1. *Which one is NOT from Phase 1 of Data Science Life Cycle*
    1. Learning the target domain
    2. Developing initial hypothesis
    3. Visualize initial hypothesis
    4. Identifying key variables

2. Which of the following is the most important language for Data Science?
    1. Ruby
    2. R
    3. Java
    4. None

3. A collection of information about a related topic is referred to as a_____
    1. Visualisation
    2. Analysis
    3. Conclusion
    4. Data

4. To find the _____ you add up all the numbers and then divide by how many numbers you have.
    1. Median
    2. Mean
    3. Mode
    4. Range

5. Which of the following is performed by Data Scientist ?
    1. Create reproducible code
    2. Challenge results
    3. Define the question
    4. All of the above

6. Which is not a tool for Statistical Data Analysis?
    1. Logistic Regression
    2. Linear & Non-linear Regression
    3. Histogram
    4. ANOVA

7. What is the mean of test scores?{70, 70, 80, 85, 85, 90, 95, 95, 100, 100}
    1. 85, 95, and 100
    2. 30
    3. 87
    4. None

8. Choose the correct keyword for this definition: A graphical representation of a data set
    1. Data Set
    2. Investigative Cycle
    3. Visualisation

4. None

9. To find the _____ you put all numbers in order from least to greatest and find the number that is in the middle.

1. Median
2. Mode
3. Mean
4. Range

10. R is an interpreted language so it can access through _____?

1. Command line interpreter
2. Disk operating system
3. Operating system
4. User interface operating system

11. Data has been collected on visitors' viewing habits at a bank's website. Which technique is used to identify pages commonly viewed during the same visit to the website?

1. Clustering
2. Classification
3. Association Rules
4. Regression

12. A relationship between two or more variables is referred to as a _____

1. Trend
2. Spike
3. All of above
4. None of above

13. A graphical representation of a data set is referred to as a _____

1. Visualization
2. Data Set
3. Investigative Cycle
4. None

14. Which of the following step is performed by data scientist AFTER acquiring the data?

1. Data Integration
2. Data Replication
3. Data Cleansing
4. All of the above

15. Data that sits outside the trend is referred to as a _____

1. Outlier
2. Trend
3. Spike
4. Both 1 & 2

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

1. Find out the question which is to be answered
2. Find only one solution for particular problem
3. Find out answer from dataset without asking question
4. None

17. Which of the following is NOT a machine learning algorithm?

1. SVG
2. Random Forest
3. SVM
4. None

18. What is Big Data?

1. Data with the word 'big' in it
2. Data about people who are big
3. Data with a large size
4. Data made with a big purpose

19. What is R an implementation of?

1. Logical Scoping
2. S Programming Language
3. Lexical Scoping
4. Q Programming Language

20. The 5 steps required to identify a problem and come up with a solution are referred to as the _____ Cycle

1. Visualization
2. Investigative
3. Conclusion
4. None

21. Which of the following is characteristic of Processed Data?

1. Hard to use for data analysis
2. Data is not ready for analysis
3. All steps should be noted
4. None of the above

\

22. Which was not mentioned as a latest trend tool_____

1. Excel
2. Pentaho
3. SPSS
4. Notepad

23. Which of the following is one of the key data science skill ?

1. Machine Learning
2. Statistics
3. Data Visualization

4. All of the above

24. Which of the following is not a stage in the Investigative Cycle?

1. Investigate
2. Analysis
3. Conclusion
4. None

25. Vectors come in two parts_____ and _____

1. Atomic vectors and list
2. Atomic vectors and array
3. Atomic vectors and matrix
4. None

26.Choose the correct keyword for this definition: A collection of information about a related topic

1. Trend
2. Spike
3. Data Set
4. None

27. The process of evaluating data through analytical and statistical tools.

1. Data Mining
2. Data Exploration
3. Data Analysis
4. Data Visualization

28. Which of the following is key characteristic of hacker ?

1. Willing to find answers on their own
2. Afraid to say they don't know the answer
3. Not Willing to find answers on their own
4. All of the mentioned

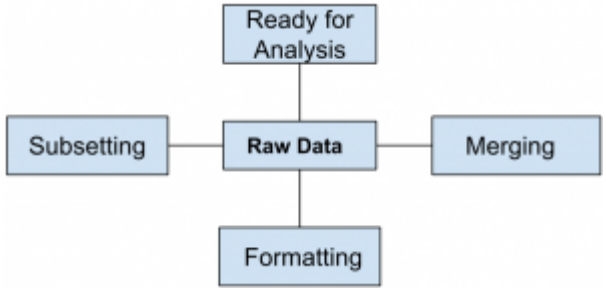
29. Which of the following characteristic of big data is relatively more concerned to data science ?

1. Variety
2. Volume
3. Velocity
4. None

30. R is an_____ programming language?

1. GPL
2. Open source
3. Closed source
4. Definite source

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



a) Subsetting
b) Raw data
c) Ready for analysis
d) None of the mentioned

32. 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

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

34. 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

35. 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

36. Which of the following data is put into a formula to produce commonly accepted results?
a) Raw
b) Processed
c) Synchronized
d) All of the Mentioned

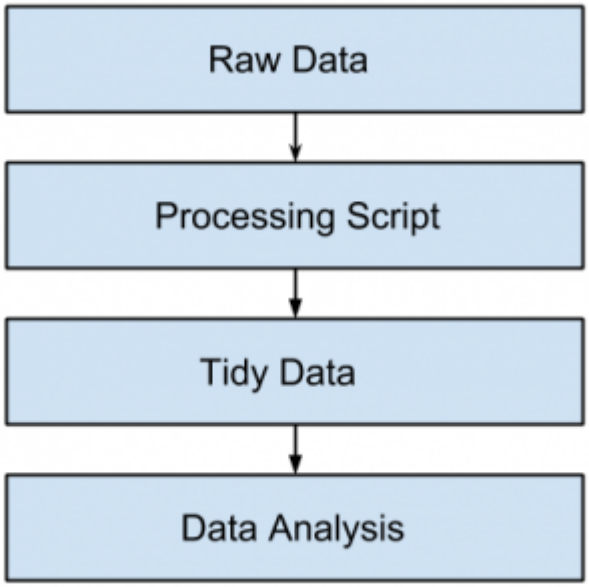
37. Processing data includes subsetting, formatting and merging only.
a) True
b) False

38. Which of the following is another name for raw data?
a) destination data
b) eggy data
c) secondary

d) machine learning

39. Which type of data is generated by POS terminal in a busy supermarket each day?
a) Source
b) Processed
c) Synchronized
d) All of the mentioned

40. Following figure represents correct sequence of steps in performing data analysis.



a) True
b) False

41. Which of the following function is good for the automatic splitting of names?
a) split
b) strsplit
c) autsplit
d) none of the mentioned

42. Point out the correct statement.
a) gsub is used for fixing character vectors
b) sub is used for finding values like grep
c) grep is used for fixing character vectors
d) none of the mentioned

43. Which of the following function is used for fixing character vectors?
a) tolower
b) toUPPER
c) toLOWER
d) all of the mentioned

44. Which of the following metacharacter is used to refer to any character?
a) %
b) @
c) .
d) All of the mentioned

45. Point out the wrong statement.
a) Variables with character values should be made less descriptive
b) Variables with character values should usually be made into factor variable
c) Common variables are used to apply transforms
d) All of the mentioned

46. Which of the following is used for specifying character class with metacharacter?
a) []
b) {}
c) /+
d) All of the mentioned

47. Regular expressions can be thought of as a combination of literals and metacharacters.
a) True
b) False

48. Which of the following signs are used to indicate repetition?
a) #
b) *
c) –
d) All of the mentioned

49. Which of the following function is used for searching text strings by means of regular expression?
a) grepd
b) grepl
c) gepexpr
d) all of the mentioned

50. merge function is used for merging data frames.
a) True
b) False

51. The expected value or _____ of a random variable is the center of its distribution.
a) mode
b) median
c) mean
d) bayesian inference

52. 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

53. 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

54. The square root of the variance is called the _____ deviation.
a) empirical

b) mean
c) continuous
d) standard

55. Point out the wrong statement.
a) A percentile is simply a quantile with expressed as a percent
b) There are two types of random variable
c) R cannot approximate quantiles for you for common distributions
d) None of the mentioned

56. Which of the following inequality is useful for interpreting variances?
a) Chebyshev
b) Stautaory
c) Testory
d) All of the mentioned

57. For continuous random variables, the CDF is the derivative of the PDF.
a) True
b) False

58. Chebyshev's inequality states that the probability of a "Six Sigma" event is less than _____
a) 10%
b) 20%
c) 30%
d) 3%

59. Which of the following random variables are the default model for random samples?
a) iid
b) id
c) pmd
d) all of the mentioned

60. Cumulative distribution functions are used to specify the distribution of multivariate random variables.

a) True
b) False

61. Predicting with trees evaluate _____ within each group of data.

a) equality
b) homogeneity
c) heterogeneity
d) all of the mentioned

62. Point out the wrong statement.

a) Training and testing data must be processed in different way
b) Test transformation would mostly be imperfect
c) The first goal is statistical and second is data compression in PCA
d) All of the mentioned

63. Which of the following method options is provided by train function for bagging?

a) bagEarth
b) treebag
c) bagFDA
d) all of the mentioned

64. Which of the following is correct with respect to random forest?

a) Random forest are difficult to interpret but often very accurate
b) Random forest are easy to interpret but often very accurate
c) Random forest are difficult to interpret but very less accurate
d) None of the mentioned

65. Point out the correct statement.

a) Prediction with regression is easy to implement
b) Prediction with regression is easy to interpret
c) Prediction with regression performs well when linear model is correct
d) All of the mentioned

66. Which of the following library is used for boosting generalized additive models?

a) gamBoost
b) gbm
c) ada
d) all of the mentioned

67. The principal components are equal to left singular values if you first scale the variables.

a) True
b) False

68. Which of the following is statistical boosting based on additive logistic regression?

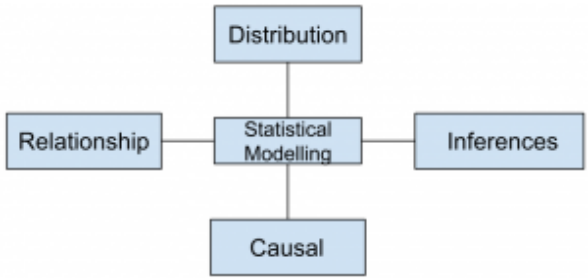
a) gamBoost
b) gbm
c) ada
d) mboost

69. Which of the following is one of the largest boost subclass in boosting?

a) variance boosting
b) gradient boosting
c) mean boosting
d) all of the mentioned

70. PCA is most useful for non linear type models.
a) True
b) False

71. Which of the following goal is incorrectly represented in the below figure?



a) Relationship between variables
b) Distribution of variables
c) Inference about relationships
d) Causal

72. Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log-normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

73. Which of the following is incorrect with respect to use of Poisson distribution?
a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

74. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

75. Point out the wrong statement.
a) The normal distribution is asymmetric and peaked about its mode
b) A constant times a normally distributed random variable is also normally distributed
c) Sample means of normally distributed random variables are again normally distributed
d) None of the mentioned

76. Which of the following form the basis for frequency interpretation of probabilities?
a) Asymptotics
b) Symptotics
c) Asymmetry
d) All of the mentioned

77. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

78. The _____ basically states that the sample mean is consistent.
a) LAN
b) LLN
c) LWN
d) None of the mentioned

79. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

80. The binomial random variables are obtained as the sum of iid Gaussian trials.
a) True
b) False

81. Which of the following is contained in NumPy library?
a) n-dimensional array object
b) tools for integrating C/C++ and Fortran code
c) fourier transform
d) all of the mentioned

82. Point out the wrong statement.
a) ipython is an enhanced interactive Python shell
b) matplotlib will enable you to plot graphics
c) rPy provides a lot of scientific routines that work on top of NumPy
d) all of the mentioned

83. The _____ function returns its argument with a modified shape, whereas the _____ method modifies the array itself.
a) reshape, resize
b) resize, reshape
c) reshape2, resize
d) all of the mentioned

84. To create sequences of numbers, NumPy provides a function _____ analogous to range that returns arrays instead of lists.
a) arange
b) aspace
c) aline
d) all of the mentioned

85. Point out the correct statement.
a) NumPy main object is the homogeneous multidimensional array
b) In Numpy, dimensions are called axes
c) Numpy array class is called ndarray
d) All of the mentioned

86. Which of the following function stacks 1D arrays as columns into a 2D array?
a) row_stack
b) column_stack
c) com_stack
d) all of the mentioned

87. ndarray is also known as the alias array.
a) True
b) False

88. Which of the following method creates a new array object that looks at the same data?
a) view
b) copy
c) paste
d) all of the mentioned

89. Which of the following function can be used to combine different vectors so as to obtain the result for each n-uplet?
a) iid_
b) ix_
c) ixd_
d) all of the mentioned

90. ndarray.dataitemSize is the buffer containing the actual elements of the array.
a) True
b) False