CS112-B
Semester Project
Deadline: 15th May 2023
Group Size 3 Persons.

Problem Statement:

In this project, you have been given students data recorded of Lahore and Peshawar campuses of a university (Note: the data is in camouflage).

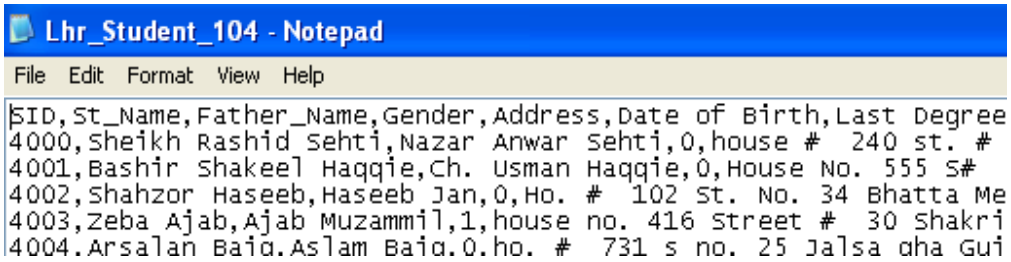
At each campus university has two degree programs BS and MS.

Four disciplines at BS level are:
1) Computer Science (CS)
2) Computer Engineering (CE)
3) System Engineering (SE) and
4) Telecommunication (TC).

Four disciplines at MS level are:
1) Computer Science (MS-CS)
2) Software Project Mgmt. (MS-SPM)
3) Networking (MS-NW) and
4) Telecommunication (MS-TC).

Properties of Students' Data from Lahore Campus

- Data at Lahore campus is stored in Text files
- To store data regarding one complete batch, two text files are used:
  - Lhr_Student_batch (Student record)
  - Lhr_Detail_batch (Course Reg. record)
- You are given a total of 22 text files for 11 BS batches and 8 text files for 4 MS batches



Structure of the Student's data:

- SID: Student ID
  - A numerical value, starting from 0
  - Starts from 0 individually for both degrees BS & MS
  - It is unique within a degree (BS/MS) but not unique across the degrees
  - Combination of SID and degree is always unique within a campus
- St_Name: Student name
- Father_Name: Father name
- Gender:
  - 0 for Male

    - 1 for Female
- Address: Permanent Address
- [Date of Birth]:
    - 14-Apr-1980
- [Reg Date]: Date on which student was enrolled
- [Reg Status]:
    - ‘A’ if student was enrolled as new Admission
    - ‘T’ if student was enrolled as Transfer case
- [Degree Status]:
    - ‘C’ (complete) if student has graduated
    - ‘I’ for incomplete degree
- [Last Degree]:
    - F.Sc. / A level for BS
    - M.Sc. / BS / BE for MS

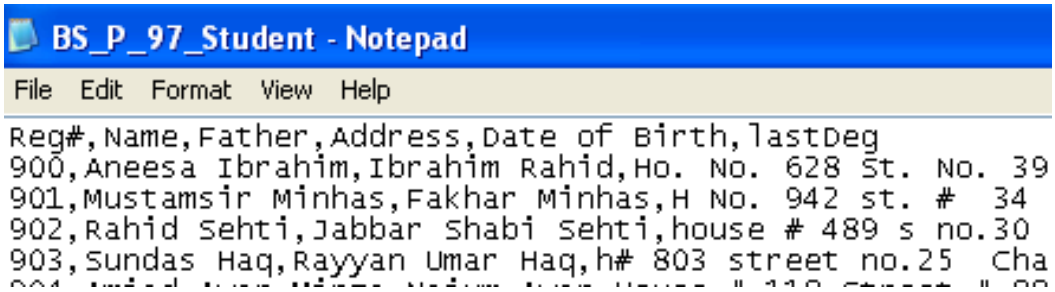
Properties of Courses’ Data from Lahore Campus

SID:
Degree: BS/MS
Semester: e.g. Fall04
Course: Course code
Marks: Out of 100
Discipline: CS/TC/SE/CE

Properties of Students’ Data from Peshawar Campus

Data at Peshawar campus is stored in Text files

- To store data regarding one complete batch 2 text files are used
- Lhr_Student_batch (Student record)
- Lhr_Detail_batch (Course Reg. record)
- You are given 22 text files for 11 BS batches and 8 text files for 4 MS batches



Structure of Peshawar campus students’ data

- TableReg#: Student identity
- Name: Student name
- Father: Father name
- Address: Permanent address
- Date of Birth: Date of Birth
- lastDeg: Last degree achieved
- Reg Date: Date of Enrollment
- Reg Status: Status of Enrollment (A/T)

- Degree Status: Status of Degree (C/I)

Structure of Peshawar campus courses' data

- Reg#:
- Courses: Course code
- Score: Out of 100
- Program: CS/TC/SE/CE
- Sem: Fall/Spring
- Year: YYYY e.g. 1999

Tasks to do:

Create classes Student and Course with appropriate features and functions to store the above mentioned data.

Read the files one-by-one using C++ code and combine the two datasets (Lahore and Peshawar campus data) in a single source.

Once you have the data in the instances of the **Student** and **Course** classes, perform data profiling for all the fields. By data profiling, computation of following statistics is meant:

- No. of unique values (for each column)
- No. of nulls (for each column)
- Invalid values (for each column)
- Total no. of courses
- Total no. of female students Vs male students
- Total no of people who has taken more than 5 courses in a semester.
- The relationship between total no. of unique student ID's and total no. of students.
- Average no of people per semester of each campus.
- Average no of students in every batch of each campus.

After the profiling, you should have identified the anomalies and data cleansing issues.

Here are some of the issues you need to address during your cleansing work by writing intelligent code.

- Separate first and last names both for the student and father. Standardize the first and last names. (Hint: Find all the unique names in data and create a lookup table with two columns i.e. correct_name, variation etc. Use the lookup file to update the name fields with standardized names. Never hardcode names)

- Introduce a new column i.e. city_name. It will involve extracting the city from the address and then the standardization of the city names. (Use the city names in the telephone directory as standard)

- Bring the gender information into a consistent representation. Use 'M' and 'F'

- Since the gender information is missing for the Peshawar campus, you can use the student names to figure out the gender. (Hint: Find all distinct male and female names and create a lookup file/array)

- Validate all the dates against the business rules e.g. the DOB should be smaller than the Reg. Date and Graduation Date should be greater than Reg. Date. The data might be having anomalies like exchanged DOB and Reg. Also, be careful with invalid dates i.e. 31st Feb. or 29th Feb. in a non leap year.

- May be in a campus, the degree information is missing. Devise some technique to figure it out and update records with empty degree fields. Validate other business rules for each field. One example is that marks should be in the range 0 to 100 inclusive.

Develop a GUI-based application for a 10% extra credit.

Enjoy coding!