Homework Assignment 1

IT 531, Summer 2018

Due Date – June 29

**Question 1**

Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 23 | 23 | 27 | 27 | 32 | 41 | 41 | 47 | 50 |
| %fat | 9 | 26 | 7 | 17 | 31 | 25 | 27 | 27 | 31 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 52 | 54 | 54 | 56 | 57 | 57 | 58 | 60 | 61 |
| %fat | 34.6 | 42 | 28 | 30 | 31.4 | 34 | 32 | 41 | 25 |

1. Calculate the mean, median, and standard deviation of age and %fat.

Age: mean = 45.55555556, median = 51, standard deviation=13.52654133

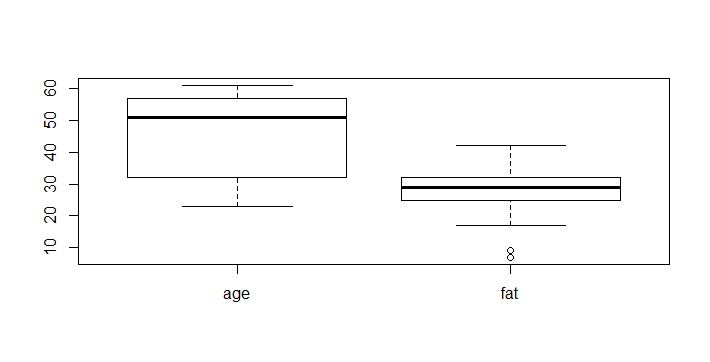
%fat: mean = 27.66666667 , median = 29, standard deviation=9.200767231

1. Draw the boxplots for age and %fat.

Answer:

Excel format

R studio format



**Question 2**

Consider 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}. Answer following:

1. 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.

Answer:

|  |  |  |
| --- | --- | --- |
| Step1 | Sort the data | The data has been sorted already |
| Step2 | Partition the data into equal-frequency bins of size 3 | Bin1: 13, 15, 16  Bin2: 16, 19, 20  Bin3: 20, 21, 22  Bin4: 22, 25, 25  Bin5: 25, 25, 30  Bin6: 33, 33, 35  Bin7: 35, 35, 35  Bin8: 36, 40, 45  Bin9: 46, 52, 70 |
| Step3 | Calculate the means of bins | Bin1: 14.3  Bin2: 18.3  Bin3: 21  Bin4: 24  Bin5: 26.6  Bin6: 33.6  Bin7: 35  Bin8: 40.3  Bin9: 56 |
| Step4 | Smoothing by bin means | Bin1: 14.3, 14.3, 14.3  Bin2: 18.3, 18.3, 18.3  Bin3: 21, 21, 21  Bin4: 24, 24, 24  Bin5: 26.6, 26.6, 26.6  Bin6: 33.6, 33.6, 33.6  Bin7: 35, 35, 35  Bin8: 40.3, 40.3, 40.3  Bin9: 56, 56, 56 |

It can smooth the data in every bins, but if we use the higher depth, the effect will be better.

1. How might you determine outliers in the data?

|  |  |  |
| --- | --- | --- |
|  | z-score | > 1.5 iqr |
| 13 | -1.310678438 | FALSE |
| 15 | -1.156144299 | FALSE |
| 16 | -1.07887723 | FALSE |
| 16 | -1.07887723 | FALSE |
| 19 | -0.847076021 | FALSE |
| 20 | -0.769808952 | FALSE |
| 20 | -0.769808952 | FALSE |
| 21 | -0.692541882 | FALSE |
| 22 | -0.615274813 | TRUE |
| 22 | -0.615274813 | TRUE |
| 25 | -0.383473604 | TRUE |
| 25 | -0.383473604 | TRUE |
| 25 | -0.383473604 | TRUE |
| 25 | -0.383473604 | TRUE |
| 30 | 0.002861743 | TRUE |
| 33 | 0.234662952 | TRUE |
| 33 | 0.234662952 | TRUE |
| 35 | 0.389197091 | TRUE |
| 35 | 0.389197091 | TRUE |
| 35 | 0.389197091 | TRUE |
| 35 | 0.389197091 | TRUE |
| 36 | 0.46646416 | TRUE |
| 40 | 0.775532438 | TRUE |
| 45 | 1.161867786 | TRUE |
| 46 | 1.239134855 | TRUE |
| 52 | 1.702737272 | TRUE |
| 70 | 3.093544523 | TRUE |
| mean | 29.96296296 |  |
| standard deviation | 12.94212407 |  |
| IQR | 14.5 |  |
| 1.5\*IQR | 21.75 |  |

Answer:

We can use IQR range,Z-score and clustering methods.

|  |  |
| --- | --- |
| IQR | We can find lots of data exceed the range of 1.5\* IQR, so it is not susceptible for this situation |
| Z score | If we determine the outliers are z-score >3 or z-score <-3 , we can treat 70 as an outlier |
| clustering | We can group by this data, we can find that there are both only one sample in the age of 50~60 group and 60~70 group, so we can treat 52,70 as outliers |

1. What other methods are there for data smoothing?

Answer : in binning methods, we can also use smoothing by median or smoothing by boundaries, or we can use regression method to predict the value of attributes.

**Question 3**

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

(a) Compute the Euclidean distance between the two objects.

(b) Compute the Manhattan distance between the two objects.

(c) Compute the Minkowski distance between the two objects, using h=3.

(d) Compute the supremum distance between the two objects.

Answer :

(a) Euclidean distance is ~ 6.40

1. Manhattan distance is 9
2. Minkowski distance is ~3.87 using h=3
3. supremum distance is the maximum distance between two objects, which is (42-36) = 6

**Question 4**

Use following methods to normalize the following group of data: 200,300,400,600,1000

(a) min-max normalization by setting min = 0 and max = 1

(b) z-score normalization

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

(d) normalization by decimal scaling

Answer :

1. min-max normalization (min=0;max=1) are 0,0.125,0.25,0.5,1
2. z-score normalization are -0.94,-0.63,-0.31,0.31,1.58
3. z-score normalization using the mean absolute deviation are -1.25,-0.83,-0.41,0.41,2.08
4. decimal scaling are 0.02,0.03,0.04,0.06,0.1 (to make the maximum absolute number 1000 after normalization to <1)

**Question 5**

A database has five transactions. Let min sup = 60% and min conf = 80%

|  |  |
| --- | --- |
| TID | Items\_bought |
| T100 | {M, O, N, K, E, Y} |
| T200 | {D, O, N, K, E, Y} |
| T300 | {M, A, K, E} |
| T400 | {M, U, C, K, Y} |
| T500 | {C, O, O, K, I, E} |

Find all frequent itemsets using Apriori and FP-growth

Answer:

1. Apriori

C1

|  |  |
| --- | --- |
| {M} | 3 |
| {O} | 3 |
| ~~{N}~~ | ~~2~~ |
| {K} | 5 |
| {E} | 4 |
| {Y} | 3 |
| ~~{D}~~ | ~~1~~ |
| ~~{A}~~ | ~~1~~ |
| ~~{U}~~ | ~~1~~ |
| ~~{C}~~ | ~~2~~ |
| ~~{I}~~ | ~~1~~ |

L1

|  |  |
| --- | --- |
| {M} | 3 |
| {O} | 3 |
| {K} | 5 |
| {E} | 4 |
| {Y} | 3 |

C2

|  |  |
| --- | --- |
| ~~{M,O}~~ | ~~1~~ |
| {M,K} | 3 |
| ~~{M,E}~~ | ~~2~~ |
| ~~{M,Y}~~ | ~~2~~ |
| {O,K} | 3 |
| {O,E} | 3 |
| ~~{O,Y}~~ | ~~2~~ |
| {K,E} | 4 |
| {K,Y} | 3 |
| ~~{E,Y}~~ | ~~2~~ |

L2

|  |  |
| --- | --- |
| {M,K} | 3 |
| {O,K} | 3 |
| {O,E} | 3 |
| {K,E} | 4 |
| {K,Y} | 3 |

C3

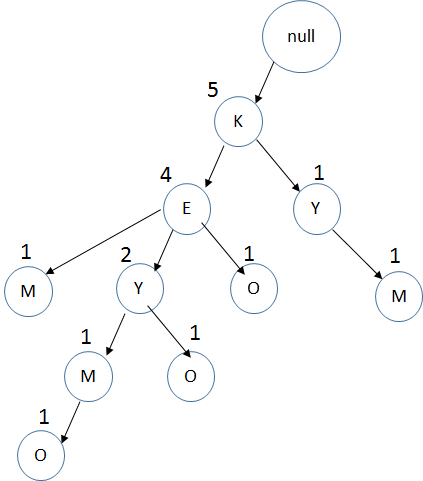
|  |  |
| --- | --- |
| ~~{M,K,O}~~ | ~~1~~ |
| ~~{M,K,E}~~ | ~~2~~ |
| ~~{M,K,Y}~~ | ~~1~~ |
| {O,K,E} | 3 |
| ~~{O,K,Y}~~ | ~~2~~ |
| ~~{K,E,Y}~~ | ~~2~~ |

L3

|  |  |
| --- | --- |
| {O,K,E} | 3 |

1. FP-growth

|  |  |
| --- | --- |
| {K} | 5 |
| {E} | 4 |
| {Y} | 3 |
| {M} | 3 |
| {O} | 3 |
| ~~{N}~~ | ~~2~~ |
| ~~{C}~~ | ~~2~~ |
| ~~{D}~~ | ~~1~~ |
| ~~{A}~~ | ~~1~~ |
| ~~{U}~~ | ~~1~~ |
| ~~{I}~~ | ~~1~~ |



|  |  |  |  |
| --- | --- | --- | --- |
| item | Conditional pattern base | Conditional FP tree | Frequent pattern generated |
| O | {K,E,Y,M:1}  {K,E,Y:1}  {K,E:1} | (K:3)  (E:3) | {K,O:3}  {E,O:3}  {K,O,E:3} |
| M | {K,E,Y:1}  {K,E:1}  {K,Y,M:1} | (K:3) | {K,M:3} |
| Y | {K,E:2}  {K:1} | (K:3) | {K:Y:3} |
| E | {K:4} | (K:4) | {K:E:4} |