

Assignment 8

Due Nov 12 at 11:59pm	Points 100	Questions 23
Available Oct 30 at 6am - Nov 15 at 11:59pm	Time Limit None	
Allowed Attempts 2		

Instructions

(no late submission accepted) -- Complete the assessment covering all materials covered thus far with a focus on NoSQL databases. There is no time limit for the test and you may use any available resources. However, you may not collaborate with anyone. You will have two attempts and the highest grade counts.

This quiz was locked Nov 15 at 11:59pm.

Attempt History

	Attempt	Time	Score
KEPT	Attempt 2	16 minutes	100 out of 100
LATEST	Attempt 2	16 minutes	100 out of 100
	Attempt 1	27 minutes	92.92 out of 100

⚠️ Correct answers are hidden.

Score for this attempt: **100** out of 100
Submitted Nov 12 at 11:26pm
This attempt took 16 minutes.

Question 1	4 / 4 pts

What should be done during EDA?

- ☒ Generate a large quantity of questions
- ☒ Visualize distributions, means, variances, etc
- ☒ Identify typical values and outliers (unusual values)
- ☒ Transform data for visualization and statistical modeling
- ☒ Refine the questions you asked
- ☐ Build a sophisticated model to answer the questions

Question 2

4 / 4 pts

To understand the infrastructure of a data repository, which of these must be considered? Check all that apply.

- ☒ What is the volume and complexity of the data?
- ☒ Who will manage the data store?
- ☒ What product will be used to house the data?
- ☐ What is the relational model for the data?

Question 3**5 / 5 pts**

True or False? *PostgreSQL* is a relational database engine that does not use SQL as its primary query language and is thus part of the NoSQL set of databases.

☐ True☒ False**Question 4****5 / 5 pts**

_____ is a database built on Hadoop and is designed for scaling across clusters of inexpensive computers with support for versioning and data compression.

☒ HBase☐ MapReduce☐ Cassandra☐ MongoDB**Question 5****5 / 5 pts**

A columnar database is most like an R _____ object.

- ☐ list
- ☒ data frame
- ☐ matrix
- ☐ array
- ☐ mode
- ☐ factor

Question 6

5 / 5 pts

Both MongoDB and CouchDB primarily use _____ as its native query language.

Java Script

Question 7

5 / 5 pts

Which of the databases below will work with R? Check all that apply.

☒ filehash☒ MySQL☒ HBase☒ Cassandra☒ Redis☒ Riak**Question 8****5 / 5 pts**

True or False? MemcachedDB is a key-value database.

☒ True☐ False**Question 9****5 / 5 pts**

Sue Ann, one of the newest Data Scientists hired by Social Dynamix has been tasked with analyzing the employee base of a large defense contractor. In particular, the client is interested in understanding how the employees are "socially connected". Data from their internal web sites and databases as well as external data from

Facebook, LinkedIn, Reddit, among others are expected to be mined in this effort. Which database would you recommend that Sue Ann consider as her primary analytic data store?

- ☒ Neo4J
- ☐ CouchDB
- ☐ PostgreSQL
- ☐ HBase
- ☐ Cassandra
- ☐ HyperTable

Question 10

5 / 5 pts

Which of the following is not a relational database management system?

- ☒ Cassandra
- ☐ Postgres
- ☐ Microsoft Access
- ☐ HSQLDB
- ☐ MySQL

☐ H2

Question 11

5 / 5 pts

One of the primary strengths of graph databases is dealing with _____ data where the database consists of _____ and _____ between them.

- ☒ highly interconnected, nodes, relationships
- ☐ sparse, records, relations
- ☐ connected, edges, links
- ☐ diverse, multiple data types, dependencies

Question 12

5 / 5 pts

Which of the following is a graph database?

- ☒ OrientDB
- ☐ CouchDB
- ☐ MongoDB
- ☐ HBase

Question 13**4 / 4 pts**

Each class can be considered a _____ for storing data about a set of similar things.

Question 14**4 / 4 pts**

Every row in a table is uniquely identified by its _____.

☒ primary key☐ key☐ alternate key☐ id**Question 15****4 / 4 pts**

A table is created with the SQL _____ statement.

☒ CREATE TABLE

☐ TABLE

☐ RELATION

☐ CREATE INDEX

Question 16

4 / 4 pts

A table has a field "Major" that can have multiple values separated by commas. This violates _____.

☒ 1st Normal Form

☐ 2nd Normal Form

☐ 3rd Normal Form

☐ referential integrity

Question 17

4 / 4 pts

What are some common problems with poorly designed databases?
Check all that apply.



database contains a single table that mimics a data set object or spreadsheet

☒ failure to deal with keywords and/or categories properly

☒ database contains repeated information

☐ database contains too many tables

Question 18

5 / 5 pts

What is the correct order of the steps necessary to scrape data from a web page?

1

retrieve the web page

2

scan the HTML tree for

3

extract the data from the

4

clean the data

Question 19

5 / 5 pts

Which of these packages is required to retrieve documents via HTTP?

☒ RCurl

☐ Rhttp

☐ libURL

☐ XML

Question 20

3 / 3 pts

The data below is untidy. Specify the operations that need to be applied to transform it to tidy data.

StudentId	CS2500	CS1800	CS2510
001	A	A	A
002	B	B	B
003	B	A	B

☒

gather the course names, since the course columns contain values of the course variable.

☐ gather followed by scatter

☐ spread the course names

- ☐ scatter by spread of the course names.

Question 21**3 / 3 pts**

The dataset below tracks the number of students who added and dropped a class during a specific semester. The data is untidy. What operation needs to be done to transform the data into a tidy dataset?

Course	Operation	Count	Semester
CS1800	Add	75	Fall 2017
CS1800	Drop	30	Fall 2017
CS1800	Add	100	Spring 2017
CS1800	Drop	10	Spring 2017

- ☒ Spread the Operation variable, since each observation (course within a semester) should track the number of students that added the course and the number of students that dropped the course.
- ☐ Spread the Count variable, since each observation (course within a semester) should track the number of students that added the course and the number of students that dropped the course.
- ☐ Spread the Operation variable, since each observation (course within a semester) should track the number of students that added the course and the number of students that dropped the course. Followed by a gather of the Count variable



Spread the Count variable, since each observation (course within a semester) should track the number of students that added the course and the number of students that dropped the course. Followed by a gather of the Operation variable.

Question 22**3 / 3 pts**

What does the following R expression do : `str_detect(Sentences, "[0-9]+")`



It returns a vector of character strings having the same length as Sentences, where each element in the returned vector is NA or the first substring of contiguous digits found in the corresponding string in Sentences.



It returns a vector of logical values, where each element is TRUE if the corresponding element in Sentences has at least one digit otherwise it returns FALSE for the corresponding element



It returns a vector of characters, of all the contiguous digits found in the strings in Sentences



It returns all elements of Sentence that contain at least one digit

Question 23**3 / 3 pts**

The command `str_subset(vocabulary, "^x")`

☐

returns a logical vector where the value is set to TRUE if the corresponding element in the vocabulary vector ends in the letter x.

☐

returns a logical vector where the value is set to TRUE if the corresponding element in the vocabulary vector begins with the letter x.

☐

returns a string vector containing all of the elements from the vocabulary vector ending in the letter x.

☒

returns a string vector containing all of the elements from the vocabulary vector beginning with the letter x.

Quiz Score: **100** out of 100