Lecture 3
Data Reduction & Transformation

Dr. Amira Rezk
dramirarezk@gmail.com
Sara S. Elhishi
sarashaker161@gmail.com
Information Systems Dept.

Data Reduction

More is not always better.

Obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

Dimensionality Reduction	Numerosity Reduction	Data Compression.

Dimensionality Reduction

- Reduce the number of attributes under consideration
- Methods include:
  - *wavelet transforms*
  - *principal components analysis (PCA),*
  - *Attribute subset selection*

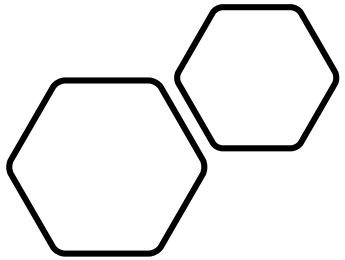
Numerosity Reduction Techniques

- Data are replaced or estimated by alternative.
- **parametric methods**, a model is used to estimate the data (PCA)
- **Nonparametric methods** histograms, clustering, sampling, and data cube aggregation

Data Compression

- Reducing the amount of capacity required to store data.
- *lossless* : No loss of information (e.g. Text)
- Lossy: the size of the file is reduced by eliminating data in the file (e.g. Image)

Take a Closer Look …



Explain some methods in Details

Attribute Subset Selection

- How can we find a ‘good’ subset of the original attributes?
- For n attributes, there are 2^n possible subsets!!!
- Solution: Heuristic (Greedy) methods
  - while searching for attribute subsets, they always make what looks to be the best choice at the time.

Heuristic :
Stepwise forward selection

- Start with empty set of attributes as reduced set.
- The best of the attributes is determined and added to the reduced set.
  - “best” is determined by some pre-determined criteria

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

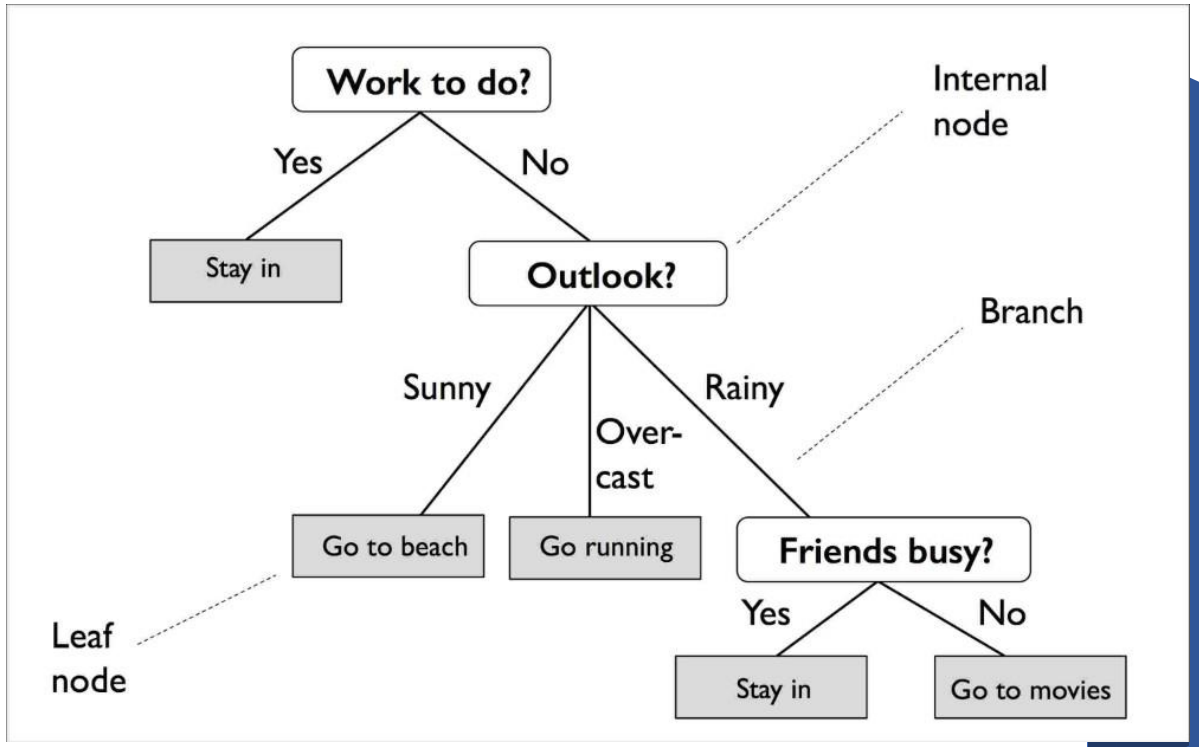
Initial reduced set:
$\{\}$
$\Rightarrow \{A_1\}$
$\Rightarrow \{A_1, A_4\}$
$\Rightarrow$ Reduced attribute set:
$\{A_1, A_4, A_6\}$

Heuristic :
Stepwise backward selection

- start with the full set of attributes.
- At each step, remove the worst attribute remaining in the set

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

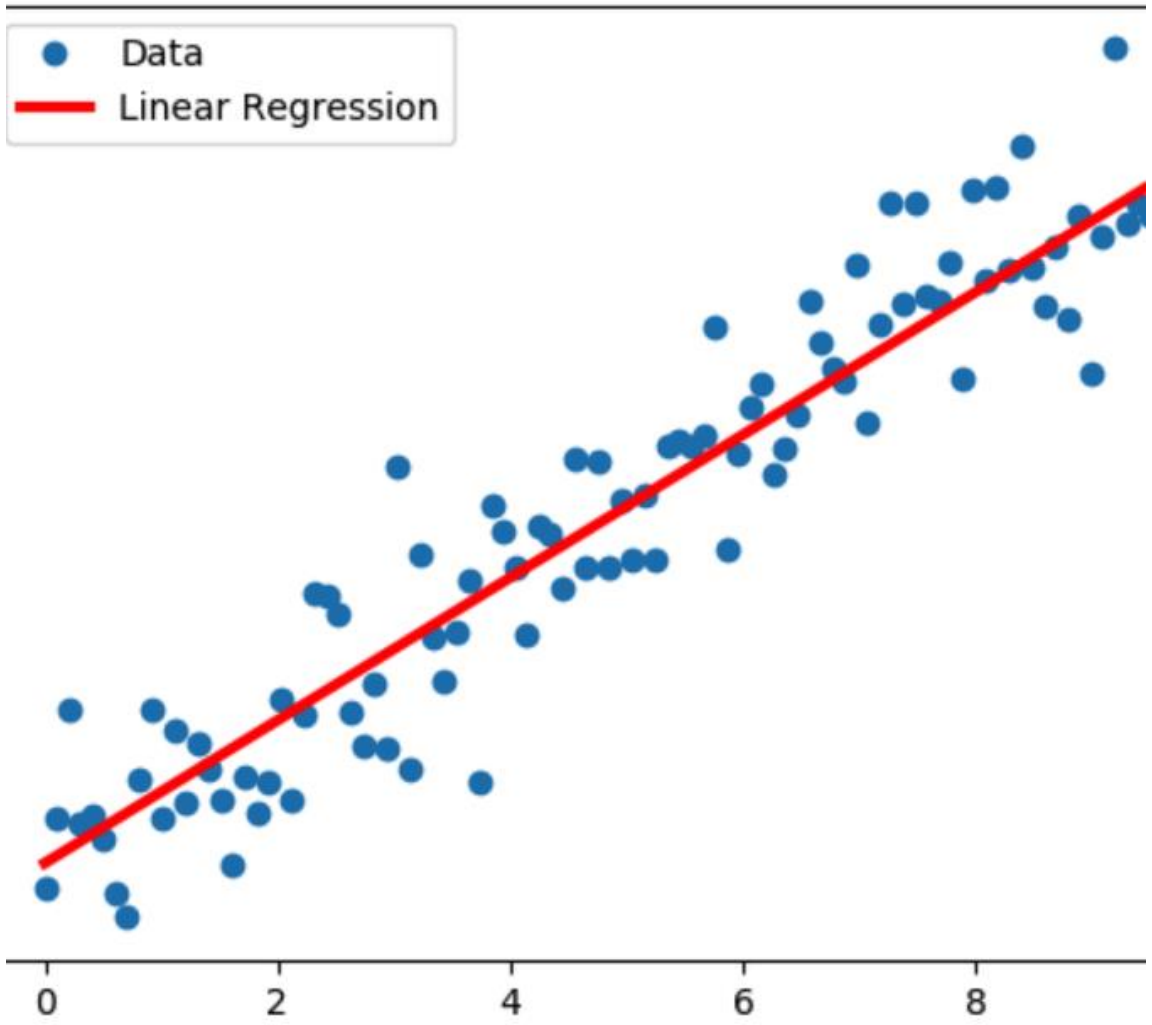
=> $\{A_1, A_3, A_4, A_5, A_6\}$
=> $\{A_1, A_4, A_5, A_6\}$
=> Reduced attribute set:
$\{A_1, A_4, A_6\}$



Heuristic: Decision tree induction

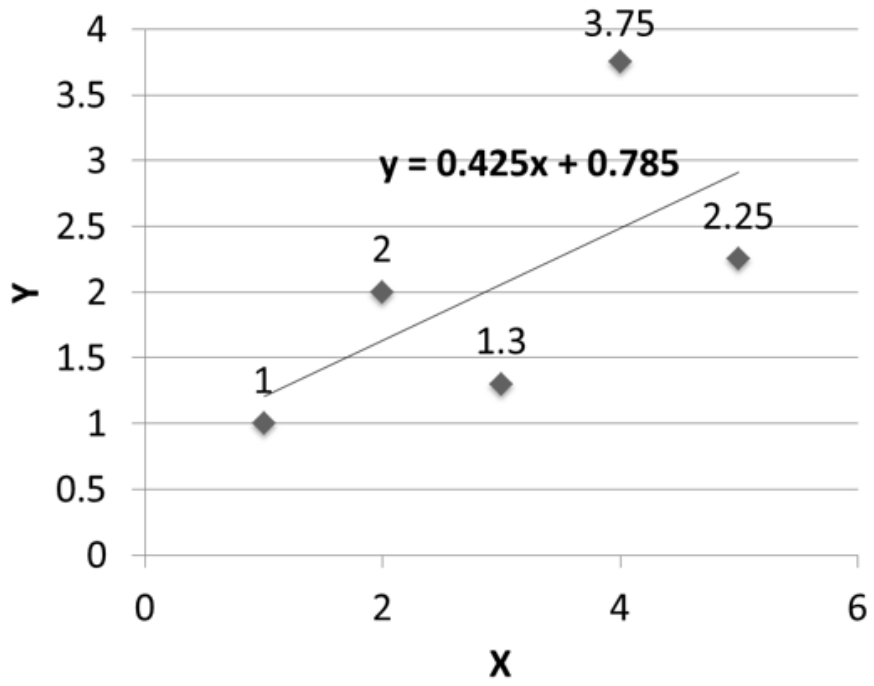
Regression

- $y = wx + b$
- y (response variable), can be modeled as a linear function of x (predictor variable)
- W (slope) and b (intercept) could be optimized to get the best fitting

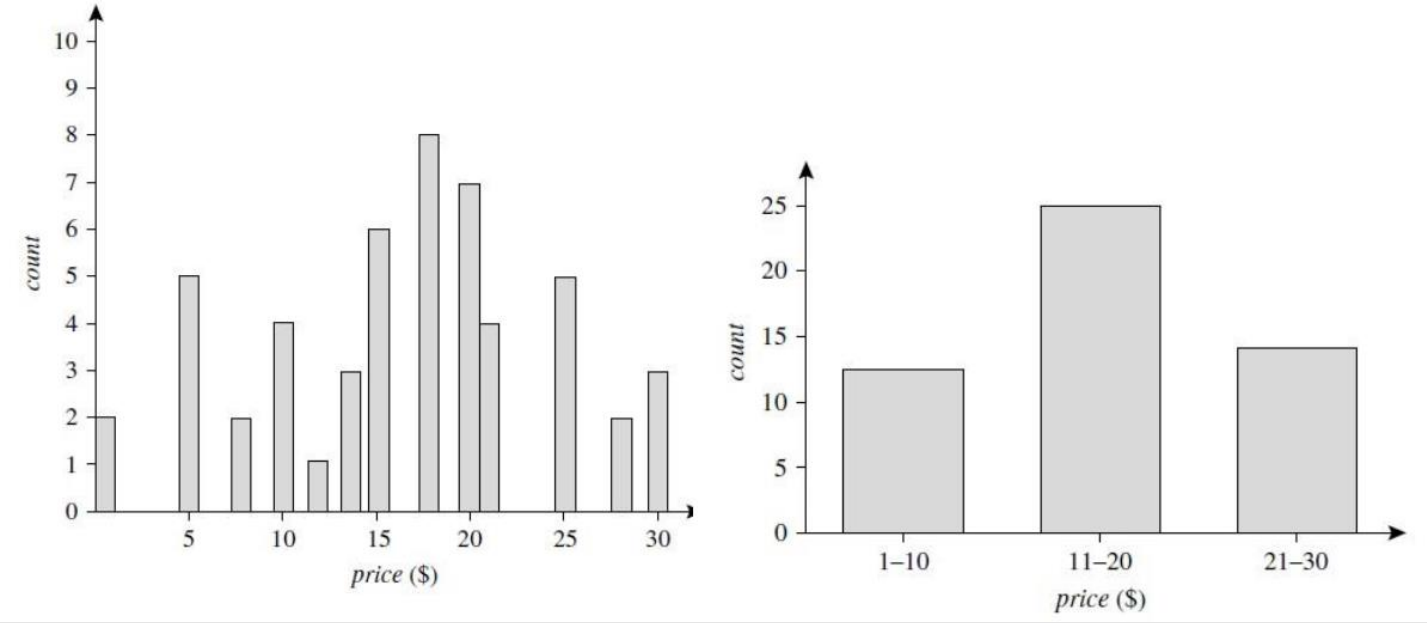


Regression

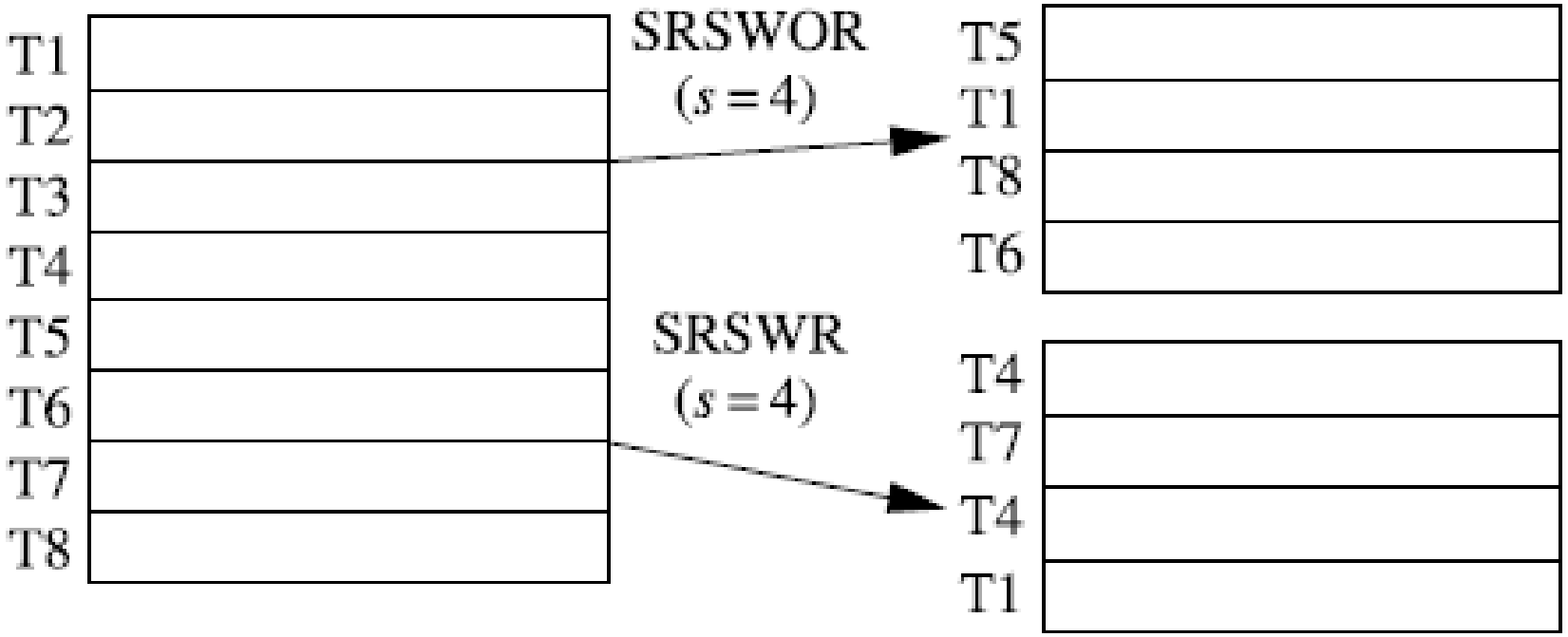
X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25



Histograms (binning)

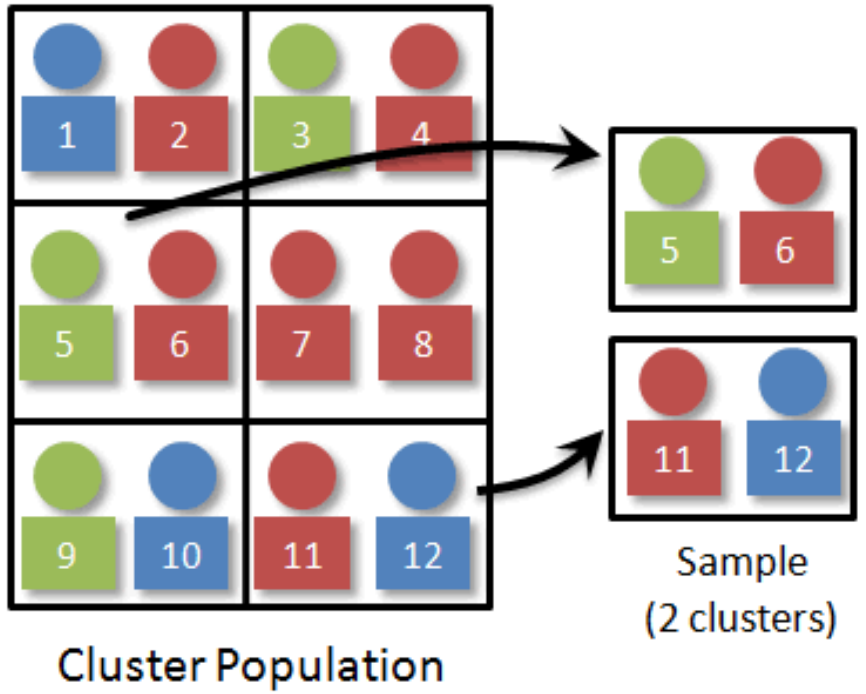


- The following data are a list of AllElectronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted:
- 1, 1, 5, 5, 5,5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,30, 30.



Sampling

- Obtain (smaller) subsets of the dataset called data sample.
- **Simple random sample without replacement (SRSWOR) of size s**: all tuples are equally likely to be sampled.
- **Simple random sample with replacement (SRSWR) of size s:** similar to SRSWOR, but a tuple is drawn recorded then placed back so it may be drawn again



Sampling

- **Cluster sample** : non overlapping
- **Stratified sample** : if the tuples are divided into strata (overlapping)

Data Transformation

Data are transformed into forms appropriate for mining.

Transformation Strategies

- Smoothing
- Attribute Selection
- Aggregation For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.
- Normalization: scaling values
- Discretization: (e.g., *age*) are replaced by interval labels (e.g., 0–10, 11–20, etc.)
- Concept Hierarchy: *street* can be generalized to higher-level concepts, like *city* or *country*

Transformation by Normalization

- To help avoid dependence on the choice of measurement units
- Normalizing the data attempts to give all attributes an equal weight
  - Min-max normalization
  - Z-score normalization

Min-Max Normalization

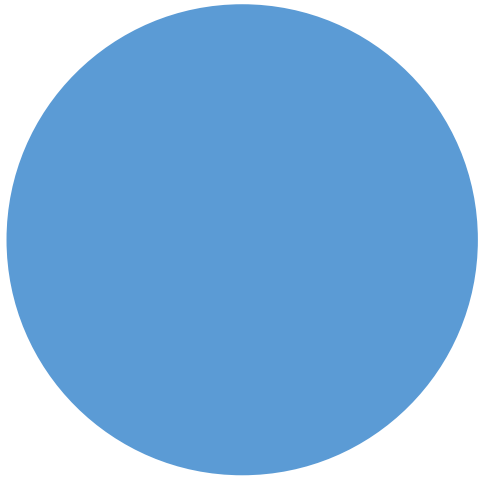
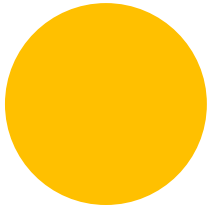
- $v = \frac{v-min}{max-min}(new_{max} - new_{min}) + new_{min}$
- Suppose that the minimum and maximum values for the attribute age are 13 and 70, respectively. We would like to map age to the range [0.0, 1.0].
- By min-max normalization, a value of 35 for age is transformed to $map(35) = \frac{35-13}{70-13}(1-0)+0 = 0.39$

Z-score normalization (zero-mean)

- Normalized based on the mean and standard deviation .
- $v = \frac{v - mean}{standard\ deviation}$
- Useful when the actual minimum and maximum of attribute A are unknown, or
- when there are outliers that dominate the min-max normalization

Concept Hierarchy Generation

- It Recursively reduce data by replacing low level concepts (e.g. age values) by higher level concepts (e.g. age groups: youth, adult, or senior).
- explicitly specified by domain experts
- formed for both numeric and nominal data

Thanks