Data_Sharkleaner

A pandas project to clean a shark attack database (kaggle)

# INDEX:

1. Set working directory, load modules
2. Load raw data frame
3. Explore basic properties (shape, info..)

25,723 rows
and 24 columns

# 4. Cleaning steps

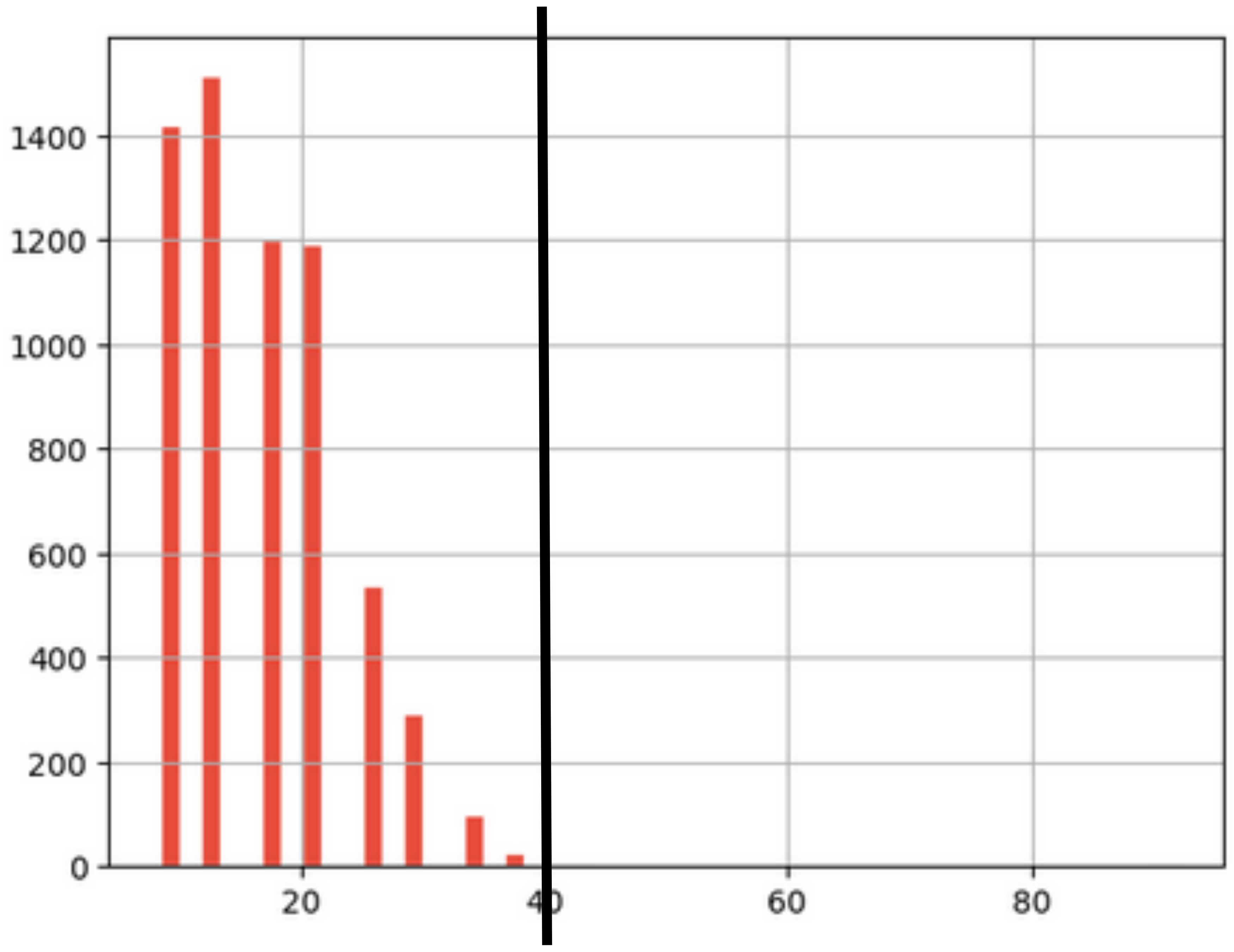
4.0 Rename columns if necessary

I replaced the column name "se" for "sex"

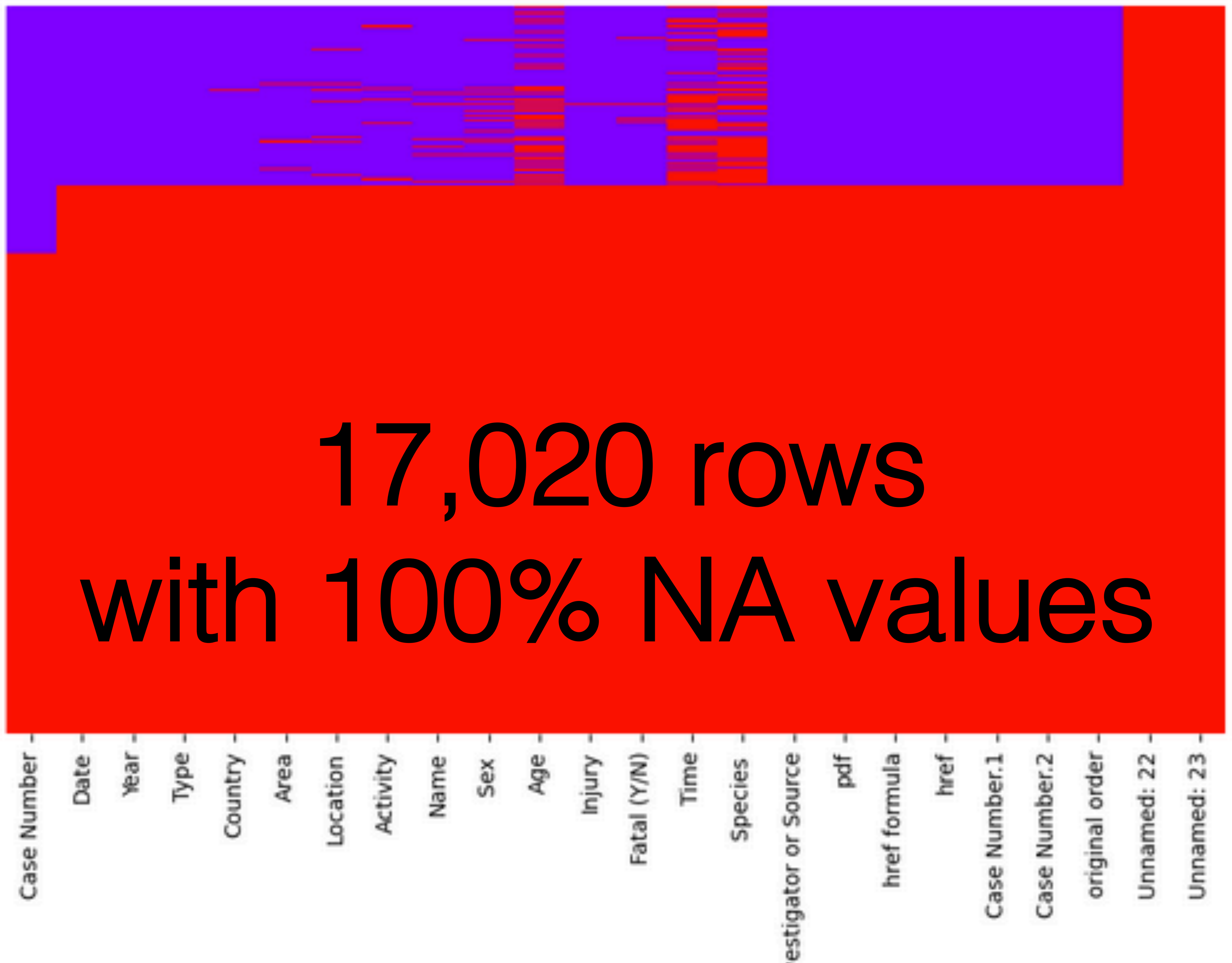
4.1 NA'S per row - drop rows with 100% of NA's

4.2 Check and drop duplicated rows

Histogram of % NA per row

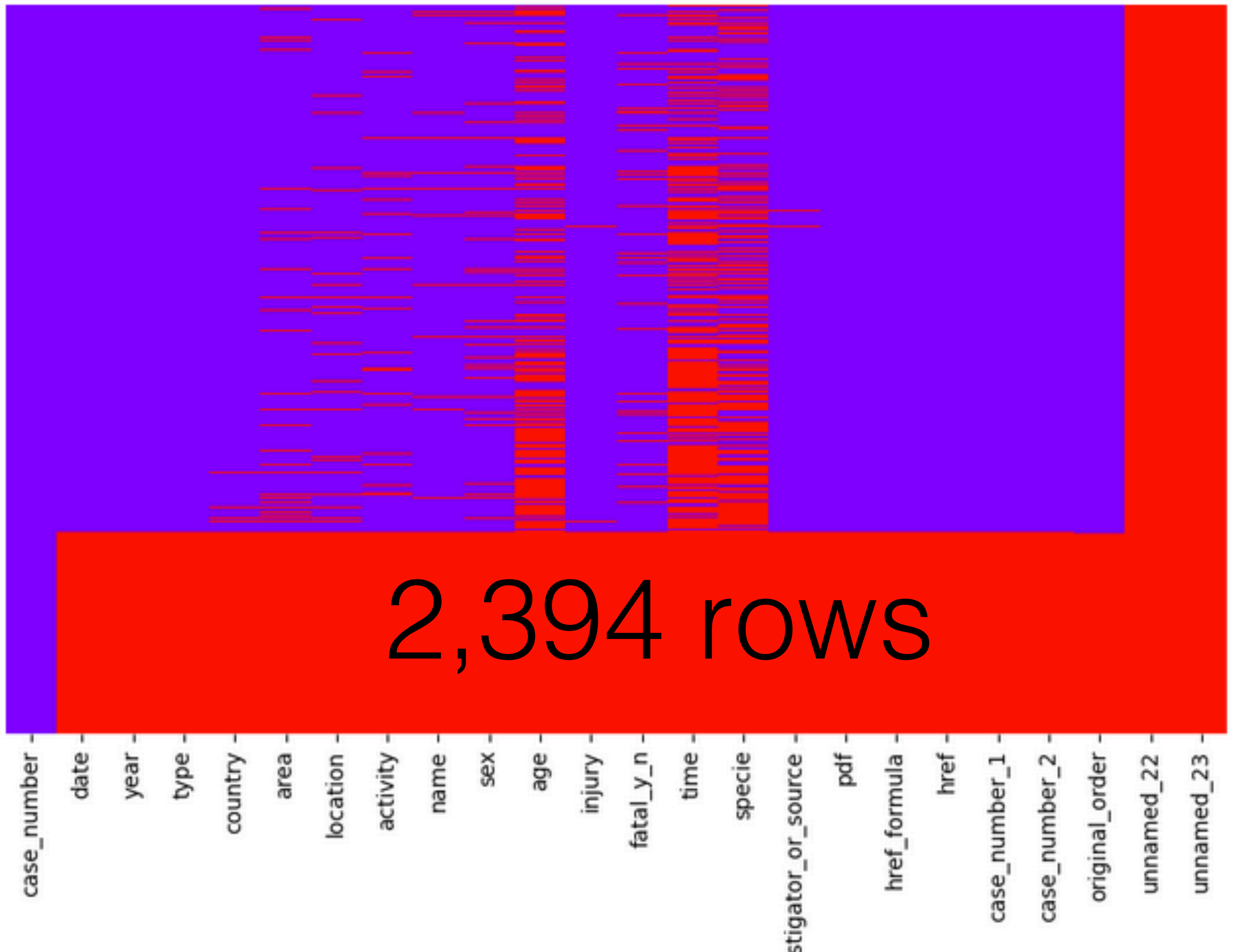


Drop rows with 100% of NAS'



17,020 rows
with 100% NA values

Drop rows with >40% of NAS'



2,394 rows

## Investigate the columns by their unique_count/freq ratio

Top rows: indicate there are MANY LEVELS with very LOW FREQ:

'case_number' should be an identifier and not present duplicated values!!!

That behavior is expected for 'case_number' an 'date' data types

According to wikipedia, there only exists 195 countries in the world but this column contains 212 unique values!!!! Check in the next cell bellow.

Same suspicion wiht "species" or "activity"

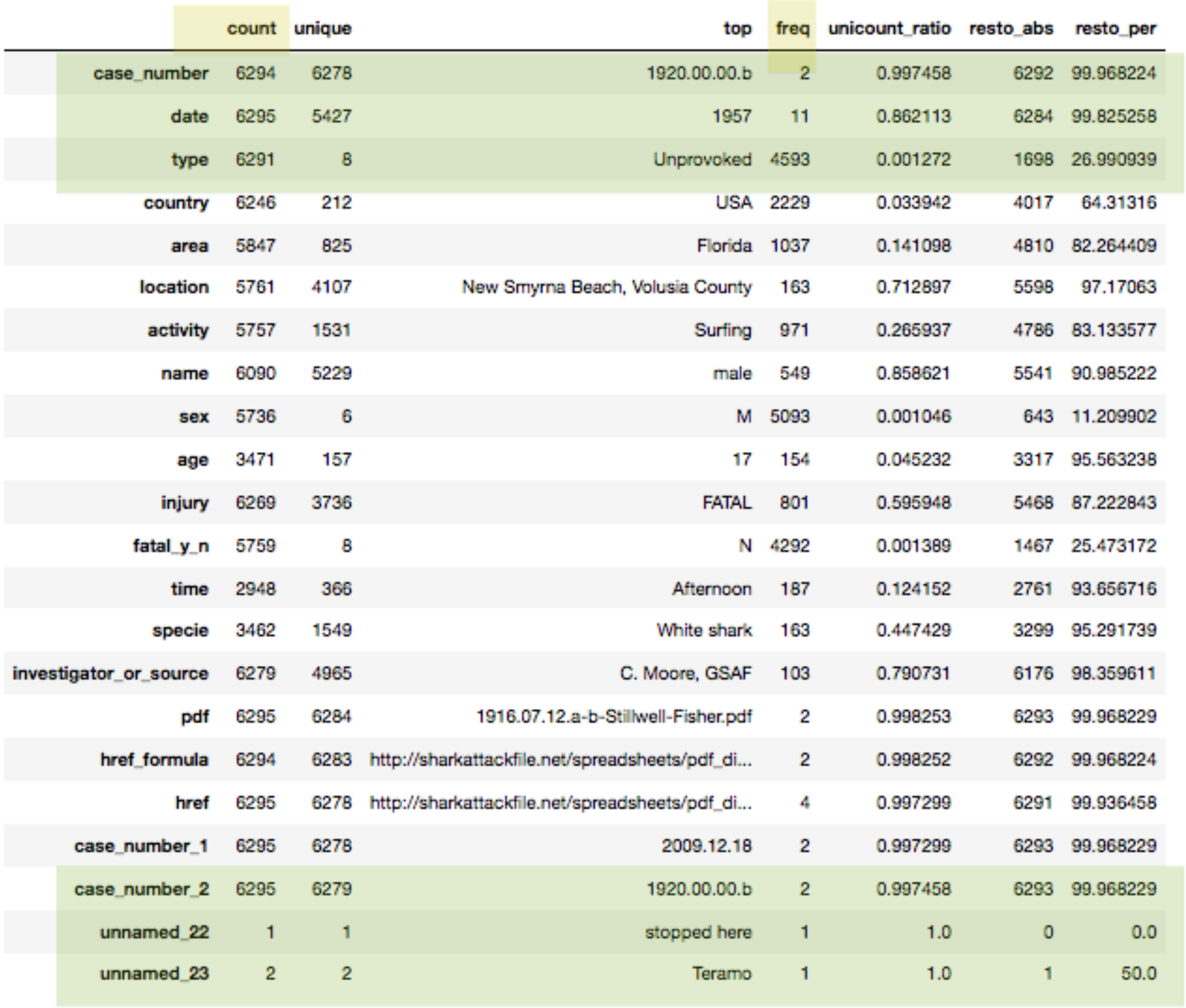
Bottom rows: indicate there is ONE LEVEL with EXCESIVE FREQ:

"unnamed_22" and "unnamed_23" should be deleted in a real job task

"case_number_1" and "case_number_2" should also be deleted cause they seem copies of "case_number"

For the moment, I will also ignore "href" and "href_formula" cause they seems uninformative link

```
df["resto_abs"] = (df["count”] - df["freq"])
df["resto_per"] = (df["resto_abs"]*100) / df["count"]
```



	count	unique	top	freq	unicount_ratio	resto_abs	resto_per
case_number	6294	6278	1920.00.00.b	2	0.997458	6292	99.968224
date	6295	5427	1957	11	0.862113	6284	99.825258
type	6291	8	Unprovoked	4593	0.001272	1698	26.990939
country	6246	212	USA	2229	0.033942	4017	64.31316
area	5847	825	Florida	1037	0.141098	4810	82.264409
location	5761	4107	New Smyrna Beach, Volusia County	163	0.712897	5598	97.17063
activity	5757	1531	Surfing	971	0.265937	4786	83.133577
name	6090	5229	male	549	0.858621	5541	90.985222
sex	5736	6	M	5093	0.001046	643	11.209902
age	3471	157	17	154	0.045232	3317	95.563238
injury	6269	3736	FATAL	801	0.595948	5468	87.222843
fatal_y_n	5759	8	N	4292	0.001389	1467	25.473172
time	2948	366	Afternoon	187	0.124152	2761	93.656716
specie	3462	1549	White shark	163	0.447429	3299	95.291739
investigator_or_source	6279	4965	C. Moore, GSAF	103	0.790731	6176	98.359611
pdf	6295	6284	1916.07.12.a-b-Stillwell-Fisher.pdf	2	0.998253	6293	99.968229
href_formula	6294	6283	http://sharkattackfile.net/spreadsheets/pdf_di...	2	0.998252	6292	99.968224
href	6295	6278	http://sharkattackfile.net/spreadsheets/pdf_di...	4	0.997299	6291	99.936458
case_number_1	6295	6278	2009.12.18	2	0.997299	6293	99.968229
case_number_2	6295	6279	1920.00.00.b	2	0.997458	6293	99.968229
unnamed_22	1	1	stopped here	1	1.0	0	0.0
unnamed_23	2	2	Teramo	1	1.0	1	50.0

### Investigate the relationship between:

"case_number", "case_number_1", "case_number_2" and "original_order”

### 4.4 Correct "date" column

* 4.4.1 Remove "Reported”

* 4.4.2 Transform to "uncertain" the cells including the following keywords:
"Before" or "No date", " or ", "A.D”

* 4.4.3 Clean terms as Ca.

* 4.4.4 Drop "uncertain" values

* 4.4.5 Keep it on-hold and continue cleaning other columns

### 4.5 Correct "type" column

* 4.5.1 Unify the following keywords:
Boating == Boat == Boatomg

* 4.5.2 Transform to uncertain's the cells including the following keywords:
Questionable, Invalid

* 4.5.3 Drop uncertain values

* 4.5.4 Keep it on-hold and continue cleaning other columns

### 4.6 Correct "country" column

* 4.6.1 Clean when the name start with spaces

* 4.6.2 Transform to uncertain cells including "/", "?"

* 4.6.3 Drop records referring to countries mentioned less than 20 times since these could noy be statistically compared against anything

* 4.6.4 Drop "uncertain" and "nan" values

* 4.6.5 Keep it on-hold and continue cleaning other columns

### 4.7 Correct "age" column

* 4.7.1 Drop NAN's

* 4.7.2 Transform to "uncertain" the cells including NON DIGIT"

* 4.7.3 Drop "uncertain" values

Keep it on-hold and continue cleaning other columns

### 4.8 Correct "fatal_y_n" column

* 4.8.1 Transform to "UNKNOWN" the cells = " N", "M" and "2017"

* 4.8.2 Drop "UNKNOWN" values

Keep it on-hold and continue cleaning other columns

AT THIS POINT,
I HAVE A RELATIVELY
CLEAN DATAFRAME
WITH:

2,869 rows and 24 columns!!

AT THIS POINT,
I HAVE A RELATIVELY CLEAN DATAFRAME WITH:
2,869 rows and 24 columns!!

FROM NOW ON I HAVE MODIFIED VALUES WITHOUT DROPING CELLS

### 4.9 Clean misspelled SEX column

### 4.10 Clean long ACTIVITY descriptions (i.e., > word by cell)

### 4.11 Clean "time" column

### 4.12 Clean "injury" column to keep only the top 5 types of lessions

### 4.13 Transform redundant columns into constant NA columns

### 4.14 Change NANs to zeroes those columns that I would prefer to cast as numeric

### 4.15 Downcast the dataframe to decrease memory use

### 4.16 Save this file as first task ---> data/sharks_clean1.csv

### 5 Data Analysis

Previous to the analysis itself:

* I will filter only the relevant columns

* Transform sex and fatal_y_n to "binary" (0/1) columns

* To define a simple analysis with enough statistical power, I will focus on columns with few levels with high frequency

* get_dummy variables from categorical columns

Exemplary analyses:

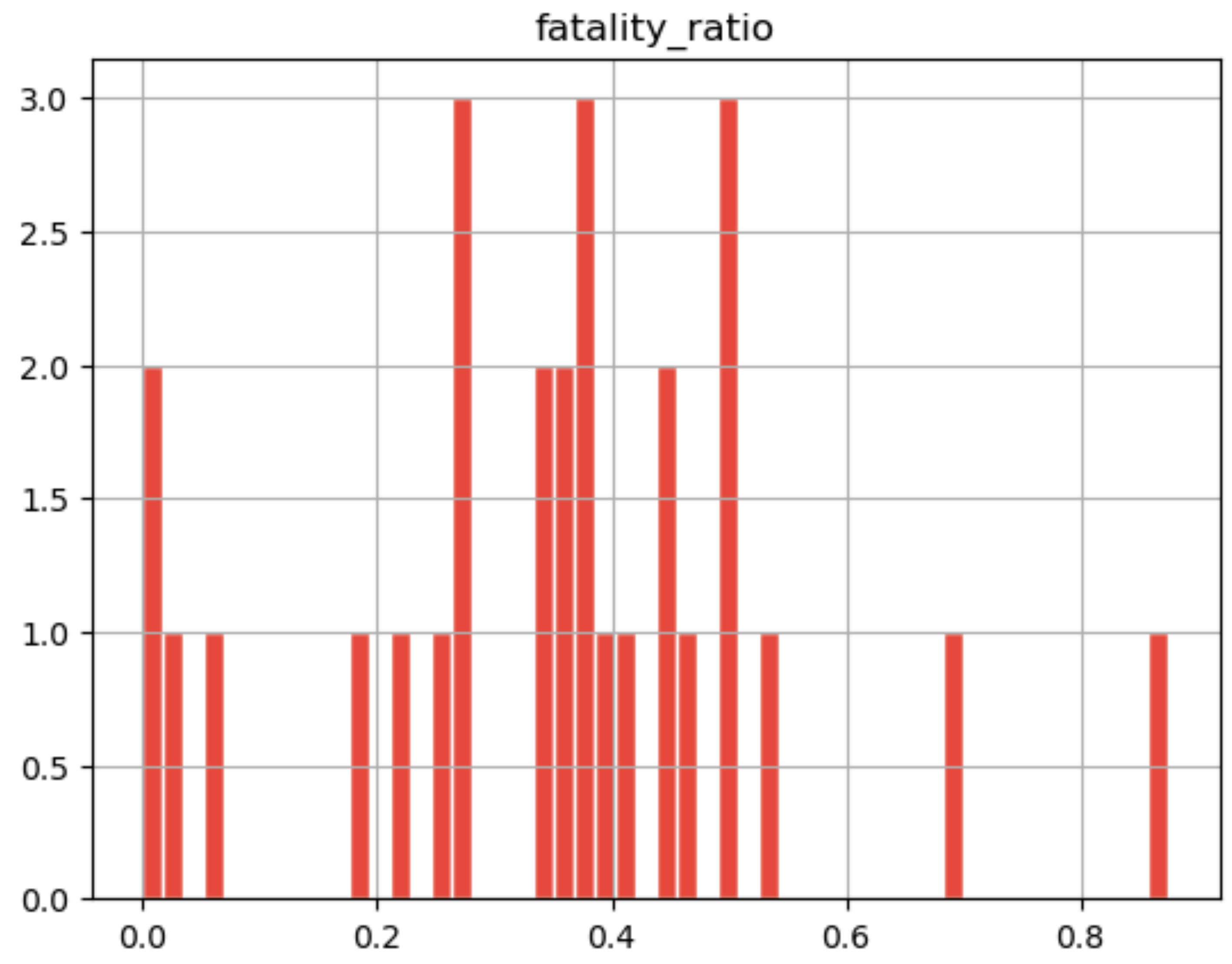
- a) Top 20 deadliest countries
- b1) Correlation between quantitative and/or binary variables
- b2) Correlation between dummy variables

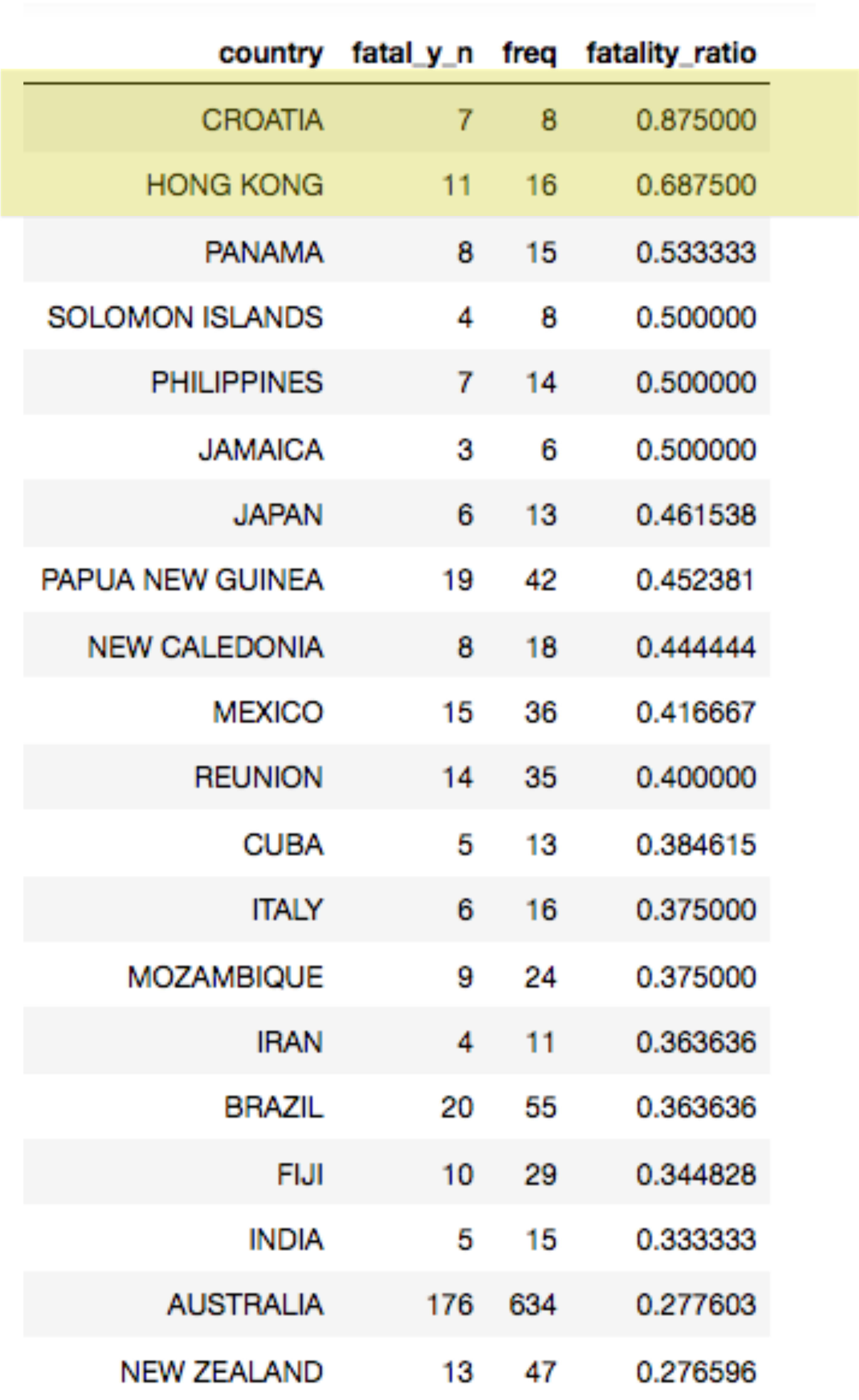
Top 20 deadliest countries

Countries with highest fatality ratio

CAUTION! The fatality ratio might excessively high in countries with low sample size

However, it serves as example



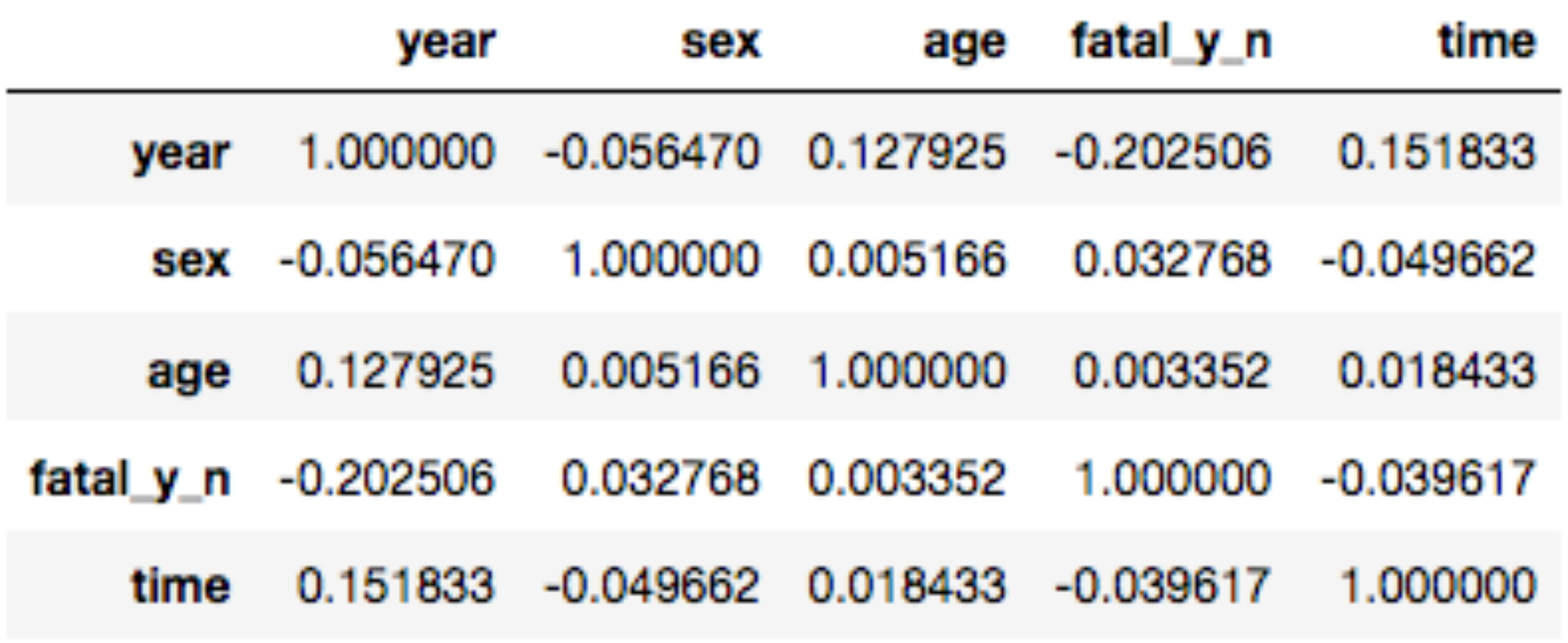


country	fatal_y_n	freq	fatality_ratio
CROATIA	7	8	0.875000
HONG KONG	11	16	0.687500
PANAMA	8	15	0.533333
SOLOMON ISLANDS	4	8	0.500000
PHILIPPINES	7	14	0.500000
JAMAICA	3	6	0.500000
JAPAN	6	13	0.461538
PAPUA NEW GUINEA	19	42	0.452381
NEW CALEDONIA	8	18	0.444444
MEXICO	15	36	0.416667
REUNION	14	35	0.400000
CUBA	5	13	0.384615
ITALY	6	16	0.375000
MOZAMBIQUE	9	24	0.375000
IRAN	4	11	0.363636
BRAZIL	20	55	0.363636
FIJI	10	29	0.344828
INDIA	5	15	0.333333
AUSTRALIA	176	634	0.277603
NEW ZEALAND	13	47	0.276596

Correlation

Between quantitative and/or binary variables

I don't find any relevant correlation between quantitative and/or binary variables.



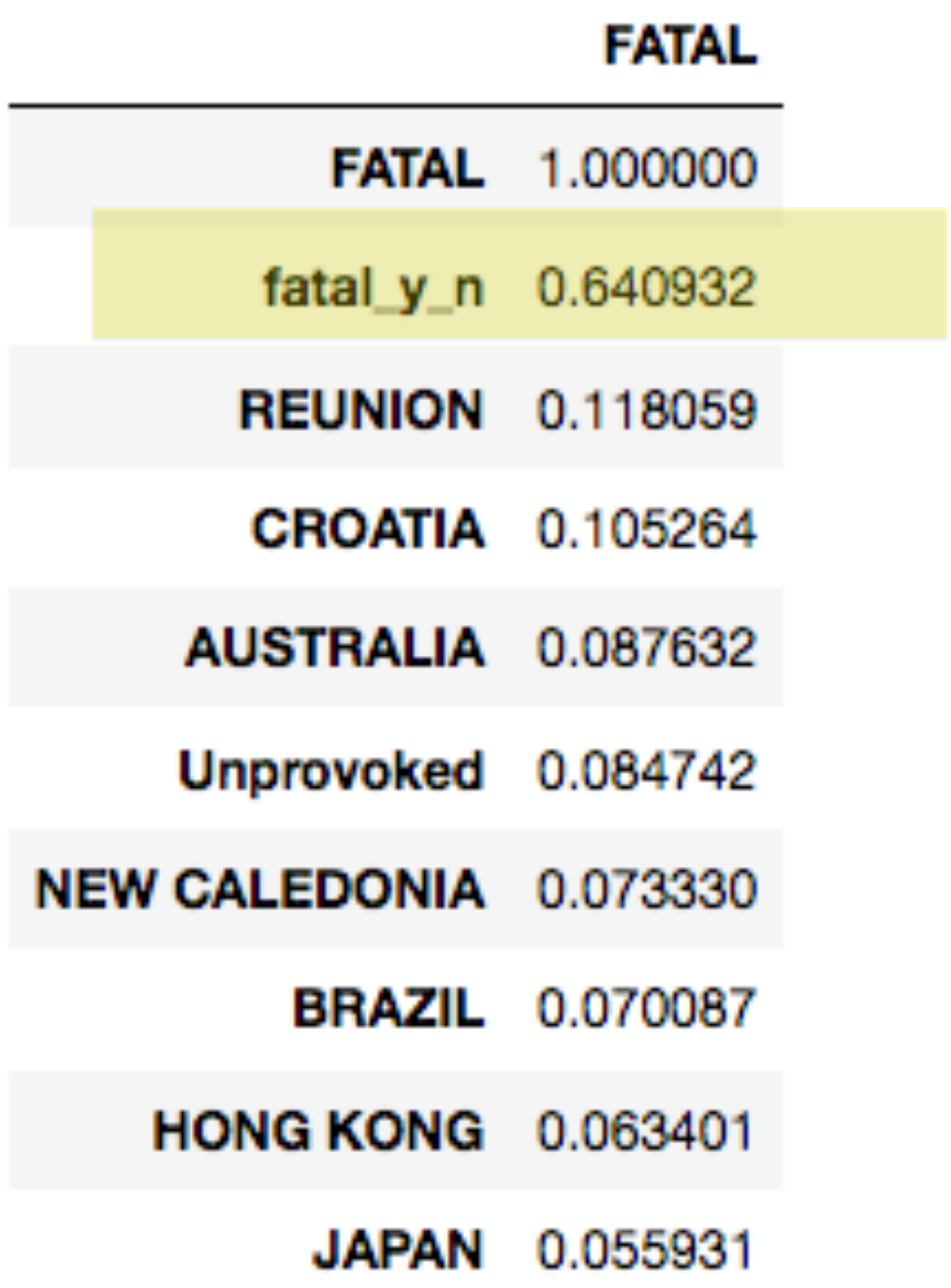
	year	sex	age	fatal_y_n	time
year	1.000000	-0.056470	0.127925	-0.202506	0.151833
sex	-0.056470	1.000000	0.005166	0.032768	-0.049662
age	0.127925	0.005166	1.000000	0.003352	0.018433
fatal_y_n	-0.202506	0.032768	0.003352	1.000000	-0.039617
time	0.151833	-0.049662	0.018433	-0.039617	1.000000

Correlation between dummy variables

These kind of analyses make no sense, we should perform other kind of stat test

In a completely cleaned dataframe, injuries == FATAL should perfectly correlate with fatal_y_n ==Yes.

However, this correlation table indicates we could have further cleaning to make ...



	FATAL
FATAL	1.000000
fatal_y_n	0.640932
REUNION	0.118059
CROATIA	0.105264
AUSTRALIA	0.087632
Unprovoked	0.084742
NEW CALEDONIA	0.073330
BRAZIL	0.070087
HONG KONG	0.063401
JAPAN	0.055931