

Lecture 2

Data Cleaning & Integration

Dr. Amira Rezk

dramirarezk@gmail.com

Sara S. Elhishi

sarashaker161@gmail.com

Information Systems Dept.

The Knowledge Discovery Process



DATA CLEANING
remove noise



DATA INTEGRATION
FROM MULTIPLE
SOURCES



DATA SELECTION
DATA RELEVANT
TO THE ANALYSIS
TASK



DATA TRANSFORMATION
INTO FORMS
APPROPRIATE FOR
MINING



DATA MINING
EXTRACT DATA
PATTERNS



PATTERN EVALUATION
INTERESTINGNESS
MEASURES



KNOWLEDGE PRESENTATION
VISUALIZATION
TO USERS

Data Preprocessing

- Datasets are highly susceptible to noisy, missing, and inconsistent data.
- Low-quality data will lead to low-quality mining results.
- Data quality factors includes:
 - accuracy,
 - completeness,
 - consistency,
 - timeliness,
 - believability, and
 - interpretability.

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Reasons of Low Data Quality

Inaccurate data

- Having incorrect attribute values (e.g., by choosing the default value “January 1” displayed for birthday)

Incomplete data

- Missing data (may not always be available or of interest)

Inconsistent data

- different assessments of the quality depends on the intended use of the data

Timeliness

- (e.g. month-end data are not updated in a timely fashion has a negative impact on the data quality.)

Believability

- reflects how much the data are trusted by users

Interpretability

- reflects how easy the data are understood (e.g. sales codes)

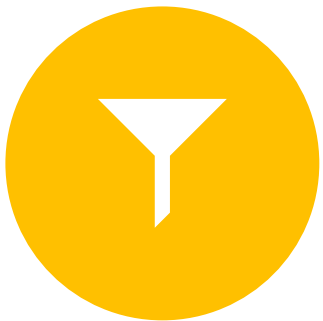
Preprocessing Tasks That Improve Data Quality



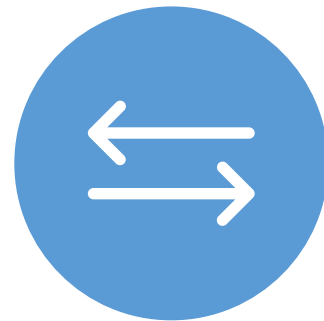
Data cleaning: missing values, noisy data, outliers



Data integration: data from multiple sources



Data reduction: reduced representation of the data set



Data transformation: data scaled to fall within a smaller range like 0.0 to 1.0

Data Cleaning

- Missing: empty, null, NaN, or Na (e.g. Occupation=" ")
- Noisy: contain errors (e.g. salary = "-1000")
- Inconsistent: (e.g. rating was "1, 2, 3", now rating "A, B, C")
- Intentional: (e.g., by choosing the default value "January 1" displayed for birthday)



Handling Missing Values

Ignore

Ignore the tuple

Fill in

Fill in the missing value manually

Use

Use a global constant to fill in the missing value

Use

Use a measure of central tendency for the attribute to fill in the missing value

Use

Use the attribute mean or median for all samples belonging to the same class as the given tuple

Use

Use the most probable value to fill in the missing value

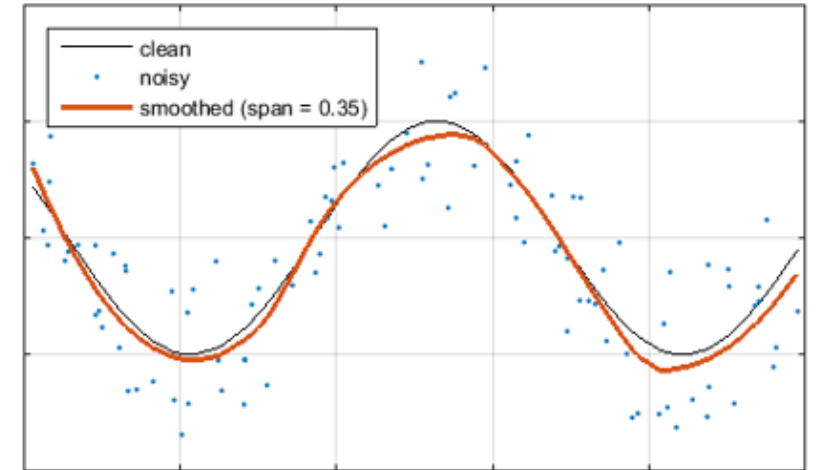


Note

- A missing value may not imply an error in the data!
- For example, when applying for a credit card, candidates may be asked to supply their driver's license number. Candidates who do not have a driver's license may naturally leave this field blank.
- Forms should allow respondents to specify values such as "not applicable." Software routines may also be used to uncover other null values (e.g., "don't know," "?" or "none")

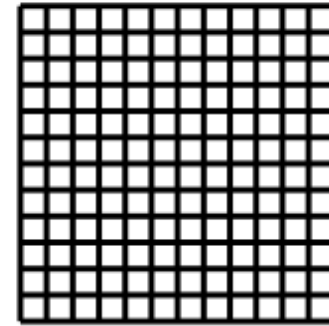
Noisy Data

- Noise is a random error or variance in a measured variable.
- “smooth” out the data to remove the noise .
- Smoothing Techniques:
 - Binning
 - Regression
 - Outlier Analysis (comes later)

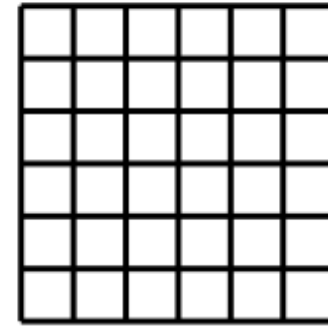




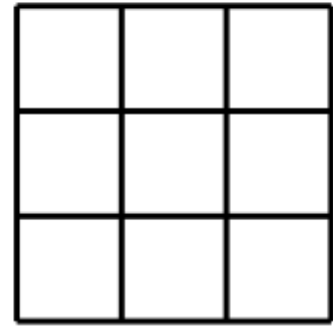
1. Binning



No binning
(144 pixels)



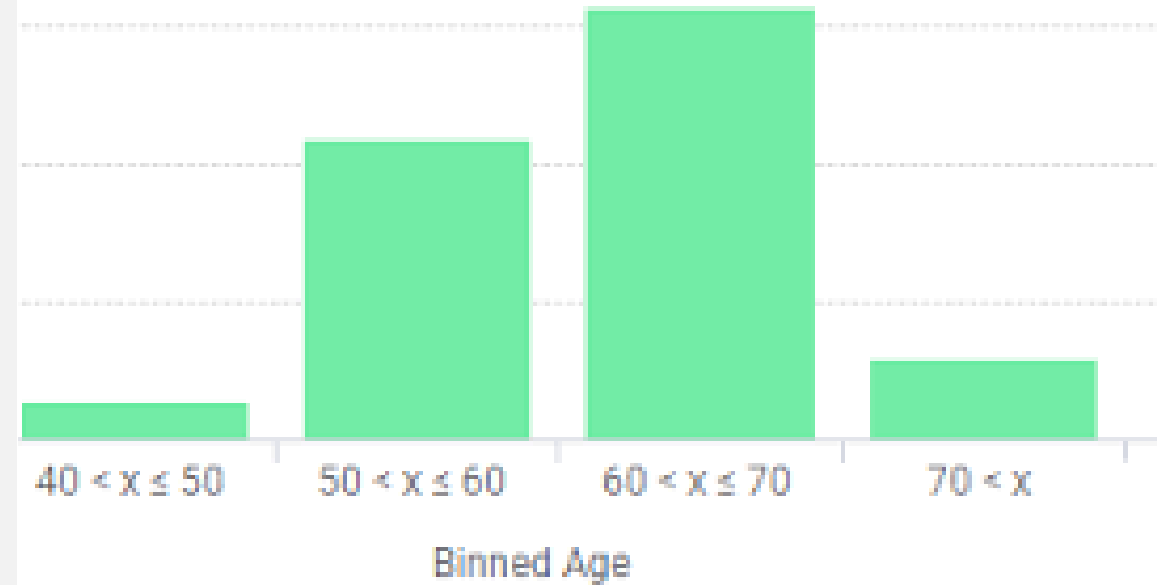
2x binning
(36 pixels)



4x binning
(9 pixels)

- The original data values are divided into “buckets” known as “bins” and then they are replaced by a general value calculated for that bin.
- In the context of image processing, binning is the procedure of combining a cluster of pixels into a single pixel.

Binning



- “Partition” sorted data by 2 methods:
 - Equal depth (frequency) bins: each bin has same number of values
 - Equal width bins: interval range of values per bin is equal
- “Smooth” each bin by:
 - **bin means**: each bin value is replaced by the bin mean
 - **bin medians**: each bin value is replaced by the bin median
 - **bin boundaries**: each bin value is replaced by the closest boundary value (min & max in a bin are bin boundaries)

Example



- Use smoothing to smooth the Age data, Partition them into three bins by
 - equal-frequency and
 - equal-width Partitioning
- Age: 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 56, 57, 58, 60, 61

Partition

- Age data are first sorted,
- 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 56, 57, 58, 60, 61
- Partitioned into three equal frequency bins of size 6
- Or Partitioned into 3-equal interval bins

Partition into (equal-frequency
bins:

Bin 1: 23, 23, 27, 27, 39, 41

Bin 2: 47, 49, 50, 52, 54, 54

Bin 3: 56, 56, 57, 58, 60, 61

Partition into (equal-width)
bins:

Bin 1: 23, 23, 27, 27

Bin 2: 39, 41, 47, 49, 50

Bin 3: 52, 54, 54, 56, 56, 57, 58,
60, 61

Smoothing (on Equal-frequency)

- In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
- In smoothing by bin boundaries, each bin value is then replaced by the closest boundary value.

Smoothing by bin boundaries:

Bin 1: 23, 23, 23, 23, 41, 41

Bin 2: 47, 47, 47, 54, 54, 54

Bin 3: 56, 56, 56, 56, 61, 61

Smoothing by bin means:

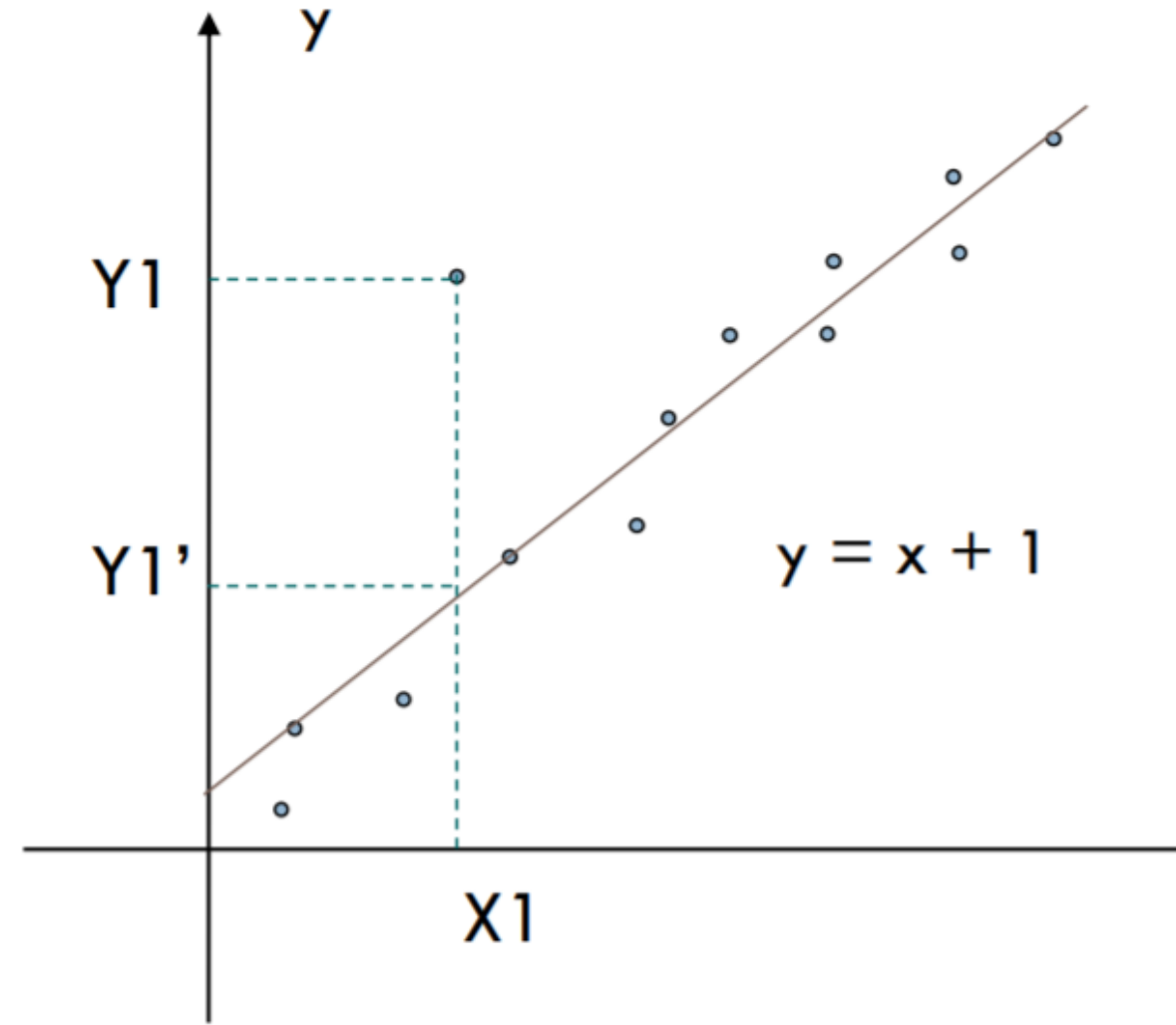
Bin 1: 30, 30, 30, 30, 30, 30

Bin 2: 51, 51, 51, 51, 51, 51

Bin 3: 58, 58, 58, 58, 58, 58,

2. Regression

- Linear regression involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.
- Replace noisy or missing values by predicted values.



Half-Time

—

Notebook review

Data Integration

- Merging of data from multiple data stores.
- Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set



Entity Identification Problem

- How can equivalent real-world entities from multiple data sources be matched up?
- For example, how can the data analyst or the computer be sure that customer id in one database and cust_number in another refer to the same attribute?
- **metadata** can be used to help avoid errors in schema integration.





Redundancy and Correlation Analysis

- An attribute may be redundant if it can be “derived” from another attribute or set of attributes.
- annual revenue, for instance
- Correlation analysis, given two attributes, such analysis can measure how strongly one attribute implies the other.
- There are 2 test:
 - Chi-Square for Nominal Data
 - Covariance for Numeric

Correlation Test for Nominal Data

- a correlation relationship between two attributes, A and B, can be discovered by a χ^2 (chi-square) test
- Where o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j) which can be computed as
- tests the hypothesis that A and B are independent, that is, there is no correlation between them.

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n},$$

Example

- Determine whether there is a relationship between gender and getting in trouble at school (both nominal variables).
- The **null hypothesis** is that the two variables are independent (i.e. no relationship or correlation) and
- the **research hypothesis** is that the two variables are related.

	Got in Trouble	Did Not Get in Trouble	Total
Boys	46	71	117
Girls	37	83	120
Total	83	154	237

Calculate
the
expected
frequency
for each cell

- For example $e_{11} = \frac{117 \times 83}{237} = 40.97$
- We do the same thing for the other three cells and end up with the following expected counts (in parentheses next to each raw score):

	Got in Trouble	Did Not Get in Trouble	Total
Boys	46 (40.97)	71 (76.02)	117
Girls	37 (42.03)	83 (77.97)	120
Total	83	154	237

Calculate the chi- square statistic

- For each cell, we square the difference between the observed frequency and the expected frequency and divide that number by the expected frequency. Then we add all of the terms
- note: a chi-square statistic can't be negative

$$\chi^2 = \frac{(46-40.97)^2}{40.97} + \frac{(37-42.03)^2}{42.03} + \frac{(71-76.03)^2}{76.03} + \frac{(83-77.97)^2}{77.97}$$

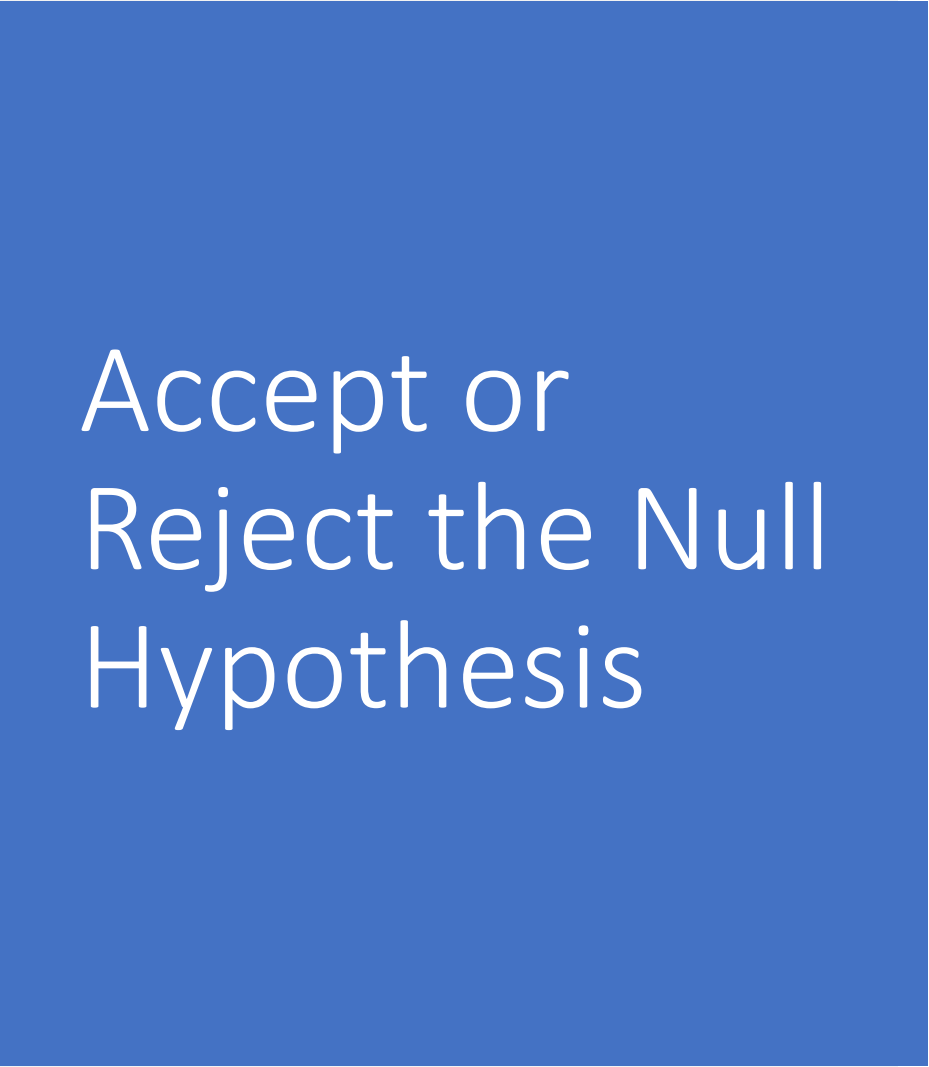
$$\chi^2 = 1.87$$

- before we can come to a conclusion, we need to find our critical statistic, which entails finding our degrees of freedom.

Compute critical statistic

- $DF = (\#rows - 1)(\#cols - 1) = (2 - 1)(2 - 1) = 1$
- Compare our obtained statistic (1.87) to our critical statistic found on the chi-square table
- The critical statistic for an alpha level of 0.05 and one degree of freedom is 3.84

	P										
DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322



Accept or Reject the Null Hypothesis

- Critical (3.841) > obtained (1.87)
- Because the critical statistic is greater than our obtained statistic, we Accept (can't reject) our null hypothesis.
- There is no relationship between gender and getting in trouble at school

Correlation Coefficient for Numeric Data

- The correlation coefficient between two attributes, A and B, is
- If $r_{A,B}$ is *greater* than 0, then A and B are *positively* correlated, The higher the value, the stronger the correlation
- If $r_{A,B} = 0$, then A and B are *independent*
- Note that Correlation does not imply causality! That is, if A and B are correlated, this does not necessarily imply that A causes B or that B causes A.

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

Covariance

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

- Measures how two things change together .
- *Covariance is +ve* $\rightarrow A$ & B change together, and if $A > \bar{A}$ then $B > \bar{B}$
- *Covariance is -ve* \rightarrow one is above its mean and one is below
- If A and B are independent $\rightarrow \text{Covariance} = 0$



Example

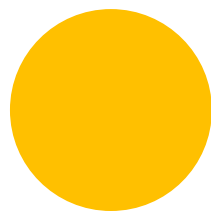
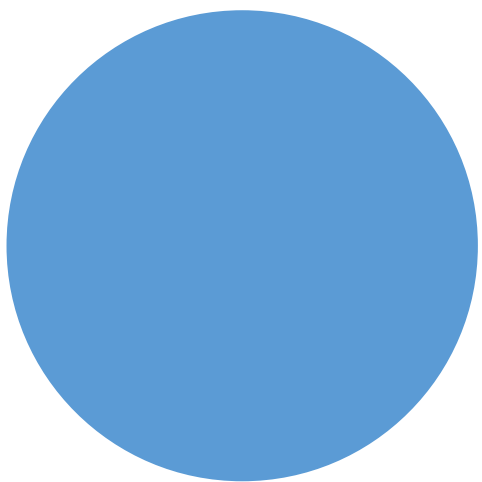
- Assume that you gathered data on six different days
- Compute Covariance

Temperature	No. of Customers
98	15
87	12
90	10
85	10
95	16
75	7



First find the mean of each variable. Then $E(X,Y)$, Finally the $Cov(X,Y)$

- X (Temperature) and Y (No. of Customers)
- the mean of x is $(98+87+90+85+95+75)/6= 88.33$.
The mean of y is $(15+12+10+10+16+7)/6= 11.67$
- $E(X.Y) = \frac{(98 \times 15) + (87 \times 12) + \dots + (75 \times 7)}{6} = 1051.5$
- $Cov(X,Y) = E(X.Y) - \bar{X}\bar{Y} = 1051.5 - (88.33 * 11.67) = 20.94$
- The number is positive, so we can state that the two variables do have a positive relationship; as temperature rises, the number of customers in the store also rises.



Thanks

