Question 1

Graph 1: (a) a stacked bar chart shows each group and subgroup's value.

(b) I don't think this graph is correct; at least it is confusing.

(c) The issue comes from the data values. For example, in the stacked bar of "men's basketball", the summation of "510,000", "640,000", and "750,000" is not "3.3 million". This makes the graph hard to understand.

Graph 2: (a) a line chart shows the evolution of the data w.r.t. time.

(b) This chart is good and impressionable. One concern may be that it uses 3D; nonetheless, it uses trailer trucks and farm production to leave a unique impression on readers' minds. This has a special presentation effect, especially for magazines.

(c) This chart shows the net farm income (y-axis) from 1973 to 1981 (x-axis).

Graph 3: (a) It is a bubble map that uses circles of different sizes to represent a value on a territory.

(b) This graph is good.

(c) It shows the COVID occurrence density in Canada and the USA. The comparison is apparent: Canada just has several dense occurrences, while the USA is almost all red. This graph clearly supports the claim in the title, "We need national leadership... like Canada".

Graph 4: (a) a Sankey diagram displays flows from one entity to another.

(b) I don't think such data needs to be presented using the Sankey diagram. The horizontal bar chart is better than this.

(c) A better representation is changing to a horizontal bar chart with the same color.

Graph 5: (a) a scatter plot shows the relationship between the x-axis and the y-axis variables.

(b) This graph is ugly.

(c) The issue comes from the background lines. They are too apparent, which affects the presentation of data points. Such lines should be in much lighter grey or just removed.

Graph 6: (a) It is a heat map that uses different colors to represent different values of a matrix. Such a matrix is generally the relation between the x-axis and the y-axis.

(b) This graph is not correct.

(c) The issue comes from the legend and data. This graph's data range should be in $[-1, 1]$ as it is the relationship (or correlation). Therefore, the value 1.2 in the legend should be 1, and the -2.5 and -3.0 in the graph are wrong.

Question 2

(a) Edward Tufte's principles:

- About the data-ink ratio: it is the amount of ink used for data compared to the amount of ink used for graphics. In principle, this ratio should be maximized, i.e. the decorative elements should be as simple as possible.
- About the chartjunk: it is the unnecessary decorative elements, such as 3d effects or meaningless colors. In principle, visuals should be straightforward and free of distraction.
- About the lie factor: it is defined as "size of actual change ÷ size of change on the graph." In principle, the lie factor should be close to 1 to ensure the honest representation of the data.
- About scaling: it is the misrepresentation of the scale of data. A common mistake in bar charts: each bar should have the same density, i.e., the height of a bar ÷ its actual value. Otherwise, it is misleading. This means the start of the axes should be 0.

(b) Calculate the lie factors:

Graph 1:

The percentage value in 1975 is 14%, and the percentage value in 1964 is 25%, we have $r_{\text{actral}} = (25 - 14)/25 \approx 0.36$. Meanwhile, the size of the 1964 doctor is 10cm, and the size of the 1975 doctor is 6cm; we have $r_{\text{graph}} = (10 - 6)/10 = 0.4$. Therefore, the lie factor is

$$\frac{\text{size of actual change}}{\text{size of change on graph}} \approx \frac{0.36}{0.4} = 0.9.$$

In the same way, we can calculate the lie factor from 1975 to 1990 as $0.28/0.33 \approx 0.85$.

Graph 2:

Since the lie factor of Chart A is 1, it can be seen as the true change.

From condition 1 to condition 2, the size of change on Chart B is about 10 times that of Chart A, while they share the same actual change. Therefore, If Chart A's lie factor is 1, then Chart B's lie factor is

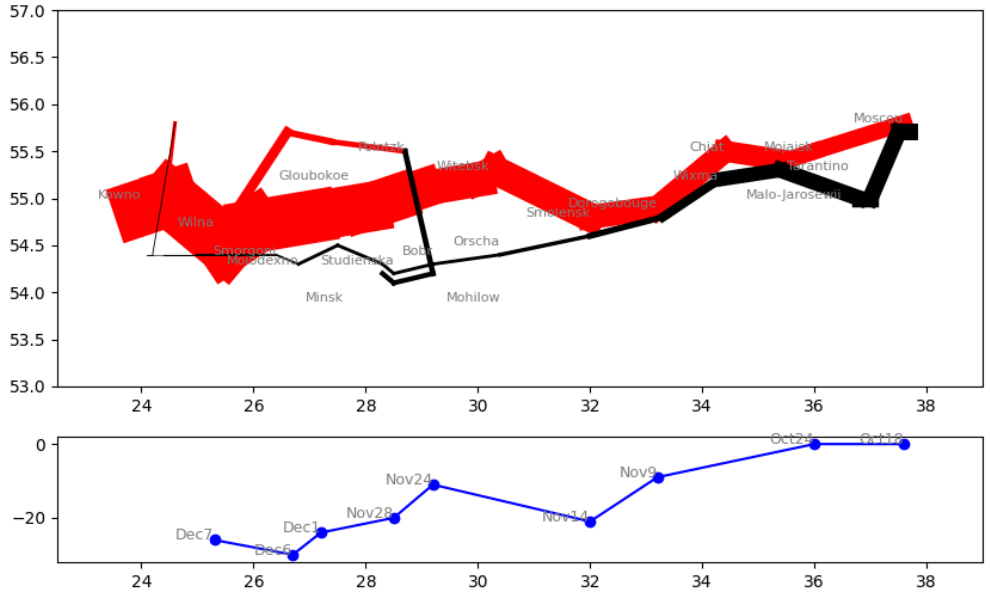
$$\frac{\text{size of change on Chart A}}{\text{size of change on Chart B}} \approx \frac{1}{10} = 0.1.$$

Question 3

(a) This graph shows Napoleon's losses during the Russian campaign of 1812. It was drawn by Charles Joseph Minard in 1869. When this graph was made, it was in the Golden Age of data visualization.

(b) The colored flow line represents *the size* of Napoleon's army at different location points. This flow also traces the *route* of the army from Poland to Moscow and back; along with it, there are *directions* and *locations* of Napoleon's army. At the bottom of the graph, the line subgraph shows the *temperature* and *retreat date*.

(c) The replicated figure is as



The code is attached along with this file.

Question 4

According to the reading material of Lin and Thorton 2022:

- *Vibrant* colors are more beautiful and popular in data visualization. (page 12)
- The most common types of charts are: bar plot, line plot, map, scatter plot, heatmap, and pie chart. (page 32)
- The most beautiful versions of the graphs were plotted with high image resolution, legible font size, Sans font type, and saturated color. (page 42)
- Graph beauty increases people's trust. Therefore, "fooled by beautiful data" means the beauty of the graph *biases* what information people tend to endorse. (page 14)