Data Acquisition and Exploration of Boat Sales

Rabbani Nur Kumoro
(21/472599/PA/20310)
Department of Computer Science and Electronics, Universitas Gadjah Mada
Building C, 4th Floor North
Yogyakarta, Indonesia
rabbani.nur.kumoro@mail.ugm.ac.id

***Abstract* — Data acquisition and exploration involves gathering, analyzing, and visualizing patterns and hidden insights from various sources of data. In this report, we are going to explore the characteristics of the boat listings in the last 7 days and investigate the sales of each boat. The report also explores common features among the most viewed boats. The findings indicate that certain features such as boat size, age, and brand are common among the most viewed boats. Moreover, the report highlights that the most expensive boats are not necessarily the most viewed. The insights generated from this report can be utilized by the marketing team to help sellers get more views of their boats and stay on top of market trends by using various machine learning methods.**

Keywords — Boat Sales, Data Acquisition, Exploratory Data Analysis, Machine Learning

I. Dataset

This report uses a dataset from DataCamp's case study through the Kaggle platform to gain insights into the characteristics of the most viewed boat listings on Nearly New Nautical, a website for advertising used boats. The dataset includes sales data for boats made within the last 7 days.

The task aims to provide valuable insights for sellers to improve their listings and stay on top of market trends. Identifying the characteristics of the most viewed boats can help sellers optimize their listings for better visibility and attract more buyers. This information can also help the marketing team develop effective strategies to increase views and sales on the platform.

This dataset has 10 columns and 9888 rows. It has details of the boats regarding the year it was built the type of boat, the price, and much more.

Here are some descriptions of features from the dataset:

Table 1. *Explanation of Features in the Dataset*

Features	Data Type	Explanation
Price	Character	Boat prices are listed in different currencies such as EUR, Â£, CHF, DKK, etc. on the website.
Boat Type	Character	Types of the Boat
Manufacturer	Character	Manufacturers of the Boat
Type	Character	Condition of the Boat and Engine Type such as 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	Materials used for the Boat such as 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

The first 5 datasets from ***Figure 1*** are generated using *'df.head'* in Python. Just from a glance, we can see that there are some NaN values in the Material column.

	Price	Boat Type	Manufacturer	Type	Year Built	Length	Width	Material	Location	Number of views last 7 days
0	CHF 3337	Motor Yacht	Rigiflex power boats	new boat from stock	2017	4.00	1.90	NaN	Switzerland Â» Lake Geneva Â» VÃ©senaz	226
1	EUR 3490	Center console boat	Terhi power boats	new boat from stock	2020	4.00	1.50	Thermoplastic	Germany Â» BÃ¶nningstedt	75
2	CHF 3770	Sport Boat	Marine power boats	new boat from stock	0	3.69	1.42	Aluminium	Switzerland Â» Lake of Zurich Â» StÃ¤fa ZH	124
3	DKK 25900	Sport Boat	Pioner power boats	new boat from stock	2020	3.00	1.00	NaN	Denmark Â» Svendborg	64
4	EUR 3399	Fishing Boat	Linder power boats	new boat from stock	2019	3.55	1.46	Aluminium	Germany Â» Bayern Â» MÃ¼nchen	58

Figure 1. *Example of the First 5 Datasets*

There seems to be a considerable amount of missing data, for 6 columns in ***Figure 2***.

```
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
dtype: int64
```

Figure 2. *Missing Value from the Dataset*

II. Data Visualization

In this section, we are going to tackle business questions, to help sellers get more views of their boats. So, we would like to answer these types of questions, such as finding out which type of boat gets the most views, common features among the most viewed boat, and highlights of the most expensive boats.

Initially, we would like to know what type of boat has the most views, therefore boat sales are divided into 2 types:

I. New boats with an average price of around 40k - 75k euros,
II. Used boats with an average price of around 10k - 30k euros.

From the price and views mentioned in ***Figure 3***, the used boat tends to get more views without considering the price. Therefore higher priced boats do not necessarily get more views.

Figure 3. *Scatter Plots Comparison*

For newer boats, the cheap and expensive both get high views as we can see from ***Figure 4***.

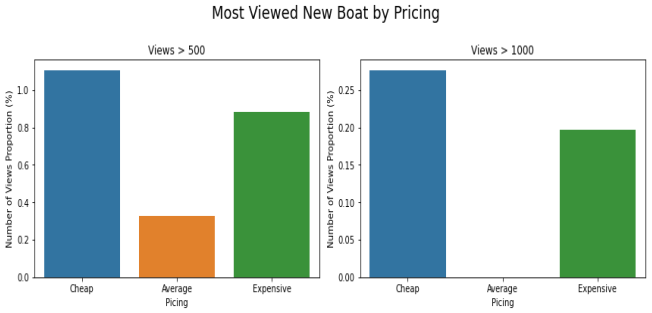

Figure 4. *Bar Charts of Most Viewed New Boats*

For used boats, it seems that there is an inverse relationship between price and the number of views they receive as it was mentioned in ***Figure 5***.

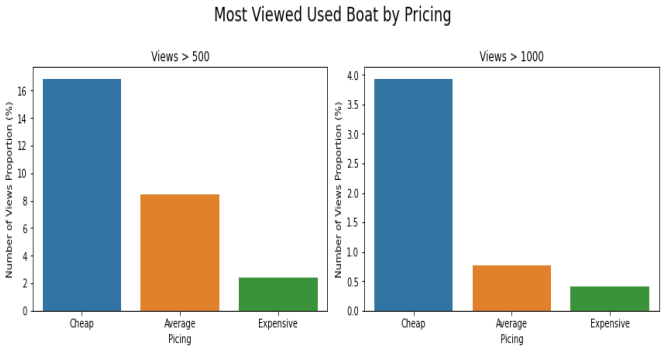

Figure 5. *Bar Charts of Most Viewed Used Boats*

From the graph in ***Figure 6*** we can conclude that the used boats are more popular, boats under 5 years are the most trending, and the second one is between 6–16 years.

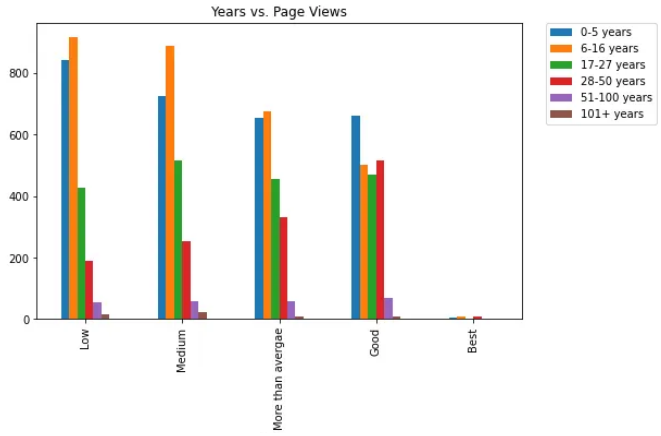

Figure 6. *Bar Chart Comparison of Years and Page Views*

In ***Figure 7***, a lot of consumers are looking for Sports Boats, as they are suitable for individuals who love water sports.

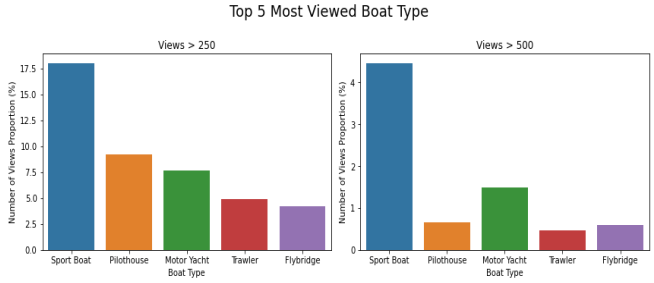

Figure 7. *Bar Chart of Most Viewed Boat Type*

From ***Figure 8***, we can conclude that out of the most popular boats, the ones built in the years 1996 and 2020 are the highest in number and the most used material is GRP. Boat made with GRP material is significantly hot in the current market as GRP is lighter and stronger. It's easier for boat owners to repair boats.

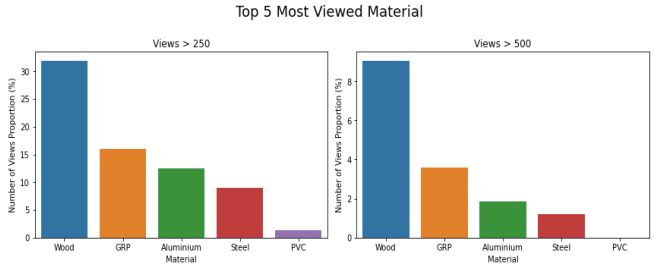

Figure 8. *Bar Chart Most Viewed Materials*

Moreover, we want to know if there are common features among the most viewed boats. As we can see in ***Figure 9***, the newer the year built, the fewer the views of used boats. New boat production is mainly from 2017 - 2021. It doesn't have a high relationship. But, from the proportion > 250 views, it tends to increase by the year.

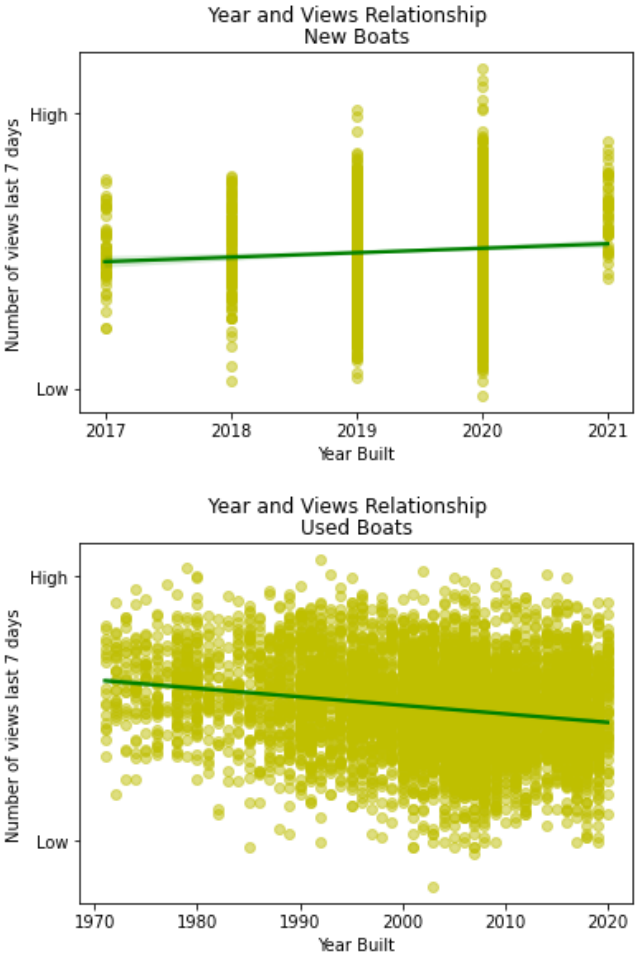

Figure 9. *Scatter Plot Comparison of Year and Views Relationship between New and Used Boats*

High views of used boats have a narrower area of the boats. Meanwhile, the area doesn't affect the views of new boats as it was seen in ***Figure 10***.

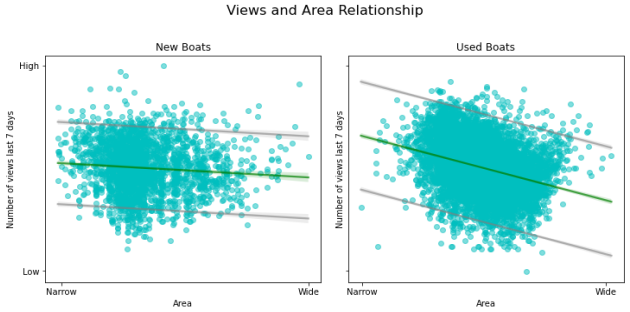

Figure 10. *Scatter Plot of Views and Area Relationship of New and Used Boats*

In summary, cheaper used boats tend to get more views than expensive ones, but both new cheap and new expensive boats are highly viewed. Additionally, the most viewed boats tend to share features such as being sport boats, made of wood, and having a lower area. Lastly, used boats are

more viewed in older years while new boats are more viewed in newer years.

III. Data Analysis

After conducting an analysis of the dataset mentioned earlier, we have determined that the dataset is of good quality. However, there is still room for improvement, as some data is incomplete and requires cleaning to obtain more accurate results.

Despite this, we believe that the dataset is ready to be utilized for machine learning tasks. With some minor data pre-processing and feature engineering, the dataset can be transformed into a strong candidate for classification tasks.

Therefore, we recommend performing these steps before proceeding with any machine learning tasks using this dataset. By doing so, we can ensure that the dataset is optimized for the task at hand and that the results obtained are accurate and reliable.

In conclusion, while the dataset is already of good quality, it can be improved further through careful data pre-processing and feature engineering. Once these steps are completed, the dataset will be an excellent choice for machine learning tasks, particularly classification tasks.

Reference

[1] Bhandary, K. *Boat Sales*. Kaggle, Datacamp, Available at: https://www.kaggle.com/datasets/karthikbhandary2/boat-sales (Accessed: February 27, 2023)

[2] Chandra, A., 2019. *Memahami Data Dengan Exploratory Data Analysis*. Data Folks Indonesia, Medium, Available at: https://medium.com/data-folks-indonesia/memahami-data-dengan-exploratory-data-analysis-a53b230cce84 (Accessed: March 2, 2023)

[3] Krishnan, M., 2020. *Exploratory data analysis using supermarket sales data in Python*. Towards Data Science, Medium, Available at: https://towardsdatascience.com/exploratory-data-analysis-using-spermarket-sales-data-in-python-e99d329a07fc (Accessed: March 1, 2023)

[4] Patil, P. 2018. *What is Exploratory Data Analysis?*. Towards Data Science, Medium, Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15 (Accessed: March 2, 2023)