National College of Ireland

MSCDAD_A_JAN25I A/B/C, PGDDA_JAN25

February 2025

A. Staikopoulos, Namita Agarwal, Sidra Aleem, John Kelly

Analytics Programming & Data Visualization

Continuous Assessment (CA) Type: In-class Assessment

Weight: The assignment will be marked out of 100. This assessment contributes towards a maximum of 30% of the module's marks.

Instructions:

- Attempt all questions.
- You may consult any of the notes and code from the classes or the labs.
- You may access online documentation.
- **Do not use generative AI, CHATGPT, or similar**
- All code must be fully commented.
- Upload your code to Moodle via the CA Submission link.
- This is an individual assessment – there should be no communication (verbal, electronic, or otherwise) between students during the assessment.

SUBMISSION DETAILS:

Submit all code file(s)/notebooks to the Moodle CA Submission link before the deadline.

TURNITIN: All report submissions will be electronically screened for evidence of academic misconduct (i.e., plagiarism and collusion)

Duration of Continuous Assessment: 2 hours

[If Applicable] Attachments: The data files for Q1 and Q2 are available on the Moodle page and should be downloaded.

nobel_prizes.json

Football Team Stats.csv

Question 1: File I/O, Functions, Exceptions

The "**nobel_prizes**" dataset is provided to you with a JSON file on Moodle (nobel_prizes.json). The dateset contains information about the Nobel Prize winners on different categories, years etc. For example, the dataset contains the following information


```
"prizes": [
 {
  "year": "2024",
  "category": "chemistry",
  "laureates": [
   {
    "id": "1039",
    "firstname": "David",
    "surname": "Baker",
    "motivation": "\"for computational protein design\"",
    "share": "2"
   },
   {
    "id": "1041",
    "firstname": "John",
    "surname": "Jumper",
    "motivation": "\"for protein structure prediction\"",
    "share": "4"
   }
  ]
 },
```


a) Create a function that accepts two arguments. The first argument is the name of the file you will try to load/parse, and the second argument is an integer that indicates the number of Nobel Prizes in a specific category and year (for example, Physics in 2006). The default value is set to 50. Ideally, you should retrieve a random sample of that size. The function should return a data structure (such as a list) that contains the specified number of random Nobel Prizes in the given category and year. Your function should include appropriate exception handling clauses to recover from various common problems. Explain the handlers used and the effect of their declaration order. For this task, you are not allowed to use the Pandas library.

[7 marks - Function]
[8 marks – Exception Handling]
[5 marks - Random]

b) Using a loop structure, find the Nobel prizes after or equal to 1950 on physics and based upon the following values

1. *year*
2. category
3. laureates

print the following formatted message with the winners

2006 physics
1) John C. Mather in "for their discovery of the blackbody form and anisotropy of the cosmic microwave background radiation"
2) George F. Smoot in "for their discovery of the blackbody form and anisotropy of the cosmic microwave background radiation"
1986 physics
1) Ernst Ruska in "for his fundamental work in electron optics, and for the design of the first electron microscope"
2) Gerd Binnig in "for their design of the scanning tunneling microscope"
3) Heinrich Rohrer in "for their design of the scanning tunneling microscope"
….

[14 marks – loop & condition]
[6 marks - formatted output]

c) Using a loop structure, extract the following information from your random data structure (step a)

1. *year*
2. category
3. laureates

and write it to a csv file using the following structure and column names.

YEAR, CATEGORY, LAUREATES
1973, physics, Leo Esaki **&** Ivar Giaever **&** Brian D. Josephson

For this task you can reuse previous steps.
Note, you are not allowed to use the pandas library.

[10 marks – CSV Writing]

Question 2: Pandas

Create a pandas data frame by loading the provided **Football Team Stats.csv** file. The data of the file looks like this

	Rk	Squad	Country	LgRk	MP	W	D	L	GF	GA	GD	Pts	Attendance
0	1	Barcelona	ESP	1	29	23	4	2	53	9	44	73	83148
1	2	Napoli	ITA	1	30	24	3	3	66	21	45	75	25662
2	3	Paris S-G	FRA	1	32	24	3	5	75	31	44	75	40508

a) Use pandas to find the Squads where the Attendance in less than 10000 and greater than 50000.

[10 marks]

b) Use pandas to find for each Country the Squad with maximum wins (W)

[10 marks]

Question 3: Numpy

a) Create a NumPy array containing all numbers between -100 (included) and 100 (excluded) and with a step factor of 10. Next, modify the array so that the data is organized as a 2-dimensional array where the number of rows are 5.

[5 marks]

b) Create a 2D NumPy array (4x4) and slice it to extract the submatrix that includes rows 1 and 2 and columns 2 and 3.

[5 marks]

c) Create a 2D NumPy array and find the maximum values along both axis (both row-wise and column-wise).

[5 marks]

Question 4: RE

a) Find all words in text with exactly two or three letters. Test your function with different text inputs.

[5 marks]

b) Create a python program that would extract repeated (duplicate) words from a sentence. The words are separated with a space or hyphen (-). For example, the following are valid matches

It is also used at the start of every knock knock joke of which there are many.
It is also used at the start of every knock-knock joke of which there are many.

Test your function with different text inputs.

[10 marks]