Keanu Gomes
A6.1 Sourcing Open Data
January 23, 2024

Advanced Analytics & Dashboard Design

6.1: Sourcing Open Data

I. Data Source

- **Data Sourcing:** Boat Sales (kaggle.com)
- **Data Collection Method(s):** Web scraping (assumed, contacted dataset owner for confirmation)
- **Data Contents:** This data set is about the sales data of boats. 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 look at the recent data and get some insights.
- **Data Relevance:** Most of the dataset contains usable specifications, time series, and geospatial information on boat sales, with a limitation of 14% nulls on some variables. It can be cleaned and prepared to showcase marketable insights for the website owners.

Why was the data set chosen?

Initially, I have a passion for yachts and boats. I previously worked for marine line cutters and have visited the major ports and docks of South Florida for installations, making the industry feel close to home. Cleaning data will be a significant aspect of my role as a data analyst. Therefore, I believe this dataset will serve as a good showcase for cleaning dirty data and illustrating the techniques I use to ensure a smooth analysis process.

Keanu Gomes
A6.1 Sourcing Open Data
January 23, 2024

II. Data Profile

Limitations and ethics

The location variable posed the most significant challenge in this analysis when it came to cleaning. Decoding and splitting the locations proved difficult due to potential lack of formatting requirements for boat sellers' location submissions on the yacht and boat website.

This led to formatting issues during the cleaning process, as various symbols like dashes and commas were used to separate location values, possibly due to the absence of specific guidelines provided by the boat website.

Another potential source of dirty data could be the scraping process. Additionally, the Location column contained personally identifiable information, such as numbers and names. There were also instances of non-location entries, stating phrases like "View by appointment only" in various languages.

Further complicating the cleaning of the Location columns post-split was the presence of duplicated countries, countries written in native languages, countries with missing variables or formatted by alphanumeric code, zip codes, phone numbers, personal names and business names.

III. Questions to explore.

From data brief:

1. Characteristics of the most viewed boat listings in the last 7 days
2. Is it the most expensive boats that get the most views?
3. Are there common features among the most viewed boats?

My questions:

1. What is the distribution of boat types in the sales data set, and are there any specific types that stand out in terms of popularity or demand?
2. How does the year-built impact the number of views? Is there a correlation between newer boats and increased attention from potential buyers?

3. Are there any noticeable trends or patterns in boat sales based on the specifications provided in the data set? For example, do boats with specific features tend to attract more views?
4. Could we explore the relationship between boat dimensions (length, width) and the number of views? Are there certain size ranges that are more appealing to potential buyers?

5. What is the influence of the condition of the boat (both in terms of year built condition and physical condition) on the number of views? Do well-maintained or newer boats tend to garner more interest?

6. Is there a correlation between the material of the boat and its popularity in terms of views? For instance, are boats made of a particular material more sought after in the market?

7. How does the pricing of boats affect their visibility? Are there optimal price ranges that attract more views, and does the currency used play a role in this?

8. Can you identify any geographical trends by analyzing the data based on the country, state/region, or specific locations where the boats are listed? Are there regions with higher or lower demand?

Hypothesis:

Given the current market trends, it is hypothesized that boats with newer year-built conditions, competitive pricing, and certain specifications may experience higher views in the last 7 days. This hypothesis could guide further analysis and investigation into specific factors influencing boat visibility and market trends.