1. 2.3 A given set of values are grouped into intervals and frequencies:

age	frequency
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700
81-110	44

$$Median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

$$Median = 21 + \left(\frac{3194/2 - 950}{1500} \right) \times 30 = 33.94$$

2. 2.6 Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

   (a) Euclidean distance: $d_{Euclid} = \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{4+1+36+4} = \sqrt{45}$;

   (b) Manhattan distance: $d_{Manh} = 2 + 1 + 6 + 2 = 11$;

   (c) Minkowski distance (q=3): $d_{Mink} =^3 \sqrt{2^3 + 1 + 6^3 + 2^3} =^3 \sqrt{233}$;

   (d) supremum distance: $d_{sup} = max(2, 1, 6, 2) = 6$;

3. 2.7 Propose several methods for median approximation. Analyze their respective complexity under different parameter settings and decide to what extent the real value can be approximated.

   (a) Methods: One can use hierarchy methods to generate a 'tree' with several branches and find the median at the bottom. Mathematically, there are several median approximation according to wikipedia, such as Hodges–Lehmann estimator and highly efficient estimator of the population median, etc.

   (b) Analysis: For the complexity estimate, even though comparison-sorting n items requires Ω(n log n) operations, selection algorithms can compute the kth-smallest of n items with only Θ(n) operations. This includes the median, which is the (n/2)th order statistic (or for an even number of samples, the average of the two middle order statistics).

4. 2.8 Select similarity measures in data analysis. Suppose we have the following 2-D data set:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

(a) Given a new data point, x=(1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

	A_1	A_2	d_{Euclid}	d_{Manh}	d_{supr}	cosine similarity
x_1	1.5	1.7	$0.1\sqrt{2}$	0.2	0.1	0.99999139
x_2	2	1.9	$0.3\sqrt{5}$	0.9	0.6	0.995752261
x_3	1.6	1.8	$0.2\sqrt{2}$	0.4	0.2	0.999969483
x_4	1.2	1.5	$0.1\sqrt{5}$	0.3	0.2	0.9990282349
x_5	1.5	1.0	$0.1\sqrt{37}$	0.7	0.6	0.96536339303

Sample calculations:

d_{Euclid}(x,x1)	$\sqrt{(1.5-1.4)^2+(1.7-1.6)^2} = 0.1\sqrt{2}$
d_{Manh}(x,x1)	$\lvert 1.5-1.4\rvert + \lvert 1.7-1.6\rvert = 0.2$
d_{supr}(x,x1)	$max(\lvert 1.5-1.4\rvert , \lvert 1.7-1.6\rvert) = 0.1$
cosine similarity	$\frac{(1.4\times1.5+1.6\times1.7)}{\sqrt{1.4^2+1.6^2}\cdot\sqrt{1.5^2+1.7^2}}$ =0.99999139

(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

For this question, generate matlab code as below:

- X=[1.4,1.6;1.5,1.7;2,1.9;1.6,1.8;1.2,1.5;1.5,1.0]
- N=normr(X)
- for k = 1:6
  - M(k,1)=norm(N(k,:)-N(1,:));
- end

	(A_1, A_2)	$norm - ed(a_1, a_2)$	d_{Euclid}
x	(1.4, 1.6)	(0.6585, 0.7526)	0
x_1	(1.5, 1.7)	(0.6616, 0.7498)	0.0041
x_2	(2, 1.9)	(0.7250, 0.6887)	0.0922
x_3	(1.6, 1.8)	(0.6644, 0.7474)	0.0078
x_4	(1.2, 1.5)	(0.6247, 0.7809)	0.0441
x_5	(1.5, 1.0)	(0.8321, 0.5547)	0.2632

5. 3.1 For accuracy, completeness, and consistency, discuss how data quality assessment can depend on the intended use of the data, giving examples. Propose two other dimensions of data quality.

(a) Accuracy: if a dataset has lots of errors like values that deviate from the expected, or NaN missing data, or lacking precision (such as mistaken order of magnitude about 2 or larger, which will affect data analysis a lot for high-accuracy required experiments in natural science). These would influence result a lot. A story that have been said from time to time is a airplane crashed caused by a decimal point error for one data.

(b) Completeness, the most popular incompleteness case is dataset with NaNs. If lacking certain amount of 'good data', one dataset may be regarded as incomplete.

(c) Consistency, if data failed to obey a certain recognizable pattern which should be expected as right, data errors in the dataset may be seen as inconsistency.

i. From the textbook, explanation for this question is cited below:

There are many possible reasons for inaccurate data (i.e., having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value "January 1" displayed for birthday). This is known as disguised missing data. Errors in data transmission can also occur. There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date). Duplicate tuples also require data cleaning.Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Other data may not be included simply because they were not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the data history or modifications may have been overlooked.Missing data, particularly for tuples with missing values for some attributes, may need to be inferred. Recall that data quality depends on the intended use of the data. Two different users may have very different assessments of the quality of a given database. For example, a marketing analyst may need to access the database mentioned before for a list of customer addresses. Some of the addresses are outdated or incorrect, yet overall, 80% of the addresses are accurate. The marketing analyst considers this to be a large customer database for target marketing purposes and is pleased with the database's accuracy, although, as sales manager, you found the data inaccurate.

(d) Two other dimensions of data quality are believability and interpretability.Believability reflects how much the data are trusted by users, while interpretability reflects how easy the data are understood.

6. 3.3 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) Smoothing(compare initial vector and smoothed vector can be seen from Fig.1):

Smoothing by bin means, bin depth of 3:

Bin#	equal-frequency	by-bin-means
1	13,15, 16	14.67, 14.67, 14.67
2	16,19,20	18.33, 18.33, 18.33
3	20,21,22	21, 21, 21
4	22,25,25	24, 24, 24
5	25,25,30	26.67, 26.67, 26.67
6	33,33,35	33.67, 33.67, 33.67
7	35,35,35	35, 35, 35
8	36,40,45	40.33, 40.33, 40.33
9	46,52,70	56, 56, 56

Illustration of steps:

Sorting data,

Partition data into bin depth of 3,

Replace numbers with the mean value of its bin.

(b) Outliers: Can be detected by clustering, values that fall outside of the set of clusters maybe considered outliers. There are several built-in functions in matlab, such as kmeans, hierarchical clustering, etc. One can detect certain outliers according to specific task.

(c) Other methods for data smoothing:

Other forms of binning(smooth by bin median/boundaries); Regression(linear regression, multiple linear regressioin); Data discretization(decision tree induction, concept hierarchies); Some methods of classification have built-in data smoothing mechanisms.

7. 3.5 Value ranges of the following normalization methods>

   (a) min-max normalization:[new_min(A), new_max(A)]. Other data points in the attribute are mapping into $\frac{A_i - min(A)}{max(A) - min(A)} \times (newmax(A) - newmin(A) + newmin(A)$

   (b) z-score normalization: $[\frac{min(A)-\bar{A}}{\sigma_A}, \frac{max(A)-\bar{A}}{\sigma_A}]$

   (c) z-score normalization using mean absolute deviation: $[\frac{min(A)-\bar{A}}{S_A}, \frac{max(A)-\bar{A}}{S_A}]$, $S_A = \frac{1}{n}(|A_1 - \bar{A}| + \cdots + |A_n - \bar{A}|)$

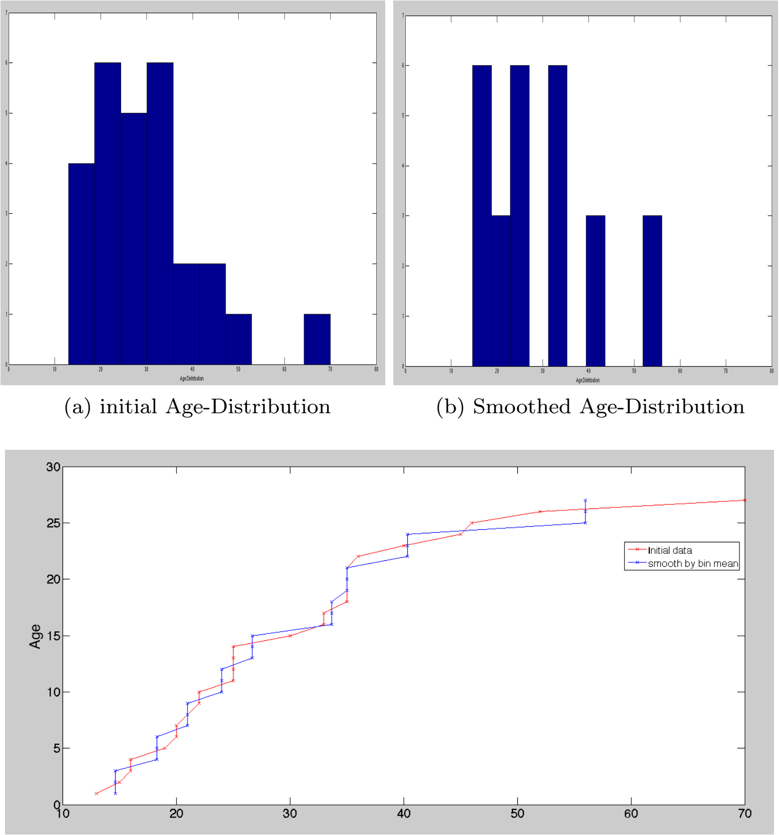
   (d) normailizaiton by decimal scaling: $[\frac{min(A)}{10^j}, \frac{max(A)}{10^j}]$, j is the smallest integer such that $max(\left|v_i^{'}\right| \leq 1$.

8. 3.7 Age distribution: X= [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) min-max normalization for 35: $\frac{35-13}{70-13} = \frac{22}{57} = 0.38596$

   (b) z-score for 35:

      i. As mean(X)= 29.963 and sqrt(var(X))=12.9421, 35 will be transformed as $\frac{35-29.963}{12.9421} = 0.3892$;



(a) initial Age-Distribution

(b) Smoothed Age-Distribution

(c) Smoothed Age-Distribution Scattering

Figure 1: Compare – initial and smoothed by bin means

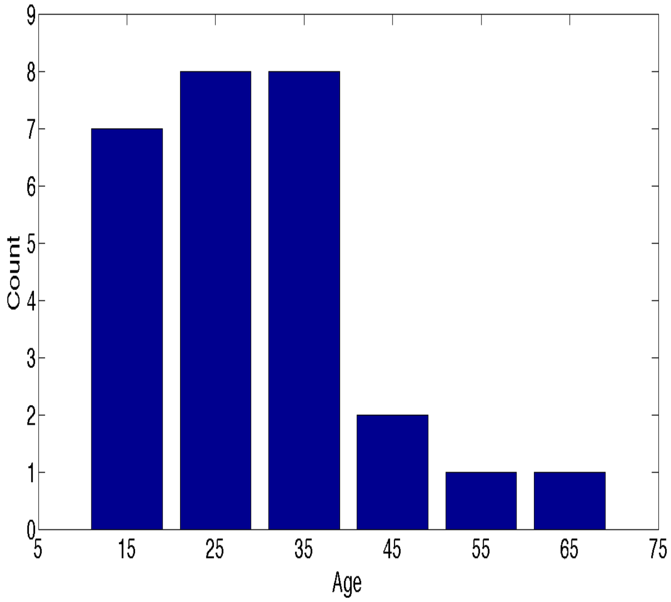


Figure 2: histogram of width=10

ii. Use mean absolute deviation: $s_A = 10.0357$ (Matlab:mad(X,0)), 35 will be transformed as $\frac{35-29.963}{10.0357} = 0.5019$

(c) decimal scaling normalization: 70→100, 35→0.35

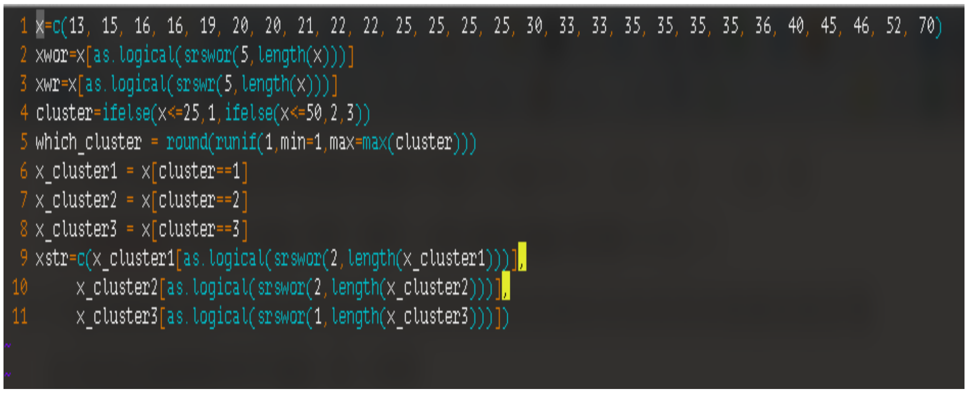
(d) It depends on the actual task/problem. If one wants to preserve the relationships among the original data values, use min-max maybe better. If the actual minimum or maximum of attribute X are unknown, or when there are outliers that dominate the min-max normalization, z-score normalization may be preferred. If one want to reduce the effect of outliers, z-score(with mean absolute deviation) normalization may be used. If one want more weight for numbers with larger absolute value, decimal scaling might be take precedence over others.

9. 3.11 Age distribution: X= [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) equal-width histogram of width 10:

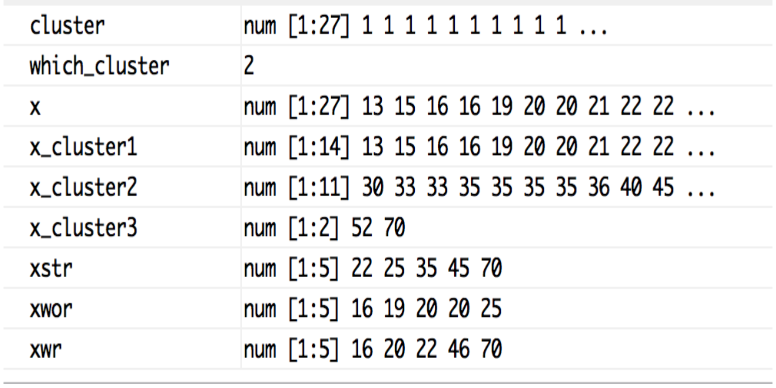
(b) Sketch examples with: SRSWOR, SRSWR, clustering sampling, and stratified sampling. Use sample of size 5 and the strata "youth", "middle-aged", and "senior".

Using R to generate these four sampling, chose(0,25],(25,50],(50,75] as three clusters, and the sampling cluster result in cluster2(middle-aged). For stratified sampling, chose 2 value from 'youth', 2 value from 'middle -aged', 1 value from 'senior'. The code in R is represented below as fig3, and sketch fig is fig3(xstor is SRSWOR sampling, xwr is SRSWR sampling, cluster 2 (which cluster) was sampled as result in clustering sampling, xstr is for stratified sampling)



```r
x=c(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)
xwor=x[as.logical(srswor(5,length(x)))]
xwr=x[as.logical(srswr(5,length(x)))]
cluster=ifelse(x<=25,1,ifelse(x<=50,2,3))
which_cluster = round(runif(1,min=1,max=max(cluster)))
x_cluster1 = x[cluster==1]
x_cluster2 = x[cluster==2]
x_cluster3 = x[cluster==3]
xstr=c(x_cluster1[as.logical(srswor(2,length(x_cluster1)))],
     x_cluster2[as.logical(srswor(2,length(x_cluster2)))],
     x_cluster3[as.logical(srswor(1,length(x_cluster3)))])
```

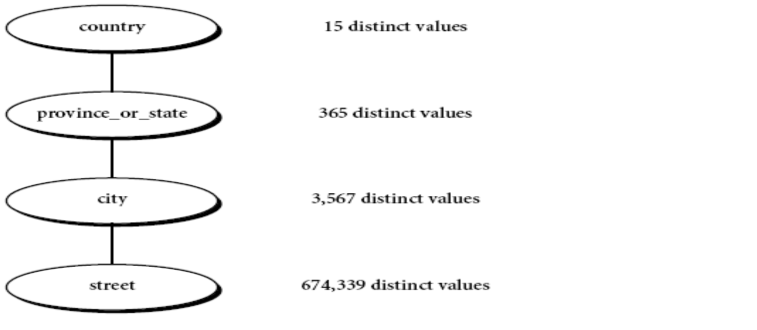
(a) R_code



cluster	num [1:27] 1 1 1 1 1 1 1 1 1 1 ...
which_cluster	2
x	num [1:27] 13 15 16 16 19 20 20 21 22 22 ...
x_cluster1	num [1:14] 13 15 16 16 19 20 20 21 22 22 ...
x_cluster2	num [1:11] 30 33 33 35 35 35 35 36 40 45 ...
x_cluster3	num [1:2] 52 70
xstr	num [1:5] 22 25 35 45 70
xwor	num [1:5] 16 19 20 20 25
xwr	num [1:5] 16 20 22 46 70

(b) Sketched sampling

Figure 3: R_code and sketched samplings



Automatic generation of a schema concept hierarchy based on the number of distinct attribute values.

Figure 4: hierarchy for AllElectrons

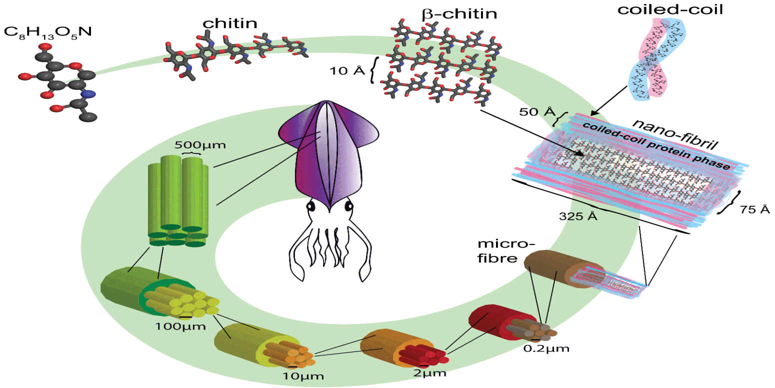


Figure 5: Bio-information or chemistry application

10. Lacking of practical or specific data, generate algorithms for AllElectronics database as shown in the book (pseudocode):

    (a) Hierarchy, nominal data, basing on distinct values.

      i. Sort(attributes, 'ascending'); # AllElectronics has attributes with distinct values: country(15), province_state(365), city(3567), streets(674,339)

      ii. Hierarchy(attributes) # According to sorted data, put country(15) the first attribute, province_state(365) second, city(3567) third, streets(674,339) the last attribute

      iii. Modifying(hierarchy) # to reflect desired semantic relationships or to target specific task

    This kind of hierarchy may also be used in bio-information:

    (b) Hierarchy, numeric data, basing on equal-width partitioning rule.

      i. Prices=[0, ... x_i]; Prices=Sort(Prices); # Sort Prices in AllElectronic sells

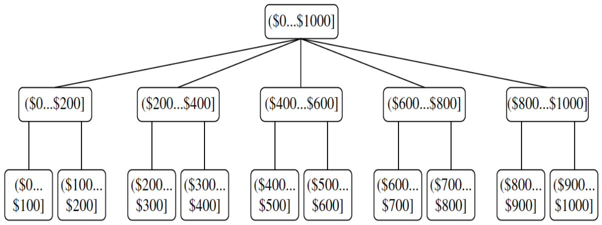


Figure 3.12 A concept hierarchy for the attribute *price*, where an interval ($*X*...$*Y*] denotes the range from $*X* (exclusive) to $*Y* (inclusive).

Figure 6: hierarchy for AllElectrons Prices

ii. Binning(Prices, interval= 200) # first-level

iii. Binning(Prices, interval=100) # Second level

iv. $\vdots$

From binning and counting, one can generate equal-width partition hierarchy(see fig 6).