Using Data Visualization to Humanize Large Datasets

Bethany Yollin, Data Scientist
Nicholas Richter, Systems Analyst II

Presenters: WSP Data Informatics Group (DIG)

Bethany Yollin
Data Scientist

Nicholas Richter
Systems Analyst II
(Part of) What We Do

— We work with a variety of large datasets:

  - Tolling transactions
  - Public transit and ridership
  - Transit asset management
  - Project financials

— We use automation, data analysis tools, and data visualization to take data and turn it into information that people can use

— We build tools to help monitor activity in datasets that cannot be manually checked

What is a “Large Dataset”?

— A “large dataset” is any amount of data that challenges your ability to understand and use its contents

— What is a “large dataset” may be a small dataset for others

<table>
<thead>
<tr>
<th></th>
<th>Basic Data Skills</th>
<th>Advanced Data Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000 Rows</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>500,000 Rows</td>
<td>😐</td>
<td>😊</td>
</tr>
<tr>
<td>750,000,000 Rows</td>
<td>😞</td>
<td>😊</td>
</tr>
</tbody>
</table>
Current Trends

The Data Tsunami

— Data of every type is being collected in greater detail and at a greater rate than previously in history

— Working with data is a challenge that organizations of all types and in every industry face

— The value of the data is constrained by several factors:
  — The quality, consistency, and coverage of the data
  — How the information is communicated to the target audience
  — The skill of the data analyst
Skills x Format

— Your starting point is a function of data and skills
— Good data is an asset; bad data is a burden
  — Good data collects what is needed, accounts for differences over time, and remains **consistently formatted**
— You don’t have to be a wizard to benefit from good data
  — …but it helps

<table>
<thead>
<tr>
<th>Basic Data Skills</th>
<th>Advanced Data Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust, Well Formatted Data</td>
<td>Skills limit its value</td>
</tr>
<tr>
<td>Bad Data</td>
<td>Data is a burden</td>
</tr>
</tbody>
</table>

Skills x Presentation

— Regardless of your data, how you present makes a huge difference
— Presenting the data using visualizations and other tools can make massive amounts of data easily accessible to people who don’t have the same data skills

<table>
<thead>
<tr>
<th>Basic Data Skills</th>
<th>Advanced Data Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presented Well</td>
<td>Gets the job done</td>
</tr>
<tr>
<td>Bad Presentation</td>
<td>Confusion</td>
</tr>
</tbody>
</table>
At a glance

Our Focus Today

— **Gauging Your Audience**
  — Understanding what level to present at

— **Picking the Right Tool**
  — Getting your data ready to be used
  — Using the right tool for the job

— **Picking the Right Visualization**
  — Using the visualization that enhances communication
  — Avoiding confusion

Section 1

Humanizing Data: Gauging Your Audience
The Spectrum of Audiences

- Generally, the higher up, the less interest in detailed, nuanced materials
  - Exceptions exist!
- Documents should be tailored to be under the curve of the audience wonkiness capacity

**Summary**

<table>
<thead>
<tr>
<th>Audience</th>
<th>Approach</th>
<th>Analogy</th>
<th>Level of Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>Initial Communication at Low to Medium level of “wonkiness”, but high level available as appendices as appropriate</td>
<td>Throw Everything on the Wall</td>
<td>Low to Medium, with high backup</td>
</tr>
</tbody>
</table>
| Co-Workers    | **For co-workers engaged in the data:** High detail, with common nomenclature  
                **For co-workers less engaged in the data:** Summaries, with common nomenclature | Initiative Text | Summaries to High Detail |
| Managers      | Medium detail for the factors critical to the topic, summaries/explanations for factors filtered from communication | Voter Pamphlet | Specific Medium |
| Leadership    | Summaries and staff conclusions about topic at hand, be ready with follow up questions | Elevator Pitch | Low |
Section 2

Humanizing Data: Picking the Right Tool

Data Wrangling & Analysis

— Data wrangling is the process of transforming data to an appropriate and usable format for further analysis and visualizations

— Very often the most time-intensive step in the data visualization pipeline

— Jeffery Heer (University of Washington) outlines five steps of data wrangling:

1. Discover content and patterns
2. Structure and format needed attributes
3. Clean to eliminate meaningless outliers
4. Enrich with data that adds context and meaning
5. Validate by testing against logical constraints
The Right Tool

— There are many data visualization programs out there. Some are even free!

— Picking the right tool should consider:
  — How much do I want to pay?
  — Can all of my data be public?
  — Does this tool allow me to create the visualizations I want?
  — How often will we be using/updating this visualization?
  — Where are my skills at with the tools I could use?
  — How much time do I have to invest in learning the tool?
  — What tools do we have right now?

— All tools have drawbacks, even industry leaders

Some Current Options

<table>
<thead>
<tr>
<th>Capabilities</th>
<th>Ease of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>R (0)</td>
<td></td>
</tr>
<tr>
<td>Javascript (0)</td>
<td></td>
</tr>
<tr>
<td>Tableau (0-$$)</td>
<td></td>
</tr>
<tr>
<td>Power BI (0-$$)</td>
<td></td>
</tr>
<tr>
<td>Google Data Studio</td>
<td></td>
</tr>
<tr>
<td>Excel</td>
<td></td>
</tr>
</tbody>
</table>

R & JS are free and extensible, but require programming experience.
Points off due to restrictions of getting the program - .gov emails restricted, Microsoft Online admin must approve.
### R

- R is a statistical computing and graphics environment that boasts a large and active open-source community.
- R and RStudio are free without restrictions, and are capable of enabling users to develop publication-quality static graphics and web-based interactive dashboards.
- R is a scripting language and requires programming skills.
  - Can have a steep learning curve for new programmers.
  - RStudio, a popular interface, "makes R easier to use" vs. "easy to use".
- A great choice for someone looking to define their career on data analysis and visualization.

### Tableau

- Tableau (made in Seattle!) is a very intuitive package for data visualization, but is pricey ($420 to $1345 per year).
- Publishing to the web requires a more expensive license. Most visualizations result in static images or PDFs.
- Offline visualizations are possible, but require people receiving the file to have a special program (Tableau Reader).
- A free version exists (Tableau Public), but comes with a major caveat: **All data loaded to Tableau Public is public**.
  - Anyone can download Tableau Public data!
  - Tableau Public can be a good fit if you are typically visualize data that is considered public domain, does not contain any sensitive information.
Excel

— Everyone is doing it

— Limited charting and data visualizations – Not as intuitive as other options

— You will grow out of it (or your data will)

— Used predominately for quick, small summaries, such as with Pivot Tables/Charts
Design principles

— Design is a search problem. What do I mean by this?
  — Design is iterative
  — Seek feedback from your colleagues and invite criticism
  — Visualize your data early and often
— There are many ways to visualize your data and you should carefully evaluate your intended message and decide the best visualization for your audience, but a few guiding principles should be observed:

  - Simplicity
  - Clarity
  - Consistency

Distribution

— Shows range and frequency of observations, as well as outliers
— Useful when trying to categorize an entire observed population
  — “My model predicted the travel time for 50,000 cars. What’s that look like?”
— Easy to see measures of central tendencies including mean, median and mode or measures of variance like range and standard deviation
Each color shows how 100% of the modeled travel times are distributed.

Distribution Example: Density

Estimated Travel Times T/Th PM Peak May 2018
Grady Way & Rainier Ave S to Tukwila Int'l Blvd Station via Southcenter Blvd
Source: Google Directions API

Distribution Example: Airline Departure Performance

Average Delay by Carrier over the Month Nov 2017

This particular box plot shows the median, the middle two quartiles (box), and all data points within 1.5x of the interquartile range.

Outliers are easy to pick out.
Part of a whole

— Shows proportions or percentages
— Avoid showing too many categories, especially in pie or doughnut charts
— Watch out for labeling challenges with treemaps

Pies and Donuts: Better as Desserts

— Pie charts are OK as “throw away” quick glances across 3 to 5 categories, but should be avoided overall.
— Pie charts are especially bad at communicating changes.
   — Comparing slices between two charts is difficult
   — Stacked bars and donut charts are better
Correlation

— Shows a relationship between two or more variables
  — Avoid over-plotting, i.e., showing too many points – sampling can help in this case
  — Alternatively, add transparency to your points
— Remember: correlation does not imply causation
  — Even though we know this, your audience may draw their own conclusions

Correlation Example: Motor Trend Correlogram

— Visualizes the strength of correlation between many variables
— Great for exploratory data analysis
Beware Spurious Correlation!

Per capita consumption of mozzarella cheese correlates with Civil engineering doctorates awarded

http://www.tylervigen.com/spurious-correlations

Evolution

— Generally used to show change over time or iterations of something

http://www.r-graph-gallery.com/
Evolution Example: Stacked Area Time Series

Flow

— Visualizes the relationships between different things and how things move between them

— Examples include how funds are allocated, migration patterns, and things with other types of relationships
Sankey Diagrams: Visualizing Distributions

— Sankey diagrams are very effective at communicating how things are distributed from one group to another

Network Map

— Network maps can be used to map relationships

— Using software, patterns based on the relationship network emerge from naïve data

— This is a program visualized to help understand which parts are dependent on other scripts or data tables

— Colors indicate different parts of the program
Maps

— Basemaps are crucial for contextualizing place
  — Free basemaps are available from OpenStreetMaps, CartoDB, etc.

— Maps are part of every day life and easily understood by most people

Map Example: ORCA Zone-to-Zone Flow
Interactive Map
Summary

— “Big” data is a challenge that all organizations face

— Mile long spreadsheets are oppressive and make data seem like a daunting problem

— The key to tackling and communicating your “big” data is a combination of:
  — Developing your data skills
  — Understanding what your audience can process
  — Selecting the right visualization/analysis
  — Selecting the right software to assist with execution
THANK YOU