INTRODUCTION TO DATA VISUALIZATION FOR RESEARCH

Alexandra Wong

Data Visualization and Analytics Librarian,
York University

wongalex@yorku.ca

3:00 PM – Introduction to Data Visualization for Research

*with **Alexandra Wong**, Data Visualization and Analytics Librarian, Open Scholarship Department, York University*

Data visualization is a powerful method to both understand our data and to convey our research and stories to others. This session will introduce data visualization fundamentals and its different types, as well as showcase the potential of using data visualization within research methods and outputs. You will learn how to apply a simple process to get started on creating your own data visualization, from first defining its purpose, to cleaning your data, choosing the appropriate tool, to creating your visualization and sharing it! No experience with data or data visualization is necessary.

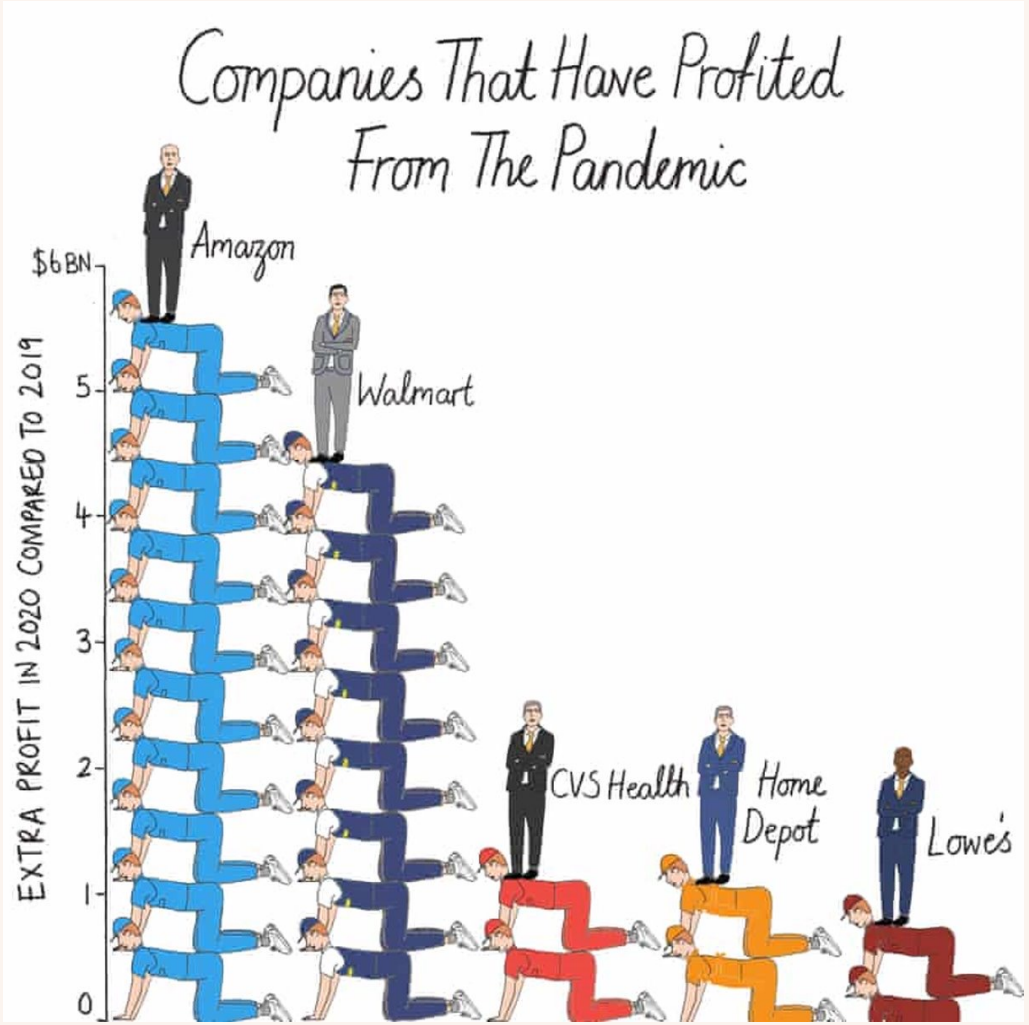
AGENDA

- INTRODUCTION
- TYPES OF DATA VISUALIZATIONS
- HOW TO CREATE A DATA VISUALIZATION
- HOW TO CREATE: AN EXAMPLE

DATA VISUALIZATION

is anything that converts data and information into a visual representation: tables, charts, graphs, maps and more.

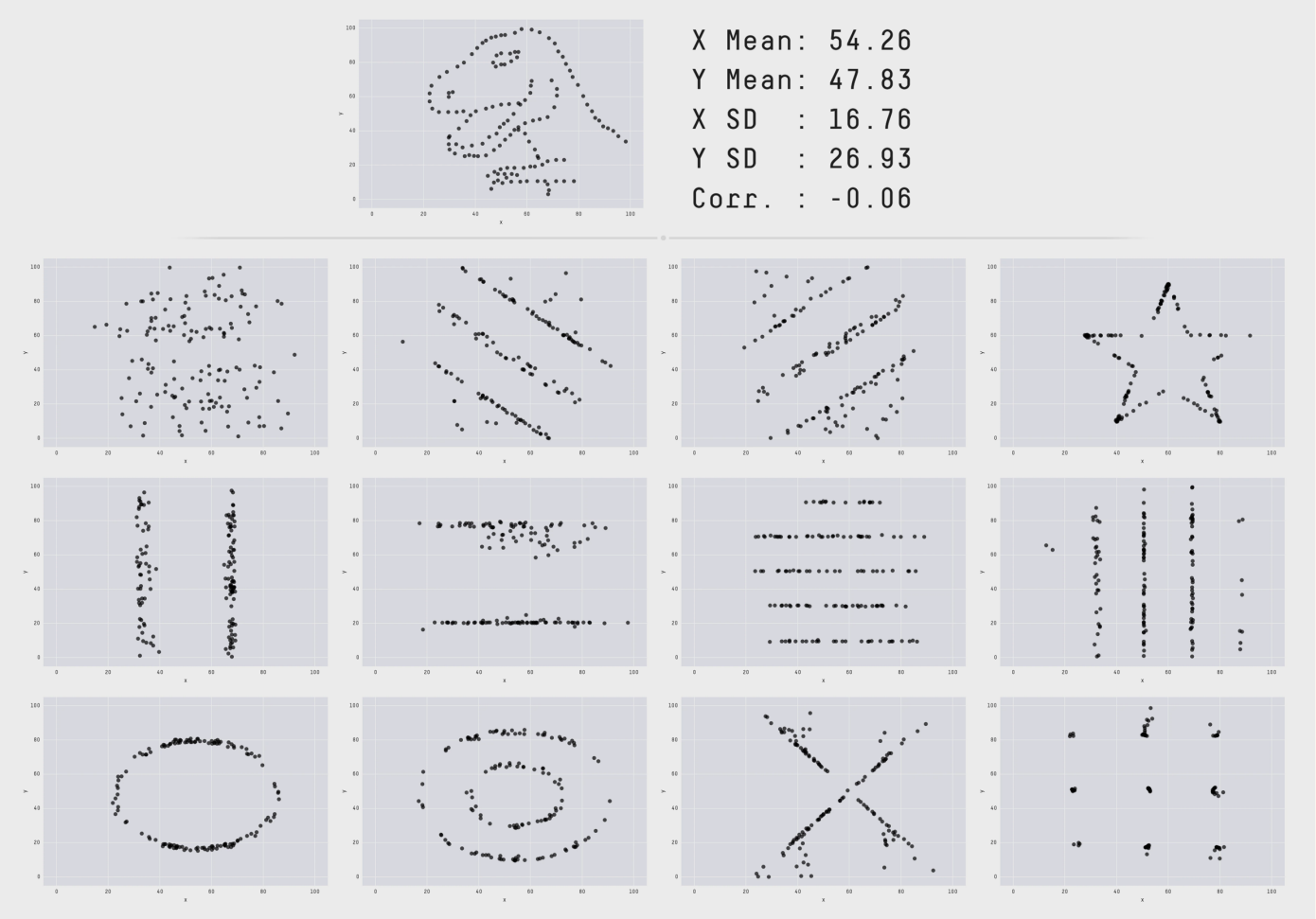
THE POWER OF VISUALIZATION: COMMUNICATING DATA



Source: Mona Chalabi

- COMMUNICATE MEMORABLE STORIES THROUGH DATA
- HIGHLIGHT NARRATIVES, PATTERNS, OUTLIERS
- EXPLAIN RESEARCH
- CREATE DATA-DRIVEN SUPPORT FOR INITIATIVES

THE POWER OF VISUALIZATION: EXPLORING DATA



- SEEING BEYOND SUMMARY STATISTICS
- SPOT PATTERNS AND OUTLIERS
- ANALYZE LARGE AMOUNTS OF DATA

Source: Autodesk Research, 2017. Retrieved from https://www.autodesk.com/research/publications/same-stats-different-graphs

LEVERAGING VISUAL PERCEPTION

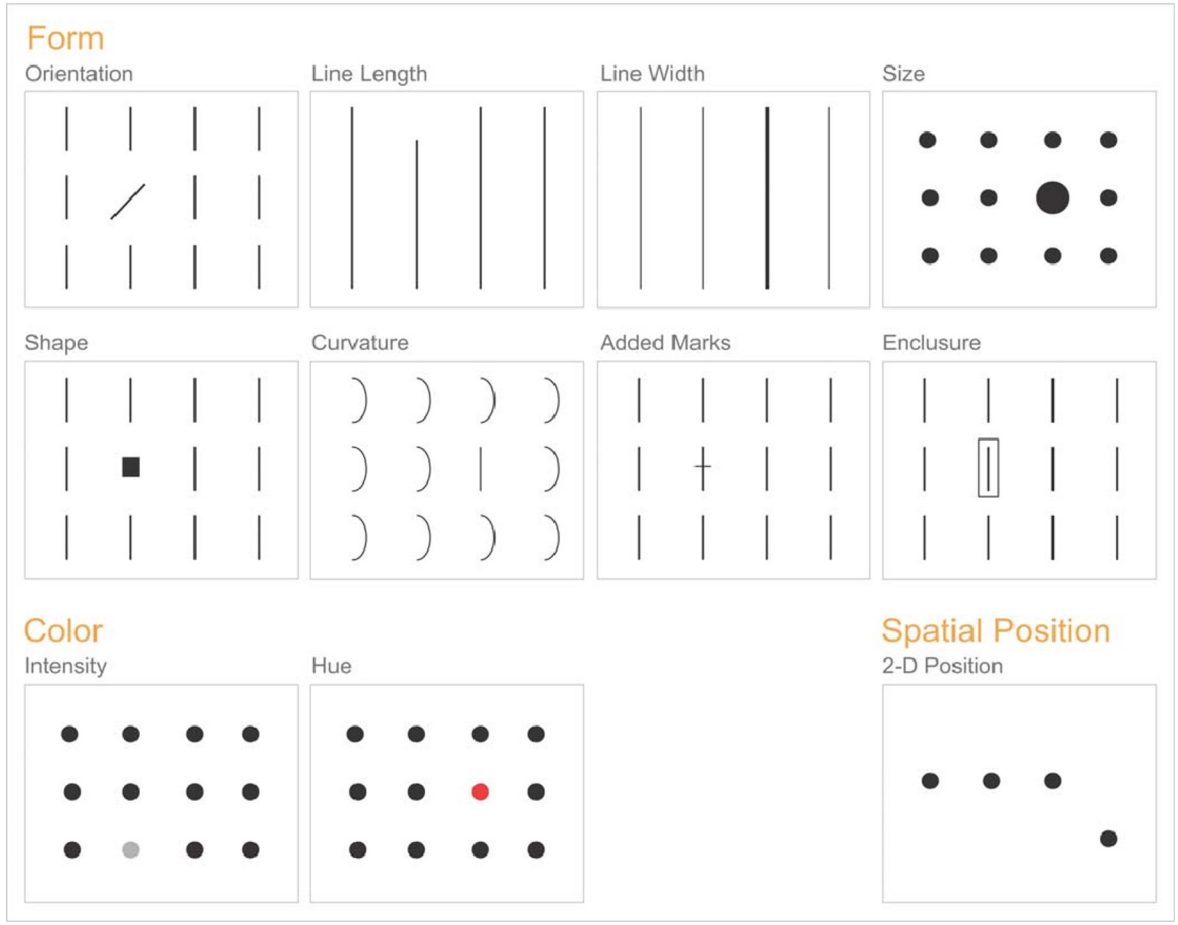
Humans naturally perceive objects as organized patterns and objects

9873497902756479028947286240
8032080290073025012702370083
2478026027037937757097073779
9279797097230972309795927509

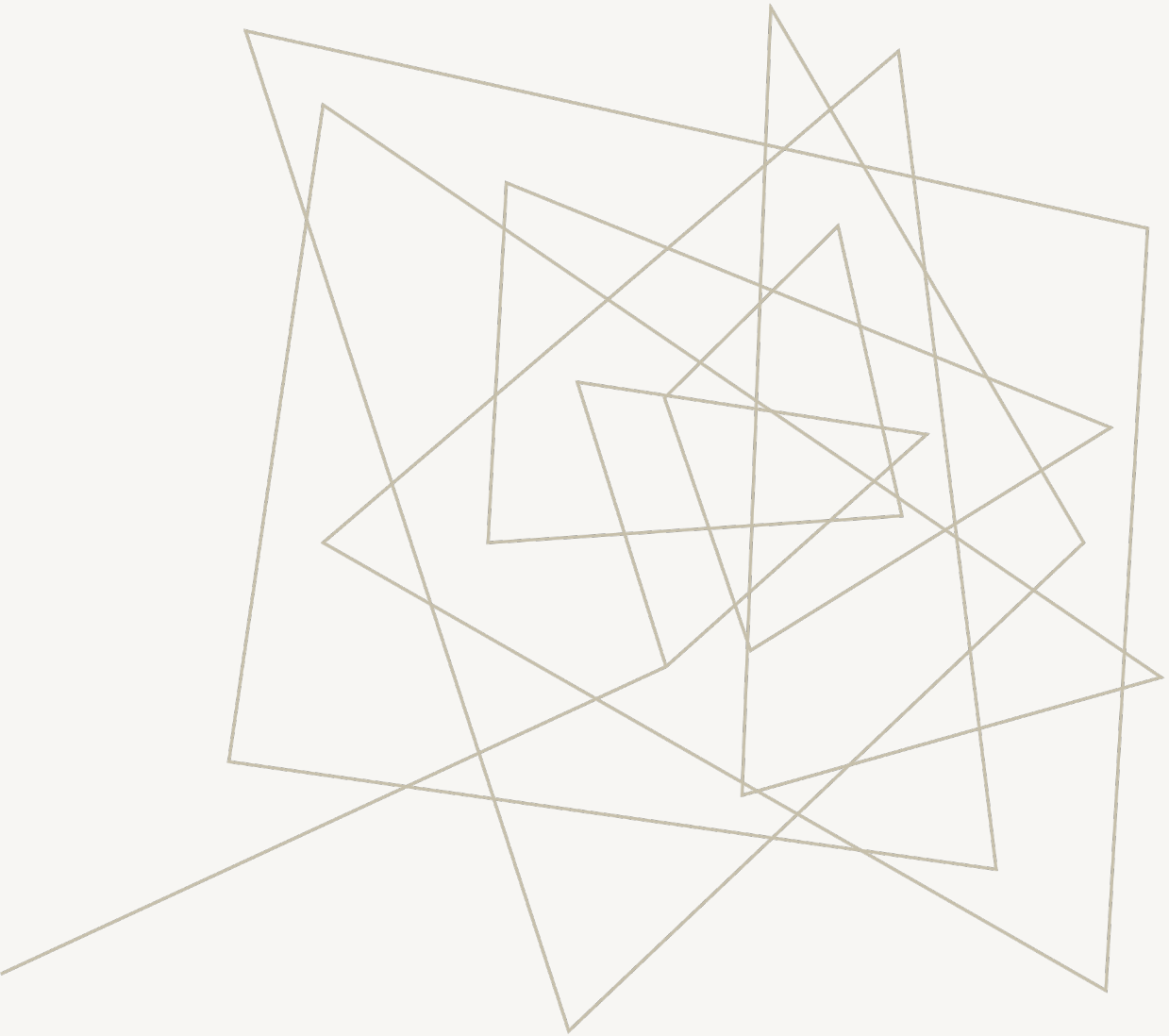
Attentive processing

9873497902756479028947286240
8032080290073025012702370083
2478026027037937757097073779
9279797097230972309795927509

Pre-attentive processing



Stephen Few, "Tapping the Power of Visual Perception"



TYPES OF DATA VISUALIZATION

TYPES OF DATA VISUALIZATIONS

Well-known type and an efficient display of data, but not easy to decode the data.

Tables

Charts

Graphs

Networks

Maps

Infographics

Dashboards

Table 6: Length of time for experience working in an academic library as a qualification.

Academic library experience length of time	***n***
Length not specified	13
1 - 2 years	10
3 - 5 years	3
5 + years	1

Source: Thielen & Marsolek, 2022.

TYPES OF DATA VISUALIZATIONS

Tables

Charts

Graphs

Networks

Maps

Infographics

Dashboards

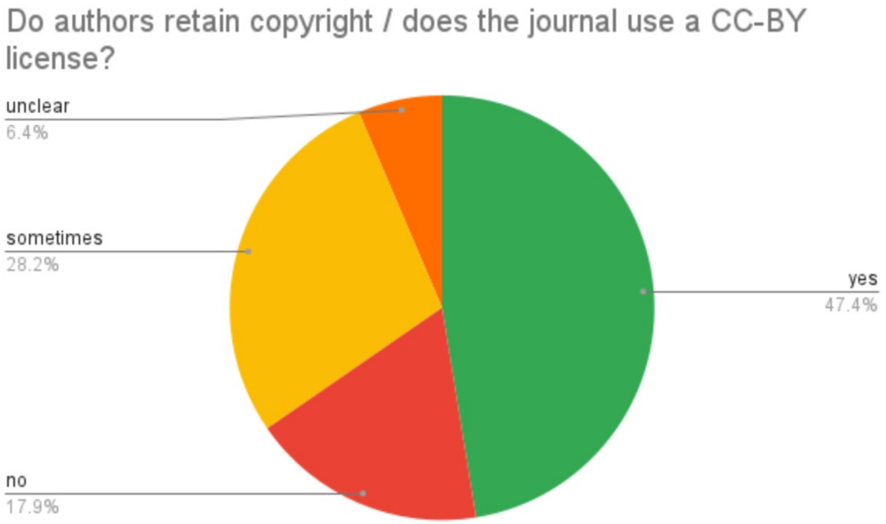
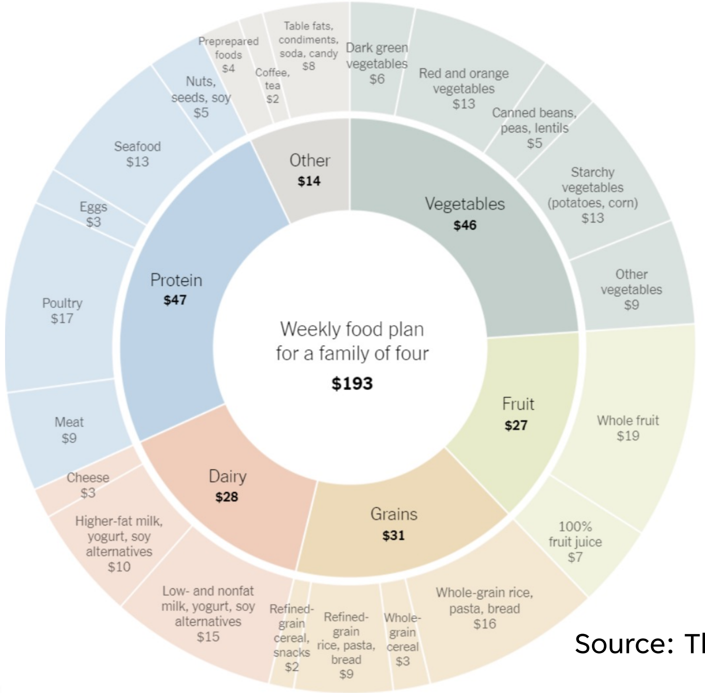


Figure 5. Status of copyright ownership and Creative Commons licensing.

Accessible equivalent of this chart as a list.

Source: Borchardt et al. 2022



Source: The New York Times

Source: Austen, Pride & Prejudice

TYPES OF DATA VISUALIZATIONS

Tables

Charts

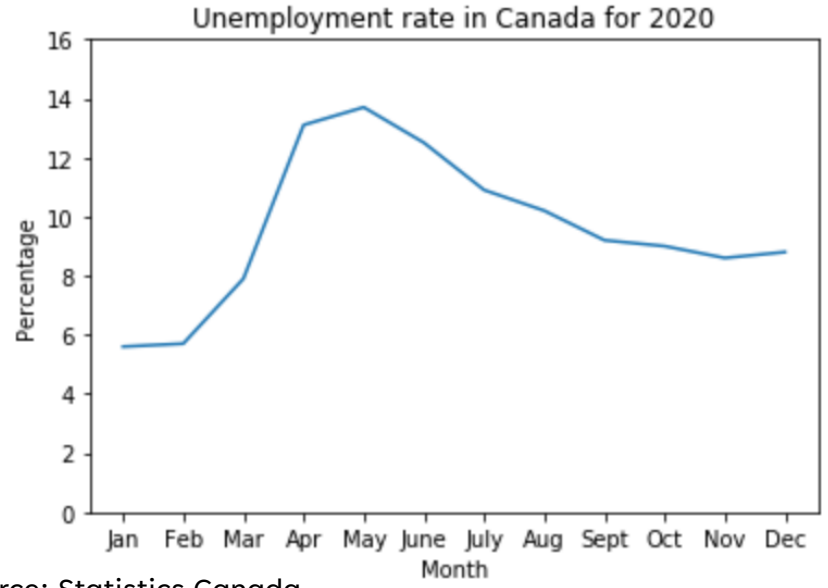
Graphs

Networks

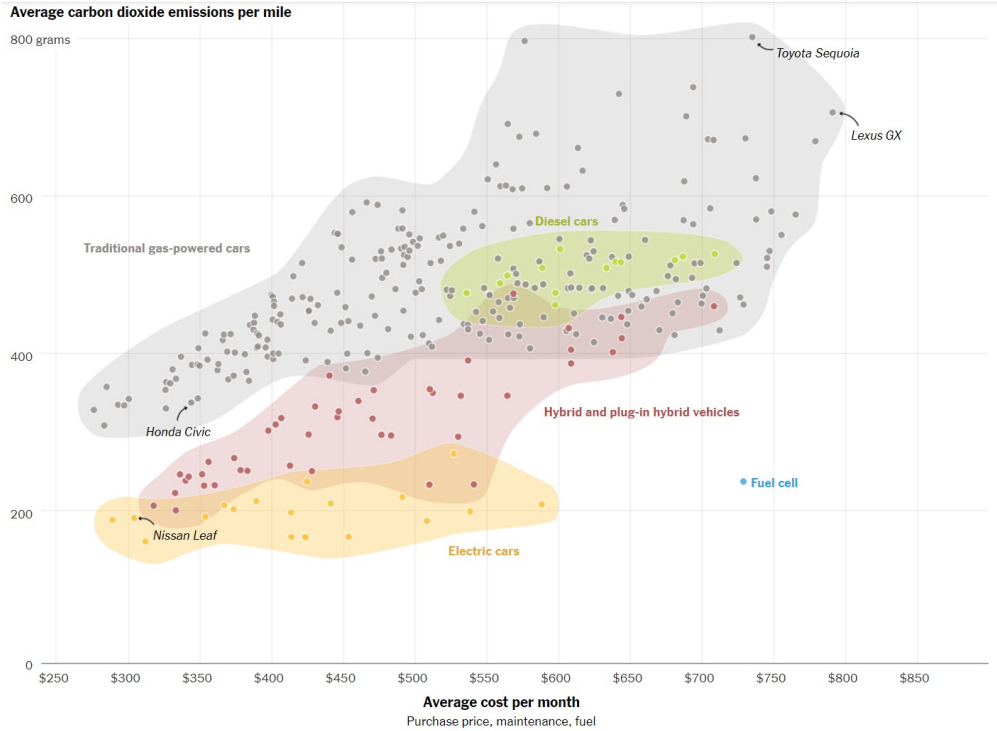
Maps

Infographics

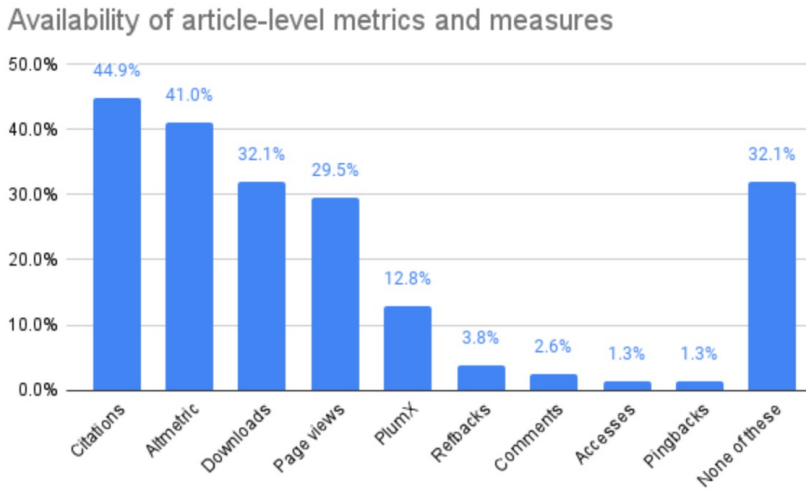
Dashboards



Source: Statistics Canada



Source: The New York Times



Source: Borchardt et al. 2022

TYPES OF DATA VISUALIZATIONS

Tables

Charts

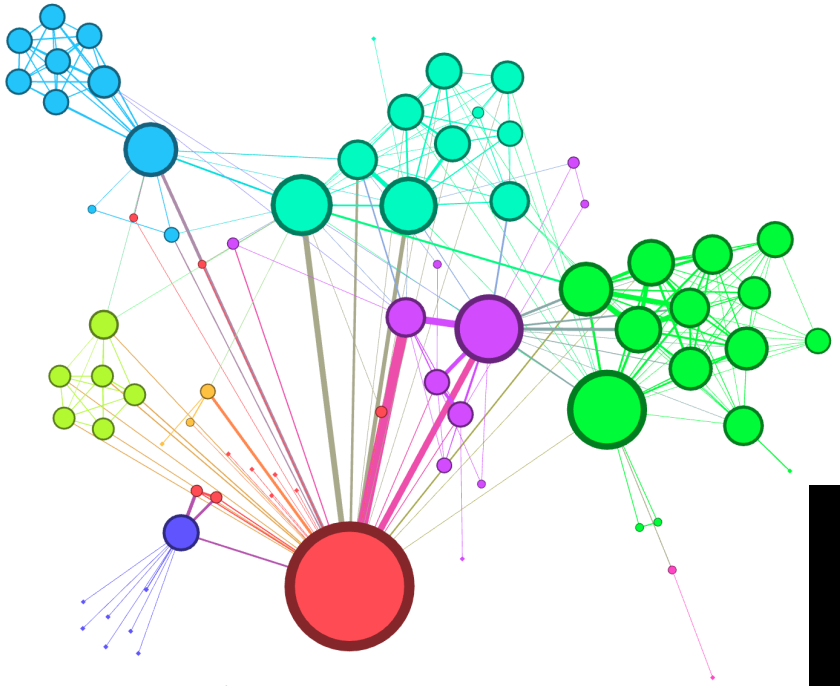
Graphs

Networks

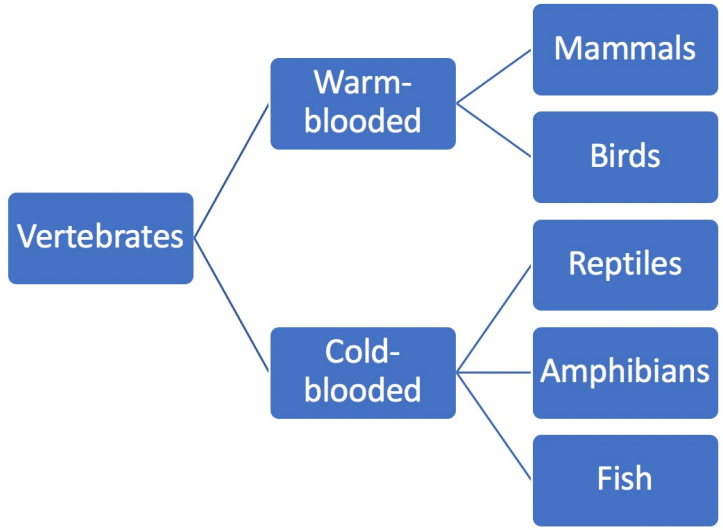
Maps

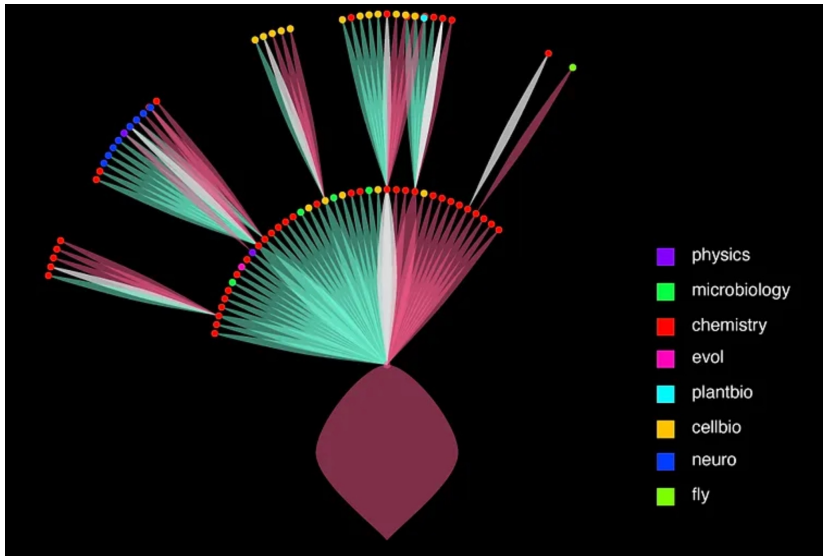
Infographics

Dashboards



Source: Gephi





Source: mentortrees.com

TYPES OF DATA VISUALIZATIONS

Tables

Charts

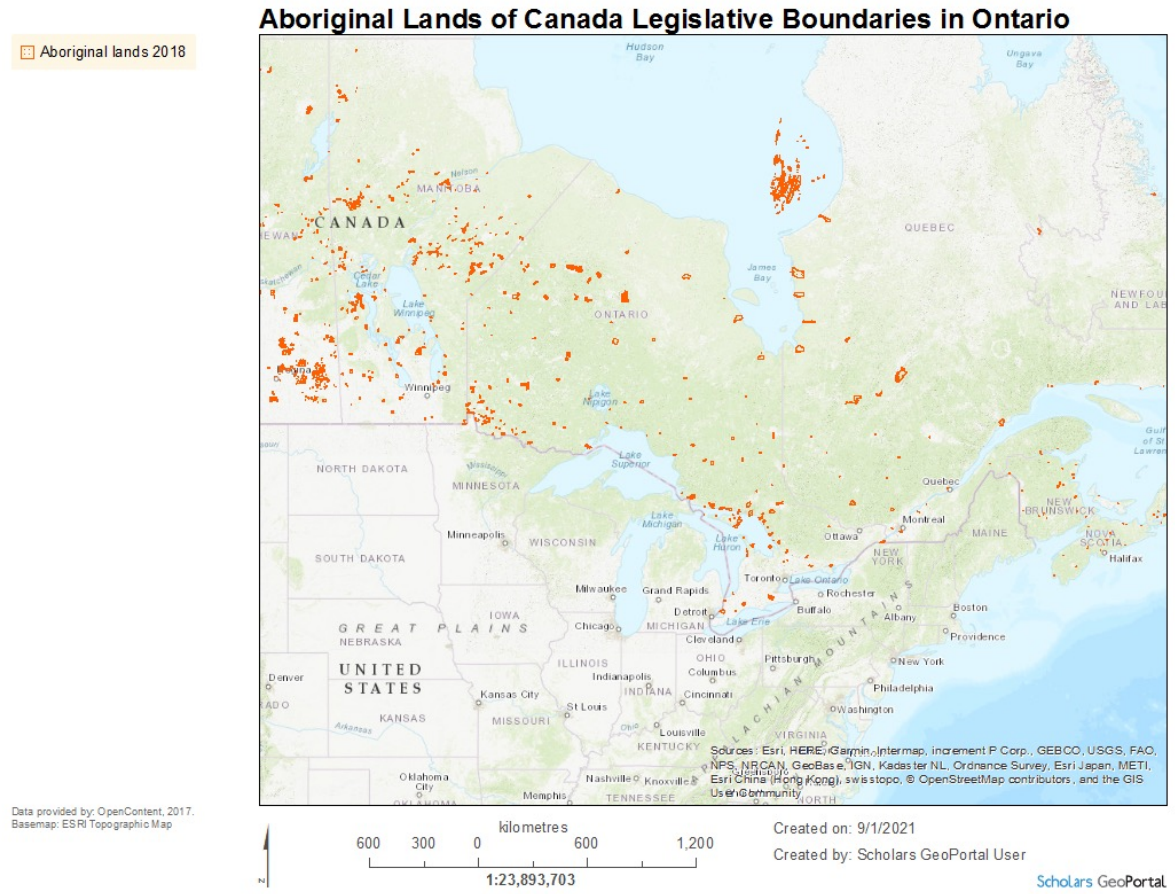
Graphs

Networks

Maps

Infographics

Dashboards



TYPES OF DATA VISUALIZATIONS

Tables

Charts

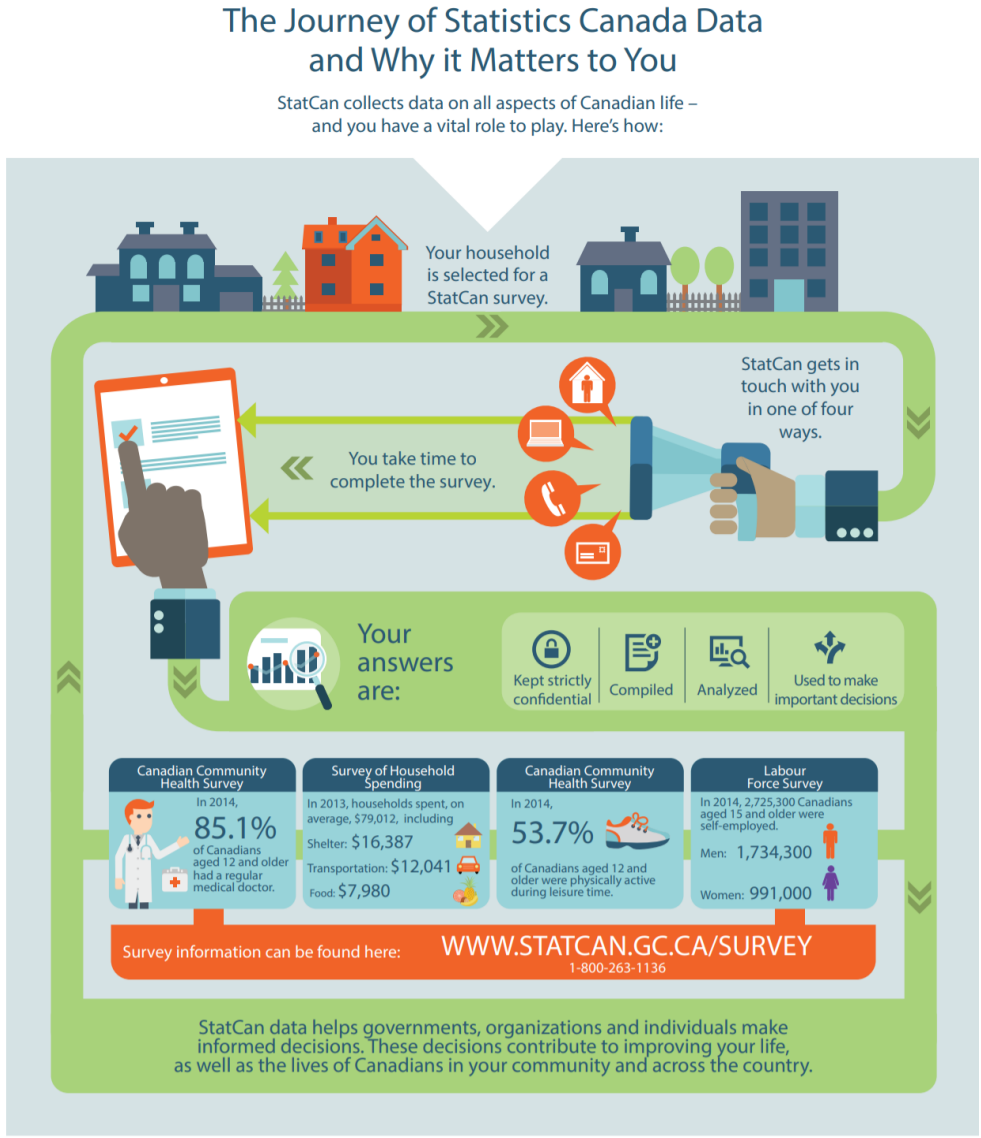
Graphs

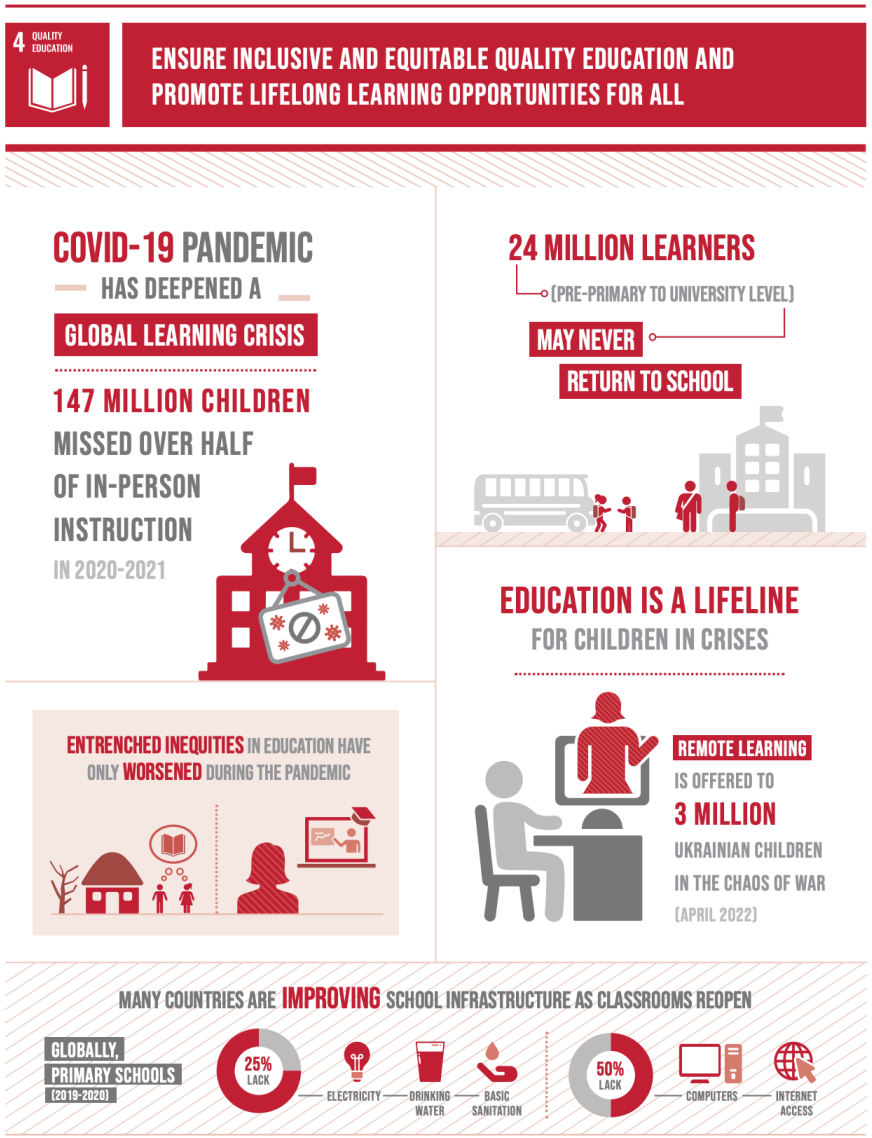
Networks

Maps

Infographics

Dashboards





Source: United Nations, 2022

TYPES OF DATA VISUALIZATIONS

Tables

Charts

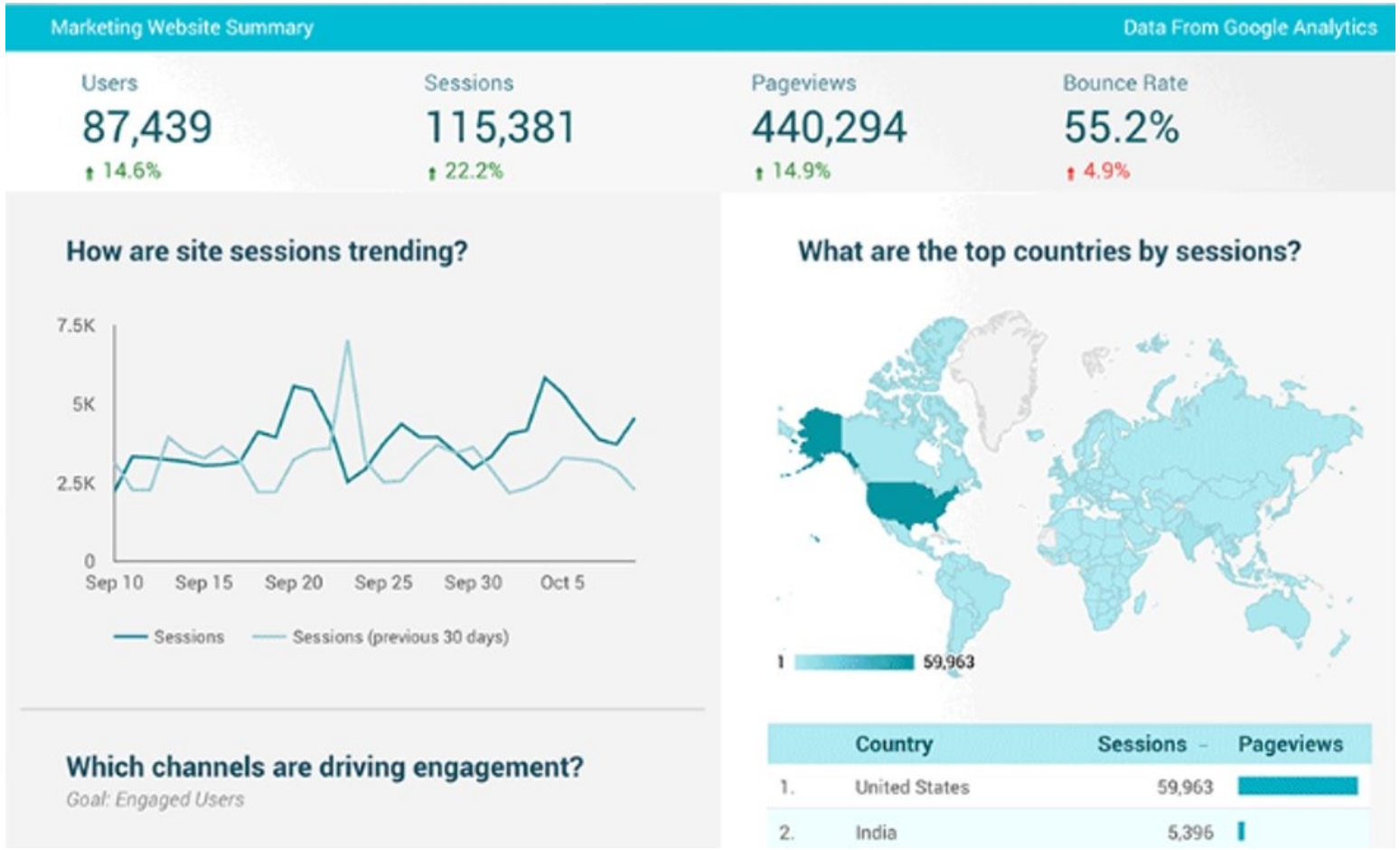
Graphs

Networks

Maps

Infographics

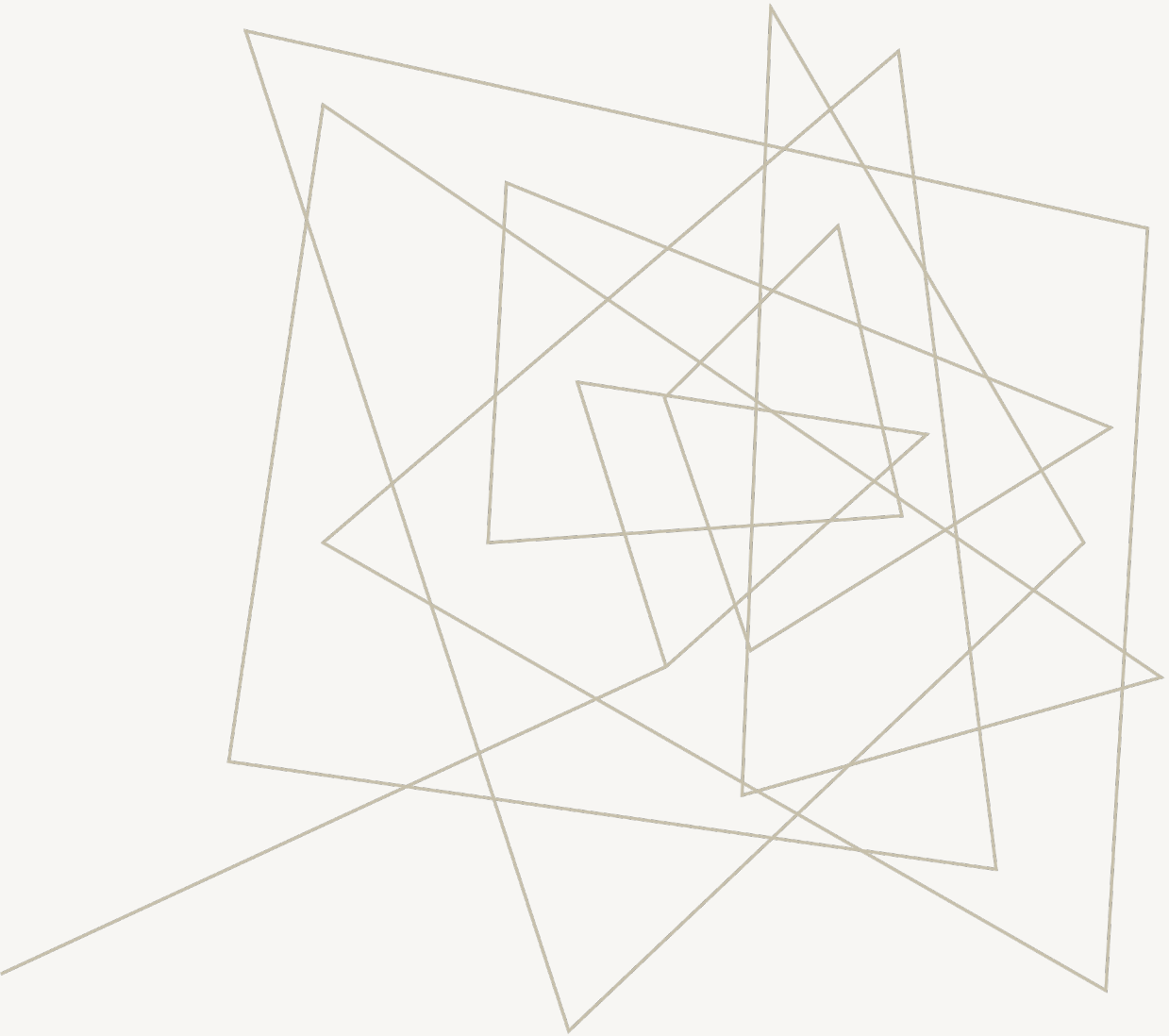
Dashboards



TYPES OF DATA VISUALIZATIONS

- Tables
- Charts
- Graphs
- Networks
- Maps
- Infographics
- Dashboards

Source: https://datavizcatalogue.com/



HOW TO CREATE A DATA VISUALIZATION

HOW TO CREATE A DATA VISUALIZATION IN 5 STEPS

1. DEFINE THE PURPOSE, AUDIENCE, AND CONTEXT

- Identify why you are creating a visualization in the first place. Is it for **exploratory** purposes (analyzing data) or **explanatory** (communicating a message)?
- What is the main **takeaway**? Factual or persuasive?
- Who is the **audience**? Demographics, data literacy?
- What is your **context** for creation? Timeline, budget, knowledge?
- **Where** will the data visualization be shared? Online, published article, poster?

2. COLLECT, EXPLORE, AND CLEAN THE DATA

COLLECT DATA

- **Data sources** (https://researchguides.library.yorku.ca/data)
  - Open data portals and research data repositories
  - Statista
  - Government websites
  - Advanced Google Searches

- Evaluate your data sources for their quality. **This checklist** can guide your evaluation:
  - Source and authority
  - Objectivity and purpose
  - Currency and geographic coverage
  - Methodology and completeness

2. COLLECT, EXPLORE, AND CLEAN THE DATA

	A	B	C	D	E	F	G	H	I
1	Serial_Num	Season	Num	Basin	Sub_basin	Name	ISO_time	Nature	Latitude
2	1851175N26	1851	1	NA	GM	UNNAMED	1851-06-25 00:00:00	TS	28
3	1851175N26	1851	1	NA	GM	UNNAMED	1851-06-25 06:00:00	TS	28
4	1851175N26	1851	1	NA	GM	UNNAMED	1851-06-25 12:00:00	TS	28
5	1851175N26	1851	1	NA	GM	UNNAMED	1851-06-25 18:00:00	TS	28.1
6	1851175N26	1851	1	NA	GM	UNNAMED	1851-06-25 21:00:00	TS	28.2
7	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-26 00:00:00	TS	28.2
8	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-26 06:00:00	TS	28.3
9	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-26 12:00:00	TS	28.4
10	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-26 18:00:00	TS	28.6
11	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-27 00:00:00	TS	29
12	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-27 06:00:00	TS	29.5
13	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-27 12:00:00	TS	30
14	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-27 18:00:00	TS	30.5
15	1851175N26	1851	1	NA	NA	UNNAMED	1851-06-28 00:00:00	TS	31
16	1851187N22	1851	3	NA	GM	UNNAMED	1851-07-05 12:00:00	TS	22.2
17	1851192N12	1851	4	NA	CS	UNNAMED	1851-07-10 12:00:00	TS	12
18	1851228N13	1851	6	NA	NA	UNNAMED	1851-08-16 00:00:00	TS	13.4
19	1851228N13	1851	6	NA	NA	UNNAMED	1851-08-16 06:00:00	TS	13.7

CLEAN DATA

- **Set-up:** Table format with no missing rows or columns, headers in your first row, no complex values within a single column
- **Cleaning:** Remove stray punctuation, spaces, test records, spelling errors, other issues that impact the data
- **Standardize**: Names, dates, missing values
- **Transforming**: Rearrange table, pivot table, etc.

2. COLLECT, EXPLORE, AND CLEAN THE DATA

CLEAN DATA: WITH SOFTWARE

Use software tools to help automate cleaning!

- **Excel**, Google Sheets, or equivalent open source
  - Basic functions and formulas to clean spreadsheet (Find & Replace, Remove Duplicates, Text to Columns, etc.)

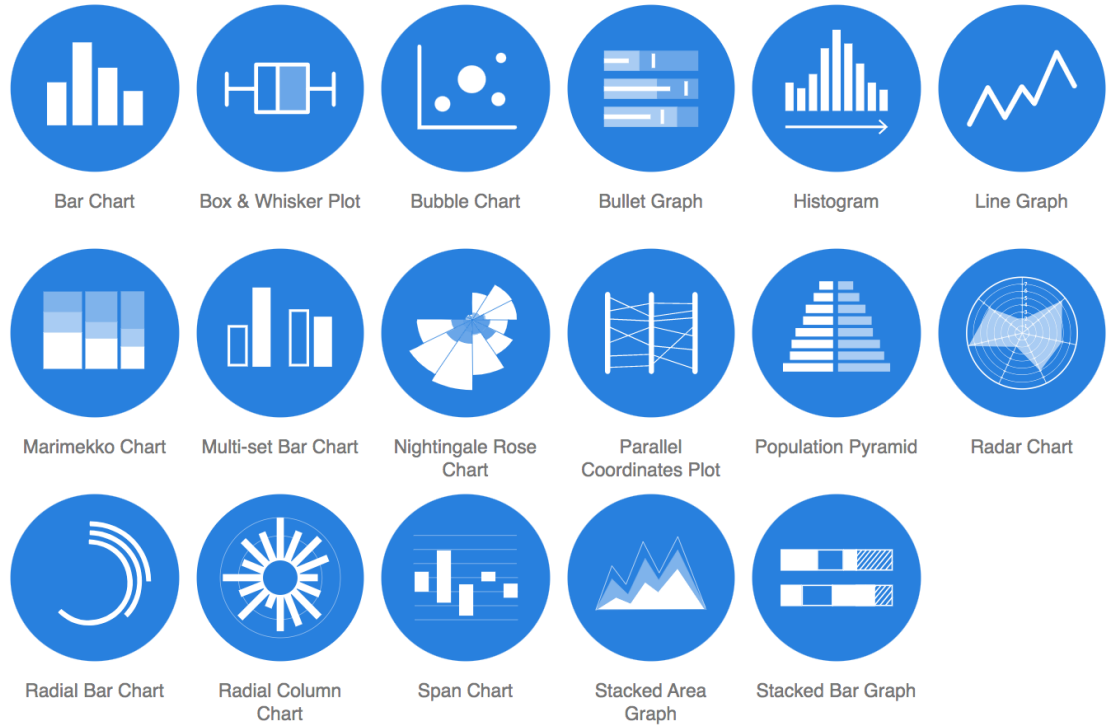
- **OpenRefine** (open source)
  - Reproducible cleaning steps and advanced features for text cleaning

- **Scripting languages like Python, R** (open source)
  - Create scripts for highly reproducible data cleaning
  - Powerful but with a learning curve

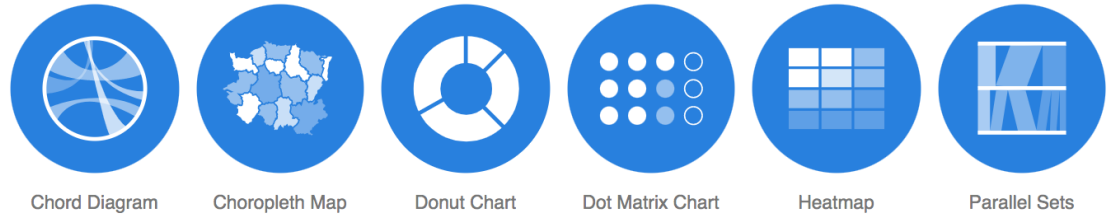
3. SELECT VISUALIZATION TYPE AND TOOL

[The Data Visualization Catalogue]

With an axis



Without an axis



3. SELECT VISUALIZATION TYPE AND TOOL

Quick and basic

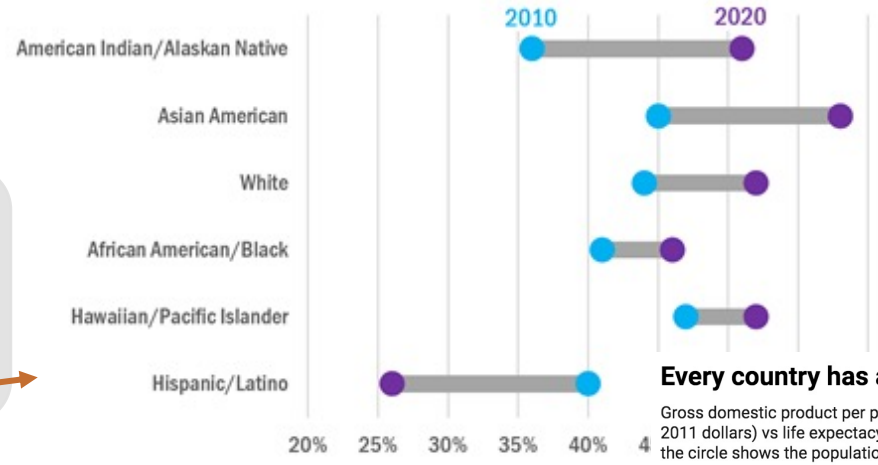
- Pen and paper
- Microsoft Excel
- Google Sheets

Robust and beginner friendly

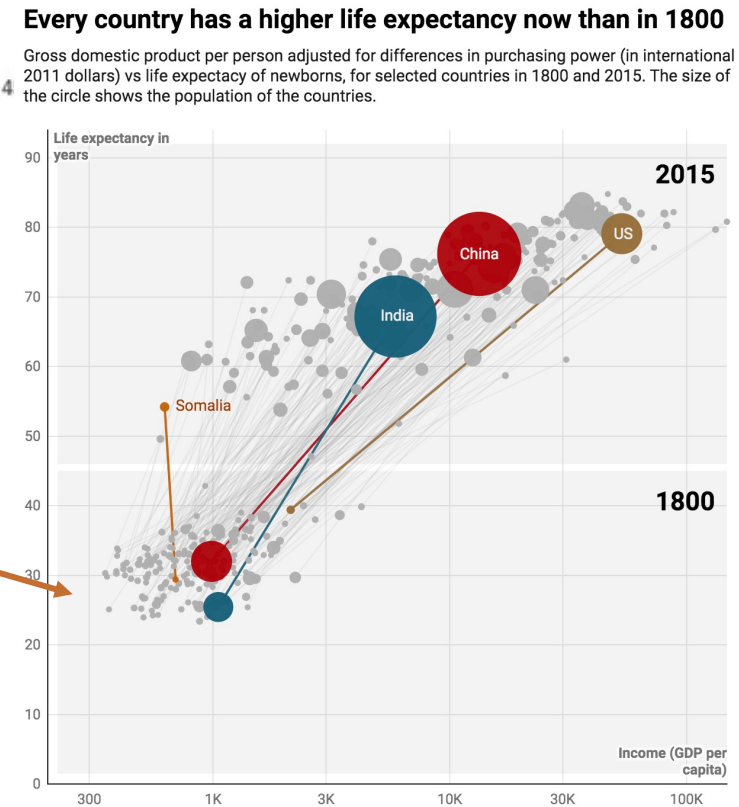
- Datawrapper
- Flourish
- Tableau
- RAWGraphs (OS)

Coding required but powerful

- R (OS)
- Python (OS)
- D3.js (OS)

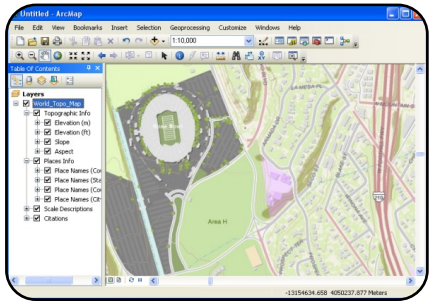


Source: Stephanie Evergreen



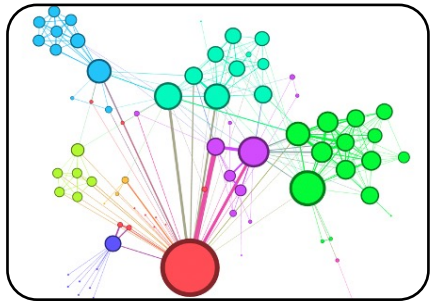
3. SELECT VISUALIZATION TYPE AND TOOL

SPECIALIZED DATA VISUALIZATION TOOLS



Maps

- ArcGIS
- QGIS

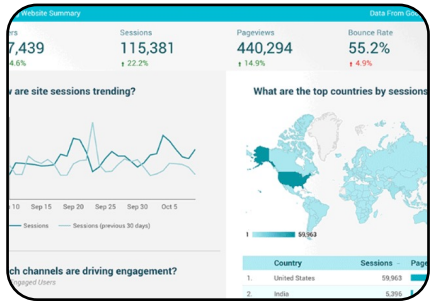


Networks

- Gephi
- Cytoscape

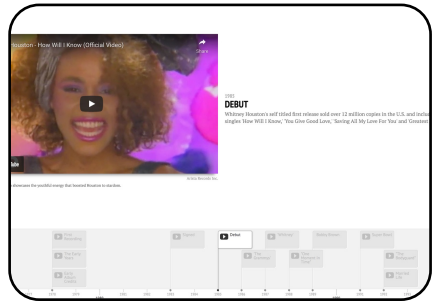
Infographics

- Canva
- Infogram
- Piktochart



Dashboards

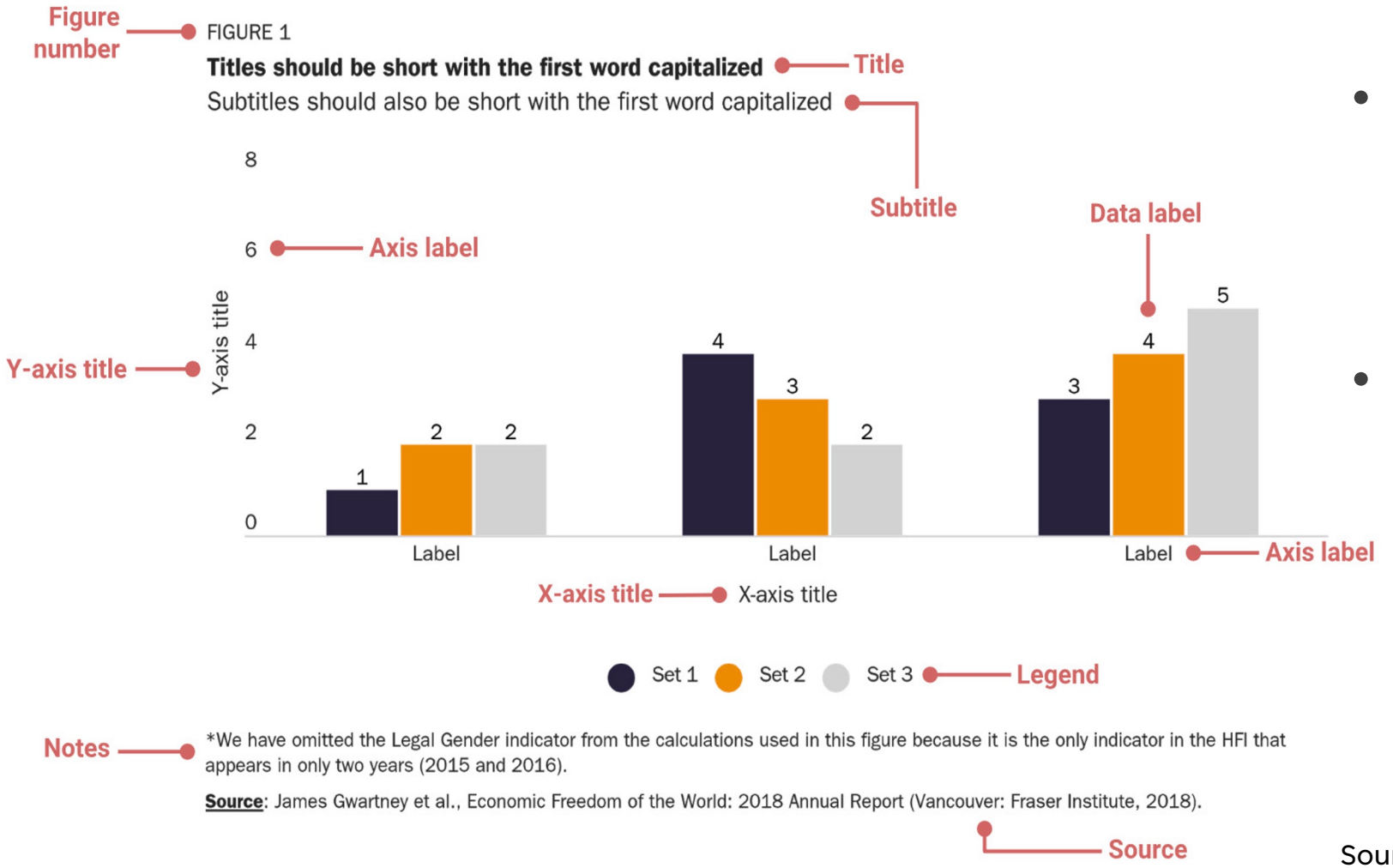
- Tableau
- Microsoft Power BI
- Google Data Studio



Timelines

- TimelineJS
- Tiki-Toki

4. TEST, REFINE, & ITERATE VISUALIZATION

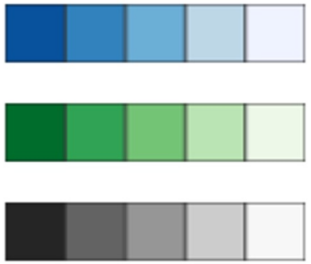


- Does this visualization **type** and **tool** best meet your needs?
- Should visualization **elements** be edited to better convey message?

Source: Abrigo, L.A. & Schneider, G.S. (2018). *Data visualization guidelines.* The Cato Institute. https://github.com/glosophy/CatoDataVizGuidelines

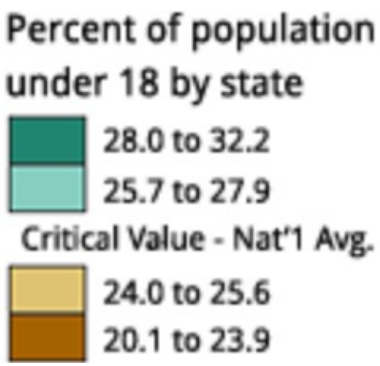
4. TEST, REFINE, & ITERATE VISUALIZATION

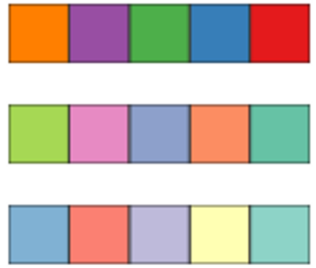
COLOUR CHOICES



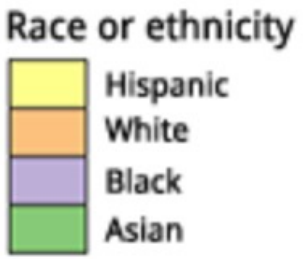
Sequential
(dark to light or light to dark)

Diverging
(dark in 1 hue to light to dark in a different hue)





Categorical



[Color scales from ColorBrewer 2.0]

4. TEST, REFINE, & ITERATE VISUALIZATION

ACCESSIBILITY

- Include the **data table**
- Write **alt text** for your data visualizations
- Verify that your visualizations are colour-blind friendly: **colour choices**, **colour contrast**, **type contrast**
- Create strong takeaway **titles**

EQUITY, DIVERSITY, AND INCLUSION

- Critically examining your data and data visualization elements and reflect on your audience
  - Ex. Missing groups: gender, race
  - Ex. Colour: avoid gradients and hierarchies for categories
- https://researchguides.library.yorku.ca/datavisualization/edi

5. SHARE AND PUBLISH VISUALIZATION

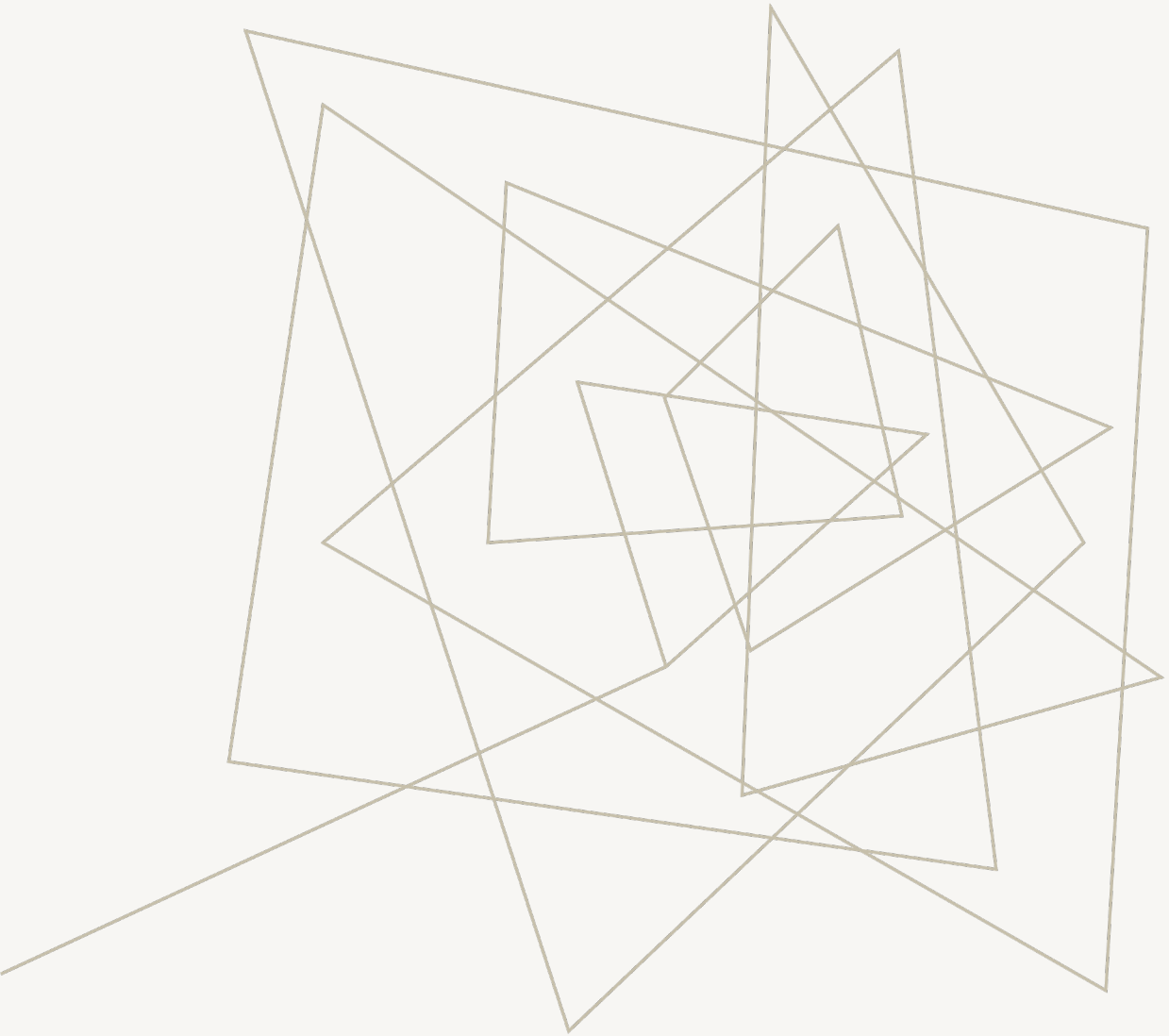
Cite if you are using someone else's data! Why?

- Give credit and helps readers find the original data source
- Increases the legitimacy of your chart
- Helps track licensing and reuse permission details

Copy the citation suggested where you got the data or use your preferred style.

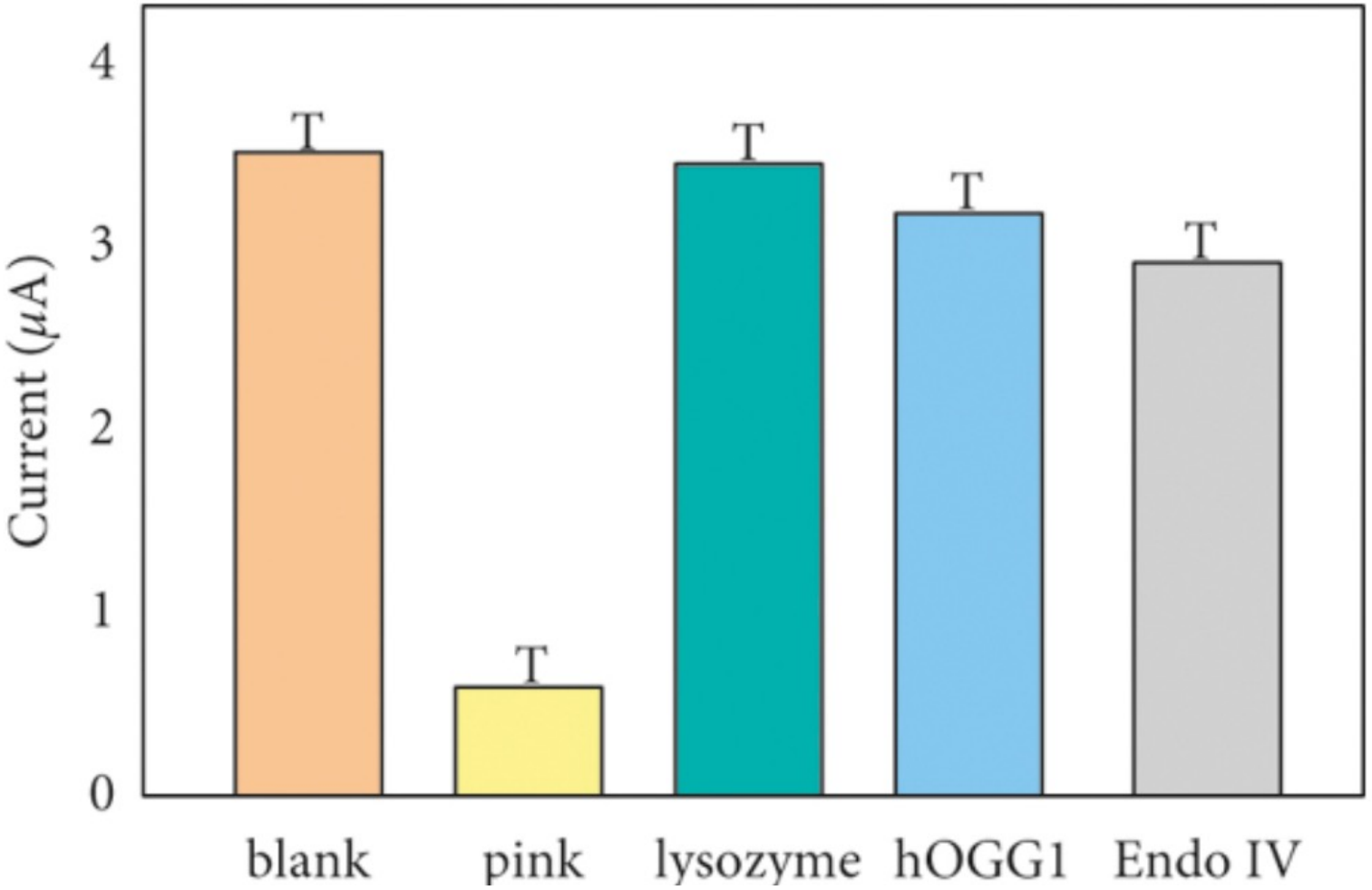
Aim for data sharing and research reproducibility

- Share your data as supplementary research, perhaps upload to Borealis
- Aim for reproducing data visualizations – raw data & cleaned data, your steps to create your data visualization

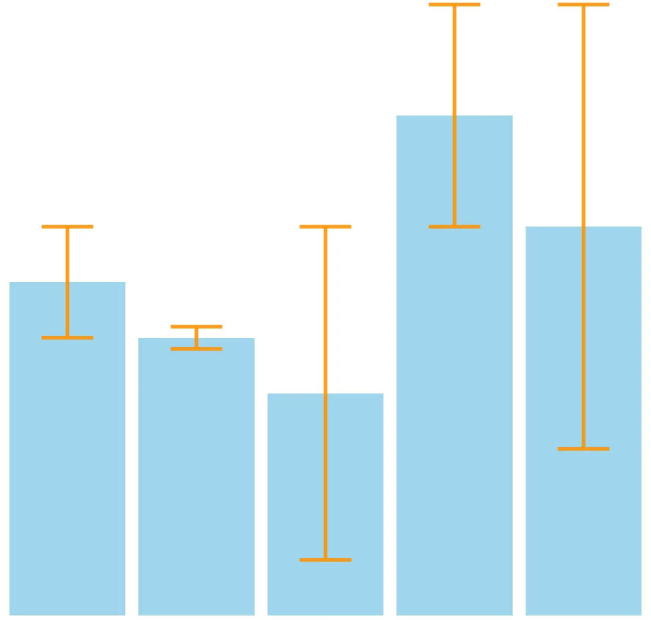


HOW TO CREATE A DATA VISUALIZATION: AN EXAMPLE

DATA VISUALIZATION IN RESEARCH: A CAUTIONARY TALE



Source: Ruyao Gong, Binghong Liu, "Monitoring of Sports Health Indicators Based on Wearable Nanobiosensors", *Advances in Materials Science and Engineering*, https://doi.org/10.1155/2022/3802603



Source: https://r-graph-gallery.com/4-barplot-with-error-bar.html

EXAMPLE

Imagine we created a survey to understand how our library system feels about X.

We asked demographic questions, including race, gender, time working within the library system.

How can we share these survey results?

1. DEFINE THE PURPOSE, AUDIENCE, AND CONTEXT

- **Purpose**: to explain the data
- **Takeaway**: factual communication of demographic data
- **Context**: Plan to submit the research as an article in the journal, Partnership: The Canadian Journal of Library & Information Practice & Research
  - Format requirements: Colour? Placement?
  - Not a lot of time nor previous knowledge on creating data visualizations
- **Audience**: Library workers reading the journal, with varying levels of data visualization literacy

2. COLLECT AND CLEAN THE DATA

ID	Gender	2SLGBTQ+	Racialized	Time in library system
1	Female	Yes	No	10 years
2	Female	No	No	8.5 years
3	Female	No	No	1 year
4	Male	Yes	No	1 month
5	Prefer not to say	No	Yes	13
6	Male	No	Yes	11 years
7	Male	No	No	3 years
8	Female	No	No	4 y
9	Female	No	Yes	5 Year
10	Female	No	Yes	7 year
11	Transgender	Yes	No	7 Year
12	A gender not listed	Yes	Yes	0.5 years
13	Female	Yes	Yes	1 years
14	Female	Yes	No	3 yers
15	Female	No	Yes	11
16	Male	No	Yes	12 years
17	Female	No	No	10 years
18	Male	No	Yes	9 years
19	Female	Yes	No	20 years
20	Female	No	Yes	19 years
21	Prefer not to say	Yes	Yes	13 years
22	Female	No	No	3 years
23	Female	Yes	No	4 years
24	Female	No	Yes	17 years
25	Male	No	No	23 tears
26	Female	Yes	No	2 years

“How many years have you worked within the library system? Please answer in number of years.”

Messy data that needs to be cleaned!

2. COLLECT AND CLEAN THE DATA

Time in library system	Time in library system (cleaned)
10 years	6-10 years
8.5 years	6-10 years
1 year	0-5 years
1 month	0-5 years
13	11-15 years
11 years	11-15 years
3 years	0-5 years
4 y	0-5 years
5 Year	0-5 years
7 year	6-10 years
7 Year	6-10 years
0.5 years	0-5 years
1 years	0-5 years
3 yers	0-5 years
11	11-15 years
12 years	11-15 years
10 years	6-10 years
9 years	6-10 years
20 years	16-20 years
19 years	0-5 years
13 years	16-20 years
3 years	0-5 years
4 years	21+ years
17 years	16-20 years
23 tears	21+ years
2 years	0-5 years

- Use a spreadsheet and add a new column for the cleaned data
- Cleaning this messy data = in the new column, manually converting the raw data to the appropriate 5 year timespan group

2. COLLECT AND CLEAN THE DATA

ID	Gender	2SLGBTQ+	Racialized	Time in library system (cleaned)
1	Female	Yes	No	6-10 years
2	Female	No	No	6-10 years
3	Female	No	No	0-5 years
4	Male	Yes	No	0-5 years
5	Prefer not to say	No	Yes	11-15 years
6	Male	No	Yes	11-15 years
7	Male	No	No	0-5 years
8	Female	No	No	0-5 years
9	Female	No	Yes	0-5 years
10	Female	No	Yes	6-10 years
11	Transgender	Yes	No	6-10 years
12	A gender not listed	Yes	Yes	0-5 years
13	Female	Yes	Yes	0-5 years
14	Female	Yes	No	0-5 years
15	Female	No	Yes	11-15 years
16	Male	No	Yes	11-15 years
17	Female	No	No	6-10 years
18	Male	No	Yes	6-10 years
19	Female	Yes	No	16-20 years
20	Female	No	Yes	0-5 years
21	Prefer not to say	Yes	Yes	16-20 years
22	Female	No	No	0-5 years
23	Female	Yes	No	21+ years
24	Female	No	Yes	16-20 years
25	Male	No	No	21+ years
26	Female	Yes	No	0-5 years

Transform the data using Excel pivot tables to group and count

Gender	Count of ID
Female	16
Male	6
Prefer not to say	2
A gender not listed	1
Transgender	1
Grand Total	**26**

2SLGBTQ+	Count of ID
No	16
Yes	10
Grand Total	**26**

Racialized	Count of ID
No	14
Yes	12
Grand Total	**26**

Time in library syste	Count of ID
0-5 years	11
6-10 years	6
11-15 years	4
16-20 years	3
21+ years	2
Grand Total	**26**

Gender	2SLGBTQ+	Racialized	Count of ID
Female	**No**	No	5
		Yes	5
	Yes	No	5
		Yes	1
Male	**No**	No	2
		Yes	3
	Yes	No	1
Prefer not to say	**No**	Yes	1
	Yes	Yes	1
Transgender	**Yes**	No	1
A gender not listed	**Yes**	Yes	1
Grand Total			**26**

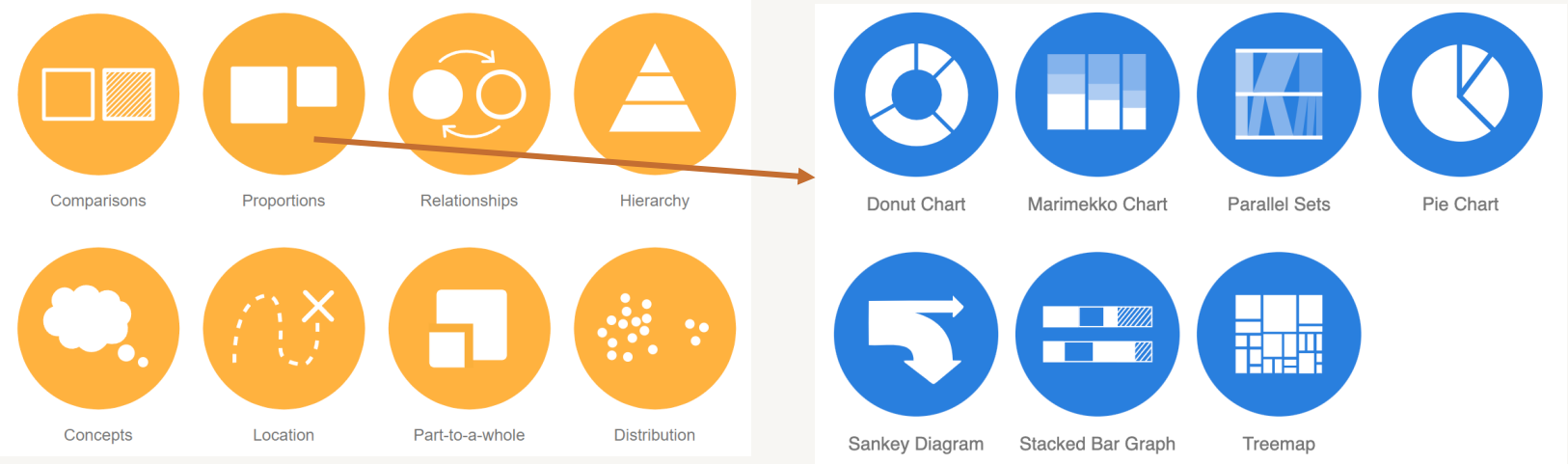
3. SELECT VISUALIZATION TYPE AND TOOL

CHOOSING THE TOOL

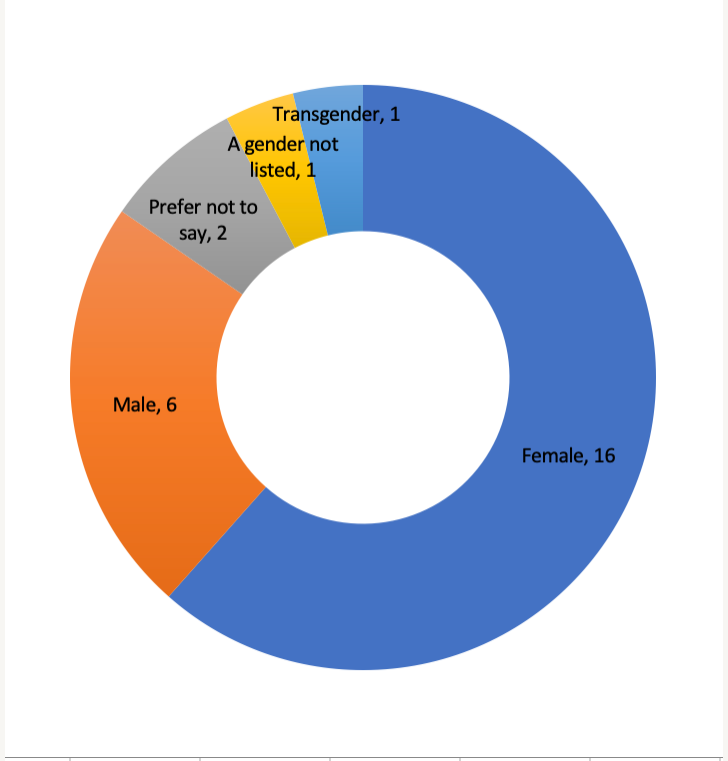
- We are not needing to create any specialized visualizations like maps or networks
- We want to take our context to help decide if we want to do a Quick & Basic tool, Robust & Beginner-friendly, or Coding Required but Powerful
- Since we don't have a lot of time or previous knowledge, let us pick Excel for now, where we were already doing our data cleaning

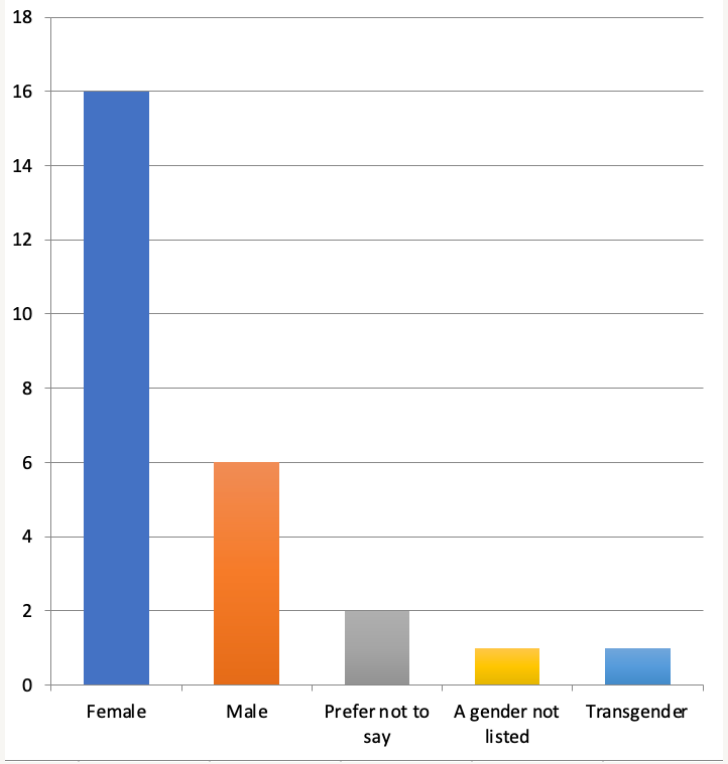
CHOOSING THE TYPE

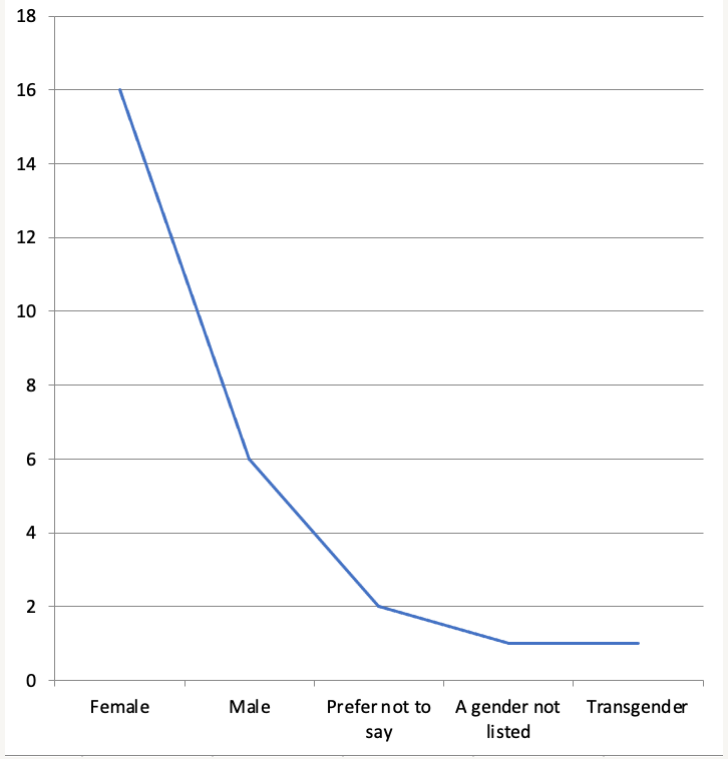
- Choosing a chart type that emphasizes proportions



4. TEST, REFINE, & ITERATE VISUALIZATION

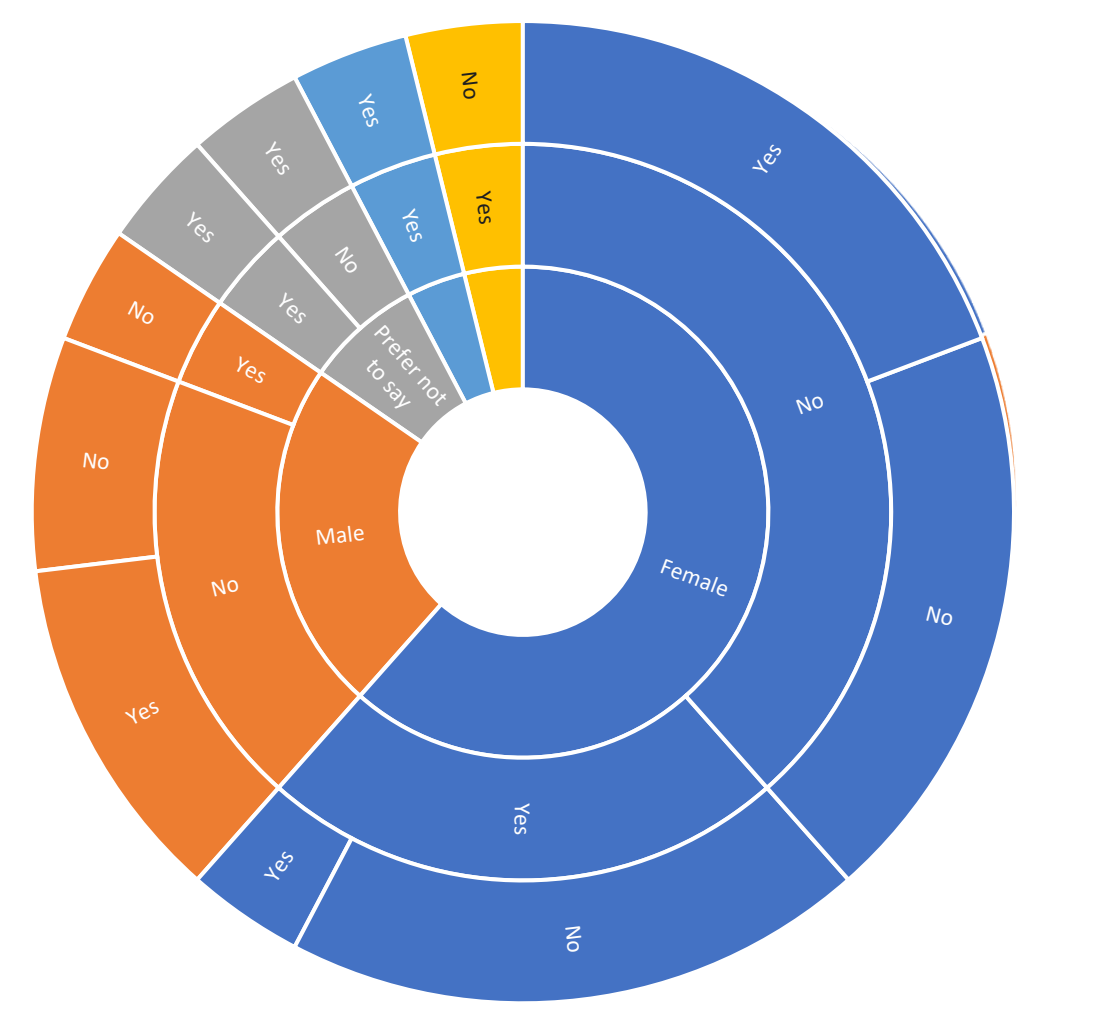




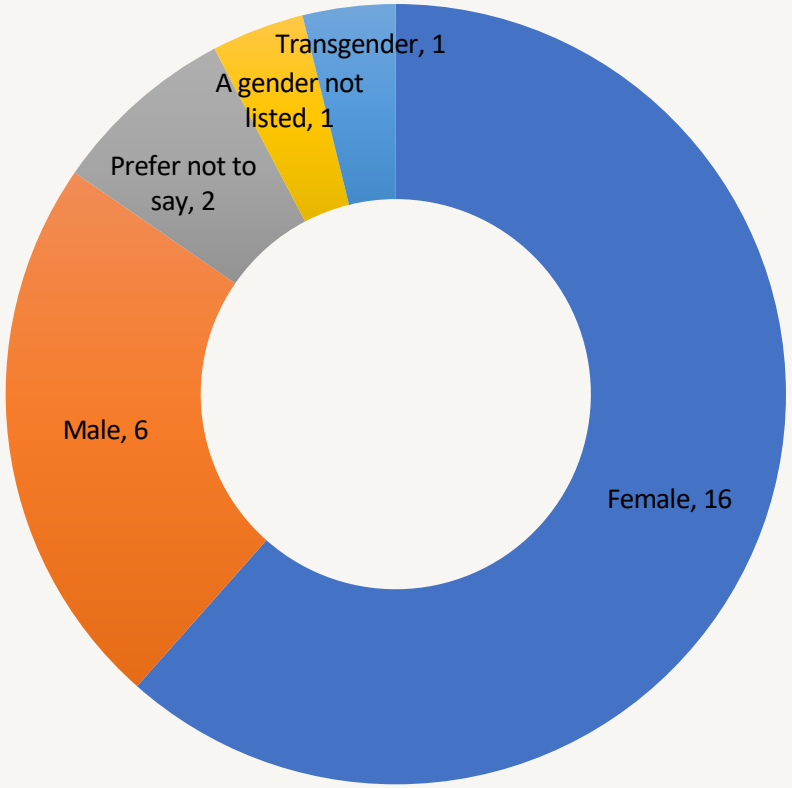


4. TEST, REFINE, & ITERATE VISUALIZATION

Gender	2SLGBTQ+	Racialized	Count of ID
Female	No	No	5
		Yes	5
	Yes	No	5
		Yes	1
Male	No	No	2
		Yes	3
	Yes	No	1
Prefer not to say	No	Yes	1
	Yes	Yes	1
Transgender	Yes	No	1
A gender not listed	Yes	Yes	1
Grand Total			26

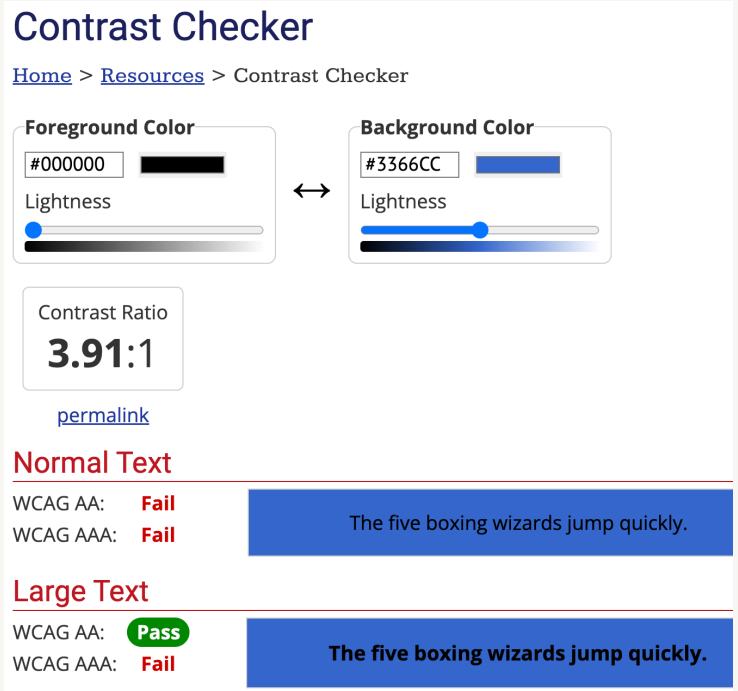


4. TEST, REFINE, & ITERATE VISUALIZATION



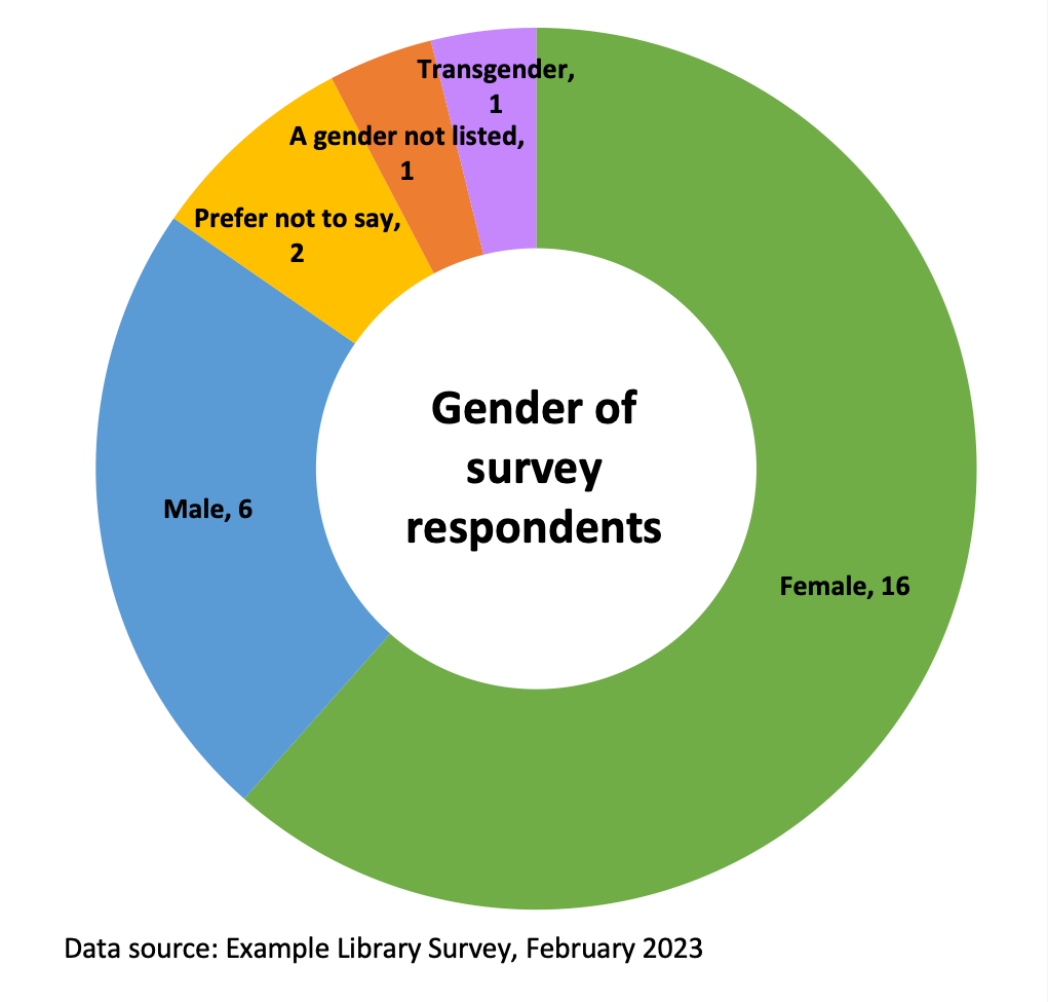
Continuing to refine:

- Are my colours accessible?
- Are my group names created with empathy?
- Can I add elements like a title, source, subtitle?



https://webaim.org/resources/contrastchecker/

5. SHARE AND PUBLISH VISUALIZATION



THANK YOU!

Alexandra Wong

Data Visualization and Analytics Librarian, York University

wongalex@yorku.ca

EXAMPLE LIBRARY SURVEY DEMOGRAPHIC DATA

ID	Gender	2SLGBTQ+	Racialized	Time in library system (cleaned)
1	Female	Yes	No	6-10 years
2	Female	No	No	6-10 years
3	Female	No	No	0-5 years
4	Male	Yes	No	0-5 years
5	Prefer not to say	No	Yes	11-15 years
6	Male	No	Yes	11-15 years
7	Male	No	No	0-5 years
8	Female	No	No	0-5 years
9	Female	No	Yes	0-5 years
10	Female	No	Yes	6-10 years
11	Transgender	Yes	No	6-10 years
12	A gender not listed	Yes	Yes	0-5 years
13	Female	Yes	Yes	0-5 years
14	Female	Yes	No	0-5 years
15	Female	No	Yes	11-15 years
16	Male	No	Yes	11-15 years
17	Female	No	No	6-10 years
18	Male	No	Yes	6-10 years
19	Female	Yes	No	16-20 years
20	Female	No	Yes	0-5 years
21	Prefer not to say	Yes	Yes	16-20 years
22	Female	No	No	0-5 years
23	Female	Yes	No	21+ years
24	Female	No	Yes	16-20 years
25	Male	No	No	21+ years
26	Female	Yes	No	0-5 years