Slide 1

- Intro

Slide 2

- Florence Nightingale recreation
- Statistician and Nurse
- Public health
- Hospital during the Crimean way
- red = disease, blue = other, green = wounds
- Source http://understandinguncertainty.org/node/214

Slide 3

- Coxcomb vs stacked bar charts

Slide 4

- Even the least interesting data set can be visually referenced and abstracted in a number of ways that impacts how the data is consumed.
- lengths in miles of a collection of US rivers
- source, built in base R data set referenced to World Almanac and Book of Facts, 1975, page 406.

Slide 5

- Tabular presentations provide granular access

Slide 6

- A stem and leaf plot provides more compact access, but involves rounding, abstracting us from the data a little bit.

Slide 7

- Shift the stem and leaf plots

Slide 8

- Moving to a dot plot gets us closer to a traditional visualizations
- y-axis has river length, x access has the number of rivers that match those lengths.
- We have trouble seeing the individual data points though.

Slide 9

- To solve for this, we can use a technique called jitter, that maintains our buckets, but puts random space between individual data points.

Slide 10

- Just how much jitter we use takes us from increased to decreased clarity.

Slide 11

- On top of jitter, we can then add colour; adding in redundancy to the message can increase the clarity of the message.
- At this stage, we have removed reference to the individual values, grouped the lengths by the number of counts, reinforced this grouping with colour, and added jitter to all for viewing of the individual data points.
- And although the scale is not exactly one to one here, we can see that they share an underlying common shape.

Slide 12

- A shape that can be further reinforced by grouping the river lengths by the hundredths and displaying this a histogram.
- This loses access to the individual data points, but highlights trends in length much more clearly.

Slide 13

- Flipping this 90 degrees gives us a more familiar presentation of a histogram, and adding in some grid lines allows for easier reading of counts.
- Through this little exercise, we can see that even with a uni-dimensional data set, one where all we have to work with is a single variable, there are a number of considerations to be taken in how we visualize that data and what we emphasize as a consequence:
  - granular data
  - rounded numeric values but increased compactness
  - representation of values with shapes and colours to see distribution
  - emphasis of the distribution by bundling data points together in a histogram

Slide 14

- Overview

Slide 15

- 3D visualizations are a topic for another discussion

- Infographics, while often lumped with data visualization, are arguably a genre of their own that incorporate data visualizations. Often these visualizations are not of the sort designed to accurately portray the underlying data, but instead emphasize visual appeal to reinforce a message.

Slide 16

- Open question

Slide 17

- Data -> visual information

Slide 18

- This is a process reliant on abstraction as objects other than the data are used to express the data

Slide 19

- This abstraction only makes sense if we have the literacies to both understand the underlying data and the abstraction used to convey those data.

Slide 20

- Open question

Slide 21

- It's generally accepted that data visualizations play 2 roles. They are used to explore and make sense of data sets that are otherwise too dense in information for us to be able to readily interpret them and identify patterns. And then, when we have relationships that we wish to communicate, they are one tool available to us.

Slide 22

- How do they help us do this exactly?
- By using a series of attributes to represent the data, we simplify the interpretive process.
- We need to pull out the relevant information for easy consumption, limiting cognitive load.

Slide 23

- As an example, how many fives are there in the following table?

Slide 24

- Using colour - in fact brightness of a colour - as an attribute to highlight these sections of the image, we lessen the interpretive barrier.

Slide 25

- The three attributes that data visualizations draw on include:
  - form
  - colour
  - position

Slide 26

- Form can take many forms, from specific shapes to ways of highlighting individual shapes, to representations of volume etc.

Slide 27

- Colour can be used in a variety of ways to highlight and differentiate either specific values or specific variables.
- Difference between a value and a variable?

Slide 28

- And position allows us to use cues of things like distance to derive information about the relationships between variables.

Slide 29

- Each of these three qualities of a visual are deliberately manipulated to reveal or obfuscate certain aspects of the underlying data the visual is trying to convey.

Slide 30

- There is virtually no end to the possible visualizations we might encounter in the sciences.
- We'll balance between what we might see in the scientific literature and what we might see in day to day communication of information to a non-descript public audience, where the overwhelming number of visualizations rely on bar charts.
- Starting with common visualizations, we'll look at a few key considerations in creating
  - bar charts
  - histograms
  - line charts

  - scatter plots or dot plots
  - and pie charts

Slide 31

- But as each type of visualization is catered to a specific kind of data, we'll first revisit the basic data types we're generally dealing with, and then look at a visualization tool we might use for this kind of data.
- Data at it's most fundamental can be readily divided into categorical data and numeric data.
- These then can be further broken down, categorical into whether or not the categories have an inherent order, and numeric into a slightly more complex matrix of those things that we count and those things that we measure one the one hand, and those things that start at 0 and those things that don't on the other.

Slide 32

- Start with categorical
- Nominal data has no order
- Ordinal data has order

Slide 33

- This impacts how we decide to group and order these data
- For nominal data, where we place the label is of less importance and we might opt for alphabetic, count etc
- For ordinal data, the order is very important in sense making and being able to draw comparisons between categories.

Slide 34

- When working with a single categorical variable, we often use bar charts, that use a count of the allowable values for that variable and volume (bars) to allow for visual comparison.

Slide 35

- We'll start with a data set. This is a data set called penguins and has some basic biological measurements across three species of penguins on three islands in Antarctica
- Source: https://allisonhorst.github.io/palmerpenguins/
- What are the categorical variables?
- Are they nominal or ordinal?
- Dates are tricky and variables, situation depending, could represent different data types; it's a complicated topic.

Slide 36

- We'll demo with the species variable

Slide 37

- Every visualization application will have a default approach. Here, it's listing our variable values alphabetically.

Slide 38

- We may be able to better share these count data if we order by overall count.
- The default isn't always the best.

Slide 39

- Another data set that we have available to us is about diamond cut and quality.
- Source: built into R, no other attribution given
- What are the categorical variables?
- Are they nominal or ordinal?

Slide 40

- We'll use the colour variable here

Slide 41

- Following on our previous example, we might be inclined to order by count

Slide 42

- The documentation tells us D is the best and j is the worst

Slide 43

- A few things to consider when making bar graphs

Slide 44

- Categories are discrete, the bars should be discreet. A Histogram is something quite different.

Slide 45

- When working with counts, we're working with a hard 0 and using volume as a means of interpreting relationships between categories. We should start at 0 so as not to be misleading.

Slide 46

- We can build in redundancy, epxressing somethign about the data with more than one visual cue.
- Consider when redundancy may or may not be beneficial.

Slide 47

- Always include an x and y axis label, and always provide a full n for reference.
- You might consider providing an n on each bar as well.

Slide 48

- Comparisons across more than one categorical variable can also be done in bar charts, either by stacking or placing content side by side, adding multiple facets to each count
- The more values a given variable has, the more difficult these can be to easily interpret.

Slide 49

- Working with the diamond data, we'll visualize both cut and colour in this way.

Slide 50

- Which of these is most appropriate will change depending on the nature of the comparison, and the number of possible values associated with each variable.

Slide 51

- Next we hit on numeric data
- We'll divide this discussion into a couple of parts, noting first that numeric data may be either counted or measured. Counting works with whole objects, and groups of observations are easily achieved. Measuring captures non-whole objects and is resistant to grouping without simplification.
- Examples of count and measure data?
  - People
  - Temperature

  - Distance
  - Number of times someone does something; steps

Slide 52

- Numeric data can also be either interval or ratio.
- Ratio data has a meaningful zero point-or a known origin-and can be represented as a ratio that we do multiplication and division on.
- Interval data has no meaningful zero or known origin; we can do addition and subtraction, and we can know that a higher number means more of whatever is being measured, but can't quantify that difference as a ratio.
- Examples
  - Height - can be 0, and we can say someone is twice as tall.
  - Temperature in Celsius or Fahrenheit - 0 is arbitrarily set to the freezing point of water.
  - Temperature in Kelvin, however, is ratio data as it's zero point is tied to the absence of molecular movement, a known origin
  - Currency has a set point of 0 and is ratio data.
  - Number of people is also ratio data.
- discrete and continuous data may be either integer or ratio.

Slide 53

- When visualizing a single numeric variable, we often use a histogram - a bar chart with fused section, representing the continuity between objects counted or measured.

Slide 54

- For this example, we'll revist the penguin data set.
- question - is flipper length discrete or continuous, and integer or ratio?

Slide 55

- Unlike categorical data, numeric data, whether discrete or continuous, blends naturally from one measure or count to the next. A histogram represents this by not having space between the bars.
- And the bars in a histogram represent buckets of information. We could plot this as dots, capturing each individual data point, but histograms are valuable in clearly communicating distribution.
- This bucketing is a critical component of histograms though.

Slide 56

- The number of buckets we use can reveal or obfuscate different apsects of the underlying data.

Slide 57

- And just like with bar charts, we can stack histograms to reveal a more nuanced understanding of the data and to 'factor' the representation on a categorical variable. In this case, species.

Slide 58

- While bar charts and histograms provide one window into our data, when looking for relationships between variables, we often turn to dot plots, scatter plots, and line graphs.

Slide 59

- This is a great way to see how values change across space or through time. In this case, we'll start with some data on life expectancy in subset of countries for 2016.

Slide 60

- Since the year is a stable value, we'll plot life expectancy against country.
- question, is the categorical data nominal or ordinal? And the numeric data, discrete, continuous, interval or ratio?

Slide 61

- Using a bar chart vs a dot plot
- Since age doesn't represent a 'count' of amalgamated objects, the volume representation is not appropriate.

Slide 62

- There are many factors we might consider when deciding on how to organize the data, some based on convention, some based on ease of access etc.

Slide 63

- When looking at changes over time - in this case life expectancy over a 60 year period in Canada - we frequently see lines as opposed dots being used, giving the visual representation of continuity.
- Filling in the voids like this, does however abstract us away from the individual data points; easily seen in peaks, but not so easily seen in smoother sections.

Slide 64

- One way to address this would be to include both shape types, again, introducing redundency, but potentially getting a more informative visual display.

Slide 65

- One of things we need to consider, especially with things like line plots, are aspect ratio. We'll also start to breifly talk a bit more about colour.

Slide 66

- We have some weather data here for Kelowna in 2020.
- We should consider how apsect ration impacts how we interpret peaks and valleys in the data.

Slide 67

- We can make this a bit more extreme

Slide 68

- And we can have a significant impact on how data is read through colour choice.

Slide 69

- When reading scientific reports, we are rarely looking at just counts of and simple comparisons between variables.
- Usually we are looking at visuals that try to convey some descriptive or inferential statistics about the data. Or are otherwise displaying a calculated version of the data.
- We may also be looking at many variables simultaneously.
- We'll look at a few examples.

Slide 70

- Our data set here is a global data set of life expectancies and GDP, including things like population size.

Slide 71

- We'll start by mapping life expectancy to GDP to see if we can find a trend.
- What graph would you recommend?

Slide 72

- The basic graph.
- Someone is pretty far afield.
- There kind of looks like a pattern here, but we're missing a lot of information.

Slide 73

- Back to the drawing board.

Slide 74

- Let's add in a bit more information to see if we can figure if there are better questions we can start to ask.
- We'll add popultion size and continent. How might we do this?

Slide 75

- Colour coded by continent

Slide 76

- Size by population

Slide 77

- Increase clarity with some opacity

Slide 78

- The visual on the right is the one I was trying to copy.
- Anyone spot the difference?

Slide 79

- Closer look.
- A pattern looks alot more obvious now.

Slide 80

- We can plot a line of best fit to the data to investigate and potentially reinforce this relationship or pattern.
- It seems to work better for some continents than others
- We can do better

Slide 81

- If we break this up by continent, we can say that yes, in general, there appears to be a relationship between life expectancy and GDP.
- However, the story is also not this simple, as indicated by Africa.
- This might encourage us to propose new research questions that we would then test.
- Does this visual help us discern anything about population size's relationship to other variables? Does it detract from the visual?

Slide 82

- Another example, returing to the penguins data set.
- A historgram does a good job of displaying the distribution of data, but there are other graph types to address this.

Slide 83

- A very common one in the literature is a box plot.
- A box plot breaks the data into interquartile ranges. The middle bar tells us the median or most common value, the box covers 50% of the data, from the 25th to the 75th quartile range, and the lines, or whiskers are the remaing 50% of the data, covering the lower and upper quarters.
- One of things it doesn't tell us is what the concentration of data points is in each interquartile range.

Slide 84

- For this, we might turn to a violin plot, which is a density plot of the distribution of the data/
- We can see that the overall shape of the plots is similar, but with the violin plot we get a sense of the data's concentration as well as its spread.
- In an ideal world, we might want the best of both

Slide 85

- And we can do that, layering plot types.

Slide 86

- We might then also want to calculate the mean, as differences between the mean and median can tell us important characteristics about our data.

Slide 87

- And finally, back to the histogram.
- Thoughts on which you prefer?

Slide 88

- Next we hit the pie chart.

Slide 89

- Perhaps accessible, but hard to interpret than a bar chart.

Slide 90

- Colour is important.
- So far, we've looked at some of the things we need to think about when visualizing data: appropriate graphs for the data being displayed as well as the audience we're communicating with, use of colour and labels to ensure accessiblity, things that can skew how the data might interpreted and the like.
- Let's situate data visualizations in a broader context of research data management, and some of the things we should be aware of as data make their way from collection to visuals for communication, and some of the concerns or expectations that surround how these processes are documented.
- This in turn has implications both for the tools used to build visualizations as well as for how one consumes these data and the stories they're purporting to tell.