

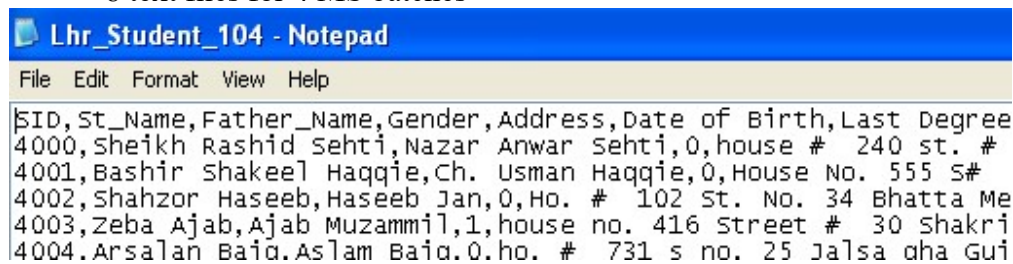
Data Mining
Assignment 1
[Data Cleansing]
Deadline: 17th November 2020 11:59 PM
Group Size 2 Persons max.

Problem Statement:

In this assignment you have been given students data recorded at the Karachi, Lahore and Peshawar campuses of a University. 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).

1. Data from Lahore Campus

- Data at Lahore 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)
- 22 text files for 11 BS batches
- 8 text files for 4 MS batches



```
Lhr_Student_104 - Notepad
File Edit Format View Help
SID,St_Name,Father_Name,Gender,Address,Date of Birth,Last Degree
4000,Sheikh Rashid Sehti,Nazar Anwar Sehti,0,house # 240 st. #
4001,Bashir Shakeel Haqqie,Ch. Usman Haqqie,0,House No. 555 S#
4002,Shahzor Haseeb,Haseeb Jan,0,Ho. # 102 St. No. 34 Bhatta Me
4003,Zeba Ajab,Ajab Muzammil,1,house no. 416 Street # 30 Shakri
4004.Arsalan Baig.Aslam Baig.0.ho. # 731 s no. 25 Jalsa dha Gui
```

Header of Student Table:

- 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

Lahore: Header of Course Reg. Table

SID:

Degree: BS/MS

Semester: e.g. Fall04

Course: Course code

Marks: Out of 100

Discipline: CS/TC/SE/CE

2. Data from Karachi Campus

Data at Karachi campus is stored in MS-Excel books

Three books are maintained

- STUDENT_KHR (Student record)
- Reg_BS_KHR (BS course Reg. record)
- Reg_MS_KHR (MS course Reg. record)

STUDENT_KHR keeps two sheets

- 'BS' for BS students records
- 'MS' for MS students records
-

	A	B	C	D	E
1	St_ID	Name	Father	DoB	M/
2	0	Mahjabeen Paracha	Salim Abdul kareem Paracha	16-Jun-79	F
3	1	Abdul Wasay Awan	Ch. Abdur Rahman Awan	13-Dec-77	M
4	2	Kina Zafar	Zafar Kamal	13-Sep-79	F
5	3	Sultan Jabir Sherazi	Sohail Sherazi	15-Sep-79	M
6	4	Suzanne Noman	Abdul Jabbar Noman	16-Jun-79	F

Header of Student Table:

- St_ID: Student identity
- Name: Student name
- Father: Father name
- DoB: Date of Birth
- M/F: Gender (M/F)
- DoReg: Date of Registration/Enrollment
- RStatus: Status of enrollment (A/T)
- DStatus: Status of Degree (C/I)
- Address: Permanent address

- Qualification: Last degree achieved

Header of Course Reg. Table

- SID:
- Courses: Course code
- Score: Out of 100
- Sem: e.g. Fall04
- Disp: CS/TC/SE/CE

Degree (BS/MS) is missing because separate books are maintained, but the issue is critical while loading data

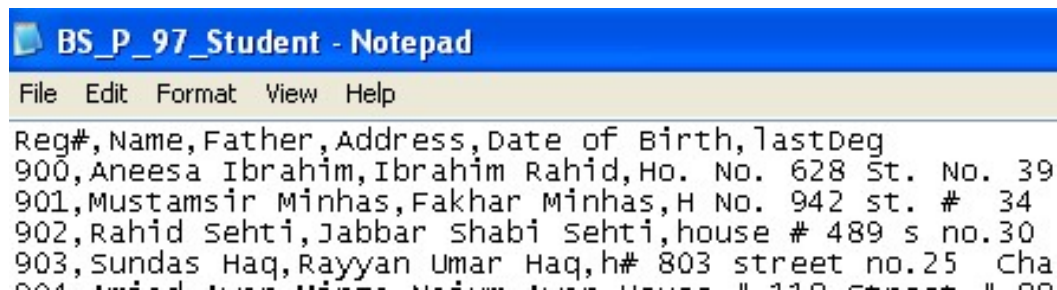
3. Data from Peshawar Campus

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)

22 text files for 11 BS batches

8 text files for 4 MS batches



```
BS_P_97_Student - Notepad
File Edit Format View Help
Reg#,Name,Father,Address,Date of Birth,lastDeg
900,Aneesa Ibrahim,Ibrahim Rahid,Ho. No. 628 St. No. 39
901,Mustamsir Minhas,Fakhar Minhas,H No. 942 st. # 34
902,Rahid Sehti,Jabbar Shabi Sehti,house # 489 s no.30
903,Sundas Haq,Rayyan Umar Haq,h# 803 street no.25 Cha
```

Header of Student

- 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)

Header of Course Reg. Table

- Reg#:
- Courses: Course code
- Score: Out of 100

- Program: CS/TC/SE/CE
- Sem: Fall/Spring
- Year: YYYY e.g. 1999

Tasks to do

1. **Combine all the three datasets in a single source, you may use any DBMS.**
2. Once you have the data in the staging area inside your DBMS, perform data profiling for all the fields. By data profiling means the following statistics:
 - 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.
3. 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.
 - 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 table to update the name fields with standardized names. Never hardcode names in your SQL.)
 - 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' with data type CHAR(1).
 - 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 table)
 - If you have been using VARCHAR or CHAR for date, port the date information into columns with the proper Date data type. The dates should be as per calendar dates.
 - 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. Date by mistake etc. Also be careful with invalid dates i.e. 31st Feb. or 29th Feb. in a non-leap year.
 - For some campuses, the degree information is missing. Devise some technique to figure it out and update the rows 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.

Submission Guidelines:

1. Zero credit for no submission
2. Submit a report on “tasks to do”