THE UNIVERSITY OF WESTERN AUSTRALIA

DESK No. | | | |

FAMILY NAME: _____

GIVEN NAMES: _____

SIGNATURE: _____

STUDENT NUMBER: | | | | | | | | |

SEMESTER 2, 2019 MIDSEMESTER EXAMINATIONS

CIT4009

Introduction to Data Science

Physics, Mathematics & Computing
Department of Computer Science & Software Engineering

This paper contains: **8** Pages **(including title)**

Time Allowed: **0:45** hours

INSTRUCTIONS:

This test consists of 20 multiple-choice questions.

Each question has four answer choices.

Read each question carefully, and select only ONE best answer.

The answers need to be marked clearly on the Multiple Choice Answer Sheet.

No answers on this question booklet will be considered.

THIS IS A CLOSED BOOK EXAMINATION

ALLOWABLE ITEMS

UWA Approved Calculator with Sticker

SUPPLIED STATIONERY

UWA Multiple Choice Answer Sheet

PLEASE NOTE

Examination candidates may only bring authorised materials into the examination room. If a supervisor finds, during the examination, that you have unauthorised material, in whatever form, in the vicinity of your desk or on your person, whether in the examination room or the toilets or en route to/from the toilets, the matter will be reported to the head of school and disciplinary action will normally be taken against you. This action may result in your being deprived of any credit for this examination or even, in some cases, for the whole unit. This will apply regardless of whether the material has been used at the time it is found. Therefore, any candidate who has brought any unauthorised material whatsoever into the examination room should declare it to the supervisor immediately. Candidates who are uncertain whether any material is authorised should ask the supervisor for clarification.

Candidates must comply with the Examination Rules of the University and with the directions of supervisors.
No electronic devices are permitted during the examination.
All question papers and answer booklets are the property of the University and remain so at all times.

This page has been left intentionally blank

Q1. In data science, which one of the following statements is *NOT* valid.

a) Big data problem could be a small data problem 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.
d) Hypothesis confirmation is the core of data science, not hypothesis generation.

Q2. Which type of R collections is created after the following statement?

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

a) Won't work as variable `a` is used twice.
b) Won't work as inputs are of different types.
c) A vector.
d) A list.

Q3. 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 bin.
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.

Q4. Which of the following geom(s) are suitable for visualising two continuous variables?

`1) geom_scatterplot  2) geom_hexbin  3) geom_smooth`

a) 1) only
b) 2) only
c) 1) and 3)
d) 2) and 3)

The following questions Q5-Q12 are related to this data frame (`df`) about student enrollment at two universities (`UniA` and `UniB`) and their ATAR score.

For ease of interpretation, the data frame can be viewed as a table below.

	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

Q5. 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")`

Q6. Which of the following statements about producing suitable plots is FALSE for comparing the `atar` distribution of the two universities?

a) The only way is to split `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`.

Q7. The boxplot stats give the following values,

```
> boxplot.stats(df$atar)[1]$stats

[1] 40 53 65 85 95
```

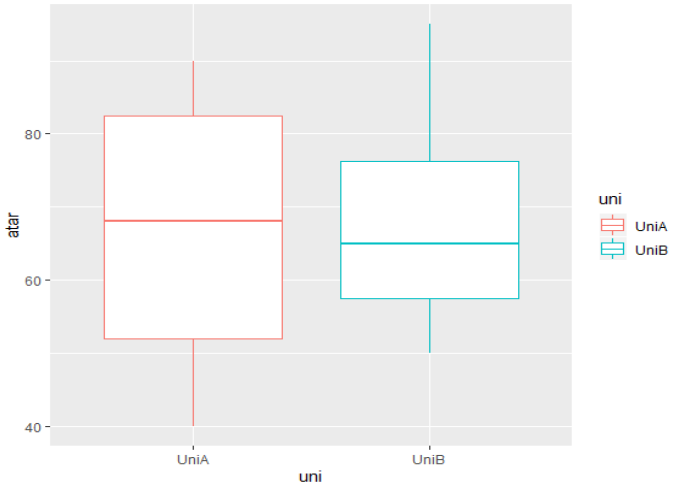
based on these five values alone, which one of the following statements is a correct interpretation?

a) `atar` below 40 and above 95 are outliers.
a) 95 and 40 is the maximum and minimum value of `atar`, respectively.
b) 25% of the students have `atar` lower than 65.
c) The mean `atar` of all students is 65.

Q8. Which geom is most suitable for displaying information in the two variables: degree and uni?

a) `histogram`
b) `geom_tile`
c) `boxplot`
d) `hexbin`

Q9. Given the above boxplot, which observation 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 Uni.
d) `UniA` has a wider spread of `atar` than `UniB`.

Q10. 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.`

Q11. What does the code below do?

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

a) It creates a new data frame with values of only the degree variable.
b) It creates a new data frame with values of every columns but the `degree` variable.
c) It won't work because you cannot start an R variable name with a `%` sign.
d) It won't work because you cannot start an R variable name with a `!` symbol.

Q12. 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 appends a new column with the mean `atar` calculated for the respective degree of that record.
d) None of the above

Q13. 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()`.
b) When the NAs 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.

Q14. 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.

Q15. 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) The normalised `age` closer to 0 signifies an unusually young customer.
b) The normalised `age` closer to 1 signifies an unusually old customer.
c) The normalised `age` is between -1 and 1.
d) The normalised `age` much less than 1 signifies an unusually young customer.

Q16. 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.

Q17. 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`

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

Q18. 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 binary 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.

The following two questions are related to the two data frames depicted in the following tables:

unit_code	unit_name
CITS4009	Data Science
CITS5508	Machine Learning
CITS1401	Python

Data Frame: units

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

Data Frame: students

Q19. Given the above data frame, `units` and `students`, how do we add a new column to the `students` table with the right mapping of `unit_name`?

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

Q20. Given the above data frame, `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**?

a) 2
b) 3
c) 4
d) 9