MAS8403 Project: Penguins!

Submit your project report in PDF form to Canvas by 16:30 on Friday 21st October 2022.

Your submission should be a single, coherent report submitted in PDF form. ***The page limit for the report is 6 pages****. Any reports going over this limit will be penalised. This limit does not include the cover page, bibliography or appendix. You do not need to include an abstract/executive summary. You do not need to include any R code.*

Palmer Penguins

The Palmer Station located in the Palmer Archipelago on Anvers Island, Antarctica, has been monitoring the ecology of the Palmer Long-Term Ecological Research (LTER) study area for over 50 years. You can see what's going on at the Palmer Station currently by clicking here. Being on Antarctica, naturally one of their keen interests is monitoring the local penguin population from which they record data in order to understand their population dynamics, responses to changing climate etc.

(a) The Palmer Research Station

(b) His name is George

The Data

The `palmerpenguins` dataset contains data measured on 333 penguins from the Palmer Archipelago. The variables observed are:

- **species**: The species of the penguin (Adelie, Chinstrap or Gentoo)
- **island**: The island on which the penguin lives (Biscoe, Dream or Torgerson)
- **bill_length_mm**: The length of the penguin's bill (in millimetres)
- **bill_depth_mm**: The depth of the penguin's bill (in millimetres)
- **flipper_length_mm**: The length of the penguin's flipper (in millimetres)
- **body_mass_g**: The penguin's body mass (in grams)
- **sex**: The sex of the penguin (male or female)
- **year**: The year the measurements were taken

Installing the Data

Install the `palmerpenguins` package and access the data

```
install.packages("palmerpenguins") # You only need to do this once
library(palmerpenguins)
data("penguins")
penguins = na.omit(penguins) # Removes missing rows
```

Run the following code to access your unique subset of the penguin dataset

```
my.student.number = 123456789 # Replace this with your student number
set.seed(my.student.number)
my.penguins = penguins[sample(nrow(penguins), 100), ]
```

the object `my.penguins` now contains the data on your 100 penguins.

The Task

You are to produce a report which comprises of an exploratory data analysis of the data on your sample of 100 penguins. **In this exploratory analysis you should include appropriate graphical and numerical summaries for your data, ensuring all summaries/figures are suitably discussed in the report**.

We would like to be able to use this sample of data to estimate probabilities/proportions for the penguin population in general. One way to do this is to fit a probability distribution to our sample, and use this distribution to estimate probabilities/proportions for the population. **For at least one of the measurement variables (bill length, bill depth, flipper length and body mass) choose an appropriate probability distribution to represent the variable, and find estimates for the parameters of the distribution for your data. Comment on the accuracy of your distribution, and whether you feel this is a good method for estimating population proportions.**

Sexing (i.e. determining the sex) of a penguin can often be very difficult without causing distress to the penguin. Researchers at the Palmer station would like to be able to estimate the sex of a penguin from measurement data, thereby avoiding the need for invasive procedures. **From your data, which variables appear to be the best at distinguishing between male and female penguins? How reliable do you think they would be at identifying the sex of a penguin?** Similarly, evolutionary biologists are interested in knowing if there is a significant difference in the physical characteristics of penguins living on different islands. **From your data, does the island the penguin is from appear to have a significant impact on any of its physical characteristics?**

Marking Criteria

Reports will be marked on the university scale. Credit will be given for:

- **Mathematical accuracy** – How well you carry out the statistical techniques in your report
- **Methodology** – An understanding of why you have chosen the techniques that you have, and what their output means in terms of your investigation
- **Critical evaluation** – A discussion of the strengths and weaknesses of your methods, how things could be improved etc
- **Report structure and presentation** – How well your report is written in terms of structure, how well it flows etc (i.e. aiming for a single, coherent piece of writing, as opposed to lots of separate answers jammed together)
- Extra credit will be given for any reading/techniques implemented from outside the scope of the module, but this is not a requirement to receive a good mark. Similarly any investigations carried out beyond what was described in the report task will also be considered for extra credit.

As such the marking scale (out of 100) will be:

- 80+ (Upper distinction) – A publication quality piece of work
- $70 - 79$ (Lower Distinction) – A very good piece of work showing strong understanding
- $60 - 69$ (Merit) – A solid attempt, mostly hitting the main points
- $50 - 59$ (Pass) – More good than bad, but clear areas to improve
- < 50 (Fail) – Significant things went wrong or weren't attempted