Analysis of worldwide shark attack

Wassim BEN YOUSSEF, Principle of Data Science, City, University of London

Introduction

According to Alex MacCormick [1], “shark attacks are still a very real threat to humans”. Indeed, shark attacks can target people doing common things like fishing, swimming or even just washing dishes in safe areas. A statistic study of these attacks might permit to understand the violent behavior of these animals by analyzing characteristics of the attacks. According to The Global Shark Attack File website [2], “by understanding when and why shark attacks occur, it is possible to lessen the likelihood of these incidents.” That is why this institution provides a complete dataset (called Global Shark Attack File or International Shark Attack File) listing all the attacks that they could identify using trustful sources (mostly journal papers). They could collect more than 5900 worldwide attacks for the last three millenniums (some Greek poems and ancient tales reported very old attacks).

This analysis tries to visualize and understand the information provided in this dataset in order to give an overview of the situation of worldwide shark attacks by using tools and methods advocated by a data scientist approach. To perform this study, I firstly decided to use the software Python but I encountered different technical problems so that I finally decided to use R. However, I kept some elements obtained using Python.

I. Description of the dataset and definition of the goals

The dataset contains 5992 observations and 16 variables:

- Case number: the ID of the observation
- Date: the date of the attack on format DD-MM-YYYY
- Year
- Type: how the attack was caused, if it was provoked, unprovoked, due to a sea disaster…
- Country: the country where happened the attack. If the assault occurred at sea, it is the name of the ocean or sea
- Area: the region of country
- Location: a more precise location that can be the beach, the city, the bay, the river…
- Activity: what the victim (or victims) was doing while she was attacked. Often just a word but sometimes can be a brief description
- Name: the name of the victim (or victims). If unknown, it is her sex
- Sex
- Age
- Injury: a short description of damages in a few words
- Fatal: takes the value Y, N or Unknown if the victim just disappeared
- Time: the approximative hour or the time of the day (morning, afternoon …)
- Species: the species of the shark or a brief description giving its approximative size in foot
- Investigator or source

This dataset is rich and provides an excellent overview of the topic. The first part of the study consists of a first analysis of the data to see the repartition of the variables, handling missing values and data cleaning and transformation. Indeed, some features do not provide clear and useful information. For instance, for the features Species and Injury, the values are sentences so that for the R software, each one is a unique value while they can be clustered to give them more sense. That is why we will have to clean and modify the dataset to extract more information about the features. After that, in order to understand the geographical repartition of attacks, new geographic zones was created to cluster countries in bigger areas. So, we will be able to see the characteristics of attacks in these regions. This clusterization permits us to calculate and visualize the probability to die when we are attacked in each zone. We also will analyze and critic a predictive model created to predict the feature Fatal. In order to tackle the problem of shark attacks, some governments installed anti-shark nets. To finish the study, I found interesting to compare results obtained for two different installations.

II. First analysis and data transformation

1. Missing values

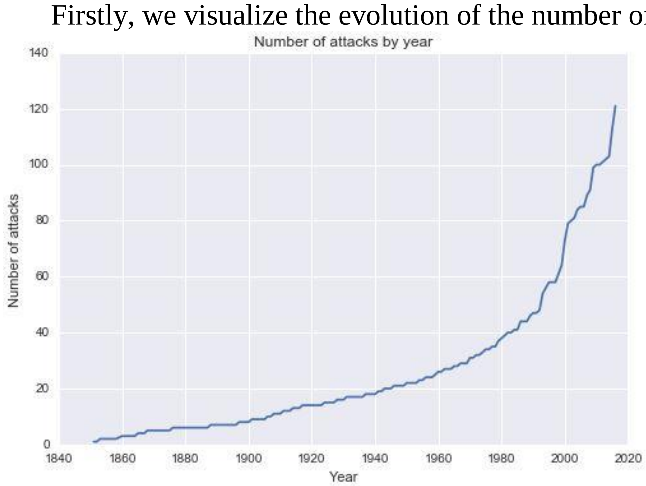
The dataset has a substantial number of missing values:

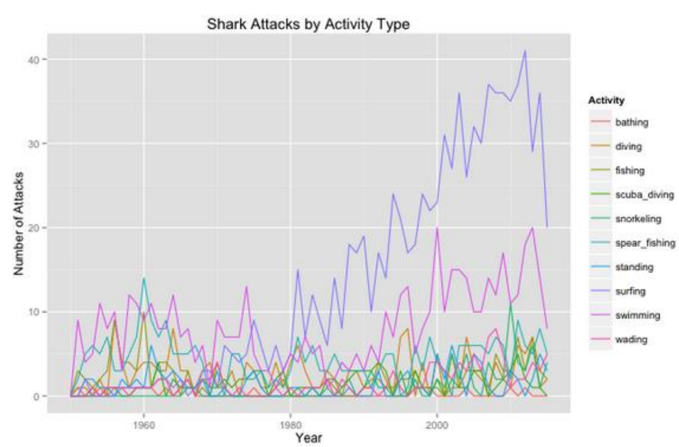
```
          Case.Number                  Date                  Year                  Type
 Country
                    0                     0                     0                     0
      43
                 Area              Location              Activity                  Name
     Sex
                  402                   496                   528                   200
     567
                  Age                Injury            Fatal..Y.N.                  Time
 Species
                 2678                    27                    19                  3215
    2940
Investigator.or.Source
                   15
```

Having a large dataset, I decided just to remove observations with missing values for features “Country”, “Area”, “Location”, “Activity”, “Name”, “Injury”, “Fatal” and “Investigator or Source”. After doing that, we still have 4286 observations which are enough. For the other variables “Species”, “Age’ and “Time”, I firstly wanted to try to predict them using a machine learning algorithm like Decision Trees, but because of the structures of these variables, it was not possible to find a suitable model. Indeed, values of “Species” are practically all different and so they are considered as unique values. So it is considered as a categorical variable with many values, which is then very tough to model. The variable “Age” sometimes takes an integer value and sometimes a character value, for instance “18 & 25”, when two people were attacked, as “Time” which can take a character value like “Afternoon” or a time value like “12h30”. I did not know how to build a good model to predict missing values of these features, so I decided not to touch its for now until these missing values are problematic. Additionally, I decided to keep only attacks which occurred after 1850. Indeed, observations before this date would be useless given that only a few attacks were reported at this time because of the low development of communication systems and poor information was given for these which was reported.

2. Visualization of features

Firstly, we visualize the evolution of the number of attacks reported over time:





In the first graph, we can see the evolution of the number of shark attacks per year from 1850 until nowadays, given that we decided to keep only the observations having its “Year” value upper than 1850. In the first part of the graph, the curve follows a natural rise resulting from the increase of the global population. But after the 1980s, the growth seems to accelerate. This is a consequence of the popularization of the surf as we can see in the second graph representing the number of attacks for the ten most common activities per year. Indeed, according to Peter Westwick [3], this corresponds to the period of which the surf was becoming very popular as a tourist activity. This explains the rise of the number of attacks against surfers and the overall increase of attacks. Also in the second graph, a decade after, we can see an increase of attacks against people swimming that can be explained by the fact that “holidays and travel are becoming accessible to ever broader of the population” according to Ueli Gyr [4]. This is more striking when you see the other activities that are mostly stagnating.

Then, we can compare the importance of each activity in the number of attacks in figures 1 to 3 in the appendix. As expected, surfing and swimming seem to be the most important ones. It is interesting to see that before 1960, surfing was only the thirteenth most common activity while after it became the first one. That confirms the importance of the risk if attack for this activity.

Now let us analyze the variable “Type”. As you can see in figure 4, the large majority of attacks were unprovoked. But a not negligible number were provoked.

Concerning the locations of attacks, the figure 5 shows that 3 countries concentrate a high number of attacks: the USA, Australia, and South Africa. These figures can be explained by the size and the location of these countries. Indeed, Australia is surrounded by the Pacific Ocean and the USA is boarded by both the Pacific and the Atlantic Ocean with also the Gulf of Mexico and Caribbean Sea between Florida and Texas which are infested with numerous species of shark according to Leighton R. Taylor [5].

The other features were complicated to visualize and interpret because of its structures; that is why we will have to make some transformations on the dataset.

3. Data transformation

According to a study by Kandel et al. [13], variables transformation is one of the most time-consuming challenges when dealing with a data analysis problem. Indeed, I spent nearly half of the time on the coursework dealing with this issue.

Firstly, as the variable “Country” counts many unique values, it can not be very relevant for a geographical study of attacks. It is the same for the variables “Area” and “Location” as you can see in figure 5. That is why I created a new variable “Zone” which clusters countries in 9 geographical zones: North-West Atlantic, South-West Atlantic, North-East Atlantic, South-East Atlantic, Oceania, North-West Pacific, Indian Ocean, Mediterranea and East Pacific. I thought that this repartition might be better than only considering the oceans or continents because this would have done too big areas regrouping very different countries in the same category. Hense, this categorization seems to be the more relevant, regrouping near and similar locations in one class, for instance, all the islands in the Oceania, all the eastern coast of Africa or all the Mediterranea. But some questionable choices had to be done, like adding the Gulf of Mexico and the Caribbean Sea to the North-West Atlantic zone or adding South Africa to the Indian Ocean. Another problem was posed by the USA. Indeed, the west part of the country is boarded by the Pacific Ocean while the other part by the Atlantic Ocean. To tackle that, the variable “Area” was useful. I could add the Eastern States like Florida or Georgia to North-West Atlantic while California was added to East Pacific. Moreover, I decided to put all east coast of the American continent in the same category because few attacks were reported in countries located in South-East Pacific Ocean.

After doing so, the variable “Zone” contained 27 missing values, due to some countries that I forgot to consider. I decided to predict these missing values using a decision tree, modelized in figure 6 and 7. This was done after having introduced the new variables described in the following paragraphs. This method has the advantages to be easily implemented and interpretable. Moreover, according to Eric Biernat and Michel Lutz [6], decision trees are quite efficient, non-parametric and non-linear. Also, they automatically select the most relevant variables, so I didn’t have to make a selection of variables when implementing it, I just put all the variables.

To tackle the variable “Species” composed of sentences (see figure 8), I created a new variable “Species_final”. If the name of one of the 26 species that I have identified in the dataset was detected in the value of Species, then “Species_final” takes the value of the name of the species. Else, it takes the value “other or unknown”. I also deleted observations where the word “not” was identified because it means that the value of “Species” was “Shark involment not confirmed” or “Shark not involved”.

I applied the same methodology to classify “Activity” (figure 9) and “Injury” (figure 10). I created two variables “Activity_Type” and “Injury_Type”. For “Activity”, I detected eight principal activities which I used to cluster all the activities, and all the others that can not be identified took the value “Other” (Figure 11). For instance, I attributed the value “Diving” for each observation in which the word “diving” was detected in the variable “Activity”, or “Surfing” when the word “Surf” was recognized so that “windsurfing” and “body surfing” were attributed the same value. For “Injury”, I clustered it in 6 categories, using the same approach.

I also created a binary variable “fatality” based on the variable “Fatal” taking the value “1” if the attack was fatal or “0” if not. I considered the value “Unknown” as “1”.

All these new features are shown in figure 12.

III. Geographical Analysis and Prediction

1. Geographic analysis

After modifying the dataset, we will proceed to a geographical approach to the dataset analysis in order to understand the characteristics of attacks in each geographic zone defined in the last section.

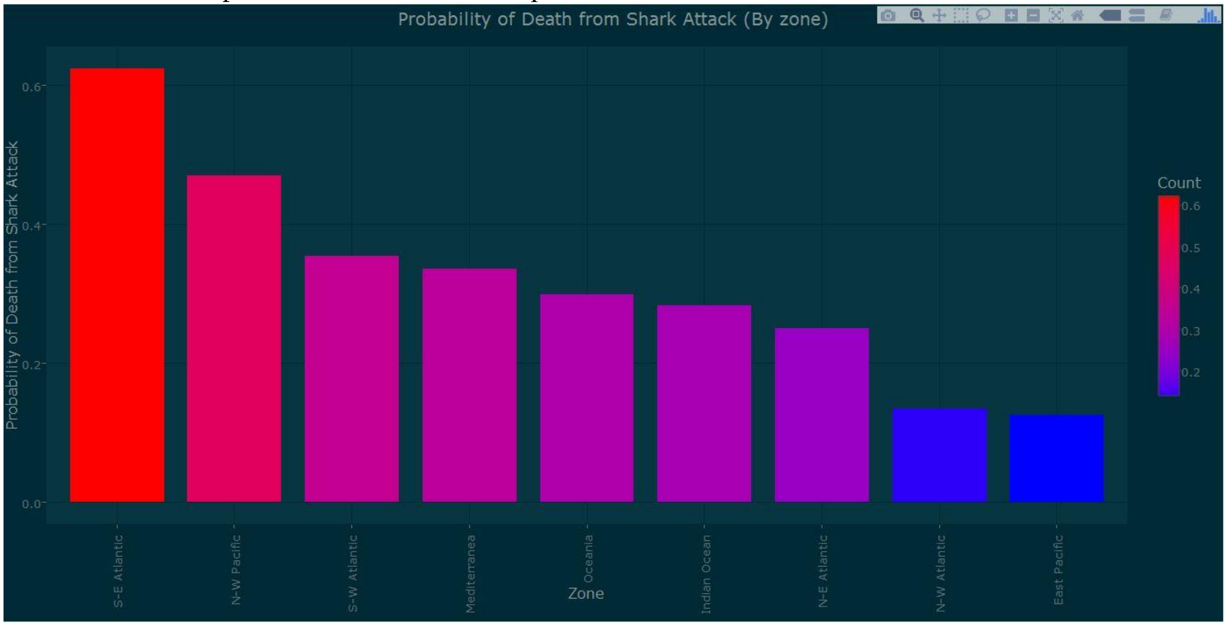
Firstly, the barplot in figure 13 shows the number of attacks for each area. This visualization allows comparing the number easily for each zone. Thus, you can see that North-West Atlantic and Oceania account the most attacks while all the east part of the Atlantic (North and South) considers only 40 attacks when adding the North and the South.

Then, we can compare the presence of dominant shark species in each zone. In figure 14, you can see that in the large majority of attacks, sharks could not be identified. But when we take into account observations when sharks could be determined, we can see that the White shark is responsible for the higher number of attacks. Tiger Shark and Bull Shark are the two other species responsible for a significant number of attacks.

Figures 15, 16 and 17 shows the repartition of attacks of these 3 species in the geographic zones. One more time, I chose to use bar plots which give a clear and quickly interpretable result. We can see that White Shark is mostly present in East Pacific, the Indian Ocean and Oceania which means from the Eastern African coast until the Western American coast while Tiger Shark was mainly detected in North-West Atlantic and Oceania, that is also the case for Bull Shark.

Now, with the bar plot in figure 18, we can see the importance of each type of injury for each zone, but this does not seem very relevant according to the fact that the repartition of injury types appears to be quite similar for every zone. However, this can be more interesting to see the proportion of injury types for each species. Figure 19 could be useful for that but it is not really visible, so plotting a pie chart for each species might be more appropriate. It seems useless to plot the pie chart for each species, so only the ones for the White Shark, Tiger Shark and Bull Shark are shown in the appendix (figure 20, 21, 22).

Another interesting issue to analyze is the repartition of the event “Fatality” in the different zones. To do that, I firstly plotted a bar plot showing the proportion of each instance of the variable “Fatality” in the geographical areas (figure 23). But another option was to visualize the probability to die when we are attacked by a shark for each zone. This probability was calculated by creating a new 5 columns data frame (figure 24). The feature “n” lists the number of values for each instance of the variable “Fatal” and for each zone, the feature “sum” just counts the number of attacks for each zone. Then, the feature “probability” is given by dividing the value of “n” by the value of “sum”, which finally give the probability of each unique value of “Fatal” for each zone. Finally, we just have to use the probabilities for “Fatal” equal “N”:



2. Prediction

I choose to try to predict the new binary variable “Fatality” created from the variable “Fatal” using a classification machine learning algorithm, in the objective to serve as a prototype for after in future work trying to resolve other issues, for instance trying to predict which species of shark attacked. Indeed, in many observations, the species was unknown so that other classification problems could be raised, for example if scientists are studying one particular species, they can try to predict if this species or no caused the attack.

Two modelisations were done: the first one using the basic features given at the beginning of the study, except for the variable “Injury” which contains the information if the attack was fatal or no. In the second modelisation, we added the new variables created during the study, except for “Injury_Type” which also contains the information if the attack was fatal. Then, we can compare if the information provided by these new was relevant for the prediction or no.

To make this prediction, we had to handle the missing values of the features “Age’ and “Time” that we firstly ignored. I choose just to delete its, given the fact that we still have nearly 1500 observations after doing so.

Now, we will see which approach was selected to obtain the modelisation (the approach was the same with both datasets used).

In the first time, the dataset was split between a training dataset, containing 70% of the observations, and the test dataset, containing 30%. To do that, we used the function “sample” which randomly generate a sample from a dataset; the size is chosen in the parameters.

Then, we had to choose an algorithm able to produce a model which predict the value of “Fatal”. And the Support Vector Machin (SVM) algorithm seems to be reluctant for this problem. Indeed, this algorithm is easy to implement in R, a function already exists to help us adjusting parameters, and the results are easily interpretable and understandable. Moreover, according to Eric Biernat and Michel Lutz [6], it is a robust algorithm able to find non-linear patterns. The principle consists in the search for an hyperplane separating individuals into two categories and in maximizing the margin between the decision border and the closest observations.

To implement this, we used the “svm” function on the training set, specifying that the type of modelisation was a “classification” and the kernel as “linear”. Then we used the function “tune” to find the best values for the hyperparameters “cost” and “gamma.” After that, we used the function “predict” on both training and testing sets to obtain predictions of “fatality” value for both. Finally, to calculate the error of each prediction, we plotted the table giving the number of “true 1”, “false 1”, “true 0” and “false 0”. The error is obtained when dividing the sum of “false 1” and “false 0” by the total number of observations. The results are summarized in the following table:

	Dataset without new features	**Dataset with new features**
Training error	4.2%	6.5%
Testing error	15.7%	12.6%

The first observation that we can make is that both results are quite good, the errors are not too high. However, the model for the dataset without the new features seems to overfit comparing to the dataset with new features. Indeed, the difference between training error and testing error is more important in the first dataset, which means that the model is too closely fit to the training dataset. The addition of the new features limits this overfitting.

IV. Benefits of the study and a case study

This kind of analysis can be very helpful for experts and institutions having for goal to find solutions to decrease shark attacks. Indeed, we can understand which areas are most at risk, which sharks are the most dangerous and in which areas they are acting. According to our study, two areas seem to have a higher number of attacks than the other zones: North-Atlantic and Oceania. Moreover, we saw that the most common activities involved in attacks were leisure activities and especially Swimming and Surfing. So, this substantial number of attacks in these zones can be explained by the fact that they are areas very appreciated by both tourists and surfers, particularly Australia, Florida and all islands in the Caribbean sea.

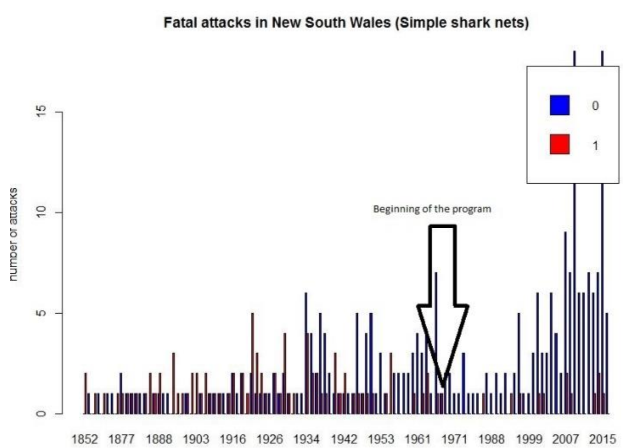
About shark species, we saw that White Shark, Tiger Shark and Bull Shark was the most often involved in attacks. That is certainly why they were listed in the book “The World’s most Dangerous animals” [7].

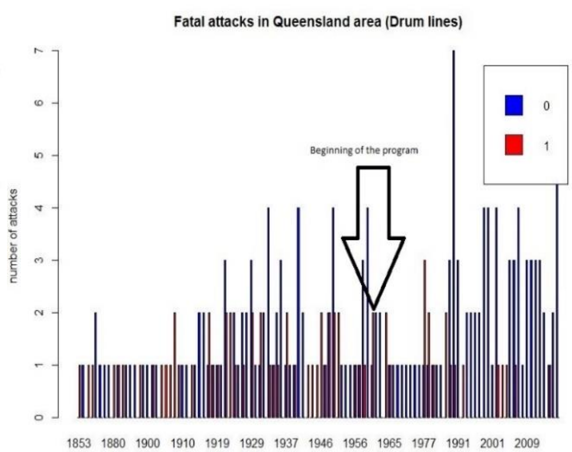
This Global Shark Attacks File was mainly studied by institutions located in risks areas. For instance, Florida Museum of Natural History based in the University of Florida [8] has created a dynamic world map in which we can see the repartition of attacks in the world, and when we move the mouse cursor on a country, we see the exact number of attacks for this country. Moreover, in clicking in this country, we have a map of this country showing the repartition of attacks in the regions of this country. Another famous study from the University of Florida used this file to understand the types of injuries and wounds caused by an attack [9].We can also quote the Australian Shark Attack File [10].

In order to tackle the problem of shark attacks, some authorities have chosen to install anti-shark systems, often composed of nets. The dataset can be useful to evaluate the efficiency of some installations. A good comparison of three programs has already be done by Dudley [11].

The most efficient one seems to be the system installed in Hong-Kong. In 1995, three swimmers were fatally attacked by sharks. The government then decided to install permanent barrier nets. Then no attacks were enumerated since 1995 as you can see in this bar plot:

In Australia, the New South Wales and Queensland regions installed two different systems. In New South Wales, authorities have decided to install simple shark nets while in Queensland drum lines were added with the nets. These installations are accompanied by various measures and programs ([11] and [12]). The following bar plots can permit to see the effects of these installations and programs established in the 1960s ([11] and [12]) for both regions on the frequency of shark attacks:





In the beginning, both programs seem to show good results in decreasing the number of attacks, especially in New South Wales where no fatal attack was reported until 1987. From this year, the number of attacks increased drastically. In Queensland, the number of attacks stagnated at 1 per year for approximately 10 years but after we see a new increase. Finally, we can affirm that the program in Queensland seems better to tackle new issues of the 21st century. Indeed, after 2000, New South Wales saw its number of attacks skyrocketing with on average six times more deaths between 2000 and 2010 than between 1980 and 1990. Queensland also knew an increase but significantly more moderated.

Conclusion and further work

This study shows an overview of how this dataset can be useful for research on shark attacks. It also provides a geographical approach by analyzing the behavior of attacks in defined geographic zones. We give a model predicting if the attack was fatal or no, which can pave the way for further predictions. But some features have not really be taken into account, especially “Age” and “Time”. Further work can be to use these variables to analyze more precisely the behavior of victims. Moreover, in the feature “Species”, it was often specified the approximate size of the shark. This could be taken into account for a prediction of the real species of the shark in the feature “Species_final” when the value is “Other or unknown”.

References

[1] A. MacCormick. *Shark Attacks: Terrifying True Accounts Of Shark Attacks Worldwide,* St. Martin's Press, 1998

[2] R. Ferrara. “Why study shark accidents? ”, Shark Research Institute, http://www.sharkattackfile.net/whystudy.htm, 2005

[3] JK. Bourne and P. Westwick. “Inside the curl: surfing’s surprising History”, National Geographic, http://news.nationalgeographic.com/news/2013/08/130803-surfing-surprising-history-hawaiian-culture-extreme-sports/, August 2013

[4] U. Gyr. “The History of Tourism: Structures on the Path to Modernity Tourism”, 2010.

[5] LR. Taylor. *The Caribbean Sea*, Blackbirch Press, 1998

[6] E. Biernat and M. Lutz. *Data Science: fondamentaux et études de cas*, 2015.

[7] K. Baker, J. Baker. *The World's Most Dangerous Animals*, 2016.

[8] Florida Museum of Natural History, https://www.flmnh.ufl.edu/fish/isaf/shark-attacks-maps-data/world/

[9] Lentz, Ashley K.; Burgess, George H.; Perrin, Karen; Brown, Jennifer A.; Mozingo, David W.; Lottenberg, Lawrence. “Mortality and Management of 96 Shark Attacks and Development of a Shark Bite Severity Scoring System”. *The American Surgeon* 76.1:101-106(6), 2010.

[10] https://taronga.org.au/conservation/conservation-science-research/australian-shark-attack-file

[11] Dudley, S. F. J. "A comparison of the shark control programs of New South Wales and Queensland (Australia) and KwaZulu-Natal (South Africa)." *Ocean & coastal management* 34(1): 1-27, 1997

[12] Paterson, R. A. "Effects of long-term anti-shark measures on target and non-target species in Queensland, Australia." *Biological Conservation* 52.2: 147-159, 1990.

[13] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. “Enterprise data analysis and visualization: An interview study ”. *IEEE Trans. on Visualization and Computer Graphics*, 18(12):2917–2926, 2012.

APPENDIX

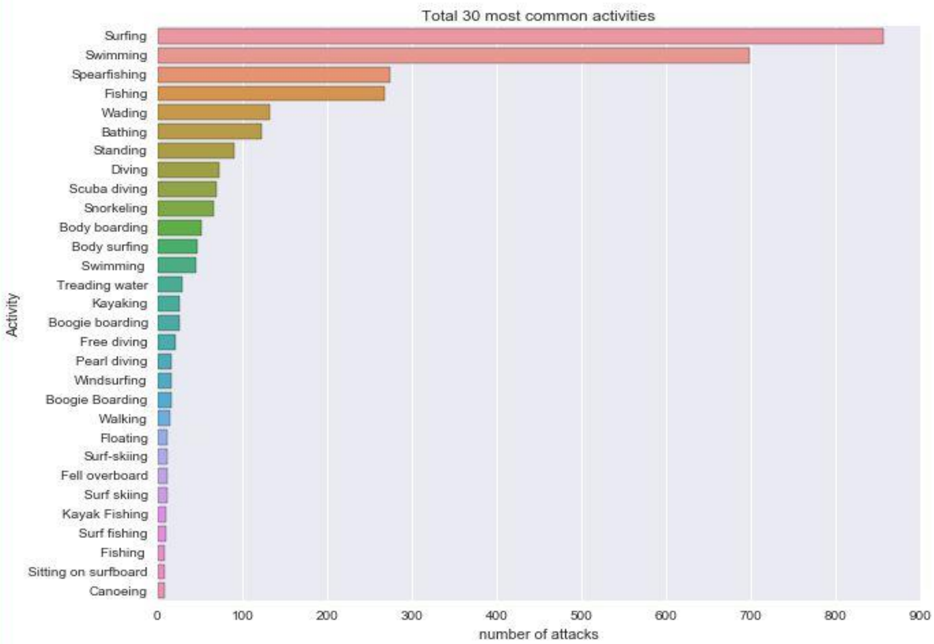


Figure 1

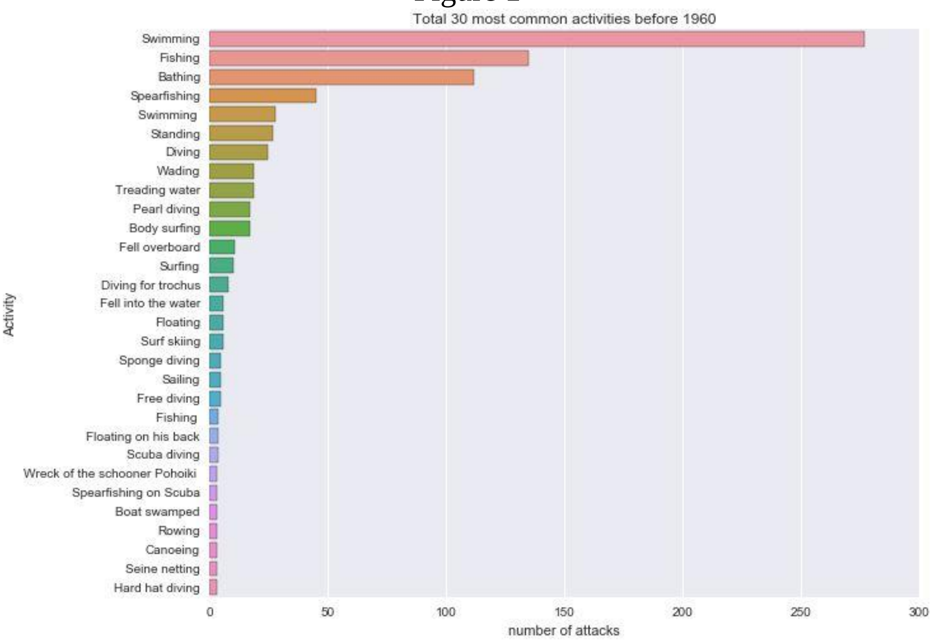


Figure 2

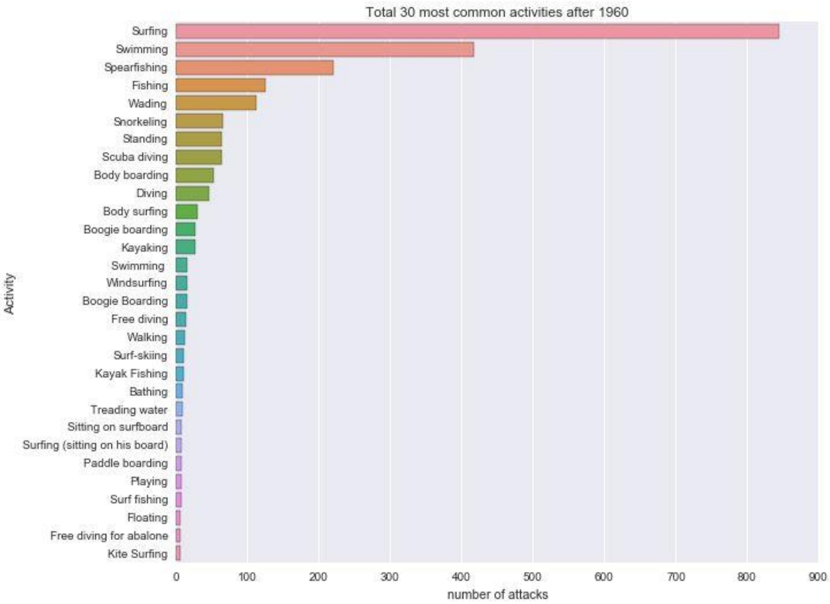


Figure 3

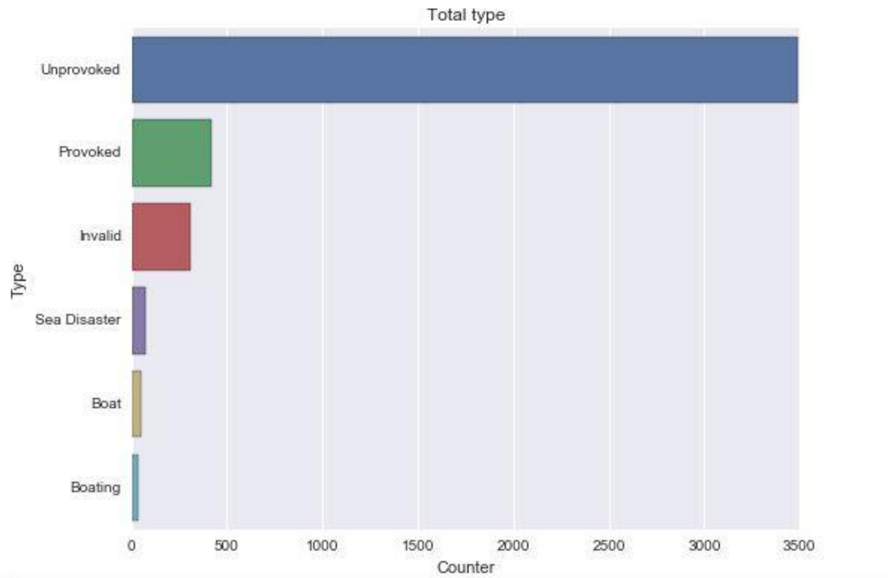


Figure 4

project_PODS.Rmd × | attacks_predict × | attacks × | shark × | Country_attack

Country	Area	Location
USA	Massachusetts	Manomet Point, Plymouth, Plymouth County
USA	Florida	Fletcher Beach, Hutchinson Island, Martin County
SPAIN	Catalonia	Salou
USA	Florida	New Smyrna Beach, Volusia County
USA	Florida	Ponce Inlet, Volusia County
USA	Maryland	Assateague National Seashore
SPAIN	Alicante	Benidorm
USA	North Carolina	Figure Eight Island, New Hanover County
USA	South Carolina	Surfside Beach, Horry County
USA	Florida	Apalachicola Bay
AUSTRALIA	Western Australia	Gnaraloo
USA	Florida	3 to 4 miles west of Indian Pass, Gulf County
USA	Florida	Lori Wilson Park, Cocoa Beach, Brevard County
USA	Louisiana	Lake Ponchartain off Southshore Harbor, New Orleans
USA	South Carolina	Folly Beach, Charleston County
USA	Florida	Cocoa Beach, Brevard County
USA	Florida	South of Cocoa Beach, Brevard County
USA	Florida	Table Beach, Brevard County
SOUTH AFRICA	Western Cape Province	Muizenberg
USA	North Carolina	Sunset Beach, Brunswick County
USA	Florida	Indialantic, Brevard County
USA	Hawaii	Paia Bay, Maui

Figure 5

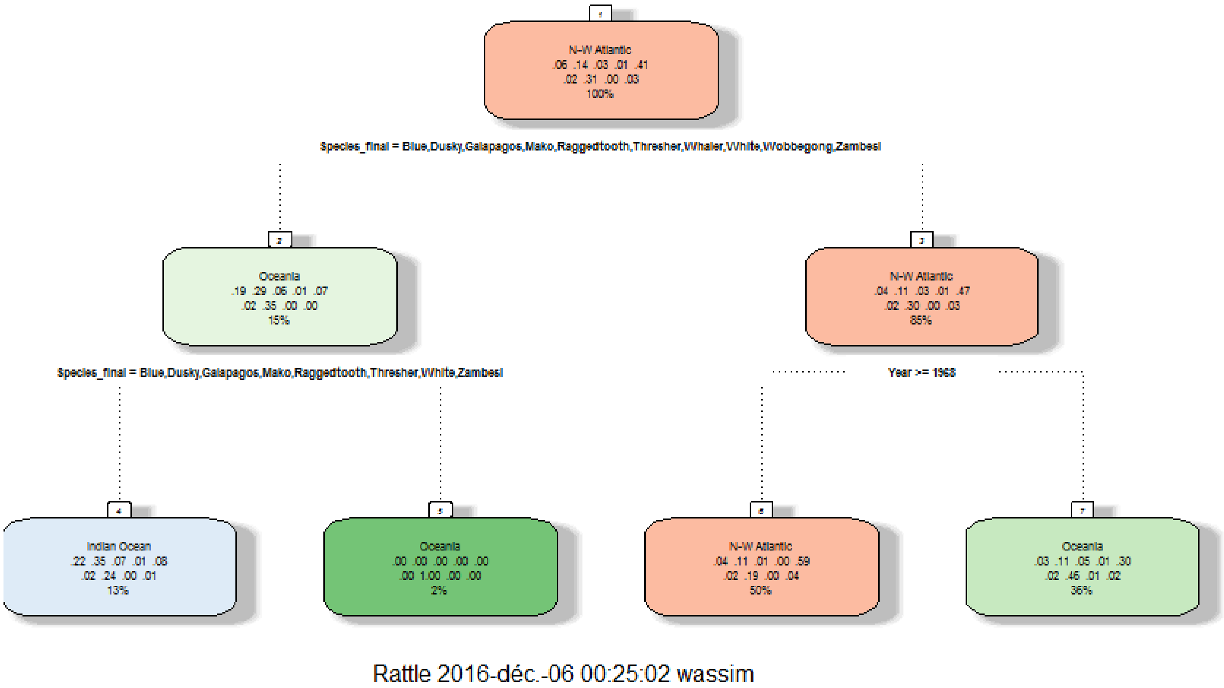


Figure 6

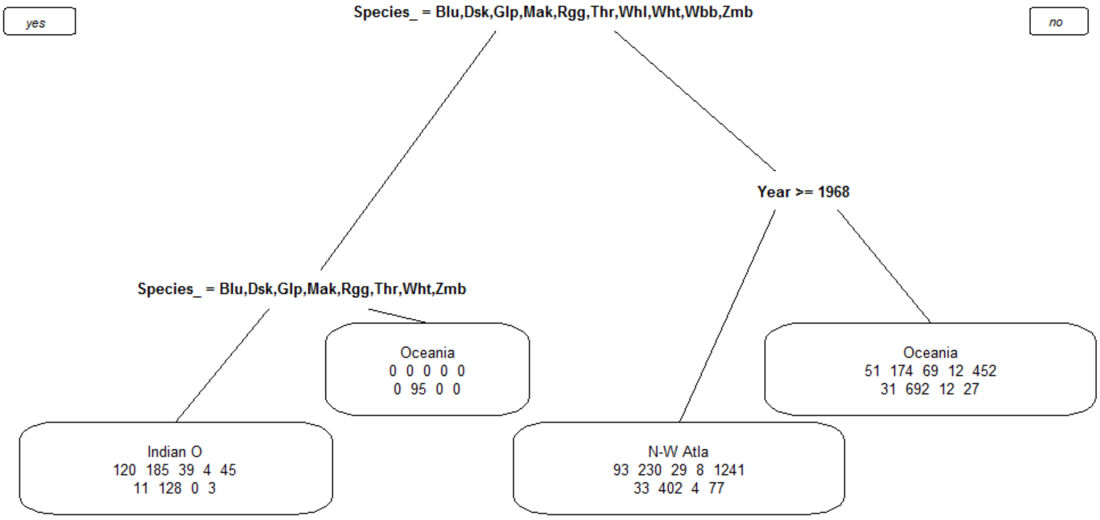
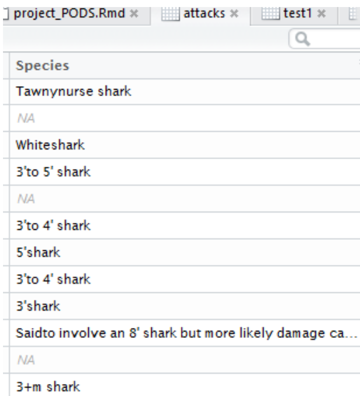
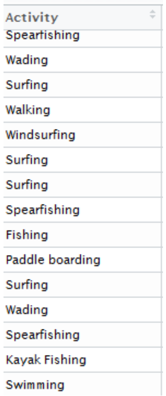


Figure 7



Species
Tawnynurse shark
NA
Whiteshark
3'to 5' shark
NA
3'to 4' shark
5'shark
3'to 4' shark
3'shark
Saidto involve an 8' shark but more likely damage ca…
NA
3+m shark

Figure 8



Activity
Spearfishing
Wading
Surfing
Walking
Windsurfing
Surfing
Surfing
Spearfishing
Fishing
Paddle boarding
Surfing
Wading
Spearfishing
Kayak Fishing
Swimming

Figure 9

Injury
Lacerations to lower leg
Struck by fin on chest & leg
No injury: Knocked off board by shark
Minor injury to arm
Severe lacerations to shoulder & forearm
FATAL
Lacerations & punctures to lower right leg
No inury, board broken in half by shark
Foot bitten
Two toes broken & lacerated
Lacerations to right foot
Minor injury to ankle
Minor injury to ankle

Figure 10

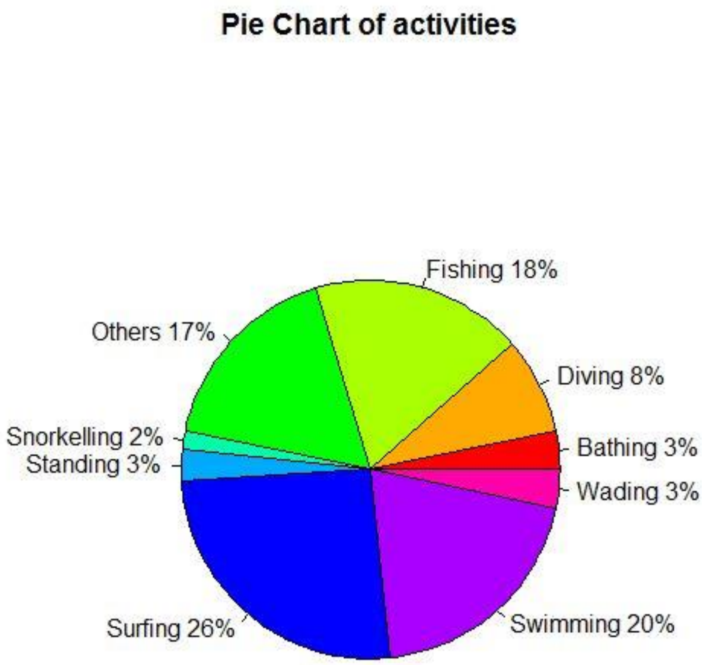


Figure 11

acks_predict × | attacks × | shark × | Country_attack_by_zone_fatal

Species_final	zone	fatality	Injury_Type	Activity_Type
White	N-W Atlantic	0	No Injury	Others
Other or unknown	N-W Atlantic	0	Bitten	Fishing
Other or unknown	Mediterranea	0	Lacerations	Others
Spinner	N-W Atlantic	0	Others	Surfing
Spinner	N-W Atlantic	0	Lacerations	Surfing
Other or unknown	N-W Atlantic	0	Lacerations	Fishing
Other or unknown	Mediterranea	0	Lacerations	Swimming
Other or unknown	N-W Atlantic	0	Lacerations	Surfing
Other or unknown	N-W Atlantic	0	Bitten	Standing
Bull	N-W Atlantic	0	Lacerations	Fishing
Reef	Oceania	0	Lacerations	Fishing
Bull	N-W Atlantic	0	Lacerations	Standing
Other or unknown	N-W Atlantic	0	Others	Swimming
Bull	N-W Atlantic	0	Lacerations	Swimming
Other or unknown	N-W Atlantic	0	Lacerations	Others
Other or unknown	N-W Atlantic	0	Lacerations	Swimming

Figure 12

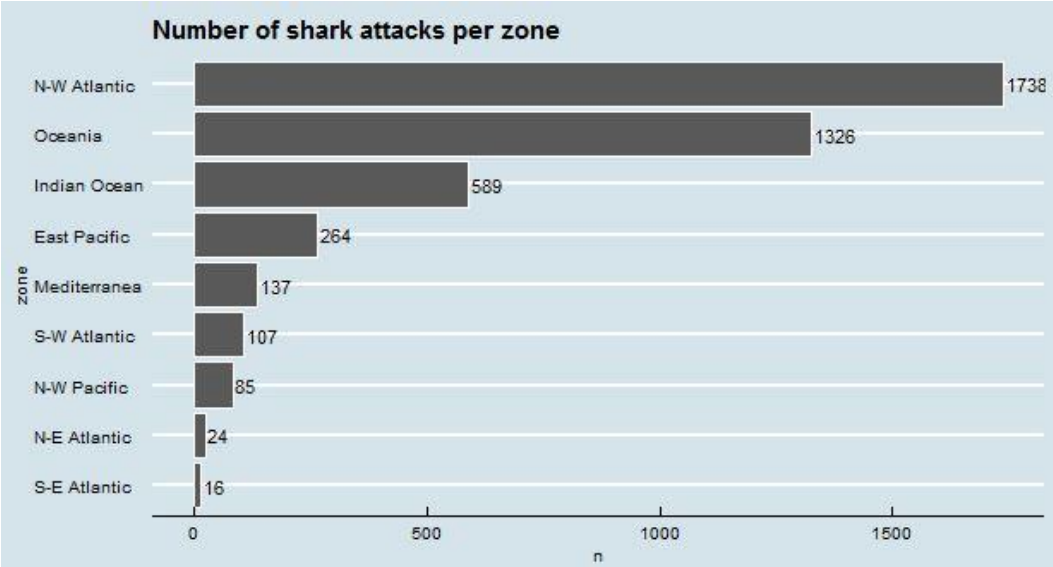


Figure 13

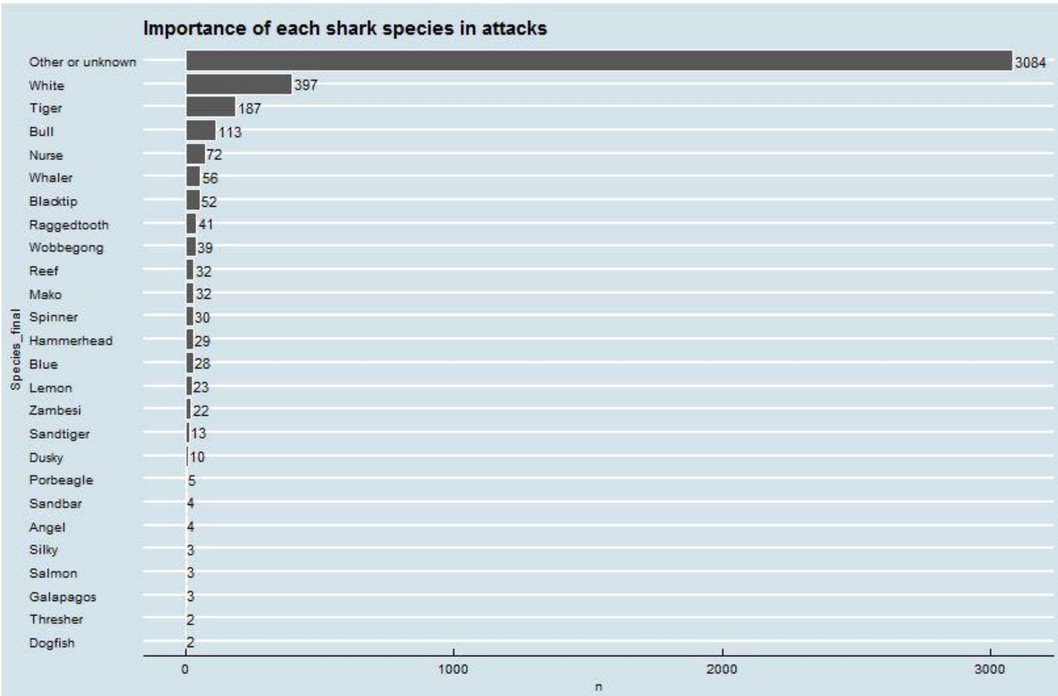


Figure 14

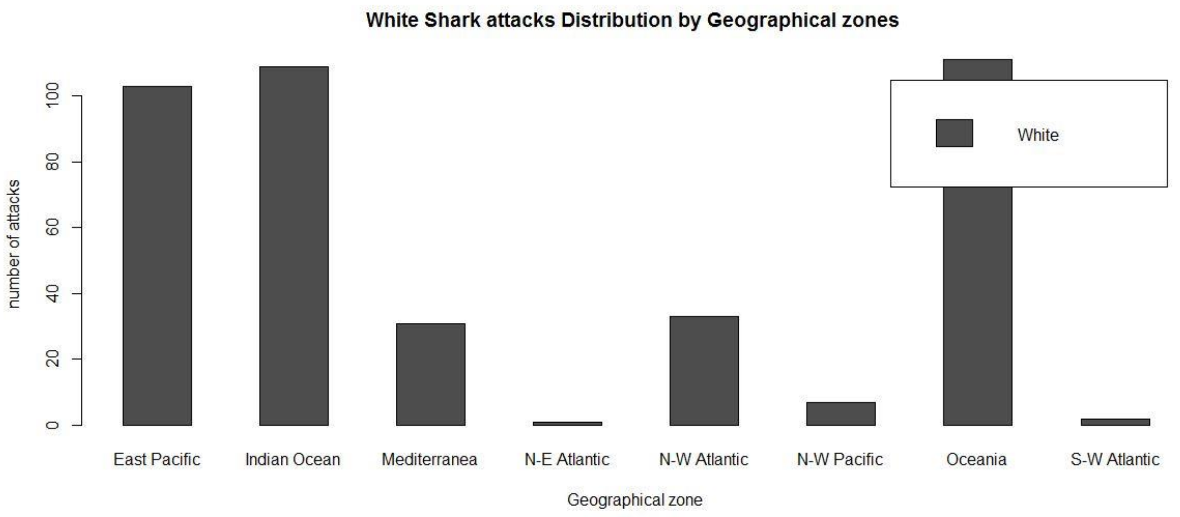


Figure 15

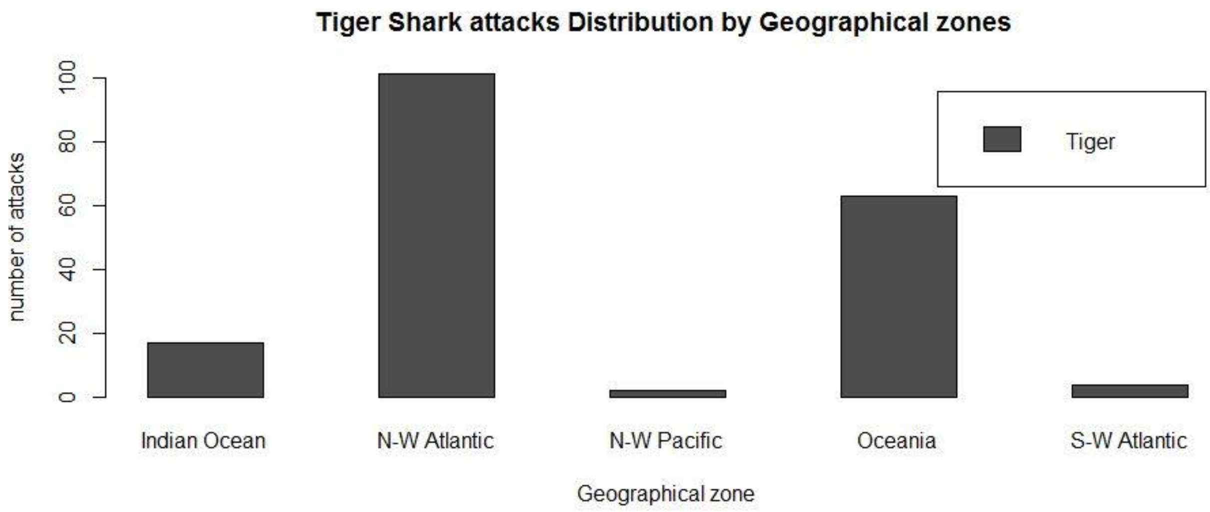


Figure 16

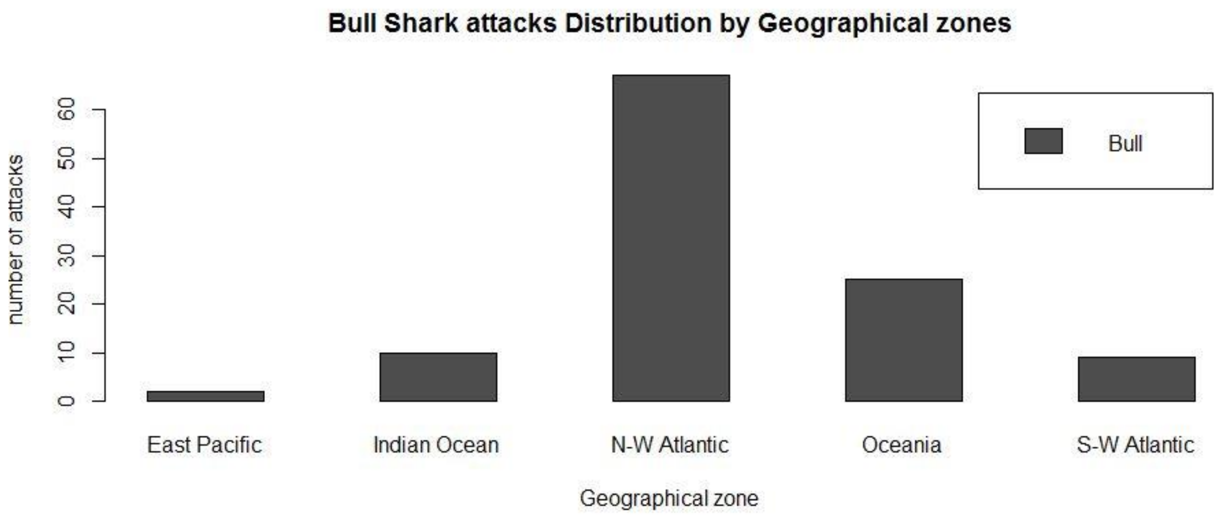


Figure 17

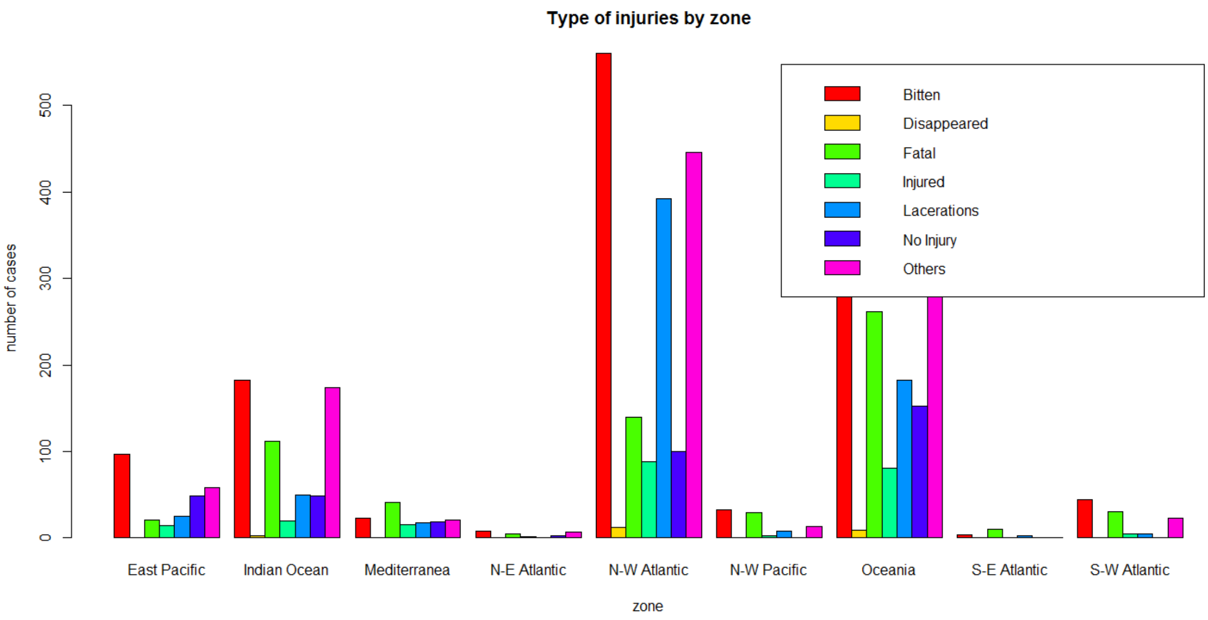


Figure 18

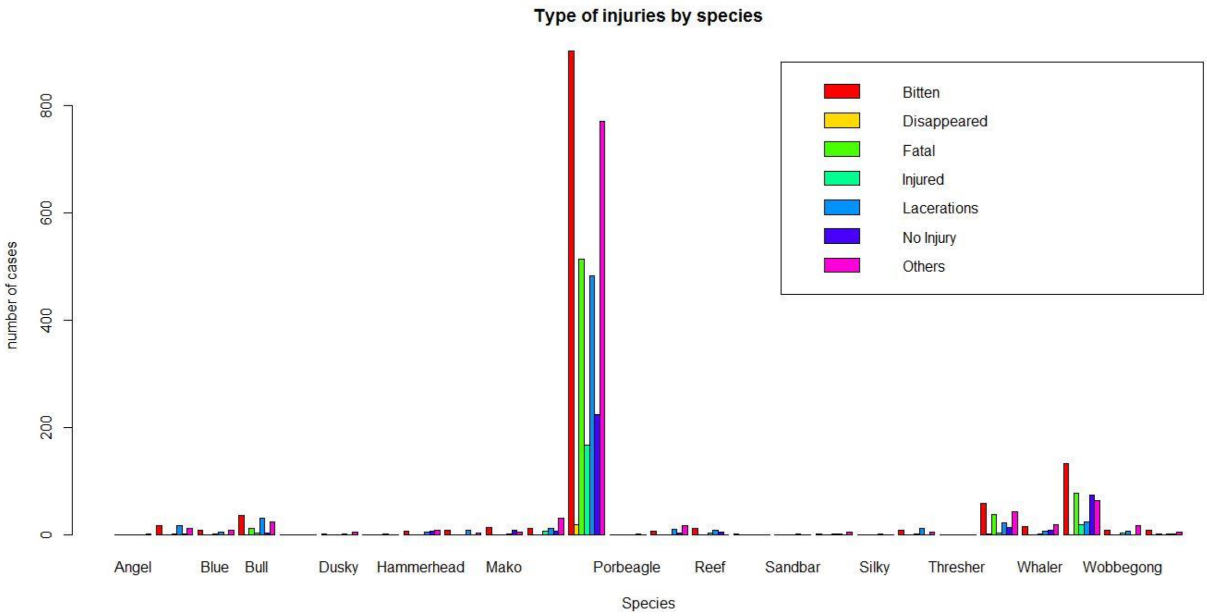


Figure 19

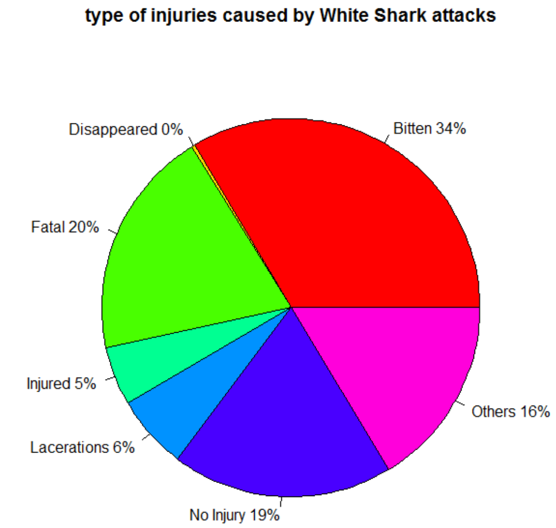


Figure 20

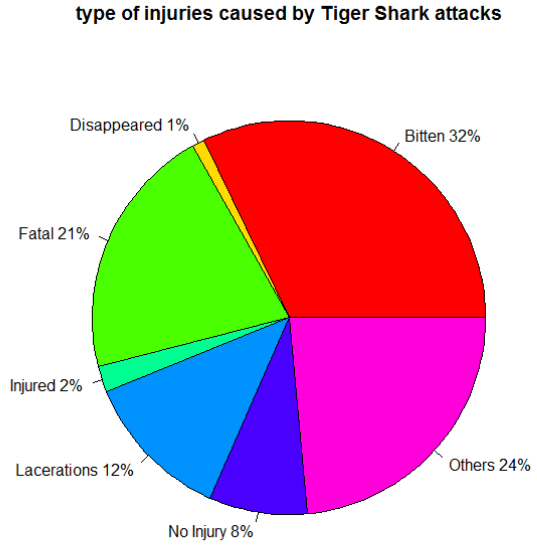


Figure 21

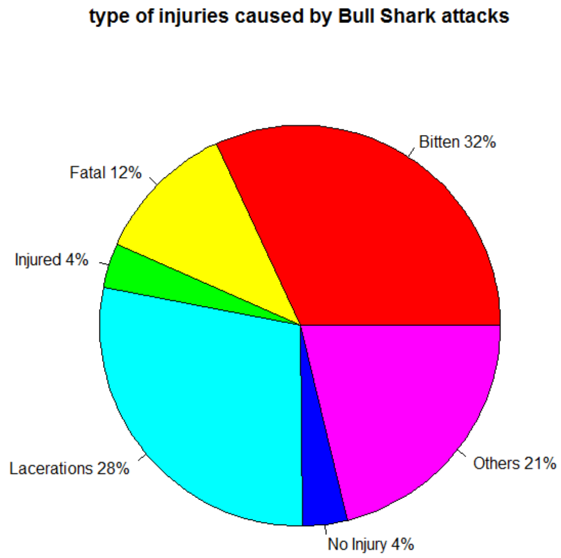


Figure 22

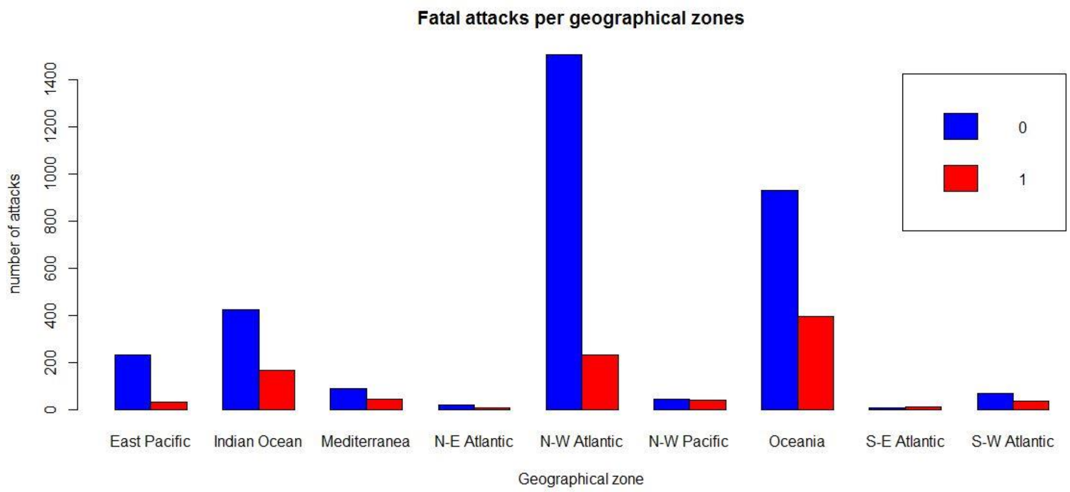


Figure 23

	zone	Fatal	n	sum	Proba
1	East Pacific	N	231	264	0.875000000
2	East Pacific	Y	33	264	0.125000000
3	Indian Ocean	N	419	589	0.711375212
4	Indian Ocean	UNKNOWN	3	589	0.005093379
5	Indian Ocean	Y	167	589	0.283531409
6	Mediterranea	N	88	137	0.642335766
7	Mediterranea	UNKNOWN	3	137	0.021897810
8	Mediterranea	Y	46	137	0.335766423
9	N-E Atlantic	N	18	24	0.750000000
10	N-E Atlantic	Y	6	24	0.250000000
11	N-W Atlantic	N	1497	1741	0.859850661
12	N-W Atlantic	UNKNOWN	7	1741	0.004020678
13	N-W Atlantic	Y	234	1741	0.134405514
14	N-W Pacific	N	44	85	0.517647059
15	N-W Pacific	UNKNOWN	1	85	0.011764706
16	N-W Pacific	Y	40	85	0.470588235
17	Oceania	N	919	1322	0.695158850
18	Oceania	UNKNOWN	7	1322	0.005295008
19	Oceania	Y	396	1322	0.299546142
20	S-E Atlantic	N	6	16	0.375000000
21	S-E Atlantic	Y	10	16	0.625000000
22	S-W Atlantic	N	69	107	0.644859813

Figure 24