…Good data science has to be aesthetic to ensure that the person who consumes your data product draws the right conclusion. Storytelling with data is a craft where mathematics and aesthetics meet. For data science to provide value, it has to be sound, useful and aesthetic. The aesthetics of data science refers to the way the results are communicated. This article discusses an informative but messy graph about the extent of sea-ice in the Arctic. The second part describes how to improve it with the ggplot2 library.

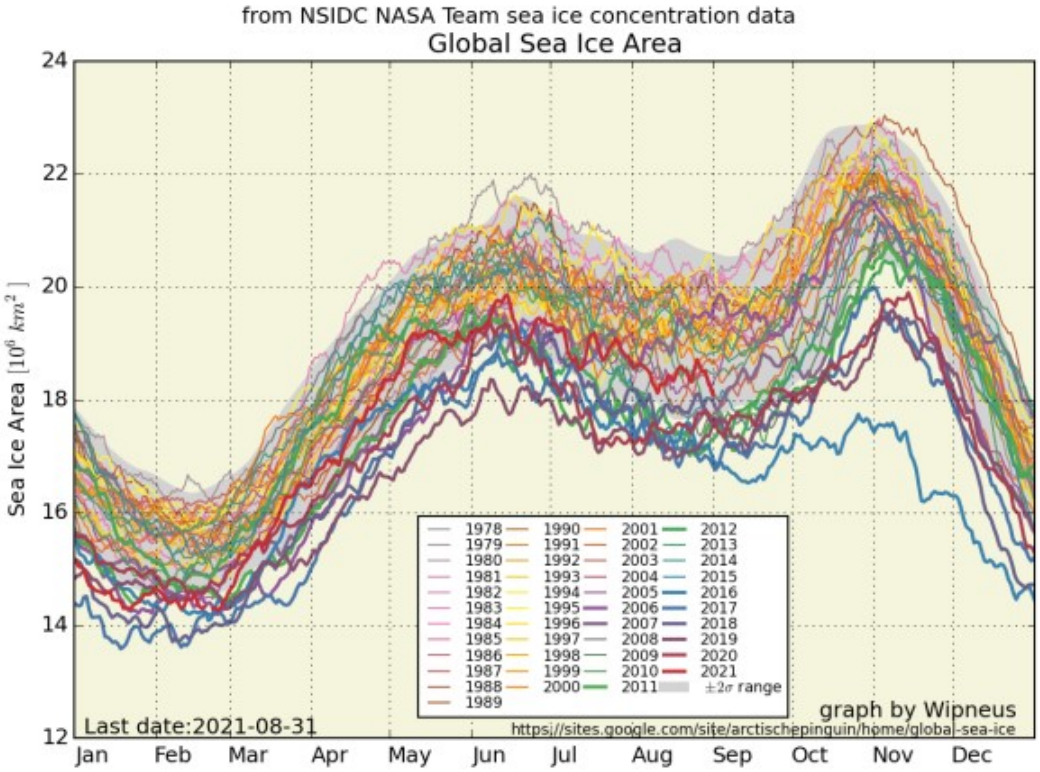
Many data scientists discuss at length which computing language they should learn. I claim that English is the most essential language an analyst can learn (or whatever other language yous peak) and the language of art. A good visualisation is a piece of data art that is composed to achieve a purpose. Whenever somebody looks at your visualisation, you want them to reach the same conclusion as you and they should be able to do so without having to dissect the information.

I subscribe to both the Data is Beautiful and Data is Ugly subreddits. While beauty might be in the eye of the beholder, aesthetic data science has to follow some rules. I subscribe to the concept that you need to maximise the data-pixel-ratio of your graphs. In other words, each pixel in your chart should ideally form part of the story. No superfluous elements, only use colour if it is part of the story. Also, the type of chart is quite important, as each geometry tells a different story. Lets put this to practice with some ggplot.

Storytelling with data case study: Receding ice sheets

A little while ago, one of the posts on the Data is Beautiful subreddit showed a disturbing graph about the changes in the extent of Arctic sea-ice from 1978 till recently. The chart is created by an anonymous citizen data scientist who calls him or herself Wipneus, which is Dutch for a snub nose and also a character in an old Dutch comic.

This graph contains all the necessary information to come to the conclusion that the total surface area of sea ice in the Arctic has reduced substantially over the past forty years. This conclusion does, however, require quite a bit of squinting and close examination of the graph.



Global Sea Ice Area in the Arctic 1978–2020.

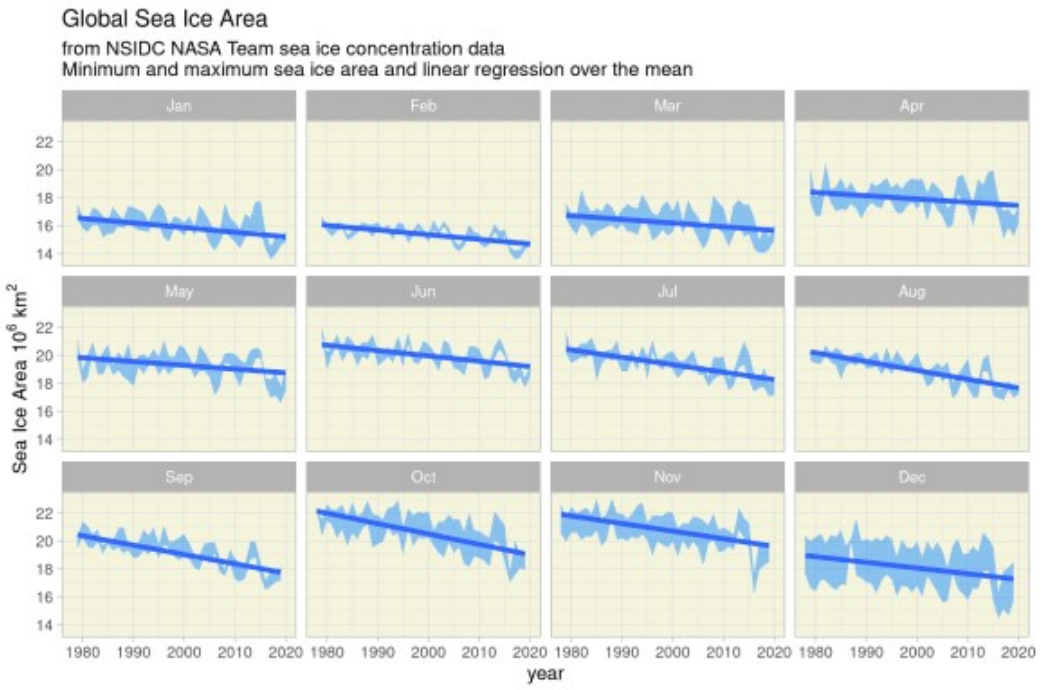
This graph is a typical multivariate time series where the colour of each line indicates a category. While this approach might work fine for one or two lines, the cacophony of colours makes it hard to distinguish which line belongs to which year. This graph would be impossible to interpret for the eight per cent of men who are colour blind. Also, the story this graph tells is confusing. While the story seems to be that ice sheets are

melting the past forty years, the chart shows the seasonal variations.

Storytelling with data and ggplot

To redesign this graph, we first need to clearly define what the story is we are telling with this data. The original chart suggests that the area of the Arctic ice sheet is receding over the past forty years. This means that our dependent variable (y-axis) is the surface area, and the independent variable (x-axis) is the year.

The original graph shows the month as the independent variable. As the ice sheet grows and recedes with the seasons, as the sinusoidal shape in the chart indicates. While this is an interesting pattern, it is not the story we want to tell with this data. To show the influence over time, the year should be the independent variable and perhaps a colour for each month. To prevent a palette with twelve colours, it is probably better to use facets for each month. The graph below is my proposed improved version. The data includes both the Arctic and Antarctic regions so we can show even more information without compromising the aesthetics.



Refactoring the ice-sheet visualisation.

Good storytelling with data requires that you first define the conclusion you want the consumer of your data product to draw. The next step is to identify the dependent and independent variables and perhaps some grouping as well. Only after you can clearly define these aspects can you choose the best geometry. There are many chart choosers available on the web that help you with this choice.

The code below loads the data from the website and transforms it into a tidy format, visualises it and saves the chart to disk.

```
  library(readr)
  library(dplyr)
  library(forcats)
  library(tidyr)
  library(ggplot2)

  url <- "https://sites.google.com/site/arctischepinguin/home/giomas/data/giomas-sumdata.csv.txt"

  antarctic <- read_csv(url, skip = 6, n_max = 40)
  arctic <- read_csv(url, skip = 49)

  sea_ice <- pivot_longer(antarctic, cols = -1, names_to = "Month", values_to =
"Area") %>%
    mutate(Location = "Antarctic") %>%
```

```
    bind_rows(
      pivot_longer(arctic, cols = -1, names_to = "Month", values_to = "Area")
%>%
      mutate(Location = "Arctic")) %>%
    mutate(Month = fct_relevel(as.factor(Month), month.abb))

  ## Facetted version - Storytelling with data
  ggplot(sea_ice, aes(Year, Area, col = Location)) +
    geom_line() +
    facet_wrap(~Month) +
    scale_color_manual(values = c("darkblue", "red")) +
    labs(title = "Global Sea Ice Area",
         subtitle = "From NSIDC NASA Team sea ice concentration data",
         y = expression(Sea~Ice~Area~10^6~km^2)) +
    theme_light(base_size = 8)

  ggsave("../../static/images/data-science/ice-data-beautiful.png", width = 6,
height = 4)
```

Strategic Data Science

If you like to know more about this topic and how to use data science to create business value, perhaps you like to read my book *Principles of Strategic Data Science*.