# Collect & Prep Your Data for Visualization and Analysis

https://github.com/BCDigSchol/coffee-code/tree/master/data-prep

Anna Kijas @anna_kijas
Sarah Melton @WorldCatLady

# What is data?

- Data is all around you
  - When you assign a value to something - you have data
  - Qualitative vs. quantitative
  - Humanities vs. Social Science Data

# Is this data?

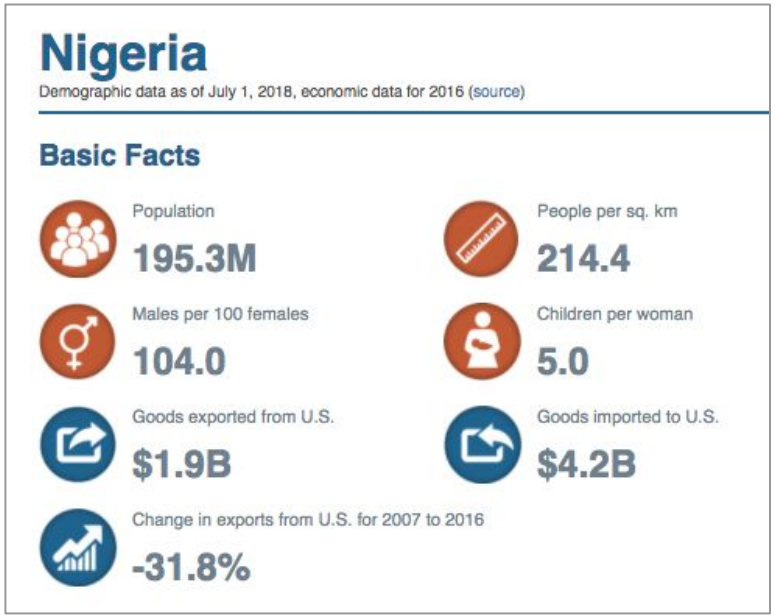
Jane Austen was born on December 16, 1775 in Steventon, Hampshire. She is British. Austen wrote the novel *Sense and Sensibility*.

Emily Dickinson was born on December 10, 1830 in Amherst, Massachusetts. She was an American poet. One of her poems is *A great Hope fell*.

# Is this data? Yes!

| Name | Birth Date (Temporal) | Place (Spatial) | Type | Nationality | Works | Form |
|---|---|---|---|---|---|---|
| Austen, Jane | 1775-12-16 | 51.228457, -1.2201168999999936 | Novelist | British | Sense and Sensibility | Novel |
| Dickinson, Emily | 1830-12-10 | 42.3732216, -72.5198537 | Poet | American | A great Hope fell | Poem |

# Is this data?

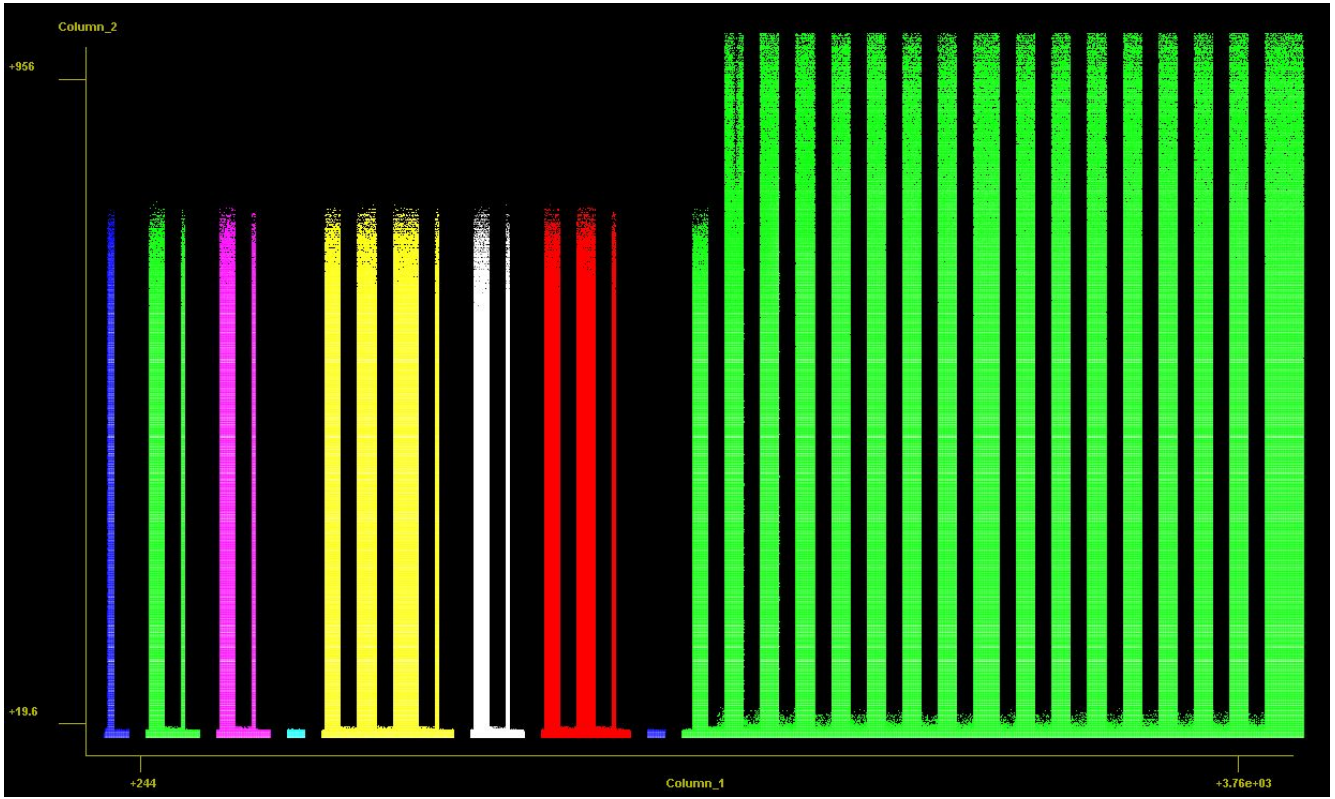




Source: https://www.census.gov/popclock/world

# Is this data? Yes!

| Country | Population | Males per 100 females | Goods exported from U.S. | People per sq. km | Children per woman | Goods imported to U.S. |
|---|---|---|---|---|---|---|
| Brazil | 208,800,000 | 97.1 | $30,100,000,000 | 25.0 | 1.7 | $26,100,000,000 |
| Nigeria | 195,300,000 | 104.0 | $1,900,000,000 | 214.4 | 5.0 | $4,200,000,000 |

# What makes data “useable”?



# What makes data “useable”?

# What makes data “useable”?

# Your Data & Visualizations

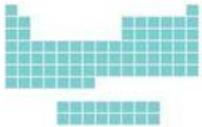
What kind of data are you working with? (Geospatial? Temporal? Textual?)

What's your audience?

What are you trying to convey? (Relationships? Scale?)

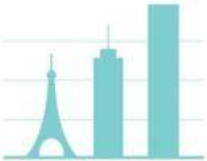
# MOST POPULAR INFOGRAPHICS

YOU CAN FIND AROUND THE WEB



PERIODIC TABLE OF SOMETHING

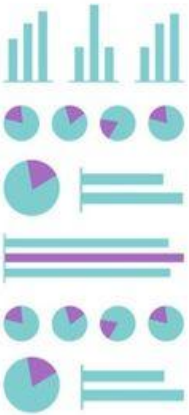
CREDIT CRISIS VISUALIZED WITH NUMBERS THAT ARE NOT EXPLAINING ANYTHING



WORLD'S TALLEST BUILDINGS PLUS SOMETHING THEY ARE BUILDING IN DUBAI

WATER, GAS, OIL OR WHATEVER UNITED STATES CONSUMES MORE THAN ANY OTHER COUNTRY

TUBE MAP OF SOMETHING



A CRAPLOAD OF IRRELEVANT DATA PUT TOGETHER IN A BIG VERTICAL IMAGE

VENN DIAGRAM OF SOMETHING


TAG CLOUD WITH RANDOM WORDS IN THE SHAPE OF SOMETHING

# Exercise 1: Analysis

https://github.com/BCDigSchol/coffee-code/tree/master/data-prep

# Tools & Methods

- OpenRefine (http://openrefine.org/)
- Google Fusion Tables (https://sites.google.com/site/fusiontablestalks/)
- Voyant (https://voyant-tools.org/)
- RAWGraphs (https://rawgraphs.io/)
- Tableau Public (https://public.tableau.com/en-us/s/)

# Things to think about during data prep...

Check to see if your data have leading or trailing whitespaces or extraneous punctuation marks

Do the values in each column match the data type?

How are your dates formatted?

Do you need to split data into separate columns?

Do you need to normalize spelling or letter case?

Do you have coordinates (lat/long)?

# Exercise 2: Data Prep & Cleaning

https://github.com/BCDigSchol/coffee-code/tree/master/data-prep

# Prep & Cleaning

- Brief overview of OpenRefine
- Review sample data (what can we normalize and why?)
- Tips & Recipes
  - Transformations
  - Faceting and clustering
  - Normalizing dates
  - Geocoding

# Normalizing Dates

Build your expression in the Transform function

**value.toDate(‘insert letter).toString(‘insert letter’)**

For day use: d or dd

For month: M or MM

For year use: yy or yyyy

# Geocode Locations

Add a column by fetching URLs based on column

- Name your column
- Change throttle delay to 1000 milliseconds

Use expression:

**“http://nominatim.openstreetmap.org/search?format=json&email=[YOUR_EMAIL_HERE]&app=google-refine&q=" + escape(value, 'url')**

# Geocode Locations

- Split your coordinates into two columns (latitude/longitude)
  - Use expression: **value.parseJson()[0].lat**

- Repeat for longitude
  - Use expression: **value.parseJson()[0].lon**

# Resources

List of Tools, Readings, and Additional Resources:
https://github.com/BCDigSchol/coffee-code/tree/master/data-prep

Data Management & Data Planning: https://libguides.bc.edu/dataplan

DMP Tool: http://www.bc.edu/sites/libraries/dmptool/

Dataverse: https://libguides.bc.edu/dataverse

Open Science Framework (OSF): https://osf.io/

Humanities Commons: https://hcommons.org/

# Thanks!

Anna Kijas - anna.kijas@bc.edu
Sarah Melton - sarah.melton@bc.edu

Find events and more at ds.bc.edu