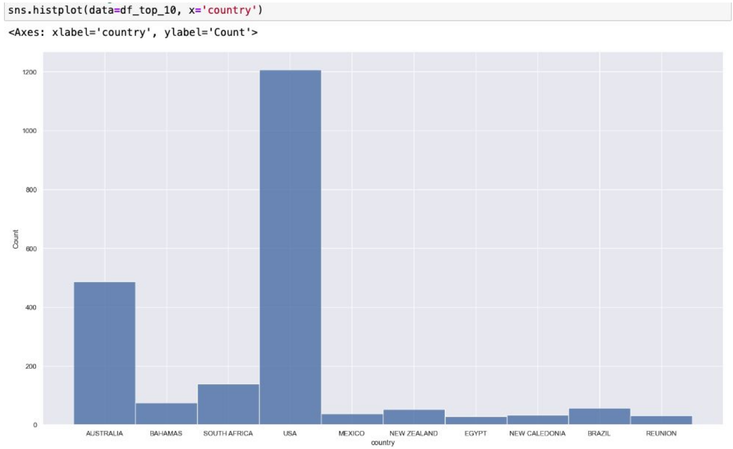
Conclusion and Insights

- H3: There are significant differences in the gender distribution of shark attack victims.
    - ○ Supported by our analysis (2092 males vs 478 females)
- H4: There are significant differences among countries regarding the number of shark attacks.
    - ○ Supported by our analysis (USA 1207, Australia 486, South Africa 140, Bahamas 75, Brazil 57, New Zealand 53, Mexico 37, New Caledonia 33, Reunion 31, Egypt 28)



```
sns.histplot(data=df, x='fatal')
```

```
<Axes: xlabel='fatal', ylabel='Count'>
```



```
sns.histplot(data=df_top_10, x='country')
```

```
<Axes: xlabel='country', ylabel='Count'>
```

Conclusion and Insights

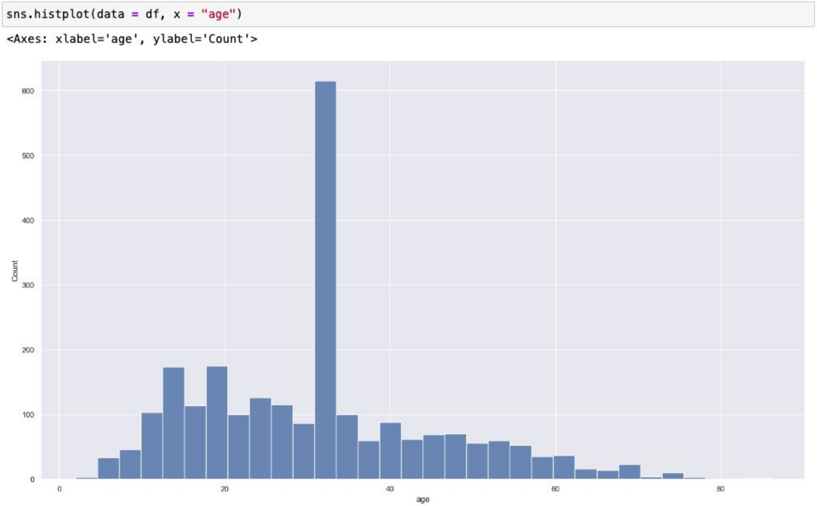
H5: There are significant differences in the age distribution of shark attack victims.

- ○ Supported by our analysis (31 years old: 519, 15 years old: 80)

H6: The activities victims were engaged in are associated with the occurrence of shark attacks.

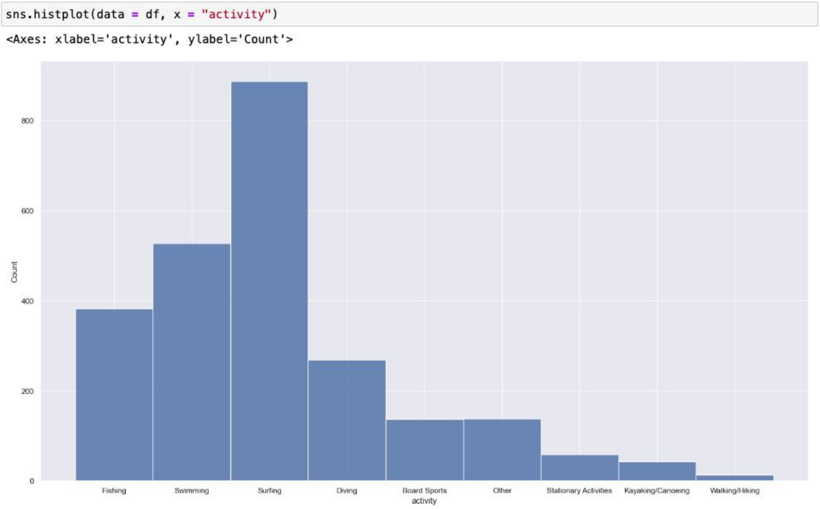
- ○ Supported by our analysis (Surfing: 887, Swimming: 527, Fishing: 382)

Check out our code



```
sns.histplot(data = df, x = "age")
```

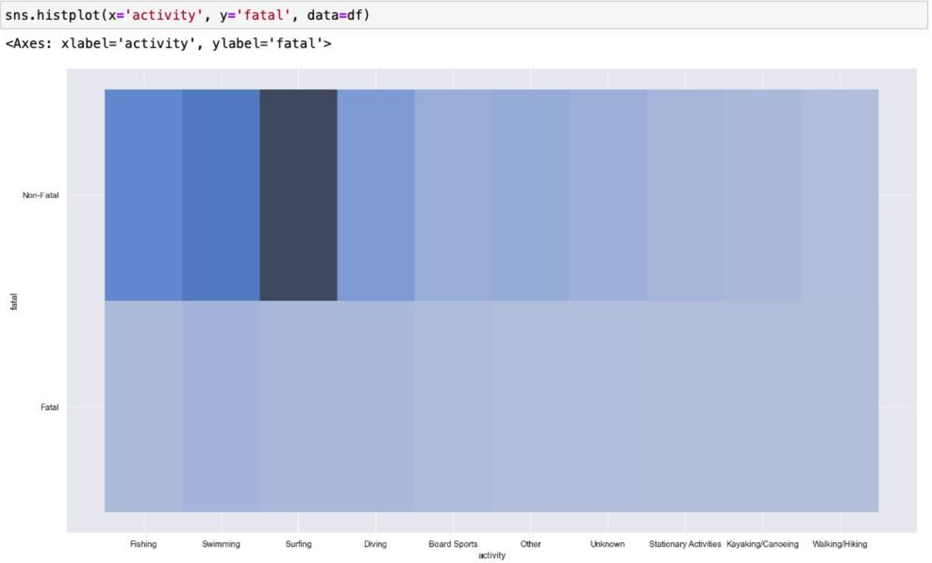
```
<Axes: xlabel='age', ylabel='Count'>
```



```
sns.histplot(data = df, x = "activity")
```

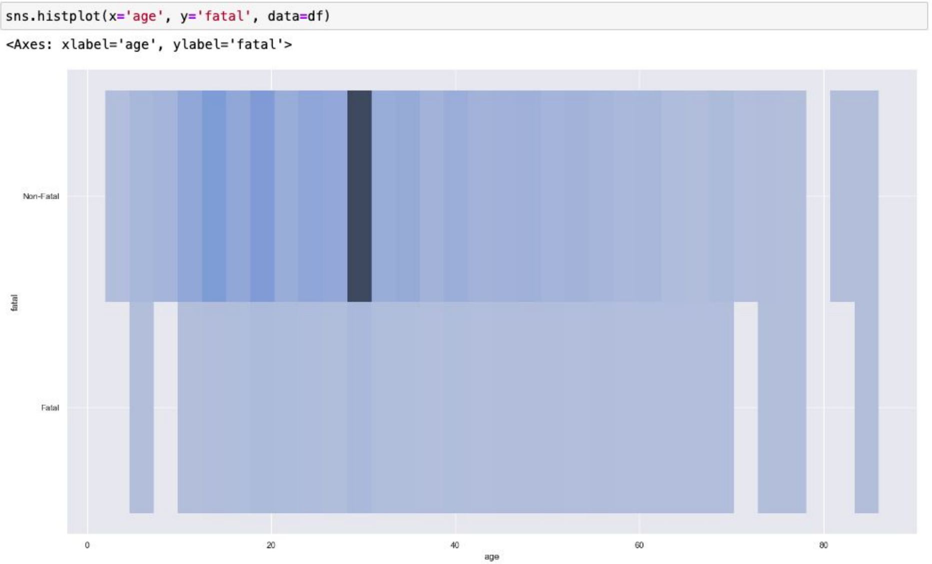
```
<Axes: xlabel='activity', ylabel='Count'>
```

Conclusion and Insights



```
sns.histplot(x='activity', y='fatal', data=df)
```

```
<Axes: xlabel='activity', ylabel='fatal'>
```



```
sns.histplot(x='age', y='fatal', data=df)
```

```
<Axes: xlabel='age', ylabel='fatal'>
```

Thank you!

Angie, Hugo, Karl, Mariana