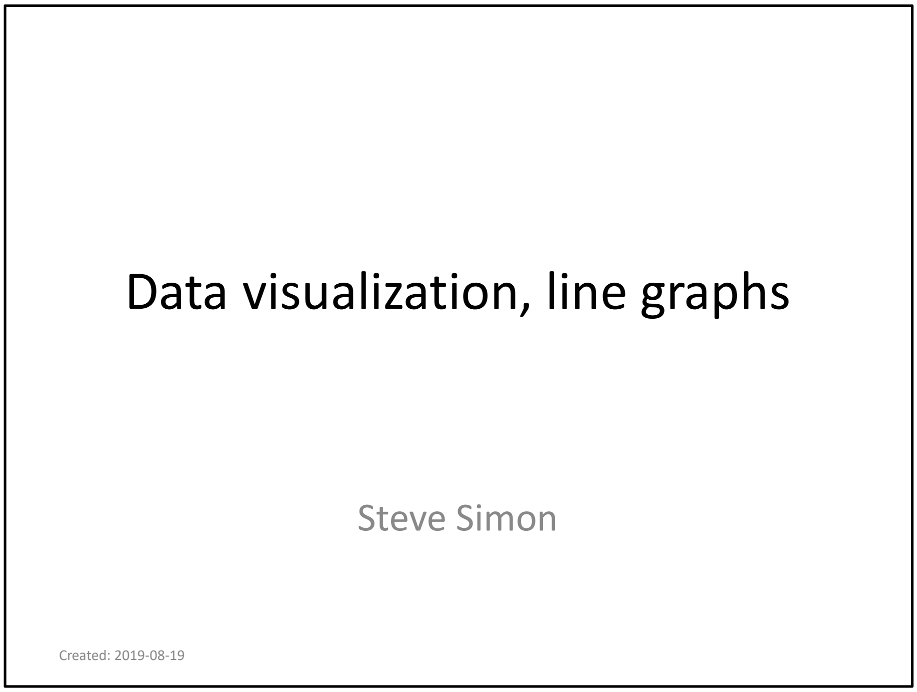
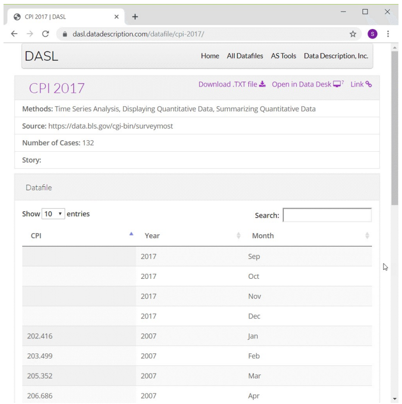


Data visualization, line graphs

Steve Simon

Created: 2019-08-19

Preparation, DASL Consumer Price Index data



CPI 2017

Methods: Time Series Analysis, Displaying Quantitative Data, Summarizing Quantitative Data

Source: https://data.bls.gov/cgi-bin/surveymost

Number of Cases: 132

Story:

Datafile

Show 10 entries

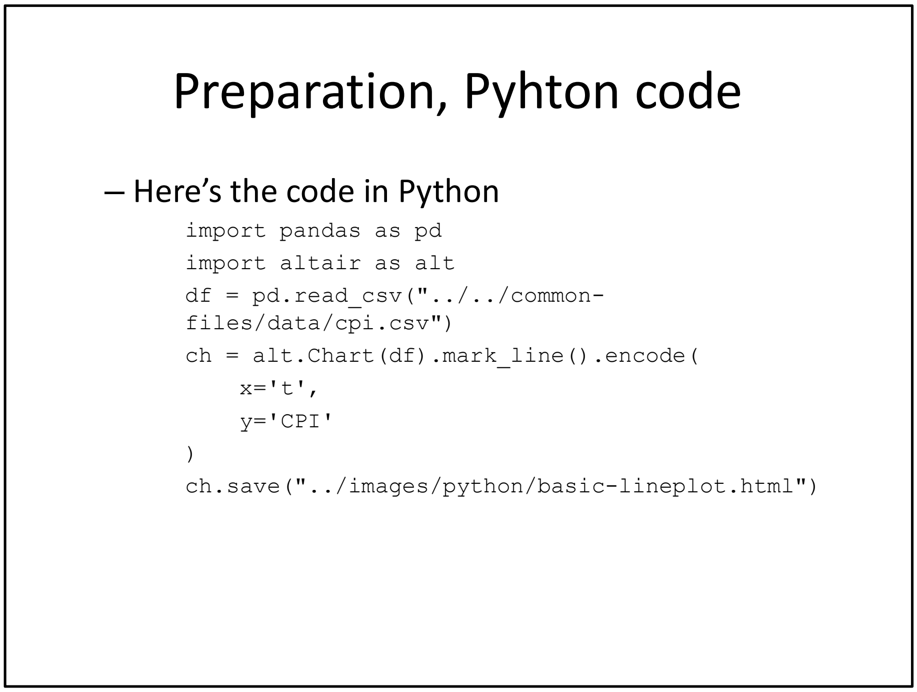
Search:

CPI	Year	Month
	2017	Sep
	2017	Oct
	2017	Nov
	2017	Dec
202.416	2007	Jan
203.499	2007	Feb
205.352	2007	Mar
206.686	2007	Apr

Screenshot of DASL website

One of the data sets that I will use in this moldule on line graphs is monthly values of the Consumer Price Index from 2007 through 2016. I have to make some minor modifications to the data, adding a continuous variable for time and deleting four rows of data.

You can find information about this data set on the DASL website, but use the csv file that I created instead of the one from the DASL website.

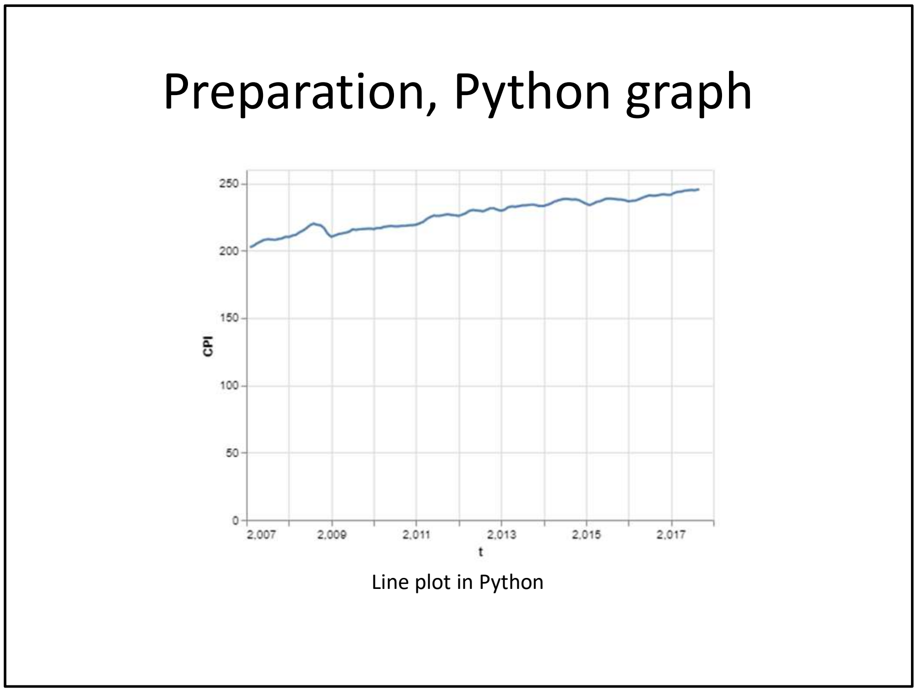


Preparation, Pyhton code

– Here’s the code in Python

```
import pandas as pd
import altair as alt
df = pd.read_csv("../../common-
files/data/cpi.csv")
ch = alt.Chart(df).mark_line().encode(
    x='t',
    y='CPI'
)
ch.save("../images/python/basic-lineplot.html")
```

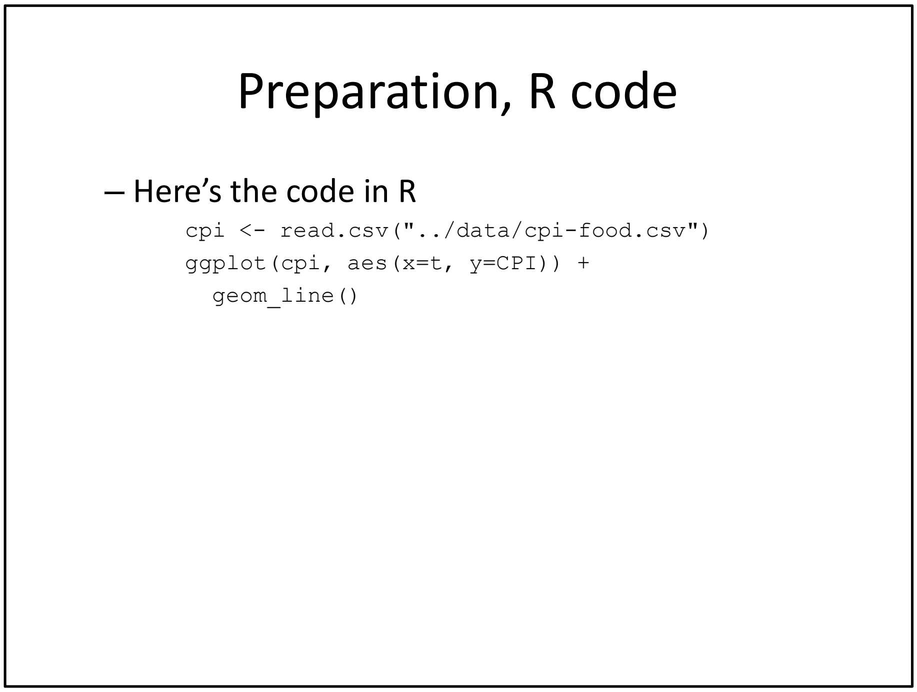
Here is the Python code to read in the data and produce a simple line graph.



Preparation, Python graph

Line plot in Python

This is what the graph would look like.

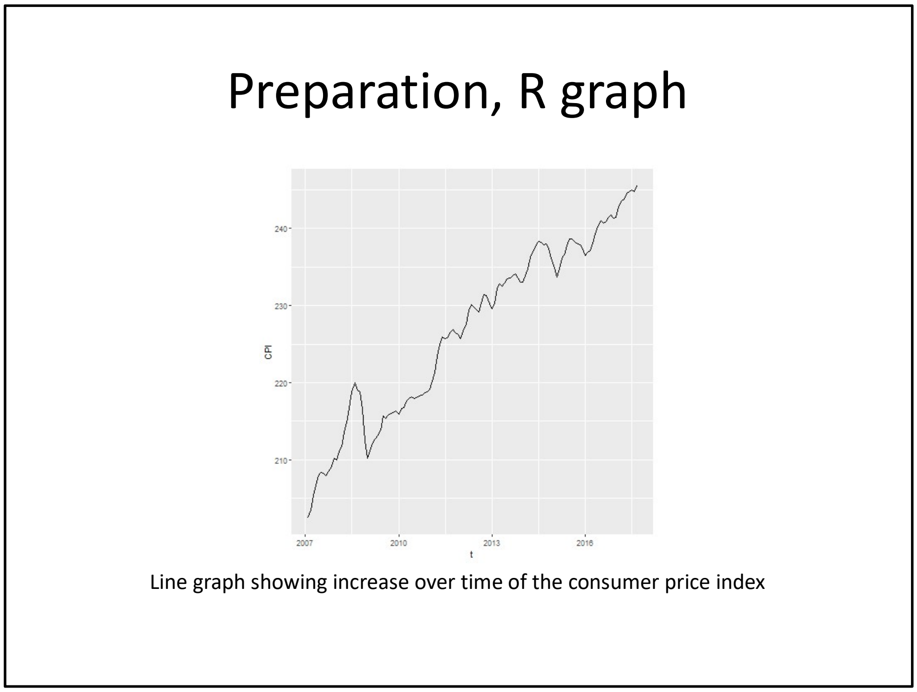


Preparation, R code

– Here’s the code in R

```
cpi <- read.csv("../data/cpi-food.csv")
ggplot(cpi, aes(x=t, y=CPI)) +
  geom_line()
```

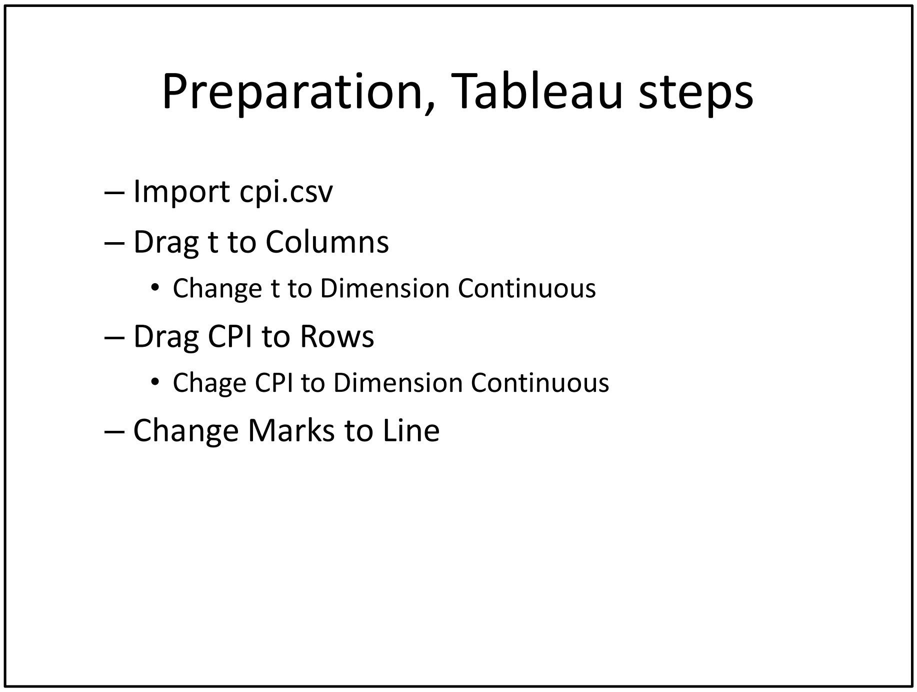
Here is the R code to read in the data and produce a simple line graph.



Preparation, R graph

Line graph showing increase over time of the consumer price index

Here is what the graph looks like in R. See if you can reproduce it.

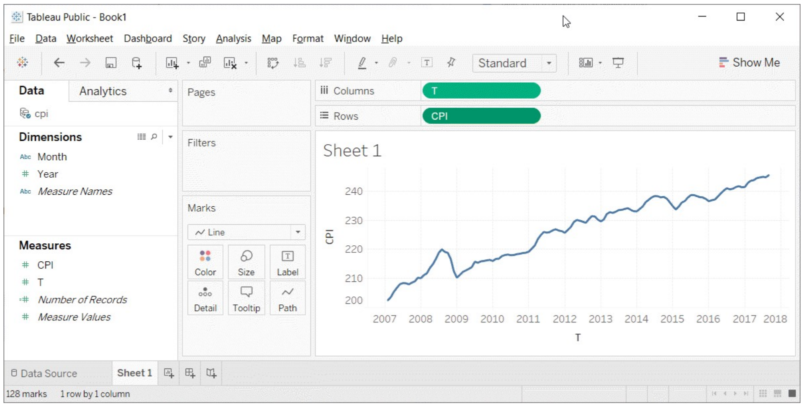


Preparation, Tableau steps

- – Import cpi.csv
- – Drag t to Columns
  - Change t to Dimension Continuous
- – Drag CPI to Rows
  - Chage CPI to Dimension Continuous
- – Change Marks to Line

Here are the steps in creating a line graph in Tableau.

Preparation, Tableau graph

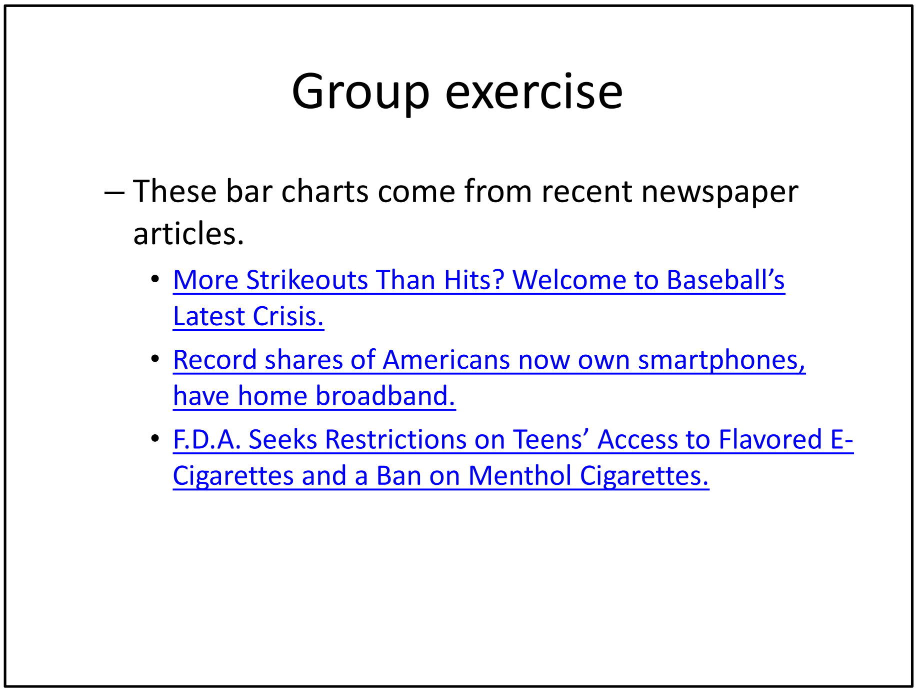


Columns: T

Rows: CPI

Line graph in Tableau

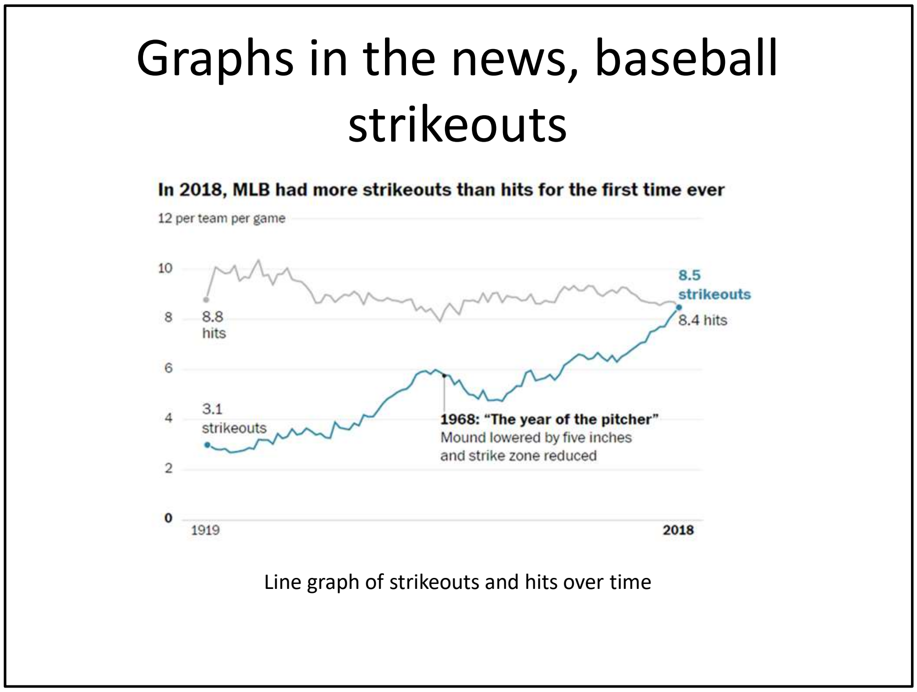
This is what the Tableau graph looks like.



Group exercise

- These bar charts come from recent newspaper articles.
  - More Strikeouts Than Hits? Welcome to Baseball’s Latest Crisis.
  - Record shares of Americans now own smartphones, have home broadband.
  - F.D.A. Seeks Restrictions on Teens’ Access to Flavored E-Cigarettes and a Ban on Menthol Cigarettes.

The following images are taken from various newspaper articles or press releases. Look at the graph and read/skim the article.

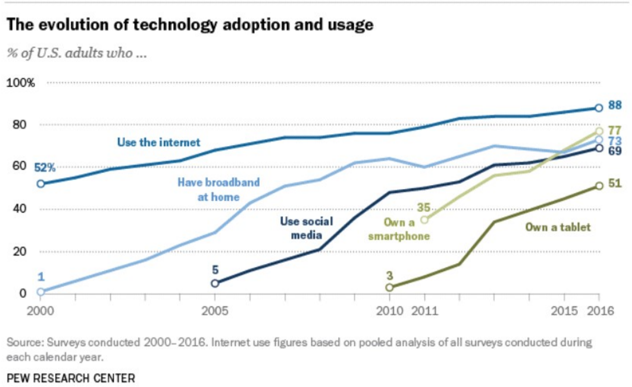


Graphs in the news, baseball strikeouts

Line graph of strikeouts and hits over time

Tyler Kepner. More Strikeouts Than Hits? Welcome to Baseball’s Latest Crisis. The New York Times, Augst 16, 2018. Retrieved from https://www.nytimes.com/2018/08/16/sports/baseball-mlb-strikeouts.html.

Graphs in the news, technology adaptation



The evolution of technology adoption and usage

% of U.S. adults who ...

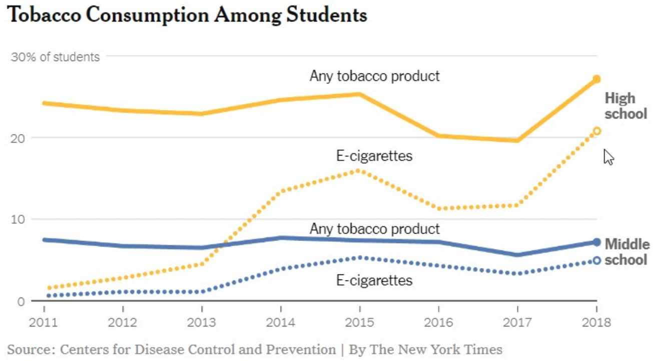
Source: Surveys conducted 2000–2016. Internet use figures based on pooled analysis of all surveys conducted during each calendar year.

PEW RESEARCH CENTER

Line graph of technology trends

Aaron Smith, Record shares of Americans now own smartphones, have home broadband. Pew Research Center, January 12, 2017. Retrived August 14, 2019 from https://www.pewresearch.org/fact-tank/2017/01/12/evolution-of-technology/.

Graphs in the news, Tobacco and e-cigarette consumption

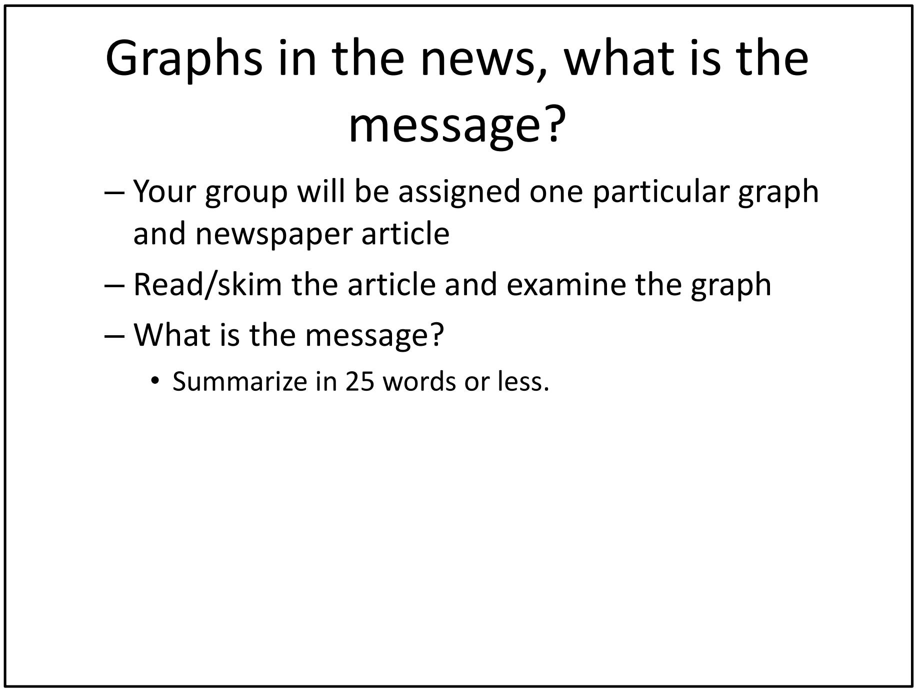


Tobacco Consumption Among Students

Source: Centers for Disease Control and Prevention | By The New York Times

Line graph of tobacco and e-cigarette consumption

Sheila Kaplan and Jan Hoffman, F.D.A. Seeks Restrictions on Teens’ Access to Flavored E-Cigarettes and a Ban on Menthol Cigarettes. The New York Times, November 15, 2018. Retrieved from http://nytimes.com/2018/11/15/health/ecigarettes-fda-flavors-ban.html

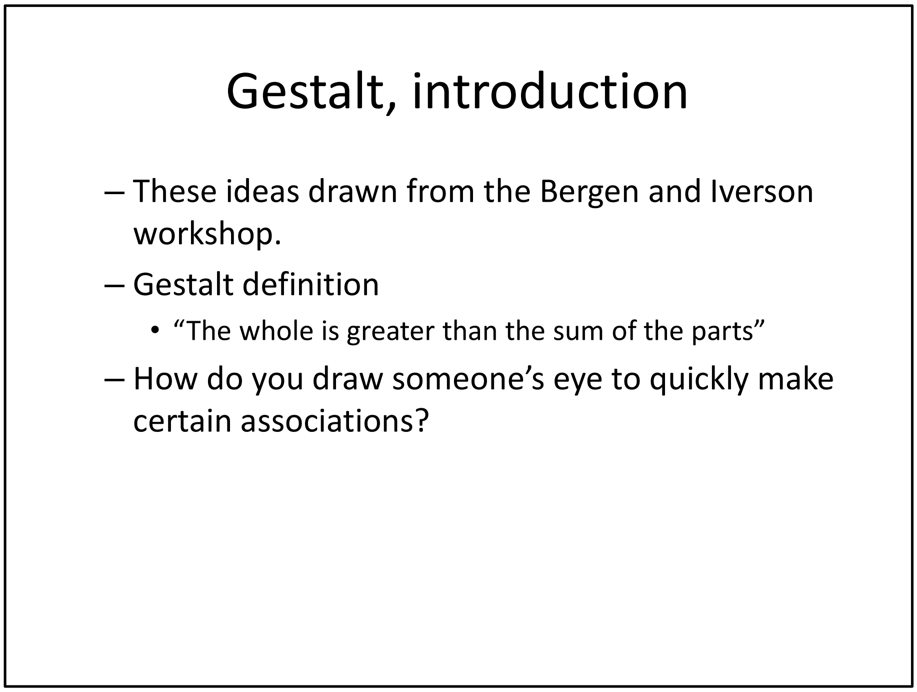


Graphs in the news, what is the message?

- Your group will be assigned one particular graph and newspaper article
- Read/skim the article and examine the graph
- What is the message?
  - Summarize in 25 words or less.

The following images are taken from various newspaper articles or press releases. Look at the graph and read/skim the article.

What message do you think the journalist is trying to convey with this graph. Summarize this message in 25 words or less.

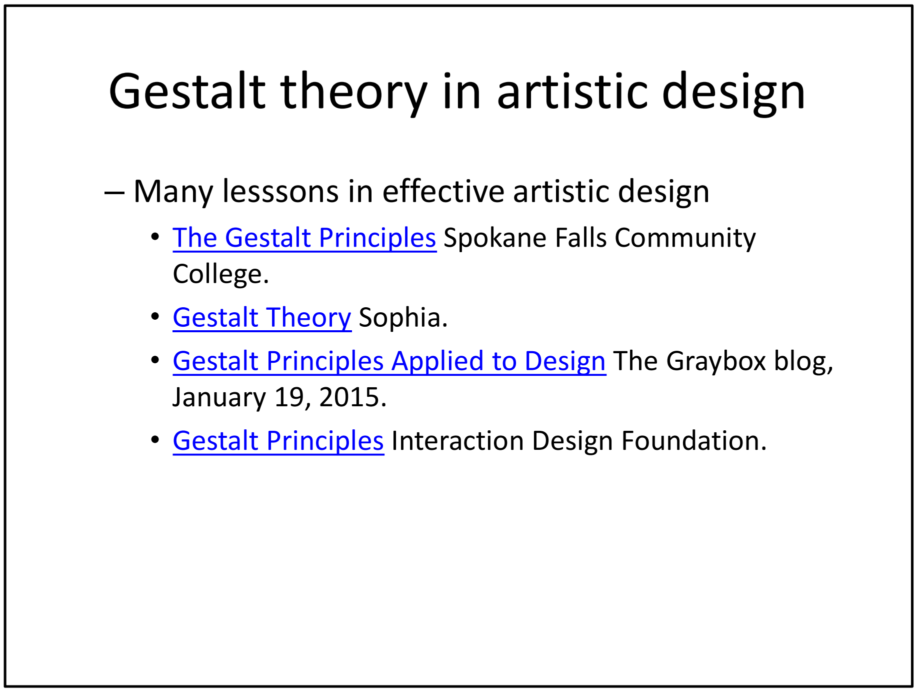


Gestalt, introduction

- These ideas drawn from the Bergen and Iverson workshop.
- Gestalt definition
  - “The whole is greater than the sum of the parts”
- How do you draw someone’s eye to quickly make certain associations?

I am borrowing heavily from https://github.com/WSU-DataScience/SDSS19-dataviz-workshop.

The Gestalt school of Psychology developed several principles that are very useful in helping to make an effective visualization. A simple definition of Gestalt is that the whole is more than just the individual items. There is a lot to Gestalt psychology, but the portion that is relevant to you is the ability to draw someone’s eye not to individual components of a graph, but to a group of related components that allow the viewer to see patterns or associations.

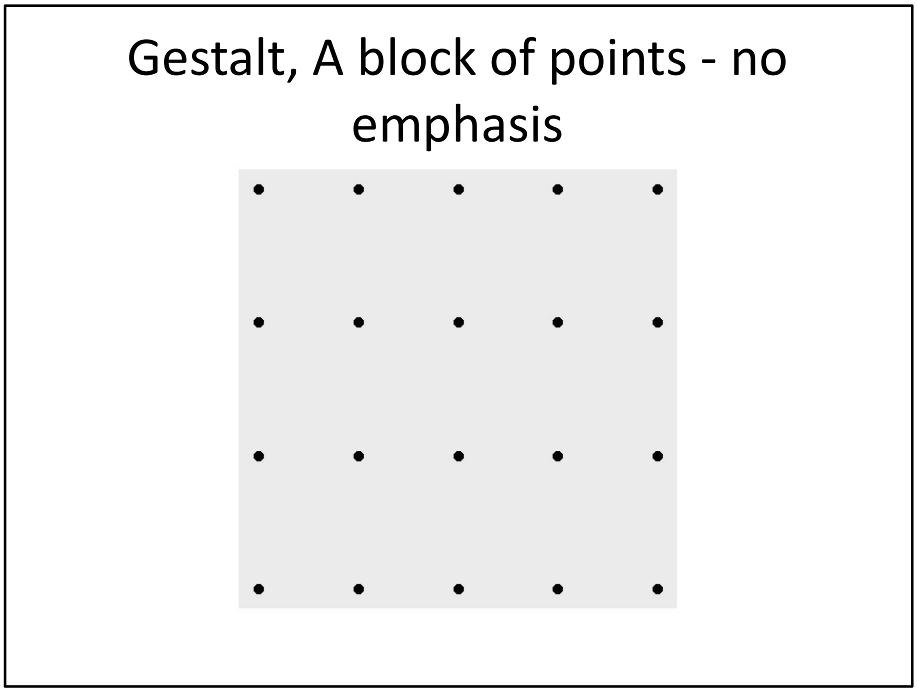


Gestalt theory in artistic design

– Many lesssons in effective artistic design
  - The Gestalt Principles Spokane Falls Community College.
  - Gestalt Theory Sophia.
  - Gestalt Principles Applied to Design The Graybox blog, January 19, 2015.
  - Gestalt Principles Interaction Design Foundation.

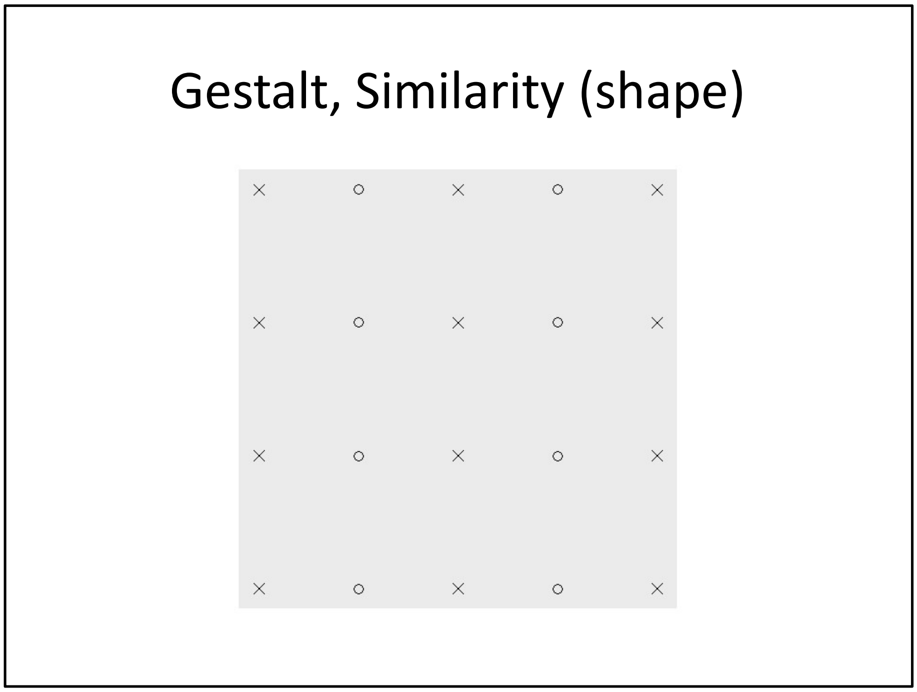
The field of artistic design is more commonly called graphic design, but I use the former term to distinguish it from statistical graphics. Artistic design is the use of typography, photography, and illustration to effectively convey a message. Artistic designers work on commercial logos, magazines and brochures, and product packaging.

These are interesting, and well worth some time viewing, even though they don't speak directly to the process of developing effective visualizations.



Gestalt, A block of points - no emphasis

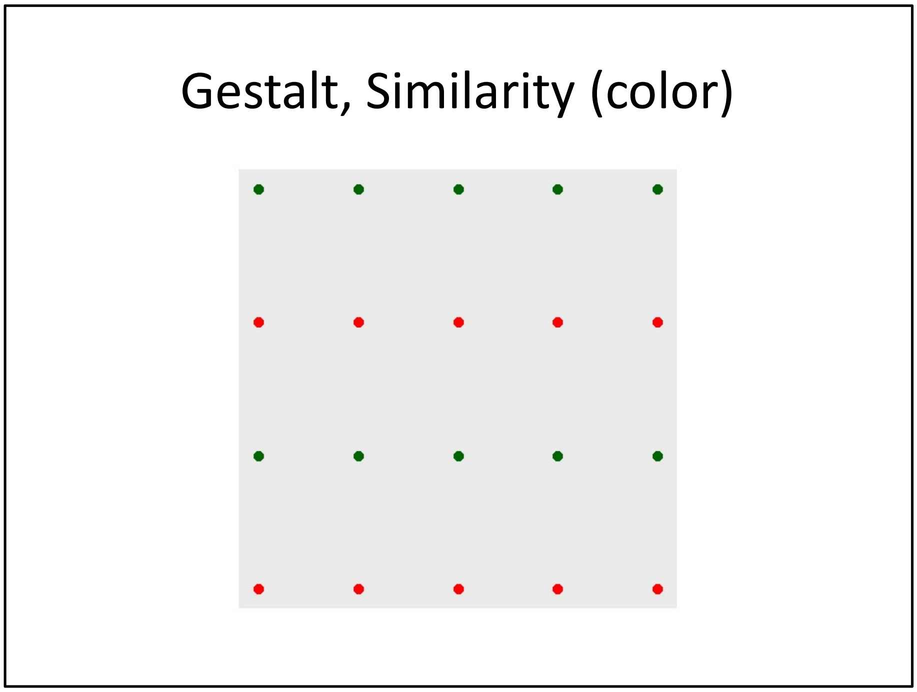
What you see in this graph is 20 individual points. When you design a graph, you want people to group things in a way to reveal patterns.



Gestalt, Similarity (shape)

Items that are similar tend to be grouped together.

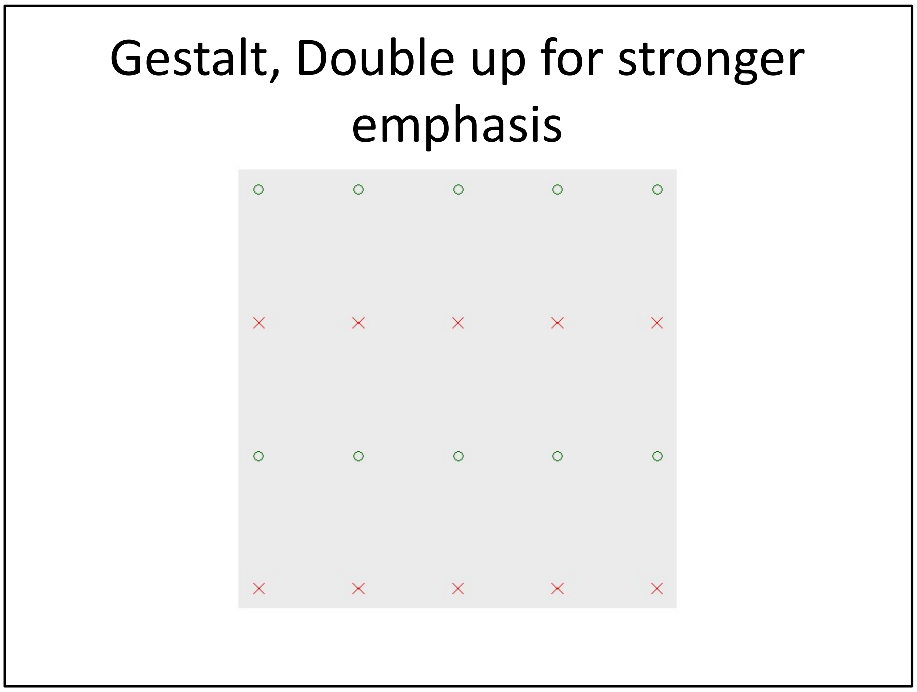
The use of common shapes causes you to see five columns rather than twenty individual data points.



Gestalt, Similarity (color)

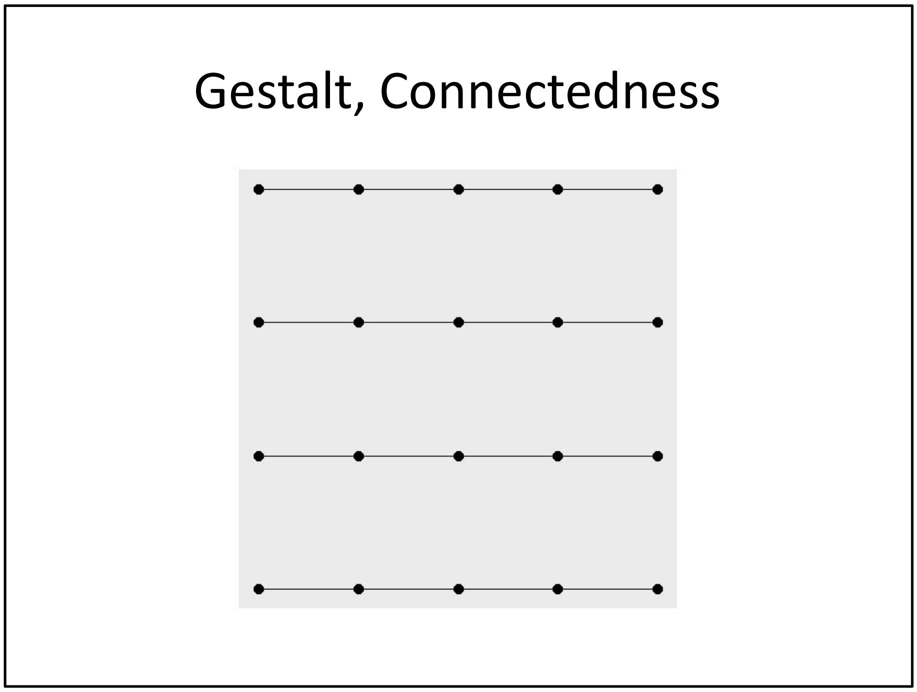
Notice how the use of color tends to draw the eye in and see the graph as not 20 separate points, but rather four rows of points.

The effect of color is a lot stronger than the effect of shape.



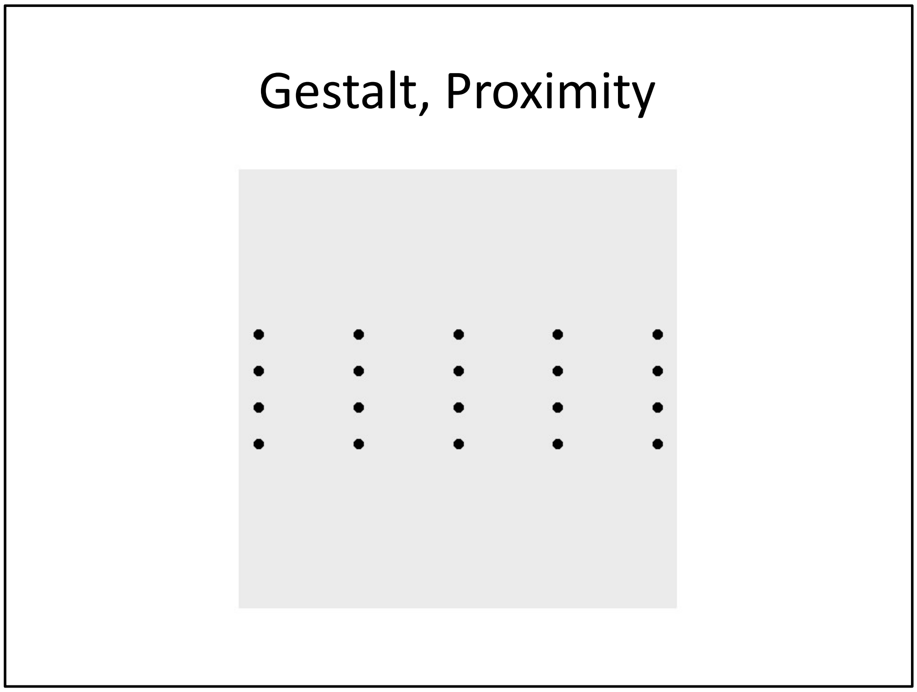
Gestalt, Double up for stronger emphasis

You can double up by using a common shape AND a common color to develop an even stronger association.



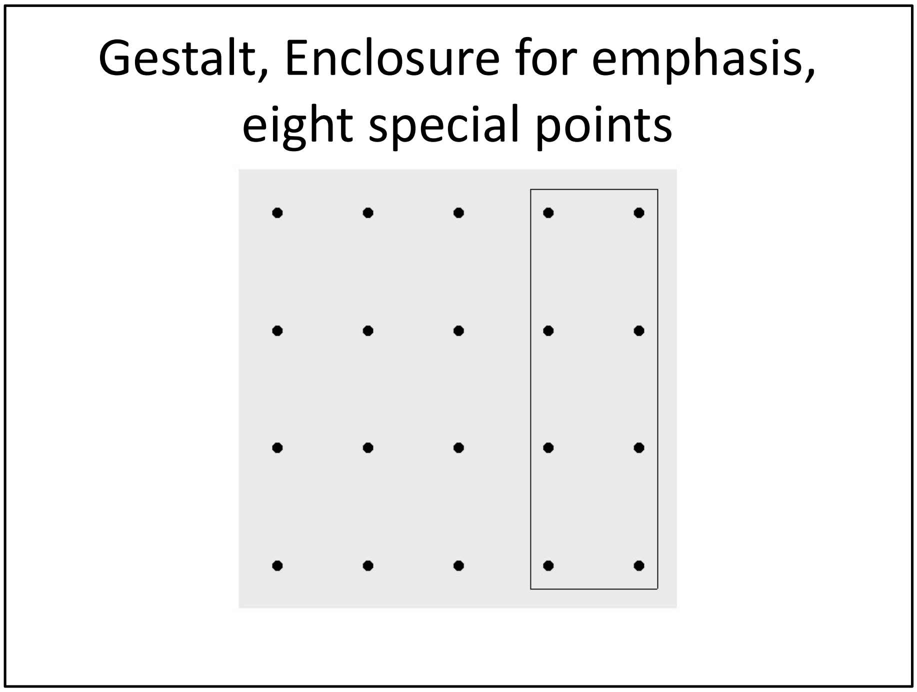
Gestalt, Connectedness

A connection between individual points causes you to perceive those connected points as part of a single group. Here, you see four rows rather than five columns.



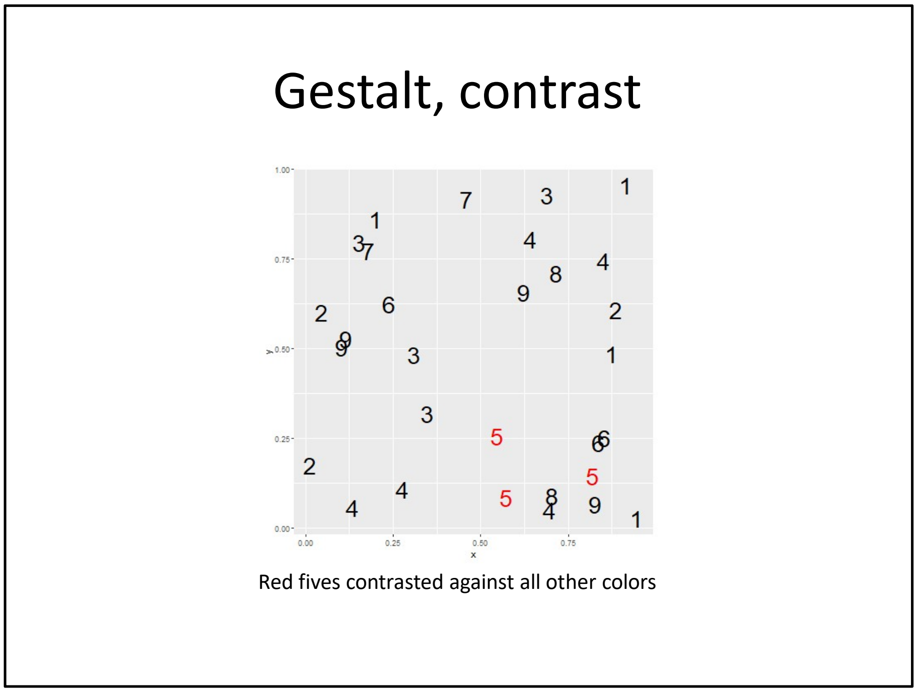
Gestalt, Proximity

Points that are close to one another also tend to encourage their grouping. Notice that the scaling change caused the points within a column to be a lot closer than the points within a row. So you tend to see this as five columns rather than four rows.



Gestalt, Enclosure for emphasis, eight special points

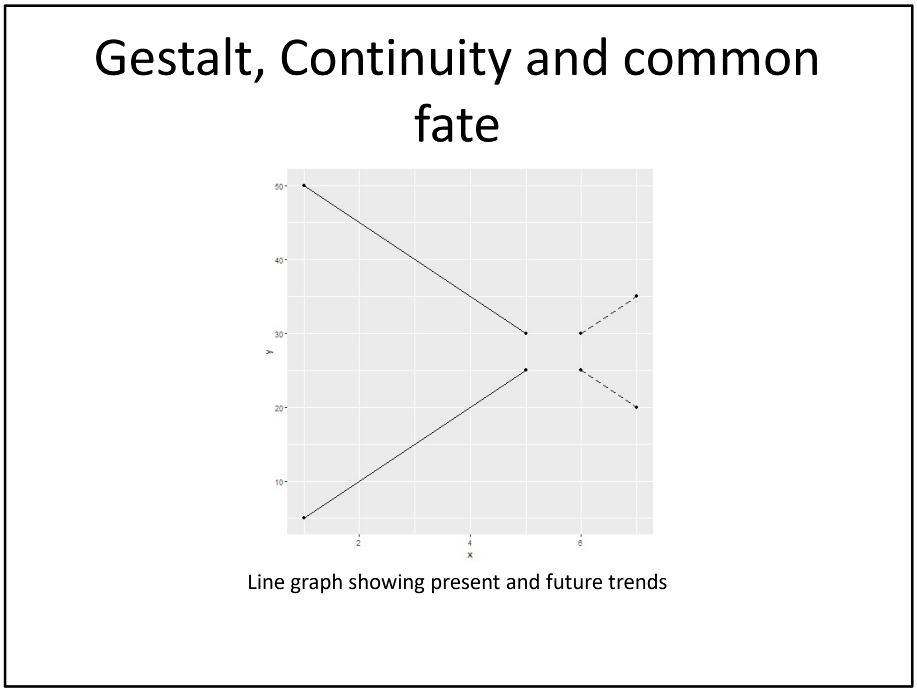
An enclosure is a strong way to emphasize grouping. Here the box provides an emphasis on a special group of eight points.



Gestalt, contrast

Red fives contrasted against all other colors

The one pattern that deviates from the others tends to draw your eye toards it. There are nine different groups here, 1 through 9, but because of the contrast, you tend to view it as 5 versus all the other numbers.

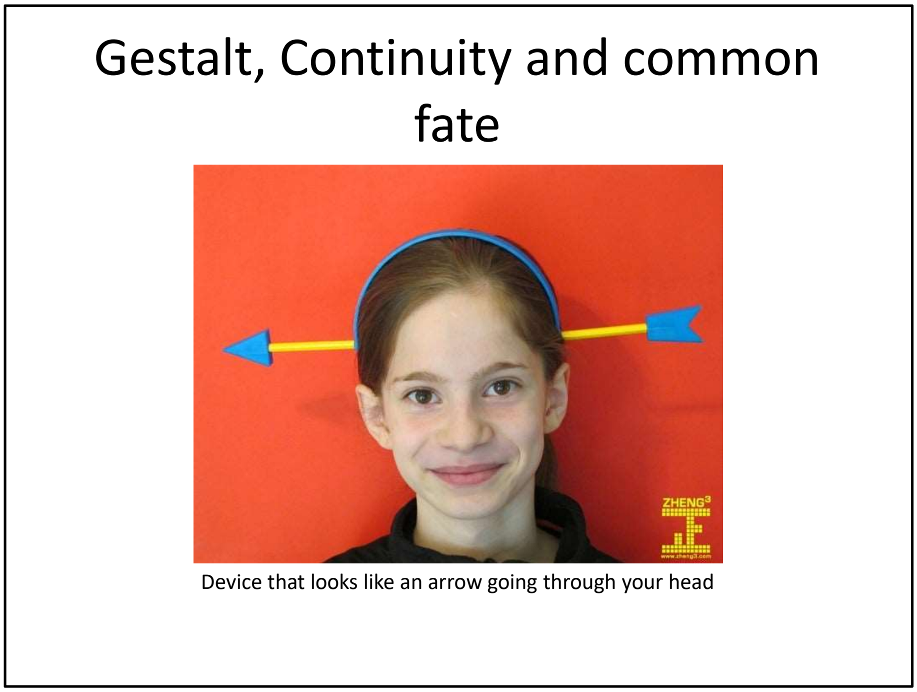


Gestalt, Continuity and common fate

Line graph showing present and future trends

A consistent trend or pattern between two elements of a graph will encourage you to group those elements together.

In this graph there are two solid lines representing the present trend in two (fictional) time series. There are two dashed lines that represent future projected trends. Your eye tends to associate the upward solid line with the upward dashed line rather than the downward dashed line. This is in spite of the fact that the distance is less between the upward solid line and the downward dashed line. There are no color or shape clues to help you. The reason you associate the solid upward trend with the dashed upward trend is that you eye wants to fill in any missing information with the simplest form possible.

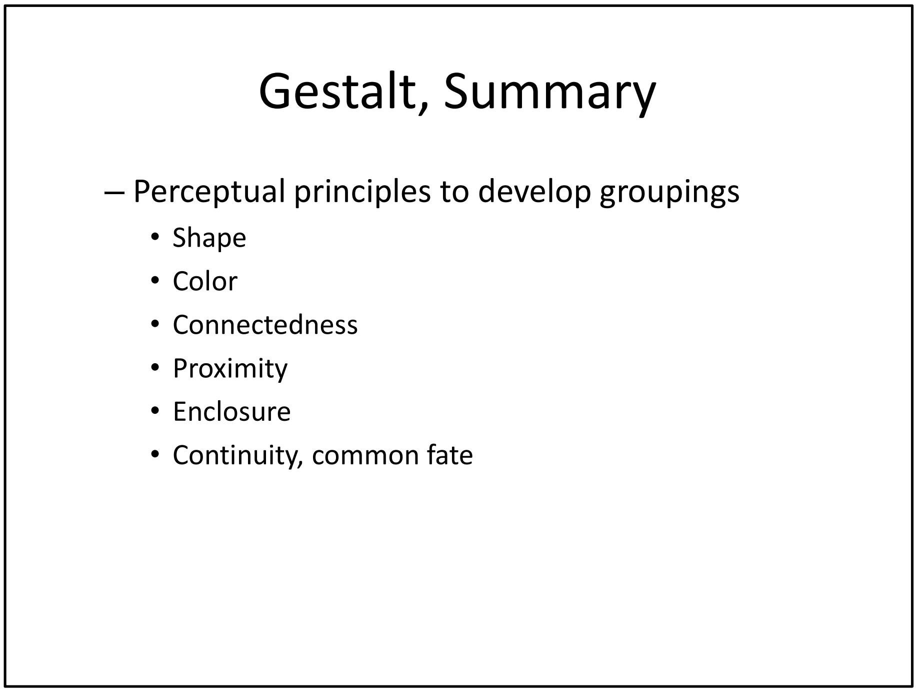


Gestalt, Continuity and common fate

Device that looks like an arrow going through your head

The reason this works is that your brain is trained to think that when part of an object is hidden by another object, you realize that this is one object rather than two.

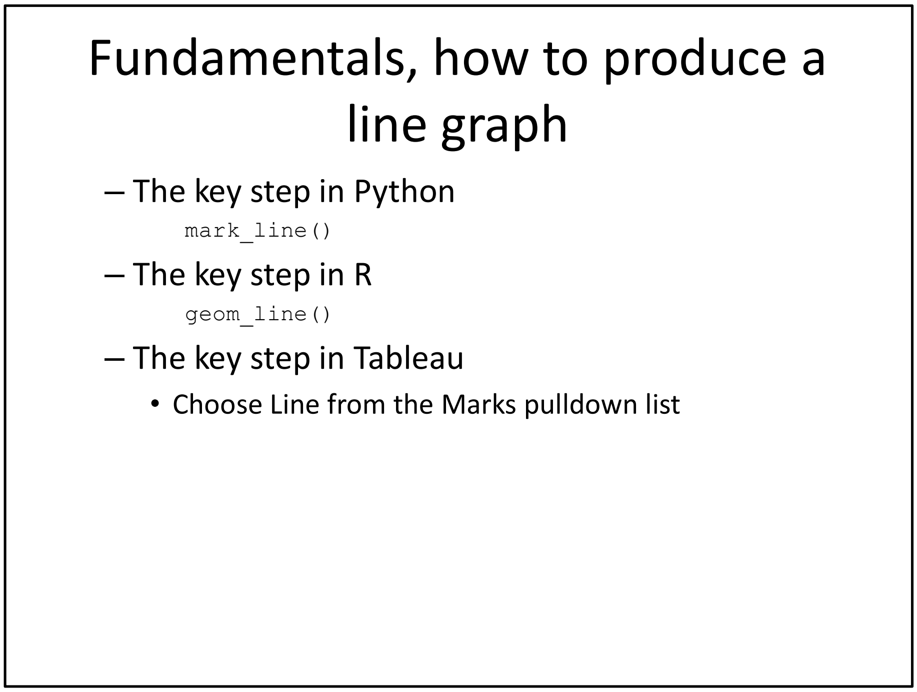
This image of an gag toy looks almost like it is real, but for the band that goes up an over the head.



Gestalt, Summary

- Perceptual principles to develop groupings
  - Shape
  - Color
  - Connectedness
  - Proximity
  - Enclosure
  - Continuity, common fate

Gestalt is a set of perceptual principles that you can use to develop groupings in your visualizations. You tend to perceive groups that have similar shape or color, that are connected, or that are close to one another. You can use enclosure to define groupings. your eye also tends to group items to preserve continuity or items that share a common fate.



Fundamentals, how to produce a line graph

- The key step in Python

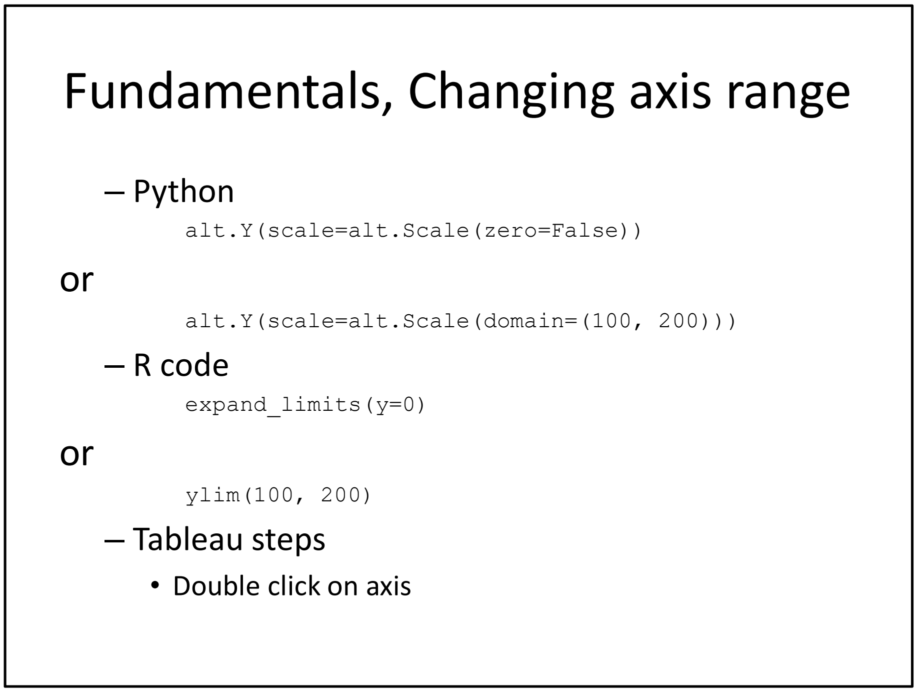
```
mark_line()
```

- The key step in R

```
geom_line()
```

- The key step in Tableau
  - Choose Line from the Marks pulldown list

The basic steps to producing a line graph are fairly easy in each of the systems. In Tableau, you use the mark_line function. In R, you use the geom_line function. In Tableau, if you are lucky, the default graph choice will be a line graph. If not, select Line from the Marks list.



Fundamentals, Changing axis range

- Python

```
alt.Y(scale=alt.Scale(zero=False))
```

or

```
alt.Y(scale=alt.Scale(domain=(100, 200)))
```

- R code

```
expand_limits(y=0)
```

or

```
ylim(100, 200)
```

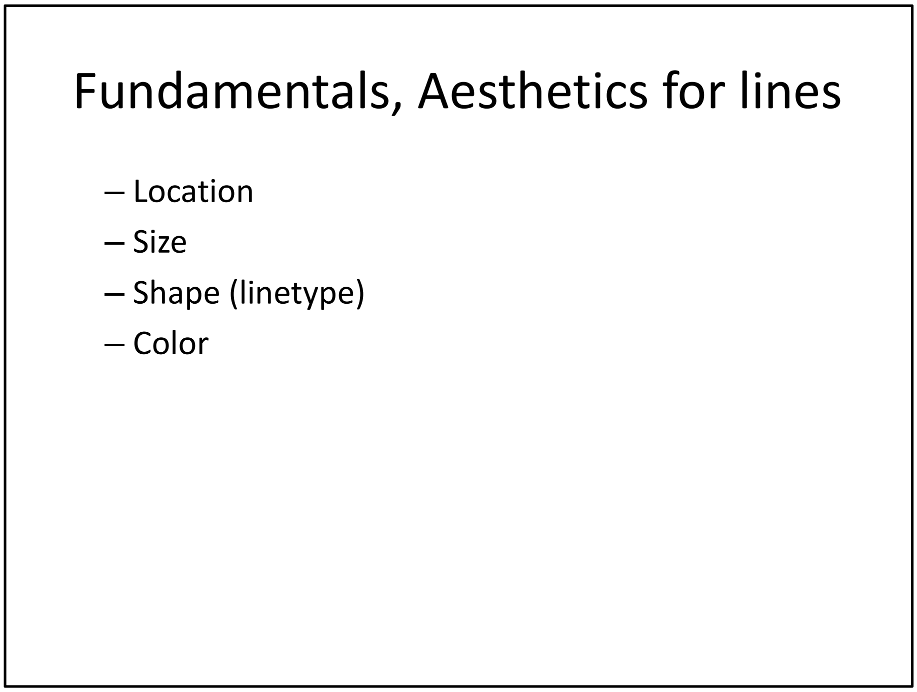
- Tableau steps
  - Double click on axis

This example is the first one where you can see a big difference between the default options. Python wants to include zero on the y-axis, no matter what. It may make sense for a barchart to always include zero, but it is less obvious for a line graph.

To override the default Python option of showing zero on the Y axis, include the zero=False statement in the Scale function. Or you can specify a particular set of limits using the domain option in the Scale function.

R does not always include zero, but if you want, you can force R to include zero with the expand_limits function. Or you can specify a lower and upper limit with the ylim function.

Tableau also does not always include zero. To modify the limits of an axis in Tableau, you double click on it to get a dialog box.

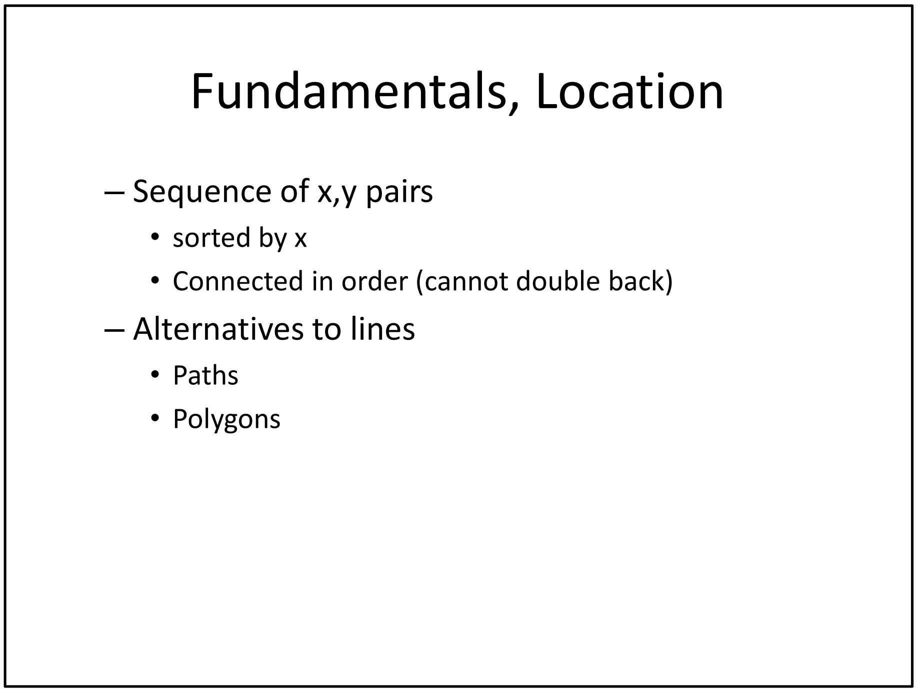


Fundamentals, Aesthetics for lines

- Location
- Size
- Shape (linetype)
- Color

A line could mean a straight line or a curved line, a single line segment, a connected series of line segments or a polygon. It’s a pretty complex thing, but generally a line represents a two dimensional relationship.

You can vary the size, shape, and color of a line. Shape is not what you think it is. It is a linetype, which can be solid, dotted, or dashed. Among the dotted lines, you have a choice of whether the dots are very close or somewhat distant. Among the dashed lines, you have a choice of the size of the dash and the size of the gap between the dashes. You can also mix dots and dashes together.



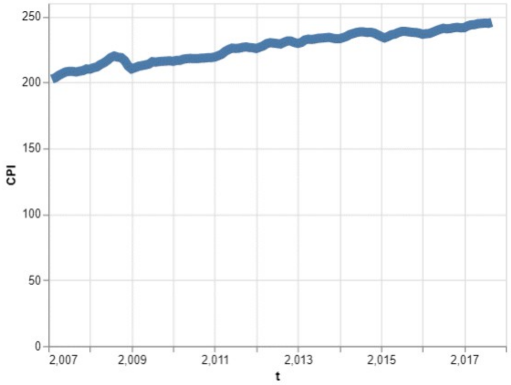
Fundamentals, Location

- Sequence of x,y pairs
  - sorted by x
  - Connected in order (cannot double back)
- Alternatives to lines
  - Paths
  - Polygons

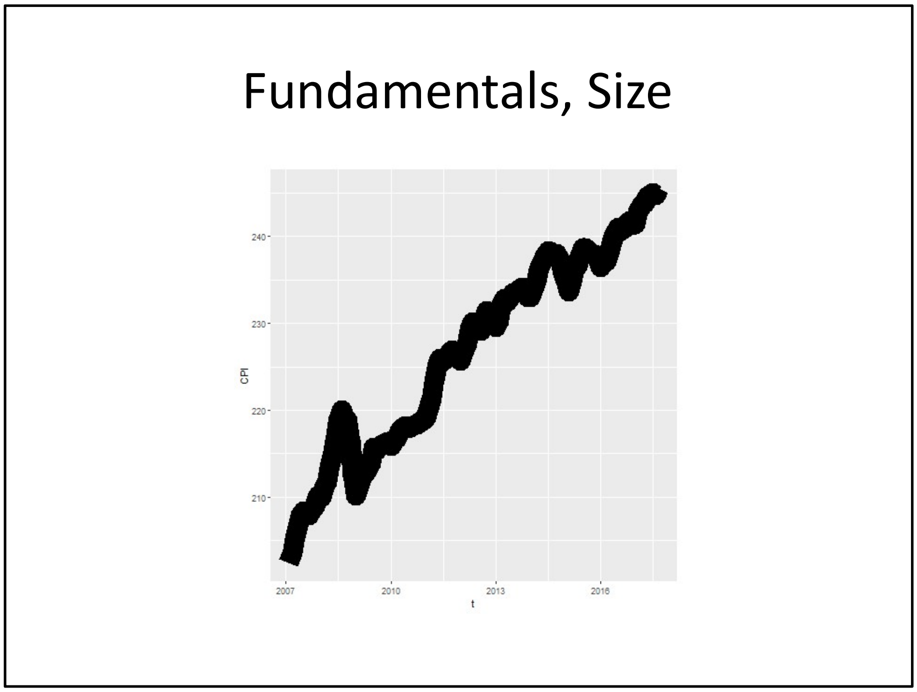
A line could mean a straight line or a curved line. You get a curve by connecting small straight line segments. If the segments are small enough, you don’t even notice. One requirement of a line is that the X values are sorted from low to high and are connected in order. This means that the line cannot double back on itself.

An alternative to a line is a path, which is allowed to meander in any direction including reversing course in the X direction one or more times. A polygon is a path that does not intersect itself and which returns at the end to the point at which it started. Paths and polygons will not be a major focus in this presentation, but the principles that apply to lines generally apply equally well to paths and polygons.

Fundamentals, Size

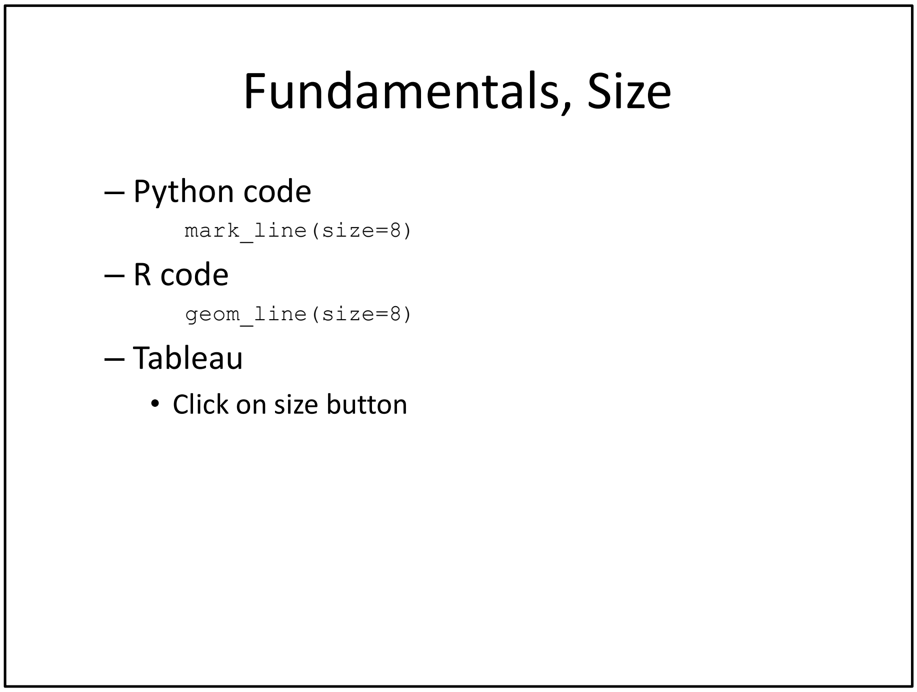


Line graph with thick line



Fundamentals, Size

Lines can change size, but some systems refer to this as line width rather than size. You might use a thick line at times to create a greater emphasis, but it does tend to obscure small sudden changes in this current graph.



Fundamentals, Size

– Python code

```
mark_line(size=8)
```

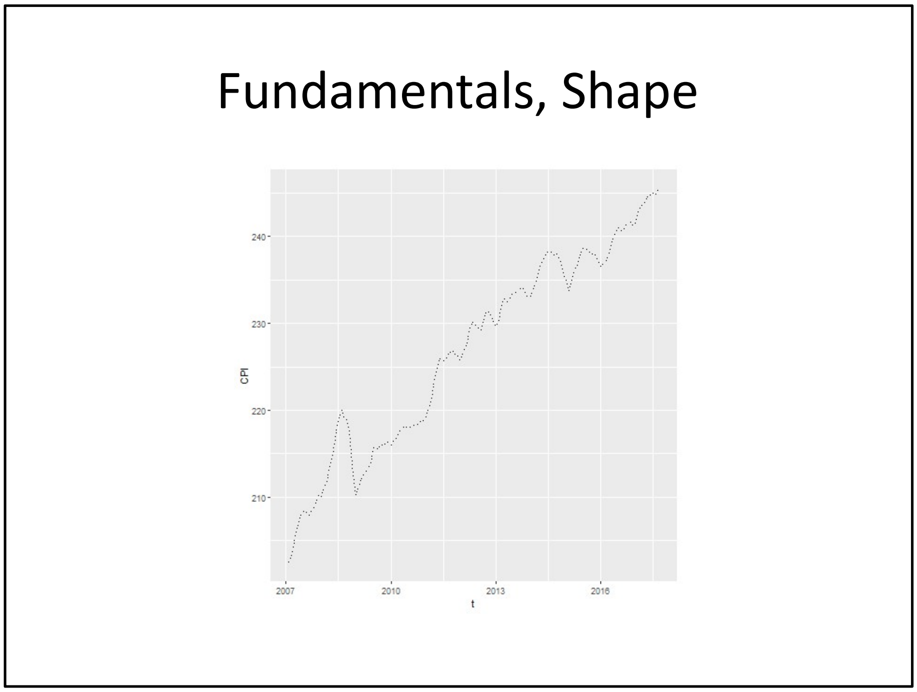
– R code

```
geom_line(size=8)
```

– Tableau

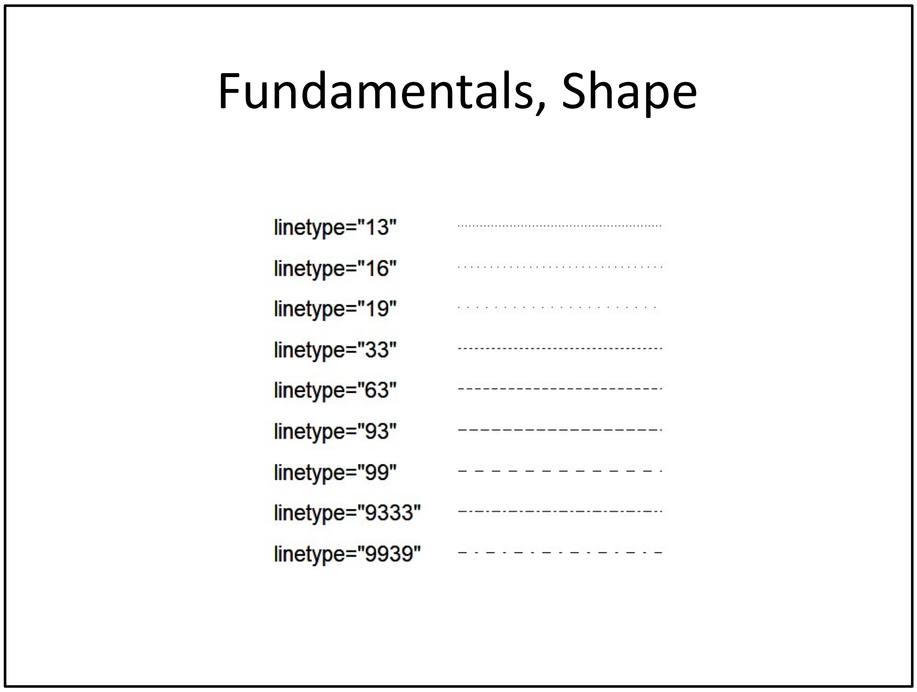
- Click on size button

To change the line thickeness in R, use the size= option inside the geom_line function.



Fundamentals, Shape

The shape of a line is not whether it is curved or straight. The shape of a line is whether it is represented using solid, dashed, or dotted lines. Most visualizations call this a line type rather than a line shape.



Fundamentals, Shape

linetype="13"

linetype="16"

linetype="19"

linetype="33"

linetype="63"

linetype="93"

linetype="99"

linetype="9333"

linetype="9939"

There is a code for dotted and dashed line patterns that works pretty much the same way for both Python and R. You specify a sequence of two (sometimes four) numbers. The odd numbered values (first, third, etc.) represent an "on" number of pixels which specifies the length of the dash. The even numbered values (second, fourth, etc.) represent an "off" number of pixels. This is the gap or space between the marks.

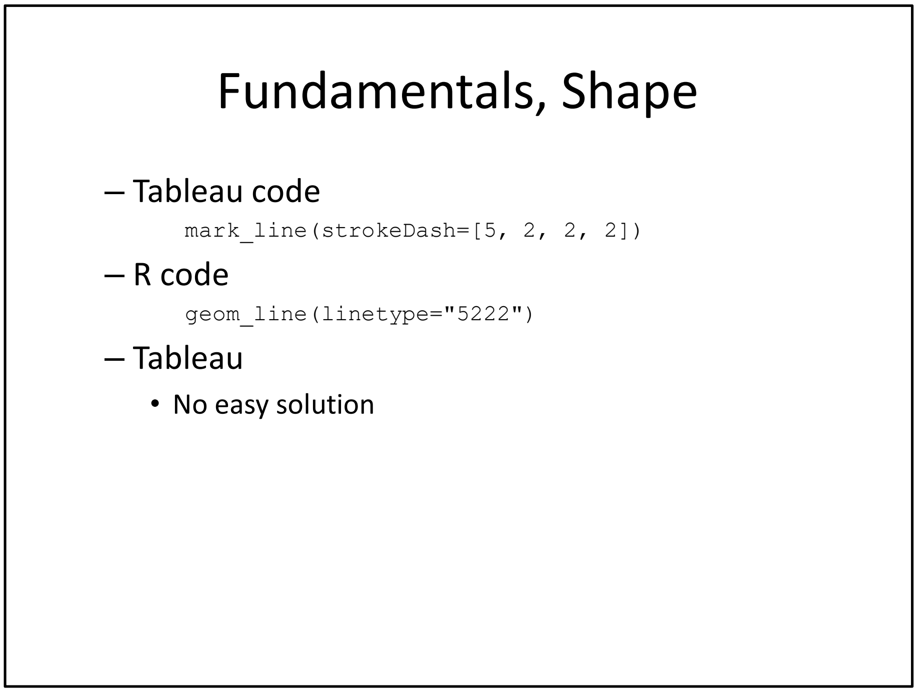
The code "13" is a tight dotted line. One pixel on represents the dot and three pixels off represents the tight space between the dots. Note that anything tighter that 3 for a space is hard to distinguish from a solid line.

The code "16" puts a moderate amount of space between the dots and "19" puts a large space between the dots.

The codes "33" and "63" and "93" create short medium and long dashes with just a small space between the dashes. The code "99" creates long dashes with large spaces between them.

A set of four numbers usually produces an alternating pattern. The pattern "9333" produces a dash-dot pattern. The "9" produces a long dash. The first "3" produces a small space. The second "3" produces a short dash and the last "3" produces a short space. The pattern “9939” produces a similar dash-dot pattern, but the “9”’s in the second a fourth position widens the spaces between the dashes and dots quite a bit.

There are an infinite number of possibilities here.



Fundamentals, Shape

– Tableau code

```
mark_line(strokeDash=[5, 2, 2, 2])
```

– R code

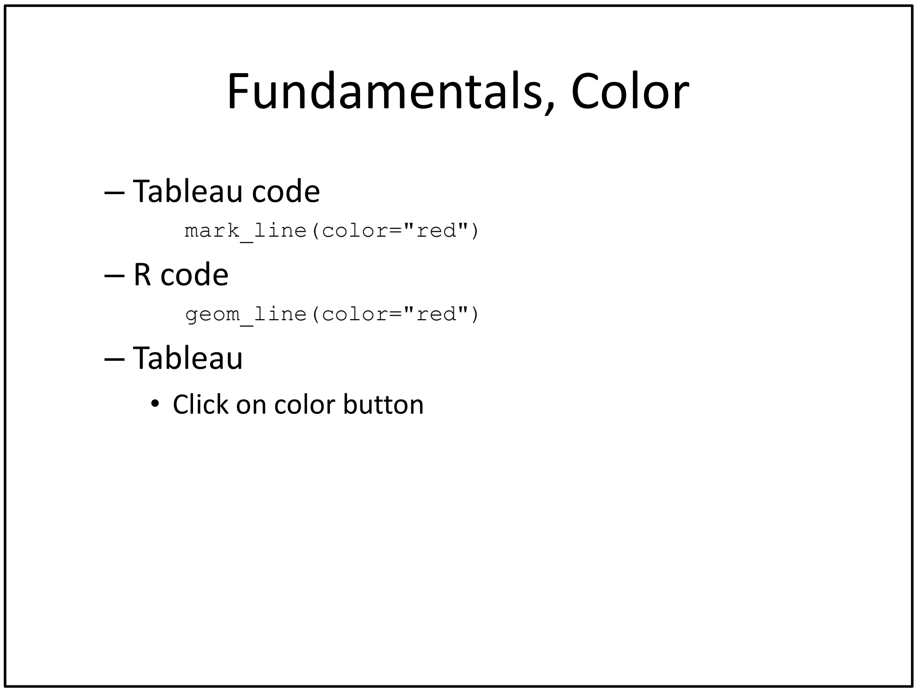
```
geom_line(linetype="5222")
```

– Tableau

- No easy solution

In Python, set the strokeDash argument to create various dotted and dashed lines. In R, use the linetype argument.

Tableau allows you a lot of freedom in choosing between dotted and dashed lines when you have two or more lines on a single graph, but it is surprisingly hard to deviate from the formal default of a solid line when you only have a single line on your graph. This is annoying, but quite honestly, there is little reason to deviate from a solid line when you only have one line on your graph.



Fundamentals, Color

- Tableau code

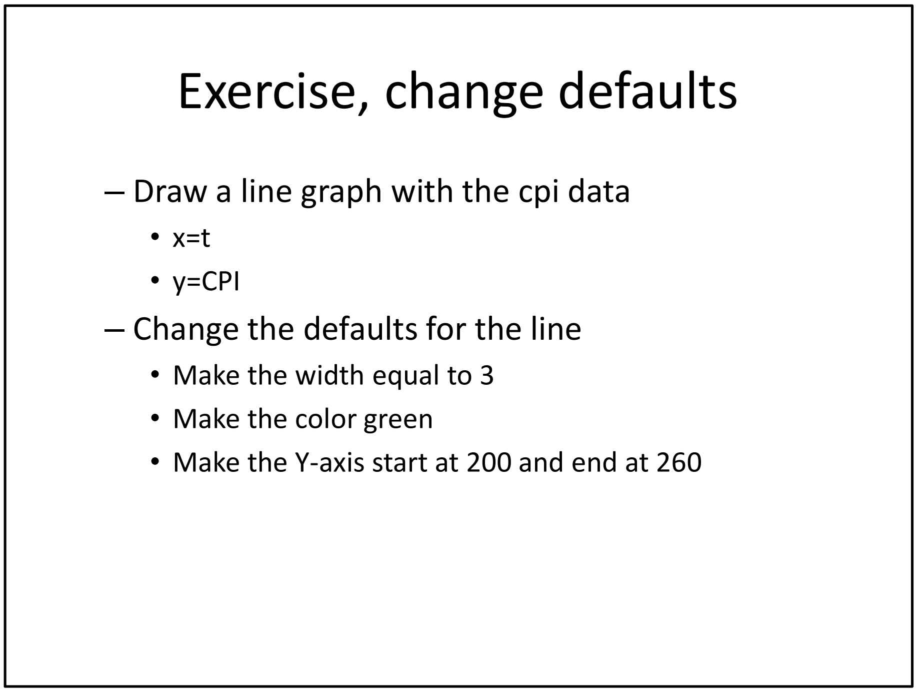
```
mark_line(color="red")
```

- R code

```
geom_line(color="red")
```

- Tableau
  - Click on color button

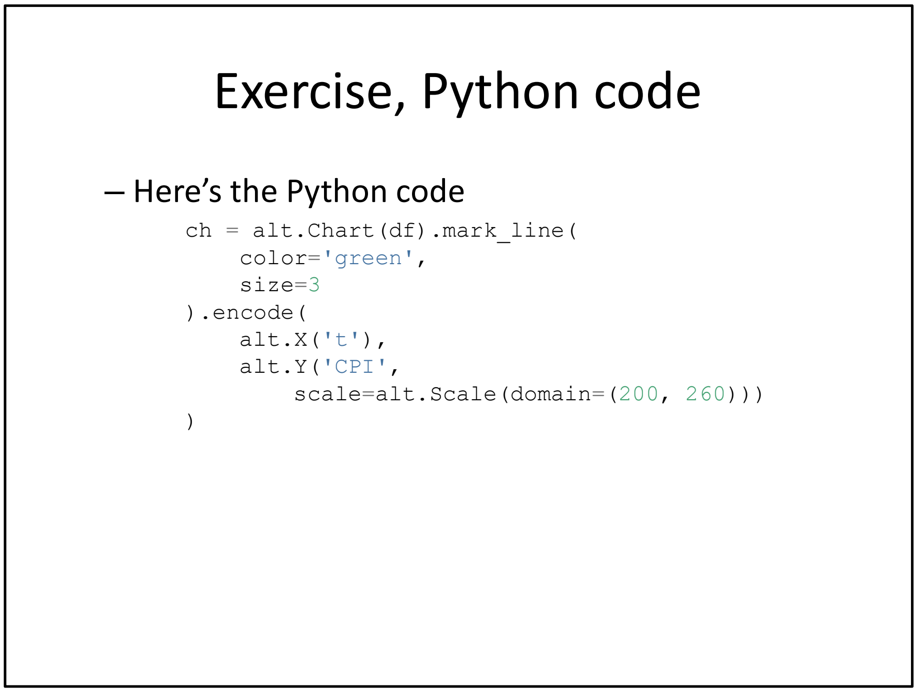
To change the default color of a single line, use the color argument in mark_line (Python) or geom_line (R). In Tableau, you change the default color by clicking on the color button.



Exercise, change defaults

- Draw a line graph with the cpi data
  - x=t
  - y=CPI
- Change the defaults for the line
  - Make the width equal to 3
  - Make the color green
  - Make the Y-axis start at 200 and end at 260

Take the CPI line graph and modify some of the default options. Make the line green and increase the thickness to three pixels. Modify the y-axis to start at 200 and end at 260.

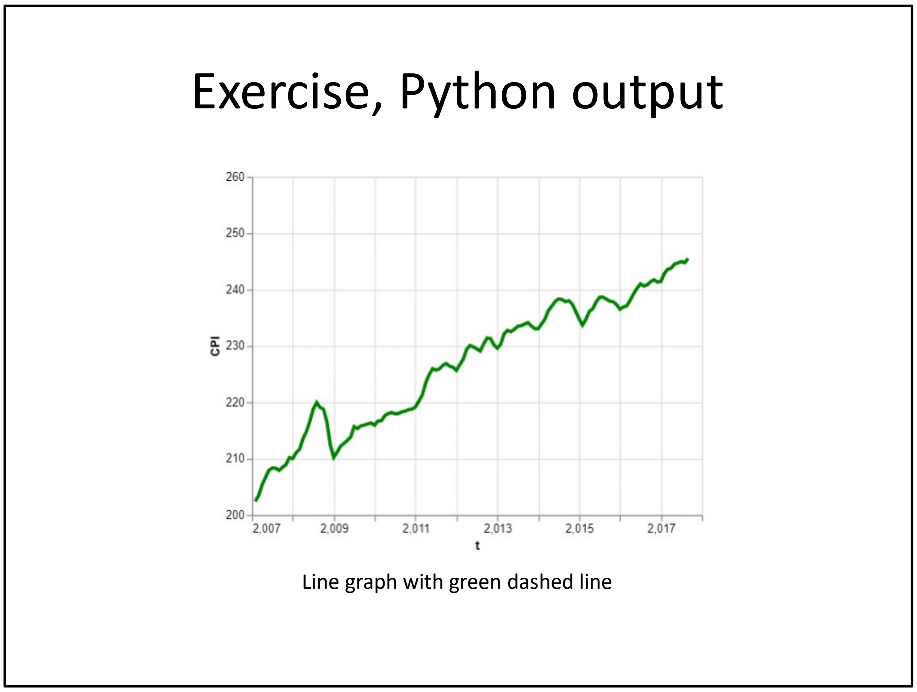


Exercise, Python code

– Here’s the Python code

```
ch = alt.Chart(df).mark_line(
    color='green',
    size=3
).encode(
    alt.X('t'),
    alt.Y('CPI',
        scale=alt.Scale(domain=(200, 260)))
)
```

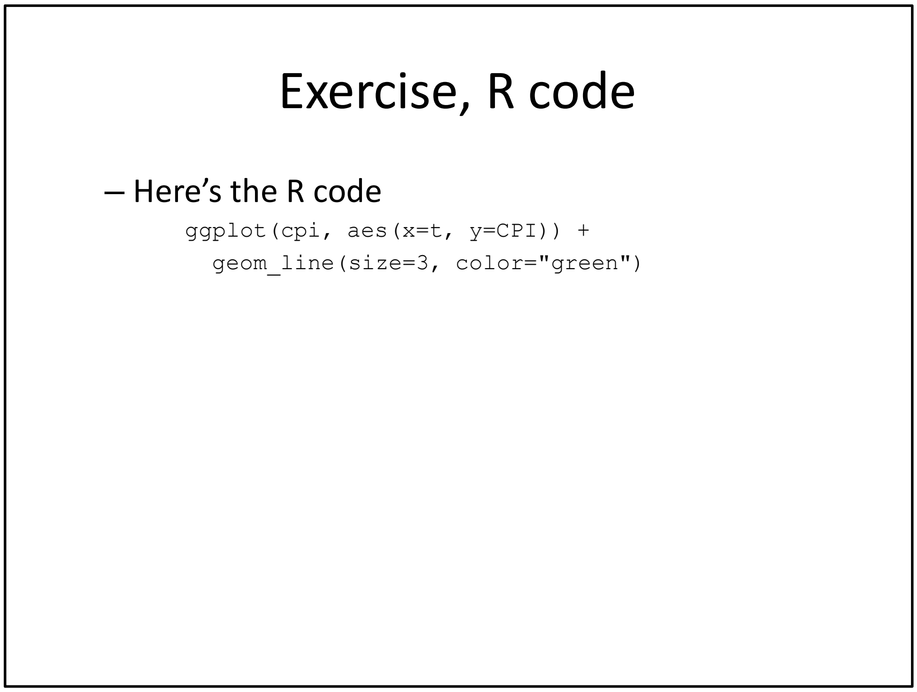
HEre is the Python code. Notice that the color and width are changed inside the mark_line function.



Exercise, Python output

Line graph with green dashed line

Here is the Python output.

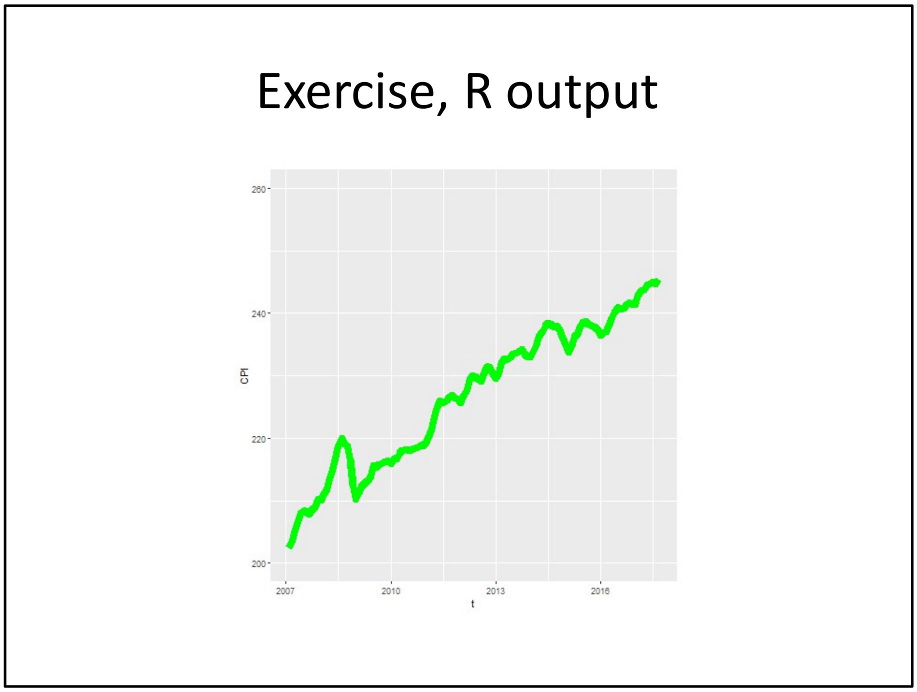


Exercise, R code

– Here’s the R code

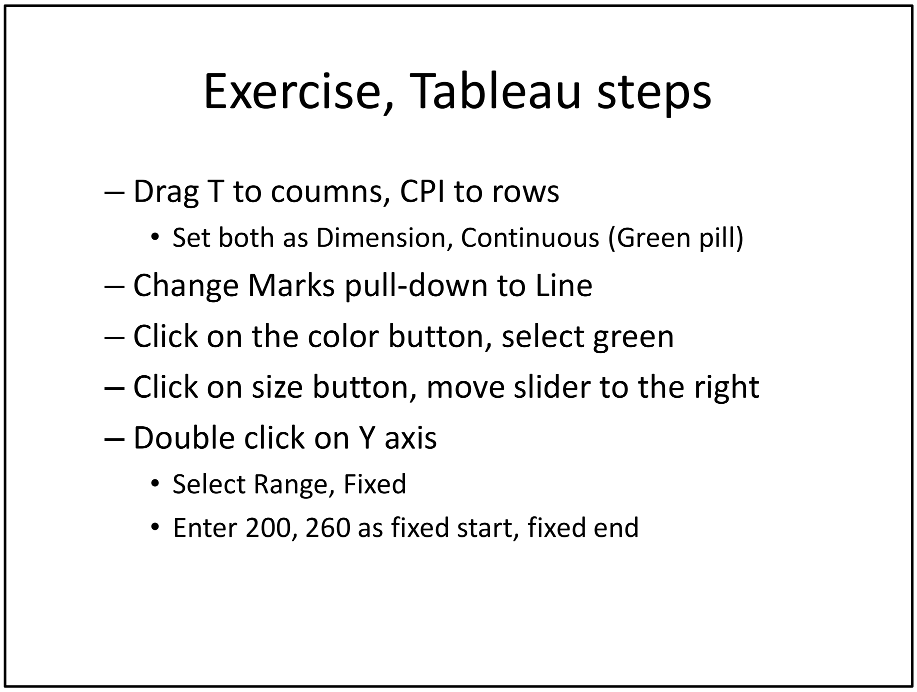
```
ggplot(cpi, aes(x=t, y=CPI)) +
  geom_line(size=3, color="green")
```

Here is the R code. Notice that you change the size and color inside the geom_line function.



Exercise, R output

Here is the R output.

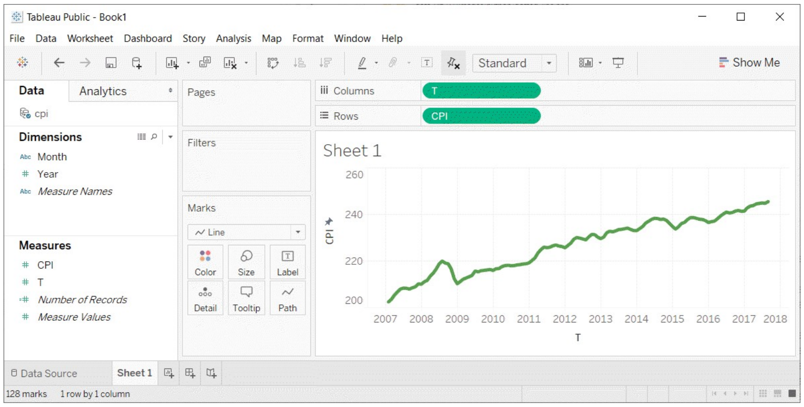


Exercise, Tableau steps

- Drag T to coumns, CPI to rows
  - Set both as Dimension, Continuous (Green pill)
- Change Marks pull-down to Line
- Click on the color button, select green
- Click on size button, move slider to the right
- Double click on Y axis
  - Select Range, Fixed
  - Enter 200, 260 as fixed start, fixed end

Here are the Tableau steps.

Exercise, Tableau output

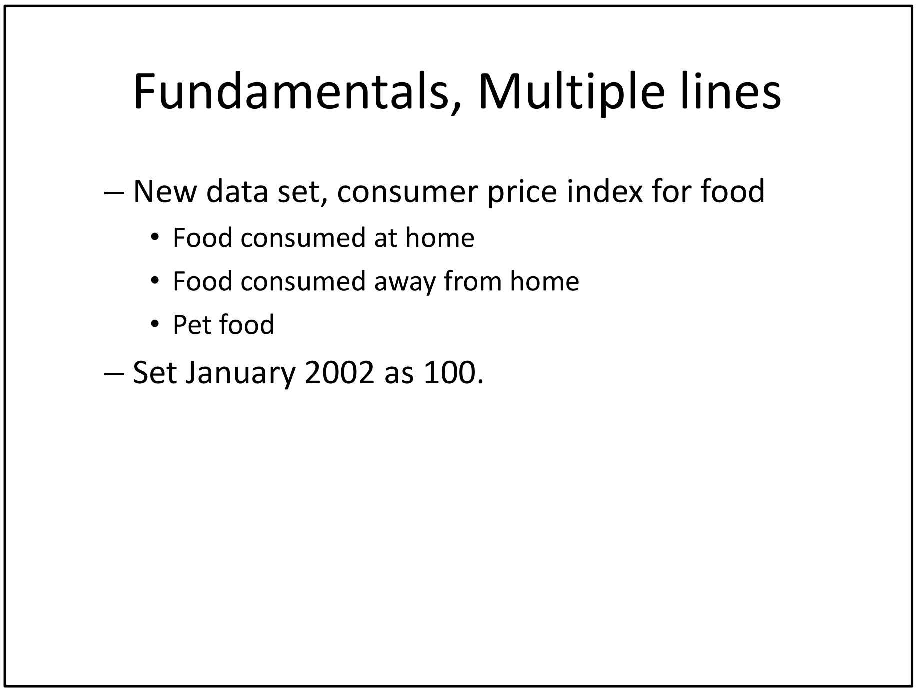


Columns: T

Rows: CPI

Line graph with thick green line

Here is the Tableau output.



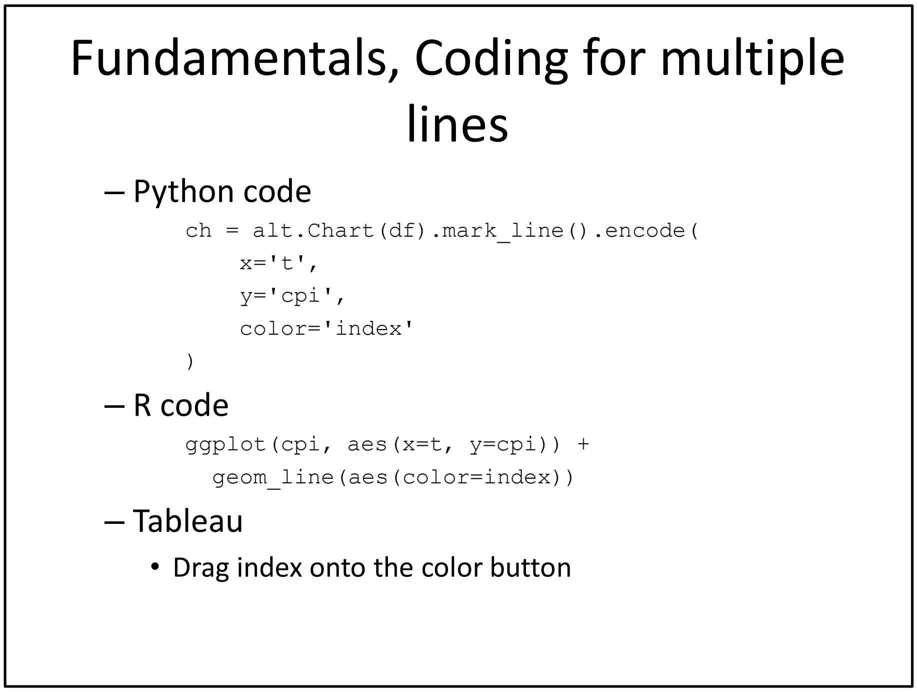
Fundamentals, Multiple lines

– New data set, consumer price index for food
  - Food consumed at home
  - Food consumed away from home
  - Pet food

– Set January 2002 as 100.

There is a similar data set that I want to switch to. It provides the consumer price index for food, separated into three series. The first series is a price index for food consumed at home. That includes things you buy at the store like Pop Tarts, assuming that you don’t try to eat them in the store. There is a second series for food consumed away from home, such as the 1300 calorie Monster Thickburger that you got at Hardee’s. There is a third series for pet food.

There’s a bad joke about this. I took my dog shopping and he was unhappy with how much the price of Alpo has risen. It was 99 cents. That may not seem like a lot to you, but you have to realize that this is almost seven dollars in dog money.



Fundamentals, Coding for multiple lines

- Python code

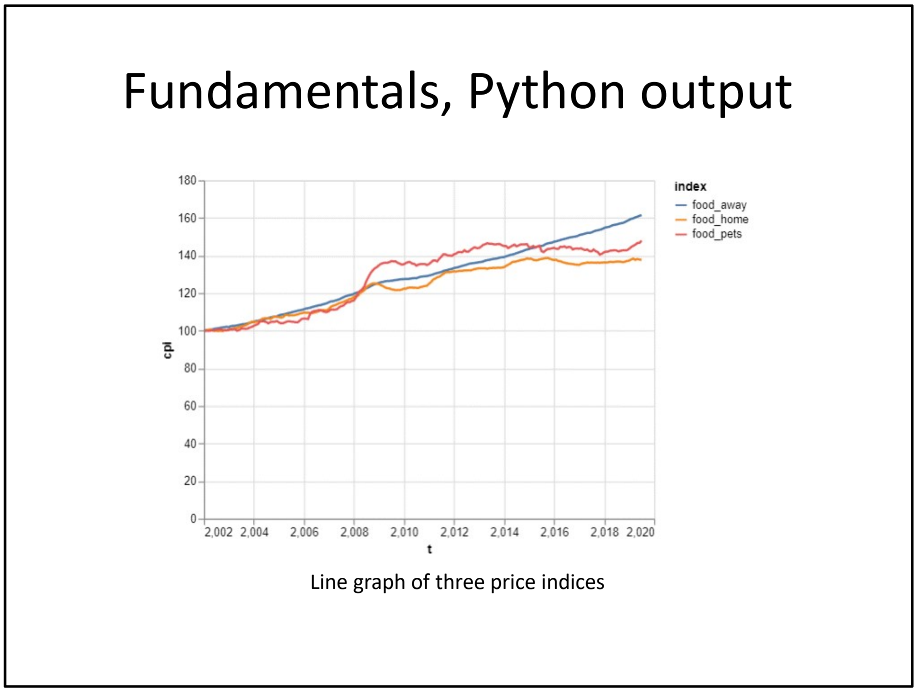
```
ch = alt.Chart(df).mark_line().encode(
    x='t',
    y='cpi',
    color='index'
)
```

- R code

```
ggplot(cpi, aes(x=t, y=cpi)) +
  geom_line(aes(color=index))
```

- Tableau
  - Drag index onto the color button

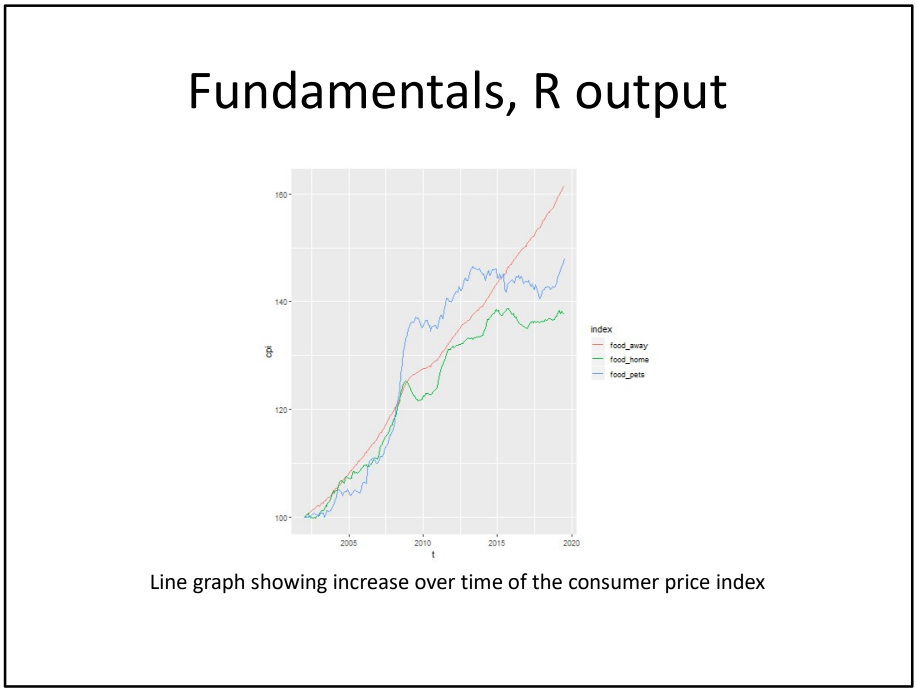
When you want to have a different line for each consumer price index series, you specify a different color for each line. Notice how this goes inside the encode function in Python and inside the aes function in R because you are encoding a third variable, index, to an aesthetic.



Fundamentals, Python output

Line graph of three price indices

This is what the Python output looks like.

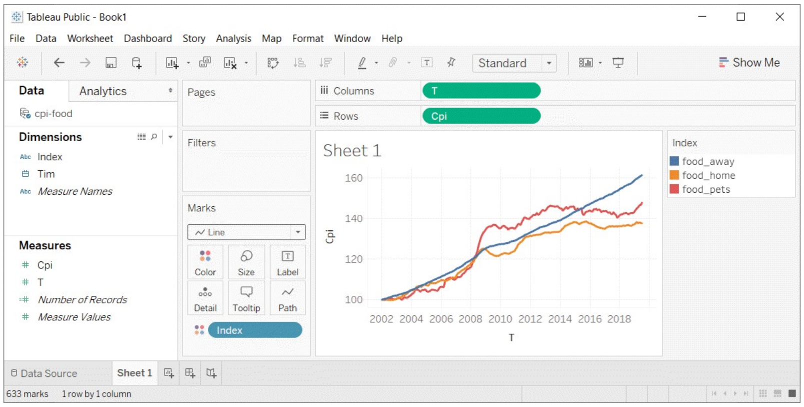


Fundamentals, R output

Line graph showing increase over time of the consumer price index

Here is what the graph looks like in R.

Fundamentals, Python output

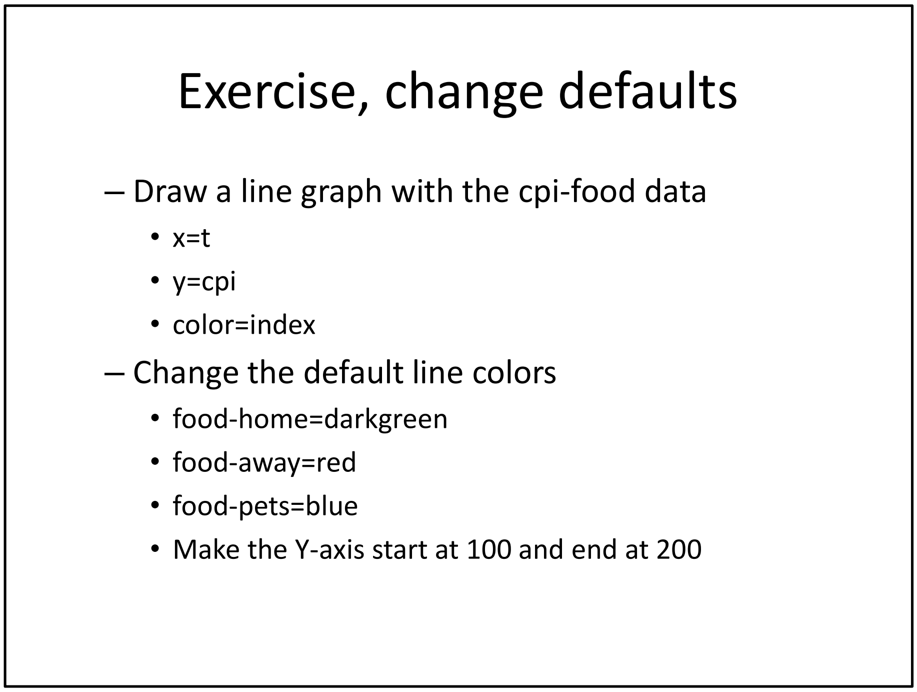


Columns: T

Rows: Cpi

Line graph in Tableau

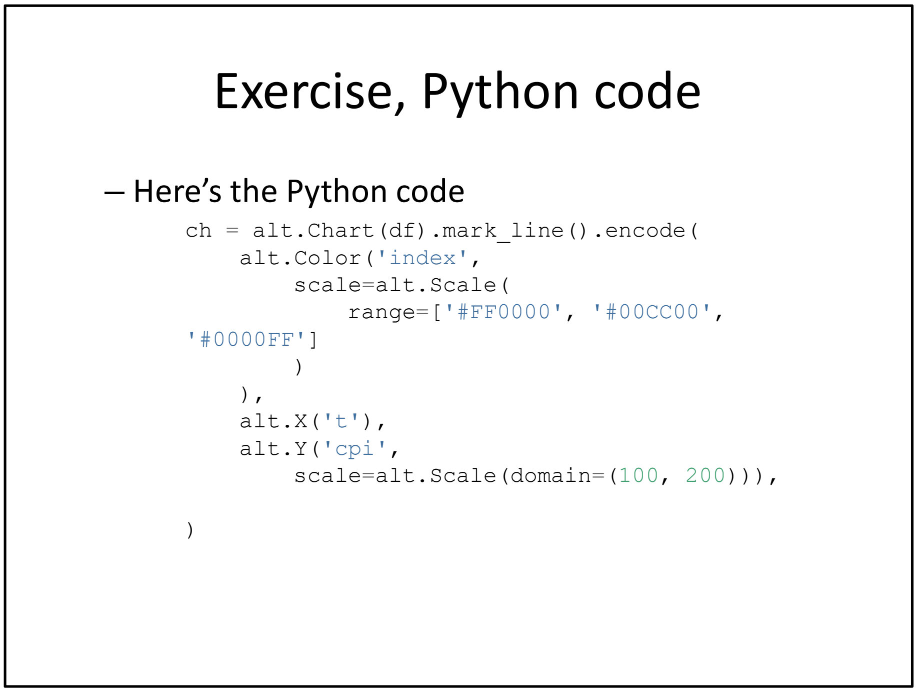
This is what the Tableau graph looks like.



Exercise, change defaults

- Draw a line graph with the cpi-food data
  - x=t
  - y=cpi
  - color=index
- Change the default line colors
  - food-home=darkgreen
  - food-away=red
  - food-pets=blue
  - Make the Y-axis start at 100 and end at 200

Take the cpi-food linegraph and modify some of the default options. Highlight the food-home in darkgreen, food-away in red, and food-pets in blue. Modify the y-axis to start at 100 and end at 200.



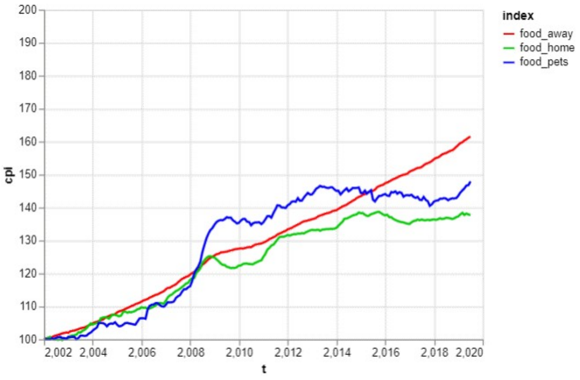
Exercise, Python code

– Here’s the Python code

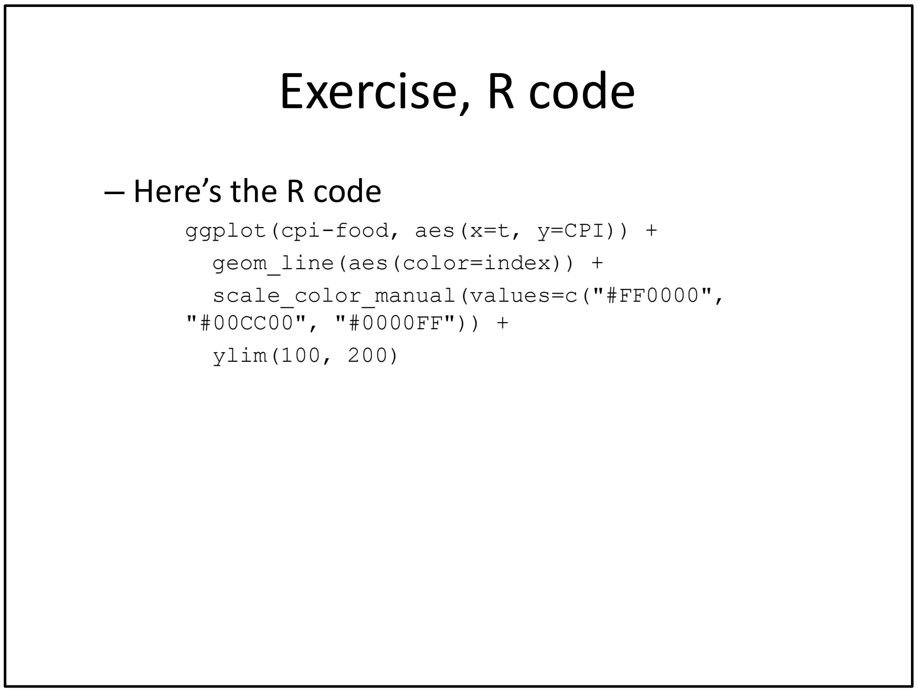
```
ch = alt.Chart(df).mark_line().encode(
    alt.Color('index',
        scale=alt.Scale(
            range=['#FF0000', '#00CC00',
'#0000FF']
        )
    ),
    alt.X('t'),
    alt.Y('cpi',
        scale=alt.Scale(domain=(100, 200))),

)
```

HEre is the Python code. Notice that the color and width are changed inside the mark_line function.

Exercise, Python output



Line graph with thick green line

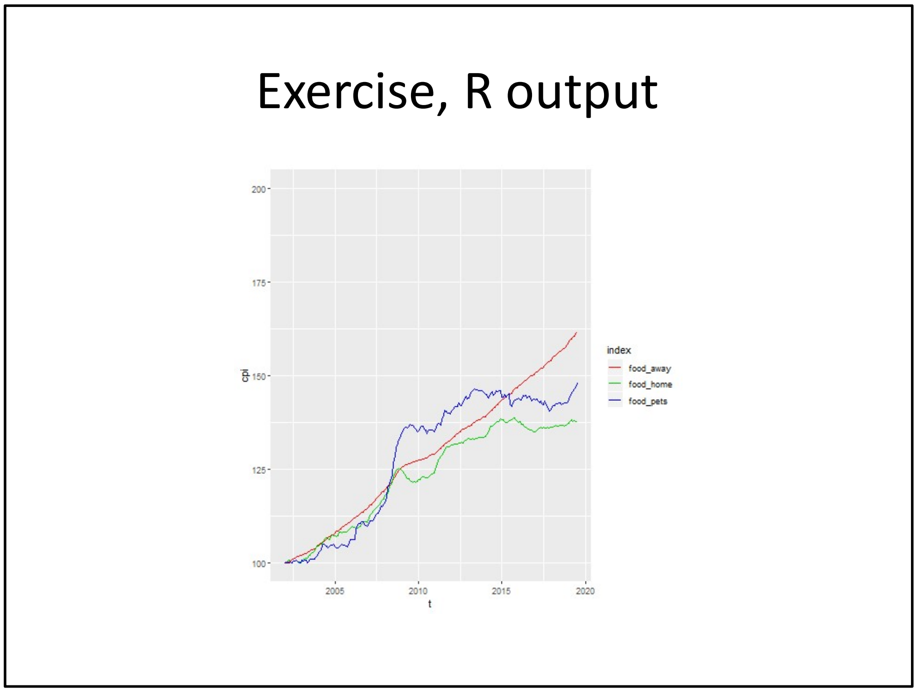


Exercise, R code

– Here’s the R code

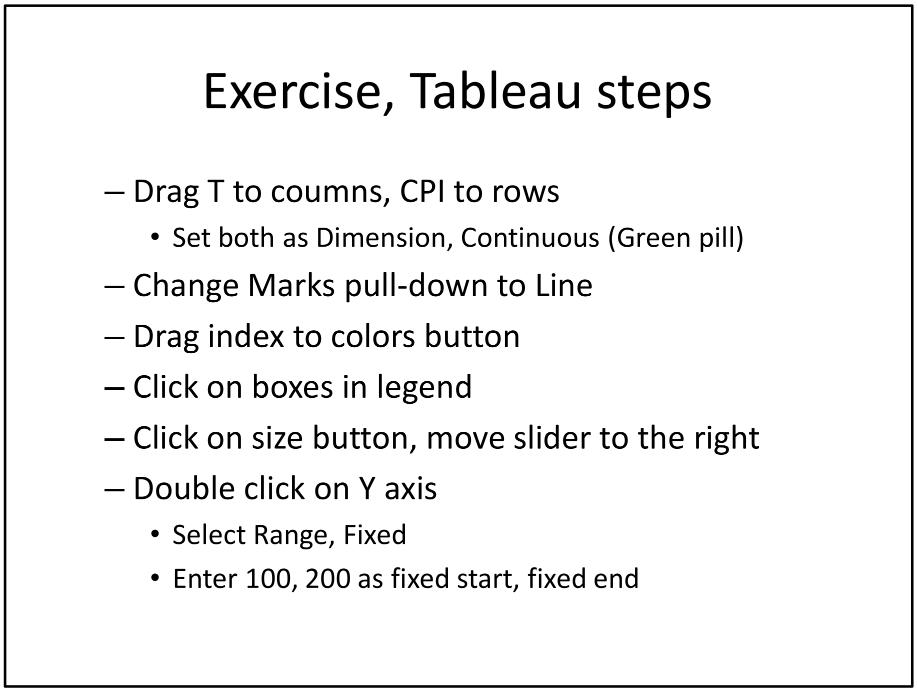
```
ggplot(cpi-food, aes(x=t, y=CPI)) +
  geom_line(aes(color=index)) +
  scale_color_manual(values=c("#FF0000",
"#00CC00", "#0000FF")) +
  ylim(100, 200)
```

Here is the R code. Notice that you change the size and color inside the geom_line function.



Exercise, R output

Here is the R output.

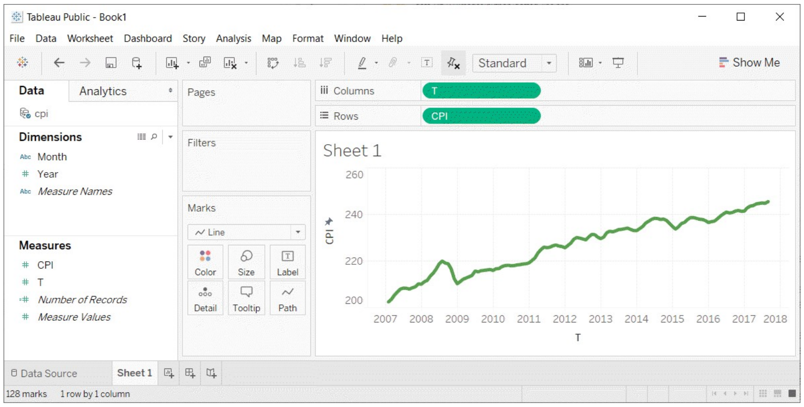


Exercise, Tableau steps

- – Drag T to coumns, CPI to rows
  - Set both as Dimension, Continuous (Green pill)
- – Change Marks pull-down to Line
- – Drag index to colors button
- – Click on boxes in legend
- – Click on size button, move slider to the right
- – Double click on Y axis
  - Select Range, Fixed
  - Enter 100, 200 as fixed start, fixed end

Here are the Tableau steps.

Exercise, Tableau output



Columns: T

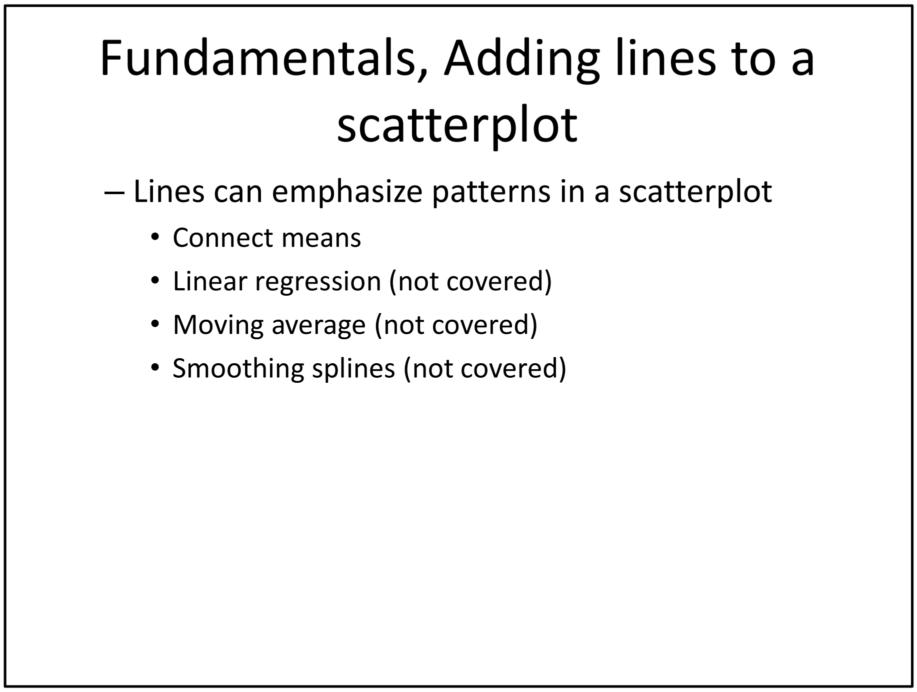
Rows: CPI

Marks: Line

128 marks 1 row by 1 column

Line graph with thick green line

Here is the Tableau output.



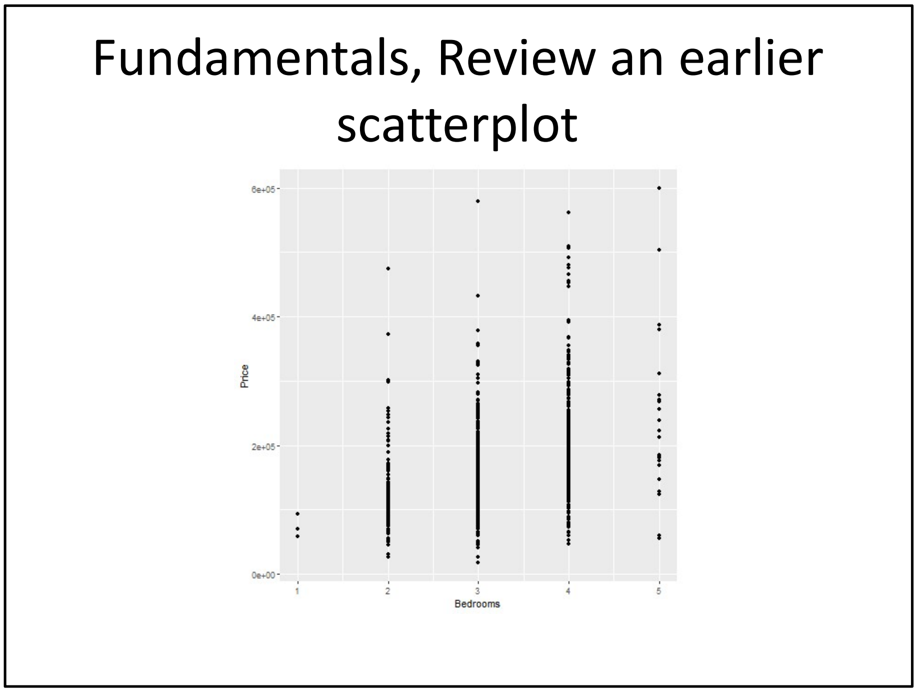
Fundamentals, Adding lines to a scatterplot

– Lines can emphasize patterns in a scatterplot
  - Connect means
  - Linear regression (not covered)
  - Moving average (not covered)
  - Smoothing splines (not covered)

If you have a lot of data, a line can sometimes emphasize patterns or trends that you might miss with just a scatterplot.

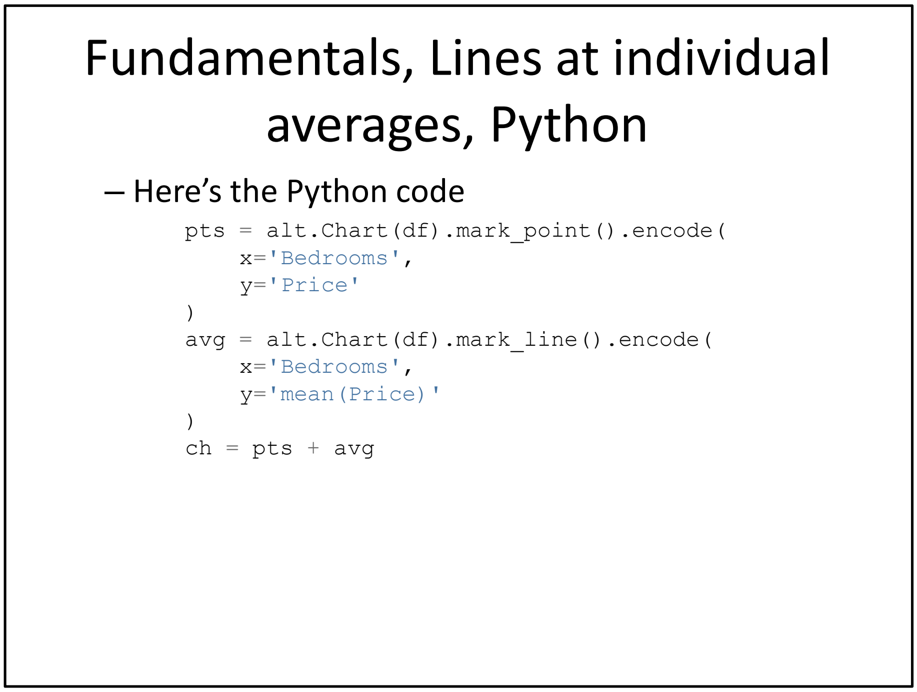
You can do this by connecting lines between individual point means, a linear regression trend line, moving averages, or smoothing splines.

We will only cover the first approach connecting means. The other approaches are quite good and definitely worth talking about. Unfortunately, these alternative approaches are implemented inconsistently across the different programs.



Fundamentals, Review an earlier scatterplot

Here’s the housing data set that we looked at earlier. This scatterplot shows the relationship between number of bedrooms and price.

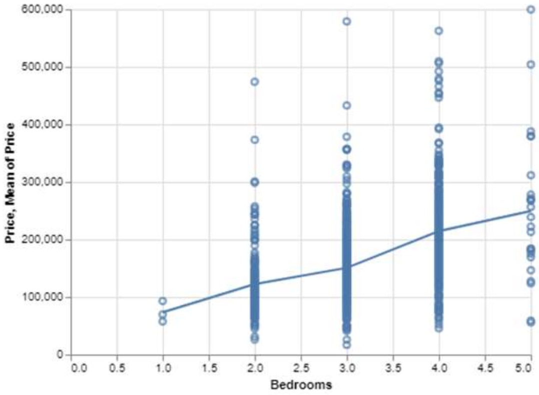


Fundamentals, Lines at individual averages, Python

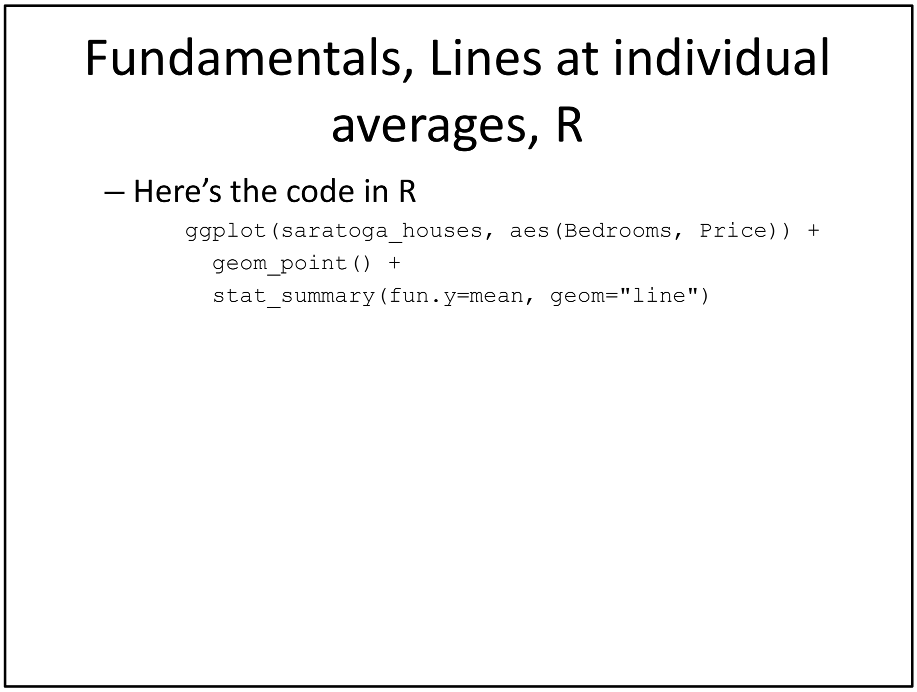
– Here’s the Python code

```
pts = alt.Chart(df).mark_point().encode(
    x='Bedrooms',
    y='Price'
)
avg = alt.Chart(df).mark_line().encode(
    x='Bedrooms',
    y='mean(Price)'
)
ch = pts + avg
```

Python graph



Python graph with average summary line

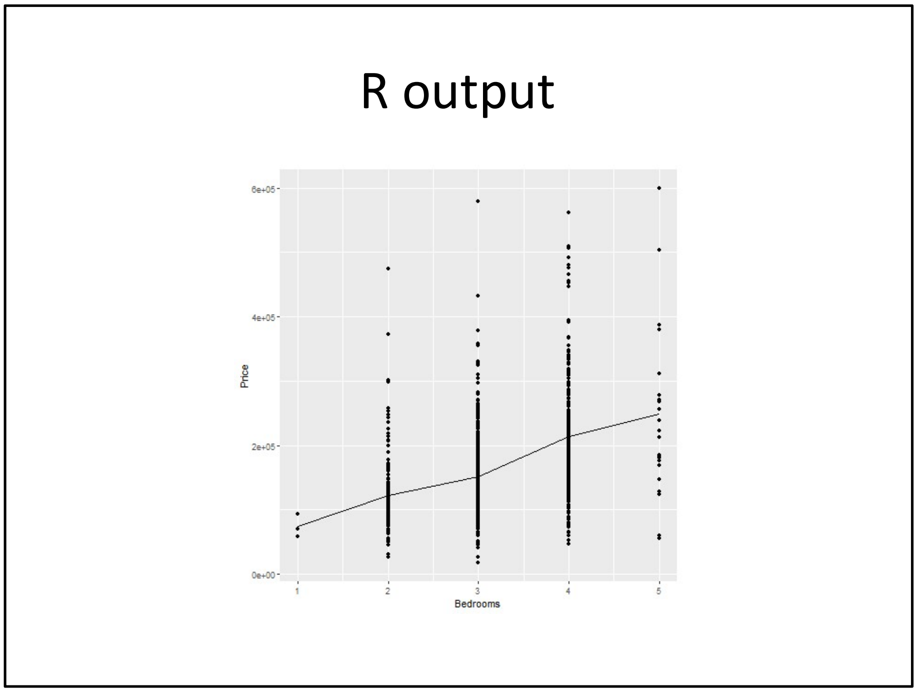


Fundamentals, Lines at individual averages, R

– Here's the code in R

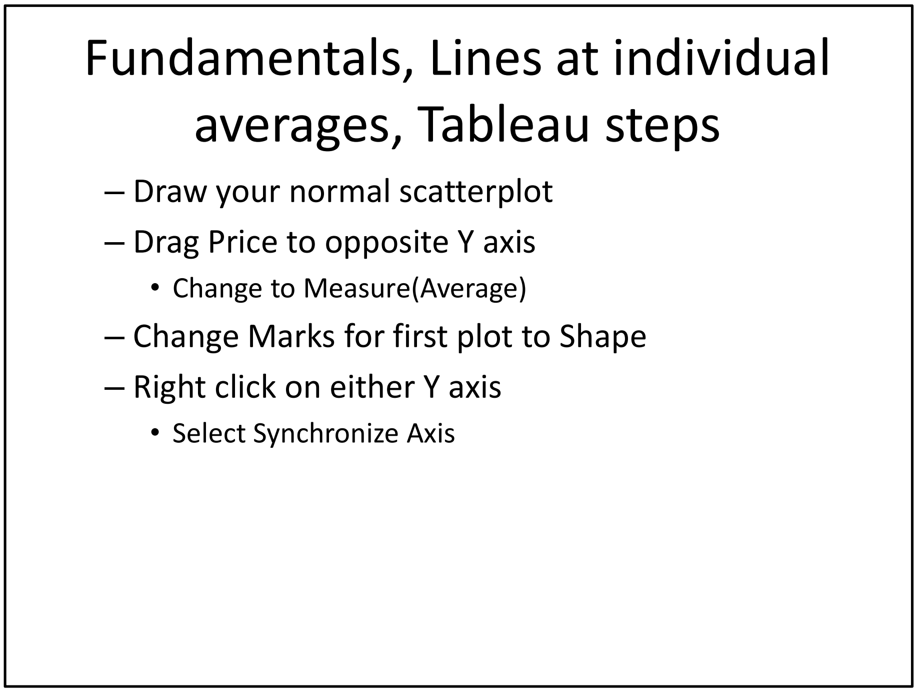
```
ggplot(saratoga_houses, aes(Bedrooms, Price)) +
  geom_point() +
  stat_summary(fun.y=mean, geom="line")
```

Here is the R code. You need the stat_summary function to create averages for y at each discrete value of X.



R output

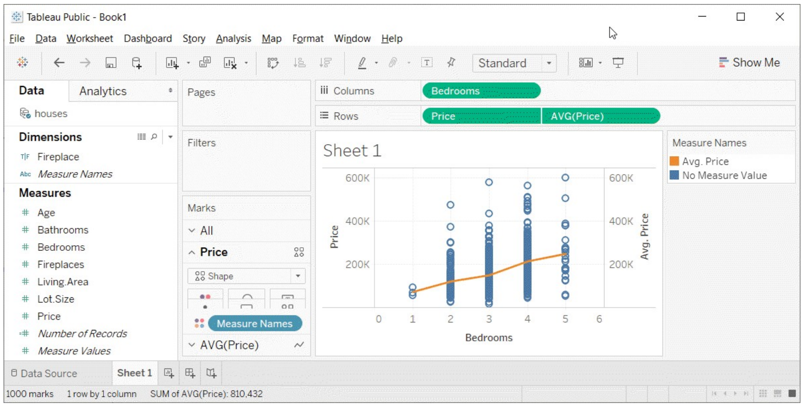
There is a clear and consistnet trend in the average price. As the number of bedrooms increase, the average price increases.



Fundamentals, Lines at individual averages, Tableau steps

- – Draw your normal scatterplot
- – Drag Price to opposite Y axis
  - Change to Measure(Average)
- – Change Marks for first plot to Shape
- – Right click on either Y axis
  - Select Synchronize Axis

Fundamentals, Lines at individual averages, Tableau plot

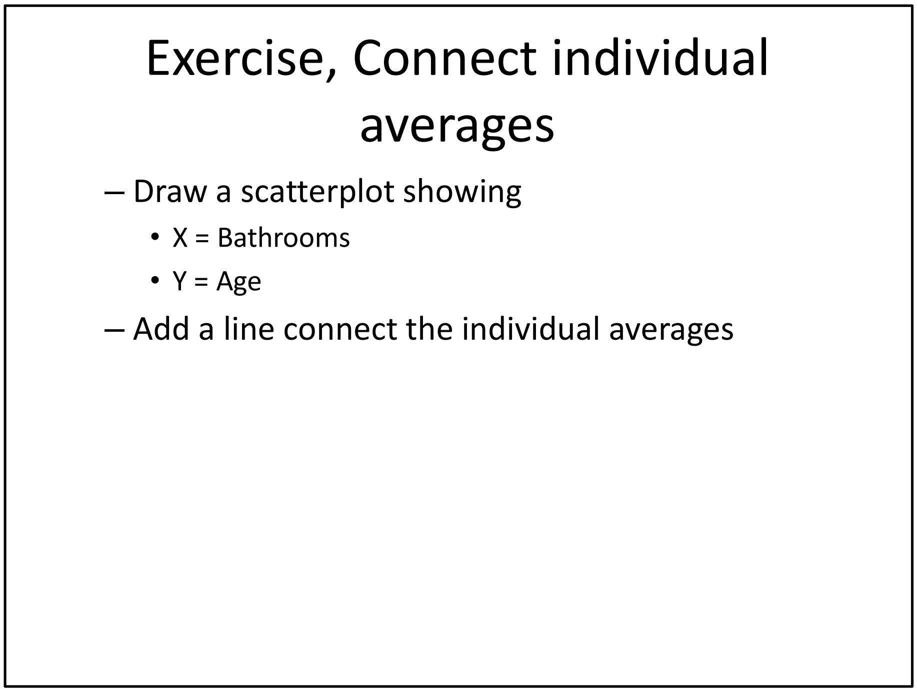


Columns: Bedrooms

Rows: Price, AVG(Price)

1000 marks 1 row by 1 column SUM of AVG(Price): 810,432

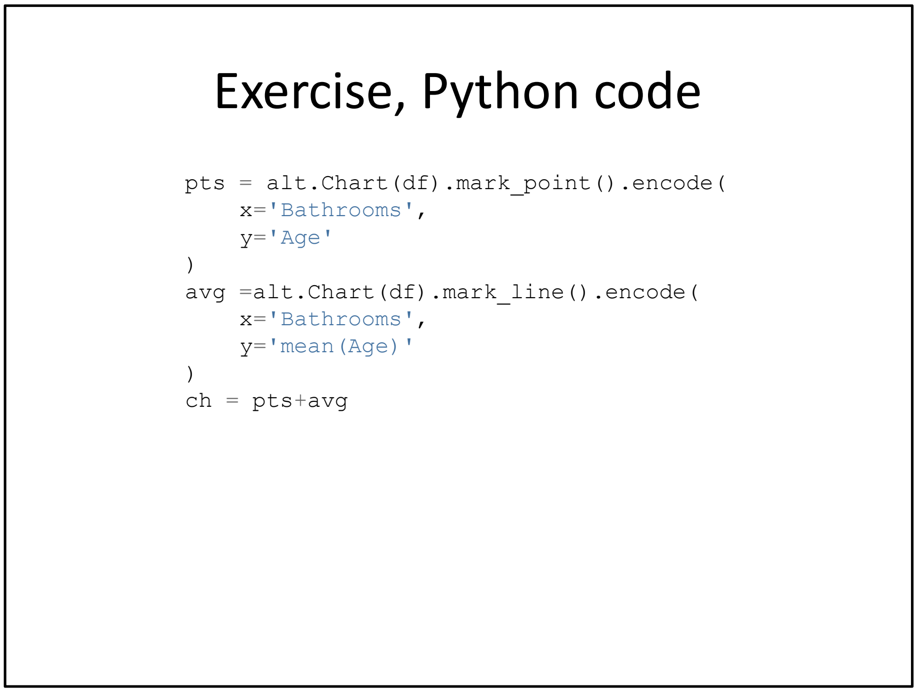
Tableau plot with lines at individual averages



Exercise, Connect individual averages

- – Draw a scatterplot showing
  - X = Bathrooms
  - Y = Age
- – Add a line connect the individual averages

On your own, draw a scatterplot. Put the number of bathrooms on the X axis and the age of the house on the Y-axis. Add a line conencting the individual averages.

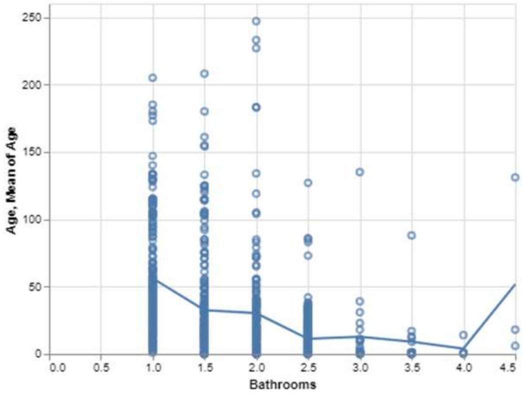


Exercise, Python code

```
pts = alt.Chart(df).mark_point().encode(
    x='Bathrooms',
    y='Age'
)
avg =alt.Chart(df).mark_line().encode(
    x='Bathrooms',
    y='mean(Age)'
)
ch = pts+avg
```

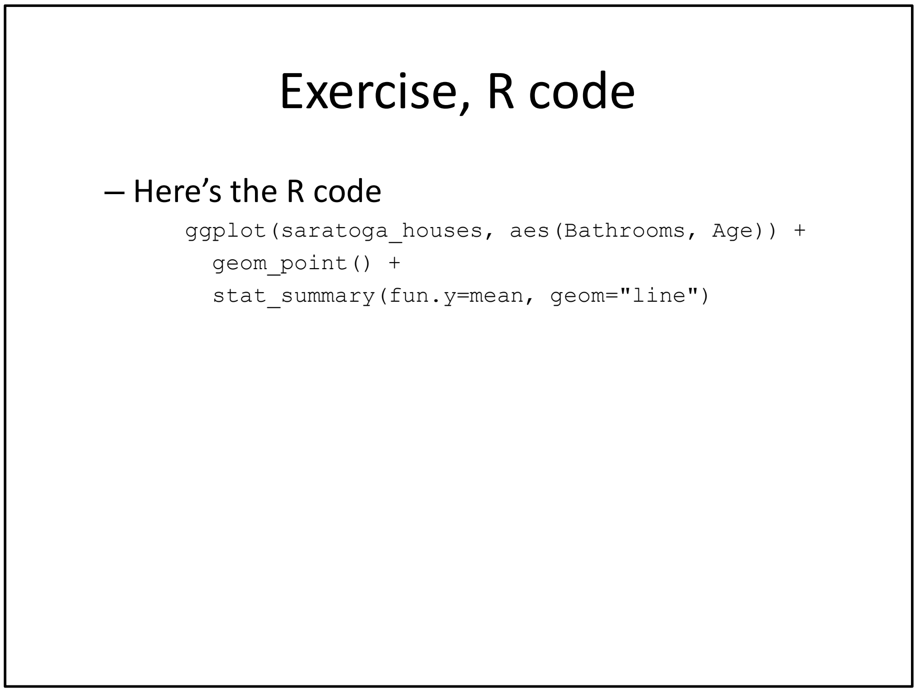
Here is the Python code. Notice how you “add” the points graph and the average line graph.

Exercise, Python output



Python plot showing trend of age versus bathrooms

Here is the Python output.

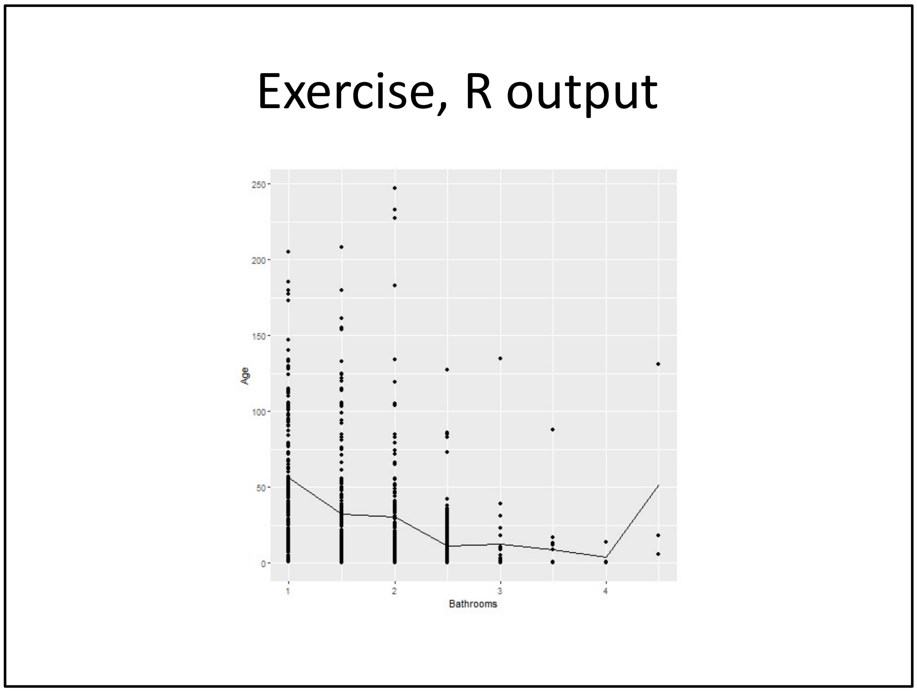


Exercise, R code

– Here’s the R code

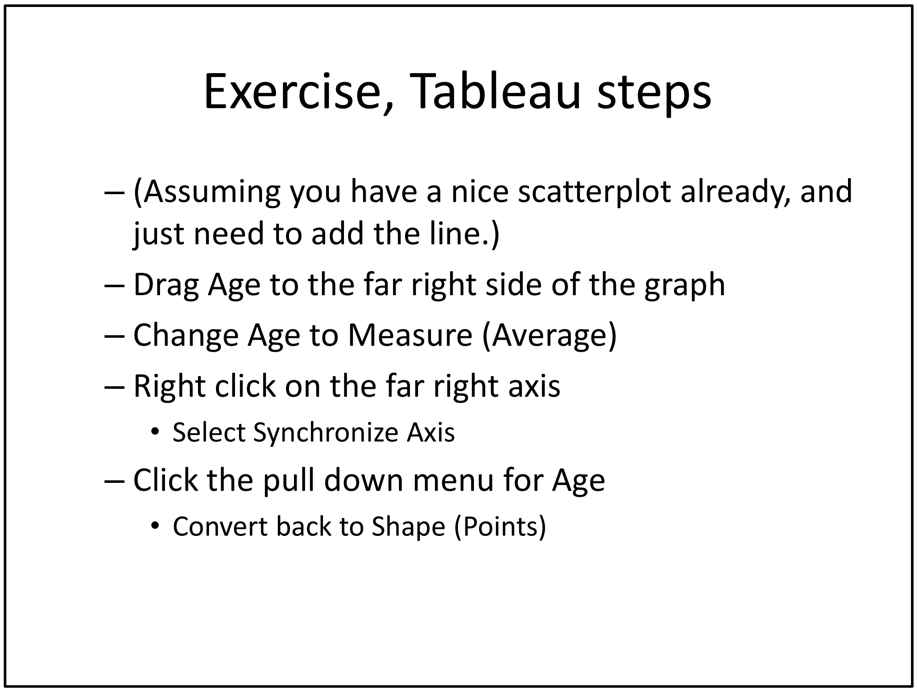
```
ggplot(saratoga_houses, aes(Bathrooms, Age)) +
  geom_point() +
  stat_summary(fun.y=mean, geom="line")
```

The geom_point function creates the scatterplot and the stat_summary function adds the line.



Exercise, R output

There is a clear and consistent trend in the average age. The average age declines as the number of bathrooms increase. There is a blip at the end, but this is probably just an artefact due to the small sample size.



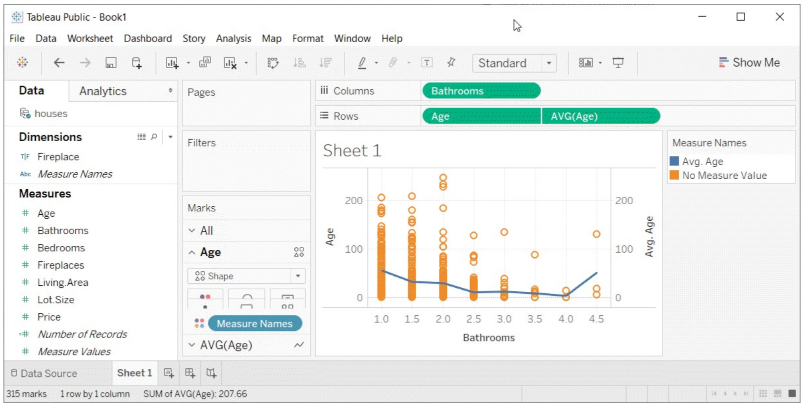
Exercise, Tableau steps

- – (Assuming you have a nice scatterplot already, and just need to add the line.)
- – Drag Age to the far right side of the graph
- – Change Age to Measure (Average)
- – Right click on the far right axis
  - Select Synchronize Axis
- – Click the pull down menu for Age
  - Convert back to Shape (Points)

The process in Tableau is a bit tricky. You need to drag age to the far right side of the graph to create a second axis. Then you need to change Age from Measure (Sum) to Measure (Average). Then you need to synchornize the left and right axes, by right clicking on the right axis.

Finally, change the line graph for Age back to points by selecting Shape from the pull-down menu.

Exercise, Tabeleau output



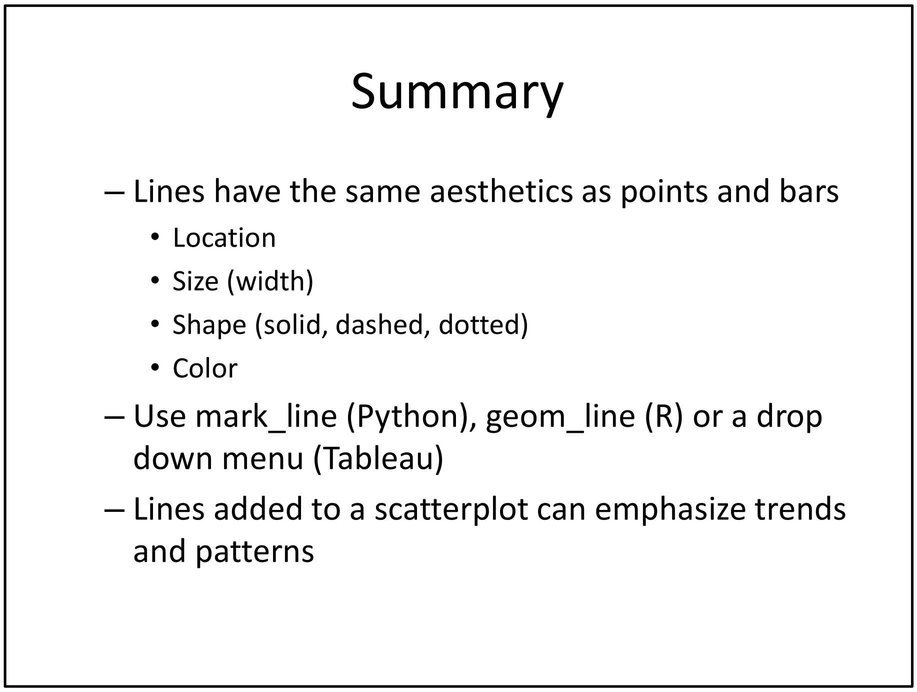
Columns: Bathrooms

Rows: Age, AVG(Age)

315 marks 1 row by 1 column SUM of AVG(Age): 207.66

Tableau plot showing trend in average age versus number of bathrooms

Here is the Tableau graph.



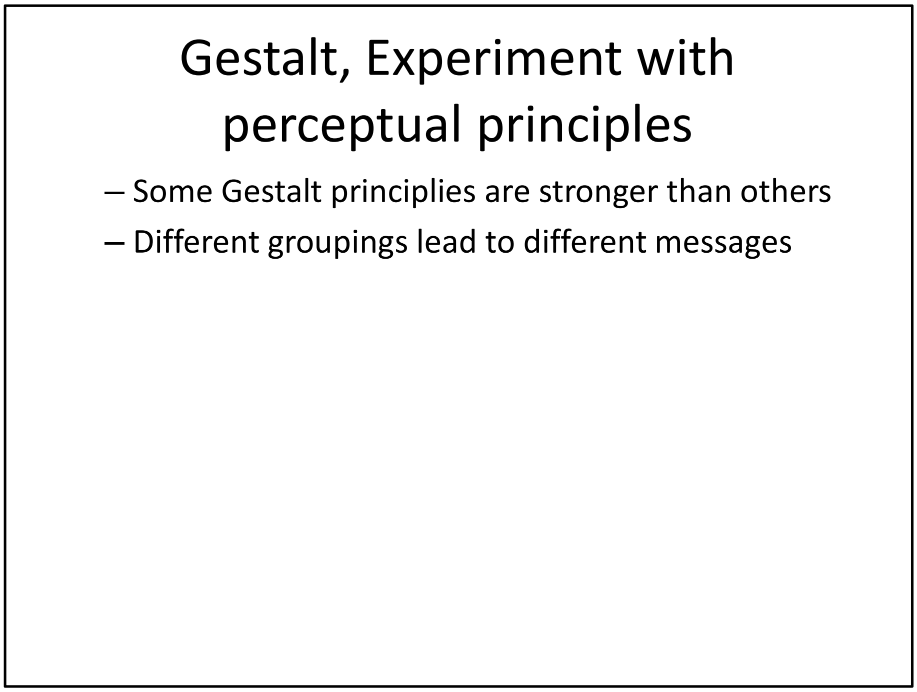
Summary

- – Lines have the same aesthetics as points and bars
  - Location
  - Size (width)
  - Shape (solid, dashed, dotted)
  - Color
- – Use mark_line (Python), geom_line (R) or a drop down menu (Tableau)
- – Lines added to a scatterplot can emphasize trends and patterns

The same aesthetics that you learned about for points and bars also apply to lines. The have an x and y location, a size (meaning width of the line), a shape (solid, dashed, dotted, and various combinations), and color.

In Python, use the mark_line function to draw a line graph. In R, use the geom_line function. In Tableau, there is a pull-down menu.

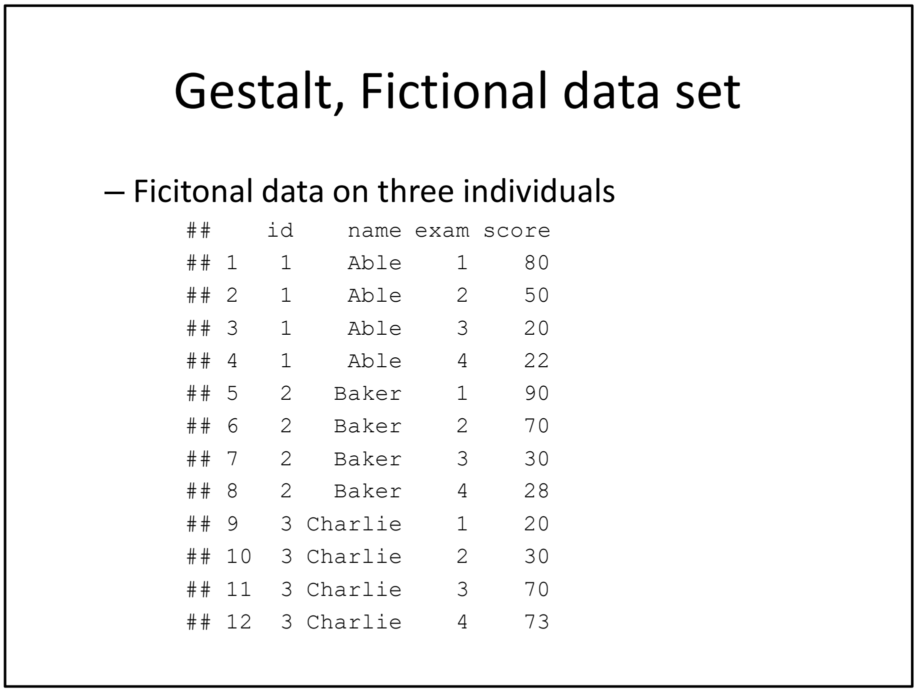
Linegraphs can show a single relationship or compare multiple relationships. You can add a line to a scatterplot to emphasize a trend or pattern.



Gestalt, Experiment with perceptual principles

– Some Gestalt principlies are stronger than others
– Different groupings lead to different messages

The Gestalt principles covered earlier help you to create the appropriate areas of emphasis in your graph. First, you need to recognize that some Gestalt features are stronger than others. Second, you can use Gestalt to create groupings that lead to different messages.

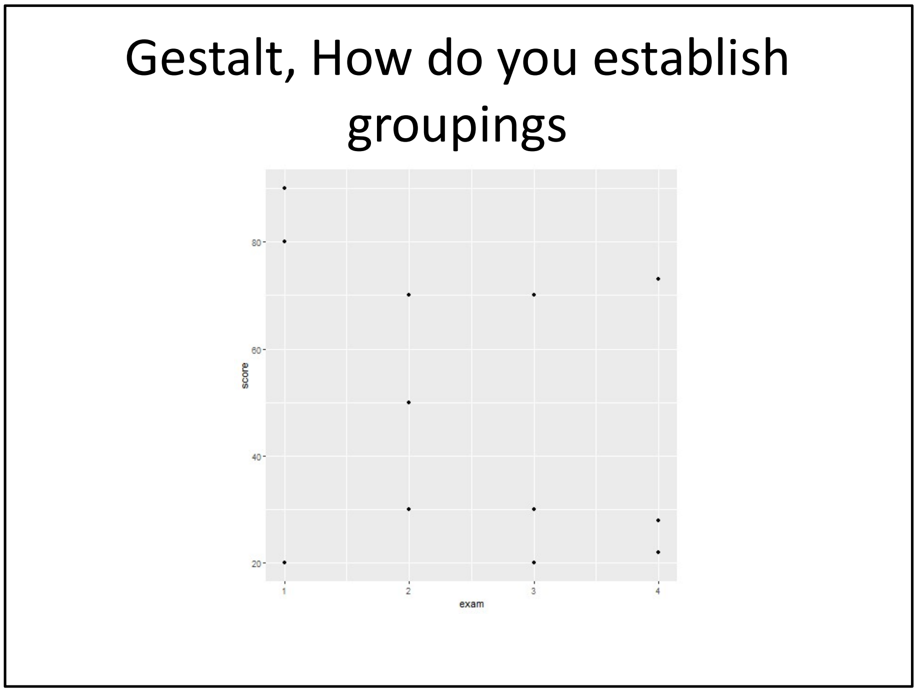


Gestalt, Fictional data set

– Ficitonal data on three individuals

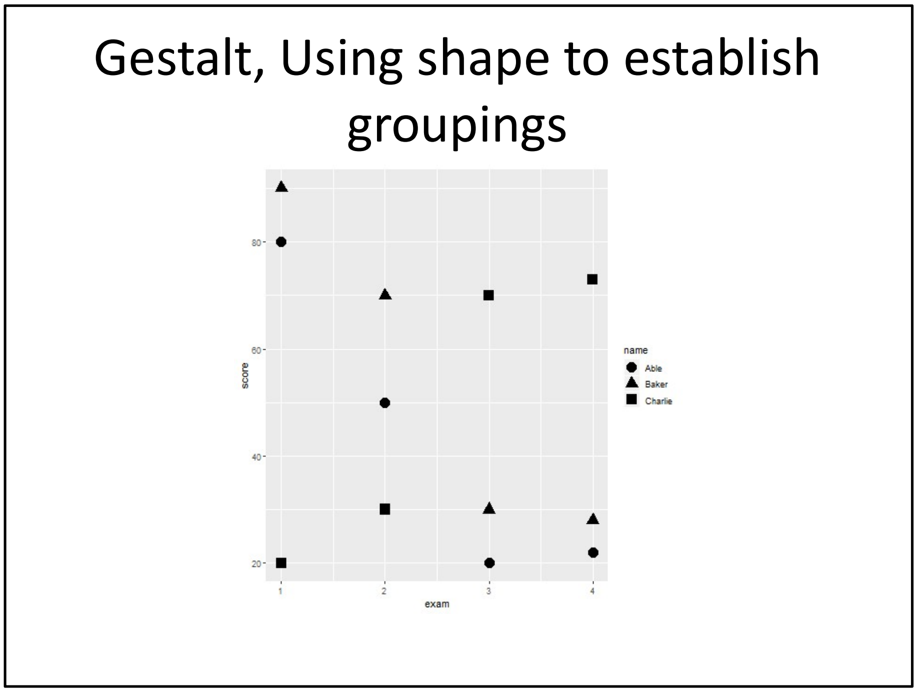
```
##    id    name exam score
## 1   1    Able    1    80
## 2   1    Able    2    50
## 3   1    Able    3    20
## 4   1    Able    4    22
## 5   2   Baker    1    90
## 6   2   Baker    2    70
## 7   2   Baker    3    30
## 8   2   Baker    4    28
## 9   3 Charlie    1    20
## 10  3 Charlie    2    30
## 11  3 Charlie    3    70
## 12  3 Charlie    4    73
```

These ideas drawn from the Bergen and Iverson workshop. The data is purely fictional, but it is a great example of how you can emphasize different features of your data in a visualization.



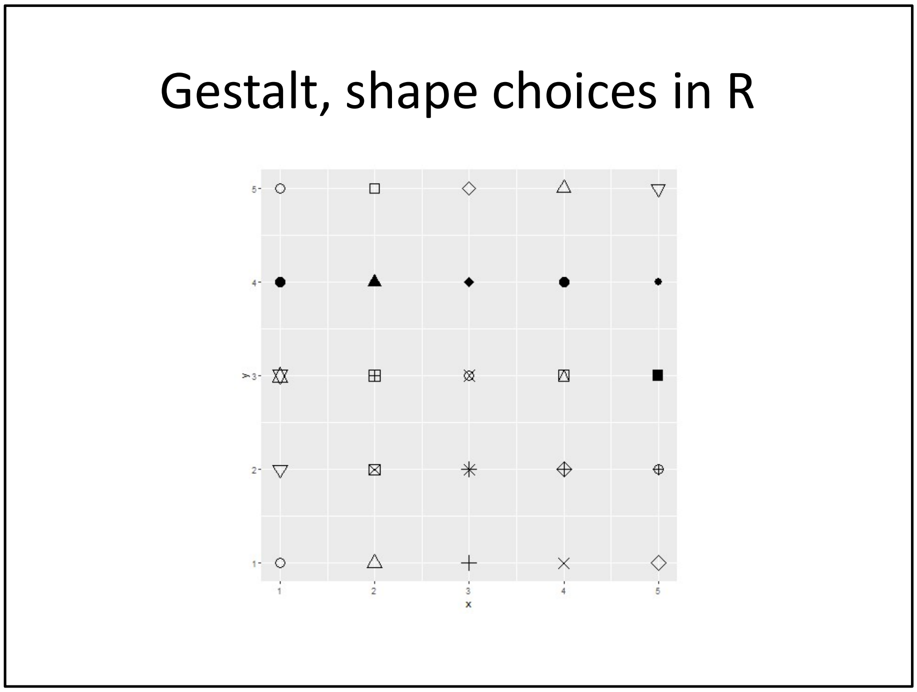
Gestalt, How do you establish groupings

Here are the data points without any indication of which values belong to which individuals. How can you best show the individual values?



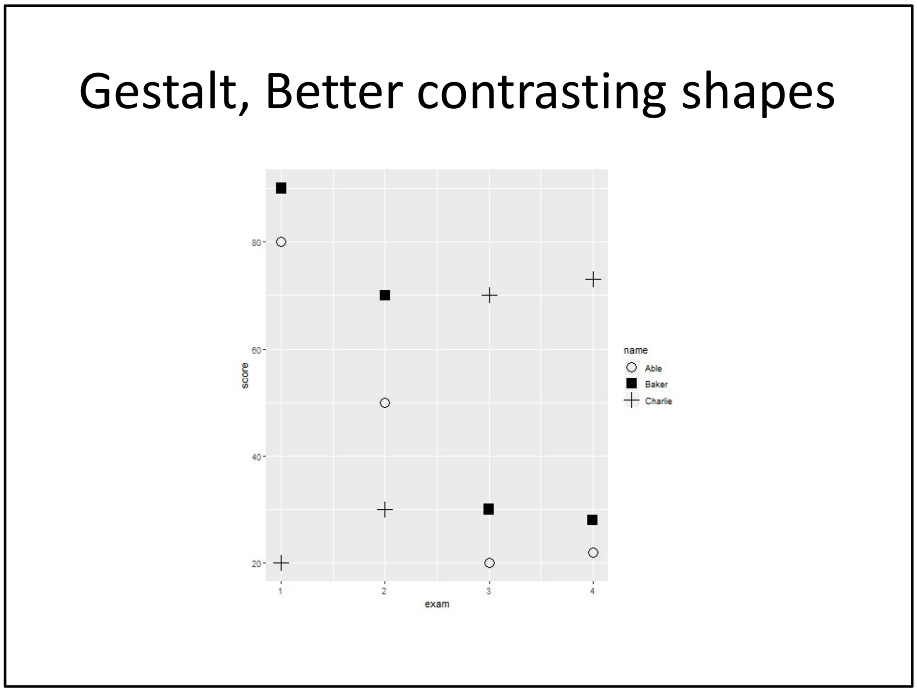
Gestalt, Using shape to establish groupings

You are relying on the principle of similarity, and there needs to be sufficient dissimilarity in the different symbols to make this work. A closed circle, a closed square, and a closed triangle all loop pretty much alike. How can you get more contrast?



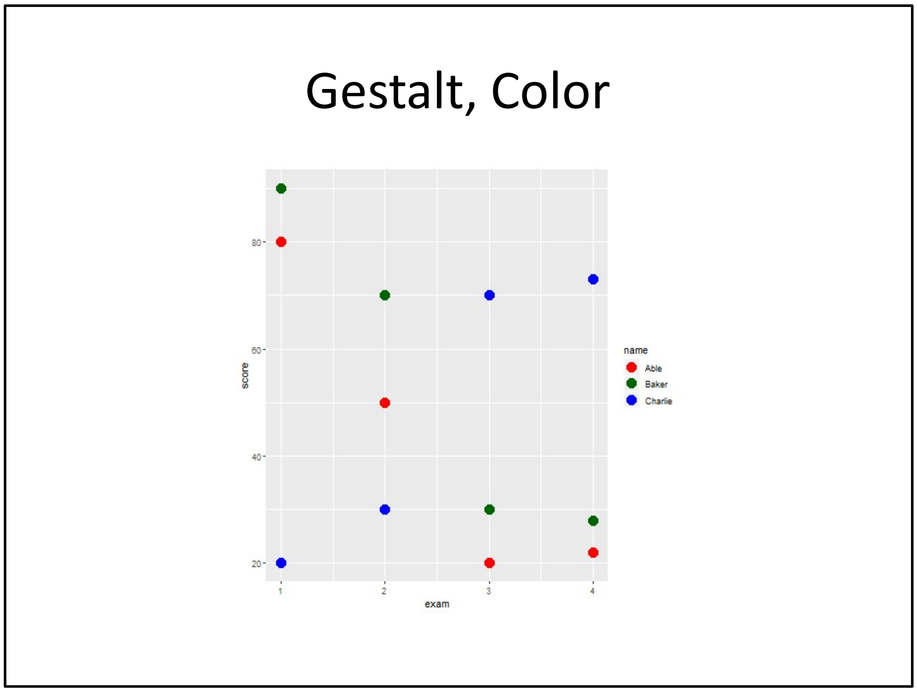
Gestalt, shape choices in R

Here are the shape choices available in R. You can get similar shapes in Python and Tableau. It should not be too hard to choose three shapes that are markedly distinct from one another.



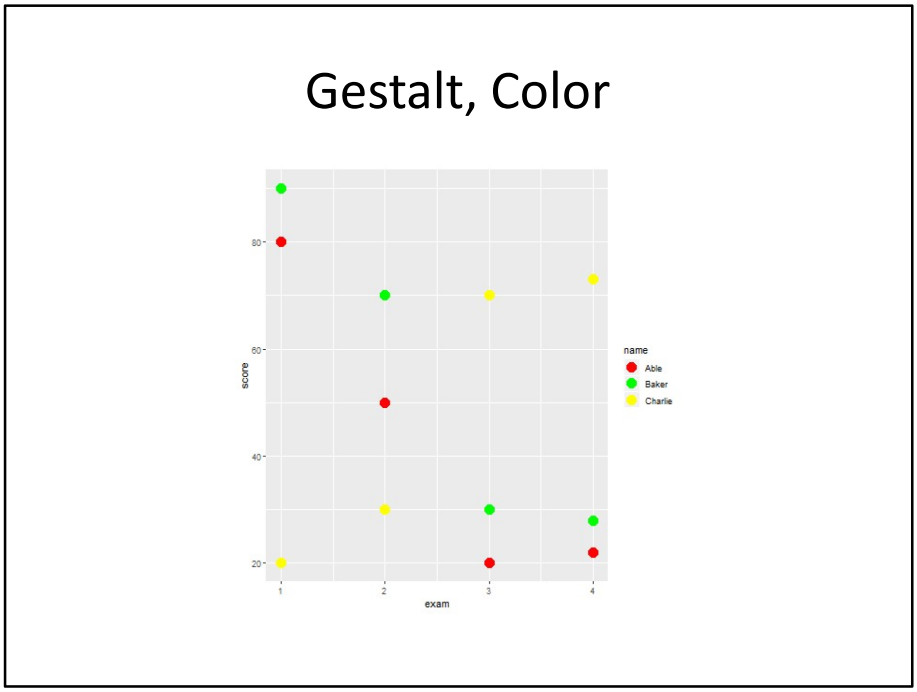
Gestalt, Better contrasting shapes

Here is one possible choice. You can more clearly see differences when you use the solid square, open circle, and plus sign. It gets more difficult if you have a fourth or fifth symbol in the mix.



Gestalt, Color

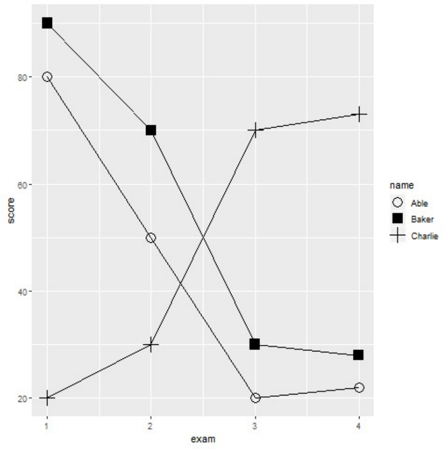
You can use color to emphasize the groupings. Here the colors are red, dark green, and blue. There is some contrast here, but there is not a lot of variation in the brightness.



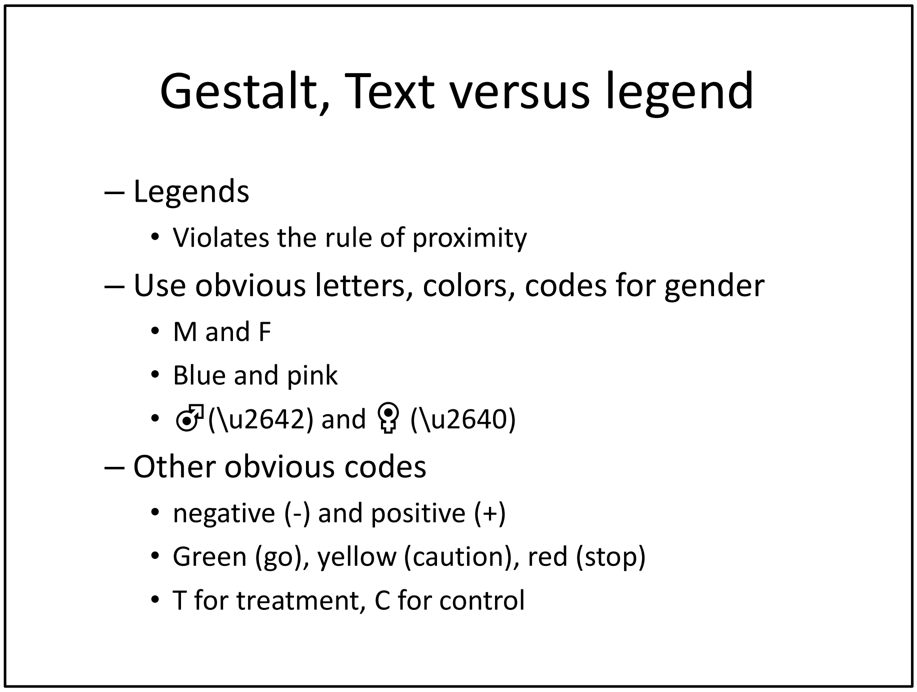
Gestalt, Color

These colors differ in brightness, and provide more of a contrast. There are times when you want this extra contrast, but there are times when you don’t. In particular, the yellow is so bright that it tends to get lost against a light gray background.

Gestalt, Connectedness



Connectedness is a very powerful Gestalt principle. This graph shows much more than the others that there are three distinct sets of points corresponding to the three subjects.



Gestalt, Text versus legend

- Legends
  - Violates the rule of proximity
- Use obvious letters, colors, codes for gender
  - M and F
  - Blue and pink
  - ♂(\u2642) and ♀ (\u2640)
- Other obvious codes
  - negative (-) and positive (+)
  - Green (go), yellow (caution), red (stop)
  - T for treatment, C for control

Legends are very common in visualization, but they violate the rule of proximity. A legend forces your eye to go back and forth between the graph and a location far outside the graph.

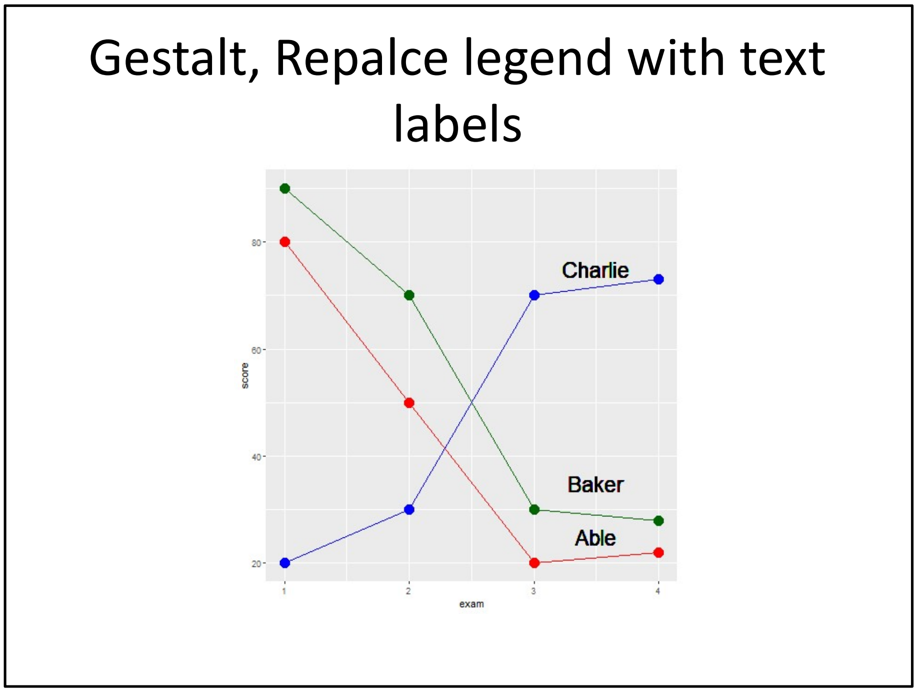
It takes some work, but if you can identify groups properly without a legend, you are much better off. There are several strategies that could help here.

Certain letter codes are expected and almost second-nature for visualizations. The letters M and F are commonly understood to represent male and female, at least among your English speaking audience. The colors blue and pink also evoke a similar recognition. You could also use the commonly recognized symbols for male and female. These are 2642 and 2640 in the Unicode font.

For something other than gender, you might be able to find well understood associations. The - and plus symbols are generally recognized as negative and positive. The colors associated with a traffic light (green, yellow, and red) are commonly associated with go, caution, and stop.

Treatment and Control groups are also readily recognized with the letters T and C.

With well recognized codes, you might be able to dispense with a legend and just point out the obvious association in a footnote to your graph.

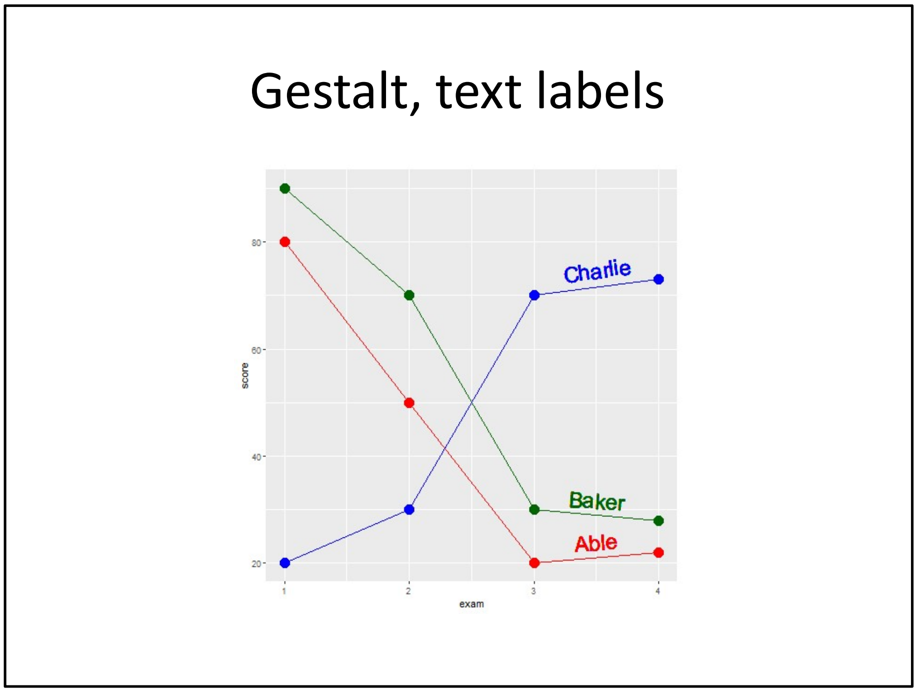


Gestalt, Repalce legend with text labels

By putting the labels directly on the graph, you minimize the distance that your eye has to travel. No more back and forth between the legend and the graph.

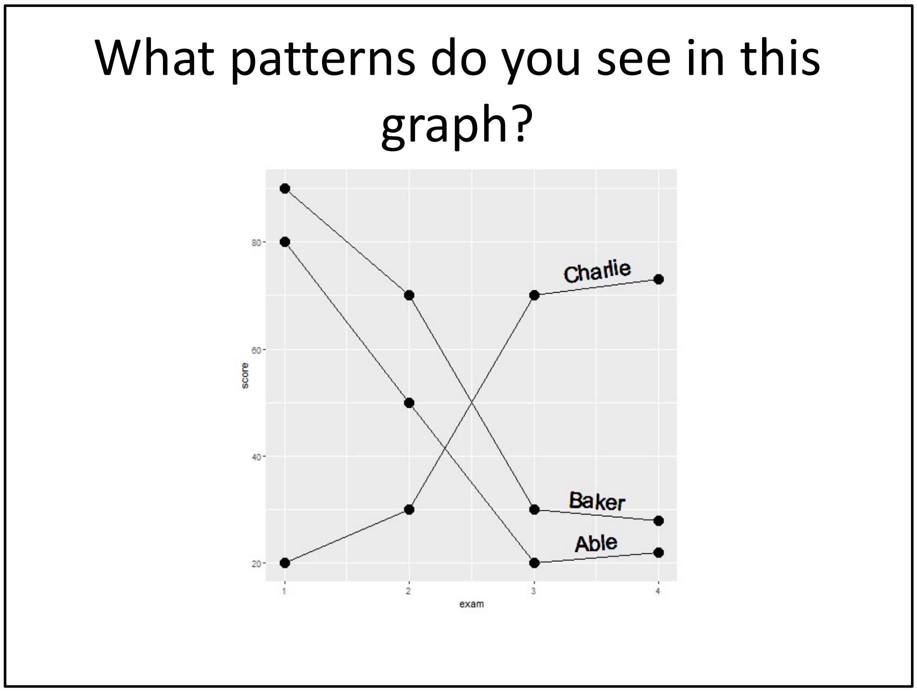
You can improve further on these labels. You might have had a moment of hesitation when you saw the label for Abel. Does it belong with the line just above or the line just below. It only slows you down for a second, because by process of elimination, you can quickly decide that it belongs with the line just below it.

But you can improve the sense of belongingness by using some of your Gestalt principles.



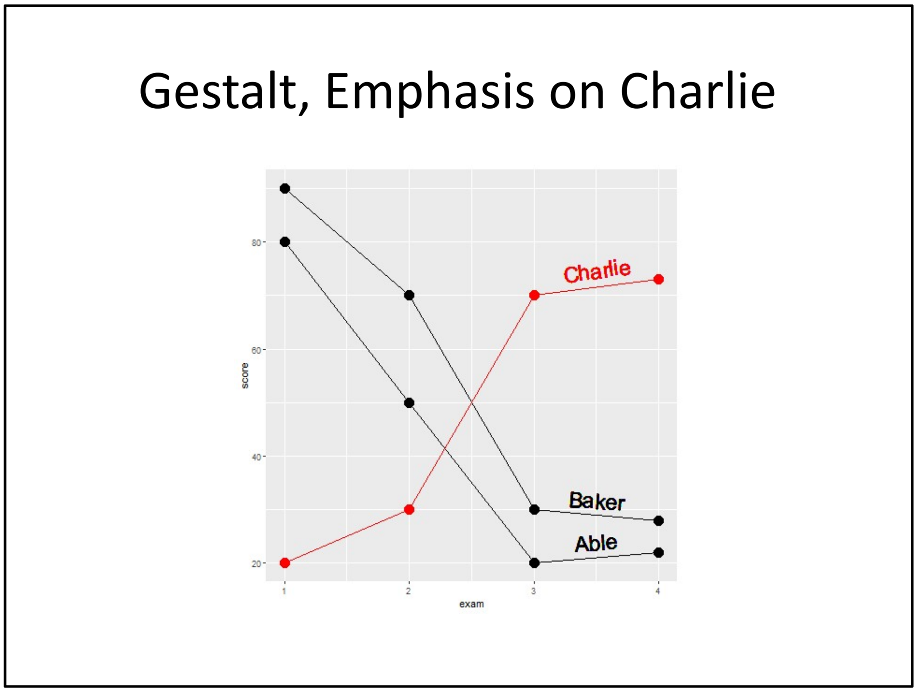
Gestalt, text labels

Notice how the labels in this example have been moved closer to the lines. The exploits the Gestalt principle of proximity. Notice how the names have the same colors as the lines. Thie exploits the Gestalt principle of similarity. It would be very hard to incorrectly assign the red label of Abel with the green line of Baker. Notice further that the labels have a slope that roughly matches the lines they are attached to. This exploits the Gestalt principle of common fate.



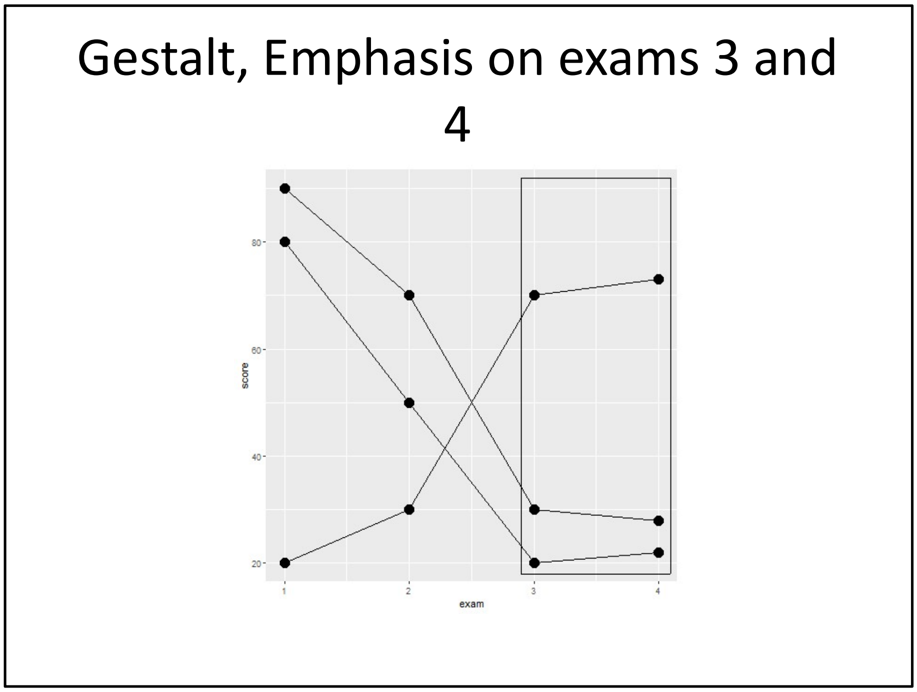
What patterns do you see in this graph?

There are at least two messages hiding in this graph. What are they?



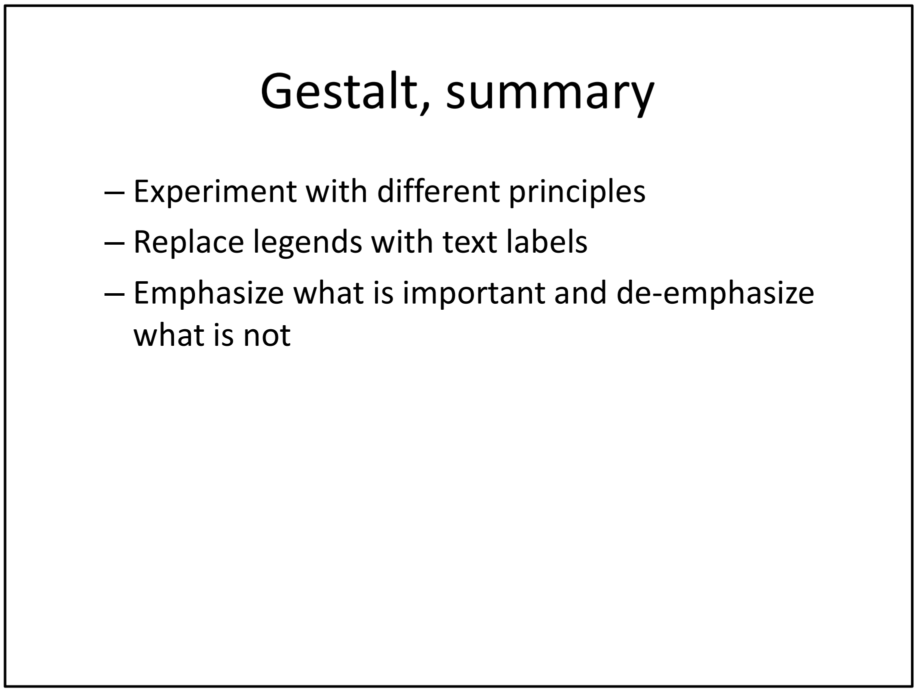
Gestalt, Emphasis on Charlie

Using a single color to emphasize the one subject who trends in the opposite direction from the other two.



Gestalt, Emphasis on exams 3 and 4

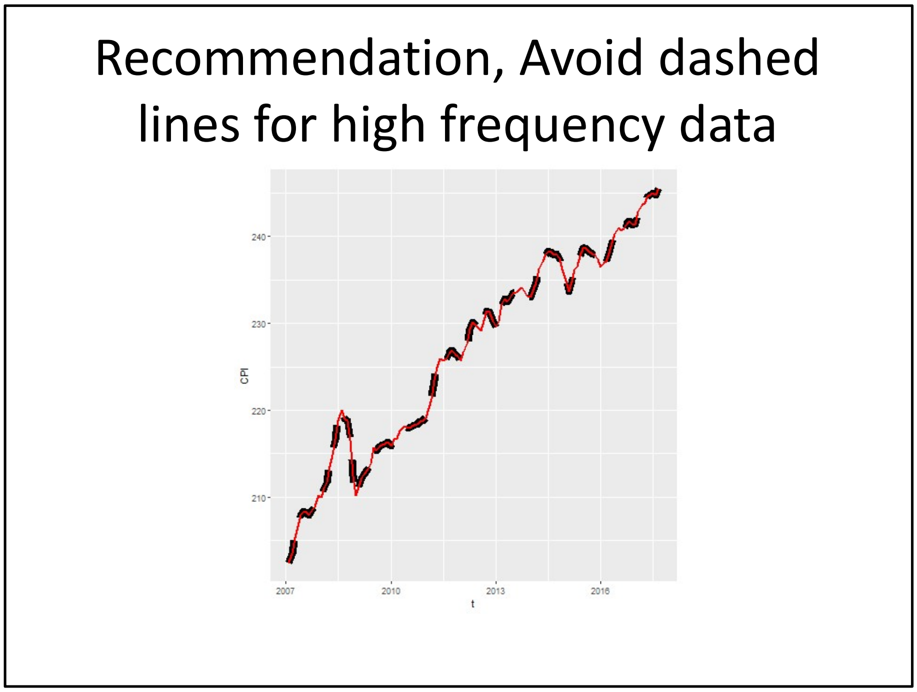
Using enclosure to emphasize the similarity of exam 3 and exam 4. Notice how the emphasis on individual subjects is lessened by keeping them the same color and removing the text labels.



Gestalt, summary

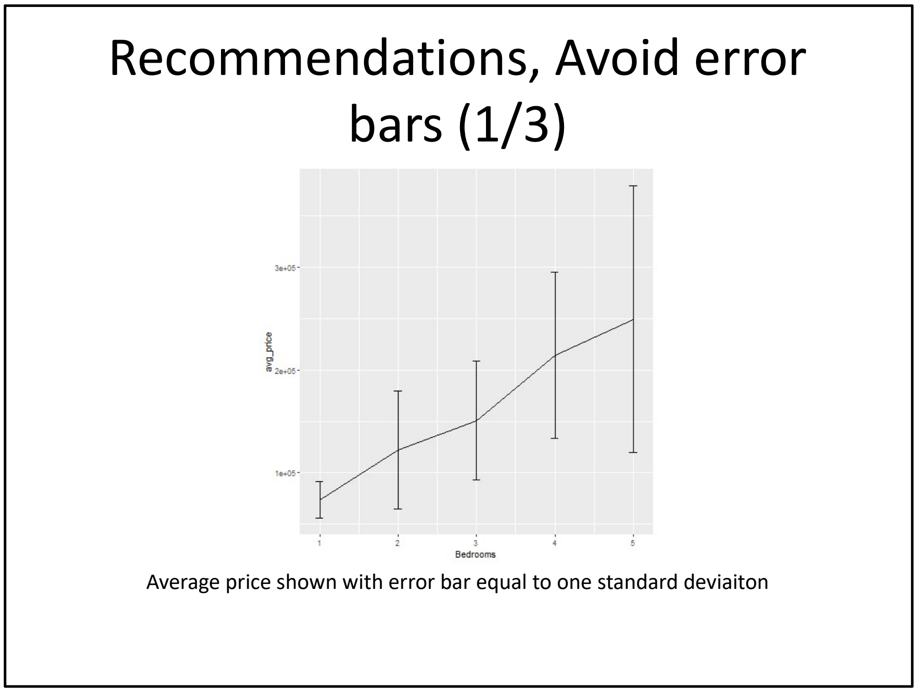
– Experiment with different principles
– Replace legends with text labels
– Emphasize what is important and de-emphasize what is not

In summary, you should experiment with different Gestalt principles. Some principles work more effectively than others. Legends are commonly used in visualization, but they violate the Gestalt principle of proximity.



Recommendation, Avoid dashed lines for high frequency data

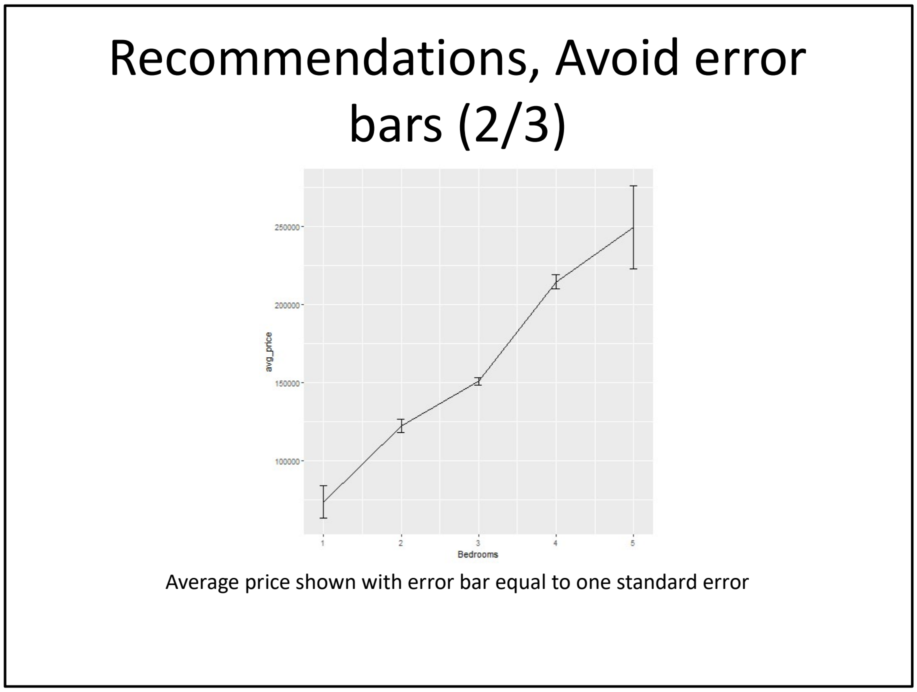
High frequency events are sudden surges and dips in a line over a narrow range. If your data has high frequency events (think stock market prices), then avoid thick lines or long dashes. These can hide some of the high frequency events.



Recommendations, Avoid error bars (1/3)

Average price shown with error bar equal to one standard deviaiton

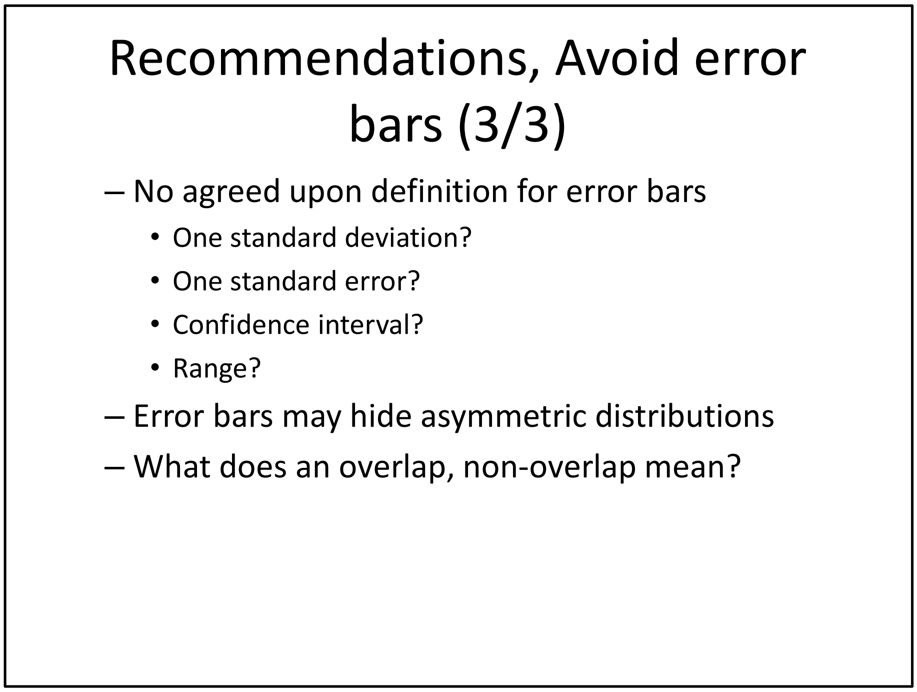
This graph shows error bars for the average price of a house given a fixed number of bedrooms. There is indeed an upward trend in prices. The more bedrooms you have, the higher the average price. But these error bars, equal to plus or minus one standard deviation, emphasize that there is a lot of overlap in individual prices.



Recommendations, Avoid error bars (2/3)

Average price shown with error bar equal to one standard error

Here is the same graph, but with the error bar representing plus or minus one standard error. The standard error is the standard deviation divided by the square root of the sample size. It shows how variable the mean is rather than how variable an individual observation is. This picture is quite different, and implies a much stronger trend. Which graph is better? Well, neither in my opinion.



Recommendations, Avoid error bars (3/3)

- – No agreed upon definition for error bars
  - One standard deviation?
  - One standard error?
  - Confidence interval?
  - Range?
- – Error bars may hide asymmetric distributions
- – What does an overlap, non-overlap mean?

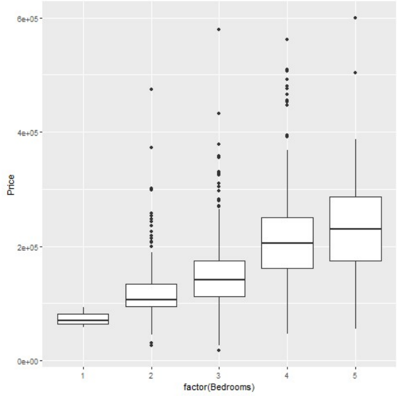
The first problem with error bars is that they do not have a commonly agreed upon definition. Is it plus or minus one standard deviation or plus or minus one standard error?

The general rule is that you use a standard deviation or a range when you want to place the emphasis on individual values and a standard error or confidence interval when you want to place the emphasis on the average.

A big problem with error bars is that they may falsely imply a symmetry to your data. If your data is skewed, then try to avoid error bars.

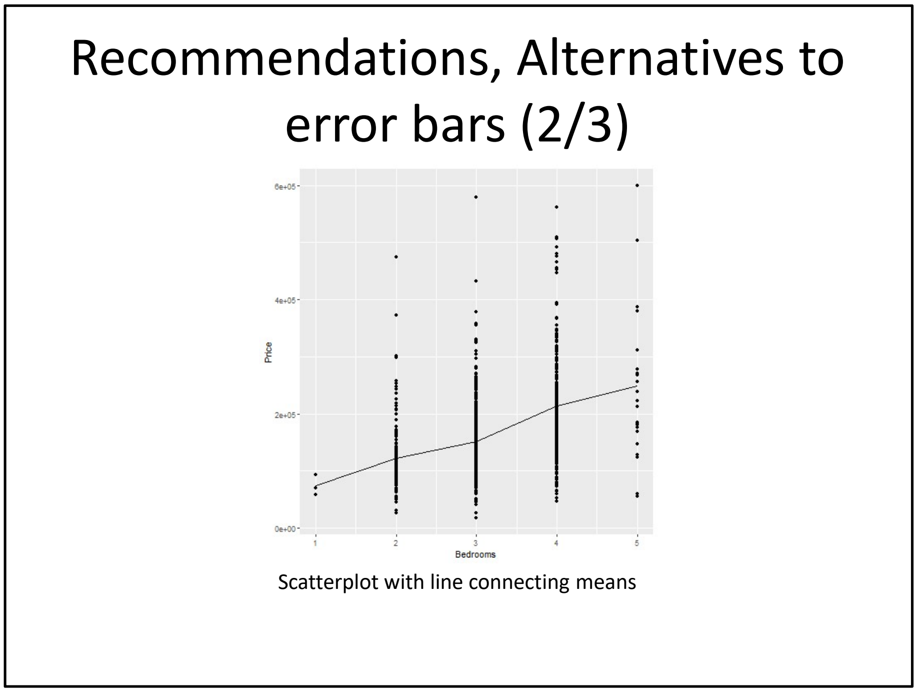
The other problem with error bars is that people make up rules, such as if two sets of error bars overlap, there is not a statistically significant difference between the two means. That only works for a confidence interval and it doesn’t quite work then even. Standard errors and confidence intervals are not additive, and a small amount of overlap between two confidence intervals proves nothing.

Recommendations, Alternatives to error bars (1/3)



Boxplot of house prices by number of bedrooms

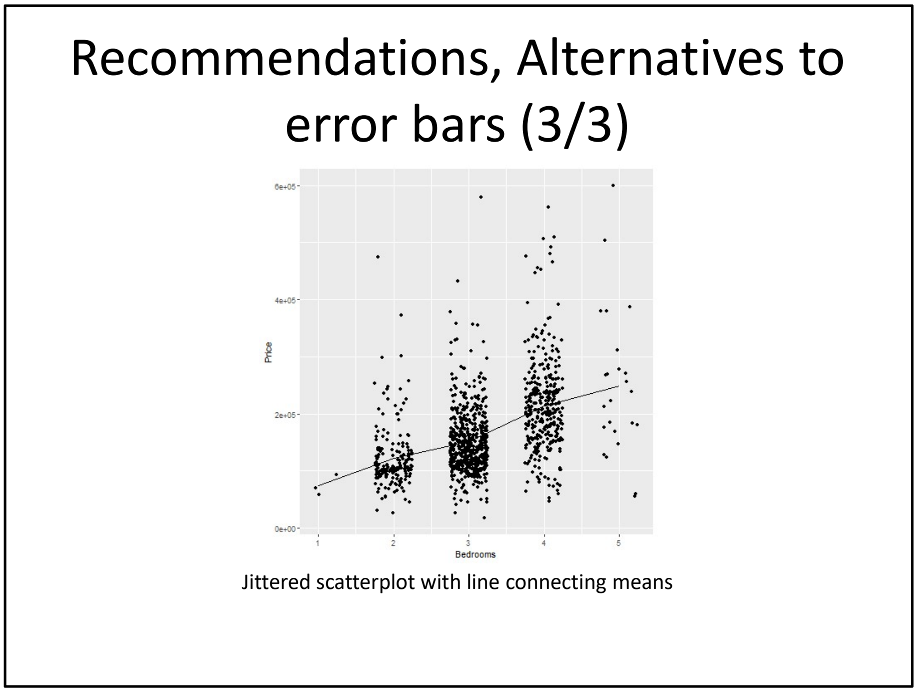
The box plot provides a more useful summary because, among other reasons, it shows whether your data is skewed in one direction or another.



Recommendations, Alternatives to error bars (2/3)

Scatterplot with line connecting means

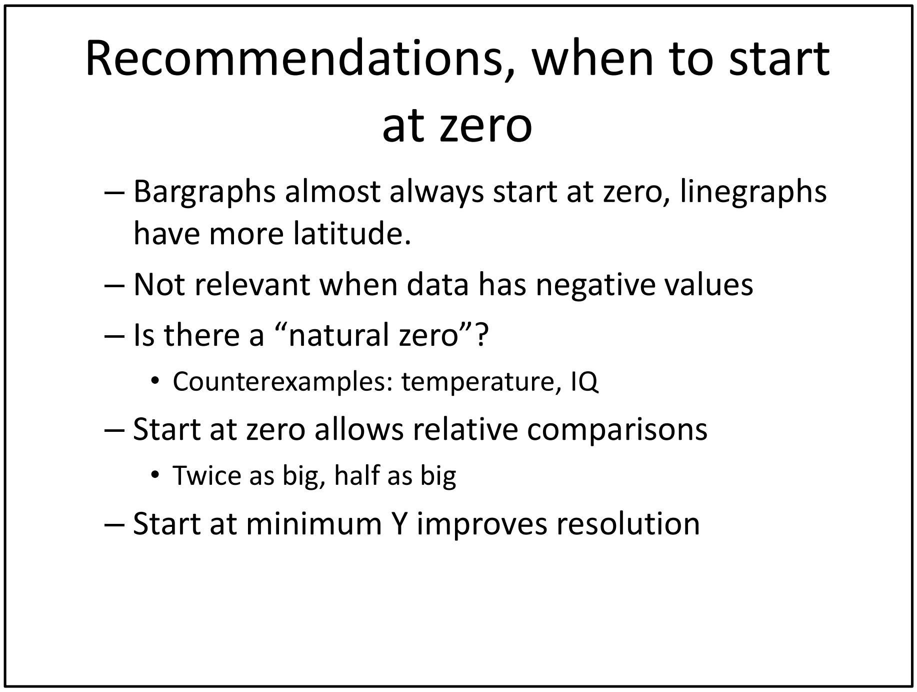
There is often no substitute, however, for showing every individual data point.



Recommendations, Alternatives to error bars (3/3)

Jittered scatterplot with line connecting means

There is a fair amount of overprinting in this graph, so jittering (randomly shifting data points around) can help you see where there are just a few points and where there are many points.



Recommendations, when to start at zero

- Bargraphs almost always start at zero, linegraphs have more latitude.
- Not relevant when data has negative values
- Is there a “natural zero”?
  - Counterexamples: temperature, IQ
- Start at zero allows relative comparisons
  - Twice as big, half as big
- Start at minimum Y improves resolution

There is a controversy that never seems to end about whether your y-axis should include the value of zero. It doesn’t seem to be as much of a controversy about the X axis, for what it’s worth. It is also not much of a controversy for bargraphs, as the strong consensus is that bargraphs should always start at zero.

You should resign yourself to the fact that if you come down on one side of this controversy, your boss is probably going to be on the other side. So get in the habit of knowing how to change from one type of graph to another.

First of all, recognize that if you have negative values in your data set, the controversy is moot. You include zero and keep on going down to include all of the negative values.

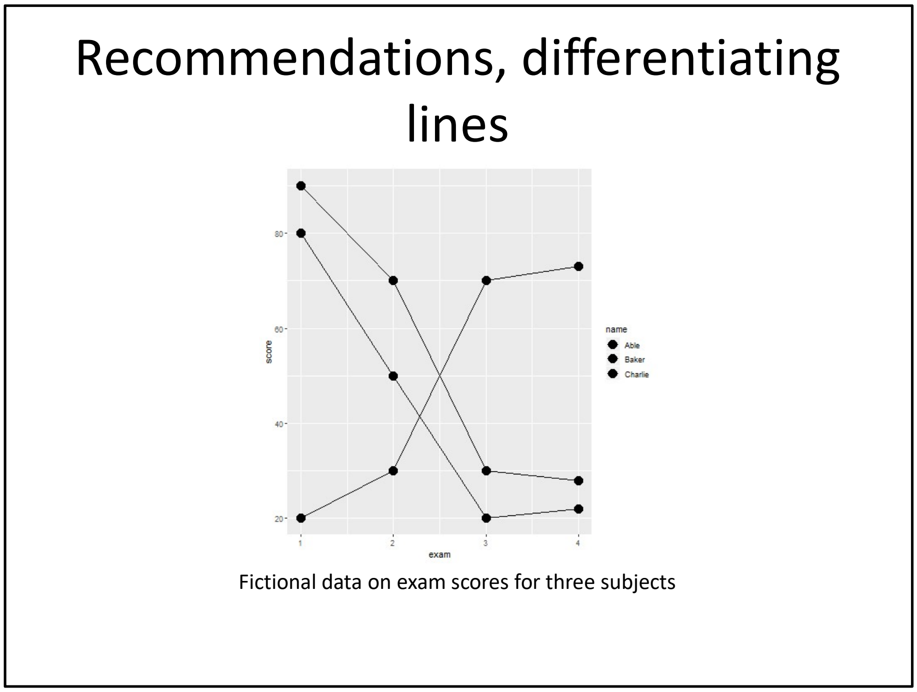
You also need to ask yourself whether your data contains a natural zero, a value that is well accepted as meaning “nothing”. Two counterexamples are temperature and IQ. Unless you are talking about the Kelvin scale where zero represents the physical concept of absolute zero, temperature does not have a natural zero. It doesn’t matter whether you are talking 0 Fahrenheit or) Celsius. Neither temperature represents the absence of all heat.

IQ also does not have a natural zero. There is no IQ rather represents the absence of all intelligence.

Starting at zero allows you to make relative comparisons. You can see when one Y value is twice as big or half as big as another value. Do relative comparisons make sense for your data. They certainly don't for IQ, because you can't say that a person with an IQ of 150 is twice as intelligent as a person with an IQ of 75.

You lose something though, when you start at zero rather than the starting at the minimum possible Y value. The extra space required by including zero forces your entire linegraph into a narrow range, making it harder for you to make absolute comparisons.

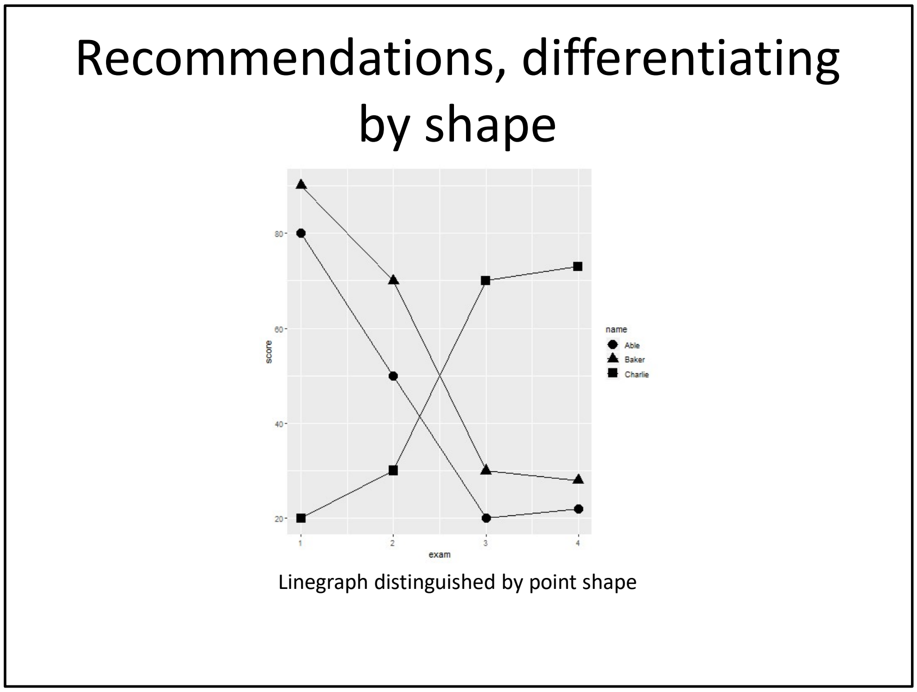
I think that debates about whether to include zero or not are often silly, but the safe thing to do is to include a zero whenever you think the reader might be interested in relative comparisons.



Recommendations, differentiating lines

Fictional data on exam scores for three subjects

How do you best differentiate among lines. In this graph, you can’t tell which line represents which subject.

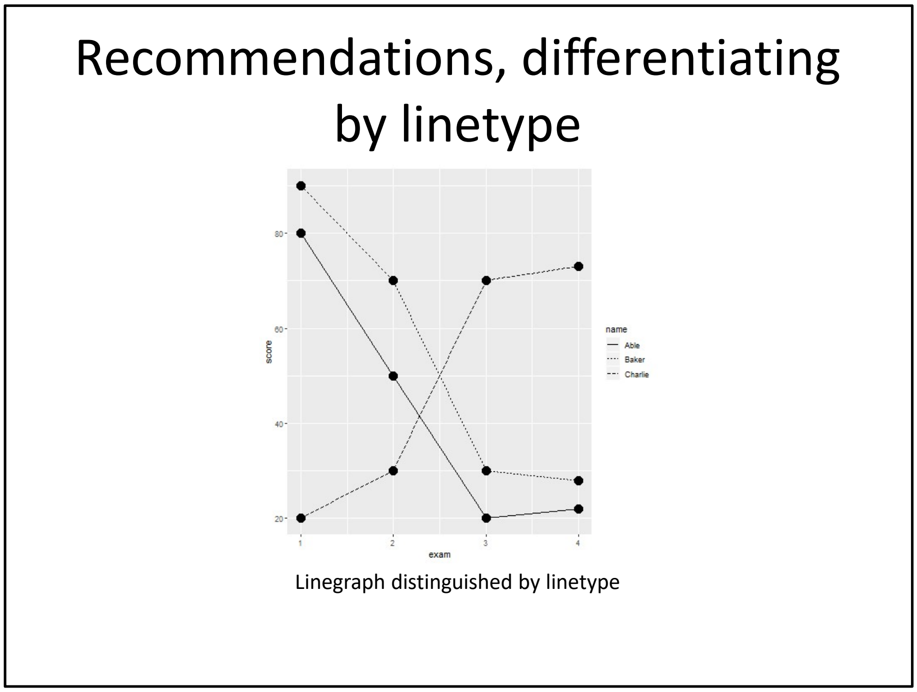


Recommendations, differentiating by shape

Linegraph distinguished by point shape

I deliberately added points to the lines because it gives you an extra chance to differentiate the lines. In this graph, the square, circle, and triangle allow you to tell which subject is which.

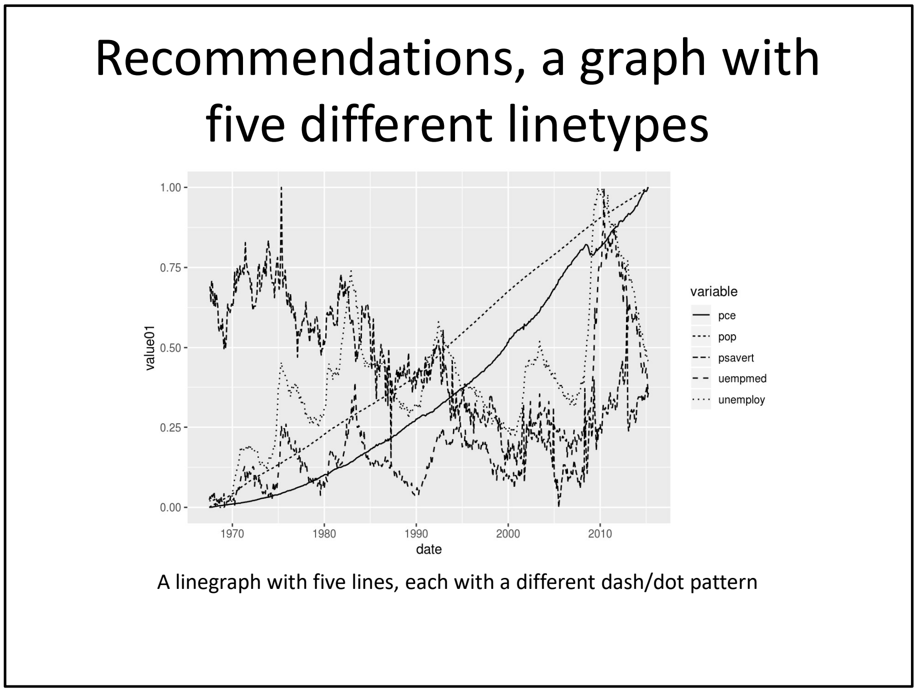
You'd be a bit better off if you compared an open symbol with a closed symbol versus an "X" or a "+" that is neither open nor closed. Even so, the use of symbols, by itself, is not very helpful in allowing you to distinguish which line is associated with which subject.



Recommendations, differentiating by linetype

Linegraph distinguished by linetype

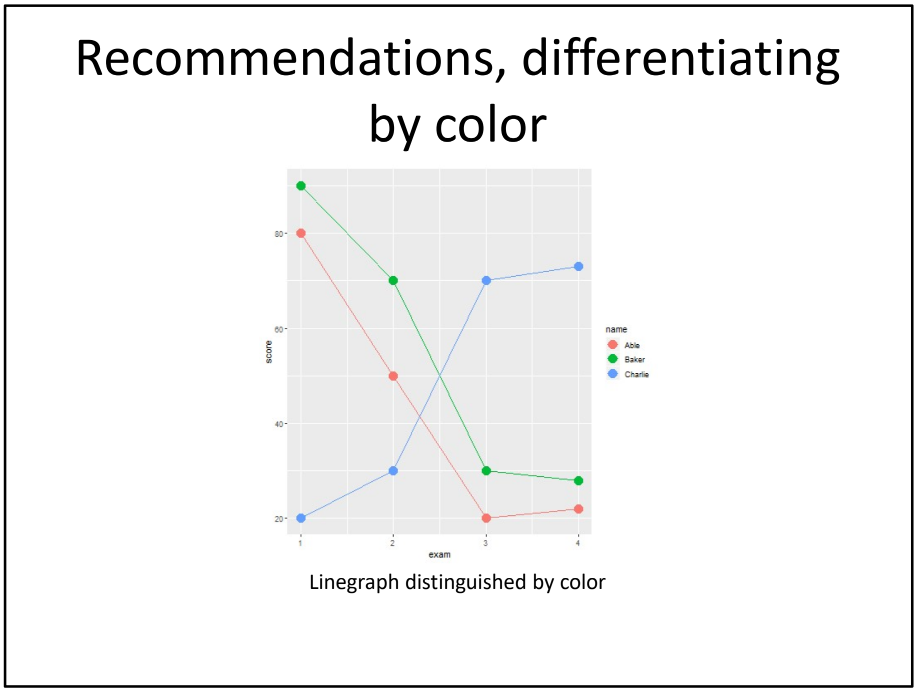
You can also differentiate by linetype. This works fine for two groups, perhaps because a solid line looks quite different than any dotted or dashed line. The problem with three or more groups is that the various dotted and dashed lines look very similar. Even this simple graph, it is a bit difficult to distinguish the short dashed line of Baker from the long dashed line of Charlie.



Recommendations, a graph with five different linetypes

A linegraph with five lines, each with a different dash/dot pattern

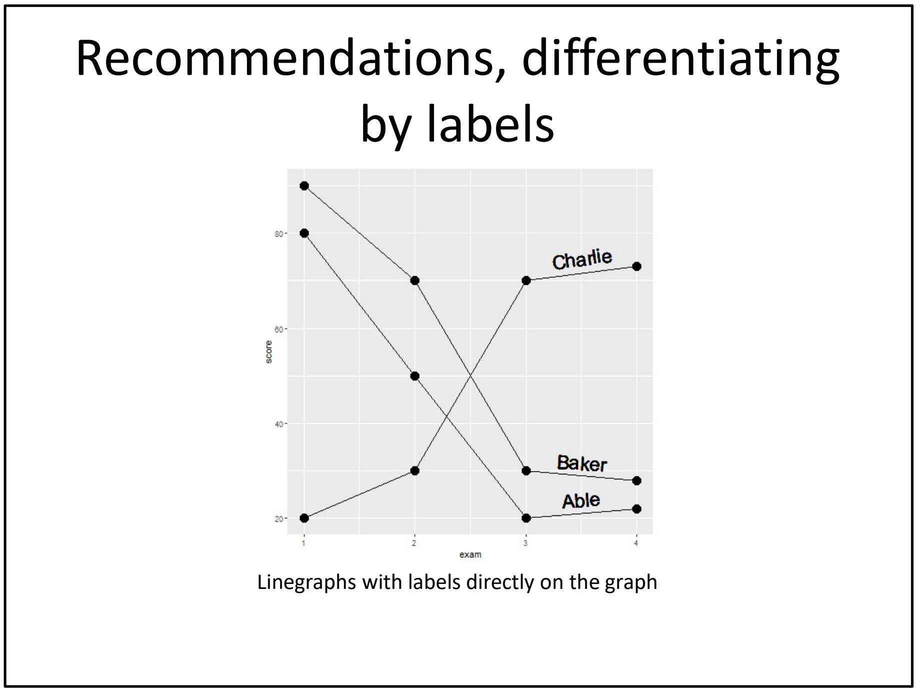
I hesitate to show this graph, because it was drawn mostly just to illustrate the variety of linetypes that R has. But it does illustrate how confusing a graph can be when the only thing you have to distinguish among five different lines is the linetype.



Recommendations, differentiating by color

Linegraph distinguished by color

Color works a little bit better and it is not hard to find four or five colors that are fairly easy to distinguish from one another.

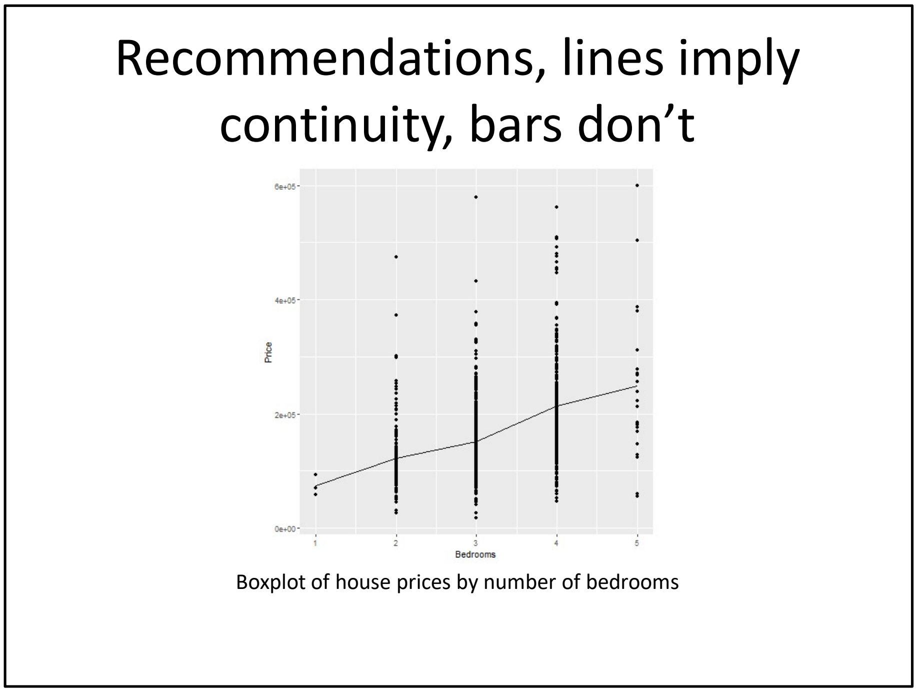


Recommendations, differentiating by labels

Linegraphs with labels directly on the graph

Labels take advantage of the Gestalt principle of proximity. They often require a bit of hand tweaking, so they are tedious. But labels are the best approach by themselves to distinguishing among different lines.

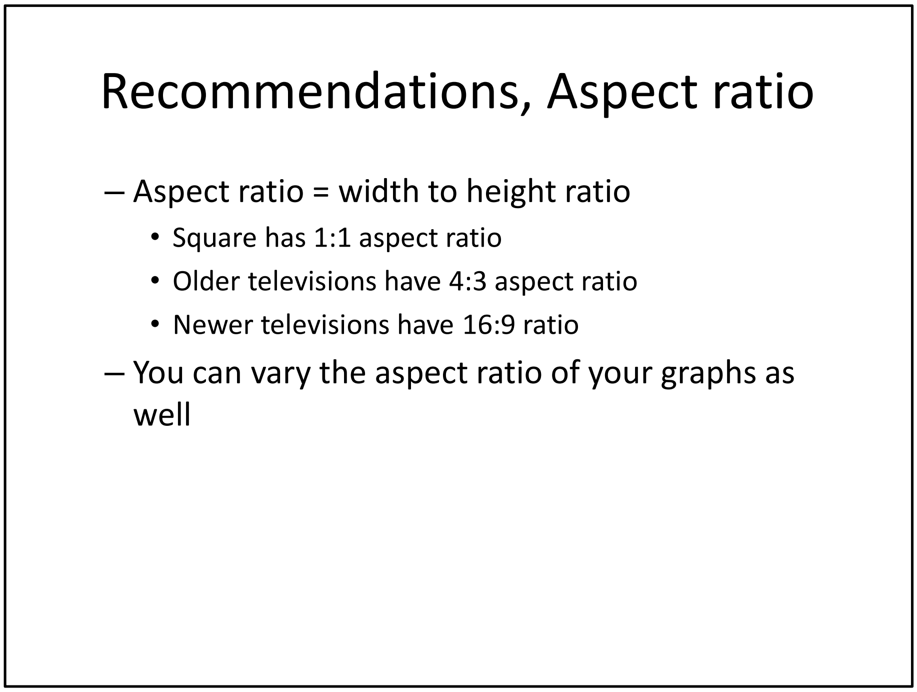
Of course, you can and should consider doubling up and combining two elements like color and linetype, if you want to make the lines readily distinguishable from one another.



Recommendations, lines imply continuity, bars don't

Boxplot of house prices by number of bedrooms

The line connecting the one bedroom house mean to the two bedroom house mean to the three bedrood house mean, etc. is thought by some people to imply that intermediate values are possible. But you can't have 2.2 bedrooms. It has to be a whole number. This does not bother me, but if it bothers you, then convert this graph to a bar chart or use discontinuous line segments.



Recommendations, Aspect ratio

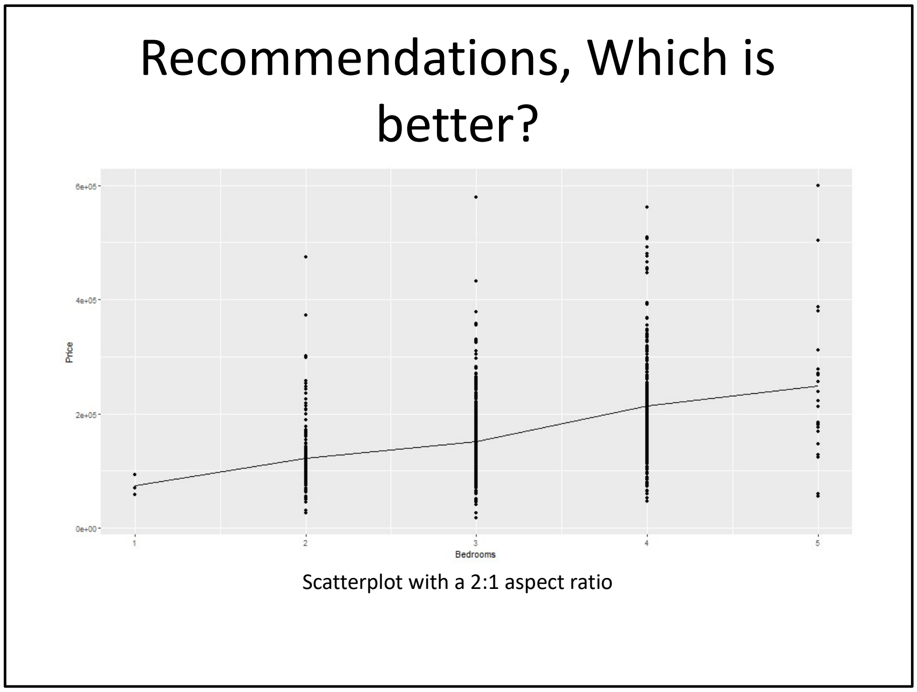
- Aspect ratio = width to height ratio
  - Square has 1:1 aspect ratio
  - Older televisions have 4:3 aspect ratio
  - Newer televisions have 16:9 ratio
- You can vary the aspect ratio of your graphs as well

The aspect ratio is the ratio of the length of a rectangle to its height. An aspect ratio of 1 to 1 is a square.

If you remember when televisions switched from a big tube to a flat panel, you will also remember that the rectangular shape of the screen changed as well. The older tube televisions had an almost square 4 to 3 ratio. They were just a little bit wider than they were tall. This was the standard from all the way back in the 1950's.

The newer televisions were much wider, like a movie screen and had an aspect ratio of 16 to 9.

You can choose a variety of different aspect ratios for your graphs as well.

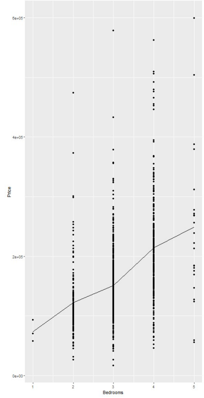


Recommendations, Which is better?

Scatterplot with a 2:1 aspect ratio

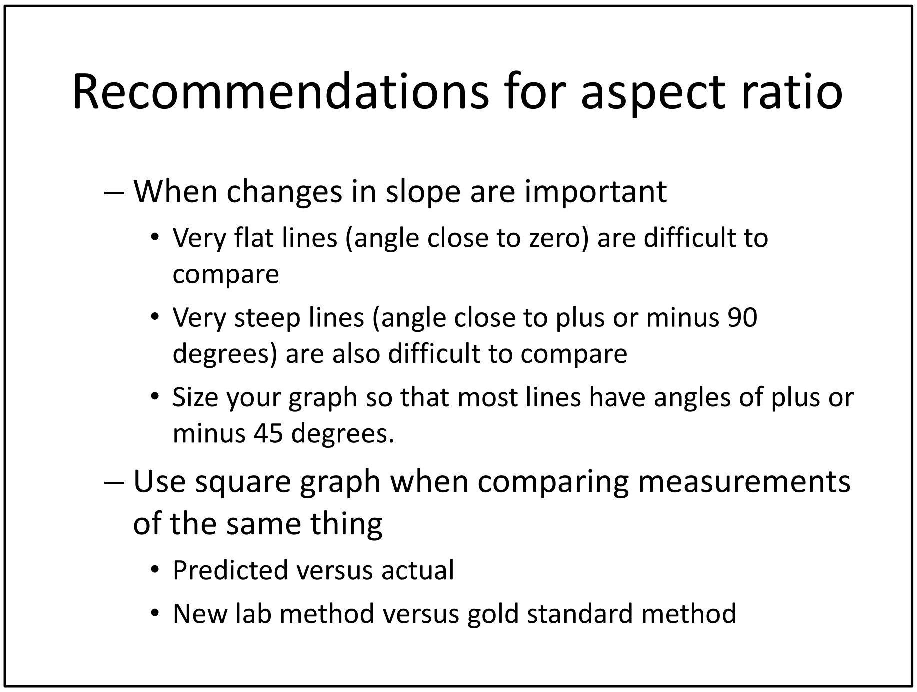
This is a graph which is (roughly) twice as wide as it is tall. Sometimes this is called landscape orientation. Compare this to the next graph.

Recommendations, Which is better?



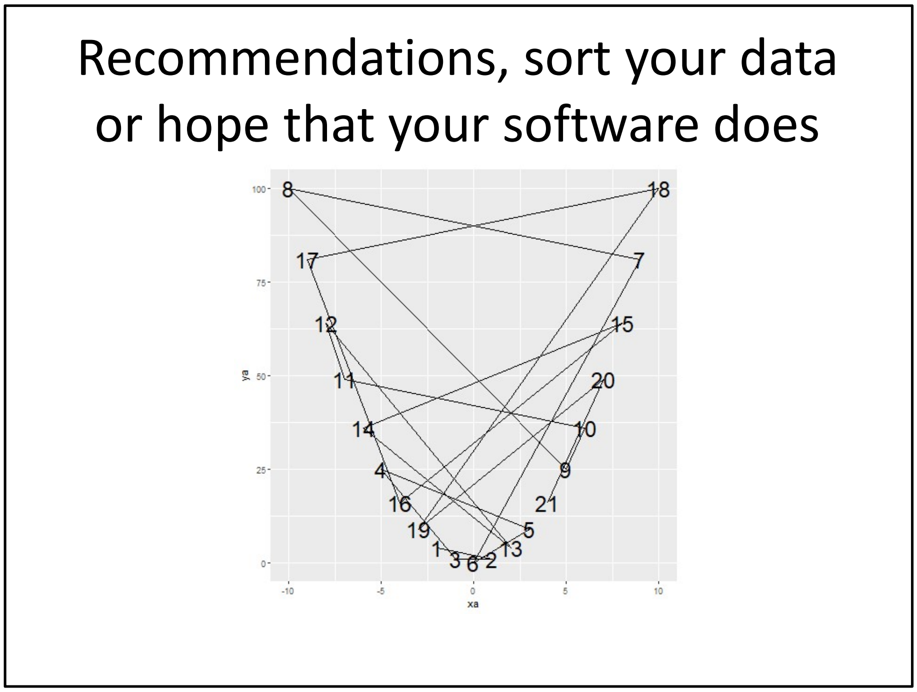
Scatterplot with a 1:2 aspect ratio

This graph has an aspect ratio of 1 to 2. It is twice as tall as it is wide. Which one is better?



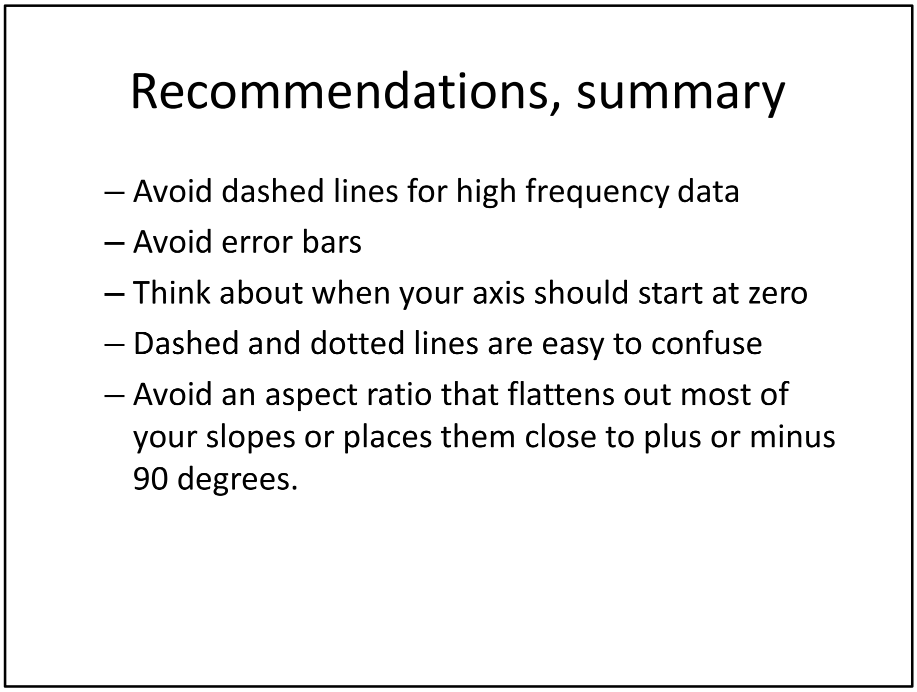
Recommendations for aspect ratio

- When changes in slope are important
  - Very flat lines (angle close to zero) are difficult to compare
  - Very steep lines (angle close to plus or minus 90 degrees) are also difficult to compare
  - Size your graph so that most lines have angles of plus or minus 45 degrees.
- Use square graph when comparing measurements of the same thing
  - Predicted versus actual
  - New lab method versus gold standard method



Recommendations, sort your data or hope that your software does

Another warning, that is not really relevant to these three graphics packages, but which you should look out for is whether a system plots a line in the order in which the data points appear or in the order in which the X-values occur. If you are not careful, you could end up with a graph like the one shown here.



Recommendations, summary

– Avoid dashed lines for high frequency data

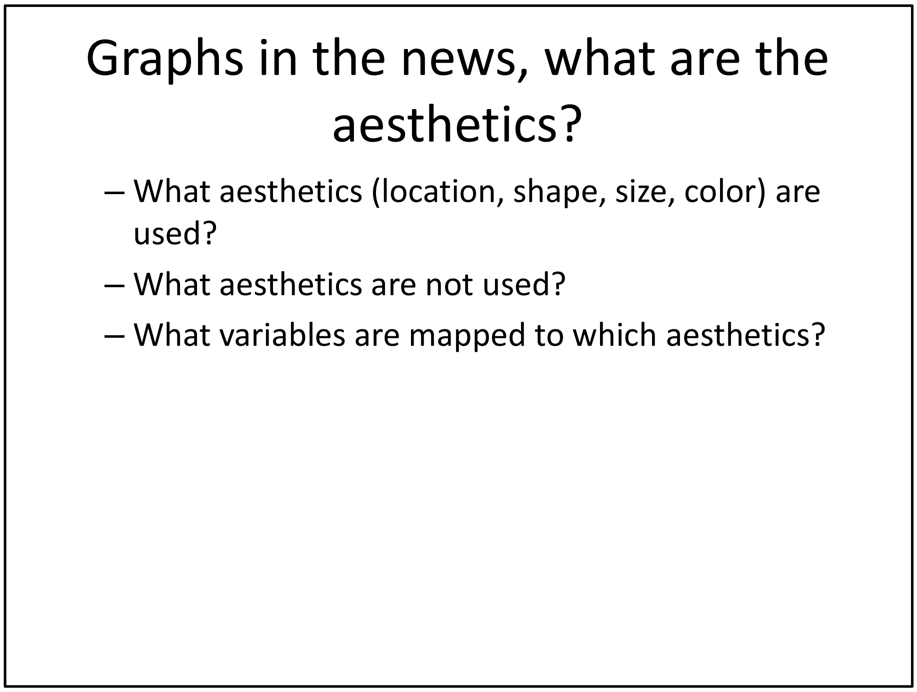
– Avoid error bars

– Think about when your axis should start at zero

– Dashed and dotted lines are easy to confuse

– Avoid an aspect ratio that flattens out most of your slopes or places them close to plus or minus 90 degrees.

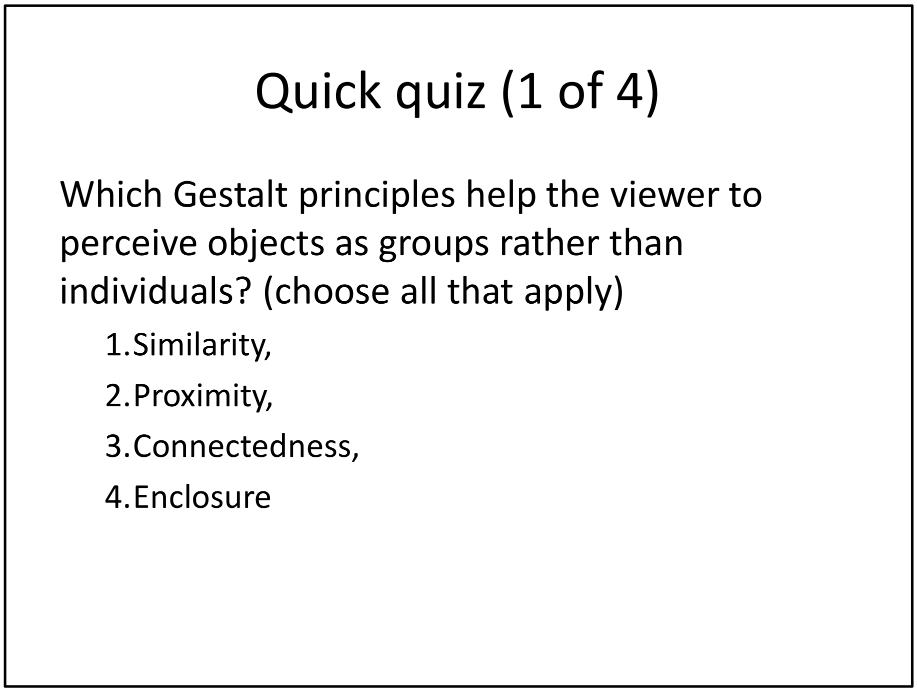
Here are the main recommendations. High frequency events can sometimes be missed by the gaps between dashes in a dashed line. Error bars are ambiguous and confusing. Sometimes it makes sense to start your y-axis at zero, and sometimes it doesn’t. A lot of different dashed and dotted lines just makes things chaotic and confusing. The aspect ratio of your graph should aim for slopes of roughly plus or minus 45 degrees. If your slopes are mostly flat or if they are nearly vertical (close to plus or minus 90 degrees) then your graph will be hard to interpret.



Graphs in the news, what are the aesthetics?

- What aesthetics (location, shape, size, color) are used?
- What aesthetics are not used?
- What variables are mapped to which aesthetics?

Review the same newspaper article and graph that you used earlier. Identify the various aesthetics that are used or are not used in your graph. What variables are associated with which aesthetics?

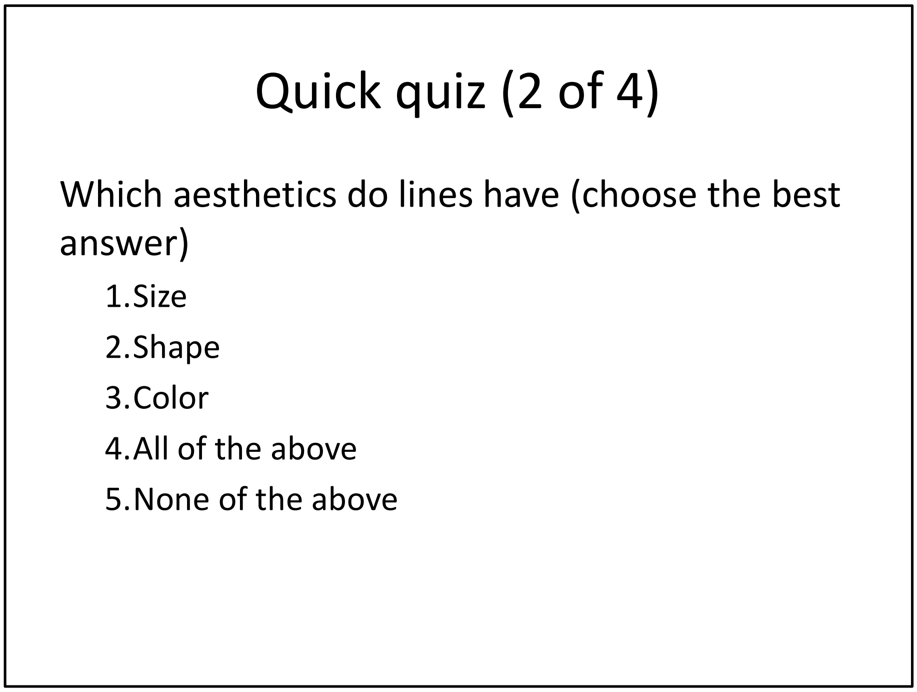


Quick quiz (1 of 4)

Which Gestalt principles help the viewer to perceive objects as groups rather than individuals? (choose all that apply)

1. Similarity,
2. Proximity,
3. Connectedness,
4. Enclosure

Answer: All of the above

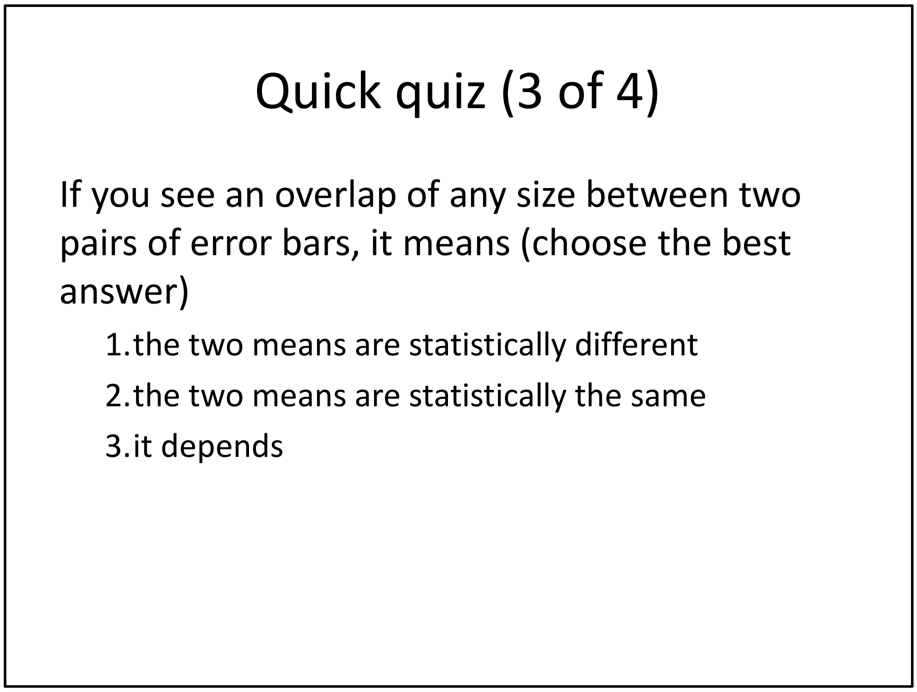


Quick quiz (2 of 4)

Which aesthetics do lines have (choose the best answer)

1. Size
2. Shape
3. Color
4. All of the above
5. None of the above

Answer: 4 (all of the above)

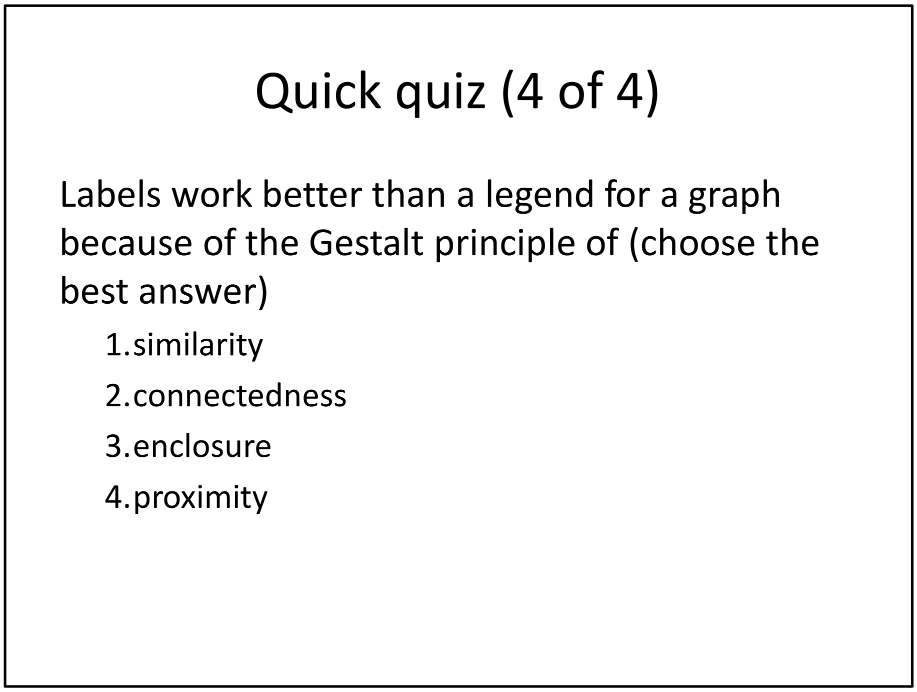


Quick quiz (3 of 4)

If you see an overlap of any size between two pairs of error bars, it means (choose the best answer)

1. the two means are statistically different
2. the two means are statistically the same
3. it depends

Answer: 3 (it depends)

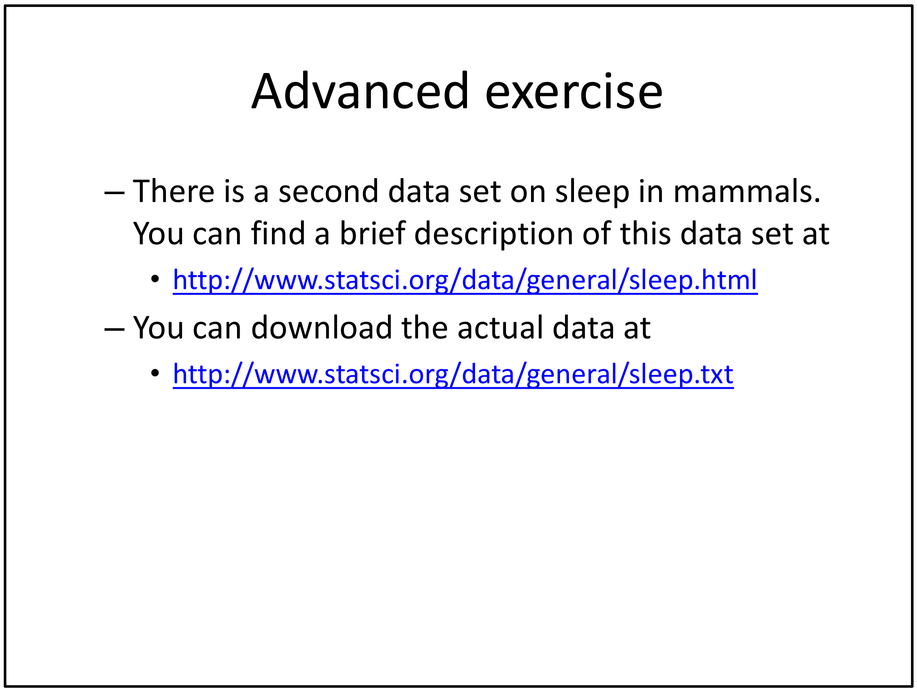


Quick quiz (4 of 4)

Labels work better than a legend for a graph because of the Gestalt principle of (choose the best answer)

1. similarity
2. connectedness
3. enclosure
4. proximity

Answer: 4 (proximity)



Advanced exercise

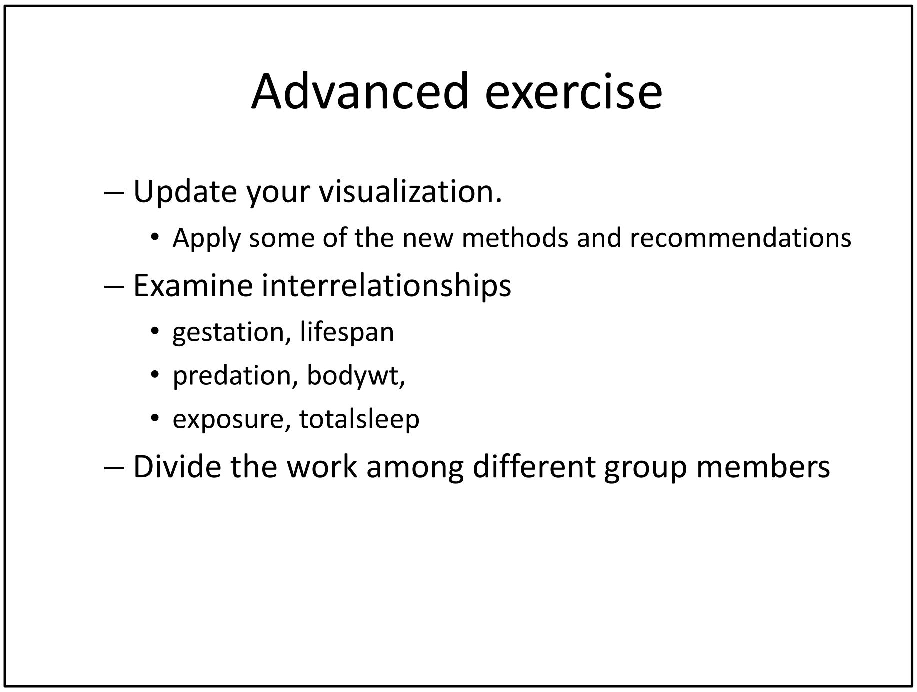
– There is a second data set on sleep in mammals. You can find a brief description of this data set at

  - http://www.statsci.org/data/general/sleep.html

– You can download the actual data at

  - http://www.statsci.org/data/general/sleep.txt

Continue your work on this data set. Review the description of each variable on the website shown here.

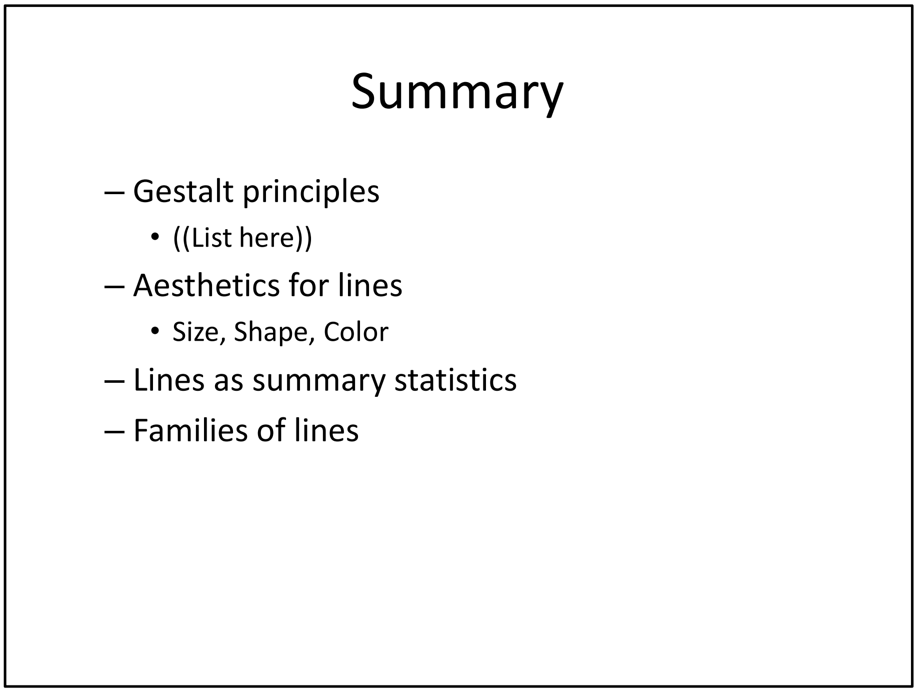


Advanced exercise

- – Update your visualization.
  - Apply some of the new methods and recommendations
- – Examine interrelationships
  - gestation, lifespan
  - predation, bodywt,
  - exposure, totalsleep
- – Divide the work among different group members

For a final assignment in this section, take the same data set on sleep in mammals. Rework and update your visualizations. Use some of the principles and recommendations presented in this section. Compare line graphs to the scatterplots that you used earlier. Which works better?

You can examine any interrelationships, but three possible areas for exploration are the relationships between gestation and lifespan, predation and bodywt, and exposure and totalsleep. Feel free to incorporare an additional variable of interest in your visualization that you feel might be relevant.



Summary

- – Gestalt principles
  - ((List here))
- – Aesthetics for lines
  - Size, Shape, Color
- – Lines as summary statistics
- – Families of lines