E20-007Q&As

Data Science and Big Data Analytics

Pass EMC E20-007 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.certbus.com/E20-007.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by EMC Official Exam Center

- **Instant Download** After Purchase
- **100% Money Back** Guarantee
- **365 Days** Free Update
- **800,000+** Satisfied Customers

QUESTION 1

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 2

Which word or phrase completes the statement; "A data scientist would consider a RDBMS is to a table as R is to a _____."?

A. Data frame

B. List

C. Matrix

D. Array

Correct Answer: A

QUESTION 3

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 4

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

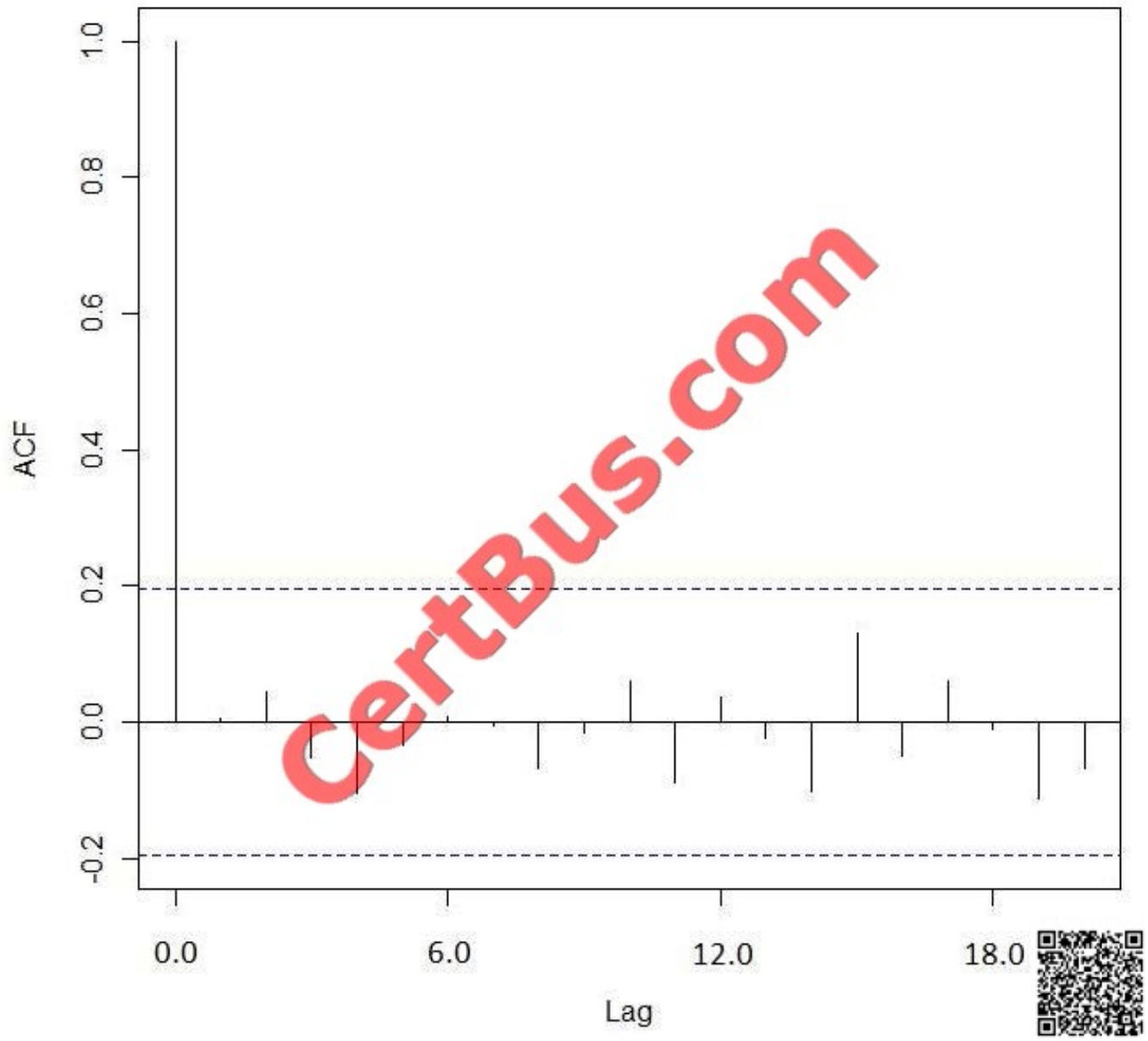
C. Show if the model will meet the SLA

D. Show the technique to be used in the production environment

Correct Answer: B

QUESTION 5

Refer to the exhibit.



In the exhibit, a correlogram is provided based on an autocorrelation analysis of a sample dataset. What can you conclude based only on this exhibit?

A. There appears to be no structure left to model in the data

B. There appears to be a seasonal component in the data

C. Lag 1 has a significant autocorrelation

D. There appears to be a cyclical component in the data

Correct Answer: A

QUESTION 6

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 7

A data scientist is given an R data frame (i.e., empdata) with the following columns:

Age

Salary

Occupation

Education Gender

The scientist wants 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 8

A business colleague who is new to Hadoop approaches you with a question. The colleague wants to know the best approach to access their data. The colleague has previously worked extensively with SQL and databases.

Which query interface should be recommended?

A. Hive

B. Pig

C. Howl

D. HBase

Correct Answer: A

QUESTION 9

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 Na?e Bayesian classification

Correct Answer: A

QUESTION 10

You have been assigned to perform a study of the daily revenue effect of a pricing model of online transactions. All data currently available to you has been loaded into your analytics database. This includes revenue data, pricing data, and online transaction data.

You discover that all data comes in different levels of granularity. The transaction data has timestamps consisting of day, hour, minutes, and seconds. Pricing is stored at the daily level and revenue data is only reported monthly.

What is the 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 the key reason in the pricing model and create a daily model based on pricing and transactions only.

Correct Answer: A

QUESTION 11

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 12

You have been assigned to perform a study of the daily revenue effect of a pricing model of online transactions. When is the analytics lifecycle considered completed?

A. When written documentation has been produced and the code has been handed off to the DBA/operations.

B. When a model has been completely developed and the results have shown statistically acceptable results.

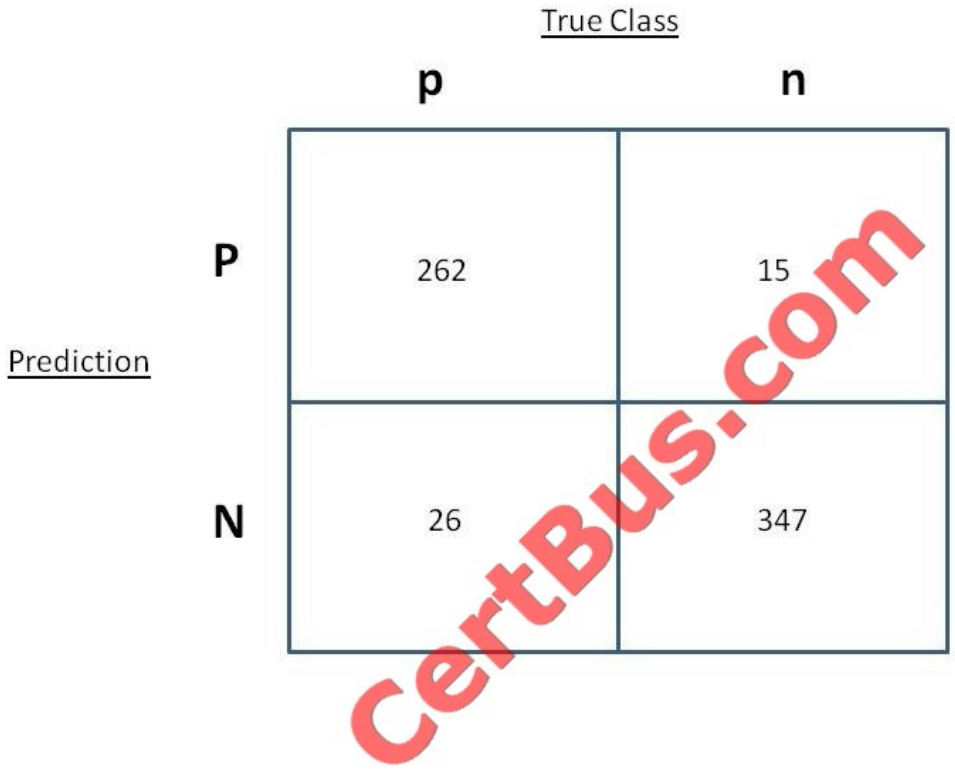
C. When the results of the model have been presented to both the internal analytics team and the business owner of the project.

D. When a model has been completely developed based on both a sample of the data and the entire set of data available.

Correct Answer: A

QUESTION 13

Refer to the exhibit.



You have scored your Naive bayesian classifier model on a hold out test data for cross validation and determined the way the samples scored and tabluated them as shown in the exhibit.

What are the Precision and Recall rate of the model?

A. Precision = 262/277 Recall = 262/288

B. Precision =262/288 Recall = 262/277

C. Precision = 277/262 Recall = 288/262

D. Precision = 288/262 Recall = 277/262

Correct Answer: A

QUESTION 14

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 15

What is the mandatory Clause that must be included when using Window functions?

A. OVER

B. RANK

C. PARTITION BY

D. RANK BY

Correct Answer: A

QUESTION 16

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 17

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 18

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 19

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 20

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 21

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 22

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 23

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 24

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 25

Refer to the exhibit.

Independent Variable	Coefficient	P-Value
A	1.23	0.001
B	-6.72	0.035
C	4.57	0.020
D	10.23	0.123

After analyzing a dataset, you report findings to your team:

1.

Variables A and C are significantly and positively impacting the dependent variable.

2.

Variable B is significantly and negatively impacting the dependent variable.

3.

Variable D is not significantly impacting the dependent variable.

After seeing your findings, the majority of your team agreed that variable B should be positively impacting the dependent variable.

What is a possible reason the coefficient for variable B was negative and not positive?

A. Variable B is interacting with another variable due to correlated inputs

B. Variable B needs a quadratic transformation due to its relationship to the dependent variable

C. The information gain from variable B is already provided by another variable

D. Variable B needs a logarithmic transformation due to its relationship to the dependent variable

Correct Answer: A

QUESTION 26

Refer to the exhibit.

Attribute	Info-Gain
Age	0.0310
Income	0.0100
Gender	0.0034
Credit Score	0.045

You are building a decision tree. In this exhibit, four variables are listed with their respective values of info-gain. Based on this information, on which attribute would you expect the next split to be in the decision tree?

A. Credit Score

B. Age

C. Income

D. Gender

Correct Answer: A

QUESTION 27

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 28

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 29

Which word or phrase completes the statement?

Business Intelligence is to ad-hoc reporting and dashboards as Data Science is to _____ .

A. Optimization and Predictive Modeling

B. Alerts and Queries

C. Structured Data and Data Sources

D. Sales and profit reporting

Correct Answer: A

QUESTION 30

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

To Read the Whole Q&As, please purchase the Complete Version from Our website.

Try 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 800,000 Satisfied Customers Worldwide
Multi-Platform capabilities - Windows, Mac, Android, iPhone, iPod, iPad, Kindle

We provide exam PDF and VCE of Cisco, Microsoft, IBM, CompTIA, Oracle and other IT Certifications.
You can view Vendor list of All Certification Exams offered:

https://www.certbus.com/allproducts

Need Help

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