Data Mining

Part 2. Data Understanding and Preparation

2.1 Data Understanding

Spring 2010

Instructor: Dr. Masoud Yaghini

Introduction

Outline

- Introduction
- Measuring the Central Tendency
- Measuring the Dispersion of Data
- Graphic Displays
- References

Introduction

- Data Understanding
  - To highlight which data values should be treated as noise or outliers.
- Measures
  - Central tendency
    - Mean, median, mode, and midrange
  - Data dispersion
    - Variance, Rang, quartiles, and interquartile range (IQR)

Introduction

- Such measures have been studied extensively in the statistical literature.
- From the data mining point of view, we need to examine how they can be computed efficiently in large databases.

Measuring the Central Tendency

Measuring the Central Tendency

- Measures of Central tendency:
  - Mean
  - Weighted mean
  - Trimmed mean
  - Median
  - Mode
  - Midrange

Mean

- Mean: The most common and most effective numerical measure of the "center" of a set of data is the (arithmetic) mean. (sample vs. population)

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Weighted (arithmetic) mean : Sometimes, each value in a set may be associated with a weight, the weights reflect the significance, importance, or occurrence frequency attached to their respective values.

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

Trimmed mean

- Disadvantage of mean
  - A major problem with the mean is its sensitivity to extreme (e.g., outlier) values.
  - Even a small number of extreme values can corrupt the mean.
- Trimmed mean
  - the trimmed mean is the mean obtained after cutting off values at the high and low extremes.
  - For example, we can sort the values and remove the top and bottom 2% before computing the mean.
  - We should avoid trimming too large a portion (such as 20%) at both ends as this can result in the loss of valuable information.

Median

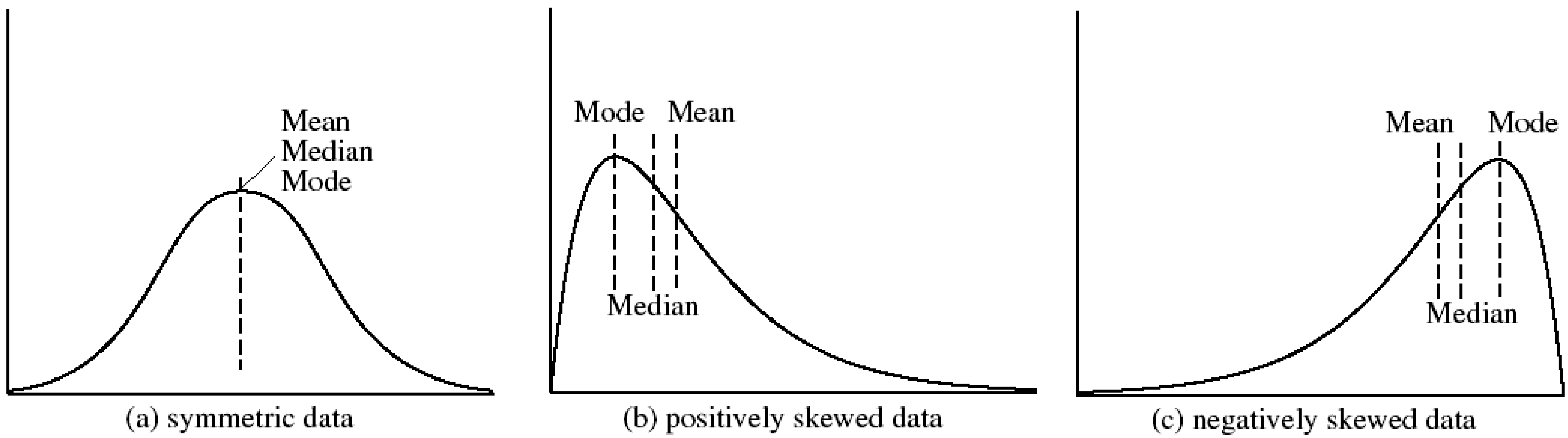
- Suppose that a given data set of *N* distinct values is sorted in numerical order.
- The median is the middle value if odd number of values, or average of the middle two values otherwise
- For skewed (asymmetric) data, a better measure of the center of data is the median.

Mode & Midrange

- Mode is the another measure of central tendency
  - The mode for a set of data is the value that occurs most frequently in the set.
  - If each data value occurs only once, then there is no mode.
- The midrange can also be used to assess the central tendency of a data set
  - It is the average of the largest and smallest values in the set.

Mean, Median, and Mode

- Mean, median, and mode of symmetric versus positively and negatively skewed data.



- Positively skewed, where the mode is smaller than the median (b), and negatively skewed, where the mode is greater than the median (c).

Measuring the Dispersion of Data

Measuring the Dispersion of Data

- The degree to which numerical data tend to spread is called the dispersion, or variance of the data.
- The measures of data dispersion:
  - Range
  - Five-number summary (based on quartiles)
  - Interquartile range (IQR)
  - Standard deviation
- Range
  - difference between highest and lowest observed values

Inter-Quartile Range

- For the remainder of this section, let's assume that the data are sorted in increasing numerical order.
- The ***kth* percentile** of a set of data in numerical order is the value x_i having the property that **k percent** of the data entries lie at or below x_i.
  - The median (discussed in the previous subsection) is the 50th percentile.
- Quartiles:
  - First quartile (Q_1): The first quartile is the value, where 25% of the values are smaller than Q_1 and 75% are larger.
  - Third quartile (Q_3): The third quartile is the value, where 75% of the values are smaller than Q_3 and 25% are larger.

Inter-Quartile Range

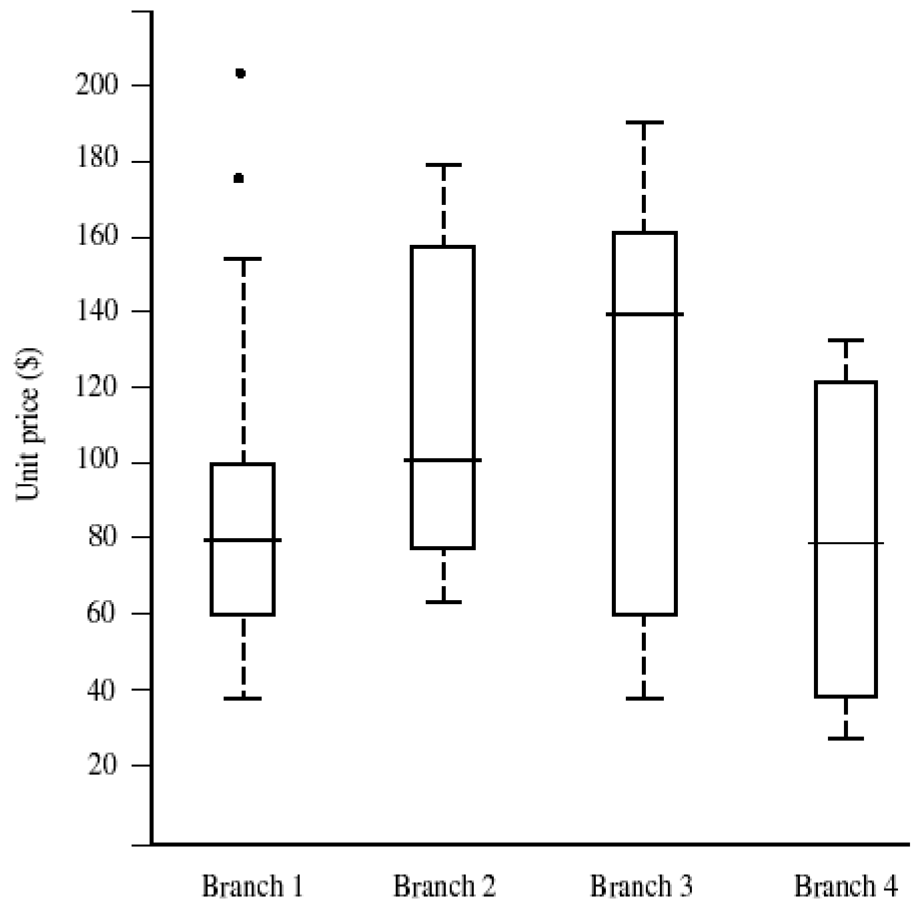
- **Inter-quartile range (IQR)**
  - IQR = Q3 – Q1
  - IQR is a simple measure of spread that gives the range covered by the middle half of the data
- **Outlier**
  - usually, values falling at least 1.5 * IQR, above the third quartile or below the first quartile.
- **Five number summary**
  - min, Q_1, Median, Q_3, max
  - Contain information about the endpoints (e.g., tails) of the data

Five Number Summary

- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum
  - To show outliers, the whiskers are extended to the extreme low and high observations only if these values are less than 1.5 * IQR beyond the quartiles.

Five Number Summary

- Boxplot for the unit price data for items sold at four branches of *AllElectronics during a given* time period.



Variance and Standard Deviation

- Variance (σ^2)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Standard deviation (σ)
  - is the square root of variance σ^2
  - σ measures spread about the mean and should be used only when the mean is chosen as the measure of center.
  - σ =0 only when there is no spread, that is, when all observations have the same value.

Graphic Displays

Graphic Displays

- There are many types of graphs for the display of data summaries and distributions, such as:
  - Bar charts
  - Pie charts
  - Line graphs
  - Boxplot
  - Histograms
  - Quantile plots
  - Scatter plots
  - Loess curves

Histogram Analysis

- Histograms or frequency histograms
  - A univariate graphical method
  - Consists of a set of **rectangles** that reflect the counts or frequencies of the classes present in the given data
  - If the attribute is categorical, then one rectangle is drawn for each known value of A, and the resulting graph is more commonly referred to as a bar chart.
  - If the attribute is numeric, the term histogram is preferred.

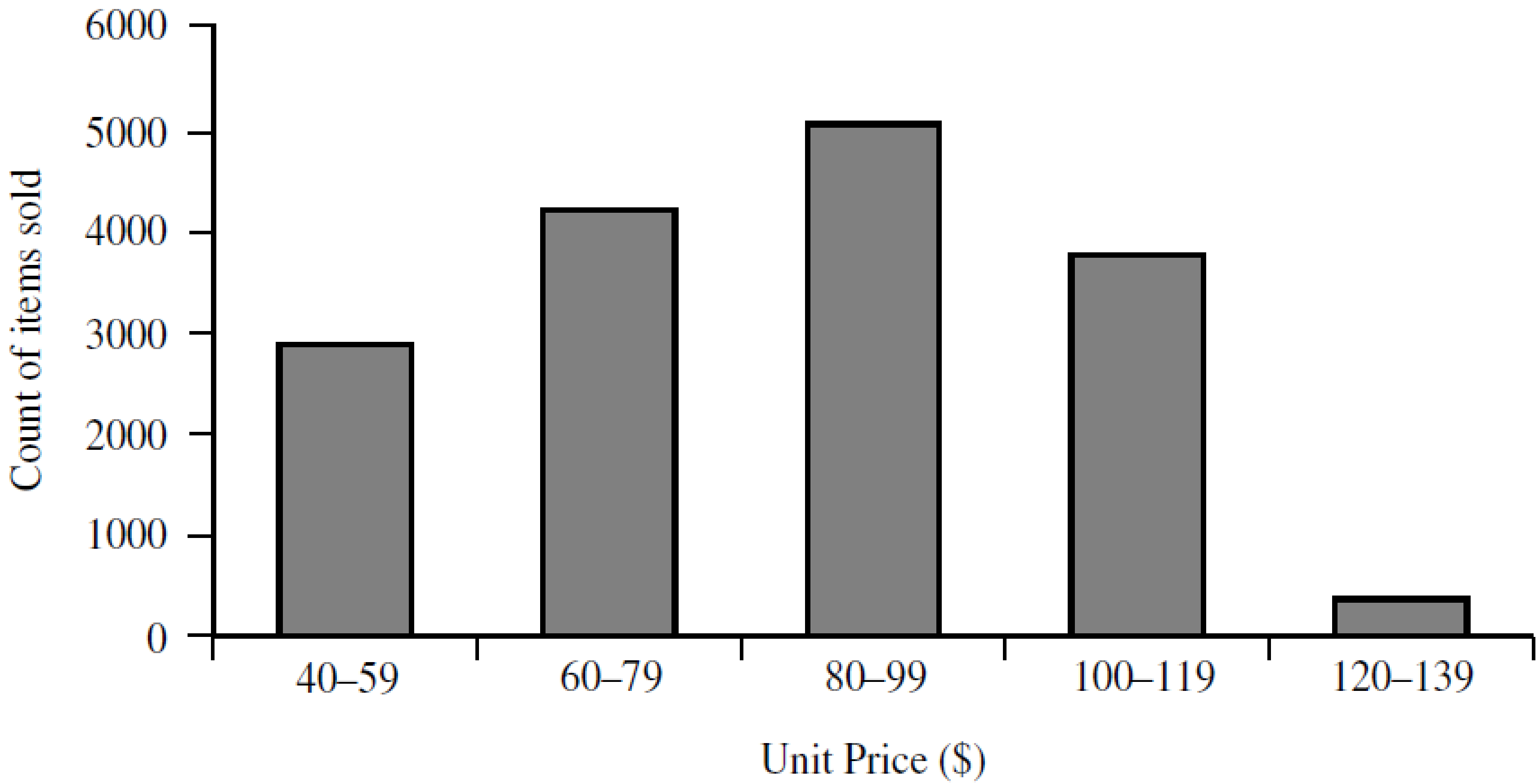
Histogram Analysis

- **Example:** A set of unit price data for items sold at a branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350

Histogram Analysis

- **Example: A histogram**

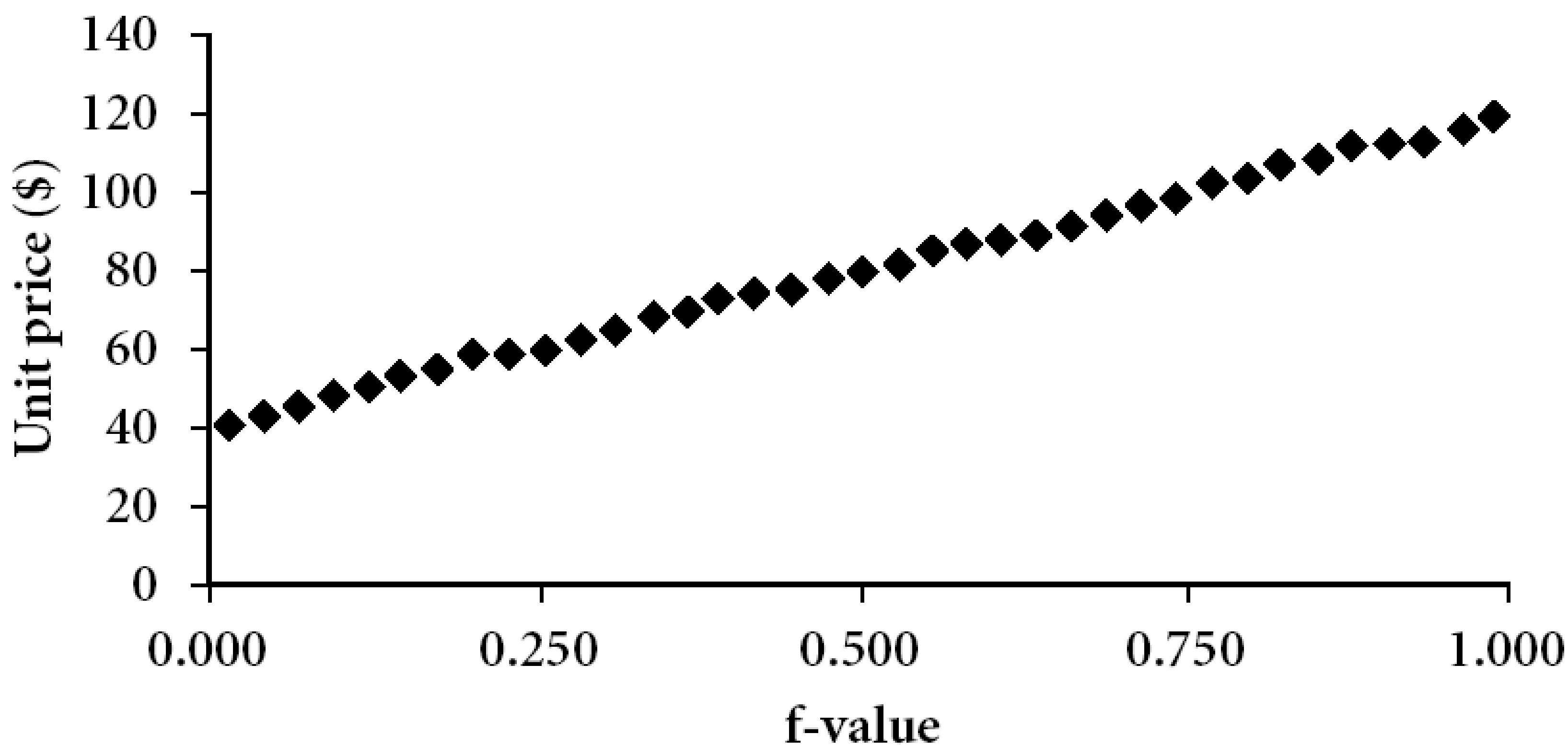


Quantile Plot

- A quantile plot is a simple and effective way to have a first look at a **univariate** data distribution.
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i
- Note that
  - the 0.25 quantile corresponds to quartile Q1,
  - the 0.50 quantile is the median, and
  - the 0.75 quantile is Q3.

Quantile Plot

- A quantile plot for the unit price data of AllElectronics.

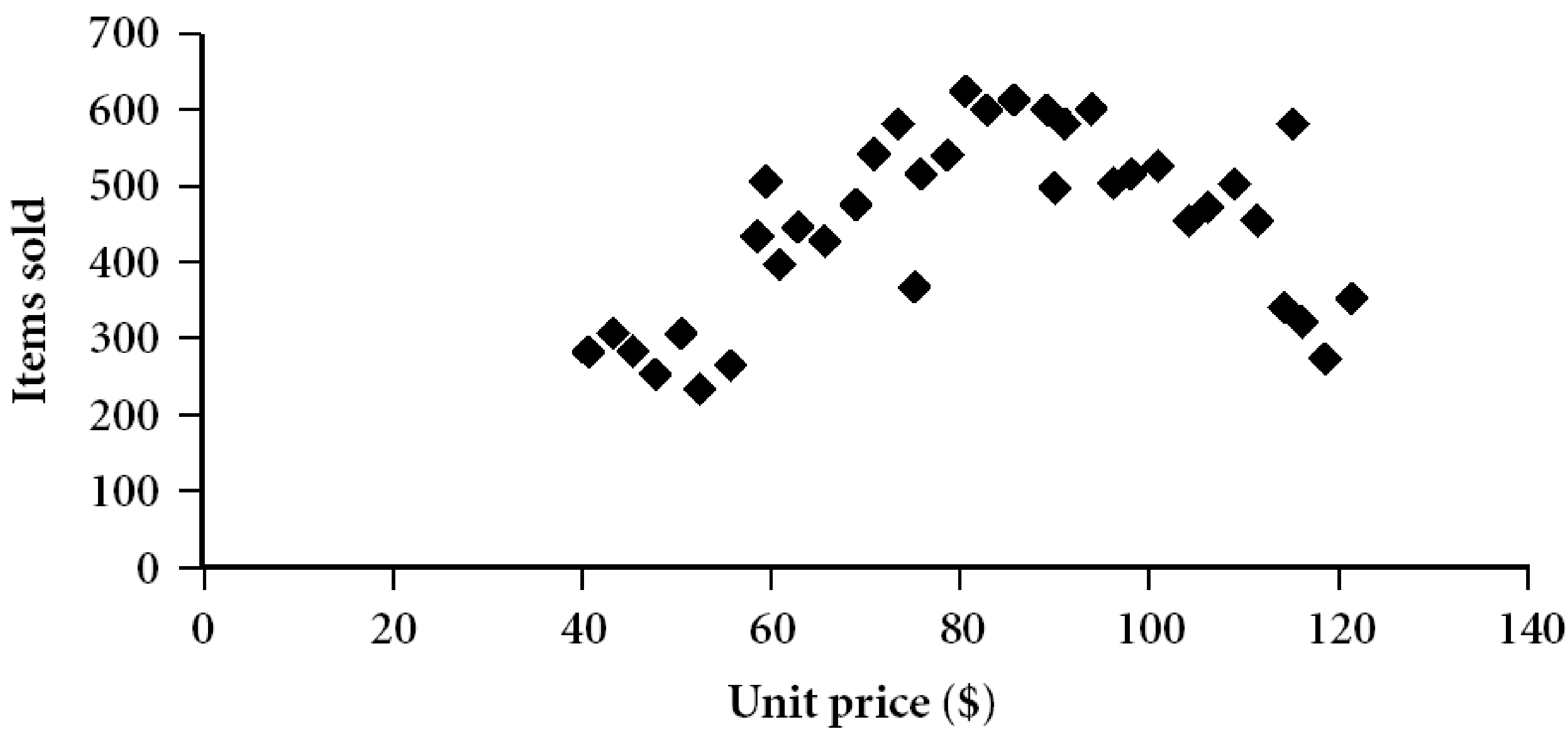


Scatter plot

- Scatter plot
  - is one of the most effective graphical methods for determining if there appears to be a **relationship**, **clusters of points**, or **outliers** between two numerical attributes.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

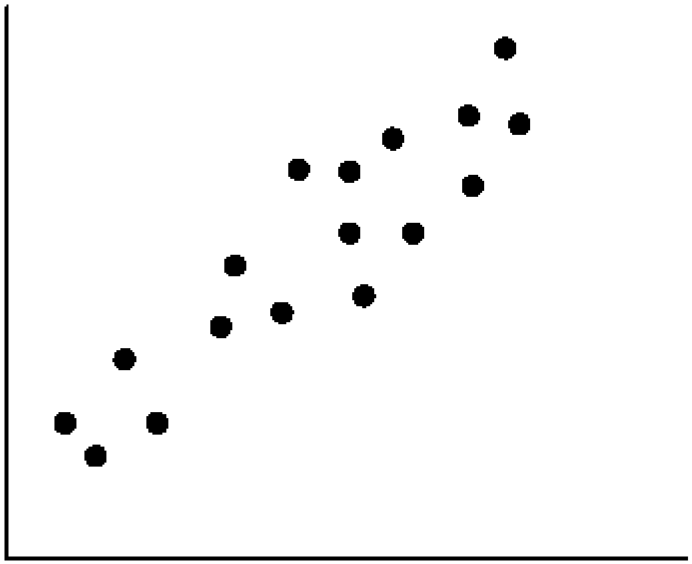
Scatter plot

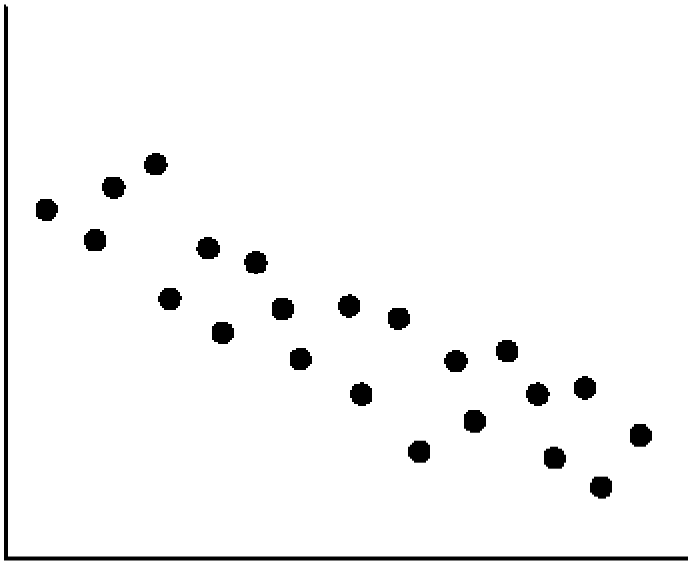
- A scatter plot for the data set of AllElectronics.



Scatter plot

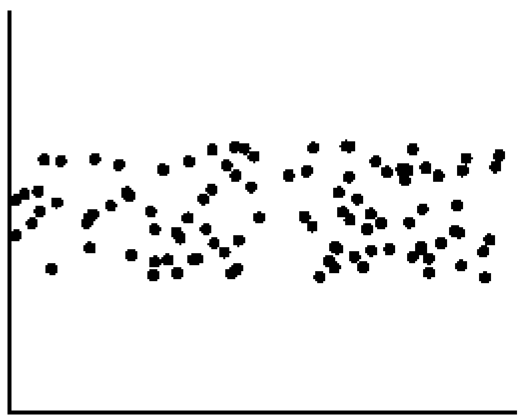
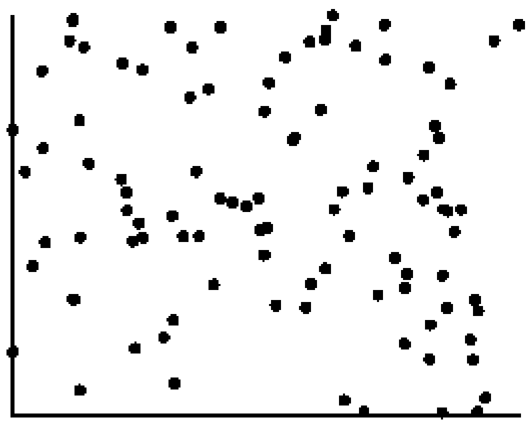
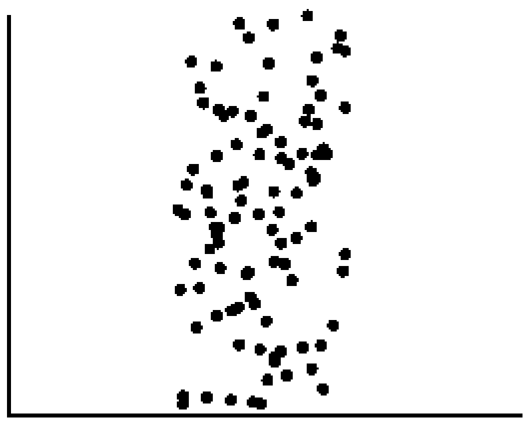
- Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.





Scatter plot

- Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

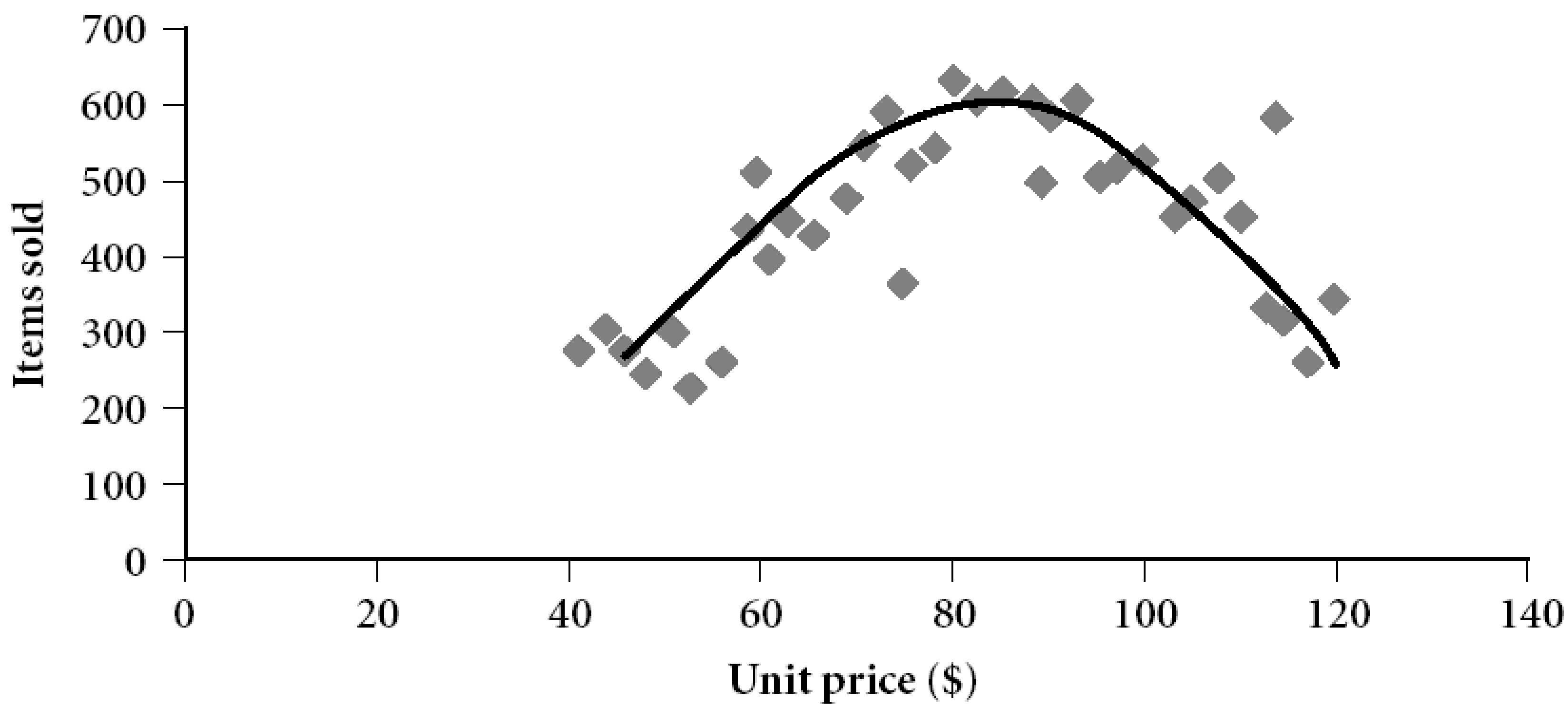




Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- The word loess is short for **local regression**.
- Loess curve is fitted by setting two parameters:
  - a smoothing parameter, and
  - the degree of the polynomials that are fitted by the regression

Loess Curve

- A loess curve for the data set of *AllElectronics*



References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 2)

The end