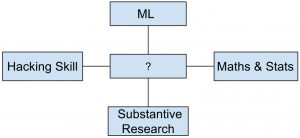
1. Which of the following would be more appropriate to be replaced with question mark in the following figure?



a) Data Analysis

b) Data Science

c) Descriptive Analytics

d) None of the mentioned

View Answer

Answer: b

Explanation: Data Science is a multidisciplinary which involves extraction of knowledge from large volumes of data that are structured or unstructured.

2. Point out the correct statement.

a) Raw data is original source of data

b) Preprocessed data is original source of data

c) Raw data is the data obtained after processing steps

d) None of the mentioned

View Answer

Answer: a

Explanation: Accounting programs are prototypical examples of data processing applications.

3. Which of the following is performed by Data Scientist?

a) Define the question

b) Create reproducible code

c) Challenge results

d) All of the mentioned

View Answer

Answer: d

Explanation: A data scientist is a job title for an employee or business intelligence (BI) consultant who excels at analyzing data, particularly large amounts of data.

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

a) Java

b) Ruby

c) R

d) None of the mentioned

View Answer

Answer: c

Explanation: R is free software for statistical computing and analysis.

5. 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

View Answer

Answer: b

Explanation: Data formatting is the organization of information according to preset specifications.

6. Which of the following approach should be used to ask Data Analysis question?

a) Find only one solution for particular problem

b) Find out the question which is to be answered

c) Find out answer from dataset without asking question

d) None of the mentioned

View Answer

Answer: b

Explanation: Data analysis has multiple facets and approaches.

7. Which of the following is one of the key data science skills?

a) Statistics

b) Machine Learning

c) Data Visualization

d) All of the mentioned

View Answer

Answer: d

Explanation: Data visualization is the presentation of data in a pictorial or graphical format.

8. Which of the following is a key characteristic of a hacker?

a) Afraid to say they don’t know the answer

b) Willing to find answers on their own

c) Not Willing to find answers on their own

d) All of the mentioned

View Answer

Answer: b

Explanation: Hacker is an expert at programming and solving problems with a computer.

9. 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

View Answer

Answer: b

Explanation: Processing includes merging, summarizing and subsetting data.

10. Raw data should be processed only one time.

a) True

b) False

View Answer

Answer: b

Explanation: Raw data may only need to be processed once.

11. Which of the following term is appropriate to the below figure?



a) Large Data

b) Big Data

c) Dark Data

d) None of the mentioned

View Answer

Answer: b

Explanation: Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.

12. Point out the correct statement.

a) Machine learning focuses on prediction, based on known properties learned from the training data

b) Data Cleaning focuses on prediction, based on known properties learned from the training data

c) Representing data in a form which both mere mortals can understand and get valuable insights is as much a science as much as it is art

d) None of the mentioned

View Answer

Answer: d

Explanation: Visualization is becoming a very important aspect.

13. Which of the following characteristic of big data is relatively more concerned to data science?

a) Velocity

b) Variety

c) Volume

d) None of the mentioned

View Answer

Answer: b

Explanation: Big data enables organizations to store, manage, and manipulate vast amounts of disparate data at the right speed and at the right time.

14. Which of the following analytical capabilities are provided by information management company?

a) Stream Computing

b) Content Management

c) Information Integration

d) All of the mentioned

View Answer

Answer: d

Explanation: With stream computing, store less, analyze more and make better decisions faster.

15. Point out the wrong statement.

a) The big volume indeed represents Big Data

b) The data growth and social media explosion have changed how we look at the data

c) Big Data is just about lots of data

d) All of the mentioned

View Answer

Answer: c

Explanation: Big Data is actually a concept providing an opportunity to find new insight into your existing data as well guidelines to capture and analysis your future data.

16. 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

View Answer

Answer: a

Explanation: Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

18. Which of the following focuses on the discovery of (previously) unknown properties on the data?

a) Data mining

b) Big Data

c) Data wrangling

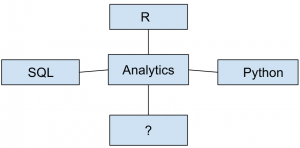
d) Machine Learning

View Answer

Answer: a

Explanation: Data munging or data wrangling is loosely the process of manually converting or mapping data from one “raw” form into another format that allows for more convenient consumption of the data with the help of semi-automated tools.

19. Which of the following language should be replaced with the question mark in the below figure?



a) Java

b) PHP

c) COBOL

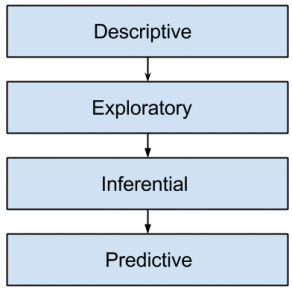
d) None of the mentioned

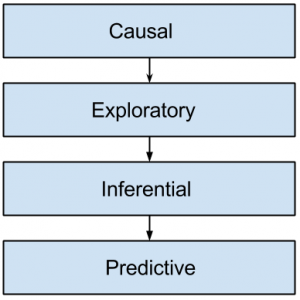
View Answer

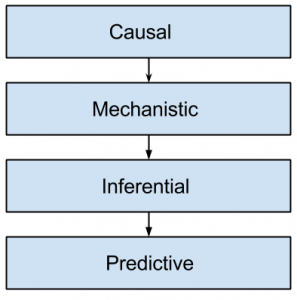
Answer: a

Explanation: Java is used for processing data in Big data Analytics.

19. Which of the following figure correctly shows approximate order of difficulty?

a) 

b) 

c) 

d) All of the mentioned

View Answer

Answer: a

Explanation: Predictive analysis is the practice of extracting information from existing data sets.

20. Point out the correct combination with regards to kind keyword for graph plotting.

a) ‘hist’ for histogram

b) ‘box’ for boxplot

c) ‘area’ for area plots

d) all of the mentioned

View Answer

Answer: d

Explanation: The kind keyword argument of plot() accepts a handful of values for plots other than the default Line plot.

21. Which of the following value is provided by kind keyword for barplot?

a) bar

b) kde

c) hexbin

d) none of the mentioned

View Answer

Answer: a

Explanation: bar can also be used for barplot.

22. You can create a scatter plot matrix using the \_\_\_\_\_\_\_\_\_\_ method in pandas.tools.plotting.

a) sca\_matrix

b) scatter\_matrix

c) DataFrame.plot

d) all of the mentioned

View Answer

Answer: b

Explanation: You can create density plots using the Series/DataFrame.plot.

23. Point out the wrong combination with regards to the kind keyword for graph plotting.

a) ‘scatter’ for scatter plots

b) ‘kde’ for hexagonal bin plots

c) ‘pie’ for pie plots

d) none of the mentioned

View Answer

Answer: b

Explanation: kde is used for density plots.

24. Which of the following plots are often used for checking randomness in time series?

a) Auto Causation

b) Autorank

c) Autocorrelation

d) None of the mentioned

View Answer

Answer: c

Explanation: If the time series is random, such autocorrelations should be near zero for any and all time-lag separations.

25. Which of the following thing can be data in Pandas?

a) a python dict

b) an ndarray

c) a scalar value

d) all of the mentioned

View Answer

Answer: d

Explanation: The passed index is a list of axis labels.

26. Point out the correct statement.

a) If data is a list, if index is passed the values in data corresponding to the labels in the index will be pulled out

b) NaN is the standard missing data marker used in pandas

c) Series acts very similarly to a array

d) None of the mentioned

View Answer

Answer: b

Explanation: If data is a dict, if index is passed the values in data corresponding to the labels in the index will be pulled out.

27. The result of an operation between unaligned Series will have the \_\_\_\_\_\_\_\_ of the indexes involved.

a) intersection

b) union

c) total

d) all of the mentioned

View Answer

Answer: b

Explanation: If a label is not found in one Series or the other, the result will be marked as missing NaN.

28. Which of the following input can be accepted by DataFrame?

a) Structured ndarray

b) Series

c) DataFrame

d) All of the mentioned

View Answer

Answer: d

Explanation: DataFrame is a 2-dimensional labeled data structure with columns of potentially different types.

29. Point out the wrong statement.

a) A DataFrame is like a fixed-size dict in that you can get and set values by index label

b) Series can be be passed into most NumPy methods expecting an ndarray

c) A key difference between Series and ndarray is that operations between Series automatically align the data based on label

d) None of the mentioned

View Answer

Answer: a

Explanation: A Series is like a fixed-size dict in that you can get and set values by index label.

30. Which of the following takes a dict of dicts or a dict of array-like sequences and returns a DataFrame?

a) DataFrame.from\_items

b) DataFrame.from\_records

c) DataFrame.from\_dict

d) All of the mentioned

View Answer

Answer: a

Explanation: DataFrame.from\_dict operates like the DataFrame constructor except for the orient parameter which is ‘columns’ by default.

31. Series is a one-dimensional labeled array capable of holding any data type.

a) True

b) False

View Answer

Answer: a

Explanation: The axis labels are collectively referred to as the index.

32. Which of the following works analogously to the form of the dict constructor?

a) DataFrame.from\_items

b) DataFrame.from\_records

c) DataFrame.from\_dict

d) All of the mentioned

View Answer

Answer: a

Explanation: DataFrame.from\_records takes a list of tuples or an ndarray with structured dtype.

33. Which of the following operation works with the same syntax as the analogous dict operations?

a) Getting columns

b) Setting columns

c) Deleting columns

d) All of the mentioned

View Answer

Answer: d

Explanation: You can treat a DataFrame semantically like a dict of like-indexed Series objects.

34. If data is an ndarray, index must be the same length as data.

a) True

b) False

View Answer

Answer: a

Explanation: If no index is passed, one will be created having values [0, …, len(data) – 1].

1. Which of the following is the base layer for all of the sparse indexed data structures?

a) SArray

b) SparseArray

c) PyArray

d) None of the mentioned

View Answer

Answer: b

Explanation: SparseArray is a 1-dimensional ndarray-like object storing only values distinct from the fill\_value.

35. Point out the correct statement.

a) All of the standard pandas data structures have a to\_sparse method

b) Any sparse object can be converted back to the standard dense form by calling to\_dense

c) The sparse objects exist for memory efficiency reasons

d) All of the mentioned

View Answer

Answer: d

Explanation: The to\_sparse method takes a kind argument and a fill\_value.

36. Which of the following is not an indexed object?

a) SparseSeries

b) SparseDataFrame

c) SparsePanel

d) None of the mentioned

View Answer

Answer: d

Explanation: SparseArray can be converted back to a regular ndarray by calling to\_dense.

37. Which of the following list-like data structure is used for managing a dynamic collection of SparseArrays?

a) SparseList

b) GeoList

c) SparseSeries

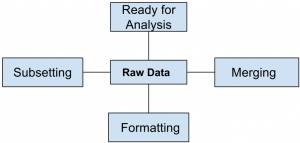
d) All of the mentioned

View Answer

Answer: a

Explanation: To create one, simply call the SparseList constructor with a fill\_value.

38. Which of the following block information is odd man out?



a) Subsetting

b) Raw data

c) Ready for analysis

d) None of the mentioned

View Answer

Answer: b

Explanation: Characteristics mentioned in the diagram are traits of processed data.

39. Point out the correct statement.

a) Data has only qualitative value

b) Data has only quantitative value

c) Data has both qualitative and quantitative value

d) None of the mentioned

View Answer

40. Data that summarize all observations in a category are called \_\_\_\_\_\_\_\_\_\_ data.

a) frequency

b) summarized

c) raw

d) none of the mentioned

View Answer

Answer: b

Explanation: The summary could be the sum of the observations, the number of occurrences, their mean value, and so on.

41. Which of the following is an example of raw data?

a) original swath files generated from a sonar system

b) initial time-series file of temperature values

c) a real-time GPS-encoded navigation file

d) all of the mentioned

View Answer

Answer: d

Explanation: Raw data refers to data that have not been changed since acquisition.

42. Which of the following is the probability calculus of beliefs, given that beliefs follow certain rules?

a) Bayesian probability

b) Frequency probability

c) Frequency inference

d) Bayesian inference

View Answer

Answer: a

Explanation: Data scientists tend to fall within shades of gray of these and various other schools of inference.

43. Point out the correct statement.

a) Bayesian inference is the use of Bayesian probability representation of beliefs to perform inference

b) NULL is the standard missing data marker used in S

c) Frequency inference is the use of Bayesian probability representation of beliefs to perform inference

d) None of the mentioned

View Answer

Answer: a

Explanation: Frequency probability is the long run proportion of times an event occurs in independent, identically distributed repetitions.

44. Which of the following can be considered as random variable?

a) The outcome from the roll of a die

b) The outcome of flip of a coin

c) The outcome of exam

d) All of the mentioned

View Answer

Answer: d

Explanation: The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values.

45. Which of the following random variable that take on only a countable number of possibilities?

a) Discrete

b) Non Discrete

c) Continuous

d) All of the mentioned

View Answer

Answer: a

Explanation: Continuous random variable can take any value on some subset of the real line.

This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on “Probability and Statistics”.

46. The expected value or \_\_\_\_\_\_\_ of a random variable is the center of its distribution.

a) mode

b) median

c) mean

d) bayesian inference

View Answer

Answer: c

Explanation: A probability model connects the data to the population using assumptions.

47. Point out the correct statement.

a) Some cumulative distribution function F is non-decreasing and right-continuous

b) Every cumulative distribution function F is decreasing and right-continuous

c) Every cumulative distribution function F is increasing and left-continuous

d) None of the mentioned

View Answer

Answer: d

Explanation: Every cumulative distribution function F is non-decreasing and right-continuous.

48. Which of the following of a random variable is a measure of spread?

a) variance

b) standard deviation

c) empirical mean

d) all of the mentioned

View Answer

Answer: a

Explanation: Densities with a higher variance are more spread out than densities with a lower variance.

49. The square root of the variance is called the \_\_\_\_\_\_\_\_ deviation.

a) empirical

b) mean

c) continuous

d) standard

View Answer

Answer: d

Explanation: Standard Deviation (SD) is the measure of spread of the numbers in a set of data from its mean value.

50. Which of the following random variables are the default model for random samples?

a) iid

b) id

c) pmd

d) all of the mentioned

View Answer

Answer: a

Explanation: Random variables are said to be iid if they are independent and identically distributed.