

# **Data Mining:**

---

## **Chapter 3**

### **Pre-mining**

# **Data Mining:**

---

## **Concepts and Techniques**

**(3<sup>rd</sup> ed.)**

### **— Chapter 3 —**

Jiawei Han, Micheline Kamber, and Jian Pei  
University of Illinois at Urbana-Champaign &  
Simon Fraser University

©2011 Han, Kamber & Pei. All rights reserved.

# Test your knowledge

- Identify the "dirt" in this dataset

A	B	C	D	E	F	G	H	I
Date_Start	Date_End	Country	Location	Type	Sub_Type	Killed	Cost	ID
102008	102008	Afghanistan	Kunduz, Balkh, Faryab	Drought	Drought		280000	2008-9475
72006	2006	Afghanistan		Drought	Drought		1900000	2006-9570
52000	2002	Afghanistan	Kandahar, Helmand,	Drought	Drought	37	2580000	2000-9186
81971	1973	Afghanistan	Central, North-West	Drought	Drought			1971-9085
11969	1969	Afghanistan	Paktia province	Drought	Drought		48000	1969-9007
29072006	29072006	Afghanistan	Imam Sahib	Earthquake (seismic activity)	Earthquake (ground shaking)	1	935	2006-0405
13122005	13122005	Afghanistan	Hindu Kush	Earthquake (seismic activity)	Earthquake (ground shaking)	5	501	2005-0686
8102005	8102005	Afghanistan	Nangahar, Jalalabad	Earthquake (seismic activity)	Earthquake (ground shaking)	1		2005-0575
18072004	18072004	Afghanistan	Paktia province	Earthquake (seismic activity)	Earthquake (ground shaking)	2	1040	2004-0436
10042003	10042003	Afghanistan	Yakabagh, Takhar	Earthquake (seismic activity)	Earthquake (ground shaking)	1	1001	2003- 0236
30112005	3122005	Albania	Vlora, Fie, Gjirokaster	Flood	General flood	3	500	2005-0696
4122004	8122004	Albania	Obot, Shkodre prefecture	Flood	General flood		2500	2004-0633
21092002	10102002	Albania	Lezha, Shkoder regions	Flood	General flood	1	66884	2002-0607
20121997	23121997	Albania	Lezhe	Flood	Storm		8000	1997-0302
27121995	27121995	Albania	Shkadra, Malesi, Modhe	Flood	General flood		2000	1995-0300
20091995	20091995	Albania	Laci, Rrogozhina, Lushnja	Flood		4	1500	1995-0234
17111992	19111992	Albania	Kruja, Lac, Lezhe, Shkodor	Flood	Flash flood	11	35000	1992-0160
7121991	7121991	Albania	Fushe Arrez	Industrial Accident	Fire	60	150000	1991-0395


# Test your knowledge

## ■ What about this set?

Date	Line	Type	Factory Location	Brand	Branch	Item	Faulty	Sales (in thousands)	Shop ID
102008	102080	Kitchen	Rumilly, France	Tefal	Tefal-FR	Airfryer		2800	2008-9475
72006	102060	Kitchen	Rumilly, France	Tefal	Tefal-FR	Frying pan		1900	2006-9570
52000	102020	Kitchen	Rumilly, France	Tefal	Tefal-FR	Kettle	37	25	2000-9186
81971	111973	Kitchen	Montpellier, France	Tefal	Tefal-FR	Airfryer			1971-9085
11969	111969	Kitchen	Montpellier, France	Tefal	Tefal-FR	Frying pan		480	1969-9007
29072006	290720	Kitchen	Osaka, Japan	Panasonic	Panasonic-JP	Rice cooker	1	935	2006-0405
13122005	131220	Kitchen	Selangor, Malaysia	Panasonic	Panasonic-MY	Rice cooker	5	501	2005-0686
8102005	180720	Kitchen	Binh Duong, Vietnam	Panasonic	Panasonic-VN	Rice cooker	1		2005-0575
18072004	281020	Kitchen	Osaka, Japan	Panasonic	Panasonic-JP	Kettle	2	1040	2004-0436
10042003	131420	Kitchen	Selangor, Malaysia	Panasonic	Panasonic-MY	Kettle	1	1001	2003- 0236
30112005	3122005	Home appliance	Seoul, South Korea	Samsung	Samsung-SK	Washing machine	3	500	2005-0696
4122004	8122004	Home appliance	Seoul, South Korea	Samsung	Samsung-SK	Refrigerator		2500	2004-0633
21092002	3102002	Home appliance	Seoul, South Korea	Samsung	Samsung-SK	Vacuum Cleaner	1	66	2002-0607
20121997	23121997	Home appliance	Thai Nguyen, Vietnam	Samsung	Samsung-VN	Washing machine	60	8000	1997-0302
27121995	27121995	Home appliance	Thai Nguyen, Vietnam	Samsung	Samsung-VN	Refrigerator		2000	1995-0300
20091995	91995	Home appliance	Uttar Pradesh, India	Samsung	Samsung-IN		4	1500	1995-0234
17111992	91992	Home appliance	Uttar Pradesh, India	Samsung	Samsung-IN	Vacuum Cleaner	11	35000	1992-0160
7121991	290710	Home appliance	Osaka, Japan	Panasonic	Panasonic-JP	Washing machine		1500	1991-0395

# Chapter 3: Data Preprocessing

---

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration 
- Data Reduction
- Data Transformation and Data Discretization
- Summary

# Data Integration

---

- **Data integration:**
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g.,  $A.cust-id \equiv B.cust-\#$ 
  - Integrate metadata from different sources
- **Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

# Handling Redundancy in Data Integration

---

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

# Correlation Analysis (Nominal Data)

---

- **X<sup>2</sup> (chi-square) test**

$$\chi^2 = \sum \frac{(\textit{Observed} - \textit{Expected})^2}{\textit{Expected}}$$

- The larger the X<sup>2</sup> value, the more likely the variables are related
- The cells that contribute the most to the X<sup>2</sup> value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population



# Chi-square

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

## Contingency Table

Observed values	Play Chess	¬ Play Chess	Total
Like Science Fiction	250	200	450
¬ Like Science Fiction	50	1000	1050
<b>Total</b>	<b>300</b>	<b>1200</b>	<b>1500</b>

Explanatory variable	Response variable		Expected values
	Category 1	Category 2	
Category of interest	$(T_A \times T_1)/T$	$(T_A \times T_2)/T$	$T_A$
Baseline category	$(T_B \times T_1)/T$	$(T_B \times T_2)/T$	$T_B$
	$T_1$	$T_2$	$T$

# Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- $\chi^2$  (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

- It shows that like\_science\_fiction and play\_chess are correlated in the group

# Covariance (Numeric Data)

- Covariance is similar to correlation

$$\text{Cov}(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

where  $n$  is the number of tuples, and  $\bar{A}$  and  $\bar{B}$  are the respective mean or **expected values** of  $A$  and  $B$ ,  $\sigma_A$  and  $\sigma_B$  are the respective standard deviation of  $A$  and  $B$ .

- **Positive covariance:** If  $\text{Cov}_{A,B} > 0$ , then  $A$  and  $B$  both tend to be larger than their expected values.
- **Negative covariance:** If  $\text{Cov}_{A,B} < 0$  then if  $A$  is larger than its expected value,  $B$  is likely to be smaller than its expected value.
- **Independence:**  $\text{Cov}_{A,B} = 0$  but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

# Co-Variance: An Example

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week:  
(2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - $E(A) = (2 + 3 + 5 + 4 + 6) / 5 = 20 / 5 = 4$
  - $E(B) = (5 + 8 + 10 + 11 + 14) / 5 = 48 / 5 = 9.6$
  - $Cov(A, B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$
- Thus, A and B rise together since  $Cov(A, B) > 0$ .

# Correlation Analysis (Numeric Data)

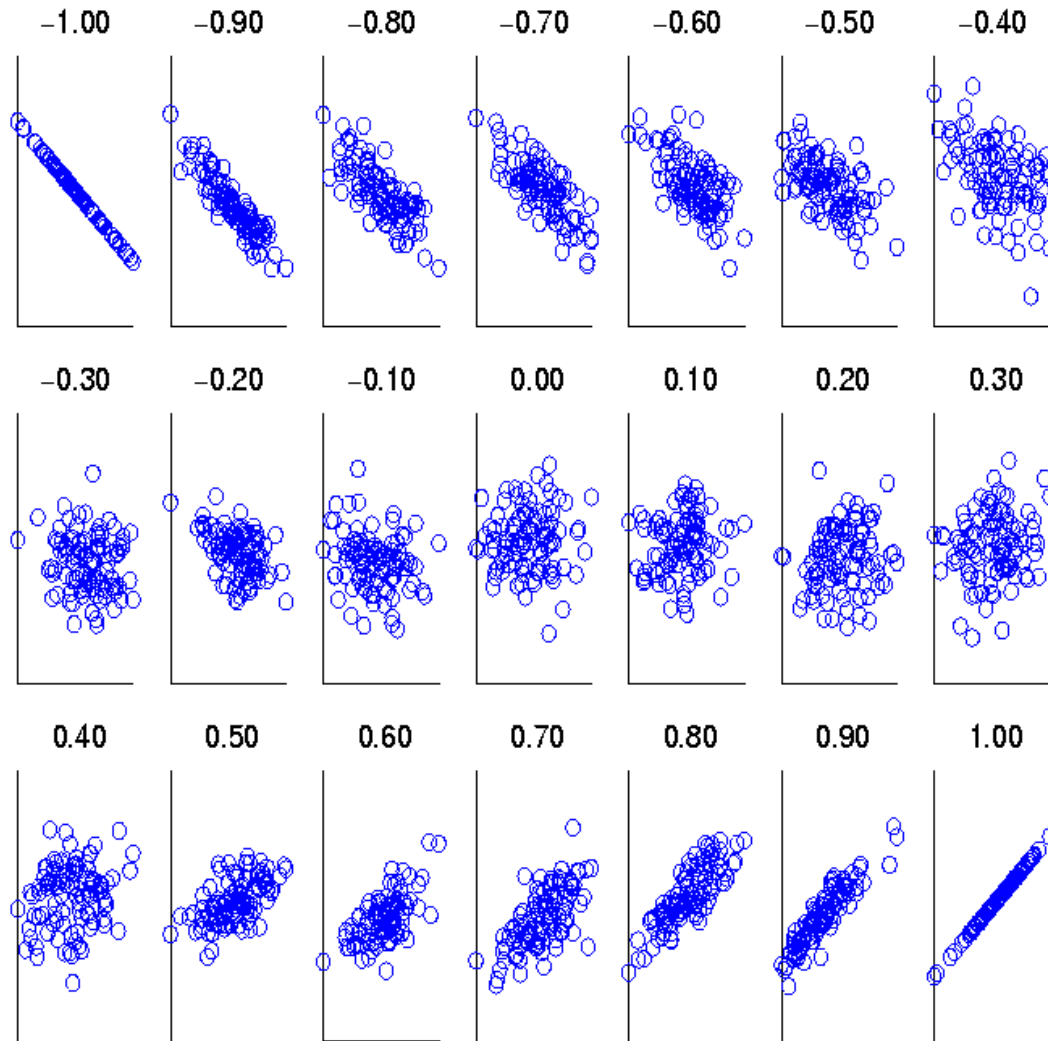
- Correlation coefficient (also called **Pearson's product moment coefficient**)

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B} \quad r_{A,B} = \frac{Cov(A, B)}{\sigma_A\sigma_B}$$

where n is the number of tuples,  $\bar{A}$  and  $\bar{B}$  are the respective means of A and B,  $\sigma_A$  and  $\sigma_B$  are the respective standard deviation of A and B, and  $\sum(a_i b_i)$  is the sum of the AB cross-product.

- If  $r_{A,B} > 0$ , A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$ : independent;  $r_{A,B} < 0$ : negatively correlated
  - Independent variables will have  $r_{A,B} = 0$ , but  $r_{A,B} = 0$  not conclusively independent

# Visually Evaluating Correlation



**Scatter plots  
showing the  
similarity from  
-1 to 1.**

# Pearson's Correlation: An Example

X	3.4	3.8	4.1	2.2	2.6	2.9	2	2.7	1.9	3.4
Y	5.5	5.9	6.5	3.3	3.6	4.6	2.9	3.6	3.1	4.9

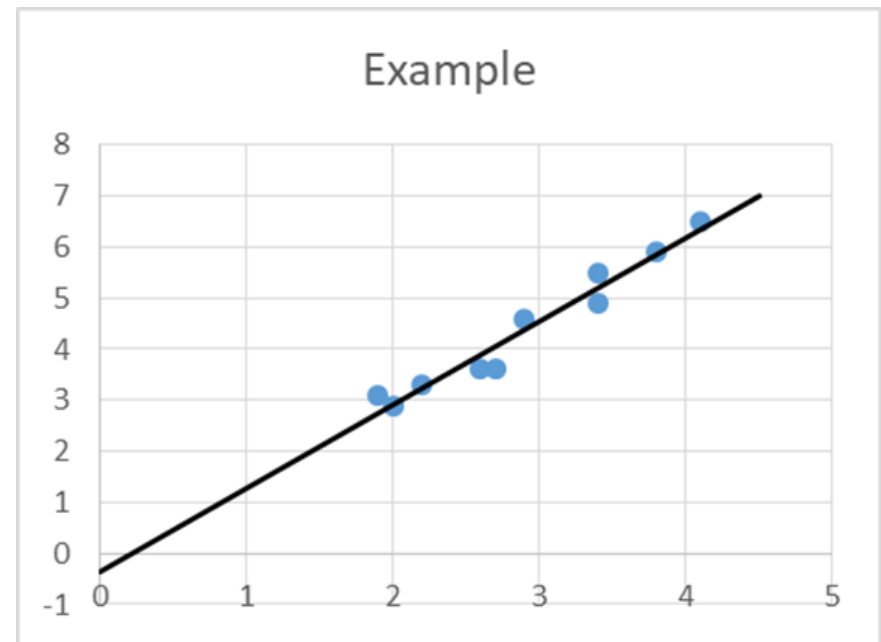
- Find the correlation  $r$ . What does the value imply?

$$\bar{x} = 2.9, \sigma_x = 0.759, \bar{y} = 4.39, \sigma_y = 1.273$$

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

=

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} =$$



# Correlation (viewed as linear relationship)

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects,  $A$  and  $B$ , and then take their dot product

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A' \bullet B'$$



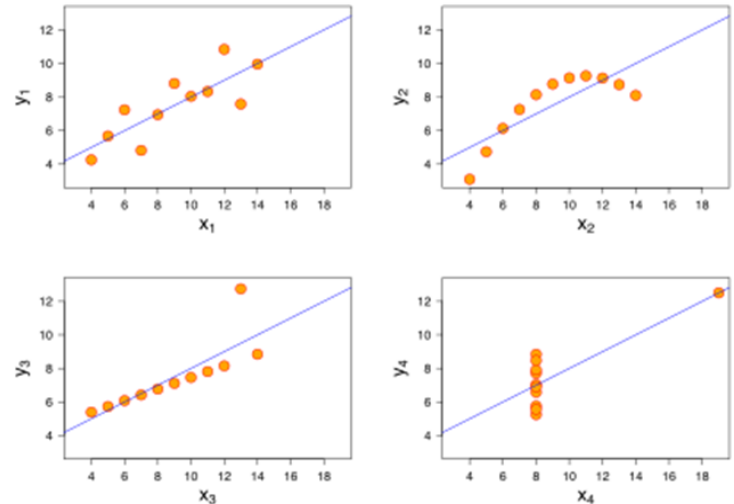
# Pearson's Drawback

- Correlation may not make sense
  - For example: Anscombe's Quartet (1973)
    - 4 datasets with identical statistics but have very different distributions when visualized

Note: use visualizations to check for non-linearity or outliers

Data set	1-3	1	2	3	4	4
Variable	x	y	y	y	x	y
Obs. no. 1 :	10.0	8.04	9.14	7.46	8.0	6.58
2 :	8.0	6.95	8.14	6.77	8.0	5.76
3 :	13.0	7.58	8.74	12.74	8.0	7.71
4 :	9.0	8.81	8.77	7.11	8.0	8.84
5 :	11.0	8.33	9.26	7.81	8.0	8.47
6 :	14.0	9.96	8.10	8.84	8.0	7.04
7 :	6.0	7.24	6.13	6.08	8.0	5.25
8 :	4.0	4.26	3.10	5.39	19.0	12.50
9 :	12.0	10.84	9.13	8.15	8.0	5.56
10 :	7.0	4.82	7.26	6.42	8.0	7.91
11 :	5.0	5.68	4.74	5.73	8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.



Property	Value
Mean x	9.0
Mean y	7.5
Variance x	11
Variance y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3 + 0.5x$

# Spearman's Rank-Order Correlation

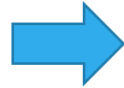
---

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad \begin{array}{l} d_i = \text{difference in paired ranks} \\ n = \text{number of cases} \end{array}$$

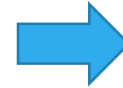
- $-1 \leq r_s \leq +1$
- $r_s = +1 \rightarrow$  perfect positive association of ranks
- $r_s = 0 \rightarrow$  no association between ranks
- $r_s = -1 \rightarrow$  perfect negative association of ranks
- The closer  $r_s$  is to 0, the weaker the association between ranks

# Spearman's Correlation: An Example

ID	Mark	
	Stats	History
01	56	66
02	75	70
03	45	40
04	71	60
05	62	65
06	64	56
07	58	59
08	80	77
09	76	67
10	61	63



Rank	
Stats	History
9	4
3	2
10	10
4	7
6	5
5	9
8	8
1	1
2	3
7	6



Difference	
$d$	$d^2$
5	25
1	1
0	0
-3	9
1	1
-4	16
0	0
0	0
-1	1
1	1
54	


$d_i$  = difference in paired ranks  
 $n$  = number of cases

$$\begin{aligned}
 r_s &= 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \\
 &= 1 - \frac{6(54)}{990} \\
 &= 1 - 0.33 \\
 &= 0.67
 \end{aligned}$$

$\therefore$  There is a positive association of ranks for the marks obtained in the stats and history exam

# Spearman's Correlation: An Example

- If there are same/tied ranks: Get average rank.

ID	Mark			Rank	
	Stats	History		Stats	History
01	56	66		9	4
02	75	70		3	2
03	45	40		10	10
04	71	60		4	7
05	61	65		6.5	5
06	64	56		5	9
07	58	59		8	8
08	80	77		1	1
09	76	67		2	3
10	61	63		6.5	6

Use full Spearman's formula:

$$r_s = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

$x_i, y_i$  = ranks

$\bar{x}, \bar{y}$  = mean ranks

# Summary

---

- **Correlation analysis:**
  - Find redundancies
  - Plot/inspect the correlation matrix
- **Pairs with strong (positive/negative) relations**
  - Good candidate for data reduction by removing one of the variables
  - E.g. pairs with correlation coefficient higher than a threshold
  - Remove one of the variables
- **Pearson's correlation**
  - Evaluates the linear relationship between two numerical variables
- **Spearman's rank-order correlation**
  - Evaluates the monotonic relationship between two numerical/ordinal variables

# References

---

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Comm. of ACM*, 42:73-78, 1999
- A. Bruce, D. Donoho, and H.-Y. Gao. Wavelet analysis. *IEEE Spectrum*, Oct 1996
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003
- J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data*. Duxbury Press, 1997.
- H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. *VLDB'01*
- M. Hua and J. Pei. Cleaning disguised missing data: A heuristic approach. *KDD'07*
- H. V. Jagadish, et al., *Special Issue on Data Reduction Techniques*. *Bulletin of the Technical Committee on Data Engineering*, 20(4), Dec. 1997
- H. Liu and H. Motoda (eds.). *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer Academic, 1998
- J. E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, *VLDB'2001*
- T. Redman. *Data Quality: The Field Guide*. Digital Press (Elsevier), 2001
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995