# Project: Boat Sales

#### **Data Summary & Source**

This data set is one of the proposed ones in the Achievement 6 Brief, and it was sourced from Kaggle, a data science and artificial intelligence free platform where users can get access and share data sets, data science work and tips, enter discussion forums, competitions, courses etc.

The information refers to sales data of a yacht and boat website, where the marketing team wants improve results by releasing a weekly newsletter with information and advice to their boat owners and is therefore in need of some data insight.

#### Data Profile

This data set is open source and recent (2021), comprised of one csv file, (size 1.15 MB, 10 columns and 9883 rows) and expected to be updated annually.

#### Boat data.csv

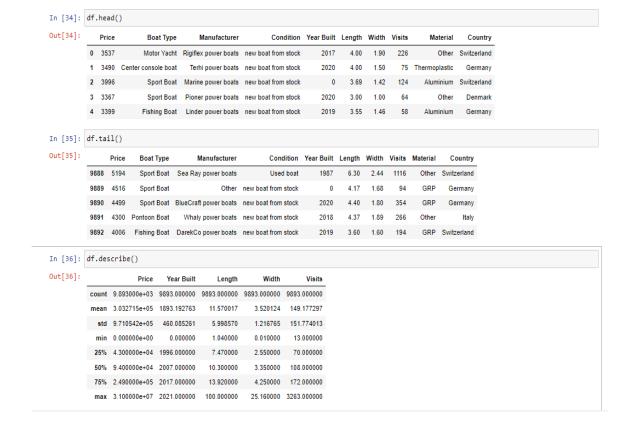
Original Data		Data Type				
Variables	Description	Time Variant/Invariant	Structured/Unstructured	Qualitative/Quantitative	Qualitative: nominal/ordinal Quantitative: discrete/continuous	
Price	Sale price ( Eur, CHF, DKK, GBP, etc)	Variant	structured	Quantitative	Continuous	
Boat Type	Type of boat/model	Invariant	structured	Qualitative	Nominal	
Manufacturer	Manufacturer identification	Invariant	structured	Qualitative	Nominal	
Туре	Condition info/Avalability info/Fuelinfo	Invariant	structured	Qualitative	Nominal	
Year Built	Year of boat built	Variant	structured	Quantitative	Discrete	
Lenght	Size: Boat length	Invariant	structured	Quantitative	Continuous	
Width	Size: Boat width	Invariant	structured	Quantitative	Continuous	
Material	Construction material	Invariant	structured	Qualitative	Nominal	
Location	Present boat location ( Country/ City / Others)	Invariant	structured	Qualitative	Nominal	
number of views last 7 days	Number of website visits on a 7 day time period	Variant	structured	Qualitative	Discrete	

## Data Cleaning & Wrangling

Prepared Data		Data Wrangling			
Variables	Variables Description		Tool	Reason	
Price	Sale price ( Eur, CHF, DKK, GBP, etc)	Convert all currency values to Euros	Excel	Data consistency checks	
Boat Type	Type of boat/model	-	-	-	
		All missing values replaced my 'Other'			
Manufacturer	Manufacturer identification	value	Jupyter Notebook	Data consistency checks	
Condition	Condition info/Avalability info/Fuelinfo	Renamed column as Condition	Renamed column as Condition Jupyter Notebook Data consiste		
		All missing values replaced my mean			
Year Built	Year of boat built	value	Jupyter Notebook	Data consistency checks	
		All missing values replaced my mean			
Lenght	Size: Boat length	value	Jupyter Notebook	Data consistency checks	
		All missing values replaced my mean			
Width	Size: Boat width	value	Jupyter Notebook	Data consistency checks	
		All missing values replaced my 'Other'			
Material	Construction material	value	Jupyter Notebook	Data consistency checks	
		split column into two :Country and City			
Country	Present boat location ( Country/ City / Others)	Dropped City column	Excel	Data consistency checks	
	Country Country	All missing values replaced my mean			
		value		]	
Visits	Number of website visits on a 7 day time period	Renamed column as Visits	Jupyter Notebook	Data consistency checks	

### Data Understanding

```
In [33]: df.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 9893 entries, 0 to 9892
         Data columns (total 10 columns):
                          Non-Null Count Dtype
         # Column
          0
              Price
                          9893 non-null
                                           int32
                          9893 non-null
              Boat Type
                                           object
          1
              Manufacturer 9893 non-null
          2
                                           object
          3
              Condition
                           9893 non-null
                                            object
             Condition 9893 non-null
Year Built 9893 non-null
          4
                                           int32
          5
              Length 9893 non-null
                                           float64
          6
             Width
                           9893 non-null
                                           float64
              Visits
                           9893 non-null
                                           int32
             Material
          8
                           9893 non-null
                                           object
             Country
                          9893 non-null
         dtypes: float64(2), int32(3), object(5)
         memory usage: 657.1+ KB
```



#### Data Ethics & Limitations:

Even though it is an open based public data source, it is very important to keep in mind that there is no available information about the data collection source or methods used, and that also the data set is dated of 30/11/2021, and it was supposed to be updated annually. On the other hand, the data collected on the Year built, years of boat construction, does seem to have some odd values, so some caution is advised on using this variable. Regarding sensible/ private information there is no PII items, so there is no need for special measures to be taken.

## Questions to explore:

- What's the higher/lower visit numbers? (Number of website visits in the last 7 days)
- Who is the most/least visited boat type?
- Where (geographically) are the most/least visit numbers?
- Do viewers prefer new or used boats?
- What price range was most/least viewed?
- Who was the most viewed manufacturer?
- What is the preferred boat material type? Most seen?