THE UNIVERSITY OF
WESTERN
AUSTRALIA

DESK No. □□□

FAMILY NAME: ____________________

GIVEN NAMES: ____________________

SIGNATURE: ____________________

STUDENT NUMBER: □□□□□□□

SEMESTER 2, 2018 EXAMINATIONS

CITS4009

Introduction to Data Science

Physics, Mathematics & Computing
EMS

.

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

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.

THIS IS A CLOSED BOOK EXAMINATION

SUPPLIED STATIONERY	**ALLOWABLE ITEMS**
1 x Answer Booklet 18 Pages	**UWA Approved Calculator with Sticker**

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) Hypothesis generation requires looking at data and apply subject-area knowledge.

b) A data science project is an iterative process.

[✓ c)] Hypothesis confirmation is also known as data exploration.

d) Models are often used for exploration.

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

```
a <- c(1,2,3)
```

[✓ a)] A vector.

b) A matrix.

c) A list.

d) An array.

method of create list/matrix/array?

Q3. Given a list of values, **x**, what does the following function do?

```
secret <- function(x){
  s <- sort(x)
  index <- floor(length(x)/2) + 1
  return(c(index, s[index]))
}
```

a) Return a list of indices and corresponding sorted values of **x**.

[✓ b)] The index and the value of median.

c) The indices and the value of lower and upper quartile.

d) A list of indices for lower quartile, median and upper quartile.

Q4. Given the following geom specifications, which are suitable for visualising a continuous variable (x) against a categorical one (y)?

```
1) geom_histogram(mapping = aes(x = x, color=y))
2) geom_histogram(mapping = aes(x = x), color=y) ✗
3) geom_boxplot(mapping = aes(x = x, y = y))
4) geom_boxplot(mapping = aes(x = y, y = x))  ✗
```

[✓ a)] 1) and 3)

b) 1) and 4)

c) 2) and 3)

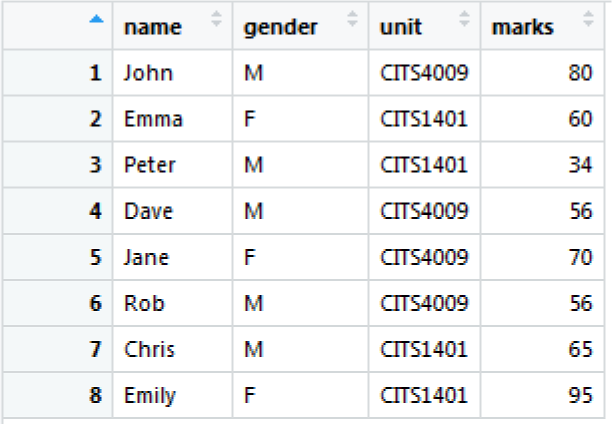
d) 2) and 4)

本题应该选b，因为boxplot的x轴应该是分类变量，y轴才是连续变量

The following questions Q5-Q12 are related to this data frame.

```
marks <- c(80,60,34,56,70,56,65,95)
name <- c("John", "Emma", "Peter", "Dave", "Jane", "Rob", "Chris", "Emily")
gender <- c("M","F","M","M","F","M","M","F")
unit <-
c("CITS4009","CITS1401","CITS1401","CITS4009","CITS4009","CITS4009","CITS1401","CITS14
01")
df <- data.frame(name=name, gender=gender, unit=unit, marks=marks)
```

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



	name	gender	unit	marks
1	John	M	CITS4009	80
2	Emma	F	CITS1401	60
3	Peter	M	CITS1401	34
4	Dave	M	CITS4009	56
5	Jane	F	CITS4009	70
6	Rob	M	CITS4009	56
7	Chris	M	CITS1401	65
8	Emily	F	CITS1401	95

Q5. Which one of the following code is equivalent to the statement below?

此题应该选a。由于subset只取出子集，a选项1代表第1列，可以；b选项name要双引号；cd选项最终显示出来一堆NA和""

```
a <- subset(df, gender==”M” & marks < 50, select=”name”)
```

a) `a <- df[gender=="M" & marks < 50, 1]` ✗

b) `a <- df[df$gender=="M" & df$marks < 50, name]` name要双引号?

[✓ c)] `a <- ifelse(df$gender=="M" & df$marks < 50, df$name, NA)`

d) `a <- ifelse(df$gender=="M" & df$marks < 50, df$name, “”)` ?

Q6. Which of the following plots will *NOT* help compare the number of male and female students in the two classes?

选a，因为六边形图需要两个连续变量才能构图，而tile热力图和条形图均能用于离散变量

[✓ a)] A hexbin plot

b) A tile plot

c) A side-by-side bar chart

d) Bar charts of gender faceted by unit code

Q7. The boxplot stats give the following values, which one of the following statements is ~~true~~ false?

```
> boxplot.stats(df$marks)
    $`stats`
    [1] 34.0 56.0 62.5 75.0 95.0
```

a) 95.0 is the maximum value of marks.

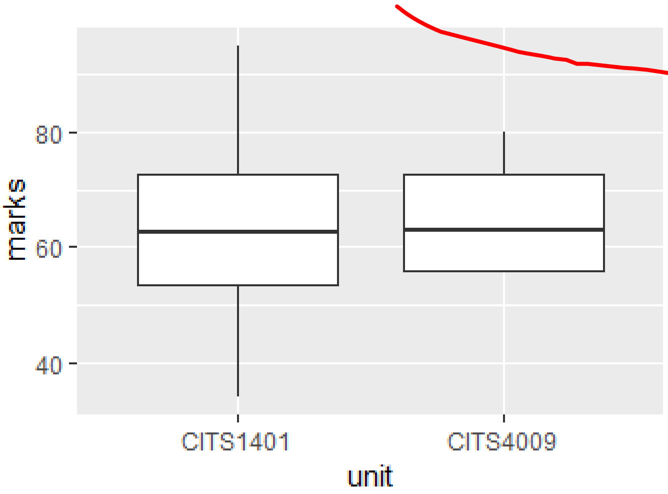
b) 34.0 is the minimum value of marks.

c) 75% percent of the students have marks lower than 75.0.

[✓ d)] The average of all students is 62.5.

Q8. Which one of the following will produce the boxplots in the figure below?

```
1) ggplot(df) + geom_boxplot(mapping=aes(x=unit, y=marks))
2) boxplot(df[df$unit=="CITS1401",]$marks, df[df$unit=="CITS4009",]$marks)
```



表示所有等于CITS1401的行取出marks那一列

a) Only 1).

b) Only 2).

[✓ c) Both 1) and 2).]

d) None of them.

Q9. Given the above boxplot, which observation is the *LEAST* sensible?

a) The two class have roughly the same median marks.

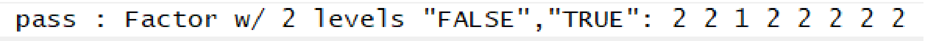
b) The distributions of marks in the two classes are similar.

c) Both units have no outlier marks.

[✓ d) CITS4009 have higher average because of no failed students.]

c选项是对的，因为没有人分数超过100或者负数的，不算异常值；d选项是错的，不一定没挂科，也不一定平均分就更高(1401有好几个高分的)

Q10. Which of the following code will insert a new column to the data frame to record a logical pass or fail status.



```
pass : Factor w/ 2 levels "FALSE","TRUE": 2 2 1 2 2 2 2 2
```

a) `df$pass <- as.factor(df$marks > 50)`

b) `df$pass <- ifelse(df$marks > 50, “TRUE”, “FALSE”)`

字符类

[✓ c) `df <- within(df, {pass=NA; pass=marks>50})`]

logi类

d) All of the above.

Q11. What does the code below do?

```
df[order(df$unit, -df$marks),]
```

a) It won’t work because you cannot have the negative sign before a column name.

b) It won’t work because it has an extra comma after the order function.

c) It sorts the data according to unit in ascending order, but remove the marks column.

[✓ d) It sorts the data first according to unit in ascending order, then marks in descending order.]

Q12. Which one is the output of the following statement?

```
aggregate(df$marks, list(df$unit), mean)
```

```
1) Group.1     x
  1 CITS1401 63.5
  2 CITS4009 65.5

2) Group.1     x
   1 CITS4009 65.5
   2 CITS1401 63.5
   3 CITS1401 63.5
   4 CITS4009 65.5
   5 CITS4009 65.5
   6 CITS4009 65.5
   7 CITS1401 63.5
   8 CITS1401 63.5

3) Group.1  mean
   1 CITS1401 63.5
   2 CITS4009 65.5
```

[✓ a)] 1)

b) 2)

c) 3)

d) None of the above

Q13. When can we use `na.omit()` to remove missing data?

[✓ a)] Only when the NAs are for roughly from the same data points and are of a small proportion of the dataset.

b) Only when the NAs are concentrated for certain variables, and are of a small proportion of the dataset.

c) When the missing data are a result of sensor errors.

d) When the data are missing systematically.

Q14. When should we consider the use of `average` to impute missing data?

a) When the data are missing systematically.

b) When the missing data are of integer type.

[✓ c)] When the missing data are concentrated on certain rows and possibly due to sensor errors.

d) When the missing data are concentrated on certain columns.

Q15. Taking the `income` variable of the `custdata` used in the lectures for example, it is highly skewed towards one side in a density plot. Which would *NOT* be a sensible transformation?

a) Use domain knowledge to set a threshold for outlier removal.

b) Use the stats returned from `boxplot.stats()` for outlier removal.

c) Apply `log10` transformation.

[✓ d)] Use z-normalisation to turn it into values between -1 and 1.

Q16. In R, what is the default reference date used for internal date storage?

a) 1900-01-01

[✓ b) 1970-01-01]

c) There is no set reference date, it should be specified using the `date()` function.

d) This depends on the operating system you use.

Q17. Given the following code, what is the likely value of `dob_num?`

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

[✓ a) -4829]

b) 4829

why this is negative?

c) "1956-10-12"

d) None of the above

Q18. How do you work out the age of this person on today?

```
dob <- as.Date("1956-10-12")
1) as.double((Sys.Date()-dob))/365
2) as.double(difftime(Sys.Date(), dob, units="days"))/365
```

a) 1) only

b) 2) only

[✓ c) Both of them]

d) None of them

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

surname	nationality	deceased
Tukey	US	yes
Venables	Australia	no
Tierney	US	no
Winton	UK	no

Data Frame: authors

name	title	other.author
Tukey	Exploratory Data Analysis	NA
Venables	Modern Applied Statistics	Ripley
Tierney	LISP-STAT	NA
Ripley	Spatial Statistics	NA
Ripley	Stochastic Simulation	NA
McNeil	Interactive Data Analysis	NA
R Core	An Introduction to R	Venables & Smith

Data Frame: books

Q19. Given the above data frame, `authors` and `books`, where the key columns are `surname` and `name` respectively. What is the number of columns in the new data frame after mutable joins?

[✓ a) 5]
b) 6
c) 3
d) None of the above

what is mutable join

Q20. Given the above data frame, `authors` and `books`, where the key columns are `surname` and `name` respectively. What is the number of records in the new data frame after full join?

全部算上，一个不落

a) 3
b) 4
c) 7
[✓ d) 8]

what is full join?