# Principles of Data Analytics

These are the assessment instructions for Principles of Data Analytics in Summer 2023/24. They cover 100% of the marks for the module. The deadline for all elements is Tuesday, 30 April 2024.

## Purpose

The purpose of the assessment is to ensure students can demonstrate the following.

1. Source and investigate sets of data.
2. Programmatically explore and visualize data.
3. Apply basic mathematical data analysis techniques to data sets.
4. Model real-world problems for analysis by computer.
5. Provide evidence in a decision-making process using a data set.
6. Appreciate the limitations of graphical representations in data intensive workflows.

## Overview

The assessment is broken into three overlapping components: a set of tasks, a small project, and a presentational component. All components should be submitted in a single Jupyter notebook in a single GitHub repository. You will be given guidance throughout delivery of the module as to how to do this.

The assessment focuses on the widely available palmerpenguins data set. You can find a copy of the data set here. The data set is interesting for several reasons, which we will cover during the module.

There are four tasks to complete, listed below. The project, also below, is essentially a fifth task but is larger than the others. These should ideally be completed as we cover the relevant topics in the first few weeks of the module.

The presentational component does not involve you giving a presentation. Rather, it refers to how your work is presented in GitHub and your Jupyter notebook. The idea is that you build a portfolio of work that can be used in future job applications and interviews. In our experience, how your portfolio is presented in GitHub is important to prospective employers.

## What to Submit

All of your work should be in the `main` branch of a single GitHub repository. Use the form on the module page to submit your repository URL. Commits in GitHub on or before the deadline will be considered. You should set up your repository and submit the URL immediately.

The repository should contain a single notebook named `penguins.ipynb`. This notebook should contain all your work on the data set.

### Tasks (40%)

1. Create a GitHub repository with a `README.md` and a `.gitignore`. Add a Jupyter notebook called `penguins.ipynb` and add a title to it.
2. Find the palmerpenguins data set online and load it into your Jupyter notebook. In your notebook, give an overview of the data set and the variables it contains.
3. Suggest the *types* of variables that should be used to model the variables in the data set in Python, explaining your rationale.
4. Create a bar chart of an appropriate variable in the data set. Then create a histogram of an appropriate variable in the data set.

### Project (40%)

Select two variables from the data set and provide an analysis of how correlated they are.

### Presentational Component (20%)

Ensure your repository is tidy, with no unnecessary items. Ensure your `README.md` and `.gitignore` files are appropriate. Make sure your notebook contains a single cohesive narrative about the data set.

## Marking Scheme

Each of the three components will be considered using the four categories below. Each category will be given equal weight. To receive a good mark in a category, your submission needs to provide evidence of meeting each of the criteria listed under it. In line with ATU policy, the examiners' overall impression of the submission may affect marks in each category.

Towards the end of the semester you may be asked to demonstrate the work to date in your repository. You will be notified by email as to when this will happen and, if called, attendance is mandatory.

### Research

- Evidence of research on topics.
- Appropriate referencing.
- Building on work in the literature.
- Comparison to similar work.

### Development

- Clear, concise, and correct code.
- Appropriate tests.
- Demonstrable knowledge of different approaches and algorithms.
- Clean architecture.

### Documentation

- Clear explanations of concepts in notebooks.
- Concise comments in code and elsewhere.
- Appropriate, standard README for a GitHub repository.

### Consistency

- Tens of commits, each representing a reasonable amount of work.
- Literature, documentation, and code evidencing work on the assessment.
- Evidence of reviewing and refactoring.

## Policies and Advice

Once completed, your repository should be readily presentable in job interviews. A technically competent person should be able to understand your work and how to interact with it, without you being there.

Please remember that you are bound by ATU policies and regulations. You should familiarize yourself with these on the Student Hub. Pay particular attention to the Policy on Plagiarism and the Student Code of Conduct. If you have any doubts about what is allowed, email me. Advice

Students sometimes struggle with the freedom given in an open-style assessment. You must decide where and how to start, what is relevant content for your submission, how much is enough, and how to make the submission your own. This is by design - we assume you have a reasonable knowledge of programming and an ability to source your own information.

Companies tell us they want graduates who can (within reason) take initiative, work independently, source information, and make design decisions without needing to ask for help. You need a plan, you cannot just start coding straight away.

## Indicative Syllabus

The following are a list of some topics we may cover in the course of the module.

Data Acquisition

- Searching for data sets
- File formats
- Data structures

Exploration and visualization

- Calculating summary statistics
- Programmatically generating plots
- Histograms, scatter plots, box plots

Data cleansing

- Character sets
- Search and replace
- Regular expressions
- Outlier identification

Mathematical techniques

- Visualization
- Ordinary linear regression
- Classification

Programming

- Programmatically applying data analysis techniques
- Automating data analysis workflows

29 January 2024