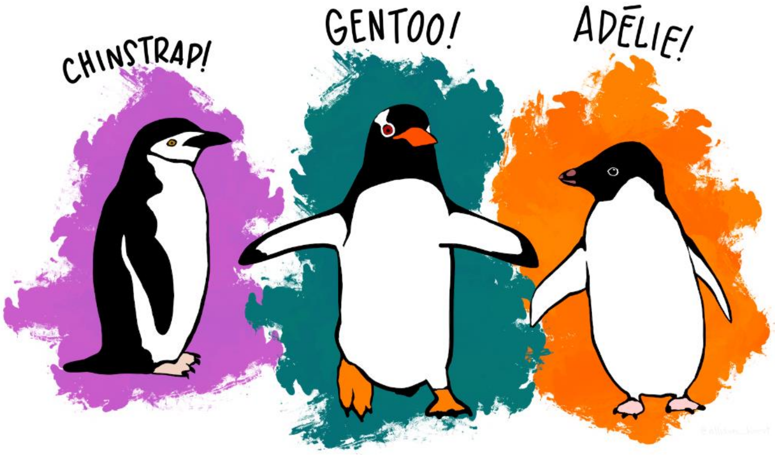
DATA2001 Assignment 1 (Weight: 20%)

Due:

The aim of this assignment is to gain practical experience in analysing structured data. You must complete this in Python using a Jupyter notebook. You will need to submit a single Jupyter notebook (.ipynb file) via Blackboard.

Dataset:

The dataset for this assignment consists of penguin's data. It contains body measurements, and isotope measurements from blood samples of three different penguin species which lived in the islands of the Palmer Archipelago, Antarctica. Data was collected from 2007 -2009, as a part of study done by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.



Artwork by @allison_horst

This dataset has 344 observations and 17 columns. A description of the columns is given below:

studyName - Sampling expedition from which data were collected, generated, etc.
Sample Number - an integer denoting the continuous numbering sequence for each sample
Species - a character string denoting the penguin species
Region - a character string denoting the region of Palmer LTER sampling grid
Island - a character string denoting the island near Palmer Station where samples were collected
Stage - a character string denoting reproductive stage at sampling
Individual ID - a character string denoting the unique ID for each individual in dataset
Clutch Completion - a character string denoting if the study nest observed with a full clutch, i.e., 2 eggs
Date Egg - a date denoting the date study nest observed with 1 egg (sampled)
Culmen Length - a number denoting the length of the dorsal ridge of a bird's bill (millimetres)
Culmen Depth - a number denoting the depth of the dorsal ridge of a bird's bill (millimetres)
Flipper Length - an integer denoting the length penguin flipper (millimetres)
Body Mass - an integer denoting the penguin body mass (grams)
Sex - a character string denoting the sex of an animal
Delta 15 N - a number denoting the measure of the ratio of stable isotopes 15N:14N
Delta 13 C - a number denoting the measure of the ratio of stable isotopes 13C:12C
Comments - a character string with text providing additional relevant information for data

More information on this data can be found in its github repository.

The submitted notebook should address 6 tasks (see marking grid for mark allocation):

1. Data Preparation: Read the dataset using the “pandas” library. Can you identify any missing data both row and column wise in the dataset? Handle them in an appropriate way. Explain how you did it along with a justification for your choices.

2. Exploratory Data Analysis (EDA): Perform a detailed univariate and bivariate EDA for all relevant columns in the dataset. Produce plots, report your observation for each plot and findings clearly.

3. Find the mean and standard deviation of each type of measurement for each species and report your findings in a table. Comment on apparent differences between the species.

4. Find correlations among the numerical columns for each species. Produce visualisations for the correlations and explain the observed results.

5. Perform k-means clustering on the data. Comment on the number of clusters chosen, on possible limitations, and on any form of uncertainty about the results. Are the results in agreement with what you observed in the EDA?

6. Perform principal component analysis on the data. Comment on the results, plot the percentage of variance explained by each principal component. Also plot the principal components which you think are of interest, report your observations and limitations.

Note: Jupyter notebook should be commented properly and written in a way easier to understand for the reader. For marking purpose, code will be rerun to verify the results.