MONASH University

Office Use Only

Semester One 2019
Examination Period

Faculty of Business and Economics

EXAM CODES: **ETC1010**

TITLE OF PAPER: **DATA MODELLING AND COMPUTING - Paper 1**

EXAM DURATION: 2 hours writing time

READING TIME: 10 minutes

THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)

☐ Caulfield ☑Clayton ☐ Parkville ☐ Peninsula
☐ Monash Extension ☐ Off Campus Learning ☐ Malaysia ☐ Sth Africa
☐ Other (specify)

During an exam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, mobile phone, smart watch/device, calculator, pencil case, or writing on any part of your body. Any authorised items are listed below. Items/materials on your desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession.

No examination materials are to be removed from the room. This includes retaining, copying, memorising or noting down content of exam material for personal use or to share with any other person by any means following your exam.

Failure to comply with the above instructions, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations, or a breach of instructions under Part 3 of the Monash University (Academic Board) Regulations.

AUTHORISED MATERIALS

OPEN BOOK	☐ **YES**	☑**NO**
CALCULATORS	☑**YES**	☐ **NO**
Only HP 10bII+ or Casio FX82 (any suffix) calculator permitted		
SPECIFICALLY PERMITTED ITEMS	☐ **YES**	☑**NO**
if yes, items permitted are:		

STUDENTS SHOULD NOT USE AN EXAM BOOKLET, AS THEY ARE REQUIRED TO WRITE THEIR ANSWERS ON THIS PAPER.

Candidates must complete this section.

STUDENT ID: ..... DESK NUMBER: .....

Instructions

There are 9 questions worth a total of 100 marks. You should attempt all of the questions.

QUESTION 1

This question is about tidy data.

species	location	class	count
emu	wilsons prom	aves	40
emu	grampians	aves	20
wombat	wilsons prom	mammalia	30
wombat	grampians	mammalia	10
wallaby	wilsons prom	mammalia	50
wallaby	grampians	mammalia	50

(a) There are three variables in this data set. Explain why this is, that is, why count is not considered a variable.

[3 marks]

What was measured at each wildlife sighting was species, class and location. Count is a statistic that was computed after the complete data set was collected, and provides a convenient way to store the data with less rows.

(b) There are 200 observations in the data. Explain why, that is, why the answer is not 6.

[3 marks]

There were 200 wildlife sightings in the data. The 6 rows is a tabular summary of the combinations of levels of the observed variables.

(c) What type of variable would location be considered to be, and what type of format would it be treated as in R? (Circle two, one from each line)

[3 marks]

quantitative qualitative temporal spatial

qualitative

numeric double logical character ordered factor

character

[Total: 9 marks]

— END OF QUESTION 1 —

QUESTION 2

This question is about data wrangling.

For each of these questions, write down the verbs that you would need to use to do the calculations to answer it using the nycflights13 data. Add explanations of why you would approach it that way if you feel it is necessary.

```
> glimpse(flights)
Observations: 336,776
Variables: 15
$ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013...
$ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 55...
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2,...
$ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 8...
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7,...
$ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6"...
$ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301...
$ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N...
$ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LG...
$ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IA...
$ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149...
$ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 73...
$ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6...
```

(i) What is the typical daily number of flights that United flies out of JFK between 8am and 9am?

[3 marks]

filter records to only keep United flights, for JFK between 8-9am, summarise to compute the average number of flights.

(ii) What hour of day should you plan to fly if you want to avoid delays as much as possible?

[3 marks]

group by hour, summarise to compute average delay, and then arrange by lowest to highest.

(iii) What plane (identified by their tail number) is responsible for the most delays?

[3 marks]

group by tailnum, summarise to compute average delay, and then arrange from highest to lowest

(iv) How do high winds impact flight operations at Newark airport?

[3 marks]

Can't be done with this table alone. Or the list of verbs provided! It would need to join the weather table, which does exist on the data set, with this data.

Full verb list: *filter, select, mutate, summarise, arrange, group_by, near, desc, starts_with, ends_with, contains, matches, rename, top_n, first, min_rank, lag, cumsum, count, tally*

[Total: 12 marks]

— END OF QUESTION 2 —

QUESTION 3

This question is about visualisation.

The hierarchy of mapping quantitative variables to plots, in order to return an accurate reading is as follows:

Mapping	Rank	Example
Area	4	mosaic plot
Shading, color	6	choropleth map
Position - nonaligned scale	2	side-by-side boxplot to compare the boxes
Volume, curvature	5	3D bubble charts??
Position - common scale	1	scatterplot
Length, direction, angle	3	piechart

(a) Fill in the rank column, labelling the the mappings 1 through 7, with 1 indicating this is the mapping method that analysts can most accurately read off the data values.

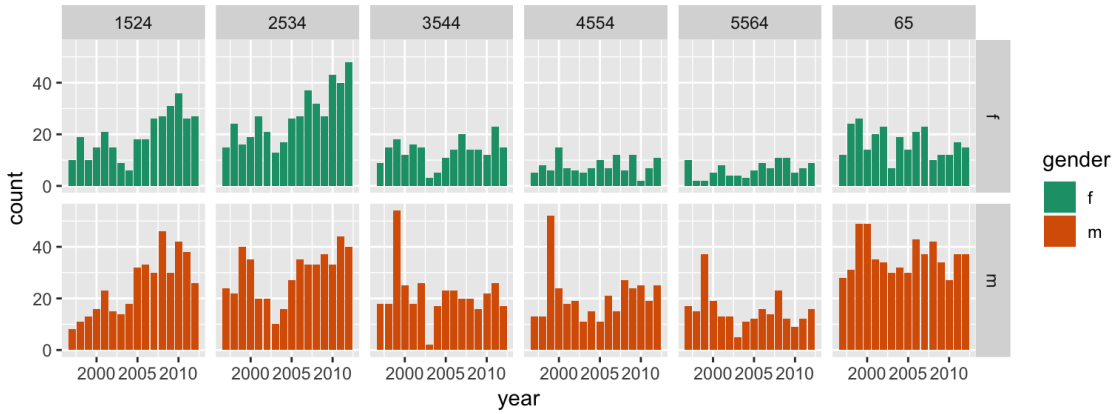
[4 marks]

(b) Given a plot type example, or a ggplot geom would also be acceptable, for each mapping. Write this into the third column.

[4 marks]

(c) The facetting in this plot may make some comparisons between groups difficult to make because of what cognitive perception issue. (Circle one)

[2 marks]



proximity colour blindness change blindness pre-attentive layering

change blindness

(d) Which variable(s) would be most affected by this? (Circle)

[2 marks]

gender age year

gender and age

[Total: 12 marks]

— END OF QUESTION 3 —

QUESTION 4

This question is about handling missing values.

(a) TRUE or FALSE: The shadow matrix can be thought of as a data table indicating where the original values are missing. TRUE

[2 marks]

(b) TRUE or FALSE: When you have missing values in a data table, it is reasonable to drop the observation before conducting your analysis. FALSE

[2 marks]

(c) Which method for imputing missing values is better in this scenario? (Circle one)

[3 marks]

There is a relationship between the variables, and all are quantitative. Several of the variables have missing values.

(i) use the mean of the complete values, separately for each variable.
(ii) set up several regression models, using the variable with missings as the response, and use complete cases to estimate.
(iii) simulate values from a normal distribution for each variable, using the mean and standard deviation of the complete cases.

ii

(d) Use the following regression model set up to impute the missing value on number of bedrooms, using complete information on number of rooms, for the Melbourne house price data, for this observation, Rooms = 8. Bedroom=NA.

[3 marks]

```
Coefficients:
(Intercept)        Rooms
    0.08523      0.96727
```

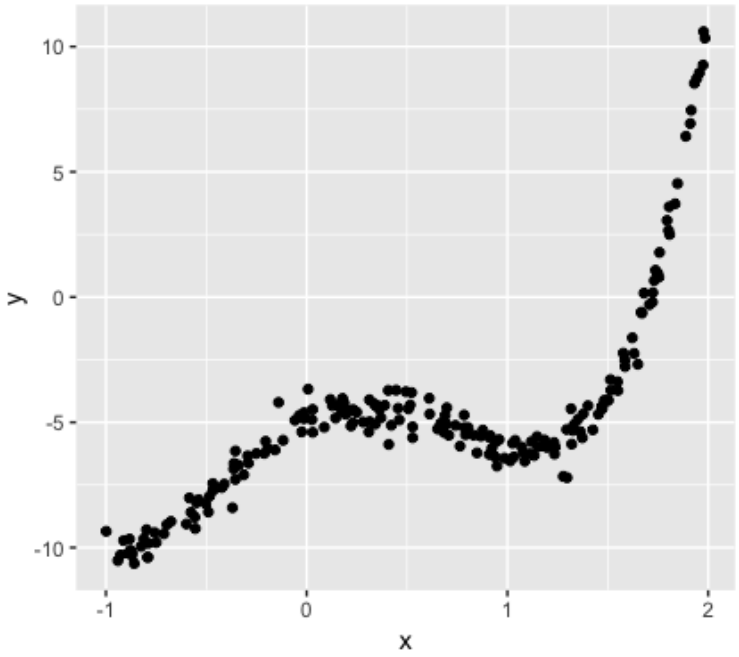
7.8

[Total: 10 marks]

— END OF QUESTION 4 —

QUESTION 5

This question is about model fitting and optimisation.



This is a simulated data set, with values generated by the code below.

```
library(tidyverse)
square_err <- function(par, data) {
  sq <- sum((data$y-(par[1]+par[2]*data$x+
                       par[3]*data$x^2+par[4]*data$x^3+
                       par[5]*data$x^4))^2)
  return(sq)
}
x <- runif(200, -1, 2)
df <- tibble(x, y=-5+4*x-6*x^2-2*x^3+3*x^4+rnorm(200)*0.5)
ggplot(df, aes(x=x, y=y)) + geom_point()

fit <- optim(c(1,1,1,1,1), square_err, data=df))
```

(a) Write down the true regression coefficients used to generate the data.

[4 marks]

(-5,4,-6,-2,3)

(b) The fitted model has these coefficients:

```
> fit$par
[1] -4.992065  3.901229 -6.064781 -1.844502  2.968980
```

Why is there a difference from the parameters used to generate the data?

[3 marks]

(These are computed on a sample of data so they are close to the parameter values but will never be exactly identical.)

(c) Write down the equation of the fitted model.

[4 marks]

$\hat{y} = -4.992065 + 3.901229x - 6.064781x^2 - 1.844502x^3 + 2.968980x^4$

(d) Predict the response value for $x = 2$.

[4 marks]

11.9

(e) If the observed value for $x = 2$ is 10, compute the residual for the fit.

[3 marks]

-1.29

(f) Why is the fitted value so much different from the observed value? Is this correct, or has there been a mistake? Explain.

[3 marks]

This is incorrect. The fitted value is close to the observed value. In this area of the predictor space it is possible to see a big difference, because there is a sharp increase in response value, but it turns out that the model predicts quite well here anyway.

[Total: 21 marks]

— END OF QUESTION 5 —

QUESTION 6

This question is about linear regression.

A simulated data set has two predictors, x_1, x_2 and one response variable (y), where x_2 is a categorical predictor with two levels, A, B. Below are summaries of two different model fits to the data.

Model I:

	term	estimate	std.error	statistic	p.value
1	(Intercept)	6.35	0.09	70.55	0.00
2	x_1	-1.65	0.08	-21.18	0.00
3	x_2B	3.02	0.14	22.21	0.00

Model II:

	term	estimate	std.error	statistic	p.value
1	(Intercept)	6.03	0.05	115.15	0.00
2	x_1	-1.03	0.05	-19.78	0.00
3	x_2B	4.01	0.09	44.96	0.00
4	x_1:x_2B	-1.96	0.09	-20.98	0.00

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
I	0.83	0.83	0.92	477.50	0.00	3	-264.77	537.55	550.74	165.37	197
II	0.95	0.95	0.51	1174.48	0.00	4	-147.06	304.12	320.61	50.96	196

(a) This is the equation for the fitted model I. There's a mistake, please fix it.

[2 marks]

$$\begin{aligned} y &= 6.35 - 1.65x_1 \quad \text{if } x_2 = A \\ &= 3.33 - 1.65x_1 \quad \text{if } x_2 = B \end{aligned}$$

3.33 should be 9.37. It should be $\hat{y}$

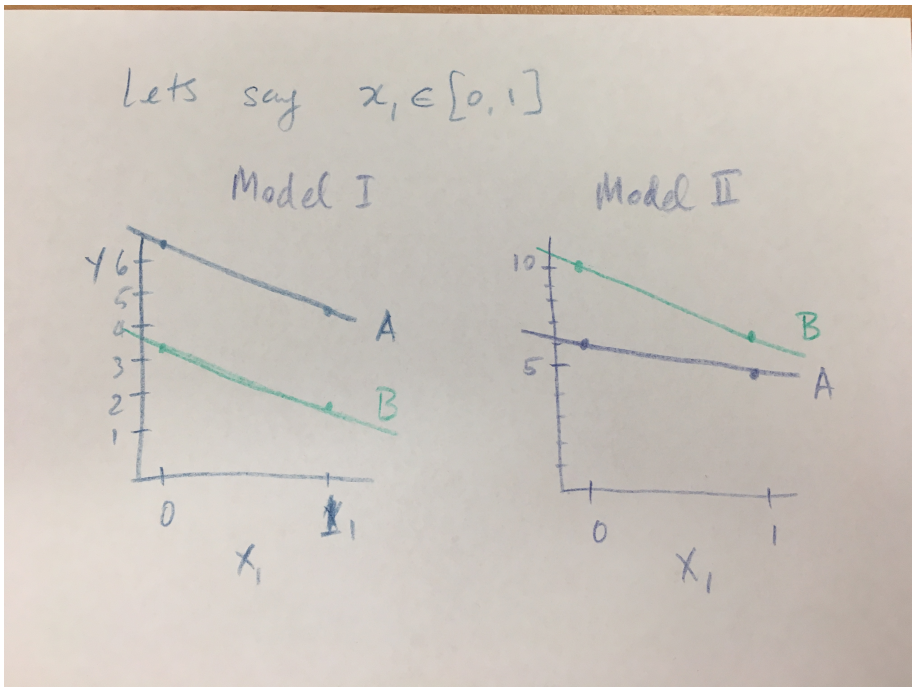
(b) Write down the equation for model 2, as simply as you can.

[3 marks]

$$\begin{aligned} \hat{y} &= 6.03 - 1.03x_1 \quad \text{if } x_2 = A \\ &= 10.04 - 2.99x_1 \quad \text{if } x_2 = B \end{aligned}$$

(c) Make a sketch of each model. Be sure to label your axes of the plots for the predictors and the response.

[4 marks]



(d) Which model fits the data better? Why?

[3 marks]

II, because R^2 is higher and deviance is smaller.

[Total: 12 marks]

— END OF QUESTION 6 —

QUESTION 7

This question is about text analysis.

(a) For this sentence:

Emma Woodhouse, handsome, clever, and rich, with a comfortable home.

(a) Indicate how it would be tokenised, for example using code like `unnest_tokens(word, text)`.
[3 marks]

Emma/ Woodhouse/ handsome/ clever/ and/ rich/ with/ a/ comfortable/ home

(b) Mark the words that would typically be considered to be "stop words".
[3 marks]

and with a

(b) Fill in the blank, from the word bank provided:
[3 marks]

The statistic "term frequency, inverse document frequency", is intended to measure how _____ a word is to a document in a collection (or corpus) of documents,

important different frequent relatively frequent common

important

[Total: 9 marks]

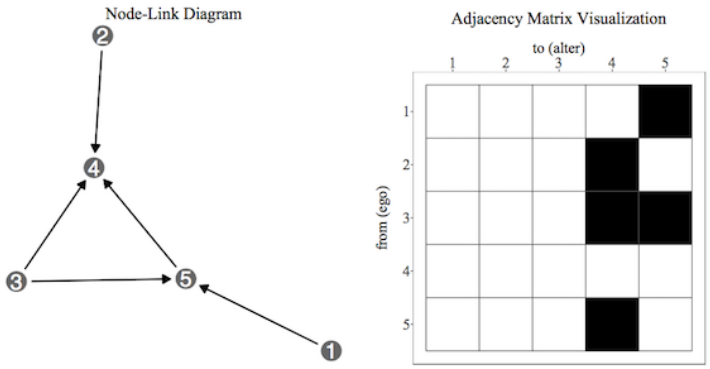
— END OF QUESTION 7 —

QUESTION 8

This question is about generating networks and their analysis.

(a) For the network visualisation show below, write down what the association matrix might be, using numbers.

[3 marks]



0	0	0	0	1
0	0	0	1	0
0	0	0	1	1
0	0	0	0	0
0	0	0	1	0

(b) For the association matrix below, showing the number of times that these five people have messaged each other in the last week.

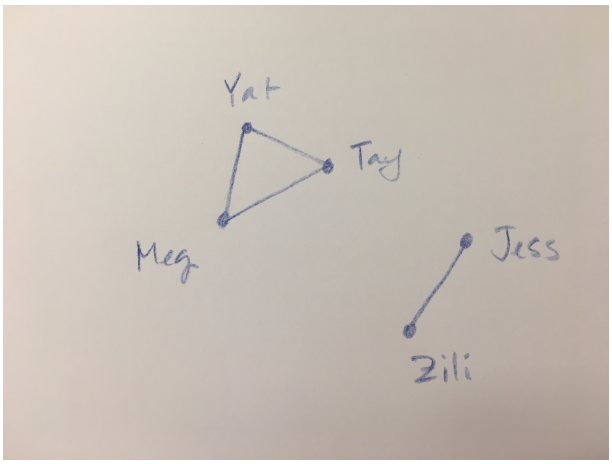
	Meg	Tay	Yat	Zili	Jess
Meg	0.00	5.00	4.00	1.00	1.00
Tay	5.00	0.00	4.00	2.00	1.00
Yat	4.00	4.00	0.00	0.00	0.00
Zili	1.00	2.00	0.00	0.00	6.00
Jess	1.00	1.00	0.00	6.00	0.00

(a) Which two people would be considered to be most similar (most connected)? Zili, Jess

[3 marks]

(b) Draw the network diagram, if you consider closeness (edges to be connected) to be 4 or more.

[3 marks]



[Total: 9 marks]

— END OF QUESTION 8 —

QUESTION 9

This question is about data collection practices.

In 2008, IBM, the most venerable computer (now data analysis) company declared vegemite to be the world's most recognisable brand.

Research conducted by IBM showed that Vegemite outranked other giants like Coca Cola, Nike, Toyota, Sony and Starbucks when it came to people searching and commenting on their favourite product online. The study, which examined more than 1.5b posts in 38 different languages, showed 479,206 mentions of Vegemite - more than any other brand globally.

(a) TRUE or FALSE: IBM used random sampling to collect their data.

[2 marks]

FALSE

(b) Explain why these survey results might not be an accurate summary of global brand recognition.

[4 marks]

Its studying only posts in online forums. Only some people participate in these, so the information may not be representative of the global community.

[Total: 6 marks]

— END OF QUESTION 9 —