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 (coursework/online test): Coursework

Individual / Group Assignment: Individual

Weighting of the Assignment: 100%

Page limit/Word count: Approximately 1,500 words, excluding captions and tables

Expected hours spent for this assignment: 20 hours

Items to be submitted online (Blackboard):

3 Items:

- A single **`.zip`** (preferred) or **`.tar.gz`** archive
  - Containing files and directories as outlined below in the Assignment Submission Requirements
- A copy of the completed **`CS2PP22_Assessment_Task1.ipynb`** in **`.pdf`** format, which displays all content (code, markdown text, figures, images, etc.).
- A copy of the completed **`CS2PP22_Assessment_Task2.ipynb`** in **`.pdf`** format, which displays all content (code, markdown text, figures, images, etc.).

Work to be submitted on-line via Blackboard Learn by: 2023 March 13th (Monday) 12:00 noon

Work will be marked and returned by: 2023 April 4th (Tuesday)

NOTES

By submitting this work, you are certifying that it is all your sentences, figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work except where explicitly the works of others have been acknowledged, quoted, and referenced. You understand that failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalized accordingly. The University's Statement of Academic Misconduct is available on the University web pages.

If your work is submitted after the deadline, *10%* of the maximum possible mark will be deducted for *each* working day (or part of) it is late. A mark of zero will be awarded if your work is submitted more than 5 working days late. You are strongly recommended to hand work in by the deadline as a late submission on one piece of work can impact on other work.

If you believe that you have a valid reason for failing to meet a deadline then you should make an Exceptional Circumstances request and submit it *before* the deadline, or as soon as is practicable afterwards, explaining why. To make such a request log on to RISIS and on the Actions tab select Exceptional Circumstance: as explained at https://www.reading.ac.uk/essentials/The-Important-Stuff/Rules-and-regulations/Exceptional-Circumstances

ASSESSMENT CLASSIFICATIONS

This coursework assesses your ability to:

- understand and use appropriate Python syntax and ecosystem;
- implement common computer science algorithms and functional programming in Python;
- understand statistical and machine learning methods for data analytics and mining in Python;
- apply appropriate statistical and machine learning techniques for data science tasks.

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 requirements:**
First Class (>=70%)	Outstanding/excellent work with correct codes and results. An outstanding work should demonstrate coding proficiency with high efficiency and based on advanced techniques. Evidence of independent research into methods used and a thorough justification of applications of these methods.
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 the 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.

ASSIGNMENT DESCRIPTION

Major Coursework (100% of module assessment)

This assignment consists of **two tasks**. Both of these will be used to assess your implementation of elements of the Data Science process, using Python as the main tool.

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

Task 1 – Data Preprocessing, Exploratory Data Analysis, and Python Classes

Using the **`cardata.csv`** file within the **`CS2PP22_Assessment_Task1.ipynb`** Jupyter notebook, you will execute several components of the data science process and design and implement a class structure that controls and compiles data about a fictional sporting event by writing Python code to perform the outlined sub-tasks detailed in the notebook. Working through this notebook, you will read, write, and manipulate data to extract specific features, design and implement functional routines, and design and implement an algorithm to select an optimal subset from a larger dataset.

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.

Task 2 – Twitter Data Analysis

Using the **`CS2PP22_Assessment_Task2.ipynb`** Jupyter notebook, you will extract data from the social media platform, Twitter, and use the data as the basis for implementing components of the data science process to build and test a regression model. You will need to extract at least 300 tweets (perhaps, the 300 most recent tweets) from at least 3 Twitter accounts.

Visualise the results concisely and **discuss** the reasons why one might prefer the use of one of your tested methods over another. As in Task 1, written responses should be provided in a set of well-formatted Markdown cells at appropriate points in your Jupyter notebook.

Additional points of consideration and example extraction methods are provided in the notebook. Efficient extraction of the tweets will require installation of at least one new Python package. The most efficient of these, `tweepy`, requires that you obtain a developer account with Twitter. Instructions for gaining the appropriate access are found in the Additional Considerations section of this document.

Project Directory and Data Description

The materials needed to complete this assessment are available in a single **`CS2PP22_Assessment.zip`** file on the CS2PP22 Blackboard space, under the **Assessment** heading, in the **Coursework 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 comma-separated values (CSV) with a header that briefly describes each column. This file will be used to work through the prompts in `CS2PP22_Assessment_Task1.ipynb` that guide analysis of the data.

In the second task, you are asked to source your own data from Twitter. Use the provided Task 2 notebook, `CS2PP22_Assessment_Task2.ipynb`, to begin this analysis.

```
CS2PP22_Assessment.zip
├── data/
│   ├── Task1/
│   │   └── cardata.csv
│   └── Task2/
│       └── < - empty - >
├── CS2PP22_Assessment.pdf
├── CS2PP22_Assessment_Task1.ipynb
└── CS2PP22_Assessment_Task2.ipynb
```

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)

Twitter Data:

As noted in the Task 2 description above, you will extract the data from 3 accounts of your choice. The format of this data will differ based on the method of extraction you choose and the specific data features you choose to extract.

Assignment Submission Requirements

"Front page" of the Submission

The following are **compulsory**. Please add these items to at the **top of your Jupyter notebooks** in a Markdown cell. To be extra helpful, please repeat this information in the **Add Comments** section of the Blackboard submission page.

Module Code:

Assignment Report Title:

Student Number (e.g., 25098635):

Date (when work was completed):

Actual hours spent on assignment:

Assignment evaluation (3 key points):

We will use information about how long you spent on the assignment when we review and balance coursework between modules for later years. An exact answer is not necessary, but please try to give a reasonable approximation.

The assignment evaluation is an opportunity for you to provide feedback on your experience with the assignment. We will use this to improve coursework for next year. You might like to comment on the following concepts:

- Were any parts of the assignment particularly fun, engaging, interesting, boring, or frustrating?
- Was the assignment too long/short/easy/difficult, or were these features simply appropriate?
- Were there any notable errors or technical problems with the materials supporting the assignment?

You will **not be penalised** for providing negative points of evaluation.

Content of the Required Work:

You must use Python (**version 3.8** or above) Jupyter Notebooks (**version 6.3.0** or above). Where possible, use the packages included in the Anaconda3 distribution used in this module (**2021.05**).

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 **3 items**: a single archive file (based on the one downloaded for this project) and separate `.pdf` copies of the notebooks, one for each of the two tasks.

You will find the submission point on the module's Blackboard page under **Assessment**. The name of the archive and `.pdf`s should be formatted with your student ID, the module code, and the tag "Assessment" (e.g., **`ce9201209_CS2PP22_Assessment.tar.gz`**).

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.

```
cz9201209_CS2PP22_Assessment_Task1.pdf
cz9201209_CS2PP22_Assessment_Task2.pdf
cz9201209_CS2PP22_Assessment.zip
├── data/
│   ├── Task1/
│   │   ├── cardata.csv
│   │   └── cardata_modified.csv
│   └── Task2/
│       ├── twitter_user1.csv
│       ├── twitter_user2.csv
│       └── twitter_user3.csv
├── CS2PP22_Assessment.pdf
├── CS2PP22_Assessment_Task1.ipynb [completed and fully executed]
├── CS2PP22_Assessment_Task2.ipynb [completed and fully executed]
├── enhanced_boxplot.png
├── popularity.png
└── [any auxiliary modules, package version notes]
```

Code Plagiarism

Copying whole tutorials, scripts or images from other sources is not allowed. Any material you borrow from other sources to build upon should be clearly referenced (use comments to reference in Python scripts); otherwise, it will be treated as plagiarism, which may lead to investigation and subsequent action.

Marking Scheme

Task	Element	Marks Available
Task 1	Organisation: Preparation and submission of all required files	**5**
	1.0: Analysis Preparation	**5**
	1.1: Data Cleaning	**15**
	1.2: Creating New Columns	**5**
	1.3: Exploratory Data Analysis	**20**
	1.4: Fuel Efficiency Tournaments	**40**
	Overall: Coding efficiency and structure, including comments and docstrings, where appropriate.	**10**
	Task 1 Total	**100**
Task 2	Organisation: Preparation and submission of all required files	**10**
	2.1: Extraction of tweet datasets	**10**
	2.2: Exploratory data analysis	**20**
	2.3: Data processing	**10**
	2.4: Regression analysis	**20**
	2.5: Model evaluation and testing	**10**
	Overall: Coding efficiency and structure, including comments and docstrings, where appropriate.	**10**
	Overall: Report structure and reasoning (format, clarity, logic, quality of written communication)	**10**
	Task 2 Total	**100**
Total	**Assessment Total**	**200**

Additional Considerations

Task 2

To extract the tweets from the accounts you have selected with the `tweepy` package, you **MUST**:

- Have or create a Twitter account.
- Request a developer Twitter account: LINK
  - As a student, you will not be making information available to a government entity.
  - To apply, you will need to verify a phone number for your account.
    - You might find the need to perform further verification. **It has been reported in the past that this process could take a week to complete. Start early!**
  - Upon submission of the application, you will need to verify the associated email address.
  - Upon email verification, you will need to name "your App" to **Get keys**.
  - **Immediately** save the following to a secure location (**These are only shown once; later requests for these will provide different values.)**:
    - API Key
    - API Secret Key
    - Bearer Token
  - Navigate to the Developer Portal, select Projects & Apps on the left, then your App. Switch to the "Keys and tokens" tab. **Generate and securely save** your:
    - Access Token
    - Access Token Secret
- Your initial **application** will provide you with **Essential** access. However, to get the data needed for this project, you will need to apply for **Elevated** access. To gain this access:
  - Go to: https://developer.twitter.com/en/portal/products/elevated
  - Select **Apply** to begin the application.
  - You will need to provide details about your account, then answer a few questions.
  - These responses recently permitted **immediate** access to the **Elevated** tier:
    - How will you use the Twitter API or Twitter Data?
      - I plan to use Twitter data and APIs to meet the aims of an Undergraduate level module at the University of Reading, which asks students to use Python to extract a small selection of Tweets to provide a dataset with which students can practice regression analysis.
    - Are you planning to analyse Twitter data?
      - **Yes**
      - The goal of this task is to explore the features that influence the number of likes for three selected Twitter accounts, to be determined.
    - Will your App use Tweet, Retweet, Like, Follow, or Direct Message functionality?
      - **No**
    - Do you plan to display Tweets or aggregate data about Twitter content outside Twitter?
      - **Yes**
      - Some Tweets or aggregate data about Twitter content might be included as part as the report on the Tweet analysis. This report will only be seen by the module leaders and markers. It will not be otherwise distributed.
    - Will your product, service, or analysis make Twitter content or derived information available to a government entity?
      - **No**

Please contact your lecturer if you encounter problems in setting up access to this service or if you have any other questions about this assignment.