Process and Workflow of Data Visualization
Part 2

John Keyser

Four Stages of a Visualization Project

- Formulating a brief
  - Definitions and requirements
- **Working with data**
  - **Characteristics and qualities of data**
- Establishing your editorial thinking
  - What will you show?
  - Bridge between data work and design work
- Developing your design solution
  - Creating the visualization itself

Kirk, *Data Visualisation: A Handbook for Data Driven Design*

What is data?

- This is all the “stuff” we have to work with
- It can take many forms
    - Values/numbers
        - Of various types
    - Labels
    - Relationships
- What gives meaning to these forms are the semantics
    - How to interpret the forms – what their meaning is
- There are a lot of different ways to think about and categorize data

Entity-Relationship Model

- One way to look at data; common in database modeling
- Entities are the things of interest
  - Can be abstract “things” – ideas,
- Relationships describe how entities are related
  - Many types: causal, temporal, hierarchical, etc.
- Attributes are attached to entities and relationships
  - Can have many dimensions

Independent vs. Dependent Variables

- A useful way to think about data
- Independent variables:
  - In an experiment, a thing that changes by the experimenter
  - Entities, items, axes
  - e.g. person, location, parameters
- Dependent variables:
  - The things that are associated with some particular set of independent variables
  - Attributes, measurements, assigned values
  - e.g. color, temperature, name, return value

Data Classes

Taxonomy defined in 1946 *Science* article by statistician S.S. Stevens: “On the theory of scales of measurement”:

- Categorical (nominal)
- Ordinal (ordered)
- Quantitative – Interval
- Quantitative – Ratio

Data Classes

- Categorical (nominal)
  - Discrete values
  - No defined ordering
  - e.g. chair/table/bed; quiz/homework/test
- Ordinal (ordered)
- Quantitative – Interval
- Quantitative – Ratio

Data Classes

- Categorical (nominal)
- Ordinal (ordered)
  - Discrete values
  - Can be ordered
  - e.g. Small/Med/Large; Bronze/Silver/Gold
- Quantitative – Interval
- Quantitative – Ratio

Data Classes

- Categorical (nominal)
- Ordinal (ordered)
- Quantitative – Interval
  - Usually “continuous” values (real or integer)
  - Differences between them are meaningful; absolute magnitude of numbers is not
  - e.g. time, degrees Fahrenheit
- Quantitative – Ratio

Data Classes

- Categorical (nominal)
- Ordinal (ordered)
- Quantitative – Interval
- Quantitative – Ratio
  - Usually “continuous” values (real or integer), with a 0 value reference
  - Can perform math, make absolute comparisons
    - e.g. A is twice the size of B
  - e.g. mass of an object; money

Some Other Data Classes

- Uncertainty
  - May be attached to quantitative data types
  - Mostly found in some scientific data
- Operations as data
  - Sometimes the things done to data are themselves data
    - e.g. A modification history
  - Can be very difficult to use, but can be a source of powerful visualizations
    - e.g. Understanding how something came to be

Ordering Data

- Sequential
  - Clear ordering from min to max
- Diverging
  - A “neutral” point (a 0) with values diverging away
- Cyclic
  - Repeating data (e.g. months of a year)

Datasets

- How the data is provided
  - Several ways possible, but there are some common ones
- Tabular data
- Network (graph) data
- Spatial data
- Collections

Datasets

- How the data is provided
- Tabular data
  - By far the most common
  - Each row is an item/entity
  - Columns identify:
    - The attributes (values associated with an item)
    - The independent variables that uniquely identify that item/entity
- Network (graph) data
- Spatial data
- Collections

Datasets

- How the data is provided
- Tabular data
- Network (graph) data
  - Describes relationships
  - Nodes and edges
    - Can be a tree, in which case there might be a hierarchy
  - Attributes can be attached to nodes and edges
- Spatial data
- Collections

Datasets

- How the data is provided
- Tabular data
- Network (graph) data
- Spatial data
  - Attributes measured across space, with geometry assumed
  - Samples taken in a regular grid pattern, or at selected locations (set of measurement points, or by region)
  - Common for scientific data
- Collections

Datasets

- How the data is provided
- Tabular data
- Network (graph) data
- Spatial data
- Collections
  - Sets
  - Ordered lists
  - Clusters

Data Formatting

- To understand data, need to know how to interpret it:
  - – Is the date 2-3-57:
    - February 3, 1957?
    - March 2, 2057?
  - – Is the time 13:25, or 1:25, or 1:25p, or 1:25 p.m.? And, which time zone?
- And remember, it might not be consistent across data sets or even within data sets!

Data Quality

- For “real world” data, it is often not “clean”
- Anything a human entered or scanned (e.g. via OCR) is likely to have errors
- Many types of issues:
  - Missing data points
  - Inconsistent references
    - e.g. TX vs. Texas
  - Misspellings, Mis-entered entries
  - Exceptional values/mis-typed entries
    - e.g. NaN, “unknown”
    - e.g. The “number” 00, or 007
  - Special/whitespace characters
  - Out-of-date info (e.g. age)
  - Capitalizations

Data Abstraction

- Understanding how the specific data fits into the more general, abstract, categories
  - This will determine what visualization methods are appropriate
- Analyze the data, ask questions about it

Questions to Answer

- What are the dataset types, what are the attribute types?
- How many entities are there?
- What is the range of each attribute
  - Number of discrete values, or range of quantitative values

Transforming Data

- Very common practice to get data into a more “usable” format
- Taking the given data and generating new data from it
  - For many possible reasons
  - Replacing or adding to the original data
- Keep in mind that as you do so, you are introducing some decision/bias/human input into what is there

Reasons to Transform data

- Transform to clean
  - “Fixing” problems in the original data
    - e.g. one value for TX, TEXAS, Texas, Tex.
  - Removing unneeded attributes/entities/etc.
    - e.g. phone number might not be used
- Transform to convert
- Transform to create
- Transform to consolidate

Reasons to Transform data

- Transform to clean
- Transform to convert
  - Convert from one data type to another that’s more usable
  - e.g. Temperature value to cold/warm/hot
  - e.g. Breaking Name out into First and Last
  - e.g. Converting text into categorical data
- Transform to create
- Transform to consolidate

Reasons to Transform data

- Transform to clean
- Transform to convert
- Transform to create
  - Generating new data from operations on original (some overlap with “converting”)
  - Derived from the original data
  - e.g. Use departure and arrival times to generate a duration of trip attribute
  - e.g. Running linear regression, other statistical techniques to get more data
- Transform to consolidate

Reasons to Transform data

- Transform to clean
- Transform to convert
- Transform to create
- Transform to consolidate
  - Combining information from multiple sources (like joins in databases)
  - Adds entities or attributes to data
  - e.g. add the department each person belongs to (allowing display by department)

Examples