Statistical Computing and Data Visualization in R

Lecture 7

Case Studies

Case study: Combining Time Series and Part-to-Whole Relationships

- Part-to-whole data refers to values that combine to form a whole compare to one another and the whole ➔ Dot/bar/pie charts usually preferred.
- Time series data refers to how values evolve over time ➔ Line plots are generally the best solution.
- In combining time series and part-to-whole data we aim at showing how both the individual parts and the whole evolve over time.
  - Time series of sales broken down by district.
  - Weekly box-office receipts from different movies.

An example: Sales data

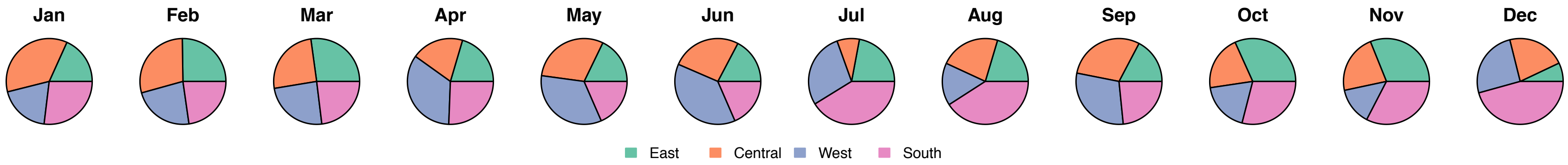
- Consider the following table containing the dollar value of sales made by four divisions of a firm:

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
East	64,162	70,172	93,657	55,056	63,631	53,040	51,974	63,272	54,110	112,284	69,697	27,437
Central	125,392	80,851	87,645	52,996	107,159	80,349	19,907	70,431	92,568	72,961	49,881	85,084
West	67,364	63,742	83,856	92,412	120,284	116,618	66,692	49,671	92,920	65,971	31,516	99,000
South	94,572	63,234	79,491	68,963	65,868	56,659	97,101	126,879	73,240	102,589	73,044	177,943
Total	$351,490	$277,999	$344,649	$269,427	$356,942	$306,666	$235,674	$310,253	$312,838	$353,805	$224,138	$389,464

- What are natural questions to ask about this data?
  - Have total sales gone up or down?
  - Has the contribution of any particular division to total sales gone up or down?
- Hard to see from table, let's use a visualization!

Solution 1: Sequence of pie charts

- Pie charts are one way in which we can represent part-to-whole relationships …
- We can use a sequence of them to represent the evolution of sales over time.
- What is wrong with this visualization?

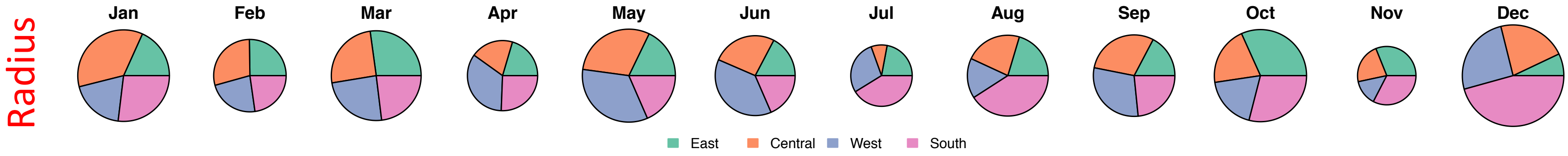


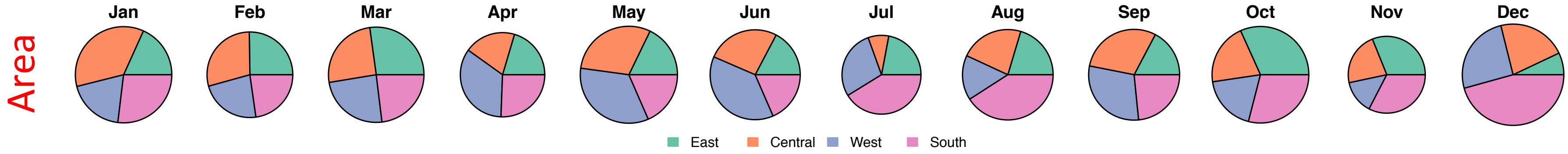
Solution 1: Sequence of pie charts

```
library(RColorBrewer)
sales             = read.table(file="sales.csv", sep=",", header=T, row.names=1)
totalsales.month = apply(sales,2,sum)
sales.cumsum      = apply(sales,2,cumsum)
colorscale        = brewer.pal(4, "Set2")

## Sequence of piecharts, equal size
quartz(width=12, height=1.2)
layout(mat=matrix(c(seq(1,12),rep(13,12)), nrow=2, ncol=12, byrow=T),
widths=rep(2,12), height=c(2,.4))
layout.show(13)
for(i in 1:12){
  par(mar=c(0,0,1,0))
  pie(sales[,i], labels="", col=colorscale, main=names(sales)[i])
}
par(mar=c(0,0,0,0))
plot(seq(1,5),seq(1,5), type="n", axes=F)
legend(3,3, legend=row.names(sales), fill=colorscale, border=colorscale, horiz=T,
bty="n", xjust=0.5, yjust=0.5)
```

Solution 2: Sequence of pie charts with total sales

- To slightly improve on the visualization, you could use the size of the pie to represent total sales …
- However, should you use the area or the radius of the pie?





Solution 2: Sequence of pie charts with total sales

```
## Sequence of piecharts, radius of the circles are proportional to total sales
quartz(width=12, height=1.2)
layout(mat=matrix(c(seq(1,12),rep(13,12)), nrow=2, ncol=12, byrow=T),
widths=rep(2,12), height=c(2,.4))
layout.show(13)
for(i in 1:12){
  par(mar=c(0,0,1,0))
  pie(sales[,i], labels="", col=colorscale, main=names(sales)[i],
radius=totalsales.month[i]/max(totalsales.month))
}
par(mar=c(0,0,0,0))
plot(seq(1,5),seq(1,5), type="n", axes=F)
legend(3,3, legend=row.names(sales), fill=colorscale, border=colorscale, horiz=T,
bty="n", xjust=0.5, yjust=0.5)
```

Radius of the pie for the month with the biggest sales is 1, rescale other radiuses accordingly

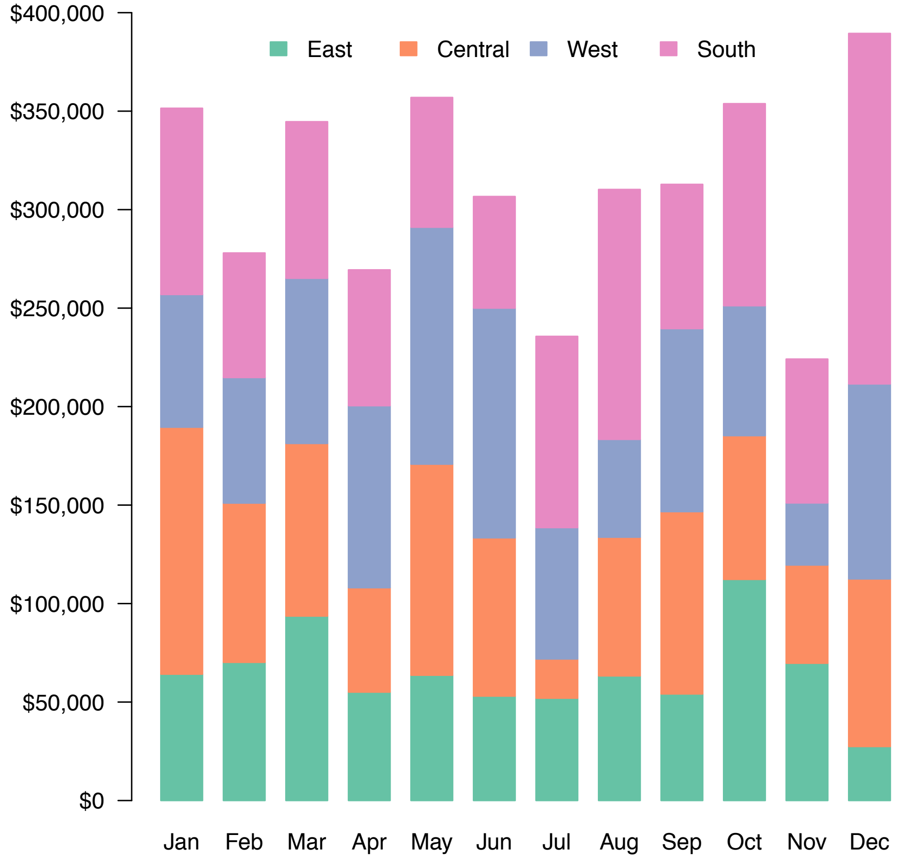
- Use this line instead if you want the **area** to be proportional to sales

```
pie(sales[,i], labels="", col=colorscale, main=names(sales)[i],
radius=sqrt(totalsales.month[i]/max(totalsales.month)))
```

Solution 3: Stacked bar plots

- Previously in the course we have emphasized that bar plots are preferable to pie charts:
  - The visual cues used by bar plots (length) is higher in the pre-attentive hierarchy than the visual cue used by pie charts (area/angle).
- Hence, it would be natural to consider stacked bar plots to represent a time series of part-to-whole data!

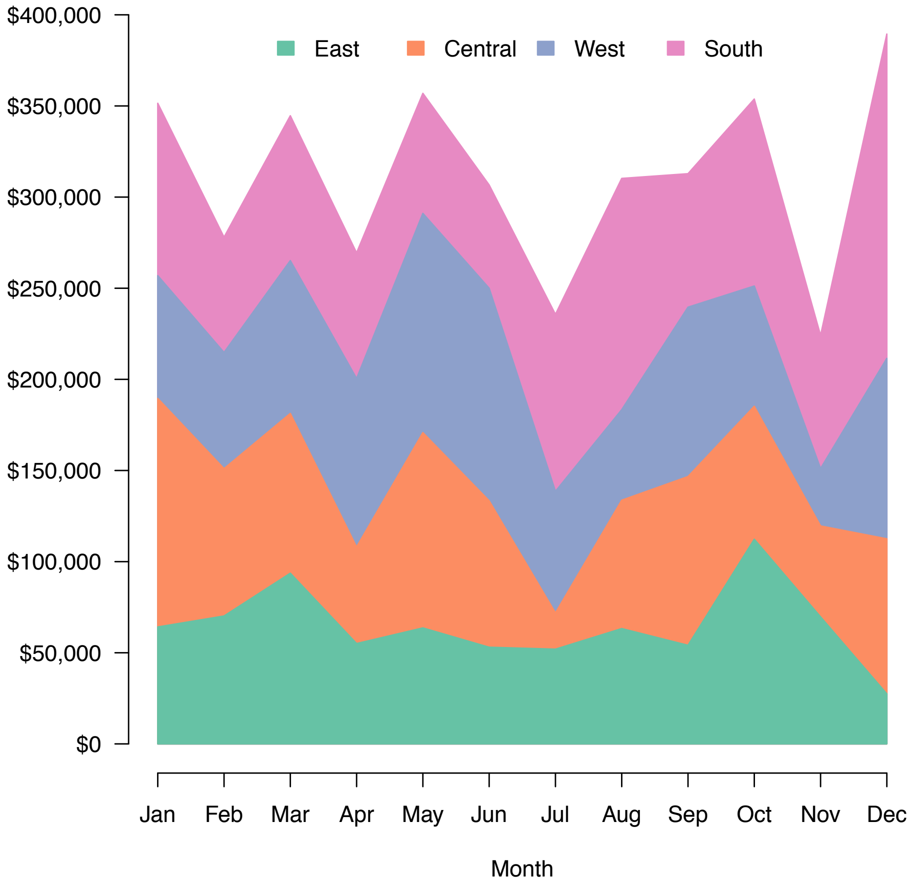
Solution 3: Stacked bar plots



Solution 3: Stacked bar plots

```
## Stacked bar plots
quartz()
par(mar=c(5,5,1,1))
barplot(as.matrix(sales), space=0.5, axes=F, col=colorscale,
border=colorscale, ylim=c(0,400000))
legend(9, 392000, legend=row.names(sales), fill=colorscale,
border=colorscale, horiz=T, bty="n", xjust=0.5, yjust=0.5)
axis(side=2, at=seq(0,400000,by=50000), labels=c("$0", "$50,000",
"$100,000", "$150,000", "$200,000", "$250,000", "$300,000", "$350,000",
"$400,000"), las=2)
```

Solution 4: Stacked area plots



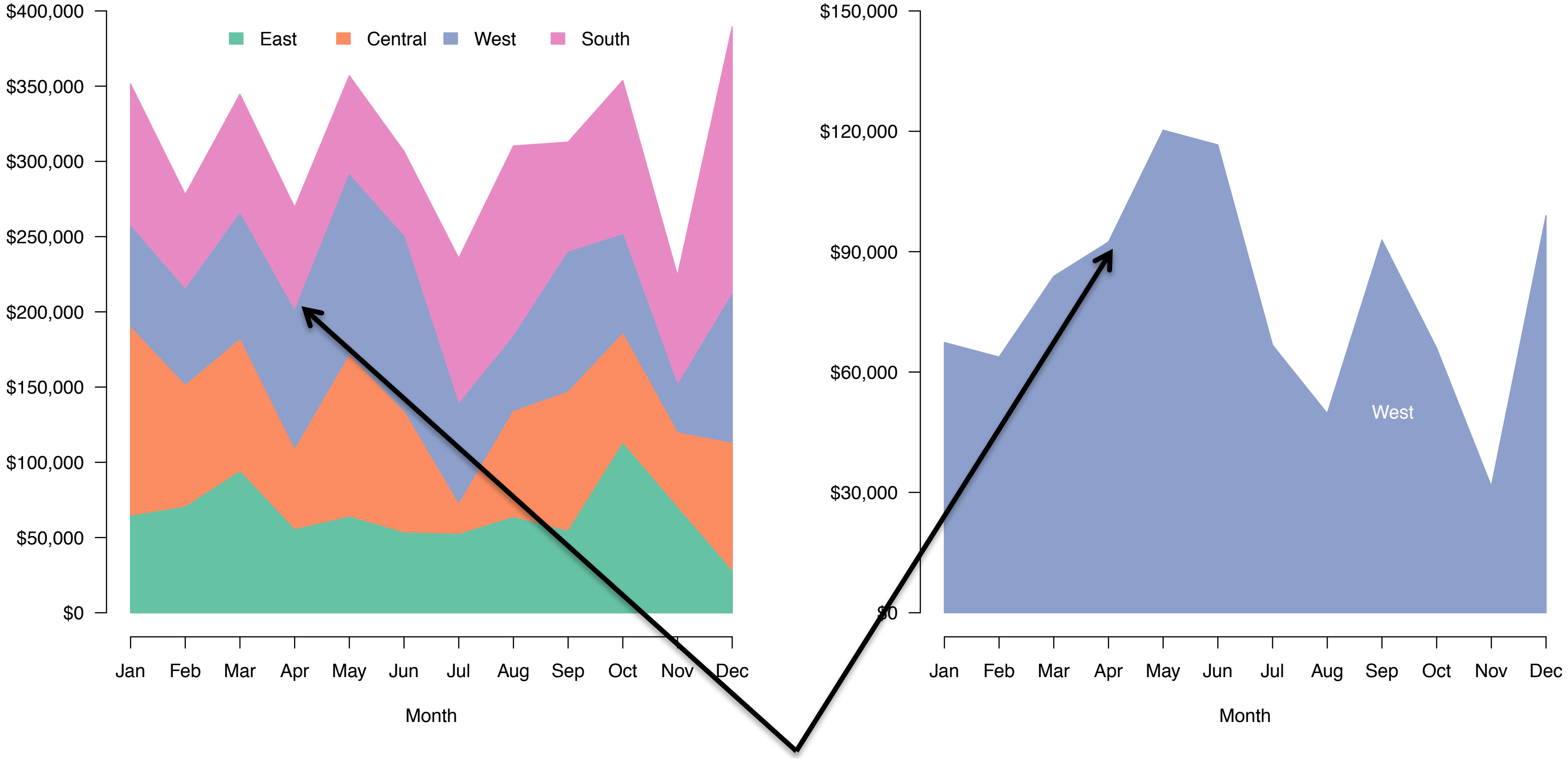
Solution 4: Stacked area plots

```
## Stacked area plots
quartz()
par(mar=c(5,5,1,1))
plot(seq(1,12), totalsales.month, axes=F, ylim=c(0,400000), type="n",
xlab="Month", ylab="")
for(i in rev(seq(1,4))){
  polygon(x=c(1,seq(1,12),12), y=c(0,sales.cumsum[i,],0),
col=colorscale[i], border=colorscale[i])
}
legend(6.5, 0.98*max(totalsales.month), legend=row.names(sales),
fill=colorscale, border=colorscale, horiz=T, bty="n", xjust=0.5, yjust=0.5)
axis(1, at=seq(1,12), labels=names(sales))
axis(side=2, at=seq(0,400000,by=50000), labels=c("$0", "$50,000",
"$100,000", "$150,000", "$200,000", "$250,000", "$300,000", "$350,000",
"$400,000"), las=2)
```

Solution 4: Stacked area plots

- Stacked area plots are a bit better than stacked bar plots to show time series of part-to-whole data.
- However, still suffer from a similar problem:
  - We can only accurately read the evolution of total sales and sales in the East region.
  - As before, the slope of the line for other regions depends on the slopes of the lines below it, so you really need to focus on the change in height of the area!!

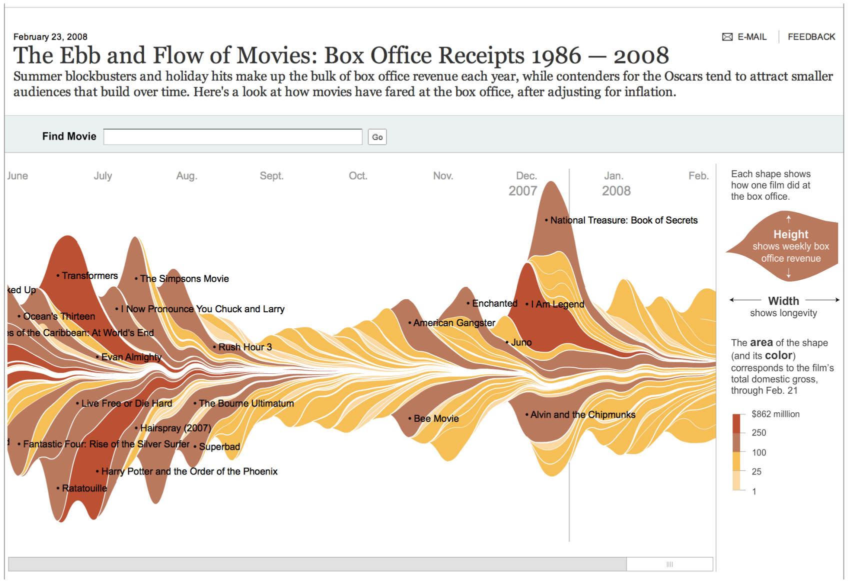
Solution 4: Stacked area plots



If you focus just on the top boundary of the area, you would think that sales went down in April, but that is not the case!!!

Solution 5: Stream graphs

- Steam graphs are a slightly more decorative version of stacked area plots!
- The one on the right corresponds to word usage in twitter over time.
- Note that, although prettier, the issues are the same (or worse)!



February 23, 2008

The Ebb and Flow of Movies: Box Office Receipts 1986 — 2008

Summer blockbusters and holiday hits make up the bulk of box office revenue each year, while contenders for the Oscars tend to attract smaller audiences that build over time. Here's a look at how movies have fared at the box office, after adjusting for inflation.

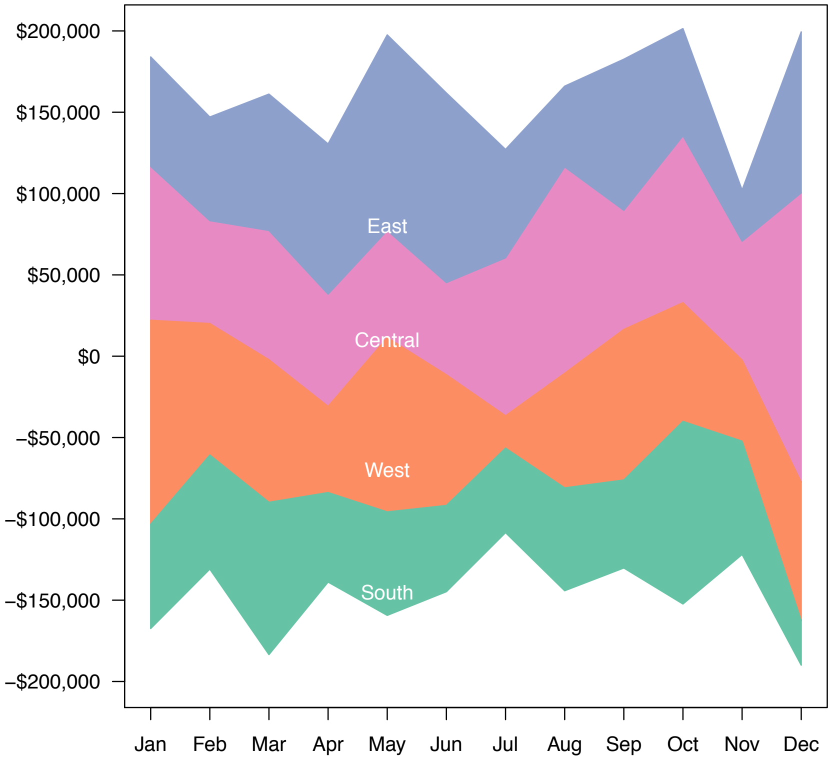
Each shape shows how one film did at the box office.

The area of the shape (and its color) corresponds to the film's total domestic gross, through Feb. 21

http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html?_r=0

Solution 5: Stream graphs

- There is no default function in R to construct stream graphs, but I am providing one in the next slide.
- Note that the stream plot is particularly bad in this setting
  - Negative number suggest losses!!).
  - We lose the few accurate comparisons we had!



Solution 5: Stream graphs

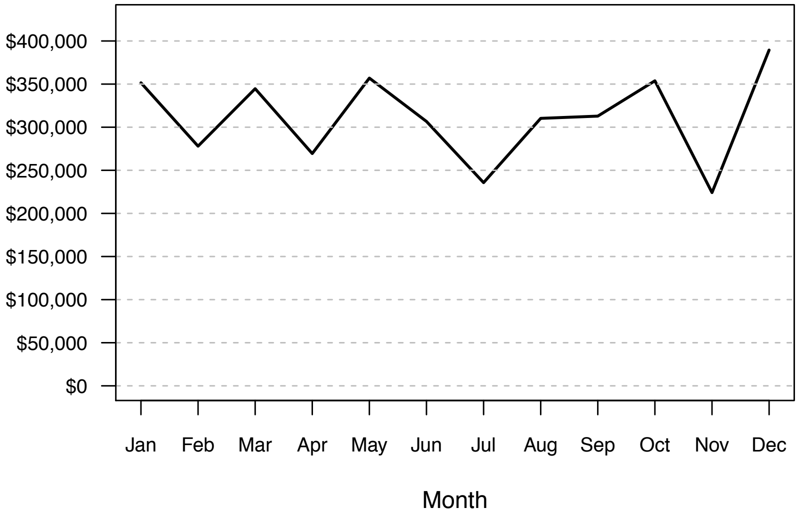
You will need the file streamplot.R!

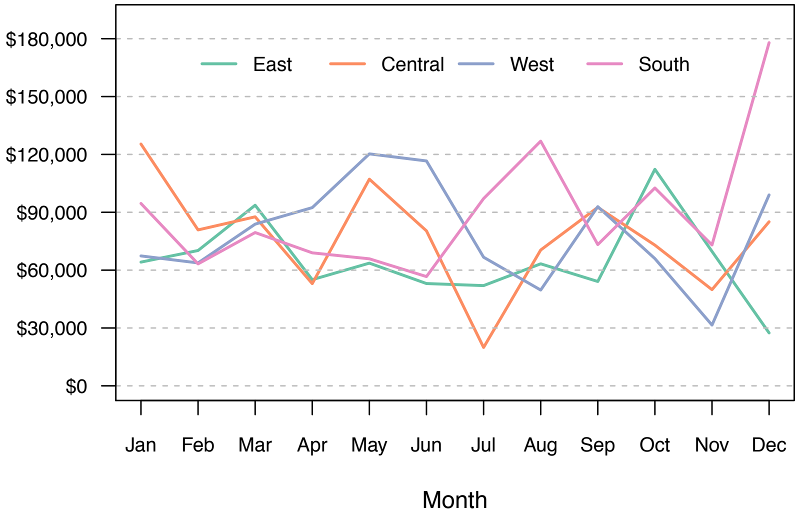
```
## Stream plots
source("streamplot.R")
quartz()
par(mar=c(5,5,1,1))
streamplot(x=seq(1,12), y=t(sales), col=colorscale, border=colorscale,
axes=F, ylim=c(-200000, 200000),order.method="min")
axis(1, at=seq(1,12), labels=names(sales))
axis(side=2, at=seq(-200000,200000,by=50000), labels=c("-$200,000", "-
$150,000", "-$100,000", "-$50,000", "$0", "$50,000", "$100,000",
"$150,000", "$200,000"), las=2)
box()
text(rep(5,4), c(-145000, -70000, 10000, 80000),
labels=rev(row.names(sales)), col="white")
```

Solution 6: Multiple line graphs

- Sometimes trying to pack too much information in a single graph is counterproductive!
  - Remember that the more multidimensional a graph is, the harder it is to understand.
- Rather than creating a single visualization, how about using a group of two (or maybe three) to display the information.

Solution 6: Multiple line graphs





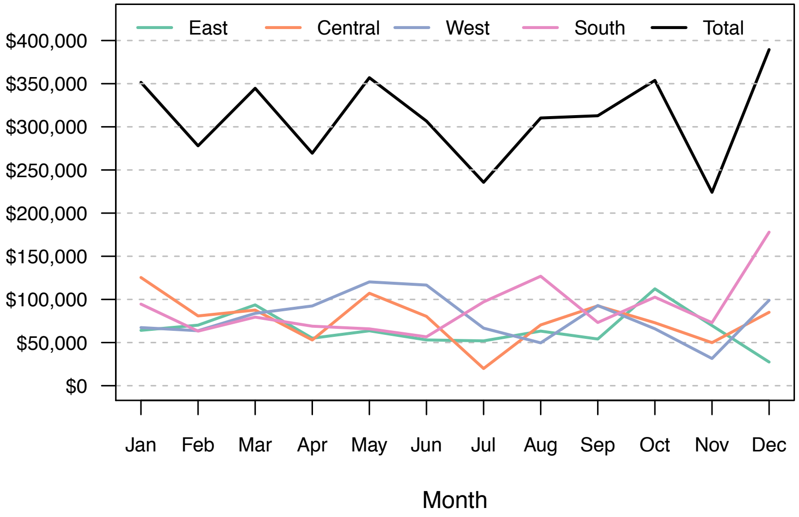
You can clearly see the evolution of sales (both total and on each market), and also how different markets have contributed to to total sales.

Solution 6: Multiple line graphs

```
## Total sales
quartz(height=4, width=6)
par(mar=c(5,5,1,1))
plot(seq(1,12), totalsales.month, type="l", xlab="Month", ylab="", axes=F, lwd=2, ylim=c(0,425000))
axis(1, at=seq(1,12), labels=names(sales), cex.axis=0.8)
axis(side=2, at=seq(0,400000,by=50000), labels=c("$0", "$50,000", "$100,000", "$150,000", "$200,000",
"$250,000", "$300,000", "$350,000", "$400,000"), las=2, cex.axis=0.8)
abline(h=seq(0,400000,by=50000), col="grey", lty=2)
box()


## Sales in each region
quartz(height=4, width=6)
par(mar=c(5,5,1,1))
plot(seq(1,12), totalsales.month, type="n", xlab="Month", ylab="", axes=F, lwd=2, ylim=c(0,190000))
for(i in 1:4){
  lines(seq(1,12), sales[i,], col=colorscale[i], lwd=2)
}
axis(1, at=seq(1,12), labels=names(sales), cex.axis=0.8)
axis(side=2, at=seq(0,180000,by=30000), labels=c("$0", "$30,000", "$60,000", "$90,000", "$120,000",
"$150,000", "$180,000"), las=2, cex.axis=0.8)
abline(h=seq(0,180000,by=30000), col="grey", lty=2)
box()
legend(6.35,167000, legend=row.names(sales), lwd=2, col=colorscale, horiz=T, bty="n", xjust=0.5,
yjust=0.5, cex=0.8)
```

Solution 6: Multiple line graphs

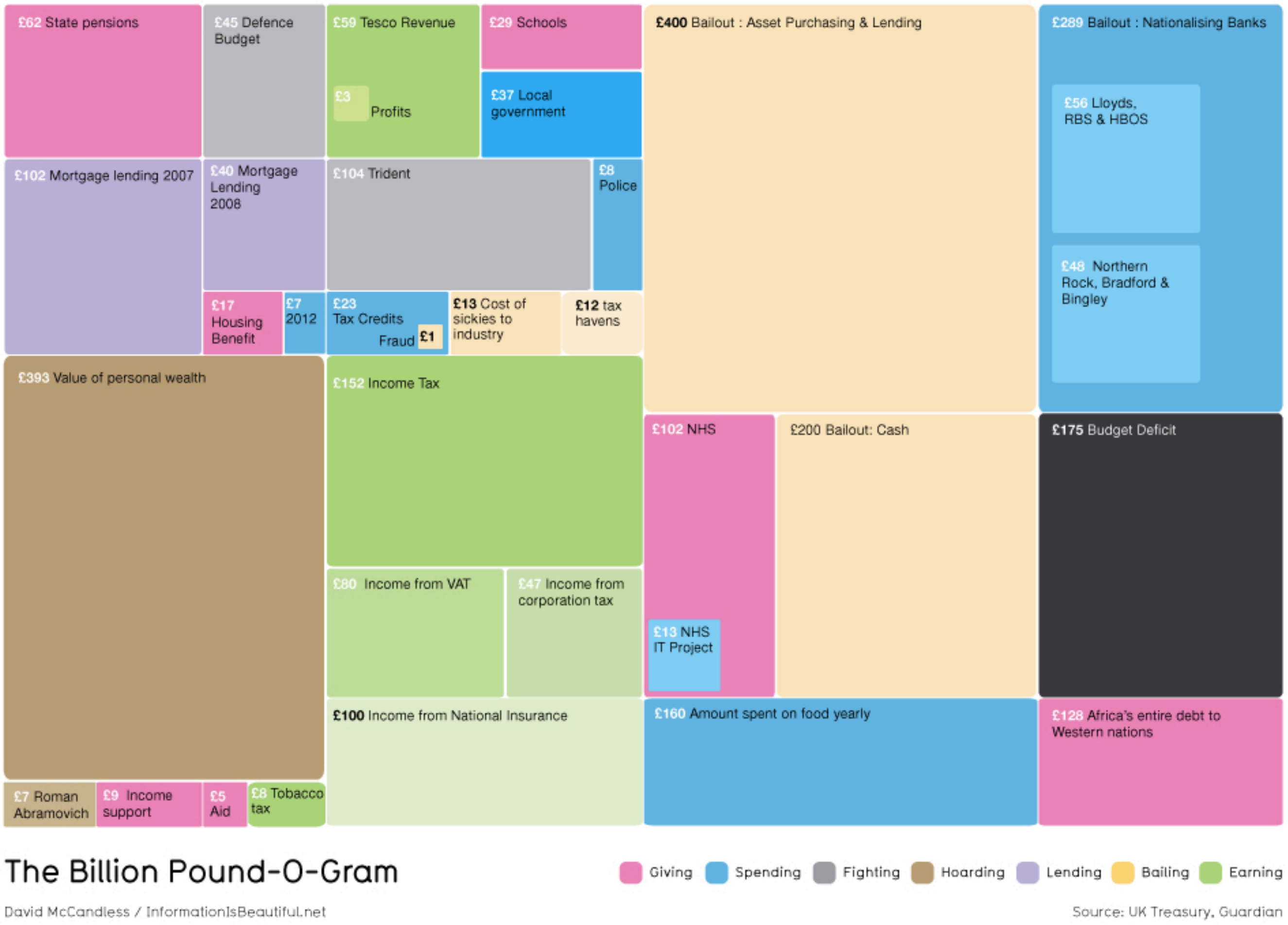
- In this case, because there are few parts, it would be possible to generate a single plot that combines total and regional sales.
- However, even in this simple case the aspect ratio for the regional sales is a poor one (most slopes are lower than 45°), so comparisons are harder.



Hierarchically classified data: The UK's budget deficit in perspective

- The UK's budget deficit in 2009 was projected to be 175 billion pounds
  - How much money is that?
  - How does it compare with the bank bailouts, or other government spending?
  - How does it compare with the wealth of some of the UK's richest men?

Solution 1: A tree map

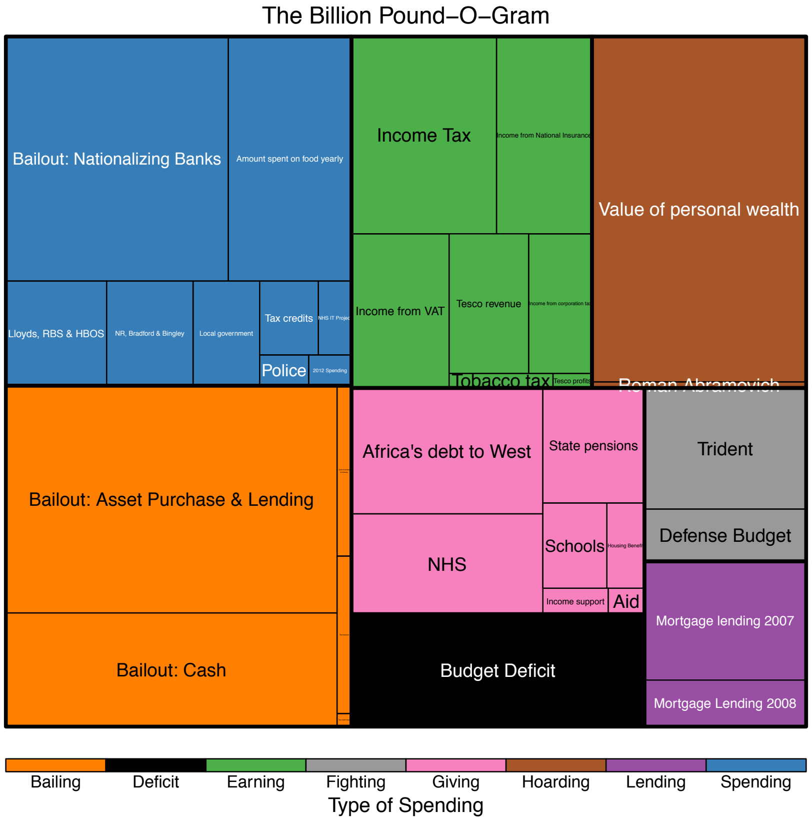


Solution 1: A tree map

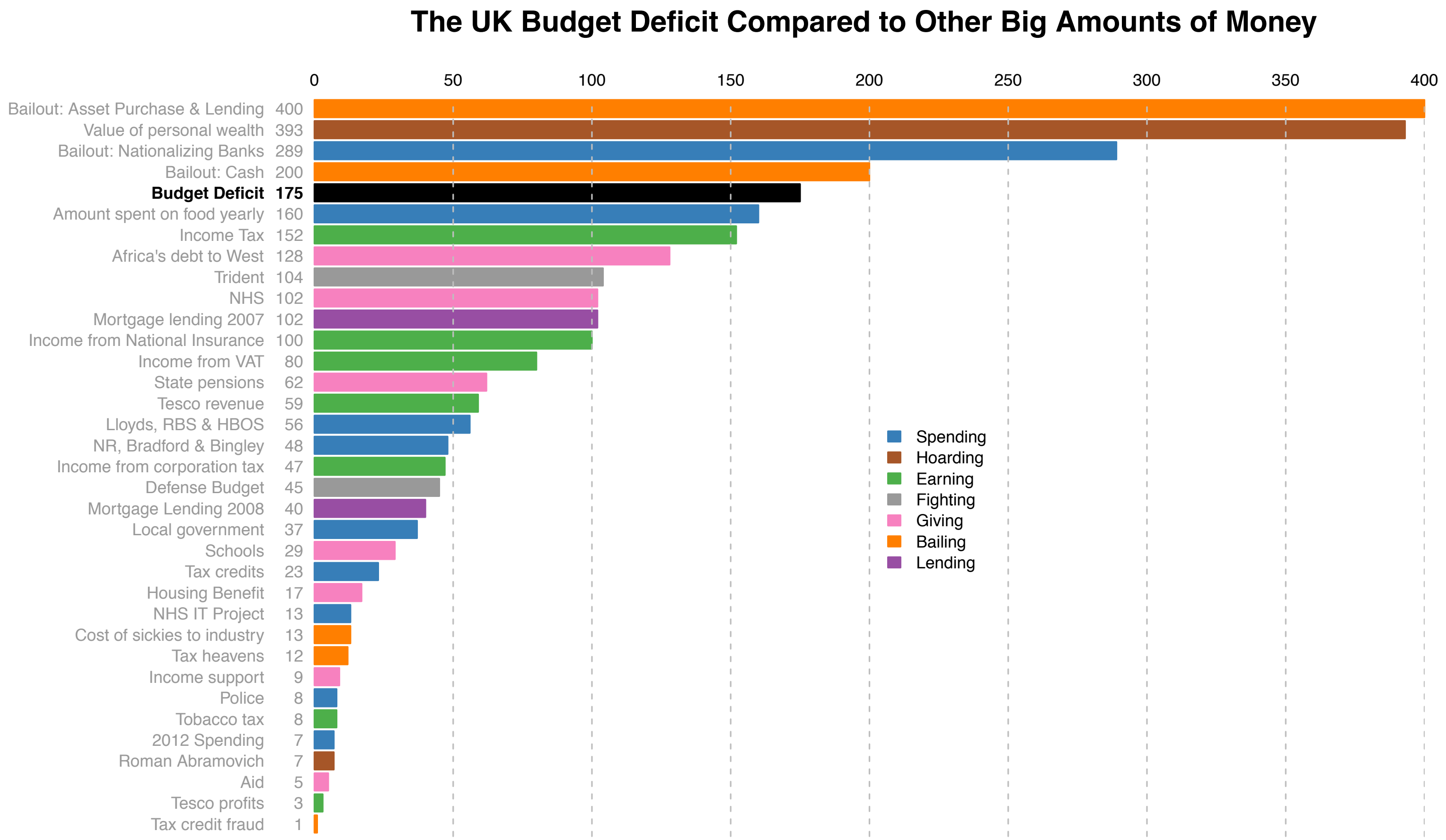
- Treemaps can be created in R using the `treemap()` command in the package treemap or the function `map.market()` in package portfolio.
- However, try answering some of these questions (without looking at the actual numbers!):
  - Which is bigger: Income Support or Police?
  - Which represents a larger amount: Mortgage Lending 2007 or NHS?
  - How much greater is Mortgage Lending 2007 than State Pensions?
  - Does State Pensions compared to Tesco Revenue look like the difference between 62 and 59?

Solution 1: A tree map

```
library(treemap)
library(RColorBrewer)
colorscale   = c(brewer.pal(9, "Set1")[5],
"black", brewer.pal(9, "Set1")
[c(3,9,8,7,4,2)])
deficit    =
read.table(file="billionpoundogram.csv",
sep=",", header=T)

quartz()
par(mar=c(4,4,1,1)+0.1)
treemap(deficit,index=c("Type","Area"),vSize=
"Amount",type="categorical", vColor="Type",
palette=colorscale, position.legend="bottom",
fontsize.labels = c(0,11),title.legend =
"Type of Spending", title="The Billion Pound-
O-Gram")
```



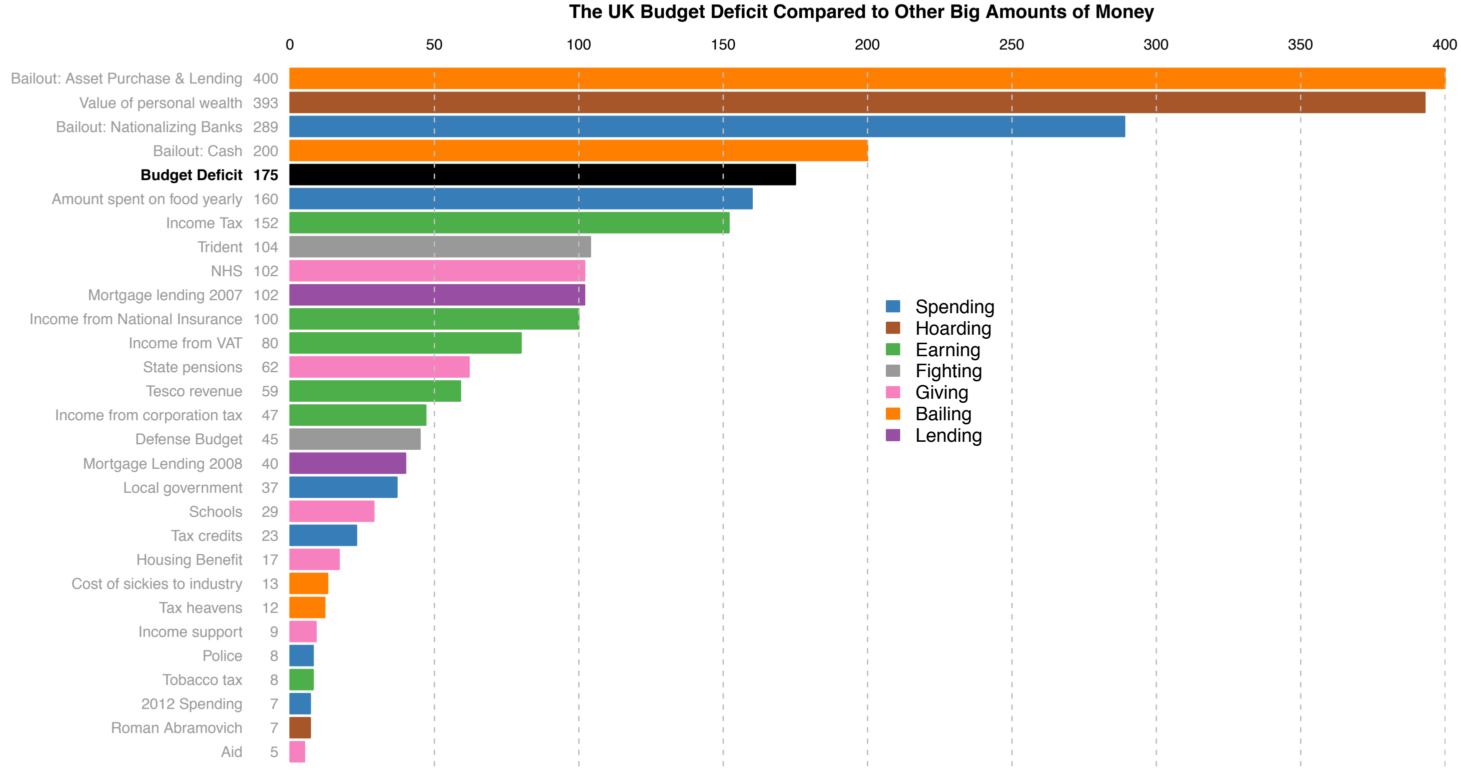
Solution 2: Bar plots



Solution 2: Bar plots

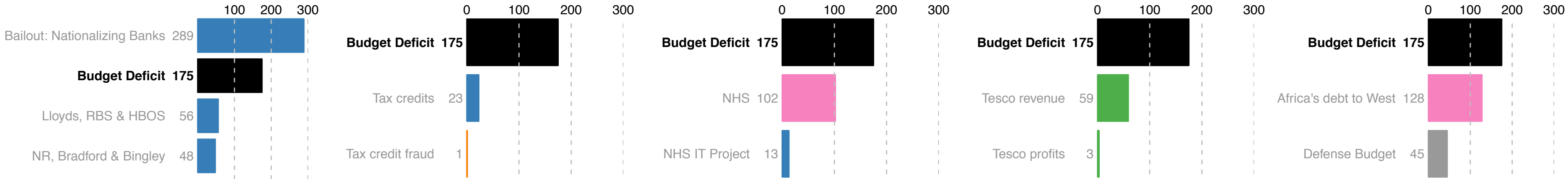
```
library(RColorBrewer)
colorscale = c("black", brewer.pal(9, "Set1")[c(2,7,3,9,8,5,4)])
deficit    = read.table(file="billionpoundogram.csv", sep=",", header=T, row.names=1)
ord        = order(deficit$Amount, decreasing=F)

## Barplot for the UK budget deficit
quartz(height=6, width=10)
par(mar=c(0,11,3,1)+0.3)
barplot(deficit$Amount[ord], horiz=T, col=colorscale[as.numeric(deficit$Color[ord])],
border=colorscale[as.numeric(deficit$Color[ord])], axes=F)
mtext(side=2, text=row.names(deficit)[ord], at=seq(from=0.7, by=1.2, length=35), las=2,
line=1.3, cex=0.70, col=c(rep(colorscale[5],30),"black",rep(colorscale[5],4)),
font=c(rep(1,30),2,rep(1,4)))
mtext(side=2, text=deficit$Amount[ord], at=seq(from=0.7, by=1.2, length=35), las=2,
line=0, cex=0.70, col=c(rep(colorscale[5],30),"black",rep(colorscale[5],4)),
font=c(rep(1,30),2,rep(1,4)))
mtext(side=3, text=seq(0,400,by=50), at=seq(0,400,by=50), cex.axis=0.8, line=-0.7,
cex=0.70)
legend(201, 24,
c("Spending","Hoarding","Earning","Fighting","Giving","Bailing","Lending"),
fill=colorscale[-1], border=colorscale[-1], bty="n",cex=0.70)
segments(x0=seq(50,400,by=50), y0=0, x1=seq(50,400,by=50), y1=42, col="grey", lty=2)
title(main="The UK Budget Deficit Compared to Other Big Amounts of Money", cex=0.70)
```

Solution 3: Combination graph



You could add text here where you make your point about the graphs.

Code in the file combined_barplots.R



The three blue bars could be replaced with a stacked bar plots.

Case study: Periodic Data

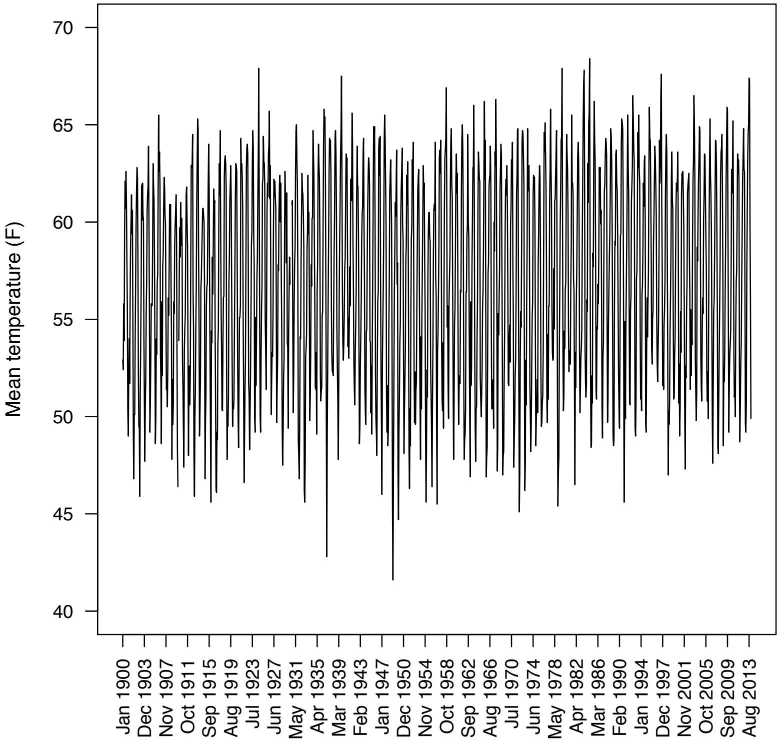
- Often we need to represent data that shows cycles (periodicities):
  - Weekly patterns (e.g., call volume, restaurant sales).
  - Monthly patterns (e.g., ATM transactions)
  - Annual patterns (e.g., temperature, precipitation).
- In some of this cases we might not know if such a pattern is present and we would like to better understand it.

Representing Periodic Data

- As an example, consider the average monthly temperatures in Santa Cruz county between 1900 and 2014.
- Some questions of interest:
  - What is the warmest month of the year?
  - What is the coldest month of the year?
  - Has Santa Cruz gotten warmer over the last 100 years? Has it grown colder?

Solution 1: A Line plot

- Given that this is a time series, the simplest, solution is to use a line plot.
- It conveys two messages well:
  - Clear periodicity.
  - Little or no overall trend.
- However, a more detailed exploration is hard because it is too cluttered.



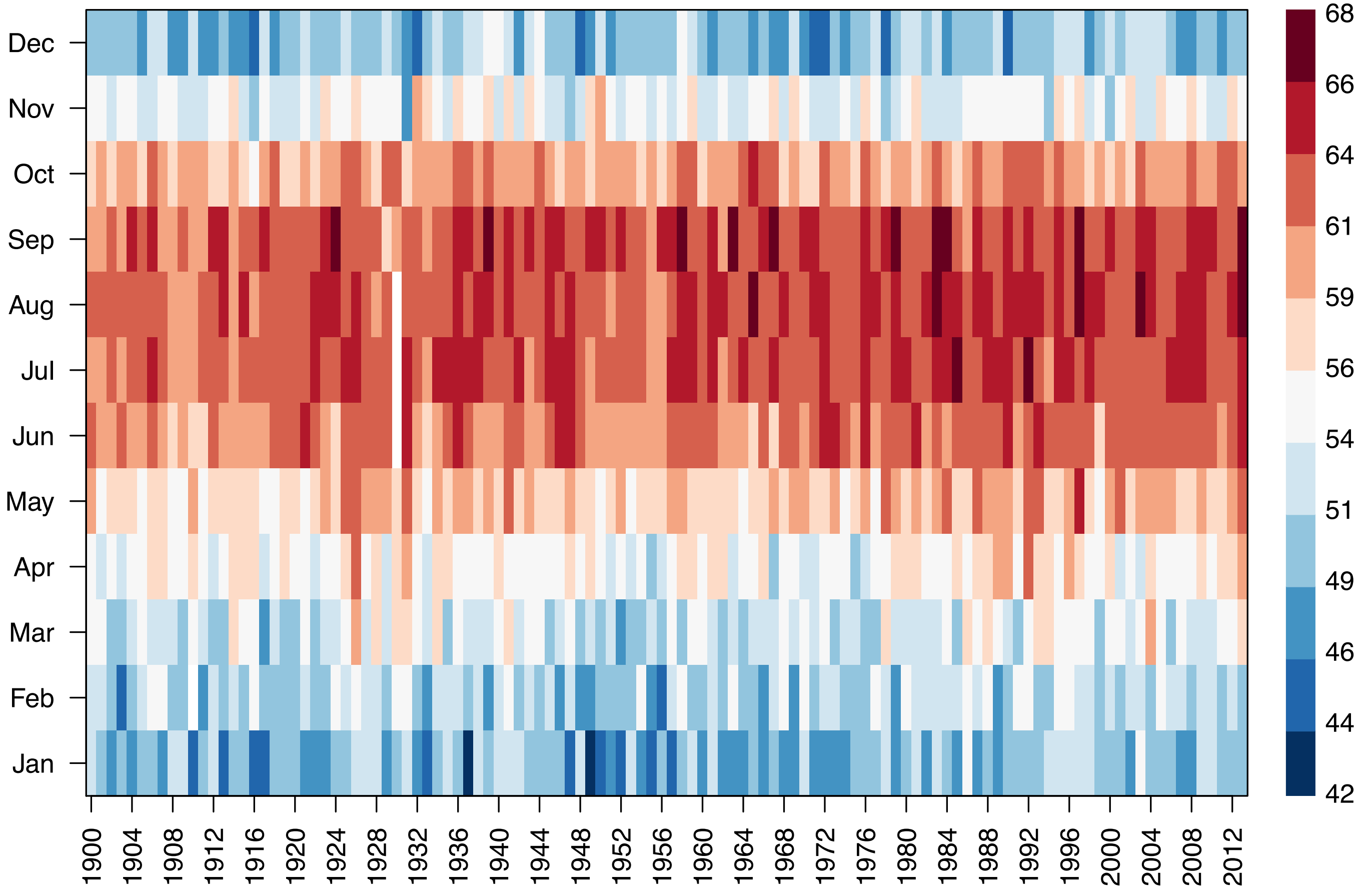
Solution 1: A Line plot

```
## Average monthly temperature in Santa Cruz
## Source
http://weather-warehouse.com/WeatherHistory/
PastWeatherData_SantaCruz_SantaCruz_CA_October.html

santacruztemperature = read.table(file="santacruztemperature.csv", sep=",", header=T,
row.names=1)

## Line plot
axislabels = paste( rep(names(santacruztemperature),114),
rep(rownames(santacruztemperature),each=12))
labelpos   = seq(1,114*12,by=47)
quartz()
par(mar=c(6,4,1,1)+01)
plot(as.vector(t(as.matrix(santacruztemperature))), type="l", ylab="Mean temperature
(F)", xlab="", axes=F, ylim=c(40,70))
axis(1, at=labelpos, labels=axislabels[labelpos], cex.axis=0.9, las=2)
axis(2, at=seq(40,70, by=5), cex.axis=0.9, las=2)
box()
```

Solution 2: Heat maps

- Heatmaps can be used to display arrayed-valued data.
  - In R, heat maps are created using the function `image()`.
  - Selecting the right color scale for a heat map is very important.
- When times series are very long and periodicity very clear, heat maps can be useful at conveying the big picture.

Solution 2: Heat maps



Solution 2: Heat maps

More flexible partitions
of a graphic display

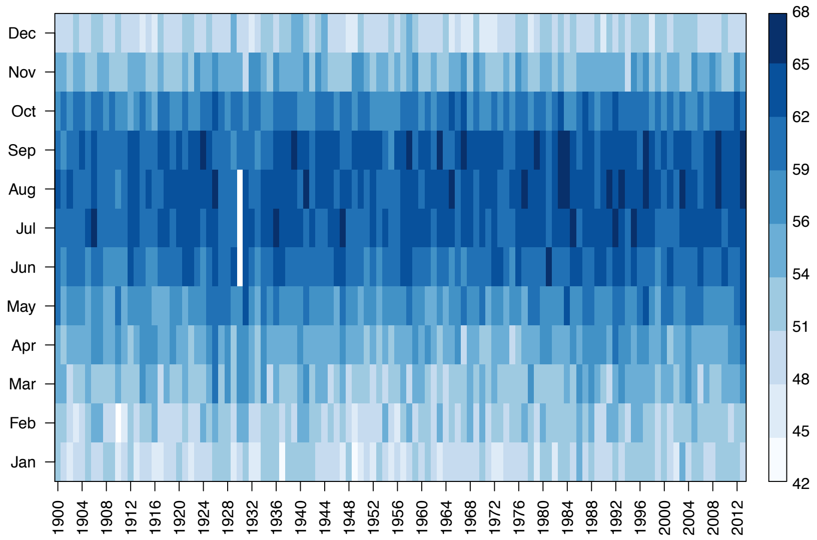
```
## Heatmap, blue to red
quartz(height=6, width=9)
layout(mat=matrix(c(1,2), nrow=1, ncol=2, byrow=T), widths=c(93,7), height=1)
layout.show(2)
par(mar=c(4,4,1,1)+0.1)
colorscale = rev(brewer.pal(11, "RdBu"))
image(seq(1,114), seq(1,12), as.matrix(santacruztemperature), col=colorscale, axes=F,
xlab="", ylab="")
axis(1, at = seq(1,114,by=4), labels=row.names(santacruztemperature)[seq(1,114,by=4)],
las=2)
axis(2, at = seq(1,12), labels=names(santacruztemperature),las=2)
box()
par(mar=c(4,0,1,2)+0.1)
plot(seq(1,5), seq(1,5), xlim=c(0,1), ylim=c(0,11) , axes=F, type="n", xlab="", ylab="")
for(i in 1:11){
  rect(xleft=0, ybottom=(i-0.5)*12/11, xright=1, ytop=(i-1.5)*12/11, col=colorscale[i],
border=colorscale[i])
}
mtext(text=round(seq(min(santacruztemperature,na.rm=T),max(santacruztemperature,na.rm=T)
,length=12),0), side=4, at=seq(-0.4,11.4,length=12), las=2, line=0.3)
```

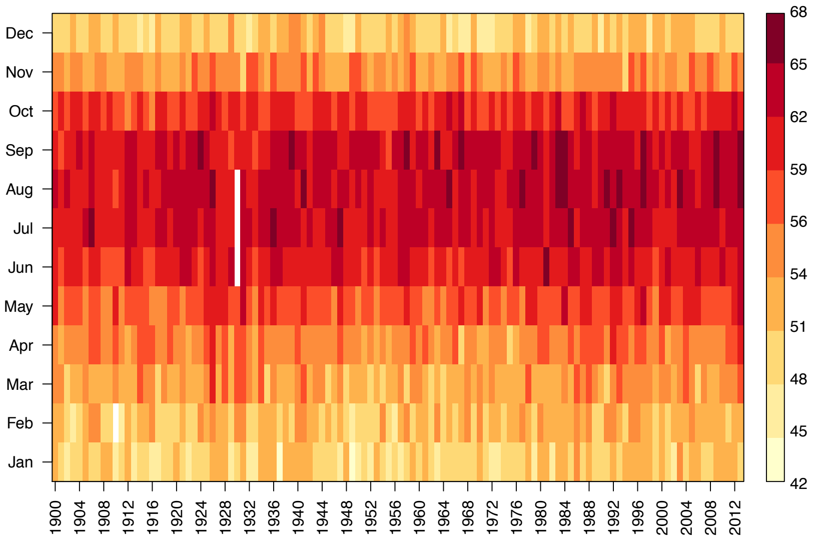
The legend is created
by tying a bunch of
rectangles together!

Solution 2: Heat maps

Although I feel about this way …

And people from outside California think of it this way …

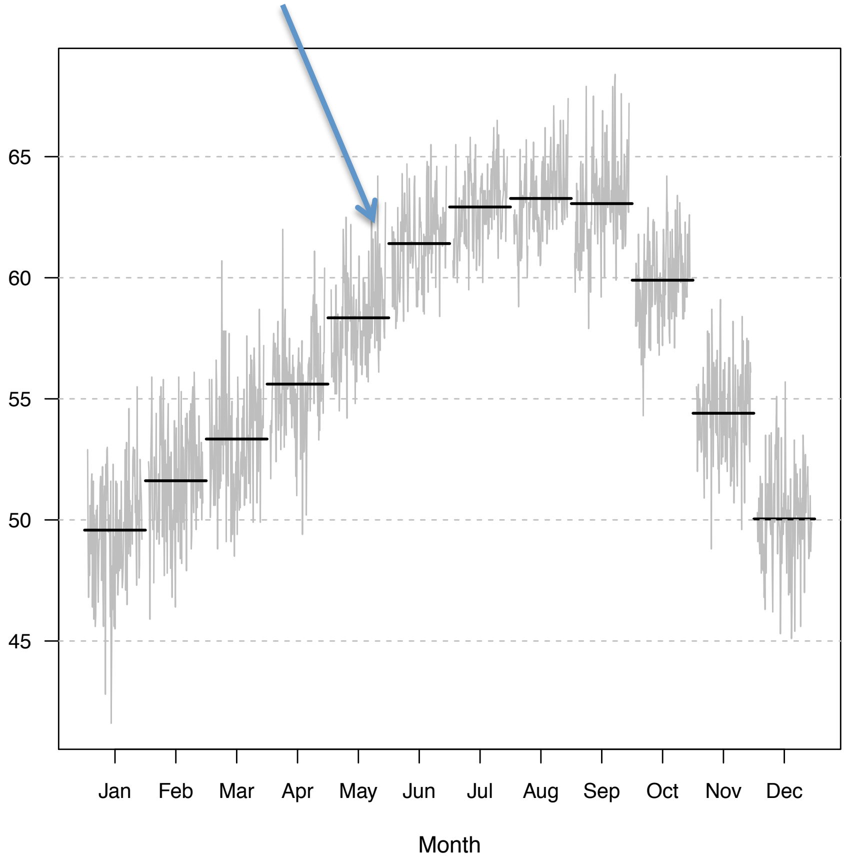




Solution 3: Cycle plots

- Cycle plots are a variant of line plots invented over 30 years, but rarely used.
- In cycle plots, the time series associated with each sub-period (in this example, each month) is plotted together, and context lines for the average temperature on each month are added to facilitate understanding of the periodicities.

Solution 3: Cycle plots



We could substitute the flat lines with a trend line

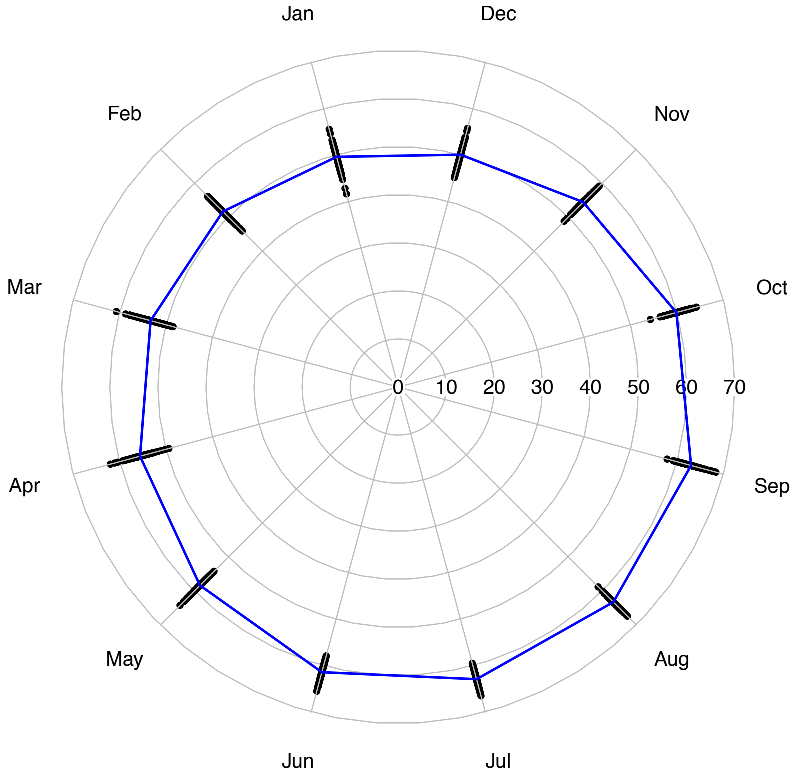
- From the graph, it is now very clear that August is the warmest month (by a slight margin) and that January is the coldest (also by a slight margin).
- Note also that temperatures fall fast, and recover more slowly.
- Warmer months seem to show a slight trend, colder months do not.

Solution 3: Cycle plots

```
## Cycle plot
meantemps = apply(santacruztemperature, 2, mean, na.rm=T)
quartz()
monthplot(t(as.matrix(santacruztemperature)), axes=F, ylab="Mean
temperature (F)", xlab="Month", col="grey")
axis(1, at=seq(1,12), labels=names(santacruztemperature), cex.axis=0.9)
axis(2, cex.axis=0.9, las=2)
for(i in 1:12){
  lines(c(0.5,1.5)+(i-1), c(meantemps[i], meantemps[i]), lwd=2)
}
abline(h=seq(45,65,by=5), col="grey", lty=2)
```

Solution 4: Polar coordinates plot

- Again, it is clear that mean temperatures are higher in September, but comparisons with August are a bit harder.
- It is apparent now that January has more variability than other months.
- We lose the ability to understand trend over multiple years.

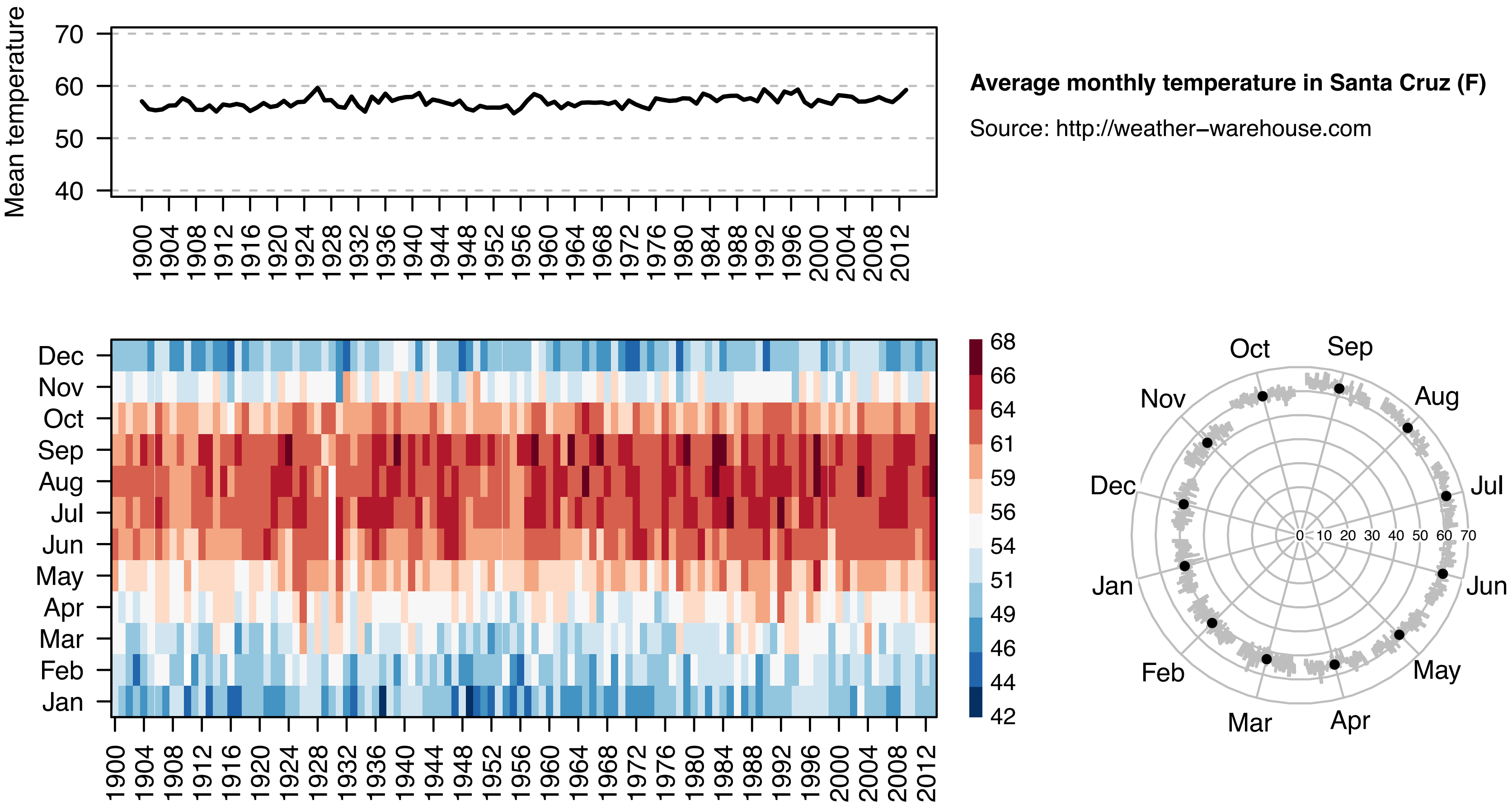


Polar coordinate plots make sense here because the data “month” is an ordinal variable with a cyclical rather than linear order.

Solution 4: Polar coordinates plot

```
## Polar coordinate plot, requires a special library that you need to
install before you can reproduce them.
library(plotrix)
quartz()
radial.plot(length=as.vector(t(as.matrix(santacruztemperature))),
radial.pos=rep(seq(0,2*pi,length=13)[-1],114)+5*pi/12,
labels=names(meantemps), label.pos=seq(0,2*pi,length=13)[-1]+5*pi/12,
radial.lim=c(0,70), line.col="blue", rp.type="s", lwd=2, point.symbols=20)
radial.plot(length=meantemps, radial.pos=seq(0,2*pi,length=13)[-1]+5*pi/12,
labels="", line.col="blue", radial.lim=c(0,70), rp.type="p", lwd=2, add=T)
```

Solution 5: Combined graph



Average monthly temperature in Santa Cruz (F)

Source: http://weather−warehouse.com

Solution 5: Polar cycle plot

```
quartz()
radial.plot(length=meantemps, radial.pos=seq(0,2*pi,length=13)[-1]+5*pi/12,
labels=names(meantemps), label.pos=seq(0,2*pi,length=13)[-1]+5*pi/12, line.col="blue",
radial.lim=c(0,70), rp.type="s", lwd=2, point.symbols=16)
radpos = seq(0,2*pi,length=13)[-1]+5*pi/12
for(i in 1:12){
  radial.plot(length=c(santacruztemperature[,i],rev(santacruztemperature[,i])),
radial.pos=c(seq(-pi/12+0.1*pi/6, pi/12-0.1*pi/6, length=114), rev(seq(-pi/12+0.07*pi/6,
pi/12-0.07*pi/6, length=114))) + radpos[i], labels="", line.col="grey",
radial.lim=c(0,70), rp.type="p", lwd=1, add=T)
}
radial.plot(length=meantemps, radial.pos=seq(0,2*pi,length=13)[-1]+5*pi/12,
line.col="blue", radial.lim=c(0,70), rp.type="s", lwd=2, point.symbols=16, add=T)
```

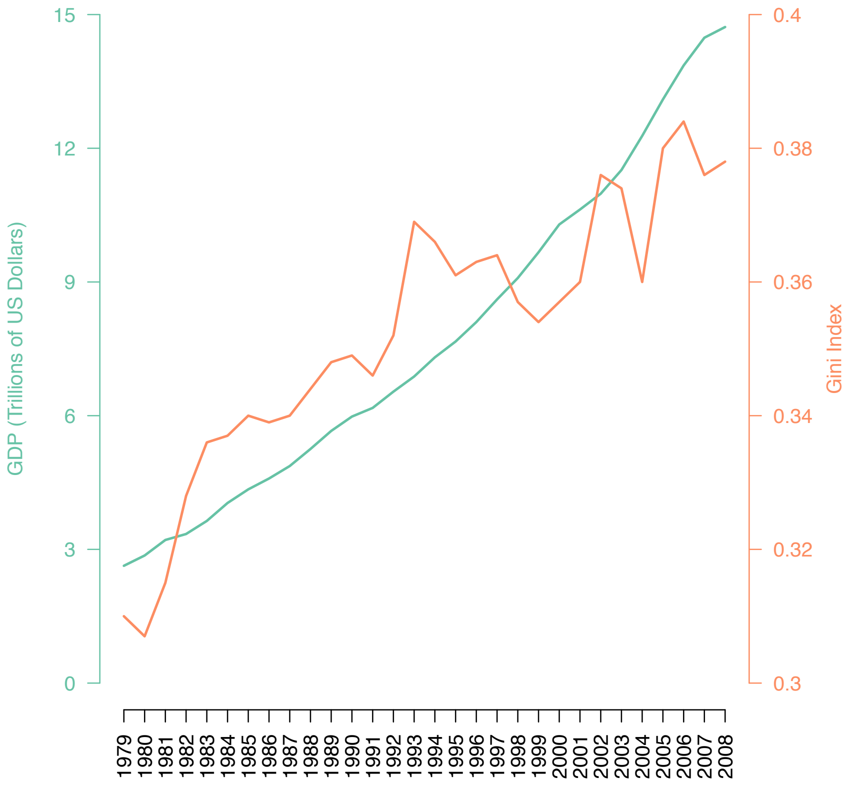
Case study: Two times series with different scales

- In many cases we are interested in looking at two (or more) time series simultaneously, with measurements being in different scales.
- Interest is both on what has happened to each variable over time and on what their association is.
- Example: Wealth (GDP) and inequality (Gini index):
  - Has the US become richer or more unequal. As it has grown richer, has it also become more unequal?
  - Data is relatively short: 1979-2008 (only 30 years).

Solution 1: Two line plots on the same graph

- Some people like to create a single graph with both time series.
- Because the two variables have very different scales, you need to have to scales on the y-axis.
- Even if you use colors cleverly, this can be extremely confusing.
  - Generally speaking, you should avoid it!

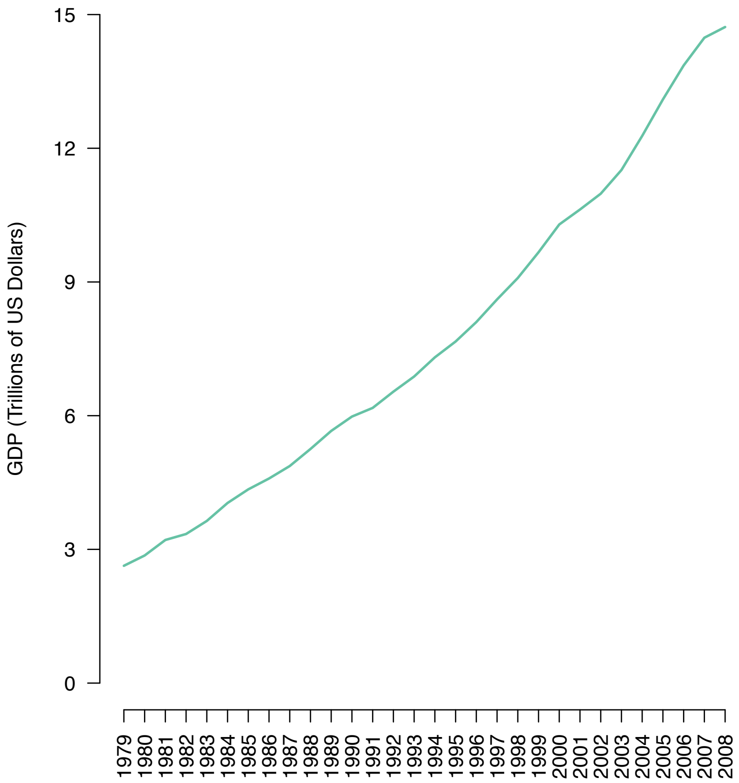
Both seem to go up with time, but the details of the relationship are not clear.

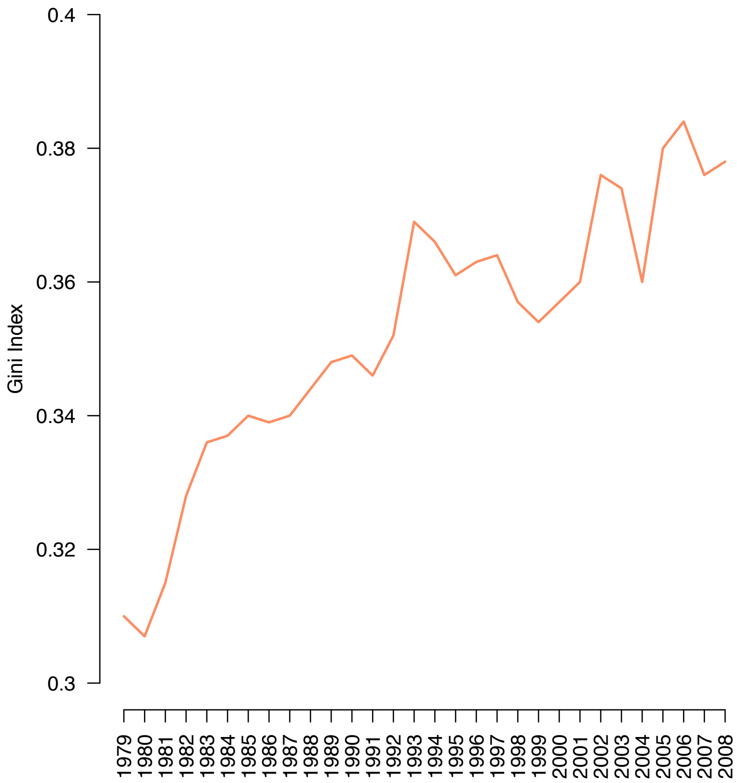


Solution 1: Two scales on a single graph

```
ushist     = read.table(file="usinequality.csv", sep=",", header=T)
colorscale = brewer.pal(3, "Set2")

## Two curves on a single graph
quartz()
par(mar=c(4,4,1,4)+0.1)
plot(ushist$Year, ushist$GDP/1000000, type="n", xlab="", ylab="", axes=F,
ylim=c(0,15))
axis(1, at=ushist$Year, labels=ushist$Year, las=2)
axis(2, at=seq(0,15,by=3), labels=rep("",6), las=2, col=colorscale[1],
col.ticks=colorscale[1])
mtext(side=2, at=seq(0,15,by=3), text=seq(0,15,by=3), las=2, col=colorscale[1],
line=1)
mtext(side=2, at=7.5, text="GDP (Trillions of US Dollars)", las=0,
col=colorscale[1], line=3)
lines(ushist$Year, ushist$GDP/1000000, col=colorscale[1], lwd=2)
axis(4, at=seq(0,15,length=6), labels=rep("",6), las=2, col=colorscale[2],
col.ticks=colorscale[2])
mtext(side=4, at=seq(0,15,length=6), text=seq(0.3,0.4,by=0.02), las=2,
col=colorscale[2], line=1)
mtext(side=4, at=7.5, text="Gini Index", las=0, col=colorscale[2], line=3)
lines(ushist$Year, 150*(ushist$Gini-0.3), col=colorscale[2], lwd=2)
```

Solution 2: Two line graphs

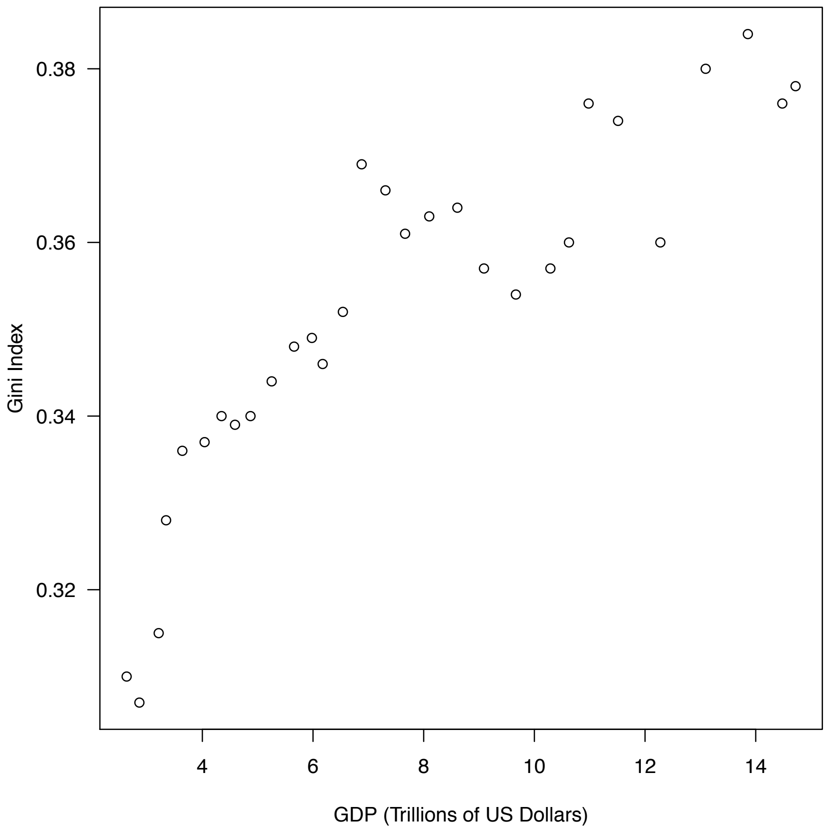




Relationship is still unclear!!!

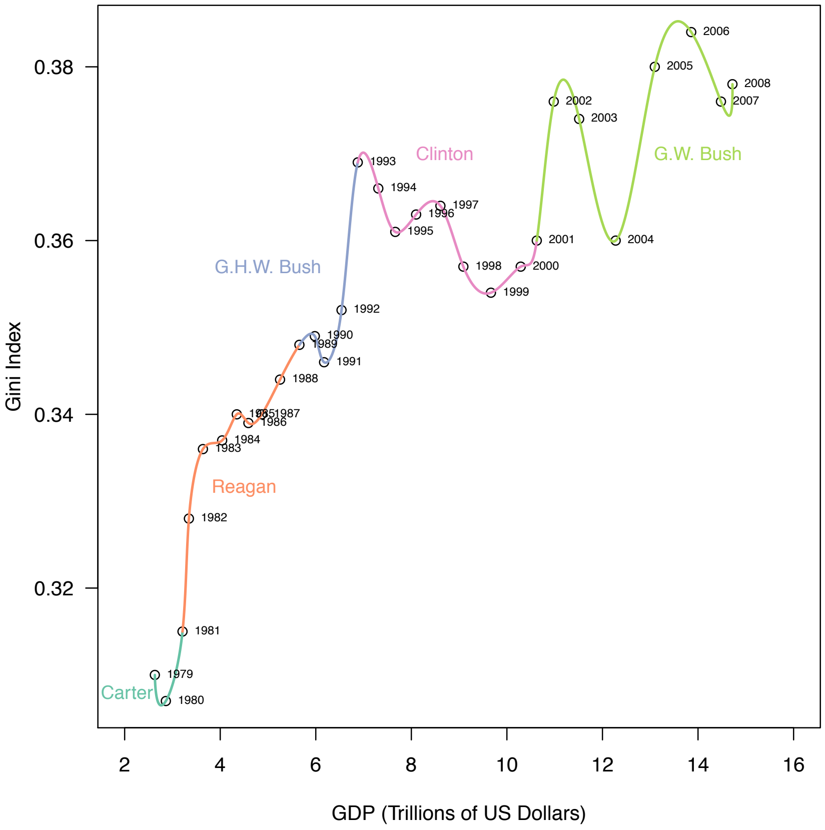
Solution 3: Scatterplot

- The natural graph to study association between pairs of variables is the scatterplot.
- It shows that the relationship seems to be more or less linear.
- However, information about the evolution over time of the two variables is lost!



Solution 4: "Interpolated" scatterplot

- Values are connected by a smooth line to provide information about the time series.
    - Curve mostly moves to the left (implying that GDP tended to only grow), while it goes up and down (meaning that the Gini index zigzagged a lot).



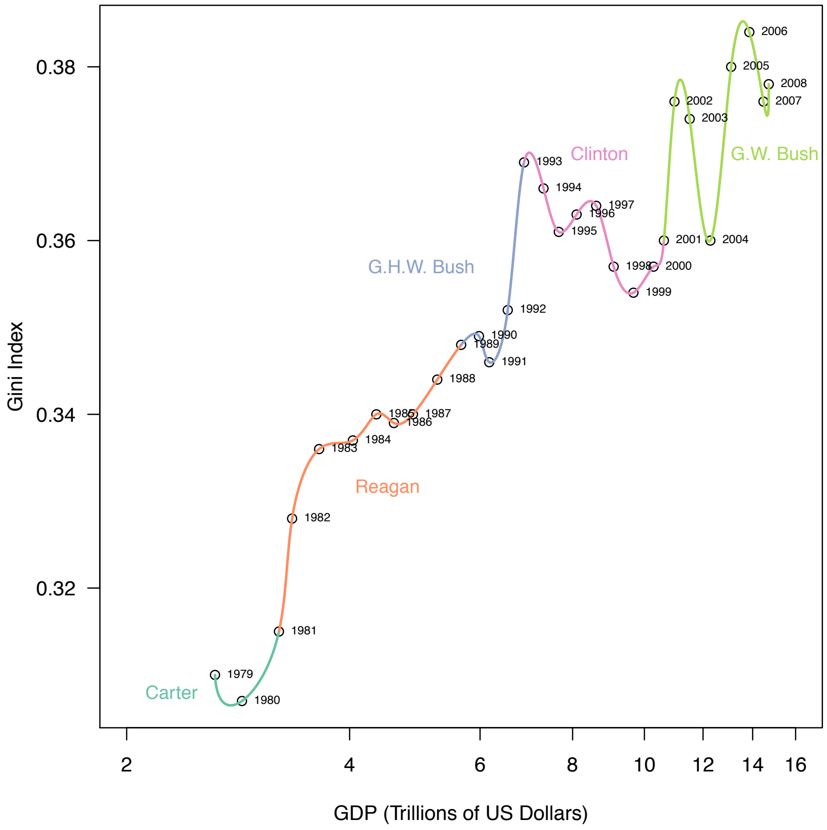
Solution 4: "Interpolated" scatterplot

```
##
colorscale = brewer.pal(5, "Set2")
quartz()
par(mar=c(4,4,1,1)+0.1)
plot(ushist$GDP/1000000, ushist$Gini, xlab="GDP (Trillions of US
Dollars)", ylab="Gini Index",las=1, xlim=c(2,16))
startdates = c(1979,1981,1989,1993,2001)
enddates   = c(1981,1989,1993,2001,2008)
for(i in 1:5){
  xx = spline(ushist$Year, ushist$GDP/1000000, n=500, method="fmm",
xmin=startdates[i], xmax=enddates[i])
  yy = spline(ushist$Year, ushist$Gini, n=500, method="fmm",
xmin=startdates[i], xmax=enddates[i])
  lines(xx$y, yy$y, lwd=2, col=colorscale[i])
}
text(ushist$GDP/1000000, ushist$Gini, labels=ushist$Year, cex=0.6,
pos=4)
text(c(2.05,4.5,5,8.7,14),c(0.308,0.3315,0.357,0.37,0.37),labels=c("Ca
rter","Reagan","G.H.W. Bush","Clinton","G.W. Bush"), col=colorscale,
cex=0.9)
```

Parameterized curve and use interpolating splines for each of the two dimensions

Solution 4: “Interpolated” scatterplot

- An improved version could also use a logarithmic scale for GDP (so that comparisons of GDP growth are in relative rather than absolute terms).
- It would be useful to have additional data!



Economic expansion was stronger in the Reagan years, and not as strong in the Bush junior years!

Case study: Representing many short time series

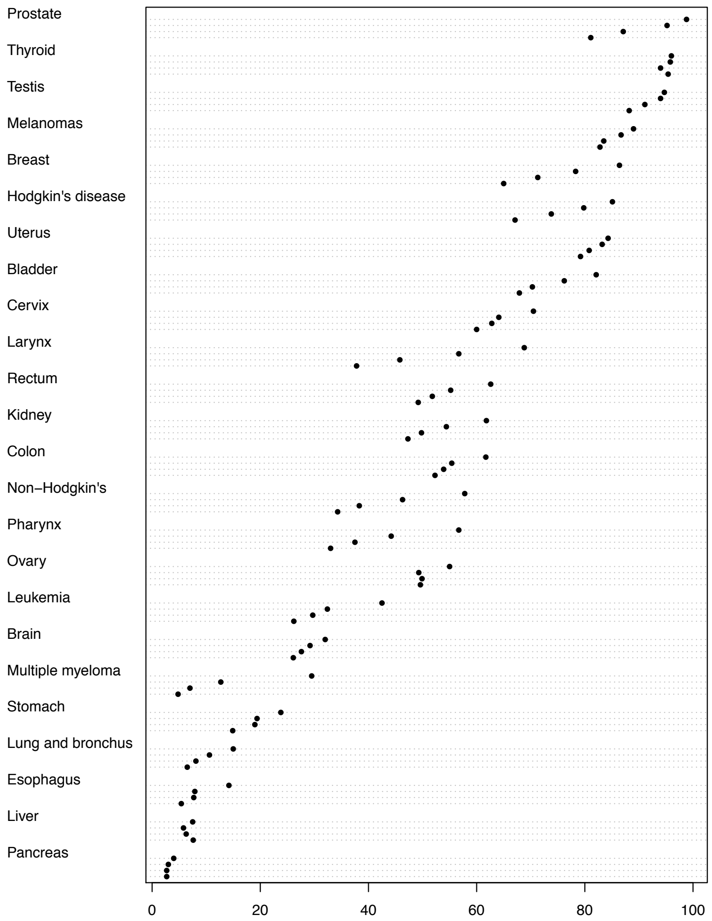
- As we discussed Line plots are the best solution for representing time series.
- Multiple time series are often represented using multiple lines in a single coordinate system.
- Traditional line plots are most helpful when representing a few long time series simultaneously.
  - For representing a large number of short time series (2 to 4 time points) a “slopegraph” is a helpful alternative.
  - For many long time series, a variant called “sparklines” can be used.

Multiple short time series

- Consider making a representation of the survival rates from cancer on different body sites after different number of years.
- How does prognosis change over time?

	5 years	10 years	15 years	20 years
Prostate	98.8	95.2	87.1	81.1
Thyroid	96	95.8	94	95.4
Testis	94.7	94	91.1	88.2
Melanomas	89	86.7	83.5	82.8
Breast	86.4	78.3	71.3	65
Hodgkin's disease	85.1	79.8	73.8	67.1
Uterus	84.3	83.2	80.8	79.2
Bladder	82.1	76.2	70.3	67.9
Cervix	70.5	64.1	62.8	60
Larynx	68.8	56.7	45.8	37.8
Rectum	62.6	55.2	51.8	49.2
Kidney	61.8	54.4	49.8	47.3
Colon	61.7	55.4	53.9	52.3
Non-Hodgkin's	57.8	46.3	38.3	34.3
Pharynx	56.7	44.2	37.5	33
Ovary	55	49.3	49.9	49.6
Leukemia	42.5	32.4	29.7	26.2
Brain	32	29.2	27.6	26.1
Multiple myeloma	29.5	12.7	7	4.8
Stomach	23.8	19.4	19	14.9
Lung and bronchus	15	10.6	8.1	6.5
Esophagus	14.2	7.9	7.7	5.4
Liver	7.5	5.8	6.3	7.6
Pancreas	4	3	2.7	2.7

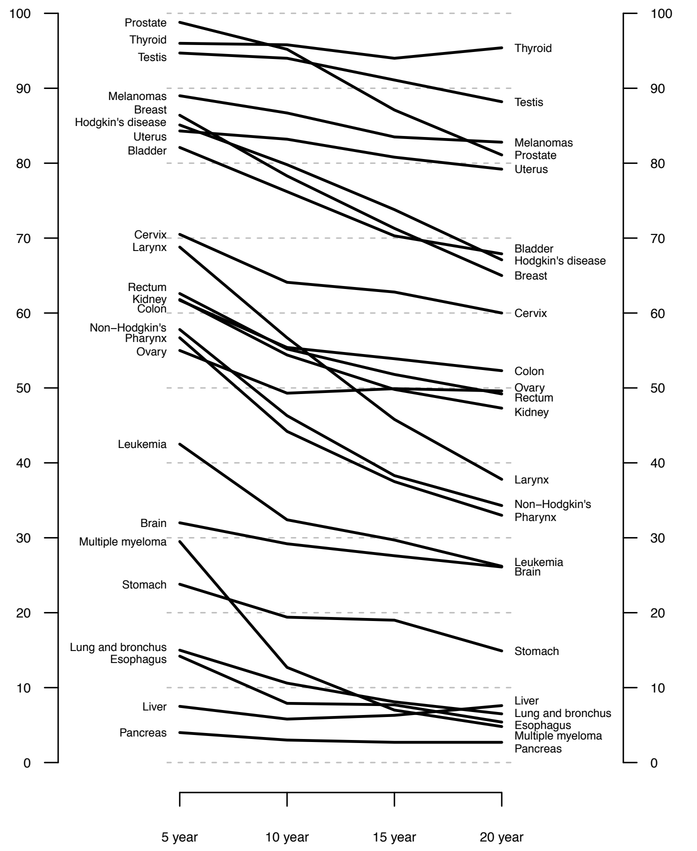
Solution 1: Dot chart



- A typical dot chart for this data is somewhat hard to read because
  - It is very dense (I could not even label the second level).
  - We are used to think of time as left to right, not up and down (we would need to rotate the graph).
  - Even rotated, we use dots rather than lines, which are hard to follow.
- The natural alternative (grouped bar plots) are not much better.

Solution 2: Slopegraph

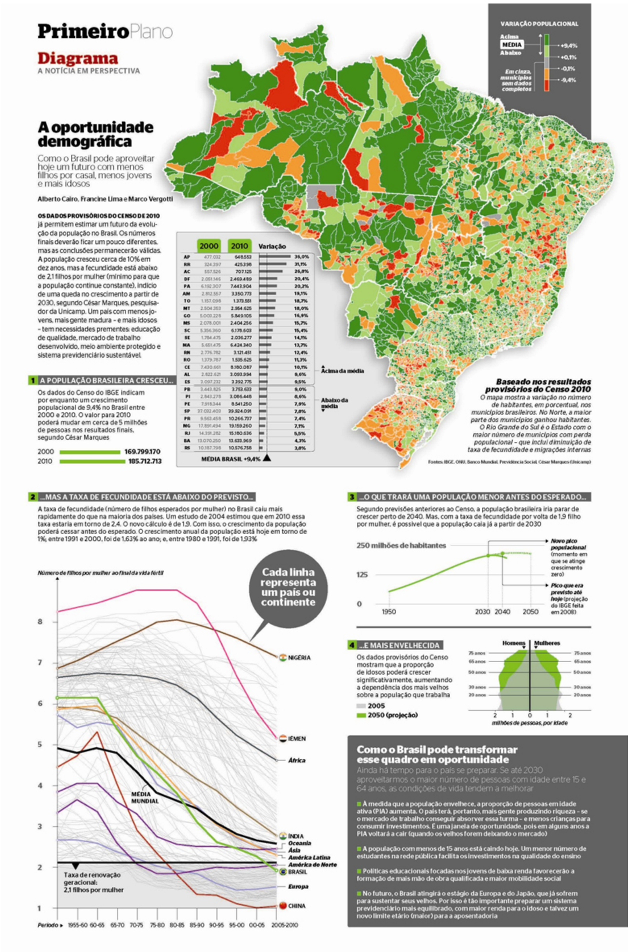
- Because the slopegraph uses lines, trends for each cancer site are very clear.
- Labels on both sides are helpful for distinguishing lines that are very close together.
- Rankings after 5 and 20 years can be read directly from the graph.



Solution 2: Slopegraph

```
quartz(height=7, width=5)
par(mar=c(4,6,1,6)+0.1)
n = dim(cancer)[1]
ord1 = order(cancer[,1], decreasing=T)
ord4 = order(cancer[,4], decreasing=T)
plot(seq(1,n), seq(1,n), xlim=c(1,4), ylim=c(0,100), xlab="", ylab="", type="n",
axes=F)
abline(h=seq(0,100,by=10), col="grey", lty=2)
for(i in ord1){
  lines(seq(1,4), cancer[i,1:4], lwd=2)
}
axis(1, at=seq(1,4), labels=c("5 year", "10 year", "15 year", "20 year"),
cex.axis=0.6)
axis(2, at=seq(0,100,by=10), line=4, las=2, cex.axis=0.6)
axis(4, at=seq(0,100,by=10), line=4, las=2, cex.axis=0.6)
offsetleft = c(0, 0.5, -0.5, 0, 0.8, 0.4, -0.7, -0.4, 0, 0, 0.9, 0.05, -1.1,
0.25, 0, -0.1, 0, 0, 0, 0, 0.4,-0.4, 0 ,0)
offsetright = c(0, 0, 0, 0, 0, 0.7, -0.1, 0, 0 , 0, 0.5, -0.5, -0.5, 0, 0.2,
-0.2, 0.6, -0.6, 0, 0.7, 0.1, -0.4, -1.2, -0.8)
mtext(side=2, line=0, at=cancer[ord1,1]+offsetleft, text=rownames(cancer)[ord1],
las=2, cex=0.5)
mtext(side=4, line=0, at=cancer[ord4,4]+offsetright, text=rownames(cancer)[ord4],
las=2, cex=0.5)
```

Other examples of plotting many time series



PrimeiroPlano

Diagrama

A NOTÍCIA EM PERSPECTIVA

A oportunidade demográfica

Como o Brasil pode aproveitar hoje um futuro com menos filhos por casal, menos jovens e mais idosos

Alberto Cairo, Francine Lima e Marco Vergotti

OS DADOS PROVISÓRIOS DO CENSO DE 2010 já permitem estimar um futuro da evolução da população no Brasil. Os números finais deverão ficar um pouco diferentes, mas as conclusões permanecerão válidas. A população cresceu cerca de 10% em dez anos, mas a fecundidade está abaixo de 2,1 filhos por mulher (mínimo para que a população continue constante), indício de uma queda no crescimento a partir de 2030, segundo César Marques, pesquisador da Unicamp. Um país com menos jovens, mais gente madura – e mais idosos – tem necessidades prementes: educação de qualidade, mercado de trabalho desenvolvido, meio ambiente protegido e sistema previdenciário sustentável

1 A POPULAÇÃO BRASILEIRA CRESCEU...

Os dados do Censo do IBGE indicam por enquanto um crescimento populacional de 9,4% no Brasil entre 2000 e 2010. O valor para 2010 poderá mudar em cerca de 5 milhões de pessoas nos resultados finais, segundo César Marques

2000	169.799.170
2010	185.712.713

	2000	2010	Variação
AP	477.032	648.553	36,0%
RR	324.397	425.398	31,1%
AC	557.526	707.125	26,8%
DF	2.051.146	2.469.489	20,4%
PA	6.192.307	7.443.904	20,2%
AM	2.812.557	3.350.773	19,1%
TO	1.157.098	1.373.551	18,7%
MT	2.504.353	2.954.625	18,0%
GO	5.003.228	5.849.105	16,9%
MS	2.078.001	2.404.256	15,7%
SC	5.356.360	6.178.603	15,4%
SE	1.784.475	2.036.277	14,1%
MA	5.651.475	6.424.340	13,7%
RN	2.776.782	3.121.451	12,4%
RO	1.379.787	1.535.625	11,3%
CE	7.430.661	8.180.087	10,1%
AL	2.822.621	3.093.994	9,6%
ES	3.097.232	3.392.775	9,5%
PB	3.443.825	3.753.633	9,0%
PI	2.843.278	3.086.448	8,6%
PE	7.918.344	8.541.250	7,9%
SP	37.032.403	39.924.091	7,8%
PR	9.563.458	10.266.737	7,4%
MG	17.891.494	19.159.260	7,1%
RJ	14.391.282	15.180.636	5,5%
BA	13.070.250	13.633.969	4,3%
RS	10.187.798	10.576.758	3,8%

MÉDIA BRASIL +9,4%

Baseado nos resultados provisórios do Censo 2010

O mapa mostra a variação no número de habitantes, em porcentual, nos municípios brasileiros. No Norte, a maior parte dos municípios ganhou habitantes. O Rio Grande do Sul é o Estado com o maior número de municípios com perda populacional – que inclui diminuição de taxa de fecundidade e migrações internas

Fontes: IBGE, ONU, Banco Mundial, Previdência Social, César Marques (Unicamp)

2 ...MAS A TAXA DE FECUNDIDADE ESTÁ ABAIXO DO PREVISTO...

A taxa de fecundidade (número de filhos esperados por mulher) no Brasil caiu mais rapidamente do que na maioria dos países. Um estudo de 2004 estimou que em 2010 essa taxa estaria em torno de 2,4. O novo cálculo é de 1,9. Com isso, o crescimento da população poderá cessar antes do esperado. O crescimento anual da população está hoje em torno de 1%; entre 1991 e 2000, foi de 1,63% ao ano; e, entre 1980 e 1991, foi de 1,93%

Número de filhos por mulher ao final da vida fértil
Cada linha representa um país ou continente
8
7
6
5
4
3
2
1
NIGÉRIA
IÊMEN
África
MÉDIA MUNDIAL
ÍNDIA
Oceania
Ásia
América Latina
América do Norte
BRASIL
Europa
CHINA
Taxa de renovação geracional: 2,1 filhos por mulher
Período ▸ 1955-60 60-65 65-70 70-75 75-80 80-85 85-90 90-95 95-00 00-05 2005-2010

3 ...O QUE TRARÁ UMA POPULAÇÃO MENOR ANTES DO ESPERADO...

Segundo previsões anteriores ao Censo, a população brasileira iria parar de crescer perto de 2040. Mas, com a taxa de fecundidade por volta de 1,9 filho por mulher, é possível que a população caia já a partir de 2030

250 milhões de habitantes
125
0
1950
2030
2040
2050
Novo pico populacional (momento em que se atinge crescimento zero)
Pico que era previsto até hoje (projeção do IBGE feita em 2008)

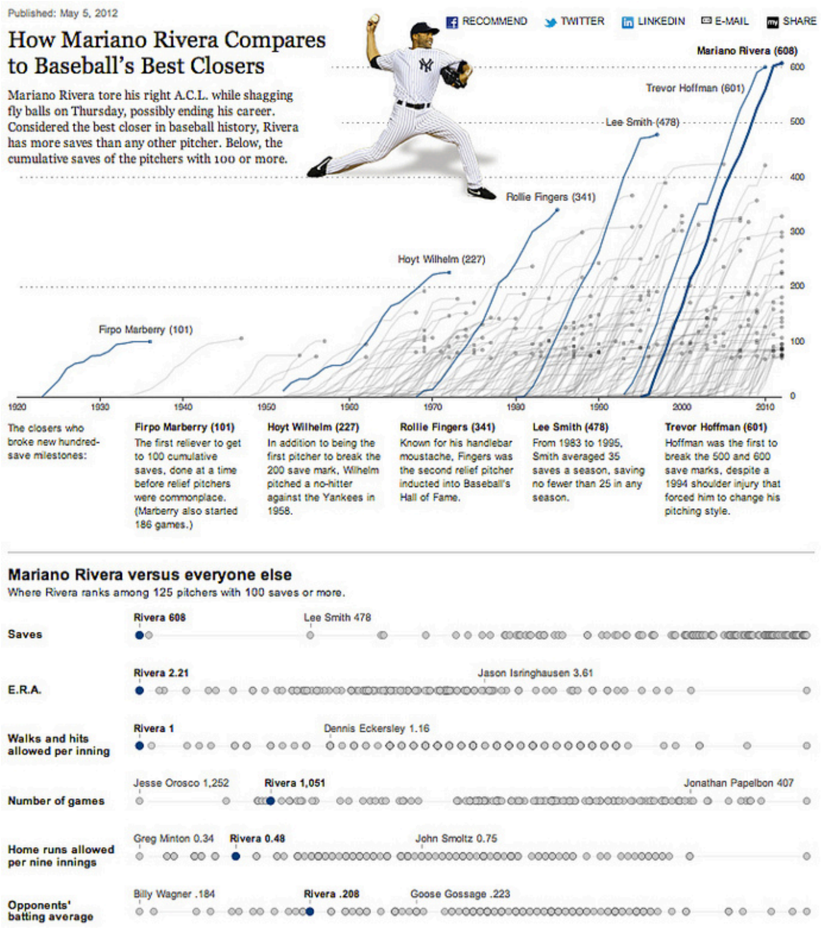
4 ...E MAIS ENVELHECIDA

Os dados provisórios do Censo mostram que a proporção de idosos poderá crescer significativamente, aumentando a dependência dos mais velhos sobre a população que trabalha

Como o Brasil pode transformar esse quadro em oportunidade

Ainda há tempo para o país se preparar. Se até 2030 aproveitarmos o maior número de pessoas com idade entre 15 e 64 anos, as condições de vida tendem a melhorar

- À medida que a população envelhece, a proporção de pessoas em idade ativa (PIA) aumenta. O país terá, portanto, mais gente produzindo riqueza – se o mercado de trabalho conseguir absorver essa turma – e menos crianças para consumir investimentos. É uma janela de oportunidade, pois em alguns anos a PIA voltará a cair (quando os velhos forem deixando o mercado)
- A população com menos de 15 anos está caindo hoje. Um menor número de estudantes na rede pública facilita os investimentos na qualidade do ensino
- Políticas educacionais focadas nos jovens de baixa renda favorecerão a formação de mais mão de obra qualificada e maior mobilidade social
- No futuro, o Brasil atingirá o estágio da Europa e do Japão, que já sofrem para sustentar seus velhos. Por isso é tão importante preparar um sistema previdenciário mais equilibrado, com maior renda para o idoso e talvez um novo limite etário (maior) para a aposentadoria



Published: May 5, 2012

How Mariano Rivera Compares to Baseball's Best Closers

Mariano Rivera tore his right A.C.L. while shagging fly balls on Thursday, possibly ending his career. Considered the best closer in baseball history, Rivera has more saves than any other pitcher. Below, the cumulative saves of the pitchers with 100 or more.

The closers who broke new hundred-save milestones:

Firpo Marberry (101) The first reliever to get to 100 cumulative saves, done at a time before relief pitchers were commonplace. (Marberry also started 186 games.)

Hoyt Wilhelm (227) In addition to being the first pitcher to break the 200 save mark, Wilhelm pitched a no-hitter against the Yankees in 1958.

Rollie Fingers (341) Known for his handlebar moustache, Fingers was the second relief pitcher inducted into Baseball's Hall of Fame.

Lee Smith (478) From 1983 to 1995, Smith averaged 35 saves a season, saving no fewer than 25 in any season.

Trevor Hoffman (601) Hoffman was the first to break the 500 and 600 save marks, despite a 1994 shoulder injury that forced him to change his pitching style.

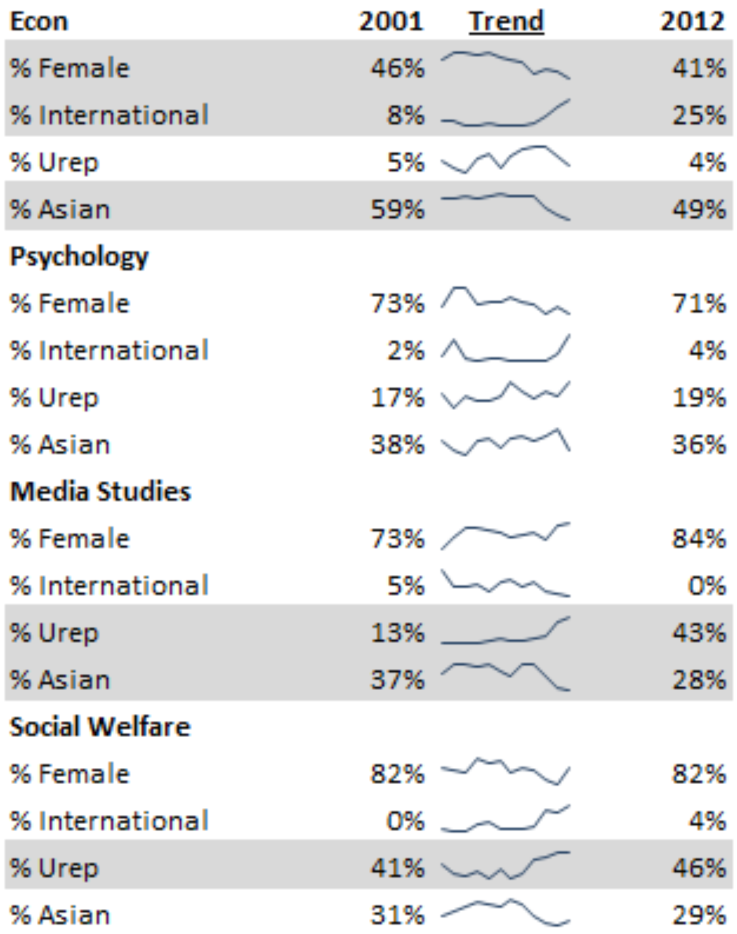
Mariano Rivera versus everyone else

Where Rivera ranks among 125 pitchers with 100 saves or more.

[Alberto Cairo, "The Functional Art"](http://chartsnthings.tumblr.com/post/22471358872/sketches-how-mariano-rivera-compares-to-baseballs)

http://chartsnthings.tumblr.com/post/22471358872/sketches-how-mariano-rivera-compares-to-baseballs

Sparklines



Econ	2001	Trend	2012
% Female	46%		41%
% International	8%		25%
% Urep	5%		4%
% Asian	59%		49%
Psychology			
% Female	73%		71%
% International	2%		4%
% Urep	17%		19%
% Asian	38%		36%
Media Studies			
% Female	73%		84%
% International	5%		0%
% Urep	13%		43%
% Asian	37%		28%
Social Welfare			
% Female	82%		82%
% International	0%		4%
% Urep	41%		46%
% Asian	31%		29%