CURRENT TOPICS IN COMPUTER SCIENCE

Business Intelligence Systems and Analytics

DATA MINING

Trong Nhan Phan, PhD

OUTLINE

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

DATA MINING OVERVIEW

KNOWLEDGE DISCOVERY

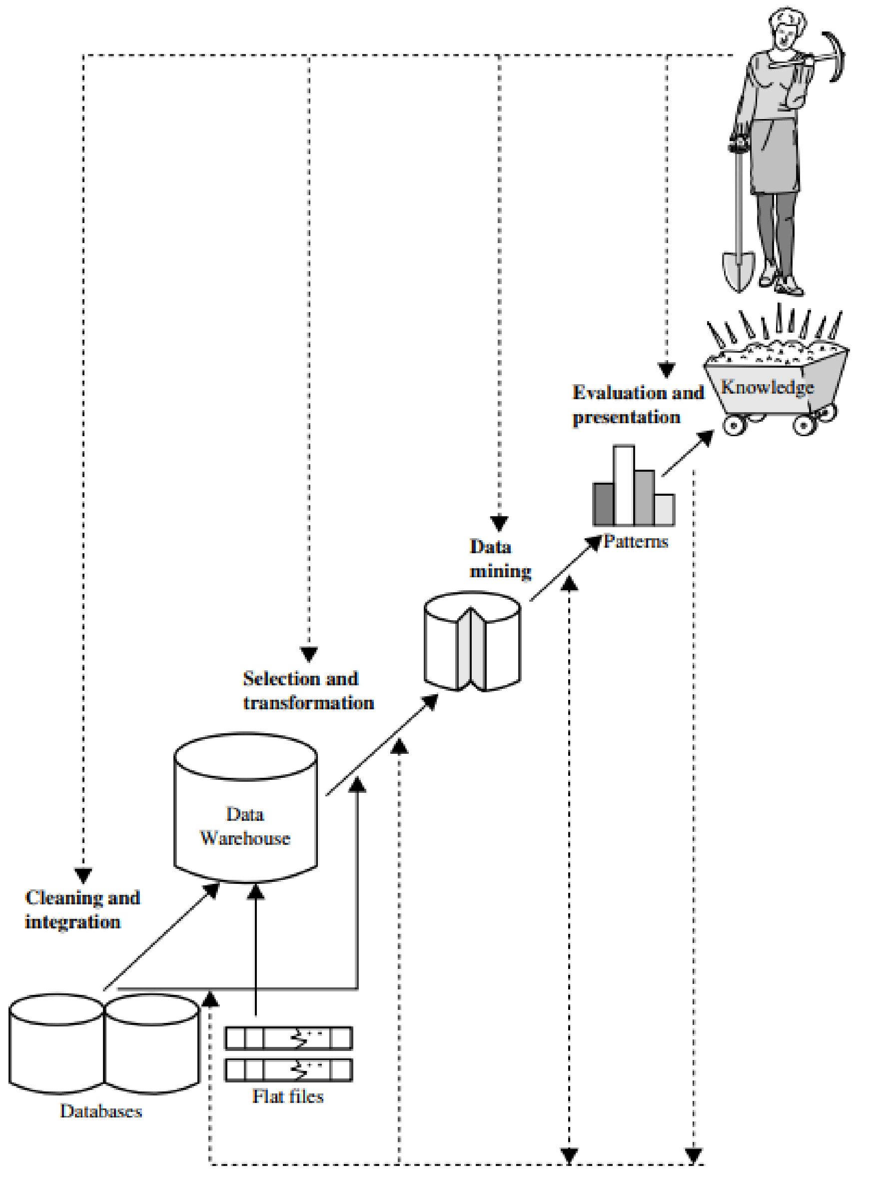


Figure 1.4 Data mining as a step in the process of knowledge discovery.

[2]

DATA MINING

- Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

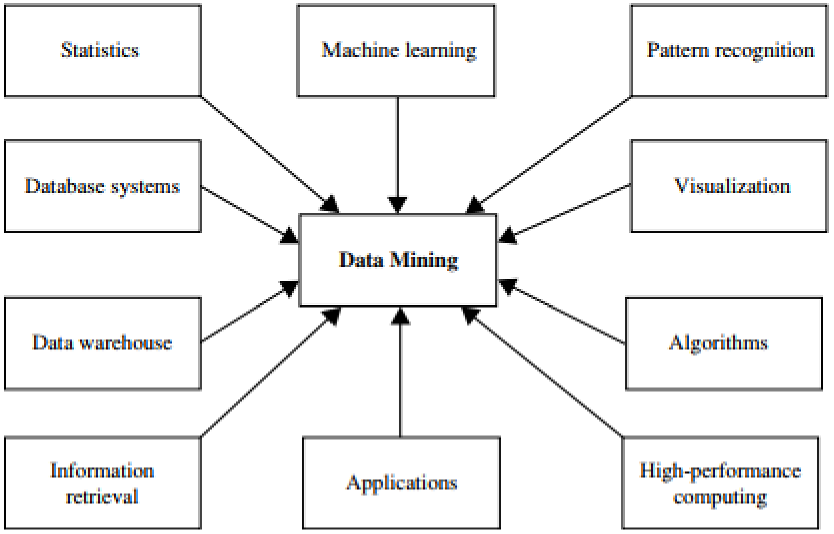


Figure 1.11 Data mining adopts techniques from many domains.

[2]

DATA MINING FUNCTIONALITY

- Descriptive tasks
  - characterize properties of the data in a target data set
- Predictive tasks
  - perform induction on the current data in order to make predictions

KINDS OF DATA TO BE MINED

- Database data
- Data warehouse data
- Transactional data
- Stream data
- Sequence data
- Graph data
- Spatial data
- Text data
- Multimedia data
- Etc.

KINDS OF PATTERNS TO BE MINED

- Class description
- Frequent patterns
- Associations
- Classification and regression
- Cluster analysis
- Outlier analysis

DATA CHARACTERIZATION

- It summarizes the data of the class under study (often called the target class) in general terms
- E.g.,
  - Summarizing the characteristics of customers who spend more than $5000 a year at AllElectronics.
  - The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.
  - The data mining system should allow the customer relationship manager to drill down on any dimension, such as on occupation to view these customers according to their type of employment.

DATA DISCRIMINATION

- It compares the target class with one or a set of comparative classes (often called the contrasting classes)
- E.g.,
  - Comparing two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year).
  - The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree. Drilling down on a dimension like occupation, or adding a new dimension like income level, may help to find even more discriminative features between the two classes.

FREQUENT PATTERNS

- They are patterns that occur frequently in data
  - Frequent itemsets: a set of items that often appear together in a transactional data set
  - Frequent subsequences: a frequently occurring subsequence
  - Frequent substructures: a frequently occurring structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.

ASSOCIATIONS

- Mining frequent patterns leads to the discovery of interesting associations within data.
- E.g.,
  - *buys(X*, "*computer*"*)* → *buys(X*, "*software*"*)* with *support* = 1% and *confidence* = 50%
  - *age(X*, "20..29") and *income(X*, "40K..49K"*)* → *buys(X*, "laptop") *with support* = 2% and *confidence* = 60%

CLASSIFICATION

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.

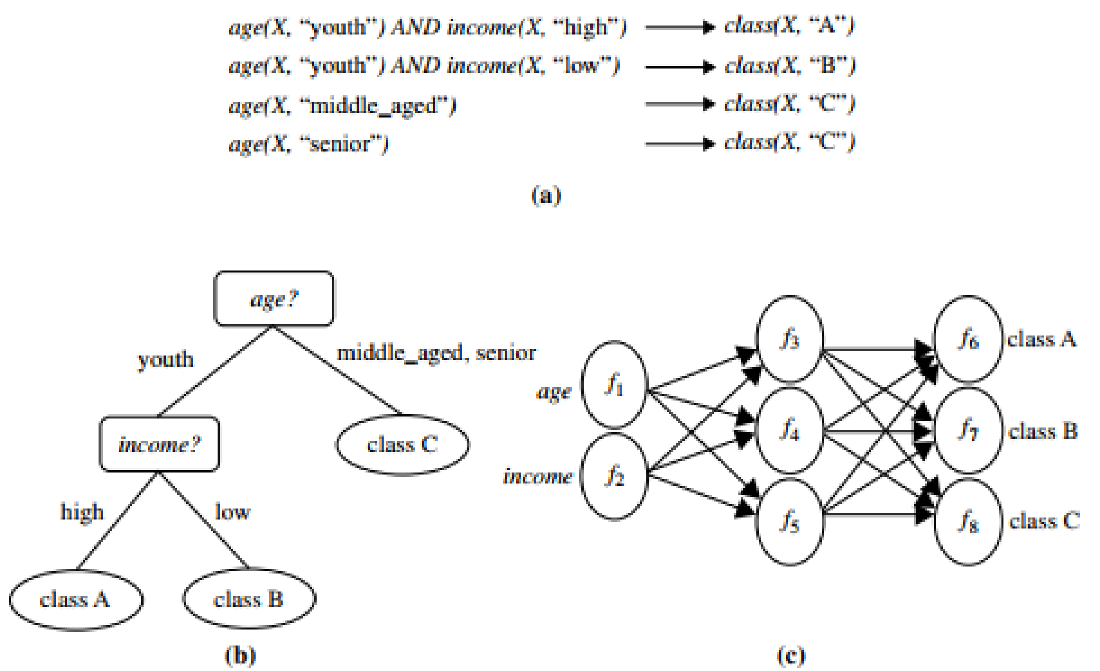


Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

[2]

REGRESSION

- Regression models continuous-valued functions. That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.

Simple linear regression:

$$Y = a + bX + u$$

Multiple linear regression:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$$

where:

Y = The dependent variable you are trying to predict or explain

X = The explanatory (independent) variable(s) you are using to predict or associate with Y

a = The y-intercept

b = (beta coefficient) is the slope of the explanatory variable(s)

u = The regression residual or error term

Internet

CLUSTER ANALYSIS

- Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.
- Clustering analyzes data objects without consulting class labels.
- E.g., interests based on locations

OUTLIER ANALYSIS

- Objects that do not comply with the general behavior or model of the data. These data objects are outliers.
- E.g., fraudulent usage of credit cards

DISCUSSION

ARE ALL PATTERNS INTERESTING?

DISCUSSION

DATA MINING VS.
MACHINE LEARNING?

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

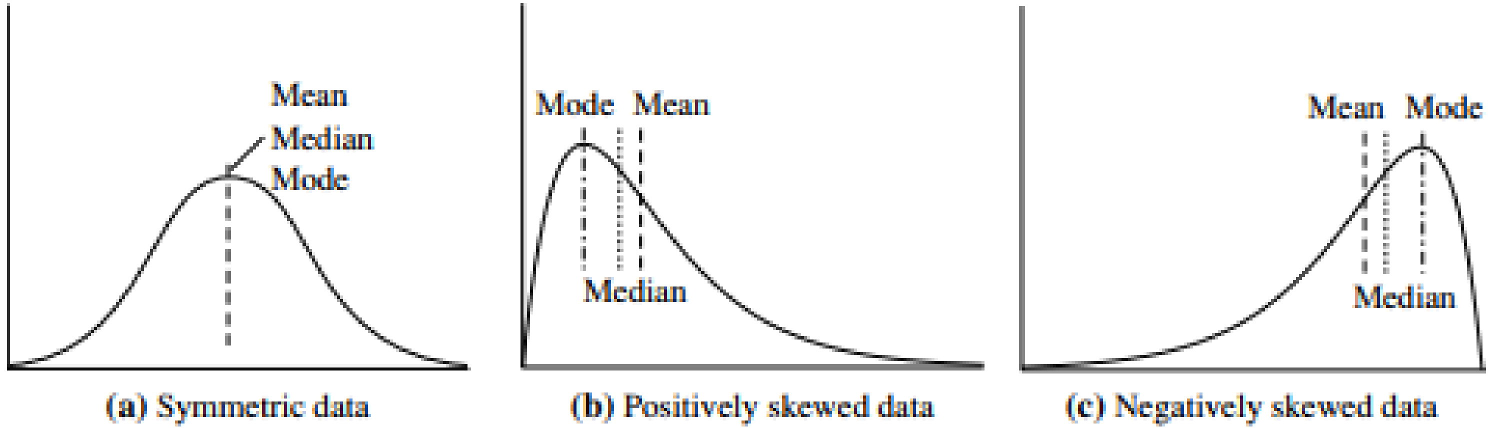
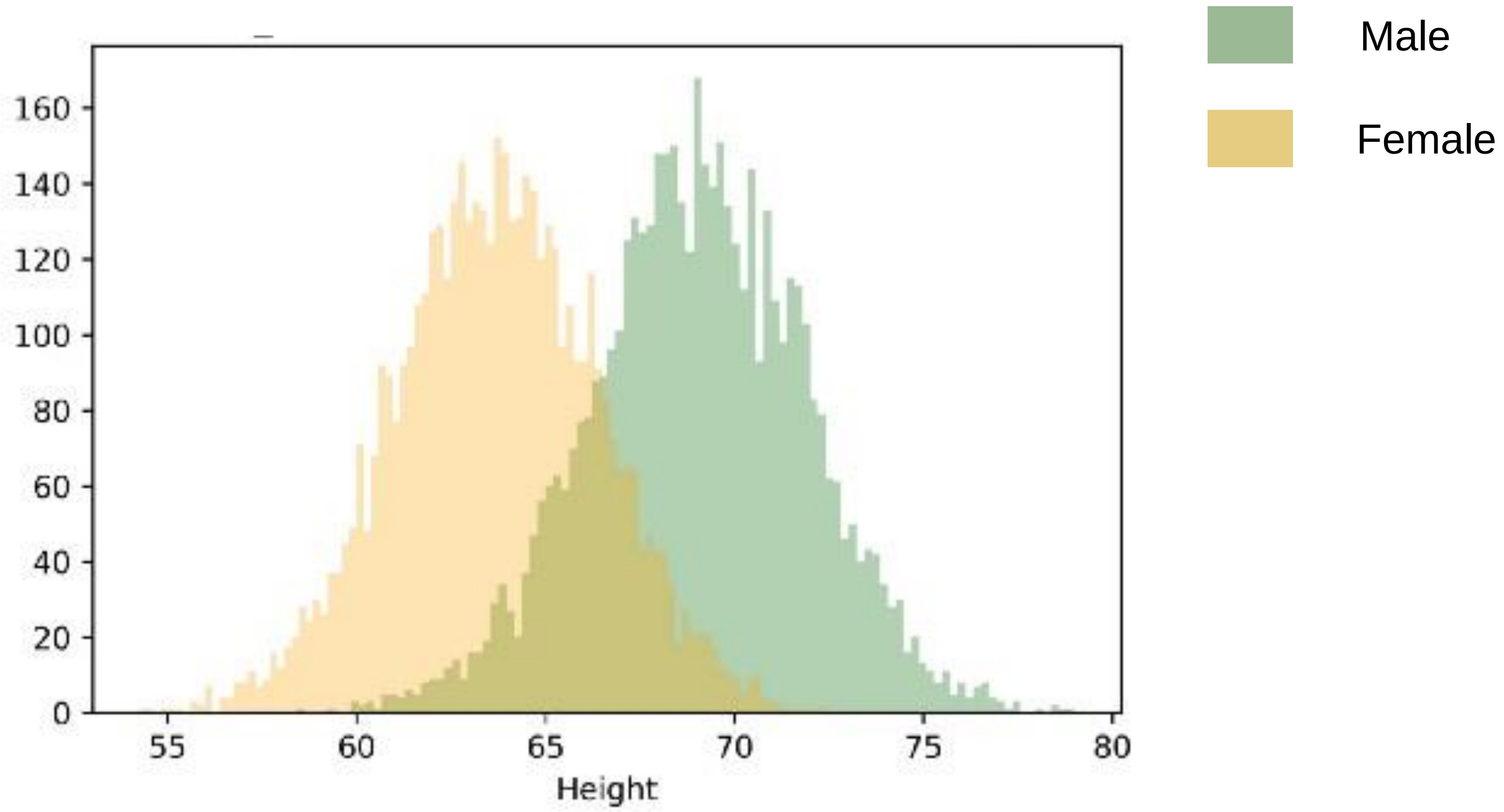


Figure 2.1 Mean, median, and mode of symmetric versus positively and negatively skewed data.

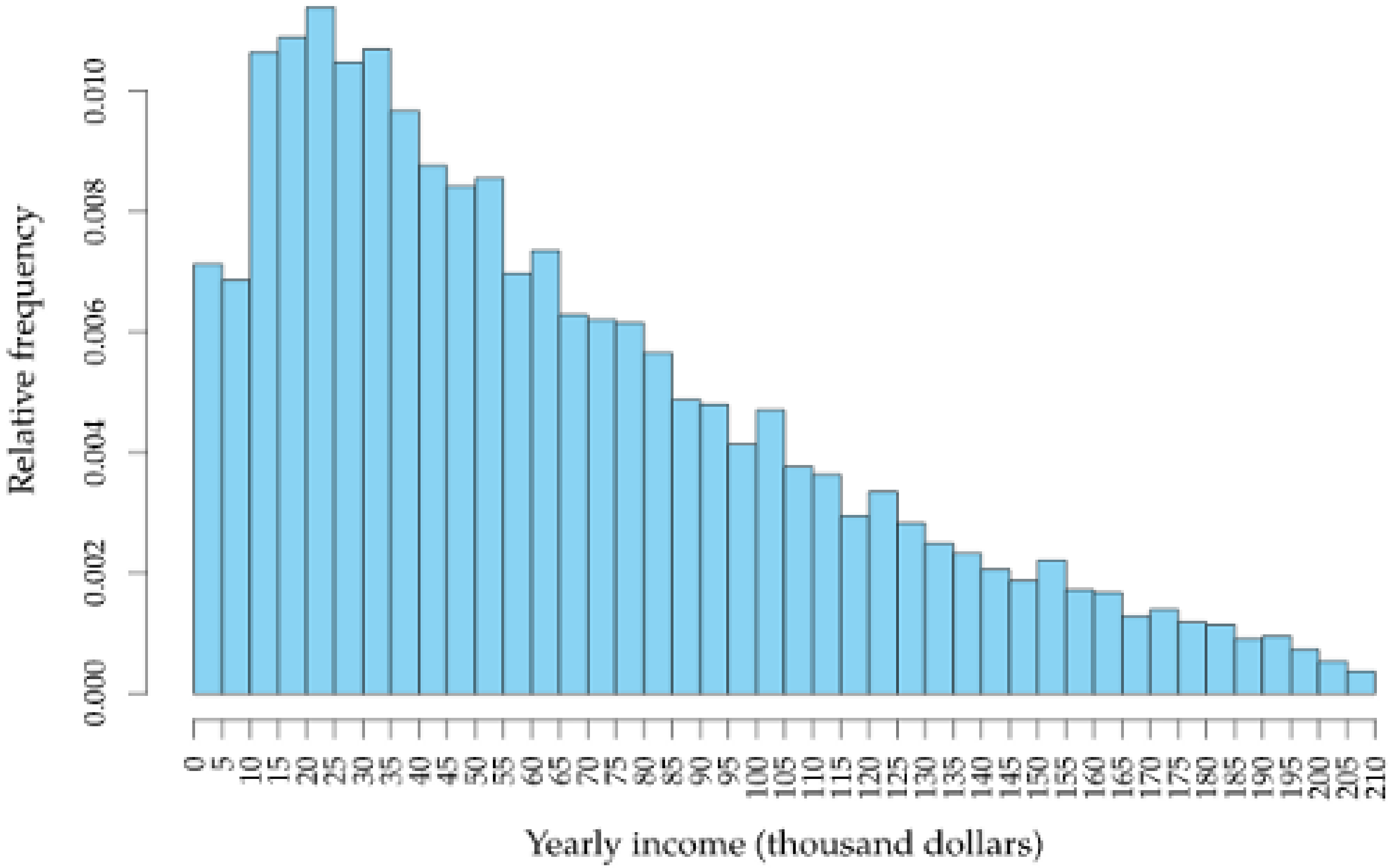
[2]

FOR INSTANCE



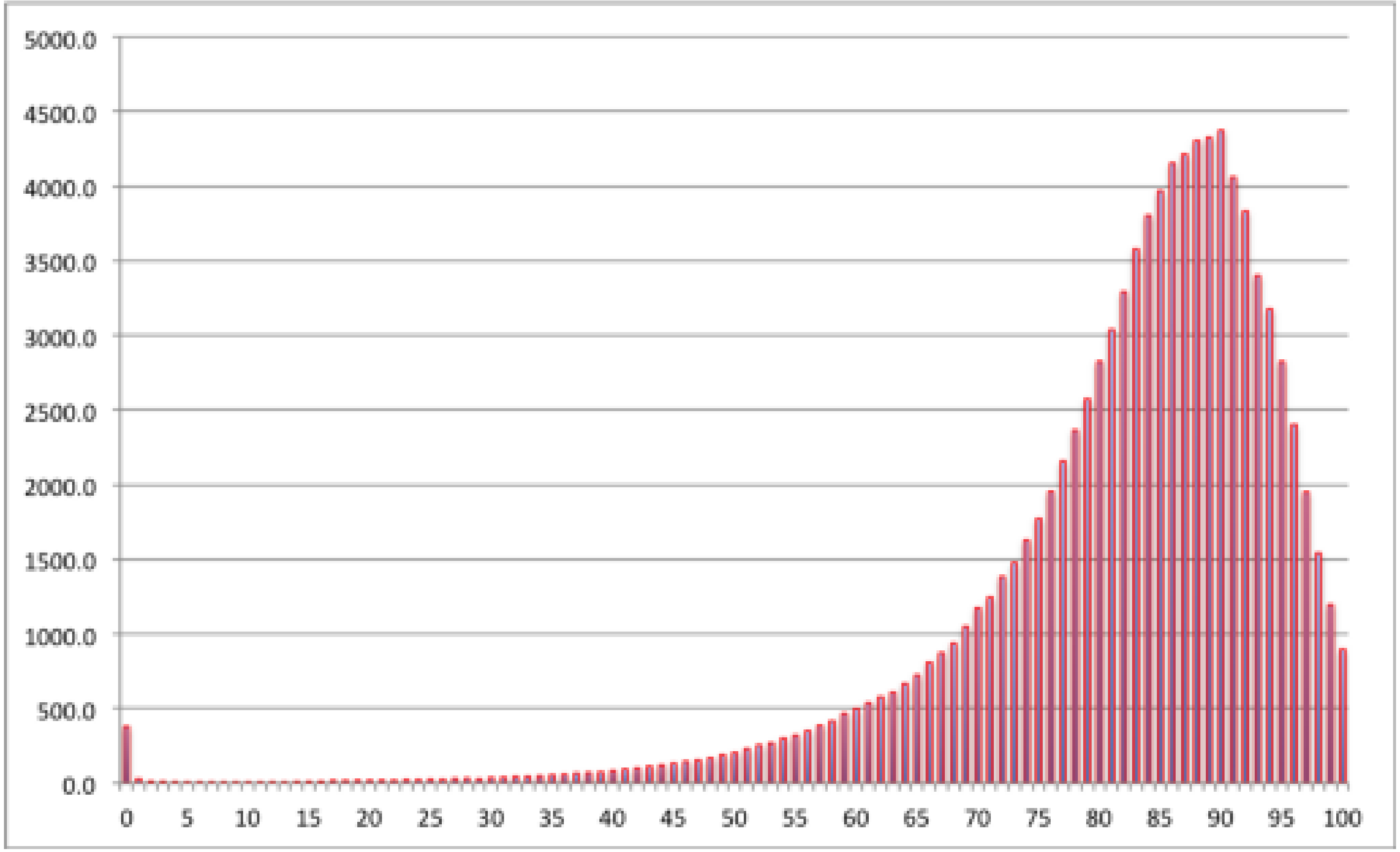
https://builtin.com/sites/www.builtin.com/files/styles/ckeditor_optimize/public/inline-images/national/1_skewed%2520data.png

FOR INSTANCE



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

FOR INSTANCE

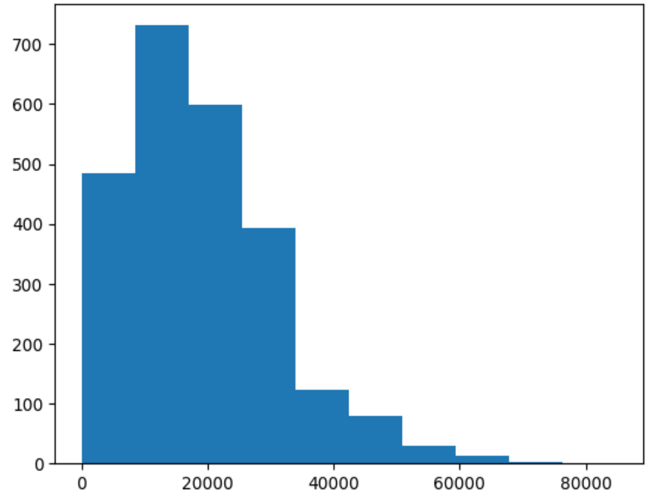


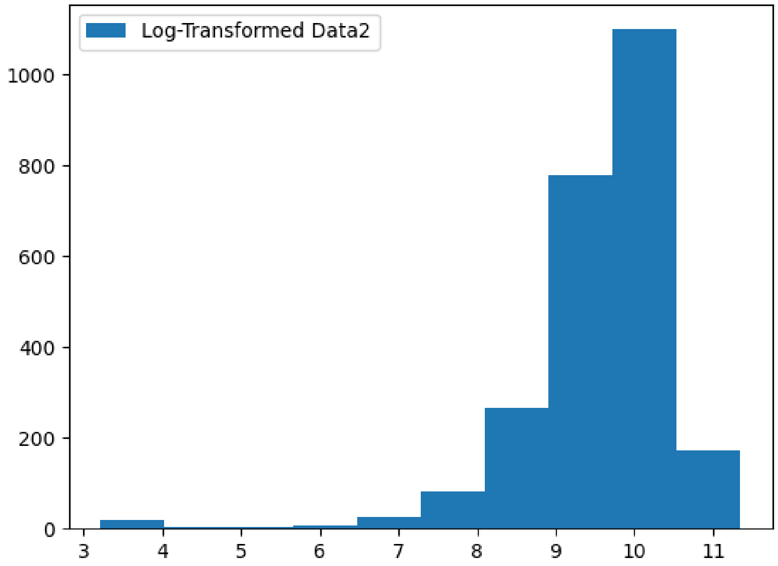
Mortality age

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.

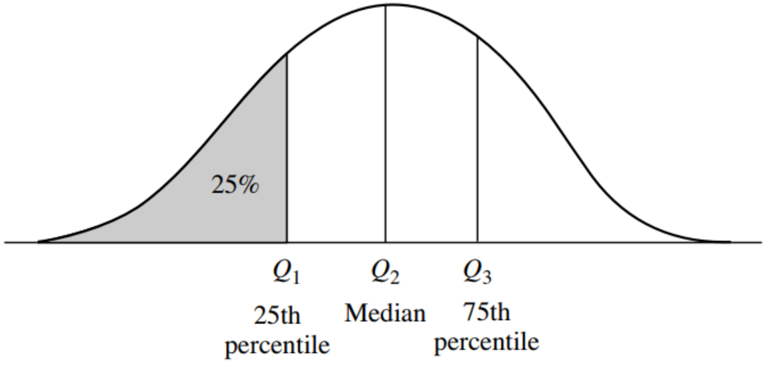


Figure 2.2 A plot of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

[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.

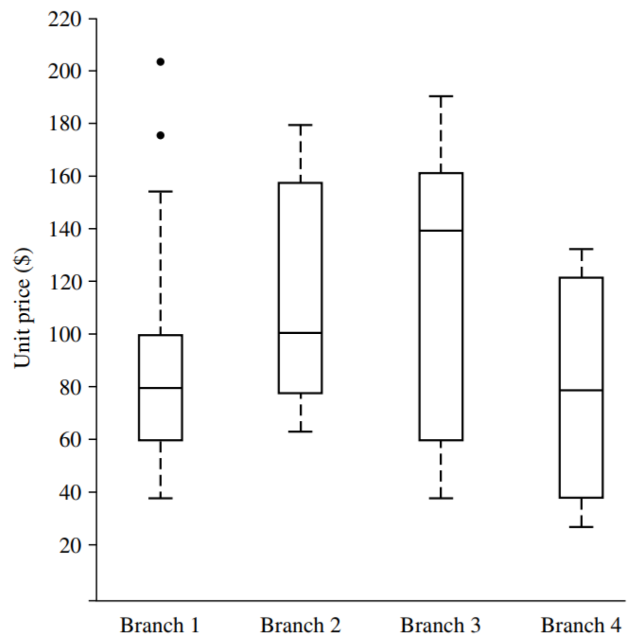


Figure 2.3 Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

[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

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

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?

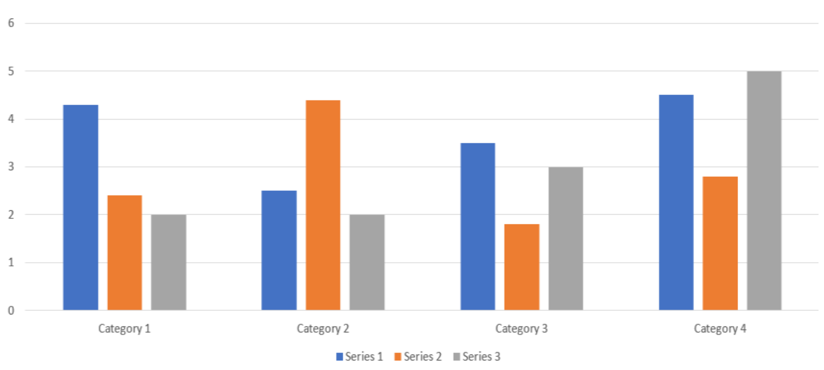
[2]

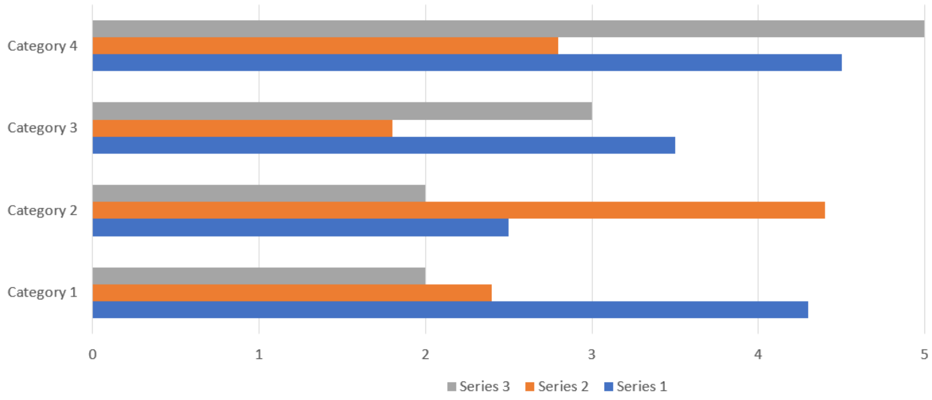
GRAPHIC DISPLAYS

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

BAR CHARTS

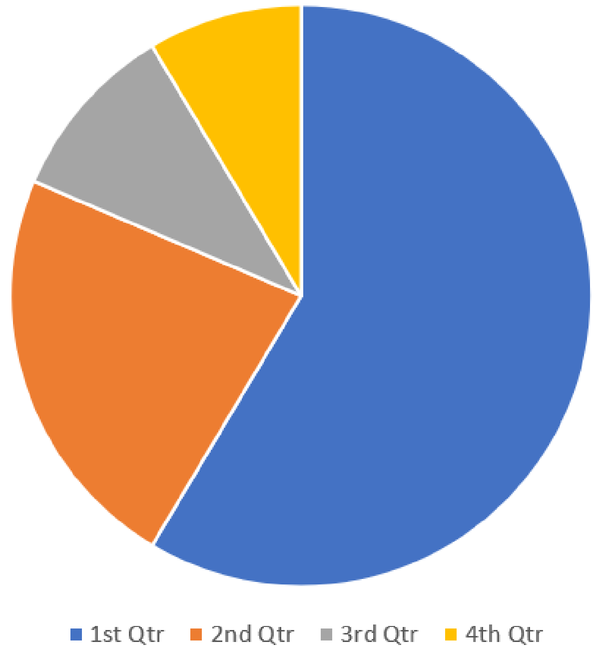
- A bar is a graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.
- The bars can be plotted vertically or horizontally.





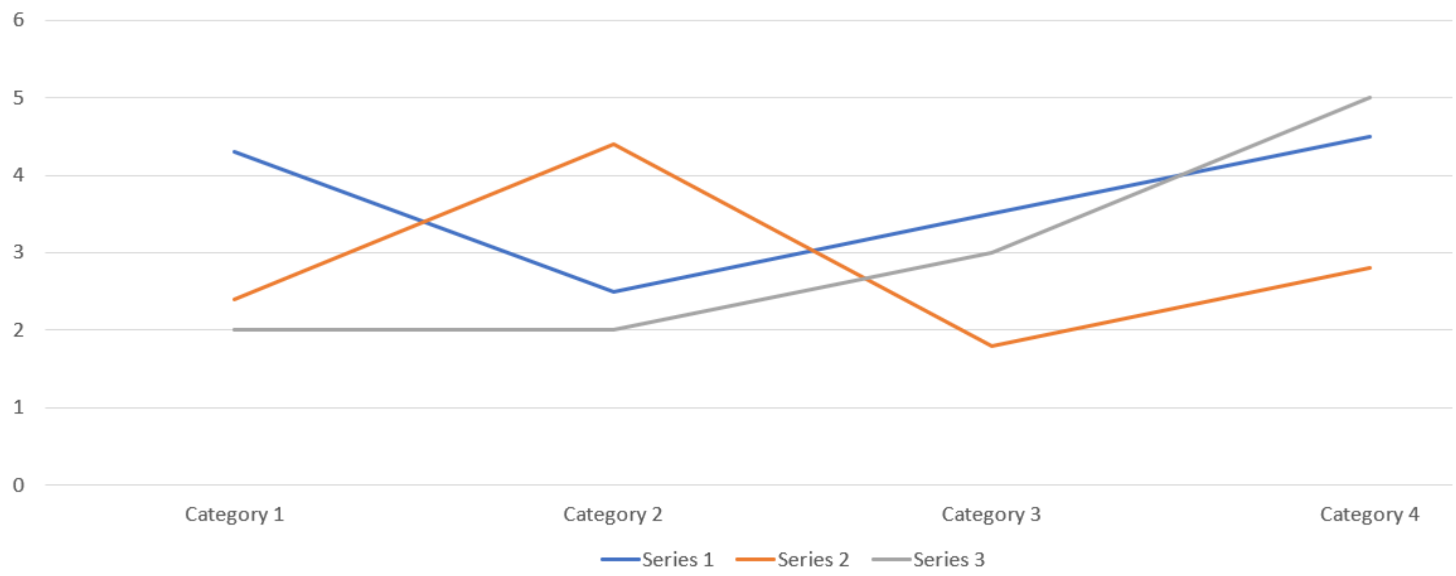
PIE CHARTS

- 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.

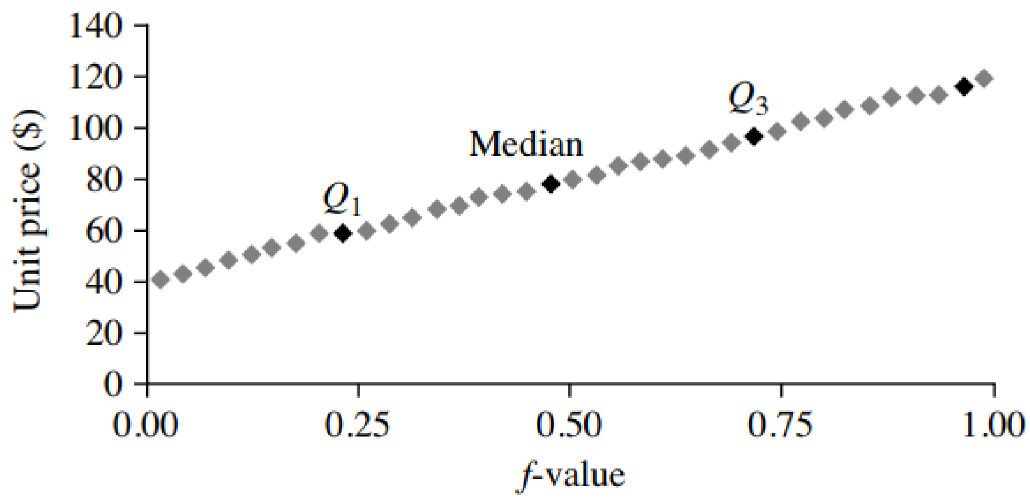
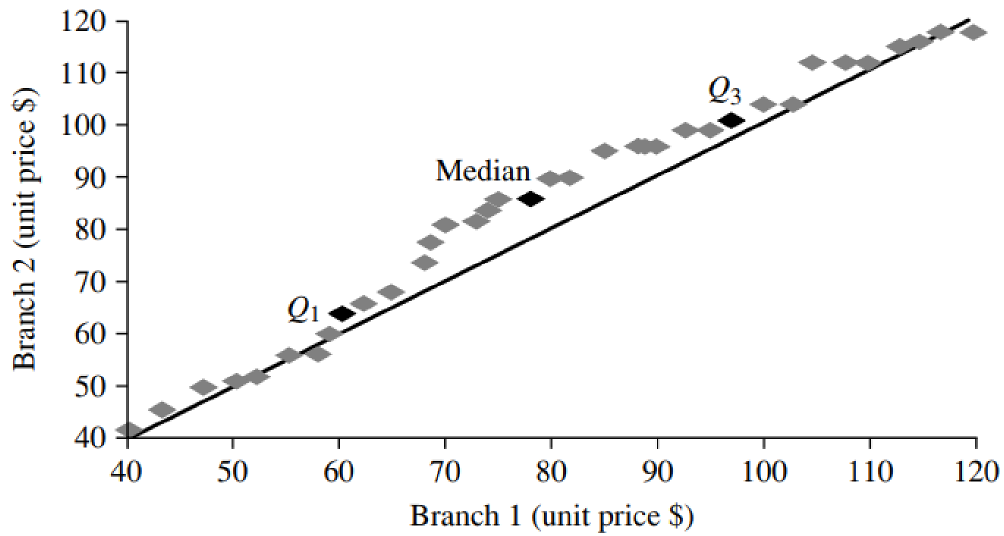


Figure 2.4 A quantile plot for the unit price data of Table 2.1.

[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.

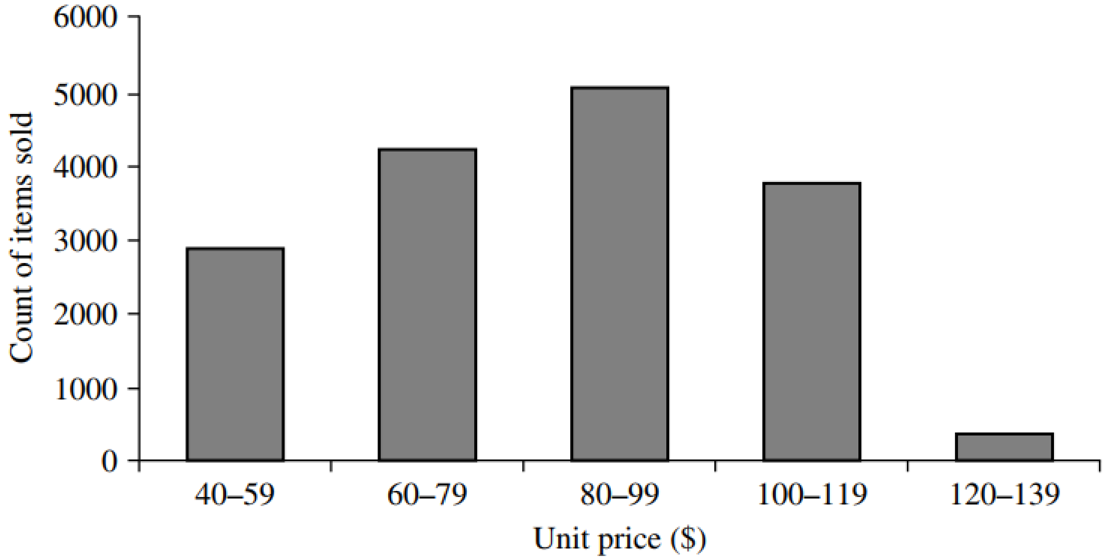


[2]

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.



[2]

Figure 2.6 A histogram for the Table 2.1 data set.

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.

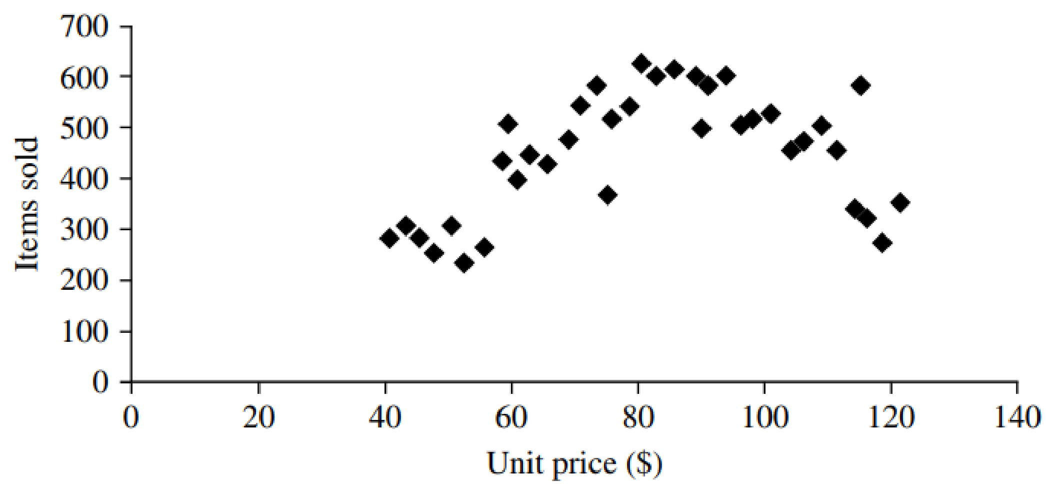


Figure 2.7 A scatter plot for the Table 2.1 data set.

[2]

DATA PRE-PROCESSING

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

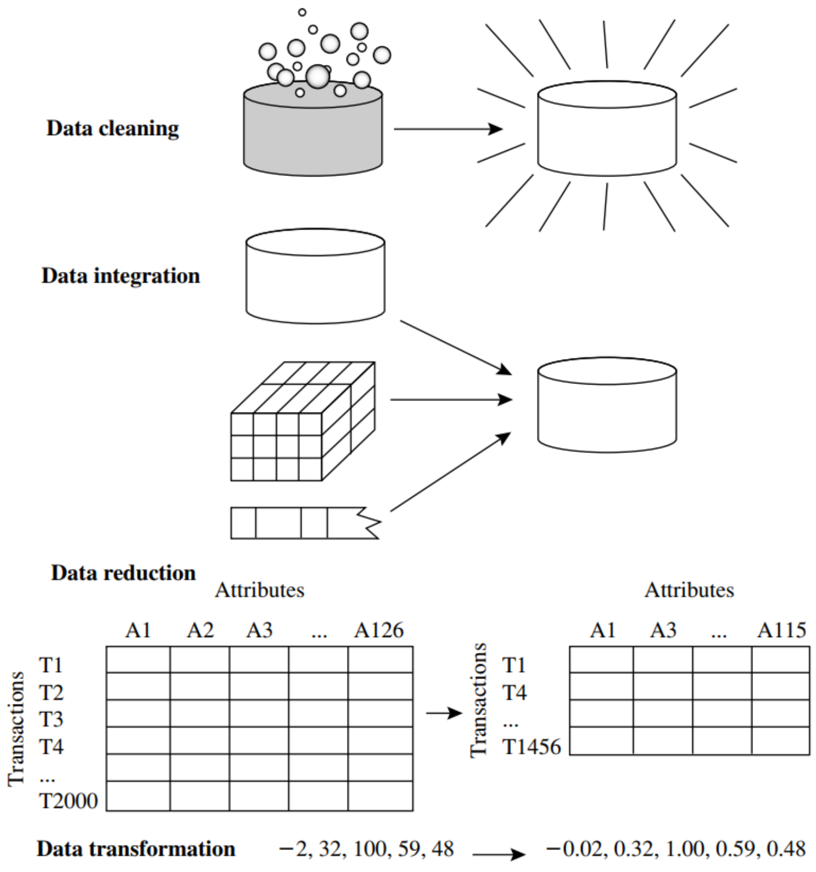


Figure 3.1 Forms of data preprocessing.

[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

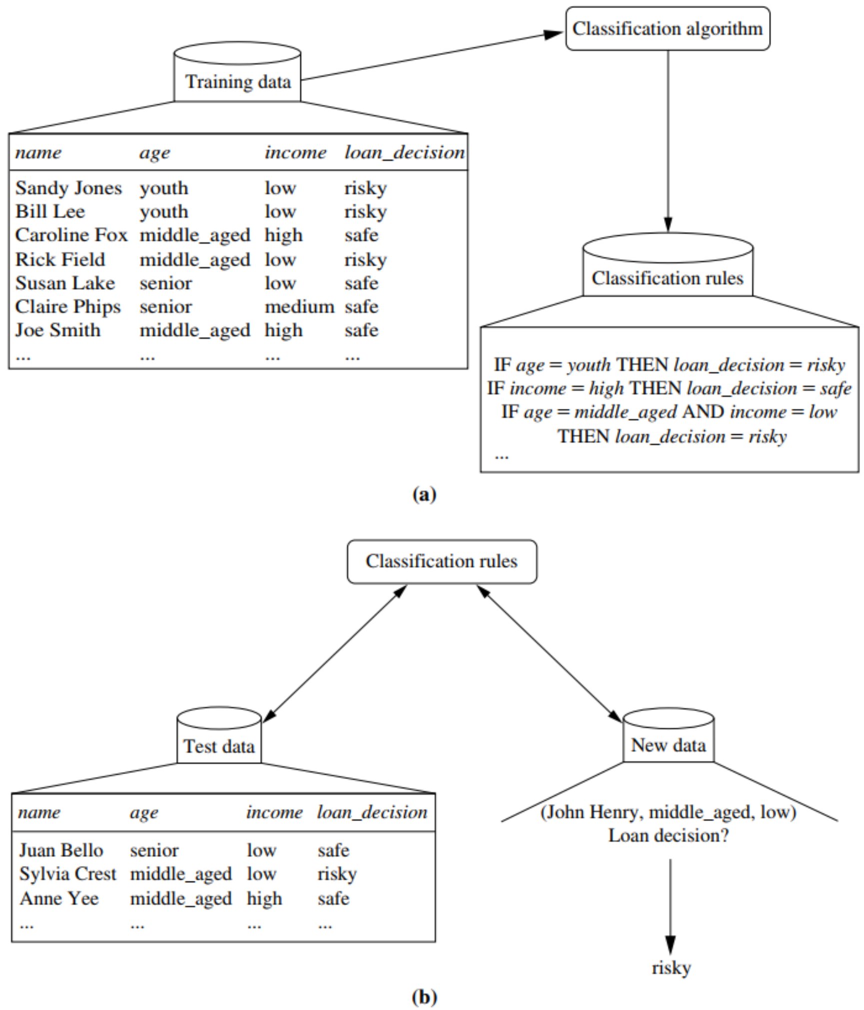
	sensor1	sensor2	sensor3	sensor4	sensor5	sensor6	sensor7	sensor8	sensor9	sensor10	sensor11
count	500	500	500	500	500	500	500	500	500	478	25
mean	0.163042	0.145753	0.164147	0.165664	0.133664	0.166875	0.156911	7.46E-17	0.009245	-0.027	0.474572
std	0.908563	0.83765	0.828226	0.838714	0.834534	0.819496	0.840272	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

CLASSIFICATION TECHNIQUES

CLASSIFICATION (1/2)

- Data classification is a two-step process:
  - A learning step where a classification model is constructed
  - A classification step where the model is used to predict class labels for given data
- E.g.,
  - Loan applicants are safe or risky
  - A customer profile to buy a computer
  - One of the treatment a patient should receive

CLASSIFICATION (2/2)



[2]

DECISION TREES (1/3)

- ID3
- C4.5
- CART

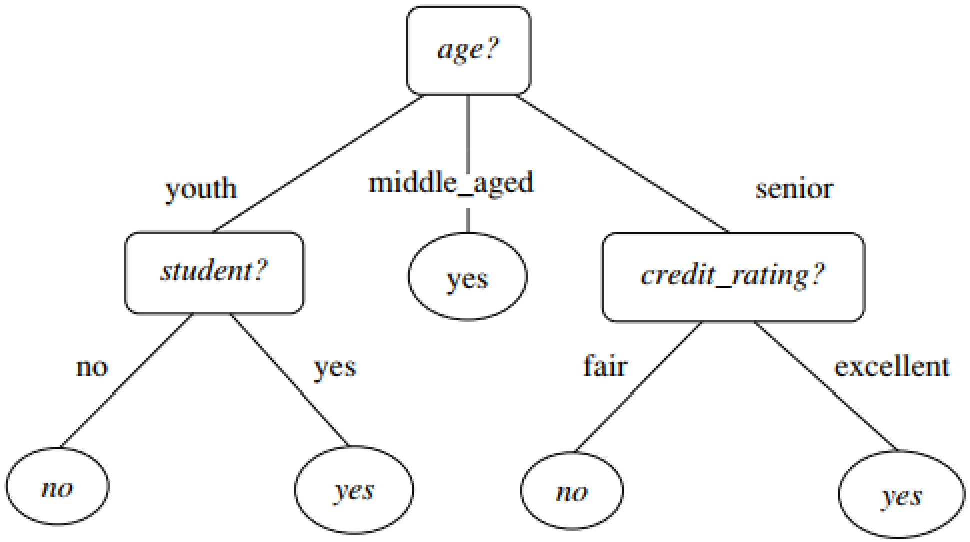


Figure 8.2 A decision tree for the concept *buys_computer*, indicating whether an *AllElectronics* customer is likely to purchase a computer. Each internal (nonleaf) node represents a test on an attribute. Each leaf node represents a class (either *buys_computer* = *yes* or *buys_computer* = *no*).

[2]

DECISION TREES (2/3)

- E.g.,

The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), \tag{8.1}$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j). \tag{8.2}$$

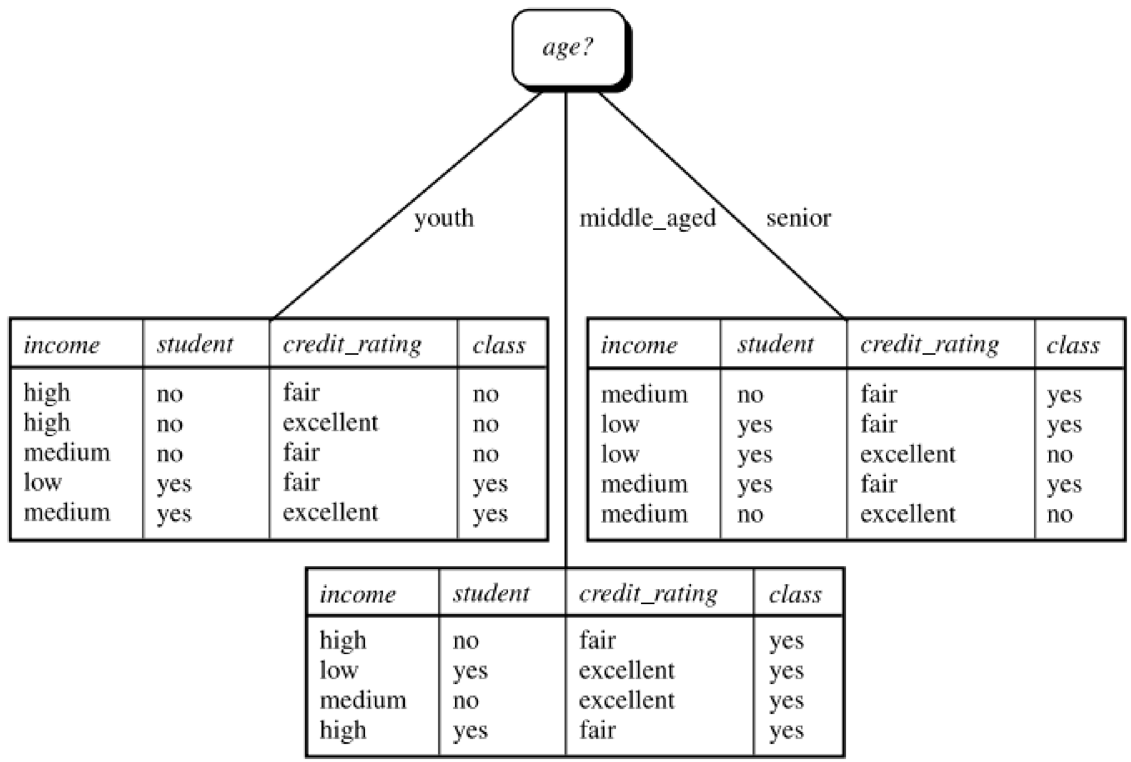
Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

[2]

DECISION TREES (3/3)

- E.g., $Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246$ bits.



income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

Figure 8.5 The attribute *age* has the highest information gain and therefore becomes the splitting attribute at the root node of the decision tree. Branches are grown for each outcome of *age*. The tuples are shown partitioned accordingly.

[2]

FOR EXAMPLE

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

$$\begin{aligned} Info_{age}(D) &= \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\ &+ \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right) \\ &+ \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \\ &= 0.694 \text{ bits.} \end{aligned}$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

$Gain(income) = 0.029$ bits, $Gain(student) = 0.151$ bits $Gain(credit\_rating) = 0.048$ bits

[2]

BAYES CLASSIFICATION METHODS (1/2)

- Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}.$$

To predict the class label of $\mathbf{X}$, $P(\mathbf{X}|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple $\mathbf{X}$ is the class C_i if and only if

$$P(\mathbf{X}|C_i)P(C_i) > P(\mathbf{X}|C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i. \qquad (8.15)$$

[2]

BAYES CLASSIFICATION METHODS (2/2)

$$X = (age = youth, income = medium, student = yes, credit\_rating = fair)$$

- P(buys_computer=yes|X) = ?
- P(buys_computer=no|X) = ?

Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

[2]

FOR INSTANCE

$$P(buys\_computer = yes) = 9/14 = 0.643$$
$$P(buys\_computer = no) = 5/14 = 0.357$$

To compute $P(\mathbf{X}|C_i)$, for $i = 1, 2$, we compute the following conditional probabilities:

$$
\begin{aligned}
P(age = youth \mid buys\_computer = yes) &= 2/9 = 0.222\\
P(age = youth \mid buys\_computer = no) &= 3/5 = 0.600\\
P(income = medium \mid buys\_computer = yes) &= 4/9 = 0.444\\
P(income = medium \mid buys\_computer = no) &= 2/5 = 0.400\\
P(student = yes \mid buys\_computer = yes) &= 6/9 = 0.667
\end{aligned}
$$

$$
\begin{aligned}
P(student = yes \mid buys\_computer = no) &= 1/5 = 0.200\\
P(credit\_rating = fair \mid buys\_computer = yes) &= 6/9 = 0.667\\
P(credit\_rating = fair \mid buys\_computer = no) &= 2/5 = 0.400
\end{aligned}
$$

Using these probabilities, we obtain

$$
\begin{aligned}
P(\mathbf{X}|buys\_computer = yes) = {} & P(age = youth \mid buys\_computer = yes)\\
& \times P(income = medium \mid buys\_computer = yes)\\
& \times P(student = yes \mid buys\_computer = yes)\\
& \times P(credit\_rating = fair \mid buys\_computer = yes)\\
= {} & 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044.
\end{aligned}
$$

Similarly,

$$P(\mathbf{X}|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

To find the class, C_i, that maximizes $P(\mathbf{X}|C_i)P(C_i)$, we compute

$$P(\mathbf{X}|buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(\mathbf{X}|buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$

Therefore, the naïve Bayesian classifier predicts $buys\_computer = yes$ for tuple $\mathbf{X}$. ■

[2]

EVALUATION METRICS (1/2)

- **True positives** *(TP)*: These refer to the positive tuples that were correctly labeled by the classifier. Let *TP* be the number of true positives.
- **True negatives** *(TN)*: These are the negative tuples that were correctly labeled by the classifier. Let *TN* be the number of true negatives.
- **False positives** *(FP)*: These are the negative tuples that were incorrectly labeled as positive. Let *FP* be the number of false positives.
- **False negatives** *(FN)*: These are the positive tuples that were mislabeled as negative. Let *FN* be the number of false negatives.

[2]

EVALUATION METRICS (2/2)

- Confusion matrix

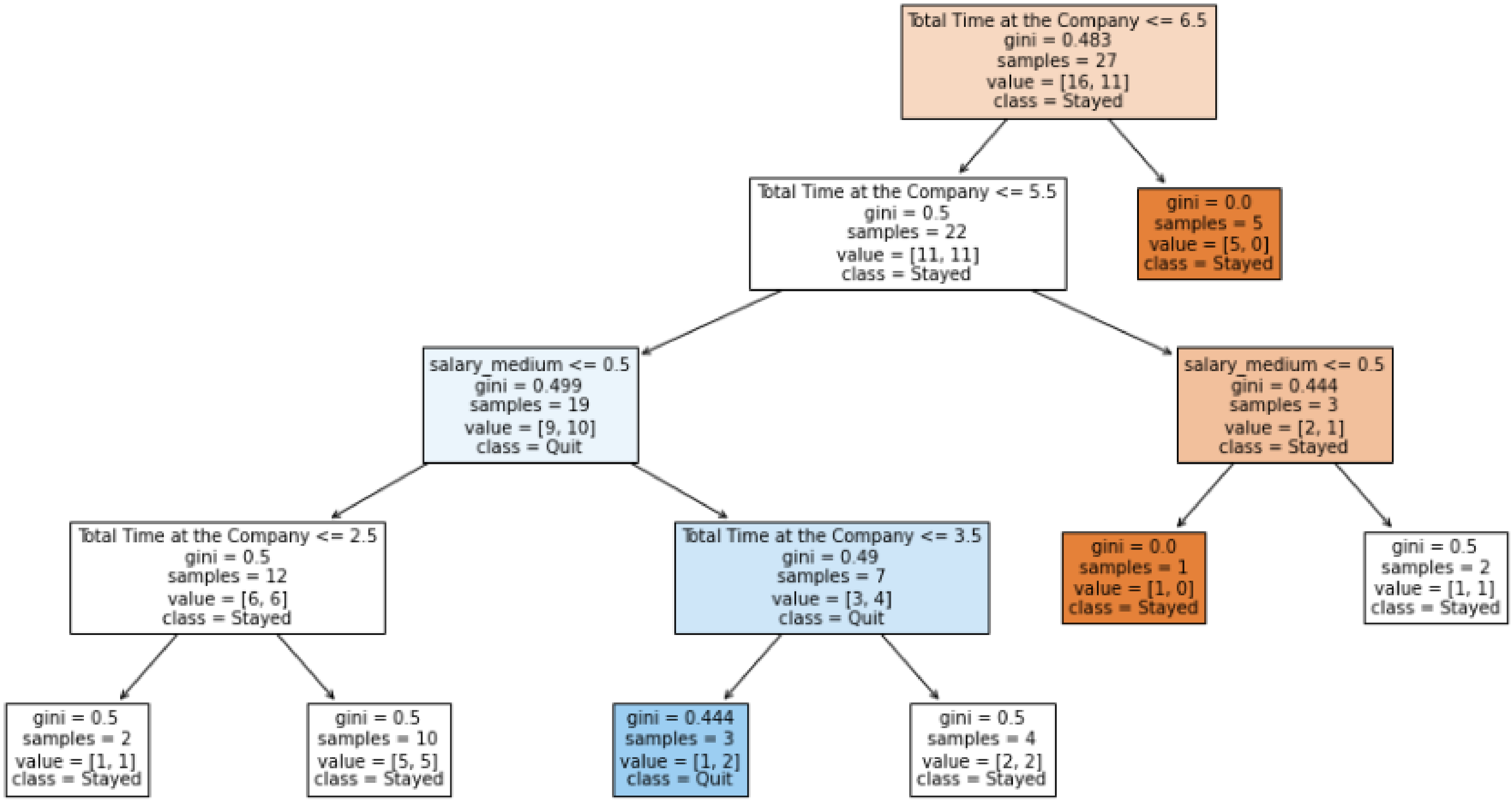
		Predicted class		
Actual class		*yes*	*no*	Total
	yes	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	$P+N$

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Figure 8.13 Evaluation measures. Note that some measures are known by more than one name. TP, TN, FP, P, N refer to the number of true positive, true negative, false positive, positive, and negative samples, respectively (see text).

[2]

COMPANY CHURN DEMO



CLUSTERING TECHNIQUES

CLUSTERING

- Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.
- E.g.,
  - Customer segmentation
  - Handwritten character recognition
  - Web search results
  - Outlier analysis

PARTITIONING METHODS

- K-means
- K-modes
- K-medoids

K-MEANS EXAMPLE

- S = {2, 3, 4, 10, 11, 12, 20, 25, 30}
- K = 2
- C1 = {2, 3, 4, 10, 11, 12)
- C2 = {20, 25, 30}

- Randomly take 2 means as centroids
  - m1 = 4
  - m2 = 12

HIERARCHICHAL METHODS

- Agglomerative
- Divisive

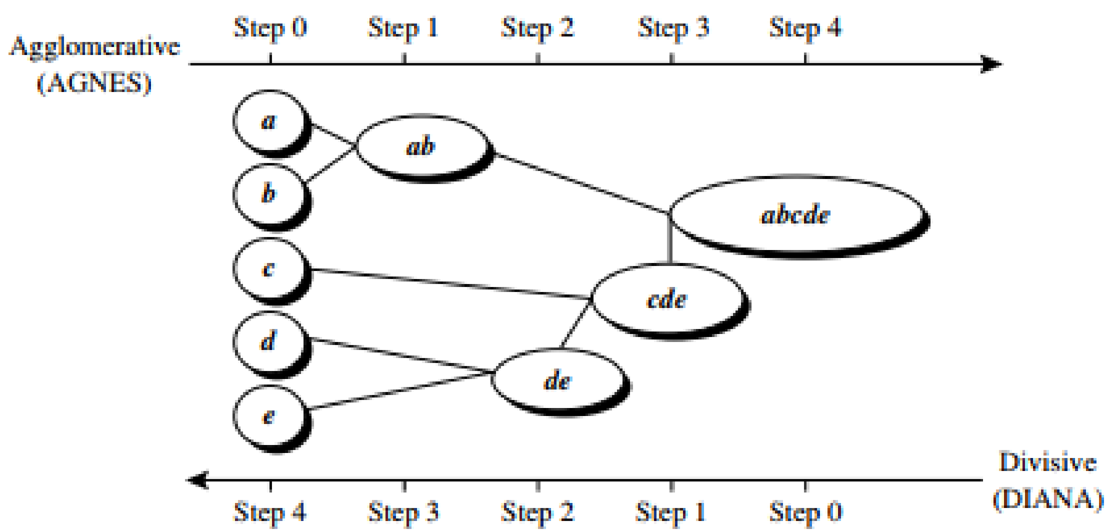
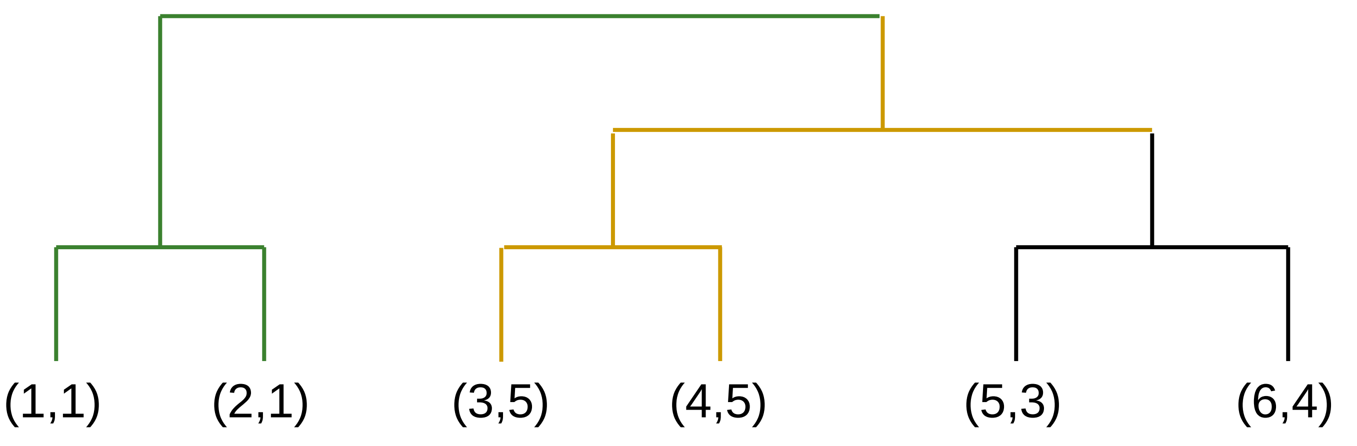


Figure 10.6 Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

[2]

AGGLOMERATIVE HIERARCHICHAL METHOD EXAMPLE

- S = {(1,1), (2,1), (3,5), (4,5), (5,3), (6,4)}



DENSITY-BASED METHODS

- We can model clusters as dense regions in the data space, separated by sparse regions, which can discover clusters of nonspherical shape.
- DBSCAN
  - Radius about each point (eps)
  - The minimum number of data points that should be around that point within that radius (MinPts)
  - E.g., (1.5, 2.5) with eps = 0.3, then the circle around the point with radius = 0.3, will contain only one other point inside it (1.2, 2.5)
- OPTICS
- DENCLUE

DBSCAN EXAMPLE

- eps = 0.6 and MinPts = 4
- The first data point (1,2)
- Cluster 1
  - (3,4), (2.5,4), (3,5), (2.8,4.5), (2.5,4.5)
- Cluster 2
  - (1,2), (1.5,2.5), (1.2,2.5), (1,3), (1,2.5)
- Outliers
  - (1,5), (5,6), (4,3)
- Example

x	y	d from (1,2)
1	2	0
3	4	2.8
2.5	4	2.5
1.5	2.5	0.7
3	5	3.6
2.8	4.5	3.08
2.5	4.5	2.9
1.2	2.5	0.53
1	3	1
1	5	3
1	2.5	0.5
5	6	5.6
4	3	3.1

GRID-BASED METHODS

- A grid-based clustering method takes a space-driven approach by partitioning the embedding space into cells independent of the distribution of the input objects.
- STING
- CLIQUE

EVALUATION METRICS

- Assessing clustering tendency so that nonrandom structure exists.
  - Hopkins statistic
- Determining the number of clusters in a data set.
  - The elbow method

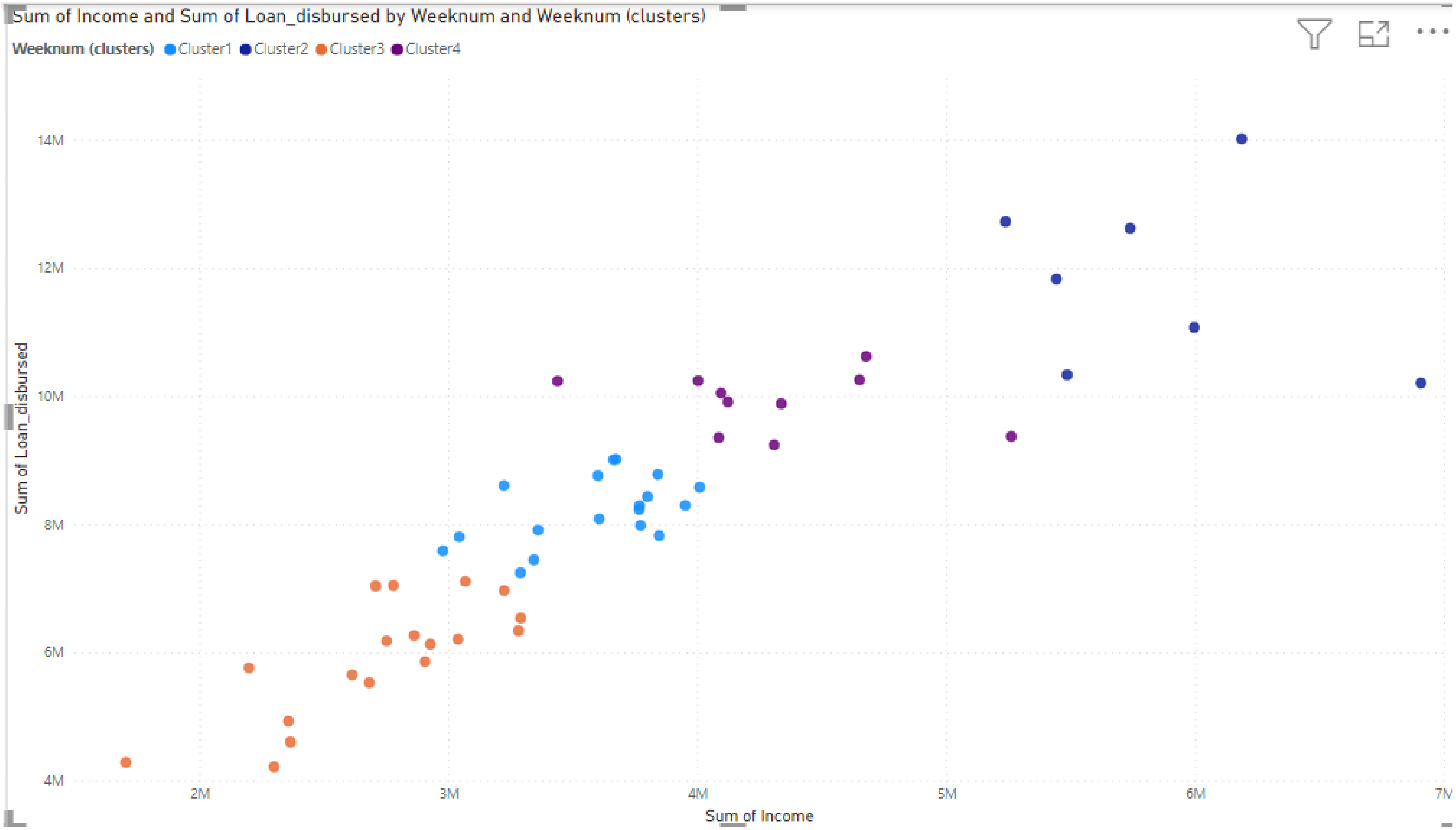
$$\text{WCSS} = \sum_{C_k}^{C_n} \left(\sum_{d_i \, in \, C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

Where,
C is the cluster centroids and d is the data point in each Cluster.

- Measuring clustering quality.
  - Extrinsic methods
  - Intrinsic methods

https://analyticsindiamag.com/beginners-guide-to-k-means-clustering/

BANK LOAN DISBURSAL CLUSTERING DEMO



ASSOCIATION RULES

ASSOCIATION RULES

- Frequent patterns and association rules are helpful for some scenario such as recommendation.
- Which patterns are interesting
  - support
  - confidence
  - lift
- Apriori algorithm
- FP-growth

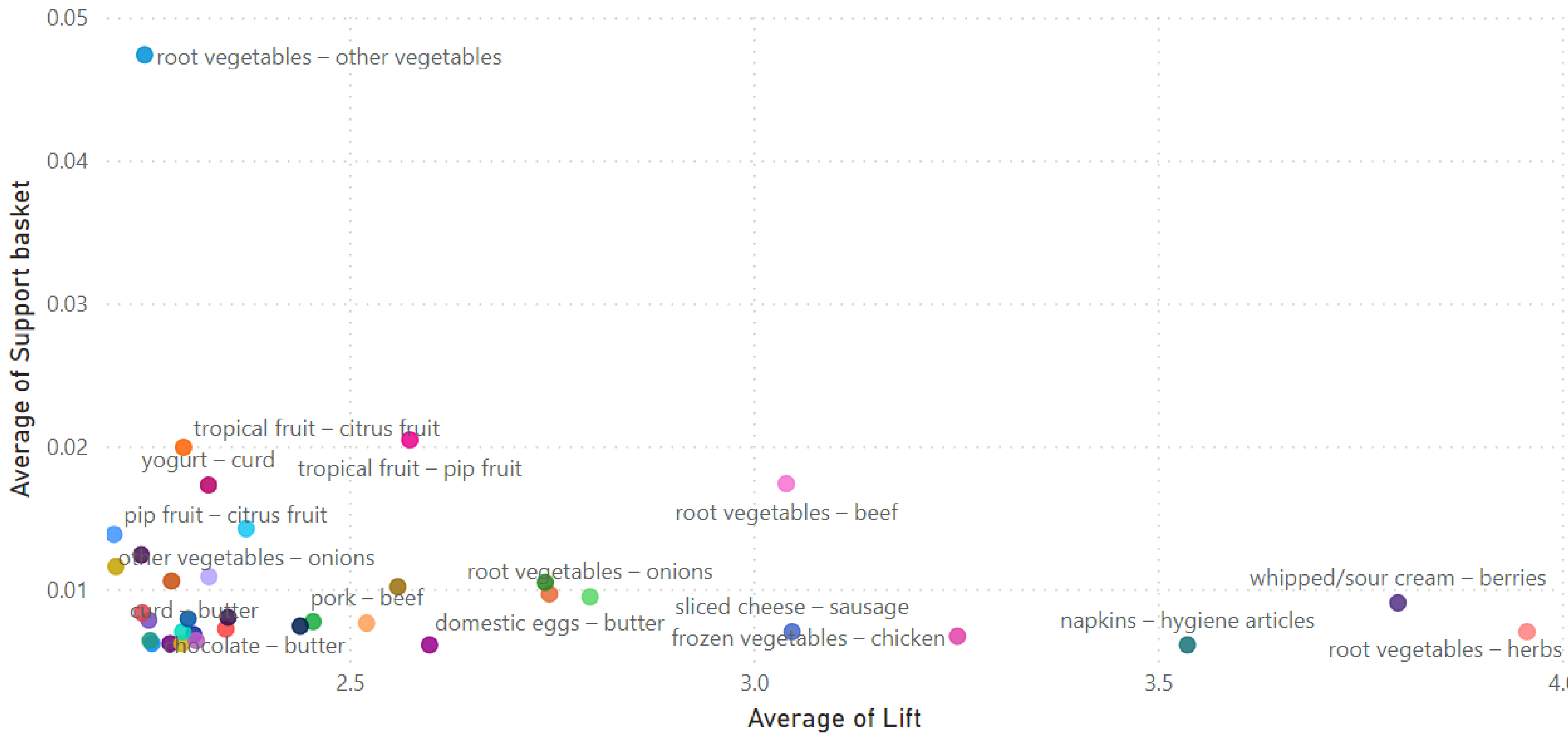
$$Support = \frac{Number\ of\ transactions\ including\ one\ or\ multiple\ products}{Total\ number\ of\ transactions}$$

$$Confidence\ of\ product\ one \rightarrow Basket = \frac{Support\ of\ basket}{Support\ of\ product\ one}$$

$$Confidence\ of\ product\ two \rightarrow Basket = \frac{Support\ of\ basket}{Support\ of\ product\ two}$$

$$Lift = \frac{Support\ of\ basket}{(Support\ of\ product\ one * Support\ of\ product\ two)}$$

BASKET ANALYSIS DEMO



LINEAR REGRESSION (1/2)

- Simple linear regression
  - Y = a + bX
    - X: independent variable
    - Y: outcome variable
    - a: Y-intercept
    - b: slope of the line
- For instance
  - Weight = 80 + 2(Height)

$$a = \frac{\left(\sum_{Y}\right)\left(\sum_{X^2}\right) - \left(\sum_{X}\right)\left(\sum_{XY}\right)}{n\left(\sum_{x^2}\right) - \left(\sum_{x}\right)^2}$$

$$b = \frac{n\left(\sum_{XY}\right) - \left(\sum_{X}\right)\left(\sum_{Y}\right)}{n\left(\sum_{x^2}\right) - \left(\sum_{x}\right)^2}$$

FOR INSTANCE

- Find a linear regression equation when given the dataset below:

x	y
2	3
4	7
6	5
8	10

LINEAR REGRESSION (2/2)

- Multiple linear regression
  - $Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$
    - $X_1 \ldots X_n$: independent variables
    - Y: outcome variable
    - a: Y-intercept
    - b: slope of the line
    - e: residuals (error)

$$b_1 = \frac{\left(\sum x_2^2\right)\left(\sum x_1 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_2 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)}$$

$$b_2 = \frac{\left(\sum x_1^2\right)\left(\sum x_2 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_1 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)}$$

$$a = b_0 = Y - b_1 X_1 - b_2 X_2$$

- For instance
  - BMI = 18.0 + 1.5 (diet score) + 1.6 (male) + 4.2 (age>20)

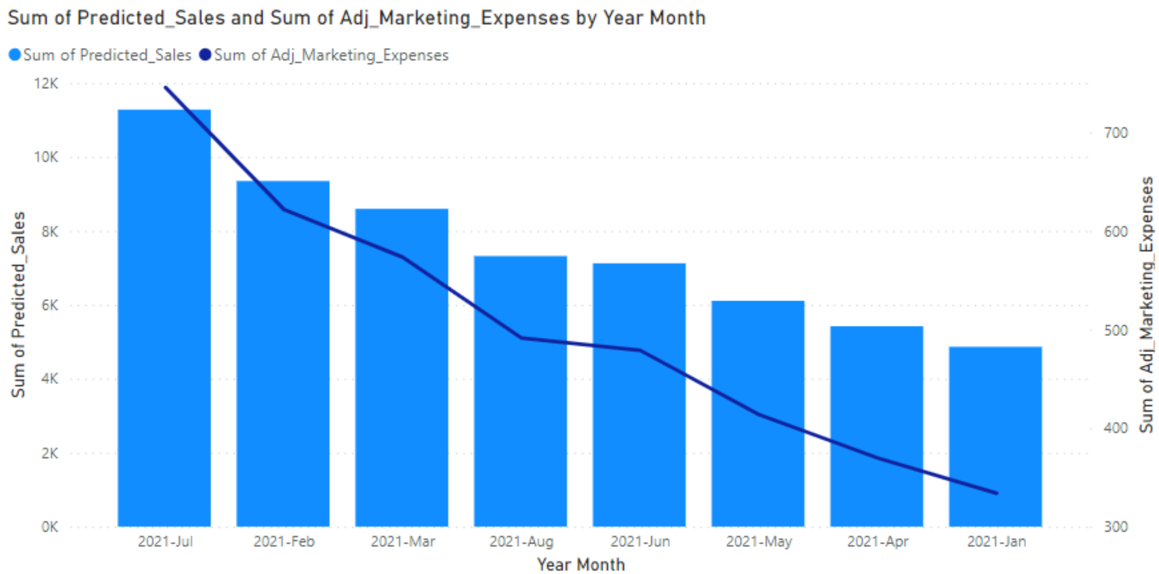
FOR INSTANCE

- Find a linear regression equation when given the dataset below:

Car	Price (thousand dollars)	Age (years)	Mileage (thousand miles)
1	29	1	18
2	25	2	25
3	21	2	50
4	18	3	68
5	15	4	75
6	15	5	65

Price = 32.46 – 1.54(Age) -0.15(Mileage)

SALES AND MARKETING EXPENSES DEMO



Marketing_Expenses_Range

All

Year Month	Sum of Budget_Marketing_Expenses	Marketing_Expenses_Range	Sum of Adj_Marketing_Expenses	Sales with Budgeted ME	Sum of Predicted_Sales	Growth
2021-Apr	7.60	0	7.60	102.35	102.35	0.00%
2021-Apr	7.60	1	8.60	102.35	117.93	15.23%
2021-Apr	7.60	2	9.60	102.35	133.52	30.45%
2021-Apr	7.60	3	10.60	102.35	149.10	45.68%
2021-Apr	7.60	4	11.60	102.35	164.68	60.91%
2021-Apr	7.60	5	12.60	102.35	180.27	76.13%
2021-Apr	7.60	6	13.60	102.35	195.85	91.36%
2021-Apr	7.60	7	14.60	102.35	211.44	106.58%
Total	**2,347.80**		**4,027.80**	**1,613.57**	**60,066.06**	**3622.56%**

SUMMARY

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

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. David L. Olson, Dursun Delen, “Advanced Data Mining Techniques,” Springer-Verlag, 2008.
4. Jeen Su Lim, John Heinrichs, “Digital Business Intelligence Management with Big Data Analytics," Kindle Edition, O'reilly Press, 2021.
5. William H. Inmon, “Building the Data Warehouse,” Fourth Edition, Wiley Publishing, Inc., 2005.
6. R. Kimball, M. Ross, "The Data Warehouse ToolKit," 3rd Edition, Wiley Publishing, Inc., 2013.
7. Turban, E., Aronson,J.E., “Decision Support Systems and Intelligent Systems” - 7th Edition, Prentice-Hall, 2005.
8. Ramesh Sharda, Dursun Delen, Efraim Turban, “Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support,” 7th Edition, Pearson Education, Inc., 2020.