List of Data Analytics MCQs

1. Data Analytics uses ___ to get insights from data.

A. Statistical figures
B. Numerical aspects
C. Statistical methods
D. None of the mentioned above

Answer: C) Statistical methods

Explanation: To gain insights from data, Data Analytics use statistical approaches. Organizations can use data analytics to uncover trends and develop insights by analyzing all of their data (real-time, historical, unstructured, structured, and qualitative).

2. Amongst which of the following is / are the branch of statistics which deals with the development of statistical methods is classified as ___.

A. Industry statistics
B. Economic statistics
C. Applied statistics
D. None of the mentioned above

Answer: C) Applied statistics

Explanation: The discipline of statistics that works with the development of statistical procedures is known as applied statistics. Planning for data collecting, maintaining data, analyzing, interpreting, and drawing conclusions from data, and finding issues, solutions, and opportunities utilizing analysis are all part of applied statistics. In data analysis and empirical research, these major fosters critical thinking and problem-solving skills.

3. Linear Regression is the supervised machine learning model in which the model finds the best fit ___ between the independent and dependent variable.

A. Linear line
B. Nonlinear line
C. Curved line
D. All of the mentioned above

Answer: A) Linear line

Explanation: Linear Regression is a supervised Machine Learning model that identifies the best fit linear line between the independent and dependent variables, i.e., the linear connection between the dependent and independent variables.

4. Amongst which of the following is / are the types of Linear Regression,

A. Simple Linear Regression
B. Multiple Linear Regression
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation: There are two forms of linear regression: simple and multiple. Simple Linear Regression is used when there is only one independent variable and the model must determine the linear connection between it and the dependent variable. Multiple Linear Regression is employed more than one independent variable in the model to determine the link.

5. Amongst which of the following is / are the true about regression analysis?

A. Describes associations within the data
B. Modeling relationships within the data
C. Answering yes/no questions about the data
D. All of the mentioned above

Answer: B) Modeling relationships within the data

Explanation: Regression analysis is used to describe relationships within data, and so it is a collection of statistical methods for estimating relationships between a dependent variable and one or more independent variables. There are various types of regression analysis, including linear, multiple linear, and nonlinear. Simple linear and multiple linear models are the most frequent. Nonlinear regression analysis is typically employed for more difficult data sets with a nonlinear connection between the dependent and independent variables.

6. Linear regression analysis is used to predict the value of a variable based on the value of another variable.

A. True
B. False

Answer: A) True

Explanation: Linear regression analysis predicts the value of one variable depending on the value of another. The variable we wish to forecast is referred to as the dependent variable. The variable we are utilizing to predict the value of the other variable is referred to as the independent variable.

7. A Linear Regression model's main aim is to find the best fit linear line and the ___ of intercept and coefficients such that the error is minimized.

A. Optimal values
B. Linear line
C. Linear polynomial
D. None of the mentioned above

Answer: A) Optimal values

Explanation: The basic goal of a Linear Regression model is to determine the best fit linear line and the ideal intercept and coefficient values such that the error is minimized. A linear regression model describes the relationship between one or more independent variables, X, and a dependent variable, y. A multiple linear regression model is a type of regression model that has numerous lines of regression. A multiple linear regression model is $yi=\beta 0+\beta 1Xi1+\beta 2Xi2+\cdots+\beta pXip+\varepsilon i, \quad i=1,\cdots,n$

8. Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

A. True
B. False

Answer: A) True

Explanation: In statistics, the actual value is the value derived from observation or measurement of the available data. It is also known as the observed value. The expected value is the predicted value of the variable based on the regression analysis. Linear regression is most commonly used to calculate model error using mean-square error (MSE). MSE is derived by measuring the distance between the observed and anticipated y-values at each value of x and then computing the mean of the squared distances.

9. The process of quantifying data is referred to as ___.

A. Decoding
B. Structure
C. Enumeration
D. Coding

Answer: C) Enumeration

Explanation:

Enumeration is the term for the process of quantifying data. Any quantifiable information that can be used for mathematical calculations or statistical analysis is referred to as quantitative data. This type of information aids in the development of real-world decisions based on mathematical derivations. To answer inquiries like how many, quantitative data is used. How often do you do it? How much is it? This information can be confirmed and validated.

10. Text Analytics, also referred to as Text Mining?

A. True
B. False

Answer: A) True

Explanation: Text analytics uses a combination of machine learning, statistical, and linguistic tools to analyze vast amounts of unstructured material (text that does not have a preset format) in order to draw insights and trends. It enables corporations, governments, researchers, and the media to make critical decisions based on the vast amounts of data available to them.

11. ___ are used when we want to visually examine the relationship between two quantitative variables.

A. Bar graph
B. Scatterplot
C. Line graph
D. Pie chart

Answer: A) Bar graph

Explanation: Dots are used to indicate values for two different numeric variables in a scatter plot, also known as a scatter chart or a scatter graph. The values for each data point are indicated by the position of each dot on the horizontal and vertical axes. Scatter plots are used to see how variables relate to one another.

12. A graph that uses vertical bars to represent data is called a ____.

A. Bar graph
B. Line graph
C. Scatterplot
D. All of the mentioned above

Answer: A) Bar graph

Explanation:A bar graph is a graph that employs vertical bars to represent data. Bar graphs are visual representations of data (usually grouped) in the shape of vertical or horizontal rectangular bars, with bar length proportional to data measure. Bar charts are another name for them. In statistics, bar graphs are one of the data management methods.

13. Data Analysis is a process of,

A. Inspecting data
B. Data Cleaning
C. Transforming of data
D. All of the mentioned above

Answer: D) All of the mentioned above
Explanation: The process of reviewing, cleansing, and manipulating data with the objective of identifying usable information, informing conclusions, and assisting decision-making is known as data analysis. Data analysis is important in today's business environment since it helps businesses make more scientific decisions and run more efficiently.

14. Least Square Method uses ___.

A. Linear polynomial
B. Linear regression
C. Linear sequence
D. None of the mentioned above

Answer: B) Linear regression
Explanation: Linear regression employs the Least Square Method. The least-squares approach is a type of mathematical regression analysis that determines the best fit line for a collection of data, displaying the relationship between the points visually. The relationship between a known independent variable and an unknown dependent variable is represented by each piece of data.

15. What is a hypothesis?

A. A statement that the researcher wants to test through the data collected in a study
B. A research questions the results will answer
C. A theory that underpins the study
D. A statistical method for calculating the extent to which the results could have happened by chance

Answer: A) A statement that the researcher wants to test through the data collected in a study.
Explanation: A hypothesis is a proposition that a researcher wishes to evaluate using data from a study. A hypothesis is a conclusion reached after considering evidence. This is the first step in any investigation, where the research questions are translated into a prediction. Variables, population, and the relationship between the variables are all included. A research hypothesis is a hypothesis that is tested to see if two or more variables have a relationship.

16. Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate ___.

A. Predictions
B. Interpretation
C. Conclusion
D. None of the mentioned above

Answer: A) Predictions
Explanation: Linear-regression models are straightforward and provide a basic mathematical method for generating predictions. Linear regression can be used in a variety of corporate and academic study.

17. Amongst which of the following is / are the applications of Linear Regression,

A. Biological
B. Behavioural
C. Social sciences
D. All of the mentioned above

Answer: D) All of the mentioned above
Explanation: Linear regression is utilized in a variety of fields, including biology, behavioural science, environmental research, and business. Linear regression models have proven to be a reliable and scientific means of forecasting the future. Because linear regression is a well-known statistical process, its properties are well understood and linear regression models may be trained quickly.

18. With reference to data, dependent and independent variables should be quantitative.

True
False

Answer: A) True
Explanation:
Dependent and independent variables should be quantitative when it comes to data. Both the dependent and independent variables should have a numerical value. Religious, major field of study and residential region categorical factors must be represented as binary variables or other sorts of contrast variables.

19. For each value of the ___, the distribution of the dependent variable must be normal.

A. Independent variable
B. Depended variable
C. Intermediate variable
D. None of the mentioned above

Answer: A) Independent variable

Explanation:

The dependent variable's distribution must be normal for each value of the independent variable. For all values of the independent variable, the variance of the dependent variable's distribution should be constant. The dependent variable should have a linear relationship with each independent variable, and all observations should be independent.

20. Residual plot helps in analyzing the model using the values of residues.

A. True
B. False

Answer: A) True

Explanation: The residue plot aids in the analysis of the model by displaying the values of the residues. It's shown as a line between the projected values and the residual. Their values are all the same. The point's distance from 0 indicates how inaccurate the prediction was for that number. If the value is positive, the probability of success is minimal. If the value is negative, the probability of success is high. A number of 0 implies that the forecast is perfect. The model can be improved by detecting residual patterns.

21. Amongst which of the following is / are not a major data analysis approach?

A. Predictive Intelligence
B. Business Intelligence
C. Text Analytics
D. Data Mining

Answer: A) Predictive Intelligence

Explanation: The practice of collecting data about consumers' and potential consumers' behaviour’s/actions from a number of sources and perhaps integrating it with profile data about their qualities is known as predictive intelligence.

22. By 2025, the volume of data will increase to,

A. TB
B. YB
C. ZB
D. EB

Answer: C) ZB

Explanation: It is projected that 2.5 quintillion bytes of data are created every day, with the volume of digital data expected to reach Zeta Byte by 2025.

23. Alternative Hypothesis is also called as?

A. Null Hypothesis
B. Research Hypothesis
C. Simple Hypothesis
D. None of the mentioned above

Answer: B) Research Hypothesis

Explanation: The alternative hypothesis is the assertion that is being tested against the null hypothesis. Ha or H1 are common abbreviations for alternative hypotheses. The alternative hypothesis is the hypothesis that is inferred from a null hypothesis that has been rejected. It is best stated as an explanation for why the null hypothesis was rejected. It is also known as the research hypothesis. Unlike the null hypothesis, the researcher is usually most interested in the alternative hypothesis.

24. If the null hypothesis is false then which of the following is accepted?

A. Alternative Hypothesis.
B. Null Hypothesis
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation: The alternative hypothesis is accepted if the null hypothesis is untrue. An alternative theory is a proposition that a researcher is testing in hypothesis testing. From the researcher's perspective, this assertion is correct, and it finally proves to reject the null hypothesis and replace it with a different one. The difference between two or more variables is anticipated in this hypothesis.

25. Amongst which of the following is / are not an example of social media?

A. Twitter
B. Instagram
C. Both A and B
D. None of the mentioned above

Answer: D) None of the mentioned above

Explanation: Social media is a type of computer-based technology that allows people to share their ideas, thoughts, and information with others via virtual networks and communities. Social media is an internet-based platform that allows people to share content such as personal information, documents, films, and images quickly and electronically.

26. Velocity is the speed at which the data is processed -

A. True
B. False

Answer: A) True

Explanation:

The rate at which data is generated, distributed, and gathered is referred to as data velocity. High data velocity is created at such a rapid rate that it necessitates the use of specialized processing techniques. The faster data can be captured and processed, the more valuable the data collected will be and the longer it will hold its worth.

27. ___ refers to the ability to turn your data useful for business.

A. Value
B. Variety
C. Velocity
D. None of the mentioned above

Answer: A) Value

Explanation:

The ability to turn our data into business value is referred to as value. The usefulness of obtained data for our business is referred to as data value. Data, regardless of its magnitude, is rarely useful on its own; to be useful, it must be transformed into insights or knowledge, which is where data processing comes in.

28. Correlation is the relationship between two variables -

A. One
B. Two
C. Zero
D. All of the mentioned above

Answer: B) Two

Explanation:

Correlation is the strength of a relationship between two variables, and the Pearson's correlation coefficient measures how strong that relationship is. The correlation of two variables is the statistical link between them. A positive correlation means that both variables move in the same direction, while a negative correlation means that when one variable's value rises, the other variable's value falls.

29. The Mean Squared Error is a measure of the average of the squares of the residuals.

A. True
B. False

Answer: A) True

Explanation:

The degree of inaccuracy in statistical models is measured by the mean squared error (MSE). The average squared difference between observed and expected values is calculated. The MSE equals zero when a model has no errors. Its value rises as the model inaccuracy rises. The mean squared deviation is another name for the mean squared deviation (MSD). The average squared residual is represented by the mean squared error in regression.

30. Logistic regression is used to find the probability of event = Success and event = ____.

A. Failure
B. Success
C. Both A and B
D. None of the mentioned above

Answer: A) Failure

Explanation:

The likelihood of event=Success and event=Failure is calculated using logistic regression. When the dependent variable is in nature, we should utilize logistic regression. For classification difficulties, logistic regression is commonly employed. There is no requirement for a linear relationship between the dependent and independent variables in logistic regression. Because it uses a non-linear log transformation on the anticipated odds ratio, it can handle a wide range of relationships.

31. A good data analytics solution includes a viable self-service ___.

A. Data mining
B. Data wrangling
C. Data warehouse
D. None of the mentioned above

Answer: B) Data wrangling

Explanation:

A smart data analytics solution incorporates self-service data wrangling and data preparation features so that data may be simply and quickly gathered from a range of incomplete, difficult, or messy data sources and cleansed for mashup and analysis.

32. To glean insights from the data, many analysts and data scientists rely on ___.

A. Data mining
B. Data visualization
C. Data warehouse
D. All of the mentioned above

Answer: B) Data visualization

Explanation:

Many analysts and data scientists use data visualization, or the graphical depiction of data, to assist individuals visually explores and finds patterns and outliers in the data in order to get insights. Data visualization features are included in a good data analytics system, making data exploration easier and faster.

33. Predictive analytics involves taking historical data -

A. True
B. False

Answer: A) True

Explanation:

The approach or practice of utilizing data to generate projections about the possibility of certain future events in your organization is known as predictive analytics, which is a form of advanced analytics. Predictive analytics models unknown future occurrences by combining historical and current data with advanced statistics and machine learning approaches. It is commonly characterized as utilizing data science and machine learning to learn from an organization's previous collective experience in order to make better decisions in the future.

34. With reference to Predictive analytics, it allows organizations to predict customer behavior -

A. True
B. False

Answer: A) True

Explanation:

Predictive analytics enables businesses to forecast consumer behavior and business results by combining historical and real-time data. Furthermore, predictive modeling is a subset of this activity that entails constructing and maintaining models, testing and iterating with existing data, and embedding models into applications.

35. Customer analytics refers -

A. Customer Relationship Management: churn analysis and prevention
B. Marketing: cross-sell, up-sell
C. Pricing: leakage monitoring, promotional effects tracking, competitive price responses
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

Customer analytics includes churn analysis and prevention, marketing: cross-sell and up-sell, and pricing: leakage monitoring, promotional effects tracking, and competitive price reactions.

36. ___ is the cyclical process of collecting and analyzing data during a research study.

A. Extremis Analysis
B. Constant analysis
C. Interim Analysis
D. All of the mentioned above

Answer: C) Interim Analysis

Explanation:

The cyclical process of gathering and assessing data throughout a research Endeavour is known as interim analysis.

37. An advantage of using computer programs for qualitative data is that they ___.

A. Can reduce time required to analyze data
B. Help in storing and organizing data
C. Make many procedures available that are rarely done by hand due to time constraints
D. All of the mentioned above

Answer: D) All of the mentioned above
Explanation: Qualitative data is that they can reduce time required to analyze data, help in storing and organizing data and make many procedures available that are rarely done by hand due to time constraints.

38. Data Modeling is the process of analyzing the data objects -

A. True
B. False

Answer: A) True
Explanation: The practice of evaluating data items and their relationships with other things is known as data modeling. It's utilized to look into the data requirements for various business activities. The data models are constructed in order to store the information in a database.

39. ___ are the basic building blocks of qualitative data.

A. Categories
B. Data chunk
C. Numeric figures
D. None of the mentioned above

Answer: A) Categories
Explanation: The fundamental building elements of qualitative data are categories. The descriptive and conceptual results gathered through surveys, interviews, or observation is referred to as qualitative data. We can explore concepts and further explain quantitative outcomes by analyzing qualitative data.

40. Metadata and data modeling tools support the creation and documentation of models -

A. True
B. False

Answer: A) True
Explanation:
Models representing the structures, flows, mappings and transformations, connections, and quality of data may be created and documented using metadata and data modeling tools.

41. The Process of describing the data that is huge and complex to store and process is known as ___.

A. Analytics mining
B. Data cleaning
C. Big data
D. None of the mentioned above

Answer: C) Big data

Explanation:Big data is a term used to describe the process of describing data that is large and difficult to store and interpret. Big data analytics is the use of advanced analytic techniques to very large, heterogeneous big data sets, which can contain structured, semi-structured, and unstructured data, as well as data from many sources and sizes ranging from terabytes to zettabytes.

42. In descriptive statistics, data from the entire population or a sample is summarized with ___.

A. Numerical descriptor
B. Decimal descriptor
C. Integer descriptor
D. All of the mentioned above

Answer: A) Numerical descriptor

Explanation: Data from the full population or a sample is summarized using numerical descriptors in descriptive statistics.

43. Customer behavior analytics is about understanding how your customers act.

A. True
B. False

Answer: A) True

Explanation:

Understanding how your customers behave across each channel and interaction point is the goal of customer behavior analytics. Understanding consumer behavior may aid in customer acquisition, engagement, and retention for your company.

44. Data Analysis is defined by the statistician?

A. John Tukey
B. Hans Peter Luhn
C. Gregory Lon
D. None of the mentioned above

Answer: A) John Tukey

Explanation:

John Tukey, a statistician, defined data analysis. Tukey began his career in statistics, and he was fascinated with data analysis challenges and methodologies. Some people remember him for pioneering exploratory data analysis, but he also made significant contributions to analysis of variance, regression, and a wide range of applications. This study examines some of the most notable contributions in these fields.

45. Amongst which of the following is / are the challenges overcome by the data strategy to make a business in a strong position -

A. Data privacy, data integrity, and data quality issues that undercut your ability to analyze data
B. Inefficient movement of data between different parts of the business
C. Lack of deep understanding of critical parts of the business
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Data strategy aids in the development of a strong firm. It also puts a company in a good position to overcome obstacles. Issues with data privacy, integrity, and quality that limit your capacity to evaluate data Lack of understanding of important business components and the processes that keep them run Inefficient data transportation between different portions of the organization, or data duplication by several business units, as well as a lack of clarity about current business needs and goals.

46. Tableau is a ___ tool.

A. Visualization
B. Analytical
C. Data Exploration
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Tableau is a visualization software program. Tableau gives data scientists a versatile front-end for data exploration with the analytical depth they need. Data scientists may execute complicated quantitative studies in Tableau and communicate visual findings to encourage improved understanding and collaboration with data by utilizing advanced computations, R and Python integration, quick cohort analysis, and predictive capabilities.

47. Big data analytics refers to collecting, processing, cleaning, and analyzing large datasets -

A. True
B. False

Answer: A) True

Explanation:

Big data analytics is the process of gathering, processing, cleaning, and analyzing enormous datasets in order to assist businesses operationalize their data.

48. Amongst which of the following is / are the features of Tableau for data analytics -

A. Data Blending
B. Real time analysis
C. Collaboration of data
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Tableau software's finest features are data blending, real-time analysis, and data collaboration. The beautiful thing about Tableau software is that it can be used without any technical or programming knowledge. The tool has piqued the curiosity of people from many walks of life, including business, researchers, and other industries.

49. ___ is a category, also called supervised machine learning methods in which the data is split on two parts.

A. Classification
B. Clustering
C. Data mining
D. None of the mentioned above

Answer: A) Classification

Explanation: Classification is a type of supervised machine learning approach in which the data is divided into two parts: a training set and a validation set. A model is trained from the training set by extracting the most discriminative characteristics that are previously connected with known outputs. This model is then tested on a test set, in which we evaluate the learnt model's efficiency by creating appropriate outputs for a particular set of input values.

50. Clustering belongs to ___ data analysis.

A. Supervised
B. Unsupervised
C. Both A and B
D. None of the mentioned above

Answer: B) Unsupervised

Explanation:

Unsupervised data analysis includes clustering. Without any prior knowledge, the data's hidden structure is discovered and emphasized. Popular clustering techniques include K-means, K-nearest neighbors, and hierarchical clustering.

51. Data analytics refers to a,

A. Analyzing Data
B. Scientific approach to get insights from the data or data sets
C. Find the hidden patterns from data
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:Data analytics refers to a scientific approach to get insights from the data or data sets. Data analytics not only analyzes the data but also explores it to find the hidden patterns from data. Data Analytics examines data sets in order to identify trends and develop conclusions about the information contained within them.

52. Data analytics applies on,

A. Raw Data or Data Set
B. Algorithm
C. Scientific method
D. None of the mentioned above

Answer: A) Raw Data or Data Set

Explanation:Data analytics applies to raw data to convert it into useful information. The science of studying raw data in order to draw conclusions about it is known as data analytics. Data analytics techniques and processes have been turned into mechanical processes and algorithms that operate on raw data for human consumption.

53. A good data analytics explores the data from different angles to get the information,

A. True
B. False

Answer: A) True

Explanation: A good data analytics explores the data from different angles to get the information which can be used in decision making.

54. Amongst which of the following is / are the significance of data analytics,

A. To collect the data, store it and put analysis
B. Analyse the data to get the fruitful insights and hidden information
C. Organizations can utilize data analytics to gain control over their data
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Companies are collecting the data, store it and put analysis on it to get fruitful insights and hidden information from the data. It can be identified and recognized that what data is saying and what will be the future prediction so that organizations existence can be established.

55. Data analytics works by analyzing large data sets with a variety of tools and methods,

A. True
B. False

Answer: A) True

Explanation:

Data analytics works by analyzing large data sets with a variety of tools and methods in order to uncover unique patterns, hidden correlations and relevant trends, and other insights that can be used to make data-driven decisions in the pursuit of improved results.

5**6. Data collection refers to,**

A. Store the data
B. Collect the data
C. Process the data
D. None of the mentioned above

Answer: B) Collect the data

Explanation: Data collection refers to collect the data from the source which is needed to analyze. This can be accomplished through a variety of means, including computers, web sources, cameras, environmental sources, human beings, etc.

5**7. Data Cleaning refers to the act of preparing data for analysis through the removal or modification of data,**

A. True
B. False

Answer: A) True

Explanation: Data cleaning is the act of preparing data for analysis through the removal or modification of data that is erroneous, incomplete, irrelevant or duplicated, or that has been incorrectly formatted.

58. Data analysis is a process to draw insights through numerous data sets.

A. True
B. False

Answer: A) True

Explanation: Data analysis is a process to draw insights through numerous data sets. Data analytics techniques and processes have been turned into mechanical processes and algorithms that operate on raw data for human consumption. A company's performance can be improved by using data analytics.

59. Amongst which of the following is / are the tools used in Data Analytics,

A. SAS
B. R
C. Python
D. All of the mentioned above

Answer: D) All of the above mentioned

Explanation: SAS, R and Python are some set of popular tools which are used in the data analytics.

60. Data analytics plays a vital role in Decision Making.

A. True
B. False

Answer: A) True

Explanation: Data analysis plays a vital role in decision-making. The goal is to make these corporate activities more time-efficient by streamlining them. Operational costs, product development, and labor planning are just a few examples.

1. Unprocessed data or processed data are observations or measurements that can be expressed as text, numbers, or other types of media?

A. True
B. False

Answer: A) True

Explanation: Data are observations or measurements (unprocessed or processed) represented as text, numbers, or multimedia. Information that has been transformed into a form that is more efficient for movement or processing is referred to as data in computing.

2. With reference to computing aspects ___ is a symbolic representation of facts or concepts from which information may be obtained with a reasonable degree of confidence.

A. Program
B. Knowledge
C. Data
D. Flowchart

Answer: A) Program

Explanation: With reference to computing aspects data is a symbolic representation of facts or concepts from which information may be obtained with a reasonable degree of confidence.

3. Which of the following can be considered to be the primary source of unstructured data among the others?

A. Facebook
B. Twitter
C. Internet webs
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Facebook, Twitter and Internet webs can be considered to be the primary source of unstructured data among the others.

4. Amongst which of the following is/are the examples of structured data -

A. Videos
B. Employee's name, employee's id, employee's age
C. Audio files
D. All of the mentioned above

Answer: B) Employee's name, employee's id, employee's age

Explanation: Structured data is extremely particular and is recorded in a set format, whereas unstructured data is a mashup of many different forms of data that are all stored in their original formats, as opposed to structured data. In above question, Employees name, employee's id, employee's age is an example of structured data.

5. Amongst which of the following step is performed by data scientist after acquiring the data?

A. Deletion
B. Data Replication
C. Data Integration
D. Data Cleansing

Answer: D) Data Cleansing

Explanation: after acquiring the data, data scientists perform Data Cleansing. Data cleansing is a critical step in preparing data for use in subsequent operations, whether in operational activities or in downstream analysis and reporting. It is most effectively accomplished with the use of data quality technologies. Depending on their purpose, these tools can perform a number of tasks ranging checking basic typographical errors to validating values against a known true reference set.

6. Quantitative data mainly deals with _____.

A. Audio data
B. Images data
C. Numeric data
D. Videos

Answer: C) Numeric data

Explanation: Quantitative data mainly deals with Numeric data Quantitative data is defined as the value of data in the form of counts or numbers, where each data-set has a unique numerical value associated with it, and where each data-set has a unique numerical value associated with it.

7. Big Data is a term that refers to data that is both too massive and impossible to be stored in _____.

A. Traditional databases
B. Big Databases
C. SQL Databases
D. All of the mentioned above

Answer: A) Traditional databases

Explanation: Big Data is a term that refers to data that is both too massive and impossible to be stored in Traditional databases. The quantities, letters, or symbols on which computer operations are done, which may be stored and conveyed in the form of electrical impulses and recorded on magnetic, optical, or mechanical storage media.

8. Big Data is a field dedicated to,

A. Storage of large collections of data
B. Processing
C. Analysis
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Big Data is a field dedicated to Storage of large collections of data, Processing and Analysis. Big data is defined as data that is so massive, quick, or complicated that it is difficult or impossible to process it using traditional methods, as opposed to little data. Having access to and keeping massive amounts of data for the purpose of analytics has been around for quite some time. The concept of big data, on the other hand, gained traction in the early 2000s.

9. Data that is less than 10 GB in size can be considered to be a little amount of data.

A. Small
B. Medium
C. Big
D. All of the mentioned above

Answer: A) Small

Explanation: Data that is less than 10 GB in size can be considered as a small data. Small data is data that is 'small' enough to be comprehended by a human being. It is information in a volume and manner that makes it easily accessible, instructive, and actionable for the intended audience.

10. Which of the following are benefits of Data Processing?

A. Cost Reduction
B. Time Reductions
C. Smarter Business Decisions
D. All of the mentioned above

Answer: D) All of the mentioned above
Explanation:

- When data is collected and transformed into useful information, this is referred to as data processing. Data processing is typically undertaken by a data scientist or team of data scientists, and it is critical that it is done correctly in order to avoid having a negative impact on the final product, or data output.
- Rather than starting with unstructured data in its raw form, data processing transforms information into a more understandable format (graphs, documents, etc.), providing it the form and context that are required for it to be processed by computers and used by personnel throughout an organization.

11. Which is the process of examining large and varied data sets?

A. Machine learning
B. Cloud computing
C. Big data analytics
D. All of the mentioned above

Answer: D) All of the mentioned above
Explanation: Big data analytics is the process of examining large and varied data sets. In the context of big data analytics, the application of advanced analytic techniques to extremely large and heterogeneous big data sets that contain structured, semi-structured, and unstructured data, from a variety of sources, and in various sizes ranging from terabytes to zettabytes is described.

12. Data Identification → Data Acquisition & Filtering → Data Extraction → Data Validation & Cleansing, are the phases of?

A. Data Analytics Lifecycle
B. System Analysis and Design
C. Software Development and Life Cycle
D. None of the mentioned above

Answer: A) Data Analytics Lifecycle
Explanation:
Data Identification, Data Acquisition & Filtering, Data Extraction, Data Validation & Cleansing are the phases of Data Analytics Lifecycle. The Data Analytics Lifecycle is a diagram that depicts these steps for professionals that are involved in data analytics projects. The phases of the Data Analytics Lifecycle are organized in a systematic manner to build a Data Analytics Lifecycle. Each phase has its own significance as well as its own set of traits.

13. Hadoop is a framework that is free and open source.

A. True
B. False

Answer: A) True
Explanation: Hadoop is an open-source platform. The Hadoop software library is a framework that enables for the distributed processing of massive data sets across clusters of computers using simple programming models. It is a component of the Apache Hadoop software library. It is intended to grow from a small number of servers to thousands of devices, each of which can do computing and storage on its own.

14. Hadoop File System is constantly required to deal with enormous amounts of data ____.

A. Network
B. Clusters
C. Data sets
D. None of the mentioned above

Answer: C) Data sets
Explanation: Hadoop File System is constantly required to deal with enormous amounts of data sets. HDFS is a distributed file system that can handle big data volumes and is designed to run on low-cost commodity computing gear. It is used to grow a single Apache Hadoop cluster to hundreds (or even thousands) of nodes by using a distributed computing model. HDFS is one of the three key components of Apache Hadoop, the other two being MapReduce and YARN. HDFS is used to store and organize data.

15. Hadoop is a framework that is used to work with _____.

A. MapReduce, Hive and HBase
B. MapReduce, MySQL and Google Apps
C. MapReduce, Hummer and Iguana
D. MapReduce, Heron and Trumpet

Answer: A) MapReduce, Hive and HBase
Explanation: Hadoop is a framework that is used to work with MapReduce, Hive and HBase. Hadoop is an open-source framework that can be used to store and process enormous datasets ranging in size from gigabytes to petabytes of data in a scalable and efficient manner. As opposed to employing a single huge computer to store and analyze all of the data, Hadoop enables for the clustering of numerous computers to analyze enormous datasets in parallel, allowing for faster analysis.

16. Amongst which of the following accurately describe Hadoop?

A. Open-source
B. Real-time
C. Java-based
D. Distributed computing approach

Answer: B) Real-time
Explanation: Hadoop is a Real-time data processing framework. Hadoop was originally intended to be used for batch processing. That is, take a large dataset as input and analyze it all at the same time, then create a large output dataset. The very concept of MapReduce is geared toward batch processing rather than real-time processing. This was true from the beginning of Hadoop's existence; today, however, there are numerous options to use Hadoop in an even more real-time manner.

17. ___ has the world's largest Hadoop cluster.

A. Apple
B. Datamatics
C. Facebook
D. None of the mentioned above

Answer: C) Facebook
Explanation: Facebook has the world’s largest Hadoop cluster.

18. Amongst which of the following is a correct statement?

A. Machine learning emphasizes on prediction, based on well-known properties learned from the training data
B. Data Cleaning emphasizes on prediction, based on well-known properties learned from the training data
C. Both a and b
D. None of the mentioned above

Answer: A) Machine learning emphasizes on prediction, based on well-known properties learned from the training data
Explanation: Machine learning emphasizes on prediction, based on well-known properties learned from the training data. Machine learning is the study of computer algorithms that can improve themselves automatically as a result of their experience and the usage of data collected from various sources. It is considered to be a component of AI. Machine learning algorithms create a model based on sample data, known as training data, in order to make predictions or choices without being explicitly taught to do so. They can accomplish this without being explicitly coded.

19. Which of the characteristics of big data is, in terms of importance, more concerned with data science?

A. Variety
B. Velocity
C. Volume
D. None of the mentioned above

Answer: A) Variety

Explanation: Variety in data is a main characteristic of big data which is more concerned with data science.

20. In which of the following areas do information management firms specialize in analytical capabilities?

A. Stream Computing
B. Content Management
C. Information Integration
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Stream Computing, Content Management and Information Integration are the areas in which information management firms specialize in analytical capabilities.

21. The use of reporting and visualization features in Data Analytics refers,

A. Processing of data
B. User friendly representation
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation:

The use of reporting and visualization features in Data Analytics refers to the processing of data and User-friendly representation. The graphical display of information and data is referred to as data visualization. Data visualization tools, which make use of visual components like as charts, graphs, and maps, make it easier to detect and analyze trends, outliers, and patterns in large amounts of information.

22. BI stands for ____.

A. Business Information
B. Business Initiation
C. Business Intelligence
D. Business Insider

Answer: C) Business Intelligence

Explanation: BI stands for Business Intelligence. Business Intelligence (BI) is concerned with complicated techniques and technology that assist end-users in analyzing data and performing decision-making activities in order to expand their businesses. Business intelligence is essential in the management of business data and the management of performance.

23. The primary introduction of Power BI was dependent on,

A. Microsoft Word
B. Microsoft Excel
C. Microsoft Outlook
D. Microsoft PowerPoint

Answer: B) Microsoft Excel

Explanation: The primary introduction of Power BI was dependent on Microsoft Excel. It is possible to consolidate self-service and enterprise data into a single view with Power BI, even when the data comes from multiple sources.

24. To consolidate inquiries in Power BI, what method do you employ?

A. Join Queries
B. Union Queries
C. Both A & B
D. None of the above

Answer: A) Join Queries
Explanation: To consolidate inquiries in Power BI, Join Queries method employ. When we combine data, we connect to two or more data sources, shape them as needed, and then consolidate them into a relevant query for the end user. The Power Query Editor in Power BI Desktop makes extensive use of the right-click menus as well as the Transform ribbon to perform complex transformations. The majority of the options available through the ribbon can also be accessed by right-clicking an object on the ribbon, such as a column, and selecting from the menu that appears.

25. What is the most effective method of preparing your data for Power BI?

A. User of a star schema
B. Load all tables
C. Include multiple objects
D. None of the above

Answer: A) User of a star schema
Explanation: The most effective method of preparing data for Power BI is a User of a star schema. Among relational data warehouses, the star schema is a mature modeling method that has been widely implemented. In order to comply with this requirement, modelers must categories their model tables as either dimensions or facts.

26. Access to Streaming Data is associated with _____.

A. System administrator
B. HDFS
C. Network System
D. None of the mentioned above

Answer: B) HDFS
Explanation: Access to Streaming Data is associated with HDFS. In the Hadoop distribution, there is an application called Hadoop streaming that may be used to stream data. Using the tool, you can construct and run Map/Reduce tasks that can use any executable or script as the mapper and/or the reducer, depending on your preferences.

27. Power BI is used by a variety of companies, including Facebook, Twilio, GitHub, and MailChimp as,

A. Online services
B. Database data sources
C. File data sources
D. None of the mentioned above

Answer: A) Online services
Explanation: Power BI is used by a variety of companies, including Facebook, Twilio, GitHub, and MailChimp as Online services. In any organization, systems generate a large amount of data, which can be measured in terabytes, petabytes, or even exabytes in some instances.
Businesses use Business Intelligence to evaluate this data and turn it into actionable information (decisions), and the entire process is referred to as business intelligence. It is undeniable that the success of the firm is dependent on the decisions that are made as a result of business intelligence.

28. When it comes to Power BI Desktop, which of the following might be regarded the most important feature?

A. Data
B. Report
C. Dashboard
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Data, Report and Dashboard are the most important features. Power BI Desktop is used to gather, organize, transform, and visualize data in various ways. With Power BI Desktop, we can connect to a variety of different data sources and merge them (a process known as modeling) into a single data model for analysis.

29. Amongst which of the following is must before using any technology to evaluate your data,

A. Study the dataset
B. Organize dataset
C. Remove impurities from data set
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: Before using any technology to evaluate your data we must study the dataset, organize dataset and remove impurities from data set. Before we begin collecting data, we must develop a detailed analysis strategy that will guide us through the various steps of the research process, from summarizing and characterizing the data to testing our hypotheses.

30. Power BI modelling refers to the relationships that exist between your data sources.

A. True
B. False

Answer: A) True
Explanation: Data Modeling is one of the aspects in a business intelligence tool that is used to connect multiple data sources through the usage of a relationship. A relationship explains how data sources are connected to one another, and we can use relationships to generate fascinating data visualizations across a variety of data sets. In Power BI, we can also see the "Relationship" between two variables in a data model.

1. Statistical data analysis deals with the usage of statistical tools.

A. True
B. False

Answer: A) True

Explanation: Statistical data analysis is concerned with the application of specific statistical tools, which necessitates the knowledge of statistics. Statistical data analysis is a procedure that entails the application of a variety of statistical operations. In quantitative research, it is a type of research that seeks to quantify the data by using some form of statistical analysis, which is typically applied. Survey data and observational data are examples of quantitative data, which is primarily comprised of descriptive data.

2. Sampling is a process used in ___.

A. Network setting
B. Statistical analysis
C. Semantic analysis
D. None of the mentioned above

Answer: B) Statistical analysis

Explanation: Sampling is a statistical analysis technique in which a predetermined number of observations are drawn from a larger population in order to conduct statistical analysis. Simple random sampling or systematic sampling may be used to select samples from a larger population, depending on the type of analysis being performed.

3. Probability sampling is a sampling technique where we set a selection of a few criteria and chooses members of a ___ randomly.

A. Population
B. Employee
C. Both A and B
D. None of the mentioned above

Answer: A) Population

Explanation:Probability sampling is a sampling technique in which a researcher selects a small number of criteria and then randomly selects members of a population from that selection. With this selection parameter, all of the members have an equal chance of becoming a part of the sample population.

4. Cluster sampling is a method where we divide the entire population into ___.

A. Sections
B. Area
C. Both A and B
D. None of the mentioned above

Answer: A) Sections

Explanation: Cluster sampling is a technique in which researchers divide a large population into sections or clusters that are representative of a population under investigation. Clusters are identified and included in a sample based on demographic parameters such as age, gender, and geographic location, among others. Because of this, it is very simple for the creator of a survey to derive useful inferences from the responses received.

5. ___ to choose the sample members of a population at regular intervals.

A. Systematic sampling
B. Random sampling
C. Stratified random sampling
D. None of the mentioned above

Answer: B) Random sampling

Explanation: Using the systematic sampling method, we can randomly select members of a population to represent a sample on a regular basis. Starting points for the sample as well as a sample size that can be repeated at regular intervals are necessary for this procedure to be successful. Due to the fact that this type of sampling method has a predetermined range, it is the least time-consuming of the sampling techniques.

6. ___ divides the population into smaller groups that don't overlap but represent the entire population.

A. Systematic sampling
B. Stratified random sampling
C. Probability sampling
D. None of the mentioned above

Answer: B) Stratified random sampling

Explanation: Stratified random sampling is a technique that divides a population into smaller groups that do not overlap but are representative of the entire population. While sampling, these groups can be organized and then a sample drawn from each group can be used to determine the results.

7. Amongst which of the following is / are the uses of probability sampling -

A. Reduce Sample Bias
B. Diverse Population
C. Create an Accurate Sample
D. All of the mentioned above

Answer: D) All of the mentioned above
Explanation: Probability sampling can be used in a variety of situations. The bias in a sample derived from a population is negligible to non-existent when the probability sampling method is used, according to the literature. The researcher's understanding and inferences are primarily reflected in the sample he or she chooses for analysis. Population with a Wide Range of Characteristics - When the population is large and diverse, it is critical to have adequate representation so that the data is not skewed towards one demographic. Additionally, it generates a precise Sample. Probability sampling aids in the planning of the research and the creation of an accurate sample. This aids in the collection of well-defined data.

8. Judgmental or purposive sampling is formed by the purpose of the study, along with the understanding of the target audience.

A. True
B. False

Answer: A) True
Explanation: Alternatively referred to as purposive sampling, this nonprobability sampling technique selects participants for a sample based solely on the knowledge and judgement of the researcher. The use of judgement sampling increases the relevance of the sample to the population of interest because only those individuals who meet specific criteria are selected for inclusion in the sample. Researchers strive to use the sample that is the most representative of the population of interest in order for a study to be carried out as efficiently and effectively as possible.

9. If the data has a singular variable, then univariate statistical data analysis can be conducted -

A. True
B. False

Answer: A) True
Explanation:
The data consists of variables that are either univariate or multivariate, and depending on the number of variables, the experts use a variety of statistical techniques to analyze the data. Once a single variable has been identified in the data, univariate statistical data analysis can be performed, such as the t-test for significance, the z test, the f test, the ANOVA one-way test, and so on. And, if the data contains a large number of variables, various multivariate techniques, such as statistical data analysis, discriminant statistical data analysis, and so on, can be used.

10. ___ sampling takes a consecutive series of items.

A. Block
B. Set
C. Heuristic
D. None of the mentioned above

Answer: A) Block
Explanation: Block sampling is a method of selecting a sample from a population by selecting a series of items from the population in a sequential manner. Block sampling is a sampling technique used in auditing that involves making a series of selections in a sequential fashion. This method is extremely efficient because it allows for the retrieval of a large number of documents from a single location. While a more random selection method would be preferable for sampling the entire population, it would be more effective in this case. A large number of sample blocks can be selected from a pool of samples when using block sampling to reduce the likelihood of sampling error

11. Measures of central tendency help to find the middle, or the ___ of a data set.

A. Average
B. Distribution
C. Integration and differential
D. None of the mentioned above

Answer: A) Average

Explanation: Measures of central tendency aid in the discovery of the middle, or the average, of a data set of observations. A single value that attempts to describe a set of data by identifying the central position within that set of data is referred to as a measure of central tendency (or central tendency index). As a result, measures of central tendency are also referred to as measures of central location in some cases.

12. Amongst which of the following is / are the measures of central tendency -

A. Mean
B. Median
C. Mode
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: The mode, median, and mean are the three most commonly used measures of central tendency. The mean is calculated by dividing the sum of all values by the total number of values. In an ordered data set, the median is the number in the middle, whereas the mode is the most frequent value.

13. A data set is a distribution of n number of scores or values.

A. True
B. False

Answer: A) True

Explanation: When all possible values for a data set are plotted on a frequency graph, the shape of the graph represents the distribution of the data set. The majority of the time, we are unable to collect all of the data for our variable of interest. As a result, we collect a sample. The results of this sample are used to draw conclusions about the entire data set.

14. In a ___, data is symmetrically distributed with no skew.

A. Normal distribution
B. Binomial distribution
C. Bernoulli distribution
D. None of the mentioned above

Answer: A) Normal distribution

Explanation: A normal distribution is characterized by data that is symmetrically distributed and does not have any skew. The majority of the values are concentrated around a central region, with values diminishing as they move further away from the centre. In a normal distribution, the mean, the mode, and the median are all exactly the same.

15. In ___, more values fall on one side of the center than the other, and the mean, median and mode all differ from each other.

A. Normal distribution
B. Skewed distribution
C. Parallel distribution
D. None of the mentioned above

Answer: B) Skewed distribution

Explanation: In skewed distributions, more values are found on one side of the center than on the other, and the mean, median, and mode are all different from one another as a result. The tail of one side is more spread out and longer than the tail of the other, with fewer scores at one end than the tail of the other. The direction of this tail indicates which side of the skew it is on. The right-hand tail of a positively skewed distribution has a cluster of lower scores on the left and a more widely spread-out tail on the right. In a negatively skewed distribution, there is a cluster of higher scores on the right side of the distribution, and a spread-out tail on the left side.

16. The mode is the most frequently occurring value in the data set. It's possible to have ___.

A. No mode
B. One mode
C. More than one mode
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation: The mode is defined as the value that appears the most frequently in the data set. It is possible to have no mode, only one mode, or multiple modes at the same time. To find the mode, sort your data set numerically or categorically, and then choose the response that occurs the most frequently from the results.

17. The range is given as the smallest and ___ observations.

A. Largest
B. Medium
C. Mean
D. None of the mentioned above

Answer: A) Largest

Explanation: The range is defined as the difference between the smallest and largest observations. This is the most straightforward way to assess variability. In statistics, a range is defined by the difference between two numbers, rather than the difference between the smallest and largest numbers. It is extremely useful for certain types of data.

18. A ___ is a statistical term that describes a division of observations into four defined intervals.

A. Quartile
B. Mean
C. Median
D. All of the mentioned above

Answer: B) Mean

Explanation:

Quarterlines are statistical terms that describe the division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observations. Quartiles are used in data analysis to describe the division of observations into four defined intervals.

19. The difference between the upper and lower quartile is known as the interquartile range.

A. True
B. False

Answer: A) True

Explanation:

The interquartile range (IQR) is a statistical measure of the spread of your data's middle half. It represents the range of values for the middle 50% of your sample. The interquartile range (IQR) can be used to determine the variability in the areas where the majority of your values are found. Larger values indicate that the central portion of your data has spread out further than the rest of your information. Smaller values, on the other hand, indicate that the middle values are more tightly clustered.

The formula for the interquartile range is given below - **Interquartile range = Upper Quartile – Lower Quartile = $Q_3 – Q_1$**

Where Q_1 represents the first quartile of the series and Q_3 represents the third quartile of the series.

20. The lowest quartile (Q1) refers the smallest quarter of values in the dataset. The upper quartile (Q4) refers the ___ quarter of values.

A. Highest
B. Lowest
C. Central one
D. None of the mentioned above

Answer: A) Highest

Explanation:

The first quartile (Q1) of your dataset contains the values that make up the smallest quarter of the total. The upper quartile (Q4) contains the values that are the highest quarter of the distribution. The interquartile range (IQR) is the half of the data that falls between the upper and lower quartiles, and it is defined as follows: If you want to visualize the interquartile range, think of your data as being divided into quarters. Quarters are referred to as quartiles by statisticians, and they are labelled as Q1, Q2, Q3, and Q4 in descending order from low to high.

Data Analysis Question 1:

Comprehension:

Directions: Consider the following data and answer questions:

S. No.	Class limit	Frequency
1.	11-20	13
2.	21-30	6
3.	31-40	12
4.	41-50	39
5.	51-60	46
6.	61-70	14
7.	71-80	10
8.	81-90	5
9.	91-100	5

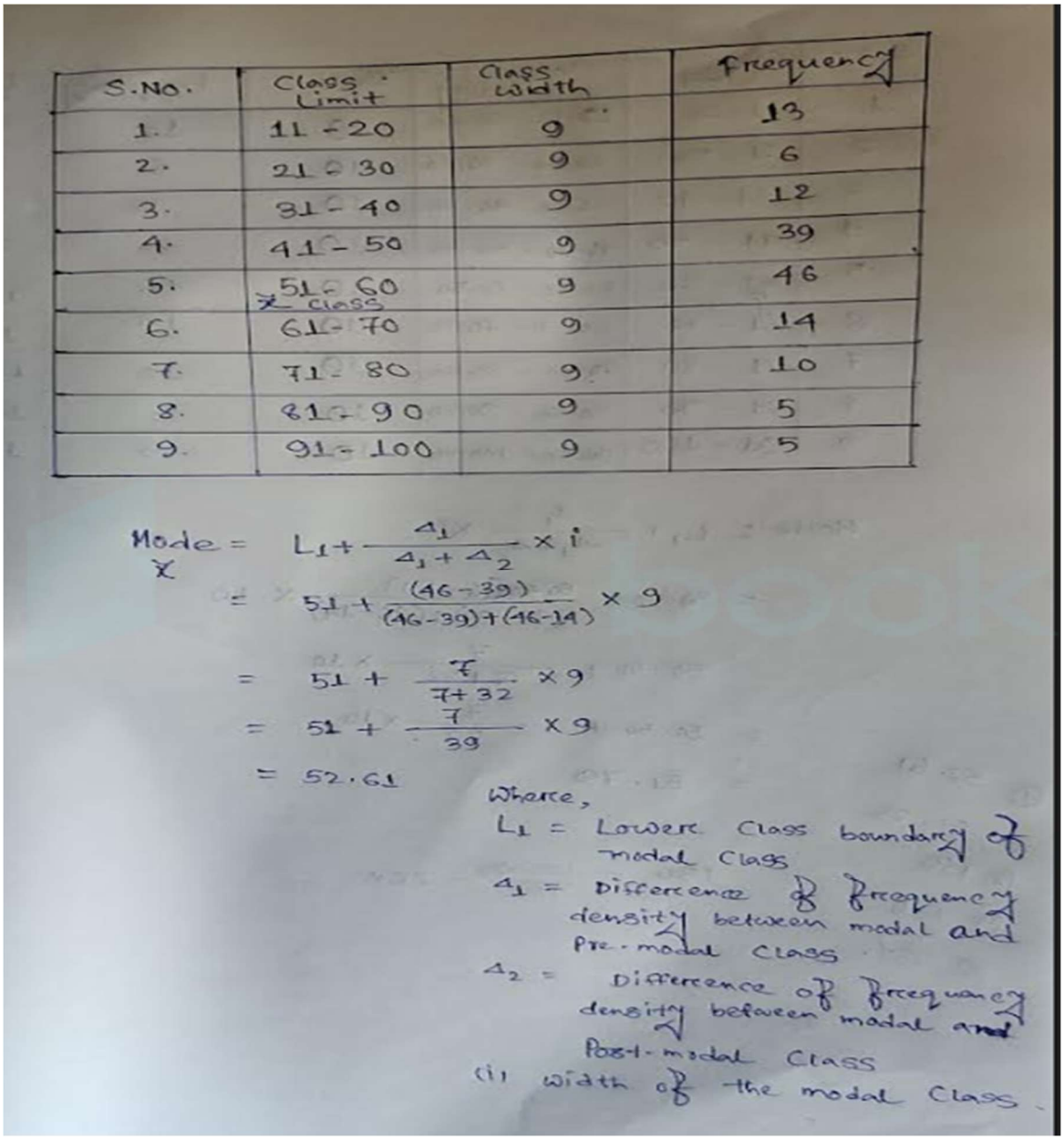
Which one of the following is the mode value for the given data set

Which one of the following is the mode value for the given data set

1. 52.61
2. 45.22
3. 39.81
4. 55.21

Answer (Detailed Solution Below)

Option 1 : 52.61



S.No.	Class Limit	Class Width	Frequency
1.	11 - 20	9	13
2.	21 - 30	9	6
3.	31 - 40	9	12
4.	41 - 50	9	39
5.	51 - 60	9	46
6.	61 - 70	9	14
7.	71 - 80	9	10
8.	81 - 90	9	5
9.	91 - 100	9	5

* Class

$$\text{Mode} = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$= 51 + \frac{(46-39)}{(46-39)+(46-14)} \times 9$$

$$= 51 + \frac{7}{7+32} \times 9$$

$$= 51 + \frac{7}{39} \times 9$$

$$= 52.61$$

Where,

L_1 = Lower Class boundary of modal Class

Δ_1 = Difference of frequency density between modal and Pre-modal Class

Δ_2 = Difference of frequency density between modal and Post-modal Class

(i) width of the modal Class.

Data Analysis Question 2:

Comprehension:

Directions: Consider the following data and answer questions:

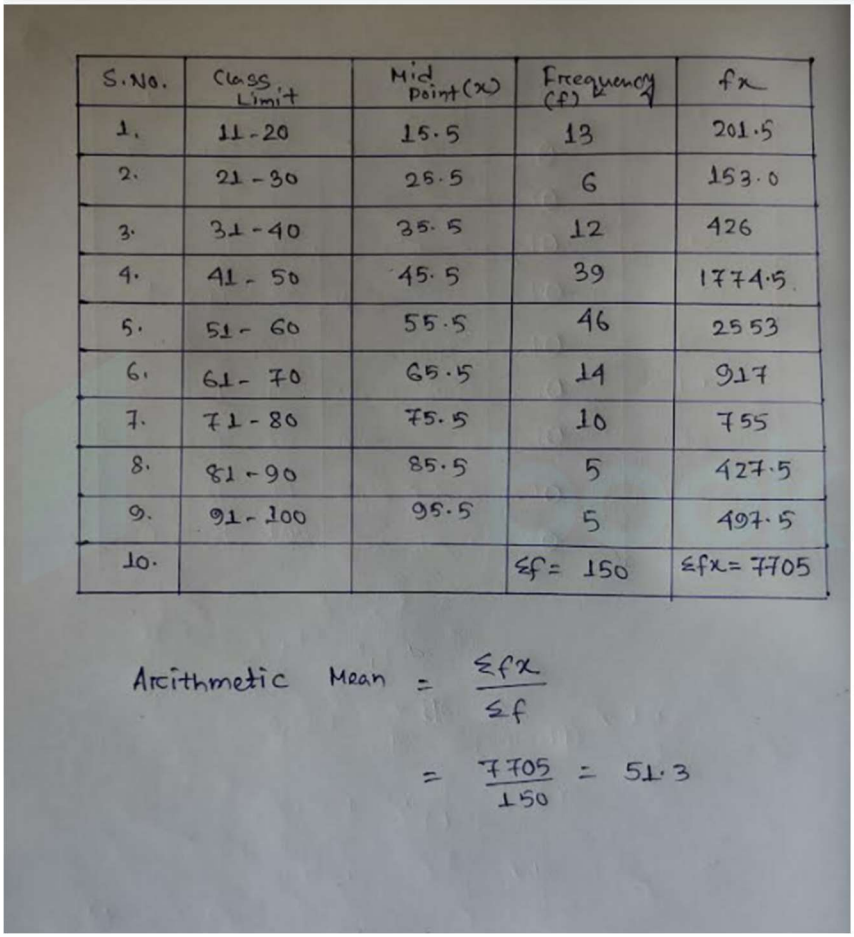
S. No.	Class limit	Frequency
1.	11-20	13
2.	21-30	6
3.	31-40	12
4.	41-50	39
5.	51-60	46
6.	61-70	14
7.	71-80	10
8.	81-90	5
9.	91-100	5

Which one of the following is the arithmetic mean value for the given data set

1. 46.87
2. 52.26
3. 49.22
4. 51.23

Answer (Detailed Solution Below)

Option 4 : 51.23



S.No.	Class Limit	Mid point (x)	Frequency (f)	fx
1.	11-20	15.5	13	201.5
2.	21-30	25.5	6	153.0
3.	31-40	35.5	12	426
4.	41-50	45.5	39	1774.5
5.	51-60	55.5	46	2553
6.	61-70	65.5	14	917
7.	71-80	75.5	10	755
8.	81-90	85.5	5	427.5
9.	91-100	95.5	5	497.5
10.			$\Sigma f = 150$	$\Sigma fx = 7705$

Arithmetic Mean $= \frac{\Sigma fx}{\Sigma f}$

$= \frac{7705}{150} = 51.3$

Data Analysis Question 3:

Comprehension:

Directions: Consider the following data and answer questions:

S. No.	Class limit	Frequency
1.	11-20	13
2.	21-30	6
3.	31-40	12
4.	41-50	39
5.	51-60	46
6.	61-70	14
7.	71-80	10
8.	81-90	5
9.	91-100	5

Which one of the following is the cumulative frequency of the entire data set

1. 150
2. 160
3. 140
4. 120

Answer (Detailed Solution Below)

Option 1 : 150

Data Analysis Question 4:

Comprehension:

Directions: Consider the following data and answer questions:

S. No.	Class limit	Frequency
1.	11-20	13
2.	21-30	6
3.	31-40	12
4.	41-50	39
5.	51-60	46
6.	61-70	14
7.	71-80	10
8.	81-90	5
9.	91-100	5

Which one of the following is the cumulative frequency for the class limit 61-70 from the given data set

1. 10
2. 116
3. 130
4. 140

Answer (Detailed Solution Below)

Option 3 : 130

Data Analysis Question 5:

Comprehension:

Directions: Consider the following data and answer questions:

S. No.	Class limit	Frequency
1.	11-20	13
2.	21-30	6
3.	31-40	12
4.	41-50	39
5.	51-60	46
6.	61-70	14
7.	71-80	10
8.	81-90	5
9.	91-100	5

Which one of the following is the relative frequency in percentage for class limit 41-50 from the given data set.

1. 42%
2. 26%
3. 16%
4. 24%

Answer (Detailed Solution Below)

Option 2 : 26%

Data Analysis Question 6

In Data Processing, what does the abbreviation SAP stand for ?

1. Systems, Applications, Products
2. Sales, Allocations, Purchases
3. Systems, Authorizations, Programs
4. Systems, Algorithms, Processes

Answer (Detailed Solution Below)

Option 1 : Systems, Applications, Products

Data Analysis Question 7

Which one of the following is a non-parametric statistic?

1. F- statistic
2. t - statistic
3. Pearson's correlation
4. Spearman's correlation

Answer (Detailed Solution Below)

Option 4 : Spearman's correlation

Data Analysis Question 8

Which of the following is a data visualization method?

1. Line
2. Circle and Triangle
3. Pie chart and Bar chart
4. Pentagon

Answer (Detailed Solution Below)

Option 3 : Pie chart and Bar chart

Data Analysis Question 9

In order to understand the classroom teaching-learning process, which of the following research tool is most appropriate?

1. Rating Scale
2. Questionnaire
3. Observation Schedule
4. Interview Schedule

Answer (Detailed Solution Below)

Option 3 : Observation Schedule

Data Analysis Question 10

Which company was recently implicated in a global data theft crime?

1. Amazon
2. Google
3. Cisco
4. Cambridge Analytica

Answer (Detailed Solution Below)

Option 4 : Cambridge Analytica

Data Analysis Question 11

If you want to compare the price of wheat over a period, which index will you use ?

1. Volume Index
2. Aggregate Index
3. Both (1) and (2)
4. Price Index

Answer (Detailed Solution Below)

Option 4 : Price Index

Data Analysis Question 12

Download Solution PDF

A researcher administers an achievement test to assess and indicate the possible effect of an independent variable in his/her study. The distribution of scores on the test is found to be negatively skewed. On the basis of this, what can be started with regard to the difficulty level of the test?

1. The test is very easy
2. The test is very difficult
3. The test is neither easy nor difficult
4. The test is easy and needs normalization

Answer (Detailed Solution Below)

Option 1 : The test is very easy

Data Analysis Question 13

Download Solution PDF

A statistical measure that indicates the extent to which changes in one factor are accompanied by changes in another

1. Standard deviation
2. Correlation coefficient
3. Quartile deviation
4. Range

Answer (Detailed Solution Below)

Option 2 : Correlation coefficient

Data Analysis Question 14

Download Solution PDF

Which among the following is a software for the analysis of qualitative data?

1. NVivo
2. R
3. SPSS
4. STATA

Answer (Detailed Solution Below)

Option 1 : NVivo

Data Analysis Question 15

Which among the following are statistical packages?

A. SPSS

B. SAS

C. ANOVA

D. STATA

E. R

Choose the correct answer from the options given below:

1. A, B, C and D only
2. A, B, C and E only
3. A, B, D and E only
4. B, C, D and E only

Answer (Detailed Solution Below)

Option 3 : A, B, D and E only

Statistical software is a specialized computer program for analysis in statistics and econometrics.

Important Points

Statistical packages	Description
SAS (Statistical Analysis System)	• SAS (previously "Statistical Analysis System") is a statistical software suite developed by SAS Institute for data management, advanced analytics, multivariate analysis, business intelligence, criminal investigation, and predictive analytics.
SPSS	• SPSS Statistics is a statistical software suite developed by IBM for data management, advanced analytics, multivariate analysis, business intelligence, criminal investigation.
STATA	• Stata is a general-purpose statistical software package developed by StataCorp for data manipulation, visualization, statistics, and automated reporting. It is used by researchers in many fields, including biomedicine, epidemiology, sociology, and science.
R	• R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. • Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners and statisticians for data analysis and developing statistical software. • Users have created packages to augment the functions of the R language.

Hence, the correct answer is A, B, D, and E only.

**********Data Analytics MCQs Set - 1**********

1. The branch of statistics which deals with development of particular statistical methods

1. industry statistics
2. economic statistics
3. applied statistics
4. applied statistics

Answer: applied statistics

2. Which of the following is true about regression analysis?

1. answering yes/no questions about the data
2. estimating numerical characteristics of the data
3. modeling relationships within the data
4. describing associations within the data

Answer: modeling relationships within the data

3. Text Analytics, also referred to as Text Mining?

1. True
2. False
3. Can be true or False
4. Cannot say

Answer: True

4. What is a hypothesis?

1. A statement that the researcher wants to test through the data collected in a study.
2. A research question the results will answer.
3. A theory that underpins the study.

Answer: A statement that the researcher wants to test through the data collected in a study.

5. What is the cyclical process of collecting and analysing data during a single research study called?

1. Interim Analysis
2. Inter analysis
3. inter item analysis
4. constant analysis

Answer: Interim Analysis

6. The process of quantifying data is referred to as ____

1. Topology
2. Digramming
3. Enumeration
4. coding

Answer: Enumeration

7. An advantage of using computer programs for qualitative data is that they
1. Can reduce time required to analyse data (i.e., after the data are transcribed)
2. Help in storing and organising data
4. All of the above
Answer: All of the Above

8. Boolean operators are words that are used to create logical combinations.
1. True
2. False
Answer: True

9. _____ are the basic building blocks of qualitative data.
1. Categories
2. Units
3. Individuals
4. None of the above

10. This is the process of transforming qualitative research data from written interviews or field notes into typed text.
1. Segmenting
2. Coding
3. Transcription
4. Mnemoning
Answer: Transcription

11. A challenge of qualitative data analysis is that it often includes data that are unwieldy and complex; it is a major challenge to make sense of the large pool of data.
1. True
2. False
Answer: True

12. Hypothesis testing and estimation are both types of descriptive statistics.
1. True
2. False
Answer: False

13. A set of data organised in a participants(rows)-by-variables(columns) format is known as a “data set.”
1. True
2. False
Answer: True

14. A graph that uses vertical bars to represent data is called a ___
1. Line graph
2. Bar graph
3. Scatterplot
4. Vertical graph
Answer: Bar graph

15. ____ are used when you want to visually examine the relationship between two Variables.
1. Bar graph
2. pie graph
3. line graph
4. Scatterplot
Answer: Scatterplot

16. The denominator (bottom) of the z-score formula is
1. Statistic
2. Hypothesis
3. Level of Significance
4. Test-Statistic
Answer: Hypothesis

17. Which of these distributions is used for a testing hypothesis?
1. Normal Distribution
2. Chi-Squared Distribution
3. Gamma Distribution
4. Poisson Distribution
Answer: Chi-Squared Distribution

18. A statement made about a population for testing purpose is called?

1. The standard deviation
2. The difference between a score and the mean
3. The range
4. The mean
Answer: The standard deviation

19. If the assumed hypothesis is tested for rejection considering it to be true is called?
1. Null Hypothesis
2. Positive Hypothesis
3. Negative Hypothesis
4. Alternative Hypothesis.
Answer: Alternative Hypothesis.

20. If the null hypothesis is false then which of the following is accepted?J

1. Null Hypothesis
2. Statistical Hypothesis
3. Simple Hypothesis
4. Composite Hypothesis
Answer: Null Hypothesis

21. Alternative Hypothesis is also called as?
1. Composite hypothesis
2. Research Hypothesis
3. Simple Hypothesis
4. Null Hypothesis
Answer: Research Hypothesis

********** Data Analytics MCQs Set – 2 **********

What is the minimum no. of variables/ features required to perform clustering?

A. 0
B. 1
C. 2
D. 3
Answer: At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram.

Which of the following algorithm is most sensitive to outliers?

1. K-means clustering algorithm
2. K-medians clustering algorithm
3. K-modes clustering algorithm
4. K-medoids clustering algorithm
Answer: K-means clustering algorithm

4. The discrete variables and continuous variables are two types of
    1. Open end classification
    2. Time series classification
    3. Qualitative classification
    4. Quantitative classification

    Answer: Quantitative classification

5. Bayesian classifiers is
    1. A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.
    2. Any mechanism employed by a learning system to constrain the search space of a hypothesis
    3. An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.
    4. None of these

    Answer: A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.

6. Classification accuracy is
    1. A subdivision of a set of examples into a number of classes
    2. Measure of the accuracy, of the classification of a concept that is given by a certain theory
    3. The task of assigning a classification to a set of examples
    4. None of these

    Answer: Measure of the accuracy, of the classification of a concept that is given by a certain theory

7. Euclidean distance measure is
    1. A stage of the KDD process in which new data is added to the existing selection.
    2. The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them
    3. The distance between two points as calculated using the Pythagoras theorem
    4. none of above

Answer: The distance between two points as calculated using the Pythagoras theorem

8. Hybrid is
    1. Combining different types of method or information
    2. Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.
    3. Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.
    4. none of above
    Answer: Combining different types of method or information

9. Decision trees use, in that they always choose the option that seems the best available at that moment
    1. Greedy Algorithms
    2. divide and conquer
    3. Backtracking
    4. Shortest path algorithm
    Answer: Greedy Algorithms

10. Discovery is
    1. It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).
    2. The process of executing implicit previously unknown and potentially useful information from data
    3. An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.
    4. None of these
    Answer: The process of executing implicit previously unknown and potentially useful information from data

12. Hidden knowledge referred to
    1. A set of databases from different vendors, possibly using different database paradigms
    2. An approach to a problem that is not guaranteed to work but performs well in most cases
    3. Information that is hidden in a database and that cannot be recovered by a simple SQL query.
    4. None of these
    Answer: Information that is hidden in a database and that cannot be recovered by a simple SQL query

13. Enrichment is

1. A stage of the KDD process in which new data is added to the existing selection
2. The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them
3. The distance between two points as calculated using the Pythagoras theorem.
4. None of these

Answer: A stage of the KDD process in which new data is added to the existing selection

14._____ are easy to implement and can execute efficiently even without prior knowledge of the data, they are among the most popular algorithms for classifying text documents.

1. 1D3
2. Naive Bayes classifiers
3. CART
4. None of above

Answer: Naive Bayes classifiers

15. High entropy means that the partitions in classification are

1. Pure
2. Not Pure
3. Usefull
4. useless

Answer: Uses a single processor or computer

16. Which of the following statements about Naive Bayes is incorrect?

A. Attributes are equally important.
B. Attributes are statistically dependent of one another given the class value.
C. Attributes are statistically independent of one another given the class value.
D. Attributes can be nominal or numeric

Answer: Attributes are statistically dependent of one another given the class value.

17. The maximum value for entropy depends on the number of classes so if we have 8 Classes what will be the max entropy.

1. Max Entropy is 1
2. Max Entropy is 2
3. Max Entropy is 3
4. Max Entropy is 4

Answer: Max Entropy is 3

18. Point out the wrong statement.

1. k-nearest neighbor is same as k-means
2. k-means clustering is a method of vector quantization
3. k-means clustering aims to partition n observations into k clusters
4. none of the mentioned

Answer: k-nearest neighbor is same as k-means

19. Consider the following example "How we can divide set of articles such that those articles have the same theme (we do not know the theme of the articles ahead of time) " is this:

1. Clustering
2. Classification
3. Regression
4. None of these

Answer: Clustering

20. Can we use K Mean Clustering to identify the objects in video?

1. Yes
2. No

Answer: Yes

21. Clustering techniques are in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters.

1. Unsupervised
2. supervised
3. Reinforcement
4, Neural network

Answer: Unsupervised

22. metric is examined to determine a reasonably optimal value of k.

A. Mean Square Error
B. Within Sum of Squares (WSS)
C. Speed
D. None of these

Answer: Within Sum of Squares (WSS)

23. If an itemset is considered frequent, then any subset of the frequent itemset must also be frequent.

1. Apriori Property
2. Downward Closure Property
3. Either 1 or 2
4. Both 1 and 2

Answer: Both 1 and 2Z

24. if {bread,eggs,milk} has a support of 0.15 and {bread,eggs} also has a support of 0.15, the confidence of rule {bread,eggs} = {milk} is

A. 1.0
B. 2.1
C. 3.2
D. 4.3

Answer: 1

25. Confidence is a measure of how X and Y are really related rather than coincidentally happening together.

1. True
2. False

Answer: False

26. recommend items based on similarity measures between users and/or items.

1. Content Based Systems
2. Hybrid System
3. Collaborative Filtering Systems
4. None of these

Answer: Collaborative Filtering Systems

27. There are major Classification of Collaborative Filtering Mechanisms

1.1
2.2
3.3
4. none of above

Answer: 2

28. Movie Recommendation to people is an example of

1. User Based Recommendation
2. Item Based Recommendation
3. Knowledge Based Recommendation
4. content-based recommendation

Answer: Item Based Recommendation

29. recommenders rely on an explicitly defined set of recommendation rules
    1. Constraint Based
    2. Case Based
    3. Content Based
    4. User Based

    Answer: Case Based

30. Parallelized hybrid recommender systems operate dependently of one another and produce separate recommendation lists.
    1. True
    2. False

    Answer: False

1. Data Analysis is a process of?

A. inspecting data
B. cleaning data
C. transforming data
D. All of the above

Ans: D

Explanation: Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.

2. Which of the following is not a major data analysis approach?

A. Data Mining
B. Predictive Intelligence
C. Business Intelligence
D. Text Analytics

Ans : B

Explanation: Predictive Analytics is major data analysis approaches not Predictive Intelligence.

3. How many main statistical methodologies are used in data analysis?

A. 2
B. 3
C. 4
D. 5

Ans : A

Explanation: In data analysis, two main statistical methodologies are used Descriptive statistics and Inferential statistics.

4. In descriptive statistics, data from the entire population or a sample is summarized with?

A. integer descriptors
B. floating descriptors
C. numerical descriptors
D. decimal descriptors

Ans : C

Explanation: In descriptive statistics, data from the entire population or a sample is summarized with numerical descriptors.

5. Data Analysis is defined by the statistician?

A. William S.
B. Hans Peter Luhn
C. Gregory Piatetsky-Shapiro
D. John Tukey

Ans : D
Explanation: Data Analysis is defined by the statistician John Tukey in 1961 as "Procedures for analyzing data.

6. Which of the following is true about hypothesis testing?

A. answering yes/no questions about the data
B. estimating numerical characteristics of the data
C. describing associations within the data
D. modeling relationships within the data

Ans : A
Explanation: answering yes/no questions about the data (hypothesis testing)

7. The goal of business intelligence is to allow easy interpretation of large volumes of data to identify new opportunities.

A. TRUE
B. FALSE
C. Can be true or false
D. Can not say

Ans : A
Explanation: The goal of business intelligence is to allow easy interpretation of large volumes of data to identify new opportunities.

8. The branch of statistics which deals with development of particular statistical methods is classified as

A. industry statistics
B. economic statistics
C. applied statistics
D. applied statistics

Ans : D

Explanation: The branch of statistics which deals with development of particular statistical methods is classified as applied statistics.

9. Which of the following is true about regression analysis?

A. answering yes/no questions about the data
B. estimating numerical characteristics of the data
C. modeling relationships within the data
D. describing associations within the data

Ans : C

Explanation: modeling relationships within the data (E.g. regression analysis).

10. Text Analytics, also referred to as Text Mining?

A. TRUE
B. FALSE
C. Can be true or false
D. Can not say

Ans : A

Explanation: Text Data Mining is the process of deriving high-quality information from text.

MCQs on Statistical Data Analysis Quiz

MCQ: The branch of statistics which deals with development of particular statistical methods is classified as

industry statistics	economic statistics
applied statistics	mathematical statistics

MCQ: The tools such decision making by nominal groups, brain storming and term buildings are all considered as

serial tools	behavioral tools
statistical tools	parallel tools

MCQ: The model which consists of management philosophy, behavioral tools and statistical methods as key steps towards improvement is considered as

serial improvement process model	behavioral improvement process model
quality improvement process model	statistics improvement process model

MCQ: The branch of statistics which deals with findings of solution in the field of medicine, education and economics is classified as

economic statistics	applied statistics
mathematical statistics	industry statistics

MCQ: The analysis based on study of price fluctuations, production of commodities and deposits in banks is classified as

sample series analysis	time series analysis
numerical analysis	experimental analysis

All Units: Data Analytic MCQs Questions

1. The branch of statistics that deals with the development of particular statistical methods are classified as

1. industry statistics
2. economic statistics
3. applied statistics
4. applied statistics

Answer: 3

2. Which of the following is true about regression analysis?

1. answering yes/no questions about the data
2. estimating numerical characteristics of the data
3. modeling relationships within the data
4. describing associations within the data

Answer: 3

3. Text Analytics, also referred to as Text Mining?

1. True
2. False
3. Can be true or False
4. Can not say

Answer: 1

4. What is a hypothesis?

1. A statement that the researcher wants to test through the data collected in a study.
2. A research question the results will answer.
3. A theory that underpins the study.
4. A statistical method for calculating the extent to which the results could have happened by chance.

Answer: 1

5. What is the cyclical process of collecting and analyzing data during a single research study called?
1. Interim Analysis
2. Inter analysis
3. inter-item analysis
4. constant analysis
Answer: 1

6. The process of quantifying data is referred to as ____
1. Topology
2. Diagramming
3. Enumeration
4. coding
Answer: 3

7. An advantage of using computer programs for qualitative data is that they _
1. Can reduce time required to analyse data (i.e., after the
data are transcribed)
2. Help in storing and organising data
3. Make many procedures available that are rarely done by hand due to time constraints
4. All of the above
Answer: 4

8 _____ are the basic building blocks of qualitative data.
1. Categories
2. Units
3. Individuals
4. None of the above

9. This is the process of transforming qualitative research data from written interviews or field notes into typed text.
1. Segmenting
2. Coding
3. Transcription
4. Mnemoning
Answer: 3

10. A graph that uses vertical bars to represent data is called a ___
1. Line graph
2. Bar graph
3. Scatterplot
4. Vertical graph
Answer: 2

11. ____ are used when you want to visually examine the relationship between two quantitative variables.
1. Bar graph
2. pie graph
3. line graph
4. Scatterplot
Answer: 4

12. The denominator (bottom) of the z-score formula is
1. The standard deviation
2. The difference between a score and the mean
3. The range
4. The mean
Answer: 1

13. Which of these distributions is used for a testing hypothesis?
1. Normal Distribution
2. Chi-Squared Distribution
3. Gamma Distribution
4. Poisson Distribution
Answer: 2

14. A statement made about a population for testing purpose is called?
1. Statistic
2. Hypothesis
3. Level of Significance
4. Test-Statistic
Answer: 2

15. If the assumed hypothesis is tested for rejection considering it to be true is called?
1. Null Hypothesis
2. Statistical Hypothesis
3. Simple Hypothesis
4. Composite Hypothesis
Answer: 1

16. If the null hypothesis is false then which of the following is accepted?
1. Null Hypothesis
2. Positive Hypothesis
3. Negative Hypothesis
4. Alternative Hypothesis.
Answer: 4

17. Alternative Hypothesis is also called as?
1. Composite hypothesis
2. Research Hypothesis
3. Simple Hypothesis
4. Null Hypothesis
Answer: 2

18. Data Analysis is a process of?
A. inspecting data
B. cleaning data
C. transforming data
D. All of the above
Answer: D

19. Which of the following is not a major data analysis approaches?
A. Data Mining
B. Predictive Intelligence
C. Business Intelligence
D. Text Analytics
Answer: B

20. How many main statistical methodologies are used in data analysis?
A. 2
B. 3
C. 4
D. 5
Answer: A

21. In descriptive statistics, data from the entire population or a sample is summarized with?
A. integer descriptors
B. floating descriptors
C. numerical descriptors
D. decimal descriptors
Answer: C

22. Data Analysis is defined by the statistician?
A. William S.
B. Hans Peter Luhn
C. Gregory Piatetsky-Shapiro
D. John Tukey
Answer: D

23. Which of the following is true about hypothesis testing?
A. answering yes/no questions about the data
B. estimating numerical characteristics of the data
C. describing associations within the data
D. modeling relationships within the data
Answer: A

24. The goal of business intelligence is to allow easy interpretation of large volumes of data to identify new opportunities.
A. TRUE
B. FALSE
C. Can be true or false
D. Can not say
Answer: A

25. The branch of statistics that deals with the development of particular statistical methods is classified as
A. industry statistics
B. economic statistics
C. applied statistics
D. mathematical statistics
Answer: D

26. Which of the following is true about regression analysis?
A. answering yes/no questions about the data
B. estimating numerical characteristics of the data
C. modeling relationships within the data
D. describing associations within the data
Answer: C

27. Text Analytics, also referred to as Text Mining?
A. TRUE
B. FALSE
C. Can be true or false
D. Can not say
Answer: A

28. What is the minimum no. of variables/ features required to perform clustering?
1. 0
2. 1
3. 2
4. 3
Answer: 2

29. For two runs of K-Mean clustering is it expected to get same clustering results?
1. Yes
2. No
Answer: 2

30. Which of the following algorithm is most sensitive to outliers?
1. K-means clustering algorithm
2. K-medians clustering algorithm
3. K-modes clustering algorithm
4. K-medoids clustering algorithm
Answer: 1

31. The discrete variables and continuous variables are two types of
1. Open end classification
2. Time series classification
3. Qualitative classification
4. Quantitative classification
Answer: 4

32. Bayesian classifiers is
1. A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.
2. Any mechanism employed by a learning system to constrain the search space of a hypothesis
3. An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.
4. None of these
Answer: 1

33. Classification accuracy is
1. A subdivision of a set of examples into a number of classes
2. Measure of the accuracy, of the classification of a concept that is given by a certain theory
3. The task of assigning a classification to a set of examples
4. None of these
Answer: 3

34. Euclidean distance measure is
1. A stage of the KDD process in which new data is added to the existing selection.
2. The process of finding a solution for a problem simply by enumerating all possible solutions according to some predefined order and then testing them
3. The distance between two points as calculated using the Pythagoras theorem
4. none of above
Answer: 3

35. Hybrid is
1. Combining different types of method or information
2. Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.
3. Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.
4. none of above
Answer: 1

36. Decision trees use _____ , in that they always choose the option that seems the best available at that moment.
1. Greedy Algorithms
2. divide and conquer
3. Backtracking
4. Shortest path algorithm
Answer: 1

37. Discovery is
1. It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).
2. The process of executing implicit previously unknown and potentially useful information from data
3. An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.
4. None of these
Answer: 2

38. Hidden knowledge referred to
1. A set of databases from different vendors, possibly using different database paradigms
2. An approach to a problem that is not guaranteed to work but performs well in most cases
3. Information that is hidden in a database and that cannot be recovered by a simple SQL query.
4. None of these
Answer: 3

39. Enrichment is
1. A stage of the KDD process in which new data is added to the existing selection
2. The process of finding a solution for a problem simply by enumerating all possible solutions according to some predefined order and then testing them
3. The distance between two points as calculated using the Pythagoras theorem.
4. None of these
Answer: 1

40.. _____ are easy to implement and can execute efficiently even without prior knowledge of the data, they are among the most popular algorithms for classifying text documents.
1. ID3
2. Naïve Bayes classifiers
3. CART
4. None of above
Answer: 2

41. High entropy means that the partitions in classification are
1. Pure
2. Not Pure
3. Usefull
4. useless
Answer: 2

42. Which of the following statements about Naive Bayes is incorrect?
1. Attributes are equally important.
2. Attributes are statistically dependent of one another given the class value.
3. Attributes are statistically independent of one another given the class value.
4. Attributes can be nominal or numeric
Answer: 2

43. The maximum value for entropy depends on the number of classes so if we have 8 Classes what will be the max entropy.
1. Max Entropy is 1
2. Max Entropy is 2
3. Max Entropy is 3
4. Max Entropy is 4
Answer: 3

44. Point out the wrong statement.
1. k-nearest neighbor is same as k-means
2. k-means clustering is a method of vector quantization
3. k-means clustering aims to partition n observations into k clusters
4. none of the mentioned
Answer: 1

45. Consider the following example “How we can divide set of articles such that those articles have the same theme (we do not
know the theme of the articles ahead of time) ” is this:
1. Clustering
2. Classification
3. Regression
4. None of these
Answer: 1

46. Clustering techniques are _____ in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters.
1. Unsupervised
2. supervised
3. Reinforcement
4. Neural network
Answer: 1

47. _____ metric is examined to determine a reasonably optimal value of k.
1. Mean Square Error
2. Within Sum of Squares (WSS)
3. Speed
4. None of these
Answer: 2

48. If an itemset is considered frequent, then any subset of the frequent itemset must also be frequent.
1. Apriori Property
2. Downward Closure Property
3. Either 1 or 2
4. Both 1 and 2
Answer: 4

49.{ bread,eggs,milk} has a support of 0.15 and {bread,eggs} also has a support of 0.15, the confidence of rule {bread,eggs}→{milk} is
1. 0
2. 1
3. 2
4. 3
Answer: 1

50. _____ recommend items based on similarity measures between users and/or items.
1. Content Based Systems
2. Hybrid System
3. Collaborative Filtering Systems
4. None of these
Answer: 3

51. There are _____ major Classification of Collaborative Filtering Mechanisms
1. 1
2. 2
3. 3
4. none of above
Answer: 2

52. Movie Recommendation to people is an example of
1. User Based Recommendation
2. Item Based Recommendation
3. Knowledge Based Recommendation
4. content based recommendation
Answer: 2

53. _____ recommenders rely on an explicitely defined set of recommendation rules
1. Constraint Based
2. Case Based
3. Content Based
4. User Based
Answer: 2

Practice Test: Question Set – 01

1. A statement made about a population for testing purpose is called?

- (A) Statistic
- (B) Hypothesis
- (C) Level of Significance
- (D) Test-Statistic

2. Data Analysis is a process of

- (A) Inspecting data
- (B) Cleaning data
- (C) Transforming data
- (D) All of Above

3. _____ have a structure but cannot be stored in a database.

- (A) Structured
- (B) Semi Structured
- (C) Unstructured
- (D) None of thes

4. HDFS Stores how much data in each clusters that can be scaled at any time?

- (A) 32
- (B) 64
- (C) 128
- (D) 256

5. The branch of statistics which deals with development of particular statistical methods is classified as

- (A) Industry statistics
- (B) Economic statistics
- (C) Applied statistics
- (D) Allied statistics

6. This is the process of transforming qualitative research data from written interviews or field notes into typed text.

- (A) Segmenting
- (B) Coding
- (C) Transcription
- (D) Mnemoning

7. If the null hypothesis is false then which of the following is accepted?

- (A) Null Hypothesis
- (B) Positive Hypothesis
- (C) Negative Hypothesis
- (D) Alternative Hypothesis

8. The Process of describing the data that is huge and complex to store and process is known as

- (A) Analytics
- (B) Data mining
- (C) Big data
- (D) Data warehouse

9. GFS consists of _____ Master and _____ Chunk Servers

- (A) Single, Single
- (B) Multiple, Single
- (C) Single, Multiple
- (D) Multiple, Multiple

10. _____ is factors considered before Adopting Big Data Technology

- (A) Validation
- (B) Verification
- (C) Data
- (D) Design

11. Which storage subsystem can support massive data volumes of increasing size.

- (A) Extensibility
- (B) Fault tolerance
- (C) Scalability
- (D) High-speed I/O capacity

12. _____ takes the grouped key-value paired data as input and runs a Reducer function on each one of them.

- (A) MAPPER
- (B) REDUCER
- (C) COMBINER
- (D) PARTITIONER

13. Movie Recommendation systems are an example of

1. Classification
2. Clustering
3. Reinforcement Learning
4. Regression

- (A) 2 only
- (B) 1 and 3
- (C) 1 and 2
- (D) 2 and 3

14. Which of the following is true about regression analysis?

- (A) Answering yes/no questions about the data
- (B) Estimating numerical characteristics of the data
- (C) Modeling relationships within the data
- (D) Describing associations within the data

15. An advantage of using computer programs for qualitative data is that they _____

- (A) Can reduce time required to analyze data (i.e., after the data are transcribed)
- (B) Help in storing and organizing data
- (C) Make many procedures available that are rarely done by hand due to time constraints
- (D) All of the above

Practice Test: Question Set - 02

1. The denominator (bottom) of the z-score formula is

- (A) The standard deviation
- (B) The difference between a score and the mean
- (C) The range
- (D) The mean

2. Which of the following is not an example of Social Media?

- (A) Twitter
- (B) Google
- (C) Instagram
- (D) Youtube

3. How many main statistical methodologies are used in data analysis?

- (A) 2
- (B) 3
- (C) 4
- (D) 5

4. Files are divided into _____ sized Chunks.

- (A) Static
- (B) Dynamic
- (C) Fixed
- (D) Variable

5. For improving supply chain management to optimize stock management, replenishment, and forecasting

- (A) Descriptive
- (B) Diagnostic
- (C) Predictive
- (D) Prescriptive

6. _____ provides performance through distribution of data and fault tolerance through replication

- (A) HDFS
- (B) PIG
- (C) HIVE
- (D) HADOOP

7. Sentiment Analysis is an example of

1. Regression
2. Classification
3. Clustering
4. Reinforcement Learning

- (A) 1, 2 and 4
- (B) 1, 2 and 3
- (C) 1 and 3
- (D) 1 and 2

8. Text Analytics, also referred to as Text Mining?

- (A) True
- (B) False
- (C) Can be true or false
- (D) Cannot say

9. The process of quantifying data is referred to as _____

- (A) Topology
- (B) Diagramming
- (C) Enumeration
- (D) Coding

10. If the assumed hypothesis is tested for rejection considering it to be true is called?

- (A) Null Hypothesis
- (B) Statistical Hypothesis
- (C) Simple Hypothesis
- (D) Composite Hypothesis

11. In descriptive statistics, data from the entire population or a sample is summarized with?

- (A) Integer descriptor
- (B) Floating descriptor
- (C) Numerical descriptor
- (D) Decimal descriptor

12. _____ phase sorts the data & _____ creates logical clusters.

- (A) Reduce, YARN
- (B) MAP, YARN
- (C) REDUCE, MAP
- (D) MAP, REDUCE

13. As an example, an expectation of using a recommendation engine would be to increase same-customer sales by adding more items into the market basket

- (A) Lowering costs
- (B) Increasing revenues
- (C) Increasing productivity
- (D) Reducing risk

14. What is a hypothesis?

- (A) A statement that the researcher wants to test through the data collected in a study
- (B) A research question the results will answer
- (C) A theory that underpins the study
- (D) A statistical method for calculating the extent to which the results could have happened by chance

15. By 2025, the volume of digital data will increase to

- (A) TB
- (B) YB
- (C) ZB
- (D) EB

Practice Test: Question Set - 03

1. _____ as a result of data accessibility, data latency, data availability, or limits on bandwidth in relation to the size of inputs

- (A) Computation-restricted throttling
- (B) Large data volumes
- (C) Data throttling
- (D) Data Parallelization

2. What is the cyclical process of collecting and analyzing data during a single research study called?

- (A) Interim Analysis
- (B) Inter analysis
- (C) Inter item analysis
- (D) Constant analysis

3. _____ refers to the ability to turn your data useful for business

- (A) Velocity
- (B) Variety
- (C) Value
- (D) Volume

4. _____ are the basic building blocks of qualitative data.

- (A) Categories
- (B) Units
- (C) Individuals
- (D) None of the above

5. _____ are used when you want to visually examine the relationship between two quantitative variables.

- (A) Bar graph
- (B) Pie graph
- (C) Line graph
- (D) Scatterplot

6. Which of these distributions is used for a testing hypothesis?

- (A) Normal Distribution
- (B) Chi-Squared Distribution
- (C) Gamma Distribution
- (D) Poisson Distribution

7. Alternative Hypothesis is also called as?

- (A) Composite hypothesis
- (B) Research Hypothesis
- (C) Simple Hypothesis
- (D) Null Hypothesis

8. Which of the following is not a major data analysis approaches?

- (A) Data Mining
- (B) Predictive Intelligence
- (C) Business Intelligence
- (D) Text Analytics

9. _____ is a type of local Reducer that groups similar data from the map phase into identifiable sets.

- (A) MAPPER
- (B) REDUCER
- (C) COMBINER
- (D) PARTITIONER

10. Data Analysis is defined by the statistician?

- (A) William S.
- (B) Hans Peter Luhn
- (C) Gregory Piatetsky-Shapiro
- (D) John Tukey

11. _____ is a programming model for writing applications that can process Big Data in parallel on multiple nodes.

- (A) HDFS
- (B) MAP REDUCE
- (C) HADOOP
- (D) HIVE

12. Which of the following is true about hypothesis testing?

- (A) Answering yes/no questions about the data
- (B) Estimating numerical characteristics of the data
- (C) Describing associations within the data
- (D) Modeling relationships within the data

13. A graph that uses vertical bars to represent data is called a _____

- (A) Line graph
- (B) Bar graph
- (C) Scatterplot
- (D) Vertical graph

14. Which among the following is not a Data mining and analytical applications?

- (A) Profile matching
- (B) Social network analysis
- (C) Facial recognition
- (D) Filtering

15. _____ is an open source framework for storing data and running application on clusters of commodity hardware.

- (A) HDFS
- (B) Hadoop
- (C) MapReduce
- (D) Cloud

16. While Installing Hadoop how many xml files are edited and list them?

- (A) core-site.xml
- (B) hdfs-site.xml
- (C) mapred.xml
- (D) yarn.xml

1.Which of the following is not a major data analysis approach?

1. Business Intelligence
2. Predictive Intelligence
3. Data Mining
4. Text Analytics

Answer: Predictive Intelligence

2. _____ is the cyclical process of collecting and analyzing data during a research study.

1. Constant analysis
2. Extremis Analysis
3. Interim Analysis
4. All of the above

Answer: Interim Analysis

3._____ are the basic building blocks of qualitative data.

1. Data chunk
2. Categories
3. Numeric figures
4. None of the above

Answer: Categories

4.The Process of describing the data that is huge and complex to store and process is known as _____

1. Analytics mining
2. Big data
3. Data cleaning
4. None of the above
5.

Answer: Big data

5.In descriptive statistics, data from the entire population or a sample is summarized with ___.

1. Decimal descriptor
2. Numerical descriptor
3. Integer descriptor
4. All of the above

Answer: Numerical descriptor

6. Data Analysis is a process of?

1. inspecting data
2. transforming data
3. cleaning data
4. All of the above

Answer: All of the above

7. A good data analytics solution includes a viable self-service _____.

1. Data warehouse
2. Data mining
3. Data wrangling
4. None of the above

Answer: Data wrangling

8. To glean insights from the data, many analysts and data scientists rely on _____.

1. Data warehouse
2. Data visualization
3. Data mining
4. All of the above

Answer: Data visualization

9. _____ refers to the ability to turn your data useful for business.

1. Velocity
2. Variety
3. Value
4. None of the above

Answer: Value

10. Correlation is the relationship between two variables_____

1. Two
2. One
3. Zero
4. None

Answer: Two

11.In descriptive statistics, data from the entire population or a sample is summarized with ?

1. numerical descriptors
2. integer descriptors
3. integer descriptors
4. decimal descriptors

Answer: numerical descriptors

12.How many main statistical methodologies are used in data analysis?

1. 3
2. 2
3. 4
4. 5

Answer: 2

13. The branch of statistics which deals with development of particular statistical methods is classified as

1. economic statistics
2. industry statistics
3. applied statistics
4. None

Answer: applied statistics

14. By 2025, the volume of data will increase to_____

1. TB
2. YB
3. ZB
4. EB

Answer: ZB

15. Alternative Hypothesis is also called as?

1. Simple Hypothesis
2. Research Hypothesis
3. Null Hypothesis
4. None of the mentioned above

Answer: Research Hypothesis

16.Data Analysis is defined by the statistician?

1. Hans Peter Luhn
2. William S.
3. Gregory Piatetsky-Shapiro
4. John Tukey

Answer: John Tukey

17. Amongst which of the following is / are not a major data analysis approach?

1. Business Intelligence
2. Text Analytics
3. Predictive Intelligence
4. Data Mining

Answer: Predictive Intelligence

18. For each value of the _____ the distribution of the dependent variable must be normal.

1. Intermediate variable
2. Depended variable
3. Independent variable
4. None of the above

Answer: Independent variable

19. Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate ___.

1. Interpretation
2. Predictions
3. Conclusion
4. None of the above

Answer: Predictions

20. Data Analysis is a process of_____

1. Data Cleaning
2. Transforming of data
3. Inspecting data
4. All of the above

Answer: All of the above

21.Amongst which of the following is / are the applications of Linear Regression_____

1. Behavioral
2. Social sciences
3. Biological
4. All of the above

Answer: All of the above

22. Least Square Method uses _____

1. Linear sequence
2. Linear regression
3. Linear polynomial
4. None of the above

Answer: Linear regression

23._____ are used when we want to visually examine the relationship between two quantitative variables.

1. Scatterplot
2. Bar graph
3. Line graph
4. Pie chart

Answer: Bar graph

24. A Linear Regression model's main aim is to find the best fit linear line and the ___ of intercept and coefficients such that the error is minimized.

1. Optimal values
2. Linear polynomial
3. Linear line
4. None of the mentioned above

Answer: Optimal values

25.The process of quantifying data is referred to as _____

1. Decoding
2. Enumeration
3. Structure
4. Coding

Answer: Enumeration

26. Amongst which of the following is / are the branch of statistics which deals with the development of statistical methods is classified as_____

1. Economic statistics
2. Industry statistics
3. Applied statistics
4. None of the mentioned above

Answer: Applied statistics

27. Amongst which of the following is / are the true about regression analysis?

1. Describes associations within the data
2. Answering yes/no questions about the data
3. Modeling relationships within the data
4. All of the above

Answer: Modeling relationships within the data

28. Linear Regression is the supervised machine learning model in which the model finds the best fit ___ between the independent and dependent variable.

1. Nonlinear line
2. Linear line
3. Curved line
4. None of the mentioned above

Answer: Linear line

29. Data Analytics uses ___ to get insights from data

1. Statistical methods
2. Statistical figures
3. Numerical aspects
4. None of the mentioned above

Answer: Statistical methods

30. Amongst which of the following is / are the types of Linear Regression_____

1. Simple Linear Regression
2. Multiple Linear Regression
3. Both of above
4. None of above

Answer: Both of above