Chapter 3 – Data Visualization
Graphs for Data Exploration

Basic Plots
  Line Graphs
  Bar Charts
  Scatterplots

Distribution Plots
  Boxplots
  Histograms
Line Graph for Time Series
95% of tracts do not border Charles River

Excel can confuse:
y-axis is actually “% of records that have a value for CATMEDV” (i.e., “% of all records”)
Scatterplot

Displays relationship between two numerical variables
Distribution Plots

- Display “how many” of each value occur in a data set

- Or, for continuous data or data with many possible values, “how many” values are in each of a series of ranges or “bins”
Histograms

Boston Housing example:

Histogram shows the distribution of the outcome variable (median house value)
Boxplots

Side-by-side boxplots are useful for comparing subgroups

Boston Housing Example: Display distribution of outcome variable (MEDV) for neighborhoods on Charles river (1) and not on Charles river (0)
Box Plot

- Top outliers defined as those above Q3+1.5(Q3-Q1).
- “max” = maximum of non-outliers
- Analogous definitions for bottom outliers and for “min”
- Details may differ across software
Heat Maps

Color conveys information

In data mining, used to visualize
   Correlations
   Missing Data
Heatmap to highlight correlations (Boston Housing)

In Excel (using conditional formatting)

In Spotfire
Multidimensional Visualization
Scatterplot with color added

Boston Housing

NOX vs. LSTAT

**Red** = low median value

**Blue** = high median value
Matrix Plot

Shows scatterplots for variable pairs

Example: scatterplots for 3 Boston Housing variables
Rescaling to log scale (on right) “uncrowds” the data
Amtrak Ridership – Monthly Data
Aggregation – Yearly Average

The graph shows the yearly average (Avg(Ridership)) of ridership from 1991 to 2004. The ridership data fluctuates over the years, with a notable peak around 2003 and a subsequent decline in 2004.
Scaling: Smaller markers, jittering, color contrast
(Universal Bank; red = accept loan)
Jittering

- Moving markers by a small random amount
- Uncrowds the data by allowing more markers to be seen
Without jittering (for comparison)
Parallel Coordinate Plot (Boston Housing)

CATMEDV = 1

CATMEDV = 0
Linked plots
(same record is highlighted in each plot)
Network Graph – eBay Auctions
(sellers on left, buyers on right)

Circle size = # of transactions for the node

Line width = # of auctions for the buyer-seller pair

Arrows point from buyer to seller
Treemap – eBay Auctions
(Hierarchical eBay data: Category > sub-category > Brand)

Rectangle size = average closing price (= item value)

Color = % sellers with negative feedback (darker = more)
Map Chart
(Comparing countries’ well-being with GDP)

Well-Being Score

Darker = higher value

GDP