**SHARK ATTACKS – DATA CLEANING**

1. Examining the data
2. Precising the purpose of data cleaning:

Travel agency is studying about best locations to organize water sports escapades.

They are asking me:

- which countries are to avoid when it comes to shark attacks danger;

- the risk that the shark attack will be fatal;

- if there is any seasonality in shark attacks, so they could avoid certain months.

So to my study, the most important columns are:

Country, Year - these cannot be nan (Year because shark migrations occur around every 20 years)

Fatal, Date - these, even if nan, I keep, because they still give me info that there was an attack in specific country

Data Cleaning Process:

1. I decided to drop columns at the end of the process, because even unnecessary columns I might need to clean the necessary columns.

2. I checked if there are exactly duplicated rows and apparently there were over 21k exact duplicates (Or I messed something up?)

3. Checking my important columns for unique values

4. Will start with Year – dropped everything before 2000 and dropped NaNs.

5. Checking Fatality column – change the column name so it doesn’t have spaces, cleaned the text, searched for the missing information in Injury column

# question in my head, shall I drop these 3 rows of nan or keep the for the sake of other info in the row? Decided to keep them as “UNKNOWN” so I don’t loose too much data.

6. I will now work on Country column – changed lower for capital to loose duplictaes, changed Nans to UNKNOWN, merged England with United Kingdom (even though missing some information)

7. Area seems OK

8. Location – it has a lot of distinct values, but seems OK

9. Date – I extracted month out of it and created separated column for unified month

10. At the end I have dropped unnecessary columns

11. I have change order of columns to make it easier to read

12. Send to csv

Obstacles encountered and lessons learned:

1. I lost a lot of time to define the purpose. I was trying few different ideas, dropping them because they seemed too difficult, starting again…
2. I realized that once we change our data, it is changed and there is no way back, it is wort to keep it in mind and make copies or push to github often.
3. Paradoxically, I find it as kind of obstacle, that pandas library is huge.

For each problem there are plenty of different functions, but when you dig deeper you find all those functions are not exactly what you are looking for, so you re spending a lot of time reading about functions that you wont use at the end. Plus I find it hard to understand which function returns what type and often had an error of mismatched types.