**Data wrangling – project report**

**Sea trades’ relationship with GDP**

**Project group members**

1. Rick Aardenburg, rag330

2. Kevin Brachthuizen, kbn285

3. Axel Ehrnrooth, aeh330

4. James Smith, jsh860

**Research question**

For this report, the considered research question is: “What is the relationship between global GDP and Marine Cargo capacity, including its relationship to global capacity of specific classes of ships (Tanker, Bulk Carrier, Container), and how has this changed over time?”.

**Data sources**

The data gathered on both GDP and cargo ships was downloaded from the United Nations Conference on Trade and Development (UNCTAD) website in the form of .csv files. More details for each specific source are given below:

GDP data URL:<https://unctadstat.unctad.org/datacentre/dataviewer/US.GDPTotal>  
 Last accessed: 22/01/2024  
Marine cargo data URL: <https://unctadstat.unctad.org/datacentre/dataviewer/US.MerchantFleet>·   
 Last accessed: 22/01/2024

**Data wrangling methods**

**Wrangling the GDP Data**

This data shows GDP Data in various forms for various economies since 1970. The data was gathered from various UN sources by UNCTAD. There are 15 columns, most of which we did not need. Because of this we dropped down to 5. After this we created a new data frame containing only the global GDP over the years called “Global\_GDP”, there was no missing data.

**Wrangling the Maritime Data.**

This data shows the number of ships and the total capacity of them per year, ranging from 1980 until 2023. The data provided by UNCTAD was gathered from 2 different sources. Before 2010 it came from Lloyd's Register Fairplay and after from Clarkson’s Research services, both of which are maritime consultancy agencies. Due to the change in data provider, we decided early on that it was worth putting a vertical marker in each graph at the year 2010, in case the change in providers was to have a noticeable impact on the data.

The data itself is formatted in a similar way to the GDP data (as the provider is the same), but with crucial differences. Each year individual economies are divided into 6 rows, these are.

1. **Total fleet** - The sum of all vessels in an economy's fleet.
2. **Oil tankers** - Vessels that are specifically designed to carry liquid cargo.
3. **Bulk carriers** - Vessels designed to take dry cargo in bulk unpacked.
4. **General cargo** - Vessels designed to take breakbulk cargo.
5. **Container ships**- Vessels designed specifically to take containers.
6. **Other types of ships** - All vessels not covered by category 2-5. This includes things like passenger vessels and specialist ships.

There were originally 23 columns in the data, a number of these were mostly empty so we dropped them. This meant we ended up with only 9 columns of useful data. For 4 of these columns data was only provided from 2011 onwards, with data before that being NaN. We decided to keep this as it was as it could prove to be useful later down the line.

The global total weights were provided in the data so we used this to create a series for the total global deadweight tons, for each category. This was then placed into a new data frame called “deadweight\_per\_year”.

The new deadweight\_per\_year data frame now contained the data we needed, but contained mostly NaN as it was copying individual rows across, not all of which had all the data. To fix this we filled all the NaN into 0, then groupby(‘Year’).sum() the DF. This gave us the data frame we needed with all the relevant info.

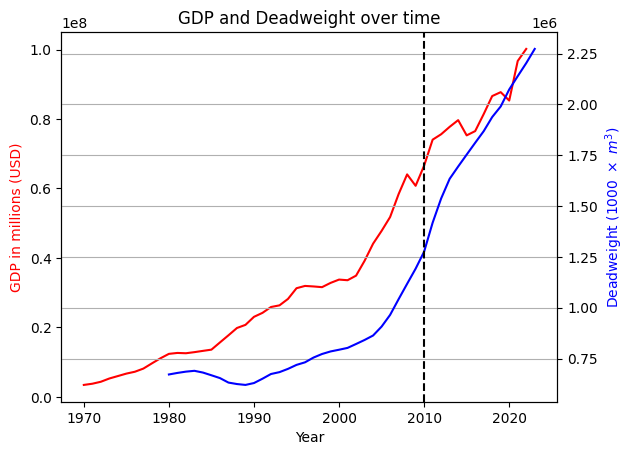
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Deadweight All** | **Deadweight Tankers** | **Deadweight Bulk** | **Deadweight General Cargo** | **Deadweight Container** | **Deadweight Other** |
| 1980 | 672142.488 | NaN | NaN | NaN | NaN | NaN |
| 1980 | NaN | 337895.557 | NaN | NaN | NaN | NaN |
| 1980 | NaN | NaN | 181880.282 | NaN | NaN | NaN |

Table 1: deadweight\_per\_year DF before formatting

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Deadweight All** | **Deadweight Tankers** | **Deadweight Bulk** | **Deadweight General Cargo** | **Deadweight Container** | **Deadweight Other** |
| 1980 | 672142.488 | 337895.557 | 181880.282 | 112840.897 | 10290.114 | 29235.638 |
| 1981 | 679704.794 | 338616.336 | 184501.243 | 114832.851 | 11060.523 | 30693.841 |
| 1982 | 686028.910 | 334237.813 | 193217.208 | 113293.837 | 12108.264 | 33171.788 |

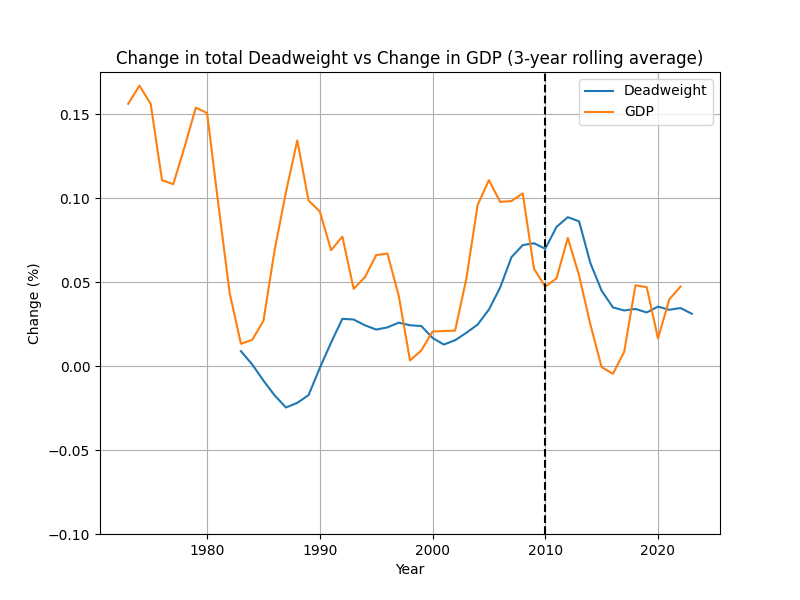
Table 2: deadweight\_per\_year DF after formatting

**Graphing the data**

We decided to keep the two data frames (Global\_GDP and deadweight\_per\_year ) separate when plotting the graphs. We did this so it was easier to use the data, avoiding confusion on which data was located where. Furthermore, we chose to plot the data using matplotlib, since we preferred the simplicity it offers when plotting. The group also considered using the seaborn library but decided against it as we preferred matplotlib’s visuals.

When looking at the plotted data in figure 1 we can see that the GDP and the total global deadweight increased in lock step. It can also be seen that deadweight is much more stable than GDP. This makes sense since a ship cannot instantaneously change or disappear. Finally, the graph shows that it is likely that GDP influences the total deadweight; The increase of deadweight is delayed with respect to that of GDP.

Figure 1: Total global GDP and deadweight over time

In order to investigate the relationship further we decided to graph the percentage change year on year, using the pct\_change() function. This was calculated for all columns then added to the relevant data frame. We compared the various maritime data sets against GDP. An important note to make about the following graphs is that the displayed change is a rolling average over 3 years. This is because we are interested mostly in trends, and not necessarily in very specific details. Taking the rolling average makes the graph more readable for this regard. However, it might remove some nuance and prevent us from making some more specific observations.

When looking at the fraction-change of both GDP and deadweight shown in figure 2, we can clearly see that from 1980 until 2000 the two are quite far apart. For example, we see quite a large increase in GDP in 1980 and 1990, but neither of these ‘spikes’ can be seen in the deadweight at any point.

From 2000 onwards we can see that the deadweight does follow the GDP. Additionally, we again see the delay in change, indicating that GDP does cause the changes in maritime deadweight, and not the other way around.

Figure 2: A rolling average of the change in global GDP and deadweight.

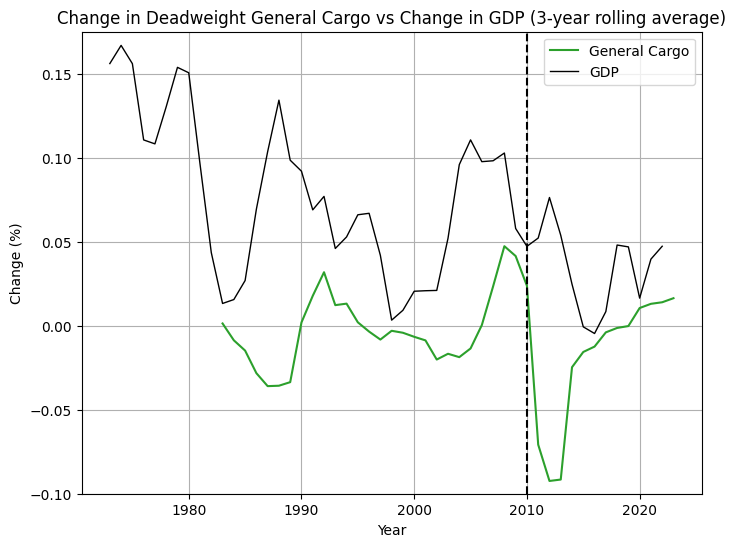
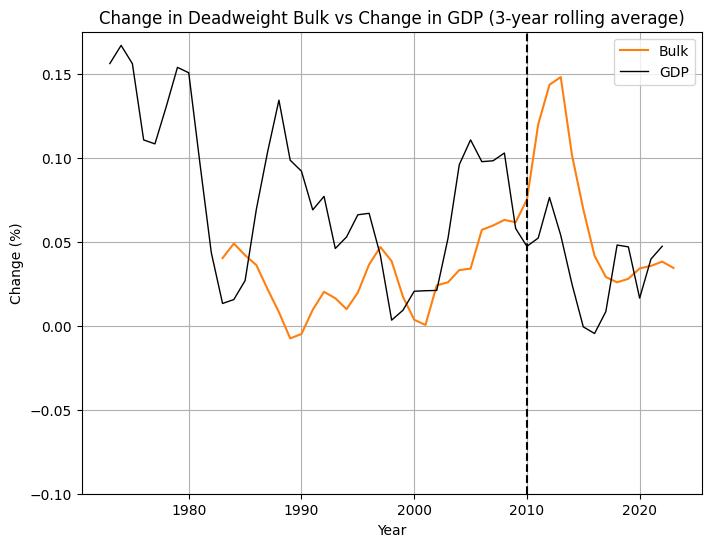
It is also interesting to break the data down into various sectors to compare how they are affected by GDP. We found that most of the observations made for the whole will also apply to the parts, but it is interesting to compare the influence of these parts on the result.

Figure 3: Rolling average of the change in GDP vs general cargo and bulk deadweight.

The first ship types, bulk and general cargo (shown in figure 3) are similar in shape, both approximately following GDP. Interesting to note is that both graphs have a very apparent peak of around 0.7 in opposing directions. The peaks being in opposite directions and about the same height indicates that this peak is likely due to the change in data provider mentioned earlier in the report. This is even further supported by the fact that this change is not in line with the change in GDP. These peaks should thus be ignored when drawing conclusions.

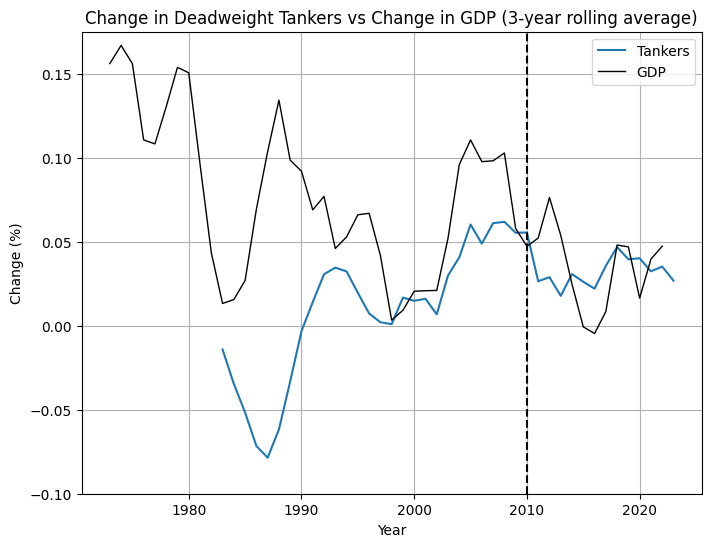
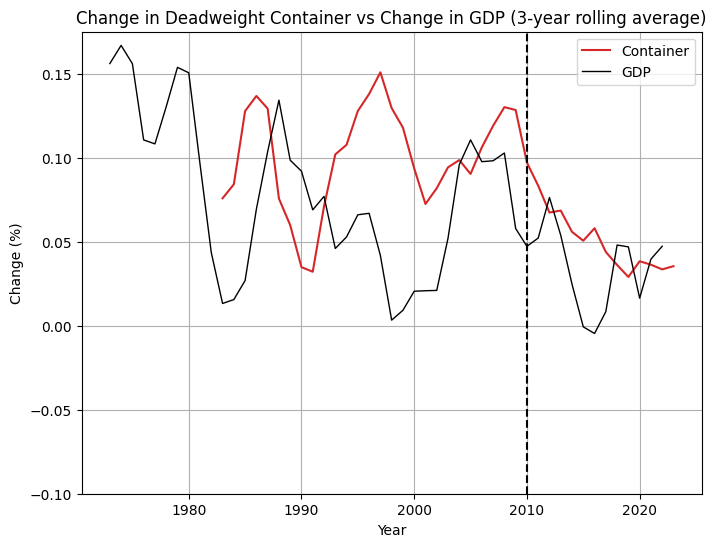
The relationship of tanker ships, shown in figure 4. We can see that this category of ships is likely the cause of the large deviation from the GDP that could be seen earlier in figure 2, since the deviation is even more extreme here. Overall, the graph looks very similar to that of the total industry. This might be an indication that this single subset of ships is a good reflection of the industry as a whole, but more research would be required to confirm such a statement.

Figure 4: Rolling average of the change in deadweight of tankers vs GDP

The final subcategory of the shipping industry is that of container ships. The change in deadweight, shown in figure 5, seems to be very dependent on the change in global GDP. In this graph, we can very clearly see that a sharp increase in GDP also leads to a relatively sharp increase in container deadweight. We can also again see the earlier observed delay in this increase, most likely due to the lead times of acquiring new ships or decommissioning them.

Figure 5: Rolling average of the change in deadweight of container ships vs GDP

**Conclusion**

There is a correlation between global GDP and Marine Cargo Capacity, meaning that the Marine Cargo Capacity follows the same trend as GDP. There is, however, usually a small delay of a few years before the Marine Cargo Capacity changes. This can be credited to the fact that it can take time for ships to be built or scrapped. Figures 3, 4, and 5 show most ship classes also follow the same pattern, but not all classes change equally. For example, the GDP has a great impact on Containers, but a smaller one on the others What can also be noted is that some ships still have a negative growth despite the positive GDP, the negative growth does become smaller at those moments in most cases, so that means there is still a positive relationship between the cargo capacity and the GDP. For the future it might be interesting to focus on specific countries and see how their GDP and Marine Cargo Capacity affect each other. This could give insight into how much different countries are affected by marine trade (think of a landlocked country versus a coastal one) and how big a part the economy plays in the country’s ability to exploit ship freight.[[1]](#footnote-1)­­

1. The code has been set up in such a way that this should be easy to implement. By changing the ‘economy’ variable data can be produced for various countries. [↑](#footnote-ref-1)