Data Visualisation Assignment 1:

Analysis of OECD Health Statistics

Khor Jing Yin (104491695)

Word Count: 1993

Table of Contents

Executive Summary

This report evaluates three visualizations from OECD Health Statistics, focusing on their effectiveness in conveying data and adherence to Edward Tufte's principles of good data visualization. Each chart is assessed for its clarity, accuracy, and ability to communicate information effectively.

- **Visualization 1**: A stacked row chart depicting work absences is criticized for its misleading legend order and complex data layering, impacting clarity. Improvements include reordering the legend, adding data labels, and enhancing gridline visibility.
- **Visualization 2**: A bar chart on female physicians' data is found to be less effective due to redundant labels and missed trend visualization. Switching to a line chart and reducing non-data ink are recommended for better trend representation.
- **Visualization 3**: A donut chart illustrating plastic production suffers from missing labels and exaggerated proportions. Adding labels and correcting scale distortions will enhance accuracy.

Overall, adhering to Tufte's principles can significantly improve data visualization clarity and effectiveness.

Introduction

Visualization is a powerful tool for representing data graphically, making it easier to identify trends, outliers, and patterns. In today's data-driven world, having clear and effective visual representations is essential for understanding complex information.

In the following document, I will examine and critique three charts from OECD Health Statistics. The evaluation focuses on the context of each visualization, adherence to data visualization guidelines, and the use of various data elements. Drawing on principles from Edward Tufte, known for his work on effective data visualization, this analysis aims to evaluate and critique these visual representations according to good data visualisation principles and guidelines.

Visualisation 1

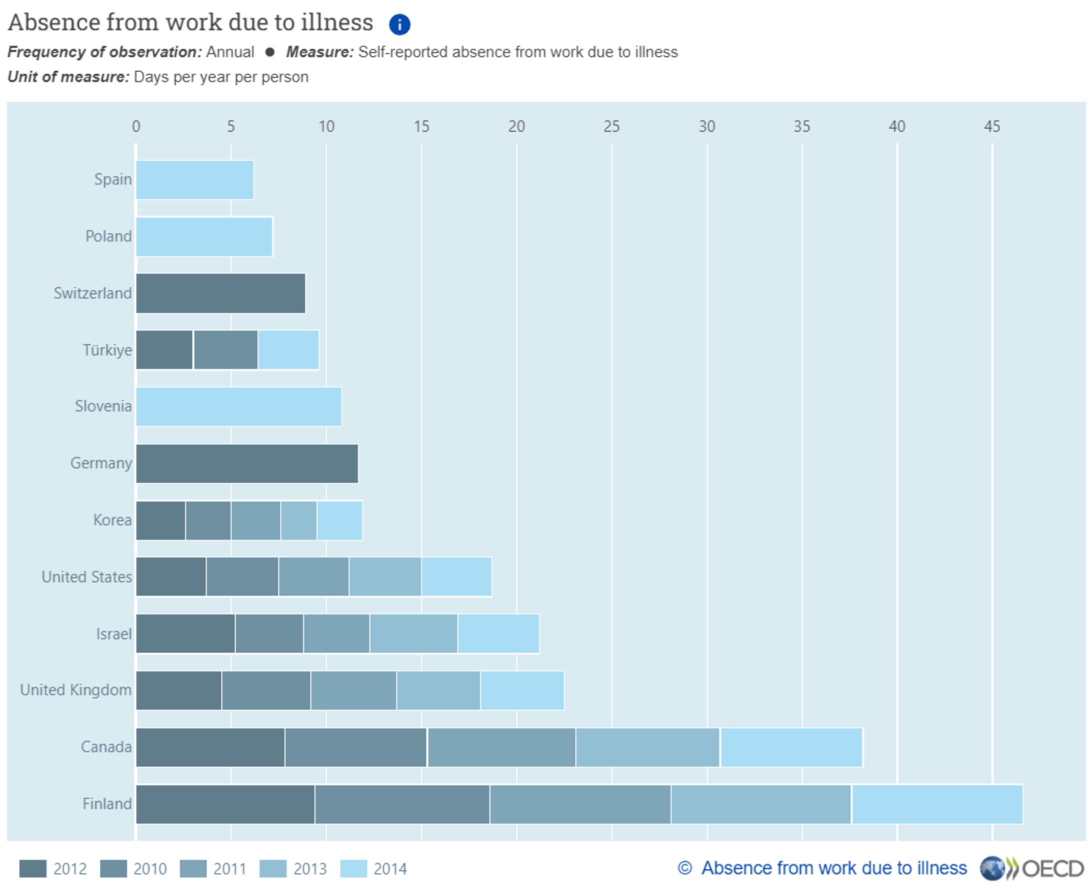


Figure 1: Absence from work due to illness

Context

The visualization in Figure 1 illustrates the self-reported work absences due to illness across various countries for the years 2010 to 2014. A stacked row chart is employed to represent this data, with the x-axis denoting the average number of days absent per person and the y-axis listing the countries under comparison.

The stacked row chart is used to simultaneous depict the average absenteeism for each year, as well as how each year's data contributes to the cumulative total across 5 years. Additionally, the chart also includes a descriptive heading that clarifies what is being measured and the unit of measurement.

The straightforward design and simple language of the chart suggests that it is merely intended for a general audience, rather than an expert one.

Critique

Upon reviewing the legend at the bottom of the chart, I noticed that while the colours are arranged from darkest to lightest shades, the years are not presented in a logical sequence. The years start with 2012, followed by 2010 and 2011, which can be misleading for viewers who may assume that the colour gradient corresponds to a chronological order. This inconsistency violates Tufte's principle of 'Show the Data', as it impairs the clarity and accuracy of the information being presented.

Additionally, I observed that the stacked format of the chart makes it challenging for viewers to accurately interpret the values for the later years. Since the data is layered, viewers must mentally subtract the previous segments to determine the values for the later years. This increases cognitive load and detracts from the chart's clarity, violating Tufte's principle of 'Layering and Separation'. Effective layering should present data in a way that enhances understanding without requiring complex mental calculations from the viewer (Andrewtk, 2020).

Lastly, the pale blue background of the graphic diminishes the visibility of the white gridlines, making it difficult to follow them. The reduced visibility of the gridlines hampers the readability of the data, further complicating the process of reading and interpreting the data values accurately.

Recommendations for Improvement

Overall, I believe that the use of the stacked row chart is effective in conveying the intended information.

To enhance its clarity based on Tufte's principles, I would recommend several improvements. First, I would reorder the legend so that the colour gradient corresponds to the years in ascending order. This would make it easier to interpret the timeline.

Additionally, I would add labels within each segment to indicate their values, as this would improve readability and make it easier for viewers to quickly grasp the values. Adding a total value for all five years at the end of each bar would also provide a clear summary of the cumulative data. With these changes, the x-axis could be removed entirely, which would also address the issue of gridline visibility (Atlassian, n.d.).

However, if the x-axis needs to remain, I would use a slightly darker blue for the background to enhance contrast and make the gridlines more visible. These adjustments should improve the chart's readability and adherence to effective data visualization principles (Scientist, 2024).

Visualisation 2

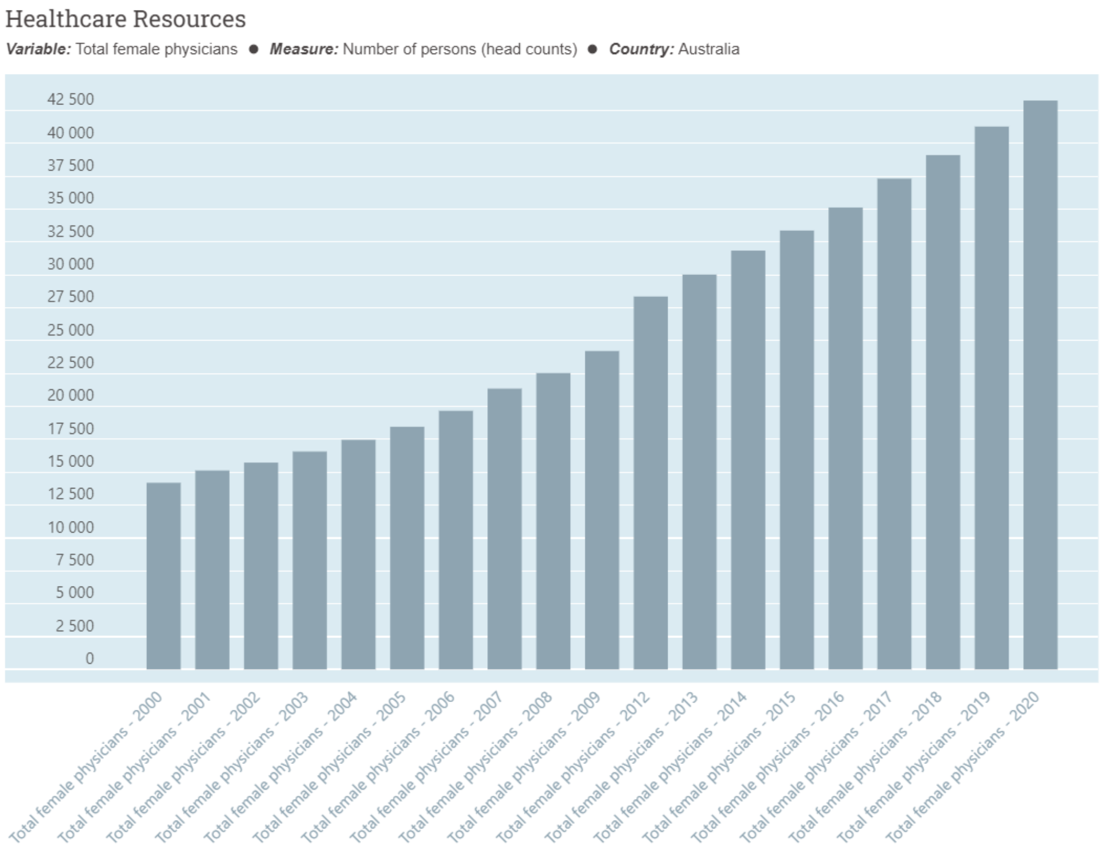


Figure 2: Number of female physicians in Australia

Context

The visualization in Figure 2 illustrates the total number of female physicians in Australia from the year 2000 to 2020. The chart uses a bar format, where the x-axis denotes the year and the y-axis represents the headcount of female physicians.

The chart was sourced from a dataset used in a paper discussing gender equality and economic growth, and it is intended for a general audience. This type of audience typically benefits from clear, accessible visualizations that effectively communicate trends and patterns without overwhelming detail or complexity.

Critique

Upon review, one of the primary issues I identified with the chart is the redundant x-axis labels, which repeatedly state "total female physicians" with only the year varying across the axis. This redundancy adds unnecessary clutter to the chart, violating Edward Tufte's principle of minimizing 'chart junk'. By streamlining these labels, we could significantly reduce the cognitive load on the viewer and make the chart easier to interpret.

Moving on, another issue relates to the type of chart used. I observed that the x-axis represents an ordinal attribute (years), while the y-axis represents a quantitative attribute (number of female physicians). In cases like this, where temporal trends are involved, a line chart is typically more suitable because it connects data points, visually emphasizing trends and the continuity of data over time. A bar chart, by contrast, is better suited for comparing discrete categories, as it segments data without implying any order between categories (Tan, n.d.).

Although the use of a bar chart here is not incorrect, it misses an opportunity to highlight trends in the data. Bar charts do effectively display the magnitude of the data and enable easy comparison between individual years. However, by not connecting the data points, the bar chart does not convey the sense of progression or change as clearly as a line chart would. Thus, in my opinion, while the current chart serves its purpose, switching to a line chart might offer a clearer view of the trends and variations across the two decades.

Finally, I believe the chart could be improved in terms of its data-ink ratio. According to Tufte's principles, most of the ink used in a graphic should be "data ink," which is ink that represents meaningful or actual data. In the present chart, the background occupies a substantial amount of non-data ink relative to the overall graphic. By reducing non-data ink and increasing data density, the chart would better adhere to best practices for data visualization, enhancing clarity and focusing more effectively on the data presented.

Recommendations for Improvement

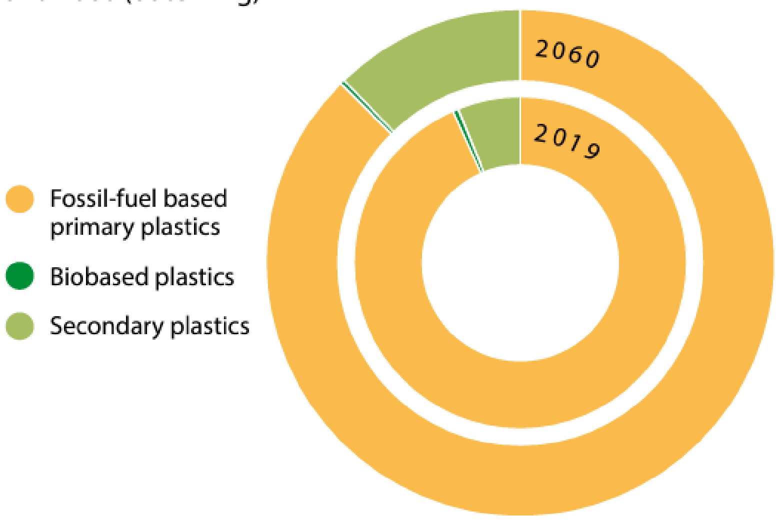
In summary, I suggest that the data in Figure 2 would be better communicated through a line chart, which would more clearly illustrate the trends and variations over time. However, if we retain the original bar chart, several adjustments could be made to improve its clarity and overall effectiveness.

First, to address the issue of the data-ink ratio, reducing non-data ink is essential. This could be achieved by subtracting ink from the horizontal scale on the columns, thereby increasing the emphasis on the data itself. Additionally, simplifying the x-axis labels to display only the year would further reduce unnecessary clutter and improve readability. These changes would lead to a cleaner, more focused visualization that communicates the information more effectively.

Visualisation 3

Figure 2. Plastic production is dominated by fossil-fuel based primary plastics

Plastics production shares by type of plastics, 2019 (inner ring) and 2060 (outer ring)



Source: *Global Plastics Outlook: Policy Scenarios to 2060* (OECD, 2022).

Figure 3: Plastics production share by type of plastics

Context

The visualization in Figure 3 depicts the distribution of plastic production by type for 2019, along with projections for 2060. A donut chart is employed to represent this data, effectively communicating the relative proportions of each plastic type.

This chart is sourced from an environmental policy report focusing on climate change mitigation and plastic pollution, and its target audience is policymakers.

As illustrated in the graphic, fossil-fuel-based primary plastics constitute the largest portion of total plastic production. The corresponding data, as referenced in the article, is summarized in the table below:

	2019	2060
Fossil-fuel based primary plastics	93%	87%
Biobased plastics	6%	12%
Secondary plastics	0.6%	0.5%

Figure 4: Plastics productions share by percentage

Critique

Upon studying the donut chart, the first issue I noticed was the absence of labels indicating the values for each proportion. This violates Tufte's principle of **maximizing data density**, which emphasizes presenting as much relevant data as possible without overwhelming the viewer. The omission of numerical values or percentages reduces the chart's ability to clearly communicate the exact distribution of plastic types, forcing viewers to estimate visually. By adding labels, the data density would be increased, making the chart more informative and allowing the audience to understand the proportions with greater precision, which is particularly important for policymakers relying on accurate information to guide decisions (ThoughtSpot, 2023).

Upon closer examination, I also noticed that for the ‘Biobased plastic’ segments, although the 2060 chart’s value (12%) is only double that of 2019 (6%), the actual graphic displays it as nearly triple in size.

$$Size\ of\ effect\ in\ data = \frac{12\% - 6\%}{6\%} = 1 = 100\%$$

$$Size\ of\ effect\ shown\ in\ graphic = \frac{3 - 1}{1} = 2 = 200\%$$

$$Lie\ factor = \frac{200\%}{100\%} = 2$$

Figure 5: Lie factor calculation

As indicated by the calculations above, the lie factor is 2, which reveals that the graphic exaggerates the data. This issue violates Tufte's principle of avoiding scale distortion, which emphasizes maintaining accurate proportional relationships between data and their visual representation. The exaggeration of the 2060 segment misleads viewers, distorting the perceived growth of biobased plastic production. This distortion undermines the integrity of the data visualization, as the visual elements no longer accurately represent their numerical values.

Recommendations for Improvement

I recommend several improvements to enhance the accuracy and effectiveness of the donut chart. First, I would add data labels directly to the chart to provide numerical values or percentages for each segment. This would increase data density, allowing viewers to quickly and precisely understand the proportions of each plastic type. Additionally, I would correct the scale distortion by ensuring that the size of each segment accurately reflects its proportion in the data. This adjustment would prevent exaggeration and ensure that the visual representation is true to the numerical values. By making these changes, the visualization will better support policymakers in making informed decisions based on accurate data.

Conclusion

In reviewing the visualizations from OECD Health Statistics, several key insights and areas for improvement have emerged. Each chart, while useful, exhibits certain weaknesses that affect its clarity and effectiveness.

Visualization 1 demonstrates how stacked row charts can convey complex data but also highlights issues such as the challenges in interpreting layered data.

Visualization 2 reveals the potential shortcomings of using bar charts for temporal data. Although the chart effectively compares discrete categories, a line chart is more recommended for illustrating trends over time.

Visualization 3 suffers from issues such as the absence of data labels and scale distortion. This highlights the importance of adding precise labels and correcting exaggerations in visual representation, especially when working with charts such as donut or pie.

Charts and visualizations are powerful tools in transforming raw data into meaningful and actionable insights. Edward Tufte's principles offer valuable guidelines for achieving this, emphasizing the importance of maximizing data density, minimizing chart junk, and maintaining accurate proportional relationships. By adhering to these principles, visualizations can better serve their intended audiences, providing a more precise and insightful understanding of the data.

References

Tan, P. S. (n.d.). *Evaluating Expressiveness and Effectiveness of Informative Charts*. Medium. https://medium.com/vitrox-publication/evaluating-expressiveness-and-effectiveness-of-informative-charts-9f0455474bf1

Andrewtk. (2020, August 28). Tufte's principles. *thedoublethink*. https://thedoublethink.com/tuftes-principles-for-visualizing-quantitative-information/

ThoughtSpot, T. (2023, November 22). *6 most useful data visualization principles for analysts*. ThoughtSpot. https://www.thoughtspot.com/data-trends/data-visualization/data-visualization-principles

Atlassian. (n.d.). *Stacked Bar Charts: A detailed breakdown | Atlassian*. https://www.atlassian.com/data/charts/stacked-bar-chart-complete-guide#:~:text=It's%20a%20good%20idea%20for,their%20sizes%20easy%20to%20read.

By. (n.d.). *The difference between line graphs and bar graphs*. Mokkup.ai. https://www.mokkup.ai/blogs/difference-between-bar-graphs-and-line-graphs/#:~:text=The%20choice%20of%20chart%20type,large%20differences%20between%20data%20points.

Scientist, S. a. D. (2024, April 5). Essential principles for Effective data Visualization. *Medium*. https://medium.com/thedeephub/essential-principles-for-effective-data-visualization-a13f05d22c39