Deree
School of Graduate
and Professional
Education

THE AMERICAN COLLEGE OF GREECE
Pierce
Deree
Alba
1875

Module	**ITC 6001 – INTRODUCTION TO BIG DATA**		
Term	**FALL SEMESTER 2023**		
Assessment	**MIDTERM**	**Weight**	**40%**
Duration			
Deliverables	*<u>2 python files</u> one for each question, <u>zipped in one file</u> + <u>data used</u>*		
Method of Submission	*Blackboard*		
Deadline:	*You can consult your notes, blackboard, and the internet. But you cannot use generative AI tools (e.g. Chat-GP)*		

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work **without improper or unauthorized assistance** *from anyone.*

Instructions

Time allocated: 2.5 hours

Open notes exam

Individual work exam

Answer all questions

Exam Instructions:

You have 15 minutes in addition to the exam time, to login in the Virtual Machines and start the relevant tools (e.g., Anaconda, mysql).

For each question: create a python file:

- Q1 → Q1-Lastname-Firstname.py
- Q2 → Q2-Lastname-Firstname.py
- When finished: zip all the files in a file named: “**LastName-id-midTerm-Exam.zip**”
- Submit the file in the blackboard: **/Assignments/Midterm-Fall23**

The files should be able to be executed just running them: Any datafiles you have used have be in the same directory as the python files, no paths should be used.

Coding:

Use python, and related libraries, e.g., Json, csv, Pandas, NumPy and a MySQL (e.g. pymysql) connector. No other framework may be used.

Grading scale: US-Scale:

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work ***without improper or unauthorized assistance*** *from anyone.*

Q1: Python: Pandas Data frames 50%

Instructions for Q1

- *Use Pandas Data frame to read the file “mall.csv”, and then write code in python for the following sub-questions. You will find mall.csv online at the place of submission.*
- *All sub-questions (1-6) should be in the same python file.*
- *For each question, before the code paste the question itself in the python file as a comment*
- *You should define a variable for each of the questions 1,2,3,…,10 named as: result1, result2, result3, …, result10 which should contain the answer and should be printed. E.g.*

```
    # Q1-1. 1. Calculate the min, max and average and the std for
the item_weights

    code  …

    more code ….

    result1=  code …

    print (result1)
```

Q1 work to be done

1. Print the min, max, average & the std for the *Item_Weight*
2. Count the number of items (based on *Item_Identifier* column per outlet type).
3. Print the unique item names (use feature 'Item_Type')
4. Print the unique *Item Identifiers*
5. Print the total number of items sold per *Outlet_Identifier*
6. Print the *item identifier* and *weight* for items heavier than 21.3
7. Compute the following pivot table that counts the number of *items* per *outlet_size*, and per *Tier*. (Hint aggregation function 'count'.). The result should be as follows:

	Item_Identifier		
Outlet_Location_Type	Tier 1	Tier 2	Tier 3
Outlet_Size			
High	NaN	NaN	932.0
Medium	930.0	NaN	1863.0
Small	1458.0	930.0	NaN

8. Print for *Outlet_Establishment_Year* (in rows), and per *Outlet_size* (in columns). You should have the total number of items sold in each cell. The result should be as follows:

	Item_Outlet_Sales		
Outlet_Size	High	Medium	Small

Outlet_Establishment_Year
1985   NaN 935.0 528.0
1987   932.0   NaN   NaN
1997   NaN   NaN 930.0
1999   NaN 930.0   NaN
2004   NaN   NaN 930.0
2009   NaN 928.0   NaN

9. Print the rows where the *Item_type* is meat or *Baking goods*
10. Remove the columns that are marked with outlet_type, and *item_identifier*

Q2: Database: 50%

Go to **/assignments/Midterm-Fall-23**. You will find files: step-1-Create-DB.py and step-2-Populate-DB.py. It contains code to create and populate the DB. The DB contains three tables: courses, coursesInstructors (who teaches what), instructors.

Instructions for Q2 *Create & Populate the database*

(Note: the code has been tested with Spyder, anaconda in Linux)

1. The connector from Python to MySQL is: pymysql
2. Create the database and tables by running the: “step-1-create-DB.py”: you will need to set username and password of your own database
3. Populate the database by running “step-2-populate-DB.py”. Make sure that the directories are right, and you provide the correct password.
4. Inspect the database to make sure that all data have been inserted in the tables (you may need to run a select * command)

Q2: ***work to be done***

Write queries on the database you have created. The queries should be in Python.

- *All sub-questions (1-10) should be in the same python file*
- *For each question, before the code, paste the question itself in the python file as comment.*

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work ***without improper or unauthorized assistance*** *from anyone.*

- *You should define a variable for each of the questions 1,2,3,.. 10, named as: result1, result2, result3, … result10, which should contain the answer and should be printed. E.g.,.*

```
# Display the names of the courses and the term they are
offered

code …

result1= more code

        print (result1)
```

Q3: ***work to be done***

1. Display the names of the courses and the term they are offered.
2. Display the names, and terms for courses that are offered in the fall or winter terms.
3. Count the number of courses offered in each semester and display them as: [term, #courses].
4. Count & display the number of instructors.
5. Display the total number of courses.
6. Count the total number of instructors that are offering course c1 or c2.
7. Display the names of courses that Jones is offering.
8. Get all instructors' names whose name end in "s".
9. Display all course names, which are offered by male instructors.
10. Update the id4 to id44 in table instructors. What is the result you get. Why?