Data Science Using R Final Test (Internship)

* Required

Data Science Using R Final Test (Internship)

All questions are compulsory. No negative marking.

Functions are defined using the _____ directive and are stored as R objects *

- ○ functions()
- ◉ function()
- ○ None of abobe
- ○ funct()

Which of the following is a categorical outcome? *

- ○ Accuracy
- ○ RSquared
- ◉ RMSE
- ○ All of the Mentioned

How do you handle missing or corrupted data in a dataset? *

- (●) All of the above.
- ( ) Replace missing values with mean/median/mode
- ( ) Drop missing rows or columns
- ( ) Assign a unique category to missing values

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

- (●) 1
- ( ) 0
- ( ) 2
- ( ) 3

MapReduce Processing is _____ *

- (●) Linear
- ( ) Serieal
- ( ) Batch
- ( ) Shared

Data frames can be converted to a matrix by calling data._____ *

- ( ) mat()
- (●) matrix()
- ( ) matr()
- ( ) All of above

For two runs of K-Mean clustering is it expected to get same clustering results? *

- ○ Yes
- ◉ NO

What is the role of exploratory graphs in data analysis? *

- ○ They are used in place of formal modeling
- ○ They are made for formal presentations
- ◉ summarize main characteristic of data
- ○ Axes, legends, and other details are clean and exactly detailed

Movie Recommendation systems are an example of: *

- ○ Clusturing
- ○ Classification
- ◉ Both 1 and 2
- ○ None of above

_____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data. *

- ◉ MapReduce
- ○ All of the mentioned
- ○ Oozie
- ○ Mahout

What type of analysis could be most effective for predicting temperature on the following type of data. *

Date	Temperature	precipitation	temperature/precipitation
12/12/12	7	0.2	35
13/12/12	9	0.123	73.1707317073
14/12/12	9.2	0.34	27.0588235294
15/12/12	10	0.453	22.0750551876
16/12/12	12	0.33	36.3636363636
17/12/12	11	0.8	13.75

- ○ Classification
- ○ Clustering
- ◉ Time Series Analysis
- ○ None of above

All of the following accurately describe Hadoop, EXCEPT: *

- ◉ Distributed computing approach
- ○ Real-time
- ○ Open source
- ○ Java-based

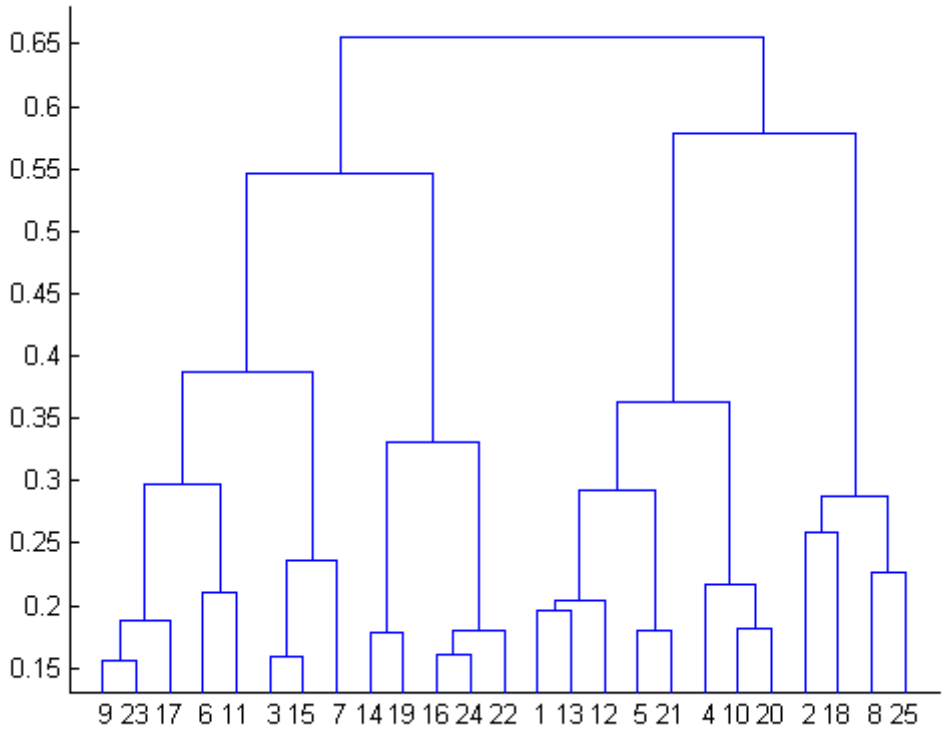
Suppose B is a categorical variable and we wish to draw a boxplot for every level of the categorical level. Which of the below commands will help us achieve that? *

data

A	B
1	Right
2	Wrong
3	Wrong
4	Right
5	Right
6	Wrong
7	Wrong
8	Right

- ○ None of the above
- ○ boxplot(A,B,data=data)
- ○ boxplot(A|B,data=data)
- ◉ boxplot(A~B,data=data)

After performing k-means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram? *



- (●) The above dendrogram interpretation is not possible for K-Means clustering analysis
- ( ) There were 28 data points in clustering analysis
- ( ) The proximity function used is Average-link clustering
- ( ) The best no. of clusters for the analyzed data points is 4

Below are the 8 actual values of target variable in the train file. [0,0,0,1,1,1,1,1] * What is the entropy of the target variable?

- (●) -(5/8 log(5/8) + 3/8 log(3/8))
- ( ) 3/8 log(5/8) + 5/8 log(3/8)
- ( ) 5/8 log(3/8) − 3/8 log(5/8)
- ( ) 5/8 log(5/8) + 3/8 log(3/8)

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

- [ ] 2
- [x] 1
- [ ] 3
- [ ] 0

Which of the following method is used for finding optimal number of clusters in the K-Mean algorithm? *

- [ ] Manhattan method
- [ ] Euclidean method
- [ ] None of these
- [x] Elbow method

In order to apply a combiner, what is one property that has to be satisfied by the values emitted from the mapper? *

- [ ] Only if the values satisfy associative and commutative property it can be done.
- [ ] Output of the mapper and output of the combiner has to be same key value pair and they can be heterogeneous
- [ ] Combiner can be applied always to any data
- [x] Output of the mapper and output of the combiner has to be same key value pair.

Which of the following is finally produced by Hierarchical Clustering? *

- ○ final estimate of cluster centroids
- ○ assignment of each point to clusters
- ◉ tree showing how close things are to each other
- ○ all of the mentioned

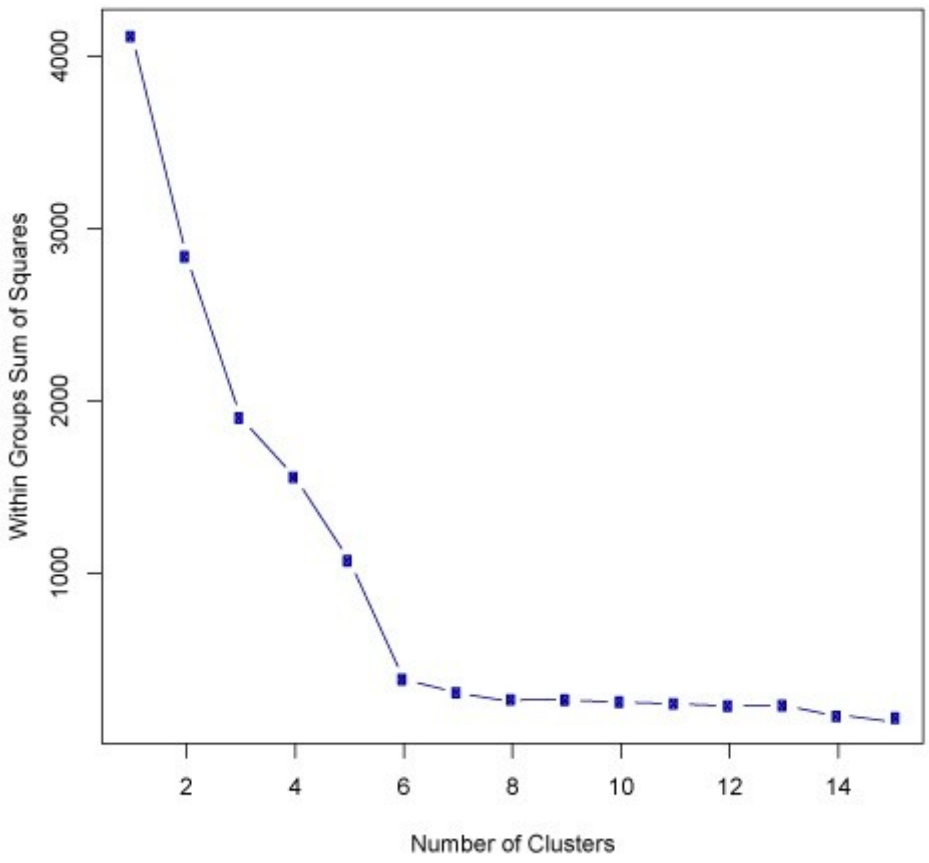
Data used to build a data mining model. *

- ○ Train Data
- ○ Test data
- ◉ Validation Data
- ○ Hidden Data

Which of the following draws plot on current graphics device? *

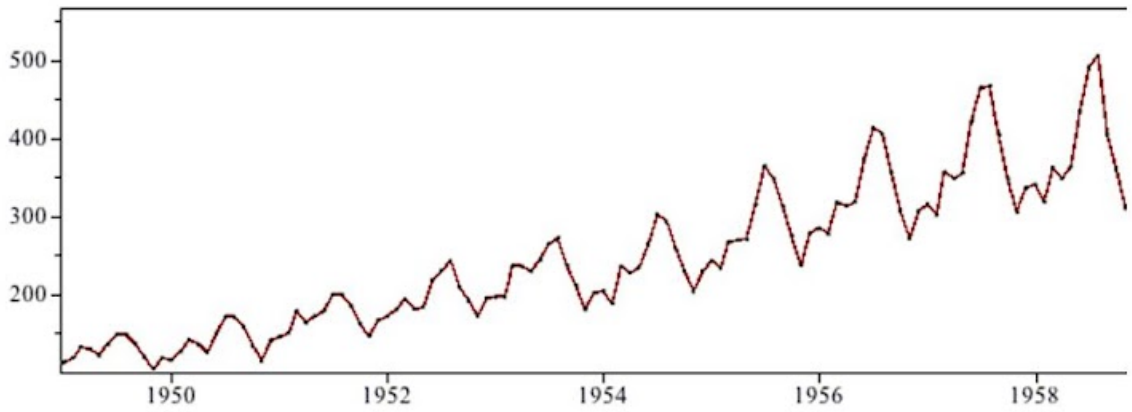
- ◉ print.ggplot
- ○ ggmissplot
- ○ printplot
- ○ ggfluctuation

What should be the best choice for number of clusters based on the following results: *



- (●) 6
- ( ) 14
- ( ) More than 14
- ( ) 5

The below time series plot contains both Cyclical and Seasonality component. *



- ○ True
- ◉ False

Which of the following step is performed by data scientist after acquiring the data ? *

- ○ Data Integration
- ◉ Data Cleansing
- ○ Data Replication
- ○ All of the Mentioned

Select the correct exploratory graph characteristic *

- ○ Quick representation of the data
- ○ All of the mentioned
- ◉ Color is used for personal information
- ○ A large number of exploratory graphs are made

Imagine, you are solving a classification problem with the highly imbalanced class. The majority class is observed 99% of the times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case? *

- ○ Precision and recall metrics are good for imbalanced class problems.
- ○ Accuracy metric is a good idea for imbalanced class problems.
- ○ Precision and recall metrics aren't good for imbalanced class problems.
- ◉ Accuracy metric is not a good idea for imbalanced class problems.

Which of the following involves predicting a categorical response ? *

- ○ Summarization
- ◉ Classification
- ○ Regression
- ○ Clustering

Hadoop is a framework that works with a variety of related tools. Common tools include: *

- ○ MapReduce, Hummer and Iguana
- ○ MapReduce, Heron and Trumpet
- ◉ MapReduce, Hive and HBase
- ○ MapReduce, MySQL and Google Apps

What could be the possible reason(s) for producing two different dendrograms using the agglomerative clustering algorithm for the same dataset? *

- ○ no. of variables used
- ○ Proximity function used
- ○ no. of data points used
- ◉ All of the above

According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop? *

- ○ Collecting and storing unstructured data
- ◉ Big data management and data mining
- ○ Data warehousing and business intelligence
- ○ Management of Hadoop clusters

Missing values in this csv file has been represented by an exclamation mark ("!") and a question mark ("?"). Which of the codes below will read the above csv file correctly into R? *

- ○ csv('Dataframe.csv',header=FALSE, sep=',',na.strings=c('?'))
- ○ csv('Dataframe.csv')
- ◉ csv2('Dataframe.csv',header=FALSE,sep=',',na.strings=c('?','!'))
- ○ dataframe('Dataframe.csv')

five-number summary does not produce which of the following information *

- ○ Mean
- ○ Median
- ○ All of the mentioned
- ◉ Mode

On which node, all the metadata related to HDFS including the information about data nodes, files stored on HDFS, and Replication, etc. are stored and maintained. *

- ○ Slave node
- ○ Data Node
- ◉ Name Node
- ○ Secondary Name Node

Match the items and choose a correct option *

Match the items from Group I with items in Group II

Group I	Group II
A. Apache Pig	1. Able to fire a single query and collects data from different storage
B. Apache HBase	2. Use for Extract, Transfer and Load data
C. Apache Drill	3. Non-relational distributed database
D. Apache Mahout	4. Machine-learning and data mining library

- ○ A-4, B-2, C-3, D-1
- ◉ A-2, B-1, C-4, D-3
- ○ A-2, B-3, C-1, D-4
- ○ A-4, B-3, C-2, D-1

Function used for linear regression in R is *

- ○ regression.linear(formula, data)
- ○ lr(formula, data)
- ○ lrm(formula, data)
- ◉ lm(formula, data)

Point out the wrong statement. *

- ○ k-means clustering aims to partition n observations into k clusters
- ○ k-means clustering is a method of vector quantization
- ◉ k-nearest neighbor is same as k-means
- ○ none of the mentioned

_____ can be used for batch processing of data and aggregation operations. *

- ○ Hive
- ○ None of the above
- ○ Oozie
- ◉ MapReduce

Which of the following syntax is used to install forecast package ? *

- ○ install.packages("cast")
- ○ All of the mentioned
- ○ install.pack("forecast")
- ◉ install.packages("forecast")

Which of the following is characteristic of best machine learning method ? *

- ○ Fast
- ○ Salable
- ○ Accurate
- ◉ All of the Mentioned

How many coefficients do you need to estimate in a simple linear regression model (One independent variable)? *

- ○ 1
- ◉ 2
- ○ 3
- ○ 4

Which is not the V of Big Data *

- ○ Velocity
- ◉ Versatile
- ○ Volume
- ○ Veriety

For which of the following hyperparameters, higher value is better for the decision tree algorithm? *

- ○ Number of samples used for split
- ◉ Cant say.
- ○ Samples for leaf
- ○ Depth of tree

A dataset has been read in R and stored in a variable “dataframe”. Missing values have been read as NA. Which of the following codes will not give the number of missing values in each column? *

A	10	Sam
B	NA	Peter
C	30	Harry
D	40	NA
E	50	Mark

- ◉ table(is.na(dataframe))
- ○ colSums(is.na(dataframe))
- ○ apply(is.na(dataframe),2,sum)
- ○ Csapply(dataframe,function(x) sum(is.na(x))

What is ggplot2 an implementation of ? *

- ○ the base plotting system in
- ○ 3D visualization system
- ◉ the Grammar of Graphics developed by Leland Wilkinson
- ○ the S language originally developed by Bell Labs

Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?1. Stemming 2. Stop word remove 3. object Standardization *

- ○ 1 and 2
- ○ 1 and 3
- ◉ 1,2 and 3
- ○ 2 and 3

In practice, Line of best fit or regression line is found when _____ *

- ◉ Sum of the square of residuals (∑ (Y-h(X))2) is minimum
- ○ Sum of residuals (∑(Y − h(X))) is minimum
- ○ Sum of the square of residuals (∑ (Y-h(X))2) is maximum
- ○ Sum of the absolute value of residuals (∑|Y-h(X)|) is maximum

_____ grammar makes a clear distinction between your data and what gets displayed on the screen or page. *

- ( ) d3.js
- (•) ggplot2
- ( ) ggplot1
- ( ) ggplot3

What would be the output of following code ? x - c("a", "b", "c", "c", "d", "a") x[c(1, 3, 4)] *

- (•) “a” “c” “c”
- ( ) All of the mentioned
- ( ) “a” “c” “b”
- ( ) “a” “b” “c”

Which will be the output of following code ? x - 3 switch(6, 2+2, mean(1:10), rnorm(5)) *

- ( ) 2
- ( ) 1
- ( ) 3
- (•) null

Which of the following is/are valid iterative strategy for treating missing values before clustering analysis? *

- (●) Imputation with Expectation Maximization algorithm
- ( ) Nearest Neighbor assignment
- ( ) All of the above
- ( ) Imputation with mean

Back Submit Clear form