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

- Overview

Slide 5

- Open question

Slide 6

- Data -> visual information

Slide 7

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

Slide 8

- 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 9

- Open question

Slide 10

- 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 11

- 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 12

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

Slide 13

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

Slide 14

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

Slide 15

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

Slide 16

- 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 17

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

Slide 18

- 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 19-20

- 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 21

- 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 22

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

Slide 23

- 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 24

- 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 25

- We'll start with a data set. This is from the Labour Force Survey, a survey run through Statistics Canada
- What are the categorical variables?
- Are they nominal or ordinal?

Slide 26

- We'll demo with the labour force status variable

Slide 27

- When working with counts, we might commonyl start with something that looks like a pivot table or frequency table

Slide 28

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

Slide 29

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

Slide 30

- We'll look at education, an ordinal variable

Slide 31

- A frequency table

Slide 32

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

Slide 33

- Sorted by education is probably more appropriate

Slide 34

- A few things to consider when making bar graphs
- Categories are discrete, the bars should be discreet. A Histogram is something quite different.

Slide 35

- 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 36

- 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 37

- 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 38

- 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 39

- For this example, we'll use hourly wages - discrete or continuous

Slide 40

- The data first as a rounded set of numbers

Slide 41

- Data tables are one way of providing access to this data

Slide 42

- We can also plot the individual points associated with that table

Slide 43

- A histogram is more commonly used; instead of showing counts, it shows the full distribution of the data

Slide 44

- Histogram

Slide 45

- Bucketing is a choice and impacts the distribution that we see

Slide 46

- Unemployment by province - not normalized

Slide 47

- Using colour to be redundant
- Issue here - too many categories to support colour blind colour options

Slide 48

- Subset of non-maratime provinces

Slide 49

- Should always start at 0

Slide 50

- 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 51

- 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 52

- 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 53

- 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 54

- 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 55

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

Slide 56

- 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.
- We have some weather data here for Kelowna in 2020.

Slide 57

- We should consider how apsect ration impacts how we interpret peaks and valleys in the data.
- We can make this a bit more extreme

Slide 58

- We can make this a bit more extreme

Slide 59

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

Slide 60

- When reading 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 61

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

Slide 62

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

Slide 63

- 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 64

- Back to the drawing board.

Slide 65

- 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 66

- Colour coded by continent

Slide 67

- Size by population

Slide 68

- Increase clarity with some opacity

Slide 69

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

Slide 70

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

Slide 71

- 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 72

- 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 73

- Next we hit the pie chart.

Slide 74

- Revisting education

Slide 75

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

Slide 76

- Colour is important.