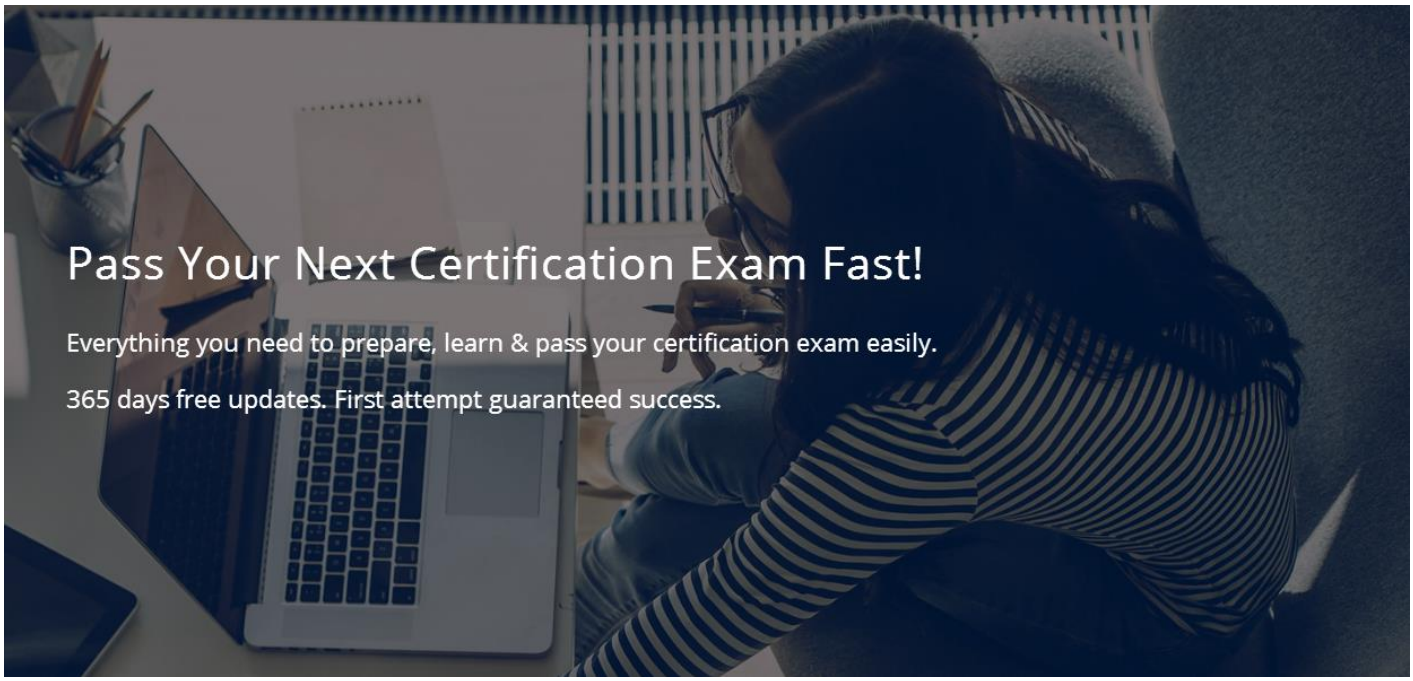
UpdateDumps



Choose the version that fits your needs	PDF Version	Desktop Test Engine	Online Test Engine
Latest and Up-to-Date exam dumps with real exam questions answers.	✓	✓	✓
Get 12-Months free updates without any extra charges.	✓	✓	✓
Experience same exam environment before appearing in the certification exam.	✗	✓	✓
100% exam passing guarante in the first attempt.	✓	✓	✓
20% discount on more than one license and 30% discount on 5+ license purchases.	✗	✓	✓
100% secure purchase on SSL.	✓	✓	✓
Completely private purchase without sharing your personal info with anyone.	✓	✓	✓

Exam : **D-DS-FN-23**

Title : Dell Data Science Foundations

Vendor : EMC

Version : DEMO

NO.1 What action occurs during feature selection in the model building phase of the data analytics lifecycle?

A. Create new combinations of attributes

B. Overfit the model to improve prediction accuracy

C. Identify the most useful input variables

D. Select a superset of variables to shorten training times

Answer: C

Explanation:

Feature selection involves identifying the most useful input variables that contribute meaningfully to the model, improving accuracy and reducing complexity during the model building phase.

NO.2 Consider this SQL statement: SELECT product, avg(prod_cost) FROM product_detail GROUP BY product.

The GROUP BY clause implies what type of function?

A. System function

B. Aggregate function

C. User defined function

D. Window function

Answer: B

Explanation:

The GROUP BY clause in the SQL statement implies the use of an aggregate function. In this case, avg (prod_cost) is an aggregate function that calculates the average of the product costs for each group of products.

NO.3 What is the primary role of a business intelligence analyst on an analytics project?

A. Extracts business data from source systems

B. Ensures business milestones are met

C. Provides business-domain expertise

D. Defines business goals for the analytics project

Answer: C

Explanation:

The primary role of a business intelligence analyst on an analytics project is to provide business-domain expertise. They bridge the gap between the technical aspects of the project and the business requirements, ensuring that the data and analysis align with the business objectives.

NO.4 When building a K-means clustering model, you notice that the clusters did not segment on variables that you expected. What should you do?

A. Decrease the value of K

B. Multiply each variable by its standard deviation

C. Add the WSS to each variable

D. Check that the data was properly scaled

Answer: D

Explanation:

When building a K-means clustering model, it's important to ensure that the data is properly scaled,

as K- means is sensitive to the scale of the variables. If the variables have different scales, the clustering may not segment as expected. Standardizing or normalizing the data can often improve the results.

NO.5 Which visualization technique should be avoided?
A. Using a small number of contrasting colors to draw distinctions
B. Using tables of numbers to present all of the data visually
C. Achieving a high data-ink ratio
D. Using visuals to illustrate key points
Answer: B
Explanation:
Using tables of numbers to present all of the data visually should be avoided, as it can overwhelm the audience and make it harder to interpret key insights. Instead, visualizations should simplify data and focus on illustrating trends or patterns effectively.

NO.6 Which chart type is intended to display time series data?
A. Bar chart
B. Pie chart
C. Line chart
D. [Histogram
Answer: C
Explanation:
A line chart is specifically designed to display time series data. It shows data points in a sequential order, making it easy to observe trends over time.

NO.7 In K-means clustering, what is a graph of the WSS versus the value of K used to help determine?
A. Optimal distance between clusters
B. Average distance between observations
C. 'Optimal number of clusters
D. Average distance between clusters
Answer: C
Explanation:
A graph of the WSS (Within-Cluster Sum of Squares) versus the value of K helps determine the optimal number of clusters. The "elbow" point on the graph represents the value of K at which adding more clusters no longer significantly improves the WSS.

NO.8 What type of variable is the dependent variable from a logistic regression?
A. Categorical
B. Continuous
C. Ratio
D. Interval
Answer: A
Explanation:

The dependent variable in logistic regression is categorical, typically binary (e.g., success/failure, yes/no), since the goal is to model the probability of an event occurring.

NO.9 What are the two data categories that represent qualitative data?
A. Ordinal and interval
B. Nominal and ordinal
C. Ratio and interval
D. Nominal and ratio
Answer: B
Explanation:
Qualitative data represents non-numeric categories. Nominal data consists of named categories without any order, while ordinal data includes categories with a meaningful order but without consistent intervals between them.

NO.10 On which type of data should you run K-means clustering?
A. Ordinal
B. Numeric
C. Text
D. Nominal
Answer: B
Explanation:
K-means clustering is best suited for numeric data, as it relies on calculating the Euclidean distance between data points, which is meaningful for numerical values.

NO.11 In hypothesis testing, when does a Type I error occur?
A. Null hypothesis is rejected when it is actually false
B. Null hypothesis is rejected when it is actually true
C. Null hypothesis is accepted when it is actually false
D. Null hypothesis is accepted when it is actually true
Answer: B
Explanation:
A Type I error occurs when the null hypothesis is rejected even though it is actually true. This is also known as a "false positive" in hypothesis testing.

NO.12 After running a density plot you realize that the data has a long tail to the right. What can you do to make the dataset more normally distributed?
A. Use a scatter plot to obtain a better picture
B. Use a histogram to obtain a better picture
C. Apply a square transformation
D. Apply a logarithmic transformation
Answer: D
Explanation:
A logarithmic transformation is commonly used to reduce right skewness (long tail to the right) and make data more normally distributed.

NO.13 When should you consider using multinomial logistic regression over binary logistic regression?

A. Dependent variable is continuous or dichotomous

B. Dependent variable is continuous or categorical

C. Dependent variable has more than two categories

D. Dependent variable is continuous only

Answer: C

Explanation:

Multinomial logistic regression should be used when the dependent variable has more than two categories (i.
e., it is categorical with more than two possible outcomes). This differs from binary logistic regression, which is used when the dependent variable is binary (i.e., has two categories).

NO.14 What is a business driver for Big Data analytics adoption?

A. Implement the latest technology and tools

B. Maintain existing data silos

C. Identify new business opportunities

D. Ensure the analysts work in isolation

Answer: C

Explanation:

A key business driver for Big Data analytics adoption is the ability to identify new business opportunities by analyzing large and complex datasets, which can provide valuable insights for decision-making and strategy development.

NO.15 What converts SQL-like commands into either Tez, Spark, or MapReduce jobs that are submitted to the Hadoop cluster?

A. Pig

B. HBase

C. Hive

D. Mahout

Answer: C

Explanation:

Hive converts SQL-like commands into execution plans that run as Tez, Spark, or MapReduce jobs on a Hadoop cluster, enabling users to query large datasets using a familiar query language.

NO.16 Which phase of the data analytic lifecycle includes conducting project sponsor interviews and drafting a problem statement?

A. Operationalize

B. Model planning

C. Model building

D. Discovery

Answer: D

Explanation:

The Discovery phase involves understanding the business problem by conducting project sponsor interviews, gathering requirements, and drafting a problem statement to define the analytic objectives clearly.

NO.17 In the data preparation phase of the data analytics lifecycle, what does the term "data conditioning" refer to?

A. Building training and testing datasets

B. Identifying relationships and correlations among variables

C. Deploying the model and monitoring its performance

D. Cleaning the data, normalizing datasets. and performing transformations

Answer: D

Explanation:

Data conditioning in the data preparation phase refers to the process of cleaning the data, normalizing datasets, and performing transformations to ensure the data is in a suitable format for analysis or modeling.

NO.18 What is the similarity between the matrix and array data structures in R?

A. Both structures can contain only integers

B. Both structures can only contain one data type

C. Both structures can store multiple data types

D. Both structures must be 2-dimensional

Answer: B

Explanation:

Both matrix and array data structures in R can only contain one data type across all their elements, ensuring consistency in the structure.

NO.19 You have the data from a popular e-commerce website. You are exploring the time spent (in seconds) on the website by 100,000 customers across 14 different product categories.
What visualization can be used to represent the relationship between time spent and product category?

A. Rug plot

B. Scatter plot

C. Box and whisker plot

D. Hexbin plot

Answer: C

Explanation:

A box and whisker plot is ideal for visualizing the relationship between time spent and product category, especially when you have multiple categories. It shows the distribution of time spent in each product category, including the median, quartiles, and any potential outliers.

NO.20 Which SQL set operator returns rows that exist in the first SELECT statement answer set but not in the second SELECT statement?

A. EXCEPT

B. UNION

C. UNION ALL
D. INTERSECT
Answer: A
Explanation:
The EXCEPT operator returns rows that exist in the result set of the first SELECT statement but not in the second. It removes duplicates and only shows the difference between the two datasets.

NO.21 What data asset is an example of quasi-structured data?
A. Excel file
B. Clickstream data
C. Relational database table
D. Comma-separated value file
Answer: A
Explanation:
An Excel file is an example of quasi-structured data because it contains structured data (like rows and columns) but may have irregularities, such as varying data formats, missing headers, or mixed data types within cells, making it less strictly structured than a relational database.

NO.22 Which Hadoop service responds to requests for compute and memory resources?
A. Application Manager
B. DataNode
C. Scheduler
D. Application Master
Answer: C
Explanation:
The Scheduler in Hadoop is responsible for allocating compute and memory resources across various applications running on the cluster. It decides how resources are distributed based on policies and availability.

NO.23 A logistic regression model is built to determine the probability of a credit card borrower defaulting on a credit loan. A threshold value of 0.3 is selected. Which statement can be used to predict a borrower will default?
A. If probability > 0.1, then predict the borrower will default
B. If probability < 0.1, then predict the borrower will default
C. If probability > 0.3, then predict the borrower will default
D. If probability < 0.3, then predict the borrower will default
Answer: C
Explanation:
In logistic regression, the threshold determines the cutoff for classifying an outcome. If the probability exceeds the threshold (0.3 in this case), the model predicts the positive class-here, that the borrower will default.

NO.24 In which programming language is Hadoop written?
A. C++

B. Scala
C. Java
D. Python
Answer: C
Explanation:
Hadoop is primarily written in Java. It is built on the Java programming language to provide a framework for distributed storage and processing of large datasets.

NO.25 What are three built-in data types in the R programming language?
A. Boolean, integer, and character
B. Boolean, table, and character
C. Boolean, table, and integer
D. List, array, and integer
Answer: A
Explanation:
Three built-in data types in R are Boolean, integer, and character. These data types are fundamental to R and represent logical values (TRUE/FALSE), whole numbers, and text data, respectively.

NO.26

	Meat	Bread	Eggs	Total
Cookies	1	100	40	141
Milk	3	120	110	233
Total	4	220	150	

Refer to the exhibit, which shows pairwise counts for items purchased together.
Consider the following association rule: Milk -> Eggs
What is value of the lift?
A. 1.18
B. 0.264
C. 120
D. 70.81
Answer: A
Explanation:

$$\text{Lift}(Milk \rightarrow Eggs) = \frac{P(Milk \cap Eggs)}{P(Milk) \times P(Eggs)}$$

P(Milk#Eggs) is the probability of both Milk and Eggs being bought together, which is the count for Milk and Eggs (110).
P(Milk)) is the probability of Milk being bought, which is the total for Milk (233) divided by the grand total (374).
P(Eggs) is the probability of Eggs being bought, which is the total for Eggs (150) divided by the grand total (374).

$$P(Milk \cap Eggs) = \frac{110}{374}$$

$$P(Milk) = \frac{233}{374}$$

$$P(Eggs) = \frac{150}{374}$$

$$\text{Lift} = \frac{\frac{110}{374}}{\frac{233}{374} \times \frac{150}{374}} = \frac{110}{233 \times 150/374}$$

The value of the lift for the association rule Milk -> Eggs is approximately 1.18.

NO.27 What is a benefit of Spark in-memory data processing as opposed to using MapReduce?

A. Avoids writing intermediate data to disk, which speeds up processing

B. Supports processing unstructured data, which MapReduce does not allow

C. Removes the need to use disks at all, which reduces cost

D. Allows parallel processing, which MapReduce does not support

Answer: A

Explanation:

Spark's in-memory data processing avoids the need to write intermediate data to disk, which significantly speeds up processing compared to MapReduce, which writes intermediate data to disk after each operation.

This makes Spark much faster for iterative tasks.

NO.28 In addition to quantitative and technical skills, what is a key aspect of the profile of a data scientist?

A. Project management and administrative skills

B. Proficient in Microsoft Project and Excel

C. Skeptical and critical thinking

D. Accounting and regulatory skills

Answer: C

Explanation:

A key aspect of a data scientist's profile is skeptical and critical thinking, which helps in questioning assumptions, validating data, and interpreting results accurately to ensure sound analysis and decision-making.

NO.29 In time series analysis, what function is examined to identify the order of the autoregressive component of an ARIMA model?

A. Logistic function

B. Lognormal distribution function

C. Partial autocorrelation function

D. Normal distribution function

Answer: C

Explanation:

The Partial Autocorrelation Function (PACF) is examined to identify the order of the autoregressive (AR) component in an ARIMA model. It helps determine how much of the autocorrelation is explained by previous lags, assisting in selecting the appropriate AR order.

NO.30 How should project results be communicated to executives and the project sponsor?

A. Focus on business outcomes and benefits

B. Demonstrate your technical prowess to establish credibility

C. Provide model performance visualizations

D. Emphasize coding details and technical requirements

Answer: A

Explanation:

When communicating project results to executives and the project sponsor, it's important to focus on business outcomes and benefits, ensuring they understand the value the project brings to the organization rather than the technical details.

NO.31 What are categorized as cluster and workflow management tools for Hadoop?

A. Flume, Sqoop, and Storm

B. Drill, Hive, and HBase

C. Spark, Tez, and Cassandra

D. Ambari, Oozie, and Zookeeper

Answer: D

Explanation:

Ambari, Oozie, and Zookeeper are tools used for cluster and workflow management in Hadoop. Ambari manages and monitors clusters, Oozie handles workflow scheduling, and Zookeeper coordinates distributed processes.

NO.32 In a user-defined aggregate function, what is FFUNC?

A. Optional final calculation function

B. Window function

C. State transition function

D. Segment-level calculation function

Answer: A

Explanation:

In a user-defined aggregate function (UDAF), FFUNC refers to the optional final calculation function. It is used to perform any final calculation or transformation on the aggregated result before it is returned to the user.

NO.33 Which R function plots a distribution of a single variable along two different axes?

A. table()

B. summaryQ

C. density ()

D. rug()

Answer: D

Explanation:

The rug() function in R adds small tick marks along the axes of a plot to represent the distribution of a variable, effectively showing it along both axes for better visual interpretation.

NO.34 In ANOVA, what is the null hypothesis for k population means?

A. All population means are equal to each other

B. At least two population means are equal

C. At least two population means are not equal

D. At most k-1 population means are equal

Answer: A

Explanation:

In ANOVA (Analysis of Variance), the null hypothesis is that all population means are equal to each other.

The goal is to test whether there is evidence to reject this hypothesis, indicating that at least one population mean differs from the others.

NO.35 What does "MAD" in MADlib stand for?

A. Magnetic Association Design

B. Magnetic Agile Deep

C. Multiple Agile Development

D. Multiple Access Design

Answer: D

Explanation:

MAD in MADlib stands for Multiple Access Design, which is a library for scalable in-database analytics, including machine learning and statistical methods.