# Module 4: Introduction to NumPy, Pandas, and Matplotlib

## Case Study – 2

# Case Study – 2

## Domain – Education

focus – Data analysis

Business challenge/requirement
You are a data analyst with the University of Cal USA (Not a machine learning expert yet as you still have not completed the ML with Python Course :-)). The University has data on Math, Physics, and Data Structure scores of sophomore students. This data is stored in different files. The University has hired a data science company to do an analysis of scores and find if there is any correlation between scores with age, ethnicity, etc. Before the data is given to the company you have to do data wrangling.

Key issues
Ensure student's identity is not revealed to the agency and only relevant data is shared

Considerations
NONE



Data volume
- In thousands, but only around 1800 records are shared in files MathScoreTerm1.csv DSScoreTerm1.csv, PhysicsScoreTerm1.csv

Additional information
- NA

Business benefits
University can get more students enrollment by improving their international ranking through personalized courses/curricula for students

Approach to Solve
You have to use the fundamentals of Numpy and Pandas covered in module 4.

1. Read the three CSV files which contain the score of the same students in term1 of each Subject
2. Remove the name and ethnicity column (to ensure confidentiality)
3. Fill missing score data with zero

4. Merge the three files
5. Change Sex(M/F) Column to 1/2 for further analysis
6. Store the data in a new file – ScoreFinal.csv

Enhancements for code

You can try these enhancements in code

1. Convert ethnicity to a numerical value
2. Fill the missing score for a student to the average of the class