Module 4: Introduction to NumPy & Pandas

Case Study II

Domain – Education (focus – Data analysis)

Business challenge/requirement

You are a data analyst with University of Cal USA (Not a machine learning expert yet as you still have not completed ML with Python Course :-)).The University has data of Math, Physics and Data Structure score of sophomore students. This data is stored in different files. The University 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 identify is not revealed to the agency and only relevant data is shared

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

Business benefits
University 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 in module 4.
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

```
import pandas as pd

# Load the CSV files
math_scores = pd.read_csv('MathScoreTerm1.csv')
physics_scores = pd.read_csv('PhysicsScoreTerm1.csv')

# Display summary information about each dataset
print("Math Scores Summary:")
print(math_scores.describe(include='all'))

print("\nPhysics Scores Summary:")
print(physics_scores.describe(include='all'))

# Remove the Name and Ethnicity columns from math_scores and physics_scores
math_scores = math_scores.drop(columns=['Name', 'Ethinicity'])
physics_scores = physics_scores.drop(columns=['Name', 'Ethinicity'])

# Fill missing score data with zero
math_scores['Score'] = math_scores['Score'].fillna(0)
physics_scores['Score'] = physics_scores['Score'].fillna(0)

# Merge the math_scores and physics_scores on ID, Age, Subject, and Sex
merged_scores = pd.merge(math_scores, physics_scores, on=['ID', 'Age', 'Sex'], suffixes=('_Math',
'_Physics'))

# Change Sex(M/F) column to 1/2 for further analysis
merged_scores['Sex'] = merged_scores['Sex'].map({'M': 1, 'F': 2})

# Store the data in a new file – ScoreFinal.csv
merged_scores.to_csv('MergeScoreFinal.csv', index=False)

print("Processing and merging completed. The final merged file is saved as MergeScoreFinal.csv.")
```

```
[john@squid use-case-II]$
[john@squid use-case-II]$ python3 use-case2-1.py
Math Scores Summary:
                   Name       Score         Age      Ethinicity Subject  Sex          ID
count               599  596.000000  599.000000             599     599  599  599.000000
unique              596         NaN         NaN               4       1    2         NaN
top     MICHAEL THOMPSON         NaN         NaN  White American   Maths    M         NaN
freq                  2         NaN         NaN             288     599  480         NaN
mean                NaN   74.348993   19.121870             NaN     NaN  NaN  300.000000
std                 NaN   16.217918    1.052234             NaN     NaN  NaN  173.060683
min                 NaN   31.000000   18.000000             NaN     NaN  NaN    1.000000
25%                 NaN   65.000000   18.000000             NaN     NaN  NaN  150.500000
50%                 NaN   78.500000   19.000000             NaN     NaN  NaN  300.000000
75%                 NaN   88.000000   20.000000             NaN     NaN  NaN  449.500000
max                 NaN   95.000000   21.000000             NaN     NaN  NaN  599.000000

Physics Scores Summary:
                   Name       Score         Age      Ethinicity  Subject  Sex          ID
count               599  593.000000  599.000000             599      599  599  599.000000
unique              596         NaN         NaN               4        1    2         NaN
top     MICHAEL THOMPSON         NaN         NaN  White American  Physics    M         NaN
freq                  2         NaN         NaN             288      599  480         NaN
mean                NaN   70.598651   19.121870             NaN      NaN  NaN  300.000000
std                 NaN   15.997280    1.052234             NaN      NaN  NaN  173.060683
min                 NaN   27.000000   18.000000             NaN      NaN  NaN    1.000000
25%                 NaN   61.000000   18.000000             NaN      NaN  NaN  150.500000
50%                 NaN   78.000000   19.000000             NaN      NaN  NaN  300.000000
75%                 NaN   84.000000   20.000000             NaN      NaN  NaN  449.500000
max                 NaN   94.000000   21.000000             NaN      NaN  NaN  599.000000
Processing and merging completed. The final merged file is saved as MergeScoreFinal.csv.
[john@squid use-case-II]$
```

MergeScoreFinal.csv > data

```
Score_Math,Age,Subject_Math,Sex,ID,Score_Physics,Subject_Physics
88.0,18,Maths,1,1,84.0,Physics
85.0,19,Maths,1,2,81.0,Physics
45.0,19,Maths,1,3,41.0,Physics
82.0,18,Maths,1,4,78.0,Physics
82.0,18,Maths,2,5,78.0,Physics
95.0,20,Maths,1,6,91.0,Physics
95.0,18,Maths,1,7,91.0,Physics
65.0,19,Maths,1,8,61.0,Physics
88.0,18,Maths,1,9,84.0,Physics
88.0,19,Maths,2,10,84.0,Physics
53.0,20,Maths,1,11,49.0,Physics
53.0,20,Maths,1,12,49.0,Physics
66.0,19,Maths,1,13,62.0,Physics
88.0,18,Maths,1,14,84.0,Physics
88.0,21,Maths,2,15,84.0,Physics
82.0,20,Maths,1,16,78.0,Physics
31.0,18,Maths,1,17,0.0,Physics
95.0,18,Maths,1,18,91.0,Physics
91.0,18,Maths,1,19,87.0,Physics
66.0,18,Maths,2,20,62.0,Physics
82.0,18,Maths,1,21,78.0,Physics
66.0,18,Maths,1,22,62.0,Physics
75.0,18,Maths,1,23,71.0,Physics
65.0,19,Maths,1,24,61.0,Physics
91.0,20,Maths,2,25,87.0,Physics
91.0,18,Maths,1,26,87.0,Physics
53.0,20,Maths,1,27,49.0,Physics
91.0,20,Maths,1,28,87.0,Physics
88.0,21,Maths,1,29,84.0,Physics
85.0,19,Maths,2,30,81.0,Physics
```