

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
Type	Condition info/Avalability info/ Fuel info	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	Description	Action	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	-	-	-
Manufacturer	Manufacturer identification	All missing values replaced my 'Other' value	Jupyter Notebook	Data consistency checks
Condition	Condition info/Avalability info/ Fuel info	Renamed column as Condition	Jupyter Notebook	Data consistency checks
Year Built	Year of boat built	All missing values replaced my mean value	Jupyter Notebook	Data consistency checks
Lenght	Size: Boat length	All missing values replaced my mean value	Jupyter Notebook	Data consistency checks
Width	Size: Boat width	All missing values replaced my mean value	Jupyter Notebook	Data consistency checks
Material	Construction material	All missing values replaced my 'Other' value	Jupyter Notebook	Data consistency checks
Country	Present boat location (Country/ City / Others)	split column into two :Country and City Dropped City column	Excel	Data consistency checks
Visits	Number of website visits on a 7 day time period	All missing values replaced my mean value 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):
#   Column          Non-Null Count  Dtype
---  ---
0   Price           9893 non-null   int32
1   Boat Type       9893 non-null   object
2   Manufacturer     9893 non-null   object
3   Condition        9893 non-null   object
4   Year Built      9893 non-null   int32
5   Length          9893 non-null   float64
6   Width           9893 non-null   float64
7   Visits          9893 non-null   int32
8   Material        9893 non-null   object
9   Country         9893 non-null   object
dtypes: float64(2), int32(3), object(5)
memory usage: 657.1+ KB
```

In [34]: df.head()

Out[34]:

	Price	Boat Type	Manufacturer	Condition	Year Built	Length	Width	Visits	Material	Country
0	3537	Motor Yacht	Rigiflex power boats	new boat from stock	2017	4.00	1.90	226	Other	Switzerland
1	3490	Center console boat	Terhi power boats	new boat from stock	2020	4.00	1.50	75	Thermoplastic	Germany
2	3996	Sport Boat	Marine power boats	new boat from stock	0	3.69	1.42	124	Aluminium	Switzerland
3	3367	Sport Boat	Pioner power boats	new boat from stock	2020	3.00	1.00	64	Other	Denmark
4	3399	Fishing Boat	Linder power boats	new boat from stock	2019	3.55	1.46	58	Aluminium	Germany

In [35]: df.tail()

Out[35]:

	Price	Boat Type	Manufacturer	Condition	Year Built	Length	Width	Visits	Material	Country
9888	5194	Sport Boat	Sea Ray power boats	Used boat	1987	6.30	2.44	1116	Other	Switzerland
9889	4516	Sport Boat	Other	new boat from stock	0	4.17	1.68	94	GRP	Germany
9890	4499	Sport Boat	BlueCraft power boats	new boat from stock	2020	4.40	1.80	354	GRP	Germany
9891	4300	Pontoon Boat	Whaly power boats	new boat from stock	2018	4.37	1.89	266	Other	Italy
9892	4006	Fishing Boat	DarekCo power boats	new boat from stock	2019	3.60	1.60	194	GRP	Switzerland

In [36]: df.describe()

Out[36]:

	Price	Year Built	Length	Width	Visits
count	9.893000e+03	9893.000000	9893.000000	9893.000000	9893.000000
mean	3.032715e+05	1893.192763	11.570017	3.520124	149.177297
std	9.710542e+05	460.085261	5.998570	1.216765	151.774013
min	0.000000e+00	0.000000	1.040000	0.010000	13.000000
25%	4.300000e+04	1996.000000	7.470000	2.550000	70.000000
50%	9.400000e+04	2007.000000	10.300000	3.350000	108.000000
75%	2.490000e+05	2017.000000	13.920000	4.250000	172.000000
max	3.100000e+07	2021.000000	100.000000	25.160000	3263.000000

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?