MONASH University

Semester Two 2026
Mid-semester PRACTICE SOLUTION

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.

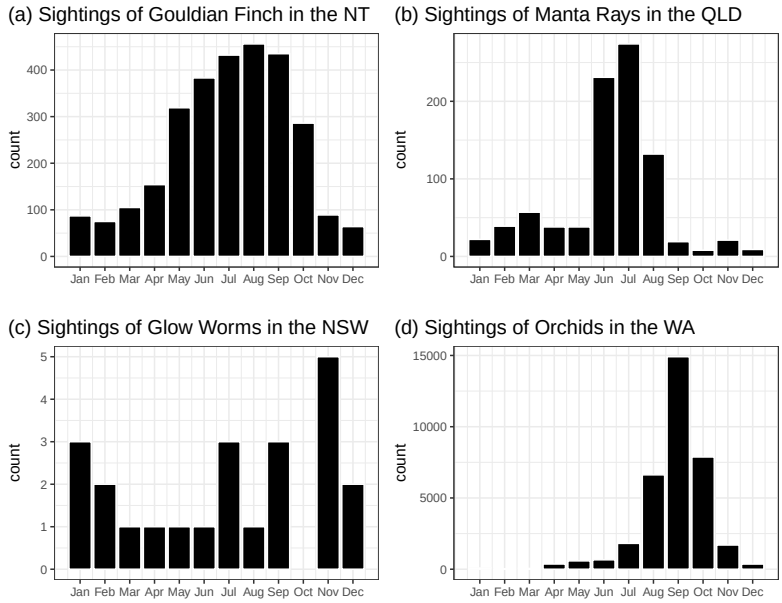
Failure to comply with the above instructions, or attempting to cheat or cheating in an assessment may constitute a breach of instructions under regulation 23 of the Monash University (Academic Board) Regulations or may constitute an act of academic misconduct under Part 7 of the Monash University (Council)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



Answer: (c) because there is no seasonal pattern.

(c) because there is no seasonal pattern.

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 | 11111111111222334455555555666777788889999
  (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: The median is the 72th score, which is 19 here.

The median is the 72th score, which is 19 here.

3. Initial data analysis (2 marks)

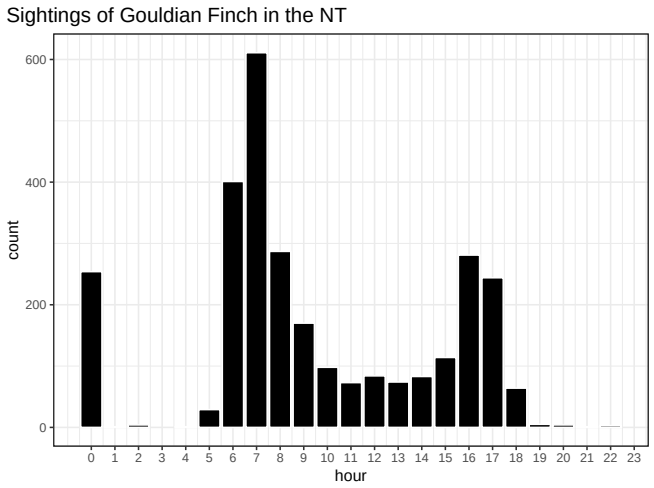
In the following data, how many missing observations are there?

Answer: Probably 3, but we only have number of missings for each variable, not whether the same ovservation is missing on all three variables.

Probably 3, but we only have number of missings for each variable, not whether the same ovservation is missing on all three variables.

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.



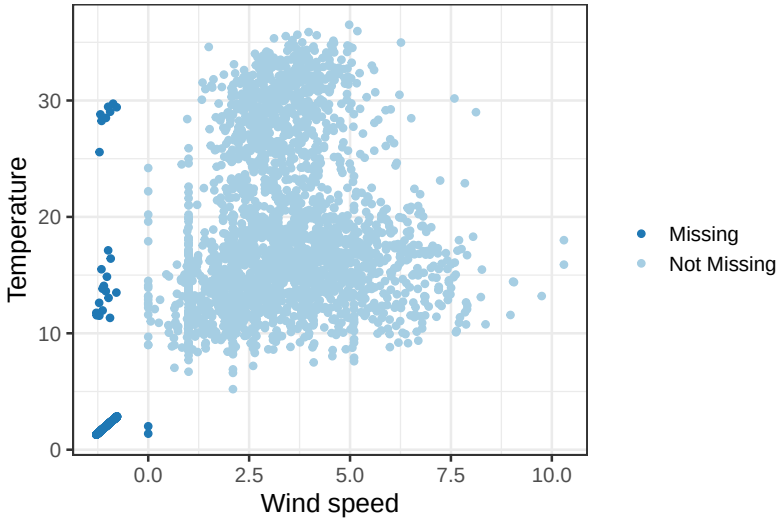
Answer: The large number of sightings between midnight and 1am is strange. It suggests that missing values have been coded as zero

The large number of sightings between midnight and 1am is strange. It suggests that missing values have been coded as zero

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.

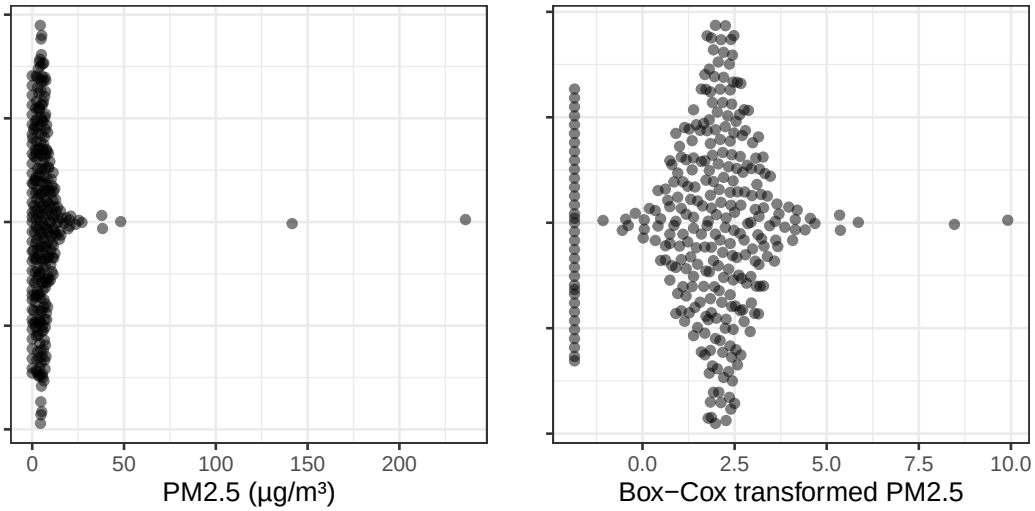


Answer: a, d are correct. Also f, but it is debatable because there are only two missings.

a, d are correct. Also f, but it is debatable because there are only two missings.

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 usnig 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.

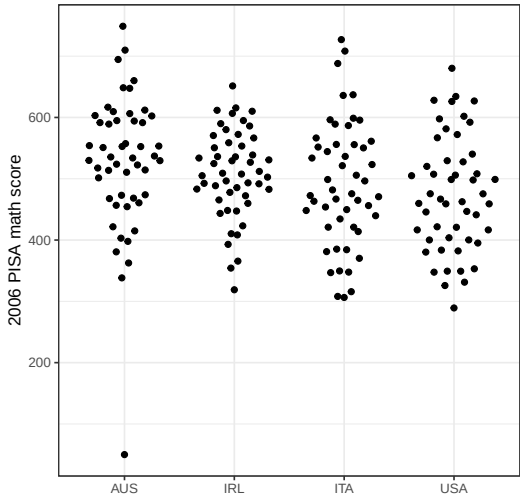


Answer: The data was transformed with a value less than 1, probably around 1/4. The data also looks like it was shifted before transforming.

The data was transformed with a value less than 1, probably around 1/4. The data also looks like it was shifted before transforming.

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: There is a very low score for Australia. It's not outside the usual bounds of test scores of 0-1000, but it is clearly unusually low. This is especially true if you remember that this is simulated data. One approach would be to provide some more evidence that this could not be a correct value, and set it formally to be an NA.

There is a very low score for Australia. It's not outside the usual bounds of test scores of 0-1000, but it is clearly unusually low. This is especially true if you remember that this is simulated data. One approach would be to provide some more evidence that this could not be a correct value, and set it formally to be an NA.

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: Yes, because the p-value is small, which suggests its very unlikely for 5 out of 21 people to identify the data plot if it was a null sample.

Yes, because the p-value is small, which suggests its very unlikely for 5 out of 21 people to identify the data plot if it was a null sample.

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.

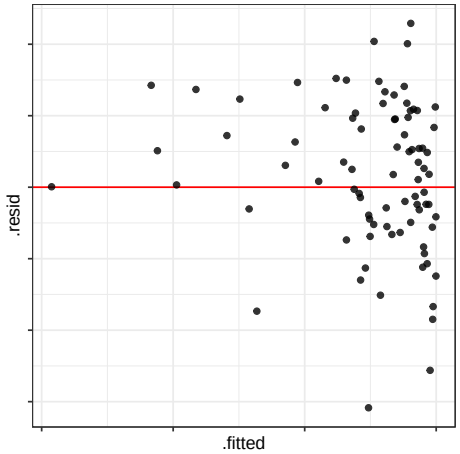
```
ggplot(lineup(null_permute("read"), stu_2022_oz),
    aes(x=television,
        y=read)) +
  geom_quasirandom(alpha=0.5) +
  facet_wrap(~.sample)
```

Answer: No. (A) overlays the mean, and a confidence interval on the plot, so it is really testing whether the means are the same across number of TVs. (B) shows only the jittered dotplot and so is testing whether the distributions are the same.

No. (A) overlays the mean, and a confidence interval on the plot, so it is really testing whether the means are the same across number of TVs. (B) shows only the jittered dotplot and so is testing whether the distributions are the same.

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: No. There isa. problem of skewness in the fitted values, with few observations having very low values. This will mean that there has to be smaller variance at low values, purely due to sample size.

No. There isa. problem of skewness in the fitted values, with few observations having very low values. This will mean that there has to be smaller variance at low values, purely due to sample size.

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: b and i are the correct answers

b and i are the correct answers

Total: 25