Practical Machine Learning – Pandas Exercises

Question 1 - Pandas – Shark Attack Dataset:

For each of the following questions you will use a dataset containing information on global shark attacks called attacks.csv.

Attribute Information:

The attributes recorded in the dataset are as follows:

0. Case Number
1. Date
2. Year
3. Type
4. Country
5. Area
6. Location
7. Activity
8. Name
9. Sex
10. Age
11. Injury
12. Fatal
13. Time
14. Species
15. Investigator or Source

Open this file using Pandas read_csv() function. The data file is stored in a different encoding format so you can use the following line to read the data into a dataframe.

```
df = pd.read_csv('attacks.csv', encoding = "ISO-8859-1")
```

Read the shark attack dataset into a Pandas Dataframe.

(i) Check how many missing values are present in this dataset. You should notice a very large number of missing values across all the columns (this is mainly due to the fact that the file has a trailing column of zeros at the end of the file). You should delete any rows that contains a significant number of missing values (a row should only be retained if it has 5 or more non-NAN values). You should notice a significant reduction in the number of rows in your data as well as the number of missing values across columns.

(ii) You will notice in the dataset that some entries in the fatality column are recorded as UNKNOWN, n, F, etc. We ignore these entries and only want to consider entries that are uppercase ‘Y’ or ‘N’. Extract all rows from the dataset that have an uppercase ‘Y’ or ‘N’ in the Fatal column. You should use the resulting dataset for all remaining questions.

(iii) What location globally has the highest number of shark attacks?

(iv) Determine the six countries that have experienced the highest number of shark attacks.

(v) Modify the code from the previous question to print out the six countries that have experienced the highest number of fatal shark attacks.

(vi) Based on the data in the Activity column are you more likely to be attacked by a shark if you are “Surfing” or “Scuba Diving”. Determine from the dataset what percentage of all recorded shark attacks were fatal.

(vii) Find out the most common activity for males that are attacked and the most common for females that are attacked.

(viii) Of all recorded shark attacks, what percentage were fatal?

(ix) For each individual country, print out the percentage of fatal shark attacks (number of fatal shark attacks expressed as a percentage of the total number of shark attacks). Some countries have recorded 0 fatal and non-fatal attacks. Your code should only consider countries where the number of non-fatal and fatal attacks are greater than 0.