Preview Test: Mid-semester Mock Test from 2019

Test Information

Description	This is a Mid-semester mock test for CITS4009 Computational Data Analysis. It is the online version of the 2019 mid-semester test paper. To get yourself familiar with the open-book online test on LMS, you should try this test paper before looking at the answers.
Instructions	You have **90 minutes** to complete the test. The timer should start to count down after you click the **begin** button. As this is a mock test, if the timer doesn't work, then use your own timer. The password for opening the test paper is: **mock-test-2019** You can attempt this test 3 times.
Timed Test	This test has a time limit of 1 hour and 30 minutes.This test will save and be submitted automatically when the time expires. Warnings appear when **half the time**, **5 minutes**, **1 minute**, and **30 seconds** remain.*[The timer does not appear when previewing this test]*
Multiple Attempts	This test allows 3 attempts. This is attempt number 1.
Force Completion	Once started, this test must be completed in one sitting. Do not leave the test before clicking **Save and Submit**. Your answers are saved automatically.

Question Completion Status:

QUESTION 1 — 1 points — Save Answer

In data science, which one of the following statements is **NOT** valid.

- a. Big data problems could be small data problems in disguise.
- b. The first thing a data scientist needs to do is to define the goal of a new data science project.
- c. In R, the = sign and <- are semantically different. why = and <- is different?
- ✓ d. *Hypothesis confirmation* is the core of data science, not *hypothesis generation*. what is hypothesis comfiremation vs generation?

QUESTION 2 — vector vs list? — 1 points — Save Answer

Which type of R collections is created after the following statement?

```
a <- c(a="Mary",age=28,married=F)
```

- a. It won't work as variable a is used twice.
- b. It won't work as inputs are of different types.
- ✓ c. A vector.
- d. A list.

QUESTION 3 — 1 points — Save Answer

Given a list of values, x, what does the following function do?

```
secret <- function(x) {
   start <- floor(min(x))
   end <- ceiling(max(x))
   counts <- c()
   for (i in start:end-1) {
       counts <- c(counts, length(x[x > i & x <= i+1]))
   }
   return(counts)
}
```

- a. It returns a list of indices for all the integer values in x.
- ✓ b. It counts the number of data points falling into each integer bin. what is integer bin? What does the formula in length() do
- c. It calculates the total number of integer values in x.
- d. It returns a list of indices for lower quartile, median, and upper quartile.

QUESTION 4 1 points Save Answer

Which of the following geom(s) are suitable for visualising two continuous variables?

1) `geom_scatterplot`
2) `geom_hex`
3) `geom_smooth`

how hex and smooth looks like in graph?

- a. 1) only
- b. 2) only
- c. 1) and 3)
- ✓ d. 2) and 3)

QUESTION 5 1 points Save Answer

The table below shows the first eight rows of the data frame (`df`). It is about student enrolment at two universities (`UniA` and `UniB`) and their ATAR score.

what is subset?

	id	degree	uni	atar
1	S1	Science	UniA	90
2	S2	Art	UniB	60
3	S3	Engineering	UniB	50
4	S4	Engineering	UniA	40
5	S5	Art	UniB	70
6	S6	Engineering	UniA	56
7	S7	Business	UniB	95
8	S8	Science	UniA	80

Which one of the following code is equivalent to the statement below?

```
a <- df[df$uni=="UniA" & df$atar < 50,]$degree
```

- a. `a <- df[uni=="UniA" & atar < 50, degree]`
- b. `a <- df[df$uni=="UniA" && df$atar < 50, "degree"]`
- c. `a <- ifelse(df$gender=="M" & df$atar < 50, df$degree, "")`
- ✓ d. `a <- subset(df, uni=="UniA" & atar < 50, select="degree")[,1]`

QUESTION 6 1 points Save Answer

Which of the following statements about producing suitable plots is `FALSE` for comparing the atar distribution of the two universities described in Question 5?

- ✓ a. The only way is to split the df into two subsets, one for each uni, then use `geom_boxplot` twice.
- b. You can use "`y=atar, group=uni`" aesthetic mapping in `geom_boxplot`.
- c. You can use "`y=atar, fill=uni`" aesthetic mapping in `geom_boxplot`.
- d. You can use "`y=atar, color=uni`" aesthetic mapping in `geom_boxplot`.

QUESTION 7 1 points Save Answer

The `boxplot.stats` function for the `atar` variable in Question 5 gives the following values:

```
> boxplot.stats(df$atar)[1]$stats
[1] 40 53 65 85 95
```

what is [1] output?

Which one of the following statements is a correct interpretation of these values?

- ✓ a. Atar below 40 and above 95 are outliers. definition of outliner?
- b. 95 and 40 are the maximum and minimum values of atar, respectively.
- c. 25% of the students have atar lower than 65.
- d. The mean atar of all students is 65. is 65 is median?

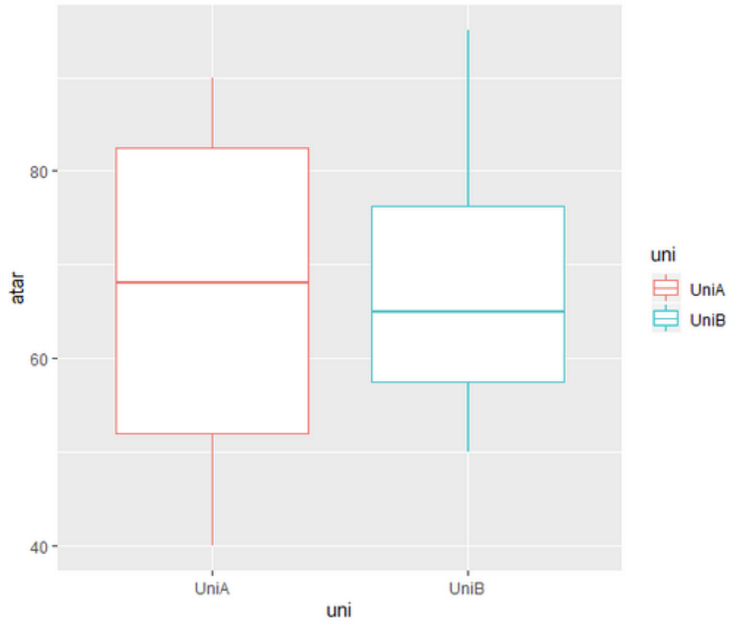
QUESTION 8 1 points Save Answer

Which geom is most suitable for displaying information stored in the two variables `degree` and `uni` for the data frame described in Question 5?

- a. histogram (or `hist`)
- ✓ b. `geom_tile`
- c. `boxplot`
- d. `hexbin` what is hexbin/geom_tile do?

QUESTION 9 1 points Save Answer

Given the boxplot below:



which observatioin is the LEAST sensible?

- a. UniA have higher median atar than UniB.
- b. UniB has more outliers than UniA. ✓
- c. No outliers are identified for either UniA or UniB.
- d. UniA has a wider spread of atar than UniB.

QUESTION 10 1 points Save Answer

Which of the following code will insert a new logical column to the data frame to record a status of meeting the atar cut off of 80.

- a. `df$meet_atar <- as.factor(df$atar >= 80)`
- b. `df$meet_atar <- ifelse(df$atar >= 80, "TRUE", "FALSE")`
- c. `df <- within(df, {meet_atar<-FALSE; meet_atar <- atar>=80})` ✓
- d. `df$atar <- df$atar >= 80`

what is within() and why a and b is incorrect?

QUESTION 11 1 points Save Answer

What does the code below do, where `df` is the original data frame described in Question 5?

```
myvars <- names(df) %in% c("id", "uni", "atar")
newdf <- df[!myvars]
```

what is names()?

- a. It creates a new data frame that contains only the values of the `degree` variable. ✓
- b. It creates a new data frame that contains the values of all columns except for the `degree` variable.
- c. It won't work because you cannot start a R variable name with a `%` sign.
- d. It won't work because you cannot start a R variable name with a `!` symbol.

what is %in%?

QUESTION 12 1 points Save Answer

Again, refer to the original data frame `df` described in Question 5.
What does the following code do?

```
mean_atar <- aggregate(df$atar, list(df$degree), mean)
merge(df, mean_atar, by.x="degree", by.y="Group.1")
```

- a. It works out the mean of each degree.
- b. It counts how many students are enrolled for each degree.
- ✓ c. It outputs a data frame like `df` with an extra column containing the mean atar calculated for the respective degree for each record.
- d. None of the above.

QUESTION 13 1 points Save Answer

Which statement about *listwise deletion* to handle missing data is `TRUE`?

- a. It is referring to deleting the columns containing `NA` values using `na.omit()`. na omit delete row
- ✓ b. When the `NA`s tend to be for the same observations, and are of a small proportion of the dataset, drop those rows.
- c. When the missing data are a result of sensor errors.
- d. When the data are missing systematically.

what should we deal with data missing systemtically?

QUESTION 14 1 points Save Answer

Which one of the following is not valid for imputing missing numerical data?

- a. Use the mean of the variable.
- b. Use the median of the variable.
- c. Use other variables with available data to build a predication model.
- ✓ d. Use the z-normalisation of the variable.

what is the z normalisation of variable?

QUESTION 15 1 points Save Answer

Taking the `age` variable of the `custdata` used in the lectures for example, which of the following about the mean transformation `custdata$age/mean(custdata$age)` is true?

- a. There should be very few customers having a normalised age value of 1.1.
- b. A normalised age value that is close to 1 signifies an unusually old customer.
- c. The normalised age values are between -1 and 1.
- ✓ d. A normalised age value that is much less than 1 signifies an unusually young customer.

QUESTION 16 1 points Save Answer

In R, which one of the following statements about the `Date` data type is `TRUE`?

- ✓ a. R stores dates internally as the number of days since `1970-01-01`.
- b. There is no set reference date in R, it should be specified using the `date()` function.
- c. The default format for inputting dates is `dd/mm/yyyy`.
- d. The default date format in R depends on the countries you are in.

QUESTION 17 1 points Save Answer

How do you work out the age of this person on today?

```
dob <- as.Date("1956-10-12")
```

1) `as.double(difftime(Sys.Date(), dob, units="days"))`
2) `julian(Sys.time(), origin = dob)/365`

what is the julian() do?

- a. 1) only
- ✓ b. 2) only
- c. Both of them
- d. None of them

QUESTION 18 1 points Save Answer

What does the following code do?

```
df[sample(1:nrow(df), 3, replace=FALSE),]
```

- a. It won't work as it contains syntax errors.
- b. It generates a list of booleans with three `TRUE` values indicating the rows to be selected from `df`.
- c. It returns 3 records, randomly sampled from `df` with replacement.
- d. It returns 3 records, randomly sampled from `df` without replacement. ✓

QUESTION 19 1 points Save Answer

Given two data frames as show below, `units` and `students`, how do we add a new column to the `students` table with the right mapping of `unit_name`?

unit_code	unit_name
CITS4009	Data Science
CITS5508	Machine Learning
CITS1401	Python

name	gender	unit
John	M	CITS4009
Emma	F	STAT1400
Peter	M	CITS1401

- a. `students <- cbind(students, units); students <- students[-3]`
- b. `students <- merge(units, students, all.x=TRUE, by.x="unit", by.y="unit_code")`
- c. `students <- merge(students, units, all.x=TRUE, by.x="unit", by.y="unit_code")` ✓
- d. `students <- merge(students, units, all.y=TRUE, by.x="unit", by.y="unit_code")`

QUESTION 20 1 points Save Answer

Given the two data frames, `units` and `students`, where the key columns are `unit_code` and `unit` respectively. What is the number of records in the new data frame after an inner join?

unit_code	unit_name
CITS4009	Data Science
CITS5508	Machine Learning
CITS1401	Python

name	gender	unit
John	M	CITS4009
Emma	F	STAT1400
Peter	M	CITS1401

- a. 2 ✓
- b. 3
- c. 4
- d. 9