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.

Solution
(10 pts for data visualization plot, 5 pts for explanation.
Note: If you directly copy data visualization picture from Internet or not attach the plot, your scores will be reduced 5 pts)

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.

Solution

(a)

The number of subjects n = 5

Sample mean $\bar{x} = \frac{\sum_i^n x_i}{n}$ (1 pts)

for texts per day, (132+52+77+115+209)/5=117 (1 pts)

for hours of sleep, (4.2+8+6+5.2+3.2)/5=5.32 (1 pts)

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

for texts per day, {[(132–117)^2 +(52–117)^2 +(77–117)^2 +(115–117)^2 +(209–117)^2]/(5–1)}1/2=60.245 (2 pts)

for hours of sleep, {[(4.2–5.32)^2 +(8–5.32)^2 +(6–5.32)^2 +(5.2–5.32)^2 +(3.2–5.32)^2]/(5–1)}1/2=1.831 (2 pts)

(b)

Correlation

$$r = \frac{\sum_i^n \left[\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)\right]}{n-1} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$= \frac{(132-117)(4.2-5.32)+\cdots+(209-117)(3.2-5.32)}{4*60.245*1.831} = -0.936$$

The correlation of -0.936 means the numbers of texts per day of teenagers has a very strong negative correlation with the hours of sleep the teenagers have. Simply speaking, this study gets a result that the more text messages a teenager sends per day, the fewer hours of sleep this teenager has. (4 pts for calculation, 3 pts for explanation)

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; (5 pts)

(b) Write a **relational algebra expression** for the query: find the names of students who have taken course "*Calculus*"; (5 pts)

Solution:
(a)
Select *StudentName*
from *Student*, *CourseEnrollment*
where *Student.StudentID* = *CourseEnrollment.StudentID* **and** *CourseEnrollment. CourseID* = "1001" **and** *Student.Department* = "Math" (5 pts)
(b)
$\Pi_{\text{StudentName}}(\sigma_{\text{CourseName="Calculus"}}$ (Student ⋈ CourseEnrollment⋈ Course)) (5 pts)

Bonus (10 point)

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

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

Solution:
We should divide by n−1 because dividing by n would give us a biased estimator of the population variance. One intuitive way to think about why the bias exists is to notice that we generally don't actually know the true population mean μ, and therefore the sample variance is being computed using the estimated mean $\bar{x}$. However the quadratic form $\sum_i^n (x_i - a)^2$ is actually minimized by a=$\bar{x}$, which means that whatever the true population mean μ is, we will always have

$$\sum_i^n (x_i - \mu)^2 \geq \sum_i^n (x_i - \bar{x})^2$$

Therefore we are underestimating the true variance because we don't know the true mean.

Mathematical Proof:

$$
\begin{aligned}
E[\sum_{i=1}^{n}(x_i - \bar{x})^2] &= E[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2] \\
&= E[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2] \\
&= nE[x_i^2] - nE[\bar{x}^2] \\
&= n(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) \\
&= (n-1)\sigma^2
\end{aligned}
$$

So dividing by n-1 gives us an unbiased estimate for the population variance.
(5pts for explanation, 5pts for mathematical proof or numerical example)

End.