Data Project: PANDAS
Global Shark Attacks Incidents

María del Carmen Rodríguez García

IRON HACK

1. PROJECT CASE

#DataProjectPandas

- **Data source**: Global Shark Attack Incidents. Data complied by the global shark attacks file – (https://www.kaggle.com/teajay/global-shark-attacks/version/1)

- **Fields:** The data table compiled the following fields: *Case Number, Date, Year, Type, Country, Area, Location, Activity, Name, Sex , Age, Injury, Fatal (Y/N), Time, Species & Investigator or Source.*

- **Initial hypothesis:** The aim will be to evidence that the countries with a greater surfing culture (e.g. USA or Australia), registered the highest numbers of shark attacks, especially during the summer months (northern and southern hemispheres) when the surfing practice is higher.

IRON
HACK

2. DATA TREATMENT

2.1 DATA TREATMENT

o 1st Step: Deleting irrelevant fields for the study, such as: Case number, location, area, name, investigator or source, pdf ...

```python
drop_cols = list(null_cols[null_cols > 5989].index)
data = data.drop(drop_cols, axis=1)
```

```python
drop_col = ["Name","Year","Type","original order","Case Number","Sex ","Age","Injury","Time","Location","Species ","pdf",
df = data.drop(drop_col,axis=1)
df
```

IRON
HACK

o   2nd Step: Cleaning the columns/fields involved in the study, mainly: Date, Country, Activity

```python
h=[]
for e in data["Country"]:
    if re.match("UNITED ARAB EMIRATES \(UAE\)",e):
        j = re.sub("UNITED ARAB EMIRATES \(UAE\)","UNITED ARAB EMIRATES",e)
        h.append(j)
    elif re.match("ST. MAARTIN",e):
        s = re.sub("ST. MAARTIN","ST. MARTIN",e)
        h.append(s)
    elif re.search("\?",e):
        g = re.sub("\?","",e)
        h.append(g)
    elif re.search("Fiji",e):
        g = re.sub("Fiji","FIJI",e)
        h.append(g)
    else:
        a = e.strip("?")
        a = e.strip()
        h.append(a)
data["Country"]=h
```

IRON
HACK

o   2nd Step: Cleaning the columns/fields involved in the study, mainly: Date, Country, Activity

```python
dataActClean=[]
for e in data["Activity"]:
    if re.search("Surf|surf",e):
        a = e.strip()
        s = a.replace(a,"Surfing")
        dataActClean.append(s)
    elif re.search("Disaster|Aircraft|aircraft|disaster|tsunami|storm|Wreck|wreck|Accident|accident|explo|Explo|Adri
        a = e.strip()
        s = a.replace(a,"Accidents or natural disasters")
        dataActClean.append(s)
    elif re.search("Playing|Canoe|canoe|kite|Kite|race|Race|Racing|racing|Competing|kayak|Kayak|Scuba|dived|Dived|di
        a = e.strip()
        s = a.replace(a,"Other Water Sports")
        dataActClean.append(s)
    elif re.search("Swimming|swimming",e):
        a = e.strip()
        s = a.replace(a,"Swimming")
        dataActClean.append(s)
    elif re.search("Fishing|fishing|Lobstering|Crabbing|crabbing|Killing|Harpoon|harpoon|Catching|Collecting|Hunting
        a = e.strip()
        s = a.replace(a,"Fishing&Hunting")
        dataActClean.append(s)
    elif re.search("Walking|walking",e):
        a = e.strip()
        s = a.replace(a,"Walking")
        dataActClean.append(s)
    elif re.search("Bathing|bathing|Floating|Standing|standing|Bather|Splash|splash|jump|Jump|Treading|Beach|beach|S
        a = e.strip()
        s = a.replace(a,"Bathing")
        dataActClean.append(s)
    elif re.search("Attempting|attempting|Escaping|horse|religious",e):
        a = e.strip()
        s = a.replace(a,"Stranger things")
        dataActClean.append(s)
    elif re.search("Shark|shark",e):
        a = e.strip()
        s = a.replace(a,"Shark related activities")
        dataActClean.append(s)
    elif re.search("Boat|boat|Boating|boating|Sailing|sailing|Cruising|raft",e):
        a = e.strip()
        s = a.replace(a,"Boating")
        dataActClean.append(s)
    elif re.search("Murder|Disappeared|disappeared|found|Unknown|male|nan|miss|details|\.",e):
        a = e.strip()
        s = a.replace(a,"UnKnown")
        dataActClean.append(s)
    else:
        a = e.strip()
        s = a.replace(a,"Random activities")
        dataActClean.append(s)
```

IRON HACK

o 2nd Step: Cleaning the columns/fields involved in the study, mainly: Date, Country, Activity

```python
data["Date"].value_counts()
meses = []
for e in data["Date"]:
    if re.search("\d{2}-[JFMAMSOND][a-z]+",e):
        a = e.strip()
        s = re.sub("\d{2}-","",a)
        s = re.sub("-\d{2}","",s)
        s = re.sub("Reported ","",s)
        s = re.sub("\d{1,2}","",s)
        s = re.sub("\.|\-|\"","",s)
        s = re.sub("Before|Between|Anniversary Day|or|&|Letter dated|Repted","",s)
        s = re.sub("Ap$|April","Apr",s)
        s = re.sub("Decp","Dec",s)
        s = re.sub("Feb\s\w+","Feb",s)
        s = re.sub("Mar.+","Mar",s)
        s = re.sub("Aug.+","Aug",s)
        s = re.sub("^\w.+[J].+|Jul.+","Jul",s)
        s = re.sub("Nov.+","Nov",s)
        s = re.sub("Oct.+","Oct",s)
        s = re.sub("Sep.+","Sep",s)
        s = re.sub("\w.+[Sep]","Sep",s)
        s=s.strip()
        meses.append(s)
    else:
        meses.append(s)
```

o   2nd Step: Cleaning the columns/fields involved in the study, mainly: Date, Country, Activity and deletion of suspicious, inconsistent or unreliable data

```python
data = data.drop(data[data["Year"]<=1950].index)
```

```python
data4 = data4.drop(data4[data["holaaa"]=='Jut'].index)
```

```python
data = data.drop(data[data["Country"]=='nan'].index)
data = data.drop(data[data["Country"]=='Between PORTUGAL & INDIA'].index)
```

IRON
HACK

o   2nd Step: Cleaning the columns/fields involved in the study, mainly: Date, Country, Activity and deletion of suspicious, inconsistent or unreliable data

	Date	Country	Area	Activity	Fatal (Y/N)
0	Sep	USA	Florida	Surfing	N
1	Sep	USA	Florida	Surfing	N
2	Sep	USA	Florida	Surfing	N
3	Sep	AUSTRALIA	Victoria	Surfing	N
4	Sep	AUSTRALIA	Victoria	Surfing	N
...	...	...	...	...	...
4223	Feb	AUSTRALIA	New South Wales	Surfing	N
4224	Feb	AUSTRALIA	New South Wales	Swimming	N
4225	Jan	SOUTH AFRICA	KwaZulu-Natal	Swimming	Y
4226	Jan	AUSTRALIA	New South Wales	UnKnown	Y
4227	Jan	NEW GUINEA	Madang Province	UnKnown	Y

4215 rows × 5 columns

IRON
HACK

o   3rd Step: Grouping of cleaned and accepted values into simple and relevant categories.

```python
y=df.groupby(['Month','Country'])['Activity'].value_counts().to_frame()
y = y.rename(columns={"Activity": "Sharks_attacks"})
y["Sharks_attacks"].sum()
y=y.sort_values(by='Sharks_attacks', ascending=False)
```

```python
z=df.groupby(['Month'])['Activity'].value_counts().to_frame()
z = z.rename(columns={"Activity": "Sharks_attacks"})
z.sort_values(by='Sharks_attacks', ascending=False)
```

```python
b=df.groupby(['Country'])['Activity'].value_counts().to_frame()
b = b.rename(columns={"Activity": "Sharks_attacks"})
b.sort_values(by='Sharks_attacks', ascending=False)
```

IRON
HACK

· DATA PROJECT
◦ 2.1 DATA TREATMENT

o   4th Step: Generation of charts and tables to contrast the initial hypothesis.

Month	Country	Activity	Sharks_attacks
Sep	USA	Surfing	112
Oct	USA	Surfing	101
Aug	USA	Surfing	81
Apr	USA	Surfing	65
Jul	USA	Surfing	65

Month	Activity	Sharks_attacks
Sep	Surfing	138
Oct	Surfing	137
Jan	Fishing&Hunting	134
Jul	Surfing	118
Aug	Surfing	107

Country	Activity	Sharks_attacks
USA	Surfing	674
	Other Water Sports	276
	Swimming	250
AUSTRALIA	Surfing	207
USA	Bathing	200

IRON HACK

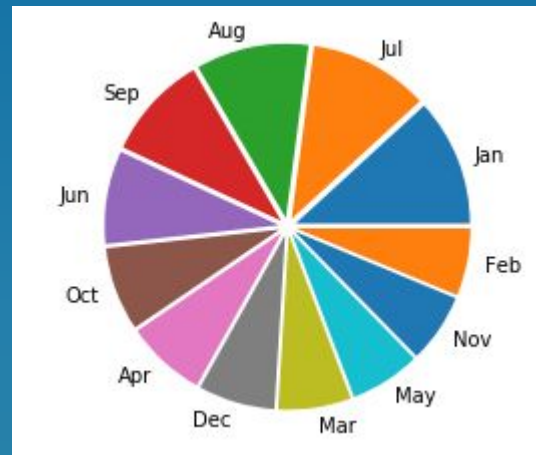o　4th Step: Generation of charts and tables to contrast the initial hypothesis.

Country	Sharks_attacks
USA	1771
AUSTRALIA	731
SOUTH AFRICA	450
PAPUA NEW GUINEA	110
BRAZIL	95

Month	Sharks_attacks
Jan	498
Jul	468
Aug	441
Sep	409
Jun	364
Oct	327

IRON
HACK

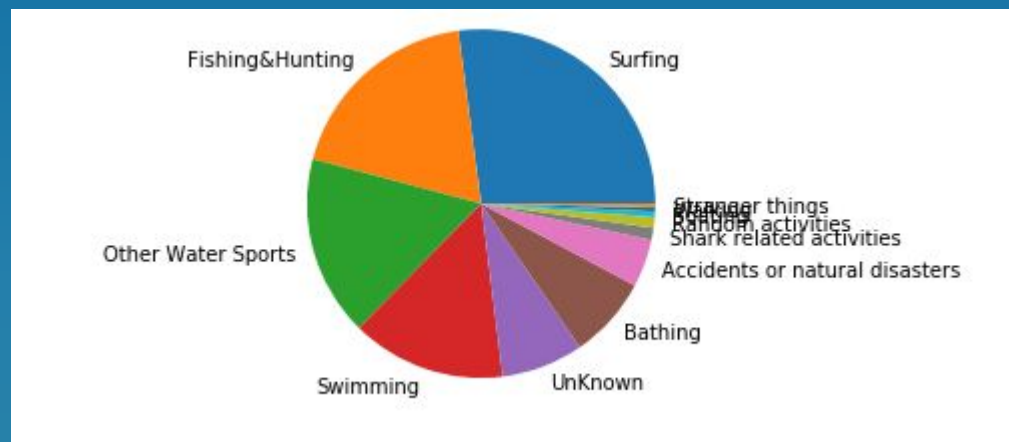○ 4th Step: Generation of charts and tables to contrast the initial hypothesis.

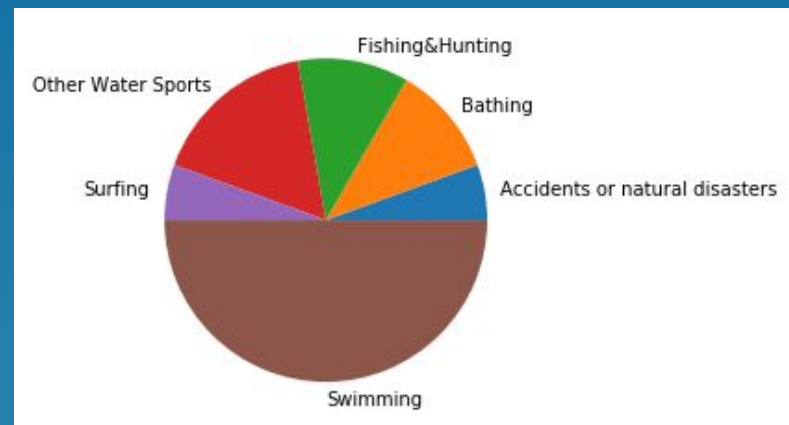| Activity	Sharks_attacks
Surfing	1142
Fishing&Hunting	792
Other Water Sports	706
Swimming	604
UnKnown	322
Bathing	321
Accidents or natural disasters	187

IRON HACK

3. DATA ANALYSIS

#DataProjectPandas

○   Curiosity:

- Data of fatality of the attack in Spain depending on the activity

Fatal (Y/N)	Activity	Sharks_attacks_Spain
N	Swimming	9
N	Other Water Sports	3
N	Bathing	2
N	Fishing&Hunting	2
N	Accidents or natural disasters	1
N	Surfing	1

Activity	Sharks_attacks_Spain
Accidents or natural disasters	1
Bathing	2
Fishing&Hunting	2
Other Water Sports	3
Surfing	1
Swimming	9

IRON
HACK

○   Others curiosities:

Year	Sharks_Year
2016	103
2015	139
2014	124
2013	122
2012	117

Fatal (Y/N)		Sharks_attacks
	Activity	
N	Surfing	1065
	Fishing&Hunting	684
	Other Water Sports	562
	Swimming	413
	Bathing	282
	UnKnown	207
Y	Swimming	188
	Other Water Sports	138
	Accidents or natural disasters	104
	Fishing&Hunting	99
	UnKnown	92
N	Accidents or natural disasters	83
Y	Surfing	68

Fatal (Y/N)		Sharks_attacks_country
	Country	
N	USA	1626
	AUSTRALIA	598
	SOUTH AFRICA	369
	BAHAMAS	76
	NEW ZEALAND	67

IRON HACK

IRON
HACK

4. CONCLUSIONS

#DataProjectPandas

○ **In conclusion I can confirm that the country with the most shark attacks since the 50's is the USA but the months with the most attacks are not July and August, as I thought a priori, if not September and October**

IRON
HACK

THE END

THANK YOU FOR YOUR
ATTENTION