Data set: Outbreaks in Toronto Healthcare Institutions1

The 2022 report, csv file.

(https://open.toronto.ca/dataset/outbreaks-in-toronto-healthcare-institutions/)

**Visualization 1 generated by R**

* **What software did you use to create your data visualization?**

Rstudio.

* **Who is your intended audience?**

General public.

* **What information or message are you trying to convey with your visualization?**

The first message is a basic idea of what types of outbreaks occurred in what settings and what the primary causative agents were. The second layer of information is that among these causative agents, COVID-19 was still the dominant one. At the same time, the audience may see that the number of outbreaks and distribution of the primary causative agents could vary in different settings. I didn’t intend to include too much emotion, but I would like to show that COVID-19 was still a dominant outbreak source in 2022.

* **What design principles (substantive, perceptual, aesthetic) did you consider when making your visualization? How did you apply these principles? With what elements of your plots?**

I tried to include all principles. For substantive, the data presented were not modified by me, and source of data was included to make the visualization accurate and reliable2. For aesthetic, I avoided unnecessary elements to make it look simpler and pleasing2. For perceptual, I added titles, labels, and caption around the bar graph to make it easier for the audience to understand the graph2.

* **How did you ensure that your data visualizations are reproducible? If the tool you used to make your data visualization is not reproducible, how will this impact your data visualization?**

I used Rstudio to generate the visualization with comments of my codes, so other people can run the same code to reproduce the visualization3. I also included the source of data.

* **How did you ensure that your data visualization is accessible?**

First, from the color aspect, I used the Viridis package to make the colors more friendly to individuals with colorblindness, and the contract of color can be maintained in grayscale which is friendly for printing4. Second, I included proper labels and caption to guide the audience and reduce their anxiety5. Thirdly, I included alt-text to describe all the labels and legends, as well as the main trend shown by the visualization to allow individuals with screen reader to understand the visualization6.

* **Who are the individuals and communities who might be impacted by your visualization?**

The general public can be impacted by this visualization to understand that COVID-19 still exist in our communities and there are still outbreaks in hospitals and long-term care or retirement homes where people tend to be more immunocompromised.

* **How did you choose which features of your chosen dataset to include or exclude from your visualization?**

I looked for columns that have relationship (i.e. can match to each other) and at the same time, the number of unique categories is manageable. Therefore, I didn’t present the institution name and address. Instead, I used the more generalized outbreak settings to group the outbreak types and number of cases.

* **What ‘underwater labour’ contributed to your final data visualization product?**

First, I went to the GitHub page of the opendatatoronto package to learn about how to import the data set. Second, I read different help documentations to figure out how to add alt-text and change positions/angle of my labels.

**Visualization 2 generated by Python**

* **What software did you use to create your data visualization?**

Python in Google Colab.

* **Who is your intended audience?**

General public.

* **What information or message are you trying to convey with your visualization?**

In 2022, there were different outbreaks caused by different causative agents occurring in both hospital and non-hospital settings, with COVID-19 being the dominant one. And across all causative agents, the outbreaks were slightly higher in non-hospital than hospital settings. This is an interactive visualization to allow audience to explore the values of each bar.

* **What design principles (substantive, perceptual, aesthetic) did you consider when making your visualization? How did you apply these principles? With what elements of your plots?**

I considered all principles. For substantive, although I processed the data, the steps are shown in the code. For aesthetic, I tried to make it pleasing to read by using bar graph which is a common chart type, and I arranged the labels on both axes to make the causative agent names more readible2. For perceptual, I had clear and straightforward labels and title make it easier for the audience to follow2.

* **How did you ensure that your data visualizations are reproducible? If the tool you used to make your data visualization is not reproducible, how will this impact your data visualization?**

I used Python to process data and generate the plot with comments on my codes, so other people can run the same code to reproduce the visualization3.

**How did you ensure that your data visualization is accessible?**

First, from the color aspect, the two colors I used are good for individuals with colorblindness7. Second, I included proper title and labels to help the audience understand the graph and reduce their anxiety5. Lastly, I set the typeface as arial which is easier for people with dyslexia8. I wanted to include alt-text and caption, but there is no embedded alt-text or caption function for the ‘plotly’ package (my Python is not as good as R).

* **Who are the individuals and communities who might be impacted by your visualization?**

Similar to the first visualization, the general public can be impacted by this visualization to understand that prevalence of COVID-19 in our communities and they may also develop the awareness that there could be more outbreaks in the non-hospital sites than the hospitals.

* **How did you choose which features of your chosen dataset to include or exclude from your visualization?**

I wanted to present something different from the last one, so I categorized the outbreak settings into hospital or non-hospital sites and tried to compare the cases of different outbreaks in this way. Therefore, I still used the same three columns but with more data processing to assign different outbreak settings to their corresponding categories.

* **What ‘underwater labour’ contributed to your final data visualization product?**

First, to assign outbreak settings to “hospital” or “non-hospital”, I generated another data frame with all the outbreak settings and their corresponding categories. Then I merged this new data frame and the original data. After that, I tried to separate the outbreak case count based on “hospital” and “non-hospital”, so I generated new data frames with counts for each category and eventually merged them to plot the graph. I also checked what color combinations are good for colorblindness.

References:

1. *Outbreaks in Toronto Healthcare Institutions*. City of Toronto Open Data Portal. (n.d.). Retrieved April 3, 2023, from <https://open.toronto.ca/dataset/outbreaks-in-toronto-healthcare-institutions/>
2. Slide 29. *Data visualization, Introduction and Overview.*
3. Slide 17, *Data visualization, First Steps: Reproducible Data Visualization.*
4. Slide 22, *Data visualization, Visualization with Purpose: Accessible Data Visualization.*
5. Slide 27, *Data visualization, Visualization with Purpose: Accessible Data Visualization.*
6. Slide 32, *Data visualization, Visualization with Purpose: Accessible Data Visualization.*
7. Nichols, D. (n.d.). *Coloring for colorblindness*. David Nichols. Retrieved April 3, 2023, from https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40
8. Slide 29, *Data visualization, Visualization with Purpose: Accessible Data Visualization.*