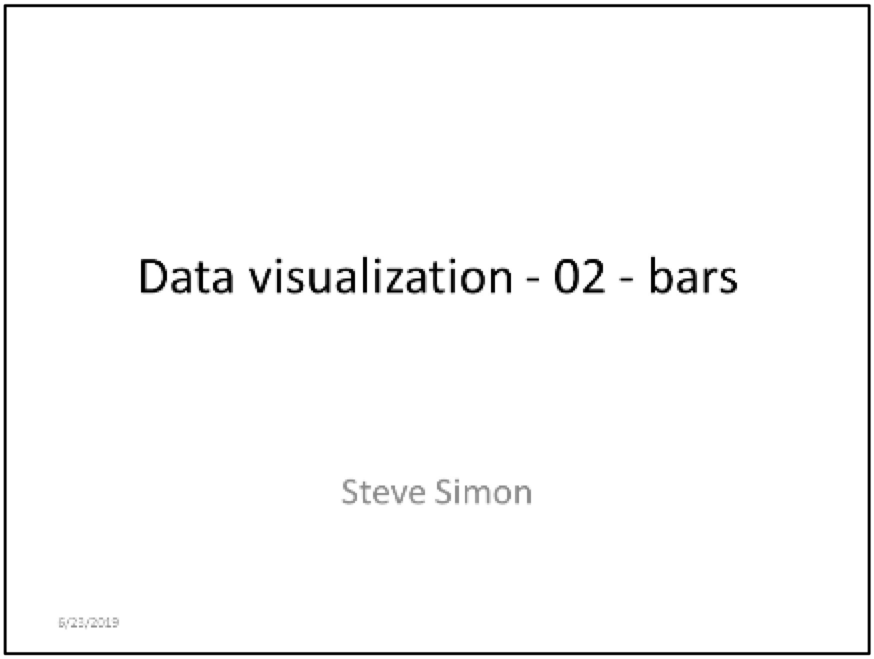
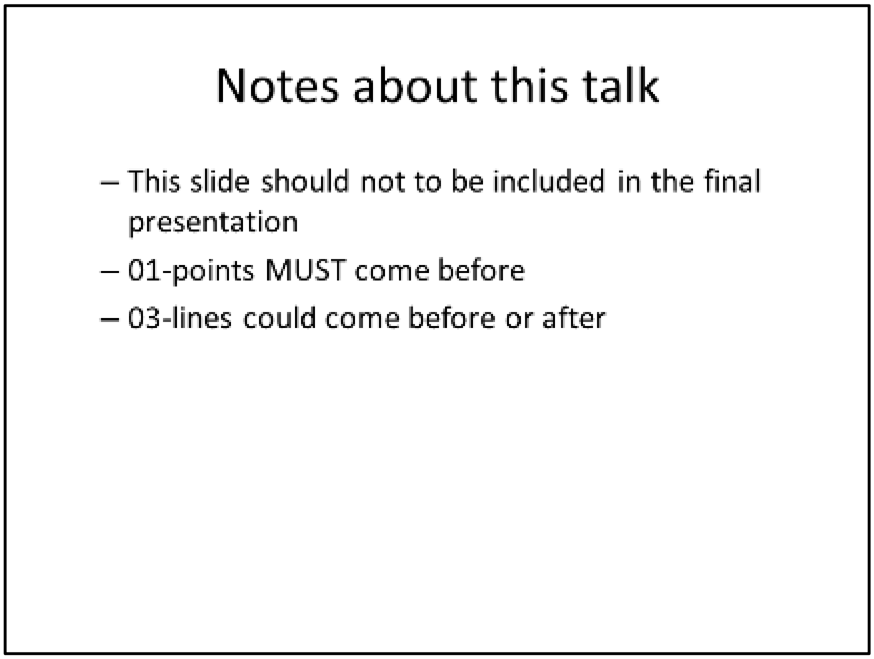


Data visualization - 02 - bars

Steve Simon

6/23/2019



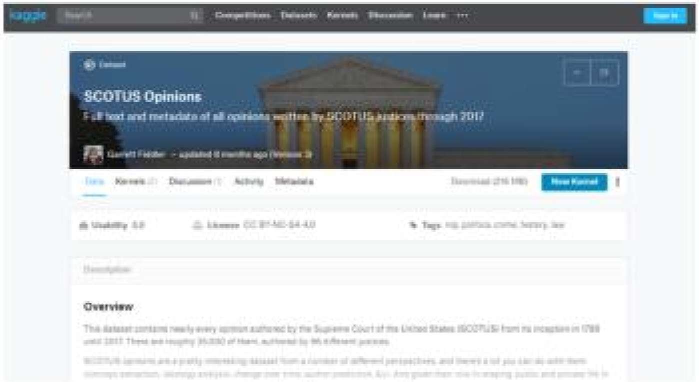
Notes about this talk

– This slide should not to be included in the final presentation

– 01-points MUST come before

– 03-lines could come before or after

To prepare for this section



SCOTUS Opinions

Full text and metadata of all opinions written by SCOTUS justices through 2017

Garrett Fiddler • updated 8 months ago (Version 3)

Usability 5.9

License CC BY-NC-SA 4.0

Tags nlp, politics, crime, history, law

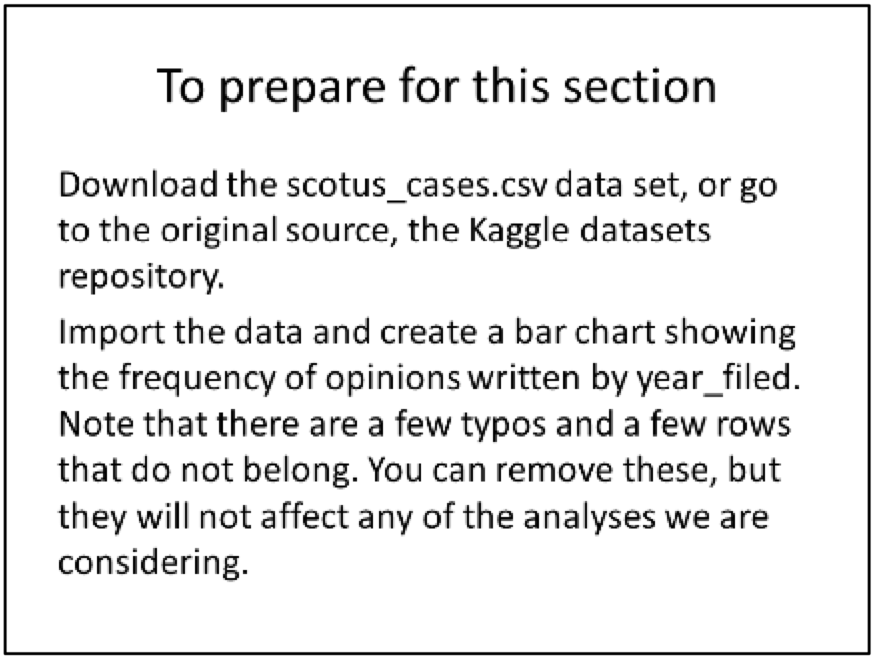
Description

Overview

This dataset contains nearly every opinion authored by the Supreme Court of the United States (SCOTUS) from its inception in 1789 until 2017. There are roughly 35,000 of them, authored by 95 different justices.

SCOTUS opinions are a pretty interesting dataset from a number of different perspectives, and there's a lot you can do with them: summary extraction, ideology analysis, change over time, author prediction, &c. And given their role in shaping public and private life in

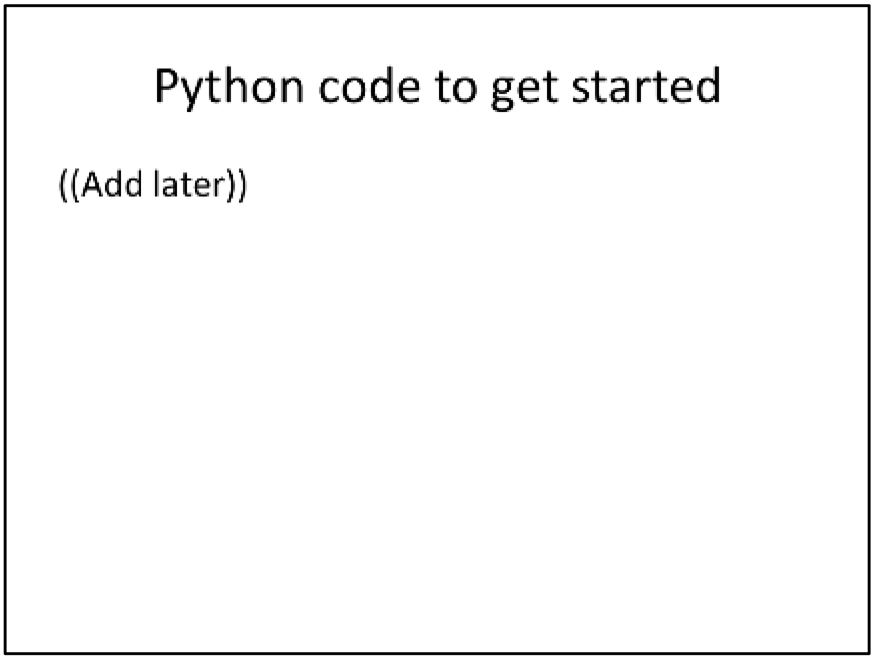
Kaggle datasets webpage



To prepare for this section

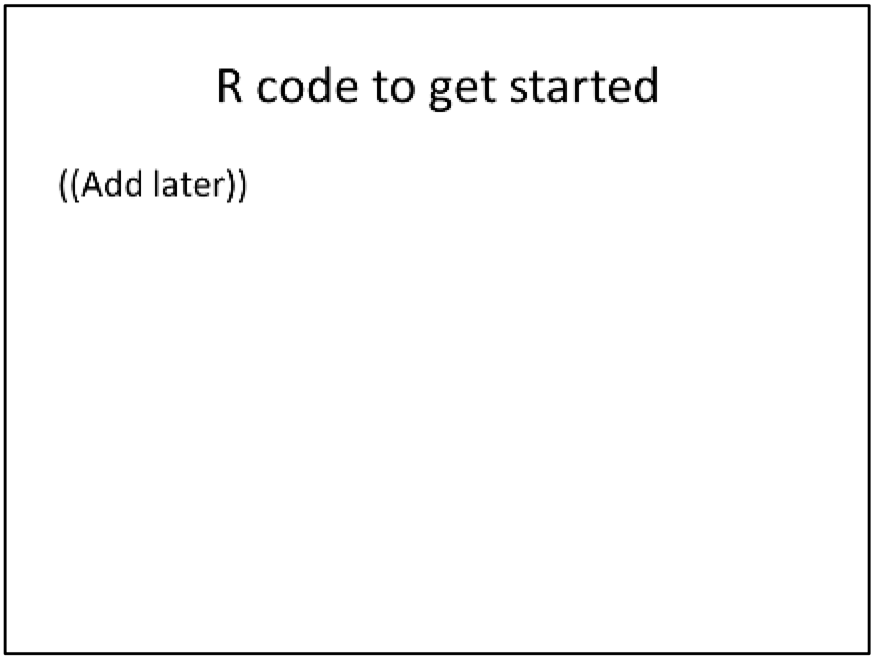
Download the scotus_cases.csv data set, or go to the original source, the Kaggle datasets repository.

Import the data and create a bar chart showing the frequency of opinions written by year_filed. Note that there are a few typos and a few rows that do not belong. You can remove these, but they will not affect any of the analyses we are considering.



Python code to get started

((Add later))



R code to get started

((Add later))

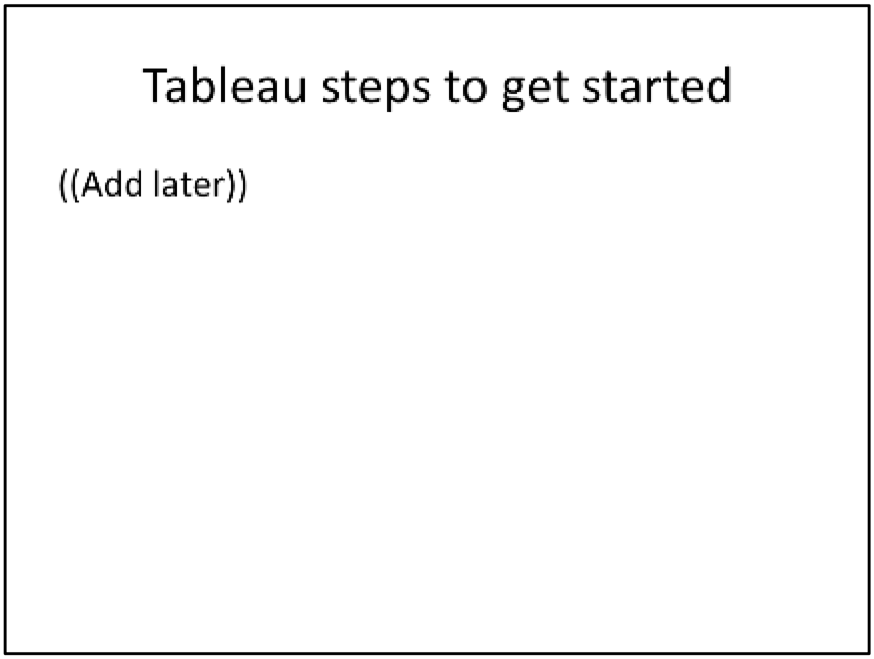
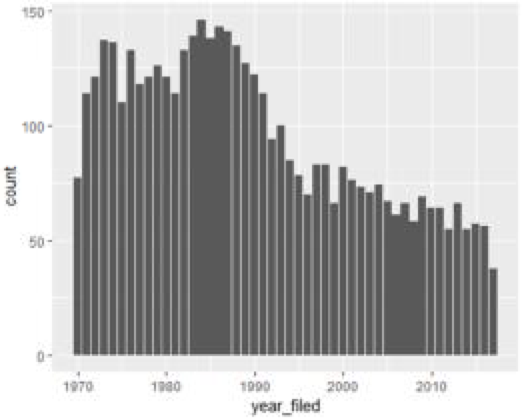
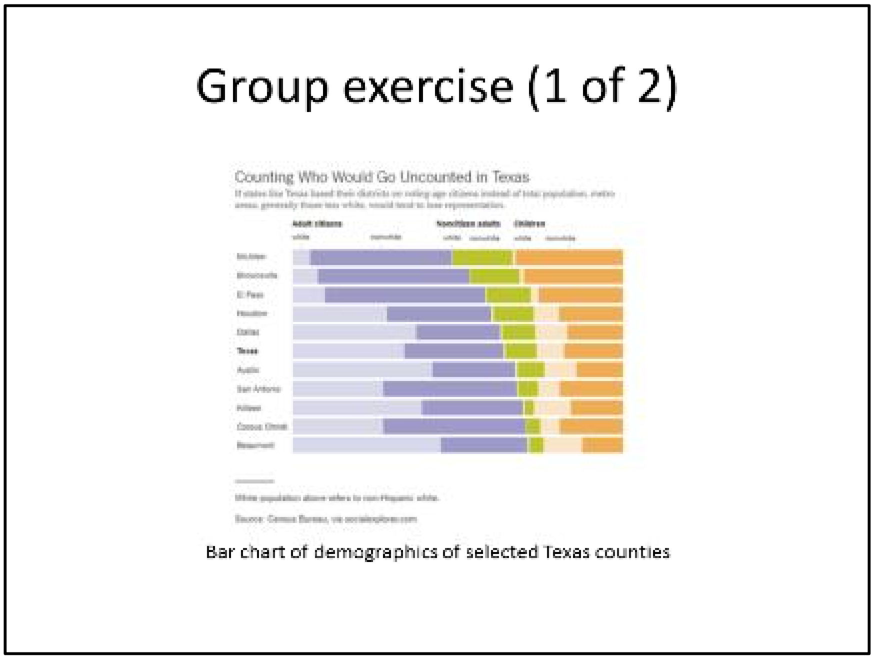


Tableau steps to get started

((Add later))

This is what your bar chart should look like





Group exercise (1 of 2)

Counting Who Would Go Uncounted in Texas

If states like Texas based their districts on voting age citizens instead of total population, metro areas, generally those less white, would tend to lose representation.

White population above refers to non-Hispanic white.

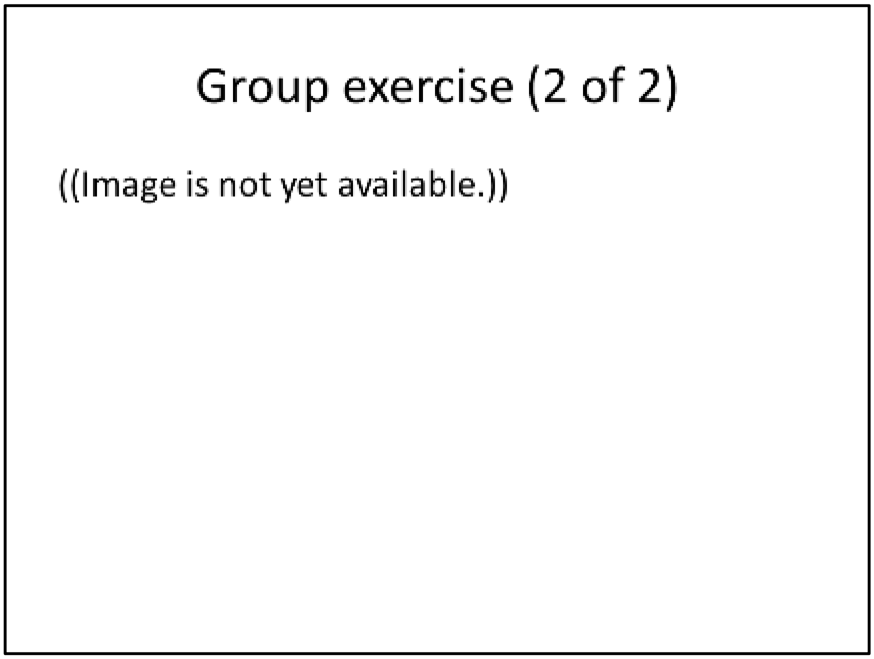
Source: Census Bureau, via socialexplorer.com

Bar chart of demographics of selected Texas counties

This is one of two graphs. It was published in

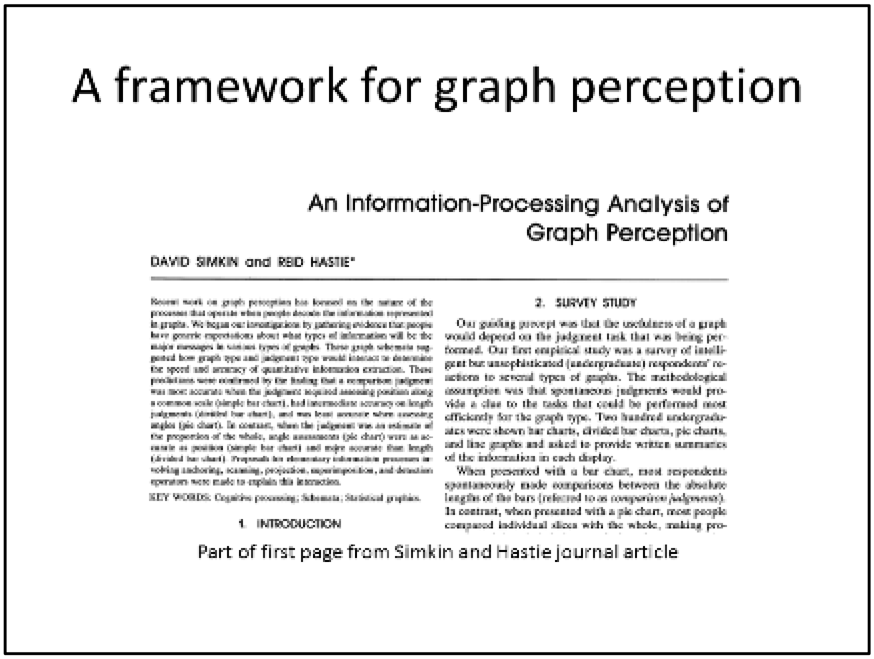
Badger, E. (2019). People Who Can't Vote Still Count Politically in America. What if That Changes?. Retrieved June 24, 2019, from The New York Times website: https://www.nytimes.com/2019/06/22/upshot/america-who-deserves-representation.html

Split into pairs. Review the article briefly (about 5 minutes) and look at the graph. Explain to your partner what the graph is trying to show. Your partner will get a different graph and do the same thing with you listening this time.



Group exercise (2 of 2)

((Image is not yet available.))



A framework for graph perception

An Information-Processing Analysis of Graph Perception

DAVID SIMKIN and REID HASTIE*

Recent work on graph perception has focused on the nature of the processes that operate when people decode the information represented in graphs. We began our investigations by gathering evidence that people have generic expectations about what types of information will be the major messages in various types of graphs. These graph schemata suggested how graph type and judgment type would interact to determine the speed and accuracy of quantitative information extraction. These predictions were confirmed by the finding that a comparison judgment was most accurate when the judgment required assessing position along a common scale (simple bar chart), had intermediate accuracy on length judgments (divided bar chart), and was least accurate when assessing angles (pie chart). In contrast, when the judgment was an estimate of the proportion of the whole, angle assessments (pie chart) were as accurate as position (simple bar chart) and more accurate than length (divided bar chart). Proposals for elementary information processes involving anchoring, scanning, projection, superimposition, and detection operators were made to explain this interaction.

KEY WORDS: Cognitive processing; Schemata; Statistical graphics.

1. INTRODUCTION

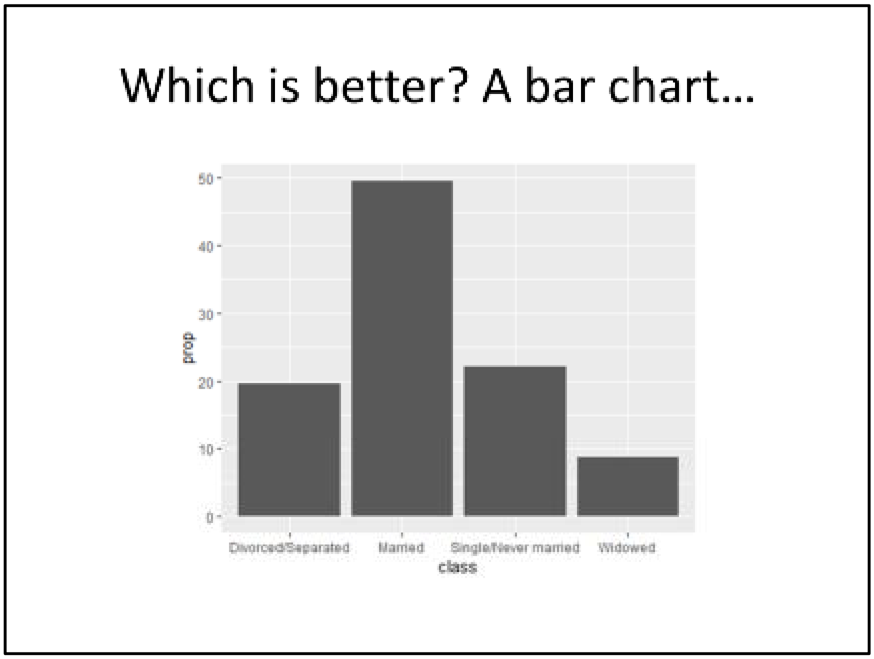
2. SURVEY STUDY

Our guiding precept was that the usefulness of a graph would depend on the judgment task that was being performed. Our first empirical study was a survey of intelligent but unsophisticated (undergraduate) respondents' reactions to several types of graphs. The methodological assumption was that spontaneous judgments would provide a clue to the tasks that could be performed most efficiently for the graph type. Two hundred undergraduates were shown bar charts, divided bar charts, pie charts, and line graphs and asked to provide written summaries of the information in each display.

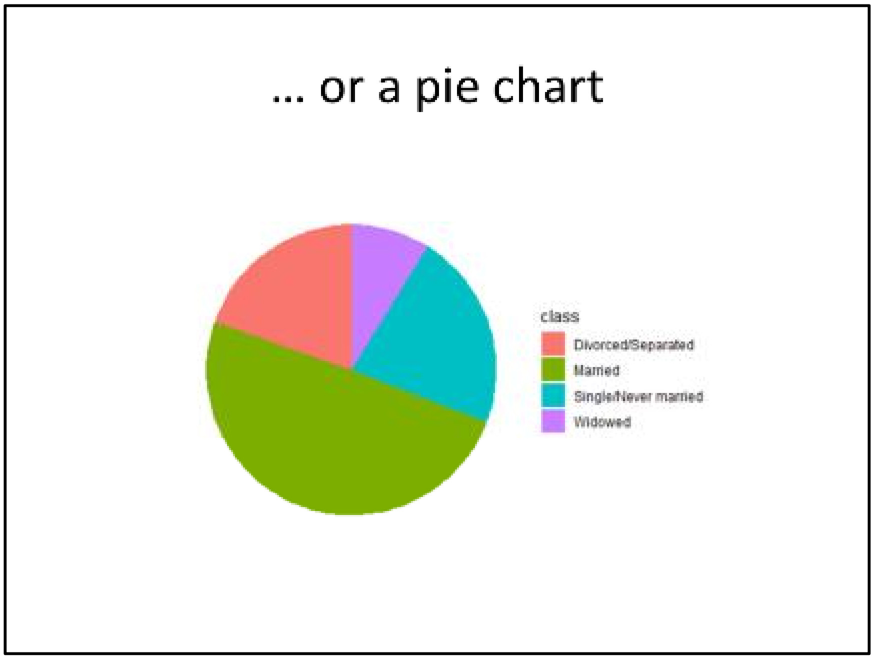
When presented with a bar chart, most respondents spontaneously made comparisons between the absolute lengths of the bars (referred to as *comparison judgments*). In contrast, when presented with a pie chart, most people compared individual slices with the whole, making pro-

Part of first page from Simkin and Hastie journal article

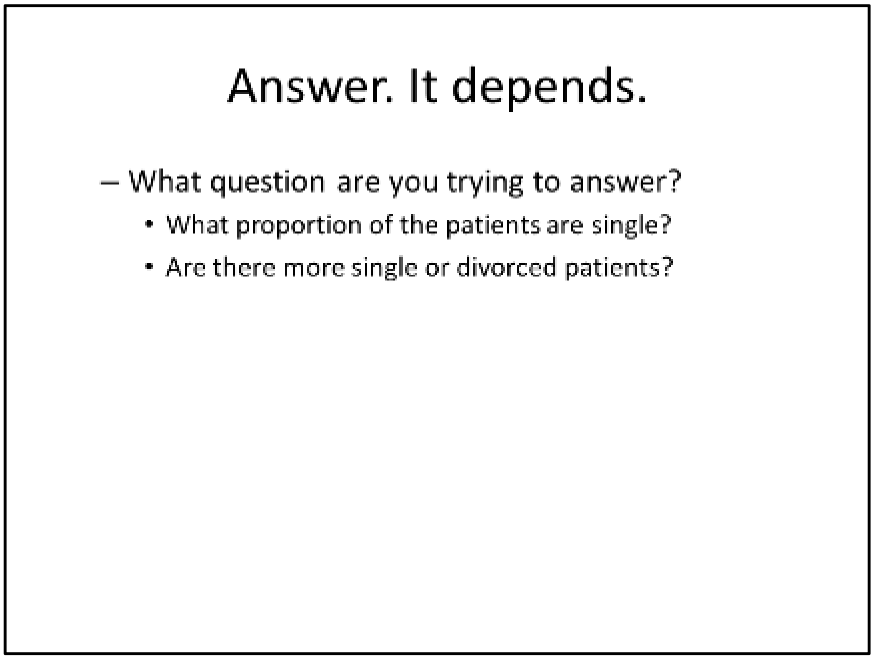
Much of the work on the psychology of perception that I will be discussing next is drawn from this 1987 article by Simkin and Hastie.



Which is better? A bar chart…



… or a pie chart

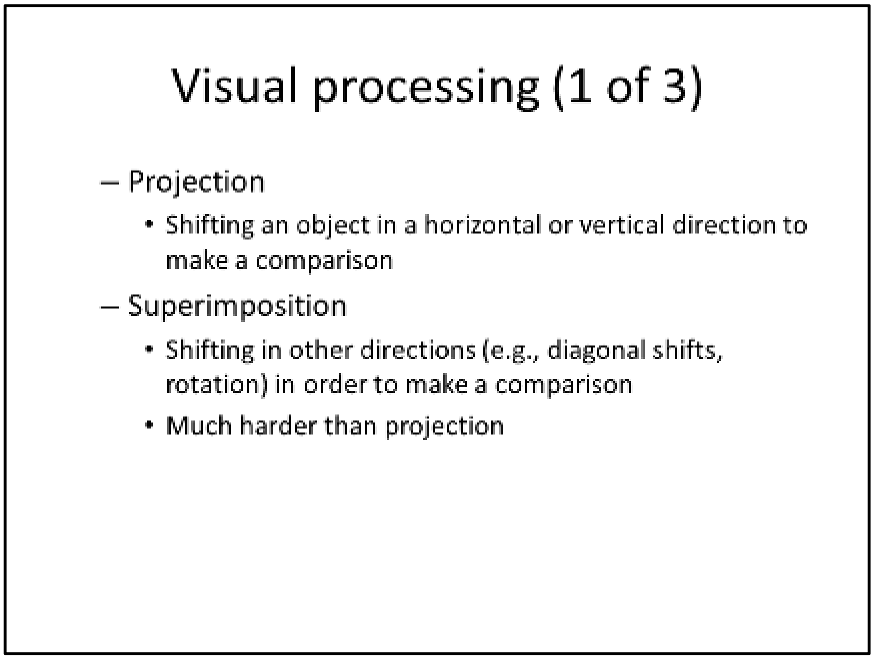


Answer. It depends.

- What question are you trying to answer?
  - What proportion of the patients are single?
  - Are there more single or divorced patients?

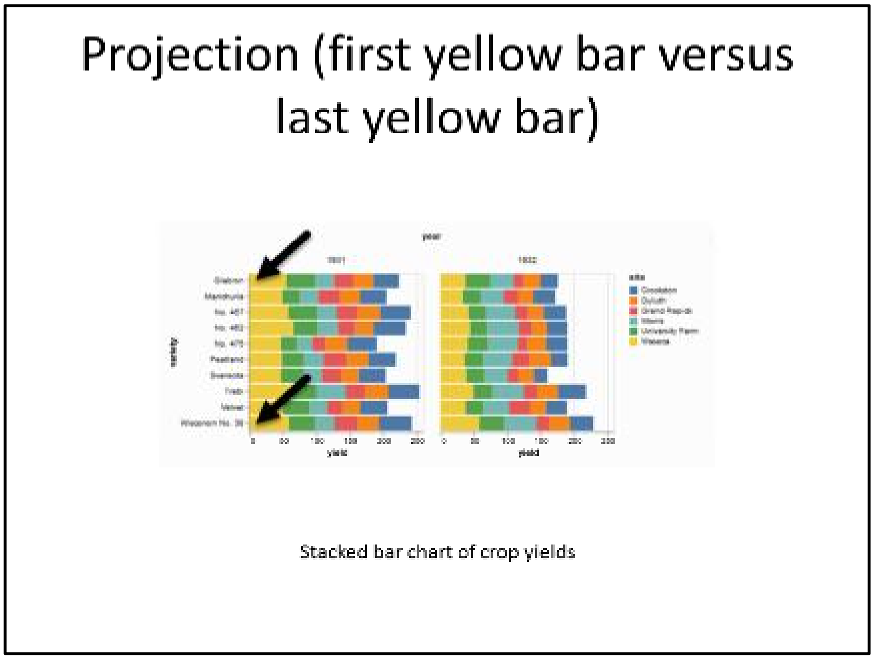
The answer really depends on what question you are asking. There are a variety of questions that you might ask. Two are illustrated above.

You can run an experiment (people have done this) where randomize and show half of them a bar chart and half of them a pie chart. Then you ask a question, like one of the questions above. Then you note the speed and accuracy of the response. Depending on the question, sometimes pie charts give faster and more accurate answers. Sometimes bar charts give faster and more accurate answers. It turns out that the results match up nicely with what we know about the psychology of perception.



Visual processing (1 of 3)

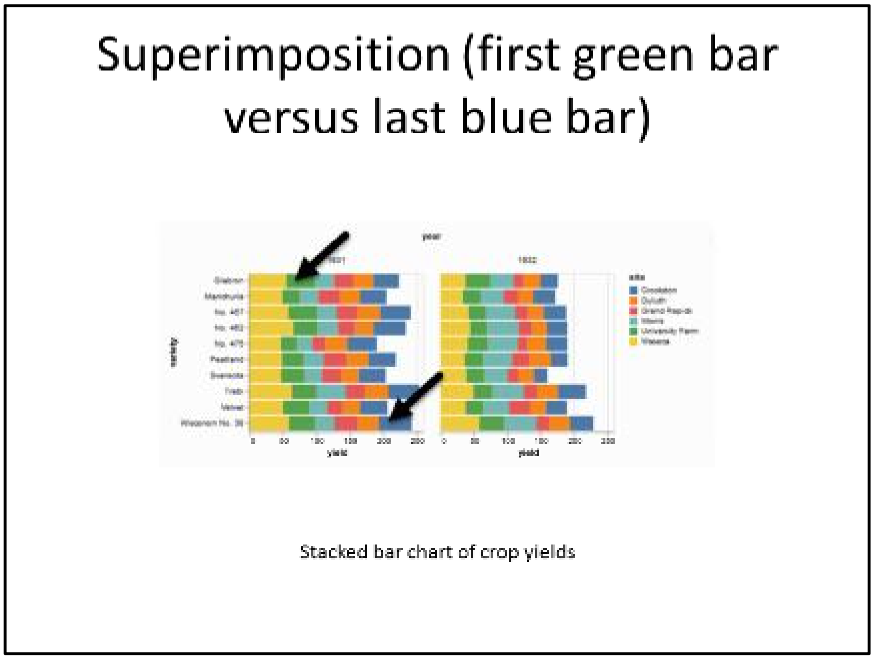
- Projection
  - Shifting an object in a horizontal or vertical direction to make a comparison
- Superimposition
  - Shifting in other directions (e.g., diagonal shifts, rotation) in order to make a comparison
  - Much harder than projection



Projection (first yellow bar versus last yellow bar)

Stacked bar chart of crop yields

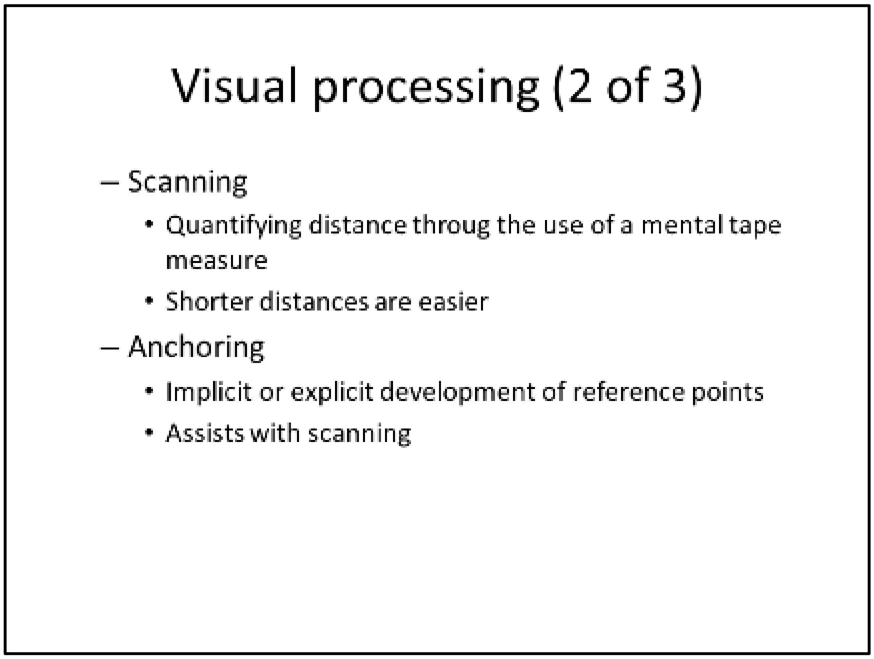
The position means the vertical or horizontal location. Does the first yellow bar in 1931 (Glabron seeds planted in Wasica) extend further to the right than the last yellow bar (Wisconsin No. 38 seeds planted in Wasica)?



Superimposition (first green bar versus last blue bar)

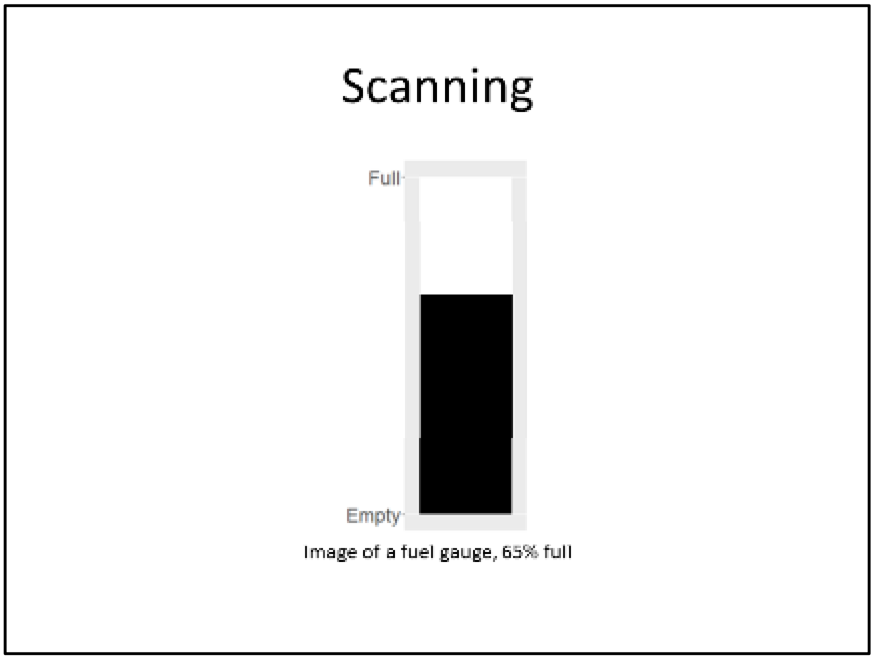
Stacked bar chart of crop yields

The length means either the width or the height. Does the first green bar in 1931 (Glabron seeds planted in University Farm) extend further to the right than the last yellow bar (Wisconsin No. 38 seeds planted in Crookston)?



Visual processing (2 of 3)

- Scanning
  - Quantifying distance throug the use of a mental tape measure
  - Shorter distances are easier
- Anchoring
  - Implicit or explicit development of reference points
  - Assists with scanning



Scanning

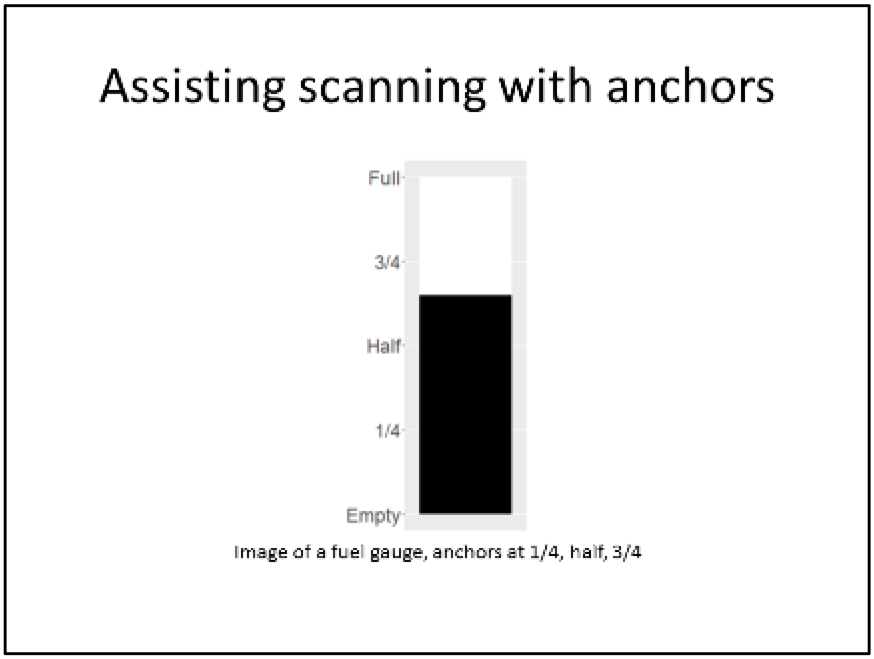
Image of a fuel gauge, 65% full

To understand scanning, think of a gas gauge. Usually it is a semicircular dial, but let’s set up the gas gauge as a rectangle. If the level is at the top, you have a full tank. If the level is at the bottom, you have an empty tank. This gauge shows a tank that is 65% full. Trust me, I drew the gauge. It is at 65%. Now how would you estimate the gas level?

You would take a mental tape measure, starting at the bottom and measure up to where the black box ends.

Now if you were smart, you’d start at the top and scan downwards. Less distance means that you can do this faster and more accurately.

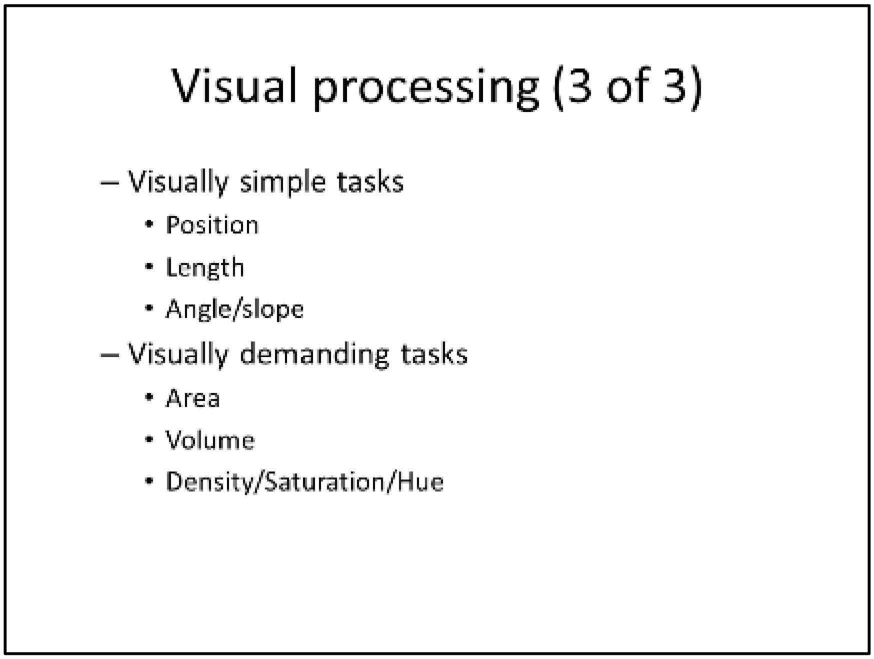
Now if you were Albert Einstein, you’d split the gauge at the halfway point and measure from the halfway point to the top of the black box. Actually, there’s a little of Albert Einstein in all of us. That halfway point is something that all of us do subconciously. You did, because you recognized almost immediately that the tank was more than half full.



Assisting scanning with anchors

Image of a fuel gauge, anchors at 1/4, half, 3/4

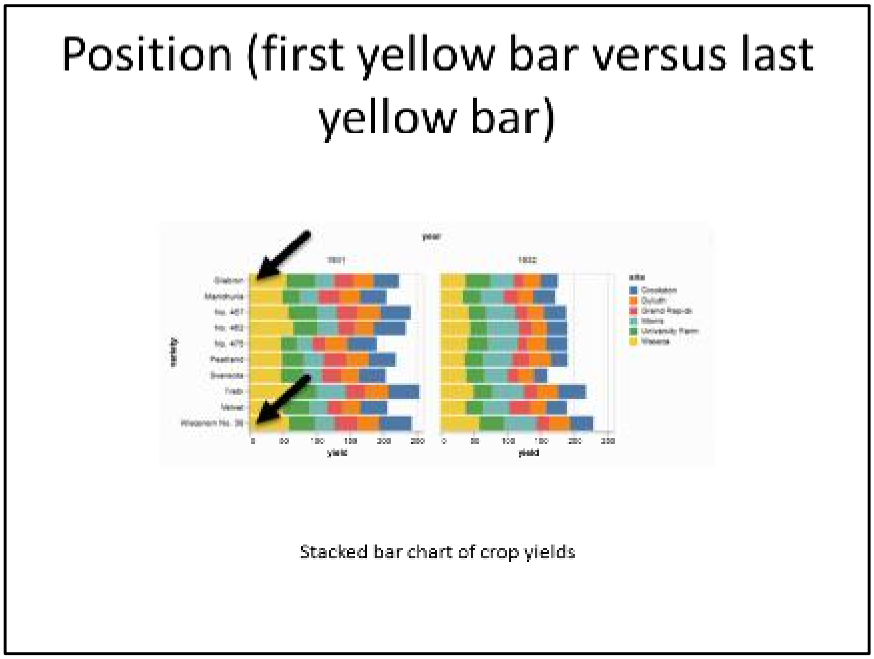
Here’s the same gas gauge, still at 65% full, but now we have added anchors at 1/4, half, and 3/4. You can read this gauge faster and more accurately, because you can scane from half up to 65% or from 3/4 down to 65%.



Visual processing (3 of 3)

- Visually simple tasks
  - Position
  - Length
  - Angle/slope
- Visually demanding tasks
  - Area
  - Volume
  - Density/Saturation/Hue

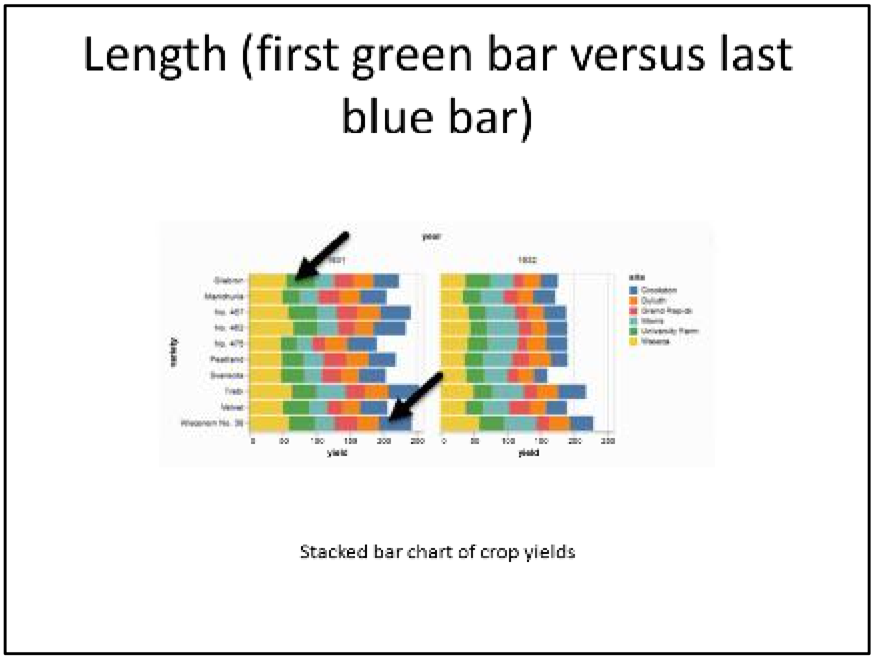
There are a variety of perceptual tasks that you use when making comparisons within an image. These are arranged on this slide roughly in order of difficulty, with the easiest tasks at the top.



Position (first yellow bar versus last yellow bar)

Stacked bar chart of crop yields

The comparison of the two yellow bars is a comparison of position. Which yellow bar extends further to the right?

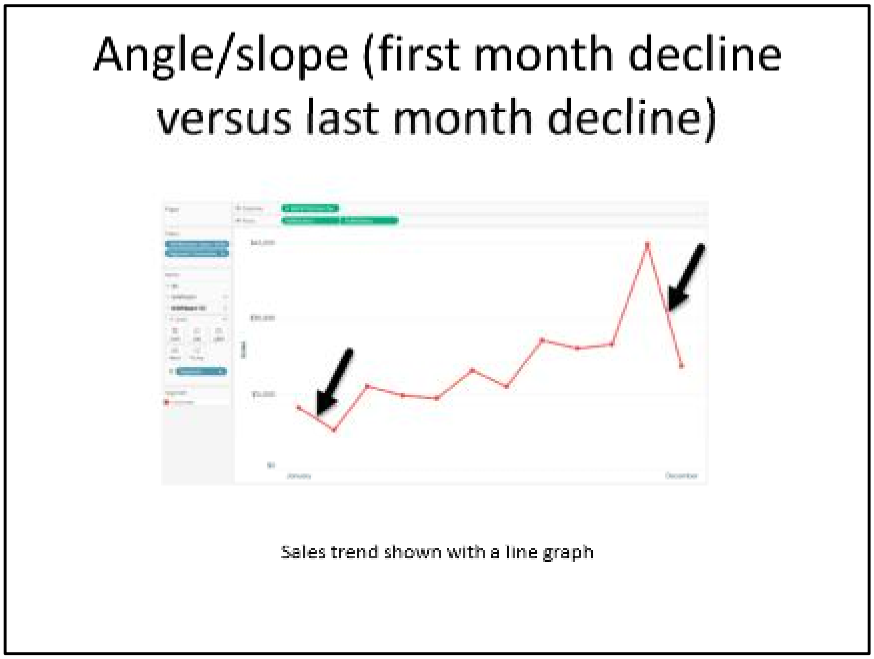


Length (first green bar versus last blue bar)

Stacked bar chart of crop yields

The comparison of the green and blue bars is a length comparison. The two bars start at different spots, so the position can't help you.

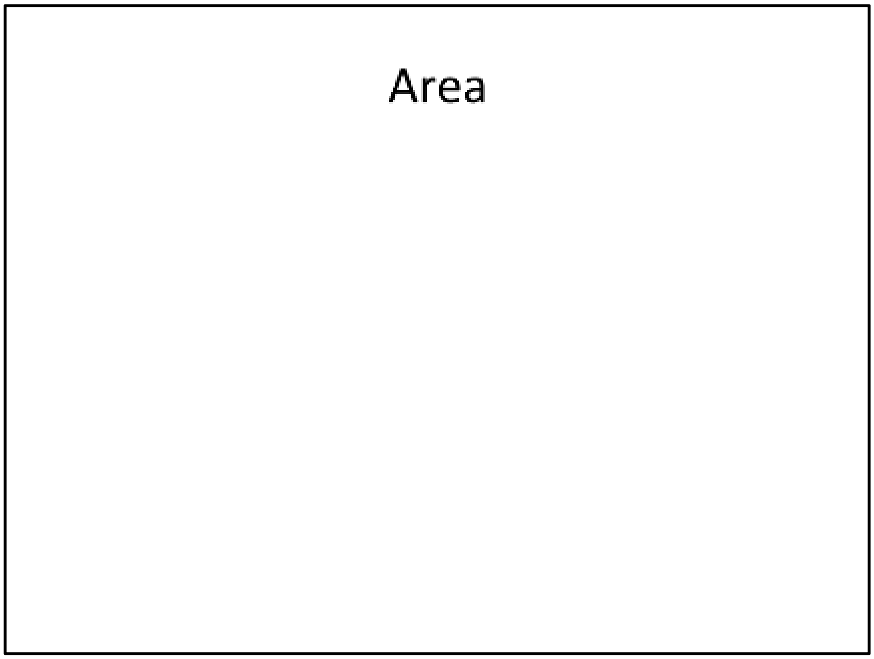
Length is harder to judge than position, because it involves a superimposition rather than a projection.



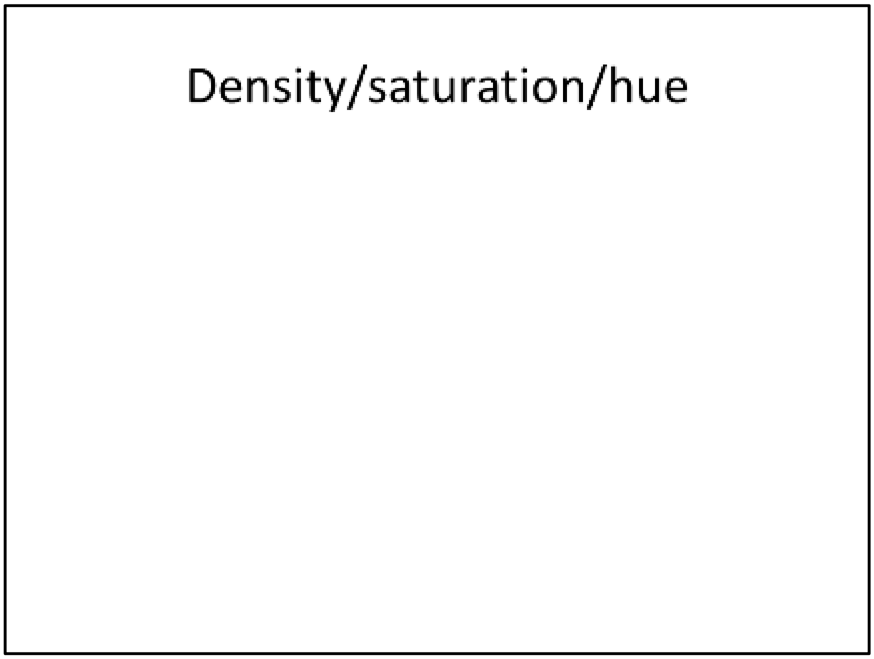
Angle/slope (first month decline versus last month decline)

Sales trend shown with a line graph

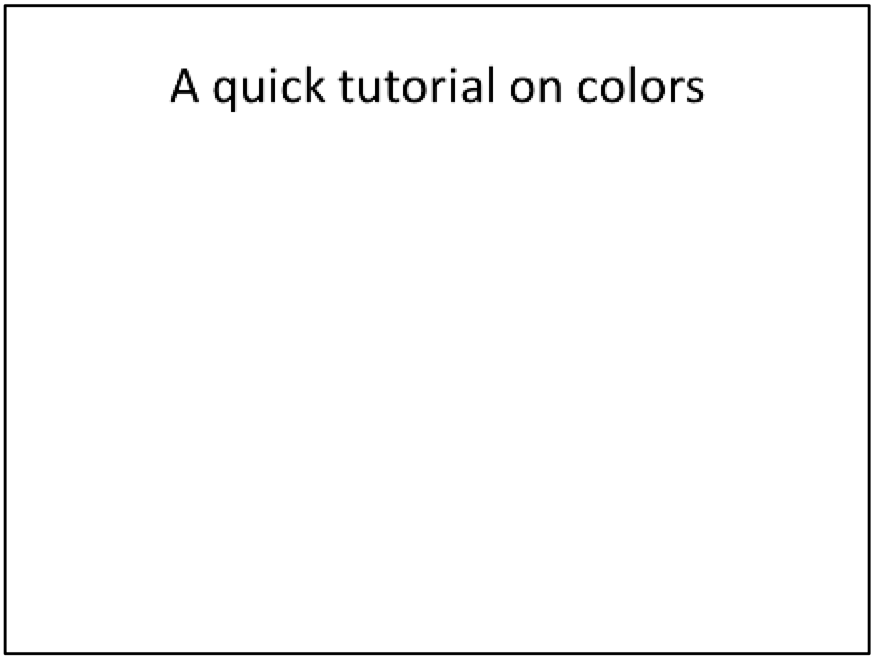
This graph shows sales trends over a twelve month span. If you want to assess whether the first month decline (the dip in sales between January and February) was worse than the last month decline (the dip in sales between November and December), you would probably do this by judging the angle of the first line segment to the angle of the second line segment. This is not quite as easy as a position or length judgement, but it isn't too bad either.



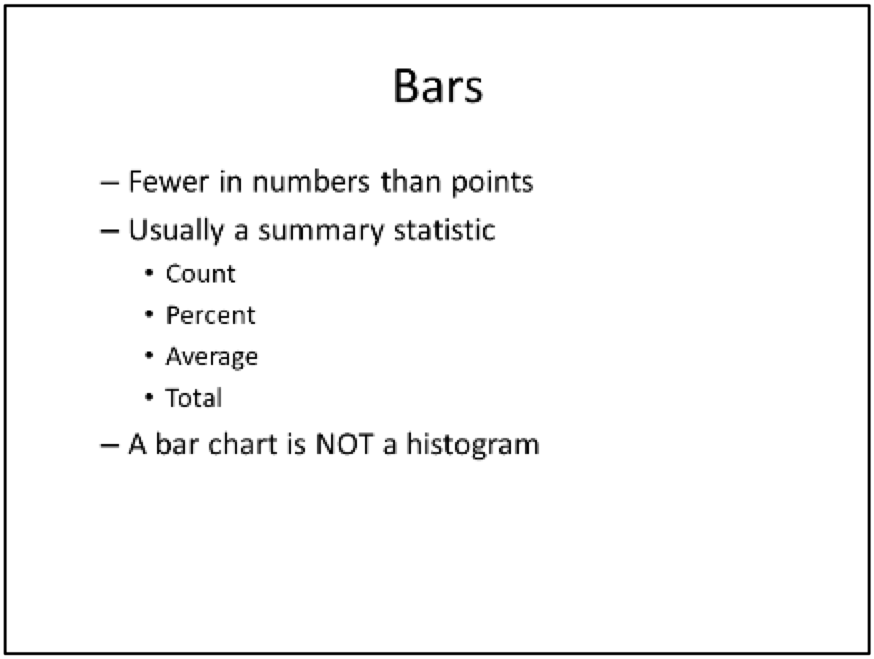
Area



Density/saturation/hue



A quick tutorial on colors

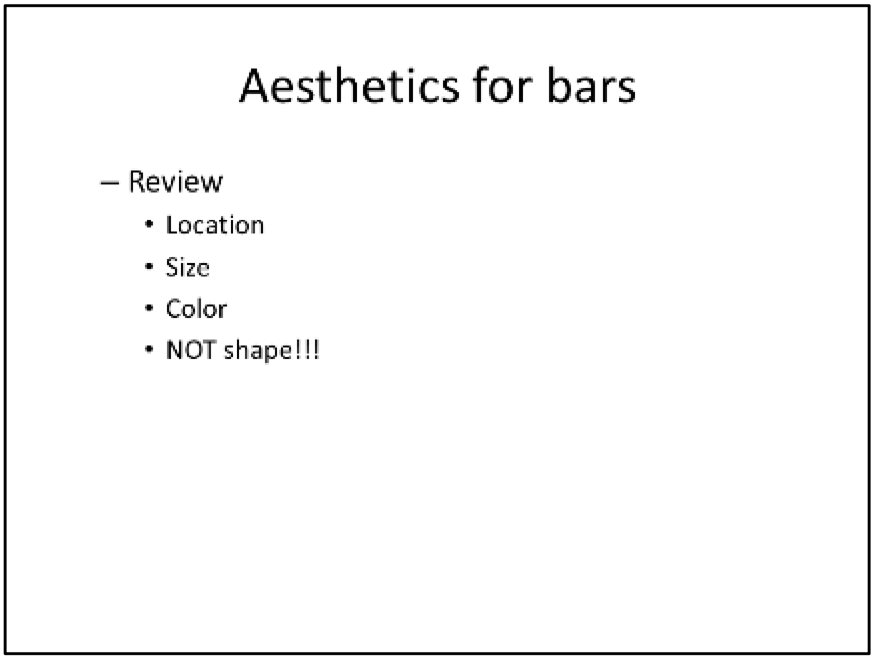


Bars

- Fewer in numbers than points
- Usually a summary statistic
  - Count
  - Percent
  - Average
  - Total
- A bar chart is NOT a histogram

Bar charts are different than point charts. There are usually only a few bars. These bars usually represent a summary statistic, like a count, percent, average, or total.

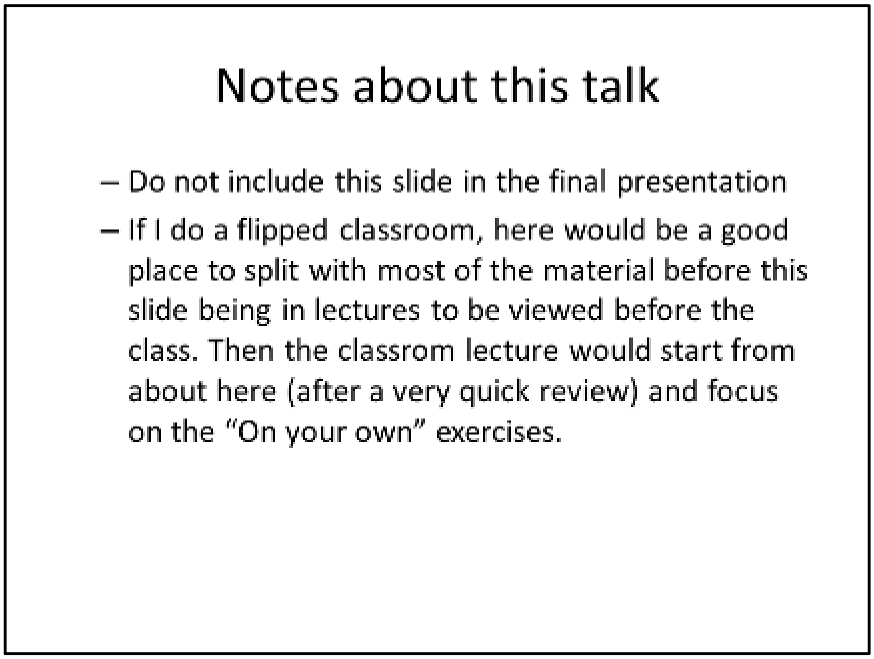
There is a technical distinction between a bar chart and a histogram. Histograms are a great diagnostic tool, but usually ends up on the cutting room floor. So I won’t be talking about it much, if at all, in this workshop.



Aesthetics for bars

- Review
  - Location
  - Size
  - Color
  - NOT shape!!!

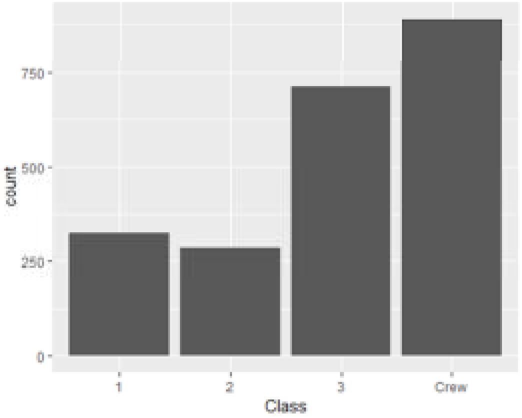
Recall that aesthetics are visual attributes associated with a geometry/mark. You can map variables to the location, size, and/or color of bars, but you cannot assign a variable to the shape of a bar.

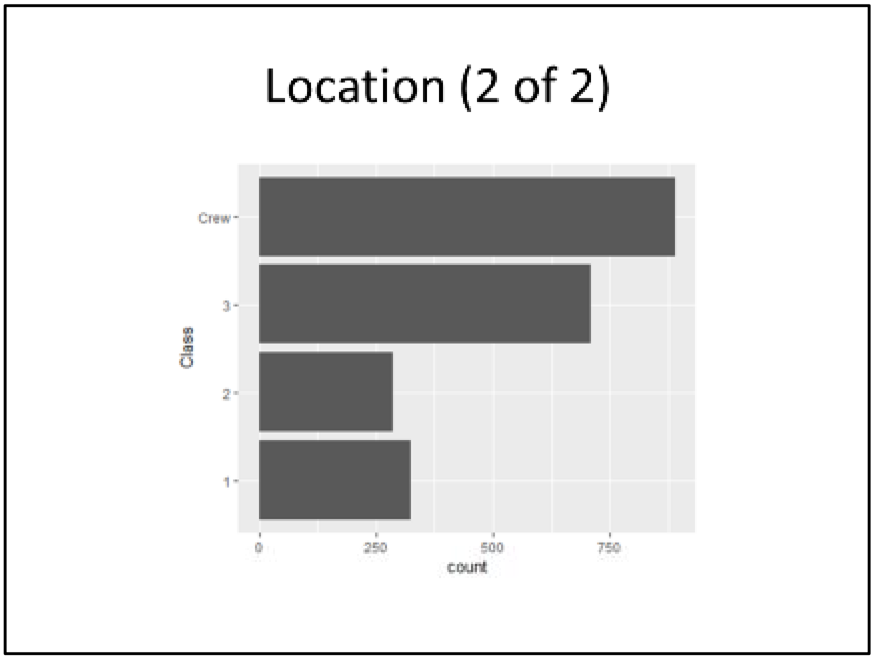


Notes about this talk

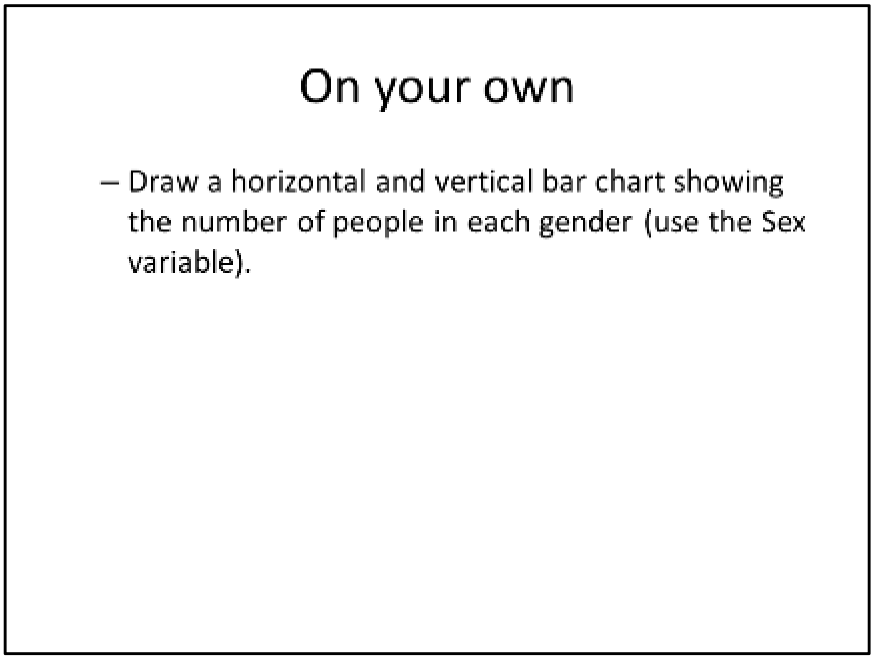
- Do not include this slide in the final presentation
- If I do a flipped classroom, here would be a good place to split with most of the material before this slide being in lectures to be viewed before the class. Then the classrom lecture would start from about here (after a very quick review) and focus on the "On your own" exercises.

Location (1 of 2)



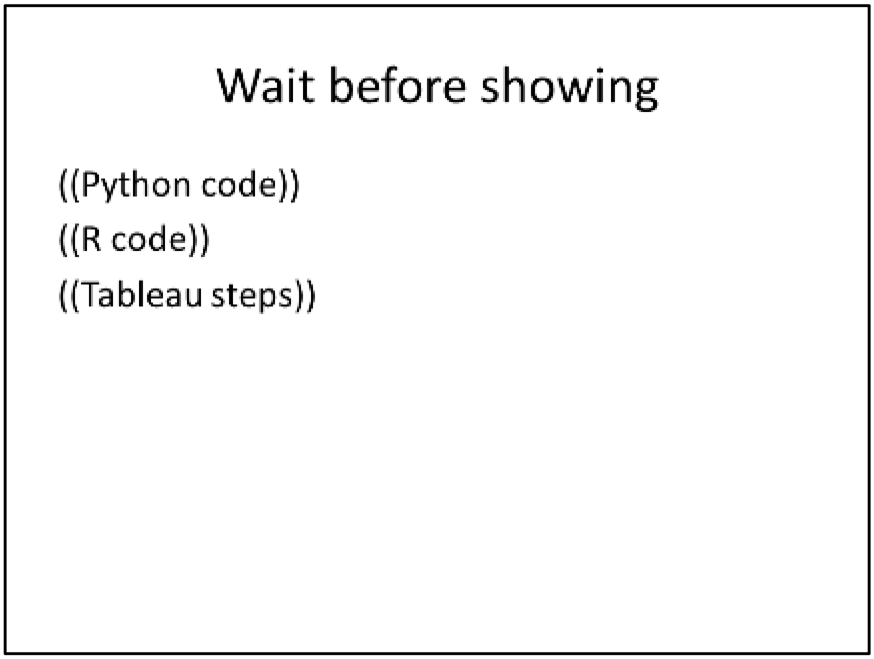


Location (2 of 2)



On your own

- Draw a horizontal and vertical bar chart showing the number of people in each gender (use the Sex variable).

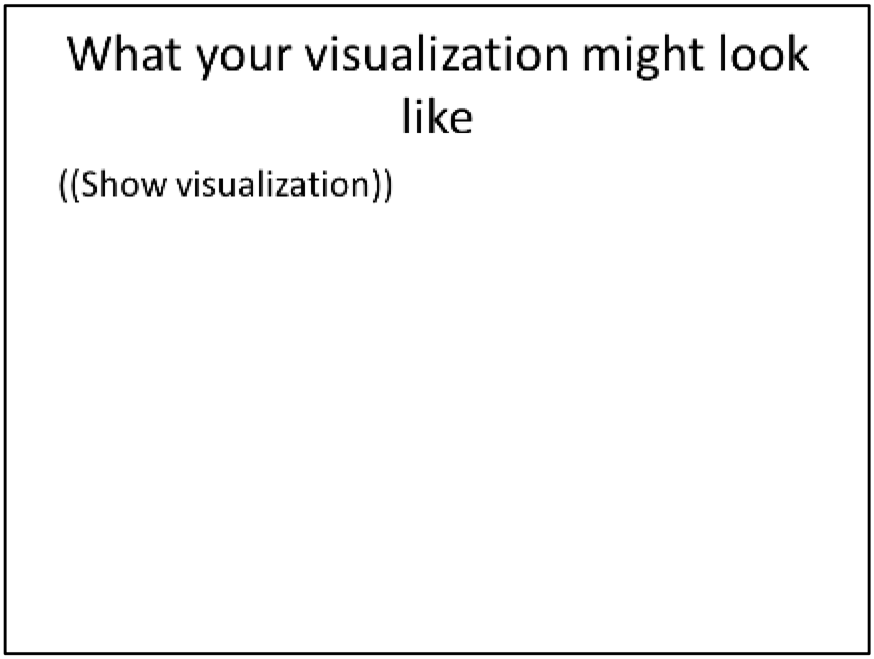


Wait before showing

((Python code))

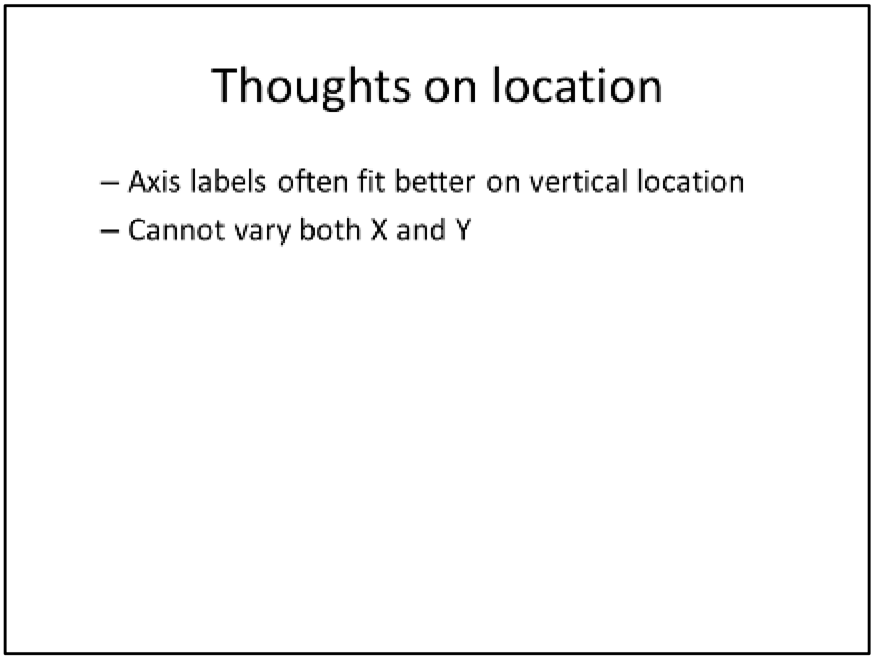
((R code))

((Tableau steps))



What your visualization might look like

((Show visualization))

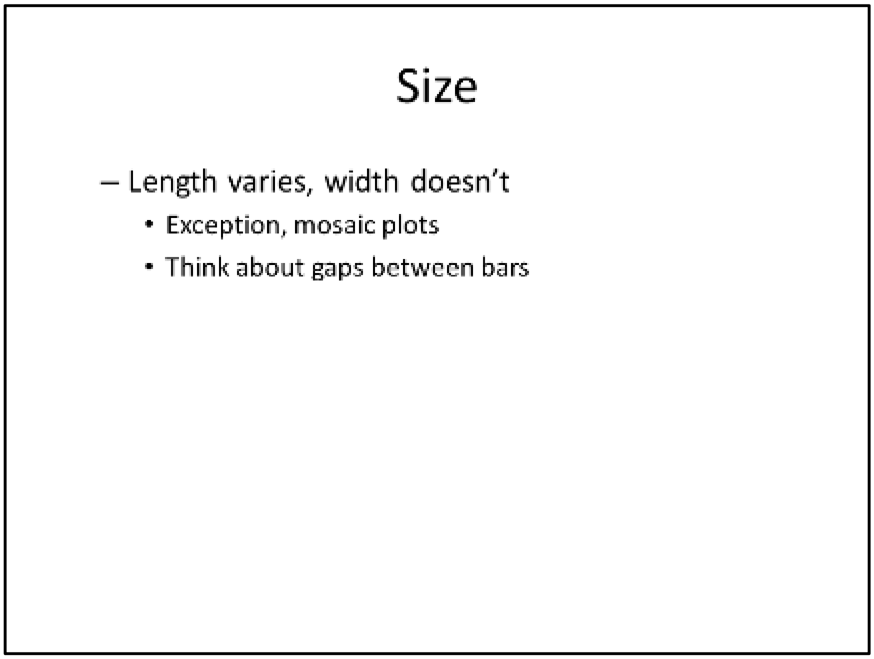


Thoughts on location

- – Axis labels often fit better on vertical location
- – Cannot vary both X and Y

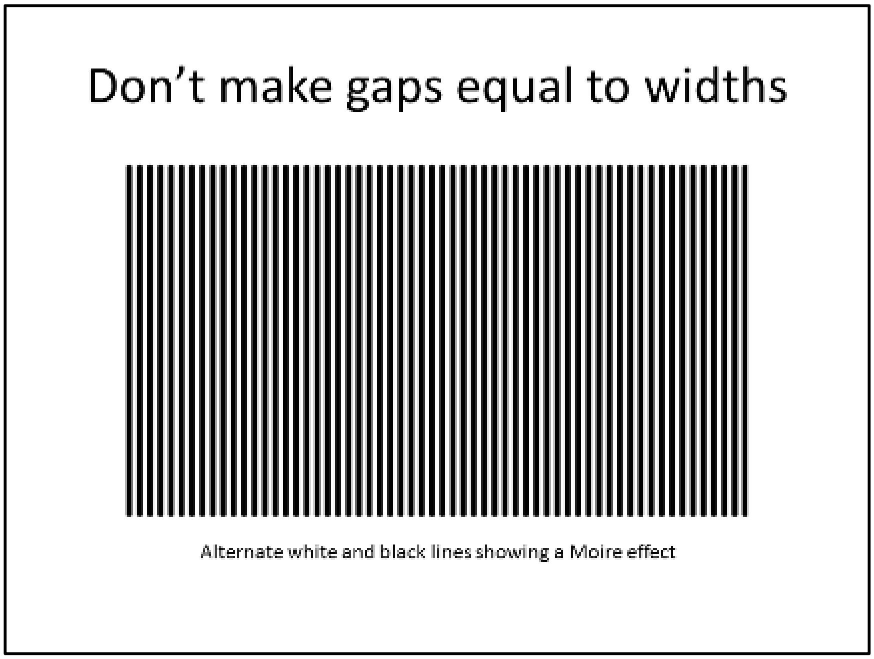
The default for most visualization software is vertical bars, but you should give thoughtful consideration to horizontal bars. The labels often fit better when the bars are horizontal. You also often have more room left to right than you do up and down in a graph, so the bars can stretch out more, allowing you to more easily discern small and subtle differences.

Remember the fault of default principle. Always try different ways of displaying your data. It costs nothing other than a few electrons to display a horizontal alternative to the typical vertical bar chart format, so why not indulge yourself?



Size

- Length varies, width doesn’t
  - Exception, mosaic plots
  - Think about gaps between bars

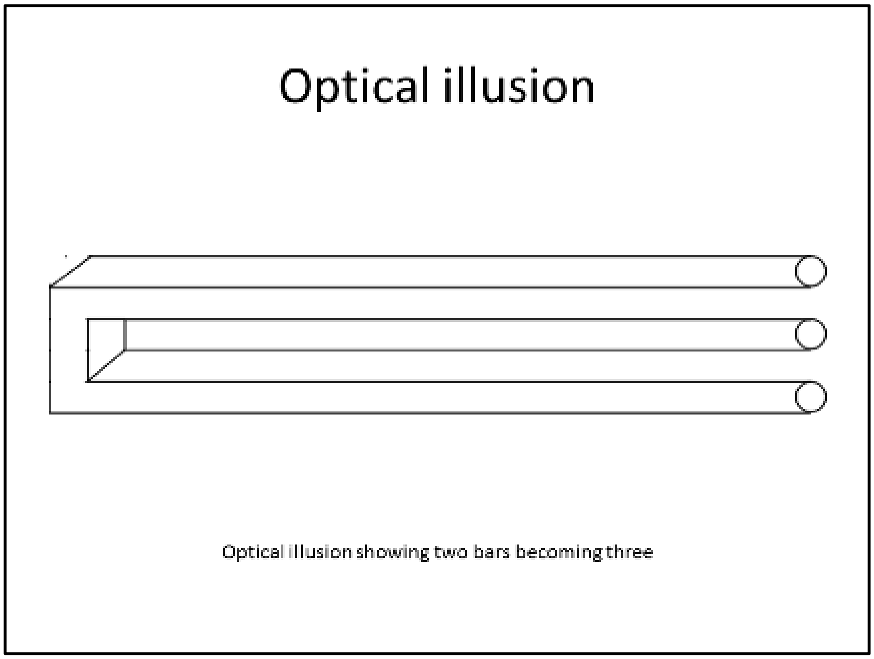


Don’t make gaps equal to widths

Alternate white and black lines showing a Moire effect

If the space between the bars is equal to the width of the bars themselves, you get an unsettling vibratory effect. This is because your eye is constantly shifting perspective. Sometimes it perceives the black as the foreground and the white as the background. Sometimes it perceives the white as the foreground and the black as the background.

“I’m ten years old, my life’s half over. And I don’t even know if I’m black with white stripes or white with black stripes.” Marty the Zebra in the movie Madagascar.

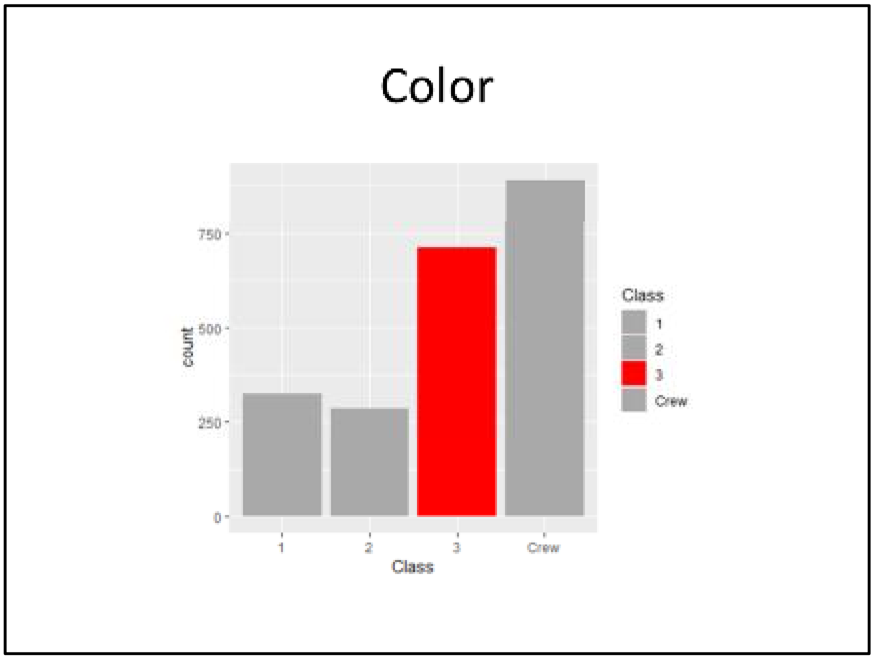


Optical illusion

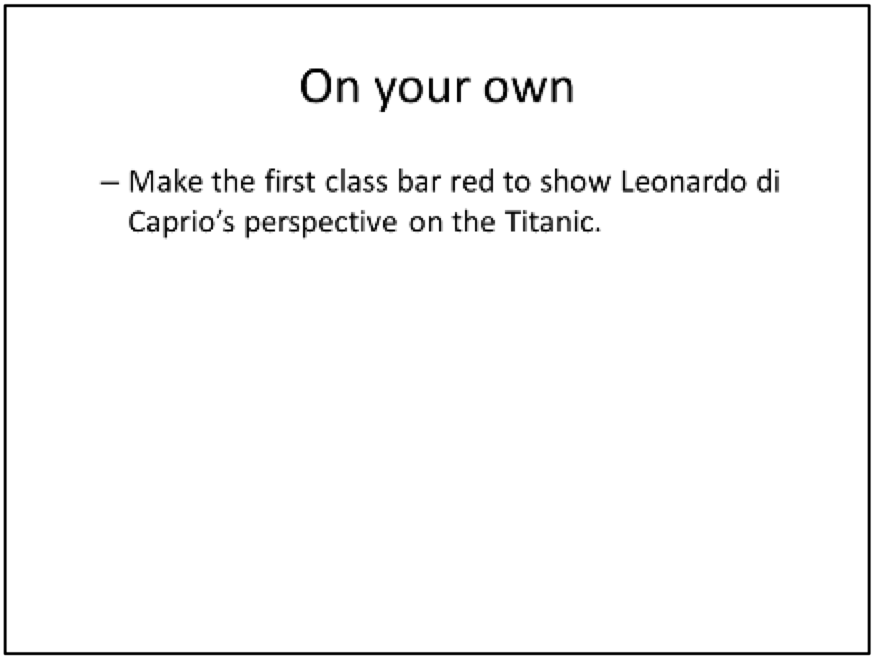
Optical illusion showing two bars becoming three

If the widths are the same and the bars are empty, then you can get a differnt problem. You might get confused as to what is the bar and what is the gap, as in this optical illusion.

As a general rule, the gap between bars should be about 10 to 20 percent of the width of the bars. Most visualization software has sensible defaults, but beware when you have a very large number of bars. There’s always a bit of rounding when you place pixels on a screen or on a page. Your software is trying to fill the available

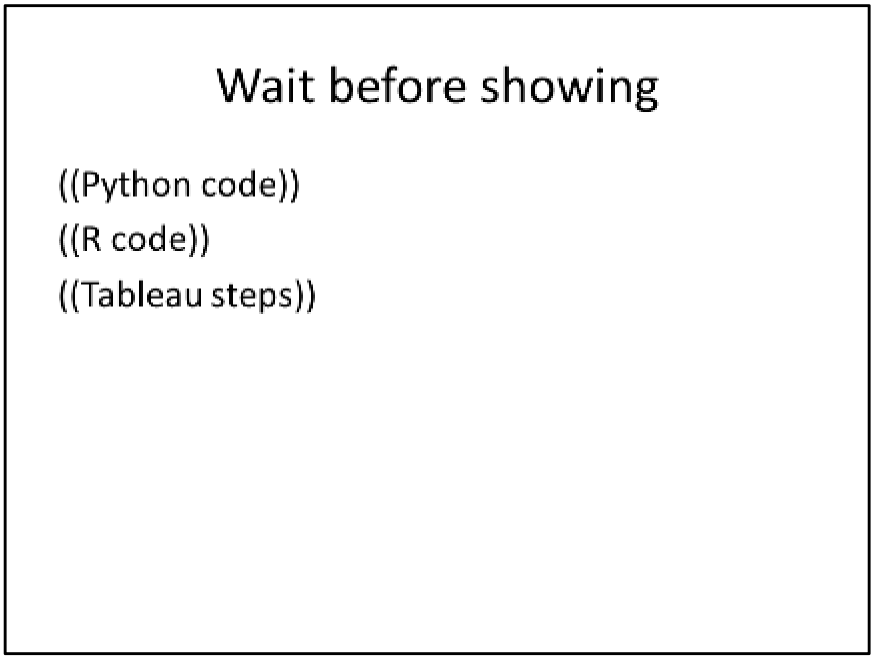


Color



On your own

– Make the first class bar red to show Leonardo di Caprio’s perspective on the Titanic.

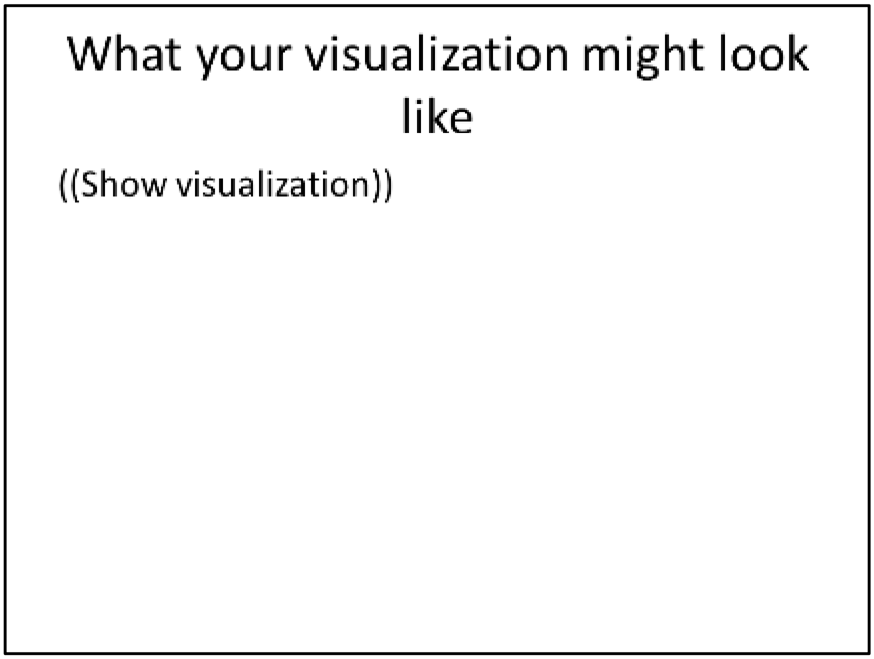


Wait before showing

((Python code))

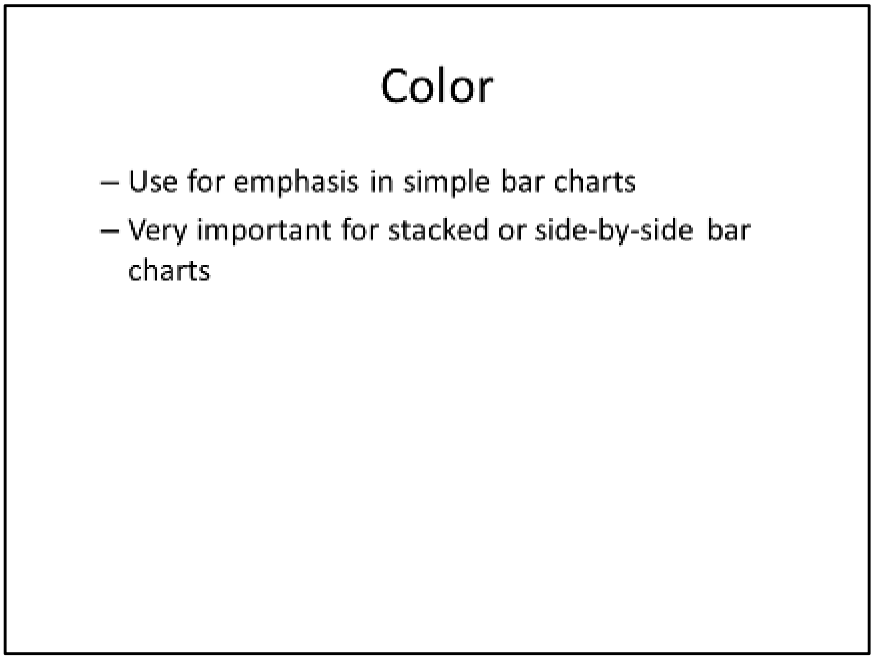
((R code))

((Tableau steps))



What your visualization might look like

((Show visualization))

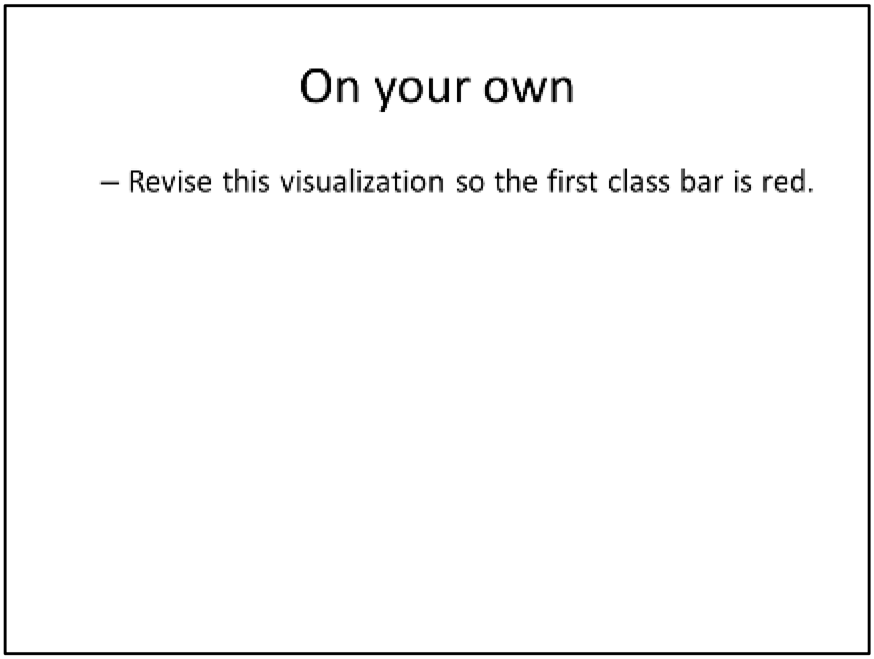


Color

- Use for emphasis in simple bar charts
- Very important for stacked or side-by-side bar charts

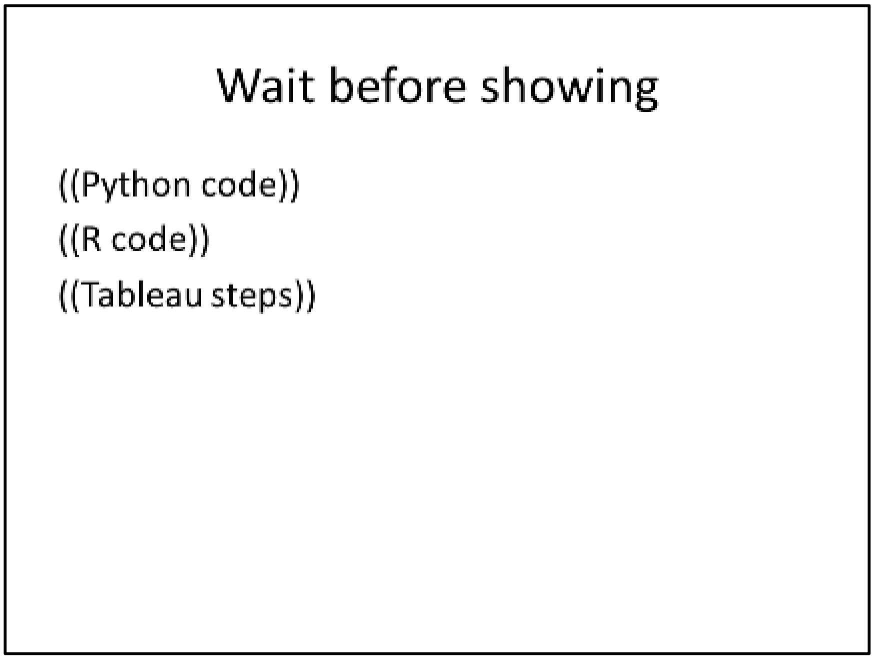
For simple bar charts like all the ones we've seen so far, color is not really needed. You already can distinguish between the passenger classes using the location. If you do use color in a simple bar chart, it is often to emphasize a point. In the previous bar chart, I used red for third class, because Kate Winslet, a rich first class passenger, found true love in third class, with the adorable Leonardo di Caprio.

Color becomes very important in just a minute when we add another layer of complexity.



On your own

– Revise this visualization so the first class bar is red.

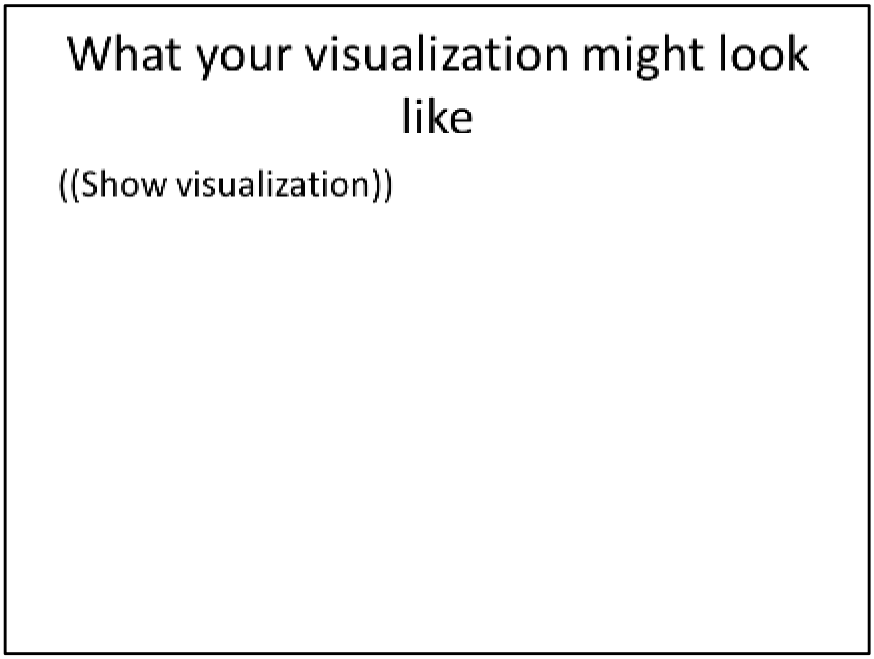


Wait before showing

((Python code))

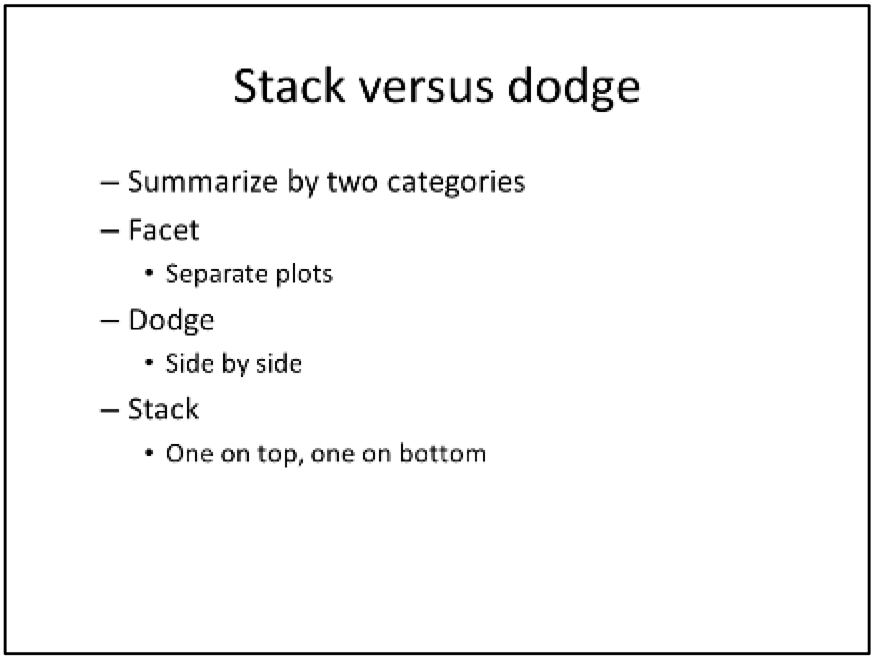
((R code))

((Tableau steps))



What your visualization might look like

((Show visualization))

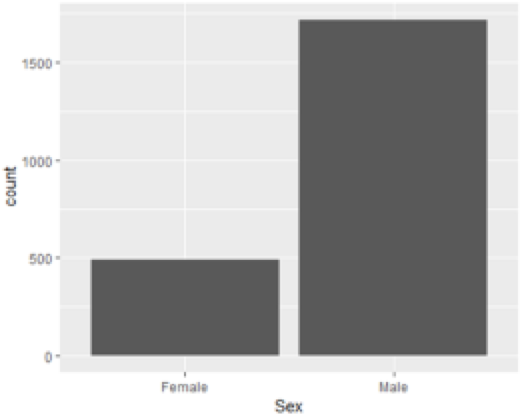


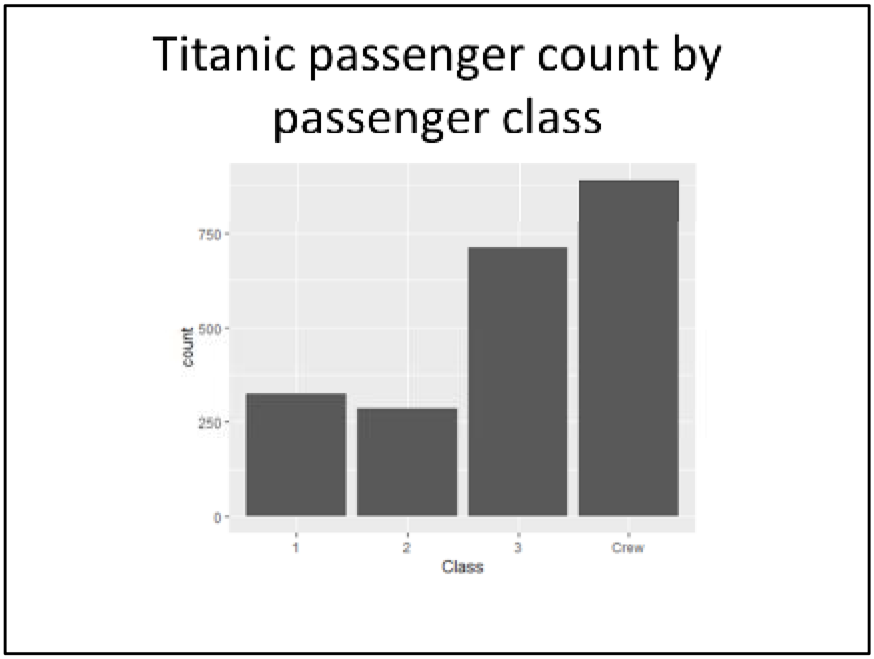
Stack versus dodge

- Summarize by two categories
- Facet
  - Separate plots
- Dodge
  - Side by side
- Stack
  - One on top, one on bottom

Bar charts that represent a summary across a single categorical variable are fairly simple and easy to handle. But when you want to summarize by two categories simultaneously, things get interesting. Interesting in a good way.

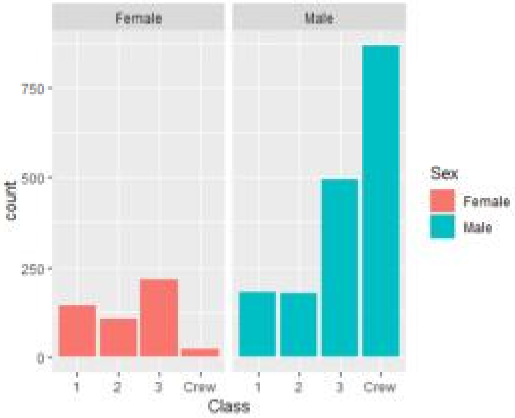
Titanic passenger count by gender



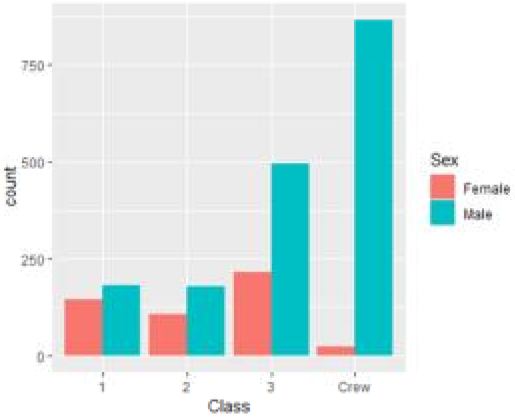


Titanic passenger count by passenger class

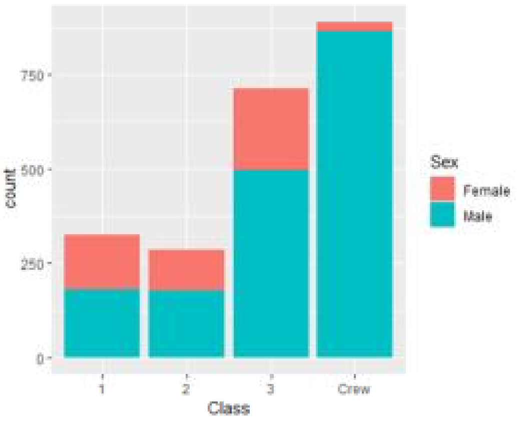
Titanic facet

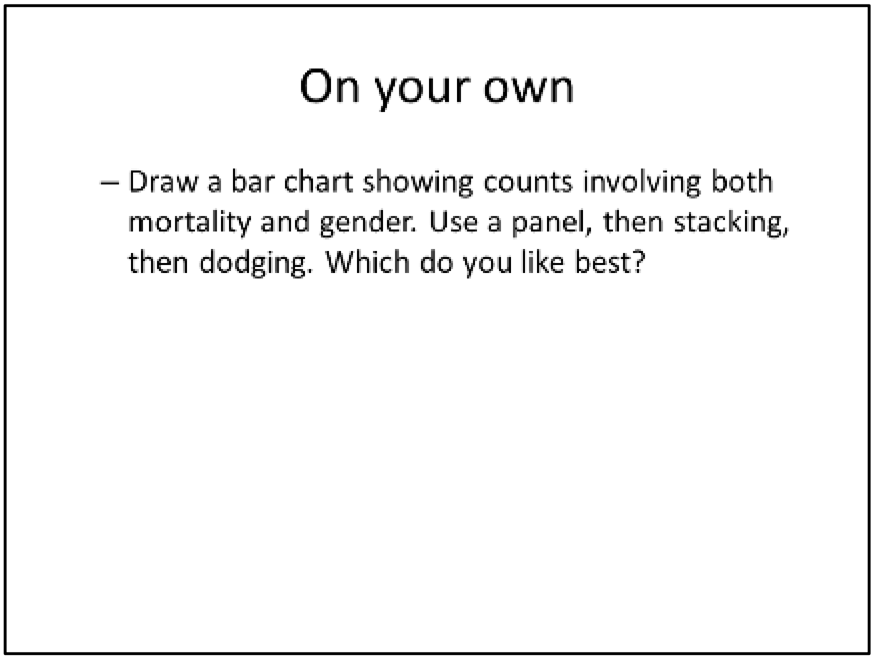


Titanic dodge



Titanic stack



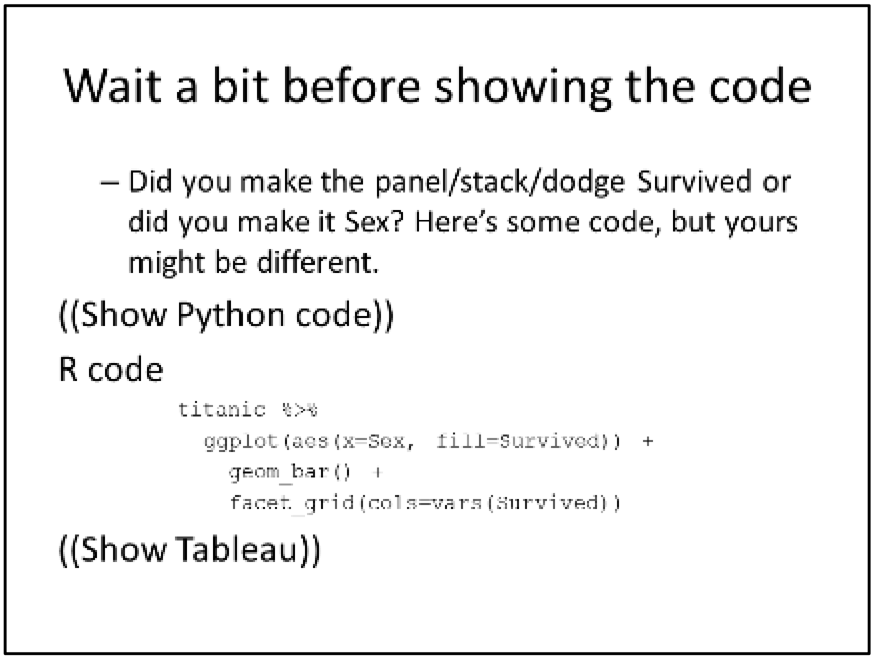


On your own

– Draw a bar chart showing counts involving both mortality and gender. Use a panel, then stacking, then dodging. Which do you like best?

Mortality on the Titanic reflects the fact that this was an era when people really did believe in the concept of “women and children first.” Someone like me, I’d be shoving the little kids aside so I could get on one of the lifeboats.

Anyway, examine how mortality is related to gender. It’s a spoiler alert, but on the Titanic, Kate survives, but Leonardo, I am so sad to say this, didn’t make it.



Wait a bit before showing the code

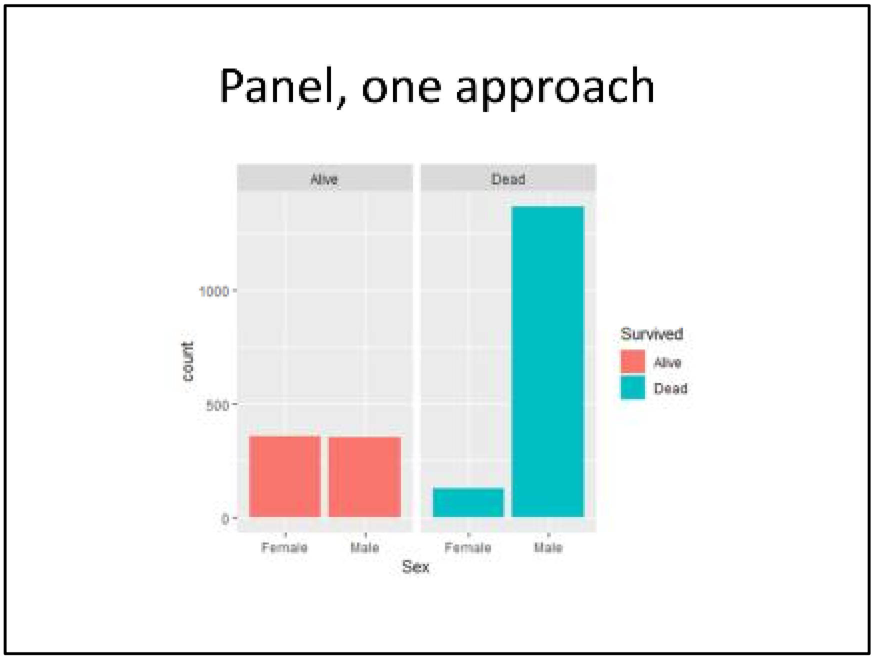
- Did you make the panel/stack/dodge Survived or did you make it Sex? Here's some code, but yours might be different.

((Show Python code))

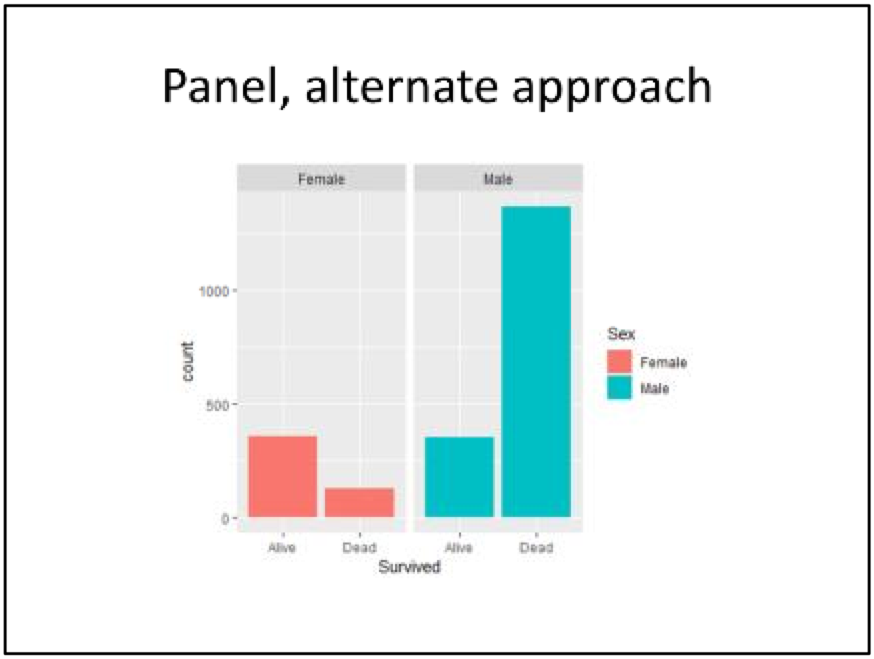
R code

```
titanic %>%
  ggplot(aes(x=Sex, fill=Survived)) +
    geom_bar() +
    facet_grid(cols=vars(Survived))
```

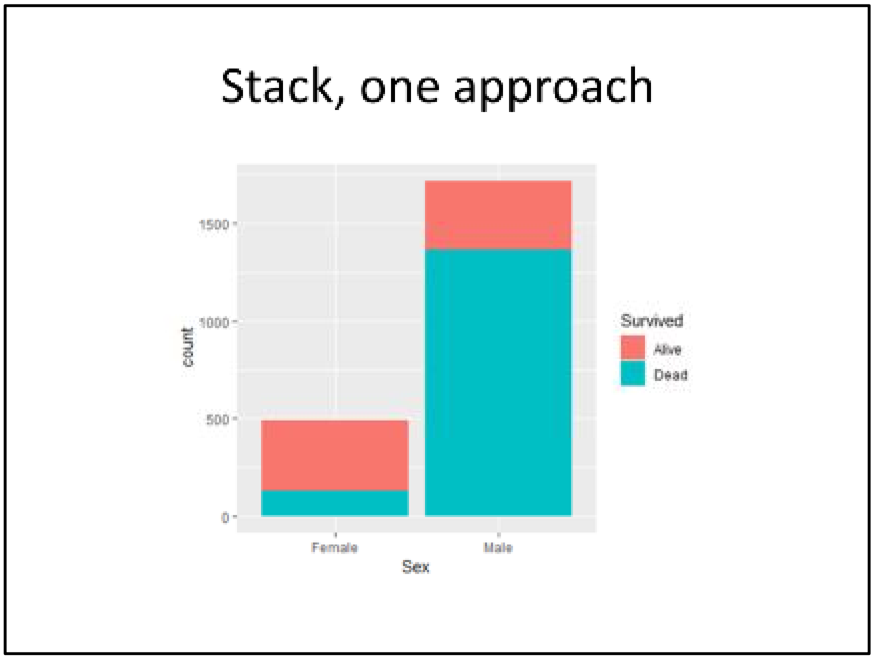
((Show Tableau))



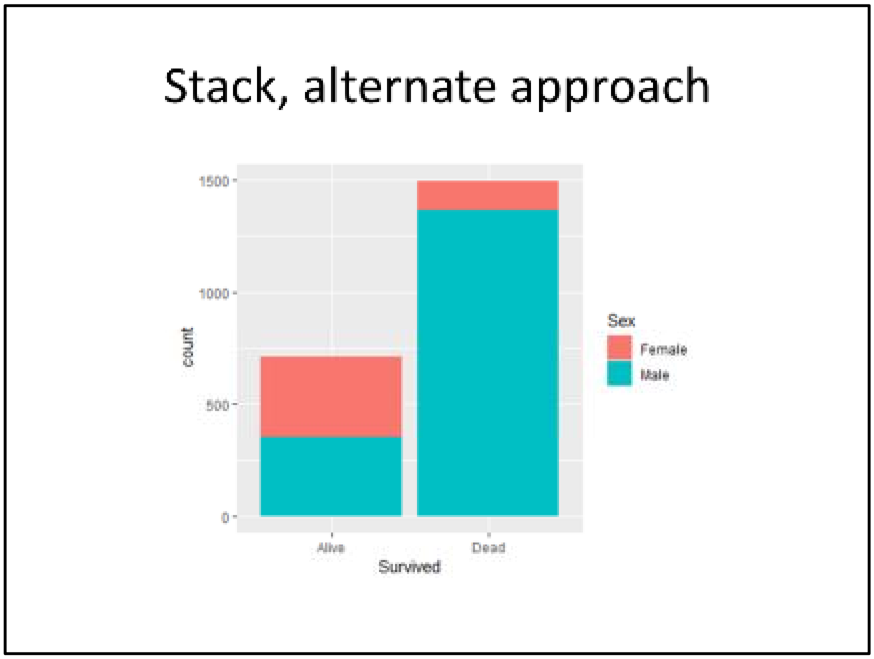
Panel, one approach



Panel, alternate approach

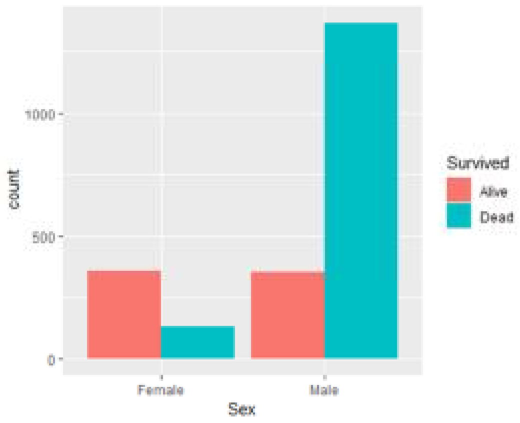


Stack, one approach

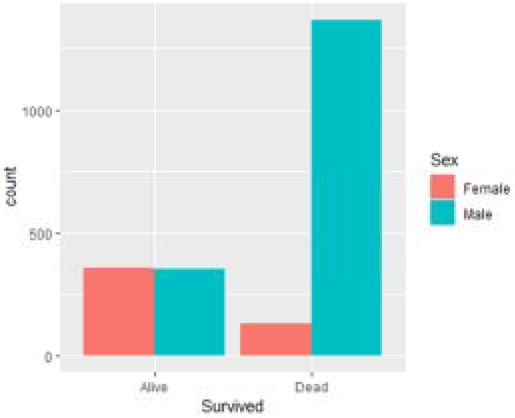


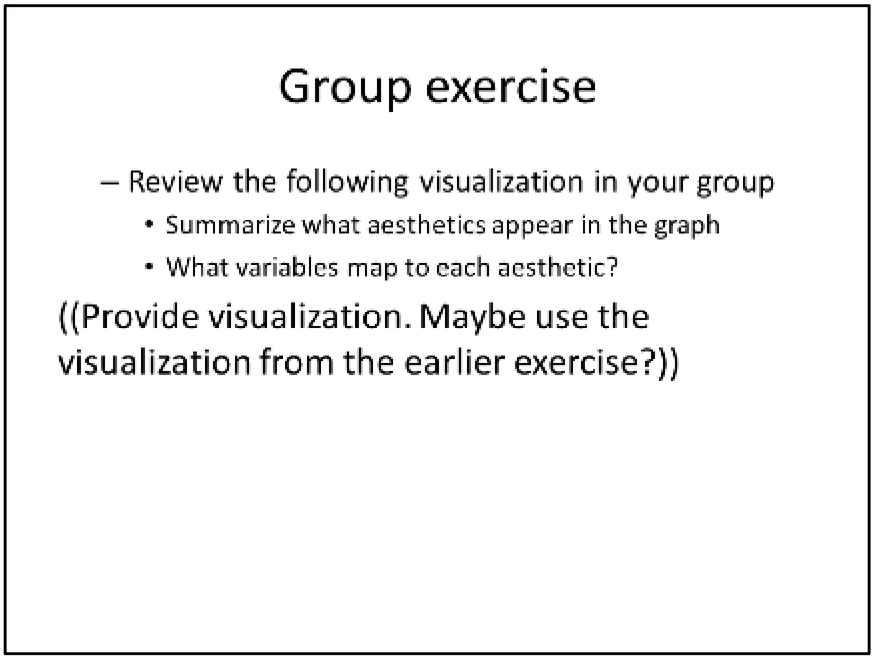
Stack, alternate approach

Dodge, one approach



Dodge, alternate approach

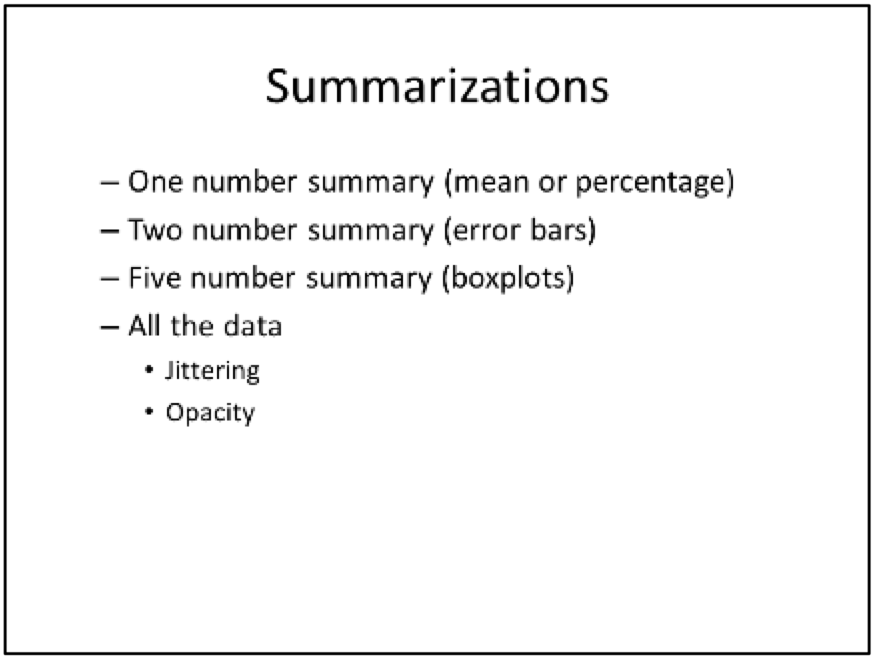




Group exercise

- Review the following visualization in your group
  - Summarize what aesthetics appear in the graph
  - What variables map to each aesthetic?

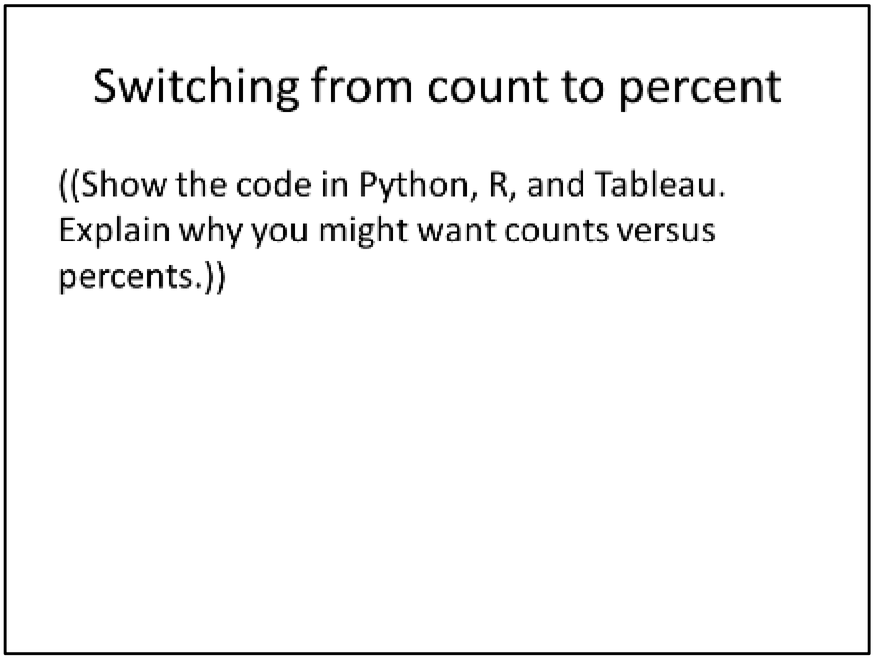
((Provide visualization. Maybe use the visualization from the earlier exercise?))



Summarizations

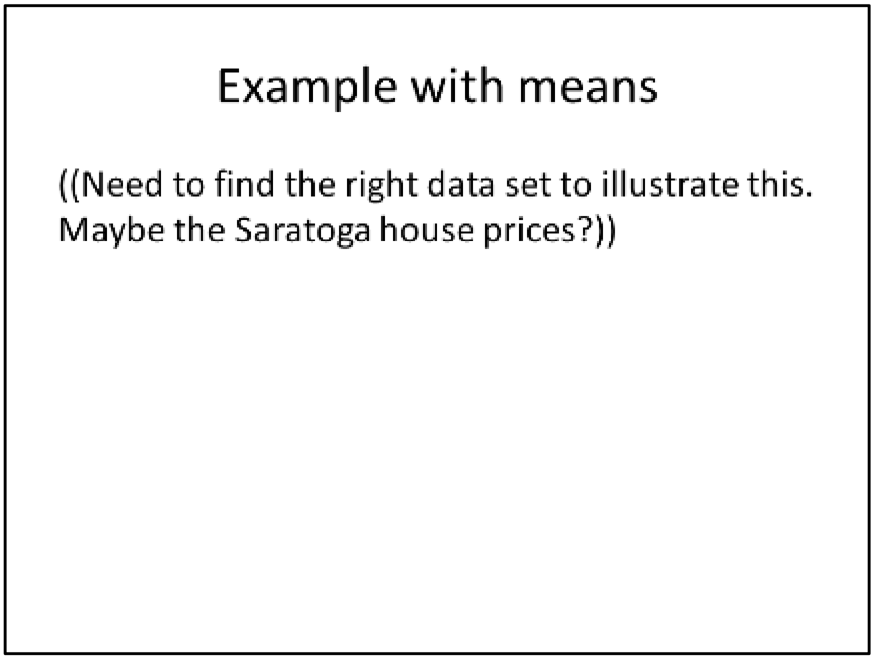
- – One number summary (mean or percentage)
- – Two number summary (error bars)
- – Five number summary (boxplots)
- – All the data
  - Jittering
  - Opacity

((note to myself: maybe this goes better in the talk about lines.))



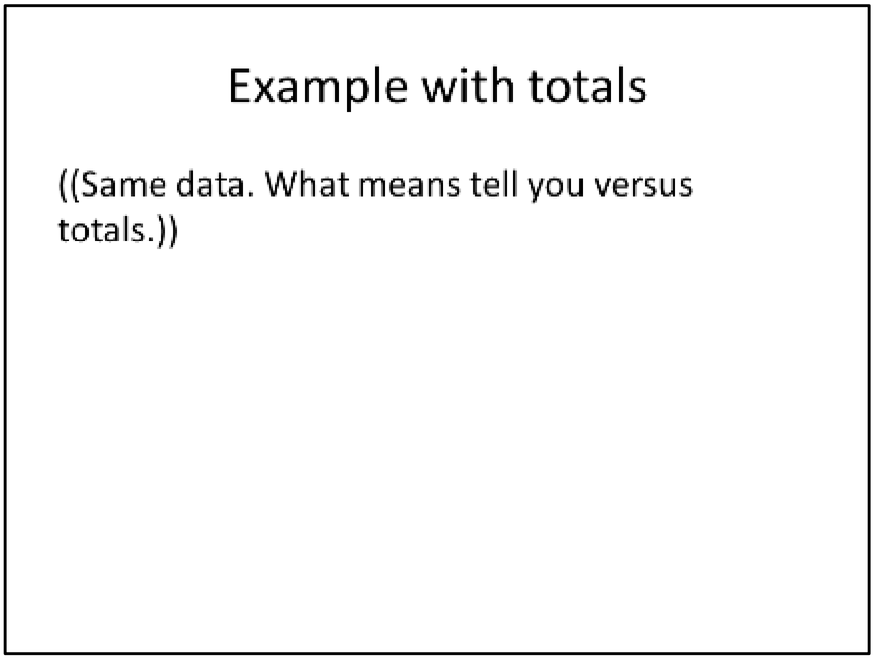
Switching from count to percent

((Show the code in Python, R, and Tableau. Explain why you might want counts versus percents.))



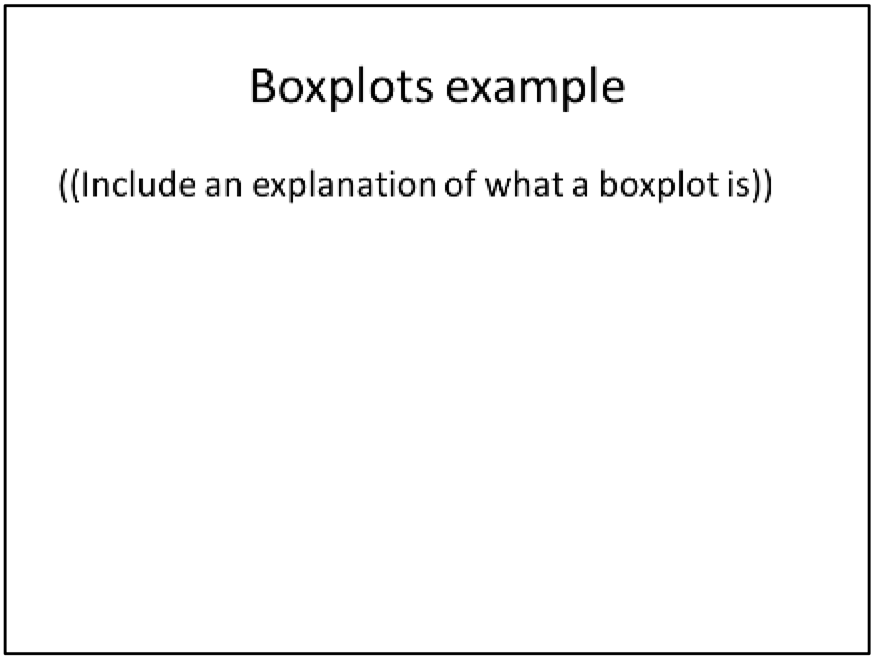
Example with means

((Need to find the right data set to illustrate this. Maybe the Saratoga house prices?))



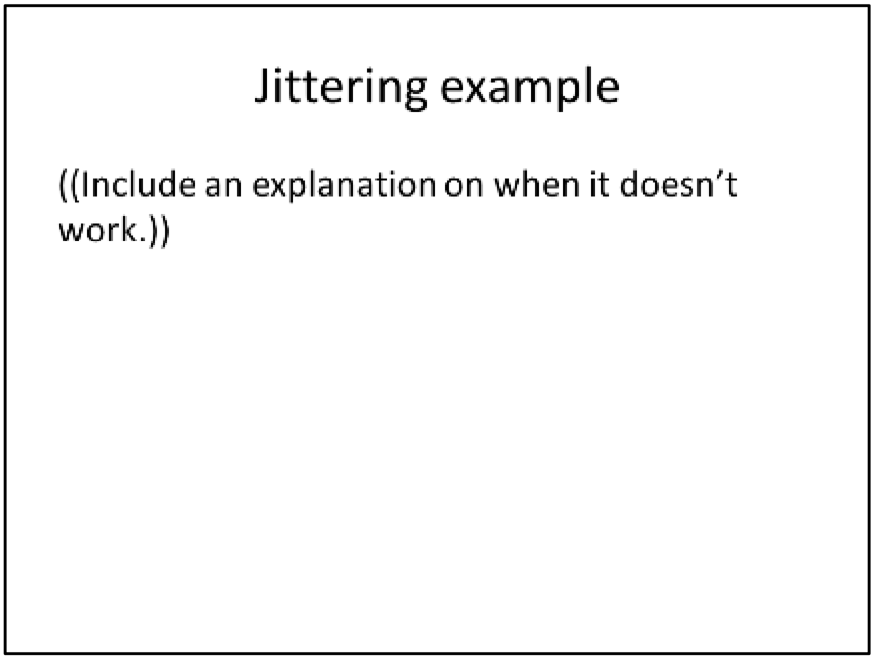
Example with totals

((Same data. What means tell you versus totals.))



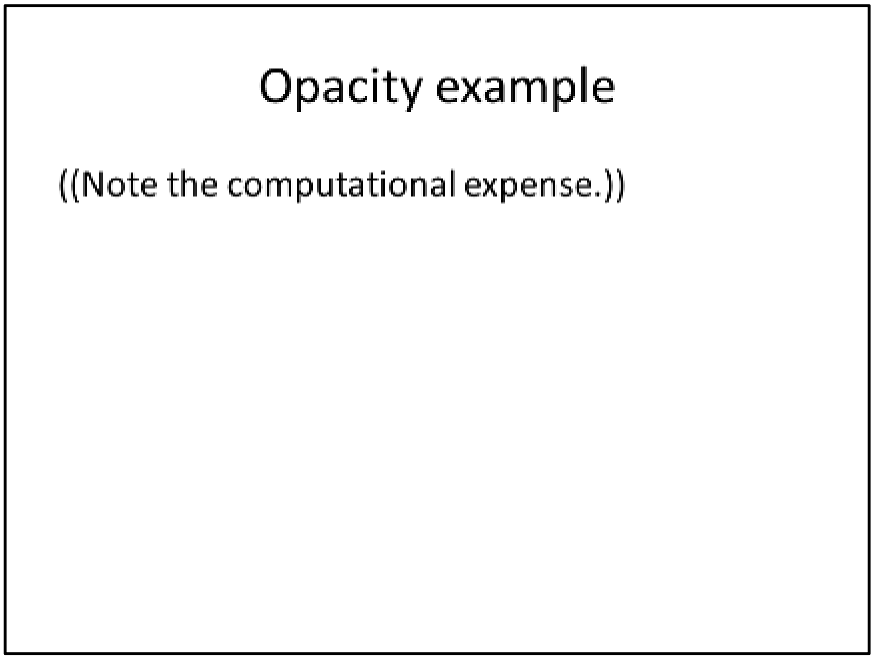
Boxplots example

((Include an explanation of what a boxplot is))



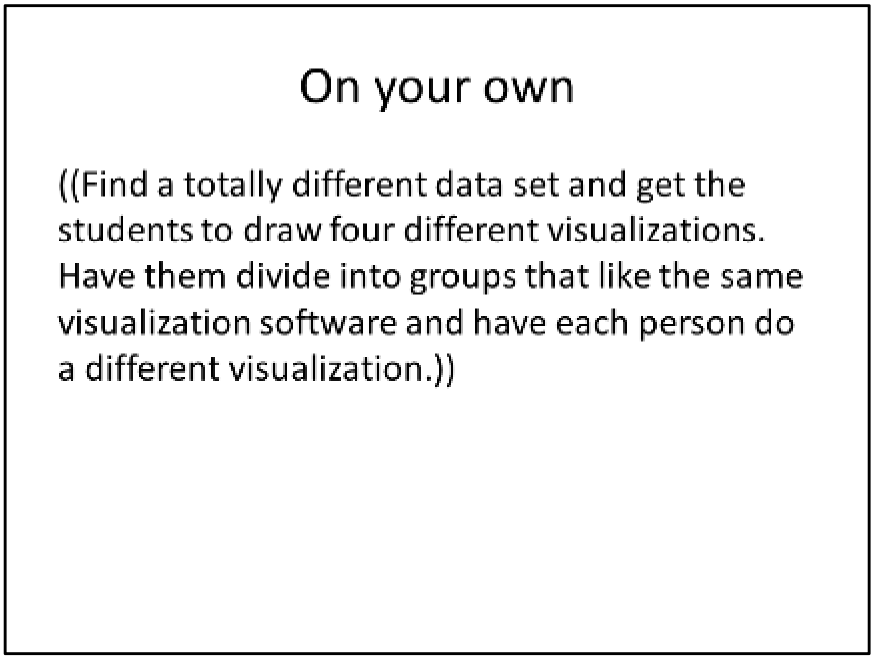
Jittering example

((Include an explanation on when it doesn't work.))



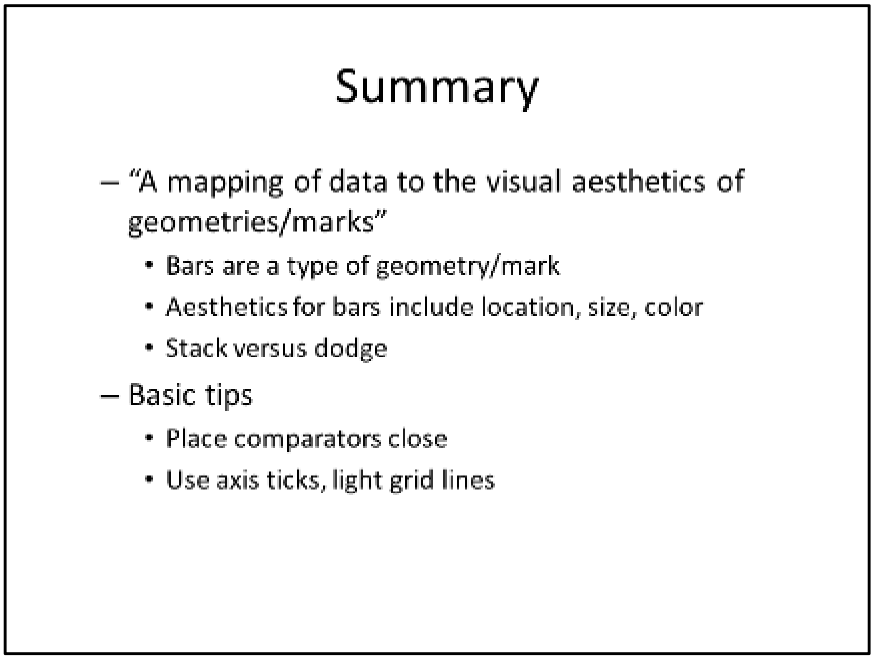
Opacity example

((Note the computational expense.))



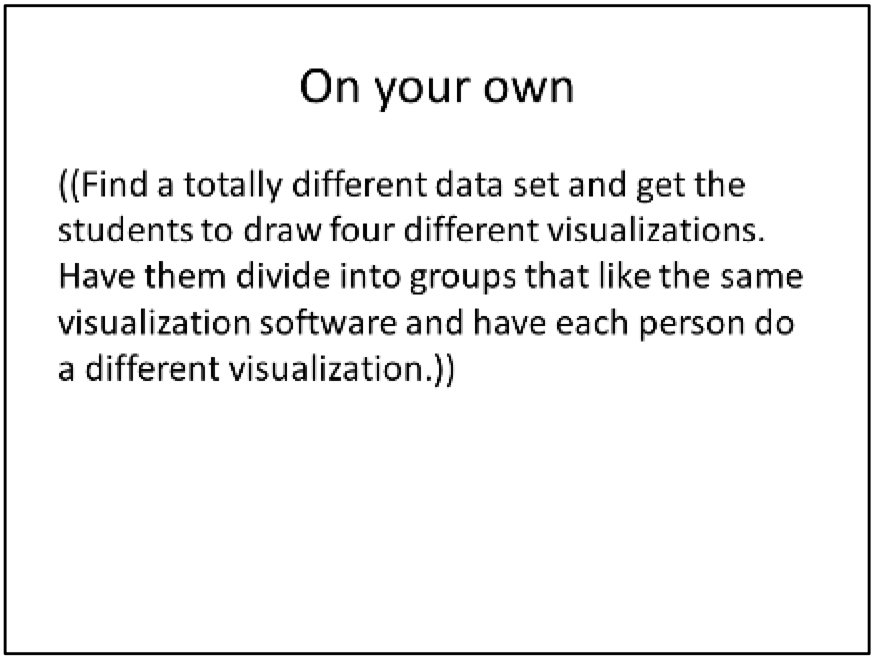
On your own

((Find a totally different data set and get the students to draw four different visualizations. Have them divide into groups that like the same visualization software and have each person do a different visualization.))



Summary

- “A mapping of data to the visual aesthetics of geometries/marks”
  - Bars are a type of geometry/mark
  - Aesthetics for bars include location, size, color
  - Stack versus dodge
- Basic tips
  - Place comparators close
  - Use axis ticks, light grid lines



On your own

((Find a totally different data set and get the students to draw four different visualizations. Have them divide into groups that like the same visualization software and have each person do a different visualization.))