Data Mining Assignment 2
Liqi Zhu

2.3

$$median = L1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

$$\frac{200 + 450 + 300 + 1500 + 700 + 44}{2} = 1597$$

So the median interval is 21-50 age group

$$medain = 21 + \left(\frac{1597 - 950}{1500} \right) \times 30 = 33.94 \approx 34$$

2.6
(22,1,42,10) (20,0,36,8)
(a)Euclidean distance

$$= \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \left(x_{i3} - x_{j3}\right)^2 + \left(x_{i4} - x_{j4}\right)^2}$$

$$= \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2}$$

$$= 6.7082$$

(b)Manhattan distance

$$= \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \left|x_{i3} - x_{j3}\right| + \left|x_{i4} - x_{j4}\right|$$

$$= |22-20| + |1-0| + |42-36| + |10-8|$$

$$= 11$$

(c)Minkowski distance

$$= \sqrt[h]{\left(x_{i1} - x_{j1}\right)^h + \left(x_{i2} - x_{j2}\right)^h + \left(x_{i3} - x_{j3}\right)^h + \left(x_{i4} - x_{j4}\right)^h}$$

$$h \geq 1$$

When h=3

Minkowski distance= $\sqrt[3]{233} = 6.1534$

(d)Supremum distance

$$= \max_f^p \left|x_{if} - x_{jf}\right| = 6$$

2.7

With all kinds of data sets, the most straight forward method to do approximation is the method used for question 2.3:

$$median = L1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

where L1 is the lower boundary of the median interval, N is the number of values in the entire data set, (freq)l is the sum of the frequencies of all of the intervals that are lower than the median interval, freqmedian is the frequency of the median interval, and width is the width of the median interval.

One other approach is that in a huge data set, divide the data into several intervals, then divide

the interval into sub intervals, and another level of sub intervals if needed until you have enough intervals, when doing calculations, use the same formula in question 2.3. Calculate the median of sub intervals from the lowest level to the first level to find the median of the data set.

As we know the number of errors decreased as the number of the intervals increases. And obviously time used to process will increase. Since the time used in the calculation is a good index of complexity, the number of errors is a good index for accuracy. So I propose that the product of time used times the number of errors is a good measure to find the most balanced one with acceptable error amount and time.

2.8

$$sim(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

(a)Cosine similarity

With the formulas in 2.6

The results:

	A1	A2	Euclidean	Manhattan	Supremum	Cosine
x	1.4	1.6	distance	Distance	distance	similarity
x1	1.5	1.7	0.141421	0.2	0.1	0.999991
x2	2	1.9	0.670820	0.9	0.6	0.995752
x3	1.6	1.8	0.282843	0.4	0.2	0.999969
x4	1.2	1.5	0.223607	0.3	0.2	0.999028
x5	1.5	1	0.608276	0.7	0.6	0.965363

Similarity using Euclidean distance: x1>x4>x3>x5>x2

Manhattan distance: x1>x4>x3>x5>x2

Supremum distance: x1>x4,x3>x5,x2

Cosine similarity: x1>x3>x4>x2>x5

(b) After normalization Data points:

	A1	A2
x	0.65850461	0.752577
x1	0.66162164	0.749838
x2	0.72499943	0.688749
x3	0.66436384	0.747409
x4	0.62469505	0.780869
x5	0.83205029	0.5547

Based on Euclidean distance formula, the results:

	A1	A2	Euclidean
x	0.65850461	0.752577	distance
x1	0.66162164	0.749838	0.004149
x2	0.72499943	0.688749	0.092171
x3	0.66436384	0.747409	0.007812
x4	0.62469505	0.780869	0.044085
x5	0.83205029	0.5547	0.263198

So the similarity after normalization: x1>x3>x4>x2>x5

3.1

Accuracy: The degree to which data correctly describes the "real world" object or event being described.

Completeness:The proportion of stored data against the potential of "100% complete".

Consistency: The absence of difference, when comparing two or more representations of a thing against a definition.
Data quality depends on the intended use of the data. For example, for any big company in the fast moving customer industry. It's so important to be consistent in data values among all departments so that every department is able to accomplish its function properly. Any mistake in database of sales volume or price can cause the false auditing and financial problems. If the data miss values then there must be additional balance sheet issues. Any problem above can cause false marketing strategy and false budget.
And also the customer's data directly reflects the status of the market. If there's error in database such as wrong value, missing value or conflicting values then it's more likely to cause loss since there's no clear image of the market. Therefore, there's companies putting so much emphasis on the quality of the data with intended use of data.
Other dimensions that can be used to assess the quality of data include timeliness, believability, value added, interpretability and accessability:
Timeliness: Data must be available within a time frame that allows it to be useful for decision making.
Believability: Data values must be within the range of possible results in order to be useful for decision making.
Value added: Data must provide additional value in terms of information that offsets the cost of collecting and accessing it.
Interpretability: Data must not be so complex that the effort to understand the information it provides exceeds the benefit of its analysis.
Accessibility: Data must be accessible so that the effort to collect it does not exceed the benefit from its use.

3.3
(a)
1.Partition into (equal-frequency) bins:
Bin 1: 13, 15, 16 Bin 4: 22, 25, 25 Bin 7: 35, 35, 35
Bin 2: 16, 19, 20 Bin 5: 25, 25, 30 Bin 8: 36, 40, 45
Bin 3: 20, 21, 22 Bin 6: 33, 33, 35 Bin 9: 46, 52, 70.

2.Calculate the mean value of each bin.
3.Replace the value with the mean value.
Smoothing by bin means:

Bin 1: 14.67, 14.67, 14.67	Bin 4: 24, 24, 24	Bin 7: 35, 35, 35
Bin 2: 18.33, 18.33, 18.33	Bin 5: 26.67, 26.67, 26.67	Bin 8: 40.33, 40.33, 40.33
Bin 3: 21, 21, 21	Bin 6: 33.67, 33.67, 33.67	Bin 9: 56, 56, 56.

(b)
Outliers may be detected by clustering, for example, where similar values are organized into groups, or 'clusters.' Intuitively, values that fall outside of the set of clusters may be considered outliers.

(c)
Other methods that can be used include forms of binning such as smoothing by bin medians or smoothing by bin boundaries. Equal-width bins can be used to all forms of binning, interval range of values in each bin is constant.

Methods other than binning include using regression techniques to smooth the data by fitting it to a function such as through linear or multiple regression. Classification techniques can be used to implement concept hierarchies that can smooth the data by rolling-up lower level concepts to higher-level concepts.

3.5

(a) min-max normalization

Value range: [new min, new max]

(b) z-score normalization

Value: $\left[\frac{v_{\min}-\overline{A}}{\sigma_A}, \frac{v_{\max}-\overline{A}}{\sigma_A}\right]$

Value Range: (-∞,+∞)

(c) z-score normalization using the mean absolute deviation instead of standard deviation

$$s_A = \frac{1}{n}(\left|v_1-\overline{A}\right|+\left|v_2-\overline{A}\right|+\left|v_3-\overline{A}\right|+\ldots+\left|v_n-\overline{A}\right|)$$

Value: $\left[\frac{v_{\min}-\overline{A}}{s_A}, \frac{v_{\max}-\overline{A}}{s_A}\right]$

Value Range: (-∞,+∞)

(c) normalization by decimal scaling

Value range: (-1,1).

3.7

(a)

$$v_i' = \frac{v_i - \min_A}{\max_A - \min_A}(new\max - new\min) + new\min = \frac{v_i - 13}{70-13}$$

vi=35 vi'=0.385964912

(b)

$$\overline{A} = \frac{\sum v_i}{27} = 29.96296$$

$$v_i' = \frac{v_i - \overline{A}}{\sigma_A}$$

$$v_i = 35, v_i' = \frac{35 - 29.96296}{12.94} = 0.38926$$

(c)

$$v_i' = \frac{v_i}{10^j}, \max\left|v_i'\right| < 1$$

$$\max v_i = 70$$

$$j = 2$$

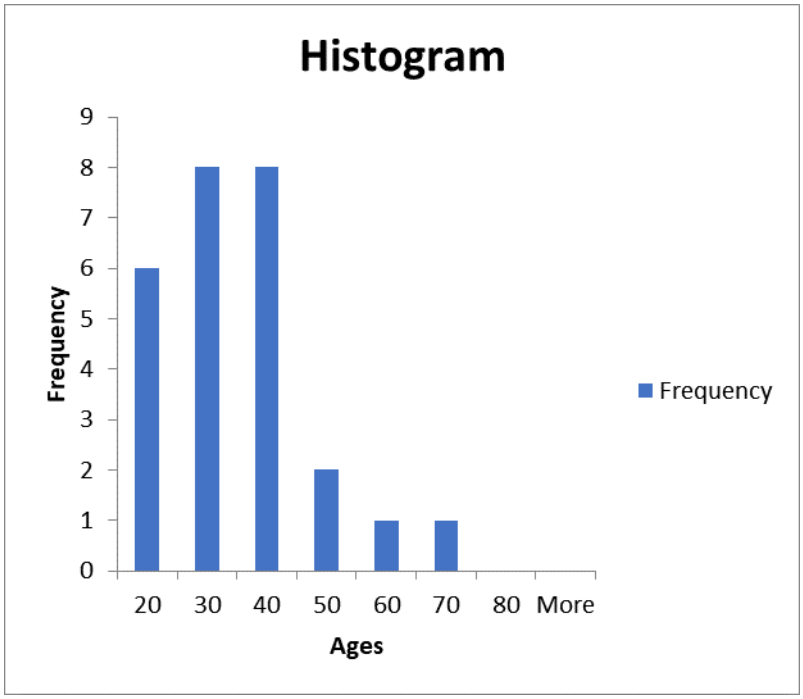
$$v_i = 35, v_i' = \frac{35}{10^2} = 0.35$$

(d)

Because this conversion will keep the data distributed and intuitively interpreted, easy to adjust if there's new value and allow further mining process in all age groups since it doesn't affect distribution. Minimal - Maximum Normalization has an undesirable effect that does not allow any future value to fall outside the current minimum and maximum values, there can be new data in the future exceeds the current value range. Z Fractional Normalization's transformation represents the distance between value and the mean, measured as a standard deviation. Compared to decimal scaling, this kind of normalization may not be intuitive enough.

3.11

(a)



(b)

T1	13
T2	15
T3	16
T4	16
T5	19
T6	20
T7	20
T8	21
T9	22

Tuples

T10	22
T11	25
T12	25
T13	25
T14	25
T15	30
T16	33
T17	33
T18	35

T19	35
T20	35
T21	35
T22	36
T23	40
T24	45
T25	46
T26	52
T27	70

SRSWOR

T3	16
T7	20
T9	22
T15	30
T21	35

SRSWR

T7	20
T14	25
T16	33
T16	33
T26	52

Cluster Sample

T1	13
T2	15
T3	16
T4	16
T5	19

T6	20
T7	20
T8	21
T9	22
T10	22

T11	25
T12	25
T13	25
T14	25
T15	30

T16	33
T17	33
T18	35
T19	35
T20	35

T21	35
T22	36
T23	40
T24	45
T25	46

T26	52
T27	70

m=2

T1	13
T2	15
T3	16
T4	16
T5	19

T21	35
T22	36
T23	40
T24	45
T25	46

Startified Sample

T1	13	young
T2	15	young
T3	16	young
T4	16	young
T5	19	young
T6	20	young
T7	20	young
T8	21	young
T9	22	young

T10	22	young
T11	25	young
T12	25	young
T13	25	young
T14	25	young
T15	30	middle age
T16	33	middle age
T17	33	middle age
T18	35	middle age

T19	35	middle age
T20	35	middle age
T21	35	middle age
T22	36	middle age
T23	40	middle age
T24	45	middle age
T25	46	middle age
T26	52	middle age
T27	70	senior

3.13

Refer to assignment2.py uploaded.