A Probability and Statistics Primer for Quantitative Finance

Week 3: Exploratory Data Analysis I

Jake Price
Instructor, Computational Finance and Risk Management
University of Washington
Slides originally produced by Kjell Konis

Outline

Outline

Population, Sample

Population	entire universe of possibilities to base our statistical study on
Sample	a subset of the population

Data: measurements on a sample of *individuals* from a population of interest

Population	Technology stocks
Sample	Technology stocks headquartered in Silicon Valley
Individual	One particular technology firm (e.g., Google)
Datum	Annualized dividend rate in 2014

We want to study a characteristic possessed by each individual in the population

This characteristic is called a *statistical variable*

Types of Variables

A variable can be *quantitative* or *qualitative*

A *quantitative* variable can be discrete or continuous

Quantitative Discrete

- yearly number of insurance claims
- number of defaults in a CDO

Quantitative Continuous

- daily return on an asset
- risk-free interest rate

Qualitative Variables

A qualitative (categorial) variable can be *nominal* or *ordinal*

Qualitative Nominal

- Sector membership
- Asset class (e.g., fixed income, equities, . . .)

Qualitative Ordinal

- Size (e.g., small cap, mid cap, large cap)
- Credit Rating (e.g., S&P, Moody's)

Often, quantitative variables are converted into qualitative variables for descriptive reasons

Outline

Exploratory Analysis of Qualitative Variables

Example: Industry Sectors of S&P 500 Component Stocks

- S&P Dow Jones Indices: S&P 500 stock market index
- 502 stocks (traded on American exchanges) issued by 500 large-cap companies
- Roughly 75 percent of USA equity market, by capitalization

Download (meta) data directly into R (source: Wikipedia 2015-03-14)

```
> SP500info <- read.csv(url(paste("http://staff.washington.edu/jrp14",
+                                 "/wordpress/wp-content/uploads",
+                                 "/2017/12/SP500.csv",sep="")))
> class(SP500info)
[1] "data.frame"
```

SP500info Data Frame

```
> head(SP500info, 5)
  Ticker             Security                 Sector
1    MMM           3M Company            Industrials
2    ABT  Abbott Laboratories            Health Care
3   ABBV          AbbVie Inc.            Health Care
4    ACN        Accenture plc Information Technology
5   ATVI  Activision Blizzard Information Technology
                         Subindustry     CIK
1         Industrial Conglomerates   66740
2            Health Care Equipment    1800
3                  Pharmaceuticals 1551152
4 IT Consulting & Other Services 1467373
5     Home Entertainment Software  718877
```

Categorical (or Factor) Data

```
> names(SP500info)
[1] "Ticker"      "Security"    "Sector"      "Subindustry"
[5] "CIK"
> sapply(SP500info, class)
     Ticker    Security      Sector Subindustry         CIK
   "factor"    "factor"    "factor"    "factor"   "integer"
> Sectors <- SP500info$Sector
> levels(Sectors)
 [1] "Consumer Discretionary"     "Consumer Staples"
 [3] "Energy"                     "Financials"
 [5] "Health Care"                "Industrials"
 [7] "Information Technology"     "Materials"
 [9] "Real Estate"                "Telecommunication Services"
[11] "Utilities"
```

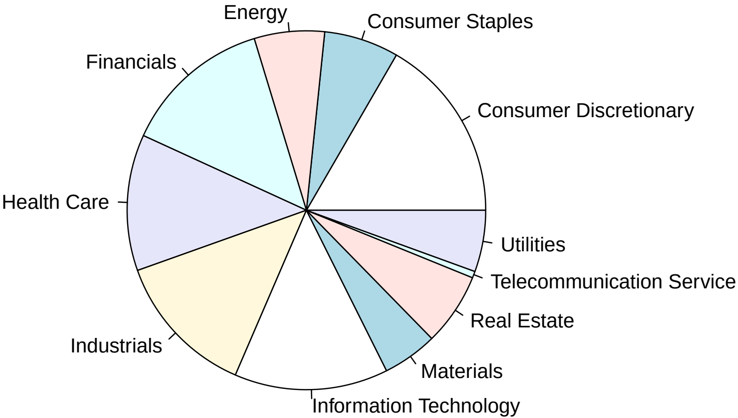
Frequency Tables

```
> Sector.freqs <- table(Sectors)
> Sector.freqs
```

	Frequency
Consumer Discretionary	84
Consumer Staples	34
Energy	32
Financials	68
Health Care	62
Industrials	66
Information Technology	70
Materials	25
Real Estate	33
Telecommunication Services	3
Utilities	28

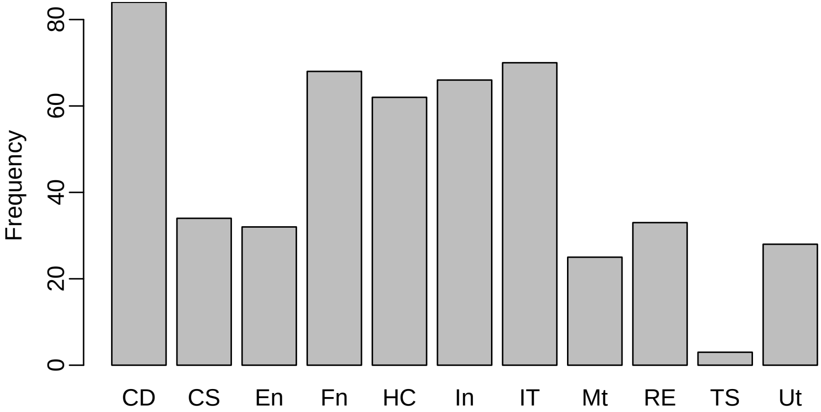
Pie Chart

```
> pie(Sector.freqs)
```



Bar Plot

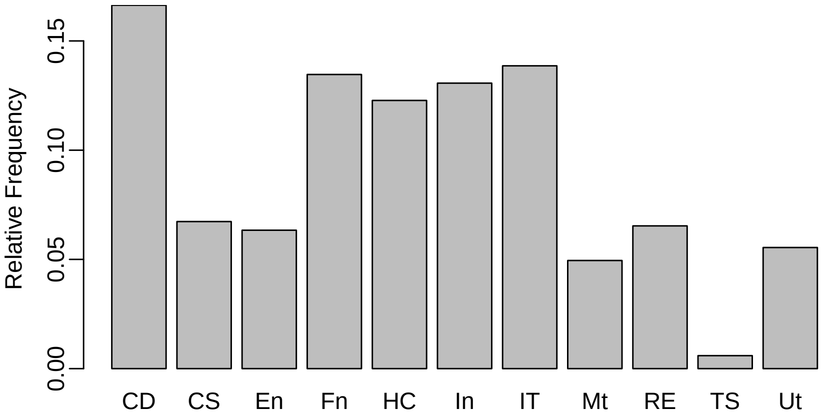
```
> bar.names <- abbreviate(names(Sector.freqs), minlength = 2)
> barplot(Sector.freqs, ylab = "Frequency", names = bar.names)
```



Bar Plot

Better to plot relative frequencies

```
> barplot(Sector.freqs / sum(Sector.freqs), ylab = "Relative Frequency",
+         names = bar.names)
```



Outline

Graphical Analysis of Quantitative Variables: Histograms

Example: 2012 Citigroup closing prices

```
> library(quantmod)

> getSymbols("C", from = "2012-01-01", to = "2012-12-31")

[1] "C"

> head(C)

           C.Open C.High C.Low C.Close C.Volume C.Adjusted
2012-01-03  27.13  28.51 27.13   28.33 58169500   27.50680
2012-01-04  28.04  28.38 27.62   28.17 41455000   27.35145
2012-01-05  27.66  29.18 27.47   28.51 66793300   27.68157
2012-01-06  28.66  29.06 28.01   28.55 48226900   27.72041
2012-01-09  28.72  29.38 28.65   29.08 35017900   28.23501
2012-01-10  29.75  30.14 29.66   30.00 47710900   29.12828
```

Graphical Analysis of Quantitative Variables: Histograms

```
> citi <- as.numeric(Cl(C["2012-5"]))
```

```
> citi
```

```
 [1] 33.60 32.70 32.48 31.60 31.67 31.32 30.45 30.65 29.35 28.14 27.79
[12] 26.92 26.41 26.01 26.25 26.92 27.15 26.66 26.47 27.02 26.00 26.51
```

For the rest of the example, use entire year of data

```
> citi <- as.numeric(Cl(C))
```

Histogram

Goal: a graphical display of the data to help us *"feel what the data are like"*

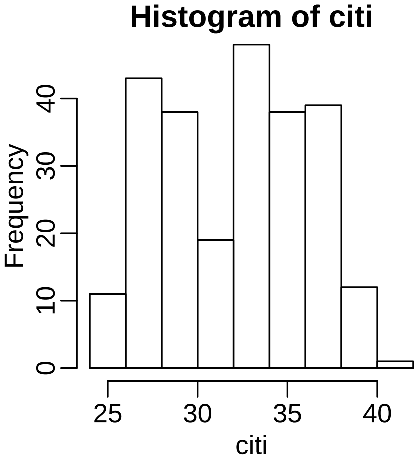
Group the data by splitting the range of the variable into k equal-length intervals

A *histogram* shows the number of observations in each group

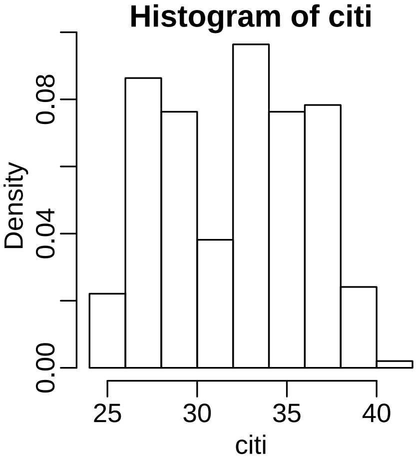
Bin	Center	Frequency	Relative Frequency
24 – 26	25	11	0.044
26 – 28	27	43	0.173
28 – 30	29	38	0.153
30 – 32	31	19	0.076
32 – 34	33	48	0.193
34 – 36	35	38	0.153
36 – 38	37	39	0.157
38 – 40	39	12	0.048
40 – 42	41	1	0.004

Histogram (continued)

```
> hist(citi)
```




```
> hist(citi, freq = FALSE)
```



Histogram (continued)

Advantages

Works well for small and large sample sizes

Disadvantages

The principal disadvantages are:

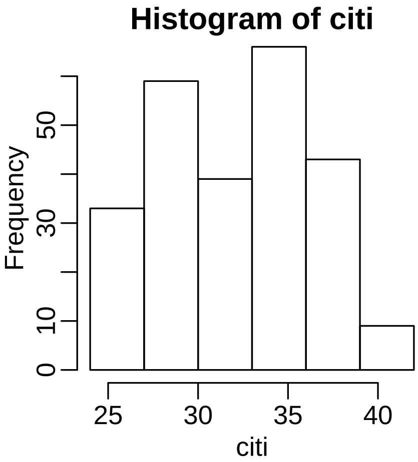
- Loss of information from binning
- Difficult to choose the number of intervals
- Interpretation difficult, not necessarily unique

Remarks

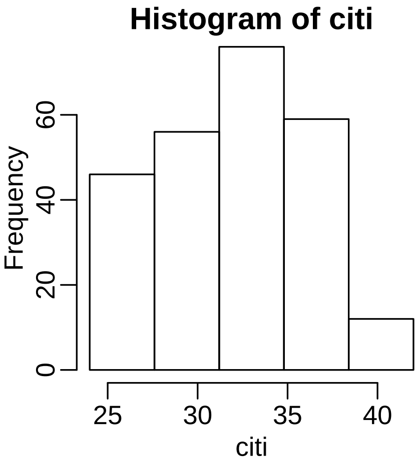
There are improvements on the histogram; for example, the kernel density estimator

Histogram: Sensitivity to Bin Choice

```
> hist(citi, seq(24, 42, len = 7))
```




```
> hist(citi, seq(24, 42, len = 6))
```



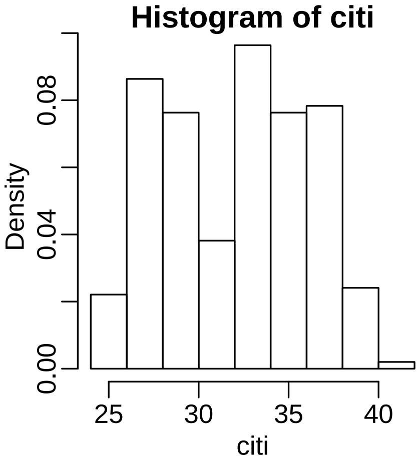
Outline

Kernel Density Estimation

Histogram: split range of data into k intervals of equal-length

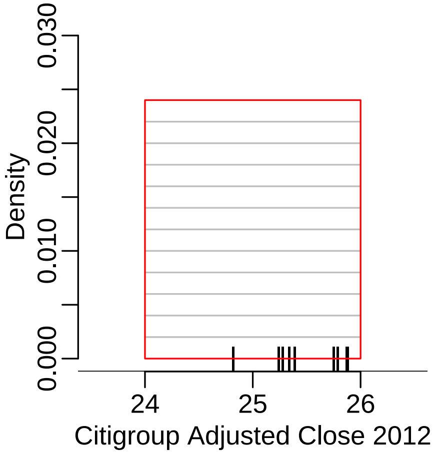
A *histogram* shows the number of observations in each interval

Bin	Center	Freq	Dens
24 - 26	25	12	0.024
26 - 28	27	42	0.084
⋮	⋮	⋮	⋮



Kernel Density Estimation

Zoom in on the first interval



For each data point in $[24, 26)$, stack a rectangle centered at 25

- Rectangle width $= 2$
- Rectangle height $= 1/2n$
- Total area $= 1$

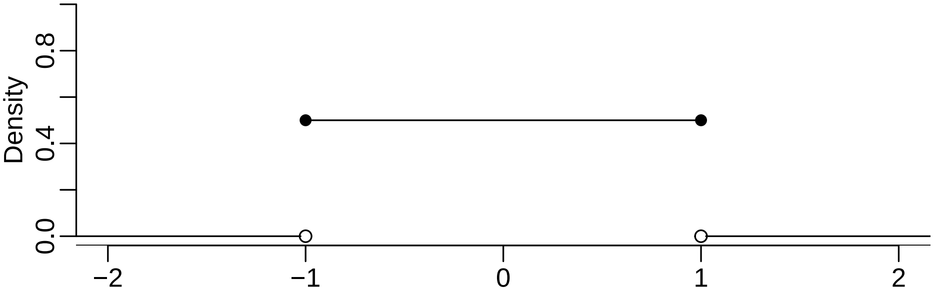
Kernel Density Estimation

- Center each rectangle on a data point instead

Kernel

Call the rectangle associated with each data point a *kernel*

$$K(x) = \begin{cases} \frac{1}{2} & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



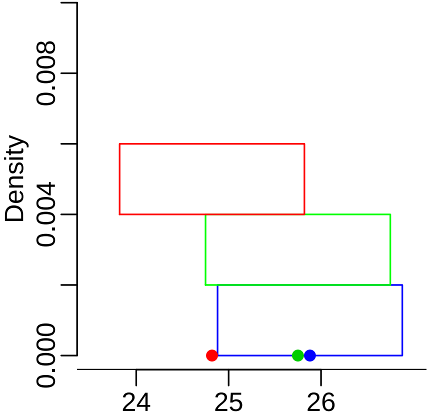
Kernel

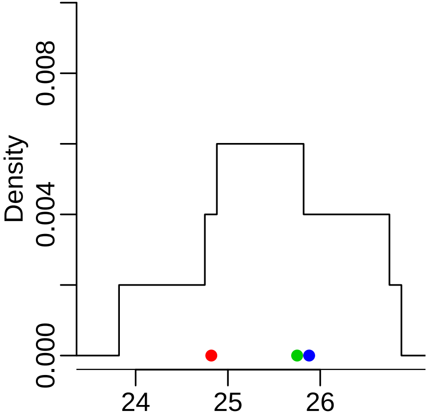
Rectangle with area 1, width $2b$, centered at x_i: $\quad \dfrac{1}{b} K \left(\dfrac{x - x_i}{b} \right)$

Kernel Density Estimator

The kernel density estimator for a set of data points $x_1, \ldots, x_n$ is

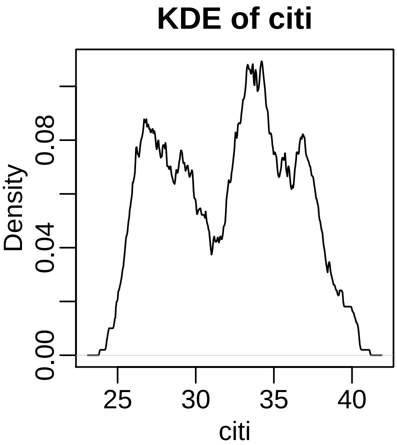
$$\hat{f}_b(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b} K\left(\frac{x - x_i}{b}\right)$$

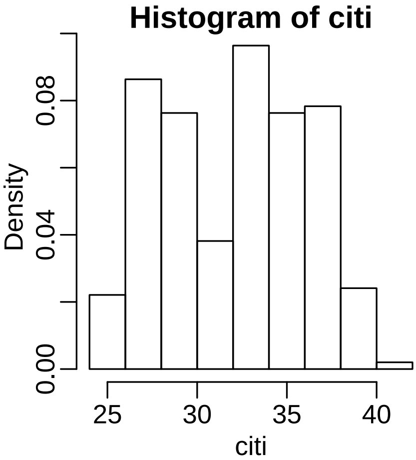




Kernel Density Estimation with R

```
> plot(density(citi, bw = 1, adjust = 1/sqrt(3), kernel = "rect"))
```





Outline

Smooth Density Estimation

Observation: kernel density estimate looks a bit ratty

Square corners of the rectangular kernel $\implies$ kde not smooth

Mathematically, the kernel K can be any function where

$$\int_{-\infty}^{\infty} K(x)\,dx = 1$$

Desirable properties:

- Nonnegative: $K(x) \geq 0 \qquad x \in \mathbb{R}$
- Symmetric: $K(x) = K(-x) \qquad x > 0$

In particular, the normal kernel satisfies these properties

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Quick Math Review

Continuous

- Conceptually, a function is continuous if small changes in the input result in small changes to the output
- Rigorously, a function g is continuous on an interval (a, b) if

$$\lim_{x \to c} g(x) = c \qquad \forall c \in (a, b)$$

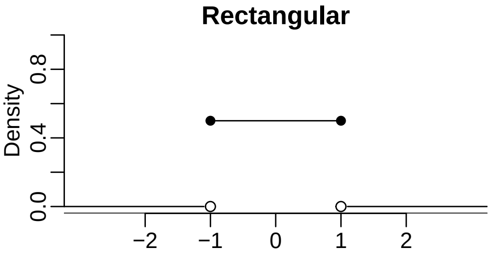
Smooth

- Conceptually, a function is smooth if it does not have any rough edges or corners
- A function g is *smooth* if its derivative g' is continuous
- Conceptually, a function is more smooth the more (continuous) higher order derivatives it has

Kernel Choice

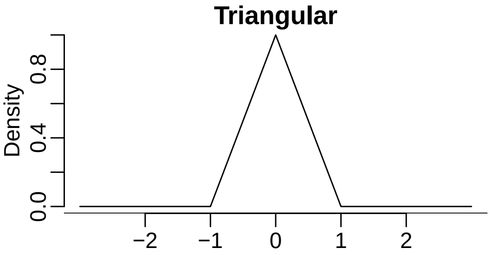
Rectangular

$$K(x) = \begin{cases} \frac{1}{2} & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



Triangular

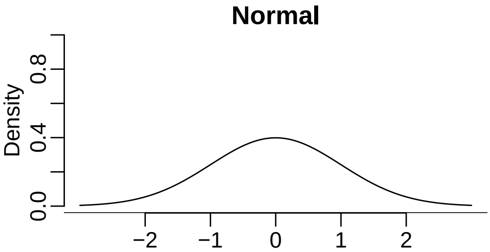
$$K(x) = \begin{cases} 1 - |x| & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



Kernel Choice

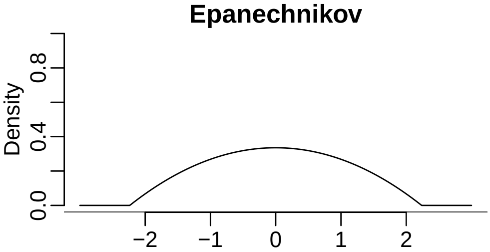
Normal

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

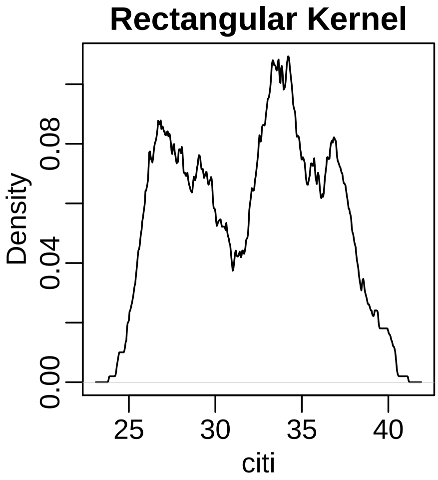


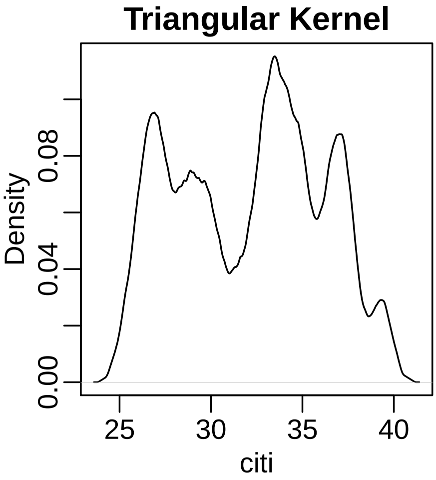
Epanechnikov

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}x^2) & |x| \leq \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

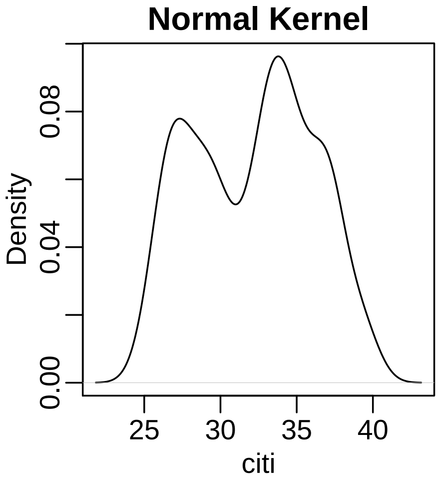


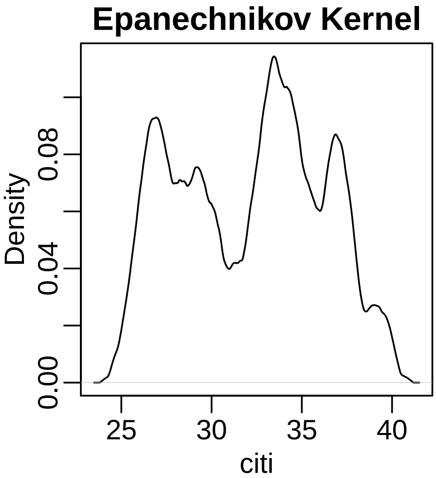
Kernel Density Estimates by Kernel





Kernel Density Estimates by Kernel

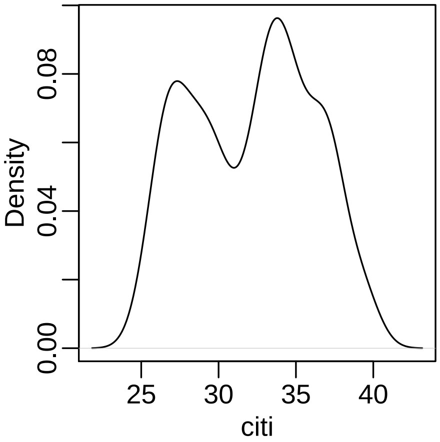




Outline

Bandwidth Selection

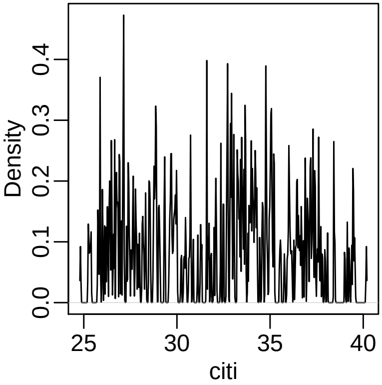
```
> plot(density(citi, bw = 1))
```



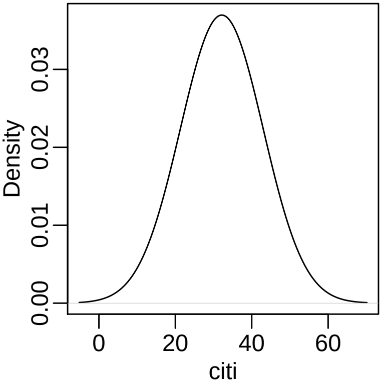
Bandwidth Selection

What happens when the bandwidth is too small or too big?

```
> plot(density(citi, bw = 0.01))
```

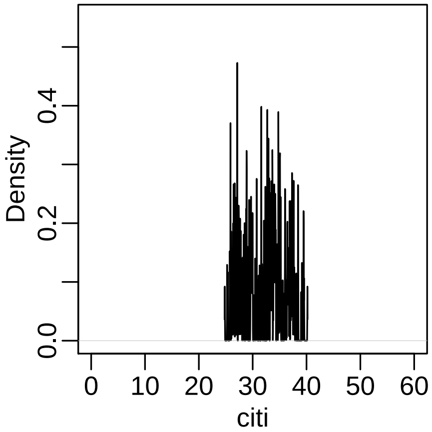


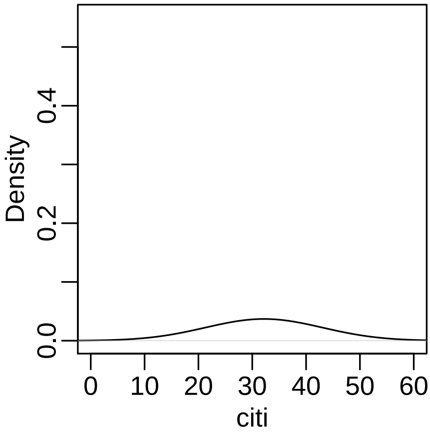

```
> plot(density(citi, bw = 10))
```



Bandwidth Selection

Difference is profound with equal scales





Smoothing

Bandwidth controls amount of smoothing

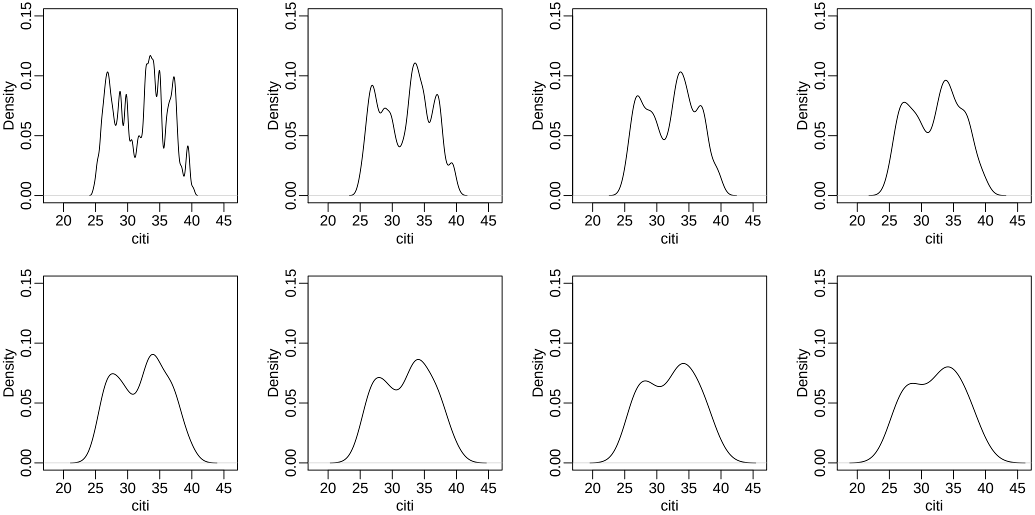
Too large a bandwidth results in *over smoothing*

Too small a bandwidth results in *under smoothing*

Have to choose the bandwidth so that the amount of smoothing is just right

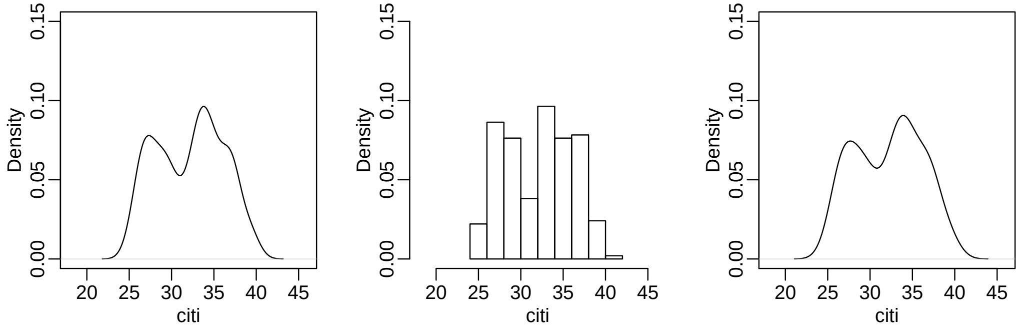
Exploratory Data Analysis $\implies$ you have to choose the amount of smoothing

Bandwidth Selection: $\frac{1}{4}$ to 2 in steps of $\frac{1}{4}$



Bandwidth Selection

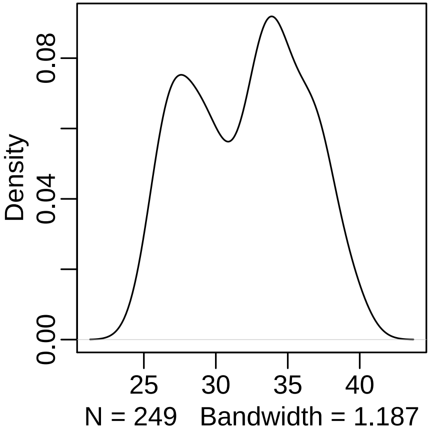
Can use a histogram to guide choice



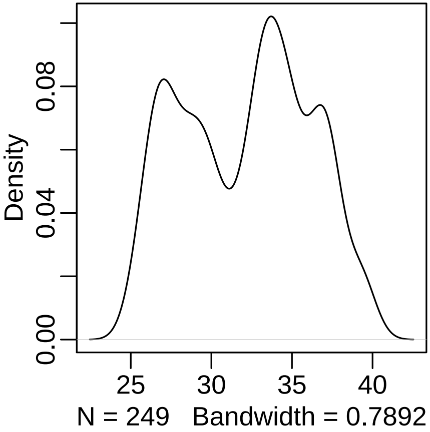
Suggests $1 < b < 1.25$ about the right amount of smoothing

Automatic Bandwidth Selection

```
> plot(density(citi))
```




```
> plot(density(citi, bw = "SJ"))
```



Summary

Histogram

- Simple, rough estimate of the distribution of the data
- Can be difficult to choose the number of bins (intervals)

Kernel Density Estimation

- Generalization of histogram
- Can get smooth estimate of distribution
- Have to choose bandwidth (also can choose kernel)

Summary

What does it mean to *"feel what the data are like?"*

Histograms and kernel density estimates reveal:

- appearances of separation into groups
- skewness: trailing off farther in on direction than the other
- the presence of unexpectedly common values
- the location (or center) of the data
- the spread of the data relative to the center

Outline

Order Statistics

The k^{th} *order statistic* of a data set is equal to its k^{th} smallest value

Example data set with 5 points

$$x_1 = 28.33,\ x_2 = 28.17, x_3 = 28.51,\ x_4 = 28.55,\ x_5 = 29.08$$

The order statistics are denoted by

$$x_{(1)} = 28.17,\ x_{(2)} = 28.33,\ x_{(3)} = 28.51,\ x_{(4)} = 28.55,\ x_{(5)} = 29.08$$

In particular

$$\begin{aligned}
\min(x_1, \ldots, x_n) &= x_{(1)} \\
\max(x_1, \ldots, x_n) &= x_{(n)} \\
\text{range}(x_1, \ldots, x_n) &= x_{(n)} - x_{(1)}
\end{aligned}$$

Sample Quantiles

For a data set $\mathbf{x} = \{x_1, \ldots, x_n\}$, the *sample quantile* of order $\alpha \in (0, 1)$ is

$$\hat{q}_\alpha(\mathbf{x}) = x_{(\lceil n\alpha \rceil)}$$

Often used to split the data into equal-sized parts, two common examples

- *quartiles*

$$\underbrace{\hat{q}_{0.25}(\mathbf{x})}_{\text{lower quartile}} \qquad \underbrace{\hat{q}_{0.50}(\mathbf{x})}_{\text{median}} \qquad \underbrace{\hat{q}_{0.75}(\mathbf{x})}_{\text{upper quartile}}$$

- *percentiles*

$$\underbrace{\hat{q}_{0.01}(\mathbf{x})}_{\text{first percentile}} \qquad \ldots$$

Example: Calculation of Quantiles

Calculate the 0.29 quantile of the following sample

$$\mathbf{x} = \{28.33, 28.17, 28.51, 28.55, 29.08, 30.00, 31.27, 31.60\}$$

In this case

$$\lceil n\alpha \rceil = \lceil 8 \times 0.29 \rceil = \lceil 2.32 \rceil = 3 \quad \Longrightarrow \quad \hat{q}(0.29) = x_{(3)} = 28.51$$

Can use quantile function in R:

```
> quantile(x, type = 1)
   0%   25%   50%   75%  100%
28.17 28.33 28.55 30.00 31.60

> quantile(x, probs = 0.29, type = 1)
  29%
28.51
```

Outline

Empirical Cumulative Distribution Function

The *empirical cumulative distribution function* for a data set $\mathbf{x} = \{x_1, \ldots, x_n\}$ is

$$\hat{F}(x) = \frac{\text{number of data points less than or equal to } x}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \leq x\}}$$

where $I_{\{a\}}$ indicates of the occurrence of a

The empirical cumulative distribution function (*ecdf*) is a right-continuous step function that increases by $\frac{1}{n}$ at each data point x_i

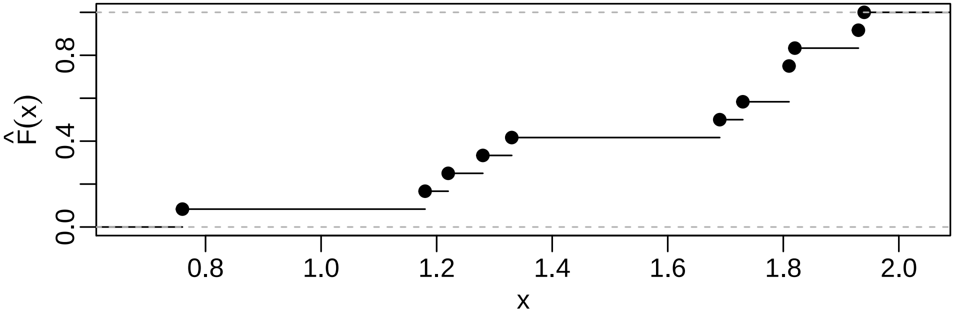
The ecdf plots the fraction of the points in the data set $\mathbf{x}$ that are less than or equal to x as a function of x

Example: Empirical Cumulative Distribution Function

Consider the following data set of 12 points and their order statistics

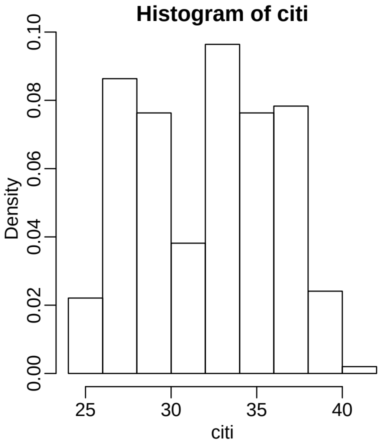
x_i	1.94	1.93	1.33	0.76	1.69	1.82	1.81	1.22	1.81	1.28	1.18	1.73
$x_{(i)}$	0.76	1.18	1.22	1.28	1.33	1.69	1.73	1.81	1.81	1.82	1.93	1.94

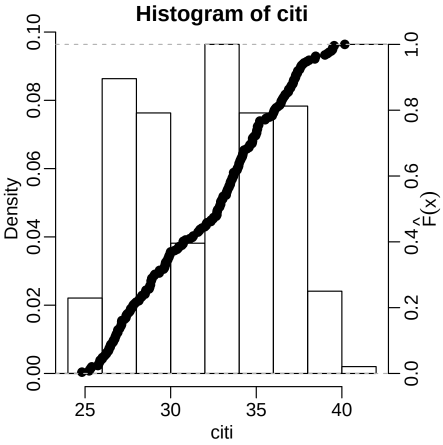
```
> plot(ecdf(x))
```



ECDF Relationship to Density Estimates

```
> hist(citi, freq = FALSE)
> plot(ecdf(citi), add = TRUE)
```





Outline

Five-Number Summary

Given a data set $\mathbf{x}$, the following five numbers

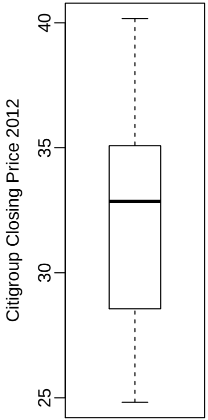
$$\min(\mathbf{x}) = x_{(1)}, \quad \hat{q}_{0.25}(\mathbf{x}), \quad \hat{q}_{0.5}(\mathbf{x}), \quad \hat{q}_{0.75}(\mathbf{x}), \quad \max(\mathbf{x}) = x_{(n)}$$

comprise the *five-number summary* of $\mathbf{x}$

- Provides a simple, practical numeric summary of a distribution
- Provides the basis for a box plot

Concepts

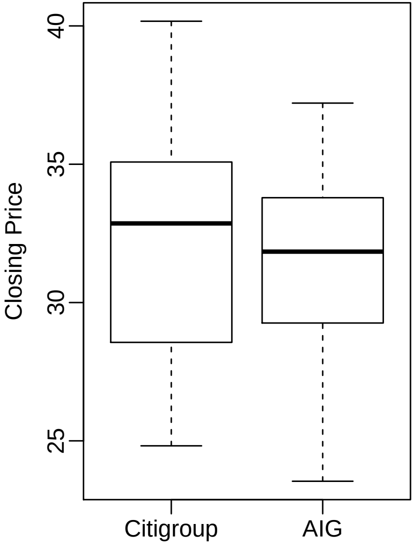
- the *center* or *location* of the data
- the *bulk* of the data
- the *tails*



boxplot of $\mathbf{x}$

Boxplot: Example

- Boxplots also useful for comparing grouped data
- For example, 2012 closing prices for Citigroup and AIG



Boxplot: Construction

Five-number summary for 2012 Citigroup Closing Prices

$$24.79,\quad 28.54,\quad 32.82,\quad 35.11,\quad 40.15$$

The 0.5-quantile provides the “center” of the boxplot

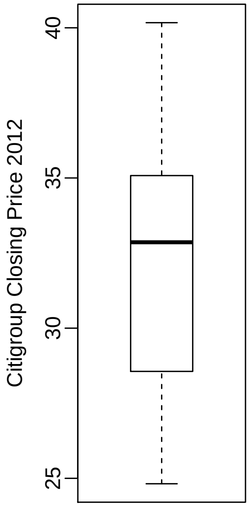
The box shows the middle half of the data

$$\begin{aligned} d &= (\hat{q}_{0.75} - \hat{q}_{0.25}) = 6.57 \\ \hat{q}(0.25) - 1.5d &= 28.54 - 9.855 = 18.685 \\ \hat{q}(0.75) + 1.5d &= 35.11 + 9.855 = 44.965 \end{aligned}$$

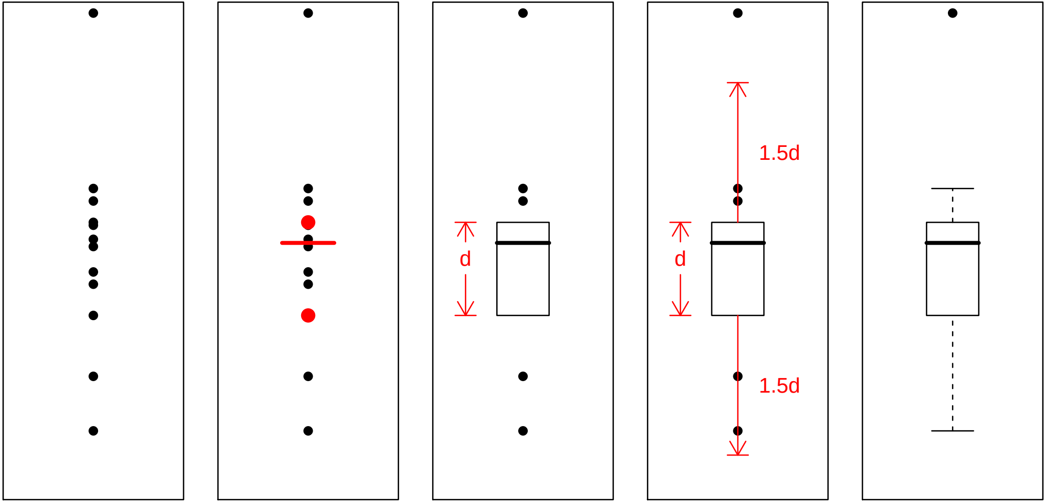
Whiskers limits are the most extreme observations in the interval

$$[\hat{q}_{0.25} - 1.5d, \hat{q}_{0.75} + 1.5d]$$

Data beyond the whiskers are plotted individually



Example: Boxplot Construction



Outline

Plotting Best Practices

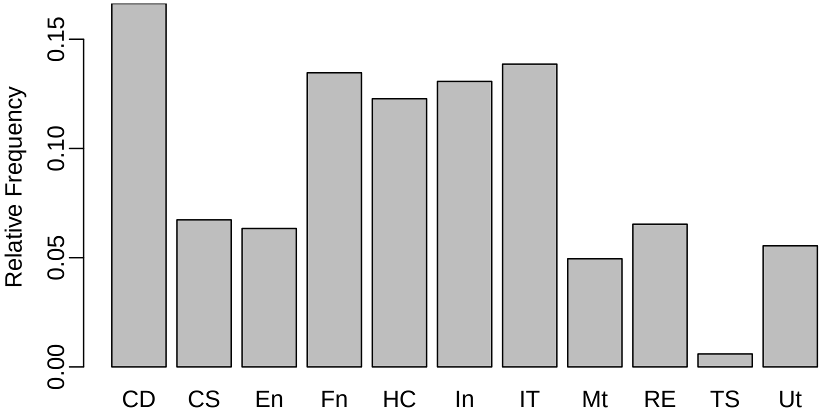
It is not easy to create good graphical displays of data (plots)

Default plots provided by standard software (e.g., Excel, R, etc.) often not very good

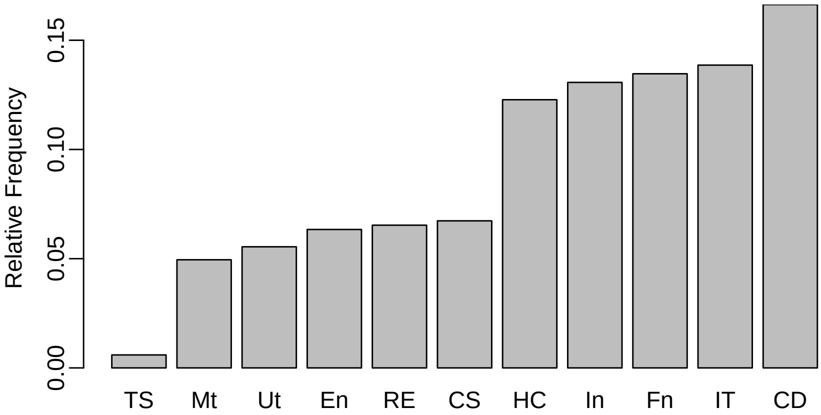
Some plotting best practices:

- Do not create the illusion of a trend when the x-axis is qualitative
- Put clear, concise labels on the axes, the legend, and the plot
- When comparing related quantities, use the same axes and intelligent positioning
- Changing the aspect ratio can reveal interesting features
- Design a plot so that departures from the expected appear as departures from linearity or distance from a *cloud* of data
- Try to show only the data — No **chart-junk**

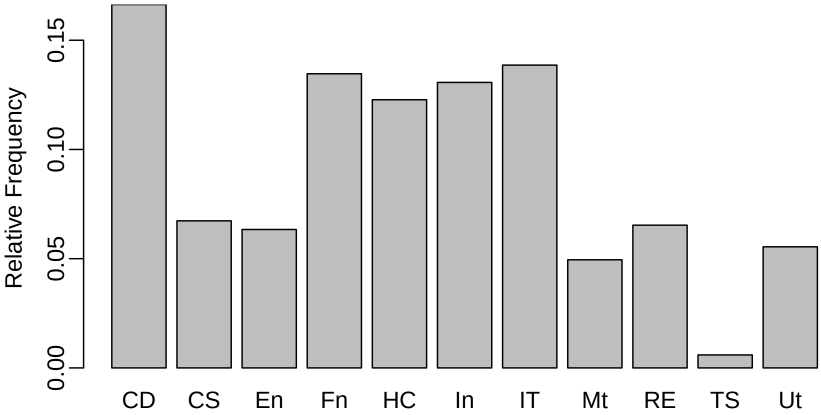
A Qualitative *x*-axis



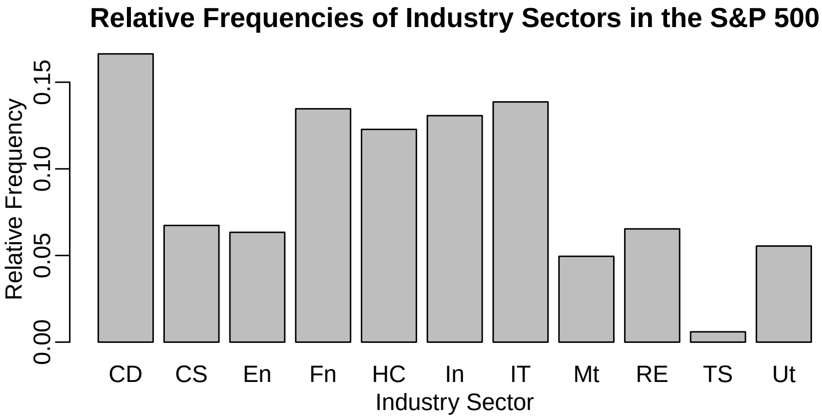
A Qualitative x-axis with the Illusion of a Trend



Clear, Concise Labels

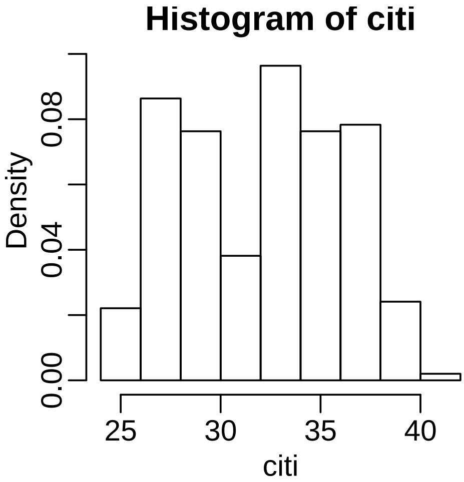


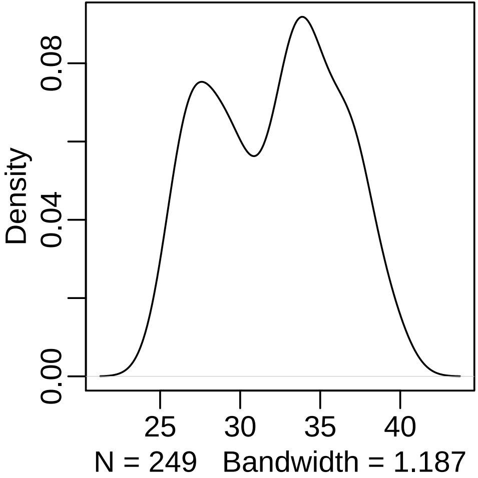
Clear, Concise Labels



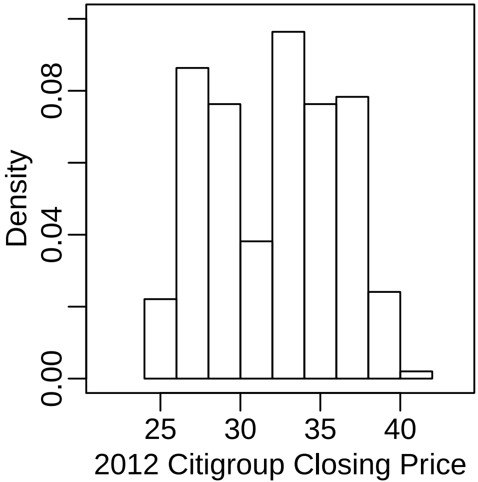
CD	Consumer Discretionary	HC	Health Care	Tel	Telecommunication Services
CS	Consumer Staples	Ind	Industrials		
E	Energy	IT	Information Technology	U	Utilities
Fin	Financials	Mat	Materials		

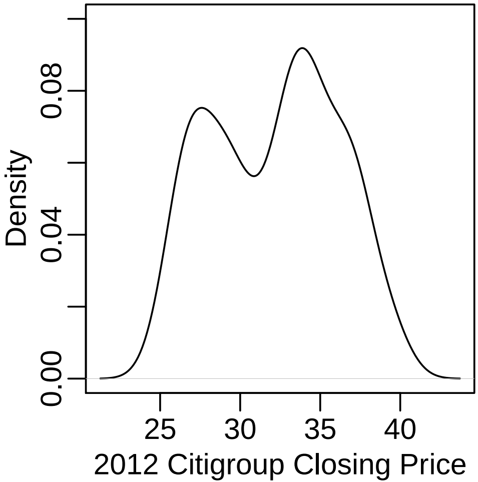
Same Axes, Intelligent Positioning



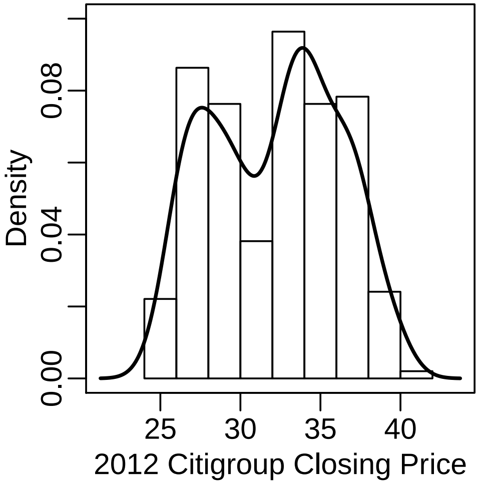


Same Axes, Intelligent Positioning





Another Possibility: **Same** Axes



Changing the Aspect Ratio Can Reveal Interesting Features

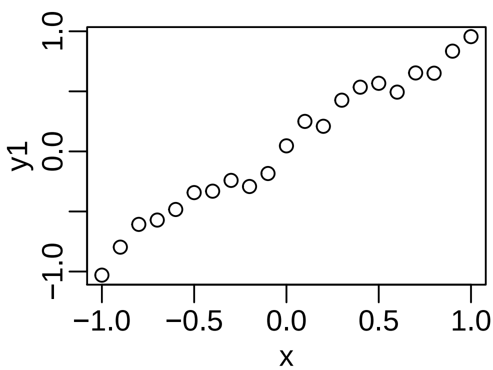
A toy example

```
> x <- seq(-1, 1, by = 0.1)
> y1 <- 0 + 1.0*x + rnorm(x, sd = 0.1)
> y2 <- 0 + 0.1*x + rnorm(x, sd = 0.01)
```

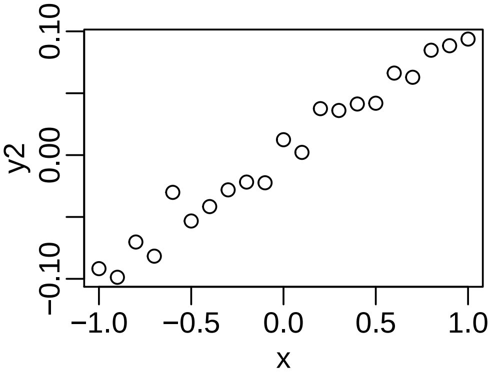
What happens when we plot `y1` vs. `x` and `y2` vs. `x`?

Changing the Aspect Ratio Can Reveal Interesting Features

```
> plot(x, y1)
```

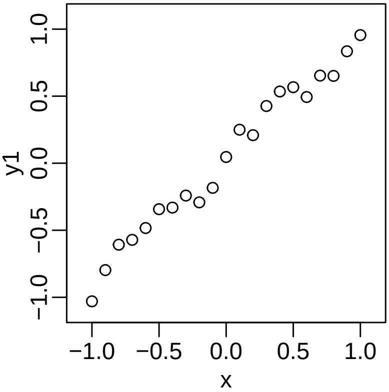



```
> plot(x, y2)
```

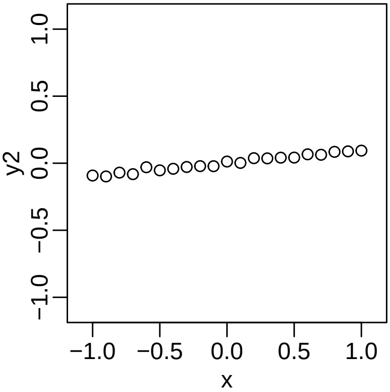


Changing the Aspect Ratio Can Reveal Interesting Features

```
> plot(x, y1, xlim = c(-1.1, 1.1),
+        ylim = c(-1.1, 1.1))
```




```
> plot(x, y2, xlim = c(-1.1, 1.1),
+        ylim = c(-1.1, 1.1))
```



Outline

Chart-Junk

E.g., plot of 5 numbers

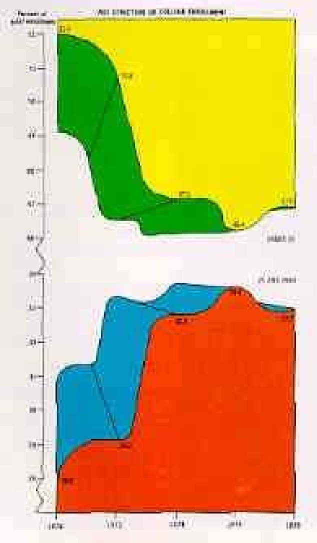


Chart-Junk and Scale

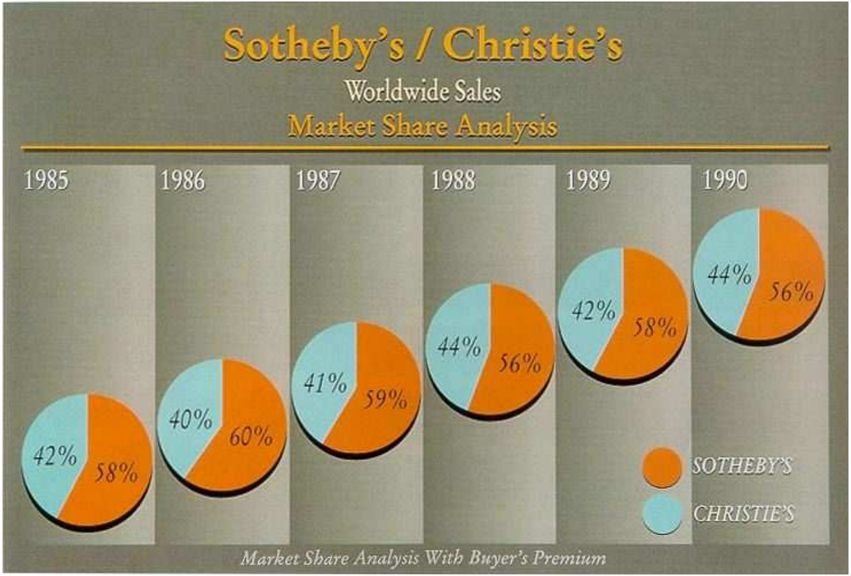
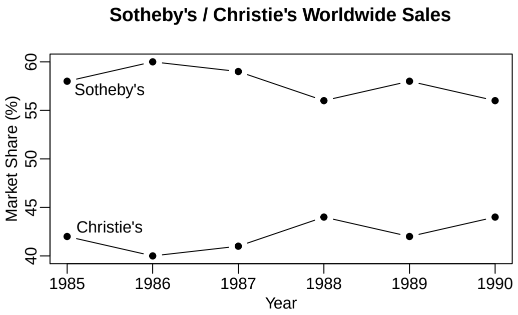
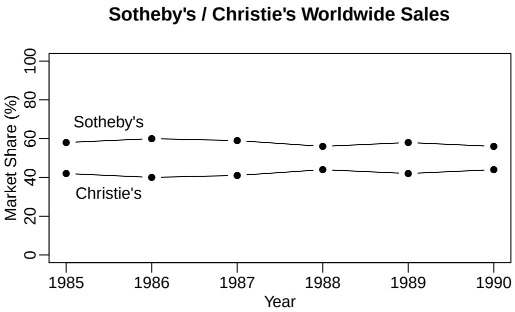


Chart-Junk and **Scale**



Sotheby's is crushing Christie's

Displaying the data on a *more honest* scale tells a different story



References and Further Reading

1. *The Visual Display of Quantitative Information* (2nd Edition). E. Tufte. Graphics Press, 2001.

2. *Envisioning Information*. E. Tufte. Graphics Press, 1990.

3. CSE 512: `http://courses.cs.washington.edu/courses/cse512/`

COMPUTATIONAL FINANCE & RISK MANAGEMENT
UNIVERSITY *of* WASHINGTON
Department of Applied Mathematics

http://computational-finance.uw.edu