

Graphics

Statistics 133 Fall 2014

Lecture 5 Sept 11 2014

Why is graphics in this course?

- Good graphics today requires the computer
- Visualization enters every step of the data analysis cycle
 - Data cleaning – are there anomalies?
 - Exploration
 - Model checking
 - Reporting results
- Plots can uncover structure in data that can't be detected with numerical summaries
- Important communication skill

R's graphics model

- There are two models in R – painter and object-oriented
- We will use the painter's model
- The other is easy to get started but hard to tweak
- Painter's model – start with a blank canvas, add/paint on it in multiple passes

Know your data types

The appropriate graphical techniques depend on the kind of data that you are working with

- Quantitative

- continuous – e.g. height, weight
- discrete – numeric data with few values, e.g. number of children in family

- Qualitative

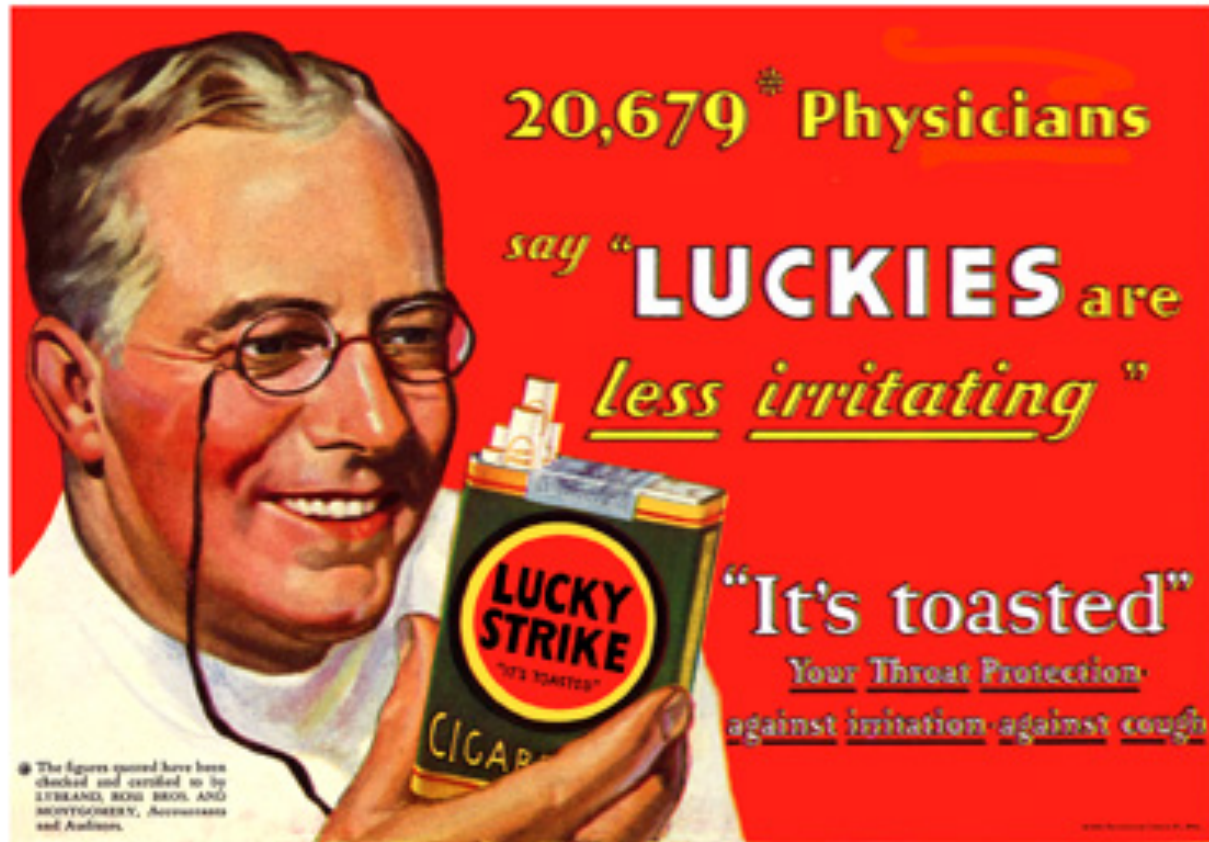
- ordered – categories with an order but no meaningful distance between, e.g. number of stars for a movie rating
- nominal - categories have no meaningful order, e.g. gender

Case: Infant Health

git pull

~src/stat133/classwork/lecture5/KaiserBabies.rda

Smoking:
Some doctors used to recommend it



Today:

SURGEON GENERAL'S WARNING: Smoking Causes Lung Cancer, Heart Disease, Emphysema, and May Complicate Pregnancy.

Kaiser Study

- Oakland Kaiser mothers
- 1960s
- Measure the babies weight (in ounces) at birth
- All babies:
 - Male
 - Single births (no twins, etc.)
 - Survived 28 days

Information collected on mother's and their babies

- Birth weight (ounces)
- Gestation (weeks)
- Parity - total number of previous pregnancies
- Mother's height and weight
- Mother's smoking status
- Mother's age, race, education level, income
- And more...

Here are the data for birth weight

What do you see?

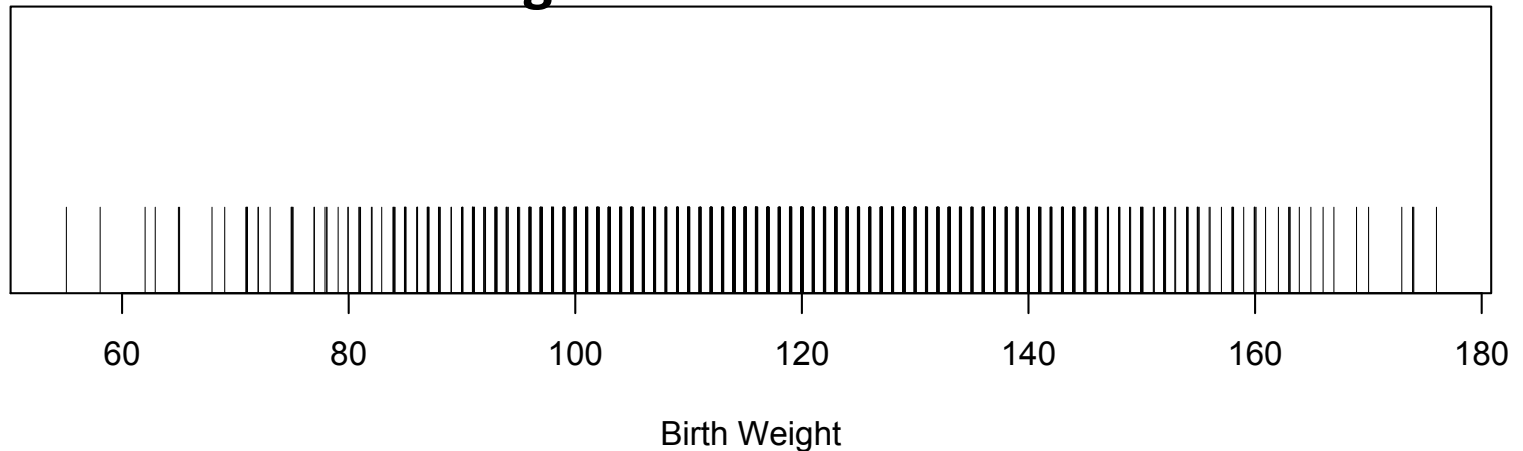
[1] 120 113 128 123 108 136 138 132 120 143 140 144 141 110 114 115 92 115 144 119 105 115 137 122 131 103 146 114 125 114 122 93 130 119 113 134
[37] 107 134 122 128 129 110 138 111 87 143 155 110 122 145 115 108 102 143 146 124 124 145 106 75 107 124 122 101 128 104 97 137 103 142 130 156
[73] 133 120 91 127 153 121 120 99 149 129 139 114 138 129 138 131 125 114] 128 134 114 92 85 135 87 125 128 105 120 119 116 107 119 133 155 126
[109] 129 137 103 125 91 134 95 118 141 131 121 100 131 118 152 121 117 115 112 94 109 132 117 101 112 128 128 117 134 127 93 122 100 147 120 144
[145] 105 136 102 160 113 126 126 115 127 119 129 123 118 133 105 134 144 111 125 135 134 116 129 113 131 126 121 121 138 136 120 122 134 101 112 132
[181] 136 113 96 124 113 131 137 133 107 96 142 136 75 125 104 130 90 118 123 137 101 142 98 124 151 109 150 119 131 101 113 127 97 117 150 85
[217] 128 105 90 115 107 121 119 117 134 117 115 110 130 140 111 93 154 125 93 122 129 126 85 173 144 114 111 154 150 111 126 122 141 142 99 113
[253] 149 117 130 106 128 125 114 130 116 81 124 125 110 125 138 142 115 102 140 133 127 104 119 152 123 143 131 141 129 113 119 109 104 131 110 148
[289] 137 117 115 98 136 121 132 91 119 85 106 132 80 109 111 143 136 110 98 108 101 71 124 93 106 101 100 104 117 117 149 135 110 121 142 104
[325] 138 112 117 109 131 120 116 140 103 120 139 123 104 131 111 122 116 129 133 110 105 93 122 133 130 104 106 120 121 118 140 114 116 129 120 127
[361] 107 71 88 107 122 106 135 107 129 126 116 124 123 145 102 129 98 110 135 101 96 104 100 154 127 126 126 127 98 127 129 131 132 127 99 115
[397] 145 102 136 121 121 120 118 127 132 102 143 118 102 163 132 116 138 139 132 87 131 130 123 115 116 119 125 144 123 120 140 120 116 120 146 112
[433] 115 132 146 122 128 119 135 116 129 116 100 118 138 123 113 129 122 132 120 114 130 117 142 144 127 115 85 99 123 112 68 102 109 102 99 78
[469] 128 107 136 101 100 109 117 88 95 119 123 127 107 124 126 98 96 104 133 93 101 118 130 125 140 115 130 114 105 101 132 112 69 114 123 129
[505] 114 115 98 128 119 119 154 127 131 129 114 110 103 117 138 126 124 111 132 103 158 146 101 132 114 71 116 108 123 129 134 113 123 147 121 125
[541] 115 101 93 109 115 130 123 111 97 122 124 129 124 107 142 129 174 105 103 124 105 133 161 105 108 153 133 115 127 128 117 123 119 141 91 116
[577] 116 121 111 102 118 126 98 131 115 103 147 123 125 117 99 115 116 118 170 104 108 144 99 97 142 85 130 117 109 147 105 135 115 123 105 154
[613] 110 119 103 117 120 145 104 123 124 129 91 109 108 79 133 114 128 129 97 103 176 143 127 107 113 106 152 150 136 151 124 123 119 122 112 93
[649] 109 136 121 150 94 120 146 129 125 124 141 96 138 127 114 103 127 141 113 99 97 116 126 158 119 123 129 117 100 131 146 84 115 115 118 91
[685] 112 115 110 117 109 99 131 136 130 134 128 150 86 115 141 78 100 116 110 109 113 136 114 121 117 166 87 120 95 132 90 131 103 144 137 124
[721] 136 117 121 116 139 110 86 133 81 133 132 132 137 84 136 92 114 129 167 71 124 105 155 125 125 125 115 174 127 113 115 139 127 111 112 143
[757] 116 155 121 110 87 132 105 129 123 91 147 144 128 137 104 120 112 138 96 134 126 112 138 110 83 112 148 119 86 110 126 125 136 127 84 131
[793] 123 96 110 123 152 127 117 125 139 114 96 124 107 113 98 119 107 117 117 144 136 121 165 120 125 137 100 134 88 108 123 141 130 139 130 113
[829] 77 62 93 109 145 92 120 135 113 126 143 128 98 110 162 116 128 111 137 134 100 160 112 134 145 116 126 111 126 109 136 119 103 124 155 122
[865] 113 122 126 116 102 110 133 125 164 133 135 124 122 121 100 129 90 128 116 86 123 87 128 120 125 118 116 131 151 88 137 127 96 129 128 85
[901] 111 124 112 115 72 122 116 127 90 99 144 138 58 109 110 129 150 128 142 115 108 108 139 115 136 163 131 77 124 104 102 94 158 112 119 97
[937] 99 115 139 144 99 105 89 129 119 114 106 122 136 121 112 112 123 139 125 105 130 146 133 147 109 122 135 107 117 138 120 119 118 105 113 136
[973] 148 140 134 120 123 102 55 103 123 105 138 128 139 104 159 118 99 144 121 117 119 105 125 119 101 105 110 100 98 127 117 122 122 118 137 120
[1009] 143 108 131 110 105 133 125 78 114 111 103 114 75 169 94 150 144 144 143 145 121 105 134 129 114 97 160 65 145 95 139 123 109 110 122 115
[1045] 117 108 120 131 136 125 96 102 102 112 135 91 129 155 109 80 125 94 148 73 123 65 118 102 120 108 122 103 105 126 145 139 124 121 126 119
[1081] 114 118 127 117 137 133 100 107 115 91 112 125 157 108 130 135 123 100 124 174 129 119 126 128 116 100 96 131 110 108 129 141 110 118 111 160
[1117] 120 121 113 117 158 128 158 133 163 128 126 127 134 140 102 100 120 98 130 104 122 137 114 63 98 99 89 117 143 106 99 156 72 75 97 106
[1153] 91 117 117 112 112 141 131 130 132 114 160 106 84 112 139 104 130 71 82 119 123 115 124 138 88 146 128 82 100 114 97 126 122 152 116 132
[1189] 84 119 104 106 124 139 103 112 96 102 120 102 97 113 130 97 116 114 127 87 141 144 116 75 138 99 118 152 97 146 81 110 135 114 124 115
[1225] 143 113 109 103 118 127 132 113 128 130 125 117

Rug plot

Each baby's weight is represented as a tickmark. The thicker lines are from multiple babies with similar weights. I added a little random noise to the weights to keep them from falling on top of each other.

What can you see now?

How are birth weights distributed?



Switch to R

Run commands given in R file. Some of the output from that code is included in the coming slides.

Distribution of Birth Weight

- The **distribution** is the pattern of variation in the birth weights.
- It provides the numerical values for birth weight and how often each value occurs.
- A **histogram/density plot** shows the shape of the distribution

Histograms

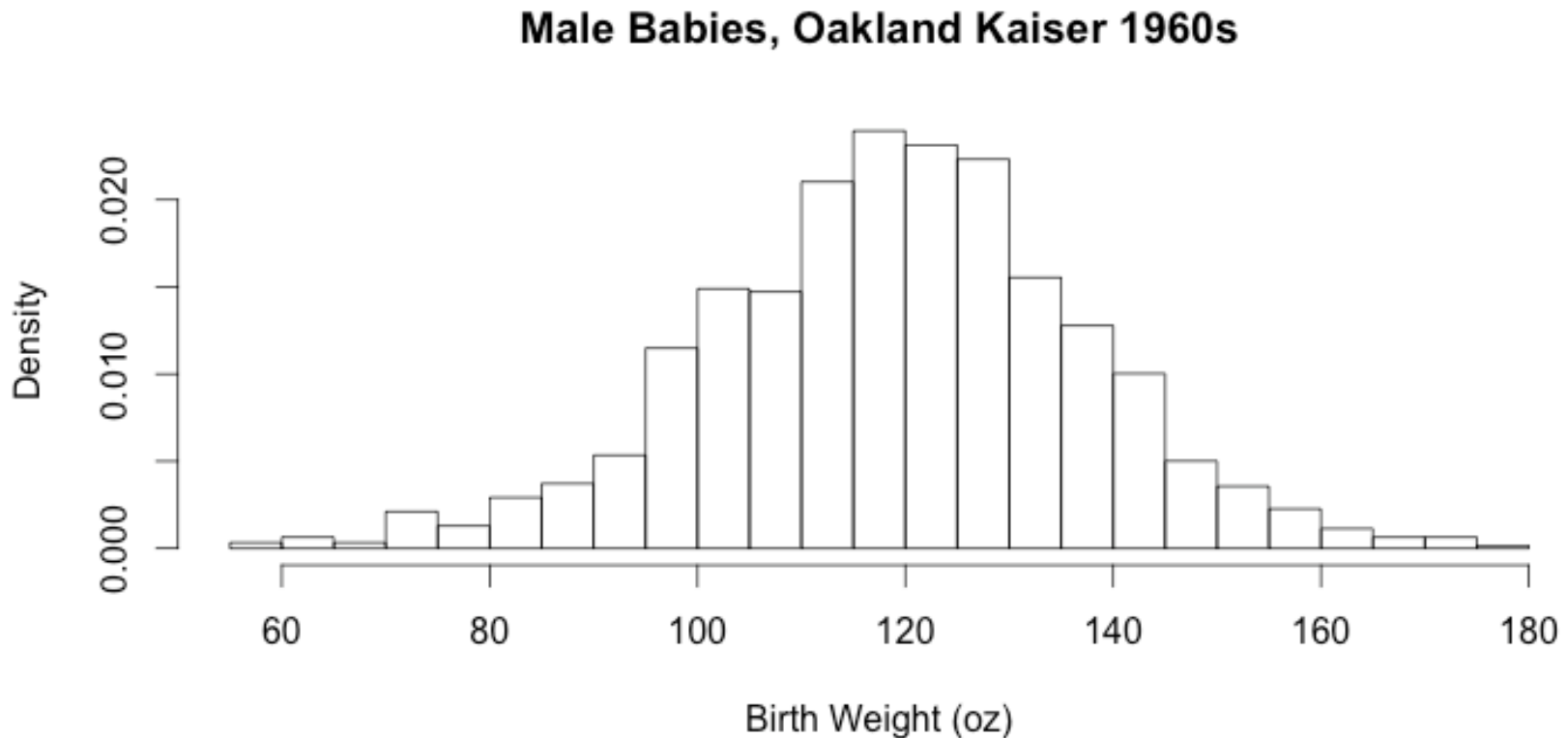
- Are a special case of density plots
- AREA = Proportion (or percent)
- The area of a bar:

$$\text{Height} * \text{Width} = \text{Area}$$

$$(\text{Proportion/oz}) * \text{oz} = \text{Proportion}$$

- Histograms are not the same as bar charts
- With bar charts, it is only the height that matters. **Bar charts are for qualitative data**

Histogram: `hist(infants$bwt)`

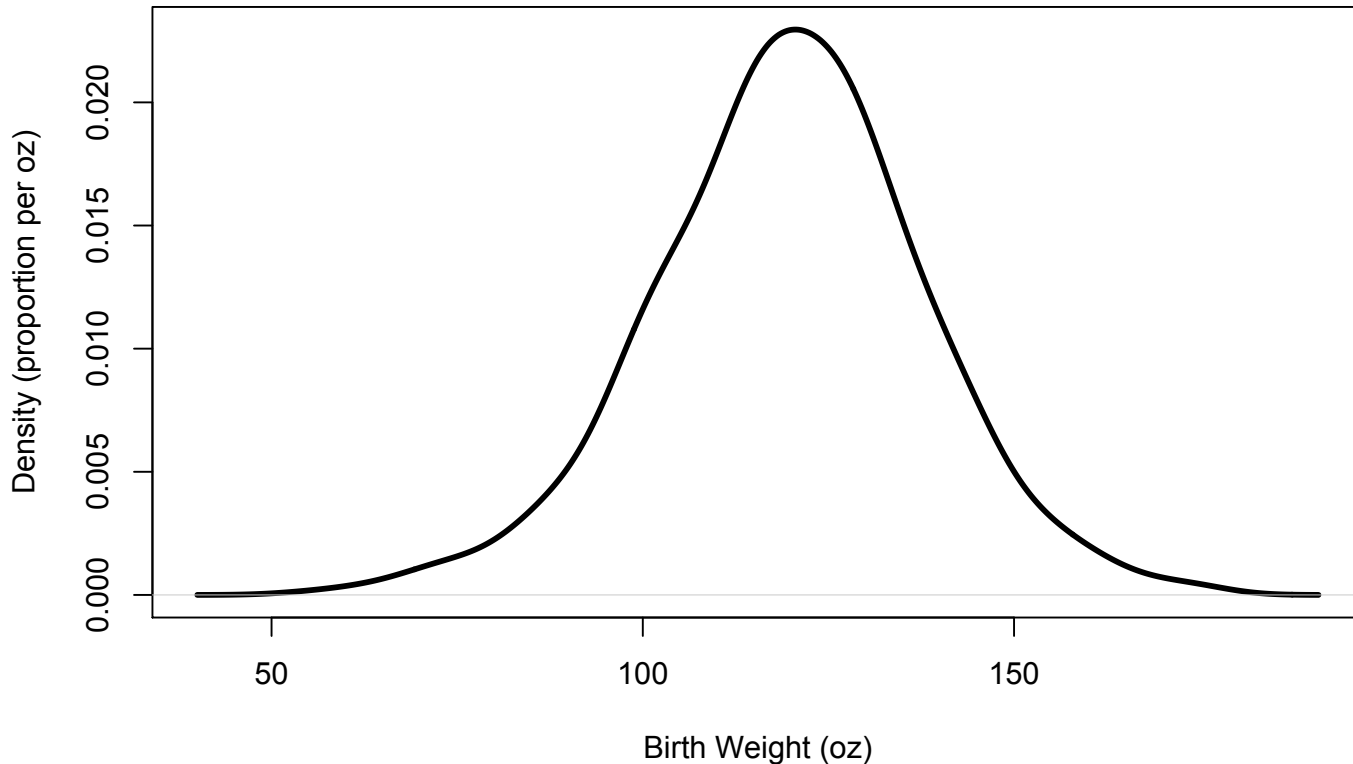


```
hist(infants$bwt , freq = FALSE,  
      xlab = "Birth Weight (oz)",  
      main = "Male Babies, Oakland Kaiser 1960s")
```

Density plot – smoothed histogram

```
plot(density(infants$bwt))
```

Male babies born at Oakland Kaiser in the 1960s

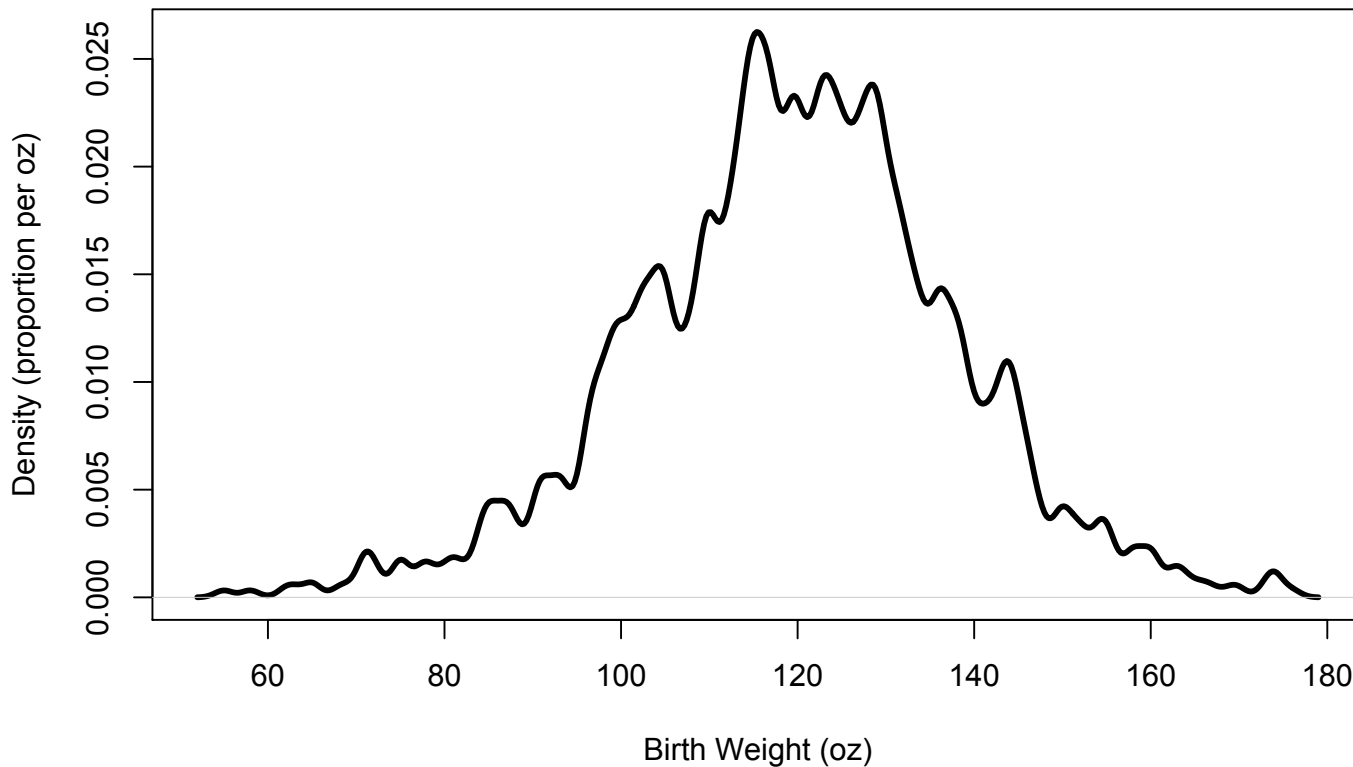


```
plot(density(infants$bwt),  
     xlab = "Birth Weight (oz)",  
     main = "Male Babies, Oakland Kaiser...")
```


Babies birth weight

```
plot(density(infants$bwt, bw = 1))
```

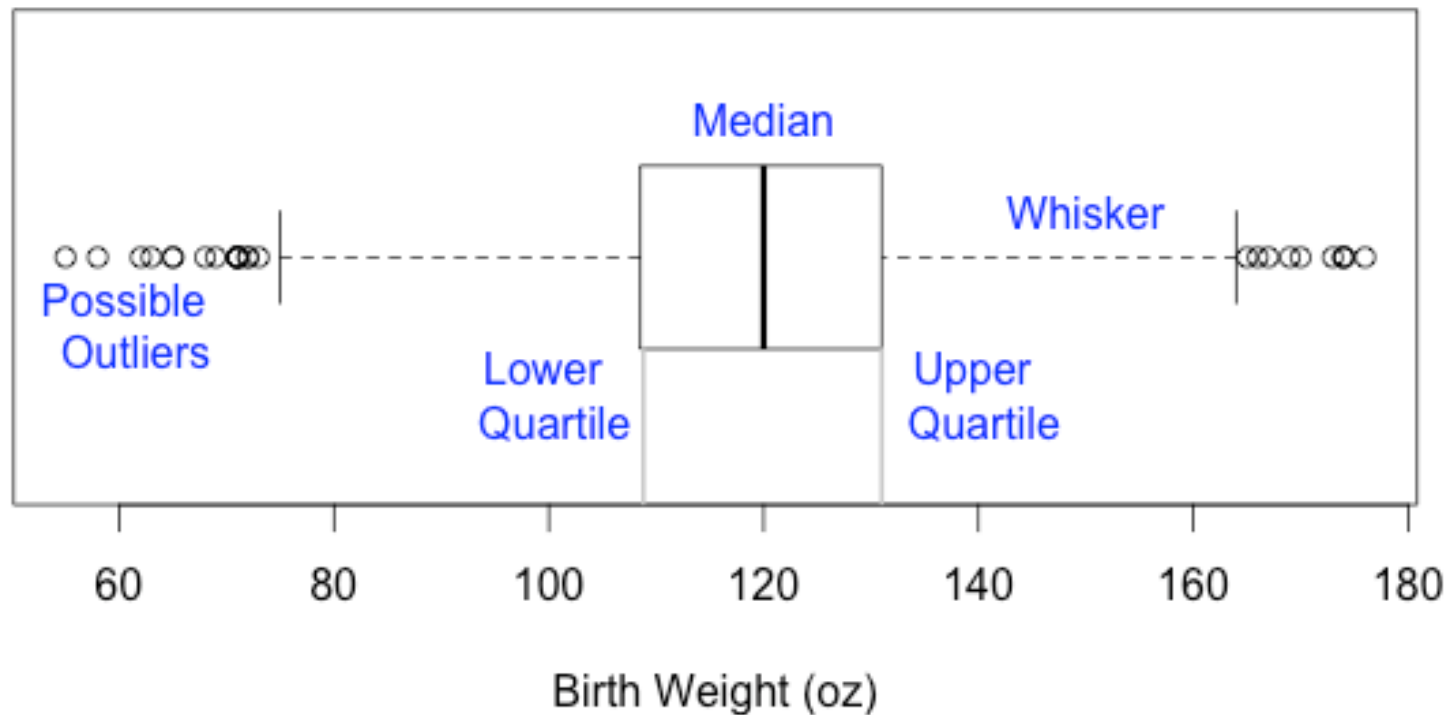
Male babies born at Oakland Kaiser in the 1960s



Selecting a **bandwidth**

- R chooses a bandwidth for you, but you can specify one if you like.
- The goal is to see the overall shape of the distribution, not the individual points.
- In a way, the density is a smooth abstraction of the distribution.

Boxplot: `boxplot(infants$bwt)`

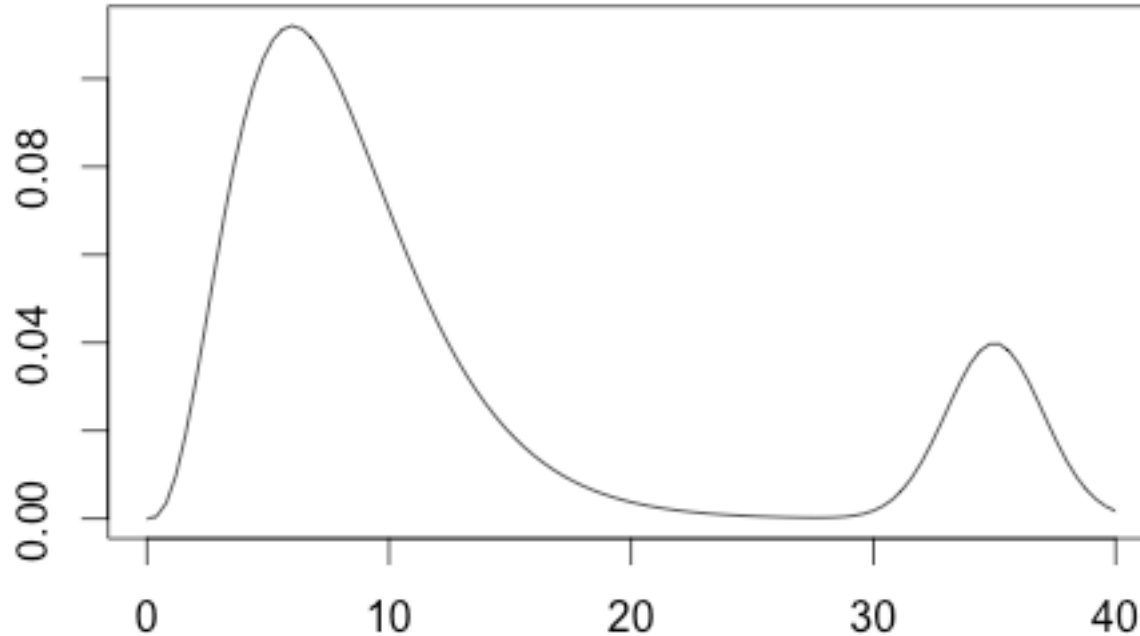


```
boxplot(infants$bwt,  
        xlab="Birth Weight (oz)")
```

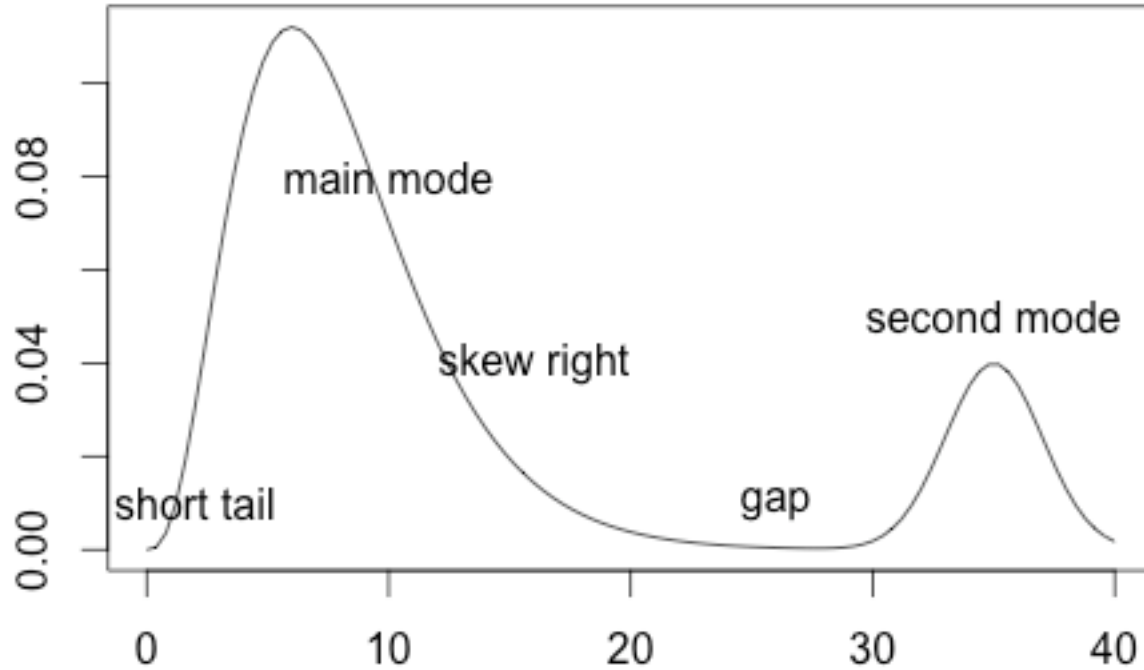
Looking for Structure: Quantitative Distribution

- **Distribution:** pattern of values for a variable
- **Mode:** high density region
- **Long Tail:** many observations far from center
- **Symmetry/Skewness:** distribution of values the left and right of the center.
- **Gaps:** places where there are no observations.
- **Outliers:** unusually large or small values that falls well beyond the overall pattern of data

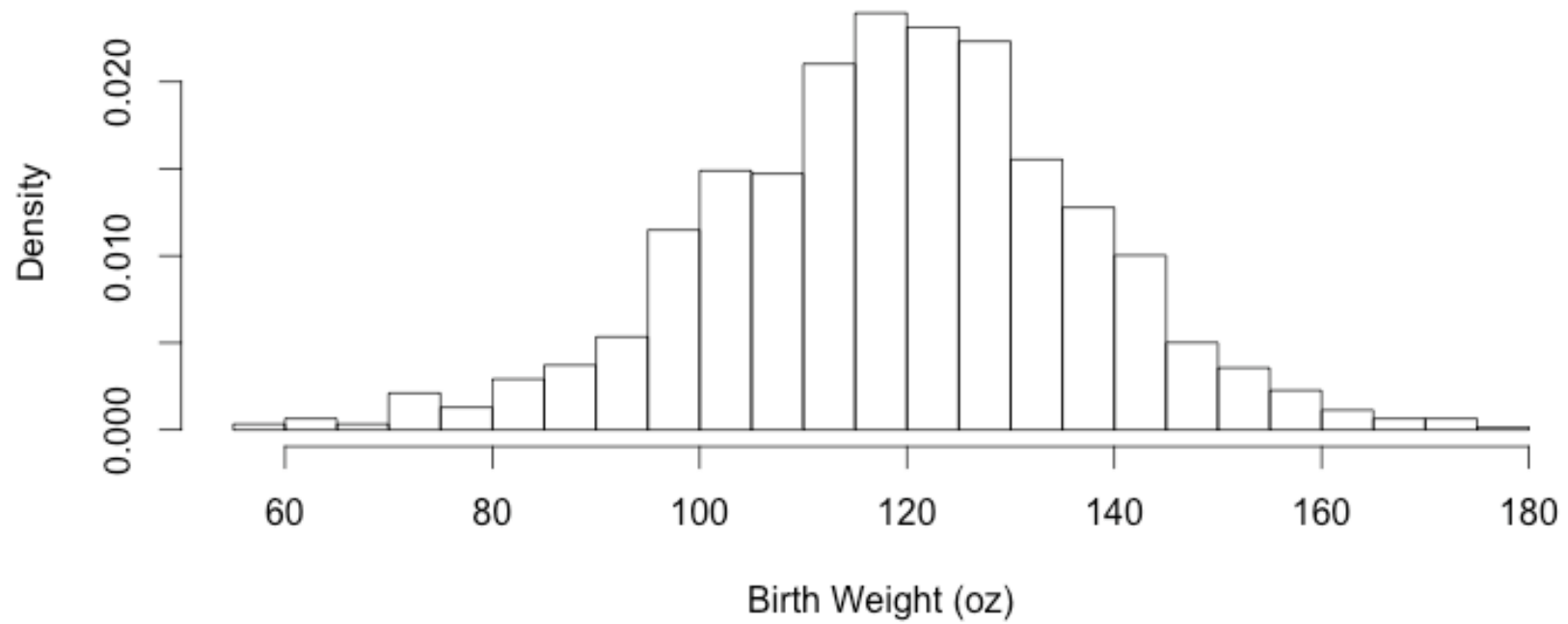
What Structure Do You See?



What Structure Do You See?



Male Babies, Oakland Kaiser 1960s



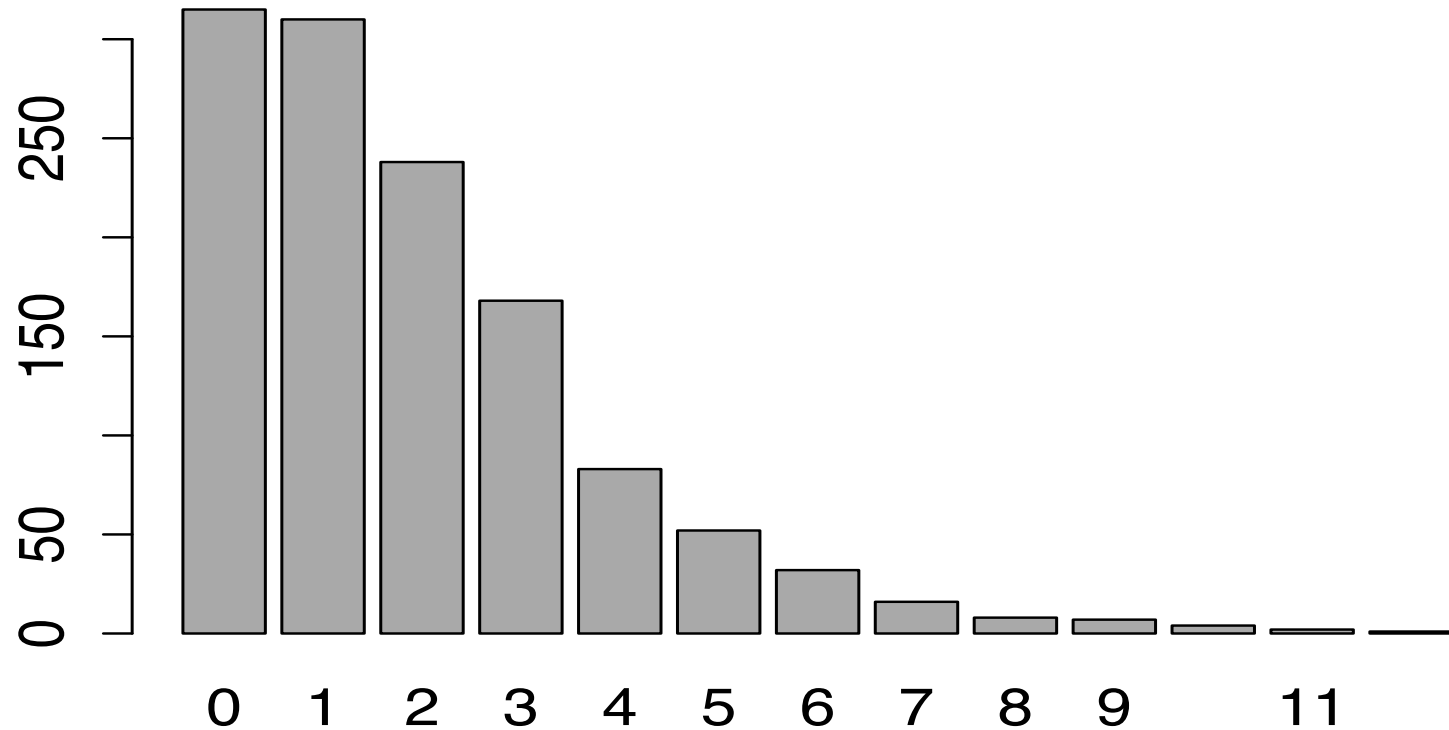
Parity: Number of siblings

- This quantitative variable is different from birth weight – there are only a few possible values, i.e. it's not possible to have 2.3 siblings, and it's highly unlikely to have 17

```
> table(infants$parity)
```

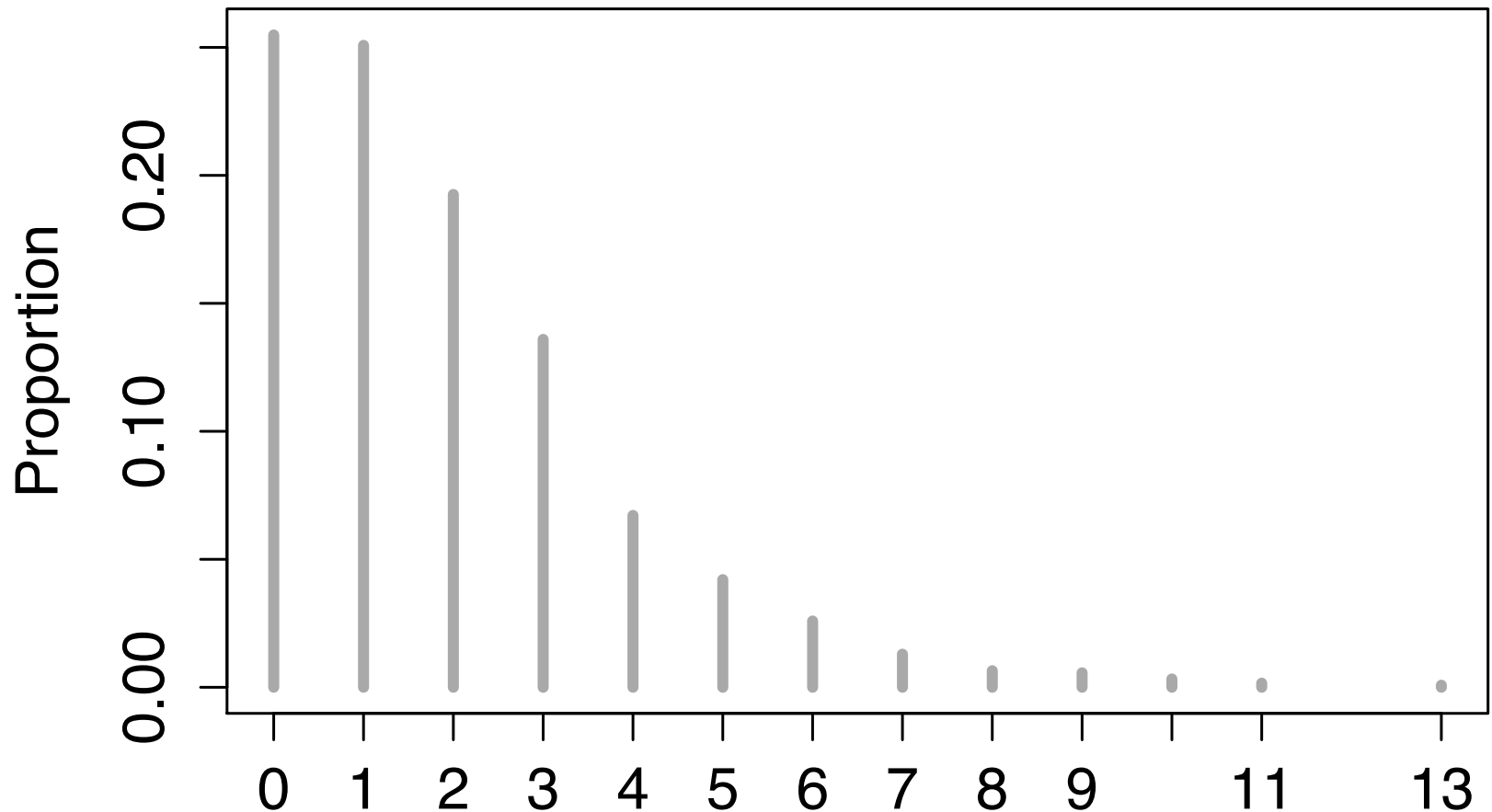
0	1	2	3	4	5	6	7	8	9	10	11	13
315	310	238	168	83	52	32	16	8	7	4	2	1

Number of Siblings



```
barplot(table(infants$parity))
```

Alternative – bar width has no meaning



```
plot(table(infants$parity),  
      type = "h", lwd = 4,  
      ylab = "Proportion", col = "darkgrey")
```

Case: College Students

git pull

~src/stat133/classwork/lecture5/videogame.rda

STAT 2 Survey

- Random Sample of 91 of 314 Cal students enrolled in Stat 2
- Survey collected the following info:
 - sex – Male/Female
 - grade – grade expected in the course (“A”, “B”, “C”, “D”, “F”)
- What type of data are these?
 - sex is qualitative (nominal)
 - grade is qualitative with an ordering

Switch to R...

```
> objects()
```

```
[1] "infants" "video"
```

```
> names(video)
```

```
[1] "time" "like" "where" "freq" "busy" "educ"
```

```
[7] "sex" "age" "home" "math" "work" "own"
```

```
[13] "cdrom" "email" "grade"
```

```
> dim(video)
```

```
[1] 91 15
```

Make tables of qualitative data

```
> table(video$grade)
```

Anything unusual
about the expected
grade?

```
F D C B A  
0 0 8 52 31
```

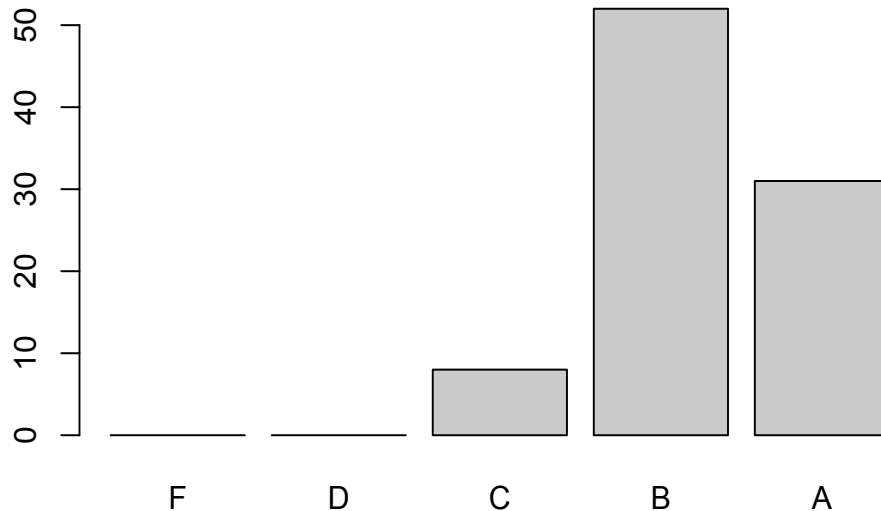
```
> table(video$grade, video$sex)
```

	Female	Male
F	0	0
D	0	0
C	8	0
B	21	31
A	9	22

Does expected
grade depend on
gender?

Expected Grade

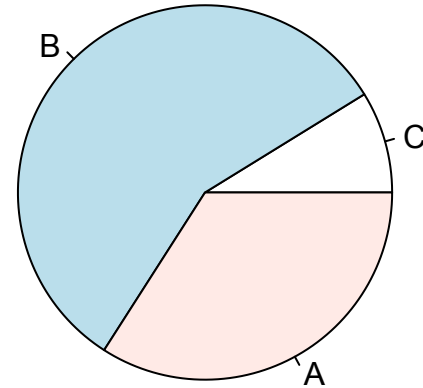
Bar chart



WIDTH of bars have no meaning

Pie chart

```
pie(table(video$grade))
```



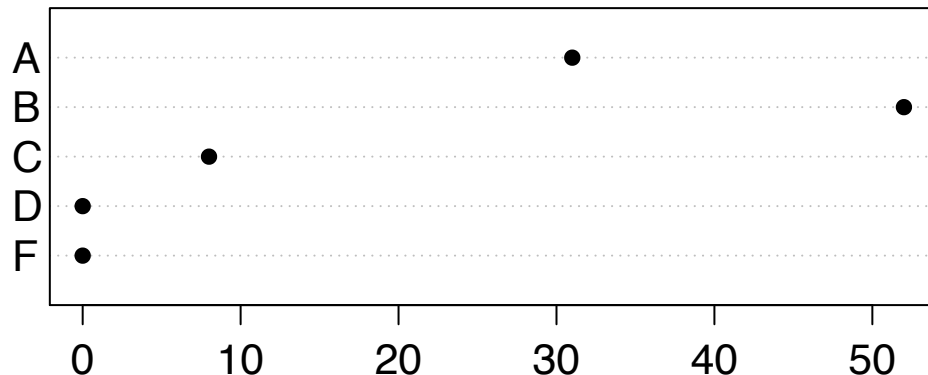
AREAS can be hard to compare

Expected Grade

Dot chart

```
dotchart(table(video$grade), pch = 19)
```

Focus on
comparison of
the values

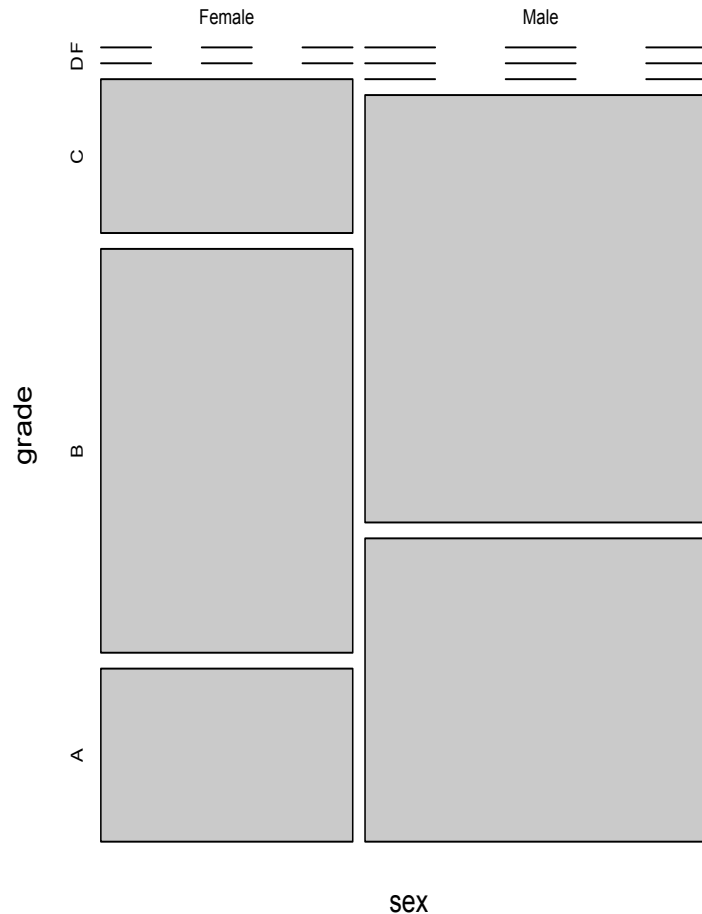


Method of Comparison

- Often, we not only want to better understand a distribution, but we want to compare the distribution for subgroups or to compare against another population or standard
- How do you think the expected grade distribution might vary with gender?

Two Qualitative variables

Stat 2 Survey



```
mosaicplot(table(video$sex, video$grade),  
            main = "Stat 2 Survey")
```

How to read a Mosaic plot

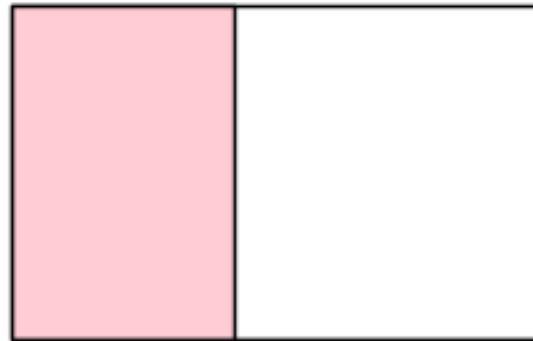
There are 91
students in the
survey.

Think of them as
spread out evenly
in the box



New Plot: Mosaic

Put all the
females on
one side of
the box.
There are 38.

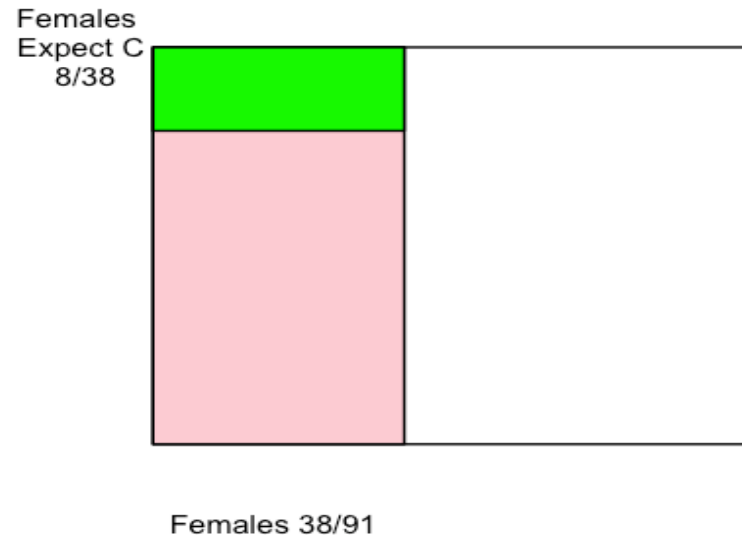


Females 38/91

New Plot: Mosaic

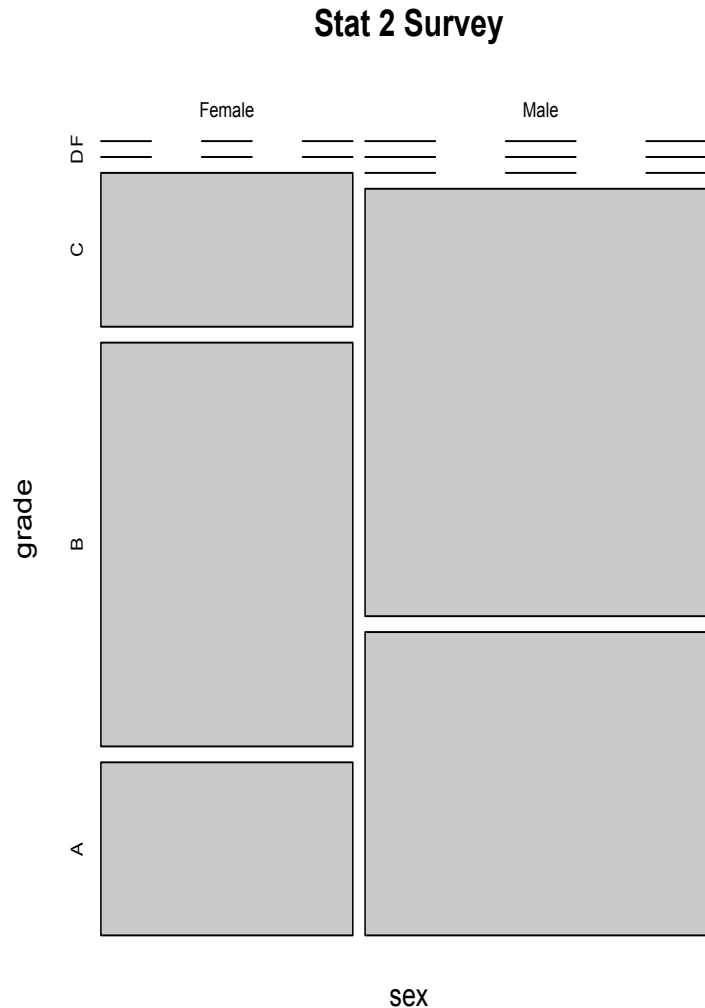
Rearrange the females
so that those who
expect the same grade
are together in the
box.

8 of the 38 expect a C



Mosaic plot

Smaller
fraction of
females expect
an A in
comparison to
Males



None of the
males expect
a C

Case: East Bay Housing Market

SFHousing.rda

Warning: It's BIG

San Francisco Chronicle listings

© FEBRU 2011



Data

- Record: house sold in a particular time period
 - Over 200,000 houses
 - Subset to a dozen cities in the East Bay – about 25,000 houses
- Variables:
- City
 - County
 - Price
 - # bedrooms
 - Lot square footage
 - and 10 more

Relationship between city and sale price

Data types:

City - factor

Sale price - numeric

Examine a subset of the cities

```
someCities = c("Albany", "Berkeley", "El  
Cerrito", "Emeryville", "Piedmont",  
"Richmond", "Lafayette", "Walnut Creek",  
"Kensington", "Alameda", "Orinda", "Moraga")
```

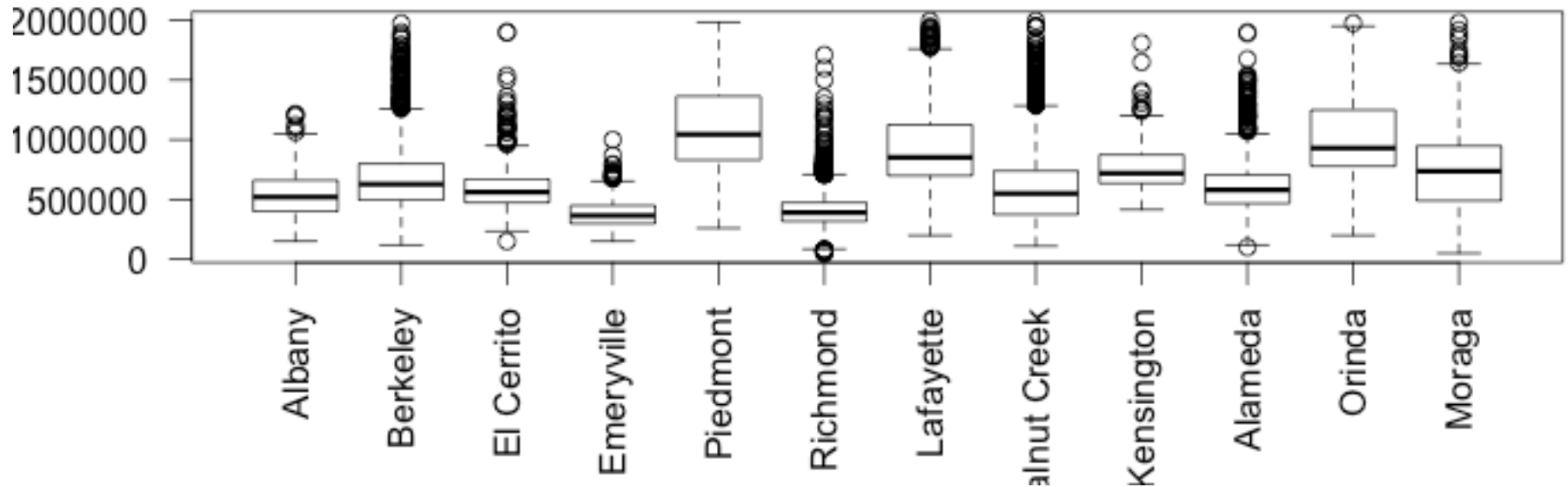
```
shousing =
```

```
housing[housing$city %in% someCities &  
housing$price < 2000000,]
```

```
dim(shousing)
```

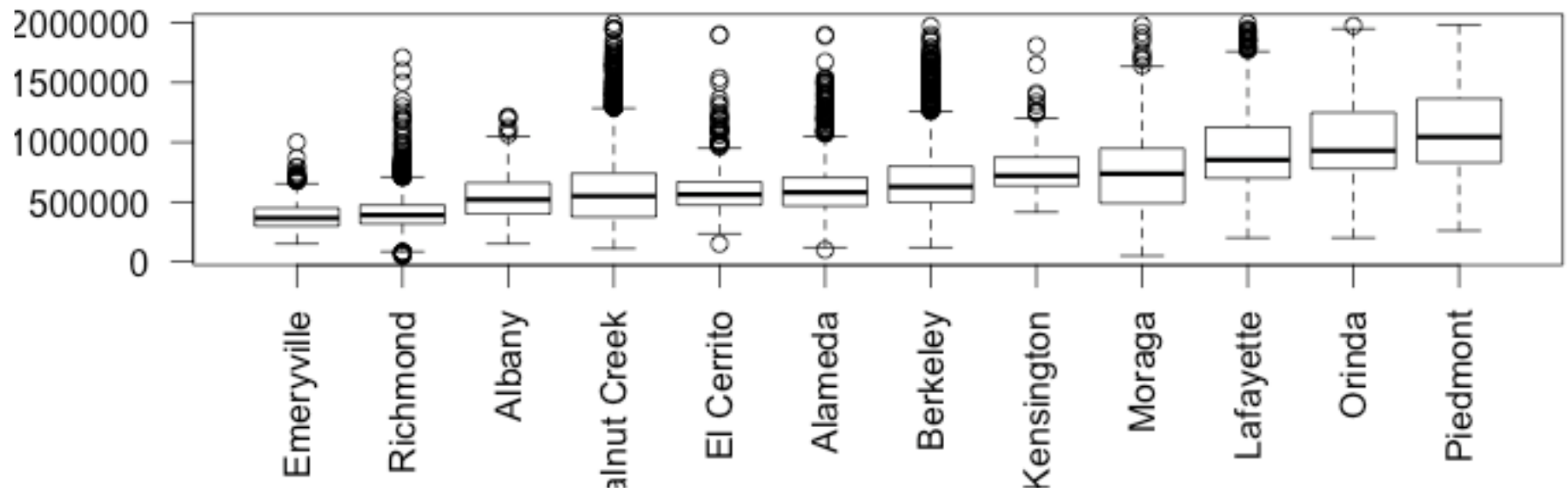
```
[1] 20415  15
```

Boxplots



```
boxplot(shousing$price ~ shousing$city,  
las = 2)
```

Cities ordered by median price

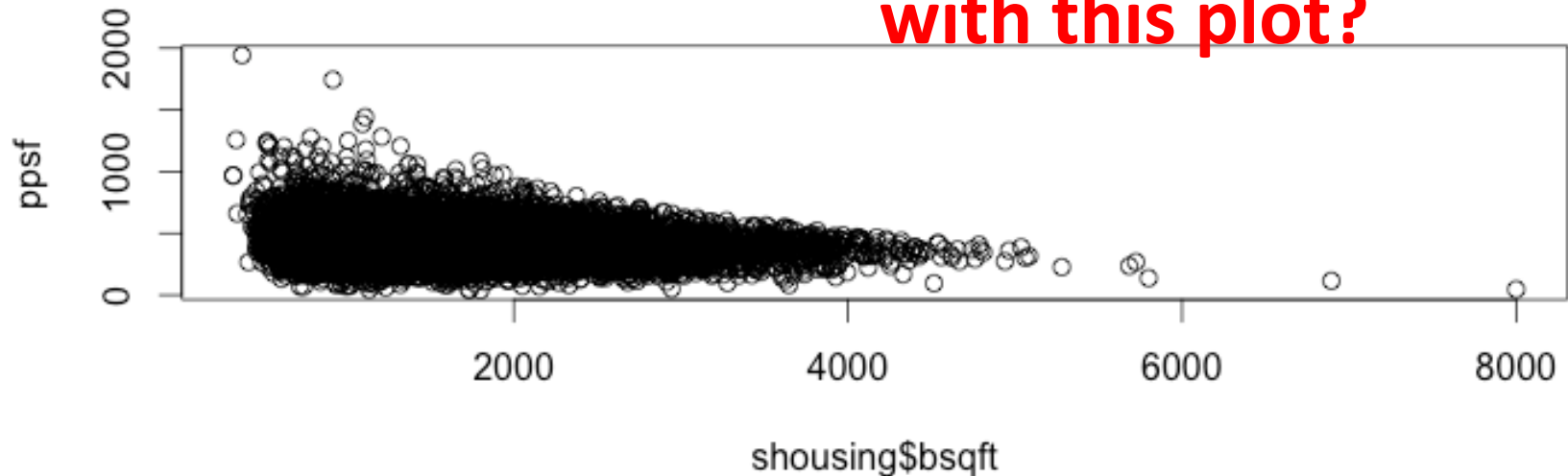


Relationship between price per square foot and total square foot

Both are quantitative

```
ppsf = shousing$price/shousing$bsqft  
plot(ppsf ~ shousing$bsqft)
```

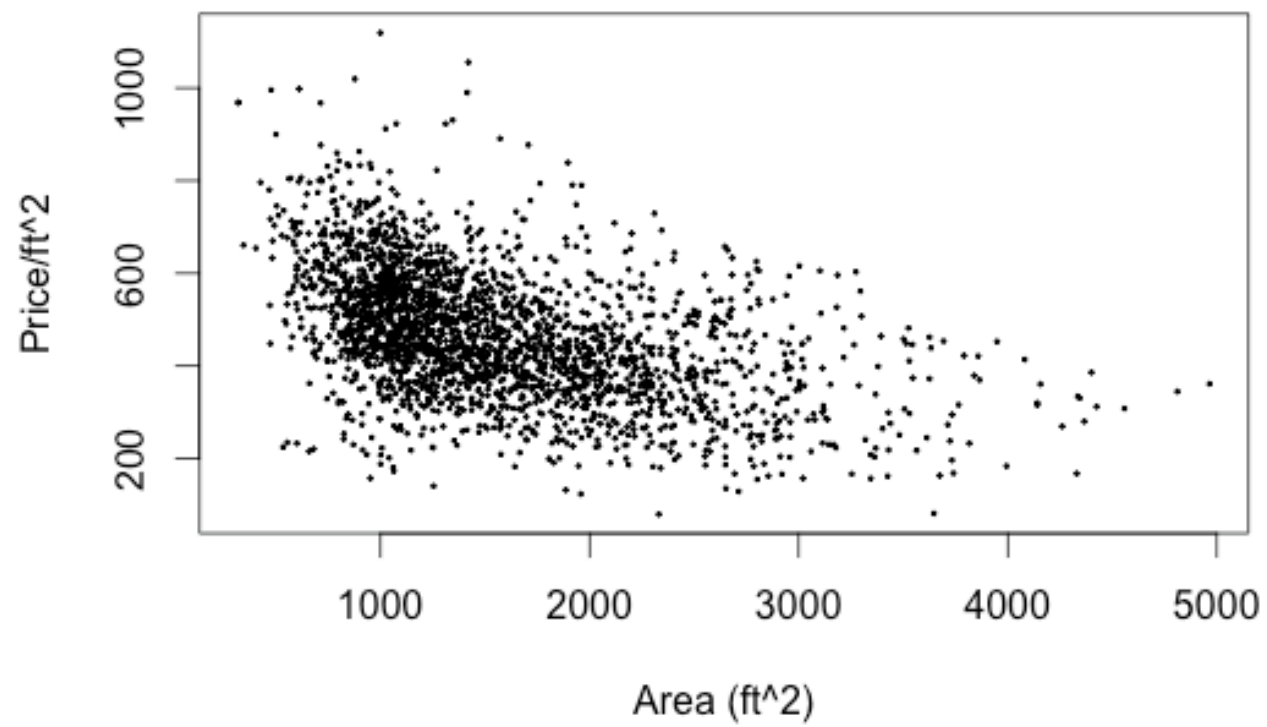
**WHAT's Wrong
with this plot?**



Scatter plot

```
plot(ppsf ~ shousing$bsqft, plot y against x  
pch=19, change plotting character to solid circle  
cex = 0.2, shrink plotting character to 20%  
subset = shousing$city == "Berkeley",  
Plot a subset of records  
main="Berkeley", title of plot  
xlab="Area (ft^2)", label for x axis  
ylab = "Price/ft^2") label for y axis
```


Berkeley

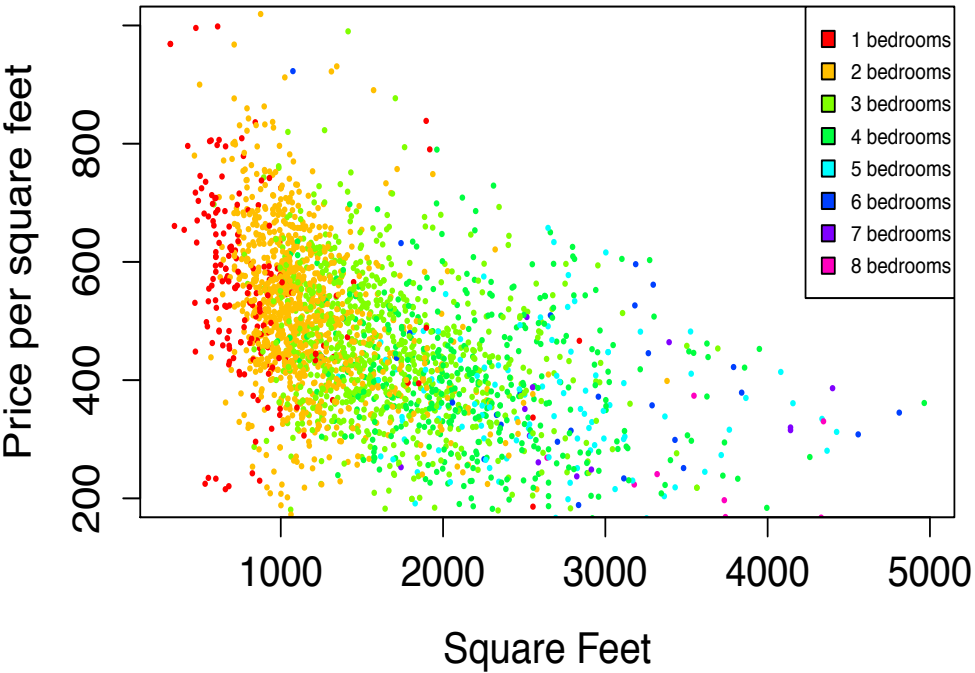


Relationships between more than 2 variables

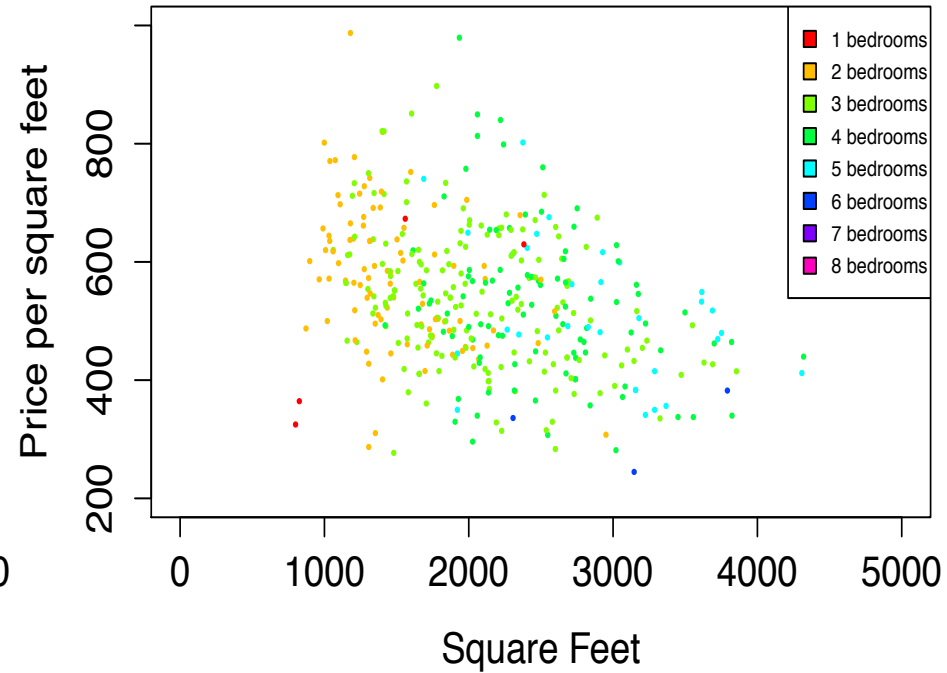
- Qualitative information can be conveyed in plots through color, plotting symbol, juxtaposed panels
- The following plot uses information from 4 variables: city, number of bedrooms, lot size (sq ft), and price per square ft

What do you see?

Berkeley



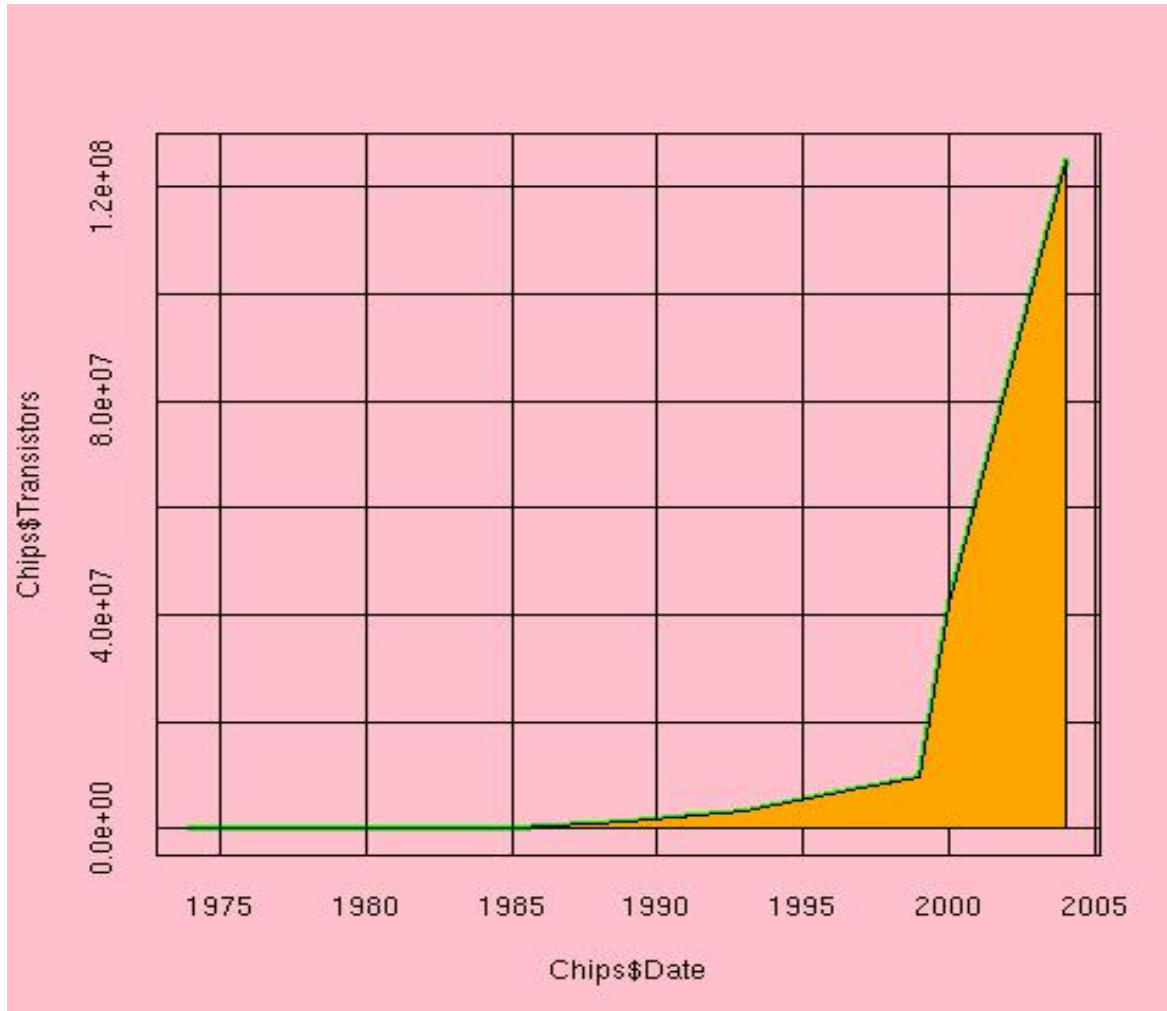
Piedmont



Summary of graph relationships between two variables

- Two Qualitative variables
 - mosaicplot, side-by-side barplots
- One Quantitative and one Qualitative
 - Boxplots, dotcharts, multiple density plots, violin plots
- Two Quantitative variables
 - Scatter plot, line plot

What do you think of this plot?



FIND 5 things that you would change

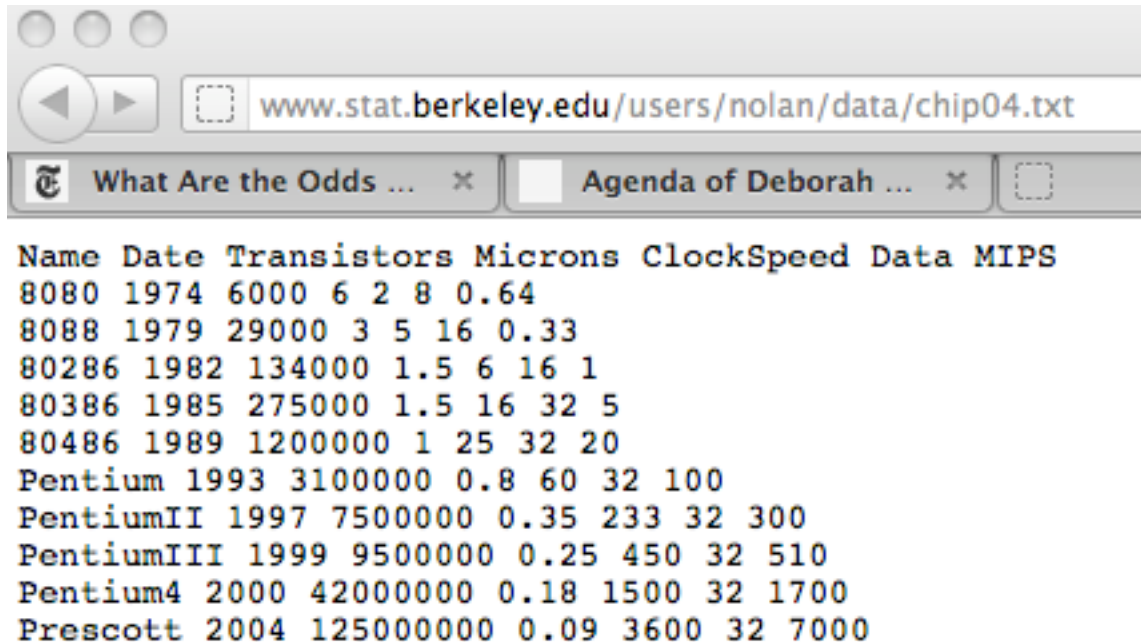
Let's fix it!

Making good plots is an iterative process

Goal is to convey a message as clearly as possible

Visit the website

<http://www.stat.berkeley.edu/users/nolan/data/chip04.txt>



The image shows a screenshot of a web browser window. The address bar displays the URL `www.stat.berkeley.edu/users/nolan/data/chip04.txt`. Below the address bar, there are two tabs: "What Are the Odds ..." and "Agenda of Deborah ...". The main content area of the browser displays the contents of the text file, which is a table of processor specifications. The table has seven columns: Name, Date, Transistors, Microns, ClockSpeed, Data, and MIPS. The data rows list various processors from 1974 to 2004, including the 8080, 8088, 80286, 80386, 80486, Pentium, PentiumII, PentiumIII, Pentium4, and Prescott.

Name	Date	Transistors	Microns	ClockSpeed	Data	MIPS
8080	1974	6000	6	2	8	0.64
8088	1979	29000	3	5	16	0.33
80286	1982	134000	1.5	6	16	1
80386	1985	275000	1.5	16	32	5
80486	1989	1200000	1	25	32	20
Pentium	1993	3100000	0.8	60	32	100
PentiumII	1997	7500000	0.35	233	32	300
PentiumIII	1999	9500000	0.25	450	32	510
Pentium4	2000	42000000	0.18	1500	32	1700
Prescott	2004	125000000	0.09	3600	32	7000

Read it into R

```
> chips = read.table("http://  
www.stat.berkeley.edu/users/nolan/data/  
chip04.txt", header = TRUE)
```

```
> class(chips)
```

```
[1] "data.frame"
```

```
> names(chips)
```

```
[1] "Name"      "Date"      "Transistors"
```

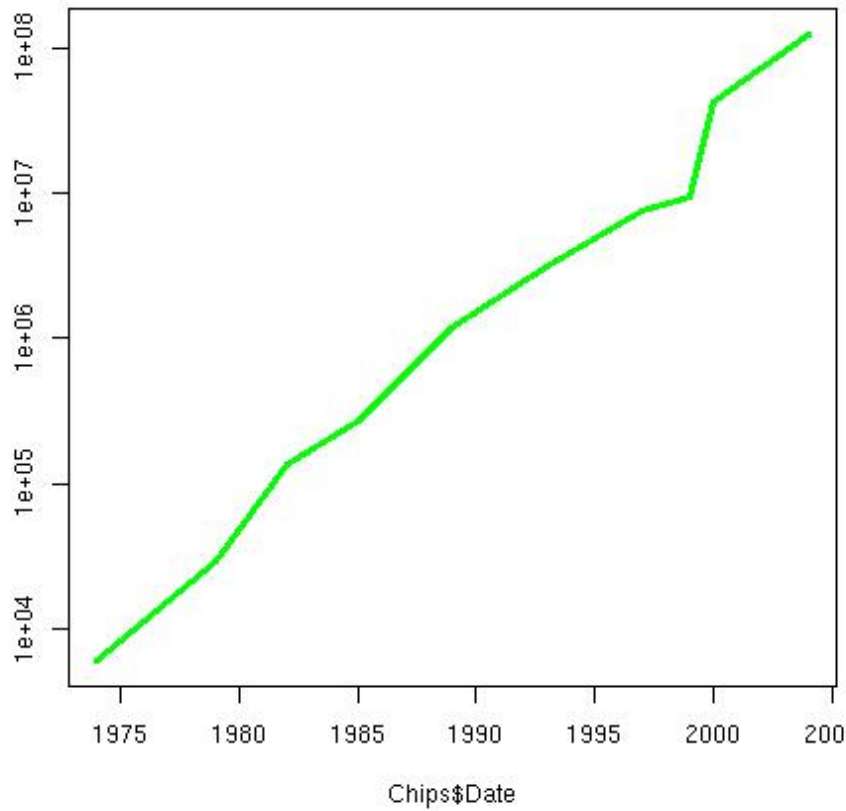
```
[4] "Microns"   "ClockSpeed" "Data"
```

```
[7] "MIPS"
```

```
> dim(chips)
```

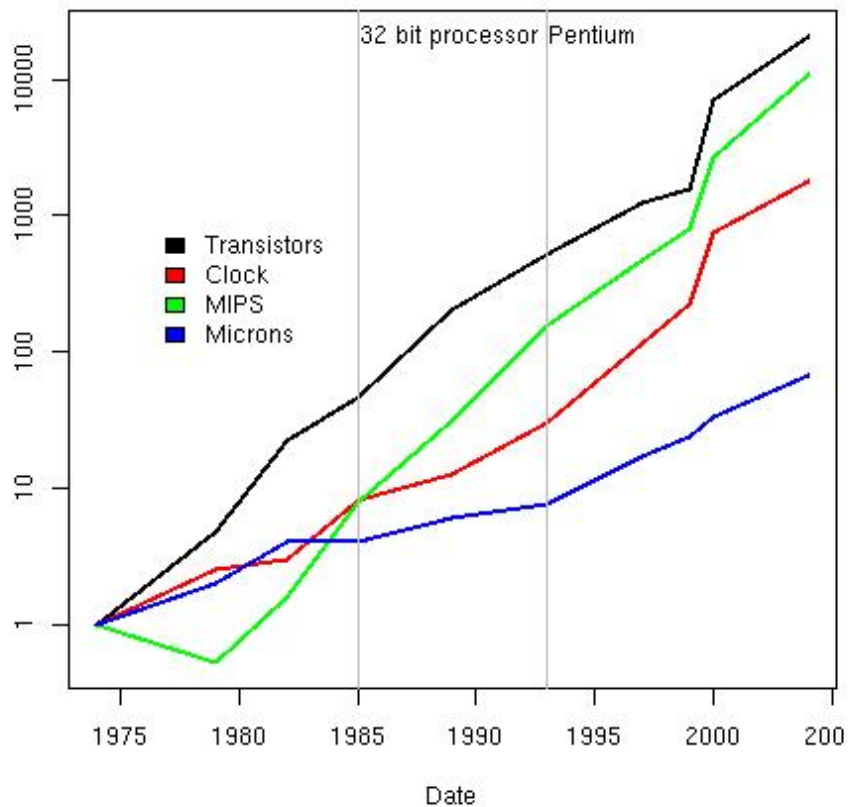
```
[1] 10 7
```


This is pretty easy to get



```
plot(chips$Date,  
     chips$Transistors,  
     type = "l",  
     lwd = 3,  
     col = "green",  
     log = "y")
```

How can we improve it
even more?



- Add more data
- Add legend for different information
- Add reference lines for important date

Review Plotting Functions

- `hist()` histogram
- `boxplot()` boxplot
- `dotchart()` dotchart
- `plot()` for scatter plots, line plots, density plots
- `barchart()`
- `pie()`
- `mosaicplot()`
- `abline()` add line to canvas
- `points()` add points to canvas
- `lines()` add line segments to canvas
- `text()` add text to canvas

Review Plot Arguments

?plot.default

- `type = "l"` "p" for points, "l" for lines, "n" for nothing
- `ylim = c(0, 1)` the range for the scale of the axis
- `xlab = "x axis label"`
- `main = "plot title"`
- `col =` vector of colors
- `log = "y"` use log scale on y axis, can be "x" or "xy"
- `lwd = 2` thickness of line
- `pch = 19` plotting character – check other numbers
- `cex = 0.5` character magnification
- `lty = 2` type of line – check other numbers
- `las = 1` 0,1,2, or 3 style of tick mark labels

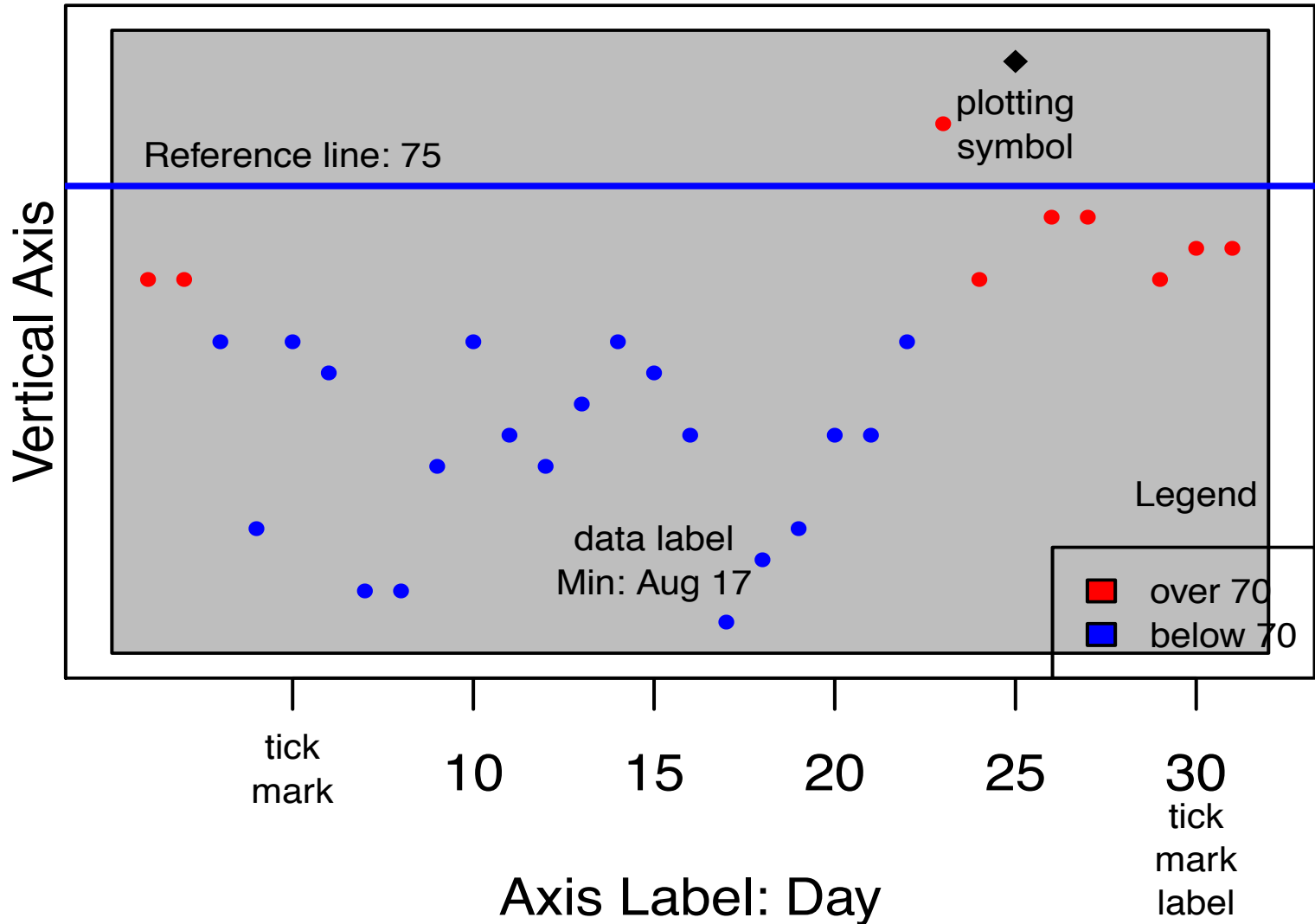
Graph Construction

Outline

- Vocabulary
- 3 Properties of good graph construction
 - Data stand out
 - Facilitate comparison
 - Information rich
- Perception
- Case studies

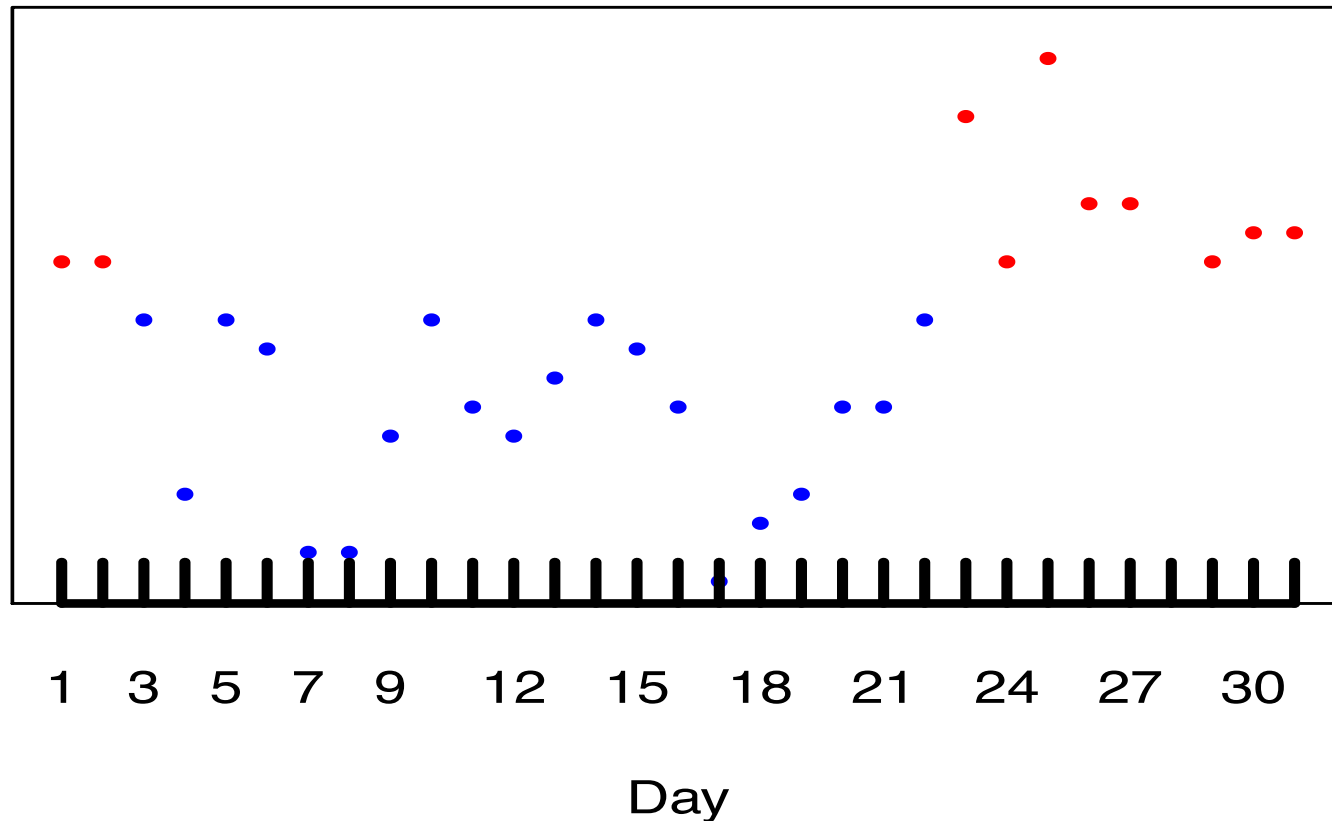
Vocabulary

Title: Temperature in August

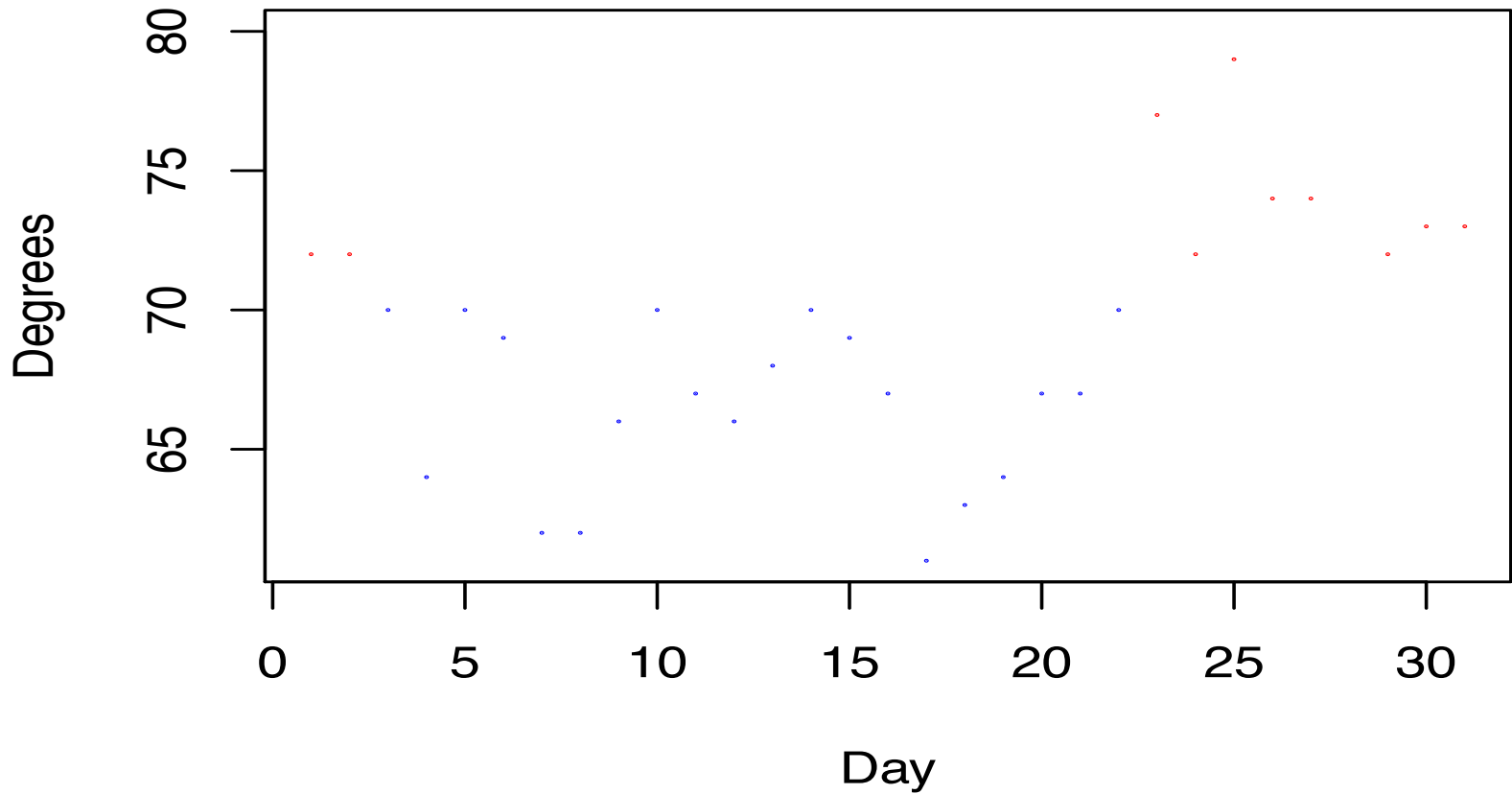


Data Stand Out

Avoid having other graph elements interfere with data



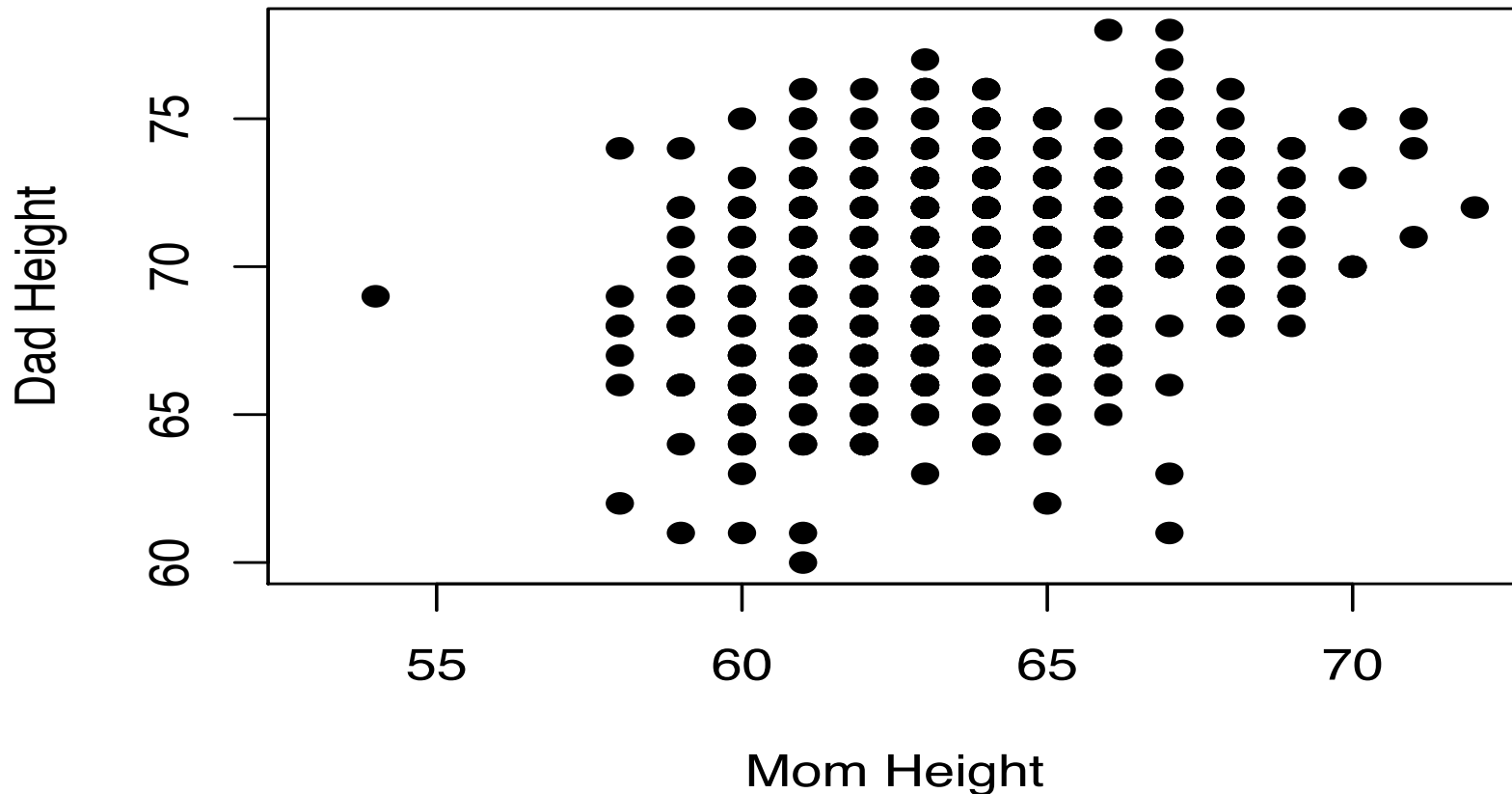
Use visually prominent symbols



Avoid over-plotting

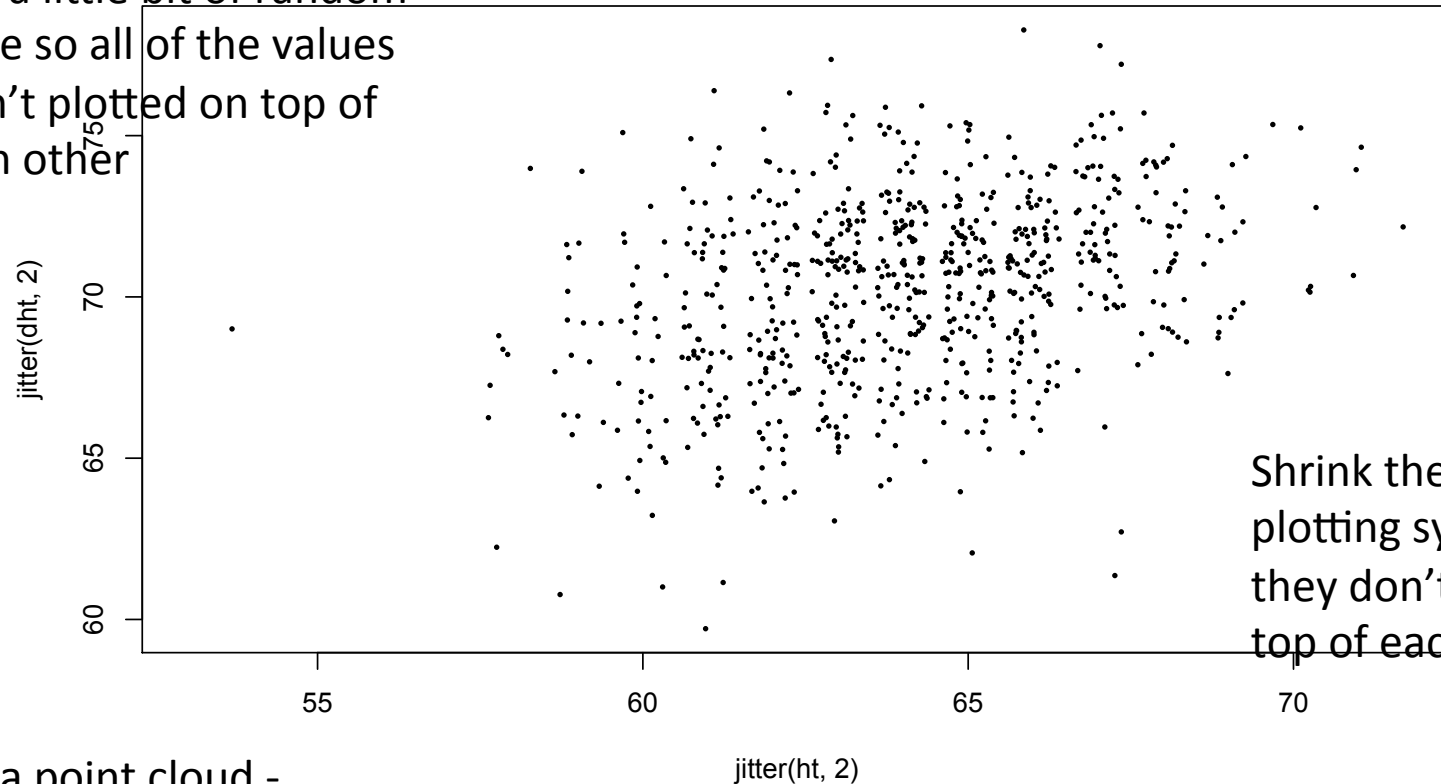
Why are there so few data points?

1200 Families



One way to avoid over plotting: Jitter the values

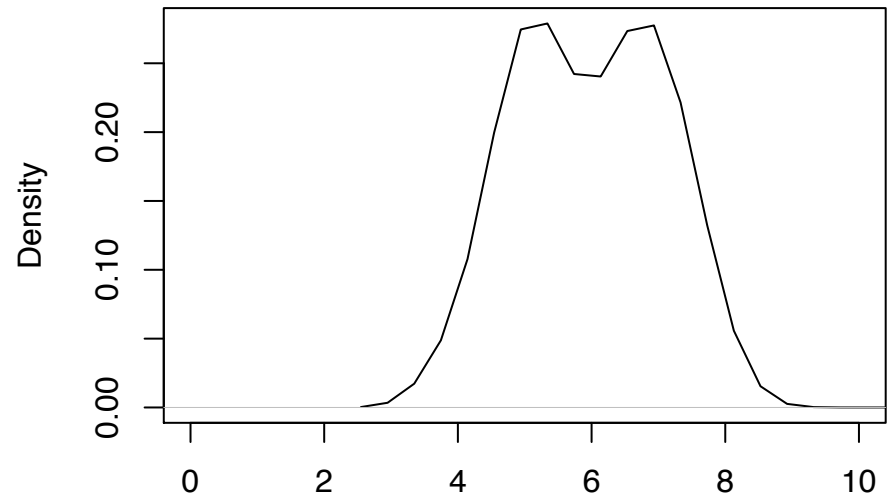
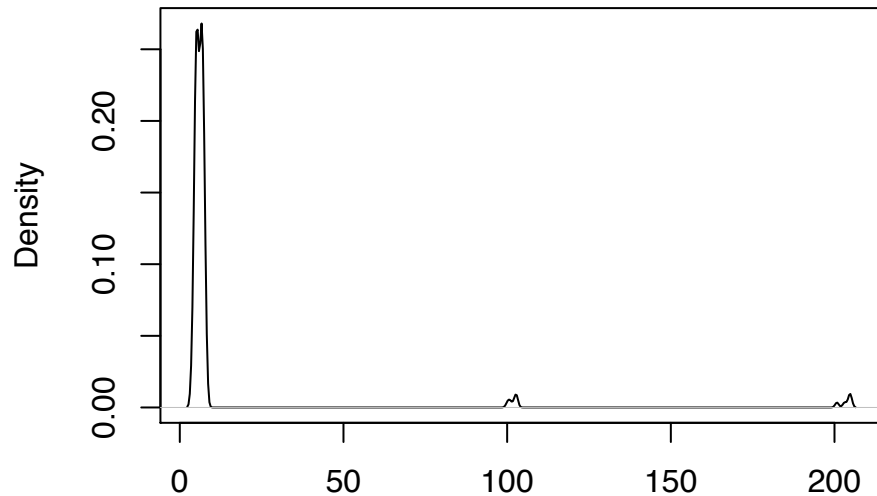
Add a little bit of random noise so all of the values aren't plotted on top of each other



Shrink the plotting symbol so they don't plot on top of each other

See a point cloud -

Different values of data may obscure each other



Most of the data are in the 0 to 10 range.
The few large values obscure the bulk of the data.
Consider mentioning these large values in a caption, instead of showing them in the plot.

Choosing the Scale of the Axis

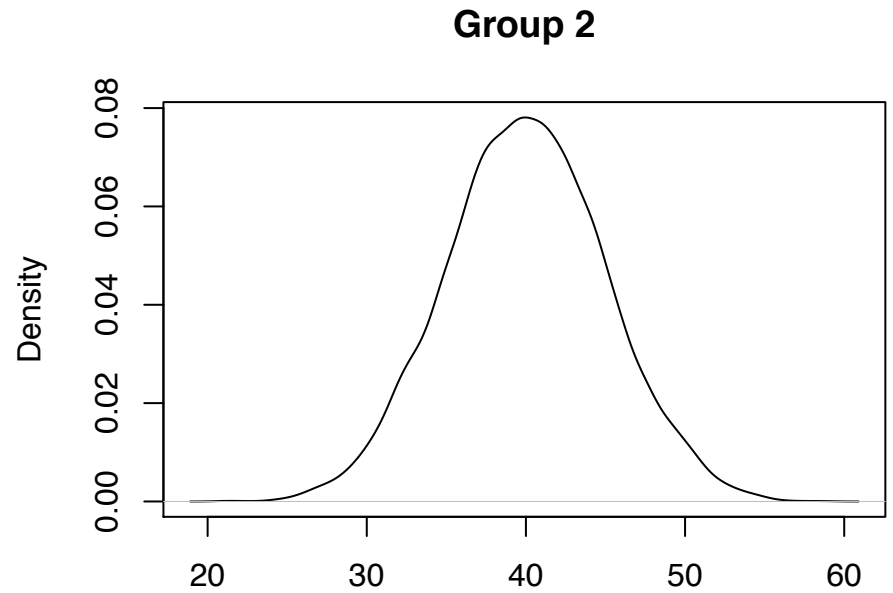
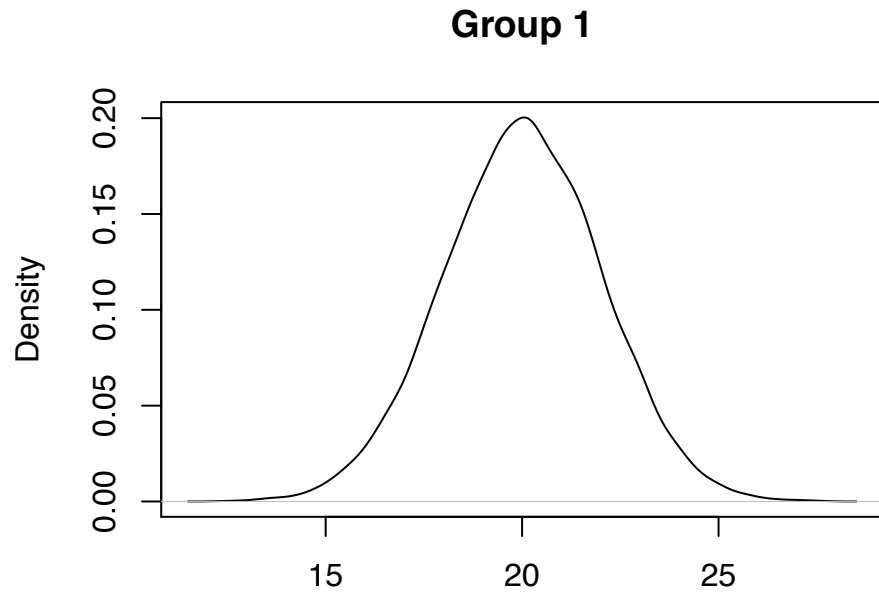
- Include all or nearly all of the data
- Fill data region
- Origin need not be on the scale
- Choose a scale that improves resolution (to be continued)

Eliminate superfluous material

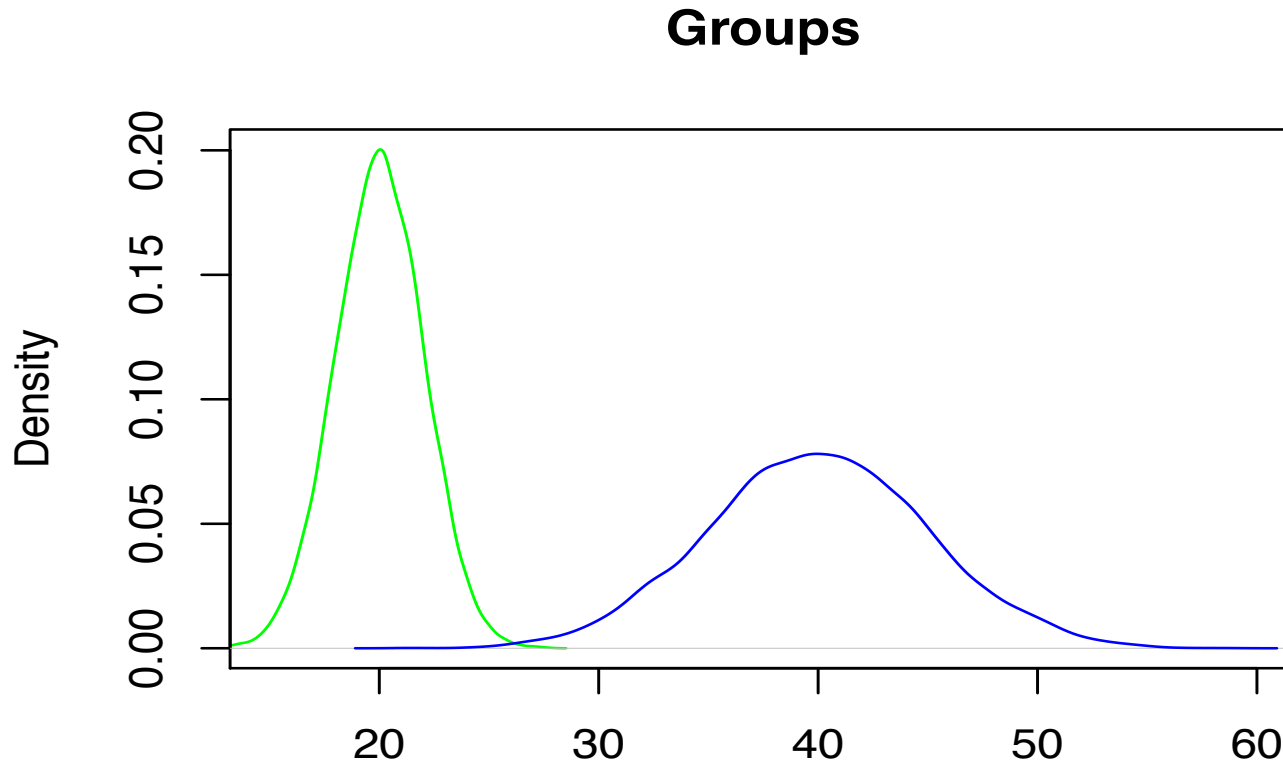
- Chart junk – stuff that adds no meaning, e.g. butterflies on top of barplots, background images
- Extra tick marks and grid lines
- Unnecessary text and arrows
- Decimal places beyond the measurement error or the level of difference

Facilitate Comparisons

Put Juxtaposed plots on same scale



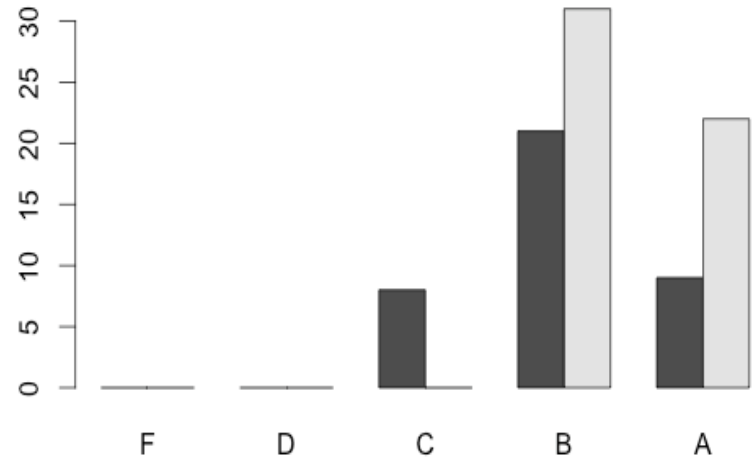
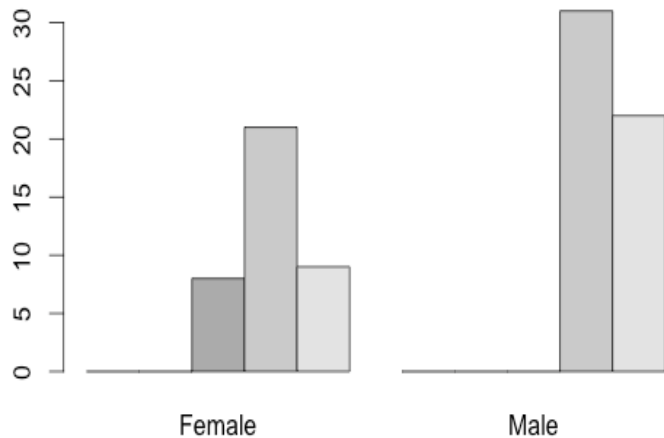
Make it easy to distinguish elements of superposed plots (e.g. color)



Choosing the Scale

- Keep scales on x and y axes the same for both plots to facilitate the comparison
- Zoom in to focus on the region that contains the bulk of the data
- These two principles may go counter to one another
- Keep the scale the same throughout the plot (i.e. don't change it mid-axis)

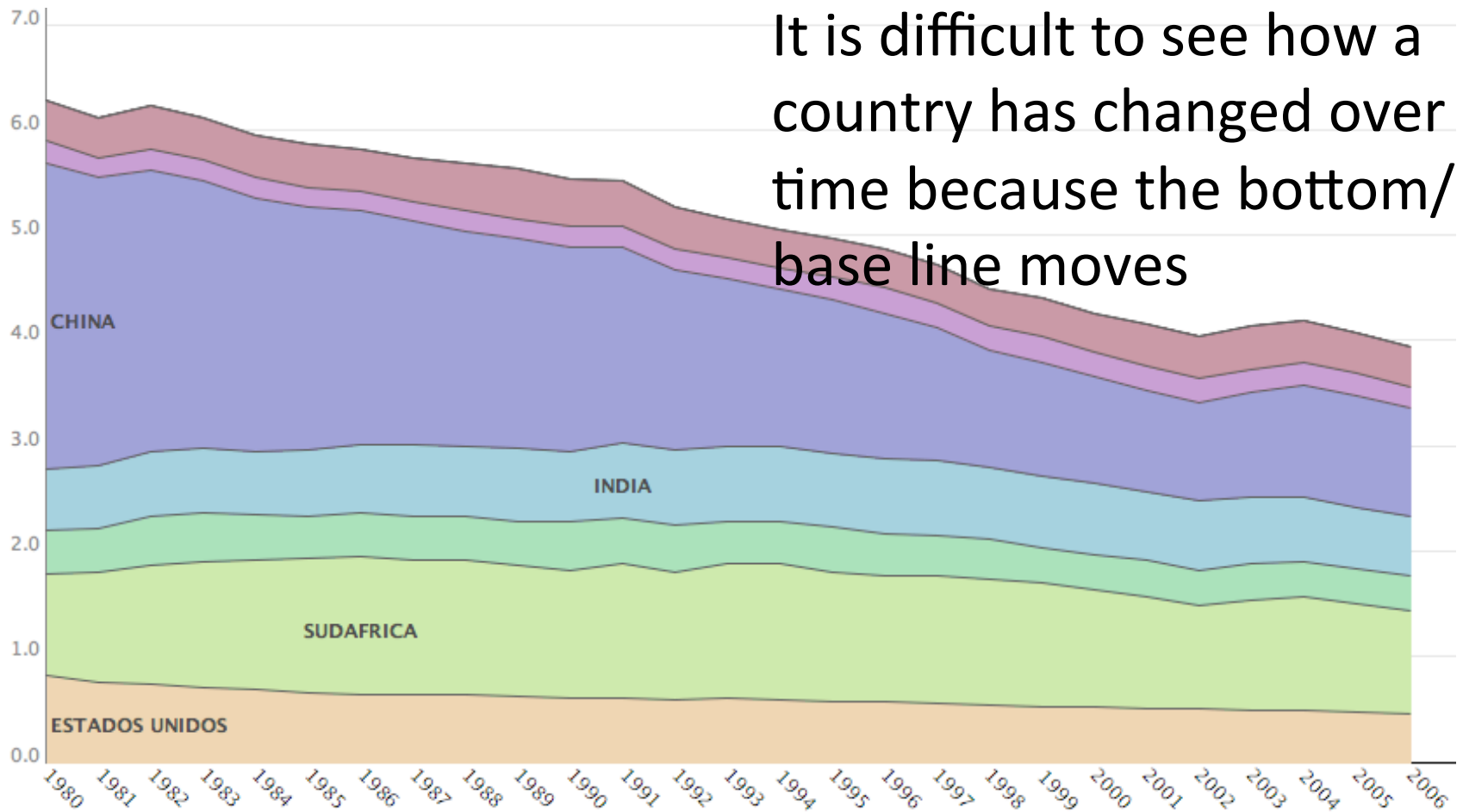
Emphasizes the important difference



Which of these side-by-side bar plots emphasizes the important difference?

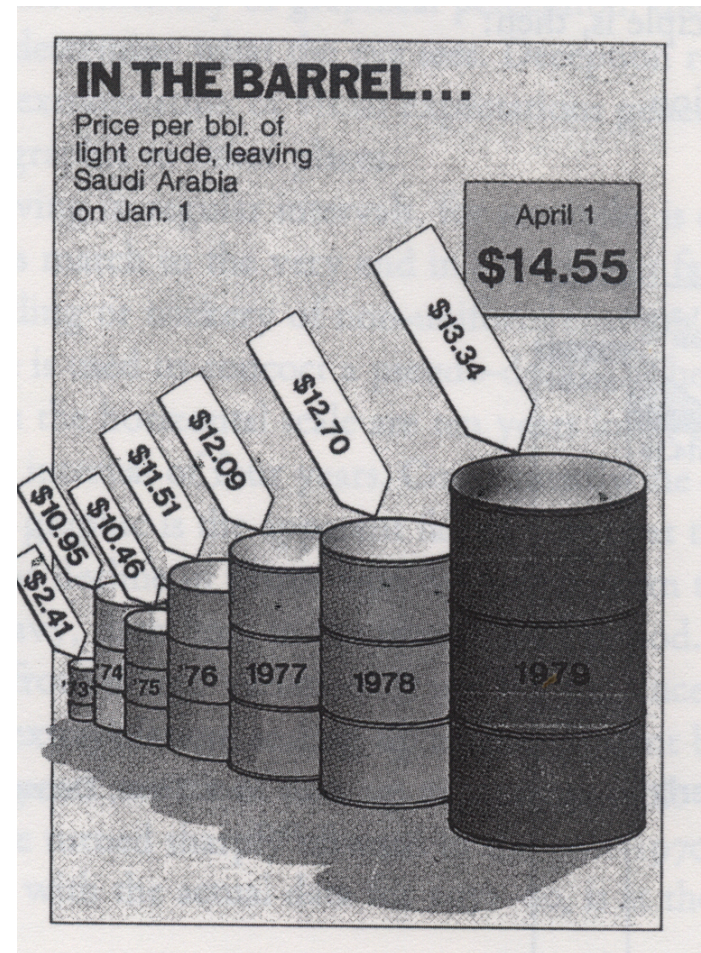
Avoid Jiggling the baseline

It is difficult to see how a country has changed over time because the bottom/base line moves



Comparison: volume, area, height

We naturally compare the volume of the barrels, but the change is really the height of the barrels



Information Rich

How to make a plot information rich

- Describe what you see in the **Caption**
- Add context with **Reference Markers** (lines and points) including text
- Add **Legends** and **Labels**
- Use color and plotting symbols to add more information
- Plot the same thing more than once in different ways/scales
- Reduce clutter

Captions

- Captions should be comprehensive
- Self-contained
- Captions should:
 - Describe what has been graphed
 - Draw attention to important features
 - Describe conclusions drawn from graph

Good Plot Making Practice

- Put major conclusions in graphical form
- Provide reference information
- Proof read for clarity and consistency
- Graphing is an iterative process
- Multiplicity is OK, i.e. two plots of the same variable may provide different messages
- Make plots data rich

Perception

Color, shape (including banking) can
affect your ability to make good
comparisons

Banking: Aspect Ratio

- The height/width of the data region was selected to be about 1 so that the trend line is at about 45 degrees.
- The Aspect ratio affects our visual decoding of the rate of change
- The banking to 45 degrees helps us see rate of change
- The ability to effectively judge rate of change allows us to see important patterns in data

Banking at 45 degrees

- Roughly: Examine the absolute value of the orientation of segments, they should be centered at 45 degrees.
- Transformations to improve the aspect ratio uncovers the structure of the relationship between variables
- Easier to see important features

Bank to 45 degrees

