SHARK

DIVING EXPERIENCE

Artur Cabral
Erik Ribeiro
J. Guilherme Nascimento
Victor Cagnin

ABOUT THE PROJECT

This research was created for "SHARKS - THE DIVING EXPERIENCE," a new company seeking the ideal location to launch its new project in the USA. The main goal is to analyze the states of California and Florida and determine where sharks are most frequently present, considering the safety level based on related incidents.

It was developed using a dataset containing an Incident Log that tracks shark-related incidents.

The original log aims to assist in shark behavior research, distinguishing provoked incidents for species and bite pattern analysis. It is regularly updated, with an invitation for researchers to join GSAF for additional data access. All individuals survived unless noted otherwise.

DATA ANALYSES

THE STRUCTURE WAS BASED ON A THE FOLLOWINGS ARGUMENTS:

LOCATION	Sorted by US/STATE/COUNTY
INCIDENT	Sorted by type of attack, highlighting fatality
SPECIES	Used to see the most common species in each location
TIME	Analyses of the last forty year, to define its life cycle.

The main techniques were used for:

- Removing Unnecessary Columns
- Cleaning and Standardizing 'State' Column:
- Filtering Rows Based on Values
- Handling Missing Values
- Handling and Analyzing Unique Values

WRANGLING & CLEANING

RAW DATA

CLEANSE

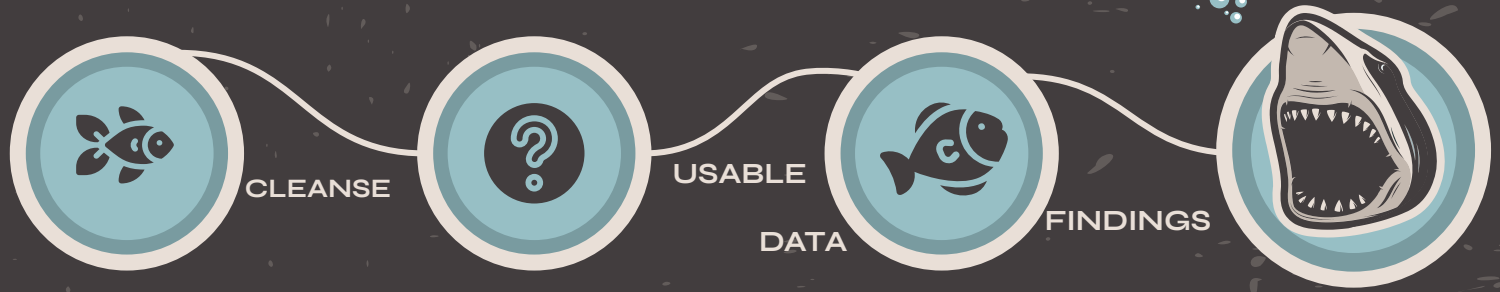
EVALUATE USABILITY

USABLE DATA

ANALIZE

FINDINGS

VIZUALIZE



A series of data cleaning and wrangling operations were applied in order to prepare the data for further visualization.

Each step addresses specific issues or requirements in the dataset, enhancing its quality and suitability for analysis.

OVERCOMINGS

standardize_column_names	Function defined to standardize column by removing spaces, converting to lowercase, and replacing spaces with underscores. df = standardize_column_names(df).
columns_to_remove	The columns specified in the list are dropped from the DataFrame using df.drop(columns=columns_to_remove, axis=1, inplace=True)
Filtering Rows in a column	Rows are filtered based on the 'country' column to include only those where the country is 'USA' using df = df[df['country'] == 'usa']
Handling Missing Values	Missing values in the 'year' column are handled by converting it to numeric format (pd.to_numeric(df['year'], errors='coerce')) and filtering for years greater than or equal to 1980.
standardize_column_names	Function defined to standardize column by removing spaces, converting to lowercase, and replacing spaces with underscores. df = standardize_column_names(df).
Filtering Rows Based on Patterns	Rows are filtered based on the 'species' column using patterns such as 'white shark' using df[df['species'].str.contains(pattern, case=False, na=False)].

EXPLORATORY DATA ANALYSIS

DATA DISTRIBUTION

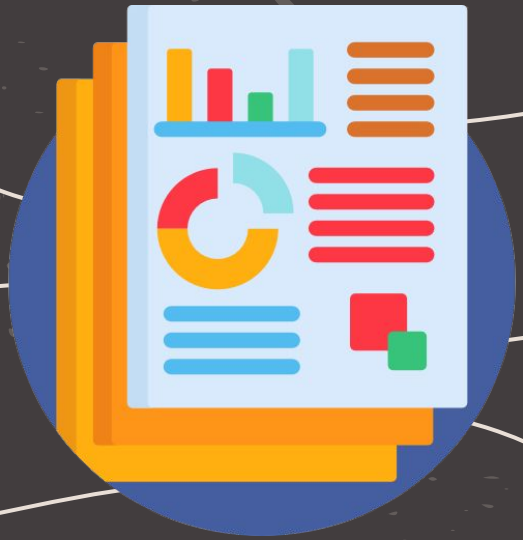
Understand the shape, central tendencies, and outliers in variable distributions.

CORRELATION ANALYSIS

Identify relationships between variables through correlation analysis.

DATA PATTERNS & TRENDS

Visualize trends and patterns, uncovering underlying dynamics.



OUTLIER DETECTION

Detect and analyze outliers, which may indicate errors or significant data points.

DATA RELATIONSHIPS

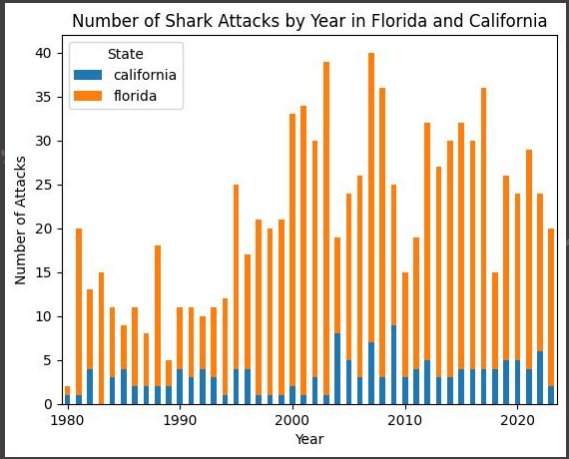
Examine how variables interact or influence each other.

DATA QUALITY ASSESSMENT

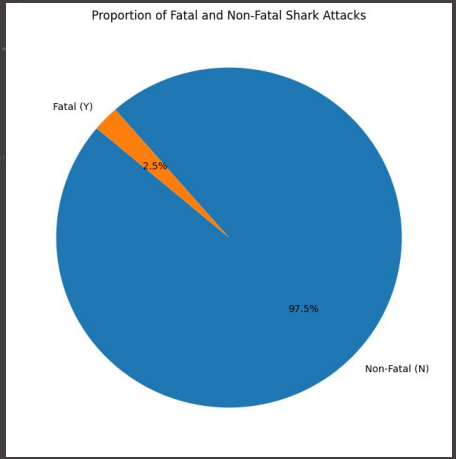
Identify missing values and assess data completeness and quality.

The best way for understanding data structure, guides decision-making, and uncovers hidden insights, forming the basis for advanced analyses.

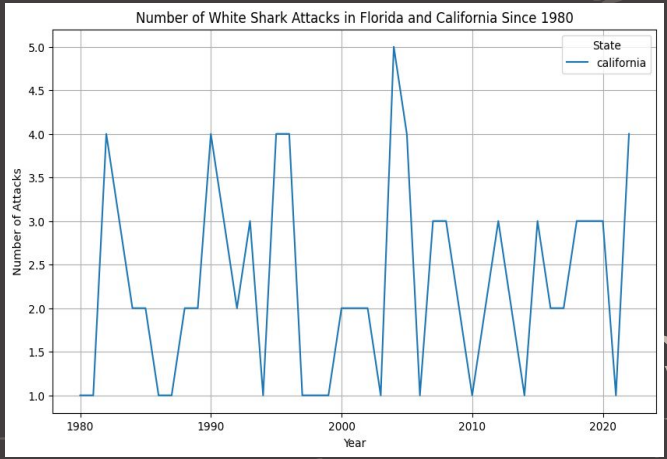
Visualization Techniques: Use visualizations for a more intuitive grasp of patterns and relationships.
Hypothesis Generation: EDA sparks hypotheses, guiding subsequent statistical tests or modeling efforts.



BAR



PIE



LINE

MAJOR OBSTACLES

MISSING DATA

- *Obstacle:* Incomplete data.
- *Solution:* Impute missing values statistically or remove affected rows/columns.

DATA SCALING AND TRANSFORMATION

- *Obstacle:* Variable scales.
- *Solution:* Normalize or standardize data.

OUTLIERS

- *Obstacle:* Distort patterns.
- *Solution:* Identify and handle outliers using data transformation or robust statistical methods.

COMPUTATIONAL CHALLENGES

- *Obstacle:* Large datasets.
- *Solution:* Use sampling, distributed computing, or code optimization.

DATA QUALITY ISSUES

- *Obstacle:* Inaccuracies.
- *Solution:* Clean data, validate against standards, and use profiling techniques.

Tackling these issues demands a blend of technical know-how, domain expertise, and ongoing validation, ensuring more accurate EDA outcomes.

FINDINGS

Pros & Cons

PROS

Potential Demand: Coastal states like Florida and California show promise for a shark-diving business, attracting enthusiasts.

Tourism Hubs: Established tourism in certain states offers a ready audience for aquatic experiences.

Thrill Factor: Areas with more sharks, such as Florida, can provide authentic and thrilling experiences.

CONS

Regulatory Challenges: Some states have stringent regulations; understanding and compliance are crucial.

Safety Concerns: States with higher shark populations pose safety challenges; robust safety measures are essential.

Seasonal Variability: Business viability may be influenced by seasonal factors.

MISSING VALUES

DATA QUALITY

TRANSFORMATION

THANKS!

Artur Cabral

Erik Ribeiro

J Guilherme Nascimento

Victor Cagnin

SHARK

DIVING EXPERIENCE