Jet Ski Inc. Shark Attack Vulnerability

Our company, Jet Ski Inc., is looking for a new home base for our Headquarters that needs to be near the water but we want to ensure our employees are safe from shark attacks. We will be assessing locations around the world, what activities may have prompted the attacks and what the gender was of the individual. These variables will help us determine both our HQ location and also which employees we will select to send out into the waters.Our team has spent the past few days gathering data on the types of shark attacks, population growth of humans, and locations. We had lots of success with the data at hand, and we are excited to share our findings.

We will be using Kaggle DB for both sets:

- **World Population Dataset**
- **Global Shark Attacks**

Our data will include the following information as columns:

Color Key: WPDS, GSA,
Joining column

- Growth Rate
- 2022 Population
- 2020 Population
- 2015 Population
- 2010 Population
- 2000 Population
- 1990 Population
- 1980 Population
- 1970 Population
- Growth Rate
- World Population Percentage
- Country
- Number of attacks
- Type of shark attack
- Country/Area/Location of the shark attack
- Activity happening during the shark attack
- Fatal
- Sex

Data Import and update

Extraction

Each data set was downloaded as a CSV file and then imported into Visual Studio Code using Pandas.

Transform/Selection

We used specific columns which had the data we were most interested in retaining in the code and dropped all the other columns. During this selection process, we determined that Years from the Shark DB was a float64 and needed to be converted into an object in order to be able to clean the data.

```
Shark_df = pd.read_csv(Shark_path, encoding="ISO-8859-1")
Shark_df.head()
```

	Case Number	Date	Year	Type	Country	Area	Location	Activity	Name	Sex	...	Species	Investigator or Source	pdf	
0	2018.06.25	25-Jun-2018	2018.0	Boating	USA	California	Oceanside, San Diego County	Paddling	Julie Wolfe	F	...	White shark	R. Collier, GSAF	2018.06.25-Wolfe.pdf	http://sharkatta
1	2018.06.18	18-Jun-2018	2018.0	Unprovoked	USA	Georgia	St. Simon Island, Glynn County	Standing	Adyson McNeely	F	...	NaN	K.McMurray, TrackingSharks.com	2018.06.18-McNeely.pdf	http://sharkatta
2	2018.06.09	09-Jun-2018	2018.0	Invalid	USA	Hawaii	Habush, Oahu	Surfing	John Denges	M	...	NaN	K.McMurray, TrackingSharks.com	2018.06.09-Denges.pdf	http://sharkatta
3	2018.06.08	08-Jun-2018	2018.0	Unprovoked	AUSTRALIA	New South Wales	Arrawarra Headland	Surfing	male	M	...	2 m shark	B. Myatt, GSAF	2018.06.08-Arrawarra.pdf	http://sharkatta
4	2018.06.04	04-Jun-2018	2018.0	Provoked	MEXICO	Colima	La Ticla	Free diving	Gustavo Ramos	M	...	Tiger shark, 3m	A .Kipper	2018.06.04-Ramos.pdf	http://sharkatta

5 rows × 24 columns

In this portion columns were dropped that weren't useful to the data we were looking for and began looking for other areas of the data where we needed to clean the data.

```
# Drop the columns of the data that will not be used
Shark_df = Shark_df.drop(columns={"Case Number","Date","Type","Area","Location","Name","Age","Injury","Time","Investigator or Source","pd
Shark_df.head()
```

	Year	Country	Activity	Sex	Fatal (Y/N)	Species
0	2018.0	USA	Paddling	F	N	White shark
1	2018.0	USA	Standing	F	N	NaN
2	2018.0	USA	Surfing	M	N	NaN
3	2018.0	AUSTRALIA	Surfing	M	N	2 m shark
4	2018.0	MEXICO	Free diving	M	N	Tiger shark, 3m

We made changes by filling the Nan with zeros in order to have information more readily available.

```
cleanshark_df = Shark_df.fillna (0)
cleanshark_df.tail()
```

	Year	Country	Activity	Sex	Fatal (Y/N)	Species
25718	0.0	0	0	0	0	0
25719	0.0	0	0	0	0	0
25720	0.0	0	0	0	0	0
25721	0.0	0	0	0	0	0
25722	0.0	0	0	0	0	0

```
cleanshark_df ['Year'] = cleanshark_df['Year'].fillna(0).astype(int)
cleanshark_df.head()
```

	Year	Country	Activity	Sex	Fatal (Y/N)	Species
0	2018	USA	Paddling	F	N	White shark
1	2018	USA	Standing	F	N	0
2	2018	USA	Surfing	M	N	0
3	2018	AUSTRALIA	Surfing	M	N	2 m shark
4	2018	MEXICO	Free diving	M	N	Tiger shark, 3m

For the last piece of data above, we changed the year to an integer in order to not have an issue with the original type of data which was a float.

Joining

We used PgAdmin to create our tables and join them. We had to arrange each of the columns in the table to run each of the table set ups. We had to rename one of the columns (Fatal Y/N) so it matched with the columns and we are able to import properly. The repo had to be updated with the new CSV, after the column name change, and with the new code. We then used the updates to import to our tables and allow them to run before we could join them together.

Issues that plagued us during the join included had to change Species, we took out the /CCA3 on Country and included VARCHAR for Growth Rate but determined it should be changed to INT. We took out Territory on the World Population CSV file and found out that while working on the Population table in SQL you need the correct Quotation marks, so if copying and pasting; erase and type in on the PGADmin4 application window.

The code written only brought up the USA when attempting to merge which we determined to be due to an issue with capitalization. We attempted to remove keys and drop table without a solution. We also corrected spelling errors and capitalization which contributed to bugs within the code. The original world population CSV and the Updated Population CSV were moved into the main repo branch - resources file with the capitalization fixed.

Queries and Loading

In order for us to determine who had the least amount of shark attacks and also the highest percent of population we queried as follows

```
select SharkAttacks.Country,count(SharkAttacks.Country), SharkAttacks
from SharkAttacks
join WorldPopulation
on SharkAttacks.country=WorldPopulation.country
group by SharkAttacks.Country,SharkAttacks.sex, SharkAttacks.activity
order by "World Population Percentage" desc;

select SharkAttacks.Country,count(SharkAttacks.Country),WorldPopulat
from SharkAttacks
join WorldPopulation
on SharkAttacks.country=WorldPopulation.country
group by SharkAttacks.Country,WorldPopulation."World Population Perc
order by "World Population Percentage" desc;

select SharkAttacks.Country,count(SharkAttacks.Country), WorldPopula
from SharkAttacks
join WorldPopulation
```

ata Output Messages Notifications

country character varying	count bigint	World Population Percentage numeric
CHINA	2	17.88
INDIA	1	17.77
USA	1599	4.24
INDONESIA	8	3.45
NIGERIA	1	2.74
BRAZIL	103	2.7
BANGLADESH	1	2.15
RUSSIA	4	1.81
MEXICO	40	1.6
JAPAN	20	1.55
PHILIPPINES	17	1.45
EGYPT	25	1.39
VIETNAM	10	1.23
IRAN	3	1.11
TURKEY	1	1.07
THAILAND	7	0.9

Within this query, we determined that China and India had the largest amount of population percentage, while also having the lowest amount of shark attacks.

	country character varying	count bigint	World Population Percentage numeric
1	USA	1599	4.24
2	AUSTRALIA	570	0.33
3	SOUTH AFRICA	363	0.75
4	BRAZIL	103	2.7
5	MEXICO	40	1.6
6	EGYPT	25	1.39
7	MOZAMBIQUE	24	0.41
8	ITALY	21	0.74
9	SPAIN	21	0.6
10	JAPAN	20	1.55
11	PHILIPPINES	17	1.45

In the query above, we saw which countries had the highest count of shark attacks and the population percent. However even if the percentage was high, we did not want a country with high shark attack counts.

Lastly, we wanted to determine the sex of the person attacked and the activity they were involved in during the attack. The following variables help us understand males were solely involved in the shark attacks in the countries we were interested in headquartering. As well, the activity involved, although informative, wasn't significant enough to determine we should not have jet ski activity in the country.

	country character varying	count bigint	sex character varying	activity character varying	World Population Percentage numeric
1	CHINA	1	M	Scuba diving in aquarium tank	17.88
2	CHINA	1	M	Swimming	17.88
3	INDIA	1	M	Swimming or surfing	17.77
4	USA	9	0	0	4.24
5	USA	1	0	Boat	4.24
6	USA	1	0	Body boarding	4.24
7	USA	1	0	Bottom fishing for lingcod & had hooked a fish	4.24
8	USA	1	0	Cruising	4.24
9	USA	3	0	Diving	4.24
10	USA	1	0	Dropping anchor	4.24
11	USA	3	0	Fishing	4.24

```
select country, count(*)
from SharkAttacks
group by country
order by "count" asc;

select country, "World Population Percentage"
FROM WorldPopulation
order by "World Population Percentage" desc;

select SharkAttacks.Country,count(SharkAttacks.Country), SharkAttacks.sex, SharkAttacks.activity, WorldPopulation."World Popula
from SharkAttacks
join WorldPopulation
on SharkAttacks.country=WorldPopulation.country
group by SharkAttacks.Country,SharkAttacks.sex, SharkAttacks.activity,WorldPopulation."World Population Percentage"
order by "World Population Percentage" desc;

select sex,count(sex),country
from SharkAttacks
group by country
order by "count" asc;
```

In Summary:

Our company Jet Ski inc. has selected our team to form five ideal locations for our new company set up before we send another team for location recon. The top two countries we have picked based off of our findings are: China and India. We picked these two because of the population, the number of shark attacks, and the activities in which these attacks had occurred. We also concluded the worst two locations are: USA, and Australia. Based on how high of the numbers there are in shark attacks. The reason we wanted to include these two, is so not only can we avoid them but hopefully make predictions on if it will ever increase or decrease.
We set a threshold to not go under the two percent in World Population Percentage on these attacks so we can try to safely assemble our teams for recon and hopefully find a new place to call home for our company. We value our employees and the natural habitat of the sharks, which is why we are looking for the balance between the two.

Future improvements:

During the process of selecting our databases we recognized data was helpful to the questions we were asking, however, there were still issues with being as specific as possible. For example, we were not able to query for coastal cities but had to leave the queries open to countries. For the next phase in the project we would like to build a table to query for weather results so we are actually placing our HQ in a location that would suit our needs for testing our jet skis.