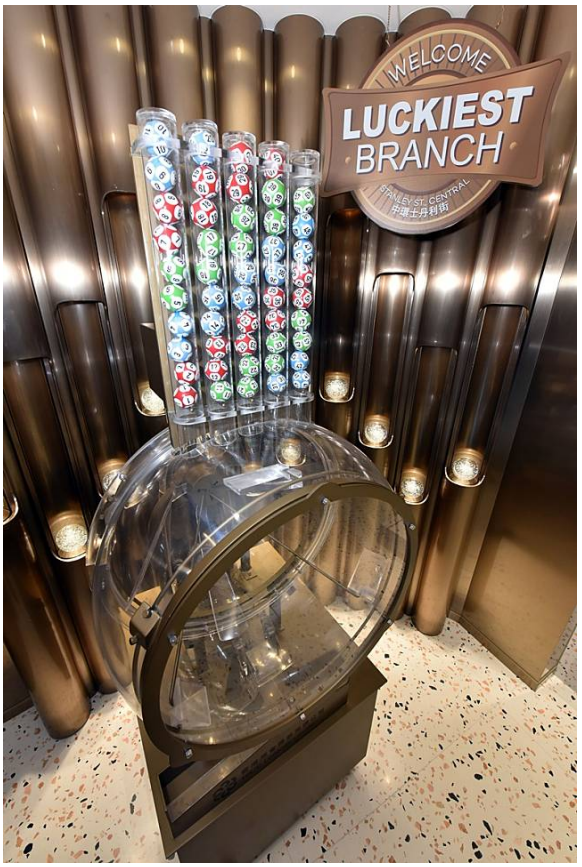
Department of Statistics, The Chinese University of Hong Kong
STAT 5106, Programming Techniques for Data Science (Term 1, 2022-23)

ASSIGNMENT 2
(Deadline: 15 Oct 2022, 2359)

[Note: Although the assignment is about Python script, we still welcome to use R script for handling those non-Datacamp questions. But please make sure the scripts are runnable and give the correct answers.]

Q1. 25%, Mark Six Simulation



Mark Six introduction
The Mark Six Lottery is a 6 out of 49 lotto game which is conducted by HKJC Lotteries Limited, a subsidiary of The Hong Kong Jockey Club.

After each draw if your entry satisfies some situations, the prize will be paid.
Mark Six Prize Qualification

For simplify the question, the prizes are assumed as following:

Prize	Unit Prize
1st	$21,531,600
2nd	$885,990
3rd	$157,500
4th	$9,600
5th	$640
6th	$320
7th	$40

Task to do

a. Write a function for inputting the list of **your entry numbers**, and the tuple with list of **the draw numbers** and the special number,
then outputting the unit prize.

Example:
Numbers of your entry: [14, 21, 25, 34, 37, 41]
Numbers of draw: ([21, 25, 34, 37, 41, 46], 14)
Unit prize of your entry: 885990 (2nd prize)

Please check in the function: ensure the input formats are as same as the example.

b. Write a function for simulating the draw numbers.
Please use the following to generate random numbers:
   1. np.random.choice
   2. np.arange
   3. np.random.seed(9527)

and output the numbers of draw, with sorting in ascending, and the structure as the draw numbers in part a: ([21, 25, 34, 37, 41, 46], 14)

Note: To perform a mark 6 drawing
- Drawing 7 numbers out, in the integer set from 1 to 49, without replacement
- the first 6 numbers are the official drawn numbers, and the last number is the extra number

Example:
The first 3 draws after random seed is set:
([5, 11, 19, 20, 40, 45], 44)
([19, 23, 29, 30, 35, 49], 45)
([4, 5, 26, 27, 35, 39], 10)

c. Combining part a, b, write a function for:
    1. inputting the list of **your entries numbers**
    2. outputting the total of prizes from all entries.

    Example:
    np.random.seed(9527)
    Numbers of your entry: [1, 10, 18, 39, 43, 45]
    Output prize = 6th prize = $320

Q2. 50% , Please attempt one of the following:

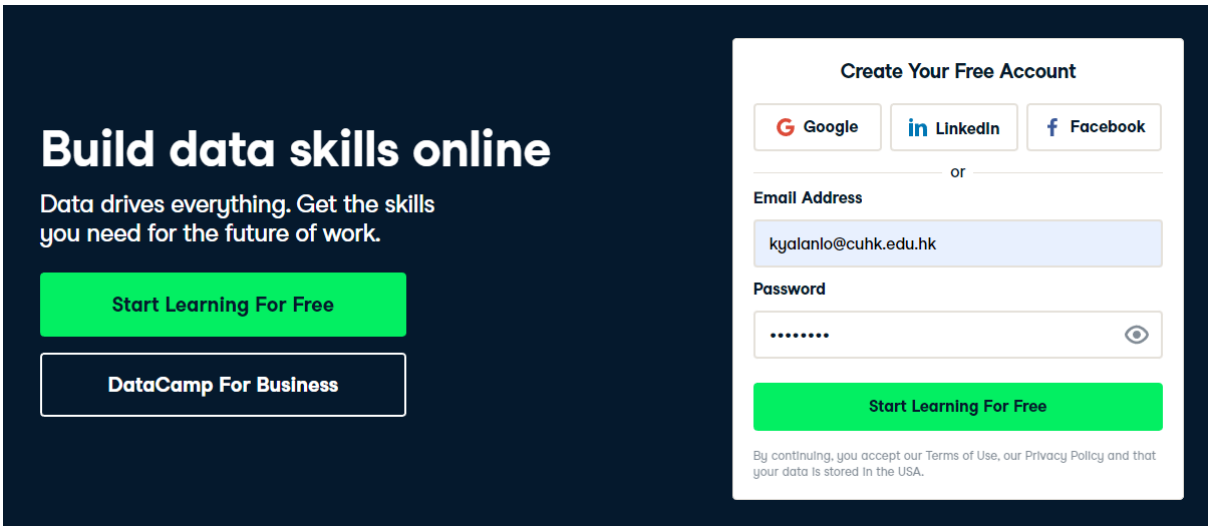
Option A. Time expectation for 5 hours. (50%)

Please complete all the following “Assignment 1” chapters in Python Fundamentals Skill Track in Datacamp.
Note for the teaching schedule I still follow to Dr. Chuck (py4e) . But if you need extra examples and training this skill track can help.
Although this is not totally covered what we learn for assignments 1 and 2, the whole skill track is still recommended to all guys completing.

Course	Chapter	Should be finished in
Introduction to Python	Python Basics	Assignment 1
Introduction to Python	Python Lists	Assignment 2
Introduction to Python	Functions and Packages	Assignment 1
Introduction to Python	Numpy	Assignment 2
Intermediate Python	Matplotlib	Do this if you want
Intermediate Python	Dictionaries and Pandas	Assignment 2
Intermediate Python	Logic, Control Flow and Filtering	Assignment 1
Intermediate Python	Loops	Assignment 1: While Loop, For Loop Assignment 2: Loop Data Structure Part 1/2
Python Data Science Toolbox (Part 1)	Writing your own functions	Assignment 1
Python Data Science Toolbox (Part 1)	Default arguments, variable-length arguments and scope	Do this if you want
Python Data Science Toolbox (Part 1)	Lambda functions and error-handling	Assignment 1: Introduction to error-handling Assignment 2: the others
Python Data Science Toolbox (Part 2)	Using iterators in PythonLand	Do this if you want
Python Data Science Toolbox (Part 2)	List comprehensions and generators	Assignment 2

To access into the datacamp classroom, please:

1. Access into www.datacamp.com
2. Register with your cuhk email account - the one ending with cuhk.edu.hk .
   (Note: don't use your personal account - since the classroom won't accept any non cuhk.edu.hk accounts)



3. After registration, check the following URL out:
   https://www.datacamp.com/groups/shared_links/d73d84de50ca199567b25860e1bb6ac11b2c36c8e99c553a88d4881eaa89d781
4. You can see assignments assigned to you.

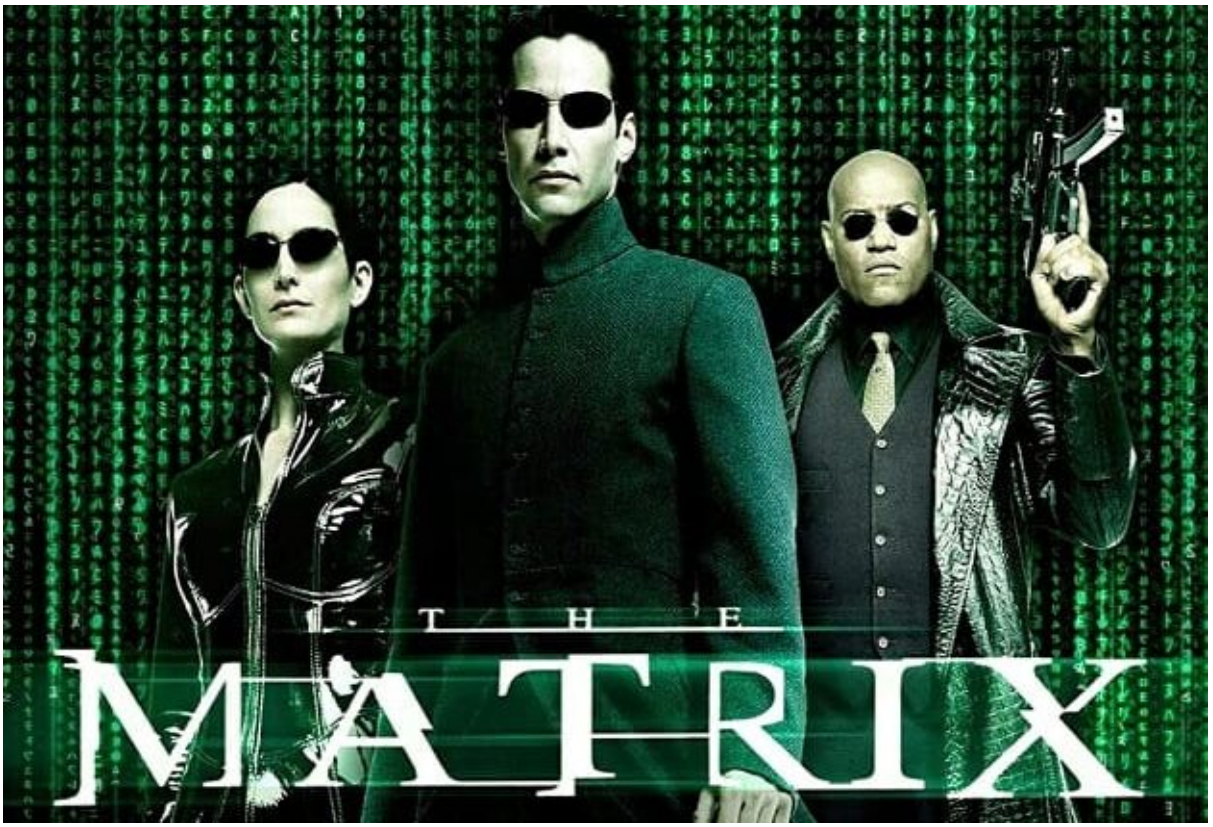
Please submit your cap screens for completion proof as the following:
(No need your name, we will check your email registered.)

1. If you have completed the Q1 Option A in Assignment 1, just cap the screen with all COMPLETED status → done.

2. If not, please attach your work to achieve the condition of this part.

Option B. Time expectation is 2 hours - given you are already familiar with Python programming, or have already started the career of Data Scientist :)

Part 1. Matrix Script (25%)



Neo has a complex matrix script. The matrix script is a N X m grid of strings. It consists of alphanumeric characters, spaces and symbols (!,@,#,$,%,&).

Matrix Script

T	s	i
h	%	x
i		#
s	M	
$	a	
#	t	%
i	r	!

Matrix Decoded

This$#is% Matrix# %!

To decode the script, Neo needs to read each column and select only the alphanumeric characters and connect them. Neo reads the column from top to bottom and starts reading from the leftmost column.

If there are symbols or spaces between two alphanumeric characters of the decoded script, then Neo replaces them with a single space ' ' for better readability.

Neo feels that there is **no need to use 'if' conditions for decoding**.

Alphanumeric characters consist of: [A-Z, a-z, and 0-9].

Task to do

Write a function with inputting a Matrix Script - list of lines, with the following format:

- The first line contains space-separated integers N (rows) and m (columns) respectively.
- The next N lines contain the row elements of the matrix script.
- Outputting the decoded matrix script

Don't use "if" - else no marks of this question will be given.

Example

Input

```
7 3
Tsi
h%x
i #
sM
$a
#t%
ir!
```

Output

```
This is Matrix#  %!
```

Note that Neo replaces the symbols or spaces between two alphanumeric characters with a single space ' ' for better readability.

Part 2. Continuing Mark Six Simulation (25%)

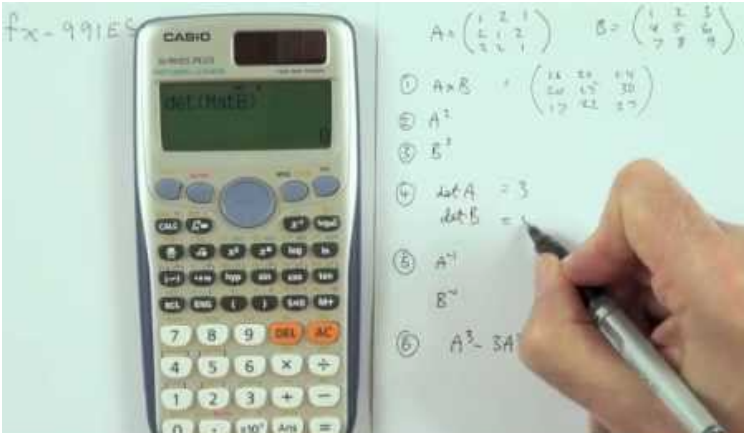
Combining part a, b, and release the length restriction (at least 6) of input entries numbers - multiple entries (複式), all combinations to make 6-number entries

Write a function for:

1. inputting the list of **your entries numbers**
2. outputting the total of prizes from all entries.

   Example:
   np.random.seed(9527)
   Numbers of your entry: [1, 10, 18, 39, 43, 45, 49]
   The draw simulated: ([4, 10, 18, 19, 39, 44], 43)
   Total of prizes: 3 6th prizes + 1 7th prizes = 320 * 3 + 60 = $1020

Q3. 10% , Matrix Calculator in Python



Create the following matrices:

$$A=\begin{bmatrix}1&2&3&4\\5&6&7&8\\9&10&11&12\\13&14&15&16\end{bmatrix} \quad B=\begin{bmatrix}4&3&2&1\\3&6&4&2\\2&4&6&3\\1&2&3&4\end{bmatrix} \quad C=\begin{bmatrix}4&3&2&1\\5&6&7&8\\8&7&6&5\\4&3&2&1\end{bmatrix}$$

Define tr(X) (stands for trace of X) as the sum of all the diagonal elements of X. For example, tr(A)=34, tr(B)=20 and tr(C)=17.

Please use numpy to verify the following:

a. tr(ABC) = tr(CAB) = tr(BCA)
b. (ABC)' = C'B'A' (Note: A' is transpose of A)
c. $(DB)^{-1} = B^{-1}D^{-1}$ where D = I + A

Q4. 15%, Extractions from Statistics Dept History table

The department website is welcome to seek our history page - nearly 40 years.

Task to do

1. download this html file and work in python - I have extracted the html of the table out.
2. Use regular expression package (re) to find out all names of chair professors.

Suggested Output:

Prof. Howell Tong
Professor Sik-yum Lee
Professor Nai Ng Chan
Professor Wing-hung Wong
Professor Yeh Lam
Professor Kim Hung Li
Professor Jian Qing Fan
Professor Ngai Hang Chan
Professor Qi Man Shao
Professor Xinyuan Song