Data Visualization for Libraries
Principles, Basics, Data Studio
Melissa Belvadi, UPEI
https://goo.gl/pZZwJ2
Agenda

- Library data - what data, why visualize, some examples
- Basic principles of data visualization, types of charts
- Steps and tips
- Google Data Studio
Revealing patterns, not telling “stories”

"Every block of stone has a statue inside it and it is the task of the sculptor to discover it." - Michelangelo

What trends, patterns, relationships are in your data that your visualization will reveal?
Library Data - What patterns do we need to reveal?

- Budget
- Circulation activity
- Collections
- Instruction
- E-Resource usage data (COUNTER, Proxy server, other)
- Interlibrary loan
- Services: virtual reference, staff-logged transactions, equipment checkouts (e.g. laptops), study rooms bookings
- Website activity (Google Analytics)
- Gate counts

Metrics that Matter!
Basic Principle of Data Visualization

Keep it Simple
Which is easier to understand?
Chartjunk

From Stephen Few Show Me the Numbers
Kinds of Data

- **Numeric data**
  - e.g., FT uses of a journal
  - can do arithmetic on - add, median/mean, min/max - meaningfully
  - may be absolute or percentage
  - don't be fooled by data that looks like numbers but isn't, see below

- **Categorical data**
  - labels like different collections of books: Stacks, Reference, Oversize
  - includes "nominal" data that looks like numbers but isn't really, like ISBNs, room numbers
  - may have a "natural" order like LC classes, rainbow colors
  - often best to present in “most to least”

- **Dates/times**
  - can be “countable” (e.g. median pub years) or categorical (month of year);
  - has natural order that should override “most to least”
Basic Principles of Data Visualization

Know what charts are best for what data

- Bar charts - comparison of categorical things (LC classes, types of patrons, databases)
  - horizontal bars for bigger labels
- Line charts - trend over time (budget, circ or any service activity, e-resource use)
- Heat maps - visualize tiers of quantities across a matrix or literal map
Basic Principles of Data Visualization

Avoid any type of chart that displays 2-d area in data values

Principle: people don't judge area as quantity very well - we're linear (length)
Bar Charts

- Vertical or horizontal - horizontal best when labels don't show well in vertical
- Good for showing numeric data in categories
- Always start Y scale at zero

Data from UPEI Library Website Google Analytics using Google Data Studio
Grouped bar chart - time as category

Questions by Time of Day

- **Reference**: 
  - 1 Morning: 66
  - 2 Afternoon: 167
  - 3 Evening: 35

- **Tech**: 
  - 1 Morning: 453
  - 2 Afternoon: 932
  - 3 Evening: 305

- **Circ**: 
  - 1 Morning: 1,110
  - 2 Afternoon: 1,874
  - 3 Evening: 883
Bad "bar" charts

People see area, obscures linear values
Stacked bar chart:
How much work to see trends in each data set?
Stacked bars and percentages

Does Web of Science have more mainstream journals (blue) than Google Scholar?

(type of documents)

- **Mainstream journals**:
  - Web of Science: 40
  - Scopus: 44
  - Microsoft Academic: 47
  - Google Scholar: 47

- **Secondary journals**:
  - Web of Science: 6
  - Scopus: 12
  - Microsoft Academic: 22
  - Google Scholar: 22

- **Book chapters**:
  - Web of Science: 1
  - Scopus: 2
  - Microsoft Academic: 7
  - Google Scholar: 20

- **Conference papers**:
  - Web of Science: 7
  - Scopus: 7
  - Microsoft Academic: 15
  - Google Scholar: 15

- **White papers (website)**: None listed

- **Software program**: None listed

- **Newsletter articles**: None listed

- **Company report**: None listed

- **Journal ranking**: None listed
Instead of stacked bar when lots of data

Panel bar chart - keep the scales the same!
Line charts

Best for showing data with continuity, usually time, especially a few sets of related data
Line Charts

- Usually start Y scale at zero or show it if go negative
- Intervals must be equal in size (e.g. years, months, quarters, etc.)
- Average slope of 45 degrees is often good

https://eagereyes.org/basics/banking-45-degrees
Why stacked area is bad

Look at the blue data - are Central sales going up?
Pie Chart

Use rarely, with very few slices, data labelled

- Checked out in first 5 years: 67.6%
- Checked out after 5 years: 14.8%
- Never checked out: 17.6%

Video: Salvaging the Pie - http://www.darkhorseanalytics.com/blog/salvaging-the-pie
Which helps you understand the data better?
Waffle Chart - if you really really want a pie chart

Only do with percentages adding to 100%

Use squares, not any other shape.

If you try this with fixed amounts per square (not percents), that's a "unit chart"
Heat Map

To illustrate patterns in a set of data that can be arranged in 2D, often a matrix or map.

Reference Desk Transaction Counts Fall 2010-Spring 2012
Radar Charts: Very bad!!

From UPEI's 2013 LibQual report.
When to just use a Table instead of a Graphic Chart

- Users want to see individual values, not just trends and patterns
- Users want to see precise values
- Values need to be shown in various groupings, e.g., subtotals
- Data includes different units, e.g., dollars, percentages, unit counts

Charts are best when it's the trend, pattern, or relationship among the data points, not their precise values, that matters most

Taken from Few, *Show Me the Numbers*
Steps to Make Good Visualizations

1. Plan what data you need and how to get it
2. Acquire raw data
3. Filter, refine, and process for use/ingest into charts/software
4. Ingest into display tools and create visualizations
Tips specific to library data

Circulation:

- Querying books with zero circ versus any more than zero circ - may take 2 queries;
- Consider median versus mean; whether to let zero circs pull down averages
- Do you count renewals?
- Eliminate collections without circ (reference) or artificially high circ (reserves) to avoid skewing results
Tips specific to library data

Transaction services (service desk, virtual reference, etc.):

- By hour of the day, day of the week, week/month of the semester?
- By academic calendar, December, Summer?
Tips specific to library data

Collections:

- Count volumes or titles?
- Use MARC for call number or other part of data in ILS (e.g. "item" record)?
- Accuracy of MARC records - fixed field/leader data, other

COUNTER data:

- PDF, HTML, or Total FT? Platform UI effects (COUNTER 5 may fix)
- Hacker impact - how to adjust for, e.g. SciHub
- BR2 - book section, comparable to a print checkout? Define/declare.
Tips specific to library data

LC Class:

- Too many for any useful graph - how to handle?

Late bloomers by LC - 5 Year

![Late bloomers by LC - 5 Year](image)
Step 2 - Acquire Raw Data

- Prefer delimited (txt vs tsv problem)
- Log files - may have to parse to delimited
- Real-time data through APIs - e.g., Google Analytics - changes your chart
Step 3 Filter, Refine, Process

- Best Practice: Copy your raw file first so you can always go back to the original!
- Clean up overall structure to get one header row only
- Remove "bad" data (e.g. pub years in ISBN field of MARC record) - sort!
- Remove outliers if suspect cause (e.g. SciHub hacks in COUNTER data) or will distort chart too much (document!)
- Remove columns that you don't need if very large data set
- Extract needed pieces of data from more specific units, e.g. LC class from entire call number; month, year from precise dates of transactions
Step 3 Filter, Refine, Process (cont.)

- Make sure numbers are stored as numbers, text is stored as text (ISSNs without hyphens), dates as date values
- Combine data from multiple tables into a single one (e.g. ebook usage and price)
- Clustering/cleaning (e.g. publisher name authority)
  - openrefine.org can be helpful for this - free desktop software
  - Pivot to find variances, Data-Filter to change values in batches
Tips for Refining Data in Google Sheets/Excel

- Copying column value down - corner "+" - column to left must have no blanks
- Spreadsheet tips for librarians - normalizing ISSNs, combining data from multiple sheets (treating spreadsheets like databases), extracting LC Class from full Call Number, sorting by LC Call Number, more in: https://goo.gl/obCsTp
- Move columns around to make contiguous for easier selection
- Date formulas to extract day of week, month, etc. - see tips file
- Formula down column then Copy / Paste Special - Values Only - reduce load
- Pivot Tables, with filters on irrelevant data, then Copy/Paste Special to another sheet - easier to make good header row, etc. for chart
Google Data Studio

https://datastudio.google.com/

Does not appear with usual list of Apps, not even "more"

Can connect to Google Sheets, Google Analytics, many other data sources
Google Data Studio linked to Website Analytics

Date filter, custom filter for audience to use

[Charts and graphs showing data analysis, including:
- Sessions: 242,308
- Visits Per Month
- Visits by Hour of Day
- How did they get here?
- Filter all results by Operating System:
  - Windows: 130.6K
  - Macintosh: 81.2K
  - iOS: 10.9K
  - Android: 8.4K
  - Chrome OS: 1.2K
  - Linux: 570
  - (not set): 246]
Steps:

1. Make sure your data is cleaned up, only one header row!
2. Create new report
3. Create new data source - select what to link to - must select one worksheet at a time from a spreadsheet with multiple worksheets/tabs
4. If needed create new "metrics" from text columns (e.g. day of week)
5. Connect the data source to the report (can add multiple sources to a single report)
6. Start to add charts, text labels, etc. Can have multiple pages.
Recommended Reading

Questions?

This presentation: https://goo.gl/pZZwJ2

Spreadsheet tips for librarians: https://goo.gl/obCsTp

My contact info: Melissa Belvadi mbelvadi@upei.ca