# Boat Sales

# This data set is about the sales data of boats.

# **Data Source**

This data set is from the Kaggle.com website.

<https://www.kaggle.com/datasets/karthikbhandary2/boat-sales/data>

Kaggle, operating under Google LLC, serves as a platform and online community for data scientists and machine learning practitioners. It facilitates users in discovering and sharing datasets, creating models within a web-based data science environment, collaborating with fellow professionals, and participating in competitions aimed at addressing data science challenges.

This data set is all about the popularity of the boats. It has 10 columns. It has details of the boats regarding the year it was built the type of boat, price and much more.

## Columns:

**Price**Character, boat price listed in different currencies (e.g. EUR, Â£, CHF etc.) on the website

**Boat Type**Character, type of the boat

**Manufacturer**Character, manufacturer of the boat

**Type**Character, condition of the boat and engine type(e.g. Diesel, Unleaded, etc.)

**Year Built**Numeric, year of the boat built

**Length**Numeric, length in meter of the boat

**Width**Numeric, width in meter of the boat

**Material**Character, material of the boat (e.g. GRP, PVC, etc.)

**Location**Character, location of the boat is listed

**Number of views last 7 days**Numeric, number of the views of the list last 7 days

## Why I have choosen this data set?

I used to sail a lot with my husband, we also had our own boat on Balaton, the largest in Hungary, that's why I chose this line of data

# **Data Profile**

## Data Cleaning

I used phyton to perform the data cleaning, with Jupiter notebook. Missing values were found in the following columns: manufacturer, material, location, width. I deleted the rows with the missing values from the database.

Number of Missing Values:

Price 0

Boat Type 0

Manufacturer 1338

Type 6

Year Built 0

Length 9

Width 56

Material 1749

Location 36

Number of views last 7 days 0

I did not find any duplicate.

When examining the outliers in the bank database, I analyzed three numerical columns using statistical methods: the ‘Year Built’, ‘Length’, and ‘Width’ columns. There was also a fourth numerical column in the database, but since it represented the number of internet views, there was no point in investigating an outlier using mathematical methods, as a very popular ship can produce extreme outlier views. Here, I only checked that there were no completely unrealistic values in the data. When examining the outliers, it became apparent that many ships had a zero written for their year of construction. This was obviously not missing data but outliers. However, in practice, we can consider this as missing data. I entered this in such a way that I filled in the median instead of zero.

## Descriptive Statistical Analysis

Columns and basic info about dataframe:

Index: 7019 entries, 1 to 9887

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Price 7019 non-null object

1 Boat Type 7019 non-null object

2 Manufacturer 7019 non-null object

3 Type 7019 non-null object

4 Year Built 7019 non-null int64

5 Length 7019 non-null float64

6 Width 7019 non-null float64

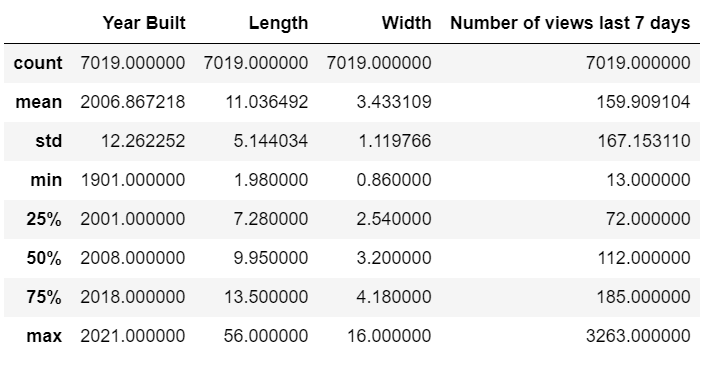
7 Material 7019 non-null object

8 Location 7019 non-null object

9 Number of views last 7 days 7019 non-null int64

dtypes: float64(2), int64(2), object(6)

Descriptive statistical analysis of numerical columns:



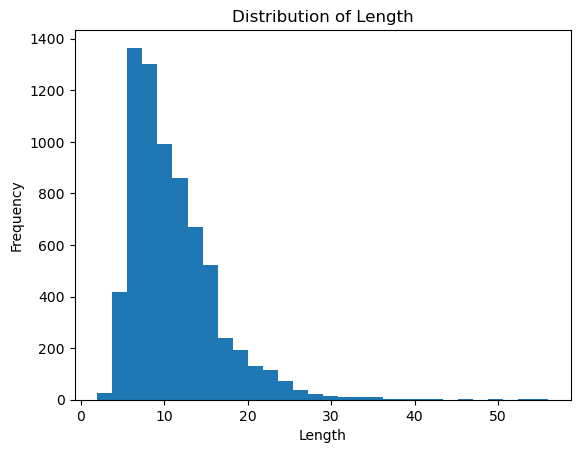
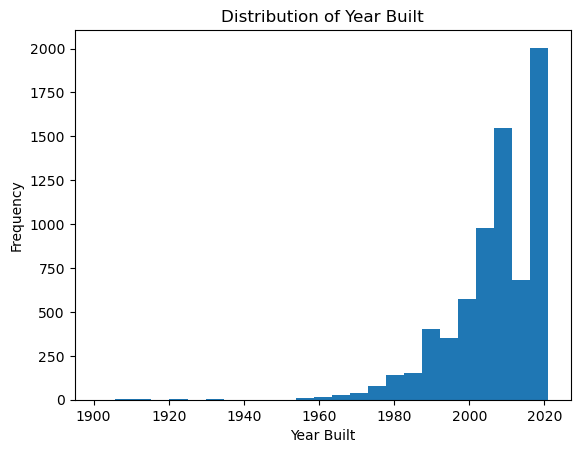
**Histograms of the numerical columns:**

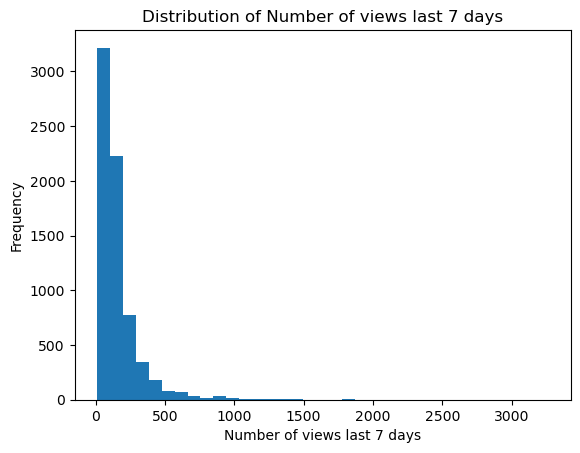
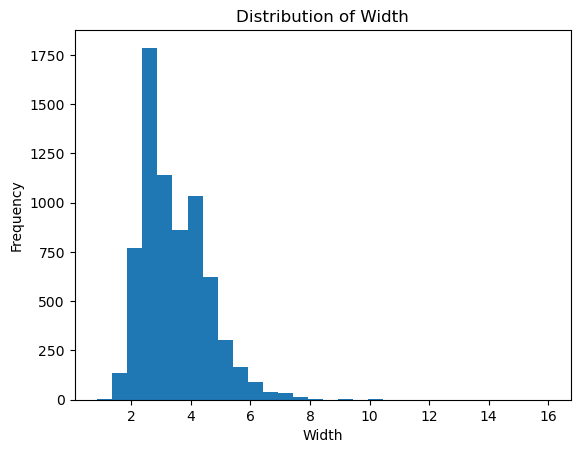
Year Built: left skewed distribution

Length: right skewed distribution

Width: right skewed distribution

Number of Views: right skewed distribution





## Limitations and Data Ethics

LIMITATIONS:

**1. Data Source Reliability:**

- The dataset is obtained from Kaggle, a platform for data scientists. While Kaggle is reputable, there might be variations in data quality based on individual contributors. No Warranty: CC0: Public Domain license includes a disclaimer stating that the work is provided "as is" without any warranties, which means the creator does not provide any guarantees or assurances regarding the quality or fitness for a particular purpose of the work.

**2. Data Cleaning Decisions:**

- The decision to delete rows with missing values might lead to a loss of potentially valuable information.

ETHICAL CONSIDERATIONS:

**1. Privacy Concerns:**

- The dataset involves information about boats and their characteristics. While it may not contain personal information directly, sharing similar datasets with more sensitive data could pose privacy concerns.

**2. Data Ownership and Attribution:**

- The Kaggle dataset has CC0: Public Domain licence. This means that the work is placed in the public domain, and anyone can use, modify, distribute, or otherwise utilize the work for any purpose, without any restrictions or requirements. The creator explicitly states that they have waived all copyright and related rights, effectively placing the work in the public domain. No Restrictions: Users can copy, modify, distribute, and perform the work, even for commercial purposes, without seeking permission from or providing attribution to the original creator. Global Applicability: The CC0 license is intended to be effective worldwide, to the extent permitted by law.

# Research Questions

You are working as a data analyst for a yacht and boat sales website. The marketing team is preparing a weekly newsletter for boat owners. The newsletter is designed to help sellers to get more views of their boat, as well as stay on top of market trends. They would like me to take a look at the recent data and get some insights.

The possible questions that we can ask ourselves are:

1. characteristics of the most viewed boats in the last 7 days
2. is it the most expensive boats that get the most views?
3. What are the geographical position of the most viewed boats?