

## 70-773 Dumps

### Analyzing Big Data with Microsoft R (beta)

<https://www.certleader.com/70-773-dumps.html>



### NEW QUESTION 1

You are running a parallel function that uses the following R code segment. (Line numbers are included for reference only.)

```
01 cp <- 0.01 xval <- 0 maxdepth <- 5
02
03 (form, data = "segmentationDataBig",maxDepth = maxdepth, cp = cp, xVal = xval, blocksPerRead = 250
```

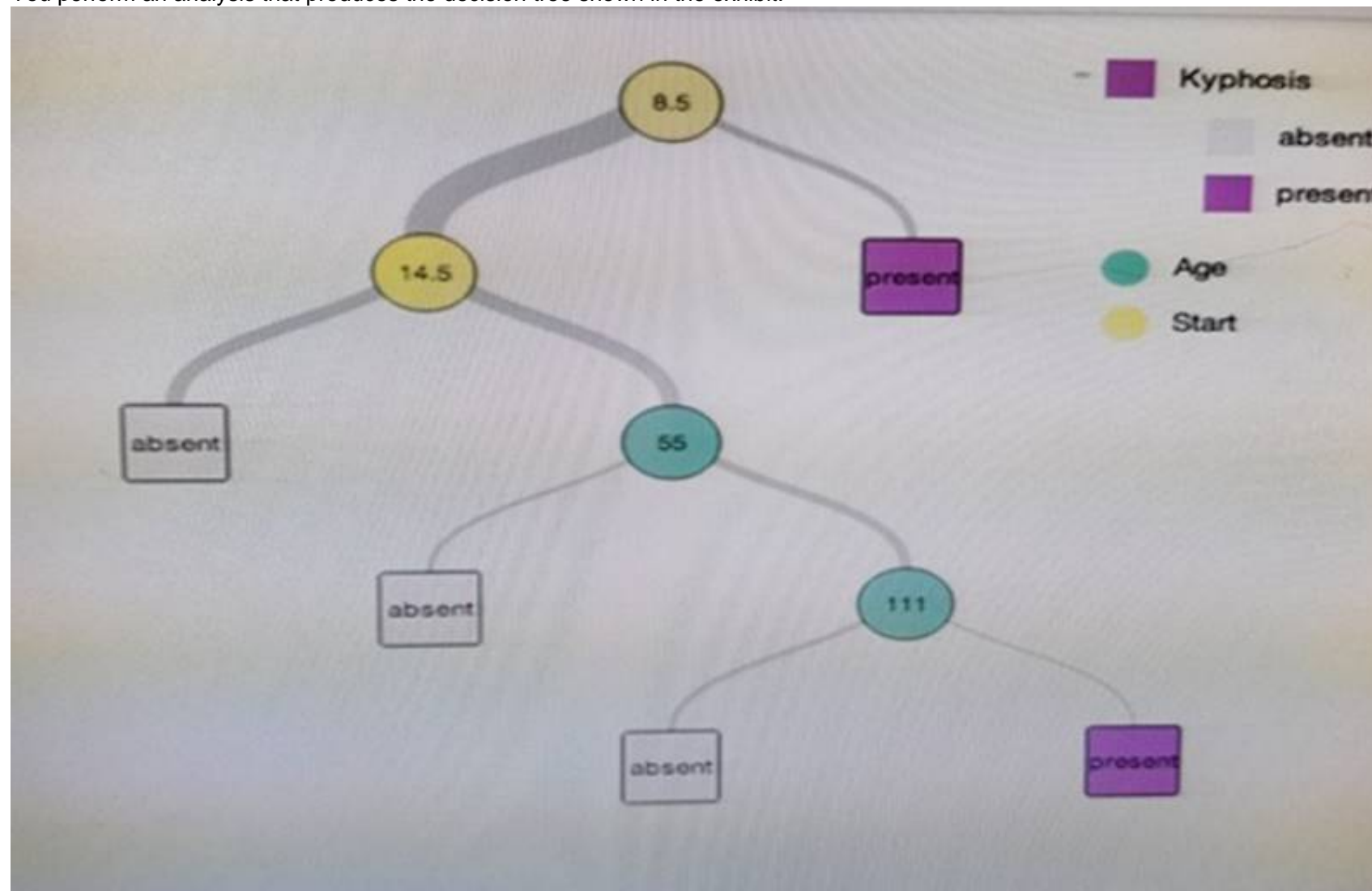
You need to complete the R code. The solution must support chunking. Which function should insert at line 02?

- A. rxBTrees
- B. rxExec
- C. rxDForest
- D. rxDTree

**Answer: C**

### NEW QUESTION 2

You perform an analysis that produces the decision tree shown in the exhibit.



How many leaf nodes are there on the tree?

- A. 2
- B. 3
- C. 5
- D. 7

**Answer: B**

### NEW QUESTION 3

You have following regression forest.

```
rxDForest(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc., data = stackloss, maxDepth = 3, nTree = 200, nTry = 2)

Type of decision forest: anova
Number of trees: 200
No. of variables tried at each split: 2
Mean of squared residuals: 44.54992 % Var explained: 65
```

Which variable contributes the most to the dependent variable?

- A. stack.loss
- B. Water.Temp
- C. Air.Flow
- D. Acid.Conc

**Answer: A**

### NEW QUESTION 4

Note: This Question is part of a series of Questions that use the same or similar answer choices. An answer choice may be correct than one question in the series.

Each question is independent of the other questions in this series. Information and details provided in a question apply only to that question.

You have a data source that is larger than memory.

You need to visualize the distribution of the values for a variable in the data source. What should you use?

- A. the Describe package
- B. the rxHistogram function
- C. the rxSummary function
- D. the rxQuantile function
- E. the rxCube function
- F. the summary function
- G. the rxCrossTabs function
- H. the ggplot2 package

**Answer: B**

#### NEW QUESTION 5

You have an Apache Hadoop Hive data warehouse. RevoScaleR is not installed. You need to sort the data according to the variables in the dataset.

What should you do?

- A. Connect to the database by using an ODBC connection, and then use the rxSort function.
- B. Create a table in the ORC file format.
- C. Connect to the database by using an ODBC connection, and then use the rxDataStep function.
- D. Execute a Hive query that sorts the data, and then reads the results.

**Answer: D**

#### NEW QUESTION 6

Note: This question is part of a series of questions that use the same or similar answer choices. An answer choice may be correct for more than one question in the series. Each question is independent of the other questions in this series. Information and details provided in a question apply only to that question.

You need to calculate a measure of central tendency and variability for the variables in a dataset that is grouped by using another categorical variable.

What should you use?

- A. the Describe package
- B. the rxHistogram function
- C. the rxSummary function
- D. the rxQuantile function
- E. the rxCube function
- F. the summary function
- G. the rxCrossTabs function
- H. the ggplot2 package

**Answer: C**

#### NEW QUESTION 7

You have a dataset that has multiple blocks and only numeric variables. You are computing in a local compute context.

You plan to lag a variable named x to create a new variable named x\_lagged by using a transform function. You will create a new element in the output of the function.

You need to minimize the number of missing values.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Assign a value to the first value of x\_lagged in the current block.
- B. Use rxSet to store the last value of x\_lagged in the current block.
- C. Use rxSet to store the last value of x in the current block.
- D. Use rxGet to retrieve the first value of x in the next block to be processed.
- E. Use rxGet to retrieve a value stored in processing of the prior block.

**Answer: ACD**

#### NEW QUESTION 8

You need to build a model that looks at the probability of an outcome. You must regulate between L1 and L2.

Which classification method should you use?

- A. Two-Class Neural Network
- B. Two-Class Support Vector Machine
- C. Two-Class Decision Forest
- D. Two-Class Logistic Regression

**Answer: A**

#### NEW QUESTION 9

Note: This question is part of a series of Questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution-After you answer a question in this section, you will NOT be able to return to it- As a result, these questions will not appear in the review screen.

You use dplyrXdf and you discover that after you exit the session, the output files that were created were deleted. You need to prevent the files from being deleted.

Solution: You use dplyrXdf with the persist verb.

Does this meet the goal?

- A. Yes
- B. No

Answer: A

NEW QUESTION 10

You are running a large logistic regression for 1,000 feature variables by using the logisticRegression0 function in the MicrosoftML package. All of the predictor variables are numeric.  
Currently, you specify the input variables separately by using the following formula.

```
Outcome ~ Feature000 + Feature001 + Feature002 + ... + Feature999
```

You discover that it takes 20 minutes to estimate each model.  
You need to reduce the amount of time required to estimate each model without losing any information in the predictors.  
What should you do?

- A. Use stepControl0 to perform stepwise regression to limit the number of variables that contribute to the model.
- B. Use selectFeatures0 to select the features that provide the most information about the outcome variable.
- C. Use princomp0 on the correlation matrix of Features, and then use only the first 100 principle components to reduce the number of input variables.
- D. Use concat0 to create a single array variable named Features, and then specify a newformula named Outcome - Features.

Answer: B

NEW QUESTION 10

You have cloud and on-premises resources that include Microsoft SQL Server and a big data environment in Apache Hadoop.  
You have 50 billion fact records.  
You need to build time series models to execute forecasting reports on the fact records. What should you use?

- A. RxSpark on the Hadoop cluster
- B. RxHadoopMR on the Hadoop cluster
- C. RxLocalseq on the SQL Server database
- D. RxLocalParallel on the SQL Server database

Answer: A

NEW QUESTION 14

DRAG DROP  
You need to set the compute context for three different target environments.  
Which Statement should you use for each environment? To answer, drag the appropriate statements to the correct execution contexts. Each statement may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.  
NOTE: Each correct selection is worth one point.

Statements

RxHadoopMR()

rxSetComputeContext('local')

rxSetComputeContext('localpar')

RxSpark()

Answer Area

Parallelized execution across the cores of the edge node server, except for rxExec calls, which are executed serially:

Parallelized execution across the cores of the edge node server:

Parallelized distributed execution via Map Reduce across the nodes of the cluster:

Statement

Statement

Statement

Answer:

Explanation:

Statements

RxHadoopMR()

rxSetComputeContext('local')

rxSetComputeContext('localpar')

RxSpark()

Answer Area

Parallelized execution across the cores of the edge server, except for rxExec calls, which are executed s

Parallelized execution across the cores of the edge

Parallelized distributed execution via Map Reduce ac

RxHadoopMR()

RxSpark()

rxSetComputeContext('local')

NEW QUESTION 17

You have one class support vector machines (SVMs).  
You have a large dataset, but you do not have enough training time to fully test the model. What is an alternative method to validate the model?

- A. Use Principal Components Analysis (PCA) Based Anomaly detection
- B. Replace the SVMs with two class SVMs.
- C. Perform feature selection.



D. Use outlier detection.

**Answer:** A

#### NEW QUESTION 20

Note: This question is part of a series of Questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution-After you answer a question in this section, you will NOT be able to return to it- As a result, these questions will not appear in the review screen.

You use dplyrXdf and you discover that after you exit the session, the output files that were created were deleted. You need to prevent the files from being deleted.

Solution: You remove all instances of the file.remove method.

Does this meet the goal?

A. Yes

B. No

**Answer:** B

#### NEW QUESTION 24

Note: This question is part of a series of Questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution-After you answer a question in this section, you will NOT be able to return to it- As a result, these questions will not appear in the review screen.

You have a Microsoft SQL Server instance that has R Services (In-Database) installed. You need to monitor the R jobs that are sent to SQL Server.

Solution: You call a function from the RevoPemaR Package.

Does this meet the goal?

A. Yes

B. No

**Answer:** B

#### NEW QUESTION 29

Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is exactly the same in each question in this series.

Start of repeated scenario

You are developing a Microsoft R Open solution that will leverage the computing power of the database server for some of your datasets.

You are performing feature engineering and data preparation for the datasets. The following is a sample of the dataset.

```
rxGetInfo(df)
head(df)
```

	age	incwage	perwt	wkswork1	state
1	50	9000	30	48	Indiana
2	41	35000	20	48	Indiana
3	55	40400	21	52	Indiana
4	56	45000	30	52	Indiana
5	46	17200	60	52	Indiana
6	49	35000	21	52	Indiana

End of repeated scenario

You have the following R code.

```
createRandomSample <- function(data)
{
  data$.rxRowSelection <- as.logical(rbinom(length(data[[1]]), 1, .10)) return(data)
}
workers <- file.path(rxGetOption("sampleDataDir"), "Workers.xdf")
df <- rxXdfToDataFrame(file = workers, transformFunc = createRandomSample, transformVars = "age")
```

Which function determines the variable?

A. transformVars

B. rxXdfToDataFrame

C. createRandomSample

D. transformFunc

**Answer:** A

#### NEW QUESTION 30

.....

## Thank You for Trying Our Product

\* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

\* One year free update

You can enjoy free update one year. 24x7 online support.

\* Trusted by Millions

We currently serve more than 30,000,000 customers.

\* Shop Securely

All transactions are protected by VeriSign!

**100% Pass Your 70-773 Exam with Our Prep Materials Via below:**

<https://www.certleader.com/70-773-dumps.html>