2D vis lecture and workshop

Lecture:

Categories of data

When are we satisfied with 2D viz? Most of the time – unless the parameters interact in some interesting way that we want to explore/explain (more on this next week)

**Representing categorical data**

1. In general: simple frequencies (i.e. counts), percentages, averages

One of the most widely used visualisations

1. Parts of a whole (i.e. don’t use pie charts)
2. Subcategories (treemap)
3. Proportions over time (stacked continuus)

1D: bar chart – y axis can’t be anything other than count, bc we don’t have any other data to plot against

2D (discrete X, continuous Y):

Bar

[Boxplot

Dot plot

Violin] 3 of the same thing, essentially – when do you want each?

2D (both discrete)

jitter

**Representing continuous data**

2D (both continuous):

Jitter

Point

Quantile

Rug

Smooth (lm)

Function:

Area

Line

Step

**Representing distributions**

1D:

Area

Density

Dotplot

Freqpoly

Histogram

2D (i.e. bivariate dist)

Bin2d

hex

**Representing time series (?)**

Potentially use these graphics to show the progression (especially good because no design is used, very few labels, but bad because no code): <https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/>

OR: it could be a workshop activity – pick top 3, what is different about how they present the data/what do you like/how does it help you take home a message

**Workshop**

NO emphasis on design choices, no changing point sizes, font sizes, or colours this week – purely learning the different data representation types

NOTE: need to hold off on submitting by one week, so next week when covering effective data viz we will get them to alter one of their plots, but need to note this on the sheet – due not next Thursday, but the Thursday after)

One dataset, create each one of those plots making a point about what type of data it is (categorical, continuous, distribution, time series) (fine to use more than one dataset if one can’t satisfy)

Potenialls: <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>

Although mpgs is boring and they have already done R for Datascience ch 3-5 which I think uses it?

Then, we give another dataset and tell them to generate all the plots again (to submit)

Then, we give another dataset and ask three specific questions that can be answered with one of the plot types, but we don’t tell them which plot type. They have to generate, justify their choice, and answer the question (to submit)