Python For Data Science Project

Project Statement & Requirement

You are a data analyst with BACL. BACL has data of Math, Physics and Data Structure score of sophomore students. This data is stored in different files. BACL has hired a data science company to do analysis of scores and find if there is any correlation of score with age, ethnicity, etc. Before the data is given to the company you have to do data wrangling.

Key issues

Ensure students 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

Business benefits

BACL can get more students enrollment by improving its international ranking through personalized course/curriculum for students

 Approach to Solve

You have to use fundamentals of Numpy and Pandas covered during the Course.

1. Read the three csv files which contains the score of same students in term1 for 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 new file – ScoreFinal.csv

Enhancements for code

You can try these enhancements in code

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