

Case Study Project: Boat Sales for Nearly New Nautical

By John Feng

Presentation for DataCamp's Data Analyst Certification

Introduction to the problem

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. The Head of Marketing has laid out an ambitious goal of increasing the number of readers by 75% this year.

They would like you to take a look at the recent data and help them learn more about the characteristics of the most viewed boat listings in the last 7 days - is it the most expensive boats that get the most views? Are there common features among the most viewed boats?

Motivation and Purpose

Which boats are most popular (according to the data)?

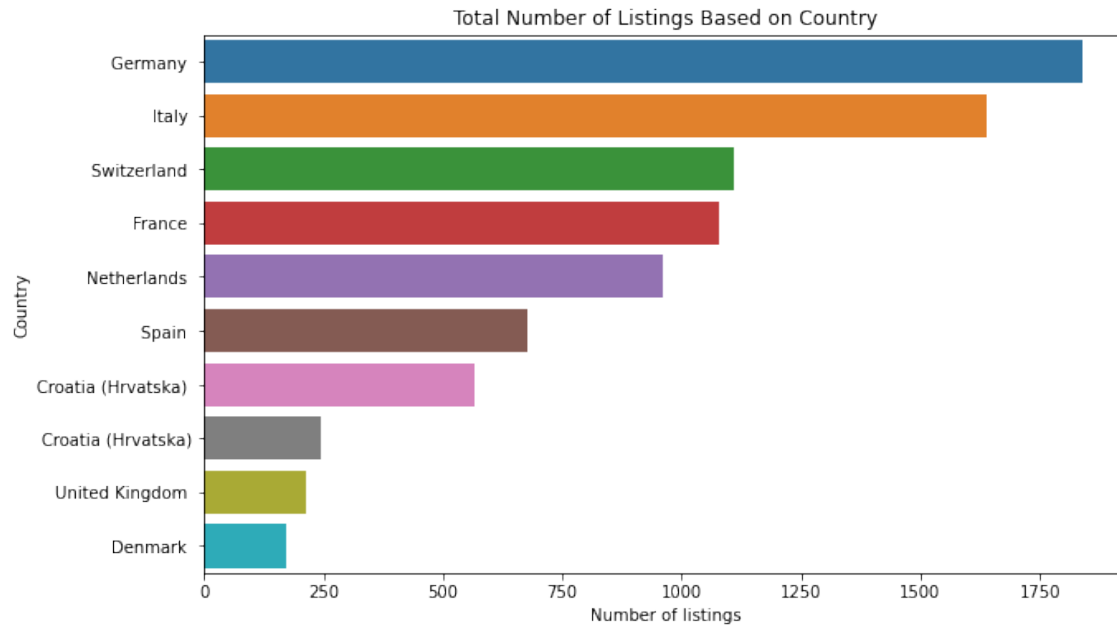
i.e. What kind of boats should be put in the weekly newsletter?

How?

- I analyze the dataset of boat listings from Nearly New Nautical, a website that allows users to post their used boats for sale.
- Primarily look at 'Number of views last 7 days' data as a measure of popularity.
- Look for trends and insights from other features, such as: Price, Boat Type, etc.

	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 » Lake Geneva » VÃ©senaz	226
1	EUR 3490	Center console boat	Terhi power boats	new boat from stock	2020	4.00	1.50	Thermoplastic	Germany » BÃ¼ttingstedt	75
2	CHF 3770	Sport Boat	Marine power boats	new boat from stock	0	3.69	1.42	Aluminium	Switzerland » Lake of Zurich » StÃ¶fa ZH	124
3	DKK 25900	Sport Boat	Pioner power boats	new boat from stock	2020	3.00	1.00	NaN	Denmark » Svendborg	64
4	EUR 3399	Fishing Boat	Linder power boats	new boat from stock	2019	3.55	1.46	Aluminium	Germany » Bayern » MÃ¼nchen	58

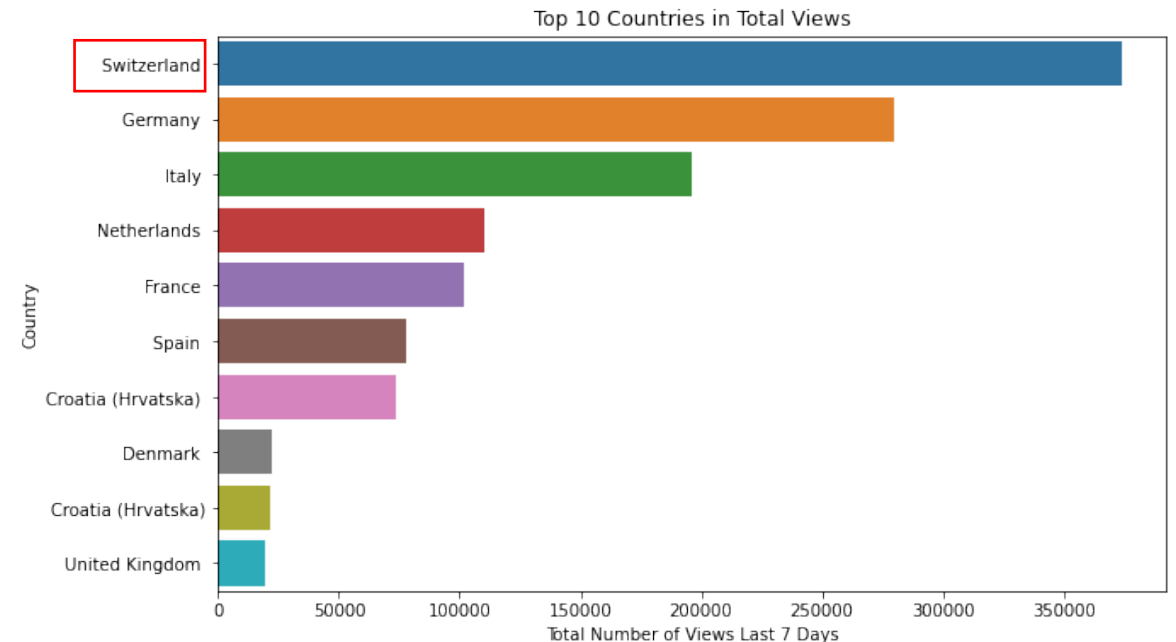
Where are boats mostly located?



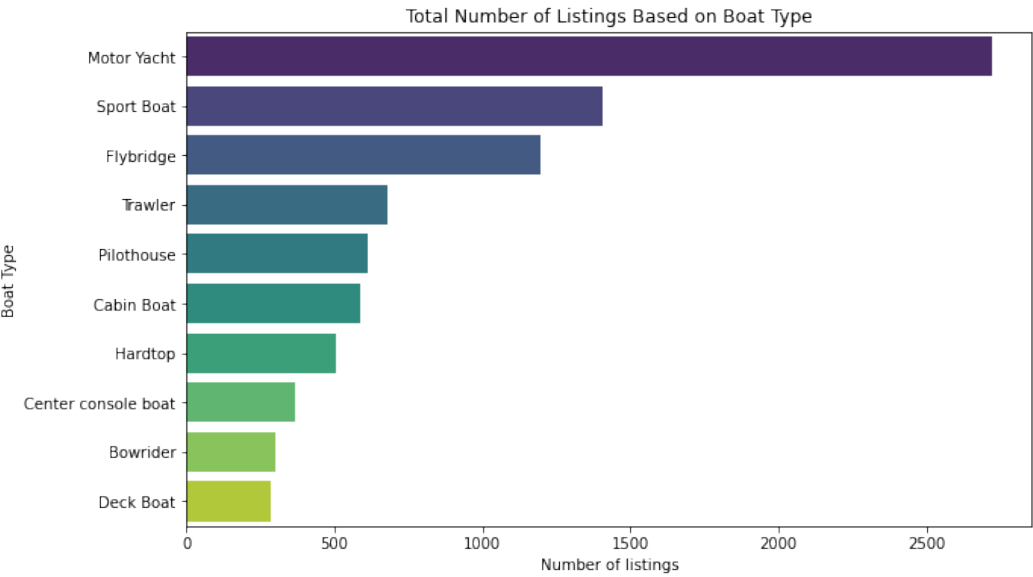
However, boats listed in Switzerland have way more total views!

Listings are located in 142 Distinct Countries for 9888 total listings

Country	Count	Frequency (%)
Germany	1839	18.6%
Italy	1638	16.6%
Switzerland	1109	11.2%
France	1077	10.9%
Netherlands	959	9.7%



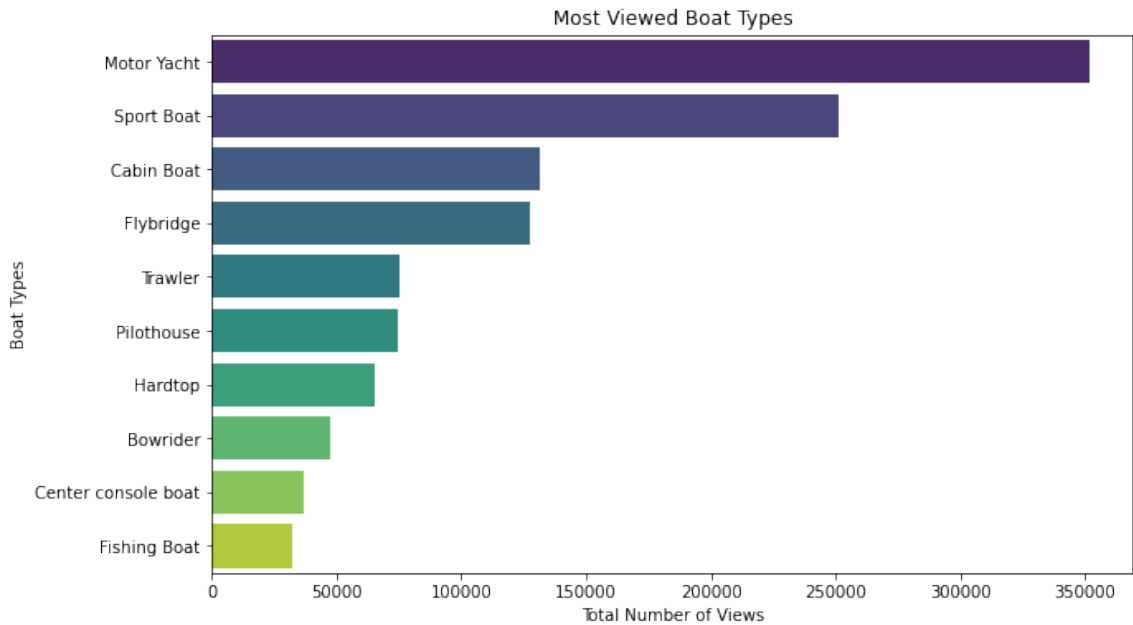
Most Popular Boat Types



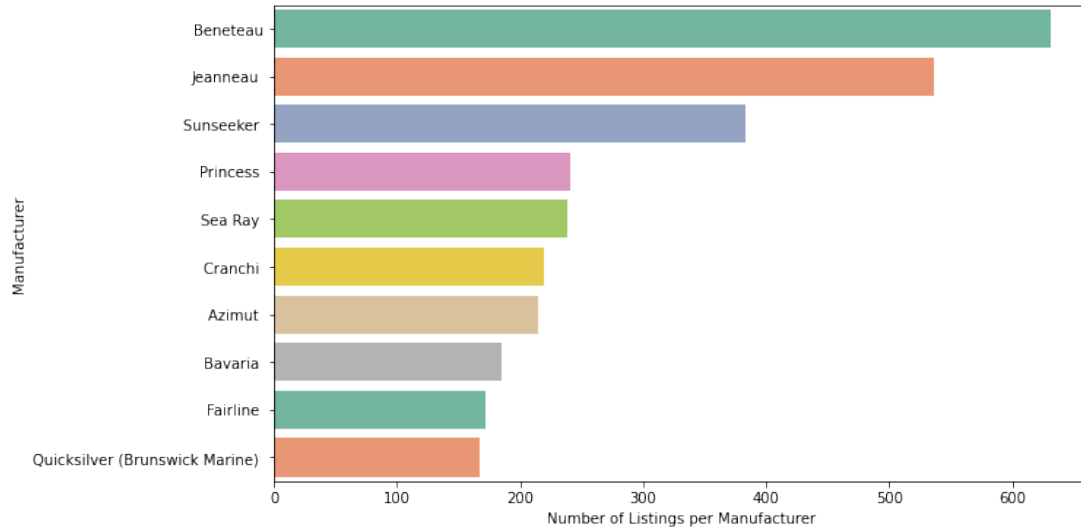
Most listed boat types also viewed more frequently.

126 Different Boat Types listed

Boat Type	Count	Frequency (%)
Motor Yacht	2720	27.5%
Sport Boat	1407	14.2%
Flybridge	1194	12.1%
Trawler	679	6.9%
Pilothouse	613	6.2%



Is there a preference for boat manufacturers?



910 Different Manufacturers for Boats

Manufacturer	Count	Frequency (%)
--------------	-------	---------------

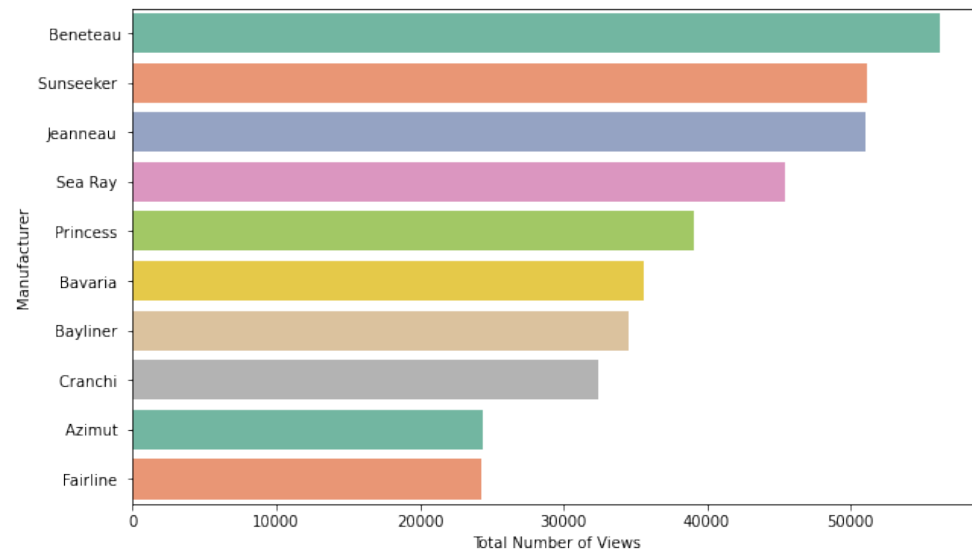
Beneteau	631	6.4%
----------	-----	------

Jeanneau	537	5.4%
----------	-----	------

Sunseeker	383	3.9%
-----------	-----	------

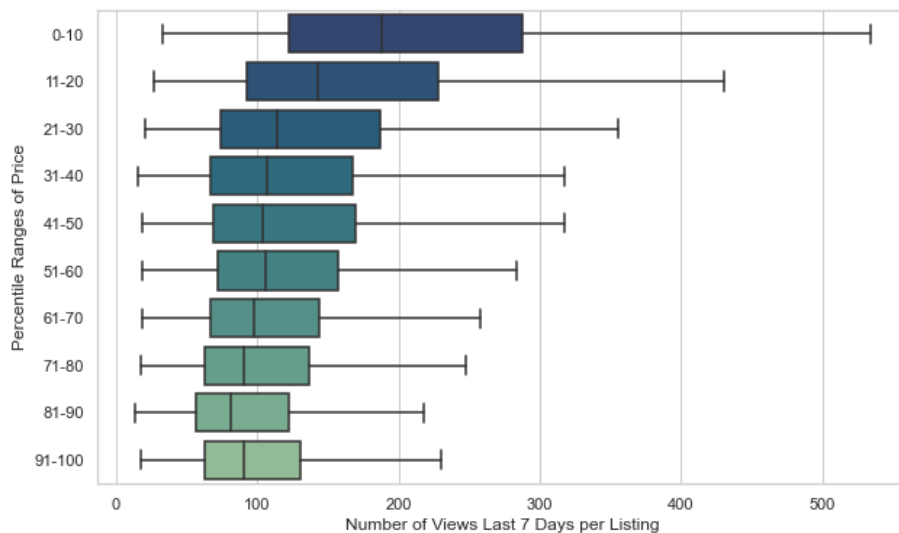
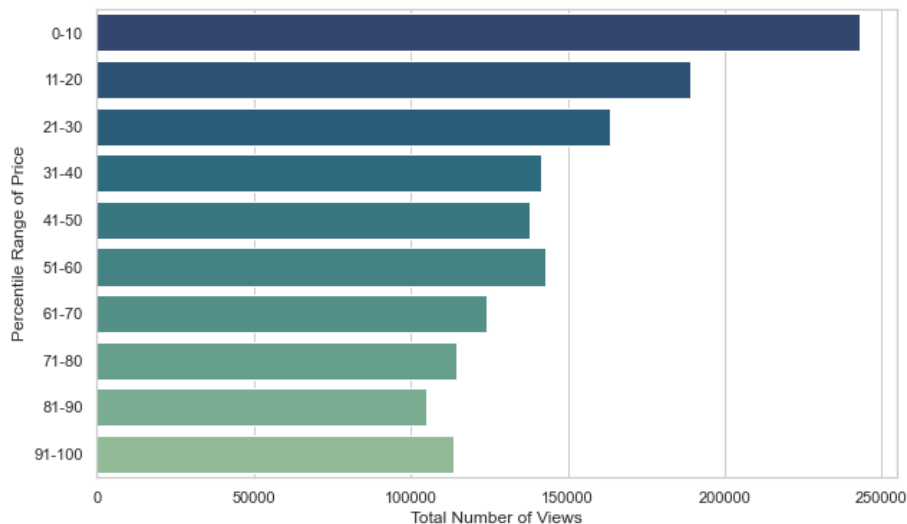
Princess	241	2.4%
----------	-----	------

Sea Ray	239	2.4%
---------	-----	------



Top boat manufacturers make up a small fraction. There is a much larger number of manufacturers.

Expensive vs Cheap Boats

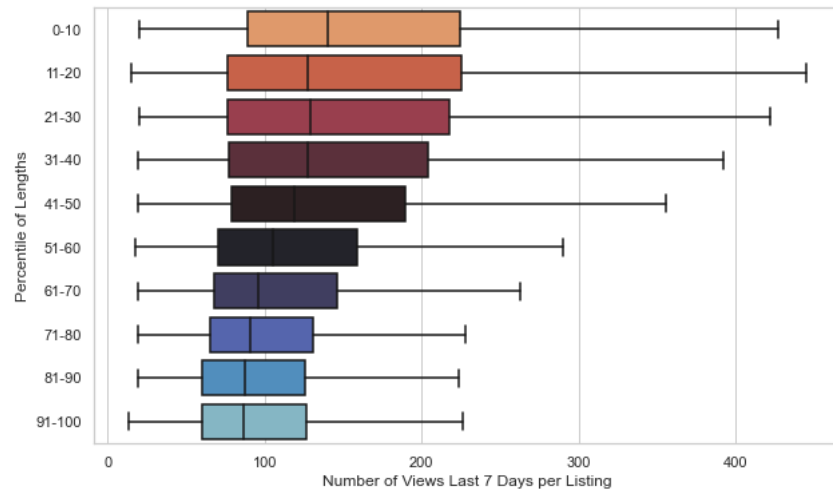
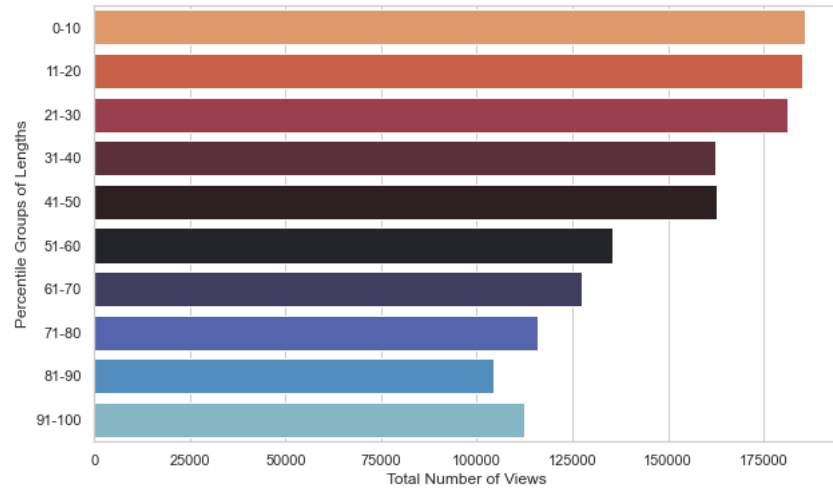


Price bins of percentile groups

Percentile Range	Min Price (Euro)	Max Price (Euro)	# of Listings in Price Range
0-10	3,223.0	21,078.9	989
11-20	21,078.9	35,147.0	989
21-30	35,147.0	49,900.0	1008
31-40	49,900.0	68,900.0	973
41-50	68,900.0	92,105.0	985
51-60	92,105.0	135,000.0	1020
61-70	135,000.0	199,000.0	1001
71-80	199,000.0	299,000.0	957
81-90	299,000.0	580,000.0	980
91-100	580,000.0	31,000,000.0	986

Cheaper boats tend to be more viewed

What about size of boats?

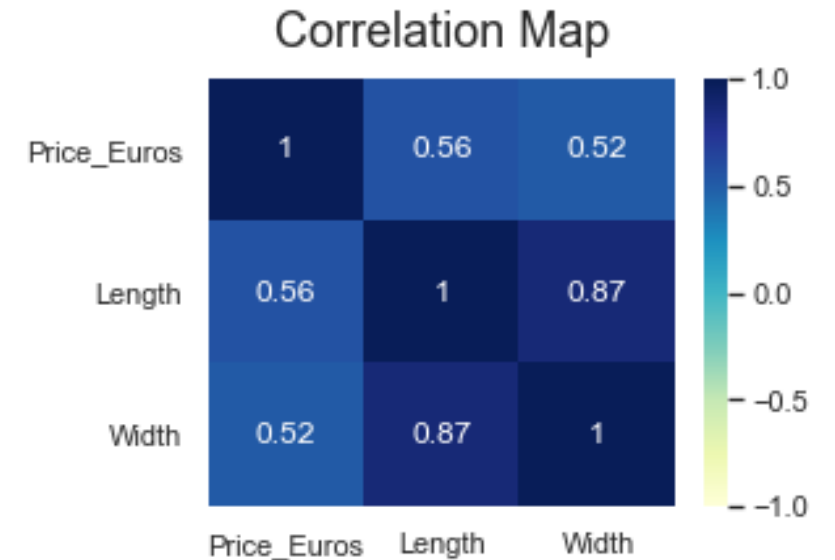


Bins of percentile groups

Percentile Range	Min Length (m)	Max Length (m)	# of Listings in Length Range
0-10	0.04	6.00	1015
11-20	6.00	6.98	962
21-30	6.98	7.92	995
31-40	7.92	9.05	980
41-50	9.05	10.28	989
51-60	10.28	11.60	996
61-70	11.60	13.11	978
71-80	13.11	14.98	995
81-90	14.98	18.80	983
91-100	18.80	100.00	986

Smaller boats are also viewed more

Strong correlations between Price, Length, and Width



Conclusion of the project

1. **Germany, Switzerland, Italy, France, Netherlands** are countries with the most boats, making up more than 50% of all listings. Therefore the newsletter should be catered for these regions.
2. **Motor yachts** and **sport boats** are the most popular amongst boat types.
3. **Most** popular manufacturers are **Beneteau, Jeanneau** and **Sunseeker**.
4. **Cheaper** and **smaller** boats tend to get more views.

Caveats:

- Number of views in last 7 days may not be an accurate measure of popularity for all listings. It's possible that a very popular boat gets sold in less than 7 days, thus having a lower view count, therefore under counting the views.
- Cheaper boats get more views because they are more affordable and more frequently purchased. However, more expensive boats are more representative of the latest boat technology and design, so its possible they will attract more readers in a newsletter.
- Other unknown factors may have a strong influence on views, and hence popularity, but is not recorded in the dataset. e.g. mileage of a boat

Technical details

- **Everything was done in Python**

- Pandas and NumPy for data cleaning and data wrangling
- Matplotlib and Seaborn for data visualization

- **Initial data cleaning and wrangling**

- The price column was in string format and in different currencies
 - Created a new column that has all the prices in Euros and in integer format
- Year column had 0 for NaN values
- Location column had non-English characters
 - The strings were replaced, and the countries data were extracted from this column
- Boat Type had redundant values
- Price had to be separated into 10 percentile groups

Sample of the Python code used for data cleaning

```
df = pd.read_csv('boat_data.csv')

price = df['Price'].values
df['currency'] = [s.split(' ')[0] for s in price]
df['amount'] = [float(s.split(' ')[1]) for s in price]

POUND2EUR = 1.16 # pound sterling to euro
DKK2EUR = 0.13 # danish krone to euro
CHF2EUR = 0.99 # swiss franc to euro

df['Price_Euros'] = np.nan
for i, cur in enumerate(df['currency']):
    if cur == 'CHF':
        df['Price_Euros'][i] = df['amount'][i]*CHF2EUR
    elif cur == 'DKK':
        df['Price_Euros'][i] = df['amount'][i]*DKK2EUR
    elif cur == 'EUR':
        df['Price_Euros'][i] = df['amount'][i]
    else:
        df['Price_Euros'][i] = df['amount'][i]*POUND2EUR

# Extract Country values out of Location column
df['Country'] = ""
for i, l in enumerate(df['Location']):
    if not df['Location'].isna()[i]: # skip NaN values in Location columns
        country = l.split('Å»')[0]
        # print(country)
        df['Country'][i] = country

# Eliminate 'power boats' from text in Manufacturer column
for i, m in enumerate(df['Manufacturer']):
    if not df['Manufacturer'].isna()[i]: # filter for non-empty values
        df['Manufacturer'][i] = m.replace('power boats', '') # remove characters 'power boats' for each row
```

Thank you for listening

- Analysis was done with Python on Datacamp Workspace
- For more details on the analysis, data cleaning, and visualization, feel free to contact me at:
johnkitfeng@gmail.com