Introduction to Data Analysis

KRAKATAU posco

KRAKATAU posco

Learning Outcomes

- Understand key elements in exploratory data analysis (EDA)
- Explain and use common summary statistics for EDA
- Plot and explain common graphs and charts for EDA

note:

- We’re focus on WHAT’s the result, not HOW it works

KRAKATAU posco

Data Analytics

- What is Data Analytics?
  - Study data to answer questions
- Why?
  - More organizations are collecting more data than ever before
    - Business, Government, Healthcare, Charity – everything!
  - This data holds many insights into their operations and beyond
  - Data Analytics is required to extract these insights
    - Requires analytical, statistical and computational skills

KRAKATAU posco

To Answer Key Questions

- What movies (or books) customers would like to watch (or read)?
- What movies to order from studio and how many?
- Who are our best customers?
- When is there a flu epidemic in region in the country?
- Which customers are most likely not to have an accident?
- When a customer is likely to jump ship & go to a competitor?
- What is the one item you want to have in your store in case of a hurricane?
- What is the one thing that will improve a lawyer’s chance to win a case?
- What are some questions one can answer with a loyalty card?
- What is the number one reason for the success of cricket player?
- … …

KRAKATAU posco

Data Analytic Tasks

- **Exploratory Data Analysis (EDA) our focus this week!**
  - A preliminary exploration of the data to better understand its characteristics
- Descriptive Modelling
  - Find human-interpretable patterns that describe the data. E.g.
    - Association Rule Discovery
    - Clustering
- Predictive Modelling
  - Use some attributes to predict unknown or future values of other attributes. E.g.
    - Classification
    - Regression

KRAKATAU posco

EDA

- If you are going to find out anything about a data set you must first understand the data
  - but most interesting data sets are too big to inspect manually
- EDA:
  - Helps to select the right tool for preprocessing or analysis
  - Makes use of humans' abilities to recognize patterns
- Focuses on:
  - Summary statistics
  - Visualization
  - Clustering and anomaly detection

1. Summary Statistics

- Getting an overall sense of the data set
- Numbers that summarize properties of the data
  - Frequency
    - frequency, mode, percentiles
  - Location
    - Min & Max, Mean & Median
  - Spread
    - Variance & Standard Deviation, Distribution

WHO	1240 earthquakes known to have caused tsunamis for which we have data or good estimates
WHAT	Magnitude (Richter scale[2]), depth (m), date, location, and other variables
WHEN	From 2000 B.C.E. to the present
WHERE	All over the earth

For example, here are the *Magnitudes* (on the Richter scale) of the 1240 earthquakes in the NGDC data:

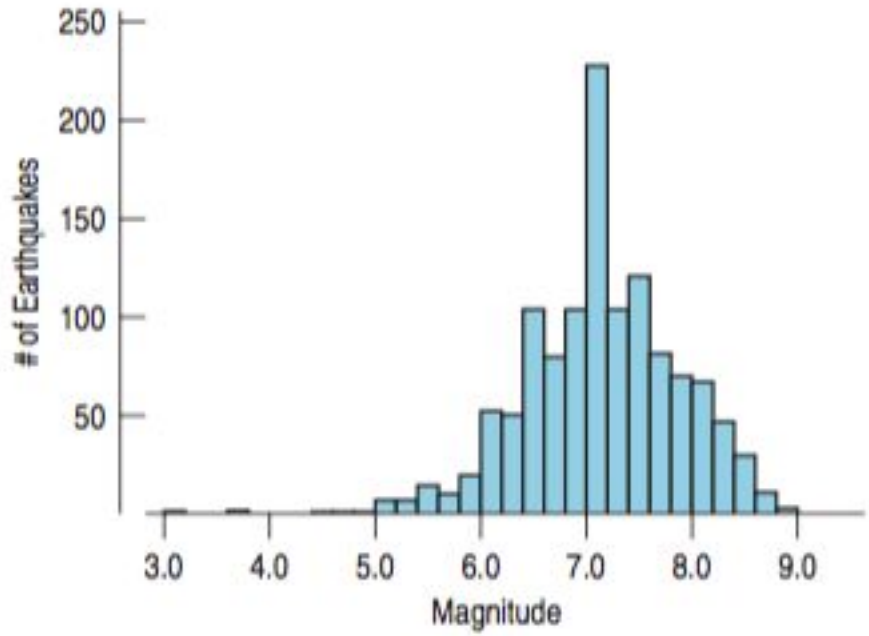


FIGURE 4.1
A histogram of earthquake magnitudes shows the number of earthquakes with magnitudes (in Richter scale units) in each bin.

histogram, the bins slice up *all the values* of the quantitative variable, so any spaces in a histogram are actual **gaps** in the data, indicating a region where there are no values.

Sometimes it is useful to make a **relative frequency histogram**, replacing the counts on the vertical axis with the *percentage* of the total number of cases falling in each bin. Of course, the shape of the histogram is exactly the same; only the vertical scale is different.

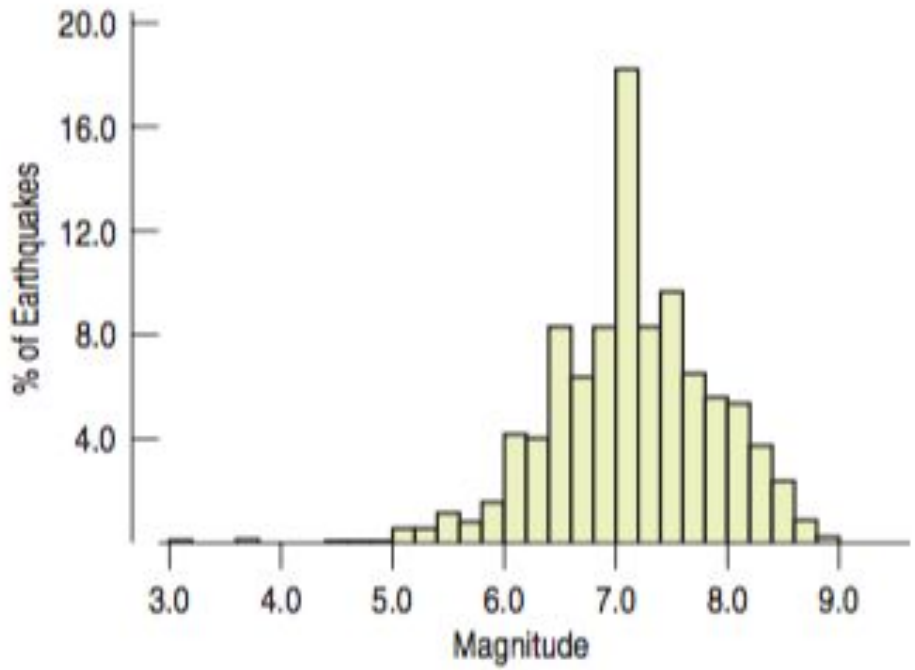


FIGURE 4.2
A relative frequency histogram looks just like a frequency histogram except for the labels on the *y*-axis, which now show the percentage of earthquakes in each bin.

1. *Does the histogram have a single, central hump or several separated humps?* These humps are called **modes**.[5] The earthquake magnitudes have a single mode at just about 7. A histogram with one peak, such as the earthquake magnitudes, is dubbed **unimodal**; histograms with two peaks are **bimodal**, and those with three or more are called **multimodal**.[6] For example, here's a bimodal histogram.

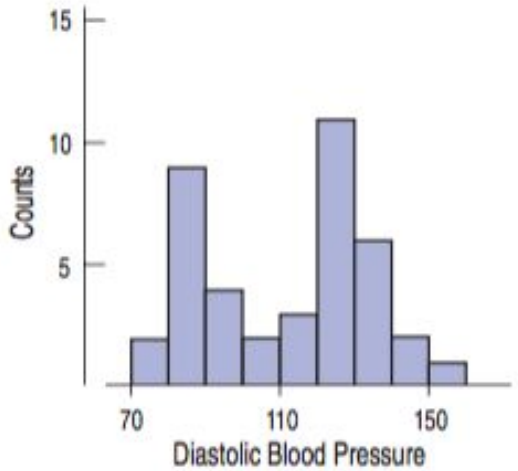


FIGURE 4.5
A bimodal histogram has two apparent peaks.

A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called **uniform**.

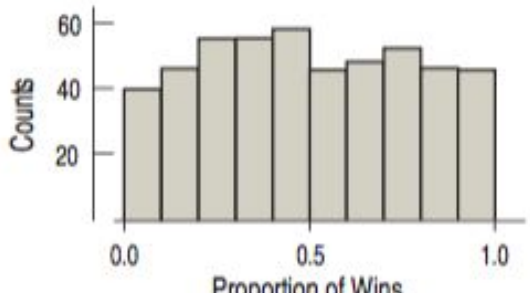


FIGURE 4.6
In a uniform histogram, the bars are all about the same height. The histogram doesn't appear to have a mode.

2. *Is the histogram* ***symmetric?*** Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?

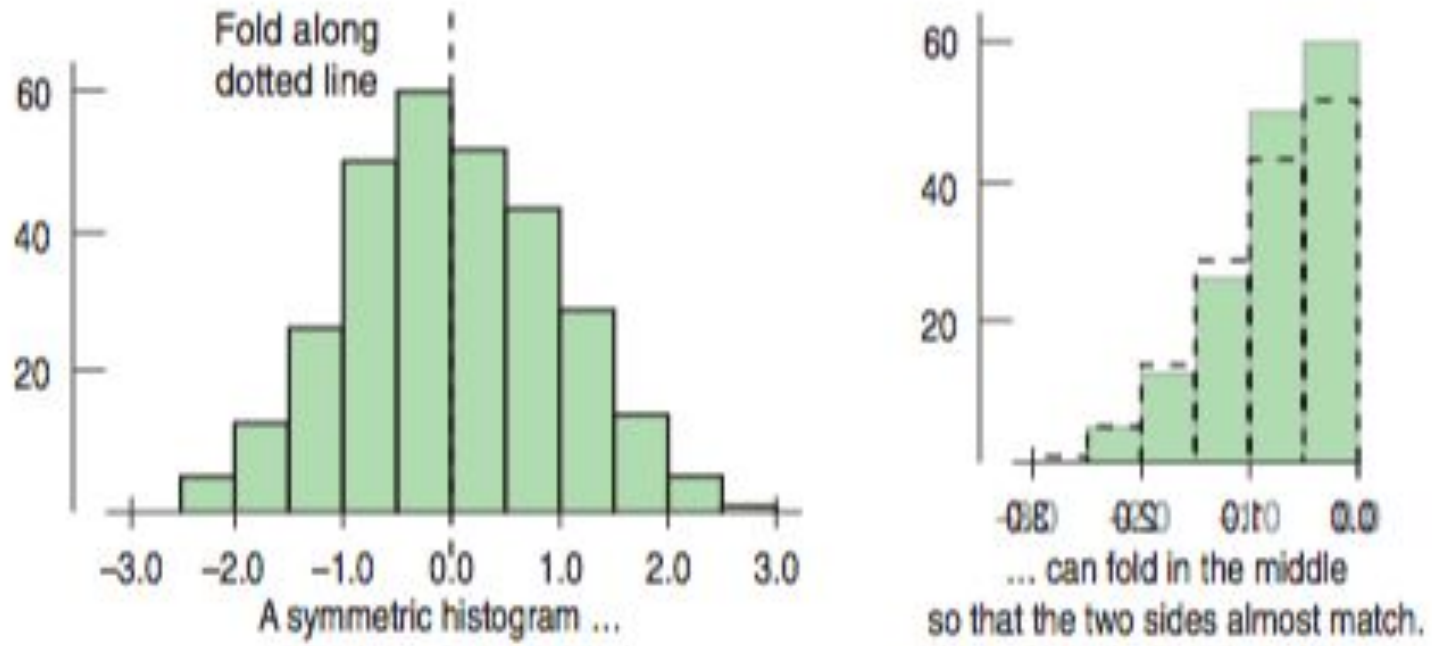


FIGURE 4.7

The (usually) thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.

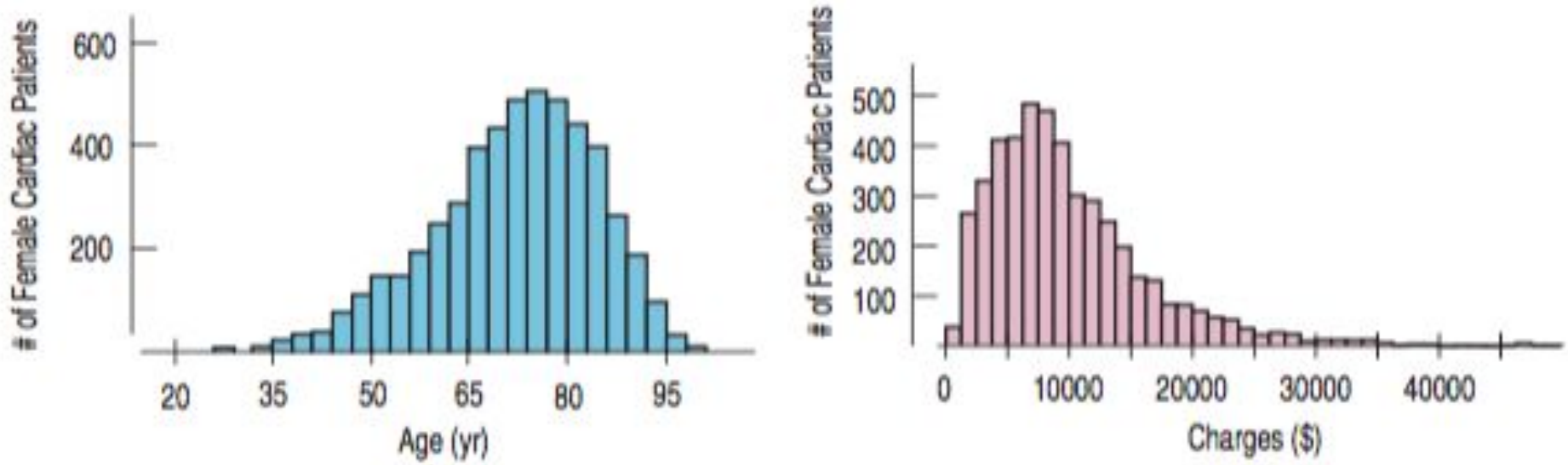


FIGURE 4.8
Two skewed histograms showing data on two variables for all female heart attack patients in New York State in one year. The blue one (age in years) is skewed to the left. The purple one (charges in $) is skewed to the right.

skewed to the right.

3. *Do any unusual features stick out?* Often such features tell us something interesting or exciting about the data. You should always mention any stragglers, or **outliers**, that stand away from the body of the distribution. If you're collecting data on nose lengths and Pinocchio is in the group, you'd probably notice him, and you'd certainly want to mention it.

   Outliers can affect almost every method we discuss in this course. So we'll always be on the lookout for them. An outlier can be the most informative part of your data. Or it might just be an error. But don't throw it away without comment. Treat it specially and discuss it when you tell about your data. Or find the error and fix it if you can. Be sure to look for outliers. Always.

In the next chapter you'll learn a handy rule of thumb for deciding when a point might be considered an outlier.

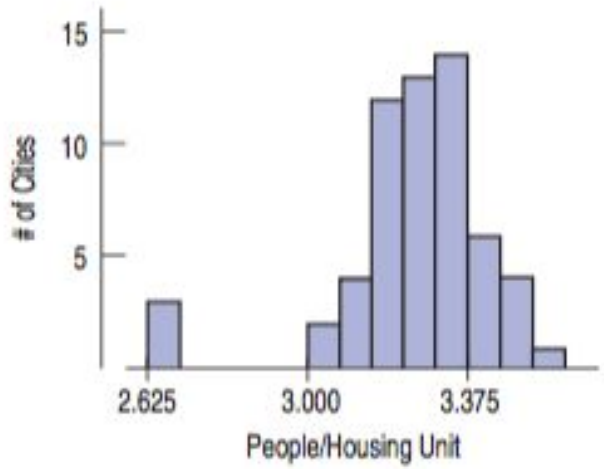


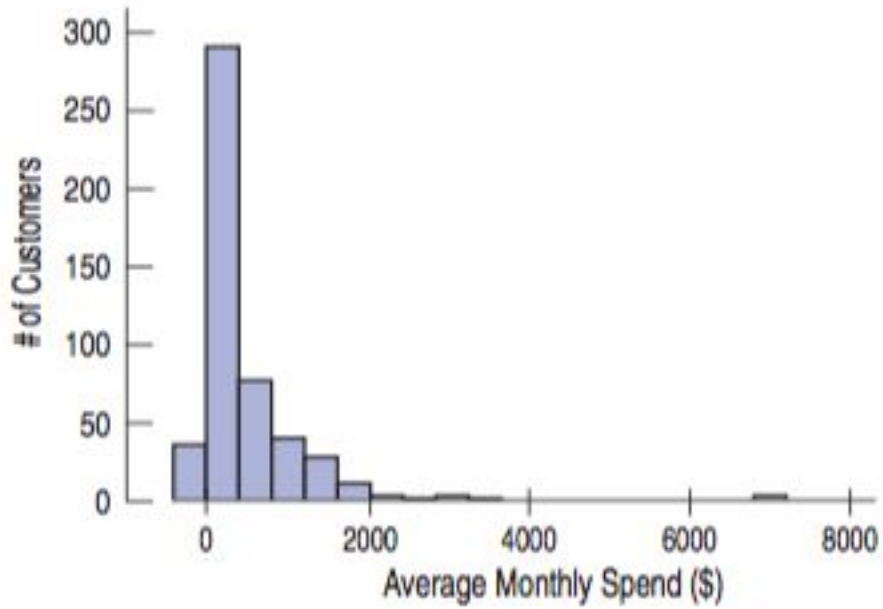
FIGURE 4.9
A histogram with outliers. There are three cities in the leftmost bar.

Credit card company

A credit card company wants to see how much customers in a particular segment of their market use their credit card. They have provided you with data[7] on the amount spent by 500 selected customers during a 3-month period and have asked you to summarize the expenditures. Of course, you begin by making a histogram.

QUESTION: Describe the shape of this distribution.

The distribution of expenditures is unimodal and skewed to the high end. There is an extraordinarily large value at about $7000, and some of the expenditures are negative.



Median

KRAKATAU posco

Let's return to the tsunami earthquakes. But this time, let's look at just 25 years of data: 176 earthquakes that occurred from 1981 through 2005. These should be more accurately measured than prehistoric quakes because seismographs were in wide use. Try to put your finger on the histogram at the value you think is typical. (Read the value from the horizontal axis and remember it.) When we think of a typical value, we usually look for the **center** of the distribution. Where do you think the center of this distribution is? For a unimodal, symmetric distribution such as these earthquake data, it's easy. We'd all agree on the center of symmetry, where we would fold the histogram to match the two sides. But when the distribution is skewed or possibly multimodal, it's not immediately clear what we even mean by the center.

One natural choice of typical value is the value that is literally in the middle, with half the values below it and half above it.

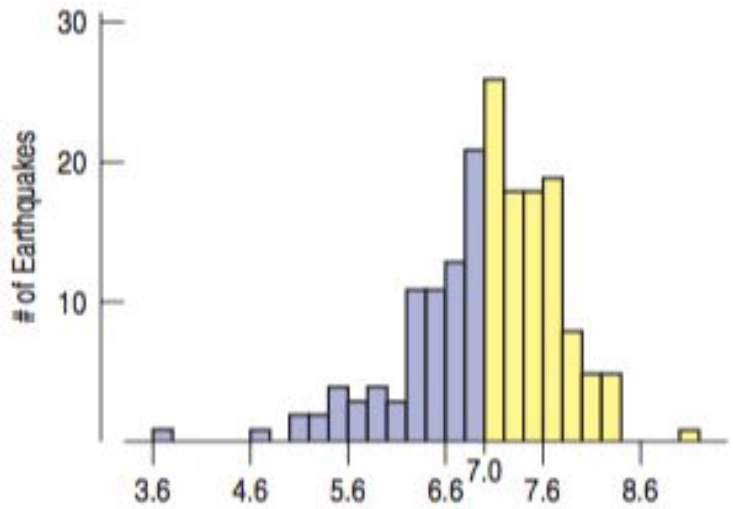


FIGURE 4.10
Tsunami-causing earthquakes (1981–2005).
The median splits the histogram into two halves of equal area.

How do medians work? Finding the median of a batch of n numbers is easy as long as you remember to order the values first. If n is odd, the median is the middle value. Counting in from the ends, we find this value in the $\frac{n+1}{2}$ position.

When n is even, there are two middle values. So, in this case, the median is the average of the two values in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

Here are two examples:

Suppose the batch has these values: 14.1, 3.2, 25.3, 2.8, −17.5, 13.9, 45.8.

First we order the values: −17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 45.8.

Since there are 7 values, the median is the $(7+1)/2 = 4$th value, counting from the top or bottom: 13.9. Notice that 3 values are lower, 3 higher.

Suppose we had the same batch with another value at 35.7. Then the ordered values are −17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7, 45.8.

The median is the average of the 8/2 or 4th, and the $(8/2)+1$, or 5th, values. So the median is $(13.9+14.1)/2 = 14.0$. Four data values are lower, and four higher.

Spread: The Interquartile Range

A better way to describe the spread of a variable might be to ignore the extremes and concentrate on the middle of the data. We could, for example, find the range of just the middle half of the data. What do we mean by the middle half? Divide the data in half at the median. Now divide both halves in half again, cutting the data into four quarters. We call these new dividing points **quartiles**. One quarter of the data lies below the **lower quartile**, and one quarter of the data lies above the **upper quartile**, so half the data lies between them. The quartiles border the middle half of the data.

The difference between the quartiles tells us how much territory the middle half of the data covers and is called the **interquartile range**. It's commonly abbreviated IQR (and pronounced "eye-cue-are"):

$$IQR = \textit{upper quartile} - \textit{lower quartile}.$$

For the earthquakes, there are 88 values below the median and 88 values above the median. The midpoint of the lower half is the average of the 44th and 45th values in the ordered data; that turns out to be 6.6. In the upper half we average the 132nd and 133rd values, finding a magnitude of 7.6 as the third quartile. The *difference* between the quartiles gives the IQR:

$$IQR = 7.6 - 6.6 = 1.0.$$

Now we know that the middle half of the earthquake magnitudes extends across a (interquartile) range of 1.0 Richter scale units. This seems like a reasonable summary of the spread of the distribution, as we can see from this histogram:

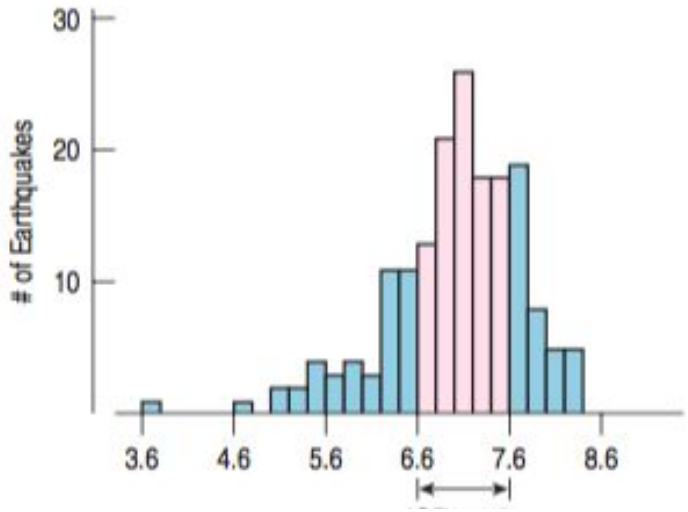


FIGURE 4.11
The quartiles bound the middle 50% of the values of the distribution. This gives a visual indication of the spread of the data. Here we see that the IQR is 1.0 Richter scale units.

Mean or Median?

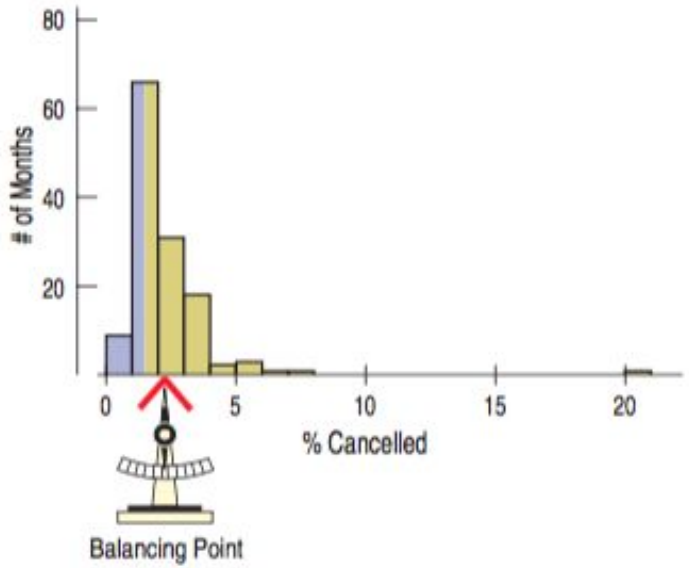


FIGURE 4.13
The median splits the area of the histogram in half at 1.755%. Because the distribution is skewed to the right, the mean (2.28%) is higher than the median. The points at the right have pulled the mean toward them away from the median.

The mean is 2.28%, but nearly 70% of months had cancellation rates below that, so the mean doesn't feel like a good overall summary. Why is the balancing point so high? The large outlying value pulls it to the right. For data like these, the median is a better summary of the center.

Because the median considers only the order of the values, it is **resistant** to values that are extraordinarily large or small; it simply notes that they are one of the "big ones" or the "small ones" and ignores their distance from the center.

False Cause

- Citing a false or remote cause to explain a situation

Example:
The increase in global warming in the past decade is because more teenager in the past using hairspray

KRAKATAU posco

Mean or Median ??

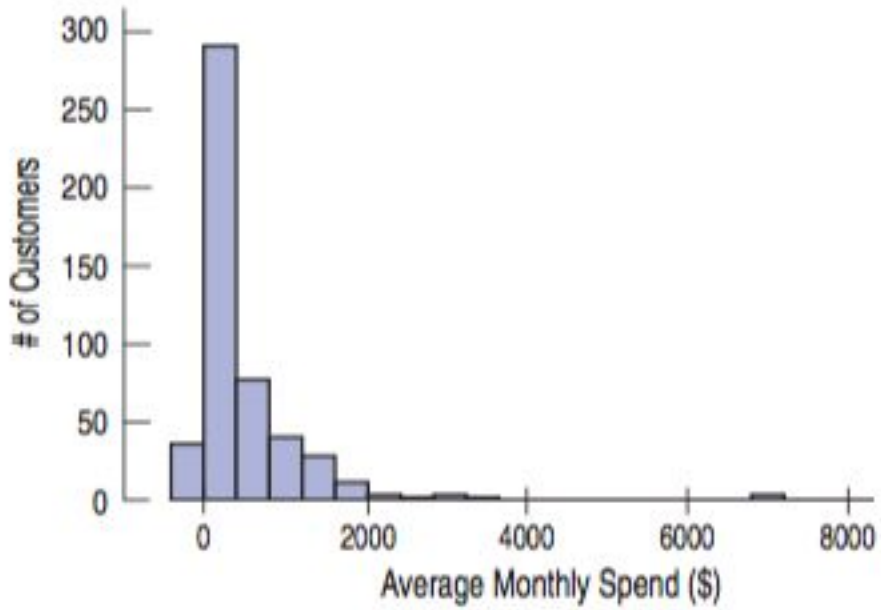
Of course, to choose between mean and median, we'll start by looking at the data. If the histogram is symmetric and there are no outliers, we'll prefer the mean. However, if the histogram is skewed or has outliers, we're usually better off with the median. If you're not sure, report both and discuss why they might differ.

KRAKATAU posco

Question: Mean or Median?

RECAP: You want to summarize the expenditures of 500 credit card company customers, and have looked at a histogram.

QUESTION: You have found the mean expenditure to be $478.19 and the median to be $216.28. Which is the more appropriate measure of center, and why?



Because the distribution of expenditures is skewed, the median is the more appropriate measure of center. Unlike the mean, it's not affected by the large outlying value or by the skewness. Half of these credit card customers had average monthly expenditures less than $216.28 and half more.

What About Spread? The Standard Deviation

The IQR is always a reasonable summary of spread, but because it uses only the two quartiles of the data, it ignores much of the information about how individual values vary. A more powerful approach uses the **standard deviation**, which takes into account how far *each* value is from the mean. Like the mean, the standard deviation is appropriate only for symmetric data.

One way to think about spread is to examine how far each data value is from the mean. This difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel each other out. So the average deviation is always zero—not very helpful.

To keep them from canceling out, we *square* each deviation. Squaring always gives a positive value, so the sum won't be zero. That's great. Squaring also emphasizes larger differences—a feature that turns out to be both good and bad.

When we add up these squared deviations and find their average (almost), we call the result the **variance**:

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}.$$

What to *Tell* About a Quantitative Variable

KRAKATAU posco

What should you *Tell* about a quantitative variable?

- Start by making a histogram or stem-and-leaf display, and discuss the shape of the distribution.
- Next, discuss the center and spread.
  - Always pair the median with the IQR and the mean with the standard deviation. It's not useful to report one without the other. Reporting a center without a spread is dangerous. You may think you know more than you do about the distribution. Reporting only the spread leaves us wondering where we are.
  - If the shape is skewed, report the median and IQR. You may want to include the mean and standard deviation as well, but you should point out why the mean and median differ.
  - If the shape is symmetric, report the mean and standard deviation and possibly the median and IQR as well. For unimodal symmetric data, the IQR is usually a bit larger than the standard deviation. If that's not true of your data set, look again to make sure that the distribution isn't skewed and there are no outliers.

KRAKATAU posco

Frequency and Mode

- Frequency: the number of times the value occurs
- Relative frequency: the percentage that the value occurs

Category	Apple	Orange	Melon	Kiwi	Plum	**Total**
Frequency	10	15	8	10	7	**50**
Relative Frequency	20%	30%	16%	20%	14%	**100%**

- Mode: the most frequent attribute value
  - (3, 5, 3, 10, 4)
- Frequency and mode are typically used with categorical data

KRAKATAU POSCO

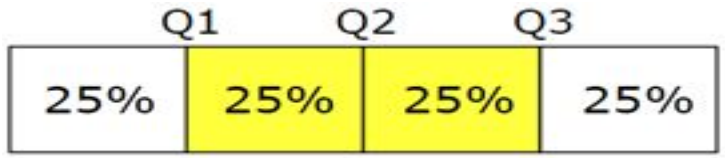
Percentiles

- For continuous data, a percentile is more useful.
- Given an ordinal or continuous attribute x, its p^{th} percentile ([0, 100]) is
  - a value x_p such that p% of the observed values are less than x_p.
- E.g. the 20th percentile is the value below which 20% of all values of the observations may be found.

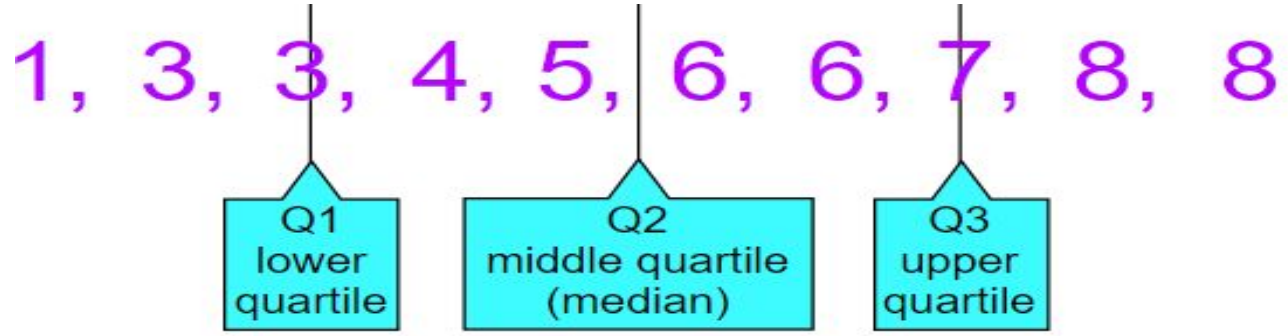
KRAKATAU POSCO

Quartiles

- Values that divide a list of numbers into quarters Q1, Q2, Q3
  - Percentiles 25^{th},50^{th}, 75^{th}



  - Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8
    - Put the list of numbers **in order**
    - Then cut the list into **four equal parts**
    - If a quartile is half way, use the average



Quartile 1 (Q1) = **3**
Quartile 2 (Q2) = **5.5**
Quartile 3 (Q3) = **7**

KRAKATAU POSCO

Mean and Median

- Measure of the locations of a set of points
  - Mean
    - average of all values

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

    - very sensitive to outliers

$$mean(1,2,5,92) = \frac{1+2+5+92}{4} = 25$$

  - Median
    - Q2 or 50th percentile: the value whose occurrence lies in the middle of a set of ordered values
    - median(1, 20, 25, 30, 31) = ?
    - median(1, 2, 5, 92) = ?

Extreme example

- Income in small town of 6 people
  - $25,000 $27,000 $29,000
  - $35,000 $37,000 $38,000
- Mean is $31,830 and median is $32,000
- Bill Gates moves to town
  - $25,000 $27,000 $29,000
  - $35,000 $37,000 $38,000 $40,000,000
- Mean is $5,741,571 median is $35,000
- Mean is pulled by the outlier while the median is not. The median is a better of measure of center for data with extreme values.
  - E.g. house prices, incomes, student exam marks, ...

KRAKATAU POSCO

Range and Variance

- Measures of data spread
  - Range: the difference between the max and min
    - (1, 2, 5, 70, 92): range = 92 – 1 = 91
  - Variance: The average of the squared differences from the Mean

$$variance(x) = s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Value	2	3	3	4	**Mean = 3**
Diff	1	0	0	1	
$Diff^2$	1	0	0	1	**Variance = 2/4 = 0.5**

  - Standard deviation: the square root of Variance
- Sensitive to outliers
  - Inter-quartile range: Q3 – Q1

2. Visualization

- Convert data into a visual or tabular format for easier understanding
- One of the most powerful and appealing techniques for data exploration
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

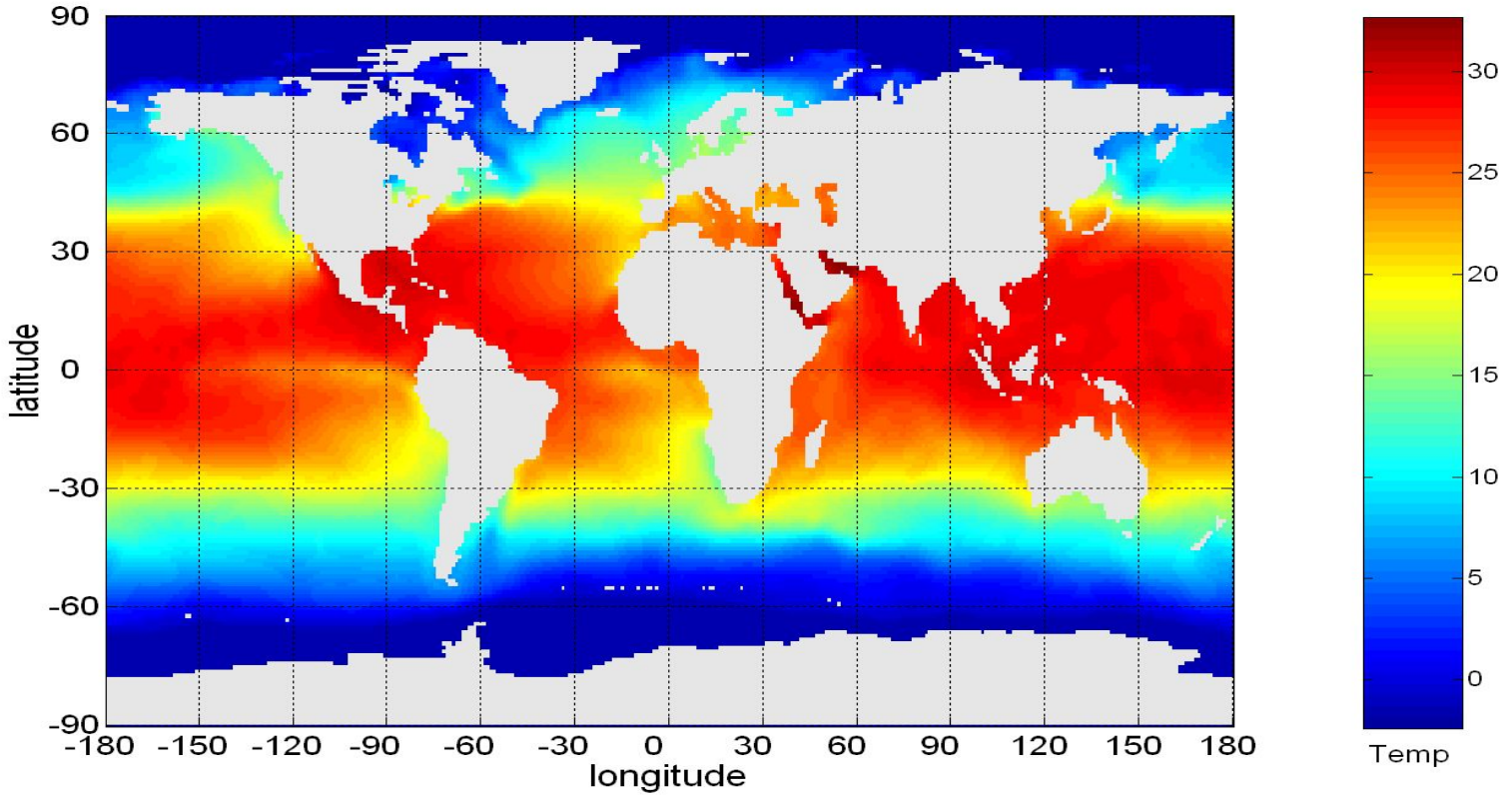
KRAKATAU posco

Visualization

- Data objects, their attributes, and the relationships are translated into graphical elements such as points, lines, shapes, and colors
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

Visualization Example

- Sea Surface Temperature for July 1982
  - Tens of thousands of data points are summarized in a single figure



KRAKATAU posco

EDA Strategy

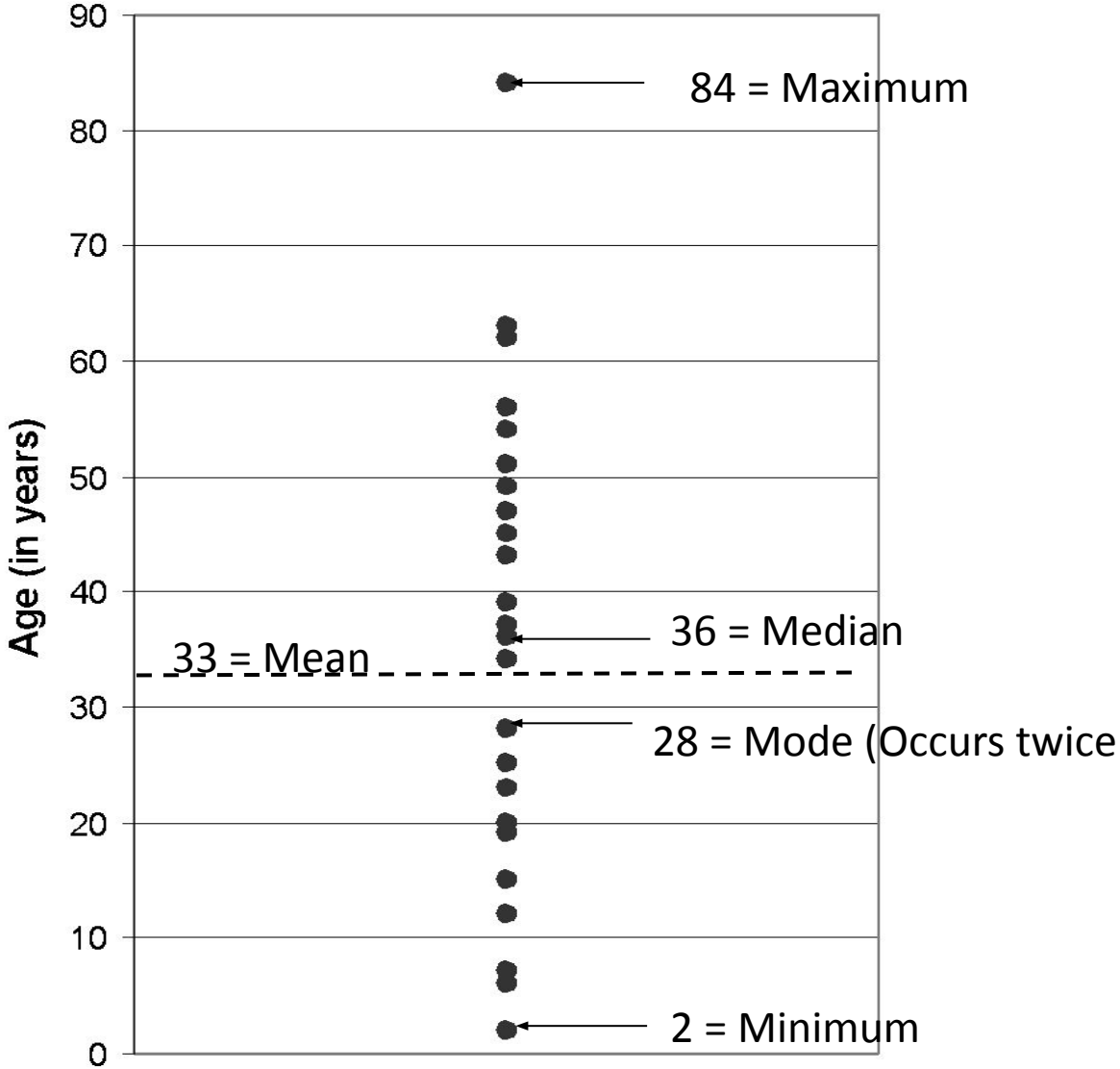
- Examine variables one by one, then look at the relationships among the different variables
- Start with graphs, then add numerical summaries of specific aspects of the data
  - Be aware of attribute types
    - Categorical vs. Numeric

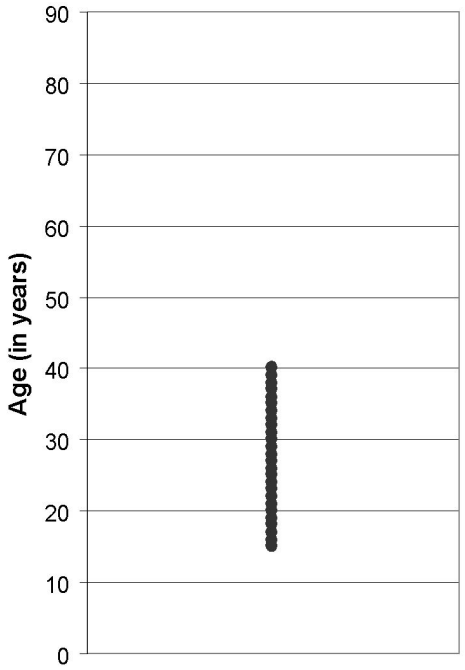
KRAKATAU posco

EDA: One Variable

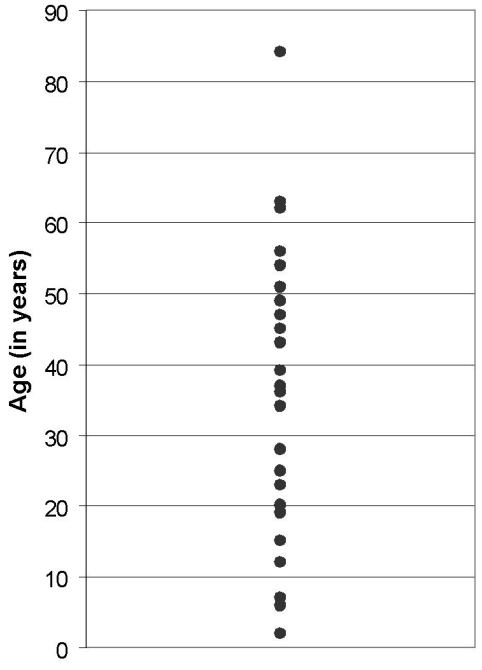
- Summary statistics
  - Categorical: frequency table
  - Numeric: mean, median, standard deviation, range, etc.
- Graphical displays
  - Categorical: bar chart, pie chart, etc.
  - Numeric: histogram, boxplot, etc.
- Probability models
  - Categorical: Binomial distribution (won’t cover)
  - Numeric: Normal curve (won’t cover)

Summary Statistics





SD = 7.6



SD = 20.4

Frequency Table

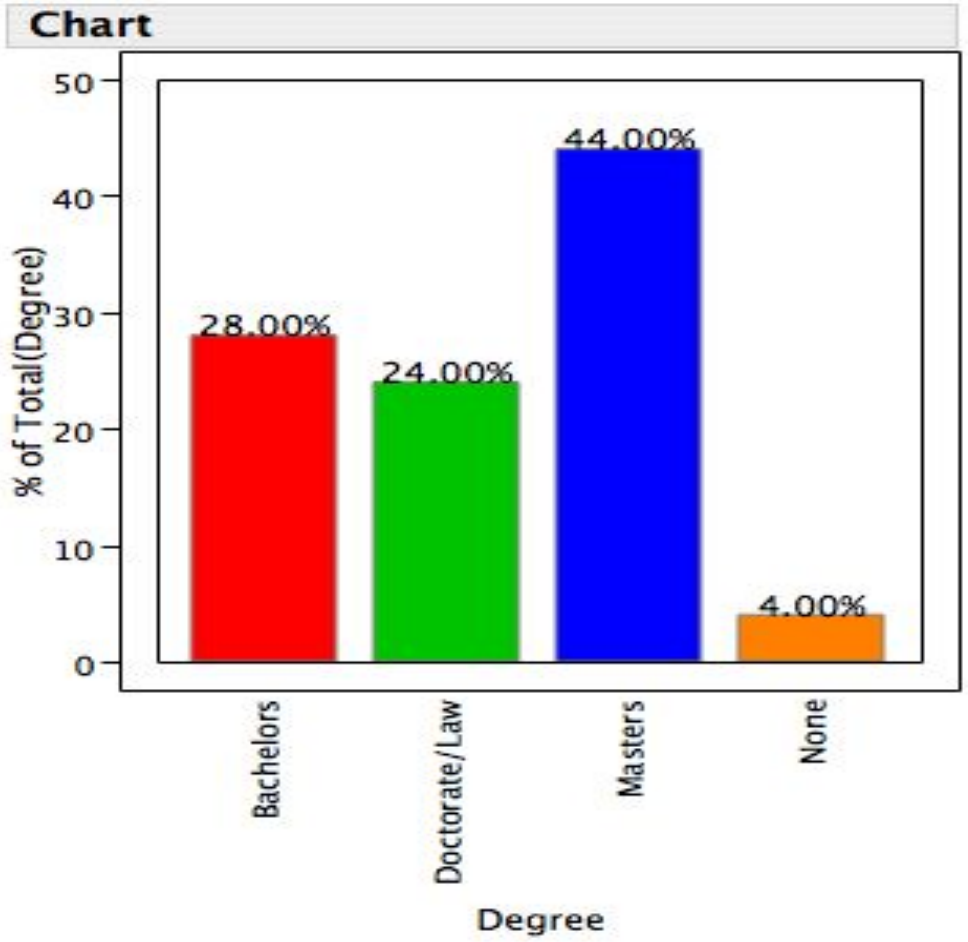
CEO	Company	Degree
Michael D. Eisner	Walt Disney	Bachelors
Mel Karmazin	CBS	Bachelors
Stephen M. Case	America Online	Bachelors
Stephen C. Hilbert	Conseco	None
Craig R. Barrett	Intel	Doctorate
Millard Drexler	Gap	Masters
John F. Welch, Jr.	General Electric	Doctorate
Thomas G. Stemberg	Staples	Masters
Henry R. Silverman	Cendant	JD (law)
Reuben Mark	Colgate-Palmolive	Masters
Philip J. Purcell	Morgan Stanley Dean Witter	Masters
Scott G. McNealy	Sun Microsystems	Masters
Margaret C. Whitman	eBay	Masters
Louis V. Gerstner, Jr.	IBM	Masters
John F. Gifford	Maxim Integrated Products	Bachelors
Robert L. Waltrip	Service Corp. International	Bachelors
M. DouglasIvester	Coca-Cola	Bachelors
Gordon M. Binder	Amgen	Masters
Charles R. Schwab	Charles Schwab	Masters
William R. Steere, Jr.	Pfizer	Bachelors
Nolan D. Archibald	Black & Decker	Masters
Charles A. Heimbold, Jr.	Bristol-Myers Squibb	LLB (law)
Wiliam L. Larson	Network Association	JD (law)
Maurice R. Greenberg	American International Group	LLB (law)
Richard Jay Kogan	Schering-Plough	Masters

Class	Frequency	Relative Frequency
Highest Degree Obtained	Number of CEOs	Proportion
None	1	0.04
Bachelors	7	0.28
Masters	11	0.44
Doctorate / Law	6	0.24
Totals	25	1.00

KRAKATAU posco

Bar Graph

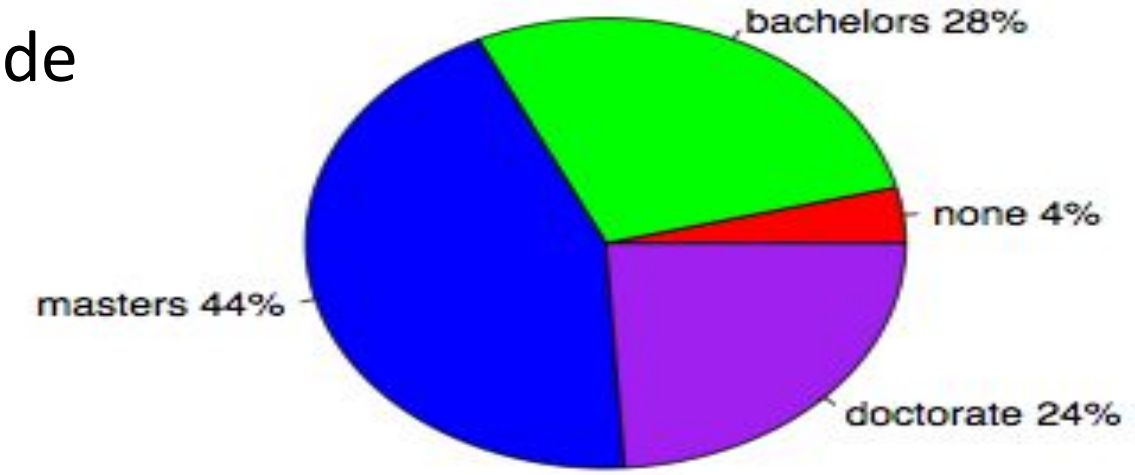
- The bar graph quickly compares the degrees of the four groups
- The heights of the four bars show the counts for the four degree categories



KRAKATAU posco

Pie Chart

- A pie chart helps us see what part of the whole group forms
- To make a pie chart, you must include all the categories that make up a whole



- WARNING: pie charts are often a poor choice – they show only relative data. A bar chart is often better.

KRAKATAU posco

Histogram

- Split data range into equal-sized bins.
- Count the number of points in each bin.

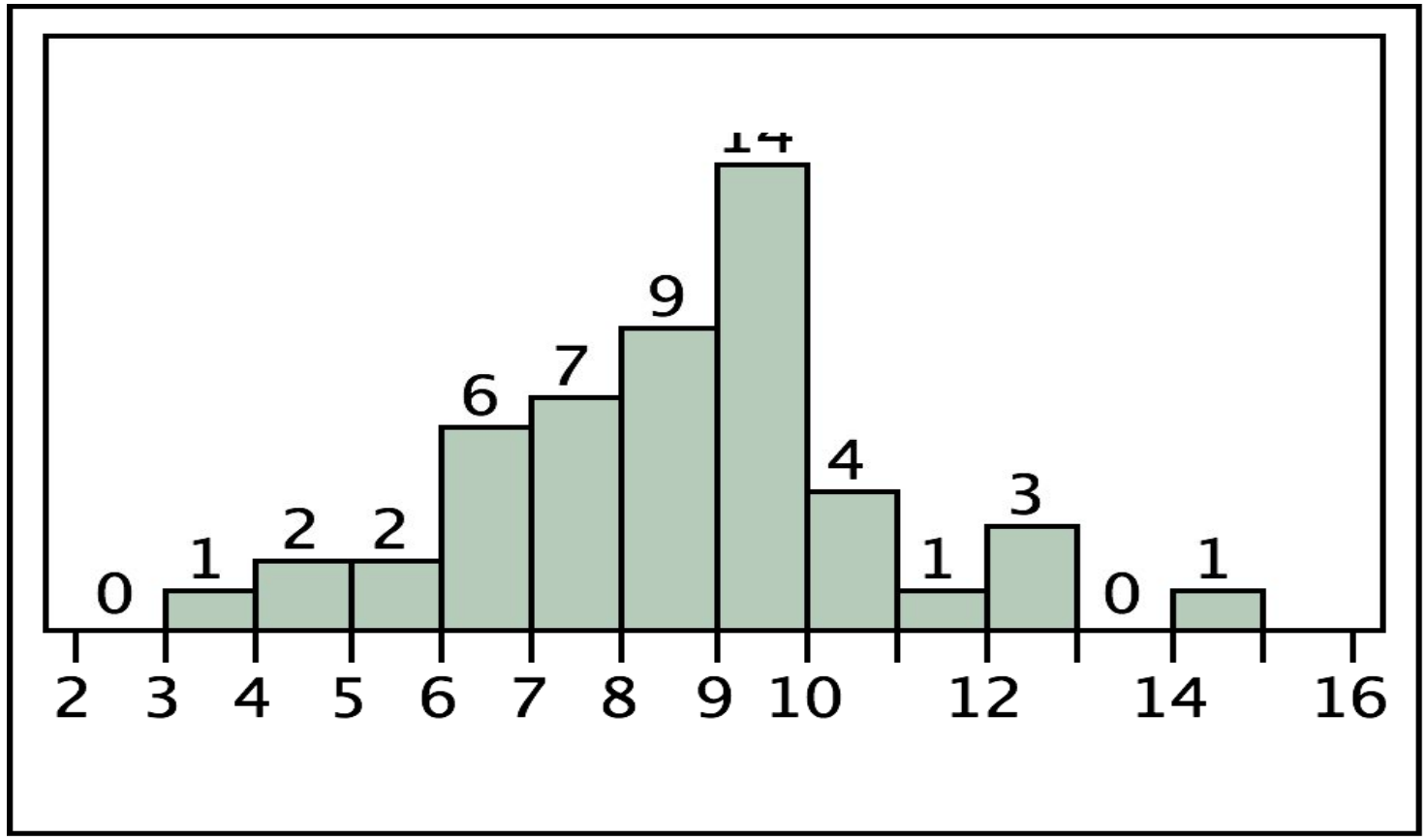
State Unemployment rates Dec 2010			
3.8	7.3	8.6	9.8
4.5	7.4	9	9.8
4.6	7.5	9	9.9
5.4	7.6	9.2	10.1
5.7	7.9	9.2	10.2
6.4	8	9.3	10.6
6.6	8.2	9.4	10.6
6.6	8.2	9.4	11.6
6.8	8.2	9.4	12
6.8	8.3	9.4	12.4
6.9	8.4	9.6	12.4
7.1	8.5	9.7	14.3
7.2	8.6		

The bins are:

$3.0 \leq rate < 4.0$
$4.0 \leq rate < 5.0$
$5.0 \leq rate < 6.0$
$6.0 \leq rate < 7.0$
$7.0 \leq rate < 8.0$
$8.0 \leq rate < 9.0$
$9.0 \leq rate < 10.0$
$10.0 \leq rate < 11.0$
$11.0 \leq rate < 12.0$
$12.0 \leq rate < 13.0$
$13.0 \leq rate < 14.0$
$14.0 \leq rate < 15.0$

KRAKATAU posco

Histograms

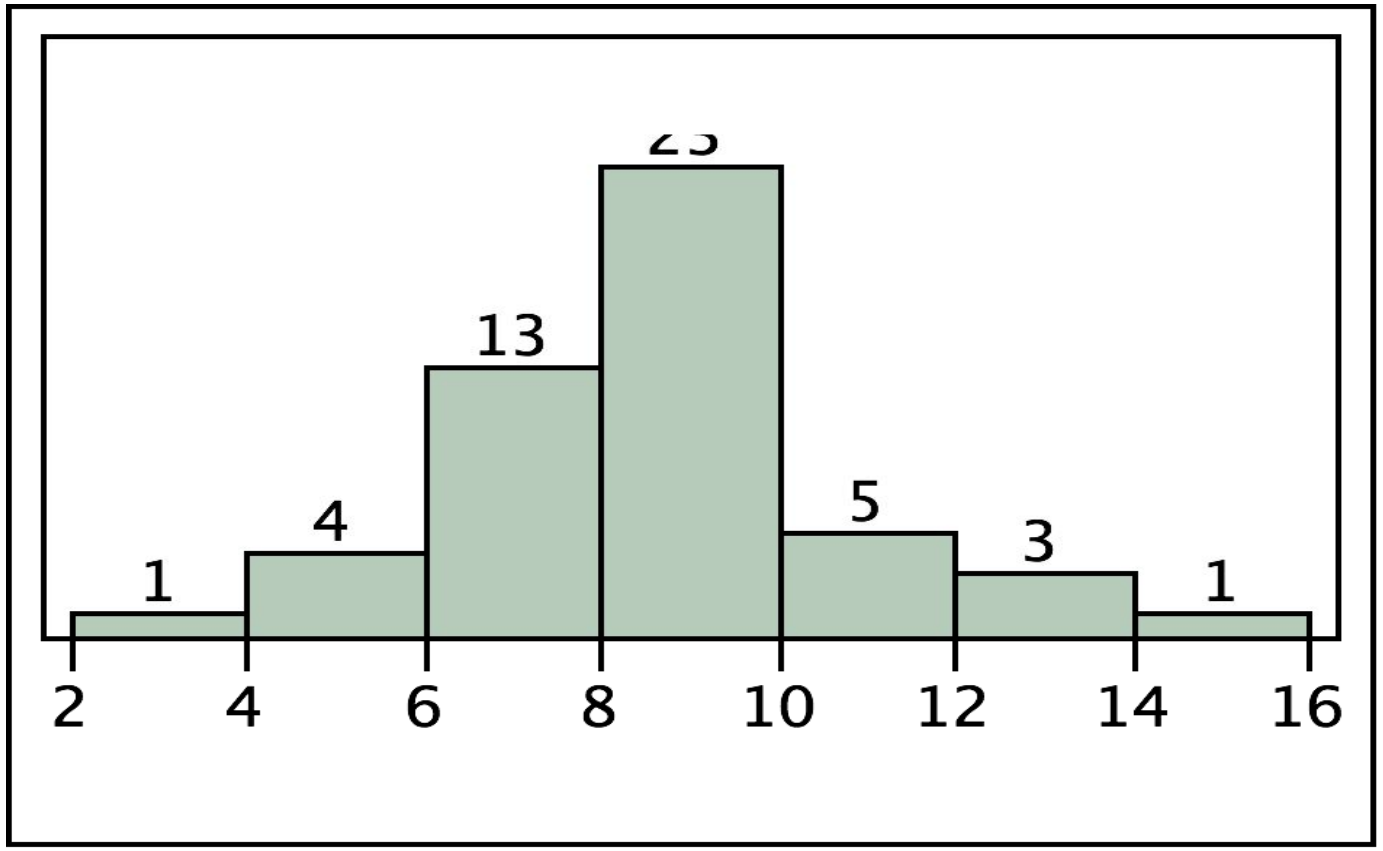


The bins are:

$3.0 \leq rate < 4.0$
$4.0 \leq rate < 5.0$
$5.0 \leq rate < 6.0$
$6.0 \leq rate < 7.0$
$7.0 \leq rate < 8.0$
$8.0 \leq rate < 9.0$
$9.0 \leq rate < 10.0$
$10.0 \leq rate < 11.0$
$11.0 \leq rate < 12.0$
$12.0 \leq rate < 13.0$
$13.0 \leq rate < 14.0$
$14.0 \leq rate < 15.0$

KRAKATAU posco

Histograms



The bins are:

$2.0 \leq rate < 4.0$
$4.0 \leq rate < 6.0$
$6.0 \leq rate < 8.0$
$8.0 \leq rate < 10.0$
$10.0 \leq rate < 12.0$
$12.0 \leq rate < 14.0$
$14.0 \leq rate < 16.0$
$16.0 \leq rate < 18.0$

KRAKATAU posco

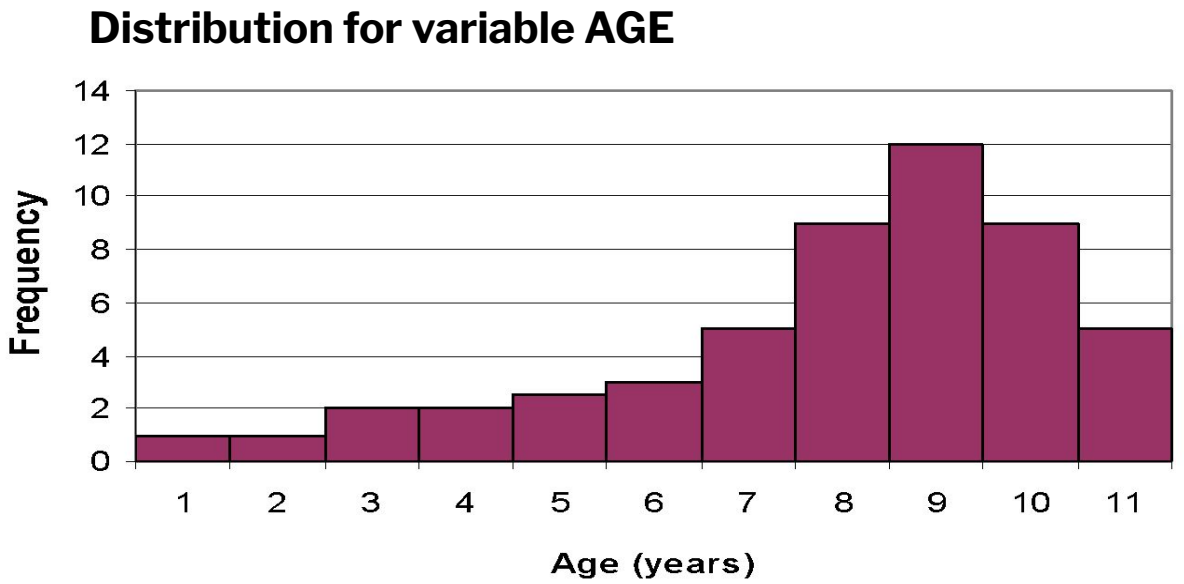
Histograms

- Where did the bins come from?
  - They were chosen rather arbitrarily
- Does choosing other bins change the picture?
  - Yes!! And sometimes dramatically
- What do we do about this?
  - Some pretty smart people have come up with some “optimal” bin widths
  - we will rely on their suggestions

KRAKATAU POSCO

Histogram

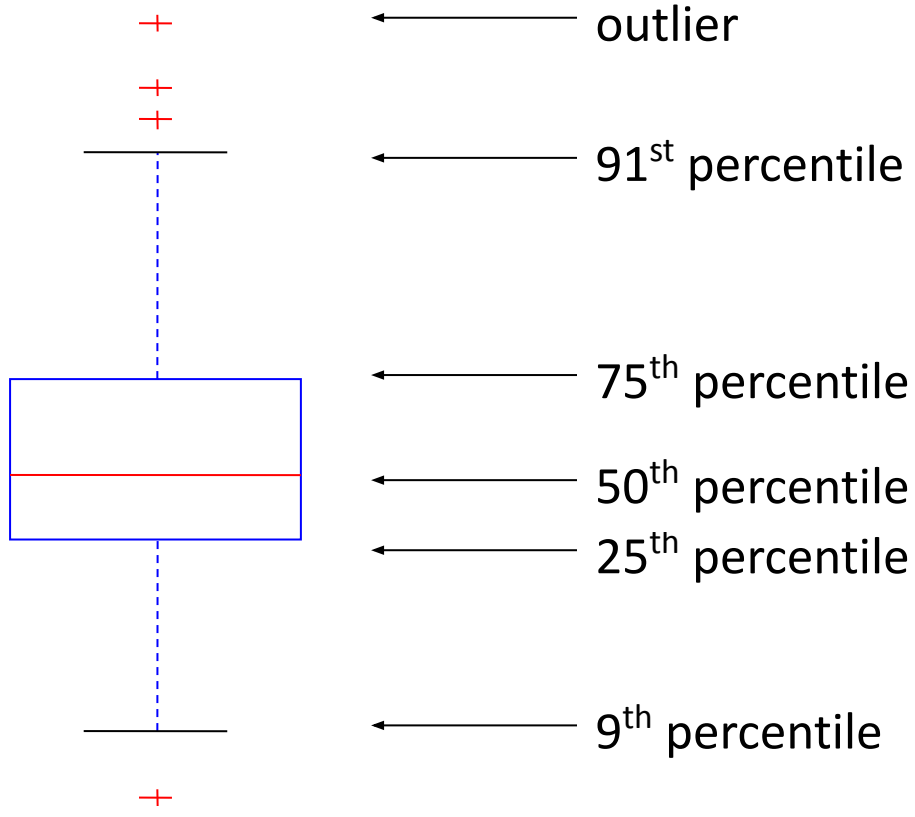
- The histogram can illustrate features related to the distribution of the data, e.g.,
  - Shape
    - Symmetric, skewed right, skewed left, bell shaped
  - presence of outliers
  - presence of multiple modes
  - "a camel with two humps" is a common grade distribution for ICT courses!



KRAKATAU POSCO

Box Plots

- Box Plots
  - Show data distribution
  - A graph of five numbers (often called the five number summary)
    - Minimum
    - 1st quartile
    - Median
    - 3rd quartile
    - Maximum
  - Various definitions for min-max
    - Min & Max of all data
    - 9th and 91st percentile
    - 2nd and 98th percentile
    - 1.5 Inter-quartile range



EDA: Two Variables

- Looking at TWO variables at once helps us to explore relationships between different attributes.
  - sometimes we can see clear relationships and clusters
  - next week we will look at clustering in more detail
  - this week we focus on exploratory analysis of two variables
- Three combinations of variables
  - 2 categorical variables
    - Contingency tables
  - 1 categorical and 1 numeric variable
    - Side-by-side box plots, counts, etc.
  - 2 numeric variables
    - Scatter plots, correlations, regressions

KRAKATAU POSCO

Contingency Tables

- Show the *frequency distribution* of one variable in rows and another in columns
- E.g. a study of sex differences in handedness with 100 samples
  - Sex (male, female) and Handedness (right, left)

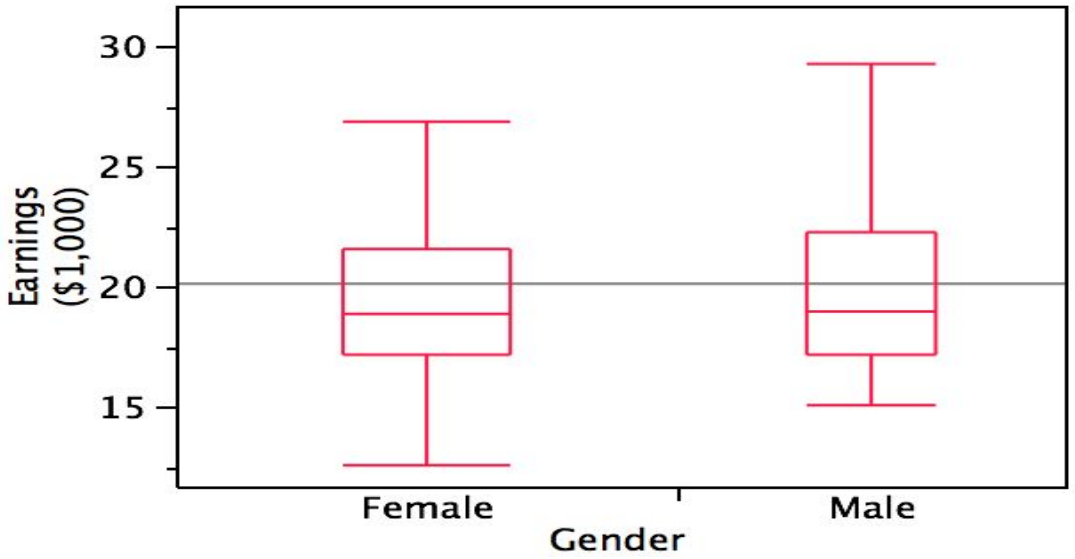
Handed-ness / Gender	Right handed	Left handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Marginal totals

Grand total

Side-by-side box plots

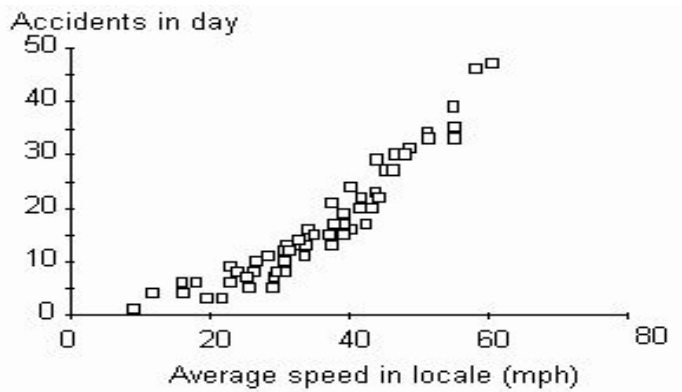
- Side-by-side box plots are graphical summaries of data when one variable is categorical and the other numeric
- Used to compare the distributions associated with the numeric variable across the levels of the categorical variable

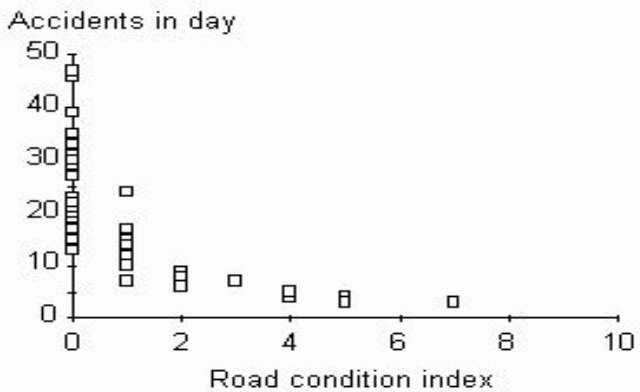


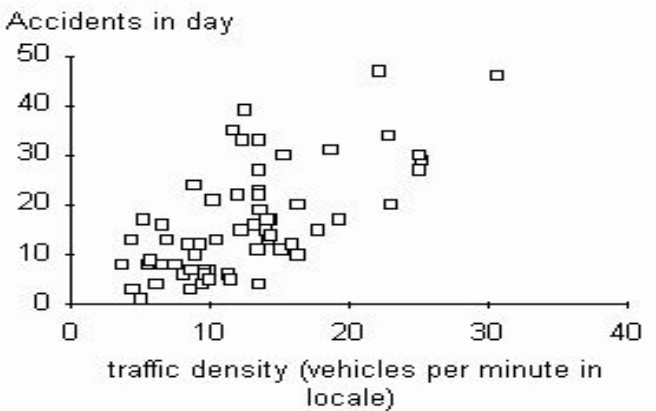
Scatter Plots

- One variable on x-axis and another on y-axis
  - Attributes values determine the position
  - Arrays of scatter plots can compactly summarize the relationships of pairs of attributes

Road condition index	Average speed (mph)	Traffic density (veh/min)	Accidents in day
1	28.4	13.4	11
1	37.6	4.3	13
0	39.4	14.3	17
7	19.6	4.4	3
1	31.0	6.8	13
5	16.2	6.1	4

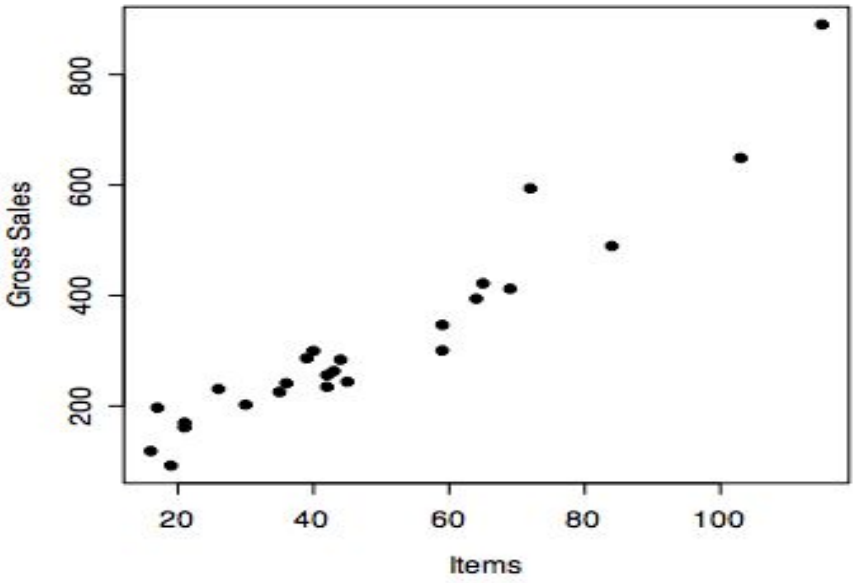






Describing scatter plots

- Form
  - Linear, quadratic, exponential
- Direction
  - Positive association
    - An increase in one variable is accompanied by an increase in the other
  - Negatively associated
    - A decrease in one variable is accompanied by an increase in the other
- Strength
  - How closely the points follow a clear form



- Form: Linear
- Direction: Positive
- Strength: Strong