PUNE VIDYARTHI GRIHA'S

COLLEGE OF SCIENCE AND TECHNOLOGY

Affiliated to University of Mumbai

Question Bank

Class: T.Y.B. Sc.CS **Semester: VI**

Subject: Data Science

Unit 1 Introduction to Data Science

1. Data types are differed on basis of---------
    a. the way of look
    b. the type of operations
    c. the type of operators used
    d. the way of storage
2. Which of the following step is performed by data scientist after acquiring the data ?
    a. Data Cleansing
    b. Data Integration
    c. Data Replication
    d. All of the Mentioned
3. Which of the following command allows you to update the repository ?
    a. Push
    b. Pop
    c. Update
    d. None of the mentioned
4. Which of the following systems record changes to a file over time ?
    a. Record Control
    b. Version Control
    c. Forecast Control
    d. None of the mentioned
5. Which of the following is good way of performing experiments in data science ?
    a. Measure variability
    b. Un Generalize to the problem
    c. Don't Have Replication
    d. Not analysing
6. Which of the following package is used for reading excel data ?
    a. Xlsx
    b. Xlsc

c. read.sheet
d. All of the Mentioned

7. Which of the following Approach should be used to ask Data Analysis Question?
    a. Data Cleansing
    b. Data Integration
    c. Data Replication
    d. All of the Mentioned
8. Binary attribute are ----------
    a. This takes only two values. In general, these values will be 0 and 1
    b. The natural environment of a certain species
    c. Systems that can be used without knowledge of internal operations
    d. None of these

9. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in following case?
    a. Feature F1 is an example of nominal variable.
    b. Feature F1 is an example of ordinal variable.
    c. It doesn't belong to any of the above category.
    d. Both of these
10. Which of the following is not a NoSQL database?
    a. SQL Server
    b. MongoDB
    c. Cassandra
    d. None of these
11. NoSQL databases is used mainly for handling large volumes of _____ data.
    a. Unstructured
    b. Structured
    c. Semi-Structured
    d. All of these
12. Most NoSQL databases support automatic _____ meaning that you get high availability and disaster recovery.
    a. Processing
    b. Scalability
    c. Replication
    d. All of the mentioned
13. Which of the following is not a data visualization Technique
    a. Histogram
    b. Pie Chart
    c. Box Plot
    d. Docx File
14. Which of these is not a High-Level Language
    a. Fortan
    b. PHP
    c. Java

d. C

15. Which of the following step is performed by data scientist after Cleaning the data ?
    a. Data Processing
    b. Data Collection
    c. Data Analysis
    d. Data Modelling
16. _____ data is difficult to manipulate and typically needs to be processed in some way before it can be used in standard data analysis software.
    a. Structured data
    b. Unstructured data
    c. Summerized data
    d. Frequency data

17. Which of the following function is used for determining missing values?
    a. any
    b. Nan
    c. not
    d. all
18. Point out the correct statement.
    a) Primary data is original source of data
    b) Secondary data is original source of data
    c) Questions are obtained after data processing steps
    d) None of the Mentioned
19. _____ measures asymmetry about the mean of the probability distribution of a random variable.
    a. skewness
    b. covariance
    c. variance
    d. Kurtosis
20. _____shows all individual data points.
    a. Box-plot
    b. scatter plot
    c. line plot
    d. pie chart
21. What is true about Data Visualization?
    a. Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts.
    b. Data Visualization helps users in analyzing a large amount of data in a simpler way.
    c. Data Visualization makes complex data more accessible, understandable, and usable.
    d. Data visualization is an IDE
22. Which of the following is known as raw data?
    a. tidy data
    b. processed data

c. clean data
d. formatted data

23. Data cannot be visualized using?
    a. graphs
    b. charts
    c. maps
    d. image
24. Data visualization is also an element of the broader _____.
    a. deliver presentation architecture
    b. data presentation architecture
    c. dataset presentation architecture
    d. data process architecture

25. Which method shows hierarchical data in a nested format?
    a. Treemaps
    b. Scatter plots
    c. Population pyramids
    d. Area charts
26. Which of the following plots are often used for checking randomness in time series?
    a. Autocausation
    b. Autorank
    c. Autocorrelation
    d. Autocustomize
27. Which one of the following is most basic and commonly used techniques?
    a. Line charts
    b. Scatter plots
    c. Population pyramids
    d. Area charts
28. Which of the intricate techniques is not used for data visualization?
    a. Bullet Graphs
    b. Bubble Clouds
    c. Fever Maps
    d. Heat Maps
29. Data science is the process of diverse set of data through ?
    a. organizing data
    b. processing data
    c. analysing data
    d. All of the above
30. What is the work of Data Architect?
    a. utilize large data sets to gather information that meets their company's needs
    b. work with businesses to determine the best usage of the information yielded from data
    c. build data solutions that are optimized for performance and design applications
    d. Only collection of data

31. Which of the following language is used in Data science?
    a. C
    b. C++
    c. R
    d. Ruby
32. Which of the following is false?
    a. Subsetting can be used to select and exclude variables and observations
    b. Raw data should be processed only one time.
    c. Merging concerns combining datasets on the same observations to produce a result with more variables
    d. Convert raw data to processed data

33. Which of the following is correct skills for a Data Scientist?
    a. Data mining
    b. Machine Learning / Deep Learning
    c. Data collection
    d. Data cleaning
34. Which of the following is not a part of data science process?
    a. Discovery
    b. Model Planning
    c. Communication Building
    d. Operationalize
35. Which of the following are the Data Sources in data science?
    a. Structured/ Unstructured
    b. Numbers
    c. images
    d. Tables
36. Which of the following is not a application for data science?
    a. Recommendation Systems
    b. Image & Speech Recognition
    c. Online Price Comparison
    d. Privacy Checker
37. Which of the following is the most important language for Data Science?
    a. Java
    b. Ruby
    c. R
    d. python
38. Point out the wrong statement.
    a. Merging concerns combining datasets on the same observations to produce a result with more variables
    b. Data visualization is the organization of information according to preset specifications
    c. Subsetting can be used to select and exclude variables and observations
    d. Raw data should be processed only one time

39. Which of the following is characteristic of Processed Data?
    a. Data is not ready for analysis
    b. All steps should be noted
    c. Hard to use for data analysis
    d. Ready data
40. Which of the following is not a step in data analysis?
    a. Obtain the data
    b. Clean the data
    c. EDA
    d. Data modification

41. MongoDB support cross platform and is written in _____language.
    a. C++
    b. R
    c. Java
    d. Python
42. MongoDB is _____ Database.
    a. SQL
    b. NoSQL
    c. RDBMS
    d. DBMS
43. Which of the following is characteristic of Processed Data?
    a) Data is not ready for analysis
    b) All steps should be noted
    c) Hard to use for data analysis
    d) None of the mentioned
44. Which of the following is characteristic of Raw Data?
    a) Data is ready for analysis
    b) Original version of data
    c) Easy to use for data analysis
    d) None of the mentioned
45. A collection and a document in MongoDB is equivalent to..... concepts respectively.
    a. Table and Column
    b. Table and Row
    c. Column and Row
    d. Database and Table
46. Which of the following is correct option ?
    a. MongoDB uses more JSON
    b. MongoDB is row-oriented data store
    c. MongoDB is a NoSQL database
    d. MongoDB is not extensible

47. A collection of related data.
    a. Information
    b. Valuable information
    c. Database
    d. Metadata
48. The restrictions placed on the data
    a. Relation
    b. Attribute
    c. Parameter
    d. Constraint
49. A level that describes data stored in a database and the relationships among the data.
    a. Physical
    b. Logical
    c. User
    d. View
50. Which of the following is false?
    a. data visualization include the ability to absorb information quickly
    b. Data visualization is another form of visual art
    c. Data visualization decrease the insights and take solwer decisions
    d. Data visualization reduce complexity in viewing

Unit 2 Data Curation

51. The _____ clause is used to list the attributes desired in the result of a query.
    a. Where
    b. Select
    c. From
    d. Distinct
52. XML Stands for ----------
    a. Extensible Markup Language
    b. XQuery Markup Language
    c. Extensive Markup Language
    d. None of these
53. Query tools are ----------
    a. A reference to the speed of an algorithm
    b. Attributes of a database table that can take only numerical values.
    c. Tools designed to query a database.
    d. Used to create data
54. XQuery is a functional query language used to retrieve information stored in --- format.
    a. Html
    b. Xml
    c. Uml
    d. Jscript
55. XPath specification has _____ types of nodes
    a. Four

b. Five
c. Six
d. Seven

56. AWS falls into which of the following cloud-computing category?
    a. Platform as a Service
    b. Software as a Service
    c. Infrastructure as a Service
    d. Back-end as a Service
57. What are the Authentication in AWS?
    a. Unlock key
    b. Face recognition
    c. Access Key/ Session Token
    d. Thumb impression

58. _____ provides an web service interface that provides resizable compute capacity in the AWS cloud.
    a. EC2
    b. S3
    c. ES2
    d. EC3
59. _____ is a billing and account management service.
    a. Amazon Mechanical Turk
    b. Amazon Elastic MapReduce
    c. Amazon DevPay
    d. Multi-Factor Authentication
60. _____ is the central application in the AWS portfolio.

    a. Amazon Simple Queue Service
    b. Amazon Elastic Compute Cloud
    c. Amazon Simple Notification Service
    d. All of the above
61. AWS reaches customers in _____countries.
    a. 137
    b. 182
    c. 190
    d. 86
62. S3 stands for
    a. Simple Storage Service
    b. Simple Software Service
    c. Simple Storage Server
    d. Simple Storage Source
63. How many buckets can you create in AWS by default?

a. 100 buckets
b. 200 buckets
c. 110buckets
d. 125 buckets

64. What are the advantages of auto-scaling?
    a. Better availability and Offers fault tolerance
    b. 100%seccurity
    c. Integrity
    d. Compatibility
65. Which of the following is a message queue or transaction system for distributed Internet-based applications?
    a. Amazon Simple Notification Service
    b. Amazon Elastic Compute Cloud
    c. Amazon Simple Queue Service
    d. Amazon Simple Storage System

66. Which of the following is an online backup and storage system?
    a. Amazon Simple Queue Service
    b. Amazon Elastic Compute Cloud
    c. Amazon Simple Notification Service
    d. Amazon Simple Storage System
67. Which of the following statement is wrong about Amazon S3?
    a. Amazon S3 provides large quantities of reliable storage
    b. Amazon S3 is highly available
    c. Amazon S3 is highly reliable
    d. Amazon S3 is highly protected
68. Which service performs the function that when an instance is healthy it is terminated and replaced with a new one?
    a. Sticky Sessions
    b. Fault Tolerance
    c. Connection Draining
    d. None of the above
69. Amazon S3 is which type of storage service?
    a. Block
    b. Object
    c. Simple
    d. Secure
70. Amazon S3 offers encryption services for ..... .
    a. Data in Drive
    b. Data in Motion
    c. Data in Rest
    d. Data in Storage
71. A virtual CloudFront user is called an OAI. This stands for what?
    a. Origin Archive Initiative

b. Origin Access Identity
c. Original Archive Identity
d. Original Accessible Initiative

72. In XQeury _____ symbol preceded before the variable name.
a. @
b. $
c. #
d. *

73. .... function used to open file in xml data.
a. file()
b. fopen()
c. doc()
d. None

74. Which is wrong in XQuery?
a. Used for transforing XML data into XHTML
b. Used for search web documents
c. Used to generate tables for XSLT
d. None of these

75. Point out the correct statement.
a. XQuery statements are case sensitive and xml is case sensitive
b. XQuery statements are case insensitive and xml is case insensitive
c. XQuery statements are case sensitive and xml is case insensitive
d. XQuery statements are case insensitive since xml is case sensitive

76. Odd Man out
a. SQL
b. TMQL
c. XQuery
d. C

77. _____ statement is not used in DDL(Data Definition Language)
a. DROP
b. SELECT
c. CREATE
d. ALTER

78. _____ statement is not used in DML(Data Manipulation Language)
a. SELECT
b. INSERT
c. CREATE
d. DELETE

79. _____ statement is not used in DCL(Data Control Language)
a. COMMIT
b. GRANT
c. REVOKE
d. None of these

80. _____ statement is not used in TCL(Transaction Control Language)

a. BEGIN
b. REVOKE
c. ROLLBACK
d. COMMIT

81. JSON stands for----------
a. JavaScript Oriented Notation
b. JavaScript Object Notation
c. Java Oriented Notation
d. Java Object Notation

82. Which of the following is not a strategy employed in web crawling?
a. Incremental Crawler
b. Focused Crawler
c. Distributed Crawler
d. Centralized Crawler

83. ---------- is an automatic process of extracting information from web
a. Web Crawling
b. Web Scraping
c. Web surfing
d. Web Service

84. ---------- is a platform for code hosting and collaborative development for systems.
a. Github
b. Hbase
c. Repository
d. Postman

85. The concept of distributed database with a goal to improve----------
a. Integrity
b. Security
c. Reliability
d. Consistency

86. A NoSQL database is also referred as -------
a. Relational database
b. non-SQL database
c. structured database
d. semi-structured database

87. AWS stands for------
a. Amazon web service
b. Amazon web series
c. Amazon web source
d. Amazon web development service

88. HBase is a distributed ---------- built on top of Hadoop file system.
a. Row-oriented database
b. Column-oriented database
c. Row-oriented collection

d. Column-oriented collection

89. ---------- is an internet-based computing service in which large group of remote servers are networked to allow centralized data storage ,and online access to computer resources or services.
    a. Distributed computing
    b. Cloud computing
    c. Grid computing
    d. Web services
90. IaaS stands for:
    a. Internet as a Service
    b. Infrastructure as a Service
    c. Information as a Service
    d. Intelligence as a Service

91. PaaS Stands for:
    a. Product as a Service
    b. Program as a Service
    c. Platform as a Service
    d. Platform as a Source
92. SaaS Stands for
    a. Software as a Service
    b. Software as Source
    c. Security as a Service
    d. Security as a Source

Unit 3 Statistical Modelling and Machine Learning

93. Ridge Regression is when data suffers from _____.
    a. Collinearity
    b. Multicollinearity
    c. Does not suffer
    d. Regression
94. Point out the wrong statement.
    a. Simple linear regression is equipped to handle more than one predictor
    b. Compound linear regression is not equipped to handle more than one predictor
    c. Linear regression consists of finding the best-fitting straight line through the points
    d. All of the mentioned
95. Bayesian Information Criterion (BIC) is related to_____.
    a. Ridge regression

b. Akaike Information Criterion (AIC)
c. Cross validation
d. Lasso Regression

96. Joins are used for combining _____ product.
a. Vector
b. Cartesian
c. Scalar
d. Euler

97. Which of the following function gives first few rows of data information from the table?
a. head
b. tail
c. summary
d. none of the mentioned

98. Data that summarize all observations in a category are called _____ data.
a. frequency
b. summarized
c. raw
d. none of the mentioned

99. Which of the following functions is used for k-means clustering?
a. k-means
b. k-mean
c. heatmap
d. none of the mentioned

100. When data are classified according to a single characteristic, it is called:
a. Quantitative classification
b. Qualitative classification
c. Area classification
d. Simple classification

101. Cluster is
a. Group of similar objects that differ significantly from other objects
b. Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm
c. Symbolic representation of facts or ideas from which information can potentially be extracted
d. None of these

102. Which of the following is required by K-means clustering?
a. defined distance metric
b. number of clusters
c. initial guess as to cluster centroids
d. all of the m

103. Suppose we would like to perform clustering on spatial data such as the geometrical locations of houses. We wish to produce clusters of many different sizes and shapes. Which of the following methods is the most appropriate?
    a. Decision Trees
    b. Density-based clustering
    c. Model-based clustering
    d. K-means clustering
104. In classification, the data are arranged according to:
    a. Similarities
    b. Differences
    c. Percentages
    d. Ratios
105. Which of the following function is used for k-means clustering?
    a. k-means
    b. k-mean
    c. heatmap
    d. none of the mentioned
106. Classification is
    a. A subdivision of a set of examples into a number of classes
    b. A measure of the accuracy, of the classification of a concept that is given by a certain theory
    c. The task of assigning a classification to a set of examples
    d. None of these
107. Which of the following curve analysis is conducted on each predictor for classification?
    a. NOC
    b. ROC
    c. COC
    d. All of the mentioned
108. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?
    a. 1
    b.2
    c.3
    d.4
109. Function used for linear regression in R is _____
    a. lm(formula, data)
    b. lr(formula, data)
    c. lrm(formula, data)
    d. regression.linear(formula, data)
110. What plot(s) are used to view the linear regression?
    a. Scatterplot
    b. Box plot
    c. Density plot
    d.Scatterplot, Boxplot, Density plot

-

111. Which of the following is used to plot multiple histograms?
    a. multi.plot()
    b. multi.hist
    c. xyplot.multi()
    d. poly()
112. If a data set or time series is random which of the following plots are used to check?
    a. Random
    b. Lag
    c. Lead
    d. Heat
113. Logistic Regression is a ---------- regression technique that is used to model data having a ---------- outcome
    a. Linear,numeric
    b. Linear ,binary
    c. Non-linear,numeric
    d. Non-linear,binary

114. Machine learning techniques differ from statistical techniques in that machine learning methods
    a. Typically assumes an underlying distribution for the data
    b. Are better able to deal with missing and noisy data.
    c. Are not able to explain their behaviour
    d. Have trouble with large-sized datasets
115. Regarding bias and variance , which of the following statements are true?
    a. Models which overfit have a high bias and underfit have high variance
    b. Models which overfit have a high bias and underfit have low variance
    c. Models which overfit have a low bias and underfit have high variance
    d. Models which overfit have a low bias and underfit have low variance
116. Regression trees are often used to model ------data
    a. Linear
    b. Non-linear
    c. Categorical
    d. Symmetrical
117. Selecting data so as to assure each class is properly represented in both the training and test set
    a. Cross validation
    b. Stratification
    c. Verification
    d. Bootstrapping
118. Simple regression assumes a ---------- relationship between input attribute and output attribute.
    a. Linear
    b. Quadratic
    c. Reciprocal
    d. Inverse

119. This supervised learning technique can process both numeric and categorical input attributes
    a. Linear regression
    b. Logistic regression
    c. Simple regression
    d. Multiple linear regression
120. This technique associates a conditional probability value with each data instance
    a. Linear regression
    b. Logistic regression
    c. Simple regression
    d. Multiple linear regression
121. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration
    a. Agglomerative clustering
    b. Conceptual clustering
    c. k-Means clustering
    d. exception maximization
122. Which of the following method make vector of repeated values?
    a. rep()
    b. data()
    c. view()
    d. read()

123. Which of the following finds the position of a quantile in a dataset?
    a) quantile()
    b) barplot()
    c) barchart()
    d) rep()

124. Which of the following function cross-tabulate tables using formulas?
    a) table
    b) stem
    c) xtabs
    d) read

125. Which of the following groups find the correlation matrix?
    a) factor.model
    b) col.max(x)
    c) stem
    d) which.max(x)

126. Which of the following is lattice command for producing a scatterplot?
    a) plot()
    b) lm()

c) xyplot()
d) anova()

127. Which of the following is used to view dataset in a spreadsheet-type format ?
a) Disp()
b) View()
c) Seq()
d) lm()

128. _____ function carries out a chi-square test.
a) chisq.test()
b) t.test()
c) prop.test()
d) fisher.test()

129. Which of the following adds marginal sums to an existing table?
a) par()
b) prop.table()
c) addmargins()
d) quantile()

130. Which of the following lists names of variables in a data.frame?
a) quantile()
b) names()
c) barchart()
d) par()

131. Which of the following is tool for chi-square distributions?
a) pchisq()
b) chisq()
c) pnorm
d) barchart()

132. Which of the following is tool for checking normality?
a) qqline()
b) qline()
c) anova()
d) lm()

133. Which of the following is lattice command for producing boxplots?
a) plot()
b) bwplot()
c) xyplot()
d) barlm()

134. Which of the following compute analysis of variance table for fitted model?
a) ecdf()
b) cum()
c) anova()
d) bwplot()

135. Which of the following is used to find variance of all values?
a) var()
b) sd()

c) mean()
d) anova()