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 2025**		
Assessment	**PROJECT**	**Weight**	**50%**
Duration			
Deliverables	1. *Report in TurnitIn* 2. *Code in Blackboard* 3. *Code in GitHub* 4. *An oral examination/ presentation of your work*		
Method of Submission	*TurinitIn and Blackboard*		
Deadline:	*Last Week of the course*		

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

General Instructions

Your project involves a series of **experiments**, **observations** coming out of the experiments, and drawing **conclusions.** Essentially you will collect data (or they will be provided by the instructor), then a programming language will be used (you are encouraged to use Python, if you intend to use anything else you need to inform the instructor) along with the appropriate libraries to process the data. Tables, diagrammes and data visualizations are essential for presenting your findings.

Deliverables: a) code in blackboard in Python, along with instructions for running it b) a report of 3000±500 words that will present your findings, and will be submitted at Turnit-in. The report must be self-contained, that is all experiments performed and all conclusions should be reported. If you need to exceed the word limit, use an appendix. c) an oral presentation d) code in GitHub
Team size: 4- persons

Grading Peer Marking :

		Person being rated		
		Person-1	**Person-2**	**Person-3**
Person doing the rating	**Person-1**	1.25	1	0.75
	Person-2	1.10	1.10	0.80
	Person-3	1	1	1
Average Rate		**1.12**	**1.03**	**0.85**
Individual score (project grade: 80%)		**89.6**	**82.4**	**68**

Grading: peer-review

Teams have 4 members. Group members will be asked to rate the relative contribution of themselves and the other group member(s). The ratings provided by each member must add up to the number of persons the group consists of (see example above). These ratings will be considered in the final grading of the project for each individual.

Example: In a group consisting of three members, each member provides a rating of all group members. As this is a three-member group, the ratings provided by each member add up to 3.00. A rating of 1.00 means that the person in question did exactly as much as expected of him/her. A rating that is less than 1.00 means that the person in question did less than expected, whereas a rating that is greater than 1.00 means that this person's contribution was greater than expected. Naturally, groups of 4 member have 4 units to allocate to team members.

Coding:

You can use python, and related libraries, e.g. Json, csv, Pandas, NumPy, a library for displaying data, and databases should they be useful. No other framework may be used.

Exploring Global Trends with World Bank Data

You will collect, clean, and analyze a large, real-world dataset related to a global issue. The datasets are provided from the World Bank. You will choose **3** to **6 indicators** from **two** of the following datasets (typically at 3 indicators per data set or topic). There is also the option to choose data sets that are not listed here, but they must be provided by the world bank.

Topic/ data set	**Example indicators**	**URLs** (world bank)
Health	Life expectancy, Mortality rate, Healthcare spending	https://data.worldbank.org/indicator/SP.DYN.LE00.IN
Economy	GDP (current US$), Inflation, Trade balance	https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
Environment	CO_2 emissions, Renewable energy %, Forest area	https://data.worldbank.org/topic/environment
Education	School enrollments, literacy rate	https://data.worldbank.org/indicator/SE.PRM.ENRR
Your choice		World bank

1. **Data harvesting**: **10%**
    a. Download the related CSV data sets that refer to indicators (e.g. GDP, inflation etc.) for each of the two topics
    b. Observe data, report on their size, handle missing data, and observe other possible problems (e.g. you may need to standardize country names).

2. **Database creation: 15%**
    a. Create tables in a database to handle the data you have, e.g.:
        i. countries (country_id, country_name, region). A region comprises multiple countries. The definition of a *reasonable* region is up to you.
        ii. indicators (indicator_id, indicator_name, unit).
        iii. values (country_id, indicator_id, year, value)
    b. write SQL queries to: (if not possible to write a query for one of the next questions, explain the reason, and use Pandas/numpty instead).

i. compute *yearly averages* and *standard deviations* for each indicator per region (the definition of region is to you as explained above).
ii. For one county compute a *10-year rolling average* for all indicators.
iii. identify *top/bottom 10 countries per region* for each indicator over the last 10 years.
iv. Draw remarks / conclusions.

3. Analysis in Pandas / Numpy: 40%

a. In a data set the elements that exhibit unusual behavior are considered as outliers. One way is to detect outliers is to use the z-score, and call outlier whatever is higher than a threshold (e.g., 3).
    i. Detect the **outliers** for each i**ndicator** for each **region** over the last 10 years.
    ii. Then select a **specific region** and get the outliers for each **indicator** for **intervals of 10 years**.

b. Use **covariance matrices** to explore relationships among all features (i.e. the indicators, e.g. GPD per capita and inflation). This has to be performed for **each region** over the **last 10 years**.

c. Build a data set that contains the indicators variables (there should be at least 6 from two different data sets) as columns, and the countries as rows. The cells should correspond to averages over the last 10 years.

    i. Perform non-negative matrix decomposition to facilitate a grouping of the **indicator variables**. What are the groupings produced?

d. **Extra work**: The countries, indicators and years form a 3D matrix, known as Tensor. Perform Tensor decomposition to find interesting patterns.

e. Draw remarks / conclusions.

4. Data Visualization: 10%

a. Create time series plots for the indicators, you will need to specify regions
b. Create heatmaps for correlations, and covariance matrices, you will need to specify regions

5. Presentation: 10%

On the last day of the course: A presentation to summarize the data set that was used, any type of preprocessing, the methodology that was applied, results and conclusion. All team members need to participate in the presentation.

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

6. **Report Quality: 15%**

The report should be self-contained. It should clearly describe the work done and should be split into meaningful sections. The use of tables and diagrammes enhances the readability. You can define any number of geographical regions. But in your report, you should present results that concern between 3 to 6 regions.