

## TYCS Sem 6 USCS Data Science(USCS606) Question Bank

1. Which of the following is known as raw data?
  - a) tidy data
  - b) unprocessed data
  - c) flat data
  - d) **all of the mentioned**
  
2. 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
  
3. Which of the following function is used for determining missing values?
  - a) **any**
  - b) Nan
  - c) not
  - d) all
  
4. Data that summarize all observations in a category are called \_\_\_\_\_ data.
  - a) frequency
  - b) **summarized**
  - c) raw
  - d) none of the mentioned
  
5. 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 above**
  
6. 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**
  
7. Which of the following function is used for k-means clustering?
  - a) **k-means**
  - b) k-mean
  - c) heatmap
  - d) K-Map

8. 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. All of the above**

9. 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. None Of the above

10. 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**

11. Which of the following language is used in Data science?

- A. C
- B. C++
- C. R
- D. Ruby

12. 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. None Of the above

13. Which of the following is correct skills for a Data Scientist?

- A. Probability & Statistics
- B. Machine Learning / Deep Learning
- C. Data Wrangling
- D. All of the above**

14. Which of the following are the Data Sources in data science?

- A. Structured
- B. UnStructured
- C. Both A and B**
- D. None Of the above

15. 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**
16. What are the Authentication in AWS?
- A. User Name/Password
  - B. Access Key
  - C. Access Key/ Session Token
  - D. **All of the above**
17. Which of the following is the most important language for Data Science?
- a) Java
  - b) Ruby
  - c) **R**
  - d) None of the mentioned
18. 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) All of the mentioned
19. 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
20. Height is example of \_\_\_\_\_.
- a) Continuous Data
  - b) Discrete Data
  - c) **Ordinal data**
  - d) Quantative Data
21. MongoDB support cross platform and is written in \_\_\_\_\_ language.
- a) C++
  - b) R
  - c) Java
  - d) Python
22. 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) **All of the above**

23. Data visualization is also an element of the broader \_\_\_\_\_.

- a) A. deliver presentation architecture
- b) **B. data presentation architecture**
- c) C. dataset presentation architecture
- d) D. data process architecture

24. 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. None Of the above

25. 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**

26. Which of the following language is used in Data science?

- A. **C**
- B. C++
- C. R
- D. Ruby

27. 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. None Of the above

28. Which of the following is correct skills for a Data Scientist?

A. Probability & Statistics

B. Machine Learning / Deep Learning

C. Data Wrangling

**D. All of the above**

29. Which of the following is not a part of data science process?

A. Discovery

B. Model Planning

**C. Communication Building**

D. Operationalize

30. Which of the following are the Data Sources in data science?

A. Structured

B. UnStructured

**C. Both A and B**

D. None Of the above

31. Which of the following is the most important language for Data Science?

a) Java

b) Ruby

c) **R**

d) .net

32. 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

33. ----- is defined as raw facts and figures collected together and stored in database

**a) Data**

- b) Knowledge
- c) File
- d) Picture

34. -----is represented using numbers or anything through which someone can measure various dimensions.
- a) Quantitative data**
  - b) Qualitative data
  - c) Nominal data
  - d) Binomial data
35. -----provides the characteristics and descriptors which cannot be easily measured.
- a) Quantitative data
  - b) Qualitative data**
  - c) Nominal data
  - d) Binomial data
36. -----means two values which are similar to binary data it can be true or false, yes or no etc.
- a) Quantitative data
  - b) Qualitative data
  - c) Nominal data
  - d) Binomial data**
37. ----- is also known as unordered data here every individual element will not have a kind of ranking but it will have some categories.
- a) Quantitative data
  - b) Qualitative data
  - c) Nominal data**
  - d) Binomial data
38. Ordinal data also known as order data here every element have some kind of order.
- a) Quantitative data
  - b) Qualitative data
  - c) Nominal data
  - d) Ordinal data**
39. ----- languages provides a set of grammar to a computer program or computing devices.
- a) Programming**
  - b) English
  - c) C++
  - d) Python
40. ----- language also known as binary or low level language consist of 0 and 1.
- a) Programming
  - b) Machine level**
  - c) Assembly
  - d) High level
41. ----- language is an intermediate language in which the instructions are written in the Assembly language.
- a) Programming

- b) Machine level
  - c) Assembly**
  - d) High level
42. ----- language keys simplest language and it is completely human readable.
- a) Programming
  - b) Machine level
  - c) Assembly
  - d) High level**
43. ----- formula type processing are related to mathematical and statistical problem solving techniques.
- a) Algebraic**
  - b) Geometric
  - c) Simple
  - d) trigonometry
44. -----data processing are used for processing the business data.
- a) Business**
  - b) Quantitative data
  - c) Qualitative data
  - d) Nominal data
45. -----is a package consisting various softwares or tools which are necessary to build a complete application .
- a) Integrated development environment**
  - b) Explanatory development Analysis
  - c) RDA
  - d) KNN
46. ---- is an approach used for Analysing datasets in order to summarise there important characteristics.
- a) EDA**
  - b) FDA
  - c) PDA
  - d) RDA
47. -----is representation which provides the distribution of numerical data.
- a) Box plot
  - b) Pie chart
  - c) Histogram**
  - d) Bar chart
48. ----- is graphical representation method use to depict groups of numerical data through there quartiles.
- a) Box plot**
  - b) Pie chart
  - c) Histogram
  - d) Bar chart
49. -----is a type of plot used to display values for mainly two variables of data set.
- a) Box plot
  - b) Pie chart
  - c) Scatter plot**
  - d) Bar chart

50. -----deals acquisition and preparation of data for various purposes in the organisation.
- Data management**
  - Data Collection
  - Data Extraction
  - Data Curation
51. ----- is a process in which data is gathered related to particular Activity.
- Data management
  - Data Collection**
  - Data Extraction
  - Data Curation
52. ----- interview is based on carefully worked interview schedule
- Structured**
  - Semi structured
  - Unstructured
  - Non structured
53. ----- type of interview the focus is to ask certain questions which provides and depth answer from the person who is being interviewed.
- Structured
  - Semi structured**
  - Unstructured
  - nonstructured
54. ----- interview or informal interview as there are no certain on time and type of question
- Structured
  - Semi structured
  - Unstructured**
  - nonstructured
55. -----interview is an alternate form of interview to the personal and face to face interview.
- Structured
  - Semi structured
  - Unstructured
  - Telephonic**
56. ----- it is also known as data cleansing and it is a process of detecting and correcting the errors or incorrect invalid entries from the data set.
- Data cleaning**
  - Data analysis
  - Data Collection
  - Data Correction
57. ----- is the process of evaluation in which someone can gain insight of organisation.
- Data cleaning
  - Data analysis**
  - Data Collection
  - Data Correction
58. -----define as the process of collecting data from various sources and integrating it into various repositories that are many more times more valuable than independent parts
- Data cleaning
  - Data analysis



- c) Data Collection
- d) Data curation**

59. preserving means collecting of data and then store and manage it.

- a) Preserving**
- b) Sharing
- c) curation
- d) forwarding

60. sharing deals with making data available according to need of authenticated user.

- a) Preserving
- b) Sharing**
- c) curation
- d) forwarding

61. ----- languages are special type of languages used for retrieving required data for information from the database.

- a) Query**
- b) Machine level
- c) Assembly
- d) High level

62. ----- query languages are used when it is known that what is to be retrieved and how it is to be retrieved from the database.

- a) Machine level
- b) Assembly
- c) non procedural**
- d) Procedural**

63. -----query languages are used when it is only known that what to retrieve but there is no idea about how to retrieve it.

- a) Machine level
- b) Assembly
- c) non procedural**
- d) Procedural

64. ---- Is used to select particular tuples from the database satisfying a given condition.

- a)  $\sigma$**
- b)  $\pi$
- c)  $\alpha$
- d)  $\beta$

65. --- operator is used to project the columns that satisfy the given condition.

- a)  $\sigma$
- b)  $\pi$**
- c)  $\alpha$
- d)  $\beta$

66. ----- operator is used to perform binary Union between two relations or tables

- a)  $\sigma$**
- b)  $\pi$

- c)  $\alpha$
- d) U

67. ----- Is used for removing the common element which are present in one relation And absent another.

- a) -
- b) B
- c)  $\alpha$
- d)  $\pi$

68. -----is used to combine two different relation or tables and makes it is a single relation.

- a) **Cartesian product**
- b) Set difference
- c) Union
- d) Project

69. -----is process for confirming the identity of the user.

- a) **Authentication**
- b) Authorization
- c) Collection
- d) auto correction

70. ----- is platform for code hosting and collaborative development for system

- a) **GitHUb**
- b) Amazon
- c) Pauer BI
- d) R tool

71. -----is type of database used to store networked data where nodes and links between nodes are done.

- a) **Graph Store**
- b) Wide column Store
- c) key value Store
- d) Amazon Web Services

72. -----Amazon cloud web hosting platform that are flexible, reliable scalable easy to use and cost effective solutions.

- a) Graph Store
- b) Wide column Store
- c) key value Store
- d) **Amazon Web Services**

73. ----- is an internet-based computing environment which consists of a large group of remote servers connected together and a centralised system is developed.

- a) **Cloud environment**
- b) Wide column Store
- c) key value Store
- d) Amazon Web Services

74. ----- cloud is the combination of both private and public cloud.

- a) Public
- b) Private

- c) **Hybrid**
  - d) hierarchical
75. IaaS stands for -----
- a) Indian Audit and Accounts Service
  - b) **infrastructure as a service**
  - c) Infrastructure as a System
  - d) Instruction as a service
76. PaaS stands for -----
- a) **platform as a service**
  - b) programming as a service
  - c) platform as a system
  - d) programming as a system
77. SaaS stands for.
- a) **software as a service**
  - b) Software as a system
  - c) Service as a system
  - d) Service as a service
78. ----- provides persistent block storage volumes for youth in Amazon EC2 instances in the cloud.
- a) Graph Store
  - b) Wide column Store
  - c) key value Store
  - d) **Elastic block Store**
79. S3 stands for
- a) **simple storage service.**
  - b) Simple storage system
  - c) storage service system
  - d) system storage service
80. EC2 stands for
- a) **elastic cloud.**
  - b) Electric cloud
  - c) Elastic circuit
  - d) Electric circuit
81. ----- deals with the perfection in the solution from the given data set how will your algorithm provides and accurate prediction is considered here.
- a) **Accuracy**
  - b) Atomicity
  - c) Atomic
  - d) Correctness
82. ----- is used as a solution to get rid out of the overfitting problem multivariate regression.
- a) **Regularization**
  - b) Authentication
  - c) Authorization
  - d) Collection
83. AIC stands for
- a) **Akaike information criterion**
  - b) Alpha information criteria
  - c) Akaike information count
  - d) Alpha information count

84. BIC stands for
- a) **bayesian information criterion**
  - b) bais information criterion
  - c) bais internet criterion
  - d) Bayesian internet criterion
85. -----is dimension reduction tool.
- a) BCA
  - b) MCA
  - c) **PCA**
  - d) DCA
86. ----- is a period over which data points for a particular resource or set of resources is collected.
- a) Dimension reduction
  - b) Feature selection
  - c) feature extraction
  - d) **Granularity**
87. -----provide time duration which determines how often a particular data has been sampled.
- a) Granularity
  - b) **Polling period**
  - c) Interval
  - d) Least square
88. -----data values are plotted and there in the function is evaluated which will fitt the maximum values and that function.
- a) **Integration**
  - b) Least square
  - c) regression tree
  - d) Time series analysis
89. In ----- the base learner are generated in the sequential manner.
- a. **SEM**
  - b. PEM
  - c. SVM
  - d. KNN
90. In ----- the base learners are generated in the parallel manner.
- a. SEM
  - b. **PEM**
  - c. SVM
  - d. KNN
91. SVM stands for -----
- a. support vector money
  - b. **support vector machine**
  - c. support voice machine
  - d. system vector machine
92. KNN stands for
- a. **K nearest neighbour**

- b. K nearest neighbor
  - c. K near neighbour
  - d. K non neighbour
93. ----- methods first derive a discretization of the space of the observations, obtaining a grid structure consisting of cells.
- a. Partition
  - b. Hierarchical
  - c. density-based
  - d. Grid
94. ----- is used to understand the structure of web using different links on different pages.
- a. Web
  - b. Content
  - c. Structure
  - d. usage
95. The analysis of cross-efficiency is based on the definition of the efficiency matrix,
- a. efficient frontier
  - b. efficiency matrix
  - c. effective frontier
  - d. effective matrix
96. The ----- of a DMU are defined as the product of the inputs used by the unit and the corresponding optimal weights.
- a. virtual inputs
  - b. virtual outputs
  - c. parameter
  - d. reference parameter
97. ----- are given by the product of the outputs of the unit and the associated optimal weights.
- a. virtual inputs
  - b. virtual outputs
  - c. parameter
  - d. reference parameter
98. ----- can be define as general mental ability for reasoning, problem solving and learning.
- a. Intelligence
  - b. Knowledge
  - c. Information
  - d. Data
99. The major purpose of an integrated logistic process is to minimize a function expressing the total -----
- a. Time
  - b. Cost
  - c. Size
  - d. People
100. The term ----- refers to the possibility that a portion of the demand due in a given period may be satisfied in a subsequent period, incurring an additional penalty cost.
- a. Log

- b. Backlog
- c. Front log
- d. None of the above