An Introduction to Data Visualization

Or, Saying Good Research Things Goodly

Martin Chandler
Data Services Librarian
martin.chandler@mcgill.ca

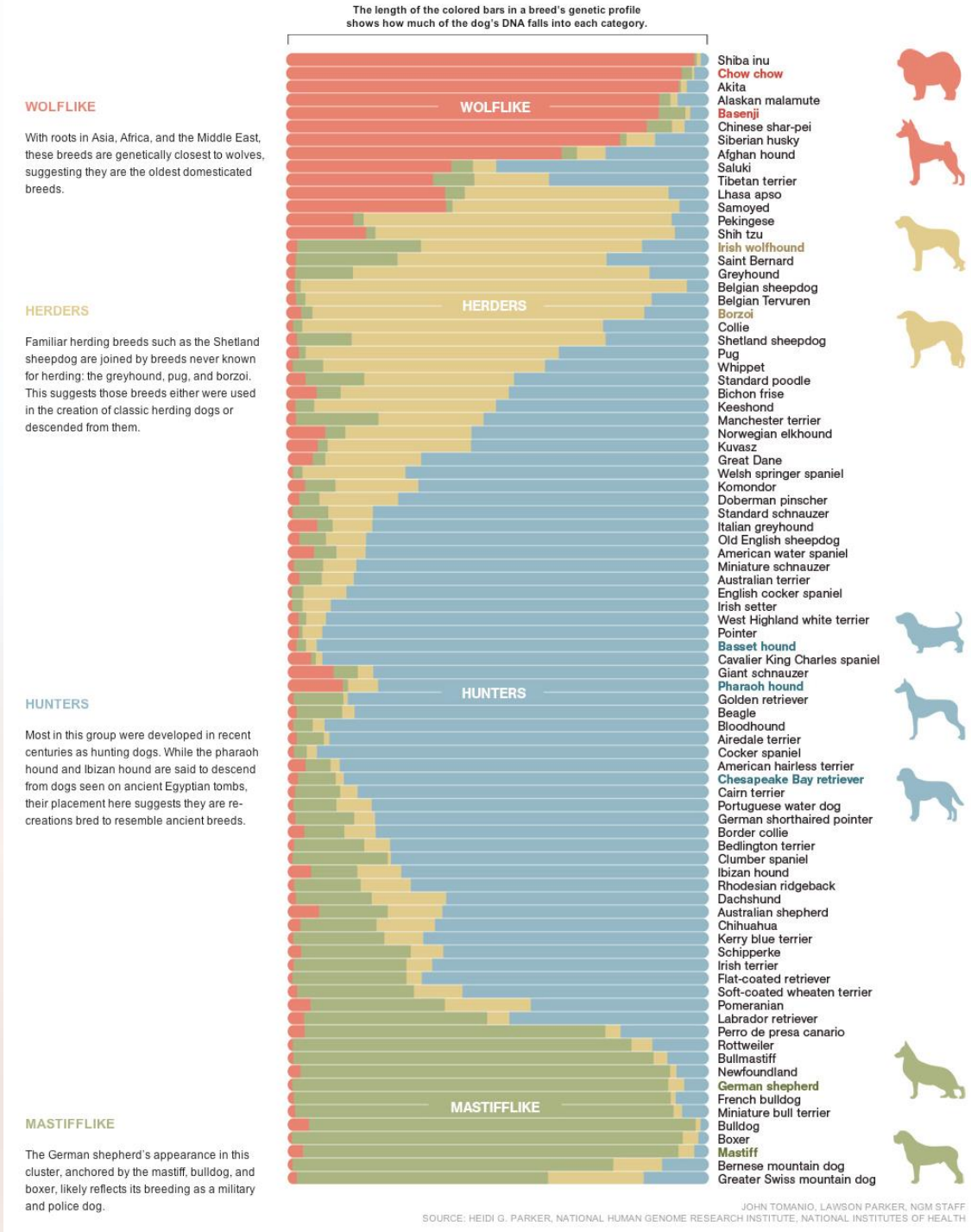
This presentation:

- https://bit.ly/2PjpIRo

Agenda

- A little about “data”
- Defining “data visualization”, and “infographic”
- Principles of Data Visualization
- Good and Bad Visualization
- Process for Visualization
- Working with Data
- Citing Data
- Resources

Let’s start:


The length of the colored bars in a breed's genetic profile shows how much of the dog's DNA falls into each category.
WOLFLIKE
With roots in Asia, Africa, and the Middle East, these breeds are genetically closest to wolves, suggesting they are the oldest domesticated breeds.
HERDERS
Familiar herding breeds such as the Shetland sheepdog are joined by breeds never known for herding: the greyhound, pug, and borzoi. This suggests those breeds either were used in the creation of classic herding dogs or descended from them.
HUNTERS
Most in this group were developed in recent centuries as hunting dogs. While the pharaoh hound and Ibizan hound are said to descend from dogs seen on ancient Egyptian tombs, their placement here suggests they are re-creations bred to resemble ancient breeds.
MASTIFFLIKE
The German shepherd's appearance in this cluster, anchored by the mastiff, bulldog, and boxer, likely reflects its breeding as a military and police dog.
WOLFLIKE
HERDERS
HUNTERS
MASTIFFLIKE
Shiba inu
Chow chow
Akita
Alaskan malamute
Basenji
Chinese shar-pei
Siberian husky
Afghan hound
Saluki
Tibetan terrier
Lhasa apso
Samoyed
Pekingese
Shih tzu
Irish wolfhound
Saint Bernard
Greyhound
Belgian sheepdog
Belgian Tervuren
Borzoi
Collie
Shetland sheepdog
Pug
Whippet
Standard poodle
Bichon frise
Keeshond
Manchester terrier
Norwegian elkhound
Kuvasz
Great Dane
Welsh springer spaniel
Komondor
Doberman pinscher
Standard schnauzer
Italian greyhound
Old English sheepdog
American water spaniel
Miniature schnauzer
Australian terrier
English cocker spaniel
Irish setter
West Highland white terrier
Pointer
Basset hound
Cavalier King Charles spaniel
Giant schnauzer
Pharaoh hound
Golden retriever
Beagle
Bloodhound
Airedale terrier
Cocker spaniel
American hairless terrier
Chesapeake Bay retriever
Cairn terrier
Portuguese water dog
German shorthaired pointer
Border collie
Bedlington terrier
Clumber spaniel
Ibizan hound
Rhodesian ridgeback
Dachshund
Australian shepherd
Chihuahua
Kerry blue terrier
Schipperke
Irish terrier
Flat-coated retriever
Soft-coated wheaten terrier
Pomeranian
Labrador retriever
Perro de presa canario
Rottweiler
Bullmastiff
Newfoundland
German shepherd
French bulldog
Miniature bull terrier
Bulldog
Boxer
Mastiff
Bernese mountain dog
Greater Swiss mountain dog
JOHN TOMANIO, LAWSON PARKER, NGM STAFF
SOURCE: HEIDI G. PARKER, NATIONAL HUMAN GENOME RESEARCH INSTITUTE, NATIONAL INSTITUTES OF HEALTH

Background information:

- Data: facts, or information, used to answer a question or make an argument
  - This can include numbers, maps, words, images, sounds, or other information – it all depends on the context.
- Dataset: A collection of facts, often gathered together to be manipulated. Includes variables (aspects being examined) and cases (each instance that contains an aspect)

Variable

Case

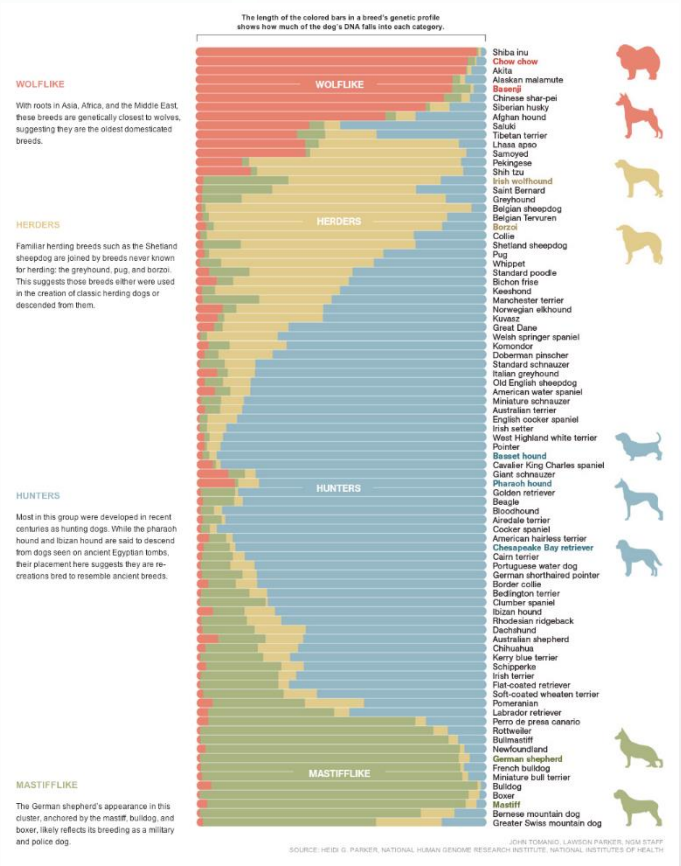
Home ID	Postal Code	Dog	Breed	Temperament
001	M6R 1N6	0	N/A	N/A
002	L2S 2A7	1	Beagle	Friendly
003	B2R 1S2	2	Dachshund; Terrier	Cute; Derpy
004	H4H 1V1	0	N/A	N/A

Background information:

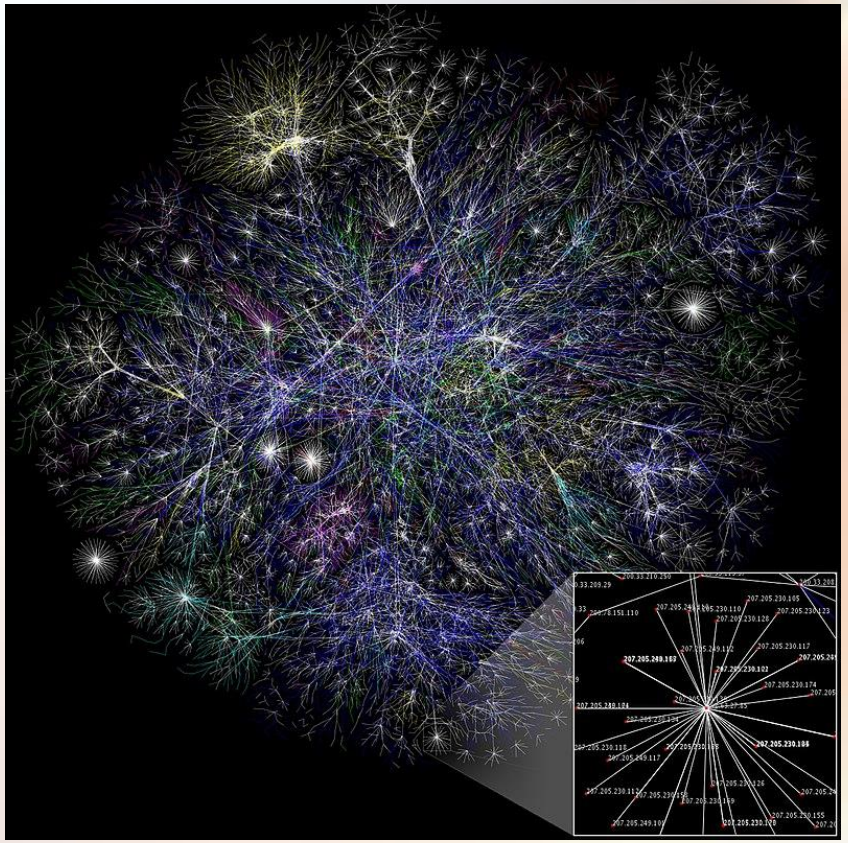
- Data Visualization: Using graphic of pictorial representations of data to explore, analyze, or communicate.
  - Scientific visualization: uses scientific data, generally with ties to physical objects or phenomena
  - Information visualization: generally involves visualizing more abstract ideas and concepts

Scientific vs Information

Genetic profiles of breeds of dog:

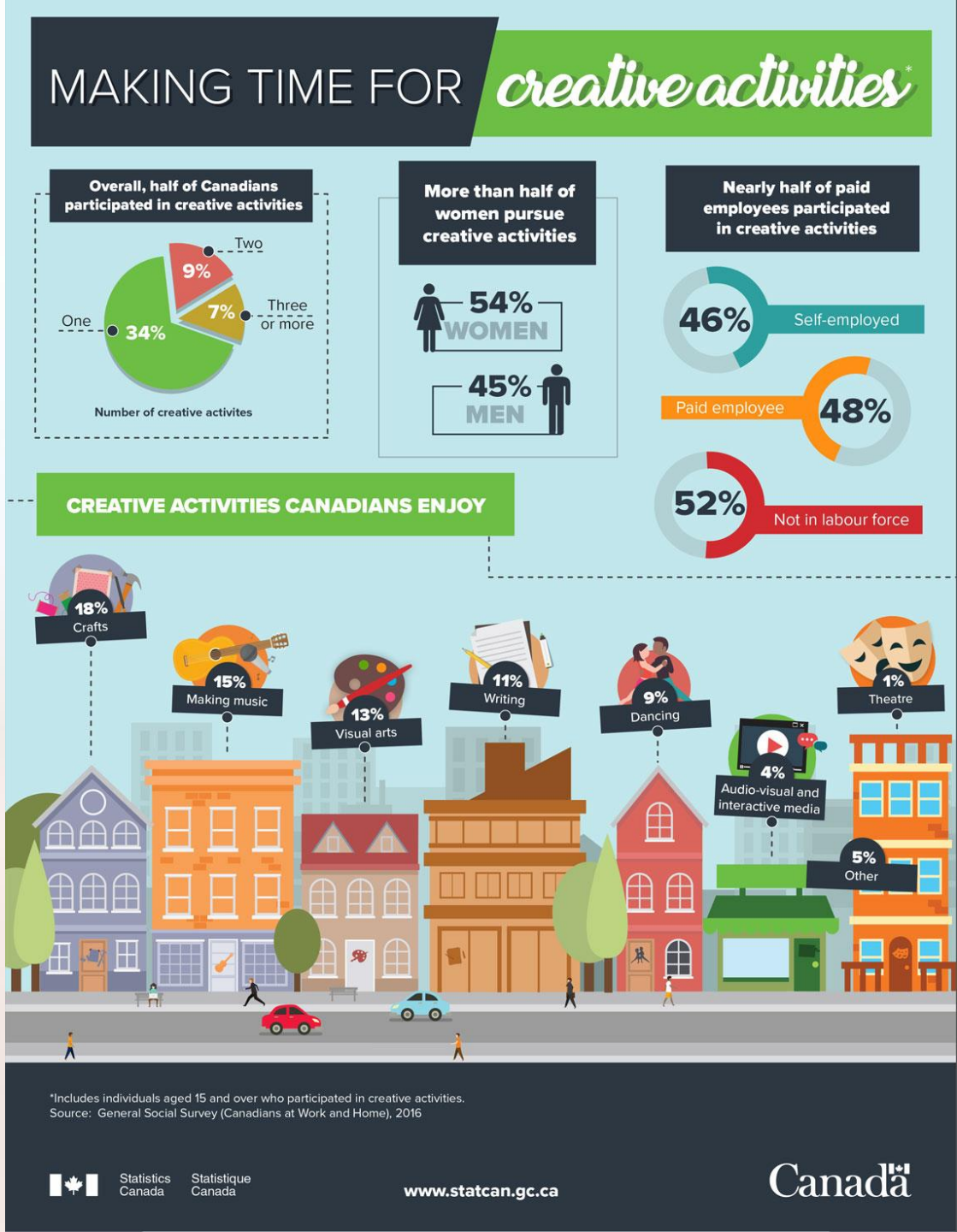


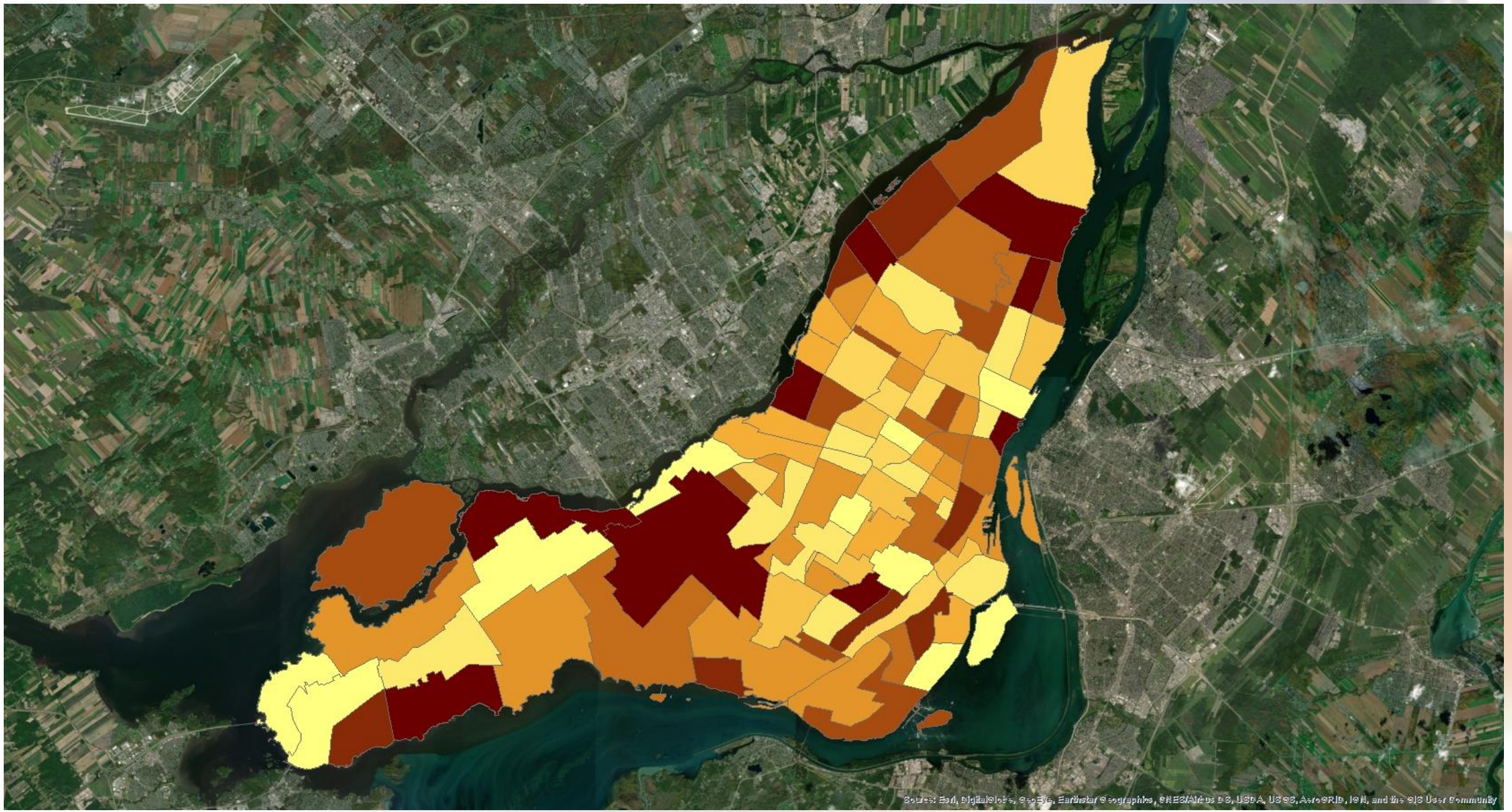
Part of the WWW in 2005:



Infographics

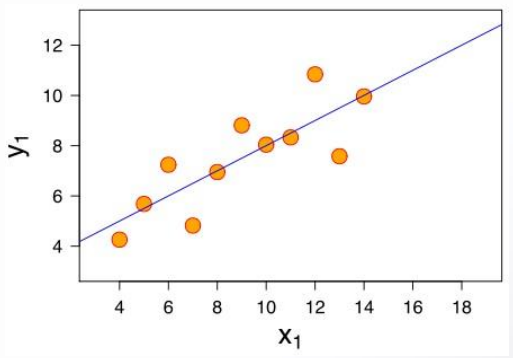
- According to Midori Nediger (2019), "**An infographic is a collection of imagery, charts, and minimal text that gives an easy-to-understand overview of a topic.**"

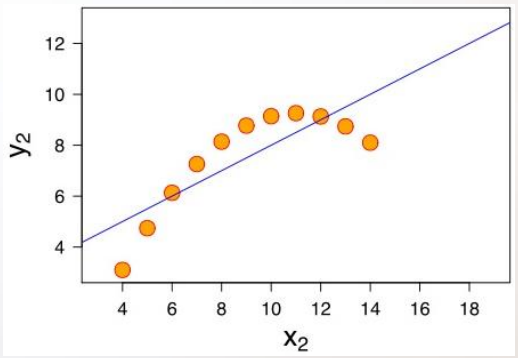


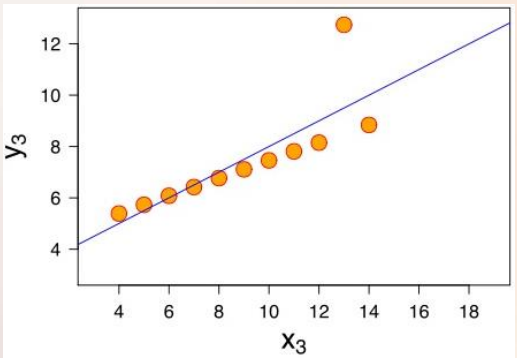

Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

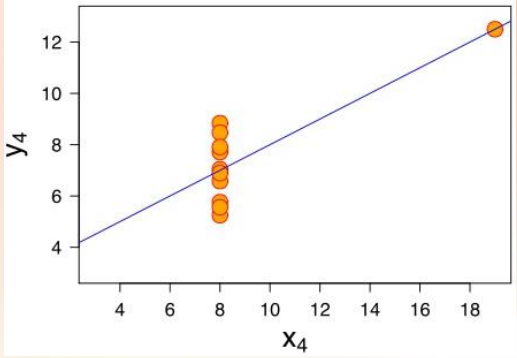
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89







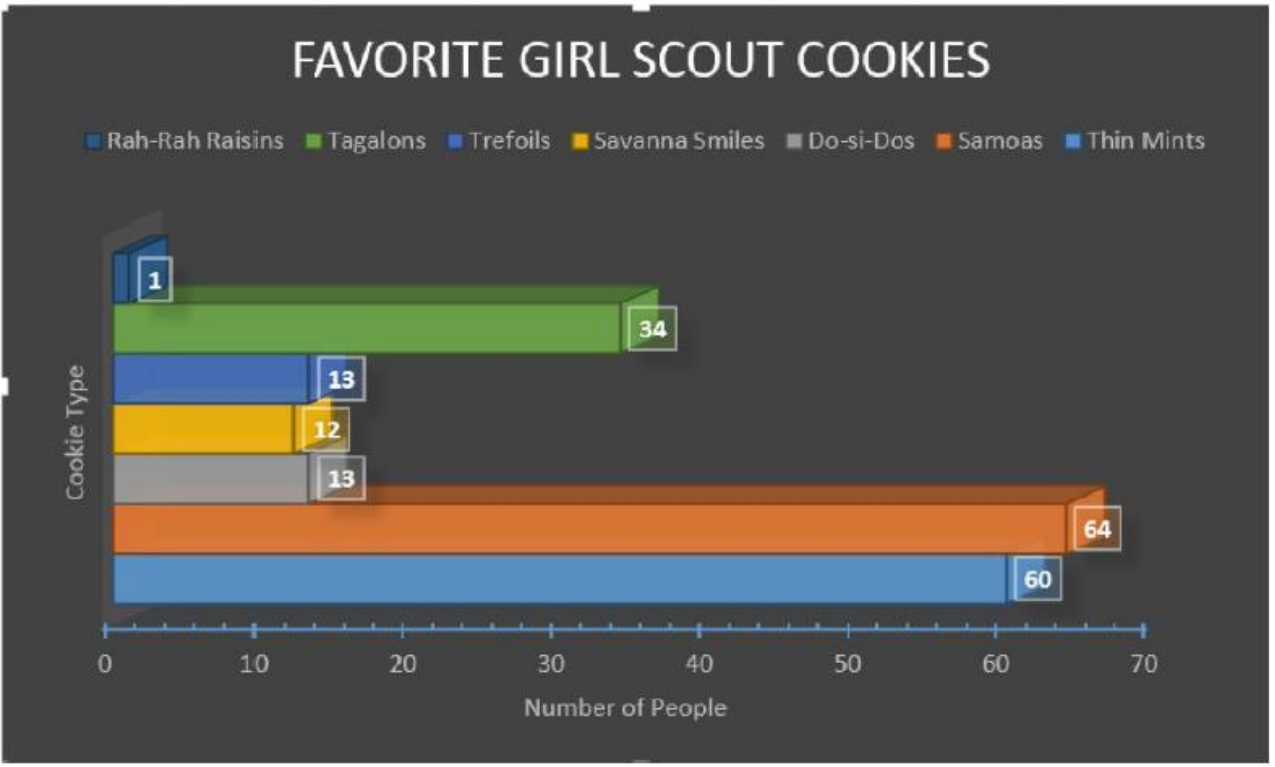


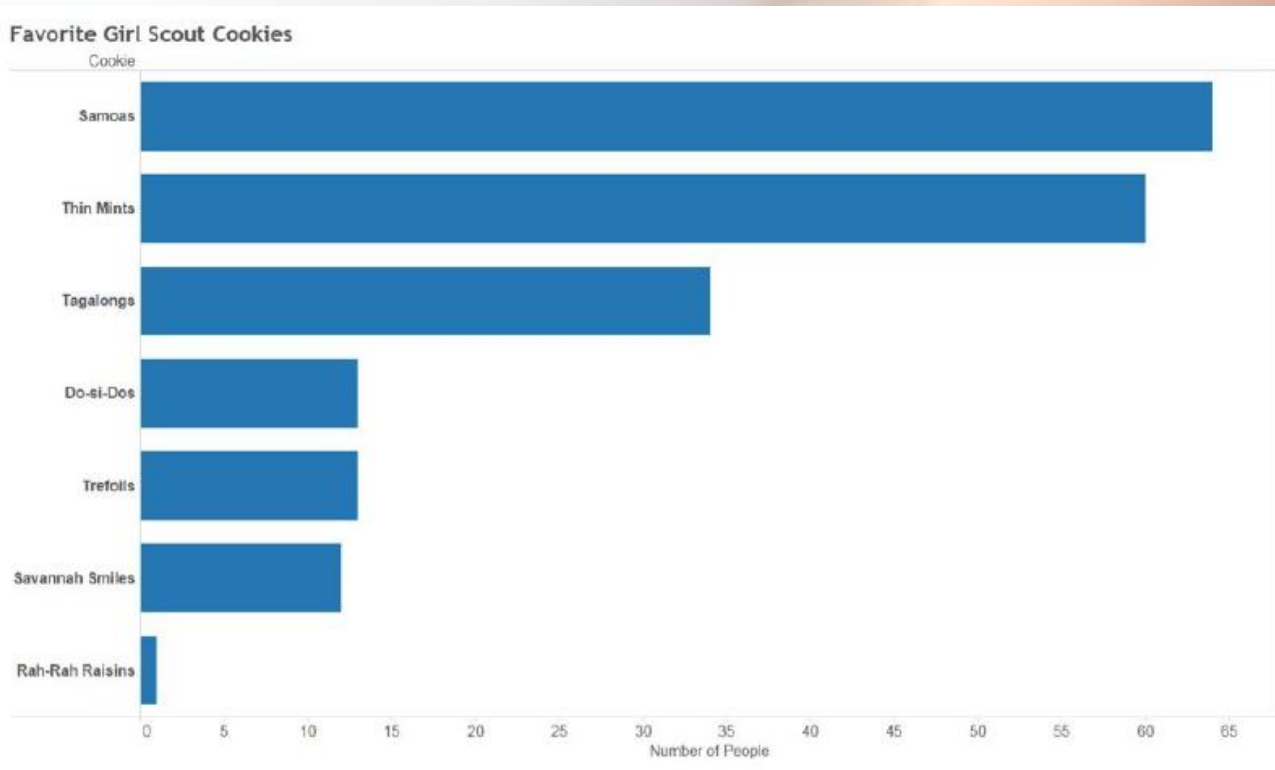
Principles of Data Visualization

Or, being real

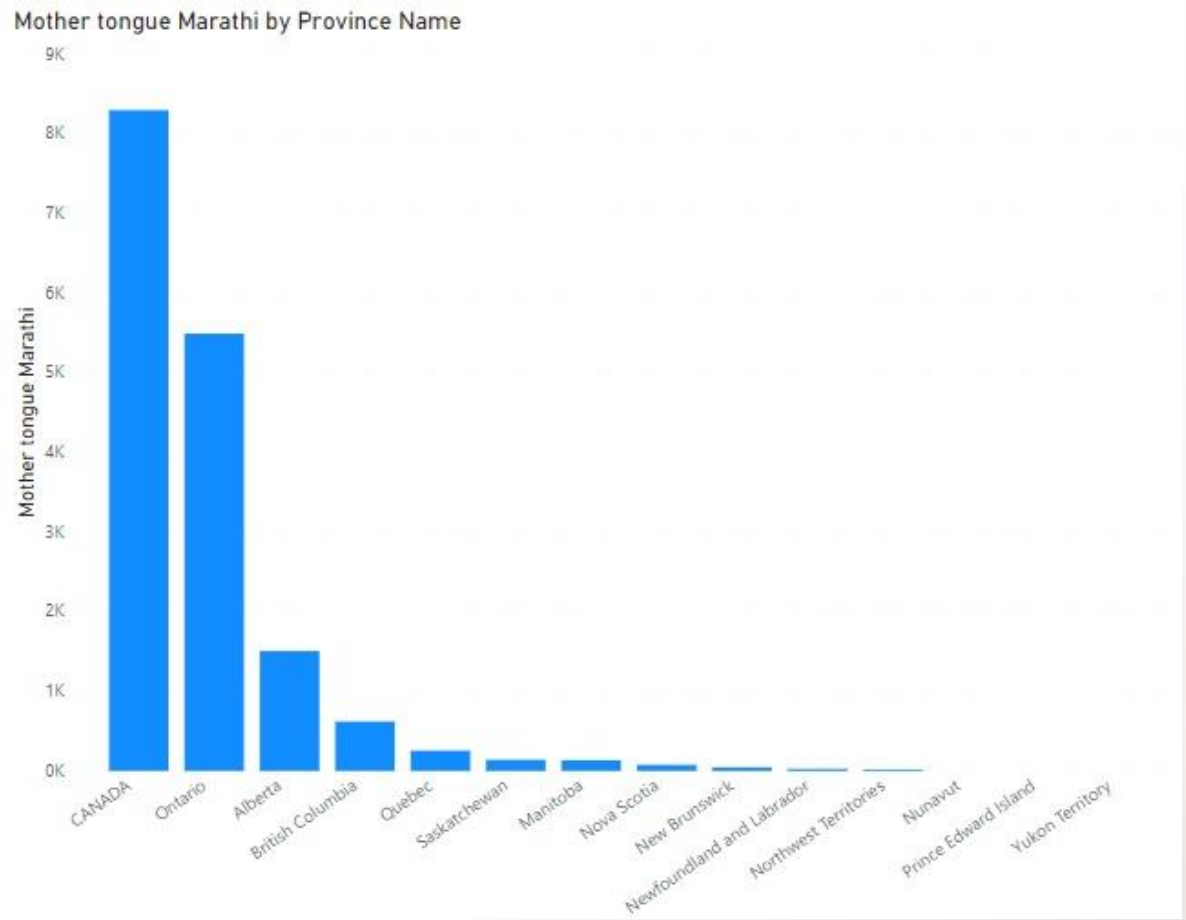
Simpler is better

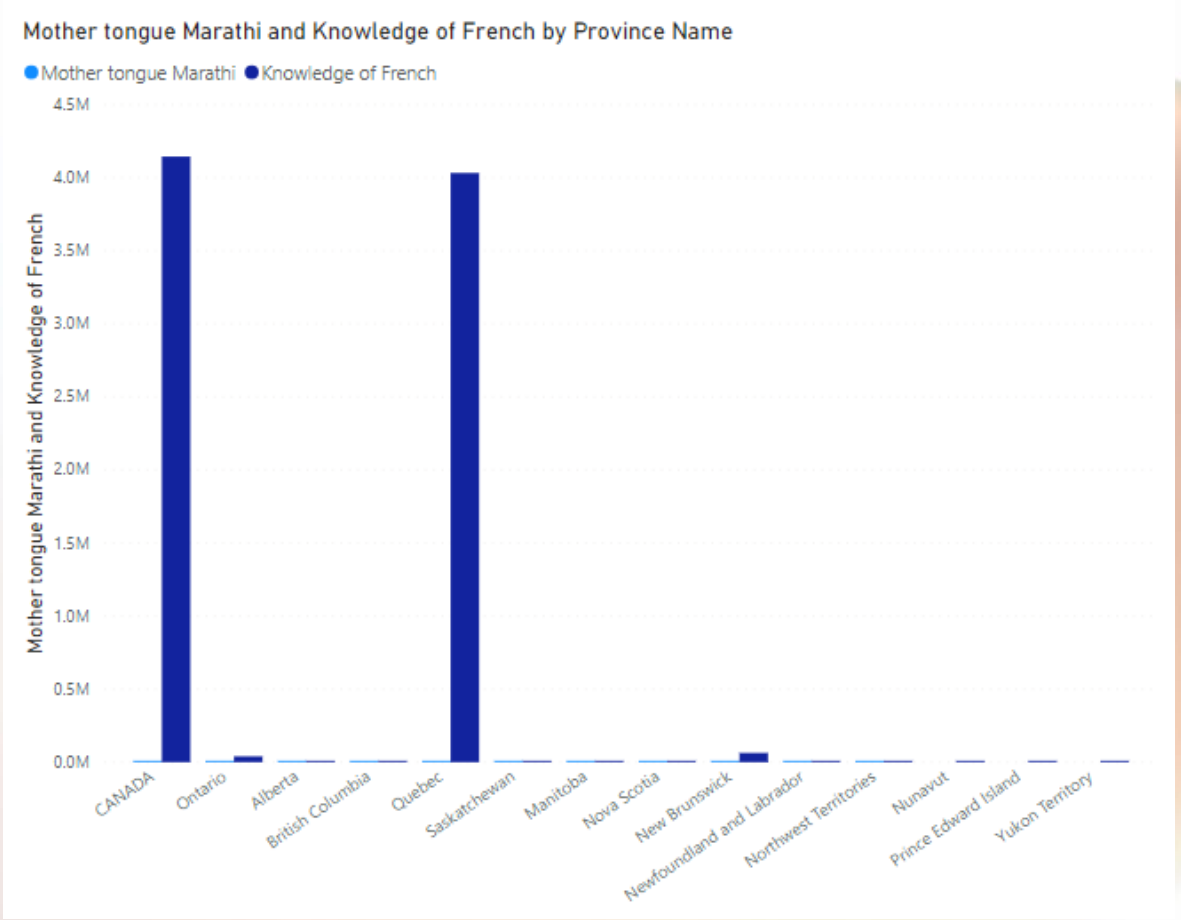
- Which is easier to understand?





Make sure your data works



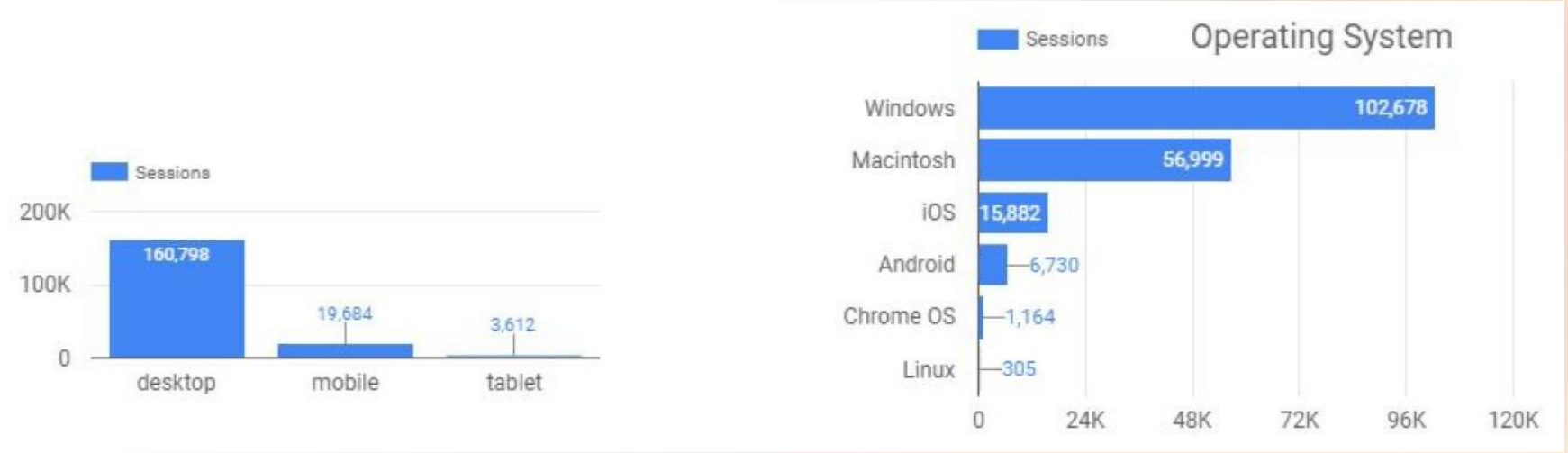


Some basic heuristics

- Bar chart: comparison of things (ex: cookie types)
- Line chart: trends with continuity, such as over time (ex: number of cookies eaten/day)
- Heat map: showing tiers of quantities, especially over space (ex: where I buy the most cookies from?)
- Scatter plot: show particular patterns
- Word cloud: good for showing usage of particular words in text (ex: number of times “cookies” appears on this slide, compared with other words) – similar to heat map

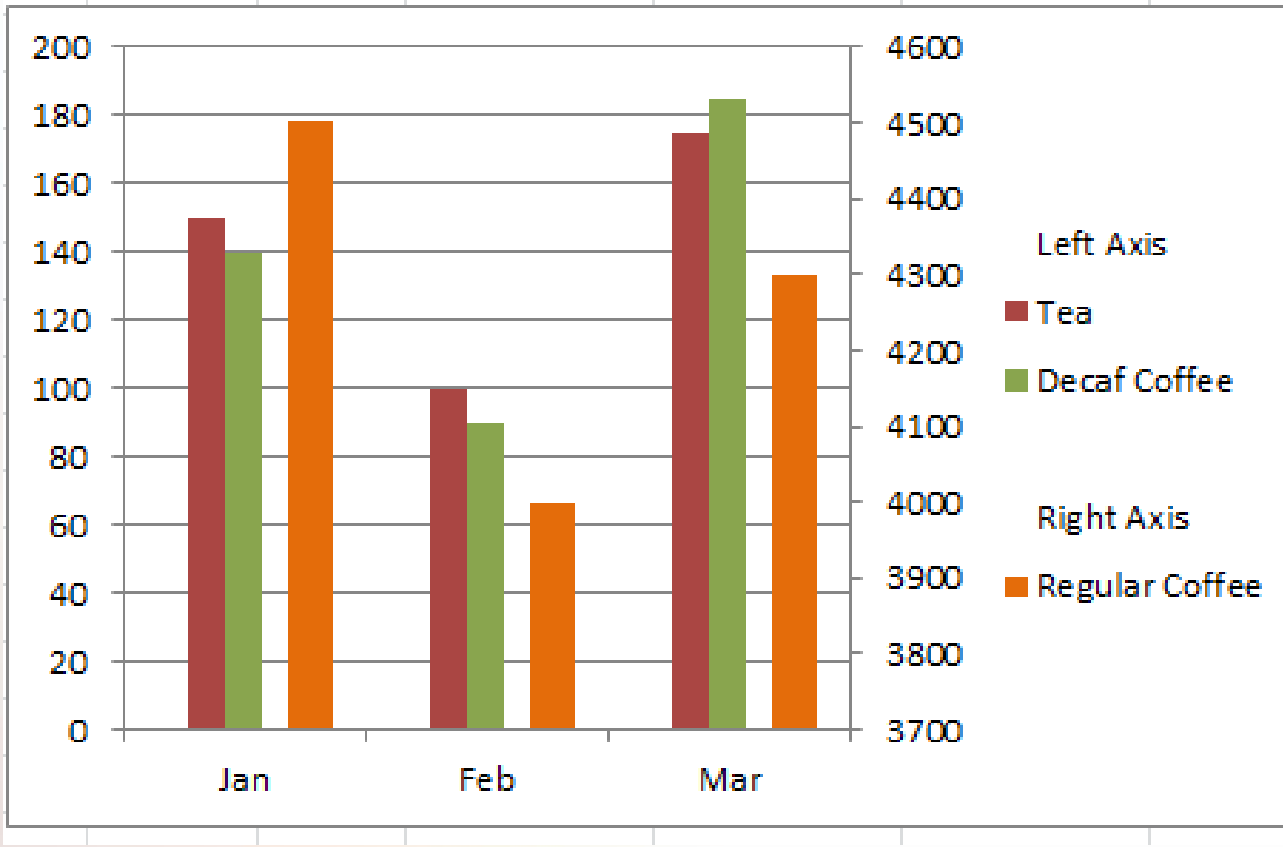
More about bar charts

- Use vertical for a few horizontal for several



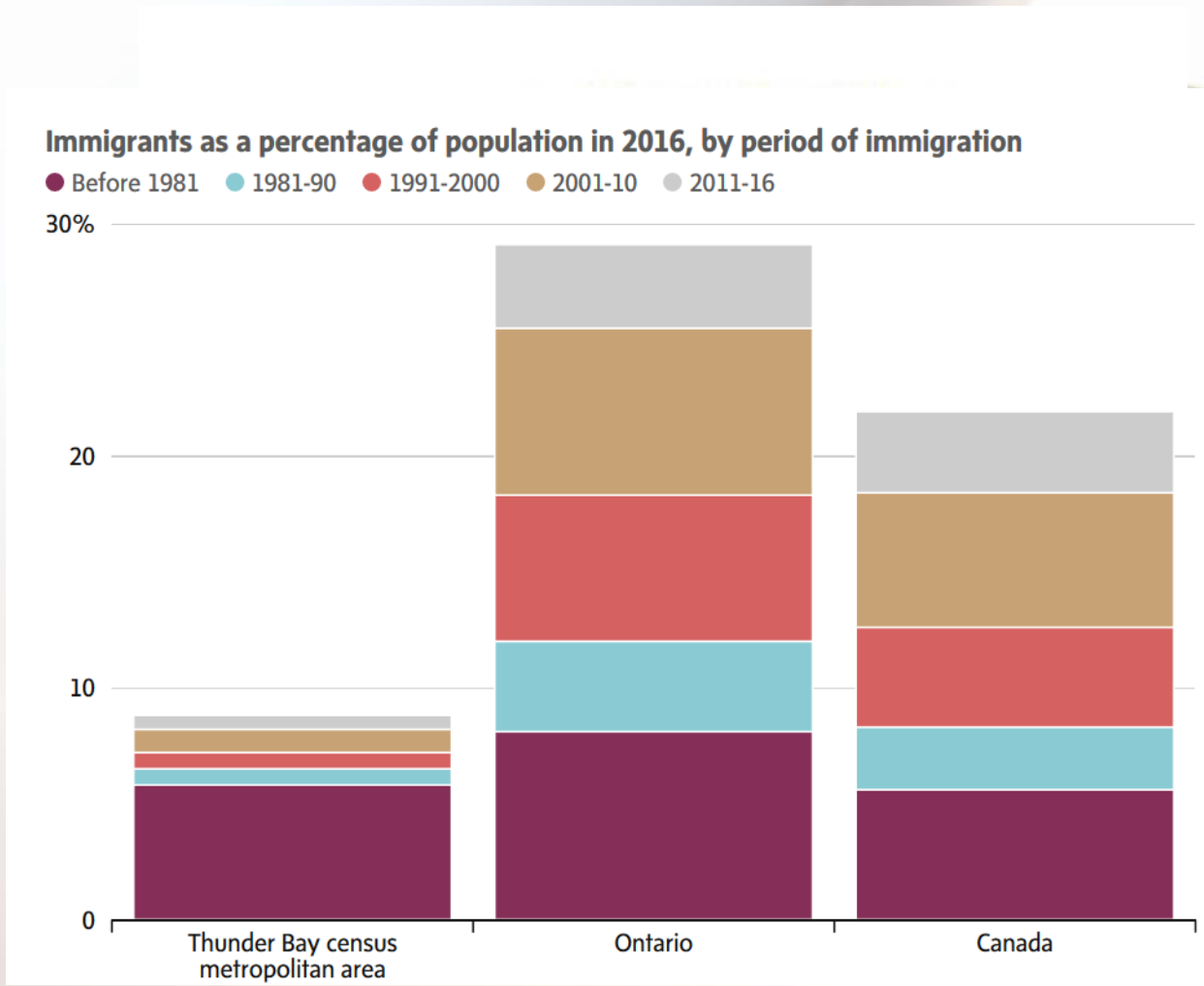
Grouped bar chart

- To show 3 different variables in one chart (but don't have to many groups, or items in a group)
- Never ever include two different variables on the same axis.



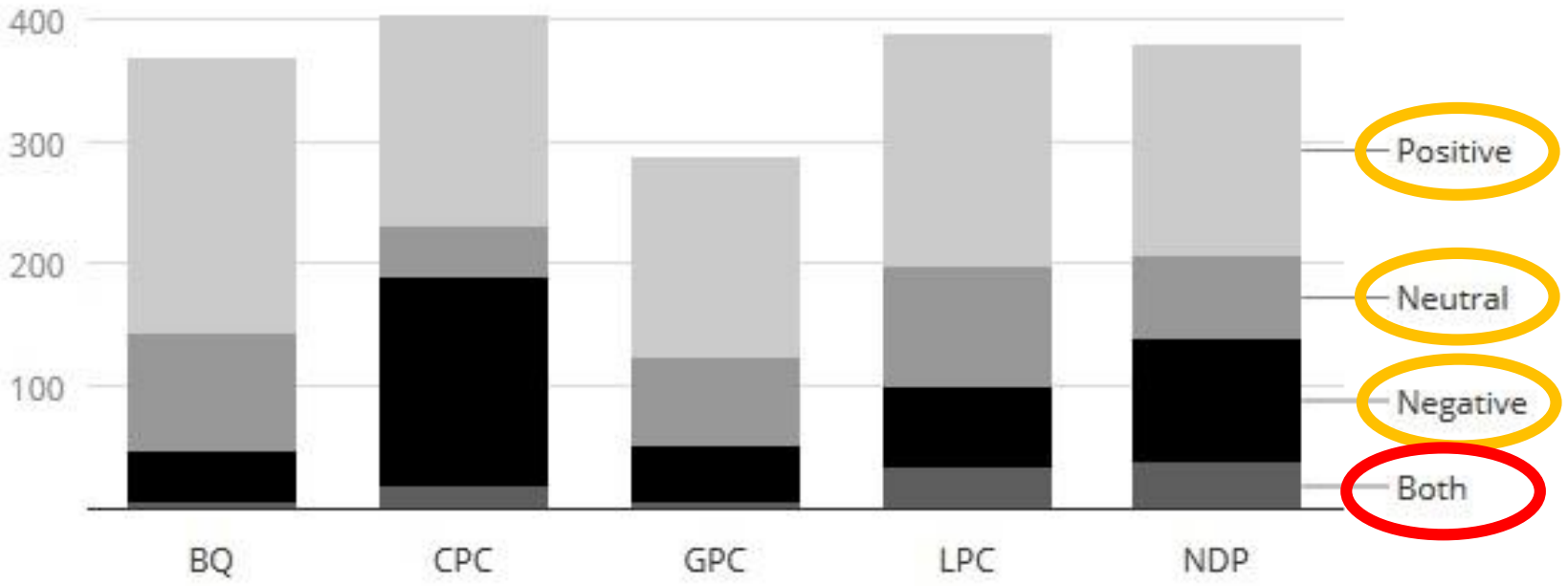
Stacked Bar Charts

- (Almost) always bad: they obscure trends/information/ comparison



Conservatives have run the most negative campaign

Chart shows number and tone of press releases and tweets from official party accounts and leaders from Sept. 11 to October 2, 2019.

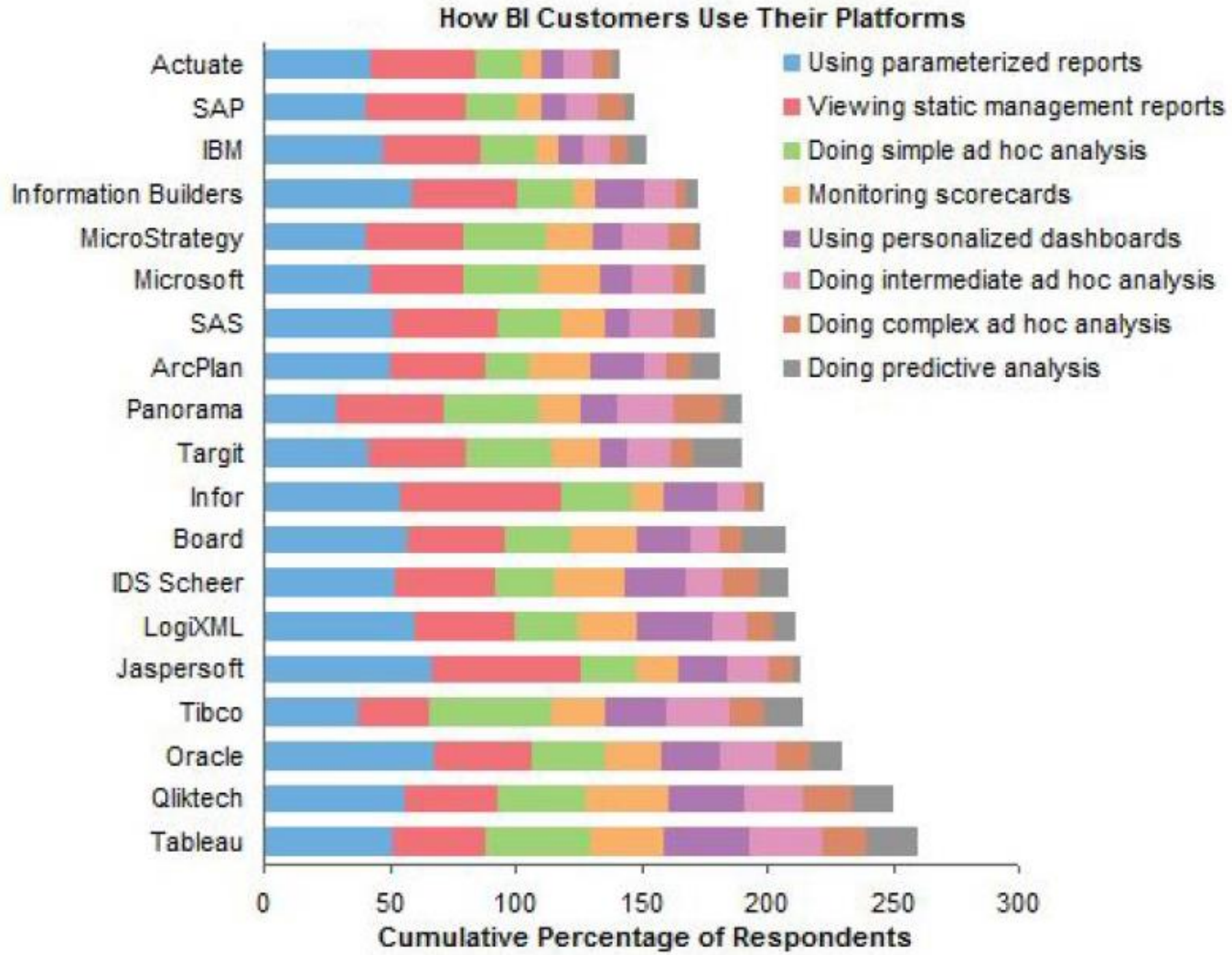


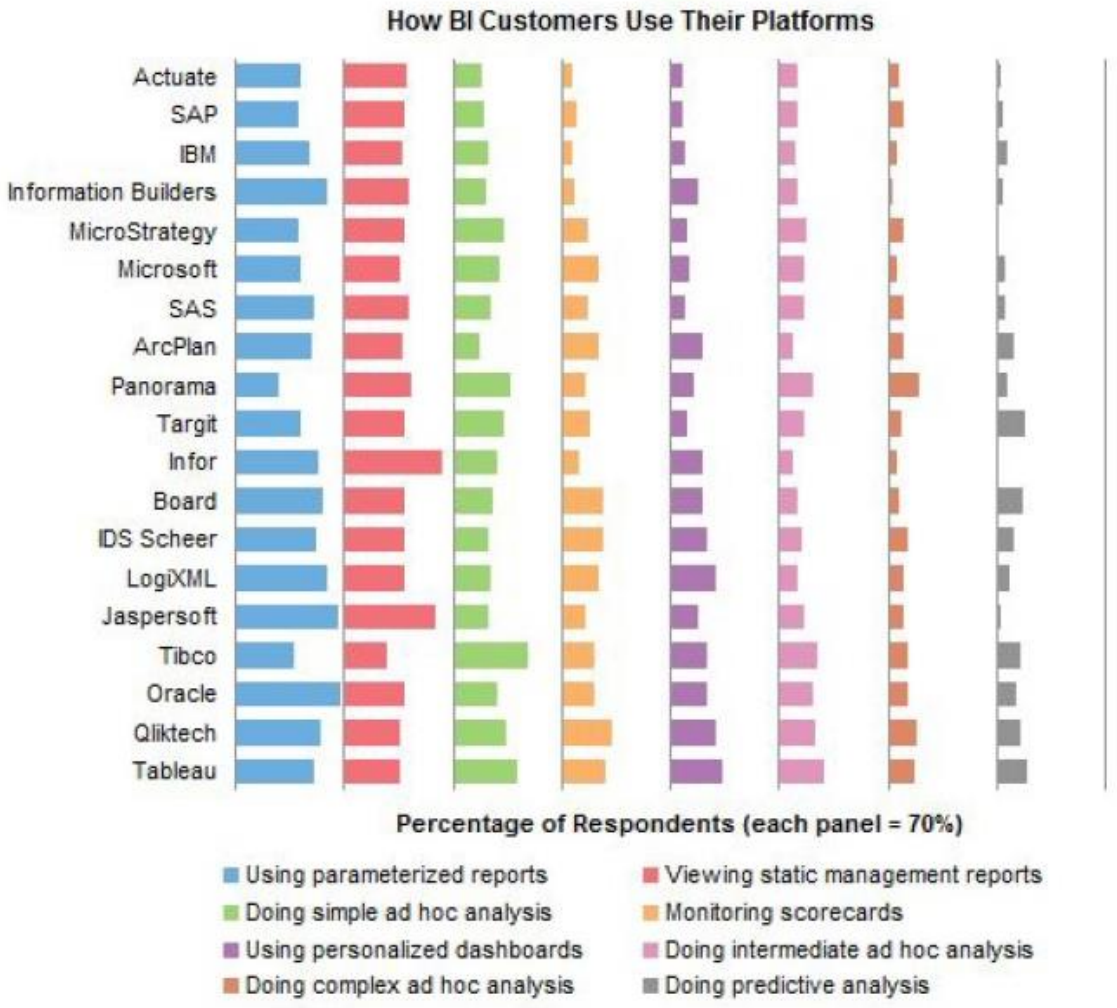
Communications include press releases and tweets from official party accounts and leaders. Does not include retweets. Counts identical tweets in both official languages as one communication.

Chart: Tara Carman • Source: CBC News

news

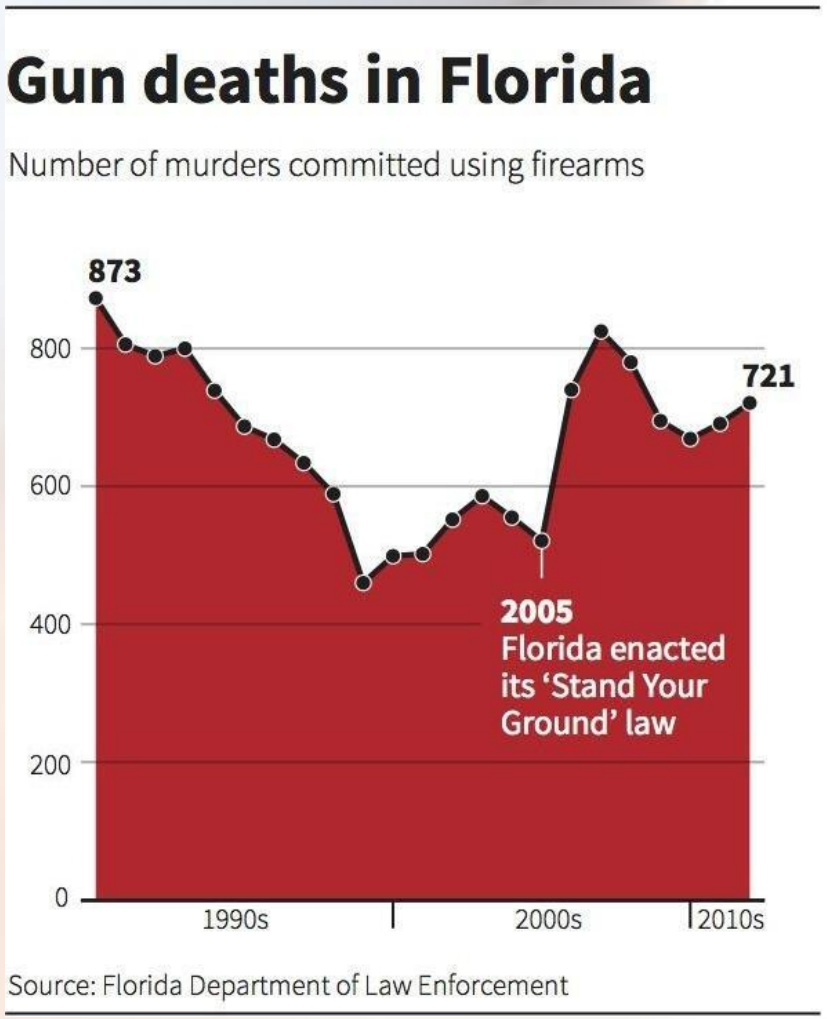
Panel bar charts are better

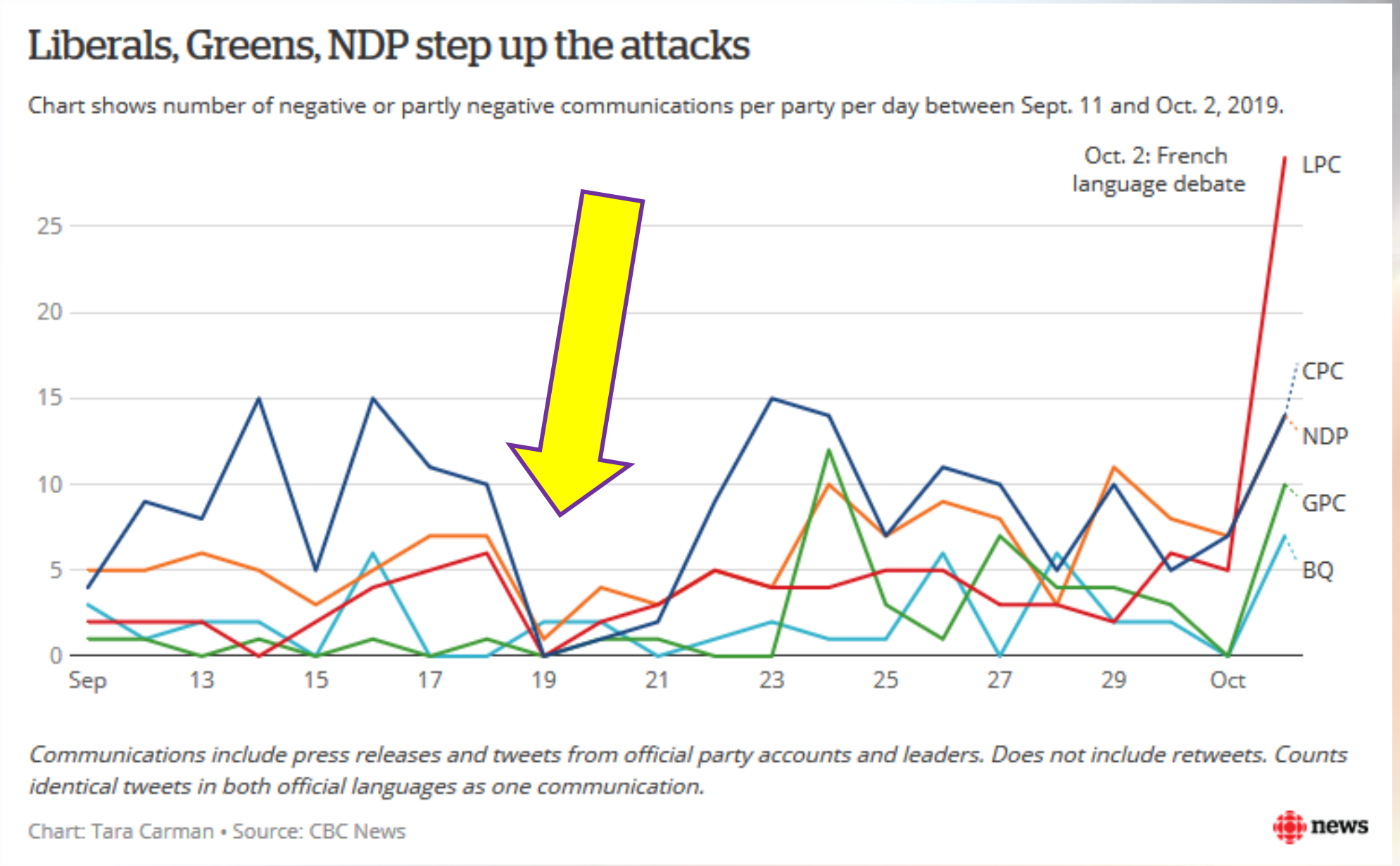




Line charts

- Make sure your data is all related
- Start your Y axis from 0
- If you go negative, show 0
- Up should always increase, down should decrease

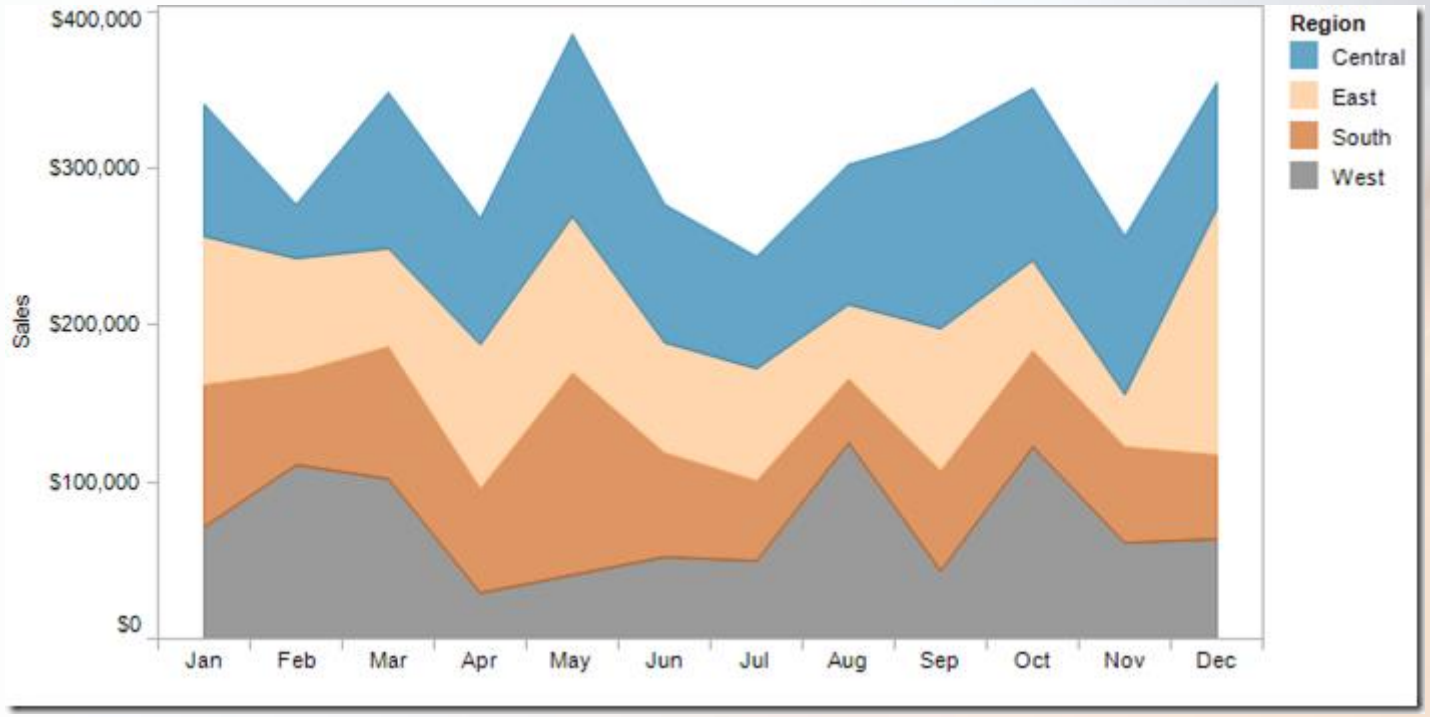




- https://www.cbc.ca/news/politics/liberal-conservative-2019-federal-election-1.5309670

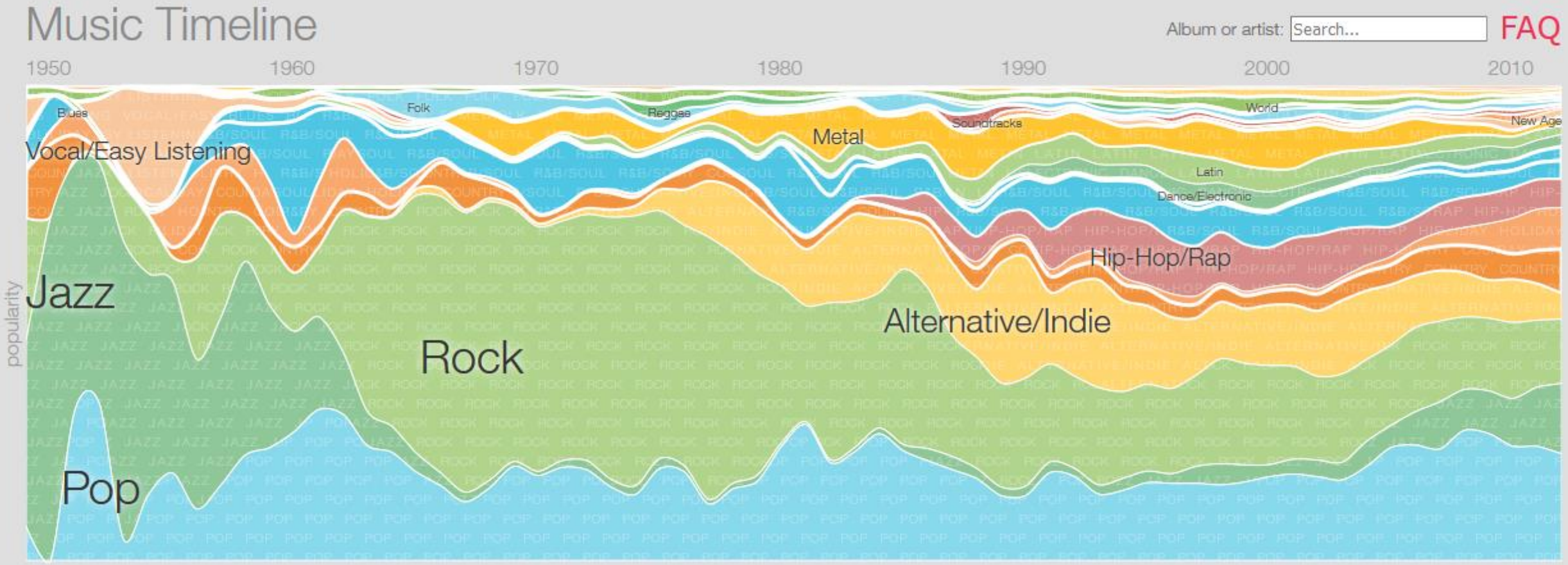
Stacked line graph

- No.



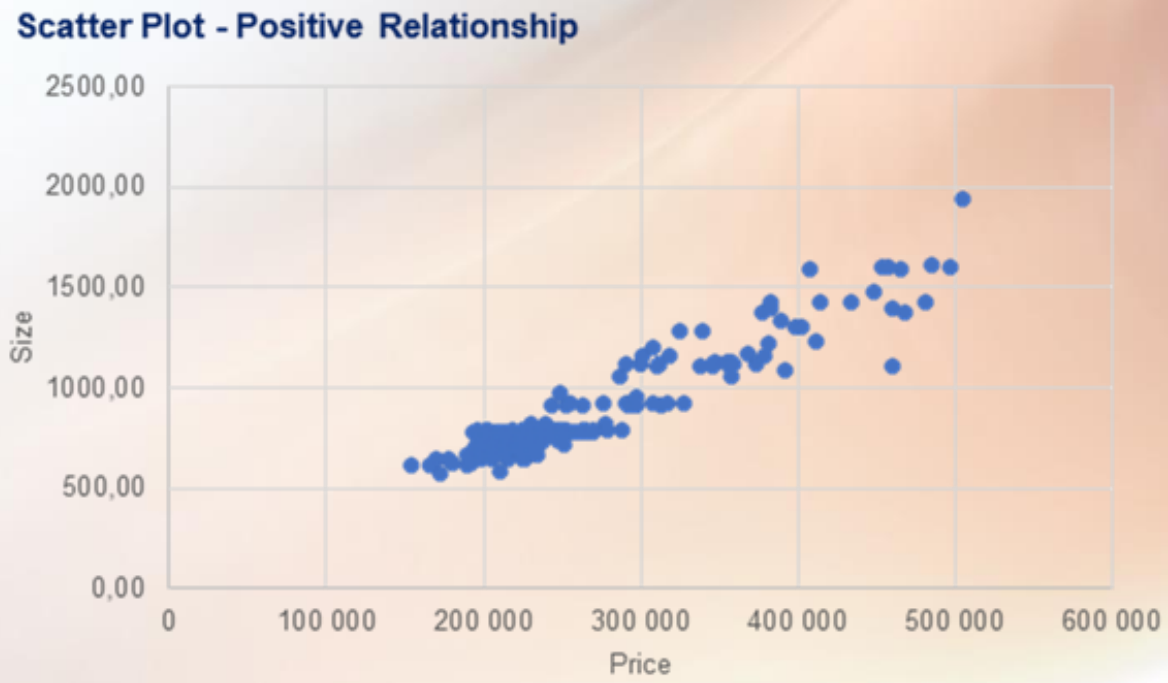
Exception:

- http://research.google.com/bigpicture/music



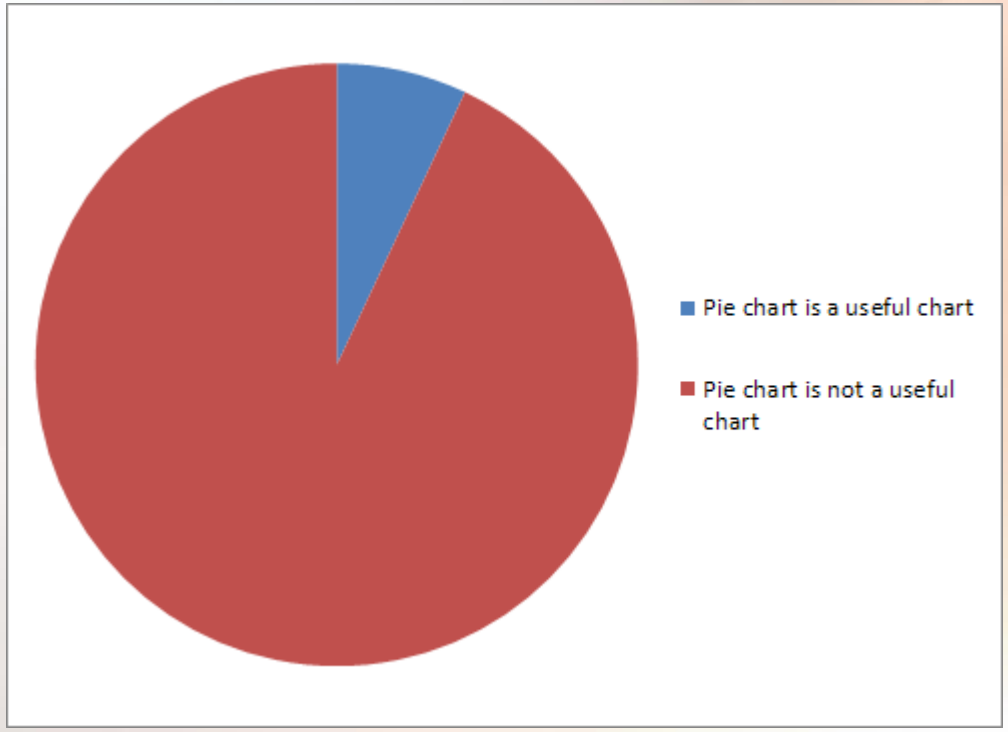
Scatter plots

- Good if your data shows a particular pattern
- However, watch out for blobs

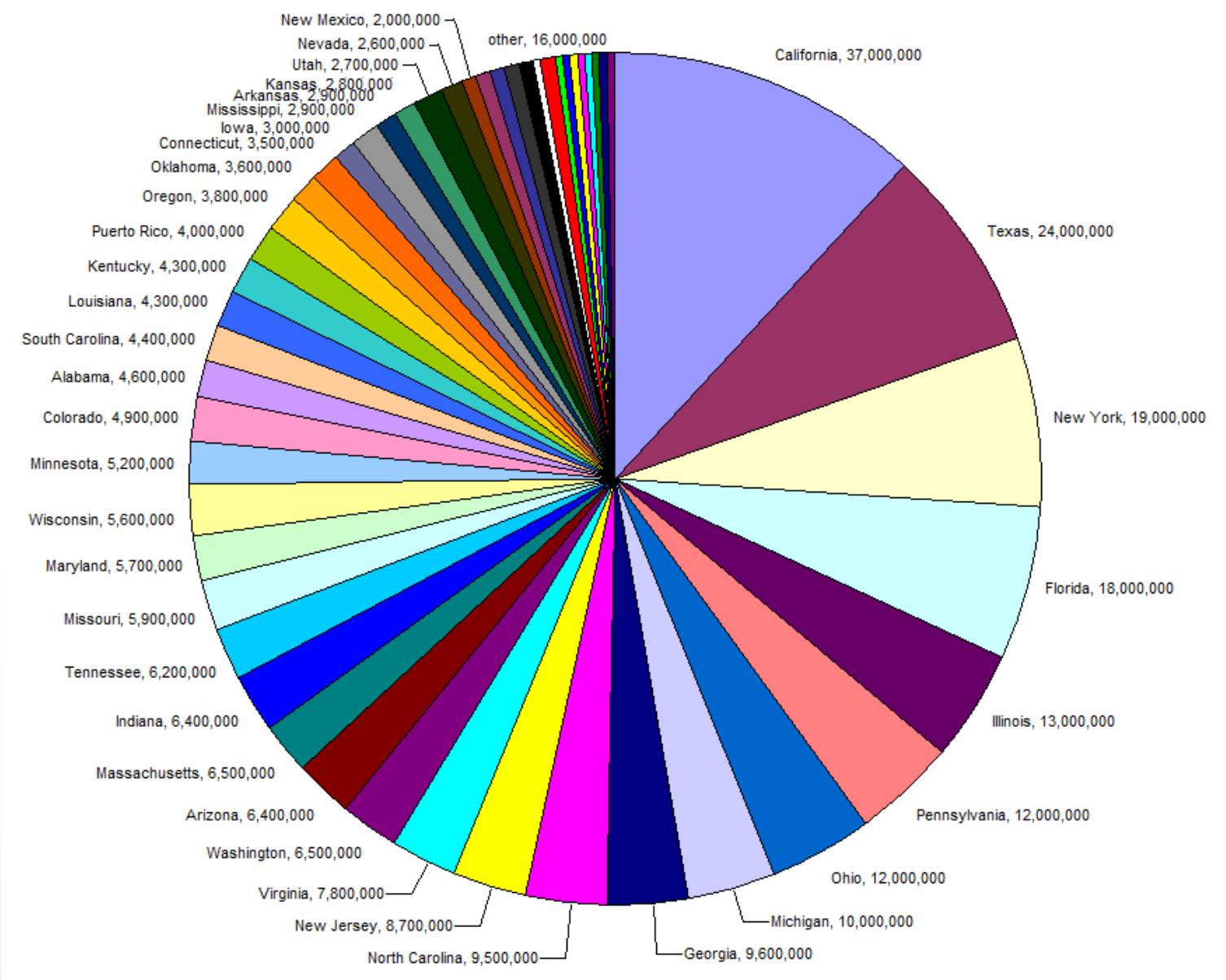


Pie charts

- Occasionally okay
- Use very few slices (max 7, and that's pushing it)
- Label data
- Should generally avoid



U.S. Population by state:



If you really want a pie chart

- Can I interest you in a waffle chart?

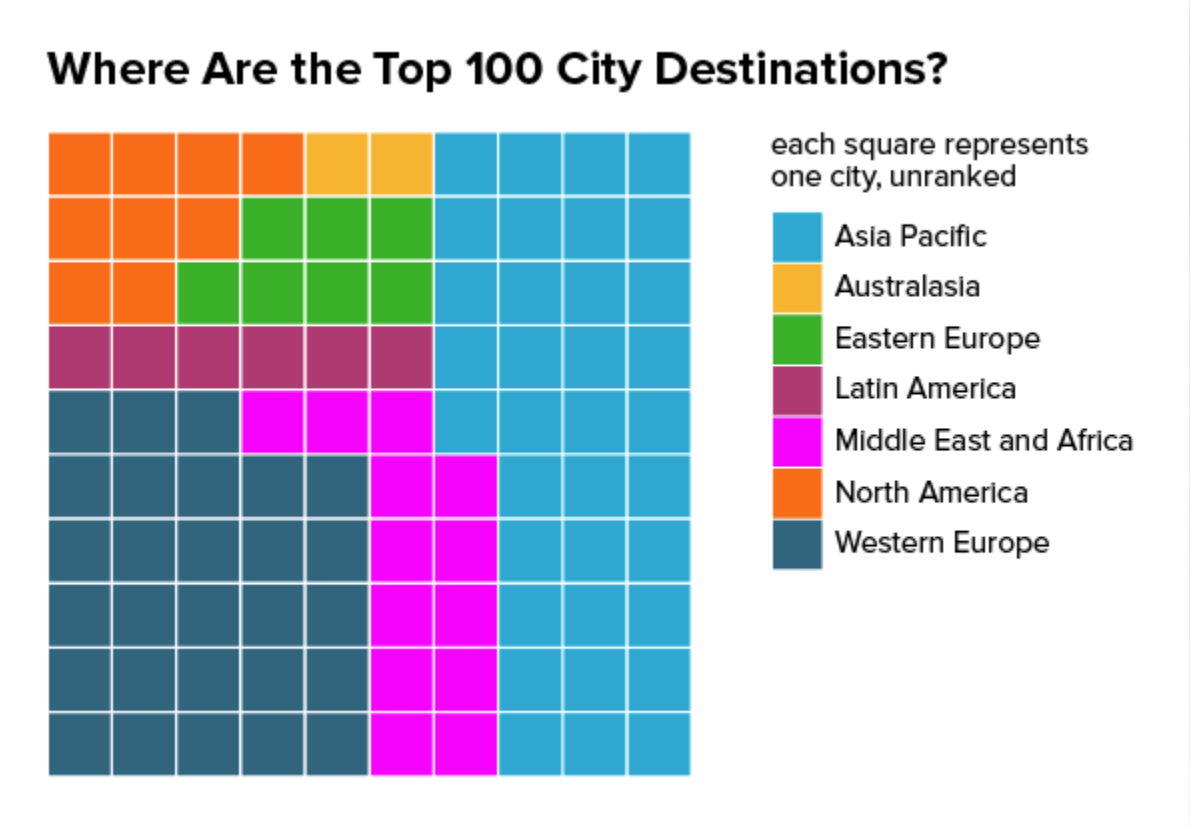
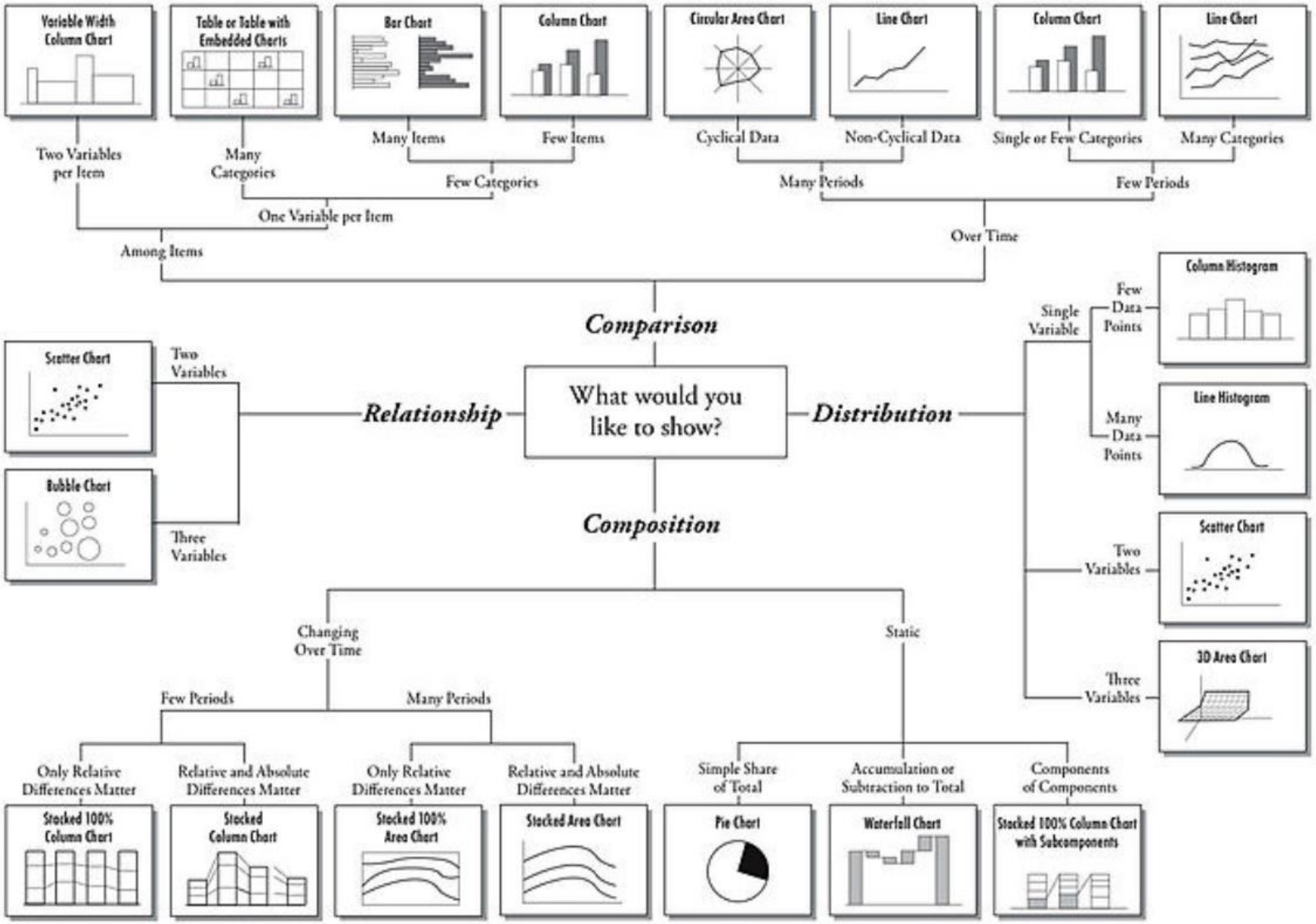


Chart Suggestions—A Thought-Starter



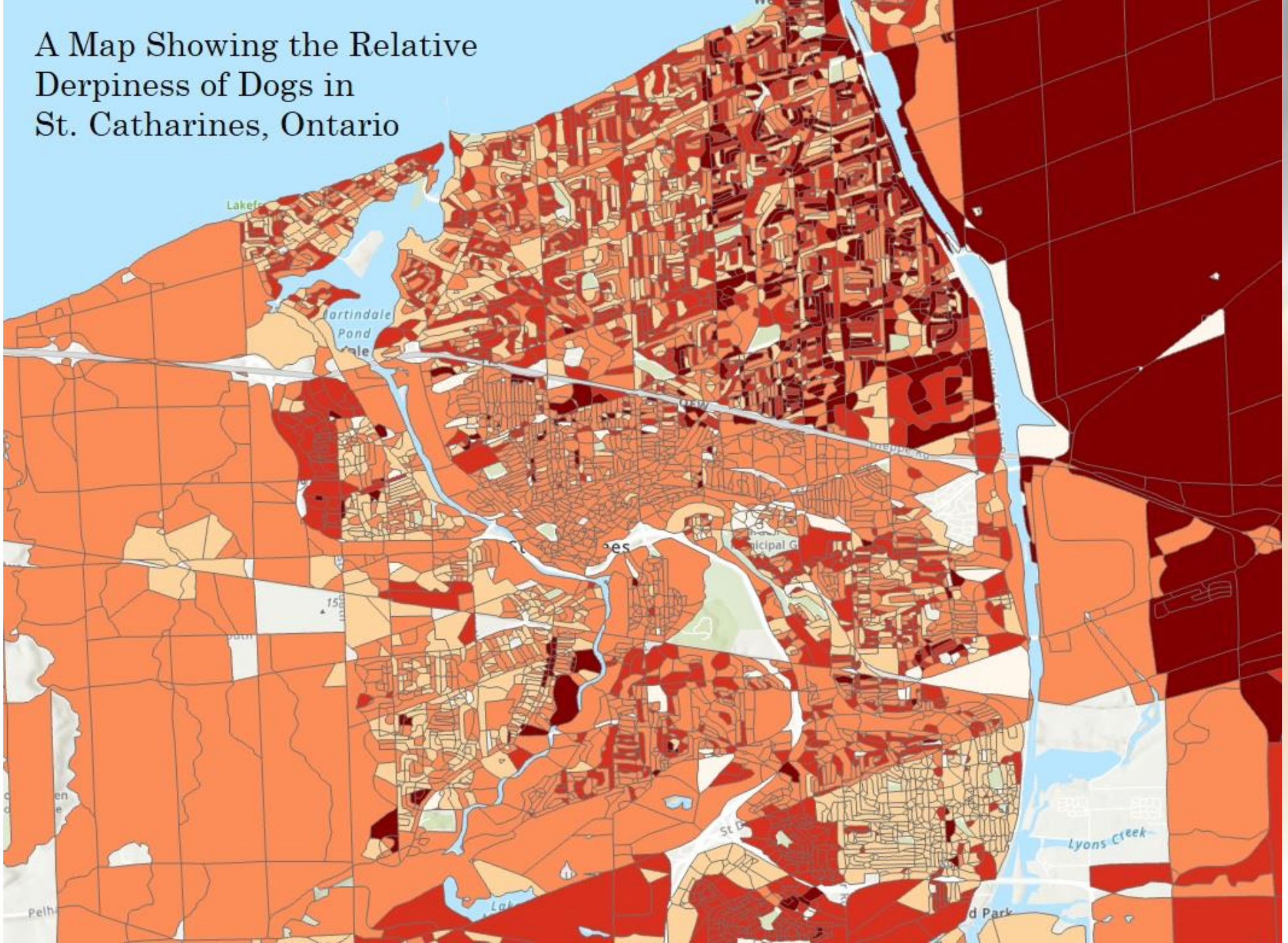
Word cloud

- Not bad, if you're working with a large corpus of text and you can pull interesting things from it
- Remember that it has to show something worthwhile


matthew
i'm
anne
marilla
little
like
said
think
diana
mrs
oh
know
don't
it's
just
going
good
time
home
went
came
girl
things
make
eyes
mr
miss
come
school
right
didn't
lynde
old
say
white
says
look
thought
green
i've
i'd
i'll
told
isn't

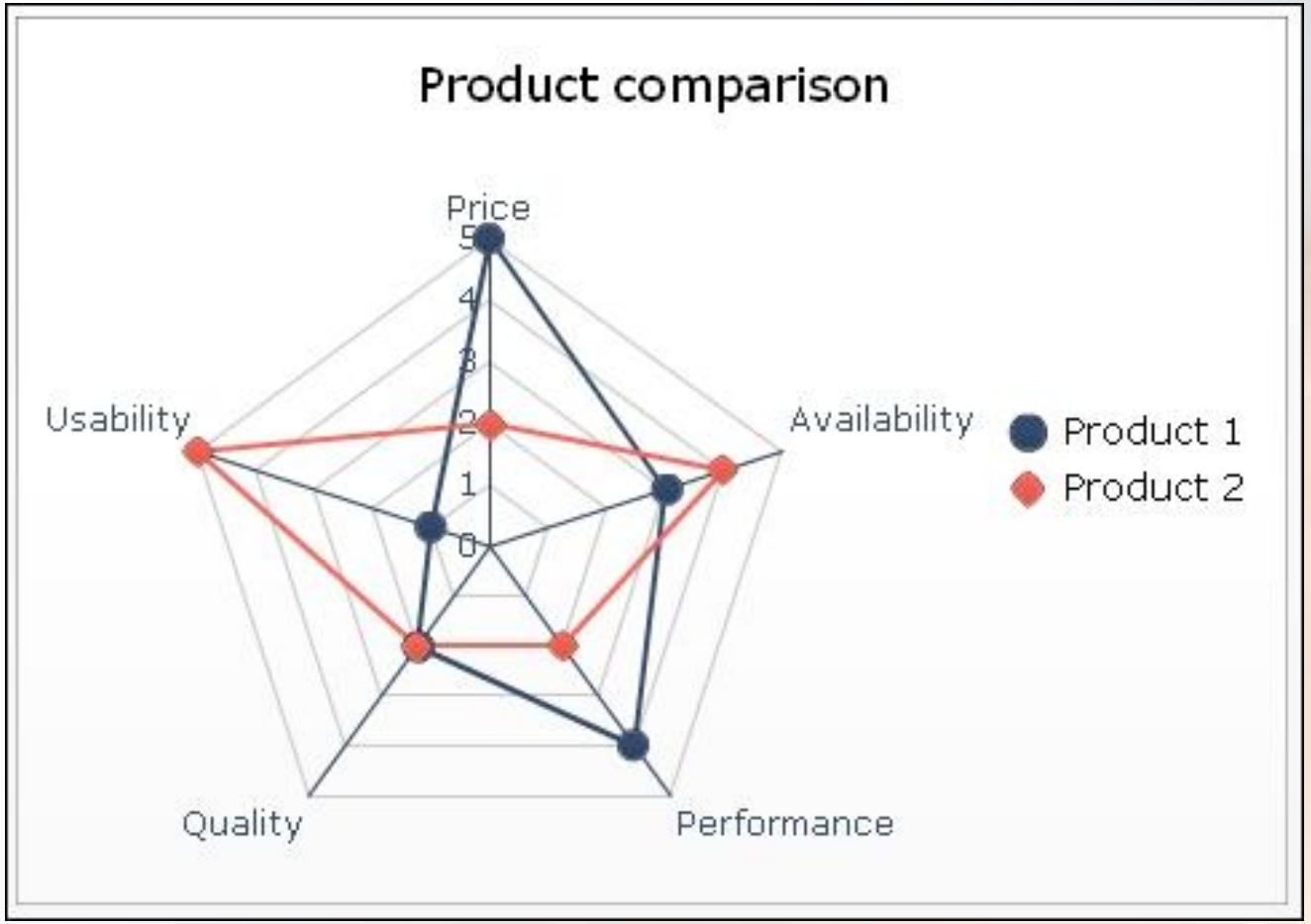
Maps

- People love maps
- Only works if you have a geographic element in your data
- And it has to be a worthwhile geographic element (eg if all of your cases have “Montreal”, yes, technically that’s a geographic element, but it doesn’t give you anything other than a map of Montreal)


A Map Showing the Relative
Derpiness of Dogs in
St. Catharines, Ontario

Radar Charts

- No.



Process for visualization

1. Acquire data
2. Clean data
3. Ingest into tool to play with
4. Play with data
5. Keep playing with data
6. Try different things to make it look good
7. Show it to a friend (preferably one who doesn't know the topic). Do they get it?
8. Profit

Acquiring Data

- Prepared data?
  - Geospatial: https://libraryguides.mcgill.ca/gis_guides
  - Numeric: https://libraryguides.mcgill.ca/data
  - Other: ???
- Collected data?
  - In a format that allows ingestion into a tool - .csv (like a spreadsheet) is the most versatile, but not the only one

Cleaning Data

- Data can be messy
  - E.g.: Say you have data about Montréal. Do you mean the island? The city? The commune in Ardèche? The river in Ontario? The river in Nunavut?
  - Is it about the whole place, or just a part? If you only want a particular section, is it a neighbourhood, or a postal code, or a census area?
  - Or maybe your data is numbers from a survey. Some people wrote it as “10.2”, some as “9,5”, and some as “1,000.4”.
  - Or maybe there’s just some extra information that could skew the results…

Cleaning Data

Cleaning Data

AnneofGreenGables - Notepad

File Edit Format View Help

```
The Project Gutenberg EBook of Anne of Green Gables, by Lucy Maud Montgomery

This eBook is for the use of anyone anywhere in the United States and most
other parts of the world at no cost and with almost no restrictions
whatsoever.  You may copy it, give it away or re-use it under the terms of
the Project Gutenberg License included with this eBook or online at
www.gutenberg.org.  If you are not located in the United States, you'll have
to check the laws of the country where you are located before using this ebook.



Title: Anne of Green Gables

Author: Lucy Maud Montgomery


Release Date: 1992 [EBook #45]
Last Updated: October 6, 2016

Language: English

Character set encoding: UTF-8

*** START OF THIS PROJECT GUTENBERG EBOOK ANNE OF GREEN GABLES ***
```

AnneofGreenGables - Notepad

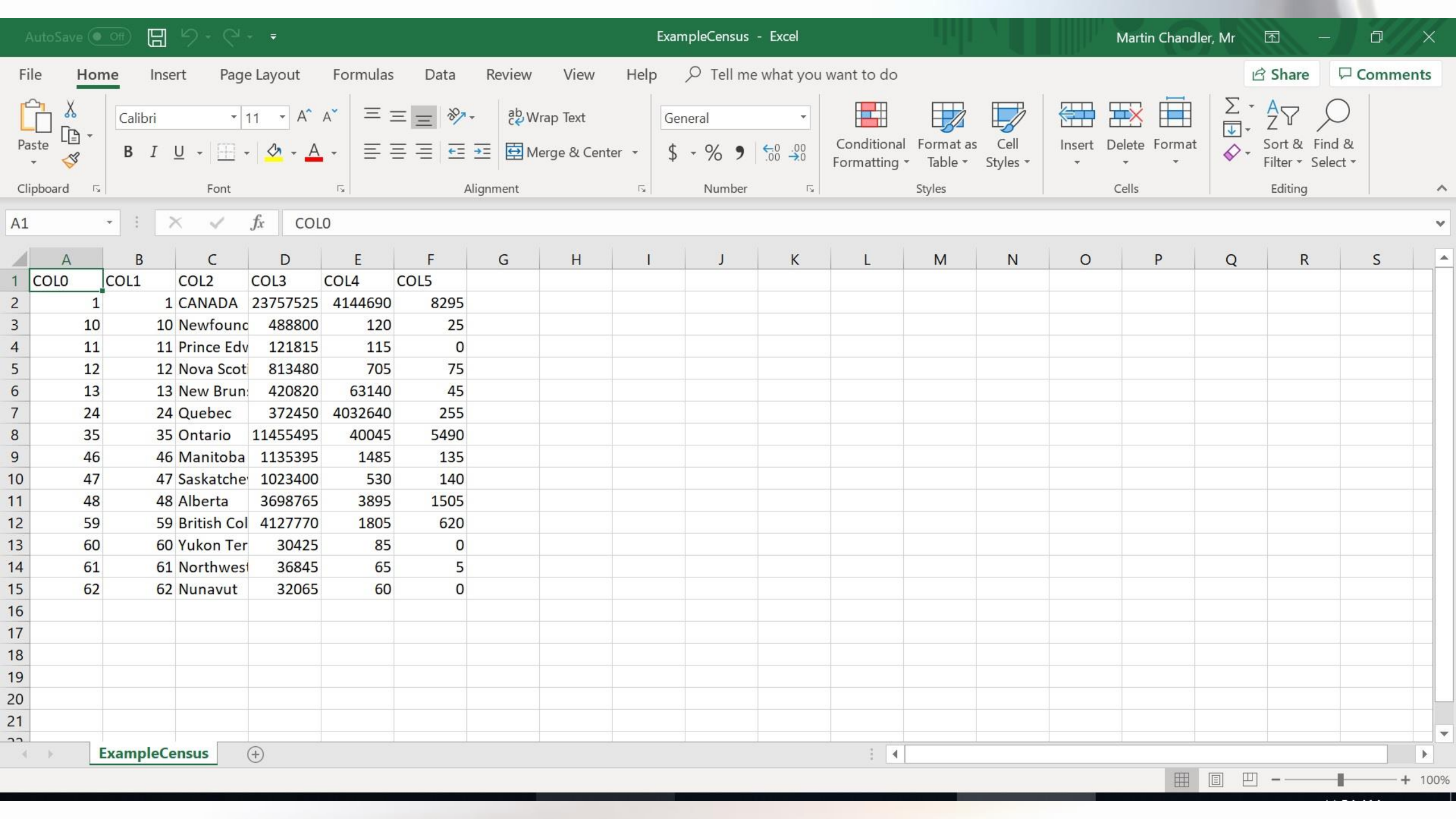
File Edit Format View Help

```
*** END OF THIS PROJECT GUTENBERG EBOOK ANNE OF GREEN GABLES ***

***** This file should be named 45-0.txt or 45-0.zip *****
This and all associated files of various formats will be found in:
        http://www.gutenberg.org/4/45/

Produced by David Widger and Charles Keller

Updated editions will replace the previous one--the old editions will
be renamed.

Creating the works from print editions not protected by U.S. copyright
law means that no one owns a United States copyright in these works,
so the Foundation (and you!) can copy and distribute it in the United
States without permission and without paying copyright
royalties. Special rules, set forth in the General Terms of Use part
of this license, apply to copying and distributing Project
Gutenberg-tm electronic works to protect the PROJECT GUTENBERG-tm
concept and trademark. Project Gutenberg is a registered trademark,
and may not be used if you charge for the eBooks, unless you receive
specific permission. If you do not charge anything for copies of this
eBook, complying with the rules is very easy. You may use this eBook
for nearly any purpose such as creation of derivative works, reports,
performances and research. They may be modified and printed and given
away--you may do practically ANYTHING in the United States with eBooks
```

- Did this cleaning change anything? Not really, in this case, because I only used the 25 most used words to make this.
- Still, data must always be cleaned. Small things can throw the whole thing off.

Some Visualization Tools

- PowerBI - General
- Tableau - General
- Raw - General
- Microsoft Excel
- Voyant - Text
- ArcGIS - Geographic
- QGIS - Geographic
- Stata - Statistics
- And many more!

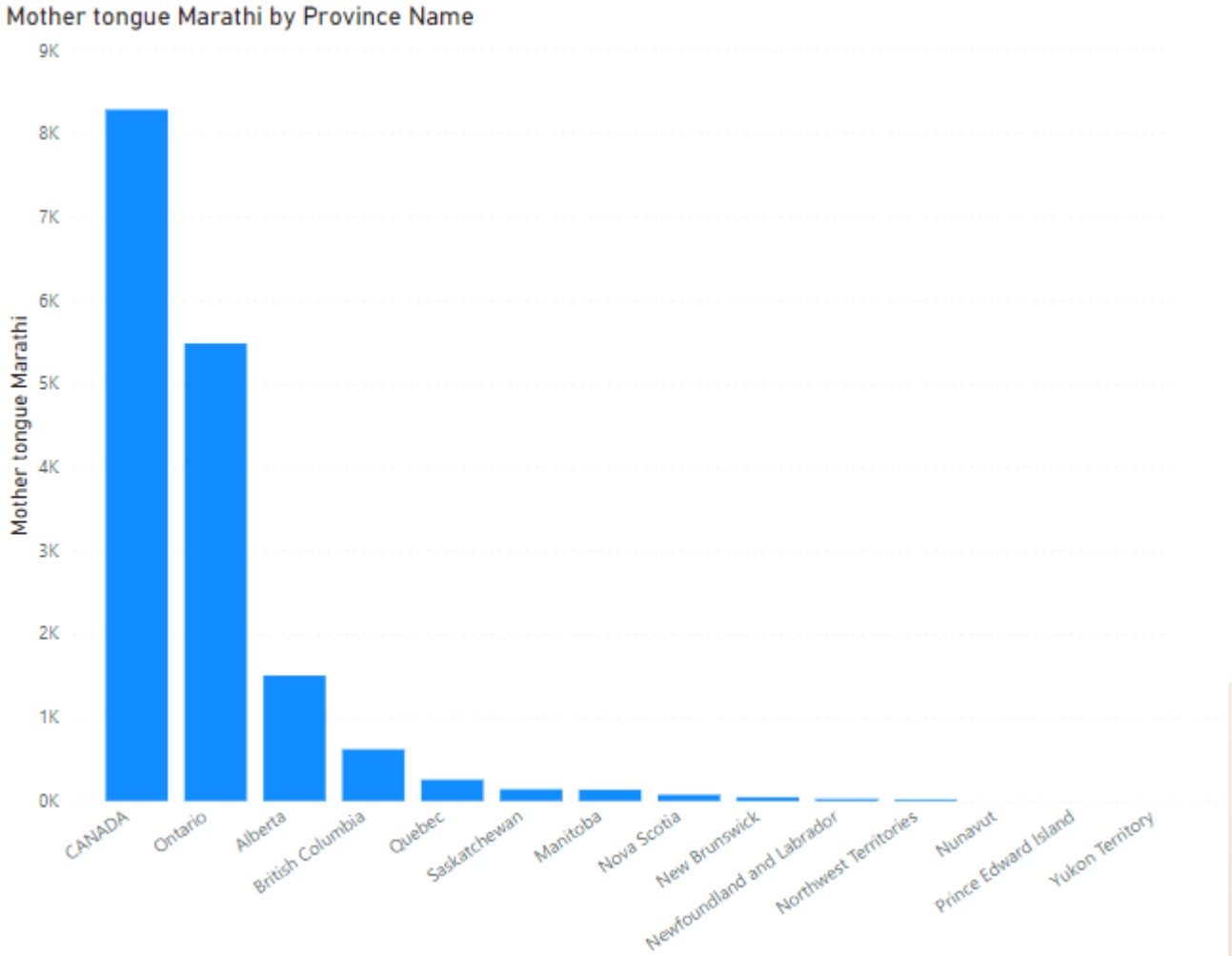


	A	B	C	D	E	F
1	COL0	COL1	COL2	COL3	COL4	COL5
2	1	1	CANADA	23757525	4144690	8295
3	10	10	Newfounc	488800	120	25
4	11	11	Prince Edv	121815	115	0
5	12	12	Nova Scot	813480	705	75
6	13	13	New Brun	420820	63140	45
7	24	24	Quebec	372450	4032640	255
8	35	35	Ontario	11455495	40045	5490
9	46	46	Manitoba	1135395	1485	135
10	47	47	Saskatche	1023400	530	140
11	48	48	Alberta	3698765	3895	1505
12	59	59	British Col	4127770	1805	620
13	60	60	Yukon Ter	30425	85	0
14	61	61	Northwest	36845	65	5
15	62	62	Nunavut	32065	60	0

- Make a copy of your master file before you do anything
- Then:

	A	B	C	D	E	F
1	GEO UID	Province Code	Province Name	Knowledge of English	Knowledge of French	Mother tongue Marathi
2	1	1	CANADA	23757525	4144690	8295
3	10	10	Newfoundland and Labrador	488800	120	25
4	11	11	Prince Edward Island	121815	115	0
5	12	12	Nova Scotia	813480	705	75
6	13	13	New Brunswick	420820	63140	45
7	24	24	Quebec	372450	4032640	255
8	35	35	Ontario	11455495	40045	5490
9	46	46	Manitoba	1135395	1485	135
10	47	47	Saskatchewan	1023400	530	140
11	48	48	Alberta	3698765	3895	1505
12	59	59	British Columbia	4127770	1805	620
13	60	60	Yukon Territory	30425	85	0
14	61	61	Northwest Territories	36845	65	5
15	62	62	Nunavut	32065	60	0
16						

Play with Data



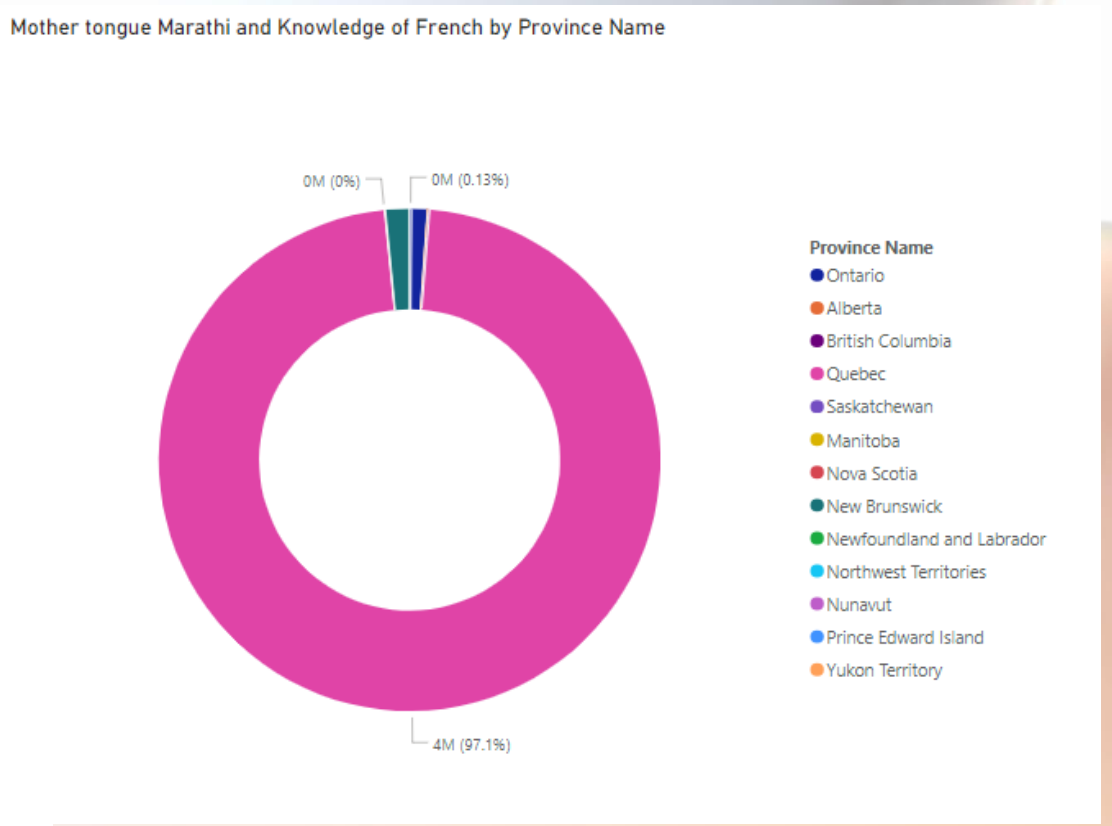


Tableau Public - Book1

File Data Worksheet Dashboard Story Analysis Map Format Window Help

Show Me

Data | Analytics

ExampleCensus1

Dimensions

- Province Code
- Province Name
- *Measure Names*

Measures

- GEO UID
- Knowledge of English
- Knowledge of French
- Mother tongue Marathi
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

Filters

- Province Name
- SUM(Mother tongue ..

Marks

Automatic

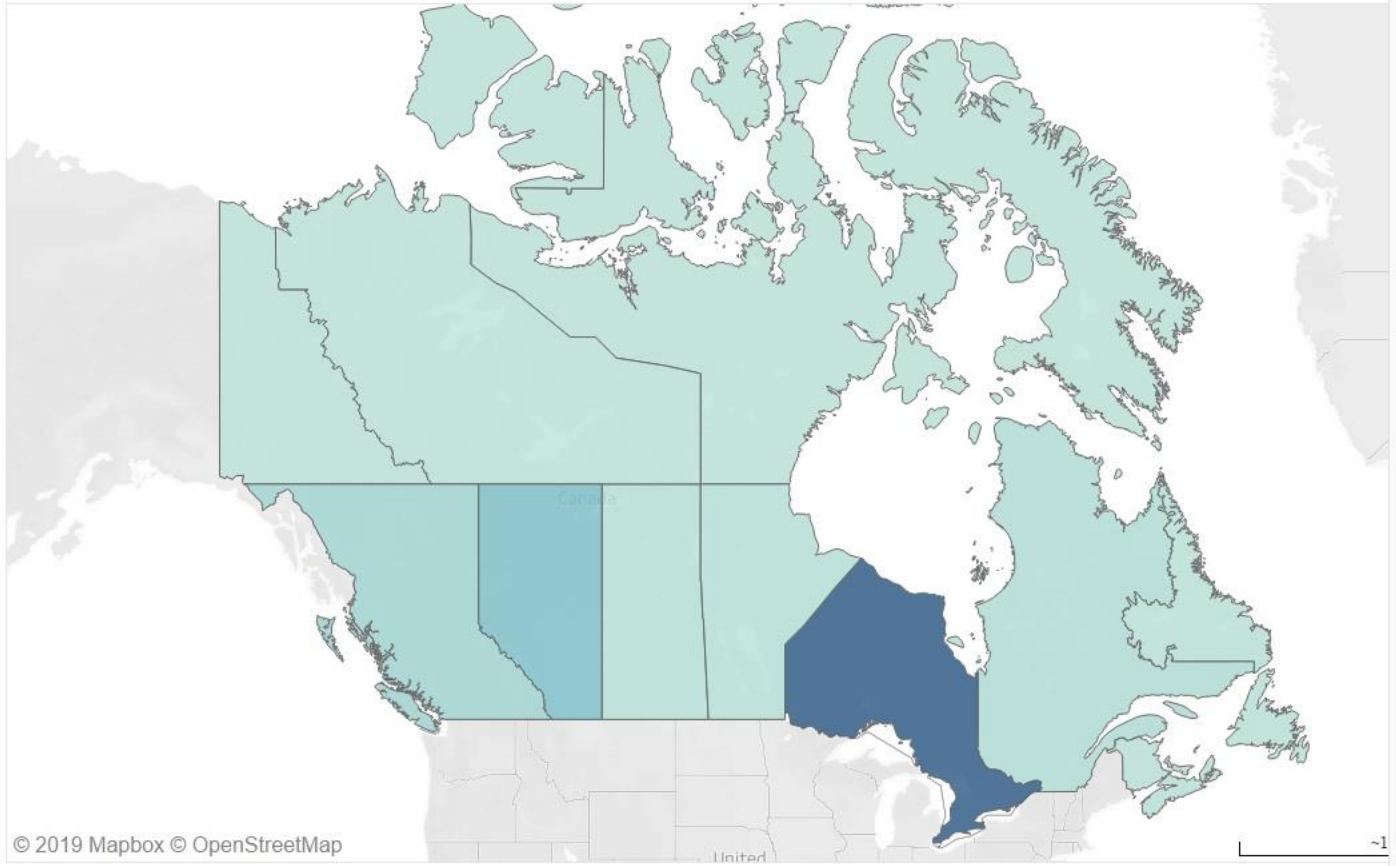
Color | Size | Label | Detail | Tooltip

- Province Name
- SUM(Mother t..

Columns: Longitude (generated)

Rows: Latitude (generated)

Sheet 1



For **symbol maps** try

1 geo Dimension

0 or more Dimensions

0 to 2 Measures

May use spatial measure in place of geo dimension

Data Source | Sheet 1

13 marks | 1 row by 1 column | SUM(Mother tongue Marathi): 8,295

Let’s look at some software

- Power BI
- Tableau
- Voyant
- ArcGIS

Citing Data - Numeric

- **Author or creator** - name(s) of each individual or organizational entity responsible for the creation of the dataset
- **Date of publication** - the year the dataset was published, posted or otherwise released to the public
- **Title or description** - complete title of the dataset **or** if no title exists, you must create a brief description of the data, including time period covered in the data as well as edition or version number, if applicable
- **Publisher and/or distributor** - organizational entity that makes the dataset available by archiving, producing, publishing, and/or distributing the dataset
- **Electronic Location or Identifier** - Web address or unique, persistent, global identifier used to locate the dataset (such as a DOI). Append the date retrieved if the title and locator are not specific to the exact instance of the data you used.
- (Sometimes) Format description e.g. data file, database, CD-ROM, computer software

- http://libraryguides.mcgill.ca/datacitation

Citing Data - Geospatial

More dependent, but generally:

- Author or Data Producer
- Map Title or scene
- Edition
- Scale
- [Series, Sheet number]
- Place of publication: publisher
- Date
- https://acmla-acacc.ca/docs/ACMLA_BestPracticesCitations.pdf

Most important:

- What are you trying to show?
- Is your visualization showing that?
- Is your visualization showing that well?

Some resources:

- https://library.duke.edu/data/data-visualization
- Steele, Julie, and Noah P. N Iliinsky. *Beautiful Visualization : [Looking at Data through the Eyes of Experts]*. 1st ed., O'Reilly, 2010. https://mcgill.on.worldcat.org/oclc/471816105
- Bertin, Jacques, and William J Berg. *Semiology of Graphics : Diagrams, Networks, Maps*. 1st ed., ESRI Press, 2011. https://mcgill.on.worldcat.org/oclc/656556106

Questions?

Martin Chandler
martin.chandler@mcgill.ca

Sources

- *Anscombe's quartet.* (n.d.). Retrieved 1 Oct 2019 from https://en.wikipedia.org/wiki/Anscombe%27s_quartet
- Belvadi, M. (2017). *Data Visualization for Libraries*. Atlantic Provinces Library Association Conference 2017. Accessed 3 Oct 2019 from https://goo.gl/pZZwJ2.
- Byrne, D., and Cook, G. *The Best American Infographics, 2013*. Mariner Books/Houghton Mifflin Harcourt, 2013.
- Carman, T. (7 Oct, 2019). "Liberals step up attacks with 2 weeks left, but Conservative campaign most negative, data shows". CBC News Online. Retrieved on 8 Oct, 2019 from https://www.cbc.ca/news/politics/liberal-conservative-2019-federal-election-1.5309670
- Chan, C. (2014). Gun deaths in Florida. *ThomsonReuters.com*. Accessed 12 June 2019 from http://graphics.thomsonreuters.com/14/02/US-FLORIDA0214.gif.
- *Choosing the right chart: Selecting among 14 chart types.* (n.d.) 365DataScience.com [website]. Accessed on 3 Oct 2019 from https://365datascience.com/chart-types-and-how-to-select-the-right-one/
- Elections Canada. (2015). *Federal Electoral Districts – Canada 2015*. Natural Resources Canada: Ottawa. Accessed 3 Oct 2019 from https://open.canada.ca/data/en/dataset/737be5ea-27cf-48a3-91d6-e835f11834b0
- "File:Internet map 1024.jpg." *Wikimedia Commons, the free media repository*. 20 Sep 2019, 20:09 UTC. Accessed 1 Oct 2019 from https://commons.wikimedia.org/w/index.php?title=File:Internet_map_1024.jpg&oldid=367359875.
- Kriebel, A. (2012). *Stacked area chart vs. Line chart – The great debate*. Vizwiz. Accessed 3 Oct 2019 from http://www.vizwiz.com/2012/10/stacked-area-chart-vs-line-chart-great.html
- Lai, D. and Hacking, X. (2015). SAP BusinessObjects Dashboards 4.1 Cookbook. Accessed 3 Oct 2019 from https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781784391959/2/ch02lvl1sec28/using-a-radar-chart
- Montgomery, L. M. *Anne of Green Gables* [Ebook]. Project Gutenberg, 1992. Accessed 1 Oct 2019 from http://www.gutenberg.org/files/45/45-0.txt
- Nediger, M. (2019). *What is an Inforgraphic? Examples, Templates & Design Tips*. Venngage. Accessed 1 Oct 2019 from https://venngage.com/blog/what-is-an-infographic/
- *Quartiers de référence en habitation*[computer file]. Montréal, QC: Ville de Montréal, 2019. Available: Portail Données Ouvertes. Accessed on 3 Oct 2019 from http://donnees.ville.montreal.qc.ca/dataset/quartiers.
- Schultz, K. (N.D.). *Data Visualization: Getting Started*. University of Toronto Map and Data Library. Accessed 1 Oct 2019 from https://mdl.library.utoronto.ca/dataviz/getting-started
- Sinclair, S. and Rockwell, G. "Cirrus." *Voyant Tools*. 2019. Web tool. Accessed 1 Oct 2019 from http://voyant-tools.org
- Statistics Canada. (2018). *Making Time for Creative Activities*. Statistics Canada Catalogue no. 11-627-M. Ottawa, Ontario. Accessed 1 Oct 2019 from https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2018010-eng.htm
- SteveEqualsTrue. (n.d.). *A Better Format For Excel Chart Secondary Axis Columns Overlap with 3 Series*. ExcelDashboardTemplates.com (website). Accessed on 3 Oct 2019 from https://www.exceldashboardtemplates.com/better-format-excel-chart-secondary-axis-columns-overlap-3-series/
- Trimble, L. (2017). Finding Canadian Statistics & Data. Presentation for graduated student library assistants, Map & Data Library, University of Toronto.
- *Waffle Chart*. DataVizProject (website). (n.d.). Accessed 3 Oct 2019 from https://datavizproject.com/data-type/percentage-grid/

Feedback

- https://bit.ly/32MyKue