BUSINESS INTELLIGENCE

LECTURE 2

OUTLINE

- Introduction
- Data mining overview
- **Data pre-processing**
- Classification techniques
- Clustering techniques
- Association rules
- References

DATA PRE-PROCESSING

WHY?

- Data quality
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Interpretability

CENTRAL TENDENCY MEASURING

- Mean
- Median
- Mode
- Midrange

MEAN

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}. \tag{2.1}$$

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$
$$= \frac{696}{12} = 58.$$

Thus, the mean salary is $58,000. ■

[2]

WEIGHTED MEAN

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}. \quad (2.2)$$

This is called the **weighted arithmetic mean** or the **weighted average.**

A major problem with the mean is its sensitivity to extreme (e.g., outlier) values.

[2]

MEDIAN

- It is the middle value in a set of ordered data values.

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is $58,000. ■

What is its median?

[2]

MODE

- The mode for a set of data is the value that occurs most frequently in the set.

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000. ■

What is its mode?

[2]

MIDRANGE

- It is the average of the largest and smallest values in the set.

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000. ■

What is its midrange?

[2]

SKEWED DATA

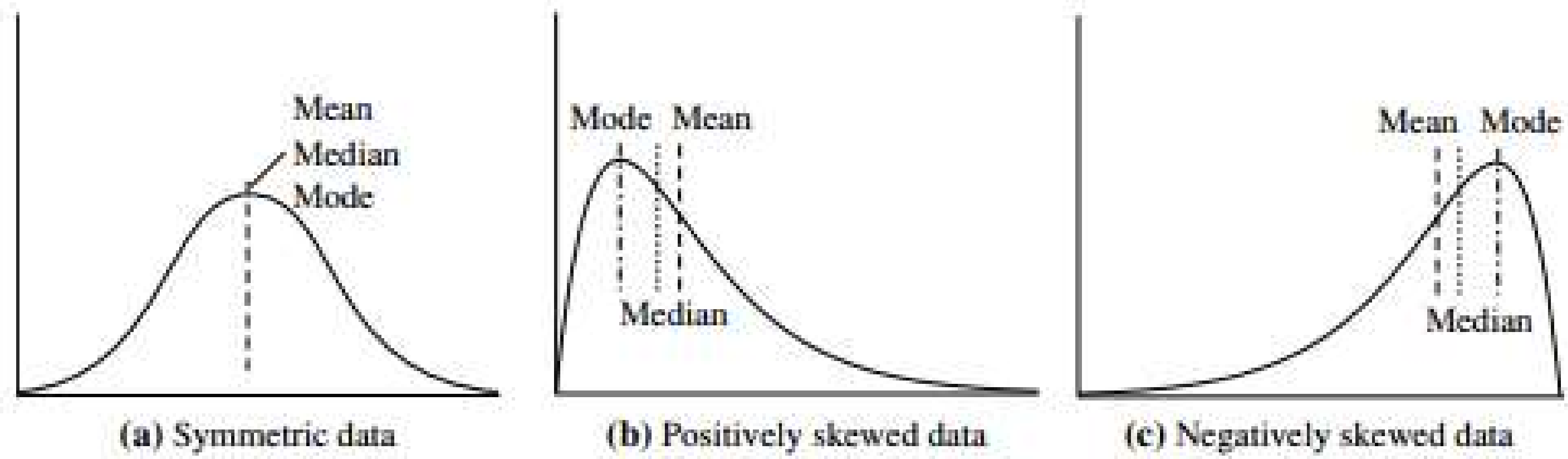
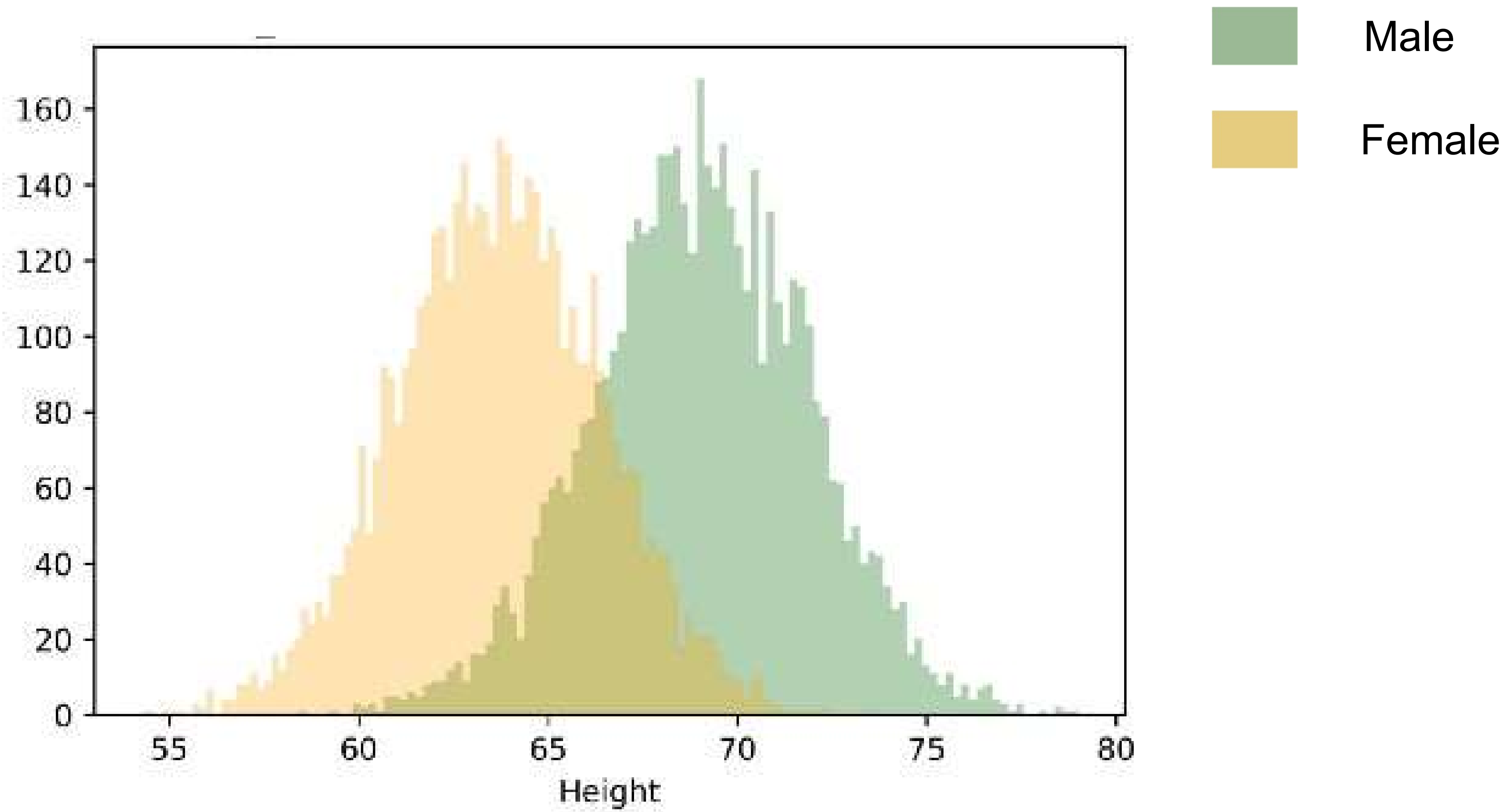


Figure 2.1 Mean, median, and mode of symmetric versus positively and negatively skewed data.

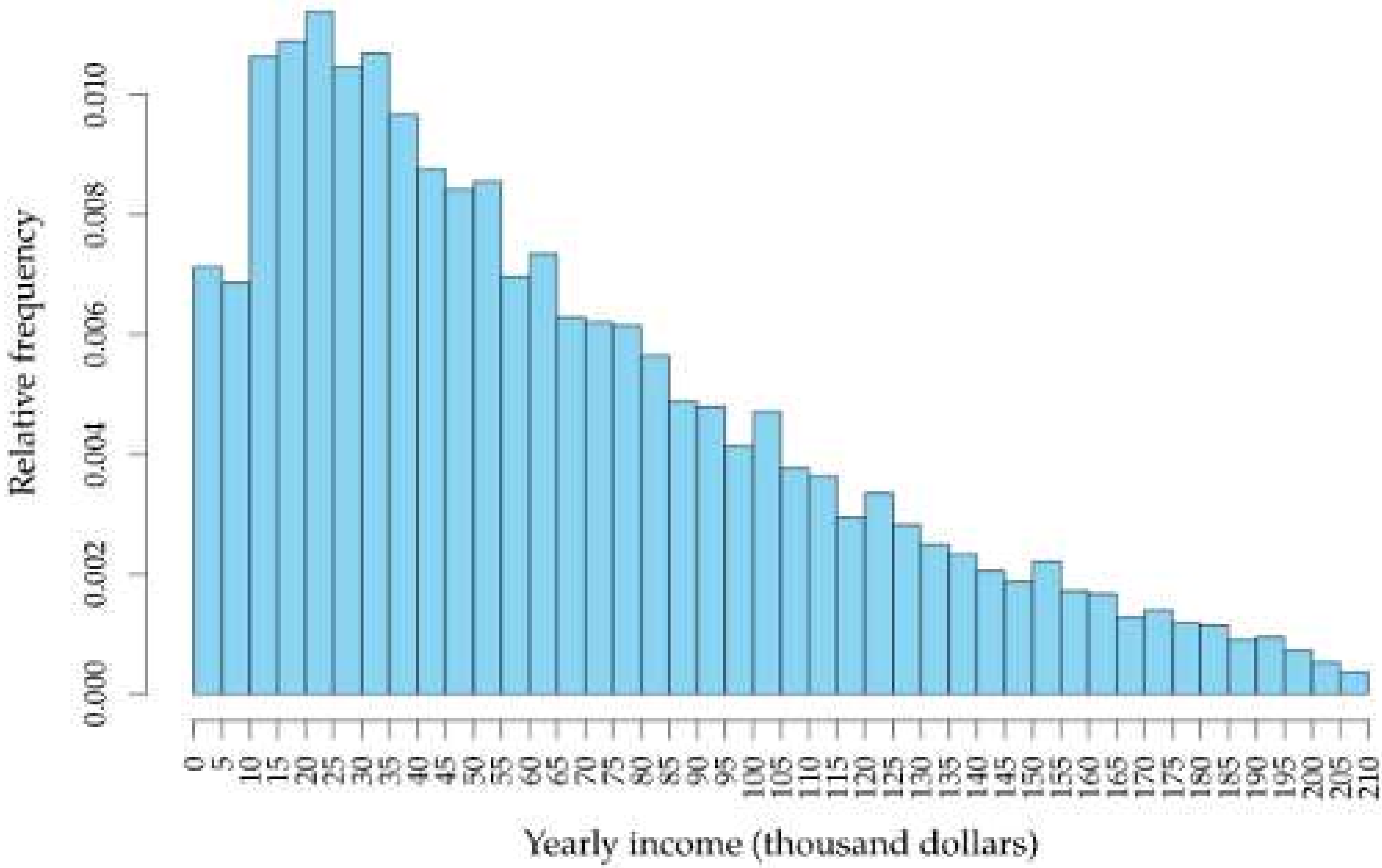
[2]

FOR INSTANCE



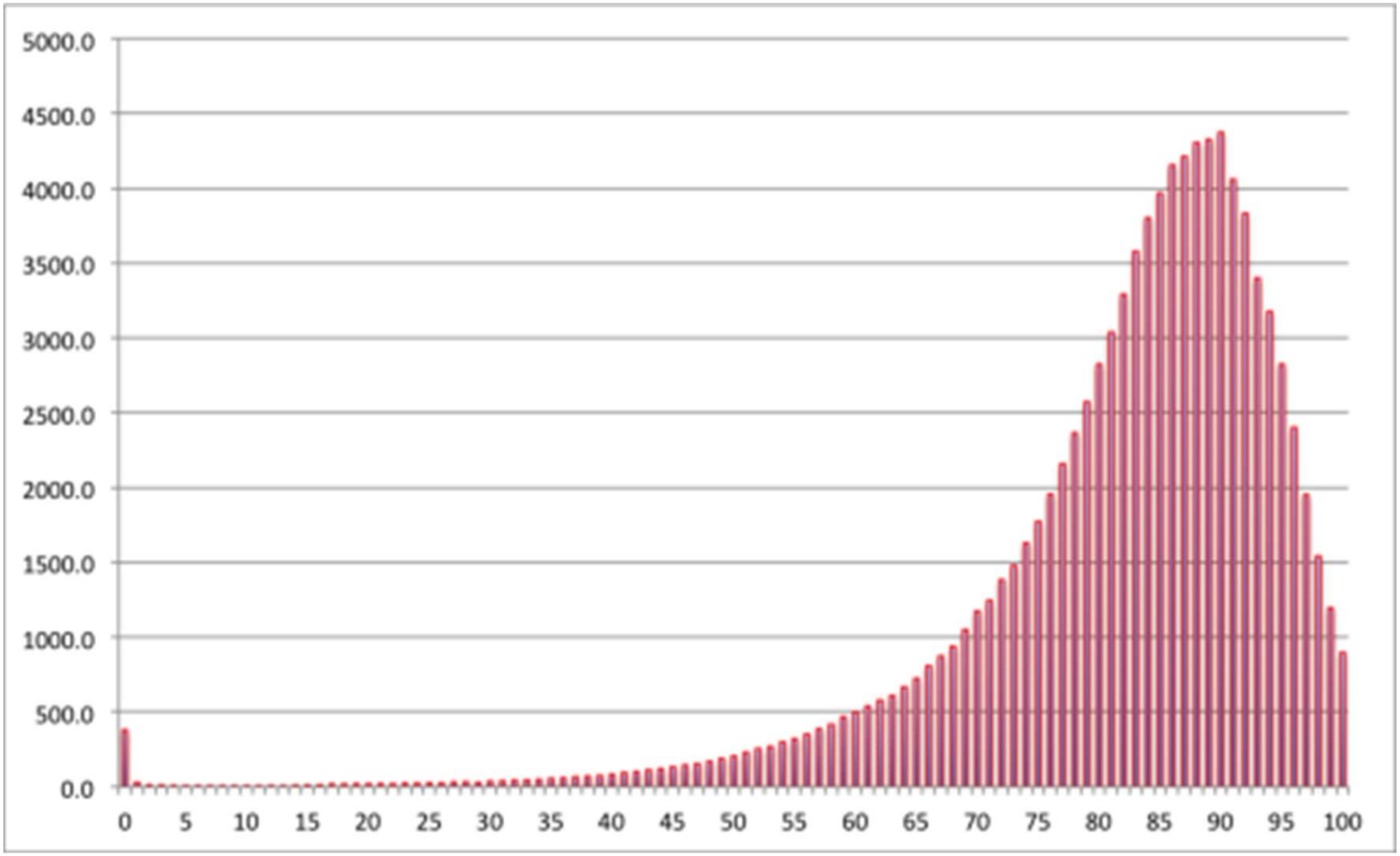
https://builtin.com/sites/www.builtin.com/files/styles/ckeditor_optimize/public/inline-images/national/1_skewed%2520data.png

FOR INSTANCE



https://builtin.com/sites/www.builtin.com/files/styles/ckeditor_optimize/public/inline-images/national/4_skewed%2520data.png

FOR INSTANCE

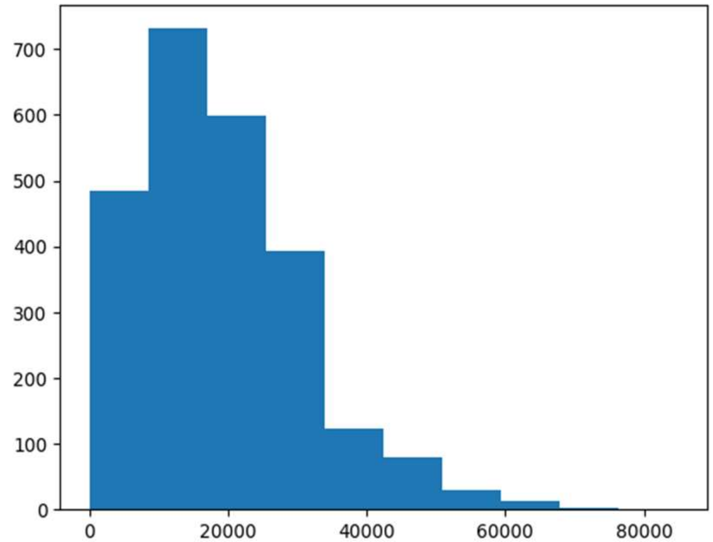


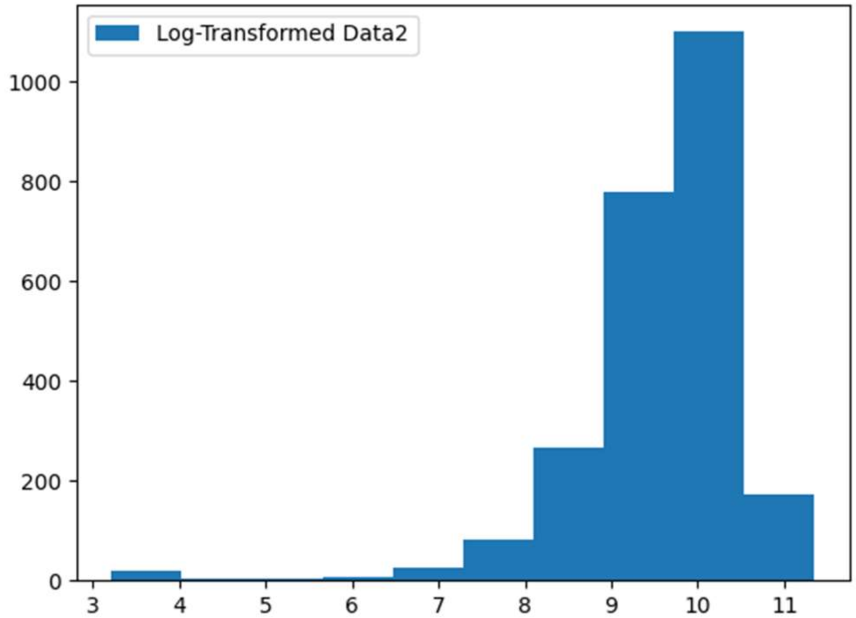
Mortality age

https://builtin.com/sites/www.builtin.com/files/styles/ckeditor_optimize/public/inline-images/national/7_skewed%2520data.png

SKEWED DATA NORMALIZATION

- Square Root Transformation
- Log Transformation
- Box-Cox Transformation
- Etc.





DATA DISPERSION

- Range
- Quartiles
- Interquartile range
- Five-number summary
- Boxplots
- Variance
- Standard deviation

RANGE

- It is the difference between the largest (max) and smallest (min) values.

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000. ■

What is its range?

[2]

QUARTILES

- They are points taken at regular intervals of a data distribution, dividing it into essentially equalsize consecutive sets.

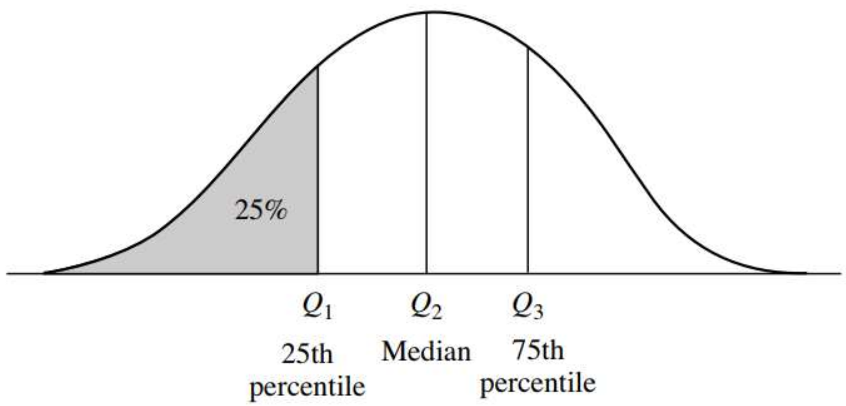


Figure 2.2 A plot of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

[2]

INTERQUARTILE RANGE

- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR)

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is $58,000. ■

What are its quartiles and interquartile range?

[2]

FIVE-NUMBER SUMMARY

- The five-number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of Minimum, Q1, Median, Q3, Maximum.

[2]

BOXPLOTS

- Boxplots are a popular way of visualizing a distribution.

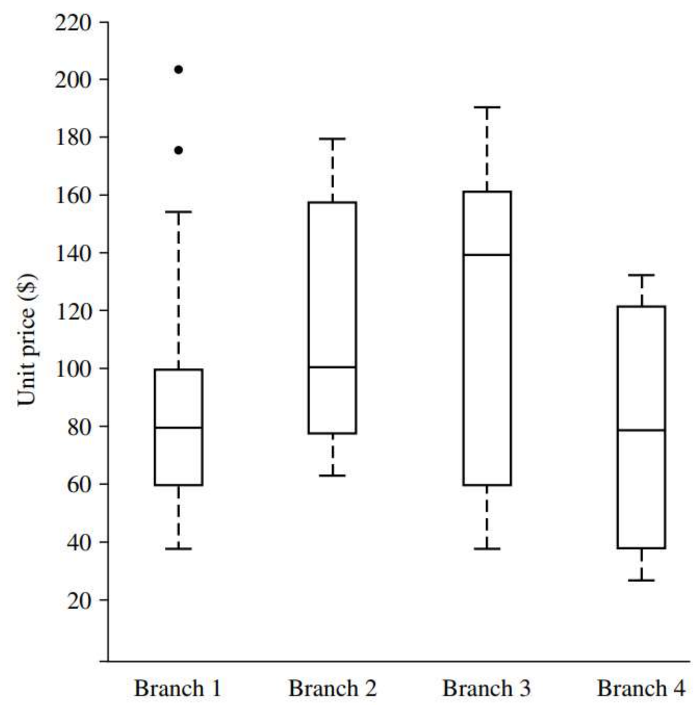


Figure 2.3 Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

[2]

VARIANCE AND STANDARD DEVIATION (1/2)

- A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2, \tag{2.6}$$

where $\bar{x}$ is the mean value of the observations, as defined in Eq. (2.1). The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

[2]

VARIANCE AND STANDARD DEVIATION (2/2)

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000. ■

What are its variance and standard deviation?

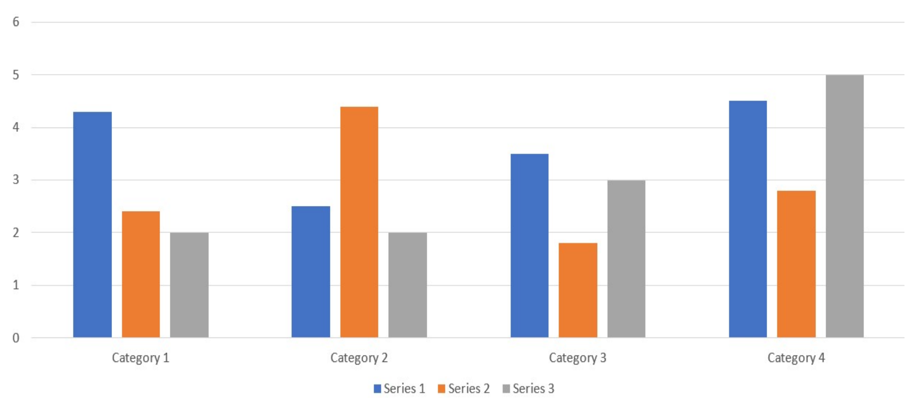
[2]

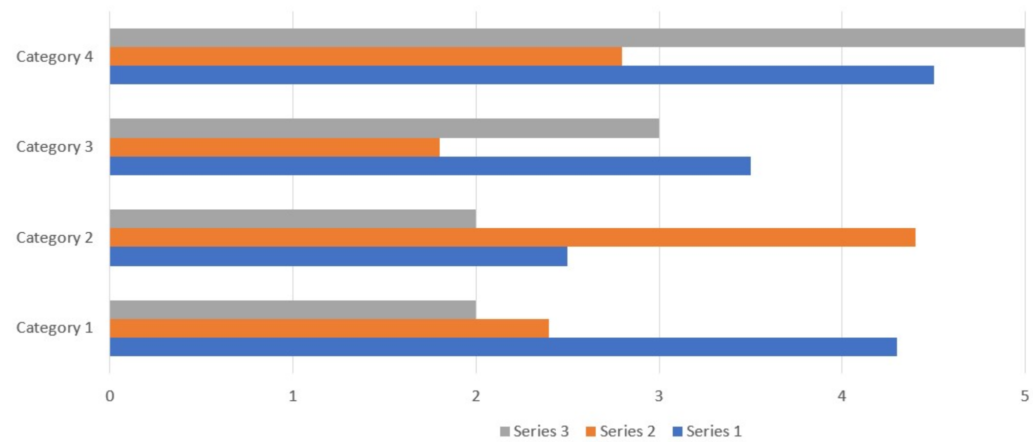
GRAPHIC DISPLAYS

- Bar charts
- Pie charts
- Line charts
- Quantile plots
- Quantile-quantile plots
- Histogram
- Scatter plots

BAR CHARTS

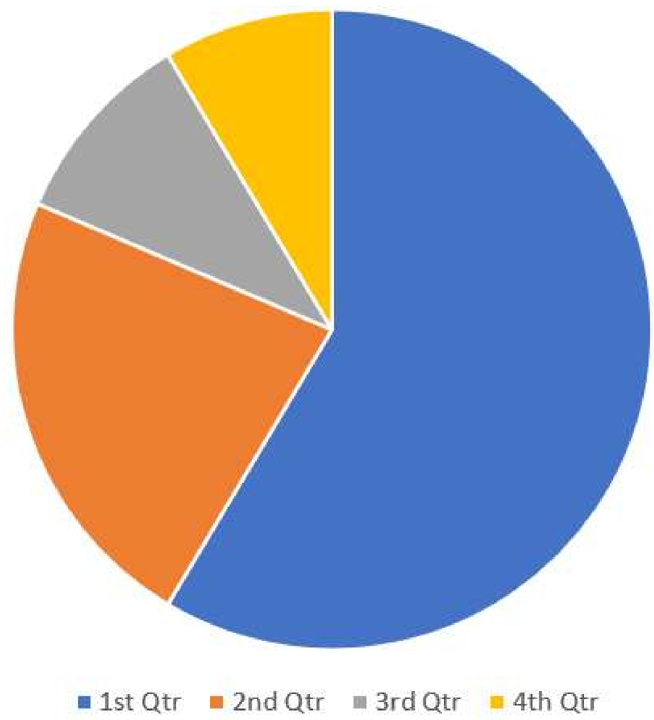
- A bar is a graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.
- The bars can be plotted vertically or horizontally.





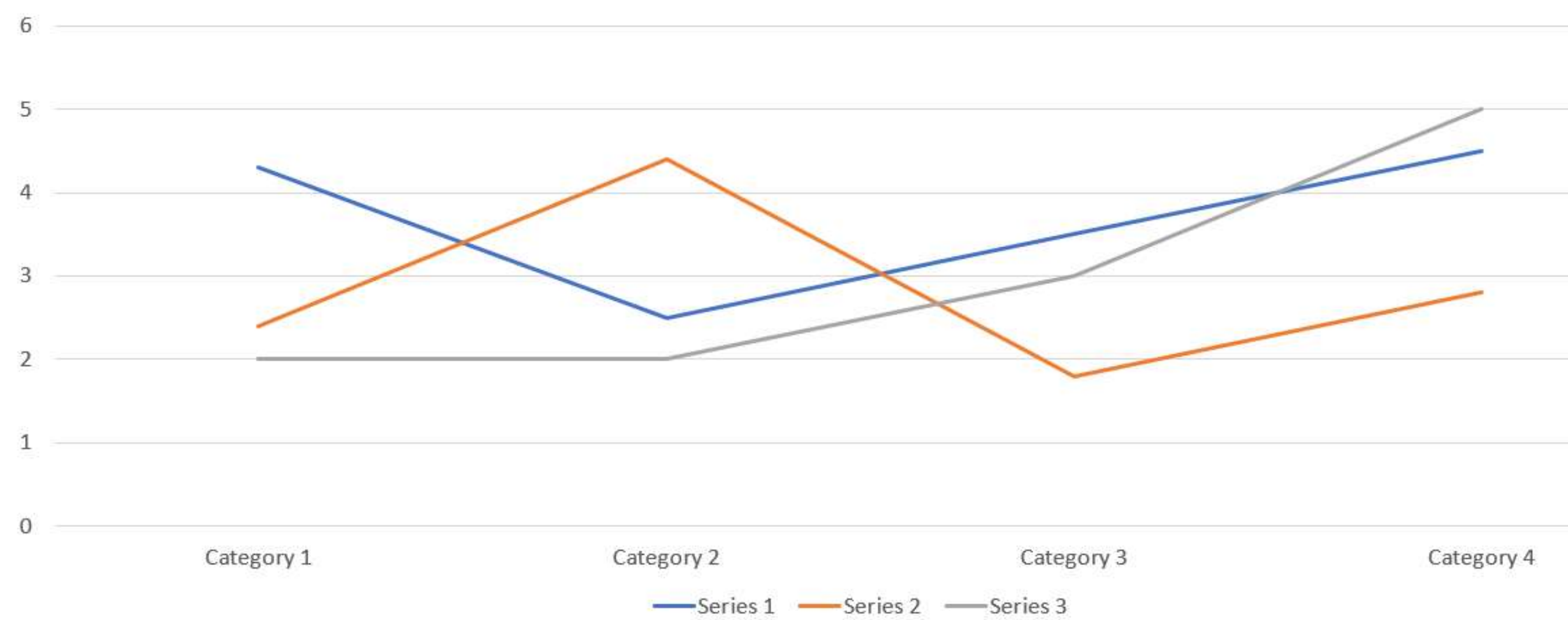
PIE CHARTS

- A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.
- In a pie chart, the arc length of each slice is proportional to the quantity it represents.



LINE CHARTS

- A line chart displays information as a series of data points connected by straight line segments.



QUANTILE PLOTS

- A quantile plot is a simple and effective way to have a first look at a univariate data distribution.

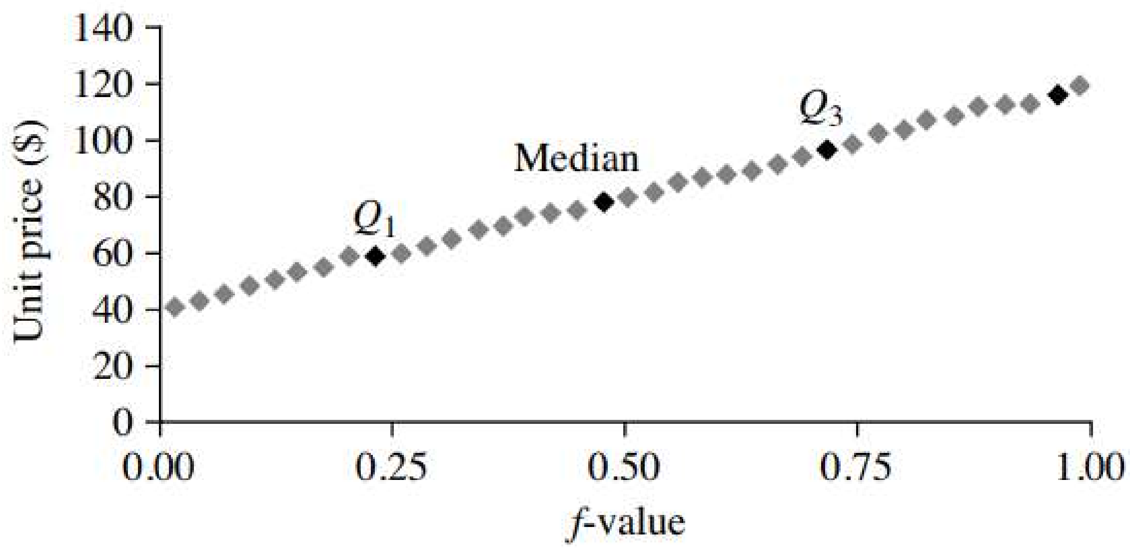
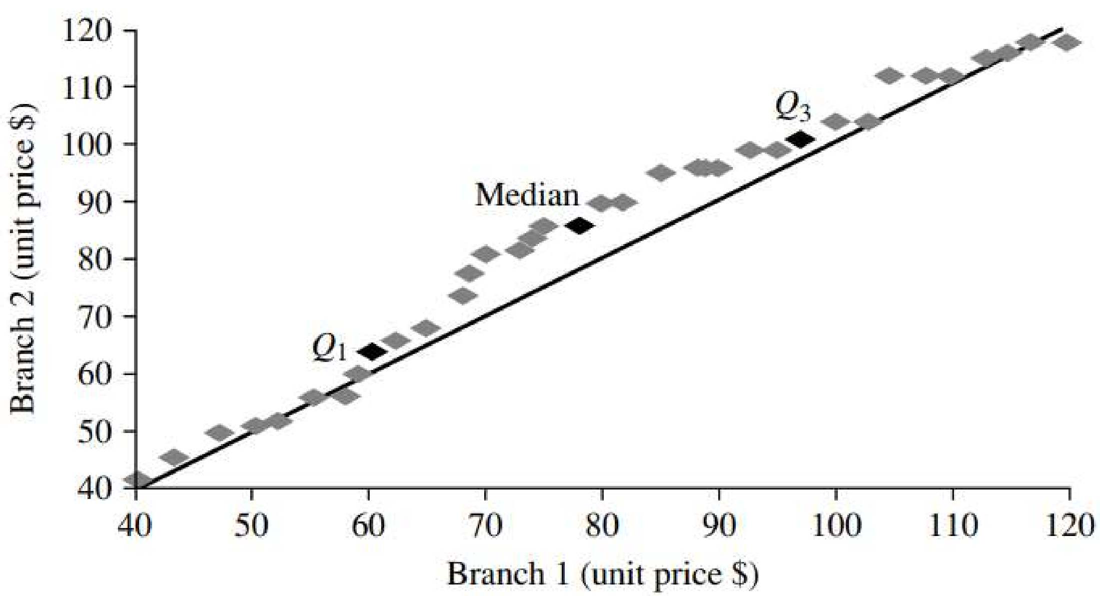


Figure 2.4 A quantile plot for the unit price data of Table 2.1.

[2]

QUANTILE-QUANTILE PLOTS

- 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.



[2]

Figure 2.5 A q-q plot for unit price data from two *AllElectronics* branches.

HISTOGRAM

- Plotting histograms is a graphical method for summarizing the distribution of a given attribute.
- The height of the bar indicates the frequency (i.e., count) of that X value.

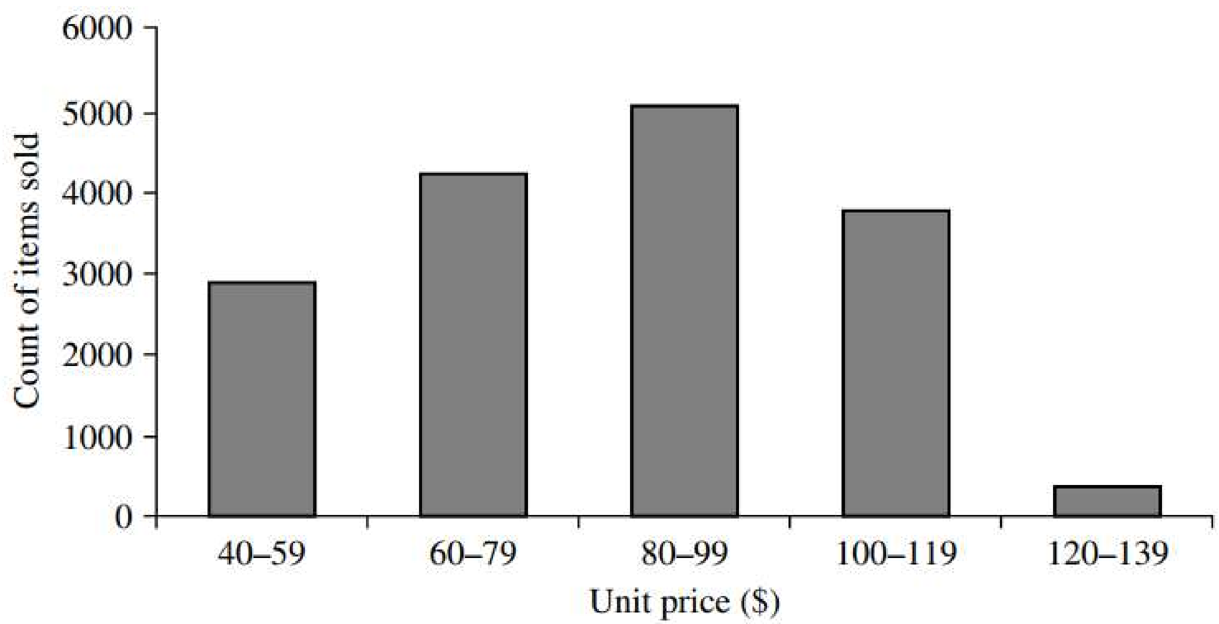


Figure 2.6 A histogram for the Table 2.1 data set.

[2]

SCATTER PLOTS

- A scatter plot determines if there appears to be a relationship, pattern, or trend between two numeric attributes.
- To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

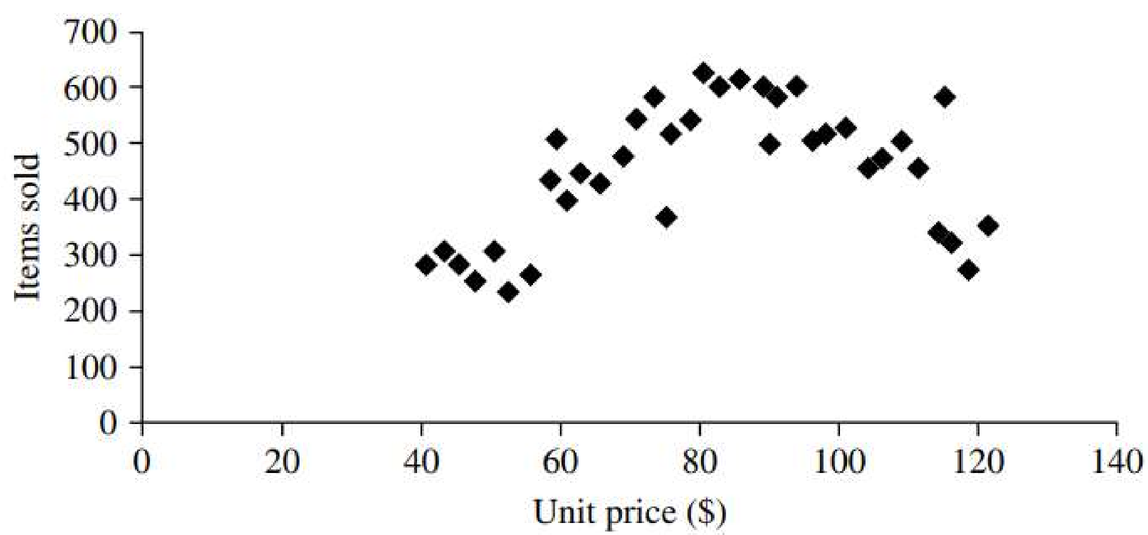


Figure 2.7 A scatter plot for the Table 2.1 data set.

[2]

DATA PRE-PROCESSING

- Data cleaning
- Data integration
- Data reduction
- Data transformation

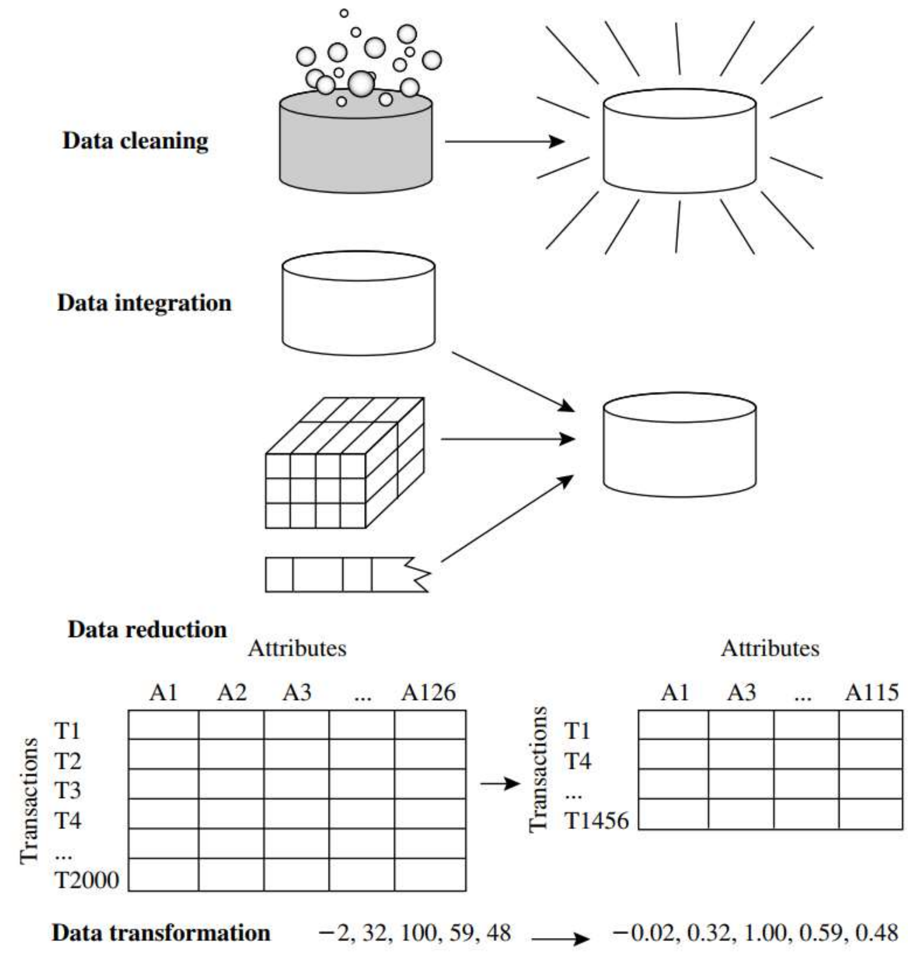


Figure 3.1 Forms of data preprocessing.

[2]

DATA CLEANING (1/3)

- Real-world data tend to be incomplete, noisy, and inconsistent.
- Data cleaning attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

DATA CLEANING (2/3)

- Missing values
  - Ignore the tuple
  - Fill in the missing values manually
  - Use a global constant
  - Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value
  - Use the attribute mean or median for all samples belonging to the same class as the given tuple
  - Use the most probable value to fill in the missing value

DATA CLEANING (3/3)

- Noisy data
  - Binning
  - Regression
  - Outlier analysis (e.g., clustering)

DATA INTEGRATION

- Entity identification problem
  - The same attribute or instance (e.g., cust_id vs. cust_no)
  - Constraints (e.g., bill discount vs. item discount)
- Redundancy and correlation analysis
  - Chi-square
  - Correlation coefficient and covariance
- Tuple duplication
- Data value conflict detection and resolution

DATA REDUCTION

- Dimensionality reduction
  - Wavelet transform
  - Principal component analysis
  - Attribute subset selection
- Numerosity reduction
  - Regression and log-linear models
  - Histograms, clustering, sampling, data cube aggregation
- Data compression

DATA TRANSFORMATION

- Smoothing
- Attribute construction
- Aggregation
- Normalization
- Discretization
- Concept hierarchy generation for nominal data

DATA PRE-PROCESSING DEMO

- Handling null values

DATA PRE-PROCESSING

	timestamp	building_name	temperature_1	temperature_2	temperature_3	temperature_4	temperature_5	pressure_1	pressure_2	pressure_3	pressure_4	pre
0	2019-01-01 10:00:00	building1	40.1746	44.2003	42.2857	48.0491	49.1427	107.4260	82.2464	68.8326	82.9828	1
1	2019-01-01 10:00:00	building2	43.5483	38.7111	44.8513	46.5925	36.1578	93.3252	107.4895	101.2728	103.6401	1
2	2019-01-01 12:04:00	building1	40.3374	36.9857	38.2883	49.7044	43.2163	95.4847	115.2700	92.5658	96.5299	1
3	2019-01-01 12:04:00	building2	44.2044	42.8381	37.6925	45.5218	46.4769	103.9656	99.8513	110.2489	81.7845	1
4	2019-01-01 14:00:00	building1	38.6388	49.3813	41.7175	39.1863	47.1067	108.2850	90.8498	113.5338	105.5288	1

	sensor1	sensor2	sensor3	sensor4	sensor5	sensor6	sensor7	sensor8	sensor9	sensor10	sensor11
count	500	500	500	500	500	500	500	500	500	478	25
mean	0.163042	0.145753	0.164147	0.165664	0.133664	0.166875	0.156911	7.46E-17	0.009245	-0.027	0.474572
std	0.908563	0.83765	0.828226	0.838714	0.834534	0.819496	0.840272	1.16E+00	1.751856	6.660626	0.263944
min	-2.6375	-2.5364	-2.1429	-2.0756	-2.015	-1.9605	-2.1858	-2.00E+00	-3.5388	-60	0.0924
25%	-0.52243	-0.44823	-0.4214	-0.45763	-0.45845	-0.44205	-0.4209	-1.00E+00	-1.47705	-0.44903	0.2381
50%	0.26885	0.21205	0.16765	0.2097	0.1895	0.25905	0.19695	0.00E+00	-0.0017	0.2298	0.4726
75%	0.872075	0.7738	0.780675	0.7923	0.75045	0.793375	0.828275	1.00E+00	1.54395	0.80935	0.6578
max	2.6526	2.2117	2.0564	1.9997	2.1503	2.1272	2.4542	2.00E+00	3.6178	100.11	0.9494

TO BE CONTINUED….

- Lecture 3: Data mining techniques
- Prepared softwares
  - PowerBI
  - Python

QUESTIONS AND ANSWERS

Picture from: http://philadelphiasculpturegym.blogspot.com/2013/09/save-date-free-talk-and-q-on-affordable.html

REFERENCES

[1] Tobias Zwingmann, AI-Powered Business Intelligence, Kindle Edition, O'reilly Press, 2022

[2] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Third Edition, Morgan Kaufmann Publishers, 2012.

[3] Jeen Su Lim, John Heinrichs, “Digital Business Intelligence Management with Big Data Analytics" Kindle Edition, O'reilly Press, 2021.

[4] David L. Olson, Dursun Delen, “Advanced Data Mining Techniques”, Springer-Verlag, 2008.

[5] Brian Larson, “Delivering Business Intelligence with Microsoft SQL Server 2016”, McGraw-Hill Education; 4 edition, 2016.

[6] Oracle, “Data Mining Concepts”, 18c, E83730-03, 2018

[7] Oracle, “Data Mining Application Developer’s Guide”, 2013.