CHƯƠNG 3
LÀM SẠCH VÀ TIỀN XỬ LÝ DỮ LIỆU

Giảng viên: Nguyễn Anh Thư

Khoa: Khoa học ứng dụng

NỘI DUNG BÀI HỌC

3.1 THĂM DÒ DỮ LIỆU

3.1.1. Phân tích đơn biến – Univariate Analysis

3.1.2. Phân tích đa biến – Multivariate Analysis

3.1.3. Distribution Fitting

3.2 LÀM SẠCH DỮ LIỆU

3.2.1. Biến đổi kiểu dữ liệu

3.2.2. Xử lý dữ liệu thiếu và các ngoại lệ

3.2.3. Phát hiện và loại bỏ trùng lặp

3.3 TIỀN XỬ LÝ DỮ LIỆU

MỤC TIÊU BÀI HỌC

Sau khi học xong bài này, cần nắm được các vấn đề sau:

- Các phương pháp phân tích đơn biến, đa biến và kiểm tra phân phối phù hợp với dữ liệu.
- Các kỹ thuật cơ bản trong việc làm sạch dữ liệu.
- Sau khi khám phá được nội hàm bên trong dữ liệu, ta thực hiện tiền xử lý dữ liệu để chuẩn bị đưa vào khai thác sâu hơn (như xây dựng mô hình,…).

3.1 THĂM DÒ DỮ LIỆU

- Khi làm việc với một tập dữ liệu, bước đầu tiên là hiểu ý nghĩa và đặc điểm chính của nó, tức là các sự kiện hoặc đối tượng mà dữ liệu đề cập đến và các thuộc tính mà nó biểu thị, cần phải hiểu được nội hàm. Mục tiêu của việc thăm dò trên một tập dữ liệu bao gồm:
  - Xác định kiểu loại của tập dữ liệu (có cấu trúc, phi cấu trúc, bán cấu trúc).
  - Nếu dữ liệu không phải phi cấu trúc, cần hiểu được schema của nó (các thuộc tính cấu thành và cách tổ chức dữ liệu).
  - Xác định kiểu miền của từng thuộc tính (danh mục, thứ tự, hay số) và đặc điểm của giá trị trong miền đó (giá trị điển hình, phạm vi giá trị, v.v.).

3.1 THĂM DÒ DỮ LIỆU

- Mục tiêu trong phần này là làm việc với dữ liệu có cấu trúc và lược đồ đã biết.
- Các bước:
  1. Phân tích từng thuộc tính riêng lẻ được gọi là phân tích đơn biến (univariate analysis) trong thống kê, tập trung vào một biến tại một thời điểm.
  2. Phân tích các mối quan hệ giữa các thuộc tính được gọi là phân tích đa biến trong thống kê. Thông thường một thuộc tính trong dữ liệu thường có liên quan đến các thuộc tính khác vì mỗi thuộc tính mô tả một khía cạnh/ đặc điểm của đối tượng, sự kiện nên nhiều khi chúng thể hiện một số đặc điểm chung.

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

- Cần xác định loại dữ liệu của thuộc tính (domain type) như categorical (danh mục), numerical (số), hoặc ordinal (thứ tự). Ngoài ra, cần xem xét liệu kiểu dữ liệu được gán cho thuộc tính (ví dụ, string, number, date) có phù hợp với loại dữ liệu của nó không.
- Ví dụ:
  - Các categorical có thể được mã hóa bằng số ('1', '2', ...), nhưng không có ý nghĩa số học hoặc thứ tự.
  - Thông tin thời gian (temporal data) thường bị nhập dưới dạng chuỗi thay vì kiểu dữ liệu ngày/thời gian, điều này gây khó khăn cho việc phân tích.
  - Thuộc tính dạng số không nên được lưu dưới dạng chuỗi (string), hoặc ngược lại, ngày tháng không nên lưu dưới dạng số.

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

- Thường đánh giá về:
  - Khoảng giá trị (min, max).
  - Xu hướng trung tâm (mean, median, mode).
  - Độ phân tán (độ lệch chuẩn, phân vị,…).

➔ Với các thước đo này ta sử dụng các hàm AGGREGATE FUNCTION.

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình - MEAN**

```
SELECT Avg(Attr) as mean
FROM Data;
```

≡

```
SELECT Sum(Attr) /
(Count(Attr) * 1.0) as mean
FROM Data;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình - MEAN**

VD1: Từ bảng StudentScores(student_id, student_name, score, exam_date, subjects_enrolled) như dưới đây, hãy tính điểm trung bình của các sinh viên.

student_id	student_name	score	exam_date	subjects_enrolled
1	Nguyen Van A	8.5	2024-01-15	3
2	Tran Thi B	9	2024-01-15	1
3	Le Van C	7	2024-01-15	2
4	Pham Thi D	8	2024-01-16	5
5	Hoang Van E	7.5	2024-01-16	4

student_id	student_name	score	exam_date	subjects_enrolled
6	Dang Thi F	6.5	2024-01-16	2
7	Vo Van G	9.5	2024-01-17	1
8	Bui Thi H	8	2024-01-17	3
9	Do Van I	7.5	2024-01-17	3
10	Hoang Bach K	7.2	2024-01-16	5

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình - MEAN**

VD1: Từ bảng StudentScores(student_id, student_name, score, exam_date, subjects_enrolled) như dưới đây, hãy tính điểm trung bình của các sinh viên.

Cách 1:

```
SELECT Avg(score) as mean
FROM StudentScores;
```

Cách 2:

```
SELECT Sum(score) / (Count(score) * 1.0) as mean
FROM StudentScores;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình - MEAN**

- Trong xác suất thống kê, nếu một giá trị x_i xuất hiện f_i lần trong tổng số N quan sát.
- Khi đó, xác suất thực nghiệm là:

$$P(x_i) \approx \frac{f_i}{N} = \frac{f_i}{\sum f_i} \qquad (1)$$

$$\sum x_i \cdot P(x_i) = \sum \left(x_i \cdot \frac{f_i}{\sum f_i} \right) \qquad (2)$$

➔ (2) là công thức TB KỲ VỌNG $\mathrm{E(X)} = \sum x_i \cdot P(x_i)$

3.1 THĂM DÒ DỮ LIỆU

❖ **Giá Trị Trung Bình –** Sử dụng xác suất

```
WITH NHistogram(Value, Prob) AS
(       SELECT Attr, sum(1.0/total)
        FROM  Data,  (SELECT  COUNT(*)  AS
                total FROM Data) AS Temp
        GROUP BY Attr
)
SELECT Sum(Value * Prob) AS mean
FROM NHistogram;
```

❖ **Giá Trị Trung Bình –** Sử dụng tần số

```
WITH Histogram(Value, Frequency) AS
(       SELECT Attr, count(*)
        FROM Data
        GROUP BY Attr
)
SELECT (1.0 * Sum(Value * Frequency)) /
Sum(Frequency) AS mean
FROM Histogram;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình - MEAN**

VD2: Từ bảng StudentScores(student_id, student_name, score, exam_date, subjects_enrolled) như dưới đây, tính số môn học trung bình mà các sinh viên tham gia (**sử dụng 2 cách xác suất và tần số**, biết số môn học tối thiểu mà sinh viên tham gia nằm trong khoảng từ 1 đến 4).

student_id	student_name	score	exam_date	subjects_enrolled
1	Nguyen Van A	8.5	2024-01-15	3
2	Tran Thi B	9	2024-01-15	1
3	Le Van C	7	2024-01-15	2
4	Pham Thi D	8	2024-01-16	5
5	Hoang Van E	7.5	2024-01-16	4

student_id	student_name	score	exam_date	subjects_enrolled
6	Dang Thi F	6.5	2024-01-16	2
7	Vo Van G	9.5	2024-01-17	1
8	Bui Thi H	8	2024-01-17	3
9	Do Van I	7.5	2024-01-17	3
10	Hoang Bach K	7.2	2024-01-16	5

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình - MEAN**

VD2: *Cách 1:*

```
WITH NHistogram(Value, Prob)
AS ( SELECT Subjects_Enrolled, sum(1.0/total)
     FROM StudentScores, (SELECT COUNT(*) AS total FROM StudentScores) AS Temp
     GROUP BY Subjects_Enrolled )
SELECT Sum(Value * Prob) AS mean FROM NHistogram;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình - MEAN**

VD2: *Cách 2:*

```
WITH Histogram(Value, Frequency)
AS ( SELECT Subjects_Enrolled, count(*)
     FROM StudentScores GROUP BY Subjects_Enrolled )
SELECT (1.0 * Sum(Value * Frequency)) / Sum(Frequency) AS mean
FROM StudentScores;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – TRIMMED MEAN**

- GTTB rất nhạy bởi các giá trị ngoại lai (outlier) và các giá trị cực trị.
- Ví dụ nếu có một giá trị rất cao hoặc rất thấp thì nó sẽ ảnh hưởng rất lớn đến giá trị trung bình.

➔ Sử dụng TRIMMED MEAN – giá trị trung bình được cắt tỉa: GTTB sẽ được tính toán sau khi bỏ qua các giá trị cực trị, thường là GTNN và GTLN, hoặc khái quát hơn là loại bỏ k% giá trị cao nhất/ thấp nhất.

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – TRIMMED MEAN**

Loại bỏ các giá trị cực trị là GTLN và GTNN

```
SELECT avg(A)
FROM Data,
    (SELECT max(A) AS Amax FROM Data) AS T1,
    (SELECT min(A) AS Amin FROM Data) AS T2
WHERE A < Amax and A > Amin;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – TRIMMED MEAN**

Loại bỏ các giá trị cực trị là GTLN và GTNN

```
SELECT AVG(score)
FROM   studentscores,
       (SELECT min(score) As min_score FROM
              studentscores) AS T1,
       (SELECT max(score) AS max_score FROM
              studentscores) as T2
WHERE min_score < score AND score < max_score;
```

avg(score)
7.837499976158142

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – TRIMMED MEAN**

Loại bỏ k% GTLN và GTNN:

Bước 1. Sắp xếp dữ liệu theo chiều tăng/giảm dần giá trị:

```
WITH RankedData
AS (SELECT A,
    ROW_NUMBER() OVER (ORDER BY A ASC) AS rank_asc,
    ROW_NUMBER() OVER (ORDER BY A DESC) AS rank_desc
    FROM Data)
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – TRIMMED MEAN**

Loại bỏ k% GTLN và GTNN:

Bước 2. Loại bỏ k% phần tử là GTLN và GTNN:

```
WITH TrimmedData
AS (SELECT A
    FROM RankedData
    WHERE rank_asc > (SELECT COUNT(*) * 0.01 * k FROM Data)
    AND rank_desc > (SELECT COUNT(*) * 0.01 * k FROM Data)
)
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – TRIMMED MEAN**

Loại bỏ k% GTLN và GTNN:

Bước 3. Tính TRIMMED MEAN:

```
SELECT AVG(A) AS TrimmedMean
FROM TrimmedData;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ Giá Trị Trung Bình – GEOMETRIC MEAN

- Geometric mean (trung bình nhân) có công thức tổng quát như sau: $\sqrt[n]{x_1 * x_2 * \cdots * x_n}$.
- Ưu điểm: ít nhạy cảm với các giá trị ngoại lại so với trung bình cộng, đặc biệt với các giá trị lớn.
- Nhược điểm: vẫn bị ảnh hưởng bởi các giá trị nhỏ.
- SQL không có hàm AGGREGATE để tính trực tiếp.

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – GEOMETRIC MEAN** $\sqrt[n]{x_1 * x_2 * \cdots * x_n}$.

- **Cách 1**: giữ nguyên công thức

Bước 1: Sử dụng hàm logarit để tính tích:

$$\boldsymbol{log(a * b) = log\, a + log\, b}$$

$$\Rightarrow \boldsymbol{a * b = \exp(\log a + \log b)}$$

```
SELECT exp(sum(log(Attr))
FROM Data;
```

Bước 2: Sử dụng hàm POW trong SQL với số mũ là $\frac{1}{n}$ để tính căn.

```
SELECT pow(exp(sum(log(Attr)), 1.0/total)
FROM Data, (SELECT count(Attr) AS total FROM Data);
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **Giá Trị Trung Bình – GEOMETRIC MEAN** $\sqrt[n]{x_1 * x_2 * \cdots * x_n}$.

- **Cách 2**: biến đổi thành công thức tương đương $\sqrt[n]{\boldsymbol{x_1 * x_2 * \cdots * x_n}} = \boldsymbol{exp}\,\frac{\boldsymbol{ln\, x_1 + ln\, x_2 + \cdots + ln\, x_n}}{\boldsymbol{n}}$.

```
SELECT exp(sum(log(Attr)) / count(Attr))
FROM Data;
```

Vì sum/count = average nên câu lệnh trên có thể thay thế thành:

```
SELECT exp(avg(log(Attr))
FROM Data;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ Giá Trị Trung Bình – GEOMETRIC MEAN

- Khi nào thì nên sử dụng trung bình nhân?
  - Hữu ích khi tính toán tỷ lệ lãi suất trong ngân hàng.
  - Đo lường tốc độ tăng trưởng/ suy giảm trung bình trong tài chính, kinh tế và sinh học.
  - Phù hợp với các tường hợp có tỷ lệ thay đổi mạnh hoặc dữ liệu có biến động lớn.

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ **MODE**

- Là giá trị xuất hiện nhiều nhất trong dữ liệu (có thể có nhiều mode).
- Trong SQL không có hàm tính mode mà phải tính toán theo các bước như sau:

```
WITH Histogram AS
 (SELECT Value AS val, count(*) AS freq FROM Data GROUP BY Attr)
SELECT val
FROM Histogram, (SELECT max(freq) AS top FROM Histogram) AS T
WHERE freq = top;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ MODE

- **VD3**: Trong các chuyến bay đến NY, hãy tính điểm đến (destination) được ghé thăm nhiều nhất (điểm đến với nhiều chuyến bay đến đó).

FlightID	Origin	Destination	FlightDate
1	NY	Los Angeles	2024-02-01
2	NY	Chicago	2024-02-02
3	NY	Los Angeles	2024-02-03
4	NY	San Francisco	2024-02-04
5	NY	Chicago	2024-02-05

FlightID	Origin	Destination	FlightDate
6	NY	Los Angeles	2024-02-06
7	NY	Miami	2024-02-07
8	NY	San Francisco	2024-02-08
9	NY	Los Angeles	2024-02-09
10	NY	Chicago	2024-02-10

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ MODE

- **VD3**: Trong các chuyến bay đến NY, hãy tính điểm đến (destination) được ghé thăm nhiều nhất (điểm đến với nhiều chuyến bay đến đó).

  Cách 1:

```
SELECT Destination, COUNT(*) AS FlightCount
FROM Flights
GROUP BY Destination
ORDER BY FlightCount DESC
LIMIT 1;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ MODE

- **VD3**: Trong các chuyến bay đến NY, hãy tính điểm đến (destination) được ghé thăm nhiều nhất (điểm đến với nhiều chuyến bay đến đó).

Cách 2:

```
WITH Histogram AS  (SELECT Destination AS val, count(*) AS freq
                    FROM Flights GROUP BY Destination)
SELECT val, COUNT(*) AS FlightCount
FROM Histogram, (SELECT max(freq) AS top FROM Histogram) AS T
WHERE freq = top;
```

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ MEDIAN

- Là giá trị xuất hiện ở vị trí giữa khi các giá trị được sắp xếp theo thứ tự nhỏ đến lớn sao cho số lượng số bên trái và phải nó là bằng nhau, được gọi là trung vị.
  - ➢ Nếu số lượng phần tử là lẻ, trung vị là giá trị ở giữa.
  - ➢ Nếu số lượng phần tử là chẵn, trung vị là trung bình của hai giá trị ở giữa.
- Trung vị ít bị ảnh hưởng bởi các giá trị ngoại lai hơn so với trung bình cộng.

3.1 THĂM DÒ DỮ LIỆU

3.1.1 PHÂN TÍCH ĐƠN BIẾN

a. VỚI CÁC THUỘC TÍNH KIỂU SỐ

❖ MEDIAN

- Tính trung vị trong SQL khá phức tạp do cần sắp xếp các giá trị (ORDER BY) và xác định vị trí ở giữa, thường sử dụng kết hợp LIMIT và OFFSET để lấy giá trị mong muốn:
  - Xác định tổng số phần tử `size`.
  - Sắp xếp dữ liệu sử dụng `ORDER BY`.
  - `MOD(size, 2)` kiểm tra tổng số phần tử chẵn hay lẻ .
  - Xác định vị trí bắt đầu của trung vị bằng cách
    - Xác định vị trí chính giữa của DL `CEIL(size / 2.0)`
    - `OFFSET` dịch con trỏ đến vị trí chính giữa của DL.
    - Tính trung bình của các phần tử chính giữa.

```
SELECT avg(Attr)
FROM (SELECT Attr FROM Data,
      (SELECT count(*) as size FROM Data)
      ORDER BY value
      LIMIT 2 - MOD(size, 2)
      OFFSET CEIL(size / 2.0) ) AS T;
```