National University of Computer and Emerging Sciences, Lahore Campus

	Course:	Data Warehousing and Data Mining	Course Code:	CS409
	Program:	BS(Computer Science)	Semester:	Fall 2018
	Out Date:	10-Oct-2018	Total Marks:	
	Due Date:	22-Oct-2018	Weight:	
	Section	CS	Page(s):	2
	Assignment:	2 (ETL)		

Data Set: The data set for this assignment should be downloaded from piazza.

Note:

- The given data set and data used for evaluation may be different, make sure you handle all the cases.
- You can use programming language of your choice (Python is our recommendation) or SSIS (**Hint:** SSIS has a module named script to add custom C# scripts into your data flow tasks).
- You need to follow object oriented approach and technology best practices for language of your choice (High weightage in evaluation).
- Make sure your source code has proper exception handling with custom written exceptions
- **Your code will checked for plagiarism.**

Task: Load Data from the source file placed at piazza under resources and load the data into destination files (Name destination files appropriately) after performing the following transformations.

De Duplication:

Remove duplicate rows in the source data. No duplicates should be present in destination file. Please note ids do not identify duplicate rows. You need to compare all the features except id to find duplicates.

Date format standardization:

Date of birth has multiple formats e.g. 16th March, 2015 can have these representations.

- 16/03/2015
- 16-03-2015
- 16 03 2015

The destination file must have a standardized descriptive format i.e. 16th March, 2015 in this case.

Derived Attribute Age:

Calculate age of the customer and place it as an attribute in the destination file.

Splitting a single attribute:

Split the column ‘address’ into three columns namely street, colony and city.

Please note that address has two formats. Street, Colony, Lahore and Street Colony Lahore.

Merge Name field:

Your destination file must have only one field Full Name with correct value of name in it.

Decode Gender Field:

Gender has no standard format.

Male has following representations: (1) Male (2) male (3) M (4) m (5) 1

Female has following representations: (1) Female (2) female (3) F (4)f (5) 0

Destination must have only one standardized format i.e. ‘male’ and ‘female’

Delete Inconsistent Rows:

Delete all the rows which do not follow the formats for gender and address.

Summarization:

Each summarization result should be placed in a separate output file.

Find City wise Sum of Ages of Customers. . I.e. The result should have city and sum of ages of customers.

Find Age Wise number of customers. I.e. The result should have age and number of customers column.

Piazza: If you have any queries regarding the assignment please redirect them to piazza.

Good Luck