Vendor: EMC

Exam Code: E20-007

Exam Name: Data Science and Big Data Analytics

Version: Demo

QUESTION 1
You are using MADlib for Linear Regression analysis. Which value does the statement return?

SELECT (linregr(depvar, indepvar)).r2 FROM zeta1;

A. Goodness of fit
B. Coefficients
C. Standard error
D. P-value

Correct Answer: A

QUESTION 2
Which data asset is an example of quasi-structured data?

A. Webserver log
B. XML data file
C. Database table
D. News article

Correct Answer: A

QUESTION 3
What would be considered "Big Data"?

A. An OLAP Cube containing customer demographic information about 100, 000, 000 customers
B. Daily Log files from a web server that receives 100, 000 hits per minute
C. Aggregated statistical data stored in a relational database table
D. Spreadsheets containing monthly sales data for a Global 100 corporation

Correct Answer: B

QUESTION 4
A data scientist plans to classify the sentiment polarity of 10, 000 product reviews collected from the Internet. What is the most appropriate model to use? Suppose labeled training data is available.

A. Naïve Bayesian classifier
B. Linear regression
C. Logistic regression
D. K-means clustering

Correct Answer: A

QUESTION 5
In which lifecycle stage are test and training data sets created?

A. Model building
B. Model planning
C. Discovery
D. Data preparation

Correct Answer: A

QUESTION 6
When creating a presentation for a technical audience, what is the main objective?

A. Show that you met the project goals
B. Show how you met the project goals
C. Show if the model will meet the SLA
D. Show the technique to be used in the production environment

Correct Answer: B

QUESTION 7
Your company has 3 different sales teams. Each team's sales manager has developed incentive offers to increase the size of each sales transaction. Any sales manager whose incentive program can be shown to increase the size of the average sales transaction will receive a bonus.

Data are available for the number and average sale amount for transactions offering one of the incentives as well as transactions offering no incentive.

The VP of Sales has asked you to determine analytically if any of the incentive programs has resulted in a demonstrable increase in the average sale amount. Which analytical technique would be appropriate in this situation?

A. One-way ANOVA
B. Multi-way ANOVA
C. Student's t-test
D. Wilcoxson Rank Sum Test

Correct Answer: A

QUESTION 8
In data visualization, what is used to focus the audience on a key part of a chart?

A. Emphasis colors
B. Detailed text
C. Pastel colors
D. A data table

Correct Answer: A

QUESTION 9
Which word or phrase completes the statement? Data-ink ratio is to data visualization as _____ .

A. Confusion matrix is to classifier
B. Data scientist is to big data
C. Seasonality is to ARIMA
D. K-means is to Naive Bayes

Correct Answer: A

QUESTION 10
Consider a database with 4 transactions:
Transaction 1: {cheese, bread, milk}
Transaction 2: {soda, bread, milk}
Transaction 3: {cheese, bread}
Transaction 4: {cheese, soda, juice}

You decide to run the association rules algorithm where minimum support is 50%. Which rule has a confidence at least 50%?

A. {cheese} => {bread}
B. {juice} => {cheese}

C. {milk} => {soda}
D. {soda} => {milk}

Correct Answer: A

QUESTION 11
You are using the Apriori algorithm to determine the likelihood that a person who owns a home has a good credit score. You have determined that the confidence for the rules used in the algorithm is > 75%. You calculate lift = 1.011 for the rule, "People with good credit are homeowners". What can you determine from the lift calculation?

A. Support for the association is low
B. Leverage of the rules is low
C. The rule is coincidental
D. The rule is true

Correct Answer: C

QUESTION 12
Consider a database with 4 transactions:
Transaction 1: {cheese, bread, milk}
Transaction 2: {soda, bread, milk}
Transaction 3: {cheese, bread}
Transaction 4: {cheese, soda, juice}

The minimum support is 25%. Which rule has a confidence equal to 50%?

A. {bread, milk} => {cheese}
B. {bread} => {milk}
C. {juice} => {soda}
D. {bread} => {cheese}

Correct Answer: A

QUESTION 13
Under which circumstance do you need to implement N-fold cross-validation after creating a regression model?

A. There is not enough data to create a test set.
B. The data is unformatted.
C. There are missing values in the data.
D. There are categorical variables in the model.

Correct Answer: A

QUESTION 14
What is an appropriate data visualization to use in a presentation for an analyst audience?

A. Pie chart
B. Area chart
C. Stacked bar chart
D. ROC curve

Correct Answer: D

QUESTION 15
When would you use GROUP BY ROLLUP clause in your OLAP query?

A. where all subtotals and grand totals are to be included in the output
B. where only the subtotals are to be included in the output
C. where only the grand totals are to be included in the output
D. where only specific subtotals and grand totals for a combination of variables are to be included in the output

Correct Answer: A

QUESTION 16
Which type of numeric value does a logistic regression model estimate?

A. Probability
B. A p-value
C. Any integer
D. Any real number

Correct Answer: A

QUESTION 17
Your colleague, who is new to Hadoop, approaches you with a question. They want to know how best to access their data. This colleague has a strong background in data flow languages and programming.

Which query interface would you recommend?

A. Pig
B. Hive
C. Howl
D. HBase

Correct Answer: A

QUESTION 18
The web analytics team uses Hadoop to process access logs. They now want to correlate this data with structured user data residing in a production single-instance JDBC database. They collaborate with the production team to import the data into Hadoop. Which tool should they use?

A. Sqoop
B. Pig
C. Chukwa
D. Scribe

Correct Answer: A

QUESTION 19
What does the R code
z <- f[1:10,]
do?

A. Assigns the first 10 rows of f to the vector z
B. Assigns the 1st 10 columns of the 1st row of f to z
C. Assigns a sequence of values from 1 to 10 to z
D. Assigns the 1st 10 columns to z

Correct Answer: A

QUESTION 20
In R, functions like plot() and hist() are known as what?

A. generic functions
B. virtual methods
C. virtual functions
D. generic methods

Correct Answer: B

QUESTION 21
Review the following code:

```
SELECT pn, vn, sum(prc*qty)
FROM sale
GROUP BY CUBE(pn, vn)
ORDER BY 1, 2, 3;
```

Which combination of subtotals do you expect to be returned by the query?

A. (pn, vn)
B. ((pn, vn), (pn))
C. ((pn, vn) , (pn), (vn))
D. ((pn, vn) , (pn), (vn) , ( ))

Correct Answer: D

QUESTION 22
In MADlib what does MAD stand for?

A. Magnetic, Agile, Deep
B. Machine Learning, Algorithms for Databases
C. Mathematical Algorithms for Databases
D. Modular, Accurate, Dependable

Correct Answer: A

QUESTION 23
The web analytics team uses Hadoop to process access logs. They now want to correlate this data with structured user data residing in their massively parallel database. Which tool should they use to export the structured data from Hadoop?

A. Sqoop
B. Pig
C. Chukwa
D. Scribe

Correct Answer: A

QUESTION 24
When would you prefer a Naive Bayes model to a logistic regression model for classification?

A. When you are using several categorical input variables with over 1000 possible values each.
B. When you need to estimate the probability of an outcome, not just which class it is in.
C. When all the input variables are numerical.
D. When some of the input variables might be correlated.

Correct Answer: A

QUESTION 25
Before you build an ARMA model, how can you tell if your time series is weakly stationary?

A. There appears to be a constant variance around a constant mean.
B. The mean of the series is close to 0.
C. The series is normally distributed.
D. There appears to be no apparent trend component.

Correct Answer: A

QUESTION 26
What is an example of a null hypothesis?

A. that a newly created model does not provide better predictions than the currently existing model
B. that a newly created model provides a prediction of a null sample mean
C. that a newly created model provides a prediction of a null population mean
D. that a newly created model provides a prediction that will be well fit to the null distribution

Correct Answer: A

QUESTION 27
You have fit a decision tree classifier using 12 input variables. The resulting tree used 7 of the 12 variables, and is 5 levels deep. Some of the nodes contain only 3 data points. The AUC of the model is 0.85. What is your evaluation of this model?

A. The tree is probably overfit. Try fitting shallower trees and using an ensemble method.
B. The AUC is high, and the small nodes are all very pure. This is an accurate model.
C. The tree did not split on all the input variables. You need a larger data set to get a more accurate model.
D. The AUC is high, so the overall model is accurate. It is not well-calibrated, because the small nodes will give poor estimates of probability.

Correct Answer: A

QUESTION 28
If your intention is to show trends over time, which chart type is the most appropriate way to depict the data?

A. Line chart
B. Bar chart
C. Stacked bar chart
D. Histogram

Correct Answer: A

QUESTION 29
You are analyzing a time series and want to determine its stationarity. You also want to determine the order of autoregressive models.

How are the autocorrelation functions used?

A. ACF as an indication of stationarity, and PACF for the correlation between Xt and Xt-k not explained by their mutual correlation with X1 through Xk-1.
B. PACF as an indication of stationarity, and ACF for the correlation between Xt and Xt-k not explained by their mutual correlation with X1 through Xk-1.
C. ACF as an indication of stationarity, and PACF to determine the correlation of X1 through Xk-1.
D. PACF as an indication of stationarity, and ACF to determine the correlation of X1 through Xk-1.

Correct Answer: A

QUESTION 30
Which word or phrase completes the statement? A spreadsheet is to a data island as a centralized database for reporting is to a _____?

A. Data Warehouse
B. Data Repository
C. Analytic Sandbox
D. Data Mart

Correct Answer: A

QUESTION 31
What is one modeling or descriptive statistical function in MADlib that is typically not provided in a standard relational database?

A. Linear regression
B. Expected value
C. Variance
D. Quantiles

Correct Answer: A

QUESTION 32
In which phase of the data analytics lifecycle do Data Scientists spend the most time in a project?

A. Discovery
B. Data Preparation
C. Model Building
D. Communicate Results

Correct Answer: B

QUESTION 33
You are testing two new weight-gain formulas for puppies. The test gives the results:

Control group: 1% weight gain
Formula A. 3% weight gain
Formula B. 4% weight gain

A one-way ANOVA returns a p-value = 0.027
What can you conclude?

A. Either Formula A or Formula B is effective at promoting weight gain.
B. Formula B is more effective at promoting weight gain than Formula A.
C. Formula A and Formula B are both effective at promoting weight gain.
D. Formula A and Formula B are about equally effective at promoting weight gain.

Correct Answer: A

QUESTION 34
Data visualization is used in the final presentation of an analytics project. For what else is this technique commonly used?

A. Data exploration
B. Descriptive statistics
C. ETLT
D. Model selection

Correct Answer: A

QUESTION 35
Which functionality do regular expressions provide?

A. text pattern matching
B. underflow prevention
C. increased numerical precision
D. decreased processing complexity

Correct Answer: A

QUESTION 36
When creating a project sponsor presentation, what is the main objective?

A. Show that you met the project goals
B. Show how you met the project goals
C. Show how well the model will meet the SLA (service level agreement)
D. Clearly describe the methods and techniques used

Correct Answer: A

QUESTION 37
The average purchase size from your online sales site is $17, 200. The customer experience team believes a certain adjustment of the website will increase sales. A pilot study on a few hundred customers showed an increase in average purchase size of $1.47, with a significance level of p=0.1. The team runs a larger study, of a few thousand customers. The second study shows an increased average purchase size of $0.74, with a significance level of 0.03. What is your assessment of this study?

A. The change in purchase size is not practically important, and the good p-value of the second study is probably a result of the large study size.
B. The change in purchase size is small, but may aggregate up to a large increase in profits over the entire customer base.
C. The difference in the change in purchase size between the two studies is troubling; The team should run another, larger study.
D. The p-value of the second study shows a statistically significant change in purchase size. The new website is an improvement.

Correct Answer: A

QUESTION 38
Which word or phrase completes the statement? Business Intelligence is to monitoring trends as Data Science is to _____ trends.

A. Predicting
B. Discarding
C. Driving
D. Optimizing

Correct Answer: A

QUESTION 39
Consider a scale that has five (5) values that range from "not important" to "very important". Which data classification best describes this data?

A. Ordinal
B. Nominal
C. Real

D. Ratio

Correct Answer: A

QUESTION 40
Which key role for a successful analytic project can provide business domain expertise with a deep understanding of the data and key performance indicators?

A. Business Intelligence Analyst
B. Project Manager
C. Project Sponsor
D. Business User

Correct Answer: A

QUESTION 41
On analyzing your time series data you suspect that the data represented as

y1, y2, y3, ... , yn-1, yn

may have a trend component that is quadratic in nature. Which pattern of data will indicate that the trend in the time series data is quadratic in nature?

A. (y3-y2) (y2-y1) = ...= (yn-yn-1)-(yn-1-yn-2)
B. (y2-y1) = (y3-y2) = ... = (yn-yn-1)
C. ((y2-y1) /y1) * 100% = ...((yn-yn-1)/yn-1) * 100%
D. (y4-y2) (y3-y1) = ...= (yn-yn-2)-(yn-1-yn-3)

Correct Answer: A

QUESTION 42
Which analytical method is considered unsupervised?

A. K-means clustering
B. Naïve Bayesian classifier
C. Decision tree
D. Linear regression

Correct Answer: A

QUESTION 43
You have used k-means clustering to classify behavior of 100, 000 customers for a retail store. You decide to use household income, age, gender and yearly purchase amount as measures. You have chosen to use 8 clusters and notice that 2 clusters only have 3 customers assigned. What should you do?

A. Decrease the number of clusters
B. Increase the number of clusters
C. Decrease the number of measures used
D. Identify additional measures to add to the analysis

Correct Answer: A

QUESTION 44
What does R code nv <- v[v < 1000] do?

A. Selects the values in vector v that are less than 1000 and assigns them to the vector nv
B. Sets nv to TRUE or FALSE depending on whether all elements of vector v are less than 1000
C. Removes elements of vector v less than 1000 and assigns the elements >= 1000 to nv

D. Selects values of vector v less than 1000, modifies v, and makes a copy to nv

Correct Answer: A

QUESTION 45
For which class of problem is MapReduce most suitable?

A. Embarrassingly parallel
B. Minimal result data
C. Simple marginalization tasks
D. Non-overlapping queries

Correct Answer: A

QUESTION 46
Which activity is performed in the Operationalize phase of the Data Analytics Lifecycle?

A. Define the process to maintain the model
B. Try different analytical techniques
C. Try different variables
D. Transform existing variables

Correct Answer: A

QUESTION 47
Since R factors are categorical variables, they are most closely related to which data classification level?

A. nominal
B. ordinal
C. interval
D. ratio

Correct Answer: A

QUESTION 48
In which phase of the analytic lifecycle would you expect to spend most of the project time?

A. Discovery
B. Data preparation
C. Communicate Results
D. Operationalize

Correct Answer: B

QUESTION 49
You are building a logistic regression model to predict whether a tax filer will be audited within the next two years. Your training set population is 1000 filers. The audit rate in your training data is 4.2%. What is the sum of the probabilities that the model assigns to all the filers in your training set that have been audited?

A. 42.0
B. 4.2
C. 0.42
D. 0.042

Correct Answer: A

QUESTION 50
Refer to exhibit.

Independent Variable	Coefficient	P-Value
A	0.45	0.0000
B	3.67	0.0000
C	1.23	0.0000

$R^2 = 0.10$

You are asked to write a report on how specific variables impact your client's sales using a data set provided to you by the client. The data includes 15 variables that the client views as directly related to sales, and you are restricted to these variables only. After a preliminary analysis of the data, the following findings were made:

1. Multicollinearity is not an issue among the variables
2. Only three variables--A, B, and C--have significant correlation with sales

You build a linear regression model on the dependent variable of sales with the independent variables of A, B, and C. The results of the regression are seen in the exhibit.

You cannot request additional datA. what is a way that you could try to increase the R2 of the model without artificially inflating it?

A. Create clusters based on the data and use them as model inputs
B. Force all 15 variables into the model as independent variables
C. Create interaction variables based only on variables A, B, and C
D. Break variables A, B, and C into their own univariate models

Correct Answer: A

QUESTION 51
You have two tables of customers in your database. Customers in cust_table_1 were sent an e-mail promotion last year, and customers in cust_table_2 received a newsletter last year. Customers can only be entered in once per table. You want to create a table that includes all customers, and any of the communications they received last year. Which type of join would you use for this table?

A. Full outer join
B. Inner join
C. Left outer join
D. Cross join

Correct Answer: A

QUESTION 52
In which lifecycle stage are initial hypotheses formed?

A. Discovery
B. Model planning
C. Model building
D. Data preparation

Correct Answer: A

QUESTION 53
You are given 10, 000, 000 user profile pages of an online dating site in XML files, and they are stored in

HDFS. You are assigned to divide the users into groups based on the content of their profiles. You have been instructed to try K-means clustering on this data. How should you proceed?

A. Run MapReduce to transform the data, and find relevant key value pairs.
B. Divide the data into sets of 1, 000 user profiles, and run K-means clustering in RHadoop iteratively.
C. Run a Naive Bayes classification as a pre-processing step in HDFS.
D. Partition the data by XML file size, and run K-means clustering in each partition.

Correct Answer: A

QUESTION 54
The Marketing department of your company wishes to track opinion on a new product that was recently introduced. Marketing would like to know how many positive and negative reviews are appearing over a given period and potentially retrieve each review for more in-depth insight.

They have identified several popular product review blogs that historically have published thousands of user reviews of your company's products.

You have been asked to provide the desired analysis. You examine the RSS feeds for each blog and determine which fields are relevant. You then craft a regular expression to match your new product's name and extract the relevant text from each matching review.

What is the next step you should take?

A. Convert the extracted text into a suitable document representation and index into a review corpus
B. Use the extracted text and your regular expression to perform a sentiment analysis based on mentions of the new product
C. Read the extracted text for each review and manually tabulate the results
D. Group the reviews using Naïve Bayesian classification

Correct Answer: A

QUESTION 55
Which word or phrase completes the statement? A Data Scientist would consider that a RDBMS is to a Table as R is to a _____ .

A. Data frame
B. List
C. Matrix
D. Array

Correct Answer: A

QUESTION 56
Which word or phrase completes the statement? Unix is to bash as Hadoop is to:

A. Pig
B. HDFS
C. Sqoop
D. NameNode

Correct Answer: A

QUESTION 57
A call center for a large electronics company handles an average of 35, 000 support calls a day. The head of the call center would like to optimize the staffing of the call center during the rollout of a new product due to recent customer complaints of long wait times. You have been asked to create a model to optimize call center costs and customer wait times.

The goals for this project include:
1. Relative to the release of a product, how does the call volume change over time?
2. How to best optimize staffing based on the call volume for the newly released product, relative to old products.
3. Historically, what time of day does the call center need to be most heavily staffed?
4. Determine the frequency of calls by both product type and customer language.

Which goals are suitable to be completed with MapReduce?

A. Goal 2 and 4
B. Goal 1 and 3
C. Goals 1, 2, 3, 4
D. Goals 2, 3, 4

Correct Answer: A

QUESTION 58
Consider the example of an analysis for fraud detection on credit card usage. You will need to ensure higher-risk transactions that may indicate fraudulent credit card activity are retained in your data for analysis, and not dropped as outliers during pre-processing. What will be your approach for loading data into the analytical sandbox for this analysis?

A. ELT
B. ETL
C. EDW
D. OLTP

Correct Answer: A

QUESTION 59
Trend, seasonal, and cyclical are components of a time series. What is another component?

A. Irregular
B. Linear
C. Quadratic
D. Exponential

Correct Answer: A

QUESTION 60
You are studying the behavior of a population, and you are provided with multidimensional data at the individual level. You have identified four specific individuals who are valuable to your study, and would like to find all users who are most similar to each individual. Which algorithm is the most appropriate for this study?

A. K-means clustering
B. Linear regression
C. Association rules
D. Decision trees

Correct Answer: A

QUESTION 61
Which R data structure allows elements to have different data types?

A. List
B. Vector
C. Matrix

D. Array

Correct Answer: A

QUESTION 62
Which key role for a successful analytic project can consult and advise the project team on the value of end results and how these will be used on a day-to-day basis?

A. Business User
B. Project Manager
C. Data Scientist
D. Business Intelligence Analyst

Correct Answer: A

QUESTION 63
A disk drive manufacturer has a defect rate of less than 1.0% with 98% confidence. A quality assurance team samples 1000 disk drives and finds 14 defective units. Which action should the team recommend?

A. The manufacturing process should be inspected for problems.
B. A larger sample size should be taken to determine if the plant is functioning properly
C. A smaller sample size should be taken to determine if the plant is functioning properly
D. The manufacturing process is functioning properly and no further action is required.

Correct Answer: A

QUESTION 64
What is required in a presentation for project sponsors?

A. The "Big Picture" takeaways for executive level stakeholders
B. Data warehouse design changes
C. Line by line review of the developed code
D. Detailed statistical basis for the modeling approach used in the project

Correct Answer: A

QUESTION 65
A data scientist wants to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level.

What is the most appropriate method for this project?

A. Logistic regression
B. Linear regression
C. K-means clustering
D. Apriori algorithm

Correct Answer: A

QUESTION 66
What are the characteristics of Big Data?

A. Data volume, processing complexity, and data structure variety.
B. Data volume, business importance, and data structure variety.
C. Data type, processing complexity, and data structure variety.
D. Data volume, processing complexity, and business importance.

Correct Answer: A

QUESTION 67
You are analyzing data in order to build a classifier model. You discover non-linear data and discontinuities that will affect the model. Which analytical method would you recommend?

A. Decision Trees
B. Logistic Regression
C. ARIMA
D. Linear Regression

Correct Answer: A

QUESTION 68
What is an appropriate data visualization to use in a presentation for a project sponsor?

A. Bar chart
B. Pie chart
C. Box and Whisker plot
D. Density plot

Correct Answer: A

QUESTION 69
In a Student's t-test, what is the meaning of the p-value?

A. it is the area under the appropriate tails of the Student's distribution
B. it is the "power" of the Student's t-test
C. it is the mean of the distribution for the null hypothesis
D. it is the mean of the distribution for the alternate hypothesis

Correct Answer: A

QUESTION 70
In addition to less data movement and the ability to use larger datasets in calculations, what is a benefit of analytical calculations in a database?

A. quicker time to insight
B. more efficient handling of categorical values
C. improved connections between disparate data sources
D. full use of data aggregation functionality

Correct Answer: A

QUESTION 71
You have been assigned to do a study of the daily revenue effect of a pricing model of online transactions. When have you completed the analytics lifecycle?

A. You have written documentation, and the code has been handed off to the Data Base Administrator and business operations.
B. You have a completely developed model, and the results have shown statistically acceptable results.
C. You have presented the results of the model to both the internal analytics team and the business owner of the project.
D. You have a completely developed model based on both a sample of the data and the entire set of data available.

Correct Answer: A

QUESTION 72

Consider these itemsets:
(hat, scarf, coat)
(hat, scarf, coat, gloves)
(hat, scarf, gloves)
(hat, gloves)
(scarf, coat, gloves)

What is the confidence of the rule (gloves -> hat)?

A. 75%
B. 60%
C. 66%
D. 80%

Correct Answer: A

QUESTION 73
What is holdout data?

A. a subset of the provided data set selected at random and used to validate the model
B. a subset of the provided data set selected at random and used to initially construct the model
C. a subset of the provided data set that is removed by the data scientist because it contains data errors
D. a subset of the provided data set that is removed by the data scientist because it contains outliers

Correct Answer: A

QUESTION 74
Which characteristic applies mainly to Data Science as opposed to Business Intelligence?

A. Advanced analytical methods
B. Robust reporting
C. Focus on structured data
D. Data dashboards

Correct Answer: A

QUESTION 75
Which word or phrase completes the statement?
Theater actor is to "Artistic and Expressive" as Data Scientist is to _____

A. "Communicative and Collaborative"
B. "Introverted and Technical"
C. "Logical and Steadfast"
D. "Independent and Intelligent"

Correct Answer: A

QUESTION 76
Which process in text analysis can be used to reduce dimensionality?

A. Stemming
B. Parsing
C. Digitizing
D. Sorting

Correct Answer: A

QUESTION 77

What is the format of the output from the Map function of MapReduce?

A. Key-value pairs
B. Binary respresentation of keys concatenated with structured data
C. Compressed index
D. Unique key record and separate records of all possible values

Correct Answer: A

QUESTION 78
Which data type value is used for the observed response variable in a logistic regression model?

A. Any positive real number
B. Any integer
C. A binary value
D. Any real number

Correct Answer: C

QUESTION 79
A data scientist is given an R data frame, "empdata", with the columns Age, Salary, Occupation, Education, and Gender. The data scientist would like to examine only the Salary and Occupation columns for ages greater than 40. Which command extracts the appropriate rows and columns from the data frame?

A. empdata[empdata$Age > 40, c("Salary", "Occupation")]
B. empdata[c("Salary", "Occupation"), empdata$Age > 40]
C. empdata[Age > 40, ("Salary", "Occupation")]
D. empdata[, c("Salary", "Occupation")]$Age > 40

Correct Answer: A

QUESTION 80
What is required in a presentation for business analysts?

A. Budgetary considerations and requests
B. Operational process changes
C. Detailed statistical explanation of the applicable modeling theory
D. The presentation author's credentials

Correct Answer: B

QUESTION 81
What is LOESS used for?

A. It fits a smoothed curve to scatterplot data, to give a general sense of the data's behavior.
B. It is a significance test for the correlation between two variables.
C. It plots a continuous variable versus a discrete variable, to compare distributions across classes.
D. It is run after a one-way ANOVA, to determine which population has the highest mean value.

Correct Answer: A

QUESTION 82
Which word or phrase completes the statement? Mahout is to Hadoop as MADlib is to _____ .

A. PostgreSQL
B. R

C. Excel
D. SAS

Correct Answer: A

QUESTION 83
In linear regression modeling, which action can be taken to improve the linearity of the relationship between the dependent and independent variables?

A. Apply a transformation to a variable
B. Use a different statistical package
C. Calculate the R-Squared value
D. Change the units of measurement on the independent variable

Correct Answer: A

QUESTION 84
Data visualization is used in the final presentation of an analytics project. For what else is this technique commonly used?

A. Assessing data quality
B. Descriptive statistics
C. ETLT
D. Model selection

Correct Answer: A

QUESTION 85
You have been assigned to do a study of the daily revenue effect of a pricing model of online transactions. All the data currently available to you has been loaded into your analytics database; revenue data, pricing data, and online transaction data. You find that all the data comes in different levels of granularity. The transaction data has timestamps (day, hour, minutes, seconds), pricing is stored at the daily level, and revenue data is only reported monthly. What is your next step?

A. Report back to the business owner that the current data model does not support the business question.
B. Interpolate a daily model for revenue from the monthly revenue data.
C. Aggregate all data to the monthly level in order to create a monthly revenue model.
D. Disregard revenue as a driver in the pricing model, and create a daily model based on pricing and transactions only.

Correct Answer: A

QUESTION 86
Which SQL OLAP extension provides all possible grouping combinations?

A. CUBE
B. ROLLUP
C. UNION ALL
D. CROSS JOIN

Correct Answer: A

QUESTION 87
What is the primary bottleneck in text classification?

A. The availablilty of tagged training data.
B. The ability to parse unstructured text data.
C. The high dimensionality of text data.

D. The fact that text corpora are dynamic.

Correct Answer: A

QUESTION 88
Which characteristic applies only to Business Intelligence as opposed to Data Science?

A. Uses only structured data
B. Supports solving "what if" scenarios
C. Uses large data sets
D. Uses predictive modeling techniques

Correct Answer: A

QUESTION 89
You have been assigned to run a linear regression model for each of 5, 000 distinct districts, and all the data is currently stored in a PostgreSQL database. Which tool/library would you use to produce these models with the least effort?

A. MADlib
B. Mahout
C. R
D. HBase

Correct Answer: A

QUESTION 90
Your customer provided you with 2, 000 unlabeled records and asked you to separate them into three groups. What is the correct analytical method to use?

A. K-means clustering
B. Linear regression
C. Naive Bayesian classification
D. Logistic regression

Correct Answer: A

QUESTION 91
You are performing a market basket analysis using the Apriori algorithm. Which measure is a ratio describing the how many more times two items are present together than would be expected if those two items are statistically independent?

A. Lift
B. Leverage
C. Support
D. Confidence

Correct Answer: A

QUESTION 92
In which lifecycle stage are appropriate analytical techniques determined?

A. Model planning
B. Model building
C. Data preparation
D. Discovery

Correct Answer: A

QUESTION 93
What is Hadoop?

A. Java classes for HDFS types and MapReduce job management and HDFS
B. Java classes for HDFS types and MapReduce job management and the MapReduce paradigm
C. MapReduce paradigm and HDFS
D. MapReduce paradigm and massive unstructured data storage on commodity hardware

Correct Answer: A

QUESTION 94
You are using k-means clustering to classify heart patients for a hospital. You have chosen Patient Sex, Height, Weight, Age and Income as measures and have used 3 clusters. When you create a pair-wise plot of the clusters, you notice that there is significant overlap between the clusters. What should you do?

A. Identify additional measures to add to the analysis
B. Remove one of the measures
C. Decrease the number of clusters
D. Increase the number of clusters

Correct Answer: C

QUESTION 95
How does Pig's use of a schema differ from that of a traditional RDBMS?

A. Pig's schema is optional
B. Pig's schema requires that the data is physically present when the schema is defined
C. Pig's schema is required for ETL
D. Pig's schema supports a single data type

Correct Answer: A

QUESTION 96
You are provided four different datasets. Initial analysis on these datasets show that they have identical mean, variance and correlation values. What should your next step in the analysis be?

A. Visualize the data to further explore the characteristics of each data set
B. Select one of the four datasets and begin planning and building a model
C. Combine the data from all four of the datasets and begin planning and bulding a model
D. Recalculate the descriptive statistics since they are unlikely to be identical for each dataset

Correct Answer: A

QUESTION 97
You are asked to create a model to predict the total number of monthly subscribers for a specific magazine. You are provided with 1 year's worth of subscription and payment data, user demographic data, and 10 years worth of content of the magazine (articles and pictures). Which algorithm is the most appropriate for building a predictive model for subscribers?

A. Linear regression
B. Logistic regression
C. Decision trees
D. TF-IDF

Correct Answer: A

QUESTION 98

Which word or phrase completes the statement? Structured data is to OLAP data as quasi-structured data is to____

A. Clickstream data
B. XML data
C. Text documents
D. Image files

Correct Answer: A

QUESTION 99
What describes a true property of Logistic Regression method?

A. It is robust with redundant variables and correlated variables.
B. It handles missing values well.
C. It works well with discrete variables that have many distinct values.
D. It works well with variables that affect the outcome in a discontinuous way.

Correct Answer: A

QUESTION 100
You have been assigned to do a study of the daily revenue effect of a pricing model of online transactions. You have tested all the theoretical models in the previous model planning stage, and all tests have yielded statistically insignificant results. What is your next step?

A. Report that the results are insignificant, and reevaluate the original business question.
B. Run all the models again against a larger sample, leveraging more historical data.
C. Move forward on the model with the highest significance scores relative to the others.
D. Modify samples used by the models and iterate until a significant result occurs.

Correct Answer: A

QUESTION 101
A data scientist is asked to implement an article recommendation feature for an on-line magazine. The magazine does not want to use client tracking technologies such as cookies or reading history. Therefore, only the style and subject matter of the current article is available for making recommendations. All of the magazine's articles are stored in a database in a format suitable for analytics.

Which method should the data scientist try first?

A. K Means Clustering
B. Naive Bayesian
C. Logistic Regression
D. Association Rules

Correct Answer: A

QUESTION 102
How are window functions different from regular aggregate functions?

A. Rows retain their separate identities and the window function can access more than the current row.
B. Rows are grouped into an output row and the window function can access more than the current row.
C. Rows retain their separate identities and the window function can only access the current row.
D. Rows are grouped into an output row and the window function can only access the current row.

Correct Answer: A

QUESTION 103

Consider these itemsets:
(hat, scarf, coat)
(hat, scarf, coat, gloves)
(hat, scarf, gloves)
(hat, gloves)
(scarf, coat, gloves)

What is the confidence of the rule (hat, scarf) -> gloves?

A. 66%
B. 40%
C. 50%
D. 60%

Correct Answer: A

QUESTION 104
In the MapReduce framework, what is the purpose of the Map Function?

A. It processes the input and generates key-value pairs
B. It collects the output of the Reduce function
C. It sorts the results of the Reduce function
D. It breaks the input into smaller components and distributes to other nodes in the cluster

Correct Answer: A

QUESTION 105
You have completed your model and are handing it off to be deployed in production. What should you deliver to the production team, along with your commented code?

A. The production team needs to understand how your model will interact with the processes they already support. Give them documentation on expected model inputs and outputs, and guidance on error-handling.
B. The production team are technical, and they need to understand how the processes that they support work, so give them the same presentation that you prepared for the analysts.
C. The production team supports the processes that run the organization, and they need context to understand how your model interacts with the processes they already support. Give them the same presentation that you prepared for the project sponsor.
D. The production team supports the processes that run the organization, and they need context to understand how your model interacts with the processes they already support. Give them the executive summary.

Correct Answer: A

QUESTION 106
While having a discussion with your colleague, this person mentions that they want to perform K- means clustering on text file data stored in HDFS.

Which tool would you recommend to this colleague?

A. Mahout
B. HBase
C. Scribe
D. Sqoop

Correct Answer: A

QUESTION 107
Which method is used to solve for coefficients b0, b1, .., bn in your linear regression model :

Y = b0 + b1x1+b2x2+....+bnxn

A. Ordinary Least squares
B. Apriori Algorithm
C. Ridge and Lasso
D. Integer programming

Correct Answer: A

QUESTION 108
What describes a true limitation of Logistic Regression method?

A. It does not handle missing values well.
B. It does not handle redundant variables well.
C. It does not handle correlated variables well.
D. It does not have explanatory values.

Correct Answer: A

QUESTION 109
You submit a MapReduce job to a Hadoop cluster and notice that although the job was successfully submitted, it is not completing. What should you do?

A. Ensure that the TaskTracker is running.
B. Ensure that the JobTracker is running
C. Ensure that the NameNode is running
D. Ensure that a DataNode is running

Correct Answer: A

QUESTION 110
A disk drive manufacturer has a defect rate of less than 1.5% with 98% confidence. A quality assurance team samples 1000 disk drives and finds 14 defective units. Which action should the team recommend?

A. The manufacturing process is functioning properly and no further action is required
B. A larger sample size should be taken to determine if the plant is operating correctly
C. A smaller sample size should be taken to determine if the plant is operating correctly
D. There is a flaw in the quality assurance process and the sample should be repeated

Correct Answer: A

QUESTION 111
What is a core deliverable at the end of the analytic project?

A. An implemented database design
B. A whitepaper describing the project and the implementation
C. A presentation for project sponsors
D. The training materials

Correct Answer: C

QUESTION 112
You have been assigned to run a logistic regression model for each of 100 countries, and all the data is currently stored in a PostgreSQL database. Which tool/library would you use to produce these models with the least effort?

A. MADlib
B. Mahout

C. RStudio
D. HBase

Correct Answer: A

QUESTION 113
Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to visitors to your website has any impact on their purchase decision.

Which analysis method should you use?

A. K-means clustering
B. Association rules
C. Student T-test
D. One-way ANOVA

Correct Answer: D

QUESTION 114
Imagine you are trying to hire a Data Scientist for your team. In addition to technical ability and quantitative background, which additional essential trait would you look for in people applying for this position?

A. Communication skill
B. Scientific background
C. Domain expertise
D. Well Organized

Correct Answer: A

QUESTION 115
What describes the use of UNION clause in a SQL statement?

A. Operates on queries and potentially increases the number of rows
B. Operates on queries and potentially decreases the number of rows
C. Operates on tables and potentially decreases the number of columns
D. Operates on both tables and queries and potentially increases both the number of rows and columns

Correct Answer: A

QUESTION 116
You have run the association rules algorithm on your data set, and the two rules {banana, apple} => {grape} and {apple, orange}=> {grape} have been found to be relevant. What else must be true?

A. {grape, apple, orange} must be a frequent itemset.
B. {banana, apple, grape, orange} must be a frequent itemset.
C. {grape} => {banana, apple} must be a relevant rule.
D. {banana, apple} => {orange} must be a relevant rule.

Correct Answer: A

QUESTION 117
When would you use a Wilcoxson Rank Sum test?

A. When you cannot make an assumption about the distribution of the populations
B. When the data can easily be sorted
C. When the populations represent the sums of other values
D. When the data cannot easily be sorted

Correct Answer: A

QUESTION 118
In the MapReduce framework, what is the purpose of the Reduce function?

A. It aggregates the results of the Map function and generates processed output
B. It distributes the input to multiple nodes for processing
C. It writes the output of the Map function to storage
D. It breaks the input into smaller components and distributes to other nodes in the cluster

Correct Answer: A

QUESTION 119
Which of the following is an example of quasi-structured data?

A. OLAP
B. OLTP
C. Customer record table
D. Clickstream data

Correct Answer: D

QUESTION 120
A Data Scientist is assigned to build a model from a reporting data warehouse. The warehouse contains data collected from many sources and transformed through a complex, multi-stage ETL process. What is a concern the data scientist should have about the data?

A. It is too processed
B. It is not structured
C. It is not normalized
D. It is too centralized

Correct Answer: A

QUESTION 121
Which word or phrase completes the statement? Emphasis color is to standard color as _____ .

A. Main message is to context
B. Main message is to key findings
C. Frequent item set is to item
D. Pie chart is to proportions

Correct Answer: A

QUESTION 122
Which data asset is an example of semi-structured data?

A. XML data file
B. Database table
C. Webserver log
D. News article

Correct Answer: A

Trying our product !

★ **100%** Guaranteed Success

★ **100%** Money Back Guarantee

★ **365 Days** Free Update

★ **Instant Download** After Purchase

★ **24x7** Customer Support

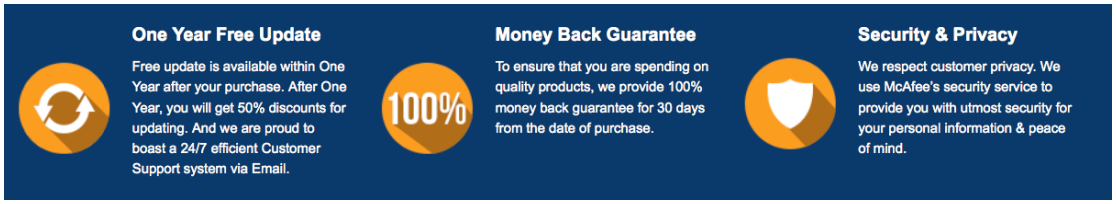
★ Average **99.9%** Success Rate

★ More than **69,000** Satisfied Customers Worldwide

★ Multi-Platform capabilities - **Windows, Mac, Android, iPhone, iPod, iPad, Kindle**

Need Help

Please provide as much detail as possible so we can best assist you.
To update a previously submitted ticket:



Guarantee & Policy | Privacy & Policy | Terms & Conditions