CSC343

Phase I Report

Dataset and Relational Schema

Shirley Qin 1004555297
Xinyi Huang 1005791047

1.0 Project Domain, Dataset and Investigative Questions

Domain chosen for this project	Data generated in daily life of students, including education, dining, device usage, survey responses, etc.	
Dataset	Dataset Name	Student Life Dataset
	Link to dataset	https://studentlife.cs.dartmouth.edu/dataset.html (original) https://github.com/qinshirl/CSC343.git (modified)
	Relevant information	Education - Courses info - Deadlines - Grades - Piazza usage Daily activity - Sleep - Dining - GPS - Exercise - Mood Survey Response - Loneliness Scale - Flourishing Scale - Perceived Stress Scale
	Learning required for Interpretation of Data	- Conversion of unix timestamps and calculation of time durations for different usage periods - Understand the concepts and scales in the responded Surveys
	Data Cleaning	- Modify and rename attributes for clearer interpretation and easier further relation merges - Remove irrelevant attributes (*such as device_ID for the specific phone types students are using*) - Remove duplicated results - Handle the missing data (example: *data for student deadline lines is missing after the date June 6th, we will be removing the portion after this date*) - Validate the data after completing the processes above
Investigative Questions	1. How does the number of courses taken, deadlines and piazza usages affect the grades of the students	
	2. How does the Mood and loneliness level of students affect their daily activity (*such as sleeping and exercising etc.*) as well as their grades	
	3. How does the number of courses taken, deadlines and grades influence the students' perceived stress and flourishing levels	

Table 1.0 Domain chosen for this project, dataset details, and Investigative questions list

2.0 Schema

2.1 Relational Schema

2.1.1 Relations

- Attend(uid, course_code, num_of_course)
    - A tuple in this relation represents a student with user ID *uid*. All courses taken by this student are listed in *course_code*, and *num_of_course* indicates the total number of courses this student is taking this semester.
- class_info(course_code, location, day, end, start)
    - A tuple in this relation represents a specific course which has a unique *course_code*. The *location*, *day*, *end* and *start* time of this course is also specified. *day* represents the day in a week (ranging from 1 to 5, which represents from Monday to Friday respectively)
- deadlines(uid, day_1, day_2, …, day_71)
    - A tuple in this relation represents a student with user ID *uid. day_1* to *day_71* shows the number of dues for each day
- grades(uid, gpa_all, gpa_13s, cs_65)
    - A tuple in this relation represents a student with user ID *uid. gpa_all* is the overall accumulated grade for this student (ranging from 0 to 4),gpa_13s is the overall grade for this student in spring 2013, and *cs_65* is the course grade for COSC065.
- piazza(uid, days_online, views, contributions, questions, notes, answers)
    - A tuple in this relation represents a student with user ID *uid.* This relation shows the usage of piazza in course COSC065 for each student (indicating the number of *days_online*, *views* of posts, post *contributions*, *questions* posted, *notes* posted and questions *answers* from every student)
- dining_uid(date, time, location, meal)
    - A tuple in this relation represents a meal a student had in a specific time that includes the *date*, *time*, *location* and *meal* type of this meal.
- Sleep_uid(hour, location, rate, resp_time, social)
    - A tuple in this relation represents the sleeping hours and quality of a student in a day. *resp_time* is the timestamp when this sleeping period occurred. *hour* indicates the length of this sleeping period. *location* is where this sleep took place (specified in latitude and longitude). *rate* is the rating for the sleeping quality for

this sleeping period. *social* is the number of times the student had trouble staying awake in class the day before.

- Exercise_uid(exercise, have, resp_time, schedule, walk）
    - A tuple in this relation represents the exercising activity of a student in a day. *exercise* is the hours the student had exercised for in specified time (noted as a timestamp in *resp_time*).
    - *have* is the student’s response to the question *”Did you do vigorous exercise today (don't include walking) such as run, swim, cycle, play a sport”*, 1 indicating yes and 2 indicating no.
    - *schedule* is the student’s response to the question “*If no did you want to but couldn't because of your schedule?*”, 1 indicating yes and 2 indicating no, and if the previous answer is yes then the data will appear to be null for this attribute.
    - *walk* is the student’s response to the question “How long did you walk for today?”. The responses are in a ratio scale from 1 to 5 (*[1]None, [2]<30 mins, [3]30-60 mins, [4]60-90 mins, [5]>90mins*)
- Mood_uid(happyornot, happy, sadornot, sad, location, resp_time)
    - A tuple in this relation represents the mood activity of a student in a day(noted as a timestamp in *resp_time*). *location* is where this mood took place for this student.
    - *happy* and *sad* is the response to questions asking the student if he/she is feeling happy at the moment, 1 indicating yes and 2 indicating no. *happyornot* and *sador not* is the response to questions asking the level of happiness or sadness of the student. The response is in an ordinal scale from 1 to 4 ([*1]a little bit, [2]somewhat, [3]very much, [4]extremely*).

- LonelinessScale(uid, type, Q_1, Q_2, …, Q20)
- PerceivedStressScale(uid, type, Q_1, Q_2, …, Q10)
- FlourishingScale(uid, type, Q_1,Q_2, …, Q_8)
    - A tuple in the 3 relations above represents the survey response from a student with student id *uid*, showing their loneliness scale/ perceived stress scale/ flourishing scale.
    - *type* indicates if it is a pre-survey response or post-survey response (represented as ‘pre’ and ‘post‘, pre being before the term being studied and post being after the term). Attributes *Q_n* (*Q_1, Q_2, …, Q_n*) are the questions asked in the

survey. Responses are in straightforward ordinal scales (such as [*never, almost never, rarely, sometimes, fairly often, very often*] in the PerceivedStressScale relation, or range from 1 to 10 in the FlourishingScale relation)

2.1.2 Integrity Constraints

- Attend[course_code] ⊆ class_info[course_code]
- LonelinessScale[type] ⊆ {‘pre’, ‘post’}
- PerceivedStressScale[type] ⊆ {‘pre’, ‘post’}
- FlourishingScale[type] ⊆ {‘pre’, ‘post’}
- LonelinessScale[Q_1, Q_2,..., Q_20] ⊆ {‘Never’, ‘Rarely’, ‘Sometimes’, ‘Often’}
- PerceivedStressScale[Q_1, Q_2, …, Q10] ⊆ {‘never’, ‘almost never’, ‘rarely’, ‘sometimes’, ‘fairly often’, ‘very often’}

2.2 Data Dictionary

- Attend(uid, course_code, num_of_course)

Attribute	Description	Type	Required	Default
uid	The ID of a student attend the class	text+int	Yes	
course_code	The course code of all course taken by the student	list of (text+int)	Yes	
num_of_course	The number of courses taken by the student	int	No	

- class_info(course_code, location, day, end, start)

Attribute	Description	Type	Required	Default
course_code	The Course Code of a course offered	text+int	Yes	
location	The location of the class	text	Yes	
day	The weekday that the lecture takes place	int	Yes	
end	The time when the class ends	Timestamp	Yes	
start	The time when the class starts	Timestamp	Yes	

- deadlines(uid, day_1, day_2, …, day_71)

Attribute	Description	Type	Required	Default
uid	The ID of a student that has deadline(s)	text+int	Yes	
day_1	The number of class deadlines for the student on March 27, 2013	int	Yes	0
day_2	The number of class deadlines for the student on March 28, 2013	int	Yes	0
…...	The number of class deadlines for the student on …...	int	Yes	0
day_71	The number of class deadlines for the student on June 5, 2013	int	Yes	0

- grades(uid, gpa_all, gpa_13s, cs_65)

Attribute	Description	Type	Required	Default
uid	The ID of a student that has grade(s)	text+int	Yes	
gpa_all	The accumulated GPA of the student	float	Yes	
gpa_13s	The GPA of the student in 2013 spring	float	Yes	
cs_65	The student's grade for COSC 065	float	Yes	

- piazza(uid, days_online, views, contributions, questions, notes, answers)

Attribute	Description	Type	Required	Default
uid	The ID of a student using piazza	text+int	Yes	
days_online	The number of days the student logged in CS65 Piazza class page	int	Yes	0
views	The number of posts the student has viewed	int	Yes	0
contributions	The number of posts, responses, edits, follow ups, and comments to follow ups	int	Yes	0
questions	The number of questions the student has asked	int	Yes	0

notes	The number of notes the student has posted	int	Yes	0
answers	The number of questions the student has answered	int	Yes	0

- dining_uid(date, time, location, meal)

Attribute	Description	Type	Required	Default
date	The date that the dining action recorded	Text	Yes	
time	The time of the dining action recorded	Timestamp	Yes	
location	The location of the dining hall	Text	Yes	
meal	The type of the meal provided	Text	Yes	

- Sleep_uid(hour, location, rate, resp_time, social)

Attribute	Description	Type	Required	Default
hour	The length of this sleeping period	int	Yes	
location	The location where this sleep took place (specified in latitude and longitude)	text	Yes	
rate	The rating for the sleeping quality for this sleeping period	int	Yes	
resp_time	The timestamp when this sleeping period occurred	Timestamp	Yes	
social	The number of times the student had trouble staying awake in class the day before	int	Yes	

- Exercise_uid(exercise, have, resp_time, schedule, walk）

Attribute	Description	Type	Required	Default
exercise	The hours the student had exercised for in specified time	int	Yes	
have	The student’s response to the question	text of an int	Yes	

resp_time	The timestamp when the student did the exercise	int	Yes	
schedule	The student’s response to the question	text of an int	Yes	
walk	The student’s response to the question	text of an int	Yes	

- Mood_uid(happyornot, happy, sadornot, sad, location, resp_time)

Attribute	Description	Type	Required	Default
happyornot	The student’s response to the question with question_id "happyornot" in "Mood"	text of an int	Yes	
happy	The student’s response to the question with question_id "happy" in "Mood"	text of an int	Yes	null
sadornot	The student’s response to the question with question_id "sadornot" in "Mood"	text of an int	Yes	
sad	The student’s response to the question with question_id "sad" in "Mood"	text of an int	Yes	null
location	The location where this survey took place (specified in latitude and longitude)	text	Yes	
resp_time	The timestamp when the survey	int	Yes	

- LonelinessScale(uid, type, Q_1, Q_2, …, Q20)

Attribute	Description	Type	Required	Default
uid	The ID of a student participate in the survey	text+int	Yes	
type	Whether it is a pre or post mental health measure of Loneliness Scale	text	Yes	

Q_1	The content of Question 1 in the survey of measuring Loneliness Scale	text	Yes	
Q_2	The content of Question 2 in the survey of measuring Loneliness Scale	text	Yes	
…...	…...	text	Yes	
Q_20	The content of Question 20 in the survey of measuring Loneliness Scale	text	Yes	

- PerceivedStressScale(uid, type, Q_1, Q_2, …, Q10)

Attribute	Description	Type	Required	Default
uid	The ID of a student participate in the survey	text+int	Yes	
type	Whether it is a pre or post mental health measure of Perceived Stress Scale	text	Yes	
Q_1	The content of Question 1 in the survey of measuring Perceived Stress Scale	text	Yes	
Q_2	The content of Question 2 in the survey of measuring Perceived Stress Scale	text	Yes	
…...	…...	text	Yes	
Q_10	The content of Question 10 in the survey of measuring Perceived Stress Scale	text	Yes	

- FlourishingScale(uid, type, Q_1,Q_2, …, Q_8)

Attribute	Description	Type	Required	Default
uid	The ID of a survey responses	text+int	Yes	
type	Whether it is a pre or post mental health measure of Flourishing Scale	text	Yes	
Q_1	The content of Question 1 in the survey of measuring Flourishing Scale	text	Yes	

Q_2	The content of Question 2 in the survey of measuring Flourishing Scale	text	Yes	
…...	…...	text	Yes	
Q_8	The content of Question 8 in the survey of measuring Flourishing Scale	text	Yes	

2.3 Justification of Design

The structure of the data is translated directly without major changes in relation table files. There are a few aspects of this dataset which are considered to be a good design and therefore do not need further major adjustments. The original dataset is relatively large (approximately 5.0 GB), which includes more than what is needed for the project purposes. Thus, it is not necessary to merge the dataset with other datasets. Moreover, this dataset has great dimensions and measures, such as using SI units and different levels of measurements in its stats (using ordinal and ratio scales), covering various dimensions of the relation. Furthermore, the original dataset includes a data dictionary, which helps with the relabeling process in future data cleaning.

However, minor modifications (such as changing specific survey questions into numbers for shorter attribute names) were made for clearer interpretations and smoother relation merges.