Department of Computer Science

Summative Coursework Set Front Page

Module Title	Programming in Python for Data Science
Module Code	CS2PP22
Lecturer responsible	Dr Todd Jones
Type of Assignment (e.g., technical report, set exercise, in-class test)	Coursework
Individual or Group Assignment	Individual
Weighting of the Assignment	40%
Word count/page limit	N/A (Complete notebook prompts)
Expected hrs spent for the assignment (set by lecturer)	8
Items to be submitted	A single **`.zip`** archive containing: 1. Fully executed Jupyter notebook (**`.ipynb`**) 2. Exported HTML copy of above (**`.html`**) 3. cardata_modified.csv 4. Separately constructed module files (optional)
Work to be submitted on-line via Blackboard Learn (**Gradescope**) by	**2024 February 9th (Friday) 12:00 (noon)**
Work will be marked and returned by	**2024 February 29th (Thursday)**

Note

By submitting this work, you are certifying that you have read the assessment guidelines, which are displayed in the folder of Assessment on the Blackboard course for this module, and that you have conformed to and understand the associated policies and practices, including those on:

- Submitting your own work, not that of other people or systems, and the associated penalties for Academic Misconduct
- Submitting by the specified deadline, and the penalties associated with late submission (if allowed)
- The exceptional circumstances system
- For students with relevant needs, attaching with a green sticker

1. Assessment classifications

This coursework assesses your ability to:

- implement common computer science algorithms in the Python programming language;
- demonstrate an understanding of the use of functional and object-oriented programming paradigms in Python;
- read and manipulate data in several formats to extract specific features;
- assemble, implement, and select appropriate data science methodologies in Python;
- employ third-party Python libraries appropriately to design and create well-structured programs for practical applications.

In general, you will gain credit for:

- preparing and submitting required files as requested;
- successful implementation of the specified coding tasks;
- writing efficient, functional code;
- providing thoughtful, clear, well-structured written analysis.

Your assignment will be marked according to the marking scheme provided below. The scheme is designed so that the collectively weighted assignment mark will correspond to the following qualitative degree classification descriptions:

The table below shows what is typically expected of the work to obtain a given mark.

Classification Range	Typically, the work should meet these specifications:
First Class (>= 70%)	Outstanding/excellent work with correct codes and results. This work demonstrates coding proficiency with high efficiency and based on advanced techniques. Evidence of independent research into the methods used and a thorough justification of applications of these methods is presented clearly.
Upper Second (60-69%)	Good work with few mistakes. Some minor tasks have not been carried out or are not completely correct. Coding with good efficiency. Evidence of good knowledge of core concepts, with good explanations and justifications.
Lower Second (50-59%)	Demonstrates knowledge of core concepts but with some mistakes. Explanations and justifications of methods used are logical but limited in depth. Coding with average efficiency. Most tasks have been carried out with sufficient accuracy.
Third (40-49%)	Some parts of the assignment are missing and/or have partially correct results. Most tasks have not been carried out with sufficient accuracy. Results may not be correct or technically sound. Mistakes in application of knowledge and shows some misunderstandings. Explanations and justifications of methods used are not clear or logical. Coding might be inefficient.
Pass (35-39%)	Some significant part of the assignment is missing and/or has partially correct results. Gaps in knowledge and many mistakes, little evidence of understanding. Methods used are not well explained or justified. Coding is notably inefficient.
Fail (0-34%)	Many aspects of the assignment are missing, or there are large gaps in knowledge and significant mistakes, also showing limited understanding. Lack of logical explanations behind the methods used.

2. Assignment description

This assignment consists of **three tasks**. Each of these will be used to assess your implementation of several components of Python and some of the Data Science Process.

A detailed breakdown of the Marking Scheme is provided later in this document.

Task 1 – Python Basics

Exploring the concept of networks, or graphs, with the **dolphins.tsv** dataset, you will design **two** implementations of code to convert the data from an edge list representation into a neighbour list representation. The two implementations will differ in efficiency. At each stage, you will verify that your code is working correctly, and you will provide a written assessment of the work you have performed. This work will all be completed within the **CS2PP22_CW1.ipynb** Jupyter notebook, where further instructions are provided.

Task 2 – Data Preprocessing, Exploratory Data Analysis

Using the **cardata.csv** file within the **CS2PP22_CW1.ipynb** Jupyter notebook, you will execute several components of the Data Science Process. Working through this notebook's prompts, you will read, write, and manipulate data to extract specific features to better understand the data.

Task 3 – Python Classes

Using the **cardata_modified.csv** file produced above, continue working in the **CS2PP22_CW1.ipynb** Jupyter notebook. There, you will find detailed specifications for the creation of a "Tournament" class in Python. This class will represent instances of competitive tournaments, with functionality to simulate competitions and report information on their outcomes.

In each of the above tasks, some sub-tasks will ask you to provide a **written explanation** of the justification behind it your coding choices. Code and written responses should be presented in a set of well-formatted code and Markdown cells at appropriate points in your Jupyter notebook. This work will require the production and submission of additional files; details about these files and how they should be submitted are provided in the notebook and the Assignment Submission Requirements.

Project Directory and Data Description

The materials needed to complete this assessment are available in a single **CS2PP22_CW1.zip** file on the CS2PP22 Blackboard space, under the **Assessment** heading, in the **Coursework 1 Description and Datasets** item. This is outlined below and contains a **data** directory with subdirectories for **Task1** and **Task2**.

The first task relies on a file consisting of tab-separated integers. The second and third tasks rely on a file consisting of comma-separated values (CSV) with a header that briefly describes each column. This file will be used to work through the prompts in `CS2PP22_CW1.ipynb` that guide analysis of the data.

```
CS2PP22_CW1.zip
├── data/
│   ├── dolphins.tsv
│   └── cardata.csv
├── images/
│   └── …a few image files…
├── CS2PP22 CW1 - Inividual.pdf
└── CS2PP22_CW1.ipynb
```

Bottlenose Dolphin Social Network Data: **dolphins.tsv**

This dataset contains a representation of a social network dataset where dolphins have links between them if they frequently associated with one another. Each line contains 2 integers separated by a tab character. Each value represents an individual dolphin, and each line represents a connection between the listed pair.

Source: http://konect.cc/networks/dolphins/

Car Features and MSRP Data: **cardata.csv**

This dataset includes car features such as make, model, year, and engine type, as scraped from Edmunds and Twitter. It is often used to develop models to predict car prices based on their other characteristics.

Source: https://www.kaggle.com/datasets/CooperUnion/cardataset

Each **row** corresponds to a single kind of vehicle.

The **columns** correspond to:

Make	Car maker
Model	Car model
Year	Car year (Marketing)
Engine Fuel Type	Type of engine fuel category
Engine HP	Engine horsepower (HP)
Engine Cylinders	Number of engine cylinders
Transmission Type	Type of transmission category
Driven_Wheels	Drive wheel category
Number of Doors	Number of doors
Market Category	Market category

`Vehicle Size`	Vehicle size category
`Vehicle Style`	Vehicle style category
`highway MPG`	Highway fuel efficiency in miles per gallon
`city mpg`	City fuel efficiency in miles per gallon
`Popularity`	Twitter-based popularity metric
`MSRP`	Manufacturer suggested retail price (USD)

3. Assignment submission requirements

"Front page" of the Submission

The following are **compulsory**. Please add these items to at the **top of your Jupyter notebook** in the provided Markdown cell.

Module Code:

Assignment report Title:

Date (when the work completed):

Actual hrs spent for the assignment:

Content of the required work

You must use Python (**version 3.10** or above) Jupyter Notebooks (**version 6.3.0** or above). Where possible, use the packages included in the Anaconda3 distribution used in this module (**2023.03**).

If you find good reason to employ **additional Python packages** in the creation of your solution, please provide an excruciatingly detailed description of the package installation procedure that includes specification of your Anaconda3, Python, and Jupyter Notebook versions, as well as the version information for your additional Python packages.

As mentioned above, your submission should take the form of a single archive file (based on the one downloaded for this project). Upload the .zip file to Gradescope via the Blackboard submission point. As you do, you should see that Gradescope automatically unzips the files as you submit. This is fine.

You will find the submission point on the module's Blackboard page under **Assessment**. The name of the archive and should be formatted with your 8-digit student ID number, the module code, and the tag "CW1" (e.g., **`12345678_CS2PP22_CW1.zip`**).

While you might find it useful to include more material (e.g., modules containing functions or classes used in the notebooks), the final content of your Blackboard submission should have, at minimum, the following structure and contents. Items in **orange** represent new files that you will produce or modify.

```
12345678_CS2PP22_CW1.zip
├── data/
│   ├── dolphin.tsv
│   ├── cardata.csv
│   └── cardata_modified.csv
│   └── Task2/
├── CS2PP22_CW1.ipynb [completed and fully executed]
└── [any auxiliary modules, package version notes]
```

- You do not need to submit the image directory or this pdf.

Code Plagiarism

This coursework is expected to be the result of your own individual effort, **not that of other people or systems**. Do not work closely with others on this coursework. Do not employ pair programming techniques. Copying whole tutorials, scripts or images from external sources is not permitted. Any material you borrow from other sources **to build upon or to support your arguments** should be clearly referenced (use comments to reference element within Python scripts and code cells and supply formal references in a Markdown section at the end); otherwise, it will be treated as plagiarism, which may lead to investigation and subsequent action. Work **inspired by** module materials is permitted, but such material should not be used without significant modification. We understand that similar lines of code are inevitable, however, very similar lines of analysis and reporting spanning significant sections of the coursework will be investigated as potential academic misconduct (e.g., it is highly unlikely that you will independently choose the same model parameters and variable names).

4. Marking scheme

Element	Marks Available
Organisation: Preparation and submission of all required files	**5**
1: Network Representations	**10**
2.0: Analysis Preparation	**5**
2.1: Data Cleaning	**15**
2.2: Creating New Columns	**5**
2.3: Exploratory Data Analysis	**15**
3.1: Data Preparation	**5**
3.2: Class Definition	**30**
3.3: Class Execution	**5**
3.4: Class Comparisons	**5**
Coursework 1 Total	**100**

Refer to the notebook for additional information about requirements and a more granular listing of marks.