Slide 1

- Title

Slide 2

- Florence Nightingale coxcomb.
- News story in the Guardian if you want to read a bit more about her.

Slide 3

- A coxcomb or pie chart as simply a rolled up bar chart and the considerations choice of representation has on how we read the data in the visualization.

Slides 4 & 5

- Overview

Slide 6

- What we're not talking about

Slides 7 - 10

- What is data visualization

Slides 11 - 13

- Explore data, look for points of further inquiry, make inferences
- Tell a story
- Multiple ways of conveying information to diverse audiences.

Slide 14 & 15

- Example of preattentive attributes.

Slides 16 - 20

- Examples of ways to leverage preattentive cognition, using
  - form
  - colour
  - position

Slide 21

- Title

Slide 22

- Overview of common graphs
  - Bar charts
  - Histograms
  - Line and dot plots

    - Scatter plots
    - Pie charts

Slide 23

- Overview of data types
    - data types will determine choice of graph and considerations for how the data is communicated

Slides 24 - 26

- Categorical data may be either ordinal or nominal.
- Ordinal data has some sort of intrinsic order, like something be more or less than or better or worse then.
- Nominal data has no order and includes characteristics like eye colour and countries.
- Categorical data is often presented as the number of individuals or observations in a given category - the number of people respectively with blue eyes, hazel eyes, etc - and are plotted in bar plots.

Slides 27 - 28

- Labour Force Survery intro
- More detail about the survey here

Slide 29

- Example plotting of educational levels as reported in the 2020 Labour Force Survey data collected by Statistics Canada.
- Consider the category order - education levels are ordinal, so it makes sense to plot according to education level as opposed highest number of people in a category to lowest number.
- Visual interpretation is measured by area of bar plots, so be cognizant of starting at zero and maintaining an even width of

the bars, so the only variable changing is hieght.
- Provide a total count (n)

Slide 30

- Consider ease of accessing the labels when we choose between vertical or horizontal representations of bar charts.

Slide 31

- Bar chart design considerations

Slide 32

- Variations on bar charts. What is easier to read in this situation? Stacked or side by side?

Slides 33 - 35

- Numeric data may be either discrete or continuous, and either interval or ratio.
- Discrete data can be counted, like number of people, number of countries etc. Continuous data is measured, like temperature, time or speed.
- Interval data can say that something is greater or less than something else, but does not have a base reference of 0, so we can't say it is *x* percent greater or less than something else. Temperature in degrees Celsius is interval data. We can below 0, and we cannot sat that 20 degrees is twice as warm as 10 degrees, simply that it is 10 degrees warmer.
- Temperature in degrees Celsius is an example of continuous, interval data.
- Ratio data has a base line of 0 and we can say that one measure is *x* percent greater than another. Height starts at 0, so we can say that someone who is 180cm tall is 100% taller than someone who is 90cm tall.
- Height is an example of discrete ratio data.

- Speed is an example of continuous ratio data - it is measured, starts at 0 and we can say that 20 Kph is twice as fast as 10 Kph.
- Numeric data is often counted in binned groups in histograms

Slide 36

- A plot of reported hourly wages from a subset of the 2020 results from the Labour Force Survey

Slide 37

- Choice of binning options impacts how the data is interpreted - smaller bins results in a more detailed representation of the data than larger binning options, while larger bins are arguably more easily visually interpreted.
- Selection of bin size will be dependent on the distribution of the data, the audience, and the message being conveyed.

Slide 38

- Histogram design considerations

Slides 39 - 40

- Per capita GDP in Canada

Slide 41

- Subset of data from slide 27
- Implications of not starting at a base of 0 when working with bar charts and histograms - differences in areas are skewed.

Slide 42

- Title slide

Slides 43 - 44

- Life expectancy data
- Dots are used to plot numeric data is not being counted, but is instead plotting two variables against each other, in this case, country against life expectancy in 2016.
- Because we are not counting data, using bars doesn't make sense.

Slide 45

- When working with country data, we're working with nominal categorical data - there's no intrinsic order here, so this requires some thought in how we choose to plot this data to make it accessible and meaningful.
- Same data set with three different orderings of the categorical data.

Slides 46 - 47

- Plotting a line through the dots often makes sense when we're trying to show trend data, that is data that is continuous, such as data points over time.
- Before we were comparing data points (life expectancy) across categorical groups (countries), so a trend line seems less appropriate than dots.
- Here, we have data within a given country over time, so connecting the points with a trend line helps to convey that shift over time.
- While a line graph helps to highlight continuity, it also hides the individual data points, obscuring the actual data from the reader.

- Depending on the situation, one or the other of these displays may be more appropriate.
- We can certainly combine these two, enriching what it is that we are able to say, but also adding more information for the reader to interpret.

Slides 48 - 51

- Time sequences, weather data
- Line graphs make sense for other continuous data such as weather.
- Since we are not measuring quantities and thus not using bins to represent our data, it is not important that we start at 0, as area is not our visual cue here.
- However, aspect ratio is a key consideration - the ratio between the x and y axes. As the aspect ratio approaches 1:1 peaks and valleys are accentuated, which can misrepresent the actual relative distances between data points.
- An even smaller aspect ration.
- The impact of both colour and aspect ratio may be used to further influence how the data is consumed by the reader. Here colour and aspect ratio may make the shifts in temperature patterns seem more or less alarming.

Slide 52

- Title

Slides 53 - 55

- Scatter plots are great for large data sets when we want to be able to investigate potential correlations or show correlation.
- In these instances, we plot two variables, the y axis against the x axis looking for a a positive, negative, or no association.
- Scatter plots do not confirm associations, relationships, or cause and effect. This is a fun site to see how trends might be connected without there being any kind of relationship: Spurious correlations

Slides 56 - 60

- In the previous slide we had one layer of information, life expectancy against per capita GDP. We can build in layers to get more insights, such as colour coding countries by their continent (keeping in mind that continent divisions might be different to a politician than they are to a geologist or a geographer), and size coding countries according to their population counts.

Slides 61 - 62

- These two plots are plotting exactly the same data. On the plot on the left, there is an outlier. Representations with outliers are often manipulated mathematically to streamline the visualization, and make it linear.
- In this instance, the plot on the right has taken the base 10 log of all the GDP data; it has also skewed–or rescaled–the x scale.

Slides 63 - 64

- Adding line of fit and breaking on continent

Slides 65 - 66

- Pie charts are great for data that has a summative whole.
- They are generally very accessible as they are very common place and so people are used to seeing and interacting with them.
- Compared to bar charts (this is the same data as in our earlier slide of education levels), they can make interpreting the actual area harder, simply due to the shape. This is especially true if we plot two pie charts side by side vs two bar charts.

Slide 67

- Finding data is both easy and hard
- Easy
    - Movement toward open data
    - Lots of data collection happening
    - Ready access through the web and repositories
- Hard
    - Not well curated as a whole like literature, so discovery is more problematic
    - Issues of sensitivity (personal information, intellectual property, locations of red listed species), make sharing problematic
    - While we generate lots of data, it's not all stored or shared, so questions that should seem easy to access with data can in fact be quite difficult.

Slide 68

- CBC fatal police encounters data set as an example.
- In looking at systemic racism, journalists were interested in the characteristics of fatal police encounters.
- While this data is certainly generated, it is not collated, reported on, or shared across policing jurisdictions in Canada.
- The data had to be collected through what was reported on in the news.
- We often think of data as raw points of information, but we should be conscious of the paths taken to created that data. In the above example, it is a subset of the true data, filtered by the media and how the media reports on these encounters. In the case of survey data, this will be impacted by just how well a question captures the information it's trying to capture.

Slide 69

- Guiding questions to help you find data and know where to target your inquiries, from the closest point of creation to the furthest.

Slide 70

- There are also many data aggregators and repositories. Governmental bodies and similar organizations create and collate lost of interesting economic and vital statistics data.
- Each level of government will generate and collate different kinds of questions - we must ask ourselves about the roles or functions each of these levels play.
- The same goes for other organizations.

Slide 71

- UBC data guide with more resources

Slide 72

- UBC purchased data sets

Slide 73

- In Canada, Statistics Canada is a great source of data, both raw and synthesized.
- Stats Can data is available in three broad categories:
  - Table data made available through an online portal - this will be state and market generated data or may be aggregated, population level census data or other surveys that stats can runs. This data is usually presented at a fairly high level. And it's freely available. https://www150.statcan.gc.ca/n1/en/type/data?HPA=1
  - Public Use Microdata Files - these are non aggregated data sets that have been anonymized. This is paid product from Stats Canada, but we have access to this with UBC. Micro data information https://www.statcan.gc.ca/eng/help /microdata Abacus access https://resources.library.ubc.ca/page.php?details=abacus-data-repository&id=1114
  - Raw Census data. These are highly restricted. In fact, we're just in the process of getting a data centre at UBCO for access to these data. The servers holding this data are provided by vetted researcher access only, and basically, you

can bring nothing superfluous with you into the server rooms and you can leave with nothing that didn’t come in with.