bit.ly/UBL-DSR

UB Libraries Data Skills Retreat

Led by Rachel Starry, CLIR Postdoctoral Fellow

August 6-7, 2019

Data Visualization Essentials

Tuesday, August 6 (9:00am-12:00pm)

Learning Goals

- Build awareness of best practices for creating effective and accessible data visualizations
- Develop familiarity with a variety of chart types and resources for selecting a chart type
- Learn about RAWGraphs' functionality and value as a tool for quickly reformatting and visualizing data

Today’s Activities

- Session Introduction
- Practicing Data Viz Evaluation (30 mins)
- Intro to RAWGraphs and Gapminder Tools (25 mins)
- RAWGraphs Practice (35 mins)
- Exploring *From Data to Viz* (25 mins)

Exploratory vs. Explanatory Visualization

- **Exploratory** visualization: making sense of your data
  - Turning over 100 rocks to find 1 or 2 gems
- **Explanatory** visualization: sharing information with an audience
  - Telling a story about those 1 or 2 gems

Source: Storytelling with Data by Cole Nussbaumer Knaflic (http://www.storytellingwithdata.com/blog/2014/04/exploratory-vs-explanatory-analysis)

Data Viz in 5 Steps

1. Know your data
2. Determine your purpose
3. Choose a chart type
4. Decide on a visualization tool
5. Refine your visualization

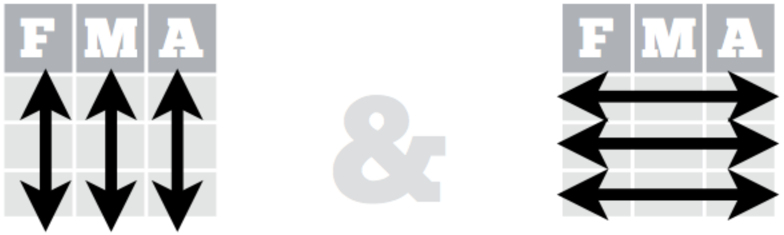
Step 1. Know your data

- Is your data ready to visualize?
  - Is it **clean**?
  - Is it in the proper **format**? (meaning – is it **tidy**?)
  - Do you know **where it came from**?
  - Do you know **what the variables mean**?

Step 1. Know your data

- Tidy Data TL;DR:
  - Each variable forms a column.
  - Each observation forms a row.

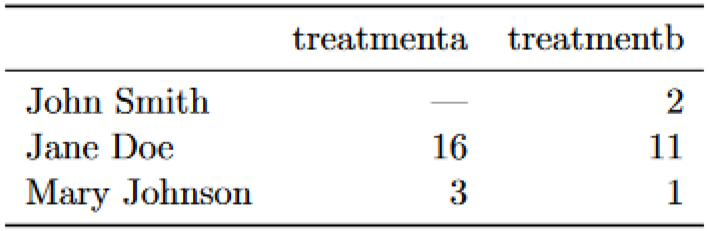
In a tidy data set:



Each **variable** is saved in its own **column**

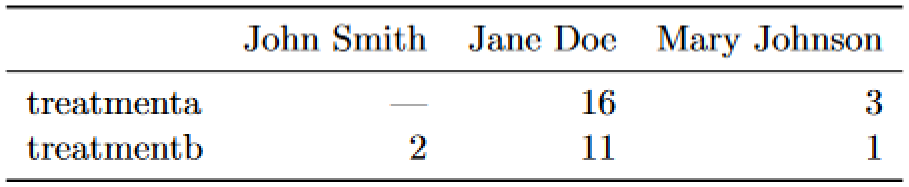
Each **observation** is saved in its own **row**

Step 1. Know your data



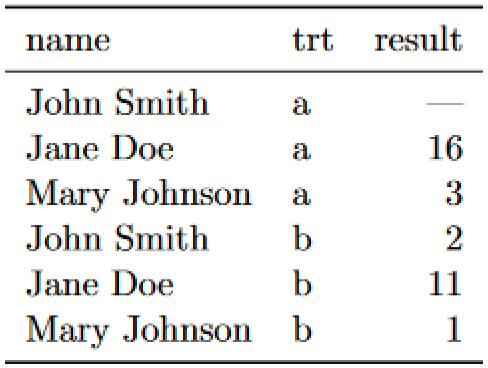
	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.



	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table 2: The same data as in Table 1 but structured differently.



name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Table 3: The same data as in Table 1 but with variables in columns and observations in rows.

Source: Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10).

Step 2. Determine your purpose

- Are you exploring? Or are you explaining?

Step 3. Choose a chart type

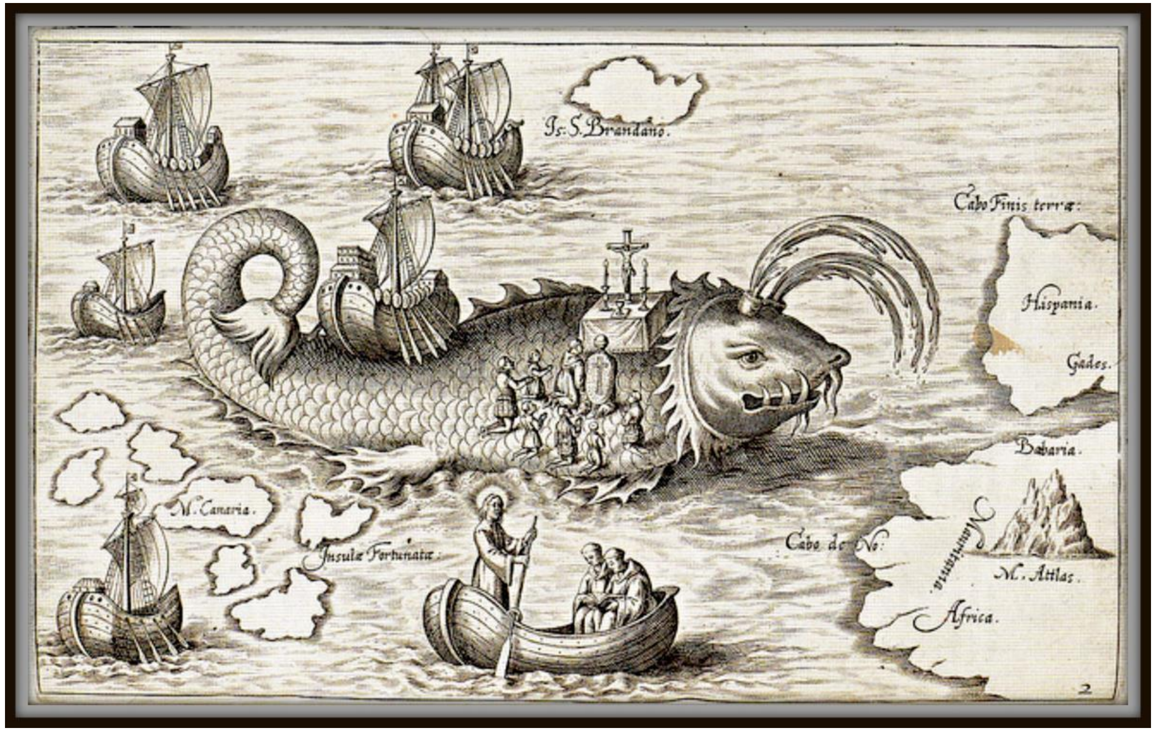
- Different charts are useful for different kinds of data, and for different purposes (exploring vs. explaining).

Step 4. Decide on a viz tool

- Many options available, some tailored to particular purposes or kinds of data.
- Trade-off between a tool’s simplicity and power.

Step 5. Refine your viz

- Here be monsters.
  (This is where data viz can go very wrong. Beware of lying with graphs!)



Data Viz: Designing for Accessibility

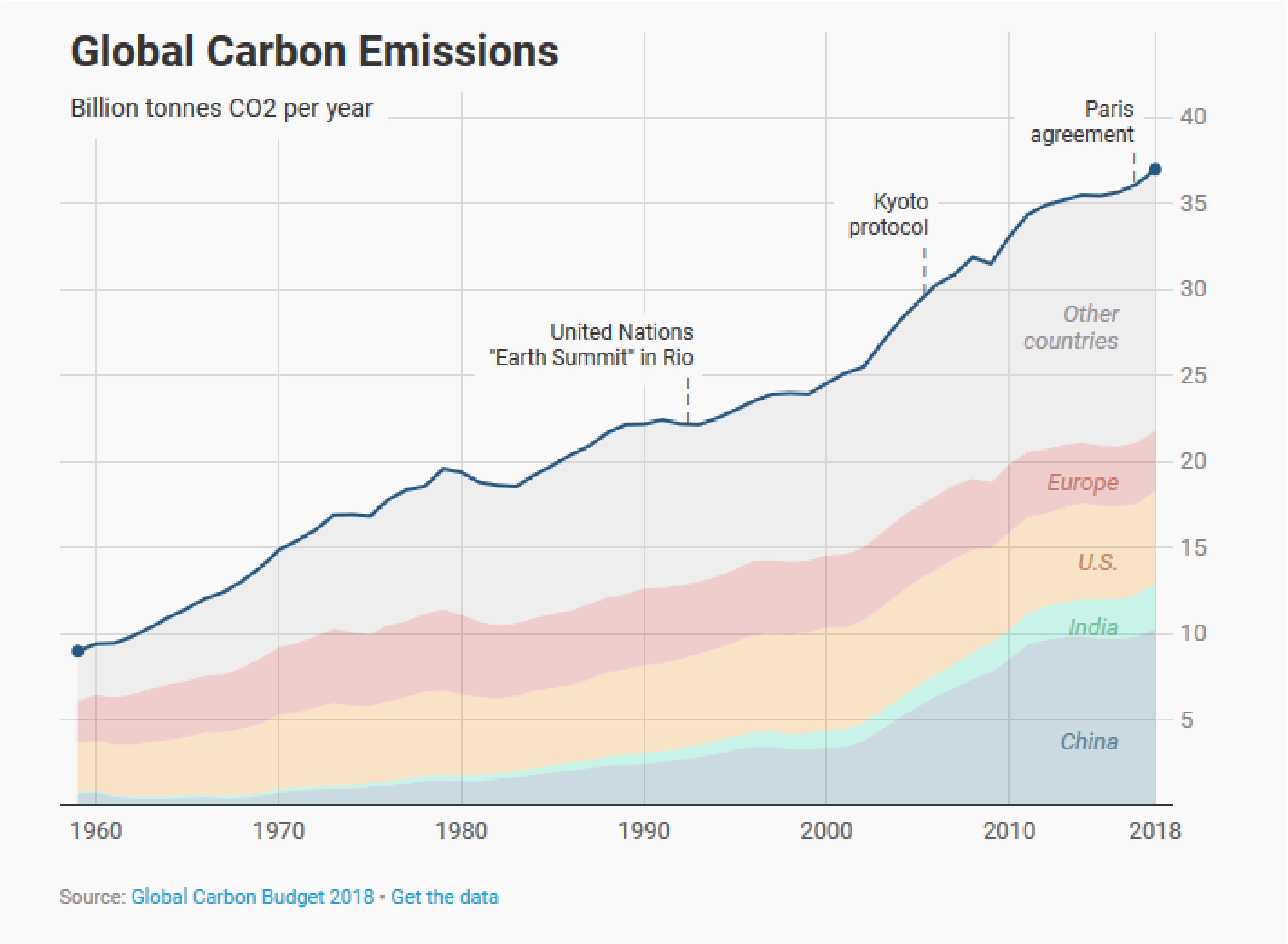
- CFPB Design Manual, on data visualization
- W3C web content accessibility guidelines also useful
- Always ask: "Does this have to be a visualization?"

- Designing for accessibility makes your visualizations more understandable by all users.

Data Viz: Designing for Accessibility

- General best practices:
  - **Provide an alternative description of your chart and the information it conveys**
  - Alt text: always describe charts shared online in image "alt text"
  - Do not use color as the sole means of conveying info
  - Use colorblind-safe color palettes
  - Adhere to w3c standards for contrast ratio (text and visual chart components)
  - Label data directly whenever possible

Data Viz: Designing for Accessibility



Global Carbon Emissions

Billion tonnes CO2 per year

Source: Global Carbon Budget 2018 • Get the data

Source: https://blog.datawrapper.de/weekly-chart-greenhouse-gas-emissions-climate-crisis/

Data Viz: Designing for Accessibility

- “This stacked area chart depicts annual global CO2 emissions from 1959-2018, showing the steady total rise from under 10 billion tons to over 35 billion tons, noting important global forums and accords on climate change (UN “Earth Summit” in 1992, Kyoto protocol in 2005, Paris agreement in 2016). China accounts for more emissions than the US and Europe combined.”

Data Viz: Designing for Accessibility

- General best practices:
  - Provide an alternative description of your chart and the information it conveys
  - **Alt text: always describe charts shared online in image "alt text"**
  - Do not use color as the sole means of conveying info
  - Use colorblind-safe color palettes
  - Adhere to w3c standards for contrast ratio (text and visual chart components)
  - Label data directly whenever possible

Data Viz: Designing for Accessibility

- General best practices:
  - Provide an alternative description of your chart and the information it conveys
  - Alt text: always describe charts shared online in image “alt text”
  - **Do not use color as the sole means of conveying info**
  - Use colorblind-safe color palettes
  - Adhere to w3c standards for contrast ratio (text and visual chart components)
  - Label data directly whenever possible

Data Viz: Designing for Accessibility

- General best practices:
  - Provide an alternative description of your chart and the information it conveys
  - Alt text: always describe charts shared online in image “alt text”
  - Do not use color as the sole means of conveying info
  - **Use colorblind-safe color palettes**
  - Adhere to w3c standards for contrast ratio (text and visual chart components)
  - Label data directly whenever possible

Viz Palette

- https://projects.susielu.com/viz-palette

Data Viz: Designing for Accessibility

- General best practices:
  - Provide an alternative description of your chart and the information it conveys
  - Alt text: always describe charts shared online in image “alt text”
  - Do not use color as the sole means of conveying info
  - Use colorblind-safe color palettes
  - **Adhere to w3c standards for contrast ratio (text and visual chart components)**
  - Label data directly whenever possible

Contrast Ratio

- https://contrast-ratio.com/

Data Viz: Designing for Accessibility

- General best practices:
  - Provide an alternative description of your chart and the information it conveys
  - Alt tags: always describe charts shared online in image “alt text”
  - Do not use color as the sole means of conveying info
  - Use colorblind-safe color palettes
  - Adhere to w3c standards for contrast ratio (text and visual chart components)
  - **Label data directly whenever possible**

Labeling Data

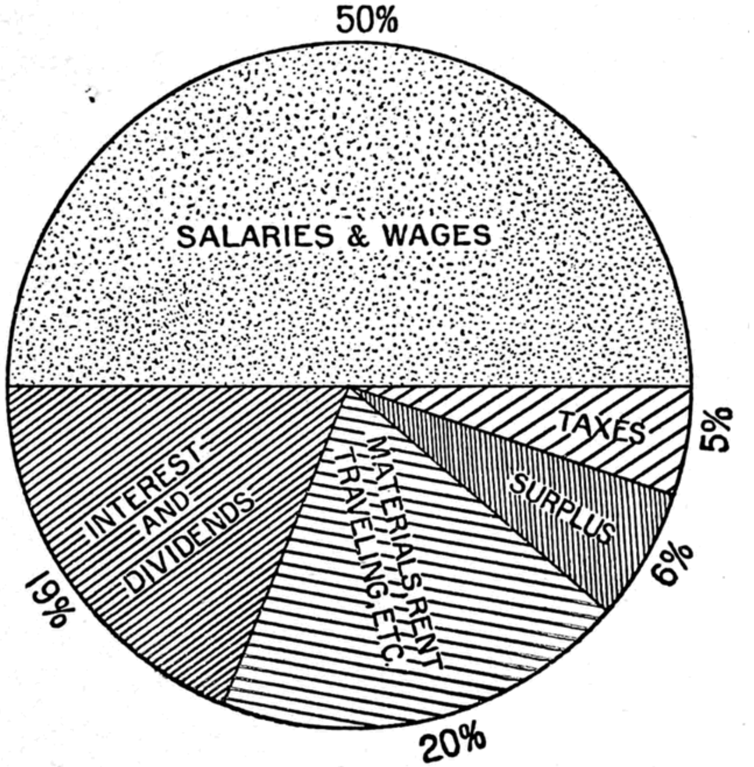
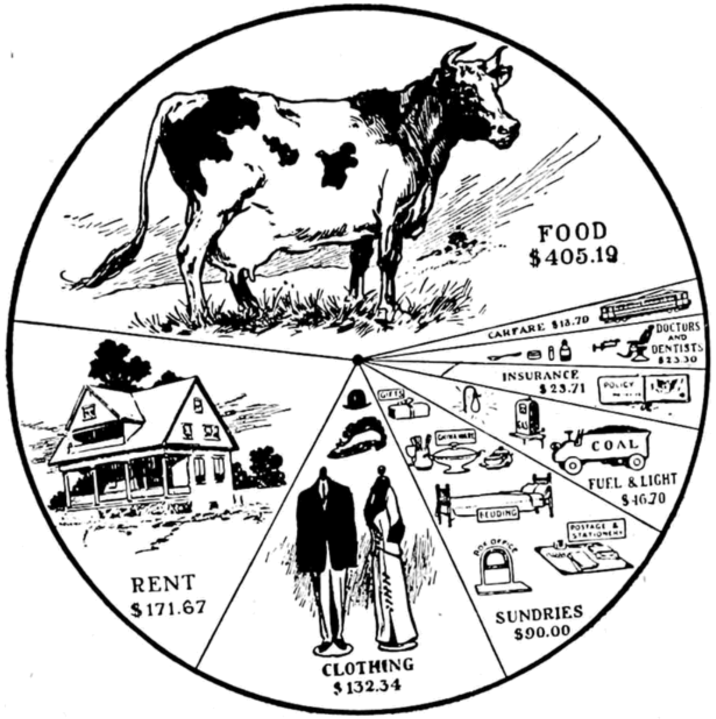


Fig. 2. Disposition of the Gross Revenue of the Bell Telephone System for the Year 1911

This chart was taken from the annual report to the stockholders of the American Telephone and Telegraph Company for the year ending December 31, 1911

The circle with sectors is not as desirable an arrangement as the horizontal bar shown in Fig. 1



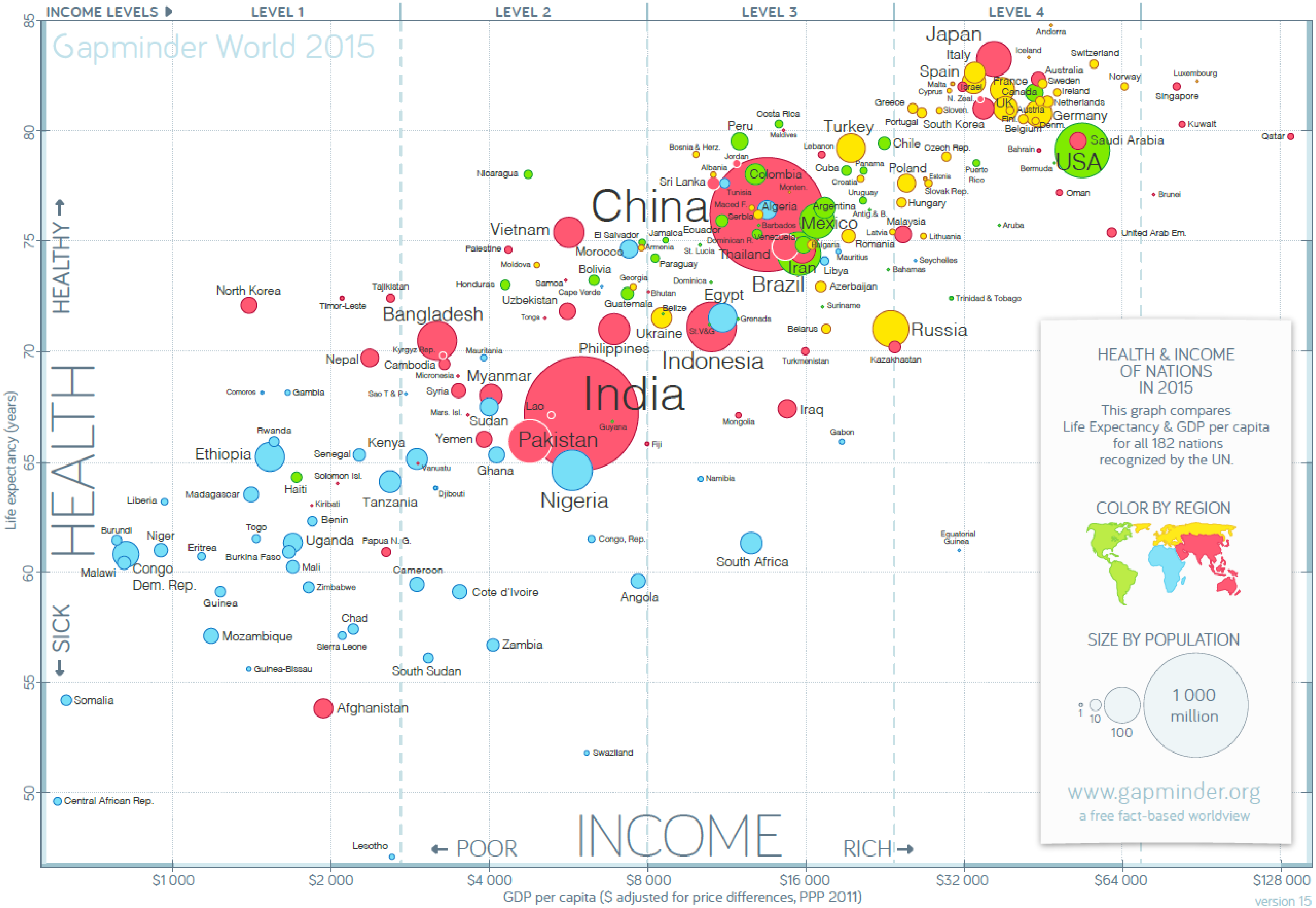
The Survey

Fig. 3. Disposition of a Family Income of from $900 to $1000

This cut shows an attempt to put figures in popular form. The eye is likely to judge by the size of the pictures rather than by the angles of the sectors

Source: https://eagereyes.org/blog/2015/ye-olde-pie-chart-debate

Gapminder World 2015



HEALTH & INCOME OF NATIONS IN 2015

This graph compares Life Expectancy & GDP per capita for all 182 nations recognized by the UN.

COLOR BY REGION

SIZE BY POPULATION

www.gapminder.org

a free fact-based worldview

version 15

DATA SOURCES—INCOME: World Bank's GDP per capita, PPP (2011 international $). Income of Syria & Cuba are Gapminder estimates. X-axis uses log-scale to make a doubling income show same distance on all levels. POPULATION: Data from UN Population Division. LIFE EXPECTANCY: IHME GBD-2015, as of Oct 2016. ANIMATING GRAPH: Go to www.gapminder.org/tools to see how this graph changed historically and compare 500 other indicators. LICENSE: Our charts are freely available under Creative Commons Attribution License. Please copy, share, modify, integrate and even sell them, as long as you mention: "Based on a free chart from www.gapminder.org".

Gapminder

- https://www.gapminder.org/answers/how-does-income-relate-to-life-expectancy/

RAWGraphs

- RAWGraphs is an open-source visualization framework, built on top of the D3 JavaScript library.
- Created and maintained by the DensityDesign Research Lab in Italy.
- RAWGraphs Gallery

Choosing a Chart Type

- Things to consider:
  - What **type of data** are you working with?
  - **How many variables** do you want to visualize at once?
  - Is your data **sequential or categorical**?
  - What **type of pattern** are you trying to explore or show?

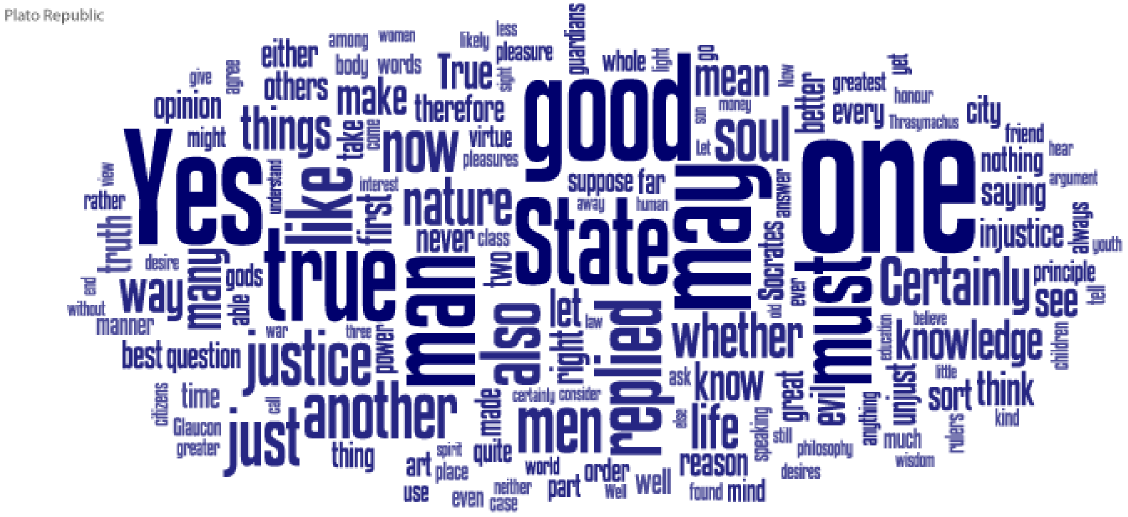
Choosing a Chart Type

- Type of Data
  - Certain types of visualizations are better suited to particular types of data
  - Types: text, numeric, networks, geospatial, time series, ...

Choosing a Chart Type

- Type of Data
  - Certain types of visualizations are better suited to particular types of data
  - Types: **text**, numeric, networks, geospatial, time series, …

Word Cloud

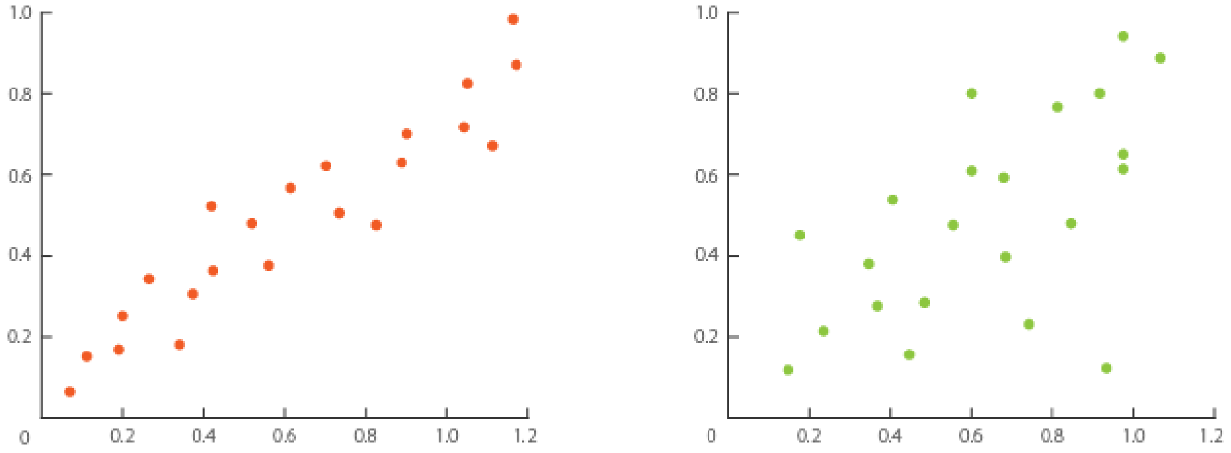


https://datavizcatalogue.com/

Choosing a Chart Type

- Type of Data
  - Certain types of visualizations are better suited to particular types of data
  - Types: text, **numeric**, networks, geospatial, time series, …

Scatterplot

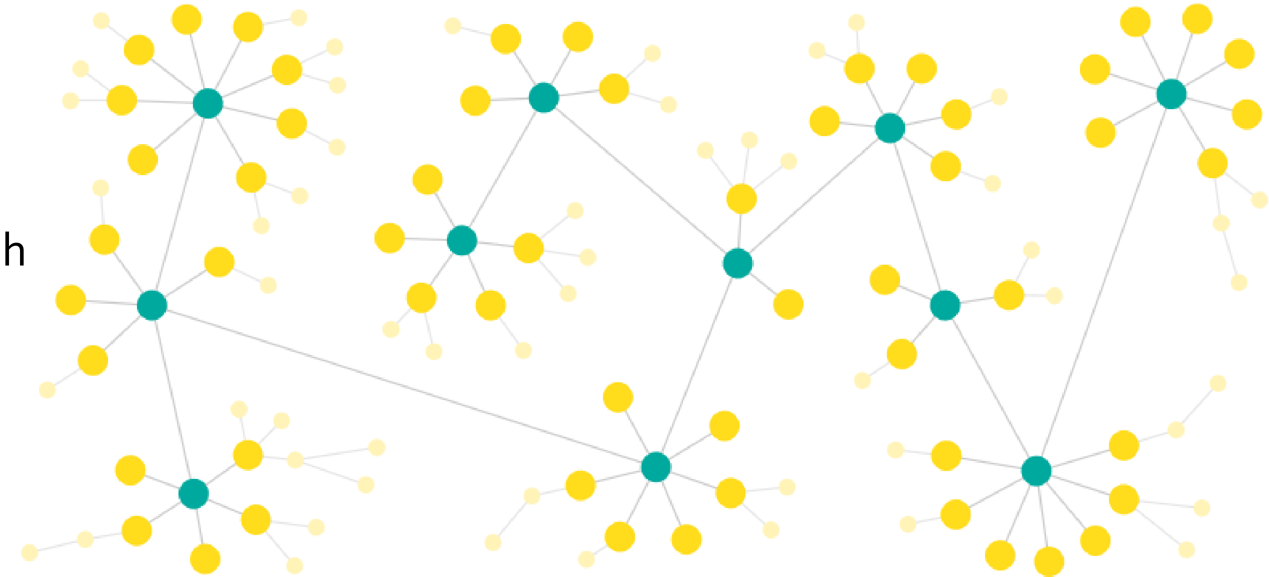


https://datavizcatalogue.com/

Choosing a Chart Type

- Type of Data
  - Certain types of visualizations are better suited to particular types of data
  - Types: text, numeric, **networks**, geospatial, time series, …

Network graph

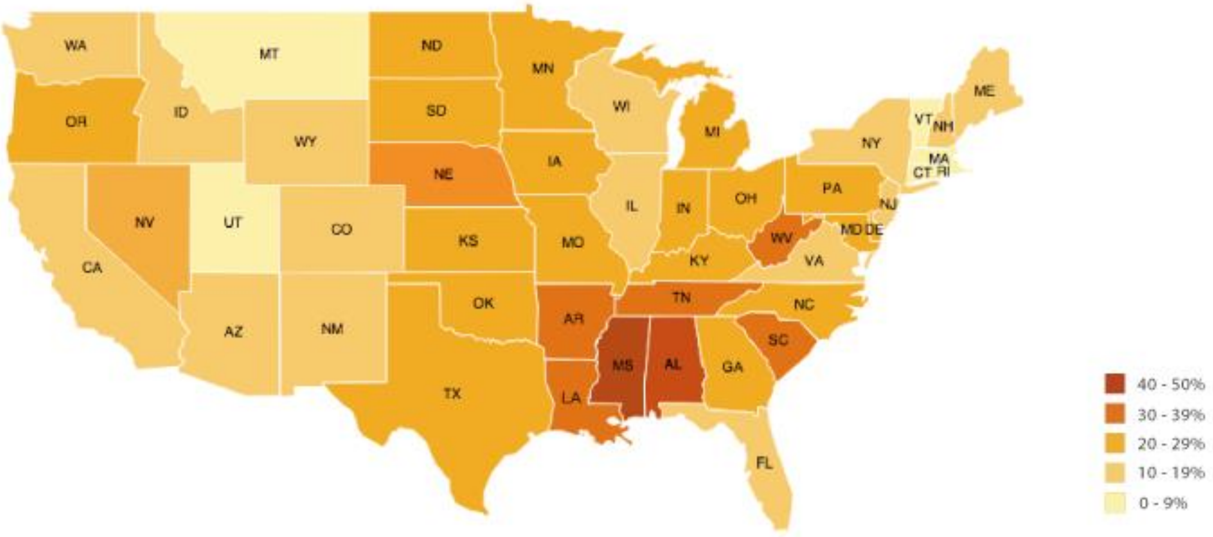


https://datavizcatalogue.com/

Choosing a Chart Type

- Type of Data
  - Certain types of visualizations are better suited to particular types of data
  - Types: text, numeric, networks, **geospatial**, time series, …

Choropleth map

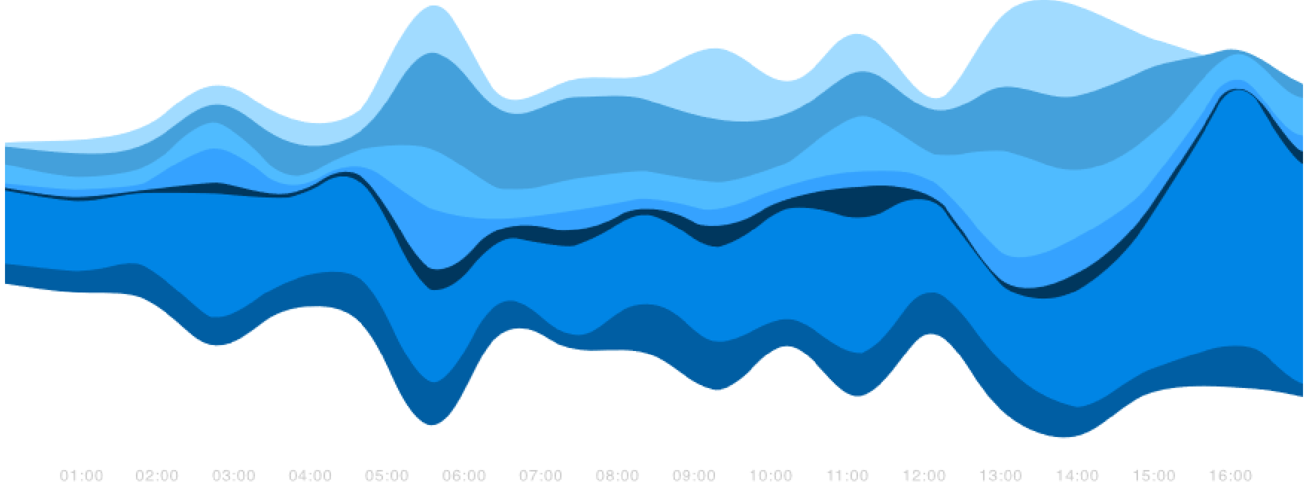


https://datavizcatalogue.com/

Choosing a Chart Type

- Type of Data
  - Certain types of visualizations are better suited to particular types of data
  - Types: text, numeric, networks, geospatial, **time series**, …

Stream graph



https://datavizcatalogue.com/

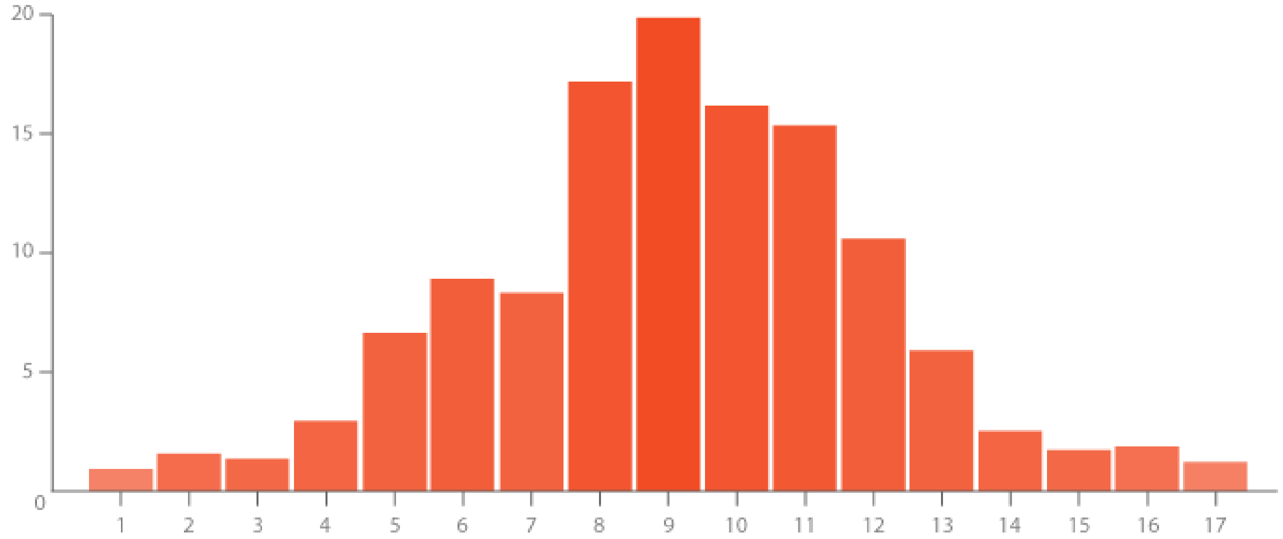
Choosing a Chart Type

- How Many Variables
  - Certain types of visualizations require specific numbers of variables
  - One, two, three, …

Choosing a Chart Type

- How Many Variables
  - Certain types of visualizations require specific numbers of variables
  - **One**, two, three, …

Histogram

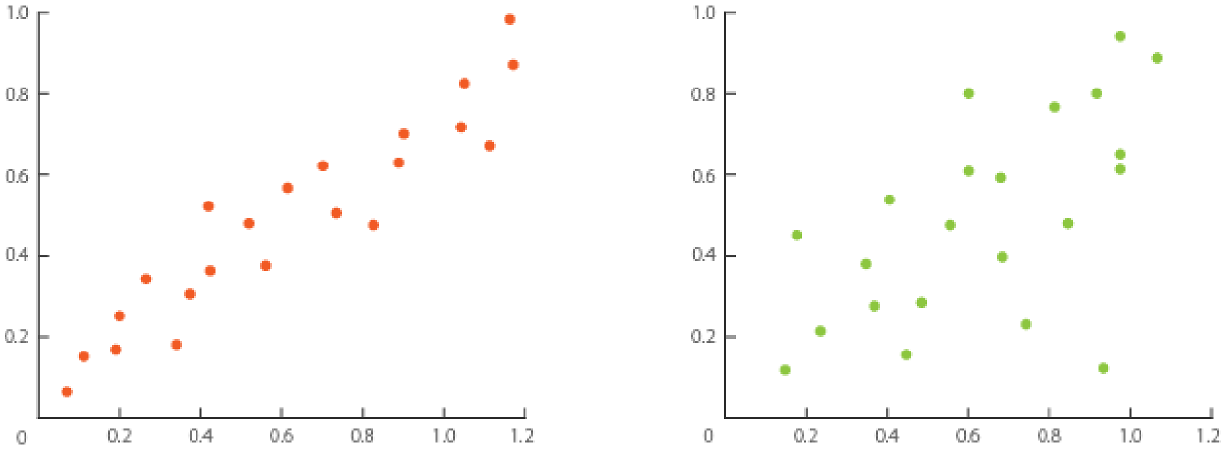


https://datavizcatalogue.com/

Choosing a Chart Type

- How Many Variables
  - Certain types of visualizations require specific numbers of variables
  - One, **two**, three, …

Scatterplot

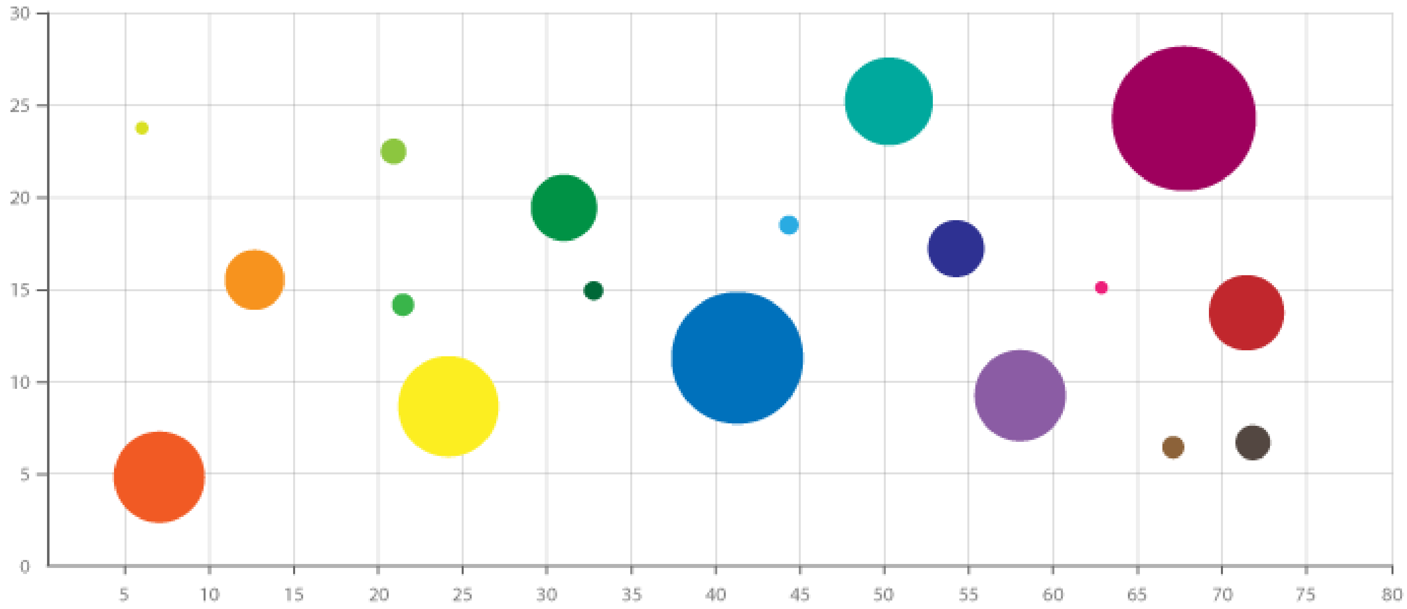


https://datavizcatalogue.com/

Choosing a Chart Type

- How Many Variables
  - Certain types of visualizations require specific numbers of variables
  - One, two, **three**, …

Bubble chart



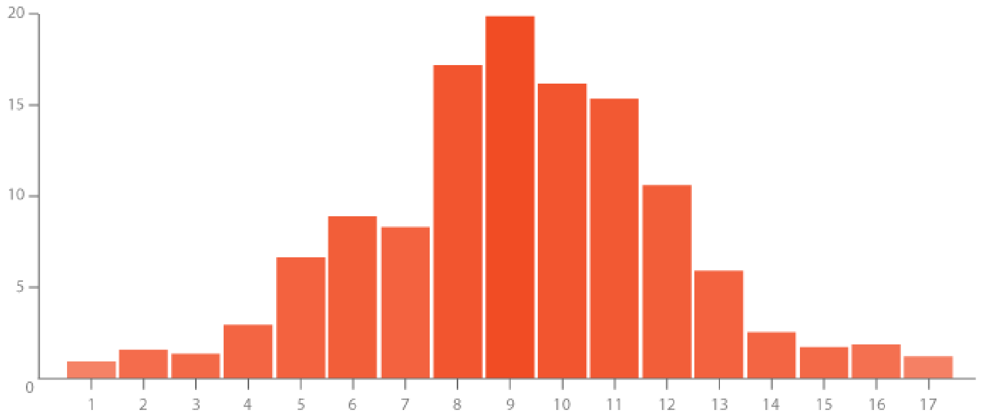
https://datavizcatalogue.com/

Choosing a Chart Type

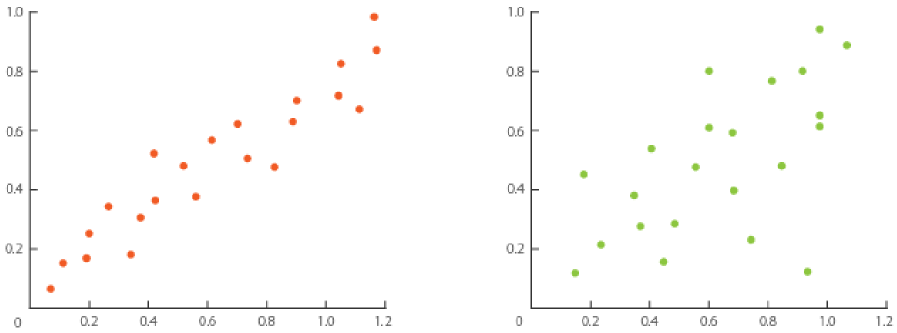
- Sequential vs. Categorical Data
    - Certain types of visualizations are better suited to continuous or discrete data
    - Sequential = continuous or continuously varying
    - Categorical = discrete or qualitative

Choosing a Chart Type

- Sequential vs. Categorical Data
  - Certain types of visualizations are better suited to continuous or discrete data
  - **Sequential** = continuous or continuously varying
  - Categorical = discrete or qualitative



Histogram

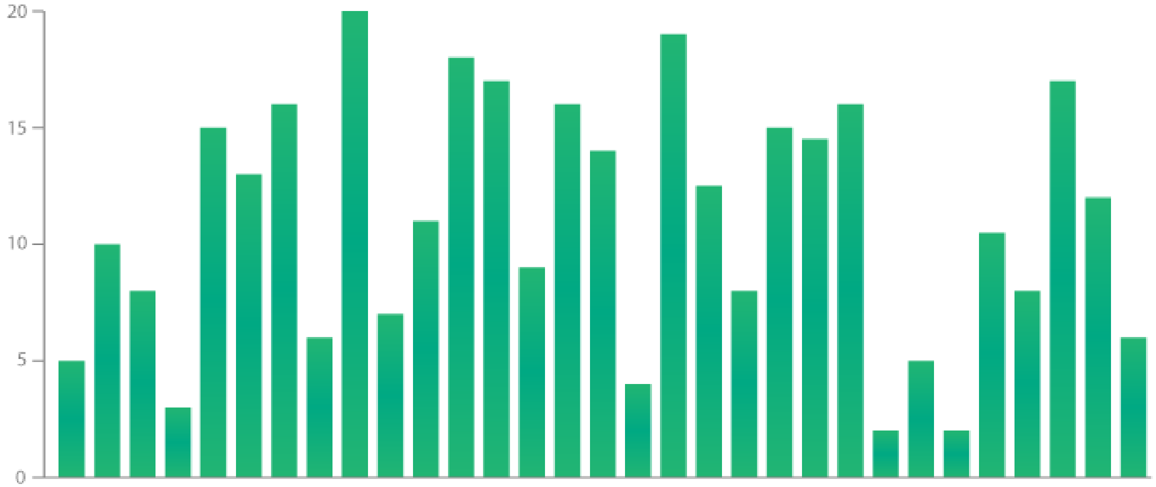


Scatterplot

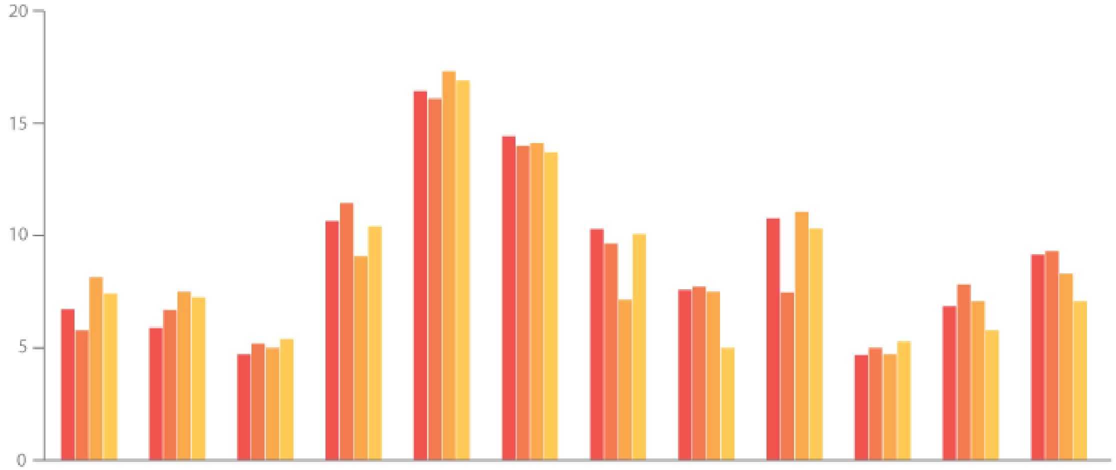
https://datavizcatalogue.com/

Choosing a Chart Type

- Sequential vs. Categorical Data
  - Certain types of visualizations are better suited to continuous or discrete data
  - Sequential = continuous or continuously varying
  - **Categorical** = discrete or qualitative



Bar chart



Grouped bar chart

https://datavizcatalogue.com/

Choosing a Chart Type

- Type of Pattern
  - Certain types of visualizations represent particular types of patterns or relationships within the data
  - Comparison, proportion/composition, part-to-whole, hierarchy, changes over time, …

- Let’s explore some examples: https://datavizcatalogue.com/

From Data to Viz

https://www.data-to-viz.com/

Questions to address:

- What **type of data** can this chart represent? (numeric, text/categorical, network, geospatial, time series)
- **How many variables** can this chart represent simultaneously?
- Does this chart require **continuous/sequential** data, or **discrete/categorical** data – or a combination?
- What **kind of pattern(s)** does this chart depict? (comparison, proportion, composition, distribution, part-to-whole, hierarchy, change over time)
- Are there any **common mistakes** to avoid when using this chart?

Reflection

Prompt:

What is one concept that you feel you now understand better? One topic that was completely new to you? One question you would like to explore further?

Data Wrangling with OpenRefine

Tuesday, August 6 (1:00-4:00pm)

Learning Goals

- Build awareness of the concept of “tidy data” and why it’s important
- Develop familiarity with best practices for cleaning data
- Learn about OpenRefine’s functionality and value as a tool for cleaning messy data, joining multiple datasets, and pivoting data

Today's Activities

- Session Introduction
- Tidy Data Introduction & Exercise (20 mins)
- OpenRefine Basics (40 mins)
- OpenRefine Practice (30 mins)
- Using OpenRefine to Join and Pivot Data (30 mins)

Exercise

What's wrong with this spreadsheet?

(Try to list every issue you see with this dataset.)

Common Spreadsheet Errors

- Using multiple tables within one spreadsheet
- Putting related data in multiple tabs
- Not filling in zeros or missing data
- Using problematic null/missing values (-999)
- Using formatting (highlighting, text color, etc.) to store information
- Using formatting to make the spreadsheet prettier (like merging cells)
- Placing comments or units in cells
- Putting more than one piece of information in a single cell
- Using problematic variable names
- Using special characters in data (line breaks, em-dashes, fancy quote marks, etc.)

Defining “Clean Data”

- Creating machine-readable spreadsheets:
  - Avoid using multiple tables within one spreadsheet.
  - Avoid spreading data across multiple tabs.
  - Don’t use formatting to convey information or make your spreadsheet look pretty.
  - Don’t include comments inside of cells.
  - Record units in column headers, not inside of cells.
  - Avoid whitespace and special characters in your data.

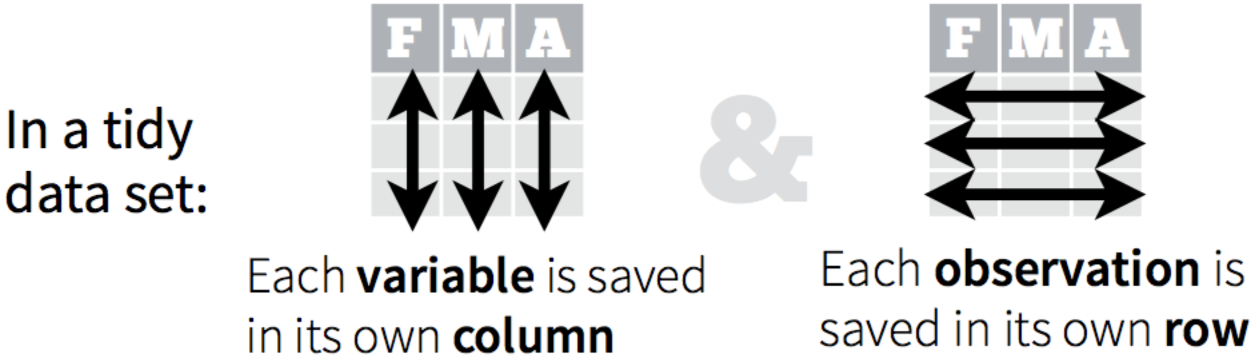
Defining “Clean Data”

- “Clean” Data…
  - treats missing values consistently (`0` or `NA`)
  - stores only one type of data in each column
  - stores only one value in each cell
  - uses appropriate variable names

See the workshop materials for “Data Organization in Spreadsheets”

Tidy Data

- Tidy Data:
  - Each variable forms a column.
  - Each observation forms a row.



In a tidy data set:

Each **variable** is saved in its own **column**

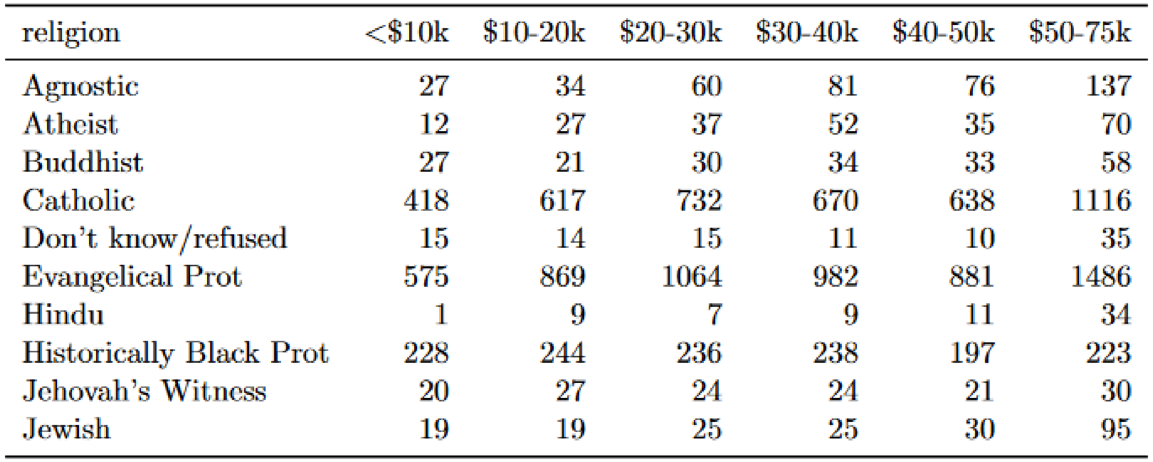
&

Each **observation** is saved in its own **row**

Source: Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10).

Tidy Data

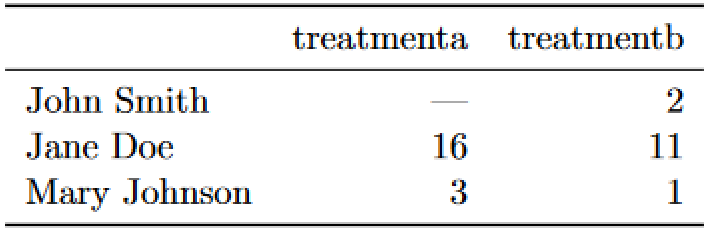
- Some definitions:
  - **Dataset**: a collection of values
  - **Value**: numbers or text
  - **Variable**: a collection of all values measuring the same attribute across units
  - **Observation**: a collection of all values measured on the same unit, across all attributes



religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

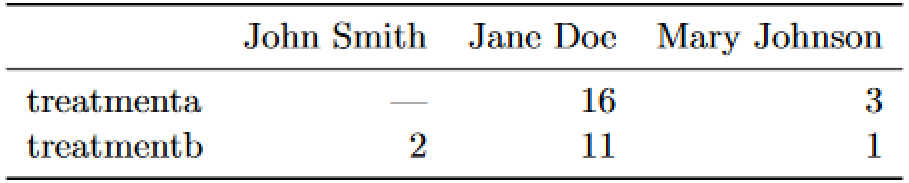
Source: Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10).

Tidy Data



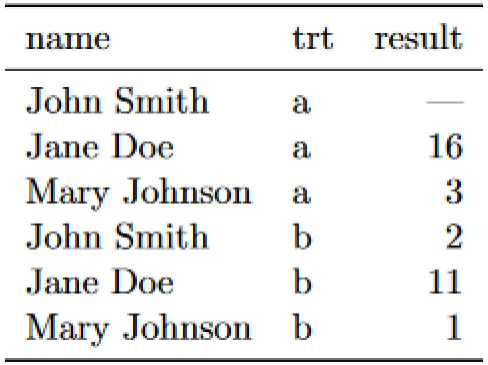
	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.



	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table 2: The same data as in Table 1 but structured differently.

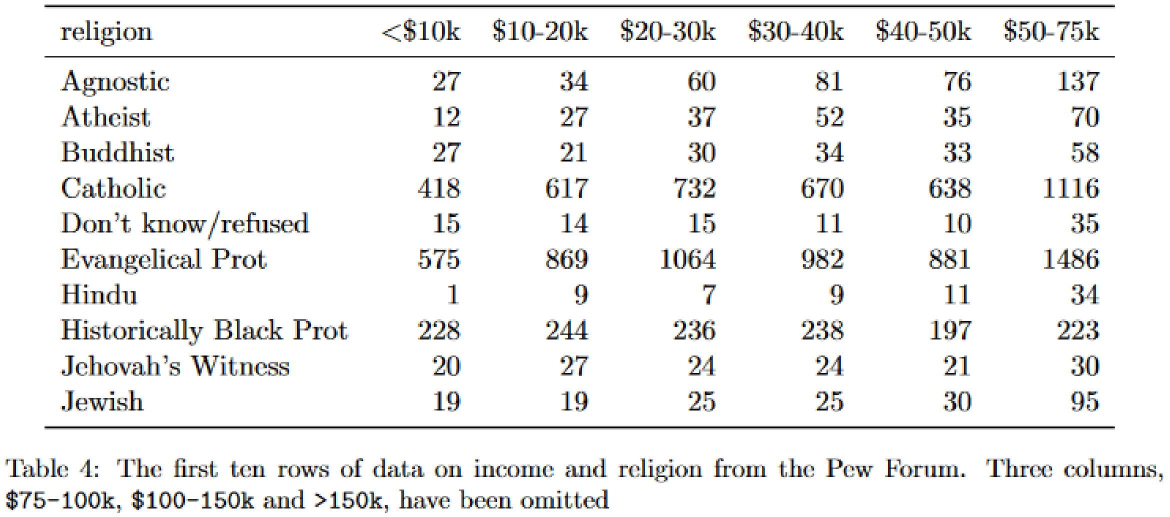


name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Table 3: The same data as in Table 1 but with variables in columns and observations in rows.

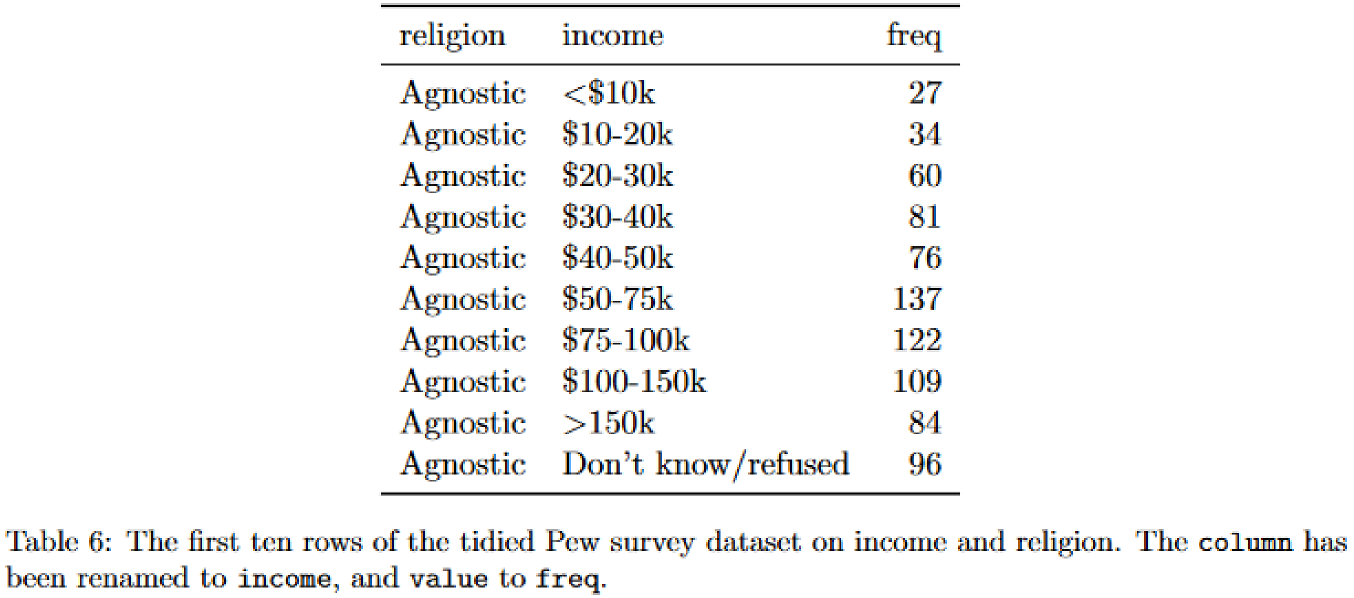
Source: Wickham, Hadley. 2014. “Tidy Data.” *Journal of Statistical Software* 59 (10).

Tidy Data



religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, `$75-100k`, `$100-150k` and `>150k`, have been omitted



religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. The `column` has been renamed to `income`, and `value` to `freq`.

Source: Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10).

Tidy Data Exercise

Questions to address:

- What are the main variables in this dataset?
- Are any values being used as variables (column headers) – if so, how would you correct this issue?
- How would you combine the data for all four years into a single table?
- How else would you clean up the spreadsheet?

OpenRefine

- http://openrefine.org/

- Formerly Freebase Gridworks → Google Refine → OpenRefine
- Open-source tool for working with messy data

- Documentation: https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- GREL: https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions

OpenRefine Practice

Things to try:

- Faceting to find typos + clustering to fix
- Faceting to find data entry errors
- Adding “Distant Learner” column based on “Status”
- Adding additional “Loan Title” columns, split on =
- Removing numbers from “Loan Author” column
- Basic clean-up
    - Remove leading/trailing whitespace
    - Remove rogue punctuation or special characters
- Creating a “Full Address” column by combining 3 separate address columns

Data Carpentry Links

Data Organization in Spreadsheets for Social Scientists: https://datacarpentry.org/spreadsheets-socialsci/

OpenRefine for Social Science Data: https://datacarpentry.org/openrefine-socialsci/

All Data Carpentry lessons: https://datacarpentry.org/lessons/

Reflection

Prompt:

What is one concept that you feel you now understand better? One topic that was completely new to you? One question you would like to explore further?

Text Mining & Analysis with Voyant Tools

Wednesday, August 7 (9:00am-12:00pm)

Learning Goals

- Build awareness of significant methods and techniques used in computational text analysis
- Learn about Voyant Tools' functionality and value as a tool for research or teaching
- Develop confidence in exploring patterns in text on your own

Today's Activities

- Session Introduction
- Dive Into Voyant (40 mins)
- Exploring Voyant-powered Analyses (20 mins)
- Exploring Methods of Text Analysis (40 mins)
- Re-Visiting Voyant as a Scholarly Tool (20 mins)

What Is Text Mining?

- Essential definition: “deriving **high-quality**, **structured data** from texts.”

What Is Text Mining?

- Common text mining tasks:
  - **Tokenization** into words and **n-grams**
  - Removing **stopwords** and punctuation
  - Computing **term and document frequencies (TF-IDF)**
    - *term frequency*: how frequently a word occurs in a document
    - *inverse document frequency*: how common a term is within a collection of documents

“Bag-of-Words” Methods

“This unflattering description captures how researchers often program computers to read books as a collection of words separated by spaces, stripped of any logical order or meaning.” (source)

Text "Mining" vs. Text "Analysis"

- Many terms are not clearly defined or used interchangeably.
- Natural Language Processing (NLP) techniques go beyond the "bag-of-words" approaches.
  - Named entity recognition
  - Part-of-speech tagging
  - Word sense disambiguation
  - Stemming and lemmatization

Research Questions for Text Analysis

- Testing a hunch, hypothesis, or thesis about an author, text, passage, genre, or period.
- Studying how an author's style changes over time or how genres develop and decay (as well as characterizing or refining how genres are even defined).
- Investigating the history of an important word, concept, or group of words or concepts over a long time span.
- Exploring the characteristic vocabulary of an author, genre, period, group of texts, a single text, or a part of a text, or comparing the characteristic vocabulary of multiple authors, genres, periods, or collections of text.
- Investigating questions of authorship attribution.
- Exploring thematic language in texts.

Source: David Hoover, 2013, "Textual Analysis,"
in *Literary Studies in the Digital Age: An Evolving Anthology,* ed. Price and Siemens

Voyant Tools

What is Voyant?

https://voyant-tools.org/

- Open-source, web-based reading and analysis environment
- Developed by Stéfan Sinclair (McGill University) and Geoffrey Rockwell (University of Alberta)
- Simple interface with over 20 tools for investigating texts
- Documentation: http://docs.voyant-tools.org/start/

Voyant Tools

- **Cirrus**: a kind of word cloud showing the most frequent terms
- **Reader**: a view into the corpus that fetches segments of text as you scroll
- **Trends**: a distribution graph showing terms across the corpus (or terms within a document)
- **Summary**: a tool that provides a simple, textual overview of the current corpus
- **Contexts**: a concordance that shows each occurrence of a keyword with a bit of surrounding context
- Full list: https://voyant-tools.org/docs/#!/guide/tools

Cirrus, Terms, & Links

- What happens when you hover your mouse over different parts of the Cirrus widget?
- What happens when you click on a word?
- What hidden buttons can you find?
- Find the "Options" button and edit or view the list of "Stopwords." Does this change your results?

- What do we learn about the text from this Word Cloud?

Trends & Document Terms

- How are words in the “trends” widget chosen for inclusion?
- What kind of information to the trend lines, x- and y-axes convey?
- What does the graph tell us about our texts?
- What information does the “document terms” widget give us?
- Bonus Questions:
  - What is the difference between relative and raw frequencies and how do you find that information?
  - Is this a helpful visualization? What might impact the value of this visualization?

Summary, Documents, & Phrases

- What kinds of document summaries are included in the summary widget?
- What information does the “document” widget give us?
- What information does the “phrases” widget give us? How is it different from the “document terms” widget we looked at earlier?

- What questions do these summaries prompt for you about our collection of texts?

Choose a Tool and Explore

Questions to address:

- What is it called?
- How do you interact with it?
- What do you learn from it?
- What is confusing about it?
- How can you save or export the data?
- Where can you find help understanding how to use it?

Voyant-powered Analyses

- http://hermeneuti.ca/

Exploring Text Analysis Methodologies

1. The Beatles
2. The Lord of the Rings
3. Billboard Top 100
4. Early American Cookbooks
5. New York Times
6. JK Rowling

Exploring Text Analysis Methodologies

Questions to address:

- Who performed this study?
- What method(s) did they use?
- What were their conclusions? (If these are not explicit, what do you think we can learn from this kind of analysis?)
- What challenges or uncertainties may be involved in this method?

Questions to Consider

- Questions about technology and methodology:
  - How complete is the text or corpus being analyzed?
  - What is the quality of the text?
  - What might Voyant help us learn about the language or style of an author?
  - What might Voyant help us understand about genre or the format of text?
  - Is large-scale text analysis the most effective way to draw meaning from words?
  - If word frequencies aren't sufficient, how else do we find patterns in texts?

Questions to Consider

- Questions about research or instruction:
  - How might you see yourself using this kind of tool to dig into texts you work with?
  - How might you see this kind of tool or text analysis more generally being useful to the students, researchers, or colleagues you work with regularly?
  - Are there particular approaches or techniques in the world of text analysis that you see as particularly interesting or useful, that you would like to explore further?

Reflection

Prompt:

What is one concept that you feel you now understand better? One topic that was completely new to you? One question you would like to explore further?

Text Processing with Regular Expressions (in R)

Wednesday, August 7 (1:00-4:00pm)

Learning Goals

- Gain familiarity with reading and writing regular expressions
- Build awareness of essential text processing concepts and practices
- Develop confidence in using a programming environment to automate text clean-up

Today's Activities

- Session Introduction
- Interactive Intro to RegEx (20 mins)
- Playing with Strings and RegEx in R (50 mins)
- Automating Text Clean-up and Processing in R (45 mins)

What Are Regular Expressions?

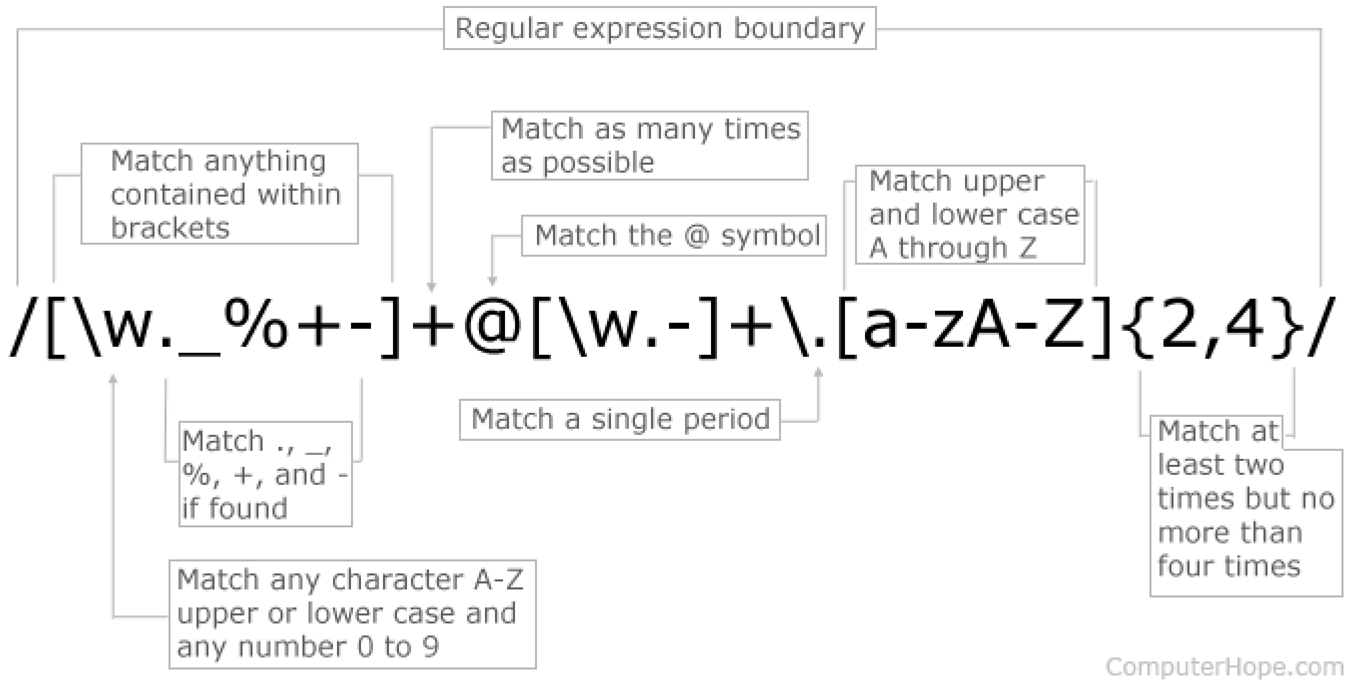
- RegEx are a way of specifying rules that describe a class of strings.
  - Find every word that starts with the letter "a" in a document
  - Extract 4-digit years from a body of text
  - Extract email addresses from a body of text

- RegEx are a "powerful mechanism for pulling structured data out of mountains of text, just by making a recipe for what patterns the text follows."

  - Friedrich Lindenberg, 2016. "A Poor Journalist's Text-Mining Toolkit"

What’s the catch?

Regular Expression E-mail Matching Example



Regular expression boundary

Match anything contained within brackets

Match as many times as possible

Match the @ symbol

Match upper and lower case A through Z

```
/[\w._%+-]+@[\w.-]+\.[a-zA-Z]{2,4}/
```

Match ., _, %, +, and - if found

Match a single period

Match at least two times but no more than four times

Match any character A-Z upper or lower case and any number 0 to 9

ComputerHope.com

“They look horrible. Regular expressions usually end up reading a bit like a three-legged cat tried swing dancing on your keyboard. If you can get over that, the world of text files is your oyster.”

- Friedrich Lindenberg, 2016.

Some Definitions

- **String**: a sequence of characters
- **Vector**: a one-dimensional array
- **Array**: a data structure that contains a group of items, which can each be accessed using the item's **index**

- Strings in R = **character vectors**

RegEx “Flavors”

- RegEx come in different “flavors” depending on the regex engine a given piece of software uses.
- Two main flavors:
  - POSIX
  - Perl-compatible (PCRE)

RegEx: The Basics

- Literal characters
- Special characters (“metacharacters”)
- Character classes
- Shorthand character classes
- Repetition
- Anchors
- Boolean matching
- Grouping

Literal Characters

- The simplest way to search is to match exact characters.
- Example: “d” matches “dog” and “red” but not “Dog”

Special Characters

- Some characters have special meanings in regex; to match literal instances of special characters, “escape” them with a backslash.
- Special characters are **\ ^ $. | ? * + ( ) [{**
  - Example: “\?” matches “hello?” and “yay?!”
- **.** matches any character except newline
  - Example: “**.**ing” matches “wedding” but not “jog” or “ingest”
- **?** says “the previous token is optional, match zero or more times”
  - Example: “colou?r” matches “color” and “colour”

Character Classes

- Character classes let you match one character out of a set or range of possibilities, using square brackets [ ].
  - Example: “[aeiou]” matches “dog” or “cat” but not “grr”
- You can use a hyphen inside [ ] to define a range, e.g. [a-z] or [0-9]
  - Example: “[A-Z]” matches “Robert” but not “general”
- You can use [^…] to specify a set of things NOT to match
  - Example: “[^0-9]” matches “Orange” but not “Red40”
- Some predefined character classes exist in POSIX-style regex:
  - [[:alnum:]]
  - [[:alpha:]]
  - [[:digit:]]
  - [[:lower:]] and [[:upper:]]
  - [[:punct:]]
  - [[:cntrl:]]
  - [[:space:]]

Shorthand Character Classes

- Because some character classes are used often, a series of shorthand classes are available.

- \w = word characters (alphanumeric plus space)
- \W = non-word characters
- \d = digits
- \D = non-digits
- \s = whitespace characters
- \S = non-whitespace characters

Repetition

- Sometimes you want to find more than one match at a time, like “find next” or “find all” in a word processor’s search.
- ***** says “match the previous token zero or more times” (lazy)
- **+** says “match the previous token one or more times” (greedy)

- Use curly braces { } to specify a specific amount of consecutive repetition.
  - Example: “n{2}” will match “anna” but not “nana” or “and”
  - Example: “[0-9]{4}” will match “1989” but not “Y2K” or “123”

Anchors

- Sometimes it's helpful to specify not only WHAT character you're looking for but also WHERE you're expecting it within a string.
- Anchors help match position within a string.

- **^** matches at the start of a string
  - Example: "^b" matches only the first "b" in "bob"
- **$** matches at the end of a string
  - Example: "[0-9]{2}$" matches "Dec91" or but not "99balloons"

Boolean Matching

- You can specify multiple alternatives if you want to match one of a number of possibilities, using the vertical pipe **|** which means “or.”
  - Example: “cat|dog” matches “cat” and “dog” but not “fish”
  - Example: “cat|dog|fish|mouse” matches “The cat in the hat” and “raining cats and dogs”

Grouping

- Parentheses help group tokens together.
    - Example: “Set(Value)?” matches “Set” and “SetValue” because the ? makes the (Value) part of the regex optional.

Interactive Intro to RegEx

- https://regexone.com/

Additional RegEx Resources

- Wikipedia entry: https://en.wikipedia.org/wiki/Regular_expression
- Regular-Expressions.info "Quick Start" guide: https://www.regular-expressions.info/quickstart.html
- RegExr – interactive tool for building/testing regex: https://regexr.com/

Reflection

Prompt:

What is one concept that you feel you now understand better? One topic that was completely new to you? One question you would like to explore further?