Chith Sabesh
chithsabesh@gmail.com

GLOBAL SHARK ATTACKS MILESTONE REPORT

OVERVIEW

The problem here deals with analysing and examining shark attacks based on various informations like country,gender etc. Project aims at coming up with a model for classifying fatal and non-fatal shark attacks based on both conventional techniques and text analysis.

THE DATA

The data is available to us from Kaggle (https://www.kaggle.com/teajay/global-shark-attacks) and hence no scraping is required. However there is a bit of cleaning and wrangling required.

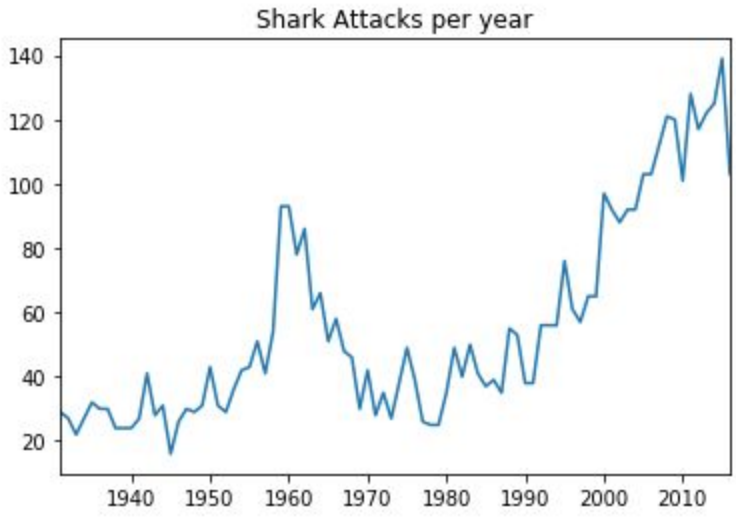
DATA WRANGLING

1) Missing values were treated appropriately for most of the features. Time was one such feature which had to be dropped since imputing time was found to be not possible.
2) Age was imputed with the mean of all ages.
3) The activities were cleaned to include only relevant activities.

EXPLORATORY DATA ANALYSIS

The following section can be divided to many other sections for ease of work.

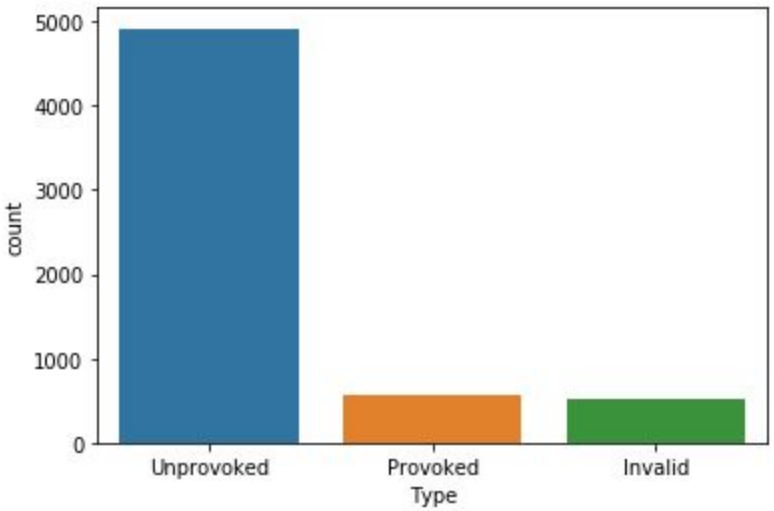
Plotting Year of attacks:



Shark attacks took a sharp increase after around 1980 can be attributed to the fact that humans started entering into the Oceans more and more and reporting of attacks became better

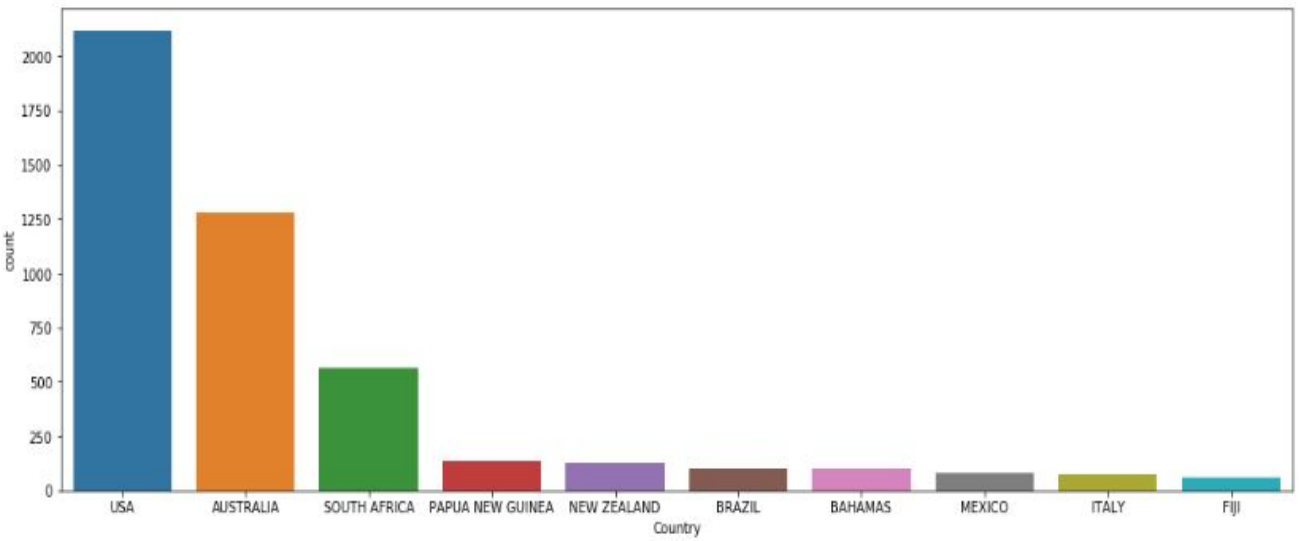
Chith Sabesh
chithsabesh@gmail.com

ATTACK TYPES



Most of the attacks are unprovoked which is reasonable since no one ever provokes a shark except maybe fishermen.

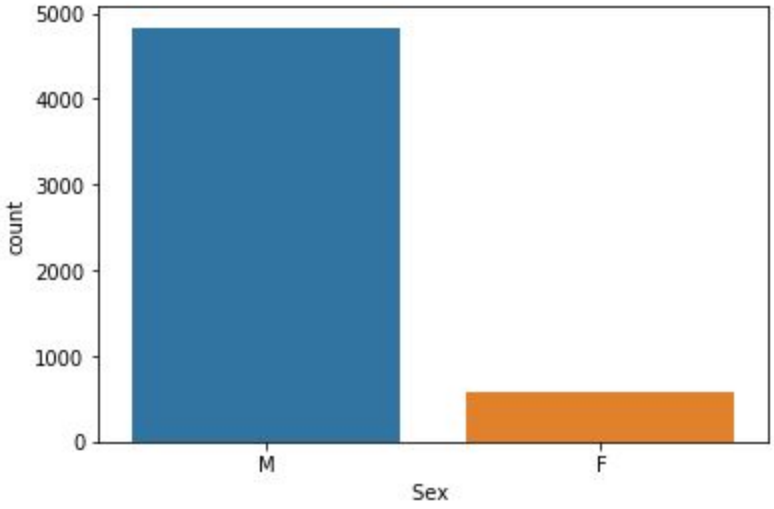
COUNTRIES



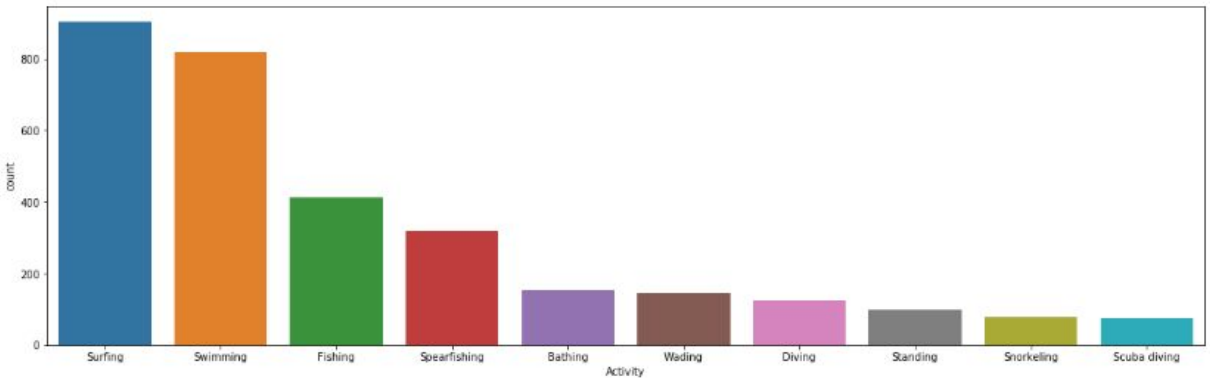
Most of these countries have a large coastline and hence large fishing industries.

Chith Sabesh
chithsabesh@gmail.com

GENDER



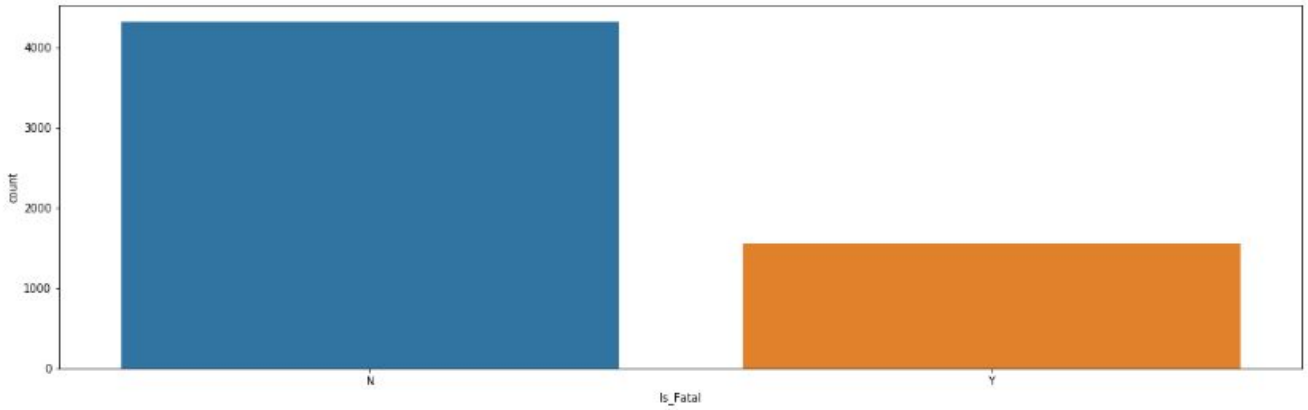
ACTIVITY



Surfing is the most attacked activity followed by swimming

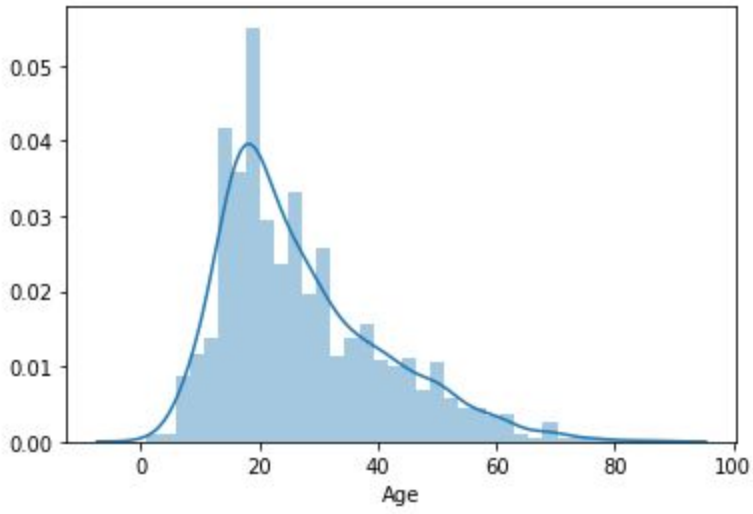
Chith Sabesh
chithsabesh@gmail.com

Fatalities to Non-Fatalities



Most of the attacks are non-fatal

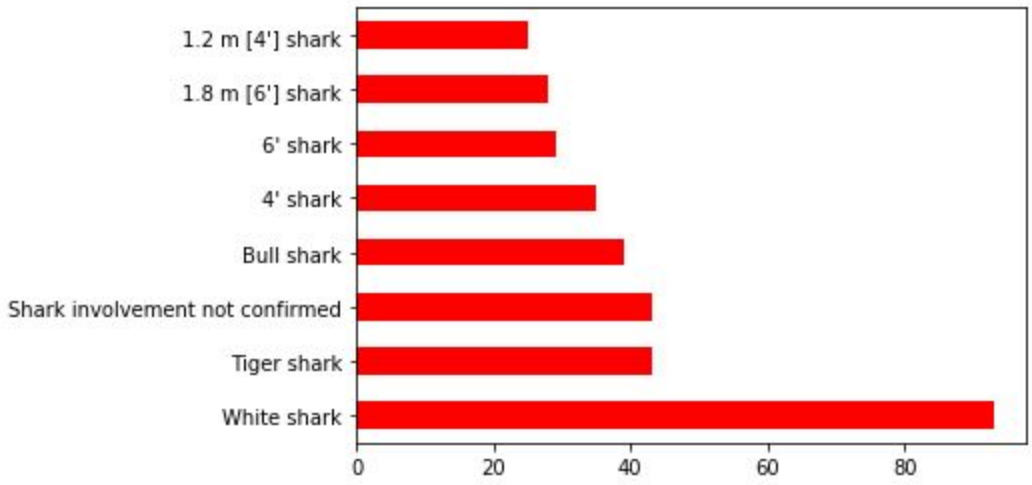
AGE DISTRIBUTION



The mean age of attacks is 27.08.There is a slight positive skew with maximum peak between 15 and 35. This is due to the fact that most of the teen and middle age adults are the ones who enter the water the most.

Chith Sabesh
chithsabesh@gmail.com

SPECIES



The species column is very unclean and hence only some analysis could be done and it was found that White Shark the biggest of all carnivorous sharks has attacked the most.

WORD CLOUDS

NON-FATAL

Chith Sabesh
chithsabesh@gmail.com

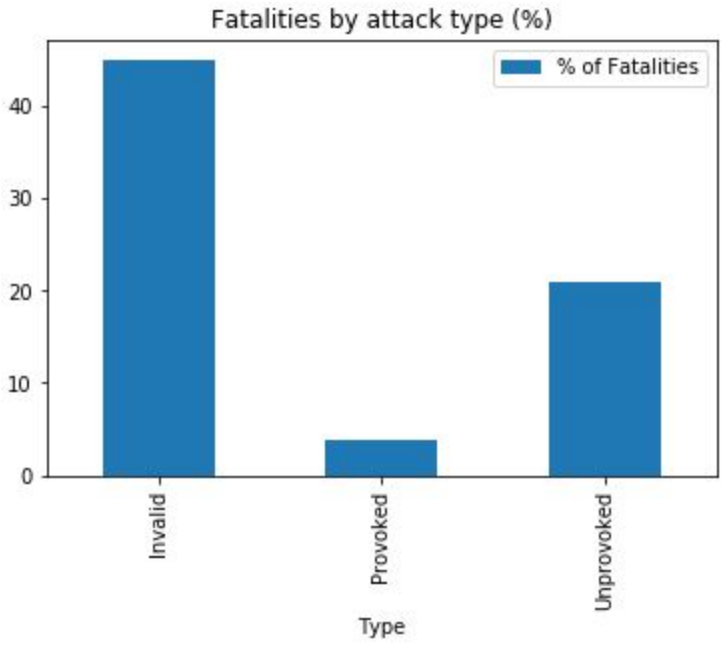
FATAL

Chith Sabesh
chithsabesh@gmail.com

Clear difference between both the word clouds

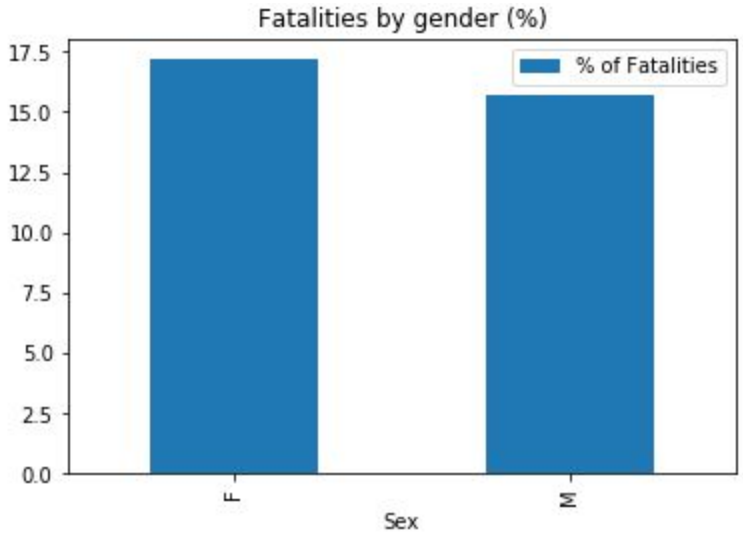
DOES PROVOCATION LEAD TO FATALITY

Chith Sabesh
chithsabesh@gmail.com



No unprovoked attacks always tend to have more fatalities here both Unprovoked and Invalid have high fatalities (Since most shark attacks happen due to mistaken identity and not due to any provocation)

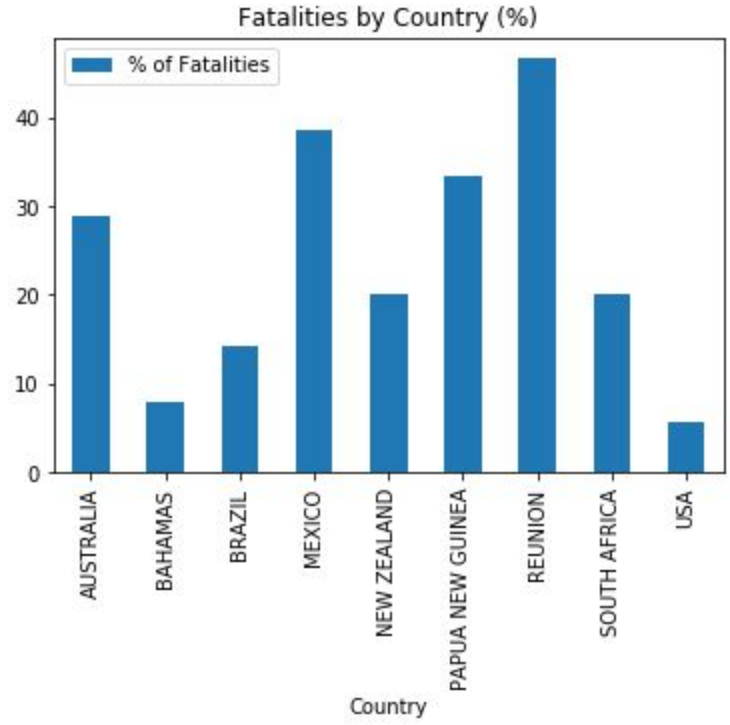
DO FEMALES DIE MORE



Strangely females have higher fatalities than males

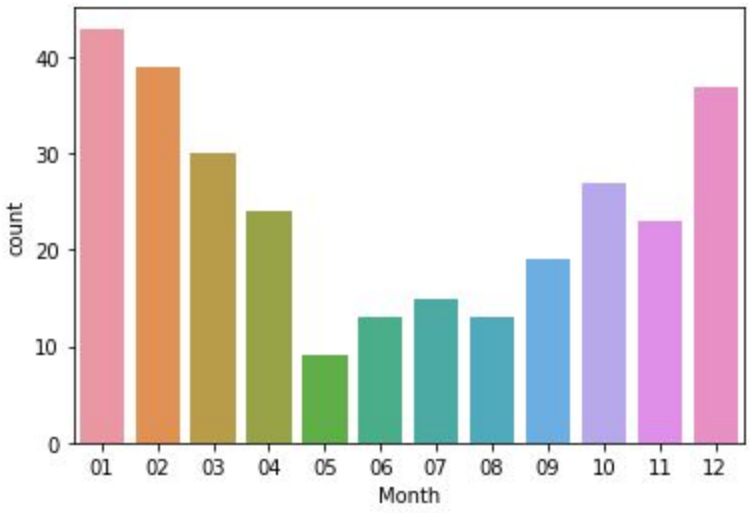
FATALITIES BY COUNTRY

Chith Sabesh
chithsabesh@gmail.com



Here it should be known that Reunion has only about 15 attacks out of which nearly 50 percent of attacks are fatal. The case is same for both Mexico and PAPUA NEW Guinea

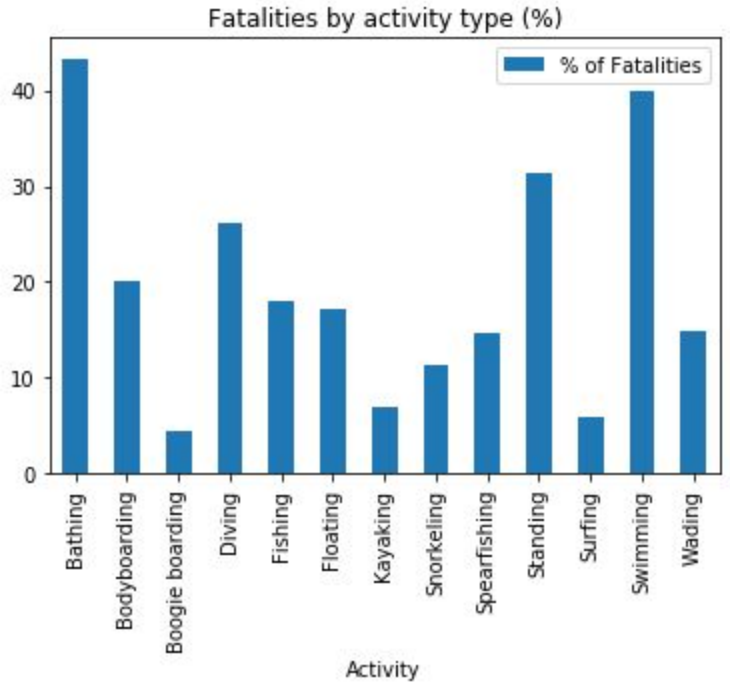
MONTH WISE ATTACKS IN AUSTRALIA



Attacks happen in the height of summer not only in Australia but all over the world and hence no point in checking for fatalities month wise

Chith Sabesh
chithsabesh@gmail.com

Fatalities By Activity



Although surfing has the most attacks it is not the most dangerous that distinction goes to bathing and swimming.Reason for this could be that sharks are attracted to splashing activities in the water

INFERENTIAL STATISTICS

This section presents the results of inferential statistics methods applied on two hypothesis tests namely:

1.) Relationship between fatality and activity
2.) Relationship between fatality and gender

Relationship between fatality and activity

This test was performed to test whether there is a relationship between fatality and activity in other words to see if the activity influences the fatality. To do this we only took the top 10 activities since there are many other activities that are big strings and they also don't make sense.Since both the variables are categorical variables we do the chi-square test to check for dependencies.We made a contingency table using the two variables

Chith Sabesh
chithsabesh@gmail.com

Is_Fatal	N	Y	All
Activity_new			
Bathing	75	61	136
Bodyboarding	95	24	119
Boogie boarding	46	2	48
Diving	260	86	346
Fishing	309	63	372
Floating	32	6	38
Kayaking	41	3	44
Playing	20	1	21
Snorkeling	64	8	72
Spearfishing	270	44	314
Standing	122	49	171
Surfing	894	53	947
Swimming	541	346	887
Wading	197	33	230
All	2966	779	3745

1.There is significant relationship between fatality and activity
2.P-value obtained was 4.14 x 10^-69

Relationship between fatality and gender
Here again we perform chi-square test for finding dependencies between the two variables

Is_Fatal	N	Y	All
Sex			
F	362	77	439
M	2600	702	3302
All	2962	779	3741

1.There is no significant relationship between gender and fatality
2.P-Value obtained here was 0.516

Chith Sabesh
chithsabesh@gmail.com

CONCLUSION

This report highlights the Wrangling, Exploratory data analysis and inferential statistics done on the global shark attacks dataset. With these insights it is now possible to move to the steps of feature engineering and Machine learning