3.1 Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how data quality assessment can depend on the intended use of the data, giving examples. Propose two other dimensions of data quality.

• Độ chính xác, trước tiên hãy xem xét hệ thống khuyến nghị mua quần áo trực tuyến. Khi nói đến ngày sinh, hệ thống có thể chỉ quan tâm đến người dùng được sinh vào năm nào, để

có thể cung cấp các lựa chọn phù hợp. Tuy nhiên, một ứng dụng trong facebook tạo lịch sinh nhật cho bạn bè phải có được ngày chính xác mà người dùng được sinh ra để tạo lịch đáng tin cậy.

• Độ hoàn thiện, người quản lý sản phẩm có thể không quan tâm lắm nếu thông tin địa chỉ của khách hàng bị thiếu trong khi nhà phân tích tiếp thị coi thông tin địa chỉ cần thiết để phân tích.

• Tính nhất quán, hãy xem xét một người quản lý cơ sở dữ liệu đang hợp nhất hai cơ sở dữ liệu thông tin phim lớn thành một. Khi anh ấy quyết định xem hai mục có đề cập đến cùng một bộ phim hay không, anh ấy có thể kiểm tra tiêu đề và ngày phát hành của mục đó. Ở đây trong một trong hai cơ sở dữ liệu, ngày phát hành phải phù hợp với tiêu đề nếu không sẽ có những vấn đề khó chịu. Nhưng khi người dùng tìm kiếm thông tin của một bộ phim chỉ để giải trí bằng cách sử dụng một trong hai cơ sở dữ liệu, liệu ngày phát hành có phù hợp với tiêu đề hay không không quá quan trọng. Người dùng thường quan tâm nhiều hơn đến nội dung của phim.

Hai thứ nguyên khác có thể được sử dụng để đánh giá chất lượng dữ liệu có thể được lấy từ những thứ sau:

tính kịp thời, đáng tin cậy, giá trị gia tăng, khả năng diễn giải và khả năng tiếp cận. Chúng có thể được sử dụng để đánh giá chất lượng liên quan đến các yếu tố sau:

• Tính kịp thời: Dữ liệu phải có sẵn trong một khung thời gian cho phép nó hữu ích cho việc ra quyết định.

• Độ tin cậy: Giá trị dữ liệu phải nằm trong phạm vi kết quả có thể có để hữu ích cho việc ra quyết định.

• Giá trị gia tăng: Dữ liệu phải cung cấp giá trị bổ sung về mặt thông tin bù đắp chi phí thu thập và truy cập nó.

• Khả năng diễn giải: Dữ liệu không được phức tạp đến mức nỗ lực hiểu thông tin mà nó cung cấp vượt quá lợi ích của việc phân tích.

• Khả năng truy cập: Dữ liệu phải có thể truy cập được để nỗ lực thu thập không vượt quá lợi ích từ việc sử dụng.

**3.3** Exercise 2.2 gave the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

Step 1: Sort the data. (This step is not required here as the data are already sorted.)

• Step 2: Partition the data into equi depth bins of depth 3.

Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22

Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35

Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70

• Step 3: Calculate the arithmetic mean of each bin.

• Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin.

Bin 1: 142/3, 142/3, 142/3 Bin 2: 181/3, 181/3, 181/3 Bin 3: 21, 21, 21

Bin 4: 24, 24, 24 Bin 5: 262/3, 262/3, 262/3 Bin 6: 332/3, 332/3, 332/3

Bin 7: 35, 35, 35 Bin 8: 401/3, 401/3, 401/3 Bin 9: 56, 56, 56

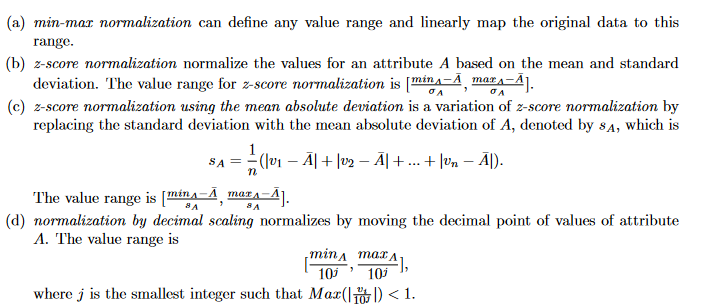
b) How might you determine outliers in the data?

Các giá trị ngoại lai trong dữ liệu có thể được phát hiện bằng cách phân nhóm, trong đó các giá trị tương tự được tổ chức thành các nhóm hoặc 'cụm'. Các giá trị nằm ngoài tập hợp các cụm có thể được coi là giá trị ngoại lệ. Ngoài ra, có thể sử dụng sự kết hợp giữa kiểm tra bằng máy tính và con người khi thực hiện phân phối dữ liệu xác định trước để cho phép máy tính xác định các ngoại lệ có thể xảy ra. Sau đó, những giá trị ngoại lệ có thể có này có thể được xác minh bằng sự kiểm tra của con người với nỗ lực ít hơn nhiều so với yêu cầu xác minh toàn bộ tập dữ liệu ban đầu.

c) What other methods are there for data smoothing?

Các phương pháp khác có thể được sử dụng để làm mịn dữ liệu bao gồm các hình thức thay thế của binning như làm mịn theo phương tiện bin hoặc làm mịn theo ranh giới bin. Ngoài ra, các thùng có chiều rộng tương đương có thể được sử dụng để thực hiện bất kỳ hình thức xếp thùng nào, trong đó phạm vi khoảng giá trị trong mỗi thùng là không đổi. Các phương pháp khác với binning bao gồm sử dụng các kỹ thuật hồi quy để làm mịn dữ liệu bằng cách điều chỉnh nó với một chức năng chẳng hạn như thông qua hồi quy tuyến tính hoặc nhiều lần. Ngoài ra, các kỹ thuật phân loại có thể được sử dụng để triển khai phân cấp khái niệm có thể làm mượt dữ liệu bằng cách cuộn các khái niệm cấp thấp hơn lên các khái niệm cấp cao hơn.

**3.5** What are the value ranges of the following normalization methods?



**3.7** Using the data for age given in Exercise 3.3, answer the following:

Using the data for age : 13, 15,16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

n = 27

sum = 809

mean = 30

(a) Use min-max normalization to transform the value 35 for age onto the range

[0.0,1.0].

0.39

(b) Use z-score normalization to transform the value 35 for age, where the standard

deviation of age is 12.94 years.

0.39

(c) Use normalization by decimal scaling to transform the value 35 for age.

0.35

(d) Comment on which method you would prefer to use for the given data, giving

reasons as to why

Với dữ liệu, người ta có thể thích chia tỷ lệ thập phân để chuẩn hóa vì một phép chuyển đổi như vậy sẽ duy trì phân phối dữ liệu và trực quan để diễn giải, trong khi vẫn cho phép khai thác trên các nhóm tuổi cụ thể. Chuẩn hóa tối đa có tác dụng không mong muốn là không cho phép bất kỳ giá trị nào trong tương lai nằm ngoài giá trị tối thiểu và tối đa hiện tại mà không gặp phải “lỗi ngoài giới hạn”. Vì có khả năng các giá trị đó có thể xuất hiện trong dữ liệu trong tương lai, nên phương pháp này ít thích hợp hơn. Ngoài ra, chuẩn hóa điểm số z biến đổi các giá trị thành các thước đo thể hiện khoảng cách của chúng so với giá trị trung bình, về độ lệch chuẩn. Có thể xảy ra rằng kiểu chuyển đổi này sẽ không làm tăng giá trị thông tin của thuộc tính về mặt trực quan đối với người dùng hoặc tính hữu ích của kết quả khai thác.

**3.9** Suppose a group of 12 sales price records has been sorted as follows:

5,10,11,13,15,35,50,55,72,92,204,215.

Partition them into three bins by each of the following methods:

(a) equal-frequency (equal-depth) partitioning

(5,10,11,13)(15,35,50,55)(72,92,204,215)

(b) equal-width partitioning

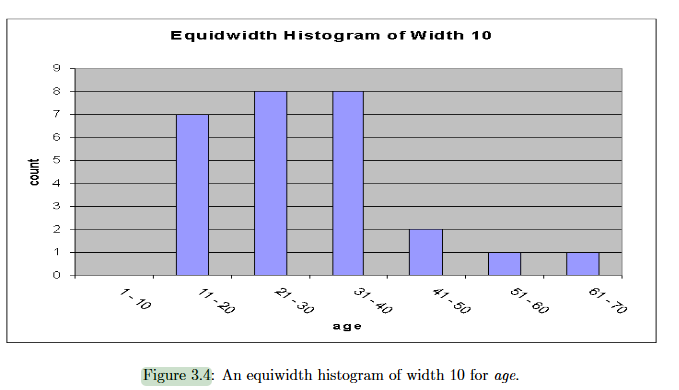
(5,10,11,13,15,35,50,55,72)(92)(204,215)

(c) clustering

(5,10,11,13,15,35)(50,55,72,92)(204,215)

3.11 Using the data for age given in Exercise 3.3,

(a) Plot an equal-width histogram of width 10.



(b) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR,

cluster sampling, and stratified sampling. Use samples of size 5 and the strata

“youth,” “middle-aged,” and “senior.”

3.13 Propose an algorithm, in pseudocode or in your favorite programming language, for the

following:

(a) The automatic generation of a concept hierarchy for nominal data based on the

number of distinct values of attributes in the given schema.

(b) The automatic generation of a concept hierarchy for numeric data based on the

equal-width partitioning rule.

3.8 Bibliographic Notes 123

(c) The automatic generation of a concept hierarchy for numeric data based on the

equal-frequency partitioning rule.