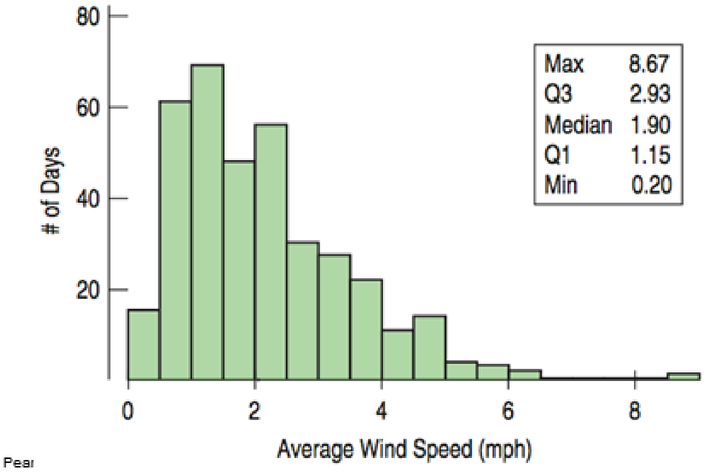
Chapter 5 - Understanding and Comparing Distributions

Obj - SWBAT select a suitable display for comparing groups, and understand the effectiveness of comparing distributions with histograms and boxplots.

The Big Picture

- We can answer much more interesting questions about variables when we compare distributions for different groups.
- Below is a histogram of the Average Wind Speed for every day in 1989.



The Five-Number Summary

- The five-number summary of a distribution reports its median, quartiles, and extremes (maximum and minimum).
- Example: The five-number summary for for the daily wind speed is:

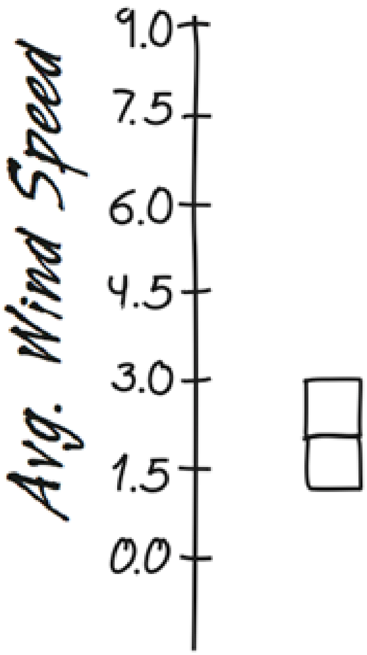
Max	8.67
Q_3	2.93
Median	1.90
Q_1	1.15
Min	0.20

Daily Wind Speed: Making Boxplots

- A boxplot is a graphical display of the five-number summary.
- Boxplots are useful when comparing groups.
- Boxplots are particularly good at pointing out outliers.

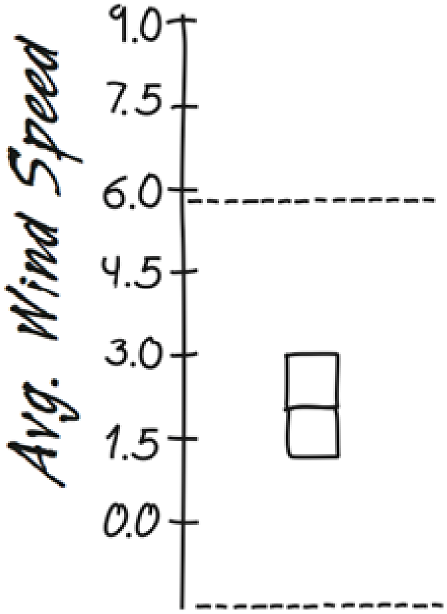
Constructing Boxplots

- Draw a single vertical axis spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.



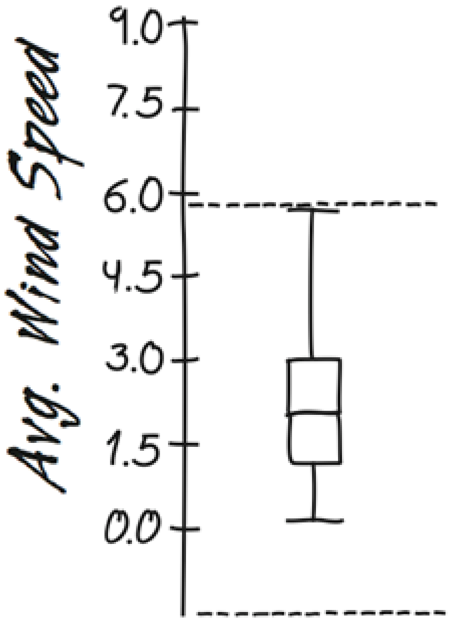
Constructing Boxplots (cont.)

- Erect “fences” around the main part of the data.



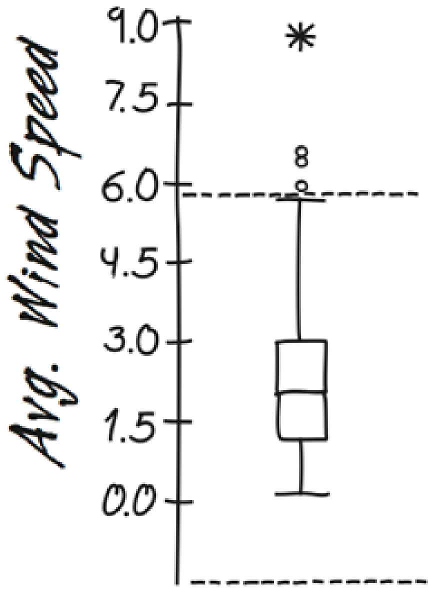
Constructing Boxplots (cont.)

- Use the fences to grow "whiskers."
- Draw lines from the ends of the box up and down to the most extreme data values found within the fences.
- If a data value falls outside one of the fences, we do not connect it with a whisker.



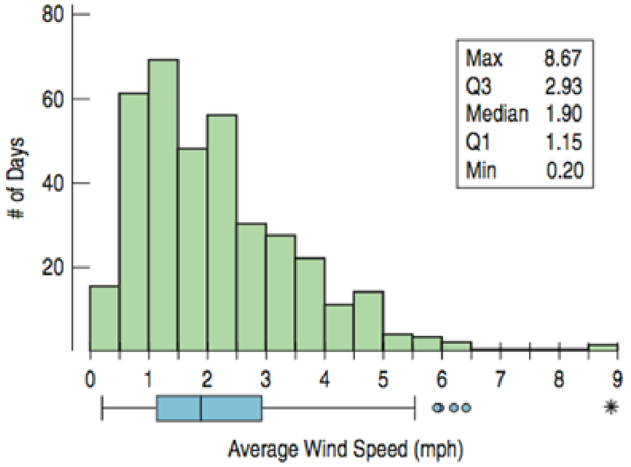
Constructing Boxplots (cont.)

- Add the outliers by displaying any data values beyond the fences with special symbols.
- We often use a different symbol for "far outliers" that are farther than 3 IQRs from the quartiles.



Wind Speed: Making Boxplots (cont.)

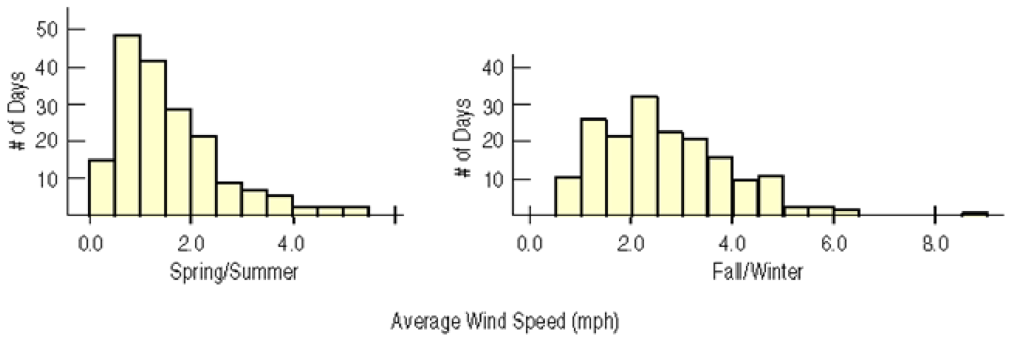
- Compare the histogram and boxplot for daily wind speeds:



- How does each display represent the distribution?

Comparing Groups

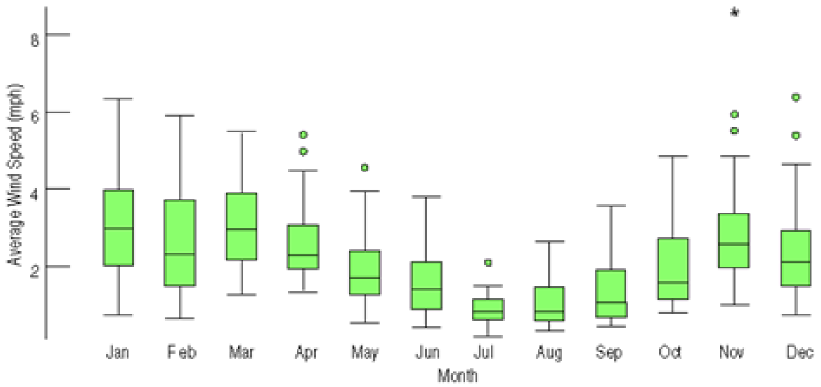
- It is almost always more interesting to compare groups.
- With histograms, note the shapes, centers, and spreads of the two distributions.



- What does this graphical display tell you?

Comparing Groups (cont.)

- Boxplots offer an ideal balance of information and simplicity, hiding the details while displaying the overall summary information.
- We often plot them side by side for groups or categories we wish to compare.



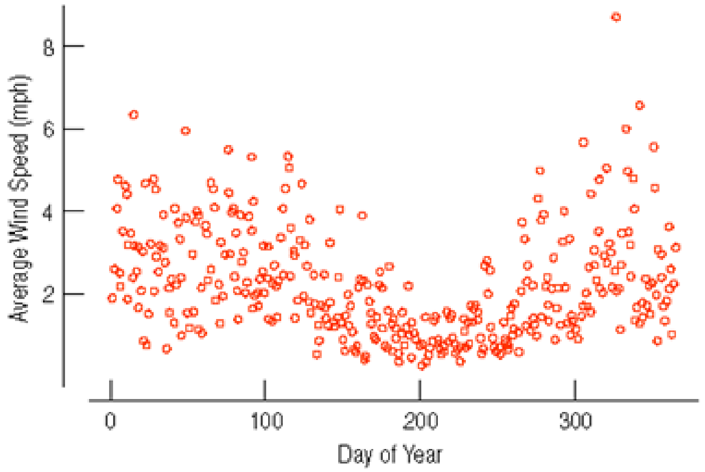
- What do these boxplots tell you?

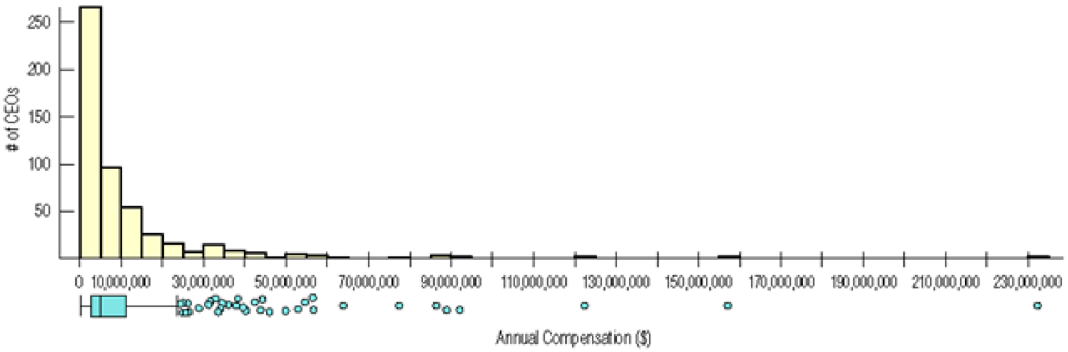
What About Outliers?

- If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be quite revealing.
- Note: The median and IQR are not likely to be affected by the outliers.

Timeplots: Order, Please!

- For some data sets, we are interested in how the data behave over time. In these cases, we construct timeplots of the data.

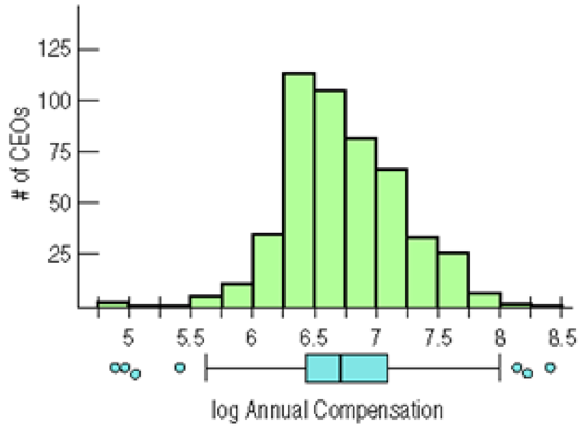




- When the data are skewed it can be hard to summarize them simply with a center and spread, and hard to decide whether the most extreme values are outliers or just part of a stretched out tail.
- How can we say anything useful about such data?

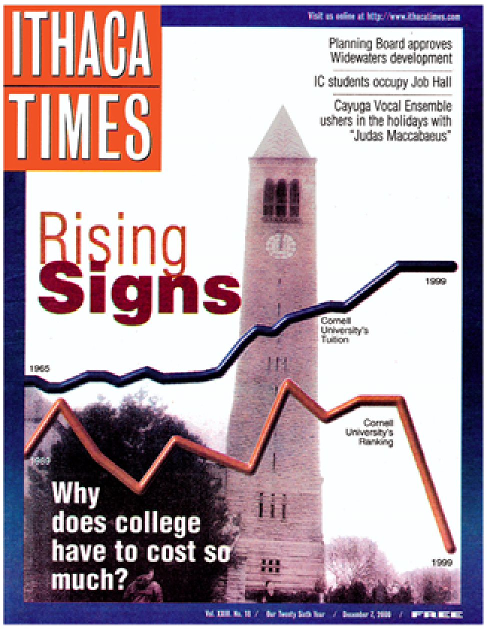
*Re-expressing Skewed Data to Improve Symmetry (cont.)

- One way to make a skewed distribution more symmetric is to re-express or transform the data by applying a simple function (e.g., logarithmic function).
- Note the change in skewness from the raw data (previous slide) to the transformed data (right):



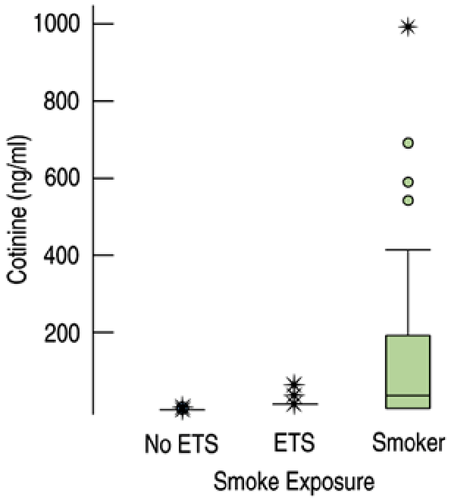
What Can Go Wrong? (cont.)

- Avoid inconsistent scales, either within the display or when comparing two displays.
- Label clearly so a reader knows what the plot displays.
- Good intentions, bad plot:


ITHACA TIMES
Visit us online at http://www.ithacatimes.com
Planning Board approves Widewaters development
IC students occupy Job Hall
Cayuga Vocal Ensemble ushers in the holidays with "Judas Maccabaeus"
Rising Signs
1999
Cornell University's Tuition
1965
Cornell University's Ranking
1999
Why does college have to cost so much?
Vol. XXIII, No. 18 / Our Twenty Sixth Year / December 7, 2000 / FREE

What Can Go Wrong? (cont.)

- Beware of outliers
- Be careful when comparing groups that have very different spreads.
- Consider these side-by-side boxplots of cotinine levels:
- Re-express . . .



What have we learned?

- We've learned the value of comparing data groups and looking for patterns among groups and over time.
- We've seen that boxplots are very effective for comparing groups graphically.
- We've experienced the value of identifying and investigating outliers.
- We've graphed data that has been measured over time against a time axis and looked for long-term trends both by eye and with a data smoother.

Collect data on the number of states each student has visited. Make a 5-number summary, boxplot, check for outliers, and compare the mean and median.

Collect data on the number of siblings each student has. Make a 5-number summary, boxplot, check for outliers, and compare the mean and the median.