Visualizing Data: Basic Plot Types

Data Science 101

Stanford University, Department of Statistics

Agenda

- Today's lecture focuses on these basic plot types:
  - bar charts
  - histograms
  - boxplots
  - scatter plots
  - densities
- Which visualization is best depends on
  - whether the data are **univariate** or **bivariate** data
  - whether the variables are **discrete** or **continuous**
  - context

Discrete Variables (a.k.a. Categorical Variables)

Roughly speaking, **discrete** variables take only a few unique values (we might turn them into **factors** in R).

- `Win`, `Lose`, `Tie`
- `Treatment` vs `Control`
- Can be numeric, e.g. cars grouped by 4, 6, or 8 cylinders.

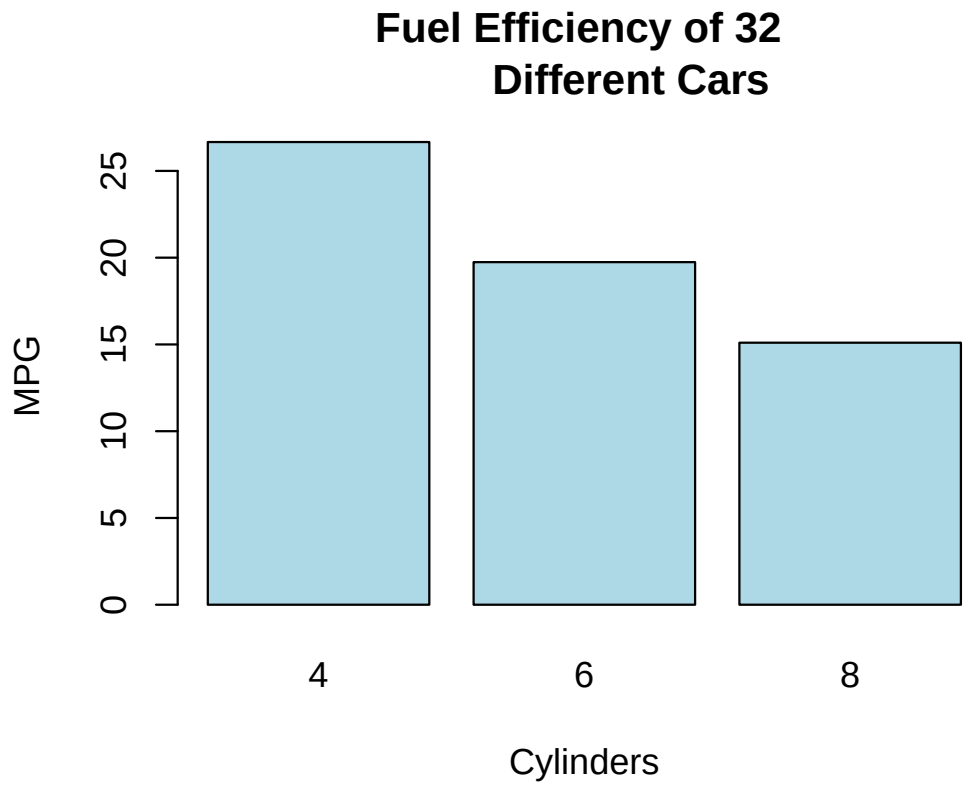
```
by(mtcars$mpg, mtcars$cyl, mean)
```

```
mtcars$cyl: 4
[1] 26.66364
------------------------------------------------------------
mtcars$cyl: 6
[1] 19.74286
------------------------------------------------------------
mtcars$cyl: 8
[1] 15.1
```

Bar chart

Bar chart: height of the bar is proportional to the number of data falling in that category, or proportional to some **summary statistic** such as the mean.

```
barplot(by(mtcars$mpg, mtcars$cyl, mean),
        main = "Fuel Efficiency of 32
        Different Cars",
        xlab = "Cylinders", ylab="MPG",
        col="lightblue")
```



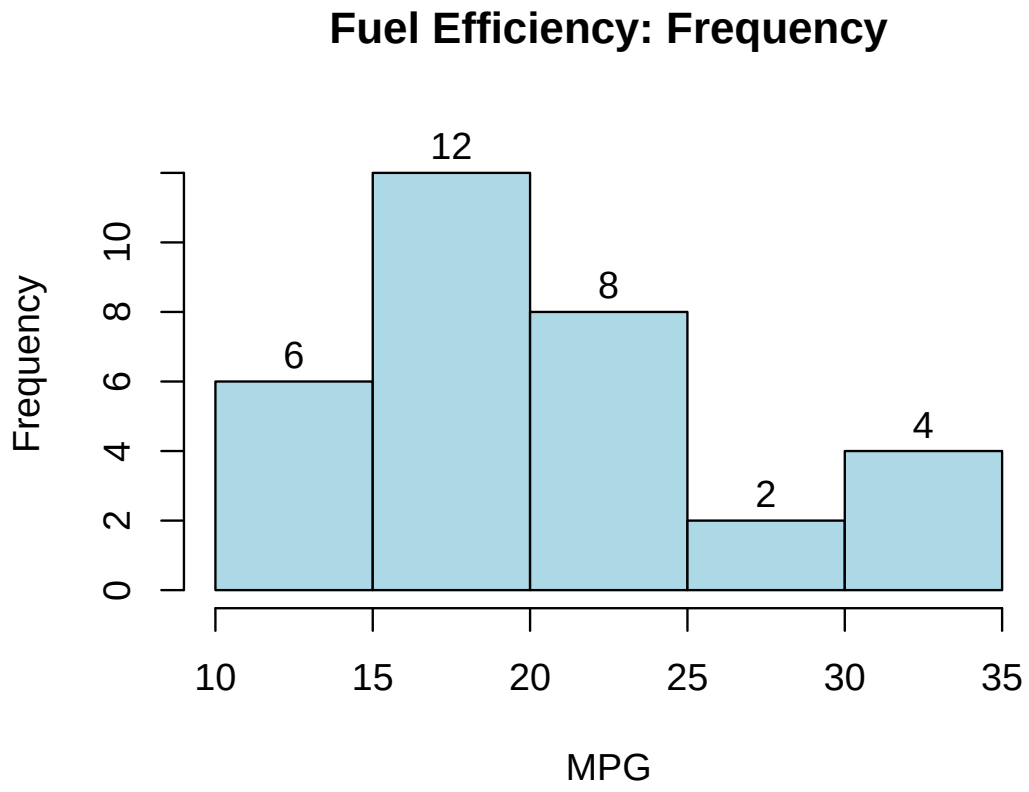
Continuous Variables

- In theory, **continuous** variables may take infinitely many values (though in practice resolution is limited by the measuring apparatus' precision).
- For example, **mpg** is a continuous variable: a car's average fuel economy could be any number between 0 and ∞.
- Histograms and density plots are often used to display continuous univariate data.

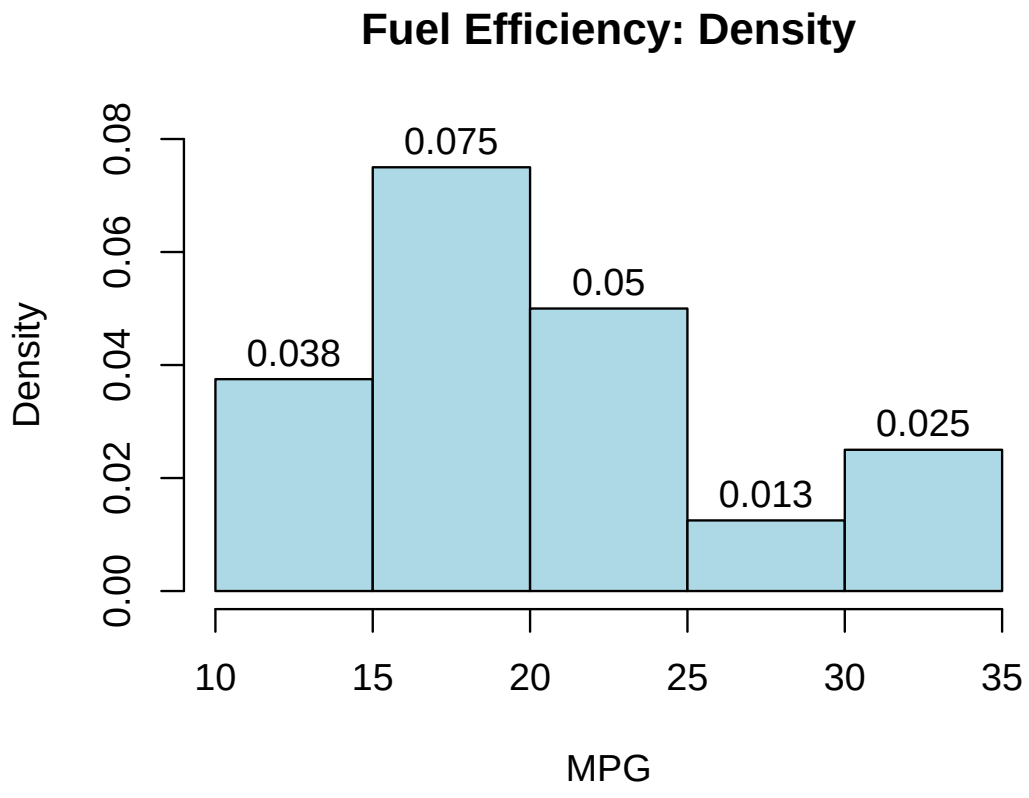
Histogram

A **Histogram** displays quantitative data like a bar graph but it allows for unequal block lengths.

```
hist(mtcars$mpg, main = "Fuel Efficiency: Frequency",
     xlab="MPG", col="lightblue",
     labels = TRUE, freq = TRUE,
     ylim = c(0, 13))
```

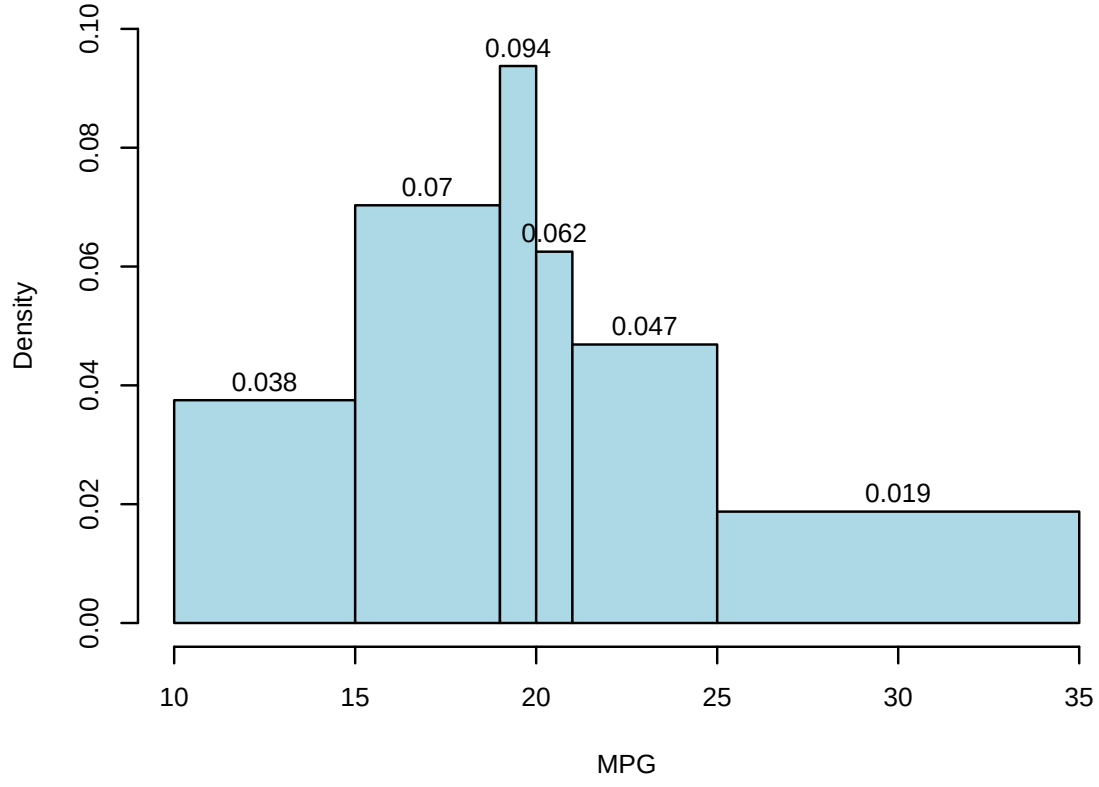


```
hist(mtcars$mpg, main = "Fuel Efficiency: Density",
     xlab="MPG", col="lightblue",
     labels = TRUE, freq = FALSE, ylim = c(0, 0.08))
```



With more breaks

```
par(cex=.7)
hist(mtcars$mpg, main = "", xlab="MPG",
     col="lightblue",
     labels=TRUE, nclass=12, freq = FALSE,
     breaks = c(10, 15, 19, 20, 21, 25, 35),
     ylim = c(0, 0.1))
```



Main Principle of the Histogram

- Area is proportional to frequency, so the percentage falling into a block can be discerned without a vertical scale (since the total area equals 100%).
- But it's helpful to have a vertical scale (*density scale*). Its unit is '% per unit', so '% per mpg' in above example.

- *Density*: The bar height tells how many cars are in one unit on the horizontal scale.
- The highest density ($0.094 = 9.4\%$) is between 19 and 20 mpg. Even though only 3 cars fall into this range, adjusting for width, more cars fall in that category than any other.
- By contrast, the density is only 0.019 for the 6 cars with $mpg > 25$.

Note: See examples for histograms with ggplot e.g. here and for density plots e.g. here.

Histograms show percentages

Percentages (relative frequencies) are given by:

$$Area = Height * Width$$

For example, 18.75% fall into the most fuel efficient category (between 25 and 35 mpg) because the corresponding area is

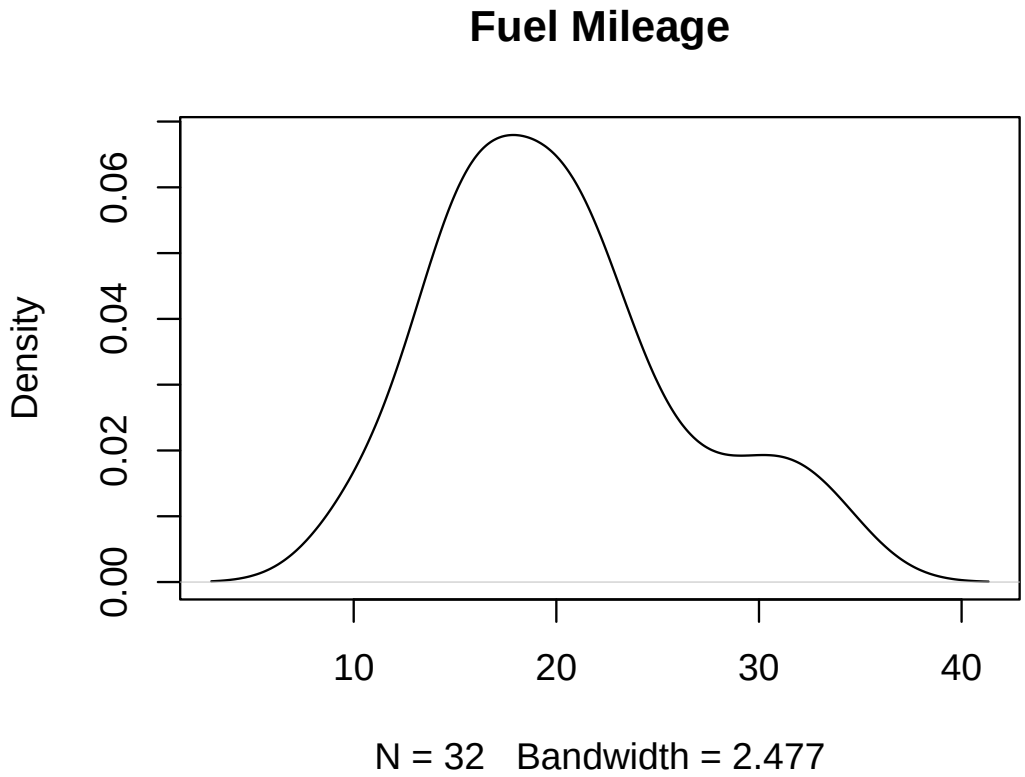
$$Area = (10\,mpg) * (0.01875\,per\,mpg)$$

Alternatively, eyeballing shows this area makes up roughly 1/5 of the total area.

Density plots

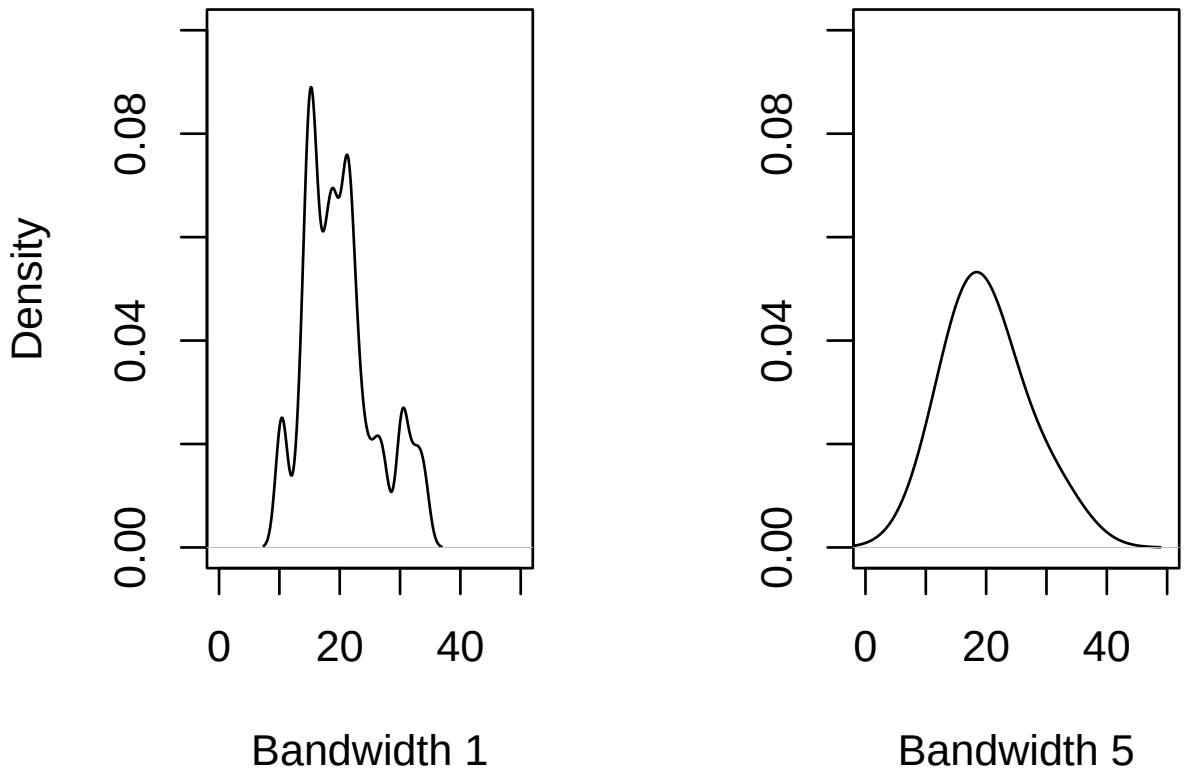
A density plot is like a "smoothed" histogram.

```
plot(density(mtcars$mpg), main = "Fuel Mileage")
```



Choice of Bandwidth: How Many Peaks?

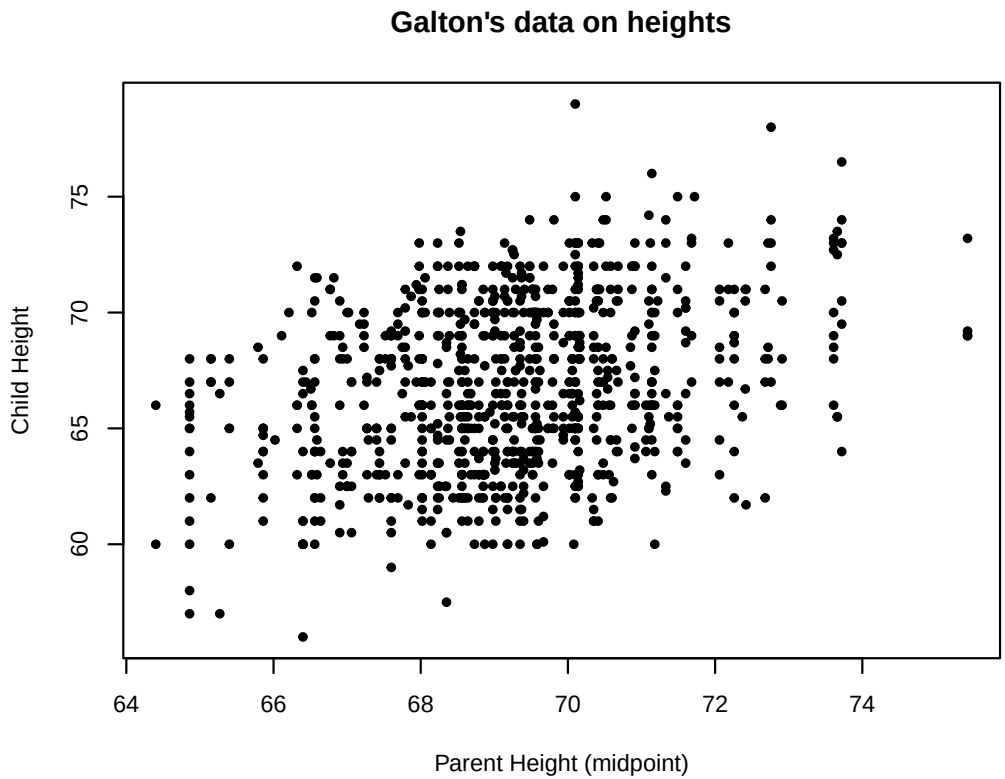
- **Small bandwidths** capture many local peaks but may be unstable ('wiggly') elsewhere.
- **Big bandwidths** are 'smoother' (fewer peaks).



Scatter plot

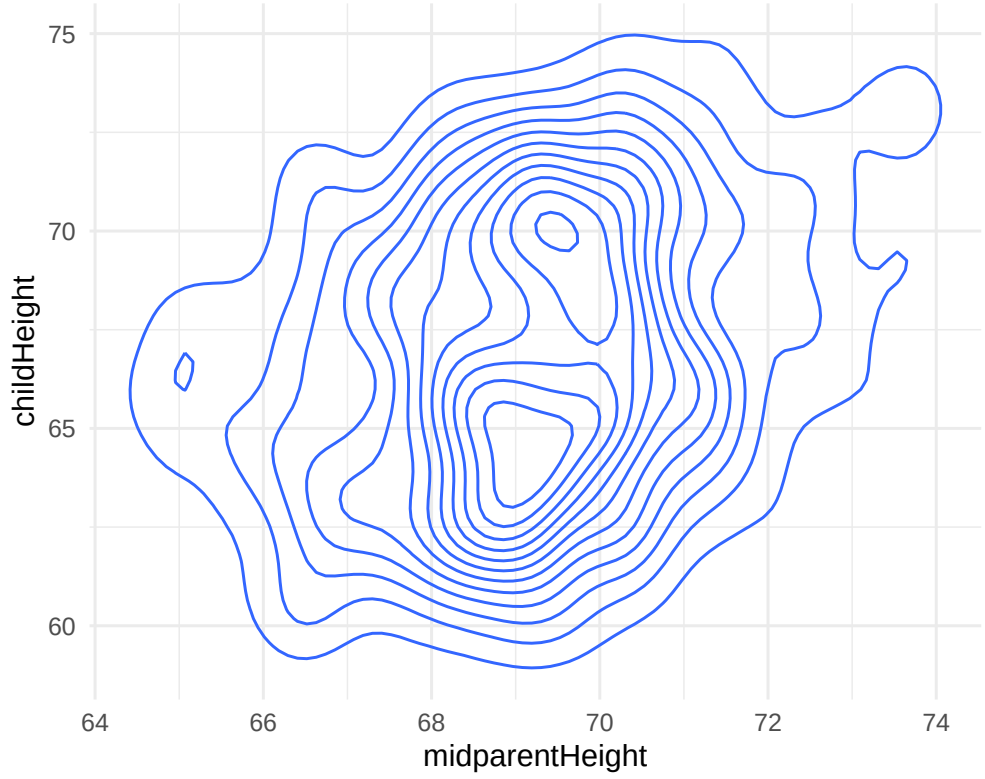
For **bivariate** data where *both* variables are *continuous*, scatterplots are the standard way to display association.

```
par(cex=.7)
plot(GaltonFamilies$midparentHeight,
     GaltonFamilies$childHeight,
     main="Galton's data on heights",
      ylab="Child Height",
     xlab="Parent Height (midpoint)",
     pch=20) # pch denotes point type
```



- There is a 2-dimensional analogue to the density plot which can be used to smooth a scatter plot:

```
ggplot(GaltonFamilies, aes(x=midparentHeight,
  y=childHeight)) + geom_density2d() + theme_minimal()
```

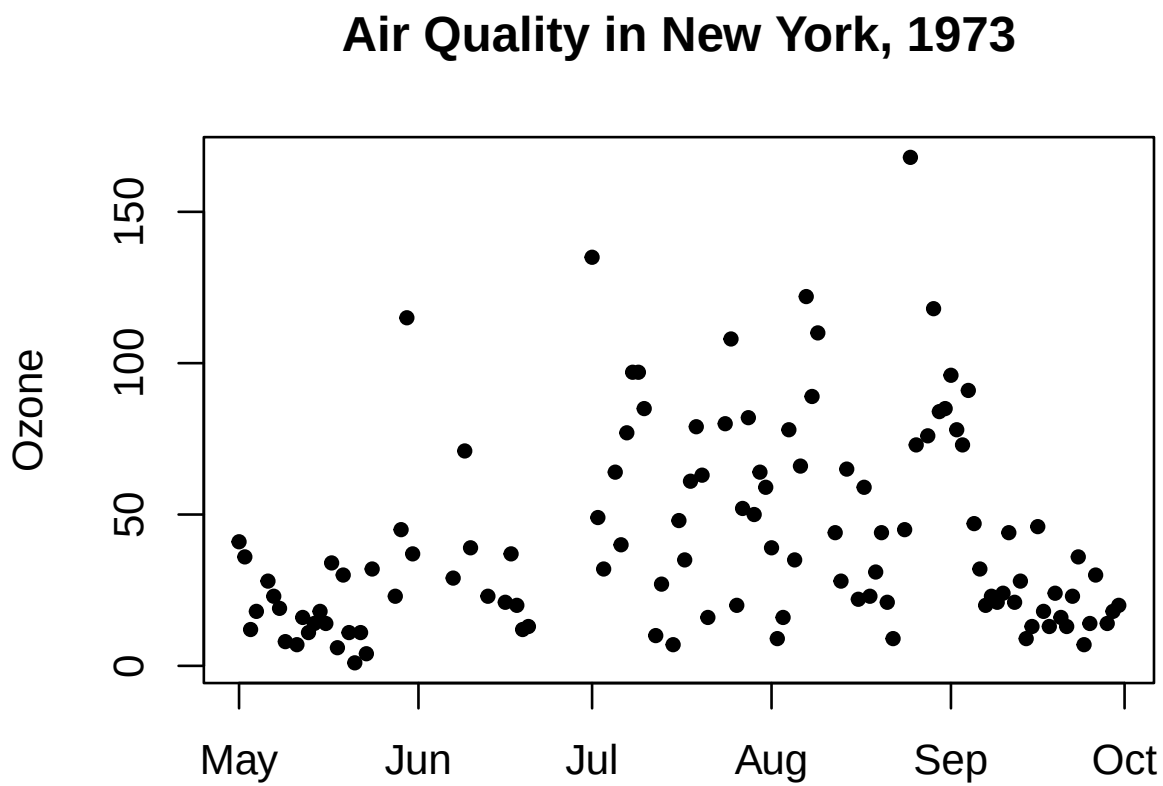


- Again there is a price to pay for the smoother appearance: we don't see the data any more.

Time Series

- A special case of the scatter plot is a 'time series', which displays a quantity that was measured at various time points.
- By convention, time is always set to be the `x` variable.

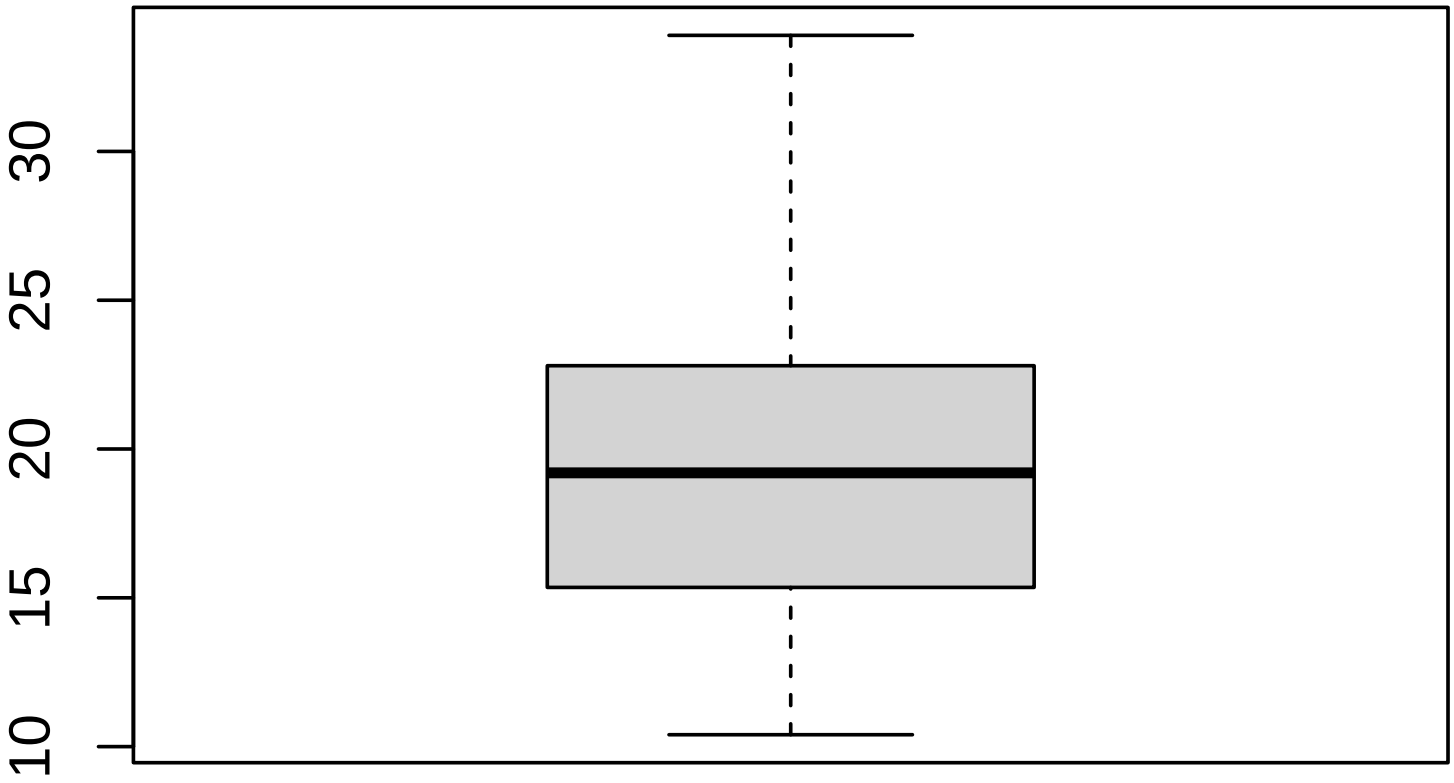
```
data("airquality")
plot(as.Date(paste(1973, airquality$Month, airquality$Day,
                   sep="-")), airquality$Ozone,
     xlab="", pch=20,
     main = "Air Quality in New York, 1973", ylab="Ozone")
```



Box Plots

Box plots provide a compact summary of a variable–both of its median as well as a depiction of the spread in the data.

```
boxplot(mtcars$mpg)
```



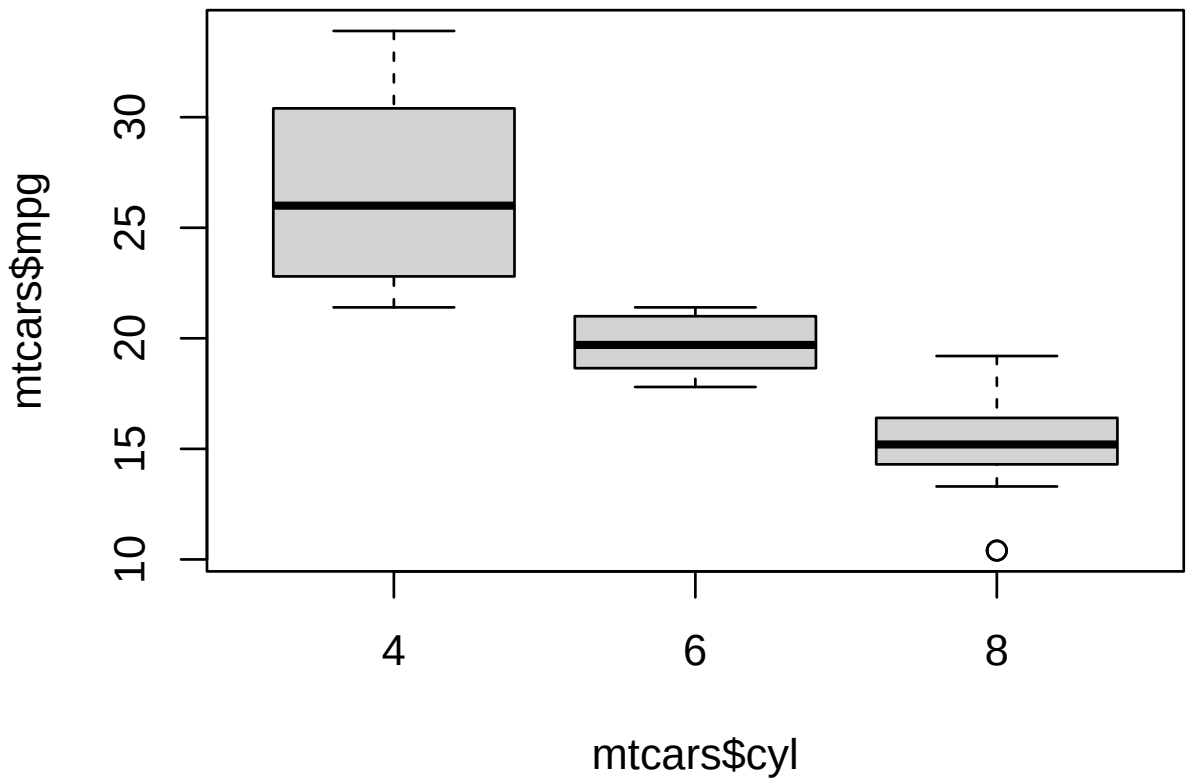
Details about the box plot

- The box plot depicts the following summary measures:
  - The median
  - The first and third quartiles; these are the **hinge** values, which represent the extent of the box
  - The minimum of the data; this is the **lower whisker** (If the minimum is far away from the box, then a different definition applies.)
  - The maximum of the data; this is the **upper whisker** (Likewise, a different definition applies if the maximum is far from the box)
  - Points that lie outside 1.5 times the interquartile range from the **hinges**, as **individual points** (sometimes denoted **outliers**)

Box plot (a.k.a. 'box-and-whisker' plot)

- The box plot is also convenient for displaying **bivariate** data, where one variable is *continuous* and the other is *categorical*.
- Box plots give useful summaries of the **continuous** variable, separately for each value of the **categorical** variable.

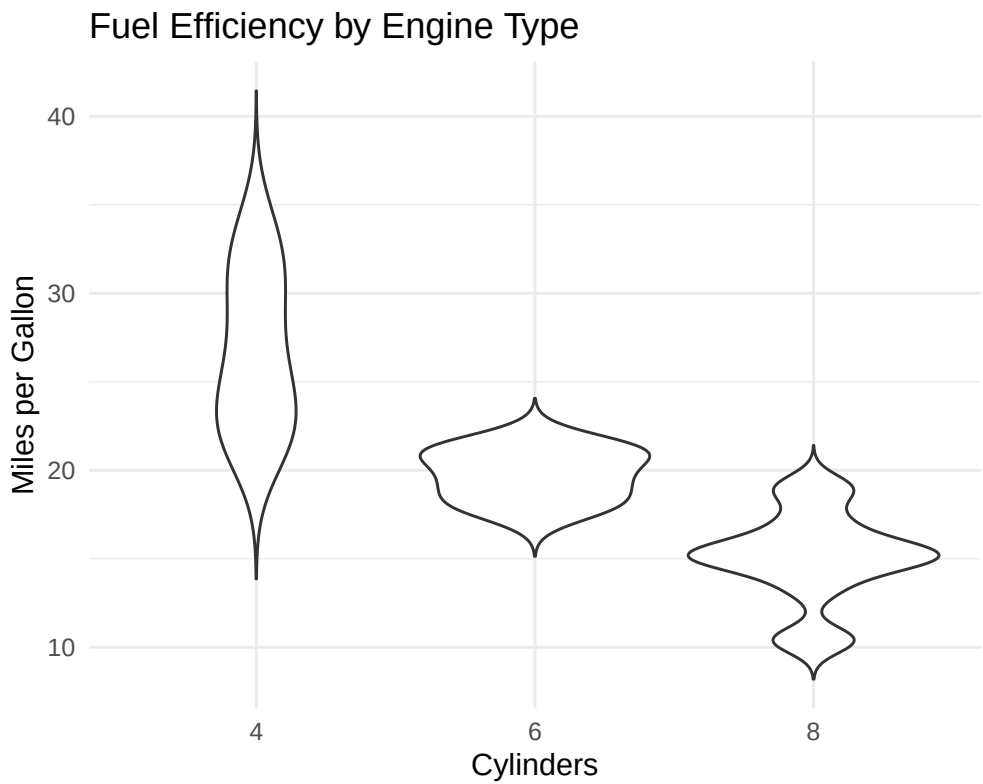
```
boxplot(mtcars$mpg ~ mtcars$cyl)
```



Violin plots

A violin plot is similar to a box plot, but has a (rotated) density plot on each side.

```
p <- ggplot(mtcars, aes(factor(cyl), mpg)) +
  geom_violin(trim = F) +  theme_minimal() +
  labs(x="Cylinders", y="Miles per Gallon",
       title="Fuel Efficiency by Engine Type")
p
```




```
# note: use e.g. + geom_boxplot(width=0.1, alpha=0.2)
# to additionally include a boxplot in the violin plot
```

Which Visualization to Use

- Visualization should be chosen based on the number and type of variables and the research question at hand
- Scatterplots can be useful to highlight the relationship between two variables
- Compared with barplots, boxplots convey more information about sampling variability (but typically require more text to explain to the reader)
- Density plots and histograms contain more detailed information about the data, but they may require some calibration for the appropriate amount of "wigliness"

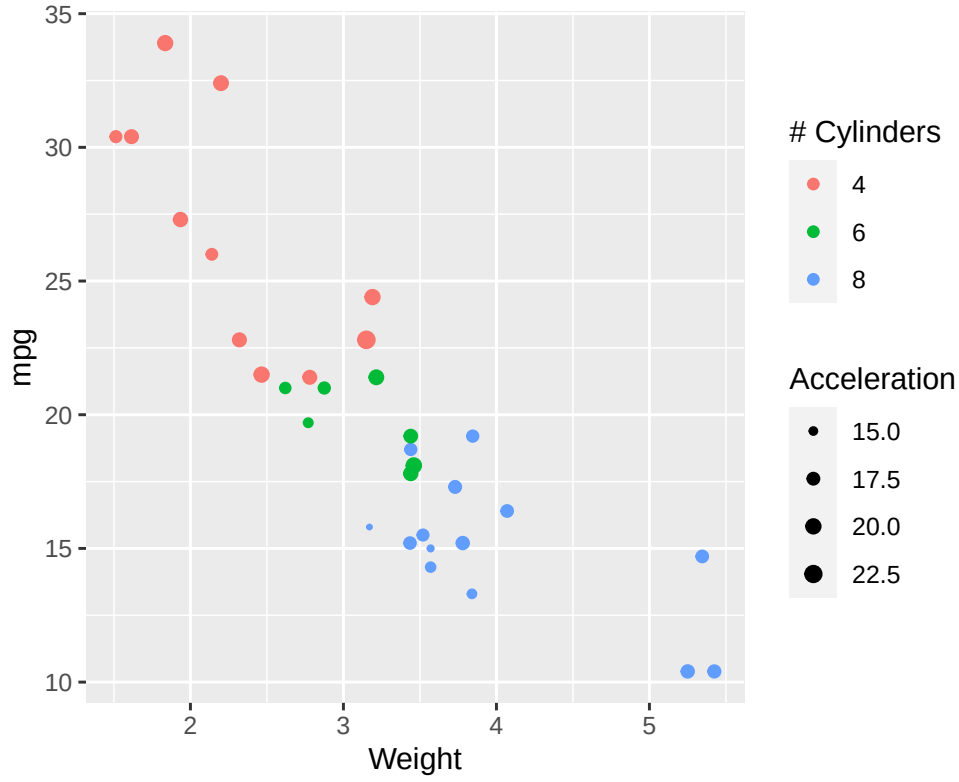
Scatter plots with more than two variables

If there are more than two variables, then these can be added to the scatter plot via **size**, **shape**, and **color**.

Example: Scatter Plot with Four Variables

A single graphic displaying fuel economy (mpg) by cylinders (cyl), weight (wt), and acceleration (qsec).

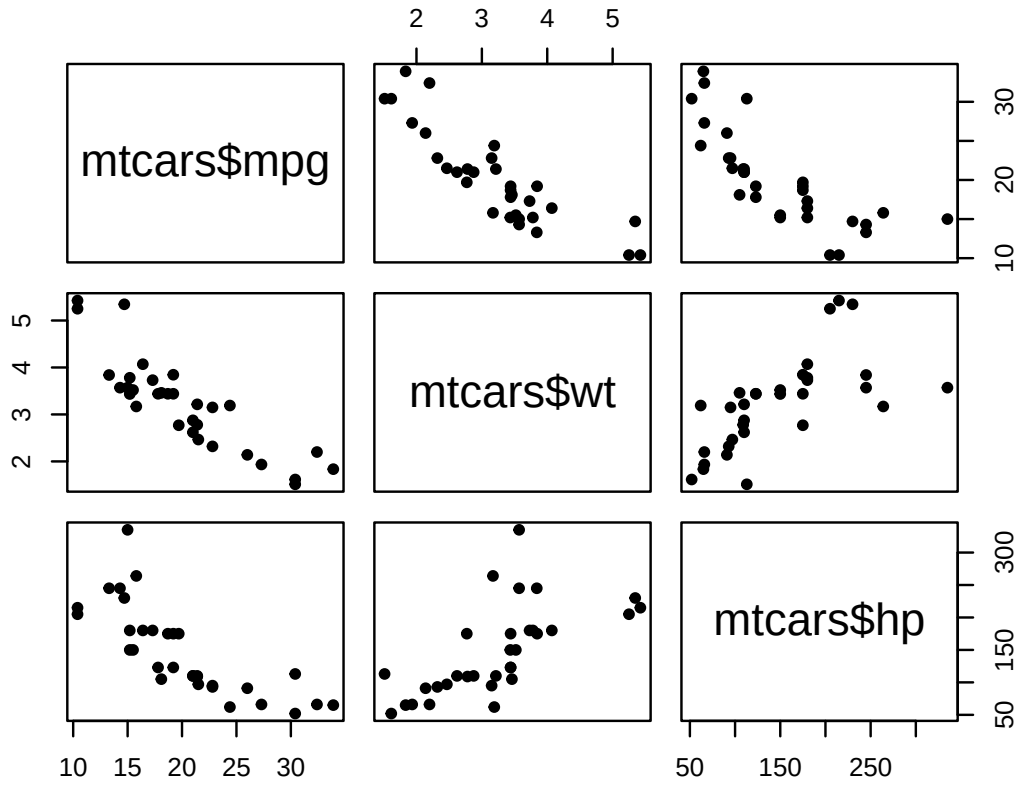
```
ggplot(mtcars,aes(x=wt,y=mpg, size=qsec, col=factor(cyl))) +
  geom_point() +
  scale_size_continuous(range = c(0.5,2.5)) +
  labs(x="Weight", size="Acceleration", col="# Cylinders")
```



Scatter plot matrices

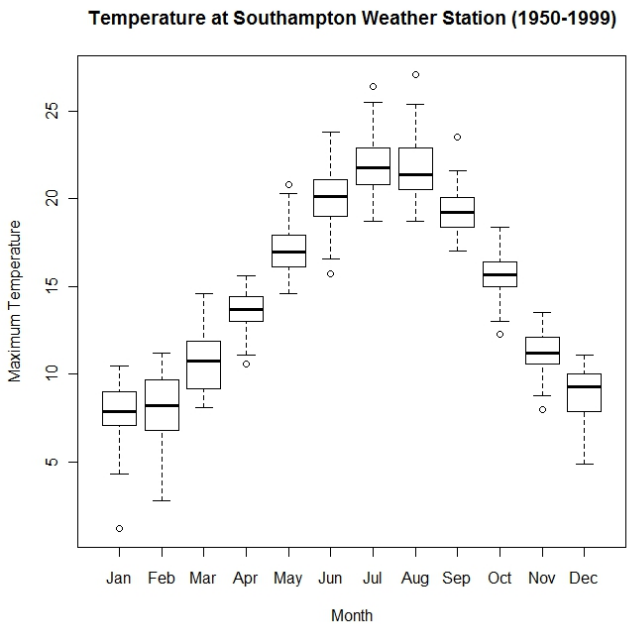
One way to visualize all pairwise scatters between the variables is with a scatter plot matrix:

```
pairs(~mtcars$mpg + mtcars$wt + mtcars$hp, pch=19, cex = 0.8)
```



Small multiples

- Small multiples refers to a series of charts using similar scales and axes, arranged in a lattice or grid.
- This is a good way to provide context to a graphic as the human brain is good at scanning the information in multiple plots.



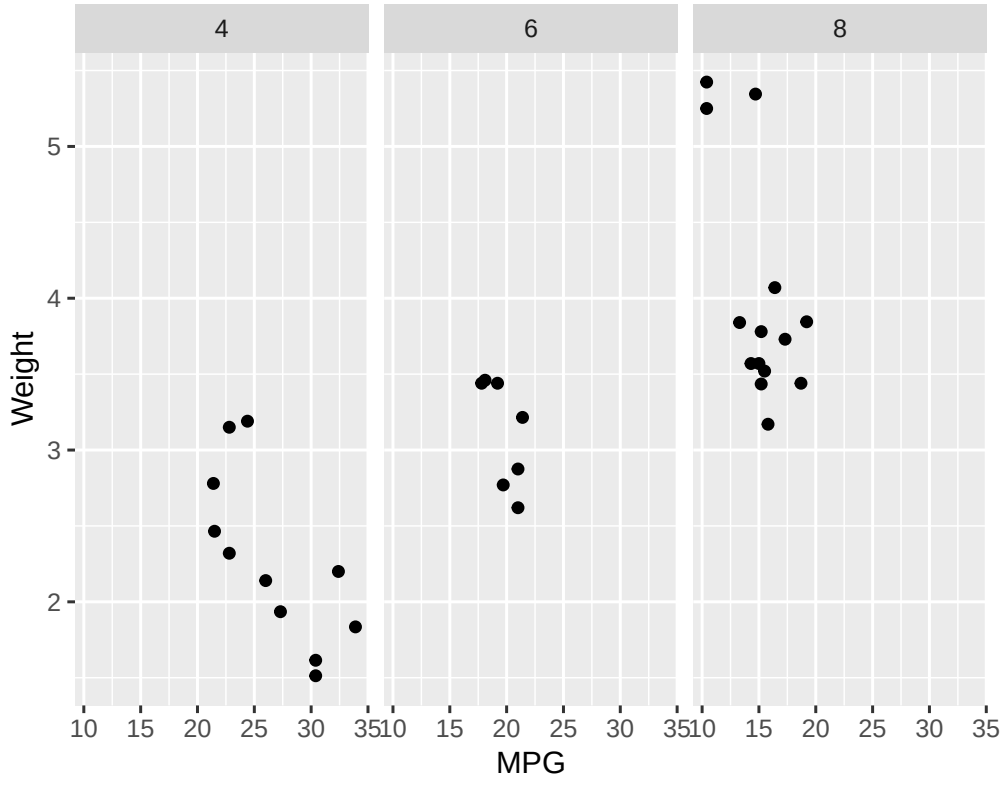
source

Producing small multiples

- A scatter plot matrix is a small multiple design
- More generally, we can produce small multiples in ggplot by **faceting**

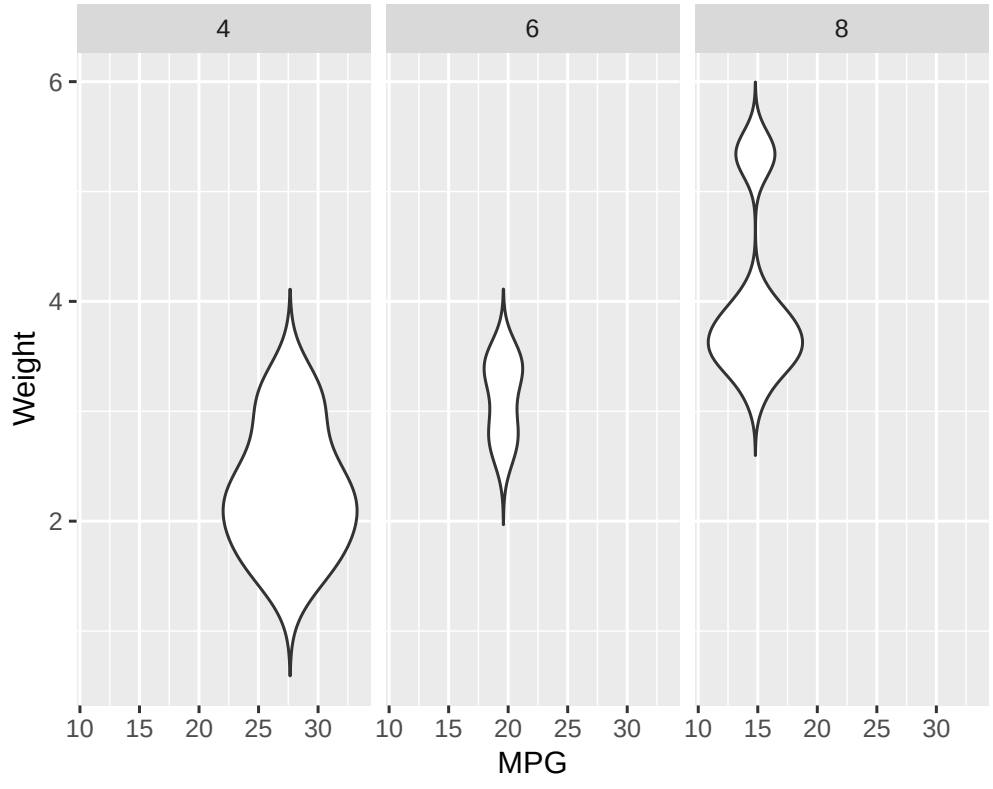
Faceting in ggplot

```
ggplot(mtcars, aes(mpg, wt)) +
  geom_point() +
  facet_grid(. ~ cyl) +
  labs(y = "Weight", x = "MPG")
```



Alternative way to facet

```
ggplot(mtcars, aes(mpg, wt)) +
  geom_violin(trim = F) +
  facet_grid(. ~ cyl) +
  labs(y = "Weight", x = "MPG")
```



Ugly Data Visualization

- "Chartjunk:" unnecessary graphics on visualizations, aspects of the visualization that don't convey additional information but distract from the point

Thursday, April 4, 2002 ◆ 3

Private gifts reach all-time high

Donations to Stanford are second largest in the nation

By KRISTIN SCHLECKER
CONTRIBUTING WRITER

Stanford received the second-highest amount of private donations to any American college or university last year, a RAND Corporation study recently reported. Stanford, with contributions totaling $469 million, slipped from its top position on the 2000 list to second behind Harvard's $683 million in 2001.

Overall, private contributions to United States colleges and universities reached an all-time high of $24.2 billion in 2001. That sum represents an increase of approximately $1 billion, or 4.3 percent, over the 2000 figure.

"The total is surprisingly large in light of the overall sharp decline in stock market values over this period," RAND's Council for Aid to Education said. The non-profit Council for Aid to Education has tracked gifts to institutes of higher education in an annual survey since 1952.

Although total giving reached a record dollar amount, the rate of increase actually slowed in 2001. In fiscal year 2000, giving had increased by 13.7 percent.

"It wasn't a banner year, but it was still a lot better than it might have been," Ann E. Kaplan, the study director, told The New York Times.

Most of the growth in donations came from non-corporate foundations, which increased their contributions by 15.1 percent in the last year, for a total of $6 billion. Foundations now comprise a quarter of all voluntary giving to colleges and universities, in part due to a growing trend on the part of families to establish their own foundations. Still, this year's surge defied even researchers' predictions.

"I expected growth in giving by foundations to be there, but I didn't expect it to be nearly as high as it was," Kaplan said. "There was so much bad news in the stock market, especially during that fiscal period, with the Standard & Poor's index, the New York Stock Exchange and the Nasdaq [index] all down."

Alumni gifts, traditionally the largest segment of private funding, held nearly steady in 2001. Alumni still edged out foundations with a 28 percent share, but grew only 0.4 percent to $6.83 billion.

Non-alumni individuals, corporations and religious and other organizations accounted for $5.9 billion, $4.95 billion, and $1.82 billion respectively.

Nationwide, the gifts were split fairly evenly between current operating expenses and capital improvement programs such as buildings or endowments. Gifts for operating purposes had the slight edge this year, with $12.2 billion versus $12.0 billion for capital purposes.

Stanford's highest giving priority now is its Campaign for Undergraduate Education, launched by University President Hennessy in 2000 with the goal to raise $1 billion over five years.

This is "the largest campaign specifically for undergraduate education ever undertaken by any university," Hennessy wrote in a 2000 letter announcing the program.

As of March 31, Stanford's campaign had already raised $733 million, although a portion of that total reflected as-yet unmet commitments to match donations.

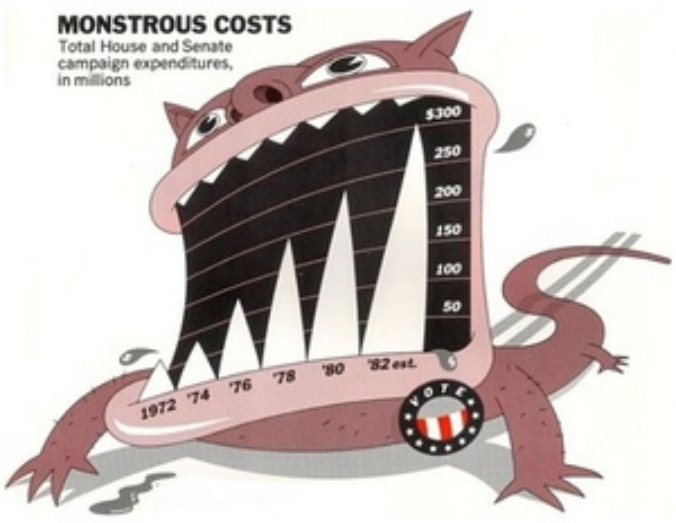
Columbia University, with $350 million in 2001, was just behind Stanford. Indiana University was the most-funded public university, in seventh place with $301 million.

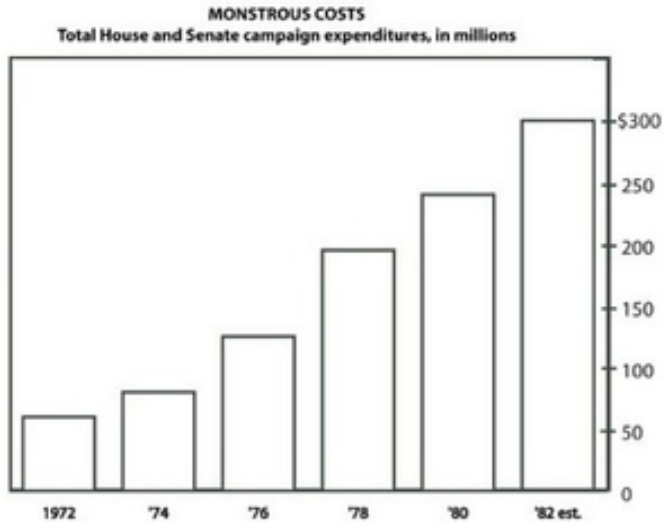
AARON SHARE/The Stanford Daily

Source: Stanford Daily, April 4, 2002.

Ugly Data Visualization

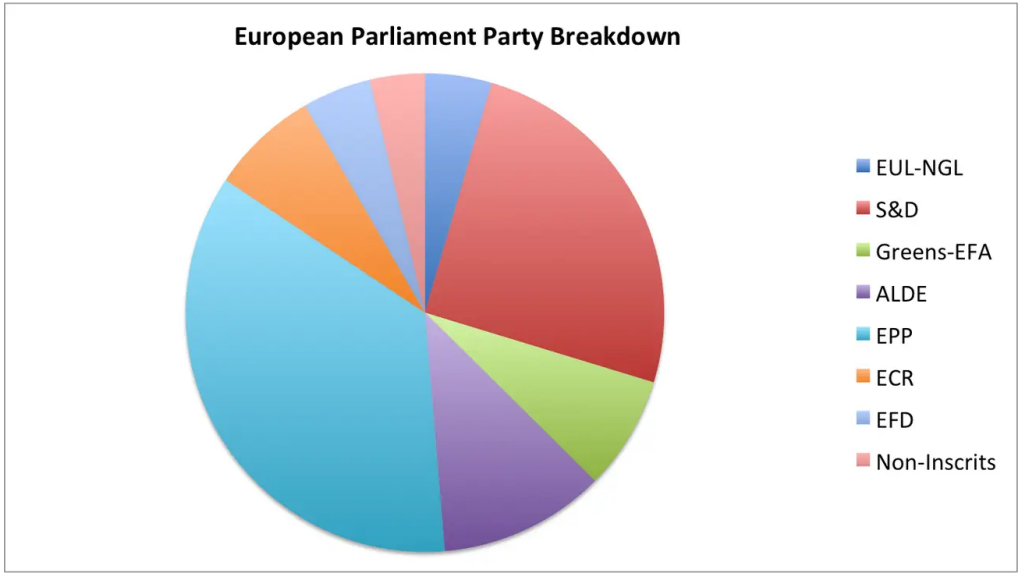
- It can be hard to resist "thematically" coordinating the "art" with the visualization. . . but please don't do so
- The chart on the left is mostly "chartjunk"





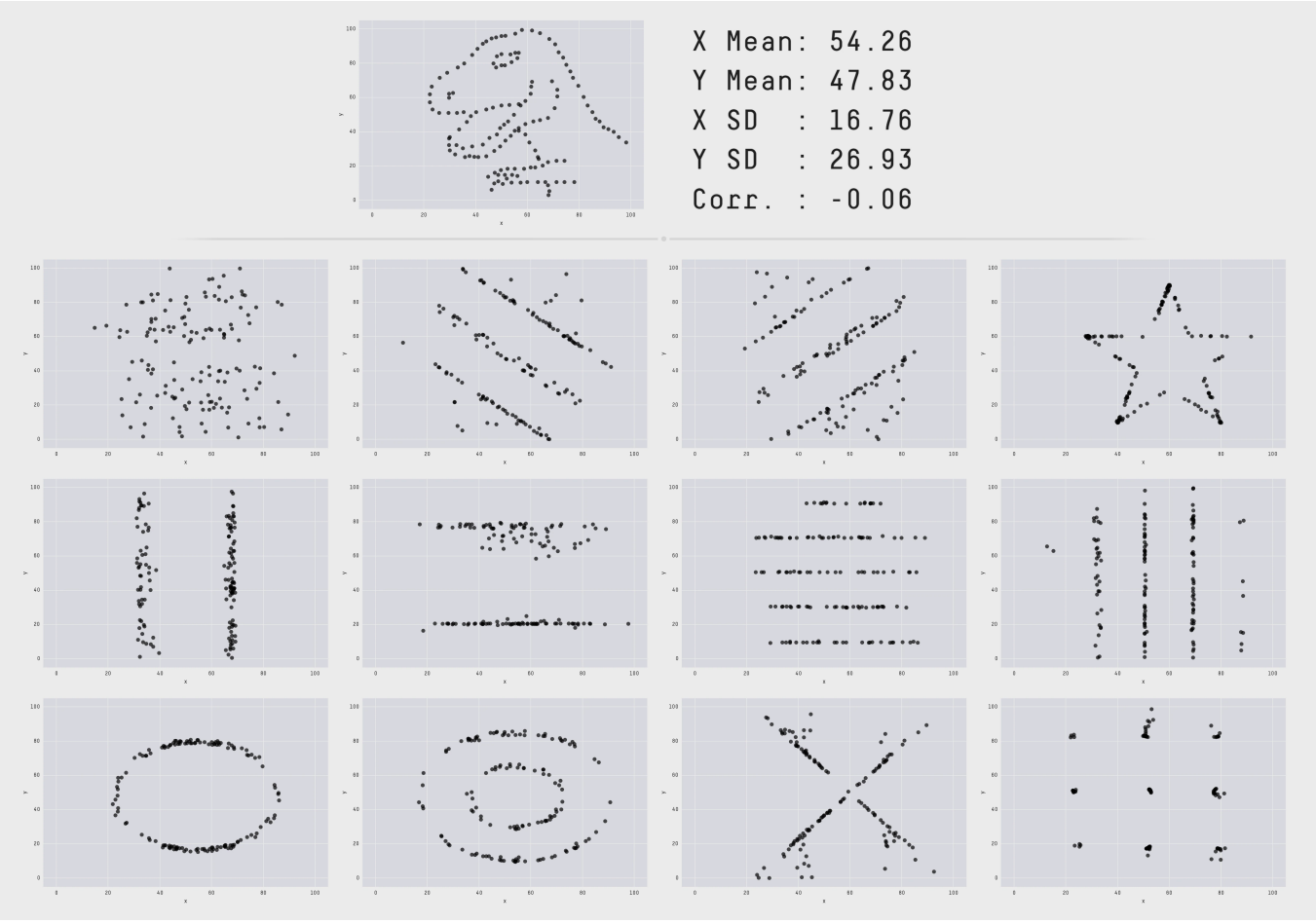
source

Ugly Data Visualization



source

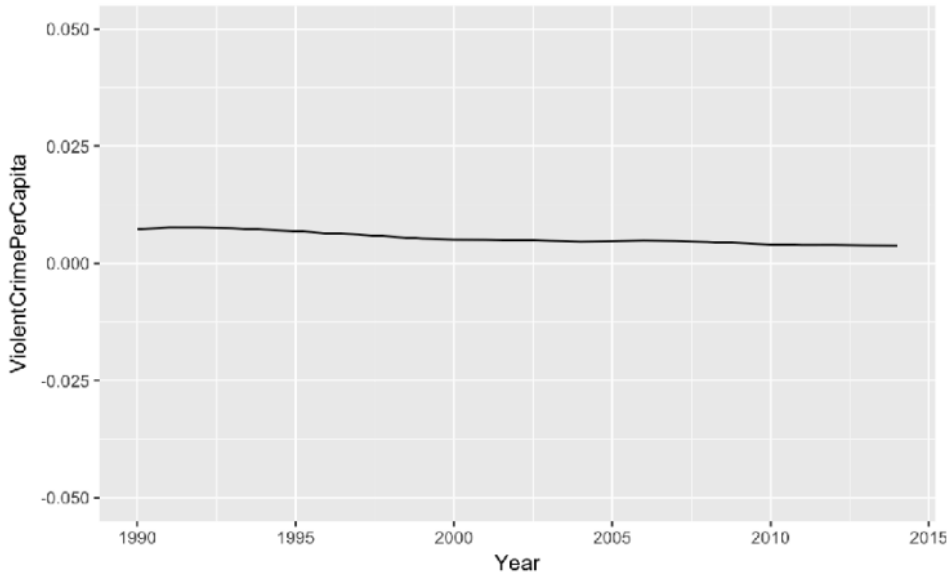
Data Visualizations



Datasaurus Dozen Source: While different in appearance, each dataset has the same summary statistics (mean, standard deviation, and Pearson's correlation) to two decimal places.

Data Visualizations

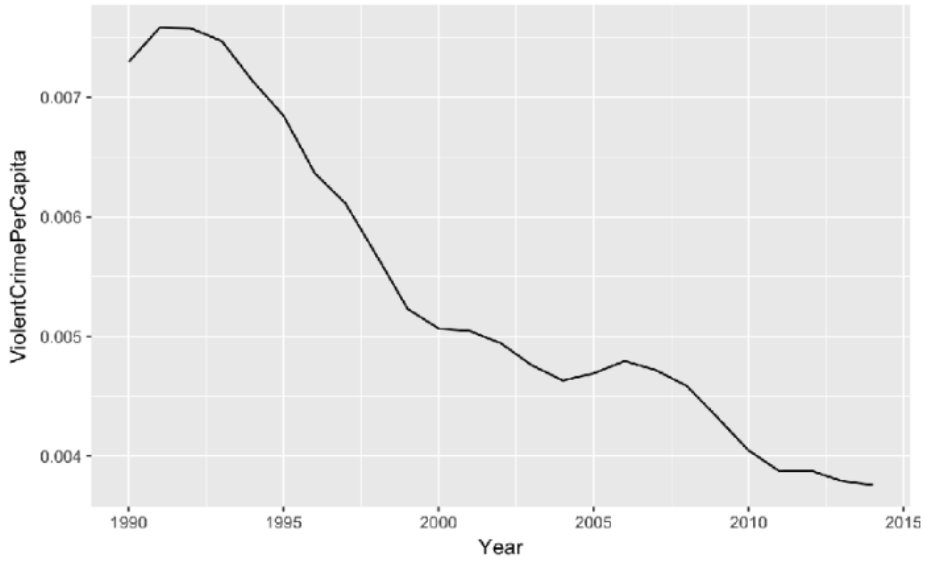
Violent crime was flat from 1990-2014.



source

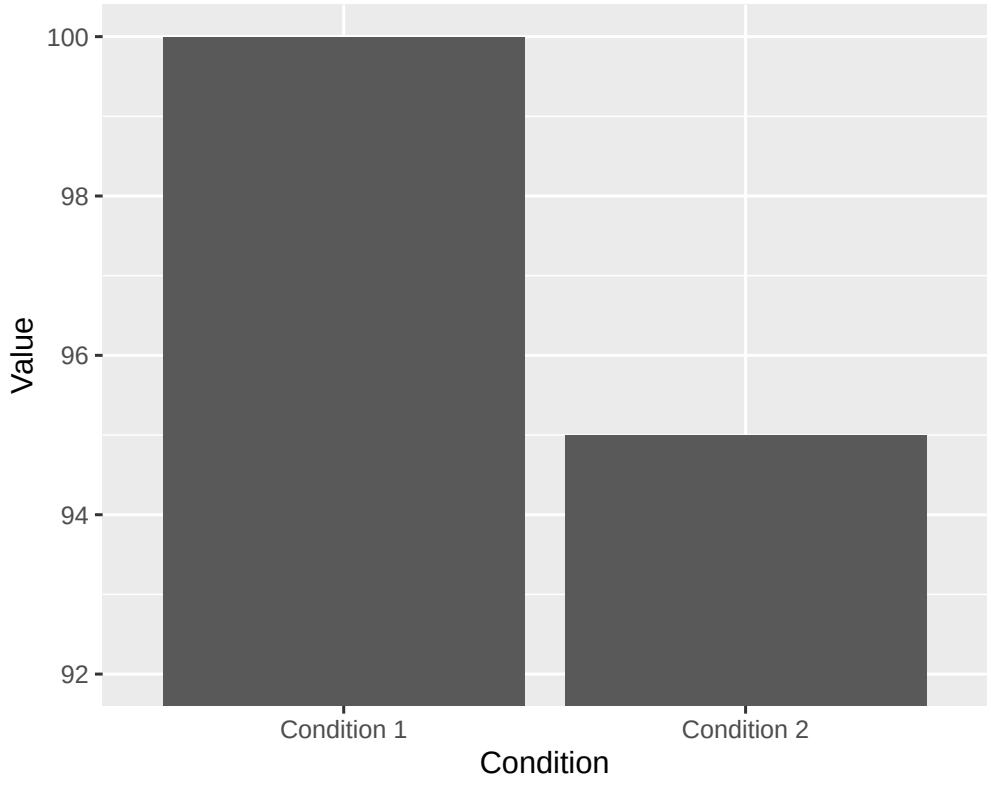
Data Visualizations

Wait . . . violent crime has plummeted since 1990!

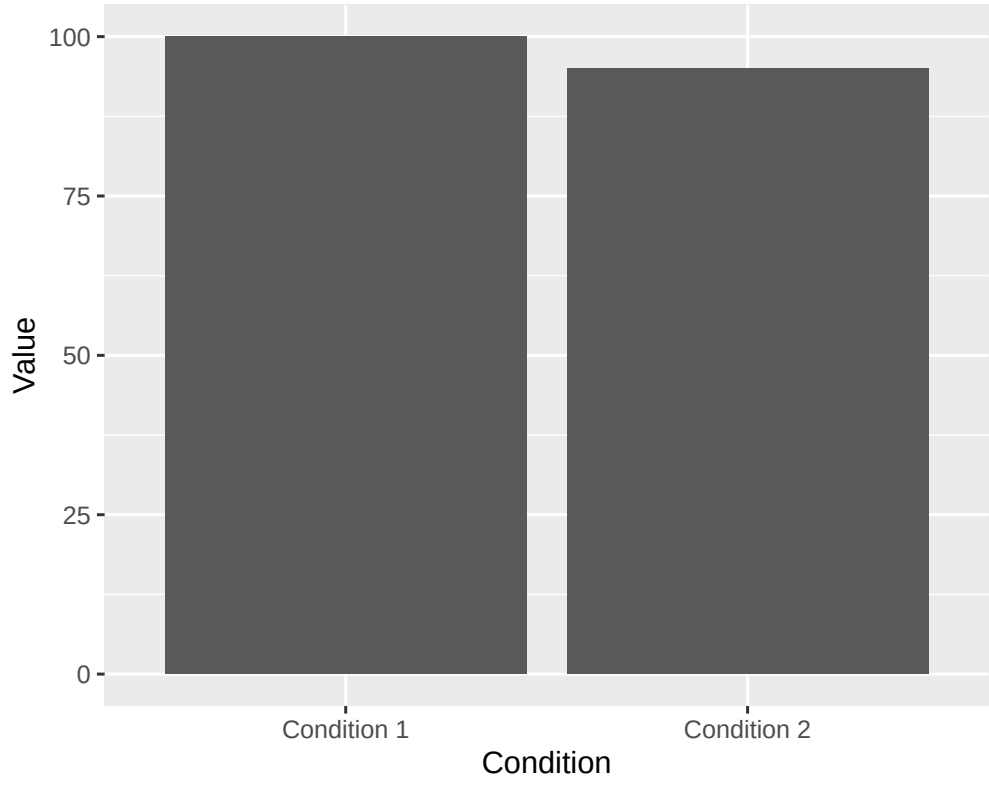


source

Data Visualizations



Data Visualizations



How to make good visualizations?

1. Focus on the data. Try to avoid clutter and chartjunk.
2. Try to avoid distorting the data and use approporiate scales.
3. When designing visualizations, try to keep human limitations in mind.
4. Focus on highlighting the underlying message of the data. Use clear and informative captions and labels. Don't misrepresent the data through graphics.