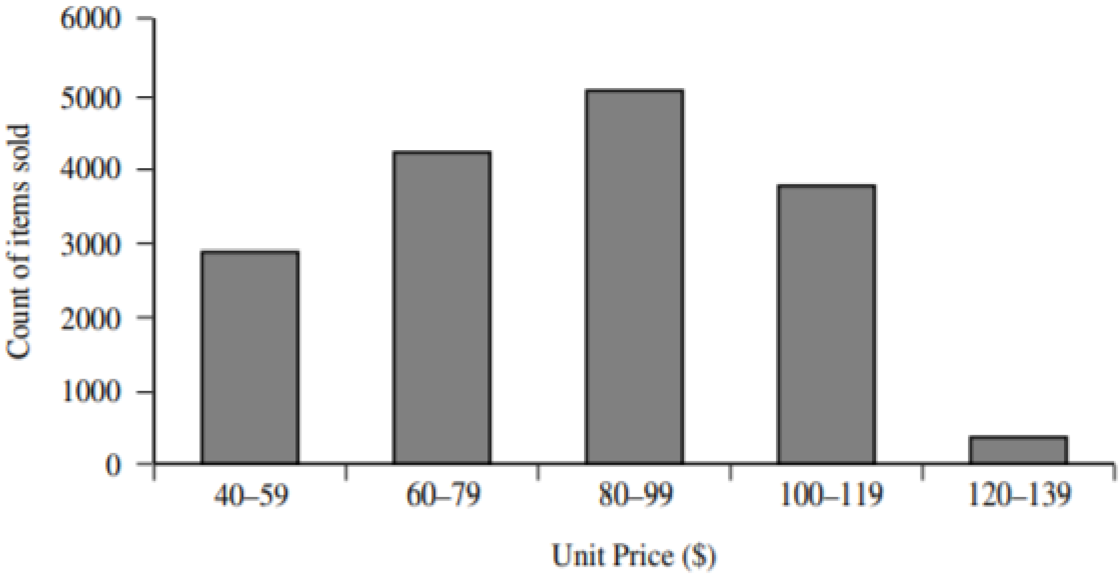
Graphic Displays of Basic Descriptive Data Summaries

1. Histograms
2. Quantile Plot (q plot)
3. Quantile-Quantile Plot (q-q plot)
4. Scatter Plot
5. Loess Curve

Graphic Displays of Basic Descriptive Data Summaries

1. Histograms

“Plotting histograms, or frequency histograms, is a graphical method for summarizing the distribution of a given attribute. A histogram for an attribute A partitions the data distribution of A into disjoint subsets, or buckets”

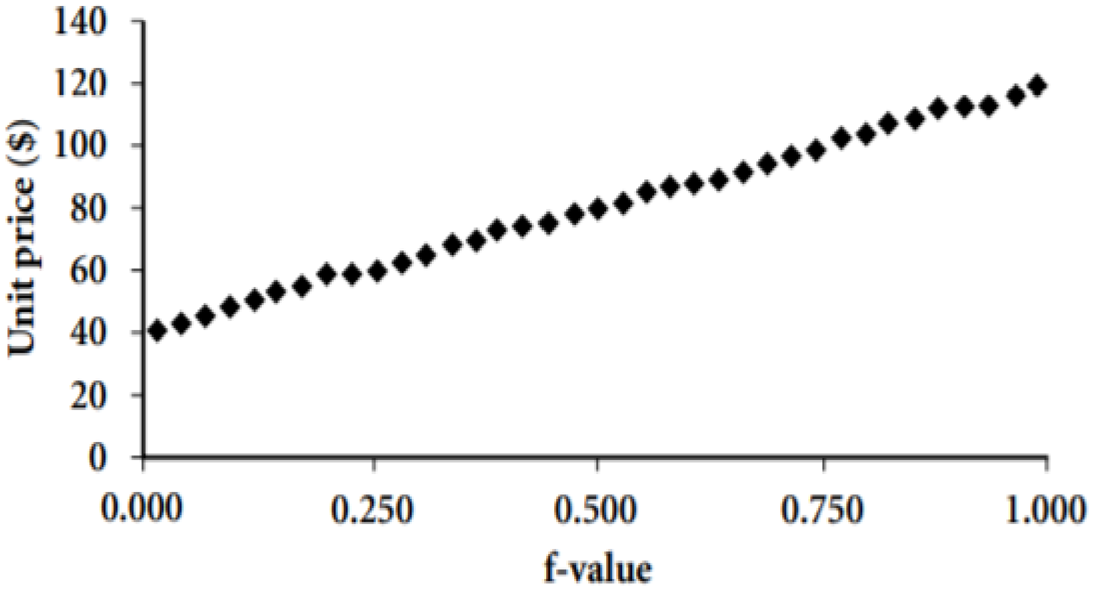


Graphic Displays of Basic Descriptive Data Summaries

Quantile Plot (q plot)

“A quantile plot is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute (allowing the user to assess both the overall behavior and unusual occurrences). Second, it plots quantile information.”

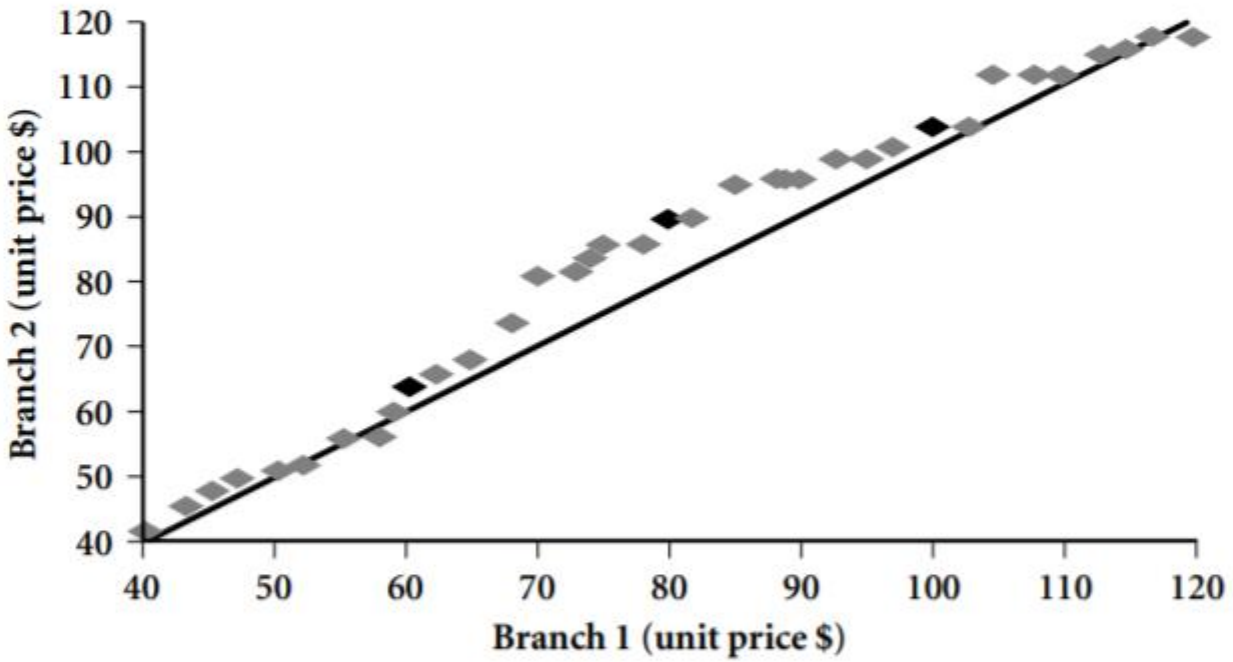
$$f_i = \frac{i - 0.5}{N}.$$



Graphic Displays of Basic Descriptive Data Summaries

Quantile-Quantile Plot (q-q plot)

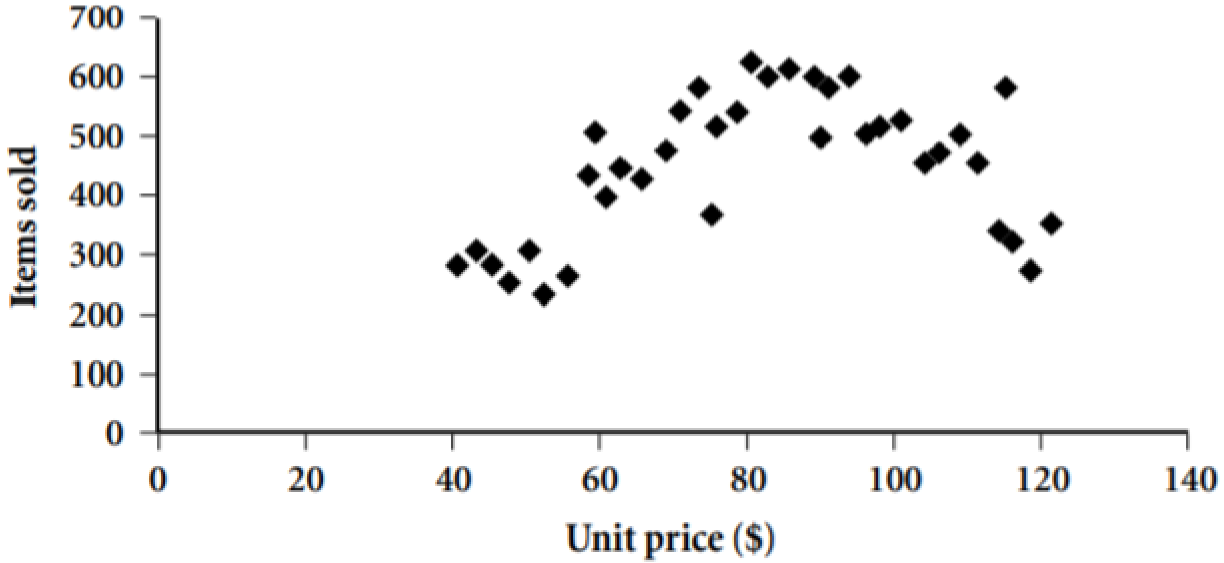
"A quantile-quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another"



Graphic Displays of Basic Descriptive Data Summaries

Scatter Plot

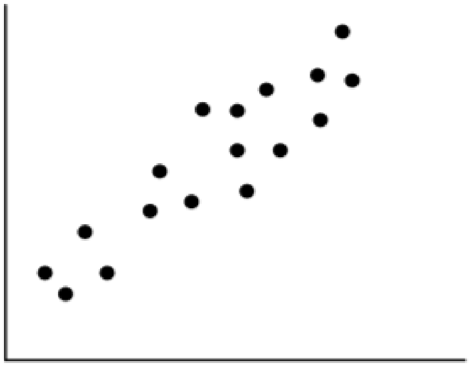
"A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numerical attributes."

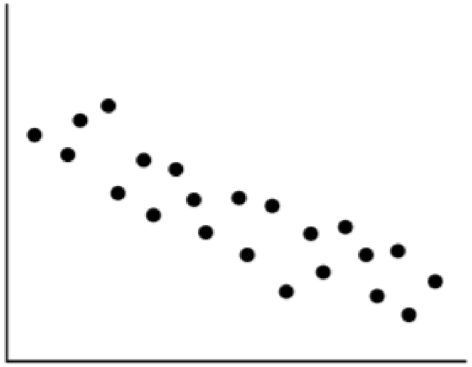


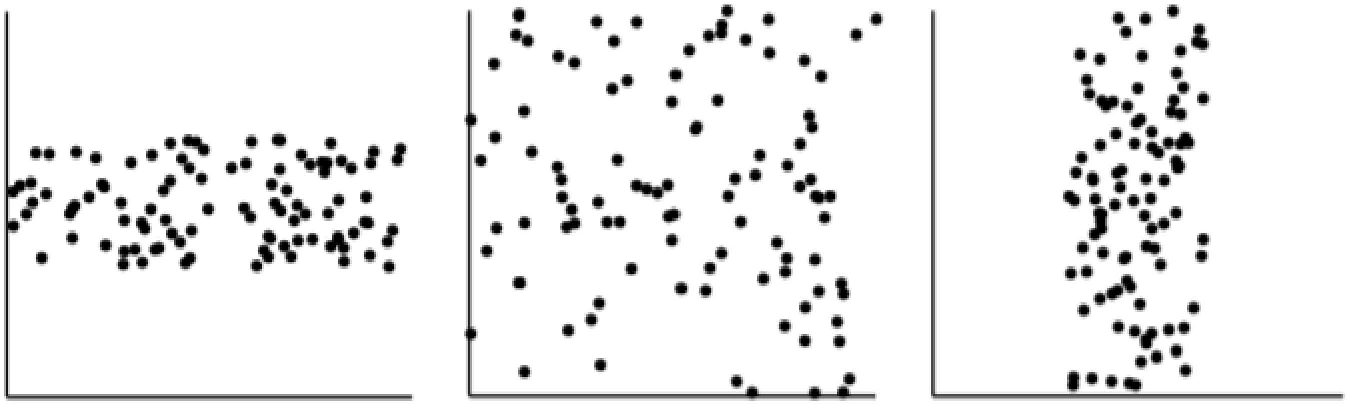
Graphic Displays of Basic Descriptive Data Summaries

Scatter Plot

- Positive Correlation
- Negative Correlation
- No Observed Correlation



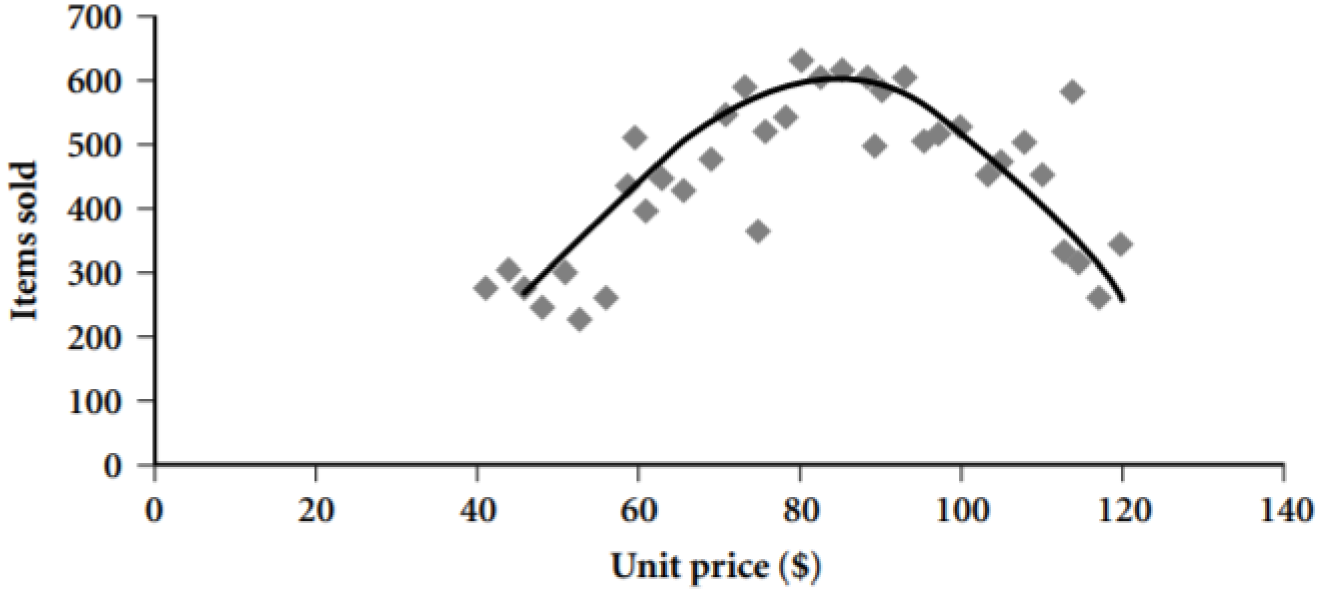




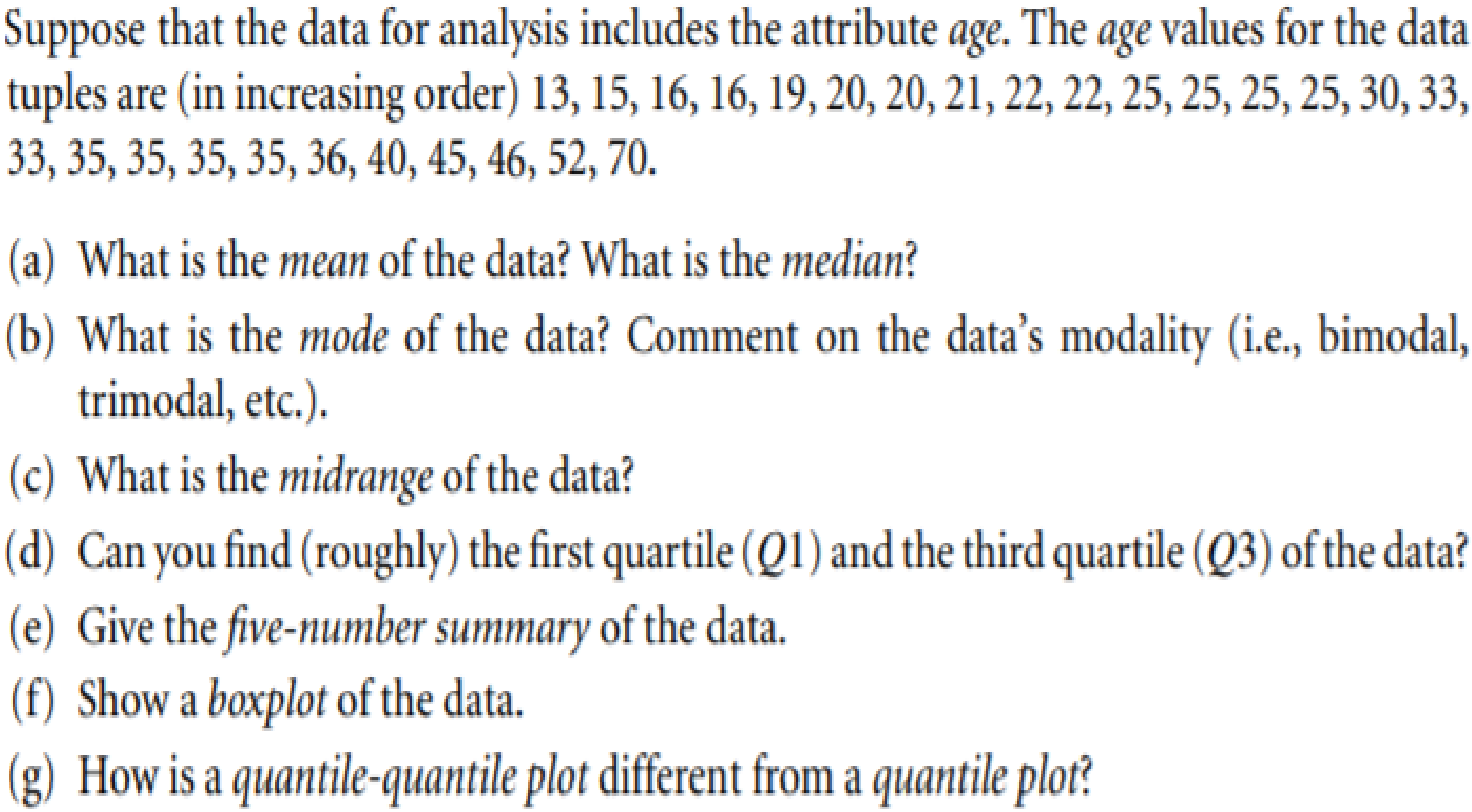
Graphic Displays of Basic Descriptive Data Summaries

Loess Curve

“adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence”



Questions



Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the *mean* of the data? What is the *median*?

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

(c) What is the *midrange* of the data?

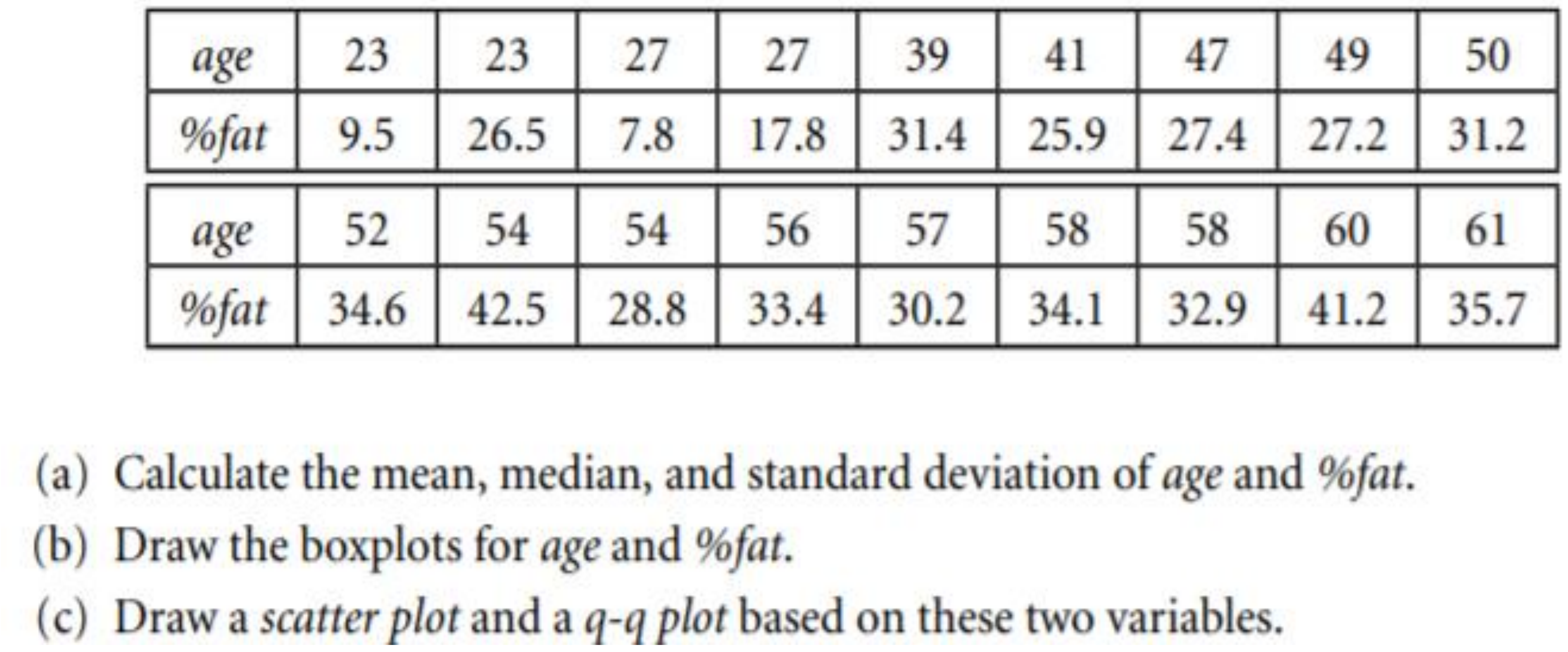
(d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

(e) Give the *five-number summary* of the data.

(f) Show a *boxplot* of the data.

(g) How is a *quantile-quantile plot* different from a *quantile plot*?

Questions



age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median, and standard deviation of *age* and *%fat*.

(b) Draw the boxplots for *age* and *%fat*.

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

Data Cleaning

To Fill in Missing Values

Smooth out Noise

Correct inconsistencies in the data

Missing Values

1. Ignore the tuple
2. Fill in the missing value manually
3. Use a global constant to fill in the missing value
4. Use the attribute mean to fill in the missing value
5. Use the attribute mean for all samples belong to the same class as the give tuple
6. Use the most probable value to fill in the missing value

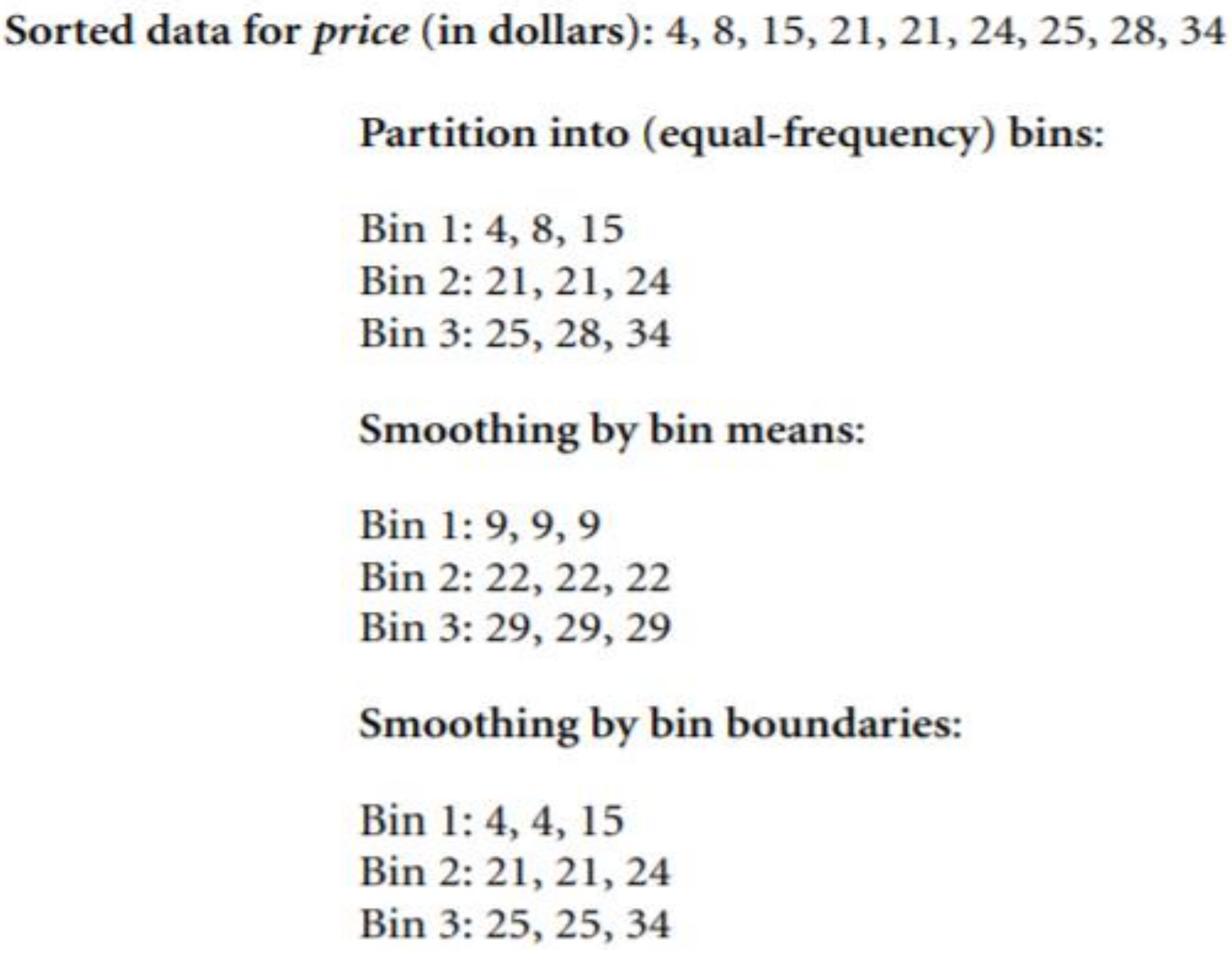
Noisy Data

“Noise is Random Error or Variance in a measured variable”

Binning Methods for data smoothing

Smoothing by bin mean
Smoothing by bin median
Smoothing by bin boundary

Noisy Data



Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34