Data Analytics: UE18CS312
Question Bank Answers

Unit-1: Exploratory Data Analysis and Visualization

Sl.No	Questions
1.	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 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) What is the mean of the data? What is the median? (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). (c) What is the midrange of the data? (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data? (e) Give the five-number summary of the data. (f) Show a boxplot of the data. (g) How is a quantile–quantile plot different from a quantile plot?
Solutions	(a) What is the *mean* of the data? What is the *median*? The (arithmetic) mean of the data is: $\bar{x} = 1/n\Sigma^{n}_{i=1}$, xi = 809/27 = 30. The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25. (b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35. (c) What is the *midrange* of the data? The midrange (average of the largest and smallest values in the data set) of the data is: (70 + 13)/2 = 41.5 (d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile
(corresponding to the 75th percentile) of the data is: 35.
(e) Give the *five-number summary* of the data.
The five number summary of a distribution consists of the minimum value, first quartile, median
value, third quartile, and maximum value. It provides a good summary of the shape of the
distribution and for this data is: 13, 20, 25, 35, 70.
(f) Show a *boxplot* of the data.
See Figure 1.
(g) How is a *quantile-quantile plot* different from a *quantile plot*?
A quantile plot is a graphical method used to show the approximate percentage of values below
or equal to the independent variable in a univariate distribution. Thus, it displays quantile
information for all the data, where the values measured for the independent variable are plotted
against their corresponding quantile.
A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the
corresponding quantiles of another univariate distribution. Both axes display the range of values
measured for their corresponding distribution, and points are plotted that correspond to the
quantile values of the two distributions. A line ($y = x$) can be added to the graph along with
points representing where the first, second and third quantiles lie, in order to increase the graph's
informational value. Points that lie above such a line indicate a correspondingly higher value for
the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same
quantile. The opposite effect is true for points lying below this line.

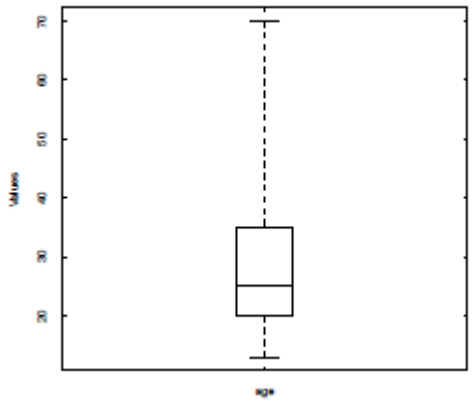


Figure 1. A boxplot of the data.

2. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

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

Compute an approximate median value for the data.

Solution $L_1 = 20,\ n = 3194,\ (\sum_f)_l = 950,\ freq\_median = 1500,\ \text{width} = 30,\ \text{median} = 30.94\ \text{years}.$

3. 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	39	41	47	49	50	52	54	54	56	57	58	58	60	61

% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	24.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median, and standard deviation of age and %fat.
(b) Draw the boxplots for age and %fat.
(c) Draw a scatter plot and a q-q plot based on these two variables.

Soln (a) Calculate the mean, median and standard deviation of age and %fat.
For the variable age the mean is 46.44, the median is 51, and the standard deviation is 12.85. For
the variable %fat the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.

b). Draw the boxplots for age and %fat.

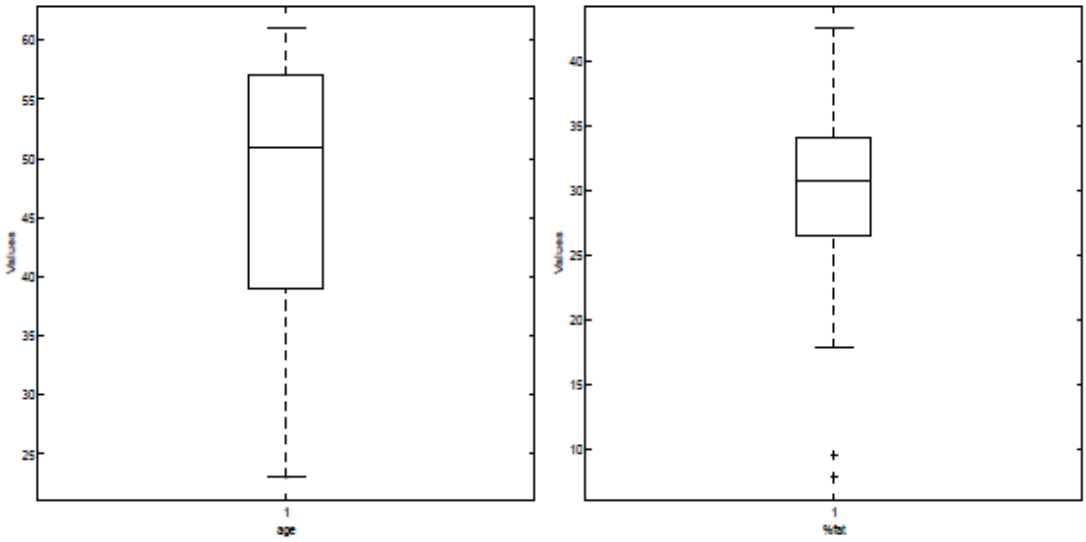


Figure 2: A boxplot of the variables age and %fat in Exercise 2.4.
(c) Draw a scatter plot and a q-q plot based on these two variables.

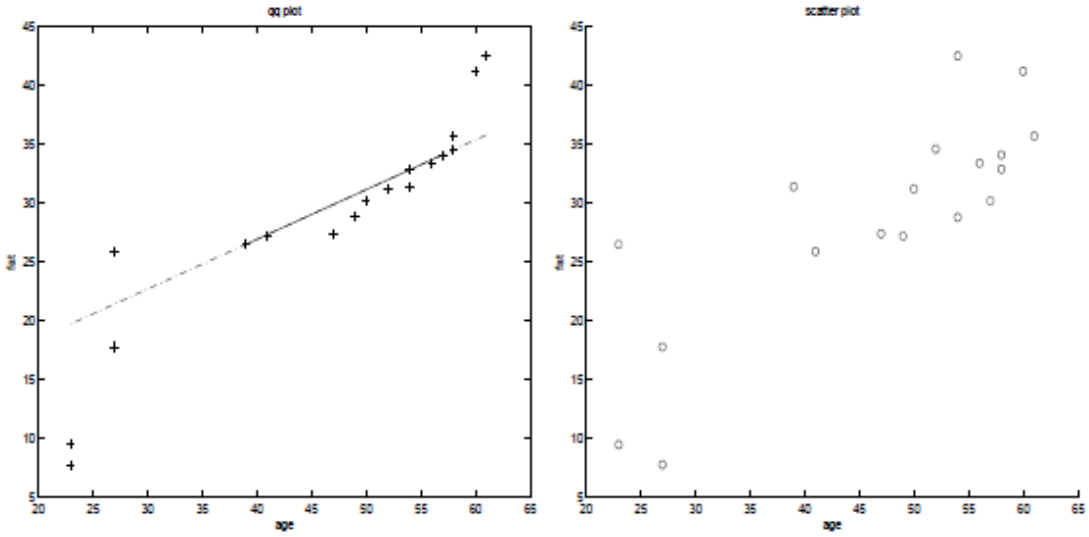


Figure 3: A q-q plot and a scatter plot of the variables age and %fat in Exercise 4.

4. 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.

Soln. There can be various examples illustrating that the assessment of data quality can depend on the
intended use of the data. Here we just give a few.

• For accuracy, first consider a recommendation system for online purchase of clothes. When it
comes to birth date, the system may only care about in which year the user was born, so that it
can provide the right choices. However, an app in facebook which makes birthday calenders for
friends must acquire the exact day on which a user was born to make a credible calendar.

• For completeness, a product manager may not care much if customers' address information is
missing while a marketing analyst considers address information essential for analysis.

• For consistency, consider a database manager who is merging two big movie information databases
into one. When he decides whether two entries refer to the same movie, he may check the entry's

title and release date. Here in either database, the release date must be consistent with the title or there will be annoying problems. But when a user is searching for a movie's information just for entertainment using either database, whether the release date is consistent with the title is not so important. A user usually cares more about the movie's content.

Two other dimensions that can be used to assess the quality of data can be taken from the following:
timeliness, believability, value added, interpretability and accessability. These can be used to assess quality with regard to the following factors:

• 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.

5. Question no. 1 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.
(b) How might you determine outliers in the data?
(c) What other methods are there for data smoothing?

Soln (a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your
steps. Comment on the effect of this technique for the given data.
The following steps are required to smooth the above data using smoothing by bin means with a
bin depth of 3.
• Step 1: Sort the data. (This step is not required here as the data are already sorted.)
• Step 2: Partition the data into equidepth 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
This method smooths a sorted data value by consulting to its ”neighborhood”. It performs local smoothing.

(b) How might you determine outliers in the data?
Outliers in the data may be detected by clustering, where similar values are organized into groups, or ‘clusters’. Values that fall outside of the set of clusters may be considered outliers. Alternatively, a combination of computer and human inspection can be used where a predetermined data distribution is implemented to allow the computer to identify possible outliers. These possible outliers can then be verified by human inspection with much less effort than would be required to verify the entire initial data set.

(c) What other methods are there for data smoothing?
Other methods that can be used for data smoothing include alternate forms of binning such as smoothing by bin medians or smoothing by bin boundaries. Alternatively, equiwidth bins can be used to implement any of the forms of binning, where the 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. Also, classification techniques can be used to implement concept hierarchies that can smooth the data by rolling-up lower level concepts to higher-level concepts.

6. What are the value ranges of the following normalization methods?
(a) min-max normalization
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation
(d) normalization by decimal scaling

Soln. (a) min-max normalization can define any value range and linearly map the original data to this
range.
(b) z-score normalization normalize the values for an attribute A based on the mean and standard deviation. The value range for z-score normalization is

$$[\frac{min_A - A}{\sigma_A}, \frac{max_A - A}{\sigma_A}].$$

(c) z-score normalization using the mean absolute deviation is a variation of z-score normalization by replacing the standard deviation with the mean absolute deviation of A, denoted by sA, which is

$$s_A = \frac{1}{n}(|v_1 - \bar{A}| + |v_2 - \bar{A}| + ... + |v_n - \bar{A}|).$$

The value range is.

$$[\frac{min_A - \bar{A}}{s_A}, \frac{max_A - \bar{A}}{s_A}].$$

(d) normalization by decimal scaling normalizes by moving the decimal point of values of attribute A.

The value range is

$$[\frac{min_A}{10^j}, \frac{max_A}{10^j}],$$

where j is the smallest integer such that

$$Max(|\frac{v_i}{10^j}|) < 1.$$

7. Use these methods to normalize the following group of data:
200, 300, 400, 600,1000
(a) min-max normalization by setting min D 0 and max D 1
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation
(d) normalization by decimal scaling

Soln (a) min-max normalization by setting min = 0 and max = 1 get the new value by computing.

$$v_i' = \frac{v_i - 200}{1000 - 200}(1 - 0) + 0.$$

The normalized data are: 0, 0.125, 0.25, 0.5, 1

(b) In z-score normalization, a value vi of A is normalized to v′
i by computing

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A},$$

where
The normalized data are:

$$\bar{A} = \frac{1}{5}(200 + 300 + 400 + 600 + 1000) = 500,$$

$$\sigma_A = \sqrt{\frac{1}{5}(200^2 + 300^2 + ... + 1000^2) - \bar{A}^2} = 282.8.$$

−1.06,−0.707,−0.354, 0.354, 1.77

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

$$\sigma_A \text{ with } s_A,$$

Where

$$s_A = \frac{1}{5}(|200 - 500| + |300 - 500| + ... + |1000 - 500|) = 240$$

The normalized data are: −1.25,−0.833,−0.417, 0.417, 2.08

(d) The smallest integer j such that

$$Max(|\tfrac{v_i}{10^j}|) < 1 \text{ is } 3.$$

After normalization by decimal scaling, the data become:
0.2, 0.3, 0.4, 0.6, 1.0

8. Using the data for age given in question no.1, answer the following:
(a) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
(b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
(c) Use normalization by decimal scaling to transform the value 35 for age.

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

Soln (a) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
Using the corresponding equation with minA = 13, maxA = 70, new minA = 0, new maxA = 1.0, then v = 35 is transformed to v′ = 0.39.

(b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

Using the corresponding equation where A = 809/27 = 29.96 and σA = 12.94, then v = 35 is transformed to v′ = 0.39.

(c) Use normalization by decimal scaling to transform the value 35 for age.
Using the corresponding equation where j = 2, v = 35 is transformed to v′ = 0.35.

(d) Comment on which method you would prefer to use for the given data, giving reasons as to why. Given the data, one may prefer decimal scaling for normalization as such a transformation would maintain the data distribution and be intuitive to interpret, while still allowing mining on specific age groups.

Min-max normalization has the undesired effect of not permitting any future values to fall outside the current minimum and maximum values without encountering an "out of bounds error". As it is probable that such values may be present in future data, this method is less appropriate. Also, z-score normalization transforms values into measures that represent their distance from the mean, in terms of standard deviations. It is probable that this type of transformation would not increase the information value of the attribute in terms of intuitiveness to users or in usefulness of mining results.

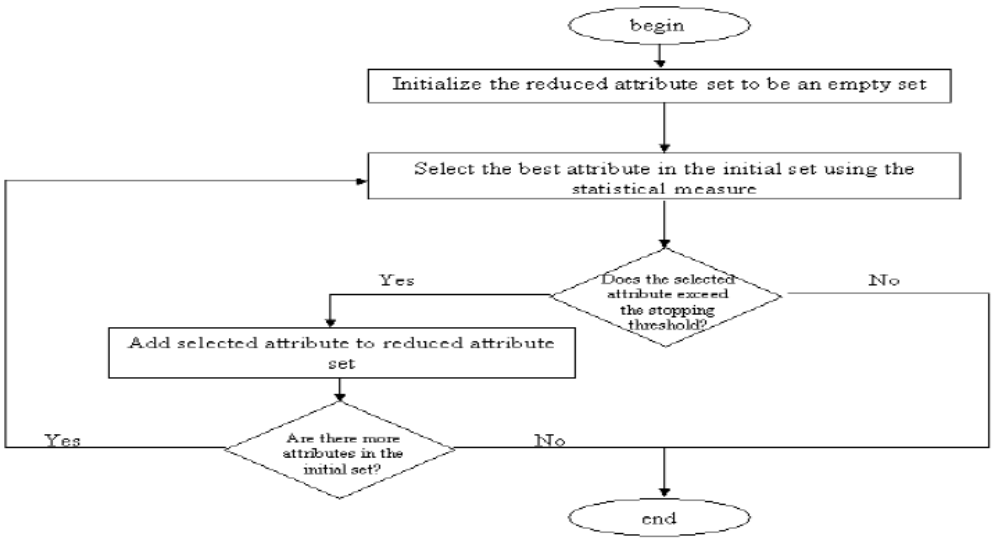
9. Suppose a group of 12 sales price records has been sorted as follows:
5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.
Partition them into three bins by each of the following methods:
(a) equal-frequency (equal-depth) partitioning
(b) equal-width partitioning
(c) clustering

Soln. (a) equal-frequency (equidepth) partitioning
Partition the data into equidepth bins of depth 4:

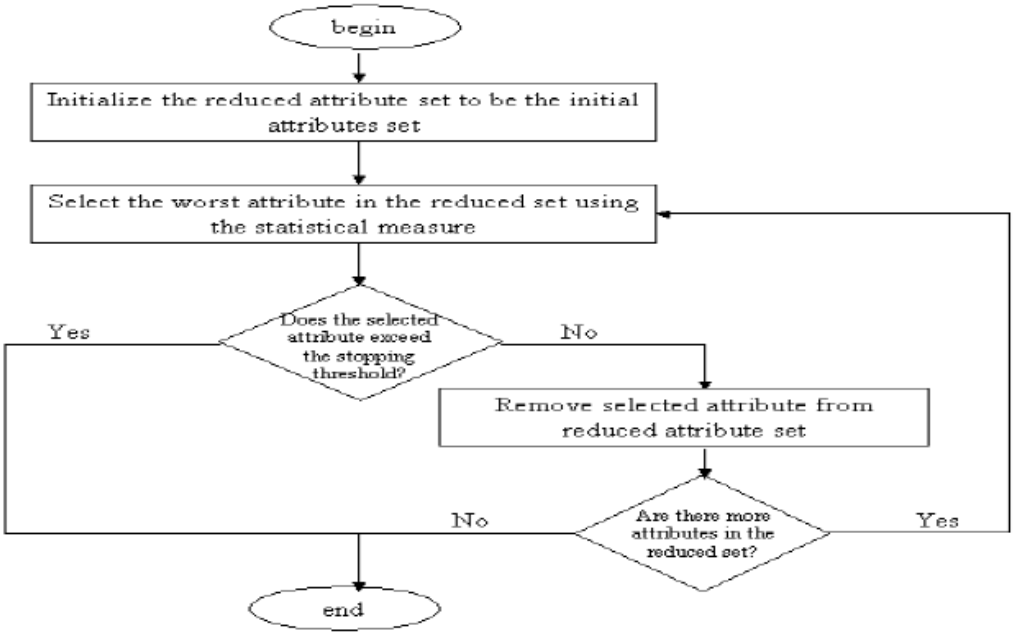
Bin 1: 1: 5, 10, 11, 13 Bin 2: 15, 35, 50, 55 Bin 3: 72, 92, 204, 215
(b) equal-width partitioning
Partitioning the data into 3 equi-width bins will require the width to be (215 − 5)/3 = 70. We
get:
Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72 Bin 2: 92 Bin 3: 204, 215
(c) clustering
Using K-means clustering to partition the data into three bins we get:
Bin 1: 5, 10, 11, 13, 15, 35 Bin 2: 50, 55, 72, 92 Bin 3: 204, 215

10. Use a flowchart to summarize the following procedures for attribute subset selection:
(a) stepwise forward selection
(b) stepwise backward elimination
(c) a combination of forward selection and backward elimination

Soln. Figure A: Stepwise forward selection.
(a) Stepwise forward selection
See Figure A.



(b) Stepwise backward elimination
See Figure B.



(c) A combination of forward selection and backward elimination
See Figure C.