Department of Computer Science and Software Engineering
The University of Western Australia

CITS1401
Computational Thinking with Python
Project 2, Semester 2, 2023
(Individual project)

Submission deadline: Friday 20th October 2023, 6:00 PM
Value: **20%** of CITS1401.

Project description:
You should construct a Python 3 program containing your solution to the following problem and submit your program electronically on Moodle. The name of the file containing your code should be your student ID e.g., 12345678.py. No other method of submission is allowed. Please note that this is an individual project. Your program will be automatically run on Moodle for sample test cases provided in the project sheet, if you click the "check" link. However, your submission will be tested thoroughly for grading purposes after the due date. Remember you need to submit the program as a single file box. You have only one attempt to submit your attempt. All open submissions at the Once your attempt is submitted, there is

You are expected to h academic conduct. In accordance with t general principles required to understar result of your own effort. Plagiarism dete tice, will therefore be used. Besides, if wh earnt little and will therefore, likely, fail th

You must submit you UWA policy, a late penalty of 5% will be c hat the assignment is submitted. No subm e except approved special consideration c

Project Overview:
The ABC research institute collected information of different organisations from all over the world for their future investment purposes. The collected dataset contains several parameters about each organisation, such as name and id of the organisation, country of the organisation registration, category of work, foundation year, number of employees, median salary, profit in 2020 and profit in 2021.

You are required to write a Python 3 program that will read a CSV file. After reading the file, your program is required to complete the following tasks:

1) Create a dictionary and store the following information in it:
    a. t-test score of profits in 2020 and 2021 for each country.
    b. Minkowski distance between the number of employees and the median salary for each country.

2) Create a nested dictionary that contains the following information for each category of organisations.
    a) organization ID's, and a list of the following data corresponding to each organization ID:
        i. Number of employees.
        ii. Percentage of profit change from 2020 to 2021 (absolute value).
        iii. Rank of the organisation within each category, with respect to the number of employees.

Requirements:

1) You are not allowed to import any external or internal module in python. While use of many of these modules, e.g., csv or math is a perfectly sensible thing to do in production setting, it takes away much of the point of different aspects of the project, which is about getting practice opening text files, processing text file data, and use of basic Python programming skills.

2) Ensure your p Calling the input() function will utomated testing system will no g program detects the call(s), it v rade).

3) Your program pt for the case of graceful term an error state and is exiting grac ionaries and print an appropriat 's outputs instead of (or in addit gram's progress in calculating su

Input:

Your program must de
`def main(csvfil`

The input argument f

- `csvfile`: T ne record of the organisations around the world. The first row the CSV file will contain the headings of the columns. A sample CSV file "Organisations.csv" is provided with project sheet on LMS and Moodle.

Output:

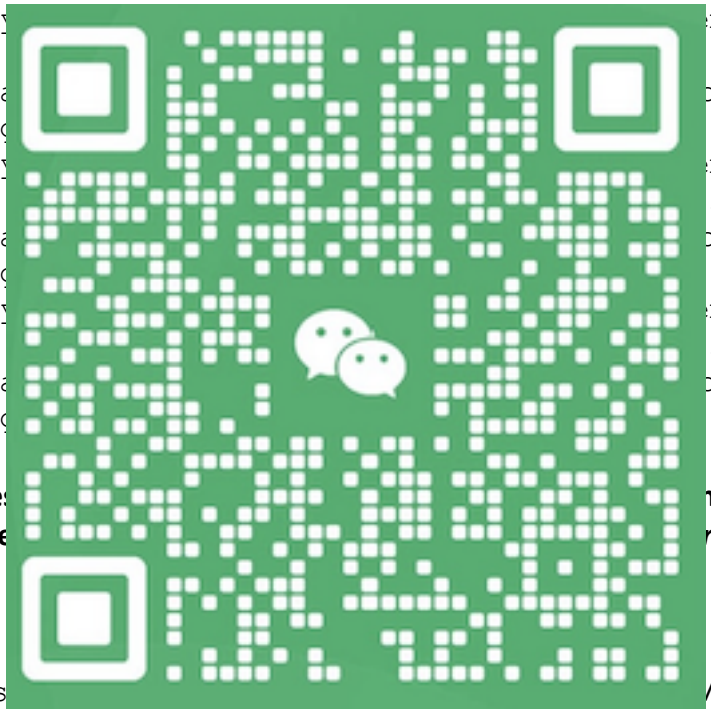
Two outputs are expected:

1) A dictionary which will have country names as keys, and the corresponding value for each country (key) will be a list containing t-test score and Minkowski distance between number of employees and median salary of the respective country. The expected output is in the following format:

```
{'country1': [t-test score, minkowski distance],
'country2': [t-test score, minkowski distance],…,
 'countryn': [t-test score, minkowski distance]}
```

2) A nested dictionary 'D' which will store the different categories of organizations (such as `'transportation'`, `'apparel'`, etc.) as keys and each corresponding value will be another dictionary 'd'. Each dictionary 'd' will store the organization IDs as keys within each category of organizations and information related to the organization IDs as values. Each value of 'd' will be a list containing the following data for each organisation:
    a. number of employees,
    b. absolute percentage of profit change from 2020 to 2021, and
    c. rank of an organisation within each category with respect to the number of employees (sort them in descending order, the organisation with the higher number of employees holds the higher rank, where the highest rank is '1'). If two organizations have the same number of employees, sort them (the tied organizations' IDs only) in descending order of their profit change. Below is the format:

```
{'category                                    mployees,
 absolute
 'organisa                                   olute
 percentag
 'category                                   mployees,
 absolute                                    …,
 'organisa                                   olute
 percentag
 'category                                   mployees,
 absolute                                    …,
 'organisa                                   olute
 percentag
```



Note: All the float re ne strings should be converted to lowe *rdered* collection of key-value pairs.

Examples:

Download `Organis` MS or Moodle. An example of how you can call your program from the Python shell (and examine the results it returns) are:

```
>>> output1, output2 = main('Organisations.csv')
```

The output variables returned are two dictionaries. Following are some examples of examining the returned dictionaries:

Example#1

```
>>> output1['brazil']
[-0.5175, 10174.3314]
```

```
>>> output2['biotechnology']
{'3c08339af3bb8c8': [8575, 36.4935, 1], 'eaf5ae0fcbcb4dd':
[6603, 78.062, 3], '139ab569bdfce4f': [3493, 62.6008, 4],
'a483cd7f7b486b4': [3427, 179.344, 5], '7ade1d82d2ac863':
[7205, 140.2845, 2], 'bf1cc30febed38c': [481, 8.9567, 6],
'bde405d2e490ebe': [92, 38.1616, 7]}
```

Example#2

```
>>> output1['afghanistan']
[0.0367, 4400.639]

>>> output2['accounting']
{'a5e8ce5cf97c2ac': [8128, 760.9484, 1], '5e2bb2dace9511e':
[7007, 73.0692, 3],'df66e70fae1aa5d': [7518, 0.5118, 2],
'795195c9db5e1c0': [6977, 96.9351, 4], 'a6bc77d5ce07c7b':
[6947, 90.4202, 5], 'b715731fa4a6cdb': [6429, 970.6279, 7],
'8f55cd0ad6dcd                                          dfba7b1d8d':
[6628, 110.683                                          01.5984, 9],
'c38cf79de2e6b                                          dce48de5ff':
[5523, 597.838                                          1.2454, 12],
'e0da4a69658ea                                          78271f3aa2':
[4288, 174.593                                          28.912, 15],
'ef7e820bc9f7e                                          05db7feee1':
[2658, 158.837                                          4.8933, 18],
'ba907c2acbc34                                          a4b5d7b2c1':
[871, 40.3551,
```

Assumptions:

Your program can ass

- The order of c sample file. Also, there can be er, rows can be in random order
- All string data ology" is same as "BIOTECHNOL consider both to be the same.
- There can be missing or invalid data in a row, and in such instance(s) the entire row(s) should be ignored. Some examples of invalid data can be: negative or zero number of employees and median salary; identical organisation IDs; null/empty values in the required columns. You need to think of other invalid cases yourself.
- The necessary formulas are provided at the end of this document.

Important grading instruction:

Note that you have not been asked to write specific functions. The task has been left to you. However, it is essential that your program defines the top-level function *main(csvfile)* (hereafter referred to as *"main()"* in the project documents to save space when writing it.
Note that when *main()* is written it still implies that it is defined with its input argument). The idea is that within *main(),* the program calls the other functions. (Of course, these functions

may then call further functions.) This is important because when your code is tested on Moodle, the testing program will call your *main()* function. So if you fail to define *main(),* the testing program will not be able to test your code and your submission will be graded zero. Don't forget the submission guidelines provided at the start of this document.

Marking rubric:
Your program will be marked out of 30.

22 out of 30 marks will be awarded automatically based on how well your program completes a number of tests, reflecting normal use of the program, and also how the program handles various states including, but not limited to, different numbers of rows in the input file and / or any error states. You need to think creatively what your program may face. Your submission will be graded by data files other than the provided data file. Therefore, you need to be creative to look into corner or worst cases. I have provided few guidelines from ACS Accreditation manual at the end of the project sheet which will help you to understand the expectations.

8 out of 30 marks will nd efficiency (3/8) "your program is well use of comments, sensible variable nam watch the lectures where this is discusse

Style Rubric:

0	Gibberis
1-2	Style is f
3-4	Style is g
5	Excellen

Your program will be large csv files) so you need to minimise data items.

Efficiency rubric:

0	Code too cor d
1	Very poor ef e()
2	Acceptable or good efficiency with some lapses
3	Excellent efficiency, should have no problem on large files, etc

Automated testing is being used so that all submitted programs are being tested the same way. Sometimes it happens that there is one mistake in the program that means that no tests are passed. If the marker is able to spot the cause and fix it readily, then they are allowed to do that and your - now fixed - program will score whatever it scores from the tests, minus 4 marks, because other students will not have had the benefit of marker intervention. Still, that's way better than getting zero. On the other hand, if the bug is hard to fix, the marker needs to move on to other submissions.

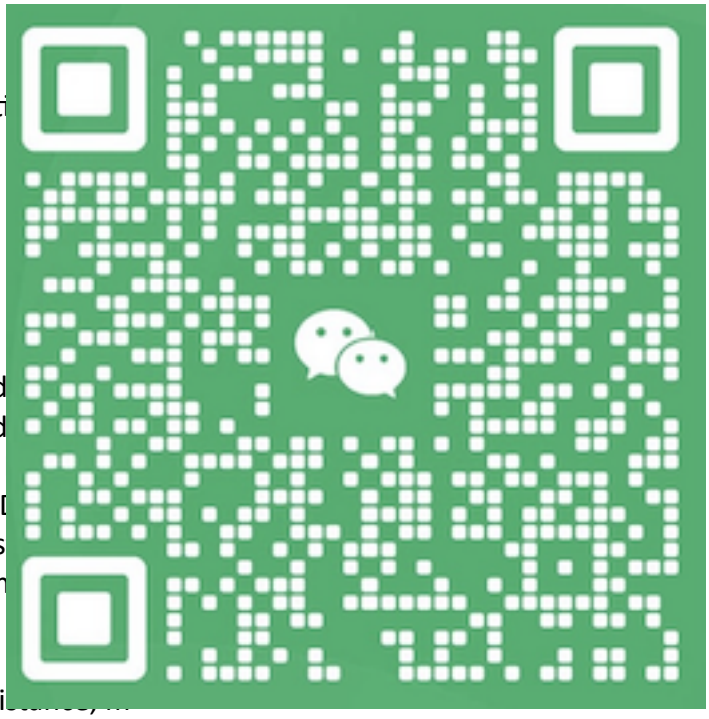
Extract from Australian Computing Society Accreditation manual 2019:

As per Seoul Accord section D, a complex computing problem will normally have some or all of the following criteria:

- involves wide-ranging or conflicting technical, computing, and other issues;
- has no obvious solution, and requires conceptual thinking and innovative analysis to formulate suitable abstract models;
- a solution requires the use of in-depth computing or domain knowledge and an analytical approach that is based on well-founded principles;
- involves infrequently-encountered issues;
- is outside problems encompassed by standards and standard practice for professional computing;
- involves diverse groups of stakeholders with widely varying needs;
- has significant consequences in a range of contexts;
- is a high-level problem possibly including many component parts or sub-problems;
- identification of a requirement or the cause of a problem is ill defined or unknown.

Necessary formulas:

1) t-test calculati



where,

$\bar{x}_1$ = Observed

$\bar{x}_2$ = Observed

s_1 = Standard

s_2= Standard D

n_1 = Size of 1s

n_2 = Size of 2n

2) Minkowski di

$$m = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

where,

x_i = first set of data.

y_i = second set of data.

n = number of samples.

p = a value which controls the level of similarity. Use p =3 in this project.

3) Percentage

Absolute profit change for an organization:

C = |profit in 2020 – profit in 2021|

Percentage = (C / profit in 2020) * 100