MULTIMEDIA UNIVERSITY®

Bridging Boundaries, Connecting Minds

Faculty of Computing and Informatics

(FCI)

TDS3301

DATA MINING

Assignment

Part 1 Exploratory Analysis

Prepared by:

Name	ID
Yap Kit Boon	1141124580
Thong Tong Lin	1132702398
Ooi Yi Jie	1131122872
Chew Siu Min	1122703126

PART 1: EXPLORATORY DATA ANALYSIS

Name: **Global Shark Attacks Dataset**
From: https://www.kaggle.com/teajay/global-shark-attacks

A. Describe the dataset in your own words.

For this assignment, we have selected the '**Global Shark Attacks**' dataset obtained from Kaggle website. The author collects this dataset from numerous accidents reported by the shark attack's victims. There are 5992 rows and 16 columns in this dataset, the rows represent observations (shark attack reports) and columns represent variables (details in shark attack report).

Some of the columns recorded are the case number, date, time, area, activity, user's information, extent of injury and the species of shark. The reports are collected over few hundred years; the earlier reports were recorded long time ago while the later reports are made via online reporting system or through investigator. As a result, the earlier reports tend to be inconsistent and contain a lot of missing values compared to the later reports. By just scanning through the dataset, we can observe that majority of the accidents happened outside Asia region, most of the victims are male, most of the accidents are non-fatal, and both 'shark species' and 'time' have most missing values compared to other variables. However, the accurate statistics of the dataset shall be obtained after data analysis.

B. What possible insights can be obtained from mining the chosen dataset?

The chosen dataset is mainly made to get insights for the shark-human interactions over the years. Based on the variables of the dataset, one of the possible insights can be obtained is the **location where shark attacks occurred the most**. The location can be categorized in 'country', 'area' and 'location' whereas 'country' will be considered the most suitable variables due to the missing values of 'area' and 'location'. A map can be plotted to show the location according to the frequency of the attacks occurred. After mining the data, the **injury type of the shark attack** victims can be classified into categories to give a clear insight on the fatality rate of the global shark attacks over the year. Next, we can get the **top 10 triggering activities** that caused shark attacks according to the 'activity' variable from the dataset. We can also get an observation on **the frequency of time periods that shark attacks occurred** in a day. This gives an insight for people who involves in water activities about the dangerous time periods that they should be aware of. A virtualized data on **frequency of shark attacks from year 1950** can be obtained in a form of graph, to get to know the trend of shark-human interactions over the years.

C. What type of data mining technique (association rule mining, classification or clustering) would be relevant? Give an example, for example, if you think classification is suitable, describe what will be classified and what the possible classes are.

In this case, **classification** would be relevant for data mining technique. Classification is about prediction on unknown class label based on other attributes of records. One of the attributes in a records will be selected to be a class in classification model. For example, in Global Shark Attacks dataset, Injury and Fatal (Y/N) attributes could be classified. In order to make it be more organized and more easier for classification, we can group **Injury** and **Fatal (Y/N)** attributes to be one attribute such as **Injured or Fatal**. In the new attribute, there are only three possible values for each record which are **No injury, Injured, and Fatal**. Classification model will categorize the class of Injured or Fatal for dataset records.

D. Describe data quality issues, and be specific. Identify which attribute (column) has issues, or if the structure of the data has problems.

There are 5992 observations and 16 variables in the dataset. There are 12 variables that contain missing value which is **Country, Area, Location, Activity, Name, Sex, Age, Injury, Fatal(Y/N), Time, Species, and Investigator or Source**. From these variables, we observed that most the missing value from variable 'Country' is reported before the 1970s. The reason for 'Area', and 'Location' to contain missing value is because some of the 'Country' don't have 'Area' and 'Location', for e.g. Aruba cannot have 'Area' and 'Location' because it is the country itself but also an island.

There are 527 observations missing value in variable 'Activity', which means that we cannot calculate the exact value of how many cases of shark attack are due to what activity. For the variable 'Name' we noticed there are quite a number of missing value, some of them have 'male' and 'female' being inputted into the field. We are making a wild guess that they don't know what is the name of the victim so they keyed in the gender of the victim instead, which will confuse the data scientist is they keyed in into the wrong variable.

There are 567 observations missing value in variable 'Sex', but as we observed that this variable is not necessary to be included in the data mining later, so that's is not important for us to record down.

There are 2676 observations missing in variable 'Age', even though it contained a lots of missing value, but we are going to remove the observations that are before the 1950s. And this variable are not important for our data mining task later.

There are 3213 observations missing value in variable 'Time', which we think that this is a problem for our data mining task, because we cannot accurately calculate the time or period that shark will be possibly attack people.

Even though there are 2934 observations missing value in variable 'Species', but we think that is not the main problem need to be include in our data mining task later.

There are little amount of missing value in variable "Injury", "Fatal(Y/N)", and "Investigator or Source". Which is not a very big problem for data mining later.

There are some typo data in the dataset that we have to fix. And also many inconsistent data, so we also have to fix it to convenience our analysis later.

E. Perform a pre-processing task on the dataset chosen.

DATA REDUCTION

1. Remove 'Case.Number', 'Age', 'Name', 'Area', 'Location', 'Fatal..Y.N.','Investigator.or.Source' ,'Species' columns
2. Remove records with year earlier than 1950
3. Remove records with unconfirmed and unknown value
4. Remove records with missing activity data

DATA INTERGRATION

1. Categorize 'Type', 'Activity' and 'Injury' columns

DATA TRANSFORMATION

1. Extract month from 'Date' column and change colname into 'Month'

DATA CLEANING

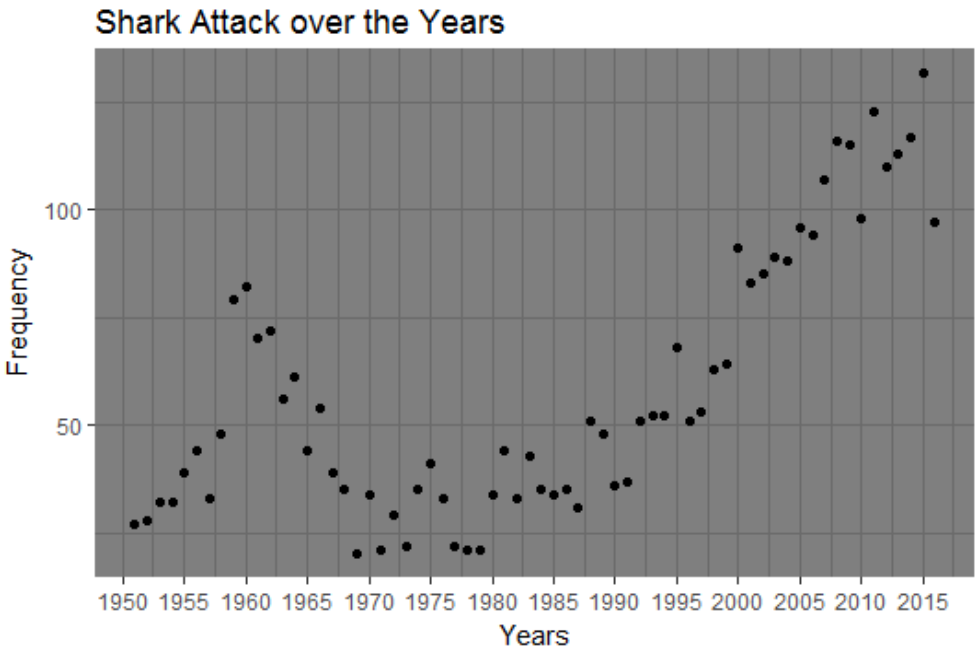
1. Fill in missing values
2. Correct fields with inappropriate representation (eg: correct "Inujury" into "Injury" in Injury column) on 'Injury' and 'Time' column

For further explanation, please refer to the .R code files.

EXPLORATORY ANALYSIS

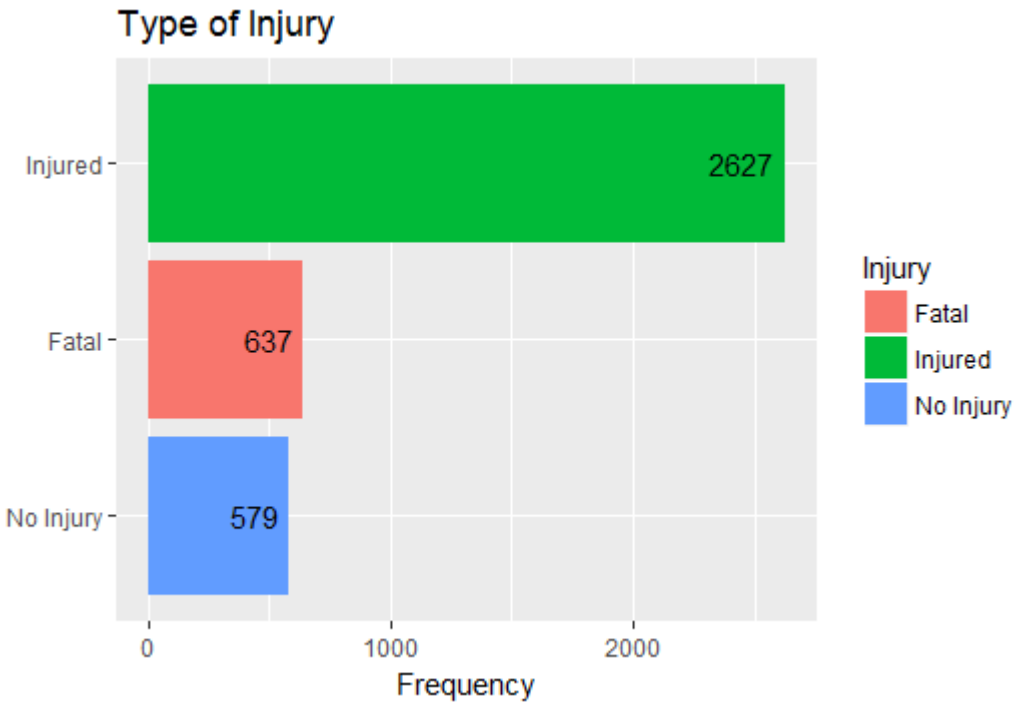
Shark Attacks over Years

After processing the data, we now have a complete dataset to conduct the exploratory analysis. The graph below is plotted based on the frequency of shack attacks over the years starts from 1950 until 2016. Based on the result, shark attacks occurred the most in 1960, then decreased gradually until 1970s and increased from 1990s until 2015. This may be caused by the increase in population of people involves in water activities throughout the years.



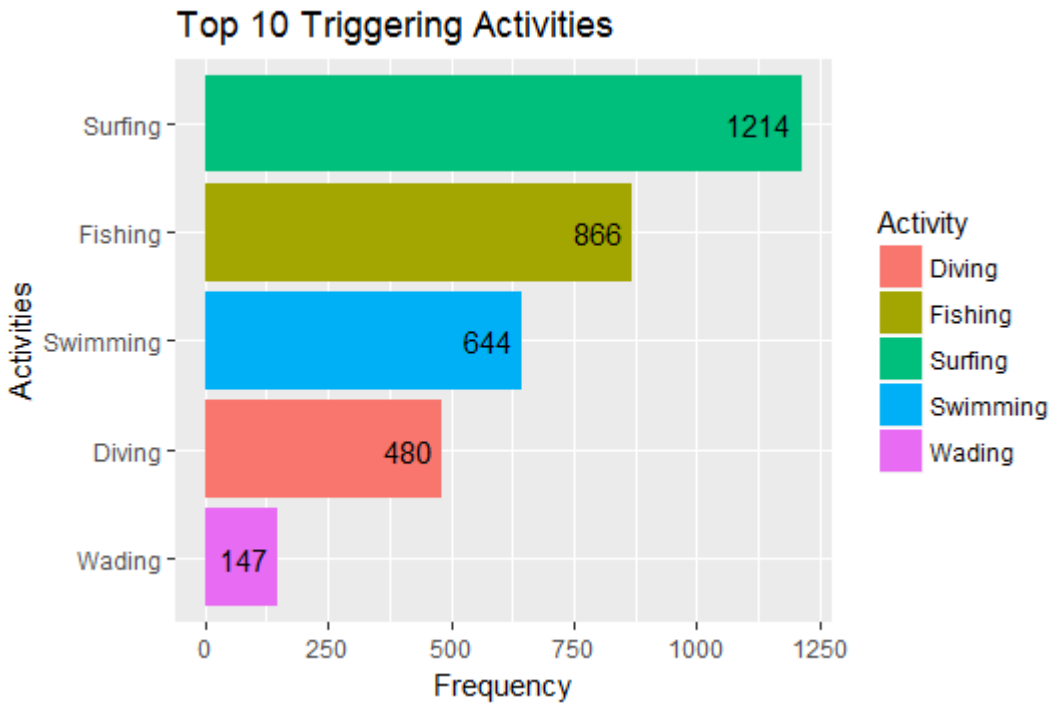
Injury Type

By classifying the injury type of the shark attack victims into categories, we now have the observation on the fatality rate of the global shark attacks. The ‘injured’ category is higher than ‘fatal’ and ‘no injury’ for approximately 2000 cases. Thus, we can predict that the majority of shark attacks are "hit-and-run" attacks where they attack, bite, and let go when they realize it's not a fish.



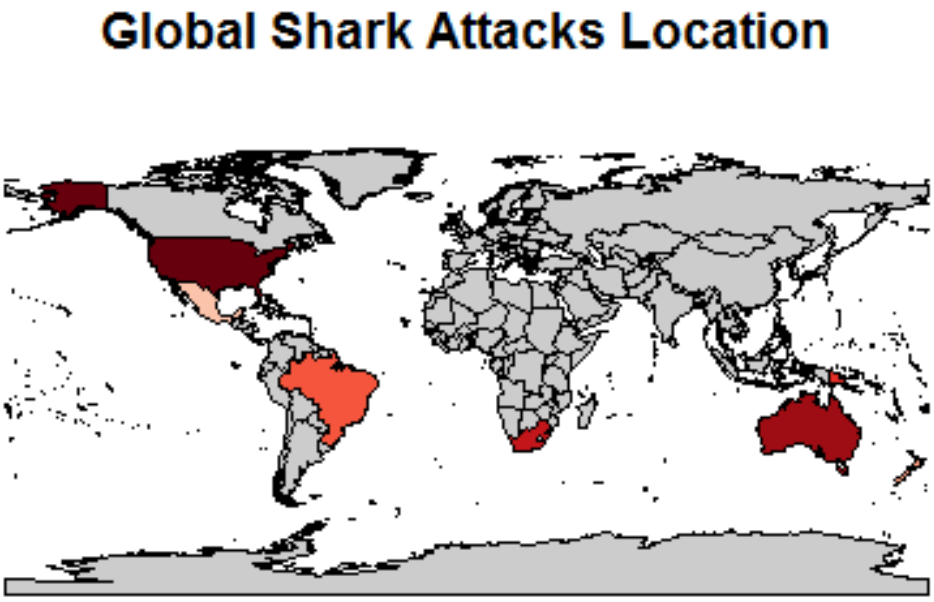
Trigerring Activities

There are various categories in triggering activities, however the others activities are not as popular as these top five.



Global Shark Attacks Location

Based on the plotted map, United States, Australia and South Africa turned out to be the countries where shark attacks occurred the most. According to the result, United States has a total of 1675 cases over the years. This is probably because of the high population of sharks in San Francisco Bay area, while there are high number of residents and visitors who participate in water activities every year. Furthermore, with the telecommunications technology in US, it is unlikely that a shark attack would go unreported.



Time of Shark Attacks in a Day

The line graph below is the frequency of shark attacks occurred based on 24 hours a day. We can see the peak hours by categorizing the time period. There is a trend of increasing shark activities from the dawn and decreasing at the dusk, while the peak hours are 11a.m. and 2p.m.. Based on the result, we can say that this phenomena is likely caused by the hotter air temperatures during daytime which means more people will hit the beach, resulting in the higher possibility of shark attacks at midday.