

# Data Preprocessing

# Real World Data

---

Any Problem?

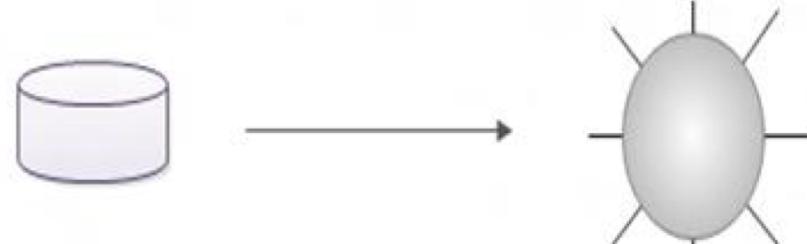
S.No	Credit_ratin g	Age	Income	Credit_card s
1	0.00	21	10000	y
2	1.0		2500	n
3	2.0	62	-500	y
4	100.012	42		n
5	yes	200	1	y
6	30	0	Seventy thousand	No

# Data Preprocessing

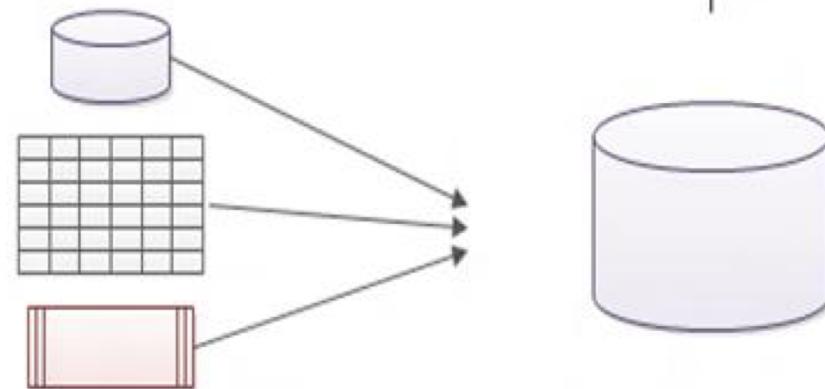
---

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation

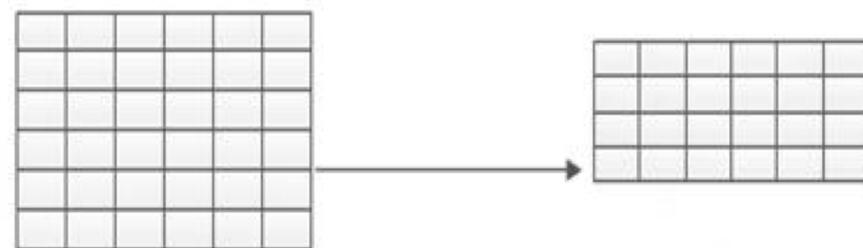
DATA CLEANING



DATA INTEGRATION



DATA REDUCTION



DATA TRANSFORMATION

-2,32,100  $\longrightarrow$  -0.02,0.32,1.00

# Data Cleaning

---

1. Missing Data
  - Central Imputation
  - KNN Imputation
2. Noisy Data
  - Smoothing
  - Clustering
1. Outlier Removal
  - Using Boxplot

company name	furigana	postal code	address		telephone number	
AlphaPurchase Co., Ltd	Alpha Purchase	107-0061	Aoyama Building 12th floor, 1-2-3, Kita-Aoyama, Minato-ku, Tokyo		03-5772-7801	
AAA Foundation	AAA	1500002	Kami-meguro, Meguro-ku X-X-X		0312345678	
BBBB, Inc.	BBBB	123	Minami-Azabu, Minato-ku XX-1-1		03(1234)9876	
company name	juridical personality	furigana	postal code	all prefectures	address	telephone number
Alpha Purchase	Co., Ltd	Alpha Purchase	1070 061	Tokyo	Aoyama Building 12th floor, 1-2-3, Kita-Aoyama, Minato-ku	0357727801
AAA	Foundation	AAA	1500 002	Tokyo	Kami-meguro, Meguro-ku X-X-X	0312345678
BBBB	Inc.	BBBB	1230 001	Tokyo	Minami-Azabu, Minato-ku XX-1-1	0312349876

# Imputation

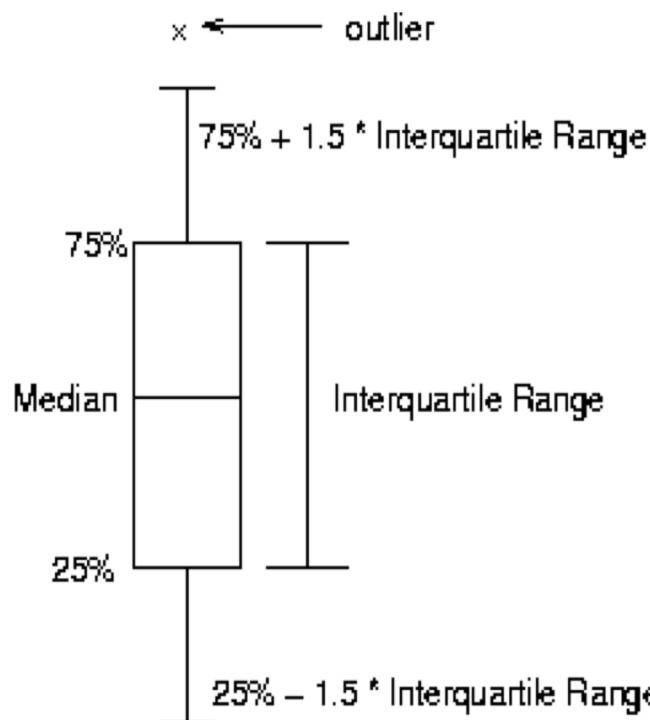
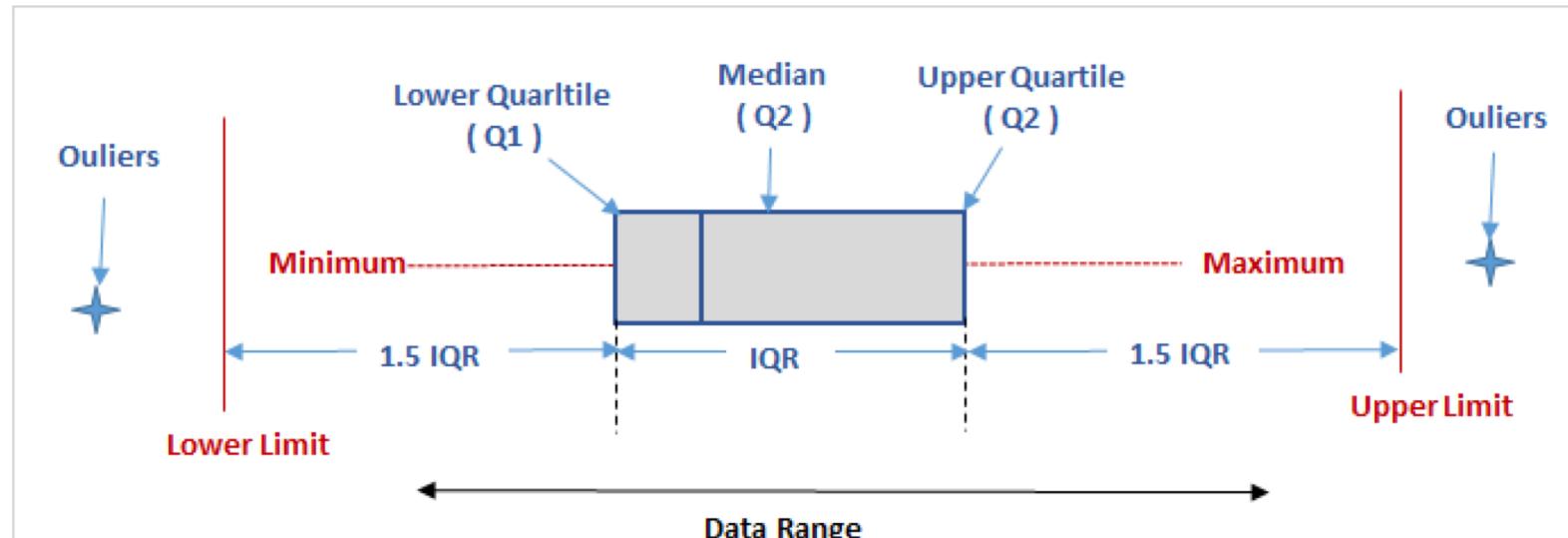
---

- Replace with mean or a median
- When to use mean?
- Replace with nearest neighbour
- How much nearest to see?

S.No	Qualification	Age	Income
1	B.Tech	25	30k
2	M.Tech	30	50k
3	B.Tech	26	32k
4	B.Tech	25	?
5	M.Tech	29	60k
6	B.Tech	?	30k

# Outlier

- BoxPlot



# Data Transformation

---

- Normalization

## Min-max normalization

1. Min Max Normalization
2. Z - Score Normalization
3. Decimal scaling

## Decimal scaling

$$v = v/10^j$$

## Normalization: Example II

- Min-Max normalization on an employee database

- max distance for salary: 100000-19000 = 81000
- max distance for age: 52-27 = 25
- New min for age and salary = 0; new max for age and salary = 1

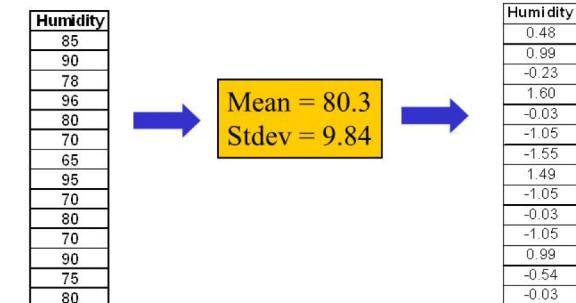
$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i} (newmax - newmin) + newmin$$

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

ID	Gender	Age	Salary
1	1	0.00	0.00
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32

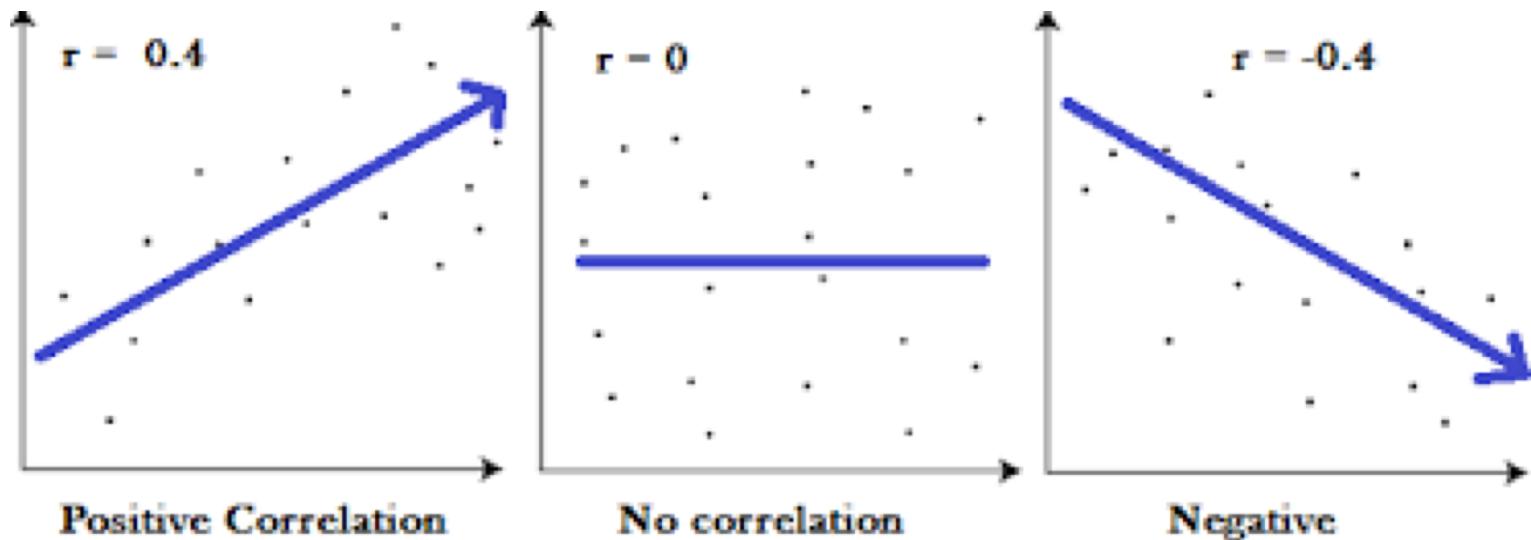
## Normalization: Example

- z-score normalization:  $v' = (v - \text{Mean}) / \text{Stdev}$
- Example: normalizing the “Humidity” attribute:



# Data Integration

- Check for correlation
- Remove uncorrelated data

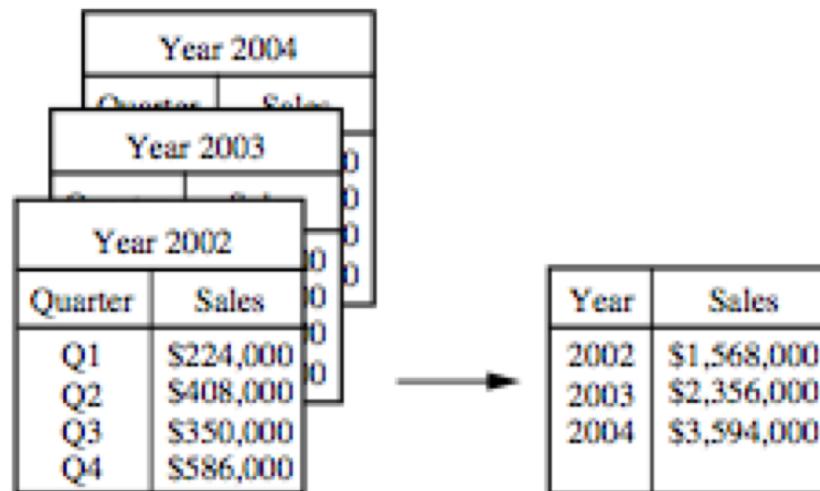


$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

# Data Reduction

---

- Data Cube Aggregation



**Figure 2.13** Sales data for a given branch of *AllElectronics* for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales.