SEEM2460 Introduction to Data Science

Assignment 2

Assignment 2 is due on **4 March, 2020**.
You need to submit your work on the blackboard website.

Question 1 (15 pts)

2019 novel coronavirus (2019-nCoV) is a newly emerged coronavirus with global implications. The attached zip file contains csv files that report the number of cumulative cases reported around the world from Jan 21- Feb 14 (source JHU CSSE https://github.com/CSSEGISandData/COVID-19).
Please choose one data visualization scheme to show what you find in these data and explain why you choose this kind of visualization.

Question 2 (15 pts)

A researcher studies text messages and sleep in teenagers. Consider the following data for text messages sent per day and average hours of sleep for five teenagers.

Texts per day	Hours of sleep
132	4.2
52	8
77	6
115	5.2
209	3.2

(a) Find the sample mean and standard deviation of texts per day and hours of sleep.
(b) Find the correlation coefficient r between texts per day and hours of sleep. Describe what your value of r means.

Question 3 (10 pts)

Consider three relations *Student, Course* and C*ourseEnrollment*, with the schema defined below.

- The *Student* relation stores students' information, including *StudentID*, *StudentName* and *Department*;
- The *Course* relation stores courses' information, including *CourseID*, *CourseName* and *Credit*;
- The *CourseEnrollment* relation stores the information of students taking courses and the grades, and a student may take multiple courses.

Student Relation

StudentID	StudentName	Department
11111	Jane Smith	Math
11112	Mike Green	Music
…	…	…

Course Relation

CourseID	CourseName	Credits
1001	Calculus	3
1002	Physics	3
…	…	…

CourseEnrollment Relation

StudentID	CourseID	Grade
11111	1001	A
11111	1002	B
11112	1001	A-
…	…	…

(a) Write a **SQL statement** for the query: find the names of students from "*Math*" department and also take the "*1001*" course;
(b) Write a **relational algebra expression** for the query: find the names of students who have taken course "*Calculus*";

Bonus (10 point)

Please explain why there is n-1 instead of n in sample standard deviation

$$s = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}}$$

End.