

Semester Two 2026
Mid-semester PRACTICE

Faculty of Business & Economics

UNIT CODES: ETC5521
TITLE OF PAPER: Diving Deeper into Data Exploration
EXAM DURATION: 1 hour 30 minutes

RULES

This is a closed book exam, however you may bring one one-sided page of notes equivalent to 12 point font. All answers are to be written on the exam script, in pen or pencil. All other devices need to be out of reach. You must comply with any instructions given to you by Monash exam staff.

As a student, and under Monash University's Student Academic Integrity procedure, you must undertake all your assessments with honesty and integrity. You must not allow anyone else to do work for you and you must not do any work for others. You must not contact, or attempt to contact, another person in an attempt to gain unfair advantage during your assessment. Assessors may take reasonable steps to check that your work displays the expected standards of academic integrity.

Failure to comply with the above instructions, or attempting to cheat or cheating in an assessment may constitute a breach of instructions under regulation 23 of the Monash University (Academic Board) Regulations or may constitute an act of academic misconduct under Part 7 of the Monash University (Council) Regulations.

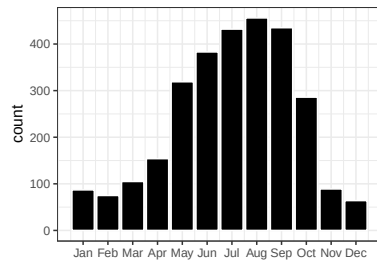
There are **11 questions** and a **total of 25 marks**.

Write your Monash ID here:

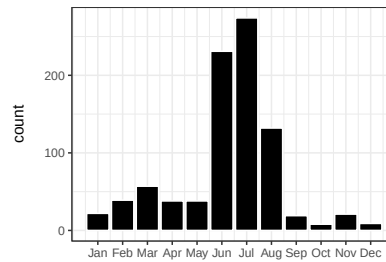
1. Exploratory data analysis (2 marks)

From these four plots, which (if any) shows the **least** interesting pattern and why?

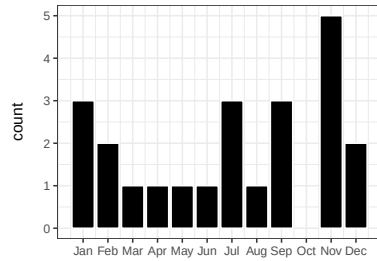
(a) Sightings of Gouldian Finch in the NT



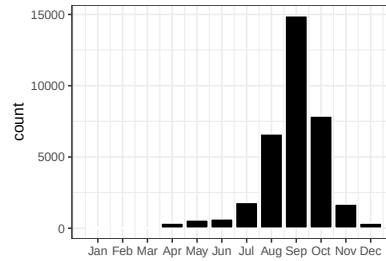
(b) Sightings of Manta Rays in the QLD



(c) Sightings of Glow Worms in the NSW



(d) Sightings of Orchids in the WA



Answer: _____

2. Exploratory data analysis (2 marks)

From the stem-and-leaf compute the median score. The leaf is at one digit, so 10 | 4 means 104.

1 | 2: represents 12

leaf unit: 1

n: 131

40	0		1111111111122233445555555566677778888999
(32)	1		00000001111222333345566667777899
59	2		0000012366788999
43	3		0012555688999
30	4		1334469
23	5		269
20	6		38
18	7		34799
13	8		267
10	9		1223
6	10		4
	11		
5	12		28
	13		
3	14		8
	15		
2	16		3
	17		
1	18		9

Answer: _____

3. Initial data analysis (2 marks)

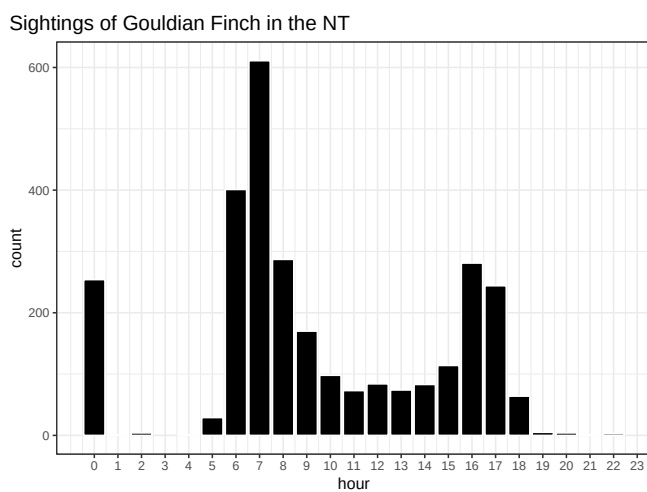
In the following data, how many missing observations are there?



Answer: _____

4. Initial data analysis (2 marks)

The main pattern here is sightings of Gouldian finch in the dawn and dusk hours. Is there anything unusual? If so, explain.

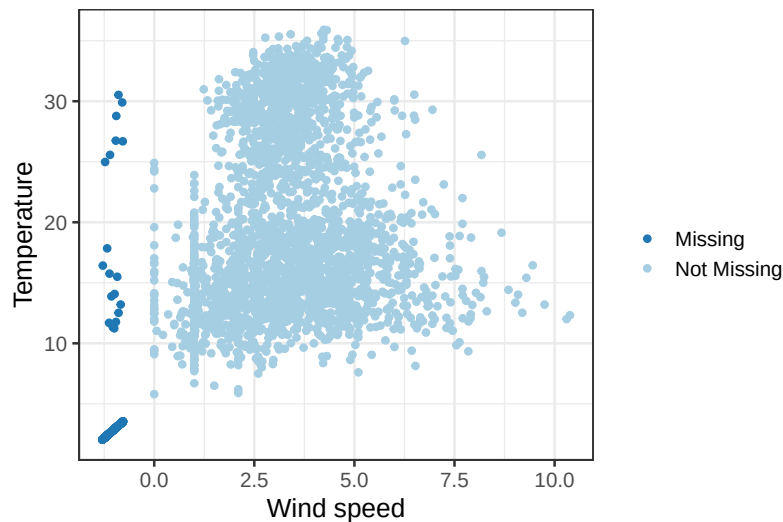


Answer: _____

5. Initial data analysis (2 marks)

This is a missing value plot of weather variables in Western Australia. What can we say about the missing values in this data? (Circle the ones that apply.)

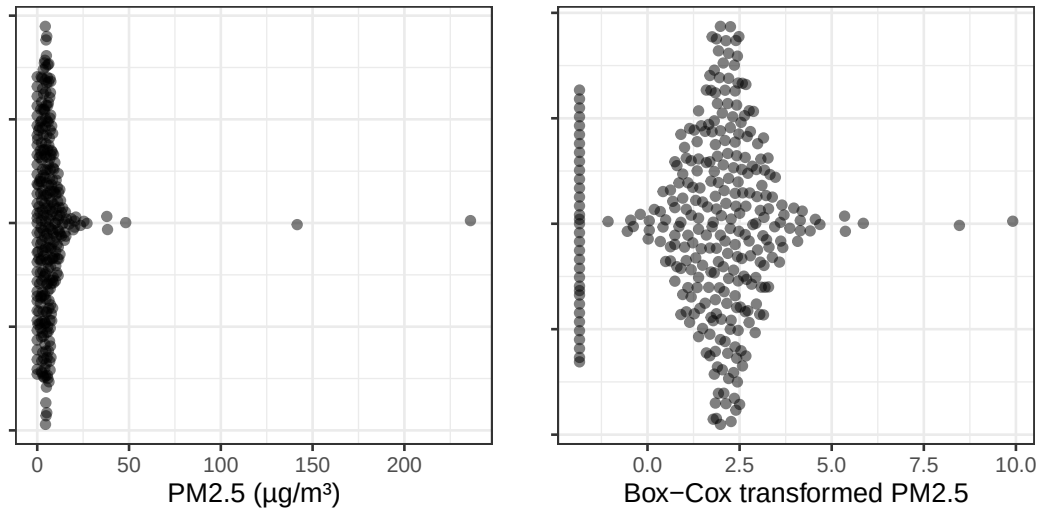
- a. There are missing values on both variables.
- b. There are missing values on only on temperature.
- c. There are missing values on only on wind speed.
- d. The missing values on wind speed have a similar temperature distribution to the complete values.
- e. The missing values on temperature have a similar wind speed distribution to the complete values.
- f. The missing values on temperature do not have a similar wind speed distribution to the complete values.
- g. The missing values on wind speed do not have a similar temperature distribution to the complete values.



Answer: _____

6. Initial data analysis (3 marks)

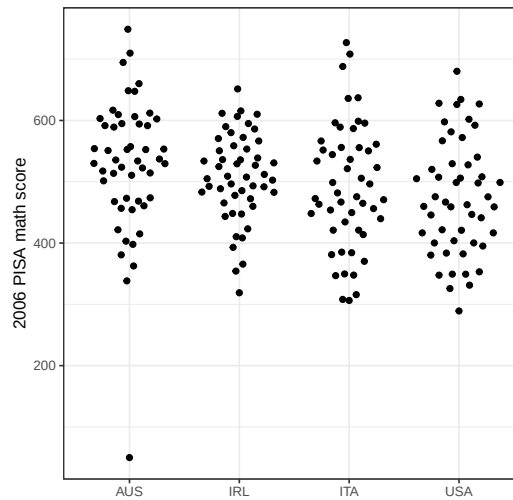
These plots show the data distribution and their distribution after transformation. It uses the ladder of transformations, -3, -2, -1, 0(log), 1/4, 1/3 1/2, 1, 2, 3 but the power function is optimised using Box-Cox. What power value do you think was used for the transformation? And explain anything else that might have been done in the calculation.



Answer: _____

7. Initial data analysis (3 marks)

This is a plot of math scores from the Program for International Student Assessment from 2006. Describe the pattern in the data, anything unusual, and how you might write a validator to catch any future problems.



Answer: _____

8. Using computational methods to assess significance of patterns (2 marks)

When showing a lineup containing 15 plots to 21 people, 5 people identified the data plot. The p -value is calculated with this code, `pvisual(5, 21, 15)`, producing the following result. Is this evidence to support that the structure in the data plot is different from null structure? (Assume that the nulls were generated correctly.)

```
x simulated binom  
[1,] 5      0.031 0.011
```

Answer: _____

9. Using computational methods to assess significance of patterns (3 marks)

Do these two plot definitions test the same null hypothesis about possible structure in the data? Explain your reasoning.

A.

```
ggplot(lineup(null_permute("read"), stu_2022_oz),  
  aes(x=television,  
    y=read)) +  
  geom_quasirandom(alpha=0.5) +  
  stat_summary(colour = "red") +  
  facet_wrap(~.sample)
```

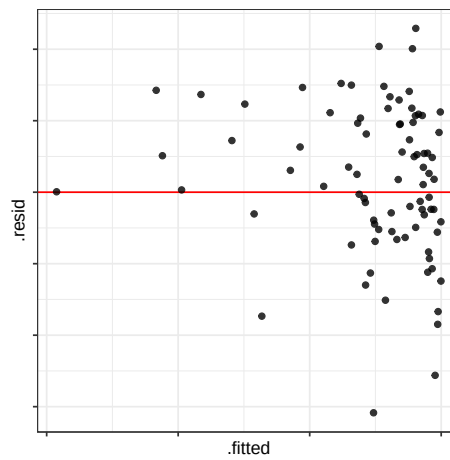
B.

```
ggplot(lineup(null_permute("read"), stu_2022_oz),  
  aes(x=television,  
    y=read)) +  
  geom_quasirandom(alpha=0.5) +  
  facet_wrap(~.sample)
```

Answer: _____

10. Using computational methods to assess significance of patterns (2 marks)

The following plot shows residuals plotted against fitted values after a linear model fit. Is there a problem with heteroskedasticity? (Note: Heteroskedasticity means that the variance of the y variable is related to the value of the x variable.)



Answer: _____

11. Using computational methods to assess significance of patterns (2 marks)

Here is a snippet of code to make a plot, where `class` is a categorical variable, and `v1` is a numeric variable.

```
ggplot(data) +  
  geom_violin(aes(x=v1, y=class))
```

Which of these is the correct null hypothesis, that the plot is testing? And what would be an appropriate method to generate null samples? (Circle your answers for each.)

Possible null hypotheses:

- a. There is a difference between the distribution of `v1` across the groups in `class`.
- b. There is no difference between the distribution of `v1` across the groups in `class`.
- c. There is a difference between the means of `v1` across the groups in `class`.
- d. There is no difference between the means of `v1` across the groups in `class`.
- e. There is a difference between the distribution of the values in `class` across the variable `v1`.
- f. There is no difference between the distribution of the values in `class` across the variable `v1`.
- g. There is no difference between the quantiles of `v1` across the groups in `class`.

Possible null generating methods:

- i. `null_permute()`: Permute the `class` variable
- ii. `null_dist(v1, "exp")`: Generate a sample from an exponential distribution for `x1`, for the different classes.
- iii. `null_dist(v1, "norm")`: Generate a sample from a normal distribution for `v1`, for the different classes.
- iv. `null_lm("rotate")`: Generate residuals by fitting `v1` to `class`, using the rotation method.
- v. `null_lm("bootstrap")`: Generate residuals by fitting `v1` to `class`, using the bootstrap method.

Answer: _____

Total: 25
