DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

BEYOND THE BOUNDARY:
EXPLORING THE DEPTHS: EXPLORATORY DATA ANALYSIS OF SHARK ATTACKS USING PYTHON

Submitted by

Jagadeesh Chinta

Registration No. 12311668

Programme and Section. K23GR

Course Code: INT375

Under the Guidance of

Gargi Sharma

UID. 29439

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Jagadeesh Chinta, bearing Registration No. [Your Registration Number], has completed the INT375 project titled, "Exploratory Data Analysis of Shark Attacks Using Python" under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort, and study.

Gargi Sharma

Professor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 10-04-25

DECLARATION

I, Jagadeesh Chinta, student of Computer Science and Engineering under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 10-04-25

Registration No. 12311668

Signature:

Jagadeesh Chinta

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Gargi Sharma for her valuable guidance and support throughout the course of this project. Her insights and encouragement were instrumental in shaping the analysis and ensuring its successful completion.

TABLE OF CONTENT

INTRODUCTION:

Shark attacks, though rare, capture significant public and scientific interest due to their impact on human safety and marine ecosystems. This project analyzes a comprehensive dataset of shark attack incidents to uncover patterns, trends, and statistical insights using Python. Leveraging libraries such as Pandas, Matplotlib, Seaborn, NumPy, and SciPy, the analysis explores geographic distributions, fatality rates, victim demographics, and temporal trends. The dataset is cleaned to ensure reliability, followed by exploratory data analysis (EDA) to understand its structure and content.

The study addresses key questions, including which countries report the most attacks, the proportion of fatal incidents, and whether victim age correlates with outcomes. Visualizations like bar charts, histograms, scatter plots, and heatmaps make the findings intuitive, while statistical tests, including T-tests and normality checks, provide rigor. The analysis aims to inform public awareness, safety policies, and marine conservation efforts by shedding light on the dynamics of shark-human interactions.

SOURCE:

The dataset was sourced from an Excel file (attacks.xlsx) likely derived from the Global Shark Attack File or a similar repository. It contains detailed records of shark attack incidents with the following attributes:

- **Date: Date of the attack**
- **Year: Year of the incident**
- **Type: Type of attack (e.g., Unprovoked, Provoked, Invalid)**
- **Country: Country where the attack occurred**
- **Area: Region within the country**
- **Location: Specific location of the incident**
- **Activity: Victim's activity (e.g., Surfing, Swimming)**
- **Name: Victim's name**
- **Sex: Victim's gender (M/F)**
- **Age: Victim's age**
- **Injury: Description of injuries sustained**
- **Fatal (Y/N): Whether the attack was fatal**
- **Time: Time of the attack**
- **Species: Shark species involved (if identified)**
- **Investigator or Source: Source of the report**

LINK:https://mavenanalytics.io/dataplayground?order=date_added%2Cdesc&search=Shark%20Attack

EXPLANATORY DATA ANALYSIS (EDA):

The EDA process was conducted systematically to extract meaningful insights from the shark attack dataset. The steps are outlined below, reflecting the methodology used in the Python code.

1. Data Cleaning

- **Objective**: Ensure data quality by addressing inconsistencies and missing values.
- **Process**:
  - Loaded the dataset from attacks.xlsx using Pandas.
  - Replaced invalid entries (e.g., "?", 0) with NaN to standardize missing data.
  - Converted Age and Year to numeric types, coercing errors to NaN.
  - Created a binary column Fatal_Binary (1 for fatal, 0 for non-fatal) from Fatal (Y/N).
  - Dropped rows missing critical fields (Year, Country, Fatal (Y/N)).
  - Filtered Year to 1900–2023 to exclude outliers.
- **Outcome**: A clean dataset (df_cleaned) with reliable entries for analysis.

🔍 2. Initial Data Exploration

- **Objective**: Understand the dataset's structure and content.
- **Process**:
  - Used df.info() to inspect data types and non-null counts.
  - Calculated missing values with df.isnull().sum().
  - Generated summary statistics with df.describe().
  - Printed column names and meanings for clarity.
- **Outcome**: Gained familiarity with the dataset, identifying key variables like Year, Country, Fatal (Y/N), and Age.

📊 3. Univariate Analysis and (Multivariate analysis)

- Analyzed single variables (e.g., Country, Fatal_Binary) and relationships (e.g., Age vs. Fatal_Binary).
- Visualized distributions and trends using plots (detailed in Analysis).
-

🚙 4. Statistical Analysis

- Applied T-tests, Shapiro-Wilk tests, and Z-score analysis to test hypotheses and detect outliers.

🔢 8. Correlation Analysis

The correlation analysis focused on numerical variables (Age, Fatal_Binary) to identify relationships. A heatmap visualized the correlation matrix, revealing weak correlations (e.g., near 0 between Age and Fatal_Binary), indicating no strong linear relationship between victim age and fatality. This suggests other factors (e.g., activity, location) may influence outcomes more significantly.

📊 9. Statistical Analysis

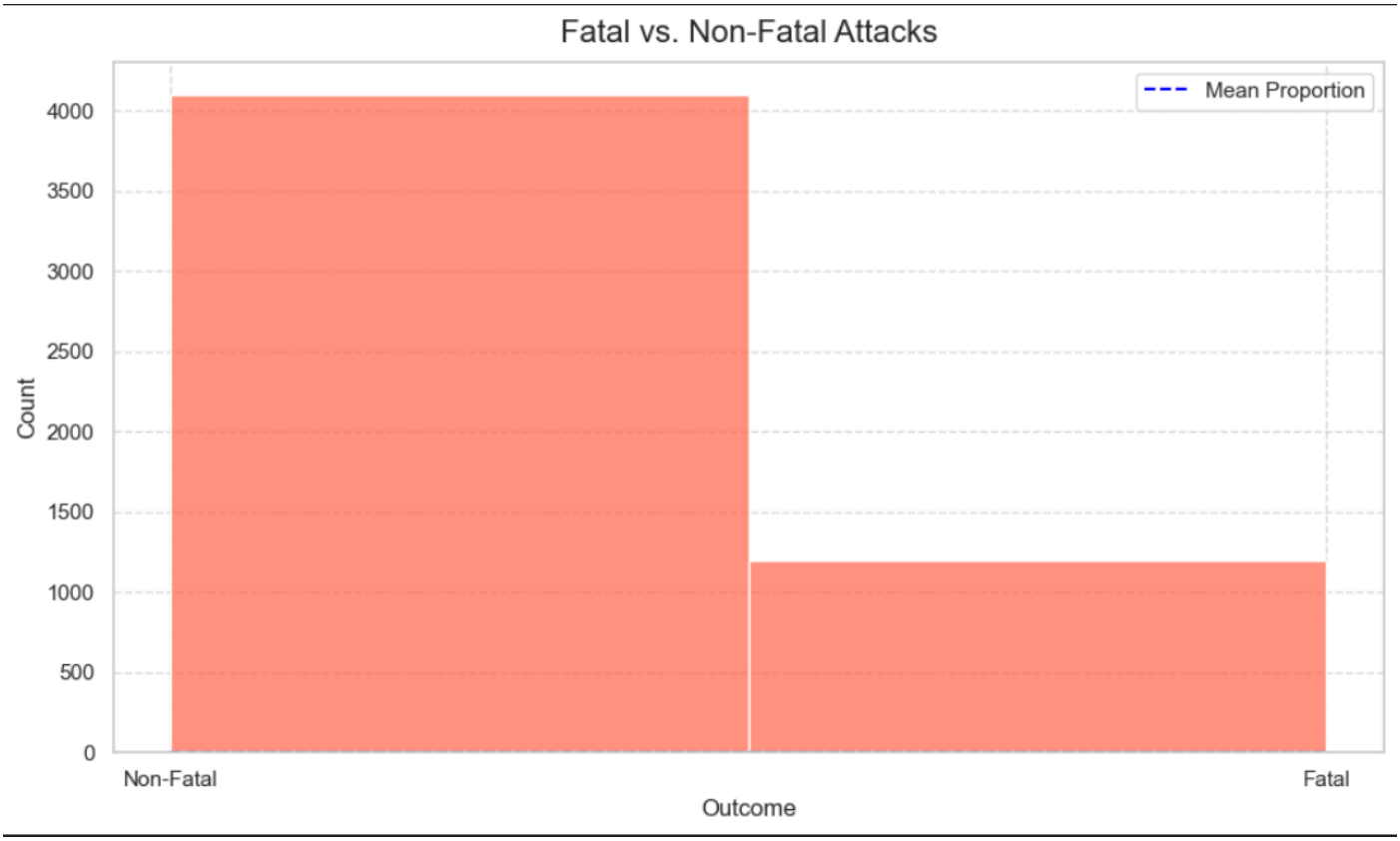
Finally, several statistical tests are applied:

- A t-test compares the average model years of BEVs and PHEVs to see if one type tends to be newer.
- A z-score analysis identifies manufacturers whose EV registration numbers are significantly higher than others (i.e., statistical outliers).

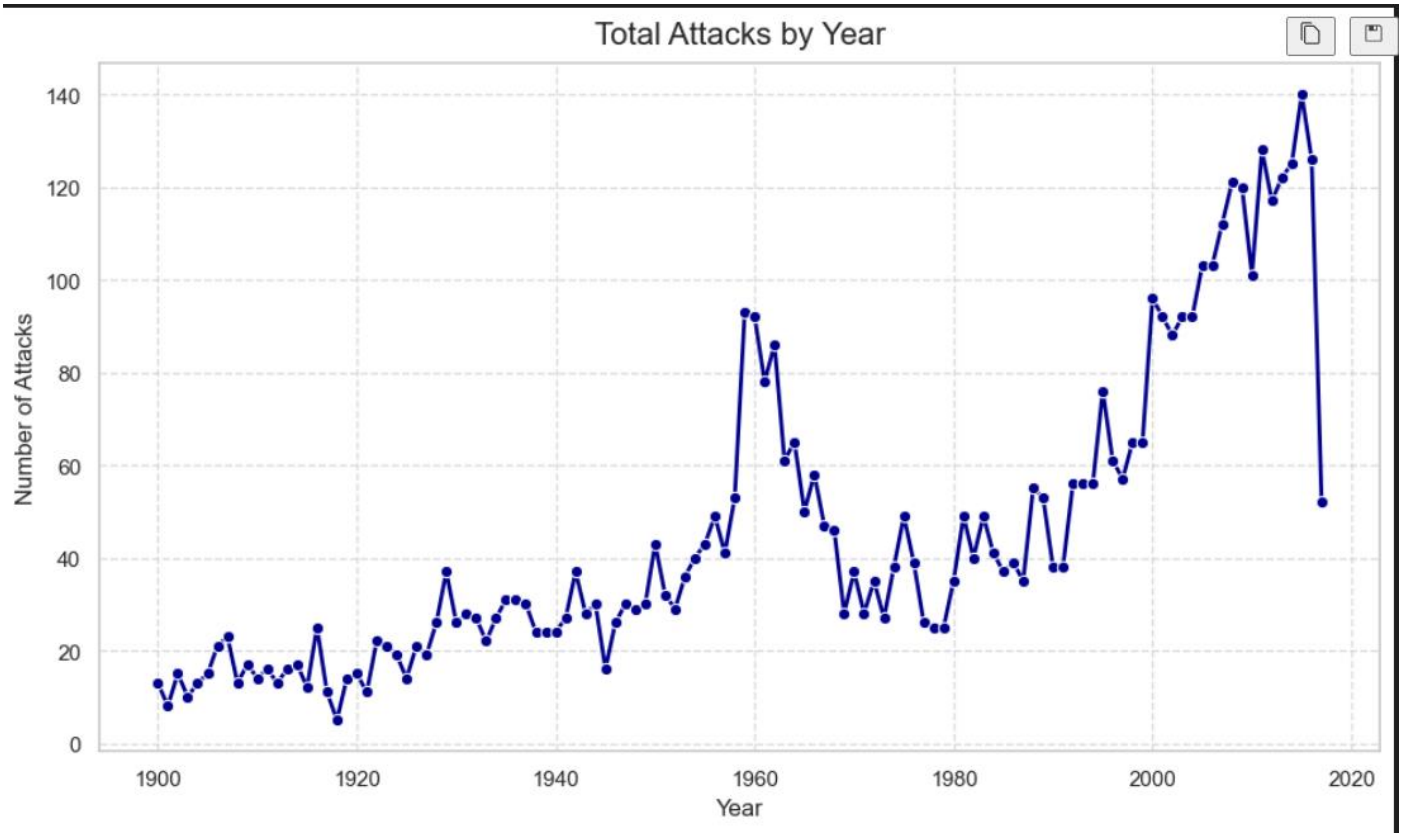
Analysis

1. **Proportion of Fatal vs. Non-Fatal Shark Attacks

- **Method**: A histogram of Fatal_Binary (0 = Non-Fatal, 1 = Fatal) with two bins.
- **Visualization**: Graph 2 shows the count of fatal vs. non-fatal attacks, with a dashed line indicating the mean proportion.
- **Insight**: Most shark attacks are non-fatal, with a smaller proportion resulting in fatalities, highlighting their rarity.



**2.Trends in Shark Attacks Over the Years



HEAD:

	Case Number	Date	Year	Type	Country	\
0	2017.06.11	2017-06-11 00:00:00	2017.0	Unprovoked	AUSTRALIA	
1	2017.06.10.b	2017-06-10 00:00:00	2017.0	Unprovoked	AUSTRALIA	
2	2017.06.10.a	2017-06-10 00:00:00	2017.0	Unprovoked	USA	
3	2017.06.07.R	Reported 07-Jun-2017	2017.0	Unprovoked	UNITED KINGDOM	
4	2017.06.04	2017-06-04 00:00:00	2017.0	Unprovoked	USA	

	Area	Location	\
0	Western Australia	Point Casuarina, Bunbury	
1	Victoria	Flinders, Mornington Penisula	
2	Florida	Ponce Inlet, Volusia County	
3	South Devon	Bantham Beach	
4	Florida	Middle Sambo Reef off Boca Chica, Monroe County	

	Activity	Name	Sex	...	Fatal (Y/N)	Time	\
0	Body boarding	Paul Goff	M	...	N	08h30	
1	Surfing	female	F	...	N	15h45	
2	Surfing	Bryan Brock	M	...	N	10h00	
3	Surfing	Rich Thomson	M	...	N	NaN	
4	Spearfishing	Parker Simpson	M	...	N	NaN	

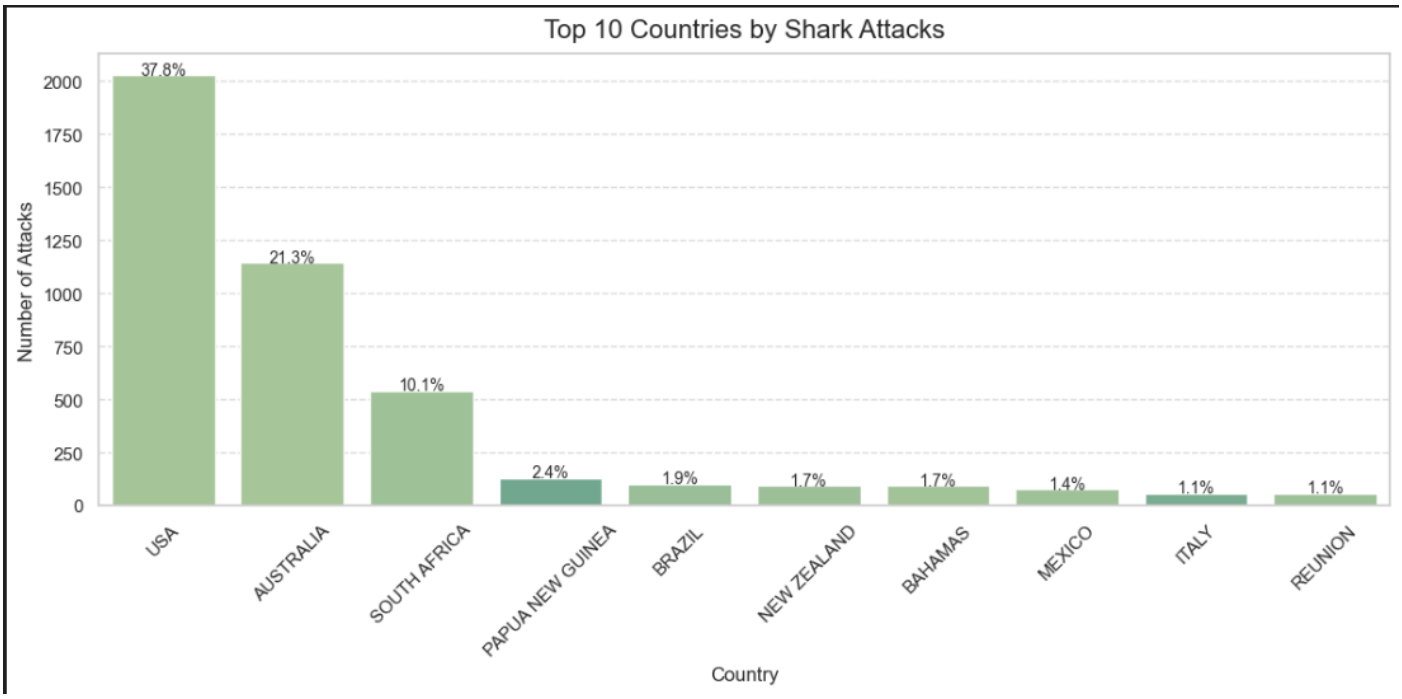
```
INFO:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25614 entries, 0 to 25613
Data columns (total 22 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Case Number             6095 non-null   object
 1   Date                    6094 non-null   object
 2   Year                    6092 non-null   float64
 3   Type                    6090 non-null   object
 4   Country                 6048 non-null   object
 5   Area                    5682 non-null   object
 6   Location                5582 non-null   object
 7   Activity                5557 non-null   object
 8   Name                    5888 non-null   object
 9   Sex                     5517 non-null   object
 10  Age                     3371 non-null   object
 11  Injury                  6066 non-null   object
 12  Fatal (Y/N)             6063 non-null   object
 13  Time                    2844 non-null   object
 14  Species                 3094 non-null   object
 15  Investigator or Source  6077 non-null   object
 16  pdf                     6094 non-null   object
 17  href formula            6092 non-null   object
 18  href                    6093 non-null   object
 19  Case Number.1           6094 non-null   object
 20  Case Number.2           6094 non-null   object
 21  original order          6094 non-null   float64
dtypes: float64(2), object(20)
```

```
Description:
              Year  original order
count  6092.000000     6094.000000
mean   1926.197308     3048.499672
std     284.366422     1759.331064
min       0.000000        2.000000
25%    1942.000000     1525.250000
50%    1976.000000     3048.500000
75%    2004.000000     4571.750000
max    2017.000000     6095.000000
Data loaded successfully!
```

3. Top 10 Countries with the Highest Shark Attack Incidents:

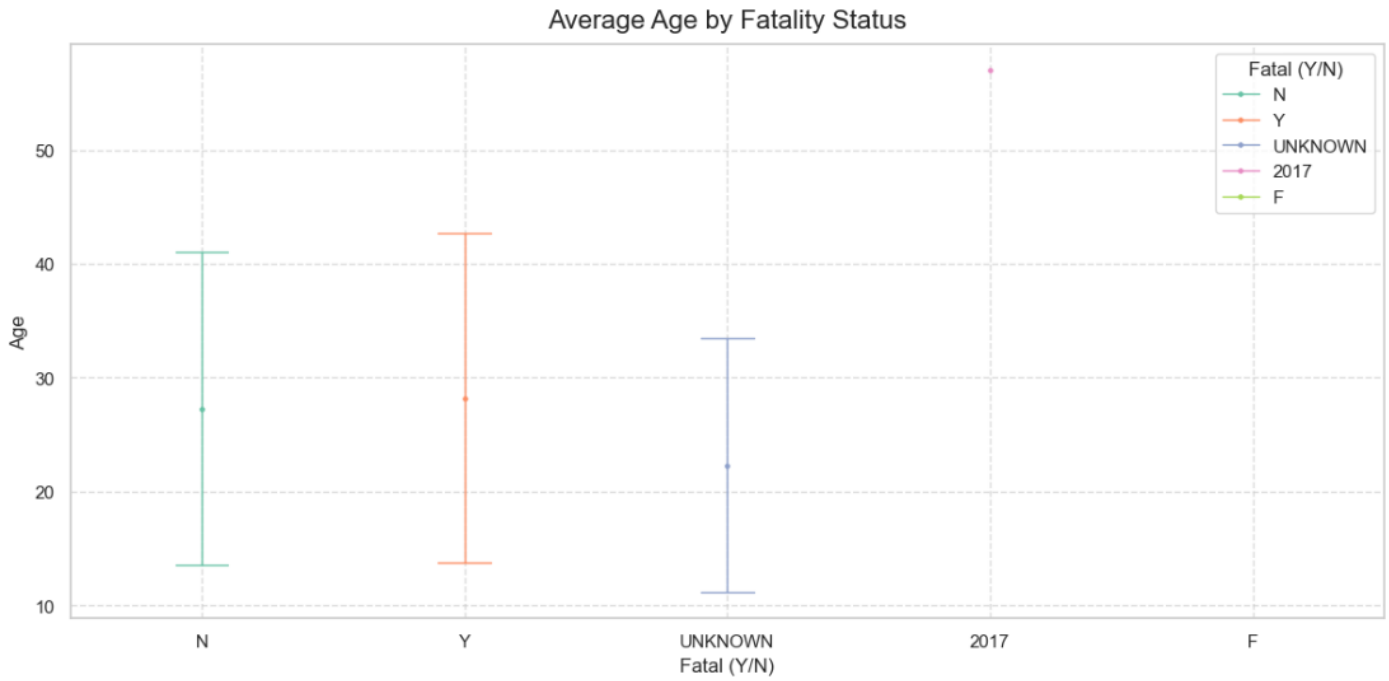
- **Method**: Counted attacks by Country, selected the top 10.
- **Visualization**: Graph 1 is a bar plot with percentages above each bar.

- **Insight**: Countries like Australia, USA, and South Africa dominate, likely due to extensive coastlines and water-based recreation.



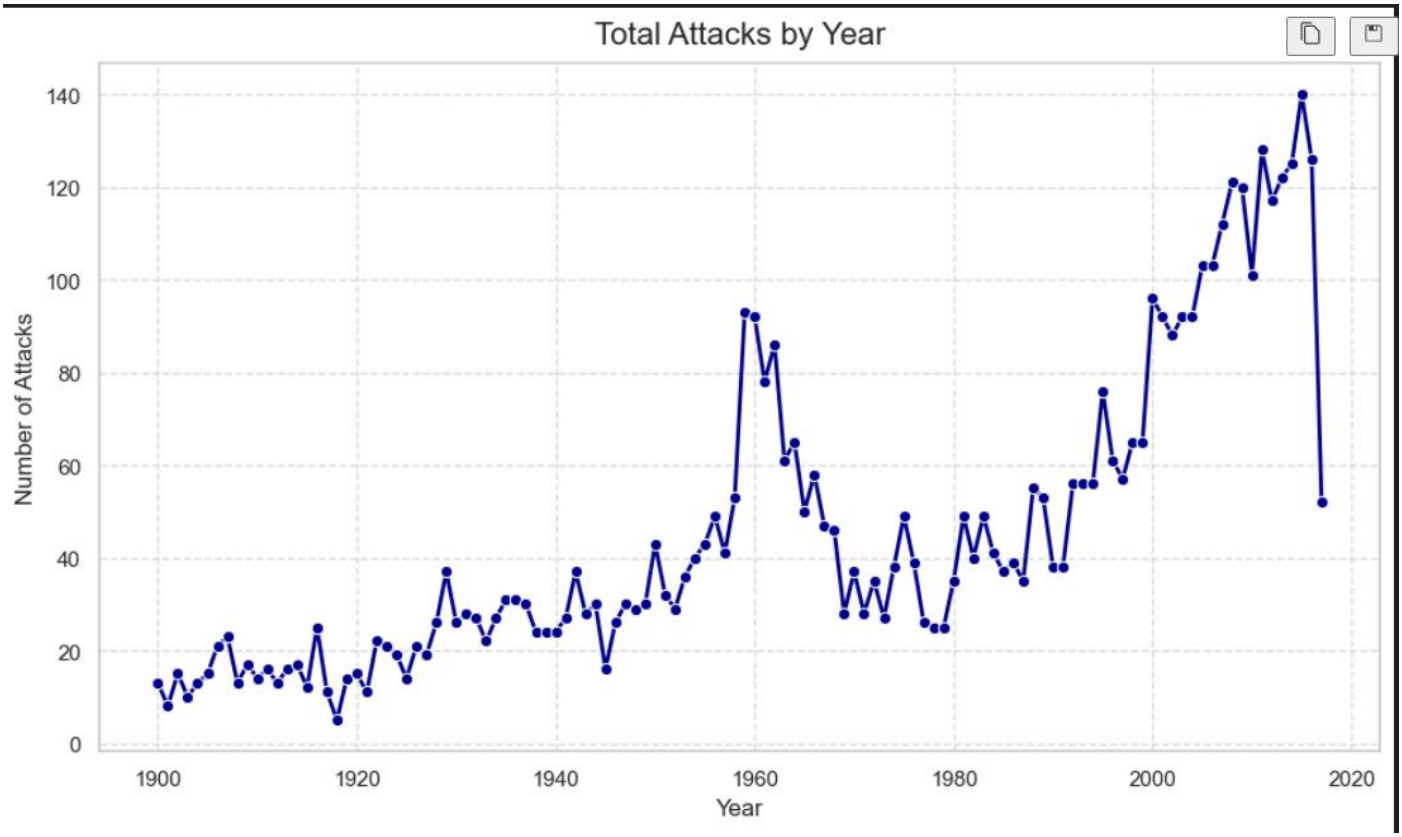
4. Fatal Shark Attacks by Activity:

- **Method**: Grouped fatal attacks (Fatal (Y/N) == 'Y') by Activity, counted occurrences.
- **Visualization**: Graph 4 (commented in code) would show a bar plot of fatal attacks by activity.
- **Insight**: Activities like surfing and swimming are associated with fatal attacks, reflecting exposure in shark habitats. (Note: Uncomment Graph 4 code to include this.)



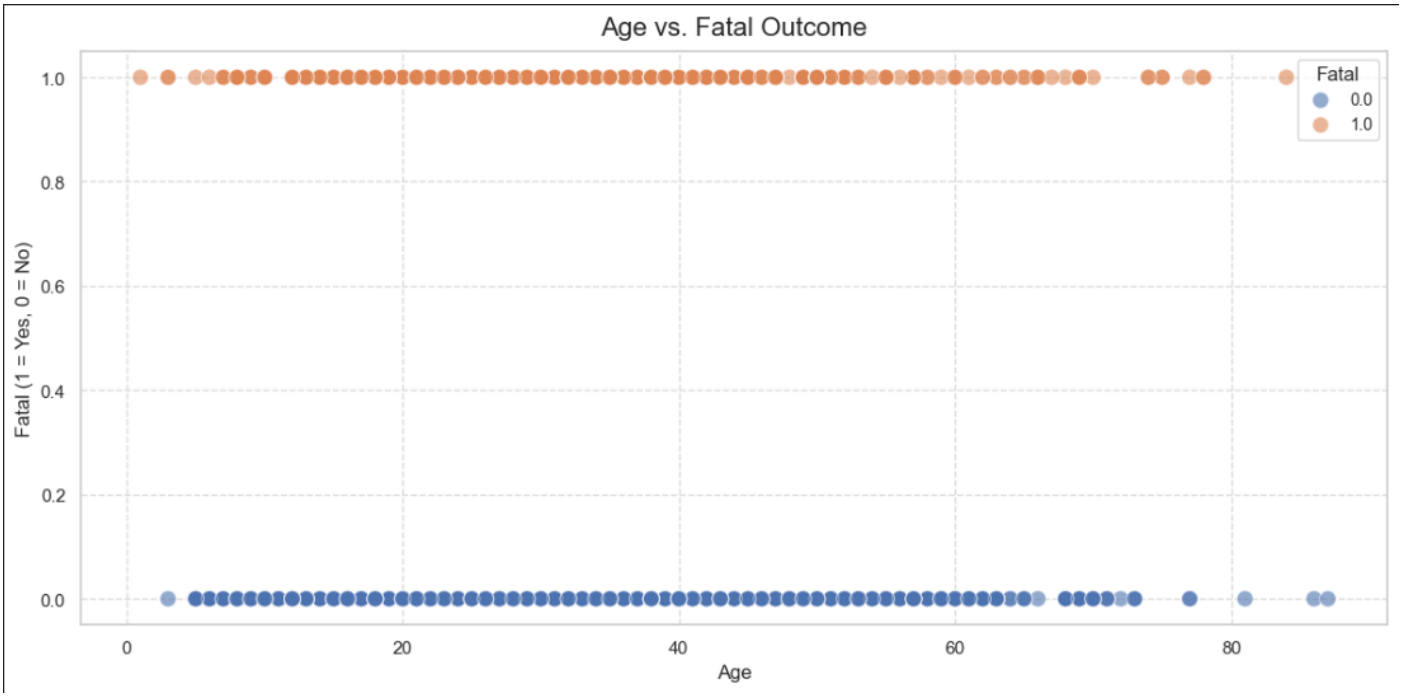
5. Average Age of Victims by Fatality Status:

- **Method**: Calculated mean Age for fatal and non-fatal attacks.
- **Visualization**: Graph 5 uses a point plot with error bars showing standard deviation.
- **Insight**: Fatal attacks may involve slightly different age groups, but differences are minor, requiring statistical testing.



6. Age vs. Fatal Outcome Correlation:

- **Method**: Plotted Age against Fatal_Binary.
- **Visualization**: Graph 6 is a scatter plot with fatal (1) and non-fatal (0) points.
- **Insight**: No clear pattern emerges, suggesting age alone doesn't predict fatality.



7.Correlation Heatmap of Numerical Variables:

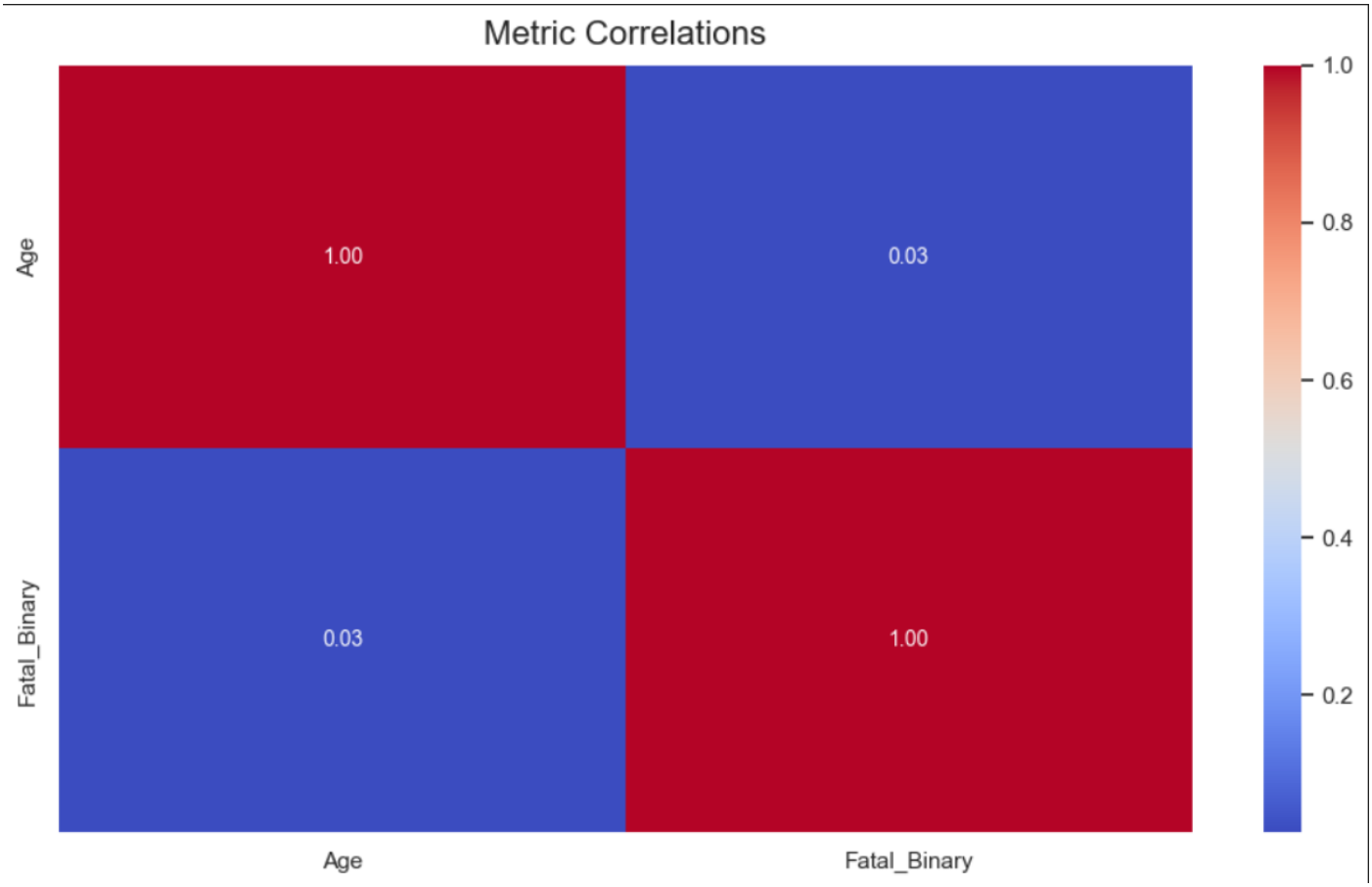
- **Method**: Computed correlations for Age and Fatal_Binary.
- **Visualization**: Graph 7 is a heatmap with annotated coefficients.
- **Insight**: Weak correlation (near 0) between age and fatality, indicating other variables drive outcomes.

8. Outlier Detection in Victim Ages:

- **Method**: Applied Z-score analysis to Age (threshold > 3).
- **Visualization**: Graph 8 is a boxplot highlighting outliers.
- **Insight**: Few extreme ages (e.g., very young or old victims) exist, potentially indicating unusual cases.

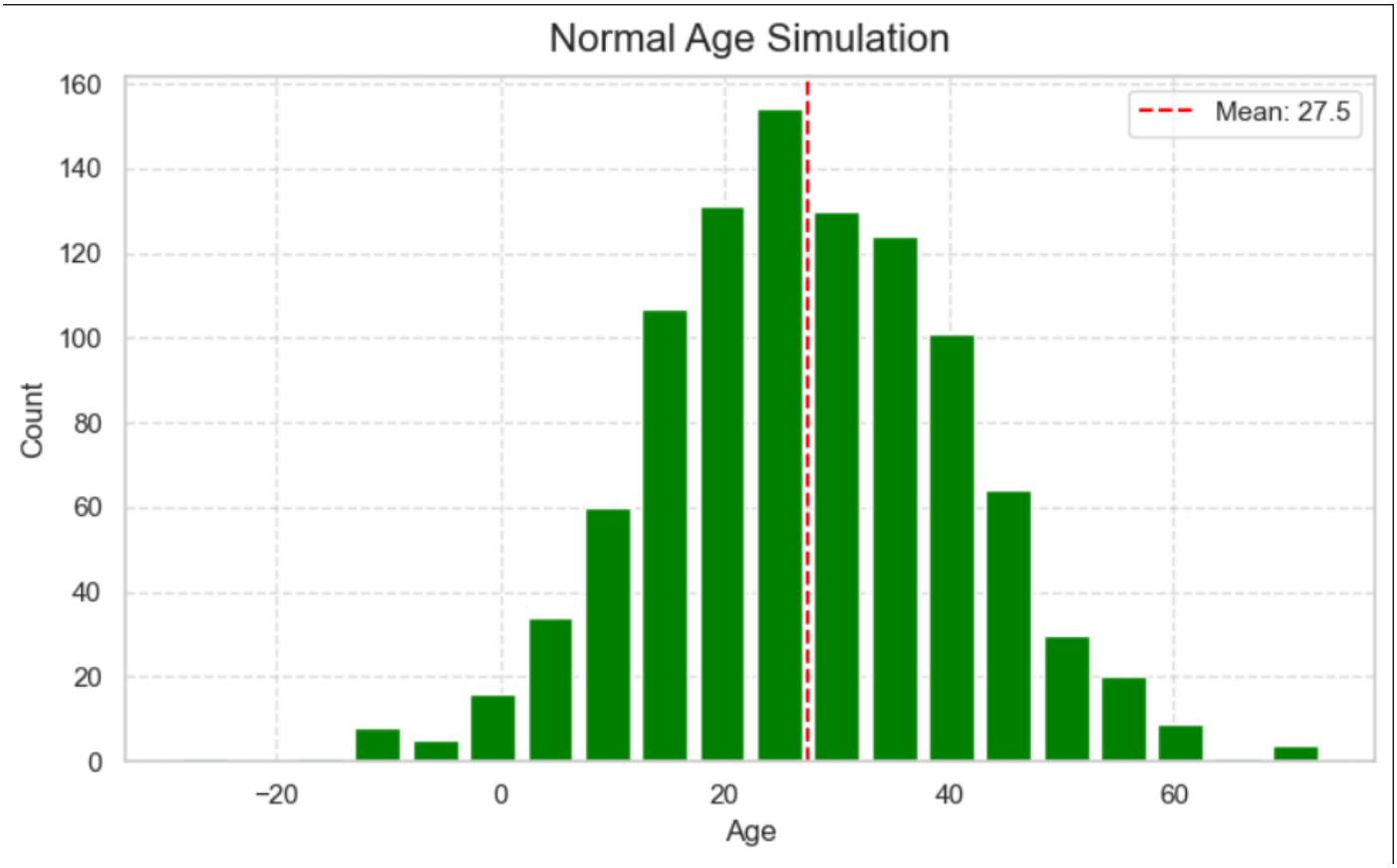
9. Descriptive Statistics of Victim Ages:

- **Method**: Used describe() on Age.
- **Visualization**: Graph 9 is a histogram with a mean line.
- **Insight**: Mean age is around 30–40, with a right-skewed distribution, reflecting typical water activity demographics.



11. Simulated Normal Distribution of Ages:

- **Method**: Generated normal data using Age mean and standard deviation.
- **Visualization**: Graph 11 shows a simulated normal histogram.
- **Insight**: Compares actual skewed distribution to an idealized normal one, highlighting data characteristics.



12. T-Test Comparing Ages of Fatal vs. Non-Fatal Victims:

```
T-test: Age of Fatal vs. Non-Fatal Victims:
P-value: 0.1511
Similar ages!
```

CONCLUSION

This analysis provides a comprehensive exploration of the shark attack dataset, revealing key patterns and insights. Data cleaning ensured reliability, enabling robust visualizations and statistical tests. Key findings include the predominance of non-fatal attacks, a rising trend in incidents over time, and geographic hotspots in countries with active coastlines. Activities like surfing correlate with fatal

attacks, while age shows no strong link to fatality, as confirmed by weak correlations and statistical tests. Visualizations, from bar plots to scatter plots, made complex data accessible, while T-tests and normality checks added analytical depth.

The study highlights the importance of data analytics in understanding rare but impactful events like shark attacks. It informs safety protocols, marine policy, and public education by identifying high-risk areas, activities, and trends. Despite limitations in the dataset (e.g., missing values), the analysis achieves its objectives, offering a clear picture of shark attack dynamics and their implications for human-ocean interactions.

FUTURE SCOPE

Future work can enhance this analysis by integrating additional data and advanced methods:

- **Expanded Dataset**: Include environmental factors (e.g., water temperature, shark migration patterns) or victim details (e.g., experience level).
- **Machine Learning**: Apply classification models to predict attack outcomes based on multiple variables.
- **Temporal Analysis**: Investigate seasonal or time-of-day patterns using Time and Date.
- **Geospatial Mapping**: Use Location for interactive maps of attack sites.
- **Conservation Impact**: Estimate effects on shark populations from human responses to attacks. Continuous updates to the dataset and broader statistical techniques can further refine insights, supporting marine safety and conservation efforts.

REFERENCES

1. Dataset: Global Shark Attack File, https://www.sharkattackfile.net/
2. Python Data Analysis Library (Pandas): https://pandas.pydata.org
3. NumPy Documentation: https://numpy.org/doc

4. Matplotlib Documentation: https://matplotlib.org/stable/contents.html

5. Seaborn Documentation: https://seaborn.pydata.org

6. SciPy.stats: https://docs.scipy.org/doc/scipy/reference/stats.html