Midterm Exam

HS616

March 26, 2015

Question 1

Which of the following is a common problem with messy datasets?

- A : Data is in the third norm form
- B : Data is in human readable format
- C : Primary and foreign keys are well defined
- D : One entity is stored in multiple tables

Question 2

When working with databases through R on your local computer, what is the advantage of working with SQLite instead of MySQL?

- A : SQLite is also suitable for a multi-user environment where hundreds of users connect to the database simultaneously.
- B : SQLite uses a flat file, as opposed to requiring a database connection.
- C : There isn't an advantage because there is no way to connect to a SQLite database in R.
- D : There are no packages to connect to a MySQL database in R.

Question 3

Generate a sequence from 122 to 154 by intervals of 2.

- A : seq(2,154,122)
- B : seq(122,2,154)
- C : seq(122,154,2)
- D : seq(154,122,2)

Question 4

Consider the following profiling results:

```
             self.time self.pct
"function_A"    278.39    86.46
"function_B"     29.32     9.10
"function_C"     14.29     4.44
```

If you make `function_B` 100 times faster, how much faster would you expect the program be?

- A : twice as fast
- B : 100 times as fast
- C : no faster
- D : less than 10% faster

Question 5

What needs to be changed in the following code for values to be arranged row wise in ascending order?

```
m <- matrix(1:20, nrow=5, ncol=4)
```

- A : `byrow = TRUE`
- B : `bycol = TRUE`
- C : No change required in the code
- D : `byrow = FALSE`

Question 6

The function head() does this:

- A : creates a header in the data frame
- B : displays the first few observations of a data frame
- C : summarizes the data in a table

Question 7

Which statement below best describes "natural join"?

- A : "natural join" uses an obviously similar column for join.
- B : "natural join" keeps only the information from second table if available
- C : "natural join" keeps only records in first table
- D : SQL does not support natural join

Question 8

xtab() does the following:

- A : crosstabulates variables
- B : is similar to table()
- C : all answers are correct
- D : can be used to easily generate a sparseMatrix

Question 9

Consider this R code showing two ways of calculating the cost of daily medicine, and select the true statement.

```
price <- c( lisonopril=106/30, crestor=204.00/30,
             clorthiazide=12.10/15, fibrosol=160/30)
dosage_day <- c( lisonopril=3, crestor=0.5,
            clorthiazide=0.5, fibrosol=1)
cost_day_1  = sum(price * dosage_day)
cost_day_1a = price %*% dosage_day
```

- A : A diagonal times a vector of that diagonal results in a squared value
- B : R is fun only for statisticians
- C : A vector times a vector is a scalar
- D : The Dot product of 2 vectors equals the sum of the element-wise products of the vectors

Question 10

To plot variables `x` and `y` along the x-axis and y-axis, respectively, one could use `plot(x, y)`. What is an alternative command that generates the same plot?

- A : `plot(x ~ y)`
- B : `plot(y % x)`
- C : `plot(y ~ x)`
- D : `plot(x %>% y)`

Question 11

The standard normal distribution has a mean of 0 and a standard deviation of 1, and the area under this curve over all possible x-values is one. What is the area under the curve of a normal probability distribution function with a standard deviation of 2?

- A : 2 pi
- B : 1
- C : 4
- D : 2

Question 12

`A` is an n by n square matrix. Identify the correct code used for augumenting the matrix A by binding an identity matrix on the right?

- A : `cbind(A, diag(1))`
- B : `rbind(A, diag(n))`
- C : `outer(A, diag(n), "+")`
- D : `cbind(A, diag(n))`

Question 13

Consider the following code, then select the correct statement regarding it.

```
maxMinusMin <- function(v, ...) max(v, ...) - min(v, ...)
apply(A, 1, maxMinusMin, na.rm=TRUE)
```

- A : If additional parameters are given to the function, they will be passed to `max` and `min`
- B : It's an invalid function that will need more parameters
- C : Typing error
- D : Function is invalid and cannot be executed

Question 14

In database management, what is meant by "Data Aggregation"?

- A : Using an inner join to extract data from a table
- B : Finding the mean of columns in a database table
- C : Normalizing the data in a database table
- D : The process by which data is gathered and summarized for further statistical analyses

Question 15

"Setting the seed", e.g. `set.seed(42)`, in R…

- A : ensures that the outcome of random number generators is *not* repeated upon re-execution of your code.
- B : ensures that someone else who runs your code does not get the same random numbers you do.
- C : has nothing to do with random number generation.
- D : ensures that the outcome of random number generators will be repeated upon re-execution of your code.

Question 16

sqldf is a fantastic tool for data scientists. Which of the following statements are true?

- A : sqldf is a useful tool for manipulation data with such statements such as: sqldf::sqldf("SELECT * FROM A JOIN B ON a=b")
- B : All of these
- C : sqldf operates on dataframes
- D : Right and full outer joins, which are unavailable in sqldf, can be accomplished with the "merge" function of base R

Question 17

If `x <- 1:4` and `y <- 5:8` what is the output of `x + y` ?

- A : A vector with values 6 8 10 12
- B : Running the statement gives an error
- C : A numeric integer with value 6
- D : A numeric integer with values 6 8 10 12

Question 18

Which of these addresses cannot be read by the built-in `url()` function?

- A : `http://ftp.ics.uci.edu/pub/machine-learning-databases/`
- B : `https://connect.usfca.edu`
- C : `http://rseek.org/`
- D : `file:///usr/share/dict/words`

Question 19

Which symbol can be used for slicing and extracting data from a vector in R?

- A : `[, c( )]`
- B : `[ ]`
- C : `$`
- D : `[[c( ) ]]`

Question 20

Which of these is not a problem with messy data

- A : Multiple variables stored in a single column
- B : Variables stored in both rows and columns
- C : Values stored in table format
- D : Multiple types of entities in the same table

Question 21

How does an ellipsis behave as a function parameter in R?

- A : Each period acts as an anonymous parameter in the function.
- B : It takes an undefined number of arguments and applies them wherever the ellipsis is used in the function, similar to a normal parameter.
- C : It takes each argument passed in by the user and applies them to undefined variables in the function based on order.

Question 22

Simulated coin-tossing can be done using different methods. Which of the following will NOT work?

- A : coin <- sample(c(“H”, “T”), 10, replace = F)
- B : ifelse(rbinom(10, 1, .5) == 1, “H”, “T”)
- C : c(“H”, “T”) [1 + rbinom(10, 1, .5)]
- D : rbinom(10, 1, .5)

Question 23

What is the name of the R function that does the equialent of SQL joins?

- A : sqlJoin
- B : join
- C : aggregate
- D : merge

Question 24

A Poisson distribution is defined as:

- A : None of the above. These answers are terrible.
- B : The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.
- C : The discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success.
- D : Is a very commonly occurring continuous probability distribution—a function that tells the probability that any real observation will fall between any two real limits or real numbers, as the curve approaches zero on either side.

Question 25

What does the `magrittr` library do in `R`?

- A : It makes the operator `%>%` in an `R` script act a lot like the pipe character `|` , in the Unix terminal, though it is not the pipe character.
- B : It allows one to use the pipe character, `|`, in an R script in the same way that it is used in the terminal.
- C : It offers a more confusing alternative to the typical passing of arguments into a function, using the `$` character. For this reason, it is becoming less popular among programmers.

Question 26

What is the correct way to vectorize the following code:

```
for(i in 1:3) x[i] <- i+i
```

- A : `x <- c(1,2,3) + c(1,2,3)`
- B : `for(i in range(1,4)) x+= [i+i]`
- C : `for(i<4) x[i] <- 2i`
- D : `while(i<4) x+= [2i]`

Question 27

Which of the following function keeps track of the function stack and tabulates how much time is spent on each function?

- A : RProf()
- B : rnorm()
- C : system.time()
- D : runif()

Question 28

What is the correct code for subtracting two dates from one another and then cast the difference to a numeric value?

- A : `as.numeric %>% (as.Date("2014-10-10" - "2014-10-1" ))`
- B : `as.Date("2014-10-10") - as.Date("2014-10-1") %>% as.numeric`
- C : `(as.Date("2014-10-10") - as.Date("2014-10-1")) %>% as.numeric`
- D : `as.Date %>% ("2014-10-10") - as.Date %>% ("2014-10-1") >%> as.numeric`

Question 29

Which keyword is used in a SQL select statement to eliminate duplicate values within a column?

- A : DISTINCT
- B : ONLY
- C : DIFFERENT
- D : can use '*'

Question 30

How may rows are returned by the following query?

```
A <- data.frame(a=1:10)
B <- data.frame(b=5:15)
sqldf::sqldf("SELECT * FROM A JOIN B ON a==b")
```

- A : 10
- B : none
- C : 6
- D : 8

Question 31

There is more than one way to "multiply" vectors. In R, `A * B` performs elementwise multiplication. What operator would you use to get the dot product of vectors `A` and `B`?

- A : `A ** B`
- B : `A %*% B`
- C : `A %>% B`
- D : `A & B`

Question 32

In the following code, what is the type of the variable returned?

```
y <- c(5, 6, 7, 8, NA)
is.na(y)
```

- A : logical
- B : numeric
- C : character
- D : integer

Question 33

What is TRUE of the following code?

```
T_shirts <- data.frame(
  sex=sample(c("M","F"), 100, replace=T),
  size=sample(c("L", "M", "S"), 100, replace=T)
)
```

- A : Always result in the same data
- B : Always result in males having more large sizes
- C : Only sometimes result in the same data, as the code does not identify a seed.
- D : Always result in females having more small sizes

Question 34

What is typically the fastest way to analyze and manipulate data using R?

- A : With vectorized functions
- B : With loops
- C : Using iteration
- D : With recursion

Question 35

Consider the following code:

```
N <-10000
x <- runif(N)
y <- runif(N)
vlength <- sqrt(x^2 +y^2)
in_circle <- vlength < 1
```

Which of the following could be the output of `head(as.integer(in_circle))` ?

- A : `1 1 1 1 1 0`
- B : `1 -1 1 0 -1 0`
- C : `0.23, ,0.34, 0.12, 0.45, 0.55, 0.79`
- D : `TRUE TRUE TRUE TRUE TRUE FALSE`