Data Analytics: UE18CS312

Question Bank Answers for Unit 1

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?
Soln	(a) What is the *mean* of the data? What is the *median*? The (arithmetic) mean of the data is: $\bar{x} = 1/n\Sigma^{n}_{i=1}$, x_i = 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 (*Q*1) and the third quartile (*Q*3) 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.

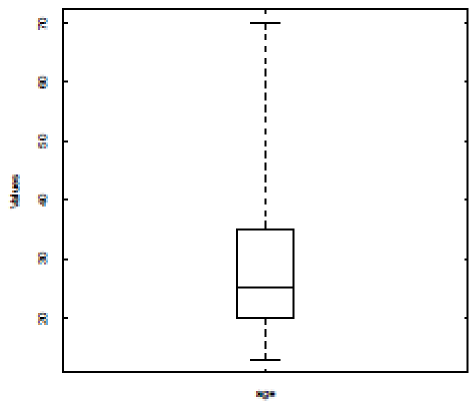


Figure 1. A boxplot of the data.

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.

Soln

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

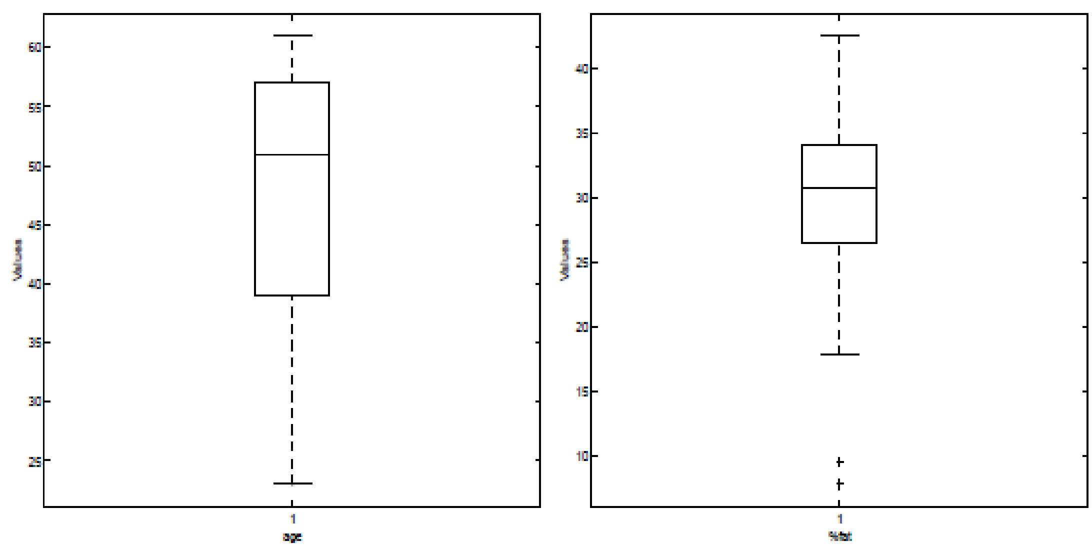
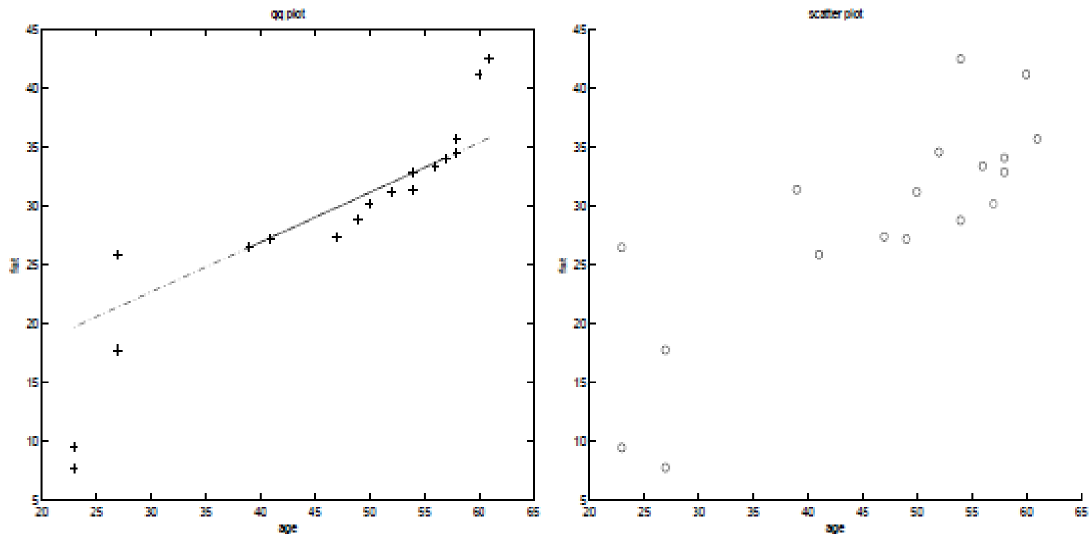


Figure 2: A boxplot of the variables age and %fat in Exercise 2.4.
(c) Draw a scatter plot and a q-q plot based on these two variables.

	 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 calendars 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 accessibility. 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. Clearly show the steps of working. 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. c+ (d-c)*(a – min(A)/ (max(A) - a)) maps data in the range of (min(A), max(A)) to the new range (c,d) (a – min(A)/ (max(A) - a)) maps values in A ranging from (min(A), max(A)) to the new range (0,1) (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 (A-mean(A))/sigma(A) (c) 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 $$Max(\lvert\frac{v_i}{10^j}\rvert) < 1.$$
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. $$v_i' = \frac{v_i - 200}{1000 - 200}(1 - 0) + 0.$$

<table>
<tr><td></td><td>
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

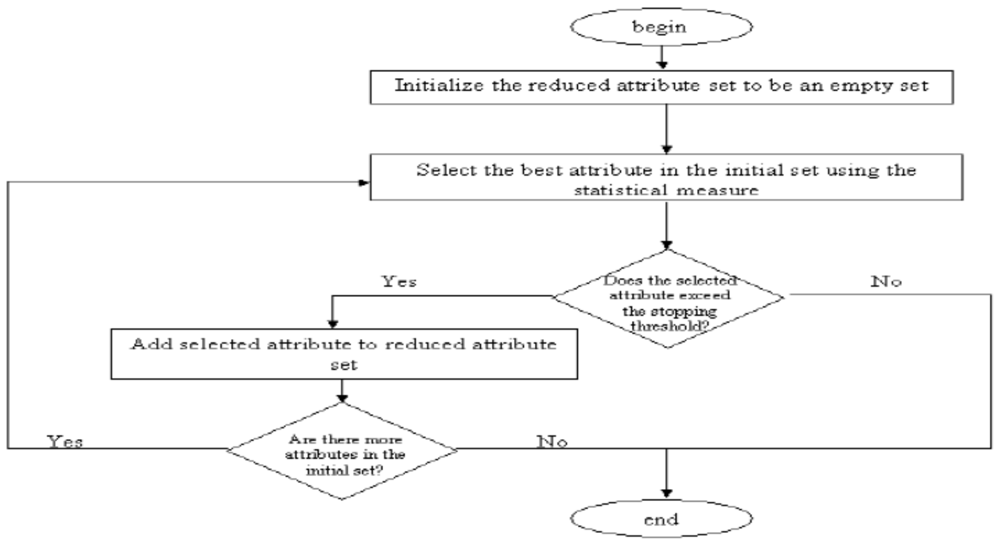
(d) The smallest integer j such that

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

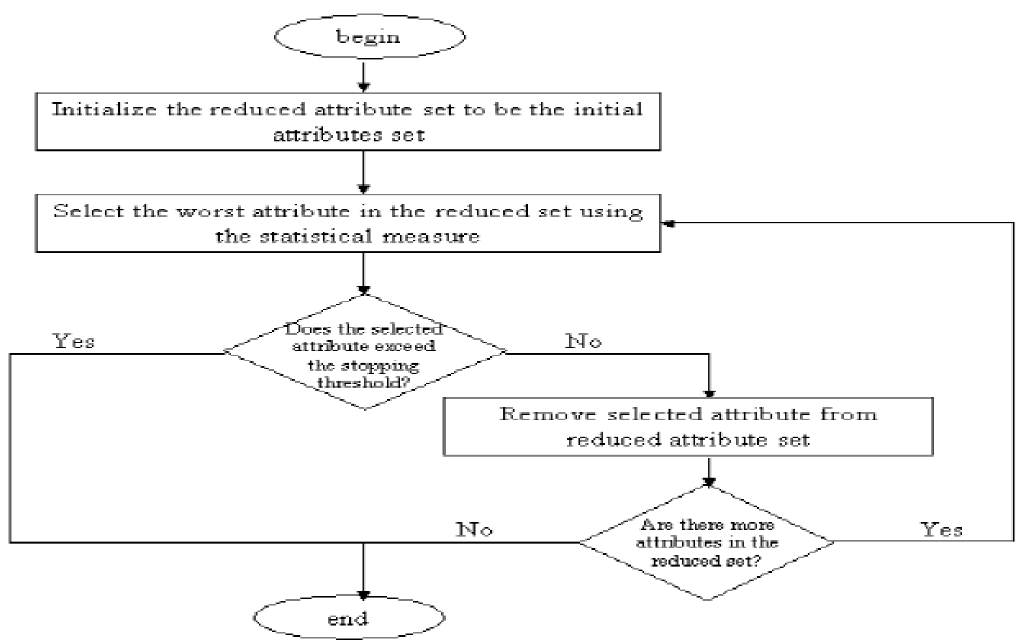
After normalization by decimal scaling, the data become:
0.2, 0.3, 0.4, 0.6, 1.0
</td></tr>
<tr><td>8.</td><td>
Using the data for age given in question no.1, answer the following:
(a) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
(b) Use z-score normalization 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.
</td></tr>
</table>

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

<table>
<tr><td></td><td>(b) stepwise backward elimination
(c) a combination of forward selection and backward elimination</td></tr>
<tr><td>Soln.</td><td>Figure A: Stepwise forward selection.
(a) Stepwise forward selection
See Figure A.

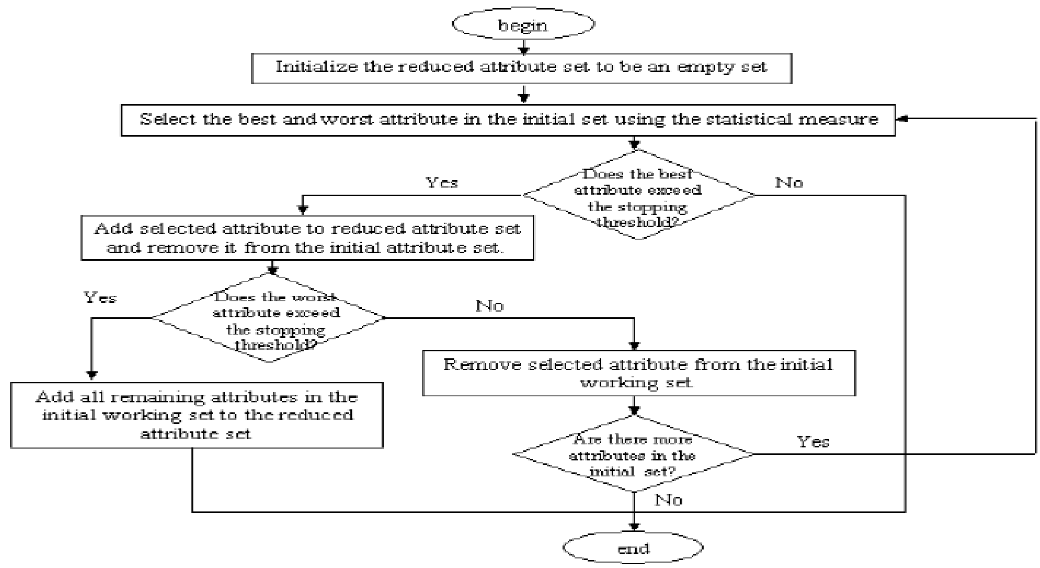


(b) Stepwise backward elimination
See Figure B.



</td></tr>
</table>

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



11

a) Given the marks secured (out of 10) by the top candidates in an entrance test, the time taken to complete that test and their percentage grade in undergrad, answer the following questions:

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Marks	4	8	7	9	7	8	7	6	5
Time (min)	24	25	30	29	28	30	27	30	27
Avg. grades (%)	80	78	89	86	86	88	84	85	79

The range of marks scored in the test and average grades in undergrad are different.

(i) Suggest a transformation that would bring data in both categories to a common range. SRN

(ii) Suggest a method to visualize the data in a single plot or chart to be able to easily select the best performing students (i.e., those who have answered the maximum questions in the least time) by looking at the visualization.

Soln

(i) xi – min/ (max - min) for each range to reduce it to 0 – 1 range
(ii) bar chart with time data reversed (i.e., $time_{max}$ – $time_i$) to ensure the interpretation is consistent or a ratio of maximum marks and time

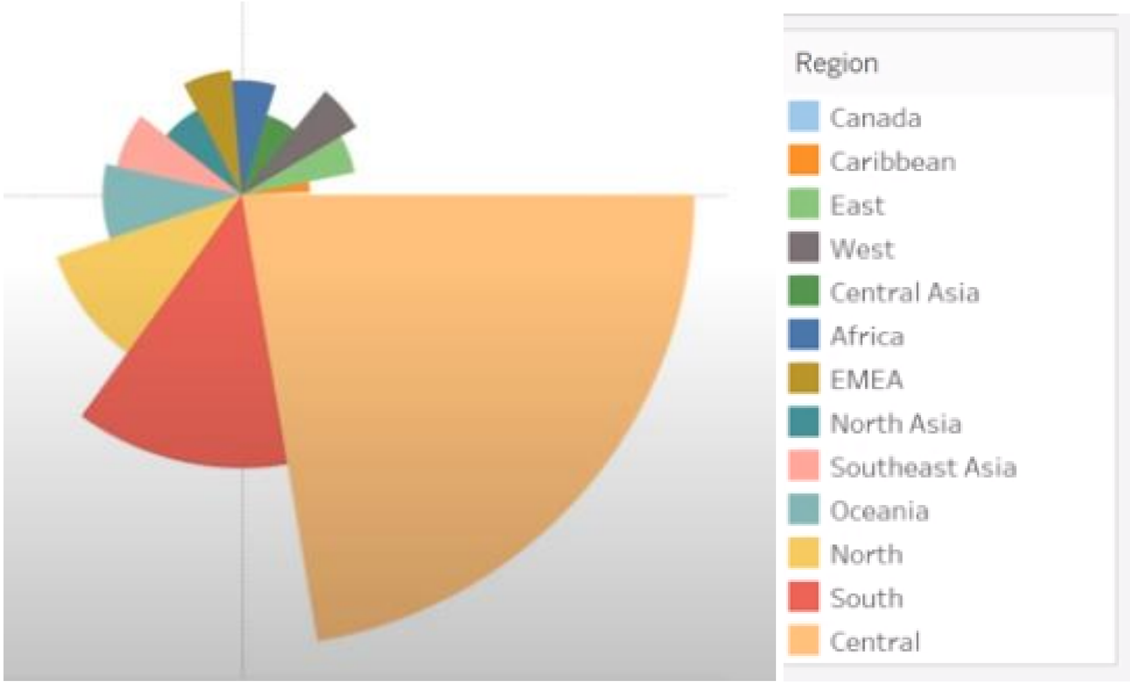
11

b) The testing agency wants to link the performance in the test to a candidate's mastery over a course they offer online. How should an experiment be set up to show the course

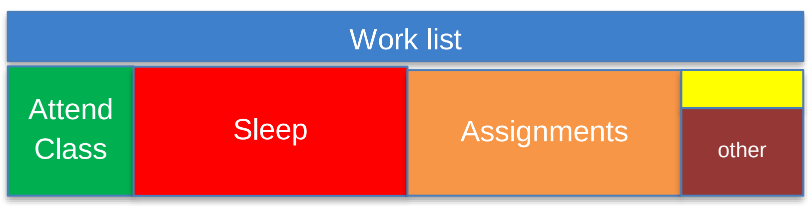
<table>
<tr><td></td><td>indeed helps a candidate do better on the test? The question can be answered in three parts:
(i) What data (any three attributes) could be collected?
(ii) From whom should the data be collected and
(iii) How should it be collected (online survey, etc.)
for the analysis to be meaningful?</td></tr>
<tr><td>Soln</td><td>Open ended:
(i) Whether a student is registered for the course or not, if yes, time spent on viewing the video lectures (recorded automatically), number of timely submissions of assignments, etc., marks scored on the test
(ii) The data for the candidate is registered for the course and marks scored on the test should be collected for everyone (so we can compare results) and the details of time spent on the video lectures and number of timely submissions collected from the students registered for the online course
These could be collected automatically (recorded from the database) so the numbers are accurate and not just ‘as perceived’ by a candidate; we could also conduct a survey to see how much time do they think they spend on a course vs how much they actually spend on it</td></tr>
<tr><td>12</td><td>The table below shows the number of samples for which data is available for five attributes:
<table><tr><td>Attribute</td><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr><tr><td>No of samples</td><td>1,00,468</td><td>2500</td><td>44,000</td><td>1765</td><td>1,14,432</td></tr></table>We intend to analyze the attributes to build a model to predict the value of attribute C in a test set based on one or more of the other attributes that are available for the same data. There are at most 2500 data points for which all attributes are available. Suggest any two ways to deal with missing values in the data to maximize utilizing data that is available and ensure the model is not wrongly biased.</td></tr>
<tr><td>Soln</td><td>Solution: Open ended
interpolation, creating multiple sets of data repeating the smaller attribute, filling in with a median with the pros and cons of whichever approach is suggested</td></tr>
<tr><td>13</td><td>Compute the chi-square statistic for the following data using the formula:
$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$</td></tr>
</table>

<table>
<tr>
<td></td>
<td>
<table>
<tr><th></th><th>Play chess</th><th>Not play chess</th></tr>
<tr><td>Like science fiction</td><td>250</td><td>200</td></tr>
<tr><td>Not like science fiction</td><td>50</td><td>1000</td></tr>
</table>

Does a larger chi-square value indicate the variables are more likely correlated or less?
</td>
</tr>
<tr>
<td>Soln</td>
<td>
$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$
Larger the chi-square value, more likely the variables are correlated.
</td>
</tr>
<tr>
<td>14</td>
<td>
An international superstore has visualized its revenue in various stores across the world in the Coxcomb plot below. Answer the following questions:
(i) Which parts of the world seem to have the highest revenue?
(ii) Which parts of the world seem to have the lowest revenue?
(iii) If we must represent the expenditure on the stores in the various parts of the world, how can it be incorporated in this chart?
</td>
</tr>
</table>

	
Soln	(i) Central Asia seems to have the highest revenue (ii) The lowest revenue seems to be in Central Asia, North Asia and Africa. (iii) The expenditure per region can be represented using a different shade (or different color altogether) as a separate band for each sector, with the area of the band representing the expenditure.
15	What is the difference between the different categories of anomalous data? (i) Inconsistent data (ii) Noisy data (iii) Incomplete data (iv) Intentional data (v) Incorrect data
Soln	(i) Inconsistent data – containing discrepancies (for example, Age 20; birth year 2018) (ii) Noisy data – contains some noise or errors (for example: salary = 10 when the range of numbers is much higher for all other entries) (iii) Incomplete data – lacking attribute values (“ “) or containing aggregates (weekly sales column contains only the annual sales figure) (iv) Intentional data – July 1st marked as everyone’s birthday (v) Incorrect data – data that does not conform to the specified type (name being numeric or phone number having nominal entries – exceptions do exist, but we assess anomalies based on the norm)
16	For the data given below, draw the tree map

<table>
<tr><td></td><td>
<table>
<tr><td>Work list</td><td></td><td></td></tr>
<tr><td>Attend class</td><td>Work list</td><td>4</td></tr>
<tr><td>Sleep</td><td>Work list</td><td>8</td></tr>
<tr><td>Exercise</td><td>Work list</td><td>1</td></tr>
<tr><td>Assignments</td><td>Work list</td><td>8</td></tr>
<tr><td>Other</td><td>Work list</td><td>3</td></tr>
</table>
</td></tr>
<tr><td>Soln</td><td>
<table>
<tr><td>Work list</td><td></td><td>24</td><td>100%</td></tr>
<tr><td>Attend class</td><td>Work list</td><td>4</td><td>16.7%</td></tr>
<tr><td>Sleep</td><td>Work list</td><td>8</td><td>33.3%</td></tr>
<tr><td>Exercise</td><td>Work list</td><td>1</td><td>4.1%</td></tr>
<tr><td>Assignments</td><td>Work list</td><td>8</td><td>33.3%</td></tr>
<tr><td>Other</td><td>Work list</td><td>3</td><td>12.6%</td></tr>
</table>



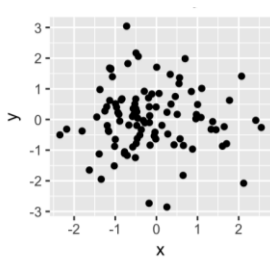
</td></tr>
<tr><td>17</td><td>
Classify the following data as qualitative/ quantitative, ordinal/interval/ ratio and continuous or discrete:

- Time in terms of AM or PM
- Brightness as measured by a light meter
- Brightness as measured by people’s judgments
- Angles as measured in degrees between 0◦ and 360
- Bronze, Silver, and Gold medals as awarded at the Olympics
- Height above sea level
- Number of patients in a hospital
- ISBN numbers for books
</td></tr>
<tr><td>Soln</td><td>

- Time in terms of AM or PM
  - Binary/ Qualitative/ Ordinal
- Brightness as measured by a light meter
  - Continuous, quantitative, ratio
- Brightness as measured by people’s judgments
  - Discrete, qualitative, ordinal
</td></tr>
</table>

	• Angles as measured in degrees between 0◦ and 360
	    • Continuous, quantitative, ratio
	• Bronze, Silver, and Gold medals as awarded at the Olympics
	    • Discrete, qualitative, ordinal
	• Height above sea level
	    • Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)
	• Number of patients in a hospital
	    • Discrete, quantitative, ratio
	• ISBN numbers for books
	    • Discrete, qualitative, nominal (ISBN numbers do have order information, though)

18	Identify the type of datacube operation applied on the Olympic medal tally: • Countries that won gold and silver in the last four Olympics • No of gold medals for each year for each country • Which events did the Gold medals come from for each year? • Winners of the Olympic Bronze in 2004 • Medals won by India, Bhutan and SriLanka in 2016 • Number of Silver medals won by Gender (all countries, all years)
Soln	• Countries that won gold and silver in the last four Olympics - slice (last four Olympics) and dice (countries that won gold and silver) • No of gold medals for each year for each country - Roll-up • Which events did the Gold medals come from for each year? - Drill down operation • Winners of the Olympic Bronze in 2004 - Slicing • Medals won by India, Bhutan and SriLanka in 2016 - Dicing • Number of Silver medals won by Gender (all countries, all years) - Drill down
19	Can we apply PCA to the following data? Briefly explain?

<table>
<tr><td></td><td>
</td></tr>
<tr><td>Soln</td><td>PCA is most suited when data shows high redundancy (i.e., strong correlation between features). This data appears to have low redundancy and hence not suited for PCA.</td></tr>
<tr><td>20</td><td>After eigen analysis, we find the following data:
Eigenvalue
Dim.1 4.124
Dim.2 1.839
Dim.3 1.239
Dim.4 0.819
Dim.5 0.702
Dim.6 0.423
Dim.7 0.303
Dim.8 0.274
Dim.9 0.155
Dim.10 0.122
How many dimensions must we select, if we want to a cumulative variance of 70% or more of the original data to be represented by the principal components?</td></tr>
<tr><td>Soln</td><td>## eigenvalue variance.percent cumulative.variance.percent
Dim.1 4.124 41.24 41.2
Dim.2 1.839 18.39 59.6
Dim.3 1.239 12.39 72.0
Dim.4 0.819 8.19 80.2
Dim.5 0.702 7.02 87.2
Dim.6 0.423 4.23 91.5
Dim.7 0.303 3.03 94.5
Dim.8 0.274 2.74 97.2
Dim.9 0.155 1.55 98.8
Dim.10 0.122 1.22 100.0

Selecting three dimensions or three principal components would suffice.</td></tr>
</table>

<table>
<tr><td>21</td><td>Following are the ages of 15 people interviewed at a shopping center - 12, 14, 15, 19, 21, 27, 31, 32, 46, 53, 56, 57, 58, 59. Describe the shape of the stem and leaf plot of the data</td></tr>
<tr><td>Soln</td><td>The stem and leaf plot is as follows. And since there are two peaks, its bimodal
1 | 2,4,5,9
2 | 1,7
3 | 1,2
4 | 6
5 | 3,6,7,8,9</td></tr>
<tr><td>22</td><td>Given the salaries of male and female research assistants at ImaginaryLab,

<table>
<tr><td></td><td>Female RA’s</td><td>Male RA’s</td></tr>
<tr><td>Number of observations</td><td>403</td><td>132</td></tr>
<tr><td>Mean salaries</td><td>Rs 17,095</td><td>Rs. 14,885</td></tr>
<tr><td>Standard deviation</td><td>6329</td><td>4676</td></tr>
<tr><td>Variance</td><td>40045241</td><td>21864976</td></tr>
</table>

Is there a statistical difference?
(T-table or relevant values will be provided on the test)</td></tr>
<tr><td>Soln</td><td>Null hypothesis: There is no relationship between gender and RA pay
Alt hypothesis: There is a statistically significant relationship between gender and RA pay
Calculate t-statistic
1) subtract the mean of the second group from the mean of the first group
17095-14885=2210
2) calculate, for each group, the variance divided by the number of observations minus 1
Female RA’s:
[40056241 / (403-1)] = [40056241 / (402)] = 99642
Male RA’s:
[21864976 / (132-1)] = [21864976 / (131)] = 166908
3) add the results obtained for each group in step two together
99642+166908=266550
4) take the square root of the results of step three
square root of 266550=516.28
5) divide the results of step one by the results of step four</td></tr>
</table>

	2210/516.28=4.28 To interpret the results, 6) calculate the degrees of freedom: number of observations -2 = 403+132-2 = 533 7) look up the value in the table (4.28) 8) interpret the value of t For a one-tailed test of t, with df=533 and p=.05, t must equal or exceed 1.645. For a two-tailed test of t, with df=533 and p=.05, t must equal or exceed 1.960. In this example, the computed t-score of 4.28 exceeds the table value of t, so we can reject the null hypothesis of no relationship between graduate assistant gender and research assistant pay, and instead accept the research hypothesis and conclude that there is a relationship between graduate assistant gender and RA's pay.
23	Identify the type of sampling methods used in the following examples: (i) Interviewing hockey players as they exit the stadium (ii) Given a restricted sample size of 100, a marketing research ensures households with and without children are equally represented (iii) A sales representative visits a random house in every other cross road of a locality (iv) A restaurant asks every 5th customer to provide a feedback of their service (v) 100 people are selected at random for feedback in every class (2-tier ac, 3-tier ac, 2sleeping berth and general) of a train and given a free t-shirt
Soln	(i) Interviewing hockey players as they exit the stadium Convenience sampling (ii) Given a restricted sample size of 100, a marketing research ensures households with and without children are equally represented Quota-based sampling (iii) A sales representative visits a random house in every other cross road of a locality Systematic sampling (every other cross) followed by simple random sampling without replacement (iv) A restaurant asks every 5th customer to provide a feedback of their service Systematic sampling (v) 100 people are selected at random for feedback in every class (2-tier ac, 3-tier ac, 2sleeping berth and general) of a train and given a free t-shirt Stratified sampling (selecting from each class) followed by simple random sampling without replacement (selecting 100 people at random)
24	List any three different types of redundancy we might see when integrating data?

Soln	Types of redundancy (i) Object duplication (a data record repeats as is) (ii) Entity duplication with different entries (Bill Clinton = William Clinton) (iii) Attribute duplication (an attribute repeats) (iv) Derivable data (one attribute can be derived from another) (v) Tuple duplication
25	What is the curse of dimensionality and how can wavelets help reduce redundancy?
Soln	When the number of attributes of data increases, the data becomes increasingly sparse. Data density is critical for clustering, etc.; with increase in dimensionality (i.e., sparsity in data) the outcome of these operations becomes less meaningful. This is called the curse of dimensionality. Wavelets are a family of transforms that helps remove redundancy by de-correlating features and providing energy compaction. By selecting only the most relevant (high energy) coefficients and essential details, we can eliminate a large number of redundant features. This is particularly useful for time series data and images.