Data Preprocessing
Data Integration

Data Integration

- Data integration is the process of combining the data from multiple sources into a coherent data store
- These sources may include multiple databases or flat files
- Example:
  - Temperature sensor, pressure sensor and rain gauge records temperature, atmospheric pressure and amount of rain at different locations
  - Each location has separate temperature, pressure and amount of rain tables (database)

Data Integration

- Data integration is the process of combining the data from multiple sources into a coherent data store
- These sources may include multiple databases or flat files
- Example:
  - Temperature sensor, pressure sensor and rain guage records t… …nount of rain at different …
  - Each loc… …ıre, pressure and amount c…

1	Dates	Station Id	Temperature	Humidity	Rain
2	08-07-2018	t10	25.46875	82.1875	6.75
3	09-07-2018	t10	26.19298	85.42	1762
4	10-07-2018	t10	25.17021	85.34043	652.5
5	11-07-2018	t10	25.1368	87.68657	963
6	08-07-2018	t11	23.53846	61.92308	3
7	09-07-2018	t11	26.8494	85.42	15
8	10-07-2018	t11	25.1368	75.07463	13583
9	11-07-2018	t11	27.35915	76.02113	19769
10	23-07-2018	t12	25.1368	94.4065	1071
11	24-07-2018	t12	24.16197	97.66901	438.8
12	25-07-2018	t12	25.29323	94.84211	13667
13	26-07-2018	t12	22.19718	99	864

Data Integration

- Data integration is the process of combining the data from multiple sources into a coherent data store
- These sources may include multiple databases or flat files
- Example:
  - Temperature sensor, pressure sensor and rain guage records temperature, pressure and amount of rain at different locations
  - Each location has separate temperature, pressure and amount of rain tables (database)

- Issues to consider during data integration:
  - Schema integration (entity matching)
  - Data value conflict
  - Redundancy

Schema Integration (Entity Matching)

- Database schema: The organization of data as a blueprint of how the database is constructed
- Entity: Each entity in real-world problem is the attribute in the database
- Addresses the question of
  - "*how can equivalent real-world entities from multiple sources be matched up?*"
  - "*how can data analysts be sure that they are same?*"
- Attribute name conflict across the multiple sources of data
  - Example: `customer_id`, `customer_num`, `cust_num`
- Entity identification problem:
  - Metadata is associated with each attribute
  - Matadata include:
    - Name, Meaning, Data type, Range of values permitted

Data Value Conflict

- Issue: Detection and resolution of data value conflicts
- For the same real-world entity, attribute values from different sources may differ
- This may be due to difference in representation, scaling, or encoding
- Example:
  - "`weight`" attribute may be stored in metric unit (gram, kilogram, etc.) in one system, British imperial unit (pound, ounce, etc.) in another system
  - In a database for hotel chain in different countries:
    - "`price of room`" attribute may be stored with price value in different currencies
  - Categorical data: "`gender`" may be stored with male and female or M and F

Redundancy

- Major issue to be addressed
- Sources of redundancy:
  - An attribute may be redundant, if it can be derived from another attribute or set of attributes
    - Example: Attribute "`Total Marks`" derived from Marks from each courses
  - Inconsistency in the attribute naming can also cause redundancy in resulting data sets
    - Example: (1) `registration_id` and `roll_num`
      (2) `customer_id` and `customer_num`
- Two types of redundancies:
  - Redundancy between the attributes
  - Redundancy at the tuple level
    - Duplication of tuples
    - Remove the duplicate tuples

Redundancy Between Attributes

- Two attributed may be related or dependent
- Detected by the correlation analysis
- Correlation analysis measures how strongly one attribute implies (related) to other, based on available data
- Correlation analysis for numerical attributes:
  - Compute correlation coefficient between two attributes `A` and `B` (e.g. Pearson's product moment coefficient i.e. Pearson's correlation coefficient)
- Correlation analysis for categorical attributes:
  - Correlation relationship between two categorical attributes `A` and `B` can be discovered by χ^2 (chi-square) test

Redundancy Between Numerical Attributes

- Pearson's correlation coefficient ($\rho_{A,B}$):

$$\rho_{A,B} = \frac{\frac{1}{N}\sum_{i=1}^{N}(a_i - \mu_A)(b_i - \mu_B)}{\sigma_A \sigma_B} = \frac{\mathrm{Cov}(A, B)}{\sigma_A \sigma_B}$$

  - N : number of tuples
  - a_i and b_i: respective values of attribute A and attribute B in tuple i
  - μ_A and μ_B: respective mean values of A and B
  - σ_A and σ_B : respective standard deviation of A and B
  - $\mathrm{Cov}(A, B)$: Covariance between A and B

- Note: $-1 \le \rho_{A,B} \le +1$

Redundancy Between Numerical Attributes: Pearson's correlation coefficient

- If $\rho_{A,B}$ is greater than 0, then attributes A and B are positively correlated
  - The values of A increases as the values of B increases or vice versa
  - The higher the value, the stronger the correlation
  - A higher correlation value may indicate that A (or B) may be removed as a redundancy
- If $\rho_{A,B}$ is equal to 0, then attributes A and B have no correlation between them (may be independent)
- If $\rho_{A,B}$ is less than 0, then attributes A and B are negatively correlated
  - The values of A increases as the values of B decreases or vice versa
  - Each attribute discourages the other

Redundancy Between Numerical Attributes: Pearson's correlation coefficient

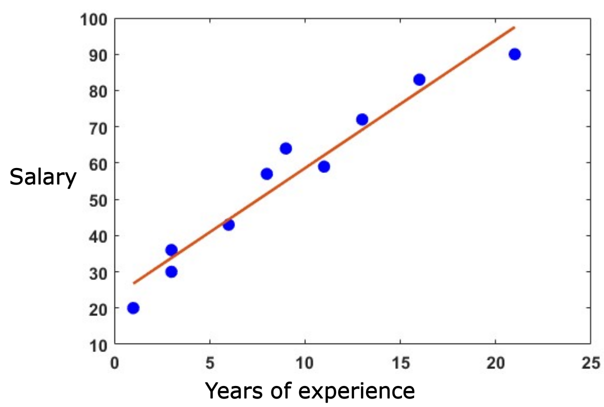
- Assumption:
  - Both attributes (variables) should be normally distributed (normally distributed variables have a bell-shaped curve)
  - Linearity: The two attributes have linear relationship
  - Homoscedasticity: Data is equally distributed about the regression line.
- Scatter plots can also be use to view correlation between the numerical attributes

Illustration of Pearson's Correlation Coefficient

Years of experience (x)	Salary (in Rs 1000) (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

$\rho_{A,B} = 0.97$

- Scatter plot



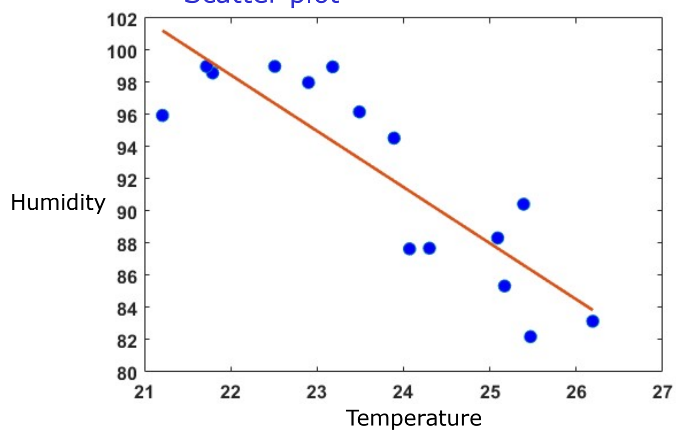
- The two attributes have linear relationship
- Data is equally distributed about the regression line (roughly)

12

Illustration of Pearson's Correlation Coefficient

Temp (x)	Humidity (y)
25.47	82.19
26.19	83.15
25.17	85.34
24.30	87.69
24.07	87.65
21.21	95.95
23.49	96.17
21.79	98.59
25.09	88.33
25.39	90.43
23.89	94.54
22.51	99.00
22.90	98.00
21.72	99.00
23.18	98.97

- Scatter plot



$\rho_{A,B}$ = -0.8648

13

Redundancy Between Numerical Attributes: Spearman Rank Correlation

- Spearman rank correlation is a non-parametric measure of rank correlation between two attributes (variables)
- Rank correlation between variables: Statistical dependence between the rankings of two variables
  - The values the variables take should be at least ordinal
- The values in the attributes should be converted into ranks of the values (ordinal values), if the attribute is not ordinal
- As it is non-parametric measure, it does not carry any assumptions about the distribution of the data
- The Spearman correlation coefficient is defined as the *Pearson correlation coefficient* between the rank variables

Redundancy Between Numerical Attributes: Spearman Rank Correlation

- Spearman correlation coefficient (ρ_{R_A, R_B}):

$$\rho_{R_A,R_B} = \frac{Cov(R_A, R_B)}{\sigma_{R_A}\sigma_{R_B}}$$

  - R_A and R_B : ranks attribute A and attribute B
  - σ_{R_A} and σ_{R_B} : respective standard deviation of ranks of A and B
  - $Cov(R_A, R_B)$: Covariance between the ranks of A and B
- Only if all N ranks are *distinct integers*, then it can be computed using the popular formula

$$\rho_{R_A,R_B} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$$

  - N : number of tuples
  - d_i: difference between the rank values of A and B in tuple i

$$-1 \le \rho_{R_A,R_B} \le +1$$

Redundancy Between Numerical Attributes: Spearman Rank Correlation

- Pearson's correlation assesses linear relationships
- Spearman's correlation assesses monotonic relationships (whether linear or not)

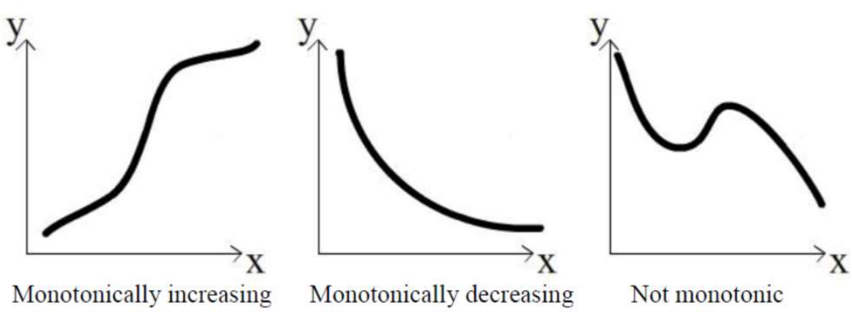


Illustration of Spearman's Correlation Coefficient

Years of experience (x)	Salary (in Rs 1000) (y)	R_x	R_y
3	30	2	2
8	57	4	5
9	64	5	7
13	72	7	8
3	36	2	3
6	43	3	4
11	59	6	6
21	90	9	10
1	20	1	1
16	83	8	9

$\rho_{R_x, R_y} = 0.9806$

- Convert the values of both attribute into rank values

17

Illustration of Spearman's Correlation Coefficient

Temp (x)	Humidity (y)	R_x	R_y
25.47	82.19	14	1
26.19	83.15	15	2
25.17	85.34	12	3
24.30	87.69	10	5
24.07	87.65	9	4
21.21	95.95	1	9
23.49	96.17	7	10
21.79	98.59	3	12
25.09	88.33	11	6
25.39	90.43	13	7
23.89	94.54	8	8
22.51	99.00	4	14
22.90	98.00	5	11
21.72	99.00	2	14
23.18	98.97	6	13

$\rho_{R_x, R_y} = -0.8523$

18