1. A fundamental concept of a data warehouse is which of the following?

a) End-users can update it.
b) It contains numerous naming conventions and formats.
c) Store the data in formats suitable for easy access for decision making.
d) It contains only current data.

2. The generic two-level data warehouse architecture includes which of the following?

a) At least one data mart
b) Data that can be extracted from numerous internal and external sources
c) Near real-time updates
d) All of the above.

3. Consider the relation

Table name: Sale

Date	Customer	Product	Vendor	VendorCity	SalesRep

Table name: Vendor

Date	Customer	Product	Vendor	VendorCity

What is the highest normal form of the sale relation shown in these tables relations?

a) 0NF
b) 1NF
c) 2NF
d) 3NF

4. Which of the following can extract or filter the data & information from the data warehouse?

a) Data redundancy
b) Data recovery tools
c) Data mining
d) Both B and C

The case of customer data, where for each customer we have a very large number of raw attributes (which could be, for example, *independent variables* in the case of a regression used to predict acquisition, purchases, or churn) ranging from demographic information, to what products they bought in the past, who their friends on various social media sites are, what websites they visit more often, how they rated a number of movies or products in general, whether they use their mobile phone mostly the weekends or in the mornings, how they responded to various surveys, etc. One can easily end up with tens if not thousands of such raw attributes, for thousands or millions of customers.

5. Which of the following CRISP-DM phase help in managing the above situation?

a) Business understanding
b) Data preparation
c) Data understanding
d) Modelling

Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city) and the price of the item.

6. Which of the business goal aligned with the data mining goal as mentioned above?

a) Increase catalogue sales to existing customers
b) Introduce catalogue to new customers
c) Understand the demographic of existing customers
d) D. All of the above

7. Before actually building a model, generate a procedure or mechanism to test the model's quality and validity
True

Contacts business analysts and domain experts later in order to discuss the data mining results in the business context

8. Which of the following activity best describe the above situation?
a) Build Model
b) Select model technique
c) Assess Model
d) Generate test design

9. Which of the following CRISP-DM phases help in unveiling additional challenges, information, or hints for future directions?

   a) Assess models
   b) Interpret the model
   c) Evaluation
   d) Modelling

10. Plan monitoring and maintenance help in identifying the following activities EXCEPT:

    a) It helps to avoid unnecessarily long periods of incorrect usage of data mining results
    b) It helps to detect the specific type of deployment
    c) It helps is determine a detailed data monitoring process
    d) It helps to produce a data preparation plan

11. What does "mode" mean?

    a) average
    b) middle
    c) repeats the most
    d) variation

12. Which is not a data cleaning method?

    a) Binning
    b) clustering
    c) regression
    d) aggregation

13. Which of the following is not a normalization method?

    a) min-max normalization
    b) z-score normalization
    c) decimal scaling normalization
    d) schema integration

14. Which of the following are binning types?

    a) smoothing by bin means
    b) smoothing by bin median
    c) smoothing by bin boundaries
    d) All of the above

15. The type of a Nominal attribute depends on which of the following properties?

    a) Distinctness & order

b) Distinctness, order & addition

c) Distinctness

d) All of the above

16. The correct way of pre-processing the data should be:

a) Imputation -> feature scaling-> training

b) Feature scaling->imputation->training

c) Feature scaling->label encoding->training

d) None of the option given

17. 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. What is the mean of the data? What is the median? Which of the following pair is correct?

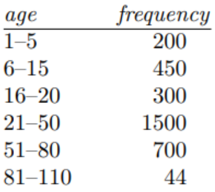
a) Mean = 25, median = 25

b) Mean = 30, median = 25

c) Mean = 30, median = 30

d) None

18. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows. What is the approximate median value for the age?



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

a) 16.73 years

b) 19.11 years

c) 30.94 years

d) 67.12 years

19. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

a) We can just simply remove the rows with missing data values if the data set is large.

b) We can substitute missing values with the mean or average of the rest of the data if the data sets is small.

c) We can substitute missing values with the median of the data if the data sets is small.

d) All of the above.

20. 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. If we partition them into three bins using equal-frequency partitioning, which of the following is correct?

a) Bin 1: 1: 5, 10, 11, 13 Bin 2: 15, 35, 50, 55 Bin 3: 72, 92, 204, 215
b) Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72 Bin 2: 92 Bin 3: 204, 215
c) Bin 1: 5, 10, 11, 13, 15, 35 Bin 2: 50, 55, 72, 92 Bin 3: 204, 215
d) None

21. What data scientist spend the most time doing?

a) Cleaning Data
b) Collecting Data
c) Modelling Data
d) Evaluating Data

22. Analytics is the process of separating a whole problem into its parts so that the parts can be critically examined.
False

23. What is likely to happen is considered as Predictive analytics
True

24. What should i do about it is known as Diagnostic analytics.
False

25. The minimum transaction rate of a hypermarket is RM5000. On a hectic day the transaction can go up to RM15000. What will be the normalized value of a transaction worth RM 7000 given a range of value of 0 to 2?

a) 0.2
b) 0.4
c) 0.6
d) None of the above

26. Which of the following is the best feature selection method to handle dataset that consist of large dimensions of numerical data?

a) step wise forwards selection
b) step wise backward selection
c) PCA
d) Decision Tree

27. A research computed positive correlations that exist between age and height of 100000 individuals from birth to 18 years.

    True

28. Following reduces the amount of information contained in the data except.

    a) Grouping the data
    b) PCA
    c) Data aggregation
    d) None of the above

29. Correlation analysis helps in identifying redundant attributes

    True

30. Suppose that blood sugar levels are normally distributed with a mean of 100 mg/dl (milligrams per deciliter) and a standard deviation of 10mg/dl. David has a blood sugar level of 85 mg/dl. Calculate David's z-score

    a) 1.5
    b) -1.5
    c) 1.05
    d) None of the above

-END OF MID-SEMESTER QUESTION-