Module 3: Data Visualization

Ellen Bledsoe

2023-02-21

A Visualization Primer

Why Does Data Visualization Matter?

On your own

Take a few minutes to write down your top 3 reasons why data visualization is important or useful to you

Small Groups

As a group, take 5 minutes to compare your notes and come to a group consensus on your top 3 reasons. Be sure to choose someone to report out!

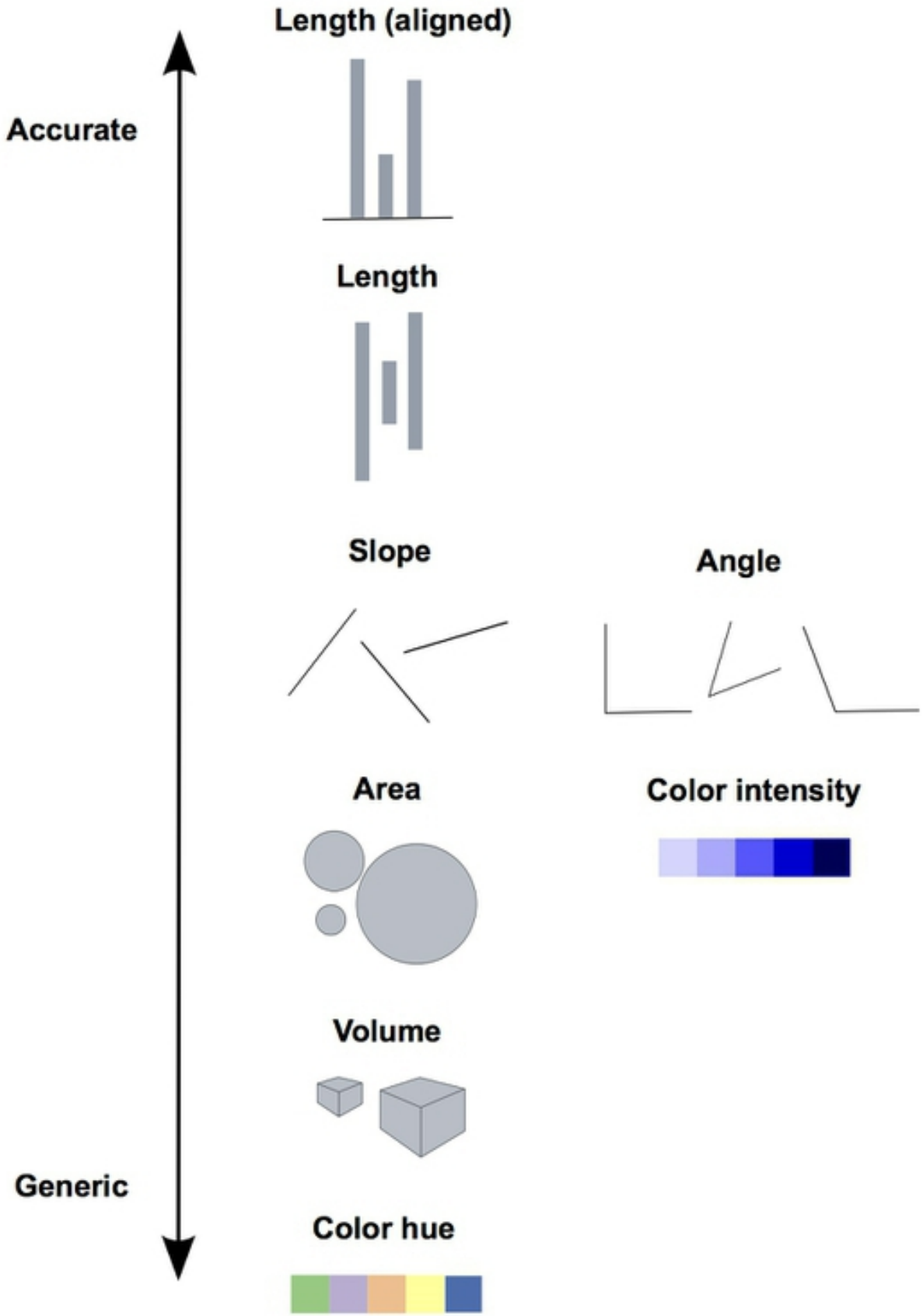
Types of Visualization

Visual cues for communicating data

Here is a decent overview of some of the core concepts of data visualization.

This website is pretty great and goes into a lot of detail about good practices in data visualization. If this is something that really piques your interest, I encourage you to check it out!

In this course, our main take-away from this website is the use of visual cues to communicate data and which ones are better than others.


Length (aligned)
Accurate
Length
Slope
Angle
Area
Color intensity
Volume
Generic
Color hue

Small group activity

In groups, discuss as many all types of visualization that come to mind. It's okay if you don't know what they are called! Make a quick list or draw them out if you prefer.

Types of data visualizations: bar graphs, bow-and-whisker plot, pie chart, line graph, histogram, pyramid chart, gantt chart, scatter plot, time series, word cloud, venn diagram, heatmap, density plot, topographic map

This website is an amazing reference for data visualization methods and when to use what. It also has examples of each type plotted in `ggplot2`.

We will talk more about how to choose the right visualization for your data now and also in the rest of the module.

Data matching activity

Still in your groups, open the PDF called "data_viz_matching_exercise"

Spend a few minutes seeing if you can match the data descriptions to the types of data visualizations. Note: these plots do not represent the data, just a type of data visualization.

This is a challenging exercise, so don't panic if you feel a little lost! Here are some things to help you think through which visualizations might be appropriate:

- how many variables are in the data description?
- how many variables are represented in the data visualization? Look at the axes, sizes, colors, etc.
- how is *one* data point represented on the plot?
- how is variation in the data represented on the plot, if at all?

Matches

1. B
    - river plots or a chord plot
    - good for showing movement from one thing to another or relationships between two categorical variables
2. D
    - 4! variables can be shown here—circles could be size of family
    - scatter plots are good at showing relationships between multiple *continuous* variables
3. C
    - Multiple samples for each treatment means we will want to show variation
    - Variation in the multiple samples is show through violin plot; show the mean *and* the distribution of the data
    - One categorical variable and one continuous variable
    - Second categorical variable could be shown through color of violin plot
4. E
    - Heatmaps are good at showing spatial distributions of data
    - x and y axes shows space; the color indicates continuous values
5. A
    - Bar plots are good for showing change through time
    - Categorical on x, continuous on y
    - mutliple stacks, 2 bars per year?

Figure Critique

First, read through this blog post on the "Dos and Don'ts of Data Visualization"

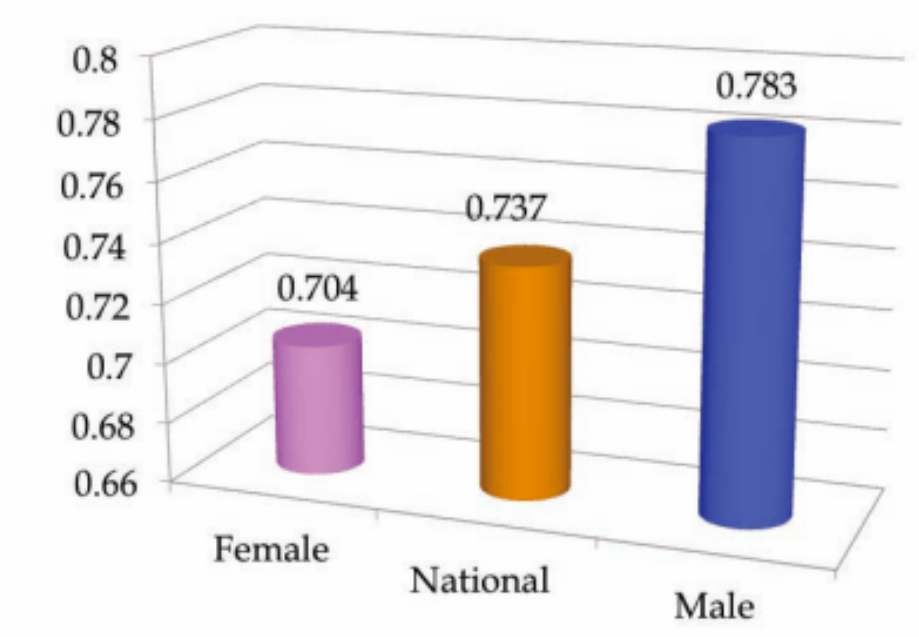
Now that you have some insights, let's critique these figures below. Yes, these are actual figures in the wild...

- what aspects don't work
- what aspects do work
- how how you present the data?

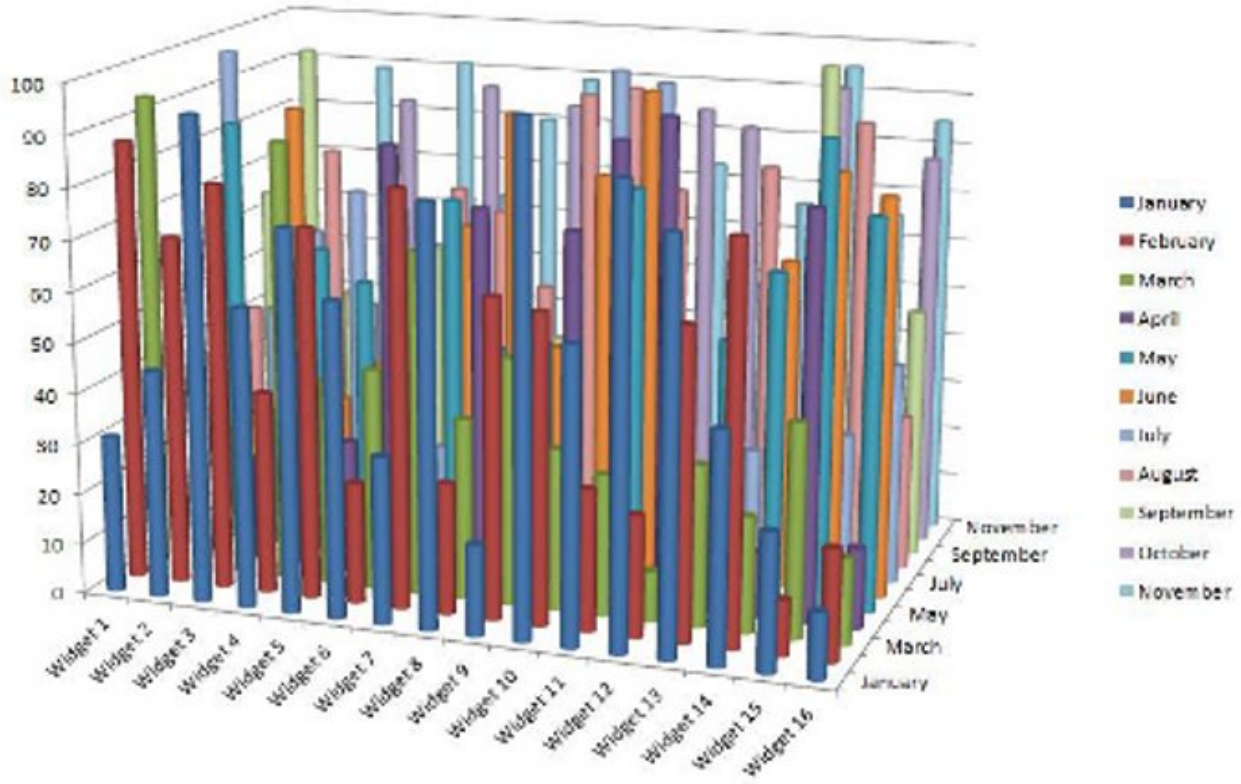
Figures:

1.

Figure 11: GNH index by gender



2.



3.

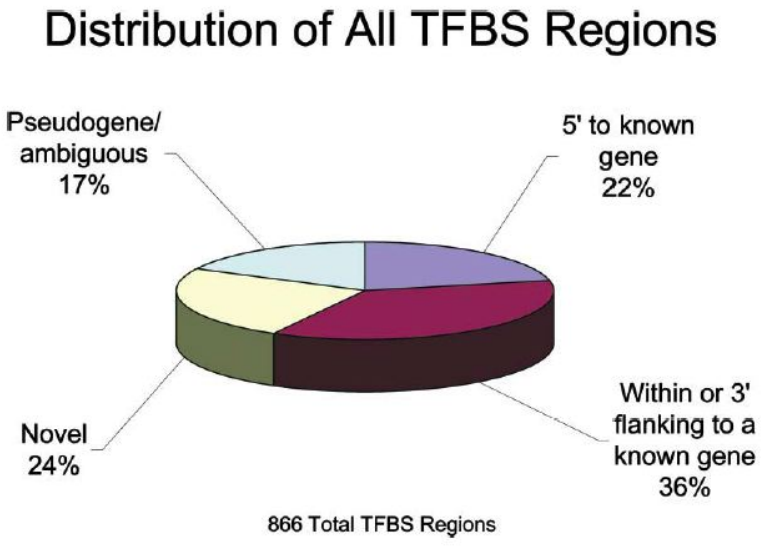


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

4.

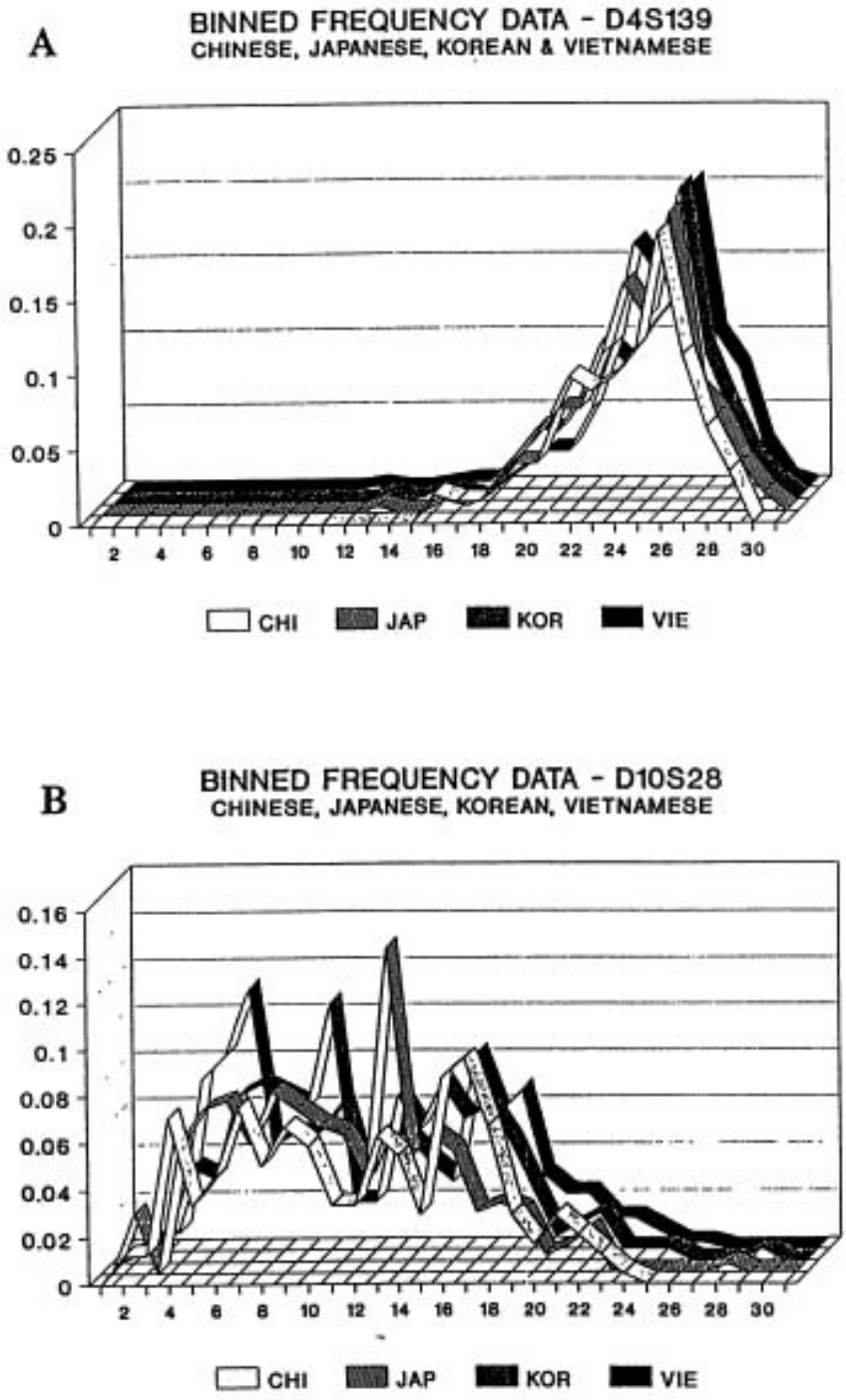


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the* 30 *bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*