Boat Sales Analysis

Data Source

Data sourcing:
Boat Sales
Nearly New Nautical is a website that allows users to advertise their used boats for sale. 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.

Data Collection:
Data is collected by Karthik Bhandary on Kaggle.

Data contents:
This dataset provides information related to boat sales. The data set contains 9888 rows and 10 columns.

Data ethics:
The Dataset does not contain any sensitive information.

Limitations:
Data set has missing values. There is no information on the data collection methods and it is unknown if the data set is trustworthy.

Data Profile

Variable	Description	Time variant/ invariant	Structured/ Unstructured	Qualitative/ Quantitative	Sub-types of qual/quant
Price	Character, boat price listed in different currencies (e.g.EUR, Â£, CHF etc.) on the website	Variant	Structured	Quantitative	Discrete
Boat Type	Character, type of the boat	Invariant	Structured	Qualitative	Nominal
Manufacturer	Character, manufacturer of the boat	Invariant	Structured	Qualitative	Nominal
Type	Character, condition of the boat and engine type(e.g. Diesel, Unleaded, etc.)	Variant	Unstructured	Qualitative	Ordinal
Year	Built Numeric, year of the boat built	Invariant	Structured	Qualitative	Discrete
Length	Numeric, length in meter of the boat	Invariant	Structured	Quantitative	Discrete
Width	Numeric, width in meter of the boat	Invariant	Structured	Quantitative	Discrete
Material	Character, material of the boat (e.g. GRP, PVC, etc.)	Invariant	Structured	Qualitative	Binary
Location	Character, location of the boat is listed	Invariant	Structured	Qualitative	Nominal
Number of views last 7 days	Numeric, number of the views of the list last 7 days	Variant	Structured	Quantitative	Discrete

Descriptive analysis

	Year Built	Length	Width	Number of views last 7 days
count	9888.000000	9879.000000	9832.000000	9888.000000
mean	1893.192860	11.570017	3.520124	149.160801
std	460.201582	6.002820	1.220534	151.819752
min	0.000000	1.040000	0.010000	13.000000
25%	1996.000000	7.470000	2.540000	70.000000
50%	2007.000000	10.280000	3.330000	108.000000
75%	2017.000000	13.930000	4.250000	172.000000
max	2021.000000	100.000000	25.160000	3263.000000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9888 entries, 0 to 9887
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Price                      9888 non-null   object
 1   Boat Type                  9888 non-null   object
 2   Manufacturer               8550 non-null   object
 3   Type                       9882 non-null   object
 4   Year Built                 9888 non-null   int64
 5   Length                     9879 non-null   float64
 6   Width                      9832 non-null   float64
 7   Material                   8139 non-null   object
 8   Location                   9852 non-null   object
 9   Number of views last 7 days  9888 non-null   int64
dtypes: float64(2), int64(2), object(6)
memory usage: 772.6+ KB
```

Data Cleaning

Consistency	Column name	Changes
Descriptive analysis	Type	Was renamed to ‘Condition’
Mixed-data column	Price	Column was split into ‘Currency’ and ‘Price’, then new column was added converting ‘Price’ into USD as a new column ‘Price_in_usd’
Inconsistency	Location	Location column was split into ‘Country’ and ‘City’, but Location column had signs which needed to be dropped, so only column ‘Country’ was kept
Null	Price, Location, Currency, City	Dropped (due to unnecessity)
Null	Material, Condition	Filled with the most met value
Null	Length, Width	Filled with median value

Questions to explore with analysis

1. Is it the most expensive boats that get the most views?
2. Characteristics of the most viewed boat listings in the last 7 days
3. Does the view depend on the condition, material and price of the boat?