Data Analysis Course 6: Share Data through the Art of Visualization

Module 1

Data visualization: the graphic representation and presentation of data.

- putting information into an image to make it easier to understand

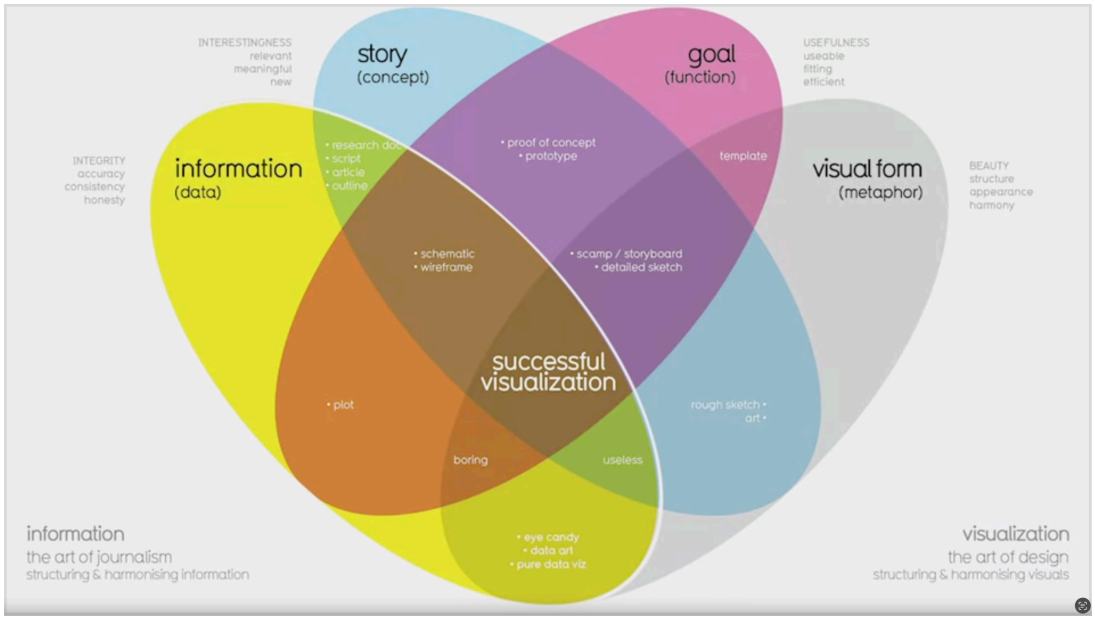
Visualizations began with maps

Scientists and mathematicians began to truly embrace the idea of arranging data visually in the 1700s and 1800s

How data analysts use visualization:

1. Looking at visuals in order to understand and draw conclusions about data
2. Creating visuals using raw data to tell a story

Rule of Thumb for Visualizations:
Creating visualizations: your audience should be able to tell what they're looking at within the first 5 seconds of looking at it.
In the next 5 seconds, they should be able to recognize the conclusion your visualization is making.



Story and data combined provide an outline of what you're trying to show.

Two frameworks for organizing data

- Frameworks help organize your thoughts about data viz and give a useful

checklist to reference as you plan and evaluate your data viz.
- McCandless method: provides terminology to isolate specific elements of a graphic.
  - Information: the data with which you're working
  - Story: a clear and compelling narrative or concept
  - Goal: a specific objective or function for the visual
  - Visual form: an effective use of metaphor or visual expression
- Kaiser Fung's Junk Charts Trifecta Checkup: questions to determine viz's effectiveness.
  - What is the practical question?
  - What does the data say?
  - What does the visual say?

Pre-attentive attributes: the elements of a data visualization that people recognize automatically and without conscious effort.
- Essential, basic building blocks that make visuals immediately understandable are called marks and channels.
  - Marks: basic visual objects such as points, lines, and shapes. Every mark can be broken down into four qualities:
    - 1. Position: Where is a specific mark in space relative to a scale or other marks?
    - 2. Size: How big, small, long, or tall is a mark?
    - 3. Shape: Does the shape of a specific object communicate something abut it?
    - 4. Color: What color is a mark?
  - Channels: visual aspects or variables that represent characteristics of the data in a visualization. Specialized marks that have been used to visualize data. Three elements determine effectiveness:
    - 1. Accuracy: Are the channels helpful in accurately estimating the values being represented?
    - 2. Popout: How easy is it to distinguish certain values from others?
    - 3. Grouping: How effective is a channel at communicating groups that exist in the data?

Bar graphs: use size contrast to compare two or more values.
Line graphs: help your audience understand shifts or changes in your data.
Pie chart: show how much each part of something makes up the whole.
Maps: help organize data geographically

One of your biggest consideration when creating a data visuzliation is where you'd like your audience to focus.

Histogram: a chart that shows how often data values fall into certain ranges.

Correlation charts: show relationships among data.

- Correlation in statistics is the measure of the degree to which two variables move in relationship to each other.
- Causation: occurs when an action/event directly leads to an outcome.
  - Correlation ≠ Causation
  - Be careful not to show causation where it doesn't exist.
  - In your data analysis, remember to:
    - Critically analyze any correlations that you find
    - Examine the data's context to determine if a causation makes sense (and can be supported by all of the data)
    - Understand the limitations of the tools that you use for analysis

Time series charts
Ranked bar charts

Static visualizations: do not change over time unless they're edited.

Dynamic visualizations: visualizations that are interactive or change over time

Tableau: a business intelligence and analytics platform that helps people see, understand, and make decisions with data.

Scatterplots show relationships between different variables. Scatterplots are typically used for two variables for a set of data, although additional variables can be displayed.

Distribution graph displays the spread of various outcomes in a dataset.

Considerations for what you want to communicate:

- Change: This is a trend or instance of observations that become different over time. A great way to measure change in data is through a line or column chart.
- Clustering: A collection of data points with similar or different values. This is best represented through a distribution graph.
- Relativity: These are observations considered in relation or in proportion to something else. You have probably seen examples of relativity data in a pie chart.
- Ranking: This is a position in a scale of achievement or status. Data that requires ranking is best represented by a column chart.
- Correlation: This shows a mutual relationship or connection between two

or more things. A scatterplot is an excellent way to represent this type of data pattern.

Decision tree: a decision-making tool that allows you to make decisions based on key questions that you can ask yourself. This can help you make decisions about critical features of your visualization.