Shark Attacks on Surfers: EDA

Peter Boyd

May 26, 2019

Introduction

As someone who enjoys surfing on the Oregon coast, I was interested in shark attacks that happen around the globe, specifically attacks on surfers.

The data set I found contains information about shark attacks around the world and can be downloaded from Kaggle (https://www.kaggle.com/teajay/global-shark-attacks#attacks.csv). The data is downloaded as a .csv containing 22 columns and 6302 rows. Each row is an attack, and the columns of interest include date, type of attack, location, activity when attacked, name of victim, sex of victim, age of victim, injury, fatal or not fatal, species of shark.

I chose to focus my project on surfing related attacks only, so after subsetting the data, we have 976 observations to work with. I found many of the columns to be rather messy, which further limited some analyses. Columns that are inherently "character" related (Name, Injury, Species, etc) offered a great deal of variability that would make analysis rather exhausting. Consequently, much of my work thus far has focused on Date, Fatal/Not Fatal attacks, Age, and Location.

A further complication was introduced by the lack of geographical coordinates. I had hoped to more cleanly map incidents around the world, but without provided coordinates, I would need to find a concise way of obtaining the coordintes from the internet for hundreds of beaches, which seemed impractical. I'd like to spend more time tweaking the location aspects of the data to allow for some map visualizations.

On this assignement, I spent roughly 1 hour on a different data set that I abandoned, 3 hours with data manipulation/idea formluation/basic plots, 4 hours with more advance plotting and new ideas, 2 hours polishing and some extra ideas, and 2 hours compiling/writing.

Ultmiately, after investigating many variables and manipulations, I was surprised and intrigued by the following three analyses.

## Visualization One: Age of Surfers

Age range of shark attack surfing victims has increased
Median Age with 25th to 75th percentiles
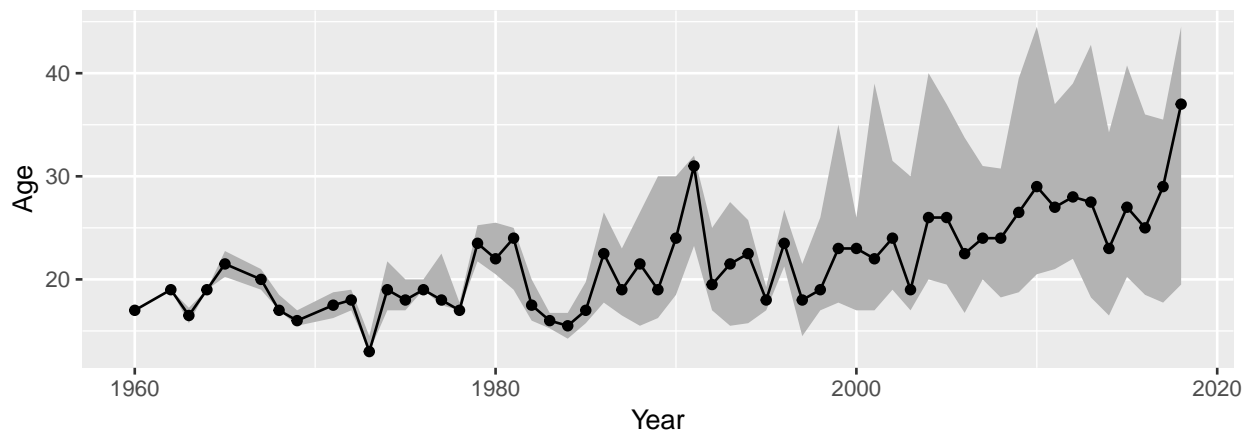
In the image above, we see median surfer's age of shark attack victims plotted for each year from 1960 - 2017. What I found interesting was the increase in variability by including a ribbon capturing the 25th and 75th percentiles. This may imply that the age of surfers in general has greater variability now than it did a few decades ago.

Visualization Two:
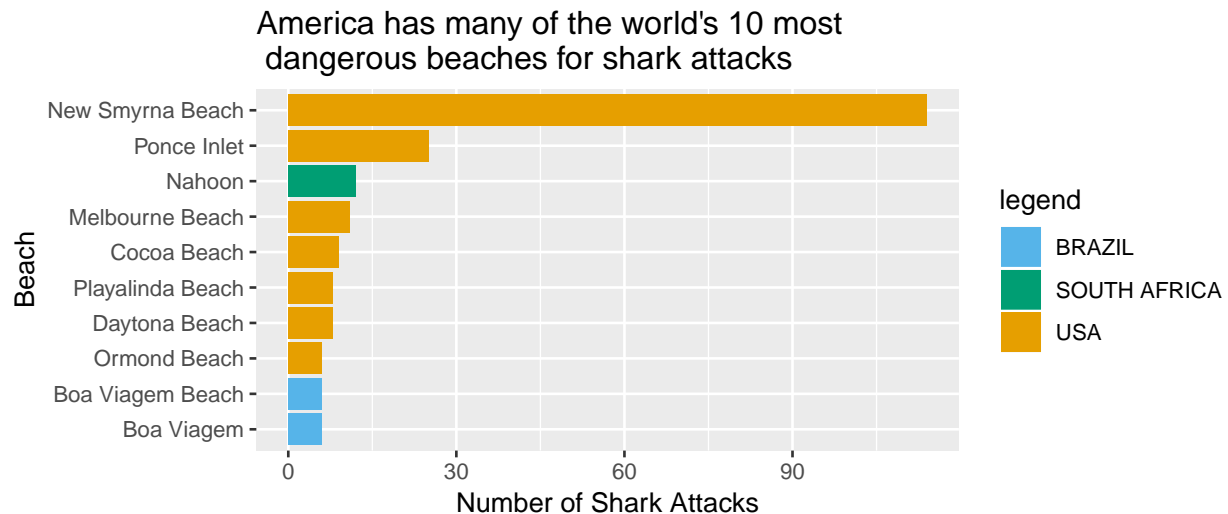
I was surprised by this visualization because I anticipated beaches in Australia to be more dangerous and an overall less dominant presence from American beaches.

Visualization Three:
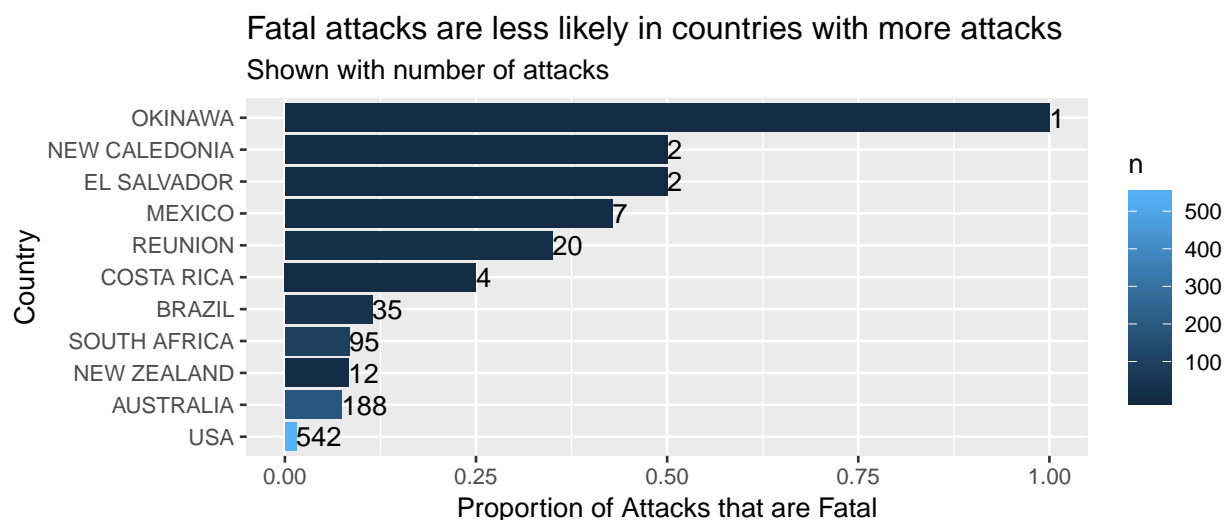
This visualization was interesting because it shows that many of the countires with very few attacks had high proportions of attacks that were in fact fatal. Meanwhile, the country with the most attacks (the US) had the lowest proportion of fatal attacks.