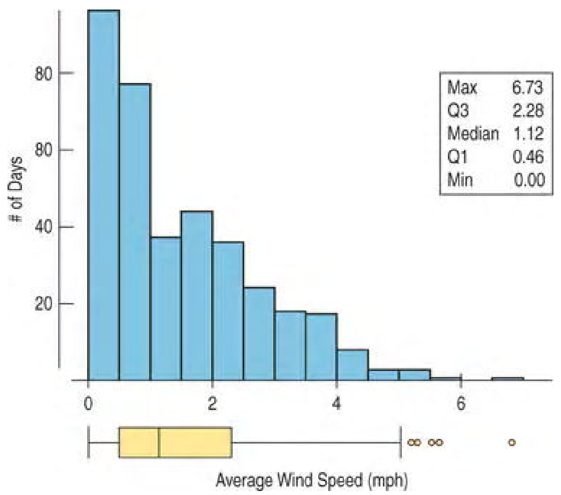
Chapter 4 – Understanding and Comparing Distributions

In Chapter 3 we learned about graphical representations and numerical summaries for Quantitative Variables. We also learned to describe the distribution of a data set in terms of the Shape, Center, and Spread.

If the distribution was symmetric we used the Mean for the Center and the Standard Deviation for the Spread

If the distribution was skewed or had outliers we used the Median for the Center and the IQR for the Spread.

Histogram and Boxplot for daily wind speeds



Skewed to the Right

unimodal

center 2-3

range 0-7

How does each represent distribution?

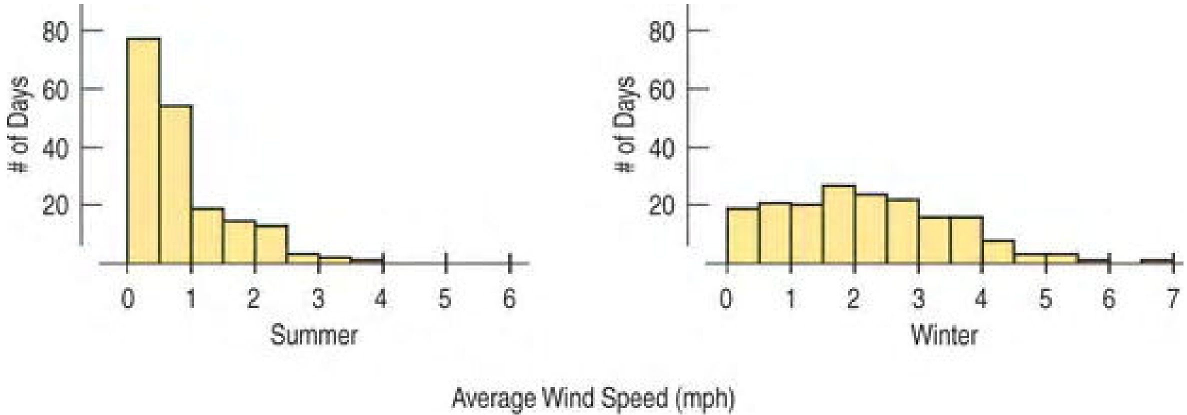
Histogram is easier to see peaks + mean

Boxplot is easier to See median.

Some outliers in box plot are not really outliers

4.1 Comparing Groups with Histograms

Axis should be similar. Note the shapes, centers, and spreads of the distributions.



skewed to the Right
unimodal
Center around 2 mph
spread is 0 – 4 mph

Less skewed
almost uniform.
center around 3mph
spread is 0-7 mph

Back to Back Stem and Leaf Plot - Stem plot used to compare two different data sets by putting one set of leaves to the left of the stems and the other set to the right of the stems.

South and West		North and Midwest
5778	8	8
12344	9	03
6667778899	9	67
02334	10	012233334
56	10	6779
	11	122444

5|8|8
85% for S&W
88% for N&MW

unimodal
Close to symmetric
center in upper 90%
spread 85% to 106%

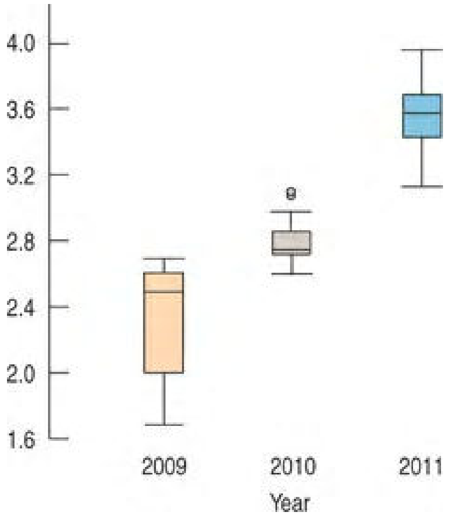
unimodal
skewed to the left
center in the Lower 100%
spread 88% to 114%

4.2 Comparing Groups with Boxplots

Boxplots are useful when comparing different groups of data. Use same axis and plot boxplots for different groups side by side or above each other.

Examples, p. 104:

26. Gas prices 2011 Here are boxplots of weekly gas prices for regular gas in the United States as reported by the U.S. Energy Information Administration for 2009, 2010, and 2011.



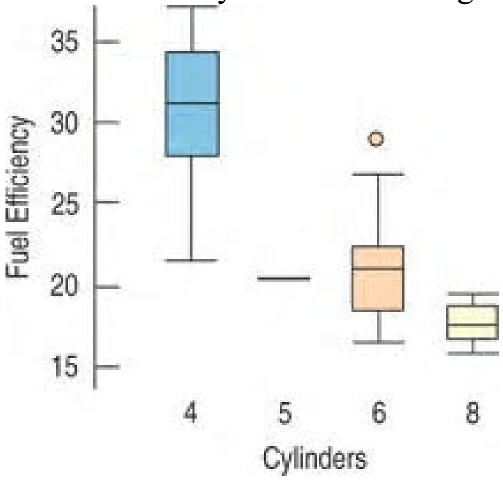
a) Compare the distribution of prices over the three years.

Average price is increasing per year.
2009 Skewed to the left
2010 slightly skewed to the right
2011 symmetric

b) In which year were the prices least stable? Explain.

2009 since it has the larger spread.

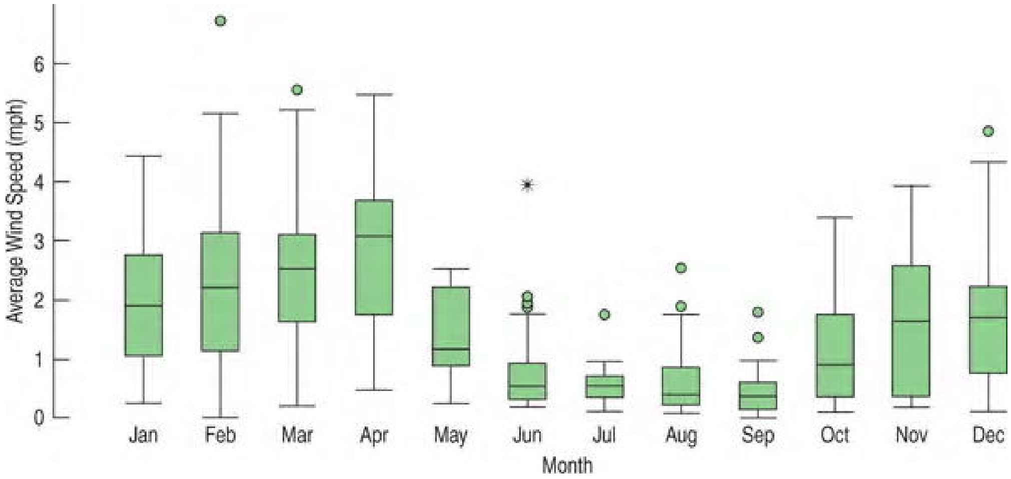
28. Fuel economy Describe what these boxplots tell you about the relationship between the number of cylinders a car's engine has and the car's fuel economy (mpg).



As Number of cylinders increase fuel economy decreases.

Spread is smaller as Number of Cylinders increases.

4.3 Outliers p.85.



Is an outlier an error? May be able to correct.

How extreme is an outlier? Outlier may not be as extreme in another context.

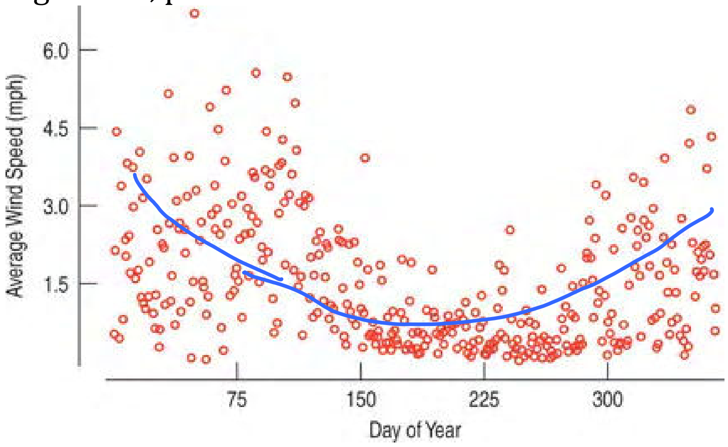
Outliers should not be ignored and treated as normal.

If an outlier is omitted from the data it should be mentioned and justified.

4.4 Timeplots

Timeplots are graphs that plot data values versus time.   Used in Stock Market values, Stock prices, Unemployment Rates, Temperature, Global Temperatures, etc.

Figure 4.4, p.90



One should resist using timeplots to predict for the future unless there is strong reasons for doing otherwise. (Path a ball follows when thrown from a certain height at a given speed, Seasonal Temperatures, Interest Rates?)